Optimal service-capacity allocation in a loss system

Refael Hassin Yair Y. Shaki Uri Yovel [*]

January 14, 2015

Abstract

We consider a loss system with a fixed budget for servers. The system owner's problem is choosing the price, and selecting the number and quality of the servers, in order to maximize profits, subject to a budget constraint. We solve the problem with identical and different service rates as well as with preemptive and non-preemptive policies. In addition, when the policy is preemptive we prove the following conservation law: the distribution of the total service time for a customer entering 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

The prevailing view is that pooling demand, i.e., combining queues into a single queue, yields a lower average waiting time. Some proofs appear in the literature (see, e.g., Rothkopf and Rech [36] and the references therein; we note 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 μ is how many servers to employ and how to allocate capacity among them in order to maximize profit. Of course, if the system allows for an infinite queue, it is best to allocate all capacity to a single server. In this paper we consider the other extreme where there is no queue, i.e., a loss system.

We consider the problem of determining optimal capacity allocation among servers in a Markovian loss system, i.e., an M/M/k/k queueing model. We assume that the system manager has a given amount of capacity μ. The objective is to maximize the profit of the system by properly selecting the number of servers, k, capacities, μ_i, $\sum_i \mu_i = \mu$, and a (state-independent) admission fee, p. Customers incur waiting costs and act strategically by joining the fastest available server if and only if the service value exceeds expected waiting costs.

Loss systems (also called *loss queues*) are queueing systems with no waiting positions, i.e., when all servers are busy new arrivals are rejected. Loss systems model various components in a wide range of applications,

[*] Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69987, Israel. Uri Yovel passed away June 2014. hassin@post.tau.ac.il, yair_sh@shoresh.org.il

such as telephone systems, computer networks, optical networks, and cloud computing (see, for example, [4, 20, 24, 42]).

We study capacity allocation in both *preemptive* and *non-preemptive* loss systems. In the preemptive case it is possible at any time to preempt service of a customer and immediately resume it with a different server from the same point. We assume no switching costs involved with reassigning customers. This assumption is often realistic, for example in cloud computing, and these costs are often negligible.

Capacity allocation has been studied in many applications as we describe below.

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 λ and a fixed budget μ for servers. Specifically,

1. The arrival process is Poisson with rate λ, and the service time of each server is exponentially distributed. Total service rate is μ.

2. Arriving customers choose between joining the fastest free server or balking. Customers are rejected if all servers are busy. The number of servers and service rates are public knowledge.

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

4. A customer joining the system pays an admission fee p and incurs cost C per unit time in the system. The customer benefit from completed service is R.

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

6. Servers are numbered so that
$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_k. \tag{1}$$
With preemption allowed, when server i becomes idle, every customer obtaining service from a slower server j immediately jockeys to server $j-1$ ($j = i+1, i+2, \ldots, k$), i.e., all customers of servers $i+1, i+2, \ldots, k$ (if any), immediately shift one server down. Consequently, at any point in time the set of active servers is $\{1, \ldots, i\}$ for some i.

Suppose the system has k servers. Optimal admission fee p can be obtained as follows. Clearly, under an optimal solution all servers are active a positive fraction of the time, meaning 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 $R \geq C\boldsymbol{E}(W^{(k)}) + p$. To achieve optimal profit, the system owner will increase the admission fee to the maximum level such that

$$p^{(k)} = R - C\boldsymbol{E}(W^{(k)}). \tag{2}$$

Note that the way in which $\boldsymbol{E}(W^{(k)})$, and with it $p^{(k)}$, is determined depends on whether or not preemption is allowed. Without preemption, $\boldsymbol{E}(W^{(k)})$ is simply the expected service time at the slowest server, whereas with preemption one has to take into consideration the possibility of being reassigned to faster servers.

Let π_k denote the steady-state probability that all k servers are busy, and let $Z^{(k)}(\cdot)$ denote profit. Then,

$$Z^{(k)}(\lambda, p, \pi_k) = \lambda p(1 - \pi_k). \tag{3}$$

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

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 Discussion of the modeling assumptions

The main assumption of our model is that customers are not willing to wait and arrivals when all servers are busy are lost. In the non preemptive model the system manager must balance the desire to fully utilize all existing capacity when the system is not empty with the desire to reduce customer loss. The first goal encourages having a small number of fast servers, while the second goal gives incentive for having many servers. In contrast, the preemptive model does allow for more flexibility and solutions such as M/M/$s/s+k$ can be implemented by having s fast servers and k additional "servers" with zero capacity.

The Poisson arrival process is commonly accepted as a good description of many real life situations. We also assume service is exponentially distributed. This assumption is harder to justify, except that it is commonly used for tractability. However, we don't use this assumption in Section 2.1 because the probability that all servers are busy is independent of service distribution, at least when the servers are identical. On the other hand, we expect the solution of the preemptive model with a general service distribution to be very different and much more difficult. The reason being that estimating expected time in the system includes estimates of residual service time of currently engaged servers, based on their number. These estimates are difficult to obtain, even in the single server case, and it is not even clear whether or not a system with more busy servers means a longer expected sojourn time. (Similar difficulties already exist in the M/G/1 queue, where the equilibrium joining strategy is not necessarily a threshold strategy. See Altman and Hassin [6] and Kerner [25].)

We assume there is a fixed amount of capacity to be allocated to servers. A more realistic scenario would assume that total capacity of the system is also a decision variable. Thus, the model needs to be complemented by assuming a cost function, $C(\mu)$, for acquiring capacity μ. However, the solution to the problem with fixed capacity is clearly all one needs in order to solve this generalization, and therefore this is what we focus on.

We assume customers to be risk neutral. This assumption may not generally hold in real life cases, but it is commonly required to keep the problem tractable. Technically, with a nonlinear utility function even the simplest formulas, like the amount a customer is ready to pay for service, become complicated expressions depending on the whole service distribution rather than only on its mean value. This means that we will not get elegant formulas for the solution as we have now.

We assume the system is observable, i.e., an arriving customer knows the service rate of its intended server in addition to the queue discipline. An unobservable system will lead to very different analyses and results. For example, we expect that, unlike in the observable model, in the unobservable case it is never optimal to allocate capacity evenly among servers. In both cases, faster servers are more highly utilized and pooling is advantageous, but in the observable case the price is determined according to the slowest server, and this provides an incentive to equally allocate capacity. In the unobservable case, the price is determined by the unconditional expected waiting time, and therefore the incentive to increase the capacity of the slowest server is not as strong. We leave the solution of the unobservable model as an interesting topic for future research.

1.3 Main results

- *The non-preemptive case:*
  - For identical servers, we find the optimal profit-maximizing number of servers (Theorem 2.4).
  - For large scale service systems with identical servers, we compute the asymptotic order of growth in the optimal number of servers and the service rate (Theorem 2.6).
  - For non-identical servers, we derive a necessary and sufficient condition for uniform allocation of service capacity to be optimal when there are two servers (Theorem 3.2).
  - For the case of more than two servers, we conjecture a necessary condition for the existence of an optimal uniform allocation and provide the numerical support (Conjecture 3.6).

- *The preemptive case:*
  - We characterize the distribution of the service time, $W^{(k)}$, of a customer who begins service at the kth server (Theorem 4.1).
  - We show that $\boldsymbol{E}(W^{(k)})$ is the same as in the case of identical servers in the non-preemptive case, irrespective of the allocation (Corollary 4.2).
  - We show that the optimal solution allocates capacity solely to one server, and that uniform allocation minimizes expected profit (Theorem 4.3).
  - We derive the optimal number of servers and show that it is consistent with the social welfare maximizing buffer size in Naor's model [34] (Theorem 4.5).

1.4 Literature review

Capacity management has been studied in many contexts, and some relevant applications of which are described in Schweitzer and Seidmann [37]. Models of capacity allocation assume fixed capacity 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 capacity allocated each station is a continuous decision variable. The first group includes Rolfe [35], Stecke and Solberg [41], Shanthikumar and Yao [39], Dallery and Stecke [12], Green and Guha [16], Hillier and So [21], Hung and Posner [23], Baron, Berman and Krass [9], and Zhang, Berman, Marcotte, and Verter [47].

The present paper belongs to the second group, which we will now describe. A main difference between these models and ours is that they assume customer routing is independent of the system's state, whereas we assume an arriving customer joins the fastest available server. We also consider the option of *reassigning* a customer when a faster server becomes available. A similar approach has been followed by Xu and Shanthikumar [45] who compute the *socially optimal* admission control policy when servers *with given service rates* are indexed (for example, in decreasing order of their service rate) and a new customer is assigned to the lowest indexed available server, if one exists. In contrast, we consider profit maximization and service rates to be decision variables. Similar to most of the literature on routing demand to servers (see, for example, Bell and Stidham [8]), Xu and Shanthikumar do not allow for customer reassignment. Reassignment is allowed in Akgun, Down, and Righter [2], but in their model capacity allocation is given, and the problem is to determine the admission policy. Kleinrock [26] finds the vector of service rates $\mu = (\mu_1, \ldots, \mu_k)$ in a Jackson network that minimizes 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 is the total available budget. The paper also reviews earlier work on capacity assignment. Wein [43] generalizes this result to general arrival and service time distributions. Ahmadi [1] and Kostami and Ward [31] deal with optimal capacity levels for rides in an amusement park during different time periods subject to a budget constraint.

