Physicists Set New Record for Network Data Transfer

An international team of physicists, computer scientists, and network engineers led by the California Institute of Technology, CERN, and the University of Michigan and partners at the University of Florida and Vanderbilt, as well as participants from Brazil (Rio de Janeiro State University, UERJ, and the State Universities of São Paulo, USP and UNESP) and Korea (Kyungpook National University, KISTI) joined forces to set new records for sustained data transfer between storage systems during the SuperComputing 2006 (SC06) Bandwidth Challenge (BWC).

The high-energy physics team's demonstration of "High Speed Data Gathering, Distribution and Analysis for Physics Discoveries at the Large Hadron Collider" achieved a peak throughput of 17.77 gigabits per second (Gbps) between clusters of servers at the show floor and at Caltech. Following the rules set for the SC06 Bandwidth Challenge, the team used a single 10-Gbps link provided by National Lambda Rail (www.nlr.net) that carried data in both directions. Sustained throughput throughout the night prior to the bandwidth challenge exceeded 16 Gbps (or two gigabytes per second) using just 10 pairs of small servers sending data at nine Gbps to Caltech from Tampa, and eight pairs of servers sending seven Gbps of data in the reverse direction.

One of the key advances in this demonstration was Fast Data Transport (FDT; http://monalisa.cern.ch/FDT), a Java application developed by Iosif Legrand of Caltech that runs on all major platforms and uses the NIO libraries to achieve stable disk reads and writes coordinated with smooth data flow across the long-range network. FDT streams a large set of files across an open TCP socket, so that a large data set composed of thousands of files, as is typical in high-energy physics applications, can be sent or received at full speed, without the network transfer restarting between files. By combining FDT with FAST TCP, developed by Steven Low of Caltech's computer science department, together with an optimized Linux kernel provided by Shawn McKee of Michigan known as the "UltraLight kernel," the team reached unprecedented throughput levels, limited only by the speeds of the disks, that correspond to nine GBytes/sec reading from, or five Gbytes/sec writing to, a single rack of 40 low-cost servers.

Overall, this year's demonstration, following the team's record memory-to-memory transfer rate of 151 Gbps using 22 10-Gbps links last year at SuperComputing 2005, represents a major milestone in providing practical, widely deployable applications. These applications exploit advances in state-of-the-art TCP-based data transport, servers (Intel Woodcrest-based systems) and the Linux kernel over the last 12 months. FDT also represents a clear advance in basic data transport capability over wide-area networks compared to last year, in that 20 Gbps could be sustained in a few streams memory-to-memory over long distances very stably for many hours, using a single 10-Gigabit Ethernet link very close to full capacity in both directions.

The two largest physics collaborations at the LHC, CMS and ATLAS, each encompass more than 2,000 physicists and engineers from 170 universities and laboratories. In order to fully exploit the potential for scientific discoveries, the many Petabytes of data produced by the experiments will be processed, distributed, and analyzed using a global Grid. The key to discovery is the analysis phase, where individual physicists and small groups repeatedly access, and sometimes extract and transport, Terabyte-scale data samples on demand, in order to optimally select the rare "signals" of new physics from potentially overwhelming "backgrounds" from already-understood particle interactions. This data will amount to many tens of Petabytes in the early years of LHC operation, rising to the Exabyte range within the coming decade.

The high-energy physics team also carried out several other demonstrations, making good use of the ten wide-area network links connected to the Caltech/CERN booth: o Vanderbilt demonstrated the capabilities of LStore, an integrated system that provides a single file-system image across many storage "depots" consisting of compact data servers distributed across wide-area networks. Reading and writing between sets of servers at the Vanderbilt booth at SC06 and Caltech, the team achieved a throughput of more than one GByte/sec. o By using five of the 10 10-Gbps links coming into SC06, the team reached an aggregate throughput of more than 75 Gbps, combining disk-to-disk and memory-to-memory transfers. During this part of the demonstrations, the links between Tampa and Jacksonville, the National Lambda Rail Framenet links, and the newly commissioned Atlantic Wave link, were often loaded to full capacity at 10 Gbps in both directions, as shown on the SCInet network "weathermap." o Of particular note was the use of FDT between Tampa and Daegu in South Korea, allowing the group from Kyungpook National University and KISTI to achieve 8.6 Gbps disk-to-disk over a single network path, using NLR's shared Packetnet via Atlanta, and the GLORIAD link between Seattle and Daejeon that was inaugurated in September 2005, shortly before SC05.

