Analysis and Modeling of Wide-Area Networks: Annual Status Report*

Hans-Werner Braun
Bilal Chinoy
Kimberly C. Claffy
George C. Polyzos

Applied Network Research
San Diego Supercomputer Center
and
Computer Systems Laboratory
University of California, San Diego

February 13, 1993

Abstract

This annual report presents the progress of Applied Network Research Group at the San Diego Supercomputer Center (SDSC) and the University of California, San Diego (UCSD) during the past year. Research topics include: analysis of the existing data and instrumentation of the NSFNET; sampling network traffic data in wide area environments; end-to-end delay and jitter across wide-area networks; routing stability and characteristics; reliability and quality of service metrics; performance evaluation of a multimedia application; and individual statistics of interest for performance evaluation and modeling of a wide area environment. We also include research efforts outside the scope of this proposal but within the scope of the larger ANR research agenda, such as our involvement with the CASA gigabit network infrastructure and our participation in NSF's NREN Engineering Group.

*The projects described in this report are supported by a grant of the National Science Foundation (NCR-9119473), a joint study agreement with IBM, and the CASA gigabit research project with CNRI, NSF, and DARPA. Any opinions, conclusions, or recommendations in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation, other supporting organizations, General Atomics, SDSC, UCSD or the SDSC Consortium members.

Contents

1 Introduction

In 1992 SDSC's Applied Network Research Group, in collaboration with UC, San Diego, embarked on a three-year project to analyze and model wide-area datagram networks. We present in this report our progress during the past year, using as a framework the research outline in our 1991 proposal to NSF[1] [6]. We also include research efforts outside the scope of this proposal but within the scope of the larger ANR research agenda.

During this past year, we focused on the deliverables planned for the first year in our proposal to NSF:

> Analysis of already accumulated and current NSFNET data.
>
> Initial study and requirements assessment.

ANR has accomplished a great deal in in pursuit of these deliverables. We have surveyed and analyzed existing data sets on the NSFNET backbone as well as active measurement of a U.S. federal interagency domain at a multiagency network interconnection point. In the process, we have defined key research issues and reevaluated our initial expectations outlined in the proposal. For a general description of the NSFNET, see Chinoy [4].

For the second year, we planned the following two deliverables, the second of which we have already begun to address:

> Creation of realistic models for NSFNET and NREN traffic.
>
> Investigation of the Interagency/NREN system.

In the coming year we look toward developing more refined data sets to allow theorists to develop realistic models for NSFNET and NREN traffic.

As indicated in the original proposal, we planned to use the third year to apply the results of our analysis and models to traffic characterization on the Internet. Our proposed deliverables were:

> Analysis and modeling research on an Internet-wide scale.
>
> Application of analysis work and resulting models to the entire system.
>
> Feedback and re-evaluation of analysis and modeling work.

One of our original project goals was to arrive at a consolidated Internet model to describe the host-network interaction using a systems theory approach. We have modified this goal based on what we have learned about the environment during the first year. Network analysis in high speed realms involves new and uncharted territory, and, as a result, maintaining a sense of direction toward our general goals has been an iterative process of defining and refining analytic models to describe systems behavior. Although we are currently only in preliminary stages towards conceptualizing models, we expect to apply our findings regarding the overall effort to the NSFNET and other components of the Internet. We also hope to apply our efforts to the ATM and gigabit research efforts in which SDSC is participating: SDSC's connection to the DOE/NASA national ATM network project and the CASA testbed project.

Our initial status report [5] presented a brief survey of the existing literature related to wide area network traffic characterization, both in the analytical and performance measurement domains. In that status report we also introduced our network-centric approach to empirical analysis. We have found that the state of networking technology, in terms of speed and traffic diversity, has advanced at a far faster rate than has the analytical and theoretical understanding of network behavior. The slower and more containable realms of years ago were amenable to characterization by closed-form mathematical expressions, which allowed prediction of performance metrics such as queue lengths and network delays.

Traditional mathematical modeling techniques, and in particular, queueing theory, have met with little success in today's networking environments. However, the need for network analysis has not diminished; on the contrary, realistic methodologies for understanding network behavior play an even more essential role in facilitating future evolution into gigabit/sec speeds and beyond. It seems inevitable, however, that simulation and empirical techniques based on observed behavior will play a larger role than traditional mathematical techniques have played in the past.

Because wide area network performance is subject to the influence of so many network components, the complexity of any realistic model often renders it intractable. Often a manageable performance model must trade off accuracy for resource-constrained usability. For example, commonly used performance measurements such as throughput or latency measurements between end-systems aggregate many unrelated influences on network performance. We advocate the development of simplified, more comprehensible, and verifiable performance expressions. A comprehensive model of performance dependencies requires greater insight into performance parameters. In the following sections, we present various studies of such depen-

[1] *Analysis and Modeling Tools for High Speed Network,* funded by NSF Grant NCR-9119473.

dencies.

Section 2 presents a discussion of the existing data and instrumentation of the NSFNET. Section 3 focuses on sampling network traffic data in wide area environments. Section 4 presents a study of asymmetries in end-to-end wide area network latencies. Section 5 discusses a study of packet spacing fluctuations across the network and its component parameters. Section 6 discusses an experiment to examine NSFNET routing stability. Section 7 presents a study of the ability of route caching to take advantage of the high degree of traffic locality in NSFNET traffic. Section 8 presents an experiment on the end-to-end reliability between SDSC and several Internet sites. Section 9 presents initial performance evaluation of a recently released packet video application. Section 12 discusses ANR modeling efforts. Section 10 discusses SDSC's role in the development of the CASA gigabit network infrastructure. Section 11 discusses individual statistics of interest in a wide area environment. Section 13 presents ANR efforts in NSF's NREN Engineering Group, and Section 14 enumerates several ANR meetings and presentations of interest. Section 15 presents conclusions and our perspective on the ANR's future research agenda.

2 Existing data and instrumentation

We have spent a significant part of the first year exploring the existing environment, with an aim toward instrumentation in support of specific measurement studies. The packet-switching nodes on the NSFNET backbone are instrumented to operationally collect certain statistics. We investigated the usefulness of these statistics in the context of our analysis and traffic characterization effort. We have also maintained contact with the NSFNET service providers and suggested changes and/or augmentation to this instrumentation.

2.1 NSFNET T1 and T3 statistics

In this section we provide an enumeration of the objects collected on both the current T3 NSFNET backbone and those which were collected on the T1 backbone during its operational lifetime.

2.1.1 Interface counters

Table 2.1.1 illustrates the SNMP objects collected on the T1 and T3 backbones.[2] The granularity of the collection process is every 15 minutes.

[2] The software packages which perform the collection on the T1 and T3 backbones are NNStat [3] and ARTS [2], respectively.

2.1.2 Traffic characterization objects

Table 2.1.2 illustrates the traffic characterization objects collected on the T1 and T3 backbones. The granularity of the collection process is every 15 minutes.

2.1.3 Other statistics collection

On the T1 NSFNET backbone Merit collected latency statistics among all backbone access points. On the T1 NSFNET backbone ANS continues to support collection of this data between the E-NSS nodes of the network.

2.2 T1 NSFNET backbone traffic characterization

In pursuit of a greater understanding of the existing instrumentation for the NSFNET backbone, we performed an in-depth study of the operationally collected metrics on the T1 backbone for May 1992. This study is available as a technical report from SDSC and will appear in the proceedings of *IEEE INFOCOM 1993* [7]. We present the highlights of the study here.

We first discussed the measurement environment and approach to data collection. We then presented some traffic characteristics of the T1 NSFNET backbone. We included both long term characterizations, essentially for the life of the T1 network, as well as more detailed results for May 1992. The long-term data were presented on a monthly basis and were obtained from publicly available summaries of measurements published by Merit Network, Inc.

The measured quantities included: long-term growth in traffic volume, including attribution to domains and protocols; trend in average packet size on the network, both over long and medium term intervals; most popular sources, destinations, and site pairs; traffic locality; international distribution of traffic; mean utilization statistics, both of the overall backbone as well as of specific links of interest; delay statistics; and assessment of downtime for the last few years.

We made the following observations about the T1 NSFNET backbone from the data. Many of the traffic characteristics apply to the T3 ANSnet backbone as well. Traffic both in packets and bytes and the number of networks (in the sense of assigned IP address families to campus and other networks served by the T1 backbone) had steadily increased since the network installation. At approximately the end of 1991, the T1 traffic volume dropped off, as the NSFNET project began to divert traffic to the T3 backbone. The increase in traffic volume in bytes is quadratic, while the increase in networks served is linear.

Table 2.1.1: SNMP objects collected per node every 15 minutes[3]		
T1 backbone	description	**T3 backbone**
ifOperStatus	operational status	
ifDescr	interface descriptors	ifDescr
ipAdEntIfIndex	IP address corresponding to interfaces	ipAdEntIfIndex
ifInErrors	incoming errors occurring interface	ifInErrors
ifOutErrors	outgoing errors occurring on interface	ifOutErrors
sysUpTime	system uptime	sysUpTime
ifInOctets	bytes entering interface	ifInOctets
ifOutOctets	bytes exiting interface	ifOutOctets
ifInUcastPkts	unicast packets entering interface	ifInUcastPkts
ifOutUcastPkts	unicast packets exiting interface	ifOutUcastPkts
	non-unicast packets entering interface	ifInNUcastPkts
	non-unicast packets exiting interface	ifOutNUcastPkts
	mapping from remote address to interface index	is-isIndex

Table 2.1.2: Packet categorization objects collected per node	
T1 backbone	**T3 backbone**
relative to exterior nodal interface	
source-destination matrix by network number (packets/bytes))	
TCP/UDP port distribution, well-known subset (packets/bytes))	
distribution of protocol over IP (e.g., TCP, UDP, ICMP) (packets/bytes)	
Packet-length histogram at a 50-byte granularity	NA
packet volume going out of backbone node	NA
NSS-centric (entire node)	
per second histogram of packet arrival rates	NA
NSS (intra-NSFNET) transit traffic volume	NA

Most of the traffic volume was within the research community, based on available network attribution data from the DDN NIC.[4] The highest volume applications are file transfer (using FTP) and network news distribution (using NNTP). A recent development of concern is that much of the traffic cannot be directly attributed to protocols and applications because of the proliferation of traffic using non-standard TCP/UDP port numbers or other transport protocols.

Monthly summaries of mean packet size do not reveal any particular trend, but fifteen-minute data shows daily cycles which are compatible with the hypothesis of bulk transfer applications, using larger packet sizes, intensifying during the U.S. off-peak hours, or, correspondingly, that interactive activity, generally characterized by smaller packet sizes, drops off during these off-peak hours. Delay statistics (the monthly median of sample packet delays obtained with `ping` at fifteen-minute intervals) revealed that typical end-to-end delays on the backbone did not exceed 100 milliseconds and typical link delays did not exceed 45 milliseconds.

As was true two decades ago in the ARPANET environment, traffic favoritism was high. For example, 0.28% of the (customer/campus/site) network pairs generated 46.9% of the traffic for May 1992. Link utilization was high, even following the diversion of a considerable proportion of the traffic to the T3 network. The mean overall utilization for the month of May 1992 was 15.4% while 5 links had more that 30% mean utilization for the month. Over fifteen-minute intervals, utilization of highly utilized links typically exceeded 50% and sometimes 80%. The most heavily used link for the month, College Park to Houston, had utilization almost always exceeding 20% (for fifteen-minute intervals) and more than 50% during the peak hours of the day. Interestingly, the reverse direction, Houston to College Park, had almost uniformly lower utilization.

