A System with a Choice of Highest-Bidder-First and FIFO Services

Tejas Bodas
Dept. of Electrical Engg.
IIT Bombay INDIA
tejaspbodas@ee.iitb.ac.in

Murtuza Ali
Dept. of Electrical Engg.
IIT Bombay INDIA
murtaza@ee.iitb.ac.in

D. Manjunath
Dept. of Electrical Engg.
IIT Bombay INDIA
dmanju@ee.iitb.ac.in

ABSTRACT

Service systems using a highest-bidder-first (HBF) policy have been studied in queueing literature for various applications and in economics literature to model corruption. Such systems have applications in modern problems like scheduling jobs in cloud computing scenarios or placement of ads on web pages. However, using a HBF service is like using a spot market and may not be preferred by many users. For such users, it may be good to provide a simple scheduler, e.g., a FIFO service. Further, in some situations it may even be necessary that a free service queue operates alongside a HBF queue. Motivated by such a scenario, we propose and analyze a service system with a FIFO server and a HBF server in parallel. Arriving customers are from a heterogeneous population with different valuations of their delay costs. They strategically choose between FIFO and HBF service; if HBF is chosen, they also choose the bid value to optimize an individual cost. We characterize the Wardrop equilibrium in such a system and analyze the revenue to the server. We see that when the total capacity is fixed and is shared between the FIFO and HBF servers, revenue is maximised when the FIFO capacity is non zero. However, if the FIFO server is added to an HBF server, then the revenue decreases with increasing FIFO capacity. We also discuss the case when customers are allowed to balk.

Keywords

Highest-bidder-first, queueing theory, game theory, Wardrop equilibrium, balking, revenue maximization.

1. INTRODUCTION

Consider a service system with a single server into which customers from a heterogeneous population arrive, wait in a queue if the server is busy and depart after receiving service. The heterogeneity of customers is captured by different costs per unit delay. In such a scenario the server can increase its revenue by providing differential service with different prices for different grades of service. One way of providing such differential service is to let arriving customers purchase their priority and the server use priority scheduling with higher prices being accorded higher priorities. The arriving customers are essentially placing a bid for the priority level that they want. This model has applications to several systems where bids can be placed and a scheduling mechanism based on the bid values can be used, e.g., [1, 5] suggest such a model for the spot market in cloud computing services.

Bidding for priorities has also been called bribing in some of the early literature. Queueing systems in which customers can purchase priorities have been studied in [14, 10, 13, 2, 7, 12]. The bribing model for priority in queues, introduced in [12] is as follows. Customers arrive according to a Poisson process of rate λ and each arrival, without observing the queue occupancy, offers a bid for service that is independent of all other offers. Service discipline is highest-bidder-first (HBF) preemptive (or non preemptive) priority. For this system, the expected delay as a function of the offered price is derived in [12]. An immediate extension is when the bid value of each customer depends on its delay cost, i.e., there is a bidding policy mapping delay cost to a bid value. In this scenario, since customers have different valuations for their delays, a natural model would be to have selfish customers that will choose their bids to minimize their total cost, i.e., the sum of delay cost and price paid for priority. This leads to a game theoretic situation in which equilibrium (or stable) bidding policies, policies from which a customer has no incentive to deviate, are of interest.

In a single server system, [12] shows that in an equilibrium bidding policy the bids are an increasing function of delay cost; closed form expressions for equilibrium bidding policies are obtained in [13] and [10]. Providing prioritized service requires additional resources and, especially in cloud computing like systems, this may be at the cost of doing 'useful work.' Further, it is not always feasible to expect that customers will bid; they may prefer obtaining a FIFO service for free or at a fixed price. This motivates the first system that we consider—a two server system in which one server uses FIFO scheduling and arrivals to the second server have to bid for priorities. An arriving customer selfishly makes two decisions—the server to use and, if it chooses the HBF

Tejas Bodas is supported by an assistantship from the Bharti Centre for Communication at IIT Bombay. Murtuza Ali was with IIT Bombay and is now at TU Eindhoven. D Manjunath was supported by the Indo-French Centre for Applied Mathematics and by GANESH, an INRIA associates program.

VALUETOOLS 2014, December 09-11, Bratislava, Slovakia

DOI 10.4108/icst.valuetools.2014.258217

server, the value of its bid. For this system, we first obtain equilibrium strategies (routing and bidding) and show that the equilibrium routing policy is of threshold type; customers with delay cost above a threshold choose the prioritizing server while those below the threshold choose to wait for their turn in the FIFO queue and obtain a free service.

In the above system, an obvious interest is in the impact of having a parallel FIFO server on the total revenue. Numerical results show that if the total capacity is fixed and is split between the FIFO and HBF servers, then having a FIFO server increases revenue substantially; a formal proof has been elusive. On the other hand, if additional capacity is allocated to a FIFO server, then we formally show that the revenue will be less than that with a single HBF server for any type profile.

