SLAC-PUB-8385
February 2000

Internet End-to-end Performance Monitoring for the High Energy Nuclear and Particle Physics Community.

Warren Matthews, Les Cottrell

Presented at Passive and Active Measurement Workshop
(PAM2000), 4/3/2000—4/4/2000, Hamilton, New Zealand

Stanford Linear Accelerator Center, Stanford University, Stanford, CA 94309

Work supported by Department of Energy contract DE–AC03–76SF00515.

Internet End-to-end Performance Monitoring for the High Energy Nuclear and Particle Physics Community.

Warren Matthews, Les Cottrell

Abstract— **Modern High Energy Nuclear and Particle Physics (HENP) experiments at Laboratories around the world present a significant challenge to wide area networks. Petabytes (10^{15}) or exabytes (10^{18}) of data will be generated during the lifetime of the experiment. Much of this data will be distributed via the Internet to the experiment's collaborators at Universities and Institutes throughout the world for analysis.**

In order to assess the feasability of the computing goals of these and future experiments, the HENP networking community is actively monitoring performance across a large part of the Internet used by its collaborators.

Since 1995, the pingER project has been collecting data on ping packet loss and round trip times. In January 2000, there are 28 monitoring sites in 15 countries gathering data on over 2000 end-to-end pairs. HENP labs such as SLAC, Fermi Lab and CERN are using Advanced Network's Surveyor project and monitoring performance from one-way delay of UDP packets. More recently several HENP sites have become involved with NLANR's active measurement program (AMP). In addition SLAC and CERN are part of the RIPE test-traffic project and SLAC is home for a NIMI machine.

The large End-to-end performance monitoring infrastructure allows the HENP networking community to chart long term trends and closely examine short term glitches across a wide range of networks and connections. The different methodologies provide opportunities to compare results based on different protocols and statistical samples. Understanding agreement and discrepencies between results provides particular insight into the nature of the network.

This paper will highlight the practical side of monitoring by reviewing the special needs of High Energy Nuclear and Particle Physics experiments and provide an overview of the experience of measuring performance across a large number of interconnected networks throughout the world with various methodologies. In particular, results from each project will be compared and disagreement will be analysed. The goal is to address issues for improving understanding for gathering and analysis of accurate monitoring data, but the outlook for the computing goals of HENP will also be examined.

Keywords— **Wide Area Network, Networking, Monitoring, End-to-end, Performance.**

I. Introduction

MODERN High Energy Nuclear and Particle Physics (HENP) experiments at Laboratories around the world present a significant challenge to wide area networks. The BaBar collaboration at the Stanford Linear Accelerator Center (SLAC), the Relativistic Heavy Ion Collider (RHIC) groups at the Brookhaven National Laboratory (BNL) and the Large Hadron Collider (LHC) projects under development at the European Center for Particle Physics (CERN) will all generate petabytes (10^{15}) or exabytes (10^{18}) of data during the lifetime of the experiment. Much of this data will be distributed via the Internet to the experiment's collaborators at Universities and Institutes throughout the world for analysis. Figure 1 shows the extent of the world wide collaboration of HENP research. The shaded countries indicate a University or Institute in that country is a collaborator on one of the experiments listed above or the D0/CDF collaborations at the Fermi National Accelerator Laboratory (FNAL) or the Zeus experiment at the Deutches Elektronen Synchrotron (DESY). In total, there are collaborators in over 50 countries. Although the collaborators are highly concentrated in North America, Europe and the Former U.S.S.R. there are a growing number in South America, the Middle East and the Far East. HENP has even begun to enter Africa. Morocco is a member of the Atlas collaboration and South Africa is an observer nation at CERN. There are a large number of other international collaborations not included in the above sample of experiments, such as the neutrino observatories in Antartica and the Astronomy collaborations based at observatories in Hawaii and Chile, however, there is no doubt the research is truly a world-wide collaborative effort.

Physicists require fast bulk transfers, smooth telnet/ssh sessions and fast database querying. In addition many collaborators use video conferencing and Voice-over-IP (VoIP) is increasing. Regular email and web is also neccessary.

Many experiments define regional centers, where the raw data gathered from the detector, or partially processed data, is made available at a geographically distant place from the experiment for better access for local researchers. The BaBar regional centers are at the Rutherford Appelton Laboratory (RAL) near Oxford in England, at IN2P3 in Lyon in France and at the INFN site at the University of Roma in Italy. Consequently network performance between SLAC and these sites is particularly important. Furthermore, there are ambitious plans to allow the LHC control room to be run from anywhere in the world. For safety reasons this system must be fail safe and will probably require some sort of quality of service (QoS) techniques.

