

## PRICING AND SCHEDULING DECISIONS

GAD ALLON  
Kellogg School of Management,  
Northwestern University,  
Evanston, Illinois

Service providers as well as make-to-order manufacturers use priority schemes in conjunction with pricing as allocation mechanisms. By carefully selecting the appropriate scheduling scheme and the optimal prices, service providers can best utilize their resources to improve their customers' social welfare or improve their overall value (through improved service quality or cost reduction). This article studies the interplay between pricing and scheduling decisions focusing on the interaction between the service provider and the delay-sensitive customers, while dealing with different types of providers (a social planner, a monopolist, or oligopolies) in different types of service settings (either when the queues are observable or when they are not), and different types of customer populations (homogeneous or heterogeneous, segmented or unsegmented).

In exploring the relationship between pricing and scheduling decisions, we give separate treatment to different settings according to the role of service providers in the economy:

- *Social Planner*. In this case, the capacity of the service process is viewed as a scarce resource, and thus granting priorities is treated as such. The role of the pricing of the service is usually viewed in this context as a transfer mechanism to guarantee that only those who need high priority (due to high cost of delay or short service time) are granted this priority. In particular, pricing is used to ensure that the pricing-priority mechanism is incentive-compatible in the sense that customers have no incentive to misrepresent their true identity

when they hold such information privately (that is to claim that their delay cost is higher or service time is shorter). The pricing may also play a role in regulating usage in the system as a whole; this is a common practice in managing congested systems.

- *Monopolistic View*. In such a setting, the firm would like to use priorities and price them so as to improve its profits by virtue of price differentiation. When facing different types of customers, with different sensitivities to delay and different values obtained from the service, serving customers according to a priority rule while charging a priority-dependent price can outperform charging a fixed price, and serving according to a first-come-first-served manner. In settings in which there is asymmetric information regarding the identity of the customers, the prices and the priority schemes should be selected so that the overall pricing-priority mechanism is incentive-compatible.
- *Competitive View*. When competing in a market in which customers are sensitive to delay, firms can choose their prices, waiting time standards, capacity levels, and priority schemes. The role of the priority scheme is to allow the firm to accommodate all demand streams (driven by both the prices and the service level of the different segments), while meeting the promised service levels. As we will show, the price charged to each segment by each firm depends on the characteristics of all firms and also on the level of priority that the firm allocates to this specific segment.

We will initially review the use of priorities in models, where a social planner sets the prices and the scheduling rules. The main focus of these models will be on the incentive compatibility of the joint pricing–scheduling decisions. We will then survey settings where profit-maximizing firms use the combination

of pricing and scheduling, and which we will use to highlight the key differences between a social planner and a monopolist. We will initially discuss a setting in which the firm is not informed about the customer's waiting time cost and value, giving rise to incentive-compatible pricing and scheduling. We will also discuss the setting in which the queues are observable, giving rise to the phenomenon of "following the crowd" (which we will describe in detail). We will complete the article with a discussion of the role of priorities and pricing in the context of a competitive market. In this context, we will also discuss how pricing and scheduling decisions affect the benefits of each segment depending on the level of priority given to it. Note that the article deals with the interplay between pricing and scheduling in settings, which are modeled using queuing systems; that is, service settings and make-to-order manufacturers, and does not address make-to-stock settings.

These decisions have been considered (jointly) also in a make-to-stock environment. For examples of such models, see Charnsirisakskul *et al.* [1]. Also, for a more complete discussion of possible scheduling rules in queuing, see the article titled *Power Indices*. For a comprehensive discussion of strategic customer behavior in queuing, see the article titled *Strategic Customer Behavior in a Single Server Queue*. Lastly, for a discussion of the interplay between pricing and lead-time decisions, see the article titled *Pricing and Lead-Time Decisions*.

#### SOCIAL WELFARE MAXIMIZATION: INCENTIVE COMPATIBILITY

In many service systems, different customers typically have rather disparate sensitivities to the delay encountered, and obtain different value from the service. In order to improve the social welfare, a social planner can exploit these differences by optimally prioritizing among the different customers while charging them the optimal prices. The social planner, who maximizes the social welfare of the players, views its limited available

capacity as a scarce resource that needs to be allocated. Using the same logic, high priority is a scarce resource that needs to be properly allocated to those who need it. However, the design of optimal price and scheduling mechanisms may require knowledge of characteristics of the customers. These characteristics are seldom known to the queue manager and must be estimated or obtained from the customers' own statements, a situation that may create an incentive for customers to declare untruthful values.

