Optimal Price/Lead-Time Menus for Queues with Customer Choice: Segmentation, Pooling, and Strategic Delay

Forthcoming in *Management Science*

Philipp Afèche
Rotman School of Management, University of Toronto, Toronto, Ontario, Canada M5S 3E6
afeche@rotman.utoronto.ca

J. Michael Pavlin
School of Business and Economics, Wilfrid Laurier University, Waterloo, Ontario, Canada N2L 3C5 mpavlin@wlu.ca

April 27 2015

How should a firm design a price/lead-time menu and scheduling policy to maximize revenues from heterogeneous time-sensitive customers with private information about their preferences? We consider a queueing system with multiple customer types that differ in their valuations for instant delivery and their delay cost rates. The distinctive feature of our model is that the ranking of customer preferences depends on lead times: patient customers are willing to pay more for long lead times than impatient ones, and vice versa for speedier service. We provide necessary and sufficient conditions, in terms of the capacity, market size, and properties of the valuation-delay cost distribution, for three features of the optimal menu and segmentation. 1. *Pricing out the middle of the delay cost spectrum* while serving both ends; 2. *Pooling* types with different delay costs into a single class; and 3. *Strategic delay* to deliberately inflate lead times.

Key words: congestion, delay, incentives, lead times, mechanism design, pooling, pricing, priorities, quality of service, queuing systems, revenue management, scheduling, service differentiation, strategic delay

1. Introduction

Firms such as Amazon, Dell or Federal Express, serve time-sensitive customers whose willingness to pay for a product or service also depends on the lead time between order placement and delivery. To exploit heterogeneous customer preferences – some value speedy service more than others – firms may offer a menu of differentiated price/lead-time options (same-day, two-day, etc.) as a revenue management tool, giving impatient customers the option to pay more for faster delivery while charging less for longer lead times. This paper studies the joint problem of designing the revenue-maximizing price/lead-time menu and the corresponding scheduling policy for a monopoly provider *who cannot tell apart individual customers* but only has aggregate information on their preferences, e.g., based on market research. We study this problem within a queueing model and consider customers who are heterogeneous in their valuations for instant delivery and their delay cost rates. This problem has recently received some attention, but as detailed below, significant gaps remain in understanding its solution for the case with *multiple delay cost types* considered here. This paper contributes to closing these gaps. We focus on the case where the valuations of types are strictly increasing and affine in their delay costs, and the valuation-to-delay cost ratio is decreasing in impatience. This yields the *simplest* model with multiple delay costs that gives rise to a *lead-time-dependent ranking of types*. That is, patient customers are willing to pay more or less than impatient customers for a given lead time, depending on whether it is above or below some indifference threshold. This is an important novel feature of our model and is critical for our results. It allows us to capture, in a one-dimensional type model, a key property of preferences over two attributes: one or the other dominates depending on the service option. To our knowledge this paper

presents the first results that specify the optimal menu for multiple types with lead-time-dependent ranking. We address three questions.

1. *Customer segmentation.* Which customer types – most, least or moderately time-sensitive – should be served?

2. *Priorities, pooling, and strategic delay.* Should the menu target a distinct price/lead-time class to each customer type based on her delay cost and prioritize types accordingly? Or, is it optimal to offer less than a full range of classes and target some types with different delay costs to be *pooled* in the *same* price/lead-time class? Should the scheduling policy be work conserving, or involve *strategic delay* to deliberately inflate some lead times above operationally feasible levels?

3. *Impact of capacity and demand attributes.* How do the optimal segmentation and menu depend on the capacity, the market size, and customer preferences?

We identify low, medium, and high capacity regimes. At low capacity it is optimal to differentiate lead times based on delay costs. At medium capacity it is optimal to pool intermediate types but sell differentiated lead times to more and less patient types. If valuations are below some threshold, then at medium capacity it is also optimal to price out some intermediate types. These medium capacity results hinge on the lead-time-dependent ranking of types. Unlike typical pooling results in screening models, our pooling result holds even if the type distribution has a monotone hazard rate, as our analysis assumes. At high capacity it is optimal to differentiate lead times either for all types, or only for sufficiently impatient ones while strategically delaying more patient ones.

1.1. Literature and Positioning

This paper bridges research streams on queueing systems and mechanism design. See Stidham (2002) for a survey of research on the analysis, design and control of queueing systems in settings where the system manager is fully informed and determines all job flows. Our analysis builds on the achievable-region approach, pioneered by Coffman and Mitrani (1980).

Mechanism design tools have been applied to many resource allocation problems under private information. Rochet and Stole (2003) survey screening studies in the economics literature. Among these, papers on the design of price-quality menus are closest to ours. In their seminal paper Mussa and Rosen (1978) consider a model with one-dimensional types. Rochet and Choné (1998) study its multidimensional version. While quality pooling is known to be potentially optimal in these "standard" screening models, they *rule out operational interdependencies among quality levels.* With this simplifying assumption their analysis can focus on interdependencies due to customer self-selection, and their results are invariant to capacity and market size. These models are therefore not designed to generate meaningful prescriptions in our setup where operational interdependencies among lead times play an important role in addition to customer self-selection. The capacity constraint and queueing effects imply externalities among service classes, and the provider controls these externalities through the price/lead-time menu and the scheduling policy. These features considerably complicate the problem as explained in Section 2.2.

Several papers that study variations of the classic price-quality design problem also ignore queueing effects. Dana and Yahalom (2008) introduce a capacity constraint. Bansal and Maglaras (2009) consider a model with multiple consumer types that are satisficing as opposed to utility maximizing, and rely on a deterministic analysis of the optimal menu, so that the queueing effects reduce to a capacity constraint. Neither study reports the phenomena we identify. Quality degradation similar to strategic delay has been studied in the damaged goods literature (Deneckere and McAfee 1996;

McAfee 2007). Anderson and Dana (2009) unify the results of the damaged goods literature and other well known pricing results, by identifying a necessary condition for price discrimination to be profitable in models with a quality constraint but ample capacity. We discuss these connections to our strategic delay results in Section 6.3.

This paper is part of a research stream on pricing and operational decisions for queueing systems with self-interested and time-sensitive customers. See Hassin and Haviv (2003) and Stidham (2009) for surveys. We consider static price/lead-time menus, unlike papers on dynamic price and/or lead time quotation (e.g., Plambeck 2004, Çelik and Maglaras 2008, Ata and Olsen 2013).

Three problem features jointly distinguish this from most papers on static price/lead-time optimization. (1) *Revenue maximization*: The objective is to maximize the provider's revenue, not the total system benefit (cf. Mendelson and Whang 1990). (2) *Customer choice over menu options*: Types have private information on their preferences and can choose their class, unlike in models that restrict each type to a single class (cf. Boyaci and Ray 2003; Maglaras and Zeevi 2005; Zhao et al. 2012). (3) *Scheduling optimization*: The provider chooses the scheduling policy, unlike in papers that *fix* the policy (cf. Naor 1969; Mendelson 1985; Rao and Petersen 1998; Afèche and Mendelson 2004). In studies that lack feature (1) or (2) neither pooling nor strategic delay can be optimal.

Only a few papers on static price/lead-time menus capture all three attributes. Afèche (2004, 2013), henceforth AF, is the first in the queueing literature to identify strategic delay and characterize its optimality. His model considers only two delay cost types, but allows heterogenous valuations for each delay cost level, unlike our model. Katta and Sethuraman (2005), henceforth KS, is the first and only other queueing paper that considers optimal pooling. Like our model, theirs also considers multiple types with perfectly correlated valuations and delay costs. However, contrary to this paper, they restrict the valuation-to-delay cost ratio to be *decreasing* in impatience. This superficially trivial distinction is important: It implies that the ranking of types is *lead time-invariant,* which rules out the key features in our findings. We further discuss the relationship to KS in Section 6.2. Maglaras et al. (2015) consider the menu-design problem for an unrestricted valuation-delay cost distribution, that is, without the perfect correlation assumed in this paper. However, unlike this paper, they provide no analytical results that identify which solution features arise in terms of the demand and capacity parameters. Rather, they focus on optimal decisions in large scale systems, that is, with large capacity and market size. They propose a deterministic relaxation (DR) of the problem that ignores queueing effects but captures the essential behavior of these large scale systems. The DR is typically not solvable in closed form in multi-type settings, although it is computationally more tractable than the original stochastic problem. They show how to construct a policy based on the DR solution that is asymptotically near-optimal in large-scale systems.

In brief, our contributions are as follows:

1. *Customer type model.* Our model is distinctive in that it jointly considers *multiple delay costs*, unlike AF, and a *lead-time-dependent ranking of types*, unlike KS. It yields novel results and offers a unifying framework for explaining the presence or absence of these results in related models. We return to these connections in Section 6.

2. *Optimal customer segmentation and price/lead-time menu.* We provide necessary and sufficient conditions on the demand and capacity parameters, for three striking solution features. (*i*) *Pricing out the middle of the delay cost spectrum* while serving both ends: This feature does not arise in KS. (*ii*) *Pooling* types with different delay costs: This feature does not arise in the two-type model of AF. Our results are more general and more informative than those in KS. Most importantly, unlike

in KS, in our model pooling arises even for delay cost distributions with a monotone hazard rate. (*iii*) *Strategic delay*: Our results complement those of AF. We identify optimal strategic delay in a setting with multiple, not just two, types, but we ignore the effect of valuation heterogeneity.

2. Model, Problem Formulation and Analysis Roadmap

We model a service or make-to-order manufacturing provider as an $M/M/1$ queueing system. Potential customers with unit demand arrive according to an exogenous Poisson process with rate or market size Λ. The system has i.i.d. exponential service times with mean $1/\mu$, where μ is the capacity. The capacity is not a decision variable, but our results specify how the optimal menu varies with μ. We normalize the marginal cost of service to zero.

Preferences. Customers differ in their valuations for immediate delivery of the product or service, and in their linear delay costs. We use the term "lead time" to refer to the entire time between order placement and delivery. We consider a continuum of customer *types* indexed by c, which denotes the customer's delay cost per unit of lead time. Types c are i.i.d. draws from a continuous distribution F with p.d.f. f, which is assumed strictly positive and continuously differentiable on the interval $\mathcal{C} \triangleq [c_{\min}, c_{\max}] \subset [0, \infty)$. Let $\overline{F} = 1 - F$. The service time and delay cost rate distributions are mutually independent and independent of the arrival process.

Valuations and delay costs are perfectly correlated. A type c customer has positive valuation $V(c)$ for immediate delivery, where $V : \mathcal{C} \to \mathbb{R}_+$ is a monotone and continuous function. The analysis focuses on $V(c) = v + c \cdot d$, where v and d are constants. The *base value* v is a scale parameter for valuations. As discussed below, the slope of the valuation-delay cost relationship d can also be viewed as a *threshold lead time* that determines the ranking of customers' willingness to pay for a given service class. The paper focuses on the case $v > 0$ and $d > 0$ since it gives rise to novel results. It also covers $v \leq 0 < d$ and $v > 0 \geq d$, in which case our model specializes to related models. Section 6.4 outlines how our results generalize if $V(c)$ is not affine.

The case of perfectly positively correlated valuations and delay cost rates ($d > 0$) is well suited for settings where delays deflate values. A variety of important phenomena lead to delay-driven value losses (cf. Afèche and Mendelson 2004), including physical decay of perishable goods during transportation delays, technological or market obsolescence of short life-cycle products such as computer chips or fashion items, and delayed information in industrial and financial markets.

A type c customer's *net* value or willingness to pay for an expected lead time w is $N(c, w) \triangleq V(c) - c \cdot w = v + c \cdot (d - w)$, and her utility at price p is $v + c \cdot (d - w) - p$. This net value function has two standard properties. First, it decreases in the lead time, i.e., the partial derivative $N_w(c, w) < 0$. This captures the notion of vertical product differentiation in that service classes can be objectively ranked from fastest to slowest, or from highest to lowest "quality". Second, the more time-sensitive a type, the sharper her net value decline as the lead time increases: $N_{cw}(c, w) < 0$. This is the single-crossing condition. The distinctive feature of our model is that *the ranking of types' net values is lead-time-dependent*. Specifically, $N_c(c, w) = d - w$, so that net values for lead times $w < d$ increase, while those for lead times $w > d$ decrease, in the time-sensitivity c. In this sense, the parameter d represents a threshold lead time. This feature describes situations where impatient customers are willing to more than patient customers for speedy service (e.g., overnight delivery) but less than patient customers for slow service (e.g., delivery in several business days).

To rule out the case where no type has a positive expected net value from service even in the absence of waiting, we assume that $\mu^{-1} < d + v/c_{\min}$ if $v > 0$, and $\mu^{-1} < d + v/c_{\max}$ if $v \leq 0$.

Information. The provider knows the arrival process, the delay cost rate distribution f, the function $V(c)$ and the service time distribution. However, a customer's type c is her private information. Service time realizations become known only once jobs are processed to completion. Only the provider observes the system queues. Customers lack this information.

Decisions. The provider announces a static price/lead-time menu and chooses a scheduling policy to maximize her long run average revenue rate. We outline the set of admissible scheduling policies below. The menu specifies for each class two attributes, the price and the expected lead time. We simply say "lead time" for the expected lead time of a class. Because we consider a continuum of types with respect to their delay preferences, the provider will optimize over a continuum of service classes. ("Class" refers to the attributes of a service option, "type" refers to those of a customer.) Because customers have private information, they can choose among all classes, and the provider must consider their choice behavior.

Formally, the provider selects a menu of lead-time/price options $\{(w, P(w)) : w \in \mathcal{W}\}$ where $\mathcal{W}$ denotes the set of offered lead times and the function $P : \mathcal{W} \to \mathbb{R}$ assigns prices to lead times.

Customer arrival times are exogenous. However, customers are strategic in their purchase decisions. We assume they are risk neutral with respect to leadtime uncertainty. Upon arrival at the facility customers decide, based on the posted menu, which service class to purchase, if any. Specifically, a customer of type c determines the lead-time/price pair $(w, P(w))$ that maximizes her expected utility from service, $U(w, P(w); c) \triangleq v + c \cdot (d - w) - P(w)$. Let $w(c) \triangleq \arg\max_{w \in \mathcal{W}} \{U(w, P(w); c)\}$ and $p(c) \triangleq P(w(c))$ denote the preferred lead-time/price pair for type c, and let $U(c) \triangleq U(w(c), p(c); c) = v + c(d - w(c)) - p(c)$ denote the corresponding expected utility from service. Customers who do not purchase balk and receive zero utility, so they only purchase if their expected utility from service is non-negative. We assume no retrials and no reneging. We keep track of purchase decisions with the acceptance function $a : \mathcal{C} \to \{0, 1\}$ where $a(c) = 1$ if type c buys service, choosing the lead-time/price pair $(w(c), p(c))$, and $a(c) = 0$ otherwise. Let $\mathcal{C}_a \triangleq \{c \in \mathcal{C} : a(c) = 1\}$ denote the set of types that buy service and $\overline{\mathcal{C}}_a = \mathcal{C} \backslash \mathcal{C}_a$ its complement. Let $\lambda \triangleq \Lambda \int_{x \in \mathcal{C}_a} f(x)\, dx$ denote the resulting arrival rate. Best responses to a menu satisfy $c \in \mathcal{C}_a$ if $U(c) > 0$ and $c \in \overline{\mathcal{C}}_a$ if $U(c) < 0$. Types with zero expected utility may or may not purchase as discussed in Section 3.

Lead times and admissible scheduling policies. We assume that customers base their decisions on the *announced* expected lead times. However, we require that the announced expected lead times equal the average steady-state lead times that are realized given the capacity μ, the scheduling policy, and the customers' purchase decisions that are induced by the menu. This consistency requirement reflects the notion that reputation effects and third party auditors instill in the provider the commitment to perform in line with her announcements (cf. Afèche 2013).

We do not assume a specific scheduling policy but rather let the provider optimize over the following set of admissible scheduling policies: 1. We focus on nonanticipative and regenerative policies. This appears to be the most general, easily described restriction under which the existence of long-run lead time averages may be verified. 2. We do not restrict attention to work conserving policies. Specifically, we allow the insertion of *strategic delay* whereby the provider artificially increases the lead times for a subset of service classes above the levels that are operationally achievable. See Afèche (2004, 2013) for a detailed discussion of strategic delay. 3. We allow preemption, which does not affect the results but simplifies the analysis. In particular, under priority scheduling, with preemption the lead time of a given class does not depend on the arrival rates to lower-priority classes. We build on the achievable region approach to multiclass scheduling problems, cf. Coffman and Mitrani (1980). Problem 1 below specifies the achievable region for these admissible policies.

2.1. Mechanism Design Formulation

We formalize the provider's problem as a mechanism design problem. Based on the revelation principle (e.g., Myerson 1979), we restrict attention, without loss of generality, to direct mechanisms in which customers have the incentive to truthfully report their type. The procedure described above, whereby customers make decisions based on a menu, is strictly speaking an indirect mechanism, but it is more descriptive of how services are sold, and it is equivalent to a direct revelation mechanism in which customers truthfully reveal their type. This requires that the functions (a, w, p) satisfy the *individual rationality* (IR) and *incentive-compatibility* (IC) constraints. IR requires that the expected utility from service be non-negative for types who are targeted for service and non-positive for all others: $U(c) \geq 0$ for $c \in \mathcal{C}_a$ and $U(c) \leq 0$ for $\in \overline{\mathcal{C}}_a$. IC requires that each type c maximizes her expected utility if it truthfully reports its type: $U(c) \geq U(w(c'), p(c'); c)$ for $c \neq c'$.

Problem 1.
$$\max_{a:\mathcal{C}\to\{0,1\},\ w:\mathcal{C}\to\mathbb{R},\ p:\mathcal{C}\to\mathbb{R}} \Lambda \int_{c_{min}}^{c_{max}} a(x)f(x)p(x)dx \tag{1}$$

$$\text{subject to} \qquad \mu > \Lambda \int_{x\in\mathcal{C}_a} f(x)\,dx, \tag{2}$$

$$\frac{\Lambda}{\mu}\int_{x\in s} f(x)w(x)dx \geq \frac{\frac{\Lambda}{\mu}\int_{x\in s} f(x)dx}{\mu - \Lambda\int_{x\in s} f(x)dx}, \qquad \forall s \subset \mathcal{C}_a, \tag{3}$$

$$U(c) = v + c(d - w(c)) - p(c) \geq 0 \qquad \forall c \in \mathcal{C}_a, \tag{4}$$

$$U(c) = v + c(d - w(c)) - p(c) \leq 0 \qquad \forall c \in \overline{\mathcal{C}}_a, \tag{5}$$

$$w(c)\cdot c + p(c) \leq w(c')\cdot c + p(c'), \qquad \forall c \neq c'. \tag{6}$$

Constraint (2) ensures that the system is stable. Constraints (3) ensure that the lead times $\{w(c) : c \in \mathcal{C}_a\}$ are *operationally achievable*. The right-hand side (RHS) of (3) is the long run average work in the system under a work conserving policy that gives all admitted customers in the set s strict preemptive priority over all others. It equals the average work in a first-in-first-out (FIFO) $M/M/1$ system with arrival rate $\Lambda\int_{x\in s} f(x)dx$ and capacity μ. A scheduling policy is *work conserving* if (3) is binding for $s = \mathcal{C}_a$. Constraints (4)-(5) capture IR and (6) capture IC. The menu corresponding to a feasible (a, p, w) satisfies $\mathcal{W} = \{w(c) : c \in \mathcal{C}\}$ and $P(w(c)) = p(c)$ for $w(c) \in \mathcal{W}$.

First-best benchmark: observable types. The first-best problem, in which the provider *observes* the types, yields a considerably simpler version of Problem 1 as the IC constraints (6) are dropped. The provider can charge each type the full amount of her net value; that is, type c pays $p(c) = v + c(d - w(c))$. In this case, a standard work conserving strict priority policy is optimal. It prioritizes admitted types by their delay costs. Hence the menu offers *all* lead times within an interval and each admitted type buys a different lead time.

2.2. Analysis Roadmap

We develop the solution of Problem 1 using the following 3-step approach.

STEP 1. *Incentive-compatible segmentation and lead times, optimal prices* (Section 3). We translate the IR and IC constraints (4)-(6) into equivalent properties that any feasible and revenue-maximizing triple (a, p, w) must satisfy. These properties yield a segmentation of customer types into three segments and also imply the optimal prices for given segmentation and lead times, reducing Problem 1 to one of choosing the arrival rates and lead times for these segments.

STEP 2. *Optimal segmentation and lead times for fixed arrival rate* (Section 4). We characterize the optimal segmentation and lead time menu depending on λ, Λ, μ, d and the distribution f.

STEP 3. *Optimal arrival rate, segmentation, and lead times* (Section 5). We characterize the solution at the *optimal* λ, for fixed capacity μ, and as a function of μ for given demand parameters.

STEP 1 is based on standard mechanism design methods, but STEPS 2 and 3 are *not.* The capacity constraint and queueing delays introduce operational interdependencies among lead times, which significantly complicate the analysis. Following the seminal work of Mussa and Rosen (1978), price quality menu design problems in the economics literature rule out operational interdependencies among quality levels, which simplifies the analysis. To be specific, if one removes the queueing-related operational constraints (3) in Problem 1 and introduces instead a quality cost function in the objective function[1], then under regularity conditions on the distribution f the problem is quickly solved point-wise for each type c, and the solution is invariant to the market size. This point-wise approach fails in the presence of queueing effects. STEPS 2 and 3 consider these effects by building on the achievable region approach, but our problem calls for three important modifications: We account for the IC and IR constraints, we optimize over arrival rates, and we allow strategic delay.

3. Incentive-Compatible Segmentation and Lead Times, Optimal Prices

Given a triple (a, w, p) we partition the set of admitted types $\mathcal{C}_a$ into the following three segments:

$$C_l \triangleq \{c \in \mathcal{C}_a : w(c) > d\}, \quad C_m \triangleq \{c \in \mathcal{C}_a : w(c) = d\}, \text{ and } C_h \triangleq \{c \in \mathcal{C}_a : w(c) < d\}. \tag{7}$$

For simplicity we suppress the dependence of C_l, C_m and C_h on a. We call classes with $w > d$ *low lead time quality* or l classes, those with $w < d$ *high lead time quality* or h classes, and the class with the threshold lead time $w = d$ the *medium lead time* or m class.

PROPOSITION 1. *Fix a triple (a, w, p). Define the marginal types c_l and c_h as follows:*

$$c_l \triangleq \begin{cases} \sup C_l & \text{if } C_l \neq \varnothing \\ c_{\min} & \text{otherwise} \end{cases}, \quad \text{and} \quad c_h \triangleq \begin{cases} \inf C_h & \text{if } C_h \neq \varnothing \\ c_{\max} & \text{otherwise} \end{cases}.$$

Suppose that (a, w, p) maximizes the revenue rate. Then (a, w, p) satisfies the IR and IC constraints (4)-(6) if and only if the following properties hold.

1. *Lead times $w(c)$ are non-increasing, prices $p(c)$ are non-decreasing, and $c_l \leq c_h$.*
2. *If there is a segment of types who buy low lead time qualities ($C_l \neq \varnothing$) then: (a) it is an interval that includes $c_{\min}$, i.e., $c < c_l \Rightarrow c \in C_l$; (b) prices and expected utilities from service satisfy:*

$$p(c) = v + c \cdot (d - w(c)) - \int_c^{c_l} (w(x) - d)\, dx, \quad \forall c \in [c_{min}, c_l], \text{ where } p(c) < v \text{ for } c < c_l, \tag{8}$$

$$U(c) = \int_c^{c_l} (w(x) - d)\, dx, \quad \forall c \in [c_{min}, c_l], \text{ where } U(c) > 0 = U(c_l) \text{ for } c < c_l. \tag{9}$$

3. *If there is a segment of types who buy high lead time qualities ($C_h \neq \varnothing$) then: (a) it is an interval that includes $c_{\max}$, i.e., $c > c_h \Rightarrow c \in C_h$; (b) prices and expected utilities from service satisfy:*

$$p(c) = v + c \cdot (d - w(c)) - \int_{c_h}^{c} (d - w(x))\, dx, \quad \forall c \in [c_h, c_{max}], \text{ where } p(c) > v \text{ for } c > c_h, \tag{10}$$

[1] The model of Mussa and Rosen (1978) is also simpler than this quality-cost version of our model, because they consider types with a quality-independent ranking, whereas in our model the ranking of types is lead-time-dependent.

$$U(c) = \int_{c_h}^{c} (d - w(x))\, dx, \quad \forall c \in [c_h, c_{max}], \text{ where } U(c) > 0 = U(c_h) \text{ for } c > c_h. \quad (11)$$

4. *If there is a segment of types who buy the medium lead time ($C_m \neq \varnothing$) then: (a) $C_m \subset [c_l, c_h]$; (b) the prices and expected utilities from service satisfy $p(c) = v$ and $U(c) = 0$ for $c \in C_m$.*
5. *Types in (c_l, c_h) buy the medium lead time or do not buy at all, i.e., $(c_l, c_h) \subset C_m \cup \overline{C}_a$, and*

$$U(c) = U(c_l) - \int_{c_l}^{c} (w(x) - d)\, dx = U(c_h) - \int_{c}^{c_h} (d - w(x))\, dx \leq 0, \quad \forall c \in [c_l, c_h], \quad (12)$$

where $U(c) = 0 \ \ \forall c \in [c_l, c_h]$ if some types buy the medium lead time ($C_m \neq \varnothing$).

