Exploiting network effects in the provisioning of large scale systems

Jayakrishnan Nair, Adam Wierman [*]
Computing & Mathematical Sciences
California Institute of Technology
{ujk,adamw}@caltech.edu

Bert Zwart [†]
CWI
Amsterdam, The Netherlands
Bert.Zwart@cwi.nl

ABSTRACT

Online services today are characterized by a highly congestion sensitive user base, that also experiences strong positive network effects. A majority of these services are supported by advertising and are offered for free to the end user. We study the problem of optimal capacity provisioning for a profit maximizing firm operating such an online service in the asymptotic regime of a large market size. We show that network effects heavily influence the optimal capacity provisioning strategy, as well as the profit of the firm. In particular, strong positive network effects allow the firm to operate the service with fewer servers, which translates to increased profit.

1. INTRODUCTION

The internet today offers a wide range of online services, and implementing these services typically requires considerable computing infrastructure, consisting of an extremely large number of servers. Therefore, how much (computing) capacity to provision is a crucial decision for the firm operating the service. Over-provisioning enhances the user-perceived quality of the service, but is also expensive. Therefore, the service provider must strategically provision the correct number of servers to maximize its profit. The goal of this paper is to provide insight into this capacity provisioning decision.

In exploring the capacity provisioning of online systems, there are three features of the online services themselves that are of particular importance.

Firstly, a majority of these online services are offered for free to the end user, the firm (or service provider) deriving its revenue via advertising. Corporations like Google and Facebook make billions of dollars in revenue annually by offering advertising supported online services.

Secondly, many online services offered today allow for interaction between users. As a result, these services exhibit strong positive network effects, i.e., users obtain an increased utility from other people using the same service [8, 4]. Examples of such services abound: social networking applications, online gaming environments, document editing services, and many others. Indeed, network effects are a primary driver of usage growth for such services.

Thirdly, users of online services today are highly delay sensitive [6]. Large delays (due to congestion) in accessing a service can adversely affect the user perceived quality of the service, potentially leading to stagnation in usage growth for the service.

The goal of this paper is to understand how the capacity provisioning decision for online services is influenced by the interplay of the three factors discussed above. More specifically, in this paper, we consider the problem of optimal capacity provisioning for a firm operating an advertising supported online service. We model both network effects and congestion sensitivity of the user base, and analyze the number of servers the firm must provision to maximize its profit as the volume of the user base (or the market size) scales to infinity.

Our analysis reveals that as the market size becomes large, the profit maximizing strategy for the service provider involves operating the service in heavy traffic, and still having almost the full market base using the service. This is made possible by the statistical economies of scale inherent in large queueing systems: the firm can run its servers at a high utilization, and simultaneously provide good quality of service to users. Moreover, the particular heavy traffic regime that emerges depends on how strong the positive network effects are. More pronounced positive network effects imply a 'heavier' traffic regime and greater profit for the firm. This is because the firm can exploit the additional utility users derive from aggregation to operate the service at a higher level of congestion. This means the firm needs to provision fewer servers to attract the user base to the service, which translates to greater profit.

Related literature

In the queueing literature, there is a large body of work analyzing systems where the arrival rate of jobs as well as the number of servers scale to infinity. Depending on how the arrival rate and the number of servers scale relative to one another, different heavy traffic regimes are possible. One well studied scaling regime is the so-called Halfin-Whitt regime [5], in which the number of servers equals the minimum number required to stably support the arrival rate, plus a 'spare' that is proportional to the square root of the arrival rate.

In a large majority of the literature on heavy traffic many-server asymptotics, the scaling regime is assumed a priori. Some of the work in this category focuses on deriving properties of different scaling regimes; see, for instance, [5, 12, 2] and the references therein. There is also considerable work that focuses on optimal routing/scheduling in the assumed asymptotic regime. Representative papers with this theme

[*] The first two authors acknowledge the support of NSF through grant CNS 0846025 and the NetSE grant CNS 0911041, ARO through MURI grant W911NF-08-1-0233, DoE through grant DE-EE0002890 and Bell Labs.

[†] Bert Zwart is also affiliated with VU University, Eurandom and Georgia Tech. His research is sponsored by an NWO VIDI grant and an IBM faculty award.

include [1, 13, 15].

