Screening and Reallocation in Usage-Priced Computing

Georgios Petropoulos*

August 23, 2026

Abstract

Cloud providers sell identical machines through a guaranteed tier and a cheaper tier that may be reclaimed mid-job. Interruption is usually read as a symptom of scarcity. We show it is an instrument: providers screen by interruption, and a seller who can set the reclamation rate rather than inherit it strictly prefers to, and prefers a rate that varies with the buyer. The reallocative account cannot carry the explanation. Screening is two-sided: the seller sets the rate *away from* the reallocative one in whichever direction sorting points: above it where exposure is free to supply, below it where it must be purchased, so manufacturing risk is one half of the instrument and not the whole of it. Under an optimal schedule exposure falls with value. The bottom of the market is held at the worst reliability consistent with buying at all and pays its entire surplus. The guaranteed rate ceases to be an instrument, becoming the residual of the schedule beneath it. Refinement beyond two or three grades is worth quadratically less, and the surplus it would create accrues to buyers rather than to the seller.

Keywords: cloud computing; spot instances; capacity allocation; mechanism design; usage-based pricing
JEL: D42; D82; L12; L86

*Petropoulos: University of Southern California, georgios@marshall.usc.edu

1 Introduction

A cloud provider sells time on a virtual machine under two plans. An on-demand customer pays a posted rate per unit of use and keeps the machine until she releases it. A spot customer pays less and accepts that the provider may reclaim the machine on short notice. The hardware can be identical; what differs is reliability. Amazon Web Services, Microsoft Azure and Google Cloud all operate variants of this arrangement.

A discount alone cannot sort these customers. If two contracts promise the same reliability, every buyer prefers the cheaper one, so a menu of rates collapses to its lowest element. Something must make the discount unattractive to the customers the provider wants to keep on the expensive plan, and in this market that something is the risk of losing the machine. This paper asks what determines that risk.

The usual answer is scarcity: the machine is taken back because someone else wants it, and the discounted tier is a way of selling capacity the provider may need. We develop that account as a benchmark, and two results say it cannot carry the explanation. Its value is bounded by the margin earned on the buyers a single posted rate would have excluded, a narrow band whenever longer jobs are not also the ones willing to pay most per machine-hour. And it does not survive scale: displacing a tenant requires that no machine be idle, while selling a discounted tier requires that some machine be, and in a large pool those two cannot hold at once.

What is left is a second function. Providers screen by interruption. A seller who can set the reclamation rate rather than inherit it strictly prefers to, and prefers a rate that varies across buyers. The reason is that interruption is costly in proportion to how long a buyer needs the machine, and how long she needs it is what her value determines, so a posted risk that looks identical to everyone is in fact a different price to each of them. That is the paper's organising claim, and the ranking that establishes it rests on revealed preference rather than on any calibration: a seller free to choose the schedule can always decline to depart from whatever demand would have generated, so choosing cannot be worse, and it is strictly better whenever the exposure she would like to deliver is not the same for every buyer.

Manufacturing risk is one half of that instrument and not the whole of it. The natural way to state the point—that providers reclaim more often than reallocation requires—names only one of the two directions. Sorting needs nothing more than that some risk be posted, because the reliability gap between two contracts widens with value whatever the level of the risk. What the level buys is the strength of the screen, and whether the discounted tier can reach buyers a single rate would exclude. So the seller sets the rate away from the reallocative one in whichever direction sorting points: above it where exposure is free to supply, since a reclamation matched to no arrival forgoes nothing, and below it where it must be purchased, since holding a tenant safe means turning arrivals away. Both directions are screening. Only one is free. Which one appears is a matter of how scarce capacity is rather than of the seller's motives, and when capacity is scarce enough the whole schedule can lie below the reallocative rate.

A second contribution concerns what a menu can achieve. It is tempting to conclude—as an

earlier version of this paper did—that efficient reallocation is unattainable because the seller never observes a valuation. That is wrong, and correcting it is what makes the rest possible. A menu makes buyers condition on their own type. If the provider posts several ordered tiers and reclaims from a tenant whenever an arriving buyer chooses a tier above hers, then each tenant faces exactly the reclamation intensity that efficiency prescribes for the marginal buyer in her own tier, and refining the menu drives that toward the efficient schedule from below, with sorting monotone at every step and no valuation ever observed. Tier choice is exactly informative enough.

Why, then, are observed menus short? Because refinement pays rapidly less. We show that the shortfall from a menu of a given length falls quadratically in the number of contracts, so the marginal value of an additional grade falls faster still, and a two- or three-element menu already captures most of what the full schedule earns over a single rate. Whether a provider stops at two or at three is a comparison of a small increment against the cost of running another product line, which the model does not price; beyond a few grades the question does not arise. The coarseness of observed menus is therefore consistent with profit maximisation and needs no appeal to informational or technological limits. The gap between the schedule a provider posts and the efficient one is a gap in rent extraction, not in information, and the surplus that finer grades would create accrues to buyers rather than to her.

Three arrangements make the argument, nested inside one another and each solved in the same environment. The first gives the seller a single posted rate and no reclamation. This is the market before any interruptible tier exists, and it already contains the feature that drives everything: she sells machine-hours rather than units, so what she faces is the demand for hours, and the optimum can be found from a rate experiment without knowing anything about individual valuations. The second lets her post a discount and reclaim only when an arriving buyer takes the expensive tier, so the interruption rate is inherited rather than chosen. This isolates the transmission function, and the rate that emerges is the efficient intensity of the marginal switcher applied to everyone, so every discounted buyer is under-exposed relative to what efficiency would prescribe for her. The arrangement is also fragile in a way worth noting: the risk being sold is generated by the buyers who decline it, so a discount deep enough to empty the expensive tier destroys the product it is pricing. The third lets her choose exposure buyer by buyer, and that is where screening operates. Optimal exposure falls with value. A positive mass of buyers at the bottom is held at exactly the worst reliability consistent with their buying at all and pays its entire surplus, which no short menu can do, and the guaranteed rate ceases to be an instrument, becoming the residual of the schedule beneath it.

Both sides of this market are firms, and the closing sections draw out what each should do. The provider can price the guaranteed tier from an elasticity she can measure; designing a full schedule is harder, because it requires knowing how long buyers need machines throughout the market, including the buyers she is currently not serving and therefore not observing. The buyer should read a published reclamation rate as a statement about the composition of other people's demand rather than about the hardware, and should evaluate reliability at her own runtime rather than

at the pool average. The two interact: buyer-side investment in restartability erodes precisely the elasticity the provider is exploiting, so improvements in checkpointing should compress discounts.

The paper proceeds from the environment and the efficiency benchmark, through the three arrangements in the order above, to the implications for both sides and the conclusion. Proofs of results stated in the body are collected in an appendix; results stated in an appendix are proved where they appear. Further appendices delimit the runtime assumption, describe the numerical implementation, treat interruption that falls short of total loss, and take up multi-unit demand, pools and privately known runtime—the settings in which the provider inherits the interruption rate rather than setting it, and therefore the boundary of the screening results.

2 Related literature

The paper draws on four strands and contributes something different to each.

Quality discrimination and damaged goods. In the tradition running from Mussa and Rosen (1978) through Deneckere and McAfee (1996), Varian (2000), Belleflamme (2005), McAfee (2007) and Bhargava and Choudhary (2008), a seller chooses a quality level and heterogeneity in marginal willingness to pay does the sorting. The physical resource here is homogeneous and the seller chooses a reclamation rate; delivered quality, measured by the probability a job completes, is determined jointly by that rate and by the buyer's own required runtime. The novelty is not that a degraded version can sort buyers but that the degradation is endogenous to the buyer: two customers offered the same contract receive different reliability. That is what allows a single degraded usage-priced tier to reach buyers whom a uniform rate excludes. It is also why the same posted contract can be a screening device and a reallocation device at once, and why the two can be separated only by comparing arrangements rather than by inspecting a single one—a distinction that does not arise where quality is not a function of how the buyer uses the good.

Probabilistic selling and stochastic mechanisms. Randomization can raise revenue with risk aversion, multiple goods or multidimensional types (Matthews, 1983; Manelli and Vincent, 2006; Pavlov, 2011; Hart and Reny, 2015), while adding nothing in the sale of a single good to a risk-neutral buyer (Riley and Zeckhauser, 1983); probabilistic products can also conceal which of several horizontally differentiated items will be delivered (Fay and Xie, 2008; Zhang et al., 2015). The buyer here is risk neutral and values one completed job, so that benchmark sharpens what is doing the work. The hazard is not a lottery over a fixed allocation. Its incidence depends on the buyer's own duration, and it is that dependence, not the randomness as such, that changes the cross-type ordering of a linear usage price. The main results require no lottery over menu entries. Risk neutrality is also doing visible work rather than serving convenience: under concave utility the risk premium on a bill scales with the square of required runtime, so duration uncertainty would create a duration-dependent wedge by itself and a bill cap would screen with no reclamation at all.

Usage pricing, priority and capacity. A long literature prices access to a shared facility by service grade (Harris and Raviv, 1981; Chao and Wilson, 1987; Wilson, 1989; Mendelson and Whang, 1990) and a more recent one studies usage pricing for information goods and cloud services (Sundararajan, 2004; Ma and Seidmann, 2015; Chen and Huang, 2016; Chen et al., 2019; Kilcioglu et al., 2017), with Bergemann and Deb (2025) and Bergemann and Wang (2025) treating robust and committed-spend cloud contracts. Closest in method are Afèche (2013) and Afèche and Pavlin (2016), where a queueing firm degrades service deliberately—strategic delay, or a menu of price and lead time—to separate customers who differ in delay cost. The logic is ours: quality is withheld to sort, not because withholding is cheap. What differs is where the degradation lands. Delay falls on the customer directly and in the same units for everyone, so the sorting statistic is her delay cost; a reclamation hazard falls on the machine, and its harm depends on how long she needs it, so the statistic is her runtime and one posted contract delivers different reliability to different buyers. The same distinction separates us from Mendelson and Whang (1990), the closest antecedent in substance, where grades in a shared computing facility differ in priority. Relative to this strand the contribution is twofold: the reliability a grade delivers can be an equilibrium object rather than a design parameter, since where the seller merely reclaims on demand the rate a tenant faces is fixed by how many buyers chose a higher grade; and where she does design it, the value of doing so can be measured against that benchmark, which is what separates screening from transmission.

Spot and preemptible computing. The immediate antecedents are Abhishek et al. (2012), Xu and Li (2013), Huang et al. (2015), Dierks et al. (2021), Chen et al. (2021) and Dierks and Seuken (2022), and three of these are close enough to require an explicit contrast. On the empirical side, Agmon Ben-Yehuda et al. (2013) reverse-engineer EC2 spot prices and conclude they were not generated by the bids they were presented as clearing, which is a reason to treat the posted terms of an interruptible tier as instruments the provider selects rather than as market outcomes—the view taken throughout this paper, and the reason Section 7 asks what a buyer can infer from published figures alone.

Huang et al. (2015) treat interruptible service as a damaged good. Their monopolist sets an upfront price for a reserved contract and a price per task for on-demand service, and chooses an interruption probability; marginal cost is negligible and capacity is unconstrained, so degradation there is deliberate and not driven by scarcity. Two modelling choices keep duration out of the story: all tasks are assumed to take the same time, so every task faces the same risk, and an interrupted task is not billed. The second is consequential for us, because charging only for completed work returns willingness to pay per billed unit to its guaranteed level and removes the force studied here.

Dierks and Seuken (2022) study a provider who posts a fixed price for non-preemptible instances and additionally sells reliably idle capacity on a preemptible spot market where users bid. A job is preempted when an idle instance is recalled or when a higher bid arrives, and each preemption costs its user a fixed increment of runtime. Their main result is a condition under which opening the spot market raises profit and is Pareto improving; the gain is a cost-side gain, since idle instances have

already incurred their fixed cost. Duration cannot sort there, since all job classes share the same mean service time and the per-preemption time loss is class independent; the private dimension is the user's waiting cost and the sorting device is the bid.

Chen et al. (2021) consider a provider whose surplus capacity, left after serving fault-intolerant demand, is sold at a discount to customers whose jobs are fault tolerant. Customers differ in willingness to pay and in sensitivity to delay, and hold a common belief about the average preemption probability that a diffusion approximation renders consistent. Preemption delays rather than destroys, and the customer is billed only while her instance runs. Their discounted price is, like ours, a rate per unit of running time, so the difference is not linearity of the tariff. It is that both the delay cost and the bill scale linearly in a job's workload, so workload cancels from the customer's comparison and a second private dimension does the sorting.

Three claims that might be mistaken for ours are therefore already in this literature, and none is claimed here: that interruption can support segmentation; that deliberate degradation can be optimal when no scarcity appears anywhere in the model (Huang et al., 2015); and that a preemptible tier alongside guaranteed service can raise profit while leaving users no worse off (Dierks and Seuken, 2022).

What is new is the separation of the two functions interruption performs. In each of these papers the interruption probability is a scalar—chosen, as in Huang et al. (2015), or imposed by capacity, as in Dierks and Seuken (2022) and Chen et al. (2021)—and the question is its level. Here it is both: a rate that demand generates, and a schedule the seller may choose instead. Separating them gives a profit ranking in which the second increment is strictly positive, so the screening function is not incidental to the arrangement, and it makes manufacturing one half of an instrument rather than the instrument itself: the seller sets the rate away from the reallocative one in either direction, and which direction appears is a matter of parameters rather than of motives, in the vocabulary of (3).

Three further consequences distinguish the account. First, we bound what the reallocative story can explain rather than assuming it away: the value of a purely demand-driven tier is limited to the margin on the buyers a single rate would exclude, and the arrangement does not survive the fluid limit, since displacement requires a fully occupied pool while a discounted tier requires an idle machine. Neither point is available in a model that fixes the interruption probability exogenously, and together they are why we read observed reclamation as chosen rather than inherited. Second, the departure from efficient reallocation has no uniform sign: a demand-driven rule under-exposes every discounted buyer, which is the opposite of the usual presumption; concentration in a pool can reverse that; and under a chosen schedule the sign varies within a single market. Whether a buyer is over- or under-exposed is therefore a property of the buyer and of how scarce capacity is, not of the provider's motives. Third, refinement is feasible and incentive compatible at every resolution but worth quadratically less, so short menus are consistent with profit maximisation without any appeal to menu costs. None of these papers gives the buyer a decision rule; we do, and it requires only quantities she can observe.

3 Environment

A seller owns one machine. Time is continuous and the machine serves at most one job.

The machine's alternative use. Requests for the machine arrive as a Poisson process of rate $\lambda \geq 0$, each drawing a completion value from the same distribution F as the incumbent's, and each yielding the seller a margin $\mu > 0$ if served. While the machine is free a request is served on arrival, so an idle machine earns $\mu\lambda$ per unit of time. While it is occupied a request can be served only by freeing the machine, that is by reclaiming the incumbent. A contract permitting reclamation at rate h therefore serves $\min\{h, \lambda\}$ requests per unit of occupied time, and the seller's opportunity cost of occupancy is

$$c(h) = \mu\big(\lambda - h\big)^+ \quad \text{per unit of occupied time.} \tag{1}$$

Two features of (1) are used throughout. It does not depend on the incumbent's type: the seller does not observe v, and she earns μ from any request she serves, so nothing in her cost can depend on who holds the machine. And it is zero for every $h \geq \lambda$. We write

$$\kappa := \mu\lambda \tag{2}$$

for the price of a machine-hour, which is what a guaranteed tenant costs since $h = 0$ in (1). A tenant reclaimed at rate h costs less, and we write $\kappa(h) := c(h) = \mu(\lambda - h)^+$ for that cost, so that $\kappa = \kappa(0)$ and $\kappa(h) = 0$ for $h \geq \lambda$. The case $\kappa = 0$, in which the machine has no valuable alternative use, is admitted throughout.

Corollary 1 (Reclamation beyond the arrival rate is costless)**.** *$c(h) = 0$ for every $h \geq \lambda$. Once the hazard is at least the arrival rate every request is already met by reclaiming, so raising the hazard further forgoes nothing: the only costs of a higher hazard are the completion value destroyed and the billed time forgone. Below λ the seller is buying reliability for the tenant at μ per request she declines to serve.*

Corollary 1 presumes that capacity freed by a reclamation is refilled. That is the commercial-pool case, and we take it as the baseline; Section 8 records what changes when it fails.

The buyer. A risk-neutral buyer privately observes the completion value v of her job, distributed with continuous density $f > 0$ on $[\underline{v}, \overline{v}]$, $0 < \underline{v} < \overline{v} < \infty$. Completion yields v; reclamation before completion yields zero, which is the appropriate specification for jobs with costly restart or lost state.[1] Write

$$R_0(v) := \frac{v}{b(v)}, \qquad \eta_b(v) := \frac{v\, b'(v)}{b(v)}.$$

[1] This is stronger than the mechanism requires. What the results depend on is that the buyer bear the cost of usage that produced no output, not that she lose the objective. Appendix D shows that interruption which merely forces a *restart* leaves every result unchanged, while interruption that *salvages* value does not, and gives the threshold at which the screen fails.

Three benchmarks, and the words for them. Interruption is compared in this paper against three different rates, and they are ordered. For any tenant on the discounted tier, with v_1 the switch cutoff of Section 5,

$$\underbrace{\alpha(v_1)}_{\text{what demand alone produces}} \leq \underbrace{\alpha(v)}_{\text{what efficiency prescribes}} \leq \underbrace{\lambda}_{\text{what any arrival could justify}}, \tag{3}$$

the first inequality because α is decreasing and $v \leq v_1$, the second because $F \geq 0$. We reserve a word for each. A posted hazard is *excess* if $h > \alpha(v_1)$, so that the seller reclaims more often than the arrival of higher-paying buyers alone would produce; type v is *over-exposed* if $h > \alpha(v)$, so that she is reclaimed more often than efficient reallocation prescribes for her; and the hazard is *manufactured* if $h > \lambda$, so that some reclamations are matched to no arrival at all and are, by Corollary 1, costless at the margin. By (3) the three are nested: manufactured implies over-exposed implies excess, and none of the converses holds. In particular a hazard can be excess and still leave the lower part of the discounted band under-exposed, which is the generic case when a single number is posted against a schedule $\alpha(\cdot)$ that varies; and uniform under-exposure, $h < \alpha(v)$ for every served v, holds if and only if $h \leq \alpha(v_1)$. Appendix E notes one qualification: in a pool the per-tenant hazard is concentrated by the reciprocal of the discounted share, so it can exceed λ with nothing manufactured, because it is the aggregate reclamation rate and not the per-tenant one that must be matched to arrivals.

Efficiency, which is a different matter from cost. A planner would free the machine only for a request worth more than the job it interrupts, and by thinning the intensity of such requests facing a tenant of type v is

$$\alpha(v) = \lambda\big(1 - F(v)\big), \tag{4}$$

constant in her elapsed runtime but strictly decreasing in her type. Write $\beta(v) := \alpha(v)b(v)$ for the exposure efficient reallocation delivers.

Two properties of this benchmark matter below. Under efficient reallocation the per-tenant reclamation time is exponential with rate $\alpha(v)$. The induced hazard is therefore constant in the tenant's elapsed runtime and blind to her identity, but strictly decreasing in her type. The exposure it generates, $\beta(v)$, rises while $\eta_b(v) > vf(v)/(1 - F(v))$ and falls thereafter; it is single-peaked whenever $vf/(1 - F) - \eta_b$ is increasing, which holds if F has an increasing generalised failure rate and η_b is non-increasing, and in particular in the uniform–power family.

Section 3 separates two properties usually imposed together. *Memorylessness* is a consequence of reallocating efficiently: Poisson demand makes the waiting time until a superior arrival exponential, so the efficient rule cannot condition on elapsed runtime even if the seller wished. *Anonymity across types* is not; efficient reallocation is type-dependent by construction. It is important that α is a welfare object and not a price. It is the rate at which a planner would want the machine back, not a rate at which the seller loses anything: her margin is μ whoever is served, and by (1) her cost is type-free. Sections 5 and 6 compare the rules she posts with (4); Lemma 5 shows that

incentive compatibility does not prevent them from coinciding, and Section 5.4 explains why the seller nonetheless does not choose them to.

Uncertain duration. The job needs a contiguous span $\widetilde{b} = b(v)\varepsilon$, where b is common knowledge and continuously differentiable and $\varepsilon > 0$ is a shock with mean one, distribution G and density g, independent of v and unobserved by both parties when the contract is signed. Let

$$L(x) := \mathbb{E}\big[e^{-x\varepsilon}\big], \qquad \mathfrak{d}(x) := \frac{x\,L(x)}{1-L(x)}, \qquad \mathfrak{M}(x) := \frac{x}{1-L(x)}, \tag{5}$$

with $\mathfrak{d}(0) = 1$, and let

$$M_L(x) := \frac{-x\,L'(x)}{L(x)\big[1-L(x)\big]} \tag{6}$$

denote the elasticity of the exposure factor. Deterministic duration is the case $\varepsilon \equiv 1$, where $L(x) = e^{-x}$ and $M_L = \mathfrak{M}$.

Two features of this specification should be separated, because they are easily conflated. The shock ε makes the realized span uncertain, but it is unknown to *both* parties and its distribution is common knowledge, so it adds risk without adding private information: two buyers with the same v remain identical objects to the seller. The substantive restriction is elsewhere, in $b(\cdot)$ being a common-knowledge function of v at all, so that value pins down required runtime. That is a convenience rather than the economic content of the model. Appendix E.6 lets (v, b) be jointly distributed with both privately observed and shows what survives: the guaranteed tier is then taken by the types above an increasing curve $v^*(b)$ rather than above a point, sorting still runs on value per machine-hour rather than on value, and the induced hazard is still an anonymous scalar, while the cutoff parametrisation and the identity $h = \alpha(v_1)$ are lost. Read in those terms, Assumption 1 below says that value and required runtime are positively but less than proportionally associated, so that ranking buyers by v and ranking them by v/b do not coincide. That is the restriction the mechanism needs, and it is testable in a way that a deterministic $b(\cdot)$ is not.

Contracts. A contract is a pair (h, w): the machine is reclaimed at the first arrival of a Poisson process of rate $h \geq 0$, and the buyer pays $w \geq 0$ per unit of realized runtime. Writing $x = hb(v)$ for cumulative exposure, a type-v buyer completes with probability $L(x)$, is billed for expected time $\tau(h, v) = (1 - L(x))/h$, and obtains

$$U(h, w; v) = L(x)\,v - w\,\tau(h, v), \tag{7}$$

so she participates exactly when

$$w \;\leq\; R_h(v) := \frac{h\,v\,L(x)}{1-L(x)} = R_0(v)\,\mathfrak{d}\big(hb(v)\big). \tag{8}$$

We adopt the seller-favourable convention at exact indifference; no strict conclusion depends on it.

A guaranteed contract sets $h = 0$: the job completes with probability one and the expected bill is $w\,b(v)\,\mathbb{E}[\varepsilon] = w\,b(v)$. Because the tariff is linear in realized time and the buyer risk neutral, only the mean of the shock enters willingness to pay, so the guaranteed side of the market is exactly as it would be under deterministic duration. The shock is idle until a hazard is posted, at which point it enters through L and M_L and governs whether exposure can screen at all.

Assumption 1 (Sub-proportional runtime)**.** $0 < \eta_b(v) < 1$ for every $v \in [\underline{v}, \overline{v}]$.

Assumption 1 says required runtime rises with completion value but less than proportionally, so b and $R_0 = v/b$ are both strictly increasing: longer jobs are worth more in total but less per machine-hour. Proposition B.1 in Appendix B shows that interruption has no screening role on either side of that interval, so the assumption marks the boundaries of the phenomenon rather than selecting a convenient case within it.

Footprint rather than duration. Nothing below requires that a job occupy one machine. If a job needs $m(v)$ machines simultaneously for a span $b(v)$, is lost when any one of them is reclaimed, and each is reclaimed independently, then it completes with probability $L\big(h\varphi(v)\big)$ and is billed for $\big(1 - L(h\varphi)\big)/h$ machine-hours, where

$$\varphi(v) := m(v)\,b(v) \tag{9}$$

is its footprint. Both expressions depend on (m, b) only through φ, so every result below holds with b replaced by φ and η_b by $\eta_\varphi = \eta_m + \eta_b$. Appendix E states this formally and draws out three consequences, of which the most useful is that the size–value link substitutes for the duration–value link: the mechanism does not require that more valuable jobs run *longer*, only that they be *larger*. We write b throughout for readability.

Two contracts sharing a hazard cannot form a menu. By (7), the payoff difference between (h, w_1) and (h, w_2) is

$$U(h, w_1; v) - U(h, w_2; v) = -(w_1 - w_2)\,\tau(h, v), \tag{10}$$

and expected billed time is strictly positive, so every type ranks rates identically and any menu of equally reliable metered contracts collapses to its lowest element. Reliability differentiation, not a transfer, is what allows more than one metered price to coexist.

Finally, define the frontier at which exposure begins to reverse the ordering of willingness to pay. Differentiating (8) logarithmically at a constant hazard,

$$\frac{d\log R_h}{d\log v} = 1 - \eta_b(v) M_L(x), \qquad x = hb(v), \tag{11}$$

so a discounted contract reaches types a guaranteed rate excludes, without tempting anyone above them, exactly where $\eta_b M_L > 1$. Write

$$K := \sup_{x>0} M_L(x) \;\in\; [1, \infty], \tag{12}$$

which is at least one because $M_L(x) \to 1$ at the origin for every mean-one shock; some hazard reverses the ordering at v if and only if $\eta_b(v)K > 1$. For a gamma shock of shape k and mean one, $L(x) = (1+x/k)^{-k}$ and $K = \max\{1, k\}$. It is useful to express the requirement as a statement about reliability: with $q_c = L(x)$ the completion probability at exposure x, the condition $\eta_b M_L(x) > 1$ becomes

$$\eta_b > \Psi_L(q_c) := \frac{1}{M_L\big(L^{-1}(q_c)\big)}, \qquad \text{and for a gamma shock,} \quad \Psi_k(q) = \frac{1-q}{k(1-q^{1/k})}, \tag{13}$$

so $\Psi_L^{-1}(\eta_b)$ is the supremum completion probability consistent with screening. Sections 5 and 6 show that (13) governs the reliability delivered at the bottom of the market, and that $\eta_b K > 1$ is a condition on *how* a discounted tier earns money rather than on whether one exists.

4 One rate: the guaranteed benchmark

We begin where the market begins: one posted rate per unit of usage and no interruption. A buyer who pays keeps the machine until her job is done. This is the arrangement every provider offers, it is the arrangement against which everything later is measured, and it already contains the two features that drive the paper— that the seller sells hours rather than units, and that how many hours a buyer needs is what her value determines.

There is no reclamation here, so there is no reallocation: by (1) occupancy costs κ per hour whatever the tenant's type, and $\alpha(\cdot)$ plays no role in this section. A buyer of type v pays w per unit of realized time, so her expected bill is $w\,b(v)$ and she buys exactly when $R_0(v) \geq w$. Under Assumption 1 the buyers are an interval at the top,

$$\{v : R_0(v) \geq w\} = [v_0(w), \overline{v}], \qquad v_0(w) := R_0^{-1}(w), \tag{14}$$

and expected billed machine-hours per arriving buyer are

$$M(w) := \int_{v_0(w)}^{\overline{v}} b(v)\, dF(v) = B\big(v_0(w)\big), \qquad B(t) := \int_t^{\overline{v}} b\, dF. \tag{15}$$

The seller's payoff is a price times a quantity of hours,

$$\Pi_H(w) = \big(w - \kappa\big)M(w), \tag{16}$$

and v_0 is not one of her instruments: it is the type her posted rate leaves indifferent, defined by (14) and used only to evaluate (15). It is convenient to record the runtime-weighted law

$$dF_b := \frac{b\, dF}{\mathbb{E}[b]}, \qquad \rho_b(t) := \frac{1 - F_b(t)}{f_b(t)} = \frac{B(t)}{b(t)f(t)}. \tag{17}$$

Assumption 2 (Regularity)**.** $\Phi(t) := t - \kappa b(t) - \big(1 - \eta_b(t)\big)\rho_b(t)$ changes sign at most once on

$[\underline{v}, \overline{v}]$, from negative to positive.

