A System Response Time Model for Local Area Networks

Dr. Norman F. Schneidewind

Code AS/Ss
Naval Postgraduate School
Monterey, CA 93943

ABSTRACT

There are several issues that confront LAN management with respect to allocating servers and files in a LAN. These are: How many servers should be used for a given number of user computers? Should files be replicated on the servers to provide better performance or should the available servers store different files in order to maximize the number and variety of programs on the LAN? What should be the acceptable access times for users, singly and simultaneously, to access an application program, with a given number of servers and user computers? These issues are analyzed as a problem in optimization to determine the optimal server/user computer ratios and degree of file replication for given access times. Analytical and empirical results are reported.

INTRODUCTION

The problem posed by this research is the following: how many servers to use on a LAN for given number of users and application requirements. In answering this question, we use the following criteria:

o We want sufficient speed to satisfy users but not so much that resources are wasted.

o We can achieve multiple servers by duplicating application program files.

o We don't want too many servers for given files because there are alternate uses of servers:

- Other files on same LAN
- Other LANs

Our criterion for deciding when to stop adding servers is the following: *stop adding servers when the decrease in response time becomes insignificant.*

EXPERIMENTAL SYSTEM

The IBM PC Network which was used for our experiment is shown in Figure 1 and consists of the following major elements:

o 25 user computers
o 3 servers
o Broadband CSMA\CD, 2 mb/s

Environment

The experimental system operates in a university laboratory environment where maximum system stress occurs when 6-15 students in a lab class access a program simultaneously. In this environment we define the response time performance measurement as follows:

Elapsed time between simultaneous requests for a program and its appearance on **each** user's screen. An important subset of this measurement is the **longest** time required for the program to appear on a user screen. Response time for this network is defined in Figure 2.

EXPERIMENT

The experiment was conducted as follows:

o Make simultaneous requests for Wordperfect 5.0.
(This program was chosen because it is typical of the programs used in the environment described above and it is representative of the size of program that users access from the servers).

o Record response times.

- First and last
- All, in some cases

o Record disk activity times: server and user.

- Vary number of simultaneous requests and servers.

- Allocate requests uniformly among the servers.
(This is the way the production network is loaded).

- Construct analytic model and compare with measurements.

MODEL

This complex analytic model was constructed in six parts, as defined by the six components of response time described below and illustrated in Figure 2.

Variables and Parameters

a : Normalized signal propagation delay on bus (dimensionless).

A : Maximum probability that a computer acquires bus and transmits.

B : Bus transmission rate (bits per second): 2 mb/s.

D : Bus cable length (meters): 33.5 m (longest cable).

N_o : Mean number of program file packets (output packets) per user request that compete for use of bus.

N_p : Number of packets required to transfer program file from server to user computer: 215.

N_u : Number of simultaneous user requests that compete for use of bus (total).

N_{us}: Number of user requests per server.

P_i : User request (input packet) size (bytes): 384 bytes.

P_o : Program file (output packet) size (bytes): 1600 bytes.

S : Normalized maximum mean throughput on bus (dimensionless).

T1 : Mean delay on bus between initiation of user request (input packet) and arrival of request at server input buffer (seconds).

T2 : Mean wait time by user request in server input buffer for hard disk access (seconds).

T3 : Mean server hard disc access time (seconds).

T4 : Mean wait time by program file packets in server output buffer for hard disk access (seconds).

T5 : Mean delay on bus between transmission of program file packets from server output buffer and their arrival in user computer RAM (seconds).

T6 : Mean user floppy disk access time (seconds).

T_r : Mean user request response time (seconds).

V : Bus signal propagation velocity: .77C, where C = $3x10^8$ m/s.

1. The elements of Ethernet inbound bus delay estimate, T1, for a user request are as follows:

o Transmission interval[1]

o Contention interval[1]

o Normalized propagation delay (a)[1]

o Maximum probability computer acquires bus and transmits (A)[1,3].

o Normalized maximum mean throughput (S)[1,2,3,4]

o Maximum mean throughput[5,6]

o Normalized maximum mean delay[2]

o Maximum mean delay[7,8,9,10]

Using the above elements we have for S:

$$S = \frac{1}{1 + 2a\ (A^{-1} - 1)} \quad (1)$$

where $a = (BD)/(VP_1)$ (2)

and $A = (1 - N_u^{-1})^{(Nu - 1)}$. (3)

From S, we estimate T1 from delay vs. throughput curves. Also we can obtain the zero collision delay or packet transmission time from P_1/B.

