

Designing and Pricing Incentive Compatible Grades of Service in Queueing Systems

Tomer Yahalom, J. Michael Harrison and Sunil Kumar

Graduate School of Business, Stanford University

January 3, 2006

Abstract

A profit-maximizing service provider (SP) confronts several classes of potential subscribers that differ in their delay sensitivity. The SP may offer several grades of service at different prices, those service grades being operationally defined by the scheduling rule used when jobs compete for capacity. We characterize the provider's optimal, incentive compatible menu of service grades and their associated prices, generalizing an earlier analysis with linear delay costs.

1 Introduction

In this paper we study a service provider (SP) who sells long-term subscriptions to a diverse group of customers having private information about their delay sensitivity and price sensitivity. A subscriber generates a stream of jobs over time, each requiring some service, and jobs may experience delay before service because of competition for limited capacity. There are several customer classes that differ in their delay sensitivity, and there is a demand curve associated with each class that specifies how many customers of that class will subscribe for any given total cost (average delay cost + price) per job.

In order to maximize profit, the SP offers a menu of grades of service (one grade for each customer class), those grades being operationally defined by a scheduling rule. Scheduling jobs for service is a particular kind of dynamic control, and throughout this paper the words “scheduling” and “control” will be used interchangeably. The SP further chooses prices so as to induce the desired subscription volumes for each grade of service. Examples for such settings would be shipping companies (such as FedEx and UPS) that offer regular and expedited services, or a computer help desk that offers customers different priorities in long-term support.

The problem described above involves monopolistic price discrimination, and more specifically, the segmentation of a market by means of multiple products that differ in their “quality.” In our setting, delay before service constitutes un-quality. There is a large literature on this topic, dating back at least to [14], [11], [13], but in most of that literature the firm can provide any grades of service it may choose (at some cost). In our problem there is a limited processing capacity used

to provide the service, and therefore some quality degradation due to queuing is inevitable. Of course, queuing delays may be more severe for customers choosing one grade of service than for those choosing another grade, depending on how jobs are scheduled. Moreover, the SP may choose to delay some jobs intentionally, beyond what is inevitable due to queuing, in order to achieve a desired segmentation.

One may also say that our service provider confronts a *screening problem*. (In the literature of information economics, “screening” and “adverse selection” are used more or less interchangeably; the former term will be used throughout this paper.) In that regard it is useful to consider the following simpler scenario. Suppose that the SP can directly observe the class (that is, directly observe the delay sensitivity) of each potential subscriber, and make different take-it-or-leave-it offers to members of different classes. Designing and pricing grades of service for the different classes in that simpler scenario will be called the *full information problem* hereafter. The following is a general result in information economics: when the SP defines and prices grades of service so as to maximize profit in the full information problem, the resulting subscription volumes and product offerings are *socially optimal*, meaning that they maximize the sum of the profit realized by the SP and the economic surplus realized by subscribers. Thus the term “social optimization,” which plays a prominent role in the literature surveyed below, can alternatively be construed as profit maximization in the full information problem.

Social optimization in the Markovian version of our problem was studied in [10] with linear delay costs, and the *cμ rule* (a static priority scheme) was shown to be the optimal scheduling rule. Van Mieghem [16] studied essentially the same problem, but with convex delay costs and weaker distributional assumptions, in the “heavy traffic” parameter regime, proving the asymptotic optimality of a more complex control policy that he called the *generalized cμ rule* (*Gcμ rule*). Later, [17] showed that the *Gcμ rule* is socially optimal in the heavy traffic problem with private information (meaning that the *Gcμ rule* is also incentive compatible), when jobs’ inter-arrival times and service durations have similar distributions across classes.

The immediate stimulus for our work was the recent paper by Afeche [1], who studied a particular case of the service provider’s screening problem, but under a *profit maximization* objective. To be specific, Afeche considered a Markovian version of the problem with two customer classes, linear delay costs, and linear demand curves. In that setting he showed that the optimal scheduling rule may involve radical departures from the *cμ rule*, including intentionally delaying some jobs in the interest of incentive compatibility (IC). Afeche’s problem is *not* a special case of the “classical” or “standard” screening problem referred to earlier [14], [11], [13], because it involves a capacity constraint, and his solution is quite different from what one sees in the classical theory. In particular, his solution may involve “upward distortion at the top”, meaning that the class most sensitive to delay is offered better service in the screening problem than it would be offered in the full information problem. (This divergence from classical screening theory is not explicitly noted in Afeche’s paper.)

In this paper we generalize Afeche’s treatment by relaxing the distributional assumptions, allowing general (downward sloping) demand curves, and most importantly, by allowing convex delay costs of the following “multiplicatively separable” form: customers of class i who are delayed d time units experience a delay cost of $a_i C(d)$, where $C(\cdot)$ is a convex, non-decreasing function

($a_1 > \dots > a_n$ without loss of generality). In this setting we use an *achievable region* approach to characterize the optimal solution to the profit-maximization problem.

In a model with linear delay costs, the only relevant performance characteristics of the service provider’s processing system are the *mean* delays experienced by customers choosing various grades of service. It is quite easy to determine the mean delay combinations that are achievable using available scheduling rules (this is what we call the “achievable region”), and to determine specific scheduling rules that achieve the Pareto efficient combinations; basically, one need only consider randomization among simple priority rules, cf. [1]. With general convex delay costs, however, the foundational queuing theory required for such an explicit treatment does not exist. For that reason we begin our analysis with an abstract model where the achievable regions (one for each vector of grade-specific subscription volumes) are treated as system data, and where no attempt is made to associate specific scheduling rules with achievable performance profiles. This level of abstraction is adequate for obtaining broad structural results, but nothing can be said about the best way to schedule jobs (that is, about how to define “service products”) without a more detailed consideration of dynamic system behavior.

To address that issue, the second stage of our analysis features a richer model structure in which achievable regions are generated from an idealized sub-model of congestion and control. To be specific, we adopt the *Brownian model* of dynamic system behavior, but only some elements of the Brownian model are really essential for our purposes. Just as Afeche concluded that a variant of the $c\mu$ rule is optimal for the SP in his model with linear delay costs, it will be shown that a variant of Van-Mieghem’s $Gc\mu$ rule, proven to minimize total delay costs in the Brownian model, is optimal in our setting. (Here the word “variant” must be interpreted broadly enough to include the phenomenon of intentional delay). However, unlike Afeche’s variants, the optimal scheduling rule in our setting can be interpreted as a $G\hat{c}\mu$ rule, which is a $Gc\mu$ rule for virtual classes (classes with virtual cost coefficients \hat{a}_i). This means the optimal solution to our private information problem can be found by solving its full information version for virtual classes.

The remainder of the paper is structured as follows. Our abstract system model, in which achievable regions are treated as given system data, is formulated and analyzed in Section 2. In the interest of simplicity, we restrict attention to the case of two customer classes in the main development, but as noted in Section 2.5, our broad structural conclusions for the abstract model carry over to the case of n classes with very little difficulty. The idealized sub-model of congestion and control is developed in section 3, which allows the sharp characterization of optimal job scheduling described above. (As noted above, the question of how to schedule jobs is essentially the question of how to define product offerings in our setting.) For concreteness, the discussion includes a complete analysis of an example with quadratic delay costs. Throughout Sections 2 and 3, our multiplicative separability assumption plays a crucial role. That assumption, which is satisfied in Afeche’s [1] setting with linear delay costs, gives a model with what economic theorists call “one-dimensional types.” In Section 3.5 we analyze an example where multiplicative separability is absent, showing that entirely new phenomena may then occur.

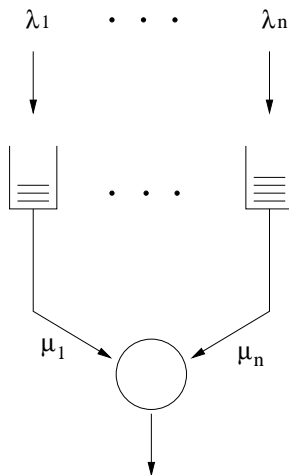


Figure 1: Multi-class queue

2 Problem Formulation and Structure of Solution

Customers of two different classes are interested in subscribing for a long-term service, in which they generate jobs to be processed by a Service Provider (SP). Customers of class i incur a delay cost of $C_i(d) = a_i C(d)$ per job they generate that is delayed for d units of time. We call this assumption *multiplicative separability*; $C(\cdot)$ is a convex and nondecreasing nominal cost function, common to all classes, and a_i is a class-specific coefficient amplifying that cost. Without loss of generality we assume $a_1 > a_2$. In addition to the cost coefficient a_i , which one might also call an “impatience parameter”, an inverse demand curve $P_i(\cdot)$ is given for each class i , which can be interpreted as follows. For $\lambda_i \in [0, u_i]$, $P_i(\lambda_i)$ gives the total cost (expected delay cost+price) per job of class i that would result in an average arrival rate λ_i of class i jobs to the system. We assume that $P_i(\cdot)$ is continuous and monotonically decreasing with $P_i(u_i) = 0$, so that u_i is the maximum class i subscription volume (see Figure 2). Jobs of class i have a mean service time $\frac{1}{\mu_i}$ (see Figure 1).

