A Fiber-Optic-Based Protocol for Manufacturing System Networks: Part II—Statistical Analysis[1]

A. Ayyagari

A. Ray
Mem. ASME

Mechanical Engineering Department,
The Pennsylvania State University,
University Park, PA 16802

Conceptual development, architecture, statistical and simulation models, and the results of test and verification of a fiber-optic-based protocol have been reported in a sequence of two papers. The protocol uses the unidirectional bus topology and is specifically designed for computer integrated manufacturing (CIM) networks. The architecture of the protocol is formulated in Part I [1]. In this second part, a statistical model of the proposed protocol is formulated for analyzing the network-induced delays and pertinent results of analysis and simulation are presented for different scenarios of network traffic. The major assumptions in formulating the statistical model are: (1) message arrival processes for real-time (RT) and non-real-time (NRT) modes are independent and Markov; and (2) message lengths follow independent general distributions with known statistical properties. The Laplace-Stieltjes transforms of probability distribution functions of queueing delays for RT and NRT messages have been derived. The analytical results for the first two moments of both RT and NRT queueing delays have been compared with those obtained from discrete event simulation.

1 Introduction

Development of a medium access control (MAC) protocol for integrated factory and office communications using a common fiber optic network medium is reported in Part I [1] of this sequence of papers. The protocol was designed on the postulation that the traffic in such networks can be classified into two broad categories of real-time (RT) and non-real-time (NRT) messages on the basis of data latency and data integrity requirements [2, 3].

Discrete-event simulation is a standard tool for evaluation of the protocol performance (e.g., statistical characteristics of network-induced delays) under different operating scenarios but it does not provide a closed form relationship between network traffic parameters and the performance variables. This can be accomplished by statistical modeling using an analytical approach which offers more compactness and better computational efficiency. However, the complexity of the proposed protocol renders an analytical model to be mathematically intractable unless certain simplifying assumptions are made toward operations of the protocol. In general, analytical and simulation models are mutually complementary and together serve as a verification procedure for the protocol performance.

This second part presents an analytical model of the proposed protocol which is derived on the basis of fundamental principles of statistics. The major assumptions in formulating the analytical model are: (1) both RT and NRT message arrival processes are independent and Markov; and (2) both RT and NRT message lengths follow general distributions with known statistics. The Laplace-Stieltjes transforms of the probability distribution functions of RT and NRT queueing delays have been derived. The analytical model serves to verify the simulation model and establish bounds on the (analytically derived) queueing delays for RT and NRT messages.

This paper is organized in five sections including the introduction. Section 2 presents a step by step derivation of the statistical model. Then the analytical results are verified in Section 3 by comparison with simulation results under identical traffic conditions and settings of protocol parameters. In Section 4, the simulation model is used for evaluating the protocol performance under operating conditions which are not restricted by the assumptions made in the analytical model. This sequence of two papers is concluded in Section 5 along with recommendations for future research.

2 Analytical Model of the Proposed Protocol

The MAC protocol, described in Part I [1], establishes conflict free data communications, i.e., access of stations to the network medium is coordinated and controlled in order to prevent more than one station from transmitting their messages at any particular instant. This imposes a queueing delay on both RT and NRT messages. The major performance measure of interest for this protocol is the queueing delays of RT and NRT messages at individual stations. Although other variables such as throughput and bandwidth utilization have been ana-

[1]This work was supported in part by National Science Foundation under Research grant No. DDM-9015173.

Contributed by the Dynamic Systems and Control Division for publication in the JOURNAL OF DYNAMIC SYSTEMS, MEASUREMENT, AND CONTROL. Manuscript received by the Dynamic Systems and Control Division March 27, 1990; revised manuscript received January 1991; Associated Editor: G. E. Young.

lyzed, this section concentrates on the queueing delays which are most difficult to model. The analytical approach is shown for two levels of priority, RT and NRT, but it can be extended to any finite number of priority levels. For example, four levels of prioritization may be necessary if voice and video data share a common channel with RT and NRT data.

It is important to note that the performance metric of mean delay (or even variance of delay) is insufficient for evaluation of medium access protocols as strict deadlines are usually needed in real-time systems [4]. That is, if the data arrives after the scheduled time, it might not have any usage. On the other hand, if the data is available before the scheduled time, it does not matter when this data arrives. This applies to many real-time problems in manufacturing such as multi-robot control systems and scheduling of machining operations. As discussed in [5, 6], if a sensor signal is received after the scheduled sampling instant, the control signal will suffer due to the network-induced delays. Since statistics of delays beyond the mean and variance are deemed necessary for performance evaluation of the protocol from the perspectives of real-time data, we aim at deriving the probability distribution functions of the queueing delays.

The following assumptions and approximations are introduced to formulate a tractable model.

1. All processes incurred in protocol operations are ergodic with probability one.

2. Finitely many active stations exist on the network, i.e., there is an upper bound on the number of active stations in the network.

3. Propagation delay and station response time are negligible, i.e., a carrier injected on the bus is detected by all stations on the medium at the same instant of time.

4. The preemption timer parameter α, described in Section 4 of Part I [1], is set to zero.

5. Each station has two buffers; one for RT messages and the other for NRT messages, and the capacity of each buffer is infinite.

6. The arrival process (for each of RT and NRT messages) at every stration is stationary Poisson and independent of each other and of processes at all other stations.

7. The bulk arrival process for each of RT and NRT messages is modeled by lumping the independent Poisson process at all stations as a single Poisson process with an equivalent arrival rate λ_{RT} and λ_{NRT}, respectively.

8. Message service times (i.e., message length divided by data transmission rate) at the individual stations are independent and identically distributed (i.i.d.) with known probability distribution functions, and these distributions are mutually independent for RT and NRT messages.

The protocol is designed such that RT messages are always transmitted in preference to NRT messages even if the NRT message arrives before the RT message. Within each class, the first-in- first-out (FIFO) policy determines the order of service. Under the *head of the line* discipline if an RT message arrives while an NRT message is being transmitted, then the ongoing transmission is allowed to be completed and any waiting RT message(s) must be served ahead of all NRT message(s) [7].

Remark 1: The effects of the propagation delay and station

Nomenclature

FDDI = Fiber-Distributed-Data-Interface
FIFO = First-In-First-Out
MAC = Medium Access Control
Mbps = Megabits per second
NRT = Non-Real-Time
RT = Real-time
α = preemption timer parameter
λ^i_{RT} = the arrival rate of real-time messages at station i
λ^i_{NRT} = the arrival rate of non-real-time messages at station i
$F_{S_{RT}}$ = probability distribution for real-time message service time
$\mu^{(1)}_{RT}$ = first moment of $F_{S_{RT}}$
$\mu^{(2)}_{RT}$ = second moment of $F_{S_{RT}}$
$\mu^{(3)}_{RT}$ = third moment of $F_{S_{RT}}$
$F_{S_{NRT}}$ = probability distribution for non-real-time message service time
$\mu^{(1)}_{NRT}$ = first moment of $F_{S_{NRT}}$
$\mu^{(2)}_{NRT}$ = second moment of $F_{S_{NRT}}$
$\mu^{(3)}_{NRT}$ = third moment of $F_{S_{NRT}}$
λ_{RT} = average arrival rate of a bulk arrival process of real-time messages as a single Poisson process
λ_{NRT} = average arrival rate of a bulk arrival process of non-real-time messages as a single Poisson process
$S(n,m)$ = state of the queueing process where n is the number of real-time message and m is the number of non-real-time messages
p_{ij} = probability that i real-time and j non-real-time messages arrive during a real-time message service period
q_{ij} = probability that i real-time and j non-real-time messages arrive during a non-real-time message service period
$P(s_1, s_2)$ = probability generating function of p_{ij}
$Q(s_1, s_2)$ = probability generating function of q_{ij}
$\tilde{S}_{RT}$ = Laplace-Stieltjes transform of $F_{S_{RT}}$
$\tilde{S}_{NRT}$ = Laplace-Stieltjes transform of $F_{S_{NRT}}$
p_{RT} = probability of a real-time message arrival
p_{NRT} = probability of a non-real-time message arrival
π_{nm} = stationary probability of occupying the state $S(n, m)$
$\Pi(s,t)$ = probability generating function of π_{nm}
F_B = distribution function of the busy period due to the service of real-time messages
$\tilde{B}$ = Laplace-Stieltjes transform of F_B
$F_{W_{RT}}$ = queueing delay distribution for real-time messages
$\tilde{W}_{RT}$ = Laplace-Stieltjes transform of $F_{W_{RT}}$
ρ_{RT} = expected value of total transmission time of all real-time messages
ρ_{NRT} = expected value of total transmission time of all real-time messages
ρ_{tot} = total sum of the offered due to both real-time and non-real-time messages
W_{NRT} = queueing delay of non-real-time messages
$\dot{W}_{NRT}$ = total time to service all real-time and non-real-time messages that are already in the queueing system
$\ddot{W}_{NRT}$ = total time to service all real-time messages that arrived during $\dot{W}_{NRT}$
$\tilde{W}_{NRT}$ = Laplace-Stieltjes transform of the queueing delay distribution for non-real-time messages
D_{RT} = data latency of a real-time message
D_{NRT} = data latency of a non-real-time message
L = message service time

response time can be compensated by incrementing each message transmission time by the average propagation delay and station response time.