High performance network connectivity throughout the world is so important to the present and future of experimental HENP research that the International Committee for Future Accelerators (ICFA) created a Standing Com-

The authors are with the Stanford Linear Accelerator Center (SLAC), a particle physics laboratory operated for the U.S. Department of Energy by Stanford University. Email them at warrenm@slac.stanford.edu or cottrell@slac.stanford.edu .

mittee for Interregional Connectiviy (SCIC) specifically to monitor and improve internetworking between regions.

In this paper, some of the network monitoring tools in use at HENP laboratories and collaborating Universities and Institutes will be reviewed, and the results from them will be compared.

II. WAN Networking for HENP.

In the United States, the Department of Energy (DoE) funded HENP labs are connected by the Energy Sciences Network (ESnet). SLAC also has a connection via Stanford Campus to the Californian regional network (CALREN2) and FNAL has a connection to the Chicago area metropolitan regional network (MREN). Most research Universities in the U.S. are connected to the vBNS or Abilene (Internet2).

Outside the U.S., most countries have a single national research network (NRN), which connects all the academic and research institutes in that country together and in many cases provides a connection to the U.S. directly. In Europe, the TEN-155 and NorduNet networks provide regional connectivity between their member countries and links to other regions including the U.S.

Peering between the networks is critical. There are over 50 separate networks involved in carrying traffic for HENP research. Many networks connect to each other in New York, although not necessarily at the same location in New York which creates problems in itself. STAR TAP in Chicago is a popular meeting point for research networks, but for European network traffic bound for the US East Coast, or for Asian network traffic bound for the US West coast, this adds significant time to the trip. Hence, many Asian networks connect at locations on the U.S. West coast. CERN is a member of Internet2, providing excellent connectivity between CERN and US research universities.

III. Monitoring Projects

The HENP computer and networking community is actively monitoring performance across a large part of the Internet used by collaborators with tools from a number of projects.

Each project has it's own motivation and methodology, but Wide-Area-Network performance is the primary concern. However, in all cases the end nodes are computers, so inevitably there will be some effect from LANs and the hosts themselves. Some projects specify the machine should be close to the external connection of the site to minimise local effects. In other cases the machines are placed closer to the working networks to get a measure of the full end-to-end performance.

The Active Measurement Program (AMP) [1] has 94 sites (in January 2000), mostly in North America, engaged in full-mesh monitoring using the well know `ping` tool. The `ping`s are sent as a poisson stream approximately once per minute. The project is aimed at sites involved in High Performance Computing (HPC), including SLAC and several BaBar collaborators; University of Colorado (colorado.edu), Colorado State University (colostate.edu), Iowa State University (iastate.edu), University of Cincinnatti (uc.edu), University of California at Irvine (uci.edu), and University of California at Santa Cruz (ucsc.edu).

The RIPE test-traffic (RIPE-TT) project [2] has 41 nodes, mostly in Europe for RIPE customers, but CERN and SLAC also have nodes. Also, nodes at the National Research Network of the Czech Republic (CESnet), the Nordic research and education network (NorduNet), Telia Networks in Sweden, SurfNet in the Nederlands, and the Slovak Academic Network (SAnet) of the Slovak Republic, are interesting to HENP because they are networks that provide connectivity to Universities and Institutes involved in experiments at SLAC or CERN.

Advanced Network's Surveyor project [3] has 59 sites mainly in North America and Europe. CERN, BNL, FNAL and SLAC all have Surveyor machines, as do many Universities that collaborate on experiments at these labs; Brown University (brown.edu), Carnegie-Mellon University (cmu.edu), University of Colorado (colorado.edu), Duke University (duke.edu), Florida State Univeristy (fsu.edu), Iowa State University (iastate.edu), John Hopkins University (jhu.edu), Penn State (psu.edu), University of Minnesota (umn.edu), University of Pennsylvania (upenn.edu), University of Washington (washington.edu) and University of Wisconsin (wisc.edu).

Surveyor and RIPE-TT require GPS antennas and send a poisson stream of one-way UDP packets.

The DoE ping End-to-end Reporting (pingER) project [4] also uses the `ping` tool, but the methodology differs significantly from the AMP project. Instead of a poisson stream, 10 `ping`s, each with 100 byte payload, are sent at one second intervals followed by 10 `ping`s each with a 1000 byte payload at one second intervals, every 30 minutes.

The project has grown significantly and now (January 2000), 593 nodes at 424 sites in 72 countries are monitored by 28 monitoring sites in 15 countries. A total of 2138 end-to-end pairs are monitored, making pingER probably the largest performance monitoring project in the world. Typically, a pingER monitoring site is a laboratory or a university interested in monitoring a certain set of other laboratories or universities that it collaborates with. The project co-ordinators have added a set of beacon sites that all monitoring sites are requested to monitor. Recently particular effort has been made to extend the monitoring of locations in East Europe and the former USSR and to Central and South America and the Middle East, reflecting the increasing reach of high energy nuclear and particle physics research.

