Performance Evaluation 154 (2022) 102282

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Optimal pricing in multi server systems☆,☆☆

Ashok Krishnan K.S. [a,*,1], Chandramani Singh [b], Siva Theja Maguluri [c], Parimal Parag [d]

[a] Qualcomm India Pvt. Ltd., Bangalore, KA 560066, India
[b] Department of ESE, Indian Institute of Science, Bangalore, KA 560012, India
[c] School of ISyE, Georgia Institute of Technology, Atlanta, GA 30332-0205, United States of America
[d] Department of ECE, Indian Institute of Science, Bangalore, KA 560012, India

ARTICLE INFO

Article history:
Received 5 May 2021
Received in revised form 15 September 2021
Accepted 21 December 2021
Available online 7 January 2022

Keywords:
Multi-server systems
Optimal pricing
Markov decision processes

ABSTRACT

We study optimal service pricing in server farms where customers arrive according to a renewal process and have independent and identical (*i.i.d.*) exponential service times and *i.i.d.* valuations of the service. The service provider charges a time varying service fee aiming at maximizing its revenue rate. The customers that find free servers and service fees lesser than their valuation join for the service else they leave without waiting. We consider both finite server and infinite server farms. We solve the optimal pricing problems using the framework of Markov decision problems. We show that the optimal prices depend on the number of free servers. We propose algorithms to compute the optimal prices. We also establish several properties of the optimal prices and the corresponding revenue rates in the case of Poisson customer arrivals. We illustrate all our findings via numerical evaluation.

1. Introduction

Server farms refer to centrally maintained collections of computer servers or processors intended to provide a service (or a class of services) to customers. Over the past decade, server farms have mushroomed to keep up with the massive demand for both data storage and computation, which continues to increase at breakneck speed. These include services such as AWS EC2 and Azure [2]. Server farms offer a cost-effective alternative to customers wherein they need not spend initial setup and maintenance of a service facility. These also allow customers to dynamically scale resource utilization and provide redundancy against failure of specific hardware. However, service providers incur considerable costs on hardware, cooling, power, security etc. Sustained proliferation of data farms is contingent on providers profiting through service charges levied on the customers.

☆ Part of the work was presented at WiOpt, 2020 (Krishnan KS et al., 2020 [1]).

☆☆ Chandramani Singh was supported in part by the Visvesvaraya Young Faculty Research Fellowship and in part by the Centre for Networked Intelligence (a Cisco corporate social responsibility (CSR) initiative) at IISc. Siva Theja Maguluri was partially supported by NSF Grant CCF – 1850439. Parimal Parag was supported in part by the Science and Engineering Research Board (SERB) under Grant DSTO-1677, in part by the Department of Telecommunications, Government of India, under Grant DOTC-0001, in part by the Centre for Networked Intelligence (a Cisco CSR initiative) at IISc, and in part by the Robert Bosch Centre for Cyber-Physical Systems at IISc.

* Corresponding author.
E-mail addresses: ashokk@alum.iisc.ac.in (Ashok Krishnan K.S.), chandra@iisc.ac.in (C. Singh), siva.theja@gatech.edu (S.T. Maguluri), parimal@iisc.ac.in (P. Parag).

[1] Ashok Krishnan K.S. was with the Indian Institute of Science when this work was done.

https://doi.org/10.1016/j.peva.2021.102282
0166-5316/

Optimal service pricing is central to the thriving operation of server farms [3,4]. Service providers' earnings come from service charges levied on the customers. Different customers may have different utilities (or, valuation) of the service. Also, in a server farm with a waiting queue, a customer's valuation will also depend on its expected waiting time, i.e., on the queue length on its arrival. The customers opt for the service only if their valuation of the service exceeds service charge. Clearly service charges directly impact service provider's revenue. These along with customers' valuation also determine servers' occupancy and congestion which in turn governs future customers' valuation. We thus see that determining optimal prices is a complex problem. The problem is further complicated by the fact that service providers cannot a priori assess customers' valuation though they often know value distributions based on historical data.

We consider a multiple server system that offers service to stochastically arriving customers. Customers' service durations are random. We do not assume any waiting queue. The service provider sells the service to customers at potentially time varying prices. Different customers also have different values of the service. The service provider does not know customers' values but knows value distribution. A customer who finds at least one idle server on arrival opts for the service if and only if its value exceeds the current service charge. The customers who find all the servers busy on arrival leave the system without getting served. The service provider aims to maximize the average revenue rate by setting appropriate prices. We derive optimal prices as a function of the number of idle servers. We also study various properties of the optimal prices and optimal revenue rate vis-a-vis total number of servers, customer arrival rate, average service time etc.

1.1. Our contribution

We assume a service provider with K servers. We further assume that the customers arrive according to a renewal process, having *i.i.d.* inter arrival times, *i.i.d.* exponential service times and *i.i.d.* valuations for the service. We formulate the optimal pricing problem that maximizes the service provider's revenue. Finding and using the optimal pricing scheme in a multi-server system may in general be challenging. The challenges are two-fold. First, finding the optimal policy is challenging given the model parameters. Second, there are practical challenges in implementing these policies, since most service providers in practice prefer simple policies. We address both these challenges in the paper. First, we study the uniform pricing problem as a sub optimal but easy to implement policy, and obtain performance bounds for this policy. Then, we obtain the revenue maximizing pricing policy by solving an associated Markov decision problem. We study the properties of the optimal solution, and compare its performance to that of sub optimal policies discussed previously. Following is a preview of our main results.

1. We observe that for the system with infinitely many servers (i.e., $K = \infty$), the optimal service prices are uniform, i.e., independent of the number of occupied servers.
2. We study optimal uniform pricing for K server system ($K < \infty$). These policies are sub optimal, but are simpler to compute and implement. We derive a bound on the revenue rate for the optimal uniform price. We observe that the potential loss in revenue by using uniform pricing, is small under low load. We also study asymptotic revenue rates for uniform pricing as arrival rates are scaled, and show that limiting revenue can go to zero for certain arrival processes.
3. We further observe that the potential loss in revenue from using a uniform pricing, may be large under heavy traffic. This suggests that the service provider has an incentive to use the optimal pricing scheme, even though it may be potentially more complex. For finite server systems, we frame the revenue rate maximization problem as a continuous time Markov control problem. We show that the optimal prices depend on the number of occupied servers, and can be obtained via solving a fixed point iteration.
4. We study the dependence of optimal prices and corresponding revenue rates on customer arrival rates, service rates, and the number of servers K, in the case of Poisson customer arrivals. We show that the optimal revenue is increasing in arrival rate, service rate and number of servers. We also show that the revenue per arrival rate, revenue per service rate and revenue per server are decreasing in their respective variables.
5. We illustrate all our findings via numerical results. Our numerical studies also provide additional insights on the behavior of optimal prices with respect to arrival and service rates.

1.2. Related work

Cloud computing facilities that host a large number of data servers face the problem of optimizing the utilization of these servers. Designing an optimal pricing policy is a crucial step in extracting the best possible revenue from the system [5,6]. Since a cloud compute facility can be modeled as a bunch of servers with an associated queueing process, the cloud pricing problem can be studied as a problem of pricing in queues. One of the earliest works that studied pricing of queues was [7], in which the entry of customers to a queue was regulated using tolls. Customers can decide to balk or join the queue, after observing the queue size. Such systems are called *observable*. Customers join the system if the difference between their valuation of the job and the cost of waiting exceeds the admission price to the queue. This translates to a threshold type policy — if the queue length is greater than the threshold, the customers balk; else they join. The optimal threshold may vary, depending on whether we want to maximize the total social utility or the revenue.

It was shown that in [7] that the socially optimal threshold was higher than the threshold for revenue maximization. A subsequent work [8] shows that, the revenue maximizing and socially optimal toll values can be the same, provided a two-part tariff is imposed. There have been a number of other works which looked at extensions of [7] or at related models. The effect of the reward variance on the performance is studied in [9]. In [10], the author examines whether it is always optimal for a profit maximizing service provider to hide the queue length from an arriving customer. It is shown that there are thresholds of arrival rates, below which it is optimal for the service provider to hide the queue state information, and above which it is optimal to reveal. These, and numerous other related works, have been summarized in [11,12].

Optimization of revenue in queueing systems has been extensively studied. In one of the first works in this direction, [13], the author studies optimal pricing for an $M/M/s$ queue with finite waiting room. He shows that the optimal prices are monotone increasing in the number of customers waiting in the system. A similar monotonicity result for the price as a function of the number of customers, for a similar system but with no waiting room, is shown in [14]. In [15], the authors look at the revenue maximization problem from the perspective of the service provider. They are interested in maximizing the expected discounted revenue, while keeping the queueing model of [7]. They obtain a revenue optimizing threshold queue length beyond which entries are not allowed into the queue. This threshold can be computed numerically. All customers who see a waiting queue length smaller than this threshold, pay a price equal to the difference between their valuation and waiting cost. In [16], an explicit form is derived for the threshold obtained in the previous work, and they characterize the earning rate asymptotically. However, both aforementioned works provide explicit solutions in the case of fixed service valuation (or simple valuation distributions, such as a valuation which takes two values). They do not provide explicit solutions for valuations with continuous support and general distributions. In [17], the authors study optimal pricing in finite capacity queueing systems. However, they consider the sub optimal class of static prices, where the prices charged by the service provider is independent of the number of customers present in the system. They find the best prices in this class, and study its variation with the number of servers. Another work which looks at optimal pricing in finite capacity queueing system is [18]. Here, under the assumption that the *generalized hazard rate* of the valuation distribution is strictly increasing, the authors obtain the optimal, revenue maximizing policy. However, this assumption does not hold for all distributions. Another work which looks at dynamic pricing in queues is [19]. The authors consider a multi server queueing system with finite waiting room. They prove that an optimal monotone policy exists, under the average reward criterion. Existence of an optimal monotone policy for a system with two tandem queues is provided in [20]. Apart from these, there is substantial literature which looks at revenue optimization of different models of queueing systems using an MDP framework and obtain existence and structural results on the optimal policy. These include works such as [3,21]. In [22], the authors study optimal pricing for a two class queueing system, and obtain structural results for the optimal prices. A comprehensive survey of different dynamic pricing techniques is available in [23]. In a recent work [24], the authors prove the existence a static pricing policy that obtains 78.9% of the optimal profit in a system with multiple reusable resources. They assume that the arrivals form a Poisson process, and further, that the revenue rate is a concave function of the arrival rate. This static pricing policy is obtained as a function of the optimal (state dependent) pricing policy.

Since explicit computation of optimal prices and revenues is difficult in general, a number of works study the pricing and revenue problem in asymptotic regimes, and obtain useful insights. That dynamic pricing can lead to lower *variability* in the revenue of pricing system, as opposed to static pricing, is shown in [25]. They use an asymptotic analysis to show that the revenue loss due to randomness is lower for dynamic pricing than static pricing, when the customer valuation is random. An asymptotically optimal pricing is obtained in [26] when the customers are delay sensitive but have fixed valuations, for a system with two classes of customers. An asymptotic approach to the dynamic pricing problem is given in [27], where the solution to an approximating diffusion control problem is used as a solution. Another asymptotic regime is the large capacity regime, explored in [28]. They aim to minimize the cost to the customers caused by delay, when the delay cost is a non-linear function of delay. The authors obtain different optimal policies, corresponding to different types of cost functions in this asymptotic regime.

As opposed to works such as [3,13,14,19–21,29] which show existence of the optimal policy and proceed to obtain structural insights, in this work, we explicitly obtain the optimal price as a solution of a fixed point equation. Moreover we consider arrival processes with general distribution, which generalizes the Poisson assumption in these works. We do not restrict ourselves to the increasing hazard rate assumption of [18], and thus have a more general result. Since we assume valuations with a general distribution, our result is more general than [15].

Notation: Before we proceed, we introduce the following notation that we use throughout in this article. We denote the set of positive integers by $\mathbb{N}$, the set of non-negative integers by $\mathbb{Z}_+$, the set of non-negative reals by $\mathbb{R}_+$, the set of first n positive integers by $[n]$, and the set of non-negative real vectors of length n by $\mathbb{R}_+^n$. A list of some commonly used symbols in this paper is given in Table 1, for easy reference.

2. System model

We model a compute cluster of K servers as a queueing system, where jobs arrive with some service time and a valuation. The price of admission into the compute cluster is updated at each job arrival. If the admission price is smaller than the valuation, then the job is admitted into the system. The job pays the admission price to the compute cluster. In

Table 1
List of notation commonly used in this paper.

Symbol	Meaning
λ	Arrival rate
μ	Service rate of one server
ρ	Load factor $\frac{\lambda}{\mu}$
V_i	Valuation of job i
$\overline{G}(u)$	$\mathbb{P}[V_1 \geq u]$
p_k	Admission price when k jobs are present in the system
$\mathbf{p}$	Price vector $(p_0, ..., p_{K-1})$
$X(t)$	Number of busy servers at time t
$[K]$	$\{1, \ldots, K\}$
$\mathcal{X}$	$\{0, \ldots, K\}$
$\mathcal{X}'$	$\{0, \ldots, K-1\}$
$R(K, \mathbf{p})$	Revenue rate for K-server system with price vector $\mathbf{p}$
$\mathbf{p}^*$	Optimal price vector for K server system
p_K^*	Optimal uniform price for K server system
π	Marginal distribution of number of busy servers seen by arriving customer
U_i	ith inter arrival time
$\phi(s)$	Laplace Stieltjes transform of interarrival time $= \mathbb{E}[e^{-sU_1}]$
β_j	$\prod_{m=1}^{j} \frac{1-\phi(m\mu)}{\phi(m\mu)}$
θ^*	Optimal revenue rate

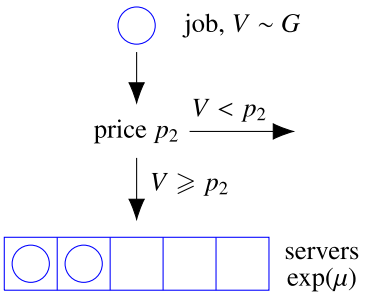


Fig. 1. We depict a 5 server system. A job with valuation V arrives when two servers are busy. The admission price is p_2 and the job joins the system if its valuation $V \geqslant p_2$.

this case, the compute cluster earns the revenue equal to the admission price, and the job leaves upon service completion. If admission price is larger than the valuation, the job leaves and never returns. A 5 server system is depicted in Fig. 1. Two servers are occupied, and a new arrival with valuation V attempts to join the system.

The arrival process is modeled as a renewal process with *i.i.d.* inter-arrival times having mean $\frac{1}{\lambda}$. Arrival processes are typically modeled by Poisson processes in the literature [12]. Our model is a generalization of this assumption, where the sequence of interarrival times $U \triangleq (U_n \in \mathbb{R}_+ : n \in \mathbb{N})$ remains *i.i.d.* however with a general distribution $F : \mathbb{R}_+ \to [0, 1]$. The sequence of arrival instants of customers is denoted by $A \triangleq (A_n \in \mathbb{R}_+ : n \in \mathbb{N})$, such that renewal instants $A_n = \sum_{i=1}^{n} U_i$. We denote the counting process associated with the arrival sequence by $N_t : \mathbb{R}_+ \to \mathbb{Z}_+$ such that

$$N_t \triangleq \sum_{n \in \mathbb{N}} \mathbb{1}_{\{A_n \leqslant t\}}$$

is the number of arrivals until time t.

Service time requirements of arriving jobs at compute clusters can be modeled as *i.i.d.* random variables with a shifted exponential distribution [30,31], with a constant start-up time and a random memoryless service time. When the job sizes are large,[2] exponential distribution is a good approximation for the job service requirement. As such, we assume that the service time requirements of arriving jobs are an *i.i.d.* random sequence $S \triangleq (S_n \in \mathbb{R}_+ : n \in \mathbb{N})$, distributed exponentially with mean $\frac{1}{\mu}$.

[2] When the job sizes are large, the mean of the memoryless service time dominates the constant start-up time.

