Routing and Staffing When Servers Are Strategic

Ragavendran Gopalakrishnan  
Xerox Research Centre India, Bangalore, Karnataka 560103, India, ragavendran.gopalakrishnan@xerox.com

Sherwin Doroudi  
Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, sdoroudi@andrew.cmu.edu

Amy R. Ward  
Marshall School of Business, University of Southern California, Los Angeles, California 90089, amyward@marshall.usc.edu

Adam Wierman  
Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California 91125, adanw@caltech.edu

Traditionally, research focusing on the design of routing and staffing policies for service systems has modeled servers as having fixed (possibly heterogeneous) service rates. However, service systems are generally staffed by people. Furthermore, people respond to workload incentives; that is, how hard a person works can depend both on how much work there is and how the work is divided between the people responsible for it. In a service system, the routing and staffing policies control such workload incentives; and so the rate servers work will be impacted by the system’s routing and staffing policies. This observation has consequences when modeling service system performance, and our objective in this paper is to investigate those consequences.

We do this in the context of the M/M/1 queue, which is the canonical model for large service systems. First, we present a model for “strategic” servers that choose their service rate to maximize a trade-off between an “effort cost,” which captures the idea that servers exert more effort when working at a faster rate, and a “value of idleness,” which assumes that servers value having idle time. Next, we characterize the symmetric Nash equilibrium service rate under any routing policy that routes based on the server idle time (such as the longest idle server first policy). We find that the system must operate in a quality-driven regime, in which servers have idle time, for an equilibrium to exist. The implication is that to have an equilibrium solution the staffing must have a first-order term that strictly exceeds that of the common square-root staffing policy. Then, within the class of policies that admit an equilibrium, we (asymptotically) solve the problem of minimizing the total cost, when there are linear staffing costs and linear waiting costs. Finally, we end by exploring the question of whether routing policies that are based on the service rate, instead of the server idle time, can improve system performance.

Keywords: service systems; staffing; routing; scheduling; routing; strategic servers.

Subject classifications: queues: applications, limit theorem; games/group decisions: noncooperative.

Area of review: Stochastic Models.

History: Received January 27, 2014; revised December 25, 2014, July 14, 2015; accepted March 4, 2016. Published online in Articles in Advance June 17, 2016.

1. Introduction

There is a broad and deep literature studying the scheduling and staffing of service systems that bridges operations research, applied probability, and computer science. This literature has had, and is continuing to have, a significant practical impact on the design of call centers (see, for example, the survey papers Gans et al. 2003 and Aksin et al. 2007), healthcare systems (see, for example, the recent book Harchol-Balter 2013), and large-scale computing systems (see, for example, the recent book Harchol-Balter 2013), among other areas. Traditionally, this literature on scheduling and staffing has modeled the servers of the system as having fixed (possibly heterogeneous) service rates and then, given these rates, scheduling and staffing policies are proposed and analyzed. However, in reality, when the servers are people, the rate a server chooses to work can be, and often is, impacted by the scheduling and staffing policies used by the system.

For example, if requests are always scheduled to the “fastest” server whenever that server is available, then this server may have the incentive to slow her rate to avoid being overloaded with work. Similarly, if extra staff is always assigned to the division of a service system that is the busiest, then servers may have the incentive to reduce their service rates to ensure their division is assigned the extra staff. The previous two examples are simplistic; however, strategic behavior has been observed in practice in service systems. For example, empirical data from call centers shows many calls that last near zero seconds (Gans et al. 2003). This strategic behavior of the servers allowed them to obtain “rest breaks” by hanging up on customers—a rather dramatic means of avoiding being overloaded with
work. For another example, academics are often guilty of strategic behavior when reviewing for journals. It is rare for reviews to be submitted before an assigned deadline since, if someone is known for reviewing papers very quickly, then they are likely to be assigned more reviews by the editor.

Clearly, the strategic behavior illustrated by the preceding examples can have a significant impact on the performance provided by a service system. One could implement a staffing or scheduling policy that is provably optimal under classical scheduling models, where servers are non-strategic, and end up with far from optimal system performance as a result of undesirable strategic incentives created by the policy. Consequently, it is crucial for service systems to be designed in a manner that provides the proper incentives for such “strategic servers.”

In practice, there are two approaches used for creating the proper incentives for strategic servers: one can either provide structured bonuses for employees depending on their job performance (performance-based payments), or one can provide incentives in how scheduling and staffing are performed that reward good job performance (incentive-aware scheduling). While there has been considerable research on how to design performance-based payments in the operations management and economics communities, the incentives created by scheduling and staffing policies are much less understood. In particular, the goal of this paper is to initiate the study of incentive-aware scheduling and staffing policies for strategic servers.

The design of incentive-aware scheduling and staffing policies is important for a wide variety of service systems. In particular, in many systems performance-based payments such as bonuses are simply not possible, e.g., in service systems staffed by volunteers such as academic reviewing. Furthermore, many service systems do not use performance-based compensation schemes; for example, the 2005 benchmark survey on call center agent compensation in the United States shows that a large fraction of call centers pay a fixed hourly wage (and have no performance-based compensation); see Anton (2005).

Even when performance-based payments are possible, the incentives created by scheduling and staffing policies impact the performance of the service system, and thus they impact the success of performance-based payments. Further, since incentive-aware scheduling and staffing do not involve monetary payments (beyond a fixed employee salary), it may be less expensive to provide incentives through scheduling and staffing than through monetary bonuses. Additionally, providing incentives through scheduling and staffing eliminates many concerns about “unfairness” that stem from differential payments to employees.

Of course, the discussion above assumes that the incentives created by scheduling and staffing can be significant enough to impact the behavior. A priori it is not clear if they are, since simply changing the scheduling and staffing policies may not provide strong enough incentives to strategic servers to significantly change service rates and thus change system performance. It is exactly this uncertainty that motivates the current paper, which seeks to understand the impact of the incentives created by scheduling and staffing and then to design incentive-aware staffing and scheduling policies that provide near-optimal system performance without the use of monetary incentives.

1.1. Contributions of This Paper

This paper makes three main contributions. We introduce a new model for the strategic behavior of servers in large service systems and, additionally, we initiate the study of staffing and routing in the context of strategic servers. Each of these contributions is described in the following.

Modeling Strategic Servers (Sections 2 and 3). The essential first step for an analysis of strategic servers is a model for server behavior that is simple enough to be analytically tractable and yet rich enough to capture the salient influences on how each server may choose her service rate. Our model is motivated by work in labor economics that identifies two main factors that impact the utility of agents: effort cost and idleness. More specifically, it is common in labor economics to model agents as having some “effort cost” function that models the decrease in utility which comes from an increase in effort (Cahuc and Zylberberg 2004). Additionally, it is a frequent empirical observation that agents in service systems engage in strategic behavior to increase the amount of idle time they have (Gans et al. 2003). The key feature of the form of the utility we propose in Section 2 is that it captures the inherent trade-off between idleness and effort. In particular, a faster service rate would mean quicker completion of jobs and might result in a higher idle time, but it would also result in a higher effort cost.

In Section 3 of this paper we apply our model in the context of a \( M/M/1 \) system, analyzing the first order condition, and provide a necessary and sufficient condition for a solution to the first order condition to be a symmetric equilibrium service rate (Theorem 4). In addition, we discuss the existence of solutions to the first order condition and provide a sufficient condition for a unique solution (Theorem 5). These results are necessary to study staffing and routing decisions, as we do in Sections 4 and 5; however, it is important to note that the model is applicable more generally as well.

Staffing Strategic Servers (Section 4). The second piece of the paper studies the impact strategic servers have on staffing policies in multiserver service systems. The decision of a staffing level for a service system has a crucial impact on the performance of the system. As such, there is a large literature focusing on this question in the classical, nonstrategic, setting, and the optimal policy is well understood. In particular, the number of servers that must be staffed to ensure stability in a conventional \( M/M/1 \) queue with arrival rate \( \lambda \) and fixed service rate \( \mu \) should be strictly
larger than the offered load, $\lambda/\mu$. However, when there are linear staffing and waiting costs, the economically optimal number of servers to staff is more. Specifically, the optimal policy employs the square root of the offered load more servers (Borst et al. 2004). This results in efficient operation, because the system loading factor $\lambda/(N\mu)$ is close to one, and the quality of service is maintained because the customer wait times are small (on the order of $1/\sqrt{\lambda}$). Thus, this is often referred to as the Quality and Efficiency Driven (QED) regime or as square-root staffing.

Our contribution in this paper is to initiate the study of staffing strategic servers. In the presence of strategic servers, the offered load depends on the arrival rate, the staffing, and the routing, through the servers’ choice of their service rate. We show that an equilibrium service rate exists only if the number of servers staffed is order $\lambda$ more than the aforementioned square-root staffing (Theorem 7). In particular, the system must operate in a quality-driven regime, in which the servers have idle time, instead of the quality-and-efficiency driven regime that arises under square-root staffing, in which servers do not have idle time. Then, within the set of policies that admit an equilibrium service rate, we (asymptotically) solve the problem of minimizing the total cost, when there are linear staffing costs and linear waiting costs (Theorem 8).

Routing to Strategic Servers (Section 5). The final piece of this paper studies the impact of strategic servers on the design of scheduling policies in multiserver service systems. When servers are not strategic, how to schedule (dispatch) jobs to servers in multiserver systems is well understood. In particular, the most commonly proposed policies for this setting include Fastest Server First (FSF), which dispatches arriving jobs to the idle server with the fastest service rate; Longest Idle Server First (LISF), which dispatches jobs to the server that has been idle for the longest period of time; and Random, which dispatches the job to each idle server with equal probability. When strategic servers are not considered, FSF is the natural choice for reducing the mean response time (though it is not optimal in general; see de Véricourt and Zhou 2005, Lin and Kumar 1984). However, in the context of strategic servers the story changes. In particular, we prove that FSF has no symmetric equilibria when strategic servers are considered, even when there are just two servers. Further, we prove that LISF, a commonly suggested policy for call centers because of its fairness properties, has the same, unique, symmetric equilibrium as random dispatching. In fact, we prove that there is a large policy-space collapse—all routing policies that are idle-time-order-based are equivalent in a very strong sense (Theorem 9).

