LHC-T3 Science/Network Requirements Workshops

Over 68 U.S. universities and national laboratories are poised to begin receiving raw data generated at the CERN Large Hadron Collider (LHC), located in Geneva, Switzerland. In addition, these institutions will receive simulation data from various Tier 1 or Tier 2 sites supporting the physicists, post-docs, and students who are analyzing LHC data. One of the major questions facing this user community is “what impact will their science needs have on the U. S. network providers at the campus, regional, state, and national levels?” To answer this question, Internet2 and US-LHC community jointly held a series of regional workshops to begin a dialog between network operators and the scientists of the US-LHC community. This short report summarizes the outcome of those workshops.

Workshop Program and Attendees

Each workshop targeted a specific geographic region of the U.S. and a common agenda was used for the workshop series, allowing workshop organizers to build a national picture of US-LHC readiness. Invitations were sent to physicists and campus network operators at every Tier 3 (T3) institution in those regions. Invitations were also sent to national network operators and the regional network and GigaPoP operators. From the potential pool of attendees, an average of 20 individuals attended each workshop. In addition to the on-site participants, several remote sites and individuals joined each workshop by H.323 video.

Each workshop opened with presentations by representatives from the US-ATLAS and US-CMS communities that highlighted the different computing models being promoted by each group. These were followed by a brief discussion on the impact of these two different models. Next, a campus LHC physicist described their needs and expectations. A series of reports from the various network operators – campus, regional, and national – offered perspectives on how networks are evolving to combine traditional switched circuit and shared IP services together into a hybrid network.

During the second half of the workshop, representatives from the TeraPath and UltraLight projects made presentations on new services and features being developed for the LHC community. These were followed by a talk on advanced network diagnostic tools and services that physicists and network operators can use to effectively resolve problems when they occur. The final presentation was from a representative of the Open Science Grid (OSG) community, explaining how T3 sites could integrate their resources into the global LHC experiment. The workshop concluded with a guided discussion to identify open issues and to generate a roadmap for how US-LHC physical science will be accomplished.

Workshop Outcomes

The workshop series succeeded in bringing T3 physicists and network operators together for some detailed discussions on the demands LHC collaborators will place on networks. Attendees agreed that the workshops helped open the lines of communication between the network operators and the LHC community. The major recommendation coming out of these workshops was that more discussions would be worthwhile and a venue should be created to support this activity. While no unknown issues or challenges were defined during these workshops, a few points are worth highlighting:

Work Models: The ATLAS group is using a hierarchical model where Tier 2 (T2) and T3 sites will contact their local national Tier 1 (T1) center for data. If the T1 center does not have a local copy, it will automatically retrieve one for the client. The CMS model, on the other hand, is much flatter. In this model, each T2 or T3 site will use metadata to determine which T1 site holds the required data and that T2/T3 site will send a request directly to the remote T1 site. These differing models will have different impacts on the campus, regional, national, and international networks.

In the ATLAS model, all data moving to or from the U.S. goes through the Brookhaven National Laboratory (BNL) T1 site. This traffic can, and should, use the LHC-OPN (LHC Optical Private Network) to move between BNL, and CERN or foreign T1 sites. T2/T3 sites will use campus, regional, and national networks to get to BNL. Current IP routing policies will direct traffic heading toward BNL over the Internet2 Network or NLR's PacketNet to an ESnet peering point. Reverse traffic will flow over ESnet to one of the regional peering points before crossing back to the Internet2 Network or PacketNet on the way to the regional/campus network.

As noted above, the CMS model will require all T2/T3 sites to contact the appropriate T1 center directly to retrieve the required data. Thus, US-CMS sites will use all the networks described above, plus they will reach directly across national boundaries. At the same time, overseas T2/T3 sites will be accessing a large fraction of their total data needs from the US-CMS T1 site at Fermilab. This will require coordination between campus, state, regional, national, and international network operators. It is expected that existing IP peering relations between the various network operators will remain, meaning that multiple networks will be involved in carrying this LHC traffic. Both shared IP services and switched circuit services may be used to meet this demand. The impact of full-scale CMS operation on the available international links has not yet been measured.

User Expectations: Physicists typically have low expectations for network performance. When asked how long it should take to move a 50 MB file across the country, the common answer was on the order of 1 minute. This is off by 1 to 2 orders of magnitude, when using FastEthernet and Gigabit Ethernet based networks respectively. Getting the T3 physicists to expect good performance and giving them tools to help find and fix real problems will be a major task for campus network operators. Internet2 and Fermilab have ongoing programs of network performance improvement.