Remark 2: An RT or NRT cycle may often consist of a single message under a very low traffic load. This is equivalent to having the contention period appended to every message, which is analogous to a fixed increase in message overhead. As the network traffic increases, a cycle is likely to be composed of more than one message, and the effects of contention period relative to any one RT (or NRT) message is monotonically decreased with increase in RT (or NRT) traffic. Therefore, a lower bound of the first moment of the queueing delay is obtained by neglecting the contention period and the corresponding upper bound by considering the complete contention period for every message transmission. ■

The queueing process characterized by a Poisson arrival and general service distribution does not have a Markov property. The Markov property is restored by reducing the continuous-time parameter process to discrete time. The discrete-time Markov process is generated by observing the queueing process only at the epochs where an RT or NRT message transmission is terminated. Under above assumptions the protocol is modeled as two interacting M/G/1 queueing processes with the *head of the line* priority scheme following the concept of prioritized queueing delay introduced by Miller [8] and Takács [9]. The analysis focuses on evaluation of the Laplace-Stieltjes transform of the probability distributions of the numbers of waiting RT and NRT messages. These Laplace-Stieltjes transforms are derived under steady-state conditions using imbedded Markov chains with general service distribution.

Remark 3: In the above analysis, the FIFO service discipline policy is more accurate under low network traffic than under high traffic. As the network traffic increases, transmission of messages always starts with the leftmost backlogged station in the network rather than in a FIFO order. Studies on the single priority queueing models have shown that the first moment of the queueing delay is independent of the service order discipline [10]. However, higher order moments of the queueing delay vary with service order disciplines: FIFO has the smallest second moment, followed by random service order, and LIFO service order has the largest. This deviation in the higher order moments becomes particularly significant in the analysis of two interacting priority queues such as RT and NRT. Under high network traffic the service order discipline of the messages is expected to lie between FIFO and LIFO. ■

2.1 An Analytical Approach for Evaluation of Queueing Delaying Statistics. The state of the queueing process is defined as $S(n, m)$ where n is the number of RT messages and m is the number of NRT messages. The transition matrices, $[p_{ij}]$ and $[q_{ij}]$, which are the probabilities that i (≥ 0) RT and j (≥ 0) NRT messages arrive during an RT and an NRT message service period, are obtained as:

$$p_{ij} = \int_0^\infty e^{-(\lambda_{RT}+\lambda_{NRT})t} \frac{(\lambda_{RT} t)^i}{i!} \frac{(\lambda_{NRT} t)^j}{j!} dF_{S_{RT}}(t),\ i \geq 0,\ j \geq 0$$

$$q_{ij} = \int_0^\infty e^{-(\lambda_{RT}+\lambda_{NRT})t} \frac{(\lambda_{RT} t)^i}{i!} \frac{(\lambda_{NRT} t)^j}{j!} dF_{S_{NRT}}(t),\ i \geq 0,\ j \geq 0 \quad (1)$$

where $F_{S_{RT}}(t)$ and $F_{S_{NRT}}(t)$ are the service time distribution functions for RT and NRT messages, respectively.

The probability generating functions [11] of p_{ij} and q_{ij} are defined as

$$P(s_1, s_2) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} p_{ij} s_1^i s_2^j;\ Q(s_1, s_2) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} q_{ij} s_1^i s_2^j \quad (2)$$

where $(s_1, s_2) \in \{|\theta_1| \leq 1,\ |\varphi_2| \leq 1: (\theta_1, \varphi_2) \in C \times C\}$

Using (1) in (2) yields

$$P(s_1, s_2) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \int_0^\infty e^{-(\lambda_{RT}+\lambda_{NRT})t} \frac{(\lambda_{RT} t)^i}{i!} \frac{(\lambda_{NRT} t)^j}{j!} s_1^i s_2^j dF_{S_{RT}}(t)$$

$$Q(s_1, s_2) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \int_0^\infty e^{-(\lambda_{RT}+\lambda_{NRT})t} \frac{(\lambda_{RT} t)^i}{i!} \frac{(\lambda_{NRT} t)^j}{j!} s_1^i s_2^j dF_{S_{RT}}(t) \quad (3)$$

Simplifying (3) yields (as derived in Appendix B of [12])

$$P(s_1, s_2) = \tilde{S}_{RT}\ (\lambda_{RT}(1-s_1) + \lambda_{NRT}(1-s_2))$$
$$Q(s_1, s_2) = \tilde{S}_{NRT}\ (\lambda_{RT}(1-s_1) + \lambda_{NRT}(1-s_2)) \quad (4)$$

where $\tilde{S}_{RT}$ and $\tilde{S}_{NRT}$ are the Laplace-Stieltjes transforms of $F_{S_{RT}}$ and $F_{S_{NRT}}$, respectively.

Let $P[S(n, m) - S(n', m')]$ be the transition probability that the network queueing system moves from $S(n, m)$ to $S(n', m')$ in one transition, i.e., during the period of an RT or an NRT message transmission. The relationships defining the complete probability transition matrix are derived below.

$$P[S(0,0) - S(i, j)] = p_{RT}\, p_{ij} + p_{NRT}\, q_{ij},\ i, j \geq 0 \quad (5)$$

where $p_{RT} = \frac{\lambda_{RT}}{\lambda_{RT} + \lambda_{NRT}}$ is the probability of an RT message arrival and $p_{NRT} = \frac{\lambda_{NRT}}{\lambda_{RT} + \lambda_{NRT}}$ is the probability of an NRT message arrival during one transition.

The transition probabilities from $S(n, m)$ with i RT and j NRT new message arrivals during the service time of an RT and an NRT message are given below by (6) and (7), respectively:

$$P[S(n,m) - S(n-1+i, m+j)] = p_{ij},\ i, j \geq 0,\ n \geq 1,\ \forall m \quad (6)$$

$$P[S(0,m) - S(i, m-1+j)] = q_{ij},\ i, j \geq 0,\ m > 0, \quad (7)$$

Next the impossible events are delineated. The transition from $S(n, m)$ to $S(n', m')$, for $n' < n-1$, and $n > 1$, is impossible because this transition represents exactly one RT transmission. Similarly, such a transition from $m' < m$, and $n \geq 1$ is impossible because an NRT message cannot be serviced while the RT message buffer is not empty. The transition from $S(0, m)$ to $S(n', m')$ for $m' < m-1$ is also impossible because this transition represents exactly one NRT transmission. Therefore,

$$P[S(n, m) - S(n', m')] = 0,\ n' < n-1,\ n > 1, \forall m, m' \quad (8)$$

$$P[S(n, m) - S(n', m')] = 0,\ m' < m,\ n > 1, \forall n' \quad (9)$$

$$P[S(0, m) - S(n', m')] = 0,\ m' < m-1,\ \forall n' \quad (10)$$

Remark 4: $S(0, 0)$ represents the condition when both RT and NRT message buffers are empty. The transition probability of moving to a new state $S(i, j)$ from $S(0, 0)$, given by (5), depends on whether the RT or the NRT message has arrived first. ■

Remark 5: The equilibrium condition guarantees the existence of a stationary distribution for the imbedded Markov chain. ■

The stationary probability of occupying the state $S(n', m')$, denoted as $\pi_{n'm'}$, is given by:

$$\pi_{n'm'} = \int_{n=0}^{\infty} \int_{m=0}^{\infty} \pi_{nm}\, P[S(n, m) - S(n', m')] \quad (11)$$

The probability generating function of $\pi_{n'm'}$ is then defined as

$$\Pi(s, t) = \int_{n'=0}^{\infty} \int_{m'=0}^{\infty} \pi_{n'm'}\, s^{n'} t^{m'} \quad (12)$$

where $(s, t) \in \{|\theta| \le 1, |\varphi| \le 1: (\theta, \varphi) \in C \times C, \forall n', m'\}$.