Proposition 1 (The optimal guaranteed rate)**.** *Under Assumptions 1 and 2:*

(i) (Neutrality of the shock.) Π_H *depends on the duration shock only through* $\mathbb{E}[\varepsilon] = 1$*; the problem is identical under deterministic duration.*

(ii) (Markup form.) *At an interior optimum*

$$\frac{w^* - \kappa}{w^*} = \frac{1}{\varepsilon(w^*)}, \qquad \varepsilon(w) := -\frac{d\log M}{d\log w} = \frac{v_0\, b(v_0) f(v_0)}{\big(1 - \eta_b(v_0)\big) B(v_0)}, \tag{18}$$

where ε *is the elasticity of billed* hours*. If* $\kappa = 0$ *the condition is* $\varepsilon(w^*) = 1$*.*

(iii) (Cutoff form.) *Equivalently* $w^* = R_0(v_0)$ *where* v_0 *is the unique zero of* Φ*,*

$$v_0 - \kappa\, b(v_0) = \big(1 - \eta_b(v_0)\big)\, \rho_b(v_0). \tag{19}$$

With $\kappa = 0$ *this is the zero-cost monopoly problem for the runtime-weighted law* F_b*, deflated by* $1 - \eta_b = d\log R_0 / d\log v$*; consequently* v_0 *lies strictly below the monopoly cutoff of* F_b*.*

(iv) (Hours against buyers.) *At the optimum*

$$\frac{\varepsilon_{buyers}}{\varepsilon_{hours}} = \frac{\mathbb{E}\big[b(v) \mid v \geq v_0\big]}{b(v_0)} > 1. \tag{20}$$

(v) (Closed form.) *If* F *is uniform on* $[\underline{v}, \overline{v}]$*,* $b(v) = v^\eta$ *with* $\eta \in (0,1)$ *and* $\kappa = 0$*, then*

$$v_0 = \overline{v}\Big(\frac{1-\eta}{2}\Big)^{\frac{1}{1+\eta}}, \quad w^* = R_0(v_0), \quad \Pi_H^* = \frac{w^* \overline{v}^{\,1+\eta}}{2(\overline{v} - \underline{v})}, \tag{21}$$

and $v_0 < \overline{v}/2 \leq \mathbb{E}[v]$*: the marginal buyer lies strictly below the mean valuation.*

(vi) (Stationary version.) *If instead the objective is long-run profit per unit of time with one machine and no queue, then* $\rho(w) = \lambda w M(w) / \big(1 + \lambda M(w)\big)$*, the optimum satisfies*

$$\varepsilon(w^*) = 1 + \lambda M(w^*) = \frac{1}{1 - U^*}, \tag{22}$$

with U *the utilisation, and the price of a machine-hour is endogenous and equal to the optimal profit rate,* $\kappa = \rho^* = w^* U^*$*. For uniform* F *and* $b = v^\eta$*,* $u := 1 - v_0^{1+\eta}$ *solves* $cu^2 + 2u - (1+\eta) = 0$ *with* $c = \lambda(1-\eta)/\big[(1+\eta)(\overline{v} - \underline{v})\big]$*.*

Four remarks, each of which recurs.

Nothing here conditions on a valuation. The seller posts a number; the cutoff is whatever her number implies. Condition (18) makes this operational: it involves the elasticity of billed hours and the price of a machine-hour and nothing else, and the elasticity is measurable from a rate change.

She need not know F or b at all, which matters because b is a function of an unobservable and cannot be estimated pointwise from her own records. The contrast with the hazard, whose design requires b at particular types because exposure is $hb(v)$, is one of the paper's practical conclusions.

Selling hours is not selling units. Part (iv) is the reason. The buyers lost at the margin have the shortest jobs among those served, so each departure costs fewer hours than an average customer supplies, and the seller optimally sits on the elastic part of the buyer-count demand curve. It is the same wedge between marginal and average runtime that appears in ρ_b, and it is why no average of valuations is a sufficient statistic for the rate: in the zero-cost case of (21) with F uniform on $[0.05, 1]$ and $\eta = 1/2$, the marginal buyer is 0.397 against a mean of 0.525, and pricing at $R_0(\mathbb{E}[v])$ costs five percent of profit.

Assumption 1 is doing structural work already. It makes R_0 invertible, hence the buy set an interval and the change of variable in (iii) legitimate. It is also what makes a single rate exclude anyone: as $\eta_b \uparrow 1$, R_0 flattens, willingness to pay per hour becomes uniform across types, and one rate extracts from the whole market. In (21) this appears as v_0 falling and w^* rising in η.

The inequality $v_0 < \mathbb{E}[v]$ is specific to uniform values. In general it holds if and only if $\Phi(\mathbb{E}[v]) > 0$, and it can fail: for $F(v) = v^{\kappa_F}$ on $(0, 1]$ with $b = v^\eta$ and $\kappa = 0$, (19) gives $v_0 = [(1-\eta)/(1+\kappa_F)]^{1/(\eta+\kappa_F)}$ against $\mathbb{E}[v] = \kappa_F/(1+\kappa_F)$, so at $\eta = 1/2$ and $\kappa_F = 0.1$ the marginal buyer is 0.269 and the mean is 0.091. The mean is contaminated by the excluded region, which (19) never touches.

5 Interruption generated by demand

We now let the seller post a second, cheaper rate whose only distinguishing feature is that the machine may be taken back, and we insist that the taking back be driven by demand alone: she reclaims when a buyer arrives who wants the machine on the expensive terms, and never otherwise. Nothing is manufactured or even excess, in the vocabulary of (3)—the hazard is exactly $\alpha(v_1)$—and nothing is conditioned on a valuation.

This section is a benchmark rather than a model of a commercial pool, and the scope restriction belongs here rather than in a footnote. It isolates what interruption is worth when the seller does not choose it, which is what the comparison of Section 6.4 requires; it is not an account of how a large provider generates reclamation, and Appendix E.4 shows why the arrangement below cannot be one. The environment of Section 3 has one machine, and for this section that is not a normalisation. Appendix E carries the arrangement into a pool. The hazard (23) acquires two corrections there, concentration and rationing, and the comparison with Section 6 survives them: a seller who sets the rate may always choose to add nothing, so the ranking of Proposition 4 holds in the pool as well, strictly wherever capacity is slack. What the pool does change is the level of the risk this section describes, and in one limit its existence: with a single free machine enough to admit any arrival, displacement and idle capacity cannot coexist, so a demand-driven hazard requires either a facility small enough for fluctuations to matter or capacity fragmented by placement constraints. The results below are stated for the former; Appendix E.5 treats the latter.

5.1 The rule and what it conditions on

The seller posts $(0, w_H)$ and (h, w_L) with $w_L < w_H$. She observes two things when a request arrives: which tier the incumbent bought, and which tier the arrival buys. The rule she applies is to reclaim the machine from a tenant on the cheap tier when an arriving buyer takes the expensive one, and otherwise do nothing.

This is a rule over posted-price choices, not over types, and it requires no commitment technology: swapping a tenant paying w_L per hour for one paying w_H per hour is profitable term by term, so the seller has no incentive to deviate, and because every guaranteed tenant pays the same rate, replacing one guaranteed tenant with another is revenue-neutral, so honouring the guarantee is costless. Contrast a maximum runtime, which requires terminating a paying customer at a foreseen date against her interest—an asymmetry we return to in Section 6.6.

Lemma 1 (The demand-driven hazard)**.** *Let S be the set of types who take the expensive tier. Then every tenant on the cheap tier faces a reclamation time that is exponential with the single rate $\lambda \Pr(v \in S)$, constant in her elapsed runtime and identical across tenants. If types sort monotonically, so that $S = [v_1, \overline{v}]$, this is*

$$h = \lambda\big(1 - F(v_1)\big) = \alpha(v_1). \tag{23}$$

Three features of (23) organize what follows. It is *anonymous*: the hazard does not depend on the incumbent's private type, so no inference about her is needed to apply it—in contrast with $\alpha(\cdot)$ itself, which does. It is *not an instrument*: the seller chooses w_H, and h follows from how buyers sort. And it is the efficient displacement intensity of *one* type—the marginal guaranteed buyer—applied to everyone, so it is the coarsest anonymous version of $\alpha(\cdot)$. The menu is the seller's measurement instrument, and a two-element menu measures value with one threshold.

One modelling choice must be stated before sorting. Admission is not rationed: an arrival is served if the machine is free, whatever contract she takes, which is the environment of Section 6.4 in which the three arrangements are compared. A single machine does ration, and rationing gives the expensive tier a second advantage—priority of access as well as reliability. Appendix E.8 works that case out, deriving the admission ratio in closed form and showing that it enlarges the expensive tier and raises the value of the second contract several-fold. We keep it out of the body because it is a feature of finite capacity rather than of the screening mechanism, and because including it in one arrangement and not the others would make the comparison of Section 6.4 meaningless.

5.2 Sorting

Write $x = hb(v)$, $U_H(v) = v - w_H b(v)$, $U_L(v) = L(x)v - w_L \tau(h, v)$ and

$$\Delta(v) := U_H(v) - U_L(v). \tag{24}$$

Assumption 3 (Single crossing)**.** On the served range, with $x = hb(v)$,

$$\eta_b(v)\Big[w_H - \big|L'(x)\big|\big(hv + w_L\big)\Big] \;<\; R_0(v)\big[1 - L(x)\big]. \tag{25}$$

Condition (25) is exactly $\Delta' > 0$ in (28), written in primitives and posted prices using $b' = \eta_b/R_0$; it is not a sufficient condition for single crossing but a restatement of it. Two readings are useful. It holds automatically whenever

$$w_H \;\le\; \big|L'(x)\big|\big(hv + w_L\big), \tag{26}$$

since the left side of (25) is then non-positive: if the guaranteed rate is below the marginal value of the exposure a switch would avoid, no type wants to move down however cheap the discount. And discarding the second term in the bracket gives the simpler but strictly stronger requirement $\eta_b w_H/R_0 \le 1 - L$, which we do *not* impose: it fails at the base calibration—at the entry cutoff its slack is -0.249—while (25) holds there with $\Delta' = +0.123$, the difference being exactly the discarded term $+0.373$. The stronger form is a poor guide because the term it drops is largest precisely at the bottom of the band, where the condition binds.

Lemma 2 (Interval structure)**.** *Under Assumptions 1 and 3, and with M_L non-decreasing, Δ is strictly increasing and there are cutoffs $v_2 \le v_1$ such that types in $[v_1, \overline{v}]$ take guaranteed service, types in $[v_2, v_1)$ take the cheap contract, and types below v_2 buy nothing. The cutoffs satisfy*

$$v_1 - w_H b(v_1) = L(x_1)v_1 - w_L\tau(h, v_1), \qquad w_L = R_h(v_2) = R_0(v_2)\,\mathfrak{d}(x_2), \tag{27}$$

with v_2 the lower root, and $\eta_b(v_2)M_L(x_2) < 1$ at the entry cutoff.

Two distinct forces push in the same direction. Differentiating (24),

$$\Delta'(v) = \underbrace{\big(1 - L(x)\big)}_{\text{reliability gap}} + b'(v)\Big[\big|L'(x)\big|\big(hv + w_L\big) - w_H\Big]. \tag{28}$$

The reliability gap rises in v because cumulative exposure $hb(v)$ does: a buyer whose job is longer is more likely to be caught by the same hazard, so protection is worth more to her. Appendix E.8 adds a second force, priority of access, which arises when a single machine rations admission; it reinforces exposure rather than offsetting it, and it accounts for most of what a demand-driven tier is worth where capacity is scarce.

Assumption 3 is a condition on the candidate menu, not on primitives: it involves the posted rates and the hazard they induce. We have not been able to reduce it to a restriction on F, b and L alone, and we do not think it can be, since the condition is exactly whether the posted risk is large enough relative to the price gap, and that is endogenous. What makes this tolerable is that the benchmark does not depend on it. Buyers select by preference whatever the menu, so an optimum exists and Lemma 1 applies to whatever set chooses the expensive tier; what (25) buys

is the closed-form characterisation, not the object being characterised. It is also checkable at any candidate menu, and is implied by the transparent (26).

Where (25) fails the damage is contained but real. Lemma 1 is unaffected: the thinning argument needs only the *measure* of the set choosing the expensive tier, not its shape, so the hazard remains $\lambda \Pr(v \in S)$ and remains anonymous and memoryless. What fails is the interval structure. Δ may cross zero more than once, the expensive tier need not be an upper interval, and the identification $\Pr(v \in S) = 1 - F(v_1)$ that turns the hazard into $\alpha(v_1)$ no longer holds—so the comparison with efficient reallocation in (31) loses its reference point, and the cutoff parametrisation used in Proposition 2 is no longer a change of variables. Section 6 is untouched, since incentive compatibility there rests on Lemma 5 rather than on (25).

Corollary 2 (Composition)**.** *Under Lemma 2: (i) every buyer on the cheap tier has a strictly shorter job and a strictly lower value per machine-hour than every buyer on the expensive tier, and* $\mathbb{E}[b \mid \textit{guaranteed}] > b(v_1) > \mathbb{E}[b \mid \textit{cheap}]$*; (ii) delivered reliability* $L(hb(v))$ *is strictly decreasing in* v*, so it is lowest for the marginal guaranteed buyer.*

Part (i) is not a feature of any calibration: under Assumption 1 low value, short job and low value per machine-hour are the same condition. Part (ii) says the buyer with most at stake is the one the hazard protects least, which is why she is the one who pays to leave.

5.3 The optimum

Parametrizing by the cutoffs through (27), which is a change of variables and not a restriction, the seller's objective is the per-arrival profit (50) of Section 6.4 at the machine-hour price κ:

$$\Pi(v_1, v_2) = \big(w_H - \kappa\big) B(v_1) + \big(w_L - \kappa\big) T(v_2, v_1), \qquad h = \lambda\big(1 - F(v_1)\big), \tag{29}$$

with $T = \int_{v_2}^{v_1} \tau(h, v)\, dF$ and $\kappa = \kappa^*$ from (51), so that the comparison with Sections 4 and 6 is between objects of the same kind. There are *two* instruments, not three: the hazard is tied to v_1 by (23).

Proposition 2 (The demand-driven optimum)**.** *Under Assumptions 1, 2 and 3, an interior optimum satisfies the following.*

(i) (Rates.) *The cheap rate is the marginal entrant's willingness to pay per hour and the expensive rate is the marginal switcher's, net of the rent her outside option leaves her:*

$$w_L^* = R_0(v_2^*)\, \mathfrak{d}\big(h^* b(v_2^*)\big), \qquad w_H^* = R_0(v_1^*) - \frac{U_L(v_1^*)}{b(v_1^*)}. \tag{30}$$

(ii) (Interruption.) $h^* = \alpha(v_1^*)$*; the exposure delivered to the marginal switcher is exactly efficient for her,* $x_1^* = \beta(v_1^*)$*; and for every* $v < v_1^*$

$$\frac{h^*}{\alpha(v)} = \frac{1 - F(v_1^*)}{1 - F(v)} < 1, \qquad h^* - \alpha(v) = -\lambda\big[F(v_1^*) - F(v)\big]. \tag{31}$$

(iii) (Margins.) *With* $S := \partial\Pi/\partial v_1$, $E := \partial\Pi/\partial v_2$ *and* $H := \partial\Pi/\partial h$ *at fixed cutoffs, the optimum solves*

$$E = 0, \qquad S = \lambda f(v_1)\, H, \qquad h = \lambda\big(1 - F(v_1)\big), \tag{32}$$

where $E = R_h'(v_2)\big[\tau_1 B(v_1)/b_1 + T\big] - (w_L - \kappa)\tau(h, v_2) f(v_2)$ *and* $S = \Delta'(v_1)B(v_1)/b_1 + f(v_1)\big[(w_L - \kappa)\tau_1 - (w_H - \kappa)b_1\big]$, *using* $\partial w_H/\partial v_1 = \Delta'(v_1)/b_1$.

(iv) (Self-consistency.) $h^* > 0$ *requires* $1 - F(v_1^*) > 0$*: the discounted tier exists only if a positive mass of buyers remains on the expensive one. Where the hazard map has several fixed points,* h^* *is the greatest of them, selected as in the selection above.*

Condition (32) is where this problem differs from any version in which the hazard is posted. There the seller would set $S = 0$ and $H = 0$ separately. Here the hazard margin does not vanish; it is *priced into* the switch margin at the rate $\lambda f(v_1)$ at which a higher expensive rate destroys the interruption risk, because $\alpha'(v_1) = -\lambda f(v_1) < 0$. At the base calibration both margins are strictly positive—$H = 0.0012$ with $\lambda f(v_1) = 1.0526$, so that $S = \lambda f H = 0.0012$ and (32) holds—so the seller would like a *higher* switch cutoff and *more* interruption, and can have the second only by conceding the first: raising the cutoff is what generates the risk she wants to sell, and it does so by moving buyers onto the expensive tier she would rather they left. The two margins are locked together by (23), which is why neither vanishes separately. Their common sign depends on the environment: in the stationary formulation of E.8, where admission is rationed and an occupied machine displaces an arrival, both are negative instead. What is invariant is the identity, not the sign. The optimal discount is correspondingly shallow, 18.5% at $\lambda = 1$, where a seller free to post the hazard would choose considerably more.

Part (iv) replaces the feasibility condition of (13) with an equilibrium condition, and the replacement is not innocuous. Because v_1 depends on h and h on v_1, the map $h \mapsto \lambda\big(1 - F(v_1(h))\big)$ is increasing—a stronger expectation of interruption pushes buyers up, which generates more interruption—so it can have several fixed points, and $h = 0$ is always one: at $h = 0$ the two contracts are equally reliable and (10) sends every buyer to the cheaper. A positive fixed point therefore requires that some type prefer the expensive contract at the largest hazard the market can generate, $h = \lambda$, which by (24) is

$$w_H \;<\; \Big[1 - L\big(\lambda b(\overline{v})\big)\Big]\left[R_0(\overline{v}) + \frac{w_L}{\lambda b(\overline{v})}\right]. \tag{33}$$

Deep discounts are therefore not merely suboptimal here; past a point they destroy the product. The multiplicity is genuine and needs a stated resolution rather than a condition, and the structure of the map supplies one.

Fix (w_H, w_L) and let $\Phi(h) := \lambda\big(1 - F(v_1(h))\big)$. A higher hazard lowers U_L pointwise and leaves U_H unchanged, so Δ rises, its zero $v_1(h)$ falls, and Φ is non-decreasing on $[0, \lambda]$ into itself. By Tarski's theorem its fixed-point set is a non-empty complete lattice with a least and a greatest element; the least is $h = 0$, since at $h = 0$ the contracts are equally reliable and (10) sends every

buyer to the cheaper. *We select the greatest*, $h^* := \max\{h : \Phi(h) = h\}$. Monotonicity makes the iteration $h_{n+1} = \Phi(h_n)$ from $h_0 = \lambda$ non-increasing and bounded below by any fixed point, so it converges to h^*; and since $\Phi(h) - h$ changes sign from positive to negative there, $\Phi'(h^*) \leq 1$ and the selection is locally stable.

Two things justify this selection rather than another. It is the equilibrium a seller would coordinate on if she could, since she posts prices before buyers choose and higher participation on the expensive tier is what sustains the product she is selling; and it is the one reached when buyers begin from the most pessimistic expectation about reliability, which is the natural starting point for a tier whose risk is not contractible. At the optimum of Table 1 the map has three fixed points—0, 0.332 and 0.514—and the iteration from $h_0 = \lambda$ converges to 0.514 with $\Phi' = 0.66$, so the selected equilibrium is stable while the interior one at 0.332 is not. Condition (33) is what rules out the degenerate selection $h = 0$; given that, the lattice structure does the rest. Condition (33) is necessary; it is sufficient when the map crosses the diagonal, which it does whenever it exceeds the identity at some h below λ.

Four comparisons with the guaranteed benchmark follow, and none needs a separate statement. Comparing with the single-rate optimum w^*, v_0, Π_H of Proposition 1: (i) $\Pi > \Pi_H$ strictly, provided $\lambda > 0$ and (33) holds; (ii) the expensive tier alone always earns less than the benchmark, $(w_H - \kappa)B(v_1) \leq \Pi_H(w_H) \leq \Pi_H$, so the entire gain is the discounted tier's revenue net of that loss; (iii) $w_L^* < w^*$ always; and (iv) $w_H^* > w^*$ if and only if $R_0(v_1) - R_0(v_0) > r(v_1)/b(v_1)$ with $r(v_1) = U_L(v_1)$, and whenever $w_H^* \geq w^*$ the expensive segment strictly shrinks, $v_1 > v_0$.

Part (ii) is the honest way to describe what a second tier does: it is not an addition to the guaranteed business but a partial substitution away from it. Part (iv) makes "the guaranteed rate rises" conditional rather than automatic: the rate rises only if the shrinkage of the segment outweighs the rent the switch type keeps because she now has somewhere to go. Both sides are close at the base calibration—0.0245 against 0.0227—so the condition holds but not comfortably, and the guaranteed rate rises only from 0.6904 to 0.6923, a third of a percent, where a posted hazard would raise it considerably more.

5.4 What the interruption risk looks like

Equation (31) is this section's sharpest statement, and it contains no primitive other than F. Every buyer strictly inside the cheap tier is *under-exposed* relative to efficiency, by exactly a ratio of survival functions, and the shortfall in rates is λ times the mass of types between the incumbent and the switch cutoff. Equality holds only at the top of the band.

The direction deserves emphasis, because the usual presumption runs the other way. In a single facility a seller who cannot see valuations and reclaims only on demand reclaims *less* than reallocation requires, and necessarily so: an arrival whose value v' lies in (v, v_1) ought to displace an incumbent of type v and does not, because she chooses the cheap tier herself and is turned away. The wedge measures the menu's resolution, not the seller's motives. In a pool it acquires a factor P_B/s by (74) and can exceed one with nothing manufactured in the sense of (3), so the sign

λ	κ^*	w^*	v_0	w_H^*	w_L^*	h^*	v_1
0.5	0.1323	0.6648	0.4420	0.6664	0.5988	0.2641	0.4982
1	0.2214	0.6904	0.4767	0.6923	0.5644	0.5146	0.5112
2	0.3370	0.7265	0.5277	0.7289	0.4976	0.9837	0.5328
3	0.4115	0.7515	0.5648	0.7515	0.4288	1.3724	0.5654

Table 1: The demand-driven optimum in the environment of Section 6.4: per arrival, with the machine-hour price κ^* endogenous by (51) and no admission priority. F uniform on $[0.05, 1]$, $b(v) = v^{1/2}$, duration deterministic; $h^* = \lambda(1 - F(v_1))$ throughout. The gain over the single rate is small—at most a few tenths of a percent at these arrival rates—which is what (53) leads one to expect: with no reversal the discounted tier extends the market only slightly below $v_0(w_H)$, and by (53) the whole gain is bounded by the margin on that band. Appendix E.8 shows that admission priority, which a single machine supplies and this environment does not, raises the figure by an order of magnitude.

holds only where blocking does not exceed the spot share; Appendix E gives the condition. The wedge closes as the menu is refined: Equations (55) and (56) construct the n-tier menu explicitly, give the rates that support any set of cutoffs, and show the delivered hazard is a decreasing step function lying below $\alpha(\cdot)$ and exact at the upper end of each band, so refinement converges to efficient reallocation from below. Sorting remains monotone under refinement provided (25) holds at each boundary, so nothing prevents this. What stops it is that the gain is exhausted quickly: successive grades are worth 6.0, 1.8, 0.7 and 0.2 percent of profit at $\lambda = 1$, and a sixth adds nothing measurable.

Whether the tier operates by re-pricing or by expansion turns on a threshold that is *not* $\eta_b K > 1$. Sorting needs only $\Delta' > 0$, which is Assumption 3 and constrains the hazard only through the exposure it delivers, never through its relation to λ—a point Section 6 exploits, since it means a seller who chooses the hazard may screen with a rate below the reallocative one as readily as above it; the reversal condition allows the cheap tier to reach types the single rate excluded without tempting anyone above them. Because exposure is bounded here by $\lambda b(v)$, a reversal requires scarcity as well as a strong duration–value link, and at the base calibration $\eta_b M_L(x_1) = 0.53$—no reversal anywhere—yet the tier exists and pays, cannibalising the band $[v_0, v_1]$ of width 0.035 against a newly served band $[v_2, v_0]$ of width 0.021. As λ rises the cannibalised band narrows and the newly served one widens, so the tier shifts from re-pricing the installed base toward serving buyers the single rate excludes. So $\eta_b K > 1$ is a condition on the tier's *mode*, not on its existence.

Table 1 records three patterns. The gain from the second tier is single-peaked in scarcity, maximal near $\lambda \approx 3$: when the machine is nearly always idle there is little to reallocate and the hazard is too weak to differentiate, and when it is nearly always busy the expensive tier already captures the surplus. The composition of the market moves little—the expensive tier holds 53% of arrivals at $\lambda = 0.5$ and 46% at $\lambda = 3$—while delivered reliability collapses, so the seller absorbs scarcity through reliability rather than through who is in which tier. And the two cutoffs move very differently: v_0 rises from 0.442 to 0.565 while v_1 rises only from 0.498 to 0.565, so the two

converge—the gap between them falls from 0.056 to 0.001 across the range—and the cannibalised band $[v_0, v_1]$ closes almost entirely. The discounted tier therefore does not merely shrink as capacity tightens; it stops re-pricing the installed base and survives only on the buyers a single rate would exclude, which is the configuration in which (53) binds most tightly. That contrast is testable, since the single-rate cutoff responds only to capacity pressure and is exactly invariant to the duration shock, whereas the tier boundary responds to the shock and barely to capacity. We can prove the two limits—as $\lambda \to 0$ the reliability gap vanishes and the menu collapses, and as $\lambda \to \infty$ the cheap contract delivers and bills nothing so $v_1 \to R_0^{-1}(w_H)$ while $v_0 \to \overline{v}$—so v_1 is non-monotone with an interior minimum, which is why it looks flat in between. The comparative static itself we do not prove.

6 A schedule of hazard contracts

In Section 5 the seller chose two prices and the interruption rate followed. We now let her choose it too, and not as a single number: she posts a schedule of contracts, each pairing a rate with a hazard, and each buyer takes the one she prefers. Guaranteed service is not a separate instrument here; it is the $h = 0$ element of the same schedule.

6.1 The problem

A schedule is a family $\{(h(v), w(v))\}$ which, by the taxation principle, we may index by the type that selects it. Write $x(v) := h(v)b(v)$ for exposure and $r(v)$ for rent. Since $r = L(x)v - w\tau$ with $\tau = b(v)(1 - L(x))/x$, payments are residual,

$$w(v) = \frac{\big[L(x)v - r(v)\big]\, x}{b(v)\big[1 - L(x)\big]}, \tag{34}$$

so the controls are the exposure schedule $x(\cdot)$, the served set, and one boundary value for the rent.

Lemma 3 (Envelope)**.** *Let* $\widetilde{\mathfrak{d}}(x) := -xL'(x)/\big(1 - L(x)\big) = L(x)M_L(x)$*, with* $\widetilde{\mathfrak{d}}(0) = 1$*. Along any incentive-compatible schedule,*

$$r'(v) = L\big(x(v)\big) - \eta_b(v)\,\widetilde{\mathfrak{d}}\big(x(v)\big)\Big(1 - \frac{r(v)}{v}\Big) = \mathcal{A}(v) + \mathcal{B}(v)\, r(v), \tag{35}$$

where $\mathcal{A} = L(x)\big[1 - \eta_b M_L(x)\big]$ *and* $\mathcal{B} = \eta_b M_L(x)L(x)/v > 0$*; the calligraphic letters distinguish these coefficients from* $B(t) = \int_t^{\overline{v}} b\, dF$ *of* (15)*.*

Equation (35) is the envelope condition of Milgrom and Segal (2002) applied to this family, differentiated once more to eliminate the payment. Three features matter. The equation is *linear* in r, so the schedule is pinned by $x(\cdot)$ and one boundary condition. The coefficient $\mathcal{B}$ is strictly positive, because a longer job pays for more billed time, so rent perturbations are *amplified* up the type space rather than damped. And the sign of the drift at zero rent is the sign of $1 - \eta_b M_L(x)$,

which is exactly the reversal test of (11): a type can be held at zero rent only where her exposure is at or below the frontier that makes screening possible. The shock enters only through L and M_L. At $x \equiv 0$, (35) reads $r' = (1-\eta_b) + \eta_b r/v$, whose solution is $r = v - w_H b(v)$: the guaranteed contract is the $x = 0$ solution of the same equation, and by (10) any interval on which $x = 0$ is served by a single rate.

