On the Design of Service Systems when Servers are Strategic

Ragavendran Gopalakrishnan
Cornell University, rg584@cornell.edu
https://people.orie.cornell.edu/ragad3/

Problem Definition: We revisit the optimal design of a service system when service rates are endogenously chosen by strategic servers, in response to the routing policy and the choice of queue configuration. We study the joint optimization of the routing policy (over a large class of "rate-based" policies) and the system configuration ("pooled" vs. "dedicated").
Academic/Practical Relevance: Many service systems are staffed by human servers who enjoy some discretion in their service times; however, queueing models with endogenous service rates have not been studied until very recently. We advance this line of research significantly by moving beyond just analyzing the simple Random routing policy. Our results highlight the importance of considering strategic server behavior, by considering servers that value idle time, while incurring a cost of effort.
Methodology: Our approach consists of a mixture of both analytical and numerical techniques. The pooled configuration (modeled as a single $M/M/N$ queue) is much less tractable, especially under heterogeneous service rates, than the dedicated configuration (modeled as N parallel $M/M/1$ queues).
Results: Under either configuration, we exhibit policies that outperform random routing in terms of the mean response/waiting times. However, such system-optimal policies also minimize servers' utilities. Moreover, these worst-case utilities are lower for servers that are "less skilled". Under system-optimal policies, servers work faster when there are more servers in the system, which adds a nontrivial angle to the classical question of optimal staffing. Finally, servers work faster in the dedicated system, but not necessarily enough to overcome the inherent inefficiency of the dedicated configuration.
Managerial Implications: In choosing the best service system design, a system manager must carefully consider the servers' skills and staffing levels. While it may be tempting to choose a system-optimal policy, it may lead to lower employee satisfaction, and less skilled servers are more vulnerable to such "exploitation".

Key words: service systems; resource pooling; routing; scheduling; strategic servers
Subject classifications: Primary: Queues: applications; secondary: Games/group decisions: noncooperative

1. Introduction

A growing body of work in the Operations Management literature, both empirical and analytical, strongly suggests that one must fundamentally rethink the design of service systems when servers are humans who make strategic decisions regarding aspects of their service over which they enjoy discretion. Such decisions include real-time or scheduled availability, as well as service efficiency (speed/quality). In today's world of the shared economy, where services are increasingly being offered by crowdworkers, e.g., ride-hailing, food/grocery delivery, etc., it is all the more important to account for strategic server behavior when making managerial decisions regarding service operations.

Our focus in this paper is on the strategic choice of service times. For example, consider prosocial service, such as peer-reviewing for a journal. Referees who accept a request to review a manuscript enjoy flexibility in scheduling their time towards working on their reports and when to submit them. Therefore, even if the actual amount of time spent by a referee on their report (and hence its quality) is fixed, the editor could observe a range of *effective* service times, since the editor only observes the *submission time* of the report. A recent field experiment performed with 1,500 referees at the Journal of Public Economics [11] strongly supports this hypothesis, finding that simply assigning a shorter deadline of 4 weeks instead of 6 improved the median review times from 55 days to about 43 days, with no effect on the quality of reviews or likelihood of acceptance (at the same journal and at other journals). Importantly, the number of pending reviews after 3 weeks (at which time a reminder email was sent) was only slightly smaller for the referees who were assigned the 4-week deadline, suggesting that most referees likely spend less than a week actively working on their reviews.

As another example, agents in a contact center may have some flexibility over their pace of service without significantly affecting quality, depending on the type of service (see, e.g., Figure 3.2 in [32]). Even more examples emerge in the shared economy with on-demand service models, e.g., a part-time in-store Instacart shopper (who is paid a fixed hourly wage [15]) choosing how fast to run their assigned grocery orders, an Uber driver choosing the route and/or driving speed [6]. Moreover, there are scenarios in which servers are machines, where such discretion is appropriate, e.g., in distributed data centers providing cloud computing services, each facility can minimize its energy costs by using dynamic speed scaling and power-down mechanisms at any given time, subject to contractual service level agreements being met [22].

A broadly applicable model for such strategic behavior is one in which (a) servers are characterized by a cost function $c(\mu)$ that specifies the cost (e.g., labor/cognitive effort, equipment operation/maintenance) they incur for operating at a particular service rate μ (while ensuring that the service quality meets an acceptable standard), and (b) servers strategically choose their service rates in order to balance this cost against the tangible benefits they get in return. While such benefits are often thought to be monetary, e.g., performance-based earnings/incentives, or social, e.g., peer ratings, they could also be purely *systemic*, which is the focus of this paper. For example, [19] and [1] consider servers that value idleness (the fraction of time they are idle in steady state), and [1] also consider servers that are "workload averse" (that is, they like seeing shorter queues). An important takeaway from these works is that a server's cost function plays a central role in determining the impact of its strategic behavior on the performance of the service system. Notably, we interpret the cost function as an indicator of how "skilled" a server is in providing service – highly skilled servers have lower $c(\mu)$ and $c'(\mu)$.

The systemic benefits that servers experience are directly influenced by the design choices of the system manager, which may include (a) the queue configuration (e.g., pooled with a single queue vs. dedicated with multiple parallel queues), (b) routing policy (which server is responsible for the next job to be served), and (c) staffing policy (what subset of servers are needed to handle demand with certain characteristics). We interpret the influence of such design choices on the servers' strategic behavior as providing *systemic incentives* (as opposed to monetary/social incentives). For example, when servers value idleness, the "Fastest Server First" routing policy may lead to servers slowing down to avoid becoming too busy. Similarly, always assigning extra staff to a busiest division of a service system can lead to servers slowing down to ensure that their division is assigned the extra staff. System policies that are provably optimal under classical models (with nonstrategic servers), could be far from optimal due to the undesirable strategic incentives created by these policies [19]. Thus, it is important for service systems to be designed in a manner that provides the proper incentives for such strategic servers. Accordingly, *our first goal is to is to study the impact of operational design choices on the performance of the system, while taking into account strategic server behavior.*

While there has been considerable research on designing performance-based rewards/penalties that augment a system's operational policies, the systemic incentives created by the policies themselves are much less understood. In many systems, performance-based payments may be counterproductive, e.g., prosocial services [5], or viewed as problematic for other reasons such as perceived unfairness from differential payments to employees [34, 28]. Even when such payments are possible, systemic incentives impact performance, and hence the success of performance-based payments. Further, when constrained by a tight budget, it may be prudent to consider systemic incentives first, before resorting to monetary incentives. This prompts a natural question as to how effective systemic incentives can be. It is not a priori clear whether incentive-aware system design alone can produce strong enough incentives for strategic servers to choose service rates that favorably improve system performance. Thus, *our second goal is to study optimal system design when servers are strategic.*

Finally, while optimal system performance is an important criterion for the manager, an equally important consideration is employee satisfaction [38]. *A third goal of this work is to identify and characterize any trade-offs between these two desiderata.*

1.1. Our Contributions We take a significant step towards these three goals, within a simple utilitarian model of strategic behavior introduced by [19], where servers value idleness when their benefits/wages are independent of their performance/productivity (beyond a contracted minimum). This model is partly supported by the emphasis on employees' perception of idleness in the literature on fair routing policies in service systems [2, 4, 37, 29]. Moreover, valuing idleness can be interpreted to mean that servers prefer having time to perform other tasks outside of the service (sub)system being modeled. This kind of reasoning readily applies to the setting of prosocial service discussed earlier. Moreover, operational policies that are blind to such systemic incentives may have serious, unintended consequences in the healthcare sector. For instance, overworked nurses strategically misreport emergency room beds as occupied, a practice known as "bed hiding" [27]. Even physicians, valuing their leisure time [10], have been observed to partake in such behavior by keeping patients longer than necessary in order to appear busier than they actually are, a practice termed "foot-dragging" [9].

The most significant contribution and the novelty of this paper is that we advance state-of-the-art research that investigates the impact of strategic server behavior on the design of service systems, by moving beyond just analyzing the simple Random routing policy. We consider a large class of routing policies called r-routing policies, that routes an incoming job to server i (operating at rate μ_i) with a probability proportional to μ_i^r. Here, r is a real number that parameterizes the policy class, which includes well-known rate-based routing policies as special cases. For example, $r = 0$ corresponds to Random routing, $r \to +\infty$ corresponds to Fastest Server First (FSF), and $r \to -\infty$ corresponds to Slowest Server First (SSF). While [19] briefly touch upon this class of policies, their analysis is extremely limited (for 2-servers in a pooled system with a centralized queue). Moreover, we focus on *both* the system performance (in terms of mean waiting and response times) as well as employee satisfaction (in terms of the servers' utilities) at symmetric (Nash) equilibrium of the noncooperative game induced between the servers. Section 2 explains all aspects of the model in detail.

Our analysis demonstrates that when servers are strategic, the choice of an r-routing policy, together with the choice of the system configuration (pooled or dedicated), is quite intricate. In particular, we show that many of the managerial insights that apply under the Random routing policy simply break down when more favorable (better performing) r-routing policies are considered. First, in Section 3, within each configuration (pooled or dedicated), we infer the following:

(a) The symmetric equilibrium service rate (when it exists), is a decreasing function of r. If $\underline{r} < 0$ denotes the least value of r for which a symmetric equilibrium exists, then choosing an r-routing policy with $r = \underline{r}$ minimizes the mean waiting time and response time at equilibrium, outperforming the Random routing policy.

(b) The optimality of $\underline{r}$-routing policies is robust in the sense that there exists no better routing policy outside the class of r-routing policies.

(c) The equilibrium utility of the servers is an increasing function of r. Therefore, while a system manager may be tempted to set $r = \underline{r}$ in order to maximize the system performance, such an action would lead to the worst possible utility for the servers. A "moral" dilemma thus emerges when the manager also cares about employee satisfaction.

(d) The worst possible utility for the servers discussed above, increases with the skill level of the servers, i.e., when the cost functions are "smaller" and "slowly increasing". Thus, less skilled servers are more vulnerable to being "exploited" by a system manager.

(e) The optimal equilibrium service rate (when it exists), increases with the number of servers. This adds an interesting complication to the question of optimal staffing level for a given arrival rate (which we do not consider in this paper).

Next, in Section 4, we compare the two configurations and infer the following:

(a) For any r-routing policy, the symmetric equilibrium service rate (when it exists) is larger in the dedicated system than in the pooled system. This is because, a strategic server in the dedicated system who is considering deviating from operating at a service rate corresponding to the symmetric equilibrium of the pooled system will have an incentive to increase her service rate, as she stands to gain more idle time per unit increase of her service rate in the dedicated system (where more idleness is up for grabs, due to the systemic inefficiency) than in the pooled system.

