Analysis and simulation of the priority scheme in token bus protocols

Seung Ho Hong and Asok Ray* present the results of a statistical analysis of the priority scheme of a token passing protocol

Token bus protocols have been widely accepted for medium access control (MAC) in computer networks. The priority scheme in token bus protocols offers multiple levels of privilege of medium access to heterogeneous traffic. Performance evaluation of the priority scheme is essential for design and operation of real-time networks such as those used in aircraft and factory communications. This paper presents the results of statistical analysis of the network-induced delays at all priority levels and their verification by simulation experiments. Use of the analytical model is illustrated for initial design of the network as well as for optimization of its performance.

Keywords: protocols, token bus, medium access control, performance evaluation, real-time networks

Computer networks connect the individual subsystems of large-scale systems, such as computer integrated manufacturing (CIM)[1, 2] and advanced aircraft[3], so that the information generated from a variety of distributed system components can be freely exchanged. The token passing bus (TPB) protocol[4, 5] is widely accepted for medium access control (MAC) in computer networking where the stations are served in a cyclic order by explicit token circulation. For example, the manufacturing automation protocol (MAP)[6] which has been largely accepted as a standard for factory communication networks by the international community is based on the IEEE 802.4 token passing bus protocol[4]. In order to handle different types of data packets, TPB employs a priority scheme to provide multiple levels of privilege of accessing the medium so that the data latency of each priority message remains below its specified bound.

Several simulation studies have been reported for performance evaluation of the IEEE 802.4 protocol[7-10], but analytical modelling of the priority scheme of token bus protocols has been so far very limited[11]. Since simulation is often time-consuming and costly, and may not provide an exhaustive means for arriving at a conclusion, it is desirable to have an analytical model which saves the cost and effort of numerous simulation runs and data assimilation, generates a direct solution and provides a better insight into the problem. Formulation of such analytical models often proves to be difficult unless certain simplifying assumptions are made.

Jayasumana *et al.*[8-10] formulated a model for evaluating the throughput at each priority level of IEEE 802.4 with respect to variations in offered traffic under fixed timer assignments, but the network-induced delay which is critical for real-time operations was not modelled. Dykeman and Bux[12] reported a model for computing throughput in the FDDI token ring which provides for up to eight different levels of priority. Apparently no significant work has been reported for analytical modelling of network-induced delays at different priority levels of a token bus protocol.

An analytical model of the priority scheme has been reported in Reference 11 for the case of *single service systems* where a station is allowed to transmit only one

Automation Systems Section, Electronics and Telecommunications Research Institute, Daedog, South Korea
*Department of Mechanical Engineering, The Pennsylvania State University, University Park, PA 16802, USA

0140-3664/90/030157-08 © 1990 Butterworth & Co (Publishers) Ltd

message frame each time it captures the token. (This is the usual practice in many applications like avionic communications networks. If a station serves several devices, the data from different devices are concatenated into a single message frame which is broadcast or multicast for the respective destination stations.) From this analytical model, one can determine the relationship between the network parameters (i.e. number of stations, message inter-arrival time, message length and token rotation timer) and network performance (i.e. average queueing delay, average data latency and average queue length) for each priority level. The major assumption in formulating the above analytical model[11] is independence of message waiting processes between individual queues[13, 14] in order to keep the analysis mathematically tractable. Since this model is not an exact representation of the protocol operations, its accuracy needs to be examined under different operating conditions of the network.

The objective is to verify the analytical model by simulation experiments for different levels of offered traffic and settings of the token rotation timer. To this effect, a discrete-event simulation model that is not subjected to the approximations of the analytical model has been formulated, and then the results of analytical and simulation models are compared. The use of the analytical model for network design is exemplified for evaluating the optimal settings of the token rotation timer under heterogeneous network traffic. This is of practical significance to network design for diverse applications. For example, in CIM, the traffic may consist of real-time sensor and actuator signals and non-real-time CAD file transfer data within the same network channel[2]. The token rotation timers in the priority scheme of the token bus protocol (e.g. IEEE 802.4[4]) need to be set such that the probability of timely arrival of high-priority data at the destination is enhanced with the best possible throughput of the low-priority data.

This paper presents the analysis of the priority scheme of a token passing protocol under heterogeneous traffic, typical of aircraft[3] and manufacturing system[1, 2] environments. The major objectives of this paper are to:

- verify the accuracy of the analytical model by simulation experiments and assess the range of operating conditions for which this model can be used as a tool for network design;
- demonstrate the use of the analytical model for optimal setting of the protocol timers for a given network traffic distribution.

DESCRIPTION OF THE PRIORITY SCHEME MODEL

The priority scheme under consideration is similar to that specified by IEEE 802.4[4] and SAE linear token bus[5] protocols. The priority of each message is assigned when the logical link control (LLC) sublayer requests the medium access control (MAC) sublayer to send a data frame. In the SAE token bus protocol, the priority classes are named 0, 1, 2 and 3, with 0 corresponding to the highest priority class and 3 to lowest. In the IEEE 802.4 token bus, the priority classes are called 0, 2, 4 and 6, with 6 corresponding to the highest priority, and 0 to the lowest. However, the basic priority mechanism is identical for both token bus protocols.

