Enabling PETASCALE Data Movement and Analysis

The flood of data being produced today and the tsunami of data expected within the decade raise a critical end-to-end question: how can researchers best access these data and use them for scientific discovery? To answer this question, the SciDAC Center for Enabling Distributed Petascale Science (CEDPS) is exploring diverse approaches — ranging from highly-scalable data-transfer tools to methods enabling server-side data analysis.

It serves little purpose to accelerate data generation if we cannot also accelerate the time required to move data into the hands of scientists.

The chair of a recent SciDAC workshop observed during his opening remarks that "if airplanes had gotten faster by the same factor as computers over the past 50 years, we would be able to cross the country in just one tenth of a second." To this, some wag at the back of the room retorted, "yes, but it would still take you two hours to get downtown."

The important truth underlying this witticism is that petascale science is an end-to-end problem. Over the past several years, investments by the U.S. Department of Energy (DOE) Office of Science and others have resulted in some of the most powerful supercomputers and scientific facilities ever known, including examples like the computers at Argonne (ANL) and Oak Ridge (ORNL) national laboratories, and facilities such as the Advanced Photon Source (APS), the Spallation Neutron Source, and the Large Hadron Collider (LHC). These facilities allow researchers to produce in a few seconds data that previously would have taken days, and to obtain in days data that were previously unthinkable. However, making sense of these data requires that users be able to access the data, and most users are located far from where the data are produced. It serves little purpose to accelerate data generation (the airplane analogy) if we cannot also accelerate the time required to move data into the hands of scientists (getting downtown).

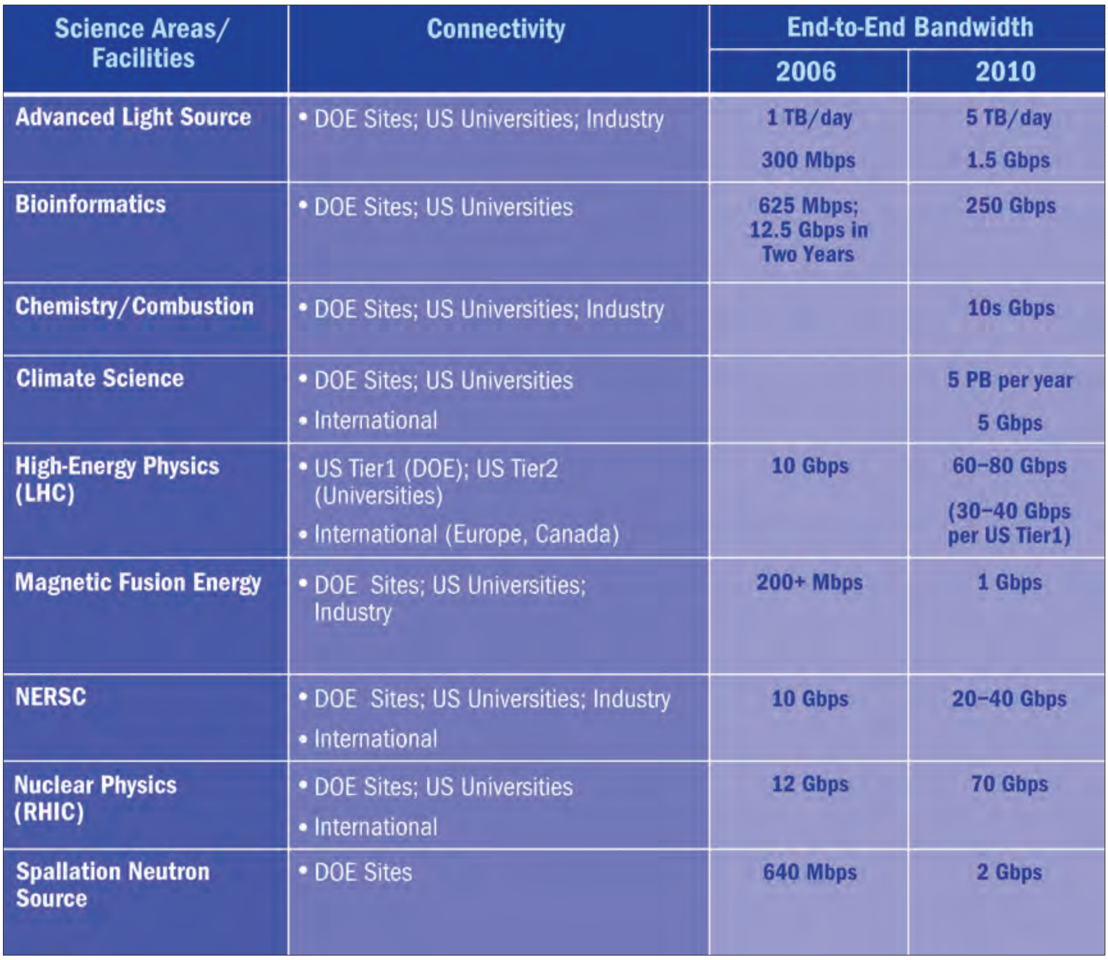
These considerations motivated the establishment of CEDPS, a SciDAC Center for Enabling Technology. CEDPS seeks to reduce time-to-discovery for DOE experimental and simulation science by accelerating access to remote data and software. CEDPS has been pursuing this goal since 2006. As described in this article, CEDPS has achieved significant successes in both the production of useful tools and the use of those tools within DOE and other projects.

