SLAC–PUB–8776
February 2001

Passive and Active Monitoring on a High Performance Research Network[*]

Warren Matthews, Les Cottrell and Davide Salomoni
Stanford Linear Accelerator Center, Stanford University, Menlo Park, CA 94025

Abstract

The bold network challenges described in ”Internet End-to-end Performance Monitoring for the High Energy and Nuclear Physics Community” presented at PAM 2000 have been tackled by the intrepid administrators and engineers providing the network services.

After less than a year, the BaBar collaboration has collected almost 100 million particle collision events in a database approaching 165TB (Tera=10^{12}). Around 20TB has been exported via the Internet to the BaBar regional center at IN2P3 in Lyon, France, for processing and around 40 TB of simulated events have been imported to SLAC from Lawrence Livermore National Laboratory (LLNL).

An unforseen challenge has arisen due to recent events and highlighted security concerns at DoE funded labs. New rules and regulations suggest it is only a matter of time before many active performance measurements may not be possible between many sites. Yet, at the same time, the importance of understanding every aspect of the network and eradicating packet loss for high throughput data transfers has become apparent. Work at SLAC to employ passive monitoring using netflow and OC3MON is underway and techniques to supplement and possibly replace the active measurements are being considered.

This paper will detail the special needs and traffic characterisation of a remarkable research project, and how the networking hurdles have been resolved (or not!) to achieve the required high data throughput. Results from active and passive measurements will be compared, and methods for achieving high throughput and the effect on the network will be assessed along with tools that directly measure throughput and applications used to actually transfer data.

Presented at Passive and Active Monitoring (PAM) 2001
Amsterdam, The Netherlands,
April 22-24, 2001

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

Passive and Active Monitoring on a High Performance Research Network.

Warren Matthews, Les Cottrell, Davide Salomoni

Abstract— **The bold network challenges described in "Internet End-to-end Performance Monitoring for the High Energy and Nuclear Physics Community" [1] presented at PAM 2000 have been tackled by the intrepid administrators and engineers providing the network services.**

After less than a year, the BaBar collaboration has collected almost 125 million interesting particle collision events in a database approaching 300 TB (Tera=10^{12}). Around 20TB has been exported via the Internet to the BaBar regional center at IN2P3 in Lyon, France, for processing and around 23 TB of simulated events have been imported to SLAC from Lawrence Livermore National Laboratory (LLNL) [2].

A new challenge has arisen due to recent events and highlighted security concerns at DoE funded labs. New rules and regulations suggest it is only a matter of time before many active performance measurements may not be possible between many sites. Yet, at the same time, the importance of understanding every aspect of the network and eradicating packet loss for high throughput data transfers has become apparent. Work at SLAC to employ passive monitoring using tools such as netflow is underway and techniques to supplement and possibly replace the active measurements are being considered.

This paper will detail the practical side of monitoring and the special needs and traffic characterisation of a remarkable research project, and how the networking hurdles have been resolved (or not!) to achieve the required high data throughput. Results from active and passive measurements will be compared, and methods for achieving high throughput and the effect on the network will be assessed along with tools that directly measure throughput and applications used to actually transfer data.

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

I. Introduction

THE BaBar experiment at SLAC exemplifies how the success of High Energy and Nuclear Physics (HENP) experiments and data intensive science in general is interwoven with the performance of networks. After less than a year the BaBar database is approaching 300TB and stores the reconstructed events of almost 125 million interesting particle collisions. BaBar collaborators plan to double data collection each year and export a third of the data to the BaBar remote computing site at the IN2P3 computer center in Lyon, France. This ambitious goal means within a few years the current SLAC WAN connections will be saturated with the transfer of database files alone and upgrades must be carefully planned and engineered.

The HENP community has a long history of monitoring performance [1]. In particular active end-to-end performance monitoring has been extensively used to discover bottlenecks in the network and identify when collaborating sites need infrastructure upgrades. With the challenge from BaBar the need for monitoring now also includes feeding back the results in order to tune and optimize the connections.

Clearly today's HENP research community's requirements are different to the typical user and the research networks that provide connectivity are engineered differently to the commercial networks. However, the history of the Internet indicates the leading edge technologies soon become standard. It is therefore conceivable that many of the challenges facing HENP now will be faced by others, and solutions being explored may be used for a myriad of tasks in the future

II. The SLAC Network and Connectivity to Collaborators

Typically, the U.S. Laboratories and research Universities have high speed (T3, OC3, OC12) connections to their services providers and the backbone networks involved are state-of-the-art high performance networks running at gigabit per second (Gbps) capacity and several are planned to have terabit per second (Tbps) capacity within a few years.

