Priority Auctions and Queue Disciplines that Depend on Processing Time

Thomas Kittsteiner and Benny Moldovanu*

July 16, 2004

Abstract

We analyze the allocation of priority in queues via simple bidding mechanisms. In our model, the stochastically arriving customers are privately informed about their own processing time. They make bids upon arrival at a queue whose length is unobservable. We consider two bidding schemes that differ in the definition of bids (these may reflect either total payments or payments per unit of time) and in the timing of payments (before, or after service). In both schemes, a customer obtains priority over all customers, waiting in the queue or arriving while he is waiting, who make lower bids. Our main results show how the convexity/concavity of the function expressing the costs of delay determines the queue-discipline (i.e., SPT, LPT) arising in a bidding equilibrium.

1 Introduction

We analyze simple auction schemes in an environment where randomly arriving, heterogenous customers need to share a processing device that can only be used

*The authors are grateful to an associate editor and two referees for excellent editorial comments on a previous version. We are grateful to Philippe Afeche, Alan Beggs, Maria Angeles de Frutos, Oliver Hart, Moshe Haviv, Paul Klemperer, Meg Meyer and Roger Myerson for helpful comments. Kittsteiner: Nuffield College, Oxford University, and University of Bonn; Moldovanu: Department of Economics, University of Bonn; mold@uni-bonn.de

sequentially. Since customers can only be served one at a time, queues may form and the customers incur a waiting cost until their job is completed. Examples include communication in congested networks, job processing by capacity-constrained computers (including services provided via the internet), and the usage of various industrial production units.

In a queueing system where all customers can be eventually served, the allocation problem reduces to a determination of the order in which customers' jobs are processed. There are two main reasons for awarding priority to some customers:

1) Giving priority to certain customers (e.g., those with shorter processing times) can lead to a decrease in overall (expected) queuing costs, and thus to overall higher welfare.

2) Since customers are willing to pay a price for being served earlier (which means avoidance of waiting costs), it is revenue-increasing for the service provider to charge customers for priority.

The efficiency aspect of priority pricing through auction-like schemes has been addressed in important papers by Glazer and Hassin[1986], Lui[1985] and Afeche and Mendelson [2001]. In their set-up, customers incur a waiting cost that is linear in waiting time, and they differ in their marginal costs which are private information (Afeche and Mendelson also consider a multiplicative structure that bundles delay costs and values). Each customer's processing time is a random variable, and it is realized when processing starts (thus it is unknown to customers a-priori). An efficient allocation (i.e., one that minimizes total delay costs) calls here for higher priority to be awarded to customers with higher marginal costs - this is the so-called "$C\mu$ rule". It has been indeed shown that customers with higher marginal costs bid more, and hence that the auction implements the efficient queue discipline. In this model the efficient discipline is unrelated to realized processing times.

In contrast, we analyze here a situation where own processing time is known to customers, and it is private information. This is a natural assumption in situations where customers know better than others the type of job they will

submit to the server.

We first study a simple bidding scheme where arriving customers submit a bid that reflects total payment for service, and then pay this bid before joining the queue. A customer is given priority over all customers in the queue who submitted lower bids upon arrival. This mechanism has several compelling properties: 1) Its rules do not depend on distributional assumptions about stochastic arrival and processing times. 2) The service provider need not enforce payments after service has been granted[1]: especially in situations with many unknown customers (e.g., services provided via the internet) ex-post enforcement may be very costly and sometimes impracticable. 3) Since auctions extract and aggregate information available to the customers, it seems a-priori likely that such a mechanism is able to implement a queue discipline with good welfare properties (e.g., better than, say, the first-come-first-serve method).

We next compare the above mechanism with another very natural scheme where arriving customers bid for a unit of time and, after being served, pay their bid multiplied by the actual monitored processing time. Thus, this scheme relies on monitoring service and on ex-post payments. It seems a-priori intuitive that the bidders' strategic manipulations can be better controlled in this second mechanism.

The main role in our analysis of both auction schemes is played by the curvature of the cost function. The point is that the increase in waiting cost that is incurred by a customer who waits one additional time unit depends on his own processing time: if the cost function is convex [concave] the magnitude of this dependence is increasing [decreasing] in own processing time. Since bids reflect the willingness to pay in order to avoid increases in waiting costs, this effect determines the form of the equilibrium bidding function: if this function is increasing [decreasing] in own processing time, the auction implements the longest-processing-time-first (LPT) [the shortest-processing-time-first

[1]The Economist [2003] reports for example that Britain's debt-collection agencies handled 20 million cases of unpaid debt in one year. The involved sum was over $8 billion, or about 7% of all unsecured debt and the recovery rate was less than 2-3%.

(SPT)] queue discipline.

In practice, delay cost functions are often non-linear. In situations with a concave cost function, initial increases in waiting time are extremely costly, but further increases are less costly. Good examples are emergency situations in capacity constrained facilities such as hospitals, fire-fighting, etc... On the other hand, convex cost function approximate the ubiquitous situations where there is a deadline (this can be real or stemming from a customer's expectation) by which a job must be processed in order for the customer to derive a value. Van Mieghem[1995] surveys several papers that describe real-life applications exhibiting non-linear costs.

We show that, depending on the curvature of the cost function, either the SPT or the LPT discipline can be implemented, but not both (the only exception is the case of linear cost functions in our first scheme). Moreover, the auction where customers have linear cost functions and make bid per units of processing time yields the SPT discipline which is then efficient (i.e., minimizes overall delay costs).

From a technical point of view, our model is relatively difficult to analyze since bids directly depend here on the customers' actual processing time, and since the queue discipline and the resulting waiting time distribution are endogenously determined by the bids (whereas, in the model where customers know only their marginal cost of waiting, bids are unrelated to processing time).

The paper is organized as follows: The next subsection reviews the relevant literature. In Section 2 we present the queueing model and the auction scheme where customers make ex-ante payments. In Section 3 we explicitly calculate bidding equilibria for the case of convex and concave cost functions, and discuss the resulting queue disciplines. We also show how Laplace-transforms can be used to derive closed-form solutions for the case of cost functions that are polynomial in waiting time. In Section 4 we relax the assumption of unanimous participation, and we discuss the welfare properties of participation decisions in equilibrium. In Section 5 we introduce a simple auction scheme where the ex-post payments depend on monitored service times, and compare its perfor-

mance to that of the auction with ex-ante payments. Section 6 gathers some concluding comments. Appendix A displays technical material about Laplace transforms, and Appendix B deals with the "non-generic" case where there are linear costs of delay.

1.1 Related Literature

Queues with a finite number of priorities (which are independent of processing times) have been first analyzed by Cobham [1954]. Phipps [1956] generalized Cobham's results to a model where jobs with a shorter processing time have a higher priority (thus there are a continuum of priorities). Phipps derived steady-state expected waiting times for the SPT discipline. Schrage and Miller [1966] allow also for preemption, and derive the Laplace Transform of the waiting time distribution in Phipps' model with the *shortest remaining time discipline* (SRPT)[2]. We extensively use this derivation for our case with non-linear delay costs. Schrage [1968] proved that the SRPT discipline minimizes expected waiting time (or, equivalently, minimizes expected cost due to delay if the cost function is linear). For the case of quadratic waiting costs, Schrage [1973] noted that queue disciplines based on *priority functions* cannot be optimal. A priority function assigns a priority to a job based solely on its own characteristics (i.e., it does not depend on characteristics of other jobs)

Kleinrock [1967] was the first to study the allocation of priorities based on payments made by customers - this small strand of the literature, together with many other strategic issues arising in queues, is well surveyed in the excellent book by Hassin and Haviv [2002]. In Kleinrock's model a new arrival offers a nonnegative payment to the queue manager (these payments were called "bribes"). This customer is then assigned a position in the queue such that all those customers who made larger payments are in front of him, and all those customers who made smaller payments are behind him. Kleinrock

[2]Their presentation is somewhat dense. For a more leisurely one see the book by Conway, Maxwell and Miller [1967].

derived steady-state expected waiting times (that depend on the bribes) and studied the resulting queue discipline for various payment functions. He also showed that payment functions that are monotone in customers' valuations of time minimize (linear) waiting costs subject to a budget constraint. Here we use Kleinrock's results for the case of a linear cost of delay. Note that Kleinrock's payment functions were exogenous, i.e., they were not determined by individual maximization or by some equilibrium condition.