Petascale Data in a Connected World

Advances in computational power, detector capabilities, and storage capacity — following Moore's law — are resulting in ever greater data volumes, from not only scientific computations but also experimental facilities. For example, while the Earth System Grid currently manages 200 terabytes (TB) of climate simulation data, including the datasets produced for the recently completed fourth Intergovernmental Panel on Climate Change assessment, the fifth assessment is expected to comprise 50 petabytes (PB) by 2013.

These enormous quantities of data are produced as a result of billions of dollars invested in scientific instrumentation and the associated science. Thus, these data have considerable cost, as well as tremendous value, to scientists worldwide. Viewing science as an end-to-end problem, we must concern ourselves with how to reduce the time that elapses between the data being generated and the distribution of new insights to the scientific community.

It used to be common that a scientist generated data at some facility, then ran analysis programs at the same facility, and finally communicated



Science Areas/ Facilities	Connectivity	End-to-End Bandwidth	
		2006	2010
Advanced Light Source	• DOE Sites; US Universities; Industry	1 TB/day 300 Mbps	5 TB/day 1.5 Gbps
Bioinformatics	• DOE Sites; US Universities	625 Mbps; 12.5 Gbps in Two Years	250 Gbps
Chemistry/Combustion	• DOE Sites; US Universities; Industry		10s Gbps
Climate Science	• DOE Sites; US Universities • International		5 PB per year 5 Gbps
High-Energy Physics (LHC)	• US Tier1 (DOE); US Tier2 (Universities) • International (Europe, Canada)	10 Gbps	60–80 Gbps (30–40 Gbps per US Tier1)
Magnetic Fusion Energy	• DOE Sites; US Universities; Industry	200+ Mbps	1 Gbps
NERSC	• DOE Sites; US Universities; Industry • International	10 Gbps	20–40 Gbps
Nuclear Physics (RHIC)	• DOE Sites; US Universities • International	12 Gbps	70 Gbps
Spallation Neutron Source	• DOE Sites	640 Mbps	2 Gbps

Source: SciDAC CEDPS Illustration: A. Tovey

Figure 1. *Connectivity and bandwidth requirements of selected DOE science areas and facilities, as determined within requirements elicitation workshops conducted by DOE's Energy Science network (ESnet). This table is adapted from a presentation given by Eli Dart of ESnet to the Advanced Scientific Computing Advisory Committee (ASCAC) Networking Subcommittee Meeting on April 13, 2007.*

results via a research publication. This mode of working posed no particular demands on the distributed computing infrastructure. For various reasons, however, this relatively simple mode of working is not always satisfactory. One reason is that, as a result of specialization of infrastructure, it may not be feasible to perform analyses at the site where the data were generated. Indeed, many experimental facilities, and even some supercomputers, provide only limited storage space and post-processing facilities. In these cases, the user has no choice but to transfer data to a home institute or another facility for analysis. In the past, this transfer was often achieved via tape. However, increased data volumes make transfer by tape increasingly problematic. A second reason is that raw data can be of interest to many researchers other than the individual or team who first generated it. Thus, we want to find ways of making that data available to many. A third reason is that, as science becomes more complex and interdisciplinary, people are eager to combine data from multiple sources. For example, climate scientists want to compare results from different climate models, and materials scientists want to image the same sample with both neutrons and photons.

