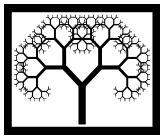
Paper 2



Civil-Comp Press, 2012
Proceedings of the Eighth International Conference on Engineering Computational Technology,
B.H.V. Topping, (Editor),
Civil-Comp Press, Stirlingshire, Scotland

Addressing the Problem of Data Mobility for Data-Intensive Science

W.E. Johnston, E. Dart and B. Tierney
ESnet and Lawrence Berkeley National Laboratory
Berkeley California, United States of America

Abstract

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

The LHC data analysis is highly distributed and this arises from the fact that the two major experiments (ATLAS and CMS) at the LHC each have a large collaboration community (more than 2900 scientists from 172 institutes in 37 countries work on the ATLAS experiment) that is scattered across Europe, N. America, and Asia, and this is increasingly the norm as scientific instruments get bigger and more expensive.

The LHC physics community has dealt with this problem methodically for a relatively long time. Their experience will be very useful for other data-intensive science disciplines, especially in the areas of highly distributed data management systems and data movement systems, and the high performance use of the intervening networks.

The distributed systems such as that used by ATLAS collaboration for data movement and analysis require network performance that is predictable like the other managed resources. This is needed for the smooth overall functioning of the system. This has given rise to a new network service that essentially provides a "virtual circuit" between specified end points, that had a guaranteed bandwidth, and that could be requested for some specific time interval in the future. This service is used to reliably interconnect the elements of the distributed systems.

On the other hand, the Internet is full of undetected "soft errors." While TCP works with such errors, over long distances they force it to operate at speeds 10-100 times slower than the link capacity.

Moving very large volumes of data over international distances requires that the network be error-free. Achieving error-free operation of the network requires constant testing and monitoring, and an international infrastructure has been put in place to do this.

Once the testing and monitoring were in place in the R&E infrastructure, it became apparent that the campuses and even the wide area R&E networks were not

well designed for moving massive amount of data.

To address the campus problem, a "Science DMZ" part of the campus network was designed that is optimized for and serves only high-performance science applications.

The wide area problems were addressed with overlay network that isolated the science traffic from the general traffic.

This paper discusses all of these issues.

Keywords: data-intensive science, large-scale, widely distributed systems, impact on the R&E Internet, moving massive quantities of data internationally, TCP is a "fragile workhorse", federated network testing and monitoring, high-throughput campus LANs, a new Internet architecture for data-intensive science.

1 The origins and common problems of data-intensive science

A collection of 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. The experiments of the LHC are centered on six detectors [2], the two largest of which are ATLAS [3] and CMS [4].

These detectors are built from thousands of tons of silicon sensors and supporting apparatus, and 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, that make up neutrons and protons. (In particular, these detectors are looking for the predicted Higgs Boson [5], which is thought to be responsible for the property of mass.)

The output of these detectors is on several thousand optical fibers, both analog and digital, at data rates that amount to about a petabyte per second. Almost all of this data is noise, and the important science – e.g. discovery of the Higgs boson – is buried in this noise. This intractably large stream of data is filtered through successive hardware and software "triggers" designed to eliminate as much noise as possible without discarding the data of interest. What remains is the data that physicists around the world carefully analyze to try and extract the very elusive science.

The filtered stream of data the physicists analyze from each of the two big detectors is of order 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 the decade that the LHC HEP community has been developing, perfecting, and tuning the data management systems, data movement systems, and the intervening networks, 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: climate research; next generation radio telescopes that will cover thousands of square kilometers (e.g. the SKA); genomics research that has seen an exponential decrease in the cost of genome sequencing machines and an associated exponential increase in their number and therefore in the data that has to be managed, etc. (These other science uses are discussed briefly in section 8.)

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.

In both cases, similar problems arise in the data management techniques and in their supporting infrastructure. Obviously not all of the problems are shared by all of the science communities due to differences in instrumentation and in the process of science. However, important commonalities show up in some of the various sub-systems that implement the data management for many disciplines.

In this paper we examine some of the problems and how they have been addressed. We use the LHC data handling as an exemplar, however this is not meant to imply that all of the areas of science that have to address managing massive amounts of data will do so in the same way. Never-the-less, there are significant area of commonality, especially in the underlying technologies, and the LHC systems provide some concrete examples of how these problems were addressed to produce what is the science world's largest distributed computing and data management system. We refer to these sorts of science as "data-intensive science" and in this paper look in particular at the data mobility issues faced by data-intensive science. (For a more detailed discussion of some of these topics see [6].)

2 Changing network usage patterns

Since about 2004 there has been a fundamental change in the traffic patterns of the Research and Education (R&E) networks of the world. This change is due to dramatically increasing volume of data in associated flows in the network – i.e. groups of flows between a small number of sites. In 2004 neither the change nor its root cause were immediately obvious, however by mid-2005 the essential characteristics of the changes were clear, and sustained large data flows were becoming a significant fraction of the total network traffic. (ESnet [7], one of the

two largest R&E networks in the U.S., is used in the examples[a].) This is illustrated in Figure 1, where 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.

From a networking point of view, these changes in traffic patterns negated the dominance of huge numbers of small flows that, up to this point, had resulted in smooth variations and predictable changes in the overall utilization of the network. One of the hints that something was changing was that the overall network traffic was becoming much more irregular than in the past, and R&E networks were not designed with this new traffic pattern in mind. This is illustrated in Figure 1, where the traffic from a single site is compared to the total traffic in ESnet, clearly demonstrating that the large flows are responsible for the large variations in the overall network traffic.