The earliest work on priority assignment based on payments that satisfy an equilibrium condition is due to Balachandran [1972]. In his model, identical customers observe the length of the queue and choose from a discrete, infinite set of possible payments. Constraints on the set of payments are exhibited under which it is an equilibrium to purchase the lowest payment that ensures being placed at the head of the queue. Tilt and Balachandran [1979] generalize this idea and derive conditions on the set of payments such that either the FCFS or the LCFS discipline are implemented in an M/M/s/N queue where arriving customers can observe the number of queuing customers (but neither their own, nor the other customers' processing time). They show how the auctioneer is able to implement each of the two opposite queue disciplines by restricting the set of possible payments in an appropriate manner.

Incentive problems that arise with privately informed customers have been first studied in the context of queues by Ghanem [1975]. In his model, customers have linear delay costs, and are privately informed about the marginal cost. In this model total waiting cost is minimized by first serving customers with a higher marginal cost- this is the "$C\mu$ rule". Ghanem assumes that there is a fixed, finite and exogenously given set of priorities, and calculates incentive compatible prices ensuring that customers sort themselves according to the (constrained) $C\mu$ rule. Mendelson and Whang [1990] assume that customers in different classes have different distributions of processing times and analyze pricing schemes that charge customers on the basis of both the declared class, and on ex-post realized processing time (that is assumed to be verifiably monitored). It is shown that incentive compatible pricing that leads to optimal

participation decisions of the customers must include a quadratic component - this is a form of non-linear pricing. Wilson [1983] offers an extensive treatment of priority pricing in this and other (non-queuing) contexts.

Glazer and Hassin [1986] and Liu [1985] revisit Kleinrock's model, but assume that customers make payments in order to minimize total cost (which is the sum of the delay cost and the bid). In their model, customers have privately known, heterogeneous marginal costs of delay (delay costs are linear in waiting time). This yields a so called "private values" auction. As in our model, customers do not get to see the actual queue length, nor the bids made by others before placing their own bid. These authors derive a bidding equilibrium, and show that a higher marginal cost leads to a higher bid. Thus, the $C\mu$ rule is implemented by the auction. The above authors also investigate whether customers gain by the introduction of the payment scheme in relation to a FCFS service discipline. This line of study and other extensions (e.g., the introduction of a minimum bid) were also pursued by Afeche and Mendelson [2001].

Note that, in contrast to the present work, the induced queue discipline in all above papers is not a function of service times since customers' bids cannot be made contingent on the yet unknown processing times.

If customers can submit a bid (or bribe), and priority is given to customers with higher bids, Hassin [1995] showed that customers' decisions to queue are socially optimal and the revenue of the service provider is maximized if the queue is unobservable. This complements the result in Edelson and Hildebrand [1975] who showed that efficient participation can be achieved by charging the revenue maximizing (fixed) price. If the queue is observable, Naor [1969] shows that the revenue maximizing toll leads to inefficient participation.

The treatment of the scheduling model (i.e., where all jobs arrive simultaneously) as a mechanism design problem with interdependent values has been introduced by Hain and Mitra [2002]. Their main result is that, for cost functions that are concave polynomials of degree less than or equal to $n-2$ (where n is the number of customers and slots), a generalized Clarke-Groves-Vickrey mechanism can be constructed that is efficient and ex-post budget balanced.

The construction of the CGV mechanism is based on general insights about efficient implementation for multi-object auctions with interdependent valuations due to Dasgupta and Maskin [2000] and Jehiel and Moldovanu [2001]. The use of CGV mechanisms for solving incentive problems in queues has been first proposed by Dolan [1978].

Kittsteiner and Moldovanu [2003] use the scheduling framework of Hain and Mitra in order to study the equilibrium of simple bidding mechanisms (e.g., auctions, auctions with bid caps, and a fixed-fee+lottery scheme). Note that in the scheduling framework the SPT discipline is always efficient, irrespective of the form of the cost function. A main result in that paper is that a lottery performs better than an auction (both from the efficiency and from the revenue point of views) if the cost functions are convex, while an auction yields the efficient queue discipline if the cost functions are strictly concave. In contrast, the present paper focuses on the complex bidding mechanics with stochastically arriving customers, and on the impact of monitoring and ex-post payments.

2 The Queueing and Bidding Models

There are infinitely many potential customers who arrive at a server according to a Poisson process with arrival rate λ. Thus, the distribution of the number of arriving customers is independent of the number of customers that have already arrived.

Each customer i has a job that needs to be processed by the server. We assume that the *processing time* t_i of customer i is drawn from a distribution F with support $[\underline{t},\overline{t}]$, $\underline{t} \geq 0$, independently of other processing times. We also assume that F has a continuous density $f > 0$. The realization t_i is only known to i, whereas the distribution F is common knowledge.

If customers are served according to a queue discipline that does not depend on the customers' types (i.e., on their private information) we obtain a standard $M/G/1$ queue. This is the case, for example, if the queue discipline is first-come first-served (FCFS).

The *waiting time* of a customer is the difference between the time when his job is finished and his arrival time to the queue. The *queuing time* is the difference between his waiting time and his processing time, i.e., it is the time a customer spends in the queue.

Each customer i derives a value of V_i if his task is processed. The valuation V_i need not be known to customers other than i.

Customers face a cost of waiting that is an increasing function of their total waiting time: a customer with processing time t_i who has to queue for T_i time units incurs a cost $C(t_i + T_i)$, where $C : \mathbb{R}^+ \mapsto \mathbb{R}^+$ is strictly increasing and differentiable.

Customer i's utility is given by $U_i = V_i - C(t_i + T_i) - m_i$, where m_i denotes a monetary payment.

Note that the queuing time T_i depends on the processing times of the customers served before i, and, therefore, that it depends on other customer's private information. We assume that customers cannot observe the queue's length and its composition prior to their arrival. We also assume that all customers join the queue, regardless of V_i. This can be rationalized if V_i is always high enough to cover the expected waiting costs (we relax this last assumption in section 4). We further assume that customers are either not allowed to leave the queue (this implies that there is no bound on the cost they are facing ex-post), or not willing to leave the queue (this implies that the marginal costs of waiting one additional time unit are bounded).

Upon arrival a customer submits a non-negative bid, and is placed in front of all customers who are presently in the queue and who submitted lower bids. Each customer has to pay his bid. In Section 5 we modify the bidding model by introducing bids per unit of processing time, and letting customers pay (after service) this bid times their actual processing time.

If two customers submit the same bid, any tie-breaking device can be used, e.g., they can be served on a FCFS basis. Any tie breaking rule can be applied also if customers arrive at the same time (this event has zero probability) and submit the same bid. The task of the customer who is currently processed is

never interrupted (hence we consider a *non-preemptive* queuing-discipline).

In contrast to standard queuing models, the distribution of an customer's waiting time is endogenously determined by his bidding behavior, and, conversely, this distribution must be taken into account when computing equilibrium bidding behavior.

We analyze the bidding behavior and the resulting allocation in the steady state of the queue. In order to ensure the existence of a steady state, we assume that $\int_{\underline{t}}^{\overline{t}} tf(t)\,dt < \frac{1}{\lambda}$, i.e., that the average processing time is smaller than the average inter-arrival time.

3 Bidding Equilibria and Queue Disciplines with Ex-Ante Bids

The shortest-processing-time (SPT) discipline [longest-processing-time discipline (LPT)] puts an arriving customer with processing time t ahead of all [behind of all] waiting customers with processing times longer than t. It is well known (see, e.g., Conway et al. [1967]) that, for the case of linear costs of delay, the SPT discipline, minimizes expected waiting time, and hence overall expected costs among all possible work conserving queue disciplines. Similarly, the LPT disciplines maximizes expected total waiting time, and hence overall expected costs.

In the linear case, e.g., $C(t) = ct$, $c \in \mathbb{R}^+$, the willingness to pay for priority depends only on expected queuing time, but not on own processing time. This "non-genericity" allows us to implement in a bidding equilibrium both the SPT and the LPT disciplines (see Appendix B for the derivation and discussion of this result).

We assume here that C is either strictly convex or strictly concave. The main effects of these assumptions are:

1. The increase in cost that is incurred by customer i who waits one additional time unit depends on his own processing time t_i : if C is convex

[concave] the magnitude of this effect is increasing [decreasing] in t_i.

2. The increase in cost that is incurred by a customer who waits one additional time unit depends on the time he has already queued.