Although the SAE and IEEE token bus protocols have four priority levels, the number of priority levels can be increased by setting different values of token rotation timer (TRT). The model developed in this paper is assumed to have $(K + 1)$ priority levels, i.e. each station has $(K + 1)$ separate priority queues, one for each priority level. Accordingly the priority classes are designated as 0, 1, 2, . . . , K with 0 corresponding to the highest and K to the lowest. Each priority queue acts as a virtual substation such that the token is passed internally from the highest priority queue to the lowest before being transferred to the next station in the logical ring.

For the highest priority message, i.e. priority 0, the opportunity to transmit is given whenever the priority 0 queue captures the token. For the lower priority messages, i.e. priorities 1 to K, access to the medium is regulated by the token rotation timer, TRT_i, of the corresponding priority i class. TRT_i is reset and restarted whenever a priority i queue captures the token. For the priority i message, $i = 1$ to K, an opportunity to transmit is given if the corresponding TRT_i is not expired when the token arrives at the priority i queue after circulating through the logical ring of all active stations in the network. The higher priority queues must have a larger value of TRT so that the probability of TRT expiration is smaller, and consequently the probability of message transmission is larger.

Large-scale systems are decomposed into several single-function subsystems. The data generated from different devices at a station are concatenated as a single message and these messages are usually transmitted one by one. Therefore, the analytical model is developed on the assumption of single-service system where each queue is allowed to transmit only one message at a time. The messages generated from a single-function subsystem are packetized to a fixed length. Therefore, it is assumed that the message generated from a priority i queue has a constant length L_i. Since the traffic pattern in the network system is bursty due to a large number of subscribers, it is assumed that message arrival process is Poisson. Since the network is designed to operate within its bandwidth capacity, it is further assumed that the probability of message rejection due to queue saturation is zero.

In this context, the queue capacity is assumed to be infinite. Also, it is assumed that messages belonging to the same priority class have the identical (average) message arrival rate and message length. Without loss of generality, it is convenient to consider the individual priority classes in each station as separate substations where the token is passed from one substation to another. Thus, the model is a system of multiple priority queues attended by a single server in a cyclic order. The analytical model is formulated on the fundamental principles of statistics as outlined in Appendix A. The detailed derivations are given in Reference 11.

DEVELOPMENT OF A DISCRETE-EVENT SIMULATION MODEL

A network protocol can be viewed as an *event-driven system*. Therefore, a *discrete-event simulation* technique has been adopted to model the operations of the priority scheme of a token bus protocol. The discrete-event simulation (DES) model follows the structure of the IEEE 802.4[4] and SAE[5] token bus protocols which have four levels of priority. The priority classes are designated as 0, 1, 2 and 3, with 0 corresponding to the highest and 3 to the lowest.

The token bus protocol consists of several interacting components where events occur simultaneously. The concurrence of different events in the protocol operations is first represented by a timed Petri net (TPN) model[15, 16] as the first step to the development of the DES model. The use of Petri net models has been reported by several investigators for the description and analysis of communication protocols in packet-switched and CSMA/CD networks[17], but it has not apparently been applied to model the priority scheme in token bus protocols. The TPN model is described in detail in Reference 11.

The DES model is developed on the basis of the TPN model of the priority scheme described above. The entities in the DES model are messages that are characterized by the following attributes:

- time of generation, i.e. the instant a message arrives at the transmitter queue;
- message length (or, message transmission time);
- source queue, i.e. the queue from which a message is generated;
- destination queue, i.e. the receiver queue at any station other than the source station;
- message priority, i.e. the priority of the queue from which a message is generated.

The DES model consists of two submodels, namely, message generation and protocol operation. The message generation submodel handles arrival of new messages at the network, and the protocol operation submodel describes the activities that occur within the network without imposing the approximations that were introduced in the analytical model.

The DES model has a modular structure and consists of several events. Following the TPN model, these events are described below.

- *Initialization*. This event reads the simulation and network parameters for a given traffic condition.
- *Power_up*. In this event, the first message arrival in each queue is scheduled.
- *Message_generation*. As new messages are created to be put into queues, this event reschedules next message generation according to the given message inter-arrival time.
- *Token_pass*. In this event, the token is passed to the successor in the logical ring. This event schedules the token_receive event.
- *Token_receive*. This event schedules the end of response time.
- *End_response_time*. At the end of response time, transmission of a priority 0 message is possible. If there is any waiting priority 0 message at this station, the waiting message is immediately transmitted by calling the *message_head_transmission* event. If there is no waiting priority 0 message, the priority 0 message cannot be transmitted and the service of priority 0 queue is completed. The token is passed to the priority 1 queue.
- *TRT_reset/restart*. This event schedules the end of TRT event. TRT is reset and restarted (i.e., a P-token is put in place *TRT$_i$_reset*). As TRT is reset, the scheduled *TRT_end* event is eliminated from the event calendar of the DES model.
- *TRT_end*. This event indicates expiration of TRT.
- *Message_head_transmission*. This event transmits the message header, and schedules its reception. Transmission of the tail of message is also scheduled at this moment to take into account the message transmission time.
- *Message_tail_transmission*. This event transmits tail of the message, and schedules its reception. The token is then passed to the next priority queue if $i = 0, 1, 2$ or to the successor station if $i = 3$.
- *Message_reception*. After a message is completely received at the destination station, the statistical data such as the expected and standard deviation of the queueing delay, and throughput for each priority class are collected.

