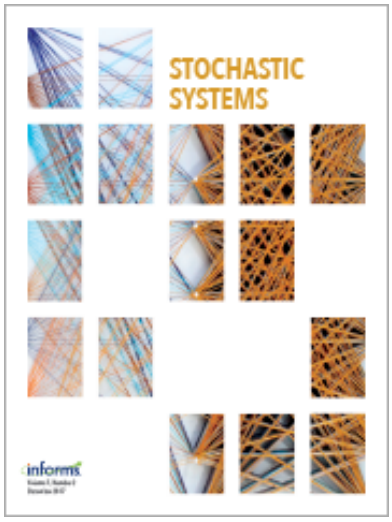


Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Delay-Minimizing Capacity Allocation in an Infinite Server-Queueing System

Refael Hassin, Liron Ravner

To cite this article:
Refael Hassin, Liron Ravner (2019) Delay-Minimizing Capacity Allocation in an Infinite Server-Queueing System. Stochastic Systems 9(1):27-46. https://doi.org/10.1287/stsy.2018.0020

http://pubsonline.informs.org/journal/stsy/

STOCHASTIC SYSTEMS
Vol. 9, No. 1, March 2019, pp. 27–46
ISSN 1946-5238 (online)

Delay-Minimizing Capacity Allocation in an Infinite Server-Queueing System

Refael Hassin,[a] **Liron Ravner**[a]

[a] Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 6997801, Israel
Contact: hassin@post.tau.ac.il, https://orcid.org/0000-0001-9631-6162 (RH); lravner@post.tau.ac.il, https://orcid.org/0000-0002-7359-9793 (LR)

Received: May 28, 2017
Revised: February 21, 2018; June 21, 2018
Accepted: August 20, 2018
Published Online in Articles in Advance: March 25, 2019

MSC2010 Subject Classification: 60K25, 68M20

https://doi.org/10.1287/stsy.2018.0020

Abstract. We consider a service system with an infinite number of exponential servers sharing a finite service capacity. The servers are ordered by their speed, and arriving customers join the fastest idle server. A capacity allocation is an infinite sequence of service rates. We study the probabilistic properties of this system by considering overflows from subsystems with a finite number of servers. Several stability measures are suggested and analyzed. The tail of the series of service rates that minimizes the average expected delay (service time) is shown to be approximately geometrically decreasing. We use this property to approximate the optimal allocation of service rates by constructing an appropriate dynamic program.

History: Former designation of this paper was SSY-2017-528.

Funding: This research was supported by the Israel Science Foundation [Grant 355/15].

Keywords: queues • capacity allocation • heterogeneous servers • infinite state dynamic programming

1. Introduction

We are interested in the optimal allocation of service rates in a system with an infinite number of parallel servers and finite service capacity. Customers join the fastest server available, without jockeying if a faster server becomes available at a later time. This model is appropriate for a system with servers at different physical locations and no possibility to accommodate waiting customers. For example, this may be the case in a distributed (cloud) computing system with jobs arriving at a central router that immediately sends them to the best available server. Other applications with ordered entry and heterogeneous servers are conveyor and storage systems. In this setting an allocation is given by an infinite sequence of service rates. Every such service-rate sequence determines the probabilistic traits of the system. Our objective is to minimize the stationary expected delay (service time) faced by arrivals.

From a practical point of view the infinite-server setting is of course aimed to be a good approximation of large-scale systems. It is important to note that in this respect the model presented here does indeed capture the behavior of such systems when the capacity allocation is a "sensible" one. In particular, for service-rate sequences that satisfy certain stability and delay conditions that will be defined in the next sections, the blocking probability from finite subsystems goes to zero very fast with the number of servers. This means that for even a moderately sized system the probability of an arrival facing a full system is negligible. We show that the blocking probability decreases exponentially with the number of servers. Moreover, we provide a framework that enables the approximation of the blocking probability for a large number of servers, and thus for the required number of servers for the probability to be smaller than a given threshold.

The model is an ordered GI/M/∞ system: independent and identically distributed interarrival times, exponential service times with heterogeneous rates and customers routed to the fastest idle server upon their arrival. An in-depth analysis of an M/M/∞ with identical servers that are ordered (geographically), including heavy traffic approximations, can be found in Newell (1984). The assumption that servers are identical implies that the service capacity is infinite and the focus in Newell (1984) is on distributional properties such as how many of the first n servers are busy. The finite-capacity system studied in this paper is very different, and the analysis relies on the probability of blocking and overflows from finite server subsystems. In particular, our methods rely on Takacs (1959), where the overflow distribution in a homogenous multiserver system with non-Markovian arrivals is characterized, and the extensions of Tahara and Nishida (1975) and Yao (1987) that account for heterogeneous

servers. A key feature for our analysis is that blocking probabilities can be written as a product of Laplace-Stieltjes transform (LTS) corresponding to the overflow times from subsets of the system. We leverage this structure to derive the expected delay in our infinite server system. Even without a queue, the number of busy servers is potentially unbounded, whereas the output rate of the system is bounded by the finite capacity, unlike in typical infinite server settings, where the output rate grows with the number of busy servers. Consequently, an important issue that arises in the infinite server model with finite service capacity is that the stability of the system depends on the allocation of service rates, and it is not enough to assume that the external arrival rate is lower than the total capacity (i.e., $\rho < 1$). This is because the system is not work conserving in the sense that fast servers may be idle while customers are being served by slower servers. Intuitively, the service allocation needs to balance between fast rates at the good (fast) servers while still leaving enough capacity to handle overflows to the bad (slow) servers. These issues are addressed in detail in Section 3, where conditions for stability and finite expected delay are established. In particular, we show that the service-rate sequence cannot decay faster than a geometric series with a decay parameter that is determined by the overflow probabilities.

The trade-off between the rate of capacity assignment, that is, how much of the remaining capacity is assigned to a server when sequentially allocating from the fastest to the slowest, and the overflow probabilities is also at the core of the delay-minimization problem. While the input of the problem is quite simple: the interarrival distribution (a single parameter if the arrival process is Poisson), the decision variable is an infinite sequence. This leads to analytical as well as computational challenges. We formulate the optimal capacity allocation problem as an infinite dynamic program. However, the dynamic program is intractable because of the elaborate state and actions spaces. To this end, we derive an asymptotically optimal geometric tail of the service-rate sequence. The asymptotic optimal geometric rate is shown to be the square-root of the term in the product representing the aforementioned overflow (blocking) probabilities. Furthermore, we use the geometric approximation of the tail to define a finite dynamic program that can be solved efficiently. Numerical analysis suggests that the optimal service-rate sequence is very close to geometric from the start. This means that instead of solving the original capacity allocation problem we can approximate the solution by the single parameter problem of finding the optimal geometric service-rate sequence. Moreover, in the special case of a Poisson arrival process the simple heuristic of choosing a geometric service-rate sequence with decay rate $\sqrt{\rho}$ is quite close to the approximate optimal solution.

The use of stochastic queueing models in order to model cloud computing, also known as distributed or parallel computing, is very common (e.g., Khazaei et al. 2012, Vilaplana et al. 2014). For example, Khazaei et al. (2012) uses a $M/G/m/m+r$ queueing system to approximate the performance. Our model is related to theirs in the case of $r = 0$. They state that a common assumption in cloud computing models is that some positive blocking probability has a predefined upper bound constraint, and our model can be conveniently used to approximate such a system. The aforementioned papers, and many others that use stochastic queueing models for cloud computing, assume homogeneous servers. However, most cloud computing systems do, in fact, have heterogeneous servers (see, e.g., Balsamo et al. 1998, Zreikat et al. 2003, Khazaei et al. 2013). Furthermore, the *Fastest Server First* is a common policy implemented in such systems (see Zreikat et al. 2003). In this paper we present a framework that allows for the analysis of constrained capacity allocation in large scale heterogeneous server systems. Another useful aspect of our model is that it provides tools to study the trade-off between blocking probabilities and expected delay in finite server systems. Other relevant applications of our model are large scale conveyor and storage systems (see Yao 1987 for a discussion).

The research of ordered service systems with heterogeneous servers has mostly focused on analysing the blocking probability in loss systems, and their minimization in particular. In Tahara and Nishida (1975) it was shown that the optimal allocation of service rates in terms of minimizing blocking in an ordered Markovian system is heterogeneous. Nath and Enns (1981) showed that the optimal sequence of service rates, in terms of minimizing blocking probabilities, is decreasing. Analysis of an ordered system with a general arrival process, along with the comparison methods for different entry order regimes, can be found in Shanthikumar and Yao (1987) and Yao (1987). An interesting observation made in Cooper and Palakurthi (1989) is that the policy of *Fastest Server First* is not necessarily optimal, for example, when the slower server has a lower variance of service time. Optimal assignment to an ordered system with heterogeneous customer types that can only be served by some of the servers was studied in Ross (2014). The work presented here is related to Hassin et al. (2015) which analyzed the capacity allocation and pricing in a loss system possibly with heterogeneous servers. The objective function considered in that paper is different from the others because the objective is maximizing profit and not minimizing blocking probabilities. This objective required analysis of the expected waiting times, which also will be important in the analysis presented here. In Rosberg and Makowski (1990), routing policies were analyzed with the goal of minimizing holding costs. Approximation analysis of ordered homogeneous-server systems can be found in Coffman et al. (1985) and Knessl (2004), and heteregenous servers with a single queue (including a waiting buffer)

under the *Fastest Server First* policy appeared in Armony (2005). Another related work is Atar et al. (2013) that considered a service capacity allocation problem for a system of parallel queues and heterogeneous customer types using a heavy-traffic approximation.

In Section 2 we present the model and mention some of its known properties. We define system stability along with necessary conditions for finite expected delay in Section 3. In Section 4 we introduce the special class of service-rate sequence that decreases geometrically. We prove that the tail of the optimal service-rate sequence is of this type. This fact is due to the product form of the blocking probabilities. In Section 5 we formally define our optimization problem as an infinite horizon dynamic program and suggest a numerical method to approximate its solution using the fact that the tail of the optimal sequence is geometric. We then proceed to present numerical analysis and examples of the optimal service-rate sequence in Section 6. The numerical results suggest that the optimal service-rate sequence is very close to geometric from the start, and not just at the tail. Finally, Section 8 features concluding remarks and a brief discussion of straightforward extensions of our analysis aimed at optimizing other performance measures of the system, apart from expected delay.

2. Model

Customers arrive at a service system according to a renewal process with mean interarrival time $ET_0 = \frac{1}{\lambda}$. The system is comprised of an infinite number of parallel exponential servers that are ordered according to service-rate; $\mu_1 \geq \mu_2 \geq \cdots$, such that $\sum_{n=1}^{\infty} \mu_n = \mu > \lambda$. Every arriving customer joins the fastest server available and does not switch servers even if a faster server later becomes available while he is still in the system. For a given μ, our goal in this paper is to find a sequence of service rates that minimizes the stationary expected sojourn time (delay) of an arriving customer.