(b) $\underline{r}$ for the dedicated configuration is smaller than that for the pooled configuration.

(c) When choosing between the best pooled and best dedicated configurations (where the respective $\underline{r}$-routing policies are chosen for each configuration), the system manager is better off picking the best dedicated configuration. Therefore, the increase in the equilibrium service rate at which servers work is large enough to overcome the systemic inefficiency of the dedicated configuration.

(d) However, a system manager may not always prefer setting $r = \underline{r}$, if she is concerned with employee satisfaction, as mentioned above. In such a situation, when the choice of r is not straightforward, we show that the choice of the optimal configuration is more complex, and depends on the skill level of the servers, as well as the number of servers in the system.

Throughout, we strive to use analytical techniques as much as possible, especially while working with the dedicated configuration, which is friendlier to closed-form analysis since it is made up of single-server queues. However, for the pooled configuration, even the simple Random routing policy is notoriously challenging [19], and we resort to numerical techniques in our endeavor to analyze arbitrary r-routing policies.

1.2. Related Literature

The question of designing the optimal *architecture* (including the optimal routing policy and server configuration) in many-server systems when servers have fixed, exogenous, service rates is well-studied. In general, this is a very difficult question, because the decisions concerning different aspects of the system architecture are often interdependent, and a joint optimization over all these parameters is quite challenging. For example, it is well-understood that when service rates are fixed, a pooled configuration in which jobs wait in a single queue for service until a server becomes idle, is more efficient than a dedicated configuration in which there are multiple parallel queues, each dedicated to being served by one of the servers [33, 17, 3]. As another example, when strategic servers are not considered, the Fastest Server First (FSF) routing policy is the natural choice for reducing the mean response time (though it is not optimal in general [26, 14]).

There has been a recent burst of activity that investigates the impact of the system architecture of many-server systems on human server behavior, both analytically [19, 1, 16] and empirically [23, 35, 31]. Our work serves as an important addition to the analytical section of this literature by moving beyond investigating the impact of strategic server behavior on system performance, and takes the additional step of optimizing the system architecture under such behavior.

The Fastest Server First routing policy mentioned earlier has already been recognized to be potentially problematic because it may be perceived as "unfair". From an operational standpoint, there is strong indication in the human resource management literature that the perception of fairness affects employee performance [13, 12]. This has motivated finding an optimal "fair" routing policy [2, 37]. Another approach, introduced by [19], is to formulate a model in which the servers choose their service rate in order 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 in which the servers act as strategic players that selfishly maximize their utility. In this work, we adopt this same model to investigate a novel problem, namely that of jointly choosing an optimal routing policy and an optimal queue configuration (dedicated vs. pooled).

Another group of literature that is closely related to our work is the literature on queueing games, surveyed by [21, 20]. 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 on settings where servers can choose their service rate, e.g., [25, 30, 24, 18, 7, 8, 36]. However, in all of the aforementioned papers, there are two firms that derive utility from some monetary compensation per job or per unit of service that they provide. In contrast, we focus on systems with more than two servers *within the same firm*, that are motivated by systemic incentives.

Finally, perhaps the closest previous work to the current paper in spirit is that of [1], who analyze a 2-server system with Random routing, and focus on whether a pooled configuration or a dedicated configuration is more efficient, under different models of work aversion by servers characterized by the amount of discretion they have in choosing their service rates. Some of the key insights from our work are also found in their models, e.g., that servers work faster in a dedicated system configuration. However, we focus on many servers, a much larger class of routing policies, and jointly choosing an optimal routing policy along with the best system configuration.

2. Model

Our goal in this work is to continue the study initiated by [19] into the effects of strategic servers on classical management decisions in service systems, in particular, their structural and operational policy design. While [19] largely focused on a class of "idle-time-order-based" routing policies and asymptotically optimal staffing under this class of policies, in this paper, we tackle the analysis of optimal "rate-based" routing policies under different system configurations. Not only do we characterize optimal routing policies *within* the class of rate-based policies, but numerical examples suggest that these policies are near-optimal among *all* routing policies.

We present our model in several parts. First, we define what we mean by "strategic servers", and recall the framework of [19]. Next, we describe the two canonical configurations of multi-server queues that we consider in this work, namely, a "pooled" system which corresponds to the classic $M/M/N$ model, and a "dedicated" system which consists of N parallel $M/M/1$ queues. Next, we define the class of routing policies that we consider, namely "r*-routing policies*" that was introduced by [19], but only superficially analyzed. We then describe the noncooperative game induced by our setting, and define the performance measures of interest.

2.1. Strategic Servers

In this section, we recall the model for a strategic server introduced by [19]. The strategic behavior of a server depends on the goal of the server, and therefore, the key aspect of the model is the utility function. This model is motivated by a service system staffed by workers who are paid a fixed wage (or obtain a fixed benefit), independent of performance. The two important factors that affect workers' utility are the amount of effort they invest and the amount of idle time they reap. This is similar to the model of "busyness aversion" proposed by [1].

By linearly combining the cost of effort and the idle time, the following form for the utility of server i in a service system with N servers is proposed:

$$U_i(\boldsymbol{\mu}) = I_i(\boldsymbol{\mu}) - c(\mu_i),\ i \in \{1,\dots,N\}, \tag{1}$$

where $\boldsymbol{\mu}$ is a vector of service rates chosen by each server, $I_i(\boldsymbol{\mu})$ is the time-average idle time experienced by server i given $\boldsymbol{\mu}$, and $c(\mu_i)$ is the effort cost of server i. The function c is assumed to be an increasing, convex function, and is identical for all servers.

The form of the utility in (1) captures the inherent tradeoff between idleness and effort. That is, while a higher service rate would result in faster completion of work and could result in more idle time, it would also incur a higher cost of effort. The system manager thus faces a challenge in routing to strategic servers, because, routing jobs to the fastest servers in order to increase throughput and decrease response times would result in a decrease in servers' utility, which reduces their willingness to maintain a fast service rate.

It is quite important to note that this particular model of strategic behavior focuses on one particular tradeoff, namely the one between the cost of effort and the idle time, whose effects on routing and system configuration we study in the rest of this paper. As mentioned earlier, [1] consider tradeoffs between the cost of effort and "busyness aversion" (closely related to the present model) as well as "workload aversion", and how they affect decisions regarding system configuration under a fixed routing policy (Random). While there are many other issues that could be considered in modeling strategic behavior, we focus away from the specifics of the model and instead significantly extend the broader message of [19] that *strategic server behavior can have a non-trivial impact on operational performance, even when the only incentives for the servers are created through system design and there is no performance-based monetary compensation*, by analyzing a more general class of routing policies under two different system configurations.

2.2. Multi-Server Service System We assume that customers arrive to a service system having N servers according to a Poisson process with rate λ. We analyze the impact of strategic servers in the following two system configurations:

(a) *Pooled Configuration*: This is the classical $M/M/N$ setting, where arriving customers are routed to an idle server (if available), and delayed customers (those that arrive to find all servers busy) wait in a central queue that is served according to the First In First Out (FIFO) discipline.

(b) *Dedicated Configuration*: Here, we have N parallel $M/M/1$ queues, each served by a single server according to the FIFO discipline, and arriving customers are routed to one of the N queues to be served immediately (if that server is idle) or wait in that queue (if that server is busy).

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. But, note that each server strategically chooses her service rate to maximize her own (steady state) utility, and so it is not a priori clear what these endogenously determined service rates will be.

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

$$U_i^{sys}(\boldsymbol{\mu};\lambda,N,R) = I_i^{sys}(\boldsymbol{\mu};\lambda,N,R) - c(\mu_i), \qquad i \in \{1,\dots,N\}, \tag{2}$$

where $\boldsymbol{\mu}$ is the vector of service rates, λ is the arrival rate, N is the number of servers, sys is the system configuration (*pooled* or *dedicated*), and R is the routing policy according to which customers are "dispatched" to an idle server or queue (we define this precisely in the following subsection). $I_i^{sys}(\boldsymbol{\mu};\lambda,N,R)$ is the steady state fraction of time that server i is idle in a sys configuration. The function $c(\mu)$ is assumed to be an increasing, convex function with $c'''(\mu) \geq 0$, and represents the server effort cost.

We wish to point out that, as compared with (1), we have emphasized the dependence on the arrival rate λ, number of servers N, the system configuration sys, and routing policy R. In the remainder of this paper, we expose or suppress the dependence on these additional parameters as relevant to the discussion. In particular, note that the steady state fraction of idle time I_i (and hence, the utility function U_i) in (2) depends on both the system configuration sys, and the routing policy R.

2.3. Routing Policy In a *pooled* system, the routing policy determines how a customer who arrives to find more than one idle server will be routed, i.e., which idle server should serve the next job in queue. In a *dedicated* system, the routing policy determines which of the N parallel queues a customer will be routed to upon arrival. In either configuration (pooled or dedicated), there are a variety of routing policies that are feasible for the system manager. In general, she may use information about the order in which the servers became idle (in a pooled system), the queue lengths (in a dedicated system), the rates at which servers have been working, etc. A routing policy could thus be as simple as Random, which chooses an idle server to route to uniformly at random (in the case of a pooled system) or any of the N queues to route to uniformly at random (in the case of a dedicated system), or more complex such as Longest/Shortest Idle Server First (LISF/SISF), Longest/Shortest Queue First, or Fastest/Slowest Server First (FSF/SSF).

In [19], the focus was largely on the class of all "idle-time-order-based" policies in a pooled system, where, quite remarkably, they showed that all such policies are equivalent to Random, and thereafter restricted their analysis to Random. However, towards the end, they briefly introduce the class of "rate-based" routing policies (which includes Random as a special case), and show that in a 2-server pooled system, there exist routing policies in this class that perform better than Random in improving the system performance (average throughput or response time). [1] focus on just random routing, but consider both pooled and dedicated 2-server systems.