IV. Results

Performance between U.S. laboratories connected to the Energy Sciences network (ESnet) and U.S. Universities connected to Internet2 is usually good. That is, packet loss is typically very low ($<< 1\%$) and delay is primarily due to transmission rather than queuing in routers.

This is because high performance research networks have clear head room between average utilization and maximum bandwidth.

However, most of the world's network do not perform as well as ESnet or Internet2. Packet loss can be extremely high in some parts of the world on poorly provisioned links. Despite significant expenditure for international connections, research groups find their packets are swamped because the traffic is sharing bandwidth with commodity traffic.

For example, monitoring between sites on ESnet and academic sites connected to the U.K.'s JAnet network[1] all show a similar pattern that track each other very well, indicating the problem is with the common trans-Atlantic link rather than within JAnet. Packet loss decreases significantly during the summer, Christmas and Easter breaks. The packets containing data from HENP experiments competes with other packets, for example from commercial US websites surfed by the students, for the UK-US bandwidth.

Figure 2 shows the packet loss measured by pingER between DESY (in Hamburg) and the U.S. and highlights the effect of a small amount of dedicated bandwidth between DESY and ESnet provided by the German DFN network. Packet loss between DESY and sites connected to ESnet is typically less than 1%, although there are peaks of high packet loss associated with intensive use on several occassions. However, even these peaks are much lower than the usual packet loss from DESY to non-ESnet sites in North America. At the time of these measurements the packet loss between DESY and sites in Canada and between DESY and non-ESnet sites in the U.S. is often unusable. A common characteristic of overloaded links is the distinctive difference between packet loss during the working week and at weekends seen in the figure.

The difference is not so large in August and September because, as with JAnet, the students are not present at the Universities. DFN upgraded the link in October 1999 by a factor of 4 in bandwidth and packet loss between DESY and non-ESnet sites dropped to closer to the rate for DESY and ESnet. The benefit of dedicated bandwidth is clear.

Figure 3 shows the Round Trip Time (RTT) between SLAC and FNAL measured by pingER and by AMP in December 1999. The bars represent the minimum to maximum range of RTT measured by 1,433 samples of 10x100 byte `ping`s taken by pingER every half an hour, and the points are the 44,263 individual `ping` measurements taken by AMP approximately every minute. A brief visual inspection of the graph indicates that the two measurements are not in complete agreement. It appears there are many higher RTT reported by AMP, especially in the second half of the month. Further analysis of the frequencies of the RTT reported shows better agreement. The minimum RTT reported by pingER is sharply peaked at 59ms with less than 1% of samples reporting a minimum of greater than 59ms. The average RTT reported by pingER is also sharply peaked, with 95% of samples reporting an average RTT of 59ms. The maximum RTT is not quite so sharply peaked, but still 82% of samples reported a maximum RTT of less than 61ms. The RTT reported by AMP, is peaked at 58ms, with over 95% of the samples reporting a RTT of 58-59ms. The slightly lower peak for the AMP samples is possibly due to a smaller payload. It appears the value reported by AMP is similar to the average RTT reported by pingER. These results are not suprising when one considers packets hopping between routers. The queues at routers can change rapidly, in fact on all but un-used links the packets will certainly experience different conditions even arriving even within seconds of each other. The range shown on the pingER measurement and the splattering of high RTT measurements from AMP highlight this. Even though the pingER measurements are separated by one second, the reported RTT can differ by hundreds of milliseconds, and the AMP measurement, perhaps taken only a few seconds after pingER reports a narrow range can differ by many standard deviations.

Similarly, the one-way delay from SLAC to CERN measured by the RIPE-TT and Surveyor boxes varies from point to point, but the overall distributions agree. Figure 4 and 5 shows the frequency distributions of the Surveyor and RIPE-TT probes for January 15-22 2000. Both distributions show strong peaks at around 86.5ms and a secondary peak around 90.5ms. In many cases observation of such a bimodal distribution would indicate a route change for some period during the interval, but in this case neither the Surveyor nor the RIPE-TT route monitoring registered a change, and the difference may be due to load on the link.

Most connections experience at least some period of unreachability. One-way probes indicate unreachability is not symmetric, so TCP/IP would break if either direction broke.

Comparison between pingER and AMP and between Surveyor and RIPE-TT is straightforward, but comparison between pingER and Surveyor is more complex. One method is to simply add the two Surveyor one-way delays between site A and site B to approximate a round trip delay. Correlation between regular *ping*s and the approximated Surveyor RTTs then shows good correlation. This correlation can be verified as being real and not a coincidence by shifting the time. It can be seen that the correlation falls sharply.

Correlation between pingER and the approximated Surveyor RTT is also good. By binning the Surveyor measurements it can be seen that monthly averages agree strongly, daily averages also agree but hourly averages are not quite so good.