The seller solves

$$\max_{x(\cdot),\, v_2} \int_{v_2}^{\overline{v}} \Big[L\big(x(v)\big)v - r(v) - c\big(v, x(v)\big) \Big] f(v)\, dv \quad \text{s.t. (35)},\ r(v_2) = 0,\ r \geq 0,\ x \geq 0, \tag{36}$$

where $c(v,x)$ is the occupancy cost per job. We solve (36) for a general c that is continuous and non-increasing in x, and two cases are used. In the environment of Section 6.4 a machine-hour is priced at the constant κ^* of (51), so

$$c(v,x) = \kappa\, \tau(h,v) = \kappa\, b(v)\, \psi(x), \qquad \psi(x) := \frac{1 - L(x)}{x}, \tag{37}$$

which is smooth in x. If instead reclamations are matched to arrivals as in (1), so that the seller is credited $\mu \min\{h, \lambda\}$ per occupied hour, the cost becomes

$$c(v,x) = \mu\big(\lambda b(v) - x\big)^{+} \psi(x), \tag{38}$$

which kinks at $x = \lambda b(v)$. The regional structure below is the same in both cases except that region (iii), the bunch at the arrival rate, is generated by the kink and is absent under (37).

Lemma 4 (Amplified survival)**.** *With $E(v) := \exp \int_{v_2}^{v} \mathcal{B}$ and $W(v) := E(v)^{-1} \int_v^{\overline{v}} E(s) f(s)\, ds$, and $\rho_W := W/f$, the costate of (36) is $p = -W$, and $W \geq 1 - F$ with equality only if $\mathcal{B} \equiv 0$.*

Two monotonicity facts are used below and are worth separating. From $p = -W$ and $p' = f - p\mathcal{B}$ we get

$$W'(v) = -f(v) - \mathcal{B}(v) W(v) \ < \ 0, \tag{39}$$

so W is strictly decreasing without further assumptions. That $\rho_W = W/f$ is decreasing is a different statement: $(W/f)' < 0$ is $f^2 + \mathcal{B} W f + W f' > 0$, which holds whenever f is non-decreasing, and in particular for the uniform family used throughout. We maintain that condition wherever "ρ_W decreasing" is invoked.

So the object that discounts rent is not $1 - F$ but W, and $\rho_W \geq (1-F)/f$ replaces the inverse hazard rate that plays this role in the standard screening problem since Myerson (1981): exposure granted low in the distribution raises the rent of every type above, compounding as it goes. The inequality is the substantive difference. In a setting where the allocation does not feed back into the rent of higher types the two coincide; here it does, because a contract's exposure changes what every longer job would have to be paid to stay put, and the discount factor is correspondingly larger. Two checks fix the interpretation. Where $x \equiv 0$ we have $\mathcal{B} = b'/b$, hence $E \propto b$ and $W = B(v)/b(v)$ with B the integral of (15), so the transversality condition of (36) collapses term by term to (19).

And in the base calibration ρ_W/ρ runs from 1.67 at the bottom of the served set to 1.06 near the top, so the correction is first order.

6.2 The optimal exposure schedule

Maximising the Hamiltonian in x and using $p = -W$ gives the condition that governs exposure. Write

$$\Theta_L(x) := \frac{\widetilde{\mathfrak{d}}'(x)}{L'(x)}, \qquad \Theta_L(0) = \tfrac{1}{2}\mathbb{E}[\varepsilon^2], \qquad \tilde{v} := v - \frac{c_x(v,x)}{L'(x)}. \tag{40}$$

Substituting $\mathcal{A}_x = L' - \eta_b\widetilde{\mathfrak{d}}'$ and $\mathcal{B}_x = \eta_b\widetilde{\mathfrak{d}}'/v$ into $\mathcal{H}_x = \big(L'v - c_x\big)f - W\big[\mathcal{A}_x + \mathcal{B}_x r\big] = 0$ and dividing by $L'f < 0$ gives, at any interior point of the served set with $r > 0$ and $x > 0$,

$$\eta_b(v)\,\Theta_L\big(x(v)\big)\Big(1 - \frac{r(v)}{v}\Big) = 1 - \frac{\tilde{v}}{\rho_W(v)}, \tag{41}$$

while $x(v) = 0$ whenever $1 - \tilde{v}/\rho_W \le \tfrac{1}{2}\eta_b\mathbb{E}[\varepsilon^2]\big(1 - r/v\big)$, the corner condition $\mathcal{H}_x \le 0$ at the origin. That inequality locates the guaranteed boundary exactly: at the base calibration it first binds at $v = 0.54772$ against a computed $v_1 = 0.54772$, and classifies every node on either side correctly. Under deterministic duration $\Theta_L = \big(\mathfrak{M}(x) - 1\big)/\big(1 - e^{-x}\big)$, an identity that holds for no other mean-one shock, so Θ_L rather than $M_L - 1$ is the object that generalises; the derivation is in Appendix A.

The left side of (41) is the rent conceded by exposing type v a little less, scaled by how far she is from full extraction; the right side is the value destroyed net of the occupancy cost released, measured against the amplified inverse hazard rate. By Lemma 5(iv), *exposure falls with value*, and so does $h(v) = x(v)/b(v)$, more steeply. A single posted hazard—whether chosen or generated by demand as in (23)—makes exposure monotone *increasing*, since $x = hb(v)$ and b rises. A schedule reverses that. Two comparative statics follow from $\Theta_L(0) = \tfrac{1}{2}\mathbb{E}[\varepsilon^2]$: a more variable duration shock widens the region on which no exposure is sold, and the deterministic case is not a safe guide to the shape of the schedule under uncertainty.

Assumption 4 (Decreasing differences)**.** The Hamiltonian of (36) has decreasing differences in (x, v) on the served set: $\mathcal{H}_{xv} \le 0$, with strict inequality where $x > 0$.

Assumption 4 is what monotone comparative statics actually require, and it is weaker than the conditions usually imposed for this purpose. By Topkis's theorem it delivers the conclusion directly.

Three things this buys. It does not require (41) to have a unique root, so it survives a non-monotone or even negative Θ_L provided the cross-partial behaves; it is a condition on the objective rather than on the duration distribution alone, which is where monotone comparative statics belong; and it is what makes the relaxation in (36) tight, since by Lemma 5 below a non-increasing h is exactly what incentive compatibility requires, so no ironing is needed. Where Assumption 4 fails, the relaxed solution may be non-monotone and must be ironed; we do not pursue that case.

Under the smooth occupancy cost (37) of the unified environment the assumption holds throughout the region we compute: evaluating $\mathcal{H}_{xv}$ on a grid of (v, x, r, W) spanning the served set gives

strictly negative values at every point. Under the kinked cost (38) it holds away from $x = \lambda b(v)$, where c_x jumps and $\mathcal{H}_{xv}$ is undefined; the bunch region (iii) is exactly the interval on which that jump binds.

The sufficient condition in Lemma 5(iv) is a restriction on the duration shock, and it is not implied by mean-one alone: Θ_L need be neither positive nor increasing for a general shock. For an even mixture of two gammas with shape 5 and means 0.1 and 1.9—mean one—Θ_L rises to 1.54 at $x = 0.89$, falls thereafter, and turns negative beyond $x = 4$. It holds for deterministic duration and, numerically, for gamma shocks with shape $k \geq 2$; it fails for gamma shocks with $k \leq 1$, where Θ_L is positive but not monotone. Since $\Theta_L(0) = \frac{1}{2}\mathbb{E}[\varepsilon^2]$ in every case, what it controls is the tail of realized-to-planned duration rather than its variance. The deterministic case is provable.

Both are provable in the case the paper computes. Under deterministic duration write $u := 1 - e^{-x}$ and $g(x) := x - 1 + e^{-x}$, so that $\Theta_L = g/u^2$; since $g(0) = 0$ and $g' = 1 - e^{-x} > 0$, we have $g > 0$ and $\Theta_L > 0$. Differentiating with $u' = 1 - u$,

$$\Theta_L' = \frac{u^2 - 2(1-u)(x-u)}{u^3},$$

whose numerator is $\varphi(x) := 1 - e^{-2x} - 2xe^{-x}$; and $\varphi(0) = 0$ with $\varphi'(x) = 2e^{-x}\left[e^{-x} + x - 1\right] = 2e^{-x}g(x) > 0$, so $\varphi > 0$ and $\Theta_L' > 0$ on $(0, \infty)$. For the second condition, (39) makes $(W/f)'$ have the sign of $-f^2 - \mathcal{B}Wf - Wf'$, strictly negative whenever $f' \geq 0$ because $f > 0$, $\mathcal{B} \geq 0$ and $W \geq 0$; for uniform F, $f' = 0$. So both hold throughout the environment we compute.

Beyond that case the assumption binds, and it is worth setting out exactly which results need it, since the model carries a general mean-one shock and most of what follows does not restrict it.

Unaffected by Assumption 4, and stated for an arbitrary mean-one ε: the guaranteed benchmark and the neutrality of the shock, Proposition 1; the reallocative schedule, Section 3; the demand-driven arrangement and its optimum, Lemmas 1 and 2 and Proposition 2 and the comparisons following it; the slope identity (11), the exposure ceiling (12) and the frontier (13), which is where the shock does its most visible work, since $K = \sup M_L$ is finite for a gamma shock and infinite under deterministic duration, so that a non-degenerate shock is what makes screening possibly *infeasible*; the envelope and costate, Lemma 3; the first-order condition (41) itself and the frontier region of Proposition 3(i); the monotone-selection criterion of Lemma 5, whose statistic is written so as to cover a general shock; the ranking and the transmission bound, Proposition 4 and (53); and Appendix D, which is a statement about $\Phi = \mathcal{Q}/\mathcal{B}$ for arbitrary L.

Requiring Assumption 4: the comparative statics of Lemma 5(iv)—that exposure falls with value—and everything that rests on the ordering of the regions in Proposition 3. Requiring deterministic duration specifically: (47), and with it the derivation of the monotone-selection condition of Lemma 5 from primitives rather than by assumption. For a general shock the mechanism-design results of Section 6 therefore hold under Assumption 4 together with the monotone-selection condition of Lemma 5, both of which are conditions on the shock and the posted prices rather than on the seller's problem, and both are automatic at $\varepsilon \equiv 1$ by the argument above. Numerically they

are also automatic for gamma shocks with shape $k \geq 2$, which we have checked but not proved. The results that invoke Assumption 4 are Lemma 5(iv) and everything that rests on exposure being monotone—the ordering of the regions in Proposition 3 and the tightness of the relaxation—while the first-order condition (41) itself does not need it.

Both Assumption 4 and the condition on f hold in every calibration reported here: duration is deterministic and F is uniform. They do the same job. Assumption 4 makes the left side of (41) increasing in x; f non-decreasing makes the right side decreasing in v; together they force the x that solves it to fall with v. It is worth being explicit about what depends on that and what does not. Unaffected are the ranking of Proposition 4, which is revealed preference over nested classes and uses no monotonicity; the bound (53), which uses only participation and $\tau \leq b$; and the frontier region of Proposition 3(i), which follows from $\mathcal{A} = 0$. What fails if either condition fails is the *shape*: (41) may then have several roots in x at a given v, exposure need not be monotone, the regional decomposition loses its ordering, and—most consequentially—h need not be non-increasing, so by Lemma 5 the envelope-only relaxation need not be tight and ironing may be required. In that case the pointwise Hamiltonian maximisation must be carried out globally rather than through the first-order condition, which is what our solver does in any event. Where Θ_L turns negative, as it does for the bimodal mixture above, $\mathcal{H}$ is not even locally concave in x and the first-order condition can identify a minimum.

6.3 Structure and free boundaries

Proposition 3 (Five regions)**.** *Impose Assumption 1, and let $\bar{x}(v)$ solve $M_L(\bar{x}) = 1/\eta_b(v)$, which exists exactly when $\eta_b K > 1$, and $\bar{q} := L(\bar{x})$. Then the schedule consists of at most five regions, as follows. Two further properties hold under an additional hypothesis rather than as part of the construction: exposure is decreasing on the interior regions by Lemma 5(iv); and the posted hazard $h(\cdot)$ is non-increasing under the condition of Lemma 5, which (47) establishes along the optimum: unconditionally on the frontier and guaranteed regions, and elsewhere exactly when the marginal completion value destroyed by exposure exceeds the occupancy cost it releases. Delivered exposure then rises on region (iii), where h is flat and b increasing, so exposure itself is not monotone across regions.*

(i) Full extraction at the frontier. *On $[v_2, v^*]$ the constraint $r \geq 0$ binds, $r \equiv 0$ and $x \equiv \bar{x}(v)$, so delivered reliability is $\bar{q} = \Psi_L^{-1}(\eta_b)$ of (13): the bottom of the market is served at exactly the supremum reliability consistent with screening and pays its entire completion value. This requires $h(v) = \bar{x}/b(v)$, falling in v, so no constant hazard can reproduce it.*

(ii) Manufactured exposure. *$x > \lambda b(v)$, with (41) and $\tilde{v} = v$.*

(iii) A bunch at the arrival rate. *On an interval $x(v) = \lambda b(v)$ exactly, so $h(v) = \lambda$: a positive mass of types is reclaimed at precisely the rate at which requests arrive. It exists because (38) kinks at $x = \lambda b(v)$, so the Hamiltonian is not concave in x and has two local maxima.*

(iv) Sub-arrival exposure. $x < \lambda b(v)$, *so the hazard is below the arrival rate and by* (3) *not manufactured, with* (41) *and* $\tilde{v} < v$.

(v) Guaranteed. *On* $[v_1, \overline{v}]$, $x = 0$ *and a single rate*

$$w_H = \frac{v_1 - r(v_1)}{b(v_1)} = R_0(v_1) - \frac{r(v_1)}{b(v_1)} \tag{42}$$

is posted, by (10).

Both free boundaries are local*:*

$$v_2 = \max\Big\{\underline{v},\ \inf\big\{v : \bar{q}\, v \geq \mu\big(\lambda b(v) - \bar{x}\big)^+ \psi(\bar{x})\big\}\Big\}, \tag{43}$$

the type at which the frontier contract just breaks even. The upper end v^* *of the frontier region is characterised in either of two ways, according to whether exposure is continuous there. Write* $x^+ := \lim_{v \downarrow v^*} x(v)$. *If* $x^+ = \bar{x}(v^*)$ *then*

$$\frac{\tilde{v}^*}{\rho_W(v^*)} = 1 - \eta_b(v^*)\, \Theta_L\big(\bar{x}(v^*)\big), \qquad \tilde{v}^* := v^* - \frac{c_x\big(v^*, \bar{x}(v^*)\big)}{L'\big(\bar{x}(v^*)\big)}, \tag{44}$$

and if $x^+ \neq \bar{x}(v^*)$ *then* v^* *is fixed by value matching,*

$$\mathcal{H}\big(v^*, \bar{x}(v^*), 0, W(v^*)\big) = \mathcal{H}\big(v^*, x^+, 0, W(v^*)\big). \tag{45}$$

Under deterministic duration the right-hand side of (44) *is* $1 - (1 - \eta_b)/(1 - \bar{q})$.

Which of the two characterisations of v^* applies is a question about structure, and both occur. The cost term in $\tilde{v}^*$ must be carried: it vanishes only where $\bar{x} \geq \lambda b(v^*)$, so that the frontier already reclaims at least as fast as requests arrive, and dropping it where it does not vanish misplaces the boundary badly—at $\lambda = 3$ the correct form gives 0.512 against a computed edge of 0.511, while the cost-free version gives 0.315. As λ rises the interior region narrows and then vanishes: at $\eta_b = 1/2$ and $\mu = 0.1$ the schedule has a frontier, an interior and a guaranteed region for $\lambda \leq 4$, with the jump in x at v^* no larger than the grid, so (44) applies; for $\lambda \geq 5$ it has only a frontier and a guaranteed region, exposure falling discontinuously from $\bar{x}$ to zero—at $\lambda = 6$, from 1.594 to 0—and (45) applies, locating v^* at 0.619, 0.650 and 0.718 for $\lambda = 5$, 6 and 8, each within one node spacing of the computed edge.

Equation (43) replaces the integral transversality condition one would expect, and the reason is (35): on region (i) the drift $\mathcal{A}$ vanishes, so the channel through which the cutoff would affect every rent above it disappears from the Hamiltonian and the condition becomes pointwise. There is no two-point problem to solve. Note also (42): once a full schedule is available the guaranteed rate ceases to be an independent instrument. It is the residual of the screening problem below it, and the seller's only lever on it is how much rent the schedule concedes to the type indifferent about leaving.

What incentive compatibility restricts is the posted hazard, and the restriction is exactly a monotonicity. Normalising by $b(v)$, a buyer's payoff is $U/b(v) = R_0(v)L(x) - w\psi(x)$, so the type enters through the two scalars $R_0(v)$ and $b(v)$, both increasing under Assumption 1. That is why single crossing is a condition rather than an implication, and the following gives it in primitives and posted prices.

Lemma 5 (Monotone selection and monotone exposure)**.** *Write* $\mathfrak{c}(x) := -xL''(x)/L'(x) > 0$, $\sigma := w/(hv)$, *and*

$$\Lambda(v) := \eta_b(v)\big(1+\sigma\big)\Big[\mathfrak{c}(x) + \widetilde{\mathfrak{d}}(x) - 1\Big]. \tag{46}$$

Along a differentiable schedule: (i) if $\Lambda < 1$ *then every incentive-compatible schedule has* $h(\cdot)$ *non-increasing; (ii) conversely, a schedule with* $h(\cdot)$ *non-increasing that satisfies* (35) *is incentive compatible; (iii) if* $\Lambda > 1$ *both conclusions reverse. Under deterministic duration* $\mathfrak{c} + \widetilde{\mathfrak{d}} = \mathfrak{M}(x)$, *so the condition reads* $\eta_b(1+\sigma)\big(\mathfrak{M}(x) - 1\big) < 1$. *(iv) Under Assumption 4 the pointwise maximiser* $x(v)$ *of* $\mathcal{H}$ *is non-increasing, and strictly decreasing where interior, so* $h(v) = x(v)/b(v)$ *is strictly decreasing wherever* $x > 0$*; a sufficient condition in primitives is that* Θ_L *be positive and increasing and* f *non-decreasing, so that the left side of* (41) *rises in* x *while its right side falls in* v. *Parts (ii) and (iv) together make the envelope-only relaxation tight, so no ironing is required.*

The condition need not be checked numerically: along the seller's own optimum it reduces to a primitive inequality.

Along the seller's own optimum, Λ reduces to an object already in hand. By (34), $w = \big[L(x)v - r\big]x/\big[b(1-L)\big]$, so with $x = hb$

$$\sigma = \frac{L(x) - r/v}{1 - L(x)}, \qquad \text{hence} \qquad \Lambda = \eta_b(v)\,\Theta_L\big(x\big)\Big(1 - \frac{r}{v}\Big) \tag{47}$$

under deterministic duration, using $1 + \sigma = (1 - r/v)/(1 - L)$ and $\mathfrak{M} - 1 = (x - 1 + L)/(1 - L)$. Three consequences follow. On the guaranteed region $x = 0$ and $\mathfrak{M}(0) = 1$, so $\Lambda = 0$. On the frontier region $r \equiv 0$ and $\eta_b M_L(\bar{x}) = 1$, so $\Lambda = (1 - \eta_b)/(1 - \bar{q})$, which is below one iff $\bar{q} < \eta_b$; and since $M_L = \mathfrak{M}$ gives $\eta_b = (1 - \bar{q})/(-\log \bar{q})$, that is $1 - q + q \log q > 0$ on $(0, 1)$, which holds because the left side vanishes at $q = 1$ and has derivative $\log q < 0$. So $\Lambda < 1$ holds *unconditionally* there. Elsewhere the right side of (41) equals Λ, so $\Lambda = 1 - \tilde{v}/\rho_W$ where the condition holds with equality and $\Lambda < 1 - \tilde{v}/\rho_W$ on the bunch, where $\mathcal{H}_x > 0$ just below the kink; in either case

$$\Lambda < 1 \iff \tilde{v} > 0 \iff L'(x)\, v < c_x(v, x),$$

the marginal completion value destroyed by exposure exceeding the occupancy cost it releases. In particular $\Lambda < 1$ throughout whenever $\kappa = 0$, since then $c_x \equiv 0$ and $\tilde{v} = v > 0$.

So $\Lambda < 1$ is not an auxiliary restriction but a property of the optimum wherever exposure is worth supplying at all, and it fails only where the seller would be raising exposure to save occupancy cost faster than she destroys value—which is not an optimum. At the base calibration Λ runs from 0.240 to 0.6275, attaining its maximum $(1 - \eta_b)/(1 - \bar{q})$ on the frontier. The condition

is also compatible with screening: combining $\Lambda < 1$ with the reversal requirement $\mathfrak{M}(x) > 1/\eta_b$ of (11) leaves a non-empty window precisely when $\sigma < \eta_b/(1-\eta_b)$. So h is non-increasing, w is non-decreasing, the envelope-only relaxation is tight, and no ironing is required.

It is worth being explicit about what this does *not* obstruct. The efficient hazard $\alpha(v) = \lambda(1-F(v))$ is decreasing, so it satisfies the requirement of Lemma 5 and is not excluded by it. Together with Section 5.4, where an ordered n-tier menu delivers $\lambda(1-F(v_{j-1}))$ to tier j with monotone sorting at every resolution, the conclusion is that efficient reallocation *is* attainable under private values. The seller does not deliver it because it does not pay: refinement toward $\alpha(\cdot)$ is feasible and incentive compatible, and its value to her falls by roughly a factor of three with each grade added, to under a quarter of a percent by the fifth. The gap between the schedule she posts and the efficient one is a rent-extraction gap, not an informational one, and the surplus refinement would create accrues to buyers rather than to her. That is the paper's organising statement, and it is a stronger claim than the impossibility one it replaces.

6.4 One environment for the three arrangements

The three arrangements must be compared in a single environment or the comparison means nothing. We use the following, which is the one Proposition 1(vi) already implies for the single rate.

A *mechanism* M posts a menu of contracts; an arriving buyer takes her preferred element or nothing. Write $\tau_M(v)$ for the machine-time a type-v buyer occupies under M and $w_M(v)$ for the rate she pays, and

$$N(M) := \int w_M(v)\,\tau_M(v)\,dF(v), \qquad T(M) := \int \tau_M(v)\,dF(v), \tag{48}$$

for revenue and occupancy per arrival. Arrivals are Poisson at rate λ and the machine serves one job at a time, so by renewal–reward the long-run profit rate is

$$\rho(M) = \frac{N(M)}{1/\lambda + T(M)}. \tag{49}$$

There is no admission priority: an arrival is served if the machine is free and not otherwise, for every contract. Appendix E takes up what changes when guaranteed arrivals may displace.

Maximising a ratio is awkward, and the standard device removes it. For a scalar κ write

$$\Pi(M;\kappa) := N(M) - \kappa\,T(M) = \int \big[w_M(v) - \kappa\big]\tau_M(v)\,dF(v), \tag{50}$$

the per-arrival profit when a machine-hour is priced at κ, and $\Pi^*_{\mathcal{M}}(\kappa) := \sup_{M\in\mathcal{M}} \Pi(M;\kappa)$ for a class $\mathcal{M}$ of mechanisms.

Maximising a ratio is awkward and the standard device removes it. For each fixed M the map $\kappa \mapsto \Pi(M;\kappa)$ is affine with slope $-T(M) \le 0$, so $\Pi^*_{\mathcal{M}}$, a pointwise supremum of affine functions, is convex, non-increasing and continuous; it is strictly positive at $\kappa = 0$ and vanishes by $\kappa = R_0(\overline{v})$,

since $\tau_M \le b$ and $w_M \le R_0$. Hence $\kappa \mapsto \Pi^*_{\mathcal{M}}(\kappa) - \kappa/\lambda$ is continuous and strictly decreasing from a positive value, and

$$\Pi^*_{\mathcal{M}}(\kappa) = \kappa/\lambda \tag{51}$$

has a unique root $\kappa^*_{\mathcal{M}}$. That root is the optimum itself: $\rho(M) \ge \kappa$ iff $N(M) \ge \kappa/\lambda + \kappa T(M)$ iff $\Pi(M;\kappa) \ge \kappa/\lambda$, so taking suprema, $\sup_{M\in\mathcal{M}} \rho(M) \ge \kappa$ exactly when $\Pi^*_{\mathcal{M}}(\kappa) \ge \kappa/\lambda$, which by monotonicity holds precisely for $\kappa \le \kappa^*_{\mathcal{M}}$. The shadow price of a machine-hour is therefore the optimal profit rate, and any maximiser of $\Pi(\cdot;\kappa^*_{\mathcal{M}})$ attains it.

Equation (51) does two things. It reduces the ratio problem (49) to the per-arrival problem (50) that Sections 4–6 solve, with the single scalar κ carrying the whole of the machine's alternative use; and it identifies that scalar as the profit rate, so it is endogenous rather than a free parameter. Setting $\kappa = \mu\lambda$ instead, as a reduced form, is a different model: it prices a machine-hour by an outside margin rather than by what the machine earns, and it degenerates as $\lambda \to \mu^{-1}$, where κ exceeds every buyer's willingness to pay, whereas $\kappa^* \le R_0(\overline{v})$ always.

Now let

$$\mathcal{M}_H \;\subset\; \mathcal{M}_2 \;\subset\; \mathcal{M}_3$$

be, respectively, the single-rate mechanisms of Section 4; the two-contract mechanisms of Section 5, in which the hazard is not posted but equals λ times the share of arrivals taking the expensive contract; and all incentive-compatible menus, which is the class of Section 6. The inclusions hold because a single rate is the two-contract menu with $w_L = w_H$, at which the cheap contract attracts nobody and the induced hazard is zero, and because a two-contract menu is a menu.

Proposition 4 (The ranking, in one environment)**.** *Let ρ^*_H, ρ^*_2 and ρ^*_3 be the optimal profit rates over $\mathcal{M}_H$, $\mathcal{M}_2$ and $\mathcal{M}_3$ in the environment above. Then*

$$\rho^*_H \;\le\; \rho^*_2 \;\le\; \rho^*_3,$$

with the second inequality strict whenever the solution of (41) at $\kappa = \kappa^*_2$ is non-constant, and the first strict whenever some type below v_0 can be served at a positive margin.

The ranking says the two increments are non-negative. The next result says considerably more about the first of them: it is bounded by a single margin, and it is zero unless the discounted tier reaches buyers the guaranteed rate excludes.

The ranking says the two increments are non-negative. Considerably more can be said about the first of them, and it does not need a theorem. Take a two-contract mechanism $(0, w_H)$, (h, w_L) with $w_L < w_H$, cutoffs $v_2 < v_1$ and entry exposure $x_2 = hb(v_2)$, and write $v_0(w) = R_0^{-1}(w)$. The switch type participates on the guaranteed contract, so $R_0(v_1) \ge w_H$ and hence $v_0(w_H) \le v_1$: the single rate w_H serves everything the guaranteed tier serves, and more. Therefore

$$\Pi_H(w_H) - \Pi_2 = (w_H - \kappa)\int_{v_0(w_H)}^{v_1} b\,dF \;-\; (w_L - \kappa)\int_{v_2}^{v_1} \tau\,dF.$$

Two consequences follow. First, if $v_2 \geq v_0(w_H)$ the second integral runs over a subset of the first with $\tau \leq b$ and $w_L < w_H$, so the single rate weakly dominates the mechanism. Since $R_h(v_2) = w_L$ and $R_h = R_0\mathfrak{d}$ give $R_0(v_2) = w_L/\mathfrak{d}(x_2)$, that condition reads

$$\mathfrak{d}(x_2) \;\leq\; \frac{w_L}{w_H}, \tag{52}$$

so *a demand-driven tier can add value only by extending the market downward, never by re-pricing alone.* Second, when $v_2 < v_0(w_H)$, splitting the integral at $v_0(w_H)$ leaves a negative piece—$(w_L - \kappa)\int_{v_0}^{v_1} \tau\, dF - (w_H - \kappa)\int_{v_0}^{v_1} b\, dF < 0$, since $w_L < w_H$ and $\tau < b$—which we may drop, giving

$$\Pi_2 - \Pi_H \;\leq\; \big(w_L - \kappa\big) \int_{v_2}^{v_0(w_H)} \tau(h, v)\, dF(v) \;\leq\; \big(w_L - \kappa\big)\, b\big(v_0(w_H)\big)\Big[F\big(v_0(w_H)\big) - F(v_2)\Big], \tag{53}$$

the margin earned on the newly served band alone, with every cannibalised type ignored.