SLAC has an OC3 (155Mbps) connection to the Energy Sciences Network (ESnet) hub in Sunnyvale, an OC12 (622Mbps) connection to CALREN2 via Stanford Campus, and an OC48 (2.4Gbps) connection to the National Transparent Optical Network (NTON) hub 20 miles away in Burlingame. Most SLAC traffic is routed via ESnet, but CALREN and certain Abilene (Internet2) connected sites are routed to via Stanford Campus. Currently the NTON link is used only for high throughput testing with collaborators at LLNL, Argonne National Laboratory (ANL) and NASA Ames Research Center (ARC) and limited traffic to the California Institute of Technology (Caltech). In addition, during the ESnet transition from the sprint-based ESnet2 to the Qwest-based ESnet3, there is a T3 connection from SLAC to the ESnet hub in Oakland providing connectivity to selected subnets at LLNL and LBNL.

The path to IN2P3 is particularly interesting to SLAC due to heavy use by BaBar. It involves the ESnet link between SLAC and Sunnyvale and the ESnet ATM cloud between Sunnyvale and STAR TAP near Chicago. The capacity of the ATM link is OC12 (622Mbps). In Chicago, traffic to IN2P3 is passed from ESnet to the control of CERN, the European Center for Particle Physics. The

Warren Matthews and Les Cottrell are with the Stanford Linear Accelerator Center (SLAC), a particle physics laboratory operated for the U.S. Department of Energy by Stanford University. Davide Salomoni, formally of SLAC, is now with Colt Telecom in the Netherlands. Please send questions and comments to iepm@slac.stanford.edu .

CERN transatlantic link is 155Mbps.

Since June 2000, the link from CERN to IN2P3 is a 34Mbps ATM link, clearly the bottleneck bandwidth. This is scheduled to be upgraded in June 2001 to 155Mbps. CERN is also a member of Internet2 and it would be possible to route between SLAC and IN2P3 across the Abilene backbone. CERN is considering a connection to NTON, and it is possible a connection to STAR LIGHT will be made at some time in the future.

There are factors other than bandwidth that dictate end-to-end performance and overloaded public peering points and poor peering arrangements are very much a cause for concern. However, typically research networks have peering arrangements away from the poorly performing public exchange points. In some cases transfer rates have been limited by the processing power of the end-node computers involved in the connection.

III. Traffic at HENP Sites

Traffic volume through the SLAC border is often overwhelmingly dominated by file transfers. In particular the data files exchanged between the particle physics databases at SLAC, IN2P3 and LLNL often dwarfs all other traffic. On occasions, secure copy (scp) and various home-grown file transfer program are also heavily used. Other protocols, including HTTP, are often negligible, being less than 20% of the file transfer traffic.

Utilizations of the various SLAC connections are also not as day/night oriented as what might be considered usual, ie busy during the local office hours and unused at other times. The widespread international collaboration means the SLAC network is constantly working, and large file transfers are often initiated during the local night time.

The contention that such a distribution is not typical is justified by comparing to University traffic. At the University of Wisconsin-Madison [3] incoming traffic is about 34% HTTP, 24% FTP, 13% Napster. Outgoing traffic is about 17% HTTP, 24% FTP, 20% Napster. It is probably not suprising that Napster traffic virtually disappears during vacation time when the students are gone.

Furthermore, a comparision to 'typical' traffic on the general Internet [4] can be made. In February 2000, from a total on about 175 Gigabytes of traffic passing through the Ames Internet Exchange, approximately 60-65% was HTTP, about 13% was NNTP, other traffic such as Email (SMTP and POP), Napster, Real Video and games traffic made up for a few percent each.

IV. Security Issues

SLAC is a US Department of Energy (DoE) single purpose tier-3 lab and is therefore exempt from many of the stringent rules imposed by the Department of Energy and other goverment bodies. However, many labs involved in more sensitive research and many companies are increasingly coming under pressure to limit their end-to-end connectivity and restrict access. In many cases this means a simple blocking traffic to non-approved application ports, and in some cases blocking of ICMP ping packets. Such blocking could dramatically reduce network support people's ability to use the 2 most common Internet trouble shooting tools, ping and traceroute. Also unless care is taken, ICMP-based active monitoring will erroneously report high packet loss and longer round trip times (RTTs). By making measurements and contacting the site network administrators, we have confirmed that a few sites are rate limiting but not entirely blocking ping packets. Such knowledge has helped us devise methods (such as probing layer 3 connectivity) to detect ping rate limiting, and in most cases avoid monitoring such sites.

