Enabling high throughput in widely distributed data management and analysis systems: Lessons from the LHC

William E. Johnston
ESnet, Lawrence Berkeley National Laboratory
e-mail: wej@es.net
Eli Dart
ESnet, Lawrence Berkeley National Laboratory
e-mail: dart@es.net
Michael Ernst
RHIC and ATLAS Computing Facility, Brookhaven National Laboratory
e-mail: mernst@bnl.gov
Brian Tierney
ESnet, Lawrence Berkeley National Laboratory
e-mail: bltierney@es.net

Paper type

Technical paper

Abstract

Today's large-scale science projects all involve world-wide collaborations that must routinely move 10s of petabytes per year between international sites in order to be successful. This is true for the two largest experiments at the Large Hadron Collider (LHC) at CERN – ATLAS and CMS – and for the climate science community. In the near future, experiments like Belle-II at the KEK accelerator in Japan, the genome science community, the Square Kilometer Array radio telescope, and ITER, the international fusion energy experiment, will all involve comparable data movement in order to accomplish their science.

The capabilities required to support this scale of data movement involve hardware and software developments at all levels: Fiber-optic signal transport, layer 2 transport, data transport (TCP is still the norm), operating system evolution, data movement and management techniques and software, and increasing sophistication in the distributed science applications. Further, ESnet's years of science requirements gathering indicates that these issues hold true across essentially all science disciplines that rely on the network for significant data transfer, even if the data quantities are modest compared to projects at the scale of the LHC experiments.

This paper will provide some of the context and state-of-the-art of each of the topics that experience has shown are essential to enable large scale "data-intensive" science.

Keywords

high performance networking, distributed application use of networks, network infrastructure to support international science collaboration

1 The origins and common problems of data-intensive science

Some science disciplines – driven by the increasing capability and sophistication of instrumentation are, or soon will be, "drowning" in data.

The high energy physics (HEP) community has dealt with this problem methodically for a relatively long time. Their efforts have culminated in the data handling systems for the experiments at the Large Hadron Collider (LHC) [1] at CERN in Switzerland, notably ATLAS [2] and CMS [3]. More high energy physics experiments are on the way, with Belle-II at KEK (Tsukuba, Japan) [4] coming on-line in a few years, and the International Linear Collider is scheduled to be operational by 2026, both of which generate more data than the LHC.

The LHC detectors are, in effect, 100,000,000 "pixel", three dimensional cameras operating at 40,000,000 "frames" per second. These detectors observe what happens when two particle beams, traveling in opposite directions at very nearly the speed of light, collide inside of the detector. The goal is to identify the sub-atomic particles, and their properties, which make up neutrons and protons.

The filtered stream of data the physicists analyze from ATLAS and CMS is of the order of 25 gigabits per second (Gb/s), steady state, per detector. This data is divided up and distributed to national LHC data centers ("Tier 1" data centers) around the world. The physicists working on the science pull the data from the Tier 1 centers into their computing clusters at universities ("Tier 2" centers) for analysis.

The management of this data – from CERN (Tier 0) to Tier 1 data center to Tier 2 analysis centers – represents the largest data management problem that any science community has ever faced.

In an example from another field of science, the Square Kilometer Array – SKA – is a radio telescope consisting of several thousand antennae spread over a million sq. km. which operate as a single instrument. [5]

The SKA is designed to explore by direct observation the very earliest stages of the evolution of the universe, identifying a set of pulsars which can be used to study the nature of space and time. Commensurate with its scale, the collaboration associated with the SKA involves some 70 institutions in 20 countries. The data rates from the individual antenna results in a total data flow, to a correlator / data processor of up to 9 Pb/s. This is reduced to the science data product which is a 100 Gb/s steady-state data flow that must be transported to the collaborators spread around the world. Finally, the transport of the 100 Gb/s data product around the world presents some interesting higher level problems. See [6].

Distributed science communities in all of these fields share the characteristic that the value of their data increases substantially when researchers can access all of it for both breadth-first and depth-first approaches to analysis. In some cases, like the LHC and probably the SKA, this data must be dispersed from the source to multiple data centers in order to make the data management problem tractable. In other cases (e.g. genomics), data must be brought together from widely dispersed sources into a framework that can be accessed and analyzed as a whole.

2 LHC as prototype for other large-scale science data problems

In the decade that the LHC HEP community has been developing, perfecting, and tuning the data management systems, data movement systems, and the intervening networks to manage hundreds of terabytes per day, a great deal has been learned that will be of use to other science communities that are facing, or will face, the same scale of data management.

2.1 Issues common to most science disciplines

In the ESnet network requirements determination process [7], certain issues are seen across essentially all science disciplines that rely on the network for significant data transfer or remote control, even if the quantities are modest compared to projects like the LHC experiments.

The common issues that have been identified, and the state-of-the-art that currently exists that enables the LHC data handling, are characterized in the following topics.

1) High capacity networks to enable the predictable, reliable movement of data to globally distributed collaborators

- State-of-the-art: Network routing, switching, and transport that today entails "coherent optical" technology providing 100 Gb/s optical channels over long distance fiber, with 80 to 100 such channels per fiber, and 100 Gb/s switches and routers.

2) Widely used data transport protocols must operate at 100 Gb/s.

- State-of-the-art: TCP – still the dominate data transport protocol – is a "fragile workhorse," and will perform well on high speed, long-haul networks only when the network is error-free, and the TCP stack appropriately tuned.

2a) Monitoring and testing of the network must detect errors and facilitate their isolation and correction.

- State-of-the-art: The perfSONAR network monitoring toolkit is widely deployed in research and education (R&E) networks and is designed for end-to-end (application-to-application) federated monitoring of the intervening network paths.

3) It must be possible to move data from the application through the operating systems and onto the network at network speeds.

- State-of-the-art: The operating systems of the platforms used for large-scale data management and analysis applications have modern, high-performance network stacks and provide support for parallelism in data transfer tools in order to drive the network at full speed.

4) In order to move and operate on the hundreds of terabytes a day of data that is involved in data-intensive science, applications must be designed and implemented so that all resources are kept operating at the highest possible efficiency.

- State-of-the-art: Automated management of data movement and error recovery by the application now achieve sustained, large-scale data transfers.

5) In order for data-intensive science experiments to operate effectively they must have access to the network resources deployed by the R&E networking community, and those resources must be structured in such a way as to be able to provide high-speed data throughput end-to-end.

- State-of-the-art: New network architectures and services have been built to interconnect the data-intensive science collaborations of the world's research universities and laboratories.

5a) Dedicated, purpose-built network infrastructure is needed to support quasi real-time data transfers, e.g. from instrument to data center

5b) A multi-domain, end-to-end virtual circuit network service is needed to provide guarantees for bandwidth sensitive applications/tasks.

5c) The typical campus LAN interface to the Wide Area Network (WAN) must be redesigned to support large, long distance data flows.

5d) Large-scale science traffic in the shared R&E network infrastructure must be explicitly managed in order to provide good service to science in ways that will not disrupt other uses of the infrastructure.

6) It is essential that the knowledge gained in the network, operating systems, and computer science research communities about how to optimize various aspects of the end-to-end data transfer problem be gotten into the hands of the science application developers and operators.

- State-of-the-art: The knowledge/experience that is built up by different segments of the community must be actively shared in order to have a broad impact. Currently it appears that a Web-based knowledge base is one effective way to do this.

Some of these topics are retrospective, and while still evolving are typically doing so slowly. Some are very new and are evolving very quickly. For example coherent optical network technology, high-throughput application design, and architectures for science-centered networks are still evolving rapidly. The remainder of this paper briefly considers the state-of-the-art for each of the topics listed above.

3 Continuous evolution of network technology

At the core of the ability of networks to transport the volume of data that is generated by science collaborations today are advances in optical transport technology and router technology.

Science data traffic volumes roughly track the size of science data sets, and exponential increase in the size of data sets (Figure 1) is closely mirrored in the observed R&E network traffic (Figure 2).

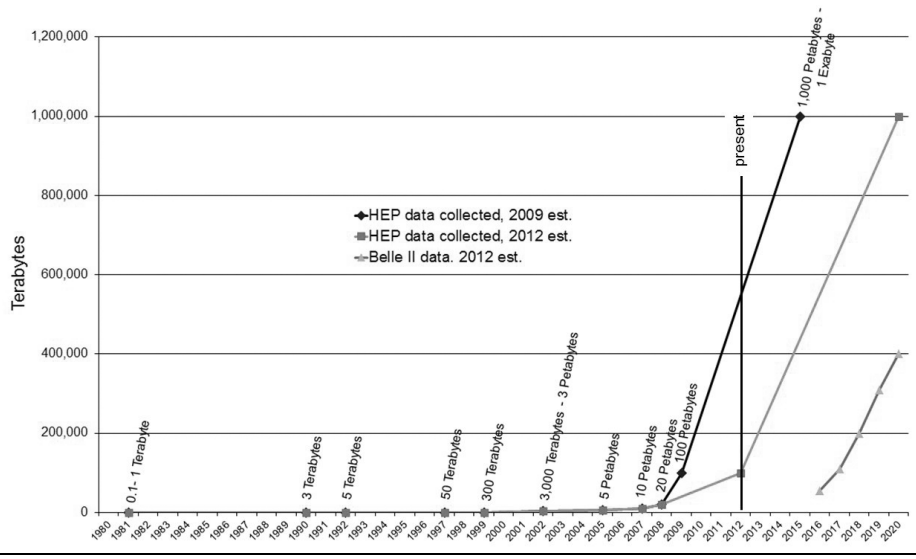

Figure 1. Historical and projected HEP data volumes (Terabytes) for leading experiments with Belle-II estimates shown separately. (Data courtesy of Harvey Newman, Caltech, and Richard Mount, SLAC and Belle II CHEP 2012 presentation.)

