Optimal Pricing in Finite Server Systems

Ashok Krishnan K. S.
Department of ECE
Indian Institute of Science
Bangalore, India
ashokk@iisc.ac.in

Chandramani Singh
Department of ESE
Indian Institute of Science
Bangalore, India
chandra@iisc.ac.in

Siva Theja Maguluri
School of ISyE
Georgia Institute of Technology
Atlanta, USA
siva.theja@gatech.edu

Parimal Parag
Department of ECE
Indian Institute of Science
Bangalore, India
parimal@iisc.ac.in

***Abstract*—We consider a system of K servers, where customers arrive according to a Poisson process, and have independent and identical (*i.i.d.*) exponential service times and *i.i.d.* valuations of the service. We consider the setting where customers leave when they find all servers busy. Service provider announces a price to an incoming customer, depending on the number of busy servers. An incoming arrival enters the system if its valuation exceeds the price. We find the optimal state dependent pricing, that maximizes the revenue rate for the service provider.**

I. 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. 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.

Optimal service pricing is central to thriving operation of server farms. 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 the 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 govern future customers' valuation. We thus see that determining optimal prices is a complex problem. The problem is further complicated when the 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.

This work was supported in part by the Department of Telecommunications, Govt. of India, under Grant DOTC-0001, in part by the RBCCPS, and in part by the Centre for Networked Intelligence (a Cisco CSR initiative) at IISc.

A. 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 [1]. One of the earliest works that studied pricing of queues was [2], 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*. Each customer has a pure strategy, which is a threshold, which is a function of the toll fixed by the service provider. If the queue length is greater than the threshold, the customer balks; else they join. It was shown that the socially optimal threshold was higher than the threshold for revenue maximization. A subsequent work [3] 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 [2] or at related models. The effect of the reward variance on the performance is studied in [4]. In [5], 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 works, have been summarized in [6].

In [7], the authors look at the problem from the perspective of the service provider. Here, they are interested in maximizing the expected discounted revenue, while keeping the queueing model of [2]. They obtain a revenue optimizing threshold queue length beyond which entries are not allowed into the queue. This threshold can be computed numerically. In [8], an explicit form is derived for the threshold, and they characterize the earning rate asymptotically. In [9], the authors study a cloud

ISBN 978-3-903176-29-4

system where the utility of the customers follows the α-fair model, and show that, discriminating between customers based on their valuation of the service does not improve the revenue, when compared to uniform pricing. A survey of available cloud pricing models is given in [10]. A system of two competing firms is studied in [11]. One firm offers a fixed cost of service, and a corresponding fixed waiting time, while the others offers customers lower waiting times proportional to higher bids. This is formulated as a game, and the equilibrium behavior is obtained. It is shown that customers with higher or lower waiting costs prefer the bidding structure, while those with moderate costs prefer the fixed pricing scheme.

B. Our Contribution

We assume a service provider with K servers. We further assume that the customers arrive according to a Poisson process, having *i.i.d.*exponential service times and *i.i.d.*values for the service. Following is a preview of our main results.

1) We observe that the system with infinitely many servers (i.e., $K = \infty$) resembles an $M/M/\infty$ queue. We show that 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 ($K < \infty$). We derive a bound on the revenue rate for the optimal uniform price. We also study asymptotic revenue rates for uniform pricing.
3) 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 obtained via solving a fixed point iteration.
4) We study how optimal prices and corresponding revenue rates vary with customer arrival rates, service rates and the number of servers K.

II. System Model

We model the system as a queuing system with K servers. Jobs arrive to this server farm as a Poisson process with rate λ. The jobs have random service time requirements that are independent and identically distributed (*i.i.d.*) exponential with mean $\frac{1}{\mu}$. Furthermore, each job has a random *i.i.d.* positive value V sampled from a continuous distribution G.

Assumption 1. We assume the distribution G such that the function $x \mapsto f(x)(1-G(x))$ has a unique finite maximum for any monotonically increasing function f.

Upon arrival, the server quotes a price for the job. The price offered by the server is state dependent, with the state variable k being the number of servers busy in the system. This price is denoted by p_k. If the quoted price p_k is less than the value V_n of the nth job, then the job joins service; otherwise it leaves. When a job joins the service in state k, the system earns a revenue p_k. We assume that $p_K = \infty$, i.e., an arrival seeing all K servers busy, leaves the system.

The state of the system is denoted by the number of busy servers, and the state space is denoted by $\mathcal{X} \triangleq \{0, \ldots, K\}$. Since $p_K = \infty$, the reduced state space is denoted by $\mathcal{X}' \triangleq \{0, \ldots, K-1\}$, and we denote a state-dependent price vector by $\underline{P} = (p_0, \ldots, p_{K-1}) \in \mathbb{R}_+^{\mathcal{X}'}$, and the revenue earned until time t by $R(t)$. The limiting revenue rate for this K server system with price vector $\underline{P}$ is denoted by

$$R(K, \underline{P}) \triangleq \lim_{t\to\infty} \frac{R(t)}{t}.$$

Our main goal is to find the pricing vector that maximizes revenue. Formally, we solve the following problem.