A natural assumption would be to assume that service time requirements affect the job valuation, i.e. higher the service time requirement, larger the valuation. However, this assumption has two caveats, the first that the job is aware of its requirements *apriori*, and the second that all jobs are valued in a homogeneous manner. In practice, jobs maybe unaware of service time requirements, and they maybe valued heterogeneously. To keep our model general and analytically tractable, we assume that each job has a random *i.i.d.* positive valuation sampled from a continuous cumulative distribution $G : \mathbb{R}_+ \to [0, 1]$. We denote the *i.i.d.* random sequence of job valuations by $V \triangleq (V_n \in \mathbb{R}_+ : n \in \mathbb{N})$ with finite mean $\mathbb{E}V_1$. We note that this remains a more general assumption, when compared to constant valuation considered in the literature [15]. Random valuation models the scenario where the customers are not identical in their assessment of the value of the job. However, they are drawn from a homogeneous population. We assume that the distribution G is known. However, in general it may be necessary to estimate this distribution. For example, see [32] where the authors use kernel density estimation methods to estimate G.

Recall that we have a finite compute cluster with K servers, and we assume that incoming jobs join a unique[3] idle[4] server if admitted. That is, a job leaves if either its valuation is lower than the admission price or all K servers are busy. We assume that the server sets a price, that depends only on the number of busy servers at any job arrival instant. That is, if we let k be the number of busy servers at a job arrival, then the admission price is p_k. The number of busy servers represents the resource crunch at the service provider. It is reasonable to expect the service provider to set its prices as a function of this number. To capture the effect of a job leaving when all K servers are busy, we can define the price $p_K \triangleq \infty$. Therefore, if there are k busy servers at arrival instant of nth job with valuation V_n, then we can indicate its admission by $\mathbb{1}_{\{V_n \geqslant p_k\}}$, and the revenue earned by the cluster by $p_k \mathbb{1}_{\{V_n \geqslant p_k\}}$. Note that in our model a customer leaves when no free server is available, or when the price posted is large. Such a model is common in the literature and is referred to as a *loss model* [3,14,24,33]. This is in agreement with majority of cloud computing modeling in literature. For example, Bouterse and Perros [34] study capacity planning of cloud infrastructure considering a finite number of application seats and no queueing. Vakilinia et al. [35] also consider resource allocation in cloud computing centers with finite number of VMs assuming that the jobs are blocked if there are not enough idle VMs to serve them. This also corresponds to a situation where the service provider is not a monopoly — there are other service providers to whom the customer can turn to, when the server under consideration is busy or expensive.

We denote the number of busy servers in the system at time t by $X(t) \in \mathcal{X} \triangleq \{0, \ldots, K\}$. Since the admission price depends only on the number of busy servers at the arrival instants, it follows from the memoryless property of service times that the number of busy servers specify the system state completely. Since we have set $p_K = \infty$, the state space $\mathcal{X}$ can be reduced to $\mathcal{X}' \triangleq \{0, \ldots, K-1\}$. We denote a state-dependent price vector by $\mathbf{p} = (p_0, \ldots, p_{K-1}) \in \mathbb{R}_+^{\mathcal{X}'}$. We denote the number of busy servers in the system seen by nth arriving customer as $Z_n \triangleq X(A_n^-)$. We denote the revenue earned by the cluster until time by $R(t)$, which can be written as

$$R(t) = \sum_{n=1}^{N_t} \sum_{k=0}^{K-1} p_k \mathbb{1}_{\{V_n \geqslant p_k\}} \mathbb{1}_{\{Z_n = k\}}. \tag{1}$$

The limiting revenue rate for this K server system with the state-dependent price vector $\mathbf{p}$ is denoted by

$$R(K, \mathbf{p}) \triangleq \lim_{t \to \infty} \frac{\mathbb{E}R(t)}{t}. \tag{2}$$

Our main goal is to find the state-dependent pricing vector $\mathbf{p}$ that maximizes revenue. Formally, we solve the following problem.

Problem 1. Find the optimal price vector $\mathbf{p}^* \in \mathbb{R}_+^{\mathcal{X}'}$ that maximizes the limiting system revenue rate $R(K, \mathbf{p})$. That is, we wish to find

$$\mathbf{p}^* \triangleq \arg\max \left\{ R(K, \mathbf{p}) : \mathbf{p} \in \mathbb{R}_+^{\mathcal{X}'} \right\}.$$

Denoting a vector of all ones by $\mathbf{1} \in \mathbb{R}_+^{\mathcal{X}'}$ and a fixed price $p \geqslant 0$, we can denote the *uniform price* vector by $p\mathbf{1}$. In this case, the price charged to a customer is independent of the state of the system. We next find the uniform price that maximizes the revenue rate.

Problem 2. Find the uniform price p that maximizes the limiting system revenue rate $R(K, p\mathbf{1})$. That is, we wish to find

$$p^* \triangleq \arg\max \{ R(K, p\mathbf{1}) : p \in \mathbb{R}_+ \}.$$

In most systems, calculating the optimal uniform price turns out to be much simpler than obtaining the optimal price vector $\mathbf{p}*$. This also provides a benchmark for comparing the optimal policy and quantifying the improvement. We denote the optimal revenue rate by $R^* = R(K, \mathbf{p}^*)$, and compare it to the revenue rate $R(K, p^*\mathbf{1})$ for the best uniform pricing.

[3] We are not considering redundant replication of jobs, which is an interesting future direction. We will see that our problem remains difficult even without redundancy.

[4] This model can be extended to the case when jobs join the queue if all K servers are busy. In this case, the price will depend on the number of people existing in the queue, and the state space of possible prices increases.

Remark 1. In this paper, we assume that the price charged does not depend on the service time. In contrast, in cloud computing systems such as Amazon EC2 and Microsoft Azure, the customers are charged based on their service time. However, the results in this paper are also applicable in such settings with p_k being interpreted as price per unit service. This can be understood as follows. Suppose S_i is the random service duration of the ith job, then its price is $p_k S_i$, and its expected value is $\frac{p_k}{\mu}$. So, the mean revenue expression, the Bellman's equation characterizing the optimal pricing etc. remain unchanged the same except for a constant scaling factor $\frac{1}{\mu}$. Consequently, the optimal pricing analysis and the properties of the optimal prices also continue to hold.

3. Computation of revenue rate

Recall that the nth customer sees $Z_n = X(A_n^-)$ busy servers in the system. We denote the indicator to the event that the job valuation of nth customer is higher than the system admission price, by $e_n \triangleq \mathbb{1}_{\{V_n \geqslant p_{Z_n}\}}$. From the memoryless property of service time requirements, state dependent admission pricing, and the *i.i.d.* nature of job valuations, it follows that the process $((Z_n, e_n) \in \mathcal{X} \times \{0,1\} : n \in \mathbb{N})$ evolves as a discrete time Markov chain with finite state space.

We define $i^* \triangleq \min\{i \in \mathcal{X} : p_i > \text{supp}(G) \text{ or } p_i = \infty\}$, where with a slight abuse of notation, we use supp(G) to denote the support of the *probability density function* of the random variable with cumulative distribution function G. Since the valuations are *i.i.d.*, it can be verified that this Markov chain is irreducible and aperiodic over the reduced state space $\{0, \ldots, i^*\} \times \{0,1\}$. It follows that this reduced Markov chain has a positive invariant distribution $\tilde{\pi}$. For ease of notation, we can extend this distribution $\tilde{\pi}$ to the entire state space $\mathcal{X} \times \{0,1\}$ by defining $\tilde{\pi}(k,u) = 0$ for all $k > i^*$ and $u \in \{0,1\}$. Since valuations are *i.i.d.*, conditioned on the number of busy servers Z_n seen by the incoming arrival, the conditional mean of the random variable $e_n \in \{0,1\}$ is $\mathbb{E}[e_n | Z_n] = \overline{G}(p_{Z_n})$. That is, $\overline{G}(p_k)$ is the admission probability of an incoming customer that sees k busy servers. Let $\pi \triangleq (\pi_k : k \in \mathcal{X})$ be the marginal distribution of the number of busy servers seen by an incoming customer. In terms of the marginal distribution π and admission probability $\overline{G}(p_k)$, we can write the joint distribution $\tilde{\pi}$ as

$$\tilde{\pi}(k,1) = \overline{G}(p_k)\pi_k, \qquad \tilde{\pi}(k,0) = G(p_k)\pi_k. \tag{3}$$

Theorem 3. *Given the marginal distribution π and the state-dependent arrival rate $\lambda_k \triangleq \lambda \overline{G}(p_k)$, the limiting mean revenue rate for the cluster with state-dependent price vectors* $\mathbf{p}$ *is*

$$R(K, \mathbf{p}) = \sum_{k=0}^{K-1} \pi_k \lambda_k p_k. \tag{4}$$

Proof. From Eq. (1) for the cumulative revenue $R(t)$ until time t, we observe that the revenue earned by the cluster for nth arriving customer is denoted by $R(Z_n, e_n) = p_{Z_n} e_n$. Since N_t is a counting process for the arrival renewal process, we have $\lim_{t\to\infty} \frac{N_t}{t} = \lambda$ almost surely. Hence, we can write,

$$\lim_{t\to\infty} \frac{R(t)}{t} = \lambda \lim_{t\to\infty} \frac{1}{N_t} \sum_{n=1}^{N_t} R(Z_n, e_n).$$

By an ergodic theorem for Markov chains (Theorem 1.10. of [36]), it follows that, almost surely,

$$\lambda \lim_{N_t\to\infty} \frac{1}{N_t} \sum_{n=1}^{N_t} R(Z_n, e_n) = \lambda \sum_{k=0}^{K-1} p_k \sum_{u\in\{0,1\}} u\tilde{\pi}(k,u).$$

From Eq. (3) for $\tilde{\pi}(k,1)$ and the definition of state-dependent arrival rate λ_k, we see that, almost surely,

$$\lim_{t\to\infty} \frac{R(t)}{t} = \sum_{k=0}^{K-1} \pi_k \lambda_k p_k.$$

Since the revenue rate is upper bounded by average valuation of all incoming customers, we get $R(t)/t \leqslant (\sum_{n=1}^{N_t} V_n)/t$. Since the valuation sequence V is independent of the interarrival sequence U, it follows from the strong law of large numbers [37, Theorem 5.4.2] that the upper bound converges to $\lambda \mathbb{E}V_1$ almost surely. From the renewal reward theorem [38, Theorem 3.6.1], we see that $\lim_{t\to\infty}(\sum_{n=1}^{N_t} V_n)/t = \lambda \mathbb{E}V_1$. It follows from [37, Theorem 4.5.4], that $(R(t)/t : t > 0)$ is a uniformly integrable family of random variables. Consequently, we have

$$\lim_{t\to\infty} \frac{\mathbb{E}R(t)}{t} = \mathbb{E} \lim_{t\to\infty} \frac{R(t)}{t} = \sum_{k=0}^{K-1} \pi_k \lambda_k p_k. \quad \square$$

We will assume that the optimal price vector defined in Problem 1 exists and is finite.

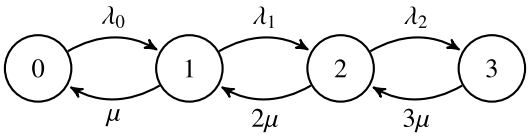


Fig. 2. Transition rate diagram of the CTMC denoting the number of busy servers with three identical servers and state dependent prices.

Assumption 4. There exists a finite optimal price $\mathbf{p}^*$ such that,

$$R(K, \mathbf{p}^*) = \max_{\mathbf{p} \in \mathbb{R}_+^{\mathscr{X}'}} R(K, \mathbf{p}).$$

We are interested in finding this $\mathbf{p}^*$ whenever it exists.

Remark 2. Consider the discrete-time discrete-state process $Z \triangleq (Z_n \in \mathcal{X} : n \in \mathbb{N})$, that denotes the number of busy servers seen by an incoming arrival. In the nth interarrival time U_n, the number of departures from $Z_n = k$ busy servers is denoted by random variable $N_k(U_n)$. Conditioned on duration U_n and $Z_n = k$, the probability of i departures is given by

$$P\{N_k(U_n) = i\} = \binom{k}{i}(1 - e^{\mu U_n})^i e^{-(k-i)\mu U_n},$$

for $i \in \{0, \ldots, k\}$. Since the interarrival time sequence U is *i.i.d.* with general distribution F, we can write the probability of $0 \leqslant i \leqslant k$ departures from k busy servers, as

$$\alpha_{k,i} \triangleq \mathbb{E}P\{N_k(U_n) = i\} = \int dF(x) P\{N_k(x) = i\}. \tag{5}$$

Then, we can write the homogeneous probability for the Markov chain Z to transition from state $k \in \mathcal{X}$ to state $j \in \{0, \ldots, \min\{k+1, K\}\}$ as

$$\overline{G}(p_k)\alpha_{k+1,k+1-j} + G(p_k)\alpha_{k,k-j}. \tag{6}$$

Therefore, one can find the transition probability matrix for the sampled Markov chain Z, for any general interarrival distribution F. It follows that the limiting distribution of the number of busy servers can be evaluated at least numerically.

Remark 3 (*Kelly [39]*)**.** The computation of marginal distribution π of the number of busy servers, is straightforward for Poisson arrivals. In this case, the evolution of the number of busy servers forms a birth–death Markov process, with transitions depicted in Fig. 2. Due to PASTA property, the distribution of number of busy servers seen by incoming customers is identical to the stationary distribution π of this Markov process. In particular, the distribution π is given in terms of the load factor $\rho \triangleq \frac{\lambda}{\mu}$ as

$$\pi_k = \begin{cases} \pi_0 \frac{\rho^k}{k!} \prod_{j=0}^{k-1} \overline{G}(p_j), & k \neq 0, \\ \left[1 + \sum_{k=1}^{K} \frac{\rho^k}{k!} \prod_{j=0}^{k-1} \overline{G}(p_j)\right]^{-1}, & k = 0. \end{cases} \tag{7}$$

We showed in Theorem 3, that the limiting revenue rate $R(K, \mathbf{p})$ can be written as a function of state-dependent price vector $\mathbf{p}$, marginal distribution π, and state-dependent arrival rates λ_k. Hence, the optimal price vector depends on the marginal distribution π. This marginal distribution is not easy to compute for the case of general inter arrival distribution, and its properties are not easy to establish even when inter arrival times are exponential. Therefore, we first consider a simple sub-class of prices, the uniform prices, where the price is independent of the state.

4. Uniform pricing

In this section, we will consider uniform pricing, not only when the number of servers is finite, but also when it is countably infinite. We show that uniform pricing is optimal in the infinite server scenario. Hence, optimizing the revenue over the simpler class of uniform prices is a reasonable solution, when the number of servers is large.

From Theorem 3, the following corollary is immediate for the revenue rate under uniform pricing.

Corollary 5. *The mean revenue rate for K-server system under uniform pricing $\mathbf{p} = p\mathbf{1}$ is*

$$R(K, p\mathbf{1}) = \lambda p \overline{G}(p)(1 - \pi_K(p)). \tag{8}$$

The revenue rate depends on the probability $1 - \pi_K(p)$ of arriving jobs seeing at least one idle server. This probability depends on the uniform price p. Hence, we obtain an expression for the blocking probability $\pi_K(p)$ to understand the revenue rate dependence on the uniform price p.

Proposition 6. *Consider a K-server system with uniform price* $\mathbf{p} = p\mathbf{1}$*. We denote the Laplace Stieltjes transform (LST) of the interarrival time U by* $\phi : \mathbb{R} \to \mathbb{R}_+$*, which is defined by* $\phi(s) \triangleq \mathbb{E}[e^{-sU}]$ *for all* $s \in \mathbb{R}$*. Defining*

$$\beta_j \triangleq \prod_{m=1}^{j} \frac{1-\phi(m\mu)}{\phi(m\mu)}, \tag{9}$$

we can write the limiting probability of finding all K servers busy as

$$\pi_K(p) = \Big(\sum_{j=0}^{K} \binom{K}{j} \overline{G}(p)^{-j} \beta_j\Big)^{-1}. \tag{10}$$

Proof. Recall that interarrival times U for jobs are *i.i.d.* with common distribution F. For uniform pricing $\mathbf{p} = p\mathbf{1}$, the admission indicator sequence $e \triangleq (e_n : n \in \mathbb{N})$ are *i.i.d.* Bernoulli with $\mathbb{E}e_n = \overline{G}(p)$. We write the number of arrivals between $(n-1)$th and nth admission as T_n, and observe that $T \triangleq (T_n : n \in \mathbb{N})$ is an *i.i.d.* geometric sequence with success probability $\overline{G}(p)$, and independent of inter-arrival sequence. We denote the inter-arrival times for admitted job as $\tilde{U} \triangleq (\tilde{U}_n : n \in \mathbb{N})$, where $\tilde{U}_n \triangleq \sum_{k=1}^{T_n} U_k$. It follows that $\tilde{U}$ is *i.i.d.* and thinned version of the original arrival process U. We can write the LST for the inter-arrival times of admitted jobs in terms of thinning probability $\overline{G}(p)$ as

$$\tilde{\phi}(x) = \sum_{n=1}^{\infty} \phi(x)^n G(p)^{n-1} \overline{G}(p) = \frac{\overline{G}(p)\phi(x)}{1 - G(p)\phi(x)}. \tag{11}$$

We observe that the evolution of the K-server pricing system under uniform pricing, is identical to that of a $G/M/K/K$ queueing system with *i.i.d.* inter-arrival times $\tilde{U}$ and K *i.i.d.* servers with exponential service rates μ. Therefore, the limiting blocking probability for this stable $G/M/K/K$ system can be written, using the Palm's formula [40], as

$$\pi_K(p) = \frac{1}{\sum_{j=0}^{K} \binom{K}{j} \prod_{m=1}^{j} \frac{1-\tilde{\phi}(m\mu)}{\tilde{\phi}(m\mu)}}.$$

Result follows from Eq. (11), which implies that $\frac{1-\tilde{\phi}(m\mu)}{\tilde{\phi}(m\mu)} = \frac{1}{\overline{G}(p)}\big(\frac{1-\phi(m\mu)}{\phi(m\mu)}\big)$. □

From above proposition, we can make the following observations for the limiting blocking probability.

