Staffing, Routing, and Payment to Trade Off Speed and Quality in Large Service Systems

Dongyuan Zhan
School of Management, University College London, London, WC1E 6BT, d.zhan@ucl.ac.uk

Amy R. Ward
Booth School of Business, The University of Chicago, Chicago, IL, 60637, amy.ward@chicagobooth.edu

Most common queueing models used for service system design assume the servers work at fixed (possibly heterogeneous) rates. However, real-life service systems are staffed by people, and people may change their service speed in response to incentives. The delicacy is that the resulting service speed is jointly affected by staffing, routing, and payment decisions. Our objective in the paper is to find a *joint* staffing, routing, and payment policy that induces optimal service system performance.

We do this under the assumption that there is a trade-off between service speed and quality, and employees are paid based on both. The employees each selfishly choose their own service speed in order to maximize their own expected utility (which depends on the staffing through their busy time). The endogenous service rate assumption leads to a centralized control problem in which the system manager jointly optimizes over the staffing, routing, and service rate. By solving the centralized control problem under fluid scaling, we find four different economically optimal operating regimes—critically loaded, efficiency-driven, quality-driven, and intentional idling (in which there is simultaneous customer abandonment and server idling). Then, we show that a simple piece-rate payment scheme can be used to solve the associated decentralized control problem under fluid scaling.

Key words: Service Operations; Queueing Games; Fluid Limits; Erlang-A; Strategic Servers

1. Introduction

The service sector occupies a central position in the U.S. economy. For example, it has grown from 53.3% of GDP in 1999 to 62.4% in 2016 (Bureau of Economic Analysis 2017). Not surprisingly, there is much research focused on service system design. One common assumption is that employees work at fixed rates. However, recent empiric work by Buell et al. (2017), Song et al. (2015) and Shunko et al. (2018) demonstrates that system-design related incentives can affect service speed and/or quality. In this paper, we build a theoretic model to investigate such an effect.

The central questions when designing a service system are: how many employees should be staffed and what should be their payment? This is because for many service systems the most significant percentage of their operating costs is labor. Hence there are many studies (e.g., Garnett et al. 2002, Borst et al. 2004, Milkovich and Newman 2004) on staffing and payment. However, these two problems are very often studied separately; for example, in the aforementioned book and papers, the studies on how to structure payment ignore staffing, and the studies on staffing ignore payment design. The issue is that the payment affects employee motivation, which affects the throughput rate of completed tasks, which affects the staffing required to handle a given workload. On the other hand, the staffing level affects how often there are customers waiting, and employees may work faster (slower) in an environment in which there are usually lines (rarely lines). Moreover, how incoming work is routed to the employees can also influence the speed at which they work. As a result, the system design questions of staffing and payment are more nuanced, and must also consider routing.

To study the joint staffing, routing, and payment problem, we begin with a classic model used to inform staffing decisions that ignores employee payment. Then, we enhance the model to incorporate the employee incentives. Specifically, we consider an M/M/N+M queueing model except we do not assume fixed service rates that are exogenously given. Instead, we assume each employee (henceforth called server, to be consistent with the queueing nomenclature) chooses the service rate that maximizes his expected utility, which equals his expected payment. This motivates solving a centralized control problem in which the system manager can control the number of servers staffed, the routing, and the rate at which each server works. The centralized control problem has a cost structure that incorporates server staffing and utilization costs, as well as costs arising from customer abandonment and service quality that deteriorates with speed. After solving the centralized control problem, the system manager can then decide on a payment structure that motivates servers to work at the desired service rate.

The implicit assumption is that the servers have discretion in how long they take to complete each service. This is natural in the professional and complex service work performed by highly skilled workers (such as engineers and physicians), as modeled in Hopp et al. (2007). This is also true in the factory environment at Coverking (a division of Shrin Corporation), which we visited, where sewing is an important component of production, and sewing times decreased under a volume based payment scheme (based on personal communication with Steve Gupta, the President of Coverking). As a final example, employees answering email in a contact center can choose to take more or less time constructing their responses (Hasija et al. 2010). Then, it is natural to assume that the longer a server spends on a service, the more likely that service is to be successful.

For a fixed staffing level, the service rates chosen by the servers emerge as a Nash equilibrium solution of the game defined through the server utility function. The main difficulty is that even when servers are individually rewarded (for example, paid for volume and quality), their decisions collectively govern the system performance measures. In particular, the utilization of each individual server, which affects the expected number of customers served in a time unit, depends on the entire vector of service rates chosen by all of the servers. In other words, the server interactions create competition between the servers for customers, which may influence their service rate. Any payment structure that uses server utilization as an input metric must account for the resulting server competition.

The main contributions of this paper are as follows.

- We propose a centralized control problem in which the system manager jointly optimizes over the staffing level, the service rate, and the routing (within the class of so-called "Idle-Time-Order-Based" rules). We establish that the solution to the centralized control problem has all servers working at the same rate. This allows us to focus on service rate vectors that are symmetric Nash equilibria. See Proposition 1.

- We solve the centralized control problem under fluid scaling, as the arrival rate becomes large, and provide conditions under which four different economically optimal operating regimes emerge (critically loaded, efficiency-driven, quality-driven, and intentional idling). See Proposition 2 and Theorem 1.
- We show that under piece-rate payment (that is, servers are paid based on individual volume and quality) there exists a symmetric equilibrium service rate, and specify payment parameters under which first best is achieved in a limiting sense. See Proposition 3 and Theorem 2.
- We develop an explicit and analytically tractable expression for the limiting server utilization when one server works at one rate and all other servers work at a different rate. This expression is useful for solving for a symmetric equilibrium service rate in a many-server queue in which the server utility function depends on the server utilization. See Proposition 4.

The remainder of this paper is organized as follows. First, we review some related literature. Then, in Section 2, we set up the queueing model with strategic servers, define the class of Idle-Time-Order-Based routing rules, and specify the resulting server utilizations. Section 3 introduces the system manager's cost structure and formulates both the centralized control problem (in which the system manager directly controls the service rate) and the decentralized control problem (in which the system manager must use payment to influence the service rate). We note that any first best contract requires first solving the centralized control problem. Section 4 solves the centralized control problem under fluid scaling, as the arrival rate becomes large, from which we see four different economically optimal operating regimes emerge. In Section 5, we provide the piece-rate payment contract that is limiting first best. From an economics perspective, that piece-rate payment contract realizes complete risk transfer when the service failure cost is linear. We make concluding remarks in Section 6. The proofs of all results in this paper are in the electronic companion (EC). There is a table of notation Table EC.1 at the beginning of the EC.

1.1. Literature Review

Our model assumes there is a trade-off between speed and quality that can be critical to the customer experience of the service. This is true in the call center application settings in the recent papers by Mehrotra et al. (2012) and Zhan and Ward (2014). This is also true in many other application settings, such as healthcare and manufacturing (Lovejoy and Sethuraman 2000, Kostami and Rajagopalan 2014, Alizamir et al. 2013). The difference between these papers and ours is that none of the aforementioned papers model the service rate as a decision made by a selfish, utility-maximizing server.

The papers Hopp et al. (2007), Lu et al. (2009) and Anand et al. (2011) all analyze models in which the server accounts for both speed and quality when choosing the service time that maximizes his utility. Hopp et al. (2007) allow for dynamic decisions, whereas Lu et al. (2009) and Anand et al. (2011) restrict to a static service rate choice, as we do. None of those papers considers the problem of how to staff systems with a large number of servers. Our focus on staffing naturally leads to a fluid analysis with a speed-quality trade-off, that is methodologically more similar to Chan et al. (2014). One main difference is that in that paper the servers are restricted to have two possible speeds, whereas we allow for a continuum of service speeds.

The service rates chosen by the servers are those that maximize their expected utility. In other words, the service rates are the solution to a queueing game. There is a large literature on queueing games, and we refer the reader to Hassin and Haviv (2003) and Hassin (2016). Much of that literature assumes fixed service rates, and focuses on how customers that strategically decide whether or not to queue, and where to queue, affect system performance. Some exceptions (that is, papers that allow the service rates to be a game equilibrium) are Kalai et al. (1992), Gilbert and Weng (1998), Cachon and Harker (2002), Cachon and Zhang (2007), Debo et al. (2008), Geng et al. (2015). Still, the maximum number of servers in all of the aforementioned papers is two.

The problem of how a system manager influences employee behavior can be thought of as a principal-agent problem, pioneered by Akerlof (1970), Spence (1973), Rothschild and Stiglitz

(1976), Holmstrom (1979). When the agents are risk neutral and only the total output is observable, although individual agents have an incentive to free ride, Holmstrom (1982) shows that group penalties can approximate a first best solution arbitrarily closely. In our service system, the outputs of each server (number of finished services and failures) are observable, and the incentive is not the main problem. The issue is that specifying a first best payment contract requires knowing the solution to the centralized control problem, and that is analytically tractable only in a limiting sense. Moreover, there is no result in the literature that establishes the existence of a symmetric equilibrium service rate in a many-server queueing system under any payment incentive structure – and that is necessary for a first best solution.

The spirit of our analysis is most similar to that in Maglaras and Zeevi (2003, 2005), Armony and Maglaras (2004), Allon and Gurvich (2010), Armony and Gurvich (2010), Allon et al. (2017), Gopalakrishnan et al. (2016), Gurvich et al. (2018), and Ibrahim (2018), all of whom use large system asymptotics to tackle service system design problems, in which either the customers or the servers have some decision-making power. However, with the exception of Gopalakrishnan et al. (2016), none of these papers is focused on the effect of many servers in the same firm competing for incoming customers. In Gopalakrishnan et al. (2016), this competition emerges in a fixed-wage or volunteer model, meaning that the service rate chosen by each server maximizes a non-monetary utility. In this paper, the competition emerges because each server's payment is increasing in the number of customers successfully served.

2. A Many Server Queue with Strategic Servers

In an $M/M/N+M$ queue, customers arrive to a service system having $N \in \{0,1,2,\ldots\}$ servers according to a Poisson process with rate $\lambda \geq 0$ per time unit. Each arriving customer independently samples from an exponential distribution with mean $1/\theta > 0$ time units to determine how long that customer is willing to wait for service before abandoning. Customers in the queue are served according to the first-come-first-served discipline (although our results do not require this, due to the exponential distributional assumptions). Each server is fully capable of handling any customer's service requirements. When a customer arrives to find more than one server available, the routing policy specifies if the customer should be delayed or should be handled immediately by one of the available servers, and if so, which server. The time required to serve each customer is independent and exponential and has mean $1/\mu > 0$ time units when the server works at rate μ. The difference in our setting is that each server strategically chooses his service rate to maximize his own utility, which equals the expected steady state payment per time unit.

The system manager must decide on the staffing level N, the server payment contract, and the routing. The staffing level decision is apparent. We discuss routing and payment respectively below.

The routing policy can know when each server last became idle, but cannot assume knowledge of the service rate. We consider routing rules that are based on the time each server has been idle. More specifically, we generalize the class of Idle-Time-Order-Based (IOB) policies proposed in Gopalakrishnan et al. (2016) for an M/M/N queue without abandonment. The generalization is to be able to lower server utilization by allowing servers to idle even when customers are waiting.

Definition 1. Let $\mathcal{I}(t)$ be the set of servers idle at time $t > 0$, and, when $\mathcal{I}(t) \neq \emptyset$, let $\boldsymbol{s}(t) = (s_1, \ldots, s_{|\mathcal{I}(t)|})$ denote the ordered vector of idle servers at time t, where server s_j became idle before server s_k whenever $j < k$. For any $m \in \{1, \ldots, N\}$, let $\mathcal{Q}^m$ be the set of all probability distributions over $\{1, \ldots, m\}$. An *Idle-Time-Order-Based* (IOB) routing rule with parameter T, or IOB(T), is a collection of probability distributions $\{p^1, p^2, \cdots, p^N\}$, with $p^m \in \mathcal{Q}^m$ for all $m \in \{1, \ldots, N\}$ that:

(i) Delays each arriving customer for $T \geq 0$ time units in a holding area, after which the customer joins the queue;

(ii) Chooses idle server $s_i \in \mathcal{I}(t+T)$, that is in the ith position in the vector $\boldsymbol{s}(t+T)$, with probability $p_i^{|\mathcal{I}(t+T)|}$, to handle the request of a customer arriving at time t if $\mathcal{I}(t+T) \neq \emptyset$, and otherwise has the customer join the end of the queue.

Finally, servers do not serve customers in the holding area, and do not idle whenever customers are present in the queue.

Servers can idle under an IOB(T) routing rule with $T > 0$ when customers are waiting in the holding area. In contrast, an IOB(0) routing rule is non-idling (that is, servers cannot idle when customers are waiting). The sub-class of non-idling IOB policies includes such common policies as randomized routing (in which the server that handles the incoming customer request is chosen at random) and longest-idle-server-first (in which the server that has idled the longest handles the incoming customer request). These can all be modified to be an IOB(T) routing rule with $T > 0$ by intentionally delaying customers as in Definition 1.

The servers are risk neutral strategic players in a non-cooperative game, each capable of completing services at any rate $\mu \in [\underline{\mu}, \overline{\mu}]$, where $0 < \underline{\mu} < \overline{\mu} < \infty$. Server $i \in \{1, \ldots, N\}$ has utility $U_i(\vec{\mu})$ that equals his expected payment when the service rate vector is $\vec{\mu}$. The service rates the servers choose will be a Nash equilibrium of this game. In particular, an equilibrium is a service rate vector $\vec{\mu}$ that satisfies

$$U_i(\vec{\mu}) = \max_{v \in [\underline{\mu}, \overline{\mu}]} U_i(\mu_1, \cdots, \mu_{i-1}, v, \mu_{i+1}, \cdots, \mu_N) \text{ for all } i \in \{1, \ldots, N\}, \tag{1}$$

and individual rationality; that is,

$$U_i(\vec{\mu}) \geq c_S \text{ for } i \in \{1, \ldots, N\}, \tag{2}$$

where $c_S > 0$ represents the expected payment from an outside employment alternative. Although server i can individually choose his service rate μ_i, server i cannot maximize his expected utility rate without worrying about the behavior of others since $U_i(\vec{\mu})$ is in general a function of the entire service rate vector $\vec{\mu}$.

We assume the expected payment to server i, $U_i(\vec{\mu})$, is a function of the system steady state performance. One important steady state performance measure is server utilization, or the percentage of time each server is busy. For example, if servers are paid per task completed, then each server's payment $U_i(\vec{\mu})$ depends on his utilization.

LEMMA 1. *In an $M/M/N+M$ queue with arrival rate λ, service rate vector $\vec{\mu}$, and impatience rate θ, all IOB(T) routing policies have the same steady state probabilities. As a consequence, all IOB(T) routing policies result in the same expected steady state utilization of server i,*

$$B_i(\vec{\mu}, N, T) := \frac{\sum_{\mathcal{I} \subseteq \{1,\ldots,N\} \setminus i} |\mathcal{I}|! \prod_{j \in \mathcal{I}} \frac{\mu_j}{\lambda \exp(-\theta T)} + \sum_{m=1}^{\infty} \prod_{k=1}^{m} \frac{\lambda \exp(-\theta T)}{k\theta + \sum_{j=1}^{N} \mu_j}}{\sum_{\mathcal{I} \subseteq \{1,\ldots,N\}} |\mathcal{I}|! \prod_{j \in \mathcal{I}} \frac{\mu_j}{\lambda \exp(-\theta T)} + \sum_{m=1}^{\infty} \prod_{k=1}^{m} \frac{\lambda \exp(-\theta T)}{k\theta + \sum_{j=1}^{N} \mu_j}}, \textit{ for } i \in \{1, \ldots, N\},$$

where $\prod_{j \in \emptyset} \frac{\mu_j}{\lambda \exp(-\theta T)} := 1$; that is, the product of elements from an empty set is 1 by convention.

Lemma 1 is remarkable in the sense that the steady state probabilities (and, consequently, the server utilizations) do not depend on the collection of probability distributions used to define IOB(T) routing in Definition 1. As a consequence, the system performance is uniquely specified by setting the IOB parameter T.

A more general formulation would allow the servers to adjust their service rates dynamically over time. The restriction to constant service rate choice ensures the control problem under consideration in the next section does not have the added complication of being a dynamic control problem.

3. The Control Problem

The equilibrium service rate in the many-server queue with strategic servers is determined by the staffing, the routing, and the payment, which are all decisions. These decisions are made to optimize system performance. Section 3.1 details the assumed cost structure. If the system manager could directly control the service rate, then the relevant optimization would be a centralized control problem to determine the staffing, the routing, and the service rate, and we formulate that problem in Section 3.2. The centralized control problem provides a lower bound on the minimum possible cost. First best is achieved when the staffing, the routing, and the payment contract are set so as to attain that lower bound cost, which requires studying the decentralized control problem detailed in Section 3.3.

3.1. The Cost Structure

The salary cost of staffing N servers that work at service rate vector $\vec{\mu}$ is $\sum_{i=1}^{N} U_i(\vec{\mu})$, which equals or exceeds $c_S N$ by the individual rationality constraint (2). The operational costs include costs for utilization, customer abandonment, and low service quality. We assume the system is operating in steady state, and discuss each cost in turn below. The notation $\beta_i = B_i(\vec{\mu}, N, T)$ refers to the expected steady state utilization of server $i \in \{1, \ldots, N\}$ given in Lemma 1.

The utilization cost for each server $g_U : [0,1] \rightarrow [0,\infty)$ captures server fatigue or machine overuse, and is often convex because such costs tend to increase more quickly as utilization becomes closer and closer to 1. The total utilization cost is

$$\sum_{i=1}^{N} g_U(\beta_i). \tag{3}$$

The customer abandonment cost is captured through a function $g_A : [0,1] \rightarrow [0,\infty)$ that represents the cost per abandoned customer. That cost depends on the steady state customer abandonment probability q_A, and can be linear or non-linear. The function should be non-linear when a small abandonment probability has almost no impact on reputation but a larger probability leads to considerable damage. When server i has expected steady state utilization β_i, the rate at which customers depart from server i is $\beta_i \mu_i$, which leads to the expected steady state number of abandonments in a time unit being $\lambda - \sum_{i=1}^{N} \beta_i \mu_i$. From the flow balance equation of the system, we must have $\lambda(1-q_A) = \sum_{i=1}^{N} \beta_i \mu_i$, or $q_A = \frac{\lambda - \sum_{i=1}^{N} \beta_i \mu_i}{\lambda}$. This results in an abandonment cost per time unit of

$$\left(\lambda - \sum_{i=1}^{N} \beta_i \mu_i\right) g_A(q_A) = \left(\lambda - \sum_{i=1}^{N} \beta_i \mu_i\right) g_A\left(\frac{\lambda - \sum_{i=1}^{N} \beta_i \mu_i}{\lambda}\right). \tag{4}$$

Service quality decreases when servers work faster, which is costly. This speed-quality trade-off is captured through a strictly decreasing function $p : [\underline{\mu}, \overline{\mu}] \rightarrow [0,1]$ that specifies the probability of successful service. The outcome of a service as either successful or failed is known when outcomes can be captured by, for example, customer complaints or customer evaluation forms (in which case a failed service does not imply the customer returns). The function $g_F : [0,1] \rightarrow [0,\infty)$ represents the cost per failed service, and depends on the steady state service failure probability q_F. As in the case of customer abandonment, g_F may be linear or non-linear, depending on whether or not having the unit cost increase in the service failure percentage is appropriate. The steady state service failure probability q_F is specified in terms of the steady state probability q_i that an arriving customer is served by server i, given that customer does not abandon; that is, $q_F = \sum_{i=1}^{N} q_i(1-p(\mu_i))$. To determine q_i, note that since customers depart from server i at rate $\beta_i \mu_i$, the flow balance of server

i gives $\lambda(1-q_A)q_i=\beta_i\mu_i$, which implies $q_i=\frac{\beta_i\mu_i}{\sum_{j=1}^N\beta_j\mu_j}$. The expected number of failed services of server i per time unit is $(1-p(\mu_i))\beta_i\mu_i$, which leads to an expected failed service cost per time unit of

$$\sum_{i=1}^N(1-p(\mu_i))\beta_i\mu_i g_F(q_F)=\sum_{i=1}^N(1-p(\mu_i))\beta_i\mu_i g_F\left(\frac{\sum_{i=1}^N(1-p(\mu_i))\beta_i\mu_i}{\sum_{i=1}^N\beta_i\mu_i}\right). \tag{5}$$

Assumption 1. *The functions g_U, g_A, g_F, and p are all continuous. The function g_U is weakly increasing and weakly convex on $[0,1]$. The function p is strictly decreasing and weakly concave on $[\underline{\mu},\overline{\mu}]$. The function g_F is weakly increasing on $[0,1]$.*

3.2. The Centralized Control Problem

A lower bound on the minimum attainable cost is found by solving a centralized control problem, in which the system manager can directly dictate the service rate vector $\vec{\mu}$ as well as the staffing level N and the parameter T that defines the IOB routing rule. The centralized control problem minimizes the sum of staffing and operational costs, subject to feasibility constraints. The staffing costs are c_SN, because the service rates do not depend on the expected payment, meaning the only requirement is to satisfy individual rationality (2). The operational costs are

$$\mathcal{C}(\vec{\mu},N,T):= \tag{6}$$
$$\sum_{i=1}^N g_U(\beta_i)+\left(\lambda-\sum_{i=1}^N\beta_i\mu_i\right)g_A\left(\frac{\lambda-\sum_{i=1}^N\beta_i\mu_i}{\lambda}\right)+\sum_{i=1}^N(1-p(\mu_i))\beta_i\mu_i g_F\left(\frac{\sum_{i=1}^N(1-p(\mu_i))\beta_i\mu_i}{\sum_{i=1}^N\beta_i\mu_i}\right),$$

for

$$\beta_i=B_i(\vec{\mu},N,T)$$

defined as in Lemma 1. Intuition suggests that an IOB routing rule that intentionally delays customers results in lower server utilization than one that does not, which follows from the below Lemma.