Substituting (11) into (12) yields

$$\Pi(s, t) = \int_{n'=0}^{\infty}\int_{m'=0}^{\infty}\int_{n=0}^{\infty}\int_{m=0}^{\infty} \pi_{nm}\, P[S(n, m) \to S(n', m')] s^{n'} t^{m'} \quad (13)$$

Following Miller [8] and Appendix B in [12], the Eq. (13) is simplified by using (1), (4) through (10)

$$\Pi(s, t) = [\pi_{00}\,(p_{RT} P(s, t) + p_{NRT}\, Q(s, t) - t^{-1} Q(s,t)) + \int_{m=0}^{\infty} \pi_{0m} t^m\,(t^{-1}\, Q(s,t) - s^{-1}\, P(s,t))]\,[1 - s^{-1} P(s,t)^{-1} \quad (14)$$

Remark 6: π_{00} is the stationary probability that both the RT and NRT buffers are empty. The time epochs when all buffers are empty can be thought of as renewal points in the imbedded Markov chain [8]. π_{0m} is the stationary probability that the RT buffer is empty. This represents the condition at the beginning of an NRT service period. π_{0m} is evaluated by imbedding a second Markov chain by considering only those service termination points from the original Markov chain. The epochs for the second Markov chain are service termination points that leave RT queues empty. On the contrary, the end of every service termination point defines the epoch for the first Markov chain. Hence the second Markov chain is imbedded within the first Markov chain. Therefore, $P[S(0, m) \to S(0, m')]$ is the transition probability of moving from one epoch to the next on the second Markov chain.

Case I: $S(0, m) \to S(0, m')$, $n=0$, $m>0$.

The two events of arrival and nonarrival of any RT messages are considered during the interval between two consecutive epochs to obtain the following transition probability.

$$P[S(m) \to S(m-1+j)] = P[S(m) \to S(m-1+j),\ \text{no RT messages arrived in between}] + P[S(m) \to S(m-1+j),\ \text{RT messages arrived in between}] \quad (15)$$

where

$$P[S(m) \to S(m-1+j),\ \text{no RT messages arrived in between}] = q_{0j} \quad (16)$$

and $P[S(m) \to S(m-1+j)$, RT messages arrived in between$] =$

$$\sum_{l=0}^{j}\sum_{n=1}^{\infty}\left[\int_0^{\infty} e^{-(\lambda_{RT}+\lambda_{NRT})u}\,\frac{(\lambda_{RT}\,u)^n}{n!}\,\frac{(\lambda_{NRT}\,u)^l}{l!}\,dF_{S_{NRT}}(u)\right]\left[\int_0^{\infty} e^{-\lambda_{NRT}v}\,\frac{(\lambda_{NRT}\,v)^{j-l}}{(j-1)!}\,dF_B^{(n)}(v)\right] \quad (17)$$

and F_B is the distribution of the busy period due to the service of only RT messages.

The distribution for the n-fold convolution $F_B^{(n)}$ has been derived by Takács [13]. The probability generating function $P(t)$ of (15) is defined below

$$P(t) = \sum_{j=0}^{\infty} P[S(m) \to S(m-1+j)]t^j,\ m>0 \quad (18)$$

where $t \in \{|\tau| \le 1: \tau \in C\}$

Substituting (16) and (17) into (18) yields the following relationship

$$P(t) = \sum_{j=0}^{\infty} q_{0j}\, t^j + \sum_{j=0}^{\infty}\sum_{l=0}^{j}\sum_{n=1}^{\infty}\left[\int_0^{\infty} e^{-(\lambda_{RT}+\lambda_{NRT})u}\,\frac{(\lambda_{RT}\,u)^n}{n!}\,\frac{(\lambda_{NRT}\,u)^l}{l!}\,dF_{S_{NRT}}(u)\right]\left[\int_0^{\infty} e^{-\lambda_{NRT}v}\,\frac{(\lambda_{NRT}\,v)^{j-1}}{(j-1)!}\,dF_B^{(n)}(v)\right]t^j \quad (19)$$

Following Miller [8] and Appendix B in [12], Eq. (19) is simplified as

$$P(t) = \tilde{S}_{NRT}\,(\lambda_{RT}\,(1 - \tilde{B}(\lambda_{NRT}\,(1-t))) + \lambda_{NRT}\,(1-t)) \quad (20)$$

where $\tilde{B}$ is the Laplace-Stieltjes transform of the distribution function F_B.

Case II: $n=0$, $m=0$, $S(0, 0) \to S(0, j)$

The first arrival in completely empty queues is considered to be an RT or an NRT message to obtain the following transition probability.

$$P[S(0) \to S(j)] = P[S(0) \to S(j),\ \text{first arrival is a RT message}] + P[S(0) \to S(j),\ \text{first arrival is a NRT message}] \quad (21)$$

where $P[S(0) \to S(j)$, first arrival is a RT message$] =$

$$p_{RT}\int_0^{\infty} e^{-\lambda}_{NRT}\,u\,\frac{(\lambda_{NRT}\,u)^j}{j!}\,dF_B(u)] \quad (22)$$

and $P[S(0) \to S(j)$, first arrival is a NRT message$] =$

$$p_{NRT}\sum_{l=0}^{j}\sum_{n=0}^{\infty}\left[\int_0^{\infty} e^{-(\lambda_{RT}+\lambda_{NRT})u}\,\frac{(\lambda_{RT}\,u)^n}{n!}\,\frac{(\lambda_{NRT}\,u)^l}{l!}\,dF_{S_{NRT}}(u)\right]\left[\int_0^{\infty} e^{-\lambda_{NRT}v}\,\frac{(\lambda_{NRT}\,v)^{j-l}}{(j-l!)}\,dF_B^{(n)}(v)\right] \quad (23)$$

The probability generating function $Q(t)$ of (21) is defined in the following

$$Q(t) = \sum_{j=0}^{\infty} P[S(0) \to S(j)]t^j \quad (24)$$

where $t \in \{|\tau| \le 1: \tau \in C\}$.

Substituting (22) and (23) into (24) yields the following relationship

$$Q(t) = p_{RT}\sum_{j=0}^{\infty}\left[\int_0^{\infty} e^{-\lambda_{NRT}u}\,\frac{(\lambda_{NRT}\,u)^j}{j!}\,dF_B(u)\right]t^j + p_{NRT}\sum_{j=0}^{\infty}\sum_{l=0}^{j}\sum_{n=0}^{\infty}\left[\int_0^{\infty} e^{-(\lambda_{RT}+\lambda_{NRT})u} \times \frac{(\lambda_{RT}\,u)^n}{n!}\,\frac{(\lambda_{NRT}\,u)^l}{l!}\,dF_{S_{NRT}}(u)\right]\left[\int_0^{\infty} e^{-\lambda_{NRT}v}\,\frac{(\lambda_{NRT}\,v)^{j-1}}{(j-1)!}\,dF_B^{(n)}(v)\right]t^j \quad (25)$$

Following Miller [8] and Appendix B in [12] Eq. (25) is simplifed as

$$Q(t) = p_{RT}\,\tilde{B}(\lambda_{NRT}\,(1-t)) + p_{NRT}\,\tilde{S}_{NRT}\,(\lambda_{RT}\,(1 - \tilde{B}(\lambda_{NRT}\,(1-t))) + \lambda_{NRT}\,(1-t))$$

The stationary probability of π_{0m}' having m' NRT messages in the queueing system of the second Markov chain is given as

$$\pi_{0m}' = \sum_{m=0}^{\infty} \pi_{0m}\, P[S(m) \to S(m')],\ \forall m' \quad (27)$$

The probability generating function of π_{0m} is denoted by $\Pi(0, t)$.

$$\Pi(0, t) = \sum_{m'=0}^{\infty} \pi_{0m'} t^{m'} \tag{28}$$

where $t \in \{|\tau| \le 1: \tau \in C\}$.

Using (27) in (28) yields

$$\Pi(0,t) = \sum_{m'=0}^{\infty} \sum_{m=0}^{\infty} \pi_{0m} P[S(m) - S(m')] t^{m'} \tag{29}$$

Following Miller [8] and Appendix in [12] further simplification of (29) is obtained as

$$\Pi(0, t) = \pi_{00} [Q(t) - t^{-1} P(t)][1 - t^{-1} P(t)]^{-1} \tag{30}$$

where $P(t)$ and $Q(t)$ are given by (20) and (26), respectively. The first Markov chain and the second Markov chain are initialized with a given value of π_{00}.

