AN ANALYSIS OF A TANDEM QUEUEING SYSTEM FOR FLOW CONTROL IN COMPUTER NETWORKS*

W. W. Chu
University of California
Los Angeles, U.S.A.

G. Fayolle
I.R.I.A.
Rocquencourt, France

D. Hibbits
University of California
Los Angeles, U.S.A.

ABSTRACT

A tandem queueing system with constant service times and threshold control is modeled and analyzed in this paper. The input to the first queue is controlled by the buffer occupancy of the second queue. When the second queue has more than N_0 customers, the input to the first queue will be rejected. The input to the second queue consists of the output from the first queue and an external input which is assumed to be Poisson distributed. The behavior of such a queueing system is analyzed and portrayed in graphs. The threshold control rejects input traffic to the first queue and avoids congestion at the second queue. As a result, the delay for an arrival to be serviced by both of the queues is much lower than the case without threshold control. As N_0 increases, the system behavior approaches the case of the system without threshold control. Such a queueing model is motivated by congestion control in a computer network. An example is given to illustrate the applications.

INTRODUCTION

An important problem in computer communication systems concerns buffer control at each node. This directly affects system performance in terms of delay in transmitting a message from the source to its destination as well as system throughput. Several control schemes based on buffer control [1,2,3] and channel traffic intensity [4] have been proposed. The basic concept is that whenever the system reaches a certain threshold state (buffer queue length or channel intensity), input arrivals that cause network congestion will be temporarily rejected in order to avoid further congestion. Using queueing representation, this can be described as a system of two queues connected in tandem.

In this paper we formulate a mathematical model for such a queueing system both with and without threshold control on input arrivals. Numerical results are portrayed in graphs to

* This research is supported in part by the Office of Naval Research, Math Science Division, Contract No. N00014-75-C-0650.

show the effect of threshold on the delay and throughput of the system. An example illustrates an application of the results.

MODEL FOR TANDEM QUEUE WITHOUT THRESHOLD CONTROL

Let us consider the tandem queueing system with Poisson arrival rates $\lambda_1(n)$ and λ_2 and constant service times D_1 and D_2 as shown in Figure 1. Let m denote the number of customers in the first queue, n denote the number of customers in the second queue, and $q(m,n)$ denote the probability of having m customers in the first queue and n customers in the second queue. We assume that $D_1 = D_2 = 1$, and $\lambda_1(n)$ is a function of n . That is,

$$\lambda_1(n) = \begin{cases} \lambda_1 & n \leq N_0 \\ \lambda_0 & n > N_0 \end{cases}$$

where N_0 is the threshold value of the queue length of the second queue.

We shall derive the steady state equations for such system. First let us consider the system without threshold control that is $\lambda_1(n) = \lambda$ for all n . By examining the state of the system at the departure instant of each customer, we obtain the following state equations:

$$q(m,0) = e^{-(\lambda_1 + \lambda_2)} \frac{\lambda_1^m}{m!} [q(0,1) + q(0,0)] \quad m \geq 0 \quad (1)$$

$$\begin{aligned} q(m,n) = e^{-(\lambda_1 + \lambda_2)} & \left[\sum_{k=0}^m \sum_{\ell=0}^{n-1} \frac{\lambda_1^k}{k!} \cdot \frac{\lambda_2^\ell}{\ell!} q(m+1-k, n-\ell) \right. \\ & + \frac{\lambda_1^m}{m!} \cdot \sum_{\ell=0}^n q(0, n+1-\ell) \frac{\lambda_2^\ell}{\ell!} + \frac{\lambda_2^{n-1}}{(n-1)!} \sum_{k=0}^m \frac{\lambda_1^k}{k!} q(m+1-k, 0) \\ & \left. + \frac{\lambda_1^m}{m!} \cdot \frac{\lambda_2^n}{n!} q(0,0) \right] \quad m \geq 0, n \geq 0. \end{aligned} \quad (2)$$