Together (52) and (53) explain why the middle increment is small, and do so without computation. A demand-driven hazard cannot reverse the ordering of willingness to pay unless $\eta_b M_L > 1$; where it cannot, $R_h < R_0$ keeps the entry cutoff close to $v_0(w_H)$, the band in (53) is narrow, and the whole value of the second contract is the margin on that band. Screening is not confined in this way: by (41) the schedule varies exposure across the entire served set, and its value is not bounded by any single band. The contrast is the analytical content of the claim that interruption is primarily a screening instrument rather than a device for transmitting demand.

Three remarks. The proof uses no property of the three classes beyond the inclusions, so it is indifferent to the shape of the optimal schedule, to whether single crossing holds, and to whether any type is reclaimed faster than the arrival rate: it is the general form of the ranking, and the three comparisons below add the structure that distinguishes the three optima. Second, the increments $\rho_2^* - \rho_H^*$ and $\rho_3^* - \rho_2^*$ are the values of relaxing one restriction at a time—first that the menu have a single element, then that the hazard be whatever demand generates—so the decomposition is a nested-constraint decomposition and each term is the marginal value of an instrument. Their relative size is not delivered by this argument. Third, nothing here requires λ to be small or capacity to be scarce; the environment is the same at every arrival rate, and κ^* adjusts.

6.5 Comparison, and why coarseness is chosen

Before the comparison, two results about menus of finite length. The first bounds the loss from coarseness; the second says how a menu of a given length is implemented.

Two facts about menus of finite length follow, and neither needs a theorem.

Refinement has quadratic returns. Suppose $x^*(\cdot)$ is Lipschitz with constant Λ_x and $|\mathcal{H}_{xx}| \leq \bar{H}$ on the served set. Partition it into n intervals of length $\ell = (\overline{v} - v_2)/n$ and let x_n be the step function equal to x^* at each midpoint; by the taxation principle x_n is an n-contract menu. Since x^* maximises $\mathcal{H}(v, \cdot)$ pointwise the first-order term vanishes, so $\mathcal{H}(v, x^*) - \mathcal{H}(v, x_n) \leq \frac{1}{2}\bar{H}(x_n - x^*)^2$

with $|x_n - x^*| \le \Lambda_x \ell/2$, giving

$$0 \;\le\; \Pi_\infty - \Pi_n \;\le\; \frac{\bar{H}\,\Lambda_x^2\,(\overline{v} - v_2)^2}{8\,n^2}. \tag{54}$$

The value of refinement therefore falls at least quadratically in the number of contracts, and the marginal value of the nth contract at rate n^{-3}.

Implementing a menu of given length. Fix cutoffs $\overline{v} = v_0 > v_1 > \cdots > v_n \ge \underline{v}$ with tier j serving $[v_j, v_{j-1})$. A tenant in tier j is displaced exactly when an arrival takes a higher tier, so thinning as in Lemma 1 gives

$$h_j = \lambda\big(1 - F(v_{j-1})\big) = \alpha(v_{j-1}), \qquad h_1 = 0, \tag{55}$$

and the supporting rates follow from the entry condition $w_n = R_{h_n}(v_n)$ and the indifference of type v_{j-1} between adjacent tiers,

$$w_{j-1} = \frac{\big[L(x_{j-1}) - L(x_j)\big]v_{j-1} + w_j\tau_j}{\tau_{j-1}}, \qquad x_i := h_i b(v_{j-1}),\ \tau_i := \tau(h_i, v_{j-1}). \tag{56}$$

Provided (25) holds at each boundary the menu induces the stated allocation, and by (55) the delivered hazard is a decreasing step function lying below $\alpha(\cdot)$ and exact at the upper end of each band, so refinement converges to efficient reallocation from below.

Beyond the ranking, the three optima differ in three respects, each a consequence of results already established rather than a new one.

Shape. The demand-driven hazard is constant at $\alpha(v_1)$ by (23); the optimal hazard is non-increasing by Lemma 5(iv), and strictly decreasing except on any interval where the occupancy cost kinks, where it is flat at λ by region (iii) of Proposition 3.

Sign. Under Section 5, $h < \alpha(v)$ for every discounted type by (31). Under a schedule the wedge varies with type: $h = 0 < \alpha(v)$ on the guaranteed region for every $v < \overline{v}$, so the top of the market is under-exposed, while $h(v) > \alpha(v)$ wherever $h(v) \ge \lambda$. That set is either empty or an interval containing the bottom of the served range, and since $x \le \bar{x}$ with $\bar{x}$ fixed by η_b through $M_L(\bar{x}) = 1/\eta_b$ and independent of λ, it is non-empty exactly when

$$\lambda \;\le\; \frac{\bar{x}}{b(\underline{v})}. \tag{57}$$

Whether a given buyer is over- or under-exposed is therefore a property of the buyer, and of how scarce capacity is, rather than of the seller's motives; the number of times h and α cross is not pinned down, and nothing below depends on it.

Extent. The served set weakly expands, strictly whenever (43) binds at $\underline{v}$; rent is driven to zero on a positive-measure set, which no two-contract menu can achieve, since there $r = \max\{v - w_H b,\ Lv - w_L\tau,\ 0\}$ and both non-trivial branches are strictly monotone; and the guaranteed rate rises further by (42).

The ranking is a revealed-preference argument: a constant schedule is admissible in (36), and

strictness follows because the solution of (41) is not constant. It is worth separating what this does and does not require. Part (i) uses no condition beyond those maintained throughout, and in particular it does not use Assumption 3. A two-element menu with selection by preference is incentive compatible by construction, whatever shape the induced allocation takes, so it lies in the feasible set of the mechanism-design problem even where single crossing fails and the expensive tier is not an interval; and the relaxation in (36) is tight under $\Lambda < 1$, which (47) derives, not under (25). Hence $\rho_3^* \geq \rho_2^*$ by Proposition 4, strictly wherever the optimal exposure varies. What failure of (25) costs is the *characterisation* of the benchmark—the cutoff parametrisation and the identity $h = \alpha(v_1)$—and not the comparison. Part (ii) is different in kind. Whether she departs *upward*—whether any type is reclaimed faster than the arrival rate—depends on (57), and manufacturing is therefore a phenomenon of abundance rather than scarcity: $\bar{x}$ is fixed by the duration–value link while $\lambda b(\underline{v})$ rises with the arrival rate, so it is when demand is thin that the seller must create the risk she wants to sell, and when demand is thick that arrivals already supply more than she wants. The same logic appears twice more in the paper—in (33), where too deep a discount empties the expensive tier and destroys the risk, and in Appendix E.9, where a slack pool has a hazard entirely of the seller's making and a saturated one leaves nothing for a schedule to do. The screening result does not depend on any of this. It is not confined to a single machine: Proposition E.8(ii) shows that a slack pool faces exactly (36) with the occupancy cost set to zero, so the same argument gives the same ranking there, and Appendix E confirms it in the pool generally, with equality only where capacity binds. The magnitudes are not comparable, however: in a slack pool the demand-driven menu collapses, so $\Pi_2 = \Pi_H$ and the screening gain is about six percent measured against the single rate, whereas the 7.0% here is Π_3/Π_2 with $\kappa = \mu\lambda$. What a schedule buys is the ability to vary exposure across *adjacent* types, which relaxes the local incentive constraint everywhere rather than at one switch point.

Part (ii) reconciles the two halves of the paper, and the reconciliation is symmetric.

Screening is two-sided, and the two sides differ in cost rather than in kind. Sorting does not require the hazard to stand in any particular relation to the arrival rate: by (25) the condition $\Delta' > 0$ involves h only through the exposure $x = hb(v)$, and nowhere compares h with λ or with α. A menu therefore screens whether the posted rate lies above or below the arrival rate, and the reversal condition of (11) governs only the further question of whether the discounted tier can reach types a single rate excludes—by (11), $d\log R_h/d\log v < 0$ requires $\eta_b M_L > 1$, and only where R_h is decreasing does a rate below $R_h(v_0)$ attract types the guaranteed rate excludes without also attracting those above. What distinguishes the two directions is price: by Corollary 1 the occupancy cost is zero for every $h \geq \lambda$ and equals $\mu(\lambda - h) > 0$, strictly decreasing in h, for every $h < \lambda$. *Screening above the arrival rate is free at the margin; screening below it is purchased.*

Which direction the optimum takes is a quantitative matter and we do not claim it in general; Table 2 exhibits both at computed optima.

At the base calibration $h(v)/\alpha(v)$ falls from 7.13 at $\underline{v}$ to 1.80 in the middle and to zero on the guaranteed region. A single number $h^* - \alpha$ is therefore not a well-defined object once the menu is

	$\eta_b = 1/2$			$\eta_b = 0.8$		
λ	1	3	6	1	3	6
h^*	3.002	3.000	6.000	1.416	1.416	1.076
h^*/λ	3.00	1.00	1.00	1.42	0.47	0.18
direction	manufactured	at the rate	at the rate	manufactured	sub-arrival	sub-arrival
Π_H	0.28003	0.18478	0.06926	0.35524	0.25348	0.11297
Π/Π_H	1.060	1.123	1.047	1.101	1.133	1.131

Table 2: Optimal two-contract menu with a freely chosen hazard, under the occupancy cost (1). F uniform on $[0.05, 1]$, $b(v) = v^{\eta_b}$, $\mu = 0.1$, duration deterministic, per-buyer accounting with $\kappa = \mu\lambda$. Both directions of distortion occur at optima, and the two middle cells sit exactly at $h = \lambda$, where (38) kinks. Menus are optimised as in Appendix C, seeded from the continuum schedule; every cell yields two clean interval boundaries with all three regimes used.

rich; the well-defined object is the crossing point.

Both directions are chosen at optima, not merely admissible. Table 2 reports the two-contract optimum with a freely chosen hazard: at $\eta_b = 1/2$ the seller manufactures at $\lambda = 1$ and sits exactly at the arrival rate at $\lambda = 3$ and $\lambda = 6$—the kink of region (iii) of Proposition 3 binding—while at $\eta_b = 0.8$ she manufactures at $\lambda = 1$ and then reclaims *below the arrival rate*, at 0.47 and 0.18 of the arrival rate, and earns 13% above the single rate while doing so. A seller who reclaims less often than requests arrive is not failing to screen; she is paying for reliability because reliability is what the discounted tier sells.

The $\eta_b = 0.8$ row also illustrates the diagnostic of Section 7.2 directly. The posted hazard is 1.416, 1.416 and 1.076 as the arrival rate rises from 1 to 3 to 6: essentially flat in absolute terms while λ rises sixfold, so its ratio to the arrival rate falls from 1.42 to 0.18 without the seller having changed what she is doing. A hazard set by the screening motive is pinned by the duration–value link and responds hardly at all to demand, whereas by (23) a demand-driven one is proportional to it.

Table 3 gives the magnitudes. The gain over demand-driven interruption is 7.0% at $\lambda = 1$ and 9.5% at $\lambda = 3$. In these calibrations the freedom to set the hazard is worth more than the freedom to add a second tier at all—the screening increment exceeds the transmission increment by a factor of 3.5 at $\lambda = 1$ and 2.2 at $\lambda = 3$—but that comparison is not delivered by Proposition 4, which orders the three levels and not the two differences, and we report it as a numerical finding rather than as a result. The schedule serves the entire market at both arrival rates, because by (43) the bottom type is profitable at the frontier contract; at higher λ that fails and exclusion returns. And most of the available gain is captured by very few contracts: at $\lambda = 1$ a two-element menu with a freely chosen hazard earns 0.296795, which is 65% of the schedule's improvement over a single rate; Table 4 gives the whole sequence. Together with Section 5.4, that is the paper's explanation of observed coarseness: refinement is feasible and incentive compatible, but the surplus it creates accrues to buyers rather than to the seller.

	Π	Π/Π_H	w_H	interruption at $\underline{v}$	served set
$\lambda = 1$, $\kappa = \mu\lambda = 0.1$					
single rate	0.280029	1.000	0.6560	—	$[0.430, 1]$
demand-driven, two tiers	0.285728	1.020	0.6599	$h = 0.502$	$[0.377, 1]$
schedule	0.305790	1.092	0.7004	$h = 7.127$	$[0.050, 1]$
$\lambda = 3$, $\kappa = 0.3$					
single rate	0.184777	1.000	0.7146	—	$[0.511, 1]$
demand-driven, two tiers	0.192957	1.044	—	$h = 1.418$	$[0.358, 1]$
schedule	0.211216	1.143	0.7512	$h = 7.127$	$[0.050, 1]$

Table 3: The three arrangements in the common environment of Section 6.4, computed at a fixed machine-hour price rather than at the endogenous κ^* of (51); the ranking of Proposition 4 does not depend on which is used, but the levels do. F uniform on $[0.05, 1]$, $b(v) = v^{1/2}$, duration deterministic, occupancy cost $\mu(\lambda - h)^+$ per hour with $\mu = 0.1$. All three rows are per-buyer profits with no admission rationing, $\gamma = 1$, so that the comparison is like for like: the schedule of Section 6 has no rationing, and imposing $\gamma = 1$ on the demand-driven arrangement is what makes the three commensurable. Table 1 instead reports that arrangement at the endogenous machine-hour price κ^*, which is why its levels differ from the fixed-κ rows here. The schedule's hazard at the bottom is $\bar{x}/b(\underline{v})$ with $\bar{x} = 1.5936$, the reversal frontier at $\eta_b = 1/2$, delivering $\bar{q} = 0.2032$.

6.6 A deadline as an alternative instrument

A natural objection is that a provider who wants to exclude long jobs should say so, since a maximum runtime screens on duration directly and destroys nothing among the jobs it admits. A *capped* contract is a pair (T, w_C): the buyer pays w_C per unit of realized time and the job is terminated at elapsed time T, yielding nothing if unfinished. There is no reclamation, so a capped tenant bears κ per hour.

A capped contract sorts on duration, and the boundary it creates is explicit. Suppose duration is deterministic. Then the capped contract is bought by

$$\big[R_0^{-1}(w_C),\ b^{-1}(T)\big] \qquad \text{and guaranteed service by} \qquad \big[\max\{b^{-1}(T),\ R_0^{-1}(w_H)\},\ \overline{v}\big],$$

so the capped contract attracts buyers if and only if

$$w_C < \min\big\{w_H,\ R_0\big(b^{-1}(T)\big)\big\}. \tag{58}$$

If in addition $w_H \le R_0\big(b^{-1}(T)\big)$ the two certain contracts are adjacent and the boundary between them is $b^{-1}(T)$, independent of every price in the menu; otherwise types in $\big(b^{-1}(T), R_0^{-1}(w_H)\big)$ buy nothing and the upper boundary is $R_0^{-1}(w_H)$.

A minimum commitment does the reverse and is dominated:

Equation (58) identifies the deadline's function, and it is not what the contract might suggest. The deadline does not exclude long jobs from the market: it denies them *cheap certainty*, forcing every job longer than T that is willing to pay w_H onto the guaranteed rate, and excluding the

n	1	2	3	4	5	6
$\lambda = 1$: *continuum* $\Pi_\infty = 0.305790$						
Π_n	0.280029	0.296795	0.302152	0.304176	0.304851	0.304851
share of gain	0.000	0.651	0.859	0.937	0.964	0.964
increment in Π	—	+5.99%	+1.80%	+0.67%	+0.22%	+0.00%
$\lambda = 3$: *continuum* $\Pi_\infty = 0.211216$						
Π_n	0.184777	0.207517	0.208890	0.210506		
share of gain	0.000	0.860	0.912	0.973		
increment in Π	—	+12.31%	+0.66%	+0.77%		

Table 4: Optimal n-contract menus against the full schedule, same primitives and accounting as Table 3. "Share of gain" is $(\Pi_n - \Pi_1)/(\Pi_\infty - \Pi_1)$. Menus are optimised over $(h_i, w_i)_{i \le n}$ with each type assigned to its preferred entry, started from discretisations of the continuum schedule and polished; see Appendix C. Every optimum contains a contract at $h = \lambda$ exactly, corroborating region (iii) of Proposition 3. The $\lambda = 3$ sequence is not convex in n, which indicates that the $n = 3$ cell is marginally under-solved rather than that the third grade is worth less than the fourth.

rest. Under Assumption 1 longer jobs are the more valuable ones, so a deadline extracts from the top, and at the optimum it destroys nothing, because every buyer who takes it completes. The hazard does the opposite, reaching buyers with long jobs and low value per machine-hour. The two condition on different things—duration alone against duration and value jointly—which is why neither substitutes for the other.

Under Assumption 1 any contract of the form $P(q) = w_R \max\{q, T\}$ with $T > 0$ is weakly dominated by a flat rate on realized time.

A minimum commitment taxes *short* jobs, and under sub-proportional runtime short jobs are the low-value ones, so the instrument falls on buyers who are already marginal. Appendix B gives the converse: where runtime rises more than proportionally and b is convex, the fixed component becomes collectable and a two-part tariff extracts the entire surplus.

We do not ask here whether a provider should add a capped contract to the schedule of Section 6, and we make no claim that one would raise revenue. The purpose of this subsection is narrower and it is twofold. It establishes that the deadline is not a substitute for the hazard, which answers the natural objection to the paper's premise; and it supplies the object that Appendix E.5 later derives from a different source, since fragmented capacity delivers the boundary of the capped-contract boundary of Section 6.6 in footprint space without any provider posting it. The commitment asymmetry is the reason the two instruments differ in kind: a deadline requires terminating a paying customer at a foreseen date against her interest, whereas the rule of Section 5 is self-enforcing term by term. Consistent with that, capped contracts are familiar where termination is externally verifiable—batch schedulers with published walltime limits—and absent from commercial pools, where the three major providers sell guaranteed and interruptible capacity and no cheaper certain contract distinguished by a maximum runtime.

7 Implications for both sides

Both sides of this market are firms. The seller chooses rates and a reliability schedule; the buyer chooses a contract and, less obviously, how much to invest in making her workload tolerant of interruption. Each rule is a corollary of the preceding sections, and the two interact.

7.1 What has to be measured, and how hard that is

Four objects govern everything above, and they differ sharply in how hard they are to obtain.

The elasticity of billed hours. By (18) this is all the guaranteed rate requires, together with the price of a machine-hour. It is measurable from a rate change and requires no knowledge of F or b at any type. This is the good case, and it is worth stating plainly because the literature's usual demand—estimate the distribution of valuations—is not needed for the largest of the seller's decisions.

The runtime elasticity η_b below the cutoff. The hazard cannot be designed from an elasticity, because exposure is $hb(v)$ and needs b pointwise. The difficulty is worst where it matters most: the gain from a discounted tier depends on the duration–value relationship *below* the guaranteed cutoff, on the segment the existing rate excludes, which generates no observations. The safe side of the frontier is estimable from a provider's own records and the profitable side is not. The asymmetry is intrinsic, not a data problem.

The exposure ceiling K. This determines the frontier reliability $\Psi_L^{-1}(\eta_b)$ of (13) and hence the reliability delivered at the bottom of the market in region (i) of Proposition 3. It is a property of the left tail of realized-to-planned duration—how often a job finishes far short of plan—and not of that ratio's variance. Estimating it from interruptible jobs is circular, because those jobs are censored by the hazard under study; the guaranteed tier is the right sample, since its jobs run to completion, but it is a selected one. For a gamma shock, $K = \max\{1, k\}$, so the exercise reduces to estimating one tail index.

The share of demand on the expensive tier. By (23), in the demand-driven arrangement the interruption rate *is* λ times this share, so the identity runs both ways: a provider who observes her own reclamation rate and arrival rate can read off the composition of her demand, and a buyer can do the same from published figures.

We do not claim that an excess component, in the sense of (3), is identified in the econometric sense. What the model supplies is a pair of restrictions rather than a signed comparison. The first is a level restriction and holds in a single facility: the demand-driven rule under-exposes by exactly (31), so an observed rate above $\lambda(1 - F(v_1))$ cannot be explained by reallocation on demand. That restriction fails in a pool, where concentration raises the per-tenant rate without any excess reclamation, and is replaced by (75), which requires the spot share and the blocking probability as well as the rate. The second is a comparative-static restriction and survives both settings: a demand-driven rate rises close to proportionally with the arrival rate—exactly so at fixed composition, and with an elasticity near one once the cutoff responds—while a screening-driven

one need not, which is the diagnostic developed in Section 7.2.

7.2 The seller

Price the guaranteed tier from an elasticity, not from a model. Solve (18). In a stationary system the target elasticity is $1/P_0$ by (22), so a provider running at 60% utilisation should be pricing where a one percent rate rise costs about two and a half percent of billed hours. Do not price off an average valuation: the marginal buyer lies below the mean under uniform values, and substituting one for the other overshoots the rate by fifteen percent and costs five percent of profit in the zero-cost case, and by ten and three percent respectively at $\kappa = \mu\lambda = 0.1$.

Do not charge the interruptible tier as though it consumed capacity. By Corollary 1 the occupancy cost is $\mu(\lambda - h)^+$, zero for every $h \geq \lambda$. A provider who charges the discounted tier at the guaranteed tier's opportunity cost will set the discount too shallow and forgo the coverage the tier exists to create.

Distinguish the two functions of interruption, and test them separately. Interruption transmits demand and screens buyers, and the two leave different signatures. Three statements must be kept apart here, and we have not always been careful to do so. *As an identity*, (23) gives $h = \lambda(1 - F(v_1))$ and (72) gives $h = \lambda(1 - F(v_1))P_B/s$, so at fixed composition a demand-driven rate is exactly proportional to the arrival rate. *As an equilibrium comparative static* this is false, because v_1 responds: in Table 1 the ratio h/λ falls from 0.528 to 0.458 as λ rises from 0.5 to 3, an elasticity of $d\log h/d\log\lambda \approx 0.92$ rather than one. A demand-driven rate therefore rises *close to* proportionally, with the shortfall equal to the response of the switch cutoff. A screening-driven rate need not do even that: by (25) it is set against the exposure the seller wants to deliver rather than against the arrivals she receives, and in Table 2 the posted hazard is almost unchanged in absolute terms—1.416, 1.416, 1.076—as λ rises from 1 to 6, so its ratio to the arrival rate falls from 1.42 to 0.18, an elasticity near zero. *As an identification strategy* neither statement is one: both are model-implied relations between two equilibrium objects, and recovering them from data would require variation in λ that is independent of the composition of demand and of the provider's pricing response. We do not provide that, and the contrast should be read as a qualitative signature—an elasticity near one against one near zero—rather than as an estimating equation. The elasticity of the reclamation rate to the arrival rate is therefore a sharper diagnostic than its level, and unlike the level it is not confounded by the concentration factor of (74), since both benchmarks scale with λ while a screening-driven hazard need not.

Recognise that a demand-driven rule need not interrupt at the efficient rate, in either direction. In a single facility, if reclamation is triggered only by on-demand arrivals then by (31) every spot tenant is protected relative to efficiency, and the protection is largest for the least valuable jobs. In a pool the sign can reverse: by (74) concentration multiplies the hazard by $1/s$,

so a tightly run pool over-exposes every discounted buyer while manufacturing nothing: every reclamation there is matched to an arrival, and it is concentration by $1/s$ that raises the per-tenant rate. The two cases are distinguished by (75) rather than by the level of h alone. A provider who wants the reliability schedule to track value must post it, which is Section 6; if she merely reclaims on demand she is delivering the efficient rate of one type to everybody.

Keep the discount shallow enough to sustain the risk. By (33) the interruption rate is generated by the buyers who decline it. A discount deep enough to empty the expensive tier destroys the reliability difference the discount is priced against, and profit falls below the single-rate benchmark. This is a constraint no analysis with an exogenous hazard can see.

Expect rapidly diminishing returns from refinement. Successive grades are worth 6.0, 1.8, 0.7 and 0.2 percent of profit at $\lambda = 1$, and a sixth is worth nothing measurable. Two grades is not a crude approximation to the optimum—it captures 65% of the gain at $\lambda = 1$ and 86% at $\lambda = 3$—and a third is worth having if a product line costs less than a percent or two of revenue to run. Beyond four the question is closed.

Check reliability against the frontier. The completion probability offered to the marginal buyer cannot exceed $\Psi_L^{-1}(\eta_b)$. A tier that completes marginal jobs more often than the frontier allows is not screening on exposure, whatever else it may be doing.

7.3 The buyer

A published reclamation rate is not a technology. By (23) it is λ times the share of demand on the expensive tier, so it moves when prices move and when the composition of demand moves, with the infrastructure unchanged. A buyer planning on a historical rate is extrapolating other buyers' choices. Her risk is created by other buyers, and their choices are strategic complements: more buyers on the expensive tier raises h, which pushes still more of them there.

Evaluate reliability at your own runtime. A pool's aggregate completion rate is an average over entrants and shorter jobs complete more often, so the published figure overstates what a long job receives. Compute $L\big(hb(v)\big)$ at your own $b(v)$. Under a schedule the bias is larger, because contracts are sorted by duration as well as differing in reliability.

The posted prices define a threshold exposure. Wherever $M_L > 1$ the function $\mathfrak{d}$ is strictly decreasing and invertible, so a buyer facing published w_L, w_H and h can compute

$$\widehat{x} := \mathfrak{d}^{-1}\big(w_L/w_H\big), \tag{59}$$

the exposure at which the posted discount stops being worth taking. If she must complete, retrying after each reclamation, her expected billed time is $b(v)/\mathfrak{d}(x)$, and dividing the job into n saved

stages replaces x by x/n, so the interruptible tier is cheaper exactly when $w_L/w_H < \mathfrak{d}(x/n)$, that is when

$$n > \frac{h\, b(v)}{\widehat{x}}. \tag{60}$$

A buyer whose job is shorter than the threshold should take the tier as it stands; one whose job is twice as long needs at least three saved stages, one three times as long at least four. Both quantities are observable to her.

What checkpointing is worth. Cost falls in n from $w_L b(v)/\mathfrak{d}(x)$ toward $w_L b(v)$, so the achievable saving is $w_L b(v)\big[1/\mathfrak{d}(x) - 1\big]$, and the saving relative to remaining on guaranteed service converges to $\big(w_H - w_L\big)b(v)$. That is the ceiling on what a buyer should spend, per job, on restartability. A deeper discount is a larger prize for checkpointing, not merely a cheaper machine.

7.4 The two sides together

The seller's screen runs on the buyer's runtime, and the buyer can change her runtime. Dividing a job into n stages replaces $M_L(x)$ by $M_L(x/n)$, so buyer-side investment in checkpointing erodes precisely the elasticity the seller exploits. Two implications follow. The discount is bounded by the buyers' cost of becoming interruption-tolerant: a discount deep enough to make checkpointing worthwhile induces the investment that destroys the screen. And improvements in checkpointing technology—finer snapshots, cheaper state serialization, restartable frameworks—should compress discounts, and in the limit of costless checkpointing eliminate the tier, since the reliability gap $1 - L(x/n)$ on which sorting rests vanishes as n grows. Note the two margins go at different rates: the expansion mode requires $\eta_b M_L(x/n) > 1$ and fails at finite n once $M_L(x/n)$ drops below $1/\eta_b$, whereas the re-pricing mode survives for every finite n, since (25) constrains h only through $x = hb(v)$, and disappears only in the limit. That is a sharper prediction than a price–reliability relation, because the driver is observable independently of the contract.

Nor is more screening socially better. Buyers lose across the three arrangements of this paper. Under a schedule the seller extracts the whole surplus from types in $[v_2, v^*]$, serving them at the worst reliability consistent with their buying at all; interior types earn strictly less rent than a coarse menu leaves them; and the guaranteed segment pays more. The gain from refinement is a transfer, and the provider's reluctance to refine is, from the buyers' side, a benefit.