To understand these effects, consider a customer who has already queued for $\widetilde{T}_i$ time units, who has to queue for another T_i time units, and who has a processing time of t_i. The increase in cost due to a marginally higher queueing time $T_i + dt$ is given by $C'\left(\widetilde{T}_i + T_i + t_i\right)$, which is increasing [decreasing] in t_i if C is convex [concave]. A consequence of 1. is that we cannot have both LPT and SPT disciplines implemented in equilibrium for a strictly convex [concave] cost function. A consequence of 2. is that the SPT discipline is not necessarily cost-minimizing since the time already spent waiting has to be taken into account by the queue discipline. For example, if C is convex, it may be possible to reduce overall cost by giving priority to a customer with a long processing time if he already waited a long time. Or, if C is concave, it may be possible to reduce cost by awarding priority to a new arrival with a longer processing time. The derivations of the optimal discipline (and the associated distribution of waiting time) in the general case is very complex and not yet known (Van Mieghem [1995] shows that a generalized $c\mu$ rule that also incorporates actual waiting time, is asymptotically optimal under heavy traffic conditions).

Clearly, simple bidding mechanisms with one-shot bids placed upon arrival cannot implement schemes where willingness to pay varies endogenously. We focus here on a simple auction, but even for this mechanism the derivation of bidding equilibria turns out to be complex since the bid of a customer with type t depends on the entire distribution of queuing time for a type t rather than solely on the expected queuing time (as is the case with linear cost functions - see Appendix B).

Let the distribution of queuing time in a non-preemptive SPT [LPT] discipline for a customer with processing time t be given by the density function $w_S(t,\cdot)$ $[w_L(t,\cdot)]$. To start with, assume that the SPT discipline is implementable in a symmetric equilibrium, i.e., that all customers bid according to

the same, strictly decreasing, equilibrium bidding function $b_S(t)$. The expected waiting cost of customer i with type t who pretends being of type $\widehat{t}$ is given by

$$C_K\left(t,\widehat{t}\right) := \int_0^\infty C\left(t+\widetilde{t}\right) w_K\left(\widehat{t},\widetilde{t}\right) d\widetilde{t}\ ,\ K=S,L$$

Hence we have

$$U_i\left(t,\widehat{t}\right) = V - C_K\left(t,\widehat{t}\right) - b_K\left(\widehat{t}\right),\ K=S,L.$$

A candidate for a (differentiable) equilibrium bidding function has to fulfill the necessary condition: $\frac{\partial}{\partial \widehat{t}} U_i\left(t,\widehat{t}\right)\Big|_{\widehat{t}=t} = 0$. Moreover this condition is sufficient[3] if, in addition, we have

$$\frac{\partial^2}{\partial t\ \partial \widehat{t}} U_i\left(t,\widehat{t}\right) = -\frac{\partial^2 C_K\left(t,\widehat{t}\right)}{\partial t\ \partial \widehat{t}} \geq 0 \text{ for all } t,\widehat{t}$$

For the SPT discipline, the above condition says that the effect of an increase in processing time t on expected cost has to be lower if a customer is placed more to the end of the queue (i.e., pretends to have a higher type). Intuitively, this holds if the cost function is concave. If the cost function is convex, we expect the opposite to hold: $\frac{\partial^2}{\partial t\ \partial \widehat{t}} \int_0^\infty C\left(t+\widetilde{t}\right) w_S\left(\widehat{t},\widetilde{t}\right) d\widetilde{t} \geq 0$ for all t, $\widehat{t}$. This implies the non-existence of a bidding equilibrium that implements the SPT-discipline in the convex case. In that case, we obtain that $\frac{\partial^2 C_L\left(t,\widehat{t}\right)}{\partial t\ \partial \widehat{t}} \leq 0$ for all t, $\widehat{t}$, and that the LPT-discipline is implemented in equilibrium (as in the static scheduling problem of Kittsteiner and Moldovanu [2003]).

We can now summarize our findings. We assume that $C_K\left(t,\widehat{t}\right) < \infty$, that it is twice continuously differentiable, and we define

$$C_{K,2}\left(t,\widehat{t}\right) := \frac{\partial C_K\left(t,\widehat{t}\right)}{\partial \widehat{t}}$$

Theorem 1 *1. If the cost function C is concave, the following equilibrium bidding function that is decreasing in processing time implements the SPT*

[3] As shown in McAfee [1991] , it suffices to show that $\frac{\partial}{\partial t \partial t_i} U_i\left(t_i,\widehat{t}\right) \geq 0$ for all $\widehat{t}$, t_i.

discipline in steady state:

$$b_S(t) = \int_t^{\overline{t}} C_{S,2}(x,x)\,dx. \tag{1}$$

Moreover, there can be no monotonically increasing equilibrium bidding function, which implies that the LPT discipline cannot be implemented in this case.

2. *If the cost function C is convex, the following equilibrium bidding function that is increasing in processing time implements the LPT discipline in steady state:*

$$b_L(t) = \int_t^{\underline{t}} C_{L,2}(x,x)\,dx. \tag{2}$$

Moreover, there can be no monotonically decreasing equilibrium bidding function, which implies that the SPT discipline cannot be implemented in this case.

Proof. We give the argument for concave C. The convex case is analogous. Sufficient conditions for a strictly decreasing function $b_S(t)$ to be an equilibrium are given by:

1. $b_S\left(\overline{t}\right) = 0$, since a customer with a processing time of $\overline{t}$ never gets priority and therefore bids zero.
2. $\frac{\partial}{\partial \widehat{t}} U_i\left(t,\widehat{t}\right)\Big|_{\widehat{t}=t} = 0$ and
3. $\frac{\partial}{\partial t\,\partial \widehat{t}} U_i\left(t,\widehat{t}\right) = -\frac{\partial^2 C_S\left(t,\widehat{t}\right)}{\partial t\,\partial \widehat{t}} \geq 0$ for all t, $\widehat{t}$.

The conditions

$$\frac{\partial}{\partial \widehat{t}} U_i\left(t,\widehat{t}\right)\Big|_{\widehat{t}=t} = -C_{S,2}(t,t) - \frac{d}{dt} b_S(t) = 0$$

and $b_S\left(\overline{t}\right) = 0$ are clearly satisfied by the definition of the bidding function in (1). Consider two customers i, j with processing times $t_i < t_j$. Since in all

possible states of the system i always gets priority over j , $i's$ waiting time distribution first-order stochastically dominates $j's$ waiting time distribution. Since C is strictly concave, we obtain that :

$$\frac{\partial}{\partial t}\int_0^\infty C\left(t+\widetilde{t}\right)w_S\left(t_i,\widetilde{t}\right)d\widetilde{t} > \frac{\partial}{\partial t}\int_0^\infty C\left(t+\widetilde{t}\right)w_S\left(t_j,\widetilde{t}\right)d\widetilde{t}.$$

This implies that $\frac{\partial^2 C_S\left(t,\widehat{t}\right)}{\partial t\ \partial\widehat{t}} < 0$ as desired.

Assume now that there exists a strictly increasing equilibrium bidding function $\widetilde{b}$ that implements the LPT discipline. The necessary first order condition $\widetilde{U}_{i,2}\left(t,\widehat{t}\right) := \frac{\partial}{\partial\widehat{t}}\widetilde{U}_i\left(t,\widehat{t}\right) = -C_{L,2}\left(t,\widehat{t}\right) - \frac{d}{dt}\widetilde{b}\left(t\right) = 0$ must be satisfied for $\widehat{t} = t$ (a.e.). Note that $\widetilde{b}\left(t\right)$ has to be continuous, since otherwise a type just above a gap could improve his payoff by lowering the bid. Hence $\widetilde{b}$ has to be differentiable a.e., and because of the first order condition, in fact everywhere. Analogously to the above calculations, we get that $\frac{\partial^2}{\partial t\ \partial\widehat{t}}\widetilde{U}_i\left(t,\widehat{t}\right) < 0$. Hence for $\widehat{t} > t$ we have

$$\begin{aligned}\int_t^{\widehat{t}}\int_x^t \frac{\partial^2}{\partial z\ \partial x}\widetilde{U}_i\left(z,x\right)dzdx \quad &> \quad 0 \Rightarrow \int_t^{\widehat{t}}\left[\widetilde{U}_{i,2}\left(t,x\right) - \underbrace{\widetilde{U}_{i,2}\left(x,x\right)}_{=0}\right]dx > 0\\ &\Rightarrow \quad \widetilde{U}_i\left(t,\widehat{t}\right) - \widetilde{U_i}\left(t,t\right) = \int_t^{\widehat{t}} U_{i,2}\left(t,x\right)dx > 0\end{aligned}$$

This shows that $\widetilde{b}$ cannot be an equilibrium bidding function. ■

In general, it is impossible to derive a closed-form solution for the distribution of waiting time in either the SPT or the LPT discipline, but it is possible to derive Laplace-transforms. The moments of the distribution with densities w_S and w_L can be derived from their Laplace transforms, and, as a consequence, equilibrium bidding functions in models with polynomial cost functions can be derived in closed form. This is demonstrated below for the quadratic (i.e., convex) cost function. The Laplace transforms of $w_K\left(t,\cdot\right)$, $t \in [\underline{t},\overline{t}]$, $K = S, L$, are given by $w_K^*\left(t,s\right) := \int_0^\infty e^{-s\widetilde{t}}w_K\left(t,\widetilde{t}\right)d\widetilde{t}$ where $s \geq 0$.