The role of pricing, in many service systems, is to modulate demand. This is the case in particular for service providers striving to maximize social welfare, where customers are asked to pay for their use in order to make sure that the optimal usage level is achieved. Naor [2], for example, showed that in the context of homogeneous customers, tolls have to be levied in order to make sure that customers do not join the system when it is too congested. When customers are heterogeneous in terms of their waiting time cost and service times, the role of pricing is not only to regulate the usage in the different priority levels, (where, within each customer class, the optimal level of usage is achieved) but also to guarantee that customers either disclose their true costs or use the priority level that was designed to their needs and do not exploit the information asymmetry to obtain a higher priority than the one they deserve.

The question that this section tries to answer is how to construct a pricing mechanism that induces all customers to make decisions, which are not only optimal from their myopic, self-interested point of view, but also optimize the overall objective function of the organization.

#### Pricing Based on Externalities

Consider a model in which customers are identical except for having different costs of waiting, and priority levels are assigned to customers according to their declared costs. The information that is available to an arriving customer includes the queue length upon his arrival, the declared time values of the customers in the queue, and the residual service time of the customer in service. Using

this information, the customer declares a value to his time and obtains priority accordingly. The goal is to induce customers to declare their true time value, so that the resulting order of service optimizes social welfare. Dolan [3] suggested the use of Clarke prices. The idea is that each customer pays for the costs that he imposes on others when joining the queue. These costs are calculated given that customers reveal their true time value, which means that a customer pays an amount that is equal to the externalities that he imposes on other customers. It follows that under this pricing rule, the strategy that prescribes declaring the true time value is an equilibrium one. If a customer overstates his time cost, he must compensate the customers, who will have to wait longer as a result of this deviation by exactly the cost he saved (due to the work-conservation principle). Since these customers have higher waiting costs than his, compensating these customers for their increased waiting time costs the customer more than what he saves.

The model of Dolan [3] focuses on the ability to use a bidding rule, which is based on externalities, geared to induce customers to reveal their true delay cost. However, the model remains silent regarding the ability to use the pricing and the priority schemes to regulate usage and to control the arrival rate to the system, either to maximize profits or social welfare. Mendelson and Whang [4] address settings in which the arrival rate to the system is a function of the prices and the delay that the customers experience in the system.

#### Incentive Compatibility of the $c\mu$ Priority Rule

We now turn to the more general case, in which the service provider has to use pricing to regulate usage while being incentive-compatible. Consider a system that is modeled as an M/M/1 queuing system with multiple classes. Each class is characterized by its delay cost per unit of time and the expected service requirement. The goal is to find a pricing mechanism that is also optimal; that is, that the arrival rates and the scheduling scheme jointly maximize the expected value of the system while allowing each customer to make individual decisions

on whether or not to purchase the service and at what priority level. In making these decisions, while the customers are assumed to maximize their own utility functions under the optimal incentive-compatible pricing and scheduling schemes, they also maximize the objective of the organization as a whole.

When minimizing the expected organization-wide delay cost per unit of time, it is well known that given an exogenously specified arrival rate, and given that the delay cost and the expected service requirements are known, the optimal scheduling scheme is the so-called  $c\mu$  rule, which gives absolute priority to customers according to the ratio of their delay cost to the expected service time required to complete their request. It is clear from the definition of the priority scheme that knowledge of these parameters is crucial for successful execution.

Mendelson and Whang [4] address the problem of resource pricing, which determines the demand rate of customers from each class, and priority pricing, which determines which customers are served at each point in time. For the case in which all customers have the same mean service times, regardless of their class, the authors show, in an analogous manner to Dolan [3], that charging each customer the externalities that he imposes on other customer classes under the optimal demand level, while using the  $c\mu$  rule to prioritize among customers, is both optimal and incentive-compatible. In particular, the authors show that the optimal joint pricing-priority scheme automatically achieves incentive compatibility.

