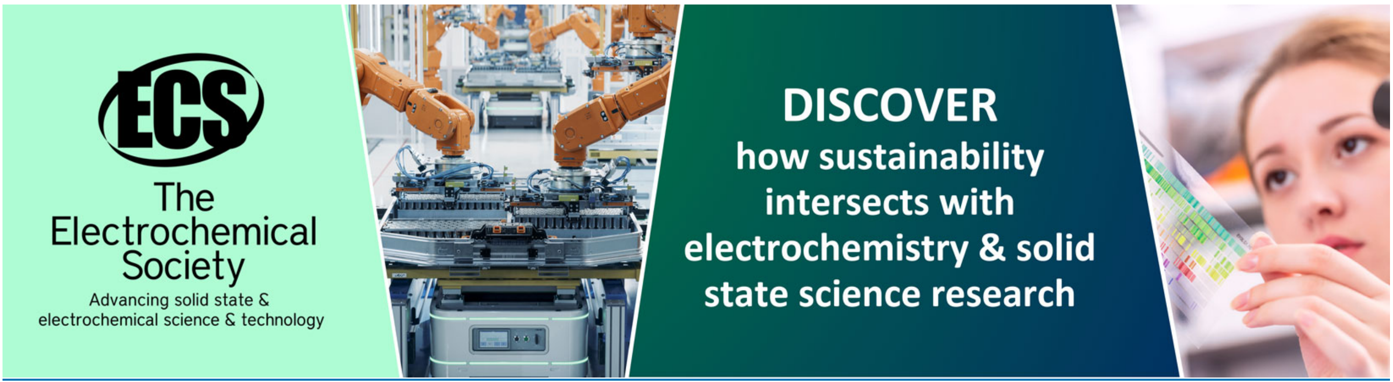
Journal of Physics: Conference Series

The IceCube Computing Infrastructure Model

To cite this article: M Merck and S Barnet 2012 *J. Phys.: Conf. Ser.* **396** 042007


ECS
The Electrochemical Society
Advancing solid state & electrochemical science & technology
DISCOVER
how sustainability intersects with electrochemistry & solid state science research

The IceCube Computing Infrastructure Model

M Merck and S Barnet

University of Wisconsin – Madison, Wisconsin IceCube Particle AstroPhysics Center,
222 W. Washington Ave Suite 500, Madison WI, 53705, USA

E-mail: barnet@wisc.edu

Abstract. In addition to the big LHC experiments, a number of mid-size experiments are coming online which need to define new computing models to meet the demands on processing and storage requirements of those experiments. We present the hybrid computing model of IceCube which leverages Grid models with a more flexible direct user model as an example of a possible solution. In IceCube a central data center at UW-Madison serves as a Tier-0 with a single Tier-1 at DESY Zeuthen.

1. Introduction

The IceCube Observatory is a kilometer-scale neutrino detector which uses the Antarctic ice sheet as a target for neutrinos and then detects the Cherenkov light resulting from particle collisions. The transparency of the ice sheet, where blue light has an absorption length of more than 100 meters, makes it an excellent medium for the detection of particle tracks, such as those from muons produced in some interactions.

The detector itself comprises 86 vertical strings, each containing 60 digital optical modules (DOMs) deployed in 2,500-meter-deep holes drilled in the ice by a hot water drill. The DOMs are photo-multiplier tubes (PMTs) with supporting electronics to digitize the waveforms and communicate with surface systems. The water in the hole refreezes producing optical contact between PMTs and ice. The 80 strings in the baseline IceCube design are deployed on a 125-meter grid covering 1 km^2 on the surface. DOMs are attached to the strings every 17 meters at depths between 1,450 and 2,450 meters. The baseline design detects muon neutrinos at energies down to about 100 GeV.