Let $\mathbf{X} = (X_1, X_2, X_3, \ldots)$ be the random sequence of server indicators, zero if idle and one if busy, at arrival times in the limit. The state space of the process can be defined as follows:[1]

$$\mathbf{X} \in \mathcal{S} = \bigcup_{n=1}^{\infty} A_n,$$

where $A_n = \{x \in \{0,1\}^{\infty} : n = \sup\{j : x_j = 1\}\}$. In words, A_n is the set of all states such that the highest indexed busy server is n. Note that $\mathcal{S}$ is a countable collection of finite sets and is therefore countable. The underlying continuous-time process is not Markovian, due to the general arrival distribution, however the embedded process at arrival moments is indeed a discrete-time Markov chain. The state space should not be confused with the uncountable set $\{0,1\}^{\infty}$ that includes states with infinitely many ones. This space is "too big" as the probability of the process being in a state $s \in \{0,1\}^{\infty}$ with an infinite number of ones is zero for any finite time, much like the queue length process of a single-server queue that is defined on the set positive integers $\mathbb{Z}^+$ and not $\mathbb{Z}^+ \cup \{\infty\}$.

Remark 1. The random variables and distributional properties discussed in this work are all with respect to the limiting distribution $\mathbf{X}$ at arrival times, which is also the stationary distribution if the process is ergodic. This distribution may be different from the limiting time average distribution and the analysis does not require Poisson Arrivals See Time Averages. Furthermore, all random variables depend on the service-rate sequence $\{\mu_n\}_{n=1}^{\infty}$, but we omit this from the notations for the sake of brevity.

Let $Y := \inf\{i : X_i = 0\}$ denote the random variable of the fastest available server, according to the limiting distribution at arrival times. Further denote by S the respective limiting delay (service time) faced by an arbitrary arriving customer. The expected delay is the expected service time at the fastest idle server upon arrival:

$$ES = \sum_{n=1}^{\infty} \frac{1}{\mu_n} \Pr(Y = n).$$

We will soon argue that this limit is well defined even when there is no stationary distribution for the underlying process, in which case ES is infinite. Specifically, each probability $\Pr(Y = n)$ is derived from the limiting distribution of a Markov chain with a finite state space, and, therefore, the infinite sequence is well defined and so is the sum.

The state of any server $n \geq 1$ only depends on the arrival process and on the service process at servers $i \leq n$. For example, if the arrival process is Poisson, then, by viewing, the first server is an isolated M/M/1/0 system:

$$\Pr(Y = 1) = \Pr(X_1 = 0) = \frac{\mu_1}{\lambda + \mu_1}.$$

The distribution of Y is obtained from the blocking probabilities of consecutive subsystems:

$$\Pr(Y = n) = \Pr(X_i = 1 \ \forall i \leq n-1) - \Pr(X_i = 1 \ \forall i \leq n), \ n \geq 2, \tag{1}$$

where $\Pr(X_i = 1\ \forall i \le n)$ is the blocking probability in a GI/M/n/n system with heterogeneous ordered servers. In the following analysis we use the more compact notation:

$$p_n := \Pr(X_i = 1\ \forall i \le n), \quad n \ge 1,$$
$$q_n := \Pr(Y = n), \quad n \ge 1.$$

We are interested in computing (1), which can be rewritten as $q_n = p_{n-1} - p_n$. If the servers are homogeneous then the well-known Erlang Loss Formula can be applied for the blocking probabilities, but this is not possible for an infinite server system with finite capacity. Otherwise, the blocking probabilities are given in Yao (1987):

$$p_n = \prod_{i=1}^{n} L_{i-1}(\mu_i) = L_{n-1}(\mu_n)p_{n-1}, \tag{2}$$

where $p_0 := 1$, $L_0(s)$ is the LST of the exogenous interarrival distribution, and

$$L_n(s) = \frac{L_{n-1}(\mu_n + s)}{1 - L_{n-1}(s) + L_{n-1}(\mu_n + s)}, \quad n \ge 1, \tag{3}$$

is the LST of the stationary time between overflows at server n, T_n. Specifically, T_n is the time between two consecutive arrivals that find the first n servers busy, and recall that T_0 is the external interarrival time. The recursive formula (3) generally relies on a Palm-type theorem for the renewal process of overflows at station n, that is, the probability at overflow times as opposed to the time-average distribution which is different as the counting process of overflows is not Poisson. The original result for homogeneous servers appeared in Takacs (1959) (see also Riordan 1962, p. 37). A generalization to heterogeneous service rates appeared in Tahara and Nishida (1975) and a similar model with an additional waiting buffer for customers that find all servers busy upon arrival was studied by Cooper (1976).

The derivative of (3) is

$$L_n'(s) = \frac{L_{n-1}'(\mu_n + s)(1 - L_{n-1}(s)) + L_{n-1}'(s)L_{n-1}(\mu_n + s)}{(1 - L_{n-1}(s) + L_{n-1}(\mu_n + s))^2}, \tag{4}$$

and thus the mean time between overflows from server n is

$$\mathrm{ET}_n = -L_n'(0) = \frac{\mathrm{ET}_{n-1}}{L_{n-1}(\mu_n)} = \frac{\mathrm{ET}_{n-2}}{L_{n-2}(\mu_{n-1})L_{n-1}(\mu_n)} = \cdots = \frac{\mathrm{ET}_0}{p_n},$$

yielding

$$\mathrm{ET}_n = \frac{1}{\lambda p_n}, \quad n \ge 1. \tag{5}$$

Next, we state a technical lemma that will be useful in the following analysis. We do not prove the lemma as these properties are straightforward extensions or rephrasing of known results.

Lemma 1. *The functions $L_i(s)$ satisfy the following properties:*

a. *$L_n(s)$ is strictly decreasing with s, $L_n(0) = 1$ and $\lim_{s\to\infty} L_n(s) = 0$.*

b. *$L_{n-1}(s) > L_n(s)$ for any $s > 0$ and $n \ge 1$ (proof in Yao 1987).*

c. *If $\mu_1 \ge \mu_2 \ge \cdots$, then the decreasing sequence p_n is convex in the discrete sense: $p_{n-1} + p_{n+1} > 2p_n$, $\forall n \ge 2$ (proof in Yao 1986).*

Remark 2. In the following sections, we use the notation $a \wedge b := \min\{a, b\}$. We will also make frequent use of the notation $a_n \approx b_n$ to indicate that the two sequence, $\{a_n\}$ and $\{b_n\}$, have the same tail behavior. Formally, this means that there exists a constant $0 < C < \infty$ such that

$$\lim_{n\to\infty} \frac{a_n}{b_n} = C,$$

and when used for the limits themselves it means that they are proportional:

$$\frac{\lim_{n\to\infty} a_n}{\lim_{n\to\infty} b_n} = C.$$

We use $a_n \ll b_n$ to indicate that $C = 0$. In the numerical analysis we will use $\approx$ when numerical results are close (in a nonaccurate sense) to some value.

3. System Stability

In an infinite-server system with infinite capacity, as the homogeneous server GI/M/∞ model, the system is always stable. In many queueing systems with finite capacity the queue-length process is stable if $\rho < 1$, in the sense that the number of customers in the system does not explode (and the underlying embedded Markov process is positive recurrent). However, this is not sufficient for our model because the service allocation may cause the effective arrival rate to a subset of the system to exceed its service capacity. For example, if the arrival process is Poisson, then $p_1 = L_0(\mu_1) = \frac{\lambda}{\lambda+\mu_1}$. Furthermore, if $\mu_1 = \mu - \epsilon$, then the effective arrival rate to the system excluding server 1 is $\lambda p_1 = \lambda \frac{\lambda}{\lambda+\mu-\epsilon}$, which is larger than ϵ when ϵ is chosen to be small enough.

This paper does not directly address the issue of positive recurrence of the process $\mathbf{X}$ at arrival times, which seems to require a different approach than the blocking probability and delay computations employed here. Rather, we define two different levels of stability: the first simply states that all subsystems have a greater capacity than their effective arrival rate, and the second is finite expected delay, a condition that may not hold even if the underlying process is positive recurrent.

As a first reasonable condition, and as we will show in Proposition 2, one that is also necessary for finite expected delay, we would like a service-rate sequence $\{\mu_n\}_{n=1}^{\infty}$ to satisfy

$$\lambda p_n < \mu - \sum_{i=1}^{n} \mu_i = \sum_{i=n+1}^{\infty} \mu_i, \quad \forall n \geq 1. \tag{6}$$

That is, the effective arrival rate into the system excluding the first n server is smaller than the remaining service capacity. In the memoryless arrival example, the first condition for $n = 1$ is $\lambda \frac{\lambda}{\lambda+\mu_1} < \mu - \mu_1$ or, equivalently,

$$\mu_1 \in \left(0, \frac{1}{2}\left(\mu - \lambda + \sqrt{(\mu-\lambda)(\mu+3\lambda)}\right)\right).$$

Definition 1. A service-rate sequence $\{\mu_n\}_{n=1}^{\infty}$ is feasible if it satisfies condition (6).

Denote the set of feasible service-rate sequence by

$$\mathcal{M} := \left\{\{\mu_n\}_{n=1}^{\infty} : \quad \lambda p_n < \sum_{i=n+1}^{\infty} \mu_i, \ \forall n \geq 1\right\}.$$

Proposition 1. *For any $\lambda \in (0, \mu)$ there exists a feasible service-rate sequence ($\mathcal{M} \neq \emptyset$) that is decreasing and satisfies $\sum_{n=1}^{\infty} \mu_n = \mu$.*

Proof. If $\lambda < \mu$, then the range for μ_1 is given by

$$\mu_1 < \mu - \lambda p_1 = \mu - \lambda L_0(\mu_1).$$

Let m_1 be the solution to $\mu_1 = \mu - \lambda L_0(\mu_1)$. Recall that $L_0(0) = 1$ and that $\lambda < \mu$. Therefore, because L_0 is an LST and hence convex, there exists a unique solution $m_1 > 0$. This argument is illustrated in Figure 1. Any point μ_1 in the interior of the interval $(0, m_1)$ satisfies the stability condition for $n = 1$, in particular $\tilde{\mu}_1 = \alpha m_1$, for any $\alpha \in (0, 1)$.