Problem 2. Find the optimal price vector $\underline{P}_K^* = (P^*, \ldots, P_{K-1}^*)$ that maximizes the limiting system revenue rate $R(K, \underline{P})$. That is, we wish to find

$$\underline{P}_K^* = \arg\max R(K, \underline{P}).$$

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 3. Find the uniform price p that maximizes the limiting system revenue rate $R(\underline{P}, K)$. That is, we wish to find

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

We denote the optimal revenue rate by $R^* = R(K, \underline{P}_K^*)$, and compare it to the revenue rate $R(K, p_K^*\mathbf{1})$ for the best uniform pricing.

Notation 4. We denote the sets of natural numbers, non-negative integers, non-negative reals, and first n positive integers by $\mathbb{N}, \mathbb{Z}_+, \mathbb{R}_+$, and $[n]$ respectively.

III. Analysis

Let $X(t)$ denote the number of busy servers in the system, at time t. It is easy to see that $(X(t) \in \mathcal{X}, t \geqslant 0)$ is a continuous time Markov chain. For this Markov process, the transition rate from state i to $i+1$ (for $i+1 \leqslant K$) is given by λ_i, where,

$$\lambda_i = \lambda \Pr\{V \geqslant p_i\} = \lambda \overline{G}(p_i),$$

where we denote the complementary distribution of value by $\overline{G}(x) = 1 - G(x)$. The transition rate from state i to state $i-1$ is given by $i\mu$. We have depicted the state space of this system for $K = 3$ in Figure 1.

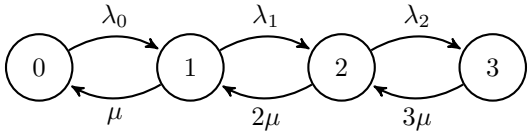

Fig. 1. Number of busy servers for state dependent price, where we have three identical servers.

Since $X(t)$ is a finite state, irreducible Markov chain, it is positive recurrent [12]. Let π denote the stationary distribution of $X(t)$. Then, the following result holds.

Theorem 5 (Kelly [13])**.** *The stationary distribution of the Markov chain $X(t)$ is given in terms of the load factor $\rho \triangleq \frac{\lambda}{\mu}$ as*

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

Proof: Since, all birth-death processes are reversible [13], it's easy to compute the equilibrium distribution of $X(t)$ using detailed balanced equations, such that

$$\pi_j \lambda_j = \pi_{j+1}(j+1)\mu.$$

Eq. (1) follows from recursively computing π_j in terms of π_0 and load factor $\rho = \frac{\lambda}{\mu}$. The equilibrium probability π_0 of being in state 0, follows from the conservation of probability. ∎

Next, we are interested in computing the mean equilibrium revenue from the system. Assuming $X(0) = 0$, we can inductively define the nth entrance to state 0 as

$$\tau_n \triangleq \inf\{t > \tau_{n-1} : X(t) = 0\}, \qquad \tau_0 = 0.$$

It follows that τ_n is a stopping time adapted to the Markov process $X(t)$. Since a continuous time Markov chain is also strongly Markov [14], it follows that $(X_{\tau_n+t} : t \geqslant 0)$ is independent of the past $(X_t : t \leqslant \tau_n)$. Further, from homogeneity of the process $X(t)$, it follows that $(X_{\tau_n+t} : t \geqslant 0)$ is an identical stochastic replica of $(X(t) : t \geqslant 0)$. Therefore, it follows that $(\tau_n : n \in \mathbb{N})$ is a sequence of renewal instants.

Denoting the state of the process at the nth jump by X_n, we recall that a continuous time Markov chain can be equivalently represented by random holding times $(H_j : j \in \mathcal{X})$ and the jump-chain $(X_n : n \in \mathbb{Z}_+)$. The holding times are *i.i.d.*exponential random variables with $\mathbb{E}H_j = 1/(\lambda_j + j\mu)$, the transition probabilities [15] of the jump chain are

$$p_{j,j+1} = \frac{\lambda_j}{\lambda_j + j\mu}, \qquad p_{j,j-1} = \frac{j\mu}{\lambda_j + j\mu}.$$

Theorem 6. *The limiting mean revenue rate for the M/M/K/K queueing system is*

$$R(K, \underline{P}) = \sum_{k=0}^{K-1} \pi_k \lambda_k p_k.$$

Proof: Let R_n be the revenue earned in the nth renewal interval $[\tau_{n-1}, \tau_n)$. Since the revenue depends only on the state and not on time, it follows that the random sequence $((R_n, \tau_n - \tau_{n-1}) : n \in \mathbb{N})$ is *iid*. Applying renewal reward theorem [15] to this sequence, we get

$$\lim_{t\to\infty} \frac{R(t)}{t} = \frac{\mathbb{E}R_1}{\mathbb{E}\tau_1} \text{ a.s..} \tag{2}$$

