Optimal service-capacity allocation in a loss system

Refael Hassin       Yair Y. Shaki       Uri Yovel *

October 27, 2013

Abstract

We consider a loss system with a fixed budget for servers. The system owner’s problem is to choose the price, and the number and quality of the servers, which maximize its profits, subject to a budget constraint. We solve the problem with identical and different service rates, preemptive and non-preemptive policies. In addition, when the policy is preemptive, we show that the distribution of the total service time of a customer that enters the slowest server is hyper-exponential with expectation equal to the average service rate, independent of the allocation of the capacity.

Keywords: Queues, capacity allocation; loss system.

1 Introduction

We consider the problem of determining the optimal capacity and service allocations in a Markovian loss system, i.e., an M/M/k/k queueing model. The objective is to maximize the profit of the system’s owner, by choosing the number of servers, $k$, their capacities, $\mu_i$, and the admission fee, $p$.

Loss systems (also called loss queues) are queueing systems with no waiting positions, i.e., new arrivals when all servers are busy are rejected. Loss systems model various components in a wide range of applications, such as telephone systems, computer networks, optical networks, and cloud computing, see, for example, [4, 17, 21, 39].

The widespread view is that pooling demand, i.e., combining queues into a single queue, yields lower average waiting time. Some proofs appear in the literature (see, e.g., Rothkopf and Rech [33] and the references therein; we remark that they also present some cases where combining queues is disadvantageous). An interesting related question that arises in a single queue system with total service capacity $\mu$ is how many servers to employ and how to allocate the capacity among them, in order to maximize profit. Of course, if the system allows an infinite queue, it is best to allocate the whole capacity to a single server. In this paper we consider the other extreme where there is no queue, i.e., a loss system.

We study capacity allocation in both preemptive and non-preemptive loss systems. In the former case, it
is possible at any time to preempt service of a customer and immediately resume it on a different server from the same point. We assume that there are no switching costs involved with reassigning customers to servers. This assumption is often realistic, for example in cloud computing, these costs are often negligible. Our main results are: The non-preemptive case: (i) for identical servers, we find the optimal profit-maximizing number of servers (Theorem 2.5). (ii) for non-identical servers, we derive a necessary and sufficient condition for a uniform allocation of service capacity to be optimal, when there are two servers (Theorem 2.20). (iii) for large scale service systems with identical servers, we compute the asymptotic order of growth of the optimal number of servers and service rate (Theorem 2.9). (iv) for the case of more than two servers, we state a conjecture a necessary condition for the existence of an optimal uniform allocation, and provide numerical support for our conjecture (Conjecture 2.25).

The preemptive case: (i) We characterize the distribution of the service time, $W^{(k)}$, of a customer who begins service at the $k$th server (Theorem 2.26). (ii) we show that $E(W^{(k)})$ is the same as in the case of identical servers in the non-preemptive case, irrespective of the allocation (Corollary 2.27). (iii) we show that the optimal solution allocates the capacity to one server, and that uniform allocation minimizes the expected profit (Theorem 2.28). (iv) we derive the optimal number of servers and show that it is consistent with the social profit maximizing buffer size in Naor’s model [31] (Theorem 2.30).

The formal definition of our model is as follows.

1.1 Model description

We consider an M/M/$k$/$k$ loss system with arrival rate $\lambda$ and a fixed budget $\mu$ for servers. The decision variables of the system’s owner are the admission fee, denoted $p$, and the number and quality (as reflected by service rate) of the servers in order to maximize its profits. Specifically,

1. The arrival process is Poisson with rate $\lambda$, and the service time of each server is exponentially distributed. The total service rate is $\mu$.

2. Each arriving customer chooses between joining the fastest free server and balking. If all servers are busy, the customer is rejected. The service rates are public knowledge.

3. Customers are risk neutral, maximizing the expected net benefit.

4. A customer who joins the system has to pay an admission fee $p$. In addition, he incurs a cost of $C$ per unit time in the system. The customer’s benefit from completed service is $R$.

5. The system’s owner has total available capacity $\mu$. Its decision variables are the number of servers $k$, ...
the service rates \( \mu_1, \mu_2, \ldots, \mu_k \) \( (\sum_{i=1}^{k} \mu_i = \mu) \), and the admission fee \( p \). Servers are numbered so that

\[
\mu_1 \geq \mu_2 \geq \cdots \geq \mu_k.
\]

Suppose that the system has \( k \) servers. The optimal admission fee \( p \) can be obtained as follows. Clearly, under an optimal solution, all servers are active, meaning that an arriving customer will enter the system even if the only free server is the slowest one. Let \( W^{(k)} \) be the total service time of a customer who begins service at the slowest server, \( k \). A customer would be ready to join the slowest server if

\[ R \geq CE(W^{(k)}) + p. \]

To achieve optimal profit, the system’s owner will increase the admission fee to the maximum level such that this inequality, i.e.,

\[ p = R - CE(W^{(k)}). \] (1)

Let \( \pi_k \) denote the steady-state probability that all the \( k \) servers are busy, and let \( Z^{(k)}(\cdot) \) denote the profit of the system’s owner. Then,

\[ Z^{(k)}(\lambda, p, \pi_k) = \lambda p (1 - \pi_k). \] (2)

We also use the following normalized system parameters: \( \rho = \lambda/\mu \), and \( \nu_e = \frac{R\mu}{\rho} \).

Throughout the paper, we use the notation, \( h(n) = \Theta(g(n)) \) if there exist \( k_1, k_2 > 0 \) and \( n_0 \) such that for all \( n > n_0 \),

\[ k_1 \cdot g(n) \leq h(n) \leq k_2 \cdot g(n). \]

### 1.2 Literature review

Capacity management has been studied in many contexts, and applications are described in Schweitzer and Seidmann [34]. Models of capacity allocation assume that there is a fixed capacity to be allocated among service stations, in order to optimize some performance measure. The literature on capacity allocation can be partitioned into discrete models of optimal allocations of servers among multi-service stations, and models where the fraction of the total capacity allocated to each station is a continuous decision variable. The first group includes Rolfe [32], Stecke and Solberg [38], Shanthikumar and Yao [36], Dallery and Stecke [10], Green and Guha [13], Hillier and So [18], Hung and Posner [20], Zhang, Berman, Marcotte, and Verter [43].

The present paper belongs to the second group, and we now describe this literature. A main difference between these models and ours is that they assume that customers routing is independent of the system’s state, whereas we assume that an arriving customer joins the fastest available server. We also consider the option of reassigning a customer when a faster server completes service. A similar approach has been followed by Xu and Shanthikumar [41] who compute the optimal admission control policy when servers are indexed (for example, in decreasing order of their service rate) and a new customer is assigned to the lowest
indexed available customer, if one exists. Similar to most of the literature on routing demand to servers (see, for example, Bell and Stidham [7]), they do not allow reassigning customers. Reassignment is allowed in by Akgun, Down, and Righter [2] but in their model the capacity allocation is given and the problem is to determine the admission policy. Kleinrock [23] finds the vector of service rates \( \mu = (\mu_1, \ldots, \mu_k) \) in a Jackson network, that minimizes the sojourn time per customer, subject to a budget constraint \( D = \sum_k d_k \mu_k \), where \( d_k \) is the unit cost of capacity at station \( k \) and \( D \) the total available budget. The paper also reviews earlier work on capacity assignment. Wein [40] generalizes this result to general arrival and service time distributions. Ahmadi [1] and Kostami and Ward [28] deal with the optimal capacity levels for the rides in an amusement park during different time periods subject to a budget constraint.

Related literature consider combined routing and capacity allocation decisions. However, in our model, the routing is determined by the strategic users, whereas in the models below, it is determined by the manager. See Shanthikumar and Xu [35], and §6 in Altman [5]. Korilis, Lazar, and Orda [27] (see also [26]) consider a finite number of users, each wishing to minimize the expected delay of its demand by splitting it among a set of parallel heterogeneous M/M/1 servers with known service rates. The manager has available an extra amount of capacity to be allocated among the servers. The minimum total equilibrium wait in this game is obtained by allocating the additional capacity exclusively to a server which has the highest initial capacity. This result resembles ours for the case of preemptive regime (see Theorem 2.28), though for a different model.

Other related literature consider dedicated servers, i.e., there are classes of customers or demands, and a server can serve a specific class. This is in contrast to our model, in which the demand is homogenous and each server can serve each customer.

Glasserman [12] considers allocating a given capacity among several items produced by a manufacturer. Production of each item follows a base-stock policy. The manufacturer’s objective is to minimize holding costs subject to a service-level constraint. Similarly, Hong and Lee [19] consider the profit maximization way of splitting a given capacity between two servers subject to linear demand function with substitution effects, waiting costs, and lateness penalties imposed when exogenously given delay guarantees are violated. De Kok [25] considers a packaging facility whose capacity must be allocated among different package sizes. Almeida, Almeida, Ardagna, Cunha, Francalanci, and Trubian [3] compute the optimal admission of multi-class demand and allocation of a given capacity among the dedicated servers serving each class, when value is generated from a customer only if its sojourn time does not exceed the service level agreement of its class. Yolken and Bambos [42] analyze capacity allocation in a Markovian queue. The system operator has a fixed capacity to be split among \( N \) users. User \( i \) dedicates the capacity allocated to it to serve its demand. Its
delay sensitivity is unknown to the system operator, and the allocation of capacity is done through a bidding mechanism.

Other works combine dedicated and non-dedicated servers. Chao, Liu, and Zheng [8] describe a capacity allocation model motivated by health-care management. There are $N$ service stations, and station $i$ has dedicated nonswitchable rate of demand. In addition, there is also a stream of non-dedicated switchable customers. Mandjes and Timmer [30] consider competition between two Internet providers. Each competitor has a fixed capacity which it can split among several subnetworks which differ by their capacity and price, and then each customer chooses one of the offered subnetworks, or balks. The authors examine the equilibrium outcomes of this game. Similarly, Shumsky and Zhang [37] study a multi-period capacity allocation model. Initially there is a given amount of capacity, and the decision maker sequentially allocates capacity after observing demand within each period.

In [22], Karsten, Slikker and van Houtum show that pooling demand is beneficial in loss systems in which servers are allowed to form coalitions. Each coalition optimizes the number of servers it employs. The cost for a coalition is the sum of costs of operating the servers and penalties on lost customers. However, it is shown that the game may have an empty core – there is no stable coalition because there is no stable way to allocate the costs.

An interesting empirical result appears in the recent paper of Lu, Musalem, Olivares and Schilkrut [29] from 2013. They conducted an experiment in a physical queue and found that customers’ decisions of joining the queue are mainly affected by its length, without adjusting it to the service speed. Therefore, since pooling multiple queues into a single queue increases the queue length, it leads to higher loss probability and lower revenues. This empirical finding conforms with our theoretical conclusions.