Finally, we briefly consider the case when arriving customers can balk, i.e., not join the queue. This can happen when the reward for receiving service is lower than the cost of obtaining it (sum of delay cost and the bid). Such a system with just the HBF server is analyzed in [13, 10] where it was shown that if customers bid strategically, those with delay costs above a threshold will balk. Our interest is to try to retain these balking customers and possibly increase the revenue by adding a free server; such an option is motivated by the argument that customers with lower delay costs will prefer the FIFO server which decreases congestion at the HBF server. This in turn incentivises some high delay cost customers to not balk and instead join for service at the HBF server. We present a preliminary analysis of such a balking system and compare its revenue to a system with only the HBF server.

The work of [11] is closely related to the work of this paper. In [11], bidding is a mechanism to self-regulate the arrival process into a single-server HBF queue and the focus is on homogeneous customers. The social welfare has customers balking if the value of service is lower than the total cost, sum of bid and waiting costs. It is shown when customers balk, at equilibrium, the social welfare objective coincides with the server profit maximizing objective. Further, when the service times are assumed exponential, it is shown that the service rate that maximizes the servers revenue is lower than the socially optimal service rate. These were shown to also be true when there are a discrete set of customer types. There are two main differences between the model of [11] and that of this paper. First, we focus much of our attention on the case when there is no reward for service and hence the customers cannot balk; the customer objective is to minimize its total cost. Second, we always assume heterogeneous customers with a continuum of classes.

There are also similarities between the model that we consider and that of [1]. The latter has an infinite server queue and a K-server queue in parallel. The K-server queue is a preemptive HBF server like the one that we have described and is the spot market. Since the infinite server queue has no waiting time, the charge for service from that queue can be high and hence attract customers with high waiting time costs. Further, the customers have a value for the service received and can balk (not join the queue) if this value is lower than the cost (sum of delay cost and charge for service). Thus in the revenue maximizing regime of [1] the arrival rate to the HBF queue is zero. In the model considered in this paper, customers with low delay costs use the free FIFO service and the revenue maximising regime shares the total capacity between the two queues suitably.

Finally, we remark that the highest-bid-first discipline allows for a continuum of prices; alternatively, one can allow a fixed number of priorities and fix the price of each of the priorities. Such systems are considered in [2, 14, 4]. Systems with multiple FIFO queues, each with its own server were considered in, among others, [9, 6] and such a system with an admission price for each queue was analyzed in [8]. We will not be interested in such systems in this paper.

The rest of the paper is organized as follows. In Section 2 we describe the notation, recap some results from literature and characterize properties of the revenue for a single HBF server. In Section 3 we first characterize the equilibrium routing and bidding policy when a FIFO and a HBF server are in parallel. Numerical results for the case when the total capacity is shared between the FIFO and the HBF server are presented. We then analyze a system in which the FIFO server is added to the HBF server. In Section 4, we analyze the system in which customers balk. We conclude with a discussion in Section 5.

2. NOTATION AND PRELIMINARIES

In this section we set up the notation and recap some results from literature for the single server system using the highest-bidder-first (HBF) priority discipline. For such a system we also characterize the revenue rate.

Customers arrive to a service system according to a homogeneous Poisson process of rate λ. Service times of customers are i.i.d. random variables with distribution $G(\cdot)$ and unit mean. Associated with each arriving customer is a random variable V, $0 \leq a \leq V \leq b < \infty$, representing its cost per unit delay. V are i.i.d. with distribution $F(v)$ that is absolutely continuous in (a, b). V is also called type of the customer and $F(v)$ is called the type profile. A single server serves with rate μ and utilization $\rho := \frac{\lambda}{\mu}$ using HBF non preemptive priority discipline. The type of a customer is private information but $G(\cdot)$, $F(v)$, λ and the service rates of the servers are assumed to be common knowledge. Further, the customer does not know its service time. Each customer is assumed to be infinitesimally small and does not affect the system dynamics on its own. Customers do not balk, i.e., all arriving customers receive service. Further, once a customer has made the choice of the queue and the bid if joining the HBF queue, then it cannot change either of these; it also does not renege and it leaves the system only after receiving the service.

Like in [10, 13, 7, 12]. we assume oblivious bidding, i.e., an arriving customer bids for its priority without observing the queue occupancy; service is non preemptive with higher priorities for higher bids. Thus on a service completion, the next customer is the one in the queue with the highest bid. Preemptive service can be analysed identically to the non preemptive case and we do not discuss it further.

$X(v)$ is the bidding policy, i.e., customers of type v bid $X(v)$.

$W(x)$ is the expected waiting time (time in queue excluding service time) of a customer that bids x. The expected total cost of receiving service for a customer of type v is

$$C(v) := X(v) + v\left(W(X(v)) + \frac{1}{\mu}\right).$$

Customers behave strategically and choose their bids to minimize their individual costs. A bidding policy is an *equilibrium policy* if no individual customer can unilaterally deviate from it and lower its total cost $C(v)$.