Lemma 2 *1. The Laplace transform $w_S^*\left(t,s\right)$ for a type-t customer in the*

SPT discipline is given by

$$\begin{aligned} w_S^*(t,s) &= \frac{1}{s}\left[\left(1-\lambda\int_{\underline{t}}^{\overline{t}} x dF(x)\right)(s+\lambda F(t)(1-g_S(t,s))) \right. \\ &\quad \left. +\lambda(1-F(t))\left(1-\frac{1}{1-F(t)}\int_t^{\overline{t}} e^{-(s+\lambda F(t)(1-g_S(t,s)))x} dF(x)\right)\right] \end{aligned}$$

where $g_S(t,\cdot)$ *is implicitly given by* $g_S(t,s)=\frac{1}{F(t)}\int_{\underline{t}}^{t} e^{-(s+\lambda F(t)(1-g_S(t,s)))x} dF(x)$.

2. *The Laplace transform* $w_L^*(t,s)$ *of type-t customer in the LPT discipline is given by*

$$\begin{aligned} w_L^*(t,s) &= \frac{1}{s}\left[\left(1-\lambda\int_{\underline{t}}^{\overline{t}} x dF(x)\right)(s+\lambda(1-F(t))(1-g_L(t,s))) \right. \\ &\quad \left. +\lambda F(t)\left(1-\frac{1}{F(t)}\int_{\underline{t}}^{t} e^{-(s+\lambda(1-F(t))(1-g_L(t,s)))x} dF(x)\right)\right] \end{aligned}$$

where $g_L(t,\cdot)$ *is implicitly given by*

$$g_L(t,s)=\frac{1}{1-F(t)}\int_t^{\overline{t}} e^{-(s+\lambda(1-F(t))(1-g_L(t,s)))x} dF(x).$$

Proof. See Appendix A. ■

Example 3 *Assume that* $C(t)=\frac{1}{2}t^2$. *We denote by* $W_L(t)$ *the actual waiting time of type* t *customer in the LPT queue, and by* $E[W_L(t)]$ *and* $E\left[W_L^2(t)\right]$ *its first and second moments, respectively. The expected waiting cost of a type* t *who pretends being of type* $\widehat{t}$ *is given by*

$$\frac{1}{2}\int_0^{\infty}\left(t+\widetilde{t}\right)^2 w_L\left(\widehat{t},\widetilde{t}\right)d\widetilde{t}=\frac{1}{2}t^2+tE\left[W_L\left(\widehat{t}\right)\right]+\frac{1}{2}E\left[W_L^2\left(\widehat{t}\right)\right].$$

Hence, we have

$$C_{L,2}(x,x)=x\left.\frac{d}{dt}E[W_L(t)]\right|_{t=x}+\frac{1}{2}\left.\frac{d}{dt}E\left[W_L^2(t)\right]\right|_{t=x}.$$

and we obtain

$$b_L(t)=-\int_{\underline{t}}^{t} x\left.\frac{d}{dt}E[W_L(t)]\right|_{t=x}dx-\frac{1}{2}\left(E\left[W_L^2(t)\right]-E\left[W_L^2(\underline{t})\right]\right). \qquad (3)$$

The moments of $W_L(t)$ are obtained by differentiation of $w_L^*(t,s)$ as follows:

$$E[W_L(t)] = -\frac{d}{ds}w_L^*(t,s)\Big|_{s=0}, \qquad E\left[W_L^2(t)\right] = \frac{d^2}{ds^2}w_L^*(t,s)\Big|_{s=0}.$$

The calculations are tedious (but straightforward), and we only sketch them here. By differentiation it can be shown that:

$$\begin{aligned}
\frac{d}{ds}g_L(t,0) &= -\frac{\frac{1}{1-F(t)}\int_t^{\overline{t}} xdF(x)}{1-\lambda\int_t^{\overline{t}} xdF(x)};\\
\frac{d^2}{ds^2}g_L(t,0) &= \frac{\frac{1}{1-F(t)}\int_t^{\overline{t}} x^2dF(x)}{\left(1-\lambda\int_t^{\overline{t}} xdF(x)\right)^3};\\
\frac{d^3}{ds^3}g_L(t,0) &= -\frac{1}{1-F(t)}\left[\frac{\int_t^{\overline{t}} x^3dF(x)}{\left(1-\lambda\int_t^{\overline{t}} xdF(x)\right)^4} + \frac{3\lambda\left(\int_t^{\overline{t}} x^2dF(x)\right)^2}{\left(1-\lambda\int_t^{\overline{t}} xdF(x)\right)^5}\right].
\end{aligned}$$

Using the above expressions, we calculate $\frac{d}{ds}w_L^*(t,s)\big|_{s=0}$ and $\frac{d^2}{ds^2}w_L^*(t,s)\Big|_{s=0}$ by applying l'Hopital's rule twice and three times, respectively. After combining terms we obtain

$$\frac{d}{ds}w_L^*(t,s)\Big|_{s=0} = -\frac{Q_0}{\left(1-\lambda\int_t^{\overline{t}} xdF(x)\right)^2}$$

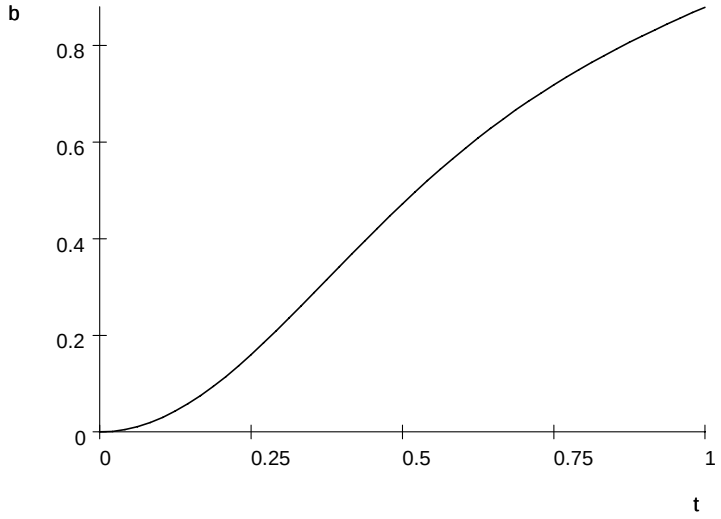
where $Q_0 := \frac{1}{2}\lambda\int_{\underline{t}}^{\overline{t}} x^2dF(x)$ denotes the average remaining processing time of the customer who is currently in service (see Appendix B for the derivation). Note that this calculation confirms a well known result for linear costs (see equation 7 in Appendix B). We also obtain:

$$\frac{d^2}{ds^2}w_L^*(t,s)\Big|_{s=0} = \frac{1}{3}\frac{\lambda\int_{\underline{t}}^{\overline{t}} x^3dF(x)}{\left(1-\lambda\int_t^{\overline{t}} xdF(x)\right)^3} + \frac{1}{2}\frac{\lambda^2\int_{\underline{t}}^{\overline{t}} x^2dF(x)\ \int_t^{\overline{t}} x^2dF(x)}{\left(1-\lambda\int_t^{\overline{t}} xdF(x)\right)^4}.$$