Proposition 7. *For the finite server system under uniform pricing, the limiting blocking probability is nonincreasing in*

(a) uniform price for a fixed number of servers,
(b) number of servers for a fixed uniform price.

Proof. We recall the form of blocking probability $\pi_K(p)$ given in Eq. (10) for K-server system under uniform price p.

(a) Blocking probability $\pi_K(p)$ is non-decreasing in $\overline{G}(p)$, and the tail probability $\overline{G}(p)$ is non-increasing in uniform price p.
(b) From the definition of $\beta_j = \prod_{m=1}^{j} \frac{1-\phi(m\mu)}{\phi(m\mu)}$ in Eq. (9), the binomial identity $\binom{K+1}{j} = \binom{K}{j} + \binom{K}{j-1}$, and positivity of all terms, we observe that

$$\pi_{K+1}(p) \leqslant \pi_K(p). \quad \square$$

Remark 4. Above proposition implies that a higher price leads to a lower blocking probability for the same number of servers, since some jobs will leave without joining. It also implies that block probability is reduced by increasing the number of servers while keeping the price fixed.

Definition 8. For a K-server system, we can define the optimal uniform price p_K^* as the price that maximizes the mean revenue rate under uniform pricing. That is,

$$p_K^* \triangleq \arg\max_{p>0} R(K, p\mathbf{1}) = \arg\max_{p>0} \lambda p \overline{G}(p)(1 - \pi_K(p)).$$

The corresponding revenue rate for this price is $R(K, p_K^*\mathbf{1})$.

4.1. Properties of revenue rate under uniform pricing

We now show that this optimal revenue increases with the number of servers.

Lemma 9. *The mean revenue rate for a finite server system under uniform pricing is increasing in the number of servers.*

Proof. Consider the optimal uniform price p_K^* for K server system. When this uniform price is applied to a $K+1$ server system, then the mean revenue rate of this system is given by Eq. (8), as

$$R(K+1, p_K^*\mathbf{1}) = \lambda p_K^* \bar{G}(p_K^*)(1-\pi_{K+1}(p_K^*)).$$

From the monotonicity of blocking probability with the number of servers in Proposition 7, for finite server system under uniform price, it follows that $\pi_{K+1}(p) \leqslant \pi_K(p)$. Therefore, we have

$$R(K, p_K^*\mathbf{1}) \leqslant \lambda p_K^* \bar{G}(p_K^*)(1-\pi_{K+1}(p_K^*)) = R(K+1, p_K^*\mathbf{1}).$$

Since the optimal uniform price for $K+1$ server system is p_{K+1}^*, we obtain that $R(K+1, p_K^*\mathbf{1}) \leqslant R(K+1, p_{K+1}^*\mathbf{1})$ and the result follows. □

Remark 5. We consider the uniform pricing for the limiting case when the number of servers grow unboundedly large. If the uniform price is p, then any arriving job with valuation higher than p joins the system. Since there is no blocking due to unavailability of servers, the mean revenue rate for the limiting system is $\lambda p\bar{G}(p)$. Therefore, the optimal uniform price for infinite server system is given by

$$p_\infty^* \triangleq \arg\max_p p\bar{G}(p). \tag{12}$$

We next see that the optimal uniform price for infinite server system is lower than the optimal uniform price for any finite server system.

Lemma 10. *Let p_∞^* defined in Eq. (12) and p_K^* defined in Eq. (8) be the optimal uniform prices for infinite and finite K-server systems respectively. Then, $p_K^* \geqslant p_\infty^*$ for all finite K.*

Proof. Let $\pi_K(p_K^*)$ and $\pi_K(p_\infty^*)$ be the blocking probabilities for K-server system with uniform prices $p_K^*\mathbf{1}$ and $p_\infty^*\mathbf{1}$ respectively. From the definition of optimal uniform price for infinite server system, it follows that $p_\infty^*\bar{G}(p_\infty^*) \geqslant p_K^*\bar{G}(p_K^*)$. From the definition of optimal uniform price for finite server systems, it follows that

$$\begin{aligned}(1-\pi_K(p_K^*))p_\infty^*\bar{G}(p_\infty^*) &\geqslant (1-\pi_K(p_K^*))p_K^*\bar{G}(p_K^*)\\ &\geqslant (1-\pi_K(p_\infty^*))p_\infty^*\bar{G}(p_\infty^*).\end{aligned}$$

Therefore, we have $\pi_K(p_\infty^*) \geqslant \pi_K(p_K^*)$. The result follows from the monotone decrease of blocking probability π_K in uniform price p from Proposition 7. □

We now establish that the mean revenue rate in the infinite server system is maximized by the optimal uniform pricing.

Proposition 11. *The optimal uniform pricing p_∞^* maximizes the mean revenue rate for infinite server system.*

Proof. By definition of the optimal revenue rate for K servers, the optimal revenue rate $R(K, \mathbf{p}^*)$ with state dependent pricing $\mathbf{p}^*$ is greater than the maximum revenue rate $R(K, p_K^*\mathbf{1})$ under uniform pricing $p_K^*\mathbf{1}$. That is,

$$R(K, p_K^*\mathbf{1}) \leqslant R(K, \mathbf{p}^*).$$

From Eq. (4) we obtain that the optimal mean revenue is a convex combination of $(\lambda p_k\bar{G}(p_k) : k \in \mathcal{X})$, where the optimal price vector is $\mathbf{p}^* = (p_0, \ldots, p_{K-1})$. From the definition of p_∞^* in Eq. (12), we get

$$R(K, \mathbf{p}^*) \leqslant \lambda \max_{k\in\mathcal{X}} p_k\bar{G}(p_k) \leqslant \lambda p_\infty^*\bar{G}(p_\infty^*).$$

From Lemma 9, the optimal revenue rate $R(K, p_K^*)$ is monotonically increasing in the number of servers K. The result follows from taking the limit $K \to \infty$ in the above equation. □

Thus, for a system with a large number of servers, choosing the optimal uniform price, is close to optimal. We note that system state for a finite server system can equivalently be represented by the number of idle servers. In an infinite server system, the number of idle servers is always infinite, and hence state dependent pricing reduces to state independent pricing. With this view, it is expected that optimal pricing for an infinite server system will be uniform. We next bound the optimal revenue rate in terms of the maximum revenue rate under uniform pricing.

Lemma 12. *Let p_∞^* and p_K^* be optimal uniform prices of infinite and finite K-server systems, and let $\mathbf{p}^*$ be the optimal state dependent price vector for the K server system. If the blocking probability of the K-server system under uniform price p_∞^* is denoted by $\pi_K(p_\infty^*)$, then*

$$R(K, p_K^*\mathbf{1}) \leqslant R(K, \mathbf{p}^*) \leqslant \frac{R(K, p_K^*\mathbf{1})}{1-\pi_K(p_\infty^*)}.$$

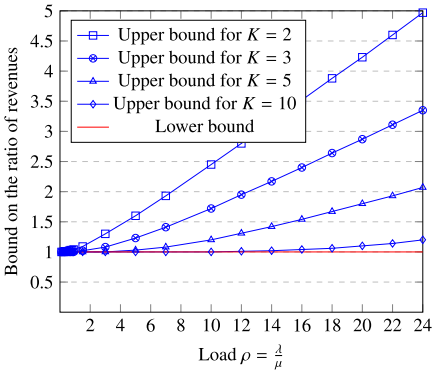


Fig. 3. Upper and lower bounds for ratio of optimal revenue to optimal uniform revenue as a function of load ρ. We have different upper bounds for different number of servers K, and the lower bound is uniformly 1.

Proof. The first inequality follows from the definition of the optimal revenue rate. To prove the second inequality, recall that optimal revenue rate under uniform pricing is increasing in the number of servers, i.e. $R(K, \mathbf{p}^*) \leqslant \lambda p^*_\infty \overline{G}(p^*_\infty)$. Multiplying both sides by $1 - \pi_K(p^*_\infty)$, we see that,

$$(1 - \pi_K(p^*_\infty))R(K, \mathbf{p}^*) \leqslant R(K, p^*_\infty \mathbf{1}).$$

Since the right hand side term is the revenue rate of a K server system with uniform price p^*_∞, it can be upper bounded by maximum revenue rate $R(K, p^*_K \mathbf{1})$ under optimal uniform price p^*_K. □

The above lemma implies that the optimal revenue rate converges to maximum revenue rate under uniform pricing as the number of servers K grows large. We show this bound in Fig. 3, by plotting the upper and lower bounds on $\frac{R(K,\mathbf{p}^*)}{R(K,p^*_K\mathbf{1})}$ for different values of load factor ρ and different number of servers K. For this plot, we have taken an exponential valuation function with parameter 1. In addition, the arrivals are assumed to be Poisson and we have taken the memoryless service rate to be 1 for each server. The upper bound tightens as we increase the number of servers. However, the bound can be made loose by scaling up the load factor to an appropriate value. This feature is captured analytically in the following corollary and the subsequent remark.

Corollary 13. *In terms of $\beta_1 = \frac{1-\phi(\mu)}{\phi(\mu)}$ defined in Eq. (9), we can upper bound the difference between the optimal revenue rate and the maximum revenue rate under uniform pricing as*

$$R(K, \mathbf{p}^*) - R(K, p^*_K \mathbf{1}) \leqslant \frac{1}{\beta_1 K} R(K, p^*_K \mathbf{1}).$$

Proof. The blocking probability of K server system given in Proposition 6 under uniform price p^*_∞, can be upper bounded as

$$\pi_K(p^*_\infty) = \frac{1}{\sum_{j=0}^{K} \binom{K}{j} \overline{G}(p^*_\infty)^{-j} \beta_j} \leqslant \frac{1}{1 + K\overline{G}(p^*_\infty)^{-1}\beta_1}.$$

The upper bound follows by taking only two positive terms corresponding to $j \in \{0, 1\}$ in the summation for $j \in \mathcal{X}$. Therefore, using the fact that $\overline{G}(p_\infty) \leqslant 1$, we get

$$\frac{1}{1 - \pi_K(p^*_\infty)} \leqslant 1 + \frac{1}{K\overline{G}(p^*_\infty)^{-1}\beta_1} \leqslant 1 + \frac{1}{\beta_1 K}.$$

We obtain the result by substituting this expression in the upper bound for optimal revenue rate $R(K, \mathbf{p}^*)$ in Lemma 12. □

Remark 6. For Poisson arrivals, $\beta_1 = \frac{1-\phi(\mu)}{\phi(\mu)} = \frac{\mu}{\lambda} = \frac{1}{\rho}$, and hence $R(K, \mathbf{p}^*) \leqslant (1 + \frac{\rho}{K})R(K, p^*_K \mathbf{1})$. It is clear that for a large enough K, the optimal uniform price is a reasonable substitute for the optimal price. However, the bound is loose for smaller values of K and higher values of ρ, corresponding to a high arrival rate.

4.2. Asymptotic behavior of revenue rate

We next address the question of maximum revenue rate scaling under uniform pricing as the arrival rate increases to infinity. For a K *i.i.d.* server system each serving at an exponential rate μ, the maximum system service rate is $K\mu$. Therefore, for a uniform price system with $\mathbf{p} = p\mathbf{1}$, the maximum revenue cannot exceed $pK\mu$. We investigate whether

we can meet this upper bound by driving the arrival rate to infinity. We observe that this is not true for all arrival distributions. In fact, for certain interarrival distributions, the revenue rate goes to zero as arrival rate increases.

Recall that $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ denotes the Laplace Stieltjes transform of the *i.i.d.* job interarrival times. To begin with, we prove the following technical lemma.

Lemma 14. *For the K-server system under uniform pricing* $\lim_{\lambda\to\infty} \phi(\theta) = 1$.

Proof. For any $\theta \in \mathbb{R}_+$, we have $e^{-\theta U_1} \leqslant 1$. Further, we observe that $f : \mathbb{R}_+ \to \mathbb{R}_+$ defined by $f(y) \triangleq e^{-\theta y}$ is a convex function. From Jensen's inequality, we have $\mathbb{E}f(U_1) \geqslant f(\mathbb{E}U_1)$. Combining both these results, we get

$$e^{-\frac{\theta}{\lambda}} = e^{-\theta \mathbb{E}U_1} \leqslant \phi(\theta) \leqslant 1, \quad \theta \in \mathbb{R}_+. \tag{13}$$

Taking the limit as arrival rate $\lambda \to \infty$, we get the result. □

Remark 7. Consider the case when $\lim_{\lambda\to\infty} \lambda(1-\phi(\mu)) = \tilde{\mu}$ exists. Then, from the definition of sequence $\beta = (\beta_j = \prod_{m=1}^{j} \frac{1-\phi(m\mu)}{\phi(\mu)} : j \in \mathbb{N})$ in Eq. (9), we get that

$$\lim_{\lambda\to\infty} \beta_j = 0, \qquad \lim_{\lambda\to\infty} \lambda\beta_j = \tilde{\mu}\mathbb{1}_{\{j=1\}}. \tag{14}$$

Theorem 15. *Consider a K server pricing system with job interarrival times being* i.i.d. *and having a Laplace Stieltjes transform* ϕ *that satisfies* $\lim_{\lambda\to\infty} \lambda(1-\phi(\mu)) = \tilde{\mu}$. *The mean revenue rate for this system under a uniform price vector* $\mathbf{p} = p\mathbf{1}$ *such that* $\overline{G}(p) > 0$, *is bounded as the arrival rate grows. In particular,* $\lim_{\lambda\to\infty} R(K, p\mathbf{1}) = \tilde{\mu}pK$.

Proof. Recall that the mean revenue rate for uniform pricing $p\mathbf{1}$ of K *i.i.d.* exponential servers is given by $R(K, p\mathbf{1}) = \lambda p\overline{G}(p)(1-\pi_K(p))$ from Eq. (8). From Proposition 6, we have the blocking probability $\pi_K(p)$ in Eq. (10) defined in terms of variables $\beta_j = \prod_{m=1}^{j} \frac{1-\phi(m\mu)}{\phi(m\mu)}$ given in Eq. (9) for $j \in [K]$. Therefore, we can write the mean revenue rate as

$$R(K, \mathbf{p}) = p\overline{G}(p) \left(\frac{\frac{K\lambda\beta_1}{\overline{G}(p)} + \lambda \sum_{j=2}^{K} \binom{K}{j} \overline{G}(p)^{-j} \beta_j}{\sum_{j=0}^{K} \binom{K}{j} \overline{G}(p)^{-j} \beta_j} \right). \tag{15}$$

Taking the limit as arrival rate $\lambda \to \infty$, substituting the limiting results from Eq. (14) in Eq. (15), we obtain the result. □

Remark 8. From the inequality on Laplace Stieltjes transform ϕ in Eq. (13) and the fact that $1 - y \leqslant e^{-y}$, we get

$$0 \leqslant \lambda(1-\phi(x)) \leqslant \lambda(1-e^{-\frac{x}{\lambda}}) \leqslant x.$$

From the definition of $\tilde{\mu} = \lim_{\lambda\to\infty} \lambda(1-\phi(\mu))$, we obtain that $0 \leqslant \tilde{\mu} \leqslant \mu$. Thus, depending on the inter arrival time distribution, the limiting revenue can lie between 0 and μpK. The quantity $\tilde{\mu}$ can be considered an asymptotic service rate per server.