Let $B^X(x)$ denote the distribution of the bids under bidding policy X, $W^X(x)$ the expected waiting time (time in queue excluding service time) of a customer that bids x, and $C^X(v)$ the total cost to customer of type v under policy X. Let W_0 be the expected waiting time added to that of an arriving customer due to residual service time of a customer in service. W_0 is the product of the residual service time and probability that an arriving customer sees a busy server, i.e.,

$$W_0 = \frac{\lambda}{2}\int_0^\infty \tau^2 dG(\mu\tau).$$

Following [10], we obtain

$$X^E(v) = \int_0^v \frac{2\rho W_0 y}{(1-\rho+\rho F(y))^3}\, dF(y) \qquad (1)$$

as an equilibrium bidding policy. Property 1 below summarizes the properties of $X^E(v)$ that have been derived in [12], [13], [10]. Note that the $X(v)$ in the following is actually $X^E(v)$ but we drop the superscript E to simplify notation. Further, $B(\cdot)$, $W(\cdot)$, and $C(\cdot)$ depend on $X(v)$. Once again, we do not explicitly capture this dependence in the notation.

Property 1. *1. $X(v)$ is continuous and strictly increasing in v. Further $B(X(v)) = F(v)$.*

2. $W(v)$ is strictly decreasing in v. Further, $W(v) = \frac{\mu^2 W_0}{(\mu-\lambda(1-F(v)))^2}$ where W_0 is as above.

3. $C(v) := \min_x\left\{x + v\left(W^X(x) + \frac{1}{\mu}\right)\right\}$ is continuous and strictly increasing concave.

4. $\frac{dC(v)}{dv} = W(x) + \frac{1}{\mu}$ and $\frac{dX(v)}{dv} + v\frac{dW(v)}{dv} = 0$

Let $R^X(\lambda, F(v))$ be the revenue rate for a system with arrival rate λ, type profile $F(v)$, and an equilibrium bidding policy $X(v)$. Also, wherever applicable, to indicate the dependence on the service rate of the server, we also use the alternative notation of $R^X(\lambda, F(v), \mu)$ to denote the revenue rate when the server serves with rate μ. Clearly,

$$R(\lambda, F(v)) = \lambda\int_v X(v)dF(v).$$

The following lemma characterizes the equilibrium revenue function where we obtain the direction of change of the revenue when one of arrival rate, service rate and type profile is changed while keeping the other parameters unchanged.

Lemma 1. *1. If $\lambda_1 < \lambda_2$, then*

$$R(\lambda_1, F(v)) \;<\; R(\lambda_2, F(v)).$$

2. Let μ_1 and μ_2 be two service rates with $\mu_1 < \mu_2$. Then

$$R(\lambda, F(v), \mu_1) > R(\lambda, F(v), \mu_2).$$

3. Let $F_1(v)$ be a type profile and let $F_2(v) = F_1(v - v_0)$ with $v_0 > 0$. For this case

$$R(\lambda, F_1(v)) < R(\lambda, F_2(v)).$$

4. Let $F_2(v) = F_1(v/c)$ with $c > 1$. Then

$$R(\lambda, F_1(v)) < R(\lambda, F_2(v)).$$

5. Let $F_1(v)$ be a type profile and define

$$F_2(v) \;:=\; \begin{cases} 0 & v < a_1 \\ \frac{\int_{a_1}^v dF_1(w)}{\int_{a_1}^b dF_1(w)} & v \geq a_1 \end{cases}$$

Then $R(\lambda, F_1(v)) < R(\lambda, F_2(v))$.

We provide informal arguments and omit formal proofs as they are quite straightforward from the analysis in [12], [13], [10]. The first and second statements are complimentary—increasing the arrival rate or decreasing the service rate increases the revenue. This happens because increasing λ (or decreasing μ) introduces more congestion and more competition. Thus all customers have to bid higher to 'stay in place.' This is also evident in (1) where the denominator of the integrand increases with increasing λ and decreasing μ. For the case of increased λ there are also more arrivals. The third statement says that if the type profile is 'shifted to the right,' i.e., the delay cost of every customer is increased by a fixed quantity v_0, then the revenue rate is increased. This is because, every one has an increased delay cost and hence every one has to bid higher than before to, once again, stay in place. This is also seen in (1) where there is a y in the integrand and the integration range is over larger values of y. The fourth statement says that 'stretching' the type profile increases the revenue rate. Once again, every one bids higher than before and the reasoning is similar to the preceding case. The last statement concerns the case when arrival rate is kept constant but the types below a cutoff are removed from the population. This also causes every one to bid higher because they are all competing with more customers of the same type.

3. A SYSTEM WITH A FIFO AND A HBF SERVER

We now analyze a service system with two servers. One server uses the non preemptive HBF discipline and serves at rate μ_1. The second server uses the FIFO discipline and serves at rate μ_2. Customers choosing the HBF server will have to bid at least M, $M \geq 0$, while the FIFO service is free. Customers arrive according to a homogeneous Poisson process of rate λ and the type profile is $F(v)$. In this section we will assume that all arrivals will have to receive service from one of the two servers and they cannot balk. Thus an arriving customer now has to make the following decisions on arrival. Which server to use, and, if it chooses the HBF server, then the value of its bid. We will assume oblivious decisions, i.e., the arrivals make the choices without observing the queue occupancy. We now characterize the structure of these choices.

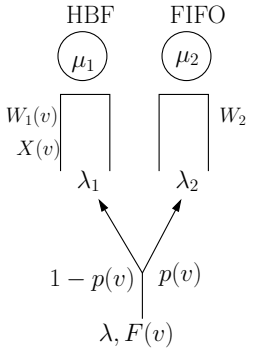

Figure 1: System with a HBF server and a FIFO server.

