Racing Queues with Strategic Servers

Andrew E. Frazelle
Jindal School of Management, The University of Texas at Dallas
andrew.frazelle@utdallas.edu

Yueyang Zhong
Management Science and Operations, London Business School
yzhong@london.edu

Abstract. We study *racing queues*, a type of multi-server queueing system in which strategic servers race against each other to complete jobs. Once any server completes the active job, that server earns a prize and all servers move to the next job in the shared queue. Each server chooses her service rate to maximize utility; serving faster wins more often but is more costly. Using exact analysis, we identify the unique symmetric equilibrium service rate in closed form across three regimes – overloaded, critically loaded, or underloaded – determined by the prize amount relative to marginal capacity cost. Notably, the equilibrium service rate can be non-monotonic in the number of servers. Under appropriate conditions, it first increases then decreases, revealing an "ideal" level of competition that maximizes individual server effort. Otherwise, reduced effort may completely eliminate any capacity gain from an additional server. We also fully characterize the equilibrium's intricate relationship with the degree of elasticity in the capacity cost function. Finally, we endogenize the prize amount. When job value is high, the optimal prize sustains a critically loaded equilibrium; otherwise, it induces an overloaded equilibrium. The profit-maximizing prize is systematically below the surplus-maximizing level, with relative surplus loss uniformly bounded above.

Key words: strategic queueing, strategic servers, racing queues, winner-take-all, cancel-on-complete, service operations
History: *This version:* May 20, 2026.

1. Introduction

In recent years, a growing stream of research endogenizes service rates in queueing systems by modeling servers as strategic agents who choose how fast to work, contrasting with traditional queueing models with exogenous service rates. However, while strategic-server queueing models typically assume that each job is assigned to exactly one server, many service environments exist where this assumption fails to hold.

A prominent example is cryptocurrency mining for block validation (Altman et al. 2020, Biais et al. 2019, Huberman et al. 2021). Each block of pending transactions is validated by miners who concurrently solve a cryptographic puzzle. Only the first to find a valid solution earns the block reward, and all others earn nothing for that block and immediately proceed to the next. Miners decide how much computing power to deploy – through hardware investment, electricity usage, and operational intensity – trading off increased odds of winning against higher costs. Similar dynamics also arise in software bug bounty programs such as those run via the HackerOne platform (on behalf of software providers), which has awarded more than $380 million to white-hat hackers (HackerOne 2026), as well as in fundamental research competitions such as the Millennium Prize Problems (Clay Mathematics Institute 2000). In these settings, multiple independent agents race to be the first to identify a software vulnerability or to solve an open mathematical problem, respectively, and the winner receives a monetary prize.

Thus motivated, we model service environments in which multiple servers simultaneously work on the same job, racing to be the first to complete it. The first server to finish earns the prize, at which point all servers immediately move on to the next job in the shared queue, regardless of their progress on the previous job; this service discipline is known as cancel-on-complete. Servers independently and non-cooperatively choose their service rates, trading off higher effort cost against a greater probability of winning each race. We coin the term *racing queue* to describe a queueing system that combines this winner-take-all, cancel-on-complete discipline with endogenous service rates.

In our context of racing queues, we address three primary research questions. First, how does racing among strategic servers affect equilibrium service rates and aggregate capacity? Unlike classical multi-server queues, where adding servers mechanically expands capacity, adding servers in a racing queue also changes each server's incentives by diluting the winning probability. Each server's equilibrium effort is governed by the balance between the marginal gain in winning probability from raising one's speed and the marginal cost of faster service. Adding competitors may strengthen or weaken this marginal gain: when strong, it induces an "arms race" in effort; when weak, servers slow down and free-ride on the larger pool. The net effect on individual effort and aggregate capacity is therefore not a priori monotone.

Second, how does the elasticity of capacity cost with respect to service rate shape equilibrium behavior? The cost elasticity governs how steeply marginal cost rises with service rate. A higher elasticity makes each incremental unit of speed more expensive, which would seem to suppress effort. This intuition applies, however, only when the service rate is sufficiently high. At low service rates, increasing the elasticity can actually decrease marginal cost, enabling higher equilibrium service rates. Even absent racing, these opposing forces can push the optimal service rate in either direction. In a winner-take-all race, the problem is further complicated by the dependence of the winning probability on all servers' rates, such that changes in cost elasticity alter not only each server's marginal cost but also the competitive return to speed.

Third, when a system designer can choose the prize amount per job, how should prizes be set to achieve system-level performance goals? A higher prize raises the value of winning and can induce faster service, but it also increases the designer's cost per completed job. Whether a higher prize improves system performance therefore depends on whether the additional throughput it generates justifies the added expense. The answer may vary with whose payoffs the designer accounts for. A designer maximizing its own profit treats prize payments as costs, whereas a planner maximizing total surplus regards them as transfers within the system. Do these two objectives agree on the optimal prize, and if not, how large is the surplus loss?

To the best of our knowledge, we are the first to study many-server queueing systems in which strategic servers race to complete jobs under a cancel-on-complete discipline. We consider a model where jobs arrive at rate λ to a system of N racing servers, and the first to complete a job earns the prize p. Each server chooses her service rate to maximize expected utility, trading off higher winning probability against effort cost: working faster increases the chance of earning the prize but entails greater cost. Our main contributions are as follows.

Equilibrium characterization (Section 3). Despite the analytical complexity of many-server queueing games, we show that a unique symmetric equilibrium service rate exists and admits an exact, piecewise closed-form expression (Proposition 1). The equilibrium falls into one of three service regimes. When the prize is sufficiently low, the equilibrium service rate is below λ/N, resulting in an overloaded system where the queue grows without bound. In this regime, the equilibrium rate depends only on the prize amount and the effort cost, independently of the arrival rate or the number of servers. The intuition is that the system is always busy, so the arrival rate is irrelevant; more competitors raise the job completion rate but proportionally lower each server's per-job winning probability, and we show that these effects exactly cancel, making each server's rate of earning prizes equal to her own service rate. When the prize is sufficiently high, servers operate above λ/N and the system is underloaded, meaning there is enough capacity to serve all arrivals. Now servers experience idle time, so working faster drains the queue sooner and partially offsets the benefit of a higher winning probability; the equilibrium service rate therefore depends on all primitives, including the arrival rate and the number of servers. Between these extremes, the equilibrium service rate sits exactly at the stability boundary λ/N (critically loaded), where aggregate capacity precisely matches demand. The thresholds separating the three regimes and the closed-form expressions within each case provide the analytical foundation for all subsequent results.

Effect of competition (Section 4). Each server's equilibrium service rate is decreasing in N for $N \geq 2$, but may jump upward from $N = 1$ to $N = 2$ (Proposition 2). We call this the *duopoly peak*: individual racing incentives are strongest not under intense many-server competition but in a head-to-head rivalry, where the sensitivity of winning probability to each server's effort is maximized. Notably, for some prize amounts, the jump can be strikingly large (e.g., a more than 10-fold increase in one of our numerical examples). In addition, aggregate service capacity is increasing in N, which is unsurprising, but more interestingly, it can be flat (Proposition 3). This occurs when both the N- and $(N + 1)$-server systems are critically loaded, so incumbents slow down just enough to offset the entrant, leaving total capacity unchanged. Unlike in classical multi-server queues, adding a server may therefore contribute zero marginal capacity.

Effect of capacity cost elasticity (Section 5). We model capacity cost as a polynomial function $c(\mu) = c_E \mu^q$, where $c_E > 0$ scales the cost level and the exponent $q > 1$ is the cost elasticity. We identify two channels through which cost elasticity affects the equilibrium service rate. The first is a marginal-cost channel operating within a fixed service regime. We establish that the equilibrium service rate rises with cost elasticity precisely when higher elasticity lowers marginal cost at the prevailing rate, and falls otherwise (Proposition 4). The second is a regime-switching channel, whereby the thresholds separating the three service regimes shift with cost elasticity, potentially moving the equilibrium across regimes. Combining these effects yields a complete monotonicity characterization (Propositions 5 and 6). When the prize is modest relative to the cost scale c_E, the equilibrium service rate is increasing in cost elasticity, as it stays low enough that higher elasticity consistently lowers marginal cost. When the prize is large and demand is high, the equilibrium initially lies

in a region where higher elasticity raises marginal cost, so effort first declines in cost elasticity; but as cost elasticity continues to grow, the equilibrium enters a region where higher elasticity lowers marginal cost, and effort recovers.

Prize design (Section 6). Finally, we examine the role of the prize p as a design variable. Each completed job generates a value $r > 0$ for the designer, who pays a prize p to the winning server of each race. A higher prize encourages faster service but reduces the designer's margin per job. We characterize the optimal prize under two objectives: profit maximization, where the designer maximizes net profit after prize payments (Proposition 7), and surplus maximization, where the designer also accounts for servers' utility (Proposition 8). Under both objectives, when job value is high, the optimal prize induces a critically loaded equilibrium where throughput matches demand; when job value is low, the optimal prize sustains an overloaded equilibrium. A key insight is that the profit-maximizing prize is always weakly below the surplus-maximizing prize (Proposition 9). The reason is that a profit-maximizing designer treats every dollar paid to servers as a cost, whereas a surplus-maximizing planner recognizes these payments as transfers and penalizes only the additional effort cost that a higher prize induces. The profit-maximizing designer therefore systematically under-prices. Despite this systematic discrepancy, we derive an upper bound on the relative surplus loss from profit-maximizing under-pricing that is purely a function of the cost elasticity q, and show that this loss never exceeds $1 - 2/e \approx 26.4\%$. These results offer guidance for blockchain protocol designers managing energy waste, bug-bounty platforms calibrating prizes to balance timeliness and cost, and research contest organizers structuring incentives to elicit productive effort.

We briefly describe the notational conventions used in this paper. We use bold letters to indicate vectors. We let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ be the vector of service rates for N servers. Similarly, for $i \in \{1, \ldots, N\}$, we let $\boldsymbol{\mu}_{-i}$ be the vector $\boldsymbol{\mu}$ excluding the i-th entry. Additionally, we let $\mathbf{0}$ denote a vector of zeros, where the length is clear from context. Finally, we use "increasing" and "decreasing" in the weak sense unless otherwise noted.

2. Literature Review

Our work lies at the intersection of several research streams: strategic servers in queues, collaboration among servers, redundancy scheduling with cancel-on-complete discipline, and the contest theory, R&D race, and blockchain mining literatures. We review each in turn.

2.1. Strategic Servers in Queues

A growing body of work endogenizes service rates by modeling servers as strategic agents who choose how fast to work. Early contributions include Gilbert and Weng (1998), who show that nonconsolidating queue structures can create stronger capacity-investment incentives for self-interested servers than pooled queues, and Hopp et al. (2007), who model discretionary task completion and demonstrate that servers' judgment about time allocation introduces quality as a fourth variability buffer alongside capacity, inventory, and time. Cachon and Zhang (2007) study performance-based demand allocation to two strategic servers who choose

their processing rates, showing that well-designed allocation policies can align incentive provision with capacity utilization. Anand et al. (2011) analyze the quality–speed tradeoff in customer-intensive services, finding that strategic servers facing more competition can paradoxically raise prices and slow down.

Building on these foundations, Gopalakrishnan et al. (2016) study routing and staffing in an $M/M/N$ queue where servers trade off effort cost against idle-time value, showing that equilibrium requires a quality-driven regime with staffing that strictly exceeds the standard square-root policy. Zhan and Ward (2019) extend this line of inquiry by jointly optimizing staffing, routing, and piece-rate payment to balance speed and quality, identifying four economically optimal service regimes under fluid scaling. Armony et al. (2021) compare pooled and dedicated queues with strategic servers, demonstrating that pooling can dilute individual effort incentives through shared (rather than individual) customer ownership. Zhong et al. (2025b) analyze strategic servers in a finite-buffer $M/M/N/k$ setting and obtain equilibrium characterizations in a many-server asymptotic regime, while Zhong et al. (2025a) study the joint equilibrium when both customers and servers behave strategically. Büke et al. (2025) characterize asymmetric equilibria among heterogeneous strategic servers in heavy traffic. Relatedly, Ibrahim (2018) studies queueing systems with servers who choose whether or not to be available for work, and Yuan (2025) studies flexible capacity management when servers face shortage constraints. Strategic servers in systems that can be modeled as queues also find a home in recent research on gig-economy workers; see, for instance, Dong and Ibrahim (2020), Hu et al. (2025b, section 4) and references therein.

A common feature of these papers is that each job is assigned to a single server. Incentives arise from piece-rate pay, idle-time valuation, or customer-ownership effects, but servers do not compete on the same job. In contrast, our model places all servers on the same job simultaneously. The winner-take-all payoff structure induces a probabilistic race for each job, fundamentally altering the strategic environment. Effort incentives therefore hinge on the marginal impact on winning probability – rather than throughput per se – leading to qualitatively different outcomes.

2.2. Collaboration Among Servers

Another related stream studies queueing systems in which multiple servers work on the same job cooperatively. Andradóttir and Ayhan (2005) analyze throughput maximization in tandem lines with flexible servers who can be reassigned across stations. Andradóttir et al. (2011) introduce a model of "synergistic" servers, whose combined service rate can be super-additive, so that joint effort exceeds the sum of individual rates. Extending this framework to settings with quality uncertainty, Hu et al. (2025a) examine optimal server assignment when servers are error-prone and may need to redo work. Rosokha and Wei (2024) study cooperation in queueing systems where servers choose effort levels and are compensated based on group throughput, highlighting the tension between collective benefits and individual free-riding incentives; they complement their theoretical analysis with laboratory experiments. Chen et al. (2025) investigate incentive mechanisms for voluntary

resource pooling among servers, each operating an independent $M/M/1$ queue, using a token-based trading mechanism to facilitate capacity sharing.

In these papers, servers benefit from one another's effort, either directly through super-additive service rates or indirectly through shared throughput rewards. A server who works harder helps her co-workers. Our setting is the opposite: the reward structure is winner-take-all, so a server who works harder reduces every rival's probability of winning each race. In the collaboration literature, the central strategic tension is free-riding: each server would like others to bear the cost of effort while sharing in the collective reward. In our model, the central tension is competition for a fixed prize: each server chooses its speed by trading off the marginal gain in winning probability against the marginal capacity cost, and one server's gain in winning probability comes entirely at the expense of its rivals. This fundamental difference – effort as a shared benefit versus effort as a rival's loss – is what distinguishes the racing queue from the collaborative service literature.

2.3. Redundancy Scheduling and Cancel-on-Complete

Our model shares the cancel-on-complete service discipline with the redundancy scheduling literature. Gardner et al. (2016) provide the first exact analysis of queueing systems with redundant requests, showing that sending copies of a job to multiple servers and canceling upon the first completion can substantially reduce response times relative to join-the-shortest-queue and probabilistic splitting policies. Gardner et al. (2017) further study how many redundant copies are needed to achieve significant delay reduction, establishing performance bounds for the "redundancy-d" policy in which each job is sent to d of N servers. Beyond delay, Nageswaran and Scheller-Wolf (2022) examine the fairness implications of redundancy when only some customers may join multiple queues, while Chen et al. (2024) analyze a customer-driven multi-listing system with horizontally differentiated servers and study its welfare and pricing implications.

The key distinction lies in the role of service rates. In the redundancy scheduling literature, service rates are fixed and exogenous, and the focus is on how routing policies affect delay. In contrast, we endogenize service rates by embedding a strategic game within the same cancel-on-complete framework. As a result, system performance depends on servers' equilibrium service rates, where each server selfishly chooses how fast to work to maximize her own utility.

2.4. Contests and R&D Races

Our model also connects to two broader literatures – contest theory and R&D races – each of which studies settings where agents compete to be first but neither of which embeds such competition within a queueing system with congestion feedback.

In the racing queue, each server's probability of completing a job first is $\mu_i / \sum_j \mu_j$, which can be viewed as a Tullock contest success function (Tullock 1980). A central result in the contest-design literature is that equilibrium effort varies non-monotonically with the number of contestants (see the comprehensive treatment by Vojnović 2016). Fullerton and McAfee (1999) study the optimal number of contestants in a

research tournament. Unlike a racing queue, where the prize is awarded only once a task is completed, their tournament awards the prize based on a pure ranking. Among other implications, this means that if there is only one contestant, they optimally exert no effort and still win the prize. They show that the expected solution quality is often maximized with two contestants, showing an interesting echo of our finding about the effort-maximizing number of servers in a racing queue. However, different from their setting, effort is always required to win prizes in a racing queue, and we also identify cases where effort is maximized with only a single server (whereas in their case, a lone contestant never exerts any effort at all). Related contest mechanisms have also been documented in other applications, such as day-to-day project management (Dawande et al. 2019).

The closest classical antecedent to a single race in our model is the R&D race literature, in which firms invest to be the first to complete an innovation. Loury (1979) introduces a stochastic-race model with memoryless (exponential) innovation times and derives a free-entry equilibrium, a structure closely paralleling a single service event in our model. Terwiesch and Xu (2008) analyze innovation contests where a seeker solicits solutions from a pool of solvers, showing that larger solver populations reduce individual effort but can benefit the seeker through solution diversity. Unlike our work, these models analyze a single isolated race, without a stochastic arrival process generating successive contests or any queueing dynamics linking one race to the next.

The key distinction from both streams of literature is the queueing feedback loop in our model. In a classical contest or R&D race, there is always a prize to compete for. In a racing queue, by contrast, jobs arrive stochastically, so the fraction of time the system is nonempty – and servers can earn prizes – depends on how fast they collectively work. When aggregate service rates do not exceed the arrival rate, serving faster not only increases one's probability of winning each race but also the rate at which races arise. This feature is absent in the contest and R&D settings.

3. Model and Equilibrium Analysis

We consider a service system with $N \geq 1$ servers who race each other to complete each job. Examples include bitcoin miners racing to validate blocks of transactions, and teams of researchers racing to solve an open problem. Jobs arrive to the system via a Poisson process with rate $\lambda > 0$. The system maintains a single FCFS queue of waiting jobs. Whenever the system is nonempty, all N servers simultaneously work on the job at the head of the line. The job departs as soon as one server finishes it. The winning server receives a prize $p > 0$, and all other servers cancel their work on that job and immediately begin working on the next waiting job (if any). If the queue is empty, all servers idle until the next arrival.

Each server i chooses a service rate $\mu_i \geq 0$. Conditional on choosing μ_i, server i's processing time for any job is exponentially distributed with rate μ_i. We assume processing times are independent across servers for the same job. This assumption is natural in settings such as proof-of-work mining, where success occurs when

a random hash satisfies a target. The time until the first completion is then the minimum of N independent exponential random variables with rates given by $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$. Thus, whenever the system is nonempty, the service time (time from start of service to job completion) is also exponentially distributed, with rate equal to $\sum_{j=1}^{N} \mu_j$. Moreover, server i wins the race with probability $\mu_i / \sum_{j=1}^{N} \mu_j$.

Operating at rate μ_i incurs a capacity cost at rate $c(\mu_i)$, where $c(\cdot)$ is strictly increasing and differentiable with $c'(\mu_i) > 0$ for all $\mu_i > 0$. The capacity cost $c(\mu_i)$ can be interpreted as the instantaneous cost flow of maintaining the computing power, electricity, staffing, and other resources required to sustain service rate μ_i. Throughout, we treat servers as risk-neutral and focus on long-run average utility per unit time.

3.1. Server Utility and Symmetric Equilibrium

The above development implies that, despite the presence of multiple servers, the congestion dynamics of the system reduce to those of an $M/M/1$ queue with arrival rate λ and service rate $\sum_{j=1}^{N} \mu_j$. When the system is nonempty, jobs are completed at rate $\sum_{j=1}^{N} \mu_j$, and the queue evolves exactly like an $M/M/1$. We denote by $B(\sum_{j=1}^{N} \mu_j; \lambda)$ the long-run fraction of time the system is busy (i.e., nonempty).

If $\sum_{j=1}^{N} \mu_j > \lambda$, the $M/M/1$ is stable and the steady-state busy fraction is $B(\sum_{j=1}^{N} \mu_j; \lambda) = \lambda / \sum_{j=1}^{N} \mu_j$. If instead $\sum_{j=1}^{N} \mu_j \le \lambda$, the $M/M/1$ is critically loaded or overloaded and has no stationary distribution. In that case, the queue length drifts to infinity and the system is asymptotically always busy; accordingly, we set the long-run busy fraction to one. This convention matches standard usage in strategic-server models that permit overload and ensures that utility rates are well-defined even in unstable regimes. Formally, we define

$$B\left(\sum_{j=1}^{N} \mu_j; \lambda\right) = \begin{cases} 1, & \sum_{j=1}^{N} \mu_j \le \lambda, \\ \dfrac{\lambda}{\sum_{j=1}^{N} \mu_j}, & \sum_{j=1}^{N} \mu_j > \lambda. \end{cases} \tag{1}$$

Since servers either all work (when the system is nonempty) or all idle (when it is empty), all servers share the same busy fraction $B(\sum_{j=1}^{N} \mu_j; \lambda)$ regardless of heterogeneity in μ_i. A given server's service-rate choice affects her own busy time and all others' busy time in the same way through the aggregate service rate $\sum_{j=1}^{N} \mu_j$.

Server i earns a prize p whenever it completes a job first and wins a race. Conditional on the system being nonempty, jobs complete at rate $\sum_{j=1}^{N} \mu_j$, and server i wins each race with probability $\mu_i / \sum_{j=1}^{N} \mu_j$. Thus, the long-run earning rate of server i is

$$p \left(\sum_{j=1}^{N} \mu_j\right) \left(\frac{\mu_i}{\sum_{j=1}^{N} \mu_j}\right) B\left(\sum_{j=1}^{N} \mu_j; \lambda\right). \tag{2}$$

The expression simplifies via cancellation[1], yielding an especially transparent form: server i's earning rate equals its own service rate times the busy fraction (multiplied by the prize), that is, $p\mu_i B\left(\sum_{j=1}^{N}\mu_j;\lambda\right)$. Subtracting the capacity cost rate $c(\mu_i)$, the long-run utility rate (net payoff per unit time) of server i is

$$U(\mu_i;\boldsymbol{\mu}_{-i},\lambda,p)=p\mu_i B\left(\sum_{j=1}^{N}\mu_j;\lambda\right)-c(\mu_i). \tag{3}$$

Like Armony et al. (2021) and other recent strategic server queueing works, we model the capacity cost as a continuous cost flow whether the server is busy or idle (see their section 3.2.2 for an explanation of this modeling choice). Note also that the utility function in (3) intentionally does not include a disutility from congestion. Different from models with customer ownership such as Armony et al. (2021), servers in our racing queue context do not internalize customer waiting costs. Unlike, e.g., a supermarket, where customers are human beings in full view of the servers, jobs in our settings are instead virtual entities like blocks of transactions. There is a single shared queue with no job "belonging to" any particular server, and rewards follow a winner-take-all rule based solely on completion order rather than waiting time. Accordingly, the utility is fully specified by the prize earnings and capacity costs.

We focus on symmetric equilibria. Suppose $N-1$ servers choose a common service rate μ and consider a tagged server (with index $i=1$ without loss of generality) that chooses μ_1. The aggregate service rate is $\mu_1+(N-1)\mu$, so the tagged server's utility becomes

$$U(\mu_1,\mu)=p\mu_1 B\left(\mu_1+(N-1)\mu;\lambda\right)-c(\mu_1). \tag{4}$$

Then, a symmetric equilibrium is a service rate $\mu^\star$ such that, when all other servers choose $\mu^\star$, it is optimal for the tagged server to choose $\mu^\star$ as well; that is,

$$\mu^\star\in\arg\max_{\mu_1\geq 0}U(\mu_1,\mu^\star). \tag{EQ}$$

3.2. The First-Order Condition

Substituting the busy-fraction expression into the tagged server's utility yields a piecewise objective that differs across service regimes. If $\mu_1+(N-1)\mu\leq\lambda$, the system is critically loaded or overloaded, is always busy, and the tagged server's earning rate is linear in μ_1. If instead $\mu_1+(N-1)\mu>\lambda$, the system is stable and throughput equals λ; in that regime, the tagged server's earning rate depends on her winning fraction $\mu_1/(\mu_1+(N-1)\mu)$. Specifically, the tagged server's utility can be written as

$$U(\mu_1,\mu)=\begin{cases}p\mu_1-c(\mu_1), & \mu_1\leq\lambda-(N-1)\mu,\\ p\lambda\left(\dfrac{\mu_1}{\mu_1+(N-1)\mu}\right)-c(\mu_1), & \mu_1>\lambda-(N-1)\mu.\end{cases} \tag{5}$$

[1] When $\boldsymbol{\mu}=\mathbf{0}$ (and thus $\sum_{j=1}^{N}\mu_j=0$), the expression in (2) is indeterminate. Naturally, if server i chooses a service rate of zero, then she will not win any prizes; accordingly, we define the earning rate as zero in this case. Conveniently, when $\boldsymbol{\mu}=\mathbf{0}$ (which implies $\mu_i=0$), $p\mu_i B\left(\sum_{j=1}^{N}\mu_j;\lambda\right)$ also evaluates to zero, so this expression for server i's earning rate remains valid in this case.

The function is continuous at the boundary $\mu_1 = \lambda - (N-1)\mu$, but not differentiable there: on the overloaded side, the busy fraction is always 1, but on the underloaded side, it is strictly decreasing in μ_1. This creates a kink in $U(\mu_1, \mu)$ at the critically loaded point that serves as the boundary between the two regimes. Such non-smoothness is responsible for the piecewise structure in determining the regime of the equilibrium.

