1995 NASA/ASEE SUMMER FACULTY FELLOWSHIP PROGRAM
JOHN F. KENNEDY SPACE CENTER
UNIVERSITY OF CENTRAL FLORIDA

55-62
2745
p. 28

CRITERIA DEVELOPMENT FOR UPGRADING COMPUTER NETWORKS

Dr. Kemal Efe
Associate Professor
Center for Advanced Computer Studies
University of Southwestern Louisiana
Lafayette, Louisiana

KSC Colleague - Ray Pecaut
Communications/Networks

Contract Number NASA-NGT-60002
Supplement 19

July 28, 1995

ACKNOWLEDGMENTS

This research would not have been possible without the support of the NASA/ASEE Summer Faculty Program. **Ray Hossler** and **Kari Stiles** have made this summer program possible as well as enjoyable. **Greg Buckhingham** and other NASA personnel have made us feel at home and this tremendously helped the congenial atmosphere that was created here.

Throughout the research efforts, I benefited from intense discussions and suggestions made by **Ray Pecaut.** I am deeply grateful for his support and encouragement.

I am also grateful to **Cindy Johnson** for sharing with me significant technical information, which was no doubt the result of many year's work, for the benefit of this research.

Furthermore, I wish to extend my appreciation to **Perry Rogers** for his support and encouragement.

Finally, I am grateful to my wife and children for understanding when I had to be absent, over and above the regular work hours, for the benefit of this research.

ABSTRACT

Being an infrastructure system, the computer network has a fundamental role in the day to day activities of personnel working at KSC. It is easily appreciated that the lack of "satisfactory" network performance can have a high "cost" for KSC. Yet, this seemingly obvious concept is quite difficult to demonstrate. At what point do we say that performance is below the lowest tolerable level? How do we know when the "cost" of using the system at the current level of degraded performance exceeds the cost of upgrading it?

In this research, we consider the cost and performance factors that may have an effect in decision making in regards to upgrading computer networks. Cost factors are detailed in terms of "direct costs" and "subjective costs". Performance factors are examined in terms of "required performance" and "offered performance." Required performance is further examined by presenting a methodology for trend analysis based on applying interpolation methods to observed traffic levels. Offered performance levels are analyzed by deriving simple equations to evaluate network performance. The results are evaluated in the light of recommended upgrade policies currently in use for telephone exchange systems, similarities and differences between the two types of services are discussed.

TABLE OF CONTENTS

I. INTRODUCTION

In almost every walk of life, we are regulated by standards and limits imposed or recommended by some organization. A truck comes with a recommended load capacity. A boat or an elevator comes with its recommended number of people to be carried. Air conditioners have recommended volumes to be cooled depending on their capacity. FDA recommendations exist about how much fat, sugar, fiber, etc. one should consume per day. We can go on giving examples of such recommended limits practically forever. An area where no organization seems to have bothered with is computer networks. There are no recommendations or guidelines that advise on how much traffic to be carried on computer networks. Computer networks are now entering every organization, some serving critical functions such as hospitals and chemical plants, where network congestion may have catastrophic results. In other organizations, network congestion may not have immediate consequences, but may never the less have long term impact in operational costs or employee productivity.

The purpose of this research is to develop simple criteria that may be used in decision making in regards to upgrading computer networks. Most computer networks in use today have been installed more than 10 years ego. Due to the recent proliferation in new networking products, network managers throughout the world are faced with the difficult question of whether to continue using their old network that served them so well in the past, or whether to replace it for a new product in the market.

After much literature research and intense discussions with colleagues here at KSC, the answer to this enormous question turns out to be extremely easy and hard simultaneously, at least in the context of the KSC network. What makes it easy is a fact, that became obvious to us in a short time after starting this research, that future applications requiring network services will saturate the existing system much beyond its capacity. The hard part was due to the fact that unless the network is saturated beyond its capacity, any level of utilization below may be interpreted as being acceptable or not based purely on interpretation. There is no single indicator that dominates all other factors to favor replacing

the network for a new one. Moreover, simple management level decisions may be made to postpone the point of catastrophy, for a cost in employee productivity or the range of network services that may be supported. Ultimately, upper management must decide when it is time to upgrade the network.

II. COST/PERFORMANCE FACTORS

When making such decisions, upper management will look for a good trade-off point for cost and performance. Clearly, it is necessary to clarify the two key words: "cost" and "performance." In doing so, we take a close look at cost factors and performance factors in detail.

II.1 COST ELEMENTS:

We can divide the involved cost factors into three categories as follows:

Direct Costs: These are the direct costs with two components: (a) direct costs for maintenance and management, and (b) direct cost of personnel time (for those personnel using the network) for time lost due to slow response time.

Older technology usually has higher maintenance cost. Contrasting to the state-of-the-art technology, such costs can be predicted with reasonable accuracy. Direct cost due to slow response time can also be computed reasonably accurately since every second of employee time has a cash value based on the salary.

Subjective Costs: These are the cost factors for lost opportunity. If the existing system does not offer the necessary capacity required for state-of-the-art facilities, the inability to use such facilities may have high cost for an organization. For example, video conferencing is one area gaining significant market in recent years. Other examples include electronic marketing, or access to digital libraries, etc.