ESnet [8], one of the two largest R&E networks in the U.S., is used in the examples in this paper; however, to be fair, even though ESnet handles a volume of traffic comparable to the largest university networks, it is not entirely typical. The ESnet sites are almost entirely large research institutes that generate and/or consume high-volume data streams. While universities also generate high-volume data streams they have a higher proportion of commodity-like traffic.

Figure 2 shows the ESnet traffic growth and the shaded tops on the bars between May 2004 and July 2006 represent the traffic in the top 1,000 flows (out of several billion). The shaded tops from January 2009 to the present represent the traffic flowing in a few dozen, site-to-site virtual circuits (described below). From August 2006 to December 2008 the trend continued, but there was no data available to quantify the trend.

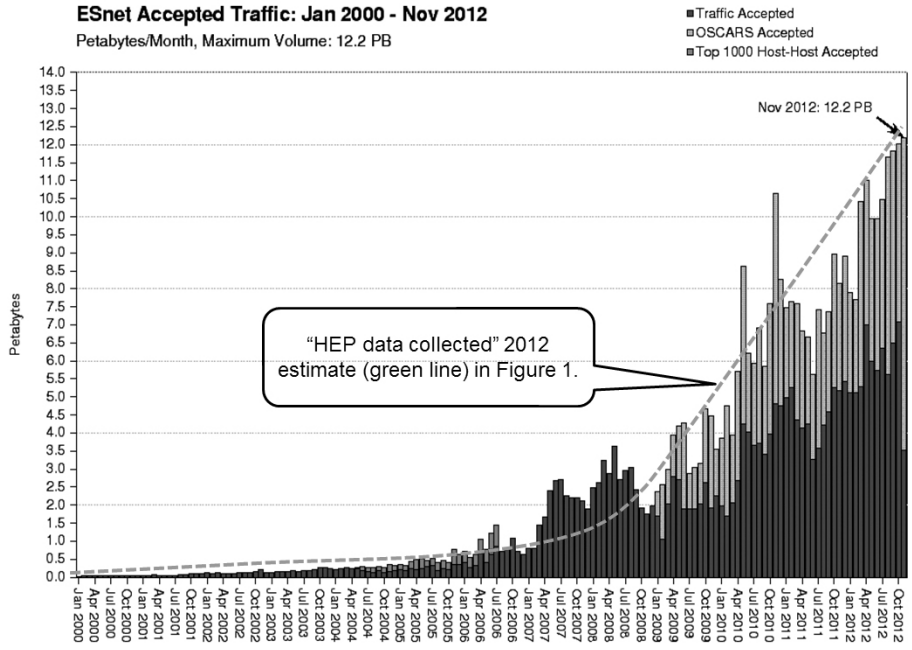

Figure 2. ESnet has seen exponential growth in traffic every year since 1990 (the traffic grows by factor of 10 about every 47 months). By March 2013 the accepted traffic was 15.6 PBy.

As seen in Figure 1, HEP data growth is expected to grow by a factor of 10 over the next several years, and the Belle-II experiment will begin contributing comparable traffic within five years. In the five-to-ten year time-frame the SKA and ITER will come on-line, and while their exact data movement patterns are not known at this time, it is known that they will be moving more data than the LHC experiments.

In response, most of the large R&E networks, including ESnet, are building next-generation networks that are 100 Gb/s per optical channel. These networks are typically based on the use of dedicated optical fiber that provides 80 to 100 channels at 100 Gb/s per channel, or 8-10 terabits/sec (8,000-10,000 gigabits/sec) per fiber.

ESnet's initial deployment of the optical-layer technology in a new network is illustrated in Figure 3.

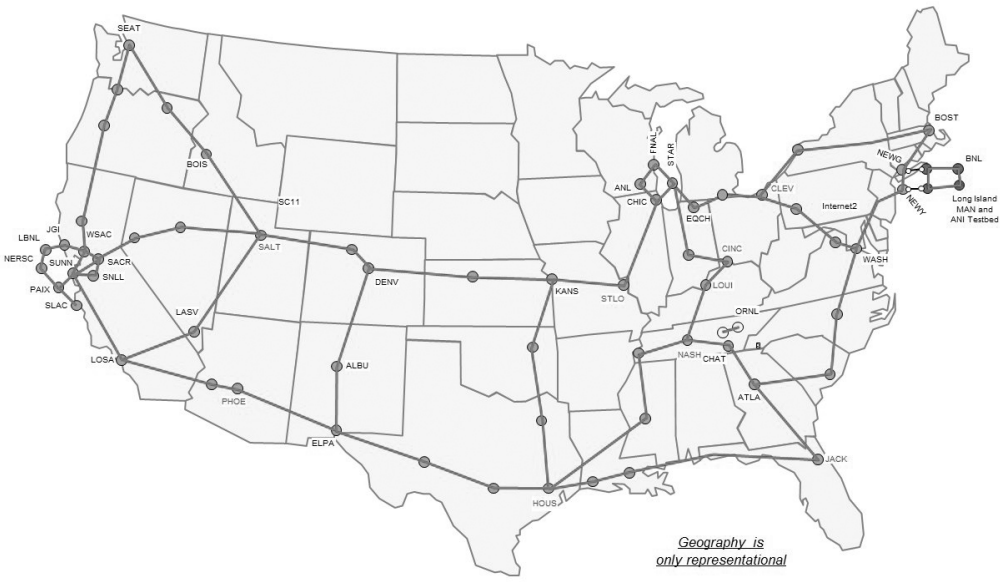

Figure 3. The ESnet5 Optical Network

The network includes approximately 13,000 miles of fiber with 200 optical amplifiers, 88 optical channels / waves, capable of 100 Gb/s each and about 60 optical add/drop sites where traffic can be inserted or extracted from the fiber (the 88 waves are shared equally with Internet2: 44 waves each).

While advances in electronics have been necessary to achieve 100 Gb/s switching and routing, these advances have been largely incremental. The advances in the underlying optical transport technology have been revolutionary.

3.1 Advances in optical transmission technology[a]

What makes it possible to transport 100 Gb/s of data on a single optical channel (colloquially called a "wave" or "lambda") is substantial advances in the technology of using optical signals to carry data.

The ESnet optical network uses Ciena's 6500 Packet-Optical Platform with WaveLogic™ coherent optical processors. Infinera's DTN with FlexCoherent™ coherent optical processors is another such system, and both achieve 100 Gb/s per channel. The technology involved is similar to other modern optical transport systems, of which there are in the order of ten on the market today.

A very brief characterization of the technology that is involved is as follows.

Optical transport systems encode information in "symbols" – the smallest optical signal that is transmitted. In the simplest case a symbol represents a single bit, which was the case in the previous generation of optical transport systems (10 Gb/s per optical channel / wave).

Symbols are transmitted in optical channels that are situated in specific and limited frequency bands in the optical spectrum. In modern dense wave division multiplexing (DWDM) transport systems these bands are about 35 nm (nanometers) wide. This spectrum (band) is divided into channels, each of which is about 0.4 nm wide. Each channel is typically used for a data transport path such as 10, 40, or 100 gigabit Ethernet.

There is a fairly narrow spectral band in an optical fiber that is useful for transmitting data. This is because the signal transmission needs to use the part of the spectrum where the transmission characteristics of the optical fiber are mostly linear and not greatly different across the usable band, and where a suitable optical amplification technology exists. While there are several such bands, the most common band for modern optical transmission is the "C" band, which is 1530 nm to 1565 nm. (The optical spectrum of human vision is about 380 nm – 700 nm, and so 1530 nm is well into the infrared.) In current DWDM systems, this band is divided up into 88 optical channels each of which is 0.4 nm wide – the ITU "grid." These 0.4 nm channels can handle a maximum symbol rate of about 50 GHz (50 giga baud[b]) (e.g. 1530 nm - 1530.4 = 195,942 GHz – 195,891.5 GHz = 50.5 GHz), which at 1 bit per "symbol" gives 50 Gb/s maximum. This maximum must also include the error correcting bits – forward error correction, "FEC" – which consume about 10% of the bandwidth. Therefore, in order to get more than about 40-45 Gb/s per channel of data transmission a more sophisticated approach must be used to encode data in symbols.

The big breakthrough in optical data transmission came with a series of developments collectively called "coherent" technology. These technologies include on the transmit side, high-order amplitude and phase modulation and polarization multiplexing; and on the receive side optical heterodyne detection (also called coherent detection with a local oscillator - see [9] and the Infinera whitepaper cited below), high-speed analogue-digital converters (ADCs), and digital signal processing to recover the original digital signal from the various encoding techniques.

How are these techniques used to achieve 100 Gb/s per channel / wave?