Very few papers take the contrasting approach of deriving the scaling regime that emerges naturally in the considered setting. Borst et al. [3] consider the problem of optimal staffing in a call center in an asymptotic regime where the call arrival rate is exogenously scaled to infinity. On the other hand, some papers (including this one) take the approach of scaling only the potential arrival rate to infinity. The actual arrival rate is a function of the price of the service, and/or the level of congestion. Papers in this category include [14, 9, 10, 11]. However, none of these papers consider network effects.

The key goal of this paper is to explicitly model network effects, and to understand their impact on the scaling regime that emerges, as well as the profit of the firm. In this sense, the work most related to ours is [7], which also models network effects and congestion in online services. However, the focus of [7] is on understanding the utility of the user base, not on capacity provisioning.

2. MODEL

In this section, we describe our model for the interaction between a profit maximizing firm (service provider) and a congestion sensitive user base. In our model, the firm implements the service by operating a cluster of servers, which serve user requests. We assume that there is a known market size, which determines the maximum possible usage of the service. The actual usage depends on the utility the service provides to the user base, as well as the congestion (or delay) experienced by the user base in accessing the service. The firm derives a revenue proportional to the usage of the service, which is characteristic of services that are supported by advertising, and incurs a cost proportional to the number of servers provisioned. The firm decides the number of servers to provision so as to maximize its own profit.

Formally, let k denote the number of servers provisioned by the firm. Λ denotes the maximum possible arrival rate of requests for the service, and thus characterizes the market size. User requests arrive according to a Poisson process with rate $\widehat{\lambda}_\Lambda(k) \leq \Lambda$. These requests are served in a First-Come-First-Served manner by a system with k parallel servers and a single queue. The processing times of requests are independent and exponentially distributed with mean $1/\mu$. Without loss of generality, we take $\mu = 1$. Note that $\widehat{\lambda}_\Lambda(k)$ captures the extent of 'usage' of the service by the user base.

We consider the following functional form for $\widehat{\lambda}_\Lambda(k)$:

$$\widehat{\lambda}_\Lambda(k) := \max\left\{ \arg\max_{\lambda \in [0,\Lambda]} \left[U(\lambda) - \lambda\xi(\lambda,k)\right] \right\}. \qquad (1)$$

Here, $U(\cdot)$ is the net utility derived by the user base as a function of the 'usage.' Clearly, network effects will determine the form of $U(\cdot)$. Specifically, more pronounced network effects will imply a larger value of $U(\cdot)$. $\xi(\lambda, k)$ is an indicator of the steady state congestion experienced by a typical request for service, with $\xi(\lambda, k) = \infty$ for $\lambda \geq k$, since the queueing system is unstable in this case. In this paper, for simplicity, we take $\xi(\lambda, k) = \mathbb{E}\left[W(\lambda, k)\right]$, where $W(\lambda, k)$ is the stationary waiting time experienced by a request. Our analysis technique, however, easily extends to a more general class of congestion indicators. If we interpret $\lambda\xi(\lambda, k)$ as the aggregate disutility experienced by the user base on account of congestion, (1) means the 'usage' of the service is set so as to maximize the social payoff.

Note that (1) corresponds to a cooperative, social optimization by the user base. Although the present paper focuses on this cooperative model of user behavior, it is also possible to analyze the following non-cooperative model. Interpret $V(\lambda) := \frac{U(\lambda)}{\lambda}$ to be the utility seen by a single (infinitesimal) user. Then an aggregate usage rate given by

$$\widehat{\lambda}_\Lambda(k) = \max\left\{\lambda \in [0, \Lambda] \mid V(\lambda) = \xi(\lambda, k)\right\} \qquad (2)$$

corresponds to a Wardrop equilibrium between the users with respect to their individual payoffs. We briefly compare our results for the cooperative model to those corresponding to the above non-cooperative model in Section 3.

We now turn to the behavioral model for the firm. By provisioning k servers, the firm derives revenue $b_1\widehat{\lambda}_\Lambda(k)$, and incurs cost b_2k per unit time. Without loss of generality, we set $b_2 = 1$. The profit maximizing firm naturally provisions capacity so as to maximize its profit. Specifically, the number of servers provisioned is given by

$$k^*_\Lambda := \max\left\{ \arg\max_k \left[b_1\widehat{\lambda}_\Lambda(k) - k\right] \right\},$$

and the corresponding request arrival rate is given by

$$\lambda^*_\Lambda := \widehat{\lambda}_\Lambda(k^*_\Lambda).$$

Since $\widehat{\lambda}_\Lambda(k) < k$, a necessary condition for the firm to make positive profit is $b_1 > 1$. Since the case $b_1 \leq 1$ is uninteresting (the firm will simply not operate in this case), we will assume hereafter that $b_1 > 1$.

