Proceedings of the 1986 Winter Simulation Conference
J. Wilson, J. Henriksen, S. Roberts (eds.)

TRANSIENT ANALYSIS OF A STORE-AND-FORWARD COMPUTER-COMMUNICATIONS NETWORK

Albert B. Garcia
Wade H. Shaw
School of Engineering
Air Force Institute of Technology
Wright-Patterson AFB, OH 45433, U.S.A.

ABSTRACT

This paper presents the results of a study of transient behavior in a store-and-forward, computer-communications network. The purpose of this paper is to discuss the nature of network performance during severe loading of a network previously operating at steady state conditions. This knowledge is critical since networks are generally designed for a maximum steady state message load. When a transient situation develops, the network may not perform as originally intended and may produce excessive message delay.

The primary problem addressed in this paper is the unknown and mathematically intractable relation between the occurrence of a transient traffic load and the resulting performance. The network design and topology are fixed with the experimental factors being the type of transient and the distribution of transient traffic precedence.

A five-node, computer-communications network is simulated under two types of transients: first, the effect of a sudden burst of message traffic and second, the effect of a sudden reduction in the interarrival time for a specified period. Also, the effect of the combined transients is investigated. The data is analyzed using ANOVA techniques for a two factor experiment design with four replications per experimental unit.

It is concluded that transient message loads have a significant and severe effect on network performance. The degradation in performance persists for a significant period of time. The transient effect is not consistent across each precedence level and erroneous conclusions on average network performance result if the interaction of transient traffic precedence and type of transient are not considered.

1. INTRODUCTION

All too often computer communication networks evolve in response to a current need without regard for future requirements. As the network grows in size, the complexity increases quickly beyond simple understanding. Sometimes computer networks are planned assuming that existing communication channels are adequate for future loads. This attractive assumption is made by computer network designers who have little or no control over communication resources. Adequate tools are needed to evaluate network performance, to determine if performance will meet objectives and to understand how performance can be improved.

The analysis of network performance is complicated by difficult and often unknown mathematical relations between attributes of network design. Few closed form solutions exist which allow analysis of queuing theory applied to communications networks. Generally, approximations and heuristics are used to guide the engineering of a network consisting of communications channels, node processors, routing algorithms and an array of communications protocols. The general rule is to design the network for maximum projected message traffic. This rule assumes the traffic patterns are known and that the network will not experience excessive loading during normal operation.

Extensive modeling and analysis have revealed basic steady state relationships such as the effect of traffic load on average message delay. Figure 1 depicts this basic relation for a typical network. It is apparent that severe message delay results when a predictable threshold of traffic is reached. This paper discusses the performance of the network when it is subjected to sudden changes in traffic load.

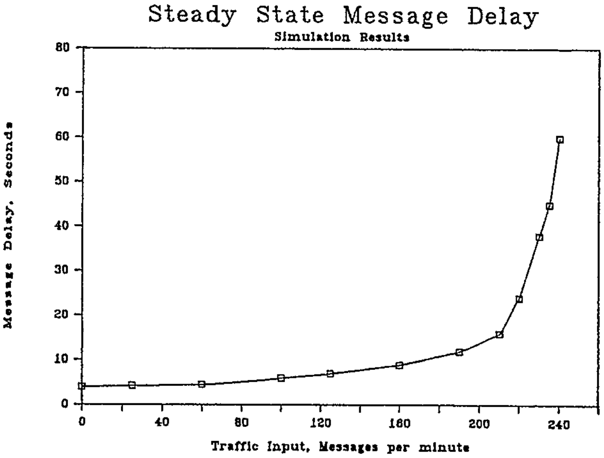

Figure 1: Steady State Message Delay

Specifically, what performance should be expected during the transition from one traffic load to a higher load over a short period of time? Clearly, the message delay may reach very high and unacceptable values.

1.1. Statement of the Problem

The primary problem addressed in this paper is the unknown and mathematically intractable relation between the occurrence of a transient traffic load and the resulting network performance. It is assumed that the engineering specifications for the network are known, that the steady state solutions can be obtained, and that the controllable factors of interest include the type and intensity of the transient, and the characteristics of the message traffic such as message lengths, message arrival rates and precedence levels during the transient.