With this in mind, one might suggest that Slowest Server First (SSF) would be a good dispatch policy, since it could incentivize servers to work fast; however, we prove that, like FSF, SSF has no symmetric equilibria (Theorem 10). However, by “softening” SSF’s bias toward slow servers, we are able to identify policies that are guaranteed to have a unique symmetric equilibrium and provide mean response times that are smaller than that under LISF and Random (Theorem 11).

A key message provided by the results described above is that scheduling policies must carefully balance two conflicting goals in the presence of strategic servers: making efficient use of the service capacity (e.g., by sending work to fast servers) while still incentivizing servers to work fast (e.g., by sending work to slow servers). While these two goals are inherently in conflict, our results show that it is possible to balance them in a way that provides improved performance over Random.

1.2. Related Work

As we have already described, the question of how to route and staff in many-server systems when servers have fixed, nonstrategic, service rates is well studied. In general, this is a very difficult question, because the routing depends on the staffing and vice versa. However, when all the servers serve at the same rate, the routing question is moot. Then, Borst et al. (2004) show that square-root staffing, first introduced in Erlang (1948) and later formalized in Halfin and Whitt (1981), is economically optimal when both staffing and waiting costs are linear. Furthermore, square root staffing is remarkably robust: there is theoretical support for why it works so well for systems of moderate size (Janssen et al. 2011), and it continues to be economically optimal both when abandonment is added to the $M/M/N$ model (Garnett et al. 2002) and when there is uncertainty in the arrival rate (Kocaga et al. 2015). Hence, to study the joint routing and staffing question for more complex systems, that include heterogeneous servers that serve at different rates and heterogeneous customers, many authors have assumed square root staffing and show how to optimize the routing for various objective functions (see, for example, Armony 2005, Gurvich and Whitt 2007, Atar 2005, Tezcan 2008, Tezcan and Dai 2010). In relation to this body of work, this paper shows that scheduling and routing results for classical many-server systems that assume fixed service rates must be revisited when servers exhibit strategic behavior. This is because they may not admit a symmetric equilibrium service rate in the case of square-root staffing (see Section 4) or be feasible in the case of Fastest Server First routing (see Section 5).

Importantly, the Fastest Server First routing policy mentioned earlier has already been recognized to be potentially problematic because it may be perceived as “unfair.” The issue from an operational standpoint is that there is strong indication in the human resource management literature that the perception of fairness affects employee performance (see Colquitt et al. 2001, Cohen-Charash and Spector 2001). This has motivated the analysis of “fair” routing policies that, for example, equalize the cumulative server idleness (Atar et al. 2011, Reed and Shaki 2014), and the desire to find an optimal “fair” routing policy (Armony and Ward 2010, Ward and Armony 2013).
Another approach is to formulate a model in which the servers choose their service rate to balance their desire for idle time (which is obtained by working faster) and the exertion required to serve faster. This leads to a noncooperative game for a $M/M/N$ queue in which the servers act as strategic players that selfishly maximize their utility.

Finally, the literature that is, perhaps, most closely related to the current paper is the literature on queueing games, which is surveyed in Hassin and Haviv (2003). The bulk of this literature focuses on the impact of customers acting strategically (e.g., deciding whether to join and which queue to join) on queueing performance. Still, there is a body of work within this literature that considers settings where servers can choose their service rate, e.g., Kalai et al. (1992), Gilbert and Weng (1998), Cachon and Harker (2002), Cachon and Zhang (2007). However, in all of the aforementioned papers, there are two servers that derive utility from some monetary compensation per job or per unit of service that they provide, and there are no staffing decisions. In contrast, our work considers systems with more than two servers, and considers servers that derive utility from idle time (and have a cost of effort). The idea that servers value idle time is most similar to the setting in Geng et al. (2013), but that paper restricts its analysis to a two-server model. Perhaps the closest previous work to the current paper in analysis spirit is Allon and Gurvich (2010), which characterizes approximate equilibria in a market with many servers that compete on price and service level. However, this is similar in theme to Kalai et al. (1992), Cachon and Harker (2002) in the sense that they consider servers as competing firms in a market. This contrasts with the current paper, where our focus is on competition between servers within the same firm.

2. A Model for Strategic Servers

The objective of this paper is to initiate an investigation into the effects of strategic servers on classical management decisions in service systems, e.g., staffing and routing. We start by, in this section, describing formally our model for the behavior of a strategic server.

The term “strategic server” could be interpreted in many ways depending on the server’s goal. Thus, the key feature of the model is the utility function for a strategic server. Our motivation comes from a service system staffed by people who are paid a fixed wage, independent of performance. In such settings, there are two key factors to the behavior of a strategic server.

\[ U_i(\mu) = I_i(\mu) - c(\mu), \quad i \in \{1, \ldots, N\}, \tag{1} \]

where $\mu$ is a vector of the rate of work chosen by each server (i.e., the service rate vector), $I_i(\mu)$ is the time-average idle time experienced by server $i$ given the service rate vector $\mu$, and $c(\mu)$ is the effort cost of server $i$. We take $c$ to be an increasing, convex function that is the same for all servers. We assume that the strategic behavior of servers (choosing a utility-maximizing service rate) is independent of the state of the system and that the server has complete information about the steady state properties of the system when choosing a rate, i.e., they know the arrival rate, scheduling policy, staffing policy, etc., and thus can optimize $U_i(\mu)$.

The key feature of the form of the utility in (1) is that it captures the inherent trade-off between idleness and effort. The idleness, and hence the utility, is a steady state quantity. In particular, a faster service rate would mean quicker completion of jobs and might result in higher idle time in steady state, but it would also result in a higher effort cost. This trade-off then creates a difficult challenge for staffing and routing in a service system. To increase throughput and decrease response times, one would like to route requests to the fastest servers, but by doing so the utility of servers decreases, making it less desirable to maintain a fast service rate. Our model should be interpreted as providing insight into the systemic incentives created by scheduling and staffing policies rather than the transitive incentives created by the stochastic behavior of the system.

Our focus in this paper will be to explore the consequences of strategic servers for staffing and routing in large service systems, specifically, in the $M/M/N$ setting. However, the model is generic and can be studied in non-queueing contexts as well.

To quickly illustrate the issues created by strategic servers, a useful example to consider is that of an $M/M/1$ queue with a strategic server.

**Example 1 (The $M/M/1$ Queue with a Strategic Server).** In a classic $M/M/1$ system, jobs arrive at rate $\lambda$ into a queue with an infinite buffer, where they wait to obtain service from a single server having fixed service rate $\mu$. When the server is strategic, instead of serving at a fixed rate $\mu$, the server chooses her service rate $\mu > \lambda$ to maximize the utility in (1). To understand what service rate will emerge, recall that in a $M/M/1$ queue with $\mu > \lambda$ the steady state fraction of time that the server is idle is given by $I(\mu) = 1 - \lambda/\mu$. Substituting this expression into (1) means that the utility of the server is given by the following concave function:

\[ U(\mu) = 1 - \frac{\lambda}{\mu} - c(\mu). \]

We now have two possible scenarios. First, suppose that $c'(\lambda) > 1/\lambda$, so that the cost function does not increase too fast. Then, $U(\mu)$ attains a maximum in $(\lambda, \infty)$ at a unique point $\mu^*$, which is the optimal (utility maximizing) operating point for the strategic server. Thus, a stable operating
point emerges, and the performance of this operating point can be derived explicitly when a specific form of a cost function is considered.

On the other hand, if \( c'(\lambda) \geq 1/\lambda \), then \( U(\mu) \) is strictly decreasing in \( (\lambda, \infty) \) and hence does not attain a maximum in this interval. We interpret this case to mean that the server’s inherent skill level (as indicated by the cost function) is such that the server must work extremely hard just to stabilize the system and therefore should not have been hired in the first place.

For example, consider the class of cost functions \( c(\mu) = c_E \mu^p \). If \( c(\lambda) < 1/p \), then \( \mu^* \) solves \( \mu^* c(\mu^*) = \lambda/p \), which gives \( \mu^* = (\lambda/(c_E p))^{1/(p+1)} > \lambda \). On the other hand, if \( c(\lambda) \geq 1/p \), then \( U(\mu) \) is strictly decreasing in \( (\lambda, \infty) \) and hence does not attain a maximum in this interval.

Before moving on to the analysis of the \( M/M/N \) model with strategic servers, it is important to point out that the model we study focuses on a linear trade-off between idleness and effort. There are certainly many generalizations that are interesting to study in future work. One particularly interesting generalization would be to consider a concave (and increasing) function of idle time in the utility function, since it is natural that the gain from improving idle time from 10%–20% would be larger than the gain from improving idle time from 80%–90%. A preliminary analysis highlights that the results in this paper would not qualitatively change in this context.\(^1\)

3. The \( M/M/N \) Queue with Strategic Servers

Our focus in this paper is on the staffing and routing decisions in large service systems, and so we adopt a classical model of this setting, the \( M/M/N \), and adjust it by considering strategic servers, as described in Section 2. The analysis of staffing and routing policies is addressed in Sections 4 and 5, but before moving to such questions, we start by formally introducing the \( M/M/N \) model and performing some preliminary analysis that is useful both in the context of staffing and routing.

3.1. Model and Notation

In a \( M/M/N \) queue, customers arrive to a service system having \( N \) servers according to a Poisson process with rate \( \lambda \). Delayed customers (those that arrive to find all servers busy) are served according to the First In First Out (FIFO) discipline. Each server is fully capable of handling any customer’s service requirements. The time required to serve each customer is independent and exponential, and has a mean of one time unit when the server works at rate one. However, each server strategically chooses her service rate to maximize her own (steady state) utility, and so it is not a priori clear what the system service rates will be.

In this setting, the utility functions that the servers seek to maximize are given by

\[
U_i(\mu; \lambda, N, R) = I_i(\mu; \lambda, N, R) - c(\mu),
\]

where \( \mu \) is the vector of service rates, \( \lambda \) is the arrival rate, \( N \) is the number of servers (staffing level), and \( R \) is the routing policy. \( I_i(\mu; \lambda, N, R) \) is the steady state fraction of time that server \( i \) is idle. \( c(\mu) \) is an increasing, convex function with \( c''(\mu) \geq 0 \), that represents the server effort cost.