This is a relatively difficult cost category to quantify. For some organizations, the need for state-of-the-art facilities may be as real as the need for a telephone on the executive director's desk. For other companies, these may be nothing more than fancy toys. Nevertheless, there is already a sizable

customer base for these facilities. Several companies offer free access to their data bases (analogous to the 800 numbers for telephone access) so that customers can easily obtain product information, place orders, etc. In the near future, a substantial increase in the number of users is expected for remote access to such on-line data which may have been stored in graphics, video, audio, or text format. AT&T and MCI are now offering up to 1.5 Mb/s bandwidth (per connection) for 800 calls for institutions that want to offer computer access to their on-line data. Various uses of such data bases in education, commerce, etc. are anticipated.

For the purposes of this research, It is assumed that the ability of KSC personnel to access such facilities offered by others has a "high" value (even though it is difficult to quantify this value due to its subjective nature).

Cost Efficiency: The amount of bandwidth purchased per dollar spent is a measure of cost efficiency. Also, the concept of cost efficiency can be applied to maintenance, i.e., the amount of bandwidth maintained per maintenance dollar spent. As defined, the maintenance cost efficiency can be easily calculated for a network by using the maintenance records.

II.2 PERFORMANCE CONSIDERATIONS:

Performance considerations are addressed under two categories: required performance, and offered performance.

Required Performance: This refers to the level of performance needed in order to satisfy certain quality of service parameters arising from applications that are deemed necessary for an organization. The amount of required bandwidth is among the major parameters that determine the quality of service. Other parameters include the delay, the probability of packet loss, the amount of jitters tolerated, etc. To do a formal study of total network capacity needed, the relationship between these parameters needs to be well understood. For the time being, it should be interesting to consider the bandwidth requirements arising from multimedia applications, for a single user, as it will allow us to do a quick and dirty analysis for the existing FDDI backbone network at KSC. The table below lists the

requirements for video, graphics, audio, and data communications, as these are typical modes of display in multimedia. (Table is compiled from reference [1]).

	Bandwidth	Bursty vs. Continuous	Tolerable Loss	Delay/Jitters Sensitivity
CBR Video	Several* Mb/s	Continuous	10^{-6}	Tight
VBR Video	Several*Mb/s	Bursty	10^{-6}	Tight
Image	Several Mb/s	Bursty	10^{-8}	Medium
Audio	Under 100Kb/s	Continuous	10^{-2}	Tight
Text/Data	Flexible	Bursty	10^{-10}	Loose

* MPEG I specified rate is 1.5 Mbit/s for CBR video, and 2.5 Mbit/s for VBR video.

Table 1: Quality of service parameters for multimedia applications.

The specified bandwidths for CBR and VBR Video are expected to increase to 6 and 19.2 Mbit/s, respectively, in the MPEG II standard for HDTV. If we use the 1.5 Mb/s value, and since the FDDI network operates at 100 Mb/s, we can allow only about 60 users at a time engaging in a video session (e.g., to search through a manufacturer's on-line catalogue). This quick and dirty computation does not consider the other traffic normally on the network. Also it does not take into account other quality of service parameters such as delays in accessing the network.

Offered Performance: Offered performance refers to the actual quality of service offered by the existing system. Network monitoring tools exist that measure various parameters such as the number of packets sent, the length of each packet, the delay taken for a packet to reach its destination, and so on. Presumably, the management would use this information to determine the bottleneck points on the network, and make decisions about upgrading the system if and when necessary. Unfortunately there are no standards organizations that make recommendations for computer networks. However, there are voluntary standards observed by telephone companies for their switching systems. In

telephone industry, grade of service is specified as the percentage of calls lost due to busy switching system (at peak time). It appears reasonable to do a detailed study of the underlying theory behind these quality of service standards used by telephone companies and investigate their suitability for computer networks. We do so in Section 4.

III. TREND ANALYSIS FOR NETWORK UTILIZATION

In marketing literature, one curve often used is the S-curve that describes percentage of customers having purchased a given product. This curve generally looks like that in figure below. When a new product (e.g. automobile, television, computer, etc.) is introduced to the market, initially 0% of the potential customers actually have that type of product. In the first few years after the product hits the market, acceptance is low, and only a small proportion of potential buyers actually acquire that product. Eventually more and more people buy the product and the percentage of people that have bought the item begins to increase sharply. After some time, the increase in the percentage begins to level off, and reaches the 100% level gradually.

This type of curve can also be applied to model the behavior of network users, and ultimately to analyze the level of required network services. Given a network, initially we anticipate a low usage as potential users have not yet acquainted themselves with benefits and conveniences of such a service. After a while more and more people use the network causing a sharp increase in the number of users. Eventually, all people that are likely to use the network become regular users, and the proportion of people using the network levels off gradually, reaching it maximum possible level in the long run.