A customer’s relationship with the service facility is long-term, and therefore we refer to customers as subscribers. These subscribers are not concerned with the delay of a particular job of theirs, but rather with the overall average delay cost per job they incur. As a means of interpreting the formal definition advanced in the next paragraph, let us suppose that the delays experienced by class i customers are distributed as a random variable D_i (of course, this delay distribution depends on the vector λ of average arrival rates and on the scheduling rule used by the SP, which effectively defines the various grades of service), and then set $\xi_i = EC(D_i)$. Obviously, ξ_i is a real-valued delay measure for class i jobs. Denoting by p_i the price charged per class i job, the triple (λ_i, p_i, ξ_i) must then satisfy the individual rationality (IR) constraint

$$P_i(\lambda_i) = p_i + a_i \xi_i.$$

The vector $\xi = (\xi_1, \xi_2)$ summarizes everything about the system performance that is relevant for customers’ decision making. Let $\Lambda = \{\lambda = (\lambda_1, \lambda_2) \mid 0 \leq \lambda_i \leq u_i, i = 1, 2\}$. Then for fixed arrival rates $\lambda \in \Lambda$, we say that a performance vector ξ is *achievable* if there exists a scheduling rule that results in those measures of delays. Because capacity is limited, not every performance vector is

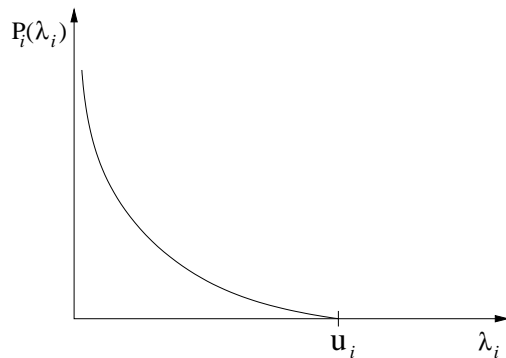


Figure 2: Inverse demand curve

achievable; for example $\xi = 0$ is not achievable for $\lambda > 0$. For a given $\lambda \in \Lambda$, let $R(\lambda)$ denote the set of achievable performance vectors ξ . We refer to $R(\lambda)$ as the *achievable region*.

2.1 The Achievable Region Approach

We take as system data a parametric family of triples $(\xi_L^\lambda, \xi_H^\lambda, r^\lambda)$, defining the achievable regions, where $0 \leq \xi_L^\lambda < \xi_U^\lambda < \infty$ and $r^\lambda : [\xi_L^\lambda, \xi_U^\lambda] \rightarrow \mathbb{R}_+$ is a convex, monotonically decreasing curve for all $\lambda \in \Lambda$. The achievable region $R(\lambda)$ can be constructed from such a triple in the following way.

$$R(\lambda) = \left\{ \xi = (\xi_1, \xi_2) \mid \exists \hat{\xi}_1 \leq \xi_1 \text{ such that } \xi_L^\lambda \leq \hat{\xi}_1 \leq \xi_U^\lambda, r^\lambda(\hat{\xi}_1) \leq \xi_2 \right\},$$

see Figure 3. Stated simply $R(\lambda)$ is the set of all points above and to the right of $r^\lambda(\xi_1)$ for $\xi_1 \in [\xi_L^\lambda, \xi_U^\lambda]$. We interpret r^λ as the *Efficient Frontier* (EF), meaning that the set $\{(\xi_1, r^\lambda(\xi_1)) \mid \xi_1 \in [\xi_L^\lambda, \xi_U^\lambda]\}$ is the set of Pareto efficient performance vectors, and indeed any point above and to the right of the EF is achievable through intentional delay. Moreover, if $\xi_1 = \xi_L^\lambda$ we say that *class 1 is given priority*, and if in addition $\xi_2 > r^\lambda(\xi_L^\lambda)$ we say that *class 2 is intentionally delayed*. If $\xi_2 = r^\lambda(\xi_U^\lambda)$ we say that *class 2 is given priority*.

This system data can be calculated for some queueing settings of interest. A concrete example of a queueing model and its achievable regions is given in Section 2.4. Furthermore, the structural results obtained by us do not rely on the explicit construction of ξ_L^λ , ξ_H^λ and r^λ ; it only relies on their existence. We believe that it can be rigorously established that most queueing models of interest generate achievable regions of this form, but we choose not to take that path.

The SP offers a menu of subscription options, corresponding to different grades of service, at different prices, so as to maximize profits. We are now ready to state the service provider's problem. We begin with its full information version.

2.2 The Full Information Problem

In this section we look at a relaxed version of the problem, where the customers' classes are observable. The SP then solves the following full information problem (FIP).

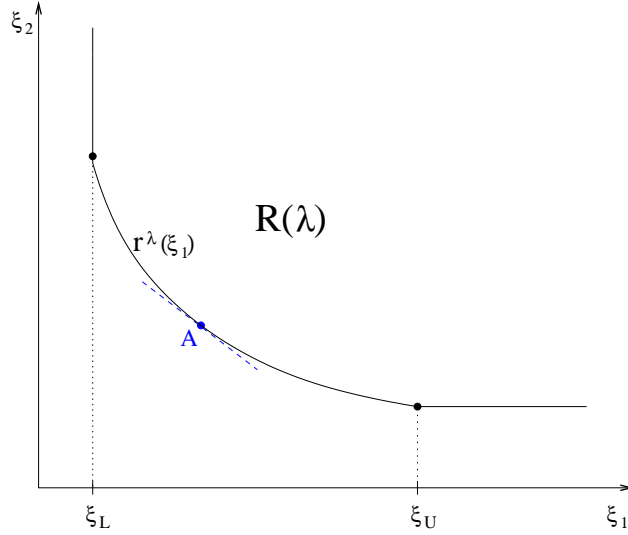


Figure 3: The achievable region

$$\begin{aligned}
 \mathbf{Problem (FIP):} \quad & \max_{\lambda, p, \xi \geq 0} \sum_{i=1}^2 \lambda_i p_i \\
 & \text{subject to } \xi \in R(\lambda) \quad \quad \quad (\text{Capacity}) \\
 & \quad \quad \quad p_i = P_i(\lambda_i) - a_i \xi_i, \quad i = 1, 2 \quad \quad \quad (\text{IR}) \\
 & \quad \quad \quad \lambda \in \Lambda.
 \end{aligned}$$

We solve this problem for fixed arrival rates $\lambda \in \Lambda$. One can later maximize over λ to obtain a closed form solution. Using the IR constraint, we can eliminate prices to obtain the following problem.

$$\begin{aligned}
 \mathbf{Problem (FIP')}: \quad & \max_{\xi_1, \xi_2 \geq 0} \sum_{i=1}^2 \lambda_i (P_i(\lambda_i) - a_i \xi_i) \\
 & \text{subject to } \xi \in R(\lambda) \quad \quad \quad (\text{Capacity}).
 \end{aligned}$$

Now, because λ is fixed, we are maximizing an objective function linear in ξ over the convex set $R(\lambda)$, and the optimal solution is marked by point A in Figure 3, where the dashed line marks the iso-profit line passing through that point. The efficient frontier is, therefore, the set of all possible full information solutions, each optimal for a different $\frac{a_1}{a_2}$ parameter. We now present the private information version of this problem.

2.3 The Private Information Problem

When customers' classes are unobservable, customers are free to selfishly choose between the different service grades offered. A class i customer, therefore, will subscribe to service grade

$$j^* = \arg \min_j \{a_i \xi_j + p_j\},$$

or not subscribe at all. A menu is said to be incentive compatible (IC), when each customer subscribes to the grade of service tailored for his or her class, or does not subscribe at all. The revelation principle guarantees that we lose nothing by restricting attention to IC menus. Therefore, grade i will refer to the grade of service tailored for class i . The SP now solves the following private information problem (PIP).

$$\begin{aligned}
\mathbf{Problem (PIP):} \quad & \max_{\lambda, p, \xi \geq 0} \sum_{i=1}^2 \lambda_i p_i \\
& \text{subject to } \xi \in R(\lambda) && \text{(Capacity)} \\
& p_2 + a_2 \xi_2 \leq p_1 + a_2 \xi_1 && (IC)_1 \\
& p_1 + a_1 \xi_1 \leq p_2 + a_1 \xi_2 && (IC)_2 \\
& p_i = P_i(\lambda_i) - a_i \xi_i, \quad i = 1, 2 && (IR) \\
& \lambda \in \Lambda.
\end{aligned}$$

We solve this problem for fixed arrival rates vector λ to obtain structural insights on the solution. One can later maximize over λ to obtain a closed form solution. Using the IR constraint, we eliminate prices to obtain the following problem.

$$\begin{aligned}
\mathbf{Problem (PIP')}: \quad & \max_{\xi_1, \xi_2 \geq 0} \sum_{i=1}^2 \lambda_i (P_i(\lambda_i) - a_i \xi_i) \\
& \text{subject to } \xi \in R(\lambda) && \text{(Capacity)} \\
& \xi_1 \leq \frac{P_1(\lambda_1) - P_2(\lambda_2)}{a_1 - a_2} && (IC)_1 \\
& \xi_2 \geq \frac{P_1(\lambda_1) - P_2(\lambda_2)}{a_1 - a_2}. && (IC)_2
\end{aligned}$$

The following proposition shows that we lose nothing by restricting our attention to arrival rates λ that satisfy $P_1(\lambda_1) \geq P_2(\lambda_2)$.