In the case of identical mean service times, the authors show that a price that depends only on the priority level is optimal and incentive-compatible. However, for the more general case, where customers differ in their delay cost as well as in their mean service time, such a pricing scheme may not be incentive-compatible anymore. In that case, the authors suggest a more general pricing scheme, which can be decomposed into two parts: a basic charge and a priority surcharge. The basic charge corresponds to the price of the lowest priority class, is equal for all customers (regardless of their class or the priority they choose), and is quadratic in the

service time. The priority charge is proportional to the service time with a coefficient that increases as the priority level goes up. One may view this pricing scheme as “externality pricing” in the following sense: when a customer buys a priority level, he pays the conditional expected externality of his decision on the other customers.

The main finding in this article is that the incentive-compatible pricing scheme induces the customers to behave in a way that is consistent with this rule, while aligning their interest with that of the overall organization. For a social planner, the prices are treated as transfer payments, whose objective is to merely be an instrument to regulate usage (both in the system as a whole and in each priority level). We next turn to ask how the pricing scheme and the scheduling decisions will change in a monopolistic setting.

#### PROFIT-MAXIMIZING PRICING AND SCHEDULING

Afeche [5] answers the question of how a capacity-constrained firm should design an incentive-compatible price-scheduling mechanism to maximize revenues from a heterogeneous pool of time-sensitive customers with private information on their willingness to pay, time-sensitivity, and processing requirements.

The article provides the following insights: First, the author shows that the familiar  $c\mu$  rule, which is known to minimize the expected delay cost and to be incentive-compatible under social welfare optimization, need not be optimal in this setting. This specific fact suggests a more general guideline: in designing incentive-compatible and revenue-maximizing scheduling policies, delay cost-minimization, which plays a prominent role in controlling and pricing queuing systems, should not be a dominant criterion. Second, the author identifies optimal scheduling policies with novel features. One such policy prioritizes the more time-sensitive customers but voluntarily delays the completed orders of low priority customers. This insertion of *strategic delays* deters time-sensitive customers

from purchasing the lower-priority class. In other situations, the author shows that it is optimal to appropriately randomize priority assignment, in one extreme case, serving customers in reverse  $c\mu$  order, which maximizes the system delay cost among all work-conserving policies. The author also shows that the optimal level of delay differentiation systematically emerges from a trade-off between operational constraints and customers’ incentives.

#### Numerical Example

In order to demonstrate the role of intentional delays, we will use a simple numerical example. Consider a firm that serves two types of customers:

1. H-type customers obtain \$100 from the service and incur \$20 per unit of time spent in the system.
2. L-type customers obtain \$30 from the service and incur \$4 per unit of time spent in the system.

We will assume that the arrival rates from each class is 0.2 H-type customers per unit of time, and 0.3 L-type customers per unit of time, independent of the price charged. We also assume that  $\mu = 1$ .

If the firm employs a  $c\mu$  rule, the sojourn time of an H-type customer is  $1/(1 - 0.2) = 1.25$ . Using a similar logic, an L-type customer’s waiting time under the  $c\mu$  rule is 2.5 time units.

If the firm tries to extract the entire consumer surplus, it would charge \$75 =  $100 - (1.25 \times 20)$  from each H-type customer, and \$20 =  $30 - (2.5 \times 4)$  from each L-type customer. This can presumably result in a revenue rate of  $(0.2 \times 75) + (0.3 \times 20) = \$21$  per unit of time.

However, this solution is not incentive-compatible: Under this price and under this priority scheme, an H-type customer would opt to purchase the service grade designed for an L-type customer; this will leave him with utility of  $100 - (2.5 \times 20) - 20 = 30$ , which is better than what he obtains by purchasing the service grade “designed” for him, which is  $100 - (1.25 \times 20) - 75$ .

Under the  $c\mu$  rule, the highest incentive-compatible price that the firm can charge an H-type customer is \$45, which results approximately in a revenue rate of \$15 per unit of time. However, if the firm chooses to inject one unit of time of intentional delays for all L-type customers, the firm can charge only \$16 from each L-type customer, but the incentive-compatible price for the H-type customers is then \$61, which results in an increased revenue rate of  $(3.2 \times 0.2) - (1.2 \times 0.3) > 0$ . This means that the firm can improve its revenues by inserting artificial unnecessary delays. By degrading the quality of the low priority service, the service provider can charge more for high priority service. Note that in this case, the idleness is increasing the overall delay cost, yet the provider shrinks the pie in order to be able to “eat a larger piece.”

