A PERFORMANCE STUDY OF A TOKEN RING PROTOCOL

S. Guptan
B. Srinivasan
N. Simha

Computer Science Telecommunications Program
University of Missouri-Kansas City

Abstract
This paper proposes a pilot analytical model of a token ring network. The model developed can be used to study the performance characteristics of interest. Past performance studies have made simplifying assumptions to make the model tractable such as station independence, stochastically identical stations, uniform distribution (spacing) of stations over the ring etc. The only assumption made in the development of this model is that all the processes involved are exponential . The model however can be easily extended to handle non-exponential processes. The model is powerful enough to refute the node independence assumption.

INTRODUCTION

The Token ring protocol is one of the commonly used Medium Access Control (MAC) protocols for a LAN where a "token" is used to control access to the transmission medium. When a station that receives a token has messages ready for transmission it immediately sends them onto the ring and then passes the token to a downstream neighbor. The token circulates around the ring among the stations and the station holding the token is allowed to transmit for the duration of the stipulated "token holding time". Each station maintains an independent queue for data. One of the features of a token ring protocol is the bounded delay that it provides in the system thus avoiding "starvation" of any station.

The mathematical model used consists of N queues (not necessarily statistically identical) corresponding to the N stations, each having an independent arrival rate, served by a single server (i.e. the token) in cyclic order. The server goes from queue to queue in some prescribed order, pausing to remove messages for the duration of the transmission time for that host. After the token leaves a queue and before it begins work on the next queue there is a period during which no host can transmit messages corresponding to the token passing time (also called the switchover time). There are three types of switching policies for the server [4]. This model implements the *limited* (to one) service system and again is easily extended to the *exhaustive* service system. In the *limited* (to one) service scheme the token is passed to the other host when the host receiving the token has no message in its buffer or after it completes the transmission of a single message. As far as the queue capacity is concerned, we are interested in the finite buffer model where arrivals to any queue that find it fully occupied are lost (i.e. a loss system). It is an open network in that there are one or more input queues and one or more exit nodes. The only assumption made is that all the processes involved, namely, the arrival and service process at each host and the token passing process are exponentially distributed.

The pilot model implements a two station token ring network. Each of the stations is assumed to have a finite buffer size of five. However the model is representative and can be easily generalized for a N station network and for any value of buffer size. The two stations or nodes are hereinafter referred to as host A and host B. The effect of various loads on this network, where this load is asymmetrically distributed over the two hosts, is studied for effective utilization, marginal distributions, blocking of arrivals, and whether the two nodes can be regarded as independent of each other.

MATHEMATICAL MODEL

The following notations are used in the ensuing discussion.

λ_A = arrival rate at host A
λ_B = arrival rate at host B
μ_A = transmission rate for a single message at host A
μ_B = transmission rate for a single message at host B
ν_{AB} = token passing rate from host A to host B
ν_{BA} = token passing rate from host B to host A
N_A = buffer size at host A
N_B = buffer size at host B

The state space diagram (see last page) gives a clear picture of the transition between the states and their rates. It also forms a base on which the steady state equations can be

deduced. The state space diagram for a buffer size of one at each host has been included to give an idea of the model.

Let the steady state probabilities be denoted by $\pi(n_A, n_B, \tau)$, where $n_A=0..N_A$ and $n_B=0..N_B$ indicate the number of customers at hosts A and B and where $\tau \in \{A,AB,B,BA\}$ indicates the location of the token: if τ=A or B, then the token is currently at A or B, while the host A or B is transmitting a message, while if τ=AB or BA, then the token is between A and B, or between B and A, and none of the hosts can transmit a message. The steady state balance equations are best presented using a vector notation. Introduce the following vectors:

$$\pi(0,0) = [\pi(0,0,AB) \quad \pi(0,0,BA)]$$
$$\pi(n_A,0) = [\pi(n_A,0,A) \quad \pi(n_A,0,AB) \quad \pi(n_A,0,BA)]$$
$$\pi(0,n_B) = [\pi(0,n_B,AB) \quad \pi(0,n_B,B) \quad \pi(0,n_B,BA)]$$
$$\pi(n_A,n_B) = \begin{bmatrix} \pi(n_A,n_B,A) & \pi(n_A,n_B,AB) & \dots \\ \dots & \pi(n_A,n_B,B) & \pi(n_A,n_B,BA) \end{bmatrix}$$

