European Journal of Operational Research 252 (2016) 477–486

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Customer equilibrium and optimal strategies in an M/M/1 queue with dynamic service control

Y. Dimitrakopoulos[a], A. N. Burnetas[b,*]

[a] Department of Informatics, Athens University of Economics and Business, 76 Patission St., Athens, Greece
[b] Department of Mathematics, National and Kapodistrian University of Athens, Panepistemiopolis, Athens, Greece

ARTICLE INFO

Article history:
Received 12 May 2015
Accepted 15 December 2015
Available online 23 December 2015

Keywords:
Queueing
Customer equilibrium strategies
Dynamic service control
Positive externalities
Price of anarchy

ABSTRACT

We consider the problem of customer equilibrium strategies in an M/M/1 queue under dynamic service control. The service rate switches between a low and a high value depending on system congestion. Arriving customers do not observe the system state at the moment of arrival. We show that due to service rate variation, the customer equilibrium strategy is not generally unique, and derive an upper bound on the number of possible equilibria. For the problem of social welfare optimization, we numerically analyze the relationship between the optimal and equilibrium arrival rates as a function of various parameter values, and assess the level of inefficiency via the price of anarchy measure. We finally derive analytic solutions for the special case where the service rate switch occurs when the queue ceases to be empty.

1. Introduction

Service rate control is a prevalent tool for dynamically adjusting the operation of a queueing system in response to varying congestion levels. Many models in the queueing and service management literature, such as dynamic staffing policies, server vacation mechanisms, etc. can be viewed as different aspects of varying the rate at which service is provided at different times. In most cases, the philosophy of a service control policy is to increase the service capacity when the congestion level is high in order to alleviate excessive delays, and reduce it at times of decreased congestion in order to lower operating costs. Viewed in this way, a service control policy provides a balance in the tradeoff between long customer delays and high service costs. However, in situations where customers are sensitive to delay and make individual decisions about entering the queue, the policy used by the service provider may also have an indirect, albeit significant, effect on the arrival stream.

In service systems with strategic customer behavior, incoming customers decide independently whether to join the system or balk, based on an individual utility function that combines the value of obtaining the service with the cost induced by the anticipated delay in the queue. Since a customer's delay depends on the decisions of other customers, a game-theoretic approach is appropriate for characterizing customer behavior and thus the properties of the incoming stream, as the result of an equilibrium strategy. Because customers are adverse to delay, it is expected that a particular customer's decision to enter the system is adversely affected as other customers also decide to enter and system congestion is increased. Such customer behavior, where in equilibrium his/her willingness to join is a decreasing function of the arrival rate, are referred to as Avoid the Crowd (ATC). On the other hand, in a system where a form of dynamic service control is implemented, the service rate is usually increased when the system is congested and it may result in Follow the Crowd (FTC) behavior, where in equilibrium arriving customers are more encouraged to join as the arrival rate increases. These concepts are discussed in Hassin and Haviv (1997), where the authors analyze customers' optimal choice between two priority levels, in order to receive service. It is shown that multiple equilibria exist, which is typical in situations where FTC behavior prevails. Both the ATC or FTC effects may appear in the same system for different values of the arrival rate. In most situations, the presence of FTC is related to a dynamic service policy employed by the service provider.

The present paper proposes a direct model of the impact of dynamic service control on customer equilibrium behavior. More specifically, we analyze the customer equilibrium behavior for joining a single server Markovian queue under a simple threshold-based service rate policy. The service rate is kept at a low value when the number of customers in the system is at or below a

* Corresponding author. Tel.: +30 210 727 6437.
E-mail addresses: dimigiannhs@aueb.gr (Y. Dimitrakopoulos), aburnetas@math.uoa.gr, aburnetas@gmail.com (A.N. Burnetas).

http://dx.doi.org/10.1016/j.ejor.2015.12.029
0377-2217/

threshold level *T*, and turns to a high value when the system congestion is above *T*. We analyze the unobservable case where arriving customers are aware of the service policy, but do not have any information on the queue length upon arrival.

The most pronounced effect of the service control policy is that due to a mixed ATC/FTC pattern the equilibrium customer strategy is not generally unique, as the expected sojourn time of a tagged customer is not monotone increasing in the arrival rate, in contrast to the simple M/M/1 queue. For example, one equilibrium strategy may be such that a small fraction of incoming customers joins the system and as a result the server works at slow rate most of the time, whereas there exists another equilibrium where the arrival rate is high and the server spends most time at the fast mode. We show that, depending on the values of the various parameters, the number of strategic equilibria may vary from one to at most three.

Moreover, we consider the problem of determining the optimal arrival rate in a system with centrally controlled arrivals which, for the given service policy, maximizes the expected net benefit of all customers, and compare the optimal arrival rates to those obtained as equilibria under selfish customer behavior. We show via computational experiments that the optimal arrival rate lies between the extreme values under equilibrium. This is related to the existence of both negative and positive externalities. In economic terms, entering customers generally induce negative externalities, by imposing increased delays to present and/or future arrivals. However, in our model it may happen that if a customer decides to enter the system, the other customers experience a benefit, because the service provider is induced to increase the service rate. Then, the net effect between the additional delay due to increased arrivals and the higher service speed due to the server's policy may be positive. In such cases, an arriving customer induces positive externalities.