Proposition 1 *If (λ, p, ξ) is a triple satisfying the constraints of Problem (PIP), then $P_1(\lambda_1) \geq P_2(\lambda_2)$.*

Proof. It is clear from constraint $(IC)_1$ in Problem (PIP') that if $P_1(\lambda_1) < P_2(\lambda_2)$, then customers of class 2 would be better off pretending to be of class 1. Therefore, if $\lambda_2 > 0$ this means the solution is not IC. However, if $\lambda_2 = 0$, then $\lambda_1 > 0$ (because clearly $\lambda = 0$ is not optimal). Because $P_1(\lambda_1) < P_2(\lambda_2) = \lim_{s \downarrow 0} P_2(s)$ and P_2 is continuous, there is a positive mass of customers in class 2 whose valuation v exceeds $P_1(\lambda_1)$. For them $v - a_2 \xi_1 - p_1 > P_1(\lambda_1) - a_1 \xi_1 - p_1 = 0$. This contradicts IR of class 2. Therefore, any solution that satisfies $P_1(\lambda_1) < P_2(\lambda_2)$ is not incentive compatible. ■

Definition 1 *An arrival rate vector λ is said to be feasible if Problem (PIP') has a feasible solution for that vector.*

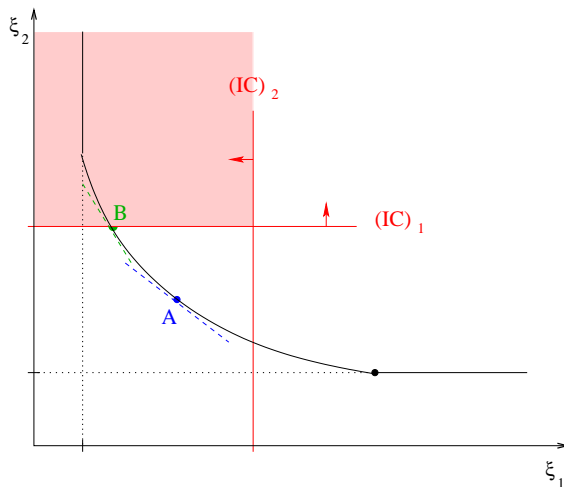


Figure 4: The private information problem

The following is the main result of this paper.

Proposition 2 *Given a fixed, feasible vector $\lambda \in \Lambda$ of arrival rates, the optimal solution to Problem (PIP') either results in Pareto efficient delay measures, or in class 1 being given priority, while class 2 is intentionally delayed.*

Proof. Let ξ denote the optimal solution to Problem (PIP'), and assume it is not on the EF, nor is class 1 being given priority, while class 2 is intentionally delayed. This means that there exists an $\epsilon > 0$, such that $\xi' = (\xi_1 - \epsilon, \xi_2) \in R(\lambda)$. Now, ξ' is a feasible solution to Problem (PIP') by construction, and it improves the objective function, contradicting the optimality of ξ . ■

Corollary 1 *The optimal scheduling rule for the queueing setting either results in Pareto efficient delay measures, in class 1 being given priority, while class 2 is intentionally delayed.*

Proof. Given a fixed, feasible vector λ of arrival rates, the corollary directly follows from Proposition 2. The objective function and constraints are all continuous in λ , and the set of feasible arrival rates is compact, and therefore an optimal vector of arrival rates λ^* exists. Now, since this structure is optimal for each feasible λ , it must be true for the optimal one as well. ■

Therefore, we conclude that the optimal scheduling results either in Pareto efficient delay measures (see Figure 4), or in class 1 being given priority and class 2 being intentionally delayed (see Figure 5). In all figures point *A* marks the solution to the full information version of the problem, and point *B* the solution to the private information version. As shown in Section 2.2, any point on the EF is optimal for some cost coefficients. Note that the distortion to the full information (as well as socially optimal) solution caused by the IC constraints is of a different form than that in the standard screening problem. In particular, we have upward distortion at the top class, meaning an improvement in ξ_1 from the full information benchmark. This is due to the limited processing capacity, as shown in [20].

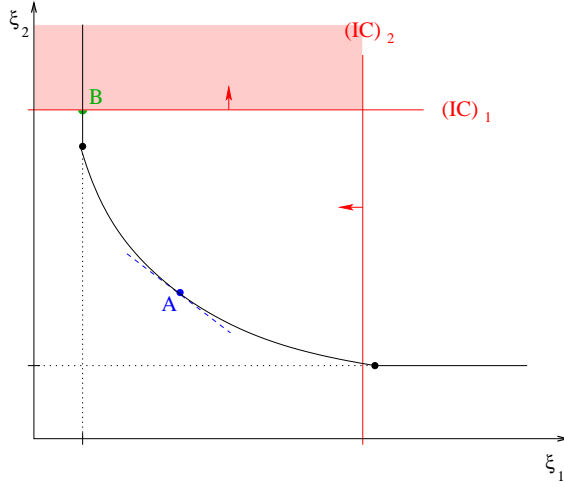


Figure 5: The private information problem: intentional delay

2.4 The Special Case Studied by Afeche (Linear Delay Costs)

We now restrict our attention to achievable regions that are defined by linear curves (Figure 6). Such achievable regions are fully characterized by the curve's endpoints. These polyhedron-shaped achievable regions arise in the queueing setting studied by Afeche in [1], where delay costs are linear. Let us characterize the achievable region for this setting with fixed arrival rates λ . We consider for concreteness a non-preemptive two-class M/M/1 queue. The two endpoints of the EF correspond to giving static priority to each of the classes, i.e. the *c μ rule* at one end, and the *reversed c μ rule* at the other. These priority rules result in the following delay measures (in this case - expected delay).

$$\begin{aligned}\xi_L &= \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{1 - \rho_1} + \frac{1}{\mu_1}; \\ \xi_U &= \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{(1 - \rho_2)(1 - \rho_1 - \rho_2)} + \frac{1}{\mu_1}; \\ r^\lambda(\xi_L) &= \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{(1 - \rho_1)(1 - \rho_1 - \rho_2)} + \frac{1}{\mu_2}; \\ r^\lambda(\xi_U) &= \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{1 - \rho_2} + \frac{1}{\mu_2}.\end{aligned}$$

Now, solutions on the EF can only be achieved by work-conserving scheduling rules, and hence must satisfy

$$\rho_1 \xi_1 + \rho_2 \xi_2 = \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{1 - \rho_1 + \rho_2}.$$

Therefore, we obtain the following linear EF, defined on $[\xi_L, \xi_U]$:

$$r^\lambda(\xi_1) = \frac{\lambda_1/\mu_1^2 + \lambda_2/\mu_2^2}{\rho_2(1 - \rho_1 + \rho_2)} - \frac{\rho_1}{\rho_2} \xi_1.$$

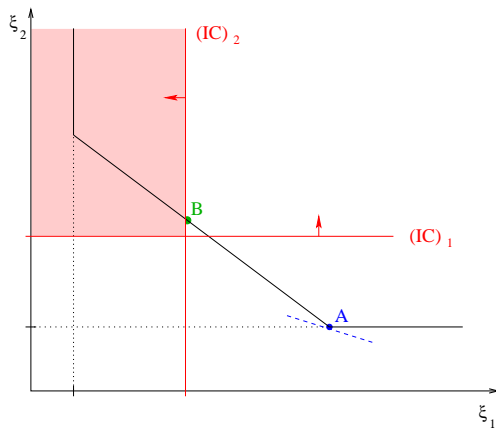


Figure 6: Afeche’s setting for $a_1 > a_2$ and $a_1\mu_1 < a_2\mu_2$

Afeche shows that, depending on the parameter regime, the optimal scheduling rule would either be the $c\mu$ rule itself with potential intentional delay (or strategic idleness, as it is referred to in [1]) used to increase the delay differentiation between the classes, or a randomized priority rule, which with probability $\alpha \in [0, 1]$ uses the $c\mu$ rule and with probability $(1 - \alpha)$ uses the *reversed* $c\mu$ rule. Corollary 1 indeed reinforces all those results, as is illustrated for the most interesting parametric case ($a_1\mu_1 < a_2\mu_2$) in Figure 6. Point A in the figure corresponds to the full information solution (the $c\mu$ rule), and point B to some randomized priority rule. As evident from the figure, if $\frac{P_1(\lambda_1) - P_2(\lambda_2)}{a_1 - a_2} > r^\lambda(\xi_L)$, intentional delay will be used to increase the delay differentiation beyond the differentiation the *reversed* $c\mu$ rule gives. Using intentional delay together with the *reversed* $c\mu$ rule never comes out optimal in [1] because pricing is allowed to increase as a function of service time. Therefore, when $\mu_1 < \mu_2$, class 1 customers will be deterred from pretending to be of class 2 because charge would be high in expectation for their longer service times. In this case, constraint $(IC)_1$ is unnecessary and therefore intentional delay will never be used. Pursuing this idea further, one can come up with more complicated pricing methods, such as non-monotone functions of the realized service time, to obtain an even further relaxation of the IC constraints, resulting in their complete elimination from the problem in extreme cases. Such pricing rules, however, are too cumbersome to implement in many applications, and hence we assume throughout this paper that pricing does not depend on service times.

The case of linear delay costs is simpler in that the EF is linear, and is generated by scheduling rules that randomize between the two static priority rules. Therefore, we not only know the achievable region $R(\lambda)$ in terms of the delay measures ξ , but also specific scheduling rules that achieve each point in it. This is a significant relaxation from queueing settings with convex delay costs, and consequently a convex EF, which is not fully characterized by its endpoints.