The performance of the network is measured in three ways. First, the average delay of a message during the transient period is measured. Second, the variability of the message delay during the transient period is used as a measure of network stability. Finally, the use of finite queues at each node processor imply that nodes operating at saturation must refuse service to some messages. The number of messages not transmitted is clearly critical.

An additional performance measure of importance is the length of time the transient effect degrades network behavior. This statistic is quite difficult to measure and was not included in this study. The difficulty is primarily the separation of the transient effect from the ambient traffic once the transient has diminished. The highly variable ambient traffic characteristics mask the transient and makes the measurement of a settle time subject to error.

The occurrence of a transient condition can be caused in two distinct ways. First, the message interarrival time could suddenly reduce to a new level due to a sudden requirement to transmit messages by each node on the network. Second, the existence of messages which are suddenly generated as a result of the tranmission of a coded message which precipitates follow-on traffic. This transient is common in military networks where a sender can suddenly transmit a group of messages to predefined receivers. Certainly, both effects could also combine to yield a third transient traffic load.

The impact of a transient can best be observed by considering a hypothesized model shown in Figure 2. The steady state delay is suddenly interrrupted at time T by one or both of the transient components. The fundamental issue is what will the transient do to network performance, and how long will it last? Figure 2 denotes the known steady state solution at the new traffic intensity as well as possible message delay curve for each transient component and for the combined effects of both components.

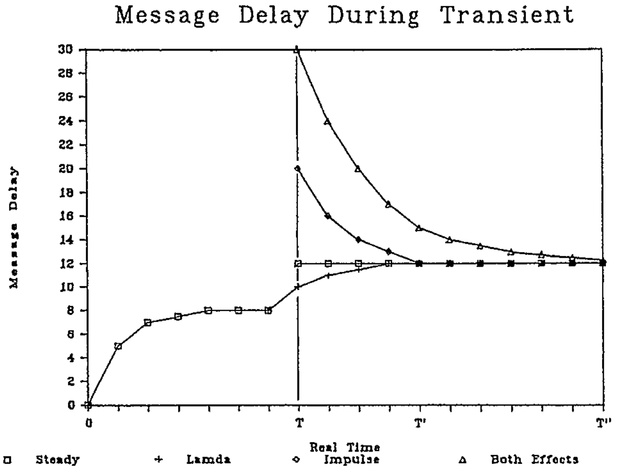

Figure 2: Message Delay During Transient

The research hypothesis tested in this paper is given as:

> There is no variabilty in average message delay, variance of message delay, or the number of lost messages explained by the type of transient, the transient messages' precedence or the interactions of these factors for the aggregate message load or any of the precedence level message loads.

This hypothesis is tested using data obtained from a simulation model where the engineering features of the network are fixed and the factors of interest are controlled by the experimenter. The data is analyzed using ANOVA techniques to determine the significance of the transient phenomenon and the factor effects.

2. RESEARCH METHOD

The Simulation Language for Alternative Modeling, SLAM (Pritsker, 1984), is used to construct the communications network depicted in Figure 3. The SLAM simulation, operating on a CDC Cyber 845 and then revalidated on a VAX 11/780, was validated by testing various network configurations with known steady state results. Also the SLAM model was used to test analytic solutions available for selected networks. The simulation model is deemed valid and generates results supported by analytic solution.

The network in Figure 3 represents a five-node, mesh-plus network with fixed link capacities, processor speeds, error rates, ACK/NAK protocol, finite node queue length and first-in-first-out queue discipline. In addition, the use of message precedence common in military networks is included to determine the effects of transients on messages categorized by precedence level. Four precedence levels were used with an ambient message load profile of 10% highest, 10% next highest, 30% medium, and 50% low precedence. Although other network topologies could be used, the mesh-plus is considered to be consistent with actual network implementations and complex enough to allow extension of the results to more elaborate networks. Five nodes are used to give reasonable complexity and realism.

