informs
http://pubsonline.informs.org/journal/Journal Name

INFORMATION SYSTEMS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0000-0000, EISSN 0000-0000

**Submitted to *Information Systems Research***

# Online Appendix: Strategic Throttling in Large Cloud Computing Platforms

Yingda Zhai
Department of Information Systems and Analytics, National University of Singapore, Singapore 119391, yingdazhai@nus.edu.sg

Maxwell B. Stinchcombe
Department of Economics, University of Texas at Austin, Texas 78712, max.stinchcombe@utexas.edu

Andrew B. Whinston
McComb School of Business, University of Texas at Austin, Texas 78705, abw@uts.cc.utexas.edu

## A.1 Construction of Hyperfinite Queues

We provide a detailed discussion of the construction of hyperfinite queues, beginning with the hyperfinite Poisson process formulated as a coin-flipping construction (Loeb 1975) in Subsection A.1.1. We then extend this construction to a hyperfinite birth–death process with a large number of traffic sources and server stations in Subsection A.1.3. The resulting service system approximates a large public cloud provider, where multi-station batch service is driven by heterogeneous traffic generated through user-specific service contracts.

The hyperfinite construction offers several advantages.

First, it models directly the large discrete queueing systems that arise in practice. Unlike fluid-limit approaches, which study limiting behavior after rescaling a sequence of systems as the scaling parameter tends to infinity, the hyperfinite formulation starts from a large but discrete system and preserves its structural features. This makes it particularly suited to environments such as cloud platforms, where individual jobs and service interactions remain economically meaningful even at large scale.

Second, the hyperfinite model provides an explicit and tractable representation of large discrete collections. It retains key finite-set properties—such as combinatorial structure and well-defined extremal elements—while simultaneously supporting continuum-style aggregation through integration. This allows us to analyze steady-state outcomes directly in a system with a large number of heterogeneous computing jobs. By contrast, fluid limits rely on rescaling time or state variables to suppress stochastic fluctuations around mean behavior, and the appropriate scaling varies across models, making economic interpretation less uniform (Whitt 2002).

Third, the hyperfinite framework naturally induces a limiting interpretation. Results obtained in this setting can be mapped to a conventional asymptotic sequence indexed by the number of independent traffic sources,

thereby providing an expansion path for a growing cloud system. This links the large-system behavior to standard convergence notions while preserving the structural clarity of the underlying discrete system.

Based on this construction, we develop three lemmas on aggregate utilization to establish the absence of aggregate uncertainty and to characterize service tranches in Subsection A.1.4.

Finally, we emphasize that our approach relies on standard tools from nonstandard analysis, including hyperfinite sets and internal structures. For readers interested in further background, we refer to Cutland (1988) and Sun (2006).

### A.1.1. Construction of the Hyperfinite Poisson Process

Let $L(\mathbf{T}) \triangleq (T, L(\mathcal{T}), L(\tau))$ be a hyperfinite timeline of size $H$:

$$T = \{n/H : n \in {}^*N \text{ and } n < H\},$$

for some infinite hyperinteger $H \in {}^*\mathbb{N} \setminus \mathbb{N}$. We consider the coin-tossing construction in Loeb (1975), where a coin is flipped over each incremental interval $[k/H, (k+1)/H)$ for $k = 0, 1, \ldots, H-1$, with probability $\lambda/H$ of heads.

We construct a hyperfinite counting space $L(\mathbf{\Omega}) \triangleq (\Omega, L(\mathcal{A}), L(\mu))$, where the internal set $\Omega = \{0,1\}^T$ contains all $2^H$ possible realizations of the coin-tossing process. Each $\omega \in \Omega$ is an internal sequence $\{\omega_t\}_{t\in T}$ with $\omega_t \in \{0,1\}$. A hyperfinite stochastic process is defined on $(\Omega \times T, L(\mathcal{A} \otimes \mathcal{T}), L(\mu \otimes \tau))$.

One can interpret the internal probability space $(\{0,1\}^T, \mathcal{A}, \mu)$ as an equivalence class $\langle(\{0,1\}^{T_n}, \mathcal{A}_n, \mu_n)\rangle$, where each finite approximation is defined on $T_n = \{0, 1, \ldots, n\}$.

For fixed $t \in T$, define $X(\omega, t) \in L(\Omega \otimes T)$ as the number of $s < t$ such that $\omega(s) = 1$, i.e.,

$$X(\omega, t) = \sum_{s<t} \mathbf{1}\{\omega(s) = 1\}.$$

This corresponds to a binomial random variable for a coin that takes value 1 (heads) with probability $\lambda/H$ and 0 (tails) with probability $1 - \lambda/H$. Then $X$ is a right lifting of a Poisson process $x$ on $L(\Omega)$.

To see this, consider an event $A \in L(\Omega)$ of exactly $k$ heads in a time interval of length $t$. There are $tH$ trials within this interval,[1] and thus

$$\begin{aligned}
L(\mu)(A) &= \binom{tH}{k} (\lambda H^{-1})^k (1 - \lambda H^{-1})^{tH-k} \\
&= \binom{tH}{k} \left(\frac{\lambda H^{-1}}{1 - \lambda H^{-1}}\right)^k \exp\{tH \cdot \ln(1 - \lambda H^{-1})\} \\
&= \frac{tH \cdot (tH-1) \cdots (tH-k+1)}{k!} \left(\frac{\lambda H^{-1}}{1 - \lambda H^{-1}}\right)^k \exp\{-\lambda t\} \\
&\simeq e^{-\lambda t} (\lambda t)^k / k!
\end{aligned}$$

[1] For simplicity, assume $H = \xi!$ for some $\xi \in {}^*\mathbb{N} \setminus \mathbb{N}$ and that $t$ is a standard rational number, so that $tH$ is an integer.

### A.1.2. Construction of the Hyperfinite Queue $M/M/\infty/\infty/S$

The $M/M/\infty/\infty/S$ queue can be represented as a birth–death process with arrival rate $\lambda_k = (S-k)\lambda$ and service rate $\mu_k = k\mu$. This process can be decomposed into the interaction of a pure birth component and a pure death component, which correspond respectively to independent Poisson processes governing arrivals and service completions.

Given our hyperfinite construction of the Poisson process, we can construct a hyperfinite birth–death process by combining the corresponding hyperfinite Poisson processes. Over an infinitesimal time interval $[k/H, (k+1)/H)$ of length $1/H$, the probability of a net increase of one unit is driven primarily by one arrival and zero departures, while other combinations (multiple arrivals, multiple departures, or simultaneous events) occur with negligible probability. To see this,

$$\begin{aligned}
&P\left(B\left(\frac{k+1}{H}\right) - B\left(\frac{k}{H}\right) = 1 \,\middle|\, B\left(\frac{k}{H}\right) = k\right) \\
&= \frac{(\lambda_k H^{-1})^1 e^{-\lambda_k H^{-1}}}{1!} \cdot \frac{(\mu_k H^{-1})^0 e^{-\lambda_k H^{-1}}}{0!} + o(H^{-1}) \\
&= \lambda_k H^{-1} \cdot e^{-(\lambda_k+\mu_k)H^{-1}} + o(H^{-1}) \\
&= \lambda_k H^{-1} \sum_{n=0}^{\infty} \frac{(-(\lambda_k+\mu_k)H^{-1})^n}{n!} + o(H^{-1}) \\
&= \lambda_k H^{-1}(1 - H^{-1}(\lambda_k+\mu_k) + \frac{1}{2}H^{-2}(\lambda_k+\mu_k)^2 - \cdots) + o(H^{-1}) \\
&= \lambda_k H^{-1} + o(H^{-1}).
\end{aligned} \tag{1}$$

Similarly, the probability of a decrease of size 1 in the incremental time interval has the form:

$$P\left(B\left(\frac{k+1}{H}\right) - B\left(\frac{k}{H}\right) = -1 \,\middle|\, B\left(\frac{k}{H}\right) = k\right) = \mu_k H^{-1} + o(H^{-1}). \tag{2}$$

