Characterizing Performance-Based Demand Allocation Policies to Stimulate Supplier Capacities in a Time-Based Competitive Environment

Xiting Gong[1,*], Qiwen Wang[1], Nagesh N. Murthy[2], Honghui Deng[3]
1 Guanghua School of Management
Peking University, Beijing 100871, China
2 Decision Sciences Department, Lundquist College of Business
468 Lillis Business Complex, University of Oregon
3 College of Business
University of Nevada Las Vegas
4505 Maryland Parkway, Las Vegas, NV 89154-6034
[*]EMAIL: gongxiting@gsm.pku.edu.cn

Abstract: Lead time performance-based demand allocation policies have been implemented in many industries where buyers fulfill their demands from competing suppliers. Such policies can be used to stimulate suppliers to further build up their capacities so as to reduce their delivery lead times. We consider a queuing framework to characterize a more general class of stationary demand allocation polices in terms of their ability to stimulate equilibrium capacity amongst suppliers, vis-à-vis those considered in the literature. These stationary policies represent situations wherein each supplier's long-run average demand is less than its capacity when suppliers' total capacity exceeds buyers' total demand, thus resulting in finite average lead times under equilibrium. We first characterize a generic policy with symmetric and concave market share functions and show that it requires the same lower-bound supplier price as all other stationary policies studied in the literature. Next, we show that when balanced allocation policy has a pure Nash equilibrium, it continues to stimulate the highest equilibrium capacity even among a wide variety of allocation policies in this class. Finally, we introduce residual proportional allocation as a new policy and show that it contains balanced allocation as a special case. In addition, it usually stimulates the highest equilibrium capacity among all allocation policies when balanced allocation does not have a pure Nash equilibrium.

Keywords: Supplier capacity management; competing suppliers; demand allocation policy; stationary policy

I. Introduction

The benefits of reduced lead times for suppliers and service providers are quite evident in sense and respond environments. When the demand entails services or make-to-order products, fast service or short delivery lead times clearly become important. Performance-based demand allocation policies have been implemented in many industries where buyers fulfill their demands from a variety of suppliers. For example, Sun Microsystems procures memory chip from multiple suppliers and allocates demand to them based on a scorecard system [4]. Air Products and Chemicals, a global manufacturer of chemicals, gas, and equipment, develops a multi-tiered system of rating suppliers and allocates more demand towards better-performing suppliers [9]. In lead time sensitive environments a buyer can adopt a performance-based allocation policy wherein a faster supplier is awarded a greater share of the demand. Since suppliers' delivery lead times depend on their capacities, demand allocation policy across competing suppliers can be used to stimulate suppliers to build up their capacities so as to reduce the delivery lead times and consequently enhance the service level. Usually, the incentives provided by demand allocation policies differ from each other, and their performance may vary considerably.Therefore, characterizing these allocation policies provides managerial insights to help understand the behavior of suppliers and furthermore utilize them more effectively.

Based on whether or not the allocation of demand depends on suppliers' real-time workloads, allocation policies can be classified into two groups: state-dependent policies (e.g., common-queue allocation, [7]) and state-independent policies (e.g., balanced allocation, [5]). In general, the suppliers are stimulated to build their capacity by their long-run average demands, regardless of whether the underlying allocation policy is state-dependent or not. Therefore, with the same incentive, a state-dependent policy generally stimulates the suppliers to build the same equilibrium capacities as a state-independent policy but provides buyers with shorter average lead times since the latter risks allocating demand to busy suppliers while keeping the others idle. However, as concluded in Gilbert and Weng [5], when their incentives differ, a state-independent policy may stimulate higher equilibrium capacities and even provide shorter average lead times than a state-dependent policy.

In the context of a queuing model with two competing servers (or suppliers), there have been a few studies that have introduced or compared different allocation policies. Kalai et al. [7] is the first study to consider a state-dependent common-queue allocation policy for competing servers.

Christ and Avi-Itzhak [3] consider a variant of common-queue allocation to include customer balking. Cachon and Zhang [2] introduce another state-dependent policy, i.e., threshold policy, aimed at converting any state-independent policy to a corresponding threshold policy with the same incentive. There are four state-independent policies studied in the literature. Balanced allocation is introduced by Gilbert and Weng [5] and further studied by Cachon and Zhang [2]. The remaining three policies, i.e., Bell-Stidham allocation [1], linear allocation, and proportional allocation are all studied by Cachon and Zhang [2].