Note that an arrival in state j indicates that an incoming customer in state j accepted the price p_j. Hence, the total revenue earned by the system in one renewal interval equals the weighted sum of the number of arrivals in state j, weighted by the price p_j of entering the system in this state. We indicate the number of visits to state j in first renewal interval $[0, \tau_1)$ by N_j. Then, the mean revenue in first renewal interval is

$$\mathbb{E}R_1 = \sum_{j=0}^{K-1} p_j \mathbb{E}N_j p_{j,j+1}. \tag{3}$$

Recall that the mean time spent in state j in first renewal interval is $\mathbb{E}N_j\mathbb{E}H_j$ from independence of jump-chain and holding times. Further, the equilibrium fraction of time spent in state j is $\pi_j = \frac{\mathbb{E}N_j\mathbb{E}H_j}{\mathbb{E}\tau_1}$ from another application of renewal reward theorem. The result follows from substituting Eq. (3) in RHS of Eq. (2), and observing that

$$\frac{\mathbb{E}N_j}{\mathbb{E}\tau_1} p_{j,j+1} = \frac{\pi_j}{\mathbb{E}H_j} p_{j,j+1} = \pi_j \lambda_j.$$

∎

IV. Uniform Pricing

We first consider the simpler case of uniform pricing for the cases when $K = \infty$ and K is finite. We show that the uniform pricing is optimal for infinite servers, and strictly sub-optimal for finite server case.

A. Infinite servers

We first consider the case when $K = \infty$. This models the case of large server systems with low customer arrival rates, such that the probability of all servers being busy is vanishingly small. For this case, we can find the equilibrium distribution of system state for general pricing scheme, and mean revenue rate for uniform pricing in closed form.

Corollary 7. *The stationary distribution of the Markov chain $X(t)$ for load factor $\rho = \frac{\lambda}{\mu}$ and $K = \infty$ is given by*

$$\pi_k = \begin{cases} \pi_0 \frac{\rho^k}{k!} \prod_{j=0}^{k-1} \overline{G}(p_j), & k \in \mathbb{N}, \\ \left[1 + \sum_{k\in\mathbb{N}} \frac{\rho^k}{k!} \prod_{j=0}^{k-1} \overline{G}(p_j)\right]^{-1}, & k = 0. \end{cases} \tag{4}$$

Remark 8. Observe that since the $\overline{G}$ are probabilities, the infinite series is bounded above by $\sum_{n\in\mathbb{Z}_+} \frac{\rho^n}{n!}$ which converges to $\exp(\rho)$, and hence the series converges. The corresponding revenue rate may be denoted by $R(\infty, \underline{P})$.

Corollary 9. *For an $M/M/\infty$ system with uniform pricing $\underline{P} = p\mathbf{1}$, the limiting mean revenue rate is*

$$R(\infty, p\mathbf{1}) = \lambda \overline{G}(p) p.$$

Proof: For uniform pricing $p_k = p$ and arrival rate $\lambda_k = \lambda\overline{G}(p)$ for all states $k \in \mathbb{N}$. Therefore, the result follows from Theorem 6. ∎

Proposition 10. *The optimal uniform pricing is the best pricing scheme for infinite servers.*

Proof: Let $p^*_\infty = \arg\max p\overline{G}(p)$ be a maximizer of mean revenue for uniform pricing scheme. For the load factor $\rho = \frac{\lambda}{\mu}$, we can write the mean revenue rate for any price vector $\underline{P}$ as

$$R(\infty, \underline{P}) = \lambda \sum_{j\in\mathbb{Z}_+} \pi_j \overline{G}(p_j) p_j \leqslant \lambda p^*_\infty \overline{G}(p^*_\infty) = R(\infty, p^*_\infty \mathbf{1}).$$

The result follows, since the uniform pricing vector is one possible price vector, and therefore $R(K, p^*_\infty\mathbf{1}) \leqslant R(K, \underline{P})$ for all K. ∎

B. Finite servers

Lemma 11. *For a finite K-server system with uniform pricing $\underline{P} = p\mathbf{1}$, the limiting mean revenue rate is*

$$R(K, p\mathbf{1}) = \lambda p \overline{G}(p) \left(1 - \frac{\frac{\rho^K \overline{G}(p)^K}{K!}}{1 + \sum_{k=1}^K \frac{\rho^k \overline{G}(p)^k}{k!}}\right).$$

Proof: For uniform pricing across states, i.e. $p_k = p$ for all states $k \in \{0, \ldots, K-1\}$, we obtain the steady state

probabilities π from Theorem 5 in terms of load factor ρ and $\overline{G}(p)$ to be

$$\pi_k = \pi_0 \frac{\rho^k \overline{G}(p)^k}{k!}, \qquad \pi_0 = \left(\sum_{k=0}^{K} \frac{\rho^k \overline{G}(p)^k}{k!} \right)^{-1}.$$

For uniform pricing, the arrival rate $\lambda_k = \lambda \overline{G}(p)$ also remains constant for all states $k \in \{0, \ldots, K-1\}$. Using this observation in Theorem 6, we obtain the mean revenue rate as

$$R(K, p\mathbf{1}) = \lambda p \overline{G}(p) \sum_{k=0}^{K-1} \pi_k = \lambda p \overline{G}(p)(1 - \pi_K).$$