As discussed, for a sufficiently small incremental time interval (i.e., $H^{-1}$ is infinitesimal),

$$P\left(\left|B\left(\frac{k+1}{H}\right) - B\left(\frac{k}{H}\right)\right| > 1 \,\middle|\, B\left(\frac{k}{H}\right) = k\right) = o(H^{-1}). \tag{3}$$

The equations (1), (2), and (3) together define a birth–death process with $\lambda_k = (S-k)\lambda$ and $\mu_k = k\mu$ for our queueing model. Therefore, we rewrite the transition probability of the birth–death process as:

$$p_{ij}(H^{-1}) = \delta_{ij} + q_{ij}H^{-1} + o(H^{-1}) \tag{4}$$

where $\delta_{ij}$ is Kronecker's delta, i.e., $\delta_{ij} = 1$ when $i = j$ and $\delta_{ij} = 0$ otherwise, and

$$q_{ij} = \begin{cases} \lambda_i, & \text{if j = i+1} \\ \mu_i, & \text{if j = i-1} \\ -(\lambda_i + \mu_i), & \text{if j = i} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

For a single $M/M/1/1/1$ queue with only one traffic source, $\lambda_k = \lambda$ only if $k = 0$ and $\mu_k = \mu$ only if $k = 1$ (and $\mu_k = 0$ elsewhere), the expression (5) degenerates to (6) in matrix form. We can then write the backward Kolmogorov differential equation to solve the steady-state distribution of $M/M/1/1/1$ queues.

### A.1.3. Hyperfinite Queue without Multi-Station Batch Service

We first study the $M/M/\infty/\infty/S$ queue without considering batch service. Each user independently generates a job that occupies at most one server and is associated with an independent service time. We represent the $M/M/\infty/\infty/S$ queue as the aggregation of $S$ independent $M/M/1/1/1$ queues, where each user $s \in 1, 2, \ldots, S$ corresponds to one such queue.

Let $L(\mathbf{T})$ of size $H$ denote the hyperfinite timeline, and let $L(\mathbf{\Omega_s})$ be the corresponding hyperfinite counting space, where $\Omega_s = 0, 1^T$ represents all possible realizations of coin flips for user $s$. For each $M/M/1/1/1$ queue, at most one job is in service at any time, with exponentially distributed inter-arrival and service times. The hyperfinite coin-toss construction can be interpreted as one coin flip over each infinitesimal interval $[k/H, (k+1)/H)$, where heads indicates that a job is served in that interval and tails indicates no service activity.

Let $B_s(t) \in 0, 1$ denote whether user $s$ has a job in service at time $t \in T$. For each realization $\omega \in \Omega_s$, the process $B_s(\omega, t)$ is a lifting of the birth–death process of the $M/M/1/1/1$ queue on the Loeb product space. More precisely, $B_s(t)$ forms a Markov chain with infinitesimal generator given by

$$\mathbf{Q} = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}. \tag{6}$$

In steady state (for $t = k/H$ with $k \in {}^*\mathbb{N} \setminus \mathbb{N}$), $B_s$ follows a Bernoulli distribution with success probability $P(B_s = 1) = \rho/(1+\rho) = \lambda/(\lambda+\mu)$. We now define the total number of busy servers in the $M/M/\infty/\infty/S$ system as $B(S, \lambda, \mu)(t) = \sum_s B_s(t)$, which aggregates the $S$ independent processes $B_s(t)$.

Since the $B_s$ are independent Bernoulli random variables, the steady-state distribution of $B(S, \lambda, \mu)$ is binomial:

$$P\left(B(S,\lambda,\mu) = k\right) = \frac{\binom{S}{k}\rho^k}{\sum_{k=0}^{S}\binom{S}{k}\rho^k} = \binom{S}{k}\left(\frac{\rho}{1+\rho}\right)^k\left(1 - \frac{\rho}{1+\rho}\right)^{S-k}, \quad k = 0, 1, \ldots, S.$$

It follows that the mean number of busy servers is $S\lambda/(\lambda+\mu)$ and the variance is $S\lambda\mu/(\lambda+\mu)^2$. Normalizing by $S$ yields the aggregate utilization $B(S, \lambda, \mu)/S$, whose dispersion vanishes as the system scales. This leads to the following aggregation result.

LEMMA EC.1. *For a large population of users $S \in {}^*\mathbb{N} \setminus \mathbb{N}$, the distribution of the server capacity utilization level has a point mass. More formally,*

$$\mathbb{E}\left(B(S,\lambda,\mu)/S\right) = \frac{\lambda}{\lambda+\mu} \text{ and } Var\left(\mathrm{B(S,}\lambda,\mu\mathrm{)/S}\right) \simeq 0.$$

Lemma EC.1 extends naturally to heterogeneous environments where each user $s \in S$ has idiosyncratic parameters $(\lambda_s, \mu_s)$.

LEMMA EC.2. *When each user has a different traffic profile $(\lambda_s, \mu_s)$ for $s \in S$, the distribution of the utilization of the servers has a point mass. Put formally, $\mathbb{E}(B(S,\lambda,\mu)/S) \simeq \int \frac{\lambda_s}{\lambda_s+\mu_s} d\mathcal{U}(s)$ and $\mathrm{Var}\left(B(S,\lambda,\mu)/S\right) \simeq 0$, where $\mathcal{U}$ denotes a uniform distribution on $\{1,2,\ldots,S\}$.*

*Proof.* The steady-state variable $B_s$ is Bernoulli with success probability $\lambda_s/(\lambda_s+\mu_s)$. Hence $B(S,\lambda,\mu)$ is a Poisson binomial random variable. Its mean and variance are given by the sums of the corresponding Bernoulli components:

$$\mathbb{E}\,\frac{B(S,\lambda,\mu)}{S} = \frac{1}{S}\sum_{s\in S}\frac{\lambda_s}{\lambda_s+\mu_s} \simeq \int \frac{\lambda_s}{\lambda_s+\mu_s} d\mathcal{U}(s),$$

and variance

$$\mathrm{Var}\left(\frac{\mathrm{B(S,}\lambda,\mu\mathrm{)}}{\mathrm{S}}\right) = \frac{1}{S^2}\sum_{s\in S}\frac{\lambda_s\mu_s}{(\lambda_s+\mu_s)^2} \simeq \frac{1}{S}\int \frac{\lambda_s\mu_s}{(\lambda_s+\mu_s)^2} d\mathcal{U}(s) \simeq 0.$$

Since $S$ users generate traffic of equal weight, where $S \in {}^*\mathbb{N} \setminus \mathbb{N}$, the average sum converges to a Riemann integral. □

Finally, since each user draws independently from a type distribution $F(\theta)$, we can equivalently index traffic by job type $\theta \in \Theta$. Rewriting the previous result in terms of types yields:

LEMMA EC.3. *When each job of type-$\theta$ has parameters $(\lambda(\theta), \mu(\theta))$, the normalized utilization converges to a point mass:*

$$\mathbb{E}\left(B(S,\lambda(\theta),\mu(\theta))/S\right) \simeq \int_\Theta \frac{\lambda(\theta)}{\lambda(\theta)+\mu(\theta)} dF(\theta), \quad \mathrm{Var}\left(B(S,\lambda(\theta),\mu(\theta))/S\right) \simeq 0.$$

### A.1.4. Hyperfinite Queue with Multi-Station Batch Service.

The $M^w/M/\infty/\infty/S$ queue with batch service allows each job to occupy multiple stations within a single $M/M/w/w$ system. Since a type-$\theta$ job can utilize $w(\theta)$ servers simultaneously, we must carefully define both total system capacity and the corresponding utilization measure.

Let the total capacity of the service system be $K = c \cdot \bar{w}S$, where $\bar{w}$ denotes the maximum number of computing units that can be allocated to any job and $c \in [0,1)$. Each user $s \in S$ draws a type $\theta$ independently from the distribution $F(\theta)$, and since $S \in {}^*\mathbb{N} \setminus \mathbb{N}$, let $S(\theta)$ denote the number of type-$\theta$ jobs in the population. Each arriving type-$\theta$ job then occupies $w(\theta)$ servers in the system.