Suppose that $\tilde{\mu}_1, \ldots, \tilde{\mu}_n$ is a decreasing sequence that satisfies (6), then by applying the product form of (2) we have that condition (6) is satisfied for $n + 1$ if

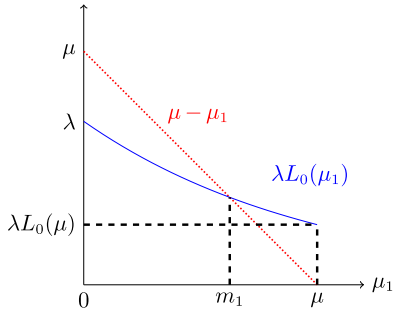
$$\lambda L_n(\mu_{n+1}) p_n < \mu - \sum_{i=1}^{n} \mu_i - \mu_{n+1},$$

or, equivalently,

$$0 < \mu_{n+1} < \mu - \sum_{i=1}^{n} \mu_i - \lambda p_n L_n(\mu_{n+1}).$$

Let $m_{n+1} > 0$ be the unique solution to $\mu_{n+1} = \mu - \sum_{i=1}^{n} \mu_i - \lambda p_n L_n(\mu_{n+1})$. Repeating the argument illustrated in Figure 1, the solution is unique and positive because $L_n(\mu_{n+1}) \in [0, 1]$ is convex and condition (6) is satisfied for n. We thus conclude that any $\mu_{n+1} \in (0, m_{n+1})$ satisfies (6). In particular, for any $\alpha \in (0, 1)$ $\tilde{\mu}_{n+1} := \alpha(m_{n+1} \wedge \mu_n)$ is non-increasing, feasible and $\sum_{n=1}^{\infty} \tilde{\mu}_n \leq \mu$. □

Figure 1. Illustration of the Feasibility Interval $(0, m_1)$ for μ_1



Recall that regardless of whether the service-rate sequence is feasible, the subsystem of the first n servers is ergodic for every finite n. Hence, the limit probabilities p_n and q_n exist for any service-rate sequence. Further observe that if condition (6) is satisfied for some N, then it is satisfied for all $n < N$ as well, as if this was not the case, that is, $\lambda L_{n-1}(0)p_{n-1} > \mu - \sum_{i=1}^{n-1}\mu_i$, then there is no $\mu_n > 0$ such that $\lambda L_{n-1}(\mu_n)p_{n-1} = \mu - \sum_{i=1}^{n-1}\mu_i - \mu_n$ (consider Figure 1 for the case that the solid convex overflow rate line starts above the dotted linear capacity allocation line).

Lemma 2. *For any service-rate sequence* $\{\mu_n\}_{n=1}^{\infty}$ *such that* $\mu_n > 0$ *for all* $n \geq 1$ *and* $\sum_{n=1}^{\infty}\mu_n = \mu$, *there exists a limit*

$$\ell := \lim_{n\to\infty} L_{n-1}(\mu_n) \in (L_0(\mu), 1].$$

Proof. Iterating the recursion of (3) yields

$$\begin{aligned} L_{n-1}(\mu_n) &= \frac{L_{n-2}(\mu_n + \mu_{n-1})}{1 - L_{n-2}(\mu_n) + L_{n-2}(\mu_n + \mu_{n-1})} \\ &= \frac{L_{n-3}(\mu_n + \mu_{n-1} + \mu_{n-2})}{[1 - L_{n-3}(\mu_n + \mu_{n-1}) + L_{n-3}(\mu_n + \mu_{n-1} + \mu_{n-2})][1 - L_{n-2}(\mu_n) + L_{n-2}(\mu_n + \mu_{n-1})]} \\ &= \frac{L_0\left(\sum_{i=1}^{n}\mu_i\right)}{\prod_{i=1}^{n-1}\left[1 - L_{i-1}\left(\sum_{k=i+1}^{n}\mu_k\right) + L_{i-1}\left(\sum_{k=i}^{n}\mu_k\right)\right]}. \end{aligned} \tag{7}$$

By Lemma 1a each term in the product in the denominator is smaller than one. Therefore, for any $n \geq 1$, $L_{n-1}(\mu_n) \geq L_0\left(\sum_{i=1}^{n}\mu_i\right)$. The lower and upper bounds are obtained using Lemma 1a: first, the fact that the LST is a decreasing function yields $L_{n-1}(\mu_n) \geq L_0\left(\sum_{i=1}^{n}\mu_i\right) \geq L_0(\mu)$, and, furthermore, every every term in the sequence $L_{n-1}(\mu_n)$ is bounded from above by 1. Hence,

$$L_0(\mu) \leq L_{n-1}(\mu_n) \leq 1.$$

By the continuity of L_0, $\lim_{n\to\infty} L_0\left(\sum_{i=1}^{n}\mu_i\right) = L_0(\mu)$, because $\sum_{i=1}^{\infty}\mu_i = \mu$. Let

$$a_{in} := 1 - L_{i-1}\left(\sum_{k=i+1}^{n}\mu_k\right) + L_{i-1}\left(\sum_{k=i}^{n}\mu_k\right),$$

then

$$L_{n-1}(\mu_n) = \frac{L_0\left(\sum_{i=1}^{n}\mu_i\right)}{\prod_{i=1}^{n-1}a_{in}} \leq 1,$$

which implies that the product in the denominator does not converge to zero. If the limit $a := \lim_{n\to\infty}\prod_{i=1}^{n-1}a_{in} > 0$ exists, then

$$\ell = \lim_{n\to\infty} L_{n-1}(\mu_n) = \frac{L_0(\mu)}{a}.$$

We will verify the existence of the limit a in three steps as outlined below:

1. We show that $a_{in} \in (0,1]$ is increasing with n and therefore has a limit $a_i := \lim_{n\to\infty} a_{in}$.
2. Let $b_{in} := |\log(a_{in})|$, then $b_{in} \in [0,\infty)$ is decreasing with n and has a limit $b_i := \lim_{n\to\infty} b_{in}$.
3. The sum $\sum_{i=1}^{n-1} b_{in} = -\log\left(\prod_{i=1}^{n-1} a_{in}\right)$ converges to a limit $b < \infty$, and, therefore, the product $\prod_{i=1}^{n-1} a_{in}$ converges to a limit $a = e^{-b}$.

Step 1. Because $\mu_n > 0$ for all $n \geq 1$ then the convexity of L_{i-1} implies that

$$L_{i-1}\left(\sum_{k=i+1}^{n} \mu_k\right) - L_{i-1}\left(\sum_{k=i}^{n} \mu_k\right) > L_{i-1}\left(\sum_{k=i+1}^{n+1} \mu_k\right) - L_{i-1}\left(\sum_{k=i}^{n+1} \mu_k\right) > 0, \quad \forall 1 \leq i \leq n < \infty.$$

Hence, a_{in} is increasing with n and bounded by one, and, thus, there exists a limit:

$$a_i := \lim_{n\to\infty} a_{in} = 1 - L_{i-1}\left(\sum_{k=i+1}^{\infty} \mu_k\right) + L_{i-1}\left(\sum_{k=i}^{\infty} \mu_k\right).$$

Step 2. Let $b_{in} := |\log(a_{in})|$ and $b_i := \lim_{n\to\infty} b_{in}$, and observe that $\log(a_{in}) \leq 0$ as $a_{in} \in [L_0(\mu), 1]$. The sum $\sum_{i=1}^{n-1} b_{in} = -\log\left(\prod_{i=1}^{n-1} a_{in}\right)$ is bounded as $n \to \infty$ because the product does not converge to zero. Moreover, as a_{in} is increasing with n, $b_{in} = |\log(a_{in})|$ is decreasing and $b_{in} \geq b_i$ for all $1 \leq i \leq n$.

Step 3. The monotonicity of b_{in} further implies that $\sum_{i=1}^{n-1} b_{in} \geq \sum_{i=1}^{n-1} b_i$. As argued in the previous step, the series $\sum_{i=1}^{n-1} b_{in}$ is bounded when taking $n \to \infty$, and, hence, $\sum_{i=1}^{n-1} b_i$ is bounded and increasing and thus converges to a limit $b < \infty$. This further implies that for every $\epsilon > 0$ there exists an N such that $\sum_{i=N}^{\infty} b_i < \frac{\epsilon}{2}$. As $\sum_{i=N}^{n-1} b_{in} \leq \sum_{i=N}^{\infty} b_{in}$, the monotone convergence theorem yields

$$\lim_{n\to\infty} \sum_{i=N}^{n-1} b_{in} \leq \lim_{n\to\infty} \sum_{i=N}^{\infty} b_{in} = \sum_{i=N}^{\infty} b_i < \frac{\epsilon}{2},$$

and we conclude that there exists some $\hat{N} \geq N$ such that $\sum_{i=N}^{n-1} b_{in} \leq \epsilon$, for all $n > \hat{N}$. Therefore, for $n > \hat{N}$,

$$\sum_{i=1}^{n-1} b_{in} = \sum_{i=1}^{N-1} b_{in} + \sum_{i=N}^{n-1} b_{in} \leq \sum_{i=1}^{N-1} b_{in} + \epsilon.$$

Furthermore, for any N there exists some $\tilde{N} \geq N$ such that $\sum_{i=1}^{N-1} b_{in} - \sum_{i=1}^{N-1} b_i \leq \epsilon$, for all $n \geq \tilde{N}$, and then for all $n \geq \max\{\tilde{N}, \hat{N}\}$,

$$\sum_{i=1}^{n-1} b_{in} - \sum_{i=1}^{n-1} b_i \leq \sum_{i=1}^{N-1} b_{in} + \epsilon - \sum_{i=1}^{n-1} b_i \leq \sum_{i=1}^{N-1} b_{in} - \sum_{i=1}^{N-1} b_i + \epsilon \leq 2\epsilon,$$

which yields

$$0 \leq \sum_{i=1}^{n-1} b_{in} - \sum_{i=1}^{n-1} b_i \leq 2\epsilon, \quad \forall n \geq \max\{\tilde{N}, \hat{N}\}.$$

The above holds for an arbitrary $\epsilon > 0$ and thus we conclude that $\lim_{n\to\infty} \sum_{i=1}^{n-1} b_{in} = \sum_{i=1}^{\infty} b_i$, and

$$\prod_{i=1}^{n-1} a_{in} = e^{-\sum_{i=1}^{n-1} b_{in}} \xrightarrow{n\to\infty} e^{-\sum_{i=1}^{\infty} b_i} =: a.$$

We conclude that there exists a limit $\ell = \lim_{n\to\infty} L_{n-1}(\mu_n) = \frac{L_0(\mu)}{a}$. □

We now turn our attention to the expected delay,

$$\mathrm{E}S = \sum_{n=1}^{\infty} \frac{q_n}{\mu_n}.$$

Definition 2. A service-rate sequence $\{\mu_n\}_{n=1}^{\infty}$ satisfies finite delay (FD) if it belongs to

$$\mathcal{FD} := \{\{\mu_n\}_{n=1}^{\infty} : \mathrm{E}S < \infty\}.$$

From (1) and (2), we have

$$q_n = p_{n-1} - p_n = (1 - L_{n-1}(\mu_n)) \prod_{i=1}^{n-1} L_{i-1}(\mu_i), \tag{8}$$

that is, a nonhomogeneous geometric distribution. For $n \geq 1$, recall that $\ell_n := L_{n-1}(\mu_n)$, $\ell := \lim_{n\to\infty} \ell_n$ and denote $p := \lim_{n\to\infty} p_n$. If $\ell < 1$ then the geometric term tends to the constant ℓ, that is, the tail of a memoryless distribution.

Lemma 3. *For any external arrival distribution T_0 and service-rate sequence $\{\mu_n\}_{n=1}^{\infty}$,*

a. $p = 0 \Leftrightarrow \sum_{n=1}^{\infty}(1 - \ell_n) = \infty$,
b. $\ell < 1 \Rightarrow p = 0$,
c. $\ell = 1 \Leftrightarrow Y$ *is heavy tailed:* $\sum_{n=1}^{\infty} q_n e^{\eta n} = \infty$, $\forall \eta > 0$,
d. $\frac{\mu_n}{p_n} \overset{n\to\infty}{\longrightarrow} 0 \Rightarrow \ell = 1$.