Briefly, lasers produce monochromatic light that originates from a stable oscillator (e.g. with a stable frequency), so it has a constant or fixed phase. If such a signal is split into two signals and one of the signals delayed by a fractional wave length amount, then you have two signals that are phase shifted with respect to each other. That is to say, they are the same frequency, but the apparent origin of the frequency variation of one is time delayed with respect to the other. This allows for controlled varying of the phase in two different signals to encode information. For a clear introduction to this see [10]; also see [11]. Lasers also produce light at a fixed polarization (a fixed angle between the E or B vectors that make up the electromagnetic wave that is light and a

---

[a] Thanks to Inder Monga and Chris Tracy, ESnet, for comments on this section.

[b] In telecommunication and electronics, baud is synonymous to symbols per second or pulses per second. It is the unit of symbol rate or modulation rate; the number of distinct symbol changes (signaling events) made to the transmission medium per second in a digitally modulated signal or a line code. (Wikipedia)

reference frame). This allows for the possibility of taking a single source and rotating it by a fixed amount and then encoding information separately on the two polarizations that are recombined for transmission.

One example of using these techniques involves optical transport using "dual polarization-quadrature phase shift keying" (DP-QPSK) technology with coherent detection. See [10], [12], and [13]. DP-QPSK involves dual polarization, which effectively provides two independent optical carriers, and quadrature phase shift keying that encodes data by changing the phase of the optical carrier and reduces the symbol rate by half, thereby sending twice as much data by using quadrature phase shift keying on each of the polarizations separately. This results in four times the data density. That is, four bits/symbol.

Together, DP and QPSK easily support a 100 Gb/s payload (plus FEC overhead) in 50 GHz of spectrum.

The way that this is implemented is by using a set of building blocks called Mach-Zehnder modulators. Each of these modulators allows for a one bit encoding. By using multiple such modulators in combination with multiple polarizations you can encode multiple, independent single bit signals. This not only provides multiple bits encoded in the same symbol, but because there are multiple digital inputs (as opposed to a single serial input), the encoding can be driven using conventional CMOS transistor technology to manage parallel bit streams. (See Infinera whitepaper noted below.)

The actual transmission rate is 7% to 20% higher than the data signal to include Forward Error Correction data that both validates the received data and allows for reconstructing erroneous data transmission. See [14].

This discussion is a substantial simplification of the optical technology involved – see the papers sited above and [15] and [16] for details. A nice, readable introduction to the collection of technologies that make up the "coherent" technologies used in modern DWDM systems is Infinera's whitepaper "Coherent DWDM Technologies" available at http://www.infinera.com/solutions/whitepapers.html.

In order to support aggregate rates higher than 100Gbps, mainly 400Gbps and beyond, many optical vendors are trying to build a group of lower rate optical channels that act and are managed as a single channel, and are calling it "super-channels." (It is also called "multiplexed transponders" – e.g. see [10].) Infinera's first steps toward a more flexible use of the optical spectrum relaxes the restriction of using the fixed optical channel bandwidth (noted above) within the super-channel clusters. See "Super-Channels: DWDM Transmission at 100 Gb/s and Beyond." This multiplexed approach is generally seen as an intermediary to serial modulation, just as 10 X 10G was an intermediary to serial 100G [10].

4 Widely used data transport protocols must operate at 100 Gb/s

Although there are other transport protocols available, TCP remains the workhorse of the Internet, including for data-intensive science, and so TCP must be made to work well.

In the extensive monitoring done by ESnet, we have observed that the Internet is full of undetected "soft errors", where TCP works, but over long distances – and, therefore, high latency paths – only at speeds 10-100 times slower than the link capacity.

Why the huge impact? TCP is a "fragile workhorse." It will not move very large volumes of data over international distances unless the network is error-free.

The reason for TCP's sensitivity to packet loss is the interaction of the buffering needed for high-speed, high-latency paths and the slow-start and congestion avoidance algorithms that were added to TCP to prevent congestion collapse of the Internet.

Congestion collapse was first observed on the early Internet in October 1986, when the US NSFnet phase-1 backbone throughput dropped three orders of magnitude from its capacity of 32 kbit/s to 40 bit/s, and this continued to occur until end-nodes started implementing Van Jacobson's congestion control algorithms in the TCP stacks between 1987 and 1988. See [17].

Packet loss is seen by TCP's congestion control algorithms as evidence of congestion, so they slow down and prevent the synchronization of the senders (which perpetuates and amplifies the congestion), leading to network throughput collapse. Network link errors also cause packet loss, so these algorithms come into play, with dramatic effect on throughput in the wide area network – hence the need for "error-free" networks.

To illustrate, consider this example: On a 10 Gb/s link a 0.0046% loss (1 packet in 22,000) was observed. In a LAN or metropolitan area network, this level of loss is barely noticeable because of how TCP works. That is, in a LAN there is relatively little buffering required to achieve full bandwidth. In a continental-scale network – 88 ms round trip time path (about that of across the US) – this seemingly insignificant rate of packet loss results in an 80x decrease in throughput for TCP due to the large buffers required to "fill the pipe," and the time required to make full use of the buffers in the face of TCP's slow-start after a bit error causes a packet drop. See [26].

While the congestion control modifications were (and still are) necessary in the general Internet, they make TCP perform poorly in high-performance environments with soft errors (occasional bit drops as opposed to circuit partition – a hard error). The Internet engineering community has been working on improvements to achieve higher performance (either further enhancements to TCP or a high-performance protocol to replace TCP) for many years with some limited success. Improvements include congestion control algorithms such as CUBIC [18] that recover more quickly from loss events, buffer auto-tuning, and TCP window scaling [19]. However, none of these enhancements have resolved the issue of the sensitivity of TCP to packet loss for long-distance high-performance use. Therefore, in the near to medium term, data-intensive science network infrastructure must be maintained error-free so that TCP-based applications perform well in high-performance science environments. This is a challenging problem for several reasons, including soft failures as described above and the number of organizations and devices involved in a typical long-distance data transfer.

4.1 Topic 2a: Monitoring and testing of the network must detect errors and facilitate their isolation and correction

The only way to keep multi-domain, international scale networks error free is to test and monitor continuously end-to-end. This led to the development and deployment of a monitoring infrastructure that could be used to detect and isolate problems that showed up in virtual circuits and in routed data paths that crossed many different network domains. (Each separately administered network is a "domain." e.g. ESnet, Internet2, GÉANT, the European NRENs, etc.)

The key to successful deployment of such a system is that it be designed to operate in a federation where each domain can maintain control over its own monitoring, and the monitoring results be easily combined at a higher level to provide a global view of the network.

perfSONAR is a network monitoring framework [18] designed for federation, to collect both passive and active network measures, to convert these to a standard format [21], and then to publish the data where it is publically accessible.

Passive measurements are information collected by the network devices – typically routers, switches, and optical transport systems. These measurements include interface error counts – bit error rates – packet loss, packet counts in and out, etc.

Active network measurements are generated by tools that measure packet delays and data transport throughput. Packet delays are obtained by measuring time-of-flight using a precision clock at both ends (typically a GPS-based clock, or comparable). These tools are OWAMP [22] and HADES [23]. Throughput is measured with BWCTL (a wrapper and controller for iperf throughput tests) by setting up a TCP connection to another perfSONAR system and measuring the achievable TCP throughput. See [26] and [24].

perfSONAR has a scheduling function that allows active testing on a scheduled basis for specific paths. The results of the tests are published in a "measurement archive" (MA) so that trends (e.g. arising from increasing soft failures that indicate developing hardware problems) can be detected. Each domain maintains complete control over who may run active tests and when.

In order to associate the measurements with network paths, perfSONAR maintains a standardized representation of the network topology.

Published data is federated by tools that discover the MAs that have data along a path of interest, and then use tools that read the MAs to produce end-to-end, multi-domain views of network performance. (See [25].)

perfSONAR measurement hosts are deployed extensively throughout the R&E international networks and in the networks and end sites used by the LHC community.

5 It must be possible to move data from the application through the operating systems and onto the network at network speeds

5.1 TCP congestion avoidance algorithms

TCP congestion control algorithms on many systems dates from the mid-1980s, and use of a modern TCP stack (the kernel implementation of the TCP protocol) is important to reduce the sensitivity to packet loss while still providing congestion avoidance. See Figure 4. [26]

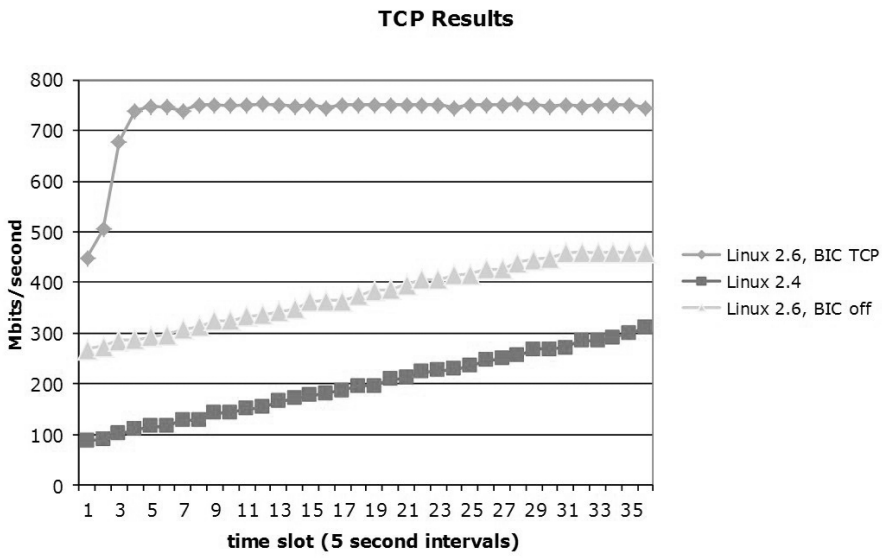

Figure 4. Modern TCP congestion avoidance algorithms have a big impact on throughput in the face of any packet loss. The "binary increase congestion" control algorithm (BIC) reaches max throughput much faster than older algorithms. (From Linux 2.6.19 the default is CUBIC, a refined version of BIC designed for high bandwidth, long paths.)

