Performance Management of Integrated Manufacturing System Networks

Suk Lee*, Asok Ray, Keum Shik Hong*, Joongsun Yoon* and Myung Chul Han***

(Received April 28, 1994)

Performance management of communication networks is critical for speed, reliability, and flexibility of information exchange between different components, subsystems, and sectors (e.g., factory, engineering design, and administration) of production process organizations in the environment of Computer Integrated Manufacturing (CIM). The objective is to improve the efficiency in handling various types of messages, e.g., control signals, sensor data and production orders, by on-line adjustment of the parameters of the network protocol. This paper presents conceptual design, development, and implementation of a performance management procedure for CIM applications. The performance management algorithm is formulated using the concepts of: (i) perturbation analysis of discrete event dynamic systems; (ii) stochastic approximation; and (iii) learning automata. The performance management procedure has been tested via emulation on a network testbed that is based on the Manufacturing Automation Protocol (MAP).

The conceptual design presented in this paper offers a step forward to bridging the gap between management standards and users' demands for efficient network operations since most standards such as ISO and IEEE address only the architecture, services, and interfaces for network management. The proposed concept for performance management can also be used as a general framework to assist design, operation, and management of flexible manufacturing systems.

Key Words: Computer Networks, Performance Management, Computer Integrated Manufacturing

1. Introduction

Availability of affordable computer hardware and software has resulted in automation of various office and factory operations. These operations include accounting, forecasting, and marketing for business administration; Computer Aided Design (CAD), Finite Element Analysis (FEA), and computer simulations for design and engineering; and, Computer Aided Process Planning (CAPP), inventory control, and monitoring & control of production processes for manufacturing. In an effort for further productivity gain, Computer Integrated Manufacturing (CIM) focuses on combining these activities into a single, closed-loop, and interactive control system. Essential to the success of CIM is computer networking which links spatially distributed islands of automation through timely exchange of relevant information such as new production orders, design modifications, and status report of manufacturing resources (Ray, 1988).

Computer networks for CIM are required to transfer information under continuous changes such as fluctuations in network traffic including addition, deletion, and failure of network components. In order to maintain an acceptable level of service, computer networks must be flexible to adapt themselves to variations in traffic load; this is the responsibility of network management. In

*Research Institute of Mechanical Technology, Pusan National University, Pusan, Korea

**Department of Mechanical Engineering, The Pennsylvania State University, University Park, PA, U.S.A.

general, network management aims to maintain reliable, flexible, and efficient network services by controlling and coordinating available resources. Among various functions of network management, three major components are configuration, fault, and performance management (Thompson, 1986). Configuration management is related to establishing the network and accommodating any configuration changes by modifying the necessary parameters while the primary functions of fault management are fault detection, isolation, and recovery. Performance management monitors a specified measure of network performance and executes the identified actions which can improve the efficiency of network operations.

Basic results have been reported on the architecture, service definition, management protocol, and distribution of network management functions (Thompson, 1986; Saydam and Sethi, 1987; and Klerer, 1988). Several standards for network management have been already published or under preparation as an important part of the integrated protocol suite (MAP/TOP Users Group, to be released, a and b; Institute of Electric and Electronic Engineers, 1990; and Society of Automobile Engineers, 1987). However, researchers have not apparently addressed the key issue of how to accomplish the network management tasks in real time as problems or disruptions arise either in the network operations or in the process that is served by the network. This has happened partly because the development of a methodology for network management is beyond the scope of standardization. Nevertheless, from the users' point of view, network management is crucial for maintaining an acceptable level of services as many functions in different sectors of CIM operations are dependent on timely and reliable exchange of information.

Network operations with a single set of protocol parameters may not result in acceptable network performance over a prolonged period of time because characteristics of network traffic and availability of network resources change from time to time, especially in the environment of flexible manufacturing. Therefore, performance management is needed for tuning the protocol parameters to maintain the acceptable performance. Furthermore, its importance is growing because a single network often must satisfy diverse requirements for different classes of messages resulting from diversity of production processes and their support services. This situation usually occurs when a high degree of systems integration is needed in complex systems by accommodating different classes of messages on a single network by virtue of elaborate priority mechanisms and/or increased network bandwidth of optical fiber media. For example, a single network for CIM should be able to deliver various messages such as database query and response, CAD file transfer, interpersonnel electronic mail, and sensor and control signals within their time limits of delivery under changing traffic patterns caused by common events like arrival of new production orders and machinery breakdown. Even though some protocols (e.g., IEEE 802.4 token bus protocol of MAP) offer mechanisms to handle various types of messages, the key parameters of the protocol which determine the network performance are at the network operator's disposal. The operator may have to adjust these parameters on the basis of ad hoc rules and individual experience because there is no established relationship between protocol parameters and network performance except for the routing and flow control problems in point-to-point networks (Bertsekas and Gallager, 1987).

This paper presents conceptual design, development, and evaluation of a performance management procedure for multiple-access computer communication networks in the CIM environment. The objective is to improve the network performance by on-line adjustment of protocol parameters. The performance management algorithm is formulated using the concepts of: (i) Perturbation Analysis (PA) of Discrete Event Dynamic Systems (DEDS); (ii) Stochastic Approximation (SA); and (iii) Learning Automata (LA). The performance management procedure has been tested via emulation on a MAP network testbed.

This paper is organized into six sections including the introduction and two appendices. Section

2 presents the main theme of the proposed performance management procedure including the concepts of PA, SA, and LA. Analytical formulation of the procedure for the specific case of a token bus protocol is presented in Section 3. Details of implementation of the performance management procedure and its emulation on a MAP network testbed are described in Section 4. The test results are presented and discussed in Section 5. Finally, the paper is summarized and concluded in Section 6. The priority mechanism of the token bus protocol considered here is briefly described in Appendix A, and an example of perturbation analysis of timer setting changes for the token bus protocol is given in Appendix B.

2. Structure of the Performance Management Procedure

Network performance, more specifically delays experienced by message packets, could be critical for dynamic performance and stability of real-time manufacturing processes. This is especially true when multiple machines are performing a task without direct connections among them. One of the examples is an intelligent welding system (Nayak, Ray and Vavreck, 1987) where a positioning table and a robot have to communicate through a network to exchange the desired table position coordinates and various messages for status report. Another example is manipulation of a bulky and flexible workpiece by more than one independent robots which initiate their own prescribed trajectories upon receiving a signal from a controller and report the completion of the trajectories back to the controller via a network. The timeliness of the transmitted data is essential because a delay could damage the workpiece or the robot's wrists and arms.

In order to maintain an acceptable level of the dynamic performance and stability of various manufacturing processes, performance management is required to manipulate adjustable protocol parameters in real time so that the network can adapt itself to the dynamic environment. This can be accomplished in two steps: (1) *performance evaluation* to find how perturbations in protocol parameters affect a selected performance measure, i.e., to determine the relationship between the performance measure and the protocol parameters; (2) *decision making* to decide on how to adjust protocol parameters, i.e., to identify the direction and magnitude of the parameter adjustment vector, utilizing pieces of information provided in the first step and the history of performance.