In this paper, we focus on a more thorough analysis of "*r-routing policies*", the class of rate-based routing policies parameterized by a number $r \in \mathbb{R}$, introduced by [19], for both dedicated and pooled systems:

(a) *r-routing policy in a pooled system*: Let $\mathcal{I}(t)$ denote the set of idle servers at time t. Under an r-routing policy, at time t, the next job in queue is assigned to idle server $i \in \mathcal{I}(t)$ with probability

$$p_i^{pooled}(\boldsymbol{\mu}, t; r) = \frac{\mu_i^r}{\displaystyle\sum_{j \in \mathcal{I}(t)} \mu_j^r}$$

(b) *r-routing policy in a dedicated system*: Under an r-routing policy, each job, upon arrival, is assigned to the queue served by server i with probability

$$p_i^{dedicated}(\boldsymbol{\mu}; r) = \frac{\mu_i^r}{\displaystyle\sum_{j=1}^{N} \mu_j^r}$$

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.

2.4. The Server Game Given the system configuration (pooled or dedicated) and the routing policy chosen by the system manager, and the form of the server utilities in (2), what ensues is a competition among the servers for the system idle time. The routing policy determines the division of idle time among the servers, and the service rates chosen by the servers determine the total amount of idle time in the system. (A rate-based routing policy would thus enable the servers to also control the division of idle time through the routing policy.)

Therefore, the servers can be modeled as strategic agents or players in a noncooperative game, and the emergent stable operating point of the system is naturally modeled as a Nash equilibrium of this game, which is given by a vector of service rates $\boldsymbol{\mu}^\star$, such that

$$U_i(\mu_i^\star, \boldsymbol{\mu}_{-i}^\star; R) = \max_{\mu_i > \frac{\lambda}{N}} U_i(\mu_i, \boldsymbol{\mu}_{-i}^\star; R), \tag{3}$$

where $\boldsymbol{\mu}_{-i}^\star = (\mu_1^\star, \ldots, \mu_{i-1}^\star, \mu_{i+1}^\star, \ldots, \mu_N^\star)$ denotes the vector of service rates of all the servers except server i. Following the same framework as [19], we exogenously impose the (symmetric) constraint that each server must work at a rate strictly greater than $\frac{\lambda}{N}$ in order to define a product action space that ensures system stability.

We restrict our attention to *symmetric* Nash equilibria. With a slight abuse of notation, we say that $\mu^\star$ is a symmetric Nash equilibrium if $\boldsymbol{\mu}^\star = (\mu^\star, \ldots, \mu^\star)$ is a Nash equilibrium (solves (3)). Throughout, the term "equilibrium service rate" means a symmetric Nash equilibrium service rate. Note that, while our exogenously imposed stability constraint facilitates steady state analysis, it does not exclude any feasible symmetric equilibria.

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. In particular, from an implementation perspective, the entire class of r-routing policies collapses to the Random routing policy *under a symmetric equilibrium.* This sort of fairness is often crucial in service organizations [13, 12, 2]. A second reason for focusing on symmetric equilibria is that even for a simple routing policy such as Random, [19] illustrates that analyzing symmetric equilibria is technically challenging, and it is not clear how to approach asymmetric equilibria. (Echoing [19], we do not rule out the existence of asymmetric equilibria; in fact, they likely exist, especially when we move away from symmetric routing policies such as Random, and it might be of academic interest to study whether they lead to better or worse system performance than their symmetric counterparts.)

2.5. Performance Measures Assuming that a symmetric equilibrium service rate $\mu^\star$ exists, we consider the following three performance measures, when all the servers operate at $\mu^\star$:

(a) *Mean Waiting Time*: Denoted by $\mathbb{E}[W]^{sys}(\mu^\star; \lambda, N)$, this is the average amount of time a customer has to wait before starting service. Mathematically,

$$\begin{aligned} \mathbb{E}[W]^{pooled}(\mu^\star; \lambda, N) &= \frac{1}{N}\left(\frac{ErlC(N, \lambda/\mu^\star)}{\mu^\star - \lambda/N}\right) \\ \mathbb{E}[W]^{dedicated}(\mu^\star; \lambda, N) &= \frac{1}{N}\left(\frac{\lambda/\mu^\star}{\mu^\star - \lambda/N}\right), \end{aligned} \tag{4}$$

where $ErlC(N, \rho)$ denotes the Erlang-C formula, given by:

$$ErlC(N, \rho) = \frac{\frac{\rho^N}{N!}\frac{N}{N-\rho}}{\sum_{j=0}^{N-1}\frac{\rho^j}{j!} + \frac{\rho^N}{N!}\frac{N}{N-\rho}}.$$

(b) *Mean Response Time*: Denoted by $\mathbb{E}[T]^{sys}(\mu^\star; \lambda, N)$, and also known as the *sojourn time*, this is the average amount of time a customer spends in the system (both waiting and service). Mathematically,

$$\mathbb{E}[T]^{sys}(\mu^\star; \lambda, N) = \mathbb{E}[W]^{sys}(\mu^\star; \lambda, N) + \frac{1}{\mu^\star}. \tag{5}$$

(c) *Server Utility*: This is the utility of any server, $U_i^{sys}(\mu^\star;\lambda,N)$ (they are all identical). Mathematically,

$$U_i^{sys}(\mu^\star;\lambda,N) = 1 - \frac{\lambda}{N\mu^\star} - c(\mu^\star). \tag{6}$$

The utility is the same in either configuration, because the steady state idle time for a server in both configurations is the same, when all servers operate at the same rate.

As mentioned earlier, when all the servers operate at a symmetric equilibrium service rate, all r-routing policies reduce to Random, and therefore, the functional form of the above performance measures do not *explicitly* depend on the routing policy. However, the symmetric equilibrium service rate $\mu^\star$ itself depends on all the other parameters of the system (including the routing policy and the system configuration); hence the analysis of these performance measures is nontrivial.

3. Symmetric Equilibrium under r-Routing Policies

In this section, we separately analyze the two configurations (pooled and dedicated) of an N-server service system with strategic servers, under an r-routing policy, with the goal of understanding the existence of a symmetric equilibrium, its uniqueness, and the system performance at equilibrium, while focusing on how the cost-of-effort function of a strategic server affects these aspects. In addition, we also investigate the optimal r-routing policy for which the mean waiting time and response times of the system at equilibrium are minimized. Section 3.1 presents our results for the dedicated configuration, most of which are analytical, with limited dependence on numerical observations. In contrast, when we focus on the pooled configuration in Section 3.2, we resort to obtaining mostly numerical results, given that the steady-state analysis of the $M/M/N$ system, even under random routing, is challenging [19]. The analytical results and numerical observations from this section then drive the discussion regarding the comparison between the dedicated and pooled configurations in Section 4.

3.1. Dedicated Configuration

In this section, we study the symmetric equilibrium service rate in a dedicated N-server system with arrival rate λ, under the class of r-routing policies. Recall that here, we have N parallel $M/M/1$ queues, each served by a single server according to the FIFO discipline, and arriving customers are routed (according to an r-routing policy; see Section 2.3) to one of the N queues to be served immediately (if that server is idle) or wait in that queue (if that server is busy). The outline of this section is as follows.

Section 3.1.1 begins the analysis of the steady state fraction of time a server is idle in such a system, and establishes certain important properties of its first and second derivatives (Theorems 1-3) necessary for equilibrium analysis. Section 3.1.2 continues the analysis by deriving the first order condition for symmetric equilibrium, and then establishes necessary conditions under which an equilibrium exists and is unique (Theorem 4). An important result here is that the symmetric equilibrium service rate is a *decreasing* function of the routing policy parameter r (Theorem 5). Then, Theorem 6 states that the resulting best system performance would also imply the worst equilibrium utility for a server, which opens up an interesting moral dilemma for the system manager. Next, Theorems 7-9 establish more characteristics of symmetric equilibria, including how small an r a system manager can choose while retaining the existence of a symmetric equilibrium, in order to minimize the mean waiting and response times of the system at equilibrium. Finally, Theorem 10 provides the complete characterization of symmetric equilibria for r-routing policies under the dedicated configuration. Section 3.1.3 concludes with numerical examples to understand the impact of the effort cost functions on the performance measures. We delegate all formal proofs to Appendix **??**, but provide proof outlines and/or discuss the key intuition.

Before characterizing the equilibrium service rate (when one exists), it is important to characterize the idle time in a dedicated system. This is the focus of the following section.

3.1.1. The Idle Time of a Tagged Server Due to our focus on symmetric equilibria, we only need to analyze a mildly heterogeneous (in terms of service rates) system. In particular, we need only understand the idle time for a "deviating server" (whom we call the tagged server) when all other servers operate at the same service rate. Our first theorem provides the expression for this idle time.

Theorem 1. *Consider a dedicated N-server system with arrival rate $\lambda > 0$, under an r-routing policy, where $N-1$ servers operate at rate $\mu > \frac{\lambda}{N}$, and a tagged server operates at rate $\mu_1 > 0$. The time-average fraction of time that the tagged server is idle is given by:*

$$I(\mu_1, \mu; \lambda, N, r) = \begin{cases} 1 - \frac{\lambda}{\mu_1} p(\mu_1, \mu; N, r), & \lambda p(\mu_1, \mu; N, r) < \mu_1 \\ 0, & \textit{otherwise} \end{cases}, \tag{7}$$

where p denotes the probability that a job, upon arrival, is assigned to the queue served by the tagged server, given by:

$$p(\mu_1, \mu; N, r) = p_1^{dedicated}((\mu_1, \mu, \ldots, \mu); r) = \frac{\mu_1^r}{\mu_1^r + (N-1)\mu^r}.$$

The proof of Theorem 1 is quite straightforward, given that the tagged server is serving an $M/M/1$ queue with arrival rate λp, by operating at a service rate μ_1. In the first case of (7), the tagged server's queue is stable, and hence, it reaps strictly positive idle time.

While the expression for the idle time given by (7) seems simple, it has a complex dependence on μ_1 through p. Therefore, in order to understand this idle time function more, we derive expressions for its first two derivatives with respect to μ_1 in the following theorem. These results are crucial to the analysis of equilibrium behavior.