Equation (1) states that the probability of having m customers in the first queue and no customers in the second queue is equal to the probability of having m customers arrive at the first queue, and no customers arrive at the second queue during one service period times the probability of the event that there are no customers in either queue immediately after service. Equation (2) states that the probability of having m customers in the first queue and n customers in the second queue is equal to: (a) the probability that k customers arrive at the first queue, and ℓ customers arrive at the second queue during the service

period in which there were $m+1-k$ customers in the first queue and $n-l$ customers in the second queue; (b) plus the probability of having m customers arrive at the first queue and l customers arrive at the second queue during the service interval in which there were no customers in the first queue and $n+1-l$ customers in the second queue; (c) plus the probability of having k customers arrive at the first queue and $n-1$ customers arrive at the second queue during the service interval in which there were $m+1-k$ customers in the first queue and no customers in the second queue; (d) plus the probability of m customers arriving at the first queue and n customers arriving at the second queue during the service time with no customers in the system. Let the moment generating function of the number of customers in the first queue conditioned by n customers in the second queue be

$$G_n(x) = \sum_{m=0}^{\infty} q(m,n) x^m. \quad (3)$$

Substituting (1) into (3) and letting $n = 0$, we obtain

$$G_0(x) = e^{\lambda_1 x} q(0,0). \quad (4)$$

Let the generating function of the joint distribution function of the number of customers in both queues be

$$F(x,y) = \sum_{n=0}^{\infty} G_n(x) y^n. \quad (5)$$

Substituting (2) and (4) into (3), we have

$$\begin{aligned} G_n(x) e^{\lambda_1 + \lambda_2 - \lambda_1 x} &= \sum_{l=0}^{n-1} \frac{\lambda_2^l}{l!} \frac{1}{x} [G_{n-l}(x) - q(0, n-l)] \\ &+ \sum_{l=0}^n q(0, n+1-l) \frac{\lambda_2^l}{l!} + \frac{\lambda_2^n}{n!} q(0,0) + \frac{\lambda_2^{n-1}}{(n-1)!} \frac{1}{x} [G_0(x) - q(0,0)] \end{aligned} \quad (6)$$

Multiplying (6) by Y^n , from (5) we have

$$\begin{aligned} [F(x,y) - G_0(x)] e^{\lambda_1 + \lambda_2 - \lambda_1 x} &= e^{\lambda_2 y} \frac{1}{y} [F(0,y) - q(0,0)] - q(0,1) + (e^{\lambda_2 y} - 1) q(0,0) + \frac{y}{x} e^{\lambda_2 y} [G_0(x) - q(0,0)] \\ &+ \frac{1}{x} e^{\lambda_2 y} [F(x,y) - G_0(x) - F(0,y) - q(0,0)] \end{aligned}$$

After simplification, we have

$$F(x,y) [e^{\lambda_1(1-x) + \lambda_2(1-y)} - \frac{1}{x}] = F(0,y) \left(\frac{1}{y} - \frac{1}{x} \right) + F(0,0) (y-1) \left(\frac{1}{y} + \frac{e^{\lambda_1 x} - 1}{x} \right) \quad (7)$$

where $F(0,y)$ can be determined by using the singular curve

$$e^{\lambda_1(1-x)+\lambda_2(1-y)} = \frac{1}{x}$$

We can derive an analytic expression for the mean number of customers in each queue of the uncontrolled system from (7). Evaluating the term $F(0,0)$ using the conservation law $F(1,1)=1$, we have

$$1 = \lim_{y \rightarrow 1} F(1,y) = \frac{F(0,1)(1-1)+F(0,0)(1-1)(1+e^{\lambda_1}-1)}{e^0 - 1} = \frac{0}{0}$$

Using L'Hospital's Rule yields

$$F(0,0) = [-\lambda_2 + F(0,1)]e^{-\lambda_1}$$

Since $F(0,1)$ is the marginal probability of having no customers in the first queue,

$$F(0,1) = 1 - \lambda_1, \text{ and } F(0,0) = (1-\lambda_1-\lambda_2)e^{-\lambda_1}.$$

The mean total number of customers in both queues is given

$$\bar{m} + \bar{n} = \frac{d}{dx} F(x,1) \Big|_{x=1} + \frac{d}{dy} F(1,y) \Big|_{y=1}$$

To eliminate the term $F(0,y)$, we can take the derivative $F(x,y)$ on the curve $y=x$, then

$$\bar{m} + \bar{n} = \frac{d}{dx} F(x,x) \Big|_{x=1} = \frac{d}{dx} \left[\frac{(1-\lambda_1-\lambda_2)e^{-\lambda_1}(x-1)e^{\lambda_1 x}}{xe^{(\lambda_1+\lambda_2)(1-x)} - 1} \right] \quad (8)$$

Using L'Hospital's Rule twice, we obtain

$$\bar{m} + \bar{n} = \frac{2\lambda_1(1-\lambda_1-\lambda_2) + 1 - (1-\lambda_1-\lambda_2)^2}{2(1-\lambda_1-\lambda_2)}$$