The analytical techniques for performance evaluation such as queueing theory (Viswanadham and Narahari, 1992) often requires unrealistic assumptions like Poisson arrival, and tend to be mathematically untractable as the structure of the performance measure becomes complex. Furthermore, network traffic statistics such as distributions of message generation interval and message length, which are required as inputs to the analytical model, are very difficult to estimate in real time. On the other hand, discrete-event simulation (Law and Kelton, 1991) is a viable alternative to analytical techniques. A major advantage of simulation over any analytical technique is that a DEDS can be modeled with much less stringent assumptions, and more complex performance measures can be handled with relative ease. However, discrete-event simulation usually suffers from significant computational burden because a single simulation run represents only one realization of a stochastic process. In order to obtain an accurate performance estimate under a given set of parameters, several independent runs (or a lengthy run if the process is ergodic) are needed, and these runs should be repeated for different sets of pertinent parameters. In order to avoid the estimation of network traffic statistics which are still required, one can record time of generation and length for each message and feed this information into a simulation model. However, this requires a large amount of information transfer from each and individual station to the computer on which the model is running, which may degrade the overall network performance.

Over the last decade, Ho and his colleagues (Ho and Cao, 1991; Ho and Li, 1988; and references therein) have developed the technique of Perturbation Analysis (PA) to circumvent the

difficulties of conventional analysis and simulation in DEDS. PA estimates the DEDS performance under perturbed conditions (with different parameter values) by observing the sequence of events occurring over a period of time in the nominal (i.e., unperturbed) system. In fact, PA constructs parts of event sequence for the perturbed system based on the nominal one. This approach has a computational advantage over repetitive simulation especially when no analytic technique is available. When the effects of n parameters on a performance measure are to be evaluated, the conventional discrete-event simulation needs $n+1$ runs (one with the nominal parameters and n runs, each with one perturbed parameter and the remaining nominal values). On the other and, PA needs only one run because it calculates the performance measure of the perturbed system based on the inherent information from the simulation with the nominal parameter. Therefore, the ratio of computation time can be approximately 1 to $n+1$ if the processing time for PA algorithms is negligible compared to that for discrete-event simulation. For performance management, PA is very suitable because this technique can directly utilize on-line observation of events. This does not require any identification of statistical parameters of the network traffic and is computationally more efficient than discrete-event simulation. Furthermore, PA still retains the inherent advantages of simulation over analytical techniques.

Decision making requires parameter optimization, and can be accomplished numerically by Stochastic Approximation (SA) which utilizes random measurements over a finite period of time to estimate the finite-difference quotient of the performance measure with respect to decision variables (Rubinstein, 1986). Since the measurement for performance measure is a random variable with an unknown distribution, the estimated quotients have a non-zero variance at every point. Consequently, the SA technique has to reduce its step size as the extremal point is approached. In many situations, however, more than one optimization algorithm may be required because any individual algorithm is likely to be efficient only in some region of the protocol parameter settings under given statistics of network traffic. An additional level of decision making is desirable to select the most appropriate optimization algorithm according to the current parameter settings and traffic statistics. This approach is likely to enhance the efficiency and credibility of performance management in a dynamic operating environment whose characteristics are unknown or partially known. Techniques like Learning Automata (LA) can be used for decision making at the upper level for selecting the most efficient and credible optimization algorithm based on the past performance (Narendra and Thatachar, 1989).

A learning automaton consists of a set of actions, a corresponding set of action probabilities, and a reinforcement scheme. The action probabilities are updated by the reinforcement scheme according to the response from the environment which reacts to the action of the learning automaton. A Performance Evaluator (PE) as a part of the environment is required to interpret the response. If the interpretation is favorable, then the probability of the chosen action is increased and those for other actions are decreased. By repeating this process, the learning automaton can select the best action under the current environment. In a dynamic environment, the performance management algorithm must know whether the network traffic statistics have changed because: (i) the optimization algorithm is likely to have its own ability of adaptation such as reduction of the step size in SA; and (ii) a change in network traffic may mislead the learning automaton in evaluating the performance of the optimization algorithms. Therefore, whenever any change in network traffic statistics takes place, the step size of the optimization algorithm and the PE may have to be reset.

3. Formulation of a Performance Management Procedure

The proposed performance management is based on the principles of Perturbation Analysis (PA), Stochastic Approximation (SA), and

Learning Automata (LA), and its overall structure is shown in Fig. 1. The Linear Token Passing Bus (LTPB) protocol (Society of Automobile Engineers, 1987) has been selected to demonstrate the efficacy of this performance management procedure for adjustment of Token Holding Timer (THT) and three Token Rotation Timers (TRT1, TRT2, and TRT3) in real time on the basis of measured network performance. The operating principle of LTPB and its priority mechanism are described in Appendix A.

3.1 Performance evaluation via perturbation Analysis

The knowledge of the relationship between a selected performance measure and the pertinent protocol parameters is a critical requirement for performance management. In this case, the behavior of the LTPB protocol with perturbed parameter settings under the same stochastic realization is constructed from a single observation with nominal parameters. The settings of the four timers of LTPB directly influence data latency (i.e., the time interval between the instant a message enters the source station's queue to the instant the last bit of the message is received at the destination station) whose statistics are often one of the most dominant factors in the network performance. How to assess the effects of change in timer settings on network operations is illustrated by a simple example of an LTPB network in Appendix B.

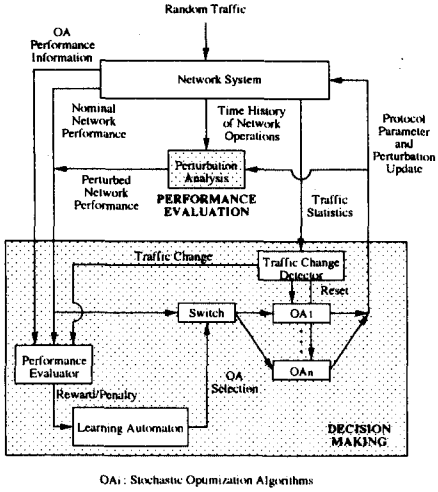

Fig. 1 Overall Structure of Performance Management

To reduce the complexity of the problem, we assume that there is no queue saturation on both nominal and perturbed paths. This assumption holds good for the majority of network applications because the network is usually designed to operate sufficiently below its full load and a safe margin is allowed in the allocation of queue capacity. The PA algorithm consists of two parts: (i) detection of any possible difference between the nominal path and a perturbed path; and (ii) construction of the portions of the perturbed path which differs from the nominal one. No additional event scheduling in discrete-event simulation is required for any one of these two parts. This is critical because a major share of computations in discrete-event simulation results from event scheduling and its execution.

Detection of Difference between Nominal and Perturbed Paths: A portion of the perturbed path begins to differ from the nominal one when the number of message transmissions from a queue on a particular token reception is different from each other. (Transfer of the right to transmit between the queues of a station is also considered as a token reception.) In order to establish the test conditions for a different number of message transmissions on the perturbed path, several notations are introduced.

- m : Number of messages just transmitted from the current queue on the nominal path, $m=0, 1, 2, \cdots$.
- q : Number of messages waiting in the current queue after m transmissions on the nominal path, $q=0, 1, 2, \cdots$.
- L_k : Transmission time of the k-th message, $k=1, 2, \cdots, m$.
- T : Nominal token circulation time of the current queue, i.e., the time interval between the previous and the current token reception instants.
- THT : Nominal length of THT.
- ΔTHT : Perturbation in THT, $THT+\Delta THT>0$.

TRT_i : Nominal length of TRTi.
ΔTRT_i : Perturbation in TRTi, $TRT_i + \Delta TRT_i > 0$.