Proof.

a. By (2), $p_n = \prod_{i=1}^{n} \ell_i$. For any positive sequence $\{a_n\}_{n=1}^{\infty}$, the convergence of the product $\prod_{n=1}^{\infty}(1 - a_n)$ to a nonzero and finite limit is equivalent to the convergence of the sum $\sum_{n=1}^{\infty} a_n$ (see Apostol 1974, p. 209). Hence, $p_n \overset{n\to\infty}{\longrightarrow} 0$ if and only if $\sum_{n=1}^{\infty}(1 - \ell_n) = \infty$.

b. If $\ell < 1$, then clearly $\sum_{n=1}^{\infty}(1 - \ell_n) = \infty$, and, hence, by the previous property, we have that $p = 0$.

c. An equivalent condition for Y being heavy tailed is given by theorem 2.6 of Foss et al. (2011):

$$-\frac{1}{n}\log \Pr(Y > n) \overset{n\to\infty}{\longrightarrow} 0.$$

As $\Pr(Y > n) = p_n$, this is equivalent by the Stolz-Cesáro Theorem (discrete version of L'Hopital's rule) to

$$\log p_n - \log p_{n+1} \overset{n\to\infty}{\longrightarrow} 0,$$

and as $p_n = \prod_{n=1}^{\infty} \ell_n$, we conclude that

$$\log p_n - \log p_{n+1} = -\log \frac{p_{n+1}}{p_n} = -\log \ell_{n+1}$$

converges to zero if and only if $\ell = 1$.

d. Recall the definition of the LST, $\ell_n = L_{n-1}(\mu_n) = \mathrm{E}e^{-\mu_n T_{n-1}}$, then by applying Jensen's inequality and (5) we conclude that

$$\ell_n = \mathrm{E}e^{-\mu_n T_{n-1}} \geq e^{-\mu_n \mathrm{E}T_{n-1}} = e^{-\frac{\mu_n}{\lambda p_{n-1}}} \geq e^{-\frac{\mu_n}{\lambda p_n}}. \quad \square$$

Lemma 3 suggests that the tail behavior of the LST sequence $L_{n-1}(\mu_n)$, and its limit in particular, is a key component in analysing the stability and expected delay in the system. The following proposition summarizes the relationship between feasibility, finite expected delay and the tail behavior of the LST sequence. In particular, we obtain a necessary and sufficient condition for finite expected delay: any feasible service-rate sequence that satisfies $\ell < 1$ with slower decay rate than ℓ. This will be useful for the optimization problem in the following sections.

Proposition 2. *Let λ and $\{\mu_n\}_{n=1}^{\infty}$ be the arrival rate and service-rate sequence, such that $\sum_{n=1}^{\infty}\mu_n = \mu > \lambda$. Then the following properties are satisfied:*

a. $\{\mu_n\}_{n=1}^{\infty} \in \mathcal{M} \Rightarrow p = 0$,
b. $\{\mu_n\}_{n=1}^{\infty} \in \mathcal{FD} \Rightarrow \ell < 1$,
c. *if* $\ell < 1$, *such that* $\ell^n \ll \mu_n$ *then* $\{\mu_n\}_{n=1}^{\infty} \in \mathcal{FD}$,
d. $\{\mu_n\}_{n=1}^{\infty} \in \mathcal{FD} \Rightarrow \{\mu_n\}_{n=1}^{\infty} \in \mathcal{M}$ *(i.e.,* $\mathcal{FD} \subseteq \mathcal{M}$*).*

Proof.

a. This can be directly seen from (6) as the right-hand side tends to zero because of the capacity constraint.

b. The series $\sum_{n=1}^{\infty}\mu_n$ converges; therefore, its tail decays faster than that of the harmonic series. Without loss of generality, as we are only interested in the tail behavior, we assume this is the case for all $n \geq 1$:

$$\mu_n < \frac{1}{n} \Leftrightarrow \frac{1}{\mu_n} > n.$$

The expected delay then satisfies

$$ES = E\frac{1}{\mu_Y} > EY.$$

If $\ell < 1$, then $q_n \approx (1-\ell)\ell^{n-1}$ by (8) and

$$EY = \sum_{n=1}^{\infty} nq_n \approx \frac{1}{1-\ell}.$$

Hence, $\ell = 1$ implies that $ES = \infty$. In other words, $\ell < 1$ is a necessary condition for finite delay.

c. If $\ell < 1$, then, by Lemma 3c, Y is not heavy tailed: there exists $\eta > 0$ such that

$$\sum_{n=1}^{\infty} q_n e^{\eta n} < \infty.$$

If we further assume that

$$ES = \sum_{n=1}^{\infty} q_n \frac{1}{\mu_n} = \infty,$$

then the tail of the service-rate sequence decays even faster than the exponential term, that is,

$$\frac{1}{\mu_n} > e^{\eta n} \Leftrightarrow \mu_n < e^{-\eta n}.$$

Equivalently, we can say that $\mu_n < \alpha^n$ for $\alpha = e^{-\eta}$. If $\mu_n = \beta^n$ such that $\ell < \beta < \alpha$, then

$$ES = \sum_{n=1}^{\infty} q_n \frac{1}{\mu_n} \approx \sum_{n=1}^{\infty} \frac{\ell^n}{\beta^n} < \infty,$$

contradicting the assumption that $ES = \infty$. Hence, if $\ell < 1$ and the service-rate sequence decays slower than ℓ^n, then the expected delay is finite.

d. Any sequence $\{\mu_n\}_{n=1}^{\infty}$ that decays at least as fast as ℓ^n induces infinite expected delay because

$$\sum_{n=1}^{\infty} \frac{\ell^n}{\mu_n} = \infty.$$

If $\{\mu_n\}_{n=1}^{\infty} \in \mathcal{M}$ such that $\mu_n \approx \ell^n$, then

$$\sum_{i=n+1}^{\infty} \mu_i \approx \sum_{i=n+1}^{\infty} \ell^i = \frac{\ell^{n+1}}{1-\ell} \approx \ell^n \approx \lambda p_n.$$

Hence, the tail of the service-rate sequence is on the boundary of the feasible range given by (6). This means that the inequality condition of (6) is satisfied for every n, although the difference converges to zero, and, moreover, any sequence decreasing at a faster rate is not feasible. We conclude that the feasibility of a sequence, $\{\mu_n\}_{n=1}^{\infty} \in \mathcal{M}$, is a necessary condition for finite delay, $\{\mu_n\}_{n=1}^{\infty} \in \mathcal{FD}$. □

Proposition 2 yields a convenient necessary and sufficient condition for a feasible service-rate sequence to satisfy $ES < \infty$, by combining parts b and c of the proposition:

$$\sum_{n=1}^{\infty} \frac{\ell^n}{\mu_n} < \infty. \tag{9}$$

We conclude this section by pointing out open questions and additional refinements of the stability analysis that can be considered in future work on this model.

Remark 3. We conjecture that a stronger result than Proposition 2 holds, namely,

$$\{\mu_n\}_{n=1}^{\infty} \in \mathcal{FD} \Leftrightarrow \ell < 1.$$

Proposition 2c establishes that if μ_n slowly decays enough, then $\ell < 1$ is sufficient for FD. Furthermore, if $\ell < 1$ and $ES = \infty$ such that $\mu_n \le \beta^n$, where $\beta < \ell$, then $\frac{\mu_n}{p_n} \le \frac{\beta^n}{\ell^n} \to 0$, and, by Lemma 3d, we conclude that $\ell_n \to 1$, a contradiction. That is, if the service-rate sequence decays faster than the blocking probability then $\ell = 1$ and the expected delay is infinite. We are left with checking the case of $\mu_n \approx p_n \approx \ell^n$, where $\ell < \beta$. We believe that in this case $\ell_n \to 1$ as well, and this belief is supported by numerical tests, but we were unable to prove this claim. In such a case $\ell_n \to 1$ at a slow rate (in the sense of Lemma 3a). If this is true then indeed $ES < \infty \Leftrightarrow \ell < 1$, but we leave this issue as an open question. A more speculative conjecture is that the extreme case on the boundary of the feasibility region, $\mu_n \approx \ell^n$, occurs when the underlying process is null-recurrent.

Remark 4. An additional open question is whether for any $\lambda < \mu$ there exists a feasible service-rate sequence $\{\mu_n\}_{n=1}^{\infty}$ such that $ES < \infty$, that is, $\mathcal{FD} \neq \emptyset$. We conjecture that this is the case but have no proof. Note that, for any finite N, it is possible to construct a sequence $\{\mu\}_{n=1}^{N}$ such that μ_n decays at a slower rate than p_n (by some positive factor), but the difficulty lies in showing that the rates don't coincide when taking $N \to \infty$.

Remark 5. Little's Law implies that a finite expected delay, $ES < \infty$, is equivalent to a finite expected number of customers in the system. This means that a feasible service-rate sequence and $\ell < 1$ are both necessary, but not sufficient, conditions for the expected number of customers in the system to be finite, which in itself is sufficient but not necessary for general system stability (in terms of positive recurrence of the underlying process). Nevertheless, the probability that a customer that arrives at server n, after being blocked by the previous servers, finds it busy is $\Pr(X_n = 1 | Y \ge n) = L_{n-1}(\mu_n)$. This can be seen by considering the blocking probability of the first server in a system with external arrival distribution L_{n-1}. From Lemma 3b, we have that for any service-rate sequence,

$$\Pr(X_n = 1 | Y \ge n) \overset{n \to \infty}{\longrightarrow} \ell > 0.$$

This gives us an interesting result: "bad" servers, that is, large n and slow service-rate μ_n, block a fixed proportion of arrivals to them.

4. Geometric Service-Rate Series

A very natural capacity allocation to consider is using a simple geometric sequence determined by a single parameter. This is especially called for in light of Proposition 2 that established that if $\ell < 1$ and the service-rate sequence decays slower than a geometric sequence with rate ℓ then the expected delay is finite. Moreover, such service-rate sequences satisfy properties that will be useful for dealing with the capacity allocation problem. Namely, the stability and finite delay conditions have a simple form and the tail of the optimal solution is indeed approximately geometric under some invariance conditions which will be elaborated.

Suppose that the service-rate sequence is determined by a single parameter representing the service capacity allocated to the first server. If we assume without loss of generality that $\mu = 1$ (and then $\rho = \lambda$), then the class of such service-rate sequence is

$$\mathcal{M}_g := \{\{\mu_n\}_{n=1}^{\infty} : \mu_n = \alpha(1-\alpha)^{n-1}, \alpha \in (0,1)\}.$$

In this formulation, the single parameter is the service allocation of the first server, $\mu_1 = \alpha$.