The mean number of customers in the first queue is the same as the M/D/1 queue,

$$\bar{m} = \frac{2\lambda_1 - \lambda_1^2}{2(1-\lambda_1)} \quad (9)$$

The mean number for all customers in the second queue is given by the difference between Eq. (8) and Eq. (9). Since the external arrivals to the second server are not distinguished from customers arriving from the first server, both have the same expected time in the second queue

D_1 is the mean time required for customers to flow through both queues after they have been accepted at the first queue. From Little's Result, we have

$$D_1 = \bar{m} / \lambda_1 + \bar{n} / (\lambda_1 + \lambda_2) \quad (10)$$

D_2 is the mean time required for customers to flow through the second queue. From Little's Result, we have

$$D_2 = \bar{n} / (\lambda_1 + \lambda_2) \quad (11)$$

MODEL FOR TANDEM QUEUES WITH THRESHOLD CONTROL

We assume that arrivals to the first queue and second queue are Poisson distributed with parameters λ_1 and λ_2 , respectively, when $n \leq N_0$. If there are more than N_0 customers in the second queue, then arrivals to the first queue will be Poisson distributed with parameter λ_0 . In our study, we let $\lambda_0 = 0$.

Because of the complexity of the tandem queue model with threshold control, we use a numerical technique for solving the state equilibrium probabilities. We use a finite-state approximation in which the waiting rooms of the first and second queues are limited to M and N respectively. This approximate model has $(M+1)(N+1)$ states whose equilibrium probabilities can be found by solving a finite set of equilibrium equations. Of course, M and N must be chosen large enough that the probability is negligible that arrivals to either queue are lost.

Let the probability of k arrivals to the first queue be $\pi_1(k)$ for $n \leq N_0$ and $\pi_0(k)$ for $n > N_0$, and let the probability of k arrivals to the second queue be $\pi_2(k)$. For Poisson arrival process, we have

$$\pi_1(k) = e^{-\lambda_1} \lambda_1^k / k!, \quad \pi_0(k) = e^{-\lambda_0} \lambda_0^k / k! \quad \text{and}$$

$$\pi_2(k) = e^{-\lambda_2} \lambda_2^k / k!$$

Let (a,b) be the state in which there are a customers in the first queue and b customers in the second queue, $a < M$, and $b < N$. Likewise, let (c,d) be the state that has c customers in the first queue and d customers in the second queue, $c < M$ and $d < N$. The transition probability from state (a,b) to state (c,d), $P[(a,b) \rightarrow (c,d)]$, is as follows:

$$P[(a,b) \rightarrow (c,d)] = \begin{cases} \pi_1(c) \pi_2(d) & \text{for } a=0, b=0 \\ \pi_1(c) \pi_2(d-b+1) & \text{for } a=0, 1 < b \leq N_0 \\ \pi_0(c) \pi_2(d-b+1) & \text{for } a=0, b > N_0 \\ \pi_1(c-a+1) \pi_2(d-1) & \text{for } a > 0, b=0 \\ \pi_1(c-a+1) \pi_2(d-b) & \text{for } a > 0, 0 < b \leq N_0 \\ \pi_0(c-a+1) \pi_2(d-b) & \text{for } a > 0, b > N_0 \end{cases} \quad (12)$$

For the case in which $c=M$, because k or more arrivals will leave M customers in the first

queue, $\pi_1(k)$ is replaced by $1 - \sum_{i=0}^{k-1} \pi_1(i)$ and $\pi_0(k)$ by $1 - \sum_{i=0}^{k-1} \pi_0(i)$ In (12).

Likewise, when $d=N$, $\pi_2(k)$ is replaced by $1 - \sum_{i=0}^{k-1} \pi_2(i)$.

The equilibrium state probability of the infinite queue system can be approximated by the corresponding finite state equilibrium state probability. That is,

$$q(m,n) \approx \begin{cases} q'(m,n) & \text{for } 0 \leq m \leq M \text{ and } 0 \leq n \leq N \\ 0 & \text{for } m > M \text{ or } n > N \end{cases}$$

where $q'(m,n)$ is the equilibrium state probability of m customers in the first queue and n in the second queue, based on the finite state model. Two measures were needed to estimate the accuracy of these approximations: 1) the probability of one or more customers being lost at the first queue during a service interval, and 2) the probability of one or more customers being lost at the second queue. These probabilities are computed from $q'(m,n)$.