We define the $\theta$-tranche as the collection of capacity allocated to type-$\theta$ jobs, given by $w(\theta)S(\theta)$. The $\theta$-tranche therefore forms a sub-queue within the overall service system.

A $\theta$-tranche consists of $S(\theta)$ independent $M/M/w(\theta)/w(\theta)$ queues, each equipped with $w(\theta)$ servers. As in the single-server case, in steady state the number of busy servers assigned to a type-$\theta$ job, denoted $B_s(\lambda(\theta), \mu(\theta), w(\theta))$, is a Bernoulli random variable. With probability $\lambda(\theta)/(\lambda(\theta)+\mu(\theta))$, all $w(\theta)$ servers are simultaneously busy for a type-$\theta$ job in service; otherwise, with probability $\mu(\theta)/(\lambda(\theta)+\mu(\theta))$, no servers are active.

The utilization of the $\theta$-tranche is defined as the total expected number of busy servers allocated to type-$\theta$ jobs, $B(\lambda(\theta), \mu(\theta), w(\theta))$, normalized by the total allocated capacity $w(\theta)S(\theta)$. That is,

$$K(\theta) = \frac{B(\lambda(\theta), \mu(\theta), w(\theta))}{w(\theta)S(\theta)}.$$

By applying Lemma EC.3, the steady-state distribution of $K(\theta)$ collapses to a point mass at its mean for every $\theta \in \Theta$. This establishes that each tranche exhibits deterministic utilization in steady state, as summarized in Proposition 1 in Section A.2.1.

### A.1.5. Transfer Principle and the Hyperfinite-to-Limit Interpretation.

A useful feature of hyperfinite queues is that statements in the hyperfinite model can be interpreted within standard limit theory for expanding service systems with a growing user base. This follows from the transfer principle, a central result in nonstandard analysis (see Robinson 2016 and Khan and Sun 1999). In our setting, the hyperfinite model implies that tranche utilization has variance equal to a positive infinitesimal. In standard limit terms, this corresponds to the fact that, once the user base is sufficiently large, pooling uncorrelated traffic eliminates aggregate demand risk.

Formally, let $P(S, \delta)$ denote a statement about queue performance in the hyperfinite model, where $S$ is a hyperfinite population size and $\delta$ is an infinitesimal bound. For instance, one such statement is that for hyperfinite $S$, there exists $\delta$ such that $\mathrm{Var}(K(\theta)) = \delta \simeq 0$. The transfer principle implies that any such hyperfinite statement admits a standard counterpart $P(N, \epsilon)$: for every $\epsilon > 0$, there exists a finite $N$ such that whenever the user base exceeds $N$, the variance of standardized utilization is less than $\epsilon$.

The transfer principle therefore provides a bridge between the hyperfinite formulation and standard asymptotic analysis, allowing us to interpret predictability results as statements about sufficiently large but finite platforms along a growth path. We use this connection throughout the paper to translate hyperfinite results into economically meaningful comparative statics.

PROPOSITION EC.1. *Given a hyperfinite statement $P(S, \delta)$, for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that the corresponding standard statement $P(N, \epsilon)$ holds.*

*Proof.* For a standard $\epsilon > 0$, let $A(\epsilon)$ be the set $A(\epsilon) = \{S \in {}^*\mathbb{N} \setminus \mathbb{N} : P(S, {}^*\epsilon)\}$. The set $A(\epsilon)$ is an internal set (see Cutland (1988) for details).[2]

[2]For example, to replace the time interval $[0, 1]$, a hyperfinite timeline can be defined as an internal probability space $\mathbf{T} = \{T, \mathscr{A}, P\}$, where $T = \{0, \delta t, 2\delta t, \ldots, \xi\delta t = 1\}$ is a hyperfinite internal set, with $\delta t = \xi^{-1}$ for some $\xi \in {}^*\mathbb{N} \setminus \mathbb{N}$, $\mathscr{A}$ is the internal algebra of all internal subsets of $T$, and $P$ is the finitely additive internal counting measure defined as $P(A) = |A|/|T| = \xi^{-1}|A|$ for $A \in \mathcal{A}$. Note that the standard map $^\circ : T \mapsto [0, 1]$ is onto, i.e., no irrational number $r \in \mathbb{R}$ is an element of $T$, but for any irrational $r$, there exists a unique $t \in T$ such that $t < r < t + \delta t$. For more details on hyperfinite counting spaces, see Albeverio (1986).

By definition, $\delta < {}^*\epsilon$ for any infinitesimal $\delta$. Hence $P(S, {}^*\epsilon)$ holds whenever $P(S, \delta)$ holds. Moreover, $A(\epsilon)$ contains all sufficiently large infinite integers for every $\epsilon > 0$. By the underflow property of internal sets, if $A(\epsilon)$ contains arbitrarily large infinite integers, it must also contain a finite integer $N \in \mathbb{N}$. The result follows. □

## A.2 Proofs of Major Results for the Monopoly Provider

### A.2.1. Proof of Proposition 1

PROPOSITION 1 (Restated). *Fix a traffic profile for type-$\theta$ jobs given by the arrival rate $\lambda(\theta)$, service rate $\mu(\theta)$, and batch size $w(\theta)$. In a sufficiently large cloud system with uncorrelated job arrivals, the steady-state utilization of the $\theta$-tranche becomes deterministic. Specifically, the fraction of total capacity devoted to type-$\theta$ jobs converges to the mean $k(\theta) = \lambda(\theta)/\big(\lambda(\theta) + \mu(\theta)\big)$, and the variance of tranche utilization converges to zero.*

*Proof.* Given $S$ users generate jobs following a distribution $f(\theta)$, we can group servers for $\theta$-type jobs together, and we know $w(\theta)S(\theta)$ are potentially allocated for $\theta$-type jobs. Given the $\theta$-tranche, each $M/M/w(\theta)/w(\theta)$ queue behaves as an independent Bernoulli random variable in the steady-state, with a mean equal to $\hat{k}(\theta) \triangleq w(\theta) \cdot \lambda(\theta)/(\lambda(\theta) + \mu(\theta))$.[3] The number of busy servers for the $\theta$-tranche is the sum of $S(\theta)$ i.i.d. Bernoulli random variables with a mean $\hat{k}(\theta)$. The expected utilization of the $\theta$-tranche is thus

$$k(\theta) \triangleq \mathbb{E}\,\frac{1}{w(\theta)S(\theta)} \sum_{s \in S(\theta)} B(\lambda(\theta), \mu(\theta), w(\theta)) = \frac{S(\theta)w(\theta)}{w(\theta)S(\theta)} \cdot \frac{\lambda(\theta)}{\lambda(\theta) + \mu(\theta)} = \frac{\lambda(\theta)}{\lambda(\theta) + \mu(\theta)}$$

and the variance of utilization is

$$\mathrm{Var}\left(\frac{B(\lambda(\theta), \mu(\theta), w(\theta))}{w(\theta)S(\theta)}\right) = \frac{S(\theta)(w(\theta))^2}{(w(\theta)S(\theta))^2} \cdot \frac{\lambda(\theta)\mu(\theta)}{(\lambda(\theta) + \mu(\theta))^2} = \frac{1}{S(\theta)} \cdot \frac{\lambda(\theta)\mu(\theta)}{(\lambda(\theta) + \mu(\theta))^2} \simeq 0$$

□

### A.2.2. Proof of Theorem 1

THEOREM 1 (Restated). *Consider a large cloud platform in which users of type $\theta \in \Theta$ generate independent Poisson job arrivals. In steady state, the utilization of each service tranche becomes effectively deterministic. As the number of users grows, random fluctuations in tranche-level utilization vanish.*

*Moreover, given the traffic profile $\{\lambda(\theta), \mu(\theta), w(\theta)\}$ and interruption policy $r(\theta)$, steady-state capacity usage must satisfy the feasibility constraint:*

$$\int_\Theta k(\theta)\big(1 - r(\theta)\big) dF(\theta) \le c, \tag{FC}$$

*where the expected utilization share generated by type-$\theta$ jobs is*

$$k(\theta) = \frac{\lambda(\theta)}{\lambda(\theta) + \mu(\theta)} \cdot \frac{w(\theta)}{\bar{w}} \cdot f(\theta).$$

[3] For the binary outcome, there is a $\lambda(\theta)/(\lambda(\theta) + \mu(\theta))$ probability that $w(\theta)$ servers are in use and a $(1 - \lambda(\theta)/(\lambda(\theta) + \mu(\theta)))$ probability that no server is in use.