These and other considerations motivate us to pursue strategies to reduce barriers that restrict access to data produced at scientific facilities. Such strategies fall into two categories:

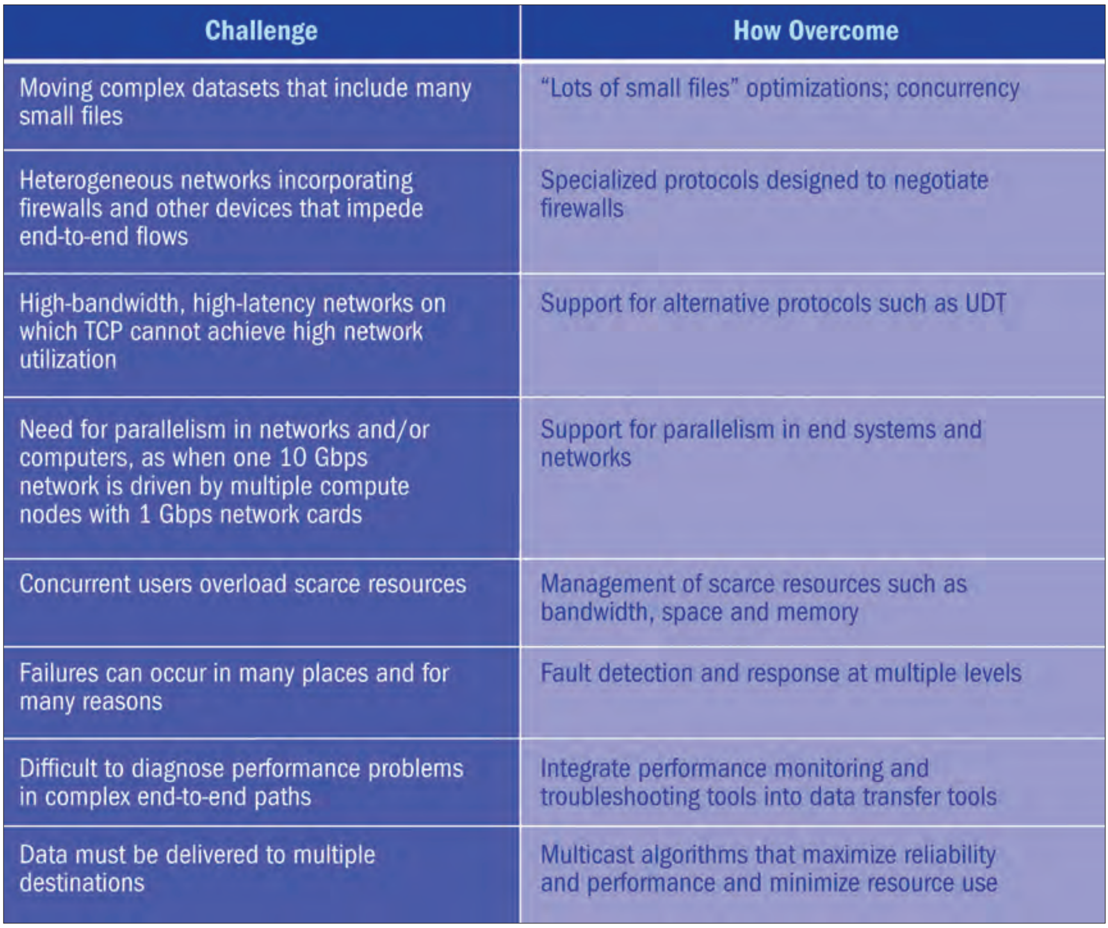
- Moving data to remote users via high-speed networks and software designed to facilitate the rapid, reliable, and secure movement of data

- Facilitating local analysis by remote users, via hardware and software designed to support rapid, reliable, and secure server-side data analysis

These strategies form two of the three CEDPS focus areas. The third area is troubleshooting, because distributed systems — especially high-performance distributed systems — are prone to failure.

CEDPS has achieved significant successes in both the production of useful tools and the use of those tools within DOE and other projects.