Professor Harvey Newman of Caltech, head of the HEP team and US CMS Collaboration Board Chair, who originated the LHC Data Grid Hierarchy concept, said, "These demonstrations allowed us to thoroughly field-test a new class of data-transport applications, together with the real-time analysis of some of the data using `ROOTlets,' a distributed form of the ROOT system (root.cern.ch) that is an essential element of high-energy physicists' arsenal of tools for large-scale data analysis.

"These demonstrations provided a new, more agile and flexible view of the globally distributed Grid system of more than 100 laboratory- and university-based computing facilities that is now being commissioned in the U.S., Europe, Asia, and Latin America in preparation for the next generation of high-energy physics experiments at CERN's Large Hadron Collider (LHC) that will begin operation in November 2007, along with several hundred computing clusters serving individual groups of physicists. By substantially reducing the difficulty of transporting Terabyte- and larger scale data sets among the sites, we are enabling physicists throughout the world to have a much greater role in the next round of physics discoveries expected soon after the LHC starts."

David Foster, head of Communications and Networking at CERN said, "The efficient use of high-speed networks to transfer large data sets is an essential component of CERN's LCG plans to deploy computing infrastructure that will enable the LHC experiments to carry out their scientific missions. This demonstration of the high-speed transfer of physics event samples and their analysis made use of equipment at Tampa, Caltech, CERN, and elsewhere, interconnected by the same network infrastructure CERN plans to use in production, and was an important milestone on the road to ensuring full capability when the LHC starts operations in 2007."

Iosif Legrand, senior software and distributed system engineer at Caltech and the technical coordinator for the MonALISA and FDT projects, said, "We demonstrated a realistic, worldwide deployment for distributed, data-intensive applications capable to effectively use and coordinate the network resources. A distributed agent-based system was used for dynamic discovery of resources and to monitor, configure, control, and orchestrate efficient data transfer between several hundreds of computers using hybrid networks."

Richard Cavanaugh of the University of Florida, technical coordinator of the UltraLight project that is developing the next generation of network-integrated grids aimed at LHC data analysis, said, "Future optical networks incorporating multiple 10-Gbps links are the foundation of the Grid system that will drive scientific discoveries at the LHC. A 'hybrid' network integrating both traditional switching and routing of packets and dynamically constructed optical paths to support the largest data flows is a central part of the near-term future vision that the scientific community has adopted to meet the challenges of data-intensive science in many fields. "By demonstrating that many 10-Gbps wavelengths can be used efficiently over continental and transoceanic distances (often in both directions simultaneously), the high-energy physics team showed that this vision of a worldwide dynamic Grid supporting many Terabyte and larger data transactions is practical."

Shawn McKee, associate research scientist in the University of Michigan department of physics and leader of the UltraLight network technical group, said, "This achievement is an impressive example of what a focused network effort can accomplish. It is an important step towards the goal of delivering a highly capable end-to-end network-aware system and architecture that meet the needs of next-generation e-Science."

Paul Sheldon of Vanderbilt University, who leads the NSF-funded Research and Education Data Depot Network (REDDnet) project that will deploy a distributed storage infrastructure of about 700TB over the next two years, commented on the innovative network storage technology that helped the group achieve such high performance in wide-area, disk-to-disk transfers.

"With IBP and the logistical network technology that Micah Beck and his group at Tennessee have developed, we were able to build middleware, L-Store, that can exploit a tremendous amount of parallelism, both in data transfers across the network and in reading and writing to disk," said Sheldon. "And since L-Store can also do efficient erasure coding in software with minimal data movement, we can build high-quality storage clusters out of commodity parts and push depot costs down to a thousand dollars a TB.

"When you combine this network-storage technology, including its cost profile, with the remarkable tools that Harvey Newman's networking team has produced, I think we are well positioned to address the incredible infrastructure demands that the LHC experiments are going to make on our community worldwide."

The team hopes this new demonstration will encourage scientists and engineers in many sectors of society to develop and plan to deploy a new generation of revolutionary Internet applications. Multigigabit/s end-to-end network performance will empower scientists to form "virtual organizations" on a planetary scale, sharing their collective computing and data resources in a flexible way. In particular, this is vital for projects on the frontiers of science and engineering, in "data-intensive" fields such as particle physics, astronomy, bioinformatics, global climate modeling, geosciences, fusion, and neutron science.