5.2 Tuning TCP

The other primary cause of poor TCP performance in science networks is incorrect configuration parameters of the host TCP implementation on the data transfer systems. (See http://fasterdata.es.net/host-tuning/ for more information on this topic.)

The appropriate configuration of TCP on data transfer nodes, e.g. TCP "window" size[a] adequate for very long round trip (RTT) (high latency) paths, can be accomplished by competent system administrators with the help of public knowledge base sites [27].

It is critical to use the optimal TCP send and receive socket buffer sizes for the RTT of the path that the applications see end-to-end. The default TCP buffer sizes are much too small for today's high speed networks. Until around 8 years ago, default TCP send/receive buffers were typically 64 KB, however the buffer size needed to fill, e.g., a California to New York 1 Gb/s path is10 Mbytes – 150X bigger than the default buffer size.

Historically TCP tuning parameters were host-global, with exceptions configured per-socket by applications. And since every application represents a unique situation, there are potentially a lot of special cases. A solution to this problem is to auto-tune TCP connections, and in modern Unix kernels this is done, though within pre-configured limits. This works, but is not a panacea because the upper limits of the auto-tuning parameters are typically not adequate for high-speed transfers on very long (e.g. international) paths, so a certain amount of hand-tuning is still needed.

5.3 Parallelism is key

It is much easier to achieve a given performance level with multiple parallel network connections than with one connection. This is partly because the OS is very good at managing multiple threads but not so good at managing sustained, maximum performance in a single thread and partly due to multiple parallel TCP flows (as opposed to a single flow transporting the data data) being somewhat less affected by errors that cause slow-start to come into play. The multiple thread advantage is also true for disk I/O. This is even more the case with modern multi-core processors.

Several tools offer parallel transfers (see below).

[a] The TCP window specifies the amount of data that can be in "flight" in the network before being acknowledged by the receiver. This needs to be large for long RTT networks so that the time to acknowledge packets does not slow down the overall transmission rate. (See the discussion of Bandwidth Delay Product in [26].)

5.4 Data transfer tools

Using the right tool is very important, and latency tolerance in the tools is critical. Many tools and protocols assume latencies typical of a LAN environment (a few milliseconds), the primary example being SCP/SFTP and HPSS mover protocols, both of which work very poorly in long path networks.

Example results are shown in Table 1.

Tool testing results on a 10 Gb/s path Berkeley, CA to Argonne, IL (near Chicago), RTT = 53 ms	
Tool	**Throughput**
scp	140 Mbps
HPN[1] patched scp	1.2 Gbps
ftp	1.4 Gbps
GridFTP, 4 streams	5.4 Gbps
GridFTP, 8 streams	6.6 Gbps
[1] PSC (Pittsburgh Supercomputer Center) has a patch set that fixes problems with SSH. This also helps rsync. See [28].	
Note that to get more than 1 Gbps (125 MB/s) disk to disk requires RAID.	

Globus GridFTP [29] is the basis of most modern high-performance data movement systems. It provides parallel network streams, buffer tuning, help in getting through firewalls (open ports), ssh, etc.

The newer Globus Online [30] incorporates all of these and small file support, pipelining, automatic error recovery, third-party transfers, etc. This is a very useful tool, especially for the application communities outside of HEP that do not have the resources to develop their own data transfer tools.

Another approach to the WAN data movement problem is Caltech's FDT (Faster Data Transfer) approach. FTD is not so much a tool as a hardware/software system designed to be a very high-speed data transfer node. FTD makes explicit parallel use of multiple disks and can fill 100 Gb/s WAN network paths. See [31].

A survey of the evolution of such data parallelism may be found in reference [32].

6 Application design

In order to move and process (e.g. analyze) the hundreds of terabytes of data per day that is involved in data-intensive science, applications must be designed and implemented so that all resources are kept operating at the highest possible efficiency.

We refer to the sciences that depend on moving massive amounts of data among globally distributed sites to accomplishing their science as “data-intensive science,” and this section looks at the data mobility issues faced by data-intensive science. (For a more detailed discussion of some of these topics see [33].)

In order to effectively move large amounts of data over the network, automated systems must be used to manage workflow and error recovery.

For example, the filtered ATLAS data rate of about 25 Gb/s is sent to 10 national Tier 1 data centers that maintain the working copy of the dataset and perform the first analysis step called “reconstruction” (transforms raw detector data to physical quantities, e.g. particle tracks, used in analysis).

The Tier 2 sites hosting the physics groups that analyze the data and carry out the science get a comparable amount of data from the Tier 1 centers. The Tier 2 sites provide most of the compute resources for analysis and they cache the data for other Tier 2 sites.

For example, the Production and Distributed Analysis system (PanDA) used by the ATLAS experiment is a highly distributed and highly automated workflow and workload management system (Figure 5). It coordinates the analysis resources and interacts with the Distributed Data Management service (DDM). Analysis jobs are submitted to the central manager that locates computer resources and matches these with dataset locations. The “best” (most easily accessible) data replica needed for a particular analysis might be in one of the Tier 1 data centers, or there may be a more available copy at one of the Tier 2 centers, where data is frequently cached while it is being used. PanDA invokes a data movement subsystem that replicates the data to the site with the available computing resources. When the data is in place PanDA launches the analysis jobs.

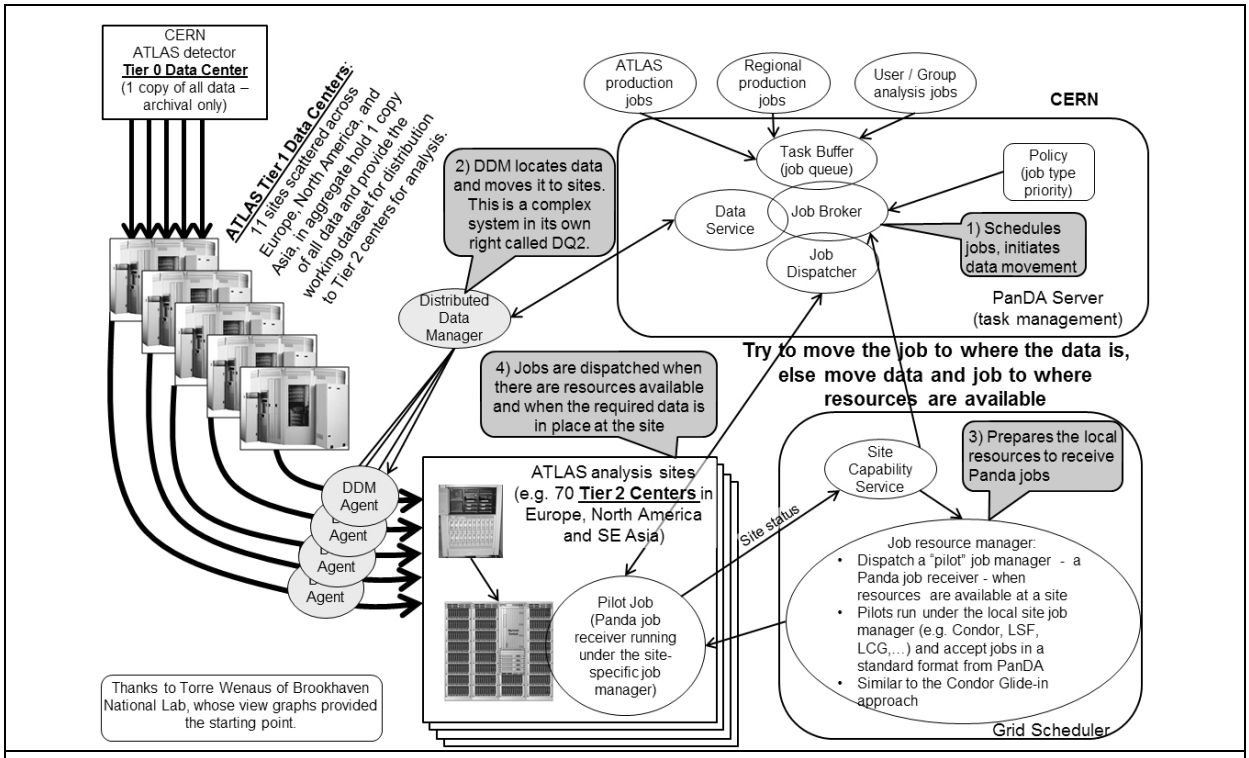

Figure 5. ATLAS PanDA (Production and Distributed Analysis) data management and analysis system. See [34] and [35].

PanDA has demonstrated its ability to manage well over 100,000 simultaneous jobs (of order a million per day). The DDM service coordinates data movement of hundreds of terabytes/day, and manages (locates, moves, stores) of the order of 100 petabytes of data per year in order to accomplish the ATLAS science.

The 100,000+ PanDA jobs cause data movement over international distances of hundreds of terabytes/day. (See Figure 6.) It is this scale of data movement and analysis jobs, going on 24 hr/day, 10 months/yr that the distributed systems and their interconnecting networks must support in order to enable this sort of large-scale science.