Note that, as compared with (1), we have emphasized the dependence on the arrival rate \( \lambda \), staffing level \( N \), and routing policy of the system, \( R \). In the remainder of this article, we expose or suppress the dependence on these additional parameters as relevant to the discussion. In particular, note that the idle time fraction \( I_i \) (and hence, the utility function \( U_i \)) in (2) depends on how arriving customers are routed to the individual servers.

There are a variety of routing policies that are feasible for the system manager. In general, the system manager may use information about the order in which the servers became idle, the rates at which servers have been working, etc. This leads to the possibility of using simple policies such as Random, which chooses an idle server to route to uniformly at random, as well as more complex policies such as Longest/Shortest Idle Server First (LISF/SISF) and Fastest/Slowest Server First (FSF/SSF). We study the impact of this decision in detail in Section 5.

Given the routing policy chosen by the system manager and the form of the server utilities in (2), the situation that emerges is a competition among the servers for the system idle time. In particular, the routing policy yields a division of idle time among the servers, and both the division and the amount of idle time will depend on the service rates chosen by the servers.

As a result, the servers can be modeled as strategic players in a noncooperative game, and thus the operating point of the system is naturally modeled as an equilibrium of this game. In particular, a Nash equilibrium of this game is a set of service rates \( \mu^* \), such that

\[
U_i(\mu^*_i, \mu^*_{i-1}, \ldots, \mu^*_{i+1}, \mu^*_{N-1}, \mu^*_N; R) = \max_{\mu_i \geq \lambda/N} U_i(\mu_i, \mu^*_{i-1}, \ldots, \mu^*_{i+1}, \mu^*_{N-1}, \mu^*_N; R),
\]

where \( \mu^*_{-i} = (\mu^*_1, \ldots, \mu^*_{i-1}, \mu^*_{i+1}, \ldots, \mu^*_N) \) denotes the vector of service rates of all the servers except server \( i \). Note that we exogenously impose the (symmetric) constraint that each server must work at a rate strictly greater than \( \lambda/N \) to define a product action space that ensures the stability of the system.\(^2\) Such a constraint is necessary to allow steady state analysis and does not eliminate any feasible symmetric equilibria. We treat this bound as exogenously fixed; however, in some situations a system manager may wish to impose quality standards on servers, which would correspond to imposing a larger lower bound (likely with correspondingly larger payments for servers). Investigating
the impact of such quality standards is an interesting topic for future work.

Our focus in this paper is on symmetric Nash equilibria. With a slight abuse of notation, we say that $\mu^*$ is a symmetric Nash equilibrium if $\mu^* = (\mu^*, \ldots, \mu^*)$ is a Nash equilibrium (solves (3)). Throughout, the term “equilibrium service rate” means a symmetric Nash equilibrium service rate.

We focus on symmetric Nash equilibria for two reasons. First, because the agents we model intrinsically have the same skill level (as quantified by the effort cost functions), a symmetric equilibrium corresponds to a fair outcome. As we have already discussed, this sort of fairness is often crucial in service organizations (Colquitt et al. 2001, Cohen-Charash and Spector 2001, Armony and Ward 2010). A second reason for focusing on symmetric equilibria is that analyzing symmetric equilibria is already technically challenging, and it is not clear how to approach asymmetric equilibria in the contexts that we consider. Note that we do not rule out the existence of asymmetric equilibria; in fact, they likely exist, and it would be interesting to study whether they lead to better or worse system performance than their symmetric counterparts.

### 3.2. The M|M|N Queue with Strategic Servers and Random Routing

Before analyzing staffing and routing in detail, we first study the M|M|N queue with strategic servers and Random routing. We focus on Random routing first because it is, perhaps, the most commonly studied policy in the classical literature on nonstrategic servers. Further, this importance is magnified by a new “policy-space collapse” result included in Section 5.1.1, which shows that all idle-time-order-based routing policies (e.g., LISF and SISF) have equivalent steady state behavior, and thus have the same steady state behavior as Random routing. We stress that this result stands on its own in the classical, nonstrategic setting of a M|M|N queue with heterogeneous service rates, but it is also crucial to analyze routing to strategic servers (Section 5).

The key goal in analyzing a queueing system with strategic servers is to understand the equilibria service rates, i.e., show conditions that guarantee their existence and characterize the equilibria when they exist. Theorems 4 and 5 of Section 3.2.2 summarize these results for the M|M|N queue with Random routing. However, to obtain such results we must first characterize the idle time in a M|M|N system to be able to understand the “best responses” for servers and thus analyze their equilibrium behavior. Such an analysis is the focus of Section 3.2.1. All proofs are deferred to the technical appendix (available as supplemental material at http://dx.doi.org/10.1287/opre.2016.1506).

#### 3.2.1. The Idle Time of a Tagged Server

To characterize the equilibria service rates, a key first step is to understand the idle time of a M|M|N queue. This is, of course, a well-studied model, and so one might expect to be able to use off-the-shelf results. While this is true when the servers are homogeneous (i.e., all the server rates are the same), for heterogeneous systems, closed form expressions are challenging to obtain in general, and the resulting forms are quite complicated; see Gumbel (1960).

To characterize equilibria, we do need to understand the idle time of heterogeneous M|M|N queues. However, because of our focus on symmetric equilibria, we only need to understand a particular, mild, form of heterogeneity. In particular, we need only understand the best response function for a “deviating server” when all other servers have the same service rate. Given this limited form of heterogeneity, the form of the idle time function simplifies, but still remains quite complicated, as the following theorem shows.

**Theorem 1.** Consider a heterogeneous M|M|N system with Random routing and arrival rate $\lambda > 0$, where $N - 1$ servers operate at rate $\mu > \lambda/N$, and a tagged server operates at rate $\mu_1 > \mu$. The steady state probability that the tagged server is idle is given by

$$I(\mu_1, \mu, \lambda, N) = \left(1 - \frac{\rho}{N}\right) \left(1 - \frac{\mu}{\mu_1}\right) \left(1 + \frac{\text{ErlC}(N, \rho)}{N - (N - 1 - \mu_1/\mu)}\right)^{-1},$$

where $\rho = \lambda/\mu$, and $\text{ErlC}(N, \rho)$ denotes the Erlang C formula, given by

$$\text{ErlC}(N, \rho) = \frac{(\rho^N/N!)(N/(N - \rho))}{\sum_{j=0}^{N-1} (\rho^j/j!)+ (\rho^N/N!)(N/(N - \rho))}.$$
convex at \( \mu_1 = \mu^\star \) to begin with, but if so, then as \( \mu_1 \) increases, it steadily becomes less convex and is eventually concave. This behavior adds considerable complication to the equilibrium analysis.

**Theorem 3.** The second derivative of the idle time satisfies the following properties.

(a) There exists a threshold \( \mu^\star_1 \in [\mu^\star_1, \infty) \) such that \( \partial^2 I / \partial \mu^2_1 > 0 \) for \( \mu^\star_1 < \mu_1 < \mu^\star_1 \), and \( \partial^2 I / \partial \mu^2_1 < 0 \) for \( \mu^\star_1 < \mu_1 < \infty \);

(b) \( s \partial^2 I / \partial \mu^2_1 > 0 \) \( \Rightarrow \partial^2 I / \partial \mu^3_1 < 0 \).

We remark that it is possible that the threshold \( \mu^\star_1 \) could be greater than \( \lambda / N \), so, restricting the service rate of server 1 to be greater than \( \lambda / N \) does not necessarily simplify the analysis.

**3.2.2. Symmetric Equilibrium Analysis for a Finite System.** The properties of the idle time function derived in the previous section provide the key tools we need to characterize the symmetric equilibria service rates under Random routing for a \( M/M/N \) system.

To characterize the symmetric equilibria, we consider the utility of a tagged server, without loss of generality, server 1, under the mildly heterogeneous setup of Theorem 1. We denote it by

\[
U(\mu_1, \mu; \lambda, N) = I(\mu_1, \mu; \lambda, N) - c(\mu_1). \tag{7}
\]

For a symmetric equilibrium in \((\lambda / N, \infty)\), we explore the first- and second-order conditions for \( U \) as a function of \( \mu_1 \) to have a maximum in \((\mu^\star_1, \infty)\).

The first order condition for an interior local maximum at \( \mu_1 \) is given by

\[
\frac{\partial U}{\partial \mu_1} = 0 \implies \frac{\partial I}{\partial \mu_1} = c'(\mu_1). \tag{8}
\]

Since we are interested in a symmetric equilibrium, we analyze the symmetric first order condition, obtained by plugging in \( \mu_1 = \mu \) in (8):

\[
\frac{\partial U}{\partial \mu_1} \bigg|_{\mu_1=\mu} = 0 \implies \frac{\lambda}{N^2 \mu^2} \left( N - \lambda + \text{ErlC} \left( N, \frac{\lambda}{\mu} \right) \right) = c'(\mu). \tag{9}
\]

Now, suppose that \( \mu^\star > \lambda / N \) satisfies the symmetric first order condition (9). Then, \( \mu_1 = \mu^\star \) is a stationary point of \( U(\mu_1, \mu^\star) \). It follows then, that \( \mu^\star \) will be a symmetric equilibrium for the servers (satisfying (3)) if and only if \( U(\mu_1, \mu^\star) \) attains a global maximum at \( \mu_1 = \mu^\star \) in the interval \((\lambda / N, \infty)\). While an obvious necessary condition for this is that \( U(\mu^\star, \mu^\star) \geq U(\lambda / N, \mu^\star) \), we show, perhaps surprisingly, that it is also sufficient, in the following theorem.

**Theorem 4.** \( \mu^\star > \lambda / N \) is a symmetric equilibrium if and only if it satisfies the symmetric first order condition (9), and the inequality \( U(\mu^\star, \mu^\star) \geq U(\lambda / N, \mu^\star) \), i.e.,

\[
c(\mu) \leq c \left( \frac{\lambda}{N} + \left( 1 - \frac{\rho}{N} \right) \right) \left( 1 + \left( 1 - \frac{\rho}{N} + \text{ErlC}(\lambda, \rho) \right)^{-1} \right)^{-1}. \tag{10}
\]

Finally, we need to understand when the symmetric first order condition (9) admits a feasible solution \( \mu^\star > \lambda / N \). Towards that, we present sufficient conditions for at least one feasible solution, as well as for a unique feasible solution.