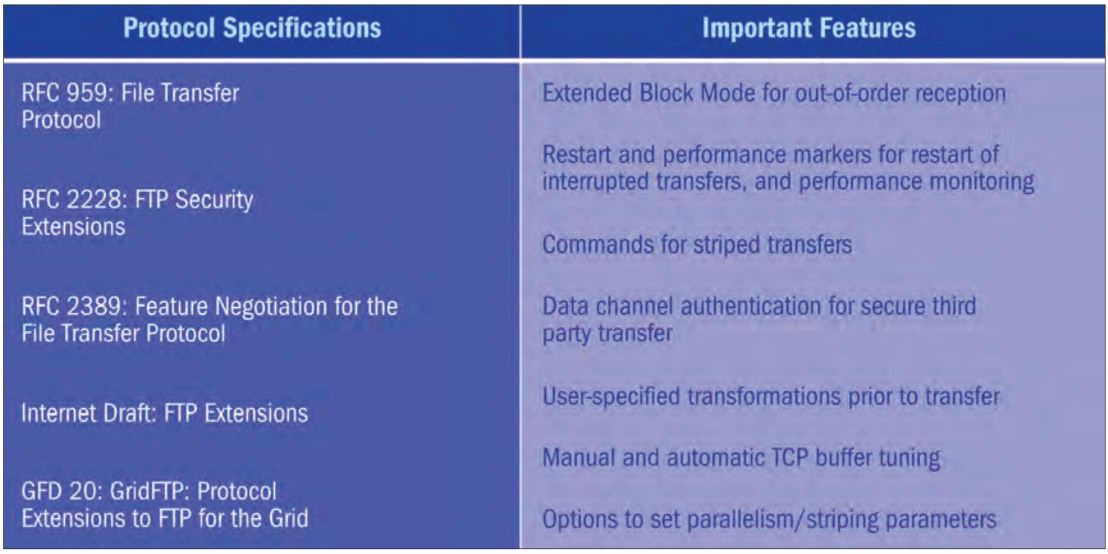
CEDPS has adopted a two-pronged strategy: scaling GridFTP to the petascale, and implementing higher-level behaviors in libraries.



Challenge	How Overcome
Moving complex datasets that include many small files	"Lots of small files" optimizations; concurrency
Heterogeneous networks incorporating firewalls and other devices that impede end-to-end flows	Specialized protocols designed to negotiate firewalls
High-bandwidth, high-latency networks on which TCP cannot achieve high network utilization	Support for alternative protocols such as UDT
Need for parallelism in networks and/or computers, as when one 10 Gbps network is driven by multiple compute nodes with 1 Gbps network cards	Support for parallelism in end systems and networks
Concurrent users overload scarce resources	Management of scarce resources such as bandwidth, space and memory
Failures can occur in many places and for many reasons	Fault detection and response at multiple levels
Difficult to diagnose performance problems in complex end-to-end paths	Integrate performance monitoring and troubleshooting tools into data transfer tools
Data must be delivered to multiple destinations	Multicast algorithms that maximize reliability and performance and minimize resource use

Source: SciDAC CEDPS Illustration: A. Tovey

Figure 2. *Challenges inherent in moving large quantities of data fast over wide-area networks, and potential solutions to those challenges.*



Protocol Specifications	Important Features
RFC 959: File Transfer Protocol RFC 2228: FTP Security Extensions RFC 2389: Feature Negotiation for the File Transfer Protocol Internet Draft: FTP Extensions GFD 20: GridFTP: Protocol Extensions to FTP for the Grid	Extended Block Mode for out-of-order reception Restart and performance markers for restart of interrupted transfers, and performance monitoring Commands for striped transfers Data channel authentication for secure third party transfer User-specified transformations prior to transfer Manual and automatic TCP buffer tuning Options to set parallelism/striping parameters

Source: SciDAC CEDPS Illustration: A. Tovey

Figure 3. *Major protocol specifications used in GridFTP and some of the major features provided by those specifications. RFCs are specifications released by the Internet Engineering Task Force; a GFD is a specification issued by the Open Grid Forum.*

It is unlikely that it will ever be possible or desirable to move all data. However, it is impressive just how much data can be moved. Over the 10 gigabit per second (Gbps) links that are common today, we can (in principle) move 100 TB in a day, if the storage systems and local area networks at the source and destination can sustain 10 Gbps speeds. Emerging 100 Gbps links can increase this number to a petabyte per day. At these speeds, it may still be faster and cheaper overall to ship data by filling a truck with disks, but the inherent complexity is so substantial that it becomes increasingly less attractive. Thus, even experiments operating at the LHC in Geneva, Switzerland, which expect to produce many petabytes per year, plan to transfer those data to the United States via the Internet. These considerations continue to drive bandwidth requirements for DOE networking ever higher, as illustrated in figure 1 (p23).