For m customers in the first queue, the probability that one or more arrived customers are lost during the next service interval in the first queue is

$$\Pr[\text{loss at first queue}|m] = \begin{cases} 1 - \sum_{k=0}^M \pi_1(k) & \text{for } m=0, n \leq N_0 \\ 1 - \sum_{k=0}^M \pi_0(k) & \text{for } m=0, n > N_0 \\ 1 - \sum_{k=0}^{M-m+1} \pi_1(k) & \text{for } m>0, n \leq N_0 \\ 1 - \sum_{k=0}^{M-m+1} \pi_0(k) & \text{for } m>0, n > N_0 \end{cases}$$

Removing the condition on m, we have

$$\Pr[\text{loss at first queue}] = \sum_{m=0}^M \sum_{n=0}^N q'(m,n) \Pr[\text{loss at first queue}|m] . \quad (13)$$

Similarly,

$$\Pr[\text{loss at second queue}|n] = \begin{cases} 1 - \sum_{k=0}^N \pi_2(k) & \text{for } m=0, n=0 \\ 1 - \sum_{k=0}^{N-1} \pi_2(k) & \text{for } m>0, n=0 \\ 1 - \sum_{k=0}^{N-n+1} \pi_2(k) & \text{for } m=0, n>0 \\ 1 - \sum_{k=0}^{N-n} \pi_2(k) & \text{for } m>0, n>0 \end{cases}$$

and

$$\Pr[\text{loss at second queue}] = \sum_{m=0}^M \sum_{n=0}^N q'(m,n) \Pr[\text{loss at second queue}|n] . \quad (14)$$

To minimize the required computation time for the finite state probabilities $q'(m,n)$, M and N used in our finite state model were varied according to the values of λ_1 , λ_2 and N_0 . In our numerical computations, M and N were chosen large enough to keep the loss probabilities of the first queue (13) and second queue (14) less than 10^{-3} .

In order to verify the result of the numerical computation from our finite writing room model, a computer program written in Fortran is used to simulate the tandem queueing system with threshold control on input arrivals. The input parameters are λ_1 , λ_2 and N_0 .

Fifty consecutive epochs of 4000 service intervals each are simulated. The first epoch is ignored to allow the queues to reach equilibrium.

Data measured during the simulation are: offered load to first queue λ_1 , carried load of first queue λ_1' , external arrival rate to second queue λ_2 , mean number in first queue $\bar{m}$, mean number in second queue $\bar{n}$, and acceptance probability = $\Pr(\text{number in second queue} \leq N_0)$. Simulation results for flow time through both queues are shown in Fig. 2a and Fig. 2b.

DISCUSSION OF RESULTS

Using the queueing models developed in the previous sections, we evaluate the behavior of the tandem queueing system. Since the queue is of finite size, the traffic offered at the queue may not always be entered and carried by the system. Therefore, the offered load equals the sum of carried load and loss traffic due to input rejection.

The expected queueing delay for an arrival to go through both queues depends on the carried load of the first queue λ_1' and external arrival rates at the second queue λ_2 . Such relationships for the cases $N_0=0$ and 5 are portrayed in Figures 2a and 2b respectively. Clearly, the minimum delay time for an arrival to go through both queues is two service times. For a given λ_2 , the queueing delay increases as the carried load of the first queue λ_1' increases. For a given λ_1' , the delay increases as λ_2 increases. Comparing Figure 2a with 2b, we notice that because of the input threshold control, the carried load of the first queue for $N_0=0$ is much lower than that of $N_0=5$. A simulation program has been written to compare the simulation results with the analytical ones. As shown in Figures 2a and 2b, the simulation and analytical results closely agree with each other.

Figures 3a and 3b present results similar to Figures 2a and 2b, except that the expected delay is expressed as a function of offered load rather than carried load of the first queue. Clearly, as N_0 increases, the system provides less and less input control, and as N_0 approaches infinity, the system reduces to a tandem queueing system without threshold control. The characteristics of such a system are shown in Figure 4. In this case, the carried load of the first queue is equal to its offered load, and the delay approaches infinity as $\lambda_1+\lambda_2$ approaches 1. Comparing Figure 4 with Figures 3a and 3b, we notice that for a given λ_1 and λ_2 , the queueing system with threshold control provides rejection to input traffic and therefore yields lower delay than the system without threshold control. Because of this threshold control, the carried load is lower than the offered load at the first queue, which prevents the system from reaching its saturation.

Figures 5a and 5b present the throughput of the first queue as a function of offered load at the first queue and second queue for $N_0=0$ and 5 respectively. Clearly, the throughput of

the first queue increases as N_o increases and as the external traffic load arrivals at the second queue λ_2 decreases. Figures 6a and 6b present the average delay incurred by an arrival at the second queue as a function λ_2 for $N_o=0$ and 5 respectively. We notice that the delay at the second queue increases as λ_2 increases. Comparing Figures 6a and 6b, for a given λ_2 , the dependence of the delay at the second queue upon λ_1 increases as N_o increases. As N_o approaches infinity, the system reduces to a tandem queueing system without threshold control. As a result, the queueing delay at the second queue varies greatly with λ_1 as shown in Figure 7.