Lemma 2. *For any $N\in\{1,2\cdots\}$ and any service rate vector $\vec{\mu}$, $B_i(\vec{\mu},N,T)$ is strictly decreasing in T on $[0,\infty)$ for each $i\in\{1,2,\ldots,N\}$.*

The centralized control problem is

$$\begin{aligned} &\min_{\vec{\mu},N,T} c_S N + \mathcal{C}(\vec{\mu},N,T) \\ &\text{subject to: } N \in \{0,1,2,\ldots\}, \mu_i \in [\underline{\mu},\overline{\mu}] \text{ for all } i \in \{1,2,\ldots,N\}, \text{ and } T \geq 0. \end{aligned} \tag{7}$$

We let $(\vec{\mu}_\star, N_\star, T_\star)$ denote a solution to (7), and $\mathcal{C}_\star := \mathcal{C}(\vec{\mu}_\star, N_\star, T_\star)$. The minimum objective function value is $c_S N_\star + \mathcal{C}_\star$.

Proposition 1. *Under Assumption 1, any solution $(\vec{\mu}_\star, N_\star, T_\star)$ to the centralized control problem (7) has all servers working at the same service rate $\vec{\mu}_{\star,i} = \mu_\star \in [\underline{\mu},\overline{\mu}]$, for all $i \in \{1,2,\cdots,N_\star\}$, and having the same utilization $B_i(\vec{\mu}_\star, N_\star, T_\star) = \beta_\star \in [0, B(\vec{\mu}_\star, N_\star, 0)]$, for all $i \in \{1,2,\ldots,N_\star\}$.*

The upper bound of $\beta_\star$ in Proposition 1 follows from Lemma 2.

We slightly abuse notation and write $B(\mu,N,T)$ and $\mathcal{C}(\mu,N,T)$ to be the server utilization and operational costs when all servers work at the same rate μ. Lemma 1 can be used to find an explicit expression for $B(\mu,N,T)$, and to verify that the utilization is the same for all servers. From (6), the operational costs are,

$$\mathcal{C}(\mu,N,T) = N g_U(\beta) + (\lambda - N\beta\mu)\, g_A\left(\frac{\lambda - N\beta\mu}{\lambda}\right) + N(1-p(\mu))\beta\mu g_F\left(1-p(\mu)\right), \text{ for } \beta = B(\mu,N,T). \tag{8}$$

Proposition 1 implies the centralized control problem is equivalently specified as

$$\begin{aligned} &\min_{\mu,N,T} c_S N + \mathcal{C}(\mu,N,T) \\ &\text{subject to: } \mu \in [\underline{\mu},\overline{\mu}], N \in \{0,1,2,\ldots\}, T \geq 0. \end{aligned} \tag{9}$$

The minimum objective function value in (9) equals $c_S N_\star + \mathcal{C}_\star$ under Assumption 1. We write $(\mu_\star, N_\star, T_\star)$ to denote a solution to (9), in which case $(\vec{\mu}_\star, N_\star, T_\star)$ solves (7), where $\vec{\mu}_{\star,i} = \mu_\star$ for all $i \in \{1,\ldots,N_\star\}$, under Assumption 1.

3.3. The Decentralized Control Problem

The system manager would like to choose the payment contracts, staffing level N, and IOB routing parameter T that ensures the equilibrium service rate will be such that the sum of staffing and operational costs equals the lower bound $c_S N_\star + \mathcal{C}_\star$. This is a decentralized control problem because each individual agent $i \in \{1,\ldots,N\}$ controls his own service rate, which is chosen so as to selfishly maximize his own expected payment U_i.

The payment contracts that produce the vector of expected server payments $\vec{U} = (U_1,\ldots,U_N)$ must be computable by relying only on observable or known elements. We assume that the system manager can observe the realized number of abandonments and number of completed and failed services for each server during any finite time interval. The manager knows the arrival rate λ, the cost functions g_U, g_A and g_F, as well as the speed quality trade-off function p. We let $\mathcal{P}$ denote the class of payment contracts specified by functions whose domains depend only on known and observable parameters over a time unit and have range that is the N-dimensional non-negative orthant. Then, $U_i = E[P_i]$, $i \in \{1,\ldots,N\}$, where $\vec{P} = (P_1,\ldots,P_N) \in \mathcal{P}$, and the expectation operator is with respect to the steady state distribution.

The differentiation between $\vec{P}$ and $\vec{U}$ is conceptually important. This is because P_i, $i \in \{1,\cdots,N\}$, must be computed based on observable or known elements, and so cannot explicitly depend on the service rate vector $\vec{\mu}$, even though U_i can. For example, if servers are paid P_S for each completed service and the random variable X_i represents the observed number of completed services by server i in a time unit, then $P_i = P_S X_i$ for $i \in \{1,\cdots,N\}$, which we can compute without knowing $\vec{\mu}$ (even though X_i implicitly depends on $\vec{\mu}$). However, the computation $U_i = P_S \mu_i B_i(\vec{\mu},N,T)$ explicitly depends on $\vec{\mu}$. We let $\mathcal{S}(\vec{P},N,T)$ denote the set of equilibrium service rate vectors (which could be the empty set) under staffing level N, IOB(T) routing, and payment contract vector $\vec{P}$.

The decentralized control problem is

$$\begin{aligned} &\min_{\vec{P},N,T} \sup_{\vec{\mu}_E \in \mathcal{S}(\vec{P},N,T)} \sum_{i=1}^{N} U_i + \mathcal{C}(\vec{\mu}_E,N,T) \\ &\text{subject to: } \vec{P} \in \mathcal{P}, N \in \{0,1,2,\ldots\}, \mathcal{S}(\vec{P},N,T) \neq \emptyset, T \geq 0, \\ &\min_{\vec{\mu}_E \in \mathcal{S}(\vec{P},N,T)} U_i = E[P_i] \geq c_S \text{ for each } i \in \{1,\ldots,N\}. \end{aligned} \tag{10}$$

The formulation (10) allows for the worst possible equilibrium (from the system manager's percepective) in the case of multiple equilibria. If $\vec{\mu}_E \in \mathcal{S}(\vec{P},N,T)$, then $(\vec{\mu}_E,N,T)$ is feasible for the centralized control problem (9). As a result, the solution to the decentralized control problem has a lower bound $c_S N_\star + \mathcal{C}_\star$. If there exists an asymmetric equilbrium, from Proposition 1 the associated cost will exceed $c_S N_\star + \mathcal{C}_\star$. This observation motivates us to focus on finding a symmetric equilibrium service rate. For the remainder of this paper, an equilibrium refers to a symmetric equilibrium, which satisfies both the equilibrium definition (1) and has all components identical, denoted by $\vec{\mu}_E = (\mu_E, \ldots, \mu_E)$. Under a symmetric equilibrium service rate, if the manager uses a common payment contract for all servers, then U_i will be identical for $i \in \{1, \ldots, N\}$ and (10) simplifies with an objective $U_1 N + \mathcal{C}(\mu, N, T)$.

The decentralized control problem can be solved via the centralized control problem by setting $N = N_\star$, $T = T_\star$, and specifying a common payment contract P that motivates each server to work at rate $\mu_\star$. Such a contract would be easy to specify if the system manager could observe the long-run average number of services completed per unit busy time, that is, the service rate $\mu_i, i \in \{1, \ldots, N_\star\}$. Then, any contract that pays strictly less than c_S for service rate not equal to $\mu_\star$, such as

$$U_i = c_S - (\mu_i - \mu_\star)^2, i \in \{1, \ldots, N_\star\}, \tag{11}$$

ensures each server working at $\mu_\star$ is a unique equilibrium, and so is first best.

The issue is that (11) requires solving the centralized control problem (9). Unfortunately, (9) is a complicated optimization that may have many local minima, because the objective function is not convex (unless additional assumptions are imposed). Hence numeric solution techniques are not a panacea. Furthermore, numeric solution techniques do not yield easy insight into solution structure. For example, we would like to understand conditions under which an optimal solution results in low customer abandonments and/or high server utilization, which does not appear possible from an exact analysis. Therefore, we perform an asymptotic analysis.

4. Asymptotic Analysis of the Centralized Control Problem

We solve (9) asymptotically, by allowing the arrival rate to become large. Section 4.1 sets up the asymptotic regime, and Section 4.2 shows the fluid control problem that (9) gives rise to in that regime. We use the solution to the aforementioned fluid control problem to propose a policy for setting the staffing level, service rate, and routing parameter. We establish that that policy is asymptotically optimal (that is, has identical asymptotic performance to a solution to (7), equivalently (9) by Proposition 1) in Section 4.3. The fluid control problem solution allows us to identify conditions on the cost structure under which different operating regimes are economically optimal, and we detail those in Section 4.4.

4.1. Preliminaries

We consider a sequence of systems with increasing arrival rate $\lambda \in (0, \infty)$. Our convention, when we refer to any quantity associated with the system with arrival rate λ, that may change with λ, is to superscript the appropriate symbol by λ.

Definition 2. A *policy* is a sequence of staffing levels, service rates, and IOB routing parameters $\{(\mu^\lambda, N^\lambda, T^\lambda) : \lambda \geq 0\}$. An *admissible policy* satisfies the constraints of the centralized control problem (9) for every λ.

We would like to find an admissible policy that has close to the minimum cost $c_S N_\star^\lambda + \mathcal{C}_\star^\lambda$ of the centralized control problem (7), which equals the minimum cost of (9) under Assumption 1 by Proposition 1.

Definition 3. An admissible policy $\{(\mu^\lambda, N^\lambda, T^\lambda) : \lambda \geq 0\}$ is *asymptotically optimal* if

$$\lim_{\lambda \to \infty} \frac{c_S N^\lambda + \mathcal{C}(\mu^\lambda, N^\lambda, T^\lambda)}{c_S N_\star^\lambda + \mathcal{C}_\star^\lambda} = 1.$$

Initially, we do not know the functional form an asymptotically optimal policy can take. The following proposition highlights that an asymptotically optimal policy should not have a staffing level that grows faster than linear in the arrival rate.

Lemma 3. *Any asymptotically optimal policy has*

$$\limsup_{\lambda \to \infty} \frac{N^\lambda}{\lambda} < \infty.$$

Lemma 3 implies that to find an asymptotically optimal policy we need only search within the class of admissible policies for which [1]

$$N^\lambda = b\lambda + o(\lambda) \text{ for some } b \geq 0. \tag{12}$$

Lemma 4. *Under IOB(T^λ) routing and staffing that satisfies (12), if $\mu^\lambda \to \mu \in [\underline{\mu}, \overline{\mu}]$ and $T^\lambda \to T \in [0, \infty]$ as $\lambda \to \infty$, then*

$$\lim_{\lambda\to\infty} B(\mu^\lambda, N^\lambda, T^\lambda) = \beta := \min\left(1, \frac{\exp(-\theta T)}{b\mu}\right),$$

and the expected steady state abandonment probability satisfies

$$\frac{\lambda - N^\lambda B(\mu^\lambda, N^\lambda, T^\lambda)\mu^\lambda}{\lambda} \to a := 1 - b\beta\mu \geq 0 \text{ as } \lambda \to \infty.$$

Lemma 4 motivates taking the limit as $\lambda \to \infty$ in the centralized control problem (9) in order to gain analytic tractability. More specifically, under the conditions given in Lemma 4, the limiting objective function value to (9) satisfies

$$\lim_{\lambda\to\infty} \frac{c_S N^\lambda + C(\mu^\lambda, N^\lambda, T^\lambda)}{\lambda} = c_S b + \hat{\mathcal{C}}(\mu, b, \beta, a), \tag{13}$$

where

$$\hat{\mathcal{C}}(\mu, b, \beta, a) := b g_U(\beta) + a g_A(a) + (1 - p(\mu))\mu\beta b g_F(1 - p(\mu)).$$

4.2. The Limiting Control Problem

The limiting control problem is

$$\begin{gathered}\min_{\mu, b, \beta, a} c_S b + \hat{\mathcal{C}}(\mu, b, \beta, a) \\ \text{subject to: } \mu \in [\underline{\mu}, \overline{\mu}], b \geq 0, \beta \in [0, 1], \text{ and } a = 1 - b\beta\mu \geq 0.\end{gathered} \tag{14}$$

We denote a solution to the limiting control problem by $(\hat{\mu}_\star, \hat{b}_\star, \hat{\beta}_\star, \hat{a}_\star)$, and let $\hat{\mathcal{C}}_\star := \hat{\mathcal{C}}(\hat{\mu}_\star, \hat{b}_\star, \hat{\beta}_\star, \hat{a}_\star)$. The minimum objective function value is $c_S \hat{b}_\star + \hat{\mathcal{C}}_\star$.

The decision variables in (14) do not include the routing parameter because given b and μ, from Lemma 4, any limiting utilization $\beta \in \left[0, \min\left(1, \frac{1}{b\mu}\right)\right]$ is achievable. In order to have limiting busy percentage $\beta \in \left[0, \min\left(1, \frac{1}{b\mu}\right)\right]$, we solve $\frac{\exp(-\theta T)}{b\mu} = \beta$ for T, and set $T^\lambda = -\log(b\mu\beta)/\theta$ for all λ.

Define

$$\hat{c}_S(\beta) := \frac{c_S + g_U(\beta)}{\beta} > 0 \text{ for all } \beta \in [0,1].$$

to be an adjusted staffing cost, that also accounts for utilization. To solve (14), we first use the equality constraint to re-write the objective function as follows

$$c_S b + \hat{\mathcal{C}}(\mu, b, \beta, a) = (1-a)\left(\frac{\hat{c}_S(\beta)}{\mu} + (1-p(\mu))g_F(1-p(\mu))\right) + a g_A(a).$$

Next, provided all minimizers are unique, we can observe

$$\hat{\beta}_\star = \arg\min_{\beta\in[0,1]} \{\hat{c}_S(\beta)\} \tag{15}$$

$$\hat{\mu}_\star = \arg\min_{\mu\in[\underline{\mu},\overline{\mu}]} \left\{ \frac{\hat{c}_S(\hat{\beta}_\star)}{\mu} + (1-p(\mu))g_F(1-p(\mu)) \right\} \tag{16}$$

$$\hat{a}_\star = \arg\min_{a\in[0,1]} \left\{ (1-a)\left(\frac{\hat{c}_S(\hat{\beta}_\star)}{\hat{\mu}_\star} + (1-p(\hat{\mu}_\star))g_F(1-p(\hat{\mu}_\star))\right) + a g_A(a) \right\} \tag{17}$$

$$\hat{b}_\star = \frac{1-\hat{a}_\star}{\hat{\beta}_\star \hat{\mu}_\star}. \tag{18}$$

Assumption 2. *The functions g_U, g_A, g_F, and p are all continuous. Furthermore, g_U is strictly convex on $[0,1]$, p is weakly concave on $[\underline{\mu},\overline{\mu}]$, g_F is weakly convex and weakly increasing on $[0,1]$, $a g_A(a)$ is strictly convex in a on $[0,1]$, and $\left(\frac{\hat{c}_S(\hat{\beta}_\star)}{\hat{\mu}_\star} + (1-p(\hat{\mu}_\star))g_F(1-p(\hat{\mu}_\star))\right) \neq g_A(1)$.*

Lemma 5. *Under Assumption 2, the minimizers in (15)-(18) are unique.*

4.3. An Asymptotically Optimal Policy

An asymptotically optimal policy uses the solution to the limiting control problem (14) to set the staffing level, the routing parameter, and the service rate.

Theorem 1. *Under Assumptions 1 and 2, any policy*

$$\left(\hat{\mu}_\star, N^\lambda_{ao}, \hat{T}_\star\right) \textit{ having } N^\lambda_{ao} = \hat{b}_\star\lambda + o(\lambda) \textit{ and } \hat{T}_\star := \begin{cases} -\log(\hat{b}_\star\hat{\mu}_\star\hat{\beta}_\star)/\theta & \textit{if } \hat{\beta}_\star < 1 \textit{ and } \hat{a}_\star > 0, \\ 0 & \textit{otherwise} \end{cases}$$

is asymptotically optimal. Furthermore,

$$\lim_{\lambda\to\infty} B\left(\hat{\mu}_\star, N^\lambda_{ao}, \hat{T}_\star\right) = \hat{\beta}_\star = \min\left(1, \frac{\exp(-\theta\hat{T}_\star)}{\hat{b}_\star\hat{\mu}_\star}\right),$$

and

$$\lim_{\lambda\to\infty} \frac{c_S N^{\lambda}_{ao} + \mathcal{C}\big(\hat{\mu}_\star, N^{\lambda}_{ao}, \hat{T}_\star\big)}{\lambda} = \lim_{\lambda\to\infty} \frac{c_S N^{\lambda}_{\star} + \mathcal{C}^{\lambda}_{\star}}{\lambda} = c_S \hat{b}_\star + \hat{\mathcal{C}}(\hat{\mu}_\star, \hat{b}_\star, \hat{\beta}_\star, \hat{a}_\star).$$

Theorem 1 motivates studying properties of the solution (15)-(18), in order to gain insight into the solution to the centralized control problem (9).

The quantity

$$\hat{c}_\star := \frac{\hat{c}_S(\hat{\beta}_\star)}{\hat{\mu}_\star} + (1 - p(\hat{\mu}_\star)) g_F(1 - p(\hat{\mu}_\star)) \tag{19}$$

represents the minimum cost to serve a customer, adjusted to include both utilization and service failure costs. If that minimum cost is too high, then the limiting solution does not staff and lets all customers abandon.

Lemma 6. *Under Assumption 2, if* $\max_{a\in[0,1]} \left(a g_A(a)\right)' = g_A(1) + g'_A(1) \le \hat{c}_\star$*, then* $\hat{a}_\star = 1$ *and* $\hat{b}_\star = 0$.

When the abandonment percentage is a, the scaled abandonment cost is $a g_A(a)$, and so the condition in Lemma 6 states in words that any marginal increase in abandonment cost cannot exceed the minimum cost to serve a customer $\hat{c}_\star$.

Provided the limiting solution does not have zero staff, we can further determine conditions for whether or not to let any customers abandon, and whether or not to push the servers to full utilization.

Proposition 2. *Suppose Assumption 2 holds,* $g_A(1) + g'_A(1) > \hat{c}_\star$.

(a) If $\min_{a\in[0,1]} \left(a g_A(a)\right)' = g_A(0) \ge \hat{c}_\star$*, then* $\hat{a}_\star = 0$*; otherwise,* $\hat{a}_\star \in (0,1)$.

(b) If $\max_{\beta\in[0,1]} \hat{c}'_S(\beta) \le 0$*, or equivalently,* $g'_U(1) - g_U(1) \le c_S$*, then* $\hat{\beta}_\star = 1$*; otherwise,* $\hat{\beta}_\star \in (0,1)$.

In words, Proposition 2(a) states no customers should abandon when the marginal abandonment cost always exceeds the minimum cost to serve a customer. Proposition 2(b) states that if the adjusted staffing cost is weakly decreasing in utilization, then full server utilization is optimal.

