Int. J. Production Economics 157 (2014) 302–312

Contents lists available at ScienceDirect

Int. J. Production Economics

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

Joint inventory and pricing decisions when customers are delay sensitive

M. Güray Güler, Taner Bilgiç, Refik Güllü *

Bogazici University, Department of Industrial Engineering, 34342 Bebek, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 26 September 2012
Accepted 29 April 2014
Available online 14 May 2014

Keywords:
Inventory
Pricing
Priority
Queueing systems
Incentive compatibility

ABSTRACT

We consider the joint pricing and inventory problem of a capacity constrained service facility with several classes of customers. The customers are differentiated with their sensitivity for waiting and their willingness to pay for the service. We model the problem using an $M/M/1$ queueing system with non-preemptive priorities. We give closed form solutions for the inventory decisions. We also show that the prices given by the first order conditions are also incentive compatible in the sense that they optimize the profit of the firm even if the firm does not know the type of an arriving customer and let the customer choose a price from the menu of prices. We approximate the problem and provide simple and explicit solutions when there is a single customer type. In numerical illustrations, we show that the customers, who are more sensitive to wait, do not enter the system until the base stock level is above a threshold. We provide extensions of our results for $M/G/1$ and $M/M/m$ systems.

1. Introduction

Companies employ various strategies in order to extract more revenue. One commonly observed strategy is providing the same service at different levels of speeds and at a different price. Many make-to-order manufacturers charge their customers based on the delivery dates, and logistics firms offer a range of services from ground shipping to same day delivery at different prices. By offering the option to pay more for faster services, the firms can extract more revenue from market segmentation (Talluri and VanRyzin, 2005).

Some companies own several brands and sell similar products at different prices. The automobile manufacturer Volkswagen group has several brands such as Audi, Volkswagen, Skoda and Seat. All brands have similar categories of products and within each category, different brands of automobiles share many common components. For example Audi A4 and VW Passat use the same engine, transmission and special features (like 4-wheel drive: quattro for Audi and 4Motion for VW). Using common components in products not only improves the design and manufacturing of the product but also provides advantage in after-sales service. After-sales service has become a major part of competition in the automotive industry. There is a global network of after-sales service facilities where all brands of Volkswagen group can get service. These facilities need to hold significant amounts of spare parts inventory to provide timely service. The owners of automobiles that require service may have different characteristics. For example an Audi owner and a Seat owner may have different budgets and valuations for their cars and for the after-sales service of their cars. The problem of the after-sales service facility is to decide how much inventory to hold and how to price the service for different customers.

Motivated by the automobile after-sales service problem we consider joint pricing and inventory decisions in a capacity constrained service facility (SF) with several customer classes. An SF that serves different brands of the same auto manufacturer while keeping a common spare parts stock for a (single) component that can be used as a replacement in any one of the brands is an example. Each customer type arrives for the replacement of the particular component. Upon arrival of a customer, if the item is available in inventory then the customer does not wait (we neglect the replacement time). Whereas, if the item is not available in inventory, the customer waits until the item is delivered through a capacitated supply system. Customer types are differentiated using priorities and the waiting times of the customers for delivery depend on the priority of the customer type. For instance, in line with our motivating example, a customer who owns a luxury brand may have a higher priority over a customer who owns an economy car. The inventory side of the SF operates under a base stock policy and decides on the price to be charged for the component for each customer type and the base stock level for

* Corresponding author.
E-mail addresses: guler@boun.edu.tr (M.G. Güler), taner@boun.edu.tr (T. Bilgiç), refik.gullu@boun.edu.tr (R. Güllü).

http://dx.doi.org/10.1016/j.ijpe.2014.04.028
0925-5273/

the spare parts inventory. Using a base stock policy for expensive components like, for example, the motor and the transmission system is common practice in the automotive industry. Customers, on the other hand, decide on joining the system or not based on their reservation price, waiting cost, and the price charged for the component. The objective of the SF is to maximize the average expected profit of the system by choosing the best set of prices and the best base stock level.

Much of the initial work on pricing problems with multiple classes of customers has been done in the context of a queueing network, where heterogeneous customers arrive with different sensitivities to delay, and pricing determines the allocation of the priorities. Our work is closely related to the priority pricing problem for the $M/M/1$ queueing system. In this problem, there is an SF like a communication network or a production line which is modeled as an $M/M/1$ queueing system. The users of the service, or customers, have different levels of *urgencies* and hence different types of customers have different sensitivities for waiting. The SF announces different levels of priorities to be chosen by the customers according to their urgency levels. The priorities can be preemptive or non-preemptive. In a queueing system with preemptive priority, a high priority job preempts a low priority job in the service. In a system with non-preemptive priorities, the service cannot be preempted and the job with the high priority waits for the service to be completed. Since every customer prefers a high priority, the aim is to construct priority schedules and pricing of these priorities to allocate service capacity and to determine the priorities optimally. Such a *priority pricing* scheme should be incentive compatible in order to make customers choose the right price designed for their type. Therefore these prices are not only optimal from the perspective of the customers, but also optimal for the entire system; the SF and the customers.