All proofs are in the Appendix. Consider the net value as a function of the lead time w and the type c, that is, $N(c, w) = V(c) - cw$. Part 1 follows because $N(c, w)$ decreases in lead time.

Parts 2-5 follow because the net value $N(c, w)$ for a fixed lead time w changes at the rate $V'(c) - w = d - w$ as the customer type increases, where d and w capture the rate of increase in valuation and delay cost, respectively. As a result, net values for lead times $w > d$ are decreasing, whereas net values for lead times $w < d$ are increasing, in customer impatience. Therefore, in Part 2 of Proposition 1, if a type c chooses to buy the lead time $w(c) > d$, then more patient types $c' < c$ have a higher net value for this lead time than c. Since IR holds for type c, more patient types must get strictly positive utility for IC to hold, i.e. they must be served. Therefore, the set C_l is an interval that includes the most patient type $c_{\min}$, and by (9) the expected utilities of customers who buy low lead time qualities decrease in their impatience. By (8) the price paid by a type $c < c_l$ decreases in the lead times of more impatient types in C_l, because the longer these lead times the more type c values them in comparison to more impatient types. Similarly, in Part 3 of Proposition 1, if a type c buys a lead time $w(c) < d$, then more impatient types $c' > c$ must be served with strictly positive utility. Therefore, the set C_h is an interval that includes the most impatient type $c_{\max}$, and by (11) the expected utilities of customers who buy high lead time qualities increase in their impatience. In contrast to the prices (8) that decrease in lead times, by (10) the price paid by a type $c > c_h$ *increases* in the lead times of more patient types in C_h. This holds because the longer these lead times the less type c values them in comparison to more patient types. By Parts 4 and 5, the set of customers C_m who purchase the threshold lead time d is a subset of $[c_l, c_h]$ that need not be an interval. If $C_m \neq \varnothing$, then the lead time d is offered at a price of v, every type has zero expected utility from this option, but only types in $[c_l, c_h]$ have no better option available and are indifferent between buying and not doing so.

By Proposition 1 it may be optimal to price the most, the least or only moderately impatient types out of the market. Furthermore, choosing C_l, C_m and C_h reduces to choosing the corresponding arrival rates, which we use to replace the acceptance function a. Let λ_l, λ_m and λ_h be the rates for l, m, and h classes, respectively, where $\lambda = \lambda_l + \lambda_m + \lambda_h$. Note that $\lambda_l = \Lambda F(c_l)$ and $\lambda_h = \Lambda \overline{F}(c_h)$ uniquely determine the corresponding function a for types $c < c_l$ and $c > c_h$, respectively, whereas λ_m only determines the *mass* of types $c \in [c_l, c_h]$ who buy the medium lead time. Any feasible triples (a, w, p) and (a', w, p) that only differ in the set, but not the mass, of types who buy the medium lead time are revenue equivalent. If $\lambda_m > 0$ then a positive mass of different types are pooled at the medium lead time. If in addition $\lambda_m < \Lambda - \lambda_l - \lambda_h$ then there are two positive masses of types in $[c_l, c_h]$ with zero expected utility, those who buy the medium lead time and those who do not buy any service. This feature arises because $V(c)$ is affine, but as outlined in Section 6.4, our main structural results remain valid for a broader class of $V(c)$ functions.

4. Optimal Segmentation and Lead Times for Fixed Arrival Rate

4.1. Virtual Delay Costs and Solution Preview

Virtual delay costs. Write $\Pi(\lambda_l, \lambda_h, \lambda, w)$ for the revenue as a function of the arrival rate λ, the segmentation characterized by λ_l and λ_h, and the lead time function w. Substituting the prices (8) and (10) into (1) yields

$$\Pi(\lambda_l, \lambda_h, \lambda, w) = \lambda v + \Lambda \int_{c_{\min}}^{c_l(\lambda_l)} f(c) f_l(c) (d - w(c)) \, dc + \Lambda \int_{c_h(\lambda_h)}^{c_{\max}} f(c) f_h(c) (d - w(c)) \, dc, \tag{13}$$

where $c_l(\lambda_l) = F^{-1}(\lambda_l/\Lambda)$, $c_h(\lambda_h) = \overline{F}^{-1}(\lambda_h/\Lambda)$ and the functions f_l and f_h are defined as follows:

$$f_l(c) \triangleq c + \frac{F(c)}{f(c)} \text{ for } c \in [c_{\min}, c_l], \tag{14}$$

$$f_h(c) \triangleq c - \frac{\overline{F}(c)}{f(c)} \text{ for } c \in [c_h, c_{\max}]. \tag{15}$$

We call f_l and f_h the *virtual delay cost functions*. They measure the marginal revenue effect from a reduction in the lead time of a given type. The virtual delay cost of a type c depends on its lead time: It is $f_l(c)$ for low quality ($w(c) > d$) and $f_h(c)$ for high quality ($w(c) < d$). The virtual delay cost consists of an *own price effect* and an *external price effect*. The own price effect is the delay cost c in (14) and (15). It simply measures how a lead time reduction for type c allows an increase in its own price. The external price effect is the second summand in (14) and (15). It measures how a lead time reduction for type c changes the prices of classes targeted to other types, in order to maintain IC. The external price effect is positive for low lead time qualities, so $f_l(c) > c$ (for $c > c_{\min}$). That is, reducing the lead time $w(c) > d$ of a type c in C_l allows price increases for classes targeted to more patient types $c' < c$ while maintaining IC. (By (8) the price paid by a type in C_l decreases in the lead times of more impatient types in C_l.) As explained in Section 3, this follows because a customer's net value for a lead time $w > d$ is higher the more patient that customer. In contrast, the external price effect is *negative* for high lead time qualities, so $f_h(c) < c$ (for $c < c_{\max}$). That is, reducing the lead time $w(c) < d$ of a type c in C_h requires price *decreases* for classes targeted to more impatient types $c' > c$ in order to maintain IC. (By (10) the price paid by a type in C_h increases in the lead times of more patient types in C_h). As explained in Section 3, this follows because a customer's net value for a lead time $w < d$ is higher the more impatient that customer. Furthermore, note that $f_h(c) < 0$ for a type c if its negative external price effect dominates its own price effect. In this case increasing the lead time $w(c) < d$ of this type increases revenues.

Solution preview. Maximizing the revenue (13) calls for lead times that are strictly decreasing in *virtual* delay costs, whereas IC requires that the lead times be appropriately ranked relative to the threshold d and non-increasing in delay costs. This gives rise to three striking solution features.

1. *Pricing out the middle of the delay cost spectrum.* It may be optimal to price out intermediate types. This is the only of the three features that also arises under the first-best menu.

2. *Pooling.* It may be optimal to target a common class with a single lead time to multiple types with different delay costs. If it is optimal to pool some types into the same class, then virtual delay costs must be decreasing over a subset of these types. This necessary condition has three variations. If pooling is (strictly) optimal at some lead time $w > d$, it must be that $f_l'(c) < 0$ for some pooled type. Similarly, $f_h'(c) < 0$ for some pooled type if pooling is optimal at a lead time shorter than d. If pooling is optimal at the threshold lead time d then $f_l(c_1) > f_h(c_2)$ for some pooled types $c_1 < c_2$.

Note that optimal pooling at the lead time threshold d can arise for *every* type distribution. Example 1 below explains why. In contrast, pooling at a lead time $w \neq d$ is *not optimal if* $f_l', f_h' > 0$, which holds for many common probability distributions, including those with log-concave density function (cf. Bagnoli and Bergstrom 2005). Examples include the uniform, normal, logistic, Laplace and power function distributions, and the gamma and Weibull distributions with shape parameter ≥ 1. However, delay cost distributions that are mixtures of unimodal distributions easily yield nonmonotone virtual delay cost functions. Such distributions might describe markets with multiple segments where across-segment delay cost differences are large relative to those within segments. We henceforth assume $f_l', f_h' > 0$. We further discuss this assumption in Section 6.1.

3. *Strategic delay.* It may be optimal to intentionally inflate the lead times of some types above operationally feasible levels, which is not work conserving. In our model doing so can only be optimal at the threshold lead time d. In this case all types c with $f_h(c) < 0$ buy the lead time d. Hence, strategic delay also implies pooling in our model, but the converse does not hold.

In contrast to the first-best solution, a menu that involves pooling, with or without strategic delay, has one or more "gaps" between the offered lead times. This implies less lead time differentiation among pooled types and more differentiation relative to neighboring types buying different classes.

4.2. Customer Segmentation and Lead Times for Fixed Arrival Rate

We now discuss STEP 2 of the solution approach outlined in Section 2.2. Let $\lambda_l^*(\lambda)$, $\lambda_m^*(\lambda)$ and $\lambda_h^*(\lambda)$ denote the optimal customer segmentation as a function of the arrival rate λ, where $\lambda_l^*(\lambda) + \lambda_m^*(\lambda) + \lambda_h^*(\lambda) = \lambda$. Write $c_l^*(\lambda) \triangleq F^{-1}(\lambda_l^*(\lambda)/\Lambda)$ and $c_h^*(\lambda) \triangleq \overline{F}^{-1}(\lambda_h^*(\lambda)/\Lambda)$ for the corresponding marginal types. Let $C_l^*(\lambda)$, $C_m^*(\lambda)$ and $C_h^*(\lambda)$ denote the sets of types buying low, medium and high lead time qualities, respectively. Finally, let the function $w^*(c;\lambda)$ denote the optimal lead times as a function of λ.

LEMMA 1. *Fix $\lambda \in (0, \Lambda] \cap (0, \mu)$. Assume strictly increasing virtual delay cost functions f_l, f_h.*

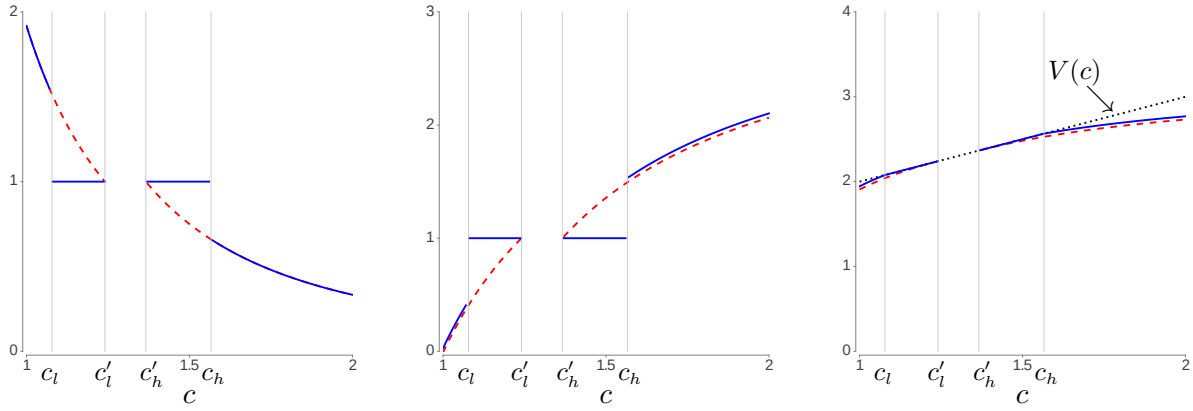
1. *The optimal lead time menu and corresponding scheduling policy have the following properties:*
   (a) *The lead times satisfy:*

$$w^*(c;\lambda) = \begin{cases} \frac{\mu}{\left(\mu - \Lambda \bar{F}(c)\right)^2} < d & \text{if } c \in C_h^*(\lambda), \\ d & \text{if } c \in C_m^*(\lambda), \\ \frac{\mu}{(\mu - [\lambda - \Lambda F(c)])^2} > d & \text{if } c \in C_l^*(\lambda). \end{cases} \tag{16}$$

   (b) *If low lead time qualities are sold ($C_l \neq \varnothing$) then the optimal policy is work conserving.*
2. *Under the optimal customer segmentation, the virtual delay costs are positive and increasing over types with low or high lead time quality: (a) $f_h(c_h^*(\lambda)) \geq 0$. (b) If $\lambda_l^*(\lambda) > 0$ and $\lambda_h^*(\lambda) > 0$ then $f_l(c_l^*(\lambda)) \leq f_h(c_h^*(\lambda))$; if in addition $\lambda_m^*(\lambda) > 0$ then $f_l(c_l^*(\lambda)) = f_h(c_h^*(\lambda))$.*

Lemma 1 provides a set of simple rules to determine the optimal lead time menu for a particular arrival rate. First, Lemma 1.1(a) limits pooling to types that buy the medium lead time d, i.e. types in $C_m^*(\lambda)$. Customers in $C_h^*(\lambda)$ receive strict priority over types in $C_m^*(\lambda)$ which receive strict priority over types in $C_l^*(\lambda)$. Different types in $C_l^*(\lambda)$ and $C_h^*(\lambda)$ buy different lead times and are strictly prioritized in the order of their delay cost. Different types in $C_m^*(\lambda)$ buy the same lead time and are pooled into a single FIFO service class. Second, Lemma 1.1(b) limits strategic delay to the case where no customers are served in the low lead time quality segment. Third, Lemma 1.2(a) rules out serving customers with negative virtual delay costs in the high quality segment; if

Figure 1 **Example 1. Optimal menu with intermediate types priced out or pooled.**



(a) Lead times $w(c)$ (b) Prices $p(c)$ (c) Full prices $p(c)+cw(c)$

Note. Optimal menu properties represented by solid lines, first-best menu properties by dashed lines. $\mu=3$, $\Lambda=2$, $\lambda=1.75$, $v=1$, $d=1$ and $f(c)$ uniform with $[c_{\min}, c_{\max}]=[1,2]$.

it is operationally feasible to serve such customers with lead time lower than d, they should instead be offered the lead time d, which yields strategic delay. Finally, Lemma 1.2(b) requires monotone virtual delay costs across admitted customers who are not pooled at lead time d.

Substituting the lead time function (16) in (13) yields the revenue function

$$\Pi(\lambda_l,\lambda_h,\lambda)\triangleq\lambda v-\Lambda\int_{c_{\min}}^{c_l(\lambda_l)}f(x)f_l(x)\left(\frac{\mu}{\left(\mu-[\lambda-\Lambda F(x)]\right)^2}-d\right)dx+\Lambda\int_{c_h(\lambda_h)}^{c_{\max}}f(x)f_h(x)\left(d-\frac{\mu}{\left(\mu-\Lambda\overline{F}(x)\right)^2}\right)dx. \tag{17}$$

Using Proposition 1 and Lemma 1, we reformulate Problem 1 to the following program.

Problem 2.

$$\max_{\lambda_l\geq 0,\lambda_h\geq 0,\lambda}\ \Pi(\lambda_l,\lambda_h,\lambda) \tag{18}$$

$$\text{subject to}\quad \lambda_l+\lambda_h\leq\lambda\leq\Lambda, \tag{19}$$

$$\lambda<\mu,$$

$$\frac{\mu}{(\mu-[\lambda-\lambda_l])(\mu-\lambda_h)}\leq d \text{ if } \lambda-\lambda_l>0, \tag{20}$$

$$\frac{\mu}{(\mu-[\lambda-\lambda_l])^2}\geq d \text{ if } \lambda_l>0. \tag{21}$$

Constraint (20) ensures that the lead times in the high and medium quality classes do not exceed d, and (21) ensures that the policy is work conserving if there are customers that are served with low lead time qualities. Constraint (20) implies the low capacity threshold $\mu=1/d$: At lower capacities, only low lead time qualities can be offered (that is, $\lambda-\lambda_l=0$), and neither pooling nor strategic delay are optimal by Lemma 1.

We illustrate Lemma 1 with the two examples shown in Figures 1 and 2. In both cases we compare the optimal menu to the first-best (FB) benchmark, which strictly prioritizes all types that are served in the order of their delay costs and operates a work conserving policy. In both figures the optimal menu properties are represented by solid lines, and FB menu properties by dashed lines.

Example 1: Pricing out or pooling intermediate types. In this example, shown in Figure 1, it is not feasible to serve all customers with a lead time shorter than d under strict priorities, because

$d = 1$ and the arrival rate $\lambda = 1.75 < \Lambda = 2$ is relatively high given the capacity $\mu = 3$. Consider the FB lead times, shown by the dashed line in Figure 1(a). The type c'_h receives $w(c) = d$ under strict priorities, and it is the lower bound on the types that are served in the high quality segment ($w < d$). The remaining types that are served must be admitted into the low quality segment, that is, their lead times exceed d. Because the net value for such lead times decreases in impatience, the optimal low quality segment is an interval that includes the most patient type, i.e., $[c_{\min}, c'_l]$. Because $\lambda < \Lambda$ the types in the remaining intermediate interval (c'_l, c'_h) are priced out.

Next consider how and why the optimal lead times involve pooling. The marginal types c'_l and c'_h under the FB lead times are close enough so that their virtual delay costs are strictly decreasing. That is, $f_l(c'_l) > f_h(c'_h)$ and $c'_l < c'_h$, which violates the necessary condition for optimality in Lemma 1.2(b). Recall from Section 4.1 that $f_l(c)$ and $f_h(c)$ have identical own price effects (equal to c), and that the external price effect of $f_l(c)$ is positive whereas that of $f_h(c)$ is negative. Because the types c'_l and c'_h are close enough, their external price effects dominate their own price effects, so that $f_l(c'_l) > f_h(c'_h)$ and, unlike in the FB menu, it is *not* optimal to serve c'_h with a shorter lead time than c'_l: Speeding up c'_l at the expense of c'_h increases revenues.[2] Starting with the FB lead times, the optimal lead times are obtained by pooling a set of customers (c_l, c'_l) from the low quality segment with a set of customers (c'_h, c_h) from the high quality segment into a common class such that $f_l(c_l) = f_h(c_h)$, as required by Lemma 1.2(b). As shown in Figure 1(a), compared to the FB lead times, the optimal lead times (solid lines) are lower for pooled types $(c_l, c'_l]$, higher for pooled types $[c'_h, c_h)$, and unchanged for the remaining strictly prioritized customers.

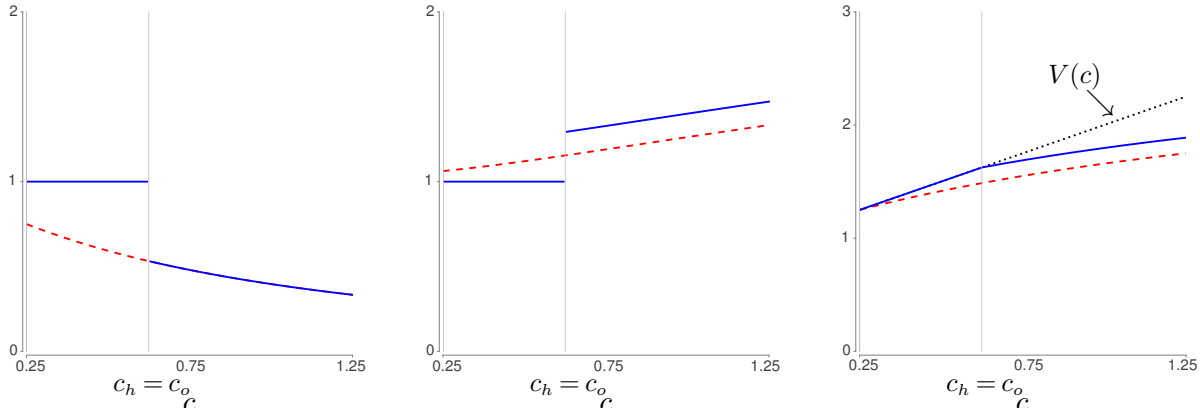
Figure 1(b) shows the prices. Compared to the FB prices, the optimal prices are lower only for the pooled types in $[c'_h, c_h)$ because their lead times are longer than in the FB menu. However, pooling increases revenue as the price reductions for these types are more than offset by price increases for all other types: Compared to the FB prices the provider can increase prices for the pooled types in $(c_l, c'_l]$ because their lead times are shorter, for the strictly prioritized patient types in $[c_{\min}, c_l]$ due to shorter lead times of the more impatient types in $(c_l, c'_l]$, and for the strictly prioritized impatient types in $[c_h, c_{\max}]$ due to the *longer* lead times of the more patient types in $[c'_h, c_h)$.

Figure 1(c) shows for both menus the value function $V(c) = v + cd$ and the full price $p(c) + cw(c)$. Their difference is the customer's utility. The full prices of the optimal menu strictly exceed those of the FB menu, resulting in lower customer utility and higher provider revenue.

Example 2: Strategic delay for the most patient types. This example, shown in Figure 2, differs in three important respects from Example 1. First, the market size $\Lambda = 1$ is smaller, so that under the FB menu all customers can be served with lead time lower than $d = 1$; refer to Figure 2(a). Second, all customers are served, i.e. $\lambda = \Lambda$. Third, types at the low end of the delay cost spectrum are so patient that their virtual delay costs are negative, i.e., $f_h(c) < 0$ for $c < c_0 = 0.625$ and $f_h(c_0) = 0$. For these types the external price effect of their virtual delay cost dominates their own price effect. That is, increasing the lead time for a type c with $f_h(c) < 0$ improves the revenue on more impatient types by more than it reduces the type-c price. By Lemma 1.2(a), optimality requires that all types in the high quality segment have a non-negative virtual delay cost, i.e., $f_h(c_h) \geq 0$. Therefore, as shown in Figure 2(a), compared to the FB lead times, it is optimal to increase the lead times of the types in $[c_{\min}, c_0]$ up to $d = 1$ by inserting *strategic delay*, and to leave the lead times of the types in $[c_0, c_{\max}]$ unchanged.

[2] If the marginal types of the low- and high-quality segments are sufficiently far apart, their virtual delay costs are non-decreasing and the FB lead times are optimal. The discussion of Figure 3 includes that case.

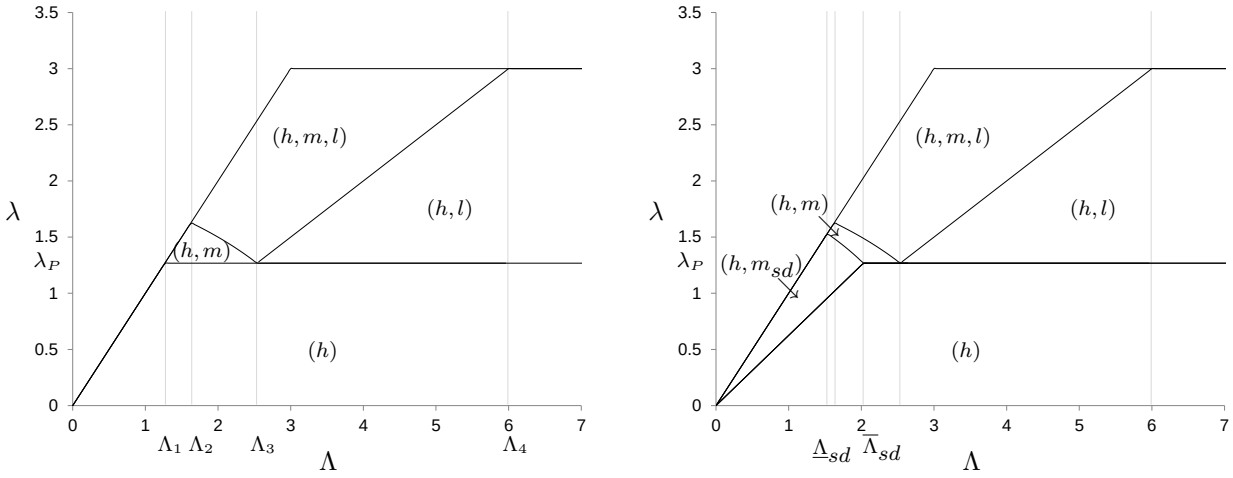
Figure 2 **Example 2. Optimal menu with low end of delay cost spectrum strategically delayed.**



(a) Lead times $w(c)$ (b) Prices $p(c)$ (c) Full prices $p(c) + cw(c)$

Note. Optimal menu properties represented by solid lines, first-best menu properties by dashed lines. $\mu = 3$, $\Lambda = 1$, $\lambda = 1$, $v = 1$, $d = 1$ and $f(c)$ uniform with $[c_{\min}, c_{\max}] = [0.25, 1.25]$ and $f_h(c_0) = 0$ for $c_0 = 0.625$.