Our asymptotic results are related to large scale service systems, introduced by Atar [6]. A collection of large scale queue systems is called a heavy traffic regime if $\rho^{(n)} \to 1$ under a second order condition (see [6] and references therein). In the first common regime, named the conventional regime (see Chen and Yao [9]), the number of the servers does not change. In contrast, the regime introduced by Halfin and Whitt [14] considers systems with a large number of servers, while the individual service rates do not change ([14]). Atar [6] introduces the $\alpha$-parametrization. To describe it, he discusses a single-class queueing model $M/M/N$ with parameters $\lambda^n$, $\mu^n$ and $N^n$ depending on some scaling parameter $n$. Let the external arrival rate increase as $\Theta(n^{\alpha})$. Given some $\alpha \in [0,1]$, assume that the number of servers is $\Theta(n^\alpha)$, while each of the individual service rates scales as $n^{1-\alpha}$. In addition, let a suitable critical load condition hold. Then the extremal cases $\alpha = 0$ and $\alpha = 1$ correspond to the conventional and Halfin-Whitt regimes, respectively. The optimal regime in our model is appropriate (as we show) for the case $\alpha = \frac{2}{7}$.
The rest of this paper is organized as follows. The main results, with their theoretical development, appear in §2. §3 provides some directions for future research. Finally, §4 contains the proofs.

2 Main results

2.1 Non-preemptive policy: identical service rates

In this section, we assume that the service rates must be equal, i.e., \( \mu_i = \mu/k, \ i = 1, \ldots, k \). We remark that the results in this section also hold for \( M/G/k/k \) loss systems.

We denote the system’s profit when using \( k \) identical servers by \( Z_k \), and the optimal number of servers by \( k^* \), i.e.,

\[
 k^* := \arg\max \{ Z_k : k = 0, 1, 2, \ldots \}.
\] (3)

Consider a fixed \( k \). Since the servers are identical, the mean service time of a customer is

\[
 E(W^{(k)}) = \frac{1}{\mu_i} \equiv \frac{k}{\mu}.
\] (4)

Substituting in (1), we obtain that the admission fee is

\[
 p = R - \frac{Ck}{\mu}.
\] (5)

Substituting in (2), the profit of the system’s owner is

\[
 Z_k = \lambda p(1 - \pi_k) = \lambda \left( R - \frac{Ck}{\mu} \right) (1 - \pi_k) = C \rho (\nu_c - k) (1 - \pi_k).
\] (6)

Recall that for identical servers, \( \pi_k \) is given by the well-known Erlang’s loss formula: (see, e.g., [24] page 105):

\[
 \pi_k = \pi_k \equiv \frac{(k \rho)^k / k!}{\sum_{l=0}^{k} (k \rho)^l / l!}.
\] (7)

(The bar symbol emphasizes the fact that the servers are identical.) Substituting \( \pi_k \) in (6) yields

\[
 Z_k = C \rho (\nu_c - k) (1 - \pi_k) = C \rho (\nu_c - k) \left( 1 - \frac{\frac{(k \rho)^k / k!}{\sum_{l=0}^{k} (k \rho)^l / l!}}{1 - \frac{\frac{(k \rho)^k / k!}{\sum_{l=0}^{k} (k \rho)^l / l!}}{1 - \pi_k}} \right).
\] (8)

It is immediate from (5) that as \( k \) increases, the admission fee \( p \) decreases and, by Lemma 2.1 below, \( 1 - \pi_k \) increases. Namely, there are opposing influences on the profit \( Z_k \). Hence, given \( \nu_c \) and \( \rho \), it is interesting to find the profit maximizing number \( k^* \) of (identical) servers.

We begin by showing that if the number of servers grows, the probability \( \pi_k \) that all servers are busy becomes smaller; recall that all proofs are given in §4.
Lemma 2.1 \( \{\pi_k\}_{k=0}^{\infty} \) is a decreasing sequence, \( \lim_{k \to \infty} \pi_k = \frac{e^{-1}}{\rho} \) if \( \rho > 1 \), and \( \lim_{k \to \infty} \pi_k = 0 \) if \( \rho \leq 1 \). In particular, for \( \rho = 1 \), \( \pi_k = \Theta(k^{-1/2}) \).

Another important property of \( \pi_k \) is convexity in \( k \), which was proved by Harel [15]; specifically:

Lemma 2.2 (Theorem 6 in [15]) For any fixed \( \rho > 0 \), \( \pi_k \) is strictly convex as a function of \( k \), that is, \( \{\pi_k - \pi_{k+1}\}_{k=0}^{\infty} \) is monotone decreasing.

The following function \( f(\cdot) \) will play a key role in our analysis; it is used for a characterization of the optimal number of servers for profit maximizing:

Definition 2.3

\[
f(k) := k + \frac{1 - \pi_{k-1}}{\pi_{k-1} - \pi_k}, \quad k = 1, 2, ...
\]

and let \( f(0) := 0 \).

Lemma 2.4 For any \( \rho > 0 \), \( \{f(k)\}_{k=1}^{\infty} \) is an increasing sequence, and \( \lim_{k \to \infty} f(k) = \infty \).

Theorem 2.5 Given \( \nu_e, \rho > 0 \), an optimal number of identical servers, \( k^* \), satisfies:

\[
f(k^*) \leq \nu_e \leq f(k^* + 1).
\] (9)

Note that Lemma 2.4 implies that (9) admits a solution, but it need not be unique: (9) admits two solutions if and only if \( \nu_e = f(l) \) for some integer \( l \); otherwise it admits a unique solution. However, in the (very unlikely) two-solutions-case, there is no preference of one solution over the other, so we choose to break ties and take the minimum feasible solution. Therefore, from now on we assume:

Assumption 2.6 \( k^* \) is the (unique) solution to

\[
f(k^*) \leq \nu_e < f(k^* + 1);
\] (10)

by Theorem 2.5, \( k^* \) is an optimal number of identical servers.

Example 2.7 For example, we derive the conditions for \( k^* = 0, 1, 2, 3 \) by writing the probabilities \( \pi_0, \pi_1, \pi_2, \pi_3 \) explicitly:

\[
\begin{align*}
\pi_0 &= 1, & \pi_1 &= \frac{\rho}{1 + \rho}, & \pi_2 &= \frac{2\rho^2}{1 + 2\rho + 2\rho^2}, & \pi_3 &= \frac{4.5\rho^3}{1 + 3\rho + 4.5\rho^2 + 4.5\rho^3}, \\
\pi_4 &= \frac{10.5^2 \rho^3}{1 + 4\rho + 8\rho^2 + 10.5^2 \rho^3 + 10.5^2 \rho^3}.
\end{align*}
\] (11)
We obtain a necessary and sufficient condition for \( k = 0 \) to be optimal (i.e., it is better to close the service system):

\[
f(0) = 0 < \nu_e < 1 = f(1),
\]

for \( k = 1 \) to be optimal:

\[
f(1) = 1 < \nu_e < \frac{1 + 2\rho + 2\rho^2}{\rho} = f(2),
\]

for \( k = 2 \) to be optimal:

\[
f(2) = 2 + \frac{1 + 2\rho + 2\rho^2}{\rho} \leq \nu_e < 3 + \frac{(1 + 3\rho + 4.5\rho^2 + 4.5\rho^3)(1 + 2\rho)}{2\rho^2 + 1.5\rho^3} = f(3), \tag{12}
\]

and for \( k = 3 \) to be optimal:

\[
f(3) = 3 + \frac{(1 + 3\rho + 4.5\rho^2 + 4.5\rho^3)(1 + 2\rho)}{2\rho^2 + 1.5\rho^3} \leq \nu_e < 4 + \frac{1 - \pi_3}{\pi_3 - \pi_4} = f(4).
\]

**Large scale systems**

We conclude this section by considering a large scale service system, that is, a collection of loss systems such that in the \( n \)th system, \( n = 1, 2, \ldots \), the total arrival and service rates are

\[
\lambda^{(n)} = \lambda \cdot n + o(n), \quad \mu^{(n)} = \mu \cdot n + o(n),
\]

and \( \lambda = \mu \).

Therefore,

\[
\rho^{(n)} \to 1, \quad \nu_e^{(n)} = \frac{R \mu^{(n)}}{C} = \Theta(\mu^{(n)}) = \Theta(n).
\]

Denote the corresponding optimal number of servers by \( k^{(n)} \). By Theorem 2.5 and Assumption 2.6, \( k^{(n)} \) satisfies the conditions

\[
f(k^{(n)}) \leq \nu_e^{(n)} < f(k^{(n)} + 1). \tag{13}
\]

We further investigate the case \( \rho = 1 \). For this purpose, we look at the sequence \( \{\pi_k \cdot k^{1/2}\} \)\(^{\infty}_{k=1} \). We will show that it converges (to a finite limit.) First of all, it is bounded, since the inequality (36) (from the proof of Lemma 2.1) is equivalent to

\[
\frac{1}{2.54} \leq \pi_k \cdot k^{1/2} \leq 4. \tag{14}
\]

Next, we will prove that its corresponding ratio sequence, \( \left\{ \frac{\pi_{k+1} \cdot (k+1)^{1/2}}{\pi_k \cdot k^{1/2}} \right\} \)\(^{\infty}_{k=1} \), converges. For this task, note that we can express its \( k \)th element as:

\[
\frac{\pi_{k+1} \cdot (k+1)^{1/2}}{\pi_k \cdot k^{1/2}} = \frac{\pi_{k+1}}{\pi_k} \cdot \left( 1 + \frac{1}{k} \right)^{1/2}.
\]

The right multiplicand converges to 1. As for the left multiplicand, we use the following Lemma, whose proof relies on another result by Harel [15]:

8
Lemma 2.8 When \( \rho = 1 \), the ratio sequence \( \{\frac{\pi_{k+1}}{\pi_k}\}_{k=1}^{\infty} \) is increasing, and \( \lim_{k \to \infty} \frac{\pi_{k+1}}{\pi_k} \leq 1 \).

Lemma 2.8 implies that \( \{\frac{\pi_{k+1}}{\pi_k}\}_{k=1}^{\infty} \) converges, hence \( \left\{\frac{\pi_{k+1} \cdot (k+1)^{1/2}}{\pi_k k^{1/2}}\right\}_{k=1}^{\infty} \) converges as well. This fact and the bounds (14) imply that the sequence \( \{\pi_k \cdot k^{1/2}\}_{k=1}^{\infty} \) converges; denote this limit by \( b \). Then \( \pi_k \approx b \cdot k^{-1/2} \), that is, \( \pi_k = b \cdot k^{-1/2} + o(k^{-1/2}) \). Hence,

\[
f(k) \approx k + \frac{1 - b}{\sqrt{k-1}} = k + \frac{1}{b} \frac{k^{1/2} - k - \sqrt{k}}{\sqrt{k - 1} - \sqrt{k}} \cdot \frac{\sqrt{k} + \sqrt{k - 1}}{\sqrt{k}} = k + \left(\frac{1}{b} \frac{k^{1/2} - k - \sqrt{k}}{(\sqrt{k} - 1)}\right) = \Theta(k^{3/2}).
\]

Therefore, \( f(k^{(n)}) + f(k^{(n)} + 1) = \Theta(k^{(n)3/2}) \), so the bounds (13) yield \( \nu_c^{(n)} = \Theta(k^{(n)3/2}) \). But \( \nu_c^{(n)} = \Theta(n) \), hence we obtain \( n = \Theta(k^{(n)3/2}) = \Theta(\mu^{(n)}) \). Thus, we have established:

**Theorem 2.9** When \( \rho = 1 \), the optimal number of servers \( k^{(n)} \) grows as \( n^{2/3} \) and individual service rate grows as \( n^{1/3} \).