Substituting the equilibrium distribution for uniform pricing in the above expression for mean revenue rate, we obtain the desired result. ∎

From the Assumption 1 on distribution G, the function $p\overline{G}(p)$ has a unique maximum. To find the optimal uniform price p for finite server systems, we need to understand how π_K changes with price p. For a fixed load ρ, we see that $\frac{\rho^k}{k!}\overline{G}(p)^k$ is monotonically decreasing in price p and state k. Further, we see that

$$\frac{1}{\pi_K} = \frac{\sum_{k=0}^{K} \frac{\rho^k \overline{G}(p)^k}{k!}}{\frac{\rho^K \overline{G}(p)^K}{K!}} = \sum_{k=0}^{K} \binom{K}{k} k! \rho^{-k} \overline{G}(p)^{-k}.$$

That is, the stationary probability $1 - \pi_K$ of K busy servers is monotonically increasing with uniform price p. Therefore, from Assumption 1, it follows that there exists a unique price p_K^* that maximizes revenue, i.e.

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

Lemma 12. *Let p_∞^* and p_K^* be maximizing prices of infinite and finite K-server systems with uniform pricing, that is*

$$p_\infty^* = \arg\max_{p>0} p\overline{G}(p), \quad p_K^* = \arg\max_{p>0} p\overline{G}(p)(1 - \pi_K).$$

Then, $p_K^* \geq p_\infty^*$.

Proof: Let $\pi(p_K^*)$ and $\pi(p_\infty^*)$ be the equilibrium distribution for K-server system with optimal uniform price p_K^* and p_∞^* respectively. Then, it is clear from the definition that

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

That is, we observe that $\pi_K(p_\infty^*) \geqslant \pi_K(p_K^*)$. Result follows from monotonic decrease of π_K with respect to price p. ∎

Lemma 13. *Let p_∞^* and p_K^* be maximizing prices of infinite and finite K-server systems with uniform pricing, and let $\underline{P}_K^*$ be the maximizing price vector for the K server system. Further, if we denote the steady state distribution of the K-server system with uniform price p_∞^* to be $\pi(p_\infty^*)$, then*

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

Proof: Since all possible price vectors include the case of uniform pricing as well, the first inequality follows from the increase in supremum over larger sets. For the second inequality, we observe that for any price vector $\underline{P}$ with corresponding equilibrium distribution π, we have

$$R(K, \underline{P}) = \lambda \sum_{k=0}^{K-1} \pi_k p_k \overline{G}(p_k) \leqslant \lambda p_\infty^* \overline{G}(p_\infty^*).$$

Hence, it follows that $R(K, \underline{P}_K^*) \leqslant \lambda p_\infty^*(p_\infty^*)$. Multiplying both sides by $(1 - \pi_K(p_\infty^*))$, we obtain

$$\begin{aligned}&(1 - \pi_K(p_\infty^*))R(K, \underline{P}_K^*) \leqslant \lambda(1 - \pi_K(p_\infty^*))p_\infty^* \overline{G}(p_\infty^*) \\ &\leqslant \lambda(1 - \pi_K(p_K^*))p_K^* \overline{G}(p_K^*) = R(K, p_K^*\mathbf{1}).\end{aligned}$$

∎

Corollary 14. *If the scaled load factor $c \triangleq \frac{\rho}{K}$, then*

$$R(K, \underline{P}_K^*) \leqslant (1 + c)R(K, p_K^*\mathbf{1}).$$

Proof: The result follows from the following observation,

$$\begin{aligned}\frac{1}{(1 - \pi_K(p_\infty^*))} &= 1 + \frac{\pi_K(p_\infty^*)}{\sum_{k=0}^{K-1} \pi_k(p_\infty^*)} \\ &= 1 + \frac{\frac{\rho}{K}\overline{G}(p_\infty^*)}{\sum_{k=0}^{K-1} \binom{K-1}{k} k! \rho^{-k} \overline{G}(p_\infty^*)^{-k}} \leqslant 1 + c.\end{aligned}$$

The last inequality follows by noting that the numerator corresponds to the $k = 0$ term in the denominator. ∎

V. Asymptotic Behavior of Revenue Rate

A direct comparison of the uniform price $p\mathbf{1}$ versus the differential price $\underline{P}$ is not easy. In this section, we make an asymptotic comparison as $\lambda \to \infty$.