In order to debug and optimize the distributed system that accomplishes the scale of the ATLAS analysis, years were spent building and testing the required software and hardware infrastructure before the LHC started producing data. Once the systems were in place, systematic testing was carried out in "service challenges" or "data challenges."

The service challenges were intended to simulate the operation of the entire distributed system just as it would operate when the LHC came on-line. The LHC Computing Grid (LCG) Project was launched in March 2002 to prepare, deploy and operate the computing environment for LHC data analysis. The final service challenges – using synthetic data based on modeling the LHC accelerator and detectors – operated at the same scale of data volume and world wide data movement that the real data would require. See, e.g., [37].

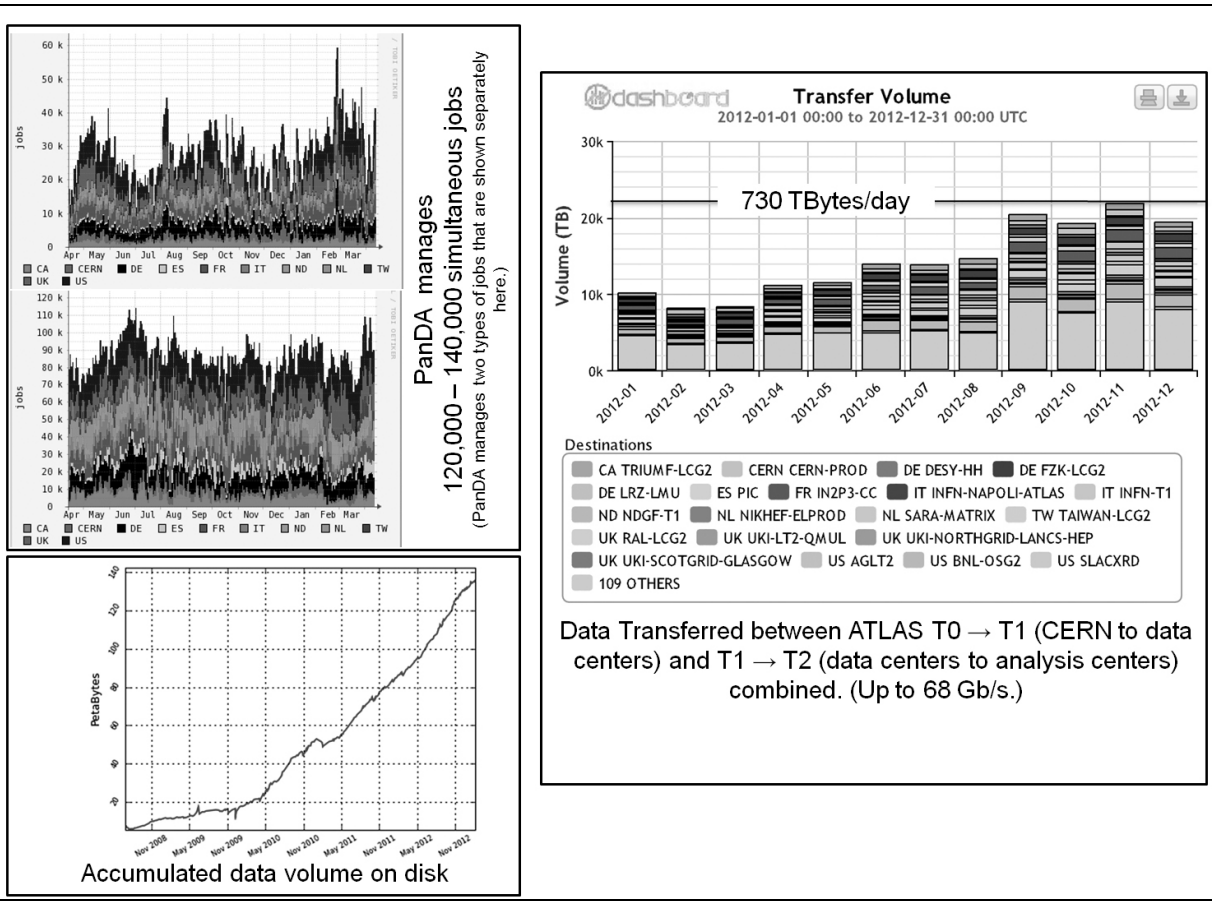

Figure 6. The scale of ATLAS data analysis [36].

This is when the impact of the LHC data movement started to become apparent in the production R&E networks – as the service challenges ramped up several years prior to LHC turn-on. Large-scale data flows started to show up in the network in about 2003 (the shaded top of the bars in Figure 2 and Figure 7). The transition from large-scale testing to actual LHC data analysis in 2010 was a continuous progression of increases – not a step function (Figure 7).

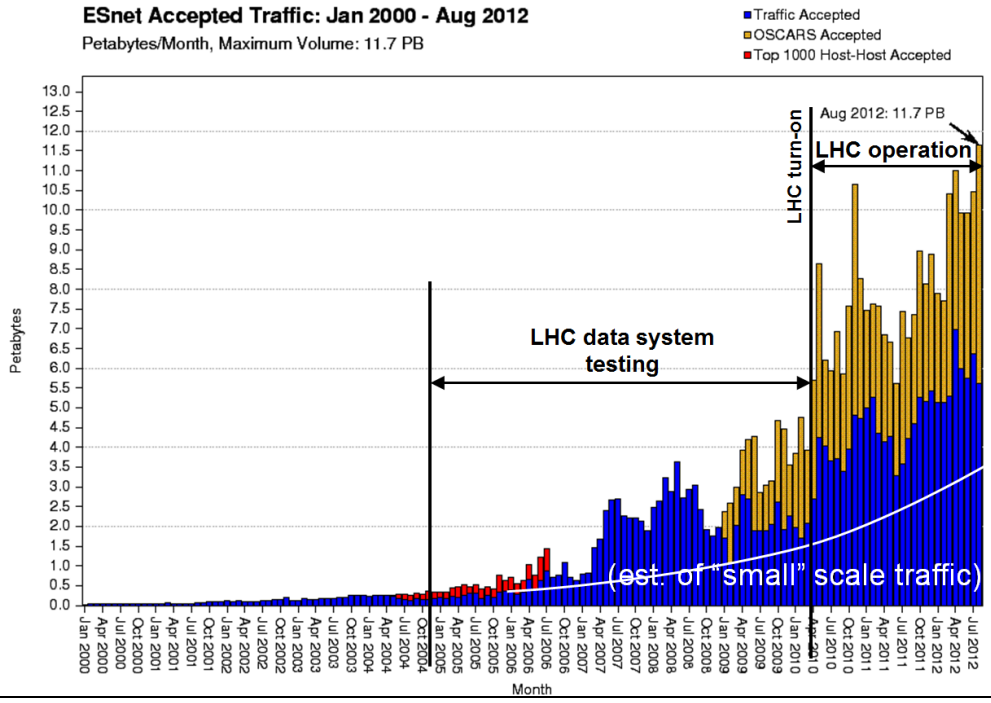

Figure 7. Ramp-up of LHC traffic in ESnet

The approach for managing this magnitude of data by such high-throughput systems is evolving as experience is gained in handling half a petabyte/day, and as disk resources turn out to be just as limiting as CPU resources in terms of maximum analysis job throughput. The original strategy was to pre-place data at various sites by "guessing" what data would be of interest and knowing which sites had the resources. This gave way to a demand-based caching scheme that places data based on analysis job requests. This is the current approach. Another big change is currently underway: the introduction of federated site data repositories and remote I/O as the data access method. This approach has arisen because the assessment of the LHC community is that the network is now capable of supporting such a model.

The point here is that the approach to massive data handling problem of the LHC is still evolving in significant ways, and some of the approaches are being considered because the capabilities of R&E network have increased significantly in the past five years.

7 Optimizing the available network resources

In order for data-intensive science to operate they must have access to the network resources deployed by the R&E networking community, and those resources must be structured in such a way as to be able to provide high-speed data throughput end-to-end.

7.1 Dedicated, purpose-built network infrastructure is needed to support quasi real-time data transfers, e.g. from instrument to data center

When instruments like the LHC produce data in essentially real time, and that data has to be distributed by networks, some sort of dedicated infrastructure is necessary to ensure adequate capacity and non-interference.

The role of the LHC Optical Private Network (LHCOPN) is to ensure that all data moves from CERN to the national Tier 1 data centers continuously. The LHCOPN is a collection of leased 10 Gb/s circuits that connect CERN with the 11 national Tier 1 data centers.

While the LHCOPN was a technically straightforward exercise – establishing 10 Gb/s links between CERN and the Tier 1s for distributing the detector output data – there were several aspects that were new to the R&E community. The issues that arose related to the fact that most sites connected to the R&E WAN infrastructure through a site firewall and the OPN was intended to bypass site firewalls in order to achieve the necessary performance. Because of this, the security issues were the primarily ones in building LHCOPN. These issues were addressed primarily by using a private address space that hosted only LHC Tier 1 systems (see [38]).

It is important to note that while in 2005 the only way to handle the CERN (T0) to Tier 1 transfers was to use dedicated, physical, 10G circuits, today, in most R&E networks, 100 Gb/s links are becoming the norm and the LHCOPN could be provided using virtual circuits with bandwidth guarantees. (See next section.)