2. The elements of server hard disk wait time estimate, T2, to access program file are as follows:

o Finite source (number of simultaneous user requests per server N_{us} for service) queuing model.

o Delay due to T1 is negligible so we can assume all requests arrive at hard disk input buffer "at same time".

o FIFO service discipline.

o Wait for copy of program file to be read from the disk.

o Estimated from mean number of user requests **per server** (N_{us}) (and hard disk access time T3) that will be serviced prior to this request.

Using the above elements we estimate the mean number of user requests that will be serviced before this request can be serviced, and multiply it by T3:

$$T2 = (N_{us} - 1)/2\ (T3) \quad (4)$$

For example if there are three requests made to a server, a given request will wait for 0, 1, or 2 service times T3, or a mean of 1.

3. The elements of server hard disk service time estimate, T3, to access and read program file are as follows:

o Distribution approximately constant because same area on disk always accessed for given program.

o IBM AT seek, track-track, rotational delay, read transfer times[11].

o Read time dominates because of large program file that is transferred (338 KB).

o Number and size of disk buffers (15 @ 512 bytes).

o Estimated from: seek + rotation + cylinder-cylinder (track-track) + read time. This time is constant because, for a given program (e.g., WordPerfect), the same part of the disk is accessed for each user request. The disk acts as a single server for its requests because a given user's requests are directed toward a specified server (for load balancing purposes). However multiple servers may be used in the network.

Using the above factors we find T3 = 8.4 seconds.

4. The elements of Ethernet outbound bus wait estimate, T4, for program file are as follows:

o Wait for program file bus transmissions that are ahead of the given program file.

o Number and size of disk buffers (number and size of packets that can be transferred is limited by disk buffer size).

o Estimated from mean number of program file packets (output packet) (and bus delay time T5) that will be transmitted prior to these packets being transmitted.

Using the above elements we estimate the mean number of program file packets (output packets) that will have to be serviced before these packets can be transmitted, and multiply it by T5:

$$T4 = (N_{us} - 1)/2 \; (T5) \qquad (5)$$

5. The elements of Ethernet outbound bus delay estimate, T5, are the same as for T1, above, except the number and size of packets are large because a copy of the large program file is required by each user and the transfer is limited by disk buffer size[12]. Thus, the queuing model is characterized by a large finite source of "customers" and packets do not arrive at the user computer RAM at the same time.

We estimate the mean number of program file packets (output packets) per user request that compete for the use of the bus (the given packet from this server plus the mean number of packets from other servers):

$$N_o = 1 + ((N_u - 1)/2 \; (N_p)) \qquad (6)$$

We compute $a = (BD)/(VP_o)$ (7)

and $A = (1 - N_o^{-1})^{(No - 1)}$. (8)

This allows us to use (1) to obtain S, as in 1. above, and to estimate T5, as described in 1. Also we can obtain the zero collision delay or packet transmission time from P_o/B. We must multiply individual packet times by N_p to obtain program file transmission time.

6. The element of interest for the user diskette (floppy) service time[13], T6, is the following:

o Estimated from: start + head settling + seek + rotation + read/write time multiplied by number of required accesses. This time is a constant because WordPerfect always writes the same set up file on the user's diskette.

Using the above factors we find T6 = 19.2 seconds.

7. Now we are able to estimate the mean user request response time:

$$T_r = T1+T2+T3+T4+T5+T6 \qquad (9)$$

RESULTS

Data were collected for various combinations of experiments: the number of user computers that simultaneously accessed the network was varied from one to ten; the number of servers was varied from one to three. Response time was measured for each user computer allowing us to calculate the minimum, mean and maximum response time for each experiment. Both server hard disk and user computer floppy disk activities were recorded so that these could be compared with times computed from the analytic model.

The six components of mean and maximum response time were estimated for each experiment, for comparison with the measured results, as pictured in Figure 2 and in accordance with the elements of each component, as identified in the MODEL section. In addition, estimates were made for using four servers.