2.5 Extension to n Classes

When extending the model to $n \geq 2$ customer classes, we still assume multiplicative separability, meaning that the delay cost function of class i customers is $C_i(\cdot) = a_i C(\cdot)$, and without loss of generality we take $a_1 > \dots > a_n$. The service provider solves the following generalized private information problem (GPIP).

$$\begin{aligned}
\textbf{Problem (GPIP): } & \max_{\lambda, p, \xi \geq 0} \sum_{i=1}^n \lambda_i p_i \\
& \text{subject to } \xi \in R(\lambda) && \text{(Capacity)} \\
& p_i + a_i \xi_i \leq p_j + a_i \xi_j, \quad \forall i \neq j && \text{(IC)} \\
& p_i = P_i(\lambda_i) - a_i \xi_i, \quad i = 1, \dots, n && \text{(IR)} \\
& 0 \leq \lambda_i \leq u_i, \quad i = 1, \dots, n,
\end{aligned}$$

Now fix a feasible vector λ . Figure 7 shows a three-dimensional example, where points a, b, c, d, e, f are the endpoints of the EF, corresponding to the $3!$ different static priority rules. As with two classes, we use the IR constraints to eliminate prices and obtain the following problem:

$$\begin{aligned}
& \max_{\xi \geq 0} \sum_{i=1}^n \lambda_i (P_i(\lambda_i) - a_i \xi_i) \\
& \text{subject to } \xi \in R(\lambda) && \text{(Capacity)} \\
& \xi_i \leq \frac{P_i(\lambda_i) - P_j(\lambda_j)}{a_i - a_j}, \quad \forall i < j && \text{(IC)}_1 \\
& \xi_j \geq \frac{P_i(\lambda_i) - P_j(\lambda_j)}{a_i - a_j}, \quad \forall i < j. && \text{(IC)}_2
\end{aligned}$$

There are now $n(n-1)$ IC constraints, making the problem seem a lot more complicated to solve as n increases. However, because the IC constraints take the form of a bound on one of the delay measures, they dominate each other and can be reduced to only $2(n-1)$ constraints, significantly reducing the complexity of the problem, as follows.

$$\begin{aligned}
\textbf{Problem (GPIP')}: & \max_{\xi \geq 0} \sum_{i=1}^n \lambda_i (P_i(\lambda_i) - a_i \xi_i) \\
& \text{subject to } \xi \in R(\lambda) && \text{(Capacity)} \\
& \xi_1 \leq \min_{j>1} \frac{P_1(\lambda_1) - P_j(\lambda_j)}{a_1 - a_j} && \text{(IC)} \\
& \max_{j<i} \frac{P_j(\lambda_j) - P_i(\lambda_i)}{a_j - a_i} \leq \xi_i \leq \min_{j>i} \frac{P_i(\lambda_i) - P_j(\lambda_j)}{a_i - a_j}, \quad \forall 1 < i < n && \text{(IC)} \\
& \xi_n \geq \max_{j<n} \frac{P_j(\lambda_j) - P_n(\lambda_n)}{a_j - a_n}. && \text{(IC)}
\end{aligned}$$

Thus the IC constraints describe a box, as in Figure 7, and we can give the obvious analogs of Propositions 1 and 2.

Proposition 3 *If (λ, p, ξ) is a triple satisfying the constraints of Problem (GPIP), then $P_i(\lambda_i) \geq P_j(\lambda_j)$ for $i < j$.*

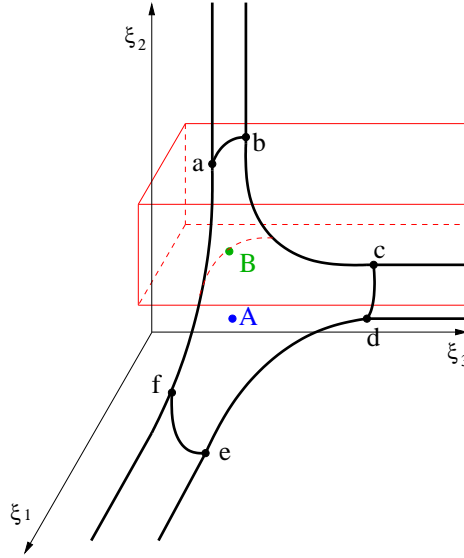


Figure 7: Solution for n customer classes

Definition 2 An arrival rate vector λ is said to be feasible if Problem (GPIP') has a feasible solution for that vector.

For a fixed, feasible arrival rate vector λ , let r^λ denote an n -dimensional EF, and

$$\xi_L = \inf \left\{ \xi_1 \mid \exists \widehat{\xi} = (\widehat{\xi}_1, \dots, \widehat{\xi}_n) \in R(\lambda) \text{ such that } \widehat{\xi}_1 = \xi_1 \right\}$$

denote the minimal achievable value of ξ_1 (achieved by points b, c in Figure 7). As in the two-dimensional case, this minimum is indeed achieved because the EF is a compact set.

Proposition 4 Given a fixed, feasible vector λ of arrival rates, the optimal solution ξ^* to Problem (GPIP') is either Pareto efficient, or it satisfies $\xi_1^* = \xi_L$.

The following is the analog of Corollary 1.

Corollary 2 The optimal scheduling rule for the queueing setting either results in Pareto efficient delay measures, or in class 1 being given priority, while some of classes $2, \dots, n$ are intentionally delayed.

The proofs of the above proposition and corollary are similar to their analogs in Section 2.3, and are therefore skipped.

3 An Idealized Delay Model with Time Scale Separation

As discussed in the previous section, the efficient frontier (EF) of an achievable region (the set of Pareto efficient solutions to the underlying queueing system) corresponds to the set of solutions to the full information version of the problem (FIP) for different cost coefficients a_1 and a_2 . However,

solving (FIP) proves extremely difficult in most cases, and we are not able to fully characterize the EF for these queueing settings, except in special cases. One such special case (linear delay costs) was discussed in Section 2.4. Another special setting is the idealized delay model with time scale separation that we study in this section.

3.1 Workload Distributions and Achievable Regions

In this setting we take as primitive a “workload distribution” $F(z|\lambda)$ that we will use to construct the efficient frontier and the achievable region $R(\lambda)$. For each feasible λ vector this is a distribution on $\mathbb{R}_+ = [0, \infty)$. We interpret z as the total hours of server work, or expected total hours of server work, in the system at a point in time, and we treat this as an adequate summary of “the state of the system” (see below). Then $F(\cdot|\lambda)$ gives a distribution of system states resulting from the capacity available and the commitments made, independently of the dynamic control policy adopted.

A control policy (or scheduling rule, or dynamic resource allocation policy) takes the form of a measurable function $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}_+^2$ with the following property:

$$\phi_1(z) + \phi_2(z) = z, \tag{1}$$

for all $z \geq 0$. One interprets $\phi_i(z)$ as the portion of the total backlog z that is held in buffer i . Given a control policy ϕ , we associate with class i a waiting time (or delay) of

$$W_i^\phi(\lambda, z) = \frac{\mu_i}{\lambda_i} \phi_i(z), \tag{2}$$

when the server workload is z , and hence associate with class i the overall delay measurement

$$\xi_i^\phi(\lambda) = \int_0^\infty C(W_i^\phi(\lambda, z)) F(dz|\lambda). \tag{3}$$

The control policies ϕ defined in (1) are work-conserving, and therefore the set $R(\lambda)$ contains all pairs $\xi^\phi(\lambda) = (\xi_1^\phi(\lambda), \xi_2^\phi(\lambda))$ that are generated from admissible scheduling rules ϕ via (2) and (3), as well as all pairs dominated by them. That is, any pair $\xi = (\xi_1, \xi_2) \in R(\lambda)$ can be achieved by using some work conserving control policy ϕ , and then intentionally delaying each job of class i further by $\delta_i \geq 0$, so that $\xi_i = \xi_i^\phi(\lambda) + \delta_i$. By construction, the achievable regions defined in this section satisfy all our assumptions on achievable regions from Section 2.1.

The key relationships (1) and (2) may be described as follows:

- (a) an instantaneous backlog redistribution idealization - meaning that the SP can hold the backlog content of work in whatever job classes she may desire, subject only to an adding-up constraint.
- (b) a deterministic idealization of delay - meaning that, given a distribution ϕ of workload among buffers, we reckon job delays by means of a fluid equilibrium logic (workload flows into buffer i at a constant rate of $\rho_i = \lambda_i/\mu_i$, and the workload content of the buffer is $\phi_i(z)$, so the residence time in the buffer with FIFO processing must be $\mu_i\phi_i(z)/\lambda_i$).

Model assumption (a) in turn has two elements. First, the SP has *complete* discretion as to how backlog is distributed between the two classes; in particular, he or she can cause *all* the backlog to

be held in one buffer if he or she wishes, meaning that delays will become negligible for the other class. Second, a redistribution of the backlog can be effective *instantaneously* once it is decided upon. This is best viewed as an assumption of time and scale separation: the system backlog z changes *slowly* relative to the time required for the SP to redistribute work.

3.2 Additional Special Structure of the Brownian Model

The “Brownian Model” of system congestion and dynamic resource allocation further assumes the following specific form of the workload distribution $F(\cdot|\lambda)$:

$$F(z|\lambda) = 1 - \exp\{-\theta(\lambda)z\}, \quad (4)$$

where $\theta(\lambda) > 0$ for each $\lambda \in \Lambda$, and the specific functional form of $\theta(\cdot)$ is irrelevant for our purposes in this paper.

The specific exponential distribution (4) is derived from the delay model that represents workload as a one-dimension reflected Brownian motion (RBM) with state space $[0, \infty)$, drift parameter $-(1 - \rho_1 - \rho_2)$ and variance parameter $\sigma^2(\lambda) > 0$ (a variability parameter that combines the effects of demand variability and service variability), *assuming it is the long-run (or steady-state) workload distribution that is of interest*. Some of the calculations done in the following section use this exponential distribution, but a different distribution $F(\cdot|\lambda)$ could be extracted from the RBM model of backlog fluctuation, if a finite time horizon were deemed relevant, and in principle the calculations would proceed in the same way.