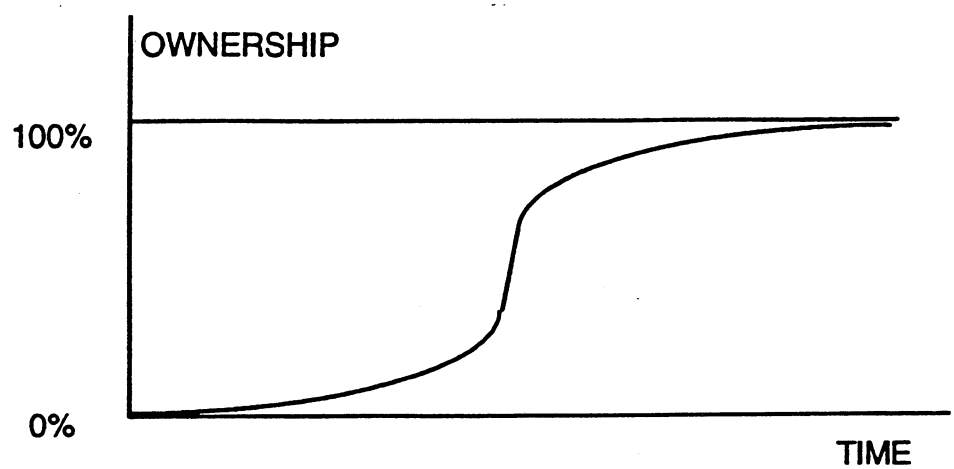

Figure 1: A typical curve representing market growth.

From the network point of view, this pattern of behavior manifests itself in the corresponding increase in network traffic. Therefore, this curve will also describe the pattern of change in network utilization and may be verified by network monitoring tools.

If we assume that this curve represents utilization patterns for computer networks, management questions to be considered are (a) what amount of traffic corresponds to the 100% level on the curve?, and (b) if that level is more than the offered capacity, then how long it will take to reach that level?.

The first question can be answered by considering the number of users, and the expected level of usage per user. Since these are difficult measures to quantify exactly, we may consider best case or worst case scenarios depending on management attitude. Best case could be defined as a conservative estimate on the usage patterns. For example, if we use the MPEG-1 bandwidth requirements for the required bandwidth per user, we can estimate the required bandwidth as 1.5N Mb/s, where N is the expected number of active users. If N=1000, then the estimated traffic would be 1.5 Gb/s. Note that N corresponds to the number of users active at any time. Thus, if for example only 20% of users are active at the same time, then 1000 active users would correspond to a total of 5000 employees. A worst case estimate can be used by considering MPEG-2 bandwidth values which is expected to be close to 20Mb/s, and assuming that a higher proportion of employees will use the network at that level. If 50% of the 5000 employees use the network at any time, requiring 20 Mb/s bandwidth each,

the total bandwidth will be 50 Gb/s. As can be seen a wide range of possibilities exist for the maximum expected traffic load, and depending on whether or not a certain application is made available to the users, different numbers may be used as the estimate for the highest traffic level.

The second question, i.e. that of how long it will take to reach a certain point on the curve, is somewhat more complicated. The reason is, although conceptually simple, this curve may come out as an aggregate function of several components, each of which having a lifetime curve by itself. The corresponding curve might look more like that in the following figure:

Figure 2: More detailed representation of market growth.

In this figure, the thick line represents the aggregate behavior of individual usage patterns. While at any time some of the commodities may have reached their peak usage, there may be other commodities that are in the mid-point of their usage and still others that are just entering the user domain. A detailed analysis and forecast method may therefore benefit from analyzing the usage levels of different items individually. It is very likely, for example, that regular email usage may have exceeded its mid-point while other applications such as WWW may be just beginning to claim its share of bandwidth.

III.1 FORECASTING

The question one normally asks at this point is how do we know where we are with respect to a particular curve. This is a very difficult question to answer as one can only know the past usage levels with absolute certainty. For future usage levels, we can only argue about approximate usage levels under certain assumptions. One assumption we will be making is that rates of increase do not change abruptly. In other words, we will be assuming that the lifetime curve for a commodity is not like those in the figure below.

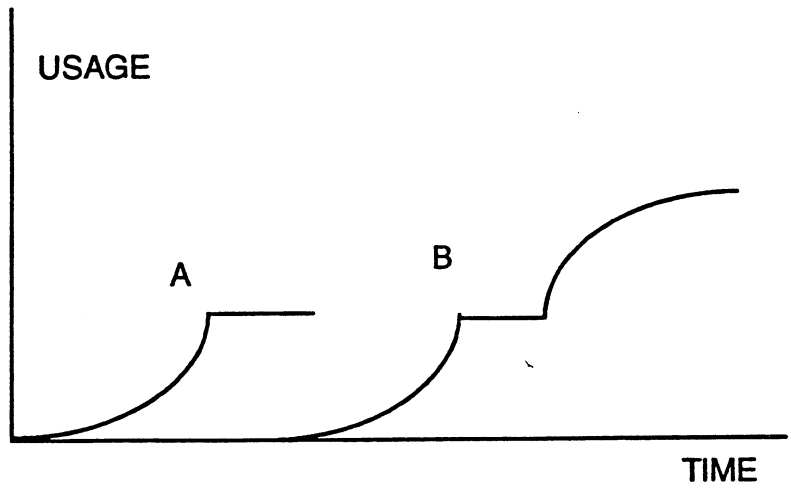

Figure 3: Some anomalous cases in market growth.