Figure 8 shows the effect of N_o on the expected delay through both queues for offered load to the first queue and for selected external arrival rates at the second queue λ_2 . Figure 9 presents the effect of N_o on the expected delay at the second queue for selected arrival rate λ_2 and offered load to Re first queue.

EXAMPLE

Consider two computer networks interconnected through a gateway as shown in Figure 10. The gateway is assumed in network A. Traffic entering the gateway consists of traffic generated from network A, and traffic generated from network B that communicates with nodes in network A. Because of distance and priorities, we assume traffic generated from network B has priority for handling at the gateway higher than that of traffic generated from network A. Thus, when the buffer occupancy at the gateway reaches a certain threshold level, in order to avoid network congestion at network A, traffic from network A entering the gateway can be rejected.

Assuming the average traffic arrival rate entering the gateway from network A is 0.4 packets/packet service times, the traffic entering the gateway from network B is 0.7 packets/packet service times. If no flow control is applied at the gateway, the gateway reaches its saturation and gradually affects the other nodes in network A. Eventually, network A will reach its saturation and greatly degrade the network throughput. If flow control is applied at the gateway with a threshold level of $N_o=5$, the expected queueing delay at the gateway is 4.7 service times (Figure 9) and the delay for packets generated from network A entering the gateway is 5.8 packet service times. The carried load at the gateway is 0.26 packets/packet service times, which is 0.14 packets/packet service times less than the offered load. By selecting the appropriate threshold level N_o at the gateway, we are able to control the desired queueing delay at the gateway.

CONCLUSIONS

From the study of the tandem queueing system with threshold control on input arrivals, we notice that the threshold control is an effective mechanism for avoiding congestion.

Further, implementation of such threshold control is simple and easy. The quantitative characteristic system behavior with threshold control presented in the paper should be useful as a design guide for flow control in packet switched computer networks.

REFERENCES

- [1] Price, W. L., "Data Network Simulation Experiments at the National Physical Laboratory 1968-1976," Computer Networks, Vol. 1, No. 4, 1977, pp. 199-210.
- [2] Lam, S. S. and M. Reisner, "Congestion Control of Store-and-Forward Networks by Input Buffer Limits," Proceedings of the NTC Conference, Los Angeles, California, December 1977, 12:1-1- 12:1-6.
- [3] Cerf, V. and B. Kahn, "A Protocol for Packet Network Intercommunications,, IEEE Transactions on Communication, Vol. COM-22, May 1974, pp. 637-648.
- [4] Chu, W. W. and M. Shen, "A Hierarchical Routing and Flow Control Policy for Packet Switched Networks," Proceedings of the Computer Performance Symposium, August 1977, pp. 485-498.

Figure 1. A Tandem Queueing System with Threshold Control on Arrivals

Figure 2a. Mean Time for An Arrival To Go Through Both Queues Vs Carried Load for $N_0 = 0$

Figure 2b. Mean Time For An Arrival To Go Through Both Queues Vs Carried Load for $N_0 = 5$

Figure 3a. Mean Time For An Arrival To Go Through Both Queues Vs Offered Load for $N_0 = 0$

Figure 4. Mean Time For An Arrival To Go Through Both Queues Vs Offered Load for the Uncontrolled Case ($N_0 = \infty$)

Figure 3b. Mean Time For An Arrival To Go Through Both Queues Vs Offered Load For $N_0 = 5$

Figure 5a. Carried Load Of The First Queue For $N_0 = 0$

Figure 5b. Carried Load Of The First Queue Vs Offered Load for $N_0 = 5$

Figure 6a. Mean Time For An Arrival To Go Through The Second Queue Vs External Arrival Rate For $N_0 = 0$

Figure 6b. Mean Time For An Arrival To Go Through The Second Queue Vs External Arrival Rate For $N_0 = 5$

Figure 7. Mean Time For An Arrival To Go Through The Second Queue Vs External Arrival Rate For Uncontrolled Case ($N_0 = \infty$)

Figure 8. Mean Time For An Arrival To Go Through Both Queues Vs N_0 for $\lambda_2 = .4$

Figure 9. Mean Time For An Arrival To Go Through The Second Queue Vs N_0 for $\lambda_1 = .4$

Figure 10. An Example Of A Tandem Queueing System With Flow Control Input Rejection