#### Bidding for Priorities in Observable Systems

We now turn to a discussion of the role of priority in systems in which the queue is observable to an arriving customer. We will first discuss several models that focus on the consumer behavior in such systems (fixing the priorities and the prices) and then discuss the behavior of a monopolist in such a market.

Consider an observable M/M/1 queue in which arriving customers choose from a discrete set of possible payments. In this model, customers are assigned priority levels according to their payments. All customers arriving to the system must obtain service (i.e., balking is not allowed), and customers are not informed on the payments made by others. Balachandran [6] shows that the equilibrium strategy of all customers is to choose the lowest price that will guarantee joining the head of the queue, which induces a “last-come-first-served” scheduling policy, when implemented by all customers. The rationale behind this equilibrium strategy is that a customer purchases high priority level not only because he can be scheduled ahead of other customers, but also so that others are not scheduled ahead of him. In many service systems, this leads to what Hassin and Haviv [7] refer to as “follow-the-crowd” behavior in which the more customers purchase priority, the more inclined an individual is to do so himself.

Haviv and Hassin [8] study a model with multiple priority classes in which agents can decide on joining strategies. While in a single class model, customers should follow a threshold policy; that is, join as long as the queue is below a certain level and balk otherwise. They showed that this is not true in a multiclass case.

The above models fix the pricing scheme and focus on the equilibrium behavior among customers in making the decision when to pay for priority, or how much to pay given the set of possible prices and priorities and in that, they are more descriptive of customer behavior in the presence of priorities.

The question that arises is how a monopolist utilizes such a consumer behavior when deciding what priority levels to offer the customers to choose from.

Alpersteins [9] considers a similar model in which the decisions on the available priorities and the associated prices are made by a profit-maximizing firm. Alperstein showed that the profit increases with the number of priority levels. The main conclusion from this model is that the profit-maximizing pricing scheme is one that induces a threshold of one for each priority type except for the highest one. That is, an arriving customer would always choose the lowest possible priority level that is above any existing customer.

#### COMPETITIVE MODELS

We will next discuss models in which service providers use prices in conjunction with priorities, to compete in the market. Examples of such industries are numerous. Banks and credit card companies segment their customers into regular and VIP or Gold and Platinum customers. Computer software and hardware firms often segment their customers, for example, into home and home office users, small businesses, large businesses and the government, education, and health-care sectors, using an integrated pool of technical support personnel to serve the different customer segments according to a specific priority discipline; each customer segment is associated with a specific price and waiting time expectation. Finally, overnight delivery services use their planes

and trucks to deliver letters, boxes, and cargo, each with different prices and delivery time standards. In many service industries, waiting time standards are used as a primary advertised competitive instrument. For example, most major electronic brokerage firms, (e.g., Ameritrade, Fidelity, and E-trade) all prominently feature the average or median execution speed per transaction, which is monitored by independent firms.

References Loch [10], Lederer and Li [11], and Armony and Haviv [12] as well as Allon and Federgruen [13] are the only published articles we are aware of that have directly addressed competition models in which waiting-time-sensitive customers are segmented into multiple classes. We will briefly survey the first three and then discuss in greater detail the model presented in Allon and Federgruen [13].

Loch [10] considers an industry with M/M/1 service providers and two customer classes, each with a given waiting cost rate and average service time. All customers within a class select the firm, which offers the lowest full price; that is, the sum of the direct price and the waiting cost, where the total demand volume in the class is given by a known function of this full price value. Under quantity competition, the author establishes the existence of a Nash equilibrium under which the customers are prioritized according to the  $c\mu$  rule. Lederer and Li [11] generalize this model to allow for an arbitrary number of nonidentical M/G/1 firms and an arbitrary number of customer classes. Assuming the firms engage in price competition, the authors establish the existence of a Nash equilibrium, under which each firm, once again, prioritizes customers according to the  $c\mu$  rule. The existence result is based on the assumption that each class' expected waiting time at a given firm is a convex function of all of the firm's demand rates for the different customer classes. Note also that while in Afeche's [5] monopoly model, the incentive-compatible optimal policy cannot, in general, be based on the  $c\mu$  priority rule; the  $c\mu$  priority rules are part of the Nash equilibrium in Lederer and Li's [11] perfect price competition model (provided the above convexity assumption is satisfied).