Related literature considers combined routing and capacity allocation decisions. However, in our model, the routing is determined by strategic users whereas in the models below it is determined by the manager. See Shanthikumar and Xu [38], and §6 in Altman [5]. Korilis, Lazar, and Orda [30] (see also [29]) consider a finite number of users, each wishing to minimize the expected delay by splitting demand among a set of parallel heterogeneous M/M/1 servers with known service rates. The manager has an extra amount of capacity available for allocation among the servers. The minimum total equilibrium wait in this game is obtained by allocating the additional capacity exclusively to the server having the highest initial capacity. This result resembles ours for the case of preemptive regime (see Theorem 4.3), though for a different model.

Other related literature deals with *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 demand is homogenous and each server can serve each customer. Glasserman [15] 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 [22] consider the profit maximizing 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 [28] considers a packaging facility whose capacity must be allocated among different package sizes. Almeida, Almeida, Ardagna, Cunha, Francalanci, and Trubian [3] compute optimal admission of multi-class demand and allocation of a given capacity among the dedicated servers where value is generated from a customer only when sojourn time does not exceed the service level agreement of its class. Yolken and Bambos [46] assume the system operator has fixed capacity to be split among N users. User i dedicates the capacity allocated for serving its own demand. Users delay sensitivity is unknown to the system operator and allocation of capacity is done through a bidding mechanism.

Some works combine dedicated and non-dedicated servers. For example, Chao, Liu, and Zheng [10] describe a capacity allocation model motivated by health-care management. There are N service stations, with a dedicated *nonswitchable* rate of demand for every station, and also a stream of non-dedicated *switchable* customers. Mandjes and Timmer [33] consider competition between two Internet providers. Each competitor has fixed capacity which it may split among several subnetworks which differ in their capacity and price. Customers choose one of the offered subnetworks or balk. The authors examine the equilibrium outcomes of this game. Shumsky and Zhang [40] similarly 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.

An interesting empirical result appears in the recent paper of Lu, Musalem, Olivares and Schilkrut [32]. They conducted an experiment in a physical queue and found that customers' decisions of joining the queue are mainly affected by queue length without any adjustment for service speed. Since pooling multiple queues into a single queue increases queue length, this leads to higher loss probability and lower revenues. This empirical finding conforms with our theoretical conclusions.

The asymptotic results in this paper are related to large scale service systems, introduced by Atar [7]. A collection of large scale queue systems is called a *heavy traffic regime* if $\rho^{(n)} \to 1$ under a second order condition (see [7] and references therein). In the *conventional regime* (see Chen and Yao [11]), the number of servers does not change. In contrast, the regime introduced by Halfin and Whitt [17] considers systems with a large number of servers, while the individual service rates do not change ([17]). Atar [7] introduces the *α-parametrization* and in describing it discusses a single-class queueing model M/M/N with parameters λ^n, μ^n, and N^n, depending on some scaling parameter n. Let the external arrival rate increase as $\Theta(n)$. Given some $\alpha \in [0,1]$, assume the number of servers is $\Theta(n^\alpha)$, while each individual service rate scales as $n^{1-\alpha}$. In addition, let a suitable critical load condition hold and 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 $\alpha = \frac{2}{3}$.

The rest of the paper is organized as follows. The main results along with with their theoretical development, appear in §2, §3 and §4. §5 provides some directions for future research. Appendix A contains the proofs.

2 Non-preemptive policy: identical service rates

In this section, we assume that the service rates must be equal, i.e., $\mu_i = \mu/k$, $i = 1, \dots, k$. 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 $\bar{Z}_k$, and the optimal number of servers by k^*, i.e.,

$$k^* := \text{argmax}\{\ \bar{Z}_k\ :\quad k = 0, 1, 2, ...\ \}. \tag{4}$$

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

$$\boldsymbol{E}(W^{(k)}) = \frac{1}{\mu_i} \equiv \frac{k}{\mu}. \tag{5}$$

Substituting in (2), we obtain that the admission fee is

$$p^{(k)} = R - \frac{Ck}{\mu}. \tag{6}$$

Substituting in (3), the profit of the system's owner is

$$\bar{Z}_k = \lambda p^{(k)}(1-\pi_k) = \lambda\Big(R - \frac{Ck}{\mu}\Big)(1-\pi_k) = C\rho(\nu_e - k)(1-\pi_k). \tag{7}$$

For identical servers, π_k is given by the well-known Erlang's loss formula: (see, e.g., [27] page 105):

$$\pi_k = \bar{\pi}_k \equiv \frac{(k\rho)^k/k!}{\sum\limits_{l=0}^{k}(k\rho)^l/l!}. \tag{8}$$

(The bar symbol emphasizes the fact that the servers are identical.) Substituting $\bar{\pi}_k$ in (7) yields

$$\bar{Z}_k = C\rho(\nu_e - k)(1-\bar{\pi}_k) = C\rho(\nu_e - k)\left(1 - \frac{(k\rho)^k/k!}{\sum\limits_{l=0}^{k}(k\rho)^l/l!}\right). \tag{9}$$

It is immediate from (6) that as k increases, the admission fee $p^{(k)}$ decreases and, by Lemma 2.1 below, the admission probability, $1-\bar{\pi}_k$, increases. Namely, there are opposing influences on the profit $\bar{Z}_k$. Hence, given ν_e and ρ, 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 $\bar{\pi}_k$ that all servers are busy becomes smaller; recall that all proofs are given in Appendix A.

Lemma 2.1 *$\{\bar{\pi}_k\}_{k=0}^{\infty}$ is a strictly decreasing sequence, $\lim\limits_{k\to\infty}\bar{\pi}_k = \frac{\rho-1}{\rho}$ if $\rho > 1$, and $\lim\limits_{k\to\infty}\bar{\pi}_k = 0$ if $\rho \le 1$. In particular, for $\rho = 1$, $\bar{\pi}_k = \Theta(k^{-1/2})$.*

The lemma can also be understood along the following reasoning. Let j_k be the number of servers with total rate λ i.e., $j_k = \min\limits_j\{j | j(\frac{\mu}{k}) > \lambda\}$ or, $j_k = \lceil k\rho \rceil$. Assuming $\lambda < \mu$, is for every $j \ge j_k$, the birth and death chain representing the number of customers in the queue has a drift left (i.e., towards j_k). The number of states with left drift is $\lfloor k(1-\rho)\rfloor$ and it increases to infinity, as $k \to \infty$. As a result $\lim\limits_{k\to\infty}\bar{\pi}_k = 0$.

In addition, $\bar{\pi}_k$ can be interpreted as the probability of abandonment in a system with k servers and infinite queue, serving extremely impatient customers. It is known (e.g., (2.23) in [44]), that in such systems (with exponential time to abandonment) the limit for the probability of abandonment is $(\rho-1)/\rho$. Similarly, with the many server queues with abandonment can also be used to establish that when $\rho = 1$ the probability of abandonment i.e., $\bar{\pi_k}$ grows as a square root (See e.g., Theorem 4 of [14]).

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 2.2

$$f(k) := k + \frac{1 - \bar{\pi}_{k-1}}{\bar{\pi}_{k-1} - \bar{\pi}_k}, \quad k = 1, 2, ...$$

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

Lemma 2.3 *For any* $\rho > 0$, $\{f(k)\}_{k=0}^{\infty}$ *is a strictly increasing sequence, and* $\lim_{k\to\infty} f(k) = \infty$.

Theorem 2.4 *Given* $\nu_e, \rho > 0$, *an optimal number of identical servers,* k^*, *satisfies:*

$$f(k^*) \le \nu_e < f(k^* + 1). \tag{10}$$

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 nth system, $n = 1, 2, ...$, the total arrival and service rates are

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

$$\text{and} \quad \lambda = \mu.$$

Therefore,

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

Denote the corresponding optimal number of servers by $k^{(n)}$. By Theorem 2.4 $k^{(n)}$ satisfies the conditions

$$f(k^{(n)}) \le \nu_e^{(n)} < f(k^{(n)} + 1). \tag{11}$$

We further investigate the case $\rho = 1$. For this purpose, we look at the sequence $\{\bar{\pi}_k \cdot k^{\frac{1}{2}}\}_{k=1}^{\infty}$. We will show that it converges (to a finite limit.) First of all, it is bounded, since the inequality (30) (from the proof of Lemma 2.1) is equivalent to

$$\frac{1}{2.54} \le \bar{\pi}_k \cdot k^{\frac{1}{2}} \le 4. \tag{12}$$

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

$$\frac{\bar{\pi}_{k+1} \cdot (k+1)^{\frac{1}{2}}}{\bar{\pi}_k \cdot k^{\frac{1}{2}}} = \frac{\bar{\pi}_{k+1}}{\bar{\pi}_k} \cdot \left(1 + \frac{1}{k}\right)^{\frac{1}{2}}.$$

The right multiplicand converges to 1. As for the left multiplicand, we use the following Lemma, whose proof relies on a result by Harel [18]:

Lemma 2.5 *When* $\rho = 1$, *the ratio sequence* $\{\frac{\bar{\pi}_{k+1}}{\bar{\pi}_k}\}_{k=1}^{\infty}$ *is increasing, and* $\lim_{k\to\infty} \frac{\bar{\pi}_{k+1}}{\bar{\pi}_k} \le 1$.