We present examples of limiting revenue rate being μpK, zero, and between $(0, \mu pK)$ in Examples 24, 25, and 26 respectively, in Appendix A.1. We see that with the fixed uniform pricing, the limiting mean revenue rate remains bounded, even when the arrival rate λ increases unboundedly large. We next show that it is indeed possible to scale the mean revenue rate with the arrival rate, at least in the limiting regime, if the uniform pricing scales with the job arrival rate λ.

Lemma 16. *Consider a K server uniform pricing system with* i.i.d. *job interarrival times having Laplace Stieltjes transform* $\phi : \mathbb{R}_+ \to \mathbb{R}_+$. *If the limit* $\tilde{\mu} \triangleq \lim_{\lambda\to\infty} \lambda(1-\phi(\mu)) > 0$, *the value distribution G has the support* $\mathbb{R}_+$, *and the uniform price* $p \in \overline{G}^{-1}(\frac{1}{\lambda})$, *then the limiting revenue rate* $\lim_{\lambda\to\infty} R(K, p\mathbf{1}) = \infty$.

Proof. Let $p \in \overline{G}^{-1}(1/\lambda)$. Substituting $\overline{G}(p) = \frac{1}{\lambda}$ in the mean revenue rate in Eq. (8) for K server system under uniform pricing, we obtain $R(K, p\mathbf{1}) = p(1-\pi_K(p))$. From the blocking probability $\pi_K(p)$ expression in Eq. (10) in terms of the positive sequence $\beta = (\beta_j : j \in [K])$ in Eq. (9), we get

$$\pi_K(p) = \frac{1}{\sum_{j=0}^{K} \binom{K}{j} \overline{G}(p)^{-j} \beta_j} \leqslant \frac{1}{1 + K\overline{G}(p)^{-1}\beta_1},$$

Recall that $\phi(\mu) \leqslant 1$, and hence $\beta_1 = \frac{(1-\phi(\mu)}{\phi(\mu)} \geqslant 1 - \phi(\mu)$. Using this fact and substituting $\overline{G}(p) = 1/\lambda$ in the above equation, we get $\pi_K(p) \leqslant (1 + K\lambda(1-\phi(\mu)))^{-1}$. From the hypothesis $\lim_{\lambda\to\infty} \lambda(1-\phi(\mu)) = \tilde{\mu} > 0$, and the fact that

$\lim_{x\to 0}\overline{G}^{-1}(x)=\infty$,[5] we obtain

$$\lim_{\lambda\to\infty} R(K,p\mathbf{1}) = \lim_{\lambda\to 0}\overline{G}^{-1}\left(\frac{1}{\lambda}\right)\frac{\tilde{\mu}}{1+\tilde{\mu}} = \infty. \quad \square$$

Thus, an arrival rate dependent uniform pricing can scale the revenue rate to infinity, in the asymptotic regime as the arrival rate λ grows arbitrarily large. We show an example of linear increase of mean revenue rate with arrival rate λ in Example 27. Since $\overline{G}^{-1}(\frac{1}{\lambda})\to\infty$ as λ increases, we see that to extract maximum revenue, the price should be made as high as possible in the heavy traffic limit. However, letting the price grow too fast can cause the revenue rate to go to zero instead of infinity, as shown in Example 28.

5. Optimal pricing for finite servers with Poisson arrivals

In the previous section, we found the optimal uniform pricing for a finite server system. Uniform pricing is optimal when the number of servers is very large. However, this yields a sub optimal revenue rate when the number of servers is finite. From Remark 6, it seems that uniform pricing would be sub optimal in a system with few servers or with high load, i.e, arrival rate much higher than service rate. In order to compute the revenue maximizing price, we frame the optimal state dependent pricing problem as a continuous time Markov decision problem [41, Chapter 5]. We derive optimal prices and also analyze their dependence on various system parameters, e.g., the number of servers, job arrival rate, and service rate. We first formulate the MDP for the case of Poisson arrivals, and solve it. In the subsequent section, we solve the MDP for a system with general arrivals. These are dealt with separately because the formulation changes when we move from Poisson to general arrivals. Furthermore, under the Poisson assumption we are able to obtain more insights into the system behavior.

5.1. The MDP formulation

As in Section 2, we consider the number of busy servers to be the state of the system and the quoted price in any state to be the control. Correspondingly, the state space is $\mathcal{X}'$ and the control space for price $u\in\mathbb{R}_+^{\mathcal{X}'}$. The mean revenue rate given a stationary state dependent policy u is,

$$R(K,u) = \lim_{t\to\infty}\frac{\mathbb{E}R(t)}{t} = \lim_{t\to\infty}\frac{1}{t}\mathbb{E}[\int_0^t g(X_s,u(X_s))ds], \tag{16}$$

where g is the instantaneous reward. In our setup, the rewards are obtained only at the arrival instants, and equals the price u if accepted by the incoming arrival. However, the price u is changed at every transition instant. Denoting X_n as the state of the system after n transitions, we can rewrite the reward rate as

$$R(K,u) = \lim_{t\to\infty}\frac{1}{t}\mathbb{E}\sum_{n=0}^{N_t} g(X_n,u(X_n)).$$

Following the discussion in [41], this is equivalent to

$$R(K,u) = \lim_{N\to\infty}\frac{1}{\mathbb{E}t_N}\mathbb{E}\sum_{n=1}^{N} g(X_n,u(X_n)),$$

where t_N is the Nth transition epoch. We wish to find the control $u^*\in\mathbb{R}_+^{\mathcal{X}'}$ that yields the optimal reward rate $R_K^*=R(K,u^*)=\max_u R(K,u)$.

The sojourn times in various states are independent exponentially distributed random variables depending on the controls applied on transitions to those states. As soon as the state changes to state i, a price u is set. This price is accepted with probability $\overline{G}(u)$ by an incoming arrival. Therefore, the sojourn times in a state i, for price u, are exponentially distributed with parameters $\nu_i(u)=i\mu+\lambda\overline{G}(u)\mathbb{1}_{\{i\in\mathcal{X}'\}}$. The state transition probabilities are independent of the sojourn times and dependent on the price $u\in\mathbb{R}_+$, and are given by: $p_{0,1}(u)=1$ and $p_{K,K-1}(u)=1$, and for $i\in[K-1]$

$$p_{ij}(u) = \frac{\lambda\overline{G}(u)}{\nu_i(u)}\mathbb{1}_{\{j=i+1\}} + \frac{i\mu}{\nu_i(u)}\mathbb{1}_{\{j=i-1\}}. \tag{17}$$

In addition, the rewards are accrued at the state transition instants and hence we focus only on the embedded discrete time Markov chain. The duration between two transitions is referred to as a stage of the MDP. When in a state i and using control u, a single stage reward u is obtained if a job arrives and joins service leading to the state $i+1$. The mean single stage reward is

$$g(i,u) = up_{i,i+1}(u) = u\mathbb{1}_{\{i=0\}} + \frac{\lambda u\overline{G}(u)}{\nu_i(u)}\mathbb{1}_{\{i\in[K-1]\}}. \tag{18}$$

[5] From the definition of distribution functions, the complimentary distribution $\overline{G}$ is non-increasing and $\lim_{x\to\infty}\overline{G}(x)=0$. Further, since the support of G is $\mathbb{R}_+$, it follows that $\lim_{x\to 0}\overline{G}^{-1}(x)=\infty$.

5.2. Uniformization of continuous time Markov chain

Using [41, Proposition 5.3.1] to solve the average reward MDP in Eq. (16) we can write the Bellman's equation for all states i

$$h(i) = \max_u \left[g(i,u) - \frac{\theta}{\nu_i(u)} + \sum_{j=0}^{K} p_{ij}(u) h(j) \right]. \tag{19}$$

Here θ is the optimal average reward per stage independent of the initial state (see [41, Section 4.1]) and $h(i)$, has interpretation of a relative or differential reward for each state i. Defining the uniformizing transition rate $\Lambda \triangleq K\mu + \lambda$, we observe that $\nu_i < \Lambda$ for all states i and control $u \in \mathbb{R}_+$. Hence we can convert the above Markov controlled process to the one with uniform transition rate Λ by allowing fictitious self transitions such that the resulting dynamics remains unchanged. Specifically, we redefine state transition probabilities for the uniformized Markov process as follows. For all states $i \in \mathcal{X}$ and control $u \in R_+$,

$$\tilde{p}_{ij}(u) = p_{ij}(u) \frac{\nu_i(u)}{\Lambda} \mathbb{1}_{\{j \neq i\}} + \left(1 - \frac{\nu_i(u)}{\Lambda}\right) \mathbb{1}_{\{j=i\}}. \tag{20}$$

We can now view the above problem as a discrete-time average reward problem with same state and control spaces, transition probabilities $\tilde{p}_{ij}(u)$ and expected single stage rewards $g(i,u)$. The Bellman's equation for this discrete-time problem has the following form for all i

$$\tilde{h}(i) = \max_u \left[g(i,u)\nu_i(u) - \theta + \sum_{j=0}^{K} \tilde{p}_{ij}(u) \tilde{h}(j) \right]. \tag{21}$$

Remark 9. The Bellman's equations (19) and (21) are equivalent. In particular, a pair (θ, h) satisfies (19) if and only if the pair $(\theta, \tilde{h})$ satisfies (21), where $\tilde{h}(i) = \Lambda h(i)$ for all i. Moreover, for all the states, the optimal actions for the two problems (control u achieving maxima in the right hand sides of (19) and (21)) are identical.

Remark 10. Defining the uniformized reward difference $\Delta(i) \triangleq \frac{(\tilde{h}(i) - \tilde{h}(i+1))}{\Lambda}$ for all states $i \in \mathcal{X}'$, and substituting in Eq. (21), along with expressions for per stage mean reward $g(i,u)$ from Eq. (18), and transition probabilities $\tilde{p}_{ij}(u)$ from Eq. (20), we get the following set of equations for all $i \in \mathcal{X}$

$$\theta = \lambda \max_u \left\{ \overline{G}(u)(u - \Delta(i)) \right\} \mathbb{1}_{\{i \in \mathcal{X}'\}} + i\mu\Delta(i-1). \tag{22}$$

5.3. Auxiliary maps

We define the mapping $f : \mathbb{R}^2 \to \mathbb{R}$ as

$$f(B,u) \triangleq (u - B)\overline{G}(u), \ \ B, u \in \mathbb{R}. \tag{23}$$

Remark 11. We define a set valued map u^* that maps $B \in \mathbb{R}$ to $u^*(B) \subseteq \mathbb{R}_+$

$$u^*(B) \triangleq \arg\max_u f(B,u). \tag{24}$$

If the maximizer is unique, then $u^* : \mathbb{R} \to \mathbb{R}_+$ is a real valued map. The maximum value of $f(B, u^*)$ is a real valued map $m : \mathbb{R} \to \mathbb{R}$ such that

$$m(B) \triangleq f(B, u^*) = \max_u f(B,u). \tag{25}$$

Lemma 17. *Following statements are true for m and u^*.*

(a) m is non-negative and decreasing in B.
(b) m is Lipschitz-1 continuous and convex function of B.
(c) For $B_1 < B_2$, we have $\sup u^*(B_1) \leqslant \inf u^*(B_2)$. When $f(B,u)$ has a unique maximizer in u, then this maximizer u^* is non-decreasing in B.

Proof. Proof is in Appendix B.1. □

5.4. The optimal pricing

In terms of the map m, we can re-write Eq. (22) as

$$m(\Delta(i))\mathbb{1}_{\{i \in \mathcal{X}'\}} + \frac{i\mu}{\lambda}\Delta(i-1) = \frac{\theta}{\lambda}, \ \ i \in \mathcal{X}. \tag{26}$$

Observe that if $(\Delta^*(i) : i \in \mathcal{X}')$ solves Eq. (26) then the control $u_i^* \triangleq u^*(\Delta^*(i))$ achieving $m(\Delta^*(i))$ in Eq. (25) is the optimal control in each state $i \in \mathcal{X}'$.

Remark 12. Consider the limiting case of infinitely many servers, i.e., $K = \infty$. We can easily see that $\theta = \lambda m(0)$ along with $\Delta(i) = 0$ for all $i \in \mathbb{Z}_+$ is a solution to Eq. (26). In particular, uniform (state independent) pricing, $u^* = \arg\max_{u \geq 0} u\bar{g}(u)$, achieves the optimal revenue rate as readily seen in Eq. (12).

Lemma 18. *Let $(\theta, \Delta(i), i \in \mathcal{X}')$ be a solution to Eq. (26) and $\mathbf{p}_K^* = (u_0^*, \ldots, u_{K-1}^*) \in \mathbb{R}_+^{\mathcal{X}'}$ be the optimal price vector. Then*

(a) $\theta \geqslant 0$,
(b) $\Delta(i)$ are positive and increasing in $i \in \mathcal{X}'$.
(c) u_i^* are also increasing in $i \in \mathcal{X}'$.

Proof. Proof is in Appendix B.2. □

Next, we will focus on solving Eq. (26). We propose an iterative algorithm to obtain θ, which can then be used to obtain $\Delta(i)$ and also the optimal price u_i^* for all the states. Realizing that $\Delta(i)$ is a function of optimal revenue θ and state i, we denote it as $g_i(\theta) \triangleq \Delta(i)$, to rewrite Eq. (26) as

$$\theta = \lambda m(g_0(\theta)), \tag{27a}$$

$$g_{i-1}(\theta) = \frac{\theta - \lambda m(g_i(\theta))}{i\mu}, \ i \in [K-1], \tag{27b}$$

$$g_{K-1}(\theta) = \frac{\theta}{K\mu}. \tag{27c}$$

We will show that there exists a unique θ which solves Eq. (27a). We then propose Algorithm 1 that finds this unique θ in terms of which the optimal prices can be found. In particular, this algorithm iteratively generates two sequences $(\underline{\theta}_k : k \in \mathcal{X})$ and $(\bar{\theta}_k : k \in \mathcal{X})$ which converge to the unique θ.

Algorithm 1

initialize $k = 0, \underline{\theta}_0 = 0, \bar{\theta}_0 = \lambda m(g_0(0))$,
while $\bar{\theta}_k - \underline{\theta}_k > \delta$ **do** ▷ δ is the desired precision.
  $\tilde{\theta}_k = \frac{\underline{\theta}_k + \bar{\theta}_k}{2}$,
  $\underline{\theta}_{k+1} = \max\left\{\underline{\theta}_k, \min\{\tilde{\theta}_k, \lambda m(g_0(\tilde{\theta}_k))\}\right\}$,
  $\bar{\theta}_{k+1} = \min\left\{\bar{\theta}_k, \max\{\tilde{\theta}_k, \lambda m(g_0(\tilde{\theta}_k))\}\right\}$,
  $k = k + 1$

Theorem 19.

(a) The fixed point equation $\theta = \lambda m(g_0(\theta))$ has unique solution.
(b) In Algorithm 1, $\underline{\theta}_k \uparrow \theta^*$ and $\underline{\theta}_k \downarrow \theta^*$, where θ^* is the unique fixed point.

Proof. We consider Eqs. (27a), (27b), (27c).

(a) Observe that $\lambda m(g_0(0)) > 0$. We now argue that $\lambda m(g_0(\theta))$ is decreasing in θ. These two facts together yield both existence and uniqueness. From the monotonicity of function m in Lemma 17(a) and definition of g_{i-1} from Eq. (27b), it follows that g_{i-1} is increasing in θ if g_i is increasing in θ. Since $g_{K-1}(\theta) = \theta/K\mu$ is increasing in θ, it follows that $g_0(\theta)$ is increasing in θ, and hence $\lambda m(g_0(\theta))$ is decreasing in θ.
(b) See [42, Theorem 2.1]. □

Remark 13. If we assume the valuation is exponentially distributed, i.e., $\bar{G}(x) = e^{-\beta x}$, we see that the mapping $u^*(B) = B + \frac{1}{\beta}$ and $m(B) = \frac{1}{\beta}e^{-(\beta B+1)}$. The optimal prices will be $u_i^* = u^*(\Delta^*(i)) = \Delta^*(i) + \frac{1}{\beta}$.