SIMAN[18] was selected for the DES model in view of program flexibility and portability, modularity and structured programming, built-in data analysis and real-time event scheduling capacities, and verification and run-time debugging. Further details of SIMAN and its comparison with other simulation language are provided in Reference 19.

VERIFICATION OF THE ANALYTICAL MODEL BY SIMULATION EXPERIMENTS

The results of the analytical model are compared with those of simulation experiments. The objective is to determine the range of validity of the analytical model, and also to establish credibility of the simulation model. Performance of the priority scheme is expressed in terms of data latency at different offered traffic G (see Definitions A6 and A7 in Appendix A).

Three levels of offered traffic, low ($G_{tot} = 0.2$), medium ($G_{tot} = 0.5$) and high ($G_{tot} = 0.8$), are considered. Offered traffic for individual priority i class is taken to be identical, i.e., $G_1 = G_2 = G_3 = G_4 = G_{tot}/4$. (Both analytical and simulation models are capable of handling asymmetric traffic, i.e., unequal G_is.) The message arrival process is assumed to be Poisson and the message length for a given priority class is constant.

Traffic condition

1. Constant message transmission time

   $L_0 = 0.1$ ms, $L_1 = 0.2$ ms, $L_2 = 0.3$ ms, $L_3 = 0.4$ ms
2. Number of queues

   $N_0 = N_1 = N_2 = N_3 = 4$
3. Average message interarrival time ($\tau = 1/\lambda$)

   $G_{tot} = 0.2$: $\tau_0 = 8$, $\tau_1 = 16$, $\tau_2 = 24$, $\tau_3 = 32$ (ms)
   $G_{tot} = 0.5$: $\tau_0 = 3.2$, $\tau_1 = 6.4$, $\tau_2 = 9.6$, $\tau_3 = 12.8$ (ms)
   $G_{tot} = 0.8$: $\tau_0 = 2$, $\tau_1 = 4$, $\tau_2 = 6$, $\tau_3 = 8$ (ms)
4. Total ring latency due to station response time at all stations during one token circulation time, $R = 0.006$ ms
5. $TRT_2 = 2TRT_3$, $TRT_1 = 3TRT_3$

Ten different experiments were performed with different seed numbers of the random number generator, and 95% confidence interval (i.e., the probability that the exact result lies in this interval is 0.95) for each data was obtained.

Data latencies obtained from both simulation and analytical models at $G_{tot} = 0.2$, 0.5 and 0.8 are shown in

Figures 1, 2 and 3, respectively. Analytical results are represented by solid, long dashed, dashed and dotted lines, and simulation results as circle (○), diamond (◇), triangle (△) and square (□) for priorities 0, 1, 2 and 3, respectively. A pair of symbols is used to represent 95% confidence intervals in the simulation results.

The data latencies of priority 0, 1, 2 and 3 messages in Figures 1 to 3 are derived under the constraints of $TRT_2 = 2TRT_3$ and $TRT_1 = 3TRT_3$. At $G_{tot} = 0.2$ in Figure 1, data latencies for all priority classes are practically independent of the TRT values in contrast to the corresponding results for $G_{tot} = 0.5$ and $G_{tot} = 0.8$ in Figures 2 and 3, respectively. The rationale is that, at low offered traffic, most of the queues are empty when the token arrives. Since token circulation time is relatively short at low traffic, TRTs are usually in the running state, i.e., not expired, when the token arrives. Therefore, the unexpired TRTs have no significant bearing on the data latency of all priority classes at low traffic.

