Optimal Price-Lead Time Menus for Queues with Customer Choice: Priorities, Pooling & Strategic Delay

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

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

August 29 2013

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 two dimensions, 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. *Pooling* types with different delay costs into a single class; 2. *Pricing out the middle of the delay cost spectrum* while serving both ends; and 3. *Strategic delay* to deliberately inflate lead times.

1. Introduction

A range of service and manufacturing 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. In this common scenario, customers choose among all menu options based on their own self-interest, and the provider's decisions must account for their choice behavior. We study this problem within a queueing model and consider customer heterogeneity in two dimensions, their *valuations* for instant delivery and their *delay cost rates*. While this problem has recently received some attention, as detailed below, significant gaps remain in understanding its solution for the case with *multiple delay cost rates* considered here. We address the following questions on the optimal menu, scheduling policy and customer segmentation.

1. *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?

2. *Customer segmentation.* Which customer types – most, least or moderately time-sensitive – should be served, and which, if any, should be pooled or strategically delayed?

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

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 do 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 further explained in Section 2.3.

Several papers study variations of the classic price-quality design problem of MR. Dana and Yahalom (2008) introduce a capacity constraint. Bansal and Maglaras (2009) consider a capacity constraint and customers with satisficing, not utility-maximizing, behavior. Both studies ignore queueing effects and neither reports the phenomena we identify. Quality degradation similar to strategic delay has been studied in the damaged goods literature (Deneckere and McAffee, 1996; McAffee, 2007). Anderson and Dana (2009) unify this discussion by identifying a necessary condition for price discrimination to be profitable including in the presence of a quality constraint but ample capacity. We discuss these connections to our strategic delay results in Section 6.4.

Three problem characteristics jointly distinguish this from most papers on price and lead time optimization. *Revenue maximization*: the objective is to maximize the provider's revenue or profits, not the total system benefit (cf. Mendelson and Whang, 1990). *Customer choice over menu options*: types have private information on their preferences and can choose their preferred service class, unlike settings where a single class can be targeted to each type (cf. Maglaras and Zeevi, 2005; Boyaci and Ray, 2003; Zhao et al. 2012). *Scheduling optimization*: the provider also chooses the scheduling policy instead of only optimizing prices for a *given* policy (cf. Naor, 1969; Mendelson, 1985; Rao and Petersen, 1998). In studies that lack one of these ingredients *neither pooling nor strategic delay can be optimal*. Finally, We consider static price-lead time menus, unlike papers on dynamic price and/or lead time quotation such as Plambeck (2004). Hassin and Haviv (2003) and Stidham (2009) survey research on queueing systems with self-interested and time-sensitive customers.

Only a few papers share these three attributes. Like this paper, they study static price-lead time menus for customer types that are heterogeneous in their valuations for instant delivery and in their linear delay costs: the *net* value of a type with valuation v and delay cost c for expected lead time w equals $v - c \cdot w$. Afèche (2004, 2013), henceforth AF, analyzes strategic delay in a model with two delay costs.[1] Katta and Sethuraman (2005), henceforth KS, provide a partial analysis of pooling in

[1] Yahalom et al. (2006) consider a version of this model with convex increasing delay cost functions and fixed demand rates.

a model with *multiple delay costs* but *lead time-invariant ranking of types*. We further discuss the relationship to these papers in Sections 6.2 and 6.5. Our contributions are as follows.

1. *Generalized customer type model.* The model we propose is distinctive in that it *jointly* considers *multiple delay costs*, unlike AF, and *lead time-dependent ranking of types*, unlike KS. It yields novel insights and offers an unifying framework for explaining the presence or absence of these phenomena in related studies. We return to these connections in Section 6.
2. *Insights on the optimal menu and segmentation.* We provide necessary and sufficient conditions, in terms of the capacity, market size and properties of the (v,c)-distribution, for three striking features of the optimal menu and segmentation. (*i*) *Pooling* types with different delay costs; our results are more general and more informative than those in KS. (*ii*) *Pricing out the middle of the delay cost spectrum* while serving both ends; this feature arises neither in AF nor in KS. (*iii*) *Strategic delay* to artificially inflate lead times; our results complement those of AF.

2. Model, Problem Formulation and Analysis Roadmap

2.1. Model

We model a make-to-order service or manufacturing operation as an $M/M/1$ queueing system. A pool of potential customers with unit demand arrive according to an exogenous Poisson process with rate or market size Λ per unit time. The system has i.i.d. exponential service times with mean $1/\mu$, where μ is the capacity or service rate. 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.

Customer preferences. Customers differ in two attributes, their valuations or willingness to pay for immediate delivery and their linear delay cost rates. We use the terms "delay" and "lead time" interchangeably to refer to the entire time interval between order placement and delivery. We consider a continuum of customer *types* indexed by c, which denotes the customer's delay cost rate and measures her (constant) disutility 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 a positive valuation $V(c)$ for immediate delivery of the product or service, 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 6.3 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 of 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 for every type, 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 known as 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 $w > d$ decrease, while those for lead times $w < d$ increase, in the time-sensitivity c. In this sense, the parameter d represents a threshold lead time. In the affine model of $V(c)$ this threshold is common for all types, however, we explore relaxing this assumption with concave and convex forms of $V(c)$ in Section 6.3. This feature plausibly describes situations where impatient customers get very little utility from a product or service with long delay while more patient customers with a smaller budget have lower willingness-to-pay for speedy service but are willing to pay more than impatient customers for somewhat delayed delivery. For example, an impatient customer may be willing to pay a lot for overnight delivery but only very little if it takes two or three days, while a more patient customer is not willing to buy overnight delivery but will pay more than impatient customers for 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 structure. The arrival process, the delay cost rate distribution f, the function $V(c)$ and the service time distribution are known to the provider. A customer's type c is her private information. However, a customer does not know her exact service time when making her purchase decision; it only becomes known – to her and to the provider – once her job is processed to completion. Only the provider observes the system state; customers lack this information when making their decisions and base their lead time forecasts on the posted mean lead times.

Provider and customer decisions. The provider's objective is to choose a price-lead time menu and a scheduling policy to maximize her long run average revenue rate. We study the system in steady-state under a static menu of service classes, each characterized by a per job price and an expected steady-state delay. We simply say "lead time", "mean delay" or "delay" when referring to the expected steady-state delay of a service class. Since types are private information the provider must let customers choose among all classes and consider their choice behavior in designing her menu. Customers are strategic in deciding whether to place an order and in choosing their class upon arrival at the facility. However, they do not choose their arrival time, do not affect the subsequent arrival process if they do not place an order, and cannot later renege if they place an order.

To formalize this revenue management problem, 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 $P : \mathcal{W} \to \mathbb{R}$ the function which assigns a price for service at a given lead time. Upon arrival at the facility a customer of type c determines the lead time-price pair $(w, P(w))$ which maximizes her expected utility from service $U(w, P(w); c) \triangleq v + c \cdot (d - w) - P(w)$. Formally, denote by $w(c) \triangleq \arg\max_{w \in \mathcal{W}} \{U(w, P(w); c)\}$ and $p(c) \triangleq P(w(c))$ 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 her 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 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.

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 service as discussed in Section 3.

Lead times and scheduling policy. We assume that customers are risk neutral with respect to lead time uncertainty and that they base their decisions on the *announced* mean delays, since they generally possess neither the required information (about queue lengths, scheduling policy, arrival rates, etc.) nor the analytical sophistication to reliably forecast their actual delays at the time of their order decision. However, the provider is committed to ensure that the announced lead times equal the realized mean delays given the capacity μ, the scheduling policy, and customers' equilibrium purchase decisions. We consider the following set of admissible scheduling policies:

1. We focus on nonanticipative and regenerative policies. This appears to be the most general, easily describable restriction under which the existence of long-run delay averages may be verified.
2. We do not restrict attention to work conserving policies. In particular, we allow the insertion of *strategic delay* whereby the provider artificially increases the mean lead times for a subset of service classes above the levels that are operationally achievable given the system utilization. See Afèche (2004, 2013) for a detailed discussion of strategic delay and its implementation through the control of server idleness, server speed and/or delivery delays of completed jobs.
3. We allow preemption, which does not affect results but simplifies the analysis: the lead time of a given class only depends on the arrival rates to higher priority classes, not on the total rate.

We use the achievable region approach, cf. Coffman and Mitrani (1980). Instead of determining an optimal scheduling policy, we solve the equivalent problem of finding the optimal mean lead times in the set of mean lead times that are achievable by admissible policies. The problem formulation below specifies the achievable region for the class of scheduling policies described by 1.-3.

2.2. Mechanism Design Formulation

A lead time-price menu $\{(w, P(w)) : w \in \mathcal{W}\}$ induces customers to make purchase decisions and service class choices that maximize their expected utility and are characterized by the triple of functions (a, w, p) described above. It is analytically convenient to view the provider's problem as a mechanism design problem. Based on the revelation principle, mechanism design problems restrict attention w.l.o.g. to direct revelation mechanisms in which each customer directly reports her type to the provider who then allocates products and charges customers following previously announced rules. The procedure described above is strictly speaking not a direct revelation mechanism – customers reveal their types only indirectly, but it is more descriptive of how services are sold. It is also *de facto* equivalent to a direct revelation mechanism in which all customers truthfully reveal their types. This requires that (a, w, p) satisfy the *individual rationality* (IR) and *incentive-compatibility* (IC) constraints. The IR constraints require 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 $\in \overline{\mathcal{C}}_a$. The IC constraints require that each type c maximizes her expected utility if it truthfully report its type: $U(c) \geq U(w(c'), p(c'); c)$ for $c \neq c'$.

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

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

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

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

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

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

The 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 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 FIFO $M/M/1$ system with arrival rate $\Lambda\int_{x\in s} f(x)dx$ and service rate μ. A scheduling policy is *work conserving* if (3) is binding for $s=\mathcal{C}_a$. The constraints (4)-(5) capture the IR and (6) the IC constraints. 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.

2.3. 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.

While STEP 1 is based on standard mechanism design methods, STEPS 2 and 3 are *not.* The capacity constraint and queueing delays introduce operational interdependencies among lead times, which significantly complicates 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 queueing related operational constraints (3) in Problem 1 and introduces instead a quality cost function in the objective function[2], 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 account for these effects.

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

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: (i) it is an interval that includes $c_{\min}$, i.e., $c < c_l \Rightarrow c \in C_l$; (ii) 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: (i) it is an interval that includes $c_{\max}$, i.e., $c > c_h \Rightarrow c \in C_h$; (ii) prices and expected utilities from service satisfy:*

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

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

4. *If there is a segment of types who buy the medium lead time ($C_m \neq \varnothing$) then: (i) $C_m \subset [c_l, c_h]$; (ii) 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., $(c_l, c_h) \subset C_m \cup \overline{\mathcal{C}}_a$, and*

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

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 for a given lead time w as a function of the type c, that is, $N(c,w) = V(c) - cw$. Part 1 follows because $N(c,w)$ decreases in lead time. For parts 2-5, the customer segmentation under a price-lead time menu that satisfies the IR and IC constraints hinges on the fact that the net value $N(c,w)$ for fixed lead time w changes at the rate $V'(c) - w = d - w$ as the customer type increases, where d and w capture the higher type's increase in valuation and delay cost, respectively. Therefore, if a type c chooses to buy a lead time $w(c) > d$, then more patient types $c' < c$ are willing to pay more for $w(c)$ than c. If IR constraints are satisfied for c, then utility must be strictly positive for more patient types, i.e. they must be served. This

implies part 2 of Proposition 1. The set C_l is an interval that includes the least time-sensitive type c_{min}. By (8) and (9), the price $p(c)$ for a type $c < c_l$ decreases and its utility increases with the lead times of more impatient types $x \in (c, c_l)$. The longer these lead times the more attractive they are to type c, compared to more impatient types. Similarly, for part 4 of Proposition 1, if a type c buys a lead time $w(c) < d$, then more impatient type $c' > c$ must be admitted with strictly positive utility; therefore, the set C_h is an interval that includes the most time-sensitive type c_{max}. However, in contrast to the prices (8) that are decreasing in lead times, by (10) and (11), the price $p(c)$ for a type $c > c_h$ increases, and its utility decreases in the lead times of more patient types $x \in (c_h, c)$. This holds because the longer these lead times, the less attractive they are to type c, compared to more patient types. By parts 4 and 5, the set of customers C_m that purchase the threshold lead time d need not be an interval; it is a subset of $[c_l, c_h]$. If $C_m \neq \varnothing$, then 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. Therefore it may be optimal to price the most, the least or only moderately impatient types out of the market.