### 2.2 Non-preemptive policy: different service rates

Here we allow the \( k \) servers to have different service rates. Recall that the service rates are sorted, that is,

\[
\mu_1 \geq \mu_2 \geq \cdots \geq \mu_k \quad (\mu_1 + \mu_2 + \cdots + \mu_k = \mu).
\]

We assume that the routing policy is non-preemptive, i.e., after a customer is assigned to a particular server, he cannot jockey to another server. Hence, the customer’s strategy is to enter the available server with the highest index \( i \) such that \( R \geq \frac{C}{\mu_i} + p \).

Since the \( k \)th server is the slowest, \( E(W^{(k)}) = \frac{1}{\mu_k} \), so the optimal strategy of the system’s owner is to set the admission fee by \( p = R - \frac{C}{\mu_k} \). (see (1)). We seek an optimal allocation, so we assume that all servers are active, hence \( \mu_k \geq \frac{C}{R} \) and \( p \geq 0 \). The core difficulty of our analysis lies in the computation of \( \pi_k \); Erlang’s loss formula is no longer valid, since the servers are nonidentical, i.e., \( \pi_k = \pi_k(\mu_1, \ldots, \mu_k) \). Substituting in (2), we obtain the following expression of the profit of the system’s owner:

\[
Z^{(k)}(\mu_1, \ldots, \mu_k) \equiv C \rho \left(\nu_c - \frac{\mu}{\mu_k}\right)(1 - \pi_k(\mu_1, \ldots, \mu_k)).
\]

(Note that \( \mu \) is not an argument parameter of \( Z^{(k)} \) since it is the sum of \( \mu_1, \ldots, \mu_k \); alternatively, we could have used \( Z^{(k)}(\mu_2, \ldots, \mu_k; \mu) \).)

#### 2.2.1 Normalizing parametrization

In the rest of §2.2, it will be convenient to use the parametrization obtained by normalizing the rates: Let

\[
d_1 = \frac{\mu_1}{\mu}, \ldots, d_k = \frac{\mu_k}{\mu} \quad (d_1 + \cdots + d_k = 1, \quad d_1 \geq d_2 \geq \cdots \geq d_k > 0).
\]

\( (16) \)
Note that \((\mu_1, \ldots, \mu_k) \neq \left(\frac{1}{k}, \ldots, \frac{1}{k}\right)\) if and only if \(d_k < \frac{1}{k}\).

Adjusting (15) to this parametrization yields:

\[
Z^{(k)}(d_2, \ldots, d_k) = C \rho \left( \nu_e - \frac{1}{d_k} \right) (1 - \pi_k(d_2, \ldots, d_k)),
\]

(17)

where \(d_1\) is not an argument of both \(Z^{(k)}\) and \(\pi_k\) since \(d_1 = 1 - \sum_{i=2}^{k} d_i\).

We are interested in values of \(d_k\) for which the profit is positive, hence we assume that \(\nu_e > k\) (so \(\frac{1}{
u_e} < \frac{1}{k}\)), and denote the relevant domain of \(Z^{(k)}(\cdot)\) as follows:

**Definition 2.10** Let \(I^{(k)} \equiv \left(\frac{1}{\nu_e}, \frac{1}{k}\right]\) denote the effective domain of \(Z^{(k)}(\cdot)\).

### 2.2.2 Two servers

Here, the service rates of the two servers are \(\mu_1\) and \(\mu_2 = \mu - \mu_1\), where \(\mu_1 \geq \mu_2 > 0\), i.e., \(\frac{1}{2} \mu \leq \mu_1 < \mu\). Its corresponding normalizing parametrization (see (16)) is:

\[
d_1 = \frac{\mu_1}{\mu}, \quad d_2 = \frac{\mu_2}{\mu} \quad (d_1 + d_2 = 1, \quad 0 < d_2 \leq \frac{1}{2} \leq d_1 < 1),
\]

Let \(\pi_2(d_2)\) denote the steady state probability that both servers are busy (as a function of \(d_2\)), and let \(Z(d_2)\) denote the corresponding system’s profit. Applying (17), we obtain:

\[
Z(d_2) = Z^{(2)}(d_2) = C \rho \left( \nu_e - \frac{1}{d_2} \right) (1 - \pi_2(d_2)).
\]

(18)

(For convenience, since in this subsection we only consider two servers, we omit the superscript ‘2’. ) Note that

\[
\pi(\frac{1}{2}) = \pi_2, \quad Z(\frac{1}{2}) = Z_2.
\]

Our goal is to characterize the situations in which the profit of a system with two identical servers is larger than that of a system with two nonidentical servers (for any allocation of the total capacity). We begin by deriving the exact expression for \(\pi_2(d_2)\):

**Lemma 2.11**

\[
\pi_2(d_2) = \frac{2\rho^2}{2\rho^2 + 2\rho + (4\rho + 2) \frac{d_2 - d_2^2}{\rho + d_2}}.
\]

(19)

Substituting in (18) yields:

\[
Z(d_2) = C \rho \left( \nu_e - \frac{1}{d_2} \right) \left( 1 - \frac{2\rho^2}{2\rho^2 + 2\rho + (4\rho + 2) \frac{d_2 - d_2^2}{\rho + d_2}} \right).
\]
Using multiplication and composition of functions, we can express $Z(d_2)$ as the following decomposition of three functions:

$$Z = Z_1(Z_2 \circ Z_3),$$

or explicitly,

$$Z(d_2) = Z_1(d_2)(Z_2(Z_3(d_2))),$$

where

$$Z_1(x) = C\rho \left( \nu_e - \frac{1}{x} \right), \quad Z_2(y) = 1 - \frac{2\rho^2}{2\rho^2 + 2\rho + y}, \quad Z_3(z) = (4\rho + 2)\frac{z - z^2}{\rho + z}. \quad (21)$$

It is easily verified that the derivatives of $Z_1(\cdot), Z_2(\cdot), Z_3(\cdot)$ are:

$$Z'_1(x) = C\rho \frac{x}{x^2}, \quad Z'_2(y) = \frac{2\rho^2}{(2\rho^2 + 2\rho + y)^2}, \quad Z'_3(z) = -(4\rho + 2)\frac{(z^2 + 2z\rho - \rho)}{(\rho + z)^2}. \quad (22)$$

We assume that $\nu_e > 2$ and the effective domain is $I^{(2)} \equiv \left( \frac{1}{\nu_e}, \frac{1}{2} \right]$; see Definition 2.10.

**Lemma 2.12** Let $\rho, \nu_e > 0$.

(i) $Z_1(\cdot), Z_2(\cdot)$ are monotone increasing on $I^{(2)}$,

(ii) $Z_3(\cdot)$ is monotone increasing on $\left( \frac{1}{\nu_e}, \sqrt{\rho^2 + \rho - \rho} \right]$, and monotone decreasing on $\left[ \sqrt{\rho^2 + \rho - \rho}, \frac{1}{2} \right]$.

(iii) For all $\rho, \nu_e > 0$, $Z(\cdot)$ is monotone increasing on $\left( \frac{1}{\nu_e}, \sqrt{\rho^2 + \rho - \rho} \right]$.

Part (iii) of this Lemma implies that $d^*_2$, the maximizer of $Z(d_2)$, must be in $\left[ \sqrt{\rho^2 + \rho - \rho}, \frac{1}{2} \right]$. We emphasize that $d^*_2$ need not equal $\frac{1}{2}$, i.e., an equal allocation is not always optimal. However, an equal allocation is relatively good in the following sense:

**Corollary 2.13** For all $\rho, \nu_e > 0$,

$$0 < d_2 \leq \frac{\rho}{2\rho + 1} \implies Z(d_2) < Z\left( \frac{1}{2} \right) = Z_2. \quad (23)$$

In order to further investigate $Z(\cdot)$ in $\left[ \sqrt{\rho^2 + \rho - \rho}, \frac{1}{2} \right]$, we turn to establish some concavity properties.

**Definition 2.14** Let $Z_{23} := 1 - \pi_2 \equiv Z_2 \circ Z_3$, that is,

$$Z_{23}(d_2) := 1 - \pi_2(d_2) = 1 - \frac{2\rho^2}{2\rho^2 + 2\rho + (4\rho + 2)\frac{d_2 - d_2^*}{\rho + d_2}}.$$

**Lemma 2.15** (i) $Z_1(\cdot), Z_2(\cdot), Z_3(\cdot)$ are concave on $I^{(2)}$, (ii) $Z_{23}(\cdot)$ is concave on $I^{(2)}$ and attains a maximum at $\sqrt{\rho^2 + \rho - \rho}$ (if this value is inside $I^{(2)}$).

**Corollary 2.16** $Z(\cdot)$ is concave on $\left[ \sqrt{\rho^2 + \rho - \rho}, \frac{1}{2} \right]$.
This concavity property implies the following equivalent condition for the superiority of identical servers:

**Lemma 2.17** For all $\rho, \nu_e > 0$,

\[ Z(d_2) < \bar{Z}_2 \text{ for all } 0 < d_2 < \frac{1}{2} \iff Z'(\frac{1}{2}) > 0. \]

The condition $Z'(\frac{1}{2}) > 0$ is further reduced to the following condition, which is much simpler to analyze:

**Definition 2.18** Let $g : [0, \infty) \to [0, \infty)$ be defined by:

\[ \forall \rho > 0, \quad g(\rho) := 8\rho^2 + 16\rho + 18 + \frac{8}{\rho} + \frac{1}{\rho^2}. \]

**Lemma 2.19** For all $\rho, \nu_e > 0$,

\[ Z'(\frac{1}{2}) > 0 \iff \nu_e < g(\rho). \]

Combining Lemmas 2.17 and 2.19, we obtain our main result at once – a necessary and sufficient condition for the superiority of two identical servers:

**Theorem 2.20** For all $\rho, \nu_e > 0$,

\[ Z(d_2) < \bar{Z}_2 \text{ for all } 0 < d_2 < \frac{1}{2} \iff \nu_e < g(\rho). \]

An immediate consequence of this Theorem and the fact that $Z(\cdot)$ is increasing on $(\frac{1}{\nu_e}, \sqrt{\rho^2 + \rho} - \rho)$ (by Lemma 2.12(iii)) is:

**Corollary 2.21** If $\nu_e \geq g(\rho)$, then $d^*_2$, the maximizer of $Z(d_2)$, satisfies:

\[ \frac{\rho}{2\rho + 1} \leq \frac{1}{\nu_e}, \sqrt{\rho^2 + \rho} - \rho \leq d^*_2 \leq \frac{1}{2}. \]

It is easily verified that for any $\rho > 0$, the difference $g(\rho) - f(3)$ is positive, monotone increasing (in $\rho$), and diverges; see (12) for the expression of $f(3)$. Thus, the following sufficient condition holds:

**Corollary 2.22** For all $\rho, \nu_e > 0$,

\[ \nu_e \leq f(3) \implies Z(d_2) < \bar{Z}_2 \text{ for all } 0 < d_2 < \frac{1}{2}. \]

Although this result is weaker than Theorem 2.20, we state it because we conjecture that it holds for an arbitrary number of servers as well; see Conjecture 2.24 below.