Putting Data Where You Need Them

Say you have produced 10 TB of data and want to ensure these data are available elsewhere — perhaps at a single location, or perhaps at several — rapidly, reliably, and securely.

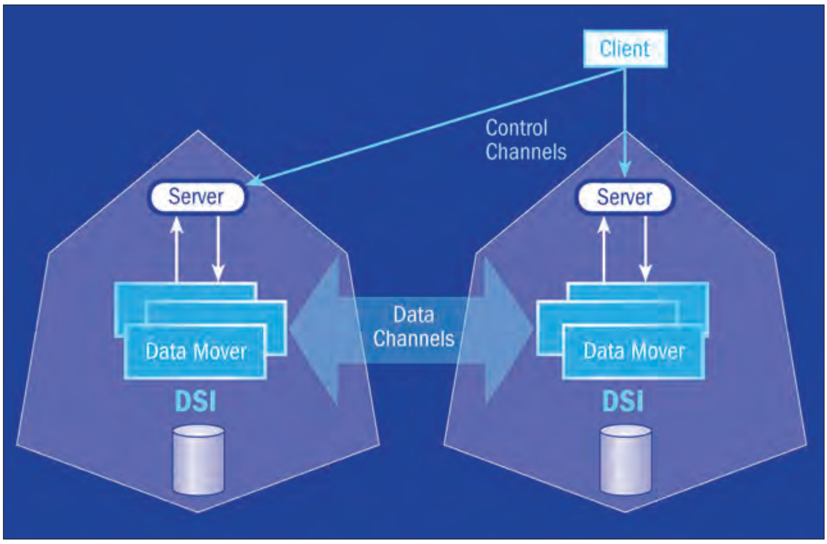
This task is not always easy to achieve. In principle, you should be able to move 10 TB over a modern 10 Gbps network in a less than three hours. In practice you may find that it takes weeks, and much pain and suffering, to achieve such a transfer. Figure 2 lists some of the problems that can arise. The first set of problems relates to performance. Far too often, some component of the end-to-end path (or the top-to-bottom configuration of application, middleware, network protocols, and hardware) is misconfigured, with the result that transfers slow to a crawl. A second set of problems relates to reliability. Even the most carefully managed storage system, firewall, router, network, or computer will occasionally fail; and in the absence of fault tolerance mechanisms, any failure in an end-to-end path will cause the entire transfer to fail.

The solutions to many of these problems (figure 2) are known, at least in principle. But packaging, configuring, and delivering them in a form where they can be used easily can introduce significant challenges. CEDPS has adopted a two-pronged strategy to achieve this goal: scaling GridFTP to the petascale, and implementing higher-level behaviors in libraries.

Scaling GridFTP to the Petascale

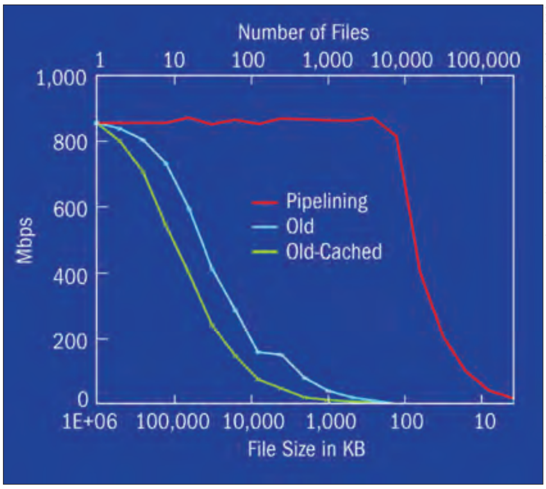
The name GridFTP is commonly used to refer both to a data movement protocol and to software that implements that protocol. The GridFTP protocol specification describes both a profile on existing specifications of and a set of extensions to the popular file transfer protocol (FTP) (figure 3). The result is a protocol that is well-suited for high-performance transport, because of its support for features such as reliability, third-party transfer, striping, and protocol tuning.

Multiple implementations of the GridFTP protocol exist. CEDPS works with two of these implementations: one developed by the Globus team at ANL, and a second developed by the dCache team at Fermilab.