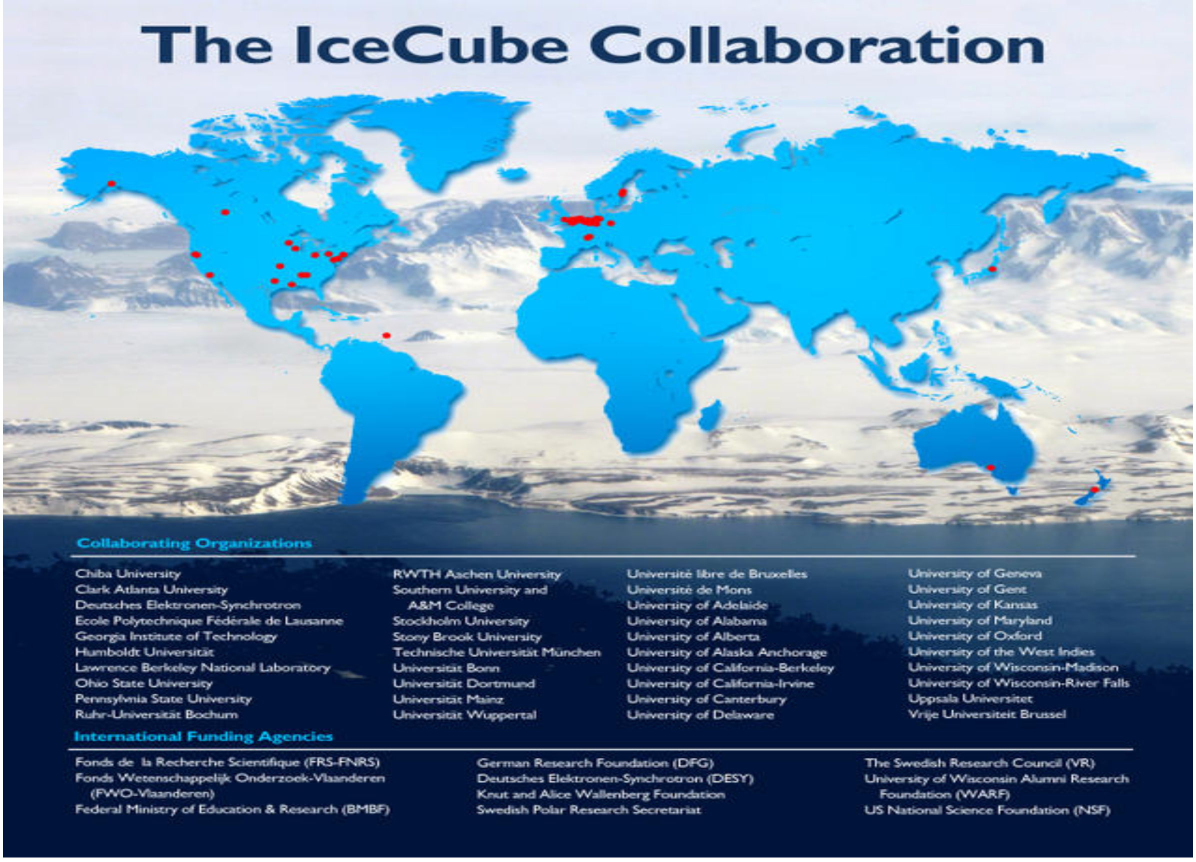
Another six strings called “DeepCore” are deployed on a more dense, 72-meter triangular grid. The DeepCore strings have 50 DOMs with 7-meter spacing at the bottom of the string, and 10 DOMs higher up serve as a veto for muons produced in the atmosphere by cosmic rays. DeepCore uses newer PMTs with higher quantum efficiency than the IceCube DOMs. The dense spacing and more efficient PMTs give DeepCore a lower energy threshold than IceCube, possibly as low as 10 GeV.

In addition to the buried DOMs, the IceCube Observatory includes a surface air shower array known as IceTop. IceTop consists of 160 ice-filled tanks, each instrumented with two IceCube DOMs. Two tanks are deployed about 10 meters apart near the top of each baseline string. IceTop detects cosmic-ray air showers with an energy threshold of about 300 TeV and will be used to study the cosmic-ray flux and composition. Further details about the IceCube instrumentation and physics goals can be found in the paper IceCube: An Instrument for Neutrino Astronomy [1].