While much work has been done, both inside and outside the LHC community, to identify the sources of end-to-end performance problems, T3 physicists don't know the status of these projects. Computing staff at the T1 and T2 sites have been working on methods to tune hosts, networks, and disk subsystems. In addition, other research groups have been developing advanced tools that simplify the troubleshooting steps an end-user must take to find a real problem. However, the knowledge of these tools and methods has not been passed down to the T3 physicist. More needs to be done to widely distribute this knowledge throughout the T3 physicists' community.

In addition to educating the T3 physicist, the campus network operators must become involved in this support activity. The campus NOC staff should be able to answer simple questions and point users to a campus advanced diagnostic server. They should also be prepared to respond to trouble reports based on the diagnostic server's test results.

Another major question that arose was what kind of network access physicists can expect from their local campus. Is Gigabit Ethernet (GigE) service common or is it a novelty? How about 10 Gb/s service to a remote site? The workshops showed that campuses are actively working in this area. Some already provide GigE to every desktop, while other sites support limited deployment of this service. A common theme was that solving the last mile issue, from the campus to the connector and from the campus edge to the physics lab, was a serious problem that few campuses have solved. More work must be done in this area to share experiences and best practices. Finally, network security issues will also present a challenge to some campuses. Having a sample security document that can help campus physicists answer security-related questions and show that systems are hardened enough to prevent intrusions would help the entire LHC community.

Finally, there is often a disconnection between the services campus network operators provide and what the physicists use. The network operators believe that they have procedures and practices in place to generate and deal with trouble reports. The physicist does not know whom to contact when a problem occurs. Bringing these two groups together can help bridge this gap and will allow physicists to focus on science, not support issues.

<u>*Local vs. Remote Computing Facilities*</u>: The physicists agreed that some combination of local and remote computing facilities will be used throughout the experiments' lifetime. One of the major problems is getting code to compile and run on remote machines. Local libraries or customized routines can mean that code will not build on a remote cluster. Local processing or remote debugging are possible solutions to this problem. Concern was expressed over being able to identify and contact application and software library experts for help when those experts may be anywhere in the world.

Another reason for using local resources is that developers will want to monitor the operation of detector elements they have built. While not part of the computing models, there is a need to send a constant stream of monitoring data back to the developers to ensure that the detector is operating correctly.

A third reason for using local resources is that small groups will start new investigations using the LHC data. One such case is a University of Mississippi group that will begin exploring black hole formation at the LHC. This project will require some special processing of the LHC data and that processing may be best performed on local computer resources.

Remote resources are available at the T1 and T2 sites. Some analysis work will be performed there, but access issues must be addressed. Either the physicist will submit a grid job and the output n-tuples will be sent back to the local site or the physicist will need to remote login to the T2 site to view the results. It was also noted that distributed file systems are being developed and they may provide some use to the LHC community.

Network Utilization: The traditional method for forecasting network utilization requirements is to look at sustained data rates needed by the experimenter. For example, moving data at 3 Gb/s would allow a physicist to move about 1.25 TByte/hour. However, 3 Gb/s is not a rate that the experiment would need to sustain for days on end. A typical LHC dataset will be in the 1-5 TB range, so the entire 5 TB could require about 4 hours at this rate. Once the dataset was downloaded, the physicist would run one or more analysis passes over this data, each pass taking hours to weeks of processing time. After performing this analysis, a new dataset would be requested and the analysis process repeated. Thus, we would see a very bursty network utilization pattern, with short (hours) periods of download and long (days or weeks) periods of low use.

A naive view would be that the network is vastly under-utilized and a much smaller access link could be used without affecting the science demands. However, the workshop participants agreed that a slower transfer rate would have a terrible impact on the scientist. Not only does this longer transfer time mean more wasted time for the user, a longer transfer time increases the possibility of faults or errors causing the transfer to fail. The workshop participants agreed that network operators should understand how the network will be used by the LHC community, i.e., peak rates are more important than sustained rates.

Conclusion

This series of workshops started the process of bringing the US-LHC physics community together with the various network operators. The T3 site representatives received useful information about dataset transfers and preparing for production. As noted above, a conclusion at each workshop was that continuing these interactions would benefit both the LHC community and the various network operators. Internet2 and the US-LHC community will use the existing Internet2 HENP Special Interest Group (http://henp.internet2.edu/) venue to continue hosting these events.

Rich Carlson
Internet2 Liaison to US-LHC