Armony and Haviv [12] analyze a two-stage competition model with two M/M/1 service providers and two customer classes, each, again, acting as a single entity in deciding what fraction of its business to assign to each of the providers. In the first stage, the two providers compete with each other by announcing their service charges. In the second stage, the customer classes compete with their allocation decisions. They show that pure price equilibrium may fail to exist in this two-stage game.

All of the previous models assume that the service organization (or the make-to-order manufacturer) has a fixed capacity and thus the role of the priority rule is to allocate the capacity among the different customers according to their cost of waiting as well as their expected service requirements.

We next present a more general model in which the firm can invest in capacity in order to accommodate the different customers, in conjunction with a priority rule, in a way that will allow the firm to position itself in the market. In such markets, firms cater to multiple customer classes or market segments with the help of shared service facilities or processes, so as to exploit pooling benefits. As in all of the settings studied in the previous section, different customer classes typically have rather disparate sensitivities to the price of service as well as the delays encountered. Thus, from the firm's perspective, it is vital to offer differentiated service charges and levels of service to different customer classes so as to maximize (long run) profits.

Allon and Federgruen [13] propose and analyze a model in which firms select all or part of the following: (i) the prices charged to all customer classes, (ii) the waiting time standards promised to all classes measured in terms of the average waiting time, (iii) the capacity level, and (iv) a priority discipline enabling the firm to meet the promised waiting time standards under the chosen capacity level. While making these decisions in a competitive setting can be quite complex, one can observe that while the price and the waiting time decisions impact all firms in the industry, the capacity and the priority (given the demand and the waiting time standard) impact only the cost of the firm itself. The

authors, thus, first present the capacity problem, in conjunction with the priority setting. The authors then embed them in the competitive setting in which firms compete in terms of their prices.

#### Capacity Choice and Associated Priority Rules

Consider a service provider, modeled as an M/M/1 queuing facility. Each customer class generates an independent Poisson stream of customers to this service provider at the rate determined by the above-mentioned demand functions. Its service times are i.i.d. with a firm- and class-dependent service rate that is proportional to the firm's capacity level. Each firm incurs a given class-dependent cost per customer, as well as a cost per unit of time, proportional to the adopted capacity level. Each firm attempts to maximize its expected profits.

Allon and Federgruen [13] derive the analytical expression of the capacity level that each firm needs to adopt to accommodate a given vector of demand volumes and waiting time standards under an optimal associated dynamic priority rule. They show that this capacity level is the maximum of a number of closed-form capacity bounds, one for each subset of the customer classes. Interestingly, for arbitrarily specified waiting time standards, the maximum may be achieved for a strict subset of the collection of all classes, the so-called *bottleneck set*, in which case strategic idleness times, that is artificial after-service delays, may be adopted for the so-called *residual* classes outside the bottleneck set. The optimal capacity level is to be complemented with a randomized absolute priority rule.

The authors study three types of competition: price competition (where waiting times are exogenously determined), waiting time competition (when the prices are exogenously determined), and simultaneous price-waiting time competition. They show that in the waiting time and the simultaneous competition models, the bottleneck set is always the full set. However, this is not necessarily true in the price competition model, where waiting times are not part of the strategic choices. Thus the need for intentional delays may arise only under price competition.

It is important to note that we again see the use of intentional delays, when selecting capacity levels and priority schemes as observed already, in a different setting and for a different purpose in Afèche [14].

#### Pooled versus Dedicated Capacity

The authors then compare the equilibria that arise when the firms use pooled capacity with those achieved when the firms service each market segment with a dedicated service process; that is, without pooling service resources. In the price competition model, for example, the equilibrium is obtained, both under service pooling and dedicated service facilities, when for each class the relative markup vis-a-vis the marginal cost equals the reciprocal of the demand elasticity. The marginal cost per customer per unit of time always consists of the variable service cost plus the marginal capacity cost (per unit of time). When service is provided with dedicated facilities, the marginal increase in the required capacity equals the expected amount of work per customer of the considered class. Under service pooling, it is zero for residual classes and less (or more) than this benchmark value, depending upon whether the customer class receives worse (or better) than average service.

