Applied Mathematical Modelling 35 (2011) 4507–4515

Contents lists available at ScienceDirect

Applied Mathematical Modelling

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

A tandem network with a sharing buffer

Wenhui Zhou [a,*], Zhaotong Lian [b]

[a] School of Business Administration, South China University of Technology, China
[b] Faculty of Business Administration University of Macau, Macau SAR, China

ARTICLE INFO

Article history:
Received 15 December 2009
Received in revised form 1 March 2011
Accepted 8 March 2011
Available online 15 March 2011

Keywords:
Buffer sharing
Tandem network
Loss probability
Sojourn time

ABSTRACT

In this paper, we consider a two-stage tandem network. The customers waiting in these two stages share one finite buffer. By constructing a Markov process, we derive the stationary probability distribution of the system and the sojourn time distribution. Given some constraints on the minimum loss probability and the maximum waiting time, we also derive the optimal buffer size and the shared-buffer size by minimizing the total buffer costs. Numerical results show that, by adopting the buffer-sharing policy, the customer acceptance fraction and the delivery reliability are more sensitive to buffer size comparing with the buffer-allocation policy.

1. Introduction

Consider a two-stage tandem network. People usually assume that each stage has an independent buffer (finite or infinite). In reality, in order to utilize the facility effectively, two stages may actually share one buffer (see Fig. 1). Based on our knowledge, we have not found any literature considering the buffer-sharing tandem network so far.

There are many literature evolving around two-stage tandem network. When both buffers have infinite capacity in the two-stage service systems, and both stages are quasi-reversibility, the network has product-form solution (Theorem 4.3 of [1]) and is analytically tractable. When the quasi-reversibility is not follow, most of researchers define the number of customer in the first queue as a level and the number of customer in the second queue as a phase. The system is then involved in a two-dimensional infinite Quasi-Birth-and-Death (QBD) process with infinite blocks (see [2,3]). It is difficult to derive the system performance measures, for example, the customer sojourn time and the busy period. Lian and Liu [4] define the total number of customers in the system as a level, and the total number of customers in the first queue as a phase, so that the system can be handled by constructing a one-dimensional level-dependent QBD process.

When the buffer of the system is finite, the upstream server may be blocked because the downstream buffer is full. Tandem service systems with blocking have received considerable attention in the queueing, communications and manufacturing literature because of their pervasiveness and significance in real life (see the survey papers [5,6]). There are two classes of two-stage network with finite buffer. The first class is that the tandem network is with infinite upstream buffer and finite downstream buffer (for example [7–9]). The second class is that both buffers are finite. When two buffers are finite, not only the upstream server will be blocked when the downstream buffer is full, but also the arriving customers will be reject when the upstream buffer is full. These rejected customers would be more likely to switch to the other service providers if they are rejected frequently. Most of researches in this field consider it as a buffer allocation problem (see [10–12]).

* Corresponding author.
E-mail address: whzhou@scut.edu.cn (W. Zhou).

0307-904X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.apm.2011.03.018

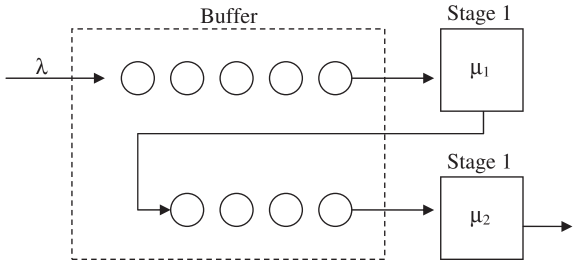

Fig. 1. A service system with a sharing buffer.

In this paper we address two finite buffers and assume that customers in different stages can share the whole or part of the finite buffer. Our objective is to investigate how the buffer sharing policy affects the performance of the system. The rest of this paper is organized as follows. In Section 2, we describe the buffer sharing policy and construct the analytical model. In Section 3, we derive the stationary probability distribution of the system. In Section 4, the customer sojourn time distribution is derived. We provide an optimal buffer design in Section 5 and conduct some numerical experiments in Section 6. Section 7 conclude the paper.

2. The analytical model

We consider a two-stage tandem network. Each stage has a single exponential server with a finite buffer. The service rates of these two servers are μ_1 and μ_2, respectively. We denote $\mu = \mu_1 + \mu_2$. Customers arrive to stage 1 following a Poisson process with parameter λ.

The total physical capacity of the system buffer is finite (denoted by N), and customers are served in an FCFS fashion. The two stages share the same buffer, in which, the buffer threshold for the first stage is N_1, and buffer threshold for the second stage is N_2 ($N_1 \leqslant N$, $N_2 \leqslant N$). We can see that

(1) When $N_1 + N_2 = N$, there is no buffer sharing, and it becomes a buffer allocation problem.
(2) When $N_1 + N_2 > N$ and $N_1 < N$, $N_2 < N$, the sharing policy is called Partial Buffer Sharing.
(3) When $N_1 = N$, $N_2 = N$, the sharing policy is called Complete Buffer Sharing.

As the buffer is finite, the arrival customer will be rejected if the upstream buffer is full, and the upstream server will be blocked if the downstream buffer is full. Throughout this paper, we assume that the "Block After Serve (BAS)" policy is used, that is, if a customer completes service at upstream stage but there is no space in the downstream stage, the upstream server will be blocked, i.e., the customer will remain in upstream stage and prevent that server from working. The blocked server will not provide service to next customer until there is a space available in the downstream stage.

3. Joint probability distribution of queue length

Consider a Markov process $\{(N_1(t), N_2(t), t \geqslant 0\}$ where $N_i(t)$ is the number of customers in the ith station ($i = 1,2$) including the customer in the corresponding server. We define $N_2(t)$ as a level and $N_1(t)$ as a phase. We arrange the states in the standard ascending order as follows:

$$\text{Level } 0 : (0,0), (1,0), \ldots, (N_1+1, 0);$$
$$\text{Level } 1 : (0,1), (1,1), \ldots, (\widetilde{N}_1, 1);$$
$$\ldots, \ldots;$$
$$\text{Level } N_2+1 : (0, N_2+1), (1, N_2+1), \ldots, (\widetilde{N}_{N_2+1}, N_2+1);$$
$$\text{Level } N_2+2 : (1, N_2+2), (2, N_2+2), \ldots, (\widetilde{N}_{N_2+1}, N_2+2),$$

where the level $N_2 + 2$ represents the customer who has finished the service from the first server is blocked, because there are N_2 customers in the second buffer, and

$$\widetilde{N}_i = Min\{N_1+1, N+2-i\}. \quad (1)$$

We denote the state space as S, and the size of S is given by:

$$\|S\| = (N_1+2)(N_2+2) + N - N2 + 1 - \frac{1}{2}(1 + N_1 + N_2 - N)(N_1 + N_2 - N). \quad (2)$$

The state transition diagrams are given below: