Comparison of Input Buffer Scheduling Algorithms in $M_1 + M_2/M/1/k$ System*

FAN Zhong and ZHENG Yingping

(Institute of Automation, Chinese Academy of Sciences · Beijing, 100080, PRC)

Abstract: Space priority and time priority are two categories in priority queueing systems. In the paper, we focused on time priority, and analyzed some buffer scheduling algorithms such as buffer sharing, partial buffer sharing, and buffer separation. In the result, we find that buffer separation can effectively decrease the expiration probability of real-time customers with low realization complexity.

Key words: queueing system; priority; buffer; scheduling algorithm

1 Introduction

According to different Quality of Service (QoS), arriving customers can be divided into several priorities in a queueing system. For example, the current ATM-based Broadband Integrate Service Digital Network (B-ISDN) offers a unique bearer to service for different sources such as voice, data and video. Real-time sources like voice, video must be transfer from origination to destination in a restricted session, but nonreal-time source like data has not strict requirement for delay in the network. If only one kind of QoS is provided for various sources, a lot of network resources that are allocated to low QoS sources will be wasted when strict QoS requirements for some sources are satisfied. If various sources have been provided different QoS according to their priority, network can acquire higher throughput. Buffer control is an effective way to enforce priority of services.

There are two categories of dividing priority classes of customers, space priority and time priority disciplines. Space priority discipline is a kind of buffer admission policy that ensures their different loss probability by discarding selected customers with low priority. Time priority discipline is a kind of customers scheduling policy that rearranges the order of the customers which will be served to ensures their different sojourn time requirements in the system. Both of the disciplines may be also used at the same time.

Kröner has analyzed $M/D/1/k$ system with space priority, and compared the performances of three kinds of buffer scheduling algorithms such as push-out mechanism, partial buffer sharing and route separation in [1]. This paper is concerned with time priority discipline, which can be used to ensure the different delay requirements for various sources in telecommunications network. While a real-time customer's sojourn time in the system is larger than a threshold $T(T>0)$, it will expire, and be discarded by the system. For decreasing

* This paper is supported by National Natural Science Foundation of China (69635030).
Manuscript received Sep. 10, 1996, revised May 9, 1997.

expiration probability of real-time customers, they are marked higher time priority. Therefore customers can be divided into two classes according to their different requirements of sojourn time in the system. Class 1 stands for real-time customers with higher priority, and Class 2 stands for nonreal-time customers with lower priority.

In this paper, we have compared the buffer scheduling schemes such as buffer sharing, partial buffer sharing and buffer separation. In Section 2, some basic assumptions of the paper have been defined. The formulas of loss probability and expiration probability are deduced in Section 3. In Section 4, some numerical examples are illustrated. A brief summary of main results is given in Section 5.

2 Assumption

It is assumed that the waiting room for arriving customers is k in the system. Newly arriving customers only those who find the system with strictly less than k customers will be allowed entry. Further arriving customers will be refused to enter into the system and depart immediately without being served while the buffer is full. Arriving customers are divided into two classes. Class 1 stands for real-time customers, which has a strict delay requirement, that is, its sojourn time in the system can not be larger than a threshold $T(T>0)$. Class 2 stands for nonreal-time customers, which has no such requirement. Class 1 is signed as high priority, and can immediately interrupt service of Class 2 to gain the server, called preemptive priority. Assumed that customers of both classes will be generated according to two independent Poisson processes with parameter λ_1 and λ_2, respectively. There is a server with negative-exponential distribution at service rate μ for both classes. Therefore, the system can be abbreviated to the form of $M_1 + M_2/M/1/k$.

3 Scheduling Algorithm Analysis

Some buffer scheduling algorithms, such as FCFS, buffer sharing, partial buffer sharing and buffer separation, are analyzed and compared in this section.

In the following, we denote that $\rho = (\lambda_1 + \lambda_2)/\mu, \rho_1 = \lambda_1/\mu, \rho_2 = \lambda_2/\mu$.

3.1 FCFS

This is a kind of simple queueing system with first comer first service (FCFS) rule. In the system, there is no priority control for arriving customers.