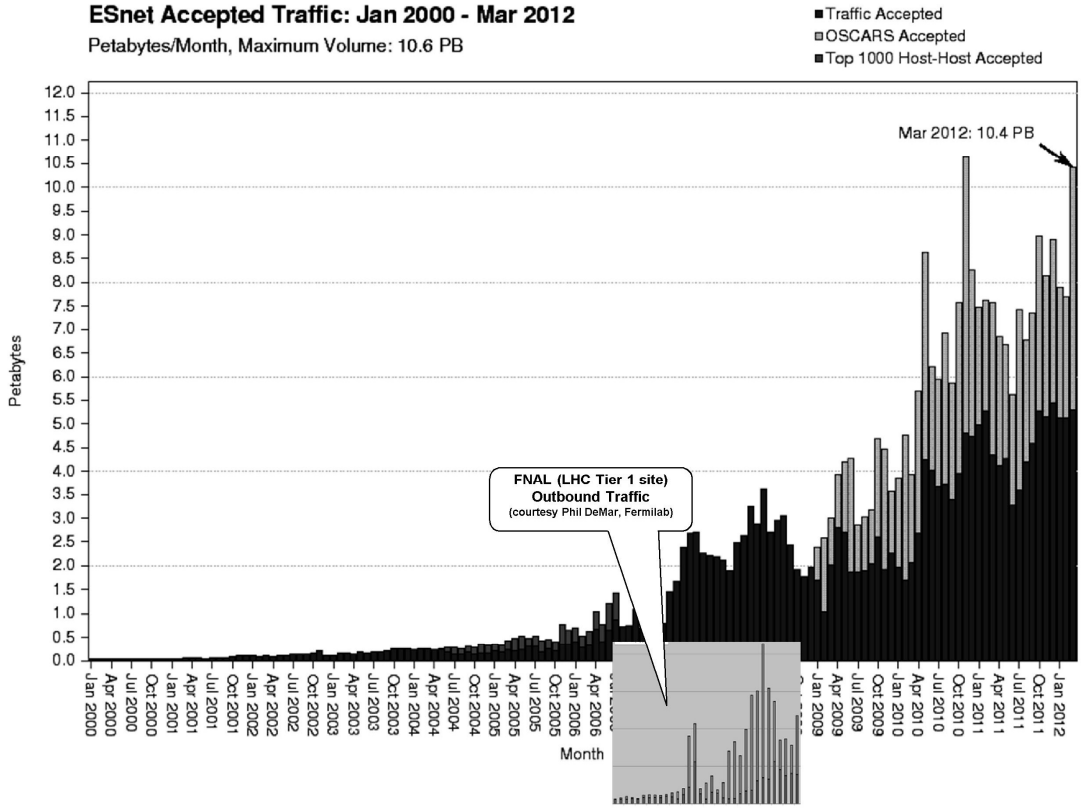


Figure 1. Evolving traffic patterns and the impact of small numbers of very large data flows on overall network traffic.

[a] 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 mostly large research institutes that generate high-volume data streams. While universities also generate high-volume data streams they have a higher proportion of commodity-like traffic.

It became clear that the driver of this change was an increasing use of the network by large-scale, data-intensive science projects, in particular those of the High Energy Physics (HEP) community.

From the time that this change was first noticed, this sort of use of the network has evolved and increased rapidly. The HEP community is now using the network for production data movement in support of the Large Hadron Collider (LHC) [1] at CERN, and several other large science projects are now doing the same sort of thing.

3 Moving 500 Terabytes/day between fifty international sites

The LHC data analysis is highly distributed and this arises from the fact that the two major experiments (ATLAS and CMS) at the LHC each have a large, international collaboration community. The network data transfers are driven by the fact that the collaborators provide most of the considerable computing capacity needed to do the data analysis[a]. The data used by the ATLAS analysis system resides in some 11 national data centers that the analysis systems (35 centers in 14 countries) access to move data to local caches (which themselves are frequently distributed) where the analysis programs operate on the data. (While ATLAS is used as the exemplar here, CMS resources and data are comparable in every way.)

The analysis systems are sophisticated workflow management systems that couple distributed databases, distributed schedulers, and distributed computing resources into a coherent system that executes tens of thousands of analysis jobs/day, which consumes terabytes of data/day from the distributed databases. This has resulted in the ATLAS analysis system moving some 7 petabytes over a seven month period in 2010 and sustaining world-wide data flows of some 350-500 terabytes/day (Figure 2). The smooth functioning of this large, distributed system is necessary to keep up with the data generated by the detector at the LHC.

During the process of designing, building, and testing these systems that manage and analyze their data, the HEP community was addressing the problem of data mobility and transfers: They developed techniques for automating the management of their highly distributed data analysis; they defined the network services needed to access the distributed data, and; they were learning how to tune the data transfer systems to be able to drive the network at high-speeds over intercontinental distances, etc. PanDA, the ATLAS version of such a system, is illustrated in Figure 3.

This system is routinely capable of running 55,000 to 65,000 simultaneous jobs that cause data movement over international distances of 350-500 terabytes/day. It is this scale of data movement and analysis jobs (Figure 2), 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

[a] These institutions contribute large numbers computing systems, disk farms, and tape systems – currently providing about 28,000 multi-core computers, 39 petabytes of disk, and 50 petabytes of tape storage. [39] (These numbers are for 2010 and are projected to increase by 75-100% over the next several years.)

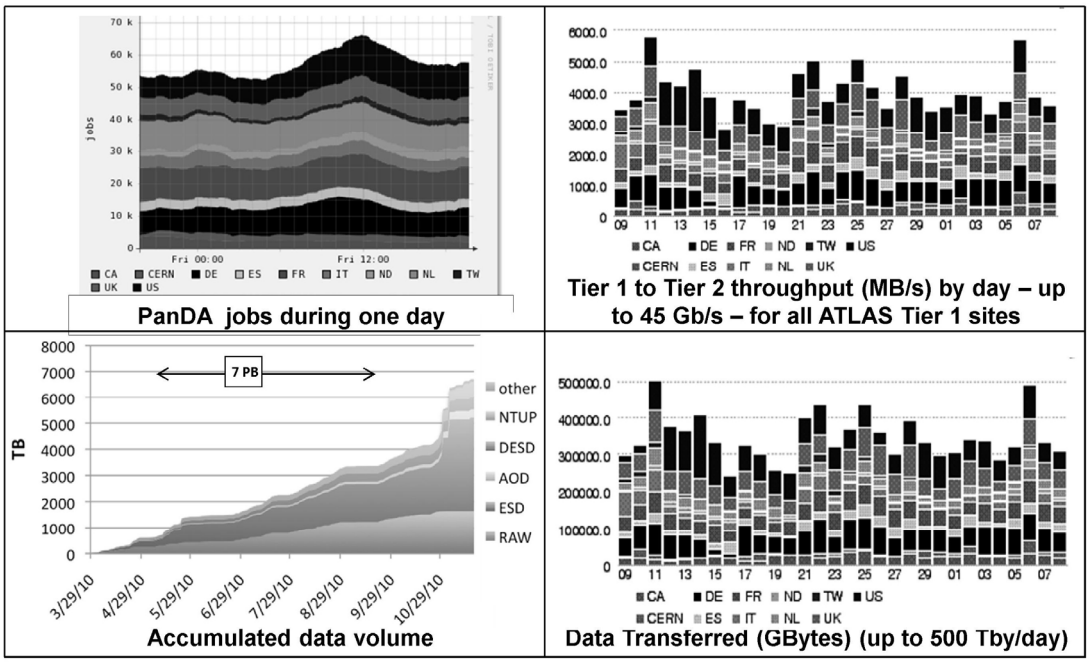


Figure 2. The scale of ATLAS data analysis. [40]