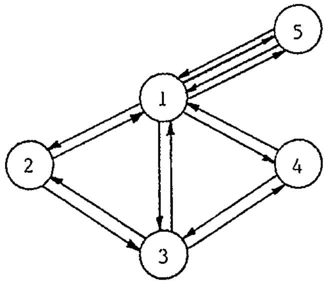

Figure 3: Network Topology

2.1. The Simulation Model

The simulation was written using a top-down modular approach, then validated by comparison with published results of analytic models and a detailed walk-through of the SLAM source code. The network simulates message switching in a store-and-forward method for relaying messages. Messages are completely received at one node before they begin transmission to the next node. Outgoing transmissions are not preempted.

The network analyzed in this paper uses link capacities of 2400 bps. Message lengths are exponentially distributed around a mean length of 6400 bits, approximately equivalent to one-half of a common video screen display in ASCII code.

The SLAM language is used because its process oriented statements are ideally suited for modeling a computer-communications network. Maximum use is made of the global variable feature to achieve a generalized simulation. The SLAM program code is highly visible and easily modified for maximum flexibility.

The modular structure of the program, similiar to that of Chlamtac and Franta (1982), is shown in Figure 4. Dividing the network into functional modules facilitates writing the simulation in a logical and controlled manner and assisted the validation process since each module was checked for correctness individually. The modular simulation allows changing one portion of the model without affecting others, thus permitting flexible experimentation and use of the simulation. The SLAM simulation modules segments described in the following paragraphs indicate how the simulation operates.

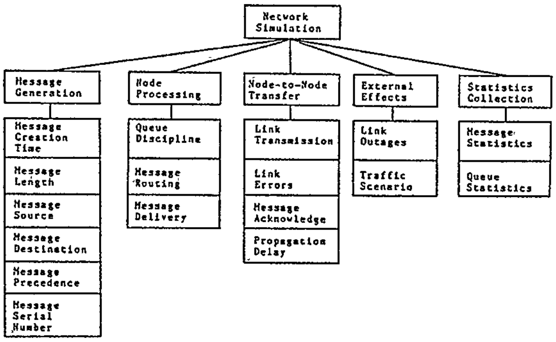

Figure 4: Modular Simulation Structure

Message Generation Module. This module randomly generates messages, fills in the message attributes and places the messages in an outgoing queue. All messages contain the attributes listed in Table 1. Message interarrival times have an exponential probability distribution. In this 5-node network there are 20 message generation module segments like that shown in Figure 5.

Table 1: Messsage Attribute Definitions

Attribute	Definition	Units
1	Creation Time	seconds
2	Destination Node	integer
3	Type	integer
4	Precedence	integer
5	Origin Node	integer
6	Serial Number	integer
7	Length	bits

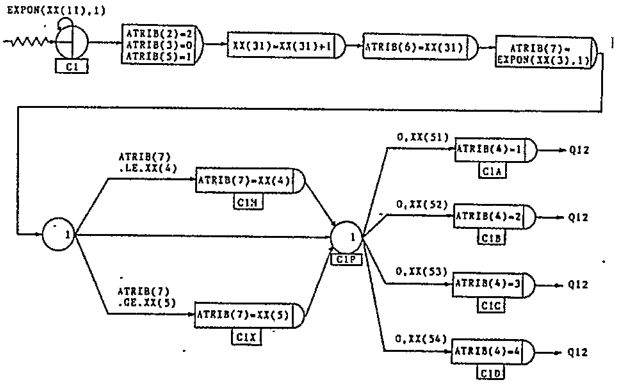

Figure 5: Message Generation Module

Node Processing Module. At each node in the network a switching computer performs node processing tasks. The time required to accomplish these tasks is simulated as an overhead processing time. Queue discipline is highest-precedence-first based on a four-level message precedence scheme. All messages are sorted by destination. Messages that have arrived at their destination are sent to the statistics collection module. Messages to be relayed are placed in the appropriate outgoing queue according to a fixed routing table that depends on network topology. Routing is accomplished using node labels. Figure 6 shows one node processing module segment.

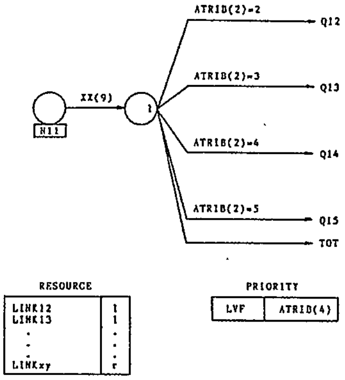

Figure 6: Node Processing Module

The communication links are simulated using SLAM resource statements, and an initialization module is used to alter the resources so that they correspond to the network connectivity.