It is useful to interpret the two regimes. In the overloaded regime, the tagged server's utility takes the simple form $p\mu_1 - c(\mu_1)$; since the $\sum_j \mu_j$ terms cancel out of the earning rate (2) and the busy fraction is 1, it is as if the tagged server is paid per completion at service rate μ_1, so an incremental increase in μ_1 increases earning rate one-for-one (by p) regardless of competitors. However, the set of μ_1 values that correspond to overload depends on other servers' rates through $\lambda - (N-1)\mu$: the effort of other servers creates a stabilization cushion.

In the underloaded regime, serving faster does not increase the overall throughput (which is fixed at λ), but it does increase the tagged server's winning probability. The potential gain from an increased winning probability can incentivize speeding up even in an already stable system, in contrast to a single-server system where increasing μ beyond λ only increases idle time.

Differentiating $U(\mu_1, \mu)$ (for μ_1 strictly inside each regime) yields the first-order condition (FOC):

$$0 = \frac{\partial U(\mu_1, \mu)}{\partial \mu_1} = \begin{cases} p - c'(\mu_1), & \mu_1 < \lambda - (N-1)\mu, \\ p\left(\dfrac{\lambda}{\mu_1 + (N-1)\mu}\right)\left(1 - \dfrac{\mu_1}{\mu_1 + (N-1)\mu}\right) - c'(\mu_1), & \mu_1 > \lambda - (N-1)\mu. \end{cases} \tag{6}$$

Note that when $N = 1$, the derivative for the underloaded branch – which corresponds to $\mu_1 > \lambda$ in this case – simplifies to $\frac{\partial U(\mu_1,\mu)}{\partial \mu_1} = -c'(\mu_1) < 0$. Thus, a lone server never finds it optimal to operate above the critical rate λ: doing so increases cost without increasing earning rate because the server already wins every race. This observation clarifies that the incentive to exert enough effort to push the system past critically loaded into the underloaded regime arises only from competition.

3.3. Equilibrium Characterization

To obtain clean comparative statics while retaining sufficient flexibility, we adopt a particular convex capacity cost specification which has precedent in the queueing literature with strategic servers (see, e.g., Armony et al. (2021), Zhong et al. (2025a)). The following assumption applies for the remainder of the paper.

ASSUMPTION 1 **(Capacity cost function)**. $c(\mu; c_E, q) = c_E \mu^q$ *for* $c_E > 0$ *and* $q > 1$.

The parameter q governs the elasticity of marginal cost (Armony et al. 2021) and c_E scales cost level. This specification nests linear marginal cost ($q \to 1+$) and increasingly convex cost (larger q).

Under Assumption 1, the tagged server's utility is strictly concave in μ_1 within each regime, and we next show that the best response has a simple structure: either the FOC has a unique interior solution or the optimum occurs at the regime boundary.

LEMMA 1 **(Tagged server optimization)**. *Fix* $\mu \geq 0$ *and* $N \geq 1$. *Exactly one of the following holds:*

(i) the FOC (6) *has a unique solution over* $\mu_1 \geq 0$*, which is also the unique global maximizer in* μ_1 *of* $U(\mu_1, \mu)$*, and moreover this solution is strictly positive; or*

(ii) the FOC (6) *has no solution over* $\mu_1 \geq 0$*, and the unique global maximizer is* $\mu_1 = \lambda - (N-1)\mu$*.*

We now characterize the symmetric equilibrium. Substituting $\mu_1 = \mu$ into the FOC (6) yields a piecewise symmetric FOC: when $\mu < \lambda/N$, the system is overloaded under symmetry, and when $\mu > \lambda/N$, it is underloaded. Therefore, the symmetric FOC is given by

$$0 = \left.\frac{\partial U(\mu_1, \mu)}{\partial \mu_1}\right|_{\mu_1=\mu} = \begin{cases} p - c'(\mu), & \mu < \dfrac{\lambda}{N}, \\ p\dfrac{\lambda}{N\mu}\left(1 - \dfrac{1}{N}\right) - c'(\mu), & \mu > \dfrac{\lambda}{N}. \end{cases} \tag{7}$$

As shown in the proof of Lemma 1, (7) has a downward jump at $\mu = \lambda/N$ (see Figure 1 for illustration), which rules out the possibility of multiple equilibria.

For later use, define the underloaded and overloaded candidate equilibrium service rates:

$$\mu_{\mathrm{U}} := \left(p\frac{\lambda}{Nqc_E}\left(1 - \frac{1}{N}\right)\right)^{\frac{1}{q}}, \quad \text{and} \quad \mu_0 := \left(\frac{p}{qc_E}\right)^{\frac{1}{q-1}}. \tag{8}$$

The equilibrium service rate is one of three possibilities: the underloaded candidate μ_{U}, the overloaded candidate μ_0, or the boundary rate λ/N. The next proposition – our first major result – provides a complete and closed-form characterization.

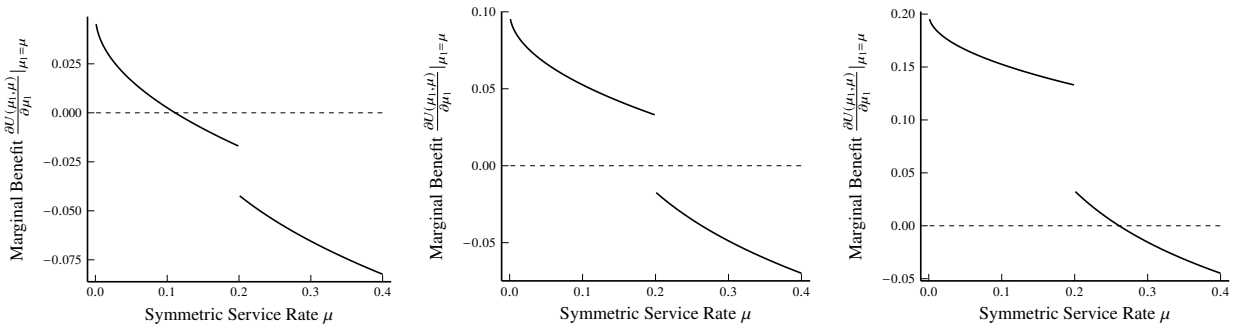
Proposition 1 **(Equilibrium service rate)**. *A symmetric equilibrium service rate* $\mu^\star$ *exists and is unique. Moreover, it satisfies the following:*

(i) (Overloaded equilibrium) If $p < c'(\frac{\lambda}{N})$*, then* $\mu^\star = \mu_0 < \frac{\lambda}{N}$*;*

(ii) (Critically loaded equilibrium) If $c'(\frac{\lambda}{N}) \leq p \leq \frac{N}{N-1}c'(\frac{\lambda}{N})$*, then* $\mu^\star = \frac{\lambda}{N}$*;*

(iii) (Underloaded equilibrium) If $p > \frac{N}{N-1}c'(\frac{\lambda}{N})$*, then* $\mu^\star = \mu_{\mathrm{U}} > \frac{\lambda}{N}$*.*

Proposition 1 reveals a sharp three-regime structure. When the prize p is small relative to marginal capacity cost at the critical rate (case (i)), servers choose an interior rate μ_0 that does not depend on λ or N; the system is thus overloaded and throughput equals $N\mu_0$. When p is large enough (case (iii)), servers operate above λ/N and the system is underloaded, so throughput equals λ and servers' speeds only affect their winning probability from that throughput. Between these regions (case (ii)), the equilibrium sits exactly at the stability boundary: servers collectively provide just enough capacity to match arrivals, and the system is critically loaded. In this range, the equilibrium effort does not change as the prize amount increases, because the return to faster speed (marginal earning rate) is discontinuously lower in the underloaded regime. To incentivize servers to stabilize the system thus requires a discrete jump in the prize amount. Figure 1 plots the derivative $\left.\frac{\partial U(\mu_1,\mu)}{\partial \mu_1}\right|_{\mu_1=\mu}$ – see equation (7) – as a function of the symmetric service rate μ. The panels cover the three cases implied by Proposition 1. In each panel, the first piece of the curve corresponds to the overloaded



(a) $p = 0.05$ ($\mu^\star = \mu_0$, overloaded) (b) $p = 0.1$ ($\mu^\star = \lambda/N$, critical) (c) $p = 0.2$ ($\mu^\star = \mu_{\mathrm{U}}$, underloaded)

Figure 1 Numerical illustrations of the first-order condition under three different values of p. ($\lambda = 0.4$, $q = 1.5$, $c_E = 0.1$, $N = 2$). In this figure, the two thresholds for p in Proposition 1 are $c'(\frac{\lambda}{N}) = 0.067$ and $\frac{N}{N-1}c'(\frac{\lambda}{N}) = 0.134$.

interval, the second piece to the underloaded interval, and a downward jump occurs at the critically loaded point λ/N. Which regime the equilibrium obeys is determined by where this derivative crosses (or jumps past) zero.

The thresholds that separate the three regimes have intuitive interpretations. The lower threshold $p = c'(\frac{\lambda}{N})$ equates the prize amount to the marginal capacity cost at $\mu = \lambda/N$. If the prize is below this marginal cost, servers do not supply enough capacity to reach stability. The upper threshold $p = \frac{N}{N-1}c'(\frac{\lambda}{N})$ is larger. This is because in the underloaded regime, increasing μ above λ/N yields only a fractional increase in winning probability, proportional to $(N-1)/N$. Thus, relative to the overloaded regime, the effective marginal benefit of speed is reduced, and a larger prize is required to induce $\mu > \lambda/N$.

The single-server case ($N = 1$) is a special limiting case: the underloaded region disappears because $N/(N-1)$ diverges. This matches our earlier observation that, without competition, no prize induces $\mu > \lambda$ because additional effort cannot increase the winning probability beyond one.

4. Effect of Competition

Having characterized the equilibrium, we now study how competition, captured by the number of servers N, affects equilibrium service rates and aggregate system capacity. This is a central question for the design of winner-take-all service environments: does adding competitors to a racing queue improve system performance, and if so, by how much? In classical non-strategic multi-server queues, each additional server linearly increases total service capacity. In our setting, however, adding a server changes not only the physical capacity but also the strategic incentives of every server, creating the possibility that individual effort declines enough to dampen, or even negate, the capacity expansion.

A natural starting point is the comparison between a monopoly server ($N = 1$) and a competitive environment ($N \geq 2$). Does a server work faster when she is the sole provider and captures all prizes, or when she must race against others? The answer is a priori ambiguous because increasing the number of servers simultaneously

introduces three distinct forces. First, it dilutes each server's expected prize earning rate, because with more competitors, any given server wins a smaller share of races. Second, it creates a racing incentive; that is, servers can increase their winning probability by working faster than rivals, an effect absent under monopoly. Third, it provides a stabilization cushion, with other servers contributing to aggregate capacity, which makes it easier to meet demand and potentially allows each server to reduce individual effort. The overall impact depends on the interplay among these forces, which is governed by the prize amount, capacity cost, and level of competition.

To emphasize the dependence on N, in this section, we denote by $\mu^\star(N)$ the unique equilibrium service rate from Proposition 1, and by $\mu_{\mathrm{U}}(N)$ and $\mu_{\mathrm{O}}(N)$ the underloaded and overloaded candidate equilibrium service rates defined in (8). Where appropriate, our results apply to the continuous extension (in N) of the relevant quantities.

4.1. The Duopoly Peak

We begin by characterizing the candidate equilibrium service rates.

LEMMA 2 **(Candidate equilibrium service rates versus N).** *The underloaded candidate equilibrium service rate $\mu_{\mathrm{U}}(N)$ is strictly increasing in N for $1 \le N \le 2$ and strictly decreasing in N for $N > 2$. The critically loaded candidate equilibrium service rate λ/N is strictly decreasing in N, and the overloaded candidate equilibrium service rate $\mu_{\mathrm{O}}(N)$ is constant in N.*

Lemma 2 shows that, in the underloaded regime, the equilibrium service rate is maximized at $N = 2$. The key mechanism operates through the marginal winning probability. When the system is underloaded, a server working at speed μ_1 earns at rate $p\lambda$ multiplied by its winning probability $\mu_1/(\mu_1 + (N-1)\mu)$. Under symmetry ($\mu_1 = \mu$), each server wins with probability $1/N$. The marginal effect of a unilateral increase in μ_1 on this winning probability is

$$\frac{\partial}{\partial \mu_1}\left(\frac{\mu_1}{\mu_1 + (N-1)\mu}\right)\bigg|_{\mu_1=\mu} = \frac{(N-1)\mu}{(\mu_1 + (N-1)\mu)^2}\bigg|_{\mu_1=\mu} = \frac{N-1}{N^2\mu}. \tag{9}$$

The factor $(N-1)/N^2$, which captures the marginal sensitivity of winning probability to individual speed, is maximized at $N = 2$ (where it equals $1/4$) and strictly decreases for all $N > 2$. Hence, holding other primitives fixed, the incentive to speed up is strongest when there is exactly one rival ($N = 2$), as illustrated in Figure 2. Note also that the full marginal sensitivity $(N-1)/(N^2\mu)$ is decreasing in μ, implying that the slower servers work, the more a unilateral speed increase improves winning odds. Notably, the duopoly peak phenomenon is agnostic to other primitives besides N. Changes in other parameters do affect the *value* of the marginal winning probability for each N through the equilibrium service rate, but this does not shift the peak away from $N = 2$.

This just-enough-competition mechanism differs fundamentally from standard product-market competition, where adding firms monotonically intensifies competitive pressure, pushing prices down towards the marginal

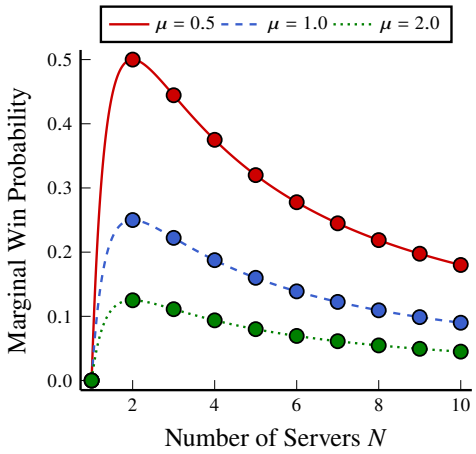


Figure 2 Marginal sensitivity of a server's winning probability to its own speed under symmetric effort, $(N-1)/(N^2\mu)$, as a function of the number of servers N.

cost. In a racing queue, competition is mediated by a probabilistic contest, which features a non-monotonic marginal winning probability. With too many competitors, the incremental chance of winning from working harder becomes too small to justify the marginal cost of effort, leading to effort retrenchment. The duopoly ($N = 2$) sits at the sweet spot, in which competition is sufficient for speeding up to meaningfully increase the probability of winning, yet limited enough that the gains are not diluted across many rivals.

4.2. Equilibrium Monotonicity and Aggregate Capacity

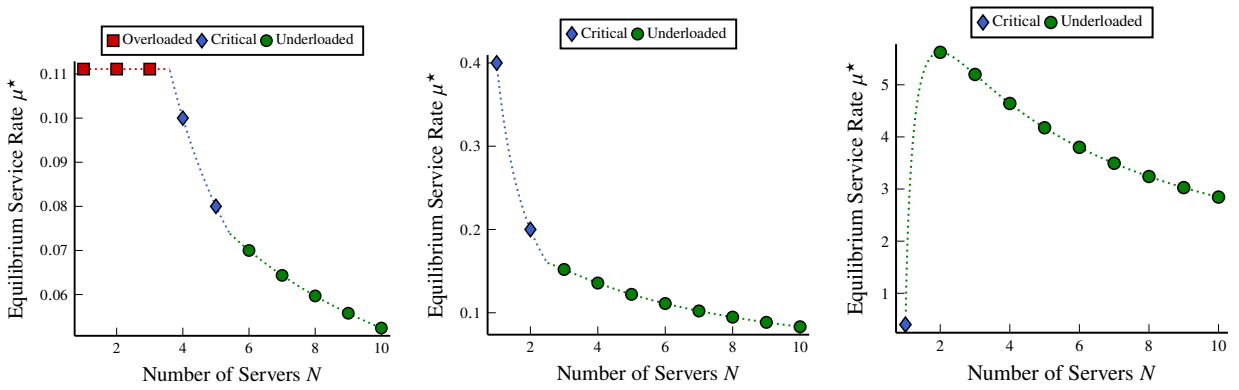
Lemma 2 does not by itself determine the behavior of $\mu^\star(N)$, since the equilibrium can traverse multiple service regimes depending on the prize amount and capacity cost. The following proposition fully characterizes how the equilibrium service rate responds to changes in the number of servers.

PROPOSITION 2 **(Equilibrium service rate versus** N**).** *(i) For* $N \geq 2$, $\mu^\star(N) \geq \mu^\star(N+1)$.
(ii) $\mu^\star(1) < \mu^\star(2)$ *if and only if* $c'(\lambda) < \frac{1}{4}p$.

Proposition 2 delivers two core messages. First, once there are at least two servers, adding additional competitors *always* (weakly) reduces individual equilibrium effort. This monotonicity follows directly from Lemma 2: the overloaded candidate rate is constant in N, the critically loaded candidate rate λ/N is strictly decreasing in N, and the underloaded candidate rate peaks at $N = 2$ and strictly decreases thereafter.

Second, unlike the monotone decline for $N \geq 2$, introducing the *first* competitor from the monopoly ($N = 1$) to the duopoly ($N = 2$) can either raise or lower equilibrium effort, depending on whether the marginal earning rate is large enough. The monopoly ($N = 1$) exhibits no racing incentive. The lone server wins every prize regardless of speed and, by Proposition 1, never finds it optimal to exceed the critically loaded rate, so $\mu^\star(1) \leq \lambda$. Introducing a second server creates a race for the first time, and whether this competition raises or lowers effort is determined by the cutoff $c'(\lambda) = \frac{1}{4}p$. To interpret this condition, suppose a second

server enters and both servers initially work at the monopolist's fastest speed $\mu = \lambda$. The left-hand side $c'(\lambda)$ is the marginal capacity cost at this speed. The right-hand side is the marginal earning rate from a unilateral speed increase at the same point: from (9), the marginal winning probability at $N = 2$ and $\mu = \lambda$ is $(N-1)/(N^2\mu)\big|_{N=2,\,\mu=\lambda} = 1/(4\lambda)$, and the throughput at critical loading is λ, so the marginal earning rate is $p\lambda \cdot 1/(4\lambda) = p/4$. If this marginal earning rate exceeds the marginal cost (i.e., $c'(\lambda) < \frac{1}{4}p$), each server finds it profitable to deviate upward, so the monopoly speed cannot be a duopoly equilibrium and $\mu^\star(2) > \mu^\star(1)$. Conversely, when $c'(\lambda) \geq \frac{1}{4}p$, the marginal capacity cost of speeding up is too large for the competitive incentive to dominate. The stabilization cushion prevails, as two servers can jointly sustain throughput λ with less individual effort, so each server reduces its effort when a competitor enters. In effect, the servers free-ride on each other's capacity contributions, even though they are competing for prizes rather than collaborating toward a shared goal. Practically speaking, this result shows that the racing incentive of competing with another server dominates the stabilization cushion only if the prize is sufficiently large.



(a) Low prize $p = 0.05 < 4c'(\lambda)$ (b) Medium prize $p = 0.1 < 4c'(\lambda)$ (c) High prize $p = 20 > 4c'(\lambda)$

Figure 3 Equilibrium service rate $\mu^\star(N)$ versus number of servers N, for three prize levels ($\lambda = 0.4$, $q = 1.5$, $c_E = 0.1$). In this figure, $c'(\lambda) \approx 0.095$.

Figure 3 illustrates Proposition 2 for three prize levels ($\lambda = 0.4$, $q = 1.5$, $c_E = 0.1$, so $c'(\lambda) \approx 0.095$ and $4c'(\lambda) \approx 0.38$). Panels (a) and (b) fall in the regime $p < 4c'(\lambda)$, where the stabilization cushion dominates; that is, the marginal earning from racing at the monopolist's speed is below the marginal capacity cost, so effort is decreasing for *all* $N \geq 1$. In panel (a) ($p = 0.05$), the prize is so low that the monopolist is overloaded, and the system traverses all three regimes as N grows – overloaded, then critically loaded, then underloaded – with effort declining throughout. In panel (b) ($p = 0.1$), the prize is high enough to sustain a critically loaded rate λ/N at $N = 1$. Panel (c) ($p = 20$) falls in the opposite regime $p > 4c'(\lambda)$, where the marginal earning far exceeds the marginal capacity cost, so by the profitable-deviation logic underlying Proposition 2(ii), introducing a second server increases the equilibrium rate by more than an order of magnitude, placing the

duopoly deep in the underloaded regime. Since the only parameter changing is N, this increase owes entirely to the introduction of a racing incentive from $N = 1$ to $N = 2$, and the size of the increase demonstrates just how strong the competitive force can be in a racing queue. For $N \geq 2$, however, the monotone decline of Proposition 2(i) takes over, and effort decreases steadily. (It does, though, remain much higher than under no competition when $N = 1$.)

Although Proposition 2 shows that individual effort $\mu^\star(N)$ can be non-monotone in N, a natural follow-up question is what this implies for aggregate system capacity $N\mu^\star(N)$. We characterize this with the next result.

PROPOSITION 3 **(Aggregate capacity)**. *For all $N \geq 1$, $(N+1)\mu^\star(N+1) \geq N\mu^\star(N)$, with equality if and only if $\mu^\star(N) = \lambda/N$ and $\mu^\star(N+1) = \lambda/(N+1)$, which is equivalent to $c'(\frac{\lambda}{N}) \leq p \leq \frac{N+1}{N} c'(\frac{\lambda}{N+1})$.*

The most striking implication of Proposition 3 is that adding a server may contribute *zero* marginal capacity. This occurs when both $\mu^\star(N)$ and $\mu^\star(N+1)$ are critically loaded; that is, the prize is high enough to sustain stability with N servers (lower bound on p) but not so high that an $(N+1)$-st server pushes the system into the underloaded regime (upper bound on p). In this case, every incumbent server slows down just enough to absorb the new entrant, keeping total capacity pinned at λ. In the other two regimes, aggregate capacity strictly increases. In the overloaded regime, $\mu^\star(N) = \mu_0(N)$ is constant in N, so total capacity $N\mu_0$ scales linearly, closing the gap toward λ; in the underloaded regime, although individual effort $\mu_{\mathrm{U}}(N)$ decreases for $N > 2$ (Lemma 2), the additional server more than compensates, and $N\mu_{\mathrm{U}}(N)$ is strictly increasing. Still, the findings for the critical and underloaded cases contrast sharply with classical multi-server queues, where each additional server increases capacity by a fixed amount. In the underloaded case, although aggregate capacity does increase with additional servers, after $N = 2$, the individual slowdown with each additional server implies diminishing returns to scale. And as noted, even more stark is that in the critically loaded regime, an extra server does not increase capacity at all.

5. Effect of Capacity Cost Elasticity

This section studies how the equilibrium service rate responds to changes in the capacity cost elasticity q, the exponent in the cost function $c(\mu) = c_E \mu^q$. Since the elasticity of capacity cost with respect to service rate is defined as $\frac{\mathrm{d}c(\mu)}{\mathrm{d}\mu} \cdot \frac{\mu}{c(\mu)} = q$ (see, e.g., Mas-Colell et al. 1995, ch. 2), a one-percent increase in service rate raises capacity cost by q percent, so q governs how steeply marginal capacity cost rises with service rate. Practically speaking, high q represents environments where scaling up is expensive (e.g., operators with limited infrastructure), while low q represents more scalable operations where marginal capacity costs rise gently. The response of the equilibrium service rate to capacity cost elasticity is a priori unclear: our analysis reveals that a higher capacity cost elasticity can, perhaps surprisingly, increase the equilibrium service rate, and the overall response can be non-monotone.

This counterintuitive possibility arises because q affects the equilibrium through two simultaneous channels, in contrast to the number of servers N, which affects the equilibrium primarily through a single *competition*

channel (Section 4). The first channel of q operates within a fixed regime: changing q alters the marginal capacity cost, shifting the candidate equilibrium service rate up or down in that regime. We call this the *marginal cost channel* and develop it in Section 5.1. The second operates across regimes: changing q moves the thresholds in Proposition 1, potentially causing the equilibrium to switch from one regime to another. We call this the *regime-switching channel* and develop it in Section 5.2. Section 5.3 then combines both channels to characterize the overall monotonicity of $\mu^\star$.

To emphasize the dependence on q, in the rest of this section, we write $\mu_{\mathrm{U}}(q)$, $\mu_0(q)$, and $\mu^\star(q)$ for the candidate and equilibrium service rates from Proposition 1.

5.1. Marginal Cost Channel

Within a fixed equilibrium regime (i.e., overloaded, critically loaded, or underloaded), how does the equilibrium service rate respond to a change in q? In equilibrium, the FOC (7) equates each server's marginal capacity cost $c'(\mu^\star)$ to a marginal earning rate that depends on the prize and the competitive environment but not on q. If raising q increases marginal capacity cost at the prevailing rate ($\frac{\partial c'}{\partial q}\big|_{\mu=\mu^\star} > 0$), the server is now overpaying for its current speed and must slow down to restore optimality; conversely, if raising q decreases marginal capacity cost ($\frac{\partial c'}{\partial q}\big|_{\mu=\mu^\star} < 0$), the server finds its current speed cheaper than before and thus speeds up. In other words, the response of $\mu^\star$ to q is opposite in sign to the response of marginal capacity cost $c'(\mu^\star)$ to q. The following result formalizes this.

PROPOSITION 4 **(Sign rule)**. *When* $\mu^\star(q) \neq \lambda/N$, $\frac{\mathrm{d}\mu^\star(q)}{\mathrm{d}q} \cdot \frac{\partial c'}{\partial q}\big|_{\mu=\mu^\star(q)} \leq 0$, *with equality if and only if* $\mu^\star(q) = e^{-1/q}$.