We conclude our analysis by noting that for any $\nu_e$, the condition of Theorem 2.20 holds for $\rho$ sufficiently large. Therefore:

**Theorem 2.23** For any $\nu_e > 0$, \[ \lim_{\rho \to \infty} \argmax_{d_2 \in (\frac{1}{\nu_e}, \frac{1}{2})} Z(d_2) = \frac{1}{2}. \]
2.2.3 An arbitrary number of servers

As we have seen so far, the various computations for the relatively simple case of \( k = 2 \) servers are fairly involved. In this section we propose two conjectures for an arbitrary number of servers. We provide some preliminary evidence that motivates them.

Our first conjecture is a straightforward generalization of Corollary 2.22, and provides a sufficient condition for the superiority of identical servers when their number is fixed. We use the normalizing parametrization (16) and its corresponding profit expression (17); recall that \((\mu_1,...,\mu_k) \neq (\frac{1}{k},...,\frac{1}{k})\) if and only if \(d_k < \frac{1}{k}\).

**Conjecture 2.24** For any fixed number \( k \) of servers and \( \rho > 0 \), in a non-preemptive policy,

\[
\nu_e \leq f(k + 1) \implies Z^{(k)}(d_2,...,d_k) < Z_k \text{ for all } 0 < d_k < \frac{1}{k}.
\]

In addition to its correctness for two servers (Corollary 2.22), further support for its correctness stems from numerical evidence for the case of \( k = 3 \) servers. Specifically, by (17),

\[
Z^{(3)}(d_2,d_3) = C\rho \left(\nu_e - \frac{1}{d_3}\right) \left(1 - \pi_3(d_2,d_3)\right),
\]

where \(\pi_3(d_2,d_3)\) is the probability that all three servers are busy. Using numerical computations, we can see that Conjecture 2.24 holds in this case, namely, for all \( \rho > 0 \),

\[
\nu_e \leq f(4) \implies Z^{(3)}(d_2,d_3) < Z_3 \text{ for all } 0 < d_3 < \frac{1}{3}.
\]

Our second conjecture concerns situations where \( k \) is not specified, i.e., it is a decision variable (as well as \( \mu_1,...,\mu_k \)). We conjecture that in this case, the optimal solution consists of identical servers:

**Conjecture 2.25** For any \( \rho,\nu_e > 0 \), when the policy is non-preemptive, the profit maximizing solution consists of identical servers; their number, \( k^* \), is given in Theorem 2.5.

We provide some motivation for its correctness by a special case. Let \( \rho > 0 \), and assume that \( \nu_e \leq f(4) \) and \( k \leq 3 \). We will see that in this case, the optimal solution consists of identical servers. We distinguish two cases depending on the value of \( \nu_e \):

- **Case 1.** \( \nu_e \leq f(3) \): By Corollary 2.22 , for any \( 0 < d_2 < \frac{1}{2} \),

\[
Z^{(2)}(d_2) < Z^{(2)}(\frac{1}{2}) = Z_2.
\]

Similarly, since \( \nu_e \leq f(4) \), (24) implies that for any \( 0 < d_3 < \frac{1}{3} \),

\[
Z^{(3)}(d_3) < Z^{(3)}(\frac{1}{3}) = Z_3.
\]
Let $k^*$ be the optimal number of identical servers corresponding to $\rho, \nu_e$ (see (3)). By Theorem 2.5, $k^* \leq 3$ (as $\nu_e \leq f(4)$). The claim follows.

- Case 2. $f(3) < \nu_e \leq f(4)$: By Theorem 2.5, $k^* = 3$. First of all, as in Case 1, for any $0 < d_3 < \frac{1}{3}$,

$$Z(3)(d_3) < Z(3)(\frac{1}{3}) = Z_3.$$ 

At this point, it remains to show that no solution of two non-identical servers yields a higher profit, i.e., for any $0 < d_2 < \frac{1}{2}$,

$$Z(2)(d_2) < Z_3.$$ 

The approach used in Case 1 turns out to be only partially beneficial; specifically, we verified numerically that there exists $0.1952 < \rho_0 < 0.1953$ such that $\rho > \rho_0 \iff f(4) < g(\rho)$. Thus, if $\rho > \rho_0$, then $\nu_e < g(\rho)$, so Theorem 2.20 implies that for any $0 < d_2 < \frac{1}{2}$,

$$Z(2)(d_2) < Z(2)(\frac{1}{2}) = Z_2 < Z_3,$$

hence the claim follows.

So assume that $\rho < \rho_0$. Then altogether we have $f(4) > \nu_e > g(\rho) > f(3) > 3$. Substituting (6),(18), we obtain that (25) is equivalent to:

$$\left(\nu_e - \frac{1}{d_2}\right)(1 - \pi_2(d_2)) < (\nu_e - 3)(1 - \pi_3).$$

Note that this inequality involves $\nu_e, d_2, \rho$ (see (19),(11) for the explicit expressions of $\pi_2(d_2), \pi_3$ respectively), so it is not clear how to verify its correctness. However, we verified numerically that if we let $d_2 = \sqrt{\rho^2 + \rho - \rho}$, which is the maximizer of $1 - \pi_2(d_2)$ by Lemma 2.15(ii), then this inequality holds for the two extreme possible values of $\nu_e$, i.e., $\nu_e \in \{g(\rho), f(4)\}$; specifically, for all $\frac{1}{\nu_e} < \rho < \rho_0$:

$$\left(\frac{g(\rho) - 1}{\sqrt{\rho^2 + \rho - \rho}}\right)(1 - \pi_2(\sqrt{\rho^2 + \rho - \rho})) < (g(\rho) - 3)(1 - \pi_3),$$

$$\left(\frac{f(4) - 1}{\sqrt{\rho^2 + \rho - \rho}}\right)(1 - \pi_2(\sqrt{\rho^2 + \rho - \rho})) < (f(4) - 3)(1 - \pi_3).$$

### 2.3 Loss systems with preemptive policy

In this section we consider an M/M/$k$/$k$ loss system with preemptive routing policy. Recall that the $k$ service rates are $\mu_1 \geq \mu_2 \geq ... \geq \mu_k$ ($\sum_{i=1}^k \mu_i = \mu$). Thus, the preemption assumption implies that when server $i$ becomes idle, every customer that obtains service from a slower server $j$ immediately jockeys to server $j-1$ ($j = i + 1, i + 2, ..., k$), i.e., all customers of servers $i + 1, i + 2, ..., k$ (if any), are immediately shifted one server down. Consequently, at any point in time, the set of active servers is $\{1, ..., i\}$ for some $i$. 

14
Recall that \( W^{(k)} \) is the total service time of a customer who begins service at the slowest server \( k \). In our system, \( W^{(k)} \) has the following form:

**Theorem 2.26** \( W^{(k)} \) is a hyper-exponential random variable with density

\[
f_{W^{(k)}}(x) = \sum_{l=1}^{k} u_l \left( \sum_{i=1}^{l} \mu_i \right) e^{-\sum_{i=1}^{l} \mu_i} x,
\]

where \( u_1, ..., u_k \) are parameters such that \( \sum_{l=1}^{k} u_l = 1 \).

The expected service time, \( E(W^{(k)}) \), of a customer who joins the slowest server \( k \) is easily computed in the two extreme cases. Surprisingly, they yield the same quantity, which is the same as in the identical-server case with non-preemption; see (4) in §2.1:

- **Case 1.** All servers have the same rate, i.e., \( \mu_i = \frac{\mu}{k} \) for all \( i = 1, ..., k \). This case is equivalent to the non-preemptive policy due to the memoryless property, so the expectation of a single service time with rate \( \frac{\mu}{k} \) is \( \frac{k}{\mu} \).
- **Case 2.** A unique server operates with the total rate \( \mu \) and all other servers operate with zero rate, i.e., they function as standby positions. Here, a customer who joins the \( k \)th server will leave the system after \( k \) service periods, hence his expected service time is \( \frac{k}{\mu} \) as well.

It turns out that this expected service time is the same for any allocation of the service rates:

**Corollary 2.27** For any allocation \( \mu_1 \geq \cdots \geq \mu_k \) (\( \sum_{i=1}^{k} = \mu \)),

\[
E(W^{(k)}) = \frac{k}{\mu_1 + \cdots + \mu_k} = \frac{k}{\mu}.
\]

(27)

Using this Corollary, we can substitute (27) in (1) and obtain that for any allocation \( \mu_1, ..., \mu_k \), the optimal entry fee is

\[
p = R - \frac{Ck}{\mu}.
\]

(28)

Note that this is the same as in the identical-server case studied in §2.1; see (5). Consequently, the corresponding profit of the system’s owner is:

\[
\bar{Z}^{(k)}(\mu_1, ..., \mu_k) = \lambda \left( R - \frac{Ck}{\mu} \right) (1 - \pi_k(\mu_1, ..., \mu_k)),
\]

where \( \pi_k(\mu_1, ..., \mu_k) \) is the probability that all servers are busy; the tilde symbol emphasizes that the policy is preemptive.
It turns out that the optimal allocation in the preemptive case is the exact opposite of the one in the non-preemptive case, namely, it is optimal to have a single server operating with the full service capacity $\mu$ while all the other $k-1$ servers are standby positions, and the equal-allocation is the worst. Formally:

**Theorem 2.28** For any $\mu_1 \geq \cdots \geq \mu_k$ with $\sum_{i=1}^{k} \mu_i = \mu$,

$$\tilde{Z}(k)(\mu, 0, 0, \ldots, 0) \geq \tilde{Z}(k)(\mu_1, \ldots, \mu_k) \geq Z_k,$$

where $Z_k$ is the profit of a system with equal service rates as in §2.1; see (8).

We provide the proof of Theorem 2.28 below (rather than in §4) because we believe that its key step may be of independent interest. To this end, let $\mu_1, \ldots, \mu_k$ be any allocation with $\mu_1 \geq \cdots \geq \mu_k$, $\sum_{i=1}^{k} \mu_i = \mu$.