The steady state balance equations are written by considering the cases:
For the states where $n_A = 0$ and $n_B = 0$,

$$\pi(0,0)\left\{\lambda_A I+\lambda_B I+\begin{bmatrix} \nu_{AB} & -\nu_{AB} \\ -\nu_{BA} & \nu_{BA} \end{bmatrix}\right\} = \pi(n_A,0)\begin{bmatrix} \mu_A \\ 0 \\ 0 \end{bmatrix}[1 \ 0] + \pi(0,n_B)\begin{bmatrix} 0 \\ \mu_B \\ 0 \end{bmatrix}[0 \ 1]$$

Notice, that on the left hand side of the equation there are essentially three ways to leave the state: due to an arrival at host A, an arrival at host B or the token arriving (and immediately going away) at either host. The right hand side of the equation gives the two ways in which the state can be entered: due to the departure of the only customer at host A or host B. The token is then passed to the other host on the ring.

$$\pi(n_A,0)\left\{\lambda_A I+\lambda_B I+\begin{bmatrix} \mu_A & 0 & 0 \\ 0 & \nu_{AB} & -\nu_{AB} \\ -\nu_{BA} & 0 & \nu_{BA} \end{bmatrix}\right\} = \pi(n_A,0)\lambda_A + \pi(n_A+1,0)\begin{bmatrix} \mu_A \\ 0 \\ 0 \end{bmatrix}[0 \ 1 \ 0] + \pi(n_A,1)\begin{bmatrix} 0 \\ 0 \\ \mu_B \\ 0 \end{bmatrix}[0 \ 0 \ 1]$$

For the states where $n_A=0$,

$$\pi(0,n_B)\left\{\lambda_A I+\lambda_B I+\begin{bmatrix} \nu_{AB} & -\nu_{AB} & 0 \\ 0 & \mu_B & 0 \\ -\nu_{BA} & 0 & \nu_{BA} \end{bmatrix}\right\} = \pi(n_A,n_B-1)\lambda_B + \pi(1,n_B)\begin{bmatrix} \mu_A \\ 0 \\ 0 \\ 0 \end{bmatrix}[0 \ 1 \ 0] + \pi(0,n_B+1)\begin{bmatrix} 0 \\ \mu_B \\ 0 \end{bmatrix}[0 \ 0 \ 1]$$

Finally, the steady state balance equation for the general state ($1 \le n_A < N_A$ and $1 \le n_B < N_B$) is given by

$$\pi(n_A,n_B)\left\{\lambda_A I+\lambda_B I+\begin{bmatrix} \mu_A & 0 & 0 & 0 \\ 0 & \nu_{AB} & -\nu_{AB} & 0 \\ 0 & 0 & \mu_B & 0 \\ -\nu_{BA} & 0 & 0 & \nu_{BA} \end{bmatrix}\right\} = \pi(n_A-1,n_B)\lambda_A + \pi(n_A+1,n_B)\begin{bmatrix} \mu_A \\ 0 \\ 0 \\ 0 \end{bmatrix}[0 \ 1 \ 0 \ 0] + \pi(n_A,n_B+1)\begin{bmatrix} 0 \\ 0 \\ \mu_B \\ 0 \end{bmatrix}[0 \ 0 \ 0 \ 1] + \pi(n_A,n_B-1)\lambda_B$$

The model can be modified with ease to handle the exhaustive service protocol by changing the position of 1 in the row vectors on the right hand side of the equations from [0 1 0 0] and [0 0 0 1] to [1 0 0 0] and [0 0 1 0] respectively. It can be extended to handle more than two hosts by increasing the state space which is reflected in these equations with an increase in the dimension of the matrices and vectors involved. Only the computational complexity of the problem is affected. The infinitesimal rate matrix or the Q* matrix as it is popularly called was generated [1]. The steady state probabilities were obtained by solving the system of linear equations $\pi Q^* = 0$.