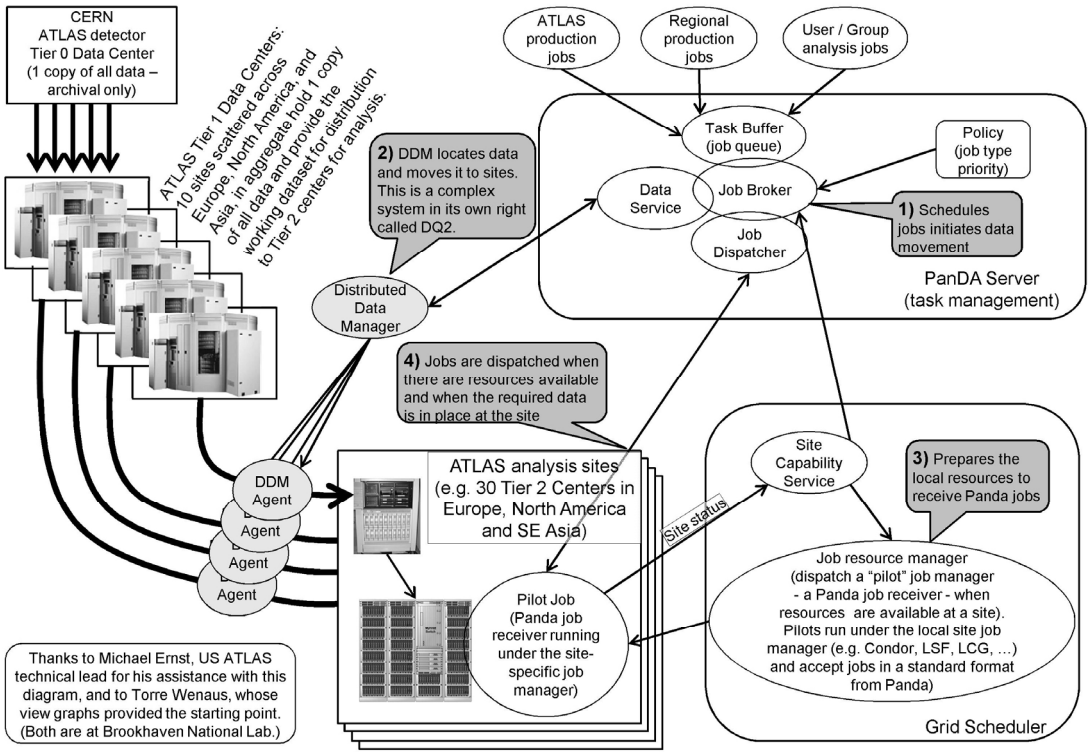


Figure 3. ATLAS PanDA (Production and Distributed Analysis) data management and analysis system. See [8] and [9].

There is a collection of technologies that must operate in concert to make this sort large-scale data movement, distributed system possible, and these are the topics of the remainder of this paper.

4 New network services

The 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 needed 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.

The network service that was developed to meet the requirements was a service that essentially provided a "virtual circuit" between specified end points, that had a guaranteed bandwidth, and that could be requested for some specific time interval in the future.

The number of network services offered to users is small. These include, for example, TCP for reliable data transport service with undefined performance characteristics, UDP for unreliable packet service, RTP for streaming media, DNS for name to address mapping, etc. The new guaranteed virtual circuit service is the first new network service to be added at the user level for some time.

This service was initially developed by ESnet for its internal use and is called OSCARS. Subsequently it was adopted in several other networks, and an international collaboration called DICE [10] 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[a] involved in the LHC data transport. Such compatible services are now deployed many of the R&E networks in Europe, the Americas, and Asia. The ad-hoc DICE effort – the InterDomain Controller Protocol (IDCP) [11] – is now moving into the Open Grid Forum (OGF) standards organization in the Network Service Interface working group [12].

The characteristics of ESnet's OSCARS service, which are similar to other, compatible services, are guaranteed, reservable bandwidth for a particular start and end time; resiliency (explicit backup paths can be requested); data transport via either layer 3 (IP) or layer 2 (Ethernet) circuits, and integrity of the established circuits [13].

Traffic isolation is 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 underlying mechanisms allow for traffic management by network operators, which means that explicit paths can be used to meet specific requirements

[a] A network domain is the infrastructure operated by a particular organization. ESnet, Internet2, GÉANT, etc. are all network domains. Each domain is an independent entity with its own policies, and engineering and operations staff.

– e.g. bypassing congested links, using higher bandwidth paths, explicit engineering of redundancy, etc.

The service provides for secure connections: The circuits are "secure" to the edges of the WAN (Wide Area Network) 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 (which, in practice, are largely the same) then the circuits do not have to transit the site firewall because the far end is known and the connection cannot be intercepted.

Flexible service semantics (not available in all implementations) provide for a user exceeding the requested bandwidth if the path has idle capacity – even if that capacity is committed (though unused). Traffic in different circuits can also be assigned different queuing priorities.

One important implication of these service semantics is that, in combination with priority based routing, provides the tools to allow sites (especially big sites like the Tier1data centers) to do their own traffic engineering on the circuits that connect them to the WAN. In particular, sites that have available multiple paths (e.g. multiple fibers) to the WAN can decide ahead of time how the bandwidth available during degraded operation (one or more circuits or paths down) is to be re-purposed. (See [14].)

5 Monitoring and testing are critical

At the same time that the network transport service was being developed, it became clear that some networks that thought that they were providing good service were, in fact, not. It was found that the Internet is full of undetected "soft errors", where TCP works, but over long distances – and therefor high latency paths – only at speeds 10-100 times slower than the link capacity. 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 other data paths that crossed many different network domains.