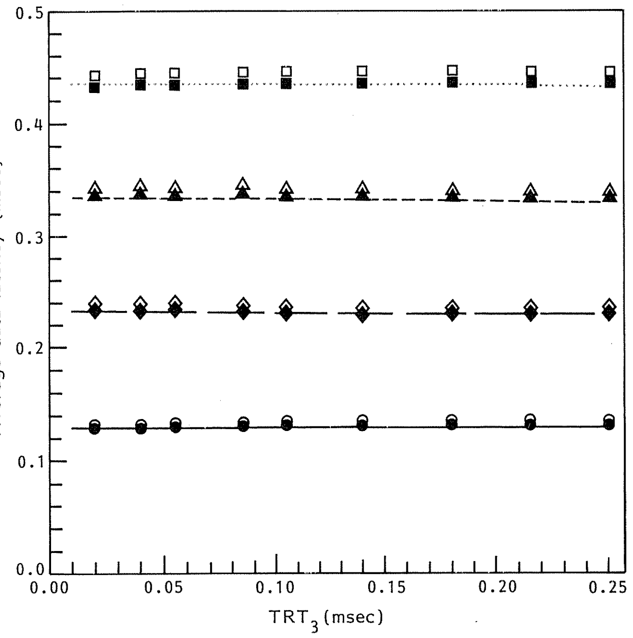


Figure 1. Average data latency for total offered traffic $G_{tot} = 0.2$

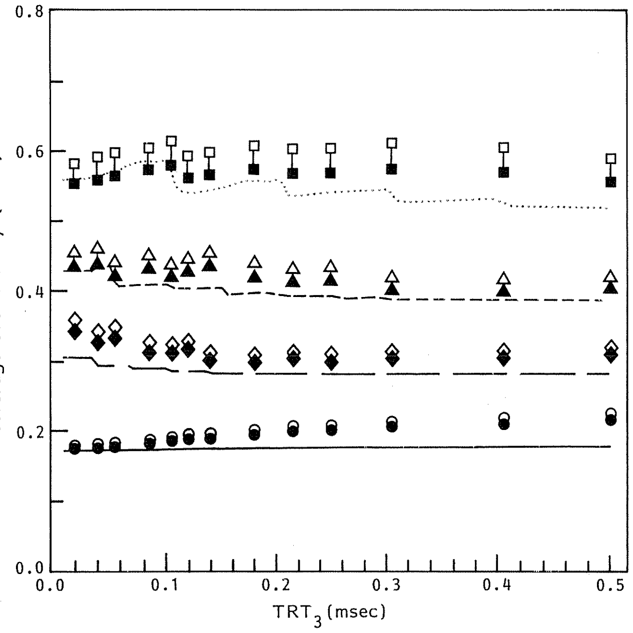


Figure 2. Average data latency for total offered traffic $G_{tot} = 0.5$

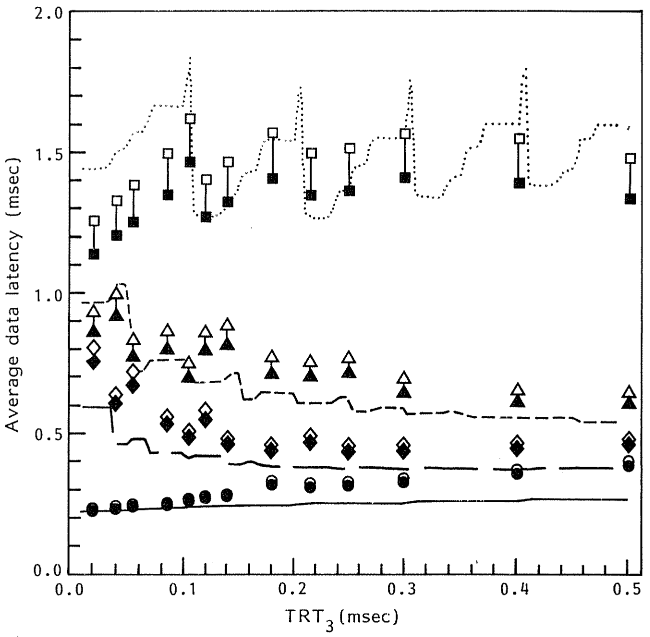


Figure 3. Average data latency for total offered traffic $G_{tot} = 0.8$

For $G_{tot} = 0.5$ and 0.8 in Figures 2 and 3, respectively, the dependence of data latency on the respective TRT values exhibits a close agreement between simulation and analytical results. When TRT_i ($i = 1, 2, 3$) is set to small values, e.g., $TRT_3 = 10\,\mu s$, data latency for priority 0 is minimum. This is because small TRTs cause frequent expiration of the priority 1, 2 and 3 messages, and the channel capacity is practically dedicated to the priority 0 class, and the lower priority queues are allowed to transmit using the leftover channel capacity. Data latency for priority 0 increases as TRT_i ($i = 1, 2, 3$) is increased because the transmission of priority i class messages are less likely to be deferred due to TRT expiration. Thus, more channel capacity is assigned to lower priority messages. This causes a modest increase in data latency of priority 0 messages.

A qualitative assessment of Figures 1 to 3 is that accuracy of the analytical model is good for low (e.g. $G_{tot} = 0.2$) to medium (e.g. $G_{tot} = 0.5$) traffic. However, when the traffic is high (e.g., $G_{tot} = 0.8$), the analytical model becomes less accurate. This is because, at high offered traffic, more messages are built up in the queue and the state of a queue is more likely to be dependent upon the states of the other queues, and the independence assumption in the analytical model generates larger errors. Also, the TRT_i expiration process at a queue is more dependent upon the message waiting process at the same queue. However, the characteristics of data latency are adequately represented by the analytical model at all traffic.

OPTIMIZATION OF PROTOCOL PARAMETERS FOR NETWORK DESIGN

This section illustrates how the analytical model can be used for optimal setting of the token rotation timers in the token bus priority scheme. Criteria for parameter optimization in network design are described in the previous section.

Criteria for network design in the initial phase

Data latency is a function of network traffic and the settings of the token rotation timers (TRT). For network design, the message length and inter-arrival time at each station are usually determined *a priori* according to their functional characteristics whereas the process of medium access is controlled by adjusting the TRT settings.

Optimization of TRT settings by simulation alone often proves to be time-consuming and very cumbersome especially if a large number of design options are available. The analytical model[11] can be directly applied to determine sub-optimal values of TRT which can be fine-tuned by simulation and perturbation analysis[20, 21].

