Breaking the 1 GByte/sec Barrier? High speed WAN data transfers for science

S. Ravot, J. Bunn, H. Newman, Y. Xia, D. Nae, X. Su,
California Institute of Technology
1200 E California Blvd, Pasadena CA 91125

O. Martin
CERN
1211 Geneva 23, Switzerland

In this paper we describe the current state of the art in equipment, software and methods for transferring large scientific datasets at high speed around the globe. We first present a short introductory history of the use of networking in HEP, some details on the evolution, current status and plans for the Caltech/CERN/DataTAG transAtlantic link, and a description of the topology and capabilities of the research networks between CERN and HEP institutes in the USA. We follow this with some detailed material on the hardware and software environments we have used in collaboration with international partners (including CERN and DataTAG) to break several Internet2 land speed records over the last couple of years. Finally we describe our recent developments in collaboration with Microsoft, Newisys, AMD, Cisco and other industrial partners, in which we are attempting to transfer HEP data files from disk servers at CERN via a 10Gbit network path to disk servers at Caltech's Center for Advanced Computing Research (a total distance of over 11,000 kilometres), at a rate exceeding 1 GByte per second. We describe some solutions being used to overcome networking and hardware performance issues. Whilst such transfers represent the bleeding edge of what is possible today, they are expected to be commonplace at the start of LHC operations in 2007.

***Index Terms*—High performance networking, High speed data transfer, TCP.**

I. Introduction

LHC experiments face unprecedented engineering challenges due to the volumes and complexity of the data, and the need for collaboration among scientists working around in the world. The massive, globally distributed datasets which will be acquired by experiments, are expected to grow to the 100 Petabyte level by 2010, and will require data throughputs on the order of Gigabytes per second between sites located around the globe.

TCP is the most common protocol used for reliable data transfer over IP networks. Since TCP was introduced in 1981[1], network topologies and capacities have evolved dramatically. Although TCP has proved its remarkable capabilities to adapt to vastly different networks, recent studies [5,6] have proved that TCP becomes inefficient when the bandwidth and the latency increase. In particular, TCP's additive increase policy limits its ability to use spare bandwidth.

In this paper we describe experiments that illustrate TCP's limitations. We report on our measurements using the LHCnet, one of the largest network testbeds available today, having 10 Gb/s links connecting CERN in Geneva, Starlight in Chicago and the Caltech campus in Pasadena.

In light of TCP's limitations, we then explain how we have tuned the end-systems and TCP Reno parameters to achieve record breaking data transfers. Finally, we present an experiment currently underway in our group to transfer High Energy Physics data files from a disk server at CERN via a 10Gb/s network path to a disk server at Caltech (a total distance of 11,000 km) at a rate exceeding 1 Giga Byte per second. Whilst such transfers represent the bleeding edge of what is possible today, they are expected to be common practice at the start of LHC experiments in 2007.

II. TCP limitations illustrated by examples

A. *TCP background*[1]

TCP is a reliable data protocol that operates across packet-switched networks. It identifies packets with sequence numbers and uses acknowledgements to allow the sender and the receiver to coordinate with one another to achieve reliable packet transfer. Concurrently, the congestion control mechanism underlying a TCP connection avoids collapses due to congestion and ensures the fairness of network usage.

TCP uses a control variable called the congestion window. The congestion window is the maximum number of unacknowledged packets a source can send, i.e., the number of packets in the pipe formed by the links and buffers along a transmission path.

Congestion control is achieved by dynamically adjusting the congestion window according to the additive-increase and multiplicative-decrease algorithm (AIMD). During the congestion avoidance phase, without any packet loss, the congestion window is incremented at a constant rate of one segment per round trip time (additive increase). Each time a loss is detected, the congestion widow is halved

[1] This is a brief and simplified description of TCP, a more complete reference is [2].

(multiplicative decrease). Note that in the most wide deployed TCP version (TCP Reno and its variants) the only feedback from the network used to adjust the congestion control algorithm is packet loss.