The first analysis involved comparing estimated mean response time with measured minimum and maximum response times to see whether the former was bounded by the latter. The results are shown in Figure 3 for three servers, with a maximum of six users. Next, measured differences in mean response time were obtained as the number of servers was increased from one to two and from two to three. Differences in response time were **estimated** for adding a fourth server. A similar analysis was performed to obtain measured and estimated differences in maximum response time. Maximum response time was estimated by using maximum rather than mean number of requests and packets in (4), (5), and (6). The purpose of the analyses was to identify the point where adding another server would not result in a significant reduction in response time.

CONCLUSIONS

Based on the data presented in the RESULTS section, we conclude the following:

o A fourth server should not be added because the reduction in response time would be insignificant.

o It pays to replicate files on servers because there is a significant reduction in response time in going from one server to two.

o Response time would be unacceptable without multiple servers.

o Disk time (server hard disk and user floppy) dominates the total response time (**as much as 90% of response time!**). It seems odd that so much attention is given in the network literature to models that analyze just **one** aspect of the **total** network performance -- Ethernet delay and throughput when, in fact, server hard disk performance is the key to response time. This is not to suggest that there are no network configurations and operational profiles where Ethernet performance would dominate. Rather, it is our conviction that the performance of the entire network must be analyzed.

o Greater performance for a given number of servers would be achieved by increasing the number of server disk buffers.

o When modeling such a network, it is very important to understand what is going on in the hardware and software! (e.g., long floppy disk access times).

o The analytic model can be used to estimate response time for larger numbers of users and servers than those shown in the performance curves. The latter were used for convenience of data collection.

o Because of the complexity involved in analyzing the performance of an entire network, a simulation model would be useful for obtaining greater estimating accuracy.

ACKNOWLEDGEMENT

Acknowledgement is due Mr. Leon Sahlman of the Naval Postgraduate School and graduate student John Riley III for assisting in the data collection.

REFERENCES

1. William Stallings, Local Networks, Macmillan, 1984.

2. Joseph L. Hammond and Peter J.P. O'Reilly, Local Computer Networks, Addison-Wesley, 1986.

3. Robert M. Metcalfe and David R. Boggs, "Ethernet: Distributed Packet Switching for Local Computer Networks", CACM, Vol. 19, No. 7, July 1976, pp. 395-404.

4. John R. Shoch and Jon A. Hupp, "Measured Performance of an Ethernet Local Network", CACM, Vol. 23, No. 12, December 1980, pp. 711-721.

5. Bart W. Stuck, "Calculating the Maximum Mean Data Rate in Local Area Networks", IEEE Computer, Vol. 16, No. 5, May 1983, pp. 72-76.

6. B.W. Stuck and E. Arthurs, A Computer & Communications Network Performance Analysis Primer, Prentice-Hall, 1985.

7. Fouad A. Tobagi and V. Bruce Hunt, "Performance Analysis of Carrier Sense Multiple Access with Collision Detection", Advances in Local Area Networks, Karl Kummerle, et al., eds., IEEE PRESS, 1987, pp. 318-339.

8. Timothy A. Gonsalves, "Measured Performance in the Ethernet", Ibid, pp. 383-410.

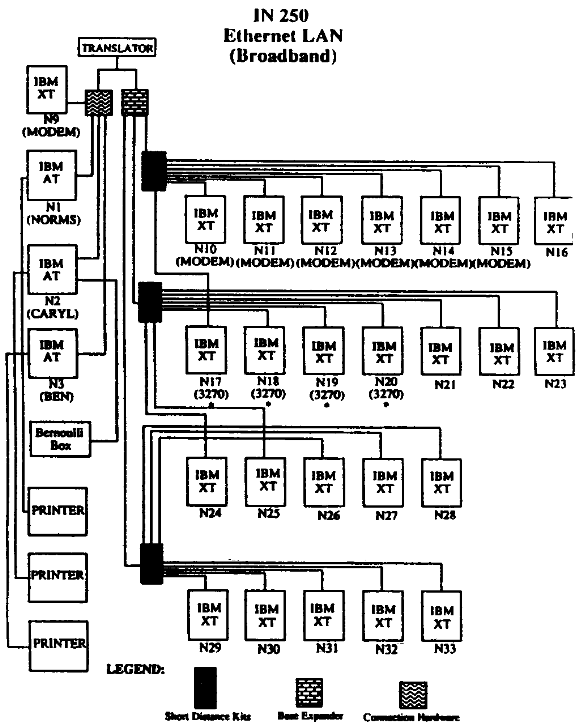
9. Werner Bux, "Local Area Subnets: A Performance Comparison", Ibid, pp. 363-382.