Node-to-Node Transfer Module. Node-to-node transfer activities include simulating link transmission time, link errors and a message acknowledgement protocol. Link transmission time is calculated for each message using the link capacity and message length.

Link errors are simulated using probability branching. An error rate can be specified for each link. Messages awaiting transmission are held in a finite queue, and messages arriving to a full queue are balked to the statistics collection module. Figure 7 represents one link in the network. Messages are retransmitted until they are received without error. Error-free messages release the link resource for the next message to begin transmission.

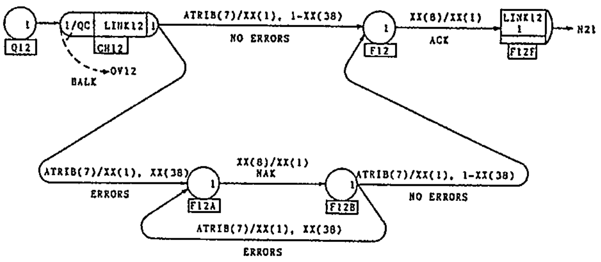

Figure 7: Node-to-Node Transfer Module

Statistics Collection Module. The SLAM summary report provides all the statistical output information for the simulation. SLAM collect nodes are used to determine statistics on message delay and queue overflows. Message delay is reported both by precedence level and as an overall statistic. The average length, maximum length and the average waiting time for each queue are computed. The number of queue overflows and the average time between queue overflows are reported for each queue. A utilization factor for each link is reported. The statistics collection module is shown in Figure 8.

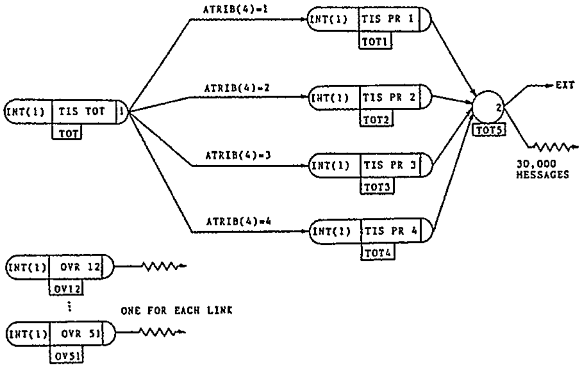

Figure 8: Statistics Collection Module

2.2. The Experiment Design

The statistical model is a designed experiment consisting of a vector of performance measures as dependent variables and a set of treatment factors set at discrete levels. The first factor is the type of transient; either none, an impulse of one minute's traffic, a 100% increase in exponentially distributed traffic for one minute (step) or a half-and-half combination of impulse and step. The second factor is the precedence profile of the transient traffic described as either the same as the ambient load, all high precedence transient, or equally likely precedence levels. The resulting design is a 4x3 factorial where the model is given as:

$$[Y]_{ijk} = u + T_i + P_j + TP_{ij} + e_{ijk} \quad (1)$$

where:

$[Y]$	is the performance vector: Average Delay, Variance of Delay, and Number of lost messages.
u	is the mean performance measure.
P_j	is the effect due to transient message precedence profile; 3 levels: Ambient, High, Equal.
T_i	is the type of transient; 4 levels: None, Impulse, Step, Both.
TP_{ij}	is the type transient x transient profile interaction.
e	is the random error component.
k	is the number of experiment replications per cell.

All effects are analyzed at the 95% significance level. Range tests are used to group the performance measures into statistically equivalent sets across the experimental factors.

Replications of the simulation are run to reduce the error component and to allow estimation of all interaction effects. Antithetic variables were used to accomplish a degree of variance reduction. The simulation model generates the observations for each cell in the experiment design. Since the simuluation runs are independent of one another except for the random number seeds, the experiment is a completely randomized design with no blocks.