5.5. Properties of the optimal solution

We now analyze how the optimal prices and the optimal revenue rate vary with arrival rate λ, service rate μ, and number of servers K. We use the fact that the optimal revenue rate θ^* is solution to Eq. (27a), from which we inductively

derive properties of the uniformized reward difference g_i using the monotonic decrease of m from Lemma 17. First, we look at the variation of the optimal revenue rate with arrival rate λ.

5.5.1. Varying arrival rate

We assume that we vary the arrival rate λ while keeping the service rate μ and number of servers K fixed.

Proposition 20. *For a K server pricing system with a fixed service rate μ, the following statements are true.*

(a) The optimal revenue rate $\theta^*(\lambda)$ increases with λ.
(b) The ratio $\theta^*(\lambda)/\lambda$ decreases with λ.

Proof. Proof is in Appendix B.3. □

Remark 14. The uniformized reward differences $\Delta(0)$ and $\Delta(K-1)$ are increasing in the arrival rate λ from Eqs. (27a) and (27c), respectively. Consequently, the optimal prices u_0^* and u_{K-1}^* are also increasing in λ. We believe that all the optimal prices $(u_i^*, i \in \mathcal{X}')$ are increasing in λ. While we have not been able to show this, we demonstrate it via numerical results in Section 7.

Next, we study how the optimal revenue rate varies as the service rate μ changes.

5.5.2. Varying service rate

Here we assume that we vary the service rate μ while keeping the arrival rate λ and number of servers K fixed. Now we express the revenue rate as $\theta^*(\mu)$ to emphasize its dependence on μ.

Proposition 21. *For a K-server pricing system with a fixed arrival rate λ, the following statements are true.*

(a) The revenue rate $\theta^*(\mu)$ increases with μ.
(b) The ratio $\theta^*(\mu)/\mu$ decreases with μ.

Proof. Proof is in Appendix B.4. □

Remark 15. Contrary to the observation in Remark 14, the uniformized reward differences $\Delta(0)$ and $\Delta(K-1)$ are decreasing in the service rate μ from Eqs. (27a) and (27c), respectively. Hence, the optimal prices u_0^* and u_{K-1}^* are decreasing in μ. We believe that all the optimal prices $(u_i^*, i \in \mathcal{X}')$ are decreasing in μ. While we have not been able to show this, we demonstrate it via numerical results in Section 7.

Finally, we study the variation of optimal revenue rate with number of servers.

5.5.3. Increasing number of servers

We assume that we vary number of servers K while keeping arrival rate λ and service rate μ fixed. Now we express the revenue rate as $\theta^*(K)$.

Proposition 22. *For a pricing system with a fixed arrival rate λ and a fixed service rate μ, the following statements are true.*

(a) The revenue rate $\theta^*(K)$ increases with K.
(b) The ratio $\theta^*(K)/K$ decreases with K.
(c) For any $i < K$, the optimal price $u_i^*(K)$ is non-increasing with K.

Proof. Proof is in Appendix B.5. □

General service times

We make an interesting observation on multiple server systems with Poisson job arrivals with rate λ and general *i.i.d.* job service times with distribution $F : \mathbb{R}_+ \to [0, 1]$ and mean $\frac{1}{\mu}$. Suppose we continue to use the optimal state dependent prices u_i^* for $i \in \mathcal{X}'$ busy servers seen by an incoming arrival, that was derived for the exponential service rate system in Section 5.4. This results in an $M/G/K/K$ system with state dependent arrivals rates $(\lambda_i \triangleq u_i^*\lambda : i \in \mathcal{X}')$. Following the insensitivity property [43, Section 8.10] of $M/G/K/K$ systems, the steady state distribution of the number of busy servers remains identical to the steady state distribution in the corresponding $M/M/K/K$ system. Moreover, the average reward rate in the $M/G/K/K$ system with state dependent prices $(u_i^* : i \in \mathcal{X}')$ will be same as the optimal average reward rate in the $M/M/K/K$ system. However, the optimal prices in the $M/G/K/K$ systems will in general be different from $(u_i^* : i \in \mathcal{X}')$. The optimal prices will depend on elapsed services times of busy servers on job arrival epochs. These optimal prices are not easy to determine following the techniques as used in this work. However, we make a non-trivial inference that the optimal average reward rate in an $M/G/K/K$ system always exceeds the optimal average reward rate in the corresponding $M/M/K/K$ system.

6. General arrival processes

In the previous section, we found the optimal pricing for a K server system with Poisson arrivals and exponential service rates. In this section, we extend the setting to K server systems with general interarrival time distribution. In particular, we assume the interarrival times $(U_n : n \in \mathbb{N})$ are *i.i.d.* with density $f : \mathbb{R}_+ \to \mathbb{R}_+$ and finite mean $1/\lambda$. We will continue to assume that the admission price is updated only at arrival instants, and hence this price depends only on the number of busy servers in the system. We assume that the price is infinite when all K servers are busy. As discussed in Section 3, the system state is modeled by the number of occupied servers seen by the arriving jobs, and the state space remains $\mathcal{X} = \{0, \ldots, K\}$, and the modified state space $\mathcal{X}' = \{0, \ldots, K_1\}$. Similarly, the control space for price remains $\mathbb{R}_+^{\mathcal{X}'}$, and we write the problem of finding optimal revenue rate as an MDP. In the Poisson arrival setting, the process $X = (X(t) : t \geqslant 0)$ sampled at all transition instants, remained Markov. In contrast, in the general arrival setting, the process X sampled only at the arrival instants, is Markov. Thus the sampled process $Z = (Z_n = X(A_n^-) : n \in \mathbb{N})$ is a controlled Markov chain. We modify the MDP in Section 5.1, to write the optimal revenue rate in the terms of sampled process Z, and the instantaneous reward at arrival instants $g(Z_n, u(Z_n)) = \mathbb{E}[u(Z_n)\mathbb{1}_{\{V > u(Z_n)\}}|Z_n] = u(Z_n)\bar{G}(Z_n)$, as

$$R(K, u) = \lim_{N\to\infty} \frac{1}{t_N}\mathbb{E}\sum_{n=1}^{N} g(Z_n, u(Z_n)).$$

The probability of $k - j$ departures from state k is given by $\alpha_{k,k-j}$ defined in Eq. (5). We recall the transition probability from state $k \in \mathcal{X}$ to state $j \in \{0, \ldots, \min\{k+1, K\}\}$ for the controlled Markov chain Z given in Eq. (6), with price p_k replaced by control map u is

$$p_{kj}(u) = \bar{G}(u)\alpha_{k+1,k+1-j} + G(u)\alpha_{k,k-j}.$$

Following similar steps as in Section 5.2, we use [41, Proposition 5.3.1] to solve the average reward MDP in the above equation. We can write the Bellman's equations for all states i

$$h(i) = \max_u \left[g(i, u) - \frac{\theta}{\lambda} + \sum_{j=0}^{K} p_{ij}(u)h(j) \right], \quad i \in \mathcal{X}'.$$

Note that the mean sojourn time $\frac{1}{\nu(i)} = \frac{1}{\lambda}$ for all states i, and θ is the optimal average reward per stage, independent of the initial state. Substituting the instantaneous reward $g(i, u) = u\bar{G}(u)$ at arrival instants, the transition probabilities for the sampled Markov chain Z in Eq. (6), and the probability distribution of number of departures between two arrival instants in Eq. (5), we get

$$\begin{aligned} h(i) = \max_u \Big[& u\bar{G}(u) - \frac{\theta}{\lambda} + \bar{G}(u)\sum_{j=0}^{i+1} \alpha_{i+1,i+1-j}h(j) \\ & + G(u)\sum_{j=0}^{i} \alpha_{i,i-j}h(j) \Big], \quad i \in \mathcal{X}'. \end{aligned} \tag{28}$$

When the number of busy servers is K, we get the boundary equation

$$h(K) = -\frac{\theta}{\lambda} + \sum_{j=0}^{K} \alpha_{K,K-j}h(j). \tag{29}$$

We first focus on states $i \in [K-1]$. To this end, we define the reward difference

$$\Delta(i) \triangleq h(i) - h(i+1), \quad i \in [K-1], \tag{30}$$

and the probability of more than $i - j$ departures from state i as

$$a_{i,j} \triangleq \sum_{l=0}^{j} \alpha_{i,i-l}, \quad i \in \mathcal{X}, j \leqslant i. \tag{31}$$

Rearranging the terms in Eq. (28), using the definition of sequences $(\Delta(i) : i \in [K-1])$ and $(a_{i,j} : j \leqslant i, i \in \mathcal{X})$, we get

$$\begin{aligned} & \max_u \left[\Big(u - \sum_{j=0}^{i-1}(a_{ij} - a_{i+1,j})\Delta(j) - \alpha_{i+1,0}\Delta(i)\Big)\bar{G}(u) \right] \\ & + \sum_{j=0}^{i-1} a_{ij}\Delta(j) = \frac{\theta}{\lambda}. \end{aligned}$$

Following similar steps for $i = K$ in Eq. (29), we get

$$\sum_{j=0}^{K-1} a_{K,j}\Delta(j) = \frac{\theta}{\lambda}.$$

For notational convenience, we define the following sequence

$$b_i \triangleq \sum_{j=0}^{i-1}(a_{i,j} - a_{i+1,j})\Delta(j) + \alpha_{i+1,0}\Delta(i), \quad i \in \mathcal{X}'. \tag{32}$$

From the definition of map $m(B) = \max_u(u - B)\overline{G}(u)$ defined in Eq. (25) and the definition of $(b_i : i \in \mathcal{X}')$ in Eq. (32), we can write the previous set of equations for the solution of average reward MDP as

$$m(b_0) = \frac{\theta}{\lambda}, \tag{33a}$$

$$m(b_i) + \sum_{j=0}^{i-1} a_{i,j}\Delta(j) = \frac{\theta}{\lambda}, i \in [K-1], \tag{33b}$$

$$\sum_{j=0}^{K-1} a_{K,j}\Delta(j) = \frac{\theta}{\lambda}. \tag{33c}$$

Theorem 23. *Let $(\theta, (\Delta(i) : i \in \mathcal{X}'))$ be a solution to Eqs. (33a)–(33c). Then, the following statements hold true.*

(a) The optimal revenue rate $\theta \geqslant 0$.
(b) The reward rate difference sequence $(\Delta(i), i \in \mathcal{X}')$ is positive and increasing in state i. The sequence $(b_i : i \in \mathcal{X}')$ is also positive and increasing in state i.
(c) The optimal price vector $(u_i^* : i \in \mathcal{X}')$ is increasing in state i.

Proof. Proof is in Appendix C. □

When the inverse map m^{-1} exists, we provide an inductive procedure to get a fixed point equation to obtain the optimal state dependent mean revenue rate θ. Given the optimal mean revenue rate θ, the reward difference $\Delta(i)$ and hence the optimal actions u_i^* can be obtained for all states $i \in \mathcal{X}'$. To show explicit dependence of the reward difference on the mean revenue rate θ, we denote the reward difference $\Delta(i) = g_i(\theta)$ for $i \in \mathcal{X}'$. Substituting this in Eqs. (33a)–(33b), we can inductively obtain

$$\begin{aligned} g_0(\theta) &= \frac{1}{\alpha_{1,0}} m^{-1}\left(\frac{\theta}{\lambda}\right), \\ g_i(\theta) &= \frac{1}{\alpha_{i+1,0}}\Bigg[m^{-1}\Big(\frac{\theta}{\lambda} - \sum_{j=0}^{i-1} a_{i,j}g_j(\theta)\Big) \\ &\quad - \sum_{j=0}^{i-1}(a_{i,j} - a_{i+1,j})g_j(\theta)\Bigg], \quad i \in \mathcal{X}' \setminus \{0\}. \end{aligned}$$

Finally, using the sequence of functions $(g_j(\theta) : j \in \mathcal{X}')$ to replace reward difference $(\Delta(j) : j \in \mathcal{X}')$ in Eq. (33c), we obtain the following fixed point equation

$$\theta = \lambda \sum_{j=0}^{K-1} a_{K,j}g_j(\theta). \tag{34}$$

We can solve the fixed point equation in Eq. (34) to obtain the optimal mean revenue rate θ and the optimal prices $(u_i^* : i \in \mathcal{X}')$.

7. Numerical evaluation

We first obtain the optimal price vectors for a 5 server system, for three different inter arrival time distributions. The arrival rate is $\lambda = 25$ and service rate $\mu = 2$ for all these systems, and the valuation distribution is $\overline{G}(p) = e^{-p}$. The optimal price as a function of the number of busy servers, for exponential, uniform and constant inter arrival time distributions, are plotted in Fig. 4. Note that the mean inter arrival time will be $\frac{1}{\lambda}$. All the prices are increasing in the number of busy

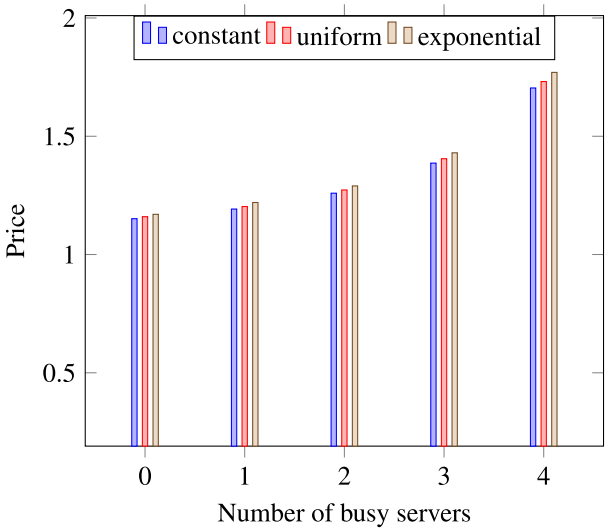
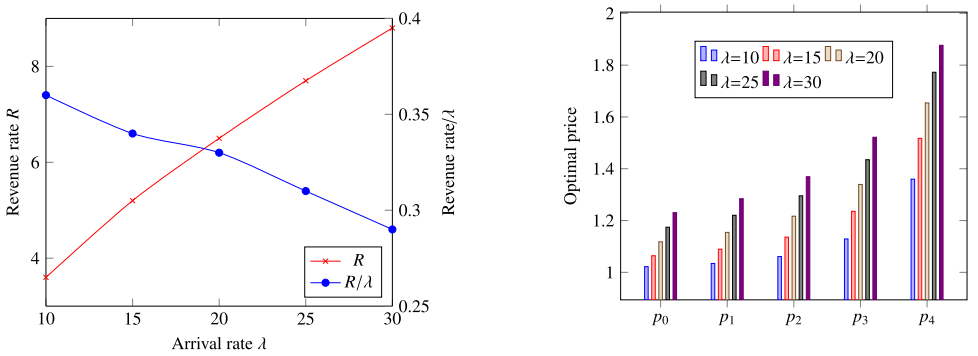


Fig. 4. Optimal price vectors for different inter arrival time distributions.



(a) Revenue rate and Revenue rate per arrival rate

(b) Optimal price vector

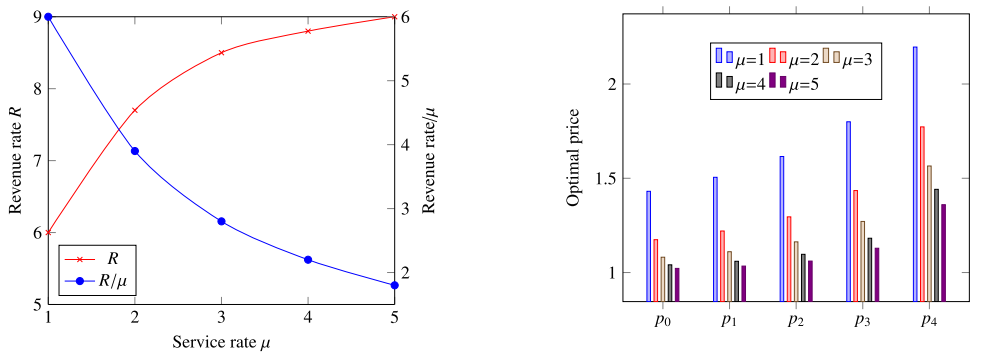
Fig. 5. Variation of optimal revenue rate, revenue rate per arrival rate and price with arrival rate.

servers, as shown before in Lemma 18(c) and Theorem 23(c). Also observe that the exponential inter arrival time attracts the highest price in any system state.