Due to the memoryless property, $\tilde{\pi}_k(\mu/k, \ldots, \mu/k) = \pi_k$ (where $\pi_k$ is the probability corresponding to equal service rates in the nonpreemptive policy; see (7)). It then follows from (29) that in order to prove the stated inequalities, it suffices to show that

$$1 - \tilde{\pi}_k(\mu, 0, \ldots, 0) \geq 1 - \tilde{\pi}_k(\mu_1, \ldots, \mu_k) \geq 1 - \tilde{\pi}_k(\mu/k, \ldots, \mu/k),$$

or equivalently,

$$\tilde{\pi}_k(\mu, 0, \ldots, 0) \leq \tilde{\pi}_k(\mu_1, \ldots, \mu_k) \leq \tilde{\pi}_k(\mu/k, \ldots, \mu/k).$$

(30)

Our method for establishing (30) is by viewing our system as a birth-death process. The shifting mechanism that was described in the beginning of this section implies that the states of this system are $0, 1, \ldots, k$, where state $i$ means that there are $i$ customers who are being served by the first (i.e., fastest) servers. It is immediate that the birth rates (of states $0, 1, \ldots, k-1$) are all $\lambda$. As for the death rates, let $S_l \sim \exp(\mu_l)$ denote the service time of a customer at the $l$th server, $l = 1, \ldots, k$. Recall that if servers $1, \ldots, i$ are busy, then the service time of the first of them to finish is distributed as $\min\{S_1, \ldots, S_i\} \sim \exp(\mu_1 + \cdots + \mu_i)$.

Therefore, those rates, corresponding to states $1, 2, \ldots, k$ respectively, are $\mu_1, \mu_1 + \mu_2, \ldots, \mu_1 + \cdots + \mu_{k-1}, \mu$; see Figure 1. Consequently, $\tilde{\pi}_k(\mu_1, \ldots, \mu_k)$ is exactly the steady-state probability of being in state $k$ (note that the preemption assumption was used in obtaining the death rates). Hence, it remains to establish:

**Lemma 2.29** For the above birth-death process, the inequalities (30) hold for any $\mu_1 \geq \cdots \geq \mu_k$ with $\sum_{i=1}^{k} \mu_i = \mu$.

See §4 for an algebraic proof of this Lemma. In addition, we provide a short intuitive argument as follows.

Let $1 \leq i < j \leq k$ and consider the $i$th and $j$th servers. Suppose that $\mu_j > 0$. Then by transferring some service rate from the $j$th server to the $i$th server, the corresponding death rates of states $i, \ldots, j-1$ increase, while all other are unchanged. Consequently, each of the $k$ death rates is maximized when $\mu_1$ is as large as
Figure 1: State diagram for our birth-death process.

possible, and is minimized when \( \mu_k \) is as large as possible. The allocation \((\mu_1, ..., \mu_k) = (\mu, 0, ..., 0)\) clearly maximizes \( \mu_1 \); the allocation \((\mu_1, ..., \mu_k) = (\mu/k, ..., \mu/k)\) maximizes \( \mu_k \) (due to the condition \( \mu_1 \geq ... \geq \mu_k \)). This establishes Lemma 2.29 and hence completes the proof of Theorem 2.28.

Thus, \( \tilde{Z}(k)(\mu, 0, 0, ..., 0) \) is the maximum profit of the system’s owner, while \( Z_k \) is the minimum. In other words, it is best to allocate the total service capacity \( \mu \) to a single server, and it is worst to allocate it equally among all servers.

Note that the results in this section assume that the number of servers, \( k \), is given and fixed. We conclude this section by analyzing the optimization problem when considering their number as well. That is, we would like to maximize the profit \( \tilde{Z}(k)(\mu_1, ..., \mu_k) \) given in (29), where \( k \) is also a decision variable. By Theorem 2.28 and (29), this problem reduces to finding the optimal number \( k^* \) of servers that maximizes:

\[
\tilde{Z}(k)(\mu, 0, 0, ..., 0) \equiv \lambda \left( R - \frac{Ck}{\mu} \right) \left( 1 - \tilde{\pi}_k(\mu, 0, 0, ..., 0) \right).
\]

Interestingly enough, this \( k^* \) is equal to the buffer size in Naor’s model [31]; this follows since in this situation, our model is equivalent to Naor’s. Specifically, we showed that the optimal number of servers consists of a single server (with the total capacity \( \mu \)) and \( k-1 \) standby positions. This is equivalent to an M/M/1 queue with buffer size \( k-1 \). Therefore, the optimal solution \( k^* \) is equal to the profit maximizing buffer size in
Naor’s model [31]. Thus, we have established:

**Theorem 2.30** In an $M/M/k/k$ loss system with preemptive policy,

$$k^* = n_r,$$

where $n_r$ is the social profit maximizing buffer size in Naor’s model [31].

### 3 Discussion and future research

In this paper, we consider a multi-server system with non-preemptive and preemptive policy and show that optimal solutions for the quality of servers are opposite for the two policies. Under the non-preemptive policy, we provide preliminary evidence that the optimal solution consists of identical servers. Under the preemptive policy, it is optimal to allocate the total service rate to a single server while the other servers operate at zero service rate, and actually serve as standby positions. Below we present discussion and some possible directions for future research.

1. In §2.2, we established a necessary and sufficient condition for the superiority of *two identical servers*, when there are two available servers. Conjecture 2.24 suggests a sufficient condition for the superiority of identical servers when their number is fixed; Conjecture 2.25 suggests that when their number is not fixed, it is always optimal to use identical servers. The corresponding birth-death equations seem fairly complicated. Thus, a challenging direction is to try to establish their correctness.

2. Suppose that the policy is preemptive as in §2.3. However, the joining customer cannot observe the service rate at a server before joining (i.e., customers cannot observe the number of busy servers), but he knows the values $\mu_1, \ldots, \mu_k$ and whether there is an available server. Of course, the optimal strategy of the system’s administrator is to employ at any time the fastest servers. Now, the customers can compute their expected service time given that there is an available server. This may mean a probabilistic equilibrium joining strategy in which some of the customers give up service without trying – as in [11]. Hence, the effective arrival rate to the system would be $P\lambda$ for some $P \leq 1$, and the customers take this equilibrium value into account. The firm takes this into account while computing its profit maximizing price. As in §2.3, the optimal solution will be a single active server with standby positions.

3. Note that if the servers have identical capacities, or the customers cannot observe the service rate of a particular server to whom they are assigned, the optimal price will be such that the expected
customer’s surplus is 0. In this case the firm’s strategy is socially optimal, as in [11]. However, when the servers have different capacities and customers can observe them (as in §2.3), the price should be determined by the slowest server, leaving customers with positive surplus. In this case the firm’s objective differs from the socially optimal one, and its profit maximizing price induces non optimal behavior. It is of interest to compute the socially optimal price.

4. Recall the shifting mechanism in the preemptive model: whenever a customer completes service from server $i$, all the customers (if any) at servers $j$ with $j > i$ are simultaneously transferred one server down. Also recall that the expected service time, $E(W^{(k)})$, is the same (identically $\mu/k$) for any capacity allocation (see Corollary 2.27).

An interesting problem is to design a different, non-shifting, mechanism that transfers customers between servers. For example, whenever a customer completes service at a server $i$ which is not the slowest (among the busy servers), we could transfer the customer from the currently slowest server to $i$. This way, we may be able to decrease $E(W^{(i)})$, and consequently increase the admission fee and hence the profit of the system’s owner (due to (1)). Note also that by doing it we decrease the gap between the monopolistic and social objectives, and the profit maximizing policy gets closer to optimize social welfare. Moreover, perhaps a more sophisticated mechanism will achieve equal expected service times, namely, for all $i = 0, \ldots, k$, $E(W^{(i)})$ will not depend on $i$ (the number of busy servers).

In this case, the profit maximizing solution is also socially optimal. Obtaining such a mechanism is a challenging combinatorial problem.

5. Other natural extensions include endogenous determination of the budget used for paying the servers, incurring additional switching costs, and, of course, allowing waiting in a queue when all servers are busy.

4 Proofs

**Proof of Lemma 2.1** Define $a_k := \pi_k^{-1}$, hence

\[
a_k = \frac{1}{(k\rho)^k/k!} \sum_{l=0}^{k} (k\rho)^l / l!
= 1 + \frac{1}{\rho} + \left(1 - \frac{1}{k}\right) \frac{1}{\rho^2} + \left(1 - \frac{1}{k}\right) \left(1 - \frac{2}{k}\right) \frac{1}{\rho^3} + \cdots + \left(1 - \frac{1}{k}\right) \cdots \left(1 - \frac{k-1}{k}\right) \frac{1}{\rho^k}.
\]

When $k$ increases, each factor $\left(1 - \frac{l}{k}\right)$ of $a_k$ increases. It follows that $\{a_k\}_{k=0}^{\infty}$ is an increasing sequence, and so $\{\pi_k\}_{k=0}^{\infty}$ is a decreasing sequence.
Suppose $\rho > 1$. Then $a_k < \sum_{l=0}^{k} \left(\frac{1}{\rho}\right)^l \rightarrow \frac{1}{1 - \frac{1}{\rho}} = \frac{\rho}{\rho - 1}$. Thus, the sequence $\{a_k\}_{k=0}^{\infty}$ is bounded and increasing, so it converges to a finite limit, denoted $\alpha$. Of course, $\alpha \leq \frac{\rho}{\rho - 1}$. Now, for every $\epsilon > 0$ and for every $N$, there exists $k$ such that $\sum_{l=0}^{N} \left(\frac{1}{\rho}\right)^l \leq \alpha < \frac{\rho}{\rho - 1}$. Since $\alpha$ is constant, letting $\epsilon \rightarrow 0$ and $k \rightarrow \infty$ gives $\alpha = \frac{\rho}{\rho - 1}$. Consequently, $\pi_k \rightarrow \frac{\rho - 1}{\rho}$.

Suppose $\rho \leq 1$. Then for every $N$, there exists $k$ such that each of the first $2N$ terms is larger than $\frac{1}{2}$, so that, $a_k > N$. Thus, the sequence $\{a_k\}_{k=0}^{\infty}$ is unbounded and increasing, so $a_k \rightarrow \infty$ and hence $\pi_k \rightarrow 0$.