Often a network has to be designed such that data latency of high priority messages at individual stations is bounded with a given confidence[22, 23]. Intuitively, the moments of data latency at a given station are directly dependent on the variance of effective service time (see Definition A2 in Appendix A). A design criterion is introduced such that the sum of the variances of effective service times of all the priority level messages is minimized. On this basis, the performance index for optimization of TRT parameters is proposed as:

$$\text{Minimize } J := \sum_{j=0}^{M} [\sigma_j'^2\,(\mathbf{TRT}) + \sigma_j''^2\,(\mathbf{TRT})] \text{ under the}$$

$$\text{constraint } D_i(\mathbf{TRT}) \leqslant (w_i/c_i)\delta_i,\ i = 0, 1, \ldots, K$$

where

$\mathbf{TRT}$: = $[\mathrm{TRT}_1\ \mathrm{TRT}_2 \ldots \mathrm{TRT}_K]^T$, K corresponds to the lowest priority (e.g. $K = 3$ in the simulation example),
M is the lowest priority level for which delays are considered to be significant ($M \leqslant K$),
$\sigma_j'^2\,(\mathbf{TRT})$ and $\sigma_j''^2\,(\mathbf{TRT})$ are variances of the conditional effective service times T_j' and T_j'', respectively, as defined in Appendix A and derived in Reference 11,
$D_i(\mathbf{TRT})$:= expected value of data latency of the priority i messages,
δ_i is the specified bound for data latency for the priority i class,
w_i is the design safety factor for the priority i class,
c_i is the compensation coefficient for the priority i class to allow for inaccuracies in the analytical model.

Results in a previous section show that the analytical model generally underestimates data latency. Therefore, it is desirable to set w_i/c_i smaller than one to take into account a safety margin of network design and underestimation of the analytical model.

Since the objective and constraint equations are formulated as nonlinear functions, a nonlinear programming technique would be a natural choice for solving the optimization problem where the constraint functions (i.e. expected values of data latencies) are not continuously differentiable with respect to the design variables (i.e. $\mathrm{TRT}_i, i = 1, \ldots, K$). This problem can be circumvented by first transforming the constrained optimization problem to an unconstrained optimization problem, and then using the Hook-Jeeve method[24] that does not require computation of derivatives. A detailed solution approach for the optimization problem is described in Appendix B of Reference 11.

An example of network design

As an example of network design, consider communication services to two types of subscribers, i.e., real-time and non-real-time data. Although occasional losses of real-time data packets can be tolerated, network-induced delays are critical for real-time operations and must not exceed the allowable bounds within a specified confidence interval. On the other hand, non-real-time data do not have to be processed within specified time-constraints but need the assurance of accurate delivery. Thus, the real-time data should receive preferential treatment at the expense of the increased data latency of non-real-time data. Figure 2 shows that data latencies are moderately sensitive to changes in TRT settings at $G_{\mathrm{tot}} = 0.5$. The usual practice is to design the network under a nominal condition of medium traffic. As an example of protocol parameter optimization, the case of medium traffic ($G_{tot} \approx 0.5$) is presented below.

The network consists of four stations where each station has four priority queues, 0, 1, 2, and 3, with 0 corresponding to the highest priority and 3 to the lowest. The priorities 0 and 1 are assigned to accommodate delay-sensitive data, i.e. $M = 1$, and priorities 2 and 3 to non-delay-sensitive data. Lengths of the priority 0, 1, 2 and 3 messages are packetized by 0.05, 0.15, 0.32 and 0.5 ms, respectively, and the corresponding average message inter-arrival times are 1.86, 6.0, 9.0 and 15 ms, and bounds for average data latencies are taken to be 0.25, 0.47, 0.9 and 1.6 ms. Design parameters w_i/c_i for all priority classes are set to 0.5.

The objective is to find the optimal settings of TRT_i, $i = 1, 2$ and 3, which minimize the sum of the variances of data latency for the delay-sensitive data, i.e. priorities 0 and 1 in this example, such that the average data latency for each of the four priority classes is bounded by the values given above. The optimal settings, TRT_1^*, TRT_2^*, TRT_3^* were determined to be 0.2 ms, 0.09 ms and 0.05 ms, respectively.

Figure 4 exhibits variations in the performance index J

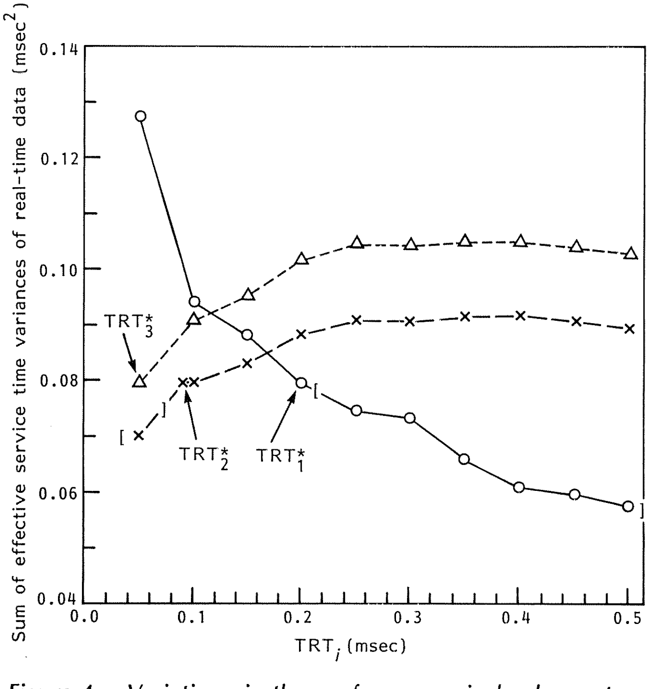


Figure 4. Variations in the performance index by perturbations in TRT_i

when the design variables, TRT_i, $i = 1, 2, 3$, are individually perturbed from the minimal point. The solid line with circle (○) represents the variation in J when TRT_1 is perturbed in several steps from 0.05 ms to 0.5 ms while TRT_2 and TRT_3 are held constant at their optimal values of 0.09 ms and 0.05 ms, respectively. Similarly, the dashed line with cross (×) and the dotted line with triangle (△) show variations in J when TRT_2 and TRT_3 are individually perturbed holding the other variables constant at their optimal settings. The intervals enclosed by square brackets in Figure 4 indicate the infeasible regions where the inequality constraints of average data latency are violated.

