Next Generation HEP Networks at Supercomputing 2012

Harvey Newman, Artur Barczyk, Azher Mughal, Sandor Rozsa,
Ramiro Voicu, Iosif Legrand, Steven Lo, Dorian Kcira
California Institute of Technology

Randall Sobie, Ian Gable, Colin Leavett-Brown, Yvan Savard
University of Victoria

Shawn Mckee, Roy Hocket, Ben Meekhof
University of Michigan

Marilyn Hay
BCNET

Thomas Tam
CANARIE

Abstract

We review the demonstration of next generation high performance 100 Gbps networks for HEP that took place at the Supercomputing 2012 (SC12) conference in Salt Lake City. Three 100 Gbps circuits were established from the California Institute of Technology (Caltech), the University of Victoria (UVic) and the University of Michigan (UMich) to the conference exhibition. We were able to efficiently utilize these circuits using a limited set of hardware, surpassing previous records established at SC11. Highlights include a record overall disk-to-disk rate using the three links of 187 Gbps, a unidirectional transfer between storage systems in Victoria and Salt Lake on one link of 96 Gbps, an 80 Gbps transfer from Caltech to a single server with two 40GE interfaces at Salt Lake, and a transfer using Remote Data Memory Access (RDMA) over Ethernet between Pasadena and Salt Lake that sustained 75 Gbps and a record total aggregate memory-to-memory transfer rate between all sites of 339 Gbps. In addition, a total of 3.8 Petabytes was transferred over the three days of the conference, including 2 Petabytes on the last day.

System Design

All three 100 GE WAN links (Figure 1) were terminated on an Alcatel SR12 router contained in the Caltech Booth (Fig. 2). A Link Aggregation Group (LAG) of 6x40GE links between SR12 and Force 10 Z9000 core switch was created to provided a path to servers communicating with UVic and UMich.

Both SSD disk servers and a *Data Direct Networks* (DDN) Lustre Storage Appliance were used to receive data incoming from the remote sites. The SSD servers included SuperMicro servers housing 24 SSD drives with three RAID disk controllers in addition to 2011 generation servers with 16 disks. The DDN appliance included SFA-12k controllers connected to five disk enclosures with a total of 288 SAS disks, each of 450 GB capacity. Eight Lustre Object Storage Servers (OSS) were connected to the DDN controllers through Infiniband. Six Lustre clients were used to read and write over the wide area network using a second Mellanox 40 GE interface.

The UMich end site consisted of a mixture of grid site production resources, including a host configured with dual port 40 GE network card connected via Host Bus Adapters to 240 conventional disks. The UVic end site was able to produce disk reads above 95 Gbps (Fig. 3) using only 4 IBM x3650 M4 servers each populated with 16 OCZ Vertex 4 SSDs. An individual IBM server was able to read from disk at 38 Gbps and write stably at 24 Gbps. Fusion-io provided Caltech with ten 4 TB ioDrive Duo PCIe storage cards. The cards were installed in two 3U SuperMicro Sandy Bridge Servers which were then connected directly to the 40 GE ports on the SR12. One server with five Fusion IO drives was able to sustain a steady disk write of 80 Gbps (Fig. 6).

Results

We have achieved a record transfer rate of 339 Gbps to the SuperComputing 2012 conference from sites in the United States and Canada using 40G and 100G network technology. The results will help establish new methods and technologies for transporting data around the globe for particle physics and other fields of research. The SC12 demonstrations achieved their goal of clearing the way to the next level of data intensive science.