In this figure, A represents an abrupt an indefinite blockage to increase of the commodity usage and B represents a temporary blockage. Our smoothness assumption appears to be valid on intuitive grounds, but examples exist to show the contrary. For example, phenomenon analogous to that in figure A has occurred in household TV dish antennas, when cable companies objected to people having their own receiving dish antennas. The case of B occurred in cellular phone sales when a user filed law suit claiming that the cellular phone has caused brain tumor.

Having pointed out a few drawbacks of our assumption, we proceed to discussion of what can be said about the future. The assumption of forbidden abrupt change allows us to at least predict where we are on the usage curve approximately. We can argue, for instance, whether we have reached the mid-point

or not. The utility of this analysis is that it allows us to predict whether to expect significant change at usage levels or a modest change in the near future.

To explain, consider three usage levels sampled at three different times. Suppose that at times t_0, t_1, t_2, we have the measured usage levels of y_0, y_1, y_2. We can use interpolation polynomials to fit a curve that crosses the three points $(t_0, y_0), (t_1, y_1), (t_2, y_2)$. Since three points define a second degree polynomial, we can use the equation:

$$y = y_0 \frac{(t-t_1)(t-t_2)}{(t_0-t_1)(t_0-t_2)} + y_1 \frac{(t-t_0)(t-t_2)}{(t_1-t_0)(t_1-t_2)} + y_2 \frac{(t-t_0)(t-t_1)}{(t_2-t_0)(t_2-t_1)}$$

to derive the desired curve. If at present time the curve has an increasing rate, then we conclude that we have not yet reached the mid-point of the lifetime curve and thus steep increases are expected in the near future. If this curve has a decreasing rate, then the expected increase will probably begin to level-off, and future increases will not be as steep as what has been observed in the past.

Example: Traffic readings on the KSC backbone network has shown that the utilization has doubled every year for the last two years. If we denote the utilization of two years ego as u, then last year's utilization is 2u, and the current utilization is 4u. If time of two years ego is $t_0 = 0$, then last year is $t_1 = 1$, and this year is $t_2 = 2$. Using these, we evaluate the above equation at points

$$(t_0, y_0) = (0, u);$$
$$(t_1, y_1) = (1, 2u);$$
$$(t_2, y_2) = (2, 4u);$$

and we obtain:

$$y = u(\frac{t^2}{2} + \frac{t}{2} + 1).$$

This equation has the derivative:

$$y' = u\left(t + \frac{1}{2}\right)$$

which is positive for every positive value of u and t. Evaluated at time t=2, the slope is y'=(5/2)u, where for t=1, it is y'=(3/2)u. This means that the increase now has a higher rate than one year ego, and even if this rate does not change in future, we can expect steep increases in the volume of communication on the network. Moreover, we also conclude that the current utilization rate is below the 50% level of the final saturation level of user demand, as beyond the 50% level the rate of increase would decline instead of increase.

IV. NETWORK PERFORMANCE ANALYSIS

The above model of trend analysis focuses on the required network usage by the user population, and does not say anything about whether or not the available system will be able to provide the required bandwidth to satisfy user community. To predict if the network will still satisfy the user needs in future, we need a method of performance analysis for the available system.

The system of computer networks at KSC is probably one of the most complex networks in the United States. It serves more than 10,000 users under three administrative authorities, using nearly 150 local network segments. All of the major vendors and networking protocols are represented, including Appletalk, Novel, Ethernet, token ring, DECnet, etc. These networks are connected together by three major backbone local area networks, which are in turn connected through the Kennedy metropolitan area network. A second metropolitan area network is dedicated to connect these three major units to other NASA centers.

No method of analysis, whether analytical or simulation-based, can handle the system of networks at KSC as a whole. While analytical models in the literature can yield very accurate results when a sufficiently detailed model of the network is considered, these models are mostly specialized for a single network running under a single protocol. Simulation models generally allow even more accurate representation of the real world environments, but they are inefficient and take a long time to run even for simulating a single network protocol in isolation. When nearly one hundred networks connected through multiple backbones is considered, the required running time of the simulation program will be prohibitive.Here we present a simple method that can be used for estimating

performance bounds in any network regardless of the communication protocol. The major advantage of the method is its remarkable simplicity.

Consider some network with 1 Mb/s bandwidth, 1000 users connected to it, and average frame size 1000 bits. Clearly, if each user generates data at rate exceeding 1 frame per second, then the total offered load exceeds the network bandwidth of 1 Mb/s, and the delay at each station will grow without bound. Below this level of traffic, we identify two regions of delay: a region of low delay where capacity is more than adequate to handle the offered load, and a region of high delay where the network becomes a bottleneck (i.e. a source of noticeable delay). The crucial question is: What region will the network operate in, based on projected offered load. To develop a method of analysis, we define five quantities:

X : the average time required to transmit one frame, once medium access has been gained by a user.

t : mean time that a user spends between two frames generated by that user.

D : system delay, which is the sum of queuing and transmission delays experienced by an average frame.

