East Asian Journal on Applied Mathematics
doi: 10.4208/eajam.231010.260711a

Vol. **1**, No. 4, pp. 379-402
November 2011

Incentive Effects of Multiple-Server Queueing Networks: The Principal-Agent Perspective

Sin-Man Choi[1,*], Ximin Huang[2], Wai-Ki Ching[3] and Min Huang[4]

[1] *Department of Industrial Engineering and Operations Research, University of California, Berkeley, US.*
[2] *College of Management, Georgia Institute of Technology, Atlanta, US.*
[3] *Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong.*
[4] *College of Information Science and Engineering, Northeastern University; State Key Laboratory of Synthetical Automation for Process Industries, (Northeastern University), Shenyang, Liaoning, 110819, China.*

Received 23 October 2010; Accepted (in revised version) 26 July 2011

Available online 23 September 2011

Abstract. A two-server service network has been studied from the principal-agent perspective. In the model, services are rendered by two independent facilities coordinated by an agency, which seeks to devise a strategy to suitably allocate customers to the facilities and to simultaneously determine compensation levels. Two possible allocation schemes were compared — viz. the common queue and separate queue schemes. The separate queue allocation scheme was shown to give more competition incentives to the independent facilities and to also induce higher service capacity. In this paper, we investigate the general case of a multiple-server queueing model, and again find that the separate queue allocation scheme creates more competition incentives for servers and induces higher service capacities. In particular, if there are no severe diseconomies associated with increasing service capacity, it gives a lower expected sojourn time in equilibrium when the compensation level is sufficiently high.

AMS subject classifications: 60K25, 68M20, 91A80

Key words: Capacity allocation, competition, incentive theory, Markovian queueing systems, Nash equilibrium, principal agent.

1. Introduction

Finding the optimal strategy and control policy for a queueing system is a traditional optimal control problem that is well studied in the literature — e.g. see [2, 11–14, 19].

*Corresponding author. *Email addresses:* kelly.smchoi@berkeley.edu (S.-M. Choi), hehe1121@gmail.com (X. Huang), wching@hku.hk (W.-K. Ching), mhuang@mail.neu.edu.cn (M. Huang)

An optimal control problem usually involves making decisions on such system parameters as service capacity, the waiting time or the sojourn time spent in the system, and the number of servers in the system under a specified cost structure (convex or concave). Service capacity is often an important competitive factor in system design — e.g. in telecommunication networks [6], data transmission systems [13], or Vendor-Managed Inventory (VMI) [3, 18] and other supply chain management [10]. In particular, current developments in supply chain management emphasize the coordination and integration of inventory and transportation logistics [4, 20]. VMI is a supply chain initiative where the distributor is responsible for all decisions regarding the selection of the retailers or agents, which creates a competitive environment for them in the market [16].

Kalai et al. [13] studied the service capacities of two servers competing for market share, assuming a Markovian queueing system. Markovian queueing systems are popular tools for modeling service systems, since they are more mathematically tractable than non-Markovian queueing systems [6, 7]. Game theory [17] is a popular and promising analytical approach [1, 5, 8, 10]. Kalai et al. [13] classified the relevant Nash equilibria into three different cases concerning the cost function and revenue per customer, with a finite waiting time and a unique symmetric equilibrium in one case. Although their model is simple, it included two important concepts. The first is the "competitive game of servers", and the second is "market share of a server in a multi-server facility". Furthermore, when the marginal cost of providing service is "high", they found there is a unique symmetric equilibrium and that the total service capacity is less than the mean demand rate. In such a case, each server actually behaves as if it were a monopolist, so there is no desirable competition. On the other hand, when the marginal cost of servicing is "low", a unique symmetric equilibrium exists and the total service capacity is greater than the mean demand rate.

In [14], a service network where a coordinating agency is responsible for satisfying the customers' total waiting and service time is studied. Two facilities (two servers) are considered, and two types of allocation policy — viz. a common queue with two servers, and two separate single-server queues. In some cases, the separate queue allocation scheme was found to have advantages over the common queue allocation scheme. In this paper, we extend the model in [13] to allow more than two servers, and are particularly interested in where the total service capacity exceeds the mean demand rate. Our analysis indicates that with multiple servers the separate queue allocation scheme gives more service incentives and induces higher service capacities. Moreover, when there are no severe diseconomies associated with increasing service capacity, the separate queue allocation scheme gives a lower expected sojourn time in equilibrium.

The remainder of this paper is structured as follows. In Section 2, we briefly review the two-server queueing system in [13] and the service system in [14]. The multiple-server common-queue model and our analysis of system performance is presented in Section 3. In Section 4, we discuss the multiple-server separate queue system and