After removing known hosts/sites deploying ICMP rate limiting or blocking, a study was conducted of about 250 hosts. monitored by PingER from SLAC. An estimate of the deployment and effect of ICMP blocking and rate limiting at firewalls or hosts was made. It was determined that in the latter half of 2000 the overall deployment of rate limiting was small at research and education sites. The most likely candidates were in Vietnam and India where it might be expected rate limiting techniqiues may be employed to resolve limited bandwidth issues. It was also concluded that the amount of data gathered providing false performance measurements was also small [5].

Methods of monitoring performance without risking blocking or limiting include: active monitoring using an approved application or at least the application's well-known port number; or to engage in passive monitoring, where bone fide traffic between sites is sniffed and analysed. Care has to be taken in the former (active) case to ensure the monitoring is not perceived as a security style attack. One can select a heavily used application/port where the extra monitoring traffic will not be noticed, or better yet one can notify the remote site administrators of the monitoring activity. The passive approach is extremely valuable in network trouble-shooting, does not introduce extra traffic, and it measures real traffic. However, it is limited in its ability emulate error scenarios, isolating the exact fault location, can generate large amounts of data and since it views packets on the network, it has its own security concerns.

V. Network Monitoring

Years of active ping monitoring between hosts at HENP sites has indicated the overall trend on research and academic networks is towards less packet loss and reduced round trip time and, by definition, better performance [1]. However, evidence that the networks are performing well in transfering individual packets is insufficient to understand optimizing throughput and extra active measurements are often neccessary.

SLAC has been conducting extensive tests to identify the maximum throughput that can be achieved between two end-hosts using the iperf tool [6]. A client run at one site connects to a server at a remote site and a stream of TCP or UDP data is sent.

The iperf tool allows the maximum TCP window to be adjusted and more than one stream can be sent in parallel. Table I shows the throughput achieved between SLAC and

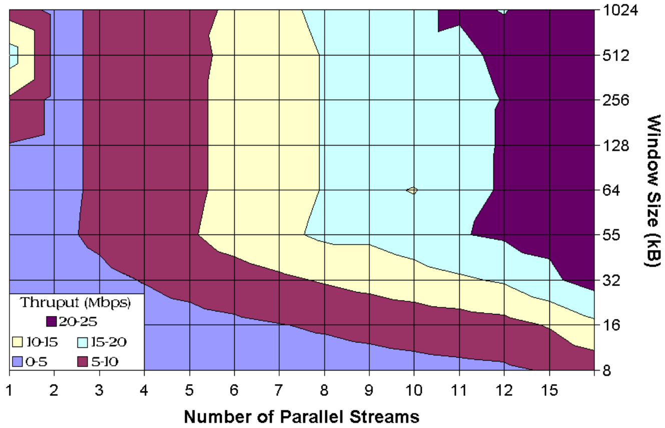

Fig. 1. Effect on throughput of varying window size and number of parallel streams in file transfer.

CERN and between SLAC and Caltech for a fixed product of window size and number of streams. It can be seen that multiple streams has a much greater effect than the window size.

TABLE I

Throughput measured between SLAC and CERN and between SLAC and Caltech using different window sizes and number of streams. Note increasing streams has a greater effect than using large windows.

Site	Window	Streams	Throughput
CERN	256 kB	2	9.45 Mbps
CERN	64 kB	8	26.8 Mbps
Caltech	256 kB	2	46.5 Mbps
Caltech	64 kB	8	63.5 Mbps

Figure 1 shows throughput measured from SLAC to IN2P3. It can be seen, in this case, that there is no improvement with window size greater than 48kB, but the number of streams continues to increase the throughput to the bottleneck bandwidth. The 48kbytes is much lower than that predicted using the product of the bottleneck bandwidth (25-30Mbits/s) and the RTT (170ms). Other sites achieve highest performance with different window sizes, but a similar pattern exists, i.e. increasing the number of parallel streams is more effective than simply increasing the maximum window size.

Further analysis indicates that each stream contributes about the same amount to the total throughput.

Figure 2 shows the throughput between CERN and Caltech tracked from CERN during a week in March 2000. It can be seen the throughput was dramatically impacted on two occasions. Comparing this to data gathered from pingER it is revealed the first hit was due to an increase in round trip time from around 175 ms to around 260 ms due to some routing problems. Packets from CERN were being sent via the Swiss Switch network and Dante rather than directly via the CERN-Abilene connection in Chicago.