By Proposition 1 choosing C_l, C_m and C_h reduces to choosing the corresponding demand rates. Let λ_l, λ_m and λ_h be the rates for l, m, and h classes, respectively, and $\lambda \triangleq \lambda_l + \lambda_m + \lambda_h$ where $\lambda_l = \Lambda F(c_l)$, $\lambda_h = \Lambda \overline{F}(c_h)$. If $\lambda_m > 0$ then *different types are pooled*. The admission function a determines c_l, c_h and $\lambda_l = \Lambda F(c_l)$, $\lambda_h = \Lambda \overline{F}(c_h)$ and $\lambda_m = \lambda - \lambda_l - \lambda_h$. While $a(c) = 1$ for $c < c_l$ and $c > c_h$, it is not uniquely determined for $c \in [c_l, c_h]$ and must only satisfy $\lambda_m = \Lambda \int_{c_l}^{c_h} a(x) f(x)\, dx$. If $C_m \neq \varnothing$ and $\lambda_m < \Lambda - \lambda_l - \lambda_h$ then only a fraction $\lambda_m / [\Lambda - \lambda_l - \lambda_h] < 1$ of types in $[c_l, c_h]$ buy the medium lead time. Any feasible triples (a, w, p) and (a', w, p) with $a \neq a'$ but the same *mass* of types $c \in [c_l, c_h]$ who buy the medium lead time are revenue equivalent.[3] Any such equilibrium has positive masses of zero-utility customers who buy and do not buy the service. While this equilibrium structure arises in our model because $V(c)$ is affine, note that our main structural insights on the optimal segmentation and menu remain valid for a broader class of $V(c)$ functions.

4. Optimal Segmentation and Lead Times for Fixed Arrival Rate

4.1. Reduced Problem, Virtual Delay Costs, and Solution Preview

Based on Proposition 1 we drop the pair (a, p) from the problem. Let $\Pi(\lambda_l, \lambda_h, \lambda, w)$ denote the expected revenue rate as a function of the total demand 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:

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

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

[3] In our model a and a' describe different pure strategies. This equilibrium can also be supported if every type $c \in [c_l, c_h]$ requests the lead time d but the provider only accepts a fraction $\lambda_m / [\Lambda - \lambda_l - \lambda_h] < 1$.

We call f_l and f_h the *virtual delay cost functions*: they measure the improvement in total revenue in response to a reduction in the lead time of a given type. Which of the virtual delay costs applies to a type c depends on its lead time: it is $f_l(c)$ if $w(c) > d$ and $f_h(c)$ if $w(c) < d$. The virtual delay cost of a type c consists of an *own effect* and an *external effect*. The own effect is the first summand in (14) and (15), the delay cost c; it simply measures how a decrease in that type's lead time $w(c)$ allows an increase in its own price. The external effect is the second summand in (14) and (15); it measures the aggregate revenue improvement from price changes necessary to maintain IC for classes targeted to other types when the lead time $w(c)$ is decreased. The external effect is positive for $f_l(c)$ but *negative* for $f_h(c)$, so $f_l(c) > c > f_h(c)$ for $c \in (c_{\min}, c_{\max})$ and $f_h(c)$ may be *negative*. Decreasing $w(c) > d$ allows price increases on classes targeted to more patient types $c' < c$ while maintaining IC, because the willingness to pay for lead times larger than d increases in customer patience (as explained in Section 3 and shown in (8), the price of a type c in C_l is decreasing in the lead times of more impatient types in C_l). In contrast, decreasing $w(c) < d$ requires price *increases* on classes targeted to more impatient types $c' > c$ because the willingness to pay for lead times shorter than d increases with impatience (as explained in Section 3 and shown in (10), the price of a type c in C_h is increasing in the lead times of more patient types in C_h).

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

1. *Pricing out the middle of the delay cost spectrum.* As outlined in Section 3, 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 types $c_1 < c_2$.

   Note that *every* type distribution *can* satisfy the necessary condition for pooling at the lead time threshold d: since $f_l(c) > c > f_h(c)$ for $c \in (c_{\min}, c_{\max})$, it cannot be optimal to sell different lead times, one larger and the other smaller than d, to types $c_1 < c_2$ with *similar* delay costs, for then $f_l(c_1) > f_h(c_2)$: For sufficiently similar types, the external effects of their virtual delay costs mainly determine how their lead times affect the revenue. Reducing the lead time of c_1 and increasing that of c_2 allows the provider to increase prices for *all* types $c' \leq c_1$ and $c'' > c_2$, whereas the price increase for c_1 and the price reduction for c_2 are relatively insignificant.

   By contrast, many common probability distributions do not satisfy the necessary conditions for pooling at $w \neq d$. If f is log-concave then $f_l' > 0$ and $f_h' > 0$, 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 nonnmonotone 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 $f_l', f_h' > 0$.

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 2.3. Let $\lambda_l^*(\lambda)$, $\lambda_m^*(\lambda)$ and $\lambda_h^*(\lambda)$ denote the optimal customer segmentation, and the function $w^*(c;\lambda)$ the optimal lead times as a function of the total 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, where $\lambda_l^*(\lambda) > 0 \Leftrightarrow c_l^*(\lambda) > c_{\min}$ and $\lambda_h^*(\lambda) > 0 \Leftrightarrow c_h^*(\lambda) < c_{\max}$, and $C_l^*(\lambda)$, $C_m^*(\lambda)$ and $C_h^*(\lambda)$ for the corresponding sets of types buying low, medium and high lead time qualities, respectively.

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:*

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

   (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 $f_l(c_l^*(\lambda)) = f_h(c_h^*(\lambda))$.*

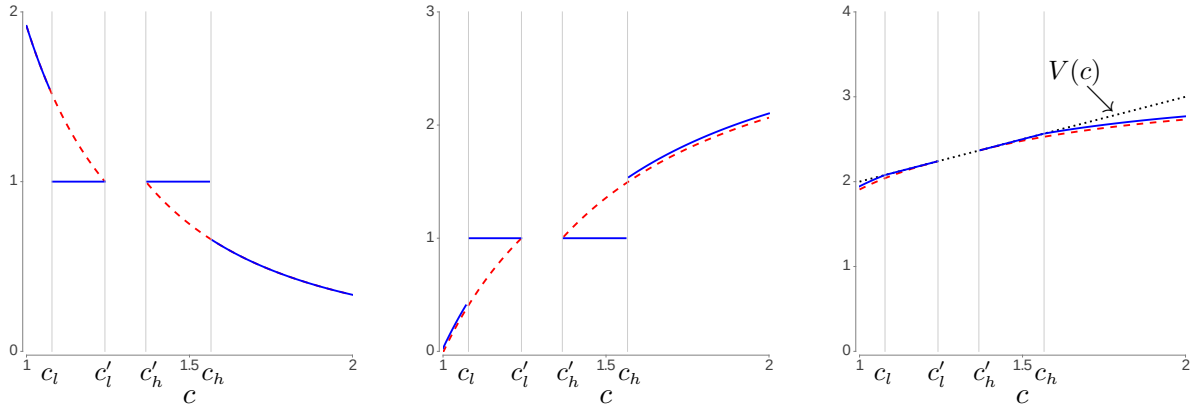
Lemma 1 provides a set of simple rules to determine the optimal lead time menu for a particular arrival rate. First, Lemma 1-1a limits pooling to types that buy the lead time d, i.e., the types in the set $C_m^*(\lambda)$. Customers in $C_h^*(\lambda)$ receive strict priority over types in $C_m^*(\lambda)$ which receive strict priority over types in $C_l^*(\lambda)$. Different types within $C_l^*(\lambda)$ and $C_h^*(\lambda)$ buy different lead times and are strictly prioritized in the order of their delay cost. Different types within $C_m^*(\lambda)$ buy the same lead time and are pooled into a single FIFO service class. Second, Lemma 1-1b limits strategic delay to the case where no customers are served in the low lead time quality segment. Third, Lemma 1-2a rules out serving customers with negative virtual delay costs in the high quality segment; if it is operationally feasible to serve such customers with lead time lower than d, they should instead be offered the lead time s, which yields strategic delay. Finally, Lemma 1-2b 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 rate as a function of $(\lambda_l, \lambda_h, \lambda)$:

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

From Proposition 1 and (16), Problem 1 is reformulated to the following program.

Figure 1 **Solution 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 denoted with solid, strictly prioritized menu dashed. $\mu=3, \Lambda=2, \lambda=1.75, v=1, d=1$ and $f(c)$ uniform with $c_{min}=1, c_{max}=2$.

Problem 2.

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

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

$$\lambda < \mu,$$

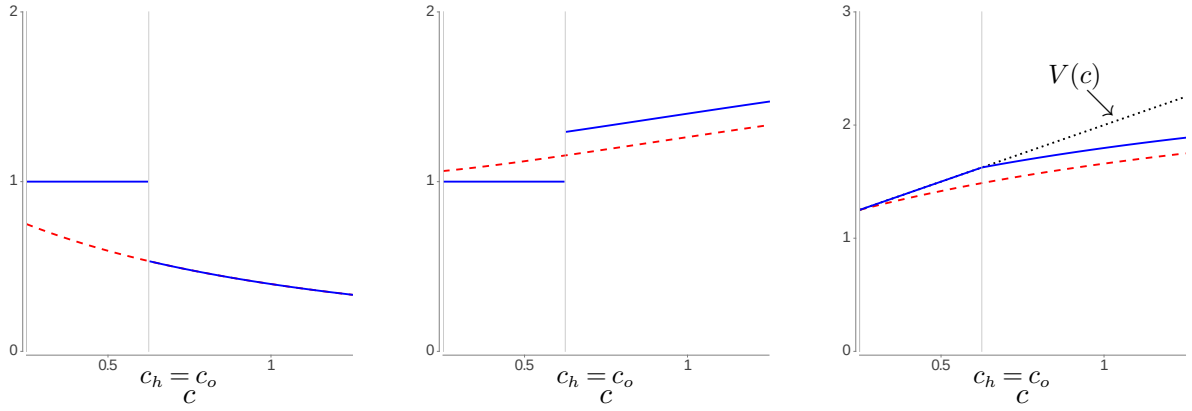
$$\frac{\mu}{(\mu-[\lambda-\lambda_l])(\mu-\lambda_h)} \leq d \text{ if } \lambda-\lambda_l>0, \tag{20}$$

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

The 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, $\lambda_h=\lambda_m=0$. It follows from Lemma 1 that neither pooling nor strategic delay are optimal.

To illustrate Lemma 1 consider Figure 1 which shows an example with capacity $\mu>1/d$ and fixed arrival rate $\lambda=1.75<\Lambda=2$. First consider the first-best benchmark, denoted by SP: it strictly prioritizes all types in the order of their delay costs and operates a work conserving policy. The dashed line in Figure 1(a) shows the lead times under the SP menu. The type c_h' is the customer that receives $w(c)=d$ under strict priorities. This type is the lower bound on types which can be strictly prioritized into the high quality segment, that is, with lead time less than d. For $\lambda=1.75$, the lead times of some types under the SP menu exceed d, that is, they must be admitted into the low quality segment ($w>d$). The net value for such lead times decreases in impatience, so it is optimal to serve the most patient remaining types, i.e., $[c_{min},c_l']$. Since $\lambda<\Lambda$ there is a remaining interval of intermediate types (c_l',c_h') who are priced out. However, the SP menu is not optimal if types are not observable. In particular, $f_l(c_l')>f_h(c_h')$ which violates the necessary condition for optimality provided in Lemma 1-2(b). Recall from Section 4.1 that $f_l(c)$ and $f_h(c)$ have identical *own effects* c however, the *external effect* of $f_l(c)$ is positive whereas that of $f_h(c)$ is negative; therefore, close enough c_l' and c_h' will have $f_l(c_l')>f_h(c_h')$. Intuitively, the external effects dominate the own effects so that speeding up c_l' at the expense of c_h' increases revenues. Starting with the SP menu, the optimal menu is obtained by pooling a set of customers (c_l,c_l') from the low quality segment with a set of

Figure 2 **Solution with low end of delay cost spectrum strategically delayed.**



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

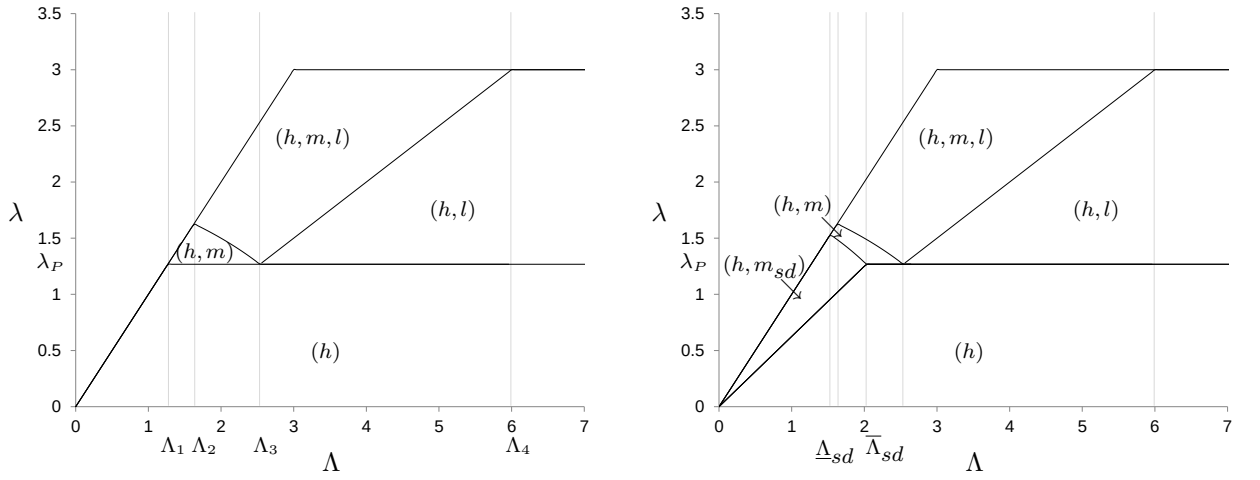
Note. Optimal menu denoted with solid, strictly prioritized menu dashed.
$\mu = 3, \Lambda = 1, \lambda = 1, v = 1, d = 1$ and $f(c)$ uniform with $c_{min} = 0.25, c_{max} = 1.25$.