Each of these detectors connects via surface cables to computing systems in the IceCube Lab (ICL). Custom-built computers and readout electronics collect the data from each string. The data acquisition system (DAQ) is a software based system which reads the waveforms from the DOMs, collects them and defines events based on a trigger condition. The events are passed to the online computing system for track reconstruction and filtering. The completed detector has an average event rate of 3000 Hz with some seasonal variation. That translates to approximately 1 TB/day of raw data which is written to tape and then further reduced to 100 GB/day to fit within the satellite bandwidth available to the South Pole station. The data is transferred to the Tier 0 data center at the University of Wisconsin – Madison where it is processed and then delivered to collaborating institutions.

2. The IceCube Collaboration

The IceCube collaboration comprises 39 institutions in 11 countries supported by the NSF and several European funding agencies. Collaborators support the collaboration by providing direct financial support in some cases, and by providing a variety of in-kind contributions typically in the form of monitoring shifts, simulation production, and writing code for data processing and analysis frameworks. Within the collaboration, working groups conduct specific analyses of interest to a variety of communities. Twice a year there is a collaboration wide meeting hosted by collaborating institutes on a rotating basis.



The IceCube Collaboration

Collaborating Organizations

Chiba University
Clark Atlanta University
Deutsches Elektronen-Synchrotron
Ecole Polytechnique Fédérale de Lausanne
Georgia Institute of Technology
Humboldt Universität
Lawrence Berkeley National Laboratory
Ohio State University
Pennsylvnia State University
Ruhr-Universität Bochum
RWTH Aachen University
Southern University and
A&M College
Stockholm University
Stony Brook University
Technische Universität München
Universität Bonn
Universität Dortmund
Universität Mainz
Universität Wuppertal
Université libre de Bruxelles
Université de Mons
University of Adelaide
University of Alabama
University of Alberta
University of Alaska Anchorage
University of California-Berkeley
University of California-Irvine
University of Canterbury
University of Delaware
University of Geneva
University of Gent
University of Kansas
University of Maryland
University of Oxford
University of the West Indies
University of Wisconsin-Madison
University of Wisconsin-River Falls
Uppsala Universitet
Vrije Universiteit Brussel

International Funding Agencies

Fonds de la Recherche Scientifique (FRS-FNRS)
Fonds Wetenschappelijk Onderzoek-Vlaanderen
(FWO-Vlaanderen)
Federal Ministry of Education & Research (BMBF)
German Research Foundation (DFG)
Deutsches Elektronen-Synchrotron (DESY)
Knut and Alice Wallenberg Foundation
Swedish Polar Research Secretariat
The Swedish Research Council (VR)
University of Wisconsin Alumni Research
Foundation (WARF)
US National Science Foundation (NSF)

Figure 1. The IceCube Collaboration Institutions

The IceCube collaboration is responsible for maintenance and operation of the detector. This includes operation of the detector at the South Pole as well as providing effort and coordination of collaboration members who develop new triggers or other detector enhancements. It also includes collecting and storing all of the experimental data as well providing the computing necessary to process data to a level that is science ready.

The Wisconsin IceCube Particle Astrophysics Center, a research center at the University of Wisconsin- Madison, acts as the lead institution for the collaboration and is responsible for storing the experimental data and processing the data to a science ready level. A current list of the IceCube collaboration institutes can be found at the IceCube web site [2].

3. Data Handling

The UW-Madison IceCube data center is physically distributed in several facilities to meet space, power, and cooling requirements. This is not optimal from the standpoint of manageability, but does provide some degree of redundancy and disaster recovery capability.

The facilities house the IceCube direct access storage and computing systems. The computing system is a typical compute cluster of 128 nodes (1284 cores) of commodity Intel x86_64 servers providing performance of 16,000 HEPSpec06. These servers are in 1U or blade form factors. The network is a standard 1Gb/s switched Ethernet network with a 10Gb/s backbone. Due to its flexibility in managing heterogeneous pools of computing resources, we chose Condor [3] as the workload management system for the cluster.

The compute system has direct access to cluster filesystems built on the Lustre filesystem [4]. There are four such filesystems holding experimental data, simulation data, analysis data sets, and working space for individual users. The largest of the filesystems are the experimental and simulation filesystems with sizes of 1 PB, and 1.2 PB respectively. The analysis and user spaces are intended to hold reduced data sets (Level 3) or individual analyses and are much smaller with analysis sized at 250 TB and user space at 90 TB. Administrative and support infrastructure is separated from the computing system (see Figure 2).