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 – very small packet loss rates result in large decreases in performance. 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. 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.

TCP is sensitive to packet loss because of changes made to avoid the congestion collapse of the Internet. (See [15].) While these modifications were necessary (and still are), they make TCP perform poorly in high-performance environments with soft errors. 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 limited success. Therefore, in the near to medium term, data-intensive science network infrastructure

must be maintained error-free in order to provide the necessary support for TCP such that TCP-based applications can 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.

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 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 [16].

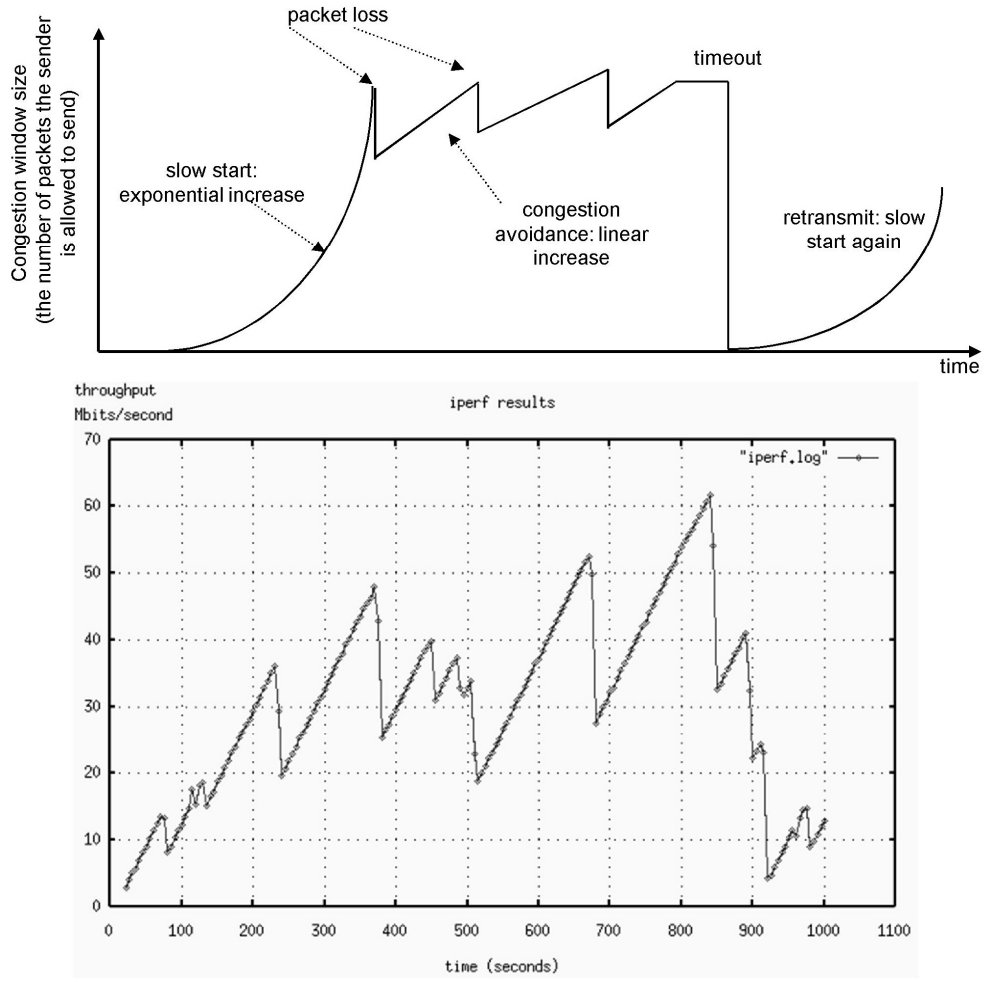


Figure 4. Top: TCP congestion avoidance in action; bottom: impact on throughput. (The path is LBL to CERN (Geneva) (RTT = 150 ms), OC-3 (155 Mb/s), in 2000.)

[a] The TCP window specifies the amount of data that can be in "flight" in the network before being acknowledges 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 [22].)

To see why TCP is so sensitive to packet loss we offer this brief overview. TCP begins a connection in a phase called Slow Start, where the data rate is gradually increased. Since TCP interprets packet loss as network congestion, when TCP encounters packet loss it assumes that there is a congested link in the network between the sending and receiving hosts and reduces its sending data rate. The reduction in sending rate varies depending on the congestion control algorithm used, but typical reduction values are 20% and 50%. After such an event, the sending rate is again gradually increased. If network conditions are such that packet loss is encountered early in slow start or loss events occur frequently, such as with failing optics, misconfigured switches, or other soft failures, TCP can spend all its time reducing its sending rate and slowly ramping back up until it encounters loss again. In these situations TCP (and by extension the data transfer nodes and the network) typically never comes close to achieving the transfer rate that would be possible in the absence of the loss. This is a common cause of poor performance.

The only way to keep multi-domain, international scale networks error free is to test and monitor continuously end-to-end.

perfSONAR is a network monitoring framework [17] designed to collect both passive and active network measures, convert these to a standard format [18] and then 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 that measure data transport throughput. Packet delays are achieved by measuring time-of-flight using a precision clock at both ends (typically a GPS clock, or comparable or derived time source). These tools are and OWAMP [19] and HADES [20]). 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 [21] and [22].

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" so that trends (e.g. due to increasing soft failures that indicate developing hardware problems) can be detected.

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

The major services provided by the perfSONAR architecture are:

- Measurement Point Service: Creates and/or publishes monitoring information related to active and passive measurements
- Measurement Archive Service: Stores and publishes monitoring information retrieved from Measurement Point Services
- Lookup Service: Registers all participating services and their capabilities
- Authentication Service: Manages domain-level access to services via tokens
- Transformation Service: Offers custom data manipulation of existing archived measurements
- Topology Service: Offers topological information on networks

PerfSONAR is designed for federated operation. Typically each domain (e.g. ESnet, Internet2, GÉANT, etc.) maintains its own Measurement Archive and has control over what data is published into it and who can query data from it.

Published data is federated by tools that discover the Measurement Archives, 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 [23].)

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

By mid-2010 a predictable circuit service ("network as a service") and monitoring and testing services more or less in place. However, the utilization of the network increasing dramatically as the LHC came on line, several more problems emerged.