Figure 4 illustrates Proposition 4. For small q, the marginal capacity cost derivative (red dotted) is positive, indicating that higher elasticity makes effort at the current rate more expensive, and the equilibrium service rate (green solid) declines. Near $q \approx 3.3$, the derivative crosses zero, and for larger q it turns negative, meaning that marginal capacity at the prevailing service rate becomes cheaper as q grows, so the equilibrium service rate rises.

Proposition 4 reduces the question of how $\mu^\star(q)$ responds to q to determining when $\frac{\partial c'}{\partial q}\big|_{\mu=\mu^\star}$ is positive or negative. Computing this derivative explicitly gives

$$\frac{\partial c'(\mu)}{\partial q} = c_E \mu^{q-1}(1 + q \log \mu), \tag{10}$$

which changes sign at $\mu = e^{-1/q}$. Since $e^{-1/q} \in (0, 1)$ for all $q > 1$, service rates above one always see their marginal capacity cost rise with q, while for service rates below one, it can go either way. When $\mu^\star > e^{-1/q}$, the derivative is positive, so by Proposition 4 the server slows down. When $\mu^\star < e^{-1/q}$, the derivative is negative, meaning that marginal capacity cost falls with higher capacity cost elasticity, and the server speeds up. Combining (10) with Proposition 4 yields an equivalent threshold characterization.

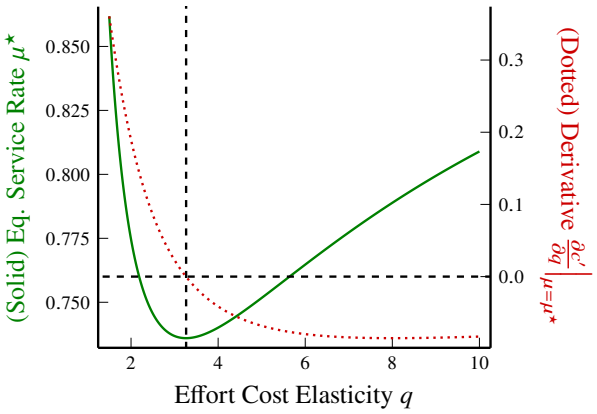


Figure 4 Equilibrium service rate $\mu^\star$ (green solid, left axis) and marginal capacity cost derivative $\frac{\partial c'}{\partial q}\big|_{\mu=\mu^\star}$ (red dotted, right axis) versus capacity cost elasticity q ($\lambda = 1.2$, $p = 2$, $c_E = 0.5$, $N = 2$). The minimum of $\mu^\star(q)$ occurs at $q \approx 3.3$, where $\frac{\partial c'}{\partial q}\big|_{\mu=\mu^\star} = 0$.

COROLLARY 1 **(Threshold form of the sign rule).** *When $\mu^\star(q) \neq \lambda/N$, $\mu^\star(q)$ is strictly increasing in q if $\mu^\star(q) < e^{-1/q}$, and strictly decreasing in q if $\mu^\star(q) > e^{-1/q}$.*

Note that the threshold $e^{-1/q}$ itself depends on q (it increases from $e^{-1} \approx 0.37$ as $q \to 1^+$ toward 1 as $q \to \infty$), so the comparison between $\mu^\star(q)$ and $e^{-1/q}$ can change as q varies. Figure 5 visualizes this using the same parameters as Figure 4. For small q, $\mu^\star(q)$ lies above $e^{-1/q}$ and declines. The two curves cross near $q \approx 3.3$, consistent with the zero-crossing of $\frac{\partial c'}{\partial q}\big|_{\mu=\mu^\star}$ observed in Figure 4. For larger q, $\mu^\star(q)$ lies below $e^{-1/q}$ and rises. (We also explicitly characterize the monotonicity properties of the candidate equilibrium service rates $\mu_0(q)$ and $\mu_U(q)$ in Lemma A.3 in Appendix A, which we use next to help establish the monotonicity of the equilibrium service rates.)

5.2. Regime-Switching Channel

The marginal cost channel developed in Section 5.1 governs how each candidate equilibrium service rate responds to q within a fixed regime. In this subsection, we examine the second channel: as q varies, the regime thresholds in Proposition 1 shift, potentially causing the equilibrium to switch between the underloaded, critically loaded, and overloaded regimes. To formalize this, note that the conditions in Proposition 1 can be equivalently expressed in terms of the auxiliary function $h(q) := q(\lambda/N)^{q-1}$:

$$\mu^\star(q) = \begin{cases} \mu_U(q), & \text{if } h(q) < \frac{p}{c_E}\frac{N-1}{N}, \\ \lambda/N, & \text{if } \frac{p}{c_E}\frac{N-1}{N} \le h(q) \le \frac{p}{c_E}, \\ \mu_0(q), & \text{if } h(q) > \frac{p}{c_E}. \end{cases} \tag{11}$$

The equilibrium regime at any given q is determined by comparing $h(q)$ against the two constant thresholds $\frac{p}{c_E}\frac{N-1}{N}$ and $\frac{p}{c_E}$. We first examine the behavior of $h(q)$, which turns out to depend on λ/N.

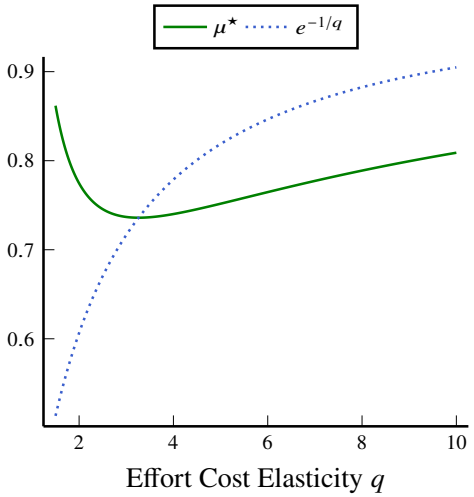


Figure 5 Equilibrium service rate $\mu^\star$ (green solid) and threshold $e^{-1/q}$ (blue dashed) versus capacity cost elasticity q ($\lambda = 1.2$, $p = 2$, $c_E = 0.5$, $N = 2$).

LEMMA 3 (**Auxiliary function**). *The function $h(q) = q(\lambda/N)^{q-1}$ satisfies $h(1) = 1$ and:*

(i) if $\lambda/N \le 1/e$, then h is strictly decreasing on $(1, \infty)$, with $h(q) \to 0$;

(ii) if $\lambda/N \in (1/e, 1)$, then h first increases to a unique maximum of $\frac{-N}{e\lambda \log(\lambda/N)}$, then decreases to 0;

(iii) if $\lambda/N \ge 1$, then h is strictly increasing on $(1, \infty)$, with $h(q) \to \infty$.

As q varies, the curve $h(q)$ can cross each of the two thresholds $\frac{p}{c_E}\frac{N-1}{N}$ and $\frac{p}{c_E}$ at most twice. When $\lambda/N \in (1/e, 1)$, the hump shape of $h(q)$ (Lemma 3(ii)) allows it to cross each threshold once on the way up and once on the way down, yielding up to four transition points. Combining this lemma with the within-regime analysis from Section 5.1, the following result provides a complete characterization of how the equilibrium service rate varies with q, potentially traversing multiple regimes.

PROPOSITION 5 (**Equilibrium service rate versus** q). *There exist $1 \le q_1 \le q_2 \le q_3 \le q_4 \le \infty$ such that $\mu^\star(q)$ satisfies the following five-phase structure:*

(i) For $q \in (1, q_1)$: $\mu^\star(q) = \mu_U(q) > \lambda/N$, the equilibrium is underloaded and strictly decreasing in q;

(ii) For $q \in (q_1, q_2)$: $\mu^\star(q) = \lambda/N$, the equilibrium is critically loaded and constant in q;

(iii) For $q \in (q_2, q_3)$: $\mu^\star(q) = \mu_O(q) < \lambda/N$, the equilibrium is overloaded. On this interval, if $p \le c_E$, then the equilibrium is strictly increasing in q; if $p > c_E$, then the equilibrium is first strictly decreasing in q, then strictly increasing in q.

(iv) For $q \in (q_3, q_4)$: $\mu^\star(q) = \lambda/N$, the equilibrium is critically loaded and constant in q;

(v) For $q \in (q_4, \infty)$: $\mu^\star(q) = \mu_U(q) > \lambda/N$, the equilibrium is underloaded and strictly increasing in q.

REMARK 1. Some phases in Proposition 5 may be degenerate by letting consecutive thresholds coincide.

When all five phases are present, the equilibrium experiences a five-phase transition "underloaded–critical–overloaded–critical–underloaded" as q increases (see the bottom left panel of Figure EC.2 in the proof of Proposition 5 in Appendix D). By Lemma 3, this requires $\lambda/N \in (1/e, 1)$ (so that $h(q)$ is hump-shaped) and both thresholds to satisfy $h(1) = 1 < \frac{p}{c_E}\frac{N-1}{N}$ and $\frac{p}{c_E} < \max_q h(q)$. The first inequality ensures $h(q)$ starts below each threshold (so the equilibrium begins underloaded), while the second ensures $h(q)$ rises above the upper threshold before descending, crossing each threshold twice and activating all five phases. When either condition fails, some phases collapse and fewer regimes are visited.

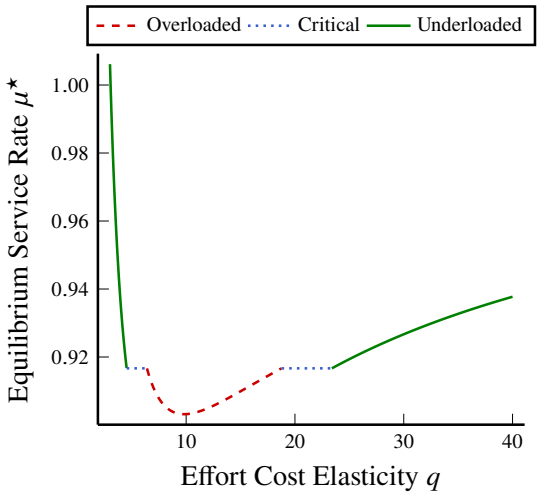


Figure 6 Equilibrium service rate $\mu^\star$ versus capacity cost elasticity q ($\lambda = 5.5$, $p = 0.4$, $c_E = 0.1$, $N = 6$).

Figure 6 illustrates all five phases of Proposition 5 for $\lambda = 5.5$, $p = 0.4$, $N = 6$, $c_E = 0.1$. Since $\lambda/N \approx 0.92 \in (1/e, 1)$, by Lemma 3(ii), $h(q) = q \cdot 0.92^{q-1}$ is hump-shaped, rising from $h(1) = 1$ to a peak of approximately 4.6 at $q \approx 11.5$ before declining to zero. The two thresholds in (11) evaluate to $\frac{N-1}{N}\frac{p}{c_E} \approx 3.33$ and $\frac{p}{c_E} = 4$. Since $h(1) = 1$ is below both thresholds, the equilibrium begins underloaded; since the peak ≈ 4.6 exceeds the upper threshold 4, the curve $h(q)$ crosses both thresholds on its way up and again on its way down, activating all five phases. The equilibrium starts underloaded and decreasing (by Corollary 1, because $\mu^\star(q) > e^{-1/q}$ in this range), passes through a critically loaded plateau as $h(q)$ crosses 3.33, enters the overloaded regime as $h(q)$ crosses 4 (with the overloaded service rate first decreasing then increasing: see Lemma A.3(i-b) in Appendix A), returns through a second critical plateau as $h(q)$ descends back through 4 and 3.33, and finally re-enters the underloaded regime on an increasing trajectory (by Corollary 1, because $\mu^\star(q) < e^{-1/q}$ in this range). Throughout, the regime-switching channel determines the phase boundaries, while the marginal cost channel governs the direction of $\mu^\star$ within each phase.

5.3. Overall Monotonicity

Proposition 5 provides a fully general characterization of the regime and the monotonicity of $\mu^\star(q)$; however, it may be difficult to digest. Focusing just on the monotonicity allows a cleaner presentation, as shown next.

PROPOSITION 6 **((Non-)Monotonicity in q).**

(i) If $p \le c_E$, then $\mu^\star(q)$ is monotonically increasing in q for all $\lambda > 0$.

(ii) If $p > c_E$, there exists a threshold $\Lambda(p) > 0$ such that:

(ii-a) if $\lambda \le \Lambda(p)$, then $\mu^\star(q)$ is monotonically increasing in q;

(ii-b) if $\lambda > \Lambda(p)$, then $\mu^\star(q)$ is first decreasing and eventually monotonically increasing in q.

In particular, we have

$$\Lambda(p) := \begin{cases} \lambda^\dagger, & c_E < p \le \frac{N}{N-1} c_E, \\ \frac{N^2}{e(N-1)} \frac{c_E}{p}, & p > \frac{N}{N-1} c_E, \end{cases} \tag{12}$$

where $\lambda^\dagger$ is the unique solution to $\frac{N}{e}\frac{c_E}{p} + \lambda \log\left(\frac{\lambda}{N}\right) = 0$.

Part (i) identifies a monotonically increasing case. When $p \le c_E$, the overloaded candidate satisfies $\mu_0 \le (1/q)^{1/(q-1)} < e^{-1/q}$ for all $q > 1$ and therefore is increasing (by Corollary 1), and the underloaded candidate is either zero ($N = 1$) or also increasing (as shown in the proof in Appendix D). Therefore, all decreasing phases in Proposition 5 are degenerate and $\mu^\star(q)$ is increasing regardless of λ.

Part (ii) delineates the two cases when $p > c_E$. The overloaded candidate $\mu_0(q)$ is now U-shaped, so it can decrease for small q. Similarly, the underloaded candidate $\mu_U(q)$ can also decrease for small q. Whether the equilibrium actually visits such a decreasing region depends on the arrival rate. When $\lambda \le \Lambda(p)$, the equilibrium service rate remains in regimes where the relevant candidate is increasing, so $\mu^\star(q)$ is increasing. When $\lambda > \Lambda(p)$, the higher demand pushes the equilibrium into a region where the candidate is decreasing, so effort initially declines. As q continues to grow, the candidate eventually enters its increasing phase, and effort recovers. The expression for $\Lambda(p)$ in (12) shows that it is decreasing in p for $p > \frac{N}{N-1} c_E$, implying that higher prizes make the first-decrease-then-increase pattern more likely.

As noted in Armony et al. (2021), as the elasticity q becomes large, the cost function becomes very kinked at 1 (extremely flat everywhere below 1 and extremely steep everywhere above 1). Thus, although the equilibrium is always increasing in q for large enough q, the limit is 1.

Figure 7 visualizes this dichotomy for $N = 2$ and $c_E = 0.9$. For $p \le c_E = 0.9$ (to the left of the leftmost vertical dotted line), the equilibrium is monotone regardless of the arrival rate, consistent with Proposition 6(i). For $p > c_E$, the boundary curve $\lambda = \Lambda(p)$ separates the monotone region (lower left) from the non-monotone region (upper right). In the high-prize regime ($p > \frac{N}{N-1} c_E = 1.8$), the boundary takes the explicit form $\Lambda(p) = N^2 c_E / (e(N-1)p)$, which implies that as the prize grows, even a modest arrival rate suffices to trigger non-monotonicity. Notably, the boundary exhibits a downward jump at $p = \frac{N}{N-1} c_E$ because the equilibrium starts in different regimes on the two sides. In particular, for p just below, it starts critically loaded at $q = 1$

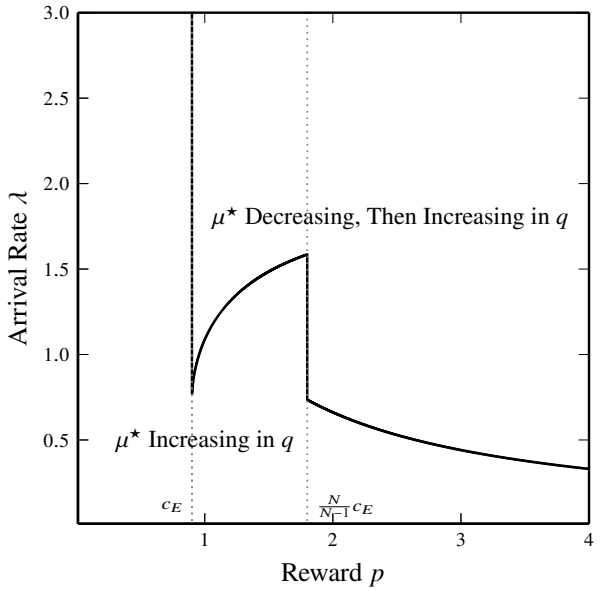


Figure 7 Monotonicity regions for $\mu^\star$ as a function of q in the (p, λ) parameter space ($N = 2$, $c_E = 0.9$). Below the curve $\lambda = \Lambda(p)$, the equilibrium service rate is monotonically increasing in q; above the curve, it is first decreasing, then increasing in q.

(since $\frac{N-1}{N}\frac{p}{c_E} \leq 1 = h(1)$ in (11)), while for p just above, it starts underloaded (since $\frac{N-1}{N}\frac{p}{c_E} > 1 = h(1)$). The conditions for non-monotonicity to emerge at some larger q differ across these two starting regimes, and as seen from (12), they translate into different arrival-rate thresholds, producing the discontinuity.

To summarize, this section delivers two findings. First, surprisingly, higher capacity cost elasticity can *increase* equilibrium effort, and does so monotonically when the prize is modest or demand is low ($p \leq c_E$ or $\lambda \leq \Lambda(p)$). Second, while a monotone increase is not guaranteed when the prize is generous and demand is high ($p > c_E$ and $\lambda > \Lambda(p)$), it is nonetheless the case that the equilibrium service rate eventually keeps increasing (in a bounded fashion) as capacity cost elasticity grows, even after an initial decline.

6. Prize Design

We now examine the role of the prize p as a design variable. Each completed job generates value $r > 0$ for the designer. Depending on the application, r may represent the economic benefit of validating a block in a blockchain network, the avoided cost of a security breach on a bug-bounty platform, or the scientific value of solving an open mathematical problem. The designer sets a prize p per completed job and takes the remaining system primitives (N, λ, q, c_E) as given.

From the equilibrium characterization in Proposition 1, the equilibrium throughput $\Theta(p)$ inherits a two-piece structure in p:

$$\Theta(p) = \begin{cases} N\left(\dfrac{p}{qc_E}\right)^{\frac{1}{q-1}}, & p < c'\left(\dfrac{\lambda}{N}\right), \\ \lambda, & p \geq c'\left(\dfrac{\lambda}{N}\right). \end{cases} \tag{13}$$

In the overloaded regime ($p < c'(\lambda/N)$), the system is always busy and throughput is strictly increasing in p, so a higher prize incentivizes faster service and raises throughput. Once the prize reaches $c'(\lambda/N)$, servers provide enough capacity to match demand, and throughput equals the arrival rate λ regardless of any further increase in p. This two-piece structure – increasing throughput below a prize threshold, flat throughput above it – shapes all prize design problems studied in this section.

Having characterized how equilibrium throughput responds to p, we turn to the prescriptive question: what prize amount is optimal? The answer depends on whose payoffs the designer accounts for. A designer who considers only its own net revenue may choose a different prize than one that also internalizes server welfare. We study both benchmarks and show that the gap between them reveals a systematic tendency to underprice.

To emphasize the dependence on p, in this section, we write $\mu^\star(p)$ for the equilibrium service rates from Proposition 1.

6.1. Profit-Maximizing Prize

Each completed job generates value r for the designer and triggers a prize payment of p to the winning server, yielding a net value of $(r - p)$ per completion. The designer's long-run net profit rate is then $(r - p)\Theta(p)$, and we denote the maximized value by $\Pi^\star = \max_{p>0}(r - p)\Theta(p)$. The profit-maximizing prize therefore solves

$$p_D^\star \in \arg\max_{p>0}(r - p)\Theta(p). \tag{14}$$

Any fixed operating costs independent of p, such as infrastructure or staffing costs, can be incorporated without affecting the optimal prize. For $p \geq c'(\lambda/N)$, throughput is constant at λ (from (13)), so the objective becomes $(r - p)\lambda$, which is strictly decreasing in p. Consequently, any prize $p > c'(\lambda/N)$ is strictly dominated by $p = c'(\lambda/N)$. The optimization therefore reduces to the overloaded domain $p \in [0, c'(\lambda/N)]$ (which includes the critically loaded boundary), where the objective function, using (13), is

$$(r - p) \cdot N\left(\frac{p}{qc_E}\right)^{\frac{1}{q-1}}, \tag{15}$$

which is a single-peaked function of p with a unique maximizer at $p = r/q$ (see the proof of Proposition 7 in Appendix E for the detailed argument). If $r/q < c'(\lambda/N)$, the unconstrained maximizer is interior, and the optimum is $p_D^\star = r/q$. If instead $r/q \geq c'(\lambda/N)$, the unconstrained maximizer is not interior, and since the objective is increasing throughout $[0, c'(\lambda/N)]$, the optimum is at the boundary $p_D^\star = c'(\lambda/N)$.

PROPOSITION 7 **(Profit-maximizing prize)**.

(i) If $r/q < c'(\lambda/N)$, then $p_D^\star = r/q$. The induced equilibrium service rate is overloaded, with $\mu^\star = \mu_0 < \lambda/N$ and $\Pi^\star = \frac{r(q-1)}{q} \cdot N \left(\frac{r}{q^2 c_E}\right)^{1/(q-1)}$.

(ii) If $r/q \geq c'(\lambda/N)$, then $p_D^\star = c'(\lambda/N)$. The induced equilibrium service rate is critically loaded, with $\mu^\star = \lambda/N$ and $\Pi^\star = (r - c'(\lambda/N))\lambda$.

When the per-job value is low ($r/q < c'(\lambda/N)$), the optimal prize $p_D^\star = r/q$ is a constant fraction $1/q$ of the job value. Notably, it is independent of the arrival rate λ, the number of servers N, and the cost parameter c_E. This invariance stems from the fact that the equilibrium is overloaded: the system is always busy, so each server earns flow payoff $p\mu^\star$. Consequently, queueing components – such as how fast jobs arrive or how many servers share the workload – do not affect equilibrium incentives. The optimal prize is therefore pinned down solely by the job value r and the cost elasticity q. The resulting net margin, $r - r/q = r(q-1)/q$, is also a constant fraction of r, implying that – conditional on being in this region – the designer extracts a fixed share of the surplus regardless of system scale.

When the per-job value is high ($r/q \geq c'(\lambda/N)$), the optimal prize instead lies at the boundary $p_D^\star = c'(\lambda/N)$, the minimum prize required to sustain full throughput λ. In this regime, the system operates at critical load, and the net margin per job is $r - c'(\lambda/N)$, which is strictly positive since $r \geq q \cdot c'(\lambda/N) > c'(\lambda/N)$ for $q > 1$. Unlike in the low-value regime, the optimal prize now depends on the system primitives λ, N, and c_E through the marginal capacity cost term $c'(\lambda/N)$.

6.2. Surplus-Maximizing Prize

The profit-maximizing prize accounts only for the designer's net revenue, ignoring servers' utility. A complementary benchmark asks what prize maximizes the total surplus, aggregating the payoffs of all participants. This broader perspective is relevant when the prize amount is set by a regulator, platform designer, or funding authority that internalizes the effects of prize design on server welfare.

The total surplus is the sum of the designer's profit and the aggregate server utility:

$$W(p) := (r - p)\,\Theta(p) + N \cdot U(\mu^\star(p), \mu^\star(p)), \tag{16}$$

where $U(\mu^\star, \mu^\star) = p\mu^\star B(N\mu^\star; \lambda) - c_E(\mu^\star)^q$ is each server's utility in equilibrium from (4). Observe that the prize is a pure transfer within this system: each prize payment is simultaneously an expense to the designer and revenue to the winning server. As a result, transfers cancel in (16), and total surplus simplifies to

$$W(p) = r\,\Theta(p) - N\,c_E\big(\mu^\star(p)\big)^q, \tag{17}$$

that is, the total value of completed jobs minus aggregate capacity costs. Consequently, the prize affects surplus only through its effect on the equilibrium service rate and the resulting throughput and capacity costs.

We denote the maximum surplus by $W^\star$ and a surplus-maximizing prize by $p_S^\star$:

$$W^\star := \max_{p>0} W(p), \qquad p_S^\star \in \arg\max_{p>0} W(p). \tag{18}$$

Using the equilibrium characterization in Proposition 1, the surplus function admits three regimes (see the proof of Proposition 8 in Appendix E for the detailed derivation):

$$W(p) = \begin{cases} N\left(\dfrac{p}{qc_E}\right)^{\frac{1}{q-1}}\left(r - \dfrac{p}{q}\right), & p < c'\left(\dfrac{\lambda}{N}\right), \\ r\lambda - Nc_E\left(\dfrac{\lambda}{N}\right)^q, & c'\left(\dfrac{\lambda}{N}\right) \le p \le \dfrac{N}{N-1}c'\left(\dfrac{\lambda}{N}\right), \\ r\lambda - \dfrac{p}{q}\lambda\left(1 - \dfrac{1}{N}\right), & p > \dfrac{N}{N-1}c'\left(\dfrac{\lambda}{N}\right). \end{cases} \tag{19}$$

Two observations substantially simplify the optimization. First, in the critically loaded regime ($c'(\lambda/N) \le p \le \frac{N}{N-1}c'(\lambda/N)$), surplus is constant at $r\lambda - Nc_E(\lambda/N)^q$. Once throughput reaches λ and as long as the equilibrium service rate remains at λ/N, changes in the prize merely redistribute surplus between the designer and servers, without affecting throughput or capacity costs. Second, in the underloaded regime ($p > \frac{N}{N-1}c'(\lambda/N)$), surplus is strictly decreasing in p. In this regime, higher prizes increase $\mu_{\mathrm{U}}(p)$ without increasing throughput (fixed at λ), thereby raising capacity costs without generating additional value. Hence, it is never optimal to set a prize that pushes the system into the underloaded regime.