The new bandwidth record was achieved through extensive use of the SCInet network infrastructure at SC06. The team used all 10 of the 10-Gbps links coming into the showfloor, connected to two Cisco Systems Catalyst 6500 Series Switches at the Caltech/CERN booth, together with computing clusters provided by Hewlett Packard and a large number of 10-gigabit Ethernet server interfaces provided by Neterion and Myricom.

The 10 10-Gbps network connections included two National Lambda Rail FrameNet links, one to Los Angeles (the official BWC wavelength) and one to StarLight two NLR PacketNet links used from Korea over GLORIAD and from the University of Michigan over MiLR; two links provided by Internet2's Abilene network used to carry traffic from Caltech and UMICH; one link from ESNET used from Brookhaven National Laboratory; and one link from FLRNET used to carry traffic from Brazil over the CHEPREO/WREN-LILA link. Also, one link provided by AtlanticWave from NYC to DC and Miami was used to carry traffic from CERN coming over the USLHCNet NYC-Geneva circuit and one link provided by UltraLight/FLR from Jacksonville.

The UltraLight/FLR circuit was the only direct WAN circuit available at SC06 and terminated directly on the Caltech equipment on the showfloor. All other circuits were connected through the SCInet infrastructure. During the test, several of the network links were shown to operate at full capacity for sustained periods. The network has been deployed through exceptional support by Cisco Systems and Nortel, as well as the network engineering staffs of National LambdaRail, Florida Lambda Rail, Internet2, ESnet, TeraGrid, CENIC, MiLR, Atlantic Wave, AMPATH, RNP and ANSP/FAPESP in Brazil, KISTI in Korea, the Starlight international peering point in Chicago, and MANLAN in New York.

As part of the SC06 demonstration, a distributed analysis of simulated LHC physics data was carried using the Grid-enabled Analysis Environment (GAE) developed at Caltech for the LHC. This demonstration involved the use of the Clarens Web Services portal developed at Caltech, the use of Root-based analysis software, and numerous architectural components developed in the framework of Caltech's "Grid Analysis Environment." The analysis made use of a new component in the Grid system: "Rootlets" hosted by Clarens servers. Each Rootlet is a full instantiation of CERN's Root tool, created on demand by the distributed clients in the Grid.

The design and deployment of the Rootlets/Clarens system was carried out under the auspices of an STTR grant for collaboration between Deep Web Technologies (www.deepwebtech.com) of New Mexico, Caltech, and Indiana University. In addition to the Rootlets/Clarens demonstration, an innovative literature and database aggregation search tool designed specifically for scientists working in the field of particle physics and developed by Deep Web Technologies was shown. This aggregation tool allowed simultaneous queries to be made on several of the most popular document databases, the results being aggregated and presented to the user in a homogeneous fashion. Deep Web's aggregation system also powers the science.gov website.

The team used Caltech's MonALISA (MONitoring Agents using a Large Integrated Services Architecture- http://monalisa.caltech.edu) system to monitor and display the real-time data for all the network links used in the demonstration. MonALISA is a Dynamic, Distributed Service System that is capable of collecting any type of information from different systems, to analyze it in near-real time, and to provide support for automated control decisions and global optimization of workflows in complex grid systems. It is currently used to monitor more than 300 sites, more than 50,000 computing nodes, and tens of thousands of concurrent jobs running on different grid systems and scientific communities.

MonALISA is a highly scalable set of autonomous, self-describing, agent-based subsystems which are able to collaborate and cooperate in performing a wide range of monitoring tasks for networks and Grid systems, as well as the scientific applications themselves. Vanderbilt demonstrated the capabilities of their L-Store middleware, a scalable, open source, and wide-area-capable form of storage virtualization that builds on the Internet Backplane Protocol (IBP) and logistical networking technology developed at the University of Tennessee. Offering both scalable metadata management and software-based fault tolerance, L-Store creates an integrated system that provides a single file-system image across many IBP storage "depots" distributed across wide-area and/or local-area networks. Reading and writing between sets of depots at the Vanderbilt booth at SC06 and Caltech in California, the team achieved a network throughput, disk to disk, of more than one GByte/sec. On the floor, the team was able to sustain throughputs of 3.5 GByte/sec between a rack of client computers and a rack of storage depots. These two racks communicated across SCinet via four 10-GigE connections.

The demonstration and the developments leading up to it were made possible through the strong support of the U.S. Department of Energy Office of Science and the National Science Foundation, in cooperation with the funding agencies of the international partners.

Further information about the demonstration may be found at: http://supercomputing.caltech.edu/

Source: Caltech