*Proof.* The proof of the theorem extends Proposition 1. To derive the capacity constraint, we express the standardized utilization level $k(\theta)$ of $\theta$-type jobs in (FC) relative to the total capacity $\bar{w}S$, rather than $S(\theta)w(\theta)$ within the $\theta$-tranche. From Proposition 1, we know that there are $S(\theta)\hat{k}(\theta)$ busy servers allocated to $\theta$-type jobs in steady state.

Similar to Lemma EC.3, the utilization level of the entire service capacity follows a Poisson binomial distribution with mean

$$\frac{1}{\bar{w}S}\sum_{s\in S}\frac{\lambda_s w_s}{\lambda_s+\mu_s} \simeq \int_\Theta \frac{\lambda(\theta)w(\theta)}{(\lambda(\theta)+\mu(\theta))\bar{w}}dF(\theta),$$

and variance

$$\frac{1}{(\bar{w}S)^2}\sum_{s\in S}\frac{\lambda_s\mu_s w_s^2}{(\lambda_s+\mu_s)^2} \simeq 0.$$

Now consider a throttling strategy $r(\theta)$, which introduces an independent source of random interruption during the service of a type-$\theta$ job, blocking access to a server with Bernoulli probability $r(\theta)$. In steady state, both the number of busy servers within each queue and the interruption events are independent Bernoulli random variables. Since the product of two independent Bernoulli random variables is also Bernoulli, the expected server capacity allocated to $\theta$-type jobs under the throttling strategy is given by $k(\theta)(1-r(\theta))$.

The total utilization of the service system, obtained by summing over all $\theta \in \Theta$, must not exceed the capacity threshold $c = K/(\bar{w}S)$. □

Furthermore, the rate of variance shrinkage is inversely proportional to the size of the user base $S$ (or equivalently the service capacity $K$), following a rate of $O(1/S)$ (or $O(1/K)$). This result is consistent with prior work in queueing theory and stochastic modeling (Halfin and Whitt 1981, Harchol-Balter 2013, Gans et al. 2003), which shows that LLN approximations are accurate in high-capacity regimes where server capacity scales proportionally with demand.

### A.2.3. Proof of Corollary 1

COROLLARY 1 (Restated). *A profit-maximizing provider benefits from operating a large-scale service system. Non-negligible aggregate uncertainty leads to a loss of efficiency and profit. In particular,*

$$\Pi^\epsilon \le \Pi^0,$$

*with strict inequality whenever the deterministic capacity constraint binds and aggregate uncertainty is non-negligible.*

*Proof.* Recall that the solution $(w^*, r^*, p^*)$ to the problem (P) is obtained under the capacity constraint (FC), where $\lambda(\theta)/(\lambda(\theta)+\mu(\theta))$ is the expected proportion of capacity working on type-$\theta$ jobs in the steady-state. Theorem 1 implies that $K(\theta)$ has a point mass at its mean $k(\theta) = \lambda(\theta)/(\lambda(\theta)+\mu(\theta))$, with $\mathrm{Var}\,(K(\theta)) \simeq 0$. The feasible set of solutions is therefore characterized by the deterministic constraint

$$\int_\Theta k(\theta)(1-r(\theta))\,dF(\theta) \le c.$$

Let $\mathcal{F}^0$ denote this feasible set and $\Pi^0$ the associated maximal profit.

When utilization is stochastic and the provider commits to a service-level agreement limiting shortfall probability to $\epsilon$, capacity planning must guard against high realizations of demand. Let $\bar{k}_\epsilon(\theta)$ denote the upper $(1-\epsilon)$-quantile of $K(\theta)$. Because $\bar{k}_\epsilon(\theta) \geq k(\theta)$, the risk-adjusted constraint

$$\int_\Theta \bar{k}_\epsilon(\theta)(1-r(\theta))\,dF(\theta) \leq c$$

defines a feasible set $\mathcal{F}^\epsilon \subseteq \mathcal{F}^0$, with strict inclusion whenever utilization variance is positive on a set of positive measure.

Since the deterministic optimum remains feasible only when the constraint does not bind or uncertainty vanishes, maximal profit under uncertainty cannot exceed $\Pi^0$. If the deterministic constraint binds and $\bar{k}_\epsilon(\theta) > k(\theta)$ on a set of positive measure, the feasible set shrinks strictly, implying $\Pi^\epsilon < \Pi^0$. □

### A.2.4. Proof of Theorem 2

THEOREM 2 (Restated). *In the provider's optimal mechanism,*

*(i) Computing speed is increasing in willingness to pay: $w^*(\nu)$ is increasing in $\nu$.*

*(ii) Interruption probability is decreasing in delay sensitivity: $r^*(\kappa)$ is decreasing in $\kappa$.*

*(iii) Compared to the socially optimal quality, the provider deliberately distorts quality downward for low types. In particular, low-$\nu$ users receive less than the efficient computing speed, and low-$\kappa$ users face higher interruption risk.*

*(iv) When user reservation utility is fixed at zero, the provider serves the entire market.*

*Proof.* With a monotonic transformation of the user's expected utility function, we write:

$$\bar{U}(\theta,\tilde{\theta}) = \nu \cdot w(\tilde{\theta})\mu - \kappa \cdot (1-r(\tilde{\theta}))^{-1} - p(\tilde{\theta}).$$

We can then rewrite the service provider's problem as:

$$\max_{w(\cdot),r(\cdot),p(\cdot)} \int_\Theta p(\theta)dF(\theta) \quad \text{subject to} \tag{P'}$$

$$\nu \cdot w(\theta)\mu - \kappa \cdot (1-r(\theta))^{-1} - p(\theta) \geq \nu \cdot w(\tilde{\theta})\mu - \kappa \cdot (1-r(\tilde{\theta}))^{-1} - p(\tilde{\theta}), \quad \forall \tilde{\theta} \in \Theta; \tag{IC'}$$

$$\nu \cdot w(\theta)\mu - \kappa \cdot (1-r(\theta))^{-1} - p(\theta) \geq 0, \quad \forall \theta \in \Theta; \tag{IR'}$$

$$\int_\Theta k(\theta)(1-r(\theta))dF(\theta) \leq c, \tag{FC}$$

$$\text{where } \lambda(\theta) = \lambda f(\theta) \text{ and } k(\theta) = \frac{\lambda(\theta)w(\theta)f(\theta)}{(\lambda(\theta)+\mu(\theta))\bar{w}}.$$

The solution to (P') is equivalent to the original problem (P) because the incentive compatibility and individual rationality constraints in (IC) and (IR) are preserved under a monotonic transformation of the utility function $U(\theta,\tilde{\theta})$. In particular, this transformation amounts to scaling terms in the utility expression without altering preference ordering.

The implications for mechanism design are twofold. First, the provider can separately sort users by $\nu$ and $\kappa$ through the instruments $w(\theta)$ and $r(\theta)$, respectively. Second, the multidimensional screening problem decomposes into two single-dimensional screening problems. As a result, the transformed utility $\bar{U}$ satisfies a single-crossing property in each dimension: it is supermodular in $(w, \nu)$ and submodular in $(r, \kappa)$ (Rochet 1985, McAfee and McMillan 1988). This implies that higher-$\nu$ users optimally select higher $w$, while higher-$\kappa$ users optimally select lower $r$. Therefore, users truthfully self-select into the menu $(w^*, r^*, p^*)$ according to their types.