Next, we study the variation of the optimal revenue rate and optimal price with respect to arrival rate, service rate and number of servers. Consider a 5 server system. The service rate $\mu = 2$ and job valuations are distributed exponentially, with $\overline{G}(p) = e^{-p}$. In Fig. 5(a), we see that the optimal revenue increases monotonically as the arrival rate increases. This is expected, since a good pricing policy will be able to extract more revenue from increased demand. However, in Fig. 5(a), we also see that the revenue per unit arrival rate is actually decreasing, as we scale up the arrival rate. This implies that the rate at which revenue can be extracted per unit arrival rate is decreasing. Both these observations validate the results of Proposition 20. In Fig. 5(b), we have plotted the price vector for different arrival rates. As the arrival rate increases, the price vector increases in all its components. Recall that this was conjectured in Remark 14.

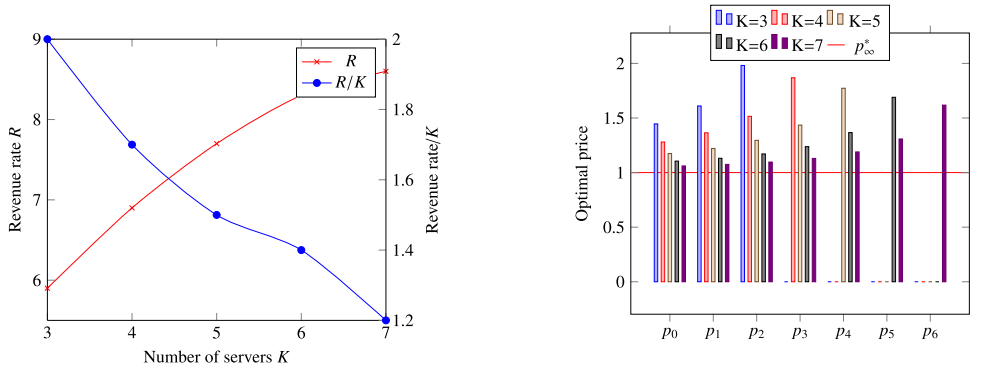
For studying the effect of service rate variation on optimal revenue rate, we again consider a 5 server system, with arrival rate $\lambda = 25$. As before, $\overline{G}(p) = e^{-p}$. In Fig. 6(a), we see that revenue scales monotonically with service rate. Thus, by increasing the service capacity, we can extract more revenue. The revenue per service rate, however, decreases as service rate increases, in Fig. 6(a). This implies that the marginal returns per unit service capacity decreases. These results are in line with Proposition 21. In Fig. 6(b), we see how the price vector decreases component wise as we increase the service rate, as expected in Remark 15.

For studying the relation between number of servers and optimal revenue/price, we consider a system with arrival rate $\lambda = 25$, service rate $\mu = 2$ and valuation distribution $\overline{G}(p) = e^{-p}$. In Fig. 7(a), we see how the optimal revenue and the optimal revenue per server vary, as we increase the number of servers. While we can extract more revenue as increase the number of servers, the revenue rate per server decreases. We also see how the price vector itself behaves, as we increase the number of servers, in Fig. 7(b). We see that the components of the price vector decrease and come closer



(a) Revenue rate and Revenue rate per service rate (b) Optimal price vector

Fig. 6. Variation of optimal revenue rate, revenue rate per service rate and price with service rate.



(a) Revenue rate and Revenue rate per server (b) Optimal price vector

Fig. 7. Variation of optimal revenue rate, revenue rate per service rate and price with number of servers.

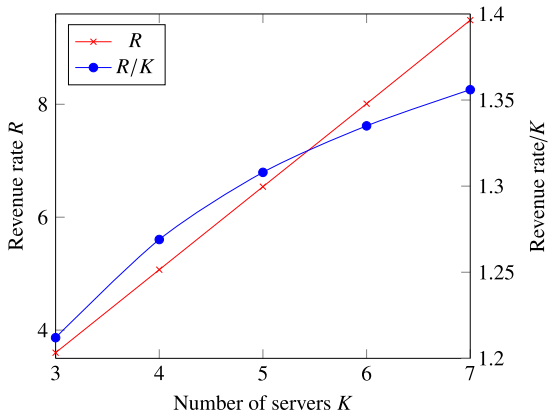
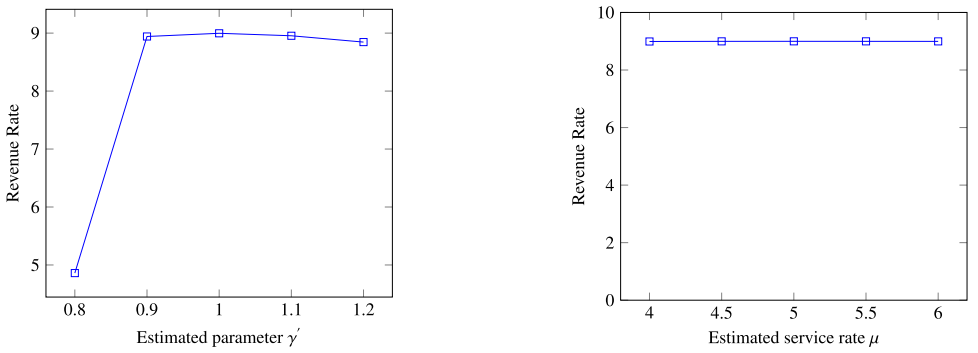


Fig. 8. Variation of optimal revenue rate and revenue rate per server, with number of servers, when arrival rate is also scaled as the number of servers increases.

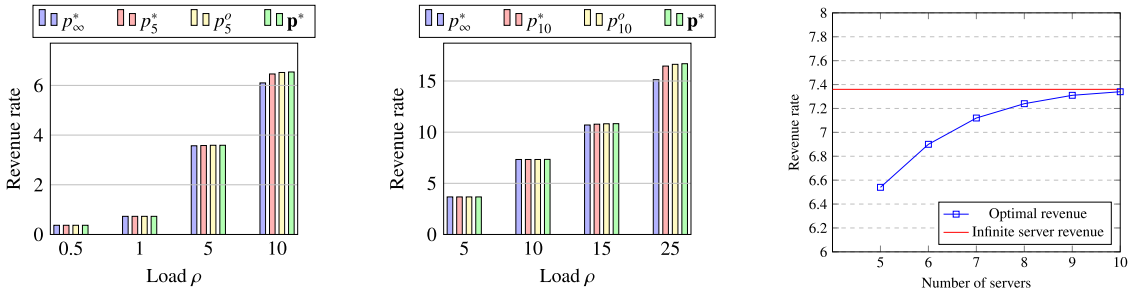
to the optimal infinite server price p^*_∞ (which equals 1 in this case), as we increase the number of servers. The trends are as predicted in Proposition 22. However, we have observed that if we increase both the number of servers and the arrival rate in the same ratio, then the revenue rate per server increases. Such an effect is shown in Fig. 8, where we look at the revenue rate as we vary number of servers as well as the arrival rate. The arrival rate is $\lambda = 4K$, and the service rate is $\mu = 2$.

To optimally price the multi-server system, the service provider would need to estimate the valuation and service distribution. We observe that an incorrect estimation of system parameters can negatively impact the revenue, and



(a) Incorrect estimation of valuation (b) Incorrect estimation of service rate

Fig. 9. Impact of incorrect estimation of system parameters on revenue rate.



(a) Revenue vs Load, 5 servers (b) Revenue vs Load, 10 servers (c) Optimal revenue vs number of servers

Fig. 10. Variation of optimal revenue and price with number of servers.

illustrate this through numerical examples. To understand the impact of incorrectly estimated valuation distribution on optimal revenue, we consider a 5 server system with Poisson arrivals of rate $\lambda = 25$ and exponential service of rate $\mu = 5$. The valuation distribution $\bar{G}(p) = e^{-\gamma p}$, where $\gamma = 1$. However, for the calculation of $\mathbf{p}^*$ by Algorithm 1, we assume that the parameter γ is incorrectly estimated as γ', leading to a different price vector. The revenue rates obtained by running Algorithm 1 with the incorrect parameter is plotted in Fig. 9(a). Note that the parameter value 1 is the true value, and hence the revenue rate obtained using any other γ' is lower than the revenue achieved with $\gamma' = 1$. It is clear that incorrect estimation of customer valuations may significantly degrade the revenue rate. In Fig. 9(b), we present the results of a similar study, this time assuming that the service rate has been incorrectly estimated. The true service rate is $\mu = 5$; however, we assume this has been estimated incorrectly. Unlike the previous case, we see that incorrect estimation of service rate does not have a major impact on the revenue rate.

We compare differential pricing and uniform pricing for a system with Poisson arrivals (or equivalently, exponential inter arrival time). We consider a 5-server system, with $\mu = 2$. For different values of load $\rho = \frac{\lambda}{\mu}$, we compare the revenue under the optimal price $\mathbf{p}^*$ with the revenue under uniform prices p^*_∞ and p^*_5, and $p_5{}^\circ$, which is the optimal step price (i.e., the optimal among prices of the form $\mathbf{p} = (p, p, \ldots, p, q, q..q)$, which is a generalization of the uniform price). The valuation function $\bar{G}(p) = e^{-p}$. The resultant values are displayed in Fig. 10(a).

At low values of arrival rates, differential pricing does not offer substantial gains over uniform pricing. At higher arrival rates, however, we begin to see that revenue rates show a significant improvement using differential pricing. One can also see that these effects are more pronounced beyond $\rho = 5$, the number of servers. The step price $p_5{}^\circ$ performs better than the uniform price, but is still sub optimal for high load values. A similar effect is seen in the case of 10 servers as well, as seen in Fig. 10(b) (all other parameters remaining same). Beyond $\rho = 10$, differential pricing begins to outperform uniform pricing and step pricing.

In Fig. 10(c), we also study how quickly the optimal differential revenue for a finite server system converges to the optimal revenue with infinite servers, for a system with Poisson arrivals. We fix $\lambda = 20$ and $\mu = 2$ and $\bar{G}(p) = e^{-p}$. It is clear that the infinite server optimal revenue, $R(\infty, p^*_\infty \mathbf{1}) = 7.36$. With as few as 10 servers, we come close to the infinite server revenue. With these many servers, optimal pricing can be closely approximated by optimal uniform pricing. Recall from Remark 6 that with $K = 10$, the optimal revenue can exceed the optimal uniform revenue by at most 5%.

8. Conclusion

We studied optimal service pricing in server farms where customers arrive according to a renewal process and have *i.i.d.* exponential service times and *i.i.d.* valuations of the service. We showed that fixed pricing achieves optimal revenue rate in infinite server systems but can guarantee only close to optimal revenue rate in finite server systems. However, fixed pricing suffices to drive revenue rate to infinity in infinite server systems as the arrival rates increase. We also showed that the optimal prices for finite server systems increase with the number of busy servers. In case of exponential interarrival times, we derived several properties of the optimal prices vis a vis arrival rates, service rates, and the number of servers in the system.

We argued that for a given service rate, the optimal revenue rates in server farms with non-exponential service times generally exceed those in server farms with exponential service times. But, the optimal prices in the former systems depend on the elapsed service times in the busy servers and cannot be obtained using the Markov control framework as we have done in this work. Similarly, optimal pricing for processor sharing systems and systems with queues where customers can wait for service are also challenging problems. These are potential topics for future research.

In addition, we observed via a numerical example, that an incorrect estimation of the valuation distribution can negatively impact the revenue. Theoretical characterization of the sensitivity of the revenue rate to the estimation error in valuation distribution is also an interesting direction for future research. Another possible extension is to consider jobs belonging to multiple classes, each with different service requirements and valuations. The service provider may choose to prioritize some jobs over others, and pool many servers together to serve high priority jobs. This is an open and complex scheduling problem.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Examples for asymptotic revenue rates

A.1. Fixed uniform pricing

We first present an example where the limiting revenue rate is μpK.

Example 24. Consider the Poisson arrival process for jobs, with rate λ. Then, the interarrival times are exponential, and the Laplace Stieltjes transform $\phi(\mu) = \frac{\lambda}{\lambda+\mu}$. We can write the limit $\lim_{\lambda\to\infty} \lambda(1-\phi(\mu)) = \lim_{\lambda\to\infty} \mu \frac{\lambda}{\lambda+\mu} = \mu$. It follows that $\tilde{\mu} = \mu$ in Theorem 15, and hence the limiting mean revenue rate $\lim_{\lambda\to\infty} R(K, p\mathbf{1}) = \mu K p$ for K server system under uniform price $p\mathbf{1}$.

We next present an example of an interarrival distribution for which the limiting revenue rate goes to 0.

Example 25. For some $m > 1$, we consider the job interarrival times $(U_n \in \left\{\sqrt{m}, \frac{1}{m}\right\} : n \in \mathbb{N})$ such that $P\left\{U_1 = \sqrt{m}\right\} = 1/m$. In this case the arrival rate $\lambda = 1/\mathbb{E}U_1$ where the mean interarrival time $\mathbb{E}U_1 = \frac{1}{\sqrt{m}} + \frac{1}{m}(1 - \frac{1}{m})$. It follows that $m \to \infty$ implies $\lambda \to \infty$. We next compute the Laplace Stieltjes transform $\phi(x) = \mathbb{E}e^{-xU_1}$ of interarrival times as

$$\phi(x) = (1 - \frac{1}{m})e^{-\frac{x}{m}} + \frac{e^{-x\sqrt{m}}}{m}.$$

Using the fact that $\lambda = 1/\mathbb{E}U_1$, we can write the limit

$$\lim_{m\to\infty} \lambda(1-\phi(x)) = \lim_{m\to\infty} \frac{\frac{m^2}{m-1} - me^{-\frac{x}{m}} - \frac{m}{m-1}e^{-x\sqrt{m}}}{1 + \frac{m\sqrt{m}}{m-1}} = 0.$$

For this interarrival distribution, it follows from Theorem 15 that the limiting mean revenue rate is zero for any uniform price p. One would expect the mean revenue rate to increase with the arrival rate, since the mean interarrival time decreases. However, for this example distribution, the mean revenue rate instead of increasing with the arrival rate, goes to zero. Such a behavior arises due to slow decay of the tail of the interarrival time distribution.

There are distributions for which the limiting revenue is non zero but strictly less than μpK, as in the following example.

Example 26. For $m > 1$, consider the *i.i.d.* job interarrival times $(U_n \in \{1, \frac{1}{m}\} : n \in \mathbb{N})$ with $P\{U_1 = 1\} = \frac{1}{m}$. That is, the mean interarrival time $1/\lambda = \mathbb{E}U_1 = \frac{2m-1}{m^2}$ and the Laplace Stieltjes transform $\phi(x) = \mathbb{E}e^{-xU_1} = \frac{1}{m}e^{-x} + (1 - \frac{1}{m})e^{-\frac{x}{m}}$ for $x \in \mathbb{R}_+$. It follows that $\lambda \to \infty$ as m grows large, and the limit

$$\tilde{\mu} = \lim_{m\to\infty} \lambda(1 - \phi(\mu)) = \frac{1 - e^{-\mu}}{2}.$$

It follows from Theorem 15 that the limiting revenue rate is $\lim_{\lambda\to\infty} R(K, p\mathbf{1}) = \dfrac{1 - e^{-\mu}}{2} pK$. Since $1 - e^{-\mu} \leqslant \mu$, the limiting revenue rate is smaller than $\dfrac{\mu pK}{2}$.

A.2. Arrival rate dependent uniform pricing

Example 27. If the arrival process is Poisson and the value distribution is Pareto, i.e. $\overline{G}(x) = \frac{\theta}{x}\mathbb{1}_{\{x\geqslant\theta\}}$, then the choice of uniform price $p(\lambda) = \overline{G}^{-1}(\frac{1}{\lambda}) = \lambda\theta$ that grows linearly with the arrival rate λ. Further, we have the Laplace Stieltjes transform or *i.i.d.* interarrival times $\phi(\mu) = \frac{\lambda}{\lambda+\mu}$, and hence $\tilde{\nu} = \lim_{\lambda\to\infty} \lambda(1 - \phi(\mu)) = \mu > 0$. Thus results in the mean revenue rate $R(K, p\mathbf{1})$ are asymptotically linearly increasing in the arrival rate λ.

Example 28. Let the arrival process be Poisson with rate λ and the complimentary value distribution $\overline{G}(x) = c_1 e^{-c_2 x^2}$. Since the arrival process is Poisson, $\tilde{\mu} = \mu$. Choosing the uniform price $p(\lambda) = \sqrt{\frac{1}{c_2}\log(c_1\lambda)}$, we see that $\lim_{\lambda\to\infty} R(K, p\mathbf{1}) = \infty$. Contrastingly, for a uniform price $p(\lambda) = \log\lambda$, we get $\lim_{\lambda\to\infty} R(K, p\mathbf{1}) = 0$.