4.4. Economically Optimal Limit Regimes

Together, Proposition 2 and Theorem 1 specify four possible operating regimes for the system, which are detailed in Table 1. This is reminiscent of the economically optimal operating regimes detailed in Borst et al. (2004) for an $M/M/N$ queue without abandonment (and under a different cost structure). Although there are papers in the literature that show conditions on a cost structure under which the critically loaded and efficiency-driven limiting regimes arise in an $M/M/N+M$ queue (see, for example, Whitt 2006, Ren and Zhou 2008, and Bassamboo and Randhawa 2010), we are not aware of any paper that has provided a cost structure under which the full spectrum of the four operating regimes detailed in Table 1 arise. Garnett et al. (2002) develop the relationship between staffing levels and performance characteristics in an $M/M/N+M$ queue to define three possible operating regimes (critically loaded, quality-driven, and efficiency-driven), but do not explicitly incorporate a cost structure, and do not discuss an intentional idling regime.

Table 1 Optimal Limiting Regimes. (The table assumes $g_A(1)+g_A'(1)>\hat{c}_\star$)

Abandonment Cost	Utilization Cost	Optimal Regime	Limiting Abandonment Percentage	Limiting Server Utilization
$g_A(0)\geq\hat{c}_\star$	$g_U'(1)-g_U(1)\leq c_S$	Critically loaded	$\hat{a}_\star=0$	$\hat{\beta}_\star=1$
$g_A(0)<\hat{c}_\star$	$g_U'(1)-g_U(1)\leq c_S$	Efficiency-driven	$\hat{a}_\star\in(0,1)$	$\hat{\beta}_\star=1$
$g_A(0)\geq\hat{c}_\star$	$g_U'(1)-g_U(1)>c_S$	Quality-driven	$\hat{a}_\star=0$	$\hat{\beta}_\star\in(0,1)$
$g_A(0)<\hat{c}_\star$	$g_U'(1)-g_U(1)>c_S$	Intentional Idling	$\hat{a}_\star\in(0,1)$	$\hat{\beta}_\star\in(0,1)$

The utilization cost drives the appearance of an intentional idling regime in which servers can idle while customers are waiting (because $\hat{T}_\star$ defined in Theorem 1 is positive). The utilization cost is one mechanism through which management can internalize the negative consequences of overworked servers becoming haggard. Then, when utilization costs are high and abandonment costs are low, the system manager may prefer to allow the servers some rest, even when customers are waiting. For example, in fixed low-wage environments (such as some call centers and government services), too high server utilization may lead to turnover, which may be more expensive than letting customers abandon. In comparison to the literature, the only other papers in which we have seen an intentional idling policy proposed are those in which the customers are heterogeneous (Afeche 2013, Afeche and Pavlin 2016, Maglaras et al. 2018).

EXAMPLE 1. We assume the utilization cost $g_U(\beta) = k\beta^r$, where $r \geq 0$, for $\beta \in [0,1]$ and the abandonment cost $g_A(a) = c_A a^s$, where $s \geq 0$, for $a \in [0,1]$. We further assume $\hat{c}_\star < g_A(1) + g_A'(1) = c_A(s+1)$ so that the staffing is non-zero. The calculus is straightforward, and so is omitted.

(a) Linear and sub-linear utilization cost ($0 \leq r \leq 1$), as well as super-linear utilization cost with small growth rate ($r > 1$ and $k \leq \frac{c_S}{r-1}$), implies $\hat{\beta}_\star = 1$. Then, the economically optimal operating regime is either critically loaded or efficiency-driven, depending on the abandonment cost growth rate; that is,

$$\hat{a}_\star = \begin{cases} 0 & \text{if } s = 0 \\ \left(\frac{\hat{c}_\star}{(s+1)c_A}\right)^{\frac{1}{s}} & \text{if } s > 0 \end{cases}. \tag{20}$$

(b) Super-linear utilization cost with high growth rate ($r > 1$ and $k > \frac{c_S}{r-1}$) implies

$$\hat{\beta}_\star = \left(\frac{c_S}{(r-1)k}\right)^{\frac{1}{r}} < 1.$$

Then, the economically optimal operating regime is either quality-driven or intentional idling, depending on the value of $a_\star$ in (20).

REMARK 1. In Example 1, if the abandonment cost is linear ($s = 0$ so that $g_A(a) = c_A$, which implies that the total abandonment cost is $ag_A(a) = c_A a$) and the utilization cost is low enough as in part (a) in Example 1, then the economically optimal limiting regime is critically loaded. This is consistent with Proposition 5 part (b) in Bassamboo and Randhawa (2010), noting that their condition $c/\mu < p$ is exactly our condition $\hat{c}_\star < c_A$ that implies staffing is non-zero. This is also consistent with Proposition 1 in Ren and Zhou (2008), and also with condition (3.22) in Proposition 1 in Whitt (2006), noting that that condition reduces to $\hat{c}_\star = c_S < c_A$ in our framework (since that paper assumes the service rate is 1, and we do not have waiting cost or revenue).

5. Limiting First Best Payment

We do not have convenient expressions for a solution to the centralized control problem (7). Instead, from Theorem 1, under Assumptions 1 and 2, we have an asymptotically optimal policy $\big(\hat{\mu}_\star, N^\lambda_{\text{ao}}, \hat{T}_\star\big)$, where $N^\lambda_{\text{ao}} = \hat{b}_\star\lambda + o(\lambda)$. Provided a symmetric equilibrium service rate μ^λ_E is close to $\hat{\mu}_\star$ when the staffing level satisfies $\hat{b}_\star\lambda + o(\lambda)$ and the routing parameter is $\hat{T}_\star$, the solution to the decentralized control problem (10) will be close to that of the centralized control problem (7).

Definition 4. Suppose $N^\lambda_{\text{ao}} = \hat{b}_\star\lambda + o(\lambda)$. *Limiting first best payment* is a sequence of contracts $\vec{P}^\lambda \in \mathcal{P}$ such that $\Big\{\big(\vec{P}^\lambda, N^\lambda_{\text{ao}}, \hat{T}_\star\big) : \lambda \geq 0\Big\}$ satisfy the constraints of the decentralized control (10) for every λ, and any sequence of symmetric equilibrium service rates satisfies

$$\lim_{\lambda\to\infty} \big|\mu^\lambda_E - \hat{\mu}_\star\big| = 0 \text{ and } \lim_{\lambda\to\infty} U^\lambda_i - c_S = 0, \quad i \in \{1, 2, \cdots, N^\lambda_{\text{ao}}\}.$$

Limiting first best payment implies that the solutions to the centralized and decentralized control problems are identical as λ becomes large; that is,

$$\lim_{\lambda\to\infty} \sup_{\mu^\lambda_E \in \mathcal{S}(\vec{P}^\lambda, N^\lambda_{\text{ao}}, T^\lambda_{\text{ao}})} \frac{\sum_{i=1}^{N^\lambda_{\text{ao}}} U^\lambda_i + \mathcal{C}(\mu^\lambda_E, N^\lambda_{\text{ao}}, T^\lambda_{\text{ao}})}{c_S N^\lambda_\star + \mathcal{C}^\lambda_\star} = 1,$$

under Assumptions 1 and 2 because from Theorem 1

$$\lim_{\lambda\to\infty} \frac{c_S N^\lambda_{\text{ao}} + \mathcal{C}\big(\hat{\mu}_\star, N^\lambda_{\text{ao}}, \hat{T}_\star\big)}{\lambda} = \lim_{\lambda\to\infty} \frac{c_S N^\lambda_\star + \mathcal{C}^\lambda_\star}{\lambda} = c_S\hat{b}_\star + \hat{\mathcal{C}}(\hat{\mu}_\star, \hat{b}_\star, \hat{\beta}_\star, \hat{a}_\star),$$

and from Definition 4

$$\lim_{\lambda\to\infty} \sup_{\mu^\lambda_E \in \mathcal{S}(\vec{P}^\lambda, N^\lambda_{\text{ao}}, T^\lambda_{\text{ao}})} \frac{\sum_{i=1}^{N^\lambda_{\text{ao}}} U^\lambda_i + \mathcal{C}(\mu^\lambda_E, N^\lambda_{\text{ao}}, T^\lambda_{\text{ao}})}{\lambda} = \lim_{\lambda\to\infty} \frac{c_S N^\lambda_{\text{ao}} + \mathcal{C}\big(\hat{\mu}_\star, N^\lambda_{\text{ao}}, \hat{T}_\star\big)}{\lambda}.$$

We would like to develop a limiting first best policy from an asymptotically optimal policy $\big(\hat{\mu}_\star, N^\lambda_{\text{ao}} = \hat{b}_\star\lambda + o(\lambda), \hat{T}_\star\big)$. One option is to adapt the payment (11) to accommodate the fact that we see the realized number of services in a time unit, but not the long-run average. In fact, there can be many payment contracts that incentivize the servers to work at or near the rate $\hat{\mu}_\star$, and so are limiting first best. The payment contract we choose to analyze is piece-rate, because that payment seems the most natural to use in practice.

The piece-rate payment contract pays each server $P_S^\lambda > 0$ per completed service and penalizes each server $P_F^\lambda \geq 0$ for each unsuccessful service. Then, when the service rate vector is $\vec{\mu}$, recalling that server $i \in \{1, 2, \cdots, N^\lambda\}$ has utilization $B_i(\vec{\mu}^\lambda, N^\lambda, T^\lambda)$ defined in Lemma 1, the expected number of completed services for server i per unit time is $\mu_i B_i(\vec{\mu}^\lambda, N^\lambda, T^\lambda)$, with $(1 - p(\mu_i))\mu_i B_i(\vec{\mu}^\lambda, N^\lambda, T^\lambda)$ expected to fail. This results in the expected payment per time unit to server i

$$U_i^\lambda := \left(P_S^\lambda - P_F^\lambda(1 - p(\mu_i))\right)\mu_i \times B_i(\vec{\mu}^\lambda, N^\lambda, T^\lambda), \text{ for each } i \in \{1, \ldots, N^\lambda\}. \tag{21}$$

The piece-rate payment in (21) is in $\mathcal{P}$ because the realized number of completed and failed services in a time unit is observable.

Since we focus on symmetric equilibrium, it is sufficient to focus on the utility of a tagged server working at rate μ_1, when all the other servers work at rate μ. Without loss of generality, we let server 1 be this tagged server, and, in a slight abuse of notation, write

$$U^\lambda(\mu_1, \mu) = \left(P_S^\lambda - P_F^\lambda(1 - p(\mu_1))\right)\mu_1 \times B\big((\mu_1, \mu), N^\lambda, T^\lambda\big),$$

where $B\big((\mu_1, \mu), N^\lambda, T^\lambda\big)$ is the utilization of server 1 when all other servers work at rate μ, which has explicit expression that follows from Lemma 1. A symmetric equilibrium is a fixed point of the best response function

$$R^\lambda(\mu) := \underset{\mu_1 \in [\underline{\mu}, \overline{\mu}]}{\arg\max}\, U^\lambda(\mu_1, \mu). \tag{22}$$

that satisfies the individual rationality contraint (2).

Proposition 3. *Assume p is strictly decreasing and weakly concave on $\left[\underline{\mu}, \overline{\mu}\right]$. Any piece-rate payment (21) having payment ratio $P_R^\lambda := P_F^\lambda / P_S^\lambda$ results in the same non-empty set of fixed points for R^λ in (22). A fixed point μ_F^λ is a symmetric equilibrium if*

$$\left(P_S^\lambda - P_F^\lambda(1 - p(\mu_F^\lambda))\right)\mu_F^\lambda \times B(\mu_F^\lambda, N^\lambda, T^\lambda) \geq c_S. \tag{23}$$

We would like to show piece-rate payment is limiting first best. One way forward is to identify payment parameters P_S^λ and P_F^λ under which $\hat{\mu}_\star$ is a fixed point of R^λ that satisfies (23) for each λ, and so a symmetric equilibrium. However, that is difficult. Moreover, we would prefer the payment parameters to result in a unique equilibrium so that there is no selection issue, but Proposition 3 does not guarantee uniqueness. That motivates us to develop an approximation for $B\big((\mu_1,\mu),N^\lambda,T^\lambda\big)$ that is more analytically tractable. We also provide intuition for that approximation in Remark 2 at the end of this section.

Proposition 4. *Fix $b \ge 0$. Under IOB(T^λ) routing and staffing N^λ that satisfies (12), if $T^\lambda \to T < \infty$ as $\lambda \to \infty$, then for any $\mu_1, \mu \in [\underline{\mu}, \overline{\mu}]$,*

$$\lim_{\lambda\to\infty} B\big((\mu_1,\mu),N^\lambda,T^\lambda\big) = \hat{B}(\mu_1,\mu), \ \textit{where } \hat{B}(\mu_1,\mu) := \frac{\mu\exp(-\theta T)}{\mu\exp(-\theta T) + \mu_1\left[b\mu - \exp(-\theta T)\right]^+}.$$

Proposition 4 implies that we would like to find piece-rate payment parameters under which the approximating best response function

$$\hat{R}(\mu) := \arg\max_{\mu_1\in[\underline{\mu},\overline{\mu}]} \hat{U}(\mu_1,\mu), \ \text{ where } \hat{U}(\mu_1,\mu) = \left(P_S - P_F(1-p(\mu_1))\right)\mu_1 \times \hat{B}(\mu_1,\mu)$$

has any desired fixed point $\mu \in [\underline{\mu},\overline{\mu}]$ (and in particular, has fixed point $\hat{\mu}_\star$), which is unique.

Lemma 7. *Assume p is strictly decreasing and weakly concave on $[\underline{\mu},\overline{\mu}]$. Given $b>0$, $\mu\in[\underline{\mu},\overline{\mu}]$, and $T>0$, define*

$$P_R(b,\mu,T) := \frac{1}{1-p(\mu)-\mu p'(\mu)\max\{b\mu\exp(\theta T),1\}}.$$

If $P_S>0$ and $P_F\ge 0$ are such that $P_F/P_S = P_R$, then μ is the unique fixed point of $\hat{R}$.

We develop a limiting first best policy from an asymptotically optimal policy $\big(\hat{\mu}_\star, N^\lambda_{\text{ao}} = \hat{b}_\star\lambda + o(\lambda), \hat{T}_\star\big)$ in Theorem 1 as follows. We define $P_R^\star := P_R(\hat{b}_\star,\hat{\mu}_\star,\hat{T}_\star)$ as in Lemma 7, which, from the relationship $\hat{\beta}_\star = \min\left(1, \frac{\exp(-\theta\hat{T}_\star)}{\hat{b}_\star\hat{\mu}_\star}\right)$ is equivalently written as

$$P_R^\star = \frac{1}{1-p(\hat{\mu}_\star)-\hat{\mu}_\star p'(\hat{\mu}_\star)/\hat{\beta}_\star}. \tag{24}$$

Let $\mathcal{S}^\lambda(P_R^\star)$ be the set of fixed points of the best response function R^λ, which is not empty by Proposition 3. Set the payment parameter P_S^λ to ensure individual rationality is satisfied for any fixed point; that is,

$$P_S^\lambda := \sup_{\mu\in\mathcal{S}^\lambda(P_R^\star)} \frac{c_S}{\big(1-(1-p(\mu))P_R^\star\big)\mu B\big(\mu, N^\lambda_{\text{ao}}, \hat{T}_\star\big)}, \ \text{for any } N^\lambda_{\text{ao}} = \hat{b}_\star\lambda + o(\lambda). \tag{25}$$

The payment ratio $P_R^\star$ then determines the payment parameter

$$P_F^\lambda := P_S^\lambda P_R^\star. \tag{26}$$

Since P_S^λ and P_F^λ are such that (23) is satisfied for all fixed points, any $\mu_E^\lambda = \mu_F^\lambda(P_R^\lambda)$ is a symmetric equilibrium.

THEOREM 2. *The piece-rate payment $\vec{U}^\lambda = (U_1^\lambda, \ldots, U_N^\lambda)$ in (21) having parameters P_S^λ and P_F^λ defined by (25) and (26) from $P_R^\star$ in (24) is limiting first best.*

The payment parameters (25) and (26) are difficult to interpret. Instead, we set

$$P_S^\star := \lim_{\lambda\to\infty} P_S^\lambda \text{ and } P_F^\star := \lim_{\lambda\to\infty} P_F^\lambda = P_S^\star P_R^\star,$$

and use the limiting parameters to provide simpler expressions. We use those expressions in Example 2 below to see how the system manager transfers her costs to the servers. For this, we take the limit as $\lambda \to \infty$ in (21) to find

$$\left(P_S^\star - P_F^\star(1-p(\hat{\mu}_\star))\hat{\mu}_\star\right)\hat{\beta}_\star = c_S,$$

or, equivalently,

$$P_S^\star\left(1 - P_R^\star(1-p(\hat{\mu}_\star))\hat{\mu}_\star\right)\hat{\beta}_\star = c_S, \tag{27}$$

where we have used Theorems 1, 2, and the definition of limiting first best payment. From (24) and (27),

$$P_S^\star = \frac{-c_S}{\hat{\mu}_\star^2 p'(\hat{\mu}_\star)}\left(1 - p(\hat{\mu}_\star) - \hat{\mu}_\star p^{'}(\hat{\mu}_\star)/\hat{\beta}_\star\right), \tag{28}$$

which implies

$$P_F^\star = \frac{-c_S}{\hat{\mu}_\star^2 p'(\hat{\mu}_\star)}. \tag{29}$$

EXAMPLE 2. Suppose each service failure costs c_F, so that $g_F(x) = c_F$ for $x \in [0,1]$. Further assume $\hat{\mu}_\star \in (\underline{\mu}, \overline{\mu})$ so that $\hat{\mu}_\star$ in (16) solves the first order condition

$$-\frac{\hat{c}_S(\hat{\beta}_\star)}{\hat{\mu}_\star^2} - c_F p^{'}(\hat{\mu}_\star) = 0,$$

and the minimum cost to serve a customer in (19) becomes

$$\hat{c}_\star = \frac{\hat{c}_S(\hat{\beta}_\star)}{\hat{\mu}_\star} + c_F(1 - p(\hat{\mu}_\star)) = c_F\big(1 - p(\hat{\mu}_\star) - \hat{\mu}_\star p^{'}(\hat{\mu}_\star)\big).$$

Substituting the above expressions into (28) and (29) yields

$$P_S^\star = \frac{c_S}{\hat{c}_S(\hat{\beta}_\star)} \frac{1 - p(\hat{\mu}_\star) - \hat{\mu}_\star p^{'}(\hat{\mu}_\star)/\hat{\beta}_\star}{1 - p(\hat{\mu}_\star) - \hat{\mu}_\star p^{'}(\hat{\mu}_\star)} \hat{c}_\star \text{ and } P_F^\star = \frac{c_S}{\hat{c}_S(\hat{\beta}_\star)} c_F. \tag{30}$$

The above expressions show that the system manager can transfer her costs to the servers in a way that induces the servers to work at rate $\hat{\mu}_\star$ (in the limit), as we explain in the two cases below.

Case 1: $g_U'(1) - g_U(1) \le c_S$. Then, $\hat{\beta}_\star = 1$ from Proposition 2, and so from (30),

$$P_S^\star = \frac{c_S}{c_S + g_U(1)} \hat{c}_\star, \quad P_F^\star = \frac{c_S}{c_S + g_U(1)} c_F.$$

If there is no utilization cost ($g_U(\cdot) = 0$), then each server is paid the minimum cost to serve a customer for each service, $P_S^\star = \hat{c}_\star$, and is penalized the cost of service failure for any failed service, $P_F^\star = c_F$. Otherwise, when $g_U(\cdot) > 0$, since only the manager experiences the utilization cost, paying $\hat{c}_\star$ for each service would lead to an expected payment higher than c_S. To overcome this, the manager reduces the payments by identical fractions, so that $P_R^\star$, and therefore the service rate is unchanged (remains $\hat{\mu}^\star$), but the individual rationality constraint continues to be satisfied with equality.

Case 2: $g_U'(1) - g_U(1) > c_S$. Then, $\hat{\beta}_\star < 1$ from Proposition 2. The servers have idle time and do not experience utilization costs and, therefore, prefer to work slower than the system manager desires in order to avoid the service failure costs, thereby increasing their utilization to one. Hence in (30) the system manager increases their payment for service completion by the factor $\frac{1 - p(\hat{\mu}_\star) - \hat{\mu}_\star p^{'}(\hat{\mu}_\star)/\hat{\beta}_\star}{1 - p(\hat{\mu}_\star) - \hat{\mu}_\star p^{'}(\hat{\mu}_\star)} > 1$ to encourage them to complete more services. That factor is the ratio between the $P_R^\star$ in (24) under $\hat{\beta}_\star$ and the $P_R^\star$ in (24) when the servers have their preferred utilization of one.