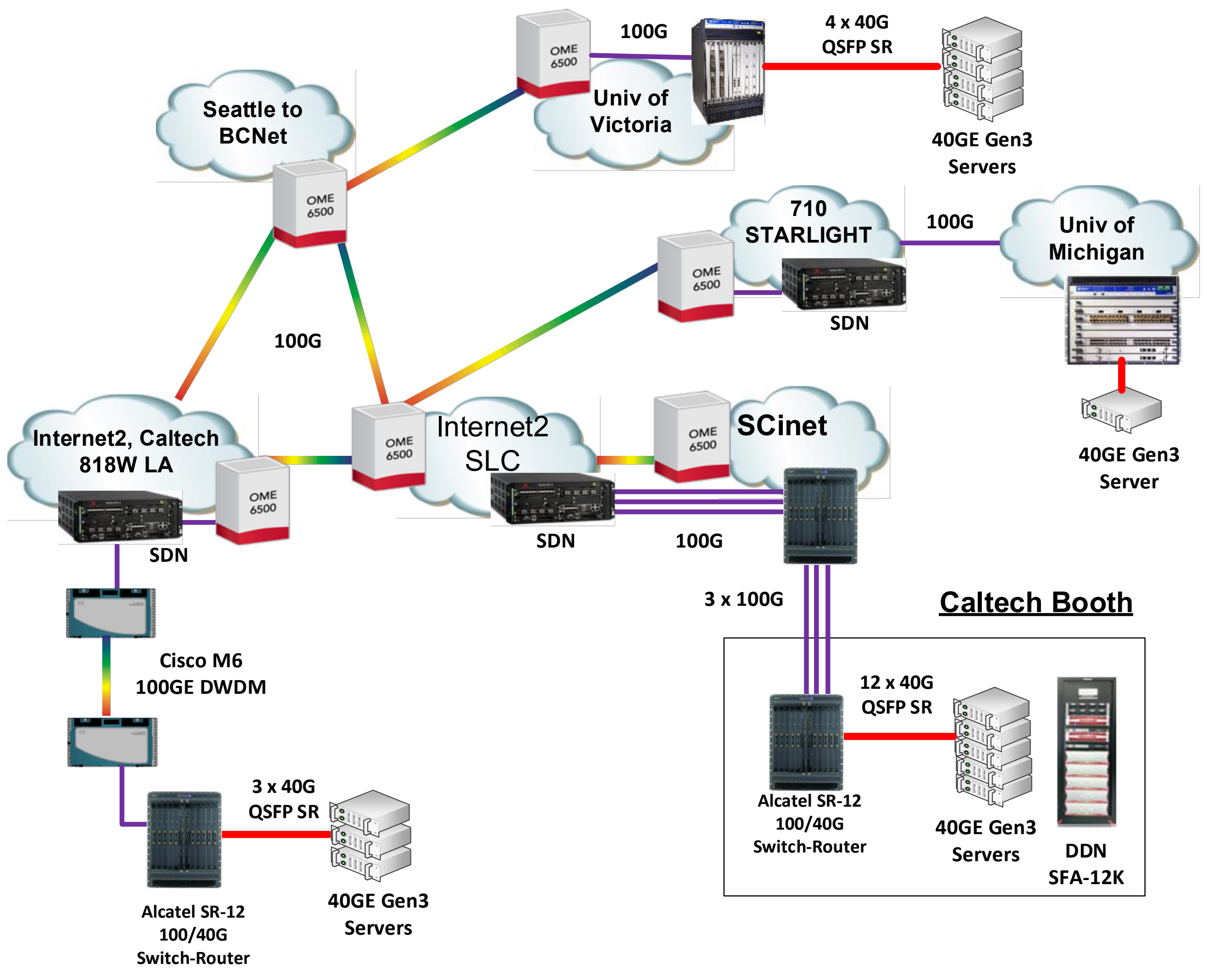


Figure 1. The wide area network diagram for SC 12 demonstration. 100 Gbps network connections were established between Caltech, the University of Victoria and the University of Michigan to the SC exhibition floor in Salt Lake City. Network links were provided by CANARIE and BCNET in Canada and Internet 2 in the US.