The following tests must be executed before passing the token to the next queue after m message transmissions.

TEST0 [test for priority class 0]

Case a. $\Delta THT > 0$ and $m > 0$:

$$\sum_{k=1}^{m} L_k < (THT + \Delta THT) \text{ and } q > 0$$

Case b. $\Delta THT < 0$ and $m > 1$:

$$\sum_{k=1}^{m-1} L_k \geq (THT + \Delta THT) \quad (1)$$

TESTi [test for priority class i, i=1,2,3]

Case a. $(TRT_i - T) > 0$, $\Delta TRT_i > 0$ and $m > 0$:

$$\sum_{k=1}^{m} L_k < (TRT_i + \Delta TRT_i - T) \text{ and } q > 0$$

Case b. $(TRT_i - T) > 0$, $\Delta TRT_i < 0$ and $m > 1$:

$$(TRT_i + \Delta TRT_i - T) \leq \sum_{k=1}^{m-1} L_k$$

Case c. $(TRT_i - T) \leq 0$ and $\Delta TRT_i > 0$:

$$(TRT_i + \Delta TRT_i - T) > 0 \text{ and } q > 0$$

Case d. $(TRT_i - T) > 0$, $\Delta TRT_i < 0$ and $m = 1$:

$$(TRT_i + \Delta TRT_i - T) \leq 0 \quad (2)$$

TEST0.a is applicable if a priority class 0 queue can transmit more message(s) on the perturbed path in addition to m transmitted messages on the nominal path. Additional transmission is possible only when THT is increased and there has been at least one message transmission on the nominal path. (No transmission on the nominal path implies that the queue is empty upon token reception.) Further, the increased THT should be long enough so that the perturbed THT is not expired even after m transmissions and the queue should not be empty for additional transmission(s). TEST0.b is opposite to TEST0.a. If TEST0.b is satisfied, then the number of transmissions on the perturbed path is less than the number m of transmissions on the nominal path. In this case, ΔTHT should be negative and the perturbed THT should expire during the (m - 1)st transmission at the latest. TESTi.a and TESTi.b are largely equivalent to TEST0.a and TEST0.b, respectively. TESTi.b represents a slightly different situation compared to TEST0.b in the following sense. TESTi.b could imply that no transmission is allowed on the perturbed path while there would be at least one transmission on the perturbed path for TEST0.b. TESTi.c implies that some transmissions are possible on the perturbed path while there has been no transmission on the nominal path due to expiration of TRT$_i$. If TESTi.d, which can be considered as a special case of TESTi.b, is satisfied, then no transmission is allowed on the perturbed path while one transmission was possible on the perturbed path.

Construction of the Perturbed Path: The construction of a perturbed path consists of three parts: maintenance of the perturbed queue status; calculation of the perturbed timer status; and propagation of the effects of perturbation on the instant of token reception. Records associated with a message such as generation time and message length are kept even after the message is transmitted on the nominal path. These records are discarded only after the message is transmitted on both nominal and perturbed paths. In this way, queue contents on the perturbed path are available for construction of the perturbed path.

For the priority level 0, THT status is independent of the token circulation time T since THT is always reset to its full value at each token reception. Therefore, upon a token reception, the remaining interval of THT on the nominal path $R(0)$ is always equal to THT during which the priority 0 queue can start to transmit its messages. On the perturbed path, the interval $R'(0)$ is

$$R'(0) = THT + \Delta THT. \quad (3)$$

For the lower priority levels, the remaining interval of TRT$_i$ on the nominal path is equal to $R(i) = \max(TRT_i - T, 0)$ upon a token reception, where $\max(\cdot, \cdot)$ denotes the larger of the two arguments. Since the interval is dependent on T, the perturbation on the instants of the current and previous token receptions at a queue should be considered, which are denoted by ΔC and ΔP, respectively. The perturbation on the remaining time when the token is received, $\Delta R(i)$, is

$$\Delta R(i)=\begin{cases}\Delta P-\Delta C+\Delta TRT_i \\ \quad \text{if}(TRT_i-T)>0, \\ \Delta P-\Delta C+\Delta TRT_i+TRT_i-T \\ \quad \text{if}(TRT_i-T)\le 0.\end{cases} \tag{4}$$

The first case of the above equation applies when the timer is not expired on the nominal path as shown in Fig. 2. The second case applies when the timer is already expired prior to the token reception as shown in Fig. 3. Then, at the token reception, the remaining interval on the perturbed path, $R'(i)$, during which priority i transmissions can start is

$$R'(i)=\max(R(i)+\Delta R(i),0). \tag{5}$$

ΔC for the next queue is yet to be obtained for construction of the perturbed path. The time $X(i)$, spent in transmitting messages from the current queue, is known from the nominal path. The time, $X'(i)$, spent for message transmissions on the perturbed path can be obtained on the basis of the perturbed queue status and the perturbed remaining time $R'(i)$. $X'(i)$ is calculated by summing up the transmission time, L_j, of the j-th message from the perturbed queue either until the perturbed queue becomes empty or until the perturbed TRTi is expired, i.e., $R'(i)-X'(i)\le 0$. While calculating $X'(i)$, the observations, like data latency on the perturbed path, can be recorded for estimating the perturbed performance.

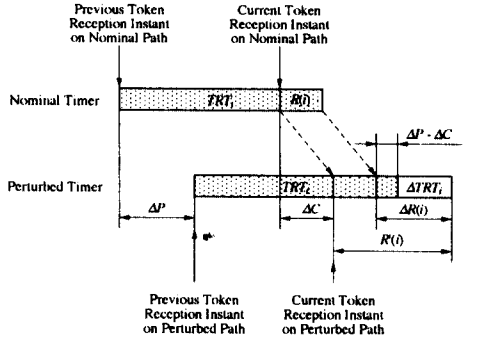

Fig. 2 Perturbed Remaining Interval of TRTi (Case 1)

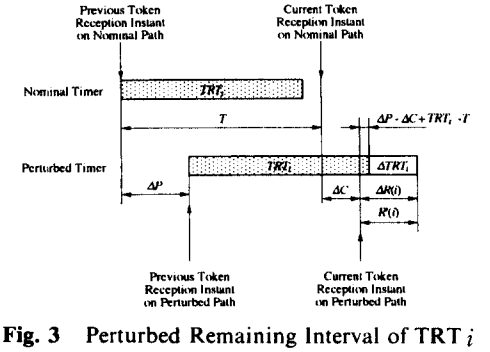

Fig. 3 Perturbed Remaining Interval of TRTi (Case 2)

Defining $\Delta X(i)=X'(i)-X(i)$, perturbation at the instant of token reception for the next queue is obtained as:

$$\Delta C(i+1)=\Delta C(i)+\Delta X(i) \tag{6}$$

where the indices i and $i+1$ which indicate the current and the next queues are based on modulo 4, i.e., they range from 0 to 3.

Summary of the PA algorithm: The algorithm for perturbation analysis of the LTPB protocol is summarized in four steps as delineated below.

Step 1. Perform appropriate tests (TEST0 and TESTi, $i=1, 2, 3$) just before passing the token to the next queue or station. Set the flag if any of the tests are satisfied.

Step 2. If the flag is not set, repeat Step 1 for the next queue or station. Otherwise, proceed to Step 3.