Surveyor is more finely grained and better for sites with good connectivity such as Universities with Internet2 connections where it is widely deployed or for assessing performance for demanding applications. One such application that has been investigated is Voice-over-IP (VoIP). Data from Surveyor machines has been used to compare the reliability of Internet telephony with the telephone network outages. 284 million Surveyor probes between SLAC,

[1]The network is currently designated SuperJanet III to recognise signifcant upgrades since the original JAnet.

FNAL, CERN and CMU between November 1998 and July 1999 the probability of no outage of 1 second in a call of length 3 minutes is 75%, and the probability rises sharply, reaching 99.6% for no outage of greater than 10 seconds.

V. Conclusions

Internet End-to-end Performance Monitoring for the HENP community is not, at least not entirely, an academic excercise. The primary motivation is to provide networks that can accomodate the needs of HENP research and allow the network engineers and managers to allocate resources appropriately. The prospect for HENP research over the Internet is good. The overall trend is for better performance because in most cases advances in infrastructure has stayed at least one step ahead of the demands of bandwidth hungry applications. The ESnet backbone will be Terabit by 2003-2005, and other networks such as the National Transparent Optical Network (NTON) also provide high bandwidth for HENP research, at least in the United States. Certainly bandwidth is not everything, but better peering arrangements have also reduced latency. In some cases dedicated bandwidth can also improve performance for HENP.

The HENP network monitoring groups will continue to use a number of tools to help extend the reach of HENP research, improve links to existing researchers, and explore the feasability of the next generation of computing models and data distribution methods. The simple tools such as pingER and AMP can provide valuable insight at very low cost, both to budgets and to the network and are useful to report a summary of performance on high performance links as well as low or overloaded links. The more fine grained tools such as Surveyor and RIPE-TT are most appropriate on high performance links to study short term glitches. Information from several sources is especially valuable and combined results can provide a method of better interpreting results.

VI. Further Work

Performance monitoring for the HENP community is work in progress. pingER is under development and there are ambitious plans to extend its functionality and configurability [5]. Continued investigation of specific events along with identification of bottlenecks to which resources can be allocated is essential. In addition, work will be conducted into quantifying the number and effect of duplicate and out-of-order packets. The effect of load on throughput and the effect of varying the TCP window size and the effect of jumbo packets (especially under IPv6) will be assessed. Work on identifying rate limiting is underway, along with understanding how to make accurate performance measurements in the prescence of differentiated services. Passive monitoring using OCxMON will also begin at SLAC.

Acknowledgments

The authors would like to acknowledge the suggestions of many people. In particular Charley Granieri and the pingER monitoring site's administrators, without whom this work would not be possible. Also we would like to thanks the co-ordinators of all the projects for allowing us to participate.

References

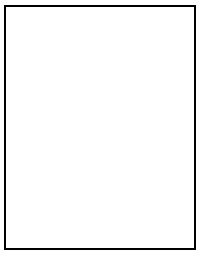
[1] The AMP Project *http://moat.nlanr.net/*, National Science Foundation Cooperative Agreement No. ANI-9807479, and the National Laboratory for Applied Network Research (NLANR).

[2] Henk Uijterwaal and Olaf Kolkman *Internet Delay Measurements using Test Traffic Design Note*, RIPE 158

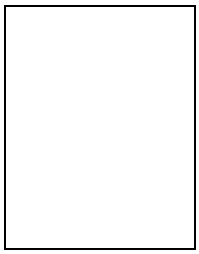
[3] The Surveyor Project *http://www.advanced.org/surveyor*, Advanced Networks.

[4] Warren Matthews and Les Cottrell, *The pingER Project: Active Internet Performance Monitoring for the HENP Community*, IEEE Communications Magazine on Network Traffic Measurements and Experiments

[5] Warren Matthews and Les Cottrell, *Field Work Proposal*, submitted to DoE/MICS



Warren Matthews obtained a PhD in particle physics then joined the operation group for an Internet Service Provider. He joined SLAC as a network specialist in 1997 as part of the DoE Internet End-to-end Performance Monitoring (IEPM) group. He is leading the development of pingER and other tools.



Les Cottrell left the University of Manchester, England in 1967 with a Ph.D. in Nuclear Physics to pursue fame and fortune on the Left Coast of the U.S.A. He joined SLAC as a research physicist in High Energy Physics, focusing on real-time data acquisition and analysis in the Nobel prize winning group that discovered the quark. In 1973/4, he spent a year's leave of absence as a visiting scientists at CERN in Geneva, Switzerland, and in 1979/80 at the IBM U.K. Laboratories at Hursley, England, where he obtained United States Patent 4,688,181 for a a dynamic graphical cursor. He is currently the Assistant Director of the SLAC Computing Services group and lead the computer networking and telecommunications areas. He is also a member of the Energy Sciences Network Site Coordinating Committee (ESCC) and the chairman of the ESnet Network Monitoring Task Force. He was a leader of the effort that, in 1994, resulted in the first Internet connection to mainland China. He is also the leader of the DoE sponsored Internet End-to-end Performance Monitoring (IEPM) effort.