The available data hold further potential for analysis which can lead to a better understanding of traffic on the network. However, there are limitations which make exploration of some interesting questions problematic, in particular those involving correlations between instantaneous network performance and traffic intensity and characteristics. Complicating the task are the enormous difficulties with methodological data collection in such a large operational environment. These problems explain, in our view, the lack of other similar studies for wide area networks. They also render any traffic data particularly valu-

[4] Defense Data Network's Network Information Center.

able and significant.

We expect to continue our research and refine our methodologies in the process of applying them to new realms. In particular, we plan to perform similar investigation of the T3 networking environment, as well as the CASA gigabit network project and its planned infrastructure.

3 Sampling network traffic data

In response to Merit's reversion to sampling on the NSFNET backbone and accompanying interest in its effect on traffic analysis, we are studying methodologies for sampling network traffic data. This investigation includes methods and accuracy requirements for the characterization of aggregated traffic on a component network. Our objective is to examine the effect of sampling on the ability to answer selected questions about network traffic characteristics. Because the high network transmission rates of current and future networks make it more difficult to capture and store the data required for traffic analysis, the sampling study has particular relevance for the NSFNET and NREN.

Using a selected packet trace from an existing highly traffic aggregating network environment, we simulated various sampling approaches, including time-driven and event-driven methods, with both random and deterministic selection patterns, at a variety of granularities. We then evaluated the sampled traces and used them to estimate several parameters of the full data set, with associated confidence intervals. The metrics we are currently exploring are the distributions of packet size and interarrival times, and we expect to also focus on network applications and distributions of the geographical locality of traffic according to network number pair matrices. We intend to evaluate not only given classes of sampling methods, but also the range of granularities, or sampling fractions, within a certain class (see fig. 2). Both the class and granularity of a method will have an effect on the confidence intervals associated with estimates of a given target metric. For example, initial results indicate that event-driven sampling, i.e., every n packets, is more effective than time-driven sampling, i.e., every n microseconds.

Another consideration is how the non-independence of traffic time-series affects the efficacy of sampling. As we are testing our methodologies on a data set from a highly traffic aggregated operational environment, a well-selected sampling interval, regardless of the sampling fraction during that interval (or how much data one actually collects), can improve the accuracy of the image of the network which a sample can offer. How a sampling method might optimally adjust for the dynamically changing levels of burstiness is an objective of our study. A less ambitious goal will be a knowledge base upon which researchers of high speed methods can base monitoring and statistical data collection decisions.

4 Assessing unidirectional latencies

We undertook specific research into measurement considerations for assessing unidirectional latencies. Our investigation elucidated a myth related to the symmetry of latencies in opposing directions between two network end sites. To investigate the validity of the symmetry hypothesis, we actively measured single direction latencies to selected destinations of the Internet utilizing a variety of paths. In our specifically designed experiment, we sent packets from a source to destinations and back, and recorded four timestamps at the reception and transmission of the packet at both the source and the destination of the traffic. Our measurements demonstrate remarkable variances in the differences between the measured timestamps on the two paths, sometimes sustained for significant periods, and often simultaneous with increased round trip delays. Heterogeneous or distant paths between end-systems especially intensify this behavior.

These results codified our belief in the hypothesis that round-trip latencies are an insufficient and sometimes misleading method to determine unidirectional delays. This claim has significant implications for high speed, multi-application networking environments, which often require predictability of delays. For example, designers of real-time applications who intend to deploy their products on such a network, or particularly wide area, high speed, highly aggregated network, should be aware that across the periphery of such an environment, symmetric delay will serve unpredictably as a model of network performance.

A report of our findings is available from SDSC as a technical report, and will appear in the *Internetworking: Research and Experience* [8].

5 Jitter and delay variance

A fundamental metric of a packet switched network is the average latency associated with end-to-end communication. This latency determines a host of secondary factors such as bulk throughput or interactive response time. Additionally, the distribution of packet transmission delays is a key determinant in the feasibility and performance of real-time constrained protocols for voice and video. We have defined a scheme to quantify the variance in end-to-end packet delays along an Internet path, and to associate this variance with the perceived quality of service over that end-to-end path. Note that we constrain ourselves to network-imposed rather than host-imposed delays and variances. This latter class include delays due to operating systems, network and application software or the end users themselves.

5.1 Motivation

The Internet is a mesh of interconnected networks that are heterogeneous in both underlying technology as well as management policy. Any given path between two hosts on the Internet may traverse many different networks at varying bandwidths and loads. A simple scheme to quantify the *quality of service* of an end-to-end path on the Internet would allow one to calibrate performance. As we developed a mechanism of our own, we imposed one key constraint on potential schemes: They could not require information unavailable to the source and destination hosts. Thus, we did not consider any scheme which depended on information provided by network operators and/or routing authorities.

We defined a scheme that an end user could use to measure, calibrate and grade end-to-end service based on the variance in packet delays throughout the life of a connection. We defined service degradation "windows" in terms of the number of successive packets in a stream that suffered high delay or loss.

We have defined most of the key elements of the scheme and written software to measure them. We now briefly describe our methodology.

5.2 Description of Scheme

As a packet travels from its source to destination, it experiences various delay components which contribute to its end-to-end delay:

- *Transmission delay* pertains to the delay incurred due to transmission and multiplexing equipment that a packet traverses. For a given path and fixed equipment, transmission delay is a fixed quantity.
- *Propagation delay* is the physical delay that electromagnetic waves experience while traversing a given media. For example, fiber-optic transmission imposes a delay of about 50 to 60% of the speed of light, 186,000 miles/sec. For a fixed path, propagation delay is a fixed quantity.
- *Store-and-forward delay* depends on the switching equipment which routes the packets. A packet switch must assemble an entire packet before it can forward the packet to the next switch. Packet assembly imposes a finite delay, the size of which depends on both the packet length and the nature of the switch. For a fixed path and packet length, this quantity is stable through time, and thus does not contribute to the variance associated with end-to-end packet delay.
- *Queueing delay* is the most critical component of the end-to-end delay along a packet-switched path. Queueing delays depend on the volume of traffic imposed on a packet switch. Such a parameter is difficult to manage, since the traffic volume is a function of the burstiness of all of the various streams of packets that are statistically multiplexed through that switch and thus require its attention. Queueing delays can result in congestion and even packet losses. Our scheme addresses this component of delay in particular.

Consider a source host A and a destination host B, as shown in Figure 1. Assume that the source sends two consecutive packets P_1 and P_2 at times t_0 and t_1, respectively, and the destination receives them at times t_2 and t_3. In this case,

$$t_2 - t_0 = Delay\ for\ packet P_1 \tag{1}$$

$$t_3 - t_1 = Delay\ for\ packet P_2 \tag{2}$$

Since all other components except the queueing delays are time invariant, we represent them with the constant C. We assume that both packets traverse the same route and are of equal size. The following relationships then hold:

$$t_2 - t_0 = C + Queueing\ delay\ for\ packet\ P_1 \tag{3}$$

$$t_3 - t_1 = C + Queueing\ delay\ for\ packet\ P_2 \tag{4}$$

Subtracting **(4)** from **(3)**:

$$(t_3 - t_2) - (t_1 - t_0) = P_2\ delay - P_1\ delay \tag{5}$$

The above difference, termed *delay differential*, represents the difference between the inter-arrival times and the inter-departure times for two successive packets. Note that this difference is a time-variant quantity, and is potentially different for different pairs of successive packets. A positive differential indicates spreading between the two packets while a negative differential indicates that the packets have bunched together. Any *relative* clock drift between the sending and receiving hosts will also affect this difference, but because the clock drift imposes only a very slow, linear change, we can filter its effect out.

The result is a measure of the difference in one-way queueing delay experienced by successive packets. These differences represent the *spreading* or the *bunching* between

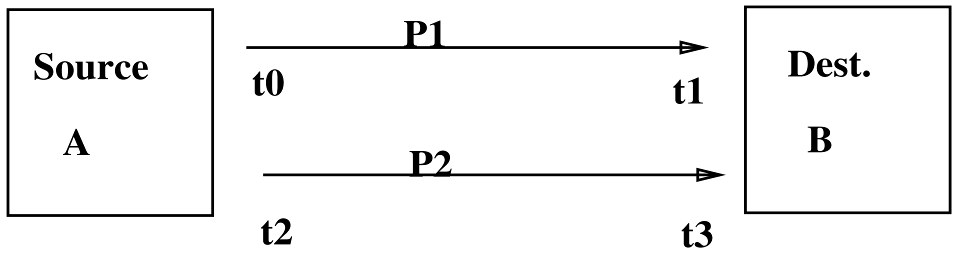

Figure 1: Packet Delays

packets due to queueing effects along the path. In the absence of queueing delays this difference would be zero and all packets would arrive with precisely the same interarrival times with which they left their source.

5.3 Results

We modified the popular `ping` software to compute the delay differential associated with each pair of packets. A sample run is graphed in Figure 2. We used this tool for a test between a source host at SDSC and a host in Japan. The first panel plots the round trip time between the two hosts. The next two panels plot the delay differential between successive packets on the y-axis as a function of the packet sequence number. The *forward* path is from SDSC to Japan, and the *reverse* from Japan to SDSC. Note the close correspondence between spikes seen in the delay differential plots and the round trip time variances. The last panel plots the lost packet sequence numbers; this panel again reveals some correlation between packet loss and spikes in the differentials. Congestion on either the forward or the reverse path is likely responsible for packets lost in this fashion.

A congestion window of length L for a *threshold* T is defined to be L successive packets that have been delayed by T or more milliseconds. This definition is relative to the first packet in the window. Note that successive packets in a congestion window cannot have negative differentials, as this would indicate a correcting action in the network. Figure 3 plots the distribution of congestion windows for the SDSC-Japan test. The plot shows that the number of congestion windows seen in the sample trace decreases as we increase the congestion threshold. Also, the average length of a congestion window decreases and equals 1 for larger values of thresholds, signifying the absence of long periods of high congestion.

5.4 Future work

We have presented a scheme to quantify the end-to-end performance of an Internet path. We need to further refine and calibrate our measurements to accurately reflect congestion windows. Our next step is to develop a simple graphical indicator of congestion along with a single number to express it. We hope this could serve as a valuable testing benchmark to express the quality of Internet end-to-end paths. A paper describing our scheme and results is currently under preparation and should be available as a GA technical report.

6 Routing stability

The Internet is a complex interconnection of networks that use a common suite (TCP/IP) of networking protocols. One of the key features of the Internet is the fact that all of these constituent networks are interconnected, thereby providing system-wide communication. Consequently, any change in network reachability is reflected at some intermediate point in the system, thereby providing a consistent up-to-date view of current connectivity. The magnitude and pattern of change in this information represent the connectivity stability of the Internet. The NSFNET backbone network provides transit services to a large portion of the global Internet and maintains routing tables reflecting this current connectivity. These routing tables are constantly updated based on information received by the attached regional networks. This study investigates the dynamics of routing information flow as presented to the NSFNET backbone network. We analyze routing changes that the backbone experiences and how they reflect the connectivity stability of the attached networks.