Now let $\rho = 1$. We will show that $a_{x^2} = \Theta(s)$. It is easy to see that $a_{x^2} = \sum_{t=0}^{s-1} \sum_{l=t+1}^{s} \prod_{i=1}^{l} (1 - \frac{i}{s^2})$. Hence,

$$a_{x^2} \leq s \sum_{t=0}^{s-1} \prod_{i=1}^{t+1} \left(1 - \frac{i}{s^2}\right),$$

and

$$s \left[ \sum_{t=1}^{s} \prod_{i=1}^{t} \left(1 - \frac{i}{s^2}\right) \right] \leq a_{x^2}.$$ (33)

By the AM–GM inequality, for every positive integers $l, n$:

$$\left(\prod_{i=1}^{l} \left(1 - \frac{i}{n}\right)\right)^{1/l} \leq \frac{1}{l} \sum_{i=1}^{l} \left(1 - \frac{i}{n}\right) = \frac{1}{l} \left(l - \frac{1}{n} \sum_{i=1}^{l} i\right) = \frac{l}{2}.$$

Substituting $n = s^2$ and $l = ts + 1$, and using the inequality $1 - x \leq e^{-x}$, we obtain:

$$\prod_{i=1}^{ts+1} \left(1 - \frac{i}{s^2}\right) \leq \left(1 - \frac{ts+2}{2s^2}\right)^{ts+1} \leq e^{-\frac{ts+2}{2s^2}(ts+1)} \leq e^{-\frac{1}{2}t^2}.$$

Using this inequality in (32) yields:

$$a_{x^2} \leq s \sum_{t=0}^{s-1} e^{-\frac{1}{2}t^2} = s \sum_{t=0}^{s-1} (e^{-\frac{1}{2}})^t \leq s \sum_{t=0}^{\infty} (e^{-\frac{1}{2}})^t = \frac{s}{1 - e^{-1/2}}.$$ (34)

Now, (33) yields:

$$a_{x^2} \geq s \left[ \sum_{t=1}^{s} \prod_{i=1}^{t} \left(1 - \frac{i}{s^2}\right) \right] = s \left[ \prod_{i=1}^{s} \left(1 - \frac{i}{s^2}\right) + \sum_{t=2}^{s} \prod_{i=1}^{t} \left(1 - \frac{i}{s^2}\right) \right] \geq s \left[ \prod_{i=1}^{s} \left(1 - \frac{i}{s^2}\right) \right] = s \left(1 - \frac{s}{2s^2}\right)^s = s \left(1 - \frac{1}{s}\right)^s.$$ (35)

Combining (34),(35), we obtain that for all $s > 2$,

$$0.25 \leq \left(1 - \frac{1}{s}\right)^s \leq \frac{a_{x^2}}{s} \leq \frac{1}{1 - e^{-1/2}} = 2.54.$$ (36)

Substituting $k = s^2$, we obtain that $0.25 \sqrt{k} \leq a_k \leq 2.54 \sqrt{k}$, which implies that $\pi_k \equiv a_k^{-1} = \Theta(k^{-1/2})$. This completes the proof. □

**Proof of Lemma 2.4** Recall that $f(k) = k + \frac{1 - \pi_k^{-1}}{\pi_k - 1 - \pi_k^{-1}}$. In order to establish the monotonicity, consider the second term, $\frac{1 - \pi_k^{-1}}{\pi_k - 1 - \pi_k^{-1}}$. The numerator is increasing (in $k$) since by Lemma 2.1, $\{\pi_k\}_{k=0}^{\infty}$ is decreasing. The
denominator is increasing since by Lemma 2.2, \( \{\pi_k - \pi_{k+1}\}_{k=0}^{\infty} \) is decreasing. It follows that \( \{f(k)\}_{k=1}^{\infty} \) is an increasing sequence.

As for the limit of \( f(k) \), note that the first term in \( f(k) \), which is \( k \), trivially converges to \( \infty \), and the second term is nonnegative. As we just proved, \( \{f(k)\}_{k=1}^{\infty} \) is an increasing sequence, hence \( \lim_{k \to \infty} f(k) = \infty. \)

\[ \square \]

**Proof of Theorem 2.5** We first show that \( k^* \) is a solution to (9). Recall that \( Z_k = C \rho (\nu_e - k)(1 - \pi_k) \) (see (8)) and that \( f(k) = k + \frac{1-\pi_{k-1}}{\pi_k} \) for \( k > 0 \) and \( f(0) = 0 \) (see Definition 2.3).

By definition of \( k^* \), it satisfies (in particular): \( Z(k^*) \geq Z(k^* - 1) \) and \( Z(k^*) \geq Z(k^* + 1) \). We derive an equivalent condition to the first of them:

\[
Z(k^*) \geq Z(k^* - 1) \iff (\nu_e - k^*)(1 - \pi_{k^*}) \geq (\nu_e - k^* + 1)(1 - \pi_{k^* - 1})
\[
\iff (\pi_{k^* - 1} - \pi_{k^*})(\nu_e - k^*) \geq 1 - \pi_{k^* - 1}
\[
\iff \nu_e \geq k + \frac{1 - \pi_{k^* - 1}}{\pi_{k^* - 1} - \pi_{k^*}} = f(k^*),
\]

that is, \( Z(k^*) \geq Z(k^* - 1) \iff \nu_e \geq f(k^*) \). By substituting \( k^* + 1 \) for \( k^* \) and reversing the inequality, we can obtain:

\[
Z(k^* + 1) \leq Z(k^*) \iff \nu_e \leq f(k^* + 1).
\]

Thus, \( k^* \) is a solution to (9): \( f(k) \leq \nu_e \leq f(k + 1). \)

\[ \square \]

**Proof of Lemma 2.8** Theorem 1 in Harel [15] states that the function

\[
f_1(k) := \frac{k!}{k^k} \sum_{l=0}^{k-1} \frac{k^l}{l!}
\]

is strictly concave in \( k \) \( (k = 1, 2, \ldots) \). Define \( a_k := \pi_k^{-1} \), that is (recall that \( \rho = 1 \)):

\[
a_k := \frac{k!}{k^k} \sum_{l=0}^{k} \frac{k^l}{l!} = \frac{k!}{k^k} \cdot \sum_{l=0}^{k-1} \frac{k^l}{l!} + \frac{k!}{k^k} \cdot \frac{k^k}{k!} = f_1(k) + 1.
\]

It follows that \( \{a_k\}_{k=1}^{\infty} \) is concave as well, that is, \( \{a_k - a_{k+1}\}_{k=1}^{\infty} \) is increasing. Equivalently, for any \( k \),

\[
2a_k > a_{k-1} + a_{k+1}. \tag{37}
\]

We will show that the ratio sequence \( \{\frac{a_{k+1}}{a_k}\}_{k=1}^{\infty} \) is decreasing. The inequality (37) implies that

\[
a_k > \frac{a_{k-1} + a_{k+1}}{2} \geq \sqrt{a_{k-1}a_{k+1}},
\]

where the second inequality is the AM–GM inequality. Consequently, \( a_k^2 > a_{k-1}a_{k+1} \), which is equivalent to

\[
\frac{a_k}{a_{k-1}} > \frac{a_{k+1}}{a_k}.
\]
Thus, \( \{\frac{a_k}{n_k}\}_{k=1}^{\infty} \) is decreasing, so \( \{\frac{\pi_{a_k}}{\pi_{n_k}}\}_{k=1}^{\infty} \) is increasing.

It remains to establish convergence. The sequence \( \{\pi_k\}_{k=1}^{\infty} \) is decreasing by Lemma 2.1, so for any \( k, \frac{\pi_{k+1}}{\pi_k} < 1 \). Thus, \( \{\frac{\pi_{k+1}}{\pi_k}\}_{k=1}^{\infty} \) is increasing and bounded from above by 1, so \( \lim_{k \to \infty} \frac{\pi_{k+1}}{\pi_k} \leq 1 \). \( \square \)

**Proof of Lemma 2.11** First, we denote by \( \pi_0(\mu_1, \mu_2), \pi_1(\mu_1, \mu_2), \pi_2(\mu_1, \mu_2) \), \( \pi_3(\mu_1, \mu_2) \), the limiting probability that the system is empty, the fast server serves a customer, the slow server serves a customer, and both servers are busy, respectively. The equations for our birth-death system are (we write in short \( \pi_i = \pi_i(\mu_1, \mu_2) \))

\[
-\lambda \pi_0 + \mu_1 \pi_1 + (\mu - \mu_1) \pi_2 = 0, \\
-(\lambda + \mu_1) \pi_1 + \lambda \pi_0 + (\mu - \mu_1) \pi_2 = 0, \\
-(\lambda + (\mu - \mu_1)) \pi_2 + \mu_1 \pi_2 = 0, \\
-\mu_2 + \lambda \pi_1 + \lambda \pi_2 = 0, \\
\pi_0 + \pi_1 + \pi_2 + \pi_2 = 1.
\]

Equivalently,

\[
-\rho \pi_0 + (1 - d_2) \pi_1 + d_2 \pi_2 = 0, \\
-(\rho + 1 - d_2) \pi_1 + \rho \pi_0 + d_2 \pi_2 = 0, \\
-(\rho + d_2) \pi_2 + (1 - d_2) \pi_2 = 0, \\
-\pi_2 + \rho \pi_1 + \rho \pi_2 = 0, \\
\pi_0 + \pi_1 + \pi_2 + \pi_2 = 1.
\]

The solution of the system for \( \pi_2(\mu_1, \mu_2) = \pi_2(d_2) \) is \( \pi_2(d_2) := \frac{2 \rho^2}{2d^2 + d_2} \). For \( \mu_1 = \mu_2 \), we have \( d_2 = \frac{1}{2} \) (see (11)), \( \pi_2(\frac{1}{2}) = \frac{2 \rho^2}{2d^2 + 2d^2+1} \). \( \square \)

**Proof of Lemma 2.12** Parts (i),(ii) follow immediately from the derivatives (22). For part (iii), note that by parts (i),(ii), all the three functions \( Z_1(\cdot), Z_2(\cdot), Z_3(\cdot) \) are monotone increasing on \( \left( \frac{1}{\sqrt{\rho}}, \sqrt{\rho^2 + \rho - \rho} \right) \). The claim now follows from the decomposition (20) of \( Z(\cdot) \) and the fact that multiplication and composition of positive functions preserve monotonicity. \( \square \)

**Proof of Corollary 2.13** Fix \( 0 < d_2 \leq \frac{\rho}{2d^2+1} \). First of all, Lemma 2.12(iii) implies that \( Z(d_2) \leq Z(\frac{\rho}{2d^2+1}) \) (it is easily verified that \( \frac{\rho}{2d^2+1} < \sqrt{\rho^2 + \rho - \rho} \)). To complete the proof, we will show that \( Z(\frac{\rho}{2d^2+1}) < Z(\frac{1}{2}) \).

It is easily verified that \( Z_3(\frac{\rho}{2d^2+1}) = Z_3(\frac{1}{2}) = 1 \), and hence \( Z_2 \circ Z_3(\frac{\rho}{2d^2+1}) = (Z_2 \circ Z_3)(\frac{1}{2}) \). However, \( Z_1(\frac{\rho}{2d^2+1}) < Z_1(\frac{1}{2}) \) by Lemma 2.12(i). Now the inequality immediately follows from the decomposition \( Z = Z_1(Z_2 \circ Z_3) \) (see (20)). This completes the proof. \( \square \)
Proof of Lemma 2.15 (i) As for $Z_1(\cdot), Z_2(\cdot)$, it is clear that $Z'_1(x), Z'_2(y)$ (see (22)) are monotone decreasing. As for $Z_3(\cdot)$, it is easily verified that the second derivative is:

$$Z''_3(z) = -\frac{4\rho(2\rho + 1)(\rho + 1)}{(\rho + z)^3},$$

which is negative for any $z > 0$.