Due to its elegant design, TCP has achieved remarkable success in efficiently using the available bandwidth, in allocating the bandwidth fairly among users, and – importantly - in reallocating bandwidth shares expediently as the use and/or available bandwidth capacity changes over time. However, recent developments have provided evidence that TCP is unable to take advantage of long-haul backbone network capacities in the 10 Gbps range. In the following section we illustrate this problem based on our experience with managing the transatlantic LHC network.

B. Testbed description

The California Institute of Technology and CERN have deployed (in the context of the DataTag[3] project) one of the largest transcontinental networking testbeds providing 10 Gigabit/s Ethernet access capabilities connecting the Starlight facilities in Chicago and the CERN computing center in Geneva through an OC-192 circuit. The testbed has been extended to the Caltech Campus at Pasadena (CA) through the shared IP backbones of Abilene and CENIC, and a 10 gigabit per second local loop dedicated to R&D traffic between downtown Los Angeles and the Caltech campus in Pasadena. This testbed, shown on Figure 1 is an ideal facility for gathering experimental data on TCP's performance.

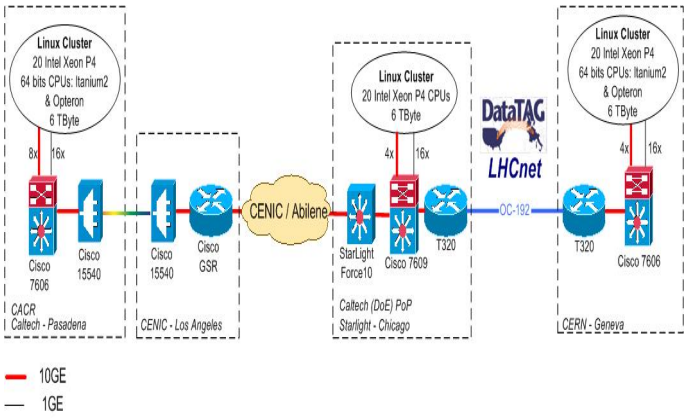

Figure 1: Transatlantic testbed

C. Poor TCP responsive

As administrators of a high speed transatlantic network, we regularly receive complaints about network performance when distributed applications are able to achieve only a small fraction of the nominal bandwidth.

There are two well-known possible reasons for such poor performance. The first reason is a high transmission error rate. While a packet loss can be caused by congestion, it can also be caused by random bit errors. Since TCP lacks an error-nature classification mechanism, the congestion window is halved unnecessarily when there are packet losses caused by bit errors, even though bandwidth is still available. Our measurements have shown that the loss error rate is zero on our testbed. We are able to transmit data at more than 6 Gbps across our un-congested testbed for several hours without experiencing a single packet loss.

The second common cause of problems is an improper setting of TCP buffer sizes. For example, [4] shows that tuned TCP buffers provide a factor of more than 20x performance gain for connections between Lawrence Berkeley National Lab in California and CERN in Geneva (which has a 180 ms Round Trip Time (RTT)). With tuned TCP buffers, the measured transfer speeds increased by more than an order of magnitude. However, buffer size adjustments are insufficient by themselves to achieve end-to-end Gb/s throughput and saturate our network.

The problem lies in the congestion algorithm itself. AIMD oscillations degrade bandwidth utilization, especially at the bottom of the AIMD saw-tooth. An additive increase by one segment per RTT after a multiplicative decrease is too conservative and substantially underutilizes the capacity of high-speed optical networks. A new parameter which describes the responsiveness of TCP is introduced in [7]. The responsiveness measures how long it takes to recover from a packet loss and eventually return to the original transmission rate (prior to the packet loss), assuming that the congestion window size is equal to the bandwidth-delay product when the packet is lost. Table 2 summarizes the recovery times on our testbed.

