HPC's Pivot to Data

S. Parete-Koon*, B. Caldwell*, S. Canon†, E. Dart‡,
J. Hick†, J. Hill*, C. Layton*, D. Pelfrey*,
G. Shipman*, D. Skinner†,H.A. Nam*, J. Wells*, J. Zurawski‡
*Oak Ridge Leadership Computing Facility
Oak Ridge National Laboratory
Oak Ridge, TN, USA
Email: {paretekoonst,blakec,hilljj,laytoncc,namha,wellsjc}@ornl.gov
†National Energy Research Scientific Computing Center
Lawrence Berkeley National Laboratory
Berkeley, CA, USA
Email: {scanon,jhick,deskinner}@lbl.gov
‡Energy Sciences Network
Lawrence Berkeley National Laboratory
Berkeley, CA, USA
Email: {dart,zurawski}@es.net

***Abstract*—Computer centers such as NERSC and OLCF have traditionally focused on delivering computational capability that enables breakthrough innovation in a wide range of science domains. Accessing that computational power has required services and tools to move the data from input and output to computation and storage. A "pivot to data" is occurring in HPC. Data transfer tools and services that were previously peripheral are becoming integral to scientific workflows. Emerging requirements from high-bandwidth detectors, high-throughput screening techniques, highly concurrent simulations, increased focus on uncertainty quantification, and an emerging open-data policy posture toward published research are among the data-drivers shaping the networks, file systems, databases, and overall compute and data environment. In this paper we explain the pivot to data in HPC through user requirements and the changing resources provided by HPC with particular focus on data movement. For WAN data transfers we present the results of a study of network performance between centers.**

***Keywords*-Data, WAN, Workflows**

I. Introduction

Research agendas often start or end with data as a crucial component to the scientific discovery process. This is true for both simulation and experiment with examples in every science domain, including particle physics [1], [2], cosmology [3], [4] climate science [5], [6], photon science [7], [8] and materials discovery [9], [10]. The boundaries of the scientific computing ecosystem are being pushed beyond single centers providing high-performance computing resources or any single experimental facility. This expansion is a result of the growing data generated as high-performance computing (HPC) reaches toward exascale and the availability of data from large instruments such as telescopes, colliders, and light sources grows exponentially.

Big Data is considered the fourth paradigm of scientific discovery along with experiment, theory and simulation. Data and data-intensive computing are expected to lead to new scientific knowledge and actionable insight. However, this focus on data does not imply a reduction in importance of large-scale simulation capabilities, rather an expansion of the scientific computing ecosystem to enable new scientific discovery through the integration of data from various sources, both simulation and experiment. To realize this new scientific opportunity, complex intra- and inter-facility workflows are necessary to move data, launch computations, provide analysis to shape experiments in real-time, archive data, and make them available to a broader community. These complex workflows require coordination between multiple facilities, common standards and software, and reliable infrastructure.

The Oak Ridge Leadership Computing Facility (OLCF) at Oak Ridge National Laboratory and the National Energy Research Scientific Computing Center (NERSC) provide some of the largest high-performance computing and storage resources in the world to enable unprecedented scientific breakthroughs. HPC centers are uniquely positioned to provide, in addition to the large-scale compute resources, data-centric resources, infrastructure, and services. Data management capabilities required by the scientific community include data redundancy, accessibility to a broader community, and long-term storage. Also, HPC centers can facilitate the new science that can be achieved from complex analysis and integration of data from various sources. ORNL and NERSC are working towards providing the high-reliability and high-performance resources and infrastructure to support and accelerate such data science discoveries. At ORNL these data-centric resources are delivered today via the Oak Ridge

Compute and Date Environment for Science (CADES). CADES provides a broad set of computatational and data infrastructure coupled with "data services" serving projects from multiple scientific domains, from materials science, to earth system modeling.

Scientific workflows increasingly involve multiple locations for data storage and processing [11]. Predicting how and why scientific workflows achieve their observed end-to-end performance is a growing challenge for scientists and network engineers. While the value of data comes primarily in the data analysis, data movement is often the underlying enabling mechanism for many parts of a data intensive workflow. The exchange of data between various sites is a significant challenge, and optimized data transfers are a necessity to ensure scientific productivity. Many tools and resources, such as high-bandwidth Wide Area Networks (WAN), specialized data transfer servers and WAN-optimized data transfer software, connect HPC centers with each other and the other remote endpoints of the multi-location HPC workflows. Data movement between HPC centers is also exemplary of workflow elements that require coordination between centers. For this reason we present an analysis of center-to-center data transfer over WAN and describe best practices so that users can predict the impact of data movement in their scientific workflows.

We discuss in Section II the scientific discovery ecosystem currently provided at ORNL and NERSC and the motivation for HPC centers to address data management rather than simply providing storage. Section III describes some of the complex workflows in the scientific ecosystem. Section IV discusses the current infrastructure for data movement and results for center-to-center transfer tests are given in Section V. Section VI discusses future data roadmaps and conclusions.

II. Motivation

The Oak Ridge Leadership Computing Facility (OLCF) and the National Energy Research Scientific Computing Center (NERSC) are two U.S. Department of Energy (DOE) Office of Science User Facilities providing high performance computing and data resources to the scientific community. The OLCF at Oak Ridge National Laboratory (ORNL) provides capability computing resources and is home to Titan [12], the world's second fastest supercomputer, a 18,688 compute node Cray XK7 system with an aggregate peak speed of approximately 27 petaflops (PF). NERSC is the Office of Science's primary scientific computing facility, providing capacity computing resources, including Hopper, a 6,384 compute node Cray XE6 with a peak performance of 1.28 PF and recently added Edison, a Cray XC30 with a peak performance of 2.57 PF. These facilities are connected through the DOE Energy Sciences Network (ESnet) [13], a high-bandwidth network providing reliable connections to over 40 national laboratories, research institutions, and universities.

Table I
Data Archived on HPSS at the OLCF as of March 2014

TBs per user	Number of Users
Less that 1TB	940
1 to 10 TB	421
10 to 100TB	245
100TB 1000 TB	82
More than 1000 TB	2

The mission of the OLCF and NERSC is to accelerate scientific discovery by providing high performance computing and data resources to the scientific community. Traditionally HPC systems have dominated the mission focus due to the overwhelming needs of the community for more compute-cycles and increased scale of simulation. Although this need continues to grow, evident in the push toward exascale computing, recent utilization trends at the OLCF and NERSC and forward-looking requirements gathering activities show an equally important need to provide data-centric systems, services and infrastructure to support the vast amounts of data from both simulation and experiment. To illustrate this, we discuss the simulation data currently stored at the OLCF and NERSC and present pertinent findings from user requirements surveys that project the HPC community′s future data needs.

A. Simulation Data

At the OLCF, the Titan supercomputer is served by a 32 petabyte (PB) Lustre parallel file system called Spider II [14], providing scratch storage capable of 1 TB/s of aggregate bandwidth. In the 4 months since commissioning Spider II in December, usage has risen to 9.5 PB, equal to the total usable capacity of the previous generation Spider I. As seen on Spider II, Lustre filesystem usage can easily grow exponentially, requiring HPC centers to enforce data management policies to purge old files to make room for new computation and analysis results. Data management policies ensure that Spider II utilization plateaus to a manageable capacity to maintain I/O performance. For a high performance filesystem like Spider II, a tradeoff is made in favor of filesystem performance over capacity. The effect of this is felt by the users in terms of shorter purge windows, which creates a necessity for archival storage not subject to the same limited-lifetime and capacity constraints.

Archival storage is provided through the High Performance Storage System (HPSS) technology at both the OLCF and NERSC. HPSS at the OLCF has been active for over 15 years and currently is storing over 34.17 PB of data. Figure 1 shows the HPSS usage at the OLCF from 2010 to present day, demonstrating a steady increase for HPSS capacity due to the accelerating data generation capabilities