Lemma 15. *For a fixed pricing policy vector $\underline{P} \in \mathbb{R}_+^K$,*

$$\lim_{\lambda \to \infty} R(K, \underline{P}) = \mu K p_{K-1}.$$

Proof: From the expressions for equilibrium distribution π in Theorem 5 and mean revenue rate in Theorem 6, for K server system with price vector $\underline{P}(K)$, we can write the limit of mean revenue rate as λ grows large, as

$$\begin{aligned}\lim_{\lambda \to \infty} \sum_{k=0}^{K-1} \pi_k \lambda p_k \overline{G}(p_k) &= \lim_{\lambda \to \infty} \frac{\lambda \sum_{k=0}^{K-1} \frac{\rho^k}{k!} p_k \prod_{j=0}^{k} \overline{G}(p_j)}{1 + \sum_{k=1}^{K} \frac{\rho^k}{k!} \prod_{j=0}^{k-1} \overline{G}(p_j)} \\ = \lim_{\lambda \to \infty} \frac{\lambda \frac{\rho^{K-1}}{(K-1)!} p_{K-1} \prod_{j=0}^{K-1} \overline{G}(p_j)}{\frac{\rho^K}{K!} \prod_{j=0}^{K-1} \overline{G}(p_j)} &= \mu K p_{K-1}.\end{aligned}$$

∎

The Lemma implies that for the uniform price $p\mathbf{1}$, the mean revenue rate $R(K, p\mathbf{1})$ goes to $Kp\mu$ as the arrival rate λ grows large. These are for fixed price policies; is it possible to vary the pricing policy as a function of λ and get the revenue rate to go to infinity?

Lemma 16. *If the value distribution $\overline{G}$ has support over $[0, \infty)$ is invertible, then $\lim_{\lambda \to \infty} R(K, p\mathbf{1}) = \infty$ for uniform price $p = \overline{G}^{-1}(\frac{1}{\lambda})$.*

Proof: Substituting $\overline{G}(p) = \frac{1}{\lambda}$ in the expression for mean revenue rate, we obtain

$$R(p\mathbf{1}, K) = \lambda p \overline{G}(p) \sum_{k=0}^{K-1} \pi_k(p) = \overline{G}^{-1}\Big(\frac{1}{\lambda}\Big) \frac{\sum_{k=0}^{K-1} \frac{1}{k! \mu^k}}{\sum_{k=0}^{K} \frac{1}{k! \mu^k}}.$$

The result follows from taking limit λ growing arbitrarily large, and observing that $\lim_{x\to 0}\overline{G}^{-1}(x)=\infty$. ∎

Thus, using prices dependent on the arrival distribution, but independent of number of servers, one can drive the revenue rate to infinity in the asymptotic regime as λ grows arbitrarily large. Since $\overline{G}^{-1}(\frac{1}{\lambda})$ increases as λ increases, we see that to extract maximum revenue, the price should be made as high as possible in heavy traffic limit. Using the following two examples, we show that the rate at which price grows is a function of the distribution.

Example 17. If the value distribution is Pareto, i.e. $\overline{G}(x)=\frac{\theta}{x}\mathbb{1}_{\{x\geqslant\theta\}}$, then for the choice of price $p(\lambda)=\overline{G}^{-1}(\frac{1}{\lambda})=\lambda\theta$, that grows linearly with λ, the revenue rate grows arbitrarily large with increase in λ.

Example 18. Consider the value distribution of the form $\overline{G}(x)=c_1e^{-c_2x^2}$. Taking 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 $\lim_{\lambda\to\infty}R(K,p\mathbf{1})=0$.

VI. Optimal Pricing for Finite Servers

We frame the optimal pricing problem as a continuous time Markov decision problem [16, Chapter 5]. We derive optimal prices and also analyze their dependence on various parameters, e.g., the number of servers, job arrival rate, and service rate.

A. The MDP formulation

As in Section II 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 $u\in\mathbb{R}_+$. The sojourn times in various states are independent exponentially distributed random variables depending on the controls applied on transitions to those states. More precisely, the sojourn times in a state i, for price u, are exponentially distributed with parameters $\nu_i(u)=i\mu+\lambda\bar{g}(u)\mathbb{1}_{\{i\in\mathcal{X}'\}}$. 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 $i\in[K-1]$

$$p_{ij}(u)=\frac{\lambda\bar{g}(u)}{\nu_i(u)}\mathbb{1}_{\{j=i+1\}}+\frac{i\mu}{\nu_i(u)}\mathbb{1}_{\{j=i-1\}}. \tag{5}$$

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

$$g(i,u)=u\mathbb{1}_{\{i=0\}}+\frac{\lambda u\bar{g}(u)}{\nu_i(u)}\mathbb{1}_{\{i\in[K-1]\}}. \tag{6}$$

B. Uniformization of continuous time Markov chain

The Bellman's equation for the average reward problem in Section VI-A takes the following form for all i

$$h(i)=\max_u\left\{g(i,u)-\frac{\theta}{\nu_i(u)}+\sum_{j=0}^{K}p_{ij}(u)h(j)\right\}. \tag{7}$$

Here $h(i)$, for each i, has interpretation of a relative or differential reward and θ is the optimal average reward per stage, independent of the initial state (see [16, Section 4.1]). Defining $\Lambda\triangleq K\mu+\lambda$, 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 one with uniform transition rate Λ by allowing fictitious self transitions such that the resulting dynamics remains unchanged. Specifically, we redefine state transition probabilities as follows. For all states $i\in\mathcal{X}$ and control $u\in R_+$,

$$\tilde{p}_{ij}(u)=p_{ij}(u)\frac{\nu_i(u)}{\Lambda}\mathbb{1}_{\{j\neq i\}}+\Big(1-\frac{\nu_i(u)}{\Lambda}\Big)\mathbb{1}_{\{j=i\}}. \tag{8}$$

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

$$\tilde{h}(i)=\max_u\left\{g(i,u)\nu_i(u)-\theta+\sum_{j=0}^{K}\tilde{p}_{ij}(u)\tilde{h}(j)\right\}. \tag{9}$$

Remark 19. The Bellman's equations (7) and (9) are equivalent. In particular, a pair (θ,h) satisfies (7) if and only if the pair $(\theta,\tilde{h})$ satisfies (9), 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 (7) and (9)) are identical.