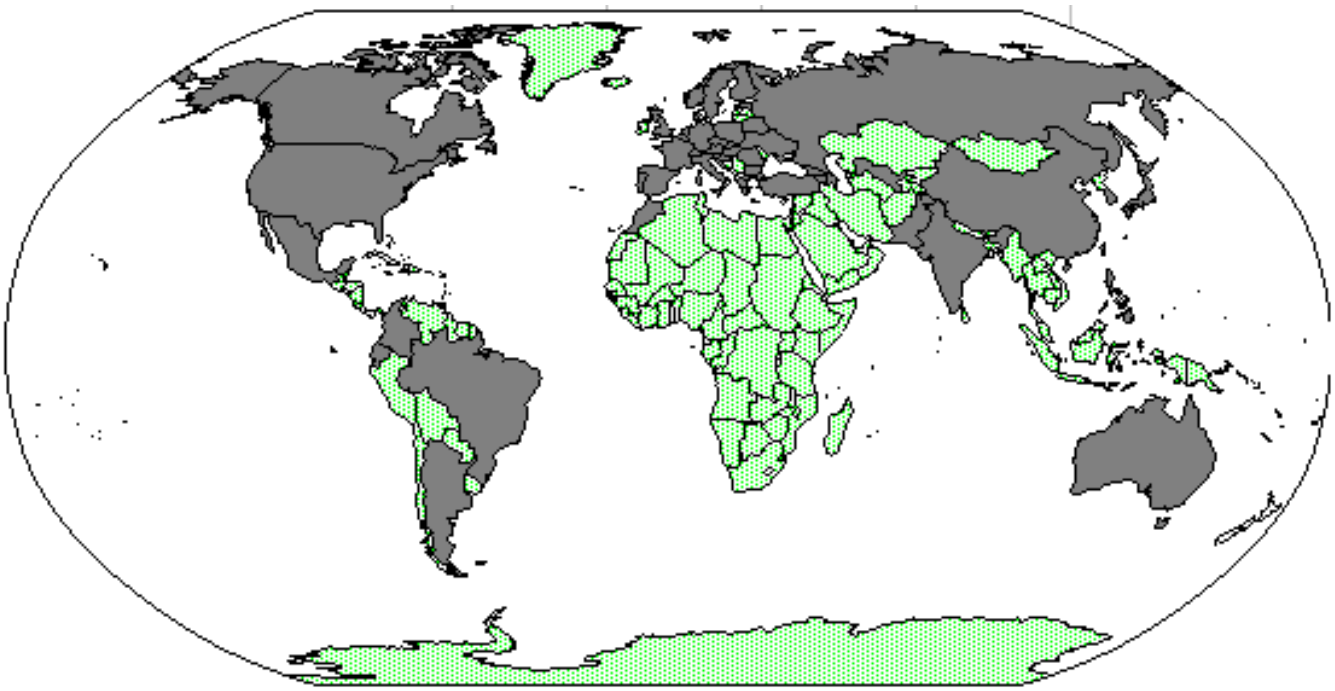


Fig. 1. Countries with HENP collaborators. Countries shaded grey have at least one University collaborating on a HENP experiment.