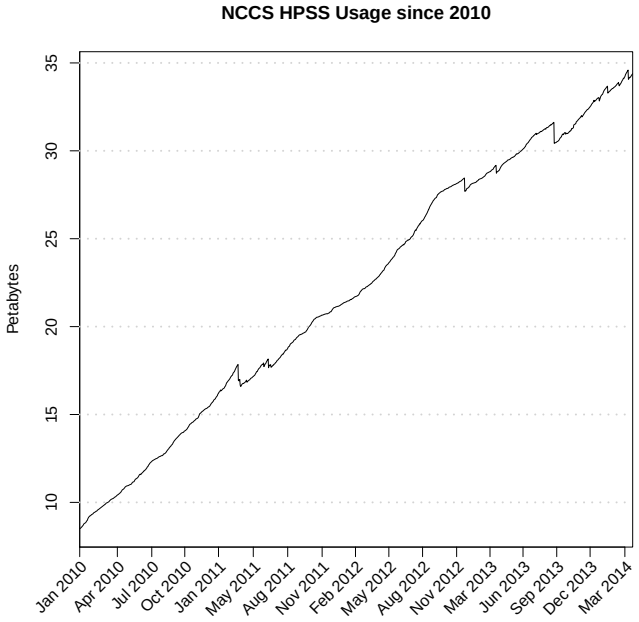

Figure 1. OLCF HPSS usage from January 2010 to March 2014

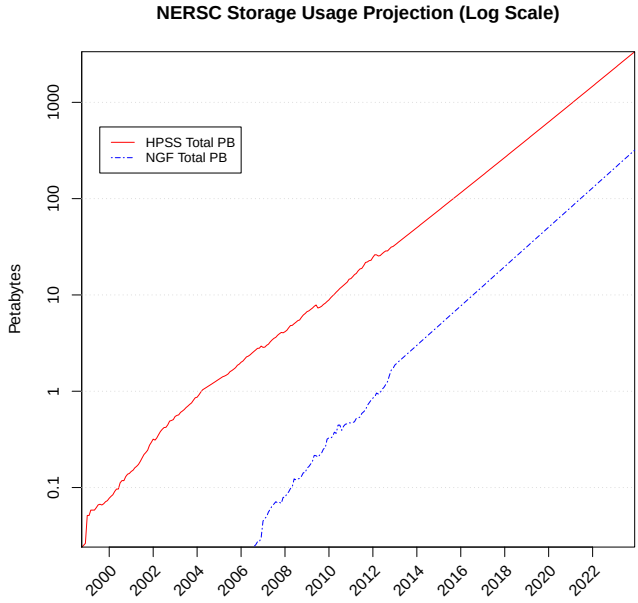

Figure 2. NERSC Global File System (NGF) and HPSS utilization and future projection from 1999 to 2024

of computation, analysis, and visualization resources. Of the 1690 current OLCF users using archival storage, the median user has only 0.5 TB stored while the average amount of data stored per user is 18 TB. This can be explained by Table I, which shows that while the majority of users have less than a terabyte stored on the HPSS, a sizable community of users store data greater than 10 TB. This distribution is partially driven by the OLCF data management policy to provide 2 TB of HPSS per user and 100 TB of HPSS in a shared project space. A survey of the quarterly reports from OLCF's flagship Innovative and Novel Computational Impact on Theory and Experiment (INCITE) allocation program, show that most projects request 100 TB of archival storage. Capacity beyond these policy limits requires special approval.

The NERSC global file systems (NGF) provide 15 PB of high bandwidth capacity. The HPSS archive and HPSS backup systems at NERSC have data dating to 1976 and contain a total of 62 PB of scientific data. Figure 2 shows past and future capacity projections for both the NGF file systems (disk) and HPSS archives (tape) [15]. The growth of HPSS archive is about 70% per year. The growing demand on disk capacity outstrips the ability to meet demand. Current plans are to increase disk capacity at about 60% per year based on technology capability. User demand for our HPSS archive as derived from scientific requirements reviews is 2 times what we are able to provide.

According to a requirements gathering exercise of the OLCF user community [16], the need for data storage and movement will increase with increasing compute power. Respondents were asked to describe their future computing needs projecting up to 2017. When users were asked to rate the importance of various hardware features, archival storage capacity ranked 4th after memory bandwidth, flops, and interconnect bandwidth, and Wide Area Network (WAN) bandwidth ranked 7th out of 12 possible features. Further, respondents also speculated that their simulation data requirements would grow in 2017, such that the aggregate storage needs would be 24 PB scratch and 164 PB archival storage with the average data lifetime to be 10 years. Clearly this indicates the importance of data movement and storage to the HPC user community. The amount of data generated will need an additional host of tools and support for data mining, visualization, and reduction, but data movement and storage needs must be met for those tools to be applied.

B. Data Movement

While the volume of stored scientific data shows significant growth, data sets are not stationary and data movement is also increasing exponentially. Bandwidth is an increasingly key commodity in the enterprise of science. Long term trends show exponential increases in realized bandwidth, evident in Figure 3, which shows the ESnet monthly traffic volume since 1990 and provides a very general view of the aggregate growth in scientific data movement. This trend is also seen from the perspective of a single HPC center, NERSC, in Figure 4). There are multiple motivations for this trend broadly, but HPC centers see two primary

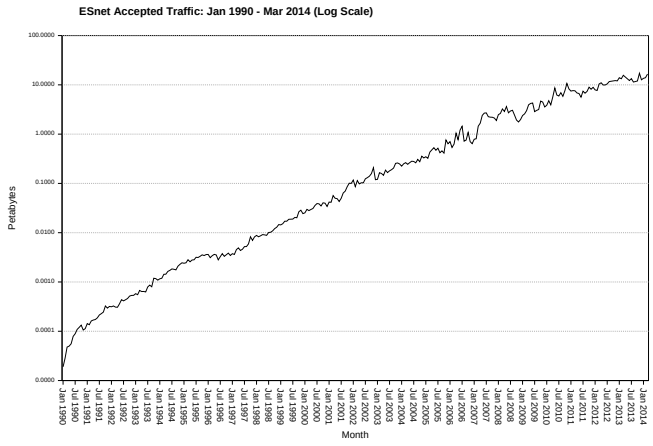

Figure 3. ESnet monthly traffic volume from January 1990 to March 2014

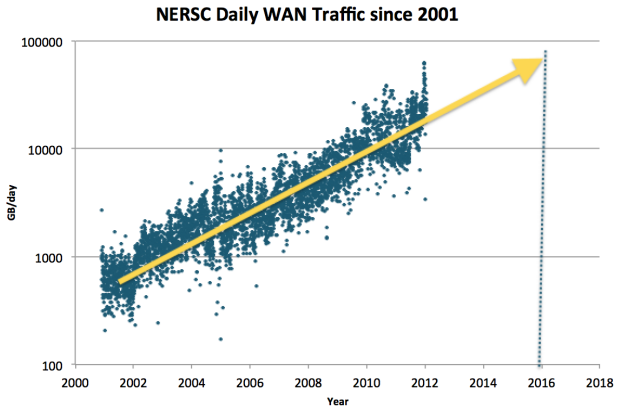

Figure 4. NERSC daily WAN traffic volume since 2001

drivers of bandwidth 1) the exponential increase in detector bandwidths from large scale scientific instrumentation and 2) the growth of Big Science collaboration to Internet scales involving large distributed teams of scientists. Increasingly HPC centers act as data hubs from which science gateways, grids, and portals provide a meeting point for research data.

III. Workflow

Deploying architectures for data centric workflows differs significantly from HPC workloads in that there exist fewer shared software standards. Data-centric workflows are often highly layered involving a richer set of software and services. Building the right infrastructure to support such workloads involves a wider set of choices and greater complexity in finding optimal performance.

The 2004 technical report on the Office of Science Data-Management Challenge [11] presents a general workflow for a computational scientific experiment shown in Figure 5. Illustrated in this picture are the layers of control flow,

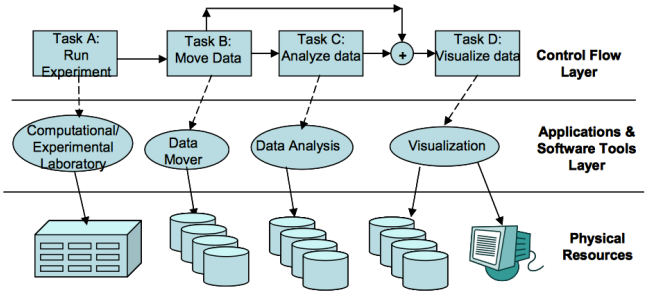

Figure 5. Example of a workflow created in the scientific investigation process, showing the three layers: control flow, applications and software tools, and physical computer hardware. Taken from the Technical Report of the Office of Science Data-Management Challenge [11].

application, and hardware needed to accomplish the data-intensive workflow. Associated with this workflow in the report, were goals for data management. Among them were the ability to better handle the massively parallel I/O that is required to allow the supercomputer to perform I/O without bottlenecks, the ability to analyze and visualize data as it is produced and the ability to transfer the data efficiently. Many of these goals have been well met for terascale computing in the decade since this report was written, however the computational capability has grown from the terascale to the petascale and continues to grow, which keeps these goals ever relevant.

A. Intra-Workflow

Inside a single center many resources exist to allow a multi-system workflow. At OLCF, data is simultaneously accessible on Titan, its associated analysis and visualization clusters, and the data transfer nodes via the center-wide Lustre filesystem. HPSS has a dedicated data transfer node for data movement to and from the Lustre filesystem. The workflow show in Figure 6 is typical during the active part of a project's allocation, with some users archiving and pulling data from HPSS every cycle. This workflow is easily automated with batch scripts because passwordless cross job submission is possible between Titan, the analysis and visualization cluster, Rhea, and the data transfer node (HSI DTN) used to access HPSS. Even projects with workflows designed to utilize only one HPC center need efficient data transfer, because data is typically moved to more permanent storage at the end of a project.