SOURCE: SCIDAC CEDPS ILLUSTRATION: A. TOVEY

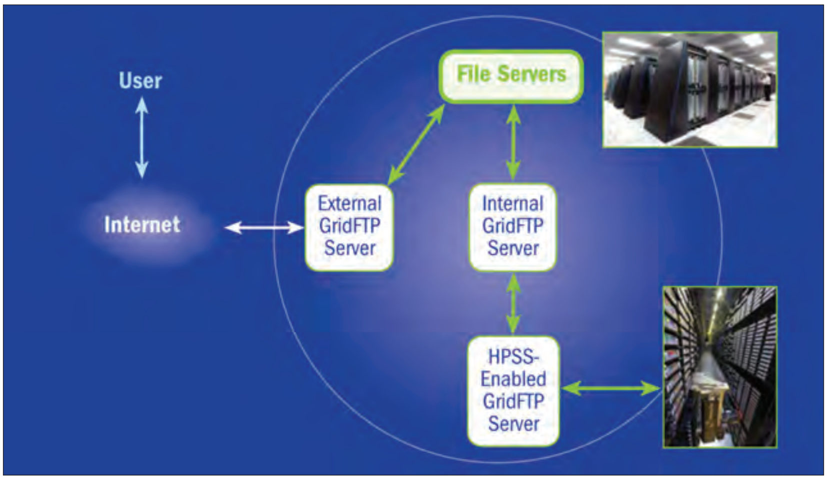
Figure 4. *The Globus GridFTP architecture. Clients can request transfers to/from a server or from one server to another (a third party transfer). A GridFTP server front end can control the operation of multiple data movers, in order to support many clients more efficiently and/or to access data striped across multiple disks. A modular data storage interface (DSI) can provide access to a variety of storage systems, including NFS, HPSS, OPENDAP, and Hadoop. The number of data movers can be varied dynamically, for example in response to changing client load.*



SOURCE: SCIDAC CEDPS ILLUSTRATION: A. TOVEY

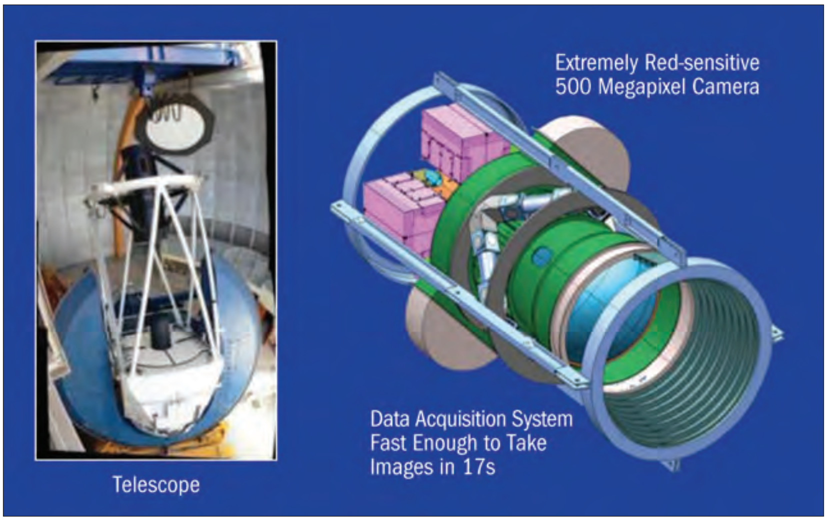
Figure 5. *GridFTP performance with and without "lots of small files" optimizations. In each case, 1 GB is transferred, with varying numbers of files (top x-axis) and file sizes (bottom x-axis). Without pipelining, data transfer performance (shown on the y-axis in Mbps) degrades once file size drops below 100 MB. With pipelining, high performance is sustained down to a file size of about 250 kB.*

The Globus implementation of GridFTP is the most full-featured and efficient available. Its modular architecture (figure 4) permits it to access data stored in a variety of data systems, move data using a variety of transport protocols, and exploit striping for concurrency and high-performance transport. These features enable Globus GridFTP to achieve high end-to-end performance over both local-area and wide-area networks. A popular