Figure 3 **Illustration of optimal segmentation as a function of λ and Λ.**



(a) Strictly positive virtual costs ($f_h(c) > 0$) (b) Segment of negative virtual costs ($f_h(c_{min}) < 0$)

Note. Capacity $\mu = 3$, $v = 1$, $d = 1$, and $f(c)$ uniform with $[c_{\min}, c_{\max}] = [1, 2]$ in panel (a) and $[c_{\min}, c_{\max}] = [0.25, 1.25]$ in panel (b).

Figure 2(b) shows that compared to the FB prices, optimal prices decrease for types in $[c_{\min}, c_0)$, but this revenue loss is more than offset by price increases for types $[c_0, c_{max}]$. Figure 2(c) shows again the loss in customer surplus under the optimal menu relative to the FB menu.

Optimal segmentation and lead times depending on the arrival rate and market size. Proposition 2 (see appendix) specifies the solution of Problem 2 as a function of λ and Λ. Figure 3 illustrates these results for the case where high lead time qualities can be offered ($\mu > 1/d$). It shows for each (Λ, λ) which lead time classes are sold in the optimal menu, high (h), medium with strategic delay (m_{sd}) or without (m), and low (l). For instance, for Example 1 where $\Lambda = 2$ and $\lambda = 1.75$, Figure 3(a) confirms all three quality classes (h, m, l) are sold. (Proposition 2 characterizes the market size thresholds Λ_1, Λ_2, Λ_3, Λ_4, $\underline{\Lambda}_{sd}$, $\overline{\Lambda}_{sd}$ shown in Figure 3.)

Pricing out the middle. Figure 3 illustrates the conditions for optimally pricing out some intermediate types while selling low and high lead time qualities to the most patient and the most impatient types, respectively. Such a menu is optimal if the market size Λ is sufficiently large, and the arrival rate is smaller than the market size ($\lambda < \Lambda$) but too large to serve all types in the high and/or medium quality segments. In Figure 3 this holds for $\lambda < \Lambda$ in the regions labeled (h,l) and (h,m,l).

Pooling at $w = d$. Figure 3(a) illustrates the conditions for optimal pooling in the case of positive virtual delay costs f_h (which rules out optimal strategic delay). At low arrival rates $\lambda \leq \lambda_P = \mu - \sqrt{\mu/d} = 1.27$, pooling is not optimal because capacity is sufficient to admit all customers into the high lead-time quality segment (h). For higher arrival rates, $\lambda > \lambda_P$, there is a threshold on the market size Λ (which depends on λ), such that for Λ below this threshold some customers are pooled (the regions labeled (h,m) and (h,m,l)), whereas for Λ above this threshold all customers are strictly prioritized (the region labeled (h,l)). This structure follows because for fixed λ, increasing Λ increases the mass of each type so that the difference between the types in the high vs. low-quality segments grows larger. Therefore, for sufficiently large market size, the marginal types of the low- and high-quality segments under the FB menu are sufficiently far apart that their virtual delay costs are non-decreasing (i.e., $f_l(c_l') \leq f_h(c_h')$), and so pooling is suboptimal by Lemma 1.

Strategic delay. Figure 3(b) illustrates the conditions for optimal strategic delay in the case where the most patient types have negative virtual delay costs f_h. Specifically, $f_h(c) < 0$ for the types $c \in [c_{\min} = 0.25, c_0 = 0.625]$, and serving them with lead time $w < d$ is suboptimal as shown in Lemma 1 and in Example 2. Strategic delay is optimal in the region of Figure 3(b) labeled (h, m_{sd}). This region reflects two conditions for optimal strategic delay. First, the arrival rate λ must be sufficiently large relative to the market size Λ, so some types with negative virtual delay costs are served. Second, λ must be sufficiently small relative to the capacity μ, so these types' "natural delays" under work conserving scheduling (FIFO, but with lower priority than types with positive virtual delay costs) are insufficient to increase their lead times up to the medium lead time d.

5. Optimal Arrival Rate, Segmentation, and Lead Times

5.1. Segmentation and Lead Times for Fixed Capacity

We turn to STEP 3 of the solution approach outlined in Section 2.2 and discuss the optimal segmentation and lead time menu at the *optimal* λ, building on Proposition 2 and the following result. (We write g_x and g_{xy} for the first- and second-order partial derivatives of a bivariate function $g(x,y)$.)

LEMMA 2. *Fix the market size Λ. Write $\Pi^*(\lambda,\mu)$ for the revenue function under the optimal segmentation and lead times.*

1. *(a) $\Pi^*_\lambda(\lambda,\mu) \geq v$ for every (λ,μ) if the optimal segmentation is (h) and $\Pi^*_\lambda(\lambda,\mu) = v$ if it is (h, m_{sd}). (b) $\Pi^*_{\lambda\lambda}(\lambda,\mu) \leq 0 \leq \Pi^*_{\lambda\mu}(\lambda,\mu)$ for every (λ,μ) if the optimal segmentation is (h) or (h, m_{sd}), and $\Pi^*_{\lambda\lambda}(\lambda,\mu) < 0 < \Pi^*_{\lambda\mu}(\lambda,\mu)$ if it is $(l), (h,m), (h,l), (h,m,l)$ or (m,l).*
2. *(a) $\Pi^*_\lambda(\lambda,\mu)$ is continuous in (λ,μ). (b) The partial derivatives $\Pi^*_{\lambda\lambda}(\lambda,\mu)$, $\Pi^*_{\lambda\mu}(\lambda,\mu)$ are continuous in (λ,μ) under each optimal segmentation; they are also continuous at every (λ,μ) where a transition takes place between two of the optimal segmentations $(l), (h,m), (h,l), (h,m,l), (m,l)$.*

For fixed λ the optimal segmentation and lead time menu do *not* depend on the base value v. The base value v scales the valuations $V(c) = v + c \cdot d$, so the profitability of all types increases in v. Write $\lambda^*(v)$ for the optimal arrival rate as a function of v. Lemma 2 implies that $\lambda^*(v)$ increases in v with $\lambda^*(v) \to \min(\mu, \Lambda)$ as $v \to \infty$. Furthermore, if $v > 0$ then selling h classes either

exclusively or together with the strategically delayed medium lead time class (the segmentations (h) and (h, m_{sd}), respectively) can only be optimal if the entire market is served, i.e., $\lambda^*(v) = \Lambda$: Under these segmentations each additional customer increases revenue by at least v. These properties imply the following result on optimal pooling and strategic delay as a function of market size.

THEOREM 1. *Fix a capacity $\mu > 0$ and assume that $f_l' > 0$ and $f_h' > 0$.*

1. *If $d \le \mu^{-1}$ or $v \le 0$, then pricing out the middle of the delay cost spectrum, pooling, and strategic delay are not optimal.*
2. *If $d > \mu^{-1}$ and $v > 0$, then there are market size thresholds $\Lambda_1 < \Lambda_2 < \Lambda_3 < \Lambda_4$ and $\underline{\Lambda}_{sd} < \Lambda_3$ such that pooling and strategic delay are optimal as follows.*

$f_h(c_{\min}) \ge 0$		$f_h(c_{\min}) < 0$	
$\Lambda \le \Lambda_1$	*no pooling, only h class*	$\Lambda < \underline{\Lambda}_{sd}$	*strategic delay with pooling iff $v > 0$*
$\Lambda \in (\Lambda_1, \Lambda_3)$	*pooling iff $v > 0$*	$\Lambda \in [\underline{\Lambda}_{sd}, \Lambda_3)$	*pooling iff $v > 0$*
$\Lambda \in (\Lambda_3, \Lambda_4)$	*pooling iff v sufficiently large*	$\Lambda \in (\Lambda_3, \Lambda_4)$	*pooling iff v sufficiently large*
$\Lambda \ge \Lambda_4$	*no pooling, h and l classes*	$\Lambda \ge \Lambda_4$	*no pooling, h and l classes*

In the case of nonnegative virtual delay costs ($f_h(c_{min}) \ge 0$), pooling is optimal only if the market size is in some intermediate range (Λ_1, Λ_4): At smaller market sizes all customers are served with strict priorities in h classes, at larger market sizes it is optimal to sell only h and l classes and price out the middle, because there are enough profitable patient and impatient types with sufficiently different virtual delay costs. In the presence of patient customers with negative virtual delay costs ($f_h(c_{min}) < 0$), the results are the same for market sizes larger than $\underline{\Lambda}_{sd}$, but strategic delay with pooling is optimal if the market size is smaller than $\underline{\Lambda}_{sd}$: In this case there is more than enough capacity to serve all customers with lead times $w \le d$, such that types with positive virtual delay costs are strictly prioritized and all others are pooled with $w = d$.

5.2. Impact of Capacity

Theorem 2 specifies how pricing out the middle, pooling, and strategic delay, depend on capacity.

THEOREM 2. *Let $\lambda^*(\mu)$ be the optimal arrival rate as a function of capacity. If $f_l', f_h' > 0$, $d > 0$ and $v > 0$, the optimal segmentation and lead times are as follows. Define the capacity thresholds*

$$\mu_{\min} \triangleq \frac{1}{d + v/c_{\min}} < \frac{1}{d} < \mu_H \triangleq \Lambda + \frac{1 + \sqrt{1 + 4d\Lambda}}{2d}. \tag{22}$$

If $f_h(c_{\min}) < 0$, let the strategic delay threshold μ_{SD} be the unique solution in $\mu \in (\Lambda + d^{-1}, \mu_H)$ of

$$\mu - \frac{\mu/d}{\mu - \Lambda} = \Lambda \overline{F}\left(f_h^{-1}(0)\right). \tag{23}$$

1. *Pricing out the middle of the delay cost spectrum is optimal iff $\mu \in (d^{-1}, \mu_A)$, where μ_A is the market coverage threshold. The optimal arrival rate $\lambda^*(\mu)$ is strictly increasing on $[\mu_{\min}, \mu_A]$ where $\mu_A > \mu_{\min}$ and $\lambda^*(\mu) = \Lambda \Leftrightarrow \mu \ge \mu_A$. If $f_h(c_{\min}) \ge 0$ then $\mu_A < \mu_H$, and if $f_h(c_{\min}) < 0$ then $\mu_A < \mu_{SD} < \mu_H$.*
2. *Pooling and strategic delay. There is a unique capacity threshold μ_P such that:*
   (a) if $f_h(c_{\min}) \ge 0$ then $\mu_P \in [d^{-1}, \mu_H)$, pooling is optimal iff $\mu \in (\mu_P, \mu_H)$, and strategic delay is suboptimal;
   (b) if $f_h(c_{\min}) < 0$ then $\mu_P \in [d^{-1}, \mu_{SD})$, pooling is optimal iff $\mu > \mu_P$, and strategic delay is optimal iff $\mu > \mu_{SD}$.

Table 1 Pooling at lead times $w > d$ or $w < d$ is optimal for nonmonotone virtual delay costs

Lead time	Necessary Conditions			Sufficient Conditions
$w < d$	$v > -dc_{\max}, d > 0$	$f_h' \not> 0$	$\mu > \max\left(\frac{1}{d}, \frac{1}{d+v/c_{\max}}\right)$	$\mu > \mu_H, v > 0$, $\exists c$ s.t. $f_h'(c) < 0 < f_h(c)$
$w > d$	$v > 0, d > \frac{-v}{c_{min}}$	$f_l' \not> 0$	$\mu \in (\mu_{min}, \mu_H)$ if $d > 0$ or $\mu > \mu_{min}$ if $d \leq 0$	$\mu > \Lambda$, v large, $d \leq 0$, $f_l' \not> 0$

3. *Effect of base value v. There are thresholds $v_P < v_A$, such that pooling is optimal for all $\mu \in (d^{-1}, \mu_H)$ iff $v \geq v_P$, and serving the entire market is optimal for all $\mu \geq d^{-1}$ iff $v \geq v_A$.*

Pricing out the middle can be optimal only for capacity in the range (d^{-1}, μ_H). In this range the provider can sell high lead time qualities to some but not all customers. For fixed capacity, increasing the arrival rate by selling service to more intermediate types generates additional revenue, but it also slows down the service that can be provided to more patient customers which reduces the revenue from these customers. The gain from additional intermediate types increases in their valuations, the loss on the more patient customers decreases in the capacity level, giving rise to two cases (Parts 1 and 3 of Theorem 2). If valuations are below a threshold ($v < v_A$) then pricing out the middle of the delay cost spectrum is optimal in the capacity range (d^{-1}, μ_A), where $\mu_A < \mu_H$: Under these conditions the revenue gain from additional intermediate types is insufficient to offset the loss on more patient customers, so market coverage is not optimal, but capacity is still sufficient to profitably sell both low and high lead time qualities. In contrast, if valuations are sufficiently high ($v \geq v_A$), then at every capacity in the range (d^{-1}, μ_H) the revenue from additional intermediate types dominates the congestion-driven revenue losses on more patient types, so market coverage is optimal (and $\mu_A \leq d^{-1}$).

Pooling at $w = d$ is optimal in the intermediate capacity range (μ_P, μ_H). In this range capacity is insufficient to serve all customers with high quality, but still sufficient to profitably serve such large low- and high-quality segments that their marginal types c_l and c_h are similar enough to be served in a single class (see Example 1).

Strategic delay is optimal for sufficiently large capacities ($\mu > \mu_{SD}$) if and only if the most patient types have negative virtual delay costs ($f_h(c_{\min}) < 0$).

6. Connections and Extensions

6.1. Pooling with Increasing vs. Non-monotone Virtual Delay Costs

We call a delay cost distribution f *irregular* if it yields nonmonotone f_h and/or f_l. In this case the conditions of Theorem 2 for pricing out the middle, and for pooling and strategic delay at the threshold lead time d, remain structurally the same. However, pooling may also be optimal within the high- and low-quality segments, and the thresholds μ_P and μ_{SD} may change.

Conditions for pooling at $w \neq d$. Table 1 lists necessary and sufficient conditions for pooling at lead times $w \neq d$. To summarize, for pooling with lead time smaller than d (greater than d), a type c where $f_h'(c) < 0$ ($f_l'(c) < 0$) must be admitted into the high-quality (low-quality) segment. Therefore, pooling at $w < d$ is optimal if there is enough capacity to serve everyone with lead time shorter than d ($d > 0$ and $\mu > \mu_H$), all types are profitable ($v > 0$), and some types c should be pooled without strategic delay ($f_h(c) > 0 > f_h'(c)$). Similarly, pooling at $w > d$ is optimal if there is enough capacity to serve everyone ($\mu > \Lambda$), only lead times longer than d can be offered ($d \leq 0$), all types are profitable (v large), and some types c should be pooled ($f_l'(c) < 0$).

Contrasting pooling at $w = d$ with pooling at $w \neq d$. Pooling at $w = d$ arises only if it is profitable to offer some lead times shorter than *and* some longer than d. This requires $v > 0$, $d > 0$, and

intermediate capacity levels – neither too low, nor too high (Theorem 2). In this case, two disjoint type intervals are served, with different virtual delay cost functions: For each type c in the low-quality (high-quality) segment, the external price effect of its virtual delay cost, $f_l(c)$ ($f_h(c)$), is positive (negative); that is, reducing its lead time increases (decreases) the prices for more patient (impatient) types. Therefore, $f_l(c) > f_h(c)$ holds for *every* delay cost distribution, and as discussed in Sections 4-5.1, pooling at the threshold lead time is optimal if the low- and high-quality segments are sufficiently close. Pooling at $w \neq d$ differs in two ways from pooling at $w = d$: First, it can also arise if it is profitable to only serve a single interval of types, all of them with lead times either shorter or longer than d. This only requires $v > 0$ or $d > 0$, but not both. Second, because all types within a segment share the same virtual delay cost function (f_l or f_h), pooling can only be optimal if the relevant virtual delay cost function is nonmonotone.

To summarize, pooling at $w = d$ can occur for *every* delay cost distribution but only at intermediate capacity. Pooling at $w \neq d$ can only occur for an irregular delay cost distribution but merely requires enough capacity so that some types with nonmonotone virtual delay cost are profitable.

6.2. Special Cases with Lead-Time-Independent Ranking of Types

Net values satisfy $N(c,w) = V(c) - cw = v + c(d-w)$. If $v > 0$ and $d > 0$ then both lead times $w < d$ and $w > d$ may be profitable, which implies a lead-time-dependent ranking of types: Their net values increase in c for $w < d$ and decrease in c for $w > d$. This property gives rise to pricing out the middle of the delay cost spectrum, work conserving pooling at $w = d$, and strategic delay. We discuss three parameter regimes that give rise to a lead-time-independent ranking of types.

Increasing V(c)/c ratio: $v \leq 0$. If $v \leq 0 < d$ then only lead times shorter than d are profitable. To contrast this case with Theorem 2 we state the following intuitive Lemma without proof.

LEMMA 3. *If $v \leq 0 < d$ then the following holds. (1) Pricing out the middle of the delay cost spectrum, pooling at $w = d$, and strategic delay are not profitable. (2) The optimal arrival rate $\lambda^*(\mu)$ increases in μ, but $\lim_{\mu\to\infty} \lambda^*(\mu) = \Lambda$ if and only if $f_h(c_{\min}) \geq |v|/d$.*

Part 1 of Lemma 3 follows because only lead times $w < d$ are profitable for $v \leq 0$. The ranking of customer types is invariant for such lead times, i.e., their net values are increasing in impatience because $N_c(c,w) = d - w > 0$. Therefore, the set of types served is contiguous and includes the most impatient ones; this also precludes pooling at the threshold lead time. Strategic delay is not profitable because types c with negative virtual delay costs $f_h(c) < 0$ are willing to pay at most $v \leq 0$, so the provider gains (does not lose) by not serving them. Therefore, with $v \leq 0 < d$, the *only* nonstandard solution feature is pooling at lead times $w < d$, which is optimal only if $f_h' \not> 0$. Part 2 of Lemma 3 follows by noting that a type c with positive virtual delay costs contributes $v + f_h(c)(d - w(c))$ to total revenues; as capacity gets large, lead times go to zero, so the condition in the lemma is required for the most patient type to have a positive revenue contribution.

With $v \leq 0 < d$ our model specializes to that for priority auctions in Afèche and Mendelson (2004)[3], and it is essentially equivalent to the model of Katta and Sethuraman (2005), henceforth KS. Nazerzadeh and Randhawa (2014) consider essentially the same demand model as Afèche and Mendelson (2004), but they focus on the asymptotic performance of menus that offer only two

[3] Afèche and Mendelson (2004) fix $\mu = 1$. They consider i.i.d. valuations x with continuous c.d.f. $\Phi(x)$ for $x \in [\underline{v}, \overline{v}]$ and perfectly correlated delay costs, given by $x\underline{d} + \underline{c}$ for $x \in [\underline{v}, \overline{v}]$, where $\underline{d} \in (0, \mu)$ and $\underline{c} \geq 0$ are constants. See Assumptions 1, 4 and 5. A simple change of variable yields $v = -\underline{c}/\underline{d} \leq 0$, $d = 1/\underline{d} > 1/\mu$, and $F(c) = \Phi(v + cd)$ for $c \in [c_{min}, c_{max}]$ where $c_{min} = \underline{v}\underline{d} + \underline{c}$ and $c_{max} = \overline{v}\underline{d} + \underline{c}$. To avoid confusion we use here $\underline{c}$, $\underline{d}$, and x for their c, d, and v, respectively.

classes, for systems with large potential demand and capacity. Afèche and Mendelson (2004) restrict attention to work conserving strict priority policies. Our analysis shows this is without loss of optimality in their model. Strategic delay is not optimal since $v \leq 0$. Pooling at lead times $w < d$ is not optimal because they assume $f_h' > 0$ (which is equivalent to their assumption that the function $\lambda \cdot \overline{\Phi}^{-1}(\lambda/\Lambda)$ is strictly concave in λ, where Φ is the c.d.f. of valuations). KS mainly analyze a discrete-type version of the model of Afèche and Mendelson (2004), but do not restrict the scheduling policy. However, they restrict attention to the case where the valuation-delay cost ratio increases in the delay cost, that is, $v_{i+1}/c_{i+1} < v_i/c_i$ where $c_{i+1} < c_i$. Although KS do not assume an affine (v_i, c_i)-relationship, their model is in essence equivalent to ours with $v < 0$. In our model, the ratio $V(c)/c = v/c + d$ increases in c if and only if $v \leq 0$. Models with increasing $V(c)/c$ ratio but nonaffine $V(c)$, such as the one of KS, yield the *same* fundamental properties as ours with $v \leq 0$: Because $V(c)/c$ is increasing if and only if $V'(c) > V(c)/c$, and $N(c,w) > 0 \Leftrightarrow V(c)/c > w$, a lead time w is profitable for type c only if $w < V'(c)$. The ranking of types is invariant for such lead times in that $N_c(c,w) = V'(c) - w > 0$. In particular, Lemma 3 applies, and pooling can be optimal only at lead times $w < V'(c)$ and if $f_h' \not> 0$. KS analyze a segmentation algorithm that computes the optimal lead times for their discrete-type model and yields pooling only if the discrete version of our virtual delay cost f_h is not monotone increasing.

Identical valuations or negative value-delay cost correlation: $d \leq 0$. If $v > 0 \geq d$ then net values decrease in c for all feasible lead times. Therefore, only lead times $w > d$ are offered, the set of types served is contiguous and includes the most patient ones, and there is no pooling at the threshold lead time. Strategic delay is not profitable because reducing the lead times of types with $w > d$ increases the prices of more patient types ($f_l > 0$ because the external price revenue effect in the low-quality segment is positive). Therefore, pooling is optimal only at lead times $w > d$ and if $f_l' \not> 0$.

Identical delay costs: $v \to -\infty, d \to \infty$. In models where types have the *same* delay cost c but heterogenous valuations, a single class is optimal, e.g., Mendelson (1985). Such a model emerges as the limiting case of ours if one lets $[c_{\min}, c_{\max}]$ get small, so that $v \to -\infty$ and $d \to \infty$.

6.3. Strategic Delay: Log Supermodularity, Damaged Goods, Queueing Effects

Anderson and Dana (2009), henceforth AD, consider a maximum quality constraint in the standard monopoly price discrimination model with quality-independent type ranking, that is, for every quality the customer surplus is increasing in the type. AD ignore capacity and queueing effects. They show that price discrimination, i.e., offering at least two quality levels, is optimal if (i) the surplus function is log supermodular, and (ii) it is not profitable to serve the lowest types at any quality. Together these conditions imply that it is optimal to degrade quality for the lowest profitable type.

In our model, log supermodularity of the net value function is equivalent[4] to the condition $v > 0$ (or $V(c) - cV'(c) > 0$ in general). This condition, together with $d > 0$ (or $V'(c) > 0$ in general) is necessary but not sufficient for optimal strategic delay. In our model, log supermodularity implies that all types are profitable if served with lead time $w = d$. For strategic delay to be optimal, we also require $f_h(c_{\min}) < 0$, i.e., the existence of types whose quality can be lowered at a net benefit; and $\mu > \mu_{SD}$, i.e., enough capacity so that their lead times must be increased artificially rather than through queueing delays.

[4] $v > 0$ implies supermodularity: let lead times be a function $w(q)$ of quality $q \in [\underline{q}, \overline{q}]$, with $w(\overline{q}) = 0$ and $w'(q) < 0$. Then the net value $N(c, w(q))$ is log supermodular in (c, q) iff $v > 0$.

To highlight the connection to AD, consider ample capacity ($\mu = \infty$). The revenue contribution of the lowest type is $v + f_h(c_{\min})(d - w)$. Given $v > 0$ and $d > 0$, condition (ii) in AD implies $v + f_h(c_{\min})d < 0$ for $w = 0$, which implies $f_h(c_{\min}) < 0$. Therefore, condition (ii) in AD is stronger, but more generally applicable than $f_h(c_{\min}) < 0$ in our model. The model and results of AD apply to the damaged goods literature (cf. Deneckere and McAfee 1996; McAfee 2007; Section 4.3 in AD). These papers assume a nonincreasing quality cost, as in our model, but ignore capacity constraints. At ample capacity, the solution in our model resembles the damaged goods solution for a model with zero costs: A high quality segment $[c_0, c_{\max}]$ pays a high price $V(c_0)$ for the lead time $w(c) = 0$, and a low quality segment $[c_{\min}, c_0]$ pays v for the 'damaged lead time' $d > 0$.

However, queueing effects play a significant role for strategic delay: Queueing increases the minimum capacity threshold for optimal strategic delay. In our model this threshold exceeds the market size, that is, $\mu_{SD} > \Lambda$ by (23). In a damaged goods model without queueing, this threshold is smaller than the market size. To see this, drop the work conservation constraints (3) in Problem 1 to eliminate queueing. Let λ_0 be the rate of customers with nonnegative virtual delay cost f_h, where $\lambda_0 < \Lambda$ if $f_h(c_{\min}) < 0$. Then for $v > 0$, $d > 0$, strategic delay is optimal if and only if $\mu > \lambda_0$, where $\lambda_0 < \mu_{SD}$.