The cost transfer in Example 2 is possible because the costs that affect the service rate decisions can be assigned (proportionally) to an individual server. This is not true for general, instead of linear, service failure cost functions. Then, the cost transfer is much more complicated. Neither the pre-limit payment parameters, P_S^λ and P_F^λ, nor the limiting payment parameters, $P_S^\star$ and $P_F^\star$, give rise to a clear intuition.

REMARK 2. When $b\mu \leq \exp(-\theta T)$, the system is overloaded, and so $\hat{B}(\mu_1, \mu) = 1$ for all $\mu_1 \in [\underline{\mu}, \overline{\mu}]$. Otherwise, when $b\mu > \exp(-\theta T)$, the system is underloaded, and the intuition for the approximation in Proposition 4 is less straightforward to see. When the system is underloaded, each server experiences alternative busy periods and idle periods. Due to the IOB routing, each server who finishes a service has to join a queue of idle servers to get the next customer and each idle period should have the same expected length. Moreover, since all the servers besides server 1 are working at μ, the average busy period length is $1/\mu$, and the utilization is about $\exp(-\theta T)/(b\mu) < 1$. We can solve for an approximate idle period length x in

$$\frac{1/\mu}{1/\mu + x} \approx \frac{\exp(-\theta T)}{b\mu}.$$

In particular, the average idle period length is approximately $b\exp(\theta T) - 1/\mu$. When server 1 works at μ_1, server 1 has the average busy period length $1/\mu_1$, so that his busy time percentage is

$$\frac{\text{Busy period length}}{\text{Busy period length} + \text{Idle period length}} \approx \frac{1/\mu_1}{1/\mu_1 + b\exp(\theta T) - 1/\mu} = \frac{\mu \exp(-\theta T)}{\mu \exp(-\theta T) + \mu_1 \left(b\mu - \exp(-\theta T)\right)}.$$

6. Concluding Remarks

The rate at which a server works is influenced by the payment structure. That observation has important consequences for staffing decisions. This is because the rate at which servers work is a first-order determinant of the staffing level required to handle a given arrival rate. The delicacy is that the staffing level, as well as the routing, can also affect the service rate. This motivates solving a joint optimization problem (i.e., a centralized control problem) to determine the staffing level, routing, and desired service rate. Depending on the system manager's cost structure, the solution to that joint optimization problem dictates whether the system manager should allow servers to idle and/or allow customers to abandon – and may require a routing policy that idles servers even when customers are waiting.

The system manager cannot control the service rates. The service rates arise as a Nash equilibrium solution to a game in which the servers each selfishly maximize their own utility, which is determined by their expected payment. Then, to achieve first best, the system manager must use a payment contract that ensures the servers work at her desired service rate. We developed a simple approximation for an equilibrium service rate under piece-rate payment in systems with large arrival rates, and used that approximation to show piece-rate payment is limiting first best. The reason to define limiting first best is that a first best payment contract requires knowledge of the solution to the centralized control problem, and the centralized control problem must be solved in the limit to find an analytic solution.

We provide conditions under which any solution to the centralized control problem has all servers working at the same rate. This would not be the case if the servers were heterogeneous in their ability levels, in which case the speed-quality trade-off function p would not be the same for all servers. Such server heterogeneity requires expanding the scope of this work to include non-symmetric equilibria.

We assume a routing policy that is based on the idle time of each server, show that any policy in this class has the same steady state probabilities as the common longest-idle-server-first rule, and allow a generalization whose purpose is to reduce the server utilization (meaning the routing policy may or may not be work conserving). All policies in this class assume customers queue in a single line and have exactly one service. Both assumptions merit further thought. First, a failed service could also mean one in which the service must be repeated. This brings into play much more complicated routing questions, such as whether or not the server responsible for the failed service should be the one to re-do the service, as studied in Lu et al. (2009) in a system with two servers. Next, the recent empiric work of Song et al. (2015) and Shunko et al. (2018) has shown that separating customers into multiple lines (i.e., "unpooling" the system) affects server behavior and can improve performance. One explanation is that servers who are responsible only for their own line have more incentive to change their service rate based on line length. This motivates future work to answer the question "to pool or not to pool", as studied in Armony et al. 2016 in a system with two servers.

Our current cost structure can be extended in two directions. On the customer side, a customer may prefer being turned away before waiting rather than abandoning after waiting (which motivated the problem formulation in Ward and Kumar 2008). This motivates adding the option of admission control, and charging a smaller cost for customers turned away in comparison with those that abandon. An admission control policy that admits each arrival with independent probability $exp(-\theta T)$ has the same customer abandonment probability as IOB(T) routing, but lower cost. On the server side, a server may prefer to have longer and less frequent breaks rather than frequent short idle periods, given the same utilization. This could be encapsulated either by modifying the utilization cost, or by changing the server utility function to reflect a non-monetary value on breaks. The technical challenge is that the resulting effect on customer wait time is complicated and potentially changes the abandonment probability. See Sun and Whitt (2018) for work in this direction in a many-server model *without* abandonment and *with* exogeneous (not endogenous) service rates, in which breaks occur randomly over time.

Another interesting direction for future research is to incorporate time-varying arrival rates or arrival rate uncertainty. There is a large literature that studies how to staff such systems (e.g., Bassamboo et al. (2010), He et al. (2016)), Liu (2018)). However, most all such papers assume fixed service rates. We wonder whether the server payments could be used to leverage the staffing decisions, by inducing the servers to speed up or slow down, depending on the arrival rate.

Finally, in our model, we assume that the number of completed services in any finite time interval can be observed, but that the long-run average service rate must be estimated. This is straightforward to handle because the servers are risk neutral, and make decisions based on their expected payment in a time unit. However, a more natural assumption is that servers are risk averse, which would require more careful consideration of the payment structure.

Endnotes

1. From Lemma 3, we know that $\frac{N^\lambda}{\lambda}$ must have a converge subsequence $\frac{N^{\lambda_i}}{\lambda_i}$. Assume the subsequence converges to b. Then we can focus on this subsequence and have $N^{\lambda_i} = b\lambda_i + o(\lambda_i)$. Therefore, without loss of generality, we need only search within the staffing policy in (12).

Acknowledgments

We would like to thank our anonymous Associate Editor and Referees for their excellent comments, which considerably improved our paper. We would like to thank Sampath Rajagopalan for helpful suggestions during the early development of the paper, Itai Ashlagi for constructive comments following an early presentation of this work, and Raga Gopalakrishnan for helpful discussions on the symmetric equilibrium existence proof. This work is supported by the William S. Fishman Faculty Research Fund at the University of Chicago Booth School of Business.

References

Afeche, P. 2013. Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing Service Oper. Management* **15**(3) 423–443.

Afeche, P., J.M. Pavlin. 2016. Optimal price/lead-time menus for queues with customer choice: Segmentation, pooling, and strategic delay. *Management Sci.* **62**(8) 2412?2436.

Akerlof, G. 1970. The market for “lemons”: Quality uncertainty and the market mechanism. *Quart. J. Econom.* **84**(3) 488–500.

Alizamir, S., F. de Vericourt, P. Sun. 2013. Diagnostic accuracy under congestion. *Management Sci.* **59**(1) 151–171.

Allon, G., A. Bassamboo, E. Cil. 2017. Skill management in large-scale service marketplaces. *Prod. Oper. Management* **26**(11) 2050–2070.

Allon, G., I. Gurvich. 2010. Pricing and dimensioning competing large-scale service providers. *Manufacturing Service Oper. Management* **12**(3) 449–469.

Anand, K., M. Pac, S. Veeraraghavan. 2011. Quality-speed conundrum: Tradeoffs in customer-intensive services. *Management Sci.* **57**(1) 40–56.

Ancker, C.J., A. Gafarian. 1962. Queueing with impatient customers who leave at random. *J. of Industrial Engineering* **13** 84–90.

Armony, M., I. Gurvich. 2010. When promotions meet operations: Cross-selling and its effect on call center performance. *Manufacturing Service Oper. Management* **12**(3) 470–488.

Armony, M., C. Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Oper. Res.* **52**(4) 527–545.

Armony, M., G. Roels, H. Song. 2016. Pooling queues with discretionary service capacity. Working Paper.

Bassamboo, A., R. Randhawa. 2010. On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Oper. Res.* **58**(5) 1398–1413.

Bassamboo, A., R. Randhawa, A. Zeevi. 2010. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Sci.* **56**(10) 1668–1686.

Borst, S., A. Mandelbaum, M.I. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* **52**(1) 17–34.

Buell, R.W., T. Kim, C.J. Tsay. 2017. Creating reciprocal value through operational transparency. *Management Sci.* **63**(6) 1673–1695.

Bureau of Economic Analysis. 2017. GDP and the economy third estimates for the third quarter of 2016. Retrieved March 2017, https://bea.gov/scb/pdf/2017/01%20January/0117_gdp_and_the_economy.pdf.

Cachon, G.P., P.T Harker. 2002. Competition and outsourcing with scale economies. *Management Sci.* **48**(10) 1314–1333.

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

Chan, C.W., G. Yom-Tov, G. Escobar. 2014. When to use speedup: An examination of service systems with returns. *Oper. Res.* **62**(2) 462–482.

Debo, L.G, L.B. Toktay, L.N. Van Wassenhove. 2008. Queuing for expert services. *Management Sci.* **54**(8) 1497–1512.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **54**(3) 208–227.

Geng, X., W.T. Huh, M. Nagarajan. 2015. Fairness among servers when capacity decisions are endogenous. *Prod. Oper. Management* **24**(6) 961–974.

Gilbert, S.M., Z.K. Weng. 1998. Incentive effects favor nonconsolidating queues in a service system: The principal-agent perspective. *Management Sci.* **44**(12) 1662–1669.

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

Gurvich, I., M. Lariviere, A. Moreno-Garcia. 2018. Operations in the on-demand economy: Staffing services with self-scheduling capacity. M. Hu, ed., *Sharing Economy: Making Supply Meet Demand*. Springer Series in Supply Chain Management.

Hasija, S., E. Pinker, R. Shumsky. 2010. Work expands to fill the time available: Capacity estimation and staffing under Parkinson's law. *Manufacturing Service Oper. Management* **12**(1) 1–18.

Hassin, R. 2016. *Rational Queueing*. CRC Press, Boca Raton, FL.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA.

He, B., Y. Liu, W. Whitt. 2016. Staffing a service system with non-poisson non-stationary arrivals. *Probab. in the Engineering and Informational Sciences* **30**(4) 593–621.

Holmstrom, B. 1979. Moral hazard and observability. *Bell J. Econom.* **10**(1) 74–91.

Holmstrom, B. 1982. Moral hazard in teams. *Bell J. Econom.* **13**(2) 324–340.

Hopp, W., S. Iravani, G. Yuen. 2007. Operations systems with discretionary task completion. *Management Sci.* **53**(1) 61–77.

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

Kalai, E., M. I. Kamien, M. Rubinovitch. 1992. Optimal service speeds in a competitive environment. *Management Sci.* **38**(8) 1154–1163.

Kostami, V., S. Rajagopalan. 2014. Speed-quality trade-offs in a dynamic model. *Manufacturing Service Oper. Management* **16**(1) 104–118.

Liu, Y. 2018. Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Oper. Res.* **66**(2) 514–534.

Lovejoy, W.S., K. Sethuraman. 2000. Congestion and complexity costs in a plant with fixed resources that strives to make schedule. *Manufacturing Service Oper. Management* **2**(3) 221–239.

Lu, L., J. Van Mieghem, C. Savaskan. 2009. Incentives for quality through endogenous routing. *Manufacturing Service Oper. Management* **11**(2) 254–273.

Maglaras, C., J. Yao, A. Zeevi. 2018. Optimal price and delay differentiation in large-scale queueing systems. *Management Sci.* **64**(5) 2427–2444.

Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Sci.* **49**(8) 1018–1038.

Maglaras, C., A. Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* **53**(2) 242–262.

Mehrotra, V., K. Ross, G. Ryder, Y. Zhou. 2012. Routing to manage resolution and waiting time in call centers with heterogeneous servers. *Manufacturing Service Oper. Management* **14**(1) 66–81.

Milkovich, G.T., J.M. Newman. 2004. *Compensation (8th ed.).* McGraw-Hill/Irwin, Boca Raton, FL.

Pagurova. 1965. An asymptotic for the incomplete Gamma function. *Zh. Vychisl. Mat. Mat. Fiz.* **5**(1) 118–121.

Ren, Z.J., Y.P. Zhou. 2008. Call center outsourcing: Coordinating staffing level and service quality. *Management Sci.* **54**(2) 369–383.

Rothschild, M., J. Stiglitz. 1976. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quart. J. Econom.* **90**(4) 629–649.

Shunko, M., J. Niederhoff, Y. Rosokha. 2018. Humans are not machines: Impact of queueing design on service time. *Management Sci.* **64**(1) 453–473.

Song, H., A. Tucker, K. Murrell. 2015. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Sci.* **61**(12) 3032–3053.

Spence, M. 1973. Job market signaling. *Quart. J. Econom.* **87**(3) 355–374.

Sun, X., W. Whitt. 2018. Creating work breaks from available idleness. To appear in *Manufacturing Service Oper. Management.*

Ward, A.R., S. Kumar. 2008. Asymptotically optimal admission control of a queue with impatient customers. *Math. Oper. Res.* **33**(1) 167–202.

Whitt, W. 2006. Staffing a call center with uncertain arrival rate and absenteeism. *Prod. Oper. Management* **15**(1) 88–102.

Zhan, D., A.R. Ward. 2014. Routing to minimize waiting and callbacks in large call centers. *Manufacturing Service Oper. Management* **16**(2) 220–237.

Zhan, D., A.R. Ward. 2018. The M/M/1+M queue with a utility-maximizing server. *Oper. Res. Letters* **46**(5) 518–522.

Staffing, Routing and Payment to Trade Off Speed and Quality in Large Service Systems: Electronic Companion

In this e-companion, we first list the notations in the main body in Table EC.1, and then give the proofs for all the lemmas, propositions and theorems, both in the order of their appearance in the main body.

Table EC.1 Notation Table

Symbol	Definition
N	The number of servers
λ	Customer arrival rate, the independent parameter that becomes large
θ	The impatience parameter
μ	The service rate
$\mathcal{I}(t)$	The set of idle servers at time t
$\boldsymbol{s}(t)$	The vector of idle servers at time t, ordered by how long they have been idle
T	The holding time parameter of IOB routing
$p(\mu)$	The probability of service success
$\underline{\mu}, \overline{\mu}$	The lower and upper bounds of the service rate
$\vec{\mu}, \mu_i$	The N-dimensional service rate vector, the service rate of sever i
$\vec{U}$	The expected steady state payment vector
$\vec{P}$	The payment contract vector
c_S	The payment per unit time of an outside employment option
$\beta_i = B_i(\vec{\mu}, N, T)$	The steady state utilization of server i under IOB(T)
q_A	The steady state customer abandonment probability
q_F	The steady state probability of service failure
$g_U(\beta_i)$	The utilization cost of server i, as a function of his steady state utilization

$g_A(q_A)$	The cost per abandonment, as a function of steady state abandonment percentage
$g_F(q_F)$	The cost per failed service, as a function of steady state service failure percentage
$\mathcal{C}(\mu, N, T)$	The system manager's expected operational cost per time unit in steady state
$\mathcal{P}$	The class of implementable payment contracts
μ_E	A symmetric equilibrium service rate
$\mathcal{S}(\vec{P}, N, T)$	The set of symmetric equilibrium service rate vectors
$(\vec{\mu}_\star, N_\star, T_\star)$	The optimal service rate vector, staffing, and routing parameter solving (7)
$\mathcal{C}_\star$	The minimum cost of (7)
$\mu_\star$	The common optimal service rate of each server under Assumption 1
$(\mu^\lambda, N^\lambda, T^\lambda)$	A policy superscripted by the system arrival rate λ
b	The staffing level parameter
$\hat{\mathcal{C}}(\mu, b, \beta, a)$	The limiting objective function
$\hat{c}_S(\beta)$	The limiting adjusted staffing cost that accounts for payment and utilization cost
$\hat{\beta}_\star, \hat{\mu}_\star, \hat{a}_\star, \hat{b}_\star$	The limiting optimal utilization, service rate, abandonment probability, staffing level
$\hat{T}_\star, \hat{c}_\star$	The limiting optimal routing parameter, cost of serving a customer
$(\hat{\mu}_\star, N^\lambda_{\text{ao}}, \hat{T}_\star)$	An asymptotically optimal policy
P^λ_S, P^λ_F	The payment per completed service, penalty per each failed service
$B\big((\mu_1, \mu), N, T\big)$	The steady state utilization of server 1 working at μ_1, when all others work at μ
$U^\lambda(\mu_1, \mu)$	The steady state utility of server 1, when all others work at μ
$R^\lambda(\mu)$	The best response function of server 1, when all others work at μ
$\hat{B}(\mu_1, \mu), \hat{U}(\mu_1, \mu)$	The approximate steady state utilization, utility of server 1
$\hat{R}(\mu)$	The approximate best response function of server 1 when others work at μ
$P^\star_R$	The limiting first best payment ratio between penalty and reward
$P^\star_S, P^\star_F$	The limit of the limiting first best payment per service, penalty per failure

A. Proof of Lemma 1

We define $\mathcal{N} := \{1,\dots,N\}$. Each arriving customer independently abandons from the holding area with probability $1-\exp(-\theta T)$. As a result, the effective arrival rate to the queue is a thinned Poisson process with rate $a = \lambda \exp(-\theta T)$. The queue state space is as follows.

1. State B is the state where all servers are busy and no customer in queue;
2. State $\mathbf{s} = (s_1, s_2, \cdots, s_{|\mathcal{I}|})$ is the ordered vector of non-empty set of idle servers $\mathcal{I} \subseteq \mathcal{N}$;
3. State m ($m \geq 1$) is the state where all servers are busy and m jobs are in the queue, in which case the queue behaves as an $M/M/1+M$ with service rate $\mu_\Sigma = \sum_{j=1}^N \mu_j$.

