Computers & Industrial Engineering 78 (2014) 26–32

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

Individually and socially optimal joining rules for an egalitarian processor-sharing queue under different information scenarios ☆,☆☆

Miaomiao Yu [a,b,*], Yinghui Tang [c], Wenqing Wu [c]

[a] School of Science, Sichuan University of Science and Engineering, 643000 Zigong, Sichuan, China
[b] Department of Electrical & Computer Engineering, University of Manitoba, R3T 5V6 Winnipeg, Manitoba, Canada
[c] School of Mathematics & Software Science, Sichuan Normal University, 610068 Chengdu, Sichuan, China

ARTICLE INFO

Article history:
Received 11 February 2014
Received in revised form 18 July 2014
Accepted 24 September 2014
Available online 6 October 2014

Keywords:
Queueing
Processor-sharing
Joining rules
Conditional sojourn time

ABSTRACT

This paper studies joining behavior of customers into an $M/M/1$ egalitarian processor-sharing (PS) queue. By constructing a left-multiplication transformation and using its matrix representation, we obtain the expected conditional sojourn time of a tagged customer. Then, in the fully observable case, we first consider the joining strategy in a decentralized manner, that is, arriving customers observe the queue size and then decide whether or not to join the queue based on the net benefit they will obtain upon the completion of service. Secondly, we derive the threshold strategy that will yield the system's maximal expected profit, to reach the so-called social welfare optimization. Finally, Nash equilibrium and socially optimal mixed strategies are derived in the fully unobservable case. Moreover, some numerical examples are provided to explore the impact of system parameters on customer's joining behavior.

1. Introduction

A packet switched network is one of the most commonly used computer networks. Most of the Internet services are supplied through such networks. The main advantage of packet switching is the fact that it permits "statistical multiplexing" on the communication lines. This also means that packets from a variety of sources can share one line at the same time. Thus, over the past few decades, PS queue was considered to be the most powerful tool in modeling how the bandwidth of a transmission link is shared among elastic traffic flows (see Altman, Avrachenkov, & Ayesta (2006)), and numerous theoretical analysis on PS queues were carried out by many researchers. In the sixties of the last century, Kleinrock (1967) introduced the simplest and best known egalitarian processor-sharing (EPS) discipline. The key property of the PS service discipline is that the common resource is fairly shared among all jobs present in the system. After the pioneering work of Kleinrock, Coffman, Muntz, and Trotter (1970) first investigated the Laplace–Stieltjes transform (LST) of the sojourn time distribution in the M/M/1 PS queue. Later, Yashkov (1983), Schassberger (1984) and Ott (1984) independently derived the LST of the sojourn time distribution in the M/G/1 PS queue. Based on their work, Morrison (1985) inverted the LST and obtained an integral expression of the sojourn time probability distribution. For the GI/M/1 PS queue, the first two moments and the LST of the unconditional sojourn time were derived by Ramaswami (1984). In addition, using the matrix-analytic method, Núñez-Queija (2001), Masuyama and Takine (2003), and Li, Lian, and Liu (2005) also discussed the block-structured PS queues with Markovian arrival process. Some effective algorithms for computing the average sojourn time were proposed in their work. For more details on PS queues, we may refer to Yashkov (1987) and references therein.

As mentioned above, a typical application of the PS model is in the performance evaluation of Internet-based data transfer system. The Internet system is a huge network made up of enormous nodes around the world. Every node is connected via TCP/IP protocol to communicate with others. Users link their computers to the node by an Internet service provider through many kinds of approaches such as a standard telephone line, TV cables, ADSL, fiber optics, or even wireless connections. Nowadays, with the rapid increase in the number of Internet users, one of the biggest problems and challenges facing the Internet is that of congestion. There are many factors that can affect the speed of access to the Internet, such as computer's hardware and software configuration, the web server speed and the distance between the client and the host computer. But except these factors, the total number of people using the

☆ This research is supported by the National Natural Science Foundation of China (Nos. 71301111, 71171138) and the FSUSE (2012RC23).
☆☆ This manuscript was processed by Area Editor Cheng-Hung Wu.
* Corresponding author at: School of Science, Sichuan University of Science and Engineering, 643000 Zigong, Sichuan, China. Tel.: +86 13540783127.
E-mail address: mmyu75@163.com (M. Yu).

http://dx.doi.org/10.1016/j.cie.2014.09.022
0360-8352/

resource within a predefined period of time (i.e. the number of concurrent users) is an another significant factor affecting the Internet access speed. And more importantly, this factor is extremely difficult to control under a competitive environment because each potential user will spare no efforts to pursue his individual self-interest in such situation. Also, according to the characteristics of PS queue, we may see that the behavior of an Internet node can be well approximated by PS discipline. Hence, discussing joining strategies of customers in such queue will be an interesting and valuable research topic. With the help of these research results, system administrators can establish some control mechanisms to prevent congestion collapse, especially in high traffic and high profit situation.