Theorem 2. *The first two partial derivatives of the idle time with respect to μ_1 are*

$$\frac{\partial I}{\partial \mu_1} = \begin{cases} \frac{1-I}{\mu_1}\left(1 - r\left(1 - p(\mu_1, \mu; N, r)\right)\right), & \lambda p(\mu_1, \mu; N, r) < \mu_1 \\ \texttt{undefined}, & \lambda p(\mu_1, \mu; N, r) = \mu_1 \\ 0, & \textit{otherwise.} \end{cases} \tag{8}$$

$$\frac{\partial^2 I}{\partial \mu_1^2} = \begin{cases} -\frac{1-I}{\mu_1^2}\left(2r^2 p^2(\mu_1, \mu; N, r) - 3r(r-1)p(\mu_1, \mu; N, r) + (r-1)(r-2)\right), & \lambda p(\mu_1, \mu; N, r) < \mu_1 \\ \texttt{undefined}, & \lambda p(\mu_1, \mu; N, r) = \mu_1 \\ 0, & \textit{otherwise.} \end{cases} \tag{9}$$

Importantly, observe that the right hand side of (8), in the first case, is always positive when $r < 1$, and therefore, the idle time is increasing in the service rate μ_1. This is quite intuitive, since r-routing policies with $r < 0$ send more jobs to slower servers, so, working at a faster rate helps a server both complete its jobs sooner and reduce the number of jobs sent its way. And when $0 < r < 1$, faster servers do get sent more jobs; however, not so much more that it outweighs the benefit of finishing jobs sooner. As we shall see in Section 3.1.2, $r \in (-\infty, 1)$ is the most interesting range when it comes to equilibrium analysis.

Note that at value(s) of μ_1 for which $\lambda p(\mu_1, \mu; N, r) = \mu_1$, the idle time $I(\mu_1, \mu; \lambda, N, r)$ is continuous, but not differentiable. When $r < 1$, it can be shown that there is a unique value $0 < \underline{\mu}_1 < \mu$ such that $\lambda p(\mu_1, \mu; N, r) < \mu_1$ (and hence the tagged server's queue is stable) if and only if $\mu_1 > \underline{\mu}_1$.

Next, we establish certain useful properties of the second derivative of the idle time.

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

(a) When $-7 \le r < 1$, $\frac{\partial^2 I}{\partial \mu_1^2} < 0$ for all $\mu_1 > \underline{\mu}_1$.

(b) *When $r < -7$, there exist thresholds $0 < \underline{\mu}_1^\dagger < \overline{\mu}_1^\dagger$ such that $\frac{\partial^2 I}{\partial \mu_1^2} > 0$ for $\underline{\mu}_1^\dagger < \mu_1 < \overline{\mu}_1^\dagger$ and $\frac{\partial^2 I}{\partial \mu_1^2} < 0$ everywhere else. These thresholds are given by:*

$$\begin{aligned} \underline{\mu}_1^\dagger &= \max\left\{ \mu \left(\frac{(r+4)(r-1) - r\sqrt{(r+7)(r-1)}}{4/(N-1)} \right)^{1/r}, \underline{\mu}_1 \right\}, \\ \overline{\mu}_1^\dagger &= \max\left\{ \mu \left(\frac{(r+4)(r-1) + r\sqrt{(r+7)(r-1)}}{4/(N-1)} \right)^{1/r}, \underline{\mu}_1 \right\} \ < \ \mu. \end{aligned} \tag{10}$$

Theorem 3 implies that when $-7 \le r < 1$, the idle time is a concave function of μ_1, and when $r < -7$, as μ_1 increases from $\underline{\mu}_1$, the idle time is either (a) concave first, then convex, and then concave throughout, or (b) convex first, and then concave throughout, or (c) concave throughout.

3.1.2. Symmetric Equilibrium Analysis for a Dedicated System

The properties of the idle time function presented in the previous section provide the necessary tools to characterize the symmetric equilibrium service rate under an r-routing policy in a dedicated N-server system.

To characterize the symmetric equilibria, we consider the utility of a tagged server, under the mildly heterogeneous setup of Theorem 1, given by:

$$U(\mu_1, \mu; \lambda, N, r) = I(\mu_1, \mu; \lambda, N, r) - c(\mu_1). \tag{11}$$

For a symmetric equilibrium in $(\frac{\lambda}{N}, \infty)$, we explore the first and second order conditions for U as a function of μ_1 to have a maximum in $(0, \infty)$, for a given $\mu > \frac{\lambda}{N}$.

The first order condition for an interior stationary point at μ_1 is given by:

$$\frac{\partial U}{\partial \mu_1} = 0 \quad \Longrightarrow \quad \frac{\partial I}{\partial \mu_1} = c'(\mu_1). \tag{12}$$

Since we are interested in a symmetric equilibrium, we analyze the symmetric first order condition, obtained by plugging in $\mu_1 = \mu$ in (12):

$$\left.\frac{\partial U}{\partial \mu_1}\right|_{\mu_1 = \mu} = 0 \quad \Longrightarrow \quad \frac{\lambda}{N}\left(1 - r\left(\frac{N-1}{N}\right)\right) = \mu^2 c'(\mu). \tag{13}$$

Moreover, since $\mu_1 = \mu$ is greater than the upper threshold $\overline{\mu}_1^\dagger$ of Theorem 3, the idle time is concave at $\mu_1 = \mu$, and hence, so is the utility function. In other words, we have:

$$\left.\frac{\partial^2 U}{\partial \mu_1^2}\right|_{\mu_1 = \mu} < 0.$$

Now, suppose that $\mu^\star > \frac{\lambda}{N}$ satisfies the symmetric first order condition (13). Then, $U(\mu_1, \mu^\star; \lambda, N, r)$ attains a local maximum at $\mu_1 = \mu^\star$. We call such $\mu^\star$ as "candidate" symmetric equilibria. The following three theorems make important observations regarding the existence, uniqueness, and monotonicity of candidate symmetric equilibria and the corresponding utility of the tagged server.

THEOREM 4. *A candidate symmetric equilibrium exists if and only if $r < \overline{r}$, where the threshold $\overline{r}$ is given by*

$$\overline{r} = \frac{N}{N-1}\left(1 - \frac{\lambda}{N} c'\left(\frac{\lambda}{N}\right)\right). \tag{14}$$

Such a candidate symmetric equilibrium is unique.

Theorem 4 follows by observing that, (a) the right hand side of (13) is increasing and convex in μ, whereas the left hand side is a constant, and (b) a candidate symmetric equilibrium must be feasible, i.e., $\mu^\star > \frac{\lambda}{N}$. Moving forward, we denote the candidate symmetric equilibrium, when it exists, as $\mu^\star(r)$ to emphasize its uniqueness and dependence on r.

THEOREM 5. *$\mu^\star(r)$ is a strictly decreasing function of r, growing unboundedly as $r \to -\infty$.*

Theorem 5 follows from the additional observation that the left hand side of (13) is decreasing in r. This is a very important result, because it tells us that in order to minimize the first two performance measures, namely the mean waiting time and mean response time, the manager should pick an r-routing policy with the smallest possible value of r for which $\mu^\star$ is a symmetric equilibrium.

THEOREM 6. *The utility of a server at a candidate symmetric equilibrium $\mu^\star(r)$, given by $U(\mu^\star(r), \mu^\star(r); \lambda, N, r)$, is an increasing function of r when $r < 0$, attains its maximum value at $r = 0$, and is a decreasing function of r when $r > 0$.*

Theorem 6 follows by first observing that, from (6), $U(\mu^\star, \mu^\star; \lambda, N, r) = 1 - \frac{\lambda}{N\mu^\star} - c(\mu^\star)$, a concave function of $\mu^\star$, and then observing that the first order condition for the maximum of this function coincides with the symmetric first order condition (13) when $r = 0$, and finally using the monotonicity of $\mu^\star(r)$ from Theorem 5. This result crucially implies that while the system manager might be able to achieve better system performance by choosing an r-routing policy with smaller $r \le 0$, the corresponding equilibrium utilities of the servers get worse. The extent to which servers are vulnerable to such "exploitation" varies with their cost-of-effort function, which is illustrated numerically in Section 3.1.3.

It remains to discuss conditions under which a candidate symmetric equilibrium will, in fact, be a symmetric equilibrium. By definition, a candidate symmetric equilibrium $\mu^\star(r) > \frac{\lambda}{N}$ is a symmetric equilibrium (satisfying (3)) if and only if $U(\mu_1, \mu^\star(r); \lambda, N, r)$ attains a *global* maximum at $\mu_1 = \mu^\star(r)$ over $\mu_1 \in (0, \infty)$. An obvious necessary condition for this is that the utility at the candidate symmetric equilibrium is at least as much as $\lim_{\mu_1 \to 0+} U(\mu_1, \mu^\star(r); \lambda, N, r)$. Evaluating this lower bound is straightforward:

$$\lim_{\mu_1 \to 0+} U(\mu_1, \mu^\star(r); \lambda, N, r) = \begin{cases} 0, & r < 1 \\ 0, & r = 1 \text{ and } \mu^\star(1) \le \frac{\lambda}{N-1} \\ 1 - \frac{\lambda}{(N-1)\mu^\star(r)}, & r = 1 \text{ and } \mu^\star(1) \ge \frac{\lambda}{N-1} \\ 1, & r > 1. \end{cases}$$

This immediately rules out symmetric equilibria for $r > 1$, since $U(\mu^\star(r), \mu^\star(r); \lambda, N, r) < 1$. Next, when $r = 1$ and $\mu^\star(1) \ge \frac{\lambda}{N-1}$, the idle time function (7) (and hence the utility function), is concave in $(0, \infty)$, and so, $\mu^\star(1)$ is a symmetric equilibrium. For the remaining two cases, the necessary condition translates to $U(\mu^\star(r), \mu^\star(r); \lambda, N, r) \ge 0$. Combined with Theorem 6, this necessary condition implies a key result establishing the continuity of the existence of symmetric equilibria:

THEOREM 7. *If there exists a symmetric equilibrium for a particular $r < 0$, then there exists a symmetric equilibrium for all $r' \in (r, 0]$. Similarly, if there exists a symmetric equilibrium for a particular $r > 0$, then there exists a symmetric equilibrium for all $r' \in [0, r]$.*

Our next result establishes, perhaps surprisingly, that the necessary condition just discussed above is also sufficient for $\mu^\star(r)$ to be a symmetric equilibrium, for $r \in [-7, 1]$.