From the local balance equations, the associated steady state probabilities must satisfy

$$\pi_m = \pi_B \prod_{k=1}^m \frac{a}{\mu_\Sigma + k\theta}, m \geq 1.$$

$$\pi_{\mathbf{s}} = \pi_B \prod_{j\in\mathcal{I}} \frac{\mu_j}{a}, \text{ for all } \mathbf{s} = (s_1, s_2, \cdots, s_{|\mathcal{I}|}), \text{ if } |\mathcal{I}| > 0, \text{ and } \pi_{\mathbf{s}} = \pi_B, \text{ if } \mathcal{I} = \emptyset,$$

and this can be verified exactly as in the proof of Theorem 9 in Gopalakrishnan et al. (2016). Since for any set $\mathcal{I}$ of idle servers, all $|\mathcal{I}|!$ permutations of ordered vectors $(s_1, s_2, \cdots, s_{|\mathcal{I}|})$ have identical steady state probabilities, the normalization condition that the steady state probabilities sum to one gives

$$\pi_B = \frac{1}{\sum_{\mathcal{I}\subseteq\mathcal{N}} |\mathcal{I}|! \prod_{j\in\mathcal{I}} \frac{\mu_j}{a} + z}, \text{ where } z := \sum_{m=1}^\infty \prod_{k=1}^m \frac{a}{\mu_\Sigma + k\theta}.$$

When server i is busy, the state $\mathbf{s}$ is a subset $\mathcal{I} \in \mathcal{N}\backslash i$, and so the probability server i is busy is

$$\pi_B \sum_{\mathcal{I}\subseteq\mathcal{N}\backslash i} |\mathcal{I}|! \prod_{j\in\mathcal{I}} \frac{\mu_j}{a} + \sum_{i=1}^\infty \pi_m,$$

which results in $B_i(\vec{\mu}, N, T)$ defined as in the statement of Lemma 1. □

B. Proof of Lemma 2

Without loss of generality, we want to show $B_1(\vec{\mu}, N, T)$ defined in Lemma 1 is strictly decreasing in T. To do this, we regard $B_1(\vec{\mu}, N, T)$ as a function of $x(T) = \frac{1}{\lambda \exp(-\theta T)}$, and show B_1 is strictly decreasing in x. Recall $\mathcal{N} := \{1, \cdots, N\}$, and define

$$\breve{A}(x) := \sum_{\emptyset\neq\mathcal{I}\subseteq\mathcal{N}\backslash 1} x^{|\mathcal{I}|} |\mathcal{I}|! \prod_{j\in\mathcal{I}} \mu_j, \breve{B}(x) := \sum_{\emptyset\neq\mathcal{I}\subseteq\mathcal{N}} x^{|\mathcal{I}|} |\mathcal{I}|! \prod_{j\in\mathcal{I}} \mu_j, \text{ and } z(x) := \sum_{m=1}^\infty \prod_{k=1}^m \frac{1}{(\mu_\Sigma + k\theta)x},$$

so that

$$B_1(x) = \frac{1 + \breve{A}(x) + z(x)}{1 + \breve{B}(x) + z(x)} \in [0,1] \text{ for } x > 0.$$

To show $B_1(x)$ is strictly decreasing in x, it is equivalent to show

$$1 - B_1(x) = \frac{\breve{B}(x) - \breve{A}(x)}{1 + \breve{B}(x) + z(x)} = \frac{1 - \frac{\breve{A}(x)}{\breve{B}(x)}}{1 + \frac{1+z(x)}{\breve{B}(x)}}$$

is strictly increasing in x. For that, it is sufficient to show the denominator in the above display is strictly decreasing in x and the numerator is strictly increasing in x. The denominator is strictly decreasing because $\breve{B}(x)$ is strictly increasing and $z(x)$ is strictly decreasing, so that $\frac{1+z(x)}{\breve{B}(x)}$ is strictly decreasing in x on $(0,\infty)$. To show the numerator is strictly increasing, it is sufficient to show that $\frac{d}{dx}\left(\breve{A}(x)/\breve{B}(x)\right) < 0$ for $x \in (0,\infty)$, or that

$$\breve{A}^{'}(x)\breve{B}(x) < \breve{A}(x)\breve{B}^{'}(x). \tag{EC.1}$$

To show (EC.1), it is helpful to define

$$c(i,1) := i! \sum_{|\mathcal{I}|=i, \mathcal{I} \subseteq \mathcal{N}\setminus 1} \prod_{j \in \mathcal{I}} \mu_j, \text{ for } i \in \mathcal{N}\setminus N, \quad c(i) := i! \sum_{|\mathcal{I}|=i, \mathcal{I} \subseteq \mathcal{N}} \prod_{j \in \mathcal{I}} \mu_j, \text{ for } j \in \mathcal{N}.$$

so that

$$\breve{A}(x) = \sum_{i=1}^{N-1} c(i,1)x^i, \quad \breve{B}(x) = \sum_{i=1}^{N} c(i)x^i.$$

Then, (EC.1) is equivalent to

$$\left(\sum_{i=1}^{N-1} ic(i,1)x^{i-1}\right)\left(\sum_{j=1}^{N} c(j)x^j\right) < \left(\sum_{i=1}^{N-1} c(i,1)x^i\right)\left(\sum_{j=1}^{N} jc(j)x^{j-1}\right),$$

or

$$\sum_{k=1}^{2N-2} \sum_{i+j=k+1, i \in \mathcal{N}\setminus N, j \in \mathcal{N}} ic(i,1)c(j)x^k < \sum_{k=1}^{2N-2} \sum_{i+j=k+1, i \in \mathcal{N}\setminus N, j \in \mathcal{N}} jc(i,1)c(j)x^k.$$

Since $x > 0$, to show (EC.1) and complete the proof, it is sufficient to show that

$$\sum_{i+j=k+1, i \in \mathcal{N}\setminus N, j \in \mathcal{N}} ic(i,1)c(j) < \sum_{i+j=k+1, i \in \mathcal{N}\setminus N, j \in \mathcal{N}} jc(i,1)c(j),$$

or

$$\sum_{i+j=k+1, i \in \mathcal{N}\setminus N, j \in \mathcal{N}} (i-j)c(i,1)c(j) < 0, \text{ for all } k \in \{1, 2, \cdots, 2N-2\}. \tag{EC.2}$$

To show (EC.2), we first observe

$$\begin{aligned}c(j) &= j!\left(\sum_{|\mathcal{I}|=j,\mathcal{I}\subseteq\mathcal{N},1\in\mathcal{I}}\prod_{j\in\mathcal{I}}\mu_j + \sum_{|\mathcal{I}|=j,\mathcal{I}\subseteq\mathcal{N},1\notin\mathcal{I}}\prod_{j\in\mathcal{I}}\mu_j\right)\\ &= j!\left(\sum_{|\mathcal{I}|=j-1,\mathcal{I}\subseteq\mathcal{N}\backslash 1}\mu_1\prod_{j\in\mathcal{I}}\mu_j + \sum_{|\mathcal{I}|=j,\mathcal{I}\subseteq\mathcal{N}\backslash 1}\prod_{j\in\mathcal{I}}\mu_j\right)\\ &= \mu_1 jc(j-1,1) + c(j,1), \text{ for } j\in\mathcal{N}\backslash\{1,N\},\end{aligned}$$

which can be extended to $j\in\mathcal{N}$ by defining $c(0,1):=1, c(N,1):=0$. Then, the left-hand side of (EC.2) satisfies

$$\sum_{i+j=k+1,i\in\mathcal{N}\backslash N,j\in\mathcal{N}}(i-j)c(i,1)c(j) = \sigma_1 + \mu_1\sigma_2, \tag{EC.3}$$

where

$$\sigma_1 := \sum_{i+j=k+1,i\in\mathcal{N}\backslash N,j\in\mathcal{N}}(i-j)c(i,1)c(j,1), \quad \sigma_2 := \sum_{i+j=k+1,i\in\mathcal{N}\backslash N,j\in\mathcal{N}}(i-j)jc(i,1)c(j-1,1).$$

Since $c(N,1)=0$, by adding zero terms,

$$\sigma_1 = \sum_{i+j=k+1,i\in\mathcal{N},j\in\mathcal{N}}(i-j)c(i,1)c(j,1).$$

Since i and j are symmetric, we can label i by j and label j by i without changing the summation; i.e.,

$$\sum_{i+j=k+1,i\in\mathcal{N},j\in\mathcal{N}}(i-j)c(i,1)c(j,1) = \sum_{j+i=k+1,i\in\mathcal{N},j\in\mathcal{N}}(j-i)c(j,1)c(i,1).$$

Summing both sides gives

$$2\sigma_1 = \sum_{i+j=k+1,i\in\mathcal{N},j\in\mathcal{N}}(i-j)c(i,1)c(j,1) + (j-i)c(j,1)c(i,1) = 0,$$

and so $\sigma_1 = 0$. Then, to show (EC.3) is negative and complete the proof, we need only show $\sigma_2 < 0$. Since $(i-j)jc(i,1)c(j-1,1) = 0$ when $j=0$, by removing these zero terms,

$$\sigma_2 = \sum_{i+j=k+1,i\in\mathcal{N}\backslash N,j\in\mathcal{N}\backslash 1}(i-j)jc(i,1)c(j-1,1).$$

Substituting $j^{'} = j-1$ yields

$$\sigma_2 = \sum_{i+j^{'}=k,i\in\mathcal{N}\backslash N,j^{'}\in\mathcal{N}\backslash N}(i-j^{'}-1)(j^{'}+1)c(i,1)c(j^{'},1).$$

Since i and $j^{'}$ are symmetric, we can label i by $j^{'}$ and label $j^{'}$ by i without changing the summation; i.e.,

$$\sum_{i+j^{'}=k, i\in\mathcal{N}\setminus N, j^{'}\in\mathcal{N}\setminus N} (i-j^{'}-1)(j^{'}+1)c(i,1)c(j^{'},1) = \sum_{j^{'}+i=k, j^{'}\in\mathcal{N}\setminus N, i\in\mathcal{N}\setminus N} (j^{'}-i-1)(i+1)c(j^{'},1)c(i,1).$$

Summing both sides gives

$$\begin{aligned} 2\sigma_2 &= \sum_{i+j^{'}=k, i\in\mathcal{N}\setminus N, j^{'}\in\mathcal{N}\setminus N} (i-j^{'}-1)(j^{'}+1)c(i,1)c(j^{'},1) + (j^{'}-i-1)(i+1)c(j^{'},1)c(i,1) \\ &= \sum_{i+j^{'}=k, i\in\mathcal{N}\setminus N, j^{'}\in\mathcal{N}\setminus N} (-(i-j^{'})^2 - i - j^{'} - 2)c(i,1)c(j^{'},1) < 0. \end{aligned}$$

and so, $\sigma_2 < 0$. □

C. Proof of Proposition 1

Suppose $(\vec{\mu} = (\mu_1, \cdots, \mu_N), N, T)$ solves the centralized control problem (7), where we have dropped the $\star$ subscript to simplify notation in this proof. Then, $N < \infty$ because otherwise the objective function in (7) is infinite. Furthermore, if $N = 0$, the service rate vector has no components, and, if $N = 1$, the service rate vector has one component; in either case, the proposition is trivially valid. Therefore, we assume $N \geq 2$. If $T = \infty$, whenever $N = 0$, $\beta_i = 0$ for all $i \in \mathcal{N}$ and $\mathcal{C}(\vec{\mu}, N, \infty) = \lambda g_A(1)$ for any service rate vector $\vec{\mu}$, which implies the objective function in (7) is minimized at $N = 0$. Hence we assume $T < \infty$.

The proof is by contradiction. Suppose $\mu_i \neq \mu_j$ for some $(i,j) \in \mathcal{N} \times \mathcal{N}$. Set

$$\mu := \frac{1}{N}\sum_{i=1}^{N} \mu_i \text{ and } \beta := \frac{1}{N\mu}\sum_{i=1}^{N} \mu_i B_i(\vec{\mu}, N, T),$$

When the service rate vector $\vec{\mu}$ has all components identical and equal to μ, from Lemma 1,

$$B(\mu, N, T_0) = \frac{\sum_{i=0}^{N-1}\left(\frac{\lambda\exp(-\theta T_0)}{\mu}\right)^{i+1}\frac{1}{Ni!} + \left(\frac{\lambda\exp(-\theta T_0)}{\mu}\right)^{N}\frac{1}{N!}\sum_{i=1}^{\infty}\prod_{k=1}^{i}\frac{\lambda\exp(-\theta T_0)}{N\mu+k\theta}}{\sum_{i=0}^{N}\left(\frac{\lambda\exp(-\theta T_0)}{\mu}\right)^{i}\frac{1}{i!} + \left(\frac{\lambda\exp(-\theta T_0)}{\mu}\right)^{N}\frac{1}{N!}\sum_{i=1}^{\infty}\prod_{k=1}^{i}\frac{\lambda\exp(-\theta T_0)}{N\mu+k\theta}}. \tag{EC.4}$$

Suppose we can show the following result.

LEMMA EC.1. *There exists $T_1 \geq 0$ such that $B(\mu, N, T_1) = \beta$.*

Then, to find the contradiction, it is enough to show

$$\mathcal{C}(\vec{\mu},N,T) > \mathcal{C}(\mu,N,T_1), \tag{EC.5}$$

where $\mathcal{C}(\vec{\mu},N,T)$ and $\mathcal{C}(\mu,N,T_1)$ are as in (6) and (8) respectively (recalling the slight abuse of notation that replaces $\vec{\mu}$ with μ when all vector components are identical). To show (EC.5), we do a term-by-term comparison.

(a) **The utilization cost.** From Lemma 1, if $\mu_i < \mu_j$, then $B_i(\vec{\mu},N,T) > B_j(\vec{\mu},N,T)$, which implies the weighted average is smaller than the unweighted average; i.e.,

$$\sum_{i=1}^{N}\left(\frac{B_i(\vec{\mu},N,T)}{\sum_{j=1}^{N} B_j(\vec{\mu},N,T)}\right)\mu_i < \frac{\sum_{i=1}^{N}\mu_i}{N}.$$

Hence from the definitions of β and μ,

$$\beta = \frac{\sum_{i=1}^{N} B_i\left(\vec{\mu},N,T\right)\mu_i}{\sum_{i=1}^{N}\mu_i} < \frac{\sum_{j=1}^{N} B_j\left(\vec{\mu},N,T\right)}{N}.$$

Since $g_U(\beta)$ is weakly convex and weakly increasingly in β

$$N g_U(\beta) \le N g_U\left(\frac{\sum_{j=1}^{N} B_j\left(\vec{\mu},N,T\right)}{N}\right) \le \sum_{j=1}^{N} g_U(B_j\left(\vec{\mu},N,T\right)). \tag{EC.6}$$

(b) **The abandonment cost.** The definitions of β and μ imply

$$\lambda - N\beta\mu = \lambda - \sum_{i=1}^{N} B_i\left(\vec{\mu},N,T\right)\mu_i, \tag{EC.7}$$

and so the second term in (6) exactly equals the second term in (8).

(c) **The service failure cost.** Suppose we can show the following result.

Lemma EC.2. *If $\mu_i > \mu_j$, then $B_i(\vec{\mu},N,T)\mu_i > B_j(\vec{\mu},N,T)\mu_j$, for $i,j\in\mathcal{N}$.*

From Lemma EC.2, the weighted average is larger than the unweighted average; i.e.,

$$\sum_{i=1}^{N}\frac{\mu_i B_i(\vec{\mu},N,T)}{\sum_{j=1}^{N}\mu_j B_j(\vec{\mu},N,T)}\mu_i > \frac{\sum_{i=1}^{N}\mu_i}{N} = \mu.$$

Since $p(x)$ is weakly concave and strictly decreasing in x,

$$\sum_{i=1}^{N}\frac{\mu_i B_i(\vec{\mu},N,T)}{\sum_{j=1}^{N}\mu_j B_j(\vec{\mu},N,T)}p(\mu_i) \le p\left(\sum_{i=1}^{N}\frac{\mu_i B_i(\vec{\mu},N,T)}{\sum_{j=1}^{N}\mu_j B_j(\vec{\mu},N,T)}\mu_i\right) < p(\mu),$$

which from the definitions of β and μ implies

$$\sum_{i=1}^{N} p(\mu_i)\mu_i B_i(\vec{\mu},N,T) < p(\mu)N\beta\mu, \text{ or } \sum_{i=1}^{N}(1-p(\mu_i))\mu_i B_i(\vec{\mu},N,T) > (1-p(\mu))N\beta\mu,$$

and also when combined with the fact that $g_F(x)$ is weakly increasing in x implies

$$g_F(1-p(\mu)) \le g_F\left(1-\frac{\sum_{i=1}^{N} p(\mu_i)\mu_i B_i\left(\vec{\mu},N,T\right)}{\sum_{i=1}^{N} B_i\left(\vec{\mu},N,T\right)\mu_i}\right).$$

The above two display together imply

$$N(1-p(\mu))\beta\mu g_F(1-p(\mu)) < \sum_{i=1}^{N}(1-p(\mu_i))\mu_i B_i\left(\vec{\mu},N,T\right) g_F\left(1-\frac{\sum_{i=1}^{N} p(\mu_i)\mu_i B_i\left(\vec{\mu},N,T\right)}{\sum_{i=1}^{N} B_i\left(\vec{\mu},N,T\right)\mu_i}\right). \tag{EC.8}$$

The contradiction (EC.5) follows from (EC.6), (EC.7), and (EC.8).

To complete the proof, we must establish Lemmas EC.1 and EC.2, which is done below.

Proof of Lemma EC.1. Recall from the first paragraph of this proof that $(\vec{\mu},N,T)$ is an assumed solution to (7) with $N \ge 2$ and $T < \infty$. Also recall the definitions of μ and β. We must show there exists $T_1 \ge 0$ such that

$$B(\mu,N,T_1) = \beta.$$

From Lemma 2, $B(\mu,N,T_1)$ is strictly decreasing in T_1. From (EC.4), $B(\mu,N,T)$ is continuous and converges to 0 (the numerator converges to 0 and the denominator converges to 1) as $T \to \infty$. Then, it is sufficient to show $B(\mu,N,T) \ge \beta$, or, equivalently,

$$\lambda - NB(\mu,N,T)\mu \le \lambda - \sum_{i=1}^{N}\mu_i B_i(\vec{\mu},N,T); \tag{EC.9}$$

that is, the abandonment rate cannot be larger. Let $\bar{Q}(\vec{\mu},N,T)$ and $\bar{Q}(\mu,N,T)$ be the respective mean queue length when the staffing level is N, the routing parameters is T, and the service rate vector is either $\vec{\mu}$ or has all components identical and equal to μ, and let $\pi_B(\vec{\mu},N,T)$ and $\pi_B(\mu,N,T)$ be the associated steady state probabilities defined in the proof of Lemma 1. From flow balance (and recalling $\mu_\Sigma = \sum_{i=1}^{N}\mu_i$ was defined in the proof of Lemma 1),

$$\lambda - \sum_{i=1}^{N}\mu_i B_i(\vec{\mu},N,T) = \theta\bar{Q}(\vec{\mu},N,T) = \pi_B(\vec{\mu},N,T)\theta\sum_{i=1}^{\infty} i\left(\prod_{k=1}^{i}\frac{\lambda\exp(-\theta T)}{\mu_\Sigma + k\theta}\right)$$

and

$$\lambda - NB(\mu,N,T)\mu = \theta\bar{Q}(\mu,N,T) = \pi_B(\mu,N,T)\theta\sum_{i=1}^{\infty} i\left(\prod_{k=1}^{i}\frac{\lambda\exp(-\theta T)}{\mu_\Sigma + k\theta}\right).$$

Hence to show (EC.9), it is equivalent to show

$$\pi_B(\mu,N,T) \le \pi_B(\vec{\mu},N,T). \tag{EC.10}$$

From the proof of Lemma 1,

$$\pi_B(\vec{\mu},N,T) = \frac{1}{\sum_{\mathcal{I}\subseteq\mathcal{N}}|\mathcal{I}|!\prod_{j\in\mathcal{I}}\frac{\mu_j}{\lambda} + z}, \text{ where } z = \sum_{m=1}^{\infty}\prod_{k=1}^{m}\frac{\lambda}{\mu_\Sigma + k\theta}.$$

Since z is identical for both sides of (EC.10), it is equivalent to show that

$$\sum_{\mathcal{I}\subseteq\mathcal{N}}|\mathcal{I}|!\prod_{i\in\mathcal{I}}\frac{\mu_i}{\lambda} \le \sum_{\mathcal{I}\subseteq\mathcal{N}}|\mathcal{I}|!\left(\frac{\mu}{\lambda}\right)^{|\mathcal{I}|}.$$

We define $x_i := \frac{\mu_i}{\lambda} > 0$, $\bar{x}_i := \frac{\sum_{j=1}^{i} x_j}{i}$, for $i\in\mathcal{N}$. Note that $\bar{x}_N := \frac{\mu}{\lambda}$. The above inequality can be written as

$$\sum_{I=1}^{N} I! \sum_{\mathcal{I}\subseteq\mathcal{N},|\mathcal{I}|=I}\prod_{i\in\mathcal{I}} x_i \le \sum_{I=1}^{N} I!\binom{N}{I}(\bar{x}_N)^I.$$

It is sufficient to show

$$\sum_{\mathcal{I}\subseteq\mathcal{N},|\mathcal{I}|=I}\prod_{i\in\mathcal{I}} x_i \le \binom{N}{I}(\bar{x}_N)^I, \text{ for all } I\in\mathcal{N},$$

or more generally, by defining $\mathcal{N}_m := \{1,2,\cdots,m\}$ for all $m\in\mathcal{N}$, to show

$$\sum_{\mathcal{I}\subseteq\mathcal{N}_m,|\mathcal{I}|=I}\prod_{i\in\mathcal{I}} x_i \le \binom{m}{I}(\bar{x}_m)^I, \text{ for all } m\in\mathcal{N} \text{ and } I\in\mathcal{N}_m, \tag{EC.11}$$

It is trivially valid for $m\in\mathcal{N}, I=1$ from the definition that $\bar{x}_m = \sum_{j=1}^{m} x_j/m$. It is also valid for $m\in\mathcal{N}, I=m$ by the fact that the geometric average of positive numbers is no larger than their arithmetic average. With (EC.11) being valid on the two boundary lines of the triangle set (I,m), we next use a two-dimensional induction to show it is valid over the whole triangle.