The bidding function becomes now:

$$\begin{aligned} b_L(t) &= 2Q_0\lambda \int_{\underline{t}}^{t} \frac{x^2}{1-\lambda\int_x^{\overline{t}} y dF(y)} dF(x) + \frac{1}{2}\left(E\left[W_L^2(\underline{t})\right] - E\left[W_L^2(t)\right]\right) \\ &= 2Q_0\lambda \int_{\underline{t}}^{t} \frac{x^2}{1-\lambda\int_x^{\overline{t}} y dF(y)} dF(x) \\ &\quad + \frac{1}{6}\frac{\lambda\int_{\underline{t}}^{\overline{t}} x^3 dF(x)}{\left(1-\lambda\int_{\underline{t}}^{\overline{t}} x dF(x)\right)^3} + \frac{1}{4}\frac{\lambda^2\left(\int_{\underline{t}}^{\overline{t}} x^2 dF(x)\right)^2}{\left(1-\lambda\int_{\underline{t}}^{\overline{t}} x dF(x)\right)^4} \\ &\quad - \frac{1}{6}\frac{\lambda\int_{\underline{t}}^{\overline{t}} x^3 dF(x)}{\left(1-\lambda\int_{t}^{\overline{t}} x dF(x)\right)^3} - \frac{1}{4}\frac{\lambda^2\int_{\underline{t}}^{\overline{t}} x^2 dF(x)\int_{t}^{\overline{t}} x^2 dF(x)}{\left(1-\lambda\int_{t}^{\overline{t}} x dF(x)\right)^4}. \end{aligned}$$

The figure displays b_L for a uniform distribution of processing times on the interval $[0,1]$ *, and for* $\lambda = 1$.



By continuity, we can infer from the linear case (see Appendix B) that the SPT discipline is approximately efficient [whereas the LPT discipline is approximately anti-efficient] if the curvature of the cost function is small. Also because of continuity, we obtain from the linear case that, as long as the curvature of the cost function is small, revenue is higher if the cost function is concave and the SPT discipline is implemented compared to the case where it is convex and

the LPT discipline is implemented.

In the present framework, the service provider cannot implement the SPT discipline if the cost function is convex (for a different situation and result see Section 5). Note that an effective artificial restriction of bids leads here to pooling: customers with different processing times submit the same bid, and thus they cannot be distinguished by their processing time as necessary for SPT (for example, Kittsteiner and Moldovanu [2003] analyze the effect of a bid-cap in a static scheduling problem with interdependent costs, and show that a bid cap can increase efficiency if costs are convex). In the model of Tilt and Balachandran [1979] LCFS is more naturally implemented since customers who observe a larger queue have a higher willingness to pay for priority (since then it is more likely that they face competitors who also observed a large queue and bid high). On the other hand, if the set of bids is restricted so that all customers submit the same bid, they are served according to FCFS.

4 Endogenous Participation

In this section we relax the assumption that all types prefer to queue. We assume that customers derive a value $V_i = V > 0$ from processing, and that some prefer not to queue in equilibrium since the value V is too low to compensate for expected waiting costs.

We conduct the analysis to the case of concave or linear cost functions $C(t)$, and comment at the end of the section on the convex case.

We will show that, in a symmetric equilibrium that implements the SPT discipline, there exists a cut-off type t^* such that only customers with a processing time $t \leq t^*$ decide to queue. This relies on the fact that the expected costs (consisting of waiting costs and payments) are increasing in t.

If only customers with processing time $\underline{t} \leq t \leq t^*$ queue, the arrival rate of queuing customers is $F(t^*)\lambda$, and their processing times are distributed according to $F(t)/F(t^*)$, $\underline{t} \leq t \leq t^*$, on the interval $[0, t^*]$. We denote by $w_S^{t^*}$ the density of the distribution of the queuing time in a (non-preemptive) SPT discipline

where only customers with processing time $\underline{t} \leq t \leq t^*$ queue. The expected queuing time depends now on t^* since we have a non-preemptive queue-discipline, where the (expected) remaining processing time of the currently processed job (upon arrival) depends on t^*. Define

$$C_S^{t^*}\left(t,\widehat{t}\right) := \int_0^\infty C\left(t+\widetilde{t}\right) w_S^{t^*}\left(\widehat{t},\widetilde{t}\right) d\widetilde{t}$$

We assume that $C_S^{t^*}\left(t,\widehat{t}\right) < \infty$, that it is twice continuously differentiable, and denote:

$$C_{S,1}^{t^*}\left(t,\widehat{t}\right) := \frac{\partial C_S^{t^*}\left(t,\widehat{t}\right)}{\partial t}, \quad C_{S,2}^{t^*}\left(t,\widehat{t}\right) := \frac{\partial C_S^{t^*}\left(t,\widehat{t}\right)}{\partial \widehat{t}}$$

Theorem 4 *There exists a type t^* such that the following strategy is a symmetric equilibrium: customers with processing time $t > t^*$ do not queue, whereas customers with processing time $t \leq t^*$ queue and submit bids according to*

$$b_S^*(t) = \int_t^{t^*} C_{S,2}^{t^*}(x,x)\, dx.$$

The marginal type t^* is uniquely defined by

$$V = C_S^{t^*}\left(t^*,t^*\right).$$

Proof. The derivation of the bidding function for $t \leq t^*$ is the same as in Theorem 1. The interim utility of a customer i with type $t \leq t^*$ who is acting as if he were of type $\widehat{t} \leq t^*$ in the candidate equilibrium is given by

$$U_i^*\left(t,\widehat{t}\right) = V - C_S^{t^*}\left(t,\widehat{t}\right) - \int_{\widehat{t}}^{t^*} C_{S,2}^{t^*}(x,x)\, dx.$$

Hence, we have $\frac{d}{dt}U_i^*(t,t) = -C_{S,1}^{t^*}(t,t) < 0$ and $U_i^*\left(t^*,t^*\right) = 0$, which shows that customers with $t < t^*$ are better-off by queuing. A customer with type $t > t^*$ who is acting as if he were of type $\widehat{t} \leq t^*$ receives a strictly lower utility than a customer with type t^* who is acting as if he was of type $\widehat{t}$. Since a customer with type t^* cannot receive a strictly positive utility, a customer with type $t > t^*$ prefers not to queue. Since $C_S^{t^*}\left(t^*,t^*\right)$ is strictly increasing in t^*, we obtain that t^* is uniquely defined. ■

The next result shows that the customers' participation decisions are not necessarily efficient: the equilibrium cut-off level t^* is different from the socially optimal cut-off level.

Corollary 5 *Let t^{opt} be the cut-off level leading to efficient participation in the SPT queue, i.e.,*

$$t^{opt} := \arg\max_{t} \int_{\underline{t}}^{t} \left(V - C_S^t(x,x)\right) f(x)\, dx,$$

then $t^{opt} \leq t^*$.

Proof. The statement follows from the fact that we have $V = C_S^{t^*}(t^*, t^*)$. Hence, for $t \geq t^*$ we obtain:

$$\begin{aligned} & \frac{d}{dt} \int_{\underline{t}}^{t} \left(V - C_S^t(x,x)\right) f(x)\, dx \\ = \quad & \left(V - C_S^t(t,t)\right) f(t) - \int_{\underline{t}}^{t} \frac{dC_S^t(x,x)}{dt} f(x)\, dx < 0. \end{aligned}$$

Thus, t^{opt}, the maximizer of $\int_{\underline{t}}^{t} \left(V - C_S^t(x,x)\right) f(x)\, dx$, must be to the left of t^*. ■

The intuition for the above result is that a participating customer with type t^* imposes a negative externality on all customers who have a lower processing time, and arrive while this customer's job is being processed. There is also a positive externality on certain customers arriving after his job is processed: for example, assume that two customers arrive consecutively while a third one is processed. Assume that the first arriving customer has a long processing time, while his follower has a short processing time. The externality on the customer who arrives last might be overall positive, since he will be served before his predecessor, which is not the case if the machine was idle upon arrival of the first customer.

By deciding on participation, customers do not internalize the externalities they impose on future arrivals. The inefficiency is caused by the non-preemptive priority rule. The argument is similar to the one in Naor [1969] where customers

are served on a FCFS basis. This point has also been made in a related model by Hassin and Haviv [2002].

The analysis for convex costs is similar. The main difference is that the marginal participating type need not be uniquely defined, and that he submits a non-zero bid since he is given priority over all customers that participate but have shorter processing time. This implies that the participation decision is even more inefficient since the marginal type exerts a negative externality on all participating customers (and not only on those arriving while he is served).