10. W. Bux, "Performance Issues in Local Area Networks", IBM Systems Journal, Vol. 23, No. 4, 1984, pp. 351-374.

11. Jon Shiell and John Markoff, "Fixed Disks and the PC AT", BYTE, Fall 1985, pp. 167-174.

12. IBM, Local Area Network Technical Reference, SC30-3383-2, November 1988.

13. IBM, Technical Reference, Options and Adapters, Vol. 1, 1984.



* The 3270 Emulation connects directly to the mainframe.

Figure 1

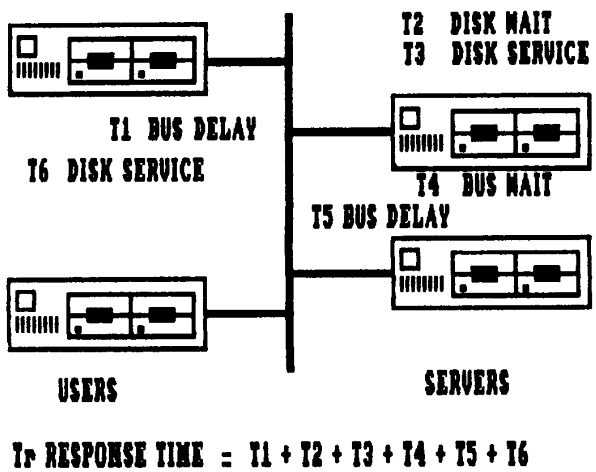


Figure 2

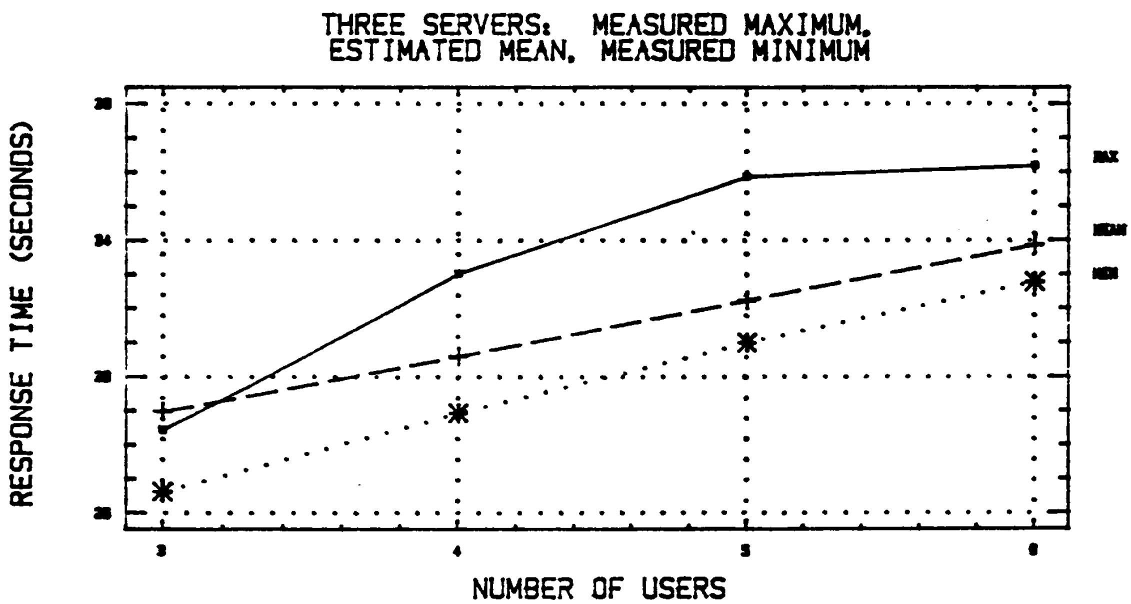


Figure 3