For any $K\in\{3,4,\cdots,N\}$, given $I\in\mathcal{N}_{K-1}$, suppose (EC.11) is valid for $m=K-2$ and $m=K-1$, we show it is valid for $m=K$. Without loss of generality, we order x_i such that

$$x_1 \le x_2 \le \cdots \le x_N, \text{ and therefore, } \bar{x}_i \le x_j, \text{ for all } i,j\in\mathcal{N} \text{ and } i\le j.$$

$$\begin{aligned}\sum_{\mathcal{I}\subseteq\mathcal{N}_K,|\mathcal{I}|=I}\prod_{i\in\mathcal{I}}x_i &= \sum_{\mathcal{I}\subseteq\mathcal{N}_K,|\mathcal{I}|=I,K\notin\mathcal{I}}\prod_{i\in\mathcal{I}}x_i+\sum_{\mathcal{I}\subseteq\mathcal{N}_K,|\mathcal{I}|=I,K\in\mathcal{I}}\prod_{i\in\mathcal{I}}x_i\\ &= \sum_{\mathcal{I}\subseteq\mathcal{N}_{K-1},|\mathcal{I}|=I}\prod_{i\in\mathcal{I}}x_i+x_K\sum_{\mathcal{I}\subseteq\mathcal{N}_{K-1},|\mathcal{I}|=I-1}\prod_{i\in\mathcal{I}}x_i\\ &\le \binom{K-1}{I}\bar{x}_{K-1}^I+x_K\binom{K-1}{I-1}\bar{x}_{K-1}^{I-1}.\end{aligned}$$

It is sufficient to show

$$\binom{K-1}{I}\bar{x}_{K-1}^I+\binom{K-1}{I-1}\bar{x}_{K-1}^{I-1}x_K\le\binom{K}{I}\bar{x}_K^I.$$

That is equivalent to

$$(K-I)+I\frac{x_K}{\bar{x}_{K-1}}\le K\left(\frac{\bar{x}_K}{\bar{x}_{K-1}}\right)^I.$$

$$\begin{aligned}K\left(\frac{\bar{x}_K}{\bar{x}_{K-1}}\right)^I &= K\left(1+\frac{\bar{x}_K-\bar{x}_{K-1}}{\bar{x}_{K-1}}\right)^I=K\left(1+I\frac{\bar{x}_K-\bar{x}_{K-1}}{\bar{x}_{K-1}}+\cdots\right)\\ &\ge K\left(1+I\frac{\bar{x}_K-\bar{x}_{K-1}}{\bar{x}_{K-1}}\right)=K\left(1+I\left(\frac{((K-1)\bar{x}_{K-1}+x_K)/K}{\bar{x}_{K-1}}-1\right)\right)\\ &= K\left(1+I\frac{x_K-\bar{x}_{K-1}}{\bar{x}_{K-1}}\right)=(K-I)+I\frac{x_K}{\bar{x}_{K-1}}.\end{aligned}$$

Now we can use induction to prove (EC.11). We know that it is valid for $(m,I)=(m,1)$ and $(m,I)=(m,m)$, for all $m\in\mathcal{N}_N$. Starting from $I=2$, (2,1) and (2,2) are both valid. From the argument in the preceding paragraph we can use induction to show $(m,2)$ is valid, for all $m\in\{2,3,\cdots,N\}$. Next, for $I=3$, we know (3,2) and (3,3) are both valid. By induction $(m,3)$ is valid for all $m\in\{3,4,\cdots,N\}$. Keeping doing this till $I=N$, we prove (EC.11) is valid.

Proof of Lemma EC.2. Without loss of generality, suppose $\mu_1>\mu_2$, which means $x_1>x_2$. We want to show $B_1<B_2$ and $x_1B_1>x_2B_2$. From Lemma 1,

$$B_i(\vec{\mu},N,T):=\frac{\sum_{\mathcal{I}\subseteq\mathcal{N}\backslash i}|\mathcal{I}|!\prod_{j\in\mathcal{I}}x_j+z}{\sum_{\mathcal{I}\subseteq\mathcal{N}}|\mathcal{I}|!\prod_{j\in\mathcal{I}}x_j+z}.$$

B_1 and B_2 has the same denominator, the numerator of B_1 is

$$\sum_{\mathcal{I}\subseteq\mathcal{N}\backslash\{1,2\}}|\mathcal{I}|!\prod_{j\in\mathcal{I}}x_j+x_2\sum_{\mathcal{I}\subseteq\mathcal{N}\backslash\{1,2\}}(|\mathcal{I}|+1)!\prod_{j\in\mathcal{I}}x_j+z,$$

and the numerator of B_2 is

$$\sum_{\mathcal{I}\subseteq\mathcal{N}\setminus\{1,2\}}|\mathcal{I}|!\prod_{j\in\mathcal{I}}x_j+x_1\sum_{\mathcal{I}\subseteq\mathcal{N}\setminus\{1,2\}}(|\mathcal{I}|+1)!\prod_{j\in\mathcal{I}}x_j+z.$$

The numerators of x_1B_1 and x_2B_2 are

$$x_1\left(\sum_{\mathcal{I}\subseteq\mathcal{N}\setminus\{1,2\}}|\mathcal{I}|!\prod_{j\in\mathcal{I}}x_j+z\right)+x_1x_2\sum_{\mathcal{I}\subseteq\mathcal{N}\setminus\{1,2\}}(|\mathcal{I}|+1)!\prod_{j\in\mathcal{I}}x_j,$$

$$x_2\left(\sum_{\mathcal{I}\subseteq\mathcal{N}\setminus\{1,2\}}|\mathcal{I}|!\prod_{j\in\mathcal{I}}x_j+z\right)+x_1x_2\sum_{\mathcal{I}\subseteq\mathcal{N}\setminus\{1,2\}}(|\mathcal{I}|+1)!\prod_{j\in\mathcal{I}}x_j.$$

We see $x_1B_1>x_2B_2$ also by comparing the numerators. □

D. Proof of Lemma 3

Any policy that does not staff ($N_0^\lambda:=0$) and let all customers abandon has, from (8), for any $\mu\in[\underline{\mu},\overline{\mu}]$,

$$\frac{c_SN_0^\lambda+\mathcal{C}(\mu_0^\lambda,N_0^\lambda,T_0^\lambda)}{\lambda}=g_A(1)\text{ for all }\lambda\geq 0.$$

Assuming $g_A(1)>0$ (otherwise zero staffing is the optimal policy), any admissible policy $(\mu^\lambda,N^\lambda,T^\lambda)$ satisfies

$$\begin{aligned}\frac{c_SN^\lambda+\mathcal{C}(\mu^\lambda,N^\lambda,T^\lambda)}{c_SN_\star^\lambda+\mathcal{C}_\star^\lambda}&=\left(\frac{c_SN^\lambda+\mathcal{C}(\mu^\lambda,N^\lambda,T^\lambda)}{\lambda}\right)\times\left(\frac{\lambda}{c_SN_0^\lambda+\mathcal{C}(\mu_0^\lambda,N_0^\lambda,T_0^\lambda)}\right)\times\left(\frac{c_SN_0^\lambda+\mathcal{C}(\mu_0^\lambda,N_0^\lambda,T_0^\lambda)}{c_SN_\star^\lambda+\mathcal{C}_\star^\lambda}\right)\\&\geq\left(c_S\frac{N^\lambda}{\lambda}\right)\times\left(\frac{\lambda}{c_SN_0^\lambda+\mathcal{C}(\mu,N_0^\lambda,T_0^\lambda)}\right).\end{aligned}$$

If $\limsup_{\lambda\to\infty}N^\lambda/\lambda=\infty$, then any subsequence λ_i on which $N^{\lambda_i}/\lambda_i=\infty$ has

$$\frac{c_SN^{\lambda_i}+\mathcal{C}\left(\mu^{\lambda_i},N^{\lambda_i},T^{\lambda_i}\right)}{c_SN_\star^{\lambda_i}+\mathcal{C}_\star^{\lambda_i}}\to\infty\text{ as }\lambda_i\to\infty,$$

and so $(\mu^\lambda,N^\lambda,T^\lambda)$ cannot be an asymptotically optimal policy. □

E. Proof of Lemma 4

Under the routing rule in Definition 1, the fraction of customers that abandon from the holding area is $1-\exp(-\theta T^\lambda)$, and the fraction of customers that enter the queue is $\exp(-\theta T^\lambda)$. The queue operates as a $M/M/N^\lambda+M$ queue with arrival rate $\lambda\exp\left(-\theta T^\lambda\right)$, and we denote the expected steady state abandonment probability from the queue by $P(Ab_Q^\lambda)$. It follows that the expected steady state abandonment probability is

$$q_A^\lambda = \left(1 - \exp\left(-\theta T^\lambda\right)\right) + \exp\left(-\theta T^\lambda\right) P(Ab_Q^\lambda). \tag{EC.12}$$

From Theorem 1 in Garnett et al. (2002) adapted to our setting, noting that

$$\frac{\lambda \exp\left(-\theta T^\lambda\right)}{N^\lambda \mu^\lambda} \to \rho_\infty := \frac{\exp(-\theta T)}{b\mu}, \text{ as } \lambda \to \infty,$$

we conclude

$$\lim_{\lambda\to\infty} P(Ab_Q^\lambda) = \begin{cases} 0 & \text{if } \rho_\infty \in [0,1] \\ 1 - \frac{b\mu}{\exp(-\theta T)} & \text{if } \rho_\infty > 1 \end{cases}. \tag{EC.13}$$

Taking the limit in (EC.12) and using (EC.13) establishes that

$$a = \lim_{\lambda\to\infty} q_A^\lambda = \begin{cases} 1 - \exp(-\theta T) \text{ if } \rho_\infty \in [0,1] \\ 1 - b\mu \text{ if } \rho_\infty > 1 \end{cases}.$$

The agent utilization must satisfy

$$B\left(\mu^\lambda, N^\lambda, T^\lambda\right) = \frac{\lambda\,(1 - q_A^\lambda)}{N^\lambda \mu^\lambda} \to \beta := \min(\rho_\infty, 1) \text{ as } \lambda \to \infty,$$

from which $a = 1 - b\beta\mu \geq 0$ follows. □

F. Proof of Lemma 5

We show the minimizer in each of (15)-(18) is unique, one-by-one, in order of their appearance.

$\hat{\beta}_\star$ **in (15).** There exists a stationary point in $(0,1)$ if and only if $\hat{c}_S^{'}(\beta) = 0$, which occurs if and only if

$$\beta g_U'(\beta) - g_U(\beta) = c_S, \text{ for some } \beta \in (0,1). \tag{EC.14}$$

Since g_U is strictly convex, $[\beta g_U'(\beta) - g_U(\beta)]' = \beta g_U''(\beta) > 0$ implies the left-hand-side of (EC.14) is strictly increasing. Hence either there exists one solution to (EC.14) or no solution. In either case, noting that $\hat{c}_S(\beta) \to \infty$ as $\beta \downarrow 0$, so that the endpoint 0 is not a candidate to be a minimizer, the minimizer $\hat{\beta}_\star$ is unique.

$\hat{\mu}_\star$ in (16). Define $f(\mu) := \hat{c}_S(\hat{\beta}_\star)/\mu + (1-p(\mu))g_F(1-p(\mu))$. Since

$$f''(\mu) =$$
$$\frac{2\hat{c}_S(\hat{\beta}_\star)}{\mu^3} + (1-p(\mu))\left(p^{'}(\mu)\right)^2 g_F^{''}(1-p(\mu)) - g_F(1-p(\mu))p^{''}(\mu) + g_F^{'}(1-p(\mu))\left(2\left(p^{'}(\mu)\right)^2 - (1-p(\mu))p^{''}(\mu)\right)$$

is positive whenever p is weakly concave and g_F is weakly convex and weakly increasing, we conclude f is strictly convex. Hence the minimizer $\hat{\mu}_\star$ is unique.

$\hat{\mathbf{a}}_\star$ in (17). The strict convexity of $ag_A(a)$ in a on $[0,1]$ implies there is at most one critical point, which, if it exists, is the minimizer. Otherwise, if no critical point exists, then the minimizer occurs either at 0 or at 1. Both 0 and 1 cannot be minimizers because $\hat{c}_\star \neq g_A(1)$ by assumption, where $\hat{c}_\star$ is defined in (19).

$\hat{\mathbf{b}}_\star$ in (18). Uniqueness is immediate from the fact that $\hat{\beta}_\star$, $\hat{\mu}_\star$, and $\hat{a}_\star$ are all unique. □

G. Proof of Theorem 1

From Lemma 4,

$$\lim_{\lambda\to\infty} B(\mu_{\text{ao}}^\lambda, N_{\text{ao}}^\lambda, T_{\text{ao}}^\lambda) = \min\left(1, \frac{\exp(-\theta\hat{T}_\star)}{\hat{b}_\star\hat{\mu}_\star}\right),$$

and therefore, from the definition of $\hat{T}_\star$,

$$\min\left(1, \frac{\exp(-\theta\hat{T}_\star)}{\hat{b}_\star\hat{\mu}_\star}\right) = \begin{cases} \hat{\beta}_\star & \text{if } \hat{\beta}_\star < 1 \text{ and } \hat{a}_\star > 0, \\ \min\left(1, \frac{1}{\hat{b}_\star\hat{\mu}_\star}\right) & \text{otherwise.} \end{cases}$$

From (18), if $\hat{\beta}_\star = 1$, then $\hat{b}_\star\hat{\mu}_\star = 1 - \hat{a}_\star \le 1$, and $\min\left(1, \frac{1}{\hat{b}_\star\hat{\mu}_\star}\right) = 1$; if $\hat{a}_\star = 0$, then $\hat{b}_\star\hat{\mu}_\star = \frac{1}{\hat{\beta}_\star} \ge 1$, and $\min\left(1, \frac{1}{\hat{b}_\star\hat{\mu}_\star}\right) = \hat{\beta}_\star$. In summary, $\lim_{\lambda\to\infty} B(\mu_{\text{ao}}^\lambda, N_{\text{ao}}^\lambda, T_{\text{ao}}^\lambda) = \hat{\beta}_\star$.

As in (13), Lemma 4 implies

$$\lim_{\lambda\to\infty} \frac{c_S N_{\text{ao}}^\lambda + \mathcal{C}(\mu_{\text{ao}}, N_{\text{ao}}^\lambda, T_{\text{ao}})}{\lambda} = c_S\hat{b}_\star + \hat{\mathcal{C}}(\hat{\mu}_\star, \hat{b}_\star, \hat{\beta}_\star, \hat{a}_\star).$$

From Proposition 1, which holds by Assumption 1, a solution to (7) has all servers working at the same rate, and so (7) and (9) have the same minimum objective function value. Then, since $(\mu_\star^\lambda, N_\star^\lambda, T_\star^\lambda)$ solves (9) for each λ,

$$\frac{c_S N_\star^\lambda + \mathcal{C}\left(\mu_\star^\lambda, N_\star^\lambda, T_\star^\lambda\right)}{\lambda} \le \frac{c_S N_{\text{ao}}^\lambda + \mathcal{C}(\mu_{\text{ao}}, N_{\text{ao}}^\lambda, T_{\text{ao}})}{\lambda} \text{ for each } \lambda.$$

Hence

$$\limsup_{\lambda\to\infty}\frac{c_S N_\star^\lambda+\mathcal{C}\left(\mu_\star^\lambda,N_\star^\lambda,T_\star^\lambda\right)}{\lambda}\le c_S\hat{b}_\star+\hat{\mathcal{C}}(\hat{\mu}_\star,\hat{b}_\star,\hat{\beta}_\star,\hat{a}_\star). \tag{EC.15}$$

We next establish

$$\liminf_{\lambda\to\infty}\frac{c_S N_\star^\lambda+\mathcal{C}\left(\mu_\star^\lambda,N_\star^\lambda,T_\star^\lambda\right)}{\lambda}\ge c_S\hat{b}_\star+\hat{\mathcal{C}}(\hat{\mu}_\star,\hat{b}_\star,\hat{\beta}_\star,\hat{a}_\star), \tag{EC.16}$$

which, together with (EC.15) implies

$$\lim_{\lambda\to\infty}\frac{c_S N_\star^\lambda+\mathcal{C}\left(\mu_\star^\lambda,N_\star^\lambda,T_\star^\lambda\right)}{\lambda}= c_S\hat{b}_\star+\hat{\mathcal{C}}(\hat{\mu}_\star,\hat{b}_\star,\hat{\beta}_\star,\hat{a}_\star),$$

and so completes the proof. If (EC.16) is not true, then there is a subsequence λ_i such that

$$\lim_{\lambda_i\to\infty}\frac{c_S N_\star^{\lambda_i}+\mathcal{C}\left(\mu_\star^{\lambda_i},N_\star^{\lambda_i},T_\star^{\lambda_i}\right)}{\lambda_i}< c_S\hat{b}_\star+\hat{\mathcal{C}}(\hat{\mu}_\star,\hat{b}_\star,\hat{\beta}_\star,\hat{a}_\star). \tag{EC.17}$$

Since $\mu_\star^{\lambda_i}\in[\underline{\mu},\overline{\mu}]$ for all λ_i, $0\le\limsup_{\lambda_i\to\infty}N_\star^{\lambda_i}/\lambda_i<\infty$ from (EC.15), and $\beta_\star^{\lambda_i}:=B(N_\star^{\lambda_i},\mu_\star^{\lambda_i},T_\star^{\lambda_i})\in[0,1]$ for all λ_i, the Bolzano-Weierstrass theorem implies there exists a further subsequence λ_{i_j} on which

$$\mu_\star^{\lambda_{i_j}}\to\mu_0,\ \frac{N_\star^{\lambda_{i_j}}}{\lambda_{i_j}}\to b_0,\ \text{ and }\beta_\star^{\lambda_{i_j}}\to\beta_0,\ \text{ as }\lambda_{i_j}\to\infty.$$

Furthermore, on that subsequence, since $\lambda_{i_j}-N_\star^{\lambda_{i_j}}\beta_\star^{\lambda_{i_j}}\mu_\star^{\lambda_{i_j}}\ge 0$ for each λ_{i_j},

$$\frac{\lambda_{i_j}-N_\star^{\lambda_{i_j}}\beta_\star^{\lambda_{i_j}}\mu_\star^{\lambda_{i_j}}}{\lambda_{i_j}}\to 1-b_0\beta_0\mu_0\ge 0\ \text{ as }\lambda_{i_j}\to\infty.$$

Then, directly from the expression for $\mathcal{C}$ in (8),

$$\lim_{\lambda_{i_j}\to\infty}\frac{c_S N_\star^{\lambda_{i_j}}+\mathcal{C}\left(\mu_\star^{\lambda_{i_j}},N_\star^{\lambda_{i_j}},T_\star^{\lambda_{i_j}}\right)}{\lambda_{i_j}}=c_S b_0+\hat{\mathcal{C}}\left(\mu_0,b_0,\beta_0,1-b_0\beta_0\mu_0\right),$$

which contradicts (EC.17) because the minimizers $\hat{\mu}_\star$, $\hat{b}_\star$, $\hat{\beta}_\star$, and $\hat{a}_\star$ are unique by Lemma 5, under Assumption 2. We conclude (EC.16) holds. □

H. Proof of Lemma 6

Define $f(a) := (1-a)\hat{c}_\star + a g_A(a)$. The first order condition $f'(a) = 0$ is equivalent to

$$a g_A'(a) + g_A(a) = \hat{c}_\star.$$

Since $a g_A(a)$ is strictly convex in a on $[0,1]$, the left-hand-side in the above display strictly increases from $g_A(0)$ to $g_A'(1) + g_A(1)$. If $g_A'(1) + g_A(1) < \hat{c}_\star$, then no solution to the first order condition exists. In that case, since $f(0) = \hat{c}_\star$, $f(1) = g_A(1)$, and $g_A'(1) \geq 0$, the minimizer $\hat{a}_\star = 1$, which from (18) implies $\hat{b}_\star = 0$. □

I. Proof of Proposition 2

Part (a). Define $f(a) := (1-a)\hat{c}_\star + a g_A(a)$. Since f is strictly convex, if a solution to the first-order condition

$$a g_A'(a) + g_A(a) = \hat{c}_\star \tag{EC.18}$$

exists, then it is a minimum. The left-hand side of (EC.18) increases from $g_A(0)$ to $g_A'(1) + g_A(1)$, and so if $g_A(0) < \hat{c}_\star$, then $a_\star \in (0,1)$. Otherwise, if $g_A(0) = \hat{c}_\star$, then $a_\star = 0$ solves (EC.18), and if $g_A(0) > \hat{c}_\star$, then no solution to (EC.18) exists. If no solution to (EC.18) exists, then $a_\star = 0$ because

$$f'(a) = -\hat{c}_\star + g_A(a) + a g_A'(a) \geq -\hat{c}_\star + \min_{a \in [0,1]} \left(a g_A(a)\right)' = -\hat{c}_S + g_A(0) > 0$$

implies f is strictly increasing.