Because waiting room is finite, it always exists an equilibrium solution in the system. We can obtain probability function distribution of queueing length in the system.

$$p_n = \begin{cases} \dfrac{1}{k+1}, & \rho = 1, \\ \dfrac{(1-\rho)\rho^n}{1-\rho^{k+1}}, & \rho \neq 1. \end{cases} \qquad n = 0,1,2,\cdots,k. \tag{1}$$

Probability density function of customer's sojourn time in the system is

$$u(t) = \begin{cases} \dfrac{1}{k}\displaystyle\sum_{n=0}^{k-1} \frac{\mu(\mu t)^n}{n!} e^{-\mu t}, & \rho = 1, \\ \displaystyle\sum_{n=0}^{k-1} \frac{(1-\rho)\rho^n \mu(\mu t)^n}{(1-\rho^k)n!} e^{-\mu t}, & \rho \neq 1. \end{cases} \qquad t \geqslant 0. \tag{2}$$

Loss probability of customers is refered to that when which a further arriving customer finds that the queueing length in the buffer is k. In FCFS system, there is no priority control for arriving customers. Hence, the loss probability of both Class 1 and Class 2 will be identical, and are only related to $\lambda_1 + \lambda_2$. Therefore, the formula of both real-time and nonreal-time customers' loss probability is

$$p_{\text{loss}} = \begin{cases} \dfrac{1}{k+1}, & \rho = 1, \\ \dfrac{(1-\rho)\rho^k}{1-\rho^{k+1}}, & \rho \neq 1. \end{cases} \tag{3}$$

The expiration probability of Class 1 is that a customer's sojourn time is larger than a threshold $T(T > 0)$. In FCFS system, the expiration probability is

$$p_{ex} = P\{u(t) > T)\} = \begin{cases} \dfrac{\mathrm{e}^{-uT}}{k} \displaystyle\sum_{n=0}^{k-1} \sum_{i=0}^{n} \frac{(\mu T)^{n-i}}{(n-i)!}, & \rho = 1, \\ \dfrac{(1-\rho)\mathrm{e}^{-uT}}{1-\rho^k} \displaystyle\sum_{n=0}^{k-1} \rho^n \sum_{i=0}^{n} \frac{(\mu T)^{n-i}}{(n-i)!}, & \rho \neq 1. \end{cases} \tag{4}$$

3.2 Buffer Sharing

In this algorithm, both real-time and nonreal-time customers are sharing a public buffer. Class 1, real-time customers, have pre-empty priority, which means Class 1 can interrupt service process of Class 2 at any time. When there is no customer of Class 1 in the system, Class 2 can get the server. No customer can interrupt service process of Class 1. Therefore, the customers of Class 1 are always queueing in front of all of Class 2 in the buffer. A newly arriving customer of Class 1 that finds buffer is full can push-out one of Class 2 in the queue and enter into the system. A newly arriving customer always queues in end of homogeneous customers which already existed in the queue. A further arriving customer that finds buffer is full with homogeneous customers will be refused into the system and depart forever.

According to buffer sharing algorithm, the loss probability of Class 1 is that when a newly arriving customer of Class 1 finds buffer is already full with homogeneous customers. Thus, we can get the loss probability of Class 1,

$$p_{\text{real-time}} = \begin{cases} \dfrac{1}{k+1}, & \rho_1 = 1, \\ \dfrac{(1-\rho_1)\rho_1^k}{1-\rho_1^{k+1}}, & \rho_1 \neq 1. \end{cases} \tag{5}$$

The loss probability of Class 2 is the combined traffic's loss probability while there is no loss of Class 1. Hence,

$$p_{\text{nonreal-time}} = p_{\text{loss}} - p_{\text{real-time}}. \tag{6}$$

We can obtain the formula of loss probability of Class 2 from equations (3) and (5),

$$p_{\text{nonreal-time}} = \begin{cases} \dfrac{1}{k+1} - \dfrac{(1-\rho_1)\rho_1^k}{1-\rho_1^{k+1}}, & \rho_1 \neq 1, \rho = 1, \\ \dfrac{(1-\rho)\rho^k}{1-\rho^{k+1}} - \dfrac{1}{k+1}, & \rho_1 = 1, \rho \neq 1, \\ \dfrac{(1-\rho)\rho^k}{1-\rho^{k+1}} - \dfrac{(1-\rho_1)\rho_1^k}{1-\rho_1^{k+1}}, & \rho_1 \neq 1, \rho \neq 1. \end{cases} \tag{7}$$

The expiration probability of Class 1 is

$$p_{ex} = \begin{cases} \dfrac{e^{-uT}}{k} \sum_{n=0}^{k-1} \sum_{i=0}^{n} \dfrac{(\mu T)^{n-i}}{(n-i)!}, & \rho = 1, \\ \dfrac{(1-\rho_1)e^{-uT}}{1-\rho_1^k} \sum_{n=0}^{k-1} \rho_1^n \sum_{i=0}^{n} \dfrac{(\mu T)^{n-i}}{(n-i)!}, & \rho \neq 1. \end{cases} \tag{8}$$

3.3 Partial Buffer Sharing

Combined traffic enters a common buffer with total k waiting room. When queueing length in the system is larger than $k_1 (k_1 < k)$, a newly arriving customer of Class 1 will be refused into the system and depart forever, but of Class 2 can still enter the system until all of the waiting room is full. We call it partial buffer sharing algorithm.

According to the rule of this algorithm, arriving customers of Class 1 will be lost while queueing length is larger than k_1 . Hence, the loss probability of Class 1 is of combined traffic with buffer size k_1. Thus,

$$p_{\text{real-time}} = \begin{cases} \dfrac{1}{k_1 + 1}, & \rho = 1, \\ \dfrac{(1-\rho)\rho^{k_1}}{1-\rho^{k_1+1}}, & \rho \neq 1. \end{cases} \tag{9}$$

The formula of the expiration probability of Class 1 is

$$p_{ex} = \begin{cases} \dfrac{e^{-uT}}{k_1} \sum_{n=0}^{k_1-1} \sum_{i=0}^{n} \dfrac{(\mu T)^{n-i}}{(n-i)!}, & \rho = 1, \\ \dfrac{(1-\rho)e^{-uT}}{1-\rho^{k_1}} \sum_{n=0}^{k_1-1} \rho^n \sum_{i=0}^{n} \dfrac{(\mu T)^{n-i}}{(n-i)!}, & \rho \neq 1. \end{cases} \tag{10}$$

The loss probability of Class 2 can be deduced by using birth-death process. Because the buffer size is finite, we can always obtain equilibrium solution. The loss probability of Class 2 is

$$p_{\text{nonreal-time}} = p_k = \rho^{k_1} \rho_2^{k-k_1} p_0, \tag{11}$$

where p_0 is given by

$$p_0 = \left[\sum_{n=0}^{k_1} \rho^n + \sum_{n=k_1+1}^{k} \rho_2^n \right]^{-1} = \left[\frac{1-\rho^{k_1+1}}{1-\rho} + \frac{\rho_2^{k_1+1} - \rho_2^{k+1}}{1-\rho_2} \right]^{-1}. \tag{12}$$

3.4 Buffer Separation

In this algorithm, buffer is divided into two separated parts B_1 and B_2 , and their capacities are k_1 and k_2 , respectively. An arriving customer of Class 1 and 2 will enter B_1 and B_2 , respectively. Hence, we know that there is no interaction each other for obtaining space of the buffer. Assume that customers of Class 1 can get server with preemptive priority, the loss probability of Class 1 is only related to k_1 with no effect of Class 2. Thus, the formula of Class 1 loss probability is

$$p_{\text{real-time},0} = \begin{cases} \dfrac{1}{k_1 + 1}, & \rho_1 = 1, \\ \dfrac{1-\rho_1}{1-\rho_1^{k_1+1}}, & \rho_1 \neq 1. \end{cases} \tag{13}$$

$$p_{\text{real-time}} = \begin{cases} \dfrac{1}{k_1+1}, & \rho_1 = 1, \\ \dfrac{(1-\rho_1)\rho_1^{k_1}}{1-\rho_1^{k_1+1}}, & \rho_1 \neq 1. \end{cases} \tag{14}$$

The expiration probability of Class 1 is

$$p_{ex} = \begin{cases} \dfrac{\mathrm{e}^{-uT}}{k_1}\sum\limits_{n=0}^{k_1-1}\sum\limits_{i=0}^{n}\dfrac{(\mu T)^{n-i}}{(n-i)!}, & \rho = 1, \\ \dfrac{(1-\rho_1)\mathrm{e}^{-uT}}{1-\rho_1^{k_1}}\sum\limits_{n=0}^{k_1-1}\rho_1^n\sum\limits_{i=0}^{n}\dfrac{(\mu T)^{n-i}}{(n-i)!}, & \rho \neq 1. \end{cases} \tag{15}$$

The loss probability of Class 2 is

$$p_{\text{nonreal-time}} = \begin{cases} \dfrac{1}{k_2+1}, & \rho_2' = 1, \\ \dfrac{(1-\rho_2')\rho_2'^{k_2}}{1-\rho_2'^{k_2+1}}, & \rho_2' \neq 1. \end{cases} \tag{16}$$

where

$$\rho_2' = \frac{\lambda_2}{p_0\mu} = \begin{cases} \dfrac{(k_1+1)\lambda_2}{\mu} \\ \dfrac{(1-\rho_1^{k_1+1})\lambda_2}{(1-\rho_1)\mu} \end{cases} = \begin{cases} (k_1+1)\rho_2, & \rho_1 = 1, \\ \dfrac{1-\rho_1^{k_1+1}}{1-\rho_1}\rho_2, & \rho_1 \neq 1. \end{cases}$$

4 Numerical Analysis

For simplifying calculation, we assume that both of classes' arrive-rate is equal, that is, $\lambda_1 = \lambda_2$. Assume that $\rho = \dfrac{\lambda_1+\lambda_2}{\mu} = 0.8$, and $k_1 = \dfrac{k}{2}, k_2 = k, \mu T = 5$.

Numerical results show that the expiration probabilities of Class 1 under both buffer sharing and buffer separation algorithm will incresae along with the augment of ratio $\lambda_1/(\lambda_1+\lambda_2)$ (see Fig. 1). There are no relation between the arriving ratio and the expiration probability of Class 1 when using other algorithms, but with a higher expiration probability than

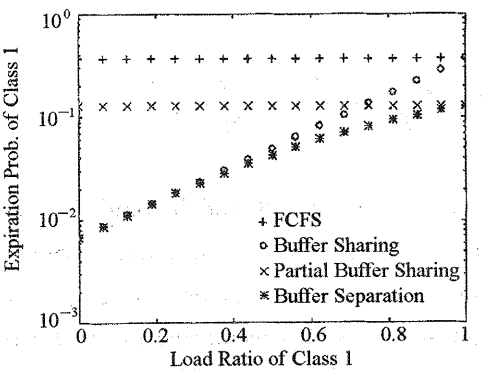

Fig. 1 In different arriving ratio at the traffic, the expiration probability of Class 1

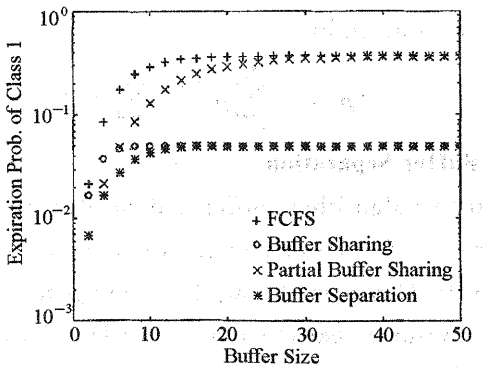

Fig. 2 In different buffer size, the expiration probability of Class 1

the above (just as the limiting value). Therefore, both buffer sharing and buffer separation can meet the requirement of real-time customer expiration probability. From the numerical results, we know that the system can acquire better real-time performance by using the algorithm of buffer separation.