Appendix B. Proofs for MDP with Poisson arrival process

B.1. Proof of Lemma 17

Let m and u^* be as defined in Eq. (25) and Eq. (24) respectively. For $B_1, B_2 \in \mathbb{R}$, we let $u_i \in u^*(B_i)$ for $i \in \{1, 2\}$. From the definition of f, we have $f(B_2, u_2) - f(B_1, u_2) = -(B_2 - B_1)\overline{G}(u_2)$. In addition, we have $f(B_1, u_1) \geqslant f(B_1, u_2)$ from the definition of m, u_1, u_2. Therefore, we can lower bound the difference

$$m(B_1) - m(B_2) \geqslant (B_2 - B_1)\overline{G}(u_2). \tag{B.1}$$

Similarly, we have $f(B_1, u_1) - f(B_2, u_1) = (B_2 - B_1)\overline{G}(u_1)$ and $f(B_2, u_2) \geqslant f(B_2, u_1)$. Therefore, we can upper bound the difference

$$m(B_1) - m(B_2) \leqslant (B_2 - B_1)\overline{G}(u_1). \tag{B.2}$$

(a) Since $f(B, B) = 0$, it follows that $m(B) \geqslant 0$ for all B. Since $\overline{G} \geqslant 0$, it follows that $m(B_2) - m(B_1) \leqslant 0$ for $B_1 < B_2$ from Eq. (B.1).

(b) Since $\overline{G} \leqslant 1$, it follows from Eq. (B.2) that $0 \leqslant m(B_1) - m(B_2) \leqslant B_2 - B_1$. Similarly for $B_2 < B_1$, we observe that $0 \leqslant m(B_2) - m(B_1) \leqslant B_1 - B_2$. Combining these, we have $|m(B_1) - m(B_2)| \leqslant |B_1 - B_2|$, implying Lipschitz-1 continuity of m. Finally, $f(B, u)$ is affine, and hence, convex in B. Hence, the maximum m of convex functions $f(B, u)$ is also convex in B [44, Section 3.2.3] .

(c) Subtracting Eq. (B.1) from Eq. (B.2), we get $(B_2 - B_1)(\overline{G}(u_2) - \overline{G}(u_1)) \leqslant 0$. This implies that if $B_2 > B_1$, then $\overline{G}(u_2) \leqslant \overline{G}(u_1)$. The monotonic decrease of $\overline{G}$ implies that $u_2 \geqslant u_1$.

B.2. Proof of Lemma 18

We assume the Lemma hypothesis.

(a) Using Eq. (26) for $i = 0$, we see that $\theta = \lambda m(\Delta(0))$. The result follows from the non-negativity of m from Lemma 17.

(b) We first prove that $\Delta(0) > 0$ via contradiction. Assume that $\Delta(0) \leqslant 0$, and assume the inductive hypothesis that $\Delta(i) \leqslant 0$ for some $i \in \mathcal{X}' \setminus \{0\}$. Then, it follows from Eq. (26)

$$m(\Delta(i)) = \frac{\theta - i\mu\Delta(i-1)}{\lambda} \geqslant \frac{\theta}{\lambda} = m(\Delta(0)) \geqslant 0.$$

From monotone decrease of m in Lemma 17(a) and the induction step, it follows that $\Delta(i) \leqslant \Delta(0) \leqslant 0$ for all $i \in \mathcal{X}'$. From Eq. (26) for $i = K$, we get $\Delta(K-1) = \frac{\theta}{K\mu} \geqslant 0$ from the non negativity of θ from part (a). This leads to a contradiction and hence we see that $\Delta(0) > 0$.
From Eq. (26) for $i = 1$ and positivity of $\Delta(0)$, we observe that

$$m(\Delta(1)) = \frac{\theta - \mu\Delta(0)}{\lambda} \leqslant \frac{\theta}{\lambda} = m(\Delta(0)).$$

Lemma 17(a) implies that m is decreasing. It follows that $\Delta(1) \geqslant \Delta(0)$. Assuming the inductive hypothesis $\Delta(i-1) \geqslant \Delta(i-2)$ for some $i \in \{2, \ldots, K-1\}$ and positivity of $\Delta(i)$s, we get from Eq. (26)

$$m(\Delta(i-1)) - m(\Delta(i)) = \frac{\mu}{\lambda}(i\Delta(i-1) - (i-1)\Delta(i-2)) \geqslant 0.$$

From monotone decrease of m and the induction step, it follows that $\Delta(i) \geqslant \Delta(i-1)$ for all $i \in \mathcal{X}' \setminus \{0\}$.

(c) This follows by combining the monotone increase of $\Delta(i)$ shown in part (b), and monotonicity of $u^*(B)$ in B shown in Lemma 17(c).

B.3. Proof of Proposition 20

Notice that the optimal revenue rate $\theta^*(\lambda)$ is the solution to Eqs. (27a)–(27c) as a function of λ, for a fixed μ and K.

(a) To begin with let us fix both θ and μ and vary λ in Eqs. (27b) and (27c). It follows that if g_i is non-increasing in λ, then $m(g_i)$ is non-decreasing in λ from its monotone decrease property. Since $g_{i-1} \propto \theta - \lambda m(g_i)$, it follows that g_{i-1} is decreasing and $m(g_{i-1})$ is increasing in λ. Since $g_{K-1} = \theta/K\mu$ is constant in λ, it follows that $m(g_i)$ is increasing in λ for all $i \in \mathcal{X}'$ and fixed θ and μ. Since $\theta^*(\lambda) = \lambda m(g_0)$ from Eq. (27a), it follows that the optimal revenue rate $\theta^*(\lambda)$ is increasing in λ for a fixed μ.

(b) The argument is via contradiction. Let $\theta^*(\lambda)/\lambda$ increase with λ. Observe that $g_{K-1}(\theta^*(\lambda)) = \frac{\theta^*(\lambda)}{K\mu}$ increases with λ. Since θ^* is the solution to Eqs. (27a)–(27c) for all $i \in \mathcal{X}$,

$$\frac{g_{i-1}(\theta^*(\lambda))}{\lambda} = \frac{\theta^*(\lambda)/\lambda - m(g_i(\theta^*(\lambda)))}{i\mu}, \quad i \in [K-1].$$

It follows that $g_{i-1}(\theta^*(\lambda))/\lambda$ is an increasing function of λ, if g_i is an increasing function of λ. It follows from induction that $g_0(\theta^*(\lambda))$ is an increasing function of λ, and hence $m(g_0(\theta^*(\lambda))) = \theta^*(\lambda)/\lambda$ is a decreasing function of λ. This leads to a contradiction.

B.4. Proof of Proposition 21

The optimal revenue rate $\theta^*(\mu)$ is the solution to Eqs. (27a)–(27c) as a function of μ, for a fixed λ and K.

(a) To begin with let us fix both θ and μ and vary λ in Eqs. (27b) and Eq. (27c). From Eq. (27b), we observe that $g_{i-1} = (\theta - \lambda m(g_i))/i\mu$ for $i \in [K-1]$. Hence, if g_i is decreasing with μ, then $m(g_i)$ is increasing in μ due to its monotone decrease property, and hence g_{i-1} is decreasing with μ. Since $g_{K-1} = \theta/K\mu$ from Eq. (27c) for $i = K$, it follows by induction that g_0 is decreasing and hence $\lambda m(g_0)$ is increasing in μ. As a result, if we increase μ keeping λ fixed, the average revenue rate $\theta^*(\mu)$, the solution to $\theta = \lambda m(g_0(\theta))$ increases in μ.

(b) The argument is via contradiction. Let $\theta^*(\mu)/\mu$ increase with μ. We obtain from Eq. (27b) for $i \in [K-1]$,

$$g_i(\theta^*(\mu)) = m^{-1}\left(i\mu\left(\frac{\theta^*(\mu)/i\mu - g_{i-1}(\theta^*(\mu))}{\lambda}\right)\right).$$

Then, it follows that if g_{i-1} is decreasing with μ, then g_i is also decreasing in μ. From Eq. (27a) for $i = 0$, we see that $g_0(\theta^*(\mu)) = m^{-1}(\frac{\theta^*(\mu)}{\lambda})$ is decreasing with μ, and hence it follows that g_{K-1} is decreasing and in μ. However $g_{K-1}(\theta^*) = \theta^*(\mu)/K\mu$ was assumed to be increasing in μ, that leads to a contradiction.

B.5. Proof of Proposition 22

We define functions

$$\bar{g}_0(\theta) \triangleq m^{-1}\left(\frac{\theta}{\lambda}\right),$$
$$\bar{g}_i(\theta) \triangleq m^{-1}\left(\frac{\theta - i\mu\bar{g}_{i-1}(\theta)}{\lambda}\right), \ i \in [K-1].$$

Following similar arguments as in the proof of Theorem 19(a) we can iteratively show that $\bar{g}_i(\theta)$ are decreasing in θ for all $i < K$.

(a) It follows that $\lambda m(\bar{g}_0) = \theta$. and $\bar{g}_{i-1} = (\theta - \lambda m(\bar{g}_i))/i\mu$ for $i \in [K-1]$. Hence, from Eqs. (27a)–(27c) it follows that the optimal average reward $\theta^*(K)$ is the solution to the fixed point equation $\theta = K\mu\bar{g}_{K-1}(\theta)$. From Lemma 18(b), we have $\bar{g}_i(\theta) > \bar{g}_{i-1}(\theta)$ for all $\theta \geqslant 0$ and $i \in \mathcal{X}'$. In particular, $(K+1)\mu\bar{g}_K(\theta) > K\mu\bar{g}_{K-1}(\theta)$ for all $\theta \geqslant 0$. Hence we can infer that $\theta^*(K)$ increases with K.

(b) Since $\bar{g}_{K-1}(\theta^*(K)) = \theta^*(K)/K\mu$, it suffices to show that $\bar{g}_{K-1}(\theta^*(K))$ is decreasing in K. We show this by contradiction. To this end, we assume that $\bar{g}_K(\theta^*(K+1)) > \bar{g}_{K-1}(\theta^*(K))$. Together with this hypothesis and monotone increase of $\bar{g}_i$ from Lemma 18(b), we obtain

$$(K+1)\bar{g}_K(\theta^*(K+1)) - K\bar{g}_{K-1}(\theta^*(K)) > \bar{g}_0(\theta^*(K+1)).$$

Multiplying both the sides by μ/λ and using definitions of $\theta^*(K)$ and $\theta^*(K+1)$, the above inequality reduces to

$$\frac{\theta^*(K+1) - \mu\bar{g}_0(\theta^*(K+1))}{\lambda} > \frac{\theta^*(K)}{\lambda}.$$

From the monotone decrease property of m and definition of $\bar{g}_1$ and $\bar{g}_0$, we obtain $\bar{g}_1(\theta^*(K+1)) < \bar{g}_0(\theta^*(K))$. We will inductively show that $\bar{g}_i(\theta^*(K+1)) < \bar{g}_{i-1}(\theta^*(K))$ for all $i \in [K]$. We have already shown the base case of $i = 1$. We assume that the inductive hypothesis holds for some $i \in [K-1]$. Further, Lemma 18(b) implies that $\bar{g}_i$ increases in i for a fixed argument. Together with inductive and initial hypothesis, we obtain

$$\begin{aligned} &K(\bar{g}_K(\theta^*(K+1)) - \bar{g}_{K-1}(\theta^*(K))) \\ &+ (\bar{g}_K(\theta^*(K+1)) - \bar{g}_i(\theta^*(K+1))) \\ &\quad > 0 > i(\bar{g}_i(\theta^*(K+1)) - \bar{g}_{i-1}(\theta^*(K))). \end{aligned}$$

Rearranging the terms, multiplying both the sides by μ/λ, using definitions of $\theta^*(K), \theta^*(K+1), \bar{g}_i, \bar{g}_{i+1}$, and from the monotone decrease of m, we get

$$\bar{g}_{i+1}(\theta^*(K+1)) < \bar{g}_i(\theta^*(K)).$$

This completes the induction step. We thus see that $\bar{g}_i(\theta^*(K+1)) < \bar{g}_{i-1}(\theta^*(K))$ for all $i \in [K]$. In particular, we get $\bar{g}_K(\theta^*(K+1)) < \bar{g}_{K-1}(\theta^*(K))$ which contradicts the initial hypothesis.

(c) Recall that the optimal price for i busy servers, when the system has K servers is given by

$$u_i^*(K) = u^*(\bar{g}_i(\theta^*(K))).$$

We know that $\theta^*(K)$ is increasing in K from part (a) of the proof, $\bar{g}_i(\theta)$ is decreasing in θ as observed in the beginning of the proof, and u^* is non-decreasing in its argument from Lemma 17(c). The result follows from the combination of these three observations.

Appendix C. Proofs for MDP with general arrival process

Lemma 29. *For the probability $\alpha_{k,j}$ of j departures from state k defined in Eq. (5), for the K server system with* i.i.d. *exponential service with rate μ and* i.i.d. *job interarrival times $(U_n : n \in \mathbb{N})$, the following statements are true for all $i \in \mathcal{X}'$.*

(a) $a_{i+1,j} \leqslant a_{i,j}$ *for all* $j \leqslant i-1$.
(b) $a_{i,i-1} + \alpha_{i,0} = 1$.
(c) $\sum_{j=0}^{i} a_{i+1,j} - \sum_{j=0}^{i-1} a_{i,j} = \alpha_{1,1}$.

Proof. Given the first interarrival time U_1, we can define a sequence of conditionally *i.i.d.* Bernoulli random variables $(\xi_r : r \in \mathbb{N})$ such that $\mathbb{E}[\xi_r|U_1] = 1 - e^{-\mu U_1}$. We define a sequence of increasing binomial random variables $(X_k : k \in \mathbb{N})$ such that $X_k \triangleq \sum_{r\in[k]} \xi_r$. We observe that

$$\mathbb{E}[\mathbb{1}_{\{X_k=k-i\}}|U_1] = \binom{k}{k-i}(1-e^{-\mu U_1})^{k-i}e^{-i\mu U_1}.$$

Therefore, it follows from Eq. (5) that the probability of $k-i$ departures from state k is $\alpha_{k,k-i} = P\{X_k = k-i\}$.

(a) Recall that $a_{i,j} = \sum_{l=0}^{j} \alpha_{i,i-l}$ and hence we can write $a_{i,j} = P\{X_i \leqslant j\}$. Since X is monotonically increasing $\{X_{k+1} \leqslant j\} \subseteq \{X_k \leqslant j\}$, and the result follows from the monotonicity of probability.
(b) From the definition of random sequence X and $a_{i,j} = \sum_{l=0}^{j} \alpha_{i,i-l}$, we have $a_{i,i} = P\{X_i \leqslant i\} = 1 = \alpha_{i,0} + a_{i,i-1}$.
(c) From the definition of $(a_{ij} : j \leqslant i)$ and monotone sequence X, we can write

$$\sum_{j=0}^{i} a_{i+1,j} = \sum_{l=1}^{i+1} lP\{X_{i+1} = l\} = \mathbb{E}X_{i+1}.$$

Since $X_{i+1} - X_i = \xi_{i+1}$ and $\mathbb{E}\xi_{i+1} = \mathbb{E}(1-e^{-\mu U_1}) = \alpha_{1,1}$, the result follows. □

C.1. Proof of Theorem 23

Recall that the map m is non-negative and monotonically decreasing from Lemma 17(a), and that m is Lipschitz-1 continuous from Lemma 17(b).

(a) From the non-negativity of m and writing $\theta = \lambda m(b_0)$ from Eq. (33a), we get the result.