THEOREM 8. *When $r \in [-7, 1]$, $\mu^\star(r) > \frac{\lambda}{N}$ is a symmetric equilibrium if and only if it satisfies the symmetric first order condition (13), and $U(\mu^\star(r), \mu^\star(r); \lambda, N, r) \ge 0$.*

From Theorem 7, it follows that when r decreases further (beyond -7), the candidate symmetric equilibrium continues to be a symmetric equilibrium (as long as $U(\mu^\star(r), \mu^\star(r); \lambda, N, r) \geq 0$) until a certain threshold value, $\underline{r} < -7$, is reached, beyond which point, it ceases to be a symmetric equilibrium. This is due to the eventual emergence of a new global maximum $\mu_1^\star < \mu^\star(\underline{r})$ for $U(\mu_1, \mu^\star(\underline{r}); \lambda, N, \underline{r})$. This is quite intuitive, because, from Theorem 5, the candidate symmetric equilibrium $\mu^\star(r)$ keeps increasing as r decreases, and at a certain point, the cost of operating at a high $\mu^\star(r)$ becomes too prohibitive that the server would rather work at a lower service rate and suffer the consequences of the resulting additional workload on its idle time. Thus, $\underline{r}$ depends crucially on the cost of effort function, as our next result shows:

Theorem 9. *There exists a unique $\underline{r} < -7$, and a corresponding unique $\mu_1^\star < \mu^\star(\underline{r})$, such that $U(\mu_1^\star, \mu^\star(\underline{r}); \lambda, N, \underline{r}) = U(\mu^\star(\underline{r}), \mu^\star(\underline{r}); \lambda, N, \underline{r})$. This pair $(\underline{r}, \mu_1^\star)$ simultaneously solves the following equations:*

$$\frac{\lambda}{\mu_1^\star} p(\mu_1^\star, \mu^\star(\underline{r}); N, r) \;=\; \frac{\lambda}{N\mu^\star(\underline{r})} + c(\mu^\star(\underline{r})) - c(\mu_1^\star) \;=\; \frac{2\lambda\mu_1^\star c'(\mu_1^\star)}{\lambda(1-\underline{r}) - \sqrt{\lambda^2(1-\underline{r})^2 + 4\lambda\underline{r}(\mu_1^\star)^2 c'(\mu_1^\star)}}. \tag{15}$$

Combining all of these results gives our final, main theorem that completely characterizes symmetric equilibria under r-routing policies under a dedicated configuration:

Theorem 10. *For any r-routing policy, $\mu^\star > \frac{\lambda}{N}$ is a symmetric equilibrium if and only if $r \in [\underline{r}, 1]$, $\mu^\star$ satisfies the symmetric first order condition (13), and $U(\mu^\star, \mu^\star; \lambda, N, r) \geq 0$.*

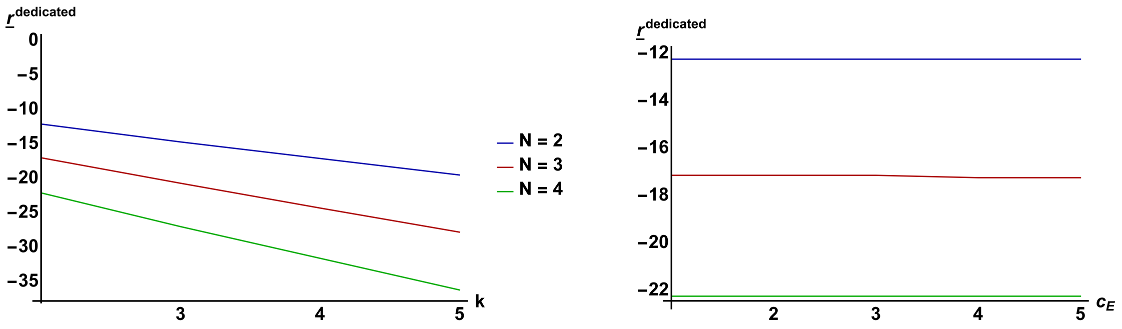
Since the characterization of $\underline{r}$ provided by Theorem 9 is complex, and it is not possible to obtain a general closed form expression for $\underline{r}$, we defer further discussion regarding $\underline{r}$ to the next section, Section 3.1.3, where we numerically study the impact of the cost function on (a) $\underline{r}$, (b) $\overline{\mu}^\star = \mu^\star(\underline{r})$, and (c) U$(\overline{\mu}^\star, \overline{\mu}^\star; \lambda, N, \underline{r})$.

3.1.3. Numerical Examples In the previous section, we established several important theoretical results, including Theorem 5, which established that the candidate symmetric equilibrium is a decreasing function of r. To the system manager, this means that for the best system performance (mean waiting time or mean response time), she must pick an r-routing policy with $r = \underline{r}$, which is the smallest possible value of r for which the corresponding candidate symmetric equilibrium $\mu^\star$ is indeed a symmetric equilibrium. In light of this insight, the unanswered question is how $\underline{r}$ depends on the system parameters (such as the number of servers), and the server characteristics (their cost of effort function). Therefore, we discuss a few numerical examples in this section to provide intuition. In addition, we point out some interesting characteristics that emerge as a consequence of strategic server behavior.

We consider the family of cost functions parameterized by $c_E \geq 1$ and $k \geq 1$, given by $c(\mu) = c_E \mu^k$. Since the idle time cannot exceed 1, μ cannot exceed $(1/c_E)^{1/k}$; otherwise, a server's utility (given by (2)) will never be positive. Therefore, we are operating within the range $\mu \in [0, 1]$, wherein we can interpret that "highly skilled" servers have a smaller c_E, and a larger k, since either of these results in a lower cost of effort when $\mu \in [0, 1]$. (An alternate interpretation of c_E is that it is a measure of the trade-off between idle time and cost.)

We are interested in how the parameters c_E and k of the cost of effort function affect the following three quantities: (a) $\underline{r}$, the minimum possible value of r for which there exists a symmetric equilibrium, (b) $\overline{\mu}^\star$, the corresponding maximum symmetric equilibrium service rate, and (c) $U(\overline{\mu}^\star, \overline{\mu}^\star; \lambda, N, \underline{r})$, the utility of a server at the maximum symmetric equilibrium service rate.

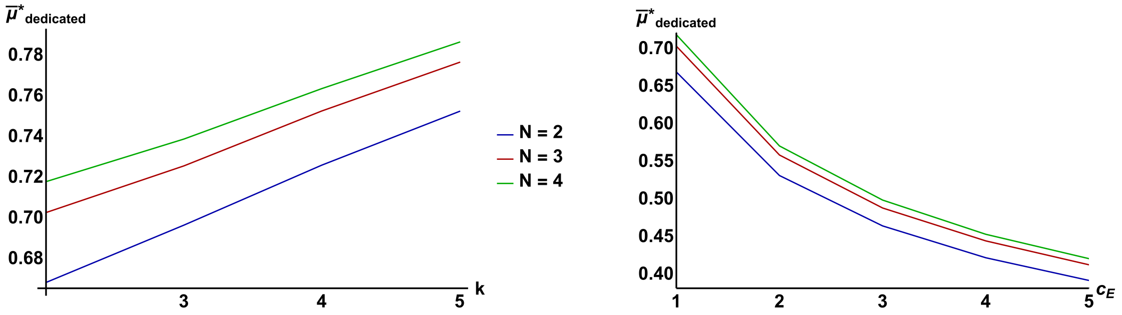
We present three sets of graphs below (Figures 1-3), one for each of the above three quantities of interest. Within each set, we have two graphs, one that illustrates the impact of k while fixing $c_E = 1$, and another that illustrates the impact of c_E while fixing $k = 2$. Finally, within each graph, we plot three curves, corresponding to $N = 2, 3, 4$ respectively. (We set $\lambda = \frac{1}{6}$, which ensures that



(a) Varying k with $c_E = 1$

(b) Varying c_E with $k = 2$

FIGURE 1. Minimum possible r for which the dedicated system configuration admits a symmetric equilibrium, denoted by $\underline{r}^{dedicated}$, for $N = 2$ (blue), $N = 3$ (red), and $N = 4$ (green).



(a) Varying k with $c_E = 1$

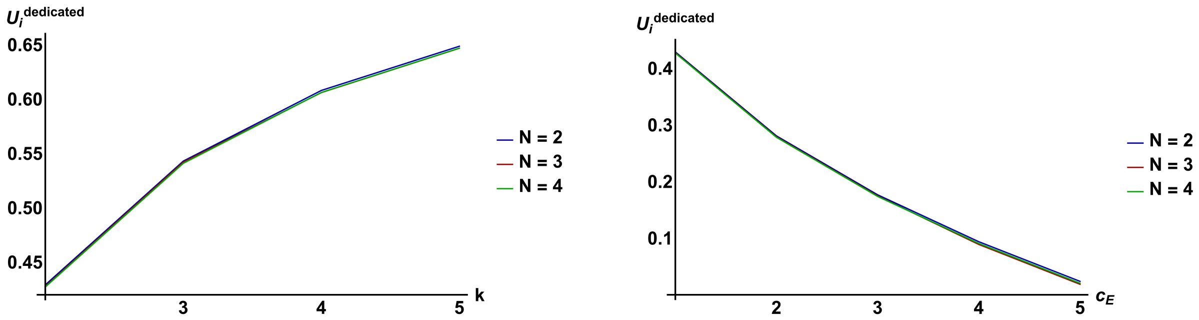
(b) Varying c_E with $k = 2$

FIGURE 2. Equilibrium service rate corresponding to $\underline{r}^{dedicated}$, denoted by $\overline{\mu}^{\star}_{dedicated}$, for $N = 2$ (blue), $N = 3$ (red), and $N = 4$ (green).

$\overline{r}$, the upper bound of Theorem 4, is greater than 1 in each of our examples, and therefore doesn't exclude the Random routing policy ($r = 0$) from the set of feasible r-routing policies.)

The first observation is that $\overline{\mu}^{\star}_{dedicated}$, the maximum possible equilibrium service rate in the dedicated configuration increases with the skill level of the servers. For example, this happens with increasing k in Figure 2(a) and with decreasing c_E in Figure 2(b). The second observation is that $U_i^{dedicated}(\overline{\mu}^{\star}_{dedicated}; \lambda, N)$, the equilibrium server utility, also increases with the skill level of the servers. Once again, this can be noticed with increasing k in Figure 3(a) and with decreasing c_E in Figure 3(b). This goes to show that in the dedicated system, servers that are less skilled are more vulnerable to being "exploited" by a system manager who desires the best possible equilibrium performance in terms of mean waiting and response times. The third observation is that as the number of servers N increases, $\overline{\mu}^{\star}_{dedicated}$ also increases (Figure 2), which means that servers are willing to work at higher service rates when there are more of them. This introduces an additional complication when considering the question of optimal staffing (which we do not deal with in this paper). Finally, something that is perhaps interesting to note, is that $\underline{r}^{dedicated}$, the smallest r for which the dedicated system admits a symmetric equilibrium, is not very sensitive to varying c_E (Figure 1(b)). In addition, the equilibrium server utilities seem insensitive to the number of servers in the system (Figure 3).