For any $\{\mu_n\} \in \mathcal{M}_g$ we have that $\sum_{i=n+1}^{\infty} \mu_i = (1-\alpha)^n$, and, therefore, the feasibility condition (6) is simply

$$\lambda p_n < (1-\alpha)^n, \quad \forall n \ge 1,$$

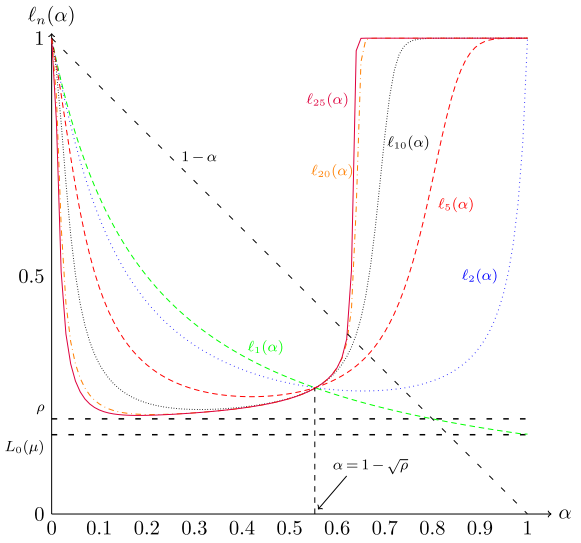
and the finite delay condition (9) is

$$\lambda p_n \ll (1-\alpha)^n \Leftrightarrow \ell < 1 - \alpha. \tag{10}$$

It is possible that the feasibility condition is met but $p_n \to (1-\alpha)$ (from below) and then condition (10) is not met.

Let $\ell_n(\alpha) := L_{n-1}(\alpha(1-\alpha)^{n-1})$ and $\ell(\alpha) = \lim_{n\to\infty} \ell_n(\alpha)$. In Figure 2 the sequence of functions $\ell_n(\alpha)$ are illustrated for the case of Poisson arrivals and $\lambda = \rho = 0.2$. Several interesting observations can be made from this figure, and all of which are robust for different values of ρ and other external interarrival distributions. For every $n \ge 2$, the function $\ell_n(\alpha)$ is unimodal (attaining a minimum) and $\ell_n(0) = \ell_n(1) = 1$. Furthermore, the slope of the functions at $\ell_n(0)$ is decreasing with n, which can be verified by recalling that the derivative of the LST at zero equals the negative of the overflow expectation given by (5): $ET_n = \frac{1}{\lambda p_n}$ (which goes to $-\infty$ as $n \to \infty$ and explains why there seems to be a downward discontinuity at zero as n gets large). This implies that for every $n \ge 1$ there exists an

Figure 2. Sequence of Functions $\ell_n(\alpha) := L_{n-1}(\alpha(1-\alpha)^{n-1})$ When the Service-Rate Sequence Is Geometric with Decay $1-\alpha$: $\mu_n = \alpha(1-\alpha)^{n-1}$



Note. The system parameters are $\mu = 1$ and Poisson arrivals with rate $\lambda = \rho = 0.2$.

$\alpha \in (0,1)$ such that $\ell_n(\alpha) < 1-\alpha$. It appears that this is the case also for the limit $\ell(\alpha)$, which implies the finite delay condition (10), but we currently have no proof to this effect. A proof of this would resolve the open question described in Remark 4 in the previous section. The limiting function $\ell(\alpha)$ appears to have an invariance region, in which the value of the function is almost constant with respect to α, specifically: the function starts at $\ell(0) = 1$, sharply decreases after zero, has an interval $\alpha \in (0, \bar{\alpha})$ which it is almost constant $\ell(\alpha) \approx \bar{\ell}$, and then sharply increases back to $\ell(\alpha) \approx 1$ for $\alpha \in [\bar{\alpha}, 1]$. In the case of a Poisson arrival process we observe that $\bar{\ell} = \rho$ and $\bar{\alpha} = 1 - \sqrt{\rho}$. The latter value is the explicit solution of $\ell_1(\alpha) = \ell_2(\alpha)$, that is, the α value, where the first and second functions intersect. Interestingly, it appears that all the functions intersect at around the same point. It is difficult to tell whether the limiting function $\ell(\alpha)$ would have an upward discontinuity to 1 at $\bar{\alpha}$ or just a sharp and continuous increase as we see for $n = 25$. We were unable to computationally explore the function for higher values.

The invariance of the limit function $\ell(\alpha)$ also has implications on the delay-minimization problem: if $\ell(\alpha) = \ell$ for all $\alpha \in (0, \bar{\alpha})$ then the tail of the delay minimization objective function has a very simple form, $\frac{\ell^n}{(1-\alpha)^n}$, and the optimal α can be computed as described below.

Suppose now that the blocking probability for all $n \geq 1$ is $p_n = \ell^n$, and consequently $q_n = (1-\ell)\ell^{n-1}$, where $\ell < 1$. We already established in Lemma 2b and Proposition 2 that this is a reasonable approximation for the tail behavior of the expected delay for any feasible service with finite delay. In the sequel (specifically in Proposition 4) we will also show that if $\{\mu_n\}_{n=1}^{\infty} \in \mathcal{FD}$ then ℓ has a certain degree of invariance to the tail of $\{\mu_n\}_{n=1}^{\infty}$, thus providing additional justification for the use of approximation of the optimal solution with a fixed ℓ. The optimal service-rate sequence for such a system is the solution to an infinite dimensional convex program on a simplex:

$$\min_{\{\mu_n\}_{n=1}^{\infty} \in \mathcal{M}} \frac{1-\ell}{\ell} \sum_{n=1}^{\infty} \frac{\ell^n}{\mu_n} \quad \text{s.t.} \left\{ \mu_n > 0,\ \forall n \geq 1,\ \sum_{n=1}^{\infty} \mu_n = 1 \right\}. \tag{11}$$

We refer to (11) as the tail approximation program (TAP). The following proposition asserts that the solution to the TAP is in $\mathcal{M}_g \cap \mathcal{M}$ with $\alpha = (1-\sqrt{\ell})$. This solution resembles the square-root optimal capacity allocation in a Jackson network (see Kleinrock 1976, p. 329),[2] but the models are not directly linked. The program is a convex infinite horizon program, in the sense of Grinold (1983) (for general optimality conditions, see Barbu and Precupanu 2012, p. 153), which allows us to find the optimal solution as a limit of finite dimensional programs.

Proposition 3. *The solution to* (11) *is* $\mu_n = (1-\ell^{\frac{1}{2}})\ell^{\frac{n-1}{2}}$, $\forall n \geq 1$.

Proof. First, we argue that the optimal service-rate sequence is nonincreasing by applying a simple interchange argument. Suppose that $\{\mu_n\}_{n=1}^{\infty}$ is an optimal service-rate sequence such that $\mu_i < \mu_j$ for some $i < j$. The contribution of elements i and j to the objective function is

$$\frac{\ell^i}{\mu_i} + \frac{\ell^j}{\mu_j}.$$

If $i < j$ then $\ell^i > \ell^j$, which means that a greater weight is given to $\frac{1}{\mu_i}$, which is bigger than $\frac{1}{\mu_j}$. Hence, we can improve the objective without deviating from the capacity constraint by switching the values of μ_i and μ_j. This contradicts the assumption that the sequence is optimal.

The objective function is an infinite sum of convex single-variable functions. We first consider the finite program for an integer M,

$$\min_{\{\mu_n\}_{n=1}^{M} \in \mathcal{M}} \frac{1-\ell}{\ell} \sum_{n=1}^{M} \frac{\ell^n}{\mu_n}$$
$$\text{s.t.} \left\{ \mu_n \geq 0,\ \forall n \geq 1,\ \sum_{n=1}^{M} \mu_n = 1 \right\}.$$

Every element of the objective function is unbounded as $\mu_n \to 0$, and, therefore, the solution is in the interior of the constraint set. This means that every element satisfies the first-order condition:

$$\frac{\ell^n}{\mu_n^2} = \kappa, \quad 1 \leq n \leq M,$$

where κ is the Lagrange multiplier for the equality constraint:

$$-\kappa \left(\sum_{n=1}^{M} \mu_n - 1 \right) = 0.$$

Simple algebra yields

$$\mu_n = \frac{1}{\sqrt{\kappa}} \ell^{\frac{n}{2}},$$

and, by applying the capacity constraint,

$$\sum_{n=1}^{M} \frac{1}{\sqrt{\kappa}} \ell^{\frac{n}{2}} = 1,$$

we derive that

$$\sqrt{\kappa} = \frac{\ell^{\frac{1}{2}} \left(1 - \ell^{\frac{M}{2}}\right)}{1 - \ell^{\frac{1}{2}}}.$$

Finally, by taking $M \to \infty$, we conclude that the optimal solution to (11) is $\mu_n = (1 - \ell^{\frac{1}{2}}) \ell^{\frac{n-1}{2}}$. □

5. Optimization and Approximation

We are interested in solving the mathematical program

$$\min_{\{\mu_n\}_{n=1}^{\infty} \in \mathcal{M}} \sum_{n=1}^{\infty} \frac{q_n}{\mu_n}. \tag{12}$$

This program can be formulated as an infinite horizon Markov decision process with a state- and an action-dependent discount factor (see Wei and Guo 2011). The idea is that at every step n we consider a new system with interarrival distribution given by the overflows from server $n-1$ and the remaining capacity constraint. The discount factor at step n will be given by $L_{n-1}(\mu_n)$, the blocking probability when a customer overflows to server n.

First, we define the mapping:

$$\hat{L}(x,L)(s) = \frac{L(s+x)}{1-L(s)+L(s+x)} : \mathbb{R} \times \mathcal{L} \to \mathcal{L},$$

where $\mathcal{L}$ is the space of nonincreasing functions from $[0,\infty)$ to $[0,1]$. For any $n \geq 1$, given the overflow distribution L_{n-1} we take advantage of the recursive form of q_n in (8) to obtain

$$q_{n+1} = \left(1 - \hat{L}(\mu_n, L_{n-1})(\mu_{n+1})\right) p_n,$$

where by (2),

$$p_n = L_{n-1}(\mu_n) p_{n-1}.$$

The objective function of (12) can be written as

$$(1-L_0(\mu_1))\frac{1}{\mu_1} + L_0(\mu_1)\left(1-\hat{L}(\mu_1,L_0)(\mu_2)\right)\frac{1}{\mu_2}$$
$$+ L_0(\mu_1)\hat{L}(\mu_1,L_0)(\mu_2)\left(1-\hat{L}(\mu_2,\hat{L}(\mu_1,L_0)(\mu_2))(\mu_3)\right)\frac{1}{\mu_3} + \dots \quad .$$

Therefore, an equivalent program to (12) is given by the Bellman equation

$$v(\mu, L) = \min_{\{x \in [0,\mu]\}} \left\{ (1-L(x))\frac{1}{x} + L(x) v(\mu - x, \hat{L}(x,L)) \right\}, \tag{13}$$

with the objective $v(\mu, L_0)$. In every step μ is the total available capacity and L is the LST defining the external arrival process to the system. Although (13) has an elegant form, it is not straightforward to solve numerically. This is due to the infinite dimensional state space $\mathcal{L}$, which is a space of continuous functions. Next, we suggest an equivalent program with a simpler state space that includes the server index and the capacities that have been allocated.

For any given exogenous arrival distribution L_0 we can compute the values of $L_n(s)$ given the sequence $\{\mu_1, \dots, \mu_{n-1}\}$ using the recursive formula (3). The program (13) with capacity constraint μ can be defined by the Bellman equation

$$v_n(\mu_1, \dots, \mu_{n-1}) = \min_{\mu_n \leq \mu - s_{n-1}} \left\{ \frac{q_n}{\mu_n} + v_{n+1}(\mu_1, \dots, \mu_n) \right\}, \quad n \geq 1, \tag{14}$$

where $s_n := \sum_{i=1}^{n} \mu_i$. The overall objective is $v_1(\emptyset)$.