3.3 The Case of Quadratic Delay Costs

For concreteness we take the nominal delay cost function to be $C(w) = w^2$, and thus for a given vector λ of arrival rates, the delay measure $\xi_i = EC(W_i(\lambda, Z))$ is the second moment of delay for class i , where expectation is with respect to a random variable Z that is distributed according to the workload steady state distribution $F(\cdot|\lambda)$.

The additional structure we have assumed on the underlying queueing model allows us to compute the EF. To do that we fix $\xi_1 \in [0, \xi_U^\lambda]$ ($\xi_L^\lambda = 0$ in the Brownian model) and look for the minimal achievable value of ξ_2 , i.e. we wish to solve the following problem:

$$\begin{aligned} \min_{\phi_1, \phi_2 \in L^2(P)} \quad & \xi_2 = \frac{E\phi_2^2(Z)}{\rho_2^2} \\ \text{subject to} \quad & \phi_1(z) + \phi_2(z) = z, \quad \forall z \geq 0 \\ & E\phi_1^2(Z) = \rho_1^2 \xi_1 \\ & \phi_1, \phi_2 \geq 0, \end{aligned} \quad (5)$$

where the first constraint is imposed because a Pareto efficient solution can only be generated by a work-conserving scheduling rule. For fixed λ , the steady state distribution of the total workload in the Brownian model is exponential with rate parameter θ .

Proposition 5 *Given a fixed, feasible vector λ of arrival rates, the efficient frontier is $r^\lambda(\xi_1) = \frac{1}{\rho_2^2} \frac{2}{\theta^2} \left(1 - \rho_1 \sqrt{\frac{\xi_1 \theta^2}{2}}\right)^2$, $0 \leq \xi_1 \leq \frac{2}{\theta^2 \rho_1^2}$.*

Proof. Solving (5) is equivalent to solving the following problem:

$$\begin{aligned}
& \min_{\phi_1 \in L^2(P)} \frac{E(Z - \phi_1(Z))^2}{(\rho - \rho_1)^2} \\
& \text{subject to } E\phi_1^2(Z) = \rho_1^2 \xi_1; \\
& 0 \leq \phi_1(z) \leq z, \quad \forall z \geq 0.
\end{aligned} \tag{6}$$

Defining $B = \{u \mid u \in L^2(P), \|u(Z)\|_2^2 \leq \rho_1^2 \xi_1\}$, we wish to minimize $\|Z - \phi_1(Z)\|_2$ for $\phi_1 \in B$. Since B is closed and convex in the Hilbert space $L^2(P)$, we know that ϕ_1^* is the unique solution if and only if it satisfies $\langle Z - \phi_1^*(Z), u(Z) - \phi_1^*(Z) \rangle \leq 0, \forall u \in B$ (see Theorem 3.12.1 in [8]). Now, let us show that $\phi_1^*(Z) = \alpha Z$, where $\alpha = \rho_1 \sqrt{\frac{\xi_1 \theta^2}{2}}$. Assume that is not the case, then there exists $u \in B$ such that

$$0 < \langle Z - \phi_1^*(Z), u(Z) - \phi_1^*(Z) \rangle = E[(Z - \alpha Z)(u(Z) - \alpha Z)] = (1 - \alpha)E[Z(u(Z) - \alpha Z)].$$

Since $\alpha < 1$, we obtain

$$\begin{aligned}
0 < E [2\alpha Z(u(Z) - \alpha Z) + (u(Z) - \alpha Z)^2 + (\alpha Z)^2 - (u(Z) - \alpha Z)^2 - (\alpha Z)^2] = \\
= Eu^2(Z) - E [(u(Z) - \alpha Z)^2 + \phi_1^{*2}(Z)],
\end{aligned}$$

or

$$\|u(Z)\|_2^2 > E(u(Z) - \alpha Z)^2 + E\phi_1^{*2}(Z) \geq E\phi_1^{*2}(Z) = \rho_1^2 \xi_1,$$

which is a contradiction to $u \in B$. Therefore, $\phi_2^*(Z) = (1 - \alpha)Z$ and $r^\lambda(\xi_1) = \frac{E\phi_2^{*2}(Z)}{\rho_2^2} = \frac{1}{\rho_2^2} \frac{2}{\theta^2} \left(1 - \rho_1 \sqrt{\frac{\xi_1 \theta^2}{2}}\right)^2$. ■

We can now express the SP's problem as follows.

$$\begin{aligned}
& \max_{\xi_1, \xi_2 \geq 0} \sum_{i=1}^2 \lambda_i (P_i(\lambda_i) - a_i \xi_i) \\
& \text{subject to } \xi_2 \geq r^\lambda(\xi_1) \tag{Capacity} \\
& \xi_i \geq 0, \quad i = 1, 2 \\
& \xi_1 \leq \frac{P_1(\lambda_1) - P_2(\lambda_2)}{a_1 - a_2} \tag{IC}_1 \\
& \xi_2 \geq \frac{P_1(\lambda_1) - P_2(\lambda_2)}{a_1 - a_2}. \tag{IC}_2
\end{aligned} \tag{7}$$

Proposition 2 in the previous section gave a general characterization of the optimal solution, but could not prescribe specific scheduling rules in the absence of a clear description of the EF and scheduling rules that generate it. The additional structure of the Brownian model allows us to look beyond the desired ξ_1, ξ_2 in search of the actual control policy (scheduling rule) needed to implement them. Under full information, it was shown in [16] that using the *generalized $c\mu$*

rule (*Gcμ rule*) for scheduling maximizes the social welfare. The *Gcμ rule* is the control policy ϕ^G that satisfies $\mu_1 C'_1(W_1^G(\lambda, z)) = \mu_2 C'_2(W_2^G(\lambda, z))$ for all $z \geq 0$, or for quadratic delay costs, $\mu_1 \phi_1^G(z)/\rho_1 = \mu_2 \phi_2^G(z)/\rho_2$. It is quite easy to see that the *Gcμ rule*, being delay cost minimizing, is also profit maximizing under full information. To see this, fix arrival rates λ_1, λ_2 to be the profit maximizing rates. Thus, the total costs to a subscriber (price and average delay cost) that would induce these rates are fixed as well. The service provider extracts only the price part of the total cost, and therefore would prefer to implement the delay costs minimizing rule, i.e. the *Gcμ rule*. The following proposition gives the values of ξ_1, ξ_2 under the *Gcμ rule*.

Proposition 6 *The second moment of delay in queue i under the *Gcμ rule* is $\xi_i^G = \frac{a_{-i}^2 \mu_{-i}^2}{(a_1 \mu_1 \rho_2 + a_2 \mu_2 \rho_1)^2} \frac{2}{\theta^2}$.*

Proof. The *Gcμ rule* satisfies for all $z \geq 0$,

$$a_1 \mu_1 \rho_2 \phi_1^G(z) = a_2 \mu_2 \rho_1 \phi_2^G(z) = a_2 \mu_2 \rho_1 (z - \phi_1^G(z)).$$

Therefore, $\phi_1(z) = \frac{a_2 \mu_2 \rho_1}{a_1 \mu_1 \rho_2 + a_2 \mu_2 \rho_1} z$, and

$$\xi_1^G = E \left(\frac{\phi_1(Z)}{\rho_1} \right)^2 = \frac{a_2^2 \mu_2^2}{(a_1 \mu_1 \rho_2 + a_2 \mu_2 \rho_1)^2} \frac{2}{\theta^2}.$$

Now, $\xi_2^G = E(Z - \phi_1(Z))$, concluding the proof. ■

The *Gcμ rule* results in delay measures on the EF for any (a_1, a_2) pair. Moreover, every point on the EF can be achieved by a *Gcμ rule* with the right delay cost coefficients. To see this, let us examine the following family of dynamic scheduling rules that allocate workload so that $C'(\frac{\phi_1}{\rho_1}) = \gamma C'(\frac{\phi_2}{\rho_2})$ for some parameter $\gamma \in [0, \infty]$. These workload allocation rules can be interpreted as the asymptotic limits of the scheduling rules that give priority to class 1 whenever $C'(\tau_1) > \gamma C'(\tau_2)$, where τ_i is the age of the oldest class i job. This family of policies generates the entire EF with $\gamma = 0$ corresponding to static priority to class 1, and $\gamma = \infty$ corresponding to static priority to class 2. Moreover, $\gamma = \frac{\mu_2 a_2}{\mu_1 a_1}$ corresponds to the *Gcμ rule*, and therefore each point on the EF can be achieved by a *Gcμ rule*, employed with a coefficient pair (a_1, a_2) , that satisfy $\gamma = \frac{\mu_2 a_2}{\mu_1 a_1}$ for the relevant γ . We call these control policies *Gcμ rules*, to indicate that they are *Gcμ rules* for virtual delay cost functions. The following proposition now follows from Proposition 2.

Proposition 7 *The profit maximizing control policy is either a $\widehat{Gc\mu}$ rule, which is the *Gcμ rule* with virtual delay cost coefficients, or giving priority to class 1 while intentionally delaying class 2.*

The following is an example where the optimal solution involves giving priority to class 1 while intentionally delaying class 2.