6.4. Non-Affine Value-Delay Cost Relationship V(c)

The affine value-delay cost relationship $V(c) = v + c \cdot d$ yields the simplest type model with multiple delay costs and lead-time-dependent ranking. It has the restriction that all types value the threshold lead-time d equally. However, our key results hold for a broader class of $V(c)$ functions, and our analysis sheds light on the underlying demand and capacity conditions. We discuss these conditions for strictly convex and strictly concave $V(c)$. In both cases, types differ in their net value for all lead-times. The discussion draws on two properties of any IC menu (see Proposition 1): The utility changes at the rate $U'(c) = V'(c) - w(c)$ and lead times decrease in impatience ($w'(c) \leq 0$).

For convex $V(c)$ it may be optimal to price out and/or pool intermediate types, as in the affine case, in both cases because $U(c)$ is convex. Specifically, for increasing $V(c)$ and intermediate capacity level, intermediate types may be the least profitable customers. In this case $U(c)$ is decreasing-increasing, so two disjoint intervals of types are served with positive utility, including the most and the least patient types. Different virtual delay cost functions, f_l and f_h, apply to these disjoint segments, hence pooling improves revenue if they are close enough, i.e., $f_l(c_l) > f_h(c_h)$. However, the threshold lead time now depends on $V(c)$, $f(c)$, and μ.

For concave $V(c)$ the segmentation may resemble or reverse the structure of the affine case, but our pooling result does not hold. The segmentation depends on the concavity of $V(c)$. For sufficiently low levels of concavity, $U(c)$ may be convex, with intermediate types priced out as in the affine case. For sufficiently concave $V(c)$, the intermediate types may be the most profitable and $U(c)$ will be concave. Contrary to the affine model, in this case it may be optimal to serve *only* intermediate types and price out both the more and the less patient types. For concave $V(c)$ pooling is suboptimal with monotone virtual delay costs, because it results in concave $U(c)$ for the pooled types. That is, unlike for (strictly) convex $V(c)$, the pooled types have positive utility *and* form a single interval (with the same virtual delay cost function). Therefore the provider can extract more surplus from these types by differentiating their lead times.

As outlined in Section 6.3, strategic delay can be optimal for every $V(c)$ function that is increasing and log supermodular, provided that $f_h(c_{\min}) < 0$ and there is enough capacity.

Finally, the menu-design problem for two-dimensional types with an *unrestricted* valuation-delay cost distribution remains open. The difficulty arises because (1) the relationship between segmentation structure and model parameters is even more intricate than under perfect correlation, and (2) the number of IC constraints is quadratic in the number of types, and local IC between neighboring types in the same segment does not ensure global IC across segments. Mechanism design problems with unrestricted multi-dimensional type distributions are notoriously cumbersome, even in the absence of queueing constraints, cf. Rochet and Choné (1998).

References

[1] Afèche, P., H. Mendelson. 2004. Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Man. Sci.* 50(7) 869-882.

[2] Afèche, P. 2004. Incentive-compatible revenue management in queueing systems: optimal strategic delay and other delay tactics. Working paper, University of Toronto.

[3] Afèche, P. 2013. Incentive-compatible revenue management in queueing systems: optimal strategic delay. *M&SOM* 15(3) 423-443.

[4] Anderson, E.T., J.D. Dana. 2009. When is price discrimination profitable? *Man. Sci.* 55(6) 980-989.

[5] Ata, B., M. Olsen. 2013. Congestion-based leadtime quotation and pricing for revenue maximization with heterogeneous customers. Queueing Systems 73(1) 35-78.

[6] Bagnoli, M., T. Bergstrom. 2005. Log-concave probability and its applications. *Econ. Theory* 26(2) 445-469.

[7] Bansal, M., C. Maglaras. 2009. Product design in a market with satisficing customers. In *Consumer-Driven Demand and Operations Models*, 37-62. Eds. S. Netessine, C.S. Tang. Springer, New York.

[8] Boyaci, T., S. Ray. 2003. Product differentiation and capacity cost interaction in time and price sensitive markets. *M&SOM* 5(1) 18-36.

[9] Çelik, S., C. Maglaras. 2008. Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. *Management Sci.* 54(6) 1132-1146.

[10] Coffman, E.G. Jr., I. Mitrani. 1980. A characterization of waiting time performance realizable by single-server queues. *Oper. Res.* 28 (3) 810-821.

[11] Dana, J.D., T. Yahalom. 2008. Price discrimination with a resource constraint. *Ec.Lett.* 100(3) 330-332.

[12] Deneckere, R.J., R.P. McAfee. Damaged goods. 1996. *J. of Econ. & Man. Strategy* 5(2) 149-174.

[13] Hassin R., M. Haviv. 2003. *To Queue or not to Queue*. Kluwer, Boston MA.

[14] Katta, A., J. Sethuraman, J. 2005. Pricing strategies and service differentiation in queues - a profit maximization perspective. Working paper, Columbia University.

[15] Maglaras, C., A, Zeevi. 2005. Pricing and design of differentiated services: approximate analysis and structural insights. *Oper. Res.* 53(2) 242-262.

[16] Maglaras C., J. Yao, A. Zeevi. 2015. Optimal price and delay differentiation in queueing systems. *Management Sci.* in press.

[17] McAfee, R.P. 2007. Pricing damaged goods. *Economics: The Open-Access Open-Assessment E-Journal* 1 (2007-1) 1-19.

[18] Mendelson, H. 1985. Pricing computer services: queueing effects. *Comm. ACM* 28(3) 312-321.

[19] Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Oper. Res.* 38(5) 870-883.

[20] Mussa, M., S. Rosen. 1978. Monopoly and product quality.*Journal of Economic Theory* 18(2) 301-317.

[21] Myerson, R.B. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47(1) 61-74.

[22] Naor, P. 1969. On the Regulation of queue size by levying tolls. *Econometrica* 37(1) 15-24.

[23] Nazerzadeh, H., R. Randhawa. 2014. Asymptotic optimality of two service grades for customer differentiation in queueing systems. Working paper, University of Southern California.

[24] Plambeck, E. 2004. Optimal leadtime differentiation via diffusion approximations. *Oper. Res.* 52(2) 213-228.

[25] Rao, S., E.R. Petersen. 1998. Optimal pricing of priority services. *Oper. Res.* 46(1) 46-56.

[26] Rochet, J., P. Choné. 1998. Ironing, sweeping, and multidimensional screening. *Econometrica* 66(4) 783-826.

[27] Rochet, J., L. Stole. 2003. The economics of multidimensional screening. In *Advances in Economics and Econometrics: Theory and Applications*, 150-198. Eds. M. Dewatripont, L. P. Hansen and S. J. Turnovsky. Cambridge University Press, Cambridge.

[28] Stidham, S. Jr. 2002. Analysis, design and control of queueing systems. *Oper. Res.* 50(1) 197-216.

[29] Stidham, S. Jr. 2009. Optimal design of queueing systems. CRC, Boca Raton FL.

[30] Van Mieghem, J.A. 2000. Price and service discrimination in queuing systems: incentive compatibility of G$c\mu$ scheduling. *Man. Sci.* 46(9) 1249-1267.

[31] Zhao, X., K.E. Stecke, A. Prasad. 2012. Lead time and price quotation mode selection: Uniform or differentiated? *POMS* 21(1) 177-193.

Appendix: Proofs

The table below summarizes the notation and indicates as applicable where it is introduced.

Λ, μ	market size (maximum arrival rate), capacity
c, v, d	delay cost (customer type), base value, threshold lead time
$c_{\min}$, $c_{\max}$	minimum and maximum delay cost
$F(c)$, $\bar{F}(c)$, $f(c)$	delay cost CDF, its complement and PDF
$a(c)$	acceptance function indicating whether type c buys service or not
$\mathcal{C}_a$, $\bar{\mathcal{C}}_a$	sets of customers who respectively buy and do not buy service
$p(c)$, $w(c)$, $U(c)$	price, expected lead time, and expected utility of type c
Π	revenue rate
λ	arrival rate of customers to all classes
λ_l, λ_m, λ_h	arrival rates to l, m, h classes
$f_l(c)$, $f_h(c)$	virtual delay cost functions of customers in C_l and C_h

Proposition 1

C_l, C_m, C_h	sets of customers who buy low (l), medium (m), high (h) classes
c_l, c_h	marginal customer types of segments buying l and h classes
$U(c';c)$	expected utility of type c who reports type c'
U_m	expected utility of a type c who buys the m class

Lemma 1

$D(\lambda_l, \lambda_h, w)$	virtual delay cost rate
λ_l^*, λ_m^*, λ_h^*, w^*	optimal arrival rates to l, m, h classes, and optimal lead times, for fixed λ
c_l^*, c_h^*	marginal types under optimal segmentation for fixed λ
C_l^*, C_h^*	sets of customers who buy l and h classes under optimal segmentation for fixed λ

Proposition 2

l, m, m_{sd}, h	indicator of positive arrival rate to classes l, m, m with strategic delay, and h
$\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_4, \underline{\Lambda}_{ml}, \bar{\Lambda}_{ml}$	market size thresholds for pooling
$\underline{\Lambda}_{sd}$, $\bar{\Lambda}_{sd}$	market size thresholds for strategic delay
λ_{mh}, $\lambda^*_{mh}(\lambda)$	total arrival rate to m and h classes, and optimal λ_{mh} as function of λ
$\Pi(\lambda_{mh}, \lambda_h)$	revenue rate as function of arrival rates λ_{mh} and λ_h
λ_F, λ_P	upper bounds, on arrival rate λ_{mh} under FIFO, and on λ_h under strict priorities
$\bar{\lambda}_h(\lambda_{mh})$, $\lambda_h(\lambda_{mh})$	maximum feasible λ_h, and optimal λ_h, as function of λ_{mh}
$c_0 = f_h^{-1}(0), \lambda_0 = \Lambda \bar{F}(c_0)$	type with zero virtual delay cost, arrival rate of higher types to h classes
$g(\lambda, \lambda_{mh})$	difference between virtual delay costs of marginal types c_l and c_h if arrival rate to h classes is $\bar{\lambda}_h(\lambda_{mh})$
λ_1, λ_2, λ_3, λ_{sd}	arrival rate thresholds (Lemmas 5, 6, and 8)

Lemma 2

$\Pi^*(\lambda, \mu)$	optimal revenue rate as function of arrival rate λ and capacity μ

Theorem 2

$\lambda^*(\mu)$	optimal arrival rate as function of capacity μ
$\mu_{\min}$	upper capacity threshold for serving all customers with l classes
μ_A, μ_P, μ_{SD}, μ_H	lower capacity thresholds for market coverage, pooling, strategic delay, and serving all customers with h classes
v_A, v_P	lower base value thresholds for market coverage, pooling over entire applicable capacity range

Proof of Proposition 1

The constraints (4)-(6) imply Parts 1-5:

Write $U(c';c)$ for the expected utility of a type c who reports type c', and $U(c) \triangleq U(c;c)$, where

$$U(c';c) \triangleq U(w(c'), p(c'); c) = v + c\,(d - w(c')) - p(c') = U(c') + (c - c')\,(d - w(c'))\,. \tag{24}$$

The IC constraints (6) require that the expected utilities from service satisfy:

$$U(c) = v + c\,(d - w(c)) - p(c) \geq U(c';c) \Leftrightarrow c \cdot w\,(c) + p\,(c) \leq c \cdot w\,(c') + p\,(c') \text{ for } \forall c \neq c'. \tag{25}$$

Similarly, we must have $U(c') \geq U(c;c')$, so from (24) the IC constraints (6) are equivalent to

$$(c - c')\,(d - w(c)) \geq U(c) - U(c') \geq (c - c')\,(d - w(c')) \text{ for } \forall c \neq c'. \tag{26}$$

It follows from (26) that the expected utility from service $U(c)$ is continuous in the type.

Part 1. It follows from (25)-(26) that $w(c)$ is nonincreasing and $p\,(c)$ is nondecreasing in c. If $c < c'$ (26) implies $w(c') \leq w(c)$, and (25) implies $p\,(c') - p\,(c) \geq c\,(w\,(c) - w\,(c')) \geq 0$. Because w is nonincreasing in c, it follows that $c_l \leq c_h$, and that w is Riemann integrable. Therefore (26) implies

$$U(c'') - U(c') = \int_{c'}^{c''} (d - w(x))dx \text{ for all } c' < c''. \tag{27}$$

Part 2. We first show (a). The case $C_l = \{c_{\min}\}$ is trivial. Suppose that $C_l \neq \varnothing$ and $c_l > c_{\min}$. Fix $c \in [c_{\min}, c_l)$. We show that $c \in C_l$. Apply (27) with $c = c'$ and $c_l = c''$ to get

$$U(c) = U(c_l) + \int_c^{c_l} (w(x) - d)dx > U(c_l) \geq 0. \tag{28}$$

The first (strict) inequality follows since $w(x) > d$ for $x < c_l$: otherwise, if $w(x) \leq d$ for some $x < c_l$ then $w(x') \leq w(x) \leq d$ for $x' > x$ since $w(c)$ is nonincreasing, contradicting that $c_l = \sup C_l$. That $U(c_l) \geq 0$ follows since $U(c)$ is continuous in c: if $U(c_l) < 0$ then by continuity it must be that $U(c) < 0$ for all c in some interval $[c_s, c_l]$ and the IR constraint (4) for $c \in [c_s, c_l]$ can only hold if $c \notin C_l$. But this contradicts that $c_l = \sup C_l$, so $U(c_l) \geq 0$. Therefore, $U(c) > 0$ for $c < c_l$; the IR constraint (5) for c only holds if $c \in \mathcal{C}_a$, and since $w(c) > d$ for $c < c_l$ it follows that $c \in C_l$. To prove (b) it remains to show that $C_l \neq \varnothing \Rightarrow U(c_l) = 0$ (which we prove with Part 5) for then (28) reduces to (9) and the price equation (8) follows since $U(c) = v + c(d - w(c)) - p(c)$.

Part 3. Follows from the same line of argument as in the proof of Part 2, by applying (27) with $c_h = c' < c = c''$ and showing that $C_h \neq \varnothing$ implies $U(c_h) = 0$, which we prove with Part 5.

Part 4. Suppose that $C_m \neq \varnothing$. Parts 2-3 imply $C_m \subset [c_l, c_h]$. Fix $c \in C_m$. Then

$$U(c') \geq U(c; c') = U(c) = v - p(c) \geq 0 \text{ for } \forall c' \neq c. \tag{29}$$

The first inequality follows from the IC constraints (6), the equalities hold since $w(c) = d$, and the IR constraint (4) implies the second inequality. It follows from (29) that $U(c') = U(c)$ for all $c' \in C_m$. Define $U_m \triangleq U(c) = v - p(c)$ as the common utility for all $c' \in C_m$.

It remains to show that $U_m = 0$, which we prove with Part 5.

Part 5. That $(c_l, c_h) \subset C_m \cup \overline{\mathcal{C}}_a$ is immediate from the definitions of c_l and c_h. The expression (12) for $U(c)$ follows from (27). We prove that $U(c) \leq 0$ for $c \in [c_l, c_h]$ for three exhaustive cases.

(i) Not all types are served ($\overline{\mathcal{C}}_a \neq \varnothing$) but some types buy the medium lead time ($C_m \neq \varnothing$). This implies that $c_l < c_h$. Then (5) and (29) imply $0 \geq U(c') \geq U(c) = U_m \geq 0$ for any types $c \in C_m$ and $c' \in \overline{\mathcal{C}}_a$; therefore we have $U(c) = 0$ for all $c \in (c_l, c_h)$. Since $U(c)$ is continuous it follows that $U(c_l) = 0 = U(c_h)$. Since $C_m \subset [c_l, c_h]$ it follows that $U(c) = U_m = 0$ for $c \in C_m$.

(ii) Not all types are served ($\overline{\mathcal{C}}_a \neq \varnothing$) and no types buy the medium lead time ($C_m = \varnothing$). It follows that $(c_l, c_h) \subset \overline{\mathcal{C}}_a$. The IR constraints (5) and the continuity of $U(c)$ imply $U(c) \leq 0$ for $c \in [c_l, c_h]$. If $C_l \neq \varnothing$ then (28) implies $U(c_l) \geq 0$ and so $U(c_l) = 0$. Similarly, $U(c_h) = 0$ if $C_h \neq \varnothing$.

(iii) All types are served ($\overline{\mathcal{C}}_a = \varnothing$). Let $U_{\min} = \min_c U(c)$. The IR constraints (4) require $U_{\min} \geq 0$, and revenue-maximization requires $U_{\min} = 0$. The proof is complete if $U(c) = U_{\min}$ for $c \in [c_l, c_h]$. First note that $U(c) = U(c_l) = U(c_h)$ for $c \in [c_l, c_h]$. If $c_l = c_h$ this is trivial. If $c_l < c_h$ this holds since then $(c_l, c_h) \subset C_m$, Part 4 implies $U(c) = U_m$ for $c \in C_m$, and by continuity $U_m = U(c_l) = U(c_h)$. By Parts 2-3 we have $U(c) > U(c_l) = U(c_h)$ for $c \notin [c_l, c_h]$ which proves that $U(c) = U_{\min}$ for $c \in [c_l, c_h]$.

Parts 1-5 imply (4)-(6): Parts 2-5 imply the IR constraints (4)-(5). The IC constraints (6) are equivalent to (26). Substituting for $U(c)$ from Parts 2-5, (26) is equivalent to

$$(c'' - c')(d - w(c'')) \geq U(c'') - U(c') = \int_{c'}^{c''} (d - w(x))dx \geq (c'' - c')(d - w(c')) \text{ for } \forall c' < c''. \tag{30}$$

By Part 1, $w(c') \geq w(x) \geq w(c'')$ for $x \in [c', c'']$, which establishes both inequalities. ■

Proof of Lemma 1

From Problem 1, Proposition 1, and the revenue rate (13), it follows that for fixed λ, the revenue-maximization problem is equivalent to minimizing the virtual delay cost rate

$$D(\lambda_l, \lambda_h, w) \triangleq \Lambda \int_{c_{\min}}^{c_l(\lambda_l)} f(x) f_l(x) (w(x) - d)\, dx + \Lambda \int_{c_h(\lambda_h)}^{c_{\max}} f(x) f_h(x) (w(x) - d)\, dx$$

over λ_l, λ_h, and the lead time function $w : \mathcal{C} \to \mathbb{R}$, subject to $\lambda_m = \lambda - \lambda_l - \lambda_h \geq 0$, increasing $w(c)$, $w(c) > d > w(c')$ for $c < c_l(\lambda_l)$ and $c' > c_h(\lambda_h)$, and

$$\Lambda \int_c^{c_{\max}} f(x) w(x) dx \geq \frac{\Lambda \overline{F}(c)}{\mu - \Lambda \overline{F}(c)}, \qquad \forall c \in [c_h(\lambda_h), c_{\max}], \tag{31}$$

$$\Lambda \int_{x \in [c, c_l] \cup [c_h, c_{max}]} f(x) w(x) dx + \lambda_m \cdot d \geq \frac{\lambda - \Lambda F(c)}{\mu - [\lambda - \Lambda F(c)]}, \qquad \forall c \in [c_{\min}, c_l(\lambda_l)]. \tag{32}$$

Recall that f_l and f_h satisfy (14)-(15), $f_l', f_h' > 0$, and that $c_l(\lambda_l) = F^{-1}(\lambda_l/\Lambda)$ and $c_h(\lambda_h) = \bar{F}^{-1}(\lambda_h/\Lambda)$ are the marginal types corresponding to λ_l and λ_h, respectively. Suppose the scalars λ_l^*, λ_h^* and the function w^* are a solution of this problem and write $c_l^* = c_l(\lambda_l^*)$and $c_h^* = c_h(\lambda_h^*)$.

The proof hinges on three necessary optimality conditions:

c1. If $c \in (c_h^*, c_{\max}]$ then $f_h(c) > 0$.

Proof of c1. We argue by contradiction. If $f_h(c_0) = 0$ for some $c_0 \in (c_h^*, c_{\max})$, where $f_h(c_{\max}) = c_{\max} > 0$, then $f_h(c) < 0$ for $c \in [c_h^*, c_0)$ since f_h is strictly increasing, and $w^*(c) < d$ for $c \in (c_h^*, c_0]$. By inspection it is clear that we can reduce the virtual delay cost rate by perturbing the lead time function from w^* to w^o, where w^o agrees with w^* except that $w^o(c) = d$ for $c \in [c_h^*, c_0]$. Then w^o is feasible and $D(\lambda_l^*, \lambda_h^*, w^o) < D(\lambda_l^*, \lambda_h^*, w^*)$. Under the menu w^o the marginal type c_h^* moves to $c_h' = c_0 > c_h^*$ and $f_h(c) > 0$ holds for $c \in (c_h', c_{\max}]$.

c2. If the set of types with high lead time qualities is nonempty ($C_h^* \neq \varnothing$) then the constraints (31) are binding for $c \in [c_h^*, c_{\max}]$, and $w^*(c_{\max}) = 1/\mu$.

Proof of c2. This is trivial if $c_h^* = c_{\max} \in C_h^*$. Suppose that $c_h^* < c_{\max}$. If the property is not satisfied, there exists a feasible perturbation w^o of w^* which reduces the virtual delay cost rate $D(\lambda_l, \lambda_h, w)$ by lowering the lead times for types $(c_2, c_2 + \epsilon_2] \subset [c_h^*, c_{\max}]$ and by increasing the lead times for lower types $(c_1, c_1 + \epsilon_1] \subset [c_h^*, c_{\max}]$, where $\epsilon_1, \epsilon_2 > 0$ and $c_1 + \epsilon_1 \leq c_2$. This holds since $f_h(c) > 0$ for $c > c_h^*$ by *c1*, and because $f_h' > 0$.

c3. If C_l^* is nonempty, then the constraints (32) bind for $c \in [c_{\min}, c_l^*]$, and $w^*(c_{\min}) = \mu/(\mu - \lambda)^2$.

Proof of c3. This follows from a similar argument as for *c2*, because $f_l > 0$ and $f_l' > 0$.

Part 1(a). Properties *c2-c3* imply optimality of the lead times shown in (16). This is immediate for $c_h^* = c_{\max} \in C_h^*$ and/or $c_{\min} = c_l^* \in C_l^*$. If $c_h^* < c_{\max}$ then by *c2* the constraints (31) are binding for $c \in [c_h^*, c_{\max}]$. Solving the resulting integral equation in w^* yields $w^*(c) = \mu/\left(\mu - \Lambda \bar{F}(c)\right)^2$. If $c_l > c_{\min}$ the RHS of (32) satisfies

$$\frac{\lambda - \Lambda F(c)}{\mu - [\lambda - \Lambda F(c)]} = \frac{[\lambda_l - \Lambda F(c)]\,\mu}{(\mu - [\lambda - \Lambda F(c)])\,(\mu - \lambda_m - \lambda_h)} + \frac{\lambda_m \mu}{(\mu - \lambda_m - \lambda_h)\,(\mu - \lambda_h)} + \frac{\lambda_h}{\mu - \lambda_h}, c \in [c_{\min}, c_l].$$

By *c2-c3* for an optimal solution the constraints (32) therefore simplify to

$$\Lambda\int_{c}^{c_l^*} f(x)w^*(x)dx + \lambda_m^* \cdot d = \frac{\Lambda\left[F(c_l^*) - F(c)\right]\mu}{(\mu - [\lambda - \Lambda F(c)])\,(\mu - \lambda_m^* - \lambda_h^*)} + \frac{\lambda_m^* \mu}{(\mu - \lambda_m^* - \lambda_h^*)\,(\mu - \lambda_h^*)}, c \in [c_{\min}, c_l^*]. \quad (33)$$

Solving this integral equation in w^* yields $w^*(c) = \mu / \left(\mu - [\lambda - \Lambda F(c)]\right)^2$.

Part 1(b). This claim follows directly from *c2-c3*. If $C_l^* \neq \varnothing$ then (32) is binding for $c = c_{\min}$, which implies that the policy is work conserving.

Part 2(a). Follows from *c1* since f_h is continuous.

Part 2(b). Follows by substituting $w^*(c)$ from (16) in the revenue function (13) and analyzing its partial derivatives with respect to λ_l and λ_h. Refer to the proof of Proposition 2. ■

Optimal Segmentation and Lead Times Depending on λ and Λ

Proposition 2 below specifies, for fixed capacity μ, how the optimal set of lead time classes changes as the arrival rate λ increases, and how these transitions depend on the market size Λ. This result yields the optimal revenue as a function of λ, Λ, and μ, which is key for characterizing the optimal menu at the *optimal* arrival rate, depending on Λ (Theorem 1) and μ (Theorem 2).