Remark 20. Defining the difference $\Delta(i)\triangleq\frac{(\tilde{h}(i)-\tilde{h}(i+1))}{\Lambda}$ for all $i\in\mathcal{X}'$, and substituting in Eq. (9), along with expressions for $g(i,u)$ from Eq. (6), and $\tilde{p}_{ij}(u)$ from Eq. (8), we get

$$\theta=\lambda\max_u\left\{\overline{G}(u)(u-\Delta(i))\right\}\mathbb{1}_{\{i\in\mathcal{X}'\}}+i\mu\Delta(i-1),\ i\in\mathcal{X}. \tag{10}$$

C. Auxiliary maps

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

$$f(B,u)\triangleq(u-B)\bar{g}(u),\ B,u\in\mathbb{R}. \tag{11}$$

We can see that $f(B,u)-u=-B\overline{G}(u)-uG(u)\leqslant 0$ for all $u,B\geqslant 0$.

Remark 21. From Assumption 1, the function $f(B,u)$ has a unique maximizer in u for all $B\in\mathbb{R}$. Hence, we can define the maps $u^*:\mathbb{R}\to\mathbb{R}$ and $m:\mathbb{R}\to\mathbb{R}$ such that

$$u^*(B)\triangleq\arg\max_u f(B,u),\qquad m(B)\triangleq f(B,u^*). \tag{12}$$

Lemma 22. *Following statements are true for m and u^*.*

(a) m is non-negative and decreasing in B.
(b) m is continuous and convex function of B.
(c) u^* is non-decreasing in B.

Proof: Let m and u^* be as defined in Eq. (12).

(a) Since $f(B,u)=0$ at $u=B$, it follows that $m(B)\geqslant 0$ for all B. Let $B_1<B_2$, and $u_i=u^*(B_i)$ for $i\in[2]$. Then, we can write

$$m(B_2)=(u_2-B_2)\bar{g}(u_2)<(u_2-B_1)\bar{g}(u_2)\leqslant m(B_1).$$

(b) For $B_1<B_2$, we can write

$$\begin{aligned}m(B_1)&=(u_1-B_2)\bar{g}(u_2)+(B_2-B_1)\bar{g}(u_1)\\&\leqslant m(B_2)+(B_2-B_1).\end{aligned}$$

Therefore, we see that $|m(B_1) - m(B_2)| \leqslant |B_1 - B_2|$, implying continuity of m. Finally, $f(B,u)$ is linear, and hence, convex in B. Hence m is also convex in B.

(c) We can add the inequalities $f(B_1,u_2) \leqslant f(B_1,u_1) = m(B_1)$ and $f(B_2,u_1) \leqslant f(B_2,u_2) = m(B_2)$, to get

$$(B_2 - B_1)(\bar{g}(u_2) - \bar{g}(u_1)) \leqslant 0.$$

This implies that if $B_2 < B_1$, then $\bar{g}(u_2) \leqslant \bar{g}(u_1)$. The monotonic decrease of $\overline{G}$ implies that $u_2 \geqslant u_1$. ■

D. The optimal pricing

In terms of the map m, we can re-write the Eq. (10) as

$$m(\Delta(i))\mathbb{1}_{\{i\in\mathcal{X}'\}} + \frac{i\mu}{\lambda}\Delta(i-1) = \frac{\theta}{\lambda},\ i \in \mathcal{X}. \tag{13}$$

Observe that if $\Delta^*(i)$ solve Eq. (13) then the control $p_i^* \triangleq u^*(\Delta^*(i))$ achieving $m(\Delta^*(i))$ in (12) is the optimal control in each state $i \in \mathcal{X}'$.

Lemma 23. *Let $(\theta, \Delta(i), i \in \mathcal{X}')$ be a solution to* (13) *and $\underline{P}_K^* = (P_0^*, \ldots, P_{K-1}^*) \in \mathbb{R}_+^{\mathcal{X}'}$ be the optimal price vector. Then*

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

Proof: We assume the Lemma hypothesis.

(a) The non-negativity of θ follows from Eq. (13) for $i = 1$, and the non-negativity of m from Lemma 22.

(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. (13)

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

From monotone decrease of m and the induction step, it follows that $\Delta(i) \leqslant 0$ for all $i \in \mathcal{X}'$. In particular, we get the contradiction that $\Delta(K-1) = \frac{\theta}{K\mu} \leqslant 0$. Hence we see that $\Delta(0) > 0$.