**Theorem 5.** If \( c'(\lambda / N) < 1 / \lambda \), then the symmetric first order condition (9) has at least one solution for \( \mu \) in \((\lambda / N, \infty)\). In addition, if \( 2(\lambda / N)c'(\lambda / N) + (\lambda / N)^2 \cdot c''(\lambda / N) \geq 1 \), then the symmetric first order condition (9) has a unique solution for \( \mu \) in \((\lambda / N, \infty)\).

In the numerical results that follow, we see instances of zero, one, and two equilibria. Interestingly, when more than one equilibrium exists, the equilibrium with the largest service rate, which leads to best system performance, also leads to highest server utility, and hence is also most preferred by the servers, as the following theorem shows.

**Theorem 6.** If the symmetric first order condition (9) has two solutions, say \( \mu^1_1 \) and \( \mu^2_1 \), with \( \mu^1_1 > \mu^2_1 > \lambda / N \), then \( U(\mu^1_1, \mu^1_1) > U(\mu^2_1, \mu^2_1) \).

**3.3. Numerical Examples**

Because of the complexity of the expression for the equilibrium service rate(s) given by the first order condition (9) and the possibility of multiple equilibria, we discuss a few numerical examples here in order to provide intuition. In addition, we point out some interesting characteristics that emerge as a consequence of strategic server behavior.

We present two sets of graphs below: one that varies the arrival rate \( \lambda \) while holding the staffing level fixed at \( N = 20 \) (Figure 1), and one that varies the staffing level \( N \) while holding the arrival rate fixed at \( \lambda = 2 \) (Figure 2). In each set, we plot the following two equilibrium quantities: (a) service rates, and (b) mean steady state waiting times. Note that the graphs in Figure 2 only show data points corresponding to integer values of \( N \); the thin line through these points is only meant as a visual tool that helps bring out the pattern. Each of the four graphs shows data for three different effort cost functions: \( c(\mu) = \mu, c(\mu) = \mu^2 \), and \( c(\mu) = \mu^3 \), which are depicted in red, blue, and green, respectively. The data points in Figure 2 marked \( \times \) and \( \circ \) correspond to special staffing levels \( N = 20 \) and \( N = 50 \) respectively, which are introduced later, in Section 4.

The first observation we make is that there are at most two equilibria. Further, for large enough values of the minimum service rate \( \lambda / N \), there is no equilibrium. (In Figure 1(a) where \( N \) is fixed, this happens for large \( \lambda \), and...
Gopalakrishnan et al.: Routing and Staffing When Servers Are Strategic
Operations Research 64(4), pp. 1033–1050, © 2016 INFORMS

Figure 1. (Color online) Equilibrium behavior as a function of the arrival rate when the staffing level is fixed at \( N = 20 \), for three different effort cost functions: linear, quadratic, and cubic.

(a) Service rates

(b) Mean steady state waiting times

Note. The dotted line in (a) is \( \mu = \lambda/N = \lambda/20 \).

Figure 2. (Color online) Equilibrium behavior as a function of the staffing level when the arrival rate is fixed at \( \lambda = 2 \), for three different effort cost functions: linear, quadratic, and cubic.

(a) Service rates

(b) Mean steady state waiting times

Notes. The dotted curve in (a) is \( \mu = \lambda/N = 2/N \). The data points marked \( \times \) and \( \circ \) correspond to \( N = 10^{0.2} \) and \( N = 10^{0.2} \) respectively.

in Figure 2(a) where \( \lambda \) is fixed, this happens for small \( N \). On the other hand, when the minimum service rate \( \lambda/N \) is small enough, there is a unique equilibrium; for this range, even if the symmetric first order condition (9) has another solution greater than \( \lambda/N \), it fails to satisfy (10). If an intermediate value of \( \lambda/N \) is small enough for (9) to have two feasible solutions, but not too small so that both solutions satisfy (10), then there are two equilibria.

The second observation we make is that the two equilibria have very different behaviors. As illustrated in Figure 1(a), the larger equilibrium service rate first increases and then decreases while the corresponding mean steady state waiting time in Figure 1(b) steadily increases. In contrast, as the smaller equilibrium service rate increases, the corresponding mean steady state waiting time decreases. The relationship between the equilibrium service rates and waiting times is similarly inconsistent in Figure 2. This behavior is not consistent with results from classical, nonstrategic models, and it could serve as a starting point to explaining empiric observations that are also not consistent with classical, nonstrategic models. For example, the nonmonotonicity of service rate in workload is consistent with behavior observed in a hospital setting in Kc and Terwiesch (2009).

4. Staffing Strategic Servers

One of the most studied questions for the design of service systems is staffing. Specifically, how many servers should
be used for a given arrival rate. In the classical, nonstrategic setting, this question is well understood. In particular, as mentioned in the introduction, square-root staffing is known to be optimal when there are linear staffing and waiting costs; see Borst et al. (2004).

In contrast, there is no previous work studying staffing in the context of strategic servers. The goal of this section is to initiate the study of the impact that strategic servers have on staffing. To get a feeling for the issues involved, consider a system with arrival rate $\lambda$ and two possible staffing policies: $N_1 = \lambda$ and $N_2 = 2\lambda$, where $N_i$ is the number of servers staffed under policy $i$ given arrival rate $\lambda$. Under $N_1$, if the servers work at any rate slightly larger than 1, then they will have almost no idle time, and so they will have no incentive to work harder. However, if servers are added, so that the provisioning is as in $N_2$, then servers will have plentiful idle time when working at rate 1; thus, they have no incentive to work harder. Thus, the staffing level has a fundamental impact on the incentives of the servers.

The above highlights that one should expect significant differences in staffing when strategic servers are considered. In particular, the key issue is that the staffing level itself creates incentives for the servers to speed up or slow down, because it influences the balance between effort and idle time. Thus, the policies that are optimal in the nonstrategic setting are likely suboptimal in the strategic setting, and vice versa.

The goal of the analysis in this section is to find the staffing level that minimizes costs when the system manager incurs linear staffing and waiting costs, within the class of policies that admit a symmetric equilibrium service rate. However, the analysis in the previous section highlights that determining the exact optimal policy is difficult, since we only have an implicit characterization of a symmetric equilibrium service rate in (9). As a result, we focus our attention on the setting where $\lambda$ is large, and we look for an asymptotically optimal policy.

As expected, the asymptotically optimal staffing policy we design for the case of strategic servers differs considerably from the optimal policies in the nonstrategic setting. In particular, for a symmetric equilibrium service rate to exist, the staffing level must be order $\lambda$ larger than the optimal staffing in the classical, nonstrategic setting. Then, the system operates in a quality-driven (QD) regime instead of the quality-and-efficiency-driven (QED) regime that results from square-root staffing. This is intuitive given that the servers value their idle time, and in the QD regime they have idle time but in the QED regime their idle time is negligible.

The remainder of this section is organized as follows. We first introduce the cost structure and define asymptotic optimality in Section 4.1. Then, in Section 4.2, we provide a simple approximation of a symmetric equilibrium service rate and an asymptotically optimal staffing policy. Finally, in Section 4.3, we compare our asymptotically optimal staffing policy for strategic servers with the square-root staffing policy that is asymptotically optimal in the nonstrategic setting. All proofs are deferred to the technical appendix.

4.1. Preliminaries

Our focus in this section is on a $M/M/N$ queue with strategic servers, as introduced in Section 3. We assume Random routing throughout this section. It follows that our results hold for any “idle-time-order-based” routing policy (as explained in the beginning of Section 3.2 and validated by Theorem 9). The cost structure we assume is consistent with the one in Borst et al. (2004), under which square-root staffing is asymptotically optimal when servers are not strategic. In their cost structure, there are linear staffing and waiting costs. One difference in our setting is that there may be multiple equilibrium service rates. In light of Theorem 6, we focus on the largest symmetric equilibrium service rate, and we assume $\overline{W}$ denotes the mean steady state waiting time in a $M/M/N$ queue with arrival rate $\lambda$, and strategic servers that serve at the largest symmetric equilibrium service rate (when there is more than one equilibrium). Then, the total system cost is

$$C^*(N, \lambda) = cS N + \overline{W} \lambda \overline{W},$$

where $cS$ is the per-unit staffing cost and $\overline{W}$ is the per-unit waiting cost. The * superscript indicates that the mean steady state waiting time, and hence, the cost function, depends on the (largest) symmetric equilibrium service rate $\mu^*$, which in turn depends on $N$ and $\lambda$.

The function $C^*(N, \lambda)$ is well defined only if a symmetric equilibrium service rate, under which the system is stable, exists. Furthermore, we would like to rule out having an unboundedly large symmetric equilibrium service rate because then the server utility (1) will be large and negative—and it is hard to imagine servers wanting to participate in such a game.

Definition 1. A staffing policy $N^\lambda$ is admissible if the following two properties hold:

(i) There exists a symmetric equilibrium $\mu^\lambda$ under which the system is stable ($\lambda < \mu^\lambda N^\lambda$) for all large enough $\lambda$.

(ii) There exists a sequence of symmetric equilibria $\{\mu^\lambda, \lambda > 0\}$ for which $\limsup_{\lambda \to \infty} \mu^\lambda < \infty$.

If the requirement (ii) in the above definition is not satisfied, then the server utility will approach $-\infty$ as the service rates become unboundedly large. The servers will not want to participate in such a game. As long as the requirement (ii) is satisfied, we can assume the server payment is sufficient to ensure that the servers have positive utility.

We let $\Pi$ denote the set of admissible staffing policies. We would like to solve for

$$N_{opt}^\lambda = \arg \min_{N \in \Pi} C^*(N, \lambda).$$

However, given the difficulty of deriving $N_{opt}^\lambda$ directly, we instead characterize the first order growth term of $N_{opt}^\lambda$ in
terms of $\lambda$. To do this, we consider a sequence of systems, indexed by the arrival rate $\lambda$, and let $\lambda$ become large.