Consider the choices made by a customer of type v. First, let $p(v) : [a, b] \to [0, 1]$ be the *routing policy*, i.e., the probability that it chooses the FIFO server. Next, let $M + X(v)$ be the bid if it chooses the HBF server. With this, the expected total cost of receiving service for a customer of type v is

$$C(v) := X(v) + M + v\left(W(X(v)) + \frac{1}{\mu_1}\right).$$

The pair $(p(v), X(v))$ is the strategy of a customer of type v and it is an *equilibrium* strategy if no customer can unilaterally deviate and lower its $C(\cdot)$.

Given $(p(v), X(v))$, we see that $\lambda_2 := \lambda \int_0^\infty p(v)dF(v)$ is the arrival rate to the FIFO server and $\lambda_1 = \lambda - \lambda_2$ is the arrival rate to the HBF server. Define $\rho_i := \lambda_i/\mu_i$. All customers that choose the HBF server experience a bid-dependent waiting time, denoted by $W_1(v)$ for a customer of type v, and all those that choose the FIFO server experience the same expected waiting time, denoted by W_2. Let $D_1(v) := W_1(v) + \frac{1}{\mu_1}$ and $D_2 := W_2 + \frac{1}{\mu_2}$ be the expected sojourn times in, respectively, the HBF and the FIFO servers. Fig. 1 illustrates this notation.

Let us now consider the equilibrium strategy for this system. Since this is a non atomic system and all customers choose individually optimal strategies, the equilibrium is a Wardrop equilibrium, first described in [15] and used extensively in transportation systems. This means that at equilibrium the following is true.

$$\begin{array}{lll} \text{If } p(v) = 1, & \text{then} & vD_2 \le M + X(v) + vD_1(X(v)). \\ \text{If } p(v) = 0, & \text{then} & vD_2 \ge M + X(v) + vD_1(X(v)). \\ \text{If } 0 < p(v) < 1, & \text{then} & vD_2 = M + X(v) + vD_1(X(v)). \end{array} \tag{2}$$

The following theorem characterizes the equilibrium strategy for this system.

Theorem 1. *Using v_1 determined below, define $p^E(v)$, $F_1(v)$, and W_0 as follows.*

$$p^E(v) = \begin{cases} 0 & \text{for } v > v_1, \\ t & \text{for } v = v_1, \\ 1 & \text{for } v < v_1. \end{cases} \tag{3}$$

$$F_1(v) := \begin{cases} 0 & v < v_1 \\ \frac{\int_{v_1}^{v} dF(x)}{\int_{v_1}^{b} dF(x)} & v_1 \le v \le b, \\ 1 & v > b, \end{cases} \tag{4}$$

$$W_0 = \frac{\lambda_1}{2}\int_0^\infty \tau^2 dG(\mu_1\tau), \tag{5}$$

$$X^E(v) = \int_0^v \frac{2\rho_1 W_0 y}{(1-\rho_1+\rho_1 F_1(y))^3}\, dF_1(y) \tag{6}$$

For the routing and bidding policy $(p^E(v), X^E(v))$ determined as above, let $D_1(v)$ be the bid-dependent expected sojourn time in the HBF server and $D_2(\lambda_2)$ be the expected sojourn time in the FIFO server when the arrival rate to it is λ_2.

v_1 is determined as follows.

- *If using $v_1 = a$ in (3)–(6) satisfies $M + aD_1(a) < aD_2$, then set $v_1 = a$.*
- *Else if using $v_1 = b$ in (3)–(6) satisfies $M + bD_1(b) > bD_2$ then set $v_1 = b$.*
- *Else find v_1 which when used in (3)–(6) satisfies*

$$M + v_1 D_1(v_1) = v_1 D_2. \tag{7}$$

$(p^E(v), X^E(v))$ is an equilibrium strategy with v_1 defined as above. Further, v_1 is unique.

Proof. We will prove that the Wardrop conditions of (2) are satisfied for the choice of $p^E(v)$ and $X^E(v)$ as described in (3)–(6).

With $p^E(v)$ as in (3), the arrival rate to the HBF server is $\lambda_1 = \lambda \int_{v_1}^b dF(\tau)$. The type profile of the customers choosing the HBF server will be as in (4). From the previous section (and from [12, 13, 10]) we know that those that decide to join the HBF server will have to use the bidding policy as in (6) for individual optimisation.

With bidding policy as in (6) for those that choose the HBF server, we will now verify that $p^E(v)$ of (3) satisfies the Wardrop condition of (2). First consider the case when $v_1 \neq a, b$. In this case, for any $v < v_1$ we see that

$$M = v_1(D_2 - D_1(v_1)) > v(D_2 - D_1(v_1)).$$

This verifies the Wardrop condition of (2) for $v < v_1$ i.e., $p^E(v) = 1$ is the optimum choice for $v < v_1$.

For a customer of type $v_1 + \epsilon$ for some $\epsilon > 0$,

$$C(v_1 + \epsilon) \le C(v_1) + \epsilon \frac{dC(v_1)}{dv} = C(v_1) + \epsilon D_1(v_1)$$

The inequality is from the concavity of $C(v)$ (item (3) of Property 1) and the equality is from item (4) in Property 1.