S : mean total throughput (number of frames sent per unit time) on the network for all users.

N : the number of users having direct access to the network.

Here we wish to use t and X as the input variables and compute upper bounds on S and N, and a lower bound on D. Before proceeding further, we also define two intermediate values in terms of the input variables t and X.

Define:

$$\lambda = \frac{1}{t}$$

$$\rho = \lambda X$$

Intuitively, λ measures the arrival rate of frames at each user cite, since a user generates one frame every t seconds. Also, ρ measures the utilization by a user of the network. It equals to the arrival rate, normalized by the amount of time taken to send one frame.

Having defined some system parameters, we can readily make some general comments about recommended usage levels for computer networks. To do so, consider the general behavior of a queuing system as shown below.

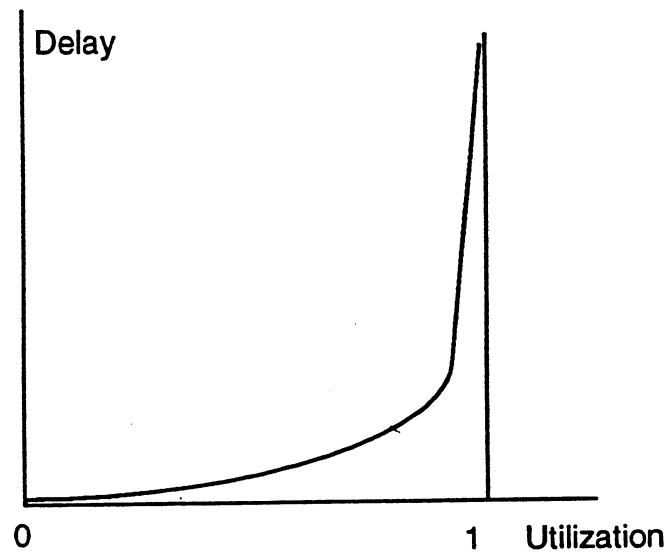

Figure 4: Typical delay curve for computer networks.

As shown, delay tends to infinity as utilization tends to 100% level. Below this level, delay may appear small, but such values may be deceiving. For example, we will show later in the case of an FDDI network that for about 50% utilization the mean delay could be around 40 microseconds. Yet the utilization level giving this average behavior frequently reaches 100% level giving rise to frequent occurrences of network congestion. The reason for this abnormal behavior can be easily explained by the Poisson property. Arrivals to a queue following poisson property are extremely irregular, so much that if the mean value of the inter-arrival times is t, its variance is t^2. (A good source is [2]).

It follows from this observation that, if t is at a level that will yield 50% utilization on the network, there are frequent occurrences of extremely long queues. To avoid frequent congestion that would arise from this phenomenon, it is necessary to keep the utilization level below 50%.

IV.1 Approximate methods for multi-access (shared) networks:

Here we give approximation methods for throughput and delay analysis for any network. In the appendix we will give extremely simple methods to derive the exact values for delay for some specific cases. To find an upper bound on the total throughput, assume that $t > 1$, and thus all the messages generated by a user are transmitted with some bounded delay. Then the throughput for one user is just $S = \lambda = 1/t$. Since each of the N users have this amount of traffic, the total throughput is:

$$S = \frac{N}{t}.$$

This upper bound cannot increase more than the total network capacity, which is:

$$S \leq \frac{1}{X}.$$

Thus, equating the right had sides of the two equations above and solving for N, we find that:

$$N \leq \frac{t}{X}$$

or, using $t = 1/\lambda$, we have:

$$N \leq \frac{1}{\lambda X} = \frac{1}{\rho}.$$

When the equality is satisfied the network is congested at the saturation point. For N less than this value the network is able to transmit all the frames submitted to it, but the amount of delay may be large if N is close to its maximum value.

We can compute a lower bound on delay by using M/G/1 approximation. The justification for the approximation is that the users can at best have access to the network like there is no interference due to network access protocols. Let

$$\rho = \frac{NX}{t}$$

denote the utilization by all users of the network. In terms of ρ delay for M/G/1 queue is given as:

$$D = X\left(1 + \frac{\rho}{1-\rho}\right);$$

for exponentially distributed service times, and,

$$D = X\left(1 + \frac{\rho}{2(1-\rho)}\right);$$

for deterministic service times (i.e. all customers have equal service time). The first of these two formulae may be used when the frame size is variable for a network, and the second may be used when the packet sizes are fixed.

An interesting property of these delay equations is that they are both in the form

$$D = Xf(\rho).$$

In other words, both of these are equal to X multiplied by some function of utilization. Since X is the time that would be needed to send an average frame, once a user gains control of the network, the function $f(\rho)$ may be used directly to obtain a comparative measure of the delay.

Here we give some examples to illustrate the use of these equations.