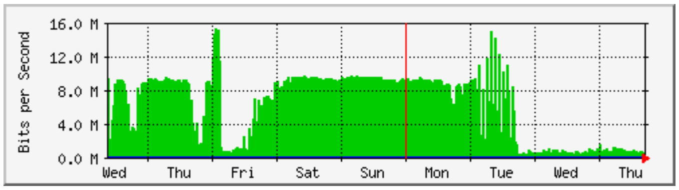

Fig. 2. Throughput between CERN and Caltech measured by the Netperf tool. Note the large drop in throughput on Tuesday.

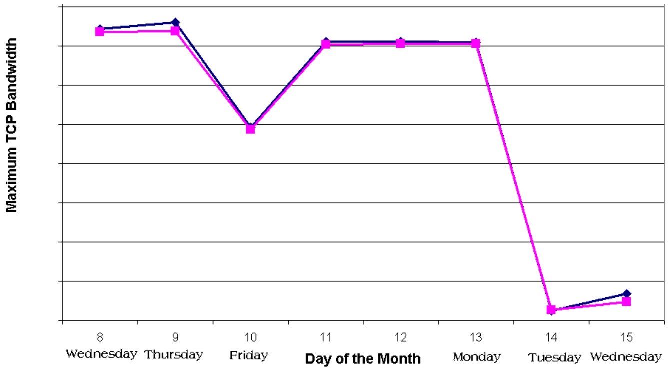

Fig. 3. Throughput between CERN and Caltech Derived ping packet loss and round trip time using the formula of Ott et al. Note the large drop similar to that Shown in figure 2. This graph contains no vertical scale because calibration was found to be unbelievable.

The interesting observation is that even small packet loss, in this case about 1.5%, caused a major impact on throughput. The packet loss is especially relevant for high bandwidth large RTT links [7].

Figure 3 shows the throughput derived from ping packet loss and round trip time using the formula of Ott et al. [8] between CERN and Caltech for the same period illustrated in figure 2. The agreement between the high-impact throughput measurement and the low-impact simple ping metrics is encouraging.

The median round trip time between SLAC and IN2P3 has typically been about 170ms, although recent changes have improved this to around 150ms. Typically there is very low packet loss, less than 0.5%. Hence the formula predicts throughput should achieve 1700kbps. Actual single-stream FTP transfers were found to be around 150kBps (1200kbps). Network conditions can change very rapidly. Even on a stable network throughput varies with a factor 3-5 from one minute to the next so regular 5 minute intervals doesn't give great accuracy. SLAC has been evaluating network simulations using the ns2 program [9]. There is good agreement between the simulated and observed throughput when the samples are measured for long periods (e.g. several measurements of 10 seconds each separated by several hours), except for links with short round trip times, or links with heavy congestion from other sources. Table II shows for several sites, the bottleneck bandwidth (BW), the minimum RTT, the product of BW and RTT, the observed maximum throughput, the square of the correlation coefficient (R^2) between the observed and ns2 predicted throughputs and the improvement achieved by using large maximum window sizes and multiple streams compared to using the default system maximum window size (typically Sun Solaris machines with an 8kByte maximum window/buffer default) and a single stream. The bottleneck bandwidths are estimated using a combination of knowledge of the network configurations and also using tools like pchar and pipechar. Each iperf throughput measurement was made for 10 seconds with a given window size and number of streams. This was repeated with 8 different window sizes from 8kbytes to 1Mbytes and about 22 different numbers of parallel streams from 1 to 40. A typical set of measurements took between 1 and 2 hours. This was repeated up to 10 times at different times of the day and week. Typically we find the maximum throughput is around 90% of the bottleneck bandwidth.

TABLE II

Agreement between measured and simulated throughput using the ns2 simulator program, where MMT is the Maximum Measured Throughput in Mbps and Imp is Improvement factor.

Site	BW Mbps	RTT ms	BW*RTT kbytes	MMT Mbps	R^2	Imp
D'sbury	10	162	203	8.2	0.89	18
IN2P3	28	180	630	26	0.85	59
Caltech	45	10	56	42	0.79	25
LBNL	30	3.4	13	30	0.2	13

Various Passive Monitoring tools have been tested at SLAC. Initial tests involved using TCPDUMP, but the segmented switched network meant there was insufficient data or too many data collection points. SLAC also deployed an OC3MON and used the coralreef tools to collect and analyze data. This area was just getting interesting when the ATM link to the SLAC controlled router was replaced with a gigabit Ethernet connection and it was not possible to continue to use OC3MON. Currently Cisco's netflow tool [10] is utilized to gather data and a home-grown analysis program compiles reports on protocols and applications crossing the SLAC WAN links. Passive tools such as netflow are useful to understand the applications used but so far a useful method to correlate active and passive measurements has eluded us.

