Omega 000 (2017) 1–17

Contents lists available at ScienceDirect

Omega

journal homepage: www.elsevier.com/locate/omega

Service capacity competition with peak arrivals and delay sensitive customers[☆]

Haiyan Wang[a], Tava Lennon Olsen[b,*], Guiqing Liu[c]

[a] Big Data Research Lab, Hitachi America, Ltd., Santa Clara, CA 95050, USA
[b] University of Auckland, New Zealand
[c] Hefei University of Technology, Anhui Sheng, China

ARTICLE INFO

Article history:
Received 25 November 2015
Accepted 7 June 2017
Available online xxx

Keywords:
Service operations
Capacity setting
Fluid model
Game theory
Seasonal demand

ABSTRACT

We study capacity competition in a service environment where the arrival rates are highly seasonal (e.g., lunch time rushes) and customers are time sensitive, so may depart without receiving service if the waiting time is too long. As a stepping stone for the competitive model, we begin by studying a monopolist's capacity decision, where the key trade-off is between the cost of extra capacity for low demand periods and the loss of revenue for high demand periods; we provide an attractive rule of thumb for capacity decisions in this setting. We then study a duopoly model, where lost demand for one firm may become increased demand for the competitor. In both models we use a fluid model for the analysis, which allows us both to provide explicit insights into the trade-offs when setting capacity and to solve for the Nash equilibrium (when it exists) in the duopoly. The canonical environment we have in mind for our modeling is a food court, but any service environment where the peak arrival rate will likely exceed available capacity is similarly appropriate.

1. Introduction

In many service environments arrival rates are highly seasonal (e.g., lunch time rushes) but capacity cannot be finely tuned (e.g., by the half hour) to the arrival pattern due to the need to install fixed capacity and the need for human capacity to be hired for some minimum shift length (a notable exception to this statement would be large call centers, which are not considered here). This often leads to an under-supply of capacity at peak arrival times, which in turn leads to long delays and significant customer balking, particularly if the service is not critical to the customer or if there are other competing providers of the service. Because we consider service environments, there is little possibility of inventorying for this peak period.

Our primary interest is the study of capacity competition in the presence of "excess" customers caused by non-stationary arrival patterns. Does competition make the overflow situation better or worse for the customer? How should firms set capacity in the presence of competition? In order to answer these questions we first provide a simple model for capacity setting for a single monopolist serving impatient customers who may balk. Although possibly too crude a model to use to design a service system, it does provide some interesting insights into the relevant trade-offs that need to be considered as well as providing an attractive rule of thumb for making such decisions. This model is then used as a building block for the duopoly setting.

We consider a non-stationary arrival pattern that rises to a peak and then declines. This type of arrival pattern is quite common in a variety of service environments. For example, around rush hour in the morning, a toll booth may face such a pattern of traffic (see, e.g., [20]. Similar patterns occur at banks (say, during the lunch break period) or post offices. In this latter case, [45] claim that the rise-peak-fall-off pattern of the average arrival rates for letters within a few hours is the main contributor to processing delays in a post office. The specific arrival pattern we study is one where the customer arrival rate grows linearly and then dissipates linearly during a certain period. As stated in [40], based on traffic studies, "There is evidence to support this assumption of triangular- or trapezoidal-shaped demand patterns"; it is noted that the trapezoid shape may come from blocking. Another typical instance for our arrival pattern is food court, which is the canonical application we use for our models. During the lunch period, the customer arrival rate will increase to a peak, somewhere around the noon hour, and then decrease (see [39], for a demand profile for a typical fast-food restaurant during the lunch period).

[☆] This manuscript was processed by Associate Editor Q. Wang.
[*] Corresponding author.
E-mail addresses: haiyan.wang@hal.hitachi.com (H. Wang), t.olsen@auckland.ac.nz (T.L. Olsen), 769450363@qq.com (G. Liu).

http://dx.doi.org/10.1016/j.omega.2017.06.001
0305-0483/© 2017 Elsevier Ltd. All rights reserved.

There is a stream of literature that studies performance analysis of non-stationary arrivals using stationary models (see, e.g., [25], and references there-in). We do not take this route. Instead, we explicitly model the non-stationary pattern of arrivals. Our focus is on explicit approximations, generation of insights, and a model that is useful for duopoly competition. Therefore, we use a relatively crude fluid model where all flows are deterministic, as are the customer balking and switching decisions. Customers will leave if their predicted delay exceeds their tolerance for wait (because the fluid system is deterministic, delays may be perfectly predicted). Although there are more accurate models of transience than this (e.g., [21,22,34,42,43,56], they will not yield the explicit expressions that we require for our duopoly model; even under our simple assumptions the expressions become surprisingly complex.

The use of fluid models for overloaded systems has been motivated by Newell Chapter 2 [44], Hall Chapter 6 [27], Kleinrock Section 2.7 [33], and [38,54–56]. The basic idea is that as the system scale becomes large the functional strong law of large numbers will apply and stochastic processes will start to look deterministic. Here, we provide no formal limit theorems, instead we begin with the assumption of a fluid model and proceed with the analysis from there. Note that the standardization of service procedures in fast food restaurants may make the assumption of deterministic processing times particularly applicable for our food court example.