Gilbert and Weng [5] first compare the performance among different policies. They find balanced allocation and common-queue allocation require the same lower-bound service price to maintain finite average lead times under equilibrium, and show that the former stimulates higher equilibrium capacity. Cachon and Zhang [2] extend the policies in Gilbert and Weng [5] and compare them with more policies. They find the above lower-bound is also required by Bell-Stidham allocation and a special case of proportional allocation. In addition, balanced allocation stimulates the highest equilibrium capacity among these policies. They further point out a serious problem of balanced allocation, i.e., it may not have a pure Nash equilibrium. They also compare two other policies: linear allocation and proportional allocation, each of which has some parameters to adjust the level of competition between the servers. They show that with appropriate parameters, both policies require a lower lower-bound service price and stimulate the highest equilibrium capacity among all allocation policies.

II. Motivation/Model:

The focus of this paper is to characterize a more general class of allocation policies under which each server's long-run average demand is less than its capacity when the servers' total capacity exceeds the buyers' total demand. We refer to such policies as *stationary* policies for brevity. In many queuing systems with multiple servers, they usually converge to stationary states when servers' total capacity exceeds the system's arrival rate, i.e., total demand, and thus, each server's long-run average demand is less than its capacity. Therefore, allocation policies embedded in these systems are stationary policies. An example would be a two-server queuing system wherein each server has exponentially distributed service times and demand arrives according to a Poisson process. Upon arrival, each demand joins the shortest queue and in case both queues have equal lengths, it joins either queue with equal probability. In this system, although analytic forms of average demand functions cannot be obtained, it can be shown that it converges to a stationary state when the servers' total capacity exceeds the total demand. Then, the allocation policy "join the shortest queue" is a stationary policy.

The motivation for characterizing stationary policies originates from the observations that all stationary policies in the literature, i.e., common-queue allocation, balanced allocation, Bell-Stidham allocation, and a special case of proportional allocation, require the same lower-bound service price to maintain finite average lead times under equilibrium. Further, among these stationary policies, balanced allocation stimulates the highest equilibrium capacity when it has a pure Nash equilibrium. Linear allocation and proportional allocation in Cachon and Zhang [2] are the only two exceptions. However, both policies may allocate one server more long-run average demand than its capacity when the servers' total capacity exceeds the buyer's total demand. For example, suppose the buyer's demand equals 1 and the servers' capacities are 0.8 and 0.4, respectively; linear allocation and proportional allocation may allocate the first server 0.9 and 0.94 of long-run average demand, respectively. Therefore, neither policy is a stationary policy.

It is important for buyers to gain insights on threshold purchase price required to maintain finite average lead time and the ability to stimulate capacity for a given stationary demand allocation policy. It is equally important to explore the envelope of potential stationary demand allocation policies (beyond those considered in the literature) that would shed more light on the ability to stimulate suppliers' capacities. Thus, there is a need to consider a more general class of stationary allocation policies and compare their performance with the best stationary policy currently in the literature (i.e., balanced allocation). Lastly, it is also important to devise new policies that perform well when balanced allocation policy does not have a Nash equilibrium. These observations motivate us to ask the following research questions:

1. *Is there a generic class of stationary allocation policies than those considered in the literature that can stimulate higher equilibrium capacity for suppliers' vis-à-vis balanced allocation, when it has a Nash equilibrium?*
2. *What is the lower-bound service price required by these generic stationary policies in order to maintain finite average lead times under equilibrium?*
3. *In case balanced allocation continues to perform the best, is there any stationary policy that stimulates higher equilibrium capacity than those in the literature when balanced allocation does not have a pure Nash equilibrium?*

To answer these research questions, we study three typical and representative stationary demand allocation policies. We first study a generic stationary policy with symmetric and

concave market share functions. It contains a group of stationary policies having two common features: a) suppliers' long-run average demand functions are symmetric in their capacities; b) each supplier's long-run average demand function is concave in its capacity. In the literature, common-queue allocation and a special case of proportional allocation are its two special cases. Under this generic policy, we characterize the equilibrium structure of the suppliers' capacity game and show that it requires the same lower-bound purchase price as all other stationary policies studied in the literature to maintain finite average lead times under equilibrium.