(b) We first prove that $\Delta(0) > 0$ via contradiction. Assume that $\Delta(0) \leqslant 0$. We will show by induction that for all $i \in \{2, \ldots, K\}$, the following two conditions hold true,

$$\Delta(i-1) < \cdots < \Delta(0) \leqslant 0, \tag{C.1a}$$

$$\sum_{j=0}^{i-1} a_{i,j}\Delta(j) < \sum_{j=0}^{i-2} a_{i-1,j}\Delta(j) \leqslant 0. \tag{C.1b}$$

The inductive hypothesis in Eqs. (C.1a)–(C.1b) at $i = K$, together with equality in Eq. (33c), we get $\frac{\theta}{\lambda} = \sum_{j=0}^{K-1} a_{K,j}\Delta(j) < 0$. This contradicts the part (a) of the theorem, which we have already established. The contradiction implies that following two conditions hold true for all $i \in [K]$,

$$\Delta(i-1) \geqslant \ldots \geqslant \Delta(0) > 0, \tag{C.2a}$$

$$\sum_{j=0}^{i-1} a_{i,j}\Delta(j) \geqslant \sum_{j=0}^{i-2} a_{i-1,j}\Delta(j) > 0. \tag{C.2b}$$

From Eqs. (33a)–(33b) and the condition (C.2b), we observe that $m(b_i) < m(b_{i-1})$. From the monotonicity of map m in Lemma 17, we observe that the sequence b is non-decreasing. Further, since $b_0 = m^{-1}(\frac{\theta}{\lambda}) \geqslant 0$, and the result follows. Therefore, it suffices to show the inductive hypothesis in Eqs. (C.1a), (C.1a) holds true for all $i \geqslant 2$.

Step 1: Base case of induction. We will first show the base case of $i = 2$ holds true for the induction. From Eq. (33a)–(33b), we get

$$m(b_1) = m(b_0) - a_{1,0}\Delta(0).$$

Since $a_{i,j}$ are the sum of probabilities, they are nonnegative. Therefore, $-a_{1,0}\Delta(0) \geqslant 0$ from the hypothesis, and the above equation implies that $m(b_1) \geqslant m(b_0)$. Since m is a nonincreasing function, it follows that $b_1 \leqslant b_0$. Further, from the Lipschitz-1 continuity of m, we get $m(b_1) - m(b_0) \leqslant |b_1 - b_0| = b_0 - b_1$. It follows that $b_1 \leqslant b_0 + a_{1,0}\Delta(0)$. From the definition of sequence b in Eq. (32) and the definition of $a_{ij} = \sum_{l=0}^{j} \alpha_{i,i-l}$, we get

$$\alpha_{2,0}\Delta(1) \leqslant (\alpha_{1,0} + \alpha_{2,2})\Delta(0).$$

Since $0 < \alpha_{2,0} < \alpha_{1,0} + \alpha_{2,2}$, we see that $\Delta(1) < \Delta(0) \leqslant 0$. Using the fact $\Delta(1) < \Delta(0) \leqslant 0$ and from the definition of $a_{ij} = \sum_{l=0}^{j} \alpha_{i,i-l}$, we can write the following inequality

$$\begin{aligned} a_{2,0}\Delta(0) + a_{2,1}\Delta(1) &< (a_{2,0} + a_{2,1})\Delta(0) \\ &= (2\alpha_{2,2} + \alpha_{2,1})\Delta(0) \\ &= 2\alpha_{1,1}\Delta(0) \leqslant a_{1,0}\Delta(0). \end{aligned}$$

Step 2: Inductive step of induction. We have shown the base case of $i = 2$ holds true for the induction, and assume the inductive hypothesis in Eqs. (C.1a), (C.1a) holds true for some $i \geqslant 2$. From Eq. (33b), we can write the difference for $i \in [K-1]$

$$m(b_i) - m(b_{i-1}) = \sum_{j=0}^{i-2} a_{i-1,j}\Delta(j) - \sum_{j=0}^{i-1} a_{i,j}\Delta(j).$$

From the inductive hypothesis for i, it follows that $m(b_i) > m(b_{i-1})$. Following the similar discussion to the base case, from the monotone nonincreasing and Lipschitz-1 continuity of the map m, it follows that $b_i < b_{i-1} + m(b_{i-1}) - m(b_i)$. Using Eq. (33b) to write the difference $m(b_i) - m(b_{i-1})$, sequence $b = (b_i : i \in \mathcal{X}')$ defined in Eq. (32) to write the difference $b_{i-1} - b_i$, and substituting $a_{i,j} = \sum_{l=0}^{j} \alpha_{i,i-l}$ where $\alpha_{k,j}$ is the probability of j departures from state k in Eq. (5), we get

$$\begin{aligned} \alpha_{i+1,0}\Delta(i) \leqslant &\sum_{j=0}^{i-2} (a_{i+1,j} - a_{i,j})\Delta(j) \\ &+ (\alpha_{i,0} + a_{i+1,i-1})\Delta(i-1). \end{aligned}$$

From Lemma 29(a), we have $a_{i+1,j} \leqslant a_{i,j}$ for all $j \leqslant i$ and from inductive hypothesis $\Delta(i-1) = \min_{j \leqslant i-1} \Delta(j) \leqslant 0$. Further, from Lemma 29(c), we have $\sum_{j=1}^{i} a_{i+1,j} - \sum_{j=1}^{i-1} a_{i,j} = \alpha_{1,1}$. Using these three facts in the above equation,

we get

$$\alpha_{i+1,0}\Delta(i) \leqslant (\alpha_{i,0} - a_{i+1,i} + a_{i,i-1} + \alpha_{1,1})\Delta(i-1).$$

From Lemma 29(b), we have $a_{i,i} = 1-\alpha_{i,0}$ for all $i \in \mathcal{X}'$ and hence we get $\alpha_{i+1,0}\Delta(i) \leqslant (\alpha_{i+1,0}+\alpha_{1,1})\Delta(i-1)$. Since $\alpha_{i,j}$ are probabilities, the first inductive result $\Delta(i) < \Delta(i-1)$ follows. Together with the inductive hypothesis in Eqs. (C.1a), (C.1a), we get $\min_{j\leqslant i}\Delta(j) = \Delta(i) \leqslant 0$. From Lemma 29(a), we have $a_{i+1,j} \leqslant a_{i,j}$ for all $i \in \mathcal{X}'$ and $j \leqslant i$. Therefore, $\sum_{j=0}^{i-1}(a_{i+1,j} - a_{i,j})\Delta(j) \leqslant \sum_{j=0}^{i-1}(a_{i+1,j} - a_{i,j})\Delta(i)$, and we can write the difference

$$\sum_{j=0}^{i} a_{i+1,j}\Delta(j) - \sum_{j=0}^{i-1} a_{i,j}\Delta(j) \leqslant (\sum_{j-0}^{i} a_{i+1,j} - \sum_{j=0}^{i-1} a_{i,j})\Delta(i).$$

From Lemma 29(c), we have $\sum_{j=0}^{i} a_{i+1,j} - \sum_{j=0}^{i-1} a_{i,j} = \alpha_{1,1}$ for all $i \in \mathcal{X}'$. Substituting this result in the above equation, we get the second inductive result, and this completes the induction.

(c) Recall that the optimal price in state i is $u_i^* = \arg\max_u f(b_i, u)$ for the map f defined in Eq. (25), where the sequence $b = (b_i : i \in \mathcal{X}')$ is positive and increasing from part (b). Therefore, it follows from Lemma 17(c) that the optimal price u_i^* is increasing with the number of busy servers i.

References

[1] A. Krishnan KS, C. Singh, S.T. Maguluri, P. Parag, Optimal pricing in finite server systems, in: 2020 18th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT), IEEE, 2020, pp. 1–8.
[2] R. Butkiene, J. Karpovic, R. Sabaliauskas, L. Sriupsa, M. Vaitkunas, G. Vilutis, Survey of open-source clouds capabilities extension, in: International Conference on Information and Software Technologies, Springer, 2020, pp. 3–13.
[3] H. Xu, B. Li, Dynamic cloud pricing for revenue maximization, IEEE Trans. Cloud Comput. 1 (2) (2013) 158–171.
[4] Y. Chi, X. Li, X. Wang, V.C. Leung, A. Shami, A fairness-aware pricing methodology for revenue enhancement in service cloud infrastructure, IEEE Syst. J. 11 (2) (2015) 1006–1017.
[5] A. Greenberg, J. Hamilton, D.A. Maltz, P. Patel, The cost of a cloud: research problems in data center networks, ACM SIGCOMM Comput. Commun. Rev. 39 (1) (2008) 68–73.
[6] C. Wu, R. Buyya, K. Ramamohanarao, Cloud pricing models: Taxonomy, survey, and interdisciplinary challenges, ACM Comput. Surv. 52 (6) (2019) 1–36.
[7] P. Naor, The regulation of queue size by levying tolls, Econometrica: J. Econ. Soc. (1969) 15–24.
[8] N.M. Edelson, D.K. Hilderbrand, Congestion tolls for Poisson queuing processes, Econometrica: J. Econ. Soc. (1975) 81–92.
[9] C. Larsen, Investigating sensitivity and the impact of information on pricing decisions in an M/M/1/∞ queueing model, Int. J. Prod. Econ. 56 (1998) 365–377.
[10] R. Hassin, Consumer information in markets with random product quality: The case of queues and balking, Econometrica: J. Econ. Soc. (1986) 1185–1195.
[11] R. Hassin, M. Haviv, To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems, Vol. 59, Springer Science & Business Media, 2003.
[12] R. Hassin, Rational Queueing, CRC Press, 2016.
[13] D.W. Low, Optimal dynamic pricing policies for an M/M/S queue, Oper. Res. 22 (3) (1974) 545–561.
[14] I.C. Paschalidis, J.N. Tsitsiklis, Congestion-dependent pricing of network services, IEEE/ACM Trans. Netw. 8 (2) (2000) 171–184.
[15] H. Chen, M.Z. Frank, State dependent pricing with a queue, IIE Trans. 33 (10) (2001) 847–860.
[16] C. Borgs, J.T. Chayes, S. Doroudi, M. Harchol-Balter, K. Xu, The optimal admission threshold in observable queues with state dependent pricing, Prob. Eng. Inf. Sci. 28 (1) (2014) 101–119.
[17] S. Ziya, H. Ayhan, R.D. Foley, Optimal prices for finite capacity queueing systems, Oper. Res. Lett. 34 (2) (2006) 214–218.
[18] I. Maoui, H. Ayhan, R.D. Foley, Congestion-dependent pricing in a stochastic service system, Adv. Appl. Prob. 39 (4) (2007) 898–921.
[19] E.A. Feinberg, F. Yang, Optimal pricing for a GI/M/k/N queue with several customer types and holding costs, Queueing Syst. 82 (1–2) (2016) 103–120.
[20] X. Wang, S. Andradóttir, H. Ayhan, Optimal pricing for tandem queues with finite buffers, Queueing Syst. 92 (3–4) (2019) 323–396.
[21] S. Yoon, M.E. Lewis, Optimal pricing and admission control in a queueing system with periodically varying parameters, Queueing Syst. 47 (3) (2004) 177–199.
[22] E.B. Çil, F. Karaesmen, E.L. Örmeci, Dynamic pricing and scheduling in a multi-class single-server queueing system, Queueing Syst. 67 (4) (2011) 305–331.
[23] A.V. den Boer, Dynamic pricing and learning: historical origins, current research, and new directions, Surv. Oper. Res. Manag. Sci. 20 (1) (2015) 1–18.
[24] O. Besbes, A.N. Elmachtoub, Y. Sun, Static pricing: Universal guarantees for reusable resources, in: Proceedings of the 2019 ACM Conference an Economics and Computation, 2019, pp. 393–394.
[25] J. Kim, R.S. Randhawa, The value of dynamic pricing in large queueing systems, Oper. Res. 66 (2) (2018) 409–425.
[26] B. Ata, T.L. Olsen, Congestion-based leadtime quotation and pricing for revenue maximization with heterogeneous customers, Queueing Syst. 73 (1) (2013) 35–78.
[27] S. Çelik, C. Maglaras, Dynamic pricing and lead-time quotation for a multiclass make-to-order queue, Manage. Sci. 54 (6) (2008) 1132–1146.
[28] B. Ata, T.L. Olsen, Near-optimal dynamic lead-time quotation and scheduling under convex-concave customer delay costs, Oper. Res. 57 (3) (2009) 753–768.
[29] J.M. George, J.M. Harrison, Dynamic control of a queue with adjustable service rate, Oper. Res. 49 (5) (2001) 720–731.
[30] R. Bitar, P. Parag, S. El Rouayheb, Minimizing latency for secure distributed computing, in: 2017 IEEE International Symposium on Information Theory (ISIT), IEEE, 2017, pp. 2900–2904.
[31] A. Badita, P. Parag, V. Aggarwal, Optimal server selection for straggler mitigation, IEEE/ACM Trans. Netw. 28 (2) (2020) 709–721.

[32] P. Cong, L. Li, J. Zhou, K. Cao, T. Wei, M. Chen, S. Hu, Developing user perceived value based pricing models for cloud markets, IEEE Trans. Parallel Distrib. Syst. 29 (12) (2018) 2742–2756.
[33] W. Ellens, J. Akkerboom, R. Litjens, H. van den Berg, et al., Performance of cloud computing centers with multiple priority classes, in: 2012 IEEE Fifth International Conference on Cloud Computing, IEEE, 2012, pp. 245–252.
[34] B. Bouterse, H. Perros, Scheduling cloud capacity for Time- Varying customer demand, in: Inter. Conf. Cloud Netw. (CLOUDNET), 2012, pp. 137–142.
[35] S. Vakilinia, M.M. Ali, D. Qiu, Modeling of the resource allocation in cloud computing centers, Comput. Netw. 91 (C) (2015) 453—470.
[36] J.R. Norris, Markov Chains, in: Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1997.
[37] K. Chung, A Course in Probability Theory, Academic Press, 2001.
[38] S.M. Ross, Stochastic Processes, John Wiley & Sons, 1996.
[39] F.P. Kelly, Reversibility and Stochastic Networks, Cambridge University Press, USA, 2011.
[40] L. Takács, On a probability problem concerning telephone traffic, Acta Math. Acad. Sci. Hungarica 8 (3–4) (1957) 319–324.
[41] D.P. Bertsekas, Dynamic Programming and Optimal Control, Vol. 2, Athena Scientific, Belmont, USA, 2007.
[42] R. Divya, A.P. Azad, C. Singh, Fair and optimal mobile assisted offloading, in: WiOpt, Paris, France, 2017, pp. 1–8.
[43] M. Zukerman, Introduction to queueing theory and stochastic teletraffic models, 2013, arXiv preprint arXiv:1307.2968.
[44] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

Ashok Krishnan K.S. received the M.Sc. (Engg.) and Ph.D. degrees in Electrical Communication Engineering from the Indian Institute of Science, Bengaluru, India, in 2020. Prior to that he had obtained his B.Tech degree from the College of Engineering, Trivandrum, India. His Ph.D dissertation focused on the areas of scheduling, routing, distributed control and optimization of wireless communication networks with queues. His research interests include wireless communications, queueing theory, optimization, control theory, applied probability and learning. He is currently employed with Qualcomm India Pvt. Ltd., Bangalore, as a Senior Engineer.

Chandramani K. Singh is an Assistant Professor in the Department of ESE at the Indian Institute of Science, Bangalore. His interests are in the areas of communication networks, data centers and smart grids. Chandramani received the M.E. and Ph.D. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, India, in 2005 and 2012, respectively. He worked at ESQUBE Communication Solutions Pvt. Ltd., Bangalore, from 2005 to 2006. He was a Research Engineer with TREC, a joint research team between INRIA Rocquencourt and ENS de Paris, from 2012 to 2013, and a Postdoctoral Research Associate at CSL, University of Illinois at Urbana Champaign, IL, USA, from 2013 to 2014.

Siva Theja Maguluri is Fouts Family Early Career Professor and Assistant Professor in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Tech. He received his B.Tech in Electrical Engineering from IIT Madras in 2008, M.S in ECE, M.S. in Applied Math and a PhD in ECE all from University of Illinois at Urbana Champaign. His research interests span the areas of Networks, Control, Optimization, Algorithms, Applied Probability and Reinforcement Learning. He is a recipient of the biennial "Best Publication in Applied Probability" award, NSF CAREER award, "CTL/BP Junior Faculty Teaching Excellence Award", and "Student Recognition of Excellence in Teaching: Class of 1934 CIOS Award."

Parimal Parag Parimal Parag (Senior Member, IEEE) received the B.Tech. and M.Tech. degrees from IIT Madras in 2004 and the Ph.D. degree from Texas A&M University in 2011, all in electrical engineering.He joined the Indian Institute of Science in 2014, where he is currently an Associate Professor with the Department of Electrical Communication Engineering. Prior to that, he was a Senior System Engineer (Research and Development) with ASSIA, Inc., Redwood City, CA, USA, from 2011 to 2014. His research interests include the design and analysis of large scale distributed systems. He was a co-author of the 2018 IEEE ISIT Student Best Paper, and a recipient of the 2017 Early Career Award from the Science and Engineering Research Board.