6.1 Motivation

Our motivation for this study derives from our recognition of the dearth of quantitative information about the dynamics of the Internet routing system, in spite of its key role in the stability of Internet connectivity. A plethora of routing protocols and techniques exist today, and most regional and campus networks are custom-engineered and configured. While some researchers have studied the be-

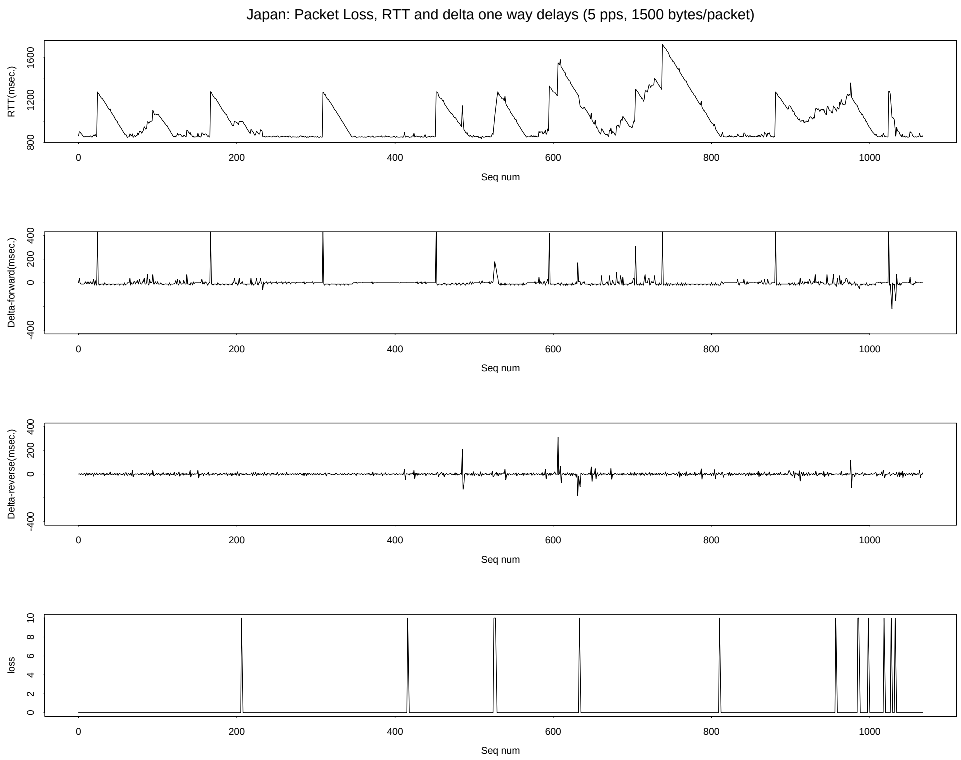

Figure 2: Delay Variances for a destination host in Japan

havior of routing protocols within the confines of a homogeneous network, no one has tackled the system. Perhaps as a substitute for rational investigation, much folklore has developed regarding the volatility of routing information and its effect on end-to-end stability. This study will, we hope, represent a first step in quantifying Internet-wide routing stability.

6.2 Internet routing

The TCP/IP Internet is organized as an interconnection of *Autonomous Systems* (hereafter known as ASs). An AS is a collection of internetwork routers managed and administered by a single authority or organization. A variety of routing schemes and protocols exist in order to maintain state information and compute paths, both within the confines of an AS and between neighboring ASs. Based on the above model, Internet routing protocols fall into two classes. *Intra-AS* protocols are used within the boundaries of an AS, and *inter-AS* protocols are used between ASs. Typically an AS uses a single routing protocol within its boundaries to generate and propagate routing information, though it is not uncommon to have a single AS use multiple routing protocols.

For routing purposes, the NSFNET backbone is modeled as a single AS into which regional network ASs connect. At each connection point, one or more regional network ASs interact with the backbone in order to share routing and reachability information. The backbone and its attached regional networks use either EGP (Exterior Gateway Protocol) or BGP (Border Gateway Protocol) as the inter-AS routing protocol. The NSFNET backbone itself uses for its intra-AS routing protocol a subset of the ANSI standard IS-IS protocol, adapted for IP networks.

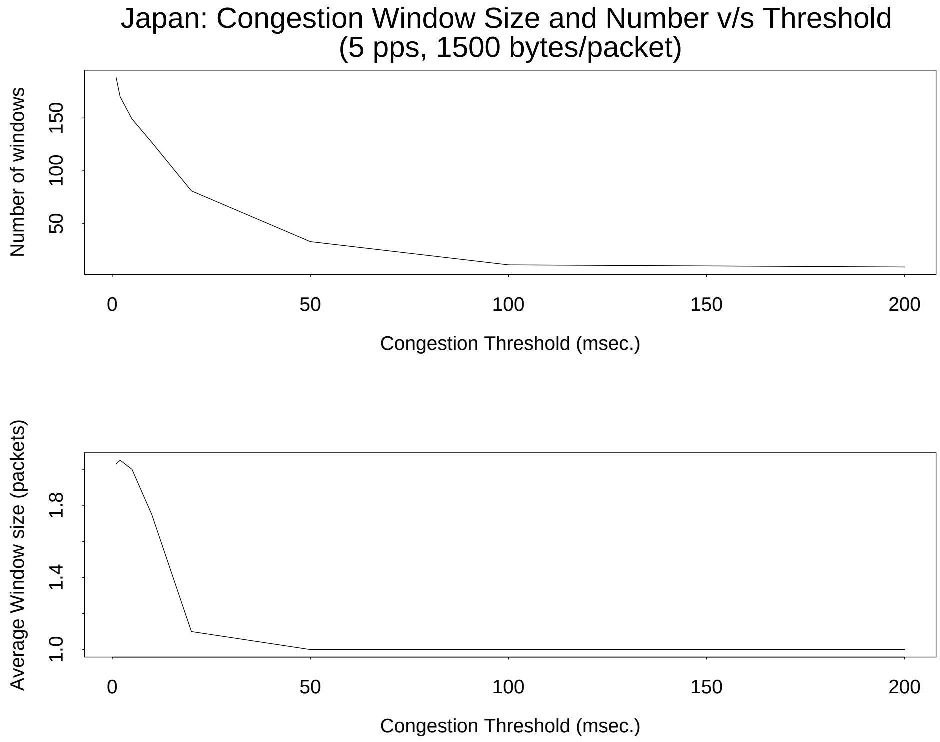

Figure 3: Congestion Windows for a destination host in Japan

6.2.1 Routing information content and flow

Every AS has a set of routers that participate in inter-AS routing on behalf of that AS, sharing knowledge about the state of that AS with the neighboring systems. These routers are called *border* routers, because they are at the borders of the parent AS.

Within the NSFNET system of networks, every IP network is required to belong to one or more ASs. This architecture allows route aggregation based on AS numbers, and indeed backbone nodes route datagrams to appropriate exit points based on the parent AS of the destination IP network address. For the EGP protocol, the information is in the form:

$$< net >< AS >< reachabilitystate(Y/N) >$$

That is, EGP only provides information about the binary state of reachability of a network. The BGP protocol added a number of improvements to EGP, notably attaching a meaningful *metric* of reachability and an *AS path* attribute for each network carried. The metric allowed border gateways to select from among multiple paths to the same network, and the AS path allowed rapid detection of routing loops. The BGP information, therefore, is in the form:

$$< net >< AS >< metric >< ASpath >$$

Both protocols are currently in use within the NSFNET system, but a transition toward using BGP alone is underway.

At every attachment point, the NSFNET backbone and attached regional network border routers periodically exchange information. The information exchange is bidirectional, with each AS updating its peer regarding reachable

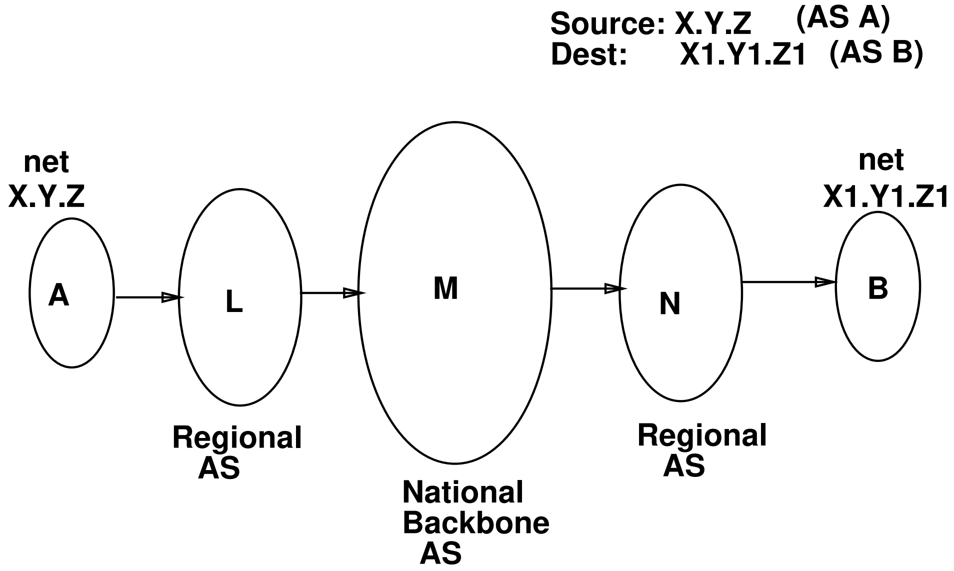

Figure 4: Routing Information Flow between AS A and B

networks. Consider the typical case shown in Figure 4, where net `X.Y.Z` belongs to AS A and net `X1.Y1.Z1` belongs to AS B. If A and B wish to exchange packets, then all routers along the path must have appropriate route entries pointing to the next router in the path, which in this example is A→L→M→N→B.[5]. ASs L and N could be regional nets and M could represent the wide area backbone connecting them.

Consider a change in state for network `X.Y.Z`. Using its selected intra-AS routing protocol, AS $\mathcal{A}$ detects the change and disseminates the information across AS boundaries until it has propagated to all ASs. Note that each protocol has finite hold-down times for state information changes, and change information is not propagated further unless protocol hold-down timers expire. This mechanism prevents short-lived changes from unnecessary propagation.

The NSFNET backbone uses the IS-IS intra-AS routing protocol to disseminate changes to all backbone routers, which then proceed to update their routing tables. As new information is incorporated into the backbone routing tables, it reaches all other attached regional networks via inter-AS protocols at the backbone boundaries. The change is propagated throughout the system of connected nets in this fashion.

[5] ASs A and L could use a default route pointing to the transit AS M, and assume that M has the appropriate entry for the destination network

6.3 Analysis

We first present a few definitions which facilitate our discussion.

A *connectivity transition* event is defined as a network event that causes a network number to either be added to the backbone routing tables when previously absent, or be deleted from the routing tables when previously present. That is, any event signalling a change in reachability of a network as seen by the NSFNET backbone is a connectivity transition. Note that a network may suffer more actual outages than the NSFNET backbone records. This is due to the fact that most routing protocols employ a *hold-down* period during which they do not propagate information about the state of a network. This is a means of damping the flow of routing information in an attempt to prevent oscillation through the set of connected networks. It is conceivable that networks could regain connectivity within the time span of routing protocol hold-down periods.

An *unreachability cycle* is defined as the interval of time between two successive connectivity transition events, the first being the deletion of a network and the next being the subsequent addition of the same network to the backbone routing tables. The length of the unreachability cycle represents the amount of time the backbone cannot route packets to that network.