Lemma 2.5 implies that $\{\frac{\bar{\pi}_{k+1}}{\bar{\pi}_k}\}_{k=1}^{\infty}$ converges, hence $\left\{\frac{\bar{\pi}_{k+1}\cdot(k+1)^{\frac{1}{2}}}{\bar{\pi}_k\cdot k^{\frac{1}{2}}}\right\}_{k=1}^{\infty}$ converges as well. This fact and the bounds (12) imply that the sequence $\{\bar{\pi}_k \cdot k^{\frac{1}{2}}\}_{k=1}^{\infty}$ converges; denote this limit by b. Then $\bar{\pi}_k \approx b \cdot k^{-\frac{1}{2}}$, that is, $\bar{\pi}_k = b \cdot k^{-\frac{1}{2}} + o(k^{-\frac{1}{2}})$. Hence,

$$\begin{aligned} f(k) \approx k + \frac{1 - \frac{b}{\sqrt{k-1}}}{\frac{b}{\sqrt{k-1}} - \frac{b}{\sqrt{k}}} &= k + \frac{\frac{1}{b}\sqrt{k^2-k} - \sqrt{k}}{\sqrt{k} - \sqrt{k-1}} \cdot \frac{\sqrt{k}+\sqrt{k-1}}{\sqrt{k}+\sqrt{k-1}} \\ &= k + (\frac{1}{b}\sqrt{k^2-k} - \sqrt{k})(\sqrt{k}+\sqrt{k-1}) = \Theta(k^{3/2}). \end{aligned}$$

Therefore, $f(k^{(n)}), f(k^{(n)}+1) = \Theta(k^{(n)^{3/2}})$, so the bounds (11) yield $\nu_e^{(n)} = \Theta(k^{(n)^{3/2}})$. But $\nu_e^{(n)} = \Theta(n)$, hence we obtain $n = \Theta(k^{(n)^{3/2}}) = \Theta(\mu^{(n)})$. Thus, we have established:

Theorem 2.6 *When $\rho = 1$, the optimal number of servers $k^{(n)}$ grows as $n^{2/3}$ and individual service rate grows as $n^{1/3}$.*

3 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 kth server is the slowest, $\boldsymbol{E}(W^{(k)}) = \frac{1}{\mu_k}$, so the optimal strategy of the system's owner is to set the admission fee by $p^{(k)} = R - \frac{C}{\mu_k}$. (see (2)). 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 π_k; Erlang's loss formula is no longer valid, since the servers are nonidentical, i.e., $\pi_k = \pi_k(\mu_1, ..., \mu_k)$. Substituting in (3), we obtain the following expression of the profit of the system's owner:

$$Z^{(k)}(\mu_1, ..., \mu_k) \equiv C\rho\Big(\nu_e - \frac{\mu}{\mu_k}\Big)(1 - \pi_k(\mu_1, ..., \mu_k)). \tag{13}$$

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

3.1 Normalizing parametrization

In the rest of §3, it will be convenient to use the parametrization obtained by normalizing the rates: Let

$$d_1 = \frac{\mu_1}{\mu}\ , \ ...\ ,\ d_k = \frac{\mu_k}{\mu} \quad (d_1 + ... + d_k = 1, \quad d_1 \geq d_2 \geq \cdots \geq d_k > 0.) \tag{14}$$

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

Adjusting (13) to this parametrization yields:

$$Z^{(k)}(d_2, ..., d_k) = C\rho\Big(\nu_e - \frac{1}{d_k}\Big)(1 - \pi_k(d_2, ..., d_k)), \tag{15}$$

where d_1 is not an argument of both $Z^{(k)}$ and π_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 $\frac{1}{\nu_e} < \frac{1}{k}$), and denote the relevant domain of $Z^{(k)}(\cdot)$ as follows:

Definition 3.1 *Let* $I^{(k)} \equiv \left(\frac{1}{\nu_e}, \frac{1}{k}\right]$ *denote the* effective domain *of* $Z^{(k)}(\cdot)$.

3.2 Two servers

Here, the service rates of the two servers are μ_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 (14)) 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 (15), we obtain:

$$Z(d_2) \equiv Z^{(2)}(d_2) = C\rho\left(\nu_e - \frac{1}{d_2}\right)(1 - \pi_2(d_2)). \tag{16}$$

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

$$\pi\left(\frac{1}{2}\right) = \bar{\pi}_2, \quad Z\left(\frac{1}{2}\right) = \bar{Z}_2.$$

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).

Theorem 3.2 *For all* $\rho, \nu_e > 0$,

$$Z(d_2) < \bar{Z}_2 \textit{ for all } 0 < d_2 < \frac{1}{2} \quad \Longleftrightarrow \quad \nu_e < g(\rho).$$

where $g(\rho) := 8\rho^2 + 16\rho + 18 + \frac{8}{\rho} + \frac{1}{\rho^2}$.

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

Corollary 3.3 *For all $\rho, \nu_e > 0$,*

$$\nu_e \leq f(3) \quad \Longrightarrow \quad Z(d_2) < \bar{Z}_2 \textit{ for all } 0 < d_2 < \frac{1}{2}.$$

Although this result is weaker than Theorem 3.2, we state it because we conjecture that it holds for an arbitrary number of servers as well; see Conjecture 3.5 below.

We conclude our analysis by noting that for any ν_e, the condition of Theorem 3.2 holds for ρ sufficiently large. Therefore:

Theorem 3.4 *For any $\nu_e > 0$, $\lim_{\rho \to \infty} \operatorname{argmax}_{d_2 \in (\frac{1}{\nu_e}, \frac{1}{2}]} Z(d_2) = \frac{1}{2}$.*

As we have seen so far, the various computations for the relatively simple case of $k = 2$ servers are fairly involved. Now, 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 3.3, and provides a sufficient condition for the superiority of identical servers *when their number is fixed*. We use the normalizing parametrization (14) and its corresponding profit expression (15); recall that $(\mu_1, ..., \mu_k) \neq (\frac{1}{k}, ..., \frac{1}{k})$ if and only if $d_k < \frac{1}{k}$.

Conjecture 3.5 *For any fixed number k of servers and $\rho > 0$, in a non-preemptive policy,*

$$\nu_e \leq f(k+1) \quad \Longrightarrow \quad Z^{(k)}(d_2, ..., d_k) < \bar{Z}_k \textit{ for all } 0 < d_k < \frac{1}{k}.$$

In addition to its correctness for two servers (Corollary 3.3), further support for its correctness stems from numerical evidence for the case of $k = 3$ servers. Specifically, by (15),

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

where $\pi_3(d_2, d_3)$ is the probability that all three servers are busy. Using numerical computations, we can see that Conjecture 3.5 holds in this case, namely, for all $\rho > 0$,

$$\nu_e \leq f(4) \quad \Longrightarrow \quad Z^{(3)}(d_2, d_3) < \bar{Z}_3 \text{ for all } 0 < d_3 < \frac{1}{3}. \tag{17}$$

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 3.6 *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 2.4.*

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 ν_e:

- Case 1. $\nu_e \leq f(3)$: By Corollary 3.3 , for any $0 < d_2 < \frac{1}{2}$,

  $$Z^{(2)}(d_2) < Z^{(2)}(\frac{1}{2}) = \bar{Z}_2.$$

  Similarly, since $\nu_e \leq f(4)$, (17) implies that for any $0 < d_3 < \frac{1}{3}$,

  $$Z^{(3)}(d_3) < Z^{(3)}(\frac{1}{3}) = \bar{Z}_3.$$

  Let k^* be the optimal number of identical servers corresponding to ρ, ν_e (see (4)). By Theorem 2.4, $k^* \leq 3$ (as $\nu_e \leq f(4)$). The claim follows.

- Case 2. $f(3) < \nu_e \leq f(4)$: By Theorem 2.4, $k^* = 3$. First of all, as in Case 1, for any $0 < d_3 < \frac{1}{3}$,

  $$Z^{(3)}(d_3) < Z^{(3)}(\frac{1}{3}) = \bar{Z}_3.$$

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

  $$Z^{(2)}(d_2) < \bar{Z}_3. \tag{18}$$

  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 $\rho > \rho_0 \Leftrightarrow f(4) < g(\rho)$. Thus, if $\rho > \rho_0$, then $\nu_e < g(\rho)$, so Theorem 3.2 implies that for any $0 < d_2 < \frac{1}{2}$,

  $$Z^{(2)}(d_2) < Z^{(2)}(\frac{1}{2}) = \bar{Z}_2 < \bar{Z}_3,$$

  hence the claim follows.