5 Monitoring and Ex-Post Payments

In many realistic situations it is technically feasible to monitor the processing of the jobs, and therefore to ex-post determine processing times. Thus, it seems that the service provider can gain complete control on bidders' incentives by conditioning payments on the ex-post observed processing times (e.g., by imposing large fines or interrupting service in case bidders have "lied"). Besides the possible inefficiencies induced by the cost of monitoring and administering interruptions and fines, there sometimes is a contractual problem with such schemes: at least a part of the transfer (that controls for incentives) needs to be paid after service is completed. Hence, the service provider faces the risk of customers defaulting on their payments. In large systems with many anonymous customers (e.g., job processing via the internet), this risk is real and potentially large. This is why we first analyzed mechanisms that do not rely on monitoring and ex-post payments.

We now analyze a simple bidding mechanism based on ex-post payments, and compare it to the scheme analyzed above. The main question is of course whether our previous insights continue to play a role.

The model is as follows: upon arrival at the queue, customer i bids a price for one unit of processing time, and after service he pays this price times (the perfectly monitored) processing time t_i. Priority is given over all other customers (waiting in the queue or arriving while the customer queues) that submitted

lower bids.

Assume that all but customer i bid according to a (strictly) decreasing function $\widetilde{b}_S$. Then, the expected utility of customer i with type t who bids $\widetilde{b}_S\left(\widehat{t}\right)$ is given by

$$U_i\left(t,\widehat{t}\right) = V - \int_0^\infty C\left(t+\widetilde{t}\right) w_S\left(\widehat{t},\widetilde{t}\right) d\widetilde{t} - \widetilde{b}_S\left(\widehat{t}\right) t.$$

There are now two main effects that influence willingness to pay:

1. As in the previous sections, willingness to pay, depends on processing time: If the cost function is concave [convex], the cost of waiting one more time unit is decreasing [increasing] in the processing time.

2. In addition, given a fixed ex-ante bid per unit of time, the ex-post payment necessarily increases in processing time, and therefore it is less costly for a customer with a lower processing time to increase his bid.

If the cost function is concave, these two effects work in the same direction, and bids are decreasing in the processing time. If the cost function is convex, the two described effects have different signs, and the form of bidding depends then on their relative magnitude. In this context, it is important to recall that the cost of waiting one more time unit is independent of own processing time for linear cost functions (see also Appendix B). Hence in this case the first effect is nil, and bids will necessarily be decreasing in processing time. Thus, we cannot anymore implement the LPT discipline if the cost function is linear and the symmetric equilibrium implements the (efficient) SPT discipline. By continuity, the equilibrium continues to implement the SPT discipline if the cost function is not "too convex" (i..e, if the first effect is small).

Theorem 6 *1. If the cost function C fulfills the following condition*

$$\widehat{t}\frac{\partial}{\partial t} C_{S,2}\left(t,\widehat{t}\right) \leq C_{S,2}\left(\widehat{t},\widehat{t}\right) \text{ for all } t,\ \widehat{t} \tag{4}$$

then the following equilibrium bidding function that is decreasing in processing time implements the SPT discipline:

$$\widetilde{b}_S(t) = \int_t^{\overline{t}} \frac{1}{x} C_{S,2}(x,x)\, dx. \tag{5}$$

Condition (4) is satisfied for concave or linear cost functions. Moreover, if this condition is satisfied, there can be no monotonically increasing equilibrium bidding function, implying that the LPT discipline cannot be implemented in this case.

2. *If the cost function C fulfills the following condition*

$$\widehat{t}\frac{\partial}{\partial t} C_{L,2}\left(t,\widehat{t}\right) \leq C_{L,2}\left(\widehat{t},\widehat{t}\right) \text{ for all } t,\ \widehat{t}$$

then the following equilibrium bidding function that is increasing in processing time implements the LPT discipline:

$$\widetilde{b}_L(t) = \int_t^{\underline{t}} \frac{1}{x} C_{L,2}(x,x)\, dx.$$

Moreover, there can be no monotonically decreasing equilibrium bidding function, implying that the SPT discipline cannot be implemented in this case.

Proof. The proof is similar to the proof of Theorem 1. We only sketch here part 1. The necessary condition

$$\frac{\partial}{\partial \widehat{t}} U_i\left(t,\widehat{t}\right)\Big|_{\widehat{t}=t} = -C_{S,2}(t,t) - t\frac{d}{dt}\widetilde{b}_S(t) = 0$$

is clearly fulfilled by the definition of the bidding function in (5). In addition we have

$$\frac{\partial^2}{\partial t\, \partial \widehat{t}} U_i\left(t,\widehat{t}\right) = -\frac{\partial}{\partial t} C_{S,2}\left(t,\widehat{t}\right) + \frac{1}{\widehat{t}} C_{S,2}\left(\widehat{t},\widehat{t}\right)$$

and therefore $\frac{\partial^2}{\partial t\, \partial \widehat{t}} U_i\left(t,\widehat{t}\right) \geq 0$ for all $t, \widehat{t}$ follows from (4). If C is concave [linear] we have that $\frac{\partial}{\partial t} C_{S,2}\left(t,\widehat{t}\right) < [=]\ 0$ and therefore (4) is satisfied. Assume

now that there exists a strictly increasing equilibrium bidding function $\widetilde{b}$ that implements the LPT discipline. Since for such a bidding function we have that $\frac{\partial^2}{\partial t\,\partial \widehat{t}}\widetilde{U}_i\left(t,\widehat{t}\right) = -\frac{\partial}{\partial t}C_{L,2}\left(t,\widehat{t}\right) - t\frac{d}{dt}\widetilde{b}\left(t\right) < 0$ we can show (see the proof of Theorem 1) that $\widetilde{b}\left(t\right)$ cannot be optimal for a type t customer. ■

If cost functions are convex but close to linear, then, by continuity, the SPT discipline will be close to being welfare-maximizing. In conjunction with Theorem 1, the above result shows, for this case, that a welfare maximizing service provider prefers to let payments depend on actual service time[4]. Interestingly, revenue is lower in the auction scheme with ex-post payments (as long as the SPT discipline is implemented):

$$\begin{aligned}\int_{\underline{t}}^{\overline{t}} t\widetilde{b}_S\left(t\right) f\left(t\right) dt &= \int_{\underline{t}}^{\overline{t}}\int_{t}^{\overline{t}} \frac{t}{x} C_{S,2}\left(x,x\right) dx f\left(t\right) dt \\ &< \int_{\underline{t}}^{\overline{t}}\int_{t}^{\overline{t}} C_{S,2}\left(x,x\right) dx f\left(t\right) dt = \int_{\underline{t}}^{\overline{t}} b_S\left(t\right) f\left(t\right) dt.\end{aligned}$$

Since an increase in bids is more costly for customers with high processing time, the auction format with ex-post payments handicaps such bidders. Hence, there is less competition for positions in the queue, resulting in lower revenue. It can readily be verified that this argument remains valid if the service provider asks customer i to pay $p\left(t_i\right) b$, where p is increasing in (monitored) processing time, and where b is the bid of i. Conversely, if the cost function is sufficiently concave, the service provider could ask for a payment $p\left(t_i\right) b$,where p is decreasing. This increases the competition from customers with high processing time, and also increases revenue.