A *cluster* of networks is a group of networks that undergo an unreachability cycle together. This represents the aggregation of networks that simultaneously suffer connec-

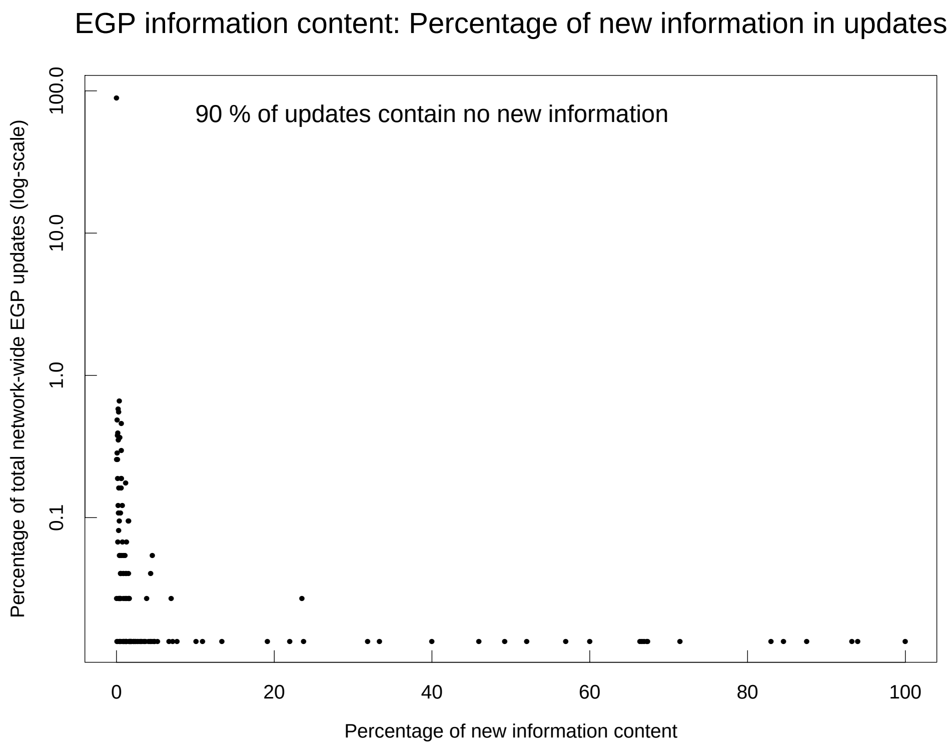

Figure 5: EGP Updates: Information content

tivity outages as seen by the NSFNET backbone. As an example, consider the case where a group of IP networks share a common path to the NSFNET backbone. If that path becomes unusable, then all these networks will appear to become unreachable simultaneously, and will appear in the backbone routing tables together when the shared path is usable again.

6.4 EGP information content

As we noted earlier, the NSFNET system of networks uses both EGP as well as BGP as inter-AS routing protocols. A significant characteristic of the use of EGP in large network environments is the large overhead associated with the update process of the protocol. EGP requires that all networks associated with the participating ASs be exchanged in every update, whether or not the state of the network changes. In other words, the absence of a network number in an EGP update signalled a unreachability event for that network.

Figure 5 charts the number of networks in an EGP update that have actually changed state, as a fraction of the total number of networks contained in the update. Note that EGP updates are synchronous with a periodicity of three minutes. The graph shows that almost 100% of the updates recorded contained close to 0% new information. Almost every EGP update conveyed absolutely no new information about the participating networks, implying a tremendous waste of network bandwidth, router CPU resources and memory.

BGP used this observation by requiring only *incremental* updates, whereby only those networks which undergo state transitions are exchanged in the routing updates.

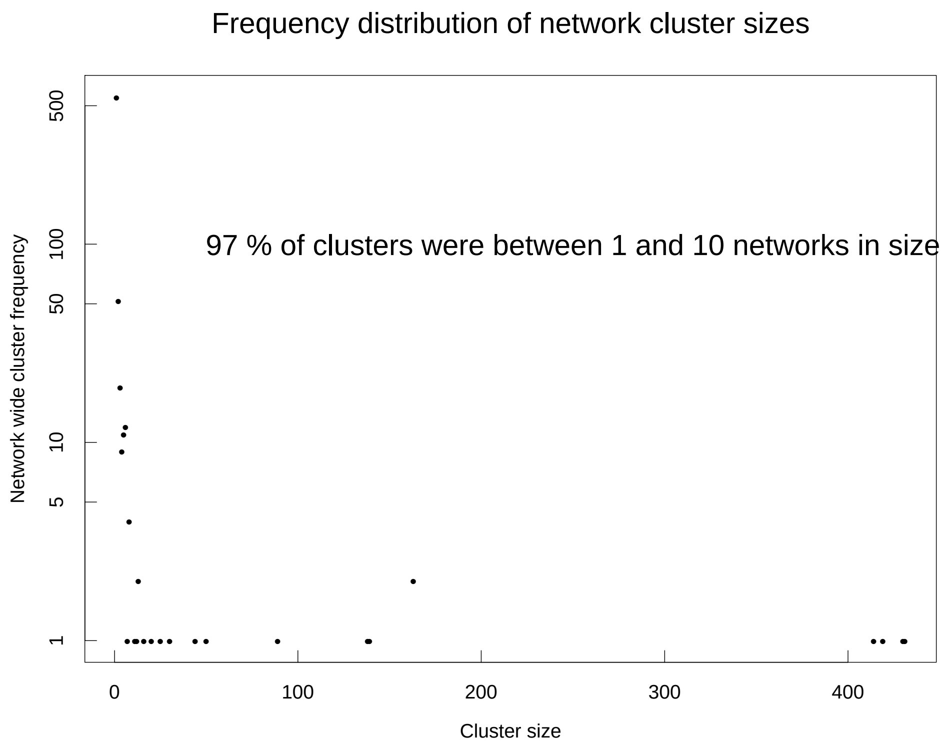

Figure 6: Cluster Size frequency distribution

6.5 Cluster size distribution

The cluster size observed in a routing fluctuation is a good indicator of the nature of the event that caused the outage. Recall that the Internet is loosely organized as a tri-level hierarchy on network connectivity. Mid-level networks that are themselves composed of smaller networks are usually directly connected to the backbone networks. Larger cluster sizes seen during a routing fluctuation would very likely indicate a problem closer to the backbone level, such as a distribution network breakdown, or an outage involving a border router serving multiple smaller ASs. Smaller cluster sizes indicate network outage situations further away from the wide area backbone.

Figure 6 shows the frequency distribution of routing fluctuations of a given cluster size. The cluster sizes seen in the trace varied from one network to a maximum of approximately 450 networks. Since the NSFNET backbone carried approximately 6500 networks at the time of the trace collection, this upper bound represents about 7% of the total set of networks. The graph clearly shows that the dominant fluctuation behavior involves less than ten networks in a cluster. Indeed, the frequency of fluctuations involving only one network was the greatest. The large majority of fluctuations, 97%, occurred in a cluster of less than ten networks. This seems to indicate that the Internet has good overall *system* stability and large scale oscillations are relatively uncommon.

6.6 Unreachability cycle distribution

It is also instructive to study the distribution of unreachability cycles seen in the sample. Figure 7 shows this distribution as a function of the cluster size. Smaller sized clusters have, for the most part, relatively shorter cycle times. The larger sized clusters, while less frequent, seem

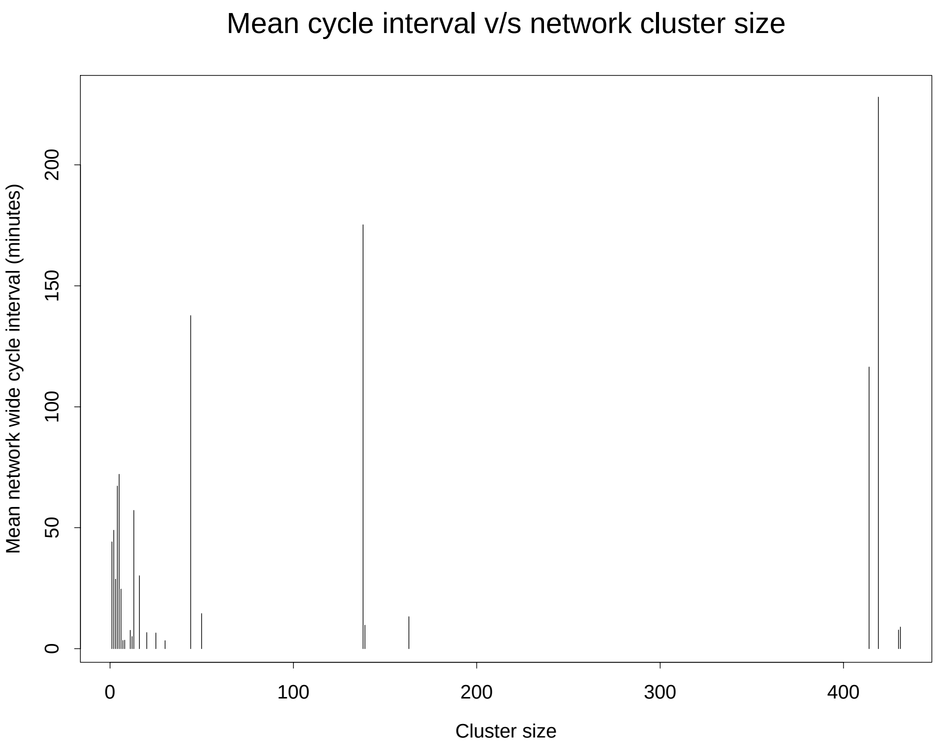

Figure 7: Cluster Size Cycle time distribution

to have a longer cycle time. That is, the more dominant behavior is where a large number of small sized clusters have a small cycle time. An additional factor that the study does not pursue is the proximity of the AS containing the cluster to the NSFNET backbone. Since the distance between an AS and the backbone influences the number of routing protocols an update must traverse, a larger AS may have a smaller cycle time if it is located closer to the backbone. Further studies should investigate this factor.

6.7 Future work

This work represents a first exploration into the analysis of system-wide routing information flow. We also deem important the sources of routing fluctuation, and would like to investigate them as well. We also wish to track the end-to-end latency associated with the flow of routing information across the Internet. This discussion has only presented latencies experienced across the T1 NSFNET, Internet-wide analysis is much more of a challenge. A paper entitled *Internet Routing Stability* is available in draft form from SDSC. Please contact bac@sdsc.edu for a copy.

7 Route caching

7.1 Introduction

Observations of Internet traffic have revealed a high degree of *concentration* in IP network destination addresses. That is, the distribution of destination addresses is highly non-uniform, with a large fraction of packets in a given sample destined to only a small fraction of the total routable destinations. This phenomenon has been ob-

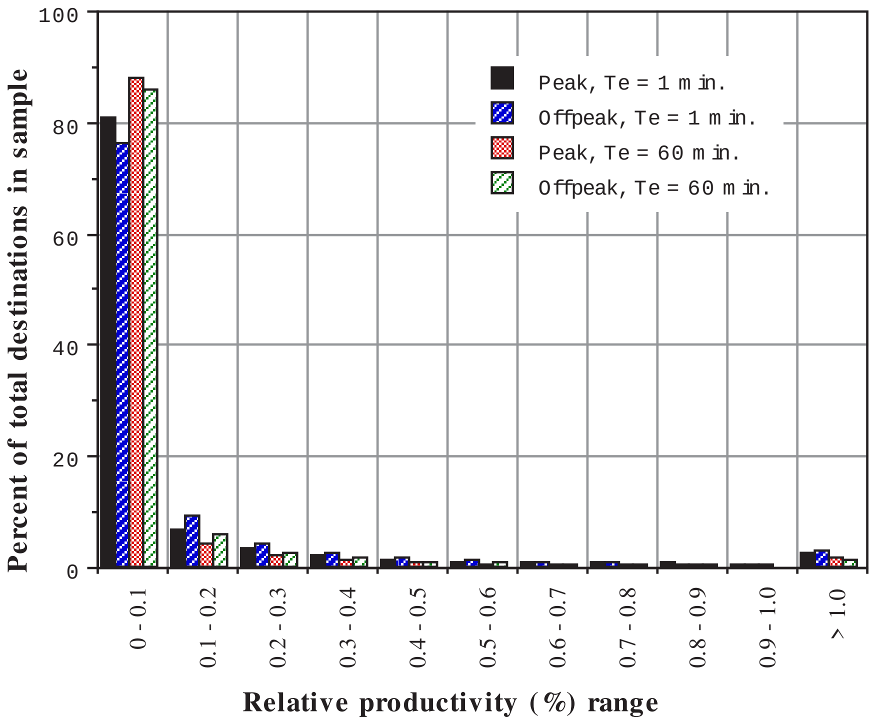

Figure 8: Histogram of route productivity

served both in a local area context as well as in a wide area environment. Traffic presented to the NSFNET backbone network also exhibits this concentration even though there are many thousands of possible destination networks.