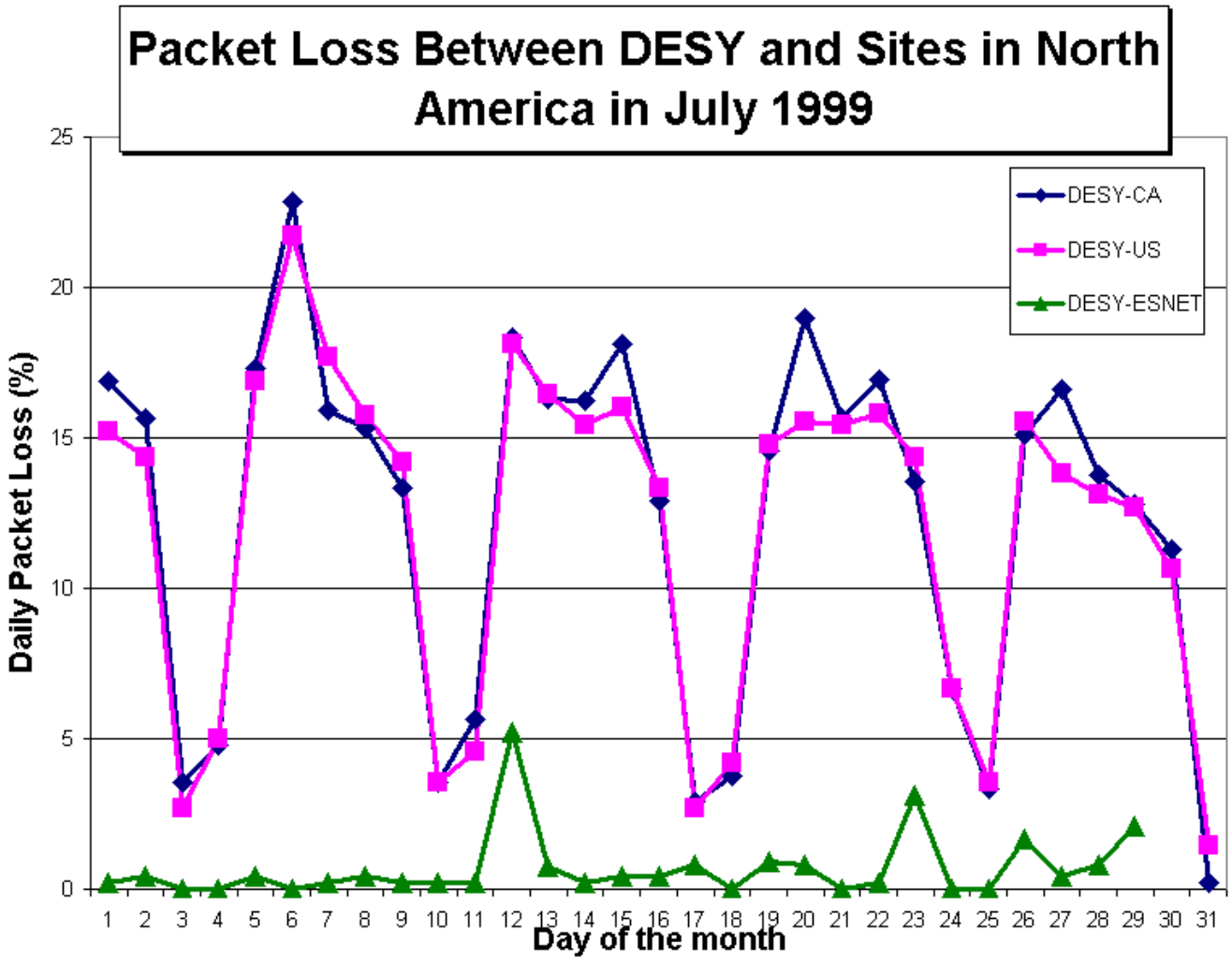


Fig. 2. The effect of dedicated bandwidth between DESY and North America. Sites on ESnet typically suffer much less packet loss because part of the DFN trans-Atlantic link is dedicated.