Our convention when we wish to refer to any process or quantity associated with the system having arrival rate $\lambda$ is to superscript the appropriate symbol by $\lambda$. In particular, $N^\lambda$ denotes the staffing level in the system having arrival rate $\lambda$, and $\mu^{\ast, \lambda}$ denotes an equilibrium service rate (assuming existence) in the system with arrival rate $\lambda$ and staffing level $N^\lambda$. We assume $W^{\ast, \lambda}$ equals the mean steady state waiting time in a $M/M/N^\lambda$ queue with arrival rate $\lambda$ when the servers work at the largest equilibrium service rate. The associated cost is

$$C^{\ast, \lambda}(N^\lambda) = c_s N^\lambda + \pi \lambda W^{\ast, \lambda}. \quad (12)$$

Given this setup, we would like to find an admissible staffing policy $N^\lambda$ that has close to the minimum cost $C^{\ast, \lambda}(N^{\text{opt}, \lambda})$.

**Definition 2.** A staffing policy $N^\lambda$ is asymptotically optimal if it is admissible ($N^\lambda \in \Pi$) and

$$\lim_{\lambda \to \infty} \frac{C^{\ast, \lambda}(N^\lambda)}{C^{\ast, \lambda}(N^{\text{opt}, \lambda})} = 1.$$  

In what follows, we use the $o$ and $\omega$ notations to denote the limiting behavior of functions. Formally, for any two real-valued functions $f(x), g(x)$ that take nonzero values for sufficiently large $x$, we say that $f(x) = o(g(x))$ (equivalently, $g(x) = \omega(f(x))$) if $\lim_{x \to \infty} (f(x)/g(x)) = 0$. In other words, $f$ is dominated by $g$ asymptotically (equivalently, $g$ dominates $f$ asymptotically).

### 4.2. An Asymptotically Optimal Staffing Policy

The class of policies we study are those that staff independently of the equilibrium service rates, which are endogenously determined according to the analysis in Section 3.2. More specifically, these are policies that choose $N^\lambda$ purely as a function of $\lambda$. Initially, it is unclear what functional form an asymptotically optimal staffing policy can take in the strategic server setting. Thus, to begin, it is important to rule out policies that cannot be asymptotically optimal. The following proposition does this, and it highlights that asymptotically optimal policies must be asymptotically linear in $\lambda$.

**Proposition 1.** Suppose $N^\lambda = f(\lambda) + o(f(\lambda))$ for some function $f$. If either $f(\lambda) = o(\lambda)$ or $f(\lambda) = \omega(\lambda)$, then the staffing policy $N^\lambda$ cannot be asymptotically optimal.

Intuitively, if $f(\lambda) = o(\lambda)$, understaffing forces the servers to work too hard, their service rates growing unboundedly (and hence their utilities approaching $-\infty$) as $\lambda$ becomes large. On the other hand, the servers may prefer to have $f(\lambda) = \omega(\lambda)$ because the overstaffing allows them to be lazier; however, the overstaffing is too expensive for the system manager.

Proposition 1 implies that to find a staffing policy that is asymptotically optimal, we need only search within the class of policies that have the following form:

$$N^\lambda = \frac{1}{a} \lambda + o(\lambda), \quad \text{for } a \in (0, \infty). \quad (13)$$

However, before we can search for the cost-minimizing $a$, we must ensure that the staffing $(13)$ guarantees the existence of a symmetric equilibria $\mu^{\ast, \lambda}$ for all large enough $\lambda$. It turns out that this is only true when $a$ satisfies certain conditions. After providing these conditions (see Theorem 7), we evaluate the cost function as $\lambda$ becomes large to find the $a^\ast$ (defined in (17)) under which (13) is an asymptotically optimal staffing policy (see Theorem 8).

#### 4.2.1. Equilibrium Characterization

The challenge in characterizing equilibria comes from the complexity of the first order condition derived in Section 3. This complexity drives our focus on the large $\lambda$ regime.

The first order condition for a symmetric equilibrium (9) is equivalently written as

$$\frac{\lambda}{N^\lambda} \left( \mu \left( 1 + \frac{\text{ErlC}(N^\lambda, \lambda/\mu)}{N^\lambda} \right) - \frac{\lambda}{N^\lambda} \right) = \mu^3 c'(\mu). \quad (14)$$

Under the staffing policy (13), when the limit $\lambda \to \infty$ is taken, this becomes

$$a(\mu - a) = \mu^3 c'(\mu). \quad (15)$$

Since $\mu^3 c'(\mu) > 0$, it follows that any solution $\mu$ has $\mu > a$. Therefore, under the optimistic assumption that a symmetric equilibrium solution $\mu^{\ast, \lambda}$ converging to the aforementioned solution $\mu$ exists, it follows that

$$\lambda/\mu^{\ast, \lambda} < \lambda/a$$

for all large enough $\lambda$. In words, the presence of strategic servers that value their idle time forces the system manager to staff order $\lambda$ more servers than the offered load $\lambda/\mu^{\ast, \lambda}$. In particular, since the growth rate of $N^\lambda$ is $\lambda/a$, the system will operate in the quality-driven regime.

The properties of the Equation (15) are easier to see when it is rewritten as

$$\frac{1}{a} = \frac{\mu^2}{\mu^3 c'(\mu)} + \frac{1}{\mu}. \quad (16)$$

Note that the left-hand side of (16) is a constant function and the right-hand side is a convex function. These functions either cross at exactly two points, at exactly one point, or never intersect, depending on $a$. That information then can be used to show whether or not there exists a solution to the first order condition (14), depending on the value of $a$ in the staffing policy (13).

**Theorem 7.** The following holds for all large enough $\lambda$.

(i) Suppose $a > 0$ is such that there exists $\mu_1 > \mu_2 > 0$ that solve (16). Then, there exist two solutions that solve (14).
Suppose \( a > 0 \) is such that there exists exactly one \( \mu_1 > 0 \) that solves (16).

(a) Suppose \( N^\lambda - \lambda/a \geq 0 \). Then, there exist two solutions that solve (14).

(b) Otherwise, if \( N^\lambda - \lambda/a < -3 \), then there does not exist a solution \( \mu^* \) to (14).

Furthermore, from continuity, for any \( \epsilon > 0 \), if \( \mu^\lambda \) solves (14), then \( |\mu^\lambda - \mu| < \epsilon \) for some \( \mu \) that solves (16).

We are not sure if there exists a solution in the case of \( N^\lambda - (1/a)\lambda \in [-3, 0) \); however, given that we are focusing on a large \( \lambda \) asymptotic regime, the range \( [-3, 0) \) is vanishingly small.

Moving forward, once the existence of a solution to the first order condition (14) is established, to conclude that solution is a symmetric equilibrium service rate also requires verifying the condition (10) in Theorem 4. This can be done for any staffing policy (13) under which the system operates in the quality driven-regime.

**Lemma 1.** For any staffing policy \( N^\lambda \) and associated \( \mu^\lambda \) that satisfies the first order condition (14), if

\[
\liminf_{\lambda \to \infty} \frac{N^\lambda \mu^\lambda}{\lambda} = d > 1 \quad \text{and} \quad \limsup_{\lambda \to \infty} \mu^\lambda < \infty,
\]

then \( \mu^\lambda = \mu^* \) is a symmetric equilibrium for all large enough \( \lambda \).

Under the conditions for the existence of a solution to the first order condition (14) in Theorem 7, it is also true that the conditions of Lemma 1 are satisfied. In particular, there exists a bounded sequence \( \{\mu^\lambda\} \) having

\[
\lim_{\lambda \to \infty} \frac{N^\lambda \mu^\lambda}{\lambda} = \lim_{\lambda \to \infty} \left[ \frac{\mu^\lambda}{a} + \frac{\lambda \sigma(a)}{\lambda} \right] > 1.
\]

This then guarantees that, for all large enough \( \lambda \), there exists a solution \( \mu^{\lambda, \lambda} \) to (14) that is a symmetric equilibrium, under the conditions of Theorem 7.

There are either multiple symmetric equilibria for each \( \lambda \) or 0, because from Theorem 7 there are either multiple or zero solutions to the first order condition (14). These symmetric equilibria will be close when there exists exactly one \( \mu \) that solves (16); however, they may not be close when there exist two \( \mu \) that solve (16). We show in the following that this does not affect what staffing policy should be asymptotically optimal.

**4.2.2. Optimal Staffing.** Given the characterization of symmetric equilibria under a staffing policy (13), we can now move to the task of optimizing the staffing level, i.e., optimizing \( a \). The first step is to characterize the associated cost, which is done in the following proposition.

**Proposition 2.** Suppose \( a > 0 \) is such that there exists \( \mu > 0 \) that solves (16). Then, under the staffing policy (13),

\[
\frac{C^{\lambda, a}(N^\lambda)}{\lambda} \to \frac{1}{a} c_E, \quad \text{as} \quad \lambda \to \infty.
\]

Proposition 2 implies that to minimize costs within the class of staffing policies that satisfy (13), the maximum \( a \) under which there exists at least one solution to (16) should be used. That is, we should choose \( a \) to be

\[
a^* := \sup \mathcal{A}, \quad \text{where} \quad \mathcal{A} := \{a > 0 : \text{there exists at least one solution } \mu > 0 \text{ to (16)}\}.
\]

**Lemma 2.** \( a^* \in \mathcal{A} \) is finite.

Importantly, this \( a^* \) is not only optimal among the class of staffing policies that satisfy (13), it is asymptotically optimal among all admissible staffing policies. In particular, the following theorem shows that as \( \lambda \) becomes unboundedly large, no other admissible staffing policy can asymptotically achieve strictly lower cost than the one in (13) with \( a = a^* \).

**Theorem 8.** If \( N^{a, \lambda} \) satisfies (13) with \( a = a^* \), then \( N^{a, \lambda} \) is admissible and asymptotically optimal. Furthermore,

\[
\lim_{\lambda \to \infty} \frac{C^{\lambda, a^*}(N^{a^*, \lambda})}{\lambda} = \lim_{\lambda \to \infty} \frac{C^{\lambda, a^*}(N^{\text{opt}, \lambda})}{\lambda} = \frac{1}{a^*}.
\]

Note that an inspection of the proof of Theorem 8 shows that it holds regardless of which equilibrium service rate is used to define \( W^{a, \lambda} \). Hence, even though we have defined \( W^{a, \lambda} \) to be the mean steady state waiting time when the servers serve at the largest equilibrium service rate, this is not necessary. The staffing policy \( N^{a, \lambda} \) in Theorem 8 will be asymptotically optimal regardless of which equilibrium service rate the servers choose.

Though the above theorem characterizes an asymptotically optimal staffing level, because the definition of \( a^* \) is implicit, it is difficult to develop intuition. To highlight the structure more clearly, the following lemma characterizes \( a^* \) for a specific class of effort cost functions.