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 SP lead times, the optimal lead times (solid lines) are lower for pooled types $(c_l, c'_l]$, higher for pooled types $[c'_h, c_h)$, but remain unchanged for the remaining strictly prioritized customers. Figure 1(b) shows the prices under the SP and optimal menus. Compared to the SP menu, increases in price in the optimal menu for $[c_{min}, c_l]$ is the result of the aggregate positive external effect from decreasing the lead times on $(c_l, c'_l]$. The price increases on customers $[c_h, c_{max}]$ similarly result from the negative external effect from *increasing* the lead time on $[c'_h, c_h)$. Figure 1(c) shows the full price $p(c) + cw(c)$ for both menus along with the value function $V(c) = v + cd$. The difference between the value function and the full price is the customer's utility. The full prices of the optimal menu strictly exceed those of the SP menu, resulting in lower customer utility and higher provider revenue.

The example shown in Figure 2 differs in three important respects from the one shown in Figure 1. First, the market size is such that all customers can be served under SP with lead time lower than d (by Figure 2(b), all lead times under the SP menu are below $d = 1$); second, all customers are served, i.e. $\lambda = \Lambda$; third, the virtual delay costs are negative at the low end of the delay cost spectrum, i.e., $f_h(c_{min}) < 0$. For such types the external effect in $f_h(c)$ dominates the own 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 part 2(a) of Lemma 1, it is necessary for optimality that all types in the high quality segment have a non-negative virtual delay cost, i.e., $f_h(c_h) \geq 0$. As shown in Figure 2(a), compared to the SP menu, the optimal menu increases the lead times of customers in $[c_{min}, c_0]$ to $d = 1$, where the type c_0 satisfies $f_h(c_0) = 0$. The highest types in $[c_0, c_{\max}]$ are served with high lead time quality, while types in $[c_{\min}, c_0]$ are pooled into the medium quality class and the lead time d is attained by inserting *strategic delay*. Figure 2(b) shows that compared to the SP prices, optimal prices decrease for customers $[c_{min}, c_0)$, but the revenue loss is more than offset by price increases for customers $[c_0, c_{max}]$. Figure 2(c) shows again the loss in customer surplus.

Proposition 2 (in the appendix) specifies the optimal menu (i.e., the solution of Problem 2) as a function of the market size Λ and arrival rate λ. Figure 3 illustrates these results with examples for the case $\mu > 1/d$. It shows for each (Λ, λ) which lead times qualities are sold under the optimal menu, high(h), medium (m) and/or low (l). Proposition 2 gives analytical expressions for the market

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 $\mu = 3$, $d = 1$, uniform delay cost rate distribution with $c_{min} = 1$ and $c_{max} = 2.\mu = 3, v = 1, d = 1$ and $f(c)$ uniform with $c_{min} = 1, c_{max} = 2$ for panel (a) and $c_{min} = 0.25, c_{max} = 1.25$ for panel (b).

size thresholds Λ_1, Λ_2, Λ_3, Λ_4, $\underline{\Lambda}_{sd}$, $\overline{\Lambda}_{sd}$ in Figure 3. For instance, for the case shown in Figure 2, where $\Lambda = 2$ and $\lambda = 1.75$, Figure 3-(a) confirms all three quality classes (h, m, l) are present.

Pooling at $w = d$: Consider the following three zones in Figure 3(a), which illustrates the case of positive virtual delay costs. In the first zone, for 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) and artificially increasing their lead times to $w = d$ is suboptimal since virtual delay costs are positive. For higher arrival rates, $\lambda > \lambda_P$, there are two zones; one for smaller Λ, where customers are pooled, (h, m) and (h, m, l), and the other for larger Λ where customers are strictly prioritized, (h, l). This 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. As a result, for sufficiently large market size, the virtual delay costs of the respective boundary types under the SP menu satisfy $f_l(c'_l) \leq f_h(c'_h)$, so that pooling is suboptimal by Lemma 1.

Pricing out the middle: If 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 (h), (h, m) and (h, m_{sd}); the subscript sd denotes strategic delay), then it is optimal to price out some intermediate types while selling low and high lead time qualities; these are the regions (h, l) and (h, m, l) in Figure 3.

Strategic delay: The example shown in Figure 3-(b) features $f_h(c) < 0$ for $[c_{min} = 0.25, c_0 = 1.25]$. As a result, there is a region (h, m_{sd}) where strategic delay is optimal. For types c with $f_h(c) < 0$, targeting a high lead time quality ($w < d$) *lowers* overall revenue. These types should be targeted for a lead time $\geq d$, even if serving them with lead time $< d$ is operationally feasible. Strategic delay is optimal under two conditions, λ must be small relative to μ so that all customers can be offered $w < d$, but large relative to the market size Λ requiring admission of a wide range of customer types including those with negative virtual delay costs. These two conditions imply the two market size thresholds $\underline{\Lambda}_{sd} < \overline{\Lambda}_{sd}$ in Figure 3-(b): Strategic delay is optimal for all sufficiently large arrival rates if $\Lambda < \underline{\Lambda}_{sd}$, for no arrival rates if $\Lambda > \overline{\Lambda_{sd}}$ and only for intermediate rates if $\Lambda \in [\underline{\Lambda}_{sd}, \overline{\Lambda}_{sd})$.

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 2.3 and discuss the optimal segmentation and lead time menu at the *optimal* λ, building on Proposition 2 and the following result.

LEMMA 2. *Fix the market size* Λ. *Write* $\Pi^*(\lambda,\mu)$ *for the maximum revenue as a function of* λ *and* μ, *under the optimal segmentation.*

1. $\Pi^*(\lambda,\mu)$ *satisfies: (i)* $\Pi^*_\lambda(\lambda,\mu) \geq v$ *for every* (λ,μ) *if the optimal segmentation is* (h) *and* $\Pi^*_\lambda(\lambda,\mu) = v$ *if it is* (h,m_{sd}). *(ii)* $\Pi^*_{\lambda\lambda}(\lambda,\mu) \leq 0 \leq \Pi^*_{\lambda\mu}(\lambda,\mu)$ *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. *(i)* $\Pi^*_\lambda(\lambda,\mu)$ *is continuous in* (λ,μ). *(ii) 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 specified in Proposition 2 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 $\mu^{-1} < d$ and $v > 0$, then selling only high quality lead times, denoted by (h), or h classes together with the strategically delayed medium lead time class, (h,m_{sd}), can only be optimal if it is optimal to serve the entire market, i.e., $\lambda^*(v) = \Lambda$: under these segmentations each additional customer brings positive revenue and is served with lowest priority. 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* $d \leq \mu^{-1}$ *or* $v \leq 0$, *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.*

$f_h(c_{\min}) \geq 0$		$f_h(c_{\min}) < 0$	
$\Lambda \leq \Lambda_1$	*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*
$\Lambda \geq \Lambda_4$	*no pooling, h and l classes*	$\Lambda \geq \Lambda_4$	*no pooling, h and l classes*

In the case of nonnegative virtual delay costs ($f_h(c_{min}) \geq 0$), 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 this threshold; in this case there is more than enough capacity to serve all customers with lead times $w \leq d$, strictly prioritizing types with positive virtual delay costs and pooling all other types with $w = d$.

5.2. Impact of Capacity

Theorem 2 specifies how the key features of the optimal menu depend on the 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 $v > 0$, then the optimal segmentation and lead times are as follows. Define the 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 $\lambda^*(\mu) = \Lambda \Leftrightarrow \mu \geq \mu_A$. If $f_h(c_{\min}) \geq 0$ 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 threshold μ_P such that:*
   (a) if $f_h(c_{\min}) \geq 0$ 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 $\mu > \mu_{SD}$.
3. *Effect of valuations. There are unique thresholds $v_P < v_A$, such that pooling is optimal for all $\mu \in (d^{-1}, \mu_H)$ iff the base value $v \geq v_P$, and serving the entire market is optimal for all $\mu \geq d^{-1}$ iff $v \geq v_A$.*

Theorem 2 identifies three capacity intervals for the key features of the optimal menu. First, pooling is optimal in the intermediate range (μ_P, μ_H), regardless of other parameters; 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. Second, pricing out the middle of the delay cost spectrum is optimal in the range $\mu \in (d^{-1}, \mu_A)$, if and only if valuations are not too high ($v < v_A$); in this range capacity is insufficient for market coverage to be optimal, but still sufficient to profitably serve both low- and high-quality segments. If customers have sufficiently high valuations ($v \geq v_A$), then pricing out intermediate types is not optimal at any capacity level (part 3 of Theorem 2). In this case it is profitable already at low capacities to serve the entire market, i.e., $\mu_A \leq d^{-1}$, with only lead times $w \geq d$ for $\mu \leq d^{-1}$, and with all lead time qualities and some pooling for all capacities in (μ_P, μ_H). Finally, strategic delay (with pooling) is optimal for sufficiently large capacities ($\mu > \mu_{SD}$) if and only if patient types have negative virtual delay costs.

6. Summary, Connections and Extensions

By Theorem 1 the conditions $v > 0$ and $d > 0$ are necessary for three nonstandard properties to arise under the optimal menu: (i) Pricing out the middle of the delay cost spectrum; (ii) work-conserving pooling at the threshold lead time $w = d$; and (iii) strategic delay. It is important to emphasize that work-conserving pooling at the threshold lead time can be optimal under any delay cost distribution, specifically, even if it has increasing virtual delay costs. For increasing f_h and f_l, Theorem 2 specifies necessary and sufficient conditions for these properties of the optimal menu in terms of the base value parameter v, the delay cost distribution $f(c)$, and the capacity μ.

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

Lead time	Necessary conditions			Sufficient
$w < d$	$v > -dc_{\max}, d > 0$	$f_h' \not> 0$	$\mu > \max\left(\frac{1}{d}, \frac{1}{d+v/c_{\max}}\right)$	$\mu > \mu_H, v > 0$, $\exists c$ 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, $d \leq 0$, $f_l' \not> 0$

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

If f_h and/or f_l are nonmonotone, then 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, but the thresholds μ_P and μ_{SD} may change.

If f_h and/or f_l are nonmonotone, pooling *may* also be optimal at lead times $w \neq d$. Table 1 shows simple and intuitive sufficient conditions. 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 ($f_h(c) > 0 > f_h'(c)$). Pooling at $w > d$ is optimal if only lead times longer than d can be offered ($d \leq 0$), there is enough capacity to serve everyone ($\mu > \Lambda$), all types are profitable (v large), and some types c should be pooled ($f_l'(c) < 0$).

The rationale and necessary optimality conditions for pooling at the threshold lead time $w = d$ differ significantly from those for pooling at shorter or longer lead times: Pooling at $w = d$ arises only if it is profitable to offer lead times shorter than and longer than d, which requires $v > 0$ and $d > 0$. In this case, customers from two disjointed segments are served (Proposition 1), and a different virtual delay cost function applies to each segment: For each type c in the low-quality (high-quality) segment, $f_l(c)$ ($f_h(c)$) has a positive (negative) external revenue effect; that is, reducing the type-c lead time increases (decreases) the price for more patient (impatient) customers. As a result, $f_l(c) > f_h(c)$ holds for any delay cost distribution, and as discussed in Sections 4-5, 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 the external revenue effects of all types within a segment have the same sign, they have the same virtual delay cost function ($f_l(c)$ or $f_h(c)$), and pooling can only be optimal if the relevant virtual delay cost function is nonmonotone.

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, or $v \leq 0$. 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)[4], and it is essentially equivalent to the model of Katta and Sethuraman (2005), henceforth KS. Afèche and Mendelson (2004) a priori restrict attention to work conserving strict priority policies, but our analysis shows that doing so 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 N-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 valuation-delay cost ratios increase in delay costs, that is, $v_{i+1}/c_{i+1} < v_i/c_i$ where $c_i > c_{i+1}$. 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 $v \leq 0$. In particular, since $V(c)/c$ 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 to 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 for their discrete type model that computes the optimal lead times; it 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 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. 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 and chooses d, v and f appropriately.

6.3. Convex or Concave Value-Delay Cost Relationship V(c)

The affine value-delay cost relationship $V(c) = v + c \cdot d$ yields the simplest model of two-dimensional types with multiple delay costs and lead-time dependent ranking. This model simplifies the analysis, but has the restriction that all types value the threshold lead time d equally. We discuss the

[4] 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.

robustness of our main results for strictly convex or strictly concave $V(c)$. In these cases, the types have different net values for every lead time, i.e., $N_c(c,w) = V'(c) - w$ varies in c for all w. The function $V(c)$ has two main effects on the solution. First, under an IC menu customer utility changes at the rate $U'(c) = V'(c) - w(c)$ where $w(c)$ is nonincreasing (Proposition 1). Therefore, to serve a set of types with increasing utility, it is necessary that $w(c) < V'(c)$ (in the affine case, this holds at some capacity only if $d > 0$); similarly $w(c) > V'(c)$ for types with decreasing utility. Second, IR implies that types with decreasing utility can only be profitable if their net value is nonnegative for $w(c) > V'(c)$, which requires $V(c) - cV'(c) > 0$; equivalently, the ratio $V(c)/c$ is decreasing (in the affine case $v > 0$, see Section 6.2), or the net value function is log supermodular (see Section 6.4).