B. Inter-Workflow

With well-coordinated systems at HPC centers like those at NERSC and OLCF, it is preferable to move computation to the data, and science collaborations typically do this whenever they can. However, it is frequently impossible to move all the computation to the data. Example cases include the use of HPC resources to analyze data produced at a remote facility, the gathering of data sets from multiple

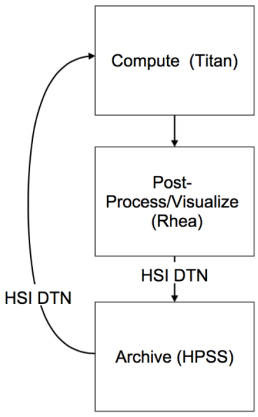

Figure 6. Typical workflow for computation at OLCF

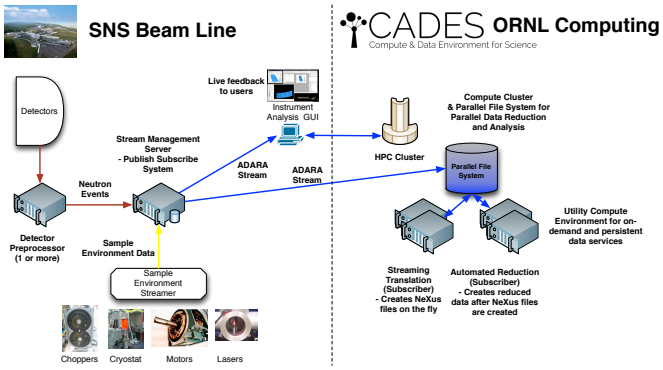

Figure 7. Architectural overview of ADARA integrating experiment with HPC.

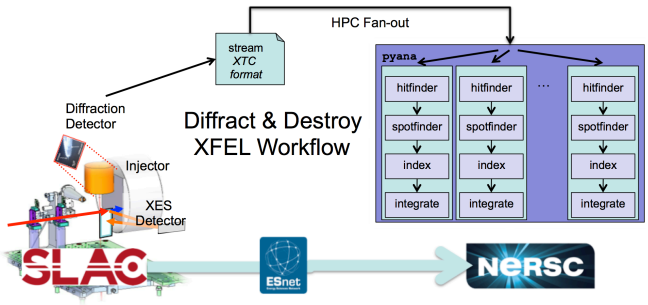

Figure 8. An Experiment run at the Stanford Linear Accelerator Center (SLAC) utilizes the resources at NERSC via ESnet.

sources for intercomparison purposes, and the transfer of data sets from a filesystem at one facility to the filesystem at another facility (e.g. because of available applications, system capabilities, or allocation availability). The following case studies stand as examples of inter-facility workflows.

1) Spallation Neutron Source: ORNL operates the world's brightest neutron source, the Spallation Neutron Source (SNS) that hosts hundreds of scientists from around the world, providing a platform to enable break-through research in materials science, sustainable energy, and basic science. Due to the high-power on target (1.4 MW), the pulsed nature of the source and the high-resolution detector technologies, instruments at the SNS are capable of generating millions of neutron events per second. A collaboration between ORNL computing and SNS data experts has developed a streaming data acquisition system and workflow management system coupled to a high-performance computing infrastructure for capturing data from the neutron detectors and the sample environment equipment. The Accelerating Data Acquisition, Reduction, and Analysis (ADARA) system provides near real-time feedback to users and data collected is instantly available to the user and for processing on a high-performance computing infrastructure. This data streaming infrastructure differs from other file-based data-movement approaches and, in the case of the SNS, was more appropriate for near real-time feedback as data can begin to be processed as it is generated. Today, all SNS data (hundreds of terabytes) is managed by ORNL's CADES.

2) Stanford Linear Accelerator Center Linac Coherent Light Source(LSCS): The most accurate structural analysis of complex biomolecular machines is currently derived through single particle x-ray crystallography. Thousands or millions of diffractive images of aerosol droplets that contain on average a single particle are composed to reconstruct angstrom scale three dimensional structures. The overall workflow involves high bandwidth collection of images at the devices, prompt analysis to determine data quality, followed by storage and analysis of tens to hundreds of terabytes of detector data. In 2013 one experiment on the photosystem-II system conducted by Nick Sauter resolved the structure of this important biological system using 130TB of image data. The data-centric aspects of that experiment are detailed in Figure 8. This LCLS diffraction experiment can generate 150TB of data which is analyzed in parallel. Data movement from the instrument to computing capability, HPC execution, and data access through a web-based gateway form the overall workflow.

3) Bellerphon: Multi-center HPC workflows often employ methods to reduce the data locally before it is transferred. For example, a collaboration using resources at NERSC, OLCF and NICS to study core-collapse supernovae [17] have developed an automated workflow control tool called Bellerphon [18]. The multi-tiered elements, including HPC center resources, web and data servers, and user-friendly data presentation clients, of this workflow are shown in figure 9. Components of this software system locally reduce the data generated by HPC core-collapse supernova simulations. The reduced data is then archived to HPSS and transferred in near real-time to a remote web and data server where it is rendered and made accessible via a web-

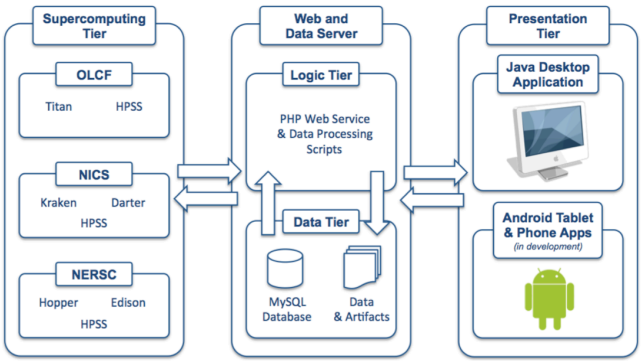

Figure 9. A schematic representation of Bellerophons n-tier architecture.

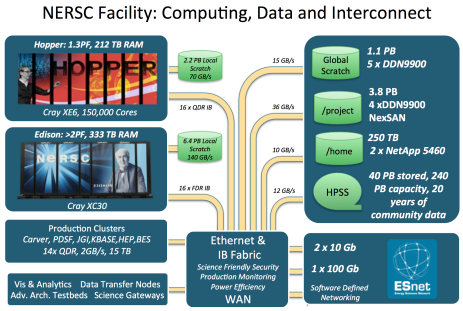

Figure 10. Compute, storage and networking systems at NERSC in 2014.

deliverable, Java application. Currently, Bellerophon uses the scp Linux utility augmented with ssh key authentication as its primary data transfer method. This workflow process has reliably delivered up-to-date data analysis artifacts for multiple core-collapse supernova models however the slow transfer speed of scp has occasionally resulted in incomplete transmissions of the data. The collaboration is now running three-dimensional supernova models, which generate one to two orders of magnitude, more data per model. The collaboration is investigating methods for visualizing the data locally at the HPC centers and only transferring the resulting images to their home institution. Thus, one of the primary future workflow challenge for this collaboration is to compatibly automatic local visualization of the data at each HPC center.

Deriving user requirements from data-centric workloads is an important but challenging undertaking. Tools such as Netlogger, Bro, LMT, IPM, and perfSONAR, all provide views into aspects of these workloads, but as yet, there is no overarching or quantitatively reliable way to globally assess data-centric workflows. As such we rely on the data available, often from disparate sources, integrating that with input from users through surveys and direct contact. We also suggest a wider and continued effort to build workload analysis into HPC more directly moving toward an enterprise level dashboard for HPC resourcing.

IV. Data Transfer

When the workflow involves resources at multiple centers, speed and reliability of the transfer is required to keep the parts of the simulation cycle in sync. Because data movement is so important, both science networks and HPC facilities put significant engineering effort into ensuring consistently high data transfer performance for data transfer, data ingest, and data export. There are significant challenges to data transfer including security, diverse workloads impeding performance, and compatibility of software at each endpoint.

Ideally, a workflow has a high degree of automation. However, each time data is transferred to a new facility, authentication must be used to gain access to the facility's systems. The rules and requirements for security and authentication at different intuitions vary, requiring researchers to customize their workflows for the allowed authentication methods of each center. Centers like the OLCF that must maintain a high degree of security, require one-time passwords (OTP) for authentication that must be manually entered. When ssh key authentication is allowed, as NERSC and CADES do, and the limited-time proxy certificates that stand in for OTP for gridFTP transfers, as both OLCF and NERSC allow, automation of transfers is possible via scripting.