Example 1 $C(w) = w^2$, $a_1 = 2$, $a_2 = 1$, $P_1(\lambda_1) = 138000 - 132000\lambda_1$, $P_2(\lambda_2) = 56000 - 56000\lambda_2$, $\mu_1 = \mu_2 = 1$, $\theta(\lambda) = \frac{1}{\lambda_1 + \lambda_2} - 1$. Using numerical analysis we find that the optimal subscription rates are $\lambda_1 = 0.659$, $\lambda_2 = 0.272$, giving traffic intensity $\rho = 0.931$, and the optimal scheduling rule results in $\xi_1 = 0$ (giving priority to class 1) and $\xi_2 = 10278.5$ (of which 4909 is due to waiting in queue, and the remaining 5369.5 is due to intentional delay).

For the cases when the $G\widehat{c}\mu$ rule is optimal, we know that the capacity constraint in (7) is binding, and we can express the problem in terms of ξ_1 alone as follows.

$$\begin{aligned}
& \max_{\xi_1 \geq 0} \lambda_1 [P_1(\lambda_1) - a_1 \xi_1] + \lambda_2 [P_2(\lambda_2) - a_2 r^\lambda(\xi_1)] \\
& \text{subject to } 0 \leq \xi_1 \leq \frac{2}{\rho_1^2 \theta^2} \\
& \xi_1 \leq \frac{P_1(\lambda_1) - P_2(\lambda_2)}{a_1 - a_2} \quad (IC)_1 \\
& \xi_1 \leq \frac{2}{\rho_1^2 \theta^2} \left[1 - \rho_2 \theta \sqrt{\frac{P_1(\lambda_1) - P_2(\lambda_2)}{2(a_1 - a_2)}} \right]^2. \quad (IC)_2
\end{aligned} \tag{8}$$

The $Gc\mu$ rule depends only on the ratio $\frac{a_1}{a_2}$, and therefore it suffices to modify only coefficient a_1 . Specifically, the virtual coefficient is $\widehat{a}_1 = a_1 + \frac{\beta_1 + \beta_2}{\lambda_1}$, where β_1, β_2 are the Lagrange multipliers associated with the IC constraints in (8). We can solve for the Lagrange multipliers to obtain

$$\begin{aligned}
\beta_1 &= \left[\frac{a_2 \rho_1 \lambda_2}{\rho_2^2 \theta (a_1 - a_2)} \left(\sqrt{\frac{2(a_1 - a_2)}{P_1(\lambda_1) - P_2(\lambda_2)}} - \rho_1 \theta \right) - \frac{a_1 \lambda_1}{a_1 - a_2} \right]^+ \mathbb{1}_{\left\{ \theta \leq \sqrt{\frac{2(a_1 - a_2)}{P_1(\lambda_1) - P_2(\lambda_2)}} \right\}} \\
\beta_2 &= \left[\frac{\rho_1 \theta}{1 - \rho_2 \theta \sqrt{\frac{P_1(\lambda_1) - P_2(\lambda_2)}{2(a_1 - a_2)}}} - a_2 \lambda_2 \frac{\rho_1^2}{\rho_2^2} - a_1 \lambda_1 \right]^+ \mathbb{1}_{\left\{ \theta \geq \sqrt{\frac{2(a_1 - a_2)}{P_1(\lambda_1) - P_2(\lambda_2)}} \right\}}
\end{aligned} \tag{9}$$

where $[x]^+ = \max\{0, x\}$.

We have concluded that the optimal scheduling rule is either a $G\widehat{c}\mu$ rule with virtual delay cost coefficient \widehat{a}_1 as described above, or full priority to class 1 while intentionally delaying class 2. This idea of virtual types was first introduced by Myerson in [13], and is explored in the presence of capacity constraints in [20]. Having reduced the problem to a full information one with virtual classes, we can prescribe an exact scheduling rule for problems whose full information version is solvable. Unfortunately, even with full information, these problems are very difficult, and remain unsolved except in the special models studied in this paper.

For some arrival rates λ , it could be the case that the full information solution is already incentive compatible, which results in virtual coefficients identical to the original ones. For example, when jobs' inter-arrival times and service durations have similar distributions across classes, [17] showed that for the social welfare maximizing arrival rates, using the $Gc\mu$ rule is incentive compatible, in the sense that subscribers only subscribe to their designated service grade, and hence the private information problem has the same solution as the full information one. The $Gc\mu$ rule, however, is not incentive compatible for all arrival rates, and in particular not necessarily for the profit maximizing rates. The following corollary characterizes the arrival rates for which the $Gc\mu$ rule is incentive compatible.

Corollary 3 *Using the $Gc\mu$ rule is incentive compatible only for arrival rates λ_1, λ_2 that satisfy*

$$\xi_1^G = \frac{a_2^2 \mu_2^2}{(a_1 \mu_1 \rho_2 + a_2 \mu_2 \rho_1)^2 \theta^2} \frac{2}{\theta^2} \leq \frac{P_1(\lambda_1) - P_2(\lambda_2)}{(a_1 - a_2)} \leq \frac{a_1^2 \mu_1^2}{(a_1 \mu_1 \rho_2 + a_2 \mu_2 \rho_1)^2 \theta^2} \frac{2}{\theta^2} = \xi_2^G. \tag{10}$$

Proof. The proof is a direct result of Proposition 6 and the incentive compatibility constraints in (7). ■

This provides us with an alternative proof for Van-Mieghem's result for quadratic delay costs, as exhibited in the following proposition.

Proposition 8 *If $F(\cdot|\lambda)$ depends on λ only through $\lambda_1 + \lambda_2$, then using the $Gc\mu$ rule is socially optimal for the Brownian model even with private information.*

Proof. Under the $Gc\mu$ rule, the social welfare is

$$\pi_S(\lambda_1, \lambda_2) = \int_0^{\lambda_1} P_1(s)ds + \int_0^{\lambda_2} P_2(s)ds - \lambda_1 a_1 \frac{a_2^2 \mu^2}{(a_1 \lambda_2 + a_2 \lambda_1)^2} \frac{2}{\theta^2} - \lambda_2 a_2 \frac{a_1^2 \mu^2}{(a_1 \lambda_2 + a_2 \lambda_1)^2} \frac{2}{\theta^2}.$$

Let $\tilde{\pi}_S(\epsilon) = \pi_S(\lambda_1^* + \epsilon, \lambda_2^* - \epsilon)$, where λ_1^*, λ_2^* are the socially optimal arrival rates and $\theta = \theta(\lambda^*)$ their corresponding workload rate parameter. Therefore, $\tilde{\pi}_S$ must satisfy $\tilde{\pi}'_S(0) = 0$, or equivalently,

$$P_1(\lambda_1^*) - P_2(\lambda_2^*) - \frac{a_1 a_2^2 \mu^2 \frac{2}{\theta^2}}{(a_1 \lambda_2^* + a_2 \lambda_1^*)^2} + \frac{2 \lambda_1 a_1 a_2^2 \mu^2 \frac{2}{\theta^2} (a_2 - a_1)}{(a_1 \lambda_2^* + a_2 \lambda_1^*)^3} + \frac{a_2 a_1^2 \mu^2 \frac{2}{\theta^2}}{(a_1 \lambda_2^* + a_2 \lambda_1^*)^2} + \frac{2 \lambda_2 a_2 a_1^2 \mu^2 \frac{2}{\theta^2} (a_2 - a_1)}{(a_1 \lambda_2^* + a_2 \lambda_1^*)^3} = 0.$$

Therefore,

$$\frac{P_1(\lambda_1^*) - P_2(\lambda_2^*)}{a_1 - a_2} = \frac{a_1 a_2 \mu^2 \frac{2}{\theta^2}}{(a_1 \lambda_2^* + a_2 \lambda_1^*)^2},$$

and using the condition in Corollary 3, completes the proof. ■

According to Proposition 2 any Pareto inefficiency of the optimal solution is due to intentionally delaying class 2. However, it is important not to confuse Pareto inefficiency with intentional delay or server idling. The following proposition shows that some Pareto inefficient solutions can be achieved by work-conserving control rules (non-idling scheduling rules with no intentional delay), simply by using bad scheduling. In fact, the proposition studies how far from the EF work-conserving rules can reach. Let $\bar{\xi}_2(\xi_1)$ denote the highest value ξ_2 can achieve among work-conserving rules, while keeping ξ_1 fixed. i.e. it is the solution to the following problem:

$$\begin{aligned} \max_{\phi_1, \phi_2 \in L^2(P)} \quad & \xi_2 = \frac{E\phi_2^2(Z)}{\rho_2^2} \\ \text{subject to} \quad & \phi_1(z) + \phi_2(z) = z, \quad \forall z \geq 0; \\ & E\phi_1^2(Z) = \rho_1^2 \xi_1; \\ & \phi_1, \phi_2 \geq 0. \end{aligned} \tag{11}$$

Proposition 9 *The highest value ξ_2 can achieve, using a work-conserving rule, while keeping ξ_1 fixed, is $\bar{\xi}_2(\xi_1) = \frac{2}{\rho_2^2 \theta^2} - \frac{\rho_1^2}{\rho_2^2} \xi_1$, for $\xi_1 \in \left[0, \frac{2}{\rho_1^2 \theta^2}\right]$.*