Part (b). Define $h(\beta) := \beta g_U'(\beta) - g_U(\beta)$, and note that $h'(\beta) = \beta g_U''(\beta) > 0$ (recalling g_U is strictly convex), so that h is strictly increasing on $[0,1]$. If $h(1) = g_U'(1) - g_U(1) \leq c_S$, then

$$\hat{c}_S'(\beta) = \frac{h(\beta) - c_S}{\beta^2} < \frac{h(1) - c_S}{\beta^2} \leq 0$$

implies $\hat{\beta}_\star = 1$. Otherwise, if $h(1) = g_U'(1) - g_U(1) > c_S$, then $h(0) = -g_U(0) < c_S < h(1)$ implies there exists a solution $\beta_0 \in (0,1)$ to the first-order condition $\hat{c}_S'(\beta) = 0$, so that $h(\beta_0) = c_S$. Since

$$\hat{c}_S''(\beta) = \frac{\beta^2 h'(\beta) - 2\beta\left(h(\beta) - c_S\right)}{\beta^4},$$

we find $\hat{c}_S''(\beta_0) = h'(\beta_0)/\beta_0^2 > 0$, meaning β_0 is a local minimum. Since β_0 is the only stationary point on $[0,1]$, we conclude $\hat{\beta}_\star = \beta_0 \in (0,1)$. □

J. Proof of Proposition 3

We drop the superscript λ for convenience in presentation. Also, we write $B(\mu_1,\mu)$ instead of $B((\mu_1,\mu),N,T)$.

We first observe that the best response function (22) is equivalently written as

$$R(\mu)=\underset{\mu_1\in[\underline{\mu},\overline{\mu}]}{\arg\max}\,(1-P_R(1-p(\mu_1)))\,\mu_1\times B(\mu_1,\mu),$$

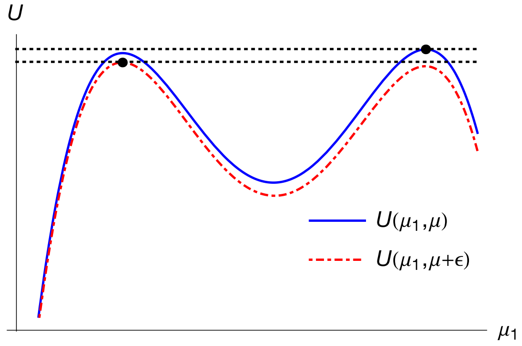
where we have substituted $P_R=P_F/P_S$ in (22). The implication is that any fixed point μ_F depends on P_F and P_S only through the payment ratio P_R. Therefore, if we set

$$P_S=\frac{c_S}{(1-P_R(1-p(\mu_F)))\,\mu_F\times B(\mu_F,\mu_F)}$$

to ensure the individual rationality constraint (23) is satisfied, and $P_F=P_S\times P_R$, then μ_F is a symmetric equilibrium service rate. In summary, we must show there exists a fixed point of (22).

The existence of a fixed point μ_F follows if $R(\mu)$ is continuous in μ. This is because the domain of R is $[\underline{\mu},\overline{\mu}]$ and the range of R is a subset of $[\underline{\mu},\overline{\mu}]$, which implies there exists at least one $\mu\in[\underline{\mu},\overline{\mu}]$ such that $R(\mu)=\mu$. A sufficient condition for the continuity of R is that the function $U(\mu_1,\mu)$ is quasiconcave in μ_1 on $[\underline{\mu},\overline{\mu}]$ for any fixed $\mu\in[\underline{\mu},\overline{\mu}]$ so it has a unique maximizer. Figure EC.1 illustrates why a small change of μ can lead to a large change in $R_1(\mu)$ if $U(\mu_1,\mu)$ is multimodal.

Figure EC.1 Illustration of Possible Discontinuity in $R_1(\mu)$



To show the desired quasiconcavity, we use the second partial derivative to argue $U(\mu_1,\mu)$ can have at most one stationary point, which is a local maximum. For

$$P(\mu):=(P_S-P_F(1-p(\mu)))\,\mu,\ \text{for}\ \mu\in[\underline{\mu},\overline{\mu}]\,, \tag{EC.19}$$

so that

$$U(\mu_1,\mu) = P(\mu) \times B(\mu_1,\mu),$$

the first partial derivative is

$$\frac{\partial U(\mu_1,\mu)}{\partial \mu_1} = P(\mu_1)\frac{\partial B(\mu_1,\mu)}{\partial \mu_1} + P'(\mu_1)B(\mu_1,\mu),$$

and the second partial derivative is

$$\begin{aligned}\frac{\partial^2 U(\mu_1,\mu)}{\partial \mu_1^2} &= P^{''}(\mu)B(\mu_1,\mu) + 2P^{'}(\mu_1)\frac{\partial B(\mu_1,\mu)}{\partial \mu_1} + P(\mu_1)\frac{\partial^2 B(\mu_1,\mu)}{\partial \mu_1^2} \\ &= P^{''}(\mu_1)B(\mu_1,\mu) + \frac{P(\mu_1)}{B(\mu_1,\mu)}\left(B(\mu_1,\mu)\frac{\partial^2 B(\mu_1,\mu)}{\partial \mu_1^2} - 2\left(\frac{\partial B(\mu_1,\mu)}{\partial \mu_1}\right)^2\right) + 2\frac{\partial U(\mu_1,\mu)}{\partial \mu_1}\frac{\frac{\partial B(\mu_1,\mu)}{\partial \mu_1}}{B(\mu_1,\mu)}.\end{aligned} \tag{EC.20}$$

The first term is negative because $P(\mu)$ is strictly concave on $\left[\underline{\mu},\overline{\mu}\right]$ by assumption on p. The second term is negative from the following result, whose proof is provided below.

Lemma EC.3. *For any $\mu_1,\mu \in \left[\underline{\mu},\overline{\mu}\right]$,*

$$B(\mu_1,\mu)\frac{\partial^2 B(\mu_1,\mu)}{\partial \mu_1^2} - 2\left(\frac{\partial B(\mu_1,\mu)}{\partial \mu_1}\right)^2 < 0.$$

For the third term, at a stationary point satisfying $\frac{\partial U(\mu_1,\mu)}{\partial \mu_1} = 0$, the third term is zero. As a result, $\frac{\partial^2 U(\mu_1,\mu)}{\partial \mu_1^2} < 0$. Hence, any stationary point is a local maximum. Since any stationary point must be a local maximum, there can exist at most one stationary point. Therefore, $U(\mu_1,\mu)$ has at most one stationary point on $\left[\underline{\mu},\overline{\mu}\right]$, which is a local maximum, implying quasiconcavity.

Proof of Lemma EC.3. Define $a := \lambda \exp(-\theta T)$, and $z := \sum_{i=1}^{\infty}\prod_{k=1}^{i}\frac{a}{(N-1)\mu+\mu_1+k\theta}$. From Lemma 1,

$$B_1(\mu_1,\mu) = \frac{\sum_{I\subseteq\mathcal{N}\backslash 1}|I|!\left(\frac{\mu}{a}\right)^{|I|} + z}{\sum_{I\subseteq\mathcal{N}\backslash 1}|I|!\left(\frac{\mu}{a}\right)^{|I|} + \sum_{I\subseteq\mathcal{N}\,\&\,1\in I}|I|!\frac{\mu_1}{a}\left(\frac{\mu}{a}\right)^{|I|-1} + z}.$$

From combinatorics,

$$\sum_{I\subseteq\mathcal{N}\backslash 1}|I|!\left(\frac{\mu}{a}\right)^{|I|} = \sum_{i=1}^{N-1}\frac{(N-1)!}{(N-1-i)!}\left(\frac{\mu}{a}\right)^{i} + 1, \qquad \sum_{I\subseteq\mathcal{N}\,\&\,1\in I}|I|!\frac{\mu_1}{a}\left(\frac{\mu}{a}\right)^{|I|-1} = \frac{\mu_1}{a}\sum_{i=1}^{N} i\frac{(N-1)!}{(N-i)!}\left(\frac{\mu}{a}\right)^{i-1}.$$

Multiplying both numerator and denominator by $\left(\frac{a}{\mu}\right)^{N-1}\frac{1}{(N-1)!}$ and so

$$B_1(\mu_1,\mu)=\frac{\sum_{j=0}^{N-1}\left(\frac{a}{\mu}\right)^j\frac{1}{j!}+\left(\frac{a}{\mu}\right)^{N-1}\frac{z}{(N-1)!}}{\sum_{j=0}^{N-1}\left(\frac{a}{\mu}\right)^j\frac{1}{j!}+\frac{\mu_1}{a}\sum_{j=0}^{N-1}\left(\frac{a}{\mu}\right)^j\frac{N-j}{j!}+\left(\frac{a}{\mu}\right)^{N-1}\frac{z}{(N-1)!}}.$$

For notation simplification we use $B_1(\mu_1\mu)$ instead of $B_1\left((\mu_1,\mu),N,T\right)$. Let X be a Poisson random variable with parameter $\frac{a}{\mu}$, and observe that by multiplying both numerator and denominator by $\exp\left(-\frac{a}{\mu}\right)$, we can rewrite B_1 as follows:

$$\begin{aligned}B_1\left((\mu_1,\mu),N,T\right)&=\frac{\Pr(X\le N-1)+P(X=N-1)z}{\Pr(X\le N-1)+\frac{\mu_1 N}{a}\Pr(X\le N-1)-\frac{\mu_1}{\mu}\Pr(X\le N-2)+\Pr(X=N-1)z}\\&=\frac{K+z}{K+\mu_1\Big(K\left(\frac{N}{a}-\frac{1}{\mu}\right)+\frac{1}{\mu}\Big)+z},\end{aligned}\tag{EC.21}$$

where

$$K:=\frac{\Pr(X\le N-1)}{\Pr(X=N-1)}.$$

We have

$$B(\mu_1,\mu)=\frac{K+z}{K+J\mu_1+z},\ \text{where } J:=K\left(\frac{N}{a}-\frac{1}{\mu}\right)+\frac{1}{\mu}.$$

Define $f_i(\mu_1):=\prod_{k=1}^{i}\frac{\lambda}{(N-1)\mu+\mu_1+k\theta}$, so that $z=\sum_{i=1}^{\infty}f_i(\mu_1)$. From calculus,

$$f_i^{'}(\mu_1)=-f_i(\mu_1)\sum_{k=1}^{i}\frac{1}{(N-1)\mu+\mu_1+k\theta}=-f_i(\mu_1)g_i(\mu_1),\ \text{for } g_i(\mu_1):=\sum_{k=1}^{i}\frac{1}{(N-1)\mu+\mu_1+k\theta},$$

and

$$g_i^{'}(\mu_1)=-h_i(\mu_1)\ \text{for } h_i(\mu_1):=\sum_{k=1}^{i}\frac{1}{((N-1)\mu+\mu_1+k\theta)^2}.$$

Assuming the interchange of summation and derivative, and dropping the arguments μ_1 from f_i,g_i,h_i for simplicity,

$$\frac{\partial B(\mu_1,\mu)}{\partial\mu_1}=-\frac{J(K+z+\mu_1\sum_{k=1}^{\infty}f_kg_k)}{(K+J\mu_1+z)^2}.\tag{EC.22}$$

Next, using the above expression and again assuming the interchange of summation and derivative,

$$\frac{\partial^2B(\mu_1,\mu)}{\partial\mu_1^2}=J\frac{2(J-\sum_{k=1}^{\infty}f_kg_k)(K+z+\mu_1\sum_{k=1}^{\infty}f_kg_k)+\mu_1(K+J\mu_1+z)\sum_{k=1}^{\infty}f_k(g_k^2+h_k)}{(K+J\mu_1+z)^3}.\tag{EC.23}$$

Substitution then shows

$$\begin{aligned}&B(\mu_1,\mu)\frac{\partial^2 B(\mu_1,\mu)}{\partial\mu_1^2}-2\left(\frac{\partial B(\mu_1,\mu)}{\partial\mu_1}\right)^2\\&=\frac{J}{(K+J\mu_1+z)^3}\left(-2(K+z)\sum_{k=1}^{\infty}f_kg_k-2\mu_1\left(\sum_{k=1}^{\infty}f_kg_k\right)^2+\mu_1(K+z)\sum_{k=1}^{\infty}f_k(g_k^2+h_k)\right).\end{aligned}$$

Suppose we can show

$$z\sum_{k=1}^{\infty}f_k(g_k^2+h_k)<2\left(\sum_{k=1}^{\infty}f_kg_k\right)^2. \tag{EC.24}$$

Then,

$$\begin{aligned}&B(\mu_1,\mu)\frac{\partial^2 B(\mu_1,\mu)}{\partial\mu_1^2}-2\left(\frac{\partial B(\mu_1,\mu)}{\partial\mu_1}\right)^2\\&<\frac{J}{(K+J\mu_1+z)^3}\left(-2(K+z)\sum_{k=1}^{\infty}f_kg_k-2\mu_1\left(\sum_{k=1}^{\infty}f_kg_k\right)^2+\frac{2\mu_1(K+z)}{z}\left(\sum_{k=1}^{\infty}f_kg_k\right)^2\right)\\&=\frac{2J\sum_{k=1}^{\infty}f_kg_k}{(K+J\mu_1+z)^3}\left(-(K+z)+\frac{\mu_1K}{z}\sum_{k=1}^{\infty}f_kg_k\right),\end{aligned}$$

so that showing

$$\mu_1K\sum_{k=1}^{\infty}f_kg_k-Kz-z^2<0 \tag{EC.25}$$

is sufficient to complete the proof. (EC.22), (EC.23) and (EC.24) are shown in Lemma EC.4.

Since $g_k\le kg_1$ for $k\ge1$, to show (EC.25) is valid, it is sufficient to show

$$K\left(g_1\mu_1\sum_{k=1}^{\infty}kf_k-z\right)-z^2<0. \tag{EC.26}$$

We divide the argument into tow cases: $a\le(N-1)\mu$ and $a>(N-1)\mu$.

- **Case 1 ($a\le(N-1)\mu$):** We show that

$$\sum_{k=1}^{\infty}\frac{kf_k}{z}-\frac{1}{g_1\mu_1}<0, \tag{EC.27}$$

which implies (EC.26) is valid. Define $\rho:=\frac{\lambda}{(N-1)\mu+\mu_1}<1$. Since $\theta>0$, for given $k>1$, we have $f_i<f_k\rho^{i-k}$ when $i>k$, and $f_i>f_k\rho^{i-k}$ when $1\le i<k$, which implies

$$\sum_{i=k}^{\infty}f_i<\sum_{i=k}^{\infty}f_k\rho^{i-k},\quad \sum_{i=1}^{k-1}f_i>\sum_{i=1}^{k-1}f_k\rho^{i-k},\ \text{for any } k>1.$$

If we view $\frac{f_k}{z}$ as a probability distribution and G as a geometric random variable with parameter ρ, then

$$\frac{\sum_{i=k}^{\infty} f_i}{z} = \frac{\sum_{i=k}^{\infty} f_i}{\sum_{i=1}^{k-1} f_i + \sum_{i=k}^{\infty} f_i} < \frac{\sum_{i=k}^{\infty} f_k \rho^{i-k}}{\sum_{i=1}^{k-1} f_k \rho^{i-k} + \sum_{i=k}^{\infty} f_k \rho^{i-k}} = \rho^{k-1} = P(G \geq k),$$

which implies stochastic dominance. Then since G has mean $\frac{1}{1-\rho}$,

$$\sum_{k=1}^{\infty} \frac{k f_k}{z} < \frac{1}{1-\rho},$$

and algebra shows that

$$\frac{1}{1-\rho} = \frac{(N-1)\mu + \mu_1}{(N-1)\mu + \mu_1 - a} < \frac{(N-1)\mu + \mu_1 + \theta}{\mu_1} = \frac{1}{g_1 \mu_1},$$

(EC.27) follows.

- **Case 2 ($a > (N-1)\mu$):** For any $1 \leq j \leq N$,

$$\frac{\Pr(X = N-j)}{\Pr(X = N-1)} = \frac{\left(\frac{a}{\mu}\right)^{N-j} \frac{1}{(N-j)!}}{\left(\frac{a}{\mu}\right)^{N-1} \frac{1}{(N-1)!}} = \left(\frac{(N-1)\mu}{a}\right)^{j-1} \frac{\prod_{k=1}^{j-1}(N-k)}{(N-1)^{j-1}} \leq \left(\frac{(N-1)\mu}{a}\right)^{j-1}, \tag{EC.28}$$

which implies

$$K = \frac{\Pr(X \leq N-1)}{\Pr(X = N-1)} \leq \sum_{j=1}^{N} \left(\frac{(N-1)\mu}{a}\right)^{j-1} < \frac{a}{a-(N-1)\mu}. \tag{EC.29}$$

Then, to establish (EC.26), it is sufficient to show

$$\frac{a}{a-(N-1)\mu} \left(g_1 \mu_1 \sum_{k=1}^{\infty} k f_k - z \right) - z^2 < 0. \tag{EC.30}$$

Define $u := \frac{(N-1)\mu}{\theta} > 0, v := \frac{a}{\theta} > 0, w := \frac{\mu_1}{\theta} > 0$. Due to $\lambda > (N-1)\mu$, $v > u$. From Equation (12) in Ancker and Gafarian (1962), regarding z as a function of u, v and w, we have

$$z(u,v,w) = \exp(v) v^{-u-w} \gamma(u+w+1, v), \tag{EC.31}$$

where $\gamma(u+w+1,v) := \int_0^v t^{u+w} \exp(-t) dt$ is the lower incomplete Gamma function. Next, we have

$$\frac{\partial f_i}{\partial v} = \frac{i v^{i-1}}{\prod_{j=0}^{i}(u+w+j)} = \frac{i f_i}{v}.$$

Then, assuming the interchange of summation and derivative,

$$\begin{aligned}\sum_{i=1}^{\infty} if_i = v\sum_{i=1}^{\infty}\frac{\partial f_i}{\partial v} = v\frac{\partial z}{\partial v} = v\frac{\partial\left(v^{-u-w}\exp(v)\gamma(u+w+1,v)\right)}{\partial v}\\ = (v-u-w)v^{-u-w}\exp(v)\gamma(u+w+1,v)+v = (v-u-w)z+v.\end{aligned} \tag{EC.32}$$

where the interchange of summation and derivative required for the second equality is shown in Lemma EC.4. Therefore, (EC.30) is equivalent to

$$\frac{v}{v-u}\left(\frac{w}{u+w+1}((v-u-w)z+v)-z\right)-z^2<0.$$

Since the quadratic having input z defines a parabola that opens downwards with one positive root and one negative root, showing the above is equivalent to showing

$$z(u,v,w) > R(u,v,w), \text{ for all } v>u>0, w>0. \tag{EC.33}$$

where

$$R(u,v,w) := v\frac{\sqrt{(1+u+w-(v-u)w+w^2)^2+4(v-u)(u+w+1)w}-(1+u+w-(v-u)w+w^2)}{2(v-u)(u+w+1)}$$

is the positive root of the aforementioned quadratic.

We first show the inequality (EC.33) at the boundary. We have

$$\lim_{v\downarrow u} R(u,v,w) = \frac{uw}{1+u+w+w^2}, \text{ for any given } u>0, w>0.$$

To show $z(u,u,w) > \lim_{v\downarrow u} R(u,v,w)$, it is sufficient to show the following lower bound

$$z(u,v,w) = \sum_{i=1}^{\infty}\prod_{j=1}^{i}\frac{v}{u+w+j} > \frac{vw}{1+u+w+w^2}, \text{ for } v\geq u>0, w>0. \tag{EC.34}$$

We denote the partial sums by $z_n := \sum_{i=1}^{n}\prod_{j=1}^{i}\frac{v}{u+w+j}$. We use induction to show that

$$z_n - \frac{vw}{1+u+w+w^2} > \frac{(n-w)u^n v}{(1+u+w+w^2)\prod_{j=1}^{n}(u+w+j)}. \tag{EC.35}$$

Showing (EC.35) implies (EC.34) because for $n_0 > w$, the right hand side of (EC.35) is positive and $z = \lim_{n\to\infty} z_n > z_{n_0}$.

- When $n=1$,

$$z_1 - \frac{vw}{1+u+w+w^2} = \frac{uv+v-wuv}{(1+u+w+w^2)(u+w+1)} > \frac{(1-w)uv}{(1+u+w+w^2)(u+w+1)}$$

  verifies (EC.35).