Convex $V(c)$. This models plausible situations where customers' valuations increase disproportionately with their impatience. Like in the affine case, in this case it may be optimal (i) to price out the middle of the delay cost spectrum, (ii) to pool (with a work-conserving policy) at a threshold lead time under any delay cost distribution, and (iii) to insert strategic delay. The basic rationale for (i) and (ii) is the same as in the affine case: The utility under an IC menu is convex in the type because $V'(c) - w(c)$ is increasing, so the sets of types with decreasing (increasing) utility, if any, are intervals including c_{min} (c_{max}), and pricing out intermediate types and pooling may be optimal. To serve a set with increasing utility requires $V'(c_{max}) > 1/\mu$; to serve one with decreasing utility requires $V(c_{min}) - c_{min}V'(c_{min}) > 0$. If either condition fails, it is optimal to serve a contiguous set of types including the most and/or least patient one, so pooling can only be optimal if the relevant virtual delay cost function is nonmonotone (see Section 6.1); strategic delay is also suboptimal if one of these condition fails (see the cases $v \le 0$, $d \le 0$ in Section 6.2). If $V'(c_{max}) > 1/\mu$ and $V(c_{min}) - c_{min}V'(c_{min}) > 0$, then both pricing out the middle and pooling at some intermediate threshold lead time may be optimal (see the discussion and examples in Section 4.2): f_l and f_h apply to customers with decreasing (increasing) utility, for $f_l', f_h' > 0$ strict priority scheduling is optimal within each segment in isolation, but pooling some customers from each segment into a common class is optimal if $f_l(c_l) > f_h(c_h)$ for the boundary types under strict priorities. These main structural properties are the same as in the affine case, but there are some differences in the convex case. First, the threshold lead time, call it W, is endogenous, with or without pooling. (The boundary types buy W with zero utility, so W must satisfy $V'(c_l) \le W \le V'(c_h)$; the affine case yields $W = d$.) Second, only the boundary types are indifferent between buying/balking; all other pooled types have strictly positive utility, and types that are priced out have strictly negative service utility. Strategic delay can only be optimal for a single service class, as in the affine case, but the lead time of the strategically delayed class is endogenous in the convex case. Moreover, in the convex case it may be optimal to serve the most patient types with strategic delay while pricing out some intermediate types. Consider ample capacity ($\mu = \infty$) for illustration. Suppose that $f_h(c_{min}) < 0$ and $f_h(c_0) = 0$ for some type c_0; then it is not optimal to serve types $[c_{min}, c_0]$ with zero lead time. This is only optimal for more impatient ones. If $V(c_{min}) - c_{min}V'(c_{min}) > 0 > V(c_0) - c_0V'(c_0)$ (this is possible if and only if $V'' > 0$), then intermediate types adjacent to $[c_0, c_{max}]$ cannot be profitably served with strategic delay, i.e., the price for the lead time $w = V'(c_0)$ is negative. However, it may still be profitable to serve a neighbourhood of c_{min} with strategic delay: This requires $V(c_{min}) - c_{min}w > 0$ for $w = (V(c_0) - V(c_{min}))/(c_0 - c_{min})$, i.e., the type c_{min} has a positive net value for the lead time that deters the impatient types $[c_0, c_{max}]$ from buying this degraded service. Note that this condition is *stronger* than log supermodularity of the net value function at c_{min}, because $w > V'(c_{min})$ is required to satisfy IC across disjoint type intervals.

Concave $V(c)$. Unlike in the convex and affine cases, for concave $V(c)$ the customer utility under an IC menu may be concave, because $V'(c)-w(c)$ may be decreasing. This property yields significant differences in the results. (i) Instead of pricing out intermediate types, it may be optimal to serve them and exclude the low and high ends of the delay cost spectrum. Specifically, net values for lead times w with $V'(c_{min}) > w > V'(c_{max})$ decrease towards both ends of the type spectrum. For example, if $V'(c_{min}) > 1/\mu > V'(c_{max})$, then intermediate types are the most profitable customers. (ii) If $f_l', f_h' > 0$ then pooling is not optimal for concave $V(c)$, because in this case, the set of pooled types is contiguous and all but the boundary type(s) have strictly positive utility: As a result, the pooled types have the same increasing virtual delay cost function, so service differentiation is profitable, and it is feasible to speed up lower and slow down higher types without violating the IR constraints. Nevertheless, assigning strict priorities may not be optimal either. For example, suppose market coverage with strict priorities results in a single, intermediate, type c_m with zero utility, and utility is decreasing for more patient types and increasing for more impatient ones. In this case $f_l(c_m) > f_h(c_m)$, and our analysis shows that it is profitable to speed up lower and slow down higher types c in a neighborhood of c_m only until their lead times equal $w(c) = V'(c)$ (further equalizing lead times would result in positive utility for these types); the result is a set of intermediate zero-utility customers around type c_m with virtual delay costs increasing across all other types. (iii) Strategic delay can also be optimal for concave $V(c)$. As in the convex case, strategic delay may be optimal without market coverage. Moreover, unlike in the affine and convex cases, strategic delay will typically be used for multiple service classes. This follows because for concave $V(c)$, the net value for lead times $w(c) = V'(c)$, that is, $V(c) - cV'(c)$, is strictly increasing. Consider ample capacity ($\mu = \infty$) for illustration. If $V(c_{min}) - c_{min}V'(c_{min}) > 0$ and $f_h(c_{min}) < 0$, it is optimal to sell differentiated lead times with strategic delay to all types with negative virtual delay costs.

6.4. Supermodularity, Damaged Goods, Queueing Effects

Anderson and Dana (2009)—henceforth AD, consider a maximum quality constraint in the standard monopoly price discrimination model with one-dimensional types, that is, the customer surplus is increasing in type for every quality level. They 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 two 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 to the condition $v > 0$[5] (or $V(c) - cV'(c) > 0$ in general). This condition, together with $d > 0$, is necessary but not sufficient for our key results, pricing out the middle, pooling at the threshold lead time, and strategic delay (see Sections 5 and 6.2)[6]. 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 such that their quality must be lowered artificially rather than through queueing delays.

[5] $v > 0$ 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 $v > 0$

[6] Pricing out the middle and pooling have no counterpart in AD because the type ranking is quality-independent in their model.

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 $d > 0$, the 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 literature on damaged goods (cf. Deneckere and McAffee, 1996; McAffee, 2007; Section 4.3 in AD). These papers assume a nonincreasing quality cost, as in our model, but they ignore capacity constraints. At ample capacity ($\mu = \infty$), the solution in our model becomes similar to the damaged goods solution for a model with zero costs: A high quality set of customers $[c_0, c_{max}]$ pays a high price $V(c_0)$ at quality $w(c) = 0$ and a low quality segment $[c_{min}, c_0]$ pays v for the 'damaged lead time' d.

However, queueing effects play a significant role under constrained capacity. The transition to service management ruled by supermodularity and $f_h(c_{min})$ occurs at the capacity threshold μ_{SD}. Beyond this threshold, only prioritization within the high quality segment distinguishes the system from a damaged goods model. To further isolate the effect of a capacity constraint in the absence of queueing, eliminate the work conservation constraints (3) in Problem 1, let λ_0 be the rate of customers with nonnegative virtual delay cost f_h, and assume that $\lambda_0 < \Lambda$. For $v > 0$, $d > 0$, it is clear that strategic delay will be optimal if and only if $\mu > \lambda_0$, in which case $w(c) = d$ for $c < c_0$ and $w(c) = 0$ otherwise. In the presence of queueing, the threshold μ_{SD} is strictly larger than λ_0. Furthermore, it is the queueing constraints in our model that give rise to the three tiered menu structure, pricing out the middle and intermediate pooling as optimal strategies.

6.5. Relation to Price-Lead Time Menu Design for Queueing Systems

Katta and Sethuraman (2005) and Afèche (2004, 2013) are probably closest to this paper in that they also focus on the design of revenue-maximizing price-lead time menus and scheduling policies with customer choice. We revisit these papers to elaborate on their relationship to ours.

KS is relevant for our results on pooling. To our knowledge, KS is the first and only other paper in the queueing literature that considers the pooling phenomenon studied here. As noted in the introduction and detailed in Sections 6.1-6.2, KS assume a lead-time-independent ranking of types, in contrast to our model, which rules out the solution features analyzed in this paper.

AF is relevant for our results on strategic delay. To our knowledge, AF is the first in the queueing literature to identify strategic delay and characterize necessary and sufficient conditions for its optimality. Unlike our model where valuations are perfectly correlated to delay costs, AF allows heterogenous valuations for each delay cost level. In his model with two possible delay costs, pooling is not optimal[7] and the notion of pricing out the middle is meaningless. Our results on optimal strategic delay complement his. In our multiple delay cost case, optimal strategic delay targets a range of types at the low end of the delay cost spectrum, and it may occur jointly with pooling. For affine and convex $V(c)$ functions, strategic delay is optimal only at a single delay cost level, like in the model of AF, whereas optimal strategic delay involves multiple service classes under concave $V(c)$. Furthermore, in contrast to our results, with valuation heterogeneity optimal strategic delay is not necessarily a "large capacity phenomenon"; AF identifies cases where strategic delay is optimal only if capacity is relatively scarce.

[7] Pooling types with the same delay cost but different valuations is known to be optimal without loss of generality. Hence it is optimal in general to offer at most one, in the binary case exactly one, class per delay cost.

From a methodological perspective, our 3-Step solution approach is based on classical mechanism design techniques (Step 1) and the approach outlined in AF (Steps 2 and 3) and followed by KS. The crux of our analysis which differentiates it from that in AF and KS is the solution of the optimal menu design and segmentation problem for fixed arrival rate (Step 2). In the two-type model of AF, this problem is straightforward; optimizing over arrival rates (Step 3) is more involved as a result of valuation heterogeneity. Because KS impose a lead-time-independent type ranking, the essence of Step 2 in their model is the segmentation algorithm that determines which types to pool within a single high-quality segment. In contrast, our model with a lead-time-dependent type ranking calls for two joint decisions at every arrival rate, which types to admit with lead times above or below the threshold d, and which types to pool with $w = d$. Unlike pooling within a single high-quality segment as in KS, pooling at this specific threshold lead time introduces an additional lead time constraint that must be explicitly considered together with those on work conservation. As a result, unique from other types of pooling, the set of pooled customers depends not only on the arrival rate and customer distribution but also on the operational capabilities. This additional constraint also implies a more complex interdependency with capacity as the segmentation structure changes, which makes Step 3 more involved. These challenges are not an artifact of the affine model: The problem is more difficult in other cases, e.g., in the convex case discussed in Section 6.3, because the threshold lead time is constrained by endogenous upper and lower bounds. Nevertheless, as shown in this paper, the affine model is analytically tractable, and the solution generates a number insights on how the results generalize.

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, Northwestern University.

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

[6] 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.

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

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

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

[10] Deneckere, R.J., R.P. McAfee. Damaged goods. 1996. *Journal of Economics & Management Strategy* 5 (2): 149-174.

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

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

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

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

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

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

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

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

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

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

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

[22] 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.

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

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

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

[26] Yahalom, T., J.M. Harrison, S. Kumar. 2006. Designing and pricing incentive compatible grades of service in queueing systems. Working Paper, Stanford University.

[27] Zhao, X., K.E. Stecke, A. Prasad. 2011. Lead time and price quotation mode selection: Uniform or differentiated?. 2012 *POMS*. 21(1) 177-193.

Appendix: Proofs

Proof of Proposition 1. See Problem 1 in Section 2.2. *The constraints (4)-(6) $\Rightarrow$ 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 for any pair of types:

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

Similarly, we must have $U(c') \geq U(c;c')$, so 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 $c < c'$ then (26) implies $w(c') \leq w(c)$, and (25) implies $p\,(c') - p\,(c) \geq c\,(w\,(c) - w\,(c')) \geq 0$. The fact $c_l \leq c_h$ follows since w is nonincreasing in c. Since w is nonincreasing it is Riemann integrable, and (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 (i). 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 (ii) 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. Let $U_m \triangleq U(c) = v - p(c)$. It follows from (29) that $U(c') = U_m$ 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. $\Rightarrow$ the constraints (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. Refer to Problem 1 and the revenue rate (13) in Section 4.1. Fix $\lambda < \mu$. Let $D(\lambda_l, \lambda_h, w)$ be the virtual delay cost rate as a function of λ_l, λ_h and the lead time function w:

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

For fixed λ, maximizing revenue is equivalent to minimizing $D(\lambda_l, \lambda_h, w)$ over λ_l, λ_h and $w : \mathcal{C} \to \mathbb{R}$, subject to $\lambda \geq \lambda_l + \lambda_h$, increasing $w(c)$, $w(c) > d > w(c')$ for $c < c_l(\lambda_l)$ and $c > c_h(\lambda_h)$ and operational constraints:

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

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

Recall that f_l, f_h satisfy (14)-(15), $f_l' > 0$, $f_h' > 0$, 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.

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

(ii) 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 $w^*(c_{\max}) = 1/\mu$. 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 *(i)*, and because $f_h' > 0$.