RESULTS AND DISCUSSION

The following characteristics were studied and the graphs obtained have been included.

Mean Effective Utilization:

This is defined as the ratio of the useful time to the total time where the useful time is the time the network is actually involved in user data transmission and the total

time includes the overheads of transmission. It is essentially the sum of probabilities of those states which are effective in transmission by a host. Eg. the utilization of host A is calculated as the sum of the probabilities of those states in which host A is involved in transmitting a message and the total utilization is obtained from the sum of the probabilities of those states in which either of the hosts is involved in data transmission.

The observation we can make from figures 1a-1 and 1a-2, which show the variation of mean effective utilization with arrival rate, is that the utilization is low when the arrival rate is low since the probability of there being zero messages in the buffer is quite high and the token is passed back and forth between the two nodes. As the arrival rate increases, the queue builds up, the utilization increases and saturates once the buffer remains full. Also note that the utilization of host A is low because of its higher service rate.

In fig. 1b it is seen that as the service rate of host B is increased the utilization of host B falls off while that of host A remains constant. This is because host B is now dispatching its messages faster and returning the token back to host A and so increasing the probability of its buffer being empty when the token comes back to it.

As the token passing rate from host A to host B is increased, the utilization of host B increases initially because the probability of messages having arrived in its buffer increases, fig 1c. This increase is rapid in the beginning (because the probability of there being zero messages in the buffer at host B is drastically reduced) but then saturates quickly (because there will always be more than one message in the buffer at host B).

Mean rate of blocked messages:

For a particular host, a message at a host will be "blocked and cleared" in this loss system if it arrives to see the buffer full. Thus the probability of being turned away is nothing but the probability of finding the host with its buffer full. So summing up the probabilities of those states in which the buffer is full for host 'i' and multiplying this value by the arrival rate gives the mean rate at which messages are turned away from that host. Its dependence on arrival rate has been summarized in fig. 2. As is to be expected with higher arrival rates, the mean number of outstanding requests increases and hence the probability of an arrival finding the buffer full increases and hence the probability of it being turned away. Thus a penalty incurred when attempting to run the system near peak utilization is that more and more customers find themselves being blocked.

Actual distribution of messages at host A:

To find the probability mass function of the messages at host A, the probabilities of the states containing 'i' messages were summed up to give the probability that the host A had 'i' messages in its buffer where 'i' ranged from 1 to 5.

The pmf of messages in the buffer at host A as the arrival rate was varied is shown in fig. 3. It is seen that for low arrival rates the probability of having fewer messages in the buffer is high and as the arrival rate increases this probability decreases and the probability of finding more number of messages in the buffer increases. Thus this graph could be used to determine the acceptable values of the arrival rates at the two hosts for a prescribed delay and desired utilization. A good value to choose for eg. would be an arrival rate of lambda_1 = lambda_2 = 4 where the probability of finding 1 message in the buffer is the highest. Thus this value would give a reasonable delay with sufficiently high value of utilization. It is also seen that for some values of arrival rates (eg. lambda_1 = lambda_2 = 8 messages/sec) with a high probability the buffer is either always full or always empty. So this is a value of arrival rate that would typically be avoided.

Outstanding requests to be serviced:

To calculate the outstanding requests those states were considered in which a host was not transmitting but had outstanding requests (messages) in the buffer. From these probabilities the mean number of outstanding requests for a host was calculated as $\Sigma\ i\ \pi(i)$, where $\pi(i)$ is the probability that 'i' messages are waiting with i ranging from 1 to 5 (maximum buffer size). A similar procedure was used to calculate the total mean number of outstanding requests in the system.

The outstanding requests at each host and the system as a whole as the arrival rate was varied is shown in fig. 4. This graph is similar in structure to that of fig. 1a and a connection can immediately made between the two. As in that case, as the arrival rate increases, the probability of finding one or more messages in the buffer increases upto the point when the buffer begins to remain full and the graph levels out.