Using (28) and (30) in (14) the probability generating function $\Pi(s, t)$ is obtained as

$$\Pi(s, t) = \pi_{00} [1 - s^{-1} P(s,t)]^{-1} \{ \rho_{RT} P(s,t) + \rho_{NRT} Q(s, t) - t^{-1} Q(s, t) + (t^{-1} Q(s, t) - s^{-1} (s, t)) \times (1 - t^{-1} P(t))^{-1} (Q(t) - t^{-1} P(t)) \} \tag{31}$$

where $P(s, t)$, $Q(s, t)$, $P(t)$, and $Q(t)$ are given by (4), (20) and (26), respectively. Taking the initial conditions as $\Pi(s, t) = 1$ at $s = 1$ and $t = 1$, in (31), π_{00} is evaluated as

$$\pi_{00} = 1 - \lambda_{RT} \mu_{RT}^{(1)} - \lambda_{NRT} - \lambda_{NRT} \mu_{NRT}^{(1)} \tag{32}$$

Queueing Delay Distribution

The queueing delay of a message is the time interval between the arrival instant of a message at a queue and the instant of transmission of the first bit. Probability of having n RT messages in the queueing system is given by $\sum_{m=0}^{\infty} \pi_{nm}$.

$$\sum_{m=0}^{\infty} \pi_{nm} = \rho_{RT} \int_0^{\infty} e^{-\lambda_{RT} t} \frac{(\lambda_{RT} t)^n}{n!} d(F_{W_{RT}} * F_{S_{RT}}(t)) + \rho_{NRT} \int_0^{\infty} e^{-\lambda_{RT} t} \frac{(\lambda_{RT} t)^n}{n!} dF_{S_{NRT}}(t) \tag{33}$$

where $F_{W_{RT}}$ is the queueing delay distribution for RT messages. $F_{W_{RT}} * F_{S_{RT}}$ defines the convolution between the queueing delay and service time distribution for RT messages. The probability generating function of (33) defining the probability of having n RT messages in the queueing system follows.

$$\sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \pi_{nm} s'^n = \rho_{RT} \sum_{n=0}^{\infty} \int_0^{\infty} e^{-\lambda_{RT} t} \frac{(\lambda_{RT} t)^n}{n!} d(F_{W_{RT}} * F_{S_{RT}}(t)) s'^n + \rho_{NRT} \int_0^{\infty} e^{-\lambda_{RT} t} \frac{(\lambda_{RT} t)^n}{n!} dF_{S_{NRT}}(t) s'^n \tag{34}$$

The Laplace-Stieltjes transform $\tilde{W}_{RT}$ of the queueing delay distribution $F_{W_{RT}}$ for RT messages is evaluated by setting $s' = (\lambda_{RT} - s)/\lambda_{RT}$, and upon simplification yields the following relationship

$$\tilde{W}_{RT}(s) = \frac{\Pi\left(\frac{(\lambda_{RT} - s)}{\lambda_{RT}}, 1\right) - \rho_{NRT} \tilde{S}_{NRT}(s)}{\rho_{RT} \tilde{S}_{RT}(s)} \tag{35}$$

where $s \in \{|\theta| \le 1: \theta \in C\}$, and $\Pi((\lambda_{RT} - s)/\lambda_{RT}, 1)$ is evaluated from (31) and (32). Substituting this result into (35) yields

$$\tilde{W}_{RT}(s) = \frac{(-1 - \rho_{RT} - \rho_{NRT}) s - \lambda_{NRT} [1 - \tilde{S}_{NRT}(s)]}{\lambda_{RT} - s - \lambda_{RT} \tilde{S}_{RT}(s)} \tag{36}$$

where $\rho_{RT} := \lambda_{RT} \mu_{RT}^{(1)}$ and $\rho_{NRT} := \lambda_{NRT} \mu_{NRT}^{(1)}$. (Note: $\mu_{RT}^{(k)}$ and $\mu_{NRT}^{(k)}$ are the kth moments of RT message service time and NRT message service time, respectively, which are known following assumption #8 at the beginning of Section 2.)

Following Miller [8] and Appendix B in [12], the first two moments of the RT queueing delay distribution evaluated from (36) are given in the following

$$E[W_{RT}] = \frac{\lambda_{RT} \mu_{RT}^{(2)} + \lambda_{NRT} \mu_{NRT}^{(2)}}{2(1 - \rho_{RT})} \tag{37}$$

$$E[W_{RT}^2] = \frac{\lambda_{RT} \mu_{RT}^{(3)} + \lambda_{NRT} \mu_{NRT}^{(3)}}{3(1 - \rho_{RT})} + \frac{\lambda_{RT} \mu_{RT}^{(2)} [\lambda_{RT} \mu_{RT}^{(2)} + \lambda_{NRT} \mu_{NRT}^{(2)}]}{2(1 - \rho_{RT})^2} \tag{38}$$

The queueing delay W_{NRT} of NRT messages is defined as the sum of two independent queueing delays W_{NRT}^{*} and W_{NRT}^{**}, respectively. W_{NRT}^{*} is the total time required to service all RT and NRT messages that are already in the queueing system. W_{NRT}^{**} is the total time required to service all RT messages that arrived during the queueing delay period W_{NRT}^{*}. The probability that the queueing delay for NRT messages is less than or equal to a given value x is given as

$$\begin{aligned} P\{W_{NRT} \le x\} &= P\{W_{NRT}^{*} + W_{NRT}^{**} \le x\} \\ &= P\{W_{NRT}^{**} \le x - W_{NRT}^{*}\} \\ &= P\{W_{NRT}^{**} \le x - y \mid W_{NRT}^{*} = y\} \\ &= \sum_{n=0}^{\infty} \left[\int_0^{\infty} e^{-\lambda_{RT} y} \frac{(\lambda_{RT}(y))^n}{n!} P\{B_n \le x - y\} \right] dP\{W_{NRT}^{*} \le y\} \end{aligned} \tag{39}$$

It is derived in Appendix B of [12] that the Laplace-Stieltjes transform of (39) is

$$\tilde{W}_{NRT}(s) = \tilde{W}_{NRT}^{*}(s + \lambda_{RT}(1 - \tilde{B}(s))) \tag{40}$$

where $s \in \{|\theta| \le 1: \theta \in C\}$.

$\tilde{B}(s)$ is the Laplace-Stieltjes transform of the distribution function F_B of the busy period due to service of RT messages. The Laplace-Stieltjes transform of the steady state queueing delay distribution for a single priority queueing model is defined by Takács [13] as follows.

$$\tilde{W}(s) = \frac{1 - \Lambda E[\sigma_{NRT}]}{1 - \Lambda \left[\frac{1 - \tilde{\sigma}_{NRT}(s)}{s} \right]} \tag{41}$$

where $\Lambda = \lambda_{RT} + \lambda_{NRT}$, and σ_{NRT} is the service time of an NRT message. The Laplace-Stieltjes transform of the distribution function of σ_{NRT} is denoted as $\tilde{\sigma}_{NRT}(s) = \rho_{RT} \tilde{S}_{RT}(s) + \rho_{NRT} \tilde{S}_{NRT}(s)$, and ρ_{RT} and ρ_{NRT} are defined in (4). Using (40) and by change of arguments in (41) it follows that

$$\tilde{W}_{NRT}(s) = \frac{1 - \Lambda E[\sigma_{NRT}]}{1 - \Lambda \left[\frac{1 - \tilde{\sigma}_{NRT}(s + \lambda_{RT}(1 - \tilde{B}(s)))}{s + \lambda_{RT}(1 - \tilde{B}(s))} \right]} = \frac{1 - \rho_{RT} - \rho_{NRT}}{1 - \frac{\sum_{i=RT,\, NRT} \lambda_i [1 - \tilde{S}_i(s + \lambda_{RT}(1 - \tilde{B}(s)))]}{s + \lambda_{RT}(1 - \tilde{B}(s))}} \tag{42}$$

Following Miller [8] and Appendix B in [12], the first two moments of the NRT queueing delay distribution are evaluating from (42) as:

$$E[W_{NRT}] = \frac{\lambda_{RT}\,\mu_{RT}^{(2)} + \lambda_{NRT}\,\mu_{NRT}^{(2)}}{2(1-\rho_{RT})(1-\rho_{RT}-\rho_{NRT})} \tag{43}$$

$$\begin{aligned} E[W_{NRT}^2] = {} & \frac{\lambda_{RT}\,\mu_{RT}^{(3)} + \lambda_{NRT}\,\mu_{NRT}^{(3)}}{3(1-\rho_{RT})^2(1-\rho_{RT}-\rho_{NRT})} \\ & + \frac{[\lambda_{RT}\,\mu_{RT}^{(2)} + \lambda_{NRT}\,\mu_{NRT}^{(2)}]^2}{2(1-\rho_{RT})^2(1-\rho_{RT}-\rho_{NRT})^2} \\ & + \frac{\lambda_{RT}\,\mu_{RT}^{(2)}[\lambda_{RT}\,\mu_{RT}^{(2)} + \lambda_{NRT}\,\mu_{NRT}^{(2)}]}{2(1-\rho_{RT})^3(1-\rho_{RT}-\rho_{NRT})} \end{aligned} \tag{44}$$

The Laplace-Stieltjes transform of RT and NRT queueing delay distributions are given by (36) and (42), respectively. Since these functions are analytic in the region $|s| \leq 1$, it is also possible to evaluate all higher order moments beyond those given in this section provided that sufficient statistics of RT and NRT message service distributions are available. Furthermore, the distribution functions of RT and NRT queueing delays can be evaluating by taking the inverse of the respective Laplace-Stieltjes transforms (36) and (42), respectively.