The problem therefore reduces to comparing the surplus in the overloaded regime, $N\left(\frac{p}{qc_E}\right)^{\frac{1}{q-1}}(r - p/q)$, with the constant surplus achieved under critical loading, $r\lambda - Nc_E(\lambda/N)^q$. The surplus in the overloaded regime is a single-peaked function of p and admits a unique maximizer at $p = r$ (see the proof of Proposition 8 in Appendix E for details). If $r < c'(\lambda/N)$, this maximizer lies in the interior of the overloaded regime, yielding $p_S^\star = r$. If instead $r \ge c'(\lambda/N)$, the surplus is increasing over $[0, c'(\lambda/N)]$, and the optimum is attained at critical loading, achieved by any $p \in [c'(\lambda/N), \frac{N}{N-1}c'\left(\frac{\lambda}{N}\right)]$.

PROPOSITION 8 **(Surplus-maximizing prize)**.

(i) If $r < c'(\lambda/N)$, then $p_S^\star = r$. The induced equilibrium service rate is overloaded, with $\mu^\star = \mu_0 < \lambda/N$ and $W^\star = \frac{r(q-1)}{q} \cdot N\left(\frac{r}{qc_E}\right)^{1/(q-1)}$.

(ii) If $r \ge c'(\lambda/N)$, then any $p_S^\star \in [c'(\lambda/N), \frac{N}{N-1}c'\left(\frac{\lambda}{N}\right)]$ is optimal. The induced equilibrium service rate is critically loaded, with $\mu^\star = \lambda/N$ and $W^\star = r\lambda - Nc_E(\lambda/N)^q$.

When the per-job value is low ($r < c'(\lambda/N)$), the unique surplus-maximizing prize $p_S^\star = r$ sets the prize equal to the full job value. This induces an overloaded system, which behaves as if servers operated independently. In this regime, setting the prize at r aligns each server's utility with the system-level surplus. The resulting surplus, $W^\star = r(q-1)/q \cdot N(r/(qc_E))^{1/(q-1)}$, retains a fraction $(q-1)/q$ of the total output value, reflecting that only a share $1/q$ of the value generated (or, equivalently, of the prize, since $p = r$) is dissipated as capacity cost.

When the per-job value is high ($r \ge c'(\lambda/N)$), a job completion is worth enough to the system to push servers beyond the overloaded regime into the critically loaded regime, but the surplus-maximizing prize is no

longer unique. Within the critically loaded regime, the prize has no effect on throughput or aggregate capacity cost and serves only to redistribute surplus between the designer and servers. Similar to the profit-maximizing case, it is strictly sub-optimal to induce servers to underload the system, since the extra capacity cost generates no additional value. Any prize amount in the interval $[c'(\lambda/N), \frac{N}{N-1}c'(\lambda/N)]$ therefore achieves the same maximal surplus $r\lambda - Nc_E(\lambda/N)^q$.

6.3. Under-Pricing under Profit Maximization

Propositions 7 and 8 reveal a systematic gap between the profit-maximizing and surplus-maximizing prizes. A designer who maximizes profit alone always sets a prize weakly below any surplus-maximizing prize (we establish this formally below). The nature and magnitude of this gap depend on the per-job value r relative to the two thresholds $c'(\lambda/N)$ and $qc'(\lambda/N)$.

When the per-job value is low ($r < c'(\lambda/N)$), both objectives induce overloaded equilibria, but the surplus-maximizing prize is strictly higher. The profit-maximizing designer chooses $p_D^\star = r/q$, whereas the surplus-maximizing prize is $p_S^\star = r$. Since $q > 1$, the designer pays servers only a fraction $1/q$ of the level required for surplus maximization. Consequently, the surplus under the profit-maximizing prize is strictly below the optimum: the surplus in the overloaded regime is unimodal with a peak at $p = r$, and the profit-maximizing prize r/q lies strictly below this peak.

When the per-job value is intermediate ($c'(\lambda/N) \le r < qc'(\lambda/N)$), the two objectives lead to qualitatively different service regimes. Surplus maximization achieves critical loading, so throughput equals the full arrival rate λ. In contrast, profit maximization sets $p_D^\star = r/q < c'(\lambda/N)$, leaving the system overloaded with throughput $N\mu_0(r/q) < \lambda$.

When the per-job value is sufficiently large ($r \ge qc'(\lambda/N)$), the profit-maximizing prize $p_D^\star = c'(\lambda/N)$ lies within the surplus-maximizing range. In this regime, both objectives induce critical loading, and profit maximization is aligned with surplus maximization.

The next result formalizes these observations and quantifies the surplus loss from the profit-maximizing prize.

PROPOSITION 9 (**Profit-maximizing prize vs. surplus-maximizing prize**).

(i) For any $p_S^\star \in \arg\max_{p>0} W(p)$, we have $p_D^\star \le p_S^\star$. Moreover, $p_D^\star \in \arg\max_{p>0} W(p)$ if and only if $r \ge qc'(\lambda/N)$, in which case the relative surplus loss is zero.

(ii) If $r < c'(\lambda/N)$ (both overloaded), then the relative surplus loss is

$$\frac{W^\star - W(p_D^\star)}{W^\star} = 1 - \frac{q+1}{q^{q/(q-1)}} > 0,$$

which is bounded above by $1 - 2/e \approx 0.264$.

(iii) If $c'(\lambda/N) \le r < qc'(\lambda/N)$ (profit-maximizing overloaded, surplus-maximizing critically loaded), then the relative surplus loss is

$$\frac{W^\star - W(p_D^\star)}{W^\star} = 1 - \frac{(q^2-1)s^{q/(q-1)}}{sq^2-1}, \quad \textit{where} \quad s := \frac{r}{qc'(\lambda/N)} \in \left[\frac{1}{q}, 1\right).$$

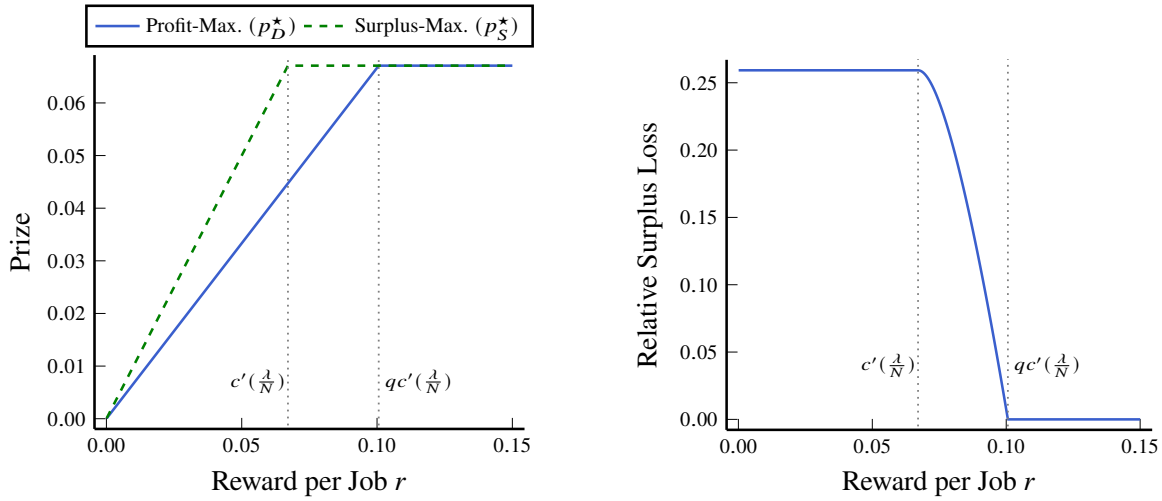
This expression coincides with part (ii) at $s = 1/q$ and decreases continuously to zero as $s \to 1$.

The under-pricing result in Proposition 9(i) reveals a fundamental discrepancy. Increasing the prize reduces the designer's margin on each completed job, which the designer treats as a cost. From the system's perspective, however, this reduction is a transfer to servers rather than a real cost. The only real cost of a higher prize is the additional capacity cost it induces due to servers speeding up. By treating transfers as costs, the profit-maximizing designer overestimates the expense of raising the prize and therefore sets it too low. This distortion arises whenever $r < qc'(\lambda/N)$; beyond this threshold, the system reaches critical loading and the surplus loss disappears.

Proposition 9(ii) and (iii) quantify this surplus loss. When the per-job value is low ($r < c'(\lambda/N)$), the relative surplus loss depends only on the cost elasticity q and is bounded above by $1 - 2/e \approx 26.4\%$. This bound is approached as $q \to 1^+$ (nearly linear costs) and decreases to zero as $q \to \infty$ (highly convex costs). For example, the relative surplus loss is $1 - 3/4 = 25\%$ when $q = 2$ and approximately $1 - 4/3^{3/2} \approx 22.8\%$ when $q = 3$. Notably, this loss is independent of r, λ, N, and c_E, being purely driven by cost elasticity. When the per-job value is intermediate ($c'(\lambda/N) \le r < qc'(\lambda/N)$), profit maximization and surplus maximization induce qualitatively different service regimes (overloaded versus critically loaded). The resulting surplus loss depends on all system primitives through the ratio $s = r/(qc'(\lambda/N))$. The loss at $r = c'(\lambda/N)$ matches that when $r < c'(\lambda/N)$ and decreases continuously to zero as r approaches $qc'(\lambda/N)$.

Figure 8 illustrates Proposition 9 for a representative parameter set. Panel (a) shows how the profit-maximizing and surplus-maximizing prizes vary with r (for surplus maximization, when multiple optimal prizes exist, we use the smallest). The two coincide for $r \ge qc'(\lambda/N)$ but diverge below this threshold. Panel (b) plots the relative surplus loss, confirming that the loss is largest at small values of r, decreases, and eventually vanishes for large enough r, once both solutions reach critical loading.

From a practical standpoint, these results highlights the role of governance or regulatory oversight in competitive service systems. In blockchain networks, under-pricing is most consequential when block-validation value is moderate relative to miners' costs: a reward of $p_D^\star = r/q$ may sustain low throughput and weaken network security, whereas an aggregate-surplus-oriented governance body would set prizes closer to $p_S^\star$. On bug-bounty platforms, prizes closer to $p_S^\star$ can improve researcher participation and sustain high-quality vulnerability discovery. In research contests or mathematics competitions, funding agencies that internalize researcher welfare would set higher prizes, sustaining greater effort even at the cost of a lower per-solution margin.



(a) Optimal prizes $p_D^\star$ and $p_S^\star$ versus per-job value r

(b) Relative surplus loss under profit-maximizing prize vs. r

Figure 8 Profit-maximizing vs. surplus-maximizing prize design ($\lambda = 0.4$, $q = 1.5$, $c_E = 0.1$, $N = 2$). In this figure, $c'(\lambda/N) \approx 0.067$ and $qc'(\lambda/N) \approx 0.101$.

7. Concluding Remarks

This paper introduces and analyzes the racing queue, a queueing model in which multiple strategic servers simultaneously race to complete each arriving job to earn a winner-take-all prize. We obtain a closed-form characterization of the unique symmetric equilibrium, which falls into one of three service regimes – overloaded, critically loaded, or underloaded – depending on the prize amount relative to marginal capacity cost. Competition, cost structure, and prize design all play distinct roles in determining servers' equilibrium effort and the resulting system performance. Adding servers always weakly reduces individual effort once there are at least two competitors, yet introducing the first competitor can sharply raise effort; in other cases, aggregate capacity can be entirely unresponsive to entry because incumbents absorb each new entrant by slowing down. The effect of cost elasticity is similarly nuanced: as elasticity increases, depending on the parameters, the marginal cost of effort at the current equilibrium can either increase or decrease, causing servers to slow down or speed up, respectively. On the design side, when job value is high, the optimal prize induces a critically loaded equilibrium that matches throughput to demand; when job value is low, it sustains an overloaded equilibrium. A profit-maximizing designer, however, systematically sets prizes below the surplus-maximizing level, because treating transfers to servers as costs leads to under-investment in incentives.

More broadly, the racing queue sits at a natural intersection of queueing theory, contest theory, and mechanism design, and we hope that the initial foray made by our work will pave the way for continued study in this exciting area. For one thing, whereas the winner-take-all prize is a natural starting point motivated from practice, other award schemes could be considered that give some payment to non-winners, which

would be an interesting mechanism design problem for queues with strategic servers. Additionally, future work could consider job abandonment. This might, however, prove more difficult to analyze as some jobs would leave the system without being served, further complicating servers' utility functions. Other possible extensions include heterogeneous servers with different cost functions or heterogeneous job types that induce state-dependent equilibrium behavior. The foundation established in this paper opens the door for this broader agenda.

References

Altman E, Menasché D, Reiffers-Masson A, Datar M, Dhamal S, Touati C, El-Azouzi R (2020) Blockchain competition between miners: A game theoretic perspective. *Frontiers in Blockchain* 2:26.

Anand KS, Paç MF, Veeraraghavan S (2011) Quality–speed conundrum: Trade-offs in customer-intensive services. *Management Science* 57(1):40–56.

Andradóttir S, Ayhan H (2005) Throughput maximization for tandem lines with two stations and flexible servers. *Operations Research* 53(3):516–531.

Andradóttir S, Ayhan H, Down DG (2011) Queueing systems with synergistic servers. *Operations Research* 59(3):772–780.

Armony M, Roels G, Song H (2021) Pooling queues with strategic servers: The effects of customer ownership. *Operations Research* 69(1):13–29.

Biais B, Bisière C, Bouvard M, Casamatta C (2019) The blockchain folk theorem. *The Review of Financial Studies* 32(5):1662–1715.

Büke B, Dos Reis G, Platonov V (2025) Many-server queueing systems with heterogeneous strategic servers in heavy traffic. *Operations Research* (Forthcoming).

Cachon GP, Zhang F (2007) Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Science* 53(3):408–420.

Chen C, Chen Y, Qian P (2025) Incentivizing resource pooling. *Management Science* (Forthcoming).

Chen Z, Ding Y, Yang L (2024) Multi-listing for horizontally differentiated services. *Working Paper, Available at SSRN 4869846* .

Clay Mathematics Institute (2000) The millennium prize problems. Available at `https://www.claymath.org/millennium-problems` (Accessed May 13, 2026).

Dawande M, Janakiraman G, Qi A, Wu Q (2019) Optimal incentive contracts in project management. *Production and Operations Management* 28(6):1431–1445.

Dong J, Ibrahim R (2020) Managing supply in the on-demand economy: Flexible workers, full-time employees, or both? *Operations Research* 68(4):1238–1264.

Fullerton RL, McAfee RP (1999) Auctioning entry into tournaments. *Journal of Political Economy* 107(3):573–605.

Gardner K, Harchol-Balter M, Scheller-Wolf A, Velednitsky M, Zbarsky S (2017) Redundancy-d: The power of d choices for redundancy. *Operations Research* 65(4):1078–1094.

Gardner K, Zbarsky S, Doroudi S, Harchol-Balter M, Hyytia E, Scheller-Wolf A (2016) Queueing with redundant requests: Exact analysis. *Queueing Systems* 83(3):227–259.

Gilbert SM, Weng ZK (1998) Incentive effects favor nonconsolidating queues in a service system: The principal–agent perspective. *Management Science* 44(12):1662–1669.

Gopalakrishnan R, Doroudi S, Ward AR, Wierman A (2016) Routing and staffing when servers are strategic. *Operations Research* 64(4):1033–1050.

HackerOne (2026) HackerOne for Hackers. `https://www.hackerone.com/hackers` (Accessed 04/28/2026).

Hopp WJ, Iravani SMR, Yuen GY (2007) Operations systems with discretionary task completion. *Management Science* 53(1):61–77.

Hu J, Andradóttir S, Ayhan H (2025a) Optimal control of queueing systems with error-prone servers. *Stochastic Systems* 15(3):220–251.

Hu M, Qin T, Wang L, Zhang ZJ (2025b) Subscription vs. spot pricing in on-demand economy. *Working Paper, Available on SSRN 4662398* .

Huberman G, Leshno JD, Moallemi C (2021) Monopoly without a monopolist: An economic analysis of the Bitcoin payment system. *The Review of Economic Studies* 88(6):3011–3040.

Ibrahim R (2018) Managing queueing systems where capacity is random and customers are impatient. *Production and Operations Management* 27(2):234–250.

Loury GC (1979) Market structure and innovation. *The Quarterly Journal of Economics* 93(3):395–410.

Mas-Colell A, Whinston MD, Green JR (1995) *Microeconomic Theory* (Oxford University Press).

Nageswaran L, Scheller-Wolf A (2022) Queues with redundancy: Is waiting in multiple lines fair? *Manufacturing & Service Operations Management* 24(4):1959–1976.

Rosokha Y, Wei C (2024) Cooperation in queueing systems. *Management Science* 70(11):7597–7616.

Terwiesch C, Xu Y (2008) Innovation contests, open innovation, and multiagent problem solving. *Management Science* 54(9):1529–1543.

Tullock G (1980) Efficient rent seeking. Buchanan JM, Tollison RD, Tullock G, eds., *Toward a Theory of the Rent-Seeking Society*, 97–112 (Texas A&M University Press).

Vojnović M (2016) *Contest Theory: Incentive Mechanisms and Ranking Methods* (Cambridge University Press).

Yuan Y (2025) Managing flexible capacity in service systems with worker shortages. *Manufacturing & Service Operations Management* 27(3):808–824.

Zhan D, Ward AR (2019) Staffing, routing, and payment to trade off speed and quality in large service systems. *Operations Research* 67(6):1738–1751.

Zhong Y, Gopalakrishnan R, Ward A (2025a) When strategic customers meet strategic servers: Individual and social optimization in many-server queueing systems. *Working Paper, Available at SSRN 5805802* .

Zhong Y, Gopalakrishnan R, Ward AR (2025b) Behavior-aware queueing: The finite-buffer setting with many strategic servers. *Operations Research* 73(1):290–310.

E-Companion to:

Racing Queues with Strategic Servers

Appendix

In Appendix A, we state and prove some auxiliary lemmas. In the remainder, we present the proofs for the theoretical results in the paper, divided by section: Section 3 (Appendix B); Section 4 (Appendix C); Section 5 (Appendix D); and Section 6 (Appendix E).

A. Auxiliary Technical Lemmas

LEMMA A.1 **(Ruling Out Service Rate of Zero)**. *If $p > c'(0)$, then $\mu = 0$ is not a symmetric equilibrium service rate. Moreover, under Assumption 1, for any $\mu \geq 0$ used by the other $N-1$ servers, it is never optimal for the tagged server to set $\mu_1 = 0$, and also the tagged server's utility $U(\mu_1, \mu)$ is strictly increasing in μ_1 at $\mu_1 = 0$.*

Proof. We begin by proving the first part of the result. Consider μ such that $0 \leq (N-1)\mu < \lambda$ (only $\mu = 0$ is relevant for the first part of the result, but it will be useful in proving the second part of the result to include all μ satisfying this inequality). We have $B(\mu_1 + (N-1)\mu; \lambda) = B(\mu_1; \lambda) = 1$ for $\mu_1 < \lambda - (N-1)\mu$. This implies that $\partial B(\mu_1 + (N-1)\mu; \lambda)/\partial \mu_1 = 0$ for $\mu_1 < \lambda - (N-1)\mu$, and in particular at $\mu_1 = 0$. Differentiating equation (4) then gives

$$\frac{\partial U(\mu_1, \mu)}{\partial \mu_1}\Big|_{\mu_1 = 0} = p - c'(0) > 0, \tag{A.1}$$

where the inequality holds by our assumption that $p > c'(0)$. Thus, there exists $\epsilon > 0$ such that $U(\epsilon, 0) > U(0, 0)$. We conclude that if $\mu = 0$, then it is strictly sub-optimal for the tagged server to set $\mu_1 = 0$. Thus, $\mu = 0$ cannot be a symmetric equilibrium, establishing the first sentence of the result.

We now move to the second part of the result, for which we impose Assumption 1.

- **Case 1:** $0 < (N-1)\mu < \lambda$**.** First, note that Assumption 1 implies $p > c'(0)$ because under Assumption 1, we have $c'(0) = c_E q(0)^{q-1} = 0$, since $q - 1 > 0$. Second, note that here, as in the above, we have $0 \leq (N-1)\mu < \lambda$. Equation (A.1) and what follows it are thus valid in this case, and we conclude that it is strictly sub-optimal for the tagged server to set $\mu_1 = 0$.
- **Case 2:** $(N-1)\mu = \lambda$**.** In this case, $\mu_1 = 0$ is at the boundary between the two pieces of $U(\mu_1, \mu)$ in (5) (the system is critically loaded for $\mu_1 = 0$ but underloaded for any $\mu_1 > 0$). Thus, the function $U(\mu_1, \mu)$ is, strictly speaking, not differentiable at $\mu_1 = 0$ in this case. Let $f_2(\mu_1, \mu)$ denote the expression for the second piece of $U(\mu_1, \mu)$. By the discussion immediately following equation (5), we know that the first and second pieces of $U(\mu_1, \mu)$ coincide at the boundary, which in this case is at $\mu_1 = 0$; this implies that $U(0, \mu) = f_2(0, \mu)$, and we already have that $U(\mu_1, \mu) = f_2(\mu_1, \mu)$ for all $\mu_1 > 0$. In particular, the right

derivative of U at $\mu_1 = 0$ coincides with the derivative of the underloaded branch f_2 evaluated at the same point.

Then, substituting $\mu_1 = 0$, and using the fact that $c'(0) = c_E q(0)^{q-1} = 0$, we get from equation (6) that the right derivative of U at $\mu_1 = 0$ satisfies

$$\frac{\partial f_2(\mu_1,\mu)}{\partial \mu_1}\Big|_{\mu_1=0} = p\left(\frac{\lambda}{(N-1)\mu}\right)\left(1 - \frac{0}{(N-1)\mu}\right) - c'(0) = p\left(\frac{\lambda}{(N-1)\mu}\right) > 0. \tag{A.2}$$

This implies that there exists $\epsilon > 0$ such that $U(\epsilon,\mu) = f_2(\epsilon,\mu) > f_2(0,\mu) = U(0,\mu)$, implying that it is strictly sub-optimal for the tagged server to set $\mu_1 = 0$.

- **Case 3:** $(N-1)\mu > \lambda$**.** In this case, the system is underloaded for any $\mu_1 > 0$. Since in this case we have $U(\mu_1,\mu) = f_2(\mu_1,\mu)$, equation (A.2) gives us that

$$\frac{\partial U(\mu_1,\mu)}{\partial \mu_1}\Big|_{\mu_1=0} = \frac{\partial f_2(\mu_1,\mu)}{\partial \mu_1}\Big|_{\mu_1=0} > 0. \tag{A.3}$$

Analogous reasoning to that used in previous cases then implies that $\mu_1 = 0$ is strictly sub-optimal for the tagged server.

The last part of the result follows from equations (A.1)–(A.3). □

LEMMA A.2 **(Continuity of Equilibrium Service Rate $\mu^\star$).** *Varying one parameter at a time and fixing the others, the unique equilibrium service rate $\mu^\star$ is continuous in $q > 1$, in (the continuous extension of) the number of servers $N \geq 1$, in the arrival rate $\lambda > 0$, in the reward $p > 0$, and in the cost coefficient $c_E > 0$.*

Proof. Throughout the proof, we consider N as a continuous quantity, as the expressions are well-defined even for non-integral values of N. As is clear from the statement, we also restrict attention to $q > 1$, $N \geq 1$, $\lambda > 0$, $p > 0$, and $c_E > 0$.

We first consider the boundary case $N = 1$. For $N = 1$, the underloaded regime is void (see Proposition 1); the only relevant transition is between the overloaded and critical regimes, which occurs at $p = c'(\lambda)$. Continuity at this transition is the $N = 1$ specialization of (A.5) below, namely $\mu_0 = (p/(qc_E))^{1/(q-1)} = \lambda$ when $p = qc_E\lambda^{q-1} = c'(\lambda)$. In the remainder of the proof we therefore assume $N \geq 2$, so that both thresholds in (A.4)–(A.5) are finite.

From Proposition 1, we have that the equilibrium $\mu^\star \in \{\mu_U, \mu_0, \lambda/N\}$. The quantity λ/N is continuous in N and λ and is constant (thus continuous) in q, p, and c_E. Also, it is clear from inspection of equation (8) that both μ_U and μ_0 are continuous (in some cases, constant) in all of these parameters. Thus, the continuity of $\mu^\star$ in all parameters can be established by showing its continuity at the boundaries where it switches from one element of the set $\{\mu_U, \mu_0, \lambda/N\}$ to another. From Proposition 1, we know that these boundaries occur when either

$$p = qc_E\left(\frac{\lambda}{N}\right)^{q-1}\left(\frac{N}{N-1}\right), \tag{A.4}$$

$$\text{or} \quad p = qc_E\left(\frac{\lambda}{N}\right)^{q-1}. \tag{A.5}$$

Specifically, when (A.4) holds, the equilibrium switches between $\mu_{\mathtt{U}}$ and λ/N, and when (A.5) holds, the equilibrium switches between μ_0 and λ/N.