Example 7 *Assume that* $C\left(t\right) = \frac{1}{2}t^2$. *We denote by* $E\left[W_S\left(t\right)\right]$ *and* $E\left[W_S^2\left(t\right)\right]$ *the first and second moments of the distribution of waiting time of a type* t *customer in the SPT queue.* $E\left[W_S\left(t\right)\right]$ *and* $E\left[W_S^2\left(t\right)\right]$ *are derived by the method described in Lemma 2 and Example 3. With* $Q_0 := \frac{1}{2}\lambda\int_{\underline{t}}^{\overline{t}} x^2 dF\left(x\right)$, *we obtain*

[4]Moreover, such a scheme eliminates the LPT equilibrium for the linear cost functions (see Appendix B)

that

$$\begin{aligned} E\left[W_S\left(t\right)\right] &= \frac{Q_0}{\left(1-\lambda\int_{\underline{t}}^{t} x dF\left(x\right)\right)^2}, \\ E\left[W_S^2\left(t\right)\right] &= \frac{1}{3}\frac{\lambda\int_{\underline{t}}^{\overline{t}} x^3 dF\left(x\right)}{\left(1-\lambda\int_{\underline{t}}^{t} x dF\left(x\right)\right)^3} + \frac{1}{2}\frac{\lambda^2\int_{\underline{t}}^{\overline{t}} x^2 dF\left(x\right)\ \int_{\underline{t}}^{t} x^2 dF\left(x\right)}{\left(1-\lambda\int_{\underline{t}}^{t} x dF\left(x\right)\right)^4}. \end{aligned}$$

Since

$$C_{S,2}\left(t,\widehat{t}\right) = t\ \frac{d}{dx}E\left[W_S\left(x\right)\right]\Big|_{x=\widehat{t}} + \frac{1}{2}\ \frac{d}{dx}E\left[W_S^2\left(x\right)\right]\Big|_{x=\widehat{t}}$$

condition (4) reduces to $\frac{1}{2}\ \frac{d}{dx}E\left[W_S^2\left(x\right)\right]\big|_{x=\widehat{t}} \geq 0$ *for all* $\widehat{t}$ *, which is satisfied by inspection. With bids that reflect payments per unit of time we can now implement the SPT discipline, whereas only the LPT discipline was implementable when bids reflected total price. Let us compare the two auction formats and the effects induced by the change in discipline: The expected waiting cost of a customer in the SPT and LPT disciplines is*

$$\begin{aligned} &\int_{\underline{t}}^{\overline{t}}\int_{0}^{\infty}\frac{1}{2}\left(t+\widetilde{t}\right)^2 w_K\left(\widetilde{t},t\right)d\widetilde{t}f\left(t\right)dt \\ =\ &\frac{1}{2}\int_{\underline{t}}^{\overline{t}} t^2 f\left(t\right)dt + \int_{\underline{t}}^{\overline{t}} tE\left[W_K\left(t\right)\right]f\left(t\right)dt + \frac{1}{2}\int_{\underline{t}}^{\overline{t}} E\left[W_K^2\left(t\right)\right]f\left(t\right)dt \end{aligned}$$

for $K = S, L$*. Since* $\int_{\underline{t}}^{\overline{t}} tE\left[W_L\left(t\right)\right]f\left(t\right)dt = \int_{\underline{t}}^{\overline{t}} tE\left[W_S\left(t\right)\right]f\left(t\right)dt$ *, the difference in average waiting costs between the SPT and the LPT discipline is* $\Delta\overline{C} = \frac{1}{2}\int_{\underline{t}}^{\overline{t}} E\left[W_S^2\left(t\right)\right]f\left(t\right)dt - \frac{1}{2}\int_{\underline{t}}^{\overline{t}} E\left[W_L^2\left(t\right)\right]f\left(t\right)dt$*. The SPT discipline is more efficient than the LPT discipline if and only if* $\Delta\overline{C}$ *is negative. A-priori, this condition seems to depend on the distribution of processing times* F*, and we do not know whether it generally holds. Numerical calculations for several different cdf's* F *suggest that we have lower expected waiting costs for the SPT discipline. Furthermore, for these cdf's we find that the revenue of the service provider is higher if bids reflect payments per unit of time instead of total payments.*

6 Concluding Remarks

We analyzed a queuing system where customers are privately informed about their processing time, and we derived the endogenous queue-disciplines that result from equilibrium bidding behavior in simple auction schemes. The main equilibrium properties were driven by the curvature of the cost function, and by the ability of the service provider to monitor and collect payments ex-post.

More realistic models that better fit actual situations will have to take into account the interplay of strategic effects caused by asymmetric information on several dimensions such as job value, processing time, due date, etc...Moreover, one would like to understand the effects of queue observability and slot trading. But it should be clear that analytical results will be hard to come by in these more complex models.

In our model the SPT-discipline [LPT-discipline] is approximately efficient [anti-efficient] when the curvature of the cost function is not "too large". The existence of mechanisms that implement better disciplines for arbitrary strictly increasing cost functions remains an open question. To our knowledge, there does not even exist a general algorithm that exactly solves the allocation problem in a model with complete information, not to mention the question of its implementability by realistic mechanisms. For the case of convex but almost linear cost functions, a random allocation of priority (or, say, the FCFS discipline) improves upon the LPT-discipline.

To some extent, our results generalize to the case where the designer can interrupt the actual job, start another, and resume the interrupted one without cost (preemptive queues). The derivation of bidding equilibria in this case is analogous to the one presented here (what changes are the formulas for the distribution of queuing times). As long as bidding functions are strictly monotonic, the service provider can indeed infer the customers' actual processing times from their bids. This is needed in order to implement the Shortest Remaining Processing Time discipline if costs are concave (or the LRPT discipline if they are convex). In addition, for the concave case, we can get efficient par-

ticipation decisions in equilibrium. This is due to the fact that the marginal type who is indifferent between participating and reneging does not exert any externalities. We focused here on an environment without preemption since, for most applications, a cost-free interruption of job processing seems unrealistic.

The impact of reneging in our model is not yet well understood and it will constitute the subject of future work. In equilibrium, reneging will be anticipated by arriving customers, and it will be incorporated in their bids since it reduces their waiting time. It is also not clear what joint condition on the distribution of processing times and the shape of cost functions would guarantee that a waiting customer's situation is constantly improving, thus making reneging irrelevant and allowing the application of our present results. (For example, in a M/M/1 model with deadlines, Hassin and Haviv[1995] showed that a customer's virtual waiting time has the increasing hazard rate property, and therefore that customers will not renege before the deadline.)

7 Appendix A: Laplace Transforms

Proof of Lemma 2: We only provide a sketch of the proof. The derivation follows and adapts the arguments used by Schrage and Miller [1966] and by Conway et al [1967], section 8-6 . We restrict attention to the case of the LPT discipline (the SPT case is very similar).

Assume that arriving customers are allocated to three different priority classes: The highest priority class (class A) consists of all customers with processing time above $t_2 > \underline{t}$; The next highest class (class B) contains all customers with processing time in $[t_2, t_1)$; The remaining types are in the lowest priority class (class C). When a job is finished, the first job from the highest priority class (that is non-empty at that point of time) is processed next. The discipline within each class is FCFS. From the perspective of a class B customer, all class C and class B customers arriving after him are the same. Furthermore, a class B customer's waiting time does not depend on the queuing discipline within classes A and C. Upon arrival, a class B customer faces either an empty system

or one of three different types of *cycles*. A type A and B cycle starts when a class A or B customer arrives at an idle machine; a class C cycle starts whenever a class C customer's job is started. Each cycle lasts until the machine is empty of class A and B customers. Conway et. al show how to compute the Laplace transform of the waiting time associated with each cycle, and the steady-state probabilities of each cycle. Hence, one can compute the Laplace transform of (unconditional) waiting time as the weighted sum of the Laplace transforms associated with the cycles. For our model, this yields:

$$w_L^*(t_1,t_2,s) = \left[\left(1-\lambda\int_{\underline{t}}^{\bar{t}} x dF(x)\right)(s+\lambda(1-F(t_2))(1-g_L(t_1,t_2,s))) + \lambda F(t_1)\left(1-\frac{1}{F(t_1)}\int_{\underline{t}}^{t_1} e^{-(s+\lambda(1-F(t_2))(1-g_L(t_1,t_2,s)))x} dF(x)\right)\right] / \left[s+(F(t_2)-F(t_1))\left(\frac{1}{F(t_2)-F(t_1)}\int_{t_1}^{t_2} e^{-(s+\lambda(1-F(t_2))(1-g_L(t_1,t_2,s)))x} dF(x)-1\right)\right],$$

where

$$g_L(t_1,t_2,s) = \frac{1}{1-F(t_2)}\int_{t_2}^{\bar{t}} e^{-(s+\lambda(1-F(t_2))(1-g_L(t_1,t_2,s)))x} dF(x).$$

The result follows by taking the limit $t_2 \to t_1 = t$. Q.E.D.

8 Appendix B: Linear costs

Throughout this section we assume that $C(t) = ct$, $c \in \mathbb{R}^+$. We show that both the LPT and SPT disciplines can result from equilibrium bidding in the auction where customers make ex-ante payments. The main intuition is that differences in the willingness to pay for priority do not depend on own processing time, but only on the expected queuing time. Therefore the profits from reducing expected queuing time by one time unit is the same for all types, and is exactly reflected by their bids.