An additional challenge in data movement requirements are the file size distributions at HPC centers. Large parallel jobs like those run on Titan can generate millions of small files very quickly, while other data sets contain files of 100 TB or greater. Both extremes of many small files and a few monolithic files pose unique technical challenges to the transfer tools and the distributed parallel filesystems on which they are stored. In the small file case, increased overhead is incurred in terms of more metadata operations for the sending and receiving filesystems and in increased CPU utilization by the tools to coordinate the transfer of each file. The overhead from metadata operations on 1MB files can easily become the bottleneck in a WAN transfer even over a 10Gbit/s network backbone. A common remedy for this slowdown is to "tar" files into a single file before transfer. However, an interruption during transfer could require restarting the transfer of the single large file, or with some parallel transfer tools, the entire data set. Thus, it is increasingly necessary to use fault tolerant transfer tools, and educate users about how to navigate the parameter space of these tools to suit their workflow.

In this section approaches, tools, and infrastructure deployed at NERSC and OLCF that enable multi-center workflows will be discussed.

A. Infrastructure

Figure 10 shows the NERSC compute, storage and networking systems provided in 2014. The NERSC environ-

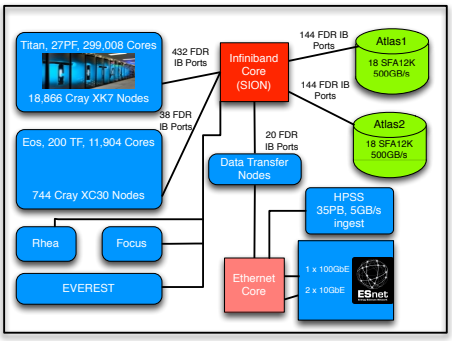

Figure 11. Compute, storage, and networking systems at OLCF in 2014.

ment reflects a goal of providing localized high performance I/O to each of the compute systems through their local scratch file systems, while maintaining an increasing capability in centralized storage for all computational systems.

The key systems at NERSC for data movement are the Data Transfer Nodes (DTNs) and the global scratch and project file systems. There are four NERSC DTNs that are connected directly via 10Gb ethernet close to the network border; experience shows this maximizes the bandwidth available to those servers. The DTNs are also highly connected to the NGF file systems such as global scratch and project with an Infiniband storage network. Transfer statistics over the past few years show that the four DTNs at NERSC move several petabytes of data each year and are second only to Hopper in the amount of data moved.

At the OLCF, an Infiniband fabric, named SION (Scalable I/O Network) is at the core of the center's I/O architecture serving Titan, multiple institutional clusters, and enabling transfers to remote sites by way of data transfer nodes shown in Figure 11. With one side of the data transfer nodes facing SION, and the other side connected to the OLCF Ethernet backbone, the DTNs provide authorized users a means to efficiently move data to or from any connected filesystem or archive. The transfer nodes adopt the Science DMZ model and have Ethernet interfaces on an IP network with a minimal number of hops and firewalls to the ESnet backbone.

Bridging two high speed networks also means the DTNs could be a bottleneck to data movement. SION's primary purpose is to provide the highest possible I/O bandwidth between Titan and Spider II with 1TB/s of aggregate capacity between the two. The other side is ESnet, built with a 100Gbit/s (12.5GB/s) backbone. Thus, a single DTN equipped with a 10Gbit/s Ethernet interface is the rate limiter in this architecture. OLCF is in the process of upgrading the DTNs with 40Gbits network interfaces, but even so, multiple nodes must be deployed and care must be taken to avoid contention for a single node's bandwidth.

In response to this challenge, the OLCF has taken a service oriented architecture to provide the most utility for common data transfer scenarios. The data transfer cluster is divided by function with 2 nodes dedicated for interactive use, 10 for batch-scheduled transfers, and 3 for transfers to the HPSS archive. Interactive node usage shows that the data transfer tools studied in this paper (scp, rsync, bbcp, and GridFTP) are commonly used. With only 2 nodes for this purpose, these are the most heavily used and the Ethernet interfaces are often fully utilized. However, these nodes are susceptible to misuse and users running CPU intensive tasks such as computation, data processing, and analysis, can impede data transfers of other users. Free from this contention, the scheduled data transfer nodes are intended for long running transfers with tools such as GridFTP for WAN transfers and dcp [19] for intra-center transfers. Transfers using these nodes can easily be integrated with workflows on Titan since they share the same batch scheduling system. The remaining 3 transfer nodes are for moving data using the *hsi* tool to interact with the HPSS. With this segregation of transfer nodes by function, users at the OLCF are able to manage contention for these nodes and match the hardware to the needed capacity of each function.

B. The Science DMZ Approach

The Science DMZ model [20] provides a set of design patterns for building data transfer infrastructure that performs consistently well. The primary components are: sufficient bandwidth to avoid congestion, a location in the network at or near the site perimeter, network test and measurement for performance verification, dedicated systems for data transfer, and high-performance data transfer tools running on those systems. Building dedicated systems (called Data Transfer Nodes or DTNs) for data transfer is already familiar to many HPC systems engineers, as is the use of high-performance data transfer tools such as Globus [21]. Experienced systems engineers understand the importance of tuning and proper tools - nobody with experience in HPC expects the default system configuration and data transfer toolset (e.g. SCP) to be useful for high-performance workloads. The same is true for networks - the network must be engineered to provide the high-performance services required to support data-intensive science. In particular, the elimination of packet loss is very important for high-performance long-distance data transfers.

Packet loss is so detrimental to TCP performance (Figure 12) that the elimination of packet loss is a central focus of performance engineering for HPC center networks and wide area networks such as ESnet. Just as HPC systems engineers design and tune for maximum performance and monitor metrics to ensure good outcomes, HPC network engineers design and tune for performance and also deploy test and measurement systems running perfSONAR [22][23] to allow them to verify that the network is performing as it should. The Science DMZ model succinctly describes the elements of a high-performance network infrastructure

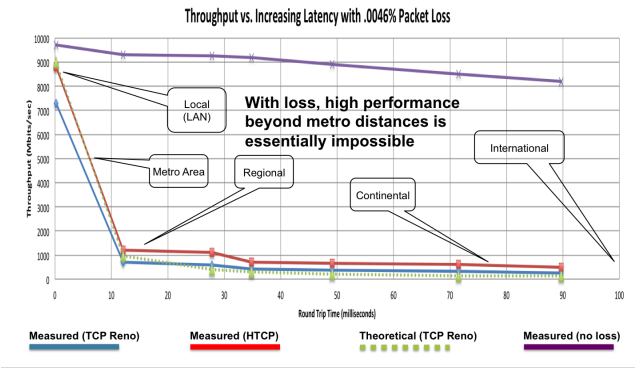

Figure 12. Impact of packet loss on data transfer performance as the distance between data transfer nodes increases. Shows need for specific engineering for high-performance data transfers, perfSONAR. Originally appeared in [20] (annotations added).

	GO	GUC	RSYNC	SCP	BBCP
Web Interface	Yes	No	No	No	No
Checksumming	Yes	Yes	Yes	No	Yes
Parallel Transfers	Yes	Yes	No	No	Yes
Requires Grid Cert	Yes	Yes	No	No	No
Adj. Network Buffer	Yes	Yes	No	No	Yes
Adj Filesystem Buffer	Yes	Yes	No	No	Yes
Default on Most Systems	No	No	Yes	Yes	No
Striped Transfers	Yes	Yes	No	No	No

Figure 13. Comparison of transfer tools.

capable of reliably supporting data-intensive science.

C. perfSONAR

Since packet loss is so detrimental to network performance, maintaining an infrastructure that supports data-intensive science requires the ability to find and fix problems quickly. Traditional network monitoring systems are very good at raising an alert when a so-called "hard failure," such as a system crash or network outage, occurs. However, "soft failures" as described in [20] cause performance degradation without drawing attention to an obvious cause. Traditional network monitoring systems are typically unable to locate the cause of the degradation for two reasons - first, nothing has obviously failed (something is just not working as well as it should be) and second, the performance degradation is often observable only at the end systems. A large number of networks, universities, laboratories, and facilities use perfSONAR to independently test and measure the network in order to find and fix performance problems. There are over 1000 perfSONAR test hosts deployed worldwide [24], and the global distribution of these test hosts is a hugely valuable resource for isolating performance problems.

We briefly describe two of the primary test and measurement tools in the perfSONAR toolkit here - one for the measurement of throughput, and the other for measuring delay and packet loss. perfSONAR manages throughput measurements by means of the Bandwidth test Controller (BWCTL) [25], which in turn runs one of several throughput test tools - typically Iperf (version 2 or version 3) [26], or nuttcp [27]. The reason BWCTL is used rather than running the throughput test directly is that BWCTL provides several capabilities that make it valuable in the larger perfSONAR context. BWCTL serializes tests, so that a given test host only runs one test at a time. In addition, BWCTL has a simple policy language and authentication mechanisms, so that tests can be controlled and managed (example limits include duration, type of test, IP address restrictions, and data rate limitation). In addition to throughput tests, perfSONAR uses one-way amplitude measurement (OWAMP) [28] [29] to measure several metrics including one-way delay, packet loss, and packet re-ordering between two test hosts. OWAMP is sensitive enough to detect queuing delay in the routers and switches in the network path, and is very useful in locating sources of packet loss that cause poor throughput performance. The perfSONAR tools can be used interactively by engineers engaged in troubleshooting, and can also be run regularly to establish a performance baseline and a historical record of performance over time. Regular tests are very useful for identifying the time at which performance changed - this often helps narrow the search for a cause (e.g. in the case of a change in performance that coincides with a maintenance event). Test results can be displayed using the native perfSONAR graphing tools (see figure 17), or aggregated into a dashboard [30].