Simple rearrangement of (A.4) gives

$$\frac{p}{c_E}\left(\frac{N-1}{N}\right) = q\left(\frac{\lambda}{N}\right)^{q-1}$$

Substituting the above into the expression for $\mu_{\mathtt{U}}$ from (8), we get

$$\mu_{\mathtt{U}} = \left(\frac{p}{c_E}\left(\frac{\lambda}{qN}\right)\left(\frac{N-1}{N}\right)\right)^{\frac{1}{q}} = \left(q\left(\frac{\lambda}{N}\right)^{q-1}\left(\frac{\lambda}{qN}\right)\right)^{\frac{1}{q}} = \left(\left(\frac{\lambda}{N}\right)^{q}\right)^{\frac{1}{q}} = \frac{\lambda}{N}.$$

The above implies that any transition for $\mu^\star$ between $\mu_{\mathtt{U}}$ and λ/N is continuous, regardless of which parameter is changing among q, N, λ, p, and c_E. Similarly, simple rearrangement of (A.5) gives

$$\frac{p}{c_E} = q\left(\frac{\lambda}{N}\right)^{q-1}.$$

Substituting the above into the expression for μ_0 from (8), we get

$$\mu_0 = \left(\frac{p}{qc_E}\right)^{\frac{1}{q-1}} = \left(q\left(\frac{\lambda}{N}\right)^{q-1}\left(\frac{1}{q}\right)\right)^{\frac{1}{q-1}} = \frac{\lambda}{N}.$$

Thus, any transition for $\mu^\star$ between μ_0 and λ/N is likewise continuous.

By Proposition 1, $\mu^\star$ never transitions directly between $\mu_{\mathtt{U}}$ and μ_0, so we conclude that all transitions of $\mu^\star$ between elements of $\{\mu_{\mathtt{U}}, \mu_0, \lambda/N\}$ are continuous, regardless of which parameter is changing, and this completes the proof. □

LEMMA A.3 (**Candidate equilibrium service rates versus** q).

(i) Regarding $\mu_0(q)$:

(i-a) If $p \le c_E$, then $\mu_0(q)$ is strictly increasing in q for all $q > 1$.

(i-b) If $p > c_E$, then $\mu_0(q)$ is strictly decreasing for $q \in (1, q^\dagger)$ and strictly increasing for $q \in (q^\dagger, \infty)$, where $q^\dagger$ is the unique solution on $(1,\infty)$ to $\log\frac{qc_E}{p} - \frac{q-1}{q} = 0$.

(ii) Regarding $\mu_{\mathtt{U}}(q)$:

(ii-a) If $N = 1$, then $\mu_{\mathtt{U}}(q) = 0$, which is constant in q.

(ii-b) If $N > 1$ and $\frac{ep\lambda(N-1)}{N^2c_E} \le 1$, then $\mu_{\mathtt{U}}(q)$ is strictly increasing in q for all $q > 1$.

(ii-c) If $N > 1$ and $\frac{ep\lambda(N-1)}{N^2c_E} > 1$, then $\mu_{\mathtt{U}}(q)$ is strictly decreasing for $q \in \left(1, \frac{ep\lambda(N-1)}{N^2c_E}\right)$ and strictly increasing for $q \in \left(\frac{ep\lambda(N-1)}{N^2c_E}, \infty\right)$.

Proof.

(i): Differentiating $\mu_0(q)$ in equation (8) yields

$$\mu_0'(q) = -\frac{\left(\frac{p}{qc_E}\right)^{\frac{1}{q-1}}}{q(q-1)^2}\left(q(1+\log\frac{p}{qc_E}) - 1\right) = -\frac{\left(\frac{p}{qc_E}\right)^{\frac{1}{q-1}}}{(q-1)^2}\left(\log\frac{p}{qc_E} + \frac{q-1}{q}\right).$$

Observing that

$$\frac{\left(\frac{p}{qc_E}\right)^{\frac{1}{q-1}}}{(q-1)^2} > 0,$$

we have that the sign of $\mu_0'(q)$ will be the same as the sign of the auxiliary function $\tau(q)$, where

$$\tau(q) := -\left(\log\frac{p}{qc_E} + \frac{q-1}{q}\right) = \log\frac{qc_E}{p} - \frac{q-1}{q}.$$

We also have $\tau'(q) = (q-1)/q^2$. For q strictly positive, we thus have $\tau'(q) \lesseqgtr 0 \iff q \lesseqgtr 1$. That is, $\tau(q)$ strictly decreases to a minimum at $q = 1$, then strictly increases thereafter. (Under Assumption 1, we are only interested in $q > 1$, but we make the preceding clarification because we soon use $\tau(1)$ as part of our argument.) Note also that $\tau(1) = \log(c_E/p)$.

- If $p \le c_E$, then $\tau(1) \ge 0$, and since $\tau(q)$ strictly increases after its minimum at $q = 1$, for all $q > 1$ we have $\tau(q) > \tau(1) \ge 0$, so $\tau(q) > 0$. Since $\tau(q)$ and $\mu_0'(q)$ have the same sign, this in turn implies that $\mu_0(q)$ is strictly increasing in q for all $q > 1$, which completes part (i-a).
- If $p > c_E$, then $\tau(1) < 0$. Observing that $\tau(q)$ increases without bound as q increases (i.e., $\lim_{q\to\infty}\tau(q) = \infty$), and again using the fact that $\tau(q)$ strictly increases after its minimum at $q = 1$, we conclude that there exists a unique $q^\dagger > 1$ such that $\tau(q) \lesseqgtr 0 \iff q \lesseqgtr q^\dagger$. Since $\tau(q)$ and $\mu_0'(q)$ have the same sign, this in turn implies that $\mu_0(q)$ is strictly decreasing in q for $q \in (1, q^\dagger)$, and strictly increasing in q for $q \in (q^\dagger, \infty)$. This completes part (i-b).

(ii): Inspection of equation (8) reveals that we have $\mu_U(q) = 0$ whenever $N = 1$, which establishes part (ii-a). For parts (ii-b) and (ii-c), which we prove next, by assumption we have $N > 1$.

Differentiating $\mu_U(q)$ in equation (8) yields

$$\mu_U'(q) = -\frac{\left(p\frac{\lambda}{Nqc_E}(1-\frac{1}{N})\right)^{\frac{1}{q}}}{q^2}\left(1 + \log\frac{p\lambda}{Nqc_E}(1-\frac{1}{N})\right).$$

Observing that

$$\frac{\left(p\frac{\lambda}{Nqc_E}(1-\frac{1}{N})\right)^{\frac{1}{q}}}{q^2} > 0,$$

we have that the sign of $\mu_U'(q)$ is the same as the sign of the auxiliary function $\eta(q)$, where

$$\eta(q) := -\left(1 + \log\frac{p\lambda}{Nqc_E}(1-\frac{1}{N})\right).$$

Note that $\eta'(q) = 1/q > 0$, so $\eta(q)$ is strictly increasing in q. Setting $\eta(q) = 0$ and isolating q gives the unique solution

$$q = \frac{ep\lambda}{Nc_E}(1-\frac{1}{N}),$$

and thus, we have $\eta(q) \lesseqgtr 0 \iff q \lesseqgtr \frac{ep\lambda}{Nc_E}(1-\frac{1}{N})$.

- If $\frac{ep\lambda}{Nc_E}(1-\frac{1}{N}) \leq 1$, then we have $\eta(q) > 0$ for all $q > 1$. Since $\eta(q)$ and $\mu_{\mathrm{U}}'(q)$ have the same sign, this implies that $\mu_{\mathrm{U}}(q)$ is strictly increasing in q for all $q > 1$. This completes part (ii-b).
- If $\frac{ep\lambda}{Nc_E}(1-\frac{1}{N}) > 1$, then the zero of $\eta(q)$ is strictly larger than 1. Analogous reasoning therefore implies that $\mu_{\mathrm{U}}(q)$ is strictly decreasing in q for $q \in (1, \frac{ep\lambda}{Nc_E}(1-\frac{1}{N}))$, and strictly increasing in q for $q \in (\frac{ep\lambda}{Nc_E}(1-\frac{1}{N}), \infty)$. This completes part (ii-c).

□

A.1. Technical Lemma for Proposition 5

LEMMA A.4. *Suppose there exist $\hat{q} > 1$ and $q_4 > \hat{q}$ such that $\mu^\star(q) = \lambda/N$ for $q \in (\hat{q}, q_4)$ and $\mu^\star(q) = \mu_{\mathrm{U}}(q)$ for all $q \geq q_4$. Then $\mu^\star(q) = \mu_{\mathrm{U}}(q)$ is strictly increasing on (q_4, ∞).*

Proof. By Lemma A.3(ii), $\mu_{\mathrm{U}}(q)$ is either constant in q and equal to zero when $N = 1$; strictly increasing in q; or first strictly decreasing and then strictly increasing in q. We know that $\mu = 0$ is never an equilibrium (see Lemma A.1), so the $N = 1$ case in Lemma A.3(ii-a) can never hold simultaneously with the hypothesis of this case, which requires that $\mu_{\mathrm{U}}(q)$ be an equilibrium for some q.

If Lemma A.3(ii-b) holds, then $\mu_{\mathrm{U}}(q)$ is strictly increasing in q for all $q > 1$, so the result follows immediately.

Finally, suppose Lemma A.3(ii-c) holds. In this case, $\mu_{\mathrm{U}}(q)$ is first strictly decreasing, then strictly increasing in q. Suppose that the hypothesis of the current result holds, and assume by way of contradiction that there exists an interval (q', q'') with $q' \geq q_4$ such that $\mu_{\mathrm{U}}(q)$ is strictly decreasing in q on this interval. Since $\mu_{\mathrm{U}}(q)$ is first strictly decreasing, then strictly increasing in q, this implies that $\mu_{\mathrm{U}}(q)$ is strictly decreasing in q for all $q \leq q''$.

By continuity of $\mu^\star$ (Lemma A.2) and the hypothesis that $\mu^\star(q) = \lambda/N$ on $(\hat{q}, q_4)$ and $\mu^\star(q) = \mu_{\mathrm{U}}(q)$ for $q \geq q_4$, we have $\mu_{\mathrm{U}}(q_4) = \mu^\star(q_4) = \lambda/N$. By continuity of $\mu_{\mathrm{U}}(q)$ and the fact that $\mu_{\mathrm{U}}(q)$ is strictly decreasing for all $q < q''$ (where recall that $q'' > q' \geq q_4$), there thus must exist $\epsilon > 0$ such that $\mu^\star(q_4 + \epsilon) = \mu_{\mathrm{U}}(q_4 + \epsilon) < \lambda/N$. But this contradicts Proposition 1 because from that result, we know that we cannot have both $\mu^\star = \mu_{\mathrm{U}}$ and $\mu_{\mathrm{U}} < \lambda/N$.

Thus, we conclude that, under the hypothesis of this case, we have $\mu^\star(q) = \mu_{\mathrm{U}}(q)$ strictly increasing in $q \in (q_4, \infty)$. □

B. Proofs from Section 3

Proof of Lemma 1. The service rate μ of the non-tagged servers is fixed in this proof, so we suppress it from our notation for brevity. For convenience, define

$$g_1(\mu_1) := p - c'(\mu_1), \quad \text{and} \quad g_2(\mu_1) := \frac{p\lambda(N-1)\mu}{(\mu_1 + (N-1)\mu)^2} - c'(\mu_1), \tag{B.1}$$

where $g_1(\mu_1)$ (for $\mu_1 < \lambda - (N-1)\mu$) and $g_2(\mu_1)$ (for $\mu_1 > \lambda - (N-1)\mu$) are equal to the first and second pieces, respectively, of $\partial U(\mu_1, \mu)/\partial \mu_1$ from equation (6), after simplification in the case of g_2. We note for

later reference that $c'(\mu_1) = qc_E\mu_1^{q-1}$ is strictly increasing because $c''(\mu_1) = q(q-1)c_E\mu_1^{q-2} > 0$,[2] so $g_1(\mu_1)$ is strictly decreasing.

Additionally, from the second piece of equation (6), we have

$$\lim_{\mu_1 \downarrow \lambda-(N-1)\mu} \frac{\partial U(\mu_1,\mu)}{\partial \mu_1} = p\left(\frac{(N-1)\mu}{\lambda}\right) - c'(\lambda-(N-1)\mu). \tag{B.2}$$

Thus, after comparing to the first piece of equation (6), we can deduce that if $\mu < \lambda/(N-1)$, then $\lim_{\mu_1\downarrow\lambda-(N-1)\mu} \partial U(\mu_1,\mu)/\partial\mu_1 < \lim_{\mu_1\uparrow\lambda-(N-1)\mu} \partial U(\mu_1,\mu)/\partial\mu_1$.

Single Server ($N = 1$): First, note that in this case, $\lambda-(N-1)\mu = \lambda$, and that we have $g_2(\mu_1) = -c'(\mu_1) = -qc_E\mu_1^{q-1} < 0$ for all $\mu_1 > \lambda$. Thus, in the single-server case, it cannot be optimal for the server to operate the system as underloaded (i.e., to set $\mu_1 > \lambda$), and we can ignore the second piece of the function in equations (5) and (6). Also, from Lemma A.1, we know that $\mu_1 = 0$ is sub-optimal and that the single server's utility is strictly increasing in μ_1 at $\mu_1 = 0$ (the proof of the lemma also shows specifically that $g_1(0) > 0$: see equation (A.1)). Thus, the optimal service rate must be on the interval $(0,\lambda)$. There are two possibilities:

- If $g_1(\lambda) \geq 0$, then since g_1 is strictly decreasing, we have for all $\mu_1 < \lambda$ that $\partial U(\mu_1,\mu)/\partial\mu_1 = g_1(\mu_1) > g_1(\lambda) \geq 0$. Thus, the tagged server's utility is strictly increasing for $\mu_1 \in [0,\lambda)$ and then decreasing thereafter (as just argued above), in which case there is no solution to the FOC, and the unique global maximizer is $\mu_1 = \lambda$. So, in this case, point (ii) holds in the statement of the result.[3]
- On the other hand, suppose $g_1(\lambda) < 0$. We have $g_1(0) = p - c'(0) = p > 0$. Then, since $g_1(\lambda) < 0$ and g_1 is strictly decreasing, the tagged server's utility increases until a unique point where $g_1(\mu_1) = 0$ (the unique solution to the FOC), then decreases thereafter, and this μ_1 is thus the unique global maximizer. So, in this case, point (i) holds in the statement of the result.

In the rest of the proof, we handle the multi-server case with $N > 1$.

Multiple Servers ($N > 1$): For $N > 1$, we first handle the case where $\mu \geq \lambda/(N-1)$. In this case, we can ignore the first pieces of equations (5) and (6) and restrict attention to the second pieces (the first branch is infeasible if the above inequality holds strictly; if equality holds, we have $\mu = \lambda/(N-1)$, and the branch boundary falls at $\mu_1 = 0$, at which point the two pieces of U coincide, so the second piece is again valid for all $\mu_1 \geq 0$). We know from Lemma A.1 and its proof that $U(\mu_1,\mu)$ is increasing in μ_1 at $\mu_1 = 0$ and that $g_2(0) > 0$ (g_2 is the derivative of the auxiliary function f_2 defined in that proof). Combined with the observation that $\lim_{\mu_1\to\infty} g_2(\mu_1) = -\infty$ and the fact that $g_2(\mu_1)$ is strictly decreasing (this follows because the fractional term in g_2 is strictly decreasing in μ_1 and so is $-c'(\mu_1)$), this implies that $U(\mu_1,\mu)$ is unimodal in μ_1, there is

[2] We have $c''(0) = 0$ for $q > 2$ and $c''(0)$ not well-defined for $1 < q < 2$; nevertheless, for any $\epsilon > 0$ and $q > 1$, $c'(\epsilon) = qc_E(\epsilon)^{q-1} > 0 = c'(0)$, so $c'(\mu_1)$ is indeed strictly increasing even when we include $\mu_1 = 0$.

[3] If $g_1(\lambda) = 0$, then we might consider the optimizer $\mu_1 = \lambda$ to be a solution to the FOC. However, strictly speaking, $U(\mu_1,\mu)$ is not differentiable in μ_1 at this point. We thus classify the $g_1(\lambda) = 0$ case as falling under point (ii) of the result, although this is arguably a matter of semantics.

exactly one solution to $g_2(\mu_1) = 0$ (and thus exactly one solution to the FOC, since the first piece of the function is irrelevant here), and this solution is the unique global maximizer.

Finally, we cover the case with $N > 1$ and $\mu < \lambda/(N-1)$. As noted above, $g_1(\mu_1)$ is strictly decreasing, and $g_1(0) > 0$. There can thus be at most one solution to $g_1(\mu_1) = 0$ for $\mu_1 \in [0, \lambda - (N-1)\mu)$. Similarly, we have that $g_2(\mu_1)$ is strictly decreasing for all $\mu_1 > 0$. Thus, there can be at most one solution to $g_2(\mu_1) = 0$ for $\mu_1 \in (\lambda - (N-1)\mu, \infty)$.

Case 1: $g_1(\lambda - (N-1)\mu) < 0$**.** Since we know $g_1(0) > 0$, in this case, there is exactly one solution (which is strictly positive) to $\partial U(\mu_1, \mu)/\partial \mu_1 = 0$ over $[0, \lambda - (N-1)\mu)$, and this solution is the unique maximizer of U on this interval. In addition, since $g_2(\mu_1) \leq g_2(\lambda - (N-1)\mu) < g_1(\lambda - (N-1)\mu) < 0$ for all $\mu_1 \geq \lambda - (N-1)\mu$ (the ordering of g_1 and g_2 at the boundary follows from equation (B.2) and its surroundings), we have that the function $U(\mu_1, \mu)$ is strictly decreasing in μ_1 for $\mu_1 \geq \lambda - (N-1)\mu$. Thus, the solution to $g_1(\mu_1) = 0$, found on $(0, \lambda - (N-1)\mu)$, is the unique solution to the FOC and the unique global maximizer in μ_1 of $U(\mu_1, \mu)$.

Case 2: $g_1(\lambda - (N-1)\mu) = 0$**.** In this case, for any $\mu_1 \geq \lambda - (N-1)\mu$, we have $g_2(\mu_1) \leq g_2(\lambda - (N-1)\mu) < g_1(\lambda - (N-1)\mu) = 0$. Thus, the utility $U(\mu_1, \mu)$ is strictly decreasing in μ_1 for all $\mu_1 \geq \lambda - (N-1)\mu$. Also, since $g_1(\mu_1)$ is strictly decreasing, we have $g_1(\mu_1) > 0$ for all $0 \leq \mu_1 < \lambda - (N-1)\mu$. Thus, $U(\mu_1, \mu)$ is strictly increasing in $\mu_1 \in [0, \lambda - (N-1)\mu)$. By definition, then, $U(\mu_1, \mu)$ is uniquely maximized in μ_1 at $\mu_1 = \lambda - (N-1)\mu$. (The function U is not differentiable in μ_1 at this point, so there is technically no solution to the FOC, and we classify this case under point (ii) of the result statement.)

Case 3: $g_1(\lambda - (N-1)\mu) > 0$ **and** $g_2(\lambda - (N-1)\mu) > 0$**.** In this case, U is strictly increasing in μ_1 over $[0, \lambda - (N-1)\mu]$ because $g_1(\mu_1)$ is strictly decreasing. The second condition for the case also implies that U continues to increase in μ_1 over $[\lambda - (N-1)\mu, \lambda - (N-1)\mu + \epsilon)$ for some $\epsilon > 0$. Since $g_2(\mu_1)$ is strictly decreasing and $\lim_{\mu_1 \to \infty} g_2(\mu_1) = -\infty$, there is exactly one solution—and this solution satisfies $\mu_1 > \lambda - (N-1)\mu$—to the FOC, which solution is also the unique global maximizer of $U(\mu_1, \mu)$.

Case 4: $g_1(\lambda - (N-1)\mu) > 0$ **and** $g_2(\lambda - (N-1)\mu) \leq 0$**.** In this case, $U(\mu_1, \mu)$ is strictly increasing in μ_1 over $[0, \lambda - (N-1)\mu]$, then strictly decreasing thereafter. Thus, the unique global maximum is the critically loaded point $\mu_1 = \lambda - (N-1)\mu$. (In the case where $g_2(\lambda - (N-1)\mu) = 0$, as in other instances, we still classify this case under point (ii) of the result statement because there is technically no solution to the FOC.) □

Proof of Proposition 1. For ease of notation, let $\text{FOC}(\mu) := \left.\frac{\partial U(\mu_1, \mu)}{\partial \mu_1}\right|_{\mu_1 = \mu}$ as given in equation (7), which is defined everywhere except at $\mu = \lambda/N$. Then, let $\text{FOC}(\frac{\lambda}{N}-) := \lim_{\mu \uparrow \frac{\lambda}{N}} \text{FOC}(\mu)$, and let $\text{FOC}(\frac{\lambda}{N}+) := \lim_{\mu \downarrow \frac{\lambda}{N}} \text{FOC}(\mu)$.

By Lemma A.1, (i) $\mu = 0$ is not a symmetric equilibrium and (ii) $\mu_1 = 0$ is not optimal for the tagged server for any $\mu \geq 0$ in use by the $N-1$ other servers. We can thus ignore $\mu = 0$ and focus on $\mu > 0$; we do so in the remainder of the proof.

Next, because $c''(\mu) = c_E q(q-1)\mu^{q-2} > 0$ for all $\mu > 0$, we can deduce from equation (7) that $\text{FOC}(\mu)$ is strictly decreasing in μ for $\mu \in (0, \frac{\lambda}{N})$, and also strictly decreasing in μ for $\mu \in (\frac{\lambda}{N}, \infty)$. Additionally, we have $\text{FOC}(\frac{\lambda}{N}-) = p - qc_E(\frac{\lambda}{N})^{q-1} > p(1-\frac{1}{N}) - qc_E(\frac{\lambda}{N})^{q-1} = \text{FOC}(\frac{\lambda}{N}+)$. To summarize, $\text{FOC}(\mu)$ is a strictly decreasing function over $(0,\infty)$, with a single hole at $\frac{\lambda}{N}$, where it jumps downward. This implies that there is at most a single value of $\mu \in (0,\infty)$ that satisfies $\text{FOC}(\mu) = 0$.

Additionally, we note that for any $\mu \in (0, \lambda/N) \cup (\lambda/N, \infty)$, $\text{FOC}(\mu) \neq 0$ implies that μ is not an equilibrium because by definition, it means that at $\mu_1 = \mu$, the tagged server has an improving direction for her utility. We next prove parts (i)-(iii) of the result statement in turn.

• Part (i): The condition $p < c'(\frac{\lambda}{N})$ is equivalent to $\text{FOC}(\frac{\lambda}{N}-) = p - qc_E(\frac{\lambda}{N})^{q-1} < 0$, and combined with the downward jump established above, this gives $\text{FOC}(\frac{\lambda}{N}+) < \text{FOC}(\frac{\lambda}{N}-) < 0$. Thus, since $\text{FOC}(\mu)$ is strictly decreasing, for any $\mu > \frac{\lambda}{N}$, we must have $\text{FOC}(\mu) < 0$, so such μ cannot be an equilibrium. Evaluating the first piece of (7) at $\mu \downarrow 0$ gives $\text{FOC}(0+) = p - c'(0) = p > 0$, and combined with FOC being strictly decreasing on $(0, \frac{\lambda}{N})$ and $\text{FOC}(\frac{\lambda}{N}-) < 0$, we conclude that there is a unique solution $\mu_0 \in (0, \frac{\lambda}{N})$ such that $\text{FOC}(\mu_0) = 0$. Since μ_0 solves the symmetric FOC in (7), it also solves the asymmetric FOC in (6) for $\mu = \mu_0$ (because (7) is merely the special case of (6) when $\mu_1 = \mu$); Lemma 1 then implies that μ_0 is the unique global maximizer in μ_1 of $U(\mu_1, \mu_0)$, i.e., μ_0 is a symmetric equilibrium service rate by our definition in (EQ).

Besides μ_0, every other $\mu \in (0, \frac{\lambda}{N})$ is not an equilibrium because an improving direction exists for the tagged server's service rate. Finally, we must rule out $\mu = \frac{\lambda}{N}$. It is straightforward to show that $\frac{\lambda}{N}$ is not an equilibrium by noting that $g_1(\frac{\lambda}{N}) = \text{FOC}(\frac{\lambda}{N}-)$ for $\mu = \frac{\lambda}{N}$. The conditions for this case therefore give us that $g_1(\frac{\lambda}{N}) < 0$, which implies that $U(\mu_1, \frac{\lambda}{N})$ is strictly decreasing in a neighborhood immediately below $\frac{\lambda}{N}$. Thus, for $\epsilon > 0$ sufficiently small, we have $U(\frac{\lambda}{N} - \epsilon, \frac{\lambda}{N}) > U(\frac{\lambda}{N}, \frac{\lambda}{N})$, implying that $\frac{\lambda}{N}$ is not an equilibrium. We conclude that μ_0, which falls on the interval $(0, \frac{\lambda}{N})$, is the unique equilibrium. The expression for μ_0 (shown in (8)) is obtained by isolating μ in the first piece of the FOC in (7).