(iii) If the set C_l^* is nonempty then the constraints (32) are binding for $c \in [c_{\min}, c_l^*]$, and $w^*(c_{\min}) = \mu/(\mu - \lambda)^2$. This follows from a similar argument as for *(ii)*, because $f_l > 0$ and $f_l' > 0$.

Part 1(a). Properties *(ii)*-*(iii)* imply (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 *(ii)* 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 $(ii) - (iii)$ for an optimal solution the constraints (32) therefore simplify to

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

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

Part 1(b). This claim follow directly from (ii)–(iii). Since (31) is binding for $c \in [c_h^*, c_{\max}]$ these types are strictly prioritized in order of their delay costs and receive strict priorities over all lower types. Since (32) is binding for $c \in [c_{\min}, c_l^*]$ these types are strictly prioritized in order of their delay costs and receive strictly lower priority than all higher types. 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 Property (i) above since f_h is continuous.

Parts 2(b)-(c). Follow by substituting $w^*(c)$ from (16) in the revenue function (13) and analyzing its partial derivatives with respect to λ_l and λ_h. The details are in the proof of Proposition 2.■

The following proposition describes structural transitions as λ increases and is an important intermediate step to determine the characteristics of the revenue function with respect to parameters.

PROPOSITION 2. *Fix a capacity $\mu > 0$ and assume that $f_l' > 0$ and $f_h' > 0$. Let h be shorthand for $\lambda_h^*(\lambda) > 0$, m for $\lambda_m^*(\lambda) > 0$, and l for $\lambda_l^*(\lambda) > 0$, where m involves pooling. 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: $\lambda_l^*(\lambda) > 0 = \lambda_m^*(\lambda) = \lambda_h^*(\lambda)$.*
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 segmentation 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)

   Segmentations (m,l) and (h, m_{sd}) are never optimal. The optimal policy is work conserving.

   (b) *Selling only medium and low quality classes, (m,l), is optimal for some (λ,Λ) if and only if $F\left(f_l^{-1}(c_{\max})\right) \cdot d < \mu^{-1} < d$. 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) if and only if $f_h(c_{\min}) < 0$ and $d > \mu^{-1}$. If $f_h(c_{\min}) < 0$ and $\mu^{-1} \leq F\left(f_l^{-1}(c_{\max})\right) \cdot d$ then (34) changes in that two thresholds $\underline{\Lambda}_{sd} < \overline{\Lambda}_{sd}$ replace Λ_1 and yield the following segmentation 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 of Proposition 2. The optimal customer segmentation and lead time menu are obtained by solving Problem 2 for fixed λ. Refer to (17)-(21) in Section 4.2.

Part 1. If $\mu^{-1} \geq d$ then (20) only holds if $\lambda_l = \lambda$; hence $\lambda_l^*(\lambda) = \lambda$ and $\lambda_m^*(\lambda) = \lambda_h^*(\lambda) = 0$.

Part 2. Suppose that $\mu^{-1} < d$. We first reformulate Problem 2 for analytical convenience.

Definitions and problem formulation. Let $\lambda_{mh} \triangleq \lambda_m + \lambda_h$ be the aggregate rate for the m class (lead time d) and h classes (lead times $< d$). For fixed λ write the revenue (17) as

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

where $\lambda_l = \lambda - \lambda_{mh}$, $c_l(x) \triangleq F^{-1}(x/\Lambda)$ with $c_l' > 0$, $c_h(x) \triangleq \overline{F}^{-1}(x/\Lambda)$ with $c_h' < 0$; and $f_l, f_l', f_h' > 0$.

By Lemma 1 types in h or l classes are prioritized by their delay costs, where h classes have lead times $< d$ and get priority over the m class (lead time $= d$) which gets priority over l classes (lead times $> d$). Let $\lambda_P \triangleq \mu - \sqrt{\mu/d}$: it is the maximum feasible rate for h classes; at this rate the maximum lead time is d under work conserving priority service: $\mu/(\mu - \lambda_P) = d$. Let $\lambda_F \triangleq \mu - 1/d$: it is the maximum feasible aggregate rate for m and h classes; at this rate the lead time is d under work conserving FIFO service: $(\mu - \lambda_F)^{-1} = d$, where $\mu d > 1$ implies $0 < \lambda_P < \lambda_F$.

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

$$\overline{\lambda}_h(\lambda_{mh}) \triangleq \min\left(\lambda_{mh}, \mu - \frac{\mu/d}{(\mu - \lambda_{mh})}\right) = \begin{cases} \lambda_{mh}, & \text{if } \lambda_{mh} \in [0, \lambda_P] \\ \mu - \frac{\mu/d}{\mu - \lambda_{mh}} < \lambda_{mh}, & \text{if } \lambda_{mh} \in [\lambda_P, \lambda_F] \end{cases}. \tag{38}$$

For $\lambda_{mh} \leq \lambda_P$ all classes can have lead times $< d$, so $\overline{\lambda}_h(\lambda_{mh}) = \lambda_{mh}$ increases on $[0, \lambda_P]$ with $\overline{\lambda}_h(\lambda_P) = \lambda_P$. For $\lambda_{mh} > \lambda_P$ only a portion $\overline{\lambda}_h(\lambda_{mh}) < \lambda_{mh}$ can be allocated to h classes, and $\overline{\lambda}_h(\lambda_{mh})$ decreases on $[\lambda_P, \lambda_F]$, with $\overline{\lambda}_h(\lambda_F) = 0$. Having $\lambda_{mh} > \lambda_F$ violates (20).

Constraint (20) in Problem 2 holds if and only if $\lambda_{mh} \leq \lambda_F$ and $\lambda_h \leq \overline{\lambda}_h(\lambda_{mh})$; 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

$$\max_{\lambda_{mh}, \lambda_h} = \Pi(\lambda_{mh}, \lambda_h) \tag{39}$$

$$\text{s.t.} \quad \min(\lambda, \lambda_P) \leq \lambda_{mh} \leq \min(\lambda, \lambda_F), \tag{40}$$

$$0 \leq \lambda_h \leq \overline{\lambda}_h(\lambda_{mh}). \tag{41}$$

The lower bound of (40) ensures a work conserving policy if $\lambda_l > 0$: $\lambda_{mh} < \min(\lambda, \lambda_P)$ is not feasible[8]. It implies strategic delay for l classes ($d > \mu/(\mu - \lambda_{mh})^2$) which is suboptimal by Lemma 1.1(c).

Overview of proof. We prove Parts 2(a)-(c) by solving (39)-(41) in the following three steps.

Step 1. For fixed λ_{mh}, we characterize the optimal rate $\lambda_h(\lambda_{mh})$.

Step 2. For fixed λ, we characterize the optimal rates $\lambda_{mh}^*(\lambda)$ and $\lambda_h^*(\lambda) = \lambda_h(\lambda_{mh}^*(\lambda))$, which determine the optimal segmentation. The conditions for $\lambda_{mh}^*(\lambda)$ and $\lambda_h^*(\lambda)$ are stated in terms of the virtual delay costs of the corresponding marginal types as summarized in (45) and (49). This segmentation problem adds an additional dimension to the arrival rate problem found in typical screening papers (c.f. Katta and Sethuraman 2005) where admitted customers form a single contiguous segment (i.e. $\lambda_h = \lambda$).

[8] We say 'feasible' with respect to (40)-(41), which are implied by the *optimal* lead times in Lemma 1.1. Choosing λ_{mh} and/or λ_h that violate these constraints may be feasible for a *suboptimal* lead time function.

Step 3. We translate the optimality conditions of Step 2 for $\lambda^*_{mh}(\lambda)$ and $\lambda^*_h(\lambda)$ into conditions on λ and Λ. These conditions imply the optimal segmentation structure specifed in Parts 2(a)-(c) of the Proposition. We organize this analysis into Lemmas 4-8 as further detailed below.

Step 1. Optimal λ_h for fixed λ_{mh}. Fix $\lambda_{mh} \in (0, \min(\lambda, \lambda_F)]$ and let $\lambda_h(\lambda_{mh}) \triangleq \arg\left\{\max_{\lambda_h} \Pi(\lambda_{mh}, \lambda_h) \text{ s.t. } 0 \le \lambda_h \le \overline{\lambda}_h(\lambda_{mh})\right\}$ be the corresponding optimal λ_h. From (37) we have

$$\frac{\partial \Pi(\lambda_{mh}, \lambda_h)}{\partial \lambda_h} = f_h(c_h(\lambda_h))\left(d - \frac{\mu}{(\mu - \lambda_h)^2}\right), \text{ for } \lambda_h \le \overline{\lambda}_h(\lambda_{mh}), \tag{42}$$

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

$$\lambda_h(\lambda_{mh}) = \begin{cases} \overline{\lambda}_h(\lambda_{mh}), & \text{if } f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) \ge 0 \\ \lambda_0 \triangleq \Lambda \overline{F}\left(f_h^{-1}(0)\right) < \overline{\lambda}_h(\lambda_{mh}), & \text{if } f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) < 0 \end{cases}. \tag{43}$$

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 (43) 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 $\lambda_0 < \overline{\lambda}_h(\lambda_{mh})$:

$$\frac{\mu}{(\mu - \lambda_{mh})(\mu - \lambda_0)} < \frac{\mu}{(\mu - \lambda_{mh})\left(\mu - \overline{\lambda}_h(\lambda_{mh})\right)} \le d. \tag{44}$$

Step 2. Optimal segmentation for fixed λ: optimal $\lambda^*_{mh}(\lambda)$ and $\lambda^*_h(\lambda)$. Let $\lambda^*_{mh}(\lambda) \triangleq \arg\left\{\max_{\lambda_{mh}} \Pi(\lambda_{mh}, \lambda_h(\lambda_{mh})) \text{ s.t.} \min(\lambda, \lambda_P) \le \lambda_{mh} \le \min(\lambda, \lambda_F)\right\}$ be the optimal λ_{mh} for fixed λ. Write $\lambda^*_h(\lambda), \lambda^*_m(\lambda)$ and $\lambda^*_l(\lambda)$ for the optimal rates of h, m and l classes for fixed λ, where $\lambda^*_h(\lambda) = \lambda_h(\lambda^*_{mh}(\lambda))$ by (43), whereas $\lambda^*_m(\lambda) = \lambda^*_{mh}(\lambda) - \lambda^*_h(\lambda)$ and $\lambda^*_l(\lambda) = \lambda - \lambda^*_{mh}(\lambda)$.
Optimal segmentation for $\lambda \le \lambda_P$. For $\lambda \le \lambda_P$ it is not optimal to sell l classes: (40) requires $\lambda_{mh} = \lambda$, so the maximizer satisfies $\lambda^*_{mh}(\lambda) = \lambda$. By (38) the entire λ can be sold to h classes, so $\overline{\lambda}_h(\lambda^*_{mh}(\lambda)) = \overline{\lambda}_h(\lambda) = \lambda$. Therefore (43) yields the following optimal segmentations, where m_{sd} denotes that the medium lead time involves strategic delay.

	Optimal segmentations for $\mu > d^{-1}$ *and* $\lambda \le \lambda_P$				
segments	*virtual delay cost condition*	$\lambda^*_{mh}(\lambda)$	$\lambda^*_h(\lambda)$	$\lambda^*_m(\lambda)$	$\lambda^*_l(\lambda)$
(h)	$f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right) = f_h(c_h(\lambda)) \ge 0$	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

(45)

Optimal segmentation for $\lambda > \lambda_P$. (40) requires $\lambda_{mh} \in [\lambda_P, \min(\lambda, \lambda_F)]$. From (37):

$$\frac{d\Pi(\lambda_{mh}, \lambda_h(\lambda_{mh}))}{d\lambda_{mh}} = \frac{\partial \Pi(\lambda_{mh}, \lambda_h(\lambda_{mh}))}{\partial \lambda_{mh}} + \frac{\partial \Pi(\lambda_{mh}, \lambda_h(\lambda_{mh}))}{\partial \lambda_h} \cdot \lambda_h'(\lambda_{mh}) =$$
$$= f_l(c_l(\lambda - \lambda_{mh}))\left(\frac{\mu}{(\mu - \lambda_{mh})^2} - d\right) + f_h(c_h(\lambda_h(\lambda_{mh})))\left(d - \frac{\mu}{(\mu - \lambda_h(\lambda_{mh}))^2}\right)\lambda_h'(\lambda_{mh}), \tag{46}$$

where $\Lambda f(c_l(x)) c_l'(x) = 1$ and $\lambda_h(\lambda_{mh})$ satisfies (43). If $f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) < 0$ then $\lambda_h'(\lambda_{mh}) = 0$. If $f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) \ge 0$, then $\lambda_h(\lambda_{mh}) = \overline{\lambda}_h(\lambda_{mh}) = \mu - \mu/d(\mu - \lambda_{mh})$ by (38) and (43); in this case $\lambda_h'(\lambda_{mh}) = -\mu/d(\mu - \lambda_{mh})^2$ and $\mu/(\mu - \lambda_h(\lambda_{mh}))^2 = d^2(\mu - \lambda_{mh})^2/\mu$. Substituting into (46) yields

$$\frac{d\Pi(\lambda_{mh}, \lambda_h(\lambda_{mh}))}{d\lambda_{mh}} = \begin{cases} f_l(c_l(\lambda - \lambda_{mh}))\left(\frac{\mu}{(\mu - \lambda_{mh})^2} - d\right), & \text{if } f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) < 0 \\ g(\lambda, \lambda_{mh})\left(\frac{\mu}{(\mu - \lambda_{mh})^2} - d\right), & \text{if } f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right) \ge 0 \end{cases}, \tag{47}$$

for $\lambda_{mh} \in [\lambda_P, \min(\lambda, \lambda_F)]$, where

$$g(\lambda, \lambda_{mh}) \triangleq f_l(c_l(\lambda - \lambda_{mh})) - f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)\right). \tag{48}$$

The function $g(\lambda, \lambda_{mh})$ is important in determining the solution. 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 the revenue derivative in (47) 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 (47) is zero at $\lambda_{mh} = \lambda_P$ and positive for $\lambda_{mh} > \lambda_P$. The maximizer $\lambda^*_{mh}(\lambda)$ is unique since the following holds 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 (38) and $f'_h c'_h < 0$.