It quickly became apparent that once you provide high quality, high performance data transfer capability in the wide area networks, then you run into bottlenecks in the general R&E networks that were not architected for this sort of data-intensive use, as well as in the campus[a] networks that are not architected to move large volumes of data into and out of the campus to the WAN.

6 Changes in campus computing systems and network architecture

Campus networks are, in general, neither designed for nor capable of supporting the data movement of data-intensive Science (DIS).

In order to address the campus-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. This is accomplished with a new campus network architecture called the "Science DMZ" [24].

The site network – the LAN – typically provides connectivity for user systems and local resources – computers, data servers, instruments, collaboration systems, etc.

In most cases, a site LAN has multiple missions. It must provide for the traffic of general users, the "mundane" business aspects of running a scientific/educational institution. It must support computer security directives for the protection of financial and personnel data (and the avoidance of embarrassing news headlines detailing security breaches), and (sometimes as an afterthought) it must support the pursuit of scientific discovery. These missions of a site network are often in direct

[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.

competition because they have very different network requirements. The LHC Tier 1 and Tier 2 sites provide examples of the DIS scenario. (The Tier 1 centers tend to be at national laboratories / research institutes and most of the Tier 2 centers are on university campuses.)

The devices and configurations typically deployed to build networks for business or general user access purposes usually don't work for DIS. Firewalls delay packets which can look like packet loss to high-speed flows, firewalls that support 10Gb/s interfaces may only support individual flows of 1 Gb/s or less, proxy servers have similar issues, low-cost switches don't have the buffering required for long RTT, high-speed traffic, and so forth.

The LAN architecture of a site that is involved in DIS must be designed to match the high-bandwidth, large data volume, long round trip time (international paths) wide area network (WAN) flows. Without this the site will impose poor performance on the entire high-speed data path, all the way back to the source.

Data-intensive science resources should be deployed in a separate portion of the network that has a different packet forwarding path and tailored security policy.

In traditional network security architectures, there is the notion of a "demilitarized zone" or DMZ network. This portion of the network contains the site resources that are routinely contacted by off-site systems – authoritative DNS servers, incoming email servers, external web servers, and so forth.

The DMZ is normally located at or near the site network perimeter – the connection between the site LAN network and the border router that connects the site to the wide area network.

A "Science DMZ" is a dedicated portion of a site or campus network, located as close to the network perimeter as possible, that serves only high-performance science applications. The equipment, configuration, and security policy of the Science DMZ are tailored specifically for science applications – not for general-purpose campus or "enterprise" Internet access.

The Science DMZ is an element of the campus network architecture. The intent of the Science DMZ is enable high-performance and data-intensive science applications that rely on high-speed networking for success.

The Science DMZ is connected directly to the border router in order to minimize the number of devices that must be configured to support high-performance data transfer and other scientific applications. Achieving high performance is very difficult to do with system and network device configuration defaults, and the location of the Science DMZ at the site perimeter simplifies the system and network tuning processes. Also, if there is a performance problem, it is much easier to troubleshoot a handful of devices rather than a large-scale LAN infrastructure.

Security for a DIS environment located on the Science DMZ can be tailored for the data transfer systems involved in DIS which typically only runs a well-defined and limited set of special-purpose applications rather than the typical array of user applications. Since the DMZ resources are assumed to interact with external systems and are isolated from, or have carefully managed access to, internal systems, the security policy for the DMZ is tailored for these functions rather than to protect in interior of the general site LAN.

A Science DMZ is built with network components suitable for high throughput science traffic – routers and switches that have packet forwarding hardware capable of steady state, high-speed data transfers, large output buffers to accommodate small "glitches" in the WAN, and can implement security appropriate to the environment (e.g. access control lists).

The Science DMZ has several key elements in addition to the network architecture and security policy. It also includes dedicated systems that are built and tuned for wide-area data transfer, and test and measurement systems for performance verification and rapid fault isolation, typically perfSONAR as described previously.

The computer systems used for wide area data transfers typically perform far better if they are purpose-built and dedicated to this function. These systems, which we call Data Transfer Nodes (DTNs) [6], are typically PC-based Linux servers built with high-quality components and configured specifically for wide area data transfer. The DTN also has access to local storage, whether it is a local high-speed disk subsystem, a connection to a local storage infrastructure such as a SAN, or the direct mount of a high-speed parallel filesystem such as Lustre [25] or GPFS [26], or a combination of these. The DTN runs software tools designed for high-speed data transfer to remote systems – typical software packages include GridFTP [27] and its service-oriented descendent Globus Online [28], discipline-specific tools such as XRootd [29], and versions of default toolsets such as SSH/SCP with high-performance patches applied [30]. (Though even with such patches, SSH is nowhere near as fast as the other tools.)

DTNs typically have high-speed network interfaces (10Gbps as of this writing, though experiments with 40Gbps interfaces are already underway, e.g. at SC11 with 40Gbps RDMA over the WAN [31]), but the key is to match the DTN to the capabilities of the site and wide area network infrastructure. So, for example, if the network connection from the site to the WAN is one gigabit Ethernet, a 10 gigabit Ethernet interface on the DTN may be counterproductive.

In addition, the Science DMZ has a test and measurement host that allows for easy fault diagnosis on the Science DMZ and for end-to-end testing with the collaborating site if they have perfSONAR installed. The perfSONAR host can run continuous checks for latency changes and packet loss using OWAMP, as well as periodic throughput tests to remote locations using BWCTL/Iperf. If a problem arises that requires a network engineer to troubleshoot the routing and switching infrastructure, the tools necessary to work the problem are already deployed – they need not be installed before troubleshooting can begin.

Finally, the default computer system configuration options (e.g. network interface tuning options and TCP parameters) are wrong for high performance data movement. These issues are addressed in knowledge bases, such as the ESnet site fasterdata.es.net, that maintain system tuning information for DIS data transfer nodes, the Science DMZ architecture, network tuning, troubleshooting, etc.

The essential components and a simple architecture for a Science DMZ are shown in Figure 5. (For a more detailed discussion see [32].)

