On the Architecture of Service Systems when Servers are Strategic

Ragavendran Gopalakrishnan
Cornell University, rg584@cornell.edu

Extended Abstract

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 strategic. In today’s world of the shared economy, with services increasingly being offered by crowdworkers, e.g., ridesharing (e.g., Lyft, Uber), grocery delivery (e.g., Instacart), etc., it becomes more important to take into account strategic server behavior when making managerial decisions regarding the service operations.

Within this literature, a commonly adopted model for strategic servers is to assume that servers are characterized by a “cost of effort function” that specifies the cost they incur for operating at a particular service rate, and that servers strategically choose their service rates in order to balance their costs of effort against the tangible benefits they get in return for their service. While such benefits are often thought to be monetary, e.g., performance-based pay, they could also be systemic. For example, Gopalakrishnan et al. (2016) and Armony et al. (2017) consider servers that value idleness (the fraction of time they are idle in steady state), and Armony et al. (2017) also consider servers that are “workload averse” (i.e., they like seeing shorter queues). An important takeaway from these works is that a server’s cost of effort function plays a central role in determining the impact of its strategic behavior on the performance of the service system.

Several aspects of system design affect the incentives garnered by strategic servers. These include the system configuration (pooled, dedicated, or hybrid), the routing policy (how jobs are routed to servers), and the staffing policy (how many servers to hire). Together, they comprise the architecture of the service system. The lofty goal of this body of research, therefore, is to find the optimal system architecture for a given set of strategic servers.

This paper takes an important step in this direction, by simultaneously investigating the impact of a large class of routing policies (“rate-based” routing policies), together with the choice of system configuration (between pooled and dedicated) on the performance of service systems with homogeneous strategic servers. Even though we limit our analysis to homogeneous servers, we focus on how different performance measures of interest vary with properties of the cost function.
such as its convexity (which can be interpreted as a proxy for a server’s skill level). This then enables us to make informed conjectures about how a system manager might handle the architecture of a service system with heterogeneous servers.

**Contributions:** The most important contribution and the novelty of this paper is that we advance the state-of-the-art research that investigates the impact of strategic server behavior on the architecture 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 a server with a probability proportional to its service rate raised to the power $r$. This class of routing policies 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 Gopalakrishnan et al. (2016) briefly touched upon this class of policies, their analysis was extremely limited (for 2-servers in a pooled system with a centralized queue). We do carry out all our analyses in their model of strategic server behavior (that is, servers value idle time and incur a cost of effort), and we focus on the performance of the system (in terms of mean waiting and response times) at symmetric (Nash) equilibrium service rates.

Our work 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, 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 $r < 0$ denotes the least value of $r$ for which a symmetric equilibrium exists, then choosing an $r$-routing policy with $r = r$ minimizes the mean waiting time and response time at equilibrium. Clearly, such an $r$-routing policy would outperform the Random routing policy.

(b) The equilibrium utility of the servers is an increasing function of $r$. Therefore, while a system manager may be tempted to set $r = r$ 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 for a system manager who may also care about employee satisfaction.

(c) 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.

(d) The symmetric equilibrium service rate (when it exists), increases with the number of servers. This adds an interesting angle to the question of optimal staffing level for a given arrival rate.
Next, 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) $r$ 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 $r$-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 $r = r$, 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.

**Future Work:** Our analysis of a system with homogeneous servers helps make the following informed conjecture when dealing with heterogeneous servers (that have different 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 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. In addition, it would be nice to characterize (asymptotically) optimal staffing policies under $r$-routing policies, and perhaps solve the joint optimization problem of staffing and routing.

**References**