Proposition 2 uses the following notation to describe these structural properties. The optimal set of lead time classes for a given arrival rate is denoted by a subset of the letters h, m, m_{sd}, and l, shown in parentheses. For instance (h, l) indicates that h and l classes have a strictly positive arrival rate, but the m and m_{sd} class have a zero arrival rate, in the optimal menu. The notation $(x) \to (y)$ indicates the existence of a threshold arrival rate such that the optimal set of classes changes from (x) to (y) as the arrival rate λ crosses the threshold from below.

PROPOSITION 2. *Fix a capacity $\mu > 0$ and assume that $f_l' > 0$ and $f_h' > 0$. The optimal customer segmentation and lead time menu depend as follows on the market size Λ and the arrival rate λ.*

1. *For $d \leq \mu^{-1}$ the segmentation (l) is optimal for all λ and Λ.*
2. *For $d > \mu^{-1}$ denote by $\lambda_P \triangleq \mu - \sqrt{\mu/d}$ and $\lambda_F \triangleq \mu - 1/d$ the arrival rates at which the maximum lead time equals d under work conserving priority and FIFO service, respectively.*
   (a) *If $f_h(c_{\min}) \geq 0$ and $\mu^{-1} \leq F\left(f_l^{-1}(c_{\max})\right) \cdot d$ then there are unique thresholds $\Lambda_1 < \Lambda_2 < \Lambda_3 < \Lambda_4$, where $\Lambda_1 = \lambda_P < \Lambda_2 < \lambda_F$ and $\mu < \Lambda_4$, which yield the following structure:*

Market Size	*Classes with positive rate as λ increases on $[0, \Lambda] \cap (0, \mu)$*
$\Lambda \in (0, \Lambda_1]$	(h)
$\Lambda \in (\Lambda_1, \Lambda_2]$	$(h) \to (h, m)$
$\Lambda \in (\Lambda_2, \Lambda_3)$	$(h) \to (h, m) \to (h, m, l)$
$\Lambda \in (\Lambda_3, \Lambda_4)$	$(h) \to (h, l) \to (h, m, l)$
$\Lambda \in [\Lambda_4, \infty)$	$(h) \to (h, l)$

(34)

   The optimal policy is work conserving.

   (b) *Selling only medium and low quality classes, (m, l), is optimal for some (λ, Λ) iff $F\left(f_l^{-1}(c_{\max})\right) \cdot d < \mu^{-1} < d$. If $f_h(c_{\min}) \geq 0$ and $F\left(f_l^{-1}(c_{\max})\right) \cdot d < \mu^{-1} < d$ then (34) is modified by additional thresholds $\underline{\Lambda}_{ml} < \overline{\Lambda}_{ml}$, where $\lambda_F < \underline{\Lambda}_{ml} < \mu < \overline{\Lambda}_{ml} < \Lambda_4$:*

Market Size	*Classes with positive rate as λ increases on $[0, \Lambda] \cap (0, \mu)$*
$\Lambda \in [\underline{\Lambda}_{ml}, \overline{\Lambda}_{ml})$	$(h) \to (h, m) \to (h, m, l) \to (m, l)$, *if* $\Lambda < \Lambda_3$ $(h) \to (h, l) \to (h, m, l) \to (m, l)$, *if* $\Lambda > \Lambda_3$

(35)

   For $\Lambda \notin [\underline{\Lambda}_{ml}, \overline{\Lambda}_{ml})$ the structure of (34) applies.

(c) *Strategic delay is optimal for some* (λ,Λ) *iff* $f_h(c_{\min})<0$ *and* $d>\mu^{-1}$. *If* $f_h(c_{\min})<0$ *and* $\mu^{-1}\leq F\left(f_l^{-1}\left(c_{\max}\right)\right)\cdot d$ *then (34) changes: two thresholds* $\underline{\Lambda}_{sd}<\overline{\Lambda}_{sd}$ *replace* Λ_1 *and yield the following structure, where* $\lambda_P<\underline{\Lambda}_{sd}\leq\Lambda_2$ *and* $\underline{\Lambda}_{sd}<\overline{\Lambda}_{sd}\leq\Lambda_3<\Lambda_4$*:*

Market Size	*Classes with positive rate as* λ *increases on* $[0,\Lambda]\cap(0,\mu)$
$\Lambda\in(0,\underline{\Lambda}_{sd})$	$(h)\to(h,m_{sd})$
$\Lambda\in[\underline{\Lambda}_{sd},\overline{\Lambda}_{sd})$	$(h)\to(h,m_{sd})\to(h,m)\,,$ *if* $\Lambda\leq\Lambda_2$ $(h)\to(h,m_{sd})\to(h,m)\to(h,m,l)\,,$ *if* $\Lambda>\Lambda_2$

(36)

where m_{sd} *indicates that the lead time* d *involves strategic delay.*

For $\Lambda\geq\overline{\Lambda}_{sd}$ *the optimal scheduling policy is work conserving and (34) applies.*

Proof.

Preliminaries. The optimal segmentation and lead time menu described in Proposition 2 are obtained by solving Problem 2 for fixed λ. Refer to (17)-(21).

Part 1 is trivial: In this case all lead times exceed d, so (20) requires $\lambda_l=\lambda$.

For Part 2 we reformulate Problem 2. Define the aggregate rate for the m class and the h classes,

$$\lambda_{mh}\triangleq\lambda_m+\lambda_h.$$

For fixed λ write the revenue (17) as

$$\Pi(\lambda_{mh},\lambda_h)\triangleq\lambda v-\Lambda\int_{c_{\min}}^{c_l(\lambda-\lambda_{mh})}f(x)f_l\left(x\right)\left(\frac{\mu}{\left(\mu-[\lambda-\Lambda F\left(x\right)]\right)^2}-d\right)dx+\Lambda\int_{c_h(\lambda_h)}^{c_{\max}}f(x)f_h\left(x\right)\left(d-\frac{\mu}{\left(\mu-\Lambda\overline{F}\left(x\right)\right)^2}\right)dx. \tag{37}$$

We define two threshold arrival rates. Define the maximum feasible rate for h classes,

$$\lambda_P\triangleq\mu-\sqrt{\mu/d}. \tag{38}$$

This is the maximum arrival rate of customers under work conserving priority service (optimal for h classes by Lemma 1) so that their lead times are shorter than d. Define the maximum aggregate rate for m and h classes

$$\lambda_F\triangleq\mu-1/d. \tag{39}$$

At this rate the lead time is d under work conserving FIFO service. $\mu d>1$ implies $0<\lambda_P<\lambda_F$.

Let $\overline{\lambda}_h\left(\lambda_{mh}\right)$ be the maximum feasible rate of h classes as a function of the total rate λ_{mh} to m and h classes, so that the medium lead time d is achievable, i.e., (20) holds. Then by (38)-(39):

$$\overline{\lambda}_h\left(\lambda_{mh}\right)\triangleq\min\left(\lambda_{mh},\mu-\frac{\mu/d}{(\mu-\lambda_{mh})}\right)=\begin{cases}\lambda_{mh}, & \text{if } \lambda_{mh}\in[0,\lambda_P]\,,\\ \mu-\frac{\mu/d}{\mu-\lambda_{mh}}\leq\lambda_P, & \text{if } \lambda_{mh}\in[\lambda_P,\lambda_F].\end{cases} \tag{40}$$

For $\lambda_{mh}\leq\lambda_P$ the entire λ_{mh} can be allocated to h classes, so $\overline{\lambda}_h\left(\lambda_{mh}\right)=\lambda_{mh}$ which increases on $[0,\lambda_P]$ with $\overline{\lambda}_h\left(\lambda_P\right)=\lambda_P$. For $\lambda_{mh}>\lambda_P$ only a portion $\overline{\lambda}_h\left(\lambda_{mh}\right)<\lambda_{mh}$ can be allocated to h classes, and $\overline{\lambda}_h\left(\lambda_{mh}\right)$ decreases on $[\lambda_P,\lambda_F]$, with $\overline{\lambda}_h\left(\lambda_F\right)=0$. Having $\lambda_{mh}>\lambda_F$ violates (20).

Constraint (20) in Problem 2 holds if and only if $\lambda_{mh}\leq\lambda_F$ and $\lambda_h\leq\overline{\lambda}_h\left(\lambda_{mh}\right)$. Constraint (21) holds if and only if $\lambda_P\leq\lambda_{mh}<\lambda$ or $\lambda_{mh}=\lambda\leq\lambda_P$. Problem 2 for fixed λ is thus equivalent to

$$\max_{\lambda_{mh},\lambda_h}=\Pi(\lambda_{mh},\lambda_h) \tag{41}$$

$$\text{s.t.}\quad\min\left(\lambda,\lambda_P\right)\leq\lambda_{mh}\leq\min\left(\lambda,\lambda_F\right), \tag{42}$$

$$0\leq\lambda_h\leq\overline{\lambda}_h\left(\lambda_{mh}\right). \tag{43}$$

The lower bound in (42) ensures work conservation if $\lambda_l > 0$: If $\lambda_{mh} < \min(\lambda, \lambda_P)$, strategic delay is required for l classes ($d > \mu/(\mu - \lambda_{mh})^2$ from (38)) which is suboptimal by Lemma 1.1(b).

Proof steps. We prove Part 2 by characterizing the solution of (41)-(43) in four steps:

1. For fixed λ_{mh}, we characterize the optimal rate $\lambda_h(\lambda_{mh})$.
2. For fixed λ, we derive the optimality conditions for the optimal segmentation, specifically, for the optimal rates $\lambda^*_{mh}(\lambda)$ and $\lambda^*_h(\lambda) = \lambda_h(\lambda^*_{mh}(\lambda))$. These conditions, summarized in (46) and (50), are stated in terms of the virtual delay costs of appropriately chosen customer types.
3. We translate the virtual delay cost conditions identified in Step 2 into conditions on λ and Λ. We organize this analysis into the technical Lemmas 4-8.
4. We prove Parts 2(a)-(c) of Proposition 2 by combining (46), (50), and Lemmas 4-8.

Step 1. Optimal λ_h for fixed λ_{mh}. For fixed $\lambda_{mh} \in (0, \min(\lambda, \lambda_F)]$, the optimal λ_h satisfies

$$\lambda_h(\lambda_{mh}) \triangleq \arg\left\{\max_{\lambda_h} \Pi(\lambda_{mh}, \lambda_h) \text{ s.t. } 0 \le \lambda_h \le \overline{\lambda}_h(\lambda_{mh})\right\}.$$

From (37) we have

$$\frac{\partial \Pi(\lambda_{mh}, \lambda_h)}{\partial \lambda_h} = f_h(c_h(\lambda_h))\left(d - \frac{\mu}{(\mu - \lambda_h)^2}\right), \text{ for } \lambda_h \le \overline{\lambda}_h(\lambda_{mh}),$$

where $\Lambda f(c_h(x)) c_h'(x) = -1$. The multiplier of $f_h(c_h(\lambda_h))$ is nonnegative by (38) and since $\overline{\lambda}_h(\lambda_{mh}) \le \lambda_P$ by (40). Since $f_h(c_h(0)) = c_{\max} > 0$ and $f_h' c_h' < 0$, the maximizer $\lambda_h(\lambda_{mh})$ is unique:

$$\lambda_h(\lambda_{mh}) = \begin{cases} \overline{\lambda}_h(\lambda_{mh}), & \text{if } f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) \ge 0, \\ \lambda_0 \triangleq \Lambda \overline{F}\left(f_h^{-1}(0)\right) < \overline{\lambda}_h(\lambda_{mh}), & \text{if } f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) < 0. \end{cases} \tag{44}$$

If $f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) \ge 0$ then it is optimal to sell the maximum possible rate $\overline{\lambda}_h(\lambda_{mh})$ to h classes and $\lambda_{mh} - \overline{\lambda}_h(\lambda_{mh}) \ge 0$ to the medium lead time class. This policy is work conserving.

If $f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) < 0$ then it is optimal to sell λ_0 to h classes, less than the maximum possible rate $\overline{\lambda}_h(\lambda_{mh})$, and $\lambda_{mh} - \lambda_0 > 0$ to the medium lead time d. At λ_0 defined in (44) the virtual delay cost of the corresponding marginal type is zero: $f_h(c_h(\lambda_0)) = f_h(\overline{F}^{-1}(\lambda_0/\Lambda)) = 0$. This policy is not work conserving: the lead time d involves strategic delay since $\lambda_0 < \overline{\lambda}_h(\lambda_{mh})$:

$$\frac{\mu}{(\mu - \lambda_{mh})(\mu - \lambda_0)} < \frac{\mu}{(\mu - \lambda_{mh})\left(\mu - \overline{\lambda}_h(\lambda_{mh})\right)} \le d. \tag{45}$$

Step 2. Optimal segmentation for fixed λ: virtual delay cost conditions. This step derives optimality conditions for the optimal arrival rates of h, m, and l classes, denoted by $\lambda^*_h(\lambda)$, $\lambda^*_m(\lambda)$, and $\lambda^*_l(\lambda)$. To this end, we characterize the optimal total arrival rate to h and m classes,

$$\lambda^*_{mh}(\lambda) \triangleq \arg\left\{\max_{\lambda_{mh}} \Pi(\lambda_{mh}, \lambda_h(\lambda_{mh})) \text{ s.t. } \min(\lambda, \lambda_P) \le \lambda_{mh} \le \min(\lambda, \lambda_F)\right\}.$$

Then $\lambda^*_h(\lambda)$ follows from (44) with $\lambda_{mh} = \lambda^*_{mh}(\lambda)$, namely, $\lambda^*_h(\lambda) = \lambda_h(\lambda^*_{mh}(\lambda))$. Furthermore, $\lambda^*_m(\lambda) = \lambda^*_{mh}(\lambda) - \lambda^*_h(\lambda)$ and $\lambda^*_l(\lambda) = \lambda - \lambda^*_{mh}(\lambda)$.

Optimal segmentation for $\lambda \leq \lambda_P$. For $\lambda \leq \lambda_P$ it is not optimal to sell l classes: (42) requires $\lambda_{mh} = \lambda$, so the maximizer satisfies $\lambda^*_{mh}(\lambda) = \lambda$. By (40) the entire λ can be sold to h classes, so $\overline{\lambda}_h(\lambda^*_{mh}(\lambda)) = \overline{\lambda}_h(\lambda) = \lambda$. Therefore (44) yields the following optimal segmentations, where m_{sd} denotes that the medium lead time involves strategic delay.

	Optimal segmentations for $\mu > d^{-1}$ *and* $\lambda \leq \lambda_P$				
segments	*virtual delay cost condition*	$\lambda^*_{mh}(\lambda)$	$\lambda^*_h(\lambda)$	$\lambda^*_m(\lambda)$	$\lambda^*_l(\lambda)$
(h)	$f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right) = f_h(c_h(\lambda)) \geq 0$	λ	λ	0	0
(h, m_{sd})	$f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right) = f_h(c_h(\lambda)) < 0$	λ	λ_0	$\lambda - \lambda_0 > 0$	0

(46)

Optimal segmentation for $\lambda > \lambda_P$. In this case $\lambda_{mh} \in [\lambda_P, \min(\lambda, \lambda_F)]$ by (42). Using (44) to substitute for λ_h into (37), the total derivative of the revenue with respect to λ_{mh} satisfies

$$\frac{d\Pi(\lambda_{mh}, \lambda_h(\lambda_{mh}))}{d\lambda_{mh}} = \frac{\partial\Pi(\lambda_{mh}, \lambda_h(\lambda_{mh}))}{\partial\lambda_{mh}} + \frac{\partial\Pi(\lambda_{mh}, \lambda_h(\lambda_{mh}))}{\partial\lambda_h} \cdot \lambda'_h(\lambda_{mh}) =$$
$$= f_l(c_l(\lambda - \lambda_{mh}))\left(\frac{\mu}{(\mu - \lambda_{mh})^2} - d\right) + f_h(c_h(\lambda_h(\lambda_{mh})))\left(d - \frac{\mu}{(\mu - \lambda_h(\lambda_{mh}))^2}\right)\lambda'_h(\lambda_{mh}), \quad (47)$$

where $\Lambda f(c_l(x))c'_l(x) = 1$ and $-\Lambda f(c_h(x))c'_h(x) = 1$. By (44), we have two cases: If $f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) < 0$ then $\lambda_h(\lambda_{mh}) = \lambda_0$ and $\lambda'_h(\lambda_{mh}) = 0$. If $f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) \geq 0$ then $\lambda_h(\lambda_{mh}) = \overline{\lambda}_h(\lambda_{mh}) = \mu - \mu/d(\mu - \lambda_{mh})$, the second equality holds by (40), so $\lambda'_h(\lambda_{mh}) = -\mu/d(\mu - \lambda_{mh})^2$.

Substituting for $\lambda_h(\lambda_{mh})$ and $\lambda'_h(\lambda_{mh})$ into (47) yields

$$\frac{d\Pi(\lambda_{mh}, \lambda_h(\lambda_{mh}))}{d\lambda_{mh}} = \begin{cases} f_l(c_l(\lambda - \lambda_{mh}))\left(\frac{\mu}{(\mu - \lambda_{mh})^2} - d\right), & \text{if } f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) < 0, \\ g(\lambda, \lambda_{mh})\left(\frac{\mu}{(\mu - \lambda_{mh})^2} - d\right), & \text{if } f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) \geq 0, \end{cases} \quad (48)$$

for $\lambda_{mh} \in [\lambda_P, \min(\lambda, \lambda_F)]$ (the feasible range by (42)), where the function

$$g(\lambda, \lambda_{mh}) \triangleq f_l(c_l(\lambda - \lambda_{mh})) - f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right). \quad (49)$$

It measures the difference between the virtual delay costs of the marginal types c_l and c_h as a function of λ and λ_{mh}, when allocating the corresponding *maximum feasible* rate $\lambda_h = \overline{\lambda}_h(\lambda_{mh})$ to h classes and $\lambda_l = \lambda - \lambda_{mh}$ to l classes. The sign of $g(\lambda, \lambda_{mh})$ is important for the optimal segmentation: $g(\lambda, \lambda_{mh}) > 0$ indicates more pooling increases profits.

The sign of the revenue derivative in (48) only depends on $f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right)$ and $g(\lambda, \lambda_{mh})$, because $f_l > 0$ and the common factor in both cases of (48) is zero at $\lambda_{mh} = \lambda_P$ and positive for $\lambda_{mh} > \lambda_P$. The maximizer $\lambda^*_{mh}(\lambda)$ is unique since the following properties hold for $\lambda_{mh} \in [\lambda_P, \min(\lambda, \lambda_F)]$:

(i) $f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right)$ increases in λ_{mh} since $\overline{\lambda}'_h(\lambda_{mh}) < 0$ by (40) and $f'_h c'_h < 0$.

(ii) $g(\lambda, \lambda_{mh})$ decreases in λ_{mh} since $f'_l c'_l > 0$, and by (i).

Properties (i)–(ii) reflect that, as λ_{mh} increases, both the maximum feasible h rate $\overline{\lambda}_h(\lambda_{mh})$ and the l rate $\lambda_l = \lambda - \lambda_{mh}$ decrease, so the corresponding marginal types and their virtual delay costs (because $f'_l, f'_h > 0$) move apart: $c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)$ and $f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right)$ increase while $c_l(\lambda - \lambda_{mh})$ and $f_l(c_l(\lambda - \lambda_{mh}))$ decrease in λ_{mh}.

Properties (i)–(ii) and (48) determine the maximizer $\lambda^*_{mh}(\lambda)$, and (44) determines $\lambda^*_h(\lambda) = \lambda_h(\lambda^*_{mh}(\lambda))$. The optimal segmentations and the corresponding conditions are summarized in (50). The optimal segmentation sets λ_{mh} and λ_h to minimize the virtual delay cost of types served, subject to the constraints (42)-(43). A segmentation is optimal if and only if it *minimizes* $\lambda_{mh} \in [\lambda_P, \min(\lambda_F, \lambda)]$, and *maximizes* $\lambda_h \leq \overline{\lambda}_h(\lambda_{mh})$, subject to two conditions. (1) The marginal h type's

virtual delay cost is nonnegative: $f_h(c_h(\lambda_h)) \geq 0$. (2) If both h and l are offered $(\lambda_h, \lambda - \lambda_{mh} > 0)$ then the virtual delay cost of the marginal h type (with the shorter lead time) is higher: $f_h(c_h(\lambda_h)) \geq f_l(c_h(\lambda - \lambda_{mh}))$. We discuss these conditions for the solutions summarized in (50), in the order $(h,l) - (h,m,l) - (m,l) - (h,m) - (h,m_{sd})$.

Optimal segmentations for $\mu > d^{-1}$ *and* $\lambda > \lambda_P$						
	conditions (other than $\lambda > \lambda_P$*)*		*arrival rates for each segment*			
segments	λ	*virtual delay costs*	$\lambda^*_{mh}(\lambda)$	$\lambda^*_h(\lambda)$	$\lambda^*_m(\lambda)$	$\lambda^*_l(\lambda)$
(h, m_{sd})	$\lambda < \lambda_F$	$f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right) < 0$	λ	λ_0	$\lambda - \lambda_0$	0
(h,m)	$\lambda < \lambda_F$	$f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right) \geq 0, g(\lambda,\lambda) \geq 0$	λ	$\overline{\lambda}_h(\lambda)$	$\lambda - \overline{\lambda}_h(\lambda)$	0
(h,l)	–	$g(\lambda,\lambda_P) \leq 0$	λ_P	λ_P	0	$\lambda - \lambda_P$
(h,m,l)	–	$g(\lambda,\lambda_P) > 0 > g(\lambda, \min(\lambda,\lambda_F))$	$\lambda^*_{mh}(\lambda)$	$\overline{\lambda}_h(\lambda^*_{mh})$	> 0	$\lambda - \lambda^*_{mh}$
(m,l)	$\lambda > \lambda_F$	$g(\lambda,\lambda_F) \geq 0$	λ_F	0	λ_F	$\lambda - \lambda_F$

(50)

Segmentation (h,l): if $g(\lambda,\lambda_P) \leq 0$, conditions $(1)-(2)$ hold at $\lambda_{mh} = \lambda_P$, which yields maximum allocations to h and l, using the entire λ. *Segmentation* (h,m,l): if $g(\lambda,\lambda_P) > 0 > g(\lambda, \min(\lambda,\lambda_F))$, then $\lambda_{mh} = \lambda_P$ violates (2) whereas $\lambda_{mh} = \min(\lambda,\lambda_F)$ satisfies (2) but does not minimize λ_{mh}. In this case increase λ_{mh} to the point where the marginal types' virtual delay costs are equal: $g(\lambda,\lambda_{mh}) = 0$ for $\lambda_{mh} = \lambda^*_{mh}(\lambda)$. This reduces the l and h rates and increases the rate λ_m.

Segmentations (m,l), (h,m) and (h,m_{sd}): in these cases $g(\lambda,\lambda_{mh}) > 0$ for every $\lambda_{mh} < \min(\lambda,\lambda_F)$, so condition (2) is violated for every segmentation that sells both h and l classes; note that $f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right) < 0$, the condition for (h,m_{sd}), implies $g(\lambda,\lambda) > 0$. In each case condition (2) is met by *not* offering both l and h classes: set $\lambda^*_{mh}(\lambda) = \min(\lambda,\lambda_F)$. If $\lambda > \lambda_F$, no h classes are sold: allocate λ_F to the m class and the rest to l classes. If $\lambda < \lambda_F$, no l classes are sold: set $\lambda^*_{mh}(\lambda) = \lambda$, then maximize $\lambda_h \leq \overline{\lambda}_h(\lambda)$ subject to condition (1), by applying (44) with $\lambda_{mh} = \lambda$. If all of $\overline{\lambda}_h(\lambda)$ can be sold to types with nonnegative f_h this yields (h,m). If not, (h,m_{sd}), *with strategic delay* is optimal: sell h classes to $\lambda_0 < \overline{\lambda}_h(\lambda)$, where $f_h(c_h(\lambda_0)) = 0$, and the m class (lead time $= d$) to the remaining $\lambda - \lambda_0$ consisting of types with *negative* f_h.

Step 3. Virtual delay cost conditions as functions of λ and Λ. The optimality conditions of Step 2, (46) for $\lambda \leq \lambda_P$ and (50) for $\lambda > \lambda_P$, are stated in terms of the signs of the virtual delay cost $f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right)$ and of the virtual delay cost difference $g(\lambda,\lambda_{mh})$ for $\lambda_{mh} = \lambda_P$ and $\lambda_{mh} = \min(\lambda,\lambda_F)$. Next, we translate these conditions into conditions on λ and Λ. For this purpose, we state the technical Lemmas 4-8 (proofs in Online Appendix):

Lemmas 4-6 characterize the signs of $g(\lambda,\lambda_P)$ and $g(\lambda,\min(\lambda,\lambda_F))$ depending on λ and Λ. These properties are important to determine whether pooling is optimal.
Lemmas 7-8 characterize, for the case $f_h(c_{\min}) < 0$, the sign of $f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right)$ depending on λ and Λ. These properties are important to determine whether strategic delay is optimal.