Path	Bandwidth	RTT (ms)	MTU (Byte)	Time to recover
Geneva-Los Angeles	1 Gb/s	180	1500	23 min
Geneva-Tokyo	1 Gb/s	300	1500	1 hr 04 min
LAN	10 Gb/s	1	1500	430 ms
Geneva–Chicago	10 Gb/s	120	1500	1 hr 32 min
Geneva-Los Angeles	10 Gb/s	180	1500	3 hr 51 min
Geneva-Los Angeles	10 Gb/s	180	9000	38 min

Table 2 : TCP's responsiveness on the assumption that the congestion window increases by one MSS each RTT.

As shown on Table 2, TCP's responsiveness is improved by larger MTUs[2]. Jumbo frames (9000 Bytes) accelerate the congestion window increase by a factor of six compared to the standard MTU (1500 Bytes). Jumbo frames not only reduce I/O overhead (i.e. CPU load) on end-hosts, they also improve the responsiveness of TCP. Unfortunately, Jumbo frames are not supported by all network equipment and by all network operators. Note that the coexistence of Jumbo and standard MTUs introduces some fairness issues [15].

The poor reactivity of TCP has a direct impact on performance in a lossy environment. TCP is much more sensitive to packet loss in a WAN than in a LAN. We used

[2] Maximum transmission unit. It defines the largest size of packets that an interface can transmit without needing to fragment.

a packet dropper [9] to measure the effect of packet loss in a LAN (RTT= 0.04ms) and in a WAN (Geneva –Chicago: RTT=120ms). Figure 2 reports the bandwidth utilization as a function of the packet loss rate. Both connections have 1 Gb/s of available bandwidth. This illustrates how TCP is much more sensitive to packet losses in a WAN than in a LAN. For example, if the loss rate is equal to 0.01%, e.g. 1 packet lost every 10000 packets transmitted, the bandwidth utilization is almost 100% in a LAN but only 1.2% in a WAN.

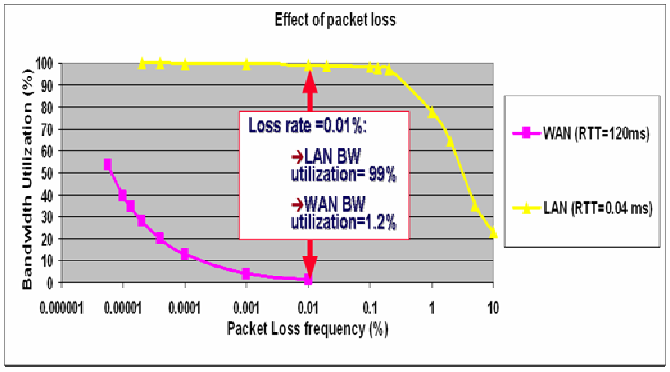

Figure 2: Effect of packet loss over a LAN and a WAN

III. RECORD-BREAKING PERFORMANCE

A. *Contest rules*

The Internet2 Land Speed Record (LSR) [8] competition for the highest-bandwidth, end-to-end networks is an open and ongoing contest. Internet2 Land Speed Record entries are judged on a combination of how much bandwidth they used and how much distance they covered end-to-end, using standard Internet (TCP/IP) protocols. All the hardware and software used along the path must be publicly available. The contest rules can be found on the LSR website. Since the year 2000, when the first record entry was filled, the records for the 4 categories (IPv4 andIPv6, single and multi-stream) have been broken several times. The current record shows a factor of 2000 increase over the initial entry.

B. *Challenges and limitations*

Probably the most restrictive rule of the contest is rule number 3, dictating the use of a standard TCP implementation as described in RFCs 791 and 793. As illustrated in the first part of this paper, the current implementation of TCP has severe limitations when it comes to "Long Fat Pipes". The limitations due to the congestion avoidance algorithm (AIMD) have a direct implication on high speed tests: **no packets can be lost during the transfer**. A single packet loss would halve the throughput and the time to recover from the loss would destroy the chances of winning the contest.

To avoid this problem, one simply needs to reduce the packet-loss rate! In our environment, packet loss is due exclusively to congestion in the network, i.e., packets are dropped when the number of unacknowledged packets exceeds the available capacity of the network. In order to reduce the packet-loss rate, we must prevent the increase of the congestion window before it reaches a congested state. Because explicit control of the congestion window is not possible, we turn to the flow-control window (TCP buffer sizing) to implicitly cap the congestion-window size to the bandwidth-delay product of the wide-area network so that the network approaches congestion but never actually reaches it.