**Lemma 3.** Suppose \( c(\mu) = c_E \mu^p \) for some \( c_E \in [1, \infty) \) and \( p \geq 1 \). Then,

\[
a^* = \left\lfloor \frac{p + 1}{c_E (p + 2)} \right\rfloor^{1/(p+1)\cdot (p+1)/p} \leq \mu^*, \quad \text{and} \quad a^* \text{ and } \mu^* \text{ are both increasing in } p.
\]

Furthermore,

\[
\text{if } a > a^*, \quad \left\{ \begin{array}{l}
\text{there are two nonnegative solutions to (16); if } a < a^*, \quad \left\{ \begin{array}{l}
\text{there is no nonnegative solution to (16); if } a = a^*, \quad \left\{ \begin{array}{l}
\text{there is exactly one solution to (16).}
\end{array} \right.
\end{array} \right.
\end{array}
\end{array}
\]

There are several interesting relationships between the effort cost function and the staffing level that follow from Lemma 3. First, for fixed \( p \),

\[
a^*(p) \downarrow 0 \quad \text{as } c_E \to \infty.
\]
In words, the system manager must staff more and more servers as effort becomes more costly. Second, for fixed $c_E$, since $a^*(p)$ is increasing in $p$, the system manager can staff fewer servers when the cost function becomes “more convex.” The lower staffing level forces the servers to work at a higher service rate since $\mu^*(p)$ is also increasing in $p$. We will revisit this idea that convexity is helpful to the system manager in Section 4.3.2.

4.3. Contrasting Staffing Policies for Strategic and Nonstrategic Servers

One of the most crucial observations that Section 4.2 makes about the impact of strategic servers on staffing is that the strategic behavior leads the system to a quality-driven regime. In this section we explore this issue in more detail, by comparing to the optimal staffing rule that arises when servers are not strategic and then attempting to implement that staffing rule.

4.3.1. Nonstrategic Servers. Recall that, for the conventional $M/M/1$ queue (without strategic servers), square-root staffing minimizes costs as $\lambda$ becomes large (see Borst et al. 2004, Equation (1), Proposition 6.2, Example 6.3).

So, we can define

$$C_\mu^N(N) = c_s N + w\lambda W_\mu^\lambda$$

to be the cost associated with staffing $N$ nonstrategic servers that work at the fixed service rate $\mu$. Further,

$$N_{\mu}^{\text{opt}, \lambda} = \arg\min_{N>0} C_\mu^N(N)$$

is the staffing level that minimizes expected cost when the system arrival rate is $\lambda$ and the service rate is fixed to be $\mu$. So, the staffing rule

$$N_{\mu}^{\text{BMR}, \lambda} = \frac{\lambda}{\mu} + y^* \sqrt{\frac{\lambda}{\mu}}$$

(18)

is asymptotically optimal in the sense that

$$\lim_{\lambda \to \infty} \frac{C_\mu^{N_{\mu}^{\text{BMR}, \lambda}}(N_{\mu}^{\text{BMR}, \lambda})}{C_\mu^{N_{\mu}^{\text{opt}, \lambda}}(N_{\mu}^{\text{opt}, \lambda})} = 1.$$ 

Here, $y^* := \arg\min_{x>0} \{c_s y + w\alpha(y)/y\}$, where $\alpha(y) = (1 + y/h(-y))^{-1}$ with $h(\cdot)$ being the hazard rate function of the standard normal distribution, namely, $h(x) := \phi'(x)/(1-\Phi(x))$ with $\phi(x) = (1/\sqrt{2\pi}) e^{-x^2/2}$ and $\Phi(x) = \int_{-\infty}^x \phi(t)\ dt$. The staffing rule (18) is the famous square-root safety staffing rule.

4.3.2. Comparing Strategic and Nonstrategic Servers. To compare the case of strategic servers to the case of nonstrategic servers, it is natural to fix $\mu$ in (18) to the limiting service rate that results from using the optimal staffing rule $N_{\mu^*}^{\text{opt}, \lambda}$ defined in Theorem 8. We see that $N_{\mu^*}^{\text{opt}, \lambda}$ staffs order $\lambda$ more servers than $N_{\mu^*}^{\text{BMR}, \lambda}$, where $\mu^*$ solves (16) for $a = a^*$, because any solution to (16) has $\alpha > \mu$.

When the effort cost function is $c(\mu) = c_E \mu^p$ for $p \geq 1$, we know from Lemma 3 and Theorem 7 (since the $a^*$ is unique) that

$$\mu^{*, \lambda} \to \mu^*$$

as $\lambda \to \infty$,

where $\mu^*$ is as given in Lemma 3. Then, the difference in the staffing levels is

$$N_{\mu}^{\text{ao}, \lambda} - N_{\mu^*}^{\text{BMR}, \lambda} = \left( \frac{1}{a^*} - \frac{1}{\mu^*} \right) \lambda + o(\lambda)$$

$$= \frac{1}{a^* (p+2)} \lambda + o(\lambda).$$

Since $a^* = a^*(p)$ is increasing in $p$ from Lemma 3, we see that the difference $N_{\mu}^{\text{ao}, \lambda} - N_{\mu^*}^{\text{BMR}, \lambda}$ decreases to 0 as the cost function becomes “more convex.” This is consistent with our observation at the end of the previous subsection that convexity is helpful to the system manager.

It is natural to wonder if a system manager can force the servers to work harder by adopting the staffing policy suggested by the analysis of nonstrategic servers, i.e.,

$$N_{\mu^*}^{\text{BMR}, \lambda} = \frac{\lambda}{\mu^*} + y^* \sqrt{\frac{\lambda}{\mu^*}}.$$ 

(19)

The interpretation of this staffing rule requires care, because the offered load $\lambda/\mu^*$ is itself a function of the staffing level (and the arrival rate) through an equilibrium service rate $\mu^{*, \lambda}$. The superscript * emphasizes this dependence.

The first question concerns whether or not the staffing policy (19) is even possible in practice, because the staffing level depends on an equilibrium service rate and vice versa. More specifically, for a given staffing level, the servers relatively quickly arrive at an equilibrium service rate. Then, when system demand grows, the system manager increases the staffing, and the servers again arrive at an equilibrium service rate. In other words, there are two games, one played on a faster time scale (that is the servers settling to an equilibrium service rate), and one played on a slower time scale (that is the servers responding to added capacity).

To analyze the staffing policy (19), note that the first order condition for a symmetric equilibrium (9) is equivalently written as

$$\frac{\lambda/\mu}{(N_{\mu^*}^{\text{BMR}, \lambda})^2} \left( N_{\mu^*}^{\text{BMR}, \lambda} - \frac{\lambda}{\mu} + \text{ErlC}(N_{\mu^*}^{\text{BMR}, \lambda}, \frac{\lambda}{\mu}) \right)$$

$$= \mu \cdot \alpha^*(\mu).$$

Then, if $\mu^*$ is a solution to the first order condition under the staffing $N_{\mu^*}^{\text{BMR}, \lambda}$ from (19), from substituting $N_{\mu^*}^{\text{BMR}, \lambda}$ into the above expression, $\mu^*$ must satisfy

$$\frac{\lambda/\mu^*}{(\lambda/\mu^* + y^* \sqrt{\lambda/\mu^*})^2}$$

$$= \mu \cdot \alpha^*(\lambda).$$
As \( \lambda \) becomes large, since \( \text{ErlC}(\lambda/\mu + y\sqrt{\lambda/\mu}, \lambda/\mu) \) is bounded above by 1, the left-hand side of the above expression has limit 0. Furthermore, the right-hand side of the above equation is nonnegative and increasing as a function of \( \mu \). Hence, any sequence of solutions \( \mu^\lambda \) to the first order condition has the limiting behavior

\[
\mu^\lambda \to 0, \quad \text{as} \quad \lambda \to \infty,
\]

which cannot be a symmetric equilibrium service rate because we require the servers to work fast enough to stabilize the system.

One possibility is to expand the definition of an equilibrium service rate in (1) to allow the servers to work exactly at the lower bound \( \lambda/N \). In fact, the system manager may now be tempted to push the servers to work even faster. However, faster service cannot be mandated for free—there must be a trade-off; for example, the service quality may suffer or the salaries should be higher.

### 4.4. Numerical Examples

To understand how well our asymptotically optimal staffing policy \( N^{\text{opt}-\lambda} \) performs in comparison with the optimal policy \( N^{\text{opt}-\lambda} \) for finite \( \lambda \), and how fast the corresponding system cost converges to the optimal cost, we present some results from numerical analysis in this section.

We consider two staffing policies: (i) \( N^{\text{opt}-\lambda} \) (defined in (11)), and (ii) \( N^{\text{opt}-\lambda} \) (defined in Theorem 8 and (17)) where we ignore the \( o(\lambda) \) term of (13). For each, we first round up the staffing level if necessary, and then we plot the following two equilibrium quantities as a function of the arrival rate \( \lambda \): (a) service rates \( \mu^{*,-\lambda} \) (if there is more than one, we pick the largest), and (b) normalized costs \( C^{*,-\lambda}/\lambda \). We calculate \( N^{\text{opt}-\lambda} \) numerically, by iterating over the staffing levels that admit equilibria (and we choose the lowest cost when there are multiple equilibria). These plots are shown in Figure 3 for three different effort cost functions: \( c(\mu) = \mu \), \( c(\mu) = \mu^2 \), and \( c(\mu) = \mu^3 \), which are depicted in red, blue, and green, respectively. For each color, the curve with the darker shade corresponds to \( N^{\text{opt}-\lambda} \) and the curve with the lighter shade corresponds to \( N^{\text{opt}-\lambda} \).

The horizontal dashed lines correspond to the limiting values as \( \lambda \to \infty \).

An immediate first observation is the jaggedness of the curves, which is a direct result of the discreteness of the staffing levels \( N^{\text{opt}-\lambda} \) and \( N^{\text{opt}-\lambda} \). In particular, as the arrival rate \( \lambda \) increases, the equilibrium service rate \( \mu^{*,-\lambda} \) decreases (respectively, the equilibrium normalized cost \( C^{*,-\lambda}/\lambda \) increases) smoothly until the staffing policy adds an additional server, which causes a sharp increase (respectively, decrease). The jaggedness is especially pronounced for smaller \( \lambda \), resulting in a complex prelimit behavior that necessitates asymptotic analysis to obtain analytic results.