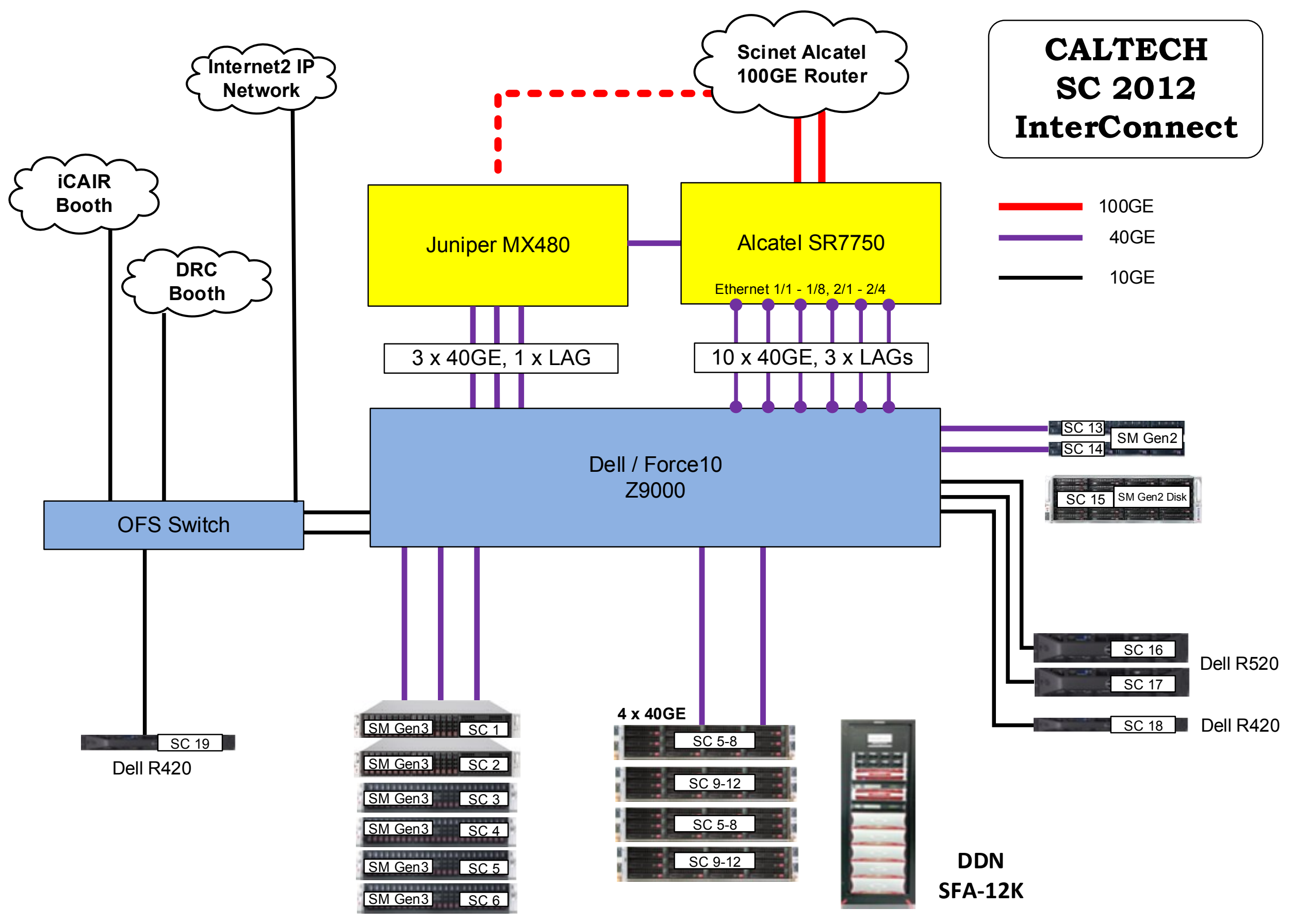