Step 3. Execute the construction procedure for the current queue before passing the token.

Step 4. Repeat Step 3 until all queues have $\Delta C=0$ and $\Delta P=0$ during a token circulation. If so, clear the flag and go to Step 1.

3.2 Stochastic approximation

Although the Stochastic Approximation (SA) technique described in (Rubinstein, 1986) has been proved to converge in a stochastic sense, it is known to have a slow convergence in many practical situations due to the fact that its step size is uniformly reduced regardless of the current value of decision variable $\mathbf{x}[k]$ (a 4x1 vector of timer settings in this paper) at the k-th iteration. To circumvent this difficulty, the conventional algorithm is modified following Ho et al. (Ho and Cao, 1983). This modified algorithm, hereafter referred to as Modified Stochastic Approximation (MSA), can be written as follows.

$$\mathbf{x}[k+1]=\mathbf{x}[k]-\Gamma[k]\mathbf{f}(\mathrm{q}^{\pm}[k], \cdots, \mathbf{q}^{\pm}[k-m_w+1]) \quad (7)$$

$\Gamma[k]$ is a diagonal matrix whose non-zero elements are not restricted to be identically equal and are dependent on the number of sign reversals in the corresponding forward or backward finite-difference quotient $q_i^{\pm}[k]$. The i-th diagonal element of $\Gamma[k]$ is

$$\gamma_i[k]=\max(c_i r^{e_i[k]}, l)$$

where c_i is a positive constant for initial step size, $r\in(0, 1)$ is a reduction factor, the $e_i[k]$ is an integer counter variable indicating the number of sign reversals in the i-th quotient $q_i^{\pm}[k]$, and l is the lower bound common to all diagonal elements. The forward or backward quotient of the observed performance $g(\mathrm{x}[k], \omega[k])$ at a sample point $\omega[k]$ is defined as

$$q_i^{\pm}[k]=\frac{g(\mathbf{x}[k]+\Delta x_i[k]\mathbf{u}_i, \omega[k])-g(\mathbf{x}[k], \omega[k])}{\Delta x_i[k]} \quad (8)$$

where $\mathbf{u}_i$ is the i-th unit vector, and a positive perturbation for i-th decision variable, $\Delta x_i[k]$ is used for forward difference while a negative $\Delta x_i[k]$ for backward difference. The MSA algorithm also utilizes the past quotients to smooth the adjustments in $\mathbf{x}[k]$. The function $\mathbf{f}$ takes a weighted average of m_w recent quotient vectors $\mathbf{q}^{\pm}[k]=\sum_{i=1}^{n} q_i^{\pm}[k]\mathbf{u}_i$. That is, the i-th element of $\mathbf{f}$ is written as

$$f_i(\mathbf{q}^{\pm}[k], \cdots, \mathbf{q}^{\pm}[k-m_w+1])=\frac{g(\mathbf{x}[k], \omega[k])}{x_i[k]}\sum_{j=1}^{m_w}\frac{w_j x_i[k-j+1]q_i^{\pm}[k-j+1]}{g(\mathbf{x}[k-j+1], \omega[k-j+1])} \quad (9)$$

where $\sum_{j=1}^{m_w} w_j=1$ and $w_j\geq 0\ \forall j$.

Remark: The basic idea behind the MSA algorithm is that the sign of the quotient changes more frequently as $\mathbf{x}[k]$ approaches its optimal point since the noise contained in the quotient is likely to determine the sign. The MSA algorithm adapts its step size based on the number of sign reversals in the quotient in contrast with uniform reduction in stochastic approximation. The weighted averaging function $\mathbf{f}$ serves to reduce the length of a period to measure the performance and to avoid alternating direction of the parameter adjustments. With an appropriate choice of the window size m_w and the weighting factors w_j, inherent noise in measurements may be reduced without sacrificing the speed of convergence. However, a rigorous proof of the convergence for the MSA algorithm has not been established. This is apparently untractable because of the dependence of the step size on past history.

3.3 Learning automata

For autonomous selection of the optimization algorithm, the Variable-Structure Stochastic Automaton (VSSA) (Narendra and Thatachar, 1989) has been adopted because greater flexibility can be exercised within a smaller structure in comparison to those in fixed or deterministic settings. The reinforcement scheme in VSSA updates the action probabilities in discrete time based on the responses from the Performance Evaluator (PE). The next action of the automaton is selected on the basis of the updated action probabilities whose sum is equal to one. A simplified VSSA is represented by the triple $\{\bar{\alpha}, \bar{\beta}, A\}$ where $\bar{\alpha}=\{\alpha_1, \alpha_2, \cdots, \alpha_a\}$, $\bar{\beta}=\{\beta_1, \beta_2, \cdots, \beta_s\}$, and A: $\bar{\alpha}\times\bar{\beta}\rightarrow\bar{\alpha}$ is the reinforcement scheme. $\bar{\alpha}$ is the set of available actions (i.e., optimization algorithms); $\alpha[k]$ denotes the action at the instant k. $\bar{\beta}$ is the set of responses (i.e., possible levels of performance of the current optimization algorithm) that are inputs to the automaton, and the response at the instant k is denoted by $\beta[k]$.

The discrete reinforcement scheme for performance management has been formulated following the concept of discrete reward/penalty (DRP) automaton (Oommen and Christensen, 1988). In this case, the automaton has two available actions (i.e., there is a choice between two alternative optimization algorithms) in response to five possible inputs, namely 0, 0.25, 0.5, 0.75, and 1, from the environment. Among these inputs, $\beta_1=0$ indicates the most favorable response while $\beta_5=1$ is the most unfavorable one. The reinforcement scheme has $M+1$ states (i.e., the action probability is allowed to assume one of the $M+1$ values) where M is an even number greater than

two. The reinforcement scheme is presented below:

$$\Delta p_1[k]=\begin{cases}\frac{2}{M} & \text{if } \alpha[k]=1 \text{ and } \beta[k]=0,\\ & \text{or if } \alpha[k]=2 \text{ and } \beta[k]=1;\\ \frac{1}{M} & \text{if } \alpha[k]=1 \text{ and } \beta[k]=0.25,\\ & \text{or if } \alpha[k]=2 \text{ and } \beta[k]=0.75;\\ 0 & \text{if } \beta[k]=0.5;\\ \frac{-1}{M} & \text{if } \alpha[k]=1 \text{ and } \beta[k]=0.75,\\ & \text{or if } \alpha[k]=2 \text{ and } \beta[k]=0.25;\\ \frac{-2}{M} & \text{if } \alpha[k]=1 \text{ and } \beta[k]=1,\\ & \text{or if } \alpha[k]=2 \text{ and } \beta[k]=0\end{cases} \tag{10}$$

and

$$p_1[k+1]=\begin{cases}\min(p_1[k]+\Delta p_1[k], 1) & \text{if } \Delta p_1[k]>0,\\ p_1[k] & \text{if } \Delta p_1[k]=0,\\ \max(p_1[k]+\Delta p_1[k], 0) & \text{if } \Delta p_1[k]<0\end{cases} \tag{11}$$

where $p_1[k]$ is the probability of selecting α_1. The probability of selecting α_2 is given as $p_2[k]=1-p_1[k]\,\forall k$.