For $v = v_1$, $X^E(v_1) = 0$ and the preceding inequality leads to

$$\begin{aligned} C(v_1+\epsilon) &\leq M + (v_1+\epsilon)D_1(v_1) \\ &= v_1 D_2 + \epsilon D_1(v_1) \ < \ v_1 D_2 + \epsilon D_2. \end{aligned} \quad (8)$$

Both the equality and inequality above follow from (7). After rearranging the terms we have

$$M + (v_1+\epsilon)D_1(v_1) \ < \ (v_1+\epsilon)D_2$$

which is substituted in (8) to get

$$C(v_1+\epsilon) < (v_1+\epsilon)D_2. \quad (9)$$

Hence a customer of type $v_1+\epsilon$ will have a lower cost with the HBF server than with the FIFO server. This verifies that $p(v_1+\epsilon) = 0$ is optimum for $v = v_1+\epsilon$. Since ϵ is arbitrary, $p^E(v) = 0$ for all $v > v_1$ is also verified.

For the cases when $v_1 = a$ (resp. $v_1 = b$) we can use similar arguments to show that $p^E(v) = 0$ (resp. $p^E(v) = 1$) is an equilibrium policy.

We will now show that v_1 satisfying (7) is unique. We need only consider the case when $v_1 \neq a, b$. Observe that $D_1(v_1)$ is a strictly decreasing function of v_1 while D_2 is strictly increasing in v_1. Further as $v_1 \neq a, b$ the following complementary conditions must be true.

$$\begin{aligned} D_1(a) + \frac{M}{a} &\geq D_2 \\ D_1(b) + \frac{M}{b} &\leq D_2 \end{aligned}$$

Hence there exists a unique v_1. □

Corollary 1. *In the class of routing policies $p(v)$ that results in $F_1(v)$ (the profile of the customers choosing the HBF server) being absolutely continuous, a Wardrop equilibrium routing policy is of the threshold type.*

Proof. From Theorem 1, $C(v_1) \leq v_1 D_2$ implies $C(v_1+\epsilon) < (v_1+\epsilon)D_2$ i.e., if a customer of type v joins the HBF server, then for a $\epsilon > 0$ a customer of type $v+\epsilon$ also joins this queue. □

Theorem 1 shows that the arrival rate to the HBF server can be less than λ and that the type profile of these customers is truncated on the left. From Lemma 1 the former reduces the revenue while the latter increases it. This prompts us to investigate the effect of the parallel FIFO server on total revenue. Two scenarios present themselves immediately. First, we can let the total service rate be fixed (say unity) and share it between the HBF and the FIFO servers, i.e., $\mu_1 + \mu_2 = 1$. The second scenario of interest is when $\mu_1 = 1$ and we add additional service capacity in the form of a FIFO queue of service rate μ_2. In both cases, we investigate the revenue as a function of μ_2.

3.1 HBF and FIFO Servers Sharing Capacity

We present a sample of the numerical results that we have obtained. We assume $M = 0$ and that the total service capacity of unity is shared between the HBF and FIFO servers, i.e., $\mu_1+\mu_2 = 1$. Service time distribution is exponential with unit mean and HBF server is non-preemptive. From (5),

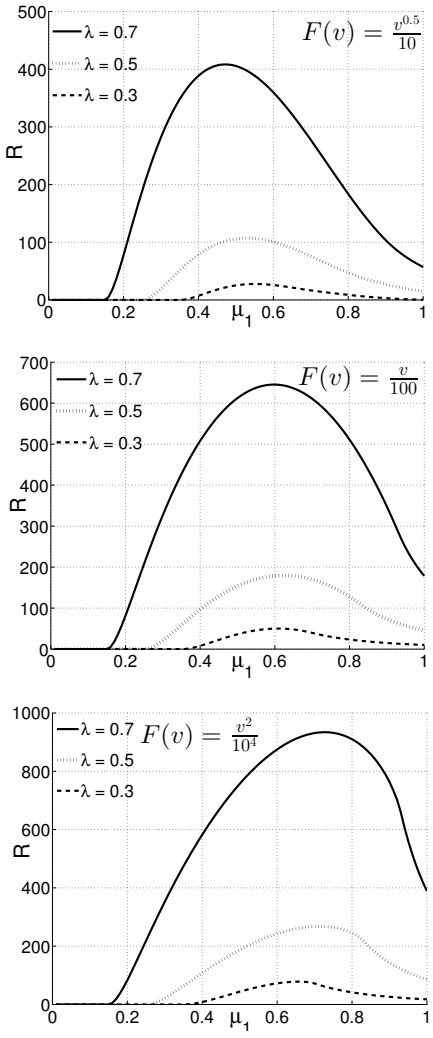

Figure 2: Revenue as a function of μ_1 for three different example $F(v)$ and different λ.

$W_0 = \frac{\lambda_1}{\mu_1^2}$. We consider the following three different $F(v)$, all of which correspond to $a = 0$ and $b = 100$.

1. $F(v) = \frac{v}{100}$ for $0 \le v \le 100$.
2. $F(v) = \frac{v^{0.5}}{10}$ for $0 \le v \le 100$
3. $F(v) = \frac{v^2}{10^4}$ for $0 \le v \le 100$

For these three examples, in Fig. 2 we plot the revenue as a function of μ_1 for $\lambda = 0.3, 0.5, 0.7$. The graphs are rather self explanatory and we will not dwell on the details, except pointing out the two observations the stand out from these examples.