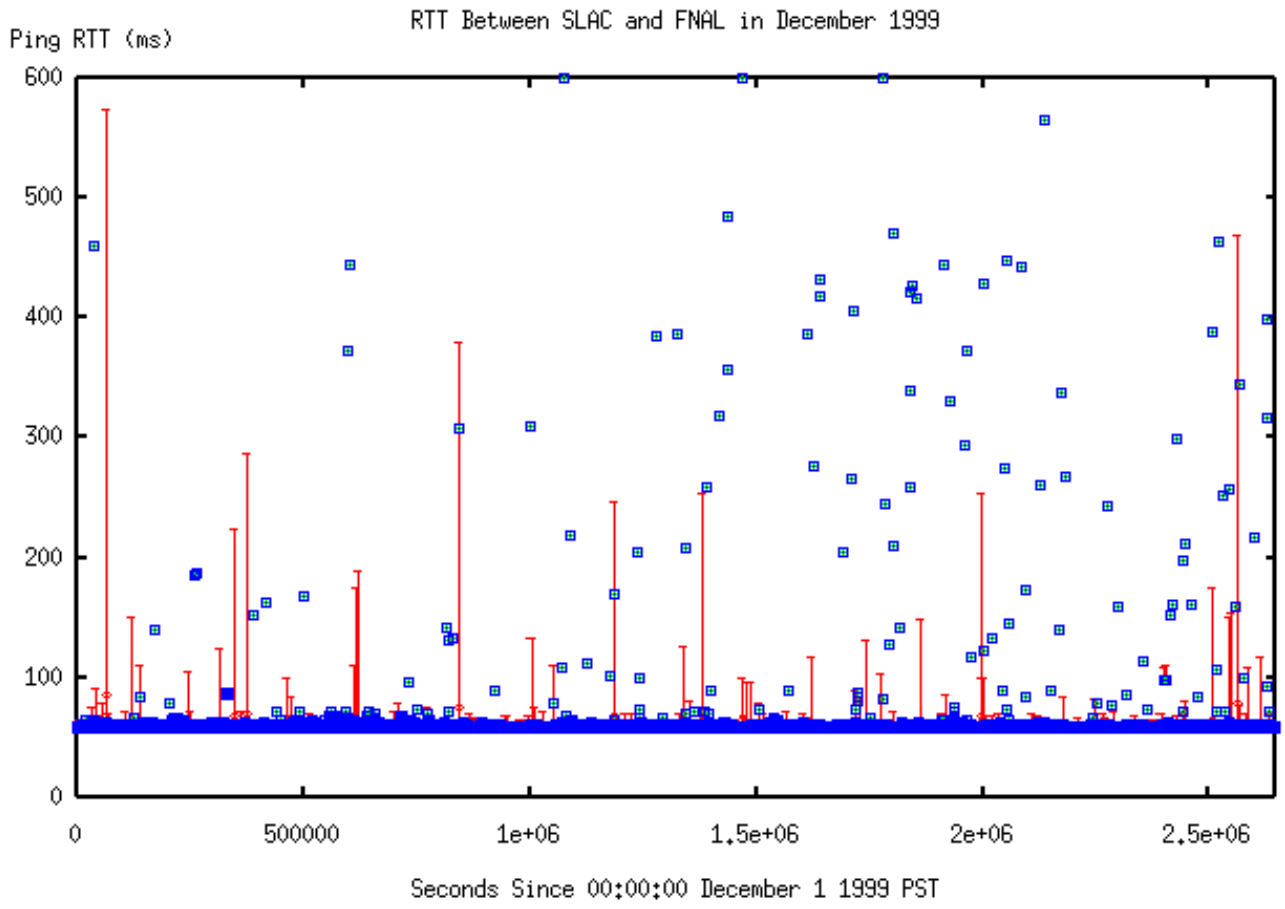


Fig. 3. RTT between SLAC and FNAL measured by pingER and by AMP.