Proof. Problem (11) is equivalent to

$$\begin{aligned} \max_{\phi_1 \in L^2(P)} \quad & \frac{E(Z - \phi_1(Z))^2}{(\rho - \rho_1)^2} = \frac{\frac{2}{\theta^2} + \rho_1^2 \xi_1 - 2E(Z\phi_1(Z))}{(\rho - \rho_1)^2} \\ \text{subject to} \quad & E\phi_1^2(Z) = \rho_1^2 \xi_1; \\ & 0 \leq \phi_1(z) \leq z, \quad \forall z \geq 0, \end{aligned} \tag{12}$$

or,

$$\begin{aligned}
& \min_{\phi_1 \in L^2(P)} E(Z\phi_1(Z)) - \rho_1^2 \xi_1 = E[\phi_1(Z)(Z - \phi_1(Z))] \\
& \text{subject to } E\phi_1^2(Z) = \rho_1^2 \xi_1; \\
& 0 \leq \phi_1(z) \leq z, \quad \forall z \geq 0.
\end{aligned} \tag{13}$$

We now show that $\phi_1^A(z) = z1_A(z)$, $\phi_2^A(z) = z1_{A^c}(z)$ is an optimal solution to (13) for any $A \subseteq [0, \infty)$ that satisfies $EZ^21_A(Z) = \xi_1$. To see that, notice that this solution gives zero as the optimal value in problem (13), and because $\phi_1(z), (z - \phi_1(z)) \geq 0$, this cannot be improved. Therefore,

$$\bar{\xi}_2(\xi_1) = \frac{E(Z^21_{A^c}(Z))}{\rho_2^2} = \frac{EZ^2 - EZ^21_A(Z)}{\rho_2^2} = \frac{\frac{2}{\theta^2} - \rho_1^2 \xi_1}{\rho_2^2}.$$

■

Therefore, the linear line connecting the efficient frontier's two endpoints gives the worst performance achievable by a work-conserving policy. The following proposition shows that this is true beyond the quadratic cost structure, and extends the result to $C(W) = W^n$ for any $n > 1$.

Proposition 10 *If $C(w) = w^n$ for some $n > 1$, then $\bar{\xi}_2(\xi_1) = \frac{EZ^n}{\rho_2^n} - \frac{\rho_1^n}{\rho_2^n} \xi_1$.*

Proof. Let ϕ_1 denote the optimal solution to the following problem:

$$\begin{aligned}
& \max_{\phi_1 \in L^n(P)} \frac{E(Z - \phi_1(Z))^n}{\rho_2^n} \\
& \text{subject to } \frac{E\phi_1^n(Z)}{\rho_1^n}; \\
& 0 \leq \phi_1(z) \leq z, \quad \forall z \geq 0.
\end{aligned} \tag{14}$$

Assume there exist a set B such that $0 < E\phi_1^n(Z)1_B(Z) < EZ^n(Z)1_B(Z)$. Take B_1, B_2 , such that $B_1 \cap B_2 = \emptyset$, $B_1 \cup B_2 = B$, and $EZ^n(Z)1_{B_1}(Z) = E\phi_1^n(Z)1_B(Z)$. We define $\hat{\phi}_1$ as follows.

$$\hat{\phi}_1(z) = \begin{cases} z, & z \in B_1 \\ 0, & z \in B_2 \\ \phi_1(z), & \text{else} \end{cases}$$

and $\hat{\phi}_2(z) = z - \hat{\phi}_1(z)$. Now, $E\hat{\phi}_2^n(Z)1_B(Z) = EZ^n(Z)1_{B_2}(Z) = EZ^n(Z)1_B(Z) - E\phi_1^n(Z)1_B(Z) > E(Z - \phi_1(Z))^n 1_B(Z) = E\phi_2^n(Z)1_B(Z)$ by convexity of $C(w) = w^n$ for $n > 1$, and because $\hat{\phi}_2 = \phi_2$ outside the set B , we conclude that $\frac{E\hat{\phi}_2^n(Z)}{\rho_2^n} > \frac{E\phi_2^n(Z)}{\rho_2^n}$. Since we constructed $\hat{\phi}_1$ to satisfy $\frac{E\hat{\phi}_1(Z)^n}{\rho_1^n} = \frac{E\phi_1^n(Z)}{\rho_1^n} = \xi_1$, we obtain a contradiction to the optimality of ϕ_1 . Therefore, no such set B exists, and we conclude that the optimal ϕ_1 is of the form $\phi_1(z) = z1_A(z)$ for some set A , such that $\frac{EZ^n1_A(Z)}{\rho_1^n} = \xi_1$. Therefore, $\bar{\xi}_2(\xi_1) = \frac{E\phi_2^n(Z)}{\rho_2^n} = \frac{EZ^n - E\phi_1^n(Z)}{\rho_2^n} = \frac{EZ^n}{\rho_2^n} - \frac{\rho_1^n}{\rho_2^n} \xi_1$. ■

3.4 An Example Lacking Multiplicative Separability

In the previous sections we have assumed multiplicative separability, meaning that classes differ only by the coefficients a_i , amplifying the nominal cost function. The optimal solution with this cost structure includes intentional delay only when class 1 receives the best possible service (static priority), and what is left for class 2 is still too attractive for class 1. In such cases, class 2 service must be degraded even further to deter deviation, and that is done by intentionally delaying class 2. We call the set of such performance measures ξ , where $\xi_1 \in [\xi_L^\lambda, \xi_U^\lambda]$ and $\xi_2 > r^\lambda(\xi_1)$, the *extension of the efficient frontier*. When multiplicative separability holds Proposition 2 says that the optimal solution is on the *extended efficient frontier* (the EF + its extension). However, when we do not have multiplicative separability, the solution can move away from the EF in other directions. Specifically, it could happen that the service of both classes is degraded compared with the full information solution. Illustrating this is difficult due to the fact that without multiplicative separability, the achievable region becomes a four-dimensional set even for the two customer classes case. This is because customers of different classes now have different delay measures for the same grade of service. Therefore, the achievable region, as well as the IC constraints consist of $\xi_{i,j} = EC_i(W_j)$ for all i, j . We illustrate this in this section through two examples; one in the Brownian model, and the other assuming a different cost structure, where delay costs are functions of the long-run average delay, and not the actual experienced delay.

We begin with the Brownian model with two customer classes, where $C_1(y) = a_1y^2 + a_2y$, $C_2(y) = a_3y$ are the delay costs with $a_1, a_2, a_3 \geq 0$. Since we are using an achievable region approach, we begin by characterizing the EF for this setting, in order to find the actual achievable region. The following Lemma will prove useful in doing so.

Lemma 1 *Let $X \in L^2(P)$ be a non-negative random variable with cumulative distribution function F . Then $X_1 = X \wedge \tau$ solves the following problem for fixed τ such that $E(X \wedge \tau) = \eta_1$.*

$$\begin{aligned} \min_{0 \leq X_1 \leq X} \quad & EX_1^2 \\ \text{subject to} \quad & EX_1 = \eta_1. \end{aligned} \tag{15}$$

Proof. Assume the optimal X_1 is of a different form, then there are two possibilities.

Case 1: $\int_0^\tau (x - X_1(x))dF(x) = \alpha > 0$.

Define $A_\alpha = \{x \mid X_1(x) > \tau\}$ to be such that $\int_{A_\alpha} (X_1(x) - \tau)dF(x) = \alpha$, and $\widehat{X}_1(x) = \begin{cases} x, & x \leq \tau \\ \tau, & x \in A_\alpha \\ X_1(x), & \text{else} \end{cases}$

Then,

$$\begin{aligned} EX_1^2 - E\widehat{X}_1^2 &= \int_0^\tau (X_1^2(x) - x^2)dF(x) + \int_{A_\alpha} (X_1^2(x) - \tau^2)dF(x) > \\ &> \int_0^\tau (X_1(x) - x)(X_1(x) + x)dF(x) + 2\tau \int_{A_\alpha} (X_1(x) - \tau)dF(x) = \\ &= \int_0^\tau (x - X_1(x))(2\tau - x - X_1(x))dF(x) \geq 0. \end{aligned} \tag{16}$$

Case 2: $\int_0^\tau (x - X_1(x))dF(x) = 0$.

Define $B = \{x \mid X_1(x) > \tau\}$, $C = \{x > \tau \mid X_1(x) < \tau\}$, and $\widehat{X}_1 = X \wedge \tau$.

Then,

$$\begin{aligned}
EX_1^2 - E\widehat{X}_1^2 &= \int_B (X_1^2(x) - \tau^2)dF(x) + \int_C (X_1^2(x) - \tau^2)dF(x) > \\
&> \int_C (X_1(x) - \tau)(X_1(x) + \tau)dF(x) + 2\tau \int_B (X_1(x) - \tau)dF(x) = \quad (17) \\
&= \int_B (X_1(x) - \tau)(X_1(x) + \tau - 2\tau)dF(x) \geq 0,
\end{aligned}$$

thus concluding the proof. ■

The following proposition characterizes the EF for this problem.

Proposition 11 *Each point on the efficient frontier is achieved by using the workload allocation rule $\phi_1(z) = z \wedge \tau$, $\phi_2(z) = z - \phi_1(z)$ for some $\tau > 0$.*

Proof. Let $\eta_i = E\phi_i(Z)$ denote the average workload held in queue i , then by work conservation $\eta_1 = \frac{1}{\theta} - \eta_2$, where $Z \sim \text{exp}(\theta)$. Therefore, a fixed $\xi_{2,2} = \frac{\eta_2}{\rho_2}$, also fixes η_1 and thus minimizing $\xi_{1,1}$ is equivalent to minimizing the second moment $E\phi_1^2(Z)$ subject to a fixed first moment η_1 . Now, Lemma 1 suggests $\phi_1(Z) = Z \wedge \tau$ is optimal with $\tau = -\frac{1}{\theta} \log(1 - \theta\eta_1) > 0$, thus completing the proof. ■

Now, using straightforward calculations, one obtains the following result.