3. EXPERIMENTAL RESULTS

Four replications of a 4x3 factorial experiment were run. Each of the experimental units represented a combination of a type transient and a transient message precedence profile. The ambient message load was set at 120 messages per minute and allowed to continue for the entire simulation run. At 100 seconds into the simulation a transient was initiated and statistics gathered for an additional 300 seconds. Pilot studies established that 100 seconds allowed reasonable steady state behavior and that 300 seconds allowed even the worst case transient to subside. Regardless of the type transient or its precedence profile, only 120 messages are generated in the transient. Such a design insures that the within and across an experimental unit tests are compatible.

Figure 9 depicts a sample simulation run where the aggregate message delay is plotted with respect to the receipt time of a message. The presence of the transient is clear. Each simulation run recorded the aggregate message delay as well as the delay for each precedence level. The variance of the message delay was also recorded as well as the total number of lost messages due to finite queues. Table 2 presents the mean performance measures for the 48 runs.

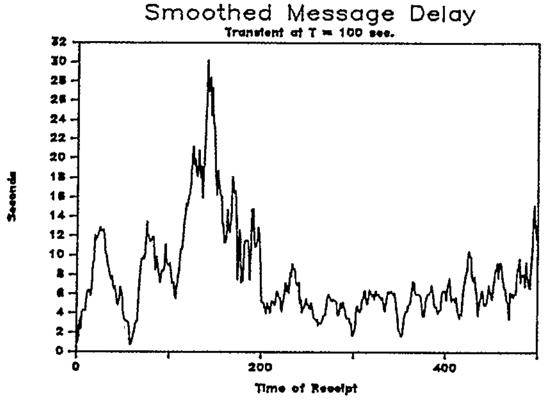

Figure 9: Smoothed Message Delay

A linear model was estimated for each performance measure using the experiment design in Equation 1. Table 3 depicts the results of the ANOVA and the corresponding F statistic for each factor for each of the linear models.

It is noteworthy that the assumption of constant variance of the error term when using the variance of message delay as the dependent variable is questionable. Nevertheless, the effects appear consistent with the analysis as a whole and considerable insight is gained. Since the purpose of this paper is expository in nature, specific contrasts were not formulated a priori.

The combination of information in tables 2 and 3 is best understood when presented as an effects chart. An effects chart for the average message delay (D), average delay for high precedence messages (D1), and lost messages is included in Figures 10, 11, and 12, showing the effects for the types of transient and precedence profiles. The statistical findings presented in Table 3 are evident.

Duncan range tests on all main effects were also computed, and the findings drawn from these tests are summarized in the conclusions. Since range tests in general do not control the experimentwise error rate, we make generalized conclusions with discretion. However, armed with these data, a number of significant and useful conclusions can be made.

Table 2: Average Performance Measures

Experimental Factors (Transient - Profile)	D	D1	D2	D3	D4	V	V1	V2	V3	V4	Lost
None - Ambient	6.17	4.8	5.4	5.7	6.9	42	21	27	28	57	0
None - Equal	6.17	4.8	5.4	5.7	6.9	42	21	27	28	57	0
None - High	6.17	4.8	5.4	5.7	6.9	42	21	27	28	57	0
Impulse - Ambient	9.41	4.9	6.1	7.2	12.4	127	14	28	49	198	25
Impulse - Equal	9.19	7.2	8.4	8.2	10.8	107	29	66	73	153	25
Impulse - High	9.35	12.5	6.1	7.2	9.6	105	112	39	75	121	25
Step - Ambient	9.51	5.6	5.9	7.1	12.6	116	23	26	39	181	3
Step - Equal	9.60	5.7	6.6	7.9	12.6	128	27	37	67	207	8
Step - High	9.57	8.0	7.0	8.3	11.8	119	45	55	72	198	7
Both - Ambient	9.97	5.4	5.6	7.4	13.4	139	20	23	46	220	15
Both - Equal	9.95	6.3	8.4	8.9	12.1	143	35	72	94	217	17
Both - High	9.60	11.6	6.3	7.4	10.5	113	87	49	64	163	18

D = average message delay in seconds, D1-D2 is for precedence levels 1-4.
V = variance of message delay, V1-V4 is for precedence levels 1-4.
Lost = the total number of lost messages.