• Part (ii): The condition for this case implies $\text{FOC}(\frac{\lambda}{N}+) \leq 0 \leq \text{FOC}(\frac{\lambda}{N}-)$ Since $\text{FOC}(\mu)$ is strictly decreasing, in this case, we have $\text{FOC}(\mu) \neq 0$ for all $\mu \neq \frac{\lambda}{N}$, and thus, all such μ are not equilibria. It remains to consider $\mu = \frac{\lambda}{N}$. Note that $\text{FOC}(\frac{\lambda}{N}-) = g_1(\frac{\lambda}{N})$ and $\text{FOC}(\frac{\lambda}{N}+) = g_2(\frac{\lambda}{N})$, where g_1 and g_2, defined in the proof of Lemma 1, are the two pieces of the partial derivative of $U(\mu_1, \mu)$ with respect to μ_1 for a particular μ, in this case $\mu = \frac{\lambda}{N}$. The same proof shows that g_1 and g_2 are both strictly decreasing functions of the tagged server's service rate μ_1. We thus have $g_1(\mu_1) > 0$ for $0 < \mu_1 < \frac{\lambda}{N}$, and $g_2(\mu_1) < 0$ for $\frac{\lambda}{N} < \mu_1$, i.e., $U(\mu_1, \frac{\lambda}{N})$ is strictly increasing in μ_1 below $\frac{\lambda}{N}$, and strictly decreasing thereafter. By definition, then, and recalling that $U(\mu_1, \mu)$ is continuous despite the kink at $\frac{\lambda}{N}$, we have that $\mu_1 = \frac{\lambda}{N}$ is the unique global maximizer in μ_1 of $U(\mu_1, \frac{\lambda}{N})$, i.e., $\mu^\star = \frac{\lambda}{N}$ is an equilibrium. It is also unique since we have eliminated above all other candidates, which completes the proof of part (ii).

• Part (iii): The condition $p > \frac{N}{N-1}c'(\frac{\lambda}{N})$ is equivalent to $\text{FOC}(\frac{\lambda}{N}+) = p(1-\frac{1}{N}) - qc_E(\frac{\lambda}{N})^{q-1} > 0$, hence by the downward jump $\text{FOC}(\frac{\lambda}{N}-) > \text{FOC}(\frac{\lambda}{N}+) > 0$. By the argument symmetric to part (i), $(0, \frac{\lambda}{N}]$ contains

no equilibrium: on $(0, \frac{\lambda}{N})$, FOC is strictly decreasing with $\text{FOC}(\frac{\lambda}{N}-) > 0$, so $\text{FOC}(\mu) > 0$ throughout; and at $\mu = \frac{\lambda}{N}$, $g_2(\frac{\lambda}{N}) = \text{FOC}(\frac{\lambda}{N}+) > 0$ implies $U(\mu_1, \frac{\lambda}{N})$ is strictly increasing in μ_1 in a right-neighborhood of $\frac{\lambda}{N}$, giving the tagged server an improving direction. It thus remains to analyze $\mu \in (\frac{\lambda}{N}, \infty)$. Note that $\lim_{\mu\to\infty} \text{FOC}(\mu) = -\infty$: in the second piece of (7), the first term vanishes as $\mu \to \infty$, while $-c'(\mu) = -qc_E\mu^{q-1} \to -\infty$. Combined with FOC being strictly decreasing on $(\frac{\lambda}{N}, \infty)$ and $\text{FOC}(\frac{\lambda}{N}+) > 0$, there is a unique solution $\mu_{\text{U}} \in (\frac{\lambda}{N}, \infty)$ to $\text{FOC}(\mu_{\text{U}}) = 0$, and the same Lemma 1 invocation as in part (i) confirms that μ_{U} is a symmetric equilibrium service rate. Combining with the exclusions above, μ_{U} is the unique symmetric equilibrium, which completes the proof of part (iii). The expression for μ_{U} (shown in (8)) is obtained by isolating μ in the second piece of the FOC in (7). For $N = 1$, parts (i) and (ii) reduce to: $\mu^\star = \mu_0$ if $p < c'(\lambda)$, and $\mu^\star = \lambda$ if $p \geq c'(\lambda)$ (consistent with the convention adopted in Section 3).

For $N = 1$, parts (i) and (ii) reduce to: $\mu^\star = \mu_0$ if $p < c'(\lambda)$, and $\mu^\star = \lambda$ if $p \geq c'(\lambda)$. □

C. Proofs from Section 4

Proof of Lemma 2. Rearrangement of (8) gives

$$\mu_{\text{U}}(N) = \left(\frac{p\lambda}{qc_E}\right)^{1/q} \left(\frac{N-1}{N^2}\right)^{1/q}.$$

The first term in parentheses is positive and constant in N, so $\mu_{\text{U}}(N)$ inherits the monotonicity of $\left((N-1)/N^2\right)^{1/q}$ in N. Letting $\nu(x) := x^{1/q}$, we have

$$\frac{\text{d}}{\text{d}N}\left(\left(\frac{N-1}{N^2}\right)^{1/q}\right) = \frac{\text{d}}{\text{d}N}\left(\nu(\frac{N-1}{N^2})\right) = \nu'(\frac{N-1}{N^2})\frac{\text{d}}{\text{d}N}\left(\frac{N-1}{N^2}\right) = \nu'(\frac{N-1}{N^2})\left(\frac{2-N}{N^3}\right).$$

Differentiating ν (and recalling that $q > 1$) yields

$$\nu'(x) = \frac{x^{-\left(\frac{q-1}{q}\right)}}{q} > 0 \text{ for } x \geq 0.^4 \tag{C.1}$$

On the interval $[1, \infty)$ for N, we have $(2-N)/N^3 \gtreqless 0 \iff N \gtreqless 2$, and since also $\nu'((N-1)/N^2) > 0$ for $N \geq 1$ by equation (C.1), we conclude that

$$\frac{\text{d}}{\text{d}N}\left(\frac{N-1}{N^2}\right)^{1/q} \lesseqgtr 0 \iff N \gtreqless 2.$$

As shown above, $\mu'_{\text{U}}(N)$ has the same sign as this derivative, and thus the proof for $\mu_{\text{U}}(N)$ is complete.

Finally, that μ_0 is constant in N can be seen immediately by inspection of equation (8). □

[4] The derivative $\nu'(x)$ grows without bound as $x \downarrow 0$, but its sign is not in doubt.

Proof of Proposition 2. From Proposition 1, the equilibrium service rate $\mu^\star(N)$ is given by

$$\mu^\star(N) = \begin{cases} \mu_{\mathrm{O}}(N) = \left(\frac{p}{qc_E}\right)^{\frac{1}{q-1}} < \frac{\lambda}{N}, & \text{if } \frac{\lambda^{q-1}qc_E}{p} > N^{q-1}, \\ \frac{\lambda}{N}, & \text{if } (N-1)N^{q-2} \le \frac{\lambda^{q-1}qc_E}{p} \le N^{q-1}, \\ \mu_{\mathrm{U}}(N) = \left(p\frac{\lambda}{Nqc_E}\left(1-\frac{1}{N}\right)\right)^{\frac{1}{q}} > \frac{\lambda}{N}, & \text{if } \frac{\lambda^{q-1}qc_E}{p} < (N-1)N^{q-2}. \end{cases} \tag{C.2}$$

Hence, as N increases, the equilibrium service rate transitions from overloaded to critically loaded, and eventually to underloaded. Observe that:

- In the overloaded regime, $\mu^\star(N)$ is independent of N.
- In the critically loaded regime, $\mu^\star(N) = \lambda/N$, which strictly decreases in N.
- In the underloaded regime, from Lemma 2, $\mu^\star(N)$ strictly increases in N for $1 \le N \le 2$, and strictly decreases for $N \ge 2$.

Define the thresholds:

$$N_1 := \left(\frac{\lambda^{q-1}qc_E}{p}\right)^{\frac{1}{q-1}},$$

and

$$N_2 := \text{the unique solution to } (N-1)N^{q-2} = \frac{\lambda^{q-1}qc_E}{p}$$

Clearly, $N_2 > 1$ since the left-hand side of the above equation is strictly increasing in N (because the derivative of the left-hand side is given by $N^{q-3}((q-1)N-(q-2)) > 0$, noting that $N \ge 1$), equals to 0 when $N = 1$, and diverges to ∞ when $N \to \infty$ (which is straightforward when $q > 2$; when $1 < q \le 2$, $\lim_{N\to\infty} \frac{N-1}{N^{2-q}} = \frac{1}{(2-q)N^{1-q}} = \frac{N^{q-1}}{2-q} = \infty$, where the first equality follows from l'Hôpital's rule.)

To proceed, we discuss the following two cases depending on the value of N_2.

Case 1: $N_2 \ge 2$**.** Equivalently, $\frac{\lambda^{q-1}qc_E}{p} \ge 2^{q-2}$. In this case, whenever the equilibrium is underloaded (i.e., when $N > N_2$), the equilibrium service rate is strictly decreasing in N because $N > N_2 \ge 2$. Since equilibria in both the critically loaded and overloaded regimes are also non-increasing in N, and because the equilibrium $\mu^\star$ is continuous in (the continuous extension of) $N \ge 1$ by Lemma A.2, it follows that $\mu^\star(N)$ is non-increasing in N for all $N \ge 1$.

Case 2: $1 < N_2 < 2$**.** Equivalently, $\frac{\lambda^{q-1}qc_E}{p} < 2^{q-2}$. Here the monotonicity depends on the relative values of $\mu^\star(1)$ and $\mu^\star(2)$:

- If $\mu^\star(1) \ge \mu^\star(2)$, then $\mu^\star(N)$ is non-increasing in N for all $N \ge 1$.
- If $\mu^\star(1) < \mu^\star(2)$, then $\mu^\star(N)$ increases from $N = 1$ to $N = 2$ and decreases thereafter, implying non-monotonicity.

Observe that, when $\frac{\lambda^{q-1}qc_E}{p} < 2^{q-2}$,

$$\mu^\star(2) = \left(p\frac{\lambda}{2qc_E}\left(1-\frac{1}{2}\right)\right)^{\frac{1}{q}} = \left(\frac{p\lambda}{4qc_E}\right)^{\frac{1}{q}}.$$

Next, we discuss two cases below.

Case 2-i: If $\frac{\lambda^{q-1}qc_E}{p} \le 1$, we have $\mu^\star(1) = \frac{\lambda}{1} = \lambda$. Then, $\mu^\star(1) < \mu^\star(2)$ becomes $\left(\frac{p\lambda}{4qc_E}\right)^{\frac{1}{q}} > \lambda$, which can be simplified as $\frac{\lambda^{q-1}qc_E}{p} < \frac{1}{4}$.

Case 2-ii: If $\frac{\lambda^{q-1}qc_E}{p} > 1$, we have $\mu^\star(1) = \mu_0(1) < \lambda$. Then, $\mu^\star(1) < \mu^\star(2)$ becomes $\frac{\lambda^{q-1}qc_E}{p} > 4^{q-1}$. Note that $4^{q-1} > 2^{q-2}$, $\frac{\lambda^{q-1}qc_E}{p} > 4^{q-1}$ can never happen when the condition of Case 2 holds.

Therefore,

- if $\frac{\lambda^{q-1}qc_E}{p} < \frac{1}{4}$, $\mu^\star(1) < \mu^\star(2)$ and $\mu^\star(N) \ge \mu^\star(N+1)$ for all $N \ge 2$.
- Otherwise, $\mu^\star(N) \ge \mu^\star(N+1)$ for all $N \ge 1$.

□

Proof of Proposition 3. We show $(N+1)\mu^\star(N+1) \ge N\mu^\star(N)$ for all $N \ge 1$. As N increases, the equilibrium can only transition from overloaded to critically loaded to underloaded, because both regime thresholds in Proposition 1—namely $c'(\lambda/N)$ and $\frac{N}{N-1}c'(\lambda/N)$—are strictly decreasing in N: the former because c' is increasing and λ/N is decreasing; the latter because $\frac{N}{N-1}c'(\lambda/N) = qc_E\lambda^{q-1}/[(N-1)N^{q-2}]$ and the denominator $(N-1)N^{q-2}$ is strictly increasing in N for $N \ge 1$ (its derivative is $N^{q-3}[(q-1)N-(q-2)] > 0$). This ordering ensures that the following six cases are exhaustive.

Case 1: Both $\mu^\star(N)$ and $\mu^\star(N+1)$ are overloaded. Then $\mu^\star(N) = \mu^\star(N+1) = \mu_0$, so $(N+1)\mu_0 > N\mu_0$.

Case 2: $\mu^\star(N)$ is overloaded and $\mu^\star(N+1)$ is critically loaded. Then $N\mu^\star(N) = N\mu_0 < \lambda$ (the system at N is overloaded) and $(N+1)\mu^\star(N+1) = \lambda$.

Case 3: $\mu^\star(N)$ is overloaded and $\mu^\star(N+1)$ is underloaded. Then $N\mu^\star(N) = N\mu_0 < \lambda < (N+1)\mu_{\mathrm{U}}(N+1) = (N+1)\mu^\star(N+1)$.

Case 4: Both $\mu^\star(N)$ and $\mu^\star(N+1)$ are critically loaded. Then $N\mu^\star(N) = \lambda = (N+1)\mu^\star(N+1)$.

Case 5: $\mu^\star(N)$ is critically loaded and $\mu^\star(N+1)$ is underloaded. Then $N\mu^\star(N) = \lambda$ and $(N+1)\mu^\star(N+1) = (N+1)\mu_{\mathrm{U}}(N+1) > (N+1)\cdot\frac{\lambda}{N+1} = \lambda$.

Case 6: Both $\mu^\star(N)$ and $\mu^\star(N+1)$ are underloaded. From (8), we have

$$N\mu_{\mathrm{U}}(N) = \left(\frac{p\lambda}{qc_E}\right)^{\frac{1}{q}} \left(N^{q-2}(N-1)\right)^{\frac{1}{q}}.$$

Define $f(N) := N^{q-2}(N-1)$. Then

$$f'(N) = N^{q-3}\big[(q-1)N-(q-2)\big].$$

For $N \ge 1$ and $q > 1$, we have $(q-1)N-(q-2) \ge (q-1)-(q-2) = 1 > 0$ and $N^{q-3} > 0$, so $f'(N) > 0$. Since f is strictly increasing and the prefactor $\left(\frac{p\lambda}{qc_E}\right)^{1/q}$ is a positive constant, $N\mu_{\mathrm{U}}(N)$ is strictly increasing in N. Hence $(N+1)\mu_{\mathrm{U}}(N+1) > N\mu_{\mathrm{U}}(N)$.

Equality condition. From Cases 1–6, equality holds if and only if both $\mu^\star(N)$ and $\mu^\star(N+1)$ are critically loaded, i.e., $\mu^\star(N) = \lambda/N$ and $\mu^\star(N+1) = \lambda/(N+1)$. By Proposition 1(ii), $\mu^\star(N) = \lambda/N$ if and only if

$c'(\lambda/N) \le p \le \frac{N}{N-1}c'(\lambda/N)$, and $\mu^\star(N+1) = \lambda/(N+1)$ if and only if $c'(\lambda/(N+1)) \le p \le \frac{N+1}{N}c'(\lambda/(N+1))$. The joint condition is the intersection of these two intervals. Since $\lambda/(N+1) < \lambda/N$ and c' is strictly increasing, $c'(\lambda/(N+1)) < c'(\lambda/N)$, so the binding lower bound is $c'(\lambda/N)$. For the upper bounds, as shown above, the threshold $\frac{N}{N-1}c'(\lambda/N)$ is strictly decreasing in N, so $\frac{N+1}{N}c'(\lambda/(N+1)) \le \frac{N}{N-1}c'(\lambda/N)$, and the binding upper bound is $\frac{N+1}{N}c'(\lambda/(N+1))$. Hence the equality condition is equivalent to $c'(\lambda/N) \le p \le \frac{N+1}{N}c'(\lambda/(N+1))$.

□

D. Proofs from Section 5

Proof of Proposition 4. We must show that $\frac{d\mu^\star(q)}{dq} \cdot \frac{\partial c'}{\partial q}\Big|_{\mu=\mu^\star(q)} \le 0$ whenever $\mu^\star(q) \ne \lambda/N$. That is, throughout we work at values of q lying in the interior of an overloaded or underloaded regime, where $\mu^\star(q)$ is the corresponding candidate equilibrium service rate and is differentiable in q; at regime boundaries, $\mu^\star(q)$ may be non-differentiable, and on critical plateaus $\mu^\star(q) \equiv \lambda/N$. We treat each regime separately.

Overloaded regime ($\mu^\star = \mu_0 > 0$). The symmetric FOC in the overloaded regime (Equation (7)) is

$$p = c'(\mu) = qc_E\mu^{q-1}.$$

The left-hand side does not depend on q. Totally differentiating both sides with respect to q while treating μ_0 as a function of q:

$$0 = \frac{d}{dq}\left[c'(\mu_0(q))\right] = \frac{\partial c'}{\partial q}\bigg|_{\mu=\mu_0} + c''(\mu_0)\cdot\mu_0'(q).$$

The second derivative of the cost function satisfies $c''(\mu) = q(q-1)c_E\mu^{q-2} > 0$ for all $\mu > 0$ and $q > 1$. Solving for $\mu_0'(q)$:

$$\mu_0'(q) = -\frac{\frac{\partial c'}{\partial q}\Big|_{\mu=\mu_0}}{c''(\mu_0)}.$$

Since $c''(\mu_0) > 0$, multiplying both sides by $\frac{\partial c'}{\partial q}\Big|_{\mu=\mu_0}$:

$$\mu_0'(q)\cdot\frac{\partial c'}{\partial q}\bigg|_{\mu=\mu_0} = -\frac{\left(\frac{\partial c'}{\partial q}\Big|_{\mu=\mu_0}\right)^2}{c''(\mu_0)} \le 0,$$

with equality if and only if $\frac{\partial c'}{\partial q}\Big|_{\mu=\mu_0} = 0$, i.e., $\mu_0 = e^{-1/q}$. This establishes the result in the overloaded regime.

Underloaded regime ($\mu^\star = \mu_{\mathrm{U}} > \lambda/N$, $N > 1$). The symmetric FOC in the underloaded regime (Equation (7)) is

$$p\frac{\lambda(N-1)}{N^2\mu} = c'(\mu) = qc_E\mu^{q-1}.$$

Define $G(\mu) := p\lambda(N-1)/(N^2\mu)$, so the FOC reads $G(\mu) = c'(\mu)$. Neither G nor the left-hand side depends on q. Totally differentiating with respect to q:

$$G'(\mu_{\mathrm{U}}) \cdot \mu_{\mathrm{U}}'(q) = \left.\frac{\partial c'}{\partial q}\right|_{\mu=\mu_{\mathrm{U}}} + c''(\mu_{\mathrm{U}}) \cdot \mu_{\mathrm{U}}'(q),$$

which gives

$$\mu_{\mathrm{U}}'(q) = \frac{\left.\frac{\partial c'}{\partial q}\right|_{\mu=\mu_{\mathrm{U}}}}{G'(\mu_{\mathrm{U}}) - c''(\mu_{\mathrm{U}})}.$$

Now $G'(\mu) = -p\lambda(N-1)/(N^2\mu^2) < 0$ and $c''(\mu) > 0$, so the denominator $G'(\mu_{\mathrm{U}}) - c''(\mu_{\mathrm{U}}) < 0$. Therefore,

$$\mu_{\mathrm{U}}'(q) \cdot \left.\frac{\partial c'}{\partial q}\right|_{\mu=\mu_{\mathrm{U}}} = \frac{\left(\left.\frac{\partial c'}{\partial q}\right|_{\mu=\mu_{\mathrm{U}}}\right)^2}{G'(\mu_{\mathrm{U}}) - c''(\mu_{\mathrm{U}})} \le 0,$$

with equality if and only if $\left.\frac{\partial c'}{\partial q}\right|_{\mu=\mu_{\mathrm{U}}} = 0$, i.e., $\mu_{\mathrm{U}} = e^{-1/q}$. This establishes the result in the underloaded regime.

□

Proof of Corollary 1. From (10), we have $\frac{\partial c'(\mu)}{\partial q} = c_E \mu^{q-1}(1 + q\log\mu)$. Since $c_E\mu^{q-1} > 0$ for $\mu > 0$, the sign of $\frac{\partial c'(\mu)}{\partial q}$ equals the sign of $1 + q\log\mu$, which is positive when $\mu > e^{-1/q}$ and negative when $\mu < e^{-1/q}$.

When $\mu^\star(q) \neq \lambda/N$, Proposition 4 gives $\frac{d\mu^\star}{dq} \cdot \left.\frac{\partial c'}{\partial q}\right|_{\mu=\mu^\star} \le 0$. As shown in the proof of Proposition 4, the inequality is strict whenever $\left.\frac{\partial c'}{\partial q}\right|_{\mu=\mu^\star} \neq 0$, i.e., whenever $\mu^\star \neq e^{-1/q}$.

- If $\mu^\star(q) < e^{-1/q}$: then $\left.\frac{\partial c'}{\partial q}\right|_{\mu=\mu^\star} < 0$, and the strict inequality forces $\frac{d\mu^\star}{dq} > 0$, so $\mu^\star(q)$ is strictly increasing in q.
- If $\mu^\star(q) > e^{-1/q}$: then $\left.\frac{\partial c'}{\partial q}\right|_{\mu=\mu^\star} > 0$, and the strict inequality $\frac{d\mu^\star}{dq} \cdot \left.\frac{\partial c'}{\partial q}\right|_{\mu=\mu^\star} < 0$ forces $\frac{d\mu^\star}{dq} < 0$, so $\mu^\star(q)$ is strictly decreasing in q.

□

Proof of Lemma 3. We have $h(q) = q(\lambda/N)^{q-1}$ and $h(1) = 1$. Differentiating:

$$h'(q) = \left(\frac{\lambda}{N}\right)^{q-1}\left(1 + q\log\frac{\lambda}{N}\right).$$

Since $(\lambda/N)^{q-1} > 0$, the sign of $h'(q)$ equals the sign of $1 + q\log(\lambda/N)$.

- If $\lambda/N \le 1/e$: then $\log(\lambda/N) \le -1$, so $1 + q\log(\lambda/N) \le 1 - q < 0$ for $q > 1$. Thus h is strictly decreasing. Moreover, $h(q) = q(\lambda/N)^{q-1} \le qe^{-(q-1)} \to 0$.
- If $\lambda/N \ge 1$: then $\log(\lambda/N) \ge 0$, so $1 + q\log(\lambda/N) \ge 1 > 0$ for all $q > 1$. Thus h is strictly increasing. Since $(\lambda/N)^{q-1} \ge 1$ and grows, $h(q) \to \infty$.
- If $\lambda/N \in (1/e, 1)$: then $\log(\lambda/N) \in (-1, 0)$. The expression $1 + q\log(\lambda/N)$ vanishes at $q^* = -1/\log(\lambda/N) > 1$, is positive for $q < q^*$ and negative for $q > q^*$. Thus h increases on $(1, q^*)$ and decreases on (q^*, ∞), with a unique maximum at q^*. The maximum value is

$$h(q^*) = q^*\left(\frac{\lambda}{N}\right)^{q^*-1} = \frac{-1}{\log(\lambda/N)} \cdot \left(\frac{\lambda}{N}\right)^{-1/\log(\lambda/N)-1} = \frac{-1}{\log(\lambda/N)} \cdot \frac{e^{-1}}{\lambda/N} = \frac{-N}{e\lambda\log(\lambda/N)},$$

where we used $(\lambda/N)^{-1/\log(\lambda/N)} = e^{-1}$. Since $h(q) \to 0$ as $q \to \infty$ (because $(\lambda/N)^{q-1} \to 0$ dominates), the claim follows.

□

Proof of Proposition 5. Throughout this proof we assume $N \geq 2$. The case $N = 1$ is immediate: the underloaded regime is then void (the upper threshold $\frac{N}{N-1}c'(\lambda/N)$ in Proposition 1 is read as $+\infty$), so by Proposition 1 the equilibrium is overloaded for $p < c'(\lambda)$ and critical for $p \geq c'(\lambda)$. In the only nontrivial regime, $\mu^\star(q) = \mu_0(q)$ is monotone in q by Lemma A.3(i), establishing the proposition with the convention $q_1 = q_2 = 1$, $q_3 = q_4 = \infty$ if $p < c'(\lambda)$, and $q_1 = q_2 = q_3 = q_4 = \infty$ otherwise.

With a slight rearrangement of the conditions in Proposition 1, we can deduce that the type of the equilibrium service rate $\mu^\star$ (namely, underloaded $\mu^\star > \frac{\lambda}{N}$, critically loaded $\mu^\star = \frac{\lambda}{N}$, or overloaded $\mu^\star < \frac{\lambda}{N}$) depends on the value of $q\left(\frac{\lambda}{N}\right)^{q-1}$; specifically,

- If $q\left(\frac{\lambda}{N}\right)^{q-1} < \frac{p}{c_E}\frac{N-1}{N}$, then the equilibrium is underloaded and we have $\mu^\star(q) = \mu_{\mathrm{U}}(q) > \frac{\lambda}{N}$;
- If $\frac{p}{c_E}\frac{N-1}{N} \leq q\left(\frac{\lambda}{N}\right)^{q-1} \leq \frac{p}{c_E}$, then the equilibrium is critically loaded ($\mu^\star = \frac{\lambda}{N}$);
- If $q\left(\frac{\lambda}{N}\right)^{q-1} > \frac{p}{c_E}$, then the equilibrium is overloaded and we have $\mu^\star(q) = \mu_0(q) < \frac{\lambda}{N}$.