- Suppose (EC.35) is valid for $n>1$. Then, (EC.35) also holds for $n+1$ because

$$\begin{aligned} z_{n+1} - \frac{vw}{1+u+w+w^2} &> \frac{(n-w)u^n v}{(1+u+w+w^2)\prod_{j=1}^{n}(u+w+j)} + \frac{v^{n+1}}{\prod_{j=1}^{n+1}(u+w+j)} \\ &> \frac{(n-w)u^n v}{(1+u+w+w^2)\prod_{j=1}^{n}(u+w+j)} + \frac{u^n v}{\prod_{j=1}^{n+1}(u+w+j)} \\ &= \frac{((n+1-w)u+n^2+n+1)u^n v}{(1+u+w+w^2)\prod_{j=1}^{n+1}(u+w+j)} \\ &> \frac{(n+1-w)u^{n+1} v}{(1+u+w+w^2)\prod_{j=1}^{n+1}(u+w+j)}. \end{aligned}$$

The last step in the proof is to show that for any given $u>0, w>0$, z should be always above $R(u,v,w)$ as v increases from u to ∞. If at some point $z=R(u,v,w)$, denote by v_0 the smallest $v>u$ that equalizes the two. Since $z(u,v,w) > \lim_{v\downarrow u} R(u,v,w)$, at the intersection, we must have

$$\left.\frac{\partial z}{\partial v}\right|_{v=v_0} < \left.\frac{\partial R}{\partial v}\right|_{v=v_0}. \tag{EC.36}$$

We use contradiction to show that such a v_0 cannot exist. Recall from (EC.32)

$$\frac{\partial z}{\partial v} = \frac{v-u-w}{v} z + 1.$$

We check two cases:

- If $v_0 \geq u+w$, then $\left.\frac{\partial z}{\partial v}\right|_{v=v_0} \geq 1$. We can use Mathematica's Reduce function (see the second to last line of code at the end of the proof) to show that

$$0 < \frac{\partial R}{\partial v} < 1, \text{ for } v>u>0, w>0.$$

  It is a contradiction to (EC.36).

- If $v_0 \in (u, u+w)$, then we have

$$\left.\frac{\partial z}{\partial v}\right|_{v=v_0} > \frac{v_0-u-w}{v_0} R(u,v_0,w) + 1.$$

We can also use Mathematica's Reduce function (see the last line of code at the end of the proof) to show that

$$\frac{\partial R}{\partial v} < \frac{v-u-w}{v}R+1, \text{ for } v>u>0, w>v-u.$$

It is also a contradiction to (EC.36). In summary, we have

$$z(u,v,w) > R(u,v,w), \text{ for } v>u>0, w>0.$$

Finally, to complete the proof, we have the following Lemma.

Lemma EC.4. *The equalities (EC.22), (EC.23), (EC.32), and the inequality (EC.24) hold.*

Proof of Lemma EC.4: The results can be shown using similar arguments to those in the proof of Lemma 1 in Zhan and Ward (2018). Specifically, the interchanges of summation and derivative leading to (EC.22), (EC.23) and (EC.32) follow as in the the last paragraph of Step 1 in that proof. Next, (EC.24) follows by observing that display (2) in Zhan and Ward (2018) holds when the μ that appears in the $B(\mu), a_i(\mu), b_i(\mu), c_i(\mu)$ defined in that paper (the second display after (1) and in the display after (2) in that paper) is replaced by $(N-1)\mu+\mu_1+\theta$.

Mathematica Code for Algebra Proof:

$$R = v(\text{Sqrt}[(1+u+w+uw-vw+w\hat{}2)\hat{}2+4(v-u)w(1+u+w)]-(1+u+w+uw-vw+w\hat{}2))$$

$$/(2(v-u)(1+u+w))$$

$$DR = \text{D}[R, v]$$

$$\text{Reduce}[0 < DR < 1\,\&\&\, w>0\,\&\&\, u>0\,\&\&\, v>u]$$

$$\text{Reduce}[DR < R(v-u-w)/v+1\,\&\&\, v>u>0\,\&\&\, w>v-u]$$

□

K. Proof of Proposition 4.

From (EC.21), adding the superscript λ in the notation,

$$B_1^\lambda\left((\mu_1,\mu),N^\lambda,T^\lambda\right)=\frac{K^\lambda+z^\lambda}{K^\lambda\left(1+\frac{\mu_1 N^\lambda}{a^\lambda}-\frac{\mu_1}{\mu}\right)+\frac{\mu_1}{\mu}+z^\lambda}, \tag{EC.37}$$

We divide the argument into two cases: $b\mu<\exp(-\theta T)$ and $b\mu\geq\exp(-\theta T)$.

Case 1: $b\mu<\exp(-\theta T)$**.** From (EC.29),

$$K^\lambda\leq\sum_{j=1}^{N^\lambda}\left(\frac{(N^\lambda-1)\mu}{a^\lambda}\right)^{j-1}<\sum_{j=1}^{\infty}\left(\frac{N^\lambda\mu}{a^\lambda}\right)^{j-1}.$$

Since $b\mu<\exp(-\theta T)$ implies $\frac{N^\lambda\mu}{a^\lambda}<1$ for all large enough λ,

$$\sum_{j=1}^{\infty}\left(\frac{N^\lambda\mu}{a^\lambda}\right)^{j-1}=\frac{1}{1-\frac{N^\lambda\mu}{a^\lambda}}\to\frac{\exp(-\theta T)}{\exp(-\theta T)-b\mu},\ \text{as } \lambda\to\infty.$$

Hence,

$$\limsup_{\lambda\to\infty}K^\lambda\leq\frac{\exp(-\theta T)}{\exp(-\theta T)-b\mu}. \tag{EC.38}$$

Next, by defining $C^\lambda:=(N^\lambda-1)\mu+\mu_1$, from (EC.31)we have

$$z^\lambda=\left(\frac{a^\lambda}{\theta}\right)^{-\frac{C^\lambda}{\theta}}\exp\left(\frac{a^\lambda}{\theta}\right)\gamma\left(\frac{C^\lambda}{\theta}+1,\frac{a^\lambda}{\theta}\right).$$

From Pagurova (1965),

$$\lim_{a\to\infty}\frac{\gamma\left(a,a+x\sqrt{a}\right)}{\Gamma(a)}=\Phi(x),$$

where $\Gamma(a):=\int_0^\infty\exp(-t)t^{a-1}dt$ is the Gamma function and $\Phi(x)$ is the c.d.f. of standard normal distribution. For any given $t>1$, $x>0$, when a is large enough, $ta>a+x\sqrt{a}$, and therefore

$$\liminf_{a\to\infty}\frac{\gamma\left(a,ta\right)}{\Gamma(a)}\geq\lim_{a\to\infty}\frac{\gamma\left(a,a+x\sqrt{a}\right)}{\Gamma(a)}=\Phi(x).$$

The above is valid for any $x>0$, implying the liminf is 1. Note $\gamma(a,x)<\Gamma(a)$ for any $x>0$, implying the limsup is also 1. Therefore,

$$\lim_{a\to\infty}\frac{\gamma\left(a,ta\right)}{\Gamma(a)}=1,\ \text{for any } t>1.$$

Note $\lim_{\lambda\to\infty} \frac{a^\lambda/\theta}{C^\lambda/\theta+1} = \frac{\exp(-\theta T)}{b\mu} > 1$, we have $\gamma\left(\frac{C^\lambda}{\theta}+1, \frac{a^\lambda}{\theta}\right) \sim \Gamma\left(\frac{C^\lambda}{\theta}+1\right) \sim \sqrt{\frac{2\pi C^\lambda}{\theta}}\left(\frac{C^\lambda}{e\theta}\right)^{\frac{C^\lambda}{\theta}}$, and

$$z^\lambda \sim \sqrt{\frac{2\pi C^\lambda}{\theta}} \exp\left(\frac{a^\lambda}{\theta} - \frac{C^\lambda}{\theta} + \frac{C^\lambda}{\theta}\log\left(\frac{C^\lambda}{a^\lambda}\right)\right) = \sqrt{\frac{2\pi C^\lambda}{\theta}} \exp\left(\frac{C^\lambda}{\theta}\left(\frac{a^\lambda}{C^\lambda} - 1 - \log\left(\frac{a^\lambda}{C^\lambda}\right)\right)\right).$$

Since $\lim_{\lambda\to\infty} \frac{a^\lambda}{C^\lambda} = \frac{\exp(-\theta T)}{b\mu} > 1$, $\lim_{\lambda\to\infty} z^\lambda = \infty$. Then, from (EC.38) we have $\lim_{\lambda\to\infty} K^\lambda z^\lambda = \infty$. From (EC.37), we have $\lim_{\lambda\to\infty} B_1^\lambda\left((\mu_1,\mu), N^\lambda, T^\lambda\right) = 1$.

Case 2: $b\mu \geq \exp(-\theta T)$**.** For any fixed $j \geq 1$, noting $N^\lambda > j$ when λ is large enough, $\lim_{\lambda\to\infty} \frac{\prod_{k=1}^{j-1}(N^\lambda - k)}{(N^\lambda - 1)^{j-1}} = 1$. Then, from (EC.28),

$$\lim_{\lambda\to\infty} \frac{\Pr(X^\lambda = N^\lambda - j)}{\Pr(X^\lambda = N^\lambda - 1)} = \left(\frac{b\mu}{\exp(-\theta T)}\right)^{j-1} \geq 1.$$

That means, given any $\epsilon \in (0,1)$, and any integer $I > 0$, there exists $\Lambda(\epsilon, I)$ such that when $\lambda > \Lambda(\epsilon, I)$, $N^\lambda > I$, and for any $1 \leq j \leq I$,

$$\frac{\Pr(X^\lambda = N^\lambda - j)}{\Pr(X^\lambda = N^\lambda - 1)} > 1 - \frac{\epsilon}{I}.$$

Therefore,

$$K^\lambda = \frac{\Pr(X^\lambda \leq N^\lambda - 1)}{\Pr(X^\lambda = N^\lambda - 1)} > \sum_{j=1}^{I} \frac{\Pr(X^\lambda = N^\lambda - j)}{\Pr(X^\lambda = N^\lambda - 1)} > \sum_{j=1}^{I}\left(1 - \frac{\epsilon}{I}\right) = I - \epsilon > I - 1.$$

Since I can be arbitrarily large, we have

$$\lim_{\lambda\to\infty} K^\lambda = \infty. \tag{EC.39}$$

If $b\mu = \exp(-\theta T)$, we have $\lim_{\lambda\to\infty} \frac{N^\lambda}{a^\lambda} = \frac{1}{\mu}$, and therefore, from (EC.37),

$$\lim_{\lambda\to\infty} B_1^\lambda\left((\mu_1,\mu), N^\lambda, T^\lambda\right) = \lim_{\lambda\to\infty} \frac{K^\lambda + z^\lambda}{K^\lambda + \frac{\mu_1}{\mu} + z^\lambda} = 1.$$

If $b\mu > \exp(-\theta T)$, since

$$z^\lambda = \sum_{i=1}^{\infty} \prod_{k=1}^{i} \frac{a^\lambda}{(N^\lambda - 1)\mu + \mu_1 + k\theta} \leq \sum_{i=1}^{\infty}\left(\frac{a^\lambda}{(N^\lambda - 1)\mu}\right)^i,$$

and $b\mu > \exp(-\theta T)$ implies $\frac{a^\lambda}{(N^\lambda - 1)\mu} < 1$ for all large enough λ, so that

$$\sum_{i=1}^{\infty}\left(\frac{a^\lambda}{(N^\lambda - 1)\mu}\right)^i = \frac{1}{1 - \frac{a^\lambda}{(N^\lambda - 1)\mu}} \to \frac{\exp(-\theta T)}{b\mu - \exp(-\theta T)}, \text{ as } \lambda \to \infty,$$

we find $\limsup_{\lambda\to\infty} z^\lambda < \infty$. Combined with $\lim_{\lambda\to\infty} K^\lambda = \infty$ in (EC.39), we have

$$\lim_{\lambda\to\infty} B_1^\lambda\left((\mu_1,\mu), N^\lambda, T^\lambda\right) = \frac{1}{1 + \mu_1 \frac{b}{\exp(-\theta T)} - \frac{\mu_1}{\mu}} = \frac{\mu \exp(-\theta)T}{\mu \exp(-\theta T) + \mu_1\left(b\mu - \exp(-\theta T)\right)}.$$

□

L. Proof of Lemma 7.

We first establish the existence of a fixed point of $\hat{R}$, second show its uniqueness, and third derive the payment ratio P_R that results in the unique fixed point being μ.

Existence. As in the third paragraph of the proof of Proposition 3, a sufficient condition for the existence of a fixed point is that $\hat{U}(\mu_1,\mu)$ is quasiconcave in μ_1 on $[\underline{\mu},\overline{\mu}]$ for any fixed $\mu \in [\underline{\mu},\overline{\mu}]$. If $b\mu \le \exp(-\theta T)$, then $\hat{B}(\mu_1,\mu)=1$, meaning $\hat{U}(\mu_1,\mu)=P(\mu_1)$ for P defined in (EC.19), and so $\hat{U}$ is strictly concave (and therefore quasiconcave) by assumption on p. If $b\mu > \exp(-\theta T)$, then as in (EC.20),

$$\frac{\partial^2 \hat{U}(\mu_1,\mu)}{\partial \mu_1^2} = P^{''}(\mu_1)\hat{B}(\mu_1,\mu) + \frac{P(\mu_1)}{\hat{B}(\mu_1,\mu)}\left(\hat{B}(\mu_1,\mu)\frac{\partial^2 \hat{B}(\mu_1,\mu)}{\partial \mu_1^2} - 2\left(\frac{\partial \hat{B}(\mu_1,\mu)}{\partial \mu_1}\right)^2\right) + 2\frac{\partial U(\mu_1,\mu)}{\partial \mu_1}\frac{\frac{\partial \hat{B}(\mu_1,\mu)}{\partial \mu_1}}{\hat{B}(\mu_1,\mu)}.$$

Straightforward calculation shows

$$2\left(\frac{\partial \hat{B}(\mu_1,\mu)}{\partial \mu_1}\right)^2 - \hat{B}(\mu_1,\mu)\frac{\partial^2 \hat{B}(\mu_1,\mu)}{\partial \mu_1^2} = 0,$$

and so the same argument as in the paragraph surrounding Lemma EC.3 shows $\hat{U}(\mu_1,\mu)$ has at most one stationary point in $[\underline{\mu},\overline{\mu}]$, which is a local maximum, implying quasiconcavity in μ_1. Similar to the proof in of Proposition 3, we conclude that $\hat{R}_1(\mu)$ is continuous in μ, and a fixed point exists.

Uniqueness. If $b\mu \le \exp(-\theta T)$, the uniqueness follows because $\hat{U}$ is strictly concave. If $b\mu > \exp(-\theta T)$, then for $P_R = P_F/P_S$ the FOC is

$$\frac{\partial \hat{U}(\mu_1,\mu)}{\partial \mu_1} = \frac{\exp(-2\theta T)(1-P_R(1-p(\mu_1))) + \exp(-\theta T)\left(\mu_1(\exp(-\theta T) + (b-\exp(-\theta T)/\mu)\mu_1)P_R p^{'}(\mu_1)\right)}{\left(\exp(-\theta T) + (b-\exp(-\theta T)/\mu)\mu_1\right)^2} = 0. \tag{EC.40}$$

Define $F(\mu_1,\mu) := \exp(-2\theta T)(1-P_R(1-p(\mu_1))) + \exp(-\theta T)\left(\mu_1(\exp(-\theta T) + (b-\exp(-\theta T)/\mu)\mu_1)P_R p^{'}(\mu_1)\right)$. Since from the implicit function theorem

$$\frac{d\mu_1}{d\mu} = -\frac{\frac{\partial F(\mu_1,\mu)}{\partial \mu}}{\frac{\partial F(\mu_1,\mu)}{\partial \mu_1}},$$

and

$$\begin{aligned}
\frac{\partial F(\mu_1,\mu)}{\partial \mu_1} &= \exp(-\theta T)P_R(\exp(-\theta T)+(b-\exp(-\theta T)/\mu)\mu_1)\times(p(\mu_1)\mu_1)^{''}<0\\
\frac{\partial F(\mu_1,\mu)}{\partial \mu} &= \frac{\exp(-2\theta T)\mu_1^2 P_R p^{'}(\mu_1)}{\mu^2}<0,
\end{aligned}$$

we conclude

$$\frac{d\mu_1}{d\mu}<0,$$

which implies (EC.40) has a unique solution.

Payment Ratio. We solve for the parameter under which

$$\left.\frac{\partial \hat{U}(\mu_1,\mu)}{\partial \mu_1}\right|_{\mu_1=\mu}=0. \tag{EC.41}$$

Noting that

$$\left.\frac{\partial \hat{U}(\mu_1,\mu)}{\partial \mu_1}\right|_{\mu_1=\mu}=\begin{cases} P^{'}(\mu)=P_S(1-P_R(1-p(\mu)-P^{'}(\mu)\mu)) & \text{if } b^{\star}\mu^{\star}\le\exp(-\theta T^{\star}),\\ \frac{F(\mu,\mu)}{(\exp(-\theta T)+b\mu-\exp(-\theta T))^2} & \text{if } b^{\star}\mu^{\star}>\exp(-\theta T^{\star}),\end{cases}$$

we find that the P_R defined in the statement of Lemma 7 solves (EC.41). □

M. Proof of Theorem 2.

As observed directly before the theorem statement, $\mu_E^{\lambda}=\mu_F^{\lambda}(P_R^{\star})$ is a symmetric equilibrium service rate (not necessarily unique). Furthermore, for each λ, the policy $\left(\hat{\mu}_{\star},N^{\lambda}=\hat{b}_{\star}\lambda+o(\lambda),\hat{T}_{\star}\right)$ that is asymptotically optimal by Theorem 1 under Assumptions 1 and 2 is such that $\left(\mu_F^{\lambda},N^{\lambda}=\hat{b}_{\star}\lambda+o(\lambda),\hat{T}_{\star}\right)$ satisfies the constraints of the decentralized control problem (10). From the definition of P_S^{λ} and P_F^{λ} in (25)-(26), $U_i^{\lambda}=c_S$ for all λ and so $\lim_{\lambda\to\infty}U_i^{\lambda}-c_S=0, i\in\mathcal{N}_{N^{\lambda}}$ holds trivially. To complete the proof, we must show any sequence of symmetric equilibrium service rates satisfies

$$\lim_{\lambda\to\infty}\left|\mu_F^{\lambda}-\hat{\mu}_{\star}\right|=0.$$

The proof is by contradiction. Suppose not. Then, there exists a subsequence λ_i on which $\mu_F^{\lambda_i}$ does not converge to $\hat{\mu}_{\star}$. Since $\mu_F^{\lambda_i}\in[\underline{\mu},\overline{\mu}]$ is a bounded sequence, it fails to converge to $\hat{\mu}_{\star}$ either because it converges to $\tilde{\mu} \neq \hat{\mu}_\star$, or because it alternates - in which case there exists a further subsequence λ_{i_j} on which $\mu_F^{\lambda_{i_j}} \to \tilde{\mu} \neq \hat{\mu}_\star$ as $\lambda_{i_j} \to \infty$. Hence we may assume there exists a subsequence λ_k on which $\mu_F^{\lambda_k} \to \tilde{\mu} \neq \hat{\mu}_\star$ as $\lambda_k \to \infty$. For each λ_k, from the definition of an equilibrium,

$$P(\mu_F^{\lambda_k}) B^{\lambda_k}(\mu_F^{\lambda_k}, \mu_F^{\lambda_k}) \geq P(\mu_1) B^{\lambda_k}(\mu_1, \mu_F^{\lambda_k}), \text{ for all } \mu_1 \in [\underline{\mu}, \overline{\mu}]. \tag{EC.42}$$

Taking limits in the above display and applying Proposition 4 shows that

$$P(\tilde{\mu}) \hat{B}(\tilde{\mu}, \tilde{\mu}) \geq P(\mu_1) \hat{B}(\mu_1, \tilde{\mu}), \text{ for all } \mu_1 \in [\underline{\mu}, \overline{\mu}],$$

which implies $\tilde{\mu}$ is also a fixed point of the approximating best response function $\hat{R}$. This contradicts the uniqueness of $\hat{\mu}_\star$ shown in Lemma 7. □