4. Implementation of the Performance Management Tool

The performance management algorithm has been implemented on a network testbed where the LTPB protocol under consideration was emulated by two interacting application processes. The testbed is operated on the IEEE 802.4 token bus protocol in the environment of the (10 Mbps broadband) Manufacturing Automation Protocol (MAP). The physical configuration of the testbed consists of a length of coaxial cable, a head-end remodulator, and three hosts. Each host computer is equipped with two network cards from Industrial Networking Incorporated (Industrial Networking Incorporated, 1987). One of the hosts operates as a network management console for initial down-loading of the protocols to the remaining two hosts. For these two hosts, a software package has been developed to emulate a number of LTPB stations by generating, transmitting, and receiving messages which are essentially packets of the Association Control Service Element (ACSE) of MAP.

4.1 Implementation strategy

Since the PA algorithm involves only logic and addition operations, it has been implemented in a distributed manner so that the algorithm is executed at each station to estimate the network performance under perturbations. The algorithm is capable of following four perturbed paths in which one of the four timer settings of the LTPB priority mechanism is perturbed. The effect of a timer perturbation (perturbation in the token reception instant) on the next station propagates through the logical ring via the token which carries this information. In this distributed implementation, additional traffic due to management operations is expected to be significantly smaller than that for a centralized PA algorithm which requires information on contents of queues and timer status from all stations. On the other hand, the decision-making module that includes stochastic approximation and learning automaton is centralized at a designated station, hereafter referred to as performance manager. Even though the network may have more than one station with partial or complete capability to execute decision-making functions, the centralized strategy allows only one active copy of each decision-making function. The rationale for selecting centralized decision making is that the distributed strategy, where every station could make decisions autonomously based only on its local performance, would result in inconsistency and conflict by having different timer settings over the network.

Network operations under the proposed performance management procedure involve a series of iterations which consist of an observation period and subsequent management actions. At the beginning, the performance manager broadcasts the initial timer settings and timer perturbation vectors to all stations. During an observation period, each station executes its own PA algorithm, and the performance manager monitors the messages flowing over the network. When the performance manager decides that enough data have been collected, it waits for the token and then broadcasts the request for a performance report to all stations. Then the performance

manager passes the token without any further message transmission. Once this request from the performance manager is received, other stations interrupt their normal operations, prepare the performance reports, and transmit these reports as soon as the token is received. After one complete token circulation, the manager receives the token again and, by this time, reports from all other stations have been received. Then, the manager processes the reports, computes new timer settings and broadcasts them with new timer perturbation vectors for the next iteration. Upon reception of the new settings and perturbation vectors, all stations set their corresponding variables and wait for the token to resume normal operations.

4.2 Implementation details

In this implementation, a measure of the network performance has been formulated on the basis of observed data latency. Throughput of the network is excluded from the performance measure since the network traffic is assumed to be below the network capacity, which implies that all messages entering the network are eventually transmitted to their destination. In other words, the throughput is equal to the offered traffic. Therefore, the main focus of the performance measure is on data latency, i.e., how long it takes for a message to reach its destination relative to the instant of its generation. This type of performance measure will reflect how timely the messages are delivered and will be suitable for many manufacturing applications such as the intelligent welding system and the manipulation of bulky objects. The network performance $g(\cdot, \cdot)$ is expressed as a function of data latency:

$$\begin{aligned} & g(\mathbf{x}[k],\ \omega[k]) \\ &= \frac{\rho}{m_a[k]} \sum_{i=0}^{3} \sum_{j=1}^{m_i[k]} F_i(\delta_j^i(\mathbf{x}[k], \omega[k])) \\ &\quad + (1-\rho)\left(\frac{1}{m_a[k]} \sum_{i=0}^{3} \sum_{j=1}^{m_i[k]} \delta_j^i(\mathbf{x}[k],\ \omega[k])\right)^2 \end{aligned} \tag{12}$$

where $\rho \in [0,\ 1]$ is a weighting factor, $m_a[k]$ is the number of messages observed during the k-th iteration, i denotes priority level, $m_i[k]$ is the number of the priority level i messages during the k-th iteration which is related to $m_a[k]$ by $m_a[k] = \sum_{i=0}^{3} m_i[k]$, and $\delta_j^i(\mathbf{x}[k],\ \omega[k])$ is the data latency of the j-th priority level i message during the k-th iteration. $F_i(\cdot)$ is a penalty function for the priority level i messages and is defined as

$$F_i(\delta_j^i(\mathbf{x}[k],\ \omega[k])) = \begin{cases} 0 \\ \text{if } \delta_j^i(\mathbf{x}[k],\ \omega[k]) \le \theta_i, \\ (\delta_j^i(\mathbf{x}[k],\ \omega[k]) - \theta_i)^2 \\ \text{if } \theta_i < \delta_j^i(\mathbf{x}[k],\ \omega[k]) \le \theta_i + b_i, \\ b_i^2 \\ \text{if } \delta_j^i(\mathbf{x}[k],\ \omega[k]) > \theta_i + b_i \end{cases} \tag{13}$$

with penalty threshold θ_i and penalty band b_i for the priority level i messages.

The first term of the performance measure represents the average penalty over all messages such that a message of which data latency exceeds the corresponding threshold is penalized according to the penalty function $F_i(\cdot)$. This is analogous to variance of data latency. This form of penalty is especially useful for messages carrying time-critical information such as control signals and sensor data, interrupt signals, and video and voice data. The second part of the performance measure takes into account the square of average data latency over all messages. For network emulation on the testbed, ρ is set to one because the two terms of the performance measure were found to have a very close correlation with each other from simulation experiments.

The Performance Evaluator (PE) assesses the performance of an optimization algorithm by observing various items after timer settings have been changed. To this end, the performance of an optimization algorithm has been formulated focusing on its stepwise behavior rather than the asymptotic one. In fact, the asymptotic performance of an algorithm may not be measurable because the performance manager usually switches from one algorithm to another. The PE maintains history of the network performance and adjustment of the timer settings in order to provide criteria for evaluation of the recent action by the learning automaton. In this implementation, records for the past five iterations are maintained. The PE compares the current network perfor-

mance with the average network performance of the past iterations. Also, the magnitude and the signs of the current timer changes are compared to those of the averaged timer changes for the past iterations. Since perturbation analysis is executed all the time, the signs of the next changes are available for the PE to compare the signs of the current changes to those of the next changes. According to these comparisons, the PE selects one value for $\beta[k]$ out of possible five values.

5. Results and Discussion

The first part of this section presents the results from simulation experiments which were conducted to investigate the accuracy of the PA algorithm which is an important ingredient of the performance management tool. The second part focuses on the results from emulation experiments on the network testbed to demonstrate the efficacy of the proposed tool.

5.1 Results of simulation experiments

An LTPB network with 10 stations has been simulated with the transmission rate of 50 megabits per second (Mbps) and queue capacity of 10 messages for every queue. The network traffic is composed of messages at four priority levels. The total network traffic takes 70 percent of the network capacity on the average: 10 percent for priority level 0 and 20 percent each for priority levels 1, 2, and 3. The PA algorithm formulated in Section 3.1 has been implemented in the simulation. This implementation maintains four separate perturbed paths and only one timer setting can be perturbed in each path. The simulation result under a typical scenario is discussed below.

Having set the timers, THT, TRT1, TRT2, and TRT3, at the nominal values of 150, 1000, 800, and 600 μsec, respectively, the PA algorithm was used to estimate the perturbed performance under four different perturbations on TRT3: -100, -50, 50, and 100 μsec. Four additional simulation experiments have also been performed to obtain the network performance with 4 sets of the perturbed timer settings without using the PA algorithm.