The ESnet part of the LHCOPN has used this approach for some five years now – in fact this is what ESnet's OSCARS virtual circuit system was originally designed for. However, such an international-scale virtual circuit infrastructure would have to be carefully tested before taking over the LHCOPN role, and the design of this approach and a testing plan are currently under development.

7.2 A multi-domain, end-to-end virtual circuit network service is needed to provide guarantees for bandwidth sensitive applications/tasks

Distributed systems like ATLAS's PanDA have a predictable pool of resources to draw from in terms of the available CPUs and disk capacity; however these resources were all coupled by the best-effort (no bandwidth guarantees) characteristics of the Internet-like R&E networks. Network performance comparably predictable to the other managed resources was needed for the smooth overall functioning of the system: The science community needs to be able to treat the network as a service that could be integrated into their analysis systems in the same way that the other resources are.

Similar requirements are found across many large-scale science collaborations in that they use distributed applications systems in order to couple existing pockets of code, data, and expertise into "systems of systems" and to break up the task of massive data analysis and use data, compute, and storage resources that are located at the collaborator's sites. See [7].

The network service that was developed to meet the requirements provided a "virtual circuit" between specified end points, that has a guaranteed bandwidth, and that can be requested for some specific time interval in the future. Traffic isolation is also provided to allow for use of high-performance, non-standard transport mechanisms that cannot co-exist with commodity TCP-based transport in the general infrastructure.

The way that packet-switched networks like the Internet provide a circuit service is with "virtual circuits" (also called pseudowires) that emulate point-to-point connections using, e.g., MPLS or OpenFlow network overlays. Both MPLS and OpenFlow create virtual circuits by adding a tag to packets that is then used to switch the packet in a router/switch that understands those tags. Defining the tag paths through a device defines a deterministic path rather than the potentially changing, "hot potato" routing of the Internet. Bandwidth guarantees can be provided by giving the traffic in the virtual circuit a higher priority that all other traffic in the network and then limiting the virtual circuit to only the bandwidth that was specified (and agree to by the circuit reservation system) when the circuit was requested.

In addition to meeting the user requirements, such a service channels big data flows into virtual circuits in ways that allow network operators to do "traffic engineering" – that is to say, to manage/optimize the use of available network resources and to keep big data flows separate from general traffic by directing the virtual circuits to specific physical network paths.

The service provides for secure connections in the sense that the circuits are "secure" to the edges of the wide area network (WAN) network (the site boundary) because they are managed by the control plane of the WAN network which is highly secure and isolated from general traffic. If the sites trust the circuit service model of all of the involved networks then the circuits do not have to transit the site firewall when the far end is known and trusted and the connection cannot be intercepted.

ESnet's OSCARS provided one of the first implementations of the virtual circuit service (see [42]): it is essentially a routing control plane (path determination and network device management) that is independent from the router/switch devices. At the network device level, OSCARS supports MPLS, Ethernet VLANs, GMPLS (a generalization of MPLS used in optical networks), and OpenFlow network technologies to provide virtual circuits in the network. What OSCARS does is now called SDN (Software Defined Networking) by the OpenFlow community.

In addition to the requirement given above, additional features that have arisen through user experience with OSCARS include flexible service semantics – e.g. allowing a user to exceed the requested bandwidth, if the path has idle capacity even if that capacity is committed but unused. This semantic turns out to have surprisingly important consequences in terms of letting users build overlay networks with specified behavior in the face of path failure. See [42].

Subsequently OSCARS has been adopted by several dozen other networks, and an international collaboration called DICE [39] was set up to define a compatible service definition so that different implementations could be used to set up end-to-end circuits across the many network domains involved in the LHC data transport. Such compatible services are now deployed in many of the R&E networks in Europe, the Americas, and Asia. The ad-hoc DICE effort – the inter domain controller protocol (IDCP) [40] – is now moving into the Open Grid Forum (OGF) standards organization in the Network Service Interface working group [41].

7.3 The typical campus LAN interface to the WAN must be redesigned to support large, long distance data flows

As other issues that inhibit performance were being addressed, it became apparent that once you provide high quality, high performance data transfer capability in the wide area networks and applications, then you run into bottlenecks in the campus[a]. Local area / site networks (LANs) were not designed to support data-intensive use, that is to say, they are not designed to move large volumes of data into and out of the campus to the wide area network (WAN).

The site network (the LAN) typically provides connectivity for local resources – computer, data, instrument, collaboration system, etc. – needed by data-intensive science. Therefore, a high performance interface between the WAN and the LAN is critical for large-scale data movement.

The problem with campus LANs arises from the devices and configurations typically deployed to build LAN networks for business and small data-flow purposes: firewalls, proxy servers, low-cost switches, and so forth, almost always impede large-scale, long-distance data flows.

To provide high data-rate access to local resources the site LAN infrastructure must be re-designed to match the high-bandwidth, large data volume, high round trip time (RTT) (international paths) of the WAN flows, otherwise the site will impose poor performance on the entire high speed data path, all the way back to the source.

[a] We use the terms "campus" and "site" interchangeably and in the general sense of a research and/or education institution: A national laboratory, a university, a research institute, etc.

In order to address the campus LAN-WAN interface bottleneck, the nature and architecture of campus networks were examined and redesigned to better accommodate moving large volumes of data across the campus boundary while not interfering with other campus traffic, and while accommodating appropriate security policies that are intended to protect the campus from dangerous aspects of the wider Internet. One way to accomplish this is with a new campus network architecture called the "Science DMZ" [43].

The computer and data resources involved in data-intensive sciences are deployed in a separate portion of the site network that has a different packet forwarding path that uses WAN-like technology and has a tailored security policy.

In traditional LAN architectures the DMZ is typically a portion of the LAN, at or near the border router that connects to the WAN provider that is reserved for services that are primarily accessed by the outside world. Externally accessed services such as the campus Web and email servers are deployed on the DMZ to keep the external traffic out of the LAN environment. Similarly, a high-performance version of the DMZ can be deployed to provide connectivity to the high-performance data moving systems used for data-intensive science – this is called the Science DMZ.

The Science DMZ typically consists of dedicated systems built and tuned for wide-area data transfer and test and measurement systems for performance verification and rapid fault isolation, typically perfSONAR. A security policy tailored for science traffic is implemented using appropriately capable hardware (e.g. routers that support high performance access control lists rather than enterprise firewalls which can severely limit WAN performance for long distance, high-speed data transfers – see [44].

The essential components and a simple architecture for a Science DMZ are shown in Figure 7. (For a more detailed discussion of variations of the ScienceDMZ architecture see [33] and [45].)

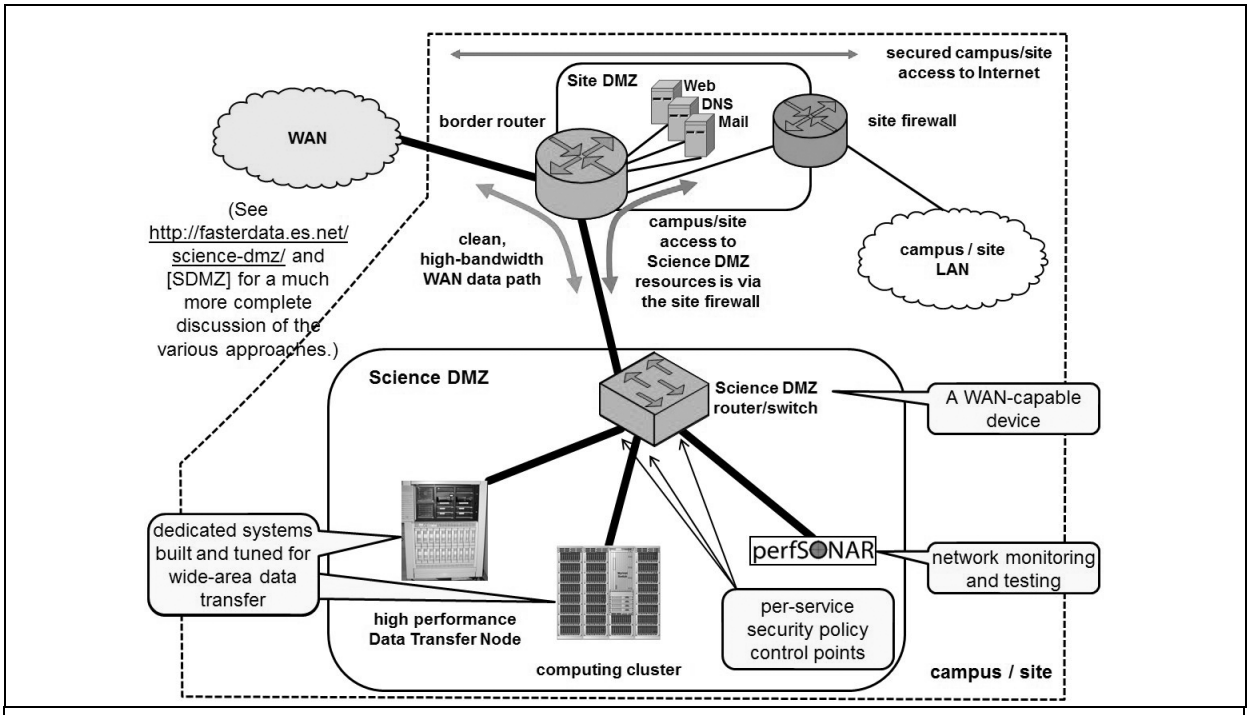

Figure 8. Simple version of a Science DMZ architecture

The success of the Science DMZ concept is demonstrated by the fact that a recent U.S. National Science Foundation (NSF) call for proposals for campus infrastructure supporting data-intensive science recommended that campuses adopt the Science DMZ architecture [46].