1. The revenue can actually increase if some of the capacity is allocated to a FIFO server. This in itself is not very surprising because there is no balking in the system and all arriving customers have to take the service. That the magnitude of the increase was substantial seems surprising.
2. The HBF server needs a minimum μ_1 before it can generate any revenue. Once again, it appears that this minimum is substantial.

3.2 Adding a FIFO Server to an HBF Server

We now compare the revenue when a FIFO server is added to the HBF server. Let S_1 be a system with a single HBF server of rate μ_1 and S_2 be a system with a HBF server of rate μ_1 and a FIFO server of rate μ_2. Customers arrive according to a Poisson process into the system, they have a type profile $F(v)$ and make selfish choices of the server and the bid. In S_1, customers follow the strategy described in [10, 13] and in S_2 they follow that outlined in Theorem 1. We need additional notation.

In all quantities of interest a superscript S_i will indicate the system that is being referred. For example, $R^{S_i}(\cdot,\cdot)$, will be the revenue in system S_i. $F^{S_1}(v) = F(v)$ and $F^{S_2}(v) = F_1(v)$ will be the type profile of customers choosing the HBF server in systems S_1 and S_2 respectively. λ_i will denote the arrival rate to the HBF server in S_i. Clearly, $\lambda_1 = \lambda$ and $\lambda_2 = \lambda(1 - F(v_1))$ where v_1 is as obtained from Theorem 1. $D^{S_2}(v), C^{S_2}(v)$ will denote the sojourn time and the total cost respectively for a customer joining the HBF server of S_2.

We first show that the mean waiting time for customers of type $v > v_1$ is lower in S_2 than in S_1 and then characterize the revenue.

Lemma 2. *For a customer of type $v > v_1$, we have $W^{S_1}(v) > W^{S_2}(v)$. Further*

$$\frac{dW^{S_1}(v)}{dv} < \frac{dW^{S_2}(v)}{dv}.$$

Proof. From item 1 of Property 1, the bids are monotonic in v. Now consider a customer of type $\tilde{v} \ge v_1$. In both S_1 and S_2, the arrival rate of customers of priority $v > \tilde{v}$ to the HBF queue is the same. Thus the denominator in the expression for $W(v)$ (see item (2) in Property 1) is the same for both S_1 and S_2 but $W_0^{S_1} > W_0^{S_2}$. Hence $W^{S_1}(v) > W^{S_2}(v)$.

Now from the expression for $W(v)$ in item (2) in Property 1 we have for $i = 1, 2$

$$\frac{dW^{S_i}(v)}{dv} = \frac{-2\mu_1^2\lambda_i W_0^{S_i} dF^{S_i}(v)}{(\mu_1 - \lambda_i(1 - F^{S_i}(v)))^3}.$$

As earlier, the arrival rate of customers of priority $v > v_1$ to the HBF queue is the same i.e.,

$$\lambda_1(1 - F(v)) = \lambda_2(1 - F_1(v)).$$

Since $W_0^{S_1} > W_0^{S_2}$, and the remaining terms are the same for both S_1 and S_2 the lemma is true. □

This sets us up for the main result of this subsection.

Theorem 2. *Adding a FIFO server of service rate μ_2 does not increase the revenue i.e.,*

$$R^{S_1}(\lambda_1, F(v)) \ge R^{S_2}(\lambda_2, F_1(v)).$$

Proof. In the first part of the proof, we will prove by contradiction that for all $v > v_1$, $X^{S_1}(v) \ge X^{S_2}(v)$. Suppose the claim is not true and there exists a $v_2 > v_1$ such that

$$X^{S_1}(v_2) < X^{S_2}(v_2). \tag{10}$$

Since the bidding policy $X^{S_i}(v)$ is increasing in v,

$$X^{S_1}(v_1) \ge X^{S_2}(v_1) = 0. \tag{11}$$

Therefore from (10) and (11), there exists a v^* such that

$$X^{S_1}(v^*) = X^{S_2}(v^*) \tag{12}$$

where for all $v < v^*$,

$$X^{S_1}(v) > X^{S_2}(v).$$

Recall that for all $v > v_1$, from Lemma 2 we have

$$\frac{dW^{S_1}(v)}{dv} < \frac{dW^{S_2}(v)}{dv}.$$

Therefore, for all $v > v_1$, from item (4) in Property 1 we get

$$\frac{dX^{S_1}(v)}{dv} > \frac{dX^{S_2}(v)}{dv}. \tag{13}$$

From (12) and (13), for all $v \ge v^*$ we have

$$X^{S_1}(v) \ge X^{S_2}(v). \tag{14}$$

But this contradicts the assumption of (10) on v_2 as $v_2 > v^*$. This implies that for all v,

$$X^{S_1}(v) \ge X^{S_2}(v). \tag{15}$$

The revenue in S_2 is given by

$$\begin{aligned} R^{S_2}(\lambda_2, F_1(v)) &= \lambda_2 \int_{v_1}^{b} \left(M + X^{S_2}(v)\right) dF_1(v) \\ &= \frac{\lambda_2}{1-F(v_1)} \int_{v_1}^{b} \left(M + X^{S_2}(v)\right) dF(v) \\ &= \lambda \int_{v_1}^{b} \left(M + X^{S_2}(v)\right) dF(v) \\ &\leq \lambda \int_{v_1}^{b} \left(M + X^{S_1}(v)\right) dF(v) \\ &\leq \lambda \int_{a}^{b} \left(M + X^{S_1}(v)\right) dF(v) \\ &= R^{S_1}(\lambda_1, F(v)). \end{aligned}$$