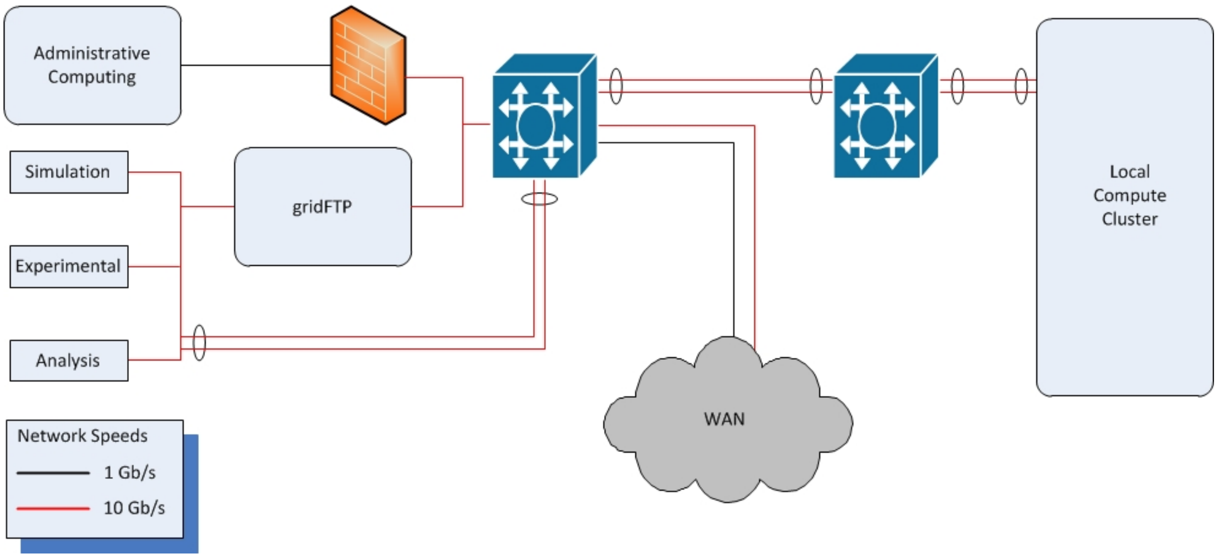


Figure 2. A high level overview of the IceCube computing infrastructure. The local computing resources and storage systems are located in separate facilities connected by a 10Gb/s network.

The experimental and simulation filesystems comprise several generations of commodity SATA disk arrays with Fibre Channel connectivity. This combination provides high density storage with acceptable price/performance characteristics. The analysis and user filesystems represent a new direction in that they are based on commodity servers populated with internal disk as the basic unit of storage. These units are connected to a 10Gb/s Ethernet fabric.

In addition to our facilities at UW-Madison, our collaborators at DESY-Zeuthen operate our only Tier 1 data center. This center provides significant storage space and computing capacity and is particularly helpful in avoiding the trans-Atlantic network hop for our European collaborators. In addition, we replicate our Level 2 data sets there for disaster recovery purposes (see Figure 3).