The following Lemma, due to Phipps [1956], gives the expected queuing time for a customer with processing time t_i in the non-preemptive SPT and

LPT disciplines. By Q_0 we denote the average remaining processing time of the customer who is currently in service. As in Phipps [1956], we have that $Q_0 = \frac{\lambda}{2}\int_{\underline{t}}^{\overline{t}} x^2 f(x)\,dx$. Here is the argument: If a customer enters the queue while another customer with processing time t is in service, the expected time of entry is $\frac{1}{2}t$. The probability that the new customer indeed arrives in this situation is $\lambda t f(t)$. Q_0 is obtained by averaging over t. (Note that this average contains zeroes for times when no one is in service).

Lemma 8 *(Phipps [1956]) The average queuing time of a customer with processing time t in the SPT discipline is given by*

$$Q_S(t) = \frac{Q_0}{\left(1 - \lambda\int_{\underline{t}}^{t} x f(x)\,dx\right)^2}. \tag{6}$$

The average queuing time of a customer with processing time t in the LPT discipline is given by

$$Q_L(t) = \frac{Q_0}{\left(1 - \lambda\int_{t}^{\overline{t}} x f(x)\,dx\right)^2}. \tag{7}$$

The analysis of Theorem 1 carries over to the linear case. Since we have that $C_K\left(t,\widehat{t}\right) = t + Q_K\left(\widehat{t}\right)$, the following holds:

Theorem 9 *Assume that cost functions are linear.*

1. *The SPT discipline is implemented by the following equilibrium bidding function that is decreasing in processing time :*

$$b_S(t) = c\left(Q_S\left(\overline{t}\right) - Q_S(t)\right).$$

2. *The LPT discipline is implemented by the following equilibrium bidding function that is increasing in processing time:*

$$b_L(t) = c\left(Q_L\left(\underline{t}\right) - Q_L(t)\right). \tag{8}$$

Note that, for $t_1 \neq t_2$, the difference in bids precisely reflects the difference in expected costs due to queuing:

$$b_S(t_2) - b_S(t_1) = c\left(Q_S(t_1) - Q_S(t_2)\right).$$

It is interesting to note that customers do not care about which equilibrium is played: their expected utility is the same in both equilibria: $V_i - ct_i - cQ_S\left(\overline{t}\right) = V_i - ct_i - cQ_L\left(\underline{t}\right)$. This means that all efficiency losses due to higher waiting costs are exactly reflected in the bids, and it implies that the revenue in the LPT equilibrium is lower than in the SPT equilibrium[5].

In the linear case it is somewhat arbitrary to make bids dependent on processing time since the willingness to pay does not depend on own processing time. Consequently, equilibrium bids may depend on processing time in a non-monotonic way (or can even depend on other sources of private information that are uncorrelated with processing time). Consider for example any measurable one-to-one function $k : \left[\underline{t}, \overline{t}\right] \to \left[\underline{k}, \overline{k}\right]$. We can implement a queue-discipline in which a customer with processing time t_i is given priority over all customers with processing time t such that $k(t) < k(t_i)$. Define the average queuing time of a customer with processing time t_i for the described queue-discipline as $\widetilde{Q}\left(k\left(t_i\right)\right)$, and assume that all customers other than i bid according to the increasing bidding function

$$\widetilde{b}\left(k\left(t\right)\right) = c\left(\widetilde{Q}\left(\overline{k}\right) - \widetilde{Q}\left(k\left(t\right)\right)\right).$$

As before, bids exactly reflect the decrease in expected waiting cost vis-a-vis the type who submits the lowest bid. Every bid in the range of $\widetilde{b}$ results in the same payoff and, in particular, $\widetilde{b}$ is an equilibrium bidding function.

References

[2001] Afeche, P. and Mendelson, H. (2001): "Priority Auctions vs. Uniform Pricing in Queueing Systems with a Generalized Delay Cost Structure", forthcoming, *Management Science*.

[1972] Balachandran, K. (1972): "Purchasing Priorities in Queues", *Management Science* **18**, 319-326.

[5]Mathematically, this follows from the the fact that expected waiting time is lower in the SPT case. i.e. $E_t\left[Q_L\right] > E_t\left[Q_S\right]$.

[1954] Cobham, A. (1954): "Priority Assignment in Waiting Line Problems," *Operations Research* **2**, 70-76.

[1967] Conway, R., Maxwell, W. and Miller, L. (1967): *Theory of Scheduling*, Reading: Addison-Wesley.

[2000] Dasgupta, P. and Maskin, E. (2000): "Efficient Auctions", *Quarterly Journal of Economics* **115**, 341-388.

[1978] Dolan, R. (1978): "Incentive Mechanisms for Priority Queueing Problems," *Bell Journal of Economics* **9**, 421-436.

[1975] Edelson, N. and Hildebrand, K. (1975): "Congestion Tolls for Poisson Queueing Processes, " *Econometrica* **43**, 81-92.

[1975] Ghanem, S. (1975): "Computing Central Optimization by a Pricing Priority Policy," *IBM Systems Journal* **14**, 272-292.

[1986] Glazer, A. and Hassin, R. (1986): "Stable Priority Purchasing in Queues," *Operations Research Letters* **4**, 285-288.

[1998] Gross, D. and Harris, C. (1998): *Fundamentals of Queueing Theory,* New York: John Wiley&Sons.

[2002] Hain, R. and Mitra, M. (2002): "Simple Sequencing Problems with Interdependent Costs" mimeo, University of Bonn.

[1995] Hassin, R. (1995): "Decentralized Regulation of a Queue", *Management Science* **41,** 163-173.

[1995] Hassin, R. and Haviv, M. (1995): "Equilibrium Strategies for Queues with Impatient Customers", *Operations Research Letters* **17,** 41-45.

[2002] Hassin, R. and Haviv, M. (2002): *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems,* Boston: Kluwer Academic Publishers.

[2001] Jehiel, P., and Moldovanu, B. (2001) "Efficient Design with Interdependent Valuations", *Econometrica* **69**, 1237-1259.

[1967] Kleinrock, L. (1967): "Optimal Bribing for Queue Positions,", *Operations Research* **15**, 304-318.

[1976] Kleinrock, L. (1976): *Queueing Systems, Vol.2: Computer Applications*, New-York: Wiley&Sons.

[2003] Kittsteiner, T., and Moldovanu, B. (2003): "Auction-based Queue Disciplines", working paper, University of Bonn

[1985] Lui, F. (1985): "An Equilibrium Queueing Model of Bribery," *Journal of Political Economy* **93**, 760-781.

[1991] McAfee, R.P. (1991): "Efficient Allocation and Continuous Quantities", *Journal of Economic Theory* **53**, 51-74.

[1990] Mendelson, H. and Whang, S (1990): "Optimal Incentive Compatible Priority Pricing for the M/M/1 Queue," *Operations Research* **38**, 870-883.

[1969] Naor, P. (1969): "The Regulation of Queue Size by Levying Tolls", *Econometrica* **37**, 15-24.

[1956] Phipps, T. (1956): "Machine Repair as a Priority Waiting Line Problem," *Operations Research* **4**, 76-85.

[1983] Ross, S. (1983): *Stochastic Processes,* New York: John Wiley&Sons.

[1968] Schrage, L. (1968): "A Proof of the Optimality of the Shortest Remaining Processing Time Discipline," *Operations Research* **16**, 687-690.

[1973] Schrage, L. (1973): "Optimal Scheduling Rules for Quadratic Waiting Costs," *Proceedings of the Seventh Annual Princeton Conference on Information Sciences and Systems*

[1966] Schrage, L. and Miller, L. (1966): "The Queue M/G/1 with the Shortest Remaining Processing Time Discipline," *Operations Research* **14**, 670-684.

[1979] Tilt, B., and Balachandran, K.R. (1979): "Stable and Superstable customer policies in queues with balking and priority options", *European Journal of Operational Research* **3,** 485-498.

[2003] *The Economist:* "Debt Collecting; Knok Knock", November 1, 2003, 39.

[1995] Van Mieghem, J.A. (1995): "Dynamic Scheduling with Convex Delay Costs: the Generalized $c\mu$ Rule", *The Annals of Applied Probability* **5**(3), 808-833.

[1983] Wilson, B. (1983): *Non-Linear Pricing*, Oxford: Oxford University Press