Unfortunately, there is an additional problem of computational complexity. Specifically, computing q_n requires computing the recursion for $L_{n-1}(\mu_n)$ which is of the magnitude of 2^n steps. In the sequel we propose a numerical approximation method that relies on the solution of (14) for a small number of steps with the TAP solution (11) as an initial condition.

Observe that there is no direct restriction for the solution of (12) to be nonincreasing, which is necessary if customers always go to the fastest server available. Next, we argue that the optimal sequence is indeed nonincreasing, even without the explicit constraint. In Proposition 3 the explicit geometric decay rate of the optimal service sequence was shown to be $\sqrt{\ell}$ for the approximation model, whereas in the general case we only know that it decreases but not at what rate.

Lemma 4. *The solution* $\{\mu_n\}_{n=1}^{\infty}$ *of* (12) *is a nonincreasing sequence.*

Proof. Suppose that $\{\mu_n\}_{n=1}^{\infty}$ is an optimal solution such that $\mu_n < \mu_{n+1}$ for some $n \geq 1$. The average expected delay is

$$\mathrm{E}S = \mathrm{E}(S|Y<n)\Pr(Y<n) + \mathrm{E}(S|Y \geq n)\Pr(Y \geq n).$$

If the rates of server n and $n+1$ are interchanged then first summand is unchanged, while the second is decreased because all blocking probabilities p_k for $k \geq n$ decrease (see Nath and Enns 1981), thus, contradicting the optimality of the sequence. □

A nice property of decreasing service-rate sequence is given to us by Lemma 1c, which states that the sequence of blocking properties p_n is discrete convex:

$$p_{n+1} + p_{n-1} > 2p_n, \quad n \geq 2.$$

Recall that $p_n = \prod_{i=1}^{n} L_{i-1}(\mu_i)$, so in terms of the LST sequence this is equivalent to

$$L_{n-1}(\mu_n)(1 - L_n(\mu_{n+1})) < 1 - L_{n-1}(\mu_n), \quad n \geq 2,$$

and thus

$$q_{n+1} = (1 - L_n(\mu_{n+1}))L_{n-1}(\mu_n) \prod_{i=1}^{n-1} L_{i-1}(\mu_i) < (1 - L_{n-1}(\mu_n)) \prod_{i=1}^{n-1} L_{i-1}(\mu_i) = q_n.$$

This means that the sequence q_n is decreasing, and, hence, the weights of the increasing sequence of expected service times, $\frac{1}{\mu_n}$, in the objective function of (12) are decreasing.

5.1. Approximate Solution

If $\ell < 1$ then (2) gives us a geometric approximation of the tail behavior of the blocking probabilities $p_n \approx \ell^n$. Thus, for large M, we set

$$q_n = p_M \ell^{n-(M+1)}(1 - \ell), \ \forall n > M.$$

Because of Proposition 3, we have that if the sequence ℓ_n does not vary by much then a good approximation for the optimal solution is given by a geometric service-rate sequence with decay rate $\sqrt{\ell}$. Specifically, the approximately optimal tail series is

$$\mu_n = \mu_{n-1}\sqrt{\ell} = \mu_M \ell^{\frac{n-M}{2}}, \quad \forall n > M, \tag{15}$$

and the approximate optimal residual is

$$\sum_{n=M+1}^{\infty} \frac{q_n}{\mu_n} \approx \frac{p_M(1-\ell)}{\mu_M \ell} \sum_{n=1}^{\infty} \left(\frac{\ell}{\sqrt{\ell}}\right)^n = \frac{p_M(1-\ell)}{\mu_M \sqrt{\ell}(1-\sqrt{\ell})} = \frac{p_M(1+\sqrt{\ell})}{\mu_M \sqrt{\ell}}.$$

For small values of M ($\leq$25) we can accurately compute $q_1, \ldots, q_M$ and approximate the optimal residual by using the TAP solution (11). This yields the approximation

$$\mathrm{ES} \approx \sum_{n=1}^{M} \frac{q_n}{\mu_n} + r_M,$$

where

$$r_M := \frac{p_M(1+\sqrt{\ell_M})}{\mu_M \sqrt{\ell_M}},$$

and $\ell_M := L_{M-1}(\mu_M)$. The term r_M represents the residual of the expected delay given by the tail approximation.

According to Proposition 2b, $\ell < 1$ for any sequence with finite delay. A finite-horizon dynamic program that approximates (14) can now be formulated: for $1 \leq n \leq M$,

$$v_n^{(M)}(\mu_1, \ldots, \mu_{n-1}) = \min_{\mu_n \leq \mu - s_{n-1}} \left\{ \frac{q_n}{\mu_n} + v_{n+1}^{(M)}(\mu_1, \ldots, \mu_n) \right\}, \tag{16}$$

with initial condition $v_{M+1}^{(M)}(\mu_1, \ldots, \mu_M) = r_M$ and the objective $v_1^{(M)}(\mu)$. We can increase M until r_M is lower than some tolerance parameter or, alternatively, until the change in the $L_{n-1}(\mu_n)$ sequence is smaller than some parameter.

The tail sequence $\{\mu_n\}_{n=M+1}^{\infty}$ needs to satisfy the capacity constraint,

$$\sum_{n=M}^{\infty} \mu_m \leq \mu - s_{M-1},$$

which according to (15) yields

$$\mu_M \le (1-\sqrt{\ell_M})(\mu - s_{M-1}). \tag{17}$$

In an optimal allocation (17) will have an equality, as there is no gain from not allocating all of the capacity. Lemma 1a implies that for every $(\mu_1,\dots,\mu_{M-1})$ there is a unique μ_M for which an equality holds. Therefore, the dynamic program effectively has $M-1$ steps.

The implementation of the approximating dynamic program is obtained using a standard fixed-point search algorithm. Let $\mu_0 := 0$ and

$$\mu_n^*(\mu_1,\dots,\mu_{n-1}) = \begin{cases} \arg\min_{\mu_1 \le \mu}\left\{\frac{q_1}{\mu_1} + v_2^{(M)}(\mu_1)\right\}, & n=1 \\ \arg\min_{\mu_n \le \mu - s_{n-1}}\left\{\frac{q_n}{\mu_n} + v_{n+1}^{(M)}(\mu_1,\dots,\mu_n)\right\}, & 2 \le n < M, \\ (1-\sqrt{\ell_M})(\mu - s_{M-1}), & n = M, \end{cases}$$

for $n = 1,\dots,M$. Then $(\mu_1,\dots,\mu_M)$ is an optimal solution if $\mu_n^*(\mu_1,\dots,\mu_{n-1}) = \mu_n$, for all $1 \le n \le M$. Note that while the constraint at step n only depends on the sum $s_{n-1} = \sum_{i=1}^{n}\mu_i$, the value function $v_n^{(M)}$ depends on the entire vector $(\mu_1,\dots,\mu_{n-1})$ through the overflow distribution L_n. A simplified description of such an algorithm is as follows:

1. For $n = M, M-1,\dots,1$ compute $\mu_n^*(\mu_1,\dots,\mu_{n-1})$ for any allocation $(\mu_1\dots,\mu_{n-1})$ and the corresponding value $v_n^{(M)}(\mu_1,\dots,\mu_n^*(\mu_1,\dots,\mu_{n-1}))$.
2. The vector

$$(\mu_1^*, \mu_2^*(\mu_1^*),\dots,\mu_M^*(\mu_1^*,\dots,\mu_{M-1}^*(\mu_1^*,\mu_2^*(\mu_1^*),\dots))),$$

satisfies the fixed-point condition and is an optimal solution.

The most naive and exhaustive way to solve the approximate program is to compute the values for all possible M-dimensional allocations on a discrete grid with increments of size $\delta > 0$. Of course, this is computationally expensive, in the magnitude of $(\frac{\mu}{\delta})^M$. The search at any stage can be carried out in a much more efficient manner, such as bisection, and then the functions do not have to be evaluated at every point in the continuous search interval for every $\mu_n \in (0, \mu - s_{n-1})$. In practice, the search finds the optimal value in a very small number of computations using various generic optimization packages and the computational bottleneck is the evaluation of $L_n()$ for increasing n.

In the TAP the blocking probability is assumed to decay at a constant rate, specifically the limit ℓ. For this to be a good approximation the tail of the sequence $\ell_n := L_{n-1}(\mu_n)$ needs to be somehow insensitive to changes in the tail of the service-rate sequence. This behavior was discussed in Section 4 and illustrated in Figure 2. To strengthen the justification of this approximation, we further show that the tail of any service-rate sequence with finite delay is decreasing and is bounded from below by an increasing sequence.

Proposition 4. *If $\{\mu_n\}_{n=1}^{\infty} \in \mathcal{FD}$ then the sequence $\{\ell_n\}$ has a decreasing tail, and is bounded from below by an increasing sequence $\{\underline{\ell}_n\}$.*

Proof. According to Lemma 2, the LST sequence has a lower bound of $L_0(\mu)$, that is, the external input LST with all the capacity. This argument can be repeated for every overflow distribution L_{n-1} into server n, when the remaining capacity is $\mu - s_{n-1}$. So at any step n of the dynamic program (16), we have

$$\ell_n = L_{n-1}(\mu_n) \ge L_{n-1}(\mu - s_{n-1}) =: \underline{\ell}_n.$$

By (3),

$$L_n(\mu - s_n) = \frac{L_{n-1}(\mu_n + \mu - s_n)}{1 - L_{n-1}(\mu - s_n) + L_{n-1}(\mu_n + \mu - s_n)},$$

and, as $\mu_n + \mu - s_n = \mu - s_{n-1}$, and the denominator is smaller than one,

$$\underline{\ell}_{n+1} = L_n(\mu - s_n) > L_{n-1}(\mu - s_{n-1}) = \underline{\ell}_n.$$

Hence, the sequence of lower bounds, $\underline{\ell}_n$, is increasing with n. Furthermore, if the sequence $\{\mu_n\}_{n=1}^{\infty}$ is nonincreasing and has finite delay then the sequence $\frac{q_n}{\mu_n}$ converges to zero and is therefore decreasing at the tail. Using (8), this implies that for large n:

$$\frac{\mu_{n+1}(1-\ell_n)p_{n-1}}{\mu_n(1-\ell_{n+1})p_n}<1,$$

which yields

$$\frac{1-\ell_n}{1-\ell_{n+1}}<\frac{\mu_n}{\mu_{n+1}}\ell_n<1.$$

The last inequality comes from the finite delay condition in Proposition 2c that demands that the decay of the service-rate sequence be slower than that of the blocking probabilities. Therefore, we conclude that $\ell_n>\ell_{n+1}$ at the tail. □

To summarize, for any reasonable service-rate sequence, that is nonincreasing and with finite delay, the sequence ℓ_n is decreasing and is also bounded from below by an increasing sequence. This shows that changing the service-rate sequence at the tail has a small, or bounded, effect on the limit ℓ, as long as the finite delay condition is met. In Figure 3 the lower-bound sequence is illustrated alongside the LST sequence for the approximate optimal solution for an example set of parameters. Indeed, ℓ_n approaches the lower-bound sequence $\underline{\ell}_n$ very quickly. In this example, we have that $\frac{\ell_{15}-\underline{\ell}_{15}}{\ell_{15}}\approx 0.025$, which indicates that the error term is very accurate even for $M=15$ (and at $M=19$ the normalized error is already smaller than 0.001). For other examples similar was observed, and, as expected, for higher levels of ρ, a bigger M is required to achieve accuracy.