The perfSONAR throughput tests are typically configured to run in one of two modes, each with a different philosophical basis. One mode is sometimes used by network operators, and is specific to the network infrastructure being monitored. The tests are designed to provide as much insight as possible into the health and functionality of the network. In the other mode, the tests are configured to resemble production data transfer traffic as closely as possible. This second mode is by far the most useful for HPC centers, and for the networks that carry data-intensive science traffic. When the throughput tests are as good a proxy as possible for production traffic, it is easier to use the tests to identify the cause of performance problems encountered on production systems.

D. Data Transfer Software

Data transfer methods must be easy to use and widely available at HPC centers since the given method must be functional and installed at both ends of the data transfer. Speed is vital for larger data transfers, but between two tools that offer adequate speed and features for a large data transfer workflow, users will often choose the tool with the best ease of use over the tool with maximum speed. Figure 13 shows several features of four frequently used transfer tools, scp, rsync, bbcp, globus-url-copy (GUC), and Globus

(GO), common to OLCF and NERSC.

scp and rsync are single tcp stream tools common to any Unix-like system and provides minimal options to confuse the user. rsync has additional options to improve performance and provide control of file transfers. rsync allows files to be "synced" by only sending the differences between the source files and existing files, which can greatly improve subsequent transfer times and give the ability to recover from a failed transfer without losing initial progress. bbcp is a multi-streaming point-to-point network file copy application created at SLAC as a tool for the BaBar collaboration [31]. bbcp uses simple SSH authentication, can be scripted into the workflow similar to rsync, and also provides a large number of options to control the file transfer performance and resiliency. globus-url-copy (GUC) and Globus are GridFTP clients. GUC is a command line tool that allows the choice of several parameters to optimize the transfer speed. Globus [21], formerly called Globus Online, is a hosted GridFTP service that allows the use of a browser to transfer files between trusted sites called endpoints. This service optimizes the transfer for users, automatically handles restarts when transfer errors occur, and offers checksum validation of the data as a default. It also offers a command-line interface so transfers can be scripted. The challenges for users of GridFTP are that each HPC center has differing policies for the method of authentication and not all centers support GridFTP.

Many factors can impact the speed of a transfer. From the network side, average Round Trip Time (RTT) of data, usually seen in ms between the hosts, number of parallel TCP streams used, and the congestion window on the sender which works in conjunction with the TCP receive window on the remote host which on modern Linux system scales during the course of the transfer. To get around single TCP stream limitations, imposed by tools like scp and rsync, tools like bbcp and GUC and globus are able to use multiple TCP streams in parallel (-s for bbcp and -p for GUC). This will still limit to the max available bandwidth of a single data transfer node, 10Gb in this use case. The GUC tool can take this even further by using two or more nodes for the transfer on both ends by using the stripe mode. This allows for multi-node transfers at speeds beyond a single 10Gb connection. The speed will be determined by the available bandwidth on both ends as well as filesystem speed. In the next section we test these tools for speed and utility.

V. Data Transfer Results and Impact

To frame the context of our results we discuss the limits of the network and filesystem I/O infrastructures and how they interact with the various data transfer applications.

When looking at the network infrastructure in isolation from the filesystem or application, we are looking at the theoretical transfer layer (layer 3 in the OSI model) throughput between DTNs at NERSC and ORNL. As described earlier, the ESnet backbone is capable of 100 Gbit/s between the border routers at each site. Each of the DTNs in these transfer tests has a single 10 Gbit/s link between it and the border routers. In a network constrained scenario, we expect, at best, approximately 10 Gbit/s in a single direction between two DTN endpoints. In the globus-url-copy (GUC) tests with a stripe count of 2, we are utilizing 2 pairs of endpoints, so the physical network capacity allows for 20 Gbit/s. However, all of these network links are shared resources and while these tests were carried out during the hours of night showing the lowest utilization, we expect that our tests were contending with traffic from other users. For a single TCP stream to achieve full network link utilization, not only would it need a quiet network, but also consideration of TCP and linux kernel tuning parameters. Such tuning advice is provided by ESnet's Fasterdata Knowledge Base [32]. Single-stream tools like scp and rsync are susceptible to inefficient link utilization, while tools that are capable of multi-stream transfers, such as bbcp, GUC, and Globus, parallelize data transfers over multiple TCP connections. This achieves a higher link utilization through a process in IP networks called statistical multiplexing, where multiple streams keep packets queued on a network device, which are then transmitted as fast as the link will allow.

While each site has a high performance parallel filesystem capable of many GB/s of aggregate throughput, the transfers of interest for this paper involve a limited number of DTNs, thus the I/O capacity between the DTN and the filesystem can be a rate-limiter. Furthermore, each DTN acts as a single client, and only a single process may be interacting with the filesystem. The Lustre filesystem has a single-client, single-process performance limitation that is demonstrated in Table IV. Tools evaluated in this paper only have a single process tasked with filesystem I/O, so this limitation is in effect. However, the single-process limitation can be circumvented by increasing the client count involved in a transfer by striping across 2 or more DTNs with GUC. Future tools such as dcp [19] coordinate multiple processes in the transfer, so higher utilization of the each link between the DTN and the filesystem could be achieved. It's worth mentioning that the I/O paths between the DTNs and their respective filesystem are subject to use by other users on the same DTN. Also NGF and Spider II are both shared center-wide filesystems, so there will be components of the DTN's I/O path that will see contention from large HPC systems like Titan and Edison.

Other filesystem operations such those between the client and metadata server, and lock management tasks must be also be completed during an application's file transfer. As average file size in a dataset decreases, the percentage of time spent relative to the actual data transfer will increase, registering a slower throughput measurement. While this effect is universal across all transfer applications tested, some are better able to pipeline file metadata operations,

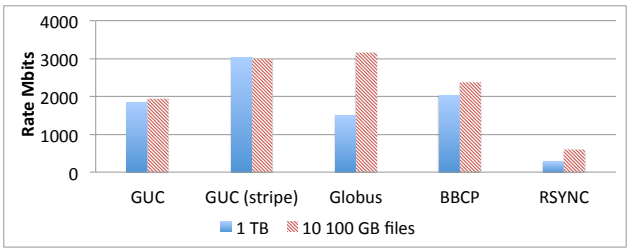

Figure 14. Average transfer rate in Mbit/s using GUC, GUC with striping, Globus, bbcp, and rsync

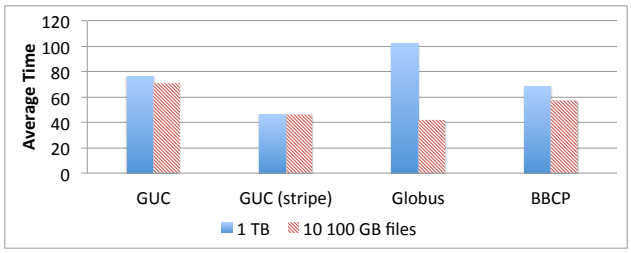

Figure 15. Average Time in minutes for transfers using GUC, GUC with striping, Globus, and bbcp

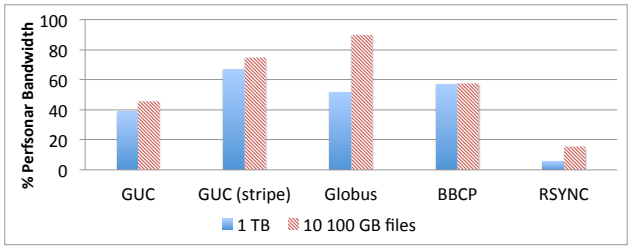

Figure 16. % perfSONAR Bandwidth

which is strongly beneficial for cases of many small files. Applications supporting striping or parallel processes have an advantage in this respect because they can make use of multiple clients, each with a separate channel for operations with a metadata server.

A. Data Transfer Tests

With these potential limitations in mind we present the rates from different transfer tools as compared to what is achievable by perfSONAR. While perfSONAR demonstrated that single-stream TCP performance between NERSC and ORNL DTNs could average 4518.9 Mbit/s, the applications we tested typically achieved results below that number. The difference is approximately the combined effect of application and I/O inefficiencies. By measuring the ORNL and NERSC filesystem throughput in workloads comparable to data I/O by the transfer application, we establish an expected filesystem rate, such that the limiting factor for the transfer can be identified, be it network, I/O, or application overhead.