Table 3: F Statistics for each Factor Effect.

Dependent Variable	Factor			Model	
	T	P	TxP	F	R-Square
D	45.3*	0.1	0.1	12.42*	0.79
D1	31.7*	82.4*	14.7*	31.63*	0.91
D2	7.5*	12.1*	3.7*	6.24*	0.66
D3	18.0*	4.1#	1.4	6.42*	0.66
D4	34.4*	4.8#	1.0	10.81*	0.77
V	29.6*	0.9	0.6	8.55*	0.72
V1	24.6*	99.4*	22.6*	37.09*	0.92
V2	2.9#	7.6	2.3	3.43*	0.51
V3	9.5*	6.3*	1.3	4.44*	0.58
V4	24.1*	1.8	1.1	7.45*	0.69
Lost	70.9*	0.8	0.4	19.67*	0.86

* indicates significance at alpha = 1%
\# indicates significance at alpha = 5%

D = Average message delay, D1-D4 are delays for precedence levels 1-4.
V = Variance of message delay, V1-V2 are variances for precedence levels 1-4.
Lost = total number of messages lost.

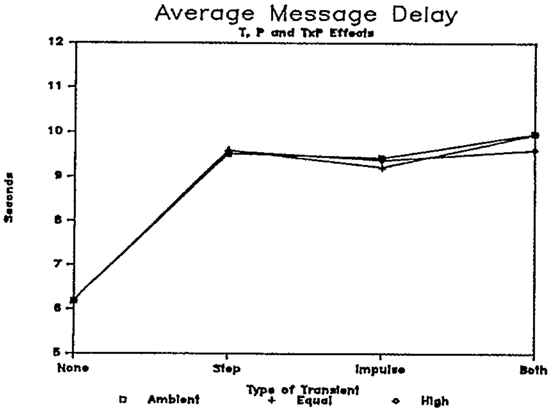

Figure 10: Average Message Delay

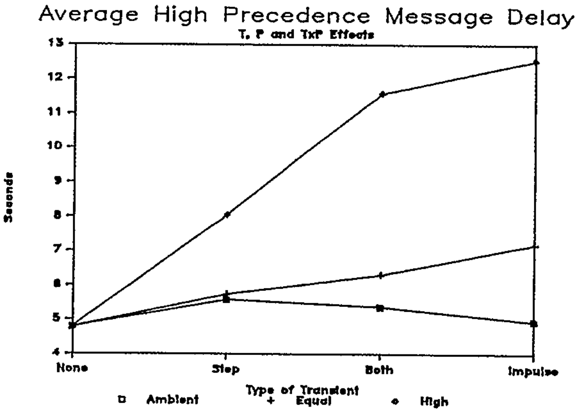

Figure 11: Average High Precedence Message Delay

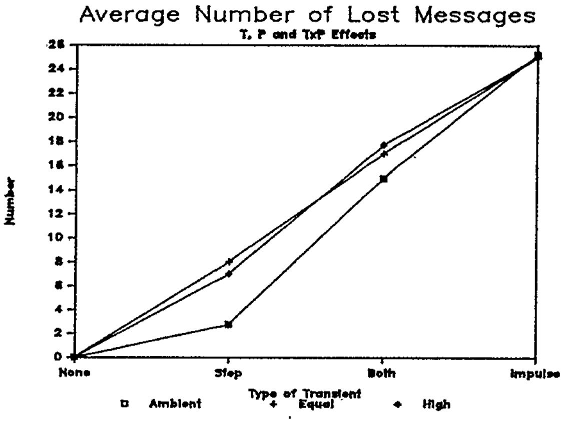

Figure 12: Average Number of Lost Messages

4. CONCLUSIONS

Several significant conclusions are evident. First, the hypothesis that transient type and transient precedence profile do not affect the performance measures is soundly rejected. The type of transient and the transient's precedence profile do explain the transient performance of the network as evidenced by the overall models' F statistics in Table 3. Every model estimated was significant at the 1% level. Also note the relatively high explanatory power of each linear model.