LEMMA 4. *Fix* $\mu > d^{-1}$ *and* $\Lambda > \lambda_P$. *Consider* $g(\lambda,\lambda_{mh})$ *for* $\lambda_{mh} = \lambda_P$ *and* $\lambda_{mh} = \min(\lambda,\lambda_F)$.

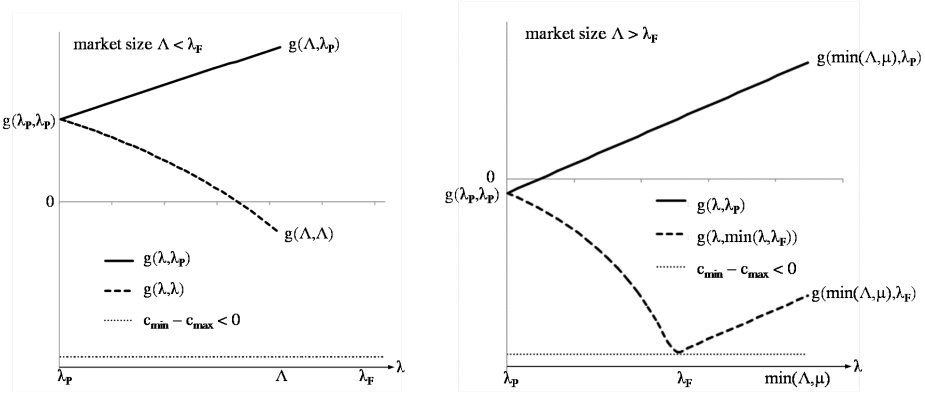
1. *The virtual delay cost difference* $g(\lambda,\lambda_P)$, *where* $\lambda_h = \lambda_P$, $\lambda_m = 0$, *increases in* $\lambda \geq \lambda_P$ *where*

$$g(\lambda,\lambda_P) = f_l(c_l(\lambda - \lambda_P)) - f_h(c_h(\lambda_P)). \tag{51}$$

2. *The virtual delay cost difference* $g(\lambda, \min(\lambda,\lambda_F))$ *varies as follows with* λ.
   (a) *It decreases in* $\lambda \in [\lambda_P, \min(\lambda_F,\Lambda)]$, *where* $\lambda_h = \overline{\lambda}_h(\lambda)$, $\lambda_m = \lambda - \overline{\lambda}_h(\lambda) > 0$, *and*

$$g(\lambda, \min(\lambda,\lambda_F)) = g(\lambda,\lambda) = c_{\min} - f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right), \quad \lambda \in [\lambda_P, \min(\lambda_F,\Lambda)]. \tag{52}$$

Figure 4 **Virtual Delay Cost Differences $g(\lambda,\lambda_P)$ and $g(\lambda,min(\lambda,\lambda_F))$ as Functions of λ**



(b) *It increases in* $\lambda\in[\lambda_F,\min(\Lambda,\mu)]$, *where* $\lambda_h=0$, $\lambda_m=\lambda_F$, $g(\lambda_F,\lambda_F)=c_{\min}-c_{\max}$, *and*

$$g(\lambda,\min(\lambda,\lambda_F))=g(\lambda,\lambda_F)=f_l(c_l(\lambda-\lambda_F))-c_{\max},\quad \lambda\in[\lambda_F,\min(\Lambda,\mu)]. \tag{53}$$

Figure 4 illustrates Lemma 4. In both panels $g(\lambda,\lambda_P)$ increases in λ (Part 1). At left $\Lambda<\lambda_F$ and $g(\lambda,\min(\lambda,\lambda_F))$ decreases in λ. At right $\Lambda>\lambda_F$ and $g(\lambda,\min(\lambda,\lambda_F))$ is minimized and equal to $c_{\min}-c_{\max}<0$ at $\lambda=\lambda_F$, where all types are served FIFO (Part 2). In both panels $g(\lambda,\lambda_P)>g(\lambda,\min(\lambda,\lambda_F))$ for fixed λ: as explained in Step 2, $g(\lambda,\lambda_{mh})$ decreases in λ_{mh} for fixed λ.

From the two panels in Figure 4, note that increasing Λ decreases $g(\lambda,\lambda_P)$ and $g(\lambda,\min(\lambda,\lambda_F))$. This holds because for fixed (λ,λ_{mh}) (where $\lambda_{mh}=\lambda_P$ or $\lambda_{mh}=\min(\lambda,\lambda_F)$), increasing Λ lowers the marginal type $c_l(\lambda-\lambda_{mh})=F^{-1}((\lambda-\lambda_{mh})/\Lambda)$, increases $c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)=\overline{F}^{-1}\left(\overline{\lambda}_h(\lambda_{mh})/\Lambda\right)$, and changes their virtual delay costs accordingly. To make the dependence on Λ explicit we write

$$g(\lambda,\lambda_{mh};\Lambda)=f_l(c_l(\lambda-\lambda_{mh};\Lambda))-f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh});\Lambda\right)\right),$$

for $\lambda\in[\lambda_P,\min(\mu,\Lambda)]$ and $\lambda_{mh}\in[\lambda_P,\min(\lambda,\lambda_F)]$.

Lemma 5 describes how the sign of $g(\lambda,\min(\lambda,\lambda_F);\Lambda)$ depends on the arrival rate λ and the market size Λ. This determines whether it is optimal to pool maximally: If $g(\lambda,\min(\lambda,\lambda_F);\Lambda)\geq 0$, it is profitable to pool the maximum feasible set of customers.

LEMMA 5. *Fix* $\mu>d^{-1}$. *Consider* $g(\lambda,\min(\lambda,\lambda_F);\Lambda)$ *as a function of* $\lambda>\lambda_P$ *and* Λ, *where*

$$g(\lambda,\min(\lambda,\lambda_F);\Lambda)=\begin{cases} g(\lambda,\lambda;\Lambda)=c_{\min}-f_h\left(\overline{F}^{-1}\left(\frac{\overline{\lambda}_h(\lambda)}{\Lambda}\right)\right),\ \lambda\in[\lambda_P,\min(\Lambda,\lambda_F)],\\ g(\lambda,\lambda_F;\Lambda)=f_l\left(F^{-1}\left(\frac{\lambda-\lambda_F}{\Lambda}\right)\right)-c_{\max},\ \lambda\in[\lambda_F,\min(\Lambda,\mu)].\end{cases} \tag{54}$$

1. *For* $\lambda\in[\lambda_P,\min(\Lambda,\lambda_F)]$ *the sign of* $g(\lambda,\min(\lambda,\lambda_F);\Lambda)=g(\lambda,\lambda;\Lambda)$ *depends on the thresholds*

$$\Lambda_2\triangleq\frac{\mu}{2}\left(\frac{1}{\overline{F}\left(f_h^{-1}(c_{\min})\right)}+1\right)-\sqrt{\frac{\mu^2}{4}\left(\frac{1}{\overline{F}\left(f_h^{-1}(c_{\min})\right)}-1\right)^2+\frac{\mu}{d\overline{F}\left(f_h^{-1}(c_{\min})\right)}}, \tag{55}$$

$$\Lambda_3\triangleq\frac{\lambda_P}{\overline{F}\left(f_h^{-1}(c_{\min})\right)}, \tag{56}$$

where $\Lambda_2\in(\lambda_P,\lambda_F)$ *satisfies* $g(\Lambda_2,\Lambda_2;\Lambda_2)=0$ *and* $\Lambda_3>\Lambda_2$ *satisfies* $g(\lambda_P,\lambda_P;\Lambda_3)=0$.

(a) *If* $\Lambda\leq\Lambda_2$ *then* $g(\lambda,\lambda;\Lambda)\geq 0$ *for* $\lambda\in[\lambda_P,\Lambda]$, *where* $\Lambda_2<\lambda_F$.

(b) *If* $\Lambda \in (\Lambda_2, \Lambda_3)$*, then there is an arrival rate threshold* $\lambda_1 \in (\lambda_P, \lambda_F)$*, given by*

$$\lambda_1 \triangleq \mu - \frac{\mu/d}{\mu - \Lambda \overline{F}(f_h^{-1}(c_{\min}))}, \text{ such that} \tag{57}$$

$$g(\lambda, \lambda; \Lambda) \begin{cases} > 0, \text{ if } \lambda \in [\lambda_P, \lambda_1), \\ = 0, \text{ if } \lambda = \lambda_1, \\ < 0, \text{ if } \lambda \in (\lambda_1, \min(\lambda_F, \Lambda)]. \end{cases} \tag{58}$$

(c) *If* $\Lambda > \Lambda_3$*, then* $g(\lambda, \lambda; \Lambda) < 0$ *for* $\lambda \in [\lambda_P, \min(\lambda_F, \Lambda)]$.

2. *For* $\lambda \in [\lambda_F, \min(\Lambda, \mu)]$ *the sign of* $g(\lambda, \min(\lambda, \lambda_F); \Lambda) = g(\lambda, \lambda_F; \Lambda)$ *depends on the thresholds*

$$\underline{\Lambda}_{ml} \triangleq \frac{\lambda_F}{\overline{F}\left(f_l^{-1}(c_{\max})\right)} > \lambda_F, \tag{59}$$

$$\overline{\Lambda}_{ml} \triangleq \frac{1}{dF\left(f_l^{-1}(c_{\max})\right)}, \tag{60}$$

where $g(\underline{\Lambda}_{ml}, \lambda_F; \underline{\Lambda}_{ml}) = 0$ *and* $g\left(\mu, \lambda_F; \overline{\Lambda}_{ml}\right) = 0$.

(a) *If* $F(f_l^{-1}(c_{\max})) \cdot d < \mu^{-1} < d$ *then* $\underline{\Lambda}_{ml} < \mu < \overline{\Lambda}_{ml}$*. If* $\Lambda \in [\underline{\Lambda}_{ml}, \overline{\Lambda}_{ml})$ *then there is a threshold*

$$\lambda_3 \triangleq \lambda_F + \Lambda F\left(f_l^{-1}(c_{\max})\right), \text{ such that} \tag{61}$$

$$g(\lambda, \lambda_F; \Lambda) \begin{cases} < 0, \text{ if } \lambda \in (\lambda_F, \lambda_3), \\ = 0, \text{ if } \lambda = \lambda_3, \\ > 0, \text{ if } \lambda > \lambda_3. \end{cases} \tag{62}$$

If $\Lambda \notin [\underline{\Lambda}_{ml}, \overline{\Lambda}_{ml})$ *then* $g(\lambda, \lambda_F; \Lambda) < 0$ *for all feasible* $\lambda > \lambda_F$.

(b) *If* $\mu^{-1} \leq F(f_l^{-1}(c_{\max})) \cdot d$ *then* $g(\lambda, \lambda_F; \Lambda) < 0$ *for all feasible* $\lambda > \lambda_F$.

Lemma 6 describes how the sign of $g(\lambda, \lambda_P; \Lambda)$ depends on the arrival rate λ and the market size Λ. This determines whether any pooling is optimal: If $g(\lambda, \lambda_P; \Lambda) \leq 0$ then pooling is not optimal.

LEMMA 6. *Fix* $\mu > d^{-1}$*. The sign of* $g(\lambda, \lambda_P; \Lambda)$ *depends as follows on* $\lambda > \lambda_P$ *and* Λ*. Let*

$$\Lambda_4 \triangleq \left\{ \Lambda \geq \lambda_P; \Lambda = \frac{\sqrt{\mu/d}}{F\left(f_l^{-1}(f_h(c_h(\lambda_P; \Lambda)))\right)} \right\}, \tag{63}$$

where Λ_4 *satisfies* $g(\mu, \lambda_P; \Lambda_4) = 0$*. Moreover,* $\Lambda_4 > \max\left(\Lambda_3, \overline{\Lambda}_{ml}, \mu\right)$.

1. *If* $\Lambda < \Lambda_3$ *then* $g(\lambda, \lambda_P; \Lambda) > 0$ *for* $\lambda > \lambda_P$.
2. *If* $\Lambda \in (\Lambda_3, \Lambda_4)$ *then there is an arrival rate threshold* $\lambda_2 > \lambda_P$ *where*

$$\lambda_2 \triangleq \lambda_P + \Lambda F\left(f_l^{-1}(f_h(c_h(\lambda_P; \Lambda)))\right), \text{ such that} \tag{64}$$

$$g(\lambda, \lambda_P; \Lambda) \begin{cases} < 0, \text{ if } \lambda \in [\lambda_P, \lambda_2), \\ = 0, \text{ if } \lambda = \lambda_2, \\ > 0, \text{ if } \lambda > \lambda_2. \end{cases} \tag{65}$$

3. *If* $\Lambda \geq \Lambda_4$ *then* $g(\lambda, \lambda_P; \Lambda) \leq 0$ *for* $\lambda > \lambda_P$.

Lemma 7 describes how an increase in λ changes the lowest virtual delay cost of types buying h classes, given we allocate the maximum feasible rate, i.e., $\overline{\lambda}_h(\lambda)$, to h classes.

LEMMA 7. *Fix $\mu > d^{-1}$ and Λ. Let $f_h(c_{\min}) < 0$. Consider the virtual delay cost $f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right)$.*
1. *It decreases in $\lambda \in [0, \min(\lambda_P, \Lambda)]$, where $\overline{\lambda}_h(\lambda) = \lambda$, the maximum $f_h(c_h(0)) = c_{\max} > 0$, and*

$$f_h(c_h(\lambda))|_{\lambda=\min(\lambda_P,\Lambda)} = \begin{cases} f_h(c_{\min}) < 0, & \Lambda \le \lambda_P, \\ f_h\left(\overline{F}^{-1}\left(\frac{\lambda_P}{\Lambda}\right)\right), & \Lambda > \lambda_P. \end{cases} \tag{66}$$

2. *If $\Lambda > \lambda_P$, it increases in $\lambda \in [\lambda_P, \min(\lambda_F, \Lambda)]$, where $\overline{\lambda}_h(\lambda) = \mu - \mu/d(\mu - \lambda)$ and*

$$f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right)\big|_{\lambda=\min(\lambda_F,\Lambda)} = \begin{cases} f_h\left(\overline{F}^{-1}\left(\frac{\overline{\lambda}_h(\Lambda)}{\Lambda}\right)\right), & \Lambda \in (\lambda_P, \lambda_F), \\ c_{\max} > 0, & \Lambda \ge \lambda_F. \end{cases} \tag{67}$$

Write $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right)$ to make the dependence on Λ explicit. For fixed λ the marginal type $c_h\left(\overline{\lambda}_h(\lambda)\right) = \overline{F}^{-1}\left(\overline{\lambda}_h(\lambda)/\Lambda\right)$ increases in Λ, hence $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right)$ increases in Λ.

Lemma 8 describes the sign of $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right)$ depending on λ and Λ. This determines whether strategic delay is optimal: If $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right) < 0$ then allocating the maximum feasible rate of customers to h classes would result in negative virtual delay costs for the marginal type.

LEMMA 8. *Fix $\mu > d^{-1}$. Let $f_h(c_{\min}) < 0$. The sign of $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right)$ depends as follows on λ and Λ. Let*

$$\underline{\Lambda}_{sd} \triangleq \frac{\mu}{2}\left(\frac{1}{\overline{F}\left(f_h^{-1}(0)\right)} + 1\right) - \sqrt{\frac{\mu^2}{4}\left(\frac{1}{\overline{F}\left(f_h^{-1}(0)\right)} - 1\right)^2 + \frac{\mu}{d\overline{F}\left(f_h^{-1}(0)\right)}}, \tag{68}$$

$$\overline{\Lambda}_{sd} \triangleq \frac{\lambda_P}{\overline{F}\left(f_h^{-1}(0)\right)}, \tag{69}$$

where $f_h\left(c_h\left(\overline{\lambda}_h(\underline{\Lambda}_{sd}); \underline{\Lambda}_{sd}\right)\right) = 0$ and $f_h\left(c_h\left(\lambda_P; \overline{\Lambda}_{sd}\right)\right) = 0$. Moreover $\lambda_P < \underline{\Lambda}_{sd} \le \Lambda_2$ and $\underline{\Lambda}_{sd} < \overline{\Lambda}_{sd} \le \Lambda_3$. Define the arrival rate thresholds

$$\lambda_0 \triangleq \Lambda\overline{F}\left(f_h^{-1}(0)\right), \tag{70}$$

$$\lambda_{sd} \triangleq \mu - \frac{\mu/d}{\mu - \lambda_0}. \tag{71}$$

1. *If $\Lambda < \underline{\Lambda}_{sd}$ then $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right) < 0$ if and only if $\lambda > \lambda_0$, where $\lambda_0 < \lambda_P$.*
2. *If $\Lambda \in [\underline{\Lambda}_{sd}, \overline{\Lambda}_{sd})$ then $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right) < 0$ if and only if $\lambda \in (\lambda_0, \lambda_{sd})$, where $\lambda_0 < \lambda_P < \lambda_{sd} < \lambda_F$.*
3. *If $\Lambda \ge \overline{\Lambda}_{sd}$ then $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right) \ge 0$ for all feasible λ.*

Step 4. Proofs of Parts 2(a)-(c) of Proposition 2. We complete the proof by combining the optimality conditions (46) and (50) from Step 2 with Lemmas 4-8 from Step 3.

Part 2(a). Suppose that $f_h(c_{\min}) \ge 0$ and $\frac{1}{\mu} \le F(f_l^{-1}(c_{\max})) \cdot d$.

The optimality conditions (46) for $\lambda \le \lambda_P$ and (50) for $\lambda > \lambda_P$, combined with Lemmas 5-6, imply the results. Since $f_h(c_{\min}) \ge 0$, segmentation (h, m_{sd}) cannot be optimal: the optimality condition is $f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right) < 0$ by (46) and (50), which cannot hold. Since $\frac{1}{\mu} \le F(f_l^{-1}(c_{\max})) \cdot d$, segmentation (m, l) cannot be optimal: the condition is $g(\lambda, \lambda_F; \Lambda) \ge 0$, see (50), and Lemma 5.2(b) rules it out.

The conditions (46) and (50), combined with Lemmas 5-6, imply the transitions among (h), (h, m), (h, l) and (h, m, l) listed in Table (34). We illustrate how for $\Lambda \in (\Lambda_2, \Lambda_3)$. By (34) the optimal

segmentation transitions as $(h) \to (h,m) \to (h,m,l)$ as λ increases. For $\lambda \le \lambda_P$ segmentation (h) is optimal by (46). For $\lambda > \lambda_P$ Lemma 5 specifies the sign of $g(\lambda, \min(\lambda, \lambda_F); \Lambda)$; Lemma 5.1(b) applies since $\Lambda \in (\Lambda_2, \Lambda_3)$, and Lemma 5.2(b) since $\frac{1}{\mu} \le F(f_l^{-1}(c_{\max})) \cdot d$. Together they specify that $g(\lambda, \min(\lambda, \lambda_F); \Lambda) \ge 0$ for $\lambda \le \lambda_1$ and $g(\lambda, \min(\lambda, \lambda_F); \Lambda) < 0$ otherwise. Lemma 6 specifies the sign of $g(\lambda, \lambda_P; \Lambda)$, where Lemma 6.1 applies since $\Lambda < \Lambda_3$: it specifies that $g(\lambda, \lambda_P; \Lambda) > 0$ for $\lambda > \lambda_P$. It follows from (50) that (h,m) is optimal for $\lambda \le \lambda_1$, and (h,m,l) is optimal for $\lambda > \lambda_1$.

Part 2(b). Suppose that $f_h(c_{\min}) \ge 0$ and $F(f_l^{-1}(c_{\max})) \cdot d < \mu^{-1} < d$. The proof follows the same logic as explained for Part 2(a). However, by Lemma 5.2(a) segmentation (m,l) is optimal if and only if $\Lambda \in [\underline{\Lambda}_{ml}, \overline{\Lambda}_{ml})$ and $\lambda \ge \lambda_3$: in this case $g(\lambda, \lambda_F; \Lambda) \ge 0$, which is the optimality condition for (m,l) by (50). The optimal segmentation for $\Lambda \in [\underline{\Lambda}_{ml}, \overline{\Lambda}_{ml})$ and $\lambda < \lambda_3$ follows from Lemma 5-6 and (50) in exactly the same way as in Part 2(a). This yields Table (35).

Part 2(c). Suppose that $f_h(c_{\min}) < 0$ and $\frac{1}{\mu} \le F(f_l^{-1}(c_{\max})) \cdot d$. The segmentation (h, m_{sd}) is optimal if and only if $f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right) < 0$ where $\overline{\lambda}_h(\lambda)$ is defined in (40). Refer to (46) for $\lambda \le \lambda_P$ and (50) for $\lambda > \lambda_P$. This condition, combined with Lemmas 5, 6, and 8, imply Table (36).

The case $\Lambda < \underline{\Lambda}_{sd}$ in Table (36) is immediate from (46), (50) and Lemma 8.1. For $\Lambda \in [\underline{\Lambda}_{sd}, \overline{\Lambda}_{sd})$ Lemma 8.2. implies the transition $(h) \to (h, m_{sd})$ for $\lambda < \lambda_{sd}$. For $\lambda \ge \lambda_{sd}$ Lemmas 5-6 and (50) imply (h,m) if $\Lambda \le \Lambda_2$ or $(h,m) \to (h,m,l)$ if $\Lambda > \Lambda_2$. ■

Proof of Lemma 2

This Lemma characterizes the partial derivatives of the maximum revenue, that is, the revenue under the optimal segmentation and menu, with respect to λ and μ.

The maximum revenue satisfies $\Pi^*(\lambda,\mu) = \Pi(\lambda, \lambda^*_{mh}, \lambda^*_h, \mu)$ where

$$\Pi(\lambda, \lambda^*_{mh}, \lambda^*_h, \mu) \triangleq \lambda v - \Lambda \int_{c_{\min}}^{c_l\left(\lambda - \lambda^*_{mh}\right)} f(x) f_l(x) \left(\frac{\mu}{\left(\mu - [\lambda - \Lambda F(x)]\right)^2} - d \right) dx + \Lambda \int_{c_h\left(\lambda^*_h\right)}^{c_{\max}} f(x) f_h(x) \left(d - \frac{\mu}{\left(\mu - \Lambda \overline{F}(x)\right)^2} \right) dx,$$

and $\lambda^*_{mh}, \lambda^*_h$ depend on (λ, μ) as tabulated below in (72). The entries for $\lambda, \lambda^*_{mh}, \lambda^*_h$ and μ are from the proof of Proposition 2; refer in particular to (46) and (50) and their discussion. They directly imply the partial derivatives of λ^*_{mh} and λ^*_h, except for (h,m,l) where we derive them below. The proof refers to these properties of λ^*_{mh} and λ^*_h and derives those of Π below. We first review some important facts. Recall from Proposition 2 and its proof: $\lambda_P = \mu - \sqrt{\mu/d}$ and $\lambda_F = \mu - 1/d$ are well defined if $\mu^{-1} < d$, so $\lambda_P < \lambda_F$; $\overline{\lambda}_h(\lambda)$ is defined in (40); and λ_0 is defined in (44).

segments	$\lambda < \mu$	μ^{-1}	λ^*_{mh}	λ^*_h	$\frac{\partial \lambda^*_{mh}}{\partial \lambda}$	$\frac{\partial \lambda^*_{mh}}{\partial \mu}$	$\frac{\partial \lambda^*_h}{\partial \lambda}$	$\frac{\partial \lambda^*_h}{\partial \mu}$
(l)	*any* λ	$\ge d$	0	0	0	0	0	0
(h)	$\le \lambda_P$	$< d$	λ	λ	1	0	1	0
(h, m_{sd})	$\in (\lambda_P, \lambda_F)$	$< d$	λ	$\lambda_0 = \Lambda \overline{F}\left(f_h^{-1}(0)\right) < \lambda_P$	1	0	0	0
(h,m)	$\in (\lambda_P, \lambda_F)$	$< d$	λ	$\lambda^*_h = \overline{\lambda}_h(\lambda) = \mu - \frac{\mu/d}{\mu - \lambda} < \lambda_P$	1	0	< 0	> 0
(h,l)	$> \lambda_P$	$< d$	λ_P	λ_P	0	> 0	0	> 0
(h,m,l)	$> \lambda_P$	$< d$	(i) $\lambda^*_h = \mu - \frac{\mu/d}{\mu - \lambda^*_{mh}} < \lambda_P < \lambda^*_{mh}$ (ii) $f_l(c_l(\lambda - \lambda^*_{mh})) = f_h(c_h(\lambda^*_h))$		$\in (0,1)$		< 0	> 0
(m,l)	$> \lambda_F$	$< d$	λ_F	0	0	> 0	0	0

(72)

We suppress the arguments of Π^*, Π, λ^*_{mh} and λ^*_h.