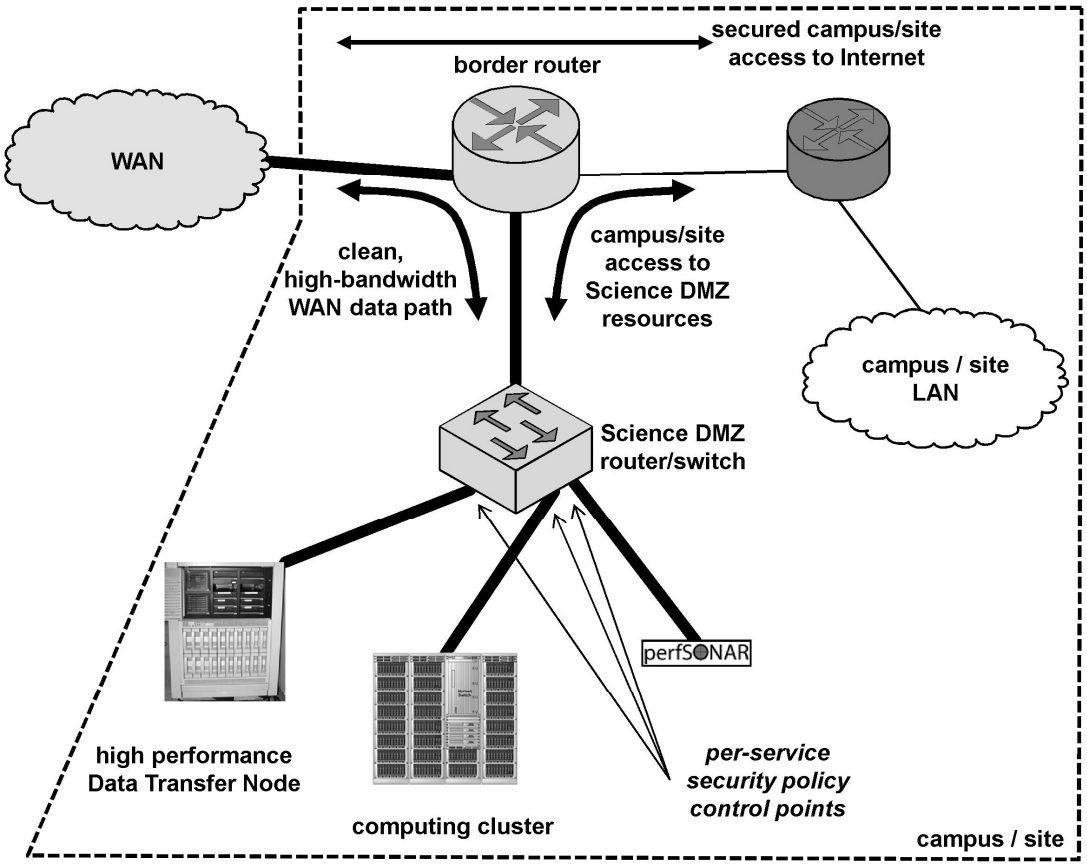


Figure 5. Simple Science DMZ architecture

In the illustrated architecture, the Data Transfer Node (DTN) is connected directly to a high-performance Science DMZ switch or router, which is connected directly to the border router. The DTN's job is to efficiently and effectively move science data to and from remote sites and facilities, and everything in the Science DMZ environment is aimed at this goal. The security policy enforcement for the DTN is done using access control lists on the Science DMZ switch or router, not on a separate firewall. To further mitigate risks, no general-purpose computing tasks are allowed on the DTN – no web browsers, no media players, no business productivity tools such as document and spreadsheet editors, no email clients, etc. which require all the security controls necessary to ensure the safety of general-purpose computing in today's environment. The DTNs are frequently only accessed by automated workflow management systems as illustrated by the LHC/ATLAS/PanDA example in section 3.

The computing systems that consume the data can be similarly configured but do not typically require the WAN performance of the DTN because they move data across the LAN. Even if users at the site access the resources on the Science DMZ through the site perimeter firewall, they will typically get good performance if the firewall can provide the required throughput. The very low latency between the Science DMZ and the on-site users results in the issues caused by the site perimeter firewall being much less of a problem in practical terms. TCP recovers quickly at

low latencies, and short-distance TCP dynamics are different enough from the TCP dynamics in long-distance transfers that packet loss that would exist if the wide area data transfers traversed the firewall may not even exist when local users access Science DMZ resources.

The success of the Science DMZ concept is demonstrated by the fact that a recent National Science Foundation call for proposals recommended that campuses adopt the Science DMZ architecture [33].

7 Changes in the global R&E network architecture

With each improvement in the network infrastructure and services to support data-intensive science, the resulting irregular large volume data flows illustrated in Figure 1 started to make themselves felt in the larger, general R&E infrastructure in ways that are, or potentially are, detrimental to the general R&E community traffic.

To address this problem an international group of R&E network engineers considered how the general architecture of the R&E Internet had to change in order to both provide good service to the science community and not to severely impact the general traffic that shared the same network.

Consider, again, the LHC as an example. From the beginning the massive, steady-state data transfers (about 50 Gb/s, 25 Gb/s for each of ATLAS and CMS, 24 hrs./day, seven days/week for about 10 months of the year) between CERN (Tier 0) and the experiment data centers (Tier 1) were carried over a purpose-built, dedicated network infrastructure called the LHCOPN (LHC Optical Private Network). This data path is a collection of 10Gb/s circuits from CERN to each of the Tier 1 data centers, and can be seen in the upper left of Figure 3. (The ATLAS data centers are located in Canada, Germany, France, the Nordic countries (a distributed Tier 1), Taiwan, the US, Switzerland (CERN), Spain, Italy, the Netherlands, and the UK.)

However, in aggregate, the Tier 2 analysis centers pull a comparable amount of data out of the Tier 1 data centers in order to do the physics analysis. Initially the Tier 2 sites were expected to get data from local (in the same country) Tier 1 centers so traffic would be confined to the national R&E infrastructure. It turned out that as soon as the physicists found out that they could access all of the data (each Tier 1 center only holds a fraction of the total data – in aggregate they have one complete copy) the Tier 2 sites began pulling data from all over the world. This had a disruptive impact on the transatlantic circuits in particular, because the capacity on those circuits was relatively small compared to the terrestrial R&E national networks.

The community approach to this issue has focused on the LHCONE project (lhcone.net), which aims to modify the R&E WAN architecture so that large science data flows can be isolated and/or managed separately from the general traffic.