3.2. Pooled Configuration In this section, we study the symmetric equilibrium service rate in a pooled N-server system with arrival rate λ, under the class of r-routing policies. Recall that here, we are in the classical $M/M/N$ setting, where arriving customers are routed to an idle server (if available) according to an r-routing policy (see Section 2.3), and delayed customers (those



(a) Varying k with $c_E = 1$

(b) Varying c_E with $k = 2$

FIGURE 3. Equilibrium server utility corresponding to $\underline{r}^{dedicated}$, given by $U_i^{dedicated}(\overline{\mu}^\star_{dedicated};\lambda,N)$, for $N=2$ (blue), $N=3$ (red), and $N=4$ (green).

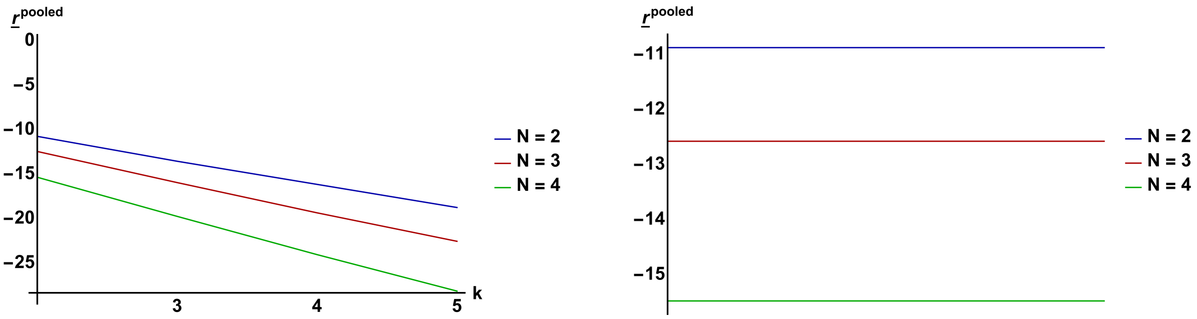
that arrive to find all servers busy) wait in a central queue that is served according to the FIFO discipline.

Analytical results concerning symmetric equilibrium service rates in an $M/M/N$ model under an r-routing policy with strategic servers are not known, except when $r=0$ (corresponding to Random routing) or when $N=2$ (the 2-server system). Both these special scenarios were studied by [19], and we refer to their results therein. Their work also illustrates the complexity of such an analysis, even for the simplest of policies, Random. However, given our eventual goal of comparing the dedicated and pooled system configurations, and the resulting necessity to obtain parallel results for the pooled system, we resort to extending the limited analytical results of [19] for 2 servers to $N>2$ servers numerically.

Due to our focus on symmetric equilibria, similar to our approach in Section 3.1.1 for the dedicated configuration, we need only focus on a mildly heterogeneous (in terms of service rates) $M/M/N$ system, where all the servers except a "deviating server" (or the tagged server), operate at the same service rate. Since it is intractable to obtain a general closed form expression for the idle time of the tagged server, we write down the corresponding Markov chain, and feed the steady-state equations into Mathematica, which is then able to solve the system (and hence obtain the idle time of the tagged server and its derivatives) in closed form, but only for a given value of N. Armed with this mechanism, we carry out the symmetric equilibrium analysis for the pooled configuration numerically, with the same objective in mind – studying the impact of the server's cost of effort function on the performance measures of the system under any r-routing policy, as well as the optimal choice of r.

3.2.1. Numerical Examples Qualitatively, we find that a pooled system mirrors some of the important characteristics of a dedicated system. For example, the candidate symmetric equilibrium service rate is decreasing in r, and there exists an $\underline{r}<0$ such that a candidate symmetric equilibrium ceases to be a symmetric equilibrium when $r<\underline{r}$. Accordingly, and for ease of comparison, we use the exact same setting in Section 3.1.3 and present the same three sets of graphs below (Figures 4-6), except that these graphs are now for the pooled system, but with the same parameters used in Section 3.1.3.

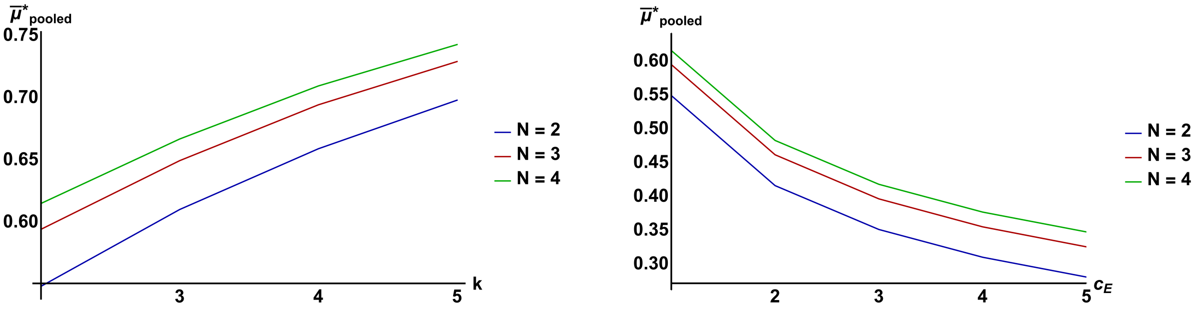
All the observations we made from the numerical examples in the dedicated system in Section 3.1.3 are qualitatively preserved here, in the pooled system. Most importantly, both $\overline{\mu}^\star_{pooled}$ and $U_i^{pooled}(\overline{\mu}^\star_{pooled};\lambda,N)$ increase with the skill level of the servers (either with increasing k; see Figures 5(a), 6(a), or with decreasing c_E; see Figures 5(b), 6(b)), and $\overline{\mu}^\star_{pooled}$ increases with the number of servers (Figure 5). Thus, the managerial implications are also preserved. In other words, less skilled servers are more vulnerable to be "exploited", and the optimal staffing problem becomes nontrivial because the best equilibrium service rate increases with the number of servers.



(a) Varying k with $c_E = 1$

(b) Varying c_E with $k = 2$

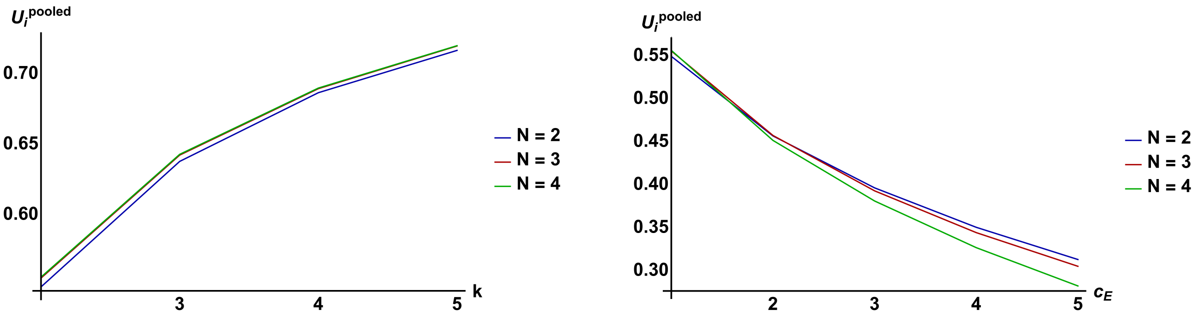
FIGURE 4. Minimum possible r for which the pooled system configuration admits a symmetric equilibrium, denoted by $\underline{r}^{pooled}$, for $N = 2$ (blue), $N = 3$ (red), and $N = 4$ (green).



(a) Varying k with $c_E = 1$

(b) Varying c_E with $k = 2$

FIGURE 5. Equilibrium service rate corresponding to $\underline{r}^{pooled}$, denoted by $\overline{\mu}^{\star}_{pooled}$, for $N = 2$ (blue), $N = 3$ (red), and $N = 4$ (green).



(a) Varying k with $c_E = 1$

(b) Varying c_E with $k = 2$

FIGURE 6. Equilibrium server utility corresponding to $\underline{r}^{pooled}$, given by $U_i^{pooled}(\overline{\mu}^{\star}_{pooled}; \lambda, N)$, for $N = 2$ (blue), $N = 3$ (red), and $N = 4$ (green).

4. Comparing Dedicated vs. Pooled Configurations

In this section, we compare the three performance measures of interest (defined in Section 2.5) across the two system configurations, i.e., pooled and dedicated, in order to understand when one configuration is better than the other. We present these comparisons in two parts. First, in Section 4.1, we compare the performance of pooled vs. dedicated systems at symmetric equilibrium, under the same r-routing policy, for different values of r. Since this kind of comparison only makes sense when considering a limited set of r-routing policies for which a symmetric equilibrium exists for *both* configurations, in Section 4.2, we also compare the performance of the *best* pooled system vs. the *best* dedicated system, where, for each configuration, we pick the r-routing policy with the maximum symmetric equilibrium service rate. Finally, in Section 4.3, we discuss what these results mean from a managerial perspective, in choosing the architecture of a service system when servers are strategic.

4.1. Dedicated vs. Pooled Configurations Under Same r-Routing Policy Our first observation is that for a given r-routing policy, the candidate equilibrium service rate (obtained by solving the first order condition for symmetric equilibrium) is always larger in the dedicated configuration than in the pooled configuration. The following theorem establishes this analytically when $N = 2$.

THEOREM 11. *For any r-routing policy, let $\mu^\star_{pooled}$ and $\mu^\star_{dedicated}$ denote the candidate symmetric equilibrium service rates for the pooled and dedicated system configurations, respectively (when they exist). When $N = 2$, $\mu^\star_{pooled} < \mu^\star_{dedicated}$.*

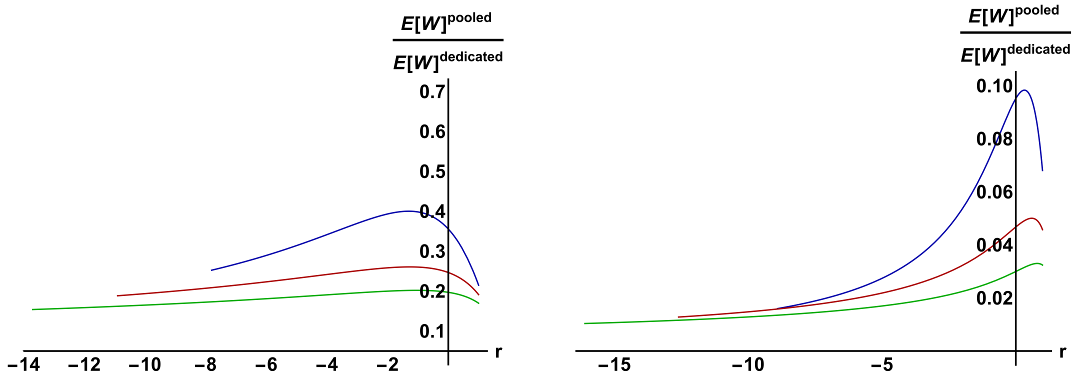
Intuitively, this is because a strategic server in the dedicated system who is considering deviating from operating at a service rate corresponding to the symmetric equilibrium of the pooled system will have an incentive to increase her service rate, because the server stands to gain more idle time per unit increase of her service rate in the dedicated system (where more idleness is up for grabs, due to the systemic inefficiency) than in the pooled system. We omit the proof due to space constraints.