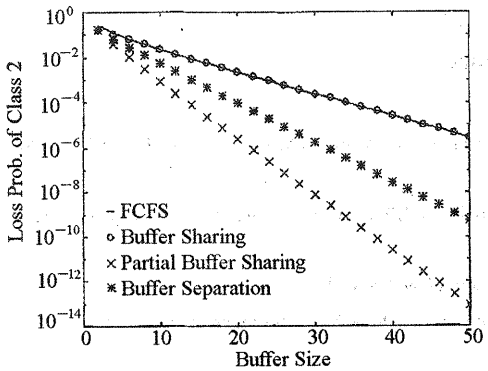

Fig. 3 In different buffer size, the loss probability of Class 2

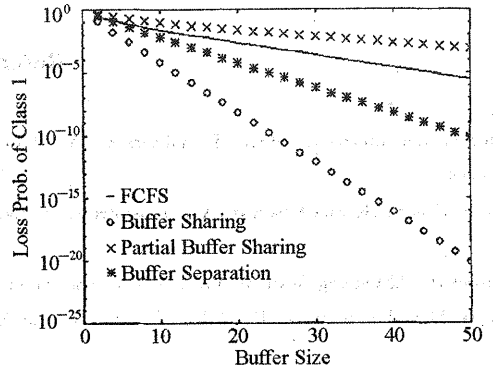

Fig. 4 In different buffer size, the loss probability of Class 1

Form Fig. 3 and 4, we can see that both classes' loss probabilities decrease along with growth of buffer size k. On the other hand, see Fig. 2, the expiration probability of Class 1 has a big decrease along with k in the case that k is not big, and becomes stable while k is big enough (such as $k > 20$). Therefore, the buffer size must have a trade-off between the loss and expiration probability.

Table 1 shows the system performances under different scheduling algorithms.

Table 1 Comparison of different scheduling algorithms

	FCFS System	Buffer Sharing	Partial Buffer Sharing	Buffer Separation
Expiration Prob. of Class 1	High	Low	High	Low
Loss Probability of Class 1	High	Low	High	Low
Loss Probability of Class 2	High	High	Low	Low
Realization Complexity	Low	High	Low	Low

5 Conclusion

In the paper, we have analyzed several buffer scheduling algorithms such as buffer sharing, partial buffer sharing and buffer separation, and compared them with FCFS system, which has no priority control for customers. We can see from the numerical results that buffer separation can meet the requirement of expiration probability of real-time customers effectively with lower realization complexity. Buffer sharing can also satisfy the requirement of expiration probability of real-time customers effectively with low loss probability, but its realization complexity is high. Therefore, buffer sharing can be used for those systems in which both loss and expiration probability of one kind of customers have been strictly required. In the algorithm of partial buffer sharing, time priority control can not be realized because the loss probability of real-time customer is high. On the other hand, becuase a strict loss probability of one kind of customers can be ensured with lower realization complexity, the partial buffer sharing algorithm can be used well in the system, which one class of customers has very strict loss probability requirement. It is coincide with the analysis results in [1].

Reference

1 Kröner, H., Hebuterne, G., Boyer, P. and Gravey, A.. Priority management in ATM switching nodes. IEEE JSAC, 1991, 9(3): 418—427

2 Cidon, I., Guerin, R. and Khamisy, A.. On protective buffer policies. IEEE/ACM Trans. Networking, 1994, 2(3): 240—246

3 Kleinrock, L.. Queueing Systems, Vol I: Theory and Application. New York: Wiley-Interscience, 1975

4 Akimaru, H. and Kawashima, K.. Teletraffic: Theory and Applications. New York: Springer-Verlag, 1993

$M_1+M_2/M/1/k$ 系统输入队列调度算法性能比较

范 中 郑应平

(中国科学院自动化研究所·北京, 100080)

摘要: 排队系统中优先级的划分方法主要有空间优先和时间优先两种类型. 本文针对时间优先级系统进行分析, 通过对缓冲器完全共享、缓冲器部分共享和分离缓冲器等三种缓冲器调度控制算法的分析比较, 我们可以看出, 分离缓冲器调度算法能够有效地减少实时性顾客的失效概率, 获得满意的控制效果, 并且其实现的复杂度也较低.

关键词: 排队系统; 优先级; 缓冲器; 调度算法

本文作者简介

范 中 1968年生. 分别于1991年, 1994年获得工学学士、工学硕士学位. 自1994年至1997年, 在中国科学院自动化研究所攻读自动控制理论与应用专业博士学位. 主要研究方向为DEDS理论及其应用, 通信系统建模及优化.

郑应平 见本刊1998年第1期第52页.