We finally consider the issue of the price of anarchy and the price of stability, which are measures of the inefficiency of the customer strategies under equilibrium compared to optimal strategies that maximize the social benefit. We show that in our framework these measures typically coincide. Furthermore, a prevailing insight from the computational experiments is that the inefficiency is significant under individual customer behavior and in those cases where the price of anarchy approaches the ideal value 1, this happens because most customers are directed to the alternative service and the proportion of those who enter the primal system approaches zero.

The implications of strategic customer behavior on the performance of a queueing system has been studied extensively in the recent years. Early works on the M/M/1 queue include Naor (1969) and Edelson and Hildebrand (1975) for the observable and unobservable models, respectively. Many variations of the original models have been studied since, and a comprehensive review of the literature until 2003 is provided in Hassin and Haviv (2003). Among the models that have been developed and analyzed, many include some varying service rate characteristics indirectly, such as vacation queues and systems in a random environment. Burnetas and Economou (2007) consider a system with server vacations and reactivation that requires a random setup time. In the vacation model of Guo and Hassin (2011), the server resumes service after a fixed number of arrivals according to a threshold service policy. Both papers analyze equilibrium behavior with respect to several levels of information about the queue length and the state of the server and involve a mixed ATC/FTC behavior. Economou and Manou (2013) analyze customer equilibrium strategies in a stochastic clearing system where the exogenous customer arrival rate and the arrival rate of the system clearing server vary according to an external environment process.

In some systems with varying service capacity, especially in transportation and telecommunication networks, counterintuitive cases have been identified regarding the behavior of the system under equilibrium. In these situations, increasing the capacity of a network link (Downs-Thomson paradox) or adding a new link (Braess paradox) may result in higher delays, since selfish choice of routes by individuals eliminates the benefit of the additional resource (cf. Afimeimounga, Solomon, & Ziedins, 2005; 2010; Braess, 1968; Calvert, 1997; Downs, 1962; Thomson, 1977).

The properties of the price of anarchy and the price of stability as the ratio between optimal and equilibrium welfares have been examined in several frameworks, most notably in problems of routing in networks. For a definition of the two measures and further discussion of their differences see Nisan, Roughgarden, Tardos, and Vazirani (2007). In queueing systems with strategic customer behavior, Haviv and Roughgarden (2007) consider a multi-server system where customers select which queue to join, with no balking option and without observing queue lengths. It is shown that the *PoA* is upper bounded by the number of servers. Gilboa-Freedman, Hassin, and Kerner (2013) consider the observable M/M/1 queue, derive bounds on the *PoA* when the arrival rate is low and show that it becomes unbounded when the potential arrival rate exceeds the service rate.

In the area of queueing control and optimization under a single decision maker, problems of service rate control have been studied extensively and in many forms. It is worth noting that the problem of social welfare maximization can be formulated as a static or dynamic admission control problem. George and Harrison (2001) consider welfare optimization in an M/M/1 system with dynamic service control and constant arrival rate. They develop an asymptotic method for computing the optimal policy under average cost minimization. Ata and Shneorson (2006) propose a model that jointly optimizes arrival and service rates, and develop a dynamic pricing strategy that induces the optimal arrival and service rates. Adusumilli and Hasenbein (2010) consider a model of joint admission and service control in an M/M/1 under an average cost criterion and show that the optimal service rates are increasing with system congestion. Dimitrakopoulos and Burnetas (2015) analyze the combined problem of dynamic control of the service rate and admission control policy in a similar model with no strategic customer behavior, and derive conditions for the service switch option to be beneficial. In Tirdad, Grassmann, and Tavakoli (2016) the service rate is controlled by altering the number of servers in a model with nonstationary arrival rates.

The rest of the paper is organized as follows. In Section 2 we introduce the model and the corresponding customer strategic behavior problem, as well as the social welfare optimization problem. In Section 3 we consider the equilibrium strategies and derive the upper bound on the number of equilibria for general *T*-threshold dynamic service policies. In Section 4 we explore numerically the socially optimal strategy and the comparison with the equilibrium problem. Since in general optimal and equilibrium arrival rates do not coincide we examine this inefficiency introducing the price of anarchy. In Section 5, we derive analytic solutions for equilibria in the special case where the service threshold $T = 1$, i.e., the service rate switches to the fast mode when the queue is not empty. Section 6 concludes.

2. Model description

We consider a single server Markovian queue under the FCFS discipline, where potential customers arrive according to a Poisson process with rate Λ. The service rate varies dynamically according to a *T*-threshold policy. Specifically, the service rate is set to μ_l whenever the number of customers in the system is less than or equal to *T*, and equal to μ_h otherwise, where $\mu_l < \mu_h$. There are no switching costs. Thus, the service policy is defined by the exogenous parameters (T, μ_l, μ_h), which are known to all arriving customers.