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

By (40)-(41) the threshold λ_P is *the smallest feasible* λ_{mh} *and the largest feasible* λ_h. The choice of λ_{mh} determines the maximum feasible h rate $\overline{\lambda}_h(\lambda_{mh})$, and the l rate $\lambda_l = \lambda - \lambda_{mh}$. The rates of both segments decrease in λ_{mh}, and their marginal types and virtual delay costs 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}. Since $\overline{\lambda}_h(\lambda_{mh}) = 0$ if $\lambda_{mh} = \lambda_F$ and $\lambda_l = 0$ if $\lambda = \lambda_{mh}$, $\min(\lambda_F, \lambda)$ is the *largest feasible* λ_{mh}.

Properties (i)–(ii) and (47) determine the maximizer $\lambda^*_{mh}(\lambda)$, and (43) determines $\lambda^*_h(\lambda) = \lambda_h(\lambda^*_{mh}(\lambda))$. The optimal segmentations and the corresponding conditions are summarized in (49). The optimal segmentation sets λ_{mh} and λ_h to minimize the virtual delay cost of types served, subject to the constraints (40)-(41). Since f_l and f_h are increasing, a segmentation is optimal if and only if it satisfies the following criteria. It *minimizes* $\lambda_{mh} \in [\lambda_P, \min(\lambda_F, \lambda)]$, and *maximizes* $\lambda_h \le \overline{\lambda}_h(\lambda_{mh})$, subject to two conditions. (1) The marginal h type's virtual delay cost is nonnegative: $f_h(c_h(\lambda_h)) \ge 0$. (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: $f_h(c_h(\lambda_h)) \ge f_l(c_h(\lambda - \lambda_{mh}))$. We review these criteria for the cases of (49) in the order $(h,l) - (h,m,l) - (m,l) - (h,m) - (h,m_{sd})$.

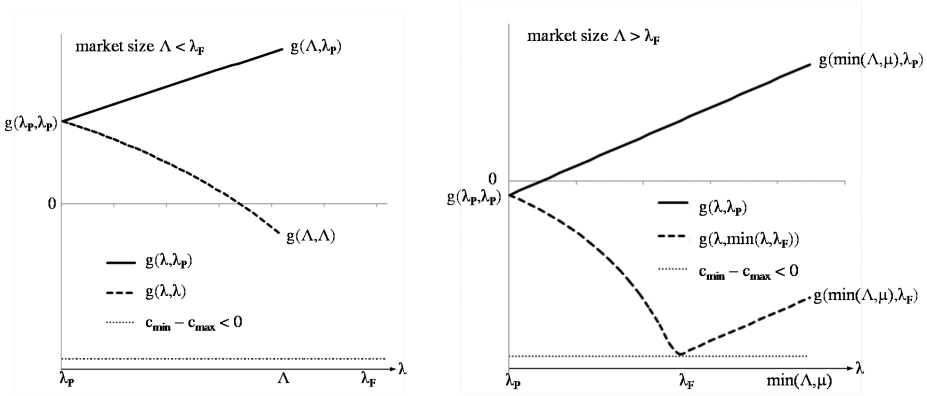
Optimal segmentations for $\mu > d^{-1}$ *and* $\lambda > \lambda_P$						
	conditions (other than $\lambda > \lambda_P$*)*		*demand rates for each segment*			
segments	λ	*virtual delay costs*	$\lambda^*_{mh}(\lambda)$	$\lambda^*_h(\lambda)$	$\lambda^*_m(\lambda)$	$\lambda^*_l(\lambda)$
(h, m_{sd})	$\lambda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

(49)

Segmentation (h,l): if $g(\lambda, \lambda_P) \le 0$, 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: allocate λ to m and h classes, then maximize $\lambda_h \le \overline{\lambda}_h(\lambda)$ subject to (1). If all of $\overline{\lambda}_h(\lambda)$ can be sold to types with nonnegative f_h this yields (h,m). If not, then (h, m_{sd}), *with strategic delay* is optimal. In this case 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

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



$\lambda-\lambda_0$ consisting of types with *negative* f_h. Since $\lambda_0<\overline{\lambda}_h(\lambda)$ the lead time d involves strategic delay: it exceeds the minimum achievable lead time given λ_0 have higher priority – see also (43).

Step 3. Optimal segmentations depending on λ and Λ. The optimality conditions of Step 2, (45) for $\lambda\leq\lambda_P$ and (49) 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. We proceed as follows. Lemmas 4-6 characterize the signs of $g(\lambda,\lambda_P)$ and $g(\lambda,\min(\lambda,\lambda_F))$ depending on λ and Λ. Together with (45) and (49) they imply Parts 2(a)-(b), in which $f_h(c_{\min})\geq 0$. Lemmas 7-8 characterize the sign of $f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right)$; these additional results imply Part 2(c) in which $f_h(c_{\min})<0$.

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*

$$g(\lambda,\lambda_P)=f_l(c_l(\lambda-\lambda_P))-f_h(c_h(\lambda_P)). \tag{50}$$

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*

$$g(\lambda,\min(\lambda_F,\lambda))=g(\lambda,\lambda)=c_{\min}-f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right),\quad \lambda\in[\lambda_P,\min(\lambda_F,\Lambda)]. \tag{51}$$

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

$$g(\lambda,\min(\lambda_F,\lambda))=g(\lambda,\lambda_F)=f_l(c_l(\lambda-\lambda_F))-c_{\max},\quad \lambda\in[\lambda_F,\min(\Lambda,\mu)]. \tag{52}$$

Proof. By (48) 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 (50). Set $\lambda_{mh}=\lambda$ and note that $c_l(0)=c_{\min}$ to get (51). Set $\lambda_{mh}=\lambda_F$ and note that $\overline{\lambda}_h(\lambda_F)=0$ and $c_h(0)=c_{\max}$ to get (52). 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 (38), while $\lambda_l=0$.■

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 $\lambda=\lambda_F$, when all types are served FIFO (Part 2). In both panels $g(\lambda,\lambda_P)>g(\lambda,\min(\lambda,\lambda_F))$ for fixed $\lambda>\lambda_P$: as explained in Step 2, $g(\lambda,\lambda_{mh})$ decreases in λ_{mh} for fixed λ.

Increasing Λ has two effects on $g(\lambda,\lambda_{mh})$ for $\lambda_{mh}=\lambda_P$ and $\lambda_{mh}=\min(\lambda,\lambda_F)$. It increases the set of feasible λ. It also decreases $g(\lambda,\lambda_{mh})$, see Figure 4: for fixed (λ,λ_{mh}) a larger market increases the

virtual delay cost difference as it "pulls apart" the marginal types $c_l(\lambda-\lambda_{mh})=F^{-1}((\lambda-\lambda_{mh})/\Lambda)$ and $c_h\left(\overline{\lambda}_h(\lambda_{mh})\right)=\overline{F}^{-1}\left(\overline{\lambda}_h(\lambda_{mh})/\Lambda\right)$. To make this dependence explicit we henceforth write

$$g(\lambda,\lambda_{mh},\Lambda)=f_l(c_l(\lambda-\lambda_{mh},\Lambda))-f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{mh}),\Lambda\right)\right), \text{ for } \lambda\in[\lambda_P,\min(\mu,\Lambda)],\ \lambda_{mh}\in[\lambda_P,\min(\lambda,\lambda_F)].$$

Lemma 5 and 6 characterize, respectively, the sign of $g(\lambda,\min(\lambda_F,\lambda),\Lambda)$ and $g(\lambda,\lambda_P,\Lambda)$.

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

$$g(\lambda,\min(\lambda_F,\lambda),\Lambda)=\begin{cases} g(\lambda,\lambda,\Lambda)=c_{\min}-f_h\left(\overline{F}^{-1}\left(\frac{\overline{\lambda}_h(\lambda)}{\Lambda}\right)\right),\ \lambda\in[\lambda_P,\min(\Lambda,\lambda_F)],\\ g(\lambda,\lambda_F,\Lambda)=f_l\left(F^{-1}\left(\frac{\lambda-\lambda_F}{\Lambda}\right)\right)-c_{\max},\ \lambda\in[\lambda_F,\min(\Lambda,\mu)].\end{cases} \tag{53}$$

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

$$\Lambda_2\triangleq\frac{\mu}{2}\left(\frac{1}{z}+1\right)-\sqrt{\frac{\mu^2}{4}\left(\frac{1}{z}-1\right)^2+\frac{\mu}{zd}} \text{ and } \Lambda_3\triangleq\frac{\lambda_P}{z},\ \text{with } z=\overline{F}\left(f_h^{-1}(c_{\min})\right)\in(0,1), \tag{54}$$

   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 $\Lambda\le\Lambda_2$ then $g(\lambda,\lambda,\Lambda)\ge 0$ for $\lambda\in[\lambda_P,\Lambda]$, where $\Lambda_2<\lambda_F$.*

   (b) *If $\Lambda\in(\Lambda_2,\Lambda_3)$, then there is a demand rate threshold $\lambda_1\in(\lambda_P,\lambda_F)$ such that*

$$g(\lambda,\lambda,\Lambda)\begin{cases}>0,\ \text{if } \lambda\in[\lambda_P,\lambda_1),\\ =0,\ \text{if } \lambda=\lambda_1\triangleq\mu-\frac{\mu/d}{\mu-\Lambda\overline{F}\left(f_h^{-1}(c_{\min})\right)},\\ <0,\ \text{if } \lambda\in(\lambda_1,\min(\lambda_F,\Lambda)].\end{cases} \tag{55}$$

   (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 $g(\lambda,\min(\lambda_F,\lambda),\Lambda)=g(\lambda,\lambda_F,\Lambda)$ depends on the thresholds*

$$\underline{\Lambda}_{ml}\triangleq\frac{\lambda_F}{\overline{F}\left(f_l^{-1}(c_{\max})\right)}>\lambda_F \text{ and } \overline{\Lambda}_{ml}\triangleq\frac{1}{dF\left(f_l^{-1}(c_{\max})\right)}, \tag{56}$$

   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*

$$g(\lambda,\lambda_F,\Lambda)\begin{cases}<0,\ \text{if } \lambda\in(\lambda_F,\lambda_3),\\ =0,\ \text{if } \lambda=\lambda_3\triangleq\lambda_F+\Lambda F\left(f_l^{-1}(c_{\max})\right),\\ >0,\ \text{if } \lambda>\lambda_3.\end{cases} \tag{57}$$

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

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

Proof. By Lemma 4.2, for fixed Λ the function $g(\lambda,\min(\lambda_F,\lambda),\Lambda)$ 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 than, the other larger than λ_F. The number of roots depends on the sign of $g(\lambda,\min(\lambda_F,\lambda),\Lambda)$ for $\lambda=\lambda_P$, for $\lambda=\min(\Lambda,\lambda_F)$ and for $\lambda=\min(\Lambda,\mu)$ when $\Lambda>\lambda_F$.

Part 1. The threshold Λ_3 determines the sign of $g(\lambda,\min(\lambda_F,\lambda),\Lambda)$ 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 $\Lambda\ge\lambda_P$. 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 (54).

The threshold Λ_2 determines the sign of $g(\lambda,\min(\lambda_F,\lambda),\Lambda)$ for $\lambda=\min(\Lambda,\lambda_F)$. It is the unique solution of $g(\Lambda,\Lambda,\Lambda)=c_{\min}-f_h(\overline{F}^{-1}\left(\overline{\lambda}_h(\Lambda)/\Lambda\right))=0$ 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 (54).

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 that 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)*. 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 (55).