8 Conclusion

Interruption in a usage-priced market performs two functions, and the paper's organising result separates them. It transmits demand, because a machine held by one tenant cannot serve another; and it screens, because how much a given rate of interruption costs a buyer depends on how long she needs the machine, which is what her value determines. The separation is a profit ranking, each step signed by revealed preference: a seller restricted to one rate does worse than one who adds a discounted tier whose risk is generated by demand alone, and that seller in turn does worse

than one who chooses the risk schedule. It is the second inequality that identifies screening as a separate function, and it is strict wherever the optimal exposure varies.

Manufacturing risk is one half of the screening instrument. Sorting needs only that some risk be posted, so the seller sets the rate away from the reallocative one in whichever direction screening points—above it where exposure is free to supply and below it where it must be paid for—and which direction appears is a matter of parameters rather than of motives. Stating the mechanism as "providers reclaim more than reallocation requires" captures one direction and misses the other, and the direction that is missed is the one that appears where the duration–value link is strong.

A separate result explains the coarseness of observed menus. An ordered menu of n tiers delivers to each tier the efficient intensity of its own marginal buyer, so the efficient reallocation schedule is attainable under private values and refinement is incentive compatible at every resolution; what stops refinement is that its value to the seller falls away quickly, while the surplus it would create accrues to buyers. The obstruction to efficiency is profit rather than information, and the provider's conservatism is a form of consumer protection.

Three arrangements make this concrete and they nest. A single rate sells machine-hours rather than units, and the optimum sets the elasticity of billed hours to the reciprocal of the idle fraction; the marginal buyer is not the average buyer, and the condition requires nothing the seller cannot observe. Two rates with reclamation triggered only by demand generate the hazard $\alpha(v_1)$—the efficient intensity of the marginal switcher applied to everyone—so every discounted buyer is *under-exposed* relative to efficiency by exactly a ratio of survival functions; the rule is ex post optimal without inference and needs no commitment, but it survives only if the discount is shallow enough that enough buyers remain on the expensive tier to generate the risk being sold. A full schedule makes exposure decreasing, and interruption then acquires a second function beyond transmitting demand: the seller sets it away from the reallocative rate in whichever direction screening points, manufacturing at the bottom of the market where exposure is free to supply—when capacity is scarce enough that the frontier no longer reaches the arrival rate, not even there—and holding it below the arrival rate at the top, where exposure must be paid for. Both directions occur at optima, including cells in which the seller reclaims well below the arrival rate and still earns substantially more than a single rate; Section 6 tabulates them. She also holds a positive mass of buyers at zero rent and at the frontier reliability, and no longer chooses the guaranteed rate at all, which becomes the residual of the schedule below it. Profit rises across the three, and the increments separate the two functions of interruption: $\rho_2^* - \rho_H^*$ measures what transmitting demand is worth and $\rho_3^* - \rho_2^*$ what the screening function adds on top of it. Both are strictly positive by Proposition 4, and because the three mechanism classes are nested each is the value of relaxing one restriction. Their relative size is a separate question, which that argument does not settle—it orders the levels, not the differences—and Section 6 reports what the computed optima give. Buyers lose across all three.

The most useful of the paper's negative results is that refinement pays rapidly less, and at a rate we can bound: the shortfall from a menu of n contracts falls quadratically in n, so a two- or three-element menu captures most of the schedule's improvement over a single rate and further

grades add little. Observed coarseness is therefore consistent with profit maximisation and needs no appeal to an informational or technological constraint, and the surplus that finer tiers would create accrues to buyers. We stop short of claiming that a third grade is worthless: it is worth having if a product line is cheap to run, and the model does not price that cost. Section 6 gives the magnitudes.

Five limitations bound the claim. First, and most consequential, the accounting of freed capacity. Corollary 1 and regions (i)–(iii) of Proposition 3 presume that a reclamation is refilled. In a system with blocking a manufactured reclamation—one matched to no arrival—empties the machine without filling it, and the manufactured component nearly vanishes: a committed hazard floor on top of demand-driven interruption is then worth almost nothing, and does not bind at all when capacity is ample. We take the refillable case as the baseline, appropriate for a commercial pool, and regard the loss system as the boundary case appropriate for a single facility; which world a given provider is in is an empirical question with a clear answer in her utilisation data. Second, sorting in Section 5 is an assumption rather than a theorem, and it is one on the candidate menu rather than on primitives; and where the hazard map admits several equilibria we select the greatest, which is stable and seller-preferred, rather than deriving uniqueness. Assumption 3 is stated sharply—it is $\Delta' > 0$ written out—and has the transparent sufficient condition (26), but it is not implied by the primitives, and where it fails the expensive tier need not be an upper interval. The hazard itself survives, since Lemma 1 needs only the measure of the set that chooses it. Third, value and runtime are tied by a common-knowledge $b(\cdot)$. Proposition E.6 relaxes this: with both privately known, the guaranteed region is bounded by an increasing curve $v^*(b)$ rather than a point, sorting still runs on value per machine-hour, and the hazard is still an anonymous scalar—but the cutoff parametrisation and the identity $h = \alpha(v_1)$ are lost. Stated in two dimensions the substantive restriction is that value and runtime be positively but less than proportionally associated, which is what Assumption 1 encodes and is testable in a way that $b(\cdot)$ is not. Fourth, the billing rule is substantive rather than conventional: if payment were contingent on completion, willingness to pay per billed unit would revert to its guaranteed level and the screen would disappear.

A fifth item is scale, and it is a finding as much as a limitation. Displacement requires that no machine be idle; a discounted tier requires that some machine be. In a large unfragmented pool those cannot hold at once, so the demand-driven account does not survive the fluid limit: the pool is either slack, in which case nobody is displaced and the menu collapses, or saturated, in which case occupancy is one and the price per occupied hour is the only question left. Read together with (53), which bounds what transmission can be worth even where it does operate, this is the paper's strongest argument that the interruption a commercial provider posts is not primarily reallocative. It also means Section 5 should be read as a transmission benchmark and not as a model of a commercial pool.

That leaves the positive question open: what does generate interruption at scale? Appendix E.5 examines the natural candidate. A request for m co-located machines can fail while many machines are free but scattered, so fragmentation raises blocking; and solving the domain occupancy by the

Kaufman–Roberts recursion shows that it does not change the limit. Blocking is $(1 - p_m)^D$ across D domains, so it vanishes geometrically, at a rate governed by p_m, the chance one domain can host the job. What survives is a statement about composition at a given scale: $1/p_m$ is a few dozen for half-domain jobs and of order 10^{10} for jobs needing a whole domain, so a provider with a few thousand domains supplies no blocking for small jobs and near-certain blocking for the largest. That is a prediction about which customers face reclamation, not a foundation for demand-driven interruption at scale. We regard the positive question—what a full finite-pool model with endogenous placement would deliver—as the most important thing the paper does not do. Appendix E works this out and examines the candidates. Fragmentation is the most promising—a request for m co-located machines can fail while many machines are free but scattered—but solving the domain occupancy shows blocking still vanishes geometrically in the number of domains. Modelling a finite pool with endogenous placement, rather than taking the fluid limit, is what remains. We regard it as the most important thing the paper does not do.

Appendix E.9 closes the gap between the two halves of that limitation by joining the chosen schedule to the pool, and the two regimes divide cleanly: a two-tier menu is dominated by a single rate whenever the pool is saturated, and whenever it is not, the hazard a spot tenant faces contains no demand-driven component at all, because displacement is required only when no machine is idle. Wherever an interruptible tier is optimal in a fluid pool, its interruption rate is therefore of the seller's own making, which is the sharpest form of the paper's central claim and is reached without any appeal to her motives.

The empirical content is a pair of restrictions rather than a decomposition. In a single facility (23) pins what a demand-driven rule delivers, so an observed rate above λ times the share of demand on the expensive tier cannot be explained by reallocation on demand; in a pool that restriction fails, because concentration raises the per-tenant rate with nothing manufactured, and it must be replaced by one requiring the discounted share and the blocking probability as well as the rate. What survives both settings is a comparative static: a demand-driven rate scales with the arrival rate and a screening-driven one need not, so the elasticity of reclamation to demand identifies the second function where its level does not. The single open question we would most like settled is the one named above: whether fragmentation, which Appendix E.5 shows can sustain a demand-driven hazard at scale and can mimic a posted deadline in footprint space, does so at the magnitudes observed. That requires the occupancy analysis we do not attempt, and it would settle both what generates interruption in a commercial pool and whether the capped contract of Section 6.6 is absent because providers already have its equivalent for free.

References

Abhishek, V., I. A. Kash, and P. Key (2012): "Fixed and Market Pricing for Cloud Services," *Proceedings IEEE INFOCOM Workshops*, 157–162.

Afèche, P. (2013): "Incentive-Compatible Revenue Management in Queueing Systems: Optimal Strategic Delay," *Manufacturing & Service Operations Management*, 15(3), 423–443.

Afèche, P., and J. M. Pavlin (2016): "Optimal Price/Lead-Time Menus for Queues with Customer Choice: Segmentation, Pooling, and Strategic Delay," *Management Science*, 62(8), 2412–2436.

Agmon Ben-Yehuda, O., M. Ben-Yehuda, A. Schuster, and D. Tsafrir (2013): "Deconstructing Amazon EC2 Spot Instance Pricing," *ACM Transactions on Economics and Computation*, 1(3), 1–20.

Belleflamme, P. (2005): "Versioning in the Information Economy: Theory and Applications," *CESifo Economic Studies*, 51(2–3), 329–358.

Bergemann, D., and R. Deb (2025): "Robust Pricing for Cloud Computing," Cowles Foundation Discussion Paper No. 2423, Yale University.

Bergemann, D., and M. C. Wang (2025): "Optimal Pricing of Cloud Services: Committed Spend under Demand Uncertainty," Cowles Foundation Discussion Paper No. 2424, Yale University.

Bhargava, H. K., and V. Choudhary (2008): "When Is Versioning Optimal for Information Goods?" *Management Science*, 54(5), 1029–1035.

Chao, H.-P., and R. Wilson (1987): "Priority Service: Pricing, Investment, and Market Organization," *American Economic Review*, 77(5), 899–916.

Chen, Y.-J., and K.-W. Huang (2016): "Pricing Data Services: Pricing by Minutes, by Gigs, or by Megabytes per Second?" *Information Systems Research*, 27(3), 596–617.

Chen, S., H. Lee, and K. Moinzadeh (2019): "Pricing Schemes in Cloud Computing: Utilization-Based vs. Reservation-Based," *Production and Operations Management*, 28(1), 82–102.

Chen, S., K. Moinzadeh, and Y. Tan (2021): "Discount Schemes for the Preemptible Service of a Cloud Platform with Unutilized Capacity," *Information Systems Research*, 32(3), 967–986.

Deneckere, R. J., and R. P. McAfee (1996): "Damaged Goods," *Journal of Economics & Management Strategy*, 5(2), 149–174.

Dierks, L., I. A. Kash, and S. Seuken (2021): "On the Cluster Admission Problem for Cloud Computing," *Journal of Artificial Intelligence Research*, 71, 1–40.

Dierks, L., and S. Seuken (2022): "Cloud Pricing: The Spot Market Strikes Back," *Management Science*, 68(1), 105–122.

Fay, S., and J. Xie (2008): "Probabilistic Goods: A Creative Way of Selling Products and Services," *Marketing Science*, 27(4), 674–690.

Harris, M., and A. Raviv (1981): "A Theory of Monopoly Pricing Schemes with Demand Uncertainty," *American Economic Review*, 71(3), 347–365.

Hart, S., and P. J. Reny (2015): "Maximal Revenue with Multiple Goods: Nonmonotonicity and Other Observations," *Theoretical Economics*, 10(3), 893–922.

Huang, J., R. J. Kauffman, and D. Ma (2015): "Pricing Strategy for Cloud Computing: A Damaged Services Perspective," *Decision Support Systems*, 78, 80–92.

Kilcioglu, C., J. M. Rao, A. Kannan, and R. P. McAfee (2017): "Usage Patterns and the Economics of the Public Cloud," *Proceedings of the 26th International Conference on World Wide Web*, 83–91.

Kurtz, T. G. (1970): "Solutions of Ordinary Differential Equations as Limits of Pure Jump Markov Processes," *Journal of Applied Probability*, 7(1), 49–58.

Ma, D., and A. Seidmann (2015): "Analyzing Software as a Service with Per-Transaction Charges," *Information Systems Research*, 26(2), 360–378.

Manelli, A. M., and D. R. Vincent (2006): "Bundling as an Optimal Selling Mechanism for a Multiple-Good Monopolist," *Journal of Economic Theory*, 127(1), 1–35.

Matthews, S. (1983): "Selling to Risk Averse Buyers with Unobservable Tastes," *Journal of Economic Theory*, 30(2), 370–400.

McAfee, R. P. (2007): "Pricing Damaged Goods," *Economics: The Open-Access, Open-Assessment E-Journal*, 1(2007-1), 1–19.

Mendelson, H., and S. Whang (1990): "Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue," *Operations Research*, 38(5), 870–883.

Milgrom, P., and I. Segal (2002): "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, 70(2), 583–601.

Mussa, M., and S. Rosen (1978): "Monopoly and Product Quality," *Journal of Economic Theory*, 18(2), 301–317.

Myerson, R. B. (1981): "Optimal Auction Design," *Mathematics of Operations Research*, 6(1), 58–73.

Oi, W. Y. (1971): "A Disneyland Dilemma: Two-Part Tariffs for a Mickey Mouse Monopoly," *Quarterly Journal of Economics*, 85(1), 77–96.

Pavlov, G. (2011): "Optimal Mechanism for Selling Two Goods," *The B.E. Journal of Theoretical Economics*, 11(1), Article 3.

Riley, J., and R. Zeckhauser (1983): "Optimal Selling Strategies: When to Haggle, When to Hold Firm," *Quarterly Journal of Economics*, 98(2), 267–289.

Sundararajan, A. (2004): "Nonlinear Pricing of Information Goods," *Management Science*, 50(12), 1660–1673.

Varian, H. R. (2000): "Versioning Information Goods," in *Internet Publishing and Beyond*, ed. by B. Kahin and H. R. Varian. MIT Press, 190–202.

Wilson, R. (1989): "Efficient and Competitive Rationing," *Econometrica*, 57(1), 1–40.

Xu, H., and B. Li (2013): "Dynamic Cloud Pricing for Revenue Maximization," *IEEE Transactions on Cloud Computing*, 1(2), 158–171.

Zhang, Z., K. Joseph, and R. Subramaniam (2015): "Probabilistic Selling in Quality-Differentiated Markets," *Management Science*, 61(8), 1959–1977.

Appendix A: Proofs

Corollary 1. Immediate from (1): $c(h) = \mu(\lambda - h)^+$ is zero for $h \geq \lambda$ and equals $\mu(\lambda - h) > 0$ for $h < \lambda$, with $\partial c/\partial h = -\mu < 0$ there. The interpretation follows from the construction of (1): a contract permitting reclamation at rate h serves $\min\{h, \lambda\}$ requests per unit of occupied time, so once $h \geq \lambda$ every request is already met and a further increase in h meets none. ■

The reallocative schedule of Section 3. Requests arrive as a Poisson process of rate λ with values drawn i.i.d. from F, independently of arrival times. Thinning by the independent event $\{v' > v\}$, which has probability $1 - F(v)$, leaves a Poisson process of rate $\alpha(v)$, so the waiting time to the first superior arrival is exponential with that rate and independent of elapsed runtime. For the shape, $\log \beta(v) = \log \lambda + \log(1 - F(v)) + \log b(v)$ has derivative $-f/(1 - F) + b'/b = \big[\eta_b(v) - vf(v)/(1 - F(v))\big]/v$. The bracket is positive at $\underline{v}$ under Assumption 1 and negative as $v \to \overline{v}$, where $vf/(1 - F) \to \infty$; it crosses once provided $vf/(1 - F) - \eta_b$ is increasing, which holds whenever F has an increasing generalised failure rate and η_b is non-increasing, and in particular for the uniform–power family used throughout. Then β is single-peaked. ■

Proposition 1. (i) The bill is $w\, b(v)\varepsilon$, linear in ε, and the buyer is risk neutral, so participation requires $v \geq w\, b(v)\mathbb{E}[\varepsilon] = wb(v)$; $h = 0$ makes completion certain, so nothing else depends on ε, and (15) uses $\mathbb{E}[\varepsilon] = 1$.

(ii) Only the boundary of (15) moves with w. Differentiating $R_0(v_0(w)) = w$ and using $R_0' = (1 - \eta_b)R_0/v$ with $R_0(v_0) = w$ gives $dv_0/dw = b(v_0)/\big(1 - \eta_b(v_0)\big) > 0$, whence $M'(w) = -b(v_0)^2 f(v_0)/\big(1-\eta_b(v_0)\big)$. Then $\Pi_H'(w) = M+(w-\kappa)M' = M\big[1-(w-\kappa)\varepsilon/w\big]$ with $\varepsilon = -wM'/M$, which after substitution is the expression in (18); setting $\Pi_H' = 0$ and using $M > 0$ gives (18), and $\kappa = 0$ gives $\varepsilon = 1$.

(iii) Since R_0 is a strictly increasing bijection, optimizing over w is optimizing over the cutoff t, with $\widetilde{\Pi}(t) = (R_0(t) - \kappa)B(t)$. Then $\widetilde{\Pi}'(t) = R_0'B - (R_0 - \kappa)bf$, and substituting $R_0' = (1 - \eta_b)R_0/t$ and dividing by $R_0bf > 0$ gives $(1-\eta_b)\rho_b(t)/t = 1-\kappa b(t)/t$, which is (19). By (17), $B/\mathbb{E}[b] = 1-F_b$ and $bf/\mathbb{E}[b] = f_b$, so ρ_b is the inverse hazard rate of F_b; with $\kappa = 0$ the condition is $t = (1-\eta_b)\rho_b(t)$, and since $1 - \eta_b < 1$ any solution lies strictly below a solution of $t = \rho_b(t)$ wherever $t - \rho_b(t)$ is increasing. Uniqueness is Assumption 2, and $\widetilde{\Pi}(\overline{v}) = 0 < \widetilde{\Pi}(\underline{v})$ rules out the upper corner.

(iv) With $N(w) := 1 - F(v_0(w))$ the same differentiation gives $N' = -b(v_0)f(v_0)/(1 - \eta_b)$, so $\varepsilon_{\text{buyers}}/\varepsilon_{\text{hours}} = (M/N)/b(v_0)$, and $M/N = \mathbb{E}[b \mid v \geq v_0] > b(v_0)$ because b is increasing.

(v) With f constant, $B(t) = (\overline{v}^{1+\eta} - t^{1+\eta})/[(1 + \eta)(\overline{v} - \underline{v})]$ and $b(t)f(t) = t^{\eta}/(\overline{v} - \underline{v})$, so (19) with $\kappa = 0$ becomes $(1 - \eta)(\overline{v}^{1+\eta} - v_0^{1+\eta}) = (1 + \eta)v_0^{1+\eta}$, i.e. $2v_0^{1+\eta} = (1 - \eta)\overline{v}^{1+\eta}$, which is (21); substituting into (15) gives $M(w^*) = \overline{v}^{1+\eta}/[2(\overline{v} - \underline{v})]$ and hence Π_H^*. Finally $v_0 < \overline{v}/2 \iff (1 - \eta)/2 < 2^{-(1+\eta)} \iff g(\eta) := (1 - \eta)2^{\eta} < 1$, and $g(0) = 1$ with $g'(\eta) = 2^{\eta}[(1 - \eta)\log 2 - 1] < 0$ on $[0, 1]$ because $(1 - \eta)\log 2 \leq \log 2 < 1$, so g is strictly decreasing and $g(\eta) < 1$ for $\eta > 0$; since $\mathbb{E}[v] = (\underline{v} + \overline{v})/2 \geq \overline{v}/2$ the last claim follows. The general statement is that $v_0 < \mathbb{E}[v]$ iff $\Phi(\mathbb{E}[v]) > 0$, which is immediate from Assumption 2.

(vi) With one machine and no queue, a cycle consists of an idle spell of mean $1/(\lambda p)$, $p = 1 - F(v_0)$, and a service of mean $\bar{B} = \mathbb{E}[b \mid v \geq v_0]$; revenue per cycle is $w\bar{B}$, and $p\bar{B} = M(w)$, so renewal–reward gives $\rho(w) = w\bar{B}/\big(1/(\lambda p)+\bar{B}\big) = \lambda wM/(1+\lambda M)$. Differentiating, $\rho' \propto M + wM' + \lambda M^2$, which vanishes iff $\varepsilon = 1+\lambda M$; utilisation is $U = \lambda M/(1+\lambda M)$, so $1+\lambda M = 1/(1-U)$. Since maximizing N/D implies the optimum maximizes $N - \rho^* D$, the shadow price of a machine-hour is ρ^*, and $\rho^* = w^*U^*$ follows from the displayed expression. For uniform F and $b = v^{\eta}$, normalised so that $\overline{v} = 1$, substituting $\varepsilon = (1 + \eta)z/[(1 - \eta)(1 - z)]$ and $M = (1 - z)/[(1 + \eta)(\overline{v} - \underline{v})]$ with $z = v_0^{1+\eta}$ into $\varepsilon = 1 + \lambda M$ and setting $u = 1 - z$ gives $cu^2 + 2u - (1 + \eta) = 0$. Since $u(c)$ is strictly decreasing, v_0 and w^* are strictly increasing in λ. ■

Lemma 1. An arrival takes the expensive tier if and only if her value lies in S, an event of probability $\Pr(v \in S)$ independent across arrivals and of arrival times. Thinning a Poisson process by an independent Bernoulli trial leaves a Poisson process of rate $\lambda \Pr(v \in S)$, so the time to its first arrival is exponential with that rate, independently of the incumbent's elapsed runtime and of her type. The argument uses only the measure of S and not its shape. Under Assumption 3 and Lemma 2, $S = [v_1, \overline{v}]$ and $\Pr(v \in S) = 1 - F(v_1)$, which is (23). ■

Lemma 2. Differentiating (24) and using $L' < 0$ gives (28). Substituting $b'(v) = \eta_b(v)/R_0(v)$ and rearranging, $\Delta'(v) > 0$ is precisely (25), so Assumption 3 gives $\Delta' > 0$ directly and Δ has at most one zero v_1, with $\Delta < 0$ below and > 0 above. Next $U_H \geq 0 \iff R_0 \geq w_H$, an upper interval by Assumption 1, on which $U_H' = 1 - \eta_b w_H/R_0 \geq 1 - \eta_b > 0$. And $U_L \geq 0 \iff w_L \leq R_h(v)$ with $d\log R_h/d\log v = 1 - \eta_b M_L(x)$ by (11); with M_L non-decreasing and $x = hb(v)$ increasing, that expression crosses zero at most once from above, so R_h is quasi-concave and $\{U_L \geq 0\} = [v_2, v_3]$ is an interval entered from below at v_2, whence $R_h'(v_2) > 0$, i.e. $\eta_b(v_2)M_L(x_2) < 1$. Above v_1: $U_H(v_1) = \gamma U_L(v_1) \geq 0$ and U_H increasing give $U_H > 0$, so guaranteed service is chosen. On $[v_2, v_1)$: $\Delta < 0$ and $U_L \geq 0$, so the cheap contract is chosen, and $v_1 \in [v_2, v_3]$ gives $v_1 \leq v_3$. Below v_2:

$U_L < 0$ and $\Delta < 0$ imply $U_H < \gamma U_L < 0$, so nothing is bought and no island of guaranteed buyers can appear. ■

Corollary 2. (i) b and R_0 are strictly increasing and the guaranteed set lies strictly above the cheap set; the conditional-mean statement follows since b is increasing. (ii) L is strictly decreasing and $x = hb(v)$ strictly increasing. ■

Proposition 2. (i) The second equation of (27) defines w_L; solving the first for w_H and using $R_0 = v/b$ gives $w_H = \big[v_1 - \gamma U_L(v_1)\big]/b(v_1) = R_0(v_1) - \gamma U_L(v_1)/b(v_1)$.

(ii) $h^* = \alpha(v_1)$ is (23), so $x_1 = h^* b(v_1) = \alpha(v_1)b(v_1) = \beta(v_1)$. For $v < v_1$, $h^*/\alpha(v) = (1 - F(v_1))/(1-F(v)) < 1$ because F is strictly increasing, and $h^* - \alpha(v) = \lambda\big[(1-F(v_1)) - (1-F(v))\big] = -\lambda\big[F(v_1) - F(v)\big]$.

(iii) The rates depend on the cutoffs through (27): writing $N := (1 - \gamma L(x_1))v_1 + \gamma w_L \tau_1$, so that $w_H = N/b_1$, and differentiating, $dN/dv_1 = (1-\gamma L_1) + \gamma |L'(x_1)| b_1'(hv_1 + w_L) = \Delta'(v_1) + w_H b_1'$ by (28), whence $\partial w_H/\partial v_1 = \Delta'(v_1)/b_1$; and $\partial w_L/\partial v_2 = R_h'(v_2)$ with $\partial w_H/\partial w_L = \gamma\tau_1/b_1$. The rates depend on the cutoffs through (27) as computed above, and B and T are differentiable in the cutoffs, so S, E and H are well defined. Since h is not free but equals $\alpha(v_1)$, the total derivative of ρ in v_1 is $S + H\alpha'(v_1) = S - \lambda f(v_1) H$; setting it and E to zero gives (32).

(iv) Immediate from (23). ■

The four comparisons of Section 5. (i) For any w_H, taking $w_L \geq w_H$ gives $U_L(v) \leq U_H(v)$ for every v, since $L \leq 1$ and $\tau \leq b$; with $\gamma \leq 1$ the cheap contract is then never strictly preferred, so the two-tier problem attains every single-rate profit and $\Pi \geq \Pi_H$. For strictness, set $w_H = w^*$ and $w_L = R_h(v_0) - \epsilon$ with $h = \alpha(v_0) > 0$: types in an interval below v_0, previously unserved, now buy at a strictly positive rate and bill strictly positive time, a first-order gain, while the mass of types above v_0 diverted is $O(\epsilon)$ and each diversion costs $O(\epsilon)$, a second-order loss.

(ii) The switch type participates on the guaranteed contract, so $R_0(v_1) \geq w_H$ and $v_0(w_H) \leq v_1$; since B is decreasing, $B(v_1) \leq B(v_0(w_H)) = M(w_H)$ and hence $(w_H - \kappa)B(v_1) \leq \Pi_H(w_H) \leq \max_w \Pi_H(w) = \Pi_H$. The inequality is strict whenever the discounted tier is non-degenerate, since U_L is strictly increasing with $U_L(v_2) = 0$, so $v_1 > v_2$ gives $U_L(v_1) > 0$ and $v_1 > v_0(w_H)$.

(iii) $v_2 < v_0$ and $\mathfrak{d}(x_2) < 1$ for $x_2 > 0$ give $w_L = R_0(v_2)\mathfrak{d}(x_2) < R_0(v_2) < R_0(v_0) = w^*$.