Figure 4 shows the absolute values of percent error in estimating the perturbed network performance with TRT3 perturbations. In Fig. 4, S denotes the number of future message generations considered by the PA algorithm. For $S=0$ where no future message is taken into consideration (shown by solid black bars), the error for the positive perturbation is much larger than that for the negative one. The reason is that a positive increment in a timer setting is likely to make the perturbation in token reception instant ΔC positive during a major part of the simulation experiment because an increased timer setting enables the corresponding queues to transmit more messages compared to the nominal path. If ΔC is positive, the PA algorithm has to construct the perturbed path with no knowledge on any potential change in the queue contents from the present time t. In order to reduce errors in estimating the perturbed performance, the PA algorithm has been modified to look into the event calendar so that the next message generation at a given queue can be considered in constructing the perturbed path. This idea has been extended to consider message generations further in the future by scheduling more than one message generation for a given queue. The shaded bars for $S=1$ in Fig. 4 show a significant reduction in the absolute values of the percent error when one future message generation at each queue is considered by

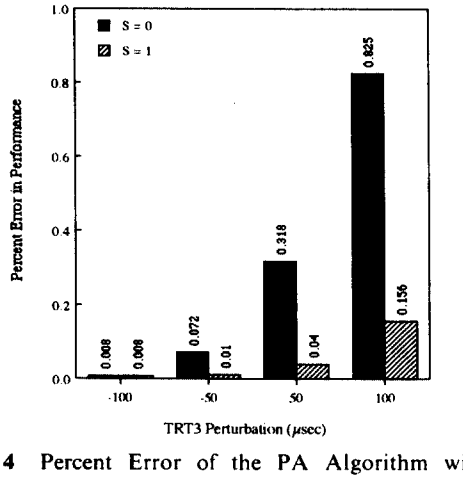

Fig. 4 Percent Error of the PA Algorithm with TRT3 Perturbation

the PA algorithm. However, this method for reducing errors is applicable only to simulation experiments. Simulation results show that the PA algorithm is capable of estimating the performance of the LTPB network under perturbed parameters by using the results of a single simulation experiment under the nominal condition. The estimation errors are found to be within 3 percent from all simulation experiments for perturbations in THT, TRT1, TRT2 and TRT3.

5.2 Results of emulation experiments on the testbed

Three experiments, conducted on the network testbed, are reported here as typical results. For each of these experiments, the emulated network was run for 100 iterations which are equivalent to 300,000 message transmissions. During the first 25 iterations, the network was operated without any performance management action in order to achieve the steady-state operations. Four timers of the LTPB priority mechanism were adjusted on line during the remaining 75 iterations. The network is loaded with 60 percent of its capacity which is evenly distributed over four priority levels. The network traffic for the priority levels 0 and 1 is intended to emulate real-time messages such as control signals and sensor data which have a stringent time limit. Both priority levels have uniformly distributed message length and interval between message generations. The messages at the priority level 0 are generated more frequently but with a smaller number of bytes compared to those at the level 1 in order to represent messages for the systems with fast dynamics. On the other hand, the priority levels 2 and 3 have exponentially distributed message length and interval between message generations with larger averages than those for the priority levels 0 and 1. This scenario mimics the network traffic of occasional information transfer for other non-real-time applications such as NC program downloading and inventory control activities.

For these experiments, two different MSA algorithms (described in Section 3.2) were used: one is referred to as *conservative* and the other as *liberal*. The conservative MSA algorithm has smaller values for initial step size c_i, reduction factor r, and weighting factor w_1 compared to the liberal algorithm. In other words, the conservative algorithm tends to adjust timers in smaller increments and to reduce its step sizes faster compared to the liberal algorithm. For the first two experiments, the two algorithms were used separately and one at a time for performance management without the learning scheme. This is referred to as the single-algorithm performance management (SPM). The conservative SPM was found to yield better cumulative performance than the liberal SPM. For the third experiment, both algorithms were employed together for performance management but one and only one of the two algorithms is selected by the learning automaton at any given iteration. This is referred to as the dual-algorithm performance management (DPM).

Figure 5 shows a comparison between the cumulative network performance of the DPM and the conservative SPM. As a reference, the network without any performance management is also shown in Fig. 5. Up to 50 iterations, both SPM and DPM yield comparable network performance which is significantly superior to that without any performance management. Then onwards, the cumulative performance using the DPM improves faster than that using the conservative SPM. The rationale is that the DPM selects the liberal SPM (which has relatively larger step sizes) more fre-

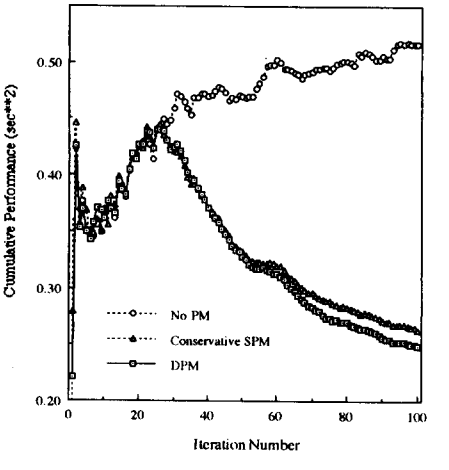

Fig. 5 Comparison of Cumulative Performance

quently to change timer settings in larger increments while the conservative SPM changes the settings in smaller increments which are reduced prematurely.

6. Summary and Conclusions

A performance management algorithm for multiple-access networks has been conceptualized, and formulated for a token bus protocol by using the principles of: (i) perturbation analysis of discrete event dynamic systems; (ii) stochastic approximation, and (iii) learning automata. The procedure is aimed to improve the performance of CIM networks in handling various types of messages by on-line adjustment of protocol parameters, and has been emulated on a network testbed. The conceptual design presented in this paper offers a step forward to bridging the gap between management standards and users' demands for efficient network operations since most standards such as ISO and IEEE address only the architecture, services, and interfaces for network management. The following major conclusions are derived from the results of simulation and emulation of performance management of Linear Token Passing Bus (LTPB) protocol.

- The perturbation analysis algorithm can estimate the performance of an LTPB network (e.g., IEEE 802.4 within the MAP architecture) by using the results of a single simulation experiment under the nominal condition. The estimation errors are found to be within 3 percent for all simulation experiments.
- The performance of a manufacturing system network can be maintained by on-line adjustment of its timers while the traffic load may change to meet the production schedules, equipment failures, or other disruptions. Most importantly, the performance management procedure does not require identification of the network traffic statistics.
- The discrete reward/penalty (DRP) reinforcement scheme is well suited as an ingredient of on-line performance management under unknown and dynamically changing environment.
- The framework developed in this research may be extended to a wide range of Discrete Event Dynamic Systems (DEDS) including FMS. According to the current status of an FMS, manufacturing resources such as robots and programmable machine tools can be dynamically allocated to various manufacturing processes of different product types.

As described in this paper, the performance management of a network has to be executed with insufficient a priori knowledge in the performance characteristics of the network and its surrounding environment. It may be effective to employ various knowledge-based techniques such as fuzzy reasoning, expert system, neural network, and genetic algorithm.