Part 2. The thresholds $\underline{\Lambda}_{ml}$ and $\overline{\Lambda}_{ml}$ determine the sign of $g(\lambda,\min(\lambda_F,\lambda),\Lambda)$ 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 (53) the virtual delay cost difference is

$$g(\min(\Lambda,\mu),\lambda_F,\Lambda)=\begin{cases} g(\Lambda,\lambda_F,\Lambda)=f_l\left(F^{-1}\left(\frac{\Lambda-\lambda_F}{\Lambda}\right)\right)-c_{\max}, & \Lambda\in[\lambda_F,\mu],\\ g(\mu,\lambda_F,\Lambda)=f_l\left(F^{-1}\left(\frac{1}{d\Lambda}\right)\right)-c_{\max}, & \Lambda\geq\mu. \end{cases} \tag{58}$$

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

$$g(\mu,\lambda_F,\mu)=f_l\left(\overline{F}\left(\frac{1}{d\mu}\right)\right)-c_{\max}>0\Leftrightarrow F(f_l^{-1}(c_{\max}))\cdot d<\mu^{-1}. \tag{59}$$

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 (56). We further have $g(\Lambda,\lambda_F,\Lambda)>0$ for $\Lambda\in(\underline{\Lambda}_{ml},\mu)$ and $g(\mu,\lambda_F,\Lambda)>0$ for $\Lambda\in[\mu,\overline{\Lambda}_{ml})$. Recalling from Lemma 4.2(b) that for fixed $\Lambda>\lambda_F$, the function $g(\lambda,\lambda_F,\Lambda)$ satisfies $g(\lambda_F,\lambda_F,\Lambda)<0$ and increases in $\lambda\geq\lambda_F$ establishes the unique root λ_3 as claimed. Solving $g(\lambda_3,\lambda_F,\Lambda)=0$ yields (57).

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 all $\lambda\geq\lambda_F$.

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 (59), together with Lemma 4.2(b) again imply that $g(\lambda,\lambda_F,\Lambda)<0$ for all feasible $\lambda\geq\lambda_F$.■

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

$$\Lambda_4\triangleq\left\{\Lambda\geq\lambda_P:\Lambda=\frac{\sqrt{\mu/d}}{F\left(f_l^{-1}\left(f_h\left(c_h\left(\lambda_P,\Lambda\right)\right)\right)\right)}\right\}, \tag{60}$$

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 a demand rate threshold $\lambda_2>\lambda_P$ such that*

$$g(\lambda,\lambda_p,\Lambda)\begin{cases} <0, \text{ if } \lambda\in[\lambda_P,\lambda_2),\\ =0, \text{ if } \lambda=\lambda_2\triangleq\lambda_P+\Lambda F\left(f_l^{-1}\left(f_h\left(c_h\left(\lambda_P,\Lambda\right)\right)\right)\right),\\ >0, \text{ if } \lambda>\lambda_2. \end{cases} \tag{61}$$

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

Proof. 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 $\lambda \in [\lambda_P, \min(\Lambda, \mu)]$.

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 (60) determines the sign of $g(\lambda, \lambda_P, \Lambda)$ for $\lambda = \min(\Lambda, \mu)$ where

$$g(\min(\Lambda, \mu), \lambda_P, \Lambda) = \begin{cases} g(\Lambda, \lambda_P, \Lambda) = f_l\left(F^{-1}\left(\frac{\Lambda - \lambda_P}{\Lambda}\right)\right) - f_h\left(\overline{F}^{-1}\left(\frac{\lambda_P}{\Lambda}\right)\right), \ \Lambda \in [\lambda_P, \mu], \\ g(\mu, \lambda_P, \Lambda) = f_l\left(F^{-1}\left(\frac{\mu - \lambda_P}{\Lambda}\right)\right) - f_h\left(\overline{F}^{-1}\left(\frac{\lambda_P}{\Lambda}\right)\right), \ \Lambda \geq \mu. \end{cases} \tag{62}$$

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 (60).

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 only consider the nontrivial case $\overline{\Lambda}_{ml} > \mu$ and show that $g\left(\mu, \lambda_P, \overline{\Lambda}_{ml}\right) > 0$. Recall that $g\left(\mu, \lambda_F, \overline{\Lambda}_{ml}\right) = 0$ 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 $g\left(\mu, \lambda_P, \overline{\Lambda}_{ml}\right) > g\left(\mu, \lambda_F, \overline{\Lambda}_{ml}\right) = 0$.

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 Proposition 2.2(a). Suppose that $f_h(c_{\min}) \geq 0$ and $\frac{1}{\mu} \leq F(f_l^{-1}(c_{\max})) \cdot d$.

The optimality conditions (45) for $\lambda \leq \lambda_P$ and (49) for $\lambda > \lambda_P$, combined with Lemmas 5-6, imply the results. Since $f_h(c_{\min}) \geq 0$, 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 (45) and (49), which cannot hold. Since $\frac{1}{\mu} \leq F(f_l^{-1}(c_{\max})) \cdot d$, segmentation (m, l) cannot be optimal: the condition is $g(\lambda, \lambda_F, \Lambda) \geq 0$, see (49), and Lemma 5.2(b) rules it out.

The conditions (45) and (49), 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 $\lambda \leq \lambda_P$ segmentation (h) is optimal by (45). 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 $\frac{1}{\mu} \leq F(f_l^{-1}(c_{\max})) \cdot d$. Together they specify that $g(\lambda, \min(\lambda, \lambda_F), \Lambda) \geq 0$ for $\lambda \leq \lambda_1$ 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 (49) that (h, m) is optimal for $\lambda \leq \lambda_1$, and (h, m, l) is optimal for $\lambda > \lambda_1$.

Proof of Proposition 2.2(b). Suppose that $f_h(c_{\min}) \geq 0$ 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 $\lambda \geq \lambda_3$: in this case $g(\lambda, \lambda_F, \Lambda) \geq 0$, which is the optimality condition for (m, l) by (49). The optimal segmentation for $\Lambda \in [\underline{\Lambda}_{ml}, \overline{\Lambda}_{ml})$ and $\lambda < \lambda_3$ follows from Lemma 5-6 and (49) in exactly the same way as in Part 2(a). This yields Table (35).

Proof of Proposition 2.2(c). Suppose that $f_h(c_{\min}) < 0$ and $\frac{1}{\mu} \leq F(f_l^{-1}(c_{\max})) \cdot d$. By the analysis of **Step 2** the segmentation (h, m_{sd}) with strategic delay 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 (38). Refer to (45) for $\lambda \leq \lambda_P$ and (49) for $\lambda > \lambda_P$.

We translate this condition into conditions on λ and Λ. To this end, Lemma 7 characterizes the slope of $f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right)$ for *fixed* Λ, and Lemma 8 specifies its sign depending on λ and Λ.

LEMMA 7. *Fix $\mu > d^{-1}$ and Λ. 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*

$$f_h(c_h(\lambda))|_{\lambda=\min(\lambda_P,\Lambda)} = \begin{cases} f_h(c_{\min}) < 0, & \Lambda \le \lambda_P \\ f_h\left(\overline{F}^{-1}\left(\frac{\lambda_P}{\Lambda}\right)\right), & \Lambda > \lambda_P \end{cases}. \tag{63}$$

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*

$$f_h\left(c_h\left(\overline{\lambda}_h(\lambda)\right)\right)\big|_{\lambda=\min(\lambda_F,\Lambda)} = \begin{cases} f_h\left(\overline{F}^{-1}\left(\frac{\overline{\lambda}_h(\Lambda)}{\Lambda}\right)\right), & \Lambda \in (\lambda_P, \lambda_F) \\ c_{\max} > 0, & \Lambda \ge \lambda_F \end{cases}. \tag{64}$$

Proof. 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 (38). In Part 1. the values of $f_h(c_h(\lambda))$ for $\lambda = 0$ and $\lambda = \min(\lambda_P, \Lambda)$ follow because $c_h(\lambda) = \overline{F}^{-1}(\lambda/\Lambda)$ 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 $\lambda = \lambda_F \le \Lambda$ holds since $\overline{\lambda}_h(\lambda_F) = 0$ by (38) and $c_h(0) = c_{\max} = f_h(c_{\max})$.■

Henceforth write $f_h\left(c_h\left(\overline{\lambda}_h(\lambda), \Lambda\right)\right)$ to emphasize the dependence on Λ. Note: 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. *Fix $\mu > d^{-1}$. The sign of $f_h\left(c_h\left(\overline{\lambda}_h(\lambda), \Lambda\right)\right)$ depends as follows on λ and Λ. Let*

$$\underline{\Lambda}_{sd} \triangleq \frac{\mu}{2}\left(\frac{1}{x} + 1\right) - \sqrt{\frac{\mu^2}{4}\left(\frac{1}{x} - 1\right)^2 + \frac{\mu}{xd}} \text{ and } \overline{\Lambda}_{sd} \triangleq \frac{\lambda_P}{x}, \text{ where } x = \overline{F}\left(f_h^{-1}(0)\right) \in (0,1), \tag{65}$$

$f_h\left(c_h\left(\overline{\lambda}_h\left(\overline{\Lambda}_{sd}\right), \overline{\Lambda}_{sd}\right)\right) = 0$ 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 demand rate thresholds

$$\lambda_0 \triangleq \Lambda \overline{F}\left(f_h^{-1}(0)\right) \text{ and } \lambda_{sd} \triangleq \mu - \frac{\mu/d}{\mu - \lambda_0}. \tag{66}$$

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.*

Proof. 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 $f_h\left(c_h\left(\overline{\lambda}_h(\lambda), \Lambda\right)\right) \ge 0$ at $\lambda = \min(\lambda_P, \Lambda)$, then Lemma 7 implies that $f_h\left(c_h\left(\overline{\lambda}_h(\lambda), \Lambda\right)\right) \ge 0$ throughout. The condition holds if and only if $\Lambda \ge \overline{\Lambda}_{sd}$: by (63) in Lemma 7, for $\lambda = \min(\lambda_P, \Lambda)$ this virtual delay cost equals $f_h(c_{\min}) < 0$ if $\Lambda \le \lambda_P$, it equals $f_h(c_h(\lambda_P, \Lambda)) = f_h(\overline{F}^{-1}(\lambda_P/\Lambda))$ and increases in $\Lambda > \lambda_P$, with $\lim_{\Lambda\to\infty} f_h(c_h(\lambda_P, \Lambda)) = f_h(c_{\max}) = c_{\max} > 0$. Solving $f_h\left(c_h\left(\lambda_P, \overline{\Lambda}_{sd}\right)\right) = 0$ yields $\overline{\Lambda}_{sd} > \lambda_P$ as in (65).

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 $\lambda \le \lambda_0$. The condition holds if and only if $\Lambda < \underline{\Lambda}_{sd}$: we have

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

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 (65). 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.■

Refer to Proposition 2.2(c). The case $\Lambda < \underline{\Lambda}_{sd}$ in Table (36) is immediate from (45), (49) and Lemma 8.1. For $\Lambda \in [\underline{\Lambda}_{sd}, \overline{\Lambda}_{sd})$ Lemma 8.2. implies the transition $(h) - (h, m_{sd})$ for $\lambda < \lambda_{sd}$. For $\lambda \geq \lambda_{sd}$ Lemmas 5-6 and (49) imply (h, m) if $\Lambda \leq \Lambda_2$ or $(h, m) - (h, m, l)$ if $\Lambda > \Lambda_2$.■

Proof of Lemma 2. The maximum revenue satisfies $\Pi^*(\lambda, \mu) = \Pi(\lambda, \lambda^*_{mh}, \lambda^*_h, \mu)$ where $\Pi(\lambda, \lambda^*_{mh}, \lambda^*_h, \mu) \triangleq \lambda v - \Lambda \int_{c_{\min}}^{c_l(\lambda - \lambda^*_{mh})} f(x) f_l(x) \left(\frac{\mu}{(\mu - [\lambda - \Lambda F(x)])^2} - d \right) dx + \Lambda \int_{c_h(\lambda^*_h)}^{c_{\max}} f(x) f_h(x) \left(d - \frac{\mu}{\left(\mu - \Lambda \overline{F}(x)\right)^2} \right) dx$, and $\lambda^*_{mh}, \lambda^*_h$ depend on (λ, μ) as tabulated in (68). Its entries for $\lambda, \lambda^*_{mh}, \lambda^*_h$ and μ are from the proof of Proposition 2; refer in particular to (45) and (49) 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: $\lambda_P \triangleq \mu - \sqrt{\mu/d}$ and $\lambda_F \triangleq \mu - 1/d$ are well defined if $\mu^{-1} < d$, so $\lambda_P < \lambda_F$; $\overline{\lambda}_h(\lambda)$ is defined in (38); and λ_0 is defined in (43).