Example 1: A workstation is attached to a 1-Mb/s network, and it generates, on average, three frames per minute, with messages averaging 500 bits. Thus, message transmission time is $X = 500\mu s$, and the mean inter-arrival time is t=20s. The maximum number of workstations that can be supported is:

$$N = \frac{t}{X} = \frac{20}{500 \times 10^{-6}} = 40{,}000.$$

Example 2: On a 100-Mb/s network, each user generates 30 frames of 36000 bits every second. This situation would arise, for instance, if video information is sent over the network. Thus $X = 36 \times 10^{-5}$ seconds, with $t = 33.33 \times 10^{-3}$ seconds. Then, using the above formula, we find that N=92.

This example illustrates that at most 92 users can use the network at the same time. In an organization, it may be reasonable to assume that only about 5-10% of users are engaged in video communication at any time. Using the 5% level, we find that total number of employees in such an organization should not exceed $92 \times 20 = 1840$.

Example 3: For the above example, suppose there are 90 users actively engaged in video communication. Then,

$$\rho = \frac{90 \times 36 \times 10^{-5}}{33.33 \times 10^{-3}} = 0.972,$$

and, by using the second delay formula, the corresponding maximum delay will be at least:

$$D = X \times \left(1 + \frac{0.972}{2(1-0.972)}\right) = X \times 18.41.$$

By using the above value for X, we can easily compute that delay is at least D=6.63 milliseconds. While this delay may appear insignificant at first, we remind the reader that it represents at least 18 times the delay that would occur on a lightly utilized network. This indicates that the network is very close to its total congestion state.

For the purpose of approximate calculations X may be replaced for a larger value X' if we wish to account for overheads due to network protocols. The following example illustrates this point.

Example 4: In an FDDI network, the end of each packet is indicated by a free token. Assume that the number of bits in the token is negligible compared to the total frame length. However, two consecutive stations are 10 Km apart, which means that time taken for the next workstation to know that it can send messages is equal to the token flight time. If we assume that network propagation delay is 1' per nanosecond, then we have $X_{ov} = 3.28 \times 10^{-5}$ seconds. Since $X' = X + X_{ov}$, replacing every occurrence of X for X', and performing the computation as above, we find that only 84 users will actually saturate the network. Here we assumed that each time the token is obtained, a user sends just one packet, even if there are many waiting to be sent.

A more detailed treatment of delay for FDDI is given in the appendix.

IV.2 Methods for switched systems:

In telephone circuits, the criterion used to upgrade the system is as follows [3]. Traffic levels are monitored continuously, and peak hours of utilization are determined. Traffic for 10 of the busiest hours (30 hours in Europe) occurring on distinct days are averaged and the corresponding probability that an arriving customer finding all the switches busy is computed. If this probability is between 0.001 and 0.01, then grade of service is considered good. If the probability exceeds 0.01, then more switches are added to the system.

To adopt a similar approach for computer networks it is necessary to have a model that is applicable to both multi-access systems and to switched systems. For this purpose, we can model the FDDI network as a time division multiplexer. If each user required 1Mb/s bandwidth, we divide the total bandwidth of 100 Mb/s into 100 switches. If each user requires 20 Mb/s, this corresponds to having 5 switches. With this analogy, we can compute the utilization level needed for the switching system to give 1% probability of queuing. Having found the utilization level, we can then compute the corresponding delay for the FDDI network. This should give a good idea about the kind of delay faced if we used the same criterion as those of telephone systems.

In queuing terms, switched systems can be modeled as M/M/m queuing systems and solved by the help of Markov Chains. Unfortunately, there appears to be no simple way to explain these models, thus in this section we just summarize the main results and explain how these formulas may be used.

The M/M/m system is a queue with m servers. Each server may correspond to a switch in a telephone exchange system. Alternatively, for data networks such as ATM, each server corresponds to an input port.

Define: $\rho = \dfrac{N\lambda X}{m}$

Here N is the number of users, λ is the arrival rate of messages generated by each user, X is the service time for each call (or packet), and m is the number of switches. For an arriving customer (or packet, or telephone call) probability of finding all servers busy, thus having to wait is:

$$P_Q = \frac{1}{1+(1-\rho)\dfrac{\sum_{k=0}^{m-1}\dfrac{(m\rho)^k}{k!}}{\dfrac{(m\rho)^m}{m!}}}.$$

The number of frames waiting is: $N_Q = P_Q \dfrac{\rho}{1-\rho}$,

and the corresponding waiting time is:

$$W = \frac{N_Q}{\lambda} = \frac{N_Q}{N} = \frac{P_Q \rho}{N(1-\rho)}.$$

These are well known equations originally due to Erlang. For a given value of P_Q it is not possible to solve for the corresponding value of ρ since P_Q is a complex function of m and ρ. We therefore write computer programs which solve for the desired values by iterative methods. Sample results of these algorithms are given below.

# Of switches	Utilization (pr.=0.01)	Utilization (pr.=0.05)	Utilization (pr.=0.1)
10	0.409	0.529	0.599
20	0.550	0.651	0.706
30	0.621	0.709	0.757
40	0.665	0.745	0.788
50	0.697	0.770	0.809
60	0.721	0.789	0.825

Table 2: Utilization values for various grade-of-service parameters and different system sizes.