Second, the significance of each factor and their interaction is not consistent across performance measures or between precedence levels. This interesting finding indicates that judgments based on aggregate network observation can be severely in error. Furthermore, an observer would conclude that average network performance does not depend on the precedence profile of the transient! Clearly, the analysis of transient performance must include an itemized study of each category of traffic.

It is noteworthy that the only significant interaction of transient type and precedence profile exists when looking at the average delay of the two highest precedence levels (D1 and D2) or the variance of the highest precedence traffic (V1). Apparently, the interaction effects those messages most who ordinarily receive priority in transmission. The remaining categories of traffic are simply delayed longer than their customary delay but not because of the presence of differences in transient precedence. Rather, the simple fact that a transient occurs delays the lower precedence messages. This observation is consistent with the intended purpose of the precedence scheme.

Duncan range tests indicate that no difference in average network performance can be specified across the types of transient given that a transient has occurred. Apparently a transient creates its effect regardless of the nature of the transient. This finding seems difficult to support conceptually. However, we feel the finding is correct and explained by the finite queue capacity of the network buffers. The effect of an impulse transient is somewhat diminished by the fact that a number of the messages will simply be lost and not allowed to further worsen the averages.

Range tests do indicate differences in the number of lost messages attributable to the type of transient. The existence of an impulse resulted in five times as many lost messages than the sudden increase in exponentially distributed arrivals and 50% more lost messages than the co-existence of an impulse and reduced arrival type transients. This finding is a function of the finite queue capacity and indicates the short-sightedness of relying on one performance measure such as message delay.

In summary, the effects of a transient are significant and severe. For our ambient conditions and degree of transient, an average additional delay of 50% was observed. The relation between type of transient and transient precedence is complex, nonlinear and varies across categories of message traffic. Major errors in judgment result if aggregate network performance is the analysis criteria. Therefore, we reject the stated research hypothesis and conclude that considerable insight can be gained on the transient performance of a network.

5. SUGGESTIONS FOR RESEARCH

Several extensions of this work are possible. First, the degree of the transient was fixed. What is the change in transient effect when the intensity of the transient is varied? Second, the ambient traffic load was fixed during the course of the simulation. It is known that steady-state performance of a network is a nonlinear function of the message load. Do transient effects also vary when the ambient message load is changed? Finally, what are the effects of the transient on different network topologies? The mesh-plus network used here provides insight generally applicable to many real-world applications. However, the transient performance of a network may be related to its topology and communications protocols. We intend to investigate these subtleties in follow-on research.

REFERENCES

Chlamtac, I. and Franta, W. R. (1982). A generalized simulator for computer networks. Simulation, October 1982, 123-132.

Kleinrock, L. (1976). Queueing Systems Volume II: Computer Applications. John Wiley & Sons, New York.

Kuo, F. F. (1981). Protocols & Techniques for Data Communication. Prentice-Hall, New Jersey, 186-189.

Pritsker, A. A. B. (1984). Introduction to Simulation and SLAM II, Second Edition. Halstead Press, New York.

Tanenbaum, A. S. (1981). Computer Networks. Prentice-Hall, New Jersey, 67-70.

AUTHORS' BIOGRAPHIES

ALBERT B. GARCIA is an Assistant Professor of Electrical and Computer Engineering at the Air Force Institute of Technology. He received his B.S.E.E. and M.S.E.E degrees from West Virginia University in 1968 and 1970, his M.B.A. from Fairleigh Dickinson University in 1980, and his Ph.D. in engineering from the University of Dayton in 1985. His teaching and research interests are computer networks, computer architecture and simulation.

Major Albert B. Garcia
AFIT/ENG
Air Force Institute of Technology
Wright-Patterson AFB, OH 45433
(513) 255-3576

WADE H. SHAW is an Assistant Professor of Electrical and Computer Engineering at the Air Force Institute of Technology. He received his Ph.D. in Engineering Management from Clemson University in 1985. He is an active member of the Institute of Industrial Engineers, the Society for Computer Simulation, the American Institute for Decision Sciences, the Operations Research Society of America and The Institute of Management Science. His teaching, research and consulting interests include simulation, statistical modeling and decision support systems.

Captain Wade H. Shaw
AFIT/ENG
Air Force Institute of Technology
Wright-Patterson AFB, OH 45433
(513) 255-3576