Users who request help with transfers often need to transfer several terabytes of data, but the average file size can vary significantly. For this reason we pick two configurations of data to transfer, both totaling 1TB: a single 1TB file and one directory with ten 100GB files. The 4 cases of multi-streaming data transfer applications that we evaluated are globus-url-copy (GUC), GUC with striping over two DTNs, Globus, and BBCP. For GUC and BBCP we use four parallel tcp streams (-s 4 for bbcp and -p 4 for GUC), the default filesystem buffer size of 1MB and allow the TCP buffer to scale as the tool permits during the transfer. For Globus we turn off the default checksum of the files to get a measure of just the transfer speed.

Figure 14 shows the average throughput achieved with each tool while transferring 1TB in a single file and in ten 100GB files. The fastest tool was clearly GUC when used with the stripe option. This option allows the transfer to utilize two pairs of DTN endpoints in parallel, so theoretical network capacity is 20Gbit/s and I/O capacity is limited by 2 filesystem clients. Since walltime is the factor needed to understand how data transfer fits into a workflow, Figure 15 shows the average duration in minutes for each multi-streaming method. For comparison, a tool that could saturate a dedicated 10 Gbit/s path between NERSC and ORNL DTNs, would complete a 1TB transfer in 14 minutes. A 1TB transfer at the real-world network capacity measured by perfSONAR would take 31 minutes. All methods of disk to disk transfers were able to transfer 1 TB in under two hours, except rsync, which is not shown on this plot because it averaged over 7 hours.

Figure 17 shows a graph of the measured bidirectional usage between ORNL and NERSC over ESnet for the month of April as measured by perfSONAR. The ORNL system dedicated to perfSONAR is on a separate 10 Gbit/s link from the DTNs used for the tests in this paper, so our tests are not reflected in this graph. Nonetheless, it gives a reasonable estimate of the useable bandwidth available on a 10 Gbit/s path between the two sites. Note that the bandwidth fluctuates between 1-5 Gbit/s even in the absence of our testing. The graphs show that the least utilization occurs late at night, but before the early morning hours.

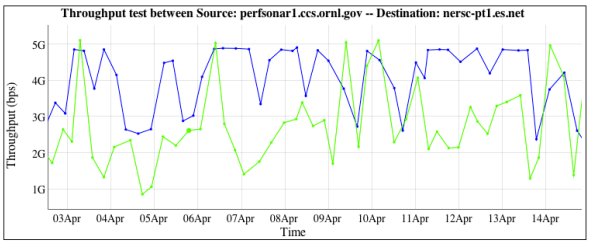

Figure 17. perfSONAR throughput between ORNL and NERSC for April 2014

The majority of the transfer tests were conducted during this quiet period.

Figure 16 shows the average throughput of each tool as a percentage of the average bandwidth for a single TCP stream as measured by perfSONAR's BWCTL tool during the transfer. All systems shown here are in production, so perfSONAR shows the best case scenario for saturating the pipe with a single TCP stream during the time of each test. From comparison of the rsync tests with perfSONAR, it is clear that a singe TCP stream tool does a poor job of saturating the bandwidth. The gap between the usable network capacity as measured by perfSONAR and the application transfer could be explained by overhead in either the application or filesystem I/O. We use tools that are capable of sending multiple TCP streams to achieve better saturation of the bandwidth. We do not expect the tools to achieve multiple times the bandwidth seen by perSONAR and figure 16 confirms that they do not.

Of the applications tested, Globus transferring the 10 100GB files achieves the best rate of transfer. Since this is a hosted service that optimizes the transfer for the user, the details of the optimizations made are unknown. The next best rate was achieved with GUC striping across 2 DTNs. This test can take advantage of two 10Gb links to ESnet and two filesystem clients. Note, for the parallel streaming tools, the file with 10 100GB files achieves better throughput than a single file. Exposing parallelism through multiple files can increase filesystem I/O rates, but further investigation is needed to determine where the efficiency gain comes from in a disk-to-disk WAN transfer.

To understand the impacts of each component (network, application, filesystem), we look at the pre-tests of our GUC transfers to verify whether each component in the disk-to-disk transfer is operating as expected and the test is not being run during times of heavy user contention. The pre-tests consist of running GUC for 30 seconds are designed to stress a particular component of the transfer while including any application overhead. The memory to memory test is most demanding of the network, while the disk to memory and memory to disk tests stress filesystem I/O at each site.

The results for GUC including the pre-tests are shown in Table II. We use the rate shown for the transfer from ORNL memory to NERSC memory as a measure of network capacity for this transfer. When incorporating one filesystem at a time on the NERSC and OLCF sides, the transfers were only able to fill 80% and 50%, respectively, of the usable network capacity. Given the results of the pre-tests, it is not surprising that the 1 TB transfer immediately following this test was able to only achieve 47% of network capacity.

Consider the same test with GUC, but using 2 endpoints at both NERSC and ORNL to conduct the transfer in parallel. This is achieved using the stripe option with GUC. Striping in the context of GUC means to segment the file into fixed-size chunks (1MB by default) and pass off the chunks to multiple DTNs to carry out the network transfer. On the remote side, a process is responsible for assembling the chunks in order and writing contiguous data to disk.

Table II
THROUGHPUT TEST USING GUC

Test	Average Rate Mbit/s	% Network Capacity
OLCF Mem to NERSC Mem	4144	100
NERSC Disk to ORNL Mem	3335	80
OLCF Disk to NERSC Mem	2078	50
1 1TB file	1932	46

Table III
THROUGHPUT TEST USING GUC STRIPE: UTILIZING TWO DTNS

Test	Average Rate Mbit/s	% Network Capacity
OLCF Mem to NERSC Mem	8874	100
NERSC Disk to ORNL Mem	5339	60
ORNL Disk to NERSC Mem	4646	52
1 1TB file	2965	33

Table III shows the improvement in transfer rates utilizing GUC striping across DTNs on each end with the default stripe size of 1MB. The end-to-end 1TB file transfer using the GUC striping option increased to 1.5 times the values seen in II. The memory-to-memory throughput increased by approximately a factor of 2, benefitting from multiple 10Gbit/s paths. The throughput of the OLCF disk-to-memory test also more than doubled from the values seen in II. However, the NERSC disk-to-memory test only increased by 60%. It is apparent that the striping option overcomes a significant filesystem related limitation for the Lustre filesystem.

A major difference with respect to the filesystem between the GUC and GUC with stripe scenarios is that striping doubles the filesystem client count. Since single client filesystem performance is limited on the Lustre filesystem, we measure the limitation with the I/O benchmarking tool fio [33]. All tests transferred 128GB to avoid caching effects on the DTNs at ORNL and NERSC with main memory sizes of 64GB and 48GB respectively. While some advanced applications take advantage of asynchronous I/O libraries, we chose the ioengine "sync" for fio, for similarity to the POSIX IO system calls used by the GUC application. Parallelism was varied by the running tests with 1, 2, and 8 processes, each transferring an equal part of the 128GB transfer. While tests were conducted on a single DTN client, increasing the parallelism through more processes approximates increasing parallelism through multiple clients.

The resulting average filesystem throughput can be seen in Table IV. The write throughput on the ORNL Lustre filesystem with a single process was limited to 2113 Mbit/s, which is just a shade above the GUC 1TB file transfer rate without striping. The read throughput on the NERSC GPFS filesystem was 14859 Mbit/s, higher than the capacity

of the DTN's 10Gbit/s network links, so we infer that the bottleneck of the single GUC 1TB file transfer is the Lustre filesystem throughput at ORNL. As the number of processes increases to two, throughput is 3684 Mbit/s, and at eight processes the throughput exceeds the capacity of a 10Gbit/s link.

Table IV
DISK THROUGHPUT MEASUREMENT WITH FIO

Site	IO Operation	Number of Processes	Average Rate Mbit/s
NERSC	read	1	14859
NERSC	read	2	13193
ORNL	write	1	2113
ORNL	write	2	3684
ORNL	write	8	11642

This filesystem benchmarking with fio shows that the performance improvement from GUC to GUC with striping comes from parallelism at the Lustre client. The distributed copy tool dcp already gives users high parallel filesystem throughput through task parallelism, but it does not have the the ability to transfer data over a WAN. Non-striped disk-to-disk transfers could achieve throughput near the memory-to-memory rates demonstrated in these tests if applications were re-written to interact with the filesystem with a degree of parallelism appropriate for the underlying filesystem technology.