  So assume that $\rho < \rho_0$. Then altogether we have $f(4) > \nu_e > g(\rho) > f(3) > 3$. Substituting (7),(16), we obtain that (18) is equivalent to:

  $$\left(\nu_e - \frac{1}{d_2}\right)(1 - \pi_2(d_2)) < (\nu_e - 3)(1 - \bar{\pi}_3). \tag{19}$$

  Note that this inequality involves ν_e, d_2, and ρ 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 A.5(ii), then this inequality holds for the two *extreme possible values* of ν_e, i.e., $\nu_e \in \{g(\rho), f(4)\}$; specifically, for all $\frac{1}{\nu_e} < \rho < \rho_0$:

  $$\begin{aligned} \left(g(\rho) - \frac{1}{\sqrt{\rho^2 + \rho} - \rho}\right)(1 - \pi_2(\sqrt{\rho^2 + \rho} - \rho)) &< (g(\rho) - 3)(1 - \bar{\pi}_3), \\ \left(f(4) - \frac{1}{\sqrt{\rho^2 + \rho} - \rho}\right)(1 - \pi_2(\sqrt{\rho^2 + \rho} - \rho)) &< (f(4) - 3)(1 - \bar{\pi}_3). \end{aligned}$$

4 Loss systems with preemptive policy

In this section we consider an M/M/k/k loss system with *preemptive* routing policy. Note that the results in this section assume that the number of servers, k, is given and fixed.

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 4.1 *$W^{(k)}$ is a hyper-exponential random variable with density*

$$f_{W^{(k)}}(x) = \sum_{l=1}^{k} u_l \left(\sum_{i=1}^{l} \mu_i \right) e^{-(\sum_{i=1}^{l} \mu_i)x}, \tag{20}$$

where $u_1, \ldots, u_k$ are parameters such that $\sum\limits_{l=1}^{k} u_l = 1$.

The expected service time, $\boldsymbol{E}(W^{(k)})$, 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 (5) in §2:

- Case 1. All servers have the same rate, i.e., $\mu_i = \frac{\mu}{k}$ for all $i = 1, \ldots, k$. 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 μ and all other servers operate with zero rate, i.e., they function as *standby positions*. Here, a customer who joins the kth 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. This property is expressed by the following conservation law:

Corollary 4.2 *For any allocation $\mu_1 \geq \cdots \geq \mu_k$ ($\sum_{i=1}^{k} = \mu$),*

$$\boldsymbol{E}(W^{(k)}) = \frac{k}{\mu_1 + \cdots + \mu_k} = \frac{k}{\mu}. \tag{21}$$

Using this corollary, we can substitute (21) in (2) and obtain that for any allocation $\mu_1, \ldots, \mu_k$, the optimal entry fee is

$$p^{(k)} = R - \frac{Ck}{\mu}. \tag{22}$$

Note that this is the same as in the identical-server case studied in §2; see (6). Consequently, the corresponding profit of the system's owner is:

$$\widetilde{Z}^{(k)}(\mu_1, \ldots, \mu_k) = \lambda \left(R - \frac{Ck}{\mu} \right) (1 - \widetilde{\pi}_k(\mu_1, \ldots, \mu_k)), \tag{23}$$

where $\widetilde{\pi}_k(\mu_1, \ldots, \mu_k)$ 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 μ while all the other $k-1$ servers are standby positions, and the equal-allocation is the worst. Formally:

Theorem 4.3 *For any $\mu_1 \geq \cdots \geq \mu_k$ with $\sum_{i=1}^k \mu_i = \mu$,*

$$\widetilde{Z}^{(k)}(\mu, 0, 0, \ldots, 0) \geq \widetilde{Z}^{(k)}(\mu_1, \ldots, \mu_k) \geq \bar{Z}_k,$$

where $\bar{Z}_k$ is the profit of a system with equal service rates as in §2; see (9).

We provide the proof of Theorem 4.3 below (rather than in Appendix A) 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, $\widetilde{\pi}_k(\mu/k, \ldots, \mu/k) = \bar{\pi}_k$ (where $\bar{\pi}_k$ is the probability corresponding to equal service rates in the *nonpreemptive* policy; see (8).) It then follows from (23) that in order to prove the stated inequalities, it suffices to show that

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

or equivalently,

$$\widetilde{\pi}_k(\mu, 0, \ldots, 0) \leq \widetilde{\pi}_k(\mu_1, \ldots, \mu_k) \leq \widetilde{\pi}_k(\mu/k, \ldots, \mu/k). \tag{24}$$

Our method for establishing (24) 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 λ. As for the death rates, let $S_l \sim \exp(\mu_l)$ denote the service time of a customer at the lth 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, $\widetilde{\pi}_k(\mu_1, \ldots, \mu_k)$ 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 4.4 *For the above birth-death process, the inequalities* (24) *hold for any $\mu_1 \geq \cdots \geq \mu_k$ with $\sum_{i=1}^k \mu_i = \mu$.*

See Appendix A 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 ith and jth servers. Suppose that $\mu_j > 0$. Then by transferring

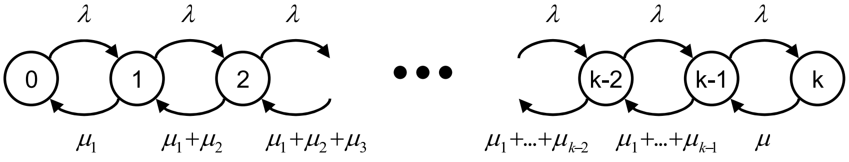

Figure 1: State diagram for our birth-death process.

some service rate from the jth server to the ith 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 μ_1 is as large as possible, and is minimized when μ_k is as large as possible. The allocation $(\mu_1, \ldots, \mu_k) = (\mu, 0, \ldots, 0)$ clearly maximizes μ_1; the allocation $(\mu_1, \ldots, \mu_k) = (\mu/k, \ldots, \mu/k)$ maximizes μ_k (due to the condition $\mu_1 \geq \cdots \geq \mu_k$). This establishes Lemma 4.4 and hence completes the proof of Theorem 4.3.

Thus, $\widetilde{Z}^{(k)}(\mu, 0, 0, \ldots, 0)$ is the maximum profit of the system's owner, while $\bar{Z}_k$ is the minimum. In other words, it is best to allocate the total service capacity μ to a single server, and it is worst to allocate it equally among all servers.

We conclude this section by analyzing the optimization problem when considering their number as well. That is, we would like to maximize the profit $\widetilde{Z}^{(k)}(\mu_1, \ldots, \mu_k)$ given in (23), where k is also a decision variable. By Theorem 4.3 and (23), this problem reduces to finding the optimal number k^* of servers that maximizes:

$$\widetilde{Z}^{(k)}(\mu, 0, 0, \ldots, 0) \equiv \lambda \left(R - \frac{Ck}{\mu} \right) (1 - \widetilde{\pi}_k(\mu, 0, 0, \ldots, 0)). \tag{25}$$

Interestingly enough, this k^* is equal to the buffer size in Naor's model [34]; 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 μ) 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 [34]. Thus, we have established:

Theorem 4.5 *In an M/M/k/k loss system with preemptive policy,*

$$k^* = n_r,$$

where n_r is the profit maximizing buffer size in Naor's model [34].

5 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 for the two policies are opposite. 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, actually serving as standby positions. Below we present discussion topics and some possible directions for future research.

1. In §3, we establish a necessary and sufficient condition for the superiority of *two identical servers* in the nonpreemptive model when there are two available servers. Conjecture 3.5 suggests a sufficient condition for the superiority of identical servers when the number of servers is fixed. Conjecture 3.6 suggests that when the number of servers is not fixed it is always optimal to use identical servers. The corresponding birth-death equations seem fairly complicated. Thus, a challenging direction would be to try to establish their correctness.

   While our conjecture is not proved, we provide support for its validity. In particular, we follow the conjecture by arguments showing that for $\rho > 0.2$, $\nu_e \leq f(4)$ and $k \leq 3$, the optimal solution is with identical servers. As we can see, $f(4)$ is rather large (for example, if $\rho = 1$ then $f(4) = 22.43$), so that in many applications, the condition $\nu_e \leq f(4)$ is satisfied.

2. Suppose the policy is preemptive, as in §4. 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 if there is an available server. Of course, the optimal strategy of the system administrator is to employ faster servers at any time the. As such, customers can compute their expected service time given there is an available server. This may mean a probabilistic equilibrium joining strategy in which some customers give up service without trying – as in [13]. 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 considers this while computing the profit maximizing price. As in §4, the optimal solution will be a single active server with standby positions.