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	$\geq d$	0	0	0	0	0	0
(h)	$\leq \lambda_P$	$< d$	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)	$\lambda^*_h = \mu - \frac{\mu/d}{\mu - \lambda^*_{mh}} < \lambda_P < \lambda^*_{mh}$	$\in (0, 1)$		< 0	> 0
			(ii)	$f_l(c_l(\lambda - \lambda^*_{mh})) = f_h(c_h(\lambda^*_h))$				
(m, l)	$> \lambda_F$	$< d$	λ_F	0	0	> 0	0	0

(68)

We suppress the arguments of Π^*, Π, λ^*_{mh} and λ^*_h. Recall: $c_l(x) = F^{-1}(x/\Lambda)$, 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

$$\Pi^*_\lambda = \frac{\partial \Pi}{\partial \lambda} + \frac{\partial \Pi}{\partial \lambda^*_{mh}} \frac{\partial \lambda^*_{mh}}{\partial \lambda} + \frac{\partial \Pi}{\partial \lambda^*_h} \frac{\partial \lambda^*_h}{\partial \lambda}, \text{ where} \tag{69}$$

$$\frac{\partial \Pi}{\partial \lambda^*_{mh}} = f_l(c_l(\lambda - \lambda^*_{mh})) \left(\frac{\mu}{(\mu - \lambda^*_{mh})^2} - d \right) \geq 0, \ \frac{\partial^2 \Pi}{\partial \lambda \partial \lambda^*_{mh}} \geq 0, \text{and } \frac{\partial^2 \Pi}{\partial \mu \partial \lambda^*_{mh}} < 0, \tag{70}$$

$$\frac{\partial \Pi}{\partial \lambda} = v - \Lambda \int_{c_{\min}}^{c_l(\lambda - \lambda^*_{mh})} \frac{f(x) f_l(x) \, 2\mu}{(\mu - \lambda + \Lambda F(x))^3} dx - \frac{\partial \Pi}{\partial \lambda^*_{mh}}, \quad \frac{\partial^2 \Pi}{\partial \lambda^2} \leq 0, \text{and } \frac{\partial^2 \Pi}{\partial \lambda \partial \mu} > 0, \tag{71}$$

$$\frac{\partial \Pi}{\partial \lambda^*_h} = f_h(c_h(\lambda^*_h)) \left(d - \frac{\mu}{(\mu - \lambda^*_h)^2} \right) \geq 0, \ \frac{\partial^2 \Pi}{\partial \lambda^{*2}_h} \leq 0, \text{ and } \frac{\partial^2 \Pi}{\partial \mu \partial \lambda^*_h} \geq 0. \tag{72}$$

Part 1(i). It follows from (68) and (69)-(72) that $\Pi^*_\lambda = v$ for (h, m_{sd}) and $\Pi^*_\lambda \geq v$ for (h).

Part 1(ii). We first show $\Pi^*_{\lambda\lambda} < 0 < \Pi^*_{\lambda\mu}$ for (h,m,l). Table (68) 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 $f'_l c'_l > 0 > f'_h c'_h$:

$$sign\left(1 - \frac{\partial\lambda^*_{mh}}{\partial\lambda}\right) = -sign\left(\frac{\partial\lambda^*_h}{\partial\lambda}\right) = sign\left(\frac{\partial\lambda^*_{mh}}{\partial\lambda}\right), \text{ so } 0 < \frac{\partial\lambda^*_{mh}}{\partial\lambda} < 1. \tag{73}$$

$$sign\left(\frac{\partial\lambda^*_{mh}}{\partial\mu}\right) = sign\left(\frac{\partial\lambda^*_h}{\partial\mu}\right) \text{ and } \frac{1-\partial\lambda^*_{mh}/\partial\mu}{\mu-\lambda^*_{mh}} + \frac{1-\partial\lambda^*_h/\partial\mu}{\mu-\lambda^*_h} = \frac{1}{\mu}, \text{ so } 0 < \frac{\partial\lambda^*_{mh}}{\partial\mu} < 1. \tag{74}$$

The first equations in (73)-(74) 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 (47) in proof of Proposition 2 where $g(\lambda, \lambda^*_{mh}) = 0$ for (h,m,l). It follows that

$$\Pi^*_\lambda = \frac{\partial\Pi}{\partial\lambda} + \frac{\partial\lambda^*_{mh}}{\partial\lambda}\left[\frac{\partial\Pi}{\partial\lambda^*_{mh}} + \frac{\partial\Pi}{\partial\lambda^*_h}\frac{\partial\lambda^*_h}{\partial\lambda^*_{mh}}\right] = \frac{\partial\Pi}{\partial\lambda} \text{ for all } (\lambda,\mu) \text{ with } (h,m,l). \tag{75}$$

We show that the following holds:

$$\Pi^*_{\lambda\lambda} = \frac{\partial^2\Pi}{\partial\lambda^2} + \frac{\partial^2\Pi}{\partial\lambda\partial\lambda^*_{mh}}\frac{\partial\lambda^*_{mh}}{\partial\lambda} \leq -\frac{f_l\left(c_l\left(\lambda-\lambda^*_{mh}\right)\right)2\mu}{\left(\mu-\lambda^*_{mh}\right)^3} - \frac{\partial^2\Pi}{\partial\lambda^*_{mh}\partial\lambda}\left(1 - \frac{\partial\lambda^*_{mh}}{\partial\lambda}\right) < 0, \tag{76}$$

$$\Pi^*_{\lambda\mu} = \frac{\partial^2\Pi}{\partial\lambda\partial\mu} + \frac{\partial^2\Pi}{\partial\lambda\partial\lambda^*_{mh}}\frac{\partial\lambda^*_{mh}}{\partial\mu} > 0. \tag{77}$$

The equations for $\Pi^*_{\lambda\lambda}, \Pi^*_{\lambda\mu}$ hold by (75) and $\partial^2\Pi/\left(\partial\lambda\partial\lambda^*_h\right) = 0$ by (72). The first inequality in (76) holds by (70)-(71); the second by (70) and (73). The inequality in (77) holds by (70)-(71) and (74).

We next show $\Pi^*_{\lambda\lambda} \leq 0 \leq \Pi^*_{\lambda\mu}$ for any segmentation *other than* (h,m,l).
First consider $\Pi^*_{\lambda\lambda}$. Since $\partial^2\Pi/\left(\partial\lambda\partial\lambda^*_h\right) = 0$ by (72) and $\partial^2\lambda^*_{mh}/\partial\lambda^2 = 0$ by (68) we have

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

The terms in brackets are nonpositive by (72) and $\partial^2\lambda^*_h/\partial\lambda^2 \leq 0$ from (68). The other terms satisfy

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

by (70)-(71). The RHS is nonpositive since $\partial^2\Pi/\partial\lambda_{mh}\partial\lambda \geq 0$ by (70) and $\partial\lambda^*_{mh}/\partial\lambda \leq 1$ by (68).

Next consider $\Pi^*_{\lambda\mu}$. By (68) 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

$$\Pi^*_{\lambda\mu} = \left(\frac{\partial^2\Pi}{\partial\lambda\partial\mu} + \frac{\partial^2\Pi}{\partial\lambda^*_{mh}\partial\mu}\frac{\partial\lambda^*_{mh}}{\partial\lambda}\right) + \left(\frac{\partial^2\Pi}{\partial\lambda\partial\lambda^*_{mh}}\frac{\partial\lambda^*_{mh}}{\partial\mu}\right) + \left(\frac{d}{d\mu}\left[\frac{\partial\Pi}{\partial\lambda^*_h}\frac{\partial\lambda^*_h}{\partial\lambda}\right]\right). \tag{78}$$

We show that each bracket is nonnegative. For the first, (68) and (70)-(71) 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 (78) is nonnegative by (68) and (70). The third bracket of (78) is also nonnegative. For segmentations other than (h,m) and (h,m,l) we have $\partial\lambda^*_h/\partial\lambda = 0$ or $= 1$ by (68) and $\partial^2\Pi/\left(\partial\lambda^*_h\partial\mu\right) \geq 0$ by (72). For (h,m) substitute for $\lambda^*_h = \mu - \mu/d\left(\mu-\lambda\right)^{-1}$ from (68) to get:

$$\frac{\partial\Pi}{\partial\lambda^*_h}\frac{\partial\lambda^*_h}{\partial\lambda} = f_h\left(c_h\left(\lambda^*_h\right)\right)\left(d - \frac{\mu}{\left(\mu-\lambda^*_h\right)^2}\right)\frac{\partial\lambda^*_h}{\partial\lambda} = v + f_h\left(c_h\left(\lambda^*_h\right)\right)\left(d - \frac{\mu}{(\mu-\lambda)^2}\right). \tag{79}$$

This expression increases in μ: the virtual delay cost $f_h(c_h(\lambda_h^*)) \geq 0$ decreases in μ since $f_h' c_h' < 0$ and $\partial \lambda_h^* / \partial \mu > 0$ by (68), and its multiplier is negative ($\lambda > \lambda_P$) and increases in μ.

We omit the remaining straightforward checks that $\Pi_{\lambda\lambda}^* < 0 < \Pi_{\lambda\mu}^*$ for (l), (h,m),(h,l), and (m,l).

Part 2. The function $\Pi^*(\lambda,\mu) = \Pi(\lambda, \lambda_{mh}^*, \lambda_h^*, \mu)$ is defined piecewise: the functions $\lambda_{mh}^*(\lambda,\mu)$ and $\lambda_h^*(\lambda,\mu)$ depend on the optimal segmentations which vary with (λ,μ) as specified by Proposition 2. By the definition of Π and table (68) all first and second order derivates of Π, λ_{mh}^* and λ_h^* with respect to (λ,μ) are continuous for *each* segmentation. Therefore so are the functions $\Pi_\lambda^*(\lambda,\mu)$, $\Pi_{\lambda\lambda}^*(\lambda,\mu)$ and $\Pi_{\lambda\mu}^*(\lambda,\mu)$. 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	$f_h(c_h(\lambda_0))=0$	8.1
(h,m_{sd})	(h,m)	$\lambda=\lambda_{sd},\ \mu>d^{-1}$	λ	$f_h\left(c_h\left(\overline{\lambda}_h(\lambda_{sd})\right)\right)=0$	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	$f_h\left(c_h\left(\overline{\lambda}_h(\lambda_1)\right)\right)=f_l(c_l(\lambda-\lambda_{mh}^*))=c_{\min}$	5.1(b)
(h,m,l)	(m,l)	$\lambda=\lambda_3,\ \mu>d^{-1}$	λ_F	$f_l(c_l(\lambda_3-\lambda_F))=c_{\max}$	5.2(a)
(h,l)	(h,m,l)	$\lambda=\lambda_2,\ \mu>d^{-1}$	λ_P	$f_h(c_h(\lambda_P))=f_l(c_l(\lambda_2-\lambda_P))$	6.2

The following facts establish (i) and (ii). 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 segmentations) agree, and ditto for $\Pi_{\lambda u}^*$. We omit these checks; they are straightforward using table (68) and the formulae for Π_λ^*, $\Pi_{\lambda\lambda}^*$ and $\Pi_{\lambda\mu}^*$ derived above.■

Proof of Theorem 1. This result is a direct implication of the optimal segmentation for an arrival rate Proposition 2 and Lemma 2.■

Proof of Theorem 2. Let $\lambda^*(\mu) = \arg\{\max_\lambda \Pi^*(\lambda,\mu) \text{ s.t. } \lambda \in [0,\Lambda] \cap [0,\mu)\}$. We first prove key properties of the thresholds μ_H, defined by (22), and μ_{SD}, defined by (23).

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 $\mu \geq \mu_H$.* Recall that $\lambda_P \triangleq \mu - \sqrt{\mu/d}$ 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 $\mu > \mu_{SD}$.* Recall the threshold $\underline{\Lambda}_{sd}$ from Proposition 2.2(c) and its proof; it is defined in (65) 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 (65) 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 diffentiable 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.* 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 $f_l(c^*_l(\Lambda)) \leq f_h(c^*_h(\Lambda))$ (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 (68). 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 that pooling is not optimal for some fixed $\mu < \mu_H$. Then pooling is also not optimal for all $\mu' < \mu$.* If $\mu \leq d^{-1}$ 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 $f_l(F^{-1}(\lambda^*_l(\mu)/\Lambda)) \leq f_h(\overline{F}^{-1}(\lambda^*_h(\mu)/\Lambda))$ by Lemma 1.2, and

$$0 = \Pi^*_\lambda(\lambda^*(\mu), \mu) \Leftrightarrow v = \Lambda \int_{c_{\min}}^{F^{-1}(\lambda^*_l(\mu)/\Lambda)} \frac{f(x) f_l(x)\, 2\mu}{(\mu - \lambda^*(\mu) + \Lambda F(x))^3} dx = \Lambda \int_{c_{\min}}^{F^{-1}(\lambda^*_l(\mu)/\Lambda)} \frac{f(x) f_l(x)\, 2\mu}{(\sqrt{\mu/d} - \lambda^*_l(\mu) + \Lambda F(x))^3} dx, \tag{80}$$

where the second equation holds since $\mu - \lambda^*(\mu) = \sqrt{\mu/d} - \lambda^*_l(\mu)$. We have $\lambda^{*\prime}_l(\mu) > 0$, since the RHS of the equation strictly decreases in μ for fixed $\lambda^*_l(\mu)$ and strictly increases in $\lambda^*_l(\mu)$ for fixed μ. Noting that $\lambda^{*\prime}_h(\mu) = 1 - 1/(2\sqrt{\mu d}) > 0$ 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 $\mu \leq \mu_P$ 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 $\mu_A \leq d^{-1}$, which holds iff $\Pi^*_\lambda(\Lambda, \mu) \geq 0$ 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 (80) yields

$$\Pi^*_\lambda(\Lambda, \mu)|_{\mu = d^{-1}} = v_A - \Lambda \int_{c_{\min}}^{c_{\max}} \frac{f(x) f_l(x)\, 2d^2}{(1 - d\Lambda \overline{F}(x))^3} dx = 0. \tag{81}$$

Threshold v_P. By Part 2, we need $\mu_P = d^{-1}$. If $v \geq v_A$ then $\lambda^*(\mu) = \Lambda$ for $\mu \geq d^{-1}$, 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 $f_l(F^{-1}(\lambda^*_l(\mu)/\Lambda)) - f_h(\overline{F}^{-1}(\lambda^*_h(\mu)/\Lambda)) \geq 0$ 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 $\lambda^*_l(\mu) \geq \Lambda F\left(f_l^{-1}(c_{\max})\right)$. Substituting in (80) 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 $\mu = d^{-1}$. ■