Allon and Federgruen [13] then show that, while all firms reduce their total cost by switching from dedicated to pooled service, these cost savings do not necessarily result in price reductions for all customer classes. Indeed, if a customer class gets better than average service (and belongs to the bottleneck set) at all firms, it is charged a higher price under service pooling than under dedicated service. The following provides some intuition behind this result: assume all classes initially get identical normalized waiting time standards; if class 1, say, subsequently bargains for a lower waiting time standard, the cost for the other customer classes increases, for which externalities class 1 is made to pay.

#### CONCLUSIONS

Many service providers use pricing and scheduling schemes in order to improve

their value proposition. In summarizing this article, it is important to highlight several important concepts that arise in the intersection between pricing and scheduling.

1. *Externalities-Based Pricing.* The concept of charging customers the externalities that they impose on other customers arises in settings in which the customers have private information regarding their delay cost as well as in settings in which firms compete while investing in a pooled capacity to accommodate different segments. The former is used to achieve incentive compatibility while the latter to make sure that different segments, those with high service levels as well as those with low service levels, are charged a competitive price.
2. *Intentional Delays.* While in many cases service organizations set priorities to minimize the total waiting time cost, via the  $c\mu$  rule, this is not always the case. Both, when a monopolist needs to set priorities that are incentive-compatible and profit-maximizing, as well as in the case in which the firm needs to choose its capacity level in conjunction with priority rules in order to accommodate different service levels, the firm might be introducing intentional delays. The idea is to delay the customers of one of the segments upon completion to achieve different goals that cannot be achieved when restricted to waiting-time-minimizing rules.
3. *Follow-the-Crowd Behaviour.* The consumer behavior in the simplest service system, the M/M/1 queue, can be characterized as “avoid-the-crowd” behaviour: customers prefer joining the queue when others do not do so. However, in the presence of priorities, the consumer behavior can be characterized as “follow-the-crowd” behaviour: the more customers purchase priorities, the more likely other customers purchase these as well. As discussed above, a monopolist can

exploit such behavior when introducing pricing and scheduling scheme to customers.

## REFERENCES

1. Charnsirisakskul K, Griffin P, Keskinocak P. Pricing and scheduling decisions with lead-time flexibility. *Eur J Oper Res* 2006;171(1): 153–169.
2. Naor P. The regulation of queue sizes by levying tolls. *Econometrica* 1969;37(1):15–24.
3. Dolan RJ. Incentive mechanisms for priority queuing problems. *Bell J Econ* 1978;9:421–436.
4. Mendelson H, Whang S. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper Res* 1990;38:870–883.
5. Afeche P. Incentive-compatible revenue management in queuing systems: Optimal strategic idleness and other delaying tactics. Working paper, Kellogg School of Management, Northwestern University; 2004.
6. Balachandran KR. Purchasing priorities in queues. *Manag Sci* 1972;18:319–326.
7. Hassin R, Haviv M. To queue or not to queue: equilibrium behavior in queuing systems. Boston (MA): Kluwer Academic Publishers; 2003.
8. Hassin R, Haviv M. Equilibrium threshold strategies: the case of queues with priorities. *Oper Res* 1997;45:966–973.
9. Alperstein H. Optimal pricing policy for the service facility offering a set of priority prices. *Manag Sci* 1988;34:666–671.
10. Loch C. Pricing in markets sensitive to delay [Ph.D. Dissertation]. Stanford (CA): Stanford University; 1991.
11. Lederer PJ, Li L. Pricing, production, scheduling, and delivery-time competition. *Oper Res* 1997;45(3):407–420.
12. Armony M, Haviv M. Price and delay competition between two service providers. *Eur J Oper Res* 2003;147(1):32–50.
13. Allon G, Federgruen A. Competition in service industries with segmented markets. *Manag Sci* 2009;55(4):619–634.
14. Afeche P. Incentive-compatible revenue management in queuing systems: optimal strategic delay and other delay tactics. Toronto: Rotman School of Management; 2006. Working paper.