(ii) By part (i), $Z_2(\cdot), Z_3(\cdot)$ are concave on $I^{(2)}$. By Lemma 2.12(i), $Z_2(\cdot)$ is increasing on $I^{(2)}$. It follows that $Z_{23} \equiv Z_2 \circ Z_3$ is concave in $I^{(2)}$ as well. As for the maximum, this follows immediately from Lemma 2.12(i),(ii), as $d_2 := \sqrt{\rho^2 + \rho - \rho} = 0$ maximizes $Z_3(\cdot)$, and $Z_2(\cdot)$ is increasing.

Proof of Corollary 2.16 We will show that $Z'(\cdot)$ is decreasing on $\left(\sqrt{\rho^2 + \rho - \rho}, \frac{3}{2}\right)$. Let $x, y \in \left(\sqrt{\rho^2 + \rho - \rho}, \frac{3}{2}\right]$ with $x < y$. We claim that the following inequality holds:

$$Z'(x) = (Z_1 Z_{23}(x))' = Z_1(x)Z'_{23}(x) + Z'_1(x)Z_{23}(x)
> Z_1(y)Z'_{23}(y) + Z'_1(y)Z_{23}(y) = Z'(y).$$

Consider the first term. $Z_{23}(\cdot)$ is concave (by Lemma 2.15(ii)), so $Z'_{23}(x) > Z'_{23}(y)$. However, it is easily verified that $Z_{23}(\sqrt{\rho^2 + \rho - \rho}) = Z'_3(\sqrt{\rho^2 + \rho - \rho}) = 0$, thus $0 \geq Z'_{23}(x) > Z'_{23}(y)$. Since $Z_1(x) < Z_1(y)$ (by Lemma 2.12(i)), it follows that $Z_1(x)Z'_{23}(x) > Z_1(y)Z'_{23}(y)$. As for the second term, $Z'_1(x) > Z'_1(y)$ by Lemma 2.15(i), and $Z_{23}(x) > Z_{23}(y)$ since in $\left(\sqrt{\rho^2 + \rho - \rho}, \frac{3}{2}\right], Z_3(\cdot)$ in decreasing and $Z_2(\cdot)$ is increasing (by Lemma 2.12(i),(ii)). This completes the proof.

Proof of Lemma 2.17 First assume that $Z'(\frac{1}{2}) \leq 0$. Then $Z(\cdot)$ is non-increasing on some interval that contains $\frac{1}{2}$, hence for $\varepsilon > 0$ sufficiently small, $Z(\frac{1}{2} - \varepsilon) \geq Z(\frac{1}{2})$.

Now assume that $Z'(\frac{1}{2}) > 0$. We show that in this case $Z(\cdot)$ is increasing on $I^{(2)}$, and so $d_2 = \frac{1}{2}$ is the unique maximizer of $Z(\cdot)$ over $I^{(2)}$. By Lemma 2.12(iii), $Z(\cdot)$ is increasing on $\left(\frac{1}{\sqrt{\rho^2 + \rho - \rho}}, \sqrt{\rho^2 + \rho - \rho}\right]$. By Corollary 2.16, $Z(\cdot)$ is concave on $\left(\sqrt{\rho^2 + \rho - \rho}, \frac{1}{2}\right]$, and this implies that $Z(\cdot)$ is increasing on $\left(\sqrt{\rho^2 + \rho - \rho}, \frac{1}{2}\right]$ as well, since $Z'(x) > Z'(\frac{1}{2}) > 0$ for any $\sqrt{\rho^2 + \rho - \rho} < x < \frac{1}{2}$. This completes the proof.

Proof of Lemma 2.19 Applying the product rule to the decomposition $Z = Z_1Z_{23}$, we obtain:

$$Z(\frac{1}{2}) = Z_1(\frac{1}{2}) Z'_{23}(\frac{1}{2}) + Z'_1(\frac{1}{2}) Z_{23}(\frac{1}{2}). \tag{38}$$

We proceed by calculating the four elements in the right hand side (using (21),(22)):

$$Z_1(\frac{1}{2}) = C\rho(\nu - 2),$$
$$Z'_1(\frac{1}{2}) = 4C\rho,$$
$$Z_{23}(\frac{1}{2}) = Z_2(Z_3(\frac{1}{2})) = Z_2(1) = 1 - \frac{2\rho^2}{2\rho^2 + 2\rho + 1} = \frac{2\rho + 1}{2\rho^2 + 2\rho + 1}.$$
and it remains to compute $Z'_{23}(\frac{1}{2})$. We specify the required auxiliary computations:

$$Z'_2(1) = \frac{2\rho^2}{(2\rho^2 + 2\rho + 1)^2}, \quad Z'_3(\frac{1}{2}) = -\frac{1}{\rho + \frac{1}{2}} = -\frac{2}{2\rho + 1}. $$

Using the chain rule:

$$Z'_{23}(\frac{1}{2}) = (Z_2 \circ Z_3)'(\frac{1}{2}) = Z'_2(Z_3(\frac{1}{2}))Z'_3(\frac{1}{2})$$

$$= Z'_2(1)Z'_3(\frac{1}{2}) = -\frac{4\rho^2}{(2\rho^2 + 2\rho + 1)^2(2\rho + 1)}. $$

Substituting in (38), we obtain:

$$Z'(\frac{1}{2}) = -\frac{4C\rho^3(\nu_e - 2)}{(2\rho^2 + 2\rho + 1)^2(2\rho + 1)} + \frac{4C\rho(2\rho + 1)}{2\rho^2 + 2\rho + 1}$$

$$= -\frac{4C\rho^3(\nu_e - 2) + 4C\rho(2\rho + 1)^2(2\rho^2 + 2\rho + 1)}{(2\rho^2 + 2\rho + 1)^2(2\rho + 1)}. $$

The sign of $Z'(\frac{1}{2})$ is determined by the numerator of the last expression. Thus,

$$0 < Z'(\frac{1}{2}) \iff 0 < -4C\rho^3(\nu_e - 2) + 4C\rho(2\rho + 1)^2(2\rho^2 + 2\rho + 1)$$

$$\iff 0 < -\rho^2(\nu_e - 2) + (2\rho + 1)^2(2\rho^2 + 2\rho + 1)$$

$$\iff \nu_e < 2 + \frac{(2\rho + 1)^2(2\rho^2 + 2\rho + 1)}{\rho^2}$$

$$\iff \nu_e < 2 + \frac{8\rho^4 + 16\rho^3 + 16\rho^2 + 8\rho + 1}{\rho^2}$$

$$\iff \nu_e < 2 + \frac{8\rho^2 + 16\rho + 18 + \frac{8}{\rho} + \frac{1}{\rho^2} \equiv g(\rho). (39)$$

Proof of Theorem 2.26 Let $f_{\text{exp}(\lambda)}(\cdot)$ denote the density function of an exponential random variable with parameter $\lambda$. We will derive the equivalent form of the density stated at the Lemma:

$$f_{W(k)}(x) = \sum_{l=1}^{k} u_l f_{\text{exp}(\sum_{i=1}^{n} \mu_i)}(x),$$

where $u_l$ are parameters such that $\sum_{l=1}^{k} u_l = 1$.

We will establish (40) by induction on $k$. Though the case $k = 1$ is trivial, we choose to describe the case $k = 2$ in detail.

Induction basis: $k = 2$. We will show that:

$$f_{W(2)}(x) = \frac{\mu_1}{\mu_2} f_{\text{exp}(\mu_1)}(x) + \left(1 - \frac{\mu_1}{\mu_2}\right) f_{\text{exp}(\mu_1 + \mu_2)}(x)$$

$$= \frac{\mu_1}{\mu_2} e^{-\mu_1 x} + \left(1 - \frac{\mu_1}{\mu_2}\right) (\mu_1 + \mu_2) e^{-(\mu_1 + \mu_2)x}. (41)$$

24
which is of the form (40) by letting \( u_1 = \frac{\mu_1}{\mu_2}, u_2 = 1 - u_1 \).

Let \( U \) be the minimum between the service times of the two servers, and \( V \) be the service time of the fast server. It is well known that, \( U \sim \exp(\mu_1 + \mu_2) \) and \( V \sim \exp(\mu_1) \). We write \( W^{(2)} = U + I_{\left( \frac{\mu_1}{\mu_1 + \mu_2} \right)} V \) where \( I_{\left( \frac{\mu_1}{\mu_1 + \mu_2} \right)} \) is Bernoulli random variable with success probability \( \frac{\mu_1}{\mu_1 + \mu_2} \), and present the distribution functions:

\[
F_U(x) = \begin{cases} 
1 - e^{-(\mu_1 + \mu_2)x} & x > 0, \\
0 & x \leq 0,
\end{cases}
\]

\[
F_{I_{\left( \frac{\mu_1}{\mu_1 + \mu_2} \right)}} (x) = \begin{cases} 
1 - \frac{\mu_1}{\mu_1 + \mu_2} e^{-\mu_1 x} & x > 0, \\
\frac{\mu_2}{\mu_1 + \mu_2} & x = 0, \\
0 & x < 0.
\end{cases}
\]

Now, we compute the density of \( W^{(2)} \) at \( x \) via convolution between

\[
f_{U - y} = \begin{cases} 
(\mu_1 + \mu_2) e^{-(\mu_1 + \mu_2)(x-y)} & x \geq y > 0, \\
0 & \text{otherwise},
\end{cases}
\]

and

\[
f_{I_{\left( \frac{\mu_1}{\mu_1 + \mu_2} \right)}} (y) = \begin{cases} 
\frac{\mu_2^2}{\mu_1 + \mu_2} e^{-\mu_1 y} & y > 0, \\
\frac{\mu_2}{\mu_1 + \mu_2} & y = 0, \\
0 & y < 0,
\end{cases}
\]

where \( x > 0 \). Since there is discontinuity at 0, and the probability \( P(I_{\left( \frac{\mu_1}{\mu_1 + \mu_2} \right)} V = 0) \) is positive, we first consider the range \((0, x]\) and then add the density of the event \( \{U = x, I_{\left( \frac{\mu_1}{\mu_1 + \mu_2} \right)} V = 0\} \). Thus:

\[
\int_{0^+}^{x} f_{U - y} f_{I_{\left( \frac{\mu_1}{\mu_1 + \mu_2} \right)}} (y) \, dy = \int_{0^+}^{x} \mu_1^2 e^{-(\mu_1 + \mu_2)(x-y)} e^{-\mu_1 y} \, dy = \mu_1^2 e^{-(\mu_1 + \mu_2)x} \int_{0^+}^{x} e^{\mu_2 y} \, dy
\]

\[
= \frac{\mu_2^2}{\mu_2} e^{-(\mu_1 + \mu_2)x} (e^{\mu_2 x} - 1) = \frac{\mu_1 \mu_2}{\mu_2} e^{-\mu_1 x} - \frac{\mu_2^2}{\mu_2} e^{-(\mu_1 + \mu_2)x},
\]

and

\[
P(I_{\left( \frac{\mu_1}{\mu_1 + \mu_2} \right)} V = 0)(\mu_1 + \mu_2) e^{-(\mu_1 + \mu_2)x} = \frac{\mu_2}{\mu_1 + \mu_2} (\mu_1 + \mu_2) e^{-(\mu_1 + \mu_2)x} = \mu_2 e^{-(\mu_1 + \mu_2)x}.
\]

Now (40) follows by summing the last two expressions.