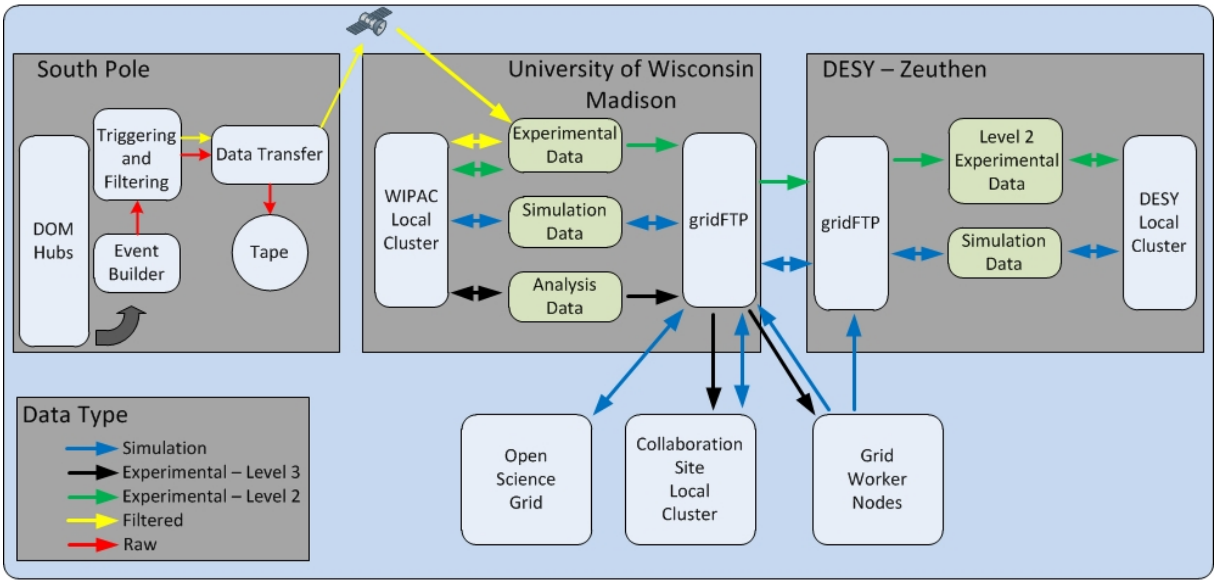


Figure 3. The flow of data from the detector through the IceCube Tier-0 and Tier-1 data centers

The available resources are distributed throughout the collaboration and vary somewhat over time. However, Table 1 provides a good approximation of the resources available to the collaboration. In most cases, these systems are part of shared computing clusters at the institution. Computing resources are reported in cores since not all sites have run HEPSpec benchmarks and so expressing resources available in those terms is not possible. The reported resources represent the maximum available resources. In practice, we have been able to use up to 6000 cores and on average, utilize about 4000 cores. Storage at the grid and collaboration sites is only temporary working storage, so storage is only reported for the Tier-0 and Tier-1 sites.

Site	Computing cores	Storage
UW-Madison – Tier 0	1500	2.5 PB
DESY Zeuthen – Tier 1	1000	750 TB
Collaboration – batch resources	8000	
Collaboration – grid resources	6200	

Table 1. Summary of resources provided to the IceCube collaboration

4. IceCube Hybrid Computing Model

The construction of IceCube began in the 2004-5 season. At that time, cluster computing was a well understood and widely deployed solution for working with large scale data sets. In addition, grid technology was emerging as a mechanism for handling computation at the LHC scale. This timing lead the IceCube project naturally toward what we now call the hybrid model.

In our hybrid model, we do not depend entirely on local computing resources or grid resources, but rather direct our workloads to the appropriate resources as needed. Local resources provide a high level of control over the hardware and software platform which provides the ability to respond quickly to changing needs. This was particularly useful during the early phases of construction. It is also a useful training ground in distributed computing and prepares our scientists for working with grid computing. The grid environment provides less control over the environment and is generally more complex than a compute cluster, but provides vastly more resources. Use of grid computing has proven far more cost effective than creating local resources of similar capacity.

4.1. Grids for Bulk Computing

The grid environment has proven an excellent tool for the production of simulation data. In particular, it was a natural fit for our simulation production team which was able to leverage earlier efforts related to the IceCube predecessor experiment AMANDA. Those efforts centered around submitting jobs to the local computing clusters at collaborating institutions, collecting the output, and tracking the results of individual simulation jobs. As grid technology matured, many aspects of job submission were simplified, and so more resources became available via grid mechanisms. This enables more effort to be directed to tracking jobs and data sets and less on the mechanics of integrating multiple clusters.

The current version of the simulation production system is a Python framework (called IceProd) which provides a convenient interface to end users and manages simulation work flows. The web based front end allows physicists to request the generation of a data set and specify many of the simulation parameters. Further fine tuning is possible by modifying the text based configuration files which are ultimately used by the simulation software. This is more complex but offers fine grained control for those who need it.