Note that the monotonicity of $q\left(\frac{\lambda}{N}\right)^{q-1}$ as a function of q depends on the value of $\frac{\lambda}{N}$. To see this, we differentiate $q\left(\frac{\lambda}{N}\right)^{q-1}$ with respect to q and obtain

$$\frac{\mathrm{d}}{\mathrm{d}q}\left(q\left(\frac{\lambda}{N}\right)^{q-1}\right) = \left(\frac{\lambda}{N}\right)^{q-1} + q\log(\frac{\lambda}{N})\left(\frac{\lambda}{N}\right)^{q-1} = \left(\frac{\lambda}{N}\right)^{q-1}\left(1 + q\log(\frac{\lambda}{N})\right).$$

Because $\left(\frac{\lambda}{N}\right)^{q-1} > 0$, the above implies that $q\left(\frac{\lambda}{N}\right)^{q-1}$ is increasing in q when $1 + q\log\left(\frac{\lambda}{N}\right) \geq 0$, and decreasing in q when $1 + q\log\left(\frac{\lambda}{N}\right) \leq 0$. Therefore, the monotonicity of $q\left(\frac{\lambda}{N}\right)^{q-1}$ can be summarized as follows:

- If $\frac{\lambda}{N} \in (0, \frac{1}{e}]$, then $\log(\frac{\lambda}{N}) \leq -1$ and thus $1 + q\log(\frac{\lambda}{N}) < 0$ always holds because $q > 1$. Hence, $q\left(\frac{\lambda}{N}\right)^{q-1}$ is monotonically decreasing in $q \in (1, \infty)$;
- If $\frac{\lambda}{N} \in (\frac{1}{e}, 1)$, then $\log(\frac{\lambda}{N}) \in (-1, 0)$. Therefore, $q\left(\frac{\lambda}{N}\right)^{q-1}$ is increasing in $q \in (1, -\left(\log\left(\frac{\lambda}{N}\right)\right)^{-1})$ and decreasing in $q \in (-\left(\log\left(\frac{\lambda}{N}\right)\right)^{-1}, \infty)$.
- If $\frac{\lambda}{N} \geq 1$, then $\log(\frac{\lambda}{N}) \geq 0$ and thus $1 + q\log(\frac{\lambda}{N}) > 0$ always holds. Hence, $q\left(\frac{\lambda}{N}\right)^{q-1}$ is monotonically increasing in $q \in (1, \infty)$;

Now we analyze the monotonicity property of the equilibrium service rate in each of the above three cases.

Case 1: If $\frac{\lambda}{N} \in (0, \frac{1}{e})$, then $q\left(\frac{\lambda}{N}\right)^{q-1}$ is monotonically decreasing in $q \in (1, \infty)$. Note that $q\left(\frac{\lambda}{N}\right)^{q-1}$ evaluates to 1 at $q = 1$, and 0 as $q \to \infty$. Figure EC.1 helps to visualize the sub-cases for this case.

- **Case 1-i:** If $\frac{p}{c_E} \leq 1$, then $\mu^\star(q)$ is overloaded when $q \in (1, q_3)$, critically loaded when $q \in (q_3, q_4)$, and underloaded when $q \in (q_4, \infty)$, where q_3 is the unique solution to $q\left(\frac{\lambda}{N}\right)^{q-1} = \frac{p}{c_E}$, and q_4 is the unique solution to $q\left(\frac{\lambda}{N}\right)^{q-1} = \frac{p}{c_E}\frac{N-1}{N}$. By Lemma A.3 and Lemma A.4, it can be easily seen that $\mu^\star(q)$ is strictly increasing in $q \in (1, q_3)$, stays constant over $q \in (q_3, q_4)$, and finally strictly increasing in $q \in (q_4, \infty)$. In this case, $q_1 = q_2 = 1$ and $q^\dagger = 1$.

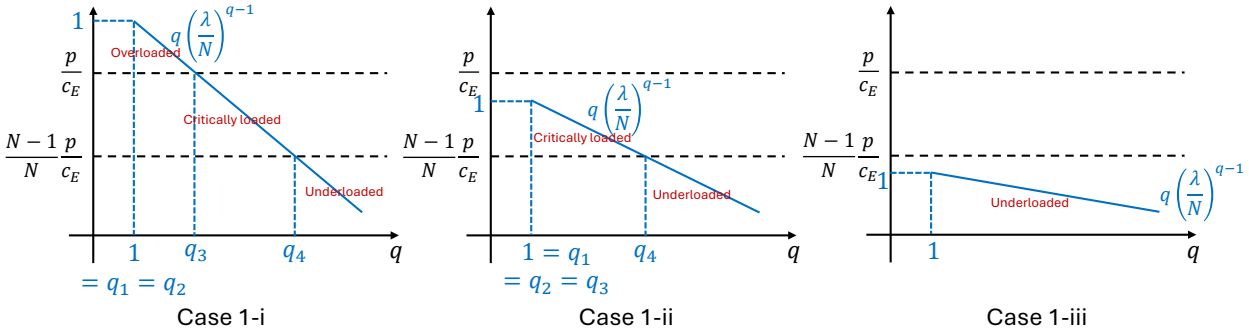


Figure EC.1 Illustration of Case 1-i to Case 1-iii

- **Case 1-ii:** If $1 < \frac{p}{c_E} \le \frac{N}{N-1}$, then similar to the latter two stages of Case 1-i, $\mu^\star(q)$ stays constant over $q \in (1, q_4)$, and then is strictly increasing in $q \in (q_4, \infty)$, where q_4 is the unique solution to $q\left(\frac{\lambda}{N}\right)^{q-1} = \frac{p}{c_E}\frac{N-1}{N}$. In this case, $q_1 = q_2 = q_3 = 1$.
- **Case 1-iii:** If $\frac{p}{c_E} > \frac{N}{N-1}$, then $\mu^\star(q)$ is first strictly decreasing in $q \in (1, \max\{1, \frac{ep\lambda}{Nc_E}(1-\frac{1}{N})\})$ and then strictly increasing in $q \in (\max\{1, \frac{ep\lambda}{Nc_E}(1-\frac{1}{N})\}, \infty)$. In this case, $q_1 = q_2 = q_3 = q_4 = \max\{1, \frac{ep\lambda}{Nc_E}(1-\frac{1}{N})\}$.
  - **Case 1-iii-a:** If $\frac{ep\lambda}{Nc_E}(1-\frac{1}{N}) \le 1$, that is, $\frac{\lambda}{N}\cdot\frac{p}{c_E} \le \frac{N}{e(N-1)}$, then $\max\{1, \frac{ep\lambda}{Nc_E}(1-\frac{1}{N})\} = 1$. Therefore, $\mu^\star(q)$ is strictly increasing in $q \in (1, \infty)$. Additionally, $q_1 = q_2 = q_3 = q_4 = 1$.
  - **Case 1-iii-b:** If $\frac{ep\lambda}{Nc_E}(1-\frac{1}{N}) > 1$, that is, $\frac{\lambda}{N}\cdot\frac{p}{c_E} > \frac{N}{e(N-1)}$, then $\max\{1, \frac{ep\lambda}{Nc_E}(1-\frac{1}{N})\} = \frac{ep\lambda}{Nc_E}(1-\frac{1}{N})$. Therefore, $\mu^\star(q)$ is first strictly decreasing in $q \in (1, \frac{ep\lambda}{Nc_E}(1-\frac{1}{N}))$ and then strictly increasing in $q \in (\frac{ep\lambda}{Nc_E}(1-\frac{1}{N}), \infty)$. Additionally, $q_1 = q_2 = q_3 = q_4 = \frac{ep\lambda}{Nc_E}(1-\frac{1}{N})$.

Case 2: If $\frac{\lambda}{N} \in (\frac{1}{e}, 1)$, then $q\left(\frac{\lambda}{N}\right)^{q-1}$ is monotonically increasing in $q \in (1, -\left(\log\left(\frac{\lambda}{N}\right)\right)^{-1})$ and monotonically decreasing in $q \in (-\left(\log\left(\frac{\lambda}{N}\right)\right)^{-1}, \infty)$. The maximum value of $q\left(\frac{\lambda}{N}\right)^{q-1}$ is $M := -\frac{N}{e\lambda\log\left(\frac{\lambda}{N}\right)}$. Note that $q\left(\frac{\lambda}{N}\right)^{q-1}$ evaluates to 1 at $q = 1$, and 0 as $q \to \infty$. Figure EC.2 helps to visualize the sub-cases of this case.

- **Case 2-i:** If $\frac{p}{c_E} \le 1$, then we fall back to Case 1-i.
- **Case 2-ii-a:** If $1 < \frac{p}{c_E} \le \frac{N}{N-1}$ and $\frac{p}{c_E} < M$, then $\mu^\star(q)$ is critically loaded when $q \in (1, q_2)$, overloaded when $q \in (q_2, q_3)$, critically loaded when $q \in (q_3, q_4)$, and underloaded when $q \in (q_4, \infty)$, where q_2 and $q_3 > q_2$ are the two solutions to $q\left(\frac{\lambda}{N}\right)^{q-1} = \frac{p}{c_E}$, and $q_4 > q_3$ is the unique solution to $q\left(\frac{\lambda}{N}\right)^{q-1} = \frac{N-1}{N}\frac{p}{c_E}$. By Lemma A.3 and Lemma A.4, it can be easily seen that $\mu^\star(q)$ stays constant over $q \in (1, q_2)$, is decreasing in $q \in (q_2, q^\dagger)$, then strictly increasing in $q \in (q^\dagger, q_3)$, then again stays constant over $q \in (q_3, q_4)$, and finally keeps increasing in $q \in (q_4, \infty)$. In this case, $q_1 = 1$.
- **Case 2-ii-b:** If $1 < \frac{p}{c_E} \le \frac{N}{N-1}$ and $\frac{p}{c_E} \ge M$, then we fall back to Case 1-ii.
- **Case 2-iii-a:** If $\frac{p}{c_E} > \frac{N}{N-1}$ and $\frac{p}{c_E} < M$, then $\mu^\star(q)$ is underloaded when $q \in (1, q_1)$, critically loaded when $q \in (q_1, q_2)$, overloaded when $q \in (q_2, q_3)$, critically loaded when $q \in (q_3, q_4)$, and underloaded when $q \in (q_4, \infty)$, where q_2 and $q_3 > q_2$ are the two solutions to $q\left(\frac{\lambda}{N}\right)^{q-1} = \frac{p}{c_E}$, and q_1 and $q_4 > q_1$ are

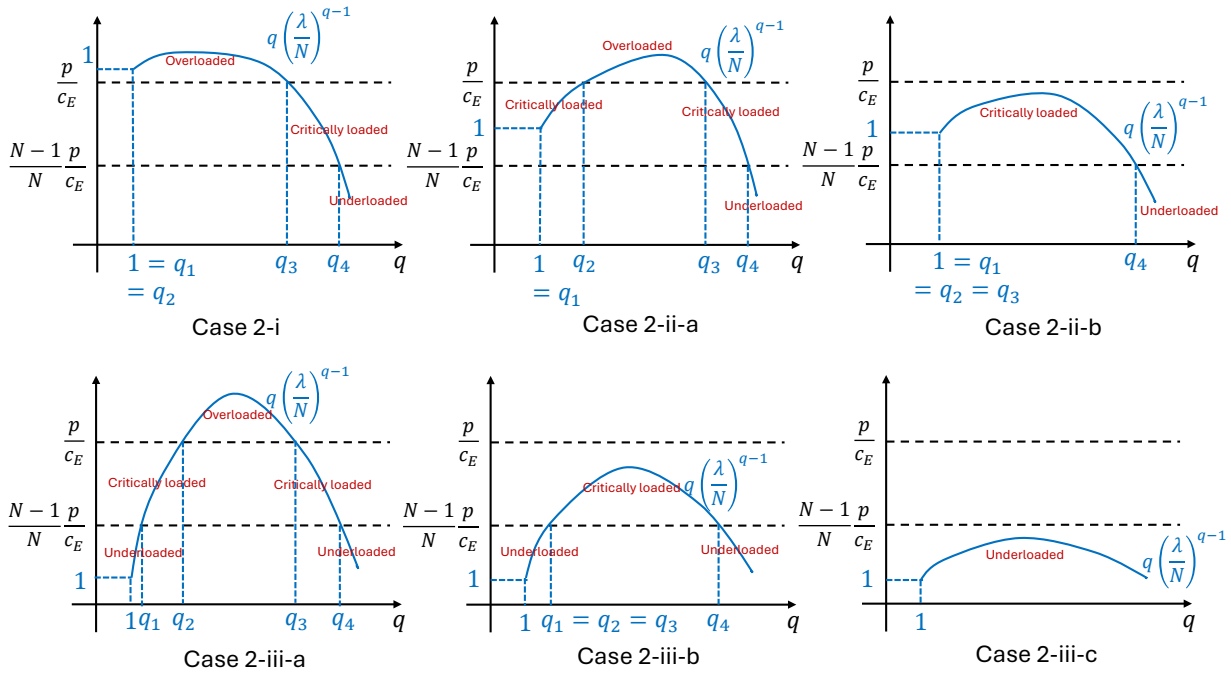


Figure EC.2 Illustration of Case 2-i to Case 2-iii-c

the two solutions to $q\left(\frac{\lambda}{N}\right)^{q-1} = \frac{N-1}{N}\frac{p}{c_E}$. For $q \in (1, q_1)$, $\mu^\star(q)$ is strictly decreasing in q, because if there were an increasing interval, then, by Lemma A.3, $\mu^\star(q) = \mu_{\mathrm{U}}$ would keep increasing and thus, by the continuity of $\mu^\star(q)$ in q established in Lemma A.2, would never come back to critically loaded. Then, $\mu^\star(q)$ stays constant over $q \in (q_1, q_2)$, is decreasing in $q \in (q_2, q^\dagger)$, then strictly increasing in $q \in (q^\dagger, q_3)$ (by Lemma A.3), then again stays constant over $q \in (q_3, q_4)$, and finally keeps increasing in $q \in (q_4, \infty)$ (by Lemma A.4).

- **Case 2-iii-b:** If $\frac{p}{c_E} > \frac{N}{N-1}$ and $M < \frac{p}{c_E} < M\frac{N}{N-1}$, then $\mu^\star(q)$ is underloaded when $q \in (1, q_1)$, critically loaded over $q \in (q_1, q_4)$, and underloaded when $q \in (q_4, \infty)$, where q_1 and $q_4 > q_1$ are the two solutions to $q\left(\frac{\lambda}{N}\right)^{q-1} = \frac{N-1}{N}\frac{p}{c_E}$. By Lemma A.3 and Lemma A.4, it can be easily seen that $\mu^\star(q)$ is strictly decreasing in $q \in (1, q_1)$, constant over $q \in (q_1, q_4)$, and strictly increasing in $q \in (q_4, \infty)$. In this case, $q_1 = q_2 = q_3$.
- **Case 2-iii-c:** If $\frac{p}{c_E} > \frac{N}{N-1}$ and $\frac{p}{c_E} > M\frac{N}{N-1}$, then we fall back to Case 1-iii.

Case 3: If $\frac{\lambda}{N} \geq 1$, then $q\left(\frac{\lambda}{N}\right)^{q-1}$ is monotonically increasing in $q \in (1, \infty)$. Note that $q\left(\frac{\lambda}{N}\right)^{q-1}$ evaluates to 1 at $q = 1$, and ∞ as $q \to \infty$. Figure EC.3 helps to visualize the sub-cases for this case.

- **Case 3-i:** If $\frac{p}{c_E} \leq 1$, then $\mu^\star(q)$ is overloaded for all $q > 1$. By Lemma A.3(i-a), $\mu^\star(q) = \mu_0(q)$ is strictly increasing in q and converges to an asymptote 1. In this case, $q_1 = q_2 = 1$, $q_3 = q_4 = \infty$.
- **Case 3-ii:** If $1 < \frac{p}{c_E} \leq \frac{N}{N-1}$, then similar to the latter two stages of Case 3-i, $\mu^\star(q)$ stays constant over $q \in (1, q_2)$, and then is strictly decreasing in $q \in (q_2, q^\dagger)$ followed by strictly increasing in $q \in (q^\dagger, \infty)$ to an asymptote 1, where q_2 is the unique solution to $q\left(\frac{\lambda}{N}\right)^{q-1} = \frac{p}{c_E}$. In this case, $q_1 = 1$, $q_3 = q_4 = \infty$.

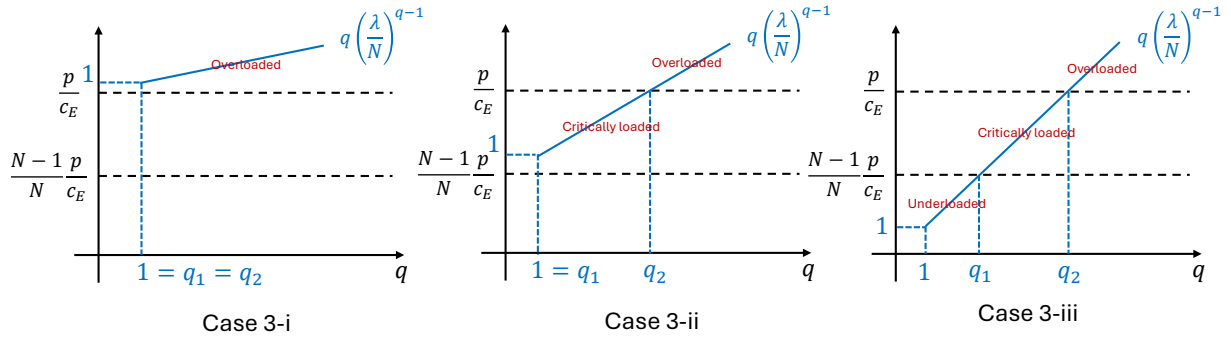


Figure EC.3 Illustration of Case 3-i to Case 3-iii

- **Case 3-iii:** If $\frac{p}{c_E} > \frac{N}{N-1}$, then $\mu^\star(q)$ is underloaded when $q \in (1, q_1)$, critically loaded when $q \in (q_1, q_2)$, and overloaded when $q \in (q_2, \infty)$, where q_1 is the unique solution to $q\left(\frac{\lambda}{N}\right)^{q-1} = \frac{p}{c_E}\frac{N-1}{N}$, and $q_2 > q_1$ is the unique solution to $q\left(\frac{\lambda}{N}\right)^{q-1} = \frac{p}{c_E}$. For $q \in (1, q_1)$, $\mu^\star(q)$ is strictly decreasing in q, because if there were an increasing interval, then, by Lemma A.3, $\mu^\star(q) = \mu_{\mathrm{U}}$ would keep increasing and thus, by the continuity of $\mu^\star(q)$ in q established in Lemma A.2, would never come back to critically loaded. Then, $\mu^\star(q)$ stays constant over $q \in (q_1, q_2)$, and, again by Lemma A.3, $\mu^\star(q)$ is strictly decreasing in $q \in (q_2, q^\dagger)$ and strictly increasing in $q \in (q^\dagger, \infty)$ to an asymptote 1. In this case, $q_3 = q_4 = \infty$.

□

Proof of Proposition 6. The result follows by inspecting the sub-cases of the proof of Proposition 5 and determining, for each parameter configuration, whether $\mu^\star(q)$ is increasing or exhibits a decreasing phase.

• **Part (i):** $p \le c_E$**.** We show that $\mu^\star(q)$ is increasing for all $\lambda > 0$ by considering each case of the proof of Proposition 5:

- Case 1-i ($\frac{\lambda}{N} < \frac{1}{e}$, $\frac{p}{c_E} \le 1$): The equilibrium traverses overloaded, critically loaded, and underloaded phases. In the overloaded phase, $\mu_0(q)$ is strictly increasing by Lemma A.3(i-a) (since $p \le c_E$). The critically loaded phase is constant at λ/N. The underloaded phase is strictly increasing (by the proof of Proposition 5, Case 1-i). Since μ_0 reaches λ/N at the transition q_3 and μ_{U} departs from λ/N at $q_4 \ge q_3$, the overall trajectory is increasing.
- Case 2-i ($\frac{\lambda}{N} \in (\frac{1}{e}, 1)$, $\frac{p}{c_E} \le 1$): Falls back to Case 1-i.
- Case 1-ii / Case 2-ii-b ($1 < \frac{p}{c_E} \le \frac{N}{N-1}$, h does not exceed $\frac{p}{c_E}$): The equilibrium is constant at λ/N then increasing.
- Case 3-i ($\frac{\lambda}{N} \ge 1$, $\frac{p}{c_E} \le 1$): The equilibrium is overloaded for all $q > 1$, and $\mu_0(q)$ is strictly increasing by Lemma A.3(i-a) (since $p \le c_E$ implies $q^\dagger = 1$, so the decreasing interval $(q_2, q^\dagger)$ is empty).

In every case, $\mu^\star(q)$ is increasing.

• **Part (ii):** $p > c_E$**.** We identify the threshold $\Lambda(p)$ by considering two sub-ranges of p.

Sub-case (A): $p > \frac{N}{N-1}c_E$. In this range, $\frac{p}{c_E} > \frac{N}{N-1}$. At $q = 1$, we have $h(1) = 1 < \frac{p}{c_E}\frac{N-1}{N}$ (since $\frac{p}{c_E}\frac{N-1}{N} > 1$), so the equilibrium starts underloaded with $\mu^\star(q) = \mu_{\mathtt{U}}(q)$. From the proof of Proposition 5, the relevant cases are:

- Cases 1-iii, 2-iii-b, 2-iii-c, 3-iii: In all of these, the equilibrium begins underloaded and $\mu^\star(q)$ is first decreasing then increasing if and only if $\mu_{\mathtt{U}}(q)$ is U-shaped in q near $q = 1$. By Lemma A.3(ii-c), $\mu_{\mathtt{U}}(q)$ is U-shaped (first decreasing then increasing) if and only if $\frac{ep\lambda(N-1)}{N^2c_E} > 1$, i.e., $\lambda > \frac{N^2c_E}{e(N-1)p}$. When $\frac{ep\lambda(N-1)}{N^2c_E} \le 1$, Lemma A.3(ii-b) gives $\mu_{\mathtt{U}}(q)$ strictly increasing, so $\mu^\star(q)$ is non-decreasing.
- Case 2-iii-a (full five-phase): The initial underloaded phase has $\mu_{\mathtt{U}}$ decreasing, which again requires $\frac{ep\lambda(N-1)}{N^2c_E} > 1$, the same condition.

Thus, $\Lambda(p) = \frac{N^2}{e(N-1)}\frac{c_E}{p}$ for $p > \frac{N}{N-1}c_E$.

Sub-case (B): $c_E < p \le \frac{N}{N-1}c_E$. In this range, $1 < \frac{p}{c_E} \le \frac{N}{N-1}$. At $q = 1$, $h(1) = 1 \in \left[\frac{p}{c_E}\frac{N-1}{N}, \frac{p}{c_E}\right]$, so the equilibrium starts critically loaded. Non-monotonicity can only arise if the equilibrium later enters the overloaded regime, where $\mu_0(q)$ is U-shaped (by Lemma A.3(i-b), since $p > c_E$). This occurs when the auxiliary function $h(q) = q(\lambda/N)^{q-1}$ exceeds $\frac{p}{c_E}$ at some $q > 1$, which requires $\max_{q>1} h(q) > \frac{p}{c_E}$.

The maximum of h depends on $\frac{\lambda}{N}$:

- If $\frac{\lambda}{N} \le \frac{1}{e}$ (Case 1): h is decreasing from $h(1) = 1$, so $\max h = 1 \le \frac{p}{c_E}$. The overloaded phase never materializes. Case 1-ii applies: $\mu^\star(q)$ is non-decreasing.
- If $\frac{\lambda}{N} \ge 1$ (Case 3): h is increasing to ∞, so h eventually exceeds $\frac{p}{c_E}$. Case 3-ii applies: the overloaded phase exists and $\mu_0(q)$ is U-shaped (first decreasing then increasing, since $p > c_E$ gives $q^\dagger > 1$, and one can verify that $q_2 < q^\dagger$ by evaluating $\tau(q_2) = \tau(p/c_E) = -(p/c_E - 1)/(p/c_E) < 0$, so μ_0 is still in its decreasing phase at the entry point). Thus $\mu^\star(q)$ is non-monotone. Since this holds for all $\lambda \ge N$, any valid threshold must satisfy $\Lambda(p) < N$.
- If $\frac{\lambda}{N} \in (\frac{1}{e}, 1)$ (Case 2): h is hump-shaped with maximum $M = -\frac{N}{e\lambda\log(\lambda/N)}$. Non-monotonicity occurs if and only if $M > \frac{p}{c_E}$.

It remains to find the critical λ at which $M = \frac{p}{c_E}$. Writing the condition $M > \frac{p}{c_E}$ explicitly:

$$-\frac{N}{e\lambda\log(\lambda/N)} > \frac{p}{c_E} \quad\Longleftrightarrow\quad \frac{N}{e}\frac{c_E}{p} + \lambda\log\left(\frac{\lambda}{N}\right) > 0.$$

Define $\varphi(\lambda) := \frac{N}{e}\frac{c_E}{p} + \lambda\log\left(\frac{\lambda}{N}\right)$. Then $\varphi'(\lambda) = \log(\lambda/N) + 1$, which vanishes at $\lambda = N/e$ and is positive for $\lambda > N/e$. Thus φ attains its minimum at $\lambda = N/e$ with value $\varphi(N/e) = \frac{N}{e}\left(\frac{c_E}{p} - 1\right) < 0$ (since $p > c_E$). Moreover, φ is strictly increasing on $(N/e, \infty)$ with $\varphi(N) = \frac{N}{e}\frac{c_E}{p} > 0$. By the intermediate value theorem, there exists a unique $\lambda^\dagger \in (N/e, N)$ such that $\varphi(\lambda^\dagger) = 0$.

For $\lambda > \lambda^\dagger$ with $\lambda/N \in (1/e, 1)$: $\varphi(\lambda) > 0$, so $M > p/c_E$, and Case 2-ii-a applies (non-monotone). Combined with Case 3 ($\lambda \ge N$, also non-monotone), $\mu^\star(q)$ is non-monotone for all $\lambda > \lambda^\dagger$.