As can be observed from this table, the utilization values increase quite high for increasing number of switches regardless of what P_Q value is used. This shows that, by the time telephone industry reaches P_Q =0.01, they are indeed receiving quite high utilization levels. This is particularly true for very large number of switches, (which is the case normally in industrial settings) when utilization approaches close to 100% even for P_Q =0.01. We have already pointed out that utilization levels above 50% are not advisable for multi-access data networks. This shows that the criteria used in telephone industry is not applicable for multi-access networks.

As an example of this, we illustrate a hypothetical situation of using the FDDI network for multimedia applications. First, recall that maximum packet length for FDDI is 36000 bits. Assume that part of the computer screen of size 512 by 512 is used for video, and the rest is used for text. The video part, in 8-bit color and 32 to 1 compression ratio comes to about 30,000 bits. Thus each screen nicely fits in a

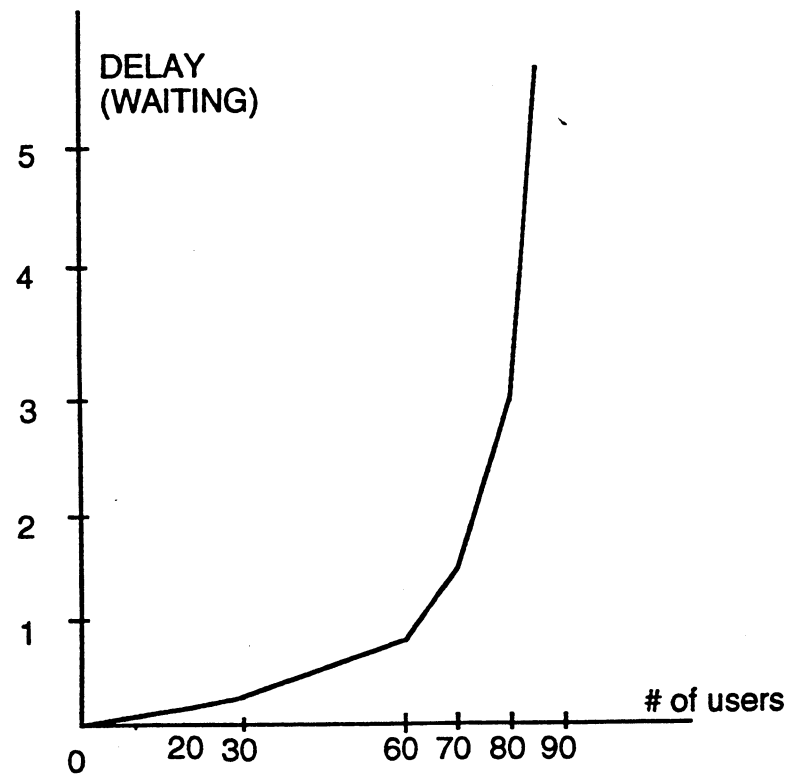

Figure 5: Delay v.s. utilization in FDDI (M/G/1 approximation with deterministic service times).

single FDDI packet, leaving ample room for the accompanying text. As a result, we assume that each user sends 30 such packets every second (corresponding to refresh rates for video pictures). The question we want to answer is how many users can be supported.

Using the delay equation for M/G/1 case (which is a generous model for FDDI network) we plotted the delay as a function of increasing number of users as shown above.

Here we used the deterministic equation since all the packets have equal length of 36000 bits. The vertical axis shows the delay in multiples of X, which is, as before, the time taken to send a single packet. X itself is very small, in the range of a few tens of microseconds. Never the less, as the number of users reaches around 90, the multiplication factor on the vertical axis reaches close to infinity. Thus, using the 50% rule of thumb, we should not really try to support more than about 45 users on that network. Contrasting the situation with the switched systems, we see that for 60 switches we reach the probability $P_Q = 0.01$ at about 72% utilization. This is clearly a forbidden region to operate the FDDI network.

V. CONCLUSIONS AND RECOMMENDATIONS

This research has addressed the problem of developing practical criteria to upgrade computer networks. The problem turns out to be extremely easy and difficult simultaneously. It is easy when network traffic generated by those applications deemed desirable far exceeds the network capacity. It is difficult, however, when the current utilization levels are far below the network saturation point.. For this reason, we suggested various cost/performance factors that may be used to asses the desrability of increasing the offered bandwidth. We also suggested methods for forecasting future traffic patterns and interpolation methods to determine the current position in a demand curve. Simplified analytical tools are described and examplified to predict network behavior under expected future utilization levels. Some specific recommendations resulting from this research include:

1) Network utilization at or above 50% levels is not recommended due to frequent congestion.

2) A method of monitoring network traffic based on application is desirable. The obtained data can better represent expected traffic growth by different applications, leading to more accurate predictions.

3) A traffic measurement method similat to that used in telephone industry is advisable. Average traffic levels can be very low, yet the traffic at busy time is a more accurate representation of what the users experience. It is advisable to record the 10 busiest traffic levels in a year for use in decision making.

4) It is advisable to use the forecasting method presented in this paper, based on past data, to see how accurately it reflects the future, using the "future" data collected later.

VI. REFERENCES:

1. J. Y. Hui, J. Zhang, and J. Li; "Quality-of-Service Control in GRAMS for ATM Local Area Networks," *IEEE Jour. SAC,* 4 (13), 700-709.