VI. Conclusions

Optimizing performance is not currently an easy task for most users. In fact many professional network administrators do little more than run their routers and servers at default settings. However, as can be seen from above there can be dramatic improvments in bulk throughput performance.

Ambitious projects such as those in the HENP arena will not succeed without significant effort to improve performance.

High capacity links are essential, nothing can be done without bandwidth. Packet loss and latency should be minimal. Optimal window size should be set and transfer should consist of multiple parallel streams. These tuning optimizations need to be automated and to achieve will require improved measurement, understanding and modifications of many layers including network, transport and application.

VII. Further Work

Research networks involved in connecting HENP sites are constantly upgrading connections and improving peering relationships. The demanding and ambitious requirements of scientists compel us to look at methods to optimize available bandwidth and improve performance and several projects are beginning that will evaluate some techniques.

The feasibility of modifying bulk throughput applications to automatically and dynamically select the maximum window sizes and number of parallel streams given measurements on the current network state at various levels of detail is being investigated. Further investigation on the benefits of non-standard flavors of TCP for high throughput data transfers will be conducted. SLAC is evaluating an early release of the software from the Web100 project [11]. It is hoped that it will provide some on-the-fly TCP tuning based on TCP measurements made by Web100.

A number of Laboratories and research Universities are proposing various new projects involving monitoring. SLAC is participating in a proposal to deploy NIMI probes into ESnet, and another proposal to simulate the network to attempt to make powerful predictions of future performance.

The HENP community looks forward to Differentiated Services and other Quality of Service (QoS) techniques. In particular a less-than-best-effort per hop behaviour (phb) dubbed 'Scavenger Service' has been proposed to allow file transfer to take up available bandwidth but will not compete with other flows such as interactive web traffic, or even other file transfers or email.

Acknowledgments

The authors would like to thank George Riley for his assistance in using and modifying the ns2 program. The authors would also like to thanks Connie Logg for her comments and suggestions regarding work with the netflow package.

References

[1] Matthews, W., and Cottrell, L., *Internet End-to-end Performance Monitoring for the High Energy Nuclear and Particle Physics Community*, PAM 2000.

[2] Thanks to Adil Hasan, BaBar Database Administrator, for the impressive statistics.

[3] Communication with Peter Couvares, University of Wisconsin.

[4] McCreary, S., and Claffy, kc, *Trends in Wide Area IP Traffic Patterns A View from Ames Internet Exchange*, Monterey, May 2000.

[5] Student Project conducted by Mit Shah. Unpublished.

[6] The Iperf Tool, *http://dast.nlanr.net/Projects/Iperf/release.html*

[7] Feng, W., and Tinnakornsriuphap, P., *The Failure of TCP in High-Performance Computational Grids*, The Proc. of SC 2000, November 2000. LA-UR00-3765

[8] Mathis, Semke, Mahdavi & Ott, *The macroscopic behavior of the TCP congestion avoidance algorithm*, Computer Communication Review, 27(3), July 1997.
[9] The ns simulator, *http://www.isi.edu/nsnam/ns*
[10] Cisco Netflow, *http://www.cisco.com/warp/public/732/netflow/*
[11] The WEB100 Project, *http://www.web100.org*

Warren Matthews is a principal network specialist at the Stanford Linear Accelerator Center (SLAC), a particle physics laboratory operated for the U.S. Department of Energy by Stanford University. Warren has a PhD in particle physics from research conducted at the OPAL experiment at CERN (the European Center for Particle Physics). He worked for an Internet Service Provider for two years before joining SLAC in 1997. Warren loses considerable sleep worrying about network performance, and he hopes new technologies such as QoS and IPv6 may improve his sleep pattern. He works with the D.o.E. sponsored Internet End-to-end Performance Monitoring (IEPM) project and the XIWT Internet Performance Working Group.

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.

Davide Salomoni graduated in Physics from the University of Bologna, Italy, in 1990 but soon after this achievement he realized he was much more willing to play with computers and networks than with electrons and positrons. After some grants from Digital Equipment in the early 1990s to work on Network Management, he spent 7 years with the National Network Center of the Italian Institute for Nuclear Physics, where he participated in the planning and deployment of several incarnations of the Italian Research Network and related services. He was also involved in European Working Groups as technical expert for the evaluation and planning of networks for the European academic and research community. In 1999 he moved to the US west coast and joined the Stanford Linear Accelerator Center as network guru; in his scarce spare time he was also involved in grid-centric networking projects. In 2001 he left California and returned to Europe to work as a network architect for Colt Telecom in the Netherlands. His interests range from high-speed network technologies to William Blake.