Recall that $c_l(x) = F^{-1}(x/\Lambda)$, so $\Lambda f(c_l(x))c_l'(x) = 1$, $c_l' > 0$; $c_h(x) = \overline{F}^{-1}(x/\Lambda)$, so $\Lambda f(c_h(x))c_h'(x) = -1$, $c_h' < 0$; $f_l, f_l', f_h' > 0$ and $f_h(c_h(\lambda_h^*)) \geq 0$. Therefore,

$$\Pi_\lambda^* = \frac{\partial \Pi}{\partial \lambda} + \frac{\partial \Pi}{\partial \lambda_{mh}^*}\frac{\partial \lambda_{mh}^*}{\partial \lambda} + \frac{\partial \Pi}{\partial \lambda_h^*}\frac{\partial \lambda_h^*}{\partial \lambda}, \text{ where} \tag{73}$$

$$\frac{\partial \Pi}{\partial \lambda_{mh}^*} = f_l(c_l(\lambda - \lambda_{mh}^*))\left(\frac{\mu}{(\mu - \lambda_{mh}^*)^2} - d\right) \geq 0,\ \frac{\partial^2 \Pi}{\partial \lambda \partial \lambda_{mh}^*} \geq 0, \text{and } \frac{\partial^2 \Pi}{\partial \mu \partial \lambda_{mh}^*} < 0, \tag{74}$$

$$\frac{\partial \Pi}{\partial \lambda} = v - \Lambda \int_{c_{\min}}^{c_l(\lambda - \lambda_{mh}^*)} \frac{f(x) f_l(x)\, 2\mu}{(\mu - \lambda + \Lambda F(x))^3} dx - \frac{\partial \Pi}{\partial \lambda_{mh}^*}, \quad \frac{\partial^2 \Pi}{\partial \lambda^2} \leq 0, \text{and } \frac{\partial^2 \Pi}{\partial \lambda \partial \mu} > 0, \tag{75}$$

$$\frac{\partial \Pi}{\partial \lambda_h^*} = f_h(c_h(\lambda_h^*))\left(d - \frac{\mu}{(\mu - \lambda_h^*)^2}\right) \geq 0,\ \frac{\partial^2 \Pi}{\partial \lambda_h^{*2}} \leq 0, \text{ and } \frac{\partial^2 \Pi}{\partial \mu \partial \lambda_h^*} \geq 0. \tag{76}$$

Part 1(a). It follows from (72) and (73)-(76) that $\Pi_\lambda^* = v$ for (h, m_{sd}) and $\Pi_\lambda^* \geq v$ for (h).

Part 1(b). We show that $\Pi_{\lambda\lambda}^* < 0 < \Pi_{\lambda\mu}^*$ for (h,m,l) and $\Pi_{\lambda\lambda}^* \leq 0 \leq \Pi_{\lambda\mu}^*$ for any segmentation *other than* (h,m,l). We omit the remaining straightforward checks that $\Pi_{\lambda\lambda}^* < 0 < \Pi_{\lambda\mu}^*$ for (l), (h,m),(h,l), and (m,l).

Proof that $\Pi_{\lambda\lambda}^* < 0 < \Pi_{\lambda\mu}^*$ *for* (h,m,l). Table (72) claims $0 < \partial \lambda_{mh}^*/\partial \lambda, \partial \lambda_{mh}^*/\partial \mu < 1$, which is implied by equations $(i) - (ii)$ in the table and the fact that $f_l' c_l' > 0 > f_h' c_h'$:

$$sign\left(1 - \frac{\partial \lambda_{mh}^*}{\partial \lambda}\right) = -sign\left(\frac{\partial \lambda_h^*}{\partial \lambda}\right) = sign\left(\frac{\partial \lambda_{mh}^*}{\partial \lambda}\right), \text{ so } 0 < \frac{\partial \lambda_{mh}^*}{\partial \lambda} < 1. \tag{77}$$

$$sign\left(\frac{\partial \lambda_{mh}^*}{\partial \mu}\right) = sign\left(\frac{\partial \lambda_h^*}{\partial \mu}\right) \text{ and } \frac{1 - \partial \lambda_{mh}^*/\partial \mu}{\mu - \lambda_{mh}^*} + \frac{1 - \partial \lambda_h^*/\partial \mu}{\mu - \lambda_h^*} = \frac{1}{\mu}, \text{ so } 0 < \frac{\partial \lambda_{mh}^*}{\partial \mu} < 1. \tag{78}$$

The first equations in (77)-(78) each follow from (ii) and $f_l' c_l' > 0 > f_h' c_h'$, the second equations each follow from (i). For fixed λ the total derivative $d\Pi/d\lambda_{mh} = 0$ at $\lambda_{mh} = \lambda_{mh}^*$: see (48) in the proof of Proposition 2 where $g(\lambda, \lambda_{mh}^*) = 0$ for (h,m,l). It follows that

$$\Pi_\lambda^* = \frac{\partial \Pi}{\partial \lambda} + \frac{\partial \lambda_{mh}^*}{\partial \lambda}\left[\frac{\partial \Pi}{\partial \lambda_{mh}^*} + \frac{\partial \Pi}{\partial \lambda_h^*}\frac{\partial \lambda_h^*}{\partial \lambda_{mh}^*}\right] = \frac{\partial \Pi}{\partial \lambda} \text{ for all } (\lambda, \mu) \text{ with } (h,m,l). \tag{79}$$

We show that the following holds:

$$\Pi_{\lambda\lambda}^* = \frac{\partial^2 \Pi}{\partial \lambda^2} + \frac{\partial^2 \Pi}{\partial \lambda \partial \lambda_{mh}^*}\frac{\partial \lambda_{mh}^*}{\partial \lambda} \leq -\frac{f_l(c_l(\lambda - \lambda_{mh}^*))\, 2\mu}{(\mu - \lambda_{mh}^*)^3} - \frac{\partial^2 \Pi}{\partial \lambda_{mh}^* \partial \lambda}\left(1 - \frac{\partial \lambda_{mh}^*}{\partial \lambda}\right) < 0, \tag{80}$$

$$\Pi_{\lambda\mu}^* = \frac{\partial^2 \Pi}{\partial \lambda \partial \mu} + \frac{\partial^2 \Pi}{\partial \lambda \partial \lambda_{mh}^*}\frac{\partial \lambda_{mh}^*}{\partial \mu} > 0. \tag{81}$$

The equations for $\Pi_{\lambda\lambda}^*$, $\Pi_{\lambda\mu}^*$ hold by (79) and $\partial^2 \Pi/(\partial \lambda \partial \lambda_h^*) = 0$ by (76). The first inequality in (80) holds by (74)-(75); the second by (74) and (77). The inequality in (81) holds by (74)-(75) and (78).

Proof that $\Pi_{\lambda\lambda}^* \leq 0 \leq \Pi_{\lambda\mu}^*$ *for segmentations other than* (h,m,l). First consider $\Pi_{\lambda\lambda}^*$. Since $\partial^2 \Pi/(\partial \lambda \partial \lambda_h^*) = 0$ by (76) and $\partial^2 \lambda_{mh}^*/\partial \lambda^2 = 0$ by (72) we have

$$\Pi_{\lambda\lambda}^* = \frac{\partial^2 \Pi}{\partial \lambda^2} + 2\frac{\partial^2 \Pi}{\partial \lambda \partial \lambda_{mh}^*}\frac{\partial \lambda_{mh}^*}{\partial \lambda} + \frac{\partial^2 \Pi}{\partial \lambda_{mh}^{*2}}\left(\frac{\partial \lambda_{mh}^*}{\partial \lambda}\right)^2 + \left[\frac{\partial^2 \Pi}{\partial \lambda_h^{*2}}\left(\frac{\partial \lambda_h^*}{\partial \lambda}\right)^2 + \frac{\partial \Pi}{\partial \lambda_h^*}\frac{\partial^2 \lambda_h^*}{\partial \lambda^2}\right].$$

The terms in brackets are nonpositive by (76) and $\partial^2\lambda_h^*/\partial\lambda^2 \leq 0$ from (72). The other terms satisfy

$$\frac{\partial^2\Pi}{\partial\lambda^2}+2\frac{\partial^2\Pi}{\partial\lambda\partial\lambda_{mh}^*}\frac{\partial\lambda_{mh}^*}{\partial\lambda}+\frac{\partial^2\Pi}{\partial\lambda_{mh}^{*2}}\left(\frac{\partial\lambda_{mh}^*}{\partial\lambda}\right)^2 \leq \left(\left(\frac{\partial\lambda_{mh}^*}{\partial\lambda}\right)^2-1\right)\frac{f_l\left(c_l\left(\lambda-\lambda_{mh}\right)\right)2\mu}{\left(\mu-\lambda_{mh}\right)^3}-\left(1-\frac{\partial\lambda_{mh}^*}{\partial\lambda}\right)^2\frac{\partial^2\Pi}{\partial\lambda_{mh}\partial\lambda}$$

by (74)-(75). The RHS is nonpositive since $\partial^2\Pi/\partial\lambda_{mh}\partial\lambda \geq 0$ by (74) and $\partial\lambda_{mh}^*/\partial\lambda \leq 1$ by (72).

Next consider $\Pi_{\lambda\mu}^*$. By (72) we have $\left(\partial\lambda_{mh}^*/\partial\lambda\right)\left(\partial\lambda_{mh}^*/\partial\mu\right)=\partial^2\lambda_{mh}^*/\left(\partial\lambda\partial\mu\right)=0$ which implies

$$\Pi_{\lambda\mu}^*=\left(\frac{\partial^2\Pi}{\partial\lambda\partial\mu}+\frac{\partial^2\Pi}{\partial\lambda_{mh}^*\partial\mu}\frac{\partial\lambda_{mh}^*}{\partial\lambda}\right)+\left(\frac{\partial^2\Pi}{\partial\lambda\partial\lambda_{mh}^*}\frac{\partial\lambda_{mh}^*}{\partial\mu}\right)+\left(\frac{d}{d\mu}\left[\frac{\partial\Pi}{\partial\lambda_h^*}\frac{\partial\lambda_h^*}{\partial\lambda}\right]\right). \tag{82}$$

We show that each bracket is nonnegative. For the first, (72) and (74)-(75) imply

$$\frac{\partial^2\Pi}{\partial\lambda\partial\mu}+\frac{\partial^2\Pi}{\partial\lambda_{mh}^*\partial\mu}\frac{\partial\lambda_{mh}^*}{\partial\lambda} \geq -\frac{\partial^2\Pi}{\partial\lambda_{mh}^*\partial\mu}\left(1-\frac{\partial\lambda_{mh}^*}{\partial\lambda}\right) \geq 0.$$

The second bracket of (82) is nonnegative by (72) and (74). The third bracket of (82) is also nonnegative. For segmentations other than (h,m) and (h,m,l) we have $\partial\lambda_h^*/\partial\lambda=0$ or $=1$ by (72) and $\partial^2\Pi/\left(\partial\lambda_h^*\partial\mu\right) \geq 0$ by (76). For (h,m) substitute for $\lambda_h^*=\mu-\mu/d\left(\mu-\lambda\right)^{-1}$ from (72) to get:

$$\frac{\partial\Pi}{\partial\lambda_h^*}\frac{\partial\lambda_h^*}{\partial\lambda}=f_h\left(c_h\left(\lambda_h^*\right)\right)\left(d-\frac{\mu}{\left(\mu-\lambda_h^*\right)^2}\right)\frac{\partial\lambda_h^*}{\partial\lambda}=v+f_h\left(c_h\left(\lambda_h^*\right)\right)\left(d-\frac{\mu}{\left(\mu-\lambda\right)^2}\right). \tag{83}$$

This expression increases in μ: the virtual delay cost $f_h\left(c_h\left(\lambda_h^*\right)\right) \geq 0$ decreases in μ since $f_h'c_h'<0$ and $\partial\lambda_h^*/\partial\mu>0$ by (72), and its multiplier is negative $(\lambda>\lambda_P)$ and increases in μ.

Part 2. The function $\Pi^*\left(\lambda,\mu\right)=\Pi(\lambda,\lambda_{mh}^*,\lambda_h^*,\mu)$ is defined piecewise: $\lambda_{mh}^*\left(\lambda,\mu\right)$ and $\lambda_h^*\left(\lambda,\mu\right)$ depend on the optimal segmentations which vary with (λ,μ) as specified by Proposition 2. We establish continuity within each piece and at the transition points.

By the definition of Π and table (72) all first and second order derivatives of Π, λ_{mh}^* and λ_h^* with respect to (λ,μ) are continuous for *each* segmentation. Therefore so are the functions $\Pi_\lambda^*\left(\lambda,\mu\right)$, $\Pi_{\lambda\lambda}^*\left(\lambda,\mu\right)$ and $\Pi_{\lambda\mu}^*\left(\lambda,\mu\right)$. It remains to show the stated properties at each (λ,μ) with a transition *between* two optimal segmentations. By Proposition 2, the possible transitions are as follows.

Transitions in optimal segmentation involving (h) or (h,m_{sd})					
from	to	at (λ,μ)	λ_{mh}^*	virtual delay cost	Lemma
(h)	(h,m_{sd})	$\lambda=\lambda_0,\ \mu>d^{-1}$	λ	$f_h\left(c_h\left(\lambda_0\right)\right)=0$	8.1
(h,m_{sd})	(h,m)	$\lambda=\lambda_{sd},\ \mu>d^{-1}$	λ	$f_h\left(c_h\left(\overline{\lambda}_h\left(\lambda_{sd}\right)\right)\right)=0$	8.2
(h)	(h,l)	$\lambda=\lambda_P,\ \mu>d^{-1}$	λ		
(h)	(h,m)	$\lambda=\lambda_P,\ \mu>d^{-1}$	λ		
Transitions in optimal segmentation involving neither (h) nor (h,m_{sd})					
from	to	at (λ,μ)	λ_{mh}^*	virtual delay cost	Lemma
(l)	(h,l)	$\lambda<\mu=d^{-1}$	0		
(l)	(m,l)	$\lambda<\mu=d^{-1}$	0		
(h,m)	(h,m,l)	$\lambda=\lambda_1,\ \mu>d^{-1}$	λ	$f_h\left(c_h\left(\overline{\lambda}_h\left(\lambda_1\right)\right)\right)=f_l\left(c_l\left(\lambda-\lambda_{mh}^*\right)\right)=c_{\min}$	5.1(b)
(h,m,l)	(m,l)	$\lambda=\lambda_3,\ \mu>d^{-1}$	λ_F	$f_l\left(c_l\left(\lambda_3-\lambda_F\right)\right)=c_{\max}$	5.2(a)
(h,l)	(h,m,l)	$\lambda=\lambda_2,\ \mu>d^{-1}$	λ_P	$f_h\left(c_h\left(\lambda_P\right)\right)=f_l\left(c_l\left(\lambda_2-\lambda_P\right)\right)$	6.2

The following facts establish (a) and (b). At (λ,μ) where the segmentation transitions from (h) or (h,m_{sd}) we have $\Pi_\lambda^*=v$. At (λ,μ) with a transition involving neither (h) nor (h,m_{sd}) the two expressions for $\Pi_{\lambda\lambda}^*$ (one for each segmentation) agree, and ditto for $\Pi_{\lambda u}^*$. We omit these checks; they are straightforward using table (72) and the formulae for Π_λ^*, $\Pi_{\lambda\lambda}^*$ and $\Pi_{\lambda\mu}^*$ derived above. ■

Proof of Theorem 1
The results follow from Proposition 2 and Lemma 2. ■

Proof of Theorem 2
As a preliminary, we prove two key properties of the thresholds μ_H, defined by (22), and μ_{SD}, defined by (23), which will simplify the proofs of Parts 1-3. The optimal arrivale rate $\lambda^*(\mu)$ satisfies

$$\lambda^*(\mu) = \arg\left\{\max_{\lambda} \Pi^*(\lambda,\mu) \text{ s.t. } \lambda \in [0,\Lambda] \cap [0,\mu)\right\}. \tag{84}$$

p1. Suppose that $f_h(c_{\min}) \geq 0$. Then (i) the optimal rate $\lambda^*(\mu) = \Lambda$ for $\mu \geq \mu_H$, and (ii) the optimal segmentation is (h) if and only if $\mu \geq \mu_H$.

Proof of p1. Recall that $\lambda_P = \mu - \sqrt{\mu/d}$ for $\mu > d^{-1}$ as defined in Proposition 2. We write $\lambda_P(\mu)$ to make its dependence on μ explicit. For fixed μ the following facts imply that *p1* holds if the condition "$\mu \geq \mu_H$" is replaced by "$\Lambda \leq \lambda_P(\mu)$". First, segmentation (h) is optimal for fixed λ if and only if $\mu > d^{-1}$ and $\lambda \leq \lambda_P(\mu)$; this holds by Proposition 2.2 and its proof. Second, $\Pi^*_\lambda(\lambda,\mu) \geq v > 0$ for all (λ,μ) where segmentation (h) is optimal, and $\Pi^*_\lambda(\lambda,\mu)$ is continuous in (λ,μ); see Lemma 2. The proof of *p1* is complete if $\mu \geq \mu_H \Leftrightarrow \Lambda \leq \lambda_P(\mu)$. This holds since $\Lambda = \lambda_P(\mu_H)$ by the definition (22), $\lambda_P(d^{-1}) = 0$, and $\lambda'_P(\mu) = 1 - 1/(2\sqrt{\mu d}) > 0$ for $\mu \geq d^{-1}$.

p2. Suppose that $f_h(c_{\min}) < 0$. Then (i) $\lambda^*(\mu) = \Lambda$ for $\mu \geq \mu_{SD}$, and (ii) the optimal segmentation is (h, m_{sd}) if and only if $\mu > \mu_{SD}$.

Proof of p2. Recall the threshold $\underline{\Lambda}_{sd}$ from Proposition 2.2(c) and its proof; it is defined in (68) of Lemma 8. We write $\underline{\Lambda}_{sd}(\mu)$ to make its dependence on μ explicit. For fixed μ the following facts imply that *p2* holds if the conditions "$\mu \geq \mu_{SD}$" and "$\mu > \mu_{SD}$" are replaced by "$\Lambda \leq \underline{\Lambda}_{sd}(\mu)$" and "$\Lambda < \underline{\Lambda}_{sd}(\mu)$", respectively. First, the optimal segmentation is (h, m_{sd}) at the largest feasible λ if and only if $\Lambda < \underline{\Lambda}_{sd}(\mu)$; this holds by Proposition 2.2(c); for details see Lemma 8. Second, the revenue Π^* satisfies $\Pi^*_\lambda(\lambda,\mu) \geq v > 0$ for all (λ,μ) where segmentation (h, m_{sd}) is optimal, $\Pi^*_{\lambda\lambda}(\lambda,\mu) \geq 0$ for all (λ,μ), and $\Pi^*_\lambda(\lambda,\mu)$ is continuous in (λ,μ); see Lemma 2. We complete the proof of *p2* by showing that $\mu = \mu_{SD}(\Lambda) \Leftrightarrow \Lambda = \underline{\Lambda}_{sd}(\mu)$ and $\mu > \mu_{SD}(\Lambda) \Leftrightarrow \Lambda < \underline{\Lambda}_{sd}(\mu)$. We write $\mu_{SD}(\Lambda)$ to emphasize that μ_{SD} depends on Λ through its defining equation (23). Fix μ and solve (23) for Λ to get $\Lambda = \underline{\Lambda}_{sd}(\mu)$ as defined in (68) of Lemma 8. Note that $\mu'_{SD}(\Lambda) > 0$ since the LHS of (23) increases in μ and decreases in Λ. For fixed Λ the fact that $\Lambda + d^{-1} < \mu_{SD} < \mu_H$ follows since the LHS of (23) is 0 for $\mu = \Lambda + d^{-1}$, and 1 for $\mu = \mu_H$ since $\lambda_P(\mu_H) = \mu_H - \sqrt{\mu_H/d} = \Lambda$.

Part 1. For $\mu \leq \mu_{\min}$, $\lambda^*(\mu) = 0$ since no type buys at a positive price: $w(c) \geq 1/\mu_{\min} = d + v/c_{\min}$ implies $v + c \cdot (d - w(c)) \leq v(1 - c/c_{\min}) \leq 0$. For $\mu = \mu_{\min}$ we have $\Pi^*_\lambda(0,\mu) = 0$. For $\mu \in (\mu_{\min}, d^{-1})$ segmentation (l) is optimal for all λ by Proposition 2.1. By Lemma 2, $\Pi^*_{\lambda\mu}(\lambda,\mu) > 0 > \Pi^*_{\lambda\lambda}(\lambda,\mu)$ under segmentation (l); therefore $\lambda^*(\mu) > 0$ for $\mu > \mu_{\min}$. Since $\Pi^*_\lambda(\lambda,\mu)$ is continuous in (λ,μ) we have $\lambda^*(\mu) < \Lambda$ for $\mu \in (\mu_{\min}, \mu_{\min} + \varepsilon)$ and small $\varepsilon > 0$.

We next show that there exists an unique threshold $\mu_A > \mu_{\min}$ such that $\lambda^*(\mu) = \Lambda$ if and only if $\mu \geq \mu_A$, where $\mu_A < \mu_H$ if $f_h(c_{\min}) \geq 0$ and $\mu_A < \mu_{SD}$ if $f_h(c_{\min}) < 0$. By Lemma 2, $\Pi^*(\lambda,\mu)$ is concave in λ for fixed μ, and $\Pi^*_{\lambda\mu}(\lambda,\mu) \geq 0$ for all (λ,μ). It follows that $\lambda^*(\mu) = \Lambda \Leftrightarrow \Pi^*_\lambda(\Lambda,\mu) \geq 0$ for any μ, and if $\lambda^*(\mu) = \Lambda$ for some μ then $\lambda^*(\mu') = \Lambda$ for all $\mu' > \mu$. It remains to show that there exists μ that satisfies $\lambda^*(\mu) = \Lambda$ and either $\mu < \mu_H$ if $f_h(c_{\min}) \geq 0$, or $\mu_A < \mu_{SD}$ if $f_h(c_{\min}) < 0$. This holds since *p1* implies that $\Pi^*_\lambda(\Lambda, \mu_H) = v > 0$, *p2* implies that $\Pi^*_\lambda(\Lambda, \mu_{SD}) = v > 0$ if $f_h(c_{\min}) < 0$, and because $\Pi^*_{\lambda\mu}(\lambda,\mu)$ is continuous in (λ,μ) by Lemma 2.2.

It remains to show that $\lambda^*(\mu)$ is strictly increasing on $[\mu_{\min}, \mu_A]$. Lemma 2.1 implies that for fixed μ, $\Pi^*(\lambda,\mu)$ has a unique maximizer $\lambda^*(\mu)$, and that if $\lambda^*(\mu)<\Lambda$ then the optimal segmentation is $(l),(h,l),(h,m),(m,l,)$or (h,m,l). Under each of these segmentations, $\Pi^*_{\lambda\lambda}(\lambda,\mu)<0<\Pi^*_{\lambda\mu}(\lambda,\mu)$ (Lemma 2.1) and $\Pi^*_{\lambda\lambda}(\lambda,\mu)$ and $\Pi^*_{\lambda\mu}(\lambda,\mu)$ are continuous in (λ,μ) (Lemma 2.2). We have $\Pi^*_{\lambda}(\lambda^*(\mu),\mu)=0$ for $\mu\in[\mu_{\min},\mu_A]$. By the implicit function theorem $\lambda^*(\mu)$ is differentiable and $\lambda^{*\prime}(\mu)=-\Pi^*_{\lambda\mu}(\lambda^*(\mu),\mu)/\Pi^*_{\lambda\lambda}(\lambda^*(\mu),\mu)>0$ for $\mu\in[\mu_{\min},\mu_A]$.

Part 2. We first prove two key properties.

p3 . If $\mu\in(d^{-1},\mu_H)$ and $\lambda^*(\mu)=\Lambda$, then pooling must be optimal.

Proof of p3. For $\mu\in(d^{-1},\mu_H)$, by *p1-p2* and Proposition 2, only one of $(h,l),(h,m),(m,l,),(h,m,l)$ or (h,m_{sd}) can be optimal. Only (h,l) has no pooling, but it cannot be optimal if $\lambda^*(\mu)=\Lambda$, for this rules out the necessary optimality condition $f_l(c_l^*(\Lambda))\le f_h(c_h^*(\Lambda))$ (Lemma 1.2). Recall that $c_l^*(\lambda)=F^{-1}(\lambda_l^*(\lambda)/\Lambda)$ and $c_h^*(\lambda)=\overline{F}^{-1}(\lambda_h^*(\lambda)/\Lambda)$, where $\lambda_l^*(\lambda)$ and $\lambda_h^*(\lambda)$ are, respectively, the optimal arrival rates to l and h classes for fixed λ. Under (h,l) with $\lambda^*(\mu)=\Lambda$, they satisfy $\lambda_h^*(\Lambda)=\mu-\sqrt{\mu/d}$ and $\lambda_l^*(\Lambda)=\Lambda-\lambda_h^*$ by (72). It follows that $c_l^*(\Lambda)=c_h^*(\Lambda)$. The proof is complete since by definition $f_l(c)>f_h(c)$ for all c.

p4. Suppose pooling is not optimal for a fixed $\mu<\mu_H$. Then pooling is not optimal for $\mu'<\mu$.