Next, based on Gilbert and Weng [5] and Cachon and Zhang [2], we further study balanced allocation and compare it with more stationary policies. We show that when balanced allocation has a pure Nash equilibrium, it stimulates the highest equilibrium capacity among a wide variety of stationary policies, including the above generic stationary policy. We also show that it is generally hard to design stationary policies to stimulate higher equilibrium capacity than balanced allocation policy when it has a pure Nash equilibrium. In addition, we show that when servers incur strictly convex capacity cost, it is usually the high service price that prevents balanced allocation from having a pure Nash equilibrium.

Finally, we introduce a new stationary policy with a parameter to adjust the level of competition between the servers. It is referred to as residual proportional allocation since servers' long-run average demands are proportional to their residual capacities, i.e., the differences between servers' capacities and their long-run average demands. We show that it contains balanced allocation as a special case and usually stimulates the highest equilibrium capacity among all allocation policies when balanced allocation does not have a pure Nash equilibrium.

The stationary policies we study are all featured as state-independent policies. The reasons we forgo directly studying state-dependent policies are as follows. First, since state-dependent policies depend on servers' real-time workloads, we can't get analytic forms of servers' long-run average demand functions. Thus, the corresponding game theoretic analysis of competing servers becomes insurmountable. Second, as mentioned before, servers' equilibrium capacities are generally determined by their long-run average demands, regardless of whether the underlying allocation policy is state-dependent or not. Moreover, as shown by Cachon and Zhang [2], any state-independent policy can be converted to the corresponding state-dependent threshold policy with the same incentive. Therefore, studying state-independent stationary policies suffices to answer the aforementioned research questions.

The contribution of this paper is mainly of fourfold. First, it shows that the same lower-bound service price is required by many stationary policies to maintain finite average lead times under equilibrium. Second, it extends Kalai et al. [7] and Cachon and Zhang [2] to characterize a generic stationary policy with symmetric and concave market share functions. Third, based on Gilbert and Weng [5] and Cachon and Zhang [2], it shows that when balanced allocation has a pure Nash equilibrium, it stimulates the highest equilibrium capacity even among a wide variety of stationary polices. Finally, a new stationary policy, i.e., residual proportional allocation, is developed. It contains balanced allocation as a special case and usually stimulates the highest equilibrium capacity among all allocation policies when balanced allocation does not have a pure Nash equilibrium.

III. Summary and Conclusions

Understanding the role of stationary allocation policies in stimulating higher equilibrium capacity amongst competing servers is critical to obtaining fast service. There is a need to get a more comprehensive understanding by broadening the subclass of stationary allocation policies considered in the literature, gain a more in-depth understanding of the equilibrium structure across these policies, and most importantly find new policies that can substitute in situations where-in a typical best policy in the literature (e.g., balanced allocation) is not an option because of lack of equilibrium.

We first introduce and study a generic stationary allocation policy with symmetric and concave market share functions for servers (termed as generic SSC policy). Common-queue allocation policy (Kalai et al., [7] and proportional allocation policy with $\beta=1$ (Cachon and Zhang [2]) are special cases of this generic SSC policy. We show that the threshold service price (R) required by generic SSC policy to maintain finite average lead times under equilibrium is identical to those required by all other stationary policies considered in the literature, including common-queue allocation, balanced allocation, Bell-Stidham allocation, and proportional allocation with $\beta=1$. We also show that when balanced allocation has a Nash equilibrium it stimulates higher equilibrium capacity than the generic SSC policy.

Given the superior performance of balanced allocation when it has a Nash equilibrium, we subsequently focus on gaining more insights on conditions that lead to absence of a Nash equilibrium for the balanced allocation policy. Zhang [11] showed that even when the service price is above the threshold, balanced allocation does not have a Nash equilibrium in the presence of a linear capacity cost function. Cachon and Zhang [2] in addition established an upper bound on service price in order for balanced allocation to have a Nash equilibrium in the presence of a quadratic capacity cost function. We generalize this insight from Cachon and Zhang

[2] to a broader class of strictly convex capacity cost functions.

Given the effectiveness of balanced allocation policy in stimulating equilibrium capacity, it is a natural progression to devise a stationary allocation policy when a Nash equilibrium does not exist for balanced allocation. We propose residual proportional allocation (RPA) as a new stationary policy that can enable the buyer to stimulate high equilibrium capacity when balanced allocation does not have a Nash equilibrium. In this policy severs' market shares are proportional to their residual capacities. RPA has a parameter β to adjust the level of competition between servers. When β=1 RPA is identical to proportional allocation. So, proportional allocation in Cachon and Zhang [2] is a special case of our RPA policy. In particular, we first summarize properties of RPA and subsequently characterize the equilibrium structure for RPA in the presence of linear and strictly convex capacity cost.