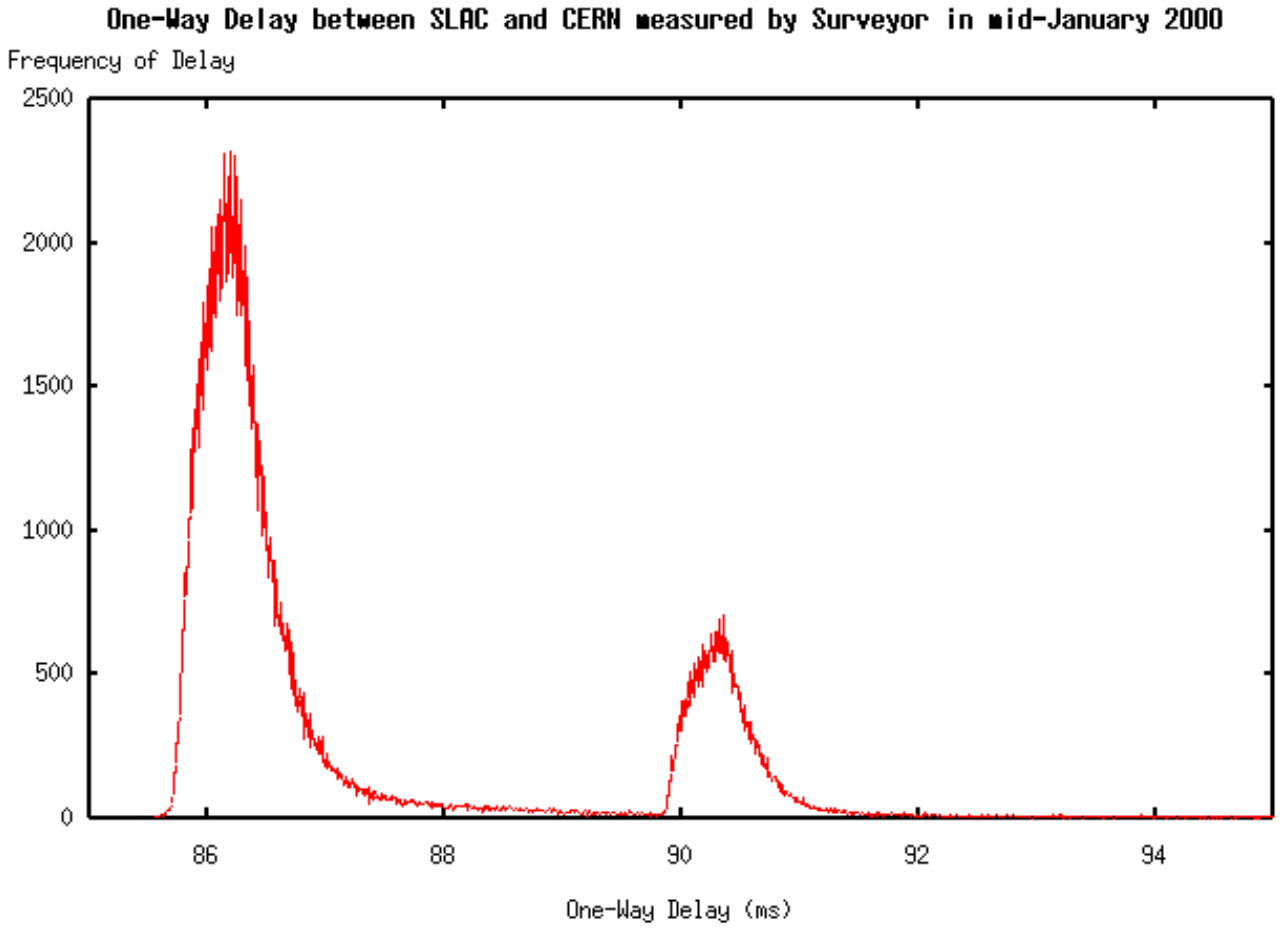


Fig. 4. One-Way Delay from SLAC to CERN measured by Surveyor.

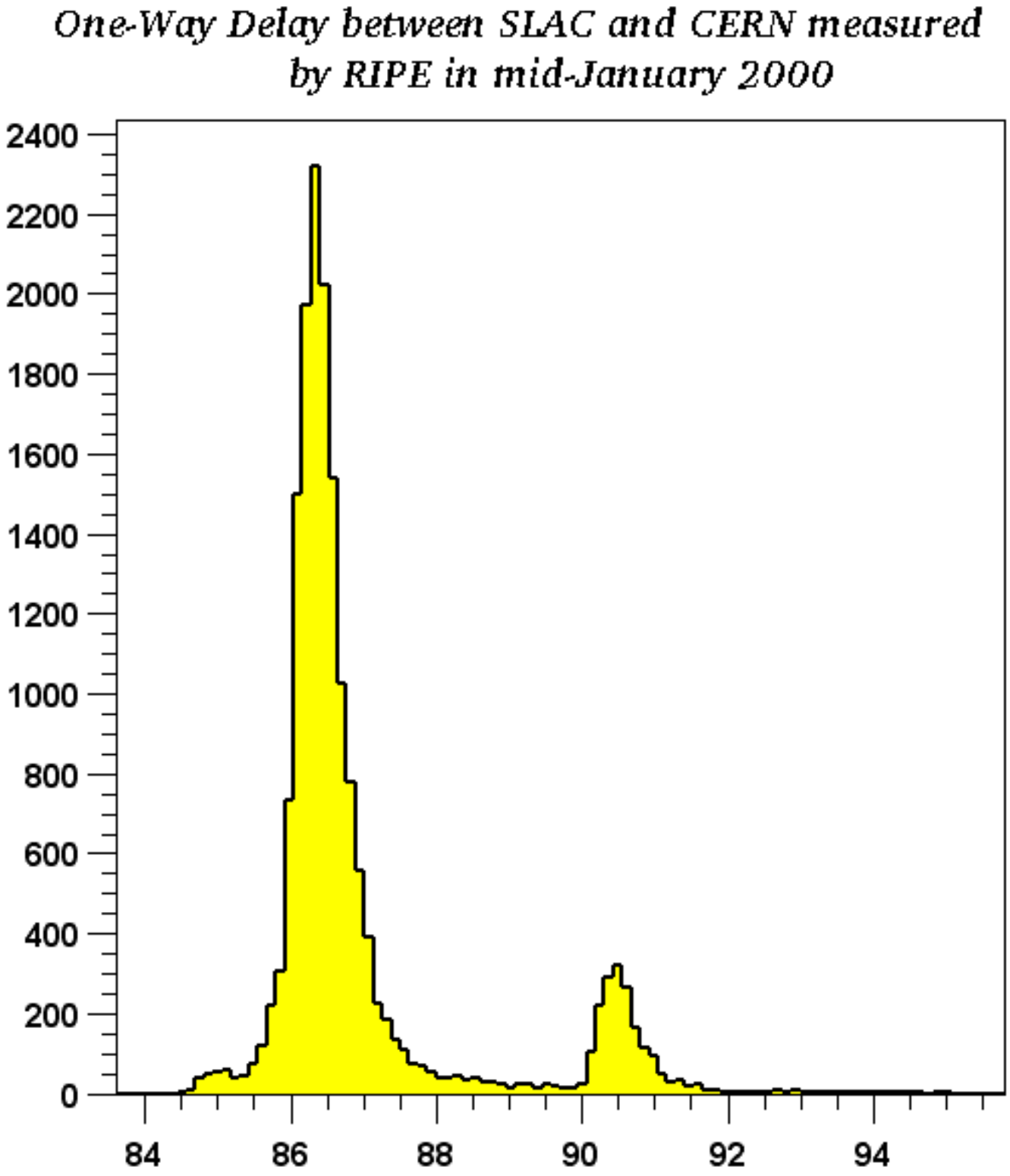


Fig. 5. One-Way Delay from SLAC to CERN measured by RIPE-TT.