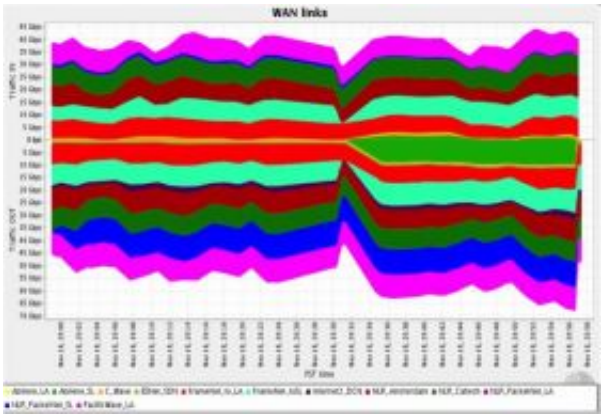
High Energy Physics Team Sets New Data-Transfer World Records

December 9 2008



A sample of the results obtained at the Caltech booth at SC08, monitored by MonALISA, flowing in and out of the servers at the booth. The general smoothness of the TCP flows on the individual links is the result of using FDT. The feature in the middle of the graph is the result of briefly losing the local session at SC08 driving some of the flows.

(PhysOrg.com) -- Building on seven years of record-breaking developments, an international team of physicists, computer scientists, and network engineers led by the California Institute of Technology--with partners from Michigan, Florida, Tennessee, Fermilab, Brookhaven, CERN, Brazil, Pakistan, Korea, and Estonia--set new records for sustained data transfer among storage systems during the SuperComputing 2008 (SC08) conference recently held in Austin, Texas.

Caltech's exhibit at SC08 by the High Energy Physics (HEP) group and the Center for Advanced Computing Research (CACR) demonstrated new applications and systems for globally distributed data analysis for the Large Hadron Collider (LHC) at CERN, along with Caltech's global monitoring system MonALISA (monalisa.caltech.edu) and its collaboration system EVO (Enabling Virtual Organizations; evo.caltech.edu), together with near real-time simulations of earthquakes in the Southern California region, experiences in time-domain astronomy with Google Sky, and recent results in multiphysics multiscale modeling. A highlight of the exhibit was the HEP team's record-breaking demonstration of storage-to-storage data transfers over wide area networks from a single rack of servers on the exhibit floor. The high-energy physics team's demonstration of "High Speed LHC Data Gathering, Distribution and Analysis Using Next Generation Networks" achieved a bidirectional peak throughput of 114 gigabits per second (Gbps) and a sustained data flow of more than 110 Gbps among clusters of servers on the show floor and at Caltech, Michigan, CERN (Geneva), Fermilab (Batavia), Brazil (Rio de Janiero, São Paulo), Korea (Daegu), Estonia, and locations in the US LHCNet network in Chicago, New York, Geneva, and Amsterdam.

Following up on the previous record transfer of more than 80 Gbps sustained among storage systems over continental and transoceanic distances in Reno, Nevada, at SC07, the team used a small fraction of the global LHC grid to sustain transfers at a total rate of 110 Gbps (114 Gbps peak) between the Tier1, Tier2, and Tier3 center facilities at the partners' sites and the Tier2-scale computing and storage facility constructed by the HEP and Caltech's Center for Advanced Computing Research team within two days on the exhibit floor. The team sustained rates of more than 40 Gbps in both directions for many hours (and up to 71 Gbps in one direction), showing that a well-designed and configured single rack of servers is now capable of saturating the highest-speed wide-area network links in production use today, which have a capacity

of 40 Gbps in each direction.

The overseas partners achieved excellent storage-to-storage results during the demonstrations: 3 Gbps (on two 1-Gbps links) with the Tier2 center in Tallinn, Estonia, and approaching 2 Gbps on two 1-Gbps links with the Tier2 Centers at UERJ (Rio) and SPRACE (São Paulo).

The record-setting demonstration was made possible through the use of 12 10-Gbps wide-area network links to SC08 provided by SCinet; National LambdaRail (6); Internet2 (3); ESnet; Pacific Wave; and the Cisco Research Wave, with onward connections provided by CENIC in California; the TransLight/StarLight link to Amsterdam; SURFNet (Netherlands) to Amsterdam and CERN; and CANARIE (Canada) to Amsterdam; as well as CENIC, Atlantic Wave and Florida LambdaRail to Gainesville and Miami; US Net to Chicago and Sunnyvale; Gloriad and KreoNet2 to Daegu in Korea; GEANT to Estonia; and the WHREN link, co-operated by FIU and the Brazilian RNP and ANSP networks, to reach the Tier2 centers in Rio and São Paulo.

Two fully populated Cisco 6500E series switch-routers, and more than 100 10-gigabit Ethernet (10GE) server interfaces provided by Myricom and Intel, as well as two fiber channel S2A9900 storage platforms provided by DataDirect Networks (DDN) equipped with 8-Gbps host bus adapters from QLogic, along with five X4500 and X4540 disk servers from Sun Microsystems, were used to set the new record. The computational nodes were 32 widely available dual-motherboard Supermicro servers housing 128 quad-core Xeon processors on 64 motherboards with a like number of 10-GE interfaces, as well as Seagate SATA II disks providing 128 terabytes of storage.

One of the key elements in this demonstration was Fast Data Transfer (monalisa.cern.ch/FDT), an open-source Java application based on TCP, developed by the Caltech team in close collaboration with the

Politehnica Bucharest team. Fast Data Transfer runs on all major platforms, and it achieves stable disk reads and writes coordinated with smooth data flow across long-range networks. The ability of FDT to sustain drive data flows at speeds reaching the capacity limits of the links, a full 10 Gbps, was shown repeatedly during the SC08 demonstrations. The FDT application works by streaming data 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, and without any packets being lost. FDT works with Caltech's MonALISA system to dynamically monitor the capability of the storage systems, as well as the network path, in real time, and sends data out to the network at a moderated rate that is matched to the capacity (measured in realtime) of long-range network paths.