However, despite the jaggedness, the plots illustrate clearly that both the equilibrium service rates and normalized costs of the optimal policy \( N^{\text{opt}-\lambda} \) converge quickly to those of the optimal policy \( N^{\text{opt}-\lambda} \), highlighting that our asymptotic results are predictive at realistically sized systems.

### 5. Routing to Strategic Servers

Thus far we have focused our discussion on staffing, assuming that jobs are routed randomly to servers when there is a choice. Of course, the decision of how to route jobs to servers is another crucial aspect of the design of service systems. As such, the analysis of routing policies has received considerable attention in the queueing literature, when servers are not strategic. In this section we begin to investigate the impact of strategic servers on the design of routing policies.

In the classical literature studying routing when servers are nonstrategic, a wide variety of policies have been considered. These include “rate-based policies” such as Fastest...
5.1. Idle-Time-Order-Based Policies

Informally, idle-time-order-based policies are those routing policies that use only the rank ordering of when servers last became idle to determine how to route incoming jobs. To describe the class of idle-time-order-based policies precisely, let \( \mathcal{J}(t) \) be the set of servers idle at time \( t > 0 \), and, when \( \mathcal{J}(t) \neq \emptyset \), let \( \mathbf{s}(t) = (s_1, \ldots, s_{|\mathcal{J}(t)|}) \) denote the ordered vector of idle servers at time \( t \), where server \( s_j \) became idle before server \( s_k \) whenever \( j < k \). For \( n \geq 1 \), let \( \mathcal{P}_n = \mathcal{A} \{(1, \ldots, n)\} \) denote the set of all probability distributions over the set \( \{1, \ldots, n\} \). An idle-time-order-based routing policy is defined by a collection of probability distributions \( \mathbb{P} = \{p^\mathcal{J}\}_{\mathcal{J} \subseteq \{1, \ldots, N\}} \), such that \( p^\mathcal{J} \in \mathcal{P}_|\mathcal{J}| \), for all \( \mathcal{J} \subseteq \{1, \ldots, N\} \backslash \emptyset \). Under this policy, at time \( t \), the next job in queue is assigned to idle server \( s_j \) with probability \( p^\mathcal{J}(j) \). Examples of idle-time-order-based routing policies are as follows.

1. Random. An arriving customer that finds more than one idle server is equally likely to be routed to any of those servers. Then, \( p^\rho = (1/|\mathcal{S}|, \ldots, 1/|\mathcal{S}|) \) for all \( \mathcal{S} \in 2^{\{1, \ldots, N\}} \backslash \emptyset \).

2. Weighted Random. Each such arriving customer is routed to one of the idle servers with probabilities that may depend on the order in which the servers became idle. For example, if

\[
p^\rho(j) = \frac{|\mathcal{S}| + 1 - j}{\sum_{i=1}^{|\mathcal{S}|} i},
\]

then the probabilities are decreasing according to the order in which the servers became idle. Note that \( \sum_j p^\rho(j) = |\mathcal{S}|(|\mathcal{S}| + 1) - \frac{1}{2} |\mathcal{S}|(|\mathcal{S}| + 1) \), which implies that \( p^\rho \) is a probability distribution.

3. Longest Idle Server First (Shortest Idle Server First). Each such arriving customer is routed to the server that has idled the longest (idled the shortest). Then, \( p^\rho = (1, 0, \ldots, 0) \) for \( \mathcal{S} = \{1, 2, \ldots, N\} \).

5.1.1. Policy-Space Collapse. Surprisingly, it turns out that all idle-time-order-based policies are “equivalent” in a very strong sense—they all lead to the same steady state probabilities, resulting in a remarkable policy-space collapse result, which we discuss in the following.

Fix \( R \) to be some idle-time-order-based routing policy, defined through the collection of probability distributions \( \mathbb{P} = \{p^\mathcal{J}\}_{\mathcal{J} \subseteq \{1, \ldots, N\}} \). The states of the associated continuous-time Markov chain are defined as follows:

- State \( B \) is the state where all servers are busy, but there are no jobs waiting in the queue.
- State \( \mathbf{s} = (s_1, s_2, \ldots, s_{|\mathcal{J}|}) \) is the ordered vector of idle servers \( \mathcal{J} \). When \( \mathcal{J} = \emptyset \), we identify the empty vector \( \mathbf{s} \) with state \( B \).
- State \( m \) \((m \geq 0)\) is the state where all servers are busy and there are \( m \) jobs waiting in the queue (i.e., there are \( N + m \) jobs in the system). We identify state \( 0 \) with state \( B \).

When all servers are busy, there is no routing, and so the system behaves exactly as a \( M/M/1 \) queue with arrival rate \( \lambda \) and service rate \( \mu_1 + \cdots \mu_N \). From the local balance equations, the associated steady state probabilities \( \pi_B \) and \( \pi_m \) for \( m = 0, 1, 2, \ldots \), must satisfy

\[
\pi_m = (\lambda/\mu)^m \pi_B m \quad \text{where } \mu = \sum_{j=1}^N \mu_j. \tag{20}
\]

One can anticipate that the remaining steady state probabilities satisfy

\[
\pi_s = \frac{\pi_B}{\lambda} \prod_{j=1}^n \frac{\mu_j}{\lambda},
\]

for all \( \mathbf{s} = (s_1, s_2, \ldots, s_{|\mathcal{J}|}) \), with \( |\mathcal{J}| > 0 \), \( (21) \)

and the following theorem verifies this by establishing that the detailed balance equations are satisfied.

THEOREM 9. All idle-time-order-based policies have the steady state probabilities that are uniquely determined by (20)–(21), together with the normalization constraint that their sum is one.

Theorem 9 is remarkable because there is no dependence on the collection of probability distributions \( \mathbb{P} \) that define \( R \). Therefore, it follows that all idle-time-order-based routing policies result in the same steady state probabilities. Note that, concurrently, a similar result has been discovered independently in the context of loss systems (see Haji and Ross 2015).

In relation to our server game, it follows from Theorem 9 that all idle-time-order-based policies have the same equilibrium behavior as Random. This is because an equilibrium service rate depends on the routing policy through the server idle time vector \( (I_1(\mathbf{\mu}; R), \ldots, I_N(\mathbf{\mu}; R)) \), which can be found from the steady state probabilities in (20)–(21). As a consequence, if there exists (does not exist) an equilibrium service rate under Random, then there exists (does not exist) an equilibrium service rate under any idle-time-order-based policy. In summary, it is not possible to achieve better performance than under Random by employing any idle-time-order-based policy.
5.2. Rate-Based Policies

Informally, a rate-based policy is one that makes routing decisions using only information about the rates of the servers. As before, let $J(t)$ denote the set of idle servers at time $t$. In a rate-based routing policy, jobs are assigned to idle servers only based on their service rates. We consider a parameterized class of rate-based routing policies that we term $r$-routing policies ($r \in \mathbb{R}$). Under an $r$-routing policy, at time $t$, the next job in queue is assigned to idle server $i \in J(t)$ with probability

$$p_i(\mu, r; r) = \frac{\mu_i^r}{\sum_{j \in J(t)} \mu_j^r}.$$ 

Notice that for special values of the parameter $r$, we recover well-known policies. For example, setting $r = 0$ results in Random; as $r \to \infty$, it approaches FSF; and as $r \to -\infty$, it approaches SSF.

To understand the performance of rate-based policies, the first step is to perform an equilibrium analysis, i.e., we need to understand what the steady state idle times look like under any $r$-routing policy. The following proposition provides us with the required expressions.

**Proposition 3.** Consider a heterogeneous M/M/2 system under an $r$-routing policy, with arrival rate $\lambda > 0$ and servers 1 and 2 operating at rates $\mu_1$ and $\mu_2$, respectively. The steady state probability that server 1 is idle is given by

$$I_1^r(\mu_1, \mu_2) = \left(\frac{\mu_1}{\mu_1 + \frac{\mu_2^r}{\mu_1^r + \mu_2^r}} + \frac{\mu_2^r}{\mu_1^r + \mu_2^r}\right)(\mu_1 + \mu_2^r) + \frac{\mu_2^r}{\mu_1^r + \mu_2^r}(\lambda \mu_1 + \lambda \mu_2^r),$$

and the steady state probability that server 2 is idle is given by $I_2^r(\mu_1, \mu_2) = I_1^r(\mu_2, \mu_1)$.

Note that we restrict ourselves to a two-server system for this analysis. This is due to the fact that there are no closed form expressions known for the resulting Markov chains for systems with more than three servers. It may be possible to extend these results to three servers using results from Mokaddis et al. (1998); but the expressions are intimidating, to say the least. However, the analysis for two servers is already enough to highlight important structure about the impact of strategic servers on policy design.

In particular, our first result concerns the FSF and SSF routing policies, which can be obtained in the limit when $r \to \infty$ and $r \to -\infty$, respectively. Recall that FSF is asymptotically optimal in the nonstrategic setting. Intuitively, however, it penalizes the servers that work the fastest by sending them more and more jobs. In a strategic setting, this might incentivize servers to decrease their service rate, which is not good for the performance of the system. One may wonder if by doing the opposite, that is, using the SSF policy, servers can be incentivized to increase their service rate. However, our next theorem (Theorem 10) shows that neither of these policies is useful if we are interested in symmetric equilibria.

Recall that our model for strategic servers already assumes an increasing, convex effort cost function with $c''(\mu) \geq 0$. For the rest of this section, in addition, we assume that $c'(\lambda/2) < 1/\lambda$. (Recall that this is identical to the sufficient condition $c'(\lambda/N) < 1/\lambda$ which we introduced in Section 3.2, on substituting $N = 2$.)

**Theorem 10.** Consider a M/M/2 queue with strategic servers. Then, FSF and SSF do not admit a symmetric equilibrium.

Moving beyond FSF and SSF, we continue our equilibrium analysis (for a finite $r$) by using the first order conditions to show that whenever an $r$-routing policy admits a symmetric equilibrium, it is unique. Furthermore, we provide an expression for the corresponding symmetric equilibrium service rate in terms of $r$, which brings out a useful monotonicity property.