Remark 7: For evaluation of the kth moments of RT and NRT queueing delays, the first $(k+1)$th moments of RT and NRT message service times are needed.

3 Comparison of the Statistical Analysis and Simulation Results

The results derived from the statistical model are presented in this section for comparison with those of the discrete-event simulation model of the protocol. The simulation model is coded in Fortran in the environment of SIMAN [14], and is structurally similar to the timed Petri net model of the protocol reported in [12]. The objectives are to determine the range of validity of the statistical model and establish credibility of the simulation model.

The simulation model is designed such that, during each run, it generates the following information:

- Number of real-time and non-real-time message transmissions;
- Expected value, standard deviation, and other cumulative relative and absolute frequency statistics of data latency for both RT and NRT messages;

The performance of the proposed protocol is expressed in terms of both RT and NRT traffic in the network. To this effect, the following definitions are introduced.

Definition 1: Offered traffic ρ_{RT} (or ρ_{NRT}) is defined as the expected value of total transmission time of all RT (or NRT) messages that arrive at the network per unit time. ■

Remark 8: ρ_{RT} and ρ_{NRT} can be expressed as

$$\rho_{RT} = \lambda_{RT}\,\mu_{RT}^{(1)} \text{ and } \rho_{NRT} = \lambda_{NRT}\mu_{NRT}^{(1)}$$

where λ_{RT} (or λ_{NRT}) is the total average arrival rate of RT (or NRT) messages in the network, $\mu_{RT}^{(1)}$ (or $\mu_{NRT}^{(1)}$) is the average transmission time for RT (or NRT) messages. ■

Definition 2: Total offered traffic ρ_{tot} in the network is defined as the sum of the offered traffic due to both RT and NRT messages, i.e., $\rho_{tot} = \rho_{RT} + \rho_{NRT}$. ■

Remark 9: The assumption of no RT and NRT message rejection at any station implies that:

- Throughput S is identically equal to offered traffic G, i.e., $S_{RT} = \rho_{RT}$ and $S_{NRT} = \rho_{NRT}$.
- The network is stable, i.e., the total offered traffic ρ_{tot} is less than the critical offered traffic [15]. ■

Definition 3: Queueing delay W_{RT} (or W_{NRT}) is the time interval between the instant of arrival of an RT (or NRT) message at the transmitter queue of the source terminal and the instant of transmission of its first bit. ■

Definition 4: Data latency D_{RT} (or D_{NRT}) is the time interval between the instant of arrival of an RT (or NRT) message at the transmitter queue of the source terminal and the instant of reception of its last bit at the destination terminal. ■

Remark 10: Neglecting the effects of propagation delay, queueing delay and data latency are related as $D = W + L$ where L is the message service time. Therefore, if L is a constant, then all higher order central moments of W and D are identical.

The performance analysis of the protocol is focused on the statistical characteristics of both D_{RT} and D_{NRT}. Simulation results are generated under the conditions that conform to the assumptions introduced in Section 2 for development of the statistical model. Five different cases with ρ_{RT} equal to 0.01, 0.05, 0.10, and 0.20 are considered. In each case ρ_{RT} was held constant and ρ_{NRT} was increased until ρ_{tot} reaches 0.6. The protocol and traffic parameters used for the performance analysis are given as follows.

- Transmission medium bandwidth capacity = 100 Mbps.
- Preemption timer parameter: α = 0. [See Section 4 of Part I [1]].
- Propagation delay and station response time = 0.
- Number of stations: #RT = #NRT = 5.
- Constant lengths of: 1000 bits for each RT message and 75000 *bits* for each NRT message, i.e., $\mu_{RT}^{(1)} = 10^{-5}\ s$, $\mu_{NRT}^{(1)} = 75\times10^{-5}\ s$, and $\mu_{RT}^{(2)} = \mu_{NRT}^{(2)} = \mu_{RT}^{(3)} = \mu_{NRT}^{(3)} = 0$.

Figures 1 to 5 exhibit a comparison of the analytical and simulation results for the cases 1 to 5, with five different values of ρ_{RT}, respectively. The expected values and standard deviations of D_{RT} and D_{NRT} are given for each case as the total offered traffic ρ_{tot} is increased. In Figs. 1 to 5 the analytical results are represented by solid lines for RT messages and by dotted lines for NRT messages. The simulation model results are signified at discrete points by plus (+) and triangle (Δ) symbols for RT and NRT messages, respectively.

Figures 1 to 5 show that the expected values generated from the analytical and simulation models are in close agreement for both RT and NRT messages in the entire range of ρ_{tot}. Whereas the standard deviations of D_{RT} generated from the two models are in agreement, those of D_{NRT} tend to deviate for NRT messages as ρ_{tot} increases. The reason for inaccuracy of the analytical model at higher traffic is explained below.

At low traffic the RT and NRT queues remain empty most of the time. Therefore, RT and NRT cycles consist of a single or very few message transmissions. As the traffic increases, the number of transmissions increase during a cycle and the service discipline in the lumped RT and NRT queues may not be FIFO. In the analytical model the service discipline is always assumed to be FIFO whereas the simulation model correctly represents the service discipline in a particular cycle which always begins with the leftmost backlogged station in the network. The standard deviation of queueing delay (and, therefore, of data latency) in the analytical model is expected to be lower than its actual value because the FIFO service discipline has the smallest second moment in a M/G/1 queue whereas the expected value of queueing delay is independent of the service discipline [10]. Therefore, the analytical model underestimates the second and higher order moments of data latency and the error increases with the offered traffic.

Figures 1 to 5 also indicate that both expected value and standard deviation of D_{NRT} increase as ρ_{RT} is increased. This happens because a larger part of the network bandwidth is utilized for RT message transmissions. The reduction in the available bandwidth results in increased D_{NRT}. A relatively

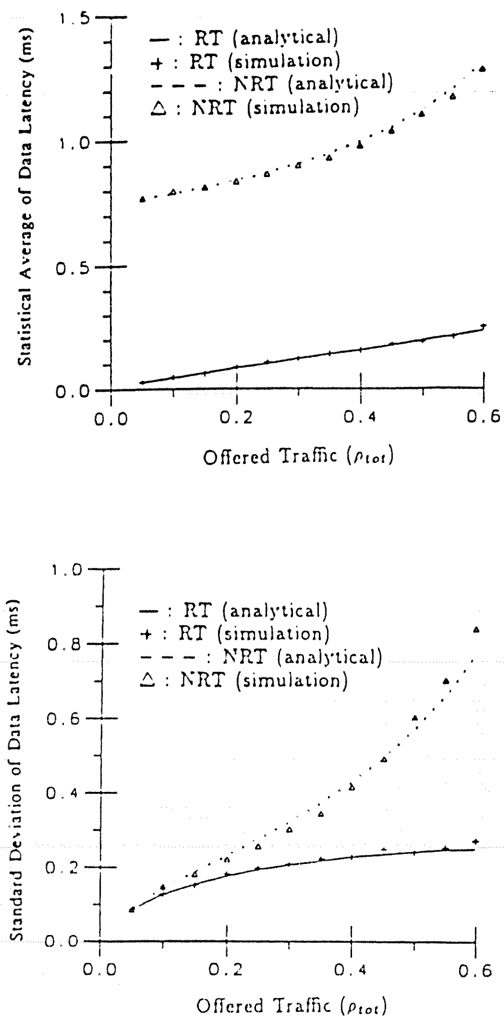

Fig. 1 Analytical and simulation results ($\rho_{RT} = 0.01$)