TRT_i^*, $i = 1, 2, 3$, obtained from the design procedure, may not be truly optimal settings of TRT_i because the analytical model is inexact but the analytically determined TRT_i^*s are expected to be close to the actual optimal values even at high offered traffic. Fine tuning can be accomplished by perturbing the individual TRT_i^*s in the simulation model where the technique of perturbation analysis in discrete-event dynamical systems would allow evaluation of the nominal and perturbed paths in a single simulation run[20, 21].

SUMMARY AND CONCLUSIONS

Performance analysis of the priority scheme in token bus protocols (e.g. IEEE 802.4 in manufacturing automation protocol (MAP) in factory environments and SAE linear token bus for aircraft control systems) is essential for design of networks that must handle heterogeneous traffic. Selection of priority timer settings solely by discrete-event simulation suffers from extensive data assimilation and computational burden which often prove to be costly and time-consuming. An analytical model serves to reduce the cost and efforts of numerous simulation runs and provides a direct insight into the problem of network design.

An analytical model has been developed to statistically evaluate the performance of the priority scheme in token bus protocols. This analytical model determines the relationship between the network parameters (i.e. number of stations, message inter-arrival time, message length and token rotation timer) and network performance (i.e. average queueing delay, average data latency and average queue length) for each priority level.

The basic network-operating assumptions used in this paper include Poisson distribution of the message arrival process, constant message length, infinite queue capacity and single-service system (i.e. one message transmission at a time) for each priority class. The restriction of constant message length can be lifted if the statistics of the message length are available and the message arrival process is independent of the message transmission time, i.e. message frame length. Because of the mathematical intractability, the processes at each queue and also the message waiting and token rotation timer expiration processes at the same queue are assumed to be independent.

The analytical model is verified by simulation experiments and pertinent results are presented where the discrete-event simulation program is based on a timed Petri net model of the token bus protocol and are not subjected to the above assumptions. Finally, use of the analytical model is demonstrated for optimal setting of the token rotation timers in the token bus priority scheme.

On the basis of comparison of the analytical and simulation results, the following conclusions are derived.

- The effects of the priority scheme become more dominant as the network traffic is increased.
- The analytical results closely agree with those of simulation experiments at low and medium traffic and are less accurate at high traffic. This inaccuracy can be attributed to the above independence assumptions which generate larger errors as the offered traffic is increased.

The above inaccuracy at high traffic is not apparently a problem for use of the analytical model as a design tool for network design because the usual practice is not to load the network to a high level to ensure stable operations and small delays.

APPENDIX A: DESCRIPTION OF THE ANALYTICAL MODEL OF THE PRIORITY SCHEME

Performance of the priority scheme is dependent on random variables, such as token circulation time and effective service time, that are directly related to network traffic. Pertinent variables, used in the analytical model, are defined below.

Definition A.1. Token circulation time T_r is the time interval between two consecutive instants at which a queue captures the token. ■

Remark A.1. T_r is a random variable and its expected value is identical relative to any queue. ■

Definition A.2. Effective service time T_j at a priority j queue is the time interval between two consecutive instants at which the priority j queue has an opportunity to transmit a message. ■

Remark A.2. For priority 0 queue, $T_0 = T_r$ because any waiting priority 0 message is transmitted whenever it captures the token. For priority 1 to K queues, an opportunity to transmit a waiting message occurs when the queue captures the token and the corresponding TRT is not expired. ■

Remark A.3. T_j, $j = 1, \ldots, K$, is a random variable and its expected value is identical relative to any priority j queue. ■

In a cyclic queueing system, several queues share a single server, e.g. the token, to transmit their messages. Therefore, the state of a queue is influenced by the states of the other queues. However, it is very difficult, if not impossible, to mathematically describe an exact relationship of the processes among the queues in a single-server network system[14]. Because of the mathematical intractability of most cyclic queueing problems, several approximate methods were suggested[13, 14]. These methods rely on certain simplifying assumptions such as the *independence assumption* under which the processes within a particular queue are considered to be independent of the processes at the other queues. The analytical model developed in this paper is based on the independence assumption.

Token circulation time T_r depends on whether a message at a queue is served or not during this circulation. From this observation, Kuehn[14] considered conditional circulation times in the non-symmetric single service system. Since Kuehn did not consider the priority scheme, i.e., every queue is allowed to transmit the waiting message whenever it captures the token, any queue has exactly one opportunity to transmit during one T_r.

T_r as observed at a queue, belonging to any priority, is classified under the two conditions of whether the queue transmits or not.