2. D. Bertsekas and R. Gallager, Data Networks, (second edition), Prentice hall, 1992.

3. G. Held, Data Communications Networking Devices, (third edition), John Wiley & Sons, 1992.

APPENDIX

Derivation of M/G/1 delay equation:

Here we offer a simple minded but exact derivation for delay equations considered in this paper, in order to provide further understanding of how these equations may be used. Consider the different components of delay in a queue with a single server. Specifically, we focus on the time W spent waiting in the queue. Once we compute W, the total delay is just:

$$D = X + W.$$

The waiting time consists of:

$$W = R + XN_Q$$

where R is the remaining time for the frame being serviced, X is the average time needed to send one frame, and N_Q is the number of frames waiting to be serviced ahead of the arriving frame. From Little's law, we know that N_Q is equal to throughput times delay, that is, $N_Q = \lambda W$. Using this in the above equation and solving for W, we find that

$$W = \frac{R}{1-\rho}.$$

Now it remains to compute R, the expected value of the remaining service time for the frame being transmitted. If all the messages have equal length (deterministic service time assumption), and the transmission has just started, then the remaining time is obviously X. If transmission is about to finish, then the remaining time is zero. The average value is thus X/2, but we must also consider the fact that this delay is experienced only while the served is busy (an even which occurs with probability ρ). If the server is not busy (which happens with probability 1- ρ) the term R is equal to zero. Thus we have

$$R = \rho\frac{X}{2} + (1-\rho)\times 0$$
$$= \frac{\rho X}{2}.$$

Using this in the above formula, for systems in which all frames have equal length, we find that

$$W = \frac{X\rho}{2(1-\rho)}.$$

If the frame lengths are not equal, then we need to know the distribution of frame lengths in order to derive the value of R above. From practical observations, it has been found that an exponential distribution can be used to represent the distribution of frame lengths in data networks. One interesting property of exponential distribution is its "memoryless" property. This means that no matter how long time has been already spent servicing a customer (or transmitting a frame in a data network) the expected value of the remaining time is statistically equal to the mean service time (or mean frame length). Many examples of such phenomena exist in real life. The most famous example is the light bulb; no matter how long a light bulb has been in use, the expected remaining time it will last before burning out is equal to the mean lifetime over all light bulbs.

With these consideration, we expect that, given that the system is busy, the length of time needed to finish a frame in transmission is just X instead of X/2. in other words, we have $R = \rho X$, and using this we find

$$W = \frac{X\rho}{1-\rho}.$$

Delay equation for FDDI network:

So far, we assumed that the amount of overhead due to network protocols is negligible. If this overhead is known, or can be approximated, then we can also replace X for X', where $X' = X + X_{ov}$. This reasoning assumes that every packet sent experiences the overhead, and such a method can be applied for and FDDI network where each time a station acquires the token it sends only one packet. In an exhaustive algorithm where a station empties its queue, the situation is more complicated. Below we treat this case in detail to give an idea about how to approach such problems.

Consider the equations we have seen before for the M/G/1 queue:

$$D = X + W$$
$$W = R + X \times N_Q.$$

In FDDI Token Ring with exhaustive policy an arriving frame waits for the time periods of R units for the current frame to finish transmission plus an additional time needed to acquire the token. That is,

$$W = R + V + X \times N_Q.$$

where V is the time needed to acquire the token, and the rest of the terms are as before. Here we do not alter the term $X \times N_Q$ since once the token has been received all the waiting frames are transmitted. We just need to add the term V to R, since R is now larger than what it would have been if the user had not waited for the token.

To compute V, let v denote the time needed to pass the token from one user to the next. In an FDDI network, this time can be approximated by the network propagation time since the idle token immediately follows the last message sent by a user. Suppose that there are N users with equal distance from one another. Then, total propagation time of the token around the network is Nv. In the worst case, a new frame arrives just after the local user has finished passing the token to its neighbor, and the new frame may need to wait for N-1 other users before it obtains the token again. In the best case, token may be already at the local station, thus no time is needed for token rotation. The average case is, however, the case of waiting for (N-1)/2 users, but all of this wait occurs only if the network is utilized by some user. This happens with probability ρ. If the network is not being used, an event which occurs with probability 1- ρ, then the arriving frame may arrive just as the token arrives to that station but has no chance of being transmitted, and its average waiting time will be Nv/2. Expressing these quantities in more formal terms, we have

$$V = \rho \frac{(N-1)v}{2} + (1-\rho)\frac{Nv}{2}.$$

After simplification, we obtain

$$V = \frac{(N-\rho)v}{2}.$$

Since we already knew how to compute the rest of the terms in delay equation (e.g. the terms R and $X \times N_Q$), combining all of these terms we obtain

$$W = \frac{X\rho}{1-\rho} + \frac{(N-\rho)v}{2(1-\rho)}.$$

This is the waiting time expression for an FDDI network when packet lengths are distributed exponentially. We simply divide the first term by 2 when packet lengths are equal. Justification for this is similar to that we gave for the M/G/1 queue. Note that, throughout the appendix we defined $\rho = \lambda X$.