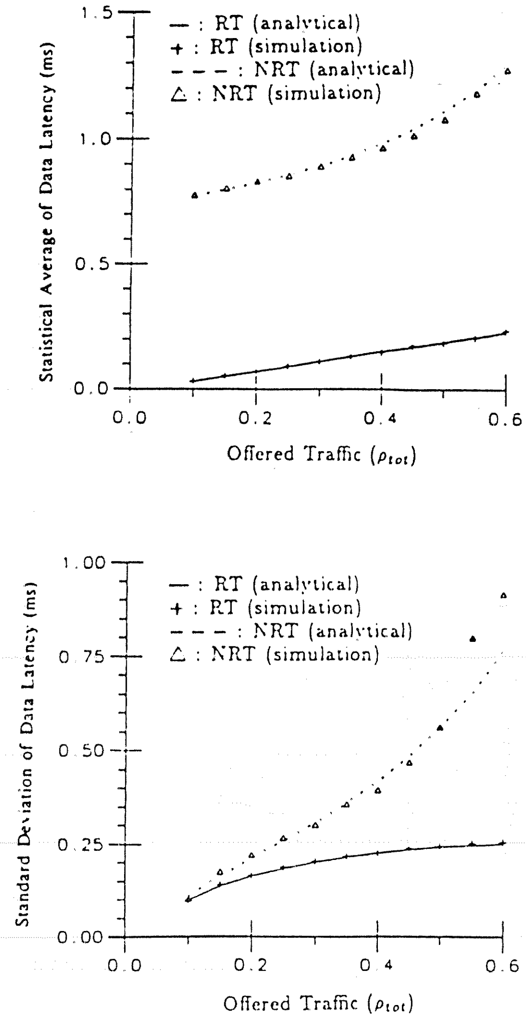

Fig. 2 Analytical and simulation results ($\rho_{RT} = 0.05$)

small increase in D_{RT} occurs as ρ_{tot} is increased. This increase accrues from a larger probability that an arriving RT message would have to wait for completion of an ongoing NRT message transmission.

The following conclusions are derived from the comparison of analytical and simulation results.

- Although the analytical model is formulated on the basis of certain approximations, it is in close agreement with the simulation model at low traffic; the standard deviation (and possible all higher moments) of the NRT message data latency tends to become inaccurate as ρ_{tot} is increased.
- The agreement of the analytical and simulation models establishes credibility of the simulation model in the sense that representations of protocol operations in the simulation are essentially independent of offered traffic whereas the analytical model is formulated on the basis of certain assumptions that are valid at low traffic. Nevertheless, the analytical model can be used for evaluating the data latency bounds if the propagation and station response delays are not neglected. This is implemented in the model by increasing the transmission time of every message by a constant to account for propagation delay, response time, and contention period.

4 Simulation Results Under a Specific Scenario

The proposed protocol is simulated for its performance evaluation under a specific scenario within an integrated office and factory communication environment. The allowable bounds for data latency of RT messages are determined by the functional characteristics of the controlled processes. A few examples where network-induced delays could be detrimental follow.

Within an integrated manufacturing environment, networking provides flexibility for coordinated control of inter-cell equipment. For example, in an intelligent welding system [16, 17] where a positioning table and a robot may not be hardwired to the same computer, the table position coordinates could be relayed to the robot controller via the network. The robot controller, in turn, may transmit back signals for a more convenient table position. This requires timely arrivals of the data at both machines. Another example is the coordinated control of two robots in a master-slave configuration while handling a bulky workpiece together. If these two robots do not belong to the same supervisory computer, they will communicate via the network so that the slave robot follows a prescribed trajectory. The timeliness of the transmitted data is essential because a delay could damage the workpiece or the robot's wrists and arms. Another example is a machine tool transfer line where several machines are assisted by robots for loading, unloading and handling of materials and parts. The timeliness of interrupt signals arriving at individual machines is critical for successful operations.

Accurate delivery of NRT messages, such as CAD drawings, material inventory and payroll files, is essential although they do not have stringent data latency requirements. The integrity of the message delivery is achieved by acknowledgement schemes that are usually implemented at higher layers in the protocol suite.

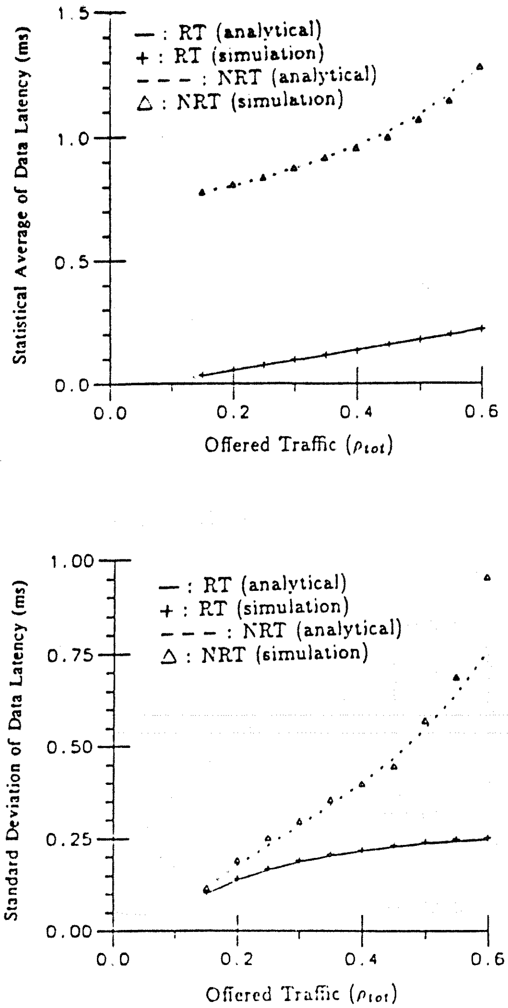

Fig. 3 Analytical and simulation results (ρ_{RT} = 0.10)

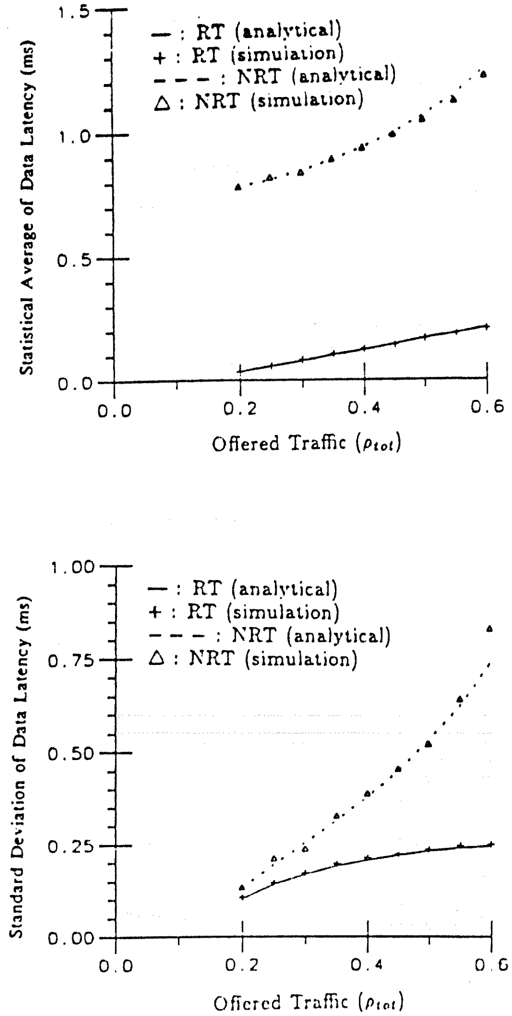

Fig. 4 Analytical and simulation results (ρ_{RT} = 0.15)

The network traffic, resulting from typical office and factory operations, is not likely to conform to the assumptions of independent Poisson processes used in the analytical model. Therefore, performance of the protocol is evaluated by simulation where two different cases of settings α = 0.0 and α = 0.5 of the preemption timer (defined in Section 4 of Part I [1]) have been considered in the following traffic scenario.

The manufacturing scenario is made up of five virtual cells where each cell is composed of a number of RT and NRT processes. The RT messages from machine tools and devices are classified into two categories: (1) periodically generated sensor and control signals; and (2) interrupt signals which may arrive in bursts and their inter-arrival time is assumed to have an exponential distribution. On the average, each cell consists of 20 machines. Each of these machines periodically generate 10 different sensor and control signals every 100 ms and randomly generate interrupt signals with a mean inter-arrival time of 10 ms. The sampling instants of the periodic signals are uniformly distributed with the period of 100 ms. On the average, one periodic and one random RT messages are generated by each machine at every 10 ms. The RT message lengths are usually very short and individual messages may vary in length in an actual manufacturing environment. However, for simplicity, the RT message length and data latency bound are taken to be 1000 bits and 10 ms, respectively, and ρ_{RT} is maintained at 0.20 in this simulation example.

The performance of the proposed protocol is evaluated by varying ρ_{NRT} while ρ_{RT} is held constant. NRT messages are usually much longer than RT messages, and widely vary in lengths. In this example, NRT message lengths are assumed to be uniformly distributed between 25000 and 75000 bits. The arrival process of NRT messages is normally bursty and, therefore, it is assumed to be Poisson in the simulation example. The average arrival rate of NRT messages are adjusted as ρ_{tot} is increased.