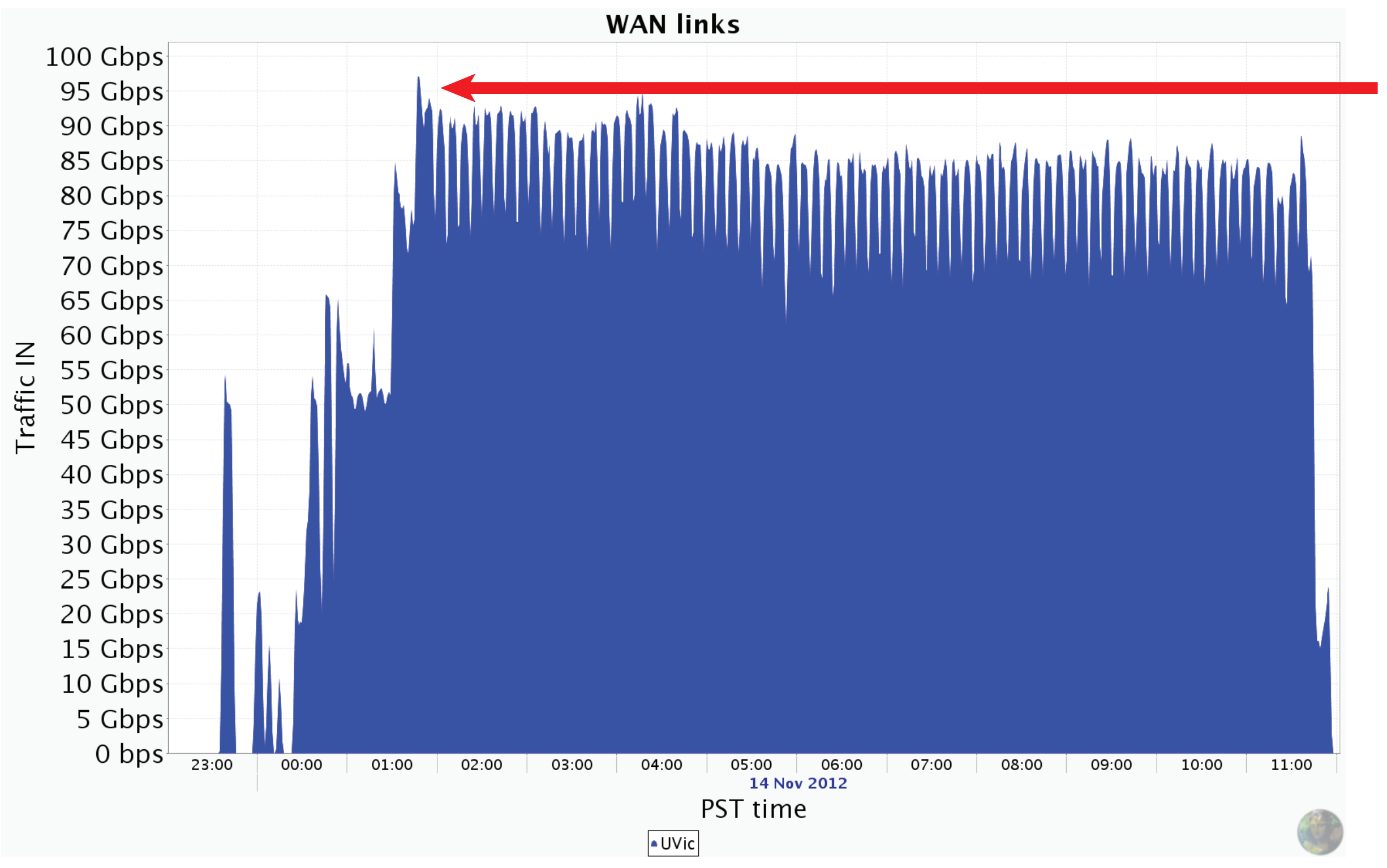