Proof of p4. If $\mu\le d^{-1}$ this follows since segmentation (l) without pooling is optimal for all $\mu'<\mu$ by Proposition 2.1. Suppose that $\mu\in(d^{-1},\mu_H)$. By *p3* segmentation (h,l) must be optimal for μ, and $\lambda^*(\mu)<\Lambda$. Let $\lambda_h^*(\mu)=\mu-\sqrt{\mu/d}$ and $\lambda_l^*(\mu)=\lambda^*(\mu)-\lambda_h^*(\mu)$ be the corresponding optimal rates. Optimality requires $f_l(F^{-1}(\lambda_l^*(\mu)/\Lambda))\le f_h(\overline{F}^{-1}(\lambda_h^*(\mu)/\Lambda))$ by Lemma 1.2, and

$$0=\Pi^*_{\lambda}(\lambda^*(\mu),\mu)\Leftrightarrow v=\Lambda\int_{c_{\min}}^{F^{-1}(\lambda_l^*(\mu)/\Lambda)}\frac{f(x)f_l(x)\,2\mu}{(\mu-\lambda^*(\mu)+\Lambda F(x))^3}dx=\Lambda\int_{c_{\min}}^{F^{-1}(\lambda_l^*(\mu)/\Lambda)}\frac{f(x)f_l(x)\,2\mu}{(\sqrt{\mu/d}-\lambda_l^*(\mu)+\Lambda F(x))^3}dx, \tag{85}$$

where the second equation holds since $\mu-\lambda^*(\mu)=\sqrt{\mu/d}-\lambda_l^*(\mu)$. We have $\lambda_l^{*\prime}(\mu)>0$, since the RHS of this equation strictly decreases in μ for fixed $\lambda_l^*(\mu)$ and strictly increases in $\lambda_l^*(\mu)$ for fixed μ. Noting that $\lambda_h^{*\prime}(\mu)=1-1/(2\sqrt{\mu d})>0$ implies that $f_l(F^{-1}(\lambda_l^*(\mu)/\Lambda))-f_h(\overline{F}^{-1}(\lambda_h^*(\mu)/\Lambda))$ strictly increases in μ. Therefore the optimality conditions for (h,l) hold for every $\mu'\in(d^{-1},\mu)$.

Part 1 and *p3-p4* imply that there is an unique $\mu_P\in[d^{-1},\mu_H)$ such that pooling is not optimal for $\mu\le\mu_P$ and optimal for $\mu\in(\mu_P,\mu_H)$. Together with *p1-p2* shown above, this proves 2(a)-(b).

Part 3. *Threshold v_A.* By Part 1 we need $\mu_A\le d^{-1}$, which holds iff $\Pi^*_{\lambda}(\Lambda,\mu)\ge 0$ for $\mu=d^{-1}$. Segmentation (l) is optimal for $\mu=d^{-1}$. Substituting $v=v_A$, $\lambda_l^*(\mu)=\Lambda$, $\lambda_h^*(\mu)=0$ in (85) yields

$$\Pi^*_{\lambda}(\Lambda,\mu)\big|_{\mu=d^{-1}}=v_A-\Lambda\int_{c_{\min}}^{c_{\max}}\frac{f(x)f_l(x)\,2d^2}{(1-d\Lambda\overline{F}(x))^3}dx=0. \tag{86}$$

Threshold v_P. By Part 2, we need $\mu_P=d^{-1}$. If $v\ge v_A$ then $\lambda^*(\mu)=\Lambda$ for $\mu\ge d^{-1}$, and it follows from *p3* that pooling is optimal for $\mu\in(d^{-1},\mu_H)$. If $v<v_A$ then $\lambda^*(\mu)<\Lambda$ for $\mu=d^{-1}$, so $\Pi^*_{\lambda}(\lambda^*(\mu),\mu)=0$ for $\mu=d^{-1}$. By *p4* pooling is optimal for $\mu\in(d^{-1},\mu_H)$ if and only if segmentation (h,l) is not optimal at any such μ. This in turn holds iff $f_l(F^{-1}(\lambda_l^*(\mu)/\Lambda))-f_h(\overline{F}^{-1}(\lambda_h^*(\mu)/\Lambda))\ge 0$ for $\mu=d^{-1}$, because this virtual delay cost difference strictly increases in μ as shown in proving *p4*. Noting that $\lambda_h^*(\mu)=0$ for $\mu=d^{-1}$, this condition is equivalent to $\lambda_l^*(\mu)\ge\Lambda F\left(f_l^{-1}(c_{\max})\right)$. Substituting in (85) the capacity $\mu=d^{-1}$, the rate for l classes $\lambda_l^*(\mu)=\Lambda F\left(f_l^{-1}(c_{\max})\right)$ and $v=v_P$ yields $\Pi^*_{\lambda}(\lambda^*(\mu),\mu)=0$. The proof is complete since $\lambda_l^*(\mu)$ strictly increases in v for $\mu=d^{-1}$. ■

Online Appendix: Proofs of Technical Lemmas 4-8

Proof of Lemma 4. By (49) we have $g(\lambda,\lambda_{mh}) = f_l(c_l(\lambda-\lambda_{mh})) - f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right)$. Set $\lambda_{mh} = \lambda_P$ and note that $\overline{\lambda}_h(\lambda_P) = \lambda_P$ to get (51). Set $\lambda_{mh} = \lambda$ and note that $c_l(0) = c_{\min}$ to get (52). Set $\lambda_{mh} = \lambda_F$ and note that $\overline{\lambda}_h(\lambda_F) = 0$ and $c_h(0) = c_{\max}$ to get (53). Parts 1. and 2(b) follow since $f_l' c_l' > 0$ and the rate of l classes $\lambda - \lambda_{mh}$ increases in λ, while λ_h is fixed. Part 2(a) follows since $f_h' c_h' < 0$ and $\overline{\lambda}_h'(\lambda) < 0$ for $\lambda \in [\lambda_P, \lambda_F]$ by (40), while $\lambda_l = 0$. □

Proof of Lemma 5. From Lemma 4.2, for fixed Λ the function $g(\lambda, \min(\lambda,\lambda_F);\Lambda)$ is decreasing in $\lambda \in [\lambda_P, \min(\Lambda,\lambda_F)]$, negative at $\lambda = \lambda_F$ and increasing in $\lambda \in [\lambda_F, \min(\Lambda,\mu)]$. Hence it has at most two roots in λ, one smaller and the other larger than λ_F. These roots are determined by the signs of $g(\lambda, \min(\lambda,\lambda_F);\Lambda)$ for $\lambda = \lambda_P$, for $\lambda = \Lambda$ when $\Lambda \in (\lambda_P, \lambda_F)$, and for $\lambda = \min(\Lambda,\mu)$ when $\Lambda > \lambda_F$. The proof identifies thresholds on Λ and λ that determine these signs.

Part 1. The threshold Λ_3 determines the sign of $g(\lambda, \min(\lambda,\lambda_F);\Lambda)$ for $\lambda = \lambda_P$. It is the unique solution of $g(\lambda_P,\lambda_P;\Lambda) = c_{\min} - f_h(\overline{F}^{-1}(\lambda_P/\Lambda)) = 0$ in $\Lambda \geq \lambda_P$. This follows because $g(\lambda_P,\lambda_P;\lambda_P) = c_{\min} - f_h(c_{\min}) > 0$ and since $g(\lambda_P,\lambda_P;\Lambda)$ decreases in Λ with $\lim_{\Lambda\to\infty} g(\lambda_P,\lambda_P;\Lambda) = c_{\min} - c_{\max} < 0$. Therefore $g(\lambda_P,\lambda_P;\Lambda) > 0$ if $\Lambda < \Lambda_3$ and conversely for $\Lambda > \Lambda_3$. Solving for Λ_3 yields (56).

The threshold Λ_2 determines the sign of $g(\lambda, \min(\lambda,\lambda_F);\Lambda)$ for $\lambda = \Lambda$. It is the unique solution of $g(\Lambda,\Lambda;\Lambda) = c_{\min} - f_h(\overline{F}^{-1}(\overline{\lambda}_h(\Lambda)/\Lambda)) = 0$ in $\Lambda \in (\lambda_P,\lambda_F)$. This follows since $g(\lambda_P,\lambda_P;\lambda_P) = c_{\min} - f_h(c_{\min}) > 0 > g(\lambda_F,\lambda_F;\lambda_F) = c_{\min} - c_{\max}$, and $g(\Lambda,\Lambda;\Lambda)$ decreases in $\Lambda \in [\lambda_P,\lambda_F]$ as the fraction $\overline{\lambda}_h(\Lambda)/\Lambda$ allocated to h classes decreases in Λ. It follows that $g(\Lambda,\Lambda;\Lambda) \geq 0$ if $\Lambda \leq \Lambda_2$ and $g(\min(\Lambda,\lambda_F), \min(\Lambda,\lambda_F);\Lambda) < 0$ if $\Lambda > \Lambda_2$. Solving for Λ_2 yields (55).

To show that $\Lambda_2 < \Lambda_3$, first note that $g(\lambda_P,\lambda_P;\Lambda_2) > g(\Lambda_2,\Lambda_2;\Lambda_2) = 0$ by Lemma 4.2(a) since $\lambda_P < \Lambda_2 < \lambda_F$. Since $g(\lambda_P,\lambda_P,\Lambda)$ decreases in Λ it follows that $\Lambda_2 < \Lambda_3$.

Parts 1(a)-(c). The above analysis implies for $\lambda \in [\lambda_P, \min(\Lambda,\lambda_F)]$ the function $g(\lambda,\lambda;\Lambda)$ is non-negative if $\Lambda \leq \Lambda_2$, as in *1(a)*, and strictly negative if $\Lambda > \Lambda_3$, as in *1(c)*. For *1(b)*, if $\Lambda \in (\Lambda_2,\Lambda_3)$ then $g(\lambda,\lambda;\Lambda)$ has an unique root $\lambda_1 \in [\lambda_P, \min(\Lambda,\lambda_F)]$ since $g(\lambda_P,\lambda_P;\Lambda) > 0 > g(\min(\Lambda,\lambda_F), \min(\Lambda,\lambda_F);\Lambda)$. Solving $g(\lambda_1,\lambda_1;\Lambda) = c_{\min} - f_h\left(c_h\left(\overline{\lambda}_h(\lambda_1);\Lambda\right)\right) = 0$ yields (57)-(58).

Part 2. The thresholds $\underline{\Lambda}_{ml}$ and $\overline{\Lambda}_{ml}$ determine the sign of $g(\lambda, \min(\lambda,\lambda_F);\Lambda)$ for $\lambda = \min(\Lambda,\mu)$ when $\Lambda > \lambda_F$. When $\Lambda > \lambda_F$ and $\lambda = \min(\Lambda,\mu)$, the maximum feasible total rate of h and m classes is λ_F: $\lambda_{mh} = \min(\lambda_F,\lambda) = \min(\lambda_F, \min(\Lambda,\mu)) = \lambda_F$. By (54) the virtual delay cost difference is

$$g(\min(\Lambda,\mu),\lambda_F;\Lambda) = \begin{cases} g(\Lambda,\lambda_F;\Lambda) = f_l\left(F^{-1}\left(\frac{\Lambda-\lambda_F}{\Lambda}\right)\right) - c_{\max}, & \Lambda \in [\lambda_F,\mu], \\ g(\mu,\lambda_F;\Lambda) = f_l\left(F^{-1}\left(\frac{1}{d\Lambda}\right)\right) - c_{\max}, & \Lambda \geq \mu. \end{cases} \tag{87}$$

The function $g(\Lambda,\lambda_F;\Lambda)$ increases in $\Lambda \in [\lambda_F,\mu]$ as the fraction $1-\lambda_F/\Lambda$ in l classes increases, $g(\mu,\lambda_F;\Lambda)$ decreases in $\Lambda \geq \mu$ as the fraction $1/(d\Lambda)$ in l classes decreases, and $\lim_{\Lambda\to\infty} g(\mu,\lambda_F;\Lambda) = c_{\min} - c_{\max} < 0$. Hence $g(\min(\Lambda,\mu),\lambda_F;\Lambda)$ has an unique maximum for $\Lambda \geq \lambda_F$ at $\Lambda = \mu$ where

$$g(\mu,\lambda_F,\mu) = f_l\left(\overline{F}\left(\frac{1}{d\mu}\right)\right) - c_{\max} > 0 \Leftrightarrow F(f_l^{-1}(c_{\max}))\cdot d < \mu^{-1}. \tag{88}$$

Part 2(a). It follows that if $F(f_l^{-1}(c_{\max}))\cdot d < \mu^{-1} < d$ then $\underline{\Lambda}_{ml} \in (\lambda_F,\mu)$ is the unique solution of $g(\Lambda,\lambda_F;\Lambda) = 0$ and $\overline{\Lambda}_{ml} > \mu$ is the unique solution of $g(\mu,\lambda_F;\Lambda) = 0$, where $\underline{\Lambda}_{ml}$ and $\overline{\Lambda}_{ml}$ satisfy (59)-(60). Furthermore, $g(\Lambda,\lambda_F;\Lambda) > 0$ for $\Lambda \in (\underline{\Lambda}_{ml},\mu)$ and $g(\mu,\lambda_F;\Lambda) > 0$ for $\Lambda \in [\mu,\overline{\Lambda}_{ml})$. Because by Lemma 4.2(b) the function $g(\lambda,\lambda_F;\Lambda)$ satisfies $g(\lambda_F,\lambda_F;\Lambda) < 0$ and increases in $\lambda \geq \lambda_F$ for fixed $\lambda_F < \Lambda$, it has an unique root λ_3 if $\Lambda \in [\underline{\Lambda}_{ml},\overline{\Lambda}_{ml})$. Solving $g(\lambda_3,\lambda_F;\Lambda) = 0$ yields (61)-(62).

If $\Lambda \notin [\underline{\Lambda}_{ml}, \overline{\Lambda}_{ml})$, the above discussion and Lemma 4.2(b) imply that $g(\lambda, \lambda_F; \Lambda) < 0$ for $\lambda \geq \lambda_F$.

Part 2(b). If $\mu^{-1} \leq F(f_l^{-1}(c_{\max})) \cdot d$ then the above analysis of $g(\min(\Lambda, \mu), \lambda_F; \Lambda)$ and (88), together with Lemma 4.2(b) again imply that $g(\lambda, \lambda_F; \Lambda) < 0$ for all feasible $\lambda \geq \lambda_F$. □

Proof of Lemma 6. By Lemma 4.1, for fixed $\Lambda > \lambda_P$ the virtual delay cost difference $g(\lambda, \lambda_P; \Lambda)$ increases in λ. Hence for fixed Λ the function $g(\lambda, \lambda_P; \Lambda)$ has at most one root in $\lambda \in [\lambda_P, \min(\Lambda, \mu)]$ which the proof characterizes.

The threshold $\Lambda_3 > \lambda_P$ is defined in Lemma 5 and determines the sign of $g(\lambda, \lambda_P; \Lambda)$ for $\lambda = \lambda_P$, where $g(\lambda_P, \lambda_P; \Lambda_3) = 0$ and $g(\lambda_P, \lambda_P; \Lambda) > 0 \, (< 0)$ if $\Lambda < (>) \Lambda_3$.

The threshold Λ_4 defined in (63) determines the sign of $g(\lambda, \lambda_P; \Lambda)$ for $\lambda = \min(\Lambda, \mu)$ where

$$g(\min(\Lambda, \mu), \lambda_P; \Lambda) = \begin{cases} g(\Lambda, \lambda_P; \Lambda) = f_l\left(F^{-1}\left(\frac{\Lambda - \lambda_P}{\Lambda}\right)\right) - f_h\left(\overline{F}^{-1}\left(\frac{\lambda_P}{\Lambda}\right)\right), \ \Lambda \in [\lambda_P, \mu], \\ g(\mu, \lambda_P; \Lambda) = f_l\left(F^{-1}\left(\frac{\mu - \lambda_P}{\Lambda}\right)\right) - f_h\left(\overline{F}^{-1}\left(\frac{\lambda_P}{\Lambda}\right)\right), \ \Lambda \geq \mu. \end{cases} \tag{89}$$

For $\Lambda \leq \mu$ and $\lambda = \min(\Lambda, \mu)$ we have $g(\Lambda, \lambda_P; \Lambda) > 0$: when all types are served the segments with high and low lead time qualities have the *same* marginal type $c_l(\Lambda - \lambda_P; \Lambda) = c_h(\lambda_P; \Lambda) = \overline{F}^{-1}(\lambda_P/\Lambda)$, and (14)-(15) imply that $f_l(c) > f_h(c)$ for $c \in [c_{\min}, c_{\max}]$. The threshold Λ_4 is the unique solution of $g(\mu, \lambda_P; \Lambda) = 0$ in $\Lambda \in [\mu, \infty)$, because $g(\mu, \lambda_P; \Lambda)$ strictly decreases in Λ with $\lim_{\Lambda \to \infty} g(\mu, \lambda_P; \Lambda) < 0$. Noting that $\mu - \lambda_P = \sqrt{\mu/d}$ and solving for Λ_4 yields (63).

To summarize, for $\lambda = \min(\Lambda, \mu)$ we have $g(\lambda, \lambda_P; \Lambda) > 0$ if $\Lambda < \Lambda_4$, and $g(\lambda, \lambda_P; \Lambda) \leq 0$ if $\Lambda \geq \Lambda_4$.

To prove $\Lambda_3 < \Lambda_4$ we show $g(\lambda, \lambda_P; \Lambda_3) > 0$ for $\lambda = \min(\Lambda_3, \mu)$. This follows since $g(\lambda_P, \lambda_P; \Lambda_3) = 0$ and $\Lambda_3 > \lambda_P$ as noted above, and because $g(\lambda, \lambda_P; \Lambda)$ increases in $\lambda \geq \lambda_P$ by Lemma 4.1.

To prove $\overline{\Lambda}_{ml} < \Lambda_4$, since $\Lambda_4 > \mu$ we show for the nontrivial case $\overline{\Lambda}_{ml} > \mu$ that $g(\mu, \lambda_P; \overline{\Lambda}_{ml}) > 0$. Recall that $g(\mu, \lambda_F; \overline{\Lambda}_{ml}) = 0$ as defined in Lemma 5.2, and that for fixed λ and Λ the virtual delay cost difference $g(\lambda, \lambda_{mh}; \Lambda)$ decreases in the rate λ_{mh} allocated to h and m classes. Setting $\lambda = \mu$ and $\Lambda = \overline{\Lambda}_{ml}$ and noting that $\lambda_P < \lambda_F$ implies that $g(\mu, \lambda_P; \overline{\Lambda}_{ml}) > g(\mu, \lambda_F; \overline{\Lambda}_{ml}) = 0$.

Parts 1-3. The claims follow from the established properties of Λ_3 and Λ_4, combined with the fact of Lemma 4.1 that $g(\lambda, \lambda_P; \Lambda)$ increases in λ for fixed $\Lambda > \lambda_P$. □

Proof of Lemma 7. The claims on the slope of $f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right)$ hold since $f_h' c_h' < 0$, and $\overline{\lambda}_h(\lambda)$ increases in $\lambda < \lambda_P$ and decreases in $\lambda \in (\lambda_P, \lambda_F)$ by (40). In Part 1 the values of $f_h(c_h(\lambda))$ for $\lambda = 0$ and $\lambda = \min(\lambda_P, \Lambda)$ follow as $c_h(\lambda) = \overline{F}^{-1}(\lambda/\Lambda)$ and $f_h(c_{\max}) = c_{\max}$. In Part 2 the fact that $f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right) = c_{\max}$ for $\lambda = \lambda_F \leq \Lambda$ holds since $\overline{\lambda}_h(\lambda_F) = 0$ by (40) and $c_h(0) = c_{\max} = f_h(c_{\max})$. □

Proof of Lemma 8. Parts 1-3 follow by Lemma 7 and since $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right)$ increases in Λ.

Part 3. For fixed Λ, if $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right) \geq 0$ at $\lambda = \min(\lambda_P, \Lambda)$, then Lemma 7 implies that $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right) \geq 0$ throughout. This condition holds if and only if $\Lambda \geq \overline{\Lambda}_{sd}$: By (66) in Lemma 7, for $\lambda = \min(\lambda_P, \Lambda)$ this virtual delay cost equals $f_h(c_{\min}) < 0$ if $\Lambda \leq \lambda_P$, otherwise it equals $f_h(\overline{F}^{-1}(\lambda_P/\Lambda))$ and increases in $\Lambda > \lambda_P$, and $\lim_{\Lambda \to \infty} f_h(c_h(\lambda_P; \Lambda)) = f_h(c_{\max}) = c_{\max} > 0$. Solving for Λ sucht that $f_h(\overline{F}^{-1}(\lambda_P/\Lambda)) = 0$ yields $\overline{\Lambda}_{sd} > \lambda_P$ as in (69).

Part 1. For fixed Λ, if $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right) < 0$ at $\lambda = \min(\lambda_F, \Lambda)$, then Lemma 7 implies that $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right)$ has an unique root λ_0, where $\lambda_0 < \lambda_P$ and $f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right)$ is nonnegative iff $\lambda \leq \lambda_0$. The condition holds if and only if $\Lambda < \underline{\Lambda}_{sd}$: we have

$$f_h\left(c_h\left(\overline{\lambda}_h(\lambda); \Lambda\right)\right)\Big|_{\lambda = \min(\lambda_F, \Lambda)} = \begin{cases} f_h(c_h(\Lambda; \Lambda)) = f_h(c_{\min}) < 0, & \Lambda \leq \lambda_P, \\ f_h\left(c_h\left(\overline{\lambda}_h(\Lambda); \Lambda\right)\right) = f_h(\overline{F}^{-1}(\overline{\lambda}_h(\Lambda)/\Lambda)), & \Lambda \in (\lambda_P, \lambda_F), \\ f_h\left(c_h\left(\overline{\lambda}_h(\lambda_F); \Lambda\right)\right) = c_{\max} > 0, & \Lambda \geq \lambda_F, \end{cases} \tag{90}$$

where $f_h\left(c_h\left(\overline{\lambda}_h(\Lambda);\Lambda\right)\right) = f_h(\overline{F}^{-1}(\overline{\lambda}_h(\Lambda)/\Lambda))$ increases in $\Lambda \in (\lambda_P, \lambda_F)$ since $\overline{\lambda}_h(\Lambda)/\Lambda$ decreases in Λ. Hence $\underline{\Lambda}_{sd} \in (\lambda_P, \lambda_F)$ is the unique solution of $f_h\left(c_h\left(\overline{\lambda}_h(\Lambda);\Lambda\right)\right) = 0$ and satisfies (68). The fact that $\underline{\Lambda}_{sd} < \overline{\Lambda}_{sd}$ follows by the properties of $\overline{\Lambda}_{sd}$.

Part 2. If $\Lambda \in [\underline{\Lambda}_{sd}, \overline{\Lambda}_{sd})$ then Parts 1 and 3 together with Lemma 7 imply the stated properties.

It remains to rank $\underline{\Lambda}_{sd}$ and $\overline{\Lambda}_{sd}$ relative to Λ_2 and Λ_3, which are defined in Lemma 5. To see that $\underline{\Lambda}_{sd} \leq \Lambda_2$, recall that $g(\Lambda_2, \Lambda_2; \Lambda_2) = 0$ which is equivalent to $f_h\left(c_h\left(\overline{\lambda}_h(\Lambda_2);\Lambda_2\right)\right) = c_{\min}$. The ranking follows since $f_h\left(c_h\left(\overline{\lambda}_h(\underline{\Lambda}_{sd}), \underline{\Lambda}_{sd}\right)\right) = 0 \leq c_{\min}$ and $f_h\left(c_h\left(\overline{\lambda}_h(\Lambda);\Lambda\right)\right)$ increases in $\Lambda \in [\lambda_P, \lambda_F]$. To see that $\overline{\Lambda}_{sd} \leq \Lambda_3$ recall that $g(\lambda_P, \lambda_P; \Lambda_3) = 0$ which is equivalent to $f_h(c_h(\lambda_P; \Lambda_3)) = c_{\min}$. The ranking follows since $f_h\left(c_h\left(\lambda_P, \overline{\Lambda}_{sd}\right)\right) = 0 \leq c_{\min}$ and $f_h(c_h(\lambda_P; \Lambda))$ increases in Λ. □