Rekhter and Chinoy [15] undertook a study of NSFNET backbone traffic in order to quantify the concentration properties of aggregate traffic and to analyze the viability of a scheme designed to limit the amount of routing information required by Internet routers. The motivation for this work stemmed from the rapid growth in reachable networks in the Internet and corresponding growth in the size of routing tables. We present the salient observations and conclusions from the study here.

7.2 Observations

We collected 48-hour traces of traffic offered into the NSFNET backbone. We then analyzed the destination references in these packets for concentration patterns. We define the *route productivity* of a route to a destination network as the ratio of the number of packets that can be forwarded using this route to the total number of packets in the observation set, for a fixed time interval. By definition, the productivity of an individual route is bounded between 0 and 1. The results show a high degree of non-uniformity, as seen in Figure 8. This histogram plots the route productivity as a function of the percent of destinations reachable using this route. The graph clearly shows that a large number of routes had relatively small productivity while a very small number of routes exhibited large productivities.

We proposed a scheme by which a router could keep only a small fraction (20%) of the total destinations reachable in its *Forwarding Information Base* and yet service a large fraction (85%) of packets offered to it. A scheme for updating based on relative route productivity allows for replacement of routes that show decaying productivity with newer routes that have increased use. Packets for which no routes are available would be encapsulated and forwarded to a designated border gateway that keeps the entire set of information. This scheme would hence dramatically reduce memory consumption on routers as well as other resources used to maintain large forwarding databases.

The paper describing the details of the scheme and other key elements was published in the *ACM Computer Communications Review* of January 1991, under the title "Injecting Inter-AS Routes Into Intra-AS System Routing: A Performance Analysis" [15].

8 End to end reliability

The end-to-end reliability of Internet network paths is a function of a variety of complex factors. The heterogeneous nature of the Internet makes the formal definition and study of this issue difficult. It is important, nevertheless, to derive reliability models based on empirical observations and data.

To this end, we have begun a study to identify the parameters involved in determining the reliability of end-to-end Internet connections. Additionally, we have initial results showing statistical observations of some of these key parameters.

8.1 Motivation

The advent of professionally managed networks and the maturity of network management ideas and software has contributed to a dramatic increase in the reliability and stability of both local as well as wide area TCP/IP networks. The interconnection of these individual networks, however, still causes some concern, as users frequently complain of broken connections or erratic performance. While an individual network operator may report a high uptime for its network, a more important parameter is the uptime for individual connections or conversations.

We attempt to obtain an initial understanding of end-to-end reliability by observing the various causes of instability and gathering statistics about them in order to build realistic models.

8.2 Background definitions

We classify a number of network-related events as factors which may contribute to *connection unreliability*. It is important to distinguish between these events and other phenomena that cause performance degradation but have no impact on the *longevity* of an Internet connection. The most common connection-oriented Internet transport protocol is TCP. An established TCP connection between two Internet hosts may break for a number of reasons; we list the two most important causes.

- Network routing changes

  While changes in the path to a given destination during the lifetime of a TCP connection should not theoretically break the connection, in actual practice routing fluctuations seem to be the most common cause of broken connections.

- *Destination Host Unreachable* messages

  An existing connection will terminate when the destination host itself becomes unreachable.

This study considers these two events as potential connection terminators. In order to simulate an ongoing and active connections, the source host in this study continually emits 1 packet per second to the destination and gathers all network related events that fall into the above categories.

8.3 Experimental Environment

We chose seven sites distributed throughout the global Internet, at varying distances from our source host, as our initial target environment. We ran the experiment for 50 days, from 1 November 1992 to 20 December 1992. We list the destination hosts below.

Source: serendip.sdsc.edu (a Sun Sparc at SDSC) *Duration:* 50 days (11/1/92 to 12/20/92) *Source frequency:* 1 packet per second to each destination *Type of packet:* UDP (to port 2000 on each destination) *Error messages:* ICMP Net and Host unreachables *Destinations:*

- cs.ucl.ac.uk (128.16.5.31): London, England.
- eng.isas.ac.jp (133.74.2.60): Tokyo, Japan.
- nsf.gov (128.150.195.1): Washington, DC.
- cs.fsu.edu (128.186.121.10): Florida State University, Florida.
- nri.reston.va.us (132.151.1.1): CNRI, Reston, Virginia.
- koala.inria.fr (138.96.24.65): INRIA, France
- edv6000.tuwien.ac.at (128.130.36.72): Vienna, Austria

In our experiment, we transmit one packet per second to each destination host in order to simulate an active, constant conversation. These packets may generate error messages from any router along the Internet path. The source host records these error and control messages with a timestamp and lists the source of the control message. We are thus able to analyze the frequency, reason, and source of the error or control message.

As an illustrative example, consider the following message:

Mon Dec 7 10:14:29 1992 len=48 addr=int-gw.ulcc.ac.uk (128.86.1.2): Destination Host Unreachable 128.86.8.10

This message indicates that a packet destined to the host cs.ucl.ac.uk (128.86.8.10) could not be forwarded by the router int-gw.ulcc.ac.uk (128.86.1.2) and was dropped. Since this is a *host unreachable* ICMP message, it implies that this network path to the destination was in tact but the last router along the way could not deliver the datagram because the destination host itself was unreachable.

8.4 Results

We tabulate some of the more important results below. We list the destination along with the number of hops from the source host to it. We also list the average number of events that pertain to the stream of packets going to that destination, as well as the average duration of the events.

As an example, consider destination cs.ucl.ac.uk. The average numbers of network unreachable events and host unreachable events per day are 10.4 and 0.4, respectively. Their average durations were 14 and 38 seconds, respectively.

The typical pattern of network unreachables messages consists of short bursts of messages lasting about 10 to 20 seconds, followed by normal activity for a few seconds. This cycle repeats itself for many hours, a pattern that may indicate load-based oscillation of routing. Such oscillation may be the result of the periodic removal and insertion, based on load, of a routing entry for a destination network. It may also represent some protocol irregularity. More investigative work is necessary to more clearly pinpoint the causal factors.

Figure 9 shows the distribution of source routers issuing network unreachable messages for cs.ucl.ac.uk. We can see from the graph that a large fraction of events for this particular host come from a router that is three AS hops away (*nsn-FIX-pe.sura.net*). This type of distribution is useful for tracing where in the interconnection of ASs problems occur.

8.5 Future work

While much work remains in formalization end-to-end reliability, this study represents some initial work. We plan to formalize some definitions for end-to-end reliability based on these initial observations and re-measure these parameters. Our goal is to provide some quantitative information about the longevity of Internet connections and the consequent reliability of end-to-end paths.

9 Packet video

IBM recently announced its latest multimedia product, called the *Person-to-Person* multimedia system. This system allows users to share audio, color video, text and graphics across an interconnecting network.

SDSC and IBM are investigating the networking aspects of using Person-to-Person across the Internet. The initial set-up involves two IBM PS/2 model 95's running Person-to-Person, one at SDSC and the other in an IBM office in Milford, Ct. The current path traverses LANs at both SDSC and Milford as well as the NSFNET T3 wide area network.

During initial tests we sustained a frame rate of approximately 5 frames/sec across the Internet path between San Diego and IBM, Milford. Our next steps for this project include:

- Investigating the effect of packet sizes and window sizes on video transmission.

  Since Person-to-Person v1.0 uses NetBios over TCP, the underlying TCP window size greatly influences throughput. The retransmission of lost packets also hinders the smooth flow of video information to the user. We will characterize the influence of window and packet size on the frame rates achievable over an Internet path.

- Investigating the effect of packet loss.

  The amount of packet loss greatly affects the received video stream and we will attempt to characterize frame loss to packet loss relationships.

10 CASA Gigabit Testbed

SDSC is one of four sites participating in the CASA gigabit network testbed project, sponsored by CNRI and jointly funded by NSF and DARPA. We describe below some of SDSC's key activities in the context of the CASA project, particularly in the areas of network infrastructure and performance measurement.

10.1 HIPPI network simulations

Researchers in SDSC's ANR group have enhanced a HIPPI link level simulator written at Los Alamos National Labs, which allows the simulation of local and wide area HIPPI networks based on the Cross Bar Interface (CBI) architecture. We have simulated a few representative scenarios with this tool, and the results have been useful in sizing CBI parameters such as required memory,

Dest .	Number of Hops		Average Events/day		Average duration (secs)	
	IP	AS	Network	Host	Network	Host
cs.ucl.ac.uk	18	5	10.4	0.4	14	38
eng.isas.ac.jp	15	3	18.1	0.2	21	93
nsf.gov	16	2	20.8	0	12	0
cs.fsu.edu	17	2	15	0	23	0
nri.reston.va.us	15	2	5.2	0	21	0
koala.inria.fr	5	24	6.8	0.3	37	72
tuwien.ac.at	17	3	7.1	0	11	0