The tuple $(\lambda^*_\Lambda, k^*_\Lambda)$ characterizes the equilibrium between the firm and the user base. We seek to understand how this tuple behaves as the market size Λ scales to ∞. In particular, we would like to discern the role played by network effects and economies of scale in the regime of large market size.

3. RESULTS

In this section, we state and interpret our results. These results, summarized in Theorem 1 below, make the following technical assumptions about the functional form of $U(\cdot)$.

Assumption 1. *$U : \mathbb{R}_+ \to \mathbb{R}_+$ is continuously differentiable over $[0, \infty)$ with $U(0) = 0$, $\lim_{\lambda\to\infty} U(\lambda) = \infty$. $U'(\cdot)$ satisfies the following properties.*

(a) There exists $\bar{\lambda} \geq 0$ such that $U'(\cdot)$ is non-decreasing over $[\bar{\lambda}, \infty)$.

(b) $\lim_{\lambda\to\infty} \frac{U'(\lambda)}{\lambda}$ exists.

(c) $\lim_{\lambda\to\infty} \frac{U'(\lambda+\nu)}{U'(\lambda)} = 1 \quad \forall\, \nu > 0$.

Condition (a) above states that $U(\lambda)$ is convex for large λ. This allows us to capture positive network effects. (b) and (c) are regularity assumptions. Note that Assumption 1 implies that

$$\alpha := \lim_{\lambda\to\infty} U'(\lambda) \in (0, \infty) \cup \{\infty\}.$$

We now state our theorem.

Theorem 1. *Suppose Assumption 1 holds. Then for large enough Λ, $\lambda^*_\Lambda \in [\Lambda - 2, \Lambda]$. As $\Lambda \uparrow \infty$, the optimal capacity provisioning is the following.*

(i) If $\alpha \in (0,\infty)$, then

$$k_\Lambda^* = \Lambda + \sqrt{\beta(\alpha)\Lambda} + o(\sqrt{\Lambda}),$$

where $\beta(\alpha) \in (0,\infty)$ is a strictly decreasing function of α.

(ii) If $\alpha = \infty$, and $\lim_{\lambda\to\infty} \frac{U'(\lambda)}{\lambda} = 0$, then

$$k_\Lambda^* = \Lambda + \sqrt{\frac{\Lambda}{U'(\Lambda)}} + o\left(\sqrt{\frac{\Lambda}{U'(\Lambda)}}\right).$$

(iii) If $\alpha = \infty$, and $\lim_{\lambda\to\infty} \frac{U'(\lambda)}{\lambda} \in (0,\infty) \cup \{\infty\}$, then

$$k_\Lambda^* = \Lambda + O(1).$$

We now highlight the key insights from Theorem 1. Firstly, it is easy see that that all α,

$$\lim_{\Lambda\to\infty} \frac{\lambda_\Lambda^*}{k_\Lambda^*} = 1.$$

This means that it is asymptotically optimal for the profit maximizing firm to operate in heavy traffic, even though the user base is congestion sensitive. This is because as the market size becomes large, the statistical economies of scale associated with large multi-server systems allow the firm to operate the service at high utilization, and still provide a good quality of service [14, 3, 9]. Moreover, the profit maximizing strategy for the firm is to provision enough capacity so as to attract (almost) the full potential market base.

Next, we observe that the heavy-traffic regime that emerges in our model, as well as the profit made by the firm, depend critically on the 'growth rate' of the social utility $U(\cdot)$. Intuitively, if the social utility is greater, the firm can attract the full potential market base by provisioning fewer servers, thereby making a higher profit. In other words, positive network effects make the user base more tolerant to congestion, allowing the firm to operate the service with fewer servers.