In other words, Theorem 11 implies that at equilibrium, servers work at a higher rate under a dedicated configuration than in a pooled configuration. However, given that a dedicated system is inherently less efficient than a pooled system, the question left unanswered is, under what circumstances can the higher equilibrium service rate in the dedicated configuration overcome the systemic inefficiency to provide a better mean waiting time or mean response time? Also, how do the equilibrium server utilities compare across the two configurations? We answer these questions with the help of numerical examples.

4.1.1. Numerical Examples We retain the same setting and system parameters used for the numerical examples in Sections 3.1.3 and 3.2.1, and present three sets of graphs below (Figures 7-9), one for each of the three performance measures of interest – the mean waiting time at equilibrium, the mean response time at equilibrium, and a server's utility at equilibrium (see Section 2.5 for their definitions). Within each set, we have two graphs (one each for $N = 2, 3$) that show the ratio of the performance measure in the pooled system to that in the dedicated system, as a function of r. Finally, within each graph, we plot three curves, corresponding to $k = 2, 3, 4$ respectively. For all these examples, we set $c_E = 1$.

The first observation is that the pooled configuration is always better than the dedicated configuration, when it comes to the mean waiting time (Figure 7). Therefore, even though the servers work at a higher service rate in the dedicated system, it is not enough to overcome the systemic inefficiency of the dedicated system, as far as the waiting time is concerned. However, the story is different when the mean response time is considered (Figure 8). While the pooled configuration still fares better than the dedicated configuration for $r = 0$ (random routing), as r becomes smaller, the dedicated configuration overtakes the pooled configuration, significantly so when the servers are less skilled ($k = 1, 2$ in Figure 8(a) and $k = 1$ in Figure 8(b)). Finally, the equilibrium server utilities are higher in the dedicated configuration for $r = 0$ (random routing), but as r becomes smaller, this gets reversed quickly (Figure 9).

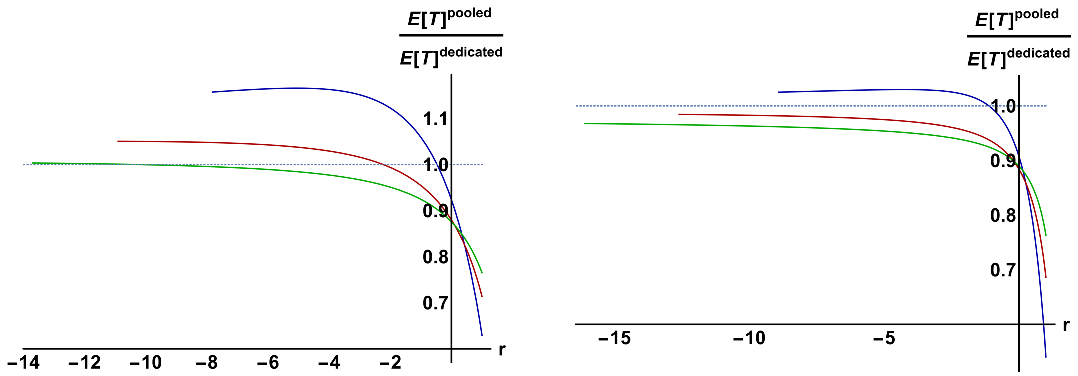
These observations suggest that a manager's decision to choose between a pooled vs. dedicated configuration is complex, and is affected by the answers to the following questions: (a) Is it important to have a smaller waiting time or a smaller response time? (b) How skilled are the servers? and (c) What r-routing policy is to be implemented?



(a) $N = 2$

(b) $N = 3$

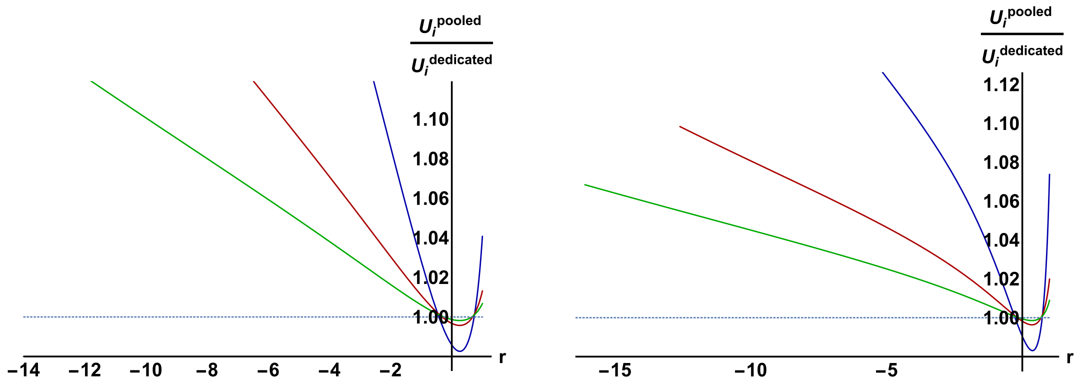
FIGURE 7. $\mathbb{E}[W]^{pooled}/\mathbb{E}[W]^{dedicated}$ as a function of r, for $k = 1$ (blue), $k = 2$ (red), and $k = 3$ (green).



(a) $N = 2$

(b) $N = 3$

FIGURE 8. $\mathbb{E}[T]^{pooled}/\mathbb{E}[T]^{dedicated}$ as a function of r, for $k = 1$ (blue), $k = 2$ (red), and $k = 3$ (green). The horizontal dotted line corresponds to "breaking even", shown for ease of interpreting the other three curves.

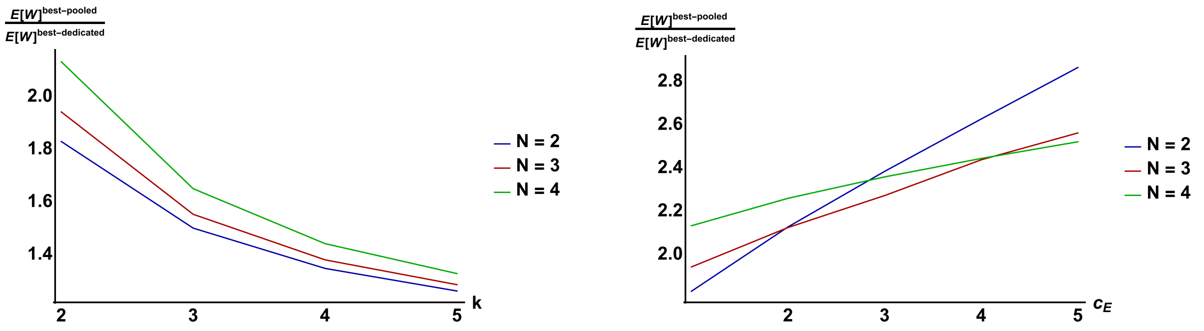


(a) $N = 2$

(b) $N = 3$

FIGURE 9. $U_i^{pooled}/U_i^{dedicated}$ as a function of r, for $k = 1$ (blue), $k = 2$ (red), and $k = 3$ (green). The horizontal dotted line corresponds to "breaking even", shown for ease of interpreting the other three curves.

4.2. Best Dedicated vs. Best Pooled Configurations While the comparison of performance measures between the dedicated and pooled configurations for the same r-routing policy is useful for isolating the effect of the system configuration, from a managerial perspective, there



(a) Varying k with $c_E = 1$

(b) Varying c_E with $k = 2$

FIGURE 10. $\mathbb{E}[W]^{best-pooled}/\mathbb{E}[W]^{best-dedicated}$, for $N = 2$ (blue), $N = 3$ (red), and $N = 4$ (green).

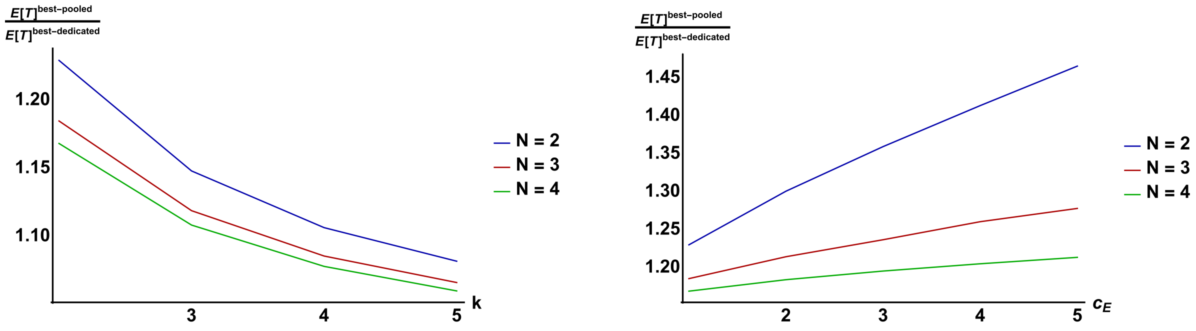
is no reason to enforce (a priori) that the same r-routing policy be considered when comparing pooled and dedicated configurations. Moreover, the previous comparison restricts the range of r to those for which a symmetric equilibrium service rate exists for *both* configurations. Instead, it may be more meaningful to the system manager to compare the performance of the pooled and dedicated configurations under their respective *best* r-routing policies, i.e., choosing $r = \underline{r}^{pooled}$ for the pooled system and $r = \underline{r}^{dedicated}$ for the dedicated system. This comparison is the focus of this section. Once again, we investigate with the help of numerical examples.