Since the objective function is increasing in $w$ and decreasing in $r$, and the capacity constraint behaves in the same monotone way, the capacity constraint (FC) binds at the optimum. The resulting allocation depends on the capacity parameter $c$. Let $(w^{**}, r^{**})$ denote the first-best allocation (without (IC')), and $(w^*, r^*)$ the second-best solution to (P'). Then $w^*(\nu, c)$ is nondecreasing in $\nu$, and $r^*(\kappa, c)$ is nonincreasing in $\kappa$. The individual rationality constraint binds only at the lowest type:

$$u(\underline{\nu}, \underline{\kappa}) = 0,$$

while $\bar{U}(\theta) \geq 0$ for all $\theta$, with strict inequality for all interior types whenever incentive constraints are binding.

We now show that exclusion cannot be optimal when the reservation utility is zero. Because service quality can be continuously adjusted by lowering $w(\theta)$ and increasing $r(\theta)$, the effective capacity usage $k(\theta)(1 - r(\theta))$ can be made arbitrarily small. Hence, for any type with $\bar{U}(\theta) < 0$, one can construct a contract with sufficiently low $w(\theta)$ and sufficiently high $r(\theta)$ such that $\bar{U}(\theta) = 0$, while consuming arbitrarily little capacity. This modification preserves (IC') and does not violate (FC), while weakly increasing profit. Therefore, exclusion is strictly dominated, and all types are served in equilibrium.

Consequently, under zero reservation utility, inefficiency arises not through exclusion but through downward distortion of service quality. □

### A.2.5. Proof of Proposition 2

PROPOSITION 2 (Restated). *Suppose the provider chooses $\pi$ such that*

$$\bar{\pi}\big(p^*(\theta)\big) = r^*(\theta), \quad \textit{for all } \theta.$$

*Then, for a user of type $\theta$, bidding $b = p^*(\theta)$ for an instance with allocation $w^*(\theta)$ is a weakly dominant strategy. The resulting interruption probability equals $r^*(\theta)$, and the user pays $p^*(\theta)$ when active.*

*Proof.* We show that, given knowledge of the price distribution $\pi$, any bid $b \neq p^*(\theta)$ yields weakly lower expected utility for a user of type $\theta$ than bidding $b = p^*(\theta)$.

First consider the case in which the user bids exactly $b = p^*(\theta)$. By construction of the distribution $\pi$, the interruption probability satisfies

$$\bar{\pi}\big(p^*(\theta)\big) = r^*(\theta),$$

where $\bar{\pi}(b) = \Pr(p \geq b)$. Hence

$$\Pr\big(p \leq p^*(\theta)\big) = 1 - \bar{\pi}\big(p^*(\theta)\big) = 1 - r^*(\theta).$$

In steady state, the job remains active with probability $1 - r^*(\theta)$ and is interrupted with probability $r^*(\theta)$. When active, the user pays the bid price $p^*(\theta)$; when interrupted, the user pays nothing. Therefore the induced allocation and payment exactly replicate the optimal contract $\big(w^*(\theta), r^*(\theta), p^*(\theta)\big)$.

Now consider a deviation.

**Case 1:** $b > p^*(\theta)$**.** When $p \geq b$, the job is inactive, as before. When $p < p^*(\theta)$, the outcome coincides with bidding $p^*(\theta)$. However, when

$$p^*(\theta) \leq p < b,$$

the job remains active but the user pays the higher bid price $b > p^*(\theta)$. This strictly increases payment without improving service quality beyond the optimal contract. Since $\big(w^*(\theta), r^*(\theta), p^*(\theta)\big)$ maximizes expected utility for type $\theta$, any bid $b > p^*(\theta)$ yields weakly lower payoff.

**Case 2:** $b < p^*(\theta)$**.** In this case, the interruption probability increases because

$$\bar{\pi}(b) > \bar{\pi}\big(p^*(\theta)\big) = r^*(\theta).$$

Thus the job is interrupted more frequently than under the optimal contract. Although the user pays a lower price when active, the higher interruption probability reduces expected utility. Again, since $\big(w^*(\theta), r^*(\theta), p^*(\theta)\big)$ is utility-maximizing for type $\theta$, any bid $b < p^*(\theta)$ yields weakly lower payoff.

Therefore, bidding $b = p^*(\theta)$ weakly dominates all other bids. By selecting $\pi$ such that $\bar{\pi}\big(p^*(\theta)\big) = r^*(\theta)$ for all $\theta$, the provider replicates the second-best allocation of the direct mechanism. □

### A.2.6. Robustness Checks

We conduct several robustness checks after characterizing the optimal direct mechanism in Theorem 2. This includes robustness of the optimal throttling strategies under alternative contract specifications. For example, those alternative specifications in which users specify the service tier and the provider then assigned computing stations; or providers impose tier-dependent per-use price.

The following corollary is a direct result of the optimal direct mechanism and the revelation principle (Myerson 1981, Rochet 1985).

COROLLARY EC.1 **(Robustness to Contract Specification)**. *Consider the monopoly environment with private types, incentive compatibility (IC), individual rationality (IR), and capacity feasibility (FC). Let* $(w^*(\theta), r^*(\theta), p^*(\theta))$ *denote the optimal direct mechanism and the profit level characterized in Theorem 2 and* $\Pi^*$ *denote the optimal profit level.*

*Any alternative static contract design that satisfies the same set of constraints, including mechanisms in which users choose only service tiers while the provider assigns computing units or imposing tier-dependent*

*pricing, yields weakly lower profit for the provider and weakly lower total surplus relative to the optimal direct mechanism.*

*Proof.* Let $\mathcal{M}$ denote the set of all direct mechanisms $\{w(\theta), r(\theta), p(\theta)\}$ for all $\theta$. that satisfy (IC), (IC), and capacity feasibility constraints (FC). By construction, problem (P) solves

$$\Pi^* \equiv \sup_{M \in \mathcal{M}} \Pi(M).$$

Now consider any alternative static mechanism $\widetilde{M}$, possibly indirect. For example, a mechanism in which users select service tiers while the provider assigns computing units ex post, is equivalent to imposing restriction on selectable service tier $r(\theta) = \tilde{r}$ for some $\tilde{r} \in [0, 1]$. A mechanism that imposes tier-dependent per-use price is the same as imposing additional constraint that only allows $p(\theta)$ varies along delay dimensionality $p(\kappa)$ while restricting along willingness to pay $\nu$.

By the Revelation Principle, for any incentive-compatible indirect mechanism $\widetilde{M}$, there exists an outcome-equivalent direct mechanism $M' \in \mathcal{M}$ that induces truthful reporting and generates identical allocation and payments in equilibrium.

Thus, any implementable static mechanism corresponds to some element of $\mathcal{M}$.

Let $\widetilde{\mathcal{M}} \subseteq \mathcal{M}$ denote a restricted class of mechanisms (e.g., mechanisms that impose functional restrictions on $w(\theta)$ or $r(\theta)$). Since $\widetilde{\mathcal{M}} \subseteq \mathcal{M}$, it follows directly that

$$\sup_{M \in \widetilde{\mathcal{M}}} \Pi(M) \leq \sup_{M \in \mathcal{M}} \Pi(M) = \Pi^*.$$

Hence, no restricted static mechanism can strictly outperform the optimal direct mechanism under the same IC, IR, and feasibility constraints. Equality holds only if the restricted mechanism class contains an allocation rule that is outcome-equivalent to $(w^*, r^*, p^*)$. □

## A.3 Proofs of Major Results under Imperfect Competition

Our main finding in the benchmark model is that, in a monopoly cloud market, a profit-maximizing platform optimally engages in strategic throttling by reducing computing speed or increasing interruption risk for low willingness-to-pay or high delay-sensitive users. This throttling strategy implements a price–quality schedule $p(q)$ that induces a downward distortion in service quality (Mussa and Rosen 1978).

In this Appendix section, we analyze two extensions of the baseline model.

First, we allow users' reservation utilities to be stochastic when making participation decisions in the cloud market. This introduces imperfectly elastic demand and yields an endogenous market share function that lies strictly between zero and one. We show that the optimal throttling strategy—implemented through downward quality distortion—remains robust, provided that user heterogeneity (e.g., $\theta_H/\theta_L$ in a two-segment environment) is not excessively large.