(iv) By (i), $w_H^* = R_0(v_1) - r(v_1)/b(v_1)$ with $r(v_1) = \gamma U_L(v_1) \geq 0$, and $w^* = R_0(v_0)$, which gives the equivalence. If $w_H^* \geq w^*$ then $R_0(v_1) \geq w_H^* \geq w^* = R_0(v_0)$, and R_0 strictly increasing gives $v_1 \geq v_0$, strictly if either inequality is strict. ■

Lemma 3. Let $\Psi(z, v)$ be type v's payoff from the contract indexed by z. Incentive compatibility gives $r(v) = \max_z \Psi(z, v)$ and, by the envelope theorem, $r'(v) = \Psi_v(v, v)$, the derivative taken at fixed (h, w). From $\Psi = L(hb(v))v - w(1 - L(hb(v)))/h$,

$$r'(v) = L(x) + vL'(x)hb'(v) + wL'(x)b'(v).$$

Substituting $hb'(v) = x\eta_b(v)/v$, $b'(v) = \eta_b b/v$ and (34) for w,

$$r' = L + x\eta_b L' + \frac{x\eta_b L'}{1-L}\Big(L - \frac{r}{v}\Big) = L + \frac{x\eta_b L'}{1-L}\Big(1 - \frac{r}{v}\Big) = L - \eta_b\widetilde{\mathfrak{d}}\Big(1 - \frac{r}{v}\Big),$$

which is (35); $\mathcal{B} > 0$ because $L' < 0$. ■

Lemma 4. The Hamiltonian of (36) is $\mathcal{H} = \big[L(x)v - r - c\big]f + p\big[\mathcal{A} + \mathcal{B}r\big]$, so $p' = -\partial\mathcal{H}/\partial r = f - pB$ with $p(\overline{v}) = 0$. With integrating factor E, $(pE)' = Ef$, and integrating from v to $\overline{v}$ gives $p(v) = -E(v)^{-1}\int_v^{\overline{v}} Ef = -W(v)$. Since $\mathcal{B} \geq 0$, E is non-decreasing, so $E(s)/E(v) \geq 1$ for $s \geq v$ and $W(v) \geq 1 - F(v)$, with equality iff $\mathcal{B} \equiv 0$. Equivalently, a perturbation δx at v_0 changes $r(v)$ for $v > v_0$ by $\big[E(v)/E(v_0)\big]\big[\mathcal{A}_x + \mathcal{B}_x r\big]\delta x\, dv$, so its total effect on $\int rf$ is $W(v_0)\big[\mathcal{A}_x + \mathcal{B}_x r\big]\delta x\, dv$. ■

Derivation of (41). Maximizing $\mathcal{H}$ in x with $p = -W$, $\mathcal{H}_x = \big(L'(x)v - c_x\big)f - W\big[\mathcal{A}_x + \mathcal{B}_x r\big] = 0$. From $\mathcal{A} = L - \eta_b\widetilde{\mathfrak{d}}$ and $\mathcal{B} = \eta_b\widetilde{\mathfrak{d}}/v$ we get $\mathcal{A}_x = L' - \eta_b\widetilde{\mathfrak{d}}'$ and $\mathcal{B}_x = \eta_b\widetilde{\mathfrak{d}}'/v$, so

$$L'(x)\big(vf - W\big) - c_x f + W\eta_b\widetilde{\mathfrak{d}}'(x)\Big(1 - \frac{r}{v}\Big) = 0.$$

Dividing by $L'f < 0$ and writing $\tilde{v} = v - c_x/L'$, $\rho_W = W/f$ and $\Theta_L = \widetilde{\mathfrak{d}}'/L'$ gives (41). For the limit, $L = 1 - x + \frac{1}{2}\mathbb{E}[\varepsilon^2]x^2 + o(x^2)$ gives $\widetilde{\mathfrak{d}} = 1 - \frac{1}{2}\mathbb{E}[\varepsilon^2]x + o(x)$ and $L'(0) = -1$, so $\Theta_L(0) = \frac{1}{2}\mathbb{E}[\varepsilon^2]$; the corner condition is $\mathcal{H}_x \leq 0$ at $x = 0$. Finally $\Theta_L = (M_L - 1)/(1 - L)$ requires $L''L = (L')^2$, i.e. $(\log L)'' = 0$, i.e. $L = e^{-cx}$. ■

Proposition 3. We prove the regional decomposition and the two free boundaries; the monotonicity of $h(\cdot)$ is Lemma 5 under its hypothesis $\Lambda < 1$ and is not re-derived here, and the monotonicity of x on the interior regions follows from (41) because its right-hand side $1 - \tilde{v}/\rho_W$ is decreasing in v while Θ_L is increasing, so the x solving it falls.

(i) If $r \equiv 0$ on an interval then $r' = 0$, so $\mathcal{A} = 0$, i.e. $\eta_b M_L(x) = 1$ and $x = \bar{x}$; conversely $x > \bar{x}$ gives $\mathcal{A} < 0$, so r would fall below zero, and $x < \bar{x}$ gives $\mathcal{A} > 0$, so rent is conceded. Delivered reliability is $L(\bar{x})$, which is $\Psi_L^{-1}(\eta_b)$ by (13). (ii) and (iv) are (41) with $c_x = 0$ and $c_x < 0$ respectively. (iii) c in (38) is continuous with $c_x = \mu[-\psi + (\lambda b - x)\psi']$ for $x < \lambda b$ and $c_x = 0$ above, so $\mathcal{H}_x$ jumps down by $\mu\psi(\lambda b)f$ as x crosses λb upward; wherever $\mathcal{H}_x > 0$ just below and < 0 just above, $x = \lambda b$ is optimal, an interval of types by continuity. (v) $x = 0$ on an interval makes all contracts there equally reliable, so (10) forces one rate, and continuity of r at v_1 gives (42). For the boundaries: with the initial point free and the initial state fixed, transversality is $\mathcal{H}(v_2) = 0$, i.e. $\big[L(x_2)v_2 - c(v_2, x_2)\big]f(v_2) = W(v_2)\mathcal{A}(v_2)$; on region (i) $\mathcal{A}(v_2) = 0$, so this reduces to $L(\bar{x})v_2 = c(v_2, \bar{x})$, which is (43), and if the bracket is positive at $\underline{v}$ the corner $v_2 = \underline{v}$ obtains. For v^*, note first that $r(v^*) = 0$ by continuity of the rent, since $r \equiv 0$ on region (i). If x is continuous at v^* then the multiplier on $r \geq 0$ vanishes there and the control is interior on both sides, so (41) holds at $\big(v^*, \bar{x}, r = 0\big)$; substituting those values—and retaining $\tilde{v}$, since c_x need not vanish at $\bar{x}$—gives (44), and under $L = e^{-x}$, $\Theta_L(\bar{x}) = (1/\eta_b - 1)/(1 - \bar{q})$ gives the stated right-hand side. If instead x jumps, then by Proposition 3 the control is the pointwise maximiser of $\mathcal{H}$ subject to $x \leq \bar{x}$, and $\mathcal{H}$ is continuous in v for each fixed x. Let $x^- = \bar{x}$ and x^+ be the limits from either side. For $v < v^*$

the maximiser is x^- and for $v > v^*$ it is x^+, so $\mathcal{H}(v, x^-) \geq \mathcal{H}(v, x^+)$ on the left and $\leq$ on the right; letting $v \to v^*$ from each side and using continuity gives equality at v^*, which is (45). This is the Weierstrass–Erdmann corner condition for the problem, and it is what (41) cannot supply, being a marginal condition at a point where the maximiser is not interior. ■

Lemma 5. Write $\Psi(z, v) = U(h(z), w(z); v)$, $\ell := L'(x) < 0$ and $\ell' := L''(x) > 0$, where $L''(x) = \mathbb{E}[\varepsilon^2 e^{-x\varepsilon}] > 0$. Incentive compatibility gives $\Psi_z(v, v) = 0$ for all v; differentiating totally, $\Psi_{zz} + \Psi_{zv} = 0$, and the second-order condition $\Psi_{zz} \leq 0$ forces $\Psi_{zv} \geq 0$. Since $\Psi_z = U_h h' + U_w w'$, the first-order condition gives $w' = -(U_h/U_w)h'$ and hence

$$\Psi_{zv} = h' \, \frac{U_{hv}U_w - U_h U_{wv}}{U_w}, \qquad U_w = -\tau < 0,$$

so with $D := U_{hv}U_w - U_h U_{wv}$ the requirement $\Psi_{zv} \geq 0$ is $h'D \leq 0$: if $D > 0$ then $h' \leq 0$, and if $D < 0$ then $h' \geq 0$.

It remains to sign D. From (7), $U_w = -\tau$ and $U_{wv} = -\partial\tau/\partial v = \ell b'$. Next $U_h = \ell b v + w\big(\ell b + \tau\big)/h$, using $\partial\tau/\partial h = [-\ell b h - (1 - L)]/h^2$ and $1 - L = h\tau$. Differentiating U_h in v and using $\partial x/\partial v = h b'$,

$$U_{hv} = \ell' b b' (h v + w) + \ell\big(b' v + b\big),$$

the two terms $\ell b' h$ and $-\ell h b'$ arising from the bracket cancelling exactly. Therefore

$$D = -\tau\Big[\ell' b b' (h v + w) + \ell(b' v + b)\Big] - \ell b'\Big[\ell b v + w(\ell b + \tau)/h\Big].$$

Substituting $b' = \eta_b b/v$, $\sigma = w/(hv)$, $x = hb$ and $h\tau = 1 - L$, and multiplying by $h/b > 0$,

$$\frac{h}{b}\, D = -\eta_b(1 + \sigma)\, x\Big[(1 - L)L'' + (L')^2\Big] - (1 - L)L'\big(1 + \eta_b + \eta_b\sigma\big). \tag{61}$$

The second term of (61) is strictly positive because $L' < 0$; the first is strictly negative because $L'' > 0$. Dividing (61) by $-(1-L)L' > 0$ and using $(L')^2/\big[-(1-L)L'\big] = \widetilde{\mathfrak{d}}(x)/x$ and $(1-L)L''/\big[-(1-L)L'\big] = \mathfrak{c}(x)/x$ gives

$$\frac{h\, D}{-b(1 - L)L'} = 1 + \eta_b(1 + \sigma) - \eta_b(1 + \sigma)\big[\mathfrak{c}(x) + \widetilde{\mathfrak{d}}(x)\big] = 1 - \Lambda(v),$$

so $D > 0$ if and only if $\Lambda < 1$, which proves (i) and (iii). For deterministic duration $L = e^{-x}$ gives $\mathfrak{c} = x$ and $\widetilde{\mathfrak{d}} = \mathfrak{d}(x)$, and $\mathfrak{d} + x = \mathfrak{M}$ by (5).

For (ii), let $G(z, s) := \partial U(h(z), w(z); s)/\partial s$. Using $r(s) = \Psi(s, s)$ and the envelope condition $r'(s) = G(s, s)$ of Lemma 3,

$$\Psi(v, v) - \Psi(v', v) = \big[\Psi(v, v) - \Psi(v', v')\big] - \big[\Psi(v', v) - \Psi(v', v')\big] = \int_{v'}^{v} \Big[G(s, s) - G(v', s)\Big] ds$$

for all v, v', both when $v > v'$ and when $v < v'$. If G is non-decreasing in its first argument the

integrand has the sign of $v - v'$ on the interval of integration in either case, so the integral is non-negative and no type gains by misreporting. And $\partial G/\partial z = \Psi_{vz} = h'D/U_w \geq 0$ exactly when $h'D \leq 0$, which under $\Lambda < 1$ is $h' \leq 0$.

For (iv), $\mathcal{H}_{xv} \leq 0$ is decreasing differences on a lattice, so by Topkis the argmax correspondence is non-increasing in v and every selection from it is non-increasing; strictness where $\mathcal{H}_{xv} < 0$ and the maximiser is interior is the strict form. Since b is increasing by Assumption 1, $h = x/b$ falls at least as fast. For the sufficient condition, where the maximiser is interior it solves (41), whose left side rises in x when Θ_L does and whose right side falls in v when ρ_W does, so the solving x falls with v. ■

Remark. Lemma 5 restricts the posted hazard, not delivered exposure: $x = h(v)b(v)$ is a product of a non-increasing and an increasing function and need not be monotone, and indeed is not on region (iii) of Proposition 3, where $h = \lambda$ is flat and $x = \lambda b(v)$ rises. Since $\alpha(\cdot)$ is decreasing it satisfies the lemma's requirement, so efficient reallocation is not excluded by incentive compatibility.

Derivation of (47). By (34), $w = \big[L(x)v - r\big]x/\big[b(1-L)\big]$, and $x = hb$ gives $\sigma = w/(hv) = \big[L(x)v - r\big]/\big[(1-L)v\big] = \big[L - r/v\big]/(1-L)$, the first identity in (47). Hence $1+\sigma = \big[(1-L)+L-r/v\big]/(1-L) = \big(1-r/v\big)/(1-L)$, and substituting into (46) with $\mathfrak{c} + \widetilde{\mathfrak{d}} = \mathfrak{M}$ under deterministic duration,

$$\Lambda = \eta_b \,\frac{1 - r/v}{1-L}\big(\mathfrak{M}(x) - 1\big) = \eta_b\big(1 - \tfrac{r}{v}\big)\frac{x - 1 + L}{(1-L)^2} = \eta_b\,\Theta_L(x)\big(1 - \tfrac{r}{v}\big),$$

using $\mathfrak{M} - 1 = (x - 1 + L)/(1-L)$ and, from (40), $\Theta_L = (x - 1 + L)/(1-L)^2$ under deterministic duration. That is the second identity.

(i) At $x = 0$, $\Theta_L(0) = \frac{1}{2}$ is finite and $1 - r/v$ bounded, but Λ is defined through $\mathfrak{c} + \widetilde{\mathfrak{d}} - 1$, which vanishes at the origin; directly, $\mathfrak{M}(0) = 1$ gives $\Lambda = 0$.

(ii) On the frontier region $r \equiv 0$ and $\eta_b M_L(\bar{x}) = 1$ by Proposition 3(i), so $\mathfrak{M}(\bar{x}) - 1 = 1/\eta_b - 1$ and $1 + \sigma = 1/(1-\bar{q})$, whence $\Lambda = \eta_b(1/\eta_b - 1)/(1-\bar{q}) = (1-\eta_b)/(1-\bar{q})$. This is below one iff $\bar{q} < \eta_b$. Now $M_L = \mathfrak{M}$ gives $\eta_b = (1-\bar{q})/\bar{x}$ with $\bar{x} = -\log\bar{q}$, so $\bar{q} < \eta_b$ is $-\bar{q}\log\bar{q} < 1 - \bar{q}$, i.e. $g(\bar{q}) > 0$ for $g(q) := 1 - q + q\log q$. But $g(1) = 0$ and $g'(q) = \log q < 0$ on $(0,1)$, so g is strictly decreasing to $g(1) = 0$ and therefore strictly positive on $(0,1)$. Hence $\bar{q} < \eta_b$ always and $\Lambda < 1$ unconditionally on this region.

(iii) Where (41) holds, its left-hand side is $\eta_b\Theta_L(x)(1 - r/v) = \Lambda$ and its right-hand side is $1 - \tilde{v}/\rho_W$, giving equality. On the bunch region (iii) of Proposition 3 the control sits at the kink with $\mathcal{H}_x > 0$ immediately below it, which after division by $L'f < 0$ reverses to $\eta_b\Theta_L(x)(1 - r/v) < 1 - \tilde{v}/\rho_W$, so the inequality is strict there. In either case $\Lambda < 1$ iff $\tilde{v} > 0$, and $\tilde{v} = v - c_x/L' > 0$ iff $L'v < c_x$ because $L' < 0$. If $\kappa = 0$ then $c \equiv 0$, so $c_x \equiv 0$, $\tilde{v} = v > 0$ and $\Lambda < 1$ throughout. ■

Proposition 4. Write $g_j(\kappa) := \Pi_j^*(\kappa) - \kappa/\lambda$ for $j \in \{H, 2, 3\}$. By the argument establishing (51) each g_j is continuous and strictly decreasing with a unique root $\kappa_j^* = \rho_j^*$. Because $\mathcal{M}_H \subset \mathcal{M}_2 \subset \mathcal{M}_3$, taking suprema over larger sets gives $\Pi_H^*(\kappa) \leq \Pi_2^*(\kappa) \leq \Pi_3^*(\kappa)$ for every κ, hence $g_H \leq g_2 \leq g_3$ pointwise. Evaluating at κ_H^*, $g_2(\kappa_H^*) \geq g_H(\kappa_H^*) = 0$, and since g_2 is strictly decreasing with root κ_2^* this forces $\kappa_2^* \geq \kappa_H^*$. The same argument at κ_2^* gives $\kappa_3^* \geq \kappa_2^*$.

For strictness of the second inequality, at $\kappa = \kappa_2^*$ the two-contract optimum delivers a constant hazard, whereas the maximiser of $\Pi(\cdot; \kappa_2^*)$ over $\mathcal{M}_3$ solves (41) pointwise; if that solution is non-constant the two differ on a set of positive measure and $\Pi_3^*(\kappa_2^*) > \Pi_2^*(\kappa_2^*)$, so $g_3(\kappa_2^*) > 0$ and $\kappa_3^* > \kappa_2^*$. For the first, at $\kappa = \kappa_H^*$ add to the optimal single rate a second contract with hazard $h = \alpha(v_0)$ and rate $R_h(v_0) - \epsilon$: types in an interval below v_0, previously unserved, are served at a rate exceeding κ_H^* for small ϵ, a first-order gain, while the mass diverted from above v_0 is $O(\epsilon)$ and each diversion costs $O(\epsilon)$. Hence $g_2(\kappa_H^*) > 0$ and $\kappa_2^* > \kappa_H^*$. ■

The three comparisons of Section 6.4. *Shape* follows from (41) and (23). *Sign*: $h = 0$ on the guaranteed region while $\alpha(v) > 0$ for $v < \overline{v}$, giving the first claim; and $h \geq \lambda \geq \alpha(v)$ wherever the manufactured set is non-empty, with strict inequality for $v > \underline{v}$ since $\alpha(v) < \lambda$ there, giving the second. By Lemma 5 h is non-increasing, so $\{h \geq \lambda\}$ is an interval containing v_2; it is $\{v : x(v) \geq \lambda b(v)\}$, and since $x \leq \bar{x}$ with $\bar{x}$ determined by η_b alone through $M_L(\bar{x}) = 1/\eta_b$, it is non-empty iff $\bar{x} \geq \lambda b(\underline{v})$, which is (57). We do not characterise the number of crossings of h and α; the statements use only the two endpoint comparisons. *Extent*: on $[v_2, v^*]$ rent is identically zero, which is impossible under a two-element menu except at a single type, since there $r = \max\{v - w_H b,\, Lv - w_L \tau,\, 0\}$ and both non-trivial branches are strictly monotone. ■

The capped-contract boundary of Section 6.6. With duration deterministic a job of type v completes under (T, w_C) if and only if $b(v) \leq T$, i.e. $v \leq b^{-1}(T)$ since b is strictly increasing. For such a type the bill is $w_C b(v)$ and the payoff $v - w_C b(v)$, which is non-negative exactly when $R_0(v) \geq w_C$, i.e. $v \geq R_0^{-1}(w_C)$, and exceeds the guaranteed payoff $v - w_H b(v)$ exactly when $w_C < w_H$; so the capped contract is bought on $[R_0^{-1}(w_C), b^{-1}(T)]$, non-empty iff $w_C < R_0(b^{-1}(T))$, and only if $w_C < w_H$, which is (58). For $v > b^{-1}(T)$ the capped contract delivers nothing and costs $w_C T > 0$, so it is strictly worse than not buying; such a type takes guaranteed service iff her payoff there is non-negative, $R_0(v) \geq w_H$, giving the set $[\max\{b^{-1}(T), R_0^{-1}(w_H)\}, \overline{v}]$. The two sets are adjacent iff $R_0^{-1}(w_H) \leq b^{-1}(T)$; otherwise the types between buy nothing. Note the participation constraint is what makes this last proviso necessary: it is not enough that a long job be denied cheap certainty, it must also be willing to pay w_H. ■

Minimum commitments are dominated. A contract $P(q) = w_R \max\{q, T\}$ charges $w_R T$ on any job with realized time $q < T$ and $w_R q$ otherwise, so it coincides with the flat rate w_R on jobs longer than T and charges strictly more on shorter ones. Under Assumption 1 the types with $b(v) < T$ are the low-value types, whose participation constraint binds first, so the extra charge falls on buyers who are marginal or excluded. Replacing P by the flat rate w_R therefore weakly increases the set of participating types at unchanged payments from every type that participated before, hence weakly raises profit. ■

Appendix B: The boundaries of Assumption 1

Assumption 1 restricts the elasticity of required runtime to $(0, 1)$. This appendix shows that interruption has no screening role on either side of that interval, so the assumption marks the

boundaries of the phenomenon rather than selecting a convenient case within it.

Proposition B.1 (No excess component outside $(0,1)$)**.**

(i) (Runtime does not rise with value.) *Suppose $b'(v) \le 0$ on $[\underline{v}, \overline{v}]$. Then for every $h > 0$ the maximum acceptable rate R_h is strictly increasing, so no rate both leaves every guaranteed buyer on her contract and attracts a previously excluded buyer. The seller reclaims only to reallocate: $h^* = \alpha$.*

(ii) (Runtime rises more than proportionally.) *Suppose b is strictly increasing and convex with $\eta_b(v) \ge 1$ on $[\underline{v}, \overline{v}]$, and that the shock is bounded by $\varepsilon \le \min_v \eta_b(v)$ almost surely. Then the menu*

$$h(v) = 0, \qquad w(v) = \frac{1}{b'(v)}, \qquad \phi(v) = v - \frac{b(v)}{b'(v)} \;\ge\; 0$$

is incentive compatible, charges every type exactly her completion value, and uses no interruption. No destructive instrument can improve on it.

Proof. (i) If $b' \le 0$ then $\eta_b(v) = vb'(v)/b(v) \le 0$, and since $M_L > 0$ we have $\eta_b(v) M_L(hb(v)) \le 0 < 1$ for every v and every $h > 0$. By (11), $d\log R_h / d\log v = 1 - \eta_b M_L > 0$, so R_h is strictly increasing and $R_h(v) < R_h(v_0)$ for every $v < v_0$. By (8) the incumbent floor is $R_h(v_0)$, so any rate that protects the guaranteed tier attracts nobody and any rate that attracts an excluded buyer is also taken by the cutoff type. No admissible tier exists at any hazard, and by Corollary 1 raising h above α then has a strictly negative payoff, since it shortens billed time and destroys completion value while forgoing no request.

(ii) Write $q = b(v)$ and $V(q) = b^{-1}(q)$, which is well defined by strict monotonicity and concave because b is convex. The tangent to V at $q = b(v)$ has slope $1/b'(v)$ and intercept $v - b(v)/b'(v)$, which is non-negative precisely because $\eta_b(v) \ge 1$ means $vb'(v) \ge b(v)$. Concavity places every tangent weakly above V, so a buyer choosing her own tangent pays exactly her value while any other tangent charges weakly more; truthful selection is therefore incentive compatible and every type is served at zero rent. Duration uncertainty does not disturb this, because payment is linear in realized time and $\mathbb{E}[\varepsilon] = 1$, so expected payment on contract j is $\phi_j + b(v)/b'(v_j)$ exactly as under deterministic duration. It does, however, create a continuation problem: at elapsed time t a buyer completes only if $v \ge w(v)\big(b(v)\varepsilon - t\big)$, which at $t = 0$ requires $\varepsilon \le \eta_b(v)$. The bound on the shock rules out abandonment, without which a buyer would participate only when the draw is favourable and retain rent. Since payments already equal completion values, no mechanism can raise revenue, and any positive hazard would strictly reduce the value delivered. ■

The argument covers the capped contract of Section 6.6 as well as the hazard, since a deadline sorts on duration alone and $b' \le 0$ makes duration decreasing in value, so it too attracts only types the guaranteed rate already serves.

Part (i) concerns a population that is reversed in its entirety, which is why it yields so stark a conclusion. It is worth distinguishing from a population containing a reversed segment alongside an ordinary one, where the two are consistent and the contrast is instructive: a reversed segment

on its own gives the seller no reason to interrupt, since a posted rate already sorts it, and what would generate an excess component is the *combination*, in which no instrument conditioning on duration alone can separate the two segments.

Part (ii) is the two-part tariff of Oi (1971) applied type by type, and super-proportionality is exactly what makes the fixed component collectable. When $0 < \eta_b < 1$ the same construction requires $\phi(v) = v\big(1 - 1/\eta_b\big) < 0$, which a seller cannot implement: a negative entry fee would be collected by a buyer who contracts and releases the machine immediately, so it is unenforceable rather than merely unobserved. Metering then leaves a coverage problem, and the question of how to reach the excluded buyers is the subject of the paper.

Appendix C: Numerical implementation

The single-rate benchmark is solved exactly: (19) is a scalar equation and (21) and the quadratic of Proposition 1(vi) are closed forms. Note that scanning a grid of rates carries an $O(dv)$ bias from integrating a masked integrand across the cutoff and returns $w = 0.6479$ where the exact value is 0.645685; the scalar equation should be solved instead.

The demand-driven optimum of Section 5 requires two nested fixed points. The inner one is in the allocation, since $h = \lambda(1 - F(v_1))$ depends on the cutoff it induces; we parametrise by the cutoffs, so that h follows from v_1 and the rates from (27), and optimise over (v_1, v_2) by a coarse grid followed by Nelder–Mead refinement to 10^{-11}. The outer one is in the machine-hour price: we iterate $\kappa \mapsto \lambda\Pi^*(\kappa)$ from $\kappa = 0$, which converges monotonically to the unique root of (51) because Π^* is non-increasing, and we check the root against a bisection on $\Pi^*(\kappa) - \kappa/\lambda$. Two checks on the inner problem: the degenerate menu $w_L = w_H$ must reproduce the single-rate optimum exactly, which is the nesting bound $\rho_2^* \geq \rho_H^*$, and the hazard map's fixed points are enumerated on a grid before the greatest is selected, so that the reported optimum is not an unstable interior one. The stationary implementation with occupancy fractions and an endogenous admission ratio, which the earlier version of this section used, is now confined to E.8, where (86) solves (85) in closed form so that no iteration on occupancy is needed at all.

The schedule of Section 6 is solved as a forward–backward sweep. The rent is integrated forward from $r(v_2) = 0$ using the integrating factor of (35); the weight W of Lemma 4 is integrated backward; and the control is obtained by *global* pointwise maximisation of $\mathcal{H}$ over the candidate set $\{0,\ \lambda b(v),\ \bar{x}(v)\}$ together with the roots of (41) on each side of the kink, each located to machine precision. Damping is applied to (r, W) rather than to x, because the argmax map is discontinuous at the kink while its integrals are not; damping the control instead does not converge. Residuals reach 10^{-11} in (r, W) within a few hundred sweeps at $N = 800$ nodes.

Three checks were run. The costate representation $p = -W$ was verified against a finite-difference perturbation of the exposure schedule, agreeing to four decimals relative. The exposure condition (41) was verified against a direct numerical maximisation of the relaxed problem that uses no first-order condition: the optimal exposure agrees to four significant figures. And incentive

compatibility was checked on a 500×500 grid of type–contract pairs, where the largest gain from misreporting is 5×10^{-8} and the recovered indirect utility matches $r(v)$ to 10^{-16}; the solution has x and h non-increasing and w non-decreasing at every computed optimum, and the induced allocation has exactly the number of boundaries the region structure implies, with no islands, across five parameterisations and menus of two to five contracts.

Two cautions apply. The profit surface is very flat near the optimum in the exposure direction, so region boundaries are better determined than the schedule's final digits, and reported cutoffs should be read to three decimals rather than five. The monotonicity condition $\Lambda < 1$ of Lemma 5 is now derived in (47) rather than checked; the computations confirm the identities there, with Λ running from 0.240 to 0.6275 and matching $\eta_b \Theta_L(x)(1 - r/v)$ to 8×10^{-15} and $(1 - \eta_b)/(1 - \bar{q})$ on the frontier at $\eta_b = 1/2$ and $\eta_b = 0.8$; the closed form (61) for the cross-partial determinant was verified against finite differences to 10^{-4} relative under deterministic and gamma shocks, and the equivalence $D > 0 \iff \Lambda < 1$ was confirmed at parameter values on both sides of the threshold, including $\Lambda = 3.32$ and 9.26 where D turns negative. The manufacturing condition (57) was checked across $\eta_b \in \{0.3, 0.5, 0.7, 0.9\}$ and $\lambda \in [0.5, 8]$: the manufactured and bunch regions are present exactly where $\lambda \leq \bar{x}/b(\underline{v})$ and absent otherwise, and where they are absent the schedule reduces to a frontier region and a guaranteed region. At $\eta_b = 0.3$ the threshold is 7.84, and the regions are present at $\lambda = 4$ and gone at $\lambda = 8$. We did not characterise the number of crossings of h and α and nothing in the paper uses it.