6. Numerical Analysis

In this section, we assume $\mu=1$ and that the external interarrival distribution is gamma$(k,k\lambda)$. In this case the overall utilization is $\rho=\frac{k\lambda}{k}=\lambda$ and the variance of the exogenous interarrival times is $\frac{1}{k\lambda^2}=\frac{1}{k\rho^2}$. Therefore, we can examine different levels of utilization and variance by changing ρ and k. For high levels of ρ, this analysis is quite general as the blocking probabilities in the heavy-traffic approximation of the GI/M/c system with many servers are known to depend on the first two moments of the arrival distribution (see Whitt 1984).

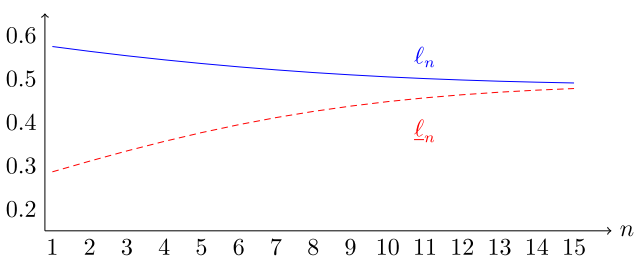
Figure 4 illustrates the approximate optimal service-rate sequence by solving (16) for different parameter values with $M=15$. The first thing to observe is that in all examples the approximate optimal service-rate sequence is very close to geometric.

In Table 1 we present the approximate optimal values of ES and the respective tail approximations r_M. For high levels of ρ, the contribution of the tail approximation is substantial and hence potentially less accurate. However, we find that the sequence of $L_{n-1}(\mu_n)$ stabilizes very quickly, and, therefore, the approximation $\ell\approx L_{M-1}(\mu_M)$ is quite accurate, suggesting the error terms provide a decent approximation. The sequence of LST of the approximate optimal service-rate sequence are illustrated in Figure 5. In all examples the sequence indeed stabilizes very fast, and, hence, the tail approximation using the value of ℓ_M is appropriate. This stability result was further verified by running computations for higher values, exact up to $M=25$ and based on a discrete event simulation of the system for $M>25$, for a large number of servers using the approximate optimal service-rate sequence. In particular, the LST sequences, for example, those displayed in Figure 5, remain almost constant when taking an n much larger than 10.

In the special case of Poisson arrivals ($k=1$), it is interesting to observe that $\mu_1\approx 1-\sqrt{\rho}$, as seen in Figure 6. Thus, a reasonable rough approximation for the optimal service-rate sequence given by

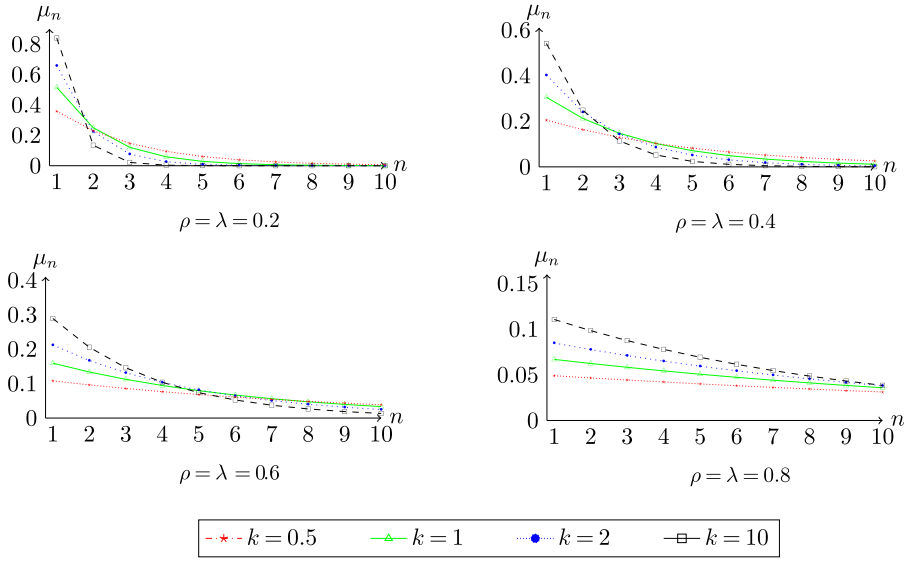
$$\mu_n=\begin{cases}1-\sqrt{\rho}, & n=1\\ \sqrt{\rho}\mu_{n-1}, & n\geq 2.\end{cases}$$

Figure 3. LST Sequence (of the Approximate Optimal Service-Rate Sequence) and the Lower-Bound Sequence



Note. Example parameters: Total capacity of $\mu=1$ and Poisson arrivals with rate $\lambda=0.4$.

Figure 4. Approximate Optimal Service-Rate Sequence for Varying Values of ρ and k



The value $\sqrt{\rho}$ appeared in two places before: (1) The solution to the equation $L_0(\alpha) = L_1(\alpha(1-\alpha))$ is $\alpha = 1 - \sqrt{\rho}$, as was elaborated in Section 4 (see also Figure 2). This seems to be a critical point for the limit function $\ell(\alpha)$ for geometric service-rate sequence (with rate α). (2) The optimal tail decay rate given in Proposition 3 is $\sqrt{\ell}$, where in the Poisson case we observe that $\ell \approx C\rho$ where C is a constant that was in the range of $(1, 1.15)$ in all examples computed.

Let $\rho_0 = \rho = \frac{\lambda}{\mu}$ and let $\rho_n := \frac{\lambda p_n}{\mu - \sum_{i=1}^{n} \mu_i}$ denote the effective utilization of the subsystem excluding the first n servers for $n \geq 1$. In Figure 7 we see that the effective utilization sequence, given the approximate optimal service-rate sequence, is decreasing with n for all parameter values. An interesting numerical result is that in all examples the tail of the utilization level sequence decays geometrically. The rate of decay is slightly higher than $1 - \mu_1$, that is to say the effective utilization decreases at a slower rate than the service-rate sequence, as expected.

6.1 Summary of Numerical Results

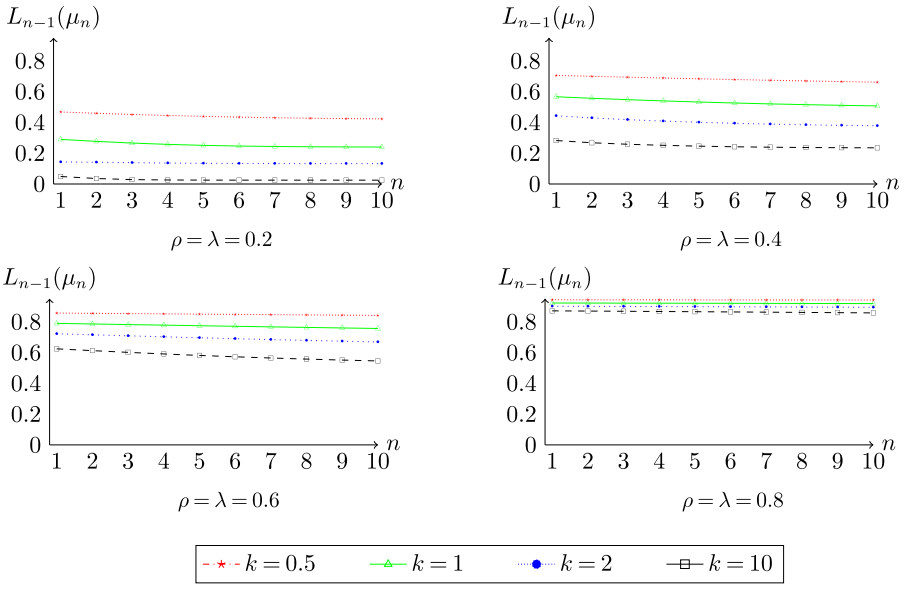
The approximate optimal service-rate sequence is close to geometric with decay rate $1 - \mu_1$. In the special case of a Poisson arrival process we observed that $\mu_1 \approx 1 - \sqrt{\rho}$ and ℓ was slightly larger than ρ. When considering a fixed ρ, lower variance of the exogenous arrival stream leads to a higher service-rate for the first server, along with a faster decline to zero. However, as ρ increases the service capacity is more "uniformly" allocated (with a lower decay rate). The effective utilization sequence of subsystems, ρ_n, under the approximate optimal service-rate sequence is decreasing. Furthermore, this sequence has a geometric tail with a slower decay rate than the optimal service-rate sequence.

Recall that for reasonable approximations we would like the sequence of overflow LST $L_{n-1}(\mu_n)$ to approach a constant at a quick rate. Indeed, we observe that it stabilizes very fast, suggesting that our approximation scheme using its limit is accurate even for a small number of DP steps M. This may provide some explanation as to why the optimal sequence seems geometric from the start. If the sequence $L_{n-1}(\mu_n)$ is more or less constant from the start then the solution to the TAP from Proposition 3 is close to optimal. It would be interesting to find an analytical explanation for why the optimal service-rate sequence comes with a stable overflow LST sequence.

Table 1. Approximate Optimal Expected Delay and the Optimal Tail Approximation: (ES, r_M)

	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$
$k = 0.5$	(5.18, 0.04)	(10.81, 0.387)	(25.72, 8.23)	(118.1, 98.38)
$k = 1$	(3.22, 4.7^{-5})	(6.86, 0.014)	(16.23, 1.57)	(69.78, 48.48)
$k = 2$	(2.23, 1.6^{-7})	(4.87, 0.001)	(11.78, 0.36)	(48.21, 27.54)
$k = 5$	(1.63, 6.9^{-10})	(3.66, 3.3^{-4})	(9.14, 0.08)	(36.19, 16.29)
$k = 10$	(1.44, 8.2^{-12})	(3.25, 8.2^{-5})	(8.25, 0.04)	(32.4, 12.9)

Figure 5. Laplace Transform Sequence ($L_{n-1}(\mu_n)$) Corresponding to the Approximate Optimal Service-Rate Sequence for Varying Values of ρ and k



7. Applications

The analysis presented here can be modified to solve several other system design problems, for example:

a. Multiobjective optimization: suppose that the system administrator can choose the number of servers as well as the capacity allocation. If n servers are used, then the system has a customer loss probability of p_n. The administrator may seek an optimal allocation with a constraint on the loss probability, for example, $p_n \le p < 1$. The other way around is also an option: minimize p_n subject to a constraint on the delay, $ES \le w$.