FDT was combined with an optimized Linux kernel, provided by Shawn McKee of Michigan, known as the "UltraLight kernel," and the FAST TCP protocol stack developed by Steven Low, professor of computer science and electrical engineering at Caltech, to reach its unprecedented sustained throughput level of 14.3 Gigabytes/sec with a single rack of servers, limited by the speed of the disks.

MonALISA's ability to monitor a worldwide global ensemble of grids and networks, from the individual process in a single processing core to the major links to the overall network topology in real time, was shown throughout the conference, running (since 2002) around the clock to keep track of more than one million parameters at 350 sites on a large overhead global display. A second major milestone was achieved by the HEP team working together with Ciena, which had just completed its first OTU-4 (112 Gbps) standard link carrying a 100-Gbps payload (or 200 Gbps bidirectional) with forward error correction. The Caltech and Ciena teams used an optical fiber cable with 10 fiber-pairs linking their neighboring booths, Ciena's system to multiplex and demultiplex 10

10-Gbps links onto the single OTU-4 wavelength running on an 80-km fiber loop, and some of Caltech's nodes used in setting the wide-area network records together with FDT, to achieve full throughput over the new link.

Thanks to FDT's high-throughput capabilities and the error-free links between the booths, the teams were able to achieve a maximum of 199.90 Gbps bidirectionally (memory-to-memory) within minutes of the start of the test, and an average of 191 Gbps during a 12-hour period that logged the transmission of 1.02 petabytes overnight.

Before dismantling the exhibit at the end of the conference, Caltech and DDN worked together to quickly reach 69.3 Gbps over the fiber cable, limited by the disk speed and the kernel, reaching 92% of the full throughput capacity of the DDN platforms. The team expects to be able to reach 100% of the storage platforms' capacity with further kernel-tuning.

The two largest physics collaborations at the LHC, CMS and ATLAS, each encompassing more than 2,000 physicists and engineers from 170 universities and laboratories, are about to embark on a new round of exploration at the frontier of high energies. When the LHC accelerator and their experiments resume operation next spring, new ground will be broken in our understanding of the nature of matter and space-time and in the search for new particles. 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 of 150 computing and storage facilities located at laboratories and universities around the world.

The key to discovery is the analysis phase, where individual physicists and small groups located at sites around the world repeatedly access, and sometimes extract and transport multiterabyte data sets 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 SC08 HEP team hopes that the demonstrations at SC08 will pave the way towards more effective distribution and use for discoveries of the masses of LHC data.

Harvey Newman, Caltech professor of physics, head of the HEP team, and chair of the US LHC Users Organization's Executive Committee, originated the LHC Data Grid Hierarchy concept. "The record-setting demonstrations at SC08 have established our continued rapid progress, advancing the state of the art in computing and storage systems for data-intensive science, and keeping pace with the leading edge of long-range optical networks," said Newman. "By sharing our methods and tools with scientists in many fields, we hope that the research community will be well-positioned to further enable their discoveries, taking full advantage of current networks, as well as next-generation networks with much greater capacity as soon as they become available. In particular, we hope that these developments will afford physicists and young students throughout the world the opportunity to participate directly in the LHC program, and potentially to make important discoveries."

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 LHC Computing Grid (LCG) infrastructure that will enable the LHC experiments to carry out their scientific missions."

"We demonstrated a realistic, worldwide deployment of distributed, data-intensive applications capable of effectively using and coordinating high-performance networks," said Iosif Legrand, senior software and distributed system engineer at Caltech, the technical coordinator of the

MonALISA and FDT projects. "A distributed agent-based system was used to dynamically discover network and storage resources, and to monitor, control, and orchestrate efficient data transfers among hundreds of computers."

Richard Cavanaugh of the University of Illinois at Chicago, technical coordinator of the UltraLight project that is developing the next generation of network-integrated grids aimed at LHC data analysis, said, "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. By also demonstrating the full use of a 100-Gbps wavelength for the first time, we can now be confident that we will be ready to fully exploit the next generation of networks as the LHC ramps up to full luminosity over the next few years, through our continued developments in the NSF-funded UltraLight and PLaNetS projects."

"This achievement is an impressive example of what a focused network and storage system effort can accomplish," said Shawn McKee, research scientist in the University of Michigan department of physics and leader of the UltraLight network technical group. "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."

Artur Barczyk, network engineer and US LHCNet team leader with Caltech, said, "The impressive capability of setting up the many light paths used in this demonstration in such a short time frame, spanning three continents and providing guaranteed bandwidth channels for applications requiring them, together with the efficient use of the provisioned bandwidth by the data transfer applications, shows the high

potential in circuit network services. The light path setup among USLHCNet, Surfnet, CANARIE, TransLight/StarLight, ESnet SDN, and Internet2 DCN, and using the MANLAN, Starlight, and Netherlight exchange points, took only days to accomplish (minutes in the case of SDN and DCN dynamic circuits). It shows how the network can already today be used as a dedicated resource in data intensive research and other fields, and demonstrates how applications can make best use of this resource basically on demand."

Paul Sheldon of Vanderbilt University, who leads the NSF-funded Research and Education Data Depot Network (REDDnet) project that is deploying a distributed storage infrastructure, noted the innovative network storage technology that helped the group achieve such high performance in wide-area, disk-to-disk transfers. "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."

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

Provided by Caltech