The network parameters used in the simulation of the protocol under the above scenario are summarized below.

- ρ_{RT} = 0.20 and ρ_{tot} is increased up to 0.6 by increasing ρ_{NRT}.
- Number of stations in the network: #RT = 5 and #N RT = 5.
- Data transmission rate in the fiber optic medium = 100 M bps, RT transmission time (constant) = 0.01 ms, and NRT transmission time distribution is uniform in (0.25, 0.75) ms.
- Real-time message data latency bound = 10.0 ms.
- Length of the network bus = 2 km, one way propagation delay = 0.01 ms.
- Station response time = 0.001 ms.
- Propagation delay between adjacent stations = 0.002 ms.
- End of cycle unjamming pattern = 5 bits, i.e, 0.00005 ms.

Figures 6 and 7 illustrate how the expected values of D_{RT} and D_{NRT} change as ρ_{tot} is increased for Case I (α = 0.0) and Case II (α = 0.5), respectively.

Table 1 shows the cumulative relative frequency of the number of transmissions whose data latency does not exceed the

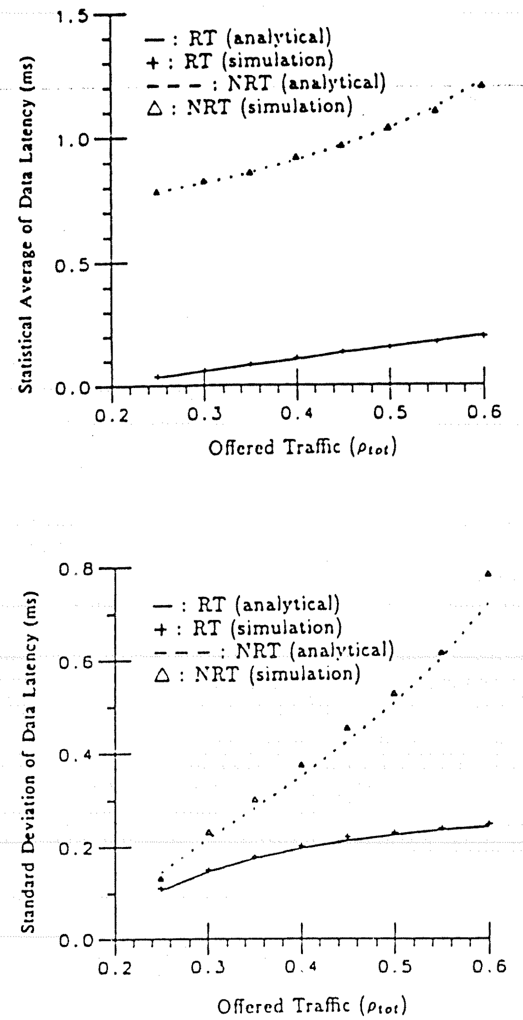

Fig. 5 Analytical and simulation results ($\rho_{RT} = 0.20$)

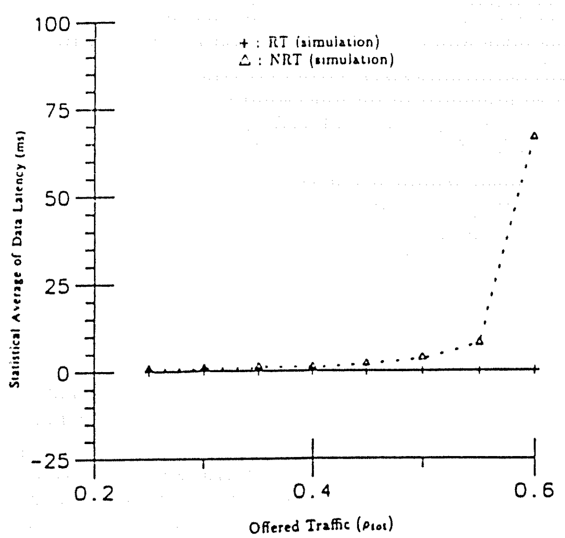

Fig. 6 Specific Scenario simulation results (Case I: $\alpha = 0.0$)

specified bounds for RT and NRT messages under different values of ρ_{tot} in Cases I and II, respectively.

RT message arrival is a mixed periodic and exponentially distributed process whereas NRT message arrival is a Poisson process. The knowledge of the frequency of RT messages for which the data latency does not exceed the specified bound is necessary to ascertain dynamic performance and stability of real-time processes. This implies the level of confidence with which the proposed protocol can maintain the data latency of messages below a certain limit at network traffic loads up to a certain level (e.g., $\rho_{tot} = 0.60$). This knowledge, although not essential for non-real-time operations, is useful for estimating the optimal buffer size to store the waiting NRT messages.

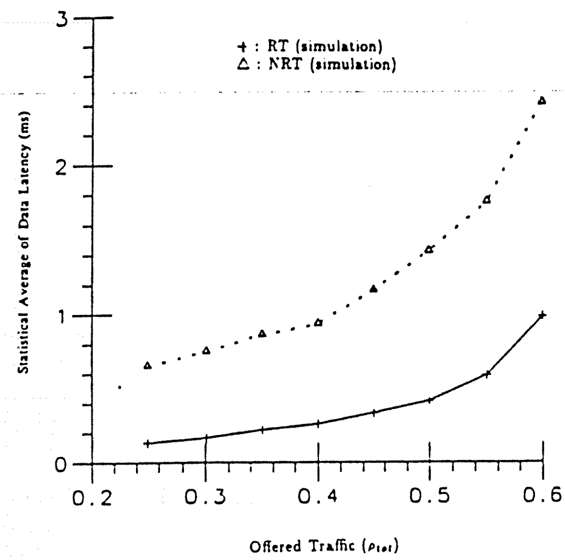

Fig. 7 Specific Scenario simulation results (Case II: $\alpha = 0.5$)

Table I Confidence level of message data latency

Case I: ($\alpha = 0.0$)

Real-time data latency bound = 0.69 ms.
Non-real-time data latency bound = 154.00 ms.

ρ_{tot}	Real-time	Non-real-time
0.25	0.9954	1.0
0.30	0.9898	1.0
0.35	0.9796	1.0
0.40	0.9762	1.0
0.45	0.9691	1.0
0.50	0.9633	1.0
0.55	0.9544	1.0
0.60	0.9503	0.9511

Case II: ($\alpha = 0.5$)

Real-time data latency bound = 3.50 ms.
Non-real-time data latency bound = 7.05 ms.

ρ_{tot}	Real-time	Non-real-time
0.25	1.0	1.0
0.30	1.0	1.0
0.35	1.0	1.0
0.40	1.0	1.0
0.45	0.9991	0.9990
0.50	0.9891	0.9940
0.55	0.9891	0.9876
0.60	0.9521	0.9517

Case I: Preemption Parameter $\alpha = 0.0$ Figure 6 shows that the expected value of D_{RT} is practically independent of ρ_{tot} whereas that of D_{NRT} gradually increases with ρ_{tot} and this increase becomes very sharp beyond $\rho_{tot} = 0.55$. Since $\alpha = 0.0$, an arriving RT message immediately preempts the bus if the protocol is executing in the NRT mode and transmits the backlogged RT message (after allowing for completion of any ongoing NRT message transmission). This results in significant reduction of D_{RT}. The periodic preemption and the additional overhead due to contention of RT cycles restrict the protocol operation in the NRT mode, resulting in delayed transmission of NRT messages. Therefore, the expected value of D_{NRT} follows an approximately exponential profile as ρ_{tot} is increased.

Similar results are expected from token passing bus and FDDI protocols under high traffic load conditions [18]. If there exists processes with very fast dynamics i.e, short data latency bound on RT messages, NRT messages would be significantly delayed and the NRT throughput be severely limited because the target token rotation timer (TTRT) must be set to a small enough value to satisfy the requirement for D_{RT}.

Case II: Preemption Parameter $\alpha = 0.5$ Figure 7 shows how the expected values of D_{RT} and D_{NRT} increase with increase in ρ_{tot}. Unlike Case I in Fig. 6, D_{RT} does not remain constant and D_{NRT} does not rise sharply. At low ρ_{tot} the setting of $\alpha = 0.5$ does not significantly influence the RT message transmissions because the bus activity timer dominates over the preemption timer. As ρ_{tot} increases, the preemption timer parameter becomes more effective and D_{RT} tends to increase. With an increase in ρ_{tot} the setting of $\alpha = 0.5$ reduces the effect of periodic preemption by RT messages and RT contention period overheads. Compared to Case I, a larger amount of bandwidth is available for NRT message transmission, which results in decreased D_{NRT}.