There has been significant work on staffing models for non-stationary arrival patterns (see [15,23,26,37,47,57] for relatively recent reviews). Much of this work specifically considers call-center staffing. Due to this application, most of this work assumes that staffing can also be adjusted (usually in intervals, e.g., every half hour) as the demand requirements change, perhaps with some requirements on shifts (e.g., [5]). However, sometimes the expense of changing capacity frequently may be very high (see, e.g., [30], for the expenses associated with adjusting work force and other aspects of production capacity and [19], for a discussion on managing overtime for full time employees).

When capacity is fixed during the service period, the service provider must trade off the potential revenue lost during the peak demand period with the cost of idle capacity during the low demand period. We consider fixed capacity here. In the food court example, the lunch time period is relatively short (probably two to four hours). It often does not make sense to change the capacity (say, staff numbers) during the shift because the time window may be shorter than a work shift. Our conversations with a particular food court provider indicate that indeed his staffing is typically for the entire lunch time block, rather than finely tuned to the arrival rate. Further, anecdotal experience suggests that long waits at food courts are typical at lunch time and significant balking and switching to one's second choice provider are common experiences, indicating that under-capacity is common during peak arrival times.

Our single firm model generates insights into how capacity should depend on basic system primitives and, more importantly, is used as a building block for our duopoly model. It provides a back-of-the-envelope solution that may be used if in fact capacity choices are limited and discrete (e.g., hire one or two servers?). In only modeling prime-time staffing, we are, in effect, assuming that there is some baseline level of capacity that is present for non-prime time (which is not modeled). This baseline capacity (say a single server) will also moderate the underlying variability during the prime time. However, for a more fine tuned answer to how capacity should be set we would recommend simulation, where a search over a single parameter (i.e., capacity) is very feasible and many more subtleties of customer behavior (e.g., stochastic arrivals, abandonment, etc.) may be included. This is the approach taken by Martinich et al. [39], who uses simulation to study when to add additional capacity in a fast food restaurant in order to deal with the peak demand. The simulation results demonstrate that scheduling additional servers a little earlier can have a dramatic impact on customer waiting times for an extended period in a non-stationary queueing system.

We find that there are two key factors in determining the server's optimal capacity level, namely, customer patience and the marginal cost to revenue ratio. Unsurprisingly, when facing impatient customers and relatively high cost to revenue ratio, we find it may be optimal for the server to forgo some customers during the peak demand period. We quantify the regions where this is the case. We show that while the optimal capacity is monotonically decreasing in the cost to revenue ratio (i.e., higher costs or lower revenue per customer lead to lower capacity being optimal), it is not always monotone in customer patience. Instead, it is unimodal where the highest optimal capacity is at “moderate” levels of patience. In other words, while high levels of customer patience will lead to less capacity being needed, which is intuitive, very low levels of patience may also lower capacity. The intuition behind this result is given in Section 4.

In our duopoly setting, customers are assumed to have an initial server preference (e.g., pizza over Chinese food) but may switch from their original choice to their second choice if the waiting time at their favorite server is too long. Therefore, the two servers compete with each other based on waiting time by strategically choosing an appropriate capacity level. We do not consider pricing competition. Instead, we assume the firms are price-takers or the prices are set by some central organization (e.g., the franchise parent company) without consideration of local competition.

We first analyze customers' switching and balking behavior and the impact on servers' profits. Because of the non-stationarity of the customer arrival process, the customers' switching and balking pattern is quite complicated. This poses a significant challenge in characterizing the Nash equilibrium in the servers' capacity competition game because it becomes intractable to explicitly express the servers' profit functions for a given pair of capacity choices in a general setting. Therefore, we develop sets of sufficient conditions that sustain the existence of a unique Nash equilibrium in the symmetric game. Intuitively speaking, as long as customers have a “sufficient” preference for their original choice, a symmetric Nash equilibrium exists at which each server works as if he were a monopolist, that is, no switching happens in the equilibrium. However, when customers do not have sufficient preference (meaning that customers have similar valuations on two services), there often does not exist a pure-strategy Nash equilibrium. The rationale is similar to that for the model of price competition over identical customers with observable queues. We study asymmetric competition numerically and find situations with no pure-strategy Nash equilibria, situations with multiple pure-strategy Nash equilibria, and situations with a unique pure-strategy Nash equilibrium; intuition behind these results is given.

As outlined above, our contributions are fourfold. First, we provide a back-of-the-envelope approach for capacity setting under seasonal demand. Second, we use that approximation to provide insights into the trade-offs to be considered when setting capacity. Third, we extend the (relatively limited, see Section 2) literature on competition with observable queues to a model with capacity competition. Finally, we provide what we believe to be the first results on capacity competition with non-stationary arrivals.

The rest of the paper is organized as follows. Related literature is surveyed in Section 2. In Section 3, we outline our assumptions regarding customers, servers, costs, and revenues, and provide some initial results. Section 4 considers the monopolist's capacity decision, providing both concrete guidance on capacity setting and a building block for the duopoly competition in Section 5. Section 6 concludes the paper. All proofs may be found in the Appendix.