The backend system manages the actual production of the data set. At the core of this system is a database which tracks jobs and the files they produce. This enables easy determination of the status of a data set and ensures that complete data sets are produced. The framework divides the data set into jobs, and submits those jobs to conventional workload management systems and registers any assigned job identifiers in the database. A plugin architecture enables IceProd to submit jobs to a variety of workload management systems and enables the easy addition of new workload management systems as they become available. It is the responsibility of workload management systems to ensure that the jobs are scheduled and executed, and also to report error conditions in the event of unrecoverable failures. When a job completes, it transfers data back to the appropriate location in either the Tier-0 or Tier-1 data center using gridFTP. When the last job completes, the data set is marked complete and made available for use.

4.2. Local Cluster for Experimental Data Processing

For handling experimental data we use local computing resources. This is a multi-level process similar to many other high energy physics experiments. The data is reformatted and offline event reconstructions are performed. This results in a science ready data set which serves as a common starting point for analysis. Occasionally, problems are identified in the Level 2 data sets and so the experimental data must be reprocessed. This processing typically requires reading an entire data set of 40-50 TB and is most efficiently handled by direct access to our Lustre filesystems. This processing demand typically accounts for 40 to 60 percent of local cluster usage.

The remaining time is used either for high priority analysis or simulation production. From the Level 2 data set, individual working groups construct their own Level 3 data sets based on filters and cuts appropriate to their investigations. This filtering typically benefits from the direct, high speed access to our Lustre filesystems. In addition, there is a significant data reduction at this level and so it is easier to copy this data set to another institution if desired. And as always there are publication, conference, and thesis deadlines to consider, so control over workload prioritization as well as platform is a significant advantage.

The Lustre filesystems have performed well for us. Our normal workloads tend to be CPU bound; an individual job reads a data file, computes for the majority of its life, and writes a much smaller result back to the storage system. Unless many jobs start simultaneously, the load on the storage systems is spread out over time and is easily handled with 10 Gb/s Ethernet. A typical example is shown in Figure 4.

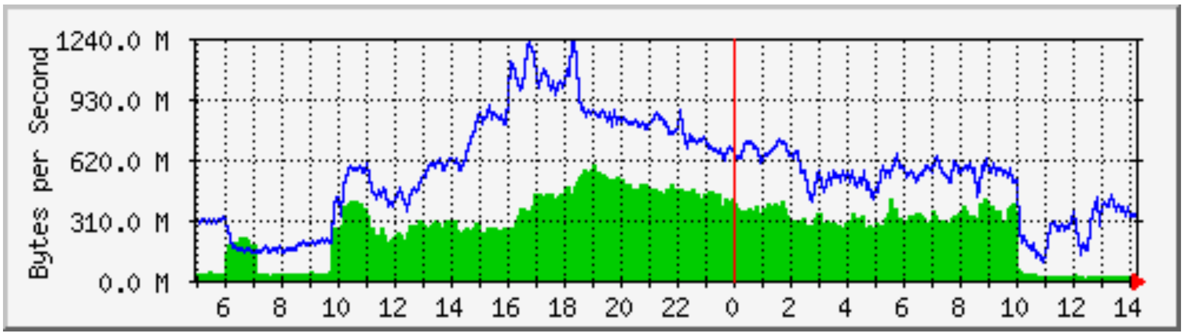


Figure 4. Typical Lustre network traffic carried over 10 Gb/s Ethernet

Occasionally there are job mixes which reverse this pattern. An example is event selection which involves scanning large portions of the Level 2 data set looking for events which are relevant to a particular analysis. These jobs are typically dominated by the time it takes to read files from the Lustre system and generate a heavy network load. Figure 5 shows an example of an event selection workload which saturates the 10 Gb/s network for approximately eight hours. During these times of heavy load, the system has remained stable.