Independence of Nodes:

The independence of nodes was investigated by calculating the marginal distributions and it was observed that the nodes are not independent of each other. In particular, with $\pi_A(n_A)$ the steady state marginal distribution at host A, and $\pi_B(n_B)$ the steady state marginal distribution at host B, we have observed that

$$\pi(n_A,n_B) \neq \pi_A(n_A)\pi_B(n_B),$$

eg: 1. $\pi(0,0) = 0.6003$

$\pi_A(0) * \pi_B(0) = 0.0865$

2. $\pi(1,1) = 0.0469$

$\pi_A(1) * \pi_B(1) = 0.0226$

thus showing that the hosts can not be considered independent. This means that many performance models currently in the literature must be questioned on their model validation part.

CONCLUSIONS

The study conducted reveals several interesting features of the token ring network. The most important conclusion that can be drawn from the above study is that the nodes are not independent of each other, an assumption made by traditional performance models. Their results should hence be carefully interpreted. This has an immediate consequence for node deletion/addition in an actual network.

In this paper, an exact analytical model was developed to represent the token ring protocol. Also, using a linear algebraic approach to queueing theory it is possible to extend the model to handle non-exponential (eg.bursty) arrivals, service times and token passing times. Future studies will address this aspect.

ACKNOWLEDGEMENTS

We wish to express our gratitude to Dr. Appie van de Liefvoort for his innovative introduction to the subject and patiently guiding us through this paper. We would also like to thank Dr. Deepankar Medhi for his encouragement in our work.

REFERENCES

[1] L.Kleinrock, *Queueing Systems Vol I - Theory* Wiley Interscience, New York, 1975.

[2] L.Kleinrock, *Queueing Systems Vol II - Computer Applications* Wiley Interscience, New York, 1976.

[3] M.Schwartz, *Telecommunication Networks: Protocols, Modelling and Analysis* Addison-Wesley, Reading, MA. 1987.

[4] H.Takagi, *Application of polling models to computer networks* Computer Networks and ISDN systems 22 (1991) 193-211.

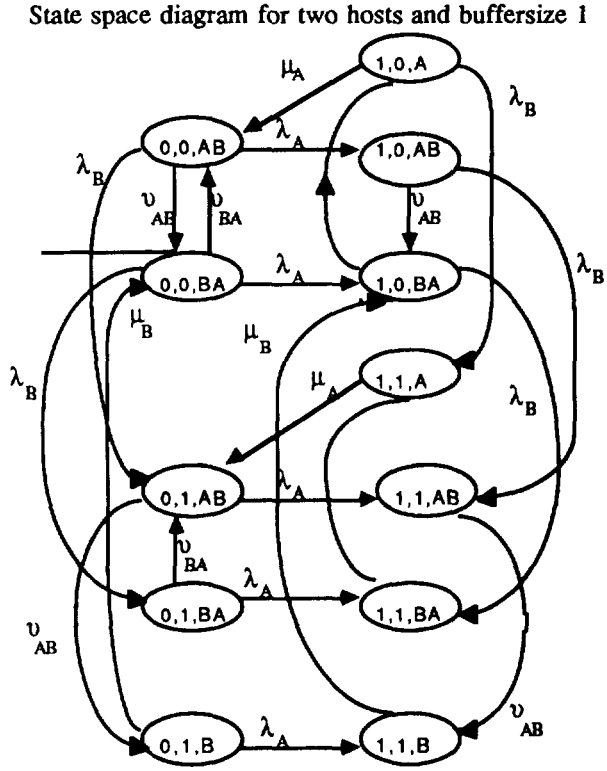
[5] W. Bux, *Local-area Subnetworks: A performance Comparison* IEEE Trans. Commun. 29 (10) (1981) 1465-1473.

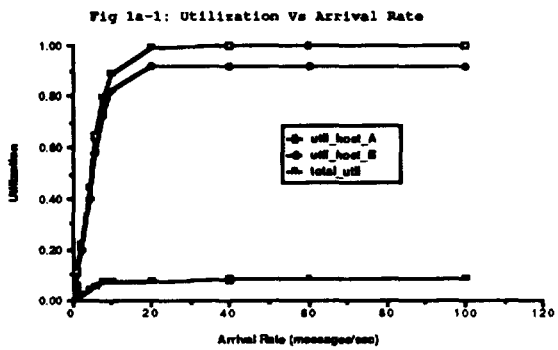
[6] W. Bux, *Token-Ring Local-area Networks and their Performance* Proc. IEEE 77 (2) (1989) 238-256.