Cases I and II, discussed above, indicate the effectiveness of the proposed protocol for timely delivery of RT messages and also of NRT messages if the medium bandwidth is used for RT transmissions. Case I with $\alpha = 0.0$ is an extremely conservative design for network traffic allocation. Therefore Case II with $\alpha = 0.5$ was considered to illustrate the effectiveness of the preemption timer in the proposed protocol for a tradeoff between a small increase in D_{RT} and a large decrease in D_{NRT}. An optimal value of α depends on both the performance index and the statistical characteristics of RT and NRT traffic.

The role of the preemption timer is, to some extent, analogous to that of the target token rotation timer (TTRT) in the FDDI protocol. However, unlike TTRT which always affects the network performance, α influences the network performance only when it is necessary, e.g., when the network system is backlogged with RT message(s) under medium or high traffic conditions. This aspect is particularly important if the network is required to handle RT messages with extremely short data latency bounds as it is the case with multi-robot control systems in inter-cell operations.

5 Summary, Conclusions, and Recommendations for Future Research

In this two-part paper we have developed and investigated the performance of a fiber-optic-based medium access control (MAC) protocol for computer integrated manufacturing (CIM) networks. An architecture of the protocol that is suited for heterogeneous traffic in factory and office environments has been proposed. A finite-state-machine model of the protocol has been constructed to describe its operations in detail. A discrete-event simulation model of the protocol has also been developed to investigate the performance, i.e., delay and throughput of both real-time and non-real-time messages under different traffic scenarios.

The performance of the protocol has been statistically analyzed under the following operating conditions: (1) Poisson distribution of message arrival; (2) general distribution of message length with known statistical properties; (3) infinite queue capacity and stable operations; and (4) the preemption timer parameter α set to zero. The principal assumptions used in developing the analytical model are negligible propagation and station response delays, FIFO service order discipline, and ergodicity of all processes with probability one. The analytical model results were compared with those of the simulation model under identical operating conditions from (1) to (4). The conclusions are as follows.

- Accuracy of the analytical model is consistently good for expected values of both real-time and non-real-time message data latency over a wide range of offered traffic;
- Accuracy of the analytical model degrades for second and higher moments of non-real-time message data latency with increase in the offered traffic beyond a certain level. This inaccuracy is attributed to the assumption of FIFO service order discipline in the analytical model. However, the simulation model which represents the actual operations of the protocol does not suffer from this inaccuracy.

The simulation model is further used to evaluate the performance of the protocol under a specific scenario that serves as a simplified example for data communication requirements in an integrated manufacturing system. The performance analysis of the protocol under the specific scenario shows that real-time messages are allowed to use the transmission medium bandwidth, and the unused bandwidth is available to the non-real-time traffic. The results from the two cases (having the preemption timer setting $\alpha = 0.0$ and 0.5) reveal that, in order to take advantage of the efficacy of the proposed protocol at high total offered traffic, an optimal value of α should be selected.

5.1 Areas of Future Research.

The future areas of research that will enhance performance analysis of the proposed protocol are delineated in the following.

- Development of an analytical model with a non-zero propagation delay and the actual service order discipline. A closed form analytical solution for the above model may not be achievable. However, the system state may be expressed by a set of recursive relationships in terms of its previous states. In that case, this analytical model shall not yield a closed form solution but the effects of the protocol and network parameters on the performance can be quantitatively evaluated. Furthermore this analytical model under normal circumstances could be used in lieu of extensive discrete-event simulation.
- Extension of the above developed analytical model for non-zero values of α.
- Development of a methodology for evaluating an optimal value of α for a given network operating condition. The first step in evaluating an optimal value of α is to identify a performance index. In case of real-time messages a weighted sum of expected value and standard deviation of the data latency may be considered as a performance index. The minimization of the first moment of the data latency is an obvious criterion for real-time messages, and that of standard deviation ensures that the time-varying nature of the network-induced delays is reduced as much as possible. The maximization of the throughput may also be included within the performance index for non-real-time messages. Having defined a performance index, the analytical model discussed above or regenerative simulation techniques [19] could be used to evaluate the optimal value of α.
- Development of network management tools for tuning the network parameters (e.g., α and frame lengths for RT and NRT messages) for efficient performance of the overall network. Techniques such as perturbation analysis [20, 21] along with a performance index defined above may be used for online tuning of the protocol and network parameters under actual network operating conditions.
- Extending the protocol model to more than one level of priority in NRT mode. This is done in the asynchronous mode of the FDDI. In the proposed protocol only one level of priority has been considered in the NRT mode and this can be extended to any finite number. The statistical model of priority queue in Section 2.1 allows for this modification.
- Application of the proposed protocol for fast dynamical processes like aircraft control. In this case the performance

of the proposed protocol needs to be evaluated against those of the protocols used in networking for process control (e.g., combined mode protocol (CP) of ARINC 629 used in commercial aircraft [22]).

- Development of fault detection, isolation and reconfiguration (FDIR) strategies for on-line operations of the protocol. This possible failures in a unidirectional dual bus network architecture can be classified into four types given below.
  - Cable breakdown;
  - Timer failure at a station;
  - Transmitter failure at a station (e.g., passive and babbling mode failure).
  - Receiver failure at a station.

The FDIR strategies should be adapted into the discrete-event simulation model to evaluate their performance and the resulting recovery delays.

References

1 Ayyagari, A., and Ray, A., "A Fiber-Optic-Based Protocol for Manufacturing System Networks: Part I—Conceptual Development and Architecture," published in this issue pp. 113-120.

2 Ray, A., and Phoha, S., "Research Directions in Networking for Manufacturing Systems," ASME *Journal of Engineering for Industry*, Vol. 111, No. 2, May 1989, pp. 109-115.

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

4 Bhattacharya, P. P., and Ephremides, A., "Optimal Scheduling with Strict Deadlines," *IEEE Trans. Aut. Contr.*, Vol. 34, No. 7, July 1989, pp. 721-728.

5 Halevi, Y., and Ray, A., "Integrated Communication and Control Systems: Part I—Analysis," ASME Journal of Dynamic Systems, Measurement, and Control, Dec. 1988, pp. 367-373.

6 Ray, A., and Halevi, Y., "Integrated Communication and Control Systems: Part II—Design Considerations," ASME Journal of Dynamic Systems, Measurement, and Control, Dec. 1988, pp. 374-381.

7 Cobham, A., "Priority Assignment in Waiting Line Problems," *Journal of Operations Research Society of America*, Vol. 2, 1954, pp. 70-76.

8 Miller, R. G., "Priority Queues," *The Annals of Mathematical Statistics*, Vol. 31, No. 1, Mar. 1960, pp. 86-103.

9 Takács, L., "Investigation of Waiting Time Problems by Reduction to Markov Process," *Acta Math., Acad. Sci. Hung.*, Vol. 6, 1955, pp. 101-129.

10 Cohen, J. W., *Single Serve Queues*, North Holland Publishing Company, Amsterdam, 1969.

11 Kleinrock, L., *Queueing Systems Volume I: Theory*, John Wiley, New York, 1975.

12 Ayyagari, A., "A Fiber-Optic-Based Protocol for Manufacturing System Networks," Ph.D. dissertation, Mechanical Engineering Department, The Pennsylvania State University, PA, 1989.

13 Takács, L., *Introduction to the Theory of Queues*, Oxford University Press, 1962.

14 Pegden, C. D., *Introduction of SIMAN, System Modeling Corporation*, State College, PA, 1986.

15 Ray, A., "Distributed Data Communication Networks for Real-Time Process Control," *Chemical Engineering Communications*, Vol. 65, March 1988, pp. 139-154.

16 Thompson, D. R., Ray, A., and Kumara, S., "A Hierarchically Structured Knowledge Based System for Welding Automation and Control," ASME *Journal of Engineering for Industry*, Feb. 1988, pp. 71-76.

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

18 Dykeman, D., and Bux, W., "Analysis and Tuning of the FDDI Media Access Control Protocol," *IEEE Journal on Selected Areas in Communications*, Vol. SAC-6, No. 6, July 1988, pp. 997-1010.

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

20 Ho, Y.-C., "Performance Evaluation and Perturbation Analysis of Discrete Event Dynamic Systems," *IEEE Trans. Aut. Contr.*, Vol. AC-32, No. 7, July 1987, pp. 563-572.

21 Suri, R., "Perturbation Analysis: The State of the Art and Research Issues Explained via the GI/G/1 Queue," *Proceedings of the IEEE*, Vol. 77, No. 1, Jan. 1989, pp. 114-137.

22 ARINC Specification 629 Multitransmitter Data Bus: Part I—Protocol Description, Aeronautical Radio, Inc., 2551 Riva Road, Annapolis, MD 21401.