For $\lambda \le \lambda^\dagger$: either $\lambda/N \le 1/e$ (Case 1-ii, non-decreasing) or $\lambda/N \in (1/e, 1)$ with $\varphi(\lambda) \le 0$ (so $M \le p/c_E$, Case 2-ii-b, which reduces to Case 1-ii, non-decreasing). In both cases, $\mu^\star(q)$ is non-decreasing.

Thus $\Lambda(p) = \lambda^\dagger$ for $c_E < p \le \frac{N}{N-1}c_E$. □

E. Proofs from Section 6

Proof of Proposition 7. We optimize the designer's objective $\Pi(p) := (r-p)\,\Theta(p)$ over $p > 0$, where $r > 0$, $q > 1$, $c_E > 0$, and $\lambda > 0$. Since $\Pi(r) = 0$ and $\Pi(p) < 0$ for $p > r$, any maximizer must lie in $(0, r]$. The throughput function $\Theta(p)$ is given by equation (13). We analyze the two regimes separately and then compare.

Overloaded regime ($0 < p < c'(\lambda/N)$). Define

$$\Pi_O(p) := (r-p) \cdot N\left(\frac{p}{qc_E}\right)^{1/(q-1)}. \tag{E.1}$$

Let $\alpha := 1/(q-1) > 0$. Then $\Pi_O(p) = N(qc_E)^{-\alpha}(r-p)\,p^{\alpha}$. Differentiating with respect to p:

$$\begin{aligned}\Pi_O'(p) &= N(qc_E)^{-\alpha}\left[-p^{\alpha} + \alpha(r-p)\,p^{\alpha-1}\right] \\ &= N(qc_E)^{-\alpha}\,p^{\alpha-1}\left[\alpha r - (1+\alpha)p\right].\end{aligned} \tag{E.2}$$

Substituting $\alpha = 1/(q-1)$ gives $1+\alpha = q/(q-1)$, so the bracketed factor becomes

$$\frac{r}{q-1} - \frac{q}{q-1}\,p \;=\; \frac{q}{q-1}\left(\frac{r}{q} - p\right).$$

Since $N(qc_E)^{-\alpha} > 0$, $p^{\alpha-1} > 0$ for $p > 0$, and $q/(q-1) > 0$, we have

$$\Pi_O'(p) \gtreqless 0 \iff p \gtreqless r/q. \tag{E.3}$$

Thus Π_O is strictly increasing on $(0, r/q)$ and strictly decreasing on $(r/q, \infty)$, with a unique global maximum at $p = r/q$. To confirm this via the second-order condition, note that at $p = r/q$ the bracketed factor in equation (E.2) vanishes, so

$$\Pi_O''(r/q) = N(qc_E)^{-\alpha}(r/q)^{\alpha-1} \cdot \left[-(1+\alpha)\right] < 0,$$

confirming a strict maximum.

Critically loaded and underloaded regime ($p \geq c'(\lambda/N)$). Define $\Pi_C(p) := (r-p)\lambda$. This is linear in p with slope $-\lambda < 0$, so Π_C is strictly decreasing and uniquely maximized at the left endpoint $p = c'(\lambda/N)$, with value

$$\Pi_C\left(c'\left(\frac{\lambda}{N}\right)\right) = \left(r - c'\left(\frac{\lambda}{N}\right)\right)\lambda. \tag{E.4}$$

Continuity at the boundary. We verify that $\lim_{p \uparrow c'(\lambda/N)} \Pi_O(p) = \Pi_C(c'(\lambda/N))$. Since $c'(\lambda/N) = qc_E(\lambda/N)^{q-1}$, we have

$$\left(\frac{c'(\lambda/N)}{qc_E}\right)^{1/(q-1)} = \left((\lambda/N)^{q-1}\right)^{1/(q-1)} = \frac{\lambda}{N}.$$

Therefore $\lim_{p \uparrow c'(\lambda/N)} \Pi_O(p) = (r - c'(\lambda/N)) \cdot N \cdot (\lambda/N) = (r - c'(\lambda/N))\lambda = \Pi_C(c'(\lambda/N))$, as required.

Global comparison. We now compare the two candidate maxima.

Case (i): $r/q < c'(\lambda/N)$. The unconstrained maximizer r/q lies strictly inside the overloaded domain $(0, c'(\lambda/N))$. By unimodality (equation (E.3)), Π_O is strictly increasing on $(0, r/q)$ and strictly decreasing on $(r/q, c'(\lambda/N))$, so Π_O attains its unique maximum on $(0, c'(\lambda/N))$ at $p = r/q$, with value

$$\Pi_O(r/q) = \frac{r(q-1)}{q} \cdot N\left(\frac{r}{q^2 c_E}\right)^{1/(q-1)}. \tag{E.5}$$

Moreover, the strict decrease on $(r/q, c'(\lambda/N))$ yields $\Pi_O(r/q) > \lim_{p\uparrow c'(\lambda/N)} \Pi_O(p) = \Pi_C(c'(\lambda/N))$. Since Π_C is maximized at $c'(\lambda/N)$ and $\Pi_O(r/q) > \Pi_C(c'(\lambda/N))$, the global maximum is $\Pi_O(r/q)$, attained at $p_D^\star = r/q$.

At $p_D^\star = r/q < c'(\lambda/N)$, Proposition 1 gives $\mu^\star = \mu_0 = (r/(q^2 c_E))^{1/(q-1)} < \lambda/N$ (overloaded), and $\Pi^\star = \Pi_O(r/q)$ as given in equation (E.5), confirming part (i).

Case (ii): $r/q \geq c'(\lambda/N)$. The unconstrained maximizer r/q of Π_O lies at or above the upper boundary $c'(\lambda/N)$ of the overloaded domain. By equation (E.3), Π_O is strictly increasing on $(0, c'(\lambda/N)) \subseteq (0, r/q)$. Hence $\Pi_O(p) < \lim_{p\uparrow c'(\lambda/N)} \Pi_O(p) = \Pi_C(c'(\lambda/N))$ for all $p \in (0, c'(\lambda/N))$. The supremum of Π_O over the open interval $(0, c'(\lambda/N))$ is not attained in that interval, but equals $\Pi_C(c'(\lambda/N))$ by continuity. Since Π_C is maximized at $c'(\lambda/N)$ over $[c'(\lambda/N), \infty)$, the global maximum of Π is $\Pi_C(c'(\lambda/N))$, attained at $p_D^\star = c'(\lambda/N)$.

At $p_D^\star = c'(\lambda/N)$, Proposition 1 gives $\mu^\star = \lambda/N$ (critically loaded), and $\Pi^\star = \Pi_C(c'(\lambda/N)) = (r - c'(\lambda/N))\lambda$, confirming part (ii). □

Proof of Proposition 8. We maximize the surplus $W(p) = (r-p)\,\Theta(p) + N \cdot U(\mu^\star(p), \mu^\star(p))$ over $p > 0$, where $r > 0$, $q > 1$, $c_E > 0$, and $\lambda > 0$.

Step 1: Reduction of the surplus function. We first verify that $W(p) = r\,\Theta(p) - N\,c_E \mu^\star(p)^q$ in all three equilibrium regimes, then derive the explicit forms given in equation (19).

Overloaded regime ($0 < p < c'(\lambda/N)$). Here $\mu^\star = \mu_0 = (p/(qc_E))^{1/(q-1)}$, $B(N\mu_0; \lambda) = 1$ (since $N\mu_0 < \lambda$), and $\Theta = N\mu_0$. By equation (4), each server's equilibrium utility is $U(\mu_0, \mu_0) = p\mu_0 - c_E \mu_0^q$. Therefore

$$W = (r-p)\,N\mu_0 + N\left(p\mu_0 - c_E\mu_0^q\right) = r\,N\mu_0 - N\,c_E\mu_0^q. \tag{E.6}$$

From the overloaded FOC $p = qc_E\mu_0^{q-1}$ (the first piece of equation (7)), we obtain

$$c_E\mu_0^q = c_E\mu_0^{q-1} \cdot \mu_0 = \frac{p}{q}\,\mu_0. \tag{E.7}$$

Substituting equation (E.7) into equation (E.6):

$$W_O(p) := r\,N\mu_0 - N\,\frac{p}{q}\,\mu_0 = N\mu_0\left(r - \frac{p}{q}\right). \tag{E.8}$$

Critically loaded regime ($c'(\lambda/N) \le p \le \frac{N}{N-1}c'(\lambda/N)$). Here $\mu^\star = \lambda/N$, $B(N\cdot\lambda/N;\lambda) = 1$, and $\Theta = \lambda$. By equation (4), each server's utility is $U(\lambda/N, \lambda/N) = p\lambda/N - c_E(\lambda/N)^q$. Therefore

$$W = (r-p)\,\lambda + N\left(\frac{p\lambda}{N} - c_E\left(\frac{\lambda}{N}\right)^q\right) = r\lambda - Nc_E\left(\frac{\lambda}{N}\right)^q. \tag{E.9}$$

Define $W_C := r\lambda - Nc_E(\lambda/N)^q$. This expression is independent of p: within the critical region, adjusting the prize redistributes surplus but does not change its value.

Underloaded regime ($p > \frac{N}{N-1}c'(\lambda/N)$). Here $\mu^\star = \mu_{\mathrm{U}} = (p\lambda(1-1/N)/(Nqc_E))^{1/q}$, $B(N\mu_{\mathrm{U}};\lambda) = \lambda/(N\mu_{\mathrm{U}})$, and $\Theta = \lambda$. By equation (4), each server's utility is

$$U(\mu_{\mathrm{U}}, \mu_{\mathrm{U}}) = p\mu_{\mathrm{U}} \cdot \frac{\lambda}{N\mu_{\mathrm{U}}} - c_E\mu_{\mathrm{U}}^q = \frac{p\lambda}{N} - c_E\mu_{\mathrm{U}}^q.$$

Therefore

$$W = (r-p)\,\lambda + N\left(\frac{p\lambda}{N} - c_E\mu_{\mathrm{U}}^q\right) = r\lambda - Nc_E\mu_{\mathrm{U}}^q. \tag{E.10}$$

Since $\mu_{\mathrm{U}} = (p\lambda(N-1)/(N^2qc_E))^{1/q}$, we have $\mu_{\mathrm{U}}^q = p\lambda(N-1)/(N^2qc_E)$, so

$$Nc_E\mu_{\mathrm{U}}^q = \frac{p\lambda(N-1)}{Nq},$$

and thus $W_U(p) = r\lambda - p\lambda(N-1)/(Nq)$. Since $\lambda(N-1)/(Nq) > 0$, W_U is strictly decreasing in p. The surplus is thus strictly decreasing over the entire underloaded regime.

Continuity at the boundaries. At $p = c'(\lambda/N)$: we have $\mu_0(c'(\lambda/N)) = (c'(\lambda/N)/(qc_E))^{1/(q-1)} = ((qc_E(\lambda/N)^{q-1})/(qc_E))^{1/(q-1)} = \lambda/N$. Therefore

$$\begin{aligned} W_O(c'(\lambda/N)^-) &= N\cdot\frac{\lambda}{N}\cdot\left(r - \frac{c'(\lambda/N)}{q}\right) = \lambda\left(r - c_E\left(\frac{\lambda}{N}\right)^{q-1}\right) \\ &= r\lambda - \lambda\cdot c_E\left(\frac{\lambda}{N}\right)^{q-1} = r\lambda - Nc_E\left(\frac{\lambda}{N}\right)^q = W_C, \end{aligned}$$

where the second-to-last equality uses $\lambda\cdot(\lambda/N)^{q-1} = N\cdot(\lambda/N)^q$.

At $p = \frac{N}{N-1}c'(\lambda/N)$: substituting into W_U gives

$$W_U = r\lambda - \frac{N}{N-1}c'\left(\frac{\lambda}{N}\right)\cdot\frac{\lambda(N-1)}{Nq} = r\lambda - \frac{c'(\lambda/N)\,\lambda}{q} = r\lambda - Nc_E\left(\frac{\lambda}{N}\right)^q = W_C.$$

Thus W is continuous across all three regimes.

Step 2: Optimization of W_O over the overloaded domain. Recall $W_O(p) = N\mu_0(p)(r - p/q)$ and $\mu_0(p) = (p/(qc_E))^{1/(q-1)}$. By logarithmic differentiation, $\mathrm{d}\mu_0/\mathrm{d}p = \mu_0/((q-1)p)$. Differentiating W_O:

$$\begin{aligned} W_O'(p) &= N\left[\frac{\mathrm{d}\mu_0}{\mathrm{d}p}\left(r - \frac{p}{q}\right) + \mu_0\left(-\frac{1}{q}\right)\right] \\ &= N\mu_0\left[\frac{r - p/q}{(q-1)p} - \frac{1}{q}\right]. \end{aligned} \tag{E.11}$$

Since $N\mu_0 > 0$ for all $p > 0$, we have $W_O'(p) = 0$ if and only if

$$\frac{r - p/q}{(q-1)p} = \frac{1}{q}.$$

Cross-multiplying: $q(r - p/q) = (q-1)p$, i.e., $qr - p = (q-1)p$, which gives $qr = qp$, hence $p = r$.

To verify the sign of W_O': the factor in brackets in equation (E.11) equals

$$\frac{r - p/q}{(q-1)p} - \frac{1}{q} = \frac{q(r-p/q) - (q-1)p}{q(q-1)p} = \frac{q(r-p)}{q(q-1)p} = \frac{r-p}{(q-1)p}.$$

Since $(q-1)p > 0$ for $p > 0$, we have:

$$W_O'(p) \gtreqless 0 \iff p \gtreqless r. \tag{E.12}$$

Thus W_O is strictly increasing on $(0, r)$ and strictly decreasing on (r, ∞), with unique global maximum at $p = r$ and value

$$W_O(r) = N\left(\frac{r}{qc_E}\right)^{1/(q-1)} \cdot \frac{r(q-1)}{q}. \tag{E.13}$$

Step 3: Global comparison. We combine the results from Steps 1 and 2. Since W_U is strictly decreasing and $W_U((\frac{N}{N-1}c'(\lambda/N))^+) = W_C$ by continuity, the underloaded regime is strictly dominated by the critically loaded regime. The global maximization therefore reduces to comparing $\max_{0<p<c'(\lambda/N)} W_O(p)$ against W_C.

Case (i): $r < c'(\lambda/N)$. The unconstrained maximizer $p = r$ lies strictly inside $(0, c'(\lambda/N))$. By the unimodality established in equation (E.12), W_O attains its unique maximum on this interval at $p = r$, with value $W_O(r)$ given by equation (E.13). Moreover, the strict decrease on $(r, c'(\lambda/N))$ gives $W_O(r) > W_O(c'(\lambda/N)^-) = W_C$. Since W_C is the best achievable outside the overloaded regime, we conclude that the unique surplus-maximizing prize is $p_S^\star = r$, with

$$W^\star = W_O(r) = N\left(\frac{r}{qc_E}\right)^{1/(q-1)} \cdot \frac{r(q-1)}{q}.$$

At $p_S^\star = r < c'(\lambda/N)$, Proposition 1 gives $\mu^\star = \mu_0 = (r/(qc_E))^{1/(q-1)} < \lambda/N$, confirming the equilibrium is overloaded. This establishes part (i).

Case (ii): $r \geq c'(\lambda/N)$. The unconstrained maximizer $p = r$ of W_O lies at or above the upper boundary $c'(\lambda/N)$ of the overloaded domain. By equation (E.12), W_O is strictly increasing on $(0, c'(\lambda/N)) \subseteq (0, r]$. Hence $W_O(p) < W_O(c'(\lambda/N)^-) = W_C$ for all $p \in (0, c'(\lambda/N))$. The supremum of W_O over $(0, c'(\lambda/N))$ is not attained in that interval, but equals W_C by continuity. Therefore, the global maximum is $W^\star = W_C = r\lambda - Nc_E(\lambda/N)^q$, achieved by any $p \in [c'(\lambda/N), \frac{N}{N-1}c'(\lambda/N)]$. This establishes part (ii). □

Proof of Proposition 9. Throughout, we use the results of Propositions 7 and 8. Recall that $q > 1$, $r > 0$, $c_E > 0$, and $\lambda > 0$.

Part (i): Under-pricing and surplus optimality.

Case A: $r < c'(\lambda/N)$**.** By Proposition 7(i), $p_D^\star = r/q$ (since $r/q < r < c'(\lambda/N)$). By Proposition 8(i), $p_S^\star = r$. Since $q > 1$, we have $p_D^\star = r/q < r = p_S^\star$.

Case B: $c'(\lambda/N) \le r < qc'(\lambda/N)$**.** We have $r/q < c'(\lambda/N)$ (since $r < qc'(\lambda/N)$), so Proposition 7(i) gives $p_D^\star = r/q$. Since $r \ge c'(\lambda/N)$, Proposition 8(ii) gives $p_S^\star \ge c'(\lambda/N)$. Therefore $p_D^\star = r/q < c'(\lambda/N) \le p_S^\star$.

Case C: $r \ge qc'(\lambda/N)$**.** We have $r/q \ge c'(\lambda/N)$, so Proposition 7(ii) gives $p_D^\star = c'(\lambda/N)$. Since $r \ge qc'(\lambda/N) \ge c'(\lambda/N)$, Proposition 8(ii) gives $p_S^\star \ge c'(\lambda/N) = p_D^\star$.

In all three cases, for any $p_S^\star \in \arg\max_{p>0} W(p)$, we have $p_D^\star \le p_S^\star$.

We next show that $p_D^\star \in \arg\max_{p>0} W(p)$ if and only if $r \ge qc'(\lambda/N)$.

Sufficiency. Suppose $r \ge qc'(\lambda/N)$. Then $r/q \ge c'(\lambda/N)$, and Proposition 7(ii) gives $p_D^\star = c'(\lambda/N)$, inducing a critically loaded equilibrium with throughput λ and surplus $W(p_D^\star) = W_C = r\lambda - Nc_E(\lambda/N)^q$. By Proposition 8(ii), W_C is the maximum surplus, so $p_D^\star$ is surplus-optimal and $p_D^\star \in \arg\max_{p>0} W(p)$.

Necessity. Suppose $r < qc'(\lambda/N)$. Then $r/q < c'(\lambda/N)$, and Proposition 7(i) gives $p_D^\star = r/q$, inducing an overloaded equilibrium. We next show $W(p_D^\star) < W^\star$ (hence $p_D^\star \notin \arg\max_{p>0} W(p)$).

If $c'(\lambda/N) \le r < qc'(\lambda/N)$, then the designer is overloaded at $p_D^\star = r/q$, with $W(p_D^\star) = W_O(r/q)$. By Proposition 8(ii), $W^\star = W_C$. We show $W_O(r/q) < W_C$. Since $r \ge c'(\lambda/N)$, the maximizer $p = r$ of W_O lies outside $(0, c'(\lambda/N))$, so W_O is strictly increasing on $(0, c'(\lambda/N))$ by equation (E.12). In particular, $r/q < c'(\lambda/N)$ implies $W_O(r/q) < W_O(c'(\lambda/N)^-) = W_C$. Hence $W(p_D^\star) < W^\star$.

If instead $r < c'(\lambda/N)$, then both designer and planner are overloaded: $p_D^\star = r/q$ and $p_S^\star = r$, both in $(0, c'(\lambda/N))$. Since W_O has its unique maximum at $p = r$ and $r/q < r$ (as $q > 1$), the strict unimodality of W_O gives $W_O(r/q) < W_O(r)$. Hence $W(p_D^\star) < W(p_S^\star) = W^\star$.

In both cases, $p_D^\star$ is not surplus-optimal. This establishes part (i).

Part (ii): Relative surplus loss when $r < c'(\lambda/N)$.

In this case both solutions are overloaded. By Proposition 8(i), $W^\star = W_O(r) = N(r/(qc_E))^{1/(q-1)} \cdot r(q-1)/q$. By Proposition 7(i), $p_D^\star = r/q$, and the surplus at the designer's price is $W(p_D^\star) = W_O(r/q) = N(r/(q^2c_E))^{1/(q-1)} \cdot r(q^2-1)/q^2$. The surplus ratio is

$$\begin{aligned}
\frac{W^\star}{W(p_D^\star)} &= \frac{(r/(qc_E))^{1/(q-1)}}{(r/(q^2c_E))^{1/(q-1)}} \cdot \frac{r(q-1)/q}{r(q^2-1)/q^2} \\
&= \left(\frac{r/(qc_E)}{r/(q^2c_E)}\right)^{1/(q-1)} \cdot \frac{(q-1)/q}{(q-1)(q+1)/q^2} \\
&= q^{1/(q-1)} \cdot \frac{q}{q+1} \\
&= \frac{q^{q/(q-1)}}{q+1}. \qquad \text{(E.14)}
\end{aligned}$$

Since $W^\star = W_O(r) > W_O(r/q) = W(p_D^\star)$ by the strict unimodality of W_O at r established in equation (E.12) (because $r/q < r$ when $q > 1$), the ratio $W^\star/W(p_D^\star) > 1$ is strict, hence $q^{q/(q-1)} > q+1$. The relative surplus loss is therefore

$$\frac{W^\star - W(p_D^\star)}{W^\star} = 1 - \frac{q+1}{q^{q/(q-1)}} > 0,$$

which depends only on the cost elasticity q and is independent of r, c_E, λ, and N.

It remains to verify the upper bound $1 - 2/e$. Define $f(q) := (q+1)/q^{q/(q-1)}$ for $q > 1$. We show that f is strictly increasing on $(1,\infty)$ with $\lim_{q\downarrow 1} f(q) = 2/e$ and $\lim_{q\to\infty} f(q) = 1$. The limit at $q \to 1^+$ follows from $q^{q/(q-1)} = \exp(\frac{q}{q-1}\log q) \to e$ (since $\frac{q\log q}{q-1} \to 1$ by l'Hôpital's rule) and $q+1 \to 2$, giving $f(q) \to 2/e$. The limit at $q \to \infty$ follows from $q^{q/(q-1)} \sim q$, giving $f(q) \to 1$. For strict monotonicity, taking logs and differentiating yields

$$(\log f)'(q) = \frac{1}{q+1} + \frac{\log q}{(q-1)^2} - \frac{1}{q-1},$$

and a direct computation shows that $(\log f)'(q) > 0$ is equivalent to $\log q > 2(q-1)/(q+1)$. The function $g(q) := \log q - 2(q-1)/(q+1)$ satisfies $g(1) = 0$ and $g'(q) = (q-1)^2/[q(q+1)^2] > 0$ for $q > 1$, so $g(q) > 0$ on $(1,\infty)$. Hence $(\log f)'(q) > 0$, and f is strictly increasing on $(1,\infty)$. Therefore $f(q) > 2/e$ for all $q > 1$ (the infimum is approached but not attained), and consequently $1 - f(q) < 1 - 2/e \approx 0.264$ for all $q > 1$.

Part (iii): Relative surplus loss when $c'(\lambda/N) \le r < qc'(\lambda/N)$.

In this case the designer is overloaded while the planner achieves critical loading. By Proposition 7(i), $p_D^\star = r/q$ with $W(p_D^\star) = W_O(r/q)$. By Proposition 8(ii), $W^\star = W_C = r\lambda - Nc_E(\lambda/N)^q$. Define $s := r/(qc'(\lambda/N)) \in [1/q,\, 1)$, so that $r = s\,q\,c'(\lambda/N)$.

We express both surplus values in terms of s and $c'(\lambda/N)$. Using $N\mu_0(p) = \lambda\big(p/c'(\lambda/N)\big)^{1/(q-1)}$ (which follows from $\mu_0(p) = (p/(qc_E))^{1/(q-1)}$ and $(c'(\lambda/N)/(qc_E))^{1/(q-1)} = \lambda/N$) and $Nc_E(\lambda/N)^q = c'(\lambda/N)\lambda/q$ (which follows from $c'(\lambda/N) = qc_E(\lambda/N)^{q-1}$):

$$W_C = r\lambda - \frac{c'(\lambda/N)\lambda}{q} = c'(\lambda/N)\lambda\left(sq - \frac{1}{q}\right) = \frac{c'(\lambda/N)\lambda}{q}\left(sq^2 - 1\right), \tag{E.15}$$

and

$$\begin{aligned} W_O(r/q) &= \lambda\left(\frac{r/q}{c'(\lambda/N)}\right)^{1/(q-1)} \cdot \left(r - \frac{r}{q^2}\right) = \lambda\, s^{1/(q-1)} \cdot sq\, c'(\lambda/N) \cdot \frac{q^2-1}{q^2} \\ &= \frac{c'(\lambda/N)\lambda}{q} \cdot s^{q/(q-1)} \cdot (q^2-1). \end{aligned} \tag{E.16}$$

The relative surplus loss is therefore

$$\frac{W^\star - W(p_D^\star)}{W^\star} = 1 - \frac{W_O(r/q)}{W_C} = 1 - \frac{(q^2-1)\,s^{q/(q-1)}}{sq^2-1}.$$

Note that $sq^2 - 1 \geq q - 1 > 0$ because $s \geq 1/q$ and $q > 1$, so the denominator of the relative-loss expression is positive. We verify the boundary values. At $s = 1/q$ (i.e., $r = c'(\lambda/N)$), we have $s^{q/(q-1)} = q^{-q/(q-1)}$ and $sq^2 - 1 = q - 1$, giving

$$1 - \frac{(q^2-1)\, q^{-q/(q-1)}}{q-1} = 1 - \frac{(q+1)}{q^{q/(q-1)}},$$

which recovers part (ii). As $s \to 1^-$ (i.e., $r \to qc'(\lambda/N)^-$), we have that $s^{q/(q-1)} \to 1$ and the ratio becomes $(q^2-1)/(q^2-1) = 1$, so the relative loss tends to 0, consistent with part (i). □