We show that competition between the servers increases in β. Further, as $\beta \to \infty$ RPA converges to balanced allocation. So, balanced allocation is a special case of RPA. Zhang [11] showed that under linear capacity cost (i.e., c(μ)=bμ) when R > 2b, balanced allocation never has a Nash equilibrium. In contrast to balanced allocation, when capacity cost is linear and R > 2b residual proportional allocation with $\beta = 2(R-b)/(R-2b)$ has a unique Nash equilibrium and stimulates the highest equilibrium capacity among all allocation policies. Thus, when servers incur a linear capacity cost, RPA perfectly substitutes balanced allocation to stimulate highest equilibrium capacity. We also show that relative superiority of the RPA policy vis-à-vis other stationary policies increases significantly with an increase purchase price R, i.e. in environments wherein the buyer can support a higher purchase price because of improved lead time performance of suppliers resulting from increased supplier capacities.

When balanced allocation does not have a Nash equilibrium in the presence of strictly convex capacity cost (i.e., $c(\mu_b) > \frac{1}{2}R\lambda$), we show that a unique $\overline{\beta}$ satisfying $c(\mu_{r\overline{\beta}}) = \frac{1}{2}R\lambda$ exists such that RPA may have Nash equilibrium $\{\mu_{r\beta}, \mu_{r\beta}\}$ only when $\beta \le \overline{\beta}$. Since $\mu_{r\beta}$ strictly increases in β, when RPA with $\beta = \overline{\beta}$ has a Nash equilibrium, it stimulates the highest equilibrium capacity among all RPAs. Moreover, when RPA with $\beta = \overline{\beta}$ has a Nash equilibrium, since servers' equilibrium profits equal zero, it actually stimulates the highest equilibrium capacity among all allocation policies. In this case, RPA also perfectly substitutes for balanced allocation to stimulate the highest equilibrium capacity.

When RPA with $\beta = \overline{\beta}$ does not have Nash equilibrium, RPA cannot stimulate the highest equilibrium capacity among all allocation policies. In this case, suppose β^* is the largest β with which RPA has Nash equilibrium. Our results show that β^* is at least greater than one and usually greater than two. Therefore, it can be verified that RPA at least stimulates higher equilibrium capacity than common-queue allocation and proportional allocation with $\beta = 1$.

Thus, this study considers a more comprehensive set of stationary allocation policies, provides insights that further unify and supplement findings in the literature vis-à-vis threshold service price and equilibrium structure, and most importantly proposes a residual allocation policy that in several cases substitutes perfectly for balanced allocation in situations where-in the latter does not have a Nash equilibrium.

Lastly, there are several avenues for future research. In this paper we have considered suppliers who are symmetric in their cost structure. It would be interesting to study the threshold service price and the equilibrium capacities in the presence of asymmetric capacity cost structures. Additional insights can be gained by also making the purchase price endogenous in the model.

References:

[1] Bell, C., S. Stidham. 1983. Individual versus social optimization in the allocation of customers to alternative servers. *Management Sci.* **29**(7) 831-839.

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

[3] Christ, D., B. Avi-Itzhak. 2002. Strategic equilibrium for a pair of competing servers with convex cost and balking. *Management Sci.* **48**(6) 813-820.

[4] Farlow, D., G. Schmidt, A. Tsay. 1996. Supplier management at Sun Microsystems (A). Stanford Business School Case, Stanford University, Stanford, CA.

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

[6] Haight, F.A. 1958. Two queues in parallel. *Biometrika* **45**(3/4) 401-410.

[7] Kalai, E., M. I. Kamian, M. Rubinovith. 1992. Optimal service speeds in a competitive environment. *Management Sci.* **38**(8) 1154-1163.

[8] Kingman, J.F.C. 1961. Two similar queues in parallel. *Ann. Math. Statist.* **32**(4) 1314-1323.

[9] Pyke, D., E. Johnson. 2003. Sourcing strategy and supplier relationships: Alliances vs. Eprocurement. C. Billington, H. Lee, J. Neale, T. Harrison, eds. *The Practice of Supply Chain Management*. Kluwer Publishing, Boston, MA, 77-89.

[10] Wilkins, C.A. 1960. On two queues in parallel. *Biometrika* **47**(1/2) 198-199.

[11] Zhang, F. 2004. Coordination of lead times in supply chains. Dissertation, University of Pennsylvania, Philadelphia, PA.