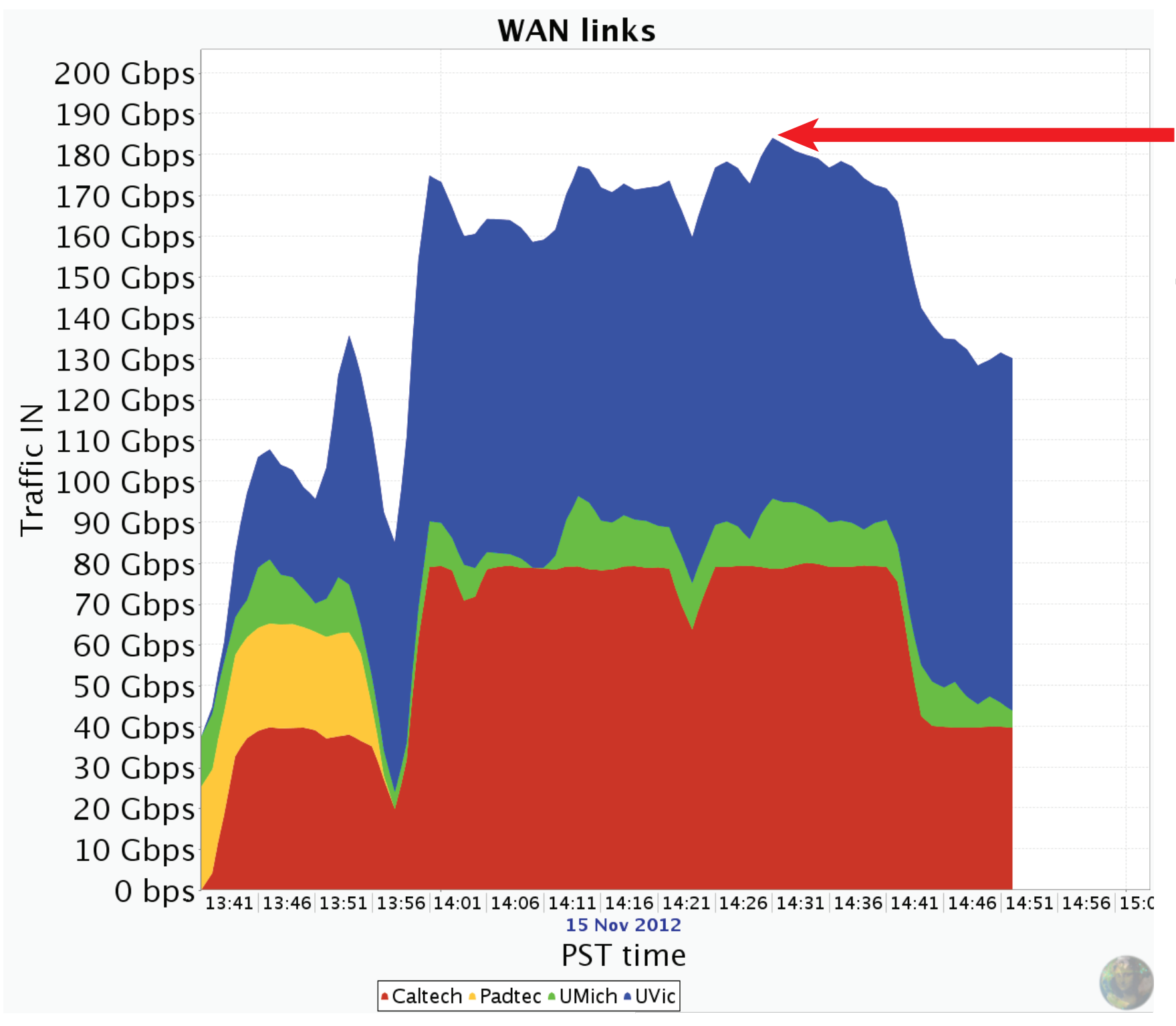