C. *Test Setup*

The network path we used crossed dedicated networks (DataTag) as well as shared networks (Abilene/CENIC). We had to take special care to not interfere with production traffic on the shared networks.

The current record, a memory-to-memory data transfer at 6,5 Gbps with a single TCP stream between Geneva and Los-Angeles, was set using an Opteron (2x Opteron 2.2 GHz Tyan 2882, 2 GB memory) as the sender and an Itanuim2 (HP rx4640, 4x 1.5GHz Itanium-2, zx1 chipset, 8GB memory) as the receiver. Both hosts were equipped with S2io[3] 10 GE network adapters. Each node ran Linux 2.6.6 with Jumbo frames and optimized buffer sizes set to be approximately the bandwidth-delay product. The network topology used to set the single stream record is shown on Figure 5. All intermediate routers/switches on the path supported 9000 byte MTU.

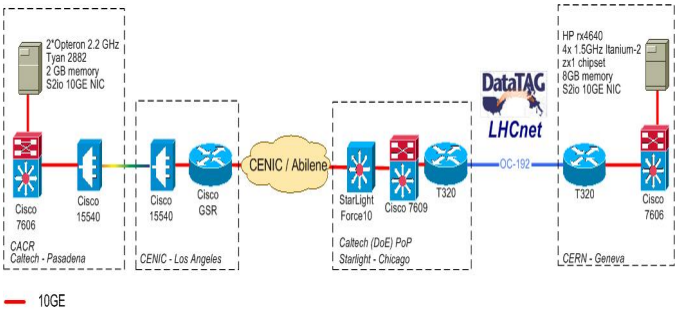

Figure 5: Internet 2 LSR - Single TCP stream at 6.5 Gb/s between CERN and Caltech

D. *LSR History*

The LSR competition has helped to establish the feasibility of multi-Gigabit per second single stream IPv4 & IPv6 data transfers. It illustrates that it is possible today, with commercial off-the-shelf components, to achieve transoceanic end-to-end Gb/s throughput across shared IP backbones such as CENIC and Abilene.

Today, the record is regularly cited as a reference in the network community. Its past evolution is useful for future networks planning and design. The achieved performance serves is an excellent benchmark reference to evaluate new TCP implementations (which should be able to reach the

[3] www.s2io.com

same level of performance across uncongested networks).

The history of IPv4 LSR is shown in Figure 6 an. Over the last two years, the IPv4 performances have increased by a factor 20. We note that these rates are much higher than Moore's Law.

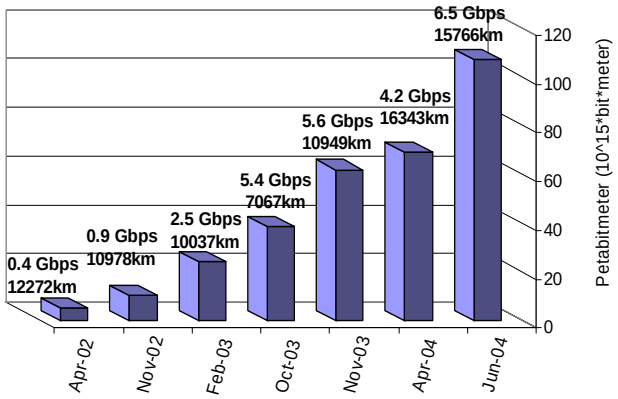

Figure 6: IPv4 LSR History. The Caltech/CERN team did not set the Apr-04 and Apr-02 records. The Jun-04 data has been submitted by Caltech/CERN to the LSR committee and probably qualifies as a new record.

Memory-to-memory transfer is only one aspect of high speed data transfers over a Wide Area Network. The next step of practical importance is disk-to-disk transfers. The introduction of storage devices in the setup adds a new degree of complexity to the challenge.