Second, allowing for random participation also facilitates a natural extension to a duopolistic setting. In this environment, two competing cloud platforms simultaneously choose price–quality schedules $p(q)$, and user participation becomes endogenous across both providers. This allows us to embed our mechanism into a standard competition framework in which platforms jointly determine market shares and quality levels.

### A.3.1. Random Reservation Utility for Users' Participation

PROPOSITION 3 (Restated). *Consider a monopoly platform facing two vertical types $\theta_L < \theta_H$ and random reservation utilities $z$ with log-concave distribution $G$. Suppose the inverse hazard rate $H(u,\theta)$ is also convex in $u$. Let $q_H^*$ and $q_L^*$ be the socially optimal quality levels. There exists a finite threshold $\bar{\phi}$ in the preference ratio $\phi = \theta_H/\theta_L$ such that:*

1. *If $\phi < \bar{\phi}$ and , the optimal direct mechanism features strategic throttling, i.e., downward quality distortion for low-end market, i.e., $q_H = q_H^*$ and $q_L < q_L^*$.*
2. *If $\phi \geq \bar{\phi}$, the optimal direct mechanism provides efficient quality for both markets and thus no quality distortion arises in equilibrium, i.e., $q_H = q_H^*$ and $q_L = q_L^*$.*

*Proof.* Following our notation, a participating user's gross utility by choosing a virtual machine with quality $q$ is $\theta q - p(q)$. A user's indirect utility is defined as $u = \max_q\{\theta q - p(q)\}$, facing platform's choices of price-quality scheme $p(q)$. Thus, a user chooses to participate the market if $u - z \geq 0$.

We also allow that a cost $c \cdot q^2$ incurred to the provider to run the virtual machines with quality $q$. Note that our benchmark setting, where we assume the marginal cost of running virtual machines is zero $c = 0$, is a special case of this specification. Here, we assume $c = 1/2$ and the provider faces a convex cost function. The provider's profit margin is written as $\pi = p - q^2/2$. The total surplus (social welfare) is thus:

$$\Pi(q,\theta) \triangleq u + \pi = \theta q - q^2/2, \qquad \text{for } \theta \in \{\theta_L, \theta_H\},$$

conditional on users participate the market. Observe that the efficient (first best) quality allocation that maximizes the total surplus is just $q^*(\theta) = \theta$.

Without the loss of generality, we again consider the direct mechanism under the revelation principle (Myerson 1986). To simplify the notation, the direct mechanism takes the form $\{q_\theta, p(\theta)\}$ for $\theta = L, H$. To be more precise, we consider the direct mechanisms provided that users decide to participate the market under the random reservation utility, given the market share function defined earlier $M(u,\theta)$. We write $u_\theta$ as the indirect utility of a user who truthfully reports his vertical preference type that determines user's participation decisions, and $M(u_\theta,\theta)$ thus denotes the probability of type-$\theta$ users who participate the markets. We follow the duality approach of Armstrong and Vickers (2001) to model providers supplying utility to the users.[4] Therefore, the monopoly provider chooses $\{q_L, u_L\}$ and $\{q_H, u_H\}$ to maximize the expected profit:

$$M(u_L,\theta_L)(\Pi(q_L,\theta_L) - u_L) + M(u_H,\theta_H)(\Pi(q_H,\theta_H) - u_H),$$

[4]The duality approach studies optimal contract designs in consumer's utility space, where the firm designs a menu of contracts to differentiate between types of consumers and extract maximum surplus while ensuring certain participation and incentive compatibility constraints. For details, please refer to Armstrong and Vickers (2001).

subject to the two incentive compatibility conditions: the upward incentive compatibility (UIC) that prevents the low-type users from choosing service contracts that are for high-type users; and the downward incentive compatibility (DIC) that prevents the high-type users from choosing the low-type contracts.

$$\theta_L q_L - p_L \geq \theta_L q_H - p_H \quad \Longleftrightarrow \quad u_H - u_L \leq q_H(\theta_H - \theta_L), \tag{UIC}$$

$$\theta_H q_H - p_H \geq \theta_H q_L - p_L \quad \Longleftrightarrow \quad u_H - u_L \geq q_L(\theta_H - \theta_L). \tag{DIC}$$

Let $\Delta\theta \triangleq \theta_H - \theta_L$, $\Delta\Pi \triangleq \Pi(q_H, \theta_H) - \Pi(q_L, \theta_L)$, $\Delta\pi = \pi_H - \pi_L$ and define the profit margins on each market $\pi_i \triangleq p_i - q_i^2/2 = \Pi(q_i, \theta_i) - u_i$ for $i = L, H$. The incentive compatibility conditions above can be summarized as:

$$q_H \Delta\theta \geq u_H - u_L \geq q_L \Delta\theta, \text{ or equivalently} \tag{7}$$

$$\Delta\Pi - q_H \Delta\theta \leq \pi_H - \pi_L \leq \Delta\Pi - q_L \Delta\theta. \tag{8}$$

In our benchmark model where $z = 0$, between the two incentive compatibility conditions, the downward incentive compatibility always binds. This yields a downward quality distortion scheme, i.e., the service throttling strategy in the equilibrium. In the other somewhat extreme case of duopolistic Bertrand competition, profit margins are driven to be zero, and thus quality distortion disappears while neither UIC nor DIC binds in equilibrium. The presence of random reservation utility for user's participation gives rise to the outcomes of quality allocation that lie between these two cases. The optimal choice of quality distortion depends on the shape of joint probability of $(\theta, z)$ (or equivalently, the inverse hazard rate $H(u, \theta)$) and the measure of users' vertical preference heterogeneity $\theta_H/\theta_L$.

To be more specific, if:

1. The inverse hazard rate $H$ is also nondecreasing with vertical preference type $\theta$,
2. Vertical and horizontal preference parameters $(\theta, z)$ are independently distributed, and $H$ is convex.
3. Users' vertical type heterogeneity $\theta_H/\theta_L$ is greater than a fixed threshold $1 + 2H'(0)$;

Quality provision is not distorted but in all other cases, quality is *downward distorted* and optimality thus admits the service throttling strategies.

The first condition ensures that the UIC is always slack in equilibrium such that if there is quality distortion, it must be downward distortion. To see this, first observe that optimal profit margins is implicitly given by the classical monopoly pricing formula:

$$\pi_i = \Pi(q_i, \theta_i) - u_i = \frac{M(u_i, \theta_i)}{M_u(u_i, \theta_i)} \triangleq H(u_i, \theta_i)$$

The reason is that if vertical preference type $\theta_i$ were observable, the provider solves the classical monopoly pricing problem:

$$\max_{q_i, u_i} M(u_i, \theta_i)(\Pi(q_i, \theta_i) - u_i), \qquad \text{for } i = L, H$$

The first-order condition with respect to quality $q_i$ yields $q_i = \arg\max_q \Pi(q, \theta_i) = \theta_i q - q^2/2 = \theta_i$. The other first-order condition with respect to $u_i$ gives $M_u(u_i, \theta_i)\pi_i - M(u_i, \theta_i) = 0$, which yields that $\pi_i = H(u_i, t_i)$. Therefore,

$$\pi_i = H(u_i, \theta_i) = H(\theta_i^2/2 - \pi_i, \theta_i)$$

Suppose that the UIC binds at optimum. We have $\Delta\pi = \Delta\Pi - q_H\Delta\theta$ from the (UIC) condition or left-hand side of (8). In order to prevent low-type users choose high-type contracts as UIC binds, necessary conditions require that quality over-provision: $q_L = \theta_L$ and $q_H \geq \theta_H$. The ride-hand side of the binding UIC is maximized at $q_H = \theta_L$ at the value of $-(\Delta\theta)^2/2$, and therefore:

$$\Delta\pi = \Delta\Pi - q_H\Delta\theta \leq -(\Delta\theta)^2/2 < 0$$

Since $\pi_i = H(u_i, \theta_i)$ is nondecreasing in both $u$ and $\theta$, $\Delta\pi \geq 0$, we conclude with a contradiction. Thus, in the equilibrium, only DIC (or downward quality distortion) can be relevant and thus $q_H = q_H^* = \theta_H$ and $q_L \leq q_L^* = \theta_L$.