Proposition 12 *Fix $\phi_1(z) = z \wedge \tau$ for $\tau > 0$ and $\phi_2(z) = z - \phi_1(z)$, then $E\phi_1(Z) = \frac{1}{\theta}(1 - e^{-\theta\tau})$, $E\phi_1^2(Z) = \frac{2}{\theta^2}(1 - e^{-\theta\tau}) - \frac{2\tau}{\theta}e^{-\theta\tau}$, $E\phi_2(Z) = \frac{1}{\theta}e^{-\theta\tau}$, and $E\phi_2^2(Z) = \frac{2}{\theta^2}e^{-\theta\tau}$.*

When $P_1(\lambda_1) - P_2(\lambda_2) > 0$, IC can be expressed as follows.

$$EC_1(W_1) - EC_2(W_1) \leq P_1(\lambda_1) - P_2(\lambda_2) \leq EC_1(W_2) - EC_2(W_2). \quad (18)$$

Since each point on the EF corresponds to a specific $\tau > 0$, the only points that are IC are those that satisfy

$$\frac{a_1}{\rho_1^2} \left[\frac{2}{\theta^2}(1 - e^{-\theta\tau}) - \frac{2\tau}{\theta}e^{-\theta\tau} \right] + \frac{(a_2 - a_3)}{\rho_1\theta}(1 - e^{-\theta\tau}) \leq P_1(\lambda_1) - P_2(\lambda_2) \leq \frac{a_1}{\rho_2^2} \frac{2}{\theta^2}e^{-\theta\tau} + \frac{(a_2 - a_3)}{\rho_2\theta}e^{-\theta\tau}. \quad (19)$$

It could happen that the entire EF is not IC. A sufficient condition for that is

$$\log \left(1 - \frac{\theta\rho_1(a_3 - a_2)}{2a_1} \right) > \left[\frac{2a_1}{\rho_2^2} - \theta(a_3 - a_2) \left(\frac{1}{\rho_1} + \frac{1}{\rho_2} \right) \right] \frac{\rho_1^2}{2a_1}.$$

However, if we go far enough along the extended efficient frontier for subscription rates that satisfy $P_1(\lambda_1) - P_2(\lambda_2) > 0$, we will eventually become IC. To see this, we define $I(W) = EC_1(W) - EC_2(W)$ for W , the random delay. Now, let us consider adding a constant intentional delay of $\delta > 0$. Then, $\frac{d}{d\delta}I(W + \delta) = 2a_1EW + (a_2 - a_3) < 0$ if and only if $EW < \frac{a_3 - a_2}{2a_1}$. Therefore, $I(W)$ is eventually increasing without bound in the intentional delay, and hence by taking $W_1 = 0$ and increasing W_2 , Equation 18 is eventually satisfied whenever $P_1(\lambda_1) - P_2(\lambda_2) > 0$. However,

there might be better ways to reach an IC point. One such way is to add intentional delay to class 1. To see this, let us consider the case where the full information solution is not IC because the first inequality in Equation 18 is not satisfied. We take $a_3 > a_2$, so that $I(W)$ is decreasing in the intentional delay δ when delay W is 'small', and increasing otherwise. Therefore, if W_1 is sufficiently 'small' under the full information solution, adding intentional delay to class 1 will increase $I(W_1)$ and improve IC. Note that when $C_1(\cdot)$ dominates $C_2(\cdot)$ (as is the case when $a_2 > a_3$), $I(W)$ is monotonically increasing in the delay, and hence adding intentional delay to class 1 only makes the IC constraints more binding. This was the case with multiplicative separability, where the IC region was a box at the upper-left corner of the space (see Figure 5). The following example shows the way the optimal solution may change in the absence of multiplicative separability.

Example 2 For the case $P_1(\lambda_1) = 600 - 200\lambda_1$, $P_2(\lambda_2) = 1000 - 300\lambda_2$, $\mu_1 = \mu_2 = 1$, $C_1(W) = 0.01W^2$ and $C_2(W) = 10W$, with Poisson arrivals and exponentially distributed service times, we obtain $\sigma^2(\lambda) = \lambda_1 + \lambda_2$, and therefore $\theta(\lambda) = \frac{1}{\lambda_1 + \lambda_2} - 1$. A numerical study shows that the optimal subscription rates are $\lambda_1 = 0.32$, $\lambda_2 = 0.628$, and the optimal scheduling rule involves using the rule suggested in Proposition 11 with $\tau = 134.7$, and then intentionally delaying class 1 for 4.06 time units. We, therefore, see that the optimal solution is not on the extended EF, and involves intentionally delaying class 1, while class 2 is not given priority.

The achievable region and IC region for this model are four-dimensional, and it is difficult to illustrate them, and therefore, in order to better understand this cost structure, we revert to a similar model, where a customer's cost is a function of the long-run average delay experienced. This model makes sense when the grades of service specify expected delays, and the customers need to make their decision based on that information alone. In this setting, the achievable region for two classes remains two-dimensional, and we can therefore better illustrate the issues arising when multiplicative separability does not hold. Defining $\hat{I}(W) = C_1(W) - C_2(W) - P_1(\lambda_1) + P_2(\lambda_2)$, the IC constraints can be expressed as follows:

$$\begin{aligned} \hat{I}(EW_1) &\leq 0 \\ \hat{I}(EW_2) &\geq 0. \end{aligned} \tag{20}$$

Now, if $\hat{I}(\cdot)$ has multiple roots, the IC region in the $EW_1 \times EW_2$ space has a checkered form, and because it is no longer a connected set, the extended efficient frontier could pass in between the "islands" of IC, as seen in Figure 8. In that case, no point on the extended efficient frontier is IC, and therefore nor is it optimal. Even if some far point on the extension to the efficient frontier is IC, there might be a different IC point (in one of the other "islands") that generates higher profit.

References

- [1] P. Afeche, Incentive Compatible Revenue Management in Queueing Systems: Optimal Strategic Idleness and other Delay Tactics, submitted for publication (2004).
- [2] J. Bulow and J. Roberts, The Simple Economics of Optimal Auctions, *Journal of Political Economy* 97 (1989), 1060-1090.
- [3] R. Chiang and C.S. Spatt, Imperfect Price Discrimination and Welfare, *The Review of Economic Studies* 49 (1982), 155-181.

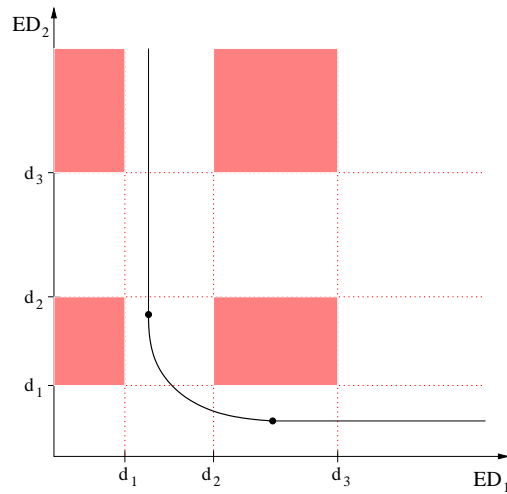


Figure 8: A different IC structure

- [4] R.J. Deneckere and R.P McAfee, Damaged Goods, *Journal of Economics and Management Strategy* 5 (1996), 149-174.
- [5] M. Harris and R.M. Townsend, Resource Allocation Under Asymmetric Information, *Econometrica* 49 (1981), 33-64.
- [6] A.K. Katta and J. Sethuraman, Pricing strategies and service differentiation in queues - A profit maximization perspective, submitted (2005).
- [7] P. Klemperer, Auction Theory: A Guide to the Literature, *Journal of Economic Surveys* 13 (1999), 227-286.
- [8] D.G. Luenberger, *Optimization by Vector Space Methods*, John Wiley and Sons, New York, 1969.
- [9] E. Maskin and J. Riley, Monopoly with Incomplete Information, *Rand Journal of Economics* 15 (1984), 171-196.
- [10] H. Mendelson and S. Whang, Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue, *Operations Research* 38 (1990), 870-883.
- [11] M. Mussa and S. Rosen, Monopoly and Product Quality, *Journal of Economic Theory* 18 (1978), 301-317.
- [12] R.B. Myerson, Incentive Compatibility and the Bargaining Problem, *Econometrica* 47 (1979), 61-73.
- [13] R.B. Myerson, Optimal Auction Design, *Mathematics of Operations Research* 6 (1981), 58-73
- [14] M. Rothschild and J.E. Stiglitz, Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information, *The Quarterly Journal of Economics* 90 (1976), 629-649.

- [15] K.T. Talluri and G.J Van Ryzin, *Revenue Management*, Kluwer Academic Publishers, Norwell, Massachusetts, 2004.
- [16] J.A. Van Mieghem, Dynamic Scheduling with Convex Delay Costs: The General $c\mu$ rule, *Annals of Applied Probability* 5 (1995), 809-833.
- [17] J.A. Van Mieghem, Price and Service Discrimination in Queueing Systems: Incentive Compatibility of $Gc\mu$ scheduling, *Management Science* 46 (2000), 1249-1267.
- [18] R.B. Wilson, Auctions of Shares, *The Quarterly Journal of Economics* 93 (1979),675-689.
- [19] R.B. Wilson, *Nonlinear Pricing*, Oxford University Press, New York, 1993.
- [20] T. Yahalom, J.M. Harrison and S. Kumar, Adverse Selection in Loss Systems with Capacity Constraints and Demand Uncertainty, working paper (2005).