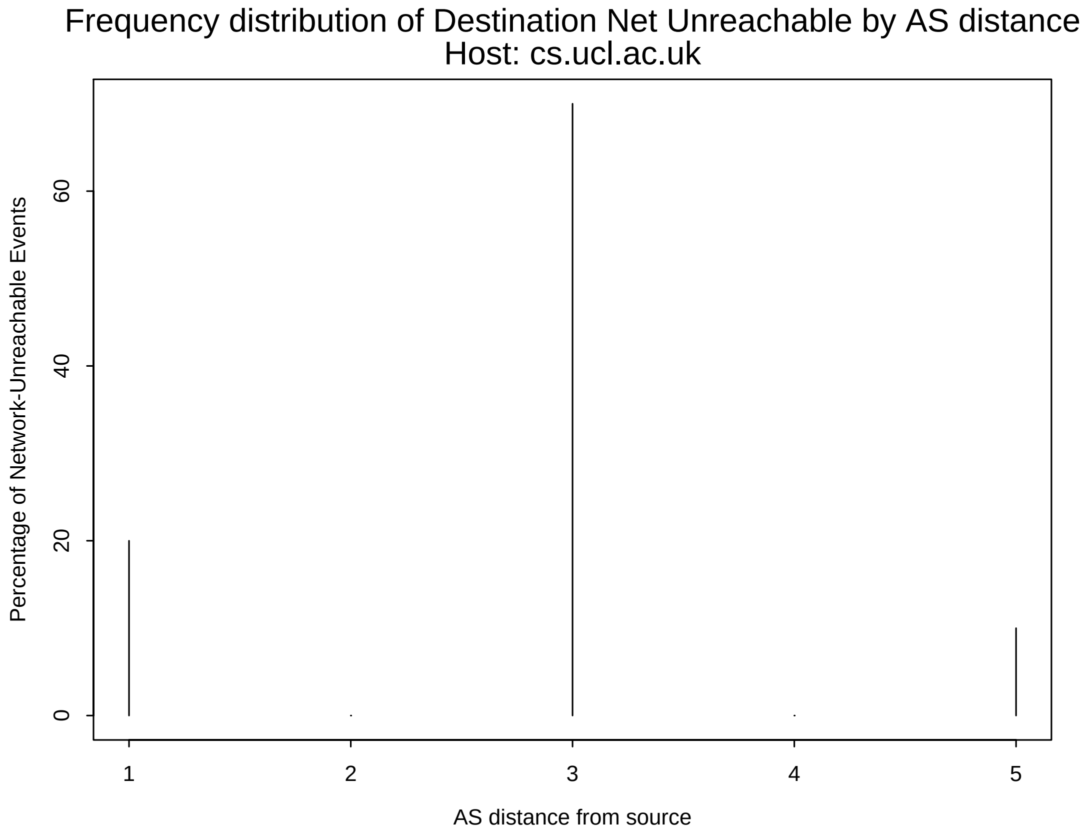

Figure 9: Unreachable Message Distribution by AS distance

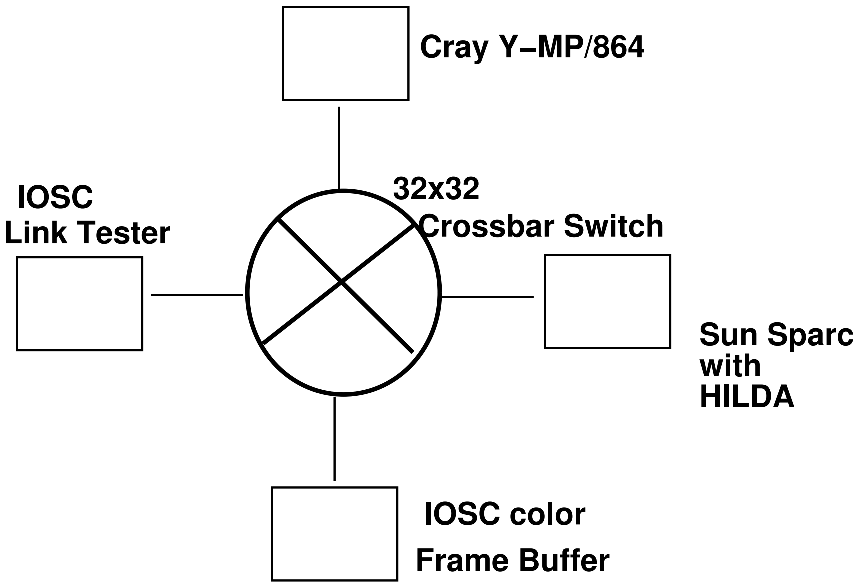

Figure 10: SDSC HIPPI LAN (12-31-92)

flow control interactions, and HIPPI contention. This tool also allows for the study of long haul lines interconnecting HIPPI LANs. The second CASA Annual report presents results of simulation scenarios, copies of which are available upon request.

10.2 SDSC HIPPI LAN

Researchers in SDSC's ANR group have recently begun to enhance the HIPPI local area network that will connect to the other CASA sites via long distance SONET circuits. At the present time we have the SDSC Cray Y/MP, a Sun SPARC-2 workstation, a HIPPI-attached frame buffer as well as a HIPPI link level tester on our LAN. A NSC PS32 HIPPI crossbar switch interconnects these machines, as shown in Figure 10.

10.2.1 Cray Y-MP/864

The Cray now supports TCP/IP over HIPPI, allowing TCP connections to other machines across the HIPPI network. Additionally, distributed applications can now set up RPC style communications across the HIPPI network by using a library of network routines written at SDSC.

We have written an initial Cray HIPPI test program, which opens the HIPPI device, reads and writes large arrays, and times the transmissions. Running loopback to the Cray, the speeds were up to about 350 Mbps, which were lower than we had expected due to the Cray's IOS (Input Output Subsystem) memory bandwidth limitations of the loopback mechanism. When in loopback mode, the IOP (Input Output Processor) transfers data between memory four times: Cray memory to IOP memory; IOP memory out the HIPPI channel; HIPPI channel into IOP memory; and IOP memory to Cray memory. Also, we expect a newer version of the IOS to have much faster unidirectional communication.

10.2.2 HILDA Network Analyzer

The Sun SPARCstation runs the HILDA (HIPPI Link Data Analyzer) system for HIPPI network analysis developed at the Microelectronics Center of North Carolina (MCNC), for the Vista gigabit testbed. The purpose of this system is to enable the capture of HIPPI frames from a LAN. This system consists of a set of HIPPI interface boards for the Sun host, and software allowing the host to capture frames that traverse the Sun. We are currently using this workstation in two ways.

The presence of the Sun provides a second machine on our HIPPI LAN capable of running TCP/IP applications, which allows us to exercise the Cray HIPPI software. This capability will be critical when the GCM application runs on the Cray and communicates across the HIPPI channel.

Additionally, a key performance parameter determining GCM speed-up in the CASA network will be the timing relationships between computation and communication. Capturing HIPPI frames using HILDA allows us to analyze the delay and variance in delay associated with the HIPPI channel. HILDA is thus facilitating the study of application performance evaluation.

10.2.3 HIPPI frame buffer

An IOSC 1280x1024 color frame buffer is part of the SDSC HIPPI LAN. This frame buffer has dual buffers and 24 bit/pixel color resolution. SDSC has written a utility called "*imhippishow*" which allows frames sent from the Cray to be displayed on the frame buffer. Variable delays between frames, as well as support for many different input image formats, allows visualization researchers to use this as a high speed rendering tool.

Initial measurements indicate throughput of about 740 Mbps for image transfers from the Cray to the Frame Buffer.

10.3 Current Status and Future Plan

The Cray and the Sun workstation are able to set up HIPPI connections through the PS32 HIPPI switch. TCP performance of a standard 4 KB window size and a maximum of 16384 bytes per packet is approximately 2 Mbps. This initial number shows that many parameters of TCP, such as window size, are not optimized for the higher bandwidths which the HIPPI network affords. We will follow this preliminary test with a comprehensive test suite which will calibrate performance based on packet sizes, window sizes and interpacket gaps at the source of traffic.

CASA grand challenge applications require strict synchronization between modules running on disparate compute platforms. The gigabit network linking these machines has to provide a well-defined set of services. In the coming months, we will characterize the performance of the CASA network in light of these requirements.

11 Statistics of importance to the Internet environment

The current operational Internet infrastructure is generally strongly focused on the real time operational and near term engineering requirements to keep the fabric alive, while ensuring short-term evolution. As a result, the community may acknowledge the importance of gathering statistics for analysis, but generally this collection must take second priority to the immediate operational requirements of the network service providers. Many times the statistics are used more for public relation purposes, such as to demonstrate the ever-increasing trend for higher bandwidth demands, or for simple periodic reports, rather than for data sets which are key to engineering, planning, and network research.

Generally, the level of insight into typical Internet networking environments is at a macroscopic level, where longer term trends are visible, but relatively little insight exists into what is really affecting the network at multiple levels of granularities. This lack makes preparation for unforeseen events, such as an event causing sudden consumption of vast networking resources, close to impossible.

11.1 Traffic characterization metrics

One of our goals is to answer the question:

> How can network analysis facilitate planning and network evolution, both short and long term? What kind of metrics can indicate where, when, and how to modify the network, in terms of connectivity, bandwidth, and other facilities?

Many times a network is characterized in only one way: "It's flaky." One rarely hears comments when the network is working well, since network services are typically invisible to end-users. This trend toward transparency will continue with new distributed applications which feign service locality rather than explicitly notify of network usage. The telephony web, which we now take for granted, has shown us how difficult it is to achieve their model: the best network is an invisible one.

In reality, end-users on data networks perceive network outages of various durations and levels of service degradations. As expectations of data services far exceed those typical of voice services, the ability to characterize network service and its degradation, becomes important. Prerequisite to this ability is a more general model of network performance.

11.2 Performance Metrics

Another question to which we seek an answer is:

> How can one define performance in large high speed wide area networks? What kind of metrics can characterize end-user transmission requirements and what kind of measurements are required?

In our traffic characterization project for the T1 backbone, described in Section 2, we learned that the set of statistics which the current NSFNET service providers currently collect for the NSFNET backbone on an operational basis allows for only a very limited and inadequate assessment of network "performance". The reason for this is a general lack of definitions for acceptable performance criteria, and subsequent measurements upon which to base judgement.

As an example, current network-performance metrics address primarily switching node throughput and only minimally address the importance of application service requirements. The network community, as well as developers of advanced applications with stringent network service requirements, need to address this issue. Sophisticated applications, in particular, those with real-time requirements, may require access to lower layer information about service qualities, such as available bandwidth and latency, to modulate their own behavior or to advise a user of the network situation. Such data, perhaps correlated with error or loss rates, could enable the network users, or even the application itself, to tailor their traffic demands to dynamically changing network capabilities.

The increasing variety of network clientele and service providers makes it necessary to parameterize network service quality from multiple viewpoints. For example, a backbone-centric environment perceives service qualities differently from an end-user. It may even be necessary to differentiate between major national backbones and service providers of a more limited geographic scope.

As an initial step, we have investigated network demand by various aggregations (e.g., by individual IP network numbers; Administrative Domains; and NSFNET nodal access points). Operational collection of statistics at granularities as fine as individual end systems or applications currently exceed the technical implementability; such a matrix would be huge. We have, however, performed snapshot investigations using short term packet traces at major networking aggregation points, such as a FIX multi-agency interconnection point and an E-NSS site providing NSFNET backbone service.

Our research thus far suggests that object definitions for individual statistics are only the first step of a more comprehensive model of network performance. Currently, many different existing definitions of network performance contribute to an incoherent picture of the quality of specific network service. We have taken steps by compounding sets of performance indicators that we expect to help us derive a better model, such as, delay, loss rates, and bandwidth.

We have experienced cases where it was necessary to define revised or additional data sets for further evaluation and optimization of the network. As one example, ANS has adopted our suggestion to routinely collect a certain SNMP object that allows us to derive the dynamic network topology during collection. New statistical objects will go beyond the standard byte volumes, throughput, and delay and will lead to a more effective, if albeit more complex, definition of network performance. An acceptable definition will depend on who is applying it: e.g., end-user versus service provider, but in any case the definition must compound multiple effects seen on the network.

We view a significant part of our effort as the definition of requirements for network analysis, many times with specific emphasis on traffic characterization, which often requires additional information and tagging. We will continue to refine the definitions of attributes for information tagging, which may be also useful for efforts outside of the direct scope of network analysis and engineering. For example, network information services may want to utilize our workload characterization results (see Section 11.4) to prepare for end-user consumptions.

11.3 Granularity of statistical objects

Presenting statistical data requires one to consider multiple granularities in various dimensions. Time series analysis offers obvious examples. A time series of a link utilization graph looks very different when displayed at daily, hourly, minute, second, etc., granularities. Traffic can also be aggregated at a variety of levels of geographic or administrative scope: by hosts (aggregating users and processes); IP network number (aggregating hosts); Primary Administrative Domain (aggregating network numbers); Domain external interfaces (roughly aggregating domains); or an aggregation of a whole backbone network itself (aggregating the traffic of all the external interfaces).