From Eq. (13) and positivity of $\Delta(0)$, we observe that

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

Since 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. (13)

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

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 22(c). ■

Next, we will focus on solving Eq. (13). We give an iterative algorithm to obtain θ, which can then be used to obtain $\Delta(i)$ and also the optimal control p_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. (13) as

$$\theta = \lambda m(g_0(\theta)), \quad g_{i-1}(\theta) = \frac{\theta - \lambda m(g_i(\theta))\mathbb{1}_{\{i\in\mathcal{X}'\}}}{i\mu},\ i \in [K]. \tag{14}$$

Let us also consider the following iterative algorithm that generates two sequences $(\underline{\theta}_k, k \in \mathcal{X})$ and $(\bar{\theta}_k, k \in \mathcal{X})$ starting with $\underline{\theta}_0 = 0$ and $\bar{\theta}_0 = \lambda m(g_0(0))$, respectively.

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 24. *(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 the Eq. (14).

(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 22(a) and definition of g_{i-1} from Eq. (14), 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 [17, Theorem 2.1]. ■

The following figure illustrates variation of prices with the number of busy servers.

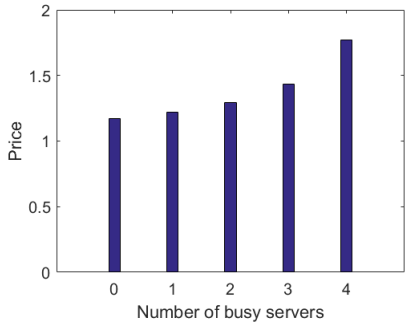

Fig. 2. State dependent optimal prices. We have set $K = 5, \lambda = 25$ and $\mu = 2$. Also, We have assumed exponential value distribution; $G(x) = 1 - \exp(-\beta x)$ with $\beta = 1$.

E. Properties of the Optimal Solution

We now analyze how the optimal prices and the optimal time average reward (or, the revenue rate) vary with various system parameters. We use the fact that the optimal revenue rate θ^* is solution to Eq. (14), from which we inductively derive property of g_i using the monotonic decrease of m from Lemma 22.

1) Varying Arrival Rate: We assume that we vary λ while keeping μ and K fixed.

Proposition 25. *(a) The revenue rate $\theta^*(\lambda)$ increases with λ. (b) The ratio $\theta^*(\lambda)/\lambda$ decreases with λ.*

Proof: Notice that $\theta^*(\lambda)$ is the solution to (14) as a function of λ, for a fixed μ.

(a) To begin with let us fix both θ and μ and vary λ. 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. (14) for $i = 0$, 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 Eq. (14) 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. ∎

2) Varying Service Rate: Here we assume that we vary μ while keeping λ and K fixed. Now we express the revenue rate as $\theta^*(\mu)$ to emphasize its dependence on μ.

Proposition 26. *(a) The revenue rate $\theta^*(\mu)$ increases with μ. (b) The ratio $\theta^*(\mu)/\mu$ decreases with μ.*

Proof: Consider the case when θ and λ remain fixed.

(a) From Eq. (14), 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. (14) 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. (14) 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. (14) 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. ∎

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

Proposition 27. *(a) The revenue rate $\theta^*(K)$ increases with K. (b) The ratio $\theta^*(K)/K$ decreases with μ. (c) For any $i < K$, the optimal price $P_i^*(K)$ is non-increasing with K.*

Proof: We define functions

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

Following similar arguments as in the proof of Theorem 24(a) we can iteratively show that $\bar{g}_i(\theta)$ are decreasing in θ for all $i < K$.

(a) It follows that $\mu m(\bar{g}_0) = \theta$. and $\bar{g}_{i-1} = (\theta - \lambda m(\bar{g}_i))/i\mu$ for $i \in [K-1]$. From Eq. (14) the optimal average reward $\theta^*(K)$ is the solution to the fixed point equation $\theta = K\mu\bar{g}_{K-1}(\theta))$. From Lemma 23(b), we have $\bar{g}_i(\theta) > \bar{g}_{i-1}(\theta)$ for all $\theta \geqslant 0$ and $i \in \mathcal{X}'$. 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 23(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 inquality 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 23(b) implies that $\bar{g}_i$ increases in i for a fixed argument. Together with inductive and initial hypothesis, we obtain

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

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

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

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 22(c). The result follows from combining these three observations. ∎

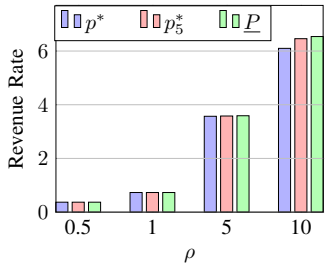

Fig. 3. Revenue rate as a function of load, for 5 servers

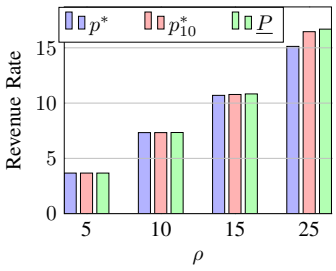

Fig. 4. Revenue rate as a function of load, for 10 servers

Remark 28. Let there be infinitely many servers, i.e., $K = \infty$. We can easily see that $\theta = \lambda m(0)$ along with $\Delta(i) = 0$ for all $i \in \mathbb{Z}_+$ satisfy Eq. (13). 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 Section IV-A.