LHCONE is essentially a collection VPN-like “islands,” each typically in a single network domain, that are interconnected to provide a global, private infrastructure for large-scale science. End user sites connect to the islands (that are frequently provided by the same network that provide the campus WAN connectivity) and the islands connect to each other over private virtual circuits. The

routing within and between the islands is such that all of the connected sites can communicate with each other.

These VPN islands – implemented with a technology called "virtual routing and forwarding" (VRF) – form a private and isolated part of the Internet using network infrastructure that is suitable for large data movement. Where the circuits of this infrastructure are shared with general R&E traffic, the two types of traffic can be separated and managed by network operators so as not to interfere with each other.

The islands are interconnected in a similar way – on infrastructure that is dedicated or shared in a carefully managed way.

Building LHCONE has required an unprecedented level of cooperation and coordination among a significant fraction of all of the world's major R&E networks. The LHCONE infrastructure illustrated in Figure 6 is largely in place and the user community is, as of April 2012, starting to gain experience in the use and behavior of this infrastructure.

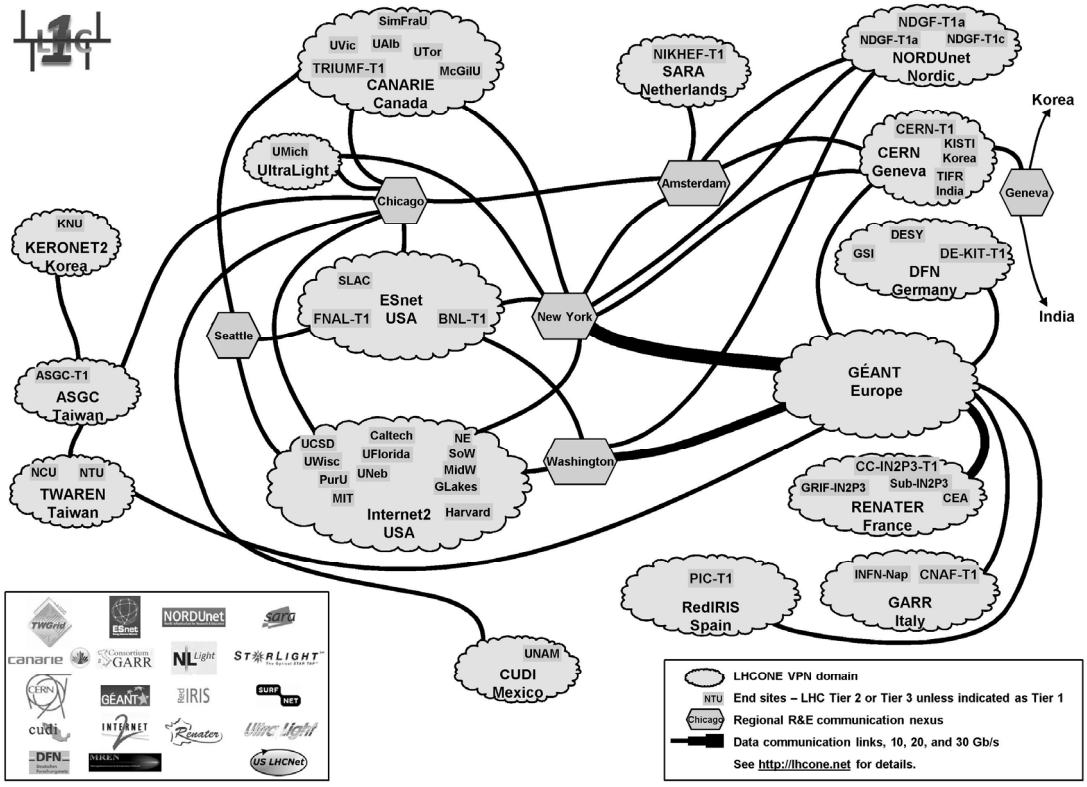


Figure 6. LHCONE: A global infrastructure for the LHC Tier1 data center – Tier 2 analysis center connectivity

8 The problem of data-intensive science in other disciplines

Several other scientific fields are now following in the footsteps of physics.

The cost of genome sequencing machines is falling dramatically while the volume of data produced is increasing rapidly. See [34] and [35]. The data must be

moved from the biology labs to data centers for processing and the resultant finished genomes entered into community databases.

Climate scientists must analyze observational data and model data, and these data – comparable in size to the LHC data – are housed at centers around the world. The requirement for productive access to those data sets has resulted in the construction of the Earth System Grid – a global data workflow infrastructure that allows climate scientists to access data sets housed at modeling centers on multiple continents including North America, Europe, Asia, and Australia. See [36].

New instruments are being deployed at X-ray synchrotrons that generate data at unprecedented resolution and at unprecedented rates – the current generation of instruments can produce 300MB/sec or more, and the next generation will be able to produce data volumes many times higher [35].

Future large-scale science projects that are data-intensive and that are currently being planned and built include, for example, ITER (the international fusion energy prototype [37]) and the Square Kilometer Array [38] – a massive scale radio telescope that will generate as much or more data than the LHC.

In all these cases, scientific productivity is governed by the ability to analyze large volumes of data with computers – because without the ability to effectively analyze the data, the science is difficult or impossible. Further, since it is often physically, politically, or organizationally impossible to put all the data storage and computational analysis resources necessary for conducting the science at the location where the data are generated, plans must be made for the distribution of large volumes of data between instruments, facilities, and analysis resources in the general case.

9 A Strawman model for an architecture of data-intensive science

One of the remaining, important pieces of a data-intensive science infrastructure is not discussed here. This is the workflow systems that are essential for both managing the analysis of massive, distributed data, and workflow for the movement of the underlying data.

The data movement workflow system is likely to be as critical for effective DIS as the topics discussed here. (In the ATLAS PanDA example (Figure 3), the data movement workflow is handled by the PanDA server and the Distributed Data Management (DDM) system.) See [6] for more information on this topic.

The view of data-intensive science presented here is a “bottom up” view, much of which resulted from years of experience by the ESnet engineering staff trying to help the ESnet user community make more effective use of high-speed networks. This experience has resulted, among other things, in the fasterdata.es.net knowledge base. Much, though certainly not all, of the driver for this has been the LHC community’s commitment to using the network as an integral part of their science process: There is no other practical way to manage the volume of steadily produced data that they have to analyze.