Figure 2. The end system configured at the Caltech conference booth on the SC exhibition floor. The system consists of a mix of SSD equipped 40 GE PCIe Gen 3 Servers and a DDN Lustre Storage Appliance.



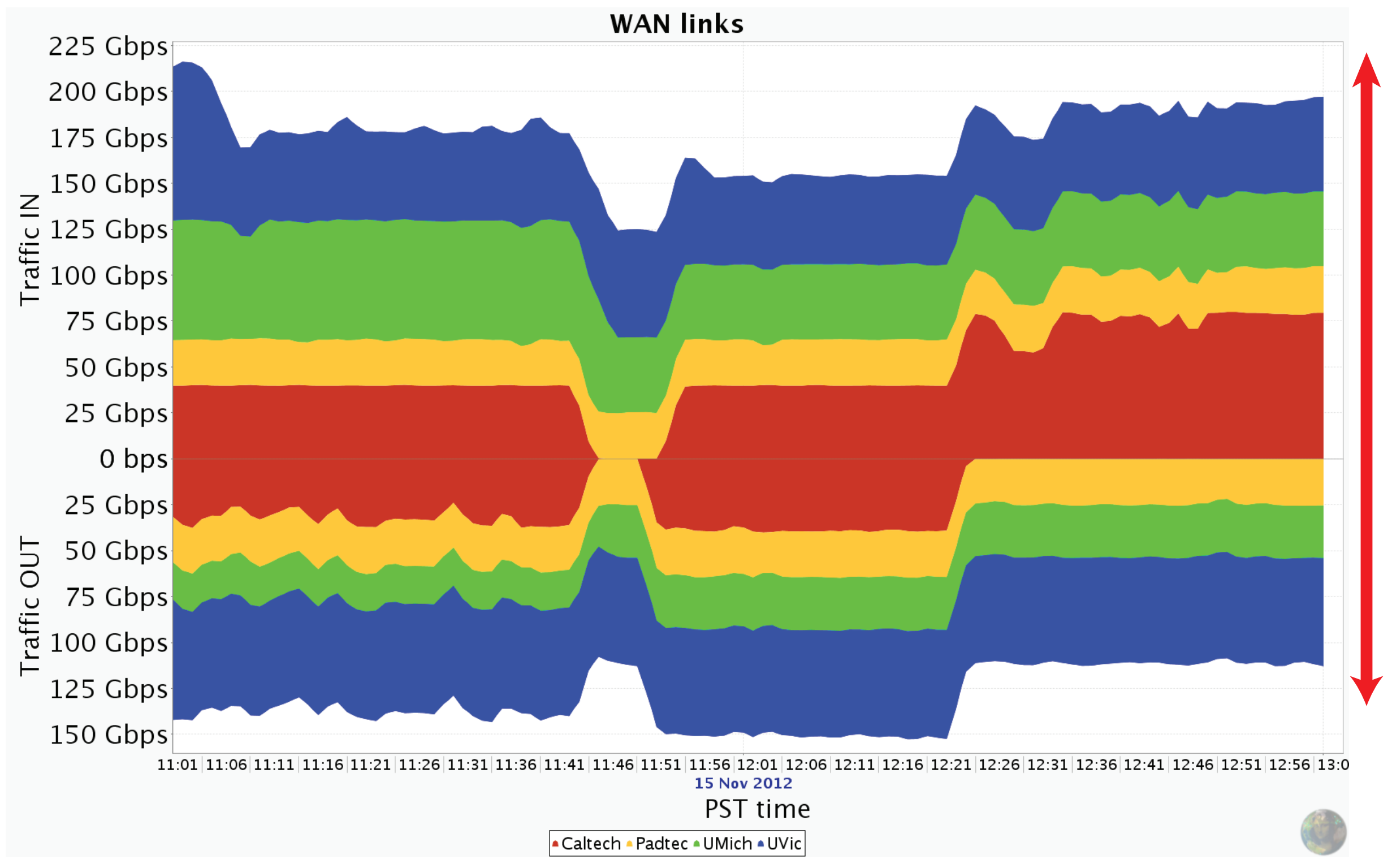
96 Gbps disk-to-disk

Figure 3. Four machines in Victoria with 16 SSDs and 40 GE network cards were used to transfer data to a combination of SSD servers and a Data Direct Networks Lustre Storage system on the exhibition floor. A peak disk-to-disk rate of 96 Gbps was achieved with a sustained rate near 85 Gbps.