The coarseness evidence should be read with (54) in front of it. The theorem gives the rate at which refinement pays less; the computations give the constant, and only for one parameterisation—uniform values, power runtime, deterministic duration, $\mu = 0.1$, two arrival rates. We have not varied the value distribution, the runtime elasticity, the duration distribution, or the salvage and checkpointing specifications of Appendix D. The claim is therefore that coarseness is consistent with profit maximisation at these primitives, not that it is optimal at all of them.

The finite-menu comparisons of Tables 4 and 2 require care, because a derivative-free optimiser over $2n$ parameters finds local optima from random starts: an earlier version of this exercise failed to beat the single-rate menu with two contracts at $\lambda = 3$. We therefore build starting menus by discretising the continuum schedule—sampling n types across the served range and reading (h, w) off $x(\cdot)$ and (34)—at several spreads, add the nested winner from $n - 1$ and the perturbed guaranteed-only solution, run differential evolution with a Sobol initialisation alongside, and polish the best repeatedly. Two checks the earlier attempt failed then pass: the sequence is monotone in n, and every $\Pi_n \leq \Pi_\infty$, as it must be since a finite menu is an admissible schedule. A grid-plus-refine version of the same exercise under-optimised the cells of Table 2 by up to nine percent in h^* while leaving the profit ratios accurate to three decimals, which is the general pattern: the objective is flat in the hazard near its optimum, so profits are far better determined than the instruments that deliver them, and reported hazards should be read to two decimals.

A systematic audit of every numbered result, re-deriving rather than re-reading, turned up two errors that had survived several passes and are worth recording as a caution about the class of

mistake to look for. The capped-contract boundary originally omitted the participation constraint, asserting that guaranteed service is taken by $\{v : b(v) > T\}$ when the correct set intersects that with $\{R_0(v) \geq w_H\}$; the error is invisible unless one checks a parameterisation with $w_H > R_0(b^{-1}(T))$, where a band of types buys nothing. And (44) originally omitted c_x, which vanishes at $\lambda = 1$ and $\lambda = 2$ in the base calibration and does not at $\lambda = 3$, where the cost-free form misplaces v^* by a factor of 1.6. Both were found by testing identities at parameter values away from the base calibration, which we recommend as the first check on any of the paper's stated conditions.

Section 5 is stated and computed in the unified environment of Section 6.4: per arrival, with the machine-hour price κ^* endogenous by (51) and no admission priority. An earlier version paired a per-buyer objective with a stationary table computed under admission priority, and the two are not commensurable; the present formulation removes the discrepancy by putting all three arrangements in one environment, and the priority case is treated separately in E.8. The quantities reported in Section 5 are on that basis: the margins satisfying (32); the non-collapse condition, $w_H = 0.6923$ against a threshold of 0.9889 at $\lambda = 1$; and the two sides of the fourth comparison, 0.0245 against 0.0227. Because κ^* is endogenous, the levels in Table 1 are not comparable with those of a fixed-κ computation such as Table 3, whose three rows are commensurable with each other and are computed at a common fixed price; the ranking of Proposition 4 is unaffected by which is used.

One further caution concerns the guaranteed-only benchmark $\max_w (w - \kappa)M(w)$. Since M vanishes identically for $w \geq R_0(\overline{v})$, the objective is flat at zero on an interval and a bounded scalar optimiser can stall there and return a spurious near-zero maximum; at $\lambda = 6$ this produced a benchmark three orders of magnitude too small before we caught it. We evaluate that maximum on a grid.

Appendix D: Interruption short of total loss

The baseline assumes a reclaimed job yields nothing. This appendix distinguishes three ways in which interruption may be costly short of total loss, and shows that they have very different consequences. Throughout, write $x = hb(v)$ and recall from (8) that a buyer accepts a rate w exactly when $w \leq hv\,\Phi(x)$, where $\Phi := \mathcal{Q}/\mathcal{B}$ with $\mathcal{Q}(x)$ the expected fraction of v delivered and $\mathcal{B}(x)/h$ the expected bill in units of time. Since Φ enters willingness to pay only through this ratio, two specifications with the same Φ are indistinguishable for every result in the paper.

Restart: every result is unchanged. Suppose a reclaimed buyer begins again, attempts are independent, and delay is costless, so that she completes with probability one but pays for the usage each failed attempt consumed. Expected billed time until completion is $\tau(h, v)/L(x)$, so

$$\mathcal{Q}(x) = 1, \qquad \mathcal{B}(x) = h \cdot \frac{\tau(h, v)}{L(x)} = \frac{1 - L(x)}{L(x)}, \qquad \Phi(x) = \frac{L(x)}{1 - L(x)},$$

which is exactly the baseline Φ obtained from $\mathcal{Q} = L$ and $\mathcal{B} = 1 - L$. The slope identity (11), the exposure elasticity M_L, the ceiling (12), the frontier (13), the participation condition (8) and every proposition therefore hold verbatim. Under deterministic duration the two coincide term by term, $\Phi(x) = 1/(e^x - 1)$.

The economics is that restart multiplies both the value delivered and the bill by the same factor, $1/L(x)$, leaving the ratio untouched. All-or-nothing completion was never the substantive assumption: the substantive assumption is that the buyer pays for unproductive usage, which is why Section 8 records the billing rule and not the completion technology as the binding restriction. One accounting consequence does change. Completing a job consumes

$$\frac{\tau^r(h,v)}{b(v)} = \frac{1 - L(x)}{x\,L(x)} = \frac{1}{\mathfrak{d}(x)}$$

times its required resources, so with a positive occupancy cost the welfare comparison of Corollary 1 must be recomputed even though the pricing analysis need not.

Salvage: the screen has a threshold. Suppose instead that a reclaimed buyer retains a fraction $\sigma \in [0, 1)$ of the completion value without redoing anything. Then

$$\mathcal{Q}_\sigma(x) = \sigma + (1 - \sigma)L(x), \qquad \mathcal{B}(x) = 1 - L(x),$$

and, differentiating $\Phi_\sigma = \mathcal{Q}_\sigma/\mathcal{B}$ logarithmically, the slope identity becomes

$$\operatorname{sign} R'(v) = \operatorname{sign}\big[1 - \eta_b(v)\, M_\sigma\big(hb(v)\big)\big], \qquad M_\sigma(x) := \frac{x\big(-L'(x)\big)}{\big(1 - L(x)\big)\big[\sigma + (1 - \sigma)L(x)\big]}, \tag{62}$$

which reduces to M_L at $\sigma = 0$. Salvage is the one perturbation considered here that makes the exposure ceiling *finite* for every positive σ: under deterministic duration $M_\sigma(x) = xe^{-x}/\big[(1 - e^{-x})(\sigma + (1 - \sigma)e^{-x})\big] \to 0$ as $x \to \infty$, so more interruption ceases to be a substitute for a weak duration–value relationship. Consequently a screening hazard exists if and only if

$$\eta_0 \sup_{x>0} M_\sigma(x) \;>\; 1, \tag{63}$$

and since the supremum decreases in σ there is a critical salvage fraction $\sigma^*(\eta_0)$ above which no hazard reverses the ordering at the cutoff. For constant-elasticity duration and deterministic runtimes, σ^* equals 0.032, 0.061, 0.102, 0.161 and 0.251 at $\eta_0 = 0.5, 0.6, 0.7, 0.8$ and 0.9. Salvage of a quarter of the value is therefore compatible with screening only where required resources are highly informative about value.

The contrast with restart is instructive. Restart lowers willingness to pay per billed unit for everyone, in proportion to exposure, which is what preserves the cross-type ordering. Salvage raises it, and raises it most for the types with the largest exposure, which is what flattens the ordering and eventually reverses the mechanism.

Preserved progress: smooth attenuation. Between the two lies interruption that destroys only the work since the last saved stage. With n equal stages, each saved on completion and each restarted independently, expected billed time to finish all n is $n\big(L(x/n)^{-1}-1\big)/h$, so Φ becomes the restart expression evaluated at x/n and the multiplier is $M_L(x/n)$. The exposure required for a reversal therefore scales linearly in the number of saved stages, and as $n\to\infty$ the multiplier tends to its value at the origin, which is one, so sufficiently fine checkpointing removes the screen. Unlike salvage this is a smooth degradation with no threshold, and unlike restart it is not neutral.

Appendix E: Multi-unit demand and pools

The body of the paper has one machine and one machine per job. Neither is innocuous, and they interact: a provider has many machines, and buyers want several at once. This appendix takes both, in that order, because the multi-unit apparatus of E.1–E.3, and in particular Proposition E.3, is what the pool analysis of E.4 draws on. The order of the argument is: multi-unit demand changes nothing (E.1); pools destroy the demand-driven mechanism (E.4); the two together do not restore it either, though they change who bears the risk and reproduce the boundary of Section 6.6 from capacity structure (E.5); time-varying demand, the other candidate, does not (E.7); E.6 lets value and runtime be separately private; E.8 works out what admission rationing adds in a single facility; and E.9 closes the gap between the pool and the chosen schedule of Section 6, with the conclusion that wherever an interruptible tier is optimal in a fluid pool its hazard is manufactured in full.

E.1 Footprint invariance

A job needs $m(v)\geq 1$ machines simultaneously for a span $b(v)\varepsilon$ and is lost if any is reclaimed before it finishes. Let m be increasing and continuously differentiable with $\eta_m(v):=vm'(v)/m(v)\geq 0$, and write

$$\varphi(v):=m(v)\,b(v),\qquad \eta_\varphi:=\eta_m+\eta_b.$$

Proposition E.1 (Footprint invariance)**.** *Suppose the m units are reclaimed independently, each at rate h. Then the job completes with probability $L\big(h\varphi(v)\big)$ and is billed for $\big(1-L(h\varphi(v))\big)/h$ machine-hours, so*

$$U(h,w;v)=L\big(h\varphi\big)v-w\,\frac{1-L(h\varphi)}{h},$$

which is (7) with b replaced by φ. Every result stated in per-buyer form—which after Section 6.4 is all of Sections 4–6 except part (vi) of Proposition 1—therefore holds with $b\mapsto\varphi$, $\eta_b\mapsto\eta_\varphi$ and $R_0\mapsto v/\varphi$; Assumption 1 becomes $\eta_m+\eta_b\in(0,1)$ and the reversal condition becomes $\eta_\varphi M_L(h\varphi)>1$. The stationary formulations—Proposition 1(vi) and the priority case of E.8—are written for a single machine and so presume $m\equiv 1$; their multi-machine counterparts are the subject of E.4 onward, where the balance is written in machines rather than in machine-hours.

Proof. The time to the first reclamation among m independent Poisson clocks of rate h is exponential with rate mh, so the job survives its span with probability $\mathbb{E}\big[e^{-mhb\varepsilon}\big]=L(hmb)=L(h\varphi)$. Billed

machine-hours are m times the elapsed time, and elapsed time is $\min\{T, b\varepsilon\}$ with $T \sim \text{Exp}(mh)$, so their expectation is $m \cdot \big(1 - L(h\varphi)\big)/(mh) = \big(1 - L(h\varphi)\big)/h$. Both expressions depend on (m, b) only through φ, and the guaranteed contract bills $w\, m\, b\, \mathbb{E}[\varepsilon] = w\varphi$. ■

Because the invariance is exact the solved models transfer rather than needing re-derivation. For $m(v) = v^{\eta_m}$, $b(v) = v^{\eta_b}$ with $\eta_\varphi < 1$ and $\kappa = 0$, Proposition 1(v) gives

$$v_0 = \overline{v}\Big(\frac{1 - \eta_\varphi}{2}\Big)^{\frac{1}{1+\eta_\varphi}}, \qquad w^* = v_0^{\,1-\eta_\varphi}, \qquad \Pi_H^* = \frac{w^* \overline{v}^{\,1+\eta_\varphi}}{2(\overline{v} - \underline{v})}, \tag{64}$$

and the schedule of Section 6 transfers with $\bar{x}$ solving $M_L(\bar{x}) = 1/\eta_\varphi$. Two checks: (64) matches direct numerical maximisation to five decimals, and $(\eta_m, \eta_b) = (0, \frac{1}{2})$ and $(\frac{1}{5}, \frac{3}{10})$—different in shape, equal in η_φ—return the identical optimum $w^* = 0.629961$, $v_0 = 0.396850$.

Two consequences. Assumption 1 is *tighter* in footprint form, since $\eta_m + \eta_b < 1$ can fail where $\eta_b < 1$ holds comfortably. But screening is *easier*, because the reversal condition asks only that the footprint rise with value. Size scales with value more reliably than runtime does, so the paper's central force does not require the empirically awkward claim that valuable jobs run longer.

E.2 Placement as an instrument

Proposition E.2 (Placement)**.** *Suppose the provider chooses how far to spread a buyer's units, so that reclamations arrive at effective rate $m^\theta h$ for $\theta \in [0, 1]$: $\theta = 1$ is independent placement across failure domains, $\theta = 0$ perfect concentration in one. Then completion occurs with probability $L(x)$ and billed machine-hours are $m^{1-\theta}\big(1 - L(x)\big)/h$ with*

$$x = h\, m(v)^\theta b(v), \tag{65}$$

willingness to pay is unchanged in form, $w \le R_0(v)\mathfrak{d}(x)$ with $R_0 = v/\varphi$, and

$$\frac{d \log R_h}{d \log v} = \big(1 - \eta_\varphi\big) + \big(1 - M_L(x)\big)\big(\theta\eta_m + \eta_b\big), \tag{66}$$

which reduces to $1 - \eta_\varphi M_L$ at $\theta = 1$.

Proof. With effective rate $m^\theta h$ the survival probability is $\mathbb{E}\big[e^{-m^\theta h b\varepsilon}\big] = L(x)$ for x as in (65), and billed machine-hours are $m\mathbb{E}[\min\{T, b\varepsilon\}]$ with $T \sim \text{Exp}(m^\theta h)$, which is $m\big(1 - L(x)\big)/(m^\theta h)$. Hence $R_h = vL/\tau = v\, m^{\theta-1} hL/(1 - L)$, and substituting $h = x/(m^\theta b)$ gives $R_h = \big(v/\varphi\big)\mathfrak{d}(x)$. Differentiating logarithmically, $d \log R_h / d \log v = (1 - \eta_\varphi) + \big(d \log \mathfrak{d}/d \log x\big) \cdot \big(d \log x/d \log v\big)$ with $d \log \mathfrak{d}/d \log x = 1 - M_L$ and $d \log x/d \log v = \theta\eta_m + \eta_b$. ■

Placement is a shape instrument and can be solved for. Since θ enters only through $x(v) = hm(v)^\theta b(v)$, with $\partial x/\partial\theta = x \log m$, the functional derivative of Section 6 gives

$$\frac{d\Pi}{d\theta} = \int_{v_2}^{\overline{v}} \mathcal{H}_x(v)\, x(v)\, \log m(v)\, dv, \qquad \mathcal{H}_x = \big(L'(x)v - c_x\big)f - W\big[\mathcal{A}_x + \mathcal{B}_x r\big], \tag{67}$$

so spreading is worthwhile exactly when the marginal value of exposure is weighted toward the larger footprints. Three implications. In the schedule problem of Section 6 placement is *redundant*, and for a reason simpler than (67): there the seller sets $h(v)$ contract by contract, so any pair $(h(\cdot), \theta)$ delivering the same exposure profile $x(\cdot)$ is the same mechanism, and θ is a reparametrisation. Placement matters only where a single hazard is posted. There, the sign is governed by regime rather than primitives: by (66) a reversal requires

$$\big(M_L(x) - 1\big)\big(\theta\eta_m + \eta_b\big) > 1 - \eta_\varphi, \tag{68}$$

whose left side increases in θ. Hence $\theta^* = 1$ whenever a reversal is required, which is the expansion regime: spreading is the cheapest way to obtain one. In the re-pricing regime no reversal is needed and the sign is that of (67), a covariance between the marginal value of exposure and $\log m$ which we do not sign in general; since $\mathcal{H}_x$ changes sign across the type space, interior optima are not excluded. And the magnitude is large: exposure at θ is $m^{1-\theta}$ times smaller than under independent placement, so the two extremes differ by the full factor m—eight for an eight-unit job—obtained by an allocation policy rather than a price. Providers publish allocation strategies that amount to choices of θ, which yields a prediction requiring no estimated primitive: discounts should be deeper where interruptible placement is diversified.

E.3 Partitionable work

Proposition E.3 (Partitionable work)**.** *Suppose losing one unit destroys only that unit's work, the remaining units continuing. Then the job is the preserved-progress case of Appendix D with $n = m$ stages, willingness to pay is*

$$w \;\le\; R_0(v)\,\mathfrak{d}\big(h\,b(v)\big), \qquad R_0 = v/\varphi, \tag{69}$$

so that its argument is exposure per unit*, $x/m = hb$, while its level is deflated by the whole footprint. Consequently*

$$\frac{d\log R_h}{d\log v} = \big(1 - \eta_\varphi\big) + \big(1 - M_L(hb)\big)\eta_b, \tag{70}$$

and a reversal requires

$$\eta_b\, M_L\big(h\,b(v)\big) \;>\; 1 - \eta_m, \tag{71}$$

which reduces to the condition of Proposition E.1 when m is constant. The contrast with gang-scheduled work is therefore sharper than a change of argument. For a gang-scheduled job the size elasticity enters multiplicatively*, $\eta_\varphi M_L(h\varphi) > 1$, so a larger footprint both raises the elasticity and raises the exposure a given hazard delivers; for partitionable work it enters* additively*, $\eta_b M_L(hb) + \eta_m > 1$, so size relaxes the threshold but no longer amplifies the instrument.*

Proof. The m units are the stages: each carries b of work, so $1/m$ of the footprint, and each is exposed at rate h for its own duration, so per-unit exposure is $hb = x/m$. A unit restarts until it

completes, so expected billed machine-hours are $m\big(1-L(hb)\big)/\big(hL(hb)\big)$ and the job completes with probability one; hence $\Phi = \mathcal{Q}/\mathcal{B} = L(hb)/\big[m(1-L(hb))\big]$ and $w \le hv\Phi = \big(v/\varphi\big)\mathfrak{d}(hb)$, which is (69). Differentiating logarithmically, $d\log R_0/d\log v = 1-\eta_\varphi$ and $d\log\mathfrak{d}/d\log v = \big(1-M_L(hb)\big)\eta_b$ since $d\log(hb)/d\log v = \eta_b$, giving (70); setting it negative and rearranging gives (71). ■

So for partitionable work what matters is exposure *per unit*, not per footprint: size still helps, but only through the level of willingness to pay and not through the exposure a hazard delivers. The screen therefore runs on whether a workload is gang-scheduled or partitionable rather than on its size, which is the sharpest workload-level prediction the paper offers—interruptible tiers taken by inference, batch and embarrassingly parallel work and avoided by gang-scheduled distributed training—and by Section 7.4 the discount is bounded by the cost of making the latter restartable.

E.4 Pools without fragmentation

Let the provider operate N identical machines with requests arriving at rate $N\lambda$, so λ is the arrival rate per machine and $N=1$ is Section 5. Write g, s and $i = 1-g-s$ for the fractions of machines holding guaranteed tenants, holding spot tenants and idle, and P_B for the stationary probability that an arrival finds no idle machine. Take $m \equiv 1$ in this subsection; E.5 restores $m(v)$; Lemma E.1 records what saturation implies for the two rates.

Displacements occur at rate $N\lambda\big(1-F(v_1)\big)P_B$ and land on sN spot tenants without regard to which, so a spot tenant faces

$$h = \frac{\lambda\big(1-F(v_1)\big)P_B}{s} = \frac{\alpha(v_1)\,P_B}{s}, \tag{72}$$

still a single number, memoryless and independent of the incumbent's type, so Lemma 2 and everything in Proposition 2 using only those properties carries over with (72) in place of (23).

So does the ranking of Proposition 4, and by the same argument. A seller who sets the rate may always choose to add nothing to what demand generates, so the demand-driven arrangement is admissible in her problem and $\rho_3^* \ge \rho_2^*$ wherever both are defined; E.9 makes that choice explicit and identifies the one case in which the inequality is not strict. What the pool changes is therefore not the direction of the comparison but the size of the gap, and it does so through the two corrections in (72), which pull in opposite directions. Displacement is *concentrated*, the requests of the whole pool falling on the discounted fraction alone, which multiplies the hazard by $1/s$; and it is *rationed*, since a guaranteed arrival preempts only when she cannot be given an idle machine, which multiplies it by P_B.

Proposition E.4 (The pool dichotomy)**.** *Let* $\omega := \lambda\big(1-F(v_1)\big)\mathbb{E}\big[b \mid v \ge v_1\big]$ *and* $\bar\tau(h) := \mathbb{E}\big[\tau(h,v) \mid v \in [v_2, v_1)\big]$*, and assume* $\omega < 1$*. In the fluid limit* $N \to \infty$ *at fixed* λ*: (i)* $g = \omega$*, the guaranteed tier occupying exactly its own offered load, being neither rationed nor displaced; (ii) if*

$$\lambda\big(F(v_1)-F(v_2)\big)\,\bar\tau\Big(\tfrac{\lambda(1-F(v_1))}{1-\omega}\Big) \;\ge\; 1-\omega \tag{73}$$

then $i = 0$, $s = 1 - \omega$, $P_B = 1$ and $h = \alpha(v_1)/(1 - \omega) > \alpha(v_1)$; (iii) otherwise $i > 0$, $P_B \to 0$, $h \to 0$ and the menu collapses by (10). For finite N the regimes join continuously, P_B being of Erlang-loss form and decaying geometrically in N at fixed load below one.

Proof. Work in the fluid limit. The occupancy process is a density-dependent Markov chain in the sense of Kurtz (1970), whose theorem gives convergence, uniformly on compact time intervals and almost surely as $N \to \infty$, of the scaled fractions to the solution of the corresponding system of ordinary differential equations; we evaluate that system at its rest point, which is what the balance equations below express. (i) Guaranteed tenants enter at rate $\lambda(1 - F(v_1))$ per machine—every such arrival is served, into an idle machine or by displacing—and leave only by completing, at rate $g/\mathbb{E}[b \mid v \geq v_1]$; equating gives $g = \omega$, which requires $\omega < 1$ and does not depend on h, s or the cheap tier. (ii)–(iii) If $i > 0$ a guaranteed arrival always finds an idle machine, so $P_B = 0$ and $h = 0$ by (72); spot tenants then leave only by completing, so $s = \lambda(F(v_1) - F(v_2))\bar{\tau}(0)$, and consistency $g + s < 1$ is the negation of (73) because $\bar{\tau}$ is decreasing in h. If $i = 0$ then $s = 1 - \omega$, every guaranteed arrival displaces, so $P_B = 1$ and (72) gives the stated h; stationarity requires the spot population be replenished as fast as depleted, $\lambda(F(v_1) - F(v_2)) \geq s/\bar{\tau}(h)$, which is (73). For finite N, P_B is the stationary all-busy probability of the corresponding loss system with preemptive priority, positive for every N and decaying geometrically at fixed load below one. ■

Lemma E.1 (Saturation requires a wide cheap tier)**.** *In the saturated regime $F(v_1) - F(v_2) \geq 1 - F(v_1)$: at least as many buyers take the cheap contract as take the expensive one.*

Proof. $\tau \leq 1/h$ gives $\lambda T \leq \lambda\big(F(v_1) - F(v_2)\big)/h$; saturation gives $\lambda T \geq s$; and $h = \lambda(1 - F(v_1))/s$. Combining and cancelling $s > 0$ gives the claim. ■

The economics is transparent. A saturated pool is full whatever the menu, so profit is the occupancy-weighted average of the posted rates and a cheap tier can only dilute it; a slack pool never displaces anyone, so the cheap tier is indistinguishable from the expensive one and the menu collapses. The mechanism of Section 5 requires idle capacity and blocking *simultaneously*, and with $m \equiv 1$ those are mutually exclusive in the fluid limit: one free machine is enough to admit any arrival. A single machine is the one case in which both hold, because with $N = 1$ the fluctuations do not average out.

Finally (72) disposes of an identification claim the single-machine model appeared to license. The comparison with efficiency becomes

$$\frac{h}{\alpha(v)} = \frac{P_B}{s} \cdot \frac{1 - F(v_1)}{1 - F(v)}, \tag{74}$$

so the under-exposure conclusion of Proposition 2(ii) holds if and only if $P_B \leq s$. In a tightly run pool $P_B = 1$ and interruption exceeds the efficient intensity throughout the band as soon as $s < \big(1 - F(v_1)\big)/\big(1 - F(v_2)\big)$, while every reclamation remains matched to an arrival and nothing is manufactured in the sense of (3): concentration by $1/s$ raises the per-tenant hazard above λ without any reclamation going unmatched, so dilution reads as manufacturing to anyone who infers it from

h alone. What survives is a test in three observables, all recorded by any scheduler,

$$\frac{h\,s}{P_B} = \lambda\big(1 - F(v_1)\big) \;\le\; \lambda. \tag{75}$$

E.5 Fragmentation, and why it does not save the mechanism

The impossibility in E.4 rests on $m \equiv 1$: with unit demand a single free machine admits any arrival, so idle capacity and blocking cannot coexist. Restore $m(v)$ and they can, because free capacity can be present but in the wrong shape.

Partition the N machines into $D = N/K$ *domains* of size K, fixed and small, within which a job's units must sit. A request for m units is admissible without displacement if and only if some domain holds m free machines. Let π_j be the fraction of domains with exactly j free machines and

$$\bar{m} := \max\{j : \pi_j > 0\}$$

the largest free block in the pool.

We make two assumptions explicit, because results (iii) and (iv) below depend on them and neither is a consequence of the structure. *Independent placement:* a tenant's units are located independently of one another, so that a seizure event and a tenant's holdings are independently placed. *Uniform targeting:* a displacing arrival assembles her slots from spot-occupied slots drawn uniformly, rather than from the domain where they are cheapest to assemble. Both are substantive. Under correlated placement a tenant's units sit together, so a single seizure takes all of them and, since one loss suffices to destroy the job, her hazard is *lower*—the $\theta = 0$ case of Proposition E.2. Under optimised targeting the provider would evict where disruption is least, which is a further instrument of the kind Proposition E.2 studies. So (iii) below should be read as the independent-placement case, and its footprint proportionality as an upper bound on the per-tenant hazard rather than as something the structure forces.

Proposition E.5 (Fragmentation, solved)**.** *Model a domain as a multi-rate loss system: class i needs b_i slots and has offered load ϱ_i, and let $q(\cdot)$ solve the Kaufman–Roberts recursion*

$$n\,q(n) = \sum_i b_i \varrho_i\, q(n - b_i), \qquad n = 1, \dots, K, \qquad q(0) = 1, \tag{76}$$

normalised, so that $\pi_j = q(K - j)$ is the fraction of domains with j free slots and $p_m := \sum_{j \ge m} \pi_j$ the chance a given domain can host an m-slot request. Then, with D domains placed independently:

(i) $q(0) > 0$, so $\pi_K > 0$ and the largest free block is $\bar{m} = K$ with probability one as $D \to \infty$. Blocking of an m-slot request is

$$P_B(m) = \big(1 - p_m\big)^D, \tag{77}$$

which is strictly positive for every finite D and vanishes geometrically *as $D \to \infty$.*

(ii) Fragmentation therefore does not restore the coexistence of idle capacity and blocking in the fluid limit. What it does is postpone the disappearance: the relevant comparison is D against $1/p_m$, and p_m falls sharply in m, so large jobs face non-negligible blocking at pool sizes at which small ones face none.

(iii) Conditional on displacement occurring, its incidence is as before: with independent placement and uniform targeting a tenant of type v faces $h(v) = m(v)\Delta/s$, so exposure is proportional to footprint and the wedge h/α is increasing in v.

Proof. (i) The recursion (76) is the stationary distribution of occupied slots in a multi-rate Erlang loss system, and $q(0) > 0$ because the empty state is reachable and positive recurrent; hence $\pi_K > 0$ and among D independent domains the chance that none is empty is $(1 - \pi_K)^D \to 0$. The same argument with p_m in place of π_K gives (77). (ii) is immediate from (77), since $P_B(m) \le e^{-Dp_m}$. (iii) is the argument of the previous version, which conditions on a displacement occurring and is unaffected. ■

The numbers make the second part concrete. With $K = 32$ slots per domain at full load and a mixture of one-slot and half-domain jobs, $1/p_m$ is about 21 for $m = K/2$, about 7.8×10^3 for $m = 3K/4$, and about 3×10^{10} for $m = K$. A provider with a few thousand domains therefore supplies essentially no blocking for small jobs, appreciable blocking for three-quarter-domain jobs, and near-certain blocking for jobs that need a whole domain. Fragmentation is thus a statement about the composition of demand at a given scale, not a mechanism that survives scale.