The second condition on the shape of the joint probability distribution of $(u, \theta)$ helps to completely characterize the equilibrium outcomes. To make our analysis more tractable, we further require that users' vertical and horizontal preferences are independent. This implies that:

$$M(u, \theta) = G(u)F(\theta), \quad \text{and moreover,} \quad H(u) = G(u)/g(u).$$

Thus $H(u)$ is completely independent of vertical preference parameter $\theta$. We can define the profit margin $\pi(\Pi)$ at the total surplus $\Pi$ as $\pi(\Pi) \triangleq \arg\max_\pi\{\pi \cdot M(\Pi - \pi, \theta)\}$. The log-concavity of $G(u)$ allows that $\pi(\Pi)$ can be uniquely defined by its first-order condition:

$$\pi(\Pi) = H(\Pi - \pi(\Pi)). \tag{9}$$

The convexity of $H$ imposes a property that the proportion of extracted surplus (i.e., the profit out of the total surplus) $\pi(\Pi)/\Pi$ is nondecreasing in $\Pi$. Like the monotonicity of $H$, this can be a plausible condition in public cloud market if we believe that the mark-up (defined as profit divided by price) is nondecreasing in service quality in cloud market.

With the third condition on the vertical preference heterogeneity, we establish the robustness of service throttling strategy: when preference heterogeneity is larger than a cut-off threshold, i.e., $\theta_H/\theta_L > 1 + 2H'(0)$, qualities are not distorted and provided at the efficient level $q_i = q_i^* = \theta_i$, while in all other cases, service qualities are downward distorted.

To see this, for $\theta_H/\theta_L$ and $\Pi_i$ such that DIC is slack (i.e., no quality distortion), necessary conditions require that $q_i = q_i^* = \theta_i$. Let $\Pi(q_L, \theta_L) = \Pi_L$ and $\phi \triangleq \theta_H/\theta_L$, we have $\Pi(q_H, \theta_H) = \theta_H^2/2 = (\theta_H/\theta_L)^2\Pi_L = \phi^2\Pi_L$. A slack DIC from the right-hand side of the inequalities (8) after replacing $\Pi_H$ and $q_L$, implies that

$$\{\pi(\phi^2\Pi_L)\} - \pi(\Pi_L)\}/\Pi_L \leq (\phi - 1)^2. \tag{10}$$

$\pi$ is convex because $H$ is convex,[5] and thus the left-hand is a nondecreasing function of $\Pi_L$. This means that there exists a function of $\bar{\Pi}(\phi)$ such that DIC is binding for all $\Pi_L \geq \bar{\Pi}(\phi)$ and DIC slacks for $\Pi_L < \bar{\Pi}(\phi)$. Therefore, for an increasing gain of participating the market $\Pi_L$, it is more likely to have a binding DIC and thus quality distortion. We now find the boundary case for a slack DIC when $\Pi_L$ reduces to zero. By L'Hôpital's rule, the left-hand side of (10) converges to $\pi'(0)(\phi^2 - 1)$.

Since $\pi(\Pi)$ is uniquely given by (9), differentiating (9) at $\Pi = 0$ yields $\pi'(0) = H'(0)/(1 + H'(0))$. Replacing $\pi'(0)$, we know that DIC is slacking for $\Pi_L$ close to zero when

$$\frac{H'(0)}{1 + H'(0)}(\phi^2 - 1) < (\phi - 1)^2,$$

which yields that when $\phi > 1 + 2H'(0)$, there exists a strictly positive $\bar{\Pi}(\phi) > 0$ such that DIC can be slack for some $\Pi_L$, where quality is not distorted and $q_i = q_i^* = \theta_i$ for both markets $i = L, H$. In all other cases, $\Pi_L \geq \bar{\Pi}(\phi)$ or $\phi \leq 1 + 2H'(0)$, DIC binds as in our benchmark case, and thus service throttling remains as the optimal strategy for a monopoly service provider. □

In order to explicitly capture user's horizontal preference, we can add a scale parameter $\gamma > 0$ for user's horizontal preference heterogeneity. Let users' reservation utility now to be $\gamma z$ with $\gamma$ ranges between 0 and 1 and $z$ remains the random variable with distribution $G$. With $\gamma$, both the benchmark model ($\gamma = 0$) and the previous case with monopoly provider ($\gamma = 1$) because special cases of this setup. With duopolistic competition, $\gamma$ can be used to represent users' preferential attachment to platforms in the horizontal space of platform's product differentiation.

Since $\gamma$ is simply a scaling factor, our observation that service throttling remains to be robust in equilibrium as long as user's vertical preference heterogeneity is not too large. The horizontal parameter $\gamma$ captures the effect of the random participation onto the platform's optimal strategy in quality allocation. It can be shown that when $\gamma$ is larger than a threshold, and $\theta_H/\theta_L$ is large enough, qualities will not be distorted in equilibrium.

To see this, the monopoly profit function in each market is $B(u, q, \theta, \gamma) = (\theta q - q^2/2 - u)G(u/\gamma)$ since the reservation utility is now $\gamma z$ instead of $z$. Due to the fact that the profit function is linear quadratic, the profit function has the property that:

$$B(u, q, \theta, \gamma) = \gamma \cdot B(q/\sqrt{\gamma}, u/\gamma, \theta/\sqrt{\gamma}, 1).$$

Therefore, the platform will choose no quality distortion if $\Pi_L < \bar{\Pi}(\phi)$, and this inequality becomes:

$$(1/2) \cdot (\theta_L/\sqrt{\gamma})^2 < \bar{\Pi}\left((\theta_H/\sqrt{\gamma})/(\theta_L/\sqrt{\gamma})\right),$$

which yields that $\gamma > \Pi_L/\bar{\Pi}(\phi)$. Therefore, when the scale parameter $\gamma$ is large enough (and so is the vertical preference heterogeneity), the monopoly platform may find it more effective to provide efficient quality allocation, given a volatile market demand due to high randomness of users' participation.

[5] Note that $\Pi = \pi(\Pi) + H^{-1}(\pi(\Pi))$ because of (9). $\pi$ is convex if and only if $H$ is convex, given a linear left-hand in $\Pi$.

### A.3.2. Throttling Strategies under Imperfect Competition

THEOREM 3 (Restated). *Consider two symmetric platforms $j = 1, 2$ competing in price–quality offering on a Hotelling line with switching-cost parameter $\gamma > 0$ and two vertical types $\theta_L < \theta_H$. Let $u_j(\theta)$ denote the equilibrium indirect utility for type $\theta$ on platform $j$. Then,*

1. ***Local monopolies*** *(high switching costs). If $\gamma \geq u_1(\theta_H) + u_2(\theta_H)$, users of both types consider only their nearest platform or non-participation. Each platform behaves as a monopolist on its side of the market. Strategic throttling emerges in the equilibrium when vertical preference heterogeneity is not sufficiently large, as in Proposition 3.*
2. ***All-out competition*** *(low switching costs). If $\gamma < u_1(\theta_L) + u_2(\theta_L)$, the two platforms fully cover the market. There exists a symmetric equilibrium in which both platforms provide efficient quality for both types, and there is no quality distortion.*
3. ***Mixed regime*** *(intermediate switching costs). Between these two regimes where $u_1(\theta_L) + u_2(\theta_L) \leq \gamma < u_1(\theta_H) + u_2(\theta_H)$, high-type users receive efficient quality, while low-type users face two local monopolies in their captive markets in which strategic throttling remains optimal when vertical heterogeneity is not too large.*

*Proof.* Let $u_j(\theta) = \max_q \theta q - p_j(q)$ be the indirect utility user obtains from platform-$j$ service. From the previous proof for Proposition 3, we know the high-type users gain (weakly) higher indirect utility than the low types from either platform, i.e., $u_j(\theta_H) \geq u_j(\theta_L)$, simply because of the the downward incentive compatibility condition (DIC) that prevents the high types from switching to the low-type contract (where by definition both $q_L \geq 0$ and $\theta_H - \theta_L > 0$).