Case (i) of Theorem 1 corresponds to an asymptotically linear growth of $U(\cdot)$. In this case, the optimal operating regime for the firm is the well known Halfin-Whitt regime; the firm provisions the minimum capacity to serve the full market size Λ, plus a 'spare capacity' approximately proportional to $\sqrt{\Lambda}$ servers. Under Case (i), the profit of the firm is given by

$$(b_1 - 1)\Lambda - \sqrt{\beta(\alpha)\Lambda} - o(\sqrt{\Lambda}).$$

Case (ii) of Theorem 1 corresponds roughly to an asymptotically super-linear, but sub-quadratic growth of $U(\cdot)$. In this case, the optimal operating regime for the firm is a 'heavier' traffic regime than the Halfin-Whitt regime: the firm provisions a 'spare capacity' of approximately $\sqrt{\frac{\Lambda}{U'(\Lambda)}}$ servers. Under Case (ii), the profit of the firm is greater than in Case (i):

$$(b_1 - 1)\Lambda - \sqrt{\frac{\Lambda}{U'(\Lambda)}} - o\left(\sqrt{\frac{\Lambda}{U'(\Lambda)}}\right).$$

Finally, Case (iii) of Theorem 1 corresponds to a roughly quadratic/super-quadratic growth of $U(\cdot)$. In this case, the firm operates the system in a very heavy-traffic regime; it only needs to provision a bounded number of 'spare servers.' Under Case (iii), the firm makes the most profit:

$$(b_1 - 1)\Lambda - O(1).$$

To summarize, the three cases of Theorem 1 formalize the central message of this paper: positive network effects can be highly profitable to the service provider.

We conclude this paper with a brief comparison of the above results with the corresponding results for non-cooperative model for user behavior given by (2). Similar to the cooperative case, the profit maximizing capacity provisioning decision for the non-cooperative model has the following characteristics. Firstly, it is optimal to provision just enough capacity to attract (almost) the full user base. Secondly, more pronounced network effects give rise to heavier traffic regimes, and increased profit for the service provider. However, compared to the cooperative model, the non-cooperative user model of (2) leads to a higher aggregate usage, and therefore a higher level of congestion. Indeed, this is what we should expect due to the 'tragedy of the commons' effect. As a result, the non-cooperative model gives rise to even heavier traffic regimes (and higher profit for the service provider) compared to the cooperative model. For example, if $V(\cdot)$ is a constant (i.e., there are no network effects), the optimal provisioning decision for the service provider is to have only a bounded number of spare servers. In contrast, the corresponding provisioning decision for the cooperative model (covered by Case (i) of Theorem 1) is to have $\Theta(\sqrt{\Lambda})+o(\sqrt{\Lambda})$ spare servers.

4. REFERENCES

[1] M. Armony. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Syst. Theory Appl.*, 51:287–329, December 2005.

[2] R. Atar. A diffusion regime with non-degenerate slowdown. *Preprint.*

[3] S. Borst, A. Mandelbaum, and M. I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, 2004.

[4] J. Farrell and P. Klemperer. Coordination and lock-in: Competition with switching costs and network effects. *Handbook of Industrial Organization*, 3:1967–2072, 2007.

[5] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[6] J. Hamilton. The cost of latency, October 2009. URL:`http://perspectives.mvdirona.com/2009/10/31/TheCostOfLatency.asp`.

[7] R. Johari and S. Kumar. Congestible services and network effects. *Preprint.*

[8] M. Katz and C. Shapiro. Network externalities, competition, and compatibility. *The American Economic Review*, 75(3):424–440, 1985.

[9] S. Kumar and R. S. Randhawa. Exploiting market size in service systems. *Manufacturing & Service Operations Management*, 12, July 2010.

[10] C. Maglaras and A. Zeevi. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science*, 49(8), 2003.

[11] R. S. Randhawa and S. Kumar. Usage restriction and subscription services: Operational benefits with rational users. *Manufacturing & Service Operations Management*, 10(3):429–447, 2008.

[12] J. Reed. The G/GI/N queue in the Halfin-Whitt regime. *Annals of Applied Probability*, 19:2211–2269, 2009.

[13] T. Tezcan. Optimal control of distributed parallel server systems under the halfin and whitt regime. *Mathematics of Operations Research*, 33(1):51–90, 2008.

[14] W. Whitt. How multiserver queues scale with growing congestion-dependent demand. *Operations Research*, 51(4):531–542, 2003.

[15] B. Zhang and B. Zwart. Steady-state analysis for multi-server queues under size-based task assignment in the quality-driven regime. *Preprint*, 2010.