B. Transfer Tools Ease of Use

It is important to note that it was difficult to get complete transfers, without manual restarts for the directories of files with most tools. Globus was the only fast tool that showed no difficulty in any transfer trial. However, Globus restarts the transfer if there are errors without user intervention. Rsync, though slow, also was able to transfer the files reliably on the first try. Only 3 in 10 attempts to transfer the directory with BBCP resulted in a clean transfer of all 10 files with no manual restarts. BBCPs documentation does not recommend recursive directory transfers; rather it recommends the use of compression when the user must transfer directories. BBCP has an option for using a pipe inside transfer syntax to employ compression tools like tar and gtar. This prevents the user from having to manually tar and untar a set of files at the ends of the transfer. The average rate of the transfer using bbcps pipe option with gtar of the 1 terabyte was 560 Mbits, a quarter of the recursive copy speed. The long running time required makes this method undesirable for large transfers, however, smaller transfer may benefit. Four tests of directories containing 1000 1MB files were transferred to NERSC successfully with BBCP using the pipe option with gtar. The average rate was 704 Mbits and average time to complete of was 200 seconds. A complete copy of this directory using the recursive bbcp option was not obtained in many tries. rsync took 439 seconds to do the recursive copy and Globus took 362 seconds to do the transfer.

Since these tests were benchmarks we did not include the time for validation of the data in the transfer rate. Globus computes checksums to verify the data integrity of the transfer - this is enabled by default, and so is included in the transfer time by default. This feature can be turned off by the user, however at data scales of hundreds of terabytes or more it is typically wise to use a tool that has integrity checking built-in (even though there is a performance cost). A checksum, using the utility md5sum, of the 1TB directory of ten 100 GB files required two hours to complete. The average time to transfer this directory with GO without its default checksum was 41 minutes. Using Globus′s built-in checksum the time for the checksum and the transfer was 100 minutes. Users can save time by using tools that have been optimized and streamlined for the whole transfer process.

Ability to script a transfer tool into a workflow is an important factor in its usability. All tools utilized here have the option of a scriptable command line interface. The issue with scripting is that manual authentication is required to gain use of the DTNs at OLCF and other facilities with similar security concerns. The Grid tools, like GUC and Globus, can make use of proxy certificates to stand in for passwords on systems that support gridFTP, but the proxy only last for a limited time. For example, OLCF users have a maximum of 12 hours to use the proxy before it must be renewed with a manual entry of the users OTP. The initial setup of the certificates that allow the proxy can be daunting. At OLCF setup requires a multi-day process of attaining an unique Open Science Grid certificate for authentication and requesting each facility to map the certificate on the data transfer nodes. Documentation of this process across the various facilities ranges from poor to copious, requiring users to search and cross-reference across websites to complete the task at both ends of the transfer.

At NERSC users may obtain a unique OSG certificate, but they also may obtain a short-lived proxy certificate without obtaining a unique OSG certificate. Globus makes it easy to use certificates issued from different institutions. However, tools like GUC become very complicated to use when their command line must specify two different certificates for the user.

bbcp and rsync can be setup to allow password-less ssh key authentication. However, not all centers allow this method. bbcp can be set up to push and pull data from one center, thus making it scriptable if one of the center′s systems allow password-less ssh keys. Pushing data to NERSC with bbcp works well in these tests, however when pulling data from NERSC, the connection timed-out before any of the 1 TB transfers could complete.

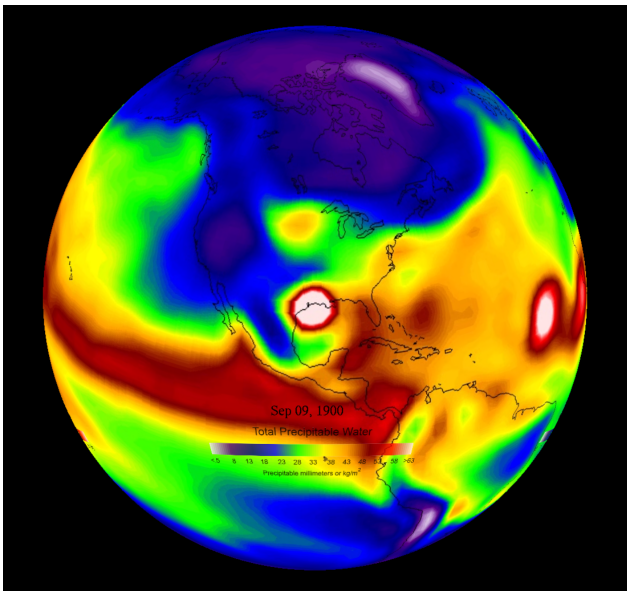

Figure 18. Image from the 20th Century Reanalysis project illustrating the 1900 Galveston hurricane.

C. Data Pace-Setters: Results From Case Studies

The following case studies illustrate how the infrastructure and tools described in the preceding sections and collaboration between OLCF, NERSC and ESnet has benefitted the workflows of users with data movement needs.

1) Combustion research challenge for multi-lab workflow: From 2007-2009, combustion research led by Jackie Chen achieved the first three dimensional numerical simulation of a turbulent non-premixed air flame [34]. The science team's work found new flame phenomena not seen in previous two dimensional numerical simulations. Scientists on the project required computational, visual, and data systems and services at a number of DOE laboratories, to include NERSC, ORNL and Sandia National Laboratory. The data required for the simulation was around 10 terabytes, and team members worked with Damian Hazen at NERSC to move the data using the NERSC DTNs from ORNL to NERSC. The data transfer was one of the first production uses of the DTNs and the success was important to prove the benefits of combining network, storage and systems expertise to the challenges of inter-facility data movement. Researchers spent less time moving their data between facilities and more time using the data towards science accomplishment.

The project also adopted a Kepler workflow system to automate their S3D simulations. The simulations generate 30 and 130 TBs of data per run and they used the workflow to help manage their data across systems and sites. Using the Kepler workflow helped them perform data intensive science simulations and could be adopted to help other data intensive workloads.

2) 20th Century Climate Reanalysis: An early use-case occurred when the 20th Century Climate Reanalysis Project, led by researcher Gil Compo, needed to transfer 40 TB of data from OLCF to NERSC.

This project uses simulations to reconstruct climate data from historical weather maps dating from 1871 to present day [5]. This data is made available on the web through a science gateway at NERSC: http://portal.nersc.gov/project/20C_Reanalysis. These reconstructions are used to validate new climate computer models, compare present extreme weather events with historical events, and understand how extreme events are changing. The project has received allocations on a number of large resources throughout the years including NERSC and OLCF. In mid-2011 the project needed to transfer data generated from simulations run on OLCF's Jaguar system to NERSC which housed the primary Science Gateway for the project. While the volume of data transferred was modest compared to some more recent examples, this transfer had some unique aspects. For example, the data was moved directly between archival storage systems at each site. To accomplish this, staff at both sites had to modify and tune the *hsi* tool which is used to interact with the archival storage system (HPSS). This tool is typically used to move data locally but is capable of initiating 3rd party transfers. In addition to this tuning, NERSC and OLCF staff were able to leverage lessons learned from the Data Transfer Working Group to facilitate moving the data.

3) Direct Numerical Simulations of Chemical Detonations: Tools and hardware must be constantly updated and the best tools for data movement require continued coordination between centers. Take for example, the hosted GridFTP service Globus Online that allows the use of a browser to transfer files between trusted sites called endpoints. Like basic GridFTP, all the challenges of this method are in the setup of the certificates used for authentication and in the fact that both ends of the transfer must support Globus Online. In 2013, OLCF began to support a public Globus endpoint on its interactive DTNs. Also in 2013, OLCF user Jacqueline Beckvermit generated 80 TB of Data from direct numerical simulations of chemical combustion. The data needed to be moved to the ALCF at ANL to continue the project at the end of their OLCF allocation. Tests at OLCF showed that GUC and BBCP spawned from the contention-free scheduled data transfer nodes were the fastest transfer methods [35], but Globus offered the most hassle-free path for the user and included a built-in checksum for data verification. Beckvermit was able to move her data to ANL at an average rate of 1000 Mbit/s, a rate which also included the time needed for the default checksum provided by Globus. Apart from documentation and minor aid provided by user assistance at ALCF and OLCF, Berckvermit was able to move the data herself without the specialized systems described in the first two case studies above. This was

possible because of the knowledge gained from cases like the previous case studies and because OLCF and ALCF had the required hardware, software and public globus endpoints already setup and working.

VI. Conclusion

Data transfer and analysis are becoming integral aspects of HPC facility architecture. Computing centers traditionally focused on large-scale simulation are expanding their repertoire to include user-facing data services. These services in turn require novel data-centric architectural thinking. This pivot is driven as much by massively concurrent petascale HPC IØdemands as it is by the flood of experimental data from advanced facilities. Existing HPC centers are the ideal place to hybridize, economize, and standardize next generation scientific data services.

Many scientific domains are increasingly reliant on scalable, high-performance data management tools and technologies. While this paper has focused on one core area, data movement, a broader set of services are required to address current and emerging needs across a breadth of scientific pursuits. Many of these services build or could be built upon core infrastructure available at major facilities such as the OLCF and NERSC but require different system software and middleware as well as different allocation policies than traditional HPC users. At ORNL, these "data services" are delivered today through the broader Oak Ridge Compute and Data Environment for Science.