3. Note that if servers have identical capacities, or 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 surplus will be 0. In this case the firm's strategy is socially optimal, as in [13]. However, when servers have different capacities which are observable to customers (as in §4), 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 the 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 customers (if any) at servers j with $j > i$ are simultaneously transferred one server down. Also recall that the expected service time, $\boldsymbol{E}(W^{(k)})$, is the same (identically μ/k) for any capacity allocation (see Corollary 4.2).

   An interesting problem is to design a different, non-shifting, mechanism for transferring customers between servers. For example, whenever a customer completes service at any server i which is not the slowest (among the busy servers), we could transfer the customer currently being served by the slowest server to i. By doing so we may be able to decrease $\boldsymbol{E}(W^{(k)})$, and consequently increase the admission fee and hence profits (due to (2)). Note also that this will decrease the gap between the monopolistic and social objectives, and the profit maximizing policy gets closer to optimizing social welfare. Perhaps a more sophisticated mechanism will achieve equal expected service times, namely, for all $i = 0, \ldots, k$, $\boldsymbol{E}(W^{(i)})$ 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 budgeting for servers, incurring switching costs, and, of course, allowing for waiting in a queue when all servers are busy.

A Appendix

Proof of Lemma 2.1 Define $a_k := \bar{\pi}_k^{-1}$, hence

$$\begin{aligned} a_k &\equiv \frac{1}{(k\rho)^k/k!}\sum_{l=0}^{k}(k\rho)^l/l! \\ &= 1+\frac{1}{\rho}+\left(1-\frac{1}{k}\right)\frac{1}{\rho^2}+\left(1-\frac{1}{k}\right)\left(1-\frac{2}{k}\right)\frac{1}{\rho^3}+\cdots+\left(1-\frac{1}{k}\right)\cdots\left(1-\frac{k-1}{k}\right)\frac{1}{\rho^k}. \end{aligned}$$

When k increases, each factor $\left(1-\frac{l}{k}\right)$ of a_k increases. It follows that $\{a_k\}_{k=0}^{\infty}$ is a strictly increasing sequence, and so $\{\bar{\pi}_k\}_{k=0}^{\infty}$ is a strictly decreasing sequence.

Suppose $\rho > 1$. Then $a_k < \sum_{l=0}^{k}\left(\frac{1}{\rho}\right)^l \to \frac{1}{1-\frac{1}{\rho}} = \frac{\rho}{\rho-1}$. Thus, the sequence $\{a_k\}_{k=0}^{\infty}$ is bounded and increasing, so it converges to a finite limit, denoted α. Of course, $\alpha \le \frac{\rho}{\rho-1}$. Now, for every $\epsilon > 0$ and for every N, there exists k such that $\sum_{l=0}^{N}\left(\frac{1}{\rho}\right)^l - \epsilon \le a_k \le \alpha \le \frac{\rho}{\rho-1}$. Since α is constant, letting $\epsilon \to 0$ and $k \to \infty$ gives $\alpha = \frac{\rho}{\rho-1}$. Consequently, $\bar{\pi}_k \to \frac{\rho-1}{\rho}$.

Suppose $\rho \le 1$. 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 $a_k \to \infty$ and hence $\bar{\pi}_k \to 0$.

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

$$a_{s^2} \le s\left[\sum_{t=0}^{s-1}\prod_{i=1}^{ts+1}\left(1-\frac{i}{s^2}\right)\right], \tag{26}$$

and

$$s\left[\sum_{t=1}^{s}\prod_{i=1}^{ts}\left(1-\frac{i}{s^2}\right)\right] \le a_{s^2}. \tag{27}$$

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

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

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

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

Using this inequality in (26) yields:

$$a_{s^2} \le s\sum_{t=0}^{s-1}e^{-\frac{1}{2}t^2} = s\sum_{t=0}^{s-1}(e^{-\frac{1}{2}})^{t^2} \le s\sum_{t=0}^{\infty}(e^{-\frac{1}{2}})^{t} = \frac{s}{1-e^{-1/2}}. \tag{28}$$

Now, (27) yields:

$$\begin{aligned} a_{s^2} &\ge s\left[\sum_{t=1}^{s}\prod_{i=1}^{ts}\left(1-\frac{i}{s^2}\right)\right] = s\left[\prod_{i=1}^{s}\left(1-\frac{i}{s^2}\right)+\sum_{t=2}^{s}\prod_{i=1}^{ts}\left(1-\frac{i}{s^2}\right)\right] \\ &\ge s\left[\prod_{i=1}^{s}\left(1-\frac{i}{s^2}\right)\right] \ge s\left[\prod_{i=1}^{s}\left(1-\frac{s}{s^2}\right)\right] = s\left(1-\frac{s}{s^2}\right)^s = s\left(1-\frac{1}{s}\right)^s. \end{aligned} \tag{29}$$

Combining (28),(29), we obtain that for all $s > 2$,

$$0.25 \leq \left(1 - \frac{1}{s}\right)^s \leq \frac{a_{s^2}}{s} \leq \frac{1}{1 - e^{-1/2}} = 2.54. \tag{30}$$

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

Proof of Lemma 2.3 First, we present the following result which was proved by Harel [18]: For any fixed $\rho > 0$, $\bar{\pi}_k$ is strictly convex as a function of k, that is, $\{\bar{\pi}_k - \bar{\pi}_{k+1}\}_{k=0}^{\infty}$ is monotone decreasing (Theorem 6 in [18]). Recall that $f(k) = k + \frac{1-\bar{\pi}_{k-1}}{\bar{\pi}_{k-1}-\bar{\pi}_k}$. In order to establish the monotonicity, consider the second term, $\frac{1-\bar{\pi}_{k-1}}{\bar{\pi}_{k-1}-\bar{\pi}_k}$. The numerator is increasing (in k) since by Lemma 2.1, $\{\bar{\pi}_k\}_{k=0}^{\infty}$ is decreasing. The denominator is increasing since by Theorem 6 [18], $\{\bar{\pi}_k - \bar{\pi}_{k+1}\}_{k=0}^{\infty}$ 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 ∞, and the second term is nonnegative. As we just proved, $\{f(k)\}_{k=1}^{\infty}$ is a strictly increasing sequence, hence $\lim_{k\to\infty} f(k) = \infty$.

□

Proof of Theorem 2.4 We first show that k^* is a solution to $f(k) \leq \nu_e \leq f(k+1)$. Recall that $\bar{Z}_k = C\rho(\nu_e - k)(1 - \bar{\pi}_k)$ (see (9)) and that $f(k) = k + \frac{1-\bar{\pi}_{k-1}}{\bar{\pi}_{k-1}-\bar{\pi}_k}$ for $k > 0$ and $f(0) = 0$ (see Definition 2.2).

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:

$$\begin{aligned}
Z(k^*) \geq Z(k^* - 1) \quad &\Longleftrightarrow \quad (\nu_e - k^*)(1 - \bar{\pi}_{k^*}) \geq (\nu_e - k^* + 1)(1 - \bar{\pi}_{k^*-1}) \\
&\Longleftrightarrow \quad (\bar{\pi}_{k^*-1} - \bar{\pi}_{k^*})(\nu_e - k^*) \geq 1 - \bar{\pi}_{k^*-1} \\
&\Longleftrightarrow \quad \nu_e \geq k + \frac{1 - \bar{\pi}_{k^*-1}}{\bar{\pi}_{k^*-1} - \bar{\pi}_{k^*}} = f(k^*),
\end{aligned}$$

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

$$Z(k^* + 1) \leq Z(k^*) \Longleftrightarrow \nu_e \leq f(k^* + 1).$$

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

We note that if $\nu_e = f(l)$ for some integer l, there are two solutions to $f(k) \leq \nu_e \leq f(k+1)$, otherwise it admits a *unique* solution. However, in the (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 that k^* is the (unique) solution to (10).

□

Proof of Lemma 2.5 Theorem 1 in Harel [18] states that the function

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

is strictly concave in k ($k = 1, 2, ...$). Define $a_k := \bar{\pi}_k^{-1}$, that is (recall that $\rho = 1$):

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

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,

$$2a_k > a_{k-1} + a_{k+1}. \tag{31}$$

We will show that the ratio sequence $\{\frac{a_{k+1}}{a_k}\}_{k=1}^{\infty}$ is decreasing. The inequality (31) 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, $\{\frac{a_{k+1}}{a_k}\}_{k=1}^{\infty}$ is decreasing, so $\{\frac{\bar{\pi}_{k+1}}{\bar{\pi}_k}\}_{k=1}^{\infty}$ is increasing.

It remains to establish convergence. The sequence $\{\bar{\pi}_k\}_{k=1}^{\infty}$ is decreasing by Lemma 2.1, so for any k, $\frac{\bar{\pi}_{k+1}}{\bar{\pi}_k} < 1$. Thus, $\{\frac{\bar{\pi}_{k+1}}{\bar{\pi}_k}\}_{k=1}^{\infty}$ is increasing and bounded from above by 1, so $\lim_{k\to\infty} \frac{\bar{\pi}_{k+1}}{\bar{\pi}_k} \leq 1$. □

Proof of Theorem 3.2 We present Lemmas A.1-A.8 (their proofs are given below), in order to prove the Theorem. We begin by deriving the exact expression for $\pi_2(d_2)$ and $Z(d_2)$.