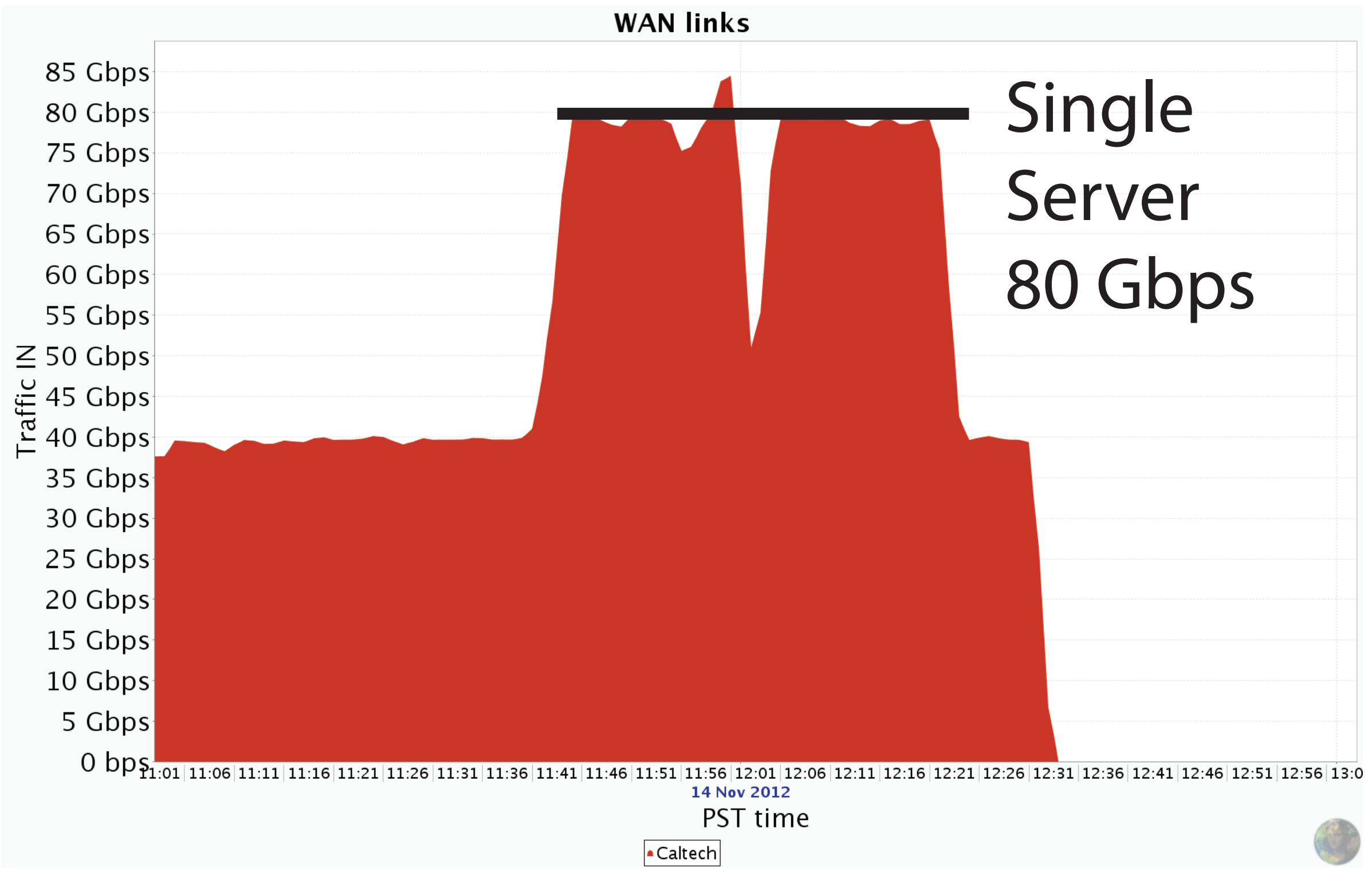
187 Gbps disk-to-disk

Figure 4. A total aggregate peak disk-to-disk throughput of 187 Gbps was achieved on Nov 15th. We were able achieve disk-to-disk what had only been achievable memory-to-memory at the SC 11 demonstration (186 Gbps).



339 Gbps mem-to-mem

Figure 5. A record total aggregate throughput between all sites of 339 Gbps was achieved on the last day of the conference.



Single Server 80 Gbps

Figure 6. The Caltech link was used to achieve an 80 Gbps transfer rate to a single server with two 40 GE interfaces at Salt Lake City with nearly 100% use of the servers' interfaces at both ends. The transfer was made using RDMA over converged Ethernet (RoCE) with a CPU load on the servers of only 5%.