Emerging services can include structured storage services such as Key Value and Graph databases. Alternative runtime environments to MPI such as Hadoop/HIVE are of value in a number of use-cases. Distributed computing building blocks such as message queues are in use for loosely coupled distributed workflows that span multiple systems or even multiple facilities. Worfklow management systems such as ADIOS and FireWorks provide the ability to orchestrate complex workflows that require HPC capabilities. At a higher-level, new data and analytic services built upon these core underlying services are emerging. These include analytic services such as data mining and data fusion as well as data management services such as metadata harvesting and management, indexing, discovery, dissemination, and data citation services.

For HPC to provision these new services we must revisit our traditional resource allocation policies which have to a large degree been built around the notion of an ephemeral resource, a compute core. Data services are built around a long-lived resource, the data itself. Both HPC architecture and allocation policies will need to adapt to this longevity.

The longer term roadmap for this transition involves considerations beyond bulk data transfer and should focus towards increasing integration of inter-facility scientific workflows.

Acknowledgment

Readers/Acknowledgements/Resources: Damian Hazen, Thomas P., Jackie Chen, Eric Lingerfelt, Jacqueline, Beckvermit, Gil Compo, and Scott Achtly.

This manuscript has been authored by an author at Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 with the U.S. Department of Energy. This work used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. The U.S. Government retains, and the publisher, by accepting the article for publication, acknowledges, that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

References

[1] G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A. Abdelalim, O. Abdinov, R. Aben, B. Abi, M. Abolins *et al.*, "Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc," *Physics Letters B*, vol. 716, no. 1, pp. 1–29, 2012.

[2] K. Tsang, F. An, Q. An, J. Bai, A. Balantekin, H. Band, W. Beriguete, M. Bishai, S. Blyth, R. Brown *et al.*, "Results from the daya bay reactor neutrino experiment," *Nuclear Physics B-Proceedings Supplements*, vol. 246, pp. 18–22, 2014.

[3] B. Dilday, D. Howell, S. Cenko, J. Silverman, P. Nugent, M. Sullivan, S. Ben-Ami, L. Bildsten, M. Bolte, M. Endl *et al.*, "Ptf 11kx: A type ia supernova with a symbiotic nova progenitor," *Science*, vol. 337, no. 6097, pp. 942–945, 2012.

[4] S. Perlmutter, "Nobel lecture: Measuring the acceleration of the cosmic expansion using supernovae," *Reviews of Modern Physics*, vol. 84, no. 3, p. 1127, 2012.

[5] G. P. Compo *et al.*, "The twentieth century reanalysis project," *Quarterly Journal of the Royal Meteorological Society*, vol. 137, no. 654, pp. 1–28, 2011. [Online]. Available: http://dx.doi.org/10.1002/qj.776

[6] D. Wuebbles, G. Meehl, K. Hayhoe, T. R. Karl, K. Kunkel, B. Santer, M. Wehner, B. Colle, E. M. Fischer, R. Fu *et al.*, "Cmip5 climate model analyses: climate extremes in the united states," *Bulletin of the American Meteorological Society*, 2013.

[7] J. Kern, R. Alonso-Mori, R. Tran, J. Hattne, R. J. Gildea, N. Echols, C. Glöckner, J. Hellmich, H. Laksmono, R. G. Sierra *et al.*, "Simultaneous femtosecond x-ray spectroscopy and diffraction of photosystem ii at room temperature," *Science*, vol. 340, no. 6131, pp. 491–495, 2013.

[8] K. T. Knox, "Enhancement of overwritten text in the archimedes palimpsest," in *Electronic Imaging 2008*. International Society for Optics and Photonics, 2008, pp. 681 004–681 004.

[9] G. Ceder and K. Persson, "How supercomputers will yield a golden age of materials science," *Scientific American, Dec*, 2013.

[10] I. E. Castelli, D. D. Landis, K. S. Thygesen, S. Dahl, I. Chorkendorff, T. F. Jaramillo, and K. W. Jacobsen, "New cubic perovskites for one-and two-photon water splitting using the computational materials repository," *Energy & Environmental Science*, vol. 5, no. 10, pp. 9034–9043, 2012.

[11] "The office of science data-management challenge." [Online]. Available: http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Final_report_v26.pdf

[12] "Introducing Titan - The World's #1 Open Science Supercomputer," 2013, http://www.olcf.ornl.gov/titan/.

[13] "Energy sciences network (esnet)," 2014, http://www.es.net/ ESnet is the high performance networking facility of the DOE Office of Science.

[14] S. Oral, D. A. Dillow, D. Fuller, J. Hill, D. Leverman, S. S. Vazhkudai, F. Wang, Y. Kim, J. Rogers, J. Simmons, and R. Miller, "Olcf's 1 tb/s, next-generation lustre file system," in *Cray Users Group 2013 Proceedings*, 2013, http://www.olcf.ornl.gov/titan/.

[15] J. Hick, "Storage Systems: 2014 and beyond," 2014. https://www.nersc.gov/assets/pubs_presos/NUG2014Storage.pdf. [Online]. Available: \url{https://www.nersc.gov/assets/pubs_presos/NUG2014Storage.pdf}

[16] V. A. et al., "Approaching Exascale: Application Requirements for OLCF Leadership Computing," Oak Ridge Leadership Computing Facility, National Center for Computational Sciences, Tech. Rep., 06 2013.

[17] S. W. Bruenn, A. Mezzacappa, W. R. Hix, E. J. Lentz, O. E. B. Messer, E. J., Lingerfelt, J. M. Blondin, E. Endeve, P. Marronetti, and K. N. Yakunin, "Axisymmetric ab initio core-collapse supernova simulations of 12-25 m stars," *The Astrophysical Journal Letters*, vol. 767, no. 1, p. L6, 2013. [Online]. Available: http://stacks.iop.org/2041-8205/767/i=1/a=L6

[18] E. Lingerfelt, O. Messer, S. Desai, C. Holt, and E. Lentz, "Near real- time data analysis of core-collapse supernova simulations with bellerophon," in *The International Conference on Computational Science 2015 Proceedings, in press*, 2014.

[19] "dcp," 2014. [Online]. Available: http://fileutils.io/

[20] E. Dart, L. Rotman, B. Tierney, M. Hester, and J. Zurawski, "The science dmz: A network design pattern for data-intensive science," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, ser. SC '13. New York, NY, USA: ACM, 2013, pp. 85:1–85:10. [Online]. Available: http://doi.acm.org/10.1145/2503210.2503245

[21] "The Globus project," 2014. [Online]. Available: http://www.globus.org/

[22] A. Hanemann, J. W. Boote, E. L. Boyd, J. Durand, L. Kudarimoti, R. Lapacz, D. M. Swany, S. Trocha, and J. Zurawski, "Perfsonar: A service oriented architecture for multi-domain network monitoring," in *Proceedings of the Third International Conference on Service-Oriented Computing*, ser. ICSOC'05. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 241–254. [Online]. Available: http://dx.doi.org/10.1007/11596141_19

[23] "perfSONAR Performance Toolkit," 2014. [Online]. Available: http://psps.perfsonar.net/

[24] "Deployments of Network Monitoring Software perfSONAR Hit 1,000," 2014. [Online]. Available: http://www.es.net/news-and-publications/esnet-news/2014/perfsonar-milestone/

[25] "BWCTL - the Bandwidth Test Controller," 2014. [Online]. Available: http://software.internet2.edu/bwctl/

[26] "Iperf and Iperf3," 2014. [Online]. Available: http://fasterdata.es.net/performance-testing/network-troubleshooting-tools/

[27] "nuttcp," 2014. [Online]. Available: http://www.nuttcp.net/

[28] "OWAMP - One-Way Active Measurement Protocol," 2014. [Online]. Available: http://software.internet2.edu/owamp/

[29] S. Shalunov, B. Teitelbaum, A. Karp, J. Boote, and M. Zekauskas, "A One-way Active Measurement Protocol (OWAMP)," RFC 4656 (Proposed Standard), Internet Engineering Task Force, September 2006. [Online]. Available: http://www.ietf.org/rfc/rfc4656.txt

[30] "perfSONAR Dashboard," 2014. [Online]. Available: http://fasterdata.es.net/performance-testing/perfsonar/perfsonar-dashboard/

[31] "bbcp," 2013, http://www.slac.stanford.edu/ abh/bbcp/.

[32] "ESnet network performance knowledge base," 2014. [Online]. Available: http://fasterdata.es.net/

[33] "fio," 2014. [Online]. Available: https://github.com/axboe/fio

[34] J. H. Chen *et al.*, "Terascale direct numerical simulations of turbulent combustion using s3d," *Computational Science & Discovery 2*, pp. 1–31, 2009. [Online]. Available: http://iopscience.iop.org/1749-4699/2/1/015001

[35] H. Nam, J. Hill, and S. Parete-Koon, "The practical obstacles of data transfer: Why researchers still love scp," *Proceedings of the Third International Workshop on Network-Aware Data Management*, 2013.