Thus, for the switching cost $\gamma$ taking values greater than $u_1(\theta_H) + u_2(\theta_H)$, lower than $u_1(\theta_L) + u_2(\theta_L)$, or in between these two cutoffs, we prove the existence for three regimes of competition: (1) local monopolies, (2) all-out competition, and (3) mixed regime in between.

*Local Monopolies.* When $\gamma$ is large enough (when $\gamma \geq u_1(\theta_H) + u_2(\theta_H)$ for $\theta = \theta_L, \theta_H$), e.g., highly loyal users or high switching cost for workloads between platforms in equilibrium, users of both type $\theta = \theta_L, \theta_H$ only consider between two options of either choosing from the nearest platform or not participating the market at all.

To see the effect of $\gamma$ on equilibrium outcome, note that when user is indifferent choosing from either platform $j = 1, 2$ can be identified by the equation:

$$u_1(\theta) - \gamma z = u_2(\theta) - \gamma(1 - z),$$

this identifies the marginal user of vertical preference $\theta = \theta_L, \theta_H$ by the horizontal location:

$$\tilde{z}(\theta) = \frac{\gamma + u_1(\theta) - u_2(\theta)}{2\gamma}.$$

If the marginal users prefer participating the markets than not, Platform 1's market share is just $\tilde{z}$ while the Platform 2's market share is the remaining measure $1-\tilde{z}$.

Local monopolies arise for the latter case, when the net utility for the marginal users from participating the market might be dominated by not participating at all. This gives the following inequality (since platforms are symmetric, taking Platform 1 for example):

$$u_1(\theta) - \gamma\tilde{z}(\theta) \leq 0,$$

replacing $\tilde{z}$ yields that $\gamma \geq u_1(\theta) + u_2(\theta)$ for both markets $\theta = \theta_L, \theta_H$.

Since $u_j(\theta_H) \geq u_j(\theta_L)$, we recover our condition.

*All-out Competition.* To show the existence of this symmetric efficient quality allocation for the two competing platforms, we need to show that either UIC or DIC constraint binds for either (symmetric) platform, e.g., Platform 1. Platform 1 maximizes its total profits for both market $L$ and $H$:

$$B_1(u_L, u_H) = G(\tilde{z}(\theta_L))F(\theta_L)\{\Pi(q_L, \theta_L) - u_L\} + G(\tilde{z}(\theta_H))F(\theta_H)\{\Pi(q_H, \theta_H) - u_H\},$$

subject to the incentive compatibility constraints, as in (7):

$$q_L\Delta\theta \leq u_H - u_L \leq q_H\Delta\theta$$

The Lagrangian function for this constrained optimization program becomes

$$\mathcal{L} = B(u_L, u_H) + \lambda_u(q_H\Delta\theta - u_H + u_L)\lambda_d(u_H - u_L - q_L\Delta\theta),$$

where $\lambda_u$ and $\lambda_d$ denote the Lagrangian multipliers for UIC and DIC constraint, respectively. Observe that $\tilde{z}(\theta_L) = 1/2 + (u_1(\theta_L) - u_2(\theta_L))/2\gamma$ and $\pi_L \triangleq \Pi(q_L, \theta_L) - u_L$ by definition, the first-order conditions with respect to $u_L$ and $u_H$ produce:

$$-G(\tilde{z}(\theta_L))F(\theta_L) + g(\tilde{z}(\theta_L))F(\theta_L)\pi_L/2\gamma - \lambda_d + \lambda_u = 0,$$

$$-G(\tilde{z}(\theta_H))F(\theta_H) + g(\tilde{z}(\theta_H))F(\theta_H)\pi_H/2\gamma + \lambda_d - \lambda_u = 0.$$

First consider the case when UIC binds and DIC slacks, thus $\lambda_u > 0$ and $\lambda_d = 0$, then we have $\pi_L < 2\gamma \cdot G(\tilde{z}(\theta_L))/g(\tilde{z}(\theta_L)) < \pi_H$. Given the increasing hazard rate $H$ and following the same arguments as in Proposition 3, UIC binds must give rise to $\Delta\pi = \pi_H - \pi - L < 0$, thus a contradiction. The same arguments run for the opposite case when UIC slacks and DIC binds when $\lambda_u = 0$ and $\lambda_d > 0$. Therefore, either incentive compatibility constraint binds at optimum, we have the case $q_L = q_L^* = \theta_L$ and $q_H = q_H^* = \theta_H$ for Platform 1. Because the platforms are symmetric, we have Platform 2 provides the exactly same level of qualities in its markets.

*Mixed Regime.* When $\gamma$ takes intermediate values between $u_1(\theta_L) + u_2(\theta_L)$ and $u_1(\theta_H) + u_2(\theta_H)$, note that the high-type users are always provided with efficient qualities due to the slackness of (UIC). Whether low-type quality is distorted, as in the monopoly case, depends on the size of vertical heterogeneity and the switching cost: when vertical preference heterogeneity is not too large, service throttling remains prevalent in the low-end market. □

## References

Albeverio S (1986) *Nonstandard Methods in Stochastic Analysis and Mathematical Physics* (Elsevier Science), ISBN 9780080874418, URL https://books.google.com/books?id=M_5RB-YGjZsC.

Armstrong M, Vickers J (2001) Competitive price discrimination. *The RAND Journal of Economics* 32(4):579–605, ISSN 07416261, URL http://www.jstor.org/stable/2696383.

Cutland N (1988) *Nonstandard Analysis and Its Applications*. Cambridge Topics in Mineral Physics & Chemistry (Cambridge University Press), ISBN 9780521351096, URL https://books.google.com/books?id=9mkW51DNABAC.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 5(2):79–141.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Operations research* 29(3):567–588.

Harchol-Balter M (2013) *Performance modeling and design of computer systems: queueing theory in action* (Cambridge University Press).

Khan M, Sun Y (1999) Non-cooperative games on hyperfinite loeb spaces1the authors respectfully dedicate this work to professor donald j. brown on the occasion of his sixtieth birthday.1. *Journal of Mathematical Economics* 31(4):455–492, ISSN 0304-4068, URL http://dx.doi.org/https://doi.org/10.1016/S0304-4068(98)00031-7.

Loeb PA (1975) Conversion from nonstandard to standard measure spaces and applications in probability theory. *Transactions of the American Mathematical Society* 211:113–122, ISSN 00029947, URL http://www.jstor.org/stable/1997222.

McAfee R, McMillan J (1988) Multidimensional incentive compatibility and mechanism design. *Journal of Economic Theory* 46(2):335 – 354, ISSN 0022-0531, URL http://dx.doi.org/http://dx.doi.org/10.1016/0022-0531(88)90135-4.

Mussa M, Rosen S (1978) Monopoly and product quality. *Journal of Economic Theory* 18(2):301–317, ISSN 0022-0531, URL http://dx.doi.org/https://doi.org/10.1016/0022-0531(78)90085-6.

Myerson RB (1981) Optimal Auction Design. *Mathematics of Operations Research* 6(1):58–73, URL https://ideas.repec.org/a/inm/ormoor/v6y1981i1p58-73.html.

Myerson RB (1986) Multistage Games with Communication. *Econometrica* 54(2):323–58, URL https://ideas.repec.org/a/ecm/emetrp/v54y1986i2p323-58.html.

Robinson A (2016) *Non-standard analysis* (Princeton University Press).

Rochet JC (1985) The taxation principle and multi-time hamilton-jacobi equations. *Journal of Mathematical Economics* 14(2):113–128, ISSN 0304-4068, URL http://dx.doi.org/https://doi.org/10.1016/0304-4068(85)90015-1.

Sun Y (2006) The exact law of large numbers via fubini extension and characterization of insurable risks. *Journal of Economic Theory* 126(1):31–69, ISSN 0022-0531.

Whitt W (2002) *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues* (Springer New York), ISBN 9780387953588, URL https://books.google.com/books?id=imq84GrwDLYC.