7.4 Large-scale science traffic in the shared R&E network infrastructure must be explicitly managed in order to provide good service to science in ways that will not disrupt other uses of the infrastructure

The traffic from the LHC Tier 1 data centers to the Tier 2 analysis centers is now large enough that it must be managed separately from the general R&E traffic. This is necessary both to ensure that the science traffic can get enough bandwidth to move the required data and because the presence of very large flow in undifferentiated infrastructure can cause the many commodity flows to behave poorly.

Both ATLAS and CMS Tier 2 analysis centers have largely abandoned the original MONARC hierarchical data distribution model of {Tier 1 → associated Tier 2 → Tier 3} in favor of a chaotic model: get whatever data you need from wherever it is available {Tier 1 ↔ any Tier 2 ↔ any Tier 2 ↔ any Tier 3}.

Because the Tier 1 ↔ Tier 2 data flows, in aggregate, are at least as large as the Tier 0 → Tier 1 flows this puts very large traffic flows onto the general infrastructure.

In the original hierarchical model, it was relatively easy to identify specific network paths that need increased capacity; in the chaotic model, the traffic can show up anywhere. In 2010 this resulted in enormous site-to-site data flows on the general IP infrastructure (including on the N. America to Europe transatlantic paths) at a scale that has previously only been seen from DDOS attacks.

Managing all possible combinations of Tier 2 – Tier 2 flows (potentially 170 × 170) cannot be done just using a virtual circuit service because that service is a relatively heavy-weight mechanism, however, the substantial flows and volume of data involved in data-intensive science make it necessary to separate this traffic from the general Internet traffic.

Special infrastructure is required for this, and the LHC's Open Network Environment – LHCONE – was designed for this purpose. LHCONE provides a private, managed infrastructure that is an overlay on the general R&E networks, and that is designed for LHC Tier 2 traffic (and likely other large-data science projects in the future). The architecture is a collection of routed "clouds" using address spaces restricted to subnets that are used by LHC systems. The clouds are mostly local to a network domain (e.g. one for each involved domain – ESnet, GÉANT ("fronts" for the NRENs), Internet2 (fronts for the US universities), etc.

The clouds (virtual routing instances in the WAN networks called VRFs) are interconnected by point-to-point circuits provided by various entities (mostly the domains involved). In this way the LHC traffic can be directed to circuits designated by the WAN network engineers.

The LHCONE could be put into place relatively "quickly" because there is capacity in the R&E community that can be made available for use by the LHC collaboration that cannot be made available for general R&E traffic, and because VRFs are relatively easy to set up in most WAN networks.

LHCONE is, therefore, essentially built as a collection of private overlay networks (like VPNs) that are interconnected by managed links to form a global infrastructure where Tier 2 traffic will get good service and not interfere with general traffic (Figure 8). From the point of view of the end sites, they see a LHC-specific environment where they can reach all other LHC sites with good performance. See LHCONE.net

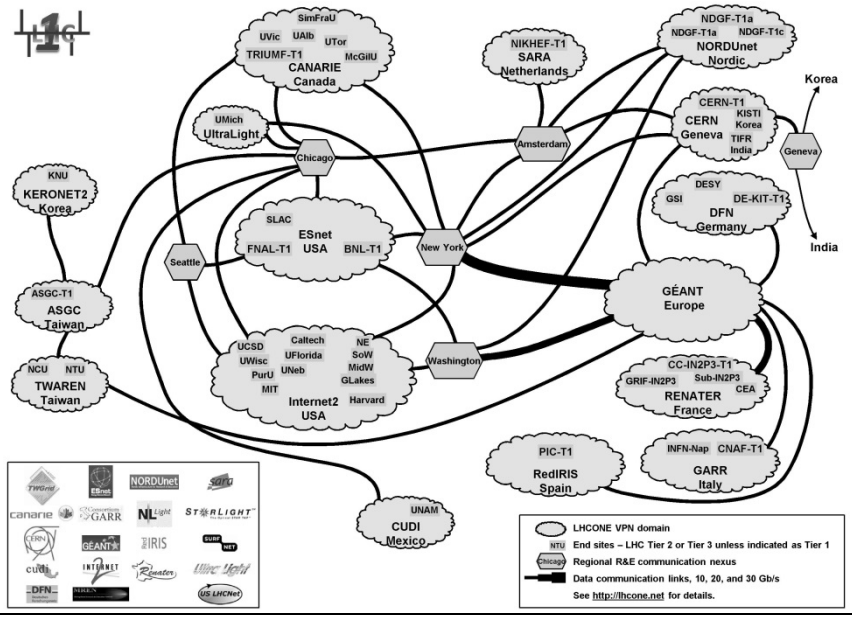

Figure 9. LHCONE in 2013

8 Disseminating knowledge

It is essential that the knowledge gained in the network, operating systems, and computer science research communities about how to optimize various aspects of the end-to-end data transfer problem is delivered into the hands of the science application developers and operators.

An example of such a knowledge base maintained by ESnet is at http://fasterdata.es.net and contains contributions from several organizations.

The ESnet knowledge base topics include, for example:

- Network architecture, including the Science DMZ model
- Host tuning
- Network tuning
- Data transfer tools
- Building a high performance data transfer node (DTN)
- Network performance testing and troubleshooting
- And special sections on: Linux TCP tuning, Cisco 6509 tuning, perfSONAR how-to, Active perfSONAR services, Say no to SCP, and TCP issues explained

9 Summary: The Message

A significant collection of issues must *all* be addressed in order to achieve the sustained data movement needed to support data-intensive science such as the LHC experiments. The approach described here provides the LHC science collaborations with the data communications underpinnings for a unique large-scale, widely distributed, very high performance data management and analysis infrastructure that is an essential component in scientific discovery at the LHC. However the issues are by no means unique to the LHC, and other disciplines that involve data-intensive science are facing, or will face, most of these issues.

10 Acknowledgements

"This work has been funded in part by the United States Department of Energy under contract numbers Contract No. DE-AC02-98CH10886 (Brookhaven National Lab) and parts of this work were supported by the Director, Office of Science, Office of Advanced Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 (ESnet and Lawrence Berkeley National Lab).

Notes and References

[1] LHC, "The Large Hadron Collider Project". http://lhc.web.cern.ch/lhc/general/gen_info.htm

[2] http://public.web.cern.ch/public/en/lhc/ATLAS-en.html http://atlas.ch/pdf/atlas_factsheet_4.pdf http://www.atlas.ch/atlas_brochures_pdf/tech_brochure-11.pdf

[3] http://public.web.cern.ch/public/en/lhc/CMS-en.html

[4] T. Kuhr, T. Hara, Belle II Computing Group, "Computing at Belle II", The International Conference on Computing in High Energy and Nuclear Physics (CHEP). http://indico.cern.ch/getFile.py/access?contribId=20&sessionId=4&resId=0& materialId=slides&confId=149557

[5] P.E. Dewdney, P.J. Hall, R.T. Schilizzi, T.J.L.W. Lazio, "The Square Kilometre Array", Proceedings of the IEEE, 97(8). http://www.skatelescope.org/publications/

[6] W.E. Johnston, R. McCool, "The Square Kilometer Array – A next generation scientific instrument and its implications for networks", ESnet, Lawrence Berkeley National Laboratory, Signal Transport and Networks, SKA Program Development Office, Jodrell Bank Centre for Astrophysics, TERENA Networking Conference 2012. https://tnc2012.terena.org/core/presentation/44

[7] See http://www.es.net/about/science-requirements/

[8] See http://www.es.net/about/

[9] http://en.wikipedia.org/wiki/Optical_heterodyne_detection

[10] I. M. Polo, "Optical Modulation for High Bit Rate Transport Technologies." Sunrise Telecom, October, 2009. www.sunrisetelecom.com/support/Intro_optical_Modulation.pdf

[11] K. Roberts, D. Beckett, D. Boertjes, J. Berthold, C. Laperle, "100G and beyond with digital coherent signal processing", Ciena Corp., Ottawa, ON, Canada, Communications Magazine, IEEE, July 2010.

[12] "100G Ultra Long Haul DWDM Framework Document", June 2009. http://www.oiforum.com/public/documents/OIF-FD-100G-DWDM-01.0.pdf

[13] C. Tracy, ESnet, "100G Deployment: Challenges & Lessons Learned from the ANI Prototype & SC11". http://www.nanog.org/meetings/nanog55/presentations/Tuesday/Tracy.pdf

[14] R. Eisenach, "'Soft decision' FEC benefits 100G." Lightwave, May 1, 2012. http://www.lightwaveonline.com/articles/print/volume-29/issue-30/features/sort-decision-fec-benefits-100g.html

[15] R. Lyons, "Quadrature Signals: Complex, but not Complicated". http://www.dspguru.com/sites/dspguru/files/QuadSignals.pdf

[16] K. Lewotsky, "100 Gb/s gets ready for prime time", SPIE Newsroom, Optoelectronics & Communications. http://spie.org/x48725.xml

[17] Van Jacobson, who ran the network research group at Lawrence Berkeley Laboratory, identified the causes of congestion collapse as poorly designed TCP implementations (the norm at the time) and proposed modifications to TCP to address the problems. (See the Wikipedia article on network congestion - http://en.wikipedia.org/wiki/Network_congestion). Van went on to build a robust implementation of his proposed TCP changes and worked with Sun Microsystems to include the changes in the Sun operating system. Over the next several years most manufacturers also picked up and implemented the changes. Vern Paxson, then a graduate student in Van's group and now a Computer Science professor at UC Berkeley, developed the techniques and tools to monitor packet streams in the Internet and deduce whether the originating system had implemented the Jacobson TCP fixes. For a number of years the IETF published lists of systems with "broken" TCP stacks as an RFC. This work was the subject of Vern's PhD thesis "Measurements and Analysis of End-to-End Internet Dynamics." (http://www.eecs.berkeley.edu/Pubs/TechRpts/1997/CSD-97-945.pdf) This work continued and evolved into the Bro network intrusion detection system (http://bro-ids.org/).