4.2.1. Numerical Examples As before, we present three sets of graphs below (Figures 10-12), one for each of the three performance measures of interest – the mean waiting time at equilibrium, the mean response time at equilibrium, and a server's utility at equilibrium (see Section 2.5 for their definitions). Within each set, we have two graphs that show the ratio of the performance measure in the *best* pooled system to that in the *best* dedicated system, one that illustrates the impact of k while fixing $c_E = 1$, and another that illustrates the impact of c_E while fixing $k = 2$. Finally, within each graph, we plot three curves, corresponding to $N = 2, 3, 4$ respectively.

From Figures 10 and 11, we observe that, the *best* dedicated configuration outperforms the *best* pooled configuration in terms of both the mean waiting time and the mean response time, with the largest advantage observed for the least skilled servers (lower k or higher c_E). This is in sharp contrast with the comparison in Section 4.1.1, where for any fixed r-routing policy, characterizing the optimal configuration was more complex. Next, from Figure 12, it can be seen that the equilibrium server utility in the best pooled configuration is more than that in the best dedicated configuration, and this advantage is more pronounced, once again, for the least skilled servers.

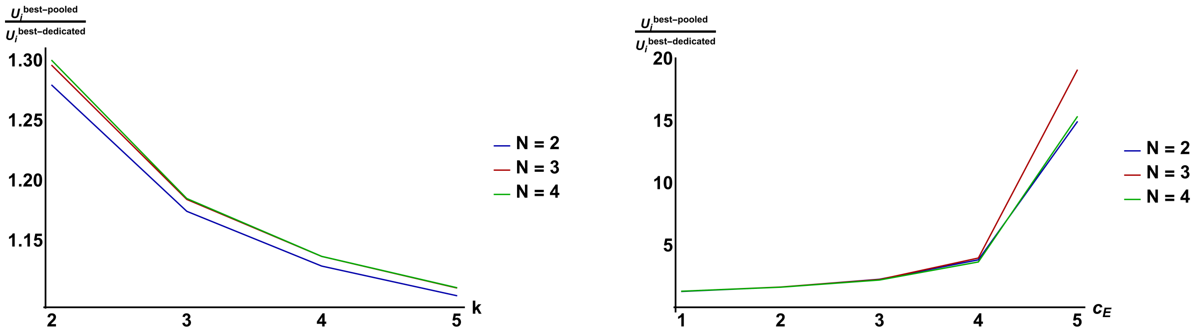
4.3. Managerial Implications Even though the observations from the comparison between the best pooled and best dedicated configurations suggest that the system manager is better off choosing a dedicated configuration, this insight is only applicable when she can pick an r-routing policy with $r = \underline{r}^{dedicated}$ without any reservation, which is unlikely to occur in practice. As we inferred from Theorem 6, and Sections 3.1.3 and 3.2.1, the system manager faces a moral dilemma in her decision regarding which r-routing policy to choose. While choosing $r = \underline{r}$ may lead to the best possible mean waiting and response times, it would result in the worst equilibrium utility for the servers and hence degrade employee satisfaction.

Therefore, in choosing an r-routing policy, the manager must carefully balance her interests concerning optimal system performance with her interest in employee satisfaction and retention. Taken together with the skill level of the servers, this decision, in turn, will affect the optimal system configuration (pooled or dedicated).



(a) Varying k with $c_E = 1$ (b) Varying c_E with $k = 2$

FIGURE 11. $\mathbb{E}[T]^{best-pooled}/\mathbb{E}[T]^{best-dedicated}$, for $N = 2$ (blue), $N = 3$ (red), and $N = 4$ (green).



(a) Varying k with $c_E = 1$ (b) Varying c_E with $k = 2$

FIGURE 12. $U_i^{best-pooled}/U_i^{best-dedicated}$, for $N = 2$ (blue), $N = 3$ (red), and $N = 4$ (green).

5. Concluding Remarks In this paper, we studied the joint optimization over the routing policy (from a large class of rate-based routing policies) and the system configuration (pooled vs. dedicated) when designing the architecture of many-server service systems, under strategic server behavior. Using both analytical and numerical techniques, we established important managerial insights that can help a system manager make an informed decision when faced with a challenging situation. In particular, we observed that the choice of optimal system configuration has a complex dependence on multiple factors that include the cost function of the server, the number of servers, and the choice of performance metric (response time vs. waiting time).

Our analysis of a system with homogeneous servers, where we observed that servers that are less skilled are better grouped into a dedicated configuration (Figure 8) to minimize the mean response time, helps make the following informed conjecture when dealing with heterogeneous servers (that have different cost functions or “skill levels”). One might expect that the optimal system configuration under a heterogeneous set of servers would consist of a hybrid of both pooled and dedicated queues, wherein a common queue feeds those servers whose skill levels (derived in an appropriate manner from their cost functions) are higher than a certain threshold, and the rest of the servers have their own dedicated queues. It would be interesting to see if this intuition can be validated.

We also did not explore the classical question of choosing the optimal staffing levels. Our result that the equilibrium service rate increases with the number of servers adds an interesting angle to this question, and it would be nice to find the optimal, or asymptotically optimal staffing policies under r-routing policies, and perhaps solve the joint optimization problem of staffing and routing.

References

[1] Armony, M., G. Roels, H. Song. 2017. Pooling queues with discretionary service capacity. Available at SSRN: https://ssrn.com/abstract=2951959.

[2] Armony, M., A. R. Ward. 2010. Fair dynamic routing in large-scale heterogeneous-server systems. *Oper. Res.* **58** 624–637.

[3] Ata, Barış, Jan A Van Mieghem. 2009. The value of partial resource pooling: Should a service network be integrated or product-focused? *Management Science* **55**(1) 115–131.

[4] Atar, R., Y. Y. Shaki, A. Shwartz. 2011. A blind policy for equalizing cumulative idleness. *Queueing Syst.* **67**(4) 275–293.

[5] Bénabou, R., J. Tirole. 2006. Incentives and prosocial behavior. *American Economic Review* **96**(5) 1652–1678.

[6] Bensinger, G. 2018. Uber drivers take riders the long way—at uber's expense. *The Wall Street Journal* URL https://www.wsj.com/articles/uber-drivers-take-riders-the-long-wayat-ubers-expense-1534152602.

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

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

[9] Chan, D. C. 2016. Teamwork and moral hazard: Evidence from the emergency department. *Journal of Political Economy* **124**(3) 734–770.

[10] Chan, D. C. 2018. The efficiency of slacking off: Evidence from the emergency department. *Econometrica* **86**(3) 997–1030.

[11] Chetty, R., E. Saez, L. Sandor. 2014. What policies increase prosocial behavior? an experiment with referees at the journal of public economics. *Journal of Economic Perspectives* **28**(3) 169–88.

[12] Cohen-Charash, Y., P. E. Spector. 2001. The role of justice in organizations: A meta-analysis. *Organ. Behav. and Hum. Dec.* **86**(2) 278–321.

[13] Colquitt, J. A., D. E. Conlon, M. J. Wesson, C. O. L. H. Porter, K. Y. Ng. 2001. Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *J. Appl. Psychol.* **86**(3) 425–445.

[14] de Véricourt, F., Y.-P. Zhou. 2005. Managing response time in a call-routing problem with service failure. *Oper. Res.* **53**(6) 968–981.

[15] Delfino, D. 2017. Make money as an instacart shopper: What to expect. *NerdWallet* URL https://www.nerdwallet.com/blog/finance/make-money-as-an-instacart-shopper-what-to-expect/.

[16] Do, H. T., M. Shunko, M. T. Lucas, D. C. Novak. 2018. Impact of Behavioral Factors on Performance of Multi-Server Queueing Systems. *Prod. Oper. Manag.* **27**(8) 1553–1573.

[17] Gans, Noah, Ger Koole, Avishai Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141.

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

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

[20] Hassin, R. 2016. *Rational Queueing*. Chapman and Hall/CRC.

[21] Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer.

[22] Jin, X., F. Zhang, A. V. Vasilakos, Z. Liu. 2016. Green data centers: A survey, perspectives, and future directions. *CoRR* **abs/1608.00687**. URL http://arxiv.org/abs/1608.00687.

[23] Jouini, Oualid, Yves Dallery, Rabie Nait-Abdallah. 2008. Analysis of the impact of team-based organizations in call center management. *Management Science* **54**(2) 400–414.

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

[25] Levhari, D., I. Luski. 1978. Duopoly pricing and waiting lines. *European Economic Review* **11**(1) 17–35.

[26] Lin, W., P. Kumar. 1984. Optimal control of a queueing system with two heterogeneous servers. *IEEE Trans. Autom. Contr.* **29**(8) 696–703.

[27] Meisel, Z., J. Pines. 2008. Waiting doom: How hospitals are killing E.R. patients. *Slate* URL https://slate.com/technology/2008/07/how-hospitals-are-killing-e-r-patients.html.

[28] Miller, S. 2016. How to counter employee perceptions of income inequality. *Society for Human Resource Management* URL https://www.shrm.org/hr-today/news/hr-magazine/0516/pages/0516-fair-compensation.asp

[29] Reed, J., Y. Shaki. 2014. A fair policy for the G/GI/N queue with multiple server pools. Preprint.

[30] Reitman, D. 1991. Endogenous Quality Differentiation in Congested Markets. *The Journal of Industrial Economics* **39**(6) 621–647.

[31] Shunko, M., J. Niederhoff, Y. Rosokha. 2018. Humans are not machines: The behavioral impact of queueing design on service time. *Management Science* **64**(1) 453–473.

[32] Sieben, I. J. P., A. de Grip, D. van Jaarsveld. 2005. *Employment and industrial relations in the Dutch call center sector*. No. 4E in ROA Reports, Researchcentrum voor Onderwijs en Arbeidsmarkt, Faculteit der Economische Wetenschappen.

[33] Smith, David R, Ward Whitt. 1981. Resource sharing for efficiency in traffic systems. *Bell Labs Technical Journal* **60**(1) 39–55.

[34] Solmon, L. C., M. Podgursky. 2000. The pros and cons of performance-based compensation. *Education Resources Information Center* URL https://eric.ed.gov/?id=ED445393.

[35] Song, Hummy, Anita L Tucker, Karen L Murrell. 2015. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* **61**(12) 3032–3053.

[36] Wang, H., T. L. Olsen, G. Liu. 2018. Service capacity competition with peak arrivals and delay sensitive customers. *Omega* **77** 80–95.

[37] Ward, A. R., M. Armony. 2013. Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Oper. Res.* **61** 228–243.

[38] Yee, R. W. Y., A. C. L. Yeung, T. C. E. Cheng. 2008. The impact of employee satisfaction on quality and profitability in high-contact service industries. *Journal of Operations Management* **26**(5) 651–668.