**Induction step:** We assume that (40) holds for \( k - 1 \) servers and show that it holds for \( k \) servers as well.

We assume that the theorem is valid for \( W^{(k-1)} \) representing the total service time of customer who starts a service at the \((k-1)\)st server. Let \( U^k \) be the minimum service time of \( k \) servers, \( U^k \sim \exp(\mu_1 + \cdots + \mu_k) \). The waiting time for \( k \) servers is given by \( W^{(k)} = U^k + I_{\left( \frac{\mu_1 + \cdots + \mu_{k-1}}{\mu} \right)} W^{(k-1)} \). We compute the density of \( W^{(k)} \):

\[
\int_{0^+}^{x} f_{U^k - y} f_{I_{\left( \frac{\mu_1 + \cdots + \mu_{k-1}}{\mu} \right)}} (y) \, dy = \int_{0^+}^{x} \mu e^{-(\mu_1 + \cdots + \mu_k)x} \frac{\mu - \mu_k}{\mu} \sum_{i=1}^{k-1} u_i \left( \sum_{i=1}^{l} \mu_i \right) e^{-\left( \sum_{i=1}^{l} \mu_i \right) y} \, dy
\]

\[
= \sum_{l=1}^{k-1} u_l (\mu - \mu_k) \left( \sum_{i=1}^{l} \mu_i \right) e^{-(\mu_1 + \cdots + \mu_k)x} \int_{0^+}^{x} e^{\mu x} \left( \sum_{i=1}^{l} \mu_i \right) \, dy
\]

\[
= \sum_{l=1}^{k-1} u_l (\mu - \mu_k) \left( \sum_{i=1}^{l} \mu_i \right) e^{-(\mu_1 + \cdots + \mu_k)x} \left( e^{-\left( \sum_{i=1}^{l} \mu_i \right)x} - e^{\mu x} \right).
\]
Since the probability $P(I_{\frac{\nu_1+\ldots+\nu_{k-1}}{\mu}} W^{(k-1)} = 0) > 0$, we have to add the density of the event $\{U^k = x, I_{\frac{\nu_1+\ldots+\nu_{k-1}}{\mu}} W^{(k-1)} = 0\}$:

$$P(I_{\frac{\nu_1+\ldots+\nu_{k-1}}{\mu}} W^{(k-1)} = 0) \mu e^{-\mu x} = \frac{\mu_k}{\mu} \mu e^{-\mu x} = \mu_k e^{-\mu x}.$$  

Hence,

$$f_{W^{(k)}}(x) = \sum_{l=1}^{k-1} u_l \left(\frac{\mu - \mu_k}{\mu} \right)^l \left(\frac{\mu - \mu_k}{\mu} \right)^{l-1} e^{-\left(\sum_{i=1}^l \mu_i\right)x} - \frac{(\mu - \mu_k) \mu \left(\sum_{i=1}^l \mu_i\right)}{\mu - \left(\sum_{i=1}^l \mu_i\right)} e^{-\mu x} + \frac{\mu_k}{\mu} \mu e^{-\mu x}$$

$$= \sum_{l=1}^k u'_l f_{\exp(\sum_{i=1}^l \mu_i)}(x),$$

where $u'_l$ are coefficients of the random variables $\exp(\sum_{i=1}^l \mu_i), l = 1, \ldots, k$. By the following equations

$$\frac{(\mu - \mu_k)}{\mu - (\sum_{i=1}^l \mu_i)} - \frac{(\mu - \mu_k) \mu (\sum_{i=1}^l \mu_i)}{\mu - (\sum_{i=1}^l \mu_i)} + \frac{\mu_k}{\mu} = 1, \quad \sum_{l=1}^{k-1} u_l = 1,$$

the sum of the coefficients $u'_l, l = 1, \ldots, k$ is equal to one. This completes the proof. \qed

**Proof of Corollary 2.27** We use induction again. By (41), $E(W^{(2)}) = \frac{2}{\mu}$. For any $k$, we assume that $E(W^{(k-1)}) = \frac{k-1}{\mu - \mu_k}$. Clearly, $E(U^k) = \frac{1}{\mu}$ and the claim follows from $E(W^{(k)}) = E(U^k) + \frac{\mu - \mu_k}{\mu} E(W^{(k-1)})$. \qed

**Proof of Lemma 2.29** For convenience, let $\tilde{\pi}_k \equiv \tilde{\pi}_k(\mu_1, \ldots, \mu_k)$. Recall that for a $k$-state birth-death process with birth rates $\alpha_0, \ldots, \alpha_{k-1}$ and death rates $\beta_1, \ldots, \beta_k$, the steady-state probabilities are:

$$\tilde{\pi}_i = \frac{\alpha_0 \alpha_1 \cdots \alpha_{i-1}}{\beta_1 \beta_2 \cdots \beta_i} \pi_0, \quad i = 1, \ldots, k,$$

where

$$\pi_0 = \left(\frac{\sum_{j=0}^k \alpha_0 \alpha_1 \cdots \alpha_{j-1}}{\beta_1 \beta_2 \cdots \beta_j} \right)^{-1}.$$

(The first term in the summation, corresponding to $j = 0$, is 1.) Thus, the inverse of $\tilde{\pi}_k$ is:

$$\tilde{\pi}_k^{-1} = \sum_{j=0}^k \frac{\alpha_0 \alpha_1 \cdots \alpha_{j-1}}{\beta_1 \beta_2 \cdots \beta_j} = \sum_{j=0}^k \sum_{j+1}^k \beta_{j+1} \beta_{j+2} \cdots \beta_k \alpha_j \alpha_{j+1} \cdots \alpha_{k-1}.$$

Substituting $\alpha_0 = \ldots = \alpha_{k-1} = \lambda$ and $\beta_j = \mu_1 + \ldots + \mu_j$ ($j = 1, \ldots, k$), we obtain:

$$\tilde{\pi}_k^{-1} = \sum_{j=0}^k \frac{(\mu_1 + \ldots + \mu_{j+1}) \cdots (\mu_1 + \ldots + \mu_k)}{\lambda^{k-j}}.$$

Thus, we would like to solve the two optimization problems:

$$\max/\min \quad H(\mu_1, \ldots, \mu_k) = \sum_{j=0}^k \lambda^{j-k} (\mu_1 + \ldots + \mu_{j+1}) \cdots (\mu_1 + \ldots + \mu_k)$$

s.t. $\mu_1 + \ldots + \mu_k = \mu$

$\mu_1 \geq \ldots \geq \mu_k \geq 0$. 

26
These two problems can be solved by standard nonlinear programming methods such as Lagrange multipliers and KKT conditions; however, we take a simpler approach using elementary tools. We will show that the maximizer and minimizer are \((\mu_1, ..., \mu_k) = (\mu, 0, 0, ..., 0)\) and \((\mu_1, ..., \mu_k) = (\mu/k, ..., \mu/k)\), respectively, by showing that they are uniformly the maximizer and minimizer of each term in the objective function. Thus, consider the optimization problems corresponding to the \(j\)th term:

\[
\begin{align*}
\max / \min \quad & h_j(\mu_1, ..., \mu_k) \equiv \lambda^{i-k}(\mu_1 + \cdots + \mu_{j+1}) \cdots (\mu_1 + \cdots + \mu_k) \\
\text{s.t.} \quad & \mu_1 + \cdots + \mu_k = \mu \\
& \mu_1 \geq \cdots \geq \mu_k \geq 0.
\end{align*}
\]

**Solution of the maximization problem**

Note that each multiplicand of \(h_j(\cdot)\) is at most \(\mu\). It follows that \((\mu_1, ..., \mu_k) = (\mu, 0, 0, ..., 0)\) is a maximizer that does not depend on \(j\). (The maximum depends on \(j\), it is \((\mu/\lambda)^{k-j} = \rho^{i-k}\).)

**Solution of the minimization problem**

Denote \(S_{j+1} := \mu_1 + \cdots + \mu_{j+1}\). Then we can rewrite:

\[
h_j(\mu_1, ..., \mu_k) = S_{j+1}(S_{j+1} + \mu_{j+2}) \cdots (S_{j+1} + \mu_{j+2} + \cdots + \mu_k).
\]

Observe that for each \(i = j + 2, ..., k\), \(\mu_i\) appears in exactly \(k - i + 1\) multiplicands. It follows that any minimizing solution must satisfy \(\mu_{j+2} \leq \mu_{j+3} \leq \cdots \leq \mu_k\) (otherwise there exists an adjacent pair \(\mu_i > \mu_{i+1}\), buy swapping their values yields a better solution). Together with the inequality conditions, we obtain that \(\mu_{j+2} = \mu_{j+3} = \cdots = \mu_k\). Since \(S_{j+1}\) appears in all \(k - j\) multiplicands, it should be minimized. But by the equality condition, \(S_{j+1} + \sum_{i=j+2}^{k} \mu_i = \mu\). Thus, minimizing \(S_{j+1}\) is equivalent to maximizing \(\sum_{i=j+2}^{k} \mu_i\), and this is achieved by assigning \(\mu_k\), and hence \(\mu_{j+2}, ..., \mu_{k-1}\), its largest possible value, which is \(\mu/k\) (this is immediate from the two conditions). Consequently, \((\mu_1, ..., \mu_k) = (\mu/k, ..., \mu/k)\) is a minimizer that does not depend on \(j\). (The minimum depends on \(j\), it is \(\frac{k!\mu}{\prod_{i=0}^{k-1}(k-i)}\).)

Since these solutions are optimal for each \(h_j(\mu_1, ..., \mu_k)\), it follows that they are optimal for \(H(\mu_1, ..., \mu_k)\) as well. Thus, for any feasible \(\mu_1, ..., \mu_k\):

\[
\bar{\pi}_k^{-1}(\mu, 0, ..., 0) \geq \bar{\pi}_k^{-1}(\mu_1, ..., \mu_k) \geq \bar{\pi}_k^{-1}(\mu/k, ..., \mu/k),
\]

which is equivalent to \((30)\).

**Acknowledgments** This research was supported by the Israel Science Foundation grants no. 526/08 and 1015/11. We thank Gail Gilboa-Freedman for her inspiring discussions.

**References**