[18] S. Ha, I. Rhee, L. Xu, "CUBIC: a new TCP-friendly high-speed TCP variant." ACM SIGOPS Operating Systems Review - Research and developments in the Linux kernel, Volume 42 Issue 5, July 2008, Pages 64-74.

[19] V. Jacobson, R. Braden, D. Borman, "TCP Extensions for High Performance." IETF RFC 1323, May 1992. http://www.ietf.org/rfc/rfc1323.txt

[20] B. Tierney, J. Boote, E. Boyd, A. Brown, M. Grigoriev, J. Metzger, M. Swany, M. Zekauskas, J. Zurawski, "perfSONAR: Instantiating a Global Network Measurement Framework", 4th Workshop on Real Overlays and Distributed Systems., October 1, 2009,

[21] "OGF → Network → Measurement → Working → Group → (NMWG)". http:// nmwg.internet2.edu/

[22] "One-Way Ping (OWAMP)". http://www.internet2.edu/performance/owamp/

[23] "HADES - Hades Active Delay Evaluation System", WiN Labor Erlangen. http://www.win-labor.dfn.de/English/mainpage.html

[24] http://www.nanog.org/meetings/nanog43/presentations/Dugan_Iperf_N43.pdf

[25] https://ecenter.fnal.gov/network, http://nettest.lbl.gov/serviceTest/index.cgi?eventType=bwctl (which comes with the perfSONAR toolkit) and https://stats.es.net/perfsonar/serviceTest/cgi-bin/index.cgi?eventType=bwctl.

[26] B. Tierney, J. Metzger, "High Performance Bulk Data Transfer", ESnet, Joint Techs, July 2010. fasterdata.es.net/assets/fasterdata/JT-201010.pdf

[27] http://fasterdata.es.net

[28] C. Rapier, M. Stevens, B. Bennett, "High Performance SSH/SCP - HPN- SSH", Pittsburgh Supercomputer Center, Carnegie Mellon University, PSC. http://www.psc.edu/networking/projects/hpn-ssh/

[29] http://www.globus.org/toolkit/docs/latest-stable/gridftp/

[30] "Why Use Globus Online?". https://www.globusonline.org/whygo/

[31] "SC 2011 bandwidth challenge results". http://monalisa.cern.ch/FDT

[32] W.E. Johnston, "High-Speed, Wide Area, Data Intensive Computing: A Ten Year Retrospective", 7th IEEE Symposium on High Performance Distributed Computing, Chicago, Ill., 29-31 July 1998. http://acs.lbl.gov/~johnston/Grids/homepage.html

[33] E. Dart, W. Johnston, "Infrastructure for Data-Intensive Science – a bottom-up approach", Energy Sciences Network (ESnet), Lawrence Berkeley National Laboratory, in K. Kleese van Dam, T. Critchlow, (Editors), "Future of Data Intensive Science", Taylor & Francis Group, Publishers. (to be published). http://es.net/news-and-publications/publications-and-presentations.

[34] M. Branco, D. Cameron, B. Gaidioz, V. Garonne, B. Koblitz, M. Lassnig, R.Rocha, P. Salgado, T. Wenaus, on behalf of the ATLAS Collaboration, “Managing ATLAS data on a petabyte-scale with DQ2”, Computing in High Energy and Nuclear Physics (CHEP), 2007. http://iopscience.iop.org/1742-6596/119/6/062017 .

[35] T. Maeno, “PanDA: Distributed production and distributed analysis system for ATLAS”, Computing in High Energy and Nuclear Physics (CHEP), 2007. http://iopscience.iop.org/1742-6596/119/6/062036

[36] Graphs are from the Atlas Dashboard – e.g. http://dashb-atlas-data- test.cern.ch/dashboard/request.py/site – and are courtesy of Michael Ernst, Brookhaven National Lab.

[37] T. Wenaus, “Challenges of the LHC: Computing”, ATLAS Experiment / LCG Applications Area, BNL / CERN. http://conferences.fnal.gov/aspen05/talks/LHC-Computing-Wenaus.pdf

[38] “LHCOPN security policy document”. https://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome

[39] DICE (DANTE-Internet2-CANARIE-ESnet) Collaboration. http://www.GÉANT2.net/server/show/nav.1227

[40] Inter Domain Controller Protocol (IDCP). http://www.controlplane.net/

[41] Network Services Interface (NSI) working group, Open Grid Forum. http://ogf.org/gf/group_info/view.php?group=nsi-wg

[42] C. Guok, D. Robertson, M. Thompson, J. Lee, B. Tierney, W. Johnston, “Intra and Interdomain Circuit Provisioning Using the OSCARS Reservation System”, Energy Sci. Network, Lawrence Berkeley National Laboratory, In BROADNETS 2006: 3rd International Conference on Broadband Communications, Networks and Systems, 2006 – IEEE. 1-5 Oct. 2006. http://es.net/news-and-publications/publications-and-presentations/ also W.E. Johnston, C. Guok, J. Metzger, B. Tierney, "Network Services for High Performance Distributed Computing and Data Management", in P. Iványi, B.H.V. Topping, (Editors), "Trends in Parallel, Distributed, Grid and Cloud Computing for Engineering", Saxe-Coburg Publications, Stirlingshire, UK, Chapter 4, 83-104, 2011. doi:10.4203/csets.27.4 and W.E. Johnston, C. Guok, E. Chaniotakis, “Motivation, Design, Deployment and Evolution of a Guaranteed Bandwidth Network Service”, ESnet and Lawrence Berkeley National Laboratory, Berkeley California, U.S.A., In TERENA Networking Conference, 2011. http://es.net/news-and-publications/publications-and- presentations/ .

[43] http://fasterdata.es.net/fasterdata/science-dmz/

[44] J. Zurawski, “Say Hello to your Frienemy – The Firewall.” Available at http://fasterdata.es.net/fasterdata/science-dmz

[45] “Achieving a Science ‘DMZ’”.http://fasterdata.es.net/assets/fasterdata/ScienceDMZ-Tutorial-Jan2012.pdf and the podcast of the talk at http://events.internet2.edu/2012/jt-loni/agenda.cfm?go=session&id=10002160&event=1223

[46] “NSF to Help Campuses Embrace 'Science DMZ' Strategy”. http://www.es.net/news-and-publications/esnet-news/2012/nsf-to-help-campuses-embrace-science-dmz-strategy

Biographies

William E. (Bill) Johnston

Bill is a Senior Scientist and advisor for ESnet.

He led ESnet for five years and led a reanalysis of the requirements of DOE's science programs that ESnet supports. As a result of this, a new network architecture and implementation approach were defined that would accommodate the massive data flows of large-scale science like that of the LHC. This new network was built in 2007 and 2008.

Bill ran the LBNL Distributed Systems Department and worked on many projects related to the application of computing in science environment s. He co-founded the Grid Forum with Ian Foster and Charlie Catlett.

For more information see www.dsd.lbl.gov/ ~ wej

Michael Ernst

From 2002 Dr. Ernst had a leading role in software research and development for the U.S. CMS Software and Computing project to enable wide-area distributed computing, which allowed connecting large compute clusters, storage facilities, etc., across North America. In 2006 Dr. Ernst was the CMS Computing Integration Coordinator. In 2007 Dr. Ernst was appointed Director of the RHIC and

ATLAS Computing Facility at Brookhaven National Laboratory, which are two leadership class scientific computing facilities that handle tens of petabytes of data from nuclear and high energy physics experiments in a worldwide user analysis environment .

Eli Dart

Eli Dart is a Network Engineer at the Energy Sciences Network (ESnet) which is the high performance networking facility of the US Department of Energy Office of Science.

Eli's works on high performance networking, network security, and network performance tuning for the use of networks as a tool to enable and enhance scientific productivity through efficient high-speed data movement, easier access to data sets, and enhanced collaboration. Eli has worked on the deployment of several major networks, including SC's SCinet.

Eli has worked in computing and networking since 1995, and has a degree in Computer Science from Oregon State University.

Brain Tierney

Brian Tierney is a Staff Scientist and group leader of the ESnet Advanced Network Technologies Group at Lawrence Berkeley National Laboratory. His interests include high-performance networking and network protocols; distributed system performance monitoring and analysis; network tuning issues; and the application of distributed computing to problems in science and engineering. He has been the PI for several DOE research projects in network and Grid monitoring systems for data intensive distributed computing. Mr. Tierney has an M.S. in Computer Science from San Francisco State University, and a B.A. in physics from the University of Iowa.