VII. Numerical Evaluation

We compare differential pricing and uniform pricing. For this, we consider a 5-server system, with $\mu = 2$. For different values of load ρ, we compare the optimal revenue under uniform pricing with price p^*, uniform optimal pricing p_K^*, and optimal differential pricing $\underline{P}$ (obtained using Algorithm 1). The value function is assumed to be exponential with parameter 1. The resultant values are displayed in Figure 3.

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. A similar effect is seen in the case of 10 servers as well, as seen in Figure 4 (all other parameters remaining same). Beyond $\rho = 10$, differential pricing begins to outperform uniform pricing. In the following table, we also study how quickly the optimal differential revenue for a finite server system converges to the optimal revenue with infinite servers. We fix $\lambda = 1$ and $\mu = 2$. The value distribution is exponential with parameter 1. It is clear that the infinite server optimal revenue, $R(\infty, p^*\mathbf{1}) = 7.36$. In Figure 5 below, we display the optimal revenue under differential pricing, as we vary the number of servers. With as few as 10 servers, we come close to the infinite server revenue. However, note that this number will be a function of the arrival rate.

VIII. Conclusion

We consider a K server system that admits customers until all the servers are busy, with a state-dependent service charges.

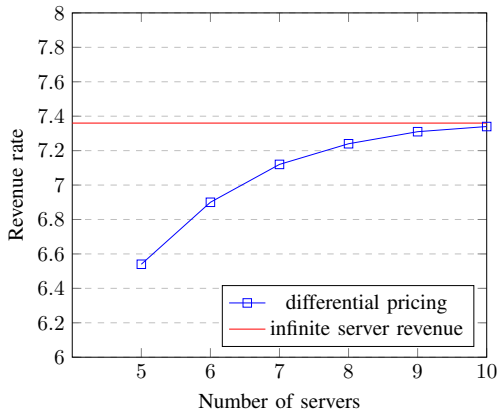

Fig. 5. Revenue as a function of number of servers

Assuming Poisson arrival for customers with *i.i.d.*service times and *i.i.d.*service valuation, we find the optimal service pricing that maximizes the server system's revenue rate. We show that the optimal pricing is uniform for an infinite server system, whereas it is increasing with number of busy servers for a finite server system.

References

[1] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," *ACM SIGCOMM computer communication review*, vol. 39, no. 1, pp. 68–73, 2008.

[2] P. Naor, "The regulation of queue size by levying tolls," *Econometrica: journal of the Econometric Society*, pp. 15–24, 1969.

[3] N. M. Edelson and D. K. Hilderbrand, "Congestion tolls for poisson queuing processes," *Econometrica: Journal of the Econometric Society*, pp. 81–92, 1975.

[4] C. Larsen, "Investigating sensitivity and the impact of information on pricing decisions in an M/M/1/∞ queueing model," *International journal of production economics*, vol. 56, pp. 365–377, 1998.

[5] R. Hassin, "Consumer information in markets with random product quality: The case of queues and balking," *Econometrica: Journal of the Econometric Society*, pp. 1185–1195, 1986.

[6] R. Hassin and M. Haviv, *To queue or not to queue: Equilibrium behavior in queueing systems*. Springer Science & Business Media, 2003, vol. 59.

[7] H. Chen and M. Z. Frank, "State dependent pricing with a queue," *Iie Transactions*, vol. 33, no. 10, pp. 847–860, 2001.

[8] C. Borgs, J. T. Chayes, S. Doroudi, M. Harchol-Balter, and K. Xu, "The optimal admission threshold in observable queues with state dependent pricing," *Probability in the Engineering and Informational Sciences*, vol. 28, no. 1, pp. 101–119, 2014.

[9] H. Xu and B. Li, "A study of pricing for cloud resources," *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 4, pp. 3–12, 2013.

[10] C. Wu, R. Buyya, and K. Ramamohanarao, "Cloud pricing models: Taxonomy, survey, and interdisciplinary challenges," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1–36, 2019.

[11] J. Gao, K. Iyer, and H. Topaloglu, "When fixed price meets priority auctions: Competing firms with different pricing and service rules," *Stochastic Systems*, vol. 9, no. 1, pp. 47–80, 2019.

[12] J. R. Norris, *Markov Chains*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

[13] F. P. Kelly, *Reversibility and Stochastic Networks*. USA: Cambridge University Press, 2011.

[14] T. M. Liggett, *Continuous Time Markov Processes: An Introduction*, ser. Graduate studies in mathematics. American Mathematical Society, 2010.

[15] S. M. Ross, *Introduction to Probability Models*, twelfth ed. Academic Press, 2019.

[16] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, USA: Athena Scientific, 2007, vol. 2.

[17] R. Divya, A. P. Azad, and C. Singh, "Fair and optimal mobile assisted offloading," in *WiOpt*, Paris, France, May 2017, pp. 1–8.