[7] J.L. Hammond and P.J.P. O'Reilly, *Performance Analysis of Local Computer Networks* Addison-Wesley, Reading, MA. 1986.

[8] L.Lipsky, *Queueing Theory - A Linear algebraic approach* Macmillan, 1991.

State space diagram for two hosts and buffersize 1





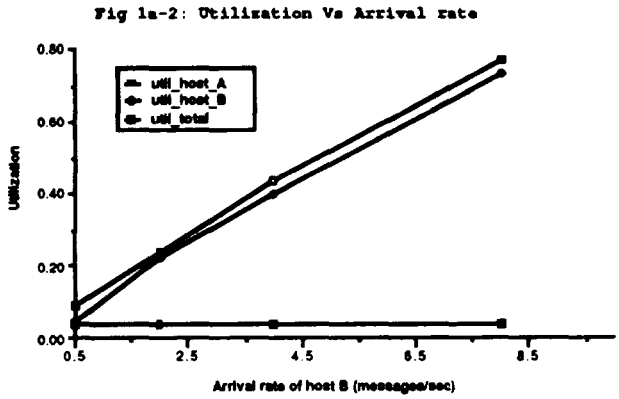


Fig 1b: Utilization Vs Service rate

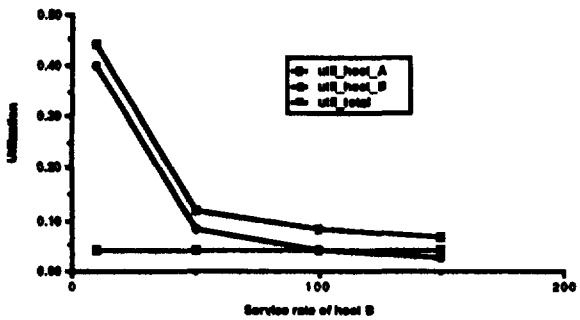


Fig 1c: Utilization Vs Token passing rate

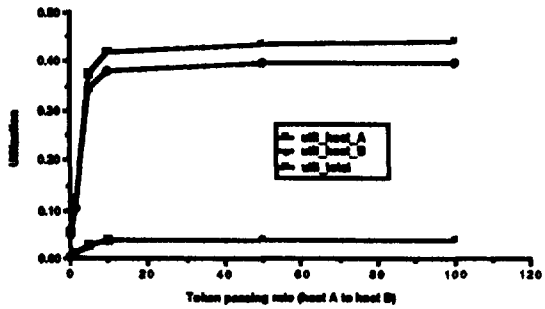


Fig 2:

Mean rejection rate Vs Arrival rate

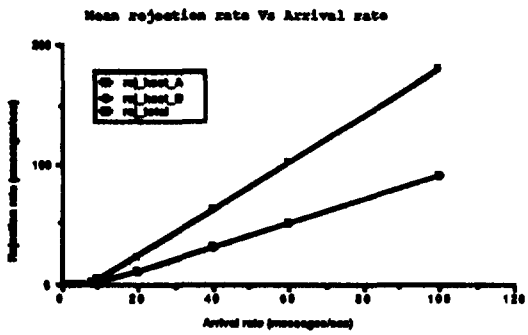


Fig 3: Actual distribution of messages at host A

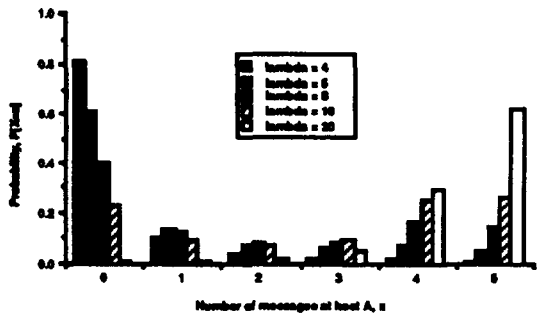


Fig 4: Outstanding requests Vs Arrival rate