References

Bertsekas, D. and Gallager, R., 1987, *Data Networks*, Prentice Hall.

Ho, Y-C. and Cao, X-R., 1983, "Perturbation Analysis and Optimization of Queueing Networks," *Journal of Optimization Theory and Applications*, Vol. 40, No. 4, pp. 559~582.

Ho, Y-C. and Cao, X-R., 1991, *Perturbation Analysis of Discrete Event Dynamic Systems*, Kluwer Academic.

Ho, Y-C. and Li, S., 1988, "Extensions of Infinitesimal Perturbation Analysis," *IEEE Transactions on Automatic Control*, Vol. AC-33, No. 5, pp. 427~438.

Industrial Networking Incorporated, 1987, MP-400 *Programming Reference Manual*.

Institute of Electric and Electronic Engineers, 1990, *ANSI/IEEE Standard 802.4 Token Passing Bus Access Method and Physical Layer Specifications*.

Klerer, S. M., 1988, "The OSI Management Architecture: An Overview," *IEEE Network*, Vol. 2, No. 2, pp. 20~29.

Law, A. M. and Kelton, W. D., 1991, *Simulation Modeling and Analysis*, 2nd ed. McGraw-Hill.

MAP/TOP Users Group, *Manufacturing Automation Protocol (MAP) 3.0, Implementation Release.*

MAP/TOP Users Group, *Technical and Office Protocols (TOP) 3.0, Implementation Release.*

Narendra, K. S. and Thatachar, M. A. S., 1989, *Learning Automata*, Prentice Hall.

Nayak, N., Ray, A. and Vavreck, A. N., 1987, "Adaptive Real-Time Intelligent Seam Tracking Systems," *Journal of Manufacturing Systems*, Vol. 6, No. 3, pp. 241~245.

Oommen, B. J. and Christensen, J. P., "ε-Optimal Discretized Linear Reward-Penalty Learning Automata," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. SMC-18, No. 3, pp. 451~458.

Ray, A., 1988, "Networking for Computer-Integrated Manufacturing," *IEEE Network*, Vol. 2, No. 3, pp. 40~47.

Rubinstein, R. Y., 1986, *Monte Carlo Optimization, Simulation and Sensitivity of Queueing Networks*, John Wiley.

Saydam,T. and Sethi, A. S., 1987, "Token Bus/Ring Local Area Network Management Concepts and Architecture," *IEEE INFOCOM*, pp. 988~993.

Society of Automobile Engineers, 1987, *Linear Token Passing Multiplexed Data Bus Standard*, Version 3.0.

Thompson, D. M., 1986, "LAN Management Standards—Architecture and Protocols," *IEEE INFOCOM*, pp. 355~363.

Viswanadham, N. and Narahari, Y., 1992, *Performance Modeling of Automated Manufacturing Systems*, Prentice Hall, Englewood Cliffs, NJ.

Appendix A
Linear Token Passing Bus Protocol

A token bus protocol is a distributed controlled-access protocol for the Medium Access Control (MAC) layer. The right to use the medium is explicitly controlled by a special bit pattern called a token, and the responsibility of controlling the use of the medium lies with every station. This appendix describes the priority mechanism of the Linear Token Passing Protocol (LTPB) which is used in this paper to demonstrate efficacy of the proposed performance management tool. A complete description and detailed specifications of the LPTB are given in the reference (Society of Automobile Engineers, 1987).

A token bus network consists of a number of stations connected via a broadcast medium on which any transmission from a station can be heard by all stations. The right to transmit a message is given to a station when it receives a special bit pattern called a token. The token is passed from a station to another following a sequence of station addresses. The last station in the sequence sends the token back to the first station to form a logical ring. A station may transmit its messages before it passes the token to the next station in the logical ring sequence. A station with the token has complete control of the medium for a finite period of time. The length of this period depends on the number of waiting messages and the status of several timers as explained later on priority mechanism. A station can transmit a number of messages or can pass the token to its successor (i.e., the next station in the logical ring sequence) without any transmission.

The LTPB protocol has a priority mechanism of four levels, namely 0, 1, 2 and 3, among which the priority level 0 has the highest privilege of medium access. Each priority level has a queue to provide temporary waiting space for messages of the corresponding priority level. A Token Holding Timer (THT) and three Token Rotation Timers, i.e., TRT1, TRT2 and TRT3 regulate message transmissions for the priority level 0, 1, 2 and 3, respectively. The priority level 0 messages are allowed to start transmission within a period equal to the length of THT. For the lower priority level messages, the initiation of a transmission must not occur beyond the residual period (i.e., the time left until its expiration) of the corresponding timer (TRTi, i = 1, 2, 3). If a timer expires while the corresponding priority message is being transmitted, the transmission will be continued to completely finish the current message and no further transmission is allowed until the instant of next token reception.

If a station receives the token, it performs a

self-diagnostics during the period of Response Time (RT) before any transmission. At the end of RT, the station resets THT to its full value and checks whether any message is waiting in the priority 0 queue. If the queue is empty, the chance of transmission is given to the priority level 1; otherwise, the station starts its THT and begins to transmit the oldest message in the priority 0 queue. At the completion of a message transmission, the station checks whether THT has expired and whether there are more messages waiting in the priority 0 queue. If THT is not expired and the queue is not empty, the station starts another message transmission. This procedure continues either until the queue becomes empty or until THT expires.

After the station finishes the above procedure for the priority level 0 messages, it checks if TRT1 has expired. If it is expired or the priority 1 queue is empty, then TRT1 is reset to its full value and restarted, and the chance of transmission is passed to the priority level 2 messages without any transmission of the priority 1 message. If TRT1 is not expired and the priority 1 queue is not empty, then THT is restarted after being reset to the remaining value of TRT1, and TRT1 is reset to its full value and restarted. The priority level 1 messages can be transmitted consecutively either until THT expires or until there is no message in the priority 1 queue.

When one of two conditions, namely, THT expiration and empty priority 1 queue is satisfied, the station begins the same procedure for TRT2 and the priority 2 queue, and continues for TRT3 and the priority 3 queue. After the station completes the procedure for the priority level 3, the token is passed to the successor station. This priority mechanism is summarized by a flowchart in Fig. A-1.

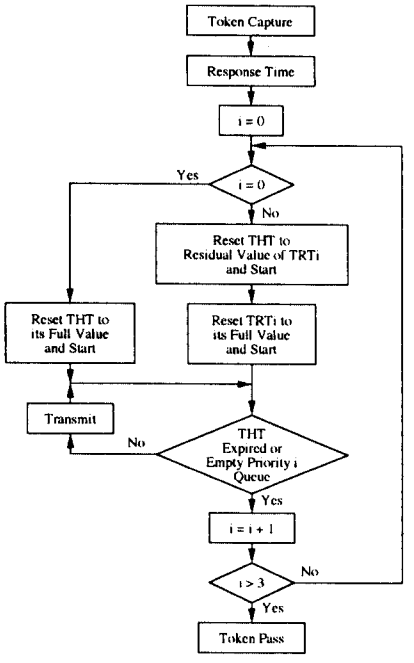

Fig. A-1 Priority Mechanism of LTPB Protocol

Appendix B
Perturbation Analysis for Timer Setting Changes of the LTPB Protocol

The network under consideration in this example consists of three stations. Each station has only one Token Rotation Timer (TRT) in addition to Token Holding Timer (THT) that is used as a dummy timer to store the remaining time of TRT. For simplicity, it is assumed that message transmissions are solely controlled by the status of TRT. Therefore, each station is considered to have only one queue. Fig. B-1 depicts evolution of the network with a nominal TRT setting (nominal path) along with evolution with a perturbed TRT setting (perturbed path). The status changes of the station i, $i=1, 2, 3$, are represented by S_i and S_i' for the nominal and the perturbed paths, respectively. S_i consists of four elements, namely, Message Generation Instant (MSG), Message Transmission Instant (XMT), Token Holding Timer status (THT) and Token Rotation Timer status (TRT). For the perturbed path, S_i' consists of three perturbed elements where MSG is excluded because the instants of message generation are identical for both.