Lemma A.1

$$\pi_2(d_2) = \frac{2\rho^2}{2\rho^2 + 2\rho + (4\rho + 2)\frac{d_2 - d_2^2}{\rho + d_2}}. \tag{32}$$

Substituting in (16) 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:

$$Z(d_2) = Z_1(d_2)(Z_2(Z_3(d_2))), \tag{33}$$

where

$$Z_1(x) = C\rho\left(\nu_e - \frac{1}{x}\right), \quad Z_2(y) = 1 - \frac{2\rho^2}{2\rho^2 + 2\rho + y}, \quad Z_3(z) = (4\rho + 2)\frac{z - z^2}{\rho + z}. \tag{34}$$

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

$$Z_1'(x) = \frac{C\rho}{x^2}, \quad Z_2'(y) = \frac{2\rho^2}{(2\rho^2+2\rho+y)^2}, \quad Z_3'(z) = -(4\rho+2)\frac{(z^2+2z\rho-\rho)}{(\rho+z)^2}. \tag{35}$$

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 3.1.

Lemma A.2 *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 A.3 *For all $\rho, \nu_e > 0$,*

$$0 < d_2 \leq \frac{\rho}{2\rho+1} \quad \Longrightarrow \quad Z(d_2) < Z(\frac{1}{2}) = \bar{Z}_2. \tag{36}$$

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 A.4 *Let $Z_{23} := 1 - \pi_2 \equiv Z_2 \circ Z_3$, that is,*

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

Lemma A.5 *(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 A.6 *$Z(\cdot)$ 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 A.7 *For all $\rho, \nu_e > 0$,*

$$Z(d_2) < \bar{Z}_2 \text{ for all } 0 < d_2 < \frac{1}{2} \quad \Longleftrightarrow \quad Z'(\frac{1}{2}) > 0.$$

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

Lemma A.8 *For all $\rho, \nu_e > 0$,*

$$Z'(\frac{1}{2}) > 0 \quad \iff \quad \nu_e < g(\rho),$$

where $g(\rho) := 8\rho^2 + 16\rho + 18 + \frac{8}{\rho} + \frac{1}{\rho^2}$.

Combining Lemmas A.7 and A.8 complete the proof.

□

Proof of Lemma A.1 First, we denote by $\pi_0(\mu_1, \mu_2), \pi_1^1(\mu_1, \mu_2), \pi_1^2(\mu_1, \mu_2)$ and $\pi_2(\mu_1, \mu_2)$, 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 $\pi_i = \pi_i(\mu_1, \mu_2)$)

$$-\lambda\pi_0 + \mu_1\pi_1^1 + (\mu - \mu_1)\pi_1^2 = 0,$$

$$-(\lambda + \mu_1)\pi_1^1 + \lambda\pi_0 + (\mu - \mu_1)\pi_2 = 0,$$

$$-(\lambda + (\mu - \mu_1))\pi_1^2 + \mu_1\pi_2 = 0,$$

$$-\mu\pi_2 + \lambda\pi_1^1 + \lambda\pi_1^2 = 0,$$

$$\pi_0 + \pi_1^1 + \pi_1^2 + \pi_2 = 1.$$

Equivalently,

$$-\rho\pi_0 + (1 - d_2)\pi_1^1 + d_2\pi_1^2 = 0,$$

$$-(\rho + 1 - d_2)\pi_1^1 + \rho\pi_0 + d_2\pi_2 = 0,$$

$$-(\rho + d_2)\pi_1^2 + (1 - d_2)\pi_2 = 0,$$

$$-\pi_2 + \rho\pi_1^1 + \rho\pi_1^2 = 0,$$

$$\pi_0 + \pi_1^1 + \pi_1^2 + \pi_2 = 1.$$

The solution of the system for $\pi_2(\mu_1, \mu_2) = \pi_2(d_2)$ is $\pi_2(d_2) := \frac{2\rho^2}{2\rho^2+2\rho+(d_2-d_2^2)\frac{4\rho+2}{\rho+d_2}}$. □

Proof of Lemma A.2 Parts (i),(ii) follow immediately from the derivatives (35). 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 $\left(\frac{1}{\nu_e}, \sqrt{\rho^2 + \rho} - \rho\right]$. The claim now follows from the decomposition (33) of $Z(\cdot)$ and the fact that multiplication and composition of positive functions preserve monotonicity. □

Proof of Corollary A.3 Fix $0 < d_2 \leq \frac{\rho}{2\rho+1}$. First of all, Lemma A.2(iii) implies that $Z(d_2) \leq Z(\frac{\rho}{2\rho+1})$ (it is easily verified that $\frac{\rho}{2\rho+1} < \sqrt{\rho^2 + \rho} - \rho$). To complete the proof, we will show that $Z(\frac{\rho}{2\rho+1}) < Z(\frac{1}{2})$. It is easily verified that $Z_3(\frac{\rho}{2\rho+1}) = Z_3(\frac{1}{2}) = 1$, and hence $(Z_2 \circ Z_3)(\frac{\rho}{2\rho+1}) = (Z_2 \circ Z_3)(\frac{1}{2})$. However,

$Z_1(\frac{\rho}{2\rho+1}) < Z_1(\frac{1}{2})$ by Lemma A.2(i). Now the inequality immediately follows from the decomposition $Z \equiv Z_1(Z_2 \circ Z_3)$ (see (33)). This completes the proof. □

Proof of Lemma A.5 (i) As for $Z_1(\cdot), Z_2(\cdot)$, it is clear that $Z_1'(x), Z_2'(y)$ (see (35)) 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 A.2(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 A.2(i),(ii), as $d_2 := \sqrt{\rho^2+\rho} - \rho$ maximizes $Z_3(\cdot)$, and $Z_2(\cdot)$ is increasing. □

Proof of Corollary A.6 We will show that $Z'(\cdot)$ is decreasing on $\left[\sqrt{\rho^2+\rho}-\rho, \frac{1}{2}\right]$. Let $x, y \in \left[\sqrt{\rho^2+\rho}-\rho, \frac{1}{2}\right]$ with $x < y$. We claim that the following inequality holds:

$$\begin{aligned} Z'(x) = (Z_1 Z_{23}(x))' &= Z_1(x)Z_{23}'(x) + Z_1'(x)Z_{23}(x) \\ &> Z_1(y)Z_{23}'(y) + Z_1'(y)Z_{23}(y) = Z'(y). \end{aligned}$$

Consider the first term. $Z_{23}(\cdot)$ is concave (by Lemma A.5(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 A.2(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 A.5(i), and $Z_{23}(x) > Z_{23}(y)$ since in $\left[\sqrt{\rho^2+\rho}-\rho, \frac{1}{2}\right]$, $Z_3(\cdot)$ in decreasing and $Z_2(\cdot)$ is increasing (by Lemma A.2(i),(ii)). This completes the proof. □

Proof of Lemma A.7 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 $\epsilon > 0$ sufficiently small, $Z(\frac{1}{2}-\epsilon) \geq Z(\frac{1}{2})$.

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 A.2(iii), $Z(\cdot)$ is increasing on $\left(\frac{1}{\nu_e}, \sqrt{\rho^2+\rho}-\rho\right]$. By Corollary A.6, $Z(\cdot)$ 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 $\sqrt{\rho^2+\rho}-\rho \leq x < \frac{1}{2}$. This completes the proof. □

Proof of Lemma A.8 Applying the product rule to the decomposition $Z = Z_1 Z_{23}$, we obtain:

$$Z'(\frac{1}{2}) = Z_1(\frac{1}{2})Z_{23}'(\frac{1}{2}) + Z_1'(\frac{1}{2})Z_{23}(\frac{1}{2}). \tag{37}$$

We proceed by calculating the four elements in the right hand side (using (34),(35)):

$$\begin{aligned}
Z_1(\frac{1}{2}) &= C\rho(\nu_e - 2), \\
Z_1'(\frac{1}{2}) &= 4C\rho, \\
Z_{23}(\frac{1}{2}) &= Z_2(Z_3(\frac{1}{2})) = Z_2(1) = 1 - \frac{2\rho^2}{2\rho^2 + 2\rho + 1} = \frac{2\rho + 1}{2\rho^2 + 2\rho + 1},
\end{aligned}$$

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:

$$\begin{aligned}
Z_{23}'(\frac{1}{2}) &= (Z_2 \circ Z_3)'(\frac{1}{2}) = Z_2'(Z_3(\frac{1}{2}))Z_3'(\frac{1}{2}) \\
&= Z_2'(1)Z_3'(\frac{1}{2}) = -\frac{4\rho^2}{(2\rho^2 + 2\rho + 1)^2(2\rho + 1)}.
\end{aligned}$$

Substituting in (37), we obtain:

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

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

$$\begin{aligned}
0 < Z'(\frac{1}{2}) &\iff 0 < -4C\rho^3(\nu_e - 2) + 4C\rho(2\rho + 1)^2(2\rho^2 + 2\rho + 1) \\
&\iff 0 < -\rho^2(\nu_e - 2) + (2\rho + 1)^2(2\rho^2 + 2\rho + 1) \\
&\iff \nu_e < 2 + \frac{(2\rho + 1)^2(2\rho^2 + 2\rho + 1)}{\rho^2} \\
&\iff \nu_e < 2 + \frac{8\rho^4 + 16\rho^3 + 16\rho^2 + 8\rho + 1}{\rho^2} \\
&\iff \nu_e < 8\rho^2 + 16\rho + 18 + \frac{8}{\rho} + \frac{1}{\rho^2} \equiv g(\rho). \qquad (38)
\end{aligned}$$

□

Proof of Theorem 4.1 We will establish (20) 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:

$$\begin{aligned}
f_{W^{(2)}}(x) &= \frac{\mu_1}{\mu_2} f_{\exp(\mu_1)}(x) + \left(1 - \frac{\mu_1}{\mu_2}\right) f_{\exp(\mu_1+\mu_2)}(x) \\
&= \frac{\mu_1}{\mu_2}\mu_1 e^{-\mu_1 x} + \left(1 - \frac{\mu_1}{\mu_2}\right)(\mu_1 + \mu_2)e^{-(\mu_1+\mu_2)x}, \qquad (39)
\end{aligned}$$

which is of the form (20) 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 $W^{(2)} = U + I_{\{\frac{\mu_1}{\mu_1+\mu_2}\}}V$ where $I_{\{\frac{\mu_1}{\mu_1+\mu_2}\}}$ is Bernoulli random variable with success probability $\frac{\mu_1}{\mu_1+\mu_2}$, and present the distribution functions:

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

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

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

$$f_U(x-y) = \begin{cases} (\mu_1+\mu_2)e^{-(\mu_1+\mu_2)(x-y)} & x \ge y > 0, \\ 0 & otherwise, \end{cases} \quad \text{and} \quad f_{I_{\{\frac{\mu_1}{\mu_1+\mu_2}\}}V}(y) = \begin{cases} \frac{\mu_1^2}{\mu_1+\mu_2}e^{-\mu_1 y} & y>0, \\ \frac{\mu_2}{\mu_1+\mu_2} & y=0, \\ 0 & y<0, \end{cases}$$

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

$$\begin{aligned} \int_{0^+}^{x} f_U(x-y) f_{I_{\{\frac{\mu_1}{\mu_1+\mu_2}\}}V}(y)dy &= \int_{0^+}^{x} \mu_1^2 e^{-(\mu_1+\mu_2)(x-y)}e^{-\mu_1 y}dy = \mu_1^2 e^{-(\mu_1+\mu_2)x}\int_{0^+}^{x} e^{\mu_2 y}dy \\ &= \frac{\mu_1^2}{\mu_2}e^{-(\mu_1+\mu_2)x}(e^{\mu_2 x}-1) = \frac{\mu_1}{\mu_2}\mu_1 e^{-\mu_1 x} - \frac{\mu_1^2}{\mu_2}e^{-(\mu_1+\mu_2)x}, \end{aligned}$$

and

$$P(I_{\{\frac{\mu_1}{\mu_1+\mu_2}\}}V=0)(\mu_1+\mu_2)e^{-(\mu_1+\mu_2)x} = \frac{\mu_2}{\mu_1+\mu_2}(\mu_1+\mu_2)e^{-(\mu_1+\mu_2)x} = \mu_2 e^{-(\mu_1+\mu_2)x}.$$

Now (20) follows by summing the last two expressions.

Induction step: We assume that (20) 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 $W^{(k)} = U^k + I_{\{\frac{\mu_1+\cdots+\mu_{k-1}}{\mu}\}}W^{(k-1)}$. We compute the density of $W^{(k)}$:

$$\begin{aligned} \int_{0^+}^{x} f_{U^k}(x-y) f_{I_{\{\frac{\mu_1+\cdots+\mu_{k-1}}{\mu}\}}W^{(k-1)}}(y)dy &= \int_{0^+}^{x} \mu e^{-\mu(x-y)} \frac{\mu-\mu_k}{\mu}\sum_{l=1}^{k-1} u_l(\sum_{i=1}^{l}\mu_i)e^{-(\sum_{i=1}^{l}\mu_i)y}dy \\ &= \sum_{l=1}^{k-1} u_l(\mu-\mu_k)(\sum_{i=1}^{l}\mu_i)e^{-\mu x}\int_{0^+}^{x} e^{[\mu-(\sum_{i=1}^{l}\mu_i)]y}dy \\ &= \sum_{l=1}^{k-1} u_l(\mu-\mu_k)\frac{\sum_{i=1}^{l}\mu_i}{\mu-(\sum_{i=1}^{l}\mu_i)}\left(e^{-(\sum_{i=1}^{l}\mu_i)x}-e^{-\mu x}\right). \end{aligned}$$

Since the probability $P(I_{\{\frac{\mu_1+\cdots+\mu_{k-1}}{\mu}\}}W^{(k-1)} = 0) > 0$, we have to add the density of the event $\{U^k = x, I_{\{\frac{\mu_1+\cdots+\mu_{k-1}}{\mu}\}}W^{(k-1)} = 0\}$:

$$P(I_{\{\frac{\mu_1+\cdots+\mu_{k-1}}{\mu}\}}W^{(k-1)} = 0)\mu e^{-\mu x} = \frac{\mu_k}{\mu}\mu e^{-\mu x}.$$

Hence,

$$\begin{aligned} f_{W^{(k)}}(x) &= \sum_{l=1}^{k-1} u_l \left(\frac{(\mu-\mu_k)\sum_{i=1}^{l}\mu_i}{\mu - (\sum_{i=1}^{l}\mu_i)} e^{-(\sum_{i=1}^{l}\mu_i)x} - \frac{(\mu-\mu_k)\sum_{i=1}^{l}\mu_i}{\mu - (\sum_{i=1}^{l}\mu_i)} e^{-\mu x} + \frac{\mu_k}{\mu}\mu e^{-\mu x} \right) \\ &= \sum_{l=1}^{k-1} u_l \left(\frac{\mu-\mu_k}{\mu - (\sum_{i=1}^{l}\mu_i)} \cdot \sum_{i=1}^{l} \mu_i e^{-(\sum_{i=1}^{l}\mu_i)x} + \frac{\mu_k - (\sum_{i=1}^{l}\mu_i)}{\mu - (\sum_{i=1}^{l}\mu_i)} \cdot \mu e^{-\mu x} \right) \\ &= \sum_{l=1}^{k} u_l' f_{exp(\sum_{i=1}^{l}\mu_i)}(x), \end{aligned}$$

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

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

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

Proof of Corollary 4.2 We use induction again. By (39), $\boldsymbol{E}(W^{(2)}) = \frac{2}{\mu}$. For any k, we assume that $\boldsymbol{E}(W^{(k-1)}) = \frac{k-1}{\mu-\mu_k}$. Clearly, $\boldsymbol{E}(U^k) = \frac{1}{\mu}$ and the claim follows from $\boldsymbol{E}(W^{(k)}) = \boldsymbol{E}(U^k) + \frac{\mu-\mu_k}{\mu}\boldsymbol{E}(W^{(k-1)})$. □

Proof of Lemma 4.4 For convenience, let $\widetilde{\pi}_k \equiv \widetilde{\pi}_k(\mu_1, ..., \mu_k)$. Recall that for a k-state birth-death process with birth rates $\alpha_0, ..., \alpha_{k-1}$ and death rates $\beta_1, ..., \beta_k$, the steady-state probabilities are:

$$\widetilde{\pi}_i = \frac{\alpha_0\alpha_1\cdots\alpha_{i-1}}{\beta_1\beta_2\cdots\beta_i}\widetilde{\pi}_0 \quad , \quad i = 1, ..., k,$$

where

$$\widetilde{\pi}_0 = \left(\sum_{j=0}^{k} \frac{\alpha_0\alpha_1\cdots\alpha_{j-1}}{\beta_1\beta_2\cdots\beta_j} \right)^{-1}.$$

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

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

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

$$\widetilde{\pi}_k^{-1} = \sum_{j=0}^{k} \frac{(\mu_1 + ... + \mu_{j+1}) \cdot ... \cdot (\mu_1 + ... + \mu_k)}{\lambda^{k-j}}. \tag{40}$$

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

$$\begin{aligned} \max / \min \quad & H(\mu_1, ..., \mu_k) \equiv \sum_{j=0}^{k} \lambda^{j-k} (\mu_1 + ... + \mu_{j+1}) \cdot ... \cdot (\mu_1 + ... + \mu_k) \\ s.t. \quad & \mu_1 + ... + \mu_k = \mu \\ & \mu_1 \geq ... \geq \mu_k \geq 0. \end{aligned}$$

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 jth term:

$$\begin{aligned} \max / \min \quad & h_j(\mu_1, ..., \mu_k) \equiv \lambda^{j-k} (\mu_1 + \cdots + \mu_{j+1}) \cdot \cdots \cdot (\mu_1 + \cdots + \mu_k) \\ s.t. \quad & \mu_1 + \cdots + \mu_k = \mu \\ & \mu_1 \geq \cdots \geq \mu_k \geq 0. \end{aligned}$$

Solution of the maximization problem

Note that each multiplicand of $h_j(\cdot)$ is at most μ. 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 $\left(\frac{\mu}{\lambda}\right)^{k-j} = \rho^{j-k}$).