IV. Disk to Disk Transfers: Breaking the 1 GByte/s Barrier

Although memory to memory tests provides significant insights on the performance of the TCP protocol, in practice the transfer of scientific data typically takes place from disk to disk. The memory-to-memory model provides a simplified setup that helps debug many of the network problems, network card driver/operating system problems (e.g. high interrupt usage for network activity), TCP AIMD algorithm problems and so on. Achieving high throughput in memory-to-memory tests is a necessary step towards high speed disk-to-disk transfers, but it does not guarantee it.

There are a number of potential bottlenecks when making host-to-host data transfers. The Wide Area Network has been the bottleneck for many years but this is no longer the case. For example, the average load on the Abilene backbone is no more than 10 % of the available bandwidth, so there is plenty of unused capacity. As described in this paper we have made significant progress in overcoming the limitations of TCP's use in high speed WANs. And new advances in TCP algorithms such as FAST TCP[11], HSTCP[12], TCP Westwood+[13] or HTCP[16] are succeeding in improving data transport speeds and reliability. The main remaining obstacle to high speed disk to disk transfers is now the storage systems.

A. Hardware limitation

End-hosts have to write/read data from disks and transmit/receive data across the network simultaneously. Those two distinct tasks share many of the same host resources (CPU, PCI-X bus, memory, chipset). The performance achievable separately from the host's memory to its disks, and that from memory to memory across the network, do not automatically equate to the same level of performance for real disk to disk transfers across network.

B. Progress in constructing systems and software to break the 1 GB/s barrier

Since April 2004 we have been collaborating with Microsoft, S2io, Newisys and AMD on the design of a prototype server system capable of sustaining 1 GB/s throughput from disk to disk across a long haul WAN. The prototype is based on the Newisys 4300 AMD system[4], a quad AMD Opteron 848 2.2GHz with three AMD-8131 chipsets and equipped with 16GB of PC3200 DDR memory. Two S2io 10GE NICs are installed in the first and second 64-bit/133MHz PCI slots. Three Supermicro DAC-SATA-MV8 controllers are installed in two 64bit/133MHz and one 64bit/66MHz PCI slots. Each Supermicro controller card has eight Western Digital 250GB SATA 150 7200RPM hard drives in two separated SATA disk cages. The 24 hard drives comprise a single RAID set offering a total capacity of 5TB. The Sprototype systems run the 64-Bit Extended Systems Edition for AMD64 of Microsoft's Windows Server 2003.

In order to avoid the 8.5 Gbps theoretical throughput limitation on the PCI-X bus, we use 802.3ad link aggregation to form a logical interface using two physical 10 GE network adapters. This provides a theoretical limit for the bundled link of 17 Gb/s (twice 8.5 Gb/s) - nominally exceeding our target of 1 GB/s. Initial back-to-back tests with a pair of these prototypes showed performance at 11.1 Gb/s from memory-to-memory, with 52.4% of the CPU utilized.

The RAID controllers used are Supermicro DAC-SATA-MV8, which we measured with eight drives achieve 445MB/s sequential read and 455MB/s sequential write. This is an 85% increase over the best write performance in the Linux systems, and is mainly due to better drivers and optimization in the Microsoft OS. With three Supermicro controllers and 24 disks, the throughput reaches 1.2GB/s in read and write with a CPU utilization of less than 20%.

Using the same prototype systems, we made tests across the 10Gb/s WAN using a single pair of 10GE network adapters. We transferred a 1 TByte file from CERN to Caltech at 4.3 Gbits/s.(536 MBytes/s).

V. Summary and Future Work

While the current TCP congestion control mechanism

[4] Newisys 4300 Enterprise-Class Server: http://www.newisys.com/products/4300.html

has proved to be scalable during the past 20 years, it is less effective on current and next generation high speed networks. In this paper we illustrated with practical examples the limitations of the additive-increase multiplicative-decrease (AIMD) algorithm that governs the performance of a TCP connection.