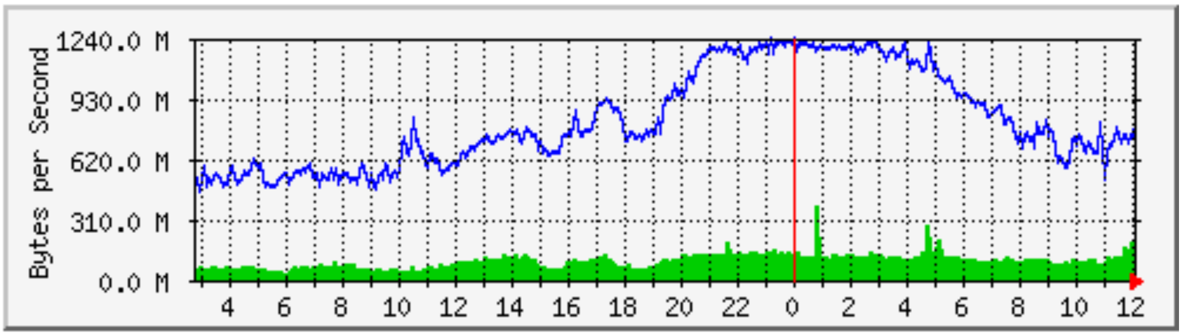


Figure 5. An event selection workload saturating the 10 Gb/s Ethernet

Upcoming improvements to our network topology will remove several significant bottlenecks and should enable us to saturate both 10 Gb/s links which connect the storage and computing systems.

5. Lessons Learned

Successfully running a hybrid computing system which supports an international collaboration requires a combination of technical and organizational tools. The local computing resources need to deliver sufficient power and storage to meet production and high priority processing requirements. Supporting grid usage requires very good external connectivity and a good understanding of the grid ecosystem.

In the technical area, our experience has been that traditional cluster computing systems on the order of 1-2% of the LHC experiments can handle an experiment of this size. Parallel or cluster filesystems are complicated, but provide the performance and capacity to deal with significant quantities of data in a reasonably cost effective manner. This sort of tightly coupled system performs well and is simple to use.

Using grid systems naturally shifts the workloads to external sites and so data movement becomes the dominant issue. In addition to local network topology issues, wide area networking becomes important to the overall performance of the storage system. As a further complication, the size of the running workload can vary quickly and assume a much larger scale than local resources. To deal with this, we have worked to ensure adequate capacity to handle our typical workloads, tuned the workload management framework to limit job submission to available capacity, and separated the grid systems from the other parts of our system to ensure that the network and disk traffic generated by those workloads does not adversely impact other systems.

Organizationally, flexibility is crucial. The University of Wisconsin provides considerable support and flexibility to individual research projects. This has included adding network capacity both on campus as well as to major research networks such as ESNet [5] and Internet 2 [6] as well as the ability to design and deploy IT and computing systems distinct from the central IT infrastructure. This has enabled us to establish the open access policies and mechanisms necessary to support a multi-institutional collaboration without requiring modifications to central IT systems, processes, or policies. In particular, we grant direct access to Tier-0 resources to IceCube collaboration members who otherwise have no affiliation with the University of Wisconsin. This direct access is typically in the form of shell accounts, grid mechanisms (based on X.509 certificates), or some combination of the two. We also retain control of our local network systems, enabling us change our local topology as needed to eliminate performance bottlenecks or change security controls to enable the deployment of new protocols. This is particularly important given the impact of firewalls and other security systems on network performance.

6. Summary

The IceCube detector opens the door to exciting new science in astro-particle physics. To support this science, we have adopted a system which combines the advantages of grid computing and local cluster computing. This approach has proven successful and provides a solid foundation for the successful collection and analysis of the IceCube detector data. Cloud computing shows promise in dealing with future challenges. Using cloud services as flash computing capacity is attractive, but the economics of data storage and transfer are not clear. Emerging technologies for creating private clouds such as OpenStack [7] provide new opportunities, particularly in enabling a variety of execution environments to share a common computing infrastructure. Currently, we are in the early stages of our evaluation of cloud computing systems, but anticipate developing a working pilot system within the year. If that pilot is successful, we anticipate that this will be a valuable addition to our existing capabilities.

References

[1] F. Halzen and S. Klein, arXiv:1007.1247 [astro-ph.HE]
[2] The IceCube Collaboration: http://icecube.wisc.edu/collaboration/collaborators
[3] Condor: http://research.cs.wisc.edu/condor/
[4] Lustre: http://www.whamcloud.com/
[5] ESnet: http://www.es.net/
[6] Internet 2: http://www.internet2.edu/
[7] OpenStack: http://openstack.org/