b. Suppose that the system administrator wants to maximize the number of users that wait less than some $\tau > 0$ time threshold. This could be an exogenous performance measure or the case if customers do not pay if their delay is too long. The objective is now

$$\min_{\{\mu_n\}_{n=1}^{\infty}} \sum_{n=1}^{\infty} q_n e^{-\mu_n \tau}.$$

c. Customers are heterogeneous with respect to the utility from the speed of service, and balk from the system if the fastest available server is slower than their value. Assume that customer values are distributed according to a continuous distribution with a convex cdf Λ such that $\Lambda(0) > 0$. If the system wants to minimize the blocking probability, then the objective becomes

$$\min_{\{\mu_n\}_{n=1}^{\infty}} \sum_{n=1}^{\infty} q_n (1 - \Lambda(\mu_n)).$$

Figure 6. Approximation of Optimal μ_1 as a Function of ρ for the Poisson Arrival Case ($k = 1$)

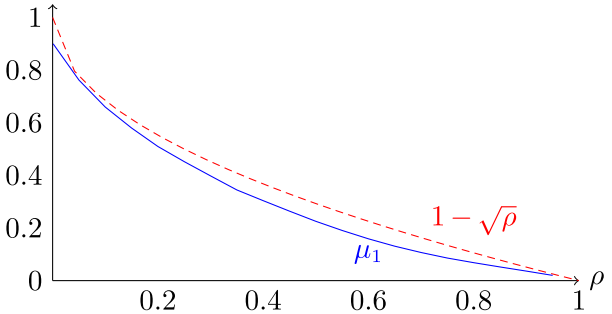
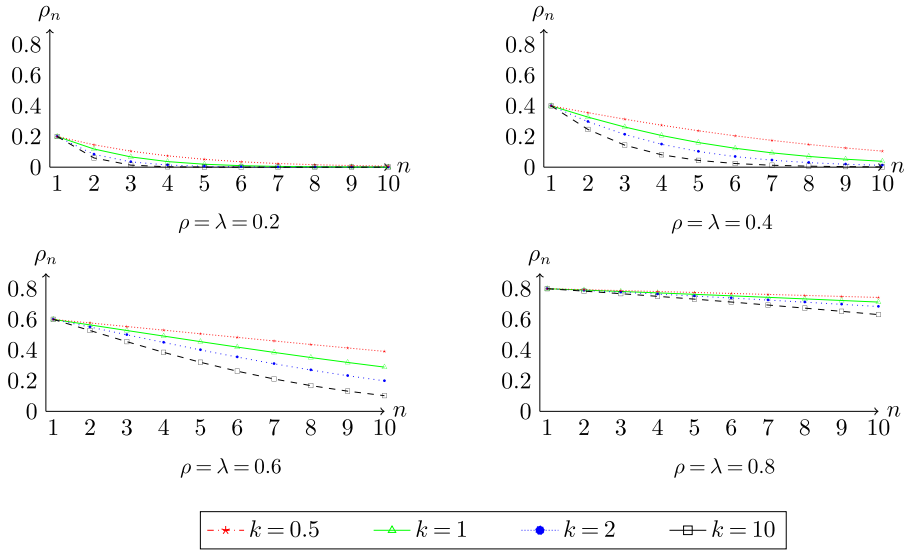


Figure 7. Effective Utilization Rate Sequence ρ_n Corresponding to the Approximate Optimal Service-Rate Sequence for Varying Values of ρ and k



In this case, the overflow distribution also requires a modification to $\tilde{L}_{n-1}(\mu_n) = L_{n-1}(\mu_n)\Lambda(\mu_n)$, to account for balking customers.

In general, our analysis can be applied to any convex function of the service rate.

8. Discussion

This paper analyzes stability and expected delay in an infinite-server system with finite service capacity. In particular, the expected service-time minimizing allocation of service-rates is examined. It is shown that the optimal service-rate sequence is geometrically decreasing at the tail. Numerical analysis suggests that the optimal allocation is very close to geometric throughout the sequence, and not just at the tail. This property is related to the product form of the blocking probabilities from finite subsystems. An interesting numerical observation is that in the Poisson case we have that $\ell \approx \rho$ (the limiting term in the blocking probability product) and consequentially the optimal tail decay is simply $\sqrt{\rho}$. An open challenge is to find analytical justification for this phenomenon.

The most important conclusion of the paper is that allocating capacity to heterogeneous servers under capacity constraints should be done with caution. Even if there is seemingly enough capacity for the incoming arrival rate, allocating too much capacity to the fast servers may lead to very long expected delay. In this paper we analyzed an infinite server system, but this conclusion is also relevant for finite server loss systems with very low blocking probability, in which case expected delay would be finite but potentially very big. Although this is most relevant for very large systems, the geometric tail behavior implies that with a "bad" allocation the expected delay can increase very fast with the number of servers, and, therefore, the conclusion is still relevant for moderately sized systems, that is, not necessarily hundreds of servers.

Stability analysis in the probabilistic sense of the underlying Markov chain $\mathbf{X}(t) \in \mathcal{S}$ is also of interest. Specifically, establishing necessary and sufficient conditions for the process to be positive recurrent. This can perhaps be achieved by considering the embedded Markov chain at arrival times. For the Markovian M/M/∞ ordered system, lower-bound and upper-bound systems with simpler dynamics can possibly be constructed and analyzed using matrix-analytic methods. Rigorous characterization of the stability conditions also could potentially shed some light on the open problems discussed at the end of Section 3, in particular, establishing the exact necessary and sufficient conditions for expected finite delay. The distinction between positive and null recurrence is potentially the additional refinement required for dealing with the case of a service sequence on the boundary of the feasibility region.

It would be interesting to study the asymptotic optimal control problem of this system using heavy-traffic analysis, that is, scaling the parameters by an appropriate rate function $r(n)$,

$$r(n)\lambda^{(n)} \overset{n\to\infty}{\longrightarrow} \lambda = \mu \overset{n\to\infty}{\longrightarrow} r(n)\mu^{(n)}, \quad \lambda^{(n)} < \mu^{(n)} \ \forall n \geq 1.$$

A question that arises is whether there is an asymptotic analog to the optimal geometric sequence we have presented here, perhaps even in closed form. For $\rho < 1$ there is always a feasible service-rate sequence, however, our approximations are less accurate as $\rho \uparrow 1$, so heavy-traffic approximations may yield better insight to such systems. Detailed heavy-traffic analysis for the homogeneous ranked M/M/∞ system can be found in Newell (1984), and approximation analysis of blocking probabilities in multiserver systems can be found in Whitt (1984). Approximations of the number of idle servers in many-server systems, such as Eschenfeldt and Gamarnik (2018), also may be useful. Approximating our model requires the analysis of systems with nonhomogeneous servers (see Atar et al. 2013).

Acknowledgments

The authors wish to thank Abhishek, Brian Fralix, and Johan van Leeuwaarden for helpful discussions and comments. They are also grateful to the associate editor and two referees for their detailed comments and suggestions, which greatly improved the paper.

Endnotes

[1] This state space description was suggested by Brian Fralix.

[2] This observation was made by Johan van Leeuwaarden.

References

Apostol TM (1974) *Mathematical Analysis* (Pearson, New York).

Armony M (2005) Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems* 51(3/4):287–329.

Atar R, Goswami A, Shwartz A (2013) Risk-sensitive control for the parallel server model. *SIAM J. Control Optim.* 51(6):4363–4386.

Balsamo S, Donatiello L, Dijk NMV (1998) Bound performance models of heterogeneous parallel processing systems. *IEEE Trans. Parallel Distributed Systems* 9(10):1041–1056.

Barbu V, Precupanu T (2012) *Convexity and Optimization in Banach Spaces* (Kluwer, Hingham, MA).

Coffman EG Jr, Kadota T, Shepp LA (1985) A stochastic model of fragmentation in dynamic storage allocation. *SIAM J. Comput.* 14(2):416–425.

Cooper RB (1976) Queues with ordered servers that work at different rates. *Opsearch* 13(2):69–78.

Cooper RB, Palakurthi S (1989) Heterogeneous-server loss systems with ordered entry: An anomaly. *Oper. Res. Lett.* 8(6):347–349.

Eschenfeldt P, Gamarnik D (2018) Join the shortest queue with many servers. The heavy-traffic asymptotics. *Math. Oper. Res.* 43(3):867–886.

Foss S, Korshunov D, Zachary S (2011) *An Introduction to Heavy-Tailed and Subexponential Distributions* (Springer, Hoboken, NJ).

Grinold R (1983) Convex infinite horizon programs. *Math. Programming* 25(1):64–82.

Hassin R, Shaki YY, Yovel U (2015) Optimal service-capacity allocation in a loss system. *Naval Res. Logist.* 62(2):81–97.

Khazaei H, Misic J, Misic VB (2012) Performance analysis of cloud computing centers using M/G/m/m+r queuing systems. *IEEE Trans. Parallel Distributed Systems* 23(5):936–943.

Khazaei H, Misic J, Misic VB (2013) A fine-grained performance model of cloud computing centers. *IEEE Trans. Parallel Distributed Systems* 24(11): 2138–2147.

Kleinrock L (1976) *Computer Applications,* Queueing Systems, vol. 2 (Wiley, Hoboken, NJ).

Knessl C (2004) Some asymptotic results for the M/M/∞queue with ranked servers. *Queueing Systems* 47(3):201–250.

Nath GB, Enns EG (1981) Optimal service rates in the multiserver loss system with heterogeneous servers. *J. Appl. Probab.* 18(3):776–781.

Newell GF (1984) *The M/M∞ Service System with Ranked Servers in Heavy Traffic* (Springer, Hoboken, NJ).

Riordan J (1962) *Stochastic Service Systems* (Wiley, New York).

Rosberg Z, Makowski AM (1990) Optimal routing to parallel heterogeneous servers-small arrival rates. *IEEE Trans. Automatic Control* 35(7): 789–796.

Ross SM (2014) Optimal server selection in a queueing loss model with heterogeneous exponential servers, discriminating arrivals, and arbitrary arrival times. *J. Appl. Probab.* 51(3):880–884.

Shanthikumar JG, Yao DD (1987) Optimal server allocation in a system of multi-server stations. *Management Sci.* 33(9):1173–1180.

Tahara A, Nishida T (1975) Optimal allocation of service rates for multi-server markovian queue. *J. Oper. Res. Soc. Japan* 18(1/2):90–95.

Takacs L (1959) On the limiting distribution of the number of coincidences concerning telephone traffic. *Ann. Math. Statist.* 30(1):134–142.

Vilaplana J, Solsona F, Teixidó I, Mateo J, Abella F, Rius J (2014) A queuing theory model for cloud computing. *J. Supercomput.* 69(1):492–507.

Wei Q, Guo X (2011) Markov decision processes with state-dependent discount factors and unbounded rewards/costs. *Oper. Res. Lett.* 39(5): 369–374.

Whitt W (1984) Heavy-traffic approximations for service systems with blocking. *AT&T Bell Labs Tech. J.* 63(5):689–708.

Yao DD (1986) Convexity properties of the overflow in an ordered-entry system with heterogeneous servers. *Oper. Res. Lett.* 5(3):145–147.

Yao DD (1987) The arrangement of servers in an ordered-entry system. *Oper. Res.* 35(5):759–763.

Zreikat A, Bolch G, Sztrik J (2003) Performance modelling of nonhomogeneous unreliable multiserver systems using mosel. *Comput. Math. Appl.* 46(2):293–312.