The rest of the paper is organized as follows. The literature review is given in Section 2. Section 3 introduces the model and presents some preliminaries. Section 4 provides solution and analysis of the model. We provide a continuous approximation of the model in Section 5. Numerical illustrations are presented in Section 6. We provide extensions of our main results to $M/M/m$ and $M/G/1$ type of queueing systems in Section 7 and present our conclusion in Section 8.

2. Literature review

For a system of homogeneous customers (i.e., when there is a single customer type) Mendelson (1985) shows that a price which captures the externality cost of congestion is necessary to induce users to submit jobs at a rate that is socially optimal. The *externality cost* is the expected delay cost per unit time, inflicted on the rest of the system by an infinitesimal unit increase in the job flow. Mendelson and Whang (1990) are among the first to study incentive compatible priority pricing strategies in queueing systems with multiple user classes. They analyze a system that is modeled as an $M/M/1$ queueing system with several user classes which differ in their valuation of the service and their waiting per unit time costs. When service requirements are homogeneous, they show that the optimal prices which maximize the net value of the system are also incentive compatible. For the non-homogeneous service requirement case, they show that these prices fail to be incentive compatible and provide an optimal pricing scheme which depends on the priorities and the service times which are shown to be incentive compatible. Mandjes (2004) analyzes a communication network where there are two types of customers, delay sensitive or delay tolerant. He assumes that customers incur the waiting cost in a non-linear way rather than the usually assumed per unit time costing. He explicitly provides the optimal prices.

The objective of the latter three papers is to optimize the net value of the whole system, which is also called the *social optimization*. The socially optimal mechanism is of interest for an operation that serves customers within an organization. However, the objective of a company which serves external customers is usually to maximize its own profit. Katta and Sethuraman (2005) study the problem of a profit maximizing firm in the context of an $M/M/1$ queueing system. The aim is to find the optimal pricing and scheduling rule. They are able to characterize the optimal policy by making some simplifying assumptions and find that pooling the customer types into a single priority class is optimal. Afèche (2004) analyzes the same problem with two types of customers. The aim is to find the optimal pricing, scheduling and also admission control, i.e., when to accept or reject the customers. They show that the celebrated $c\mu$ rule need not be optimal for setting the priorities among the customers. This rule states that the customer types should be given priorities according to their $c\mu$ values, where c is the waiting cost per unit time and μ is the expected service time needed. They show that the optimal policy may insert an idleness to the server even if there is a customer in the system. This insertion of strategic delay deters time sensitive customers from purchasing the low-priority class. Cui et al. (2009) analyze a similar problem. They use an $M/M/1$ queueing system to model a profit maximizing firm that serves customers whose waiting cost per unit time and valuation of the service can take two distinct values, low or high. In their study, the server also decides on a probabilistic admission control (i.e., there is a probability of rejecting customers). They show that a probabilistic admission may force customers to reveal their true valuations and induce the delay-sensitive customers to choose high priorities and hence enable the server to receive more revenue from the customers.

Afèche and Mendelson (2004) propose a generalized delay cost structure and capture the interdependence between delay cost and value of the service of homogeneous customers. They show that the priority auction mechanisms perform better than setting a single price under the social optimization and revenue maximization. They also show that the structure of the delay cost has a paramount effect on the system behavior.

Gilland and Warsing (2009) analyze a model where the customers have heterogeneous delay costs, picked up from a uniform distribution. The firm offers a non-preemptive prioritization for a premium. They play a Stackelberg game where the firm is the leader and decides for the premium. The customers are the followers. The total arrival rate is known and fixed. The customers decide the high priority arrival rate. They analyze the model according to the behavior of the customers. *Coordinated customers* act to minimize the total costs per unit time across all customers. *Independent customers* decide based only on the economics of that job. They show that coordinated customers submit high-priority jobs at a rate which is below the rate which minimizes total delay cost. On the other hand, independent customers choose a submission rate for high priority jobs which minimize the total delay costs.

Hassin (2007) studies an $M/M/1$ system with a single customer type where the service rate, the value of the service and the waiting cost per unit time are random variables that can take two values and are known only to the service provider. He shows that the server is motivated to reveal the information except in the case where the value of the service is random and the waiting cost is small.

The priority pricing problem for the $M/M/1$ queueing system, both under social optimization and profit maximization, is designed for service systems or production systems in which the inventory decision is not incorporated. An extensive summary and classification of queueing systems can be found in Hassin and Haviv (2003).