Despite these limitations, we described how we have established the feasibility of multi-Gigabit per second single stream intercontinental and transoceanic throughput by demonstrating a 6.5 Gb/s transfer using a single TCP stream between Los-Angeles and Geneva. We elaborated on how TCP can be fine tuned to improve its performance effectively on non-congested networks. To cope with realistic large networks where congestion does occur, we are working on the development and performance evaluation of new TCP software stacks such as FAST TCP.

The data transport protocol (TCP) is only one component of a complex systems set that together determine the end to end data transfer performance experienced by the user. Other components, such as the end systems' bus architecture, memory subsystem and disk storage configuration and technology all contribute to the achieved data rate We described how there is a factor of more than ten between memory-to-memory transfer performance and disk-to-disk transfer performance.

We are pursuing a vigorous research and development program in collaboration with partners in the industry with the goal of balancing the systems and software to achieve intercontinental file transfers at rates up to and exceeding 1GB/sec.

ACKNOWLEDGEMENTS

The transatlantic network used for conducting this research is funded by the US Line Consortium (USLIC), which is financed by DoE via Caltech (grant DE-FG03-92-ER40701), NSF (grant ANI 9730202), CERN, Canadian HEP and WHO in Switzerland. The Caltech team is funded by the DoE (grant DE-FC02-01ER25459) and NSF (ANI-0230967). The CERN team is funded by the IST Program of the European Union (grant IST-2001-32459). The authors would like to thank the network operations staff of StarLight, Abilene, NLR and CENIC for their generous support while the testbed was provisioned and the ARTI group at Cisco System for their strong financial support to the deployment of the local loop at Los-Angeles.

REFERENCES

[1] Jon Postel, "Transmission Control Protocol (TCP) - RFC 793," September 1981.
[2] W.R Stevens, "TCP/IP Illustrated, Volume 1: The Protocols," Addison-Wesley, 1994.
[3] "Research & technological development for a Data TransAtlantic Grid," See www.datatag.org.
[4] T. Dunigan, M. Mathis, B. Tierney, "TCP Tuning Daemon; In Proc. of SuperComputing 2002.
[5] W. Feng, P. Tinnakornsrisuphap, "The Failure of TCP in High-Performance Computational Grids," Supercomputing 2000.
[6] S. H. Low, F. Paganini, J. Wang and J. C. Doyle," Linear Stability of TCP/RED and a Scalable Control," Computer Networks Journal, 43(5):633-647, December 2003.
[7] J.P. Martin-Flatin and S. Ravot," TCP Congestion Control in Fast Long-Distance Networks," Technical Report CALT-68-2398, California Institute of Technology, July 2002.
[8] Internet2 Land Speed Record competition for the highest-bandwidth: http://lsr.internet2.edu
[9] The packet dropper is provided as a patch against the Linux kernel 2.4.20 and is available at https://mgmt.datatag.org/sravot/packet_dropper/
[10] A description of the problem is available at http://sravot.home.cern.ch/sravot/Networking/TCP_performance/
[11] Cheng Jin, David X. Wei and Steven H. Low "FAST TCP: motivation, architecture, algorithms, performance", IEEE Infocom, March 2004
[12] Sally Floyd "HighSpeed TCP for Large Congestion Windows" RFC 3649, Experimental, December 2003.
[13] L. A. Grieco and S. Mascolo,"A Mathematical Model of Westwood + TCP Congestion Control Algorithm",18th International Teletraffic Congress 2003 (ITC 2003).
[14] www.supermicro.com
[15] S. Ravot, Y. Xia, D. Nae, X. Su, H. Newman, J. Bunn, O. Martin, "A practical approach to TCP high speed WAN data transfers". Proc. of the First Workshop on Provisioning & Transport for Hybrid Networks (PATHNets) San José, CA, USA, Oct 2004
[16] R.N.Shorten, D.J.Leith,J.Foy, R.Kilduff, "Analysis and design of congestion control in synchronised communication networks" Proc. 12th Yale Workshop on Adaptive & Learning Systems, May 2003.