Definition A.3. The token circulation time T_{rj} for a given priority j queue is denoted as T'_{rj} if a message is not transmitted; otherwise, T_{rj} is denoted as T''_{rj}. ■

Similarly T_j is also classified under two conditions.

Definition A.4. The effective service time T_j for a given priority j queue is denoted as T'_j if a message is not transmitted; otherwise, T_j is denoted as T''_j. ■

Remark A.4. The concept of conditional token circulation time reduces the effects of independence assumption which is proposed by Hashida and Ohara[14]. Under this concept, the expected waiting time of the queue j in the nonsymmetric single service system without a priority scheme was determined by Kuehn[13] as a function of the first two moments of T'_{rj} and T''_{rj}. ■

In the priority scheme, priority j queue has exactly one opportunity to transmit during one T_j. By replacing the conditional token circulation time in Kuehn's formulation into the conditional effective service time for priority j, the queueing delay at the priority j is expressed as

$$E[W_j] = \frac{E[T'^2_j]}{2E[T'_j]} + \frac{\lambda_j E[T''^2_j]}{2(1 - \lambda_j E[T''_j])} \qquad \text{(A1)}$$

where λ_j, $j = 0, 1, \ldots, K$, denotes the average message arrival rate at the priority j queue, and $E[T'_j]$, $E[T'^2_j]$, $E[T''_j]$ and $E[T''^2_j]$ are the first and second moments of T'_j and T''_j. To determine $E[W_j]$, we need to obtain the first two moments of the conditional effective service times.

Performance analysis of the priority scheme

The first two moments of conditional effective service times $E[T'_i]$, $E[T''_i]$, $E[T'^2_i]$ and $E[T''^2_i]$ are derived in Reference 11. The average queueing delay for priority i, $E[W_i]$ can be determined from (1). $E[W_i]$ can be used to determine the average data latency which is one of the most important parameters for evaluation of the network performance. Data latency is defined as follows.

Definition A.5. Data latency is defined as the time interval between the instant of arrival of a message at the transmitter buffer of the source station and the instant of arrival of the last bit of the same message at the receiver buffer of the destination station. ■

Remark A.4. The expected value of data latency of priority i messages is given as:

$$D_i = E[W_i] + L_i + \mathscr{P}_i \qquad \text{(A2)}$$

where

L_i = average transmission time of a priority i message in unit of time, i.e. length of the message in bits divided by the data latency in bit/unit time,

W_i = queueing delay of a priority i message, and

$\mathscr{P}_i$ = average propagation delay between a source-destination pair. ■

Under stable state, every message that arrives at the transmitter queue is eventually transmitted by virtue of the assumption that there is no message rejection. Therefore, throughput is equal to the offered traffic which is defined below.

Definition A.6. Offered traffic G_i, $i = 0, \ldots, K$, for the priority i class is defined as the expected value of the total transmission time of all priority i messages per unit time and is expressed as

$$G_i = (N_i L_i)/\tau_i \qquad \text{(A3)}$$

where

N_i = number of the priority i queue in the network,

τ_i = average message interarrival time at the priority i queue ($\tau_i = 1/\lambda_i \, \mu s$),

L_i = is as defined above.

Definition A.7. Total offered traffic G_{tot} is the sum of individual offered traffic at each priority level, and is expressed as

$$G_{\text{tot}} = \sum_{i=0}^{K} G_i \qquad \text{(A4)}$$

■

Summary of the analytical model

The statistical model of networked delays at each priority level is based on the following assumptions:

- The message arrival process at each queue is Poisson which is a close approximation of scenarios for a large number of network subscribers. (Note: The queueing model cannot be expressed in a closed form without this assumption.)
- The message waiting process in a queue is independent of those in the other queues.
- The processes of message waiting and token rotation timer expiration within each priority queue are mutually independent.

The analytical model determines the probability that a message is served at the instant of token arrival at a given priority queue. The moment generation functions of the conditional token circulation times for all priority classes are determined on this basis. By using the inverse Laplace-Stieltjes transformation, the probability density functions of the conditional token circulation times are obtained. Based on the probability that the token rotation timer is not expired when the token arrives at a priority queue and the conditional token circulation times, the first and second moments of the conditional effective service times for each priority class are obtained. The average queueing delay for each priority level is determined from the first two moments of the conditional effective service times. Average data latency and average queue length for each priority class are also determined along with the stability conditions of the network. As described in Appendix A of Reference 11, any higher moments of the queueing delay at each priority level can be evaluated from the moment generation functions of the queueing delay which is, in turn, derived from the moment generating function of the conditional effective service time.

ACKNOWLEDGEMENTS

This work was supported in part by Allied-Signal Inc. under Research Contract No. 81-563164.