However, is noted, other science communities will and are finding the experience and tools of the LHC community to be important in how they conduct their science.

Acknowledgements

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 with the University of California.

Notes and References

[1] LHC - The Large Hadron Collider Project. http://lhc.web.cern.ch/lhc/general/gen_info.htm
[2] "The LHC Experiments" http://public.web.cern.ch/public/en/lhc/LHCExperiments-en.html
[3] See http://public.web.cern.ch/public/en/lhc/ATLAS-en.html , http://atlas.ch/pdf/atlas_factsheet_4.pdf , and http://www.atlas.ch/atlas_brochures_pdf/tech_brochure-11.pdf
[4] http://public.web.cern.ch/public/en/lhc/CMS-en.html
[5] "About the Higgs Boson" http://cms.web.cern.ch/node/1187
[6] "Infrastructure for Data-Intensive Science – a bottom-up approach," Eli Dart and William Johnston, Energy Sciences Network (ESnet), Lawrence Berkeley National Laboratory. To be published in *Future of Data Intensive Science*, Kerstin Kleese van Dam and Terence Critchlow, eds., Taylor & Francis Group, Publishers. Available at http://es.net/news-and-publications/publications-and-presentations.
[7] See http://www.es.net/about/ , http://www.es.net/about/our-network/ , http://www.es.net/about/esnet-history/ , and http://www.es.net/about/science-requirements/
[8] 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. Available at http://iopscience.iop.org/1742-6596/119/6/062017 .
[9] T. Maeno, "PanDA: Distributed production and distributed analysis system for ATLAS." Computing in High Energy and Nuclear Physics (CHEP), 2007. Available at http://iopscience.iop.org/1742-6596/119/6/062036
[10] DICE (DANTE-Internet2-CANARIE-ESnet) Collaboration. http://www.GÉANT2.net/server/show/nav.1227
[11] InterDomain Controller Protocol (IDCP). See http://www.controlplane.net/
[12] Network Services Interface (NSI) working group, Open Grid Forum, http://ogf.org/gf/group_info/view.php?group=nsi-wg

[13] “Intra and Interdomain Circuit Provisioning Using the OSCARS Reservation System.” Chin Guok; Robertson, D.; Thompson, M.; Lee, J.; Tierney, B.; Johnston, W., 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. Available at http://es.net/news-and-publications/publications-and-presentations/
also see
“Network Services for High Performance Distributed Computing and Data Management,” W. E. Johnston, C. Guok, J. Metzger, and B. Tierney, ESnet and Lawrence Berkeley National Laboratory, Berkeley California, U.S.A. The Second International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering,12-15 April 2011, Ajaccio - Corsica – France. Available at http://es.net/news-and-publications/publications-and-presentations/

[14] “Motivation, Design, Deployment and Evolution of a Guaranteed Bandwidth Network Service,” William E. Johnston, Chin Guok, and Evangelos Chaniotakis. ESnet and Lawrence Berkeley National Laboratory, Berkeley California, U.S.A. In TERENA Networking Conference, 2011. Available at http://es.net/news-and-publications/publications-and-presentations/ .

[15] Congestion collapse of the Internet first occurred in October 1986, when the NSFnet phase-I backbone dropped three orders of magnitude from its capacity of 32 kbit/s to 40 bit/s. Van Jacobson, who ran the network research group at Lawrence Berkeley Laboratory, identified the problem and proposed modifications to TCP to address the problem. (Those modifications are what are described in general terms in the text that points to this citation. Also 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/).

[16] E.g. see http://fasterdata.es.net

[17] See “perfSONAR Technical Overview,” http://www.perfsonar.net/overview.html

[18] See “OGF Network Measurement Working Group (NMWG),” http://nmwg.internet2.edu/

[19] “One-Way Ping (OWAMP),” http://www.internet2.edu/performance/owamp/
[20] “HADES - Hades Active Delay Evaluation System,” Web page at WiN Labor Erlangen, http://www.win-labor.dfn.de/English/mainpage.html
[21] http://fasterdata.es.net/assets/fasterdata/Tierney-bulk-data-transfer-tutorial-Sept09.pdf
[22] http://www.nanog.org/meetings/nanog43/presentations/Dugan_Iperf_N43.pdf
[23] See, e.g., 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 .
[24] http://fasterdata.es.net/fasterdata/science-dmz/
[25] http://www.lustre.org
[26] http://www.ibm.com/systems/software/gpfs
[27] http://www.globus.org/toolkit/docs/latest-stable/gridftp/
[28] See “Why Use Globus Online?” https://www.globusonline.org/whygo/
[29] “XRootD” http://xrootd.slac.stanford.edu/
[30] “High Performance SSH/SCP - HPN-SSH.” Chris Rapier (Pittsburgh Supercomputer Center), Michael Stevens (Carnegie Mellon University), Benjamin Bennett (PSC). http://www.psc.edu/networking/projects/hpn-ssh/
[31] E.g. see “Demonstrating RDMA Protocols over the WAN,” Paul Grun, Chief Scientist, SystemFabricWork, which describes a demonstration done at SC11 (http://sc11.supercomputing.org/). http://members.infinibandta.org/kwspub/home/Demonstrating_RDMA_Protocols_Over_a_40Gbs_WAN.pdf
[32] See “Achieving a Science ‘DMZ’” at 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
[33] “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 .
[34] See http://www.genome.gov/sequencingcosts/
[35] See “BES (Basic Energy Sciences) Network Requirements Workshop, September 2010.” Available at http://www.es.net/about/science-requirements/reports/
[36] See http://www.earthsystemgrid.org/about/overview.htm
[37] See http://www.iter.org/
[38] “The SquareKilometre Array.” By Peter E. Dewdney, Peter J. Hall, Richard T. Schilizzi, and T. Joseph L. W. Lazio. Proceedings of the IEEE, Vol. 97,No. 8, Available at http://www.skatelescope.org/publications/
[39] This number is based on http://lcg.web.cern.ch/LCG/Resources/WLCGResources-2010-2012_04OCT2010.pdf, which gives the CPU resources in terms of HEP-SPEC06 units - the new HEP-wide benchmark for measuring CPU performance. Modern systems seem to be about 8 HEP-SPEC06 per core, so a quad core system will deliver about 32 HEP-SPEC06. From this the number of

computing systems involved is estimated, assuming an average of 4 cores / system.

[40] 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.