Solution of the minimization problem

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

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

Observe that for each $i = j+2, ..., k$, μ_i appears in exactly $k - i + 1$ multiplicands. It follows that any minimizing solution must satisfy $\mu_{j+2} \leq \mu_{j+3} \leq ... \leq \mu_k$ (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 $\mu_{j+2} = \mu_{j+3} = ... = \mu_k$. 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 μ_k, and hence $\mu_{j+2}, ..., \mu_{k-1}$, its largest possible value, which is μ/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 $\frac{k!\mu}{j!(\lambda k)^{k-j}}$).

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 $\mu_1, ..., \mu_k$: $\widetilde{\pi}_k^{-1}(\mu, 0, ..., 0) \geq \widetilde{\pi}_k^{-1}(\mu_1, ..., \mu_k) \geq \widetilde{\pi}_k^{-1}(\mu/k, ..., \mu/k)$, which is equivalent to (24). □

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. We would like to thank the referees for various useful comments.

References

[1] Ahmadi, R. H. (1997), "Managing capacity and flow at theme parks," *Operations Research* 45, 1-13.

[2] Akgun, O. T. , D. G. Down, and R. Righter (2013), "Energy-aware scheduling on heterogeneous processors," *IEEE Transactions on Automatic Control*.

[3] Almeida, J., V. Almeida, D. Ardagna, Í. Cunha, C. Francalanci and M. Trubian (2010), "Joint admission control and resource allocation in virtualized servers," *Journal of Parallel and Distributed Computing* 70, 344-362.

[4] Alnowibet, K. A. and H. Perros (2009), "Nonstationary analysis of the loss queue and of queueing networks of loss queues," *European Journal of Operational Research* 196(3), 1015-1030.

[5] Altman, E. (1996), "Non zero-sum stochastic games in admission, service and routing control in queueing systems," *Queueing Systems* 23, 259-279.

[6] Altman, E. and R. Hassin (2002), "Non-threshold equilibrium for customers joining an M/G/1 queue," *Proceedings of the 10th International Symposium of Dynamic Games.*

[7] Atar, R. (2012), "A diffusion regime with non-degenerate slowdown," *Operations Research* 60, 490-500.

[8] Bell, C. E. and S. Stidham Jr. (1983), "Individual versus social optimization in allocation of customers to alternative servers," *Management Science* 29, 831-839.

[9] Baron, O., O. Berman, and D. Krass (2008), "Facility location with stochastic demand and constraints on waiting time," *Manufacturing and Service Operations Management*, 10(3), 484-505.

[10] Chao, X., L. Liu and S. Zheng (2003), "Resource allocation in multisite service systems with intersite customer flows," *Management Science* 49, 1739-1752.

[11] Chen, H. and D. D. Yao (2001), Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization, Springer-Verlag, New York.

[12] Dallery, Y. and K. Stecke (1990), "On the optimal allocation of servers and workloads in closed queueing networks," *Operations Research* 38, 694-703.

[13] Edelson, N. M. and D. K. Hildebrand (1975), "Congestion tolls for Poisson queuing processes," *Econometrica* 43, 81-92.

[14] Garnett, O., A. Mandelbaum and M. Reiman M. (2002), "Designing a Call Center with Impatient Customers," *Manufacturing and Service Operations Management*, 4(3), 208-227.

[15] Glasserman, P. (1996), "Allocating production capacity among multiple products," *Operations Research* 44, 724-734.

[16] Green, L. V. and D. Guha (1995), "On the efficiency of imbalance in multi-facility multi-server service systems," *Management Science* 41, 179-187.

[17] Halfin, S. and W. Whitt (1981), "Heavy-traffic limits for queues with many exponential servers," *Operations Research* 29, 567-588.

[18] Harel, A. (2011), "Convexity results for the Erlang delay and Loss formulae when the Server utilization is held constant," Operations Research 59(6), 1420-1426.

[19] Hassin, R. and M. Haviv (2003), *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*, Kluwer International Series.

[20] Hall, J. and E. Porteus (2000), "Customer service competition in capacitated systems," *manufacturing & Service Operations Management* 2, 144-165.

[21] Hillier, F. S. and K. C. So (1996), "On the simultaneous optimization of server and work allocations in production line systems with variable processing times," *Operations Research* 44, 435-443.

[22] Hong, K. and C. Lee (2012), "Integrated pricing and capacity decision for a telecommunication service provider," *Multimedia Tools and Applications*, DOI 10.1007/s11042-012-1030-3.

[23] Hung, H-C. and M. E. Posner (2007), "Allocation of jobs and identical resources with two pooling centers," *Queueing Systems* 55, 179-194.

[24] Johari, R., G.Y. Weintraub, and B. Van Roy (2010) "Investment and market structure in industries with congestion," *Operations Research* 58, 1303-1317.

[25] Kerner, Y., (2011), "Equilibrium joining probabilities for an M/G/1 queue," *Games and Economic Behavior* 71, 521-526.

[26] Kleinrock, R. L. (1964), *Communication Nets: Stochastic message flow and delay*, Dover publications, Inc, New York.

[27] Kleinrock, R. L. (1975), *Queueing Systems: Volume I: Theory*, Wiley Interscience, New York.

[28] de Kok, T. G. (2000), "Capacity allocation and outsourcing in a process industry," *Int. J. production Economics* 68, 229-239.

[29] Korilis, Y. A., A. A. Lazar and A. Orda (1995), "Architecting noncooperative networks," *IEEE Journal on Selected Areas in Communication* 13, 1241-1251.

[30] Korilis, Y. A., A. A. Lazar and A. Orda (1997), "Capacity allocation under noncooperative routing," *IEEE Transactions on Automatic Control* 42, 309-325.

[31] Kostami, V. and A. R. Ward (2009), "Managing service systems with an offline waiting option and customer abandonment," *Manufacturing & Service Operations Management* 11, 644-656.

[32] Lu, Y., A. Musalem, M. Olivares and A. Schilkrut (2013), "Measuring the effect of queues on customer purchases," *Management Science*, published online April 2013.

[33] Mandjes, M. and J. Timmer (2007), "A duopoly model with heterogeneous congestion-sensitive customers," *European Journal of Operational Research* 176, 445-467.

[34] Naor, P. (1969), "The regulation of queue size by levying tolls," *Econometrica* 37, 15-24.

[35] Rolfe, A. J. (1971), "A note on marginal allocation in multiple-server service systems," *Management Science* 17, 656-658.

[36] Rothkopf, M. H. and P. Rech (1987), "Perspectives on queues: combining queues is not always beneficial, *Operations Research* 35, 906-909.

[37] Schweitzer, P. J. and A. Seidmann (1991), "Optimizing processing rates for flexible manufacturing systems," *Management Science* 37, 454-466.

[38] Shanthikumar, J. G. and S. H. Xu (1997), "Asymptotically optimal routing and service rate allocation in a multiserver queueing system," *Operations Research* 45, 464-469.

[39] Shanthikumar, J. G. and D. D. Yao (1987), "Optimal server allocation in a system of multi-server stations," *Management Science* 33, 1173-1180.

[40] Shumsky, R. A. and F. Zhang (2009), "Dynamic capacity management with substitution," *Operations Research* 57, 671-684.

[41] Stecke, K. E. and J. J. Solberg (1985), "The optimality of unbalancing both workloads and machine group sizes in closed queueing networks of multiserver queues," *Operations Research* 33, 882-910.

[42] Tan, Y., Y. Lu and C. H. Xia (2012), "Provisioning for Large Scale Loss Network Systems with Applications in Cloud Computing," *SIGMETRICS Performance Evaluation Review* 40(3), 83-85.

[43] Wein, L. M. (1989), "Capacity allocation in generalized Jackson networks," *Operations Research Letters* 8, 143-146.

[44] Whitt, W. (2004), "Efficiency-Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments," *Management Science* 50(10), 1449-1461.

[45] Xu, S. H. and J. G. Shanthikumar (1993), "Optimal expulsion control – a dual approach to admission control of an ordered-entry system," *Operations Research* 41, 1137-1152.

[46] Yolken, B. and N. Bambos (2011), "Game based capacity allocation for utility computing environments," *Telecommunication Systems* 47, 165-181.

[47] Zhang, Y., O. Berman, P. Marcotte and V. Verter (2010), "A bilevel model for preventive healthcare facility network design with congestion," *IIE Transactions* 42, 865-880.