REFERENCES

1 **Ray, A and Phoha, S** 'Research directions in computer networking for manufacturing systems' *ASME J. Eng. Industry* (May 1989) pp 109-115

2 **Ray, A** 'Networking for computer-integrated manufacturing' *IEEE Netw. Mag. Special Issue on Communications for Manufacturing* Vol 2 No 3 (May 1988) pp 40-47

3 **Ray, A** 'Performance analysis of medium access control protocols for distributed digital avionics' *ASME J. Dynamic Syst. Measure. and Control* (December 1987) pp 370-377

4 ANSI/IEEE Standard 802.4 — 1985 'Token passing bus access method and physical layer specifications' *IEEE* (1985)

5 *SAE Linear Token Passing Multiplexed Data Bus Standard, Version 3.0* (May 1987)

6 *Manufacturing Automation Protocol (MAP) 3.0 Implementation Release* available through MAP/TOP Users Group, One SME Drive, PO Box 930, Dearborn, MI 48121, USA

7 **Archambault, J** 'An IEEE 802.4 token bus network simulation' Report No. NBSIR 84-2966, (US) National Bureau of Standards

8 **Jayasumana, A P** 'Performance analysis of IEEE 802.4 token bus' *Proc. Workshop on Analytic and Simulation Modelling of IEEE 802.4 Token Bus LAN* (US) National Bureau of Standards (April 1987)

9 **Jayasumana, A P and Jayasumana, G G** 'Simulation and performance evaluation of 802.4 priority scheme' *IEEE/ACM Symp. Simulation of Computer Networks* (August 1987)

10 **Jayasumana, A P** 'Performance analysis of a token bus priority scheme' *IEEE INFOCOM* (March 1987)

11 **Hong, S H** 'Performance analysis of token bus protocols for integrated control system networks' Doctoral Dissertation in Mechanical Engineering, Pennsylvania State University, University Park, PA (August 1989)

12 **Dykeman, D and Bux, W** 'Analysis and tuning of the FDDI media access control protocol' *IEEE J. Select. Areas Commun.* Vol SAC-6 No 6 (July 1988) pp 997-1010

13 **Kuehn, P J** 'Multiqueue systems with nonexhaustive cyclic service' *Bell Syst. Tech. J.* Vol 58 No 3 (March 1979) pp 671-689

14 **Hashida, O and Ohara, K** 'Line accommodation capacity of a communication control unit' *Review of the Electrical Communication Laboratories, Nippon Telegraph and Telephone Public Corporation* Vol 20 (March-April, 1972) pp 231-239

15 **Peterson, J M** *Petri Net Theory and the Modelling of Systems* Prentice-Hall, USA (1981)

16 **Diaz, M** 'Modelling and analysis of communication and corporation protocols using Petri net based models' *Comput. Netw.* Vol 6 (December 1982) pp 419-441

17 **Gressier, E** 'A stochastic Petri net model for Ethernet' *Proc. Int. Workshop on Timed Petri Nets* IEEE, Torino, Italy (July 1985)

18 **Pegden, C D** *Introduction to SIMAN* System Modelling Corporation, State College, PA (1986)

19 **Banks, J and Carson II, J S** 'Process-interaction simulation language' *Simulation* Vol 44 No 5 (May 1985) pp 225-235

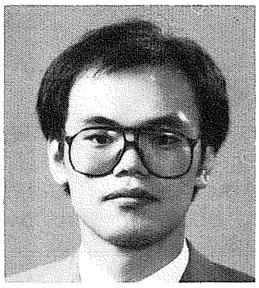
20 **Ho, Y-C and Li, S** 'Extensions of infinitesimal perturbation analysis' *IEEE Trans. Aut. Control* Vol 33 No 5 (May 1988) pp 427-438

21 **Lee, S and Ray, A** 'Perturbation analysis of a token bus protocol for network performance management' *Proc. Amer. Control. Conference* Pittsburgh, PA (June 1989) pp 518-522

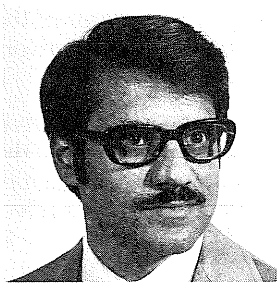
22 **Halevi, Y and Ray, A** 'Integrated communication and control systems: part I — analysis' *ASME J. Dyn. Syst. Measure. and Control* (December 1988) pp 367-373

23 **Ray, A and Halevi, Y** 'Integrated communication and control systems: part II — design considerations' *ASME J. Dyn. Syst. Measure. and Control* (December 1988) pp 374-381

24 **Bazarra, M S and Shetty, C M** *Nonlinear Programming: Theory and Algorithm* John Wiley & Sons, New York (1979)



Seung Ho Hong *received his BSc in Mechanical Engineering from Yonsei University, Seoul, Korea and MSc in Mechanical Engineering from Texas Tech University, Lubbock, Texas. After obtaining a PhD, in Mechanical Engineering from the Pennsylvania State University, University Park, PA, in 1989, he joined the Electronics and Telecommunications Research Institute, Daedog, Korea as a senior researcher. His research interests include performance analysis and design of computer networks, computer networks management, networking for factory automation and computer integrated manufacturing.*



Asok Ray *holds a PhD degree in Mechanical Engineering from Northeastern University, Boston, MA, and graduate degrees in Electrical Engineering, Computer Science and Mathematics. Dr Ray has more than ten years of research and management experience at GTE Strategic Systems Division, Charles Stark Draper Laboratory, and MITRE Corporation. He has also held research and academic positions at Carnegie-Mellon University. Massachusetts Institute of Technology, and the Pennsylvania State University. Dr Ray's research interests include control and instrumentation, networking and communication protocols, intelligent systems design, and modelling and simulation of dynamical systems as applied to aeronautics, process control, and autonomous manufacturing. He is an Associate Fellow of AIAA, a senior member of IEEE and a member of ASME.*