The second equality is from the definition of $F_1(v)$ in (4), the third equality is from the definition of λ_2 and the first inequality is from (15). This completes the proof. □

4. WHEN ARRIVALS BALK

In this section we assume that each customer receives a fixed reward for obtaining the service from the system. In this case, a type v customer whose total cost of receiving service, $C(v)$, exceeds P has no incentive to join the system and will balk. In such a system the v^* satisfying $P = X(v^*) + v^*W(v^*) + M$ is the highest type of customer receiving service; all customers with $v > v^*$ will balk.

As in the previous section, we will compare two systems S_1 and S_2 and to simplify the analysis, we will assume $F(v) = v/b$ and $a = 0$. Let $v_{u,1}$ and $v_{u,2}$ be the highest type customer joining the HBF server of, respectively, S_1 and S_2. Let $F_1(v)$ and $F_2(v)$ denote the type profile and λ_1 and λ_2 denote the arrival rate to the HBF server in S_1 and S_2 respectively. Since the arrivals balk, we have

$$\begin{aligned} F_1(v) &= \frac{F(v) - F(a)}{F(v_{u,1}) - F(a)} \\ F_2(v) &= \frac{F(v) - F(v_1)}{F(v_{u,2}) - F(v_1)} \\ \lambda_1 &= \lambda(F(v_{u,1}) - F(a)) \\ \lambda_2 &= \lambda(F(v_{u,2}) - F(v_1)). \end{aligned}$$

We first state the following lemma that compares $v_{u,1}$ and $v_{u,2}$.

Lemma 3. $v_{u,1} \leq v_{u,2}$.

PROOF. The proof is by contradiction. Suppose $v_{u,1} > v_{u,2}$. Recall the balking condition

$$P = X(v^*) + \frac{v^*}{\mu_1} + M.$$

Since P, M and μ are constant and $v_{u,1} > v_{u,2}$ we have

$$X^{S_1}(v_{u,1}) < X^{S_2}(v_{u,2}). \tag{16}$$

At $v = v_1$,

$$X^{S_1}(v_1) > X^{S_2}(v_1) = 0. \tag{17}$$

Therefore from (16) and (17), there should exist a $v_2 \leq v_{u,2}$ which satisfies

$$X^{S_1}(v_2) = X^{S_2}(v_2). \tag{18}$$

From item (4) of Property 1, we have for $i = 1, 2$

$$\frac{dW^{S_i}(v)}{dv} = \frac{-2\mu_1^2\lambda_i W_0^{S_i} dF_i(v)}{(\mu_1 - \lambda_i(1 - F_i(v)))^3}.$$

From the definition of F_1, F_2, λ_1 and λ_2 we have

$$\frac{dW^{S_1}(v)}{dv} = \frac{-2\mu_1^2\lambda W_0^{S_1} dF(v)}{(\mu - \lambda(F(v_{u,1}) - F(v)))^3} \tag{19}$$

$$\frac{dW^{S_2}(v)}{dv} = \frac{-2\mu_1^2\lambda W_0^{S_2} dF(v)}{(\mu - \lambda(F(v_{u,2}) - F(v)))^3}. \tag{20}$$

Since $v_{u,1} > v_{u,2}$ and $a < v_1$, we have $\lambda_1 > \lambda_2$ and from the definition of W_0 we have

$$W_0^{S_1} > W_0^{S_2}. \tag{21}$$

Also $v_{u,1} > v_{u,2}$ implies

$$F(v_{u,1}) - F(v) > F(v_{u,2}) - F(v). \tag{22}$$

Therefore from (19), (20), (21) and (22),

$$\frac{dW^{S_1}(v)}{dv} < \frac{dW^{S_2}(v)}{dv}.$$

Substituting the above in item (4) of Property 1, we have

$$\frac{d(X^{S_1}(v))}{dv} \geq \frac{d(X^{S_2}(v))}{dv}. \tag{23}$$

From (18) and (23), we have, for all $v > v_2$

$$X^{S_1}(v) \geq X^{S_2}(v).$$

Therefore

$$X^{S_1}(v_{u,2}) \geq X^{S_2}(v_{u,2}).$$

Now since $X(v)$ is increasing, for $v_{u,1} > v_{u,2}$,

$$X^{S_1}(v_{u,1}) \geq X^{S_1}(v_{u,2}).$$

Hence

$$X^{S_1}(v_{u,1}) \geq X^{S_2}(v_{u,2}).$$

However this contradicts (16).