While long term time series expose day/night or weekday/weekend fluctuations, the shorter term granularities expose the effects of burstiness behavior. One extreme limits the figures to averages over long time periods; the other extreme elucidates full utilization of the bandwidth at each instant of packet transmission. Some compromise in granularity is necessary, such as granularities close to packet transmission rates, which will expose fluctuations of burstiness.

11.4 Long-term change in the cross-section of applications

A further dimension in traffic characterization is differentiation between the network layer and the applications. Characterization of applications, in the case of TCP/IP as indicated by IP port numbers, allows one to quantify several different uses of the network, including trends in absolute and relative volumes over time. Newer applications using the *uncoordinated* port number space, rather than utilizing officially assigned *Well Known Port* numbers, complicate this task. The currently utilized port categorization paradigm on the NSFNET turns out to be rather insufficient to address today's environments. The "other ports" fraction offers evidence of the problem; it has now become the single largest component of the monthly traffic in the reports. We have experimented with port data traces on FIX-West and found significant components of explainable port traffic, most notably Andrew File System and game playing (MUD). The NSFNET should perhaps consider reinstrumenting the data collection mechanism to categorize these ports separately.

As the NSF supercomputing centers strive toward interconnection via a robust, high performing network, they are interested in deriving workload profiles of different network constituencies, based on available or obtainable data. A "supercomputer center workload" profile will incorporate the impact of specific applications that the supercomputing centers will introduce. A "campus profile", e.g., of a major university, likely differs from that of a supercomputing center. For example, if our analysis of SDSC and UCSD data are representative, campuses typically talk to many other IP networks, while supercomputing centers send the bulk of their traffic to a more limited clientele, in some cases with strong favoritism toward specific remote IP networks. Understanding these workload profiles may allow for better bandwidth management and general network planning in the future.

We have investigated the longer term changes in the cross-section of applications, on both relative and absolute terms. We expect to see the impact of the increased bandwidth, as well as robustness, of the T3 backbone. Instrumentation with a stronger focus on capturing networking requirements for such applications, e.g., by means of better classification of TCP network service ports, would allow tracking of new applications that are emerging in today's high speed environments. Only more carefully attended collection will allow answers to questions of how new paradigms of data transfer (like image, graphics, audio, video, other digital continuous media) will affect higher speed realms.

We have only begun investigations in this area, but we consider it important to create network instrumentation that will allow us to track usage trends over time. This includes backing away from the notion that Telnet, FTP and SMTP are the most critical network applications. Indeed, these will become "low end" services, as services such as X-Windows, AFS, NFS, NNTP, WAIS, other national file systems, and real-time continuous data streams characterize the service profiles of the future which will require committed support. Modern applications, such as data base transactions, render the network transparent to the user, who may (hopefully) perceive only a "local" service. The imminent massive introduction of multimedia applications will amplify this effect; it will not always be obvious to the user whether some massive data quantity came from a local disk, a file server, real-time across the country, or even an international location.

As an example of the application of specific service investigations, during the March 1992 IETF meeting we presented preliminary analysis results on the network impact of NNTP traffic to an NNTP developer, who immediately started to identify possible improvements of the NNTP application to ease the impact of news exchanges on both hosts and the network.

As networks become increasingly integrated, it will become more important to account for the extent of bandwidth sharing in the environment of traffic aggregated from many users. The National Research and Education Network (NREN) environment where agencies "use each other's bandwidth" requires a mechanism to measure the extent and fairness of interagency traffic exchanges. Information regarding international distribution of traffic by country across US national backbones is also of importance, if only to address certain policy concerns related to fair cost sharing for available bandwidth. Instrumenting strategic network locations to derive country-to-country traffic matrices can allow one to develop resource requirements which address some of these concerns.

12 Performance Evaluation and Modeling

As stated in our proposal [6] and Section 2, during the first year of the project we have concentrated on the investigation of existing data. We have familiarized ourselves with the details of data collection on the T1 and T3 backbone networks, as well as multi-agency interconnections, and have now reached a level of instrumentation and expertise necessary for success during the remaining two years of the project. At this point we communicate with many constituents of the Internet environment, and several Internet service providers have consulted us on data collection matters. Our focus has been on modeling the traffic, i.e., workload characterization, the first step in any system model study.

While we obtained many interesting characterizations of NSFNET traffic, it is clear that the currently collected statistics, with their focus on operational infrastructure, can not help to answer most performance evaluation questions or to produce fine grain traffic models. This is mainly due to their coarse granularity which masks most interesting phenomena. The finest grain statistics available on the NSFNET/ANSnet are averaged over 15 minute intervals, which cannot, for example, capture periods of intense activity that might lead to performance degradation.

To address these shortcomings, we have compiled and are pursuing the following prioritized list of tasks and subprojects:

1. Characterize traffic profiles and performance with direct measurements at the packet level on a limited geographic scale. This will involve a study of the NSFNET traffic through the FDDI interface of the San Diego ENSS. Our goal is to obtain statistical traffic models at various granularities, e.g., specify packet inter-arrival time distribution, autocorrelation, correlation with packet size, source, etc. We hope to augment our study with additional information from other sites, and hope to derive some indication of specific profiles for, e.g., supercomputing center, campus and government laboratory traffic.

2. Study the correlation of coarse granularity statistics with fine granularity measurements on a local scale. In particular, we will investigate whether any fine-grain characteristics are visible in the coarser description. Conversely, we will examine whether interesting points in the coarse characterization are identifiable in finer descriptions obtained by averaging the fine-grain measurements over various periods, from milliseconds, to seconds, and minutes. This approach is similar to the approach we are using to investigate sampling methodologies (See Section 3).

3. Specifically repeat the local characterization at a different NSFNET backbone site in order to investigate the level of generality of the traffic profiles obtained at the San Diego node.

4. Use the models obtained through the local characterization with parameters obtained from the coarser but global measurements in order to obtain a global traffic model. This model can then be used in performance studies.

The NSFNET does not currently monitor some specific statistics interesting to networking research, such as queue length distributions, or at least statistics about packets that are dropped because they find the output buffers full. In addition, all interface error conditions are combined into one counter. Such refined or additional statistics would be very useful for performance studies.

Furthermore, in addition to their potential for direct use, there are many indirect uses of such information. For example, one can estimate mean delay with queue length and throughput statistics using Little's law [10], or more importantly, check the validity of these measurements when obtained through different methods. Therefore, one of our main current tasks is the definition of data collection requirements in order to answer such questions.

13 NREN Engineering Group Project for NSF

NSF funded SDSC for FY92 to collaborate with NSF's NREN Engineering Group (NEG) efforts.

The NEG effort involved collaborative work among NSF (particularly Steve Wolff and Bob Aiken), LANL (Peter Ford) and SDSC (Hans-Werner Braun and some additional support from Kimberly C. Claffy). In FY93, NSF is supporting continued work in this area.

13.1 NREN Engineering Group Activities

The primary objective of NSF's NREN Engineering Group is to facilitate a graceful transition to the Interim Interagency National Research and Educational Network (NREN). This past year, the scope of work included (as outlined in last year's proposal submitted to NSF):

1. Identifying requirements for the Interim Interagency NREN;

2. Investigating high level architectural issues;

3. Identifying ancillary network services;

4. Participating in the evolution of an implementation plan;

5. Collaborating on design criteria for international connections to the Interim NREN.

6. Undertaking network engineering studies for NSF's Division of Networking and Communications Research and Infrastructure (DNCRI);

13.2 NSF IINREN Implementation Plan

Braun and Claffy contributed to the writing of NSF's Implementation Plan for the Intermediate Interagency National Research and Education Network (IINREN). They met with the NEG members, particularly Bob Aiken and Peter Ford, to discuss this plan about once a month from Sept 1991 to March 1992 in preparation of a formal document (GA TR A21174 [1]).

13.3 NSF Resolicitation

Braun and Claffy participated in several discussions and meetings relative to the anticipated NSFNET resolicitation.

13.4 Network Access Point (NAP) Architecture

Braun and Claffy contributed to an initial conceptualization of Network Access Points (NAP), and in June submitted to NSF a summary of issues surrounding prototype implementations.[6] This summary includes investigations into architectures for interconnections among multi-provider networks, and the role of route servers in managing extensive traffic flows.

13.5 General Data Collection Requirements

We are currently working with NSF, Merit and ANS to determine how the T3 NSFNET component of the ANSnet backbone network may alter the requirements for earlier NSFNET statistics collection techniques and requirements. We furthermore hope to design a multi-agency effort to support the aggregation of network statistics data from multiple service providers. Such an effort is targeted toward reducing the disparities among (1) what NSF needs; (2) what statistics service providers can provide; and (3) what network analysis and in particular traffic characterization can offer to network operability.

13.6 Regulatory Investigations

Claffy has undertaken initial investigative efforts on behalf of the NEG to research regulatory issues of the telecommunications industry. The NEG must prepare itself with sufficient information should the self-regulatory nature of the current environment create conflicts and obstacles to efficient commercialization.

In May 1992, we established contact with officials at the National Telecommunication and Information Administration. Claffy scheduled a visit to Charla Rath at NTIA and other legal specialists in the agency (22 May 1992). Rath was then moving to the FCC, and Claffy is planning to communicate with her again in several months to update her and our information base on issues of relevance to the Internet and NSF. The current trend within NTIA is to decrease the current regulation restrictions within the telecommunications industry, including repudiating the Modified Final Judgement (MFJ) restricting RBOC activities. For example, a March 1989 NTIA study entitled "The Bell Company Manufacturing Restriction and the Provision of Information Services" notes that:

> Based on our assessment of the available information,...we believe that lifting the manufacturing restriction would likely result in both an increase in R&D and, in turn, the availability of information services.[18]

We would also like to learn more about FCC and other regulatory agencies and committees which have political influence regarding these issues.

Claffy has also done background research and reading, including [14], [17], [18], [9], [11], [16], to gain a better background in regulation issues.

13.7 Network Accounting Issues

We have started to think about accounting issues surrounding infrastructural networking activities. This is of importance to us both in the context of the NEG activities, as well as our NSF project on network analysis and traffic characterization.

In particular, we are exploring the accuracy of various sampling techniques for the characterization of aggregated traffic on a component network. Motivated by the increasing difficulty of fully monitoring high speed wide area network environments, we are concerned with how accurate a picture various sampling methods can offer.

Recent contact with service providers and researchers in the telco community have revealed serious concerns on their part regarding the integrity of sampling as a tool for use in accounting purposes. We hope our investigation into sampling can lead to potential approaches for dealing with accounting in high speed datagram environments.

13.8 Multi-Protocol NREN

A further objective of the NEG is to investigate a multi-protocol Internet, specifically where OSI/CLNP coincides with IP traffic.

13.9 Related Activities

Among other committee activities, Braun participated in activities of the Engineering and Operation Working Group (EOWG) of the Federal Networking Council (FNC), the Federal Engineering Planning Group (FEPG, which supports the EOWG activities), the Intercontinental Engineering Planning Group (IEPG), and the Internet Activities Board (IAB, now called the Internet Architecture Board). Claffy gave presentations about some of our network analysis and traffic characterization to NREN constituent parties at NIST and DISA.

[6] Appendix A provides a copy of this summary.

14 Activities in Support of Project

We presented initial analysis work at a March 1992 IETF (Internet Engineering Task Force) plenary, as well as a Birds-of-a-Feather session on the traffic characterization subject. The primary focus was on community efforts toward common standards of data collection and analysis methodology.