Actually, during the last few decades, some management scientists and operation researchers have analyzed customer's joining behavior associated with congestion in service facilities. Naor (1969) first studied the situation where arriving customers are admitted or not based on the observed queue length for a single-server facility. In the several years following this article, a similar modeling approach has been extended by many others. Stidham (1985) introduced a fixed reward and a waiting cost for each job passing through the system, and studied the optimal control of admission to a queueing system. In addition, Edelson and Hildebrand (1975) reexamined the work of Naor and generalized the model by analyzing unobservable service queues. Under this situation, when a customer's need for service arises, it is not possible for him to observe the queue before his action. Further, these work spawned research into studying the strategic queueing behavior from an economic viewpoint, such as by Hassin and Haviv (2003), Burnetas and Economou (2007), Economou and Kanta (2008a, 2008b), Boudali and Economou (2012), Economou and Manou (2013), Guo and Zipkin (2007), Guo and Hassin (2011, 2012), Sun, Guo, and Tian (2010), Sun and Li (2014), Wang and Zhang (2011), Zhang, Wang, and Liu (2013), Liu, Ma, and Li (2012), Ma, Liu, and Li (2013) and Li, Wang, and Zhang (2013). However, as we have seen above, most existing research still focused on the economic analysis of the First-Come-First-Served (FCFS) queue, in which the requests of customers or users are attended to in the order that they arrived, without other biases or preferences. Until now, almost no work has been done on the individually and socially optimal joining strategies in the PS queueing system. Except the one done by Altman and Shimkin (1998), no work in this direction has come to our notice. Therefore, a main purpose of this paper is that we will extend the results of Naor for the $M/M/1$ queue under FCFS discipline to the $M/M/1$ queue under PS discipline, and extensively analyze and explore the strategic queueing behavior arising in PS queueing system.

The rest of this paper is organized as follows. In Section 2, we briefly describe the model and review some properties of the queueing system. In Section 3, we determine customer's individually and socially optimal threshold value under a linear reward-cost structure for fully observable case. Section 4 is devoted to considering the customer's equilibrium and socially optimal joining probabilities in the fully unobservable queue. Finally, Section 5 provides the conclusions.

2. Model description

In the EPS queue, each customer who enters the system equally shares the resource of service facility with other present customers. We assume that customers arrive at this system according to a Poisson process with rate λ. The distribution of the service demand is exponential such that if the customer got all the capacity of the server the service time would be exponentially distributed with parameter μ, and whenever n customers are present in the system, each customer receives service at a rate of μ/n. We further suppose that all the customers are homogeneous and delay sensitive, every customer who joins the queue receiving a reward R from service and experiencing a delay cost of C per unit time. To avoid triviality, we impose three conditions on the current model: (1) Customers are risk neutral and maximize their expected net benefit in equilibrium. Under favorable circumstance, namely when the server stays idle, a customer will desire to queue up for receiving the completion of service. This also means that the following inequality must hold

$$R \geqslant \frac{2C}{2\mu - \lambda}.$$

The above inequality can be derived from the explicit expression of the expected conditional sojourn time (see Eq. (16) of this paper). It ensures that the reward for service is not less than the expected cost for a customer who finds the system empty; (2) Intuitively speaking, if, on average, arrivals happen faster than service completions the queue size will grow indefinitely long and the system will not have a stationary distribution. Thus, we require $\rho = \frac{\lambda}{\mu} < 1$ for the queue to be stable; (3) Finally, we suppose that various stochastic processes involved in the system are assumed to be independent of each other.

3. Fully observable queue

3.1. Optimal joining rule for individual optimization

Let $N(t)$ denote the number of customers in the system at time t. Suppose that the system state information, namely $N(t)$, is communicated to customers upon their arrival. Based on the information obtained from service provider, a threshold policy is adopted by customers in a decentralized manner, that is, an arriving customer will join the queue if the queue size falls short of some fixed number; otherwise, he refuses to enter the queue. Now, let us take an arbitrary arriving customer as the tagged one. If the tagged customer enters the system, the total delay cost is determined by his mean sojourn time. Then, his expected net benefit after service completion can be expressed as $U_e = R - C\mathrm{E}[W_n]$, where $W_n (n = 1, 2, \ldots)$ is the tagged customer's sojourn time given that he finds the number of customers in the system is equal to $n-1$ just before his arrival. Thus, a newly arriving customer will join the queue if and only if $U_e = R - C\mathrm{E}[W_n] \geq 0$. The above linear cost function allows us to compute exactly the individual net benefit of the tagged customer for any value of n and to finally get the threshold for individual optimization. In addition, from the cost structure, we find that determining the expected conditional sojourn time is an important step for solving the threshold joining strategy under individual optimization. Here, we will first use the law of total expectation to calculate the Laplace transform of W_n, namely $\mathrm{E}\left[e^{-sW_n}\right]$. Then, differentiating $\mathrm{E}\left[e^{-sW_n}\right]$ with respect to s and evaluating at $s = 0$, we can directly obtain a system of difference equations for the first moment of the conditional sojourn time. Further, finding the kernel of a left-multiplication transformation, the explicit expression for $\mathrm{E}[W_n]$ can be derived.

Following the above ideas and using a first-step argument (i.e. by conditioning on the events that may occur in the next step), for $n \geqslant 2$, $\mathrm{E}\left[e^{-sW_n}\right]$ can be decomposed as

$$\begin{aligned}\mathrm{E}\left[e^{-sW_n}\right] &= \mathrm{E}\left[e^{-sW_n}\mathrm{I}_{\{A<\min(S_1,S_2,\ldots,S_n)\}}\right] \\ &\quad + \mathrm{E}\left[e^{-sW_n}\mathrm{I}_{\{A>\min(S_1,S_2,\ldots,S_n),S_n=\min(S_1,S_2,\ldots,S_n)\}}\right] \\ &\quad + \mathrm{E}\left[e^{-sW_n}\mathrm{I}_{\{A>\min(S_1,S_2,\ldots,S_n),S_n\neq\min(S_1,S_2,\ldots,S_n)\}}\right] \end{aligned} \tag{1}$$

and for $n = 1$,

$$\mathrm{E}\left[e^{-sW_1}\right] = \mathrm{E}\left[e^{-sW_1}\mathrm{I}_{\{A<S_1\}}\right] + \mathrm{E}\left[e^{-sW_1}\mathrm{I}_{\{A>S_1\}}\right], \tag{2}$$