MSG is essentially the instant of message insertion into the queue. It is denoted by a down arrow with an appropriate identifier, $M_i(j)$, for the j-th message at station i. XMT indicates the time interval of transmission and contains two kinds of transmissions. A block with a letter T denotes

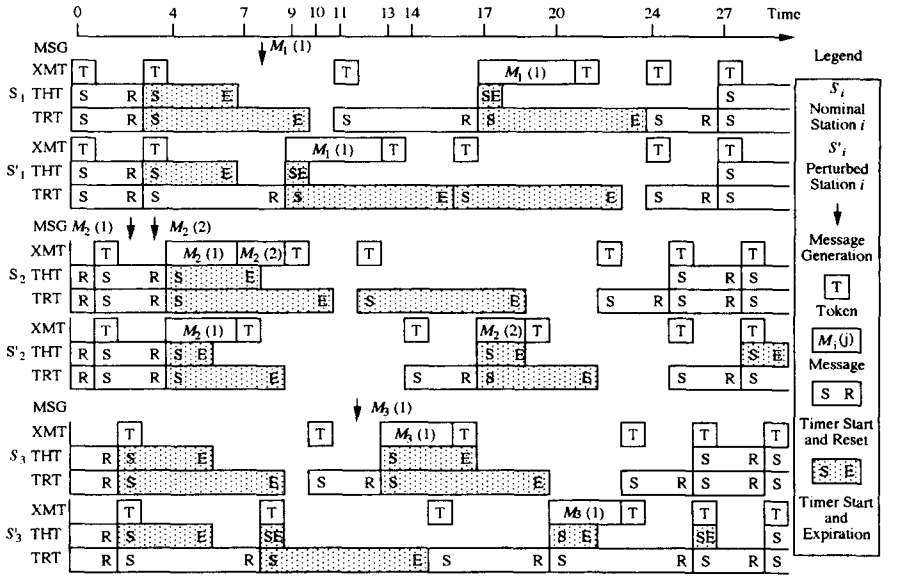

Fig. B-1 Effects of TRT Perturbation on Network Operations

transmission of the token, and message transmission is depicted by a block with an appropriate message identifier. For THT and TRT, a block represents the time period while the corresponding timer counts down, which always starts with a letter S (start). The block ends with a letter R (reset) if the timer is reset before expiration. Otherwise, i.e., if the timer is expired, the block is shaded and ends with a letter E (expire).

Figure B-1 illustrates network operations from the initial time $t = 0$ when none of the three stations have any waiting message(s) and station 1 just received the token. Each token pass is assumed to take one unit of time including the time required for the source station to transmit the token and that for the destination station to respond. The nominal length of TRT is taken to be 7 units of time.

On the nominal path (focusing on S_1, S_2 and S_3 in Fig. B-1), since the queue of station 1 is empty, the token is passed to station 2 immediately after resetting and restarting its THT and TRT. At $t=$ 1, station 2 receives the token. This process continues until station 2 receives the token again at t =4. At this moment, its queue contains two messages and TRT is not yet expired. Therefore, station 2 resets and starts its THT having the remaining time of its TRT (4 units) at disposal, and resets and starts its TRT. Almost simultaneously (it is assumed that timers are reset and started instantaneously), station 2 begins to transmit its first message $M_2(1)$ which takes 3 units of time to be transmitted. When the first message is finished, THT still has one unit of time left and the second message $M_2(2)$ is started. After finishing $M_2(2)$, the token is passed to station 3 which starts its TRT and passes the token to station 1. Upon token arrival at $t = 11$, station 1 has one waiting message. However, its transmission is not allowed because of TRT expiration at $t = 10$ before the token reception. When station 3 receives the token $t = 13$, it can transmit $M_3(1)$ of two units since the message is already in the queue and TRT still has 4 units of time left to its expiration. When the token returns to station 1 at $t = 17$, station 1 finds one unit of time left on its TRT and starts transmitting $M_1(1)$ of four units. After this transmission, the network becomes empty resulting in token circulations without any message transmission.

For the perturbed path (S_1', S_2' and S_3' in Fig. B-1), the TRT in station 2 has been perturbed by -2 units of time while keeping TRT unchanged in stations 1 and 3. This change is solely for the

purpose of illustration since a timer is usually set identically for all stations in standard protocols. Effects of the perturbation do not appear on the perturbed path until $M_2(1)$ is finished at $t = 7$ (compare S_2 and S_2' at $t = 7$). Due to the reduced length of TRT, THT of station 2 has started with only 2 units when the token is received and expires while $M_2(1)$ is being transmitted. Therefore, no further transmission is allowed and $M_2(2)$ should wait for the next opportunity. Due to the deferred transmission, stations 3 and 1 receive the token 2 time units earlier (which is required for transmission of $M_2(2)$) on the perturbed path. At $t = 9$, station 1 has one unit of remaining TRT due to earlier token reception and it can transmit $M_1(1)$. When the token returns to station 2 at $t = 14$, station 2 is again disallowed to transmit $M_2(2)$ because of TRT expiration. Station 3 also loses the opportunity to transmit $M_3(1)$ due to the same reason. Eventually, $M_2(2)$ and $M_3(1)$ are transmitted at $t = 17$ and $t = 20$, respectively.

After transmission of $M_3(1)$ on the perturbed path, the perturbed instant of the token reception by station 1 at $t = 24$ coincides with that on the nominal path even though the status of its TRT is different from each other. This implies that the instants of token reception by a given station on the nominal and the perturbed path become identical after transmission of all messages that are affected by the timer perturbation. The timer status on the perturbed path also becomes identical to that on the nominal path after one more token circulation (from $t = 27$) except station 2 where the timer is perturbed.

It follows from Fig. B-1 that it is impossible to predict the time of $M_2(2)$ transmission on the perturbed path without considering the queue contents and timer status after detecting that $M_2(2)$ cannot be transmitted at $t = 7$ due to the perturbation in TRT. This is because the order of transmissions is dependent on the status of TRT which is, in turn, dependent on the previous token circulation. Therefore, in order to compute the perturbed performance, the only choice is to construct the perturbed path, which is essentially Extended Perturbation Analysis with Brute Force Algorithm (EPA/BFA) (Ho and Li, 1988). However, the brute force construction of the perturbed path is required only for the portion of the perturbed path which differs from the nominal one because the queue contents, timer status and event sequence of the perturbed path become identical to those on the nominal one as shown in Fig. B-1. Therefore, the PA algorithm for this problem needs to check whether the perturbed path begins to differ from the nominal one while the both paths are identical. After a difference between the two paths is detected, the algorithm constructs the perturbed path until it coincides with the nominal one again.