We record this as a correction to the reading we first gave. Fragmentation is not the missing foundation for demand-driven interruption in a large pool. Taken with E.4 and E.7, the conclusion is stronger and more uncomfortable than the one we set out to reach: *no* channel we have examined generates a demand-driven hazard that survives the fluid limit. That is not a gap in the argument for screening—it is the argument. If reallocation cannot explain the interruption a large provider posts, what remains is that she chooses it.

Two things follow, the second connecting back to Section 6.6.

The incidence inverts, conditional on displacement occurring. By (iii) the sign of the wedge is not uniform: Section 5 found under-exposure everywhere, E.4 found that dilution can make it over-exposure everywhere, and fragmentation makes it over-exposure at the top of the market and under-exposure at the bottom. This part of the analysis is unaffected by (i) and (ii), since it conditions on a seizure taking place and asks only who bears it. The testable content is

$$h(v) = \frac{m(v)\,\Delta}{s}, \qquad x(v) = \frac{\varphi(v)\,\Delta}{s}, \tag{78}$$

so that delivered reliability $L\big(\varphi(v)\Delta/s\big)$ falls with footprint, so interruption rates should rise with fleet size at a given price, which is a prediction about a quantity providers already publish per instance rather than about an unobservable.

Fragmentation implements the capped contract of Section 6.6 without posting it. A job whose footprint exceeds the domain size cannot be placed at all: a buyer with $m(v) > K$ can never be

admitted to the discounted tier, since the provider has no reason to displace one spot tenant for another at the same rate, so she must buy guaranteed service whatever her value. Guaranteed service is therefore taken by

$$\Big[\big\{v:\ m(v) > K\big\}\ \cup\ \big\{v \geq v_1\big\}\Big]\ \cap\ \big\{v : R_0(v) \geq w_H\big\},$$

the intersection with the participation set being necessary for the same reason as in the capped-contract boundary of Section 6.6, and the boundary $m^{-1}(K)$ depending on no price in the menu. That is the structure of the capped-contract boundary with the footprint in place of the runtime and the domain size in place of the deadline T—and, now that Proposition E.5 has resolved the occupancy, the analogy is closer than we first thought. The earlier version of this appendix took the binding quantity to be the largest free block $\bar{m}$, an equilibrium object that moves with the composition of demand; solving the domain occupancy shows that $\bar{m} = K$ almost surely, so the binding quantity is the domain size, which is a design choice of the provider's infrastructure and exogenous to her menu exactly as a posted deadline is.

The consequence connects to Section 6.6: placement gives a provider something close to the capped contract in the size dimension, for free and without the commitment problem that distinguishes a posted deadline from the hazard—she is not promising to terminate a healthy job at a foreseen moment, merely declining to assemble capacity that does not exist in one place. Whether that is why capped contracts are absent from commercial pools we do not establish, since it would require knowing how providers choose K and how far the $m > K$ segment extends.

What we have not determined is s, the discounted share, which is a fixed point of the same occupancy process—admission fills blocks, displacement creates them—and enters (78) directly. The results above hold for any $s \in (0, 1)$ but do not pin it. Solving for s is a bin-packing question rather than a loss-system one, and it is the natural next step: it would deliver the discounted share as a function of λ, K and the footprint distribution, and with it the level of the hazard rather than only its shape. We regard the shape as the substantive finding and the level as open.

E.6 Privately known runtime

The body ties runtime to value through a common-knowledge $b(\cdot)$, so the type is scalar. In practice value and required runtime are distinct private characteristics. This subsection lets the buyer observe both, (v, b) drawn from a joint distribution on $\mathbb{R}^2_+$ with neither observed by the seller, and asks what survives.

Proposition E.6 (Two-dimensional types)**.** *Fix a menu $(0, w_H)$, (h, w_L) with $h > 0$. A buyer of type (v, b) takes the guaranteed contract if and only if*

$$v\ \geq\ v^*(b) := \frac{w_H\, b}{1 - L(hb)} - \frac{w_L}{h}, \tag{79}$$

and v^* is strictly increasing in b. Equivalently, writing

$$\psi(v,b) := \Big(v + \frac{w_L}{h}\Big)\frac{1-L(hb)}{b}, \tag{80}$$

the guaranteed contract is chosen exactly when $\psi(v,b) \geq w_H$, and ψ is decreasing in b at fixed v: the sorting statistic is a generalised value per machine-hour. Moreover the demand-driven hazard is $h = \lambda \Pr\big(v \geq v^*(b)\big)$, still a single number, memoryless and independent of the incumbent's type.

Proof. The payoff difference is $\Delta = v\big[1-L(hb)\big] - \big[w_H b - w_L \tau(h,b)\big]$ with $\tau = (1-L(hb))/h$. Setting $u := 1-L(hb) > 0$ and solving $\Delta \geq 0$ for v gives $v \geq \big[w_H b - w_L u/h\big]/u = w_H b/u - w_L/h$, which is (79); rearranging the same inequality as $\big(v + w_L/h\big)u/b \geq w_H$ gives (80). For monotonicity, $b/u = b/\big(1-L(hb)\big)$ is increasing in b, since its reciprocal u/b is the difference quotient of the concave function $z \mapsto 1-L(hz)$ at zero and therefore decreasing; hence v^* is increasing and ψ decreasing in b. The hazard is λ times the probability that an arrival lies in the guaranteed region, by the thinning argument of Lemma 1, which uses only the measure of that region and not its shape. ■

So the mechanism is not an artefact of one-dimensional types. What survives is its substance: sorting still runs on value per machine-hour rather than on value, longer jobs at a given value still take the discounted contract, and the hazard is still an anonymous scalar generated by the composition of demand. What is lost is the scalar cutoff. The guaranteed region is bounded by a curve rather than a point, so there is no v_1 and hence no identity $h = \alpha(v_1)$; the comparison with efficient reallocation in (31) has no single reference type; and the cutoff parametrisation of Proposition 2 is unavailable, so the optimum must be characterised in the two rates directly.

Two further points. The one-dimensional model is the special case in which the joint distribution is supported on the graph of $b(\cdot)$, and Assumption 1—$0 < \eta_b < 1$—is then the statement that this graph is upward-sloping in b and downward-sloping in v/b. Stated in two dimensions the economic content is clearer: what the mechanism needs is that value and runtime be positively associated but less than proportionally, so that the ordering by v and the ordering by v/b differ. A perfect deterministic relationship is a convenience; positive but imperfect association is the substantive restriction, and it is testable in a way that $b(\cdot)$ is not. Second, private footprint is the same problem in a different coordinate: by Proposition E.1 a job enters only through $\varphi = mb$, so privately known size and privately known runtime are the same two-dimensional extension, and (79) applies with φ in place of b.

E.7 Non-stationary demand

Subsection E.4 shows that the demand-driven mechanism needs idle capacity and blocking at the same time, and that a stationary fluid pool with unit demand cannot have both. E.5 asks whether fragmentation restores them and finds it does not, though it delays the loss. A second candidate is time variation: if demand is peaky the pool is saturated at peak, when displacement occurs, and slack off-peak, when spot tenants are admitted, so the two requirements might hold at different times rather than simultaneously. This subsection shows the candidate fails, and that the reason

depends on what buyers know.

Let the arrival rate alternate between λ_H and $\lambda_L < \lambda_H$, a fraction $\psi \in (0,1]$ of time at the peak, with a cycle long relative to a job so that within each phase the pool reaches the fluid state of Proposition E.4 at that arrival rate. Take

$$\lambda_L \, \mathbb{E}[b] < 1 \leq \lambda_H \, \mathbb{E}[b], \tag{81}$$

so that the trough is slack and the peak saturates: by Proposition E.4 the hazard is $h_H = \alpha(v_1)/(1-\omega_H)$ during peaks and zero during troughs, and idle capacity exists during troughs. This is exactly the coexistence E.4 rules out in the stationary case.

Buyers who observe the phase. Nothing is gained. A buyer arriving in a trough faces $h = 0$, so the cheap contract is identical to the expensive one and cheaper, and by (10) she takes the cheap one whatever her type: the menu collapses for every trough arrival. A buyer arriving at peak faces a saturated pool, so the pool is saturated within the phase, so by the argument of E.9(iii) occupancy is one whatever is posted and no menu improves on any other. Time variation therefore reproduces the dichotomy of E.4 phase by phase rather than escaping it, and the seller's optimum is a single rate in each phase—which is to say a single rate, since the phase is not a contractible event.

Buyers who do not observe the phase. This is the substantive case: a buyer designs a workload for the pool's advertised behaviour rather than for the current utilisation, so she faces a *mixture*. With a job short relative to the cycle her completion probability and expected bill are

$$Q_\psi(x) = \psi L(x) + (1-\psi), \qquad \frac{\mathcal{B}_\psi(x)}{h_H} = \frac{\psi\big(1-L(x)\big) + (1-\psi)x}{h_H}, \qquad x := h_H b(v), \tag{82}$$

so by the argument of Appendix D—willingness to pay depends on the specification only through $\Phi = \mathcal{Q}/\mathcal{B}$—she accepts a rate w exactly when $w \leq R_0(v)\,\mathfrak{d}_\psi(x)$ with

$$\mathfrak{d}_\psi(x) := \frac{x\big[\psi L(x) + 1 - \psi\big]}{\psi\big(1-L(x)\big) + (1-\psi)x}, \tag{83}$$

and $\mathfrak{d}_1 = \mathfrak{d}$, $\mathfrak{d}_0 \equiv 1$. Define M_ψ by $d\log \mathfrak{d}_\psi / d\log x = 1 - M_\psi(x)$ and $K_\psi := \sup_{x>0} M_\psi(x)$, so that by (11) the ordering of willingness to pay reverses at v if and only if $\eta_b(v) M_\psi(x) > 1$, and some hazard reverses it if and only if $\eta_b K_\psi > 1$.

Proposition E.7 (Peaky demand caps the screen)**.** *For every $\psi \in (0,1)$:*

(i) $\mathfrak{d}_\psi(0^+) = 1$, $\mathfrak{d}_\psi(\infty) = 1$, and $\mathfrak{d}_\psi(x) < 1$ for all $x > 0$. Hence $\mathfrak{d}_\psi$ attains an interior minimum and is not *monotone, in contrast to $\mathfrak{d}$, which decreases from 1 to 0.*

(ii) Consequently $M_\psi(0^+) = M_\psi(\infty) = 1$ and $K_\psi < \infty$, with $K_\psi \downarrow 1$ as $\psi \downarrow 0$. The set on which

exposure screens, $\{x : \eta_b M_\psi(x) > 1\}$, is a bounded interval: there is a maximum *exposure, and hence a maximum job length, beyond which no discount is justified.*

(iii) Screening by exposure therefore requires $\eta_b K_\psi > 1$, which given $\eta_b < 1$ imposes a strictly positive lower bound on ψ. Together with the $\psi = 1$ case, in which the pool is permanently saturated, the two-tier arrangement fails at both ends of the range: it is dominated when demand is always at peak, and infeasible when peaks are too rare.

Proof. (i) At $x \to 0$, $L = 1 - x + o(x)$ gives numerator $x + o(x)$ and denominator $x + o(x)$, so $\mathfrak{d}_\psi \to 1$. At $x \to \infty$ with $\psi < 1$, $L \to 0$ gives numerator $x(1-\psi) + o(x)$ and denominator $(1-\psi)x + \psi + o(1)$, so again $\mathfrak{d}_\psi \to 1$. For the inequality, $\mathfrak{d}_\psi(x) < 1$ is

$$x\big[\psi L + 1 - \psi\big] < \psi\big(1 - L\big) + (1-\psi)x \iff \psi x L < \psi(1-L) \iff xL < 1 - L,$$

which holds for every $x > 0$ because $1 - L(x) > xL(x)$: for deterministic duration this is $e^x - 1 > x$, and in general $1 - L(x) - xL(x)$ vanishes at $x = 0$ with derivative $-L' - L - xL' = -L - xL' > 0$ since $-xL' \geq L \cdot xM_L \geq 0$ and $L > 0$ fails only in the limit. A continuous function equal to 1 at both ends and strictly below 1 in between attains an interior minimum.

(ii) $M_\psi = 1 - d\log \mathfrak{d}_\psi / d\log x$, and by (i) the log-derivative vanishes at both ends, giving $M_\psi \to 1$; M_ψ is continuous on $(0, \infty)$ and therefore bounded, so $K_\psi < \infty$. As $\psi \downarrow 0$, $\mathfrak{d}_\psi \to 1$ uniformly on compacts, so its log-derivative tends to zero and $K_\psi \downarrow 1$. Since $\eta_b M_\psi > 1$ requires $M_\psi > 1/\eta_b > 1$ and $M_\psi \to 1$ at both ends, the set on which it holds is bounded away from 0 and ∞.

(iii) Immediate from (ii): $\eta_b < 1$ requires $K_\psi > 1/\eta_b > 1$, and K_ψ is increasing in ψ with $K_\psi \to 1$ as $\psi \to 0$, so the condition fails for small ψ. At $\psi = 1$ the mixture disappears, $\mathfrak{d}_1 = \mathfrak{d}$ and $K_1 = K$, and the pool is permanently saturated, so E.9(iii) applies. ■

The mechanism behind (i) is worth stating in words, because it is the whole content. A job long enough to be caught for certain during a peak completes with probability $1 - \psi$ *whatever its length*: reliability saturates instead of decaying. Its expected bill saturates in the same way, at $(1-\psi)w\,b(v)$, so willingness to pay per machine-hour returns to $R_0(v)$ and the discount a long job will accept vanishes. Under stationary demand the opposite holds—$\mathfrak{d}$ decreases to zero and longer jobs are always more screenable, which is what $K = \infty$ under deterministic duration records. Peakiness therefore does not merely weaken the screen; it puts a ceiling on the job length that can be screened at all.

The numbers are stark. For deterministic duration, K_ψ is 1.56, 1.22, 1.11 and 1.03 at $\psi = 0.9$, 0.7, 0.5 and 0.2, against $K = \infty$ at $\psi = 1$, and the minimum of $\mathfrak{d}_\psi$ rises from 0.36 to 0.93 across the same range. So the feasibility condition $\eta_b K_\psi > 1$ binds hard: at $\eta_b = 0.8$ it requires $\psi > 0.74$, at $\eta_b = 0.7$ it requires $\psi > 0.86$, and at $\eta_b \leq 0.65$ it requires $\psi > 0.89$; for the base calibration $\eta_b = 0.5$ it fails for every $\psi < 1$. Where it does hold the screening region is narrow—at $\eta_b = 0.8$ and $\psi = 0.74$ it is $x \in [1.14, 1.36]$—so the instrument works only for jobs in a narrow band of lengths.

Two qualifications. Sorting is not destroyed, only the reversal: the reliability gap $1 - Q_\psi(x) = \psi\big(1 - L(x)\big)$ is still increasing in v, so a menu can still re-price the installed base, and it is the

expansion mode of Section 5.4—reaching buyers a single rate excludes—that peakiness removes. And the result concerns *phase-blind* buyers with jobs short relative to the cycle; a buyer whose job spans many cycles faces the time-average hazard $\bar{h} = \psi h_H$ and is back in the stationary model with λ replaced by its average, which by (81) may then be slack.

The upshot of Proposition E.7 for the paper is negative and worth stating plainly. Of the two channels that might restore demand-driven interruption in a large pool, time variation does not: observing the phase reproduces the dichotomy, and not observing it caps the screen. Fragmentation, in Proposition E.5, postpones the disappearance rather than preventing it: solving the domain occupancy shows blocking falls geometrically in the number of domains, at a rate governed by how hard it is to assemble a job's footprint. So no channel we have examined generates a demand-driven hazard that survives the fluid limit.

E.8 Admission priority in a single facility

Section 5 assumes an arrival is served if the machine is free, whatever contract she takes. A single machine does not work that way. An arrival taking the expensive contract can be served whenever the machine is idle *or* holds a discounted tenant, since she displaces; an arrival taking the discounted contract is served only when the machine is idle. The expensive tier therefore delivers priority of access as well as reliability, and this subsection works out what that adds.

Write P_0, P_S and P_G for the long-run fractions of time the machine is idle, holds a discounted tenant and holds a guaranteed one. Poisson arrivals see time averages, so a buyer compares $(P_0 + P_S)U_H(v)$ with $P_0 U_L(v)$ and with zero; dividing,

$$\gamma := \frac{P_0}{P_0 + P_S} \in (0, 1], \tag{84}$$

and the expensive tier is chosen when $U_H(v) \geq \gamma U_L(v)$. The case $\gamma = 1$ is the body's environment. The fractions solve

$$P_G = \lambda\big(1 - F(v_1)\big)\big(P_0 + P_S\big)\,\mathbb{E}\big[b \mid v \geq v_1\big], \qquad P_S = \lambda\big(F(v_1) - F(v_2)\big)P_0\,\bar{\tau}, \qquad P_0 = 1 - P_G - P_S, \tag{85}$$

with $\bar{\tau} = \mathbb{E}[\tau(h, v) \mid v \in [v_2, v_1)]$, and the revenue rate is $\rho = w_H P_G + w_L P_S$, with no occupancy cost term: the machine's alternative use here is the arriving buyers themselves, and (85) prices it already.

Lemma E.2 (Occupancy and the admission ratio)**.** *Let* $a := \lambda\big(1 - F(v_1)\big)\mathbb{E}[b \mid v \geq v_1]$ *and* $c := \lambda\big(F(v_1) - F(v_2)\big)\bar{\tau}$*. Then* (85) *has the unique solution*

$$P_G = \frac{a}{1 + a}, \qquad P_S = \frac{c}{(1 + a)(1 + c)}, \qquad P_0 = \frac{1}{(1 + a)(1 + c)}, \qquad \gamma = \frac{1}{1 + c}, \tag{86}$$

which lies in the simplex for all $a, c \geq 0$, so no restriction on λ is needed; and

$$\rho = \frac{a\, w_H}{1+a} + \frac{c\, w_L}{(1+a)(1+c)}. \tag{87}$$

The admission ratio is governed by the discounted tier's own offered load and by nothing else.

Proof. Both a and c are non-negative. The first equation of (85) reads $P_G = a(P_0 + P_S) = a(1 - P_G)$, so $P_G = a/(1+a)$; the second gives $P_S = cP_0$; and $P_0 = 1 - P_G - P_S$ then yields $P_0(1+c) = 1/(1+a)$, which is (86). The three fractions sum to one identically and lie in the simplex for all $a, c \geq 0$, so no restriction on λ is needed. Substituting into $\rho = w_H P_G + w_L P_S$ gives (87), and $\gamma = P_0/(P_0 + P_S) = 1/(1+c)$. ■

Priority does two things, and both enlarge the arrangement. It adds the term $(1-\gamma)L(x)$ to (28), a second force separating the tiers that is positive and, despite γ being constant, not type-neutral: writing $\Delta = D + (1-\gamma)U_L$ with $D = U_H - U_L$, the value of being admitted rather than turned away is proportional to $U_L(v)$, which is larger for buyers with more surplus. Priority therefore reinforces exposure rather than offsetting it, screening on the level of rent where exposure screens on duration. And it raises w_H through (27), since the switch type now gives up access as well as reliability.

The quantitative effect is large. Solving (87) using Lemma E.2 at $\eta_b = 1/2$ and $\lambda = 1$ gives $\gamma = 0.823$, a switch cutoff of 0.706 against 0.511 in the body's environment, and a gain over the single rate of 8.0% against 0.2%. Almost all of what might be read as the value of transmitting demand is in fact the value of priority of access, which is a property of finite capacity rather than of the reclamation mechanism. That is why the body sets $\gamma = 1$: it isolates the mechanism, and it is what makes the three arrangements of Section 6.4 comparable. It also sharpens (53), which bounds the transmission increment by the margin on the newly served band: priority adds a second channel that the bound does not cover, and it is the larger of the two.

Finally, v_1 is strictly increasing in γ at fixed prices, so rationing admission strictly shrinks the discounted tier: at the prices above, raising γ to one moves the cutoff from 0.706 to 0.828. The corresponding re-optimised counterfactual runs the other way, to 0.657, because the seller reprices; only the first is a property of sorting.

E.9 A chosen hazard in a pool

Section 6 lets the seller choose the schedule but has one machine; E.4–E.7 have a pool but the hazard is whatever the structure generates. Neither is the case a provider faces, and the two never meet. This subsection joins them: the seller operates the pool of E.4 and may reclaim deliberately on top of whatever displacement occurs.

Write h_{dd} for the displacement rate per spot tenant of (72) and $h_{\mathrm{m}} \geq 0$ for the rate she adds, so a spot tenant faces $h = h_{\mathrm{dd}} + h_{\mathrm{m}}$ and exposure $x = hb(v)$. Fractions g, s, i and the fluid limit are as before, and profit is measured per machine per unit time, so that a machine holding a guaranteed

tenant earns w_H and one holding a spot tenant earns w_L:

$$\rho = g\, w_H + s\, w_L. \tag{88}$$

Proposition E.8 (Chosen hazard in a fluid pool)**.** *In the fluid limit, with* $\omega := \lambda\big(1 - F(v_1)\big)\mathbb{E}[b \mid v \geq v_1]$, $\bar{\tau}(h) := \mathbb{E}[\tau(h, v) \mid v \in [v_2, v_1)]$ *and* $\omega < 1$*:*

(i) (Guaranteed occupancy is invariant.) $g = \omega$*, independent of* h*, of* w_L *and of everything on the cheap tier.*

(ii) (Slack regime, and it is the interesting one.) *The pool is slack,* $i > 0$*, if and only if*

$$\lambda \;<\; \lambda^* := \Big[\big(1 - F(v_1)\big)\mathbb{E}[b \mid v \geq v_1] + \big(F(v_1) - F(v_2)\big)\bar{\tau}(h)\Big]^{-1}, \tag{89}$$

in which case a guaranteed arrival always finds an idle machine, so $h_{\mathrm{dd}} = 0$ *and* the entire hazard is manufactured. *Note that* λ^* *is not a primitive:* $\bar{\tau}$ *is decreasing in* h*, so the threshold rises with the hazard the seller chooses, and* (89) *is a consistency condition on* (λ, h) *rather than a bound on* λ *alone. Manufacturing therefore has a second, indirect effect the single-machine model cannot show: by shortening tenancies it keeps the pool slack, and so preserves the very regime in which it is free. Then* $s = \lambda\big(F(v_1) - F(v_2)\big)\bar{\tau}(h)$ *and*

$$\rho = \lambda \left[\int_{v_1}^{\overline{v}} w_H\, b\, dF + \int_{v_2}^{v_1} w_L\, \tau(h, v)\, dF\right], \tag{90}$$

which is λ *times the per-buyer objective of* (36) *with the occupancy cost* $c \equiv 0$*. Every result of Section 6 therefore applies, with three simplifications: the served set is the whole support, since the break-even condition* (43) *is automatic when* $c \equiv 0$*; the bunch at the arrival rate, region (iii) of Proposition 3, disappears, because it was generated by the kink in* $\mu(\lambda b - x)^+$*; and the exposure condition* (41) *holds with* $\tilde{v} = v$ *throughout.*

(iii) (Saturated regime.) *If* $\lambda \geq \lambda^*$ *then* $i = 0$ *and occupancy is one whatever is posted, so* $\rho = \sum_j w_j P_j$ *with* $\sum_j P_j = 1$*: profit is the occupancy-weighted average of the posted rates and the optimal menu is degenerate. This is the one case in which the ranking of Proposition 4 holds with equality, since when capacity binds the only remaining question is the price per occupied hour.*

Proof. (i) is Proposition E.4(i); the argument uses only that guaranteed tenants are never rationed and never displaced, which the addition of h_{m} does not affect. (ii) If $i > 0$ then every arrival finds an idle machine, so no displacement is required and $h_{\mathrm{dd}} = 0$ by (72); spot tenants enter at $\lambda(F(v_1) - F(v_2))$ and leave at $s/\bar{\tau}(h)$, giving the stated s, and $i = 1 - \omega - s > 0$ rearranges to (89). Substituting $g = \lambda \int_{v_1}^{\overline{v}} b\, dF$ and $s = \lambda \int_{v_2}^{v_1} \tau\, dF$ into (88) gives (90); since capacity is not scarce an occupied machine-hour displaces nothing, so the cost in (38) is zero and the bracket is the revenue term of (36) with $c \equiv 0$. With $c \equiv 0$ the transversality condition of Proposition 3 reads $L(\bar{x})v_2 \geq 0$,

so the corner $v_2 = \underline{v}$ obtains, and $c_x \equiv 0$ gives $\tilde{v} = v$ and removes the kink generating region (iii). (iii) If $i = 0$ every machine is occupied, so the fractions P_j sum to one and $\rho = \sum_j w_j P_j \leq \max_j w_j$, attained by placing all capacity on the highest sustainable rate; the maximiser is therefore a single contract, and a richer menu cannot improve on it. ■

Part (ii) is the substantive one, and it has a consequence that neither Section 6 nor E.4 could deliver alone: whenever the pool is slack, the hazard a discounted tenant faces contains *no* demand-driven component, and by (90) the ranking of Proposition 4 holds there with the occupancy cost set to zero. So

> *wherever an interruptible tier is optimal in a fluid pool, its interruption rate is manufactured in full.*

That is the sharpest form of the paper's original claim, and it is obtained without any appeal to the seller's motives: it follows from the fact that displacement is required only when no machine is idle, and that a menu pays only when some machine is.

At the base calibration the slack-pool optimum is the schedule of Section 6 with $c \equiv 0$. Per arrival it earns about 0.351 against 0.3316 for the best single rate, a gain of about six percent—the two ways of computing the schedule's revenue, from the solver and by integrating $w\tau$ directly, differ in the fourth decimal, so the ratio is 1.059 to 1.060 and should not be quoted more precisely; it serves the entire support; the regions present are the frontier, an interior region and the guaranteed region, with no bunch, as part (ii) predicts; the switch cutoff is $v_1 = 0.4843$ with $w_H = 0.66785$; and the hazard at the bottom of the market is $\bar{x}/b(\underline{v}) = 7.127$, so at $\lambda = 1$ the posted rate is seven times the arrival rate and, by part (ii), entirely of her own making. The slack threshold at that optimum is $\lambda^* \approx 1.67$: idle capacity is $+0.402$ of the pool at $\lambda = 1$ and $+0.103$ at $\lambda = 1.5$, and the pool saturates between $\lambda = 1.5$ and $\lambda = 2$.

Three implications, in ascending order of importance.

The two accountings of the paper are reconciled rather than merely coexisting. The occupancy cost $\mu(\lambda - h)^+$ of (1) is a reduced form for the value of a machine-hour, and (90) shows what it equals in a pool: zero when the pool is slack, because an occupied machine displaces nothing, and the whole rate when it is saturated, because then it displaces everything. The intermediate case is finite N, where $P_B \in (0, 1)$.

The screening result of Proposition 4 is not confined to a single machine. Its slack-pool counterpart is (90), and the ranking survives with the same proof, since a constant hazard remains admissible. The magnitude has to be read carefully, though, because the benchmark changes: in a slack pool the demand-driven menu collapses by Proposition E.4(iii), so $\Pi_2 = \Pi_H$ and the screening gain is measured against the single rate. It is about six percent there. The 7.0% reported in Section 6 is Π_3/Π_2 in the single-facility problem, computed per-buyer with $\kappa = \mu\lambda$ and $\gamma = 1$ as in Table 3; it is not the same ratio and the two are not directly comparable. Note also that Section 5 states its own optimum in stationary form, so its levels are comparable with neither.

And the transparency argument of Section 7.2 acquires a checkable premise. By part (ii), a provider whose pool has idle capacity cannot attribute any part of her reclamation rate to demand:

in the fluid limit the demand-driven component is exactly zero, and at finite N it is of order P_B, which is small at loads below saturation. Utilisation is the one quantity in this literature that providers do publish, so the premise is verifiable without any disclosure of internal rules—which is, we think, the strongest form the argument can take.