Hence $v_{u,1} \leq v_{u,2}$. □

Note that this lemma is true for arbitrary choice of $F(v)$ since the proof technique does not make use of $F(v) = v/b$. Having characterized the threshold v^* for the two system, we will now analyze their revenue for the case when $F(v) = v/b$.

First consider the revenue for S_1.

$$\begin{aligned} R^{S_1}(\lambda_1, F_1(v)) &= \lambda_1 \int_{0}^{v_{u,1}} \left(M + X^{S_1}(v)\right) dF_1(v) \\ &= \lambda_1 \int_{0}^{v_{u,1}} \frac{\left(M + X^{S_1}(v)\right)}{F(v_{u,1}) - F(a)} dF(v)) \\ &= \frac{\lambda}{b} \int_{0}^{v_{u,1}} \left(M + X^{S_1}(v)\right) dv. \end{aligned}$$

where the first and the second equality follow from the definition of $F_1(v)$ and λ_1 respectively.

Similarly the revenue for S_2 is given by

$$R^{S_2}(\lambda_2, F_2(v)) = \frac{\lambda}{b} \int_{v_1}^{v_{u,2}} \left(M + X^{S_2}(v) \right) dv.$$

The preceding arguments prove the following theorem that compares the revenue from the two systems.

Theorem 3. *If*

$$\int_0^{v_{u,1}} \left(M + X^{S_1}(v) \right) dv > \int_{v_1}^{v_{u,2}} \left(M + X^{S_2}(v) \right) dv,$$

then

$$R^{S_1}(\lambda_1, F_1(v)) > R^{S_2}(\lambda_2, F_2(v))$$

otherwise

$$R^{S_1}(\lambda_1, F_1(v)) \leq R^{S_2}(\lambda_2, F_2(v)).$$

5. DISCUSSION

The primary motivation for our models in this paper are from the need for new pricing and auction mechanisms in on-demand resource provisioning, e.g., cloud computing systems. An additional interest is, like in [12] and in [13], the economic aspects of bribing. While [13] assumed that the full capacity was auctioned, we investigate the economics of 'partial corruption' in which some of the population is provided an 'honest service' via the FIFO queue. In this setting M (the minimum bid) is interpreted as the 'social reward' to a customer for doing the right thing, i.e., for not bribing. The results of Fig. 2 provides the intuitively appealing interpretation that it might be more rewarding to be partially corrupt than to be fully corrupt. Further, 'small' levels of corruption does not pay.

There is of course more work to be done. Formal proofs for the numerical findings of Fig. 2, a better understanding of the system of Section 3.2, and analysing the balking system in more detail are on our immediate agenda.

An alternate system where the bid value determines the share of the server capacity in a processor sharing system has also been analysed by us in [3]. A comparitive understanding of these systems with direct application to service provisioning cloud computing and wireless communication systems are also of interest.

6. REFERENCES

[1] V. Abhishek, I. Kash, and P. Key. Fixed and market pricing for cloud services. In *Proceedings of Computer Communications Workshops (INFOCOM WORKSHOPS)*, 2012.

[2] I. Adiri and U. Yechiali. Optimal priority-purchasing and pricing decisions in nonmonopoly and monopoly queues. *Operations Research*, 22(5):1051–1066, September/October 1974.

[3] M. Ali, T. Bodas, and D. Manjunath. Optimal and equilibrium allocations in a Discriminatory Processor Sharing system. In *Proceedings of NETGCOOP*, 2014.

[4] H. Alperstein. Optimal pricing policy for the service facility offering a set of priority prices. *Management Science*, 34(5):666–671, May 1988.

[5] J. Altmann, C. Courcoubetis, G. D. Stamoulis, M. Dramitinos, T. Rayna, M. Risch, and C. Bannink. Gridecon: A market place for computing resources. In *Proceedings of the 5th International Workshop on Grid Economics and Business Models (GECON)*, pages 185–196, 2008.

[6] U. Ayesta, O. Brun, and B. Prabhu. Price of anarchy in non-cooperative load balancing. In *Proceedings of the IEEE INFOCOM*, pages 436–440, 2010.

[7] K. R. Balachandran. Purchasing priorities in queues. *Management Science*, 18(1):319–326, January 1972.

[8] T. Bodas, A. Ganesh, and D. Manjunath. Load balancing and routing games with admission price. In *Proceedings of the IEEE Conference on Decision and Control*, 2011.

[9] S. C. Borst. Optimal probabilistic allocation of customer types to servers. In *Proceedings of ACM SIGMETRICS*, pages 116–125, September 1995.

[10] A. Glazer and R. Hassin. Stable priority purchasing in queues. *Operations Research Letters*, 4:285–288, April 1986.

[11] R. Hassin. Decentralized regulation of a queue. *Management Science*, 41(1):163–173, January 1995.

[12] L. Kleinrock. Optimum bribing for queue position. *Operations Research*, 15:304–318, March/April 1967.

[13] F. T. Lui. An equilibrium queuing model of bribery. *Jour. of Political Economy*, 93:760–781, Aug 1985.

[14] S. Rao and E. Petersen. Optimal pricing of priority services. *Operations Research*, 46(1):46–56, 1998.

[15] J. G. Wardrop. Some theoretical aspects of road traffic research communication networks. *Proceedings of Industrial and Civil Engineering*, 1:325–378, 1952.