In addition, we have maintained contact with other researchers in the area of network analysis. During May 1992 Claffy and Braun visited Peter Danzig and his group at USC. In June Braun and Claffy visited networking researchers Lawrence Livermore National Laboratory, in support of traffic statistics requirements for the recently awarded NSF grant to SDSC for connecting the planned DOE/NASA national ATM-based network with the NSFNET backbone infrastructure. In May 1992, Claffy visited officials in the National Telecommunication and Information Administration to discuss Internet regulation issues. In July 1992 we visited Vern Paxson and Van Jacobson at the Lawrence Berkeley Laboratories, who are working in areas closely related to ours. On May 21, 1992, Claffy visited an IBM research site in Milford, CT, in order to present preliminary network analysis activities. In December 1992, Claffy presented network analysis and traffic characterization results to NREN constituent parties at NIST and DISA.

15 Where to Go From Here

Wide area traffic analysis offers a great opportunity to help manage rational network evolution. Current high performance services and the simultaneous strive towards ubiquitous interconnectivity make wide area traffic analysis critical to maintaining the stability of the global data communications infrastructure. The sheer number of input and output feeds and internal parameters make a large scale network more comparable to a living organism than to a simple predictable mechanism, and an evaluation of the actual network parameters becomes complex issue. New technologies have brought new constraints, which may require new paradigms in network analysis, modeling, and traffic characterization. In view of the new constraints, we pose and seek to explore, the following questions:

- What level of detail is required to sufficiently describe and model a high speed wide area network?
- What kind of network behavior can we describe within given time constraints that make such description useful?
- As the complexity of the network increases, how much must the complexity of the model increase to still offer a valid description? What are the tradeoffs between complexity and descriptive power of a model?
- What questions, which might have been relevant in smaller and slower speed realms, are no longer applicable in a network environment that attempts to support ubiquitous services at bandwidths of up to gigabits per second?
- What kind of (measurable) metrics can indicate where, when, and how to modify the network (vis a vis links, routers, etc.) and impact network planning efforts?
- How do certain network characteristics, like bandwidth, utilization, and latencies, impact network performance? What is the impact of asymmetric latencies along end-to-end network paths?
- What kinds of statistical data should high speed (T3 and beyond) networks collect that slower (T1 and below) networks did not need? How will gigabit networks differ? Does the higher speed affect the *types* of statistics that need to be collected?

In the past, the high level objective of capacity planning has been the main driving force behind statistics collection and wide area network analysis efforts. In the future, however, the complexities of the aggregation and diversity of traffic will require broader evaluation of network behavior. Only greater insight into the detail of network traffic characteristics will ensure successful and graceful evolution of such large scale environments.

Appendix A: NAP Pilot Project Summary[7]

The Network Access Point (NAP) Pilot Project (NPP) is a necessary step to investigate the design of Network Access Points for infrastructural networking. This project will require specific equipment, agreements, and collaboration of outside parties. The NPP will augment a NSF investigation into the future of Internet routing and addressing, while specifically examining results of the IAB/NSF ROAD (ROuting and ADdressing) activities for their operational applicability. The project will provide insight into the stability of the Internet's routing system and detailed network performance measurement requirements. The NPP will also enhance the network infrastructure through the alignment of DNS, X.500 and time servers at strategic NAP locations Within the context of routing management, NPP participants will work with other interested parties, such as those of the EBONE project and the BGP deployment working group in the IETF, in the specification of an intermediate data base format for routing policy requirements. This data base, in conjunction with local network policy requirements, could be translated into configuration files for individual routers. Further investigations could include considerations of multicast routing, particularly interdomain routing protocols on a shared Level-2 media, as well as specific experiments in the area of policy routing.

Following initial work on the prototype NAP, an initial operational site would ideally be at the east coast, perhaps in the Washington area where numerous Internet networks maintain a presence. This could parallel the current SURAnet FIX-East installation. The Intercontinental Engineering Planning Group (IEPG), with its requirement to investigate the need for global interconnection points has also expressed strong interest in such an east coast location. The IEPG would need an operational east coast NAP by approximately November, 1992. In selecting subsequent NAP sites, the community may naturally judge a west coast site, proximate to the CIX POP, as a promising opportunity to collaborate with the existing commercial participants. Early phases of the east coast NAP would be an appropriate time for CIX constituents to begin operational considerations for their future NAP project.

NPP goals

The NAP Pilot Project is intended to implement within a relatively short time frame enough of the requirements of a production NAP to gain experience with its components and concepts. For example, it will be important to test the new route server concepts at the pilot NAP on its existing LAN, thus avoiding the establishment of new dedicated facilities. Goals of the pilot project include:

- Involve participants from federal, commercial, and academic networks
- Examine and gain experience with new route server architecture
- Monitor, analyze and collect statistics on behavior and performance of the NAP on local networks

This pilot project is in support of NSF's NREN Engineering activities, and should result in a public report prior to the NSF Routing Arbiter and Network Access Point solicitation. This project should in no way prevent other networking constituents to additionally or collaboratively experiment with own NAP prototypes; and we in fact encourage those people to collaborate with us and exchange findings.

NPP phases

1. Preliminary investigation of routing information collection from multiple sources. SDSC, with its T1/T3 NSFNET interconnection, routing information availability from CERFnet/CIX, and other local routers, is a promising location for a NAP alpha/prototype. Eventually this may incorporate a connection to the DOE/NASA SMDS pilot project efforts.

   The first steps will include passive route listening to participating routers, and subsequent investigation of the "quality" of the collected information for use in a route server (RS). Clients of a route server may require different sets of information. We will need to discuss testing the RS serving packet forwarding engines. For example, at SDSC one could potentially peer between the RS and a CERFnet router.

   The initial phase should utilize a currently unused Ethernet controller of the T3 E-NSS at SDSC to initiate the NAP prototype. Internet numbers allocated for this are:

   - Autonomous System Number: 1909 (ALPHA-NAP-AS)
   - Class C Network Number: 192.172.226 (ALPHA-NAP)

   The NAP Ethernet can be connected to a spare RT out of the T1 NSS in order to act as a Route Server and peer with the T3 E-NSS. We will discuss with CERFnet the possibility of routing information injection of CERFnet routes into the RS as well.

[7] Working Draft Version of: 22 June 1992

A NAP will provide a mechanism by which to test many of those features of BGP which have never really been tested in a fully operational environment. For example, we need to evaluate the impact of including a NAP's AS number into a BGP AS path. We also need to determine the possible need for allowing immediate routing neighbors to request the suppression of the NAP AS insertion, perhaps even by bilateral agreement, if they choose so.

2. Implementation of a remote routing information listener. Although the initial site will be at SDSC, one could replicate them at LANL or elsewhere. A possible candidate for the routing protocol between the servers is I-BGP. Such a listener will allow an investigation of the dynamics of routing changes.

3. Preparation of an interim status report on NAP requirements and technical issues. We will provide this reports, as well as later ones, to NSF as input to their decisions in the RA/NAP solicitation.

4. Investigation of POP co-location, if feasible at this point of time.

5. Develop a plan for an operational east coast NAP, perhaps in the Washington area. SURAnet is a likely collaboration candidate, perhaps paralleling FIX-East, or if already feasible, co-located with phone company equipment to avoid the cost of extra tail circuits and to improve stability. Examples of other potential partners could include Sprint ICM, Alternet and MCI.

6. Implementation of a route server for the east-coast NAP. The implementation should be capable not only of gathering routing information from participating clients, but also of analysis and sanitization of the information, based upon the initial findings at the prototype NAP.

7. Statistics collectors for network performance data should be implemented each NAP to allow for gathering data to analyze routing and performance patterns. This will initially aid the design, planning, management and operation of NAPs. NSF and other constituents may also want to consider building desensitized generic trace data sets, with sensitive information removed, for research use.

8. NAP-East begins to provide services, including as a router server, to clients.

9. Preparation of a more comprehensive NAP status report.

10. Considerations for a collaborative west coast NAP, perhaps with the CIX constituents, and a full-scale investigation of national NAP services.

11. Replication of the NAP concept at other strategic points, including locations of high bandwidth interconnections at phone company levels, or packet-switched locations with rich connectivity.

Further activities to be considered

- general routing stability investigations
- X.500 at strategic access points
- DNS root servers at strategic access points
- Network time servers at strategic access points (HPNT project)
- Data base formats for interchangeable intermediate formats for routing policies
- Global network interconnects

Needed management equipment of operational services

It is proposed that "production" NAPs will utilize high performance workstations for the Route Servers and other equipment. These machines must support widely available tools such as gated, tcpdump, SNMP, and others.

For example, the east coast NAP will require at a minimum two machines to provide RS services and network performance data collection.

Timing considerations

- Project start
- Interim status report
- East coast considerations
- Draft document for NSF
- Meeting at Los Alamos
- Finalized document
- Solicitation release

References

[1] R. Aiken, H.-W. Braun, and P. Ford. NSF implementation plan for interagency interim NREN. Technical report, NEG, 1992. SDSC Technical Report GA-A21174.

[2] Rick Boivie. ANSnet router traffic statistics (arts), 1993. personal communication.

[3] R.T. Braden and A. DeSchon. Nnstat: Internet statistics collection package, Introduction and User Guide. Technical Report RR-88-206, ISI, USC, 1988.

[4] B. Chinoy and H.-W. Braun. The National Science Foundation Network. Technical Report GA-A21029, SDSC, 1992.

[5] K. Claffy, H.-W. Braun, B. Chinoy, and G. Polyzos. Analysis and modeling of high speed networks: Project status report. Technical Report SDSC GA-A20916, SDSC and UCSD, 1992.

[6] K. Claffy, H.-W. Braun, and G. Polyzos. Analysis and modeling tools for high speed networks: Proposal to the national science foundation. Technical report, SDSC and UCSD, 1991.

[7] K. Claffy, H.-W. Braun, and G. Polyzos. Traffic characteristics of the T1 NSFNET backbone. In *Proc. IEEE INFOCOM '93*, 1993.

[8] K. Claffy, H.-W. Braun, and G. Polyzos. Measurement considerations for assessing unidirectional latencies. *Internetworking: Research and Experience*, to appear in 1993.

[9] W. W. Hogan. Energy and information network infrastructure. In Brian Kahin, editor, *Building Information Infrastructure*, 1992.

[10] R. Jain. *The Art of Computer Systems Performance Analysis*. John Wiley & Sons, 1991.

[11] B. Kahin, editor. *Building Information Infrastructure*, 1989.

[12] K. McCloghrie and ed. M. T. Rose. Management Information Base for network management of TCP/IP-based internets, mib-ii. Internet Request for Comments Series RFC 1213, 1991.

[13] K. McCloghrie and M. T. Rose. Management Information Base for network management of TCP/IP-based internets. Internet Request for Comments Series RFC 1156, 1990.

[14] Office of Technology Assessment, editor. *Critical Connections: Communication for the Future*, 1990.

[15] Y. Rekhter and B. Chinoy. Injecting inter-autonomous system routes into intro-autonomous system routing. a performance analysis. *Computer Communcations Review*, January, 1992.

[16] Special Business Week Report. The baby bells' painful adolescense (the quiet life has ended). *Business Week*, 1992(3286):124–134., October 5 1992.

[17] National Telecommunications and Information Administration. Ntia brochure.

[18] National Telecommunications and Information Administration. The bell company manufacturing restriction and the provision of information services, March 1989.