**Theorem 11.** Consider a M/M/2 queue with strategic servers. Then, any $r$-routing policy that admits a symmetric equilibrium admits a unique symmetric equilibrium, given by $\mu^* = \varphi^{-1}(r)$, where $\varphi: (\lambda/2, \infty) \to \mathbb{R}$ is the function defined by

$$\varphi(\mu) = \frac{4(\lambda + \mu)}{\lambda(\lambda - 2\mu)}(\mu(\lambda + 2\mu)c'(\mu) - \lambda).$$

Furthermore, among all such policies, $\mu^*$ is decreasing in $r$; therefore, $\mathbb{E}[T]$, the mean response time (a.k.a. sojourn time) at symmetric equilibrium is increasing in $r$.

In light of the inverse relationship between $r$ and $\mu^*$ that is established by this theorem, the system manager would ideally choose the smallest $r$ such that the corresponding $r$-routing policy admits a symmetric equilibrium, which is in line with the intuition that a bias towards SSF (the limiting $r$-routing policy as $r \to -\infty$) incentivizes servers to work harder. However, there is a hard limit on how small an $r$ can be chosen (concurrently, how large an equilibrium service rate $\mu^*$ can be achieved) so that there exists a symmetric equilibrium, as evidenced by our next theorem.

**Theorem 12.** Consider a M/M/2 queue with strategic servers. Then there exists $\overline{\mu}, r \in \mathbb{R}$, with $r = \varphi(\overline{\mu})$, such that no service rate $\mu > \overline{\mu}$ can be a symmetric equilibrium under any $r$-routing policy, and no $r$-routing policy with $r < \overline{\mu}$ admits a symmetric equilibrium.
The proof of this theorem is constructive, and we do exhibit an \( r \); however, it is not clear whether this is tight, that is, whether there exists a symmetric equilibrium for all \( r \)-routing policies with \( r \geq r \). We provide a partial answer to this question of what \( r \)-routing policies do admit symmetric equilibria in the following theorem.

**Theorem 13.** Consider a \( M/M/2 \) queue with strategic servers. Then there exists a unique symmetric equilibrium under any \( r \)-routing policy with \( r \in \{-2, -1, 0, 1\} \).

Notice that we show equilibrium existence for four integral values of \( r \). It is challenging to show that all \( r \)-routing policies in the interval \([-2, 1]\) admit a symmetric equilibrium. This theorem provides an upper bound on the \( r \) of the previous theorem, that is, \( r \leq -2 \). Therefore, if the specific cost function \( c \) is unknown, then the system manager can guarantee better performance than Random (\( r = 0 \)), by setting \( r = -2 \). If the specific cost function is known, the system manager may be able to employ a lower \( r \) to obtain even better performance. For example, consider a two-server system with \( \lambda = 1/4 \) and one of three different effort cost functions: \( c(\mu) = \mu \), \( c(\mu) = \mu^2 \), and \( c(\mu) = \mu^3 \). Figure 4 shows the corresponding equilibrium mean response times (in red, blue, and green, respectively). It is worth noting that the more convex the effort cost function, larger the range of \( r \) (and smaller the minimum value of \( r \)) for which a symmetric equilibrium exists.

### 6. Concluding Remarks

The rate at which each server works in a service system has important consequences for service system design. However, traditional models of large service systems do not capture the fact that human servers respond to incentives created by scheduling and staffing policies, because traditional models assume each server works at a given fixed service rate. In this paper we initiate the study of a class of strategic servers that seek to optimize a utility function which values idle time and includes an effort cost.

Our focus is on the analysis of staffing and routing policies for a \( M/M/N \) queue with strategic servers, and our results highlight that strategic servers have a dramatic impact on the optimal policies in both cases. In particular, policies that are optimal in the classical, nonstrategic setting can perform quite poorly when servers act strategically.

For example, a consequence of the strategic server behavior is that the cost-minimizing staffing level is order \( \lambda \) larger than square-root staffing, the cost minimizing staffing level for systems with fixed service rate. In particular, any system with strategic servers operates in the quality-driven regime at equilibrium (as opposed to the quality-and-efficiency-driven regime that arises under square-root staffing), in which the servers all enjoy nonnegligible idle time.

The intuitive reason square-root staffing is not feasible in the context of strategic servers is that the servers do not value their idleness enough in comparison to their effort cost. This causes the servers to work too slowly, making idle time scarce. In the economics literature (see Brock 1968, Lynn 1991), it is common to assume that scarce goods are more highly valued. If we assume that the servers valued their idle time more heavily as the idle time becomes scarcer, then the servers would work faster to make sure they achieved some. This suggests the following interesting direction for future research: what is the relationship between the assumed value of idle time in (2) and the resulting cost minimizing staffing policy? Another situation in which servers may not care about idle time becoming scarce is when their compensation depends on their service volume (which is increasing in their service rate). Then it is reasonable to expect the servers prefer to have negligible idle time. It would be interesting to be able to identify a class of compensation schemes under which that is the case.

The aforementioned two future research directions become even more interesting when the class of routing policies is expanded to include rate-based policies. This paper solves the joint routing and staffing problem within the class of idle-time-order-based policies. Section 5 suggests that by expanding the class of routing policies to also include rate-based policies we should be able to achieve better system performance (although it is clear that the analysis becomes much more difficult). The richer question also aspires to understand the relationship between the server idle time value, the compensation scheme, the (potentially) rate-based routing policy, and the number of strategic servers to staff.

Finally, it is important to note that we have focused on symmetric equilibrium service rates. We have not proven that asymmetric equilibria do not exist. Thus, it is natural to wonder if there are routing and staffing policies that result in an asymmetric equilibrium. Potentially, there could be one group of servers that have low effort costs but negligible idle time and another group of servers that enjoy plentiful idle time but have high effort costs. The question of asymmetric equilibria becomes even more interesting when the servers have different utility functions. For example,
more experienced servers likely have lower effort costs than new hires. Also, different servers can value their idle time differently. How do we design routing and staffing policies that are respectful of such considerations?

Supplemental Material
Supplemental material to this paper is available at http://dx.doi.org/10.1287/opre.2016.1506.

Acknowledgments
The authors also thank Bert Zwart for an illuminating discussion regarding asymptotically optimal staffing policies. An important source of inspiration for this work came from discussions with Mor Armony regarding how to fairly assign workload to employees. This work was supported by National Science Foundation [Grant CCF-1101470], Air Force Office of Scientific Research [Grant FA9550-12-1-0359], and Office of Navy Research [Grant N00014-09-1-0751].

Endnotes
1. Specifically, if \( g(I(\mu)) \) replaces \( I(\mu) \) in (1), all the results in Section 3 characterizing equilibrium service rates are maintained so long as \( g'' < 0 \), except for Theorem 5, whose sufficient condition would have to be adjusted to accommodate \( g \). In addition, our results could be made stronger depending on the specific form of \( g \). For example, if \( g \) is such that \( \lim_{\mu \to -\infty} U(\mu) = -\infty \), then, a preliminary analysis reveals that it would not be necessary to impose the stability constraint \( \mu > \lambda/N \) exogenously. Moreover, every solution to the symmetric first order condition (9) would be a symmetric equilibrium (i.e., the sufficient condition of Theorem 4 as generalized for this case by Endnote 2 would automatically be satisfied).

2. One can imagine that servers, despite being strategic, would endogenously stabilize the system. To test this, one could study a related game where the action sets of the servers are endogenously stabilized. To test this, one could study a related game where the action sets of the servers are endogenously stabilized.

3. In general, the symmetric first order condition (9) can be rewritten as:

\[
\mu^2 c'(\mu) + \frac{\lambda}{N^2} (\rho - \text{ErlC}(N, \rho)) - \frac{\lambda}{N} = 0.
\]

Note that when the term \( \rho - \text{ErlC}(N, \rho) \) is convex in \( \mu \), it follows that the left-hand side of the above equation is also convex in \( \mu \), which implies that there are at most two symmetric equilibria.

4. Note that the staffing policy we derive in this section (Theorem 8) will be asymptotically optimal regardless of which equilibrium service rate the servers choose.

5. The sufficient condition \( c'(A/2) < 1/A \) might seem rather strong, but it can be shown that it is necessary for the symmetric first order condition to have a unique solution. This is because, if \( c'(A/2) > 1/A \), then the function \( \varphi(\mu) \), defined in (22), ceases to be monotonic, and as a result, for any given \( r \), the first order condition \( \varphi(\mu) = r \) could have more than one solution.

References


Gurvich I, Whitt W (2007) Scheduling flexible servers to the system is critically loaded.


__Ragavendran Gopalakrishnan__ is a research scientist at Xerox Research Centre India. His interests lie in the domain of algorithmic game theory, especially in its applications to utility design in distributed optimization, cost sharing mechanism design in distributed resource allocation, and systemic incentive design in multiserver service systems. He received his doctorate in computer science from the California Institute of Technology in June 2013, following which he was a research associate at the University of Colorado Boulder until February 2015.

__Sherwin Doroudi__ is a doctoral candidate in operations management in the Tepper School of Business at Carnegie Mellon University. His research focuses on modeling and analyzing resource management problems in stochastic environments. Much of his work involves the development of new analytic methodologies with an emphasis on the exact analysis of large-scale Markov chains arising in queueing-theoretic models. He will join the Department of Industrial and Systems Engineering (ISyE) at the University of Minnesota as an assistant professor in August 2016.

__Amy R. Ward__ is a professor of data sciences and operations in the Marshall School of Business at the University of Southern California. She is the current vice-chair of the Applied Probability Society of INFORMS and will become chair in November 2016. Her research focuses on the approximation and control of systems in which uncertainty and variability are present and cannot be ignored. She received the 2015 Deans Award for Research Excellence from Marshall.

__Adam Wierman__ is a professor in the Department of Computing and Mathematical Sciences at the California Institute of Technology, where he is a founding member of the Rigorous Systems Research Group. His research interests center around resource allocation and scheduling decisions in computer systems and services. He received the 2011 ACM SIGMETRICS Rising Star award, the 2014 IEEE Communications Society William R. Bennett Prize, and has been coauthor on papers that received best paper awards at ACM SIGMETRICS, IEEE INFOCOM, IFIP Performance (twice), IEEE Green Computing Conference, IEEE Power and Energy Society General Meeting, and ACM GREENMETRICS.