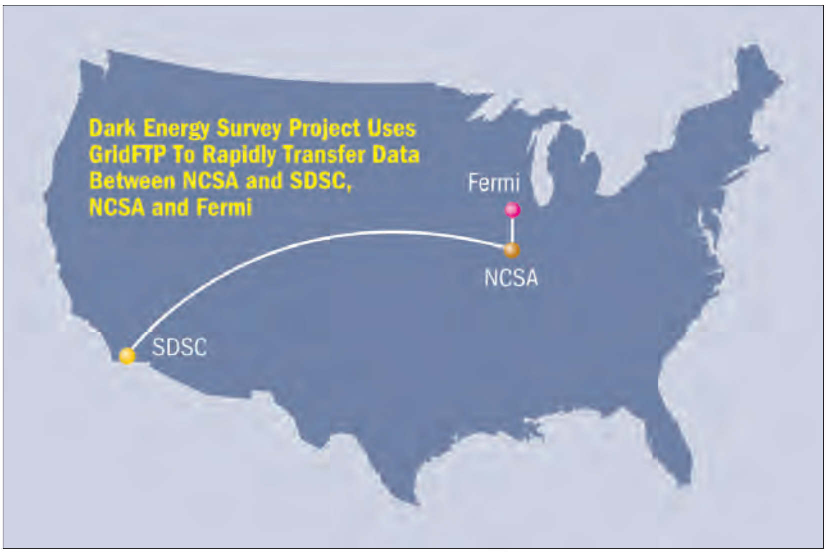
Source: SciDAC CEDPS Illustration: A. Tovey

Figure 6. *The Argonne Leadership Computing Facility (ALCF) has based its data management system on the Globus GridFTP software, using it to manage the movement of data to and from the HPSS mass store, the ALCF's high-performance file servers, and external users.*


Extremely Red-sensitive 500 Megapixel Camera
Data Acquisition System Fast Enough to Take Images in 17s
Telescope

Source: SciDAC CEDPS Illustration: A. Tovey

Figure 7. *The Dark Energy Survey project operates a four meter telescope and is constructing a new 500 megapixel optical camera for its survey.*


Dark Energy Survey Project Uses GridFTP To Rapidly Transfer Data Between NCSA and SDSC, NCSA and Fermi
Fermi
NCSA
SDSC

Source: SciDAC CEDPS Illustration: A. Tovey

Figure 8. *The Dark Energy Survey project uses GridFTP to transfer large volumes of data across the country.*

recent enhancement permits authentication using SSH credentials as an alternative to DOE X.509 certificates. Another recent enhancement achieves dramatic performance improvements for many small files, via the pipelining and concurrent execution of many transfer commands (figure 5, p25).

The following examples illustrate the impact GridFTP is having on DOE science.

- The Argonne Leadership Computing Facility (ALCF) has based its data management system on the Globus GridFTP software, using it to manage the movement of data to and from the HPSS mass store, the ALCF's high-performance file servers, and external users (figure 6). This capability provides performance of 400 megabytes per second (MB/s), superior by 10% to the alternative performance of 360 MB/s for PFTP on a local-area network when using a single data mover. Support for multiple data movers, as developed by CEDPS, increases performance to 1.72 gigabytes per second (GB/s) on 75% of available nodes and should reach 2 GB/s when all are used.

- The DOE Energy Sciences Network (ESnet) recently announced speedups of 20 times or more, to 200 MB/s, over wide area links from the National Energy Research Scientific Computing Center (NERSC) to the Oak Ridge Leadership Computing Facility (OLCF). These speedups were made possible by the use of GridFTP. "I admit to waiting more than an entire workday for a 33 GB input file to SCP and feeling extremely discouraged knowing I had 20 more to transfer," says Dr. Hai Ah Nam, a computational scientist in the OLCF Scientific Computing Group researching the nuclear properties of carbon-14. She now transfers 40 TB of information between NERSC and OLCF for each nucleus she studies — and each such transfer takes less than three days.

- APS users in Australia report similar speedups relative to the SCP tool, which makes it possible for data from experiments to be transferred by network rather than courier. Also, APS is using GridFTP for automated data movement between its beamline data acquisition machine and its HPC cluster, where the acquired data are processed (figure 7). Earlier, this data movement was manual, and APS was getting a data rate of 23 MB/s using a windows native protocol. With GridFTP, the data rates are significantly better — about 110 MB/s, a five-fold improvement. It allows the entire acquired dataset to be moved in the downtime between samples; thus, the samples do not build up on the acquisition machine.

- The Dark Energy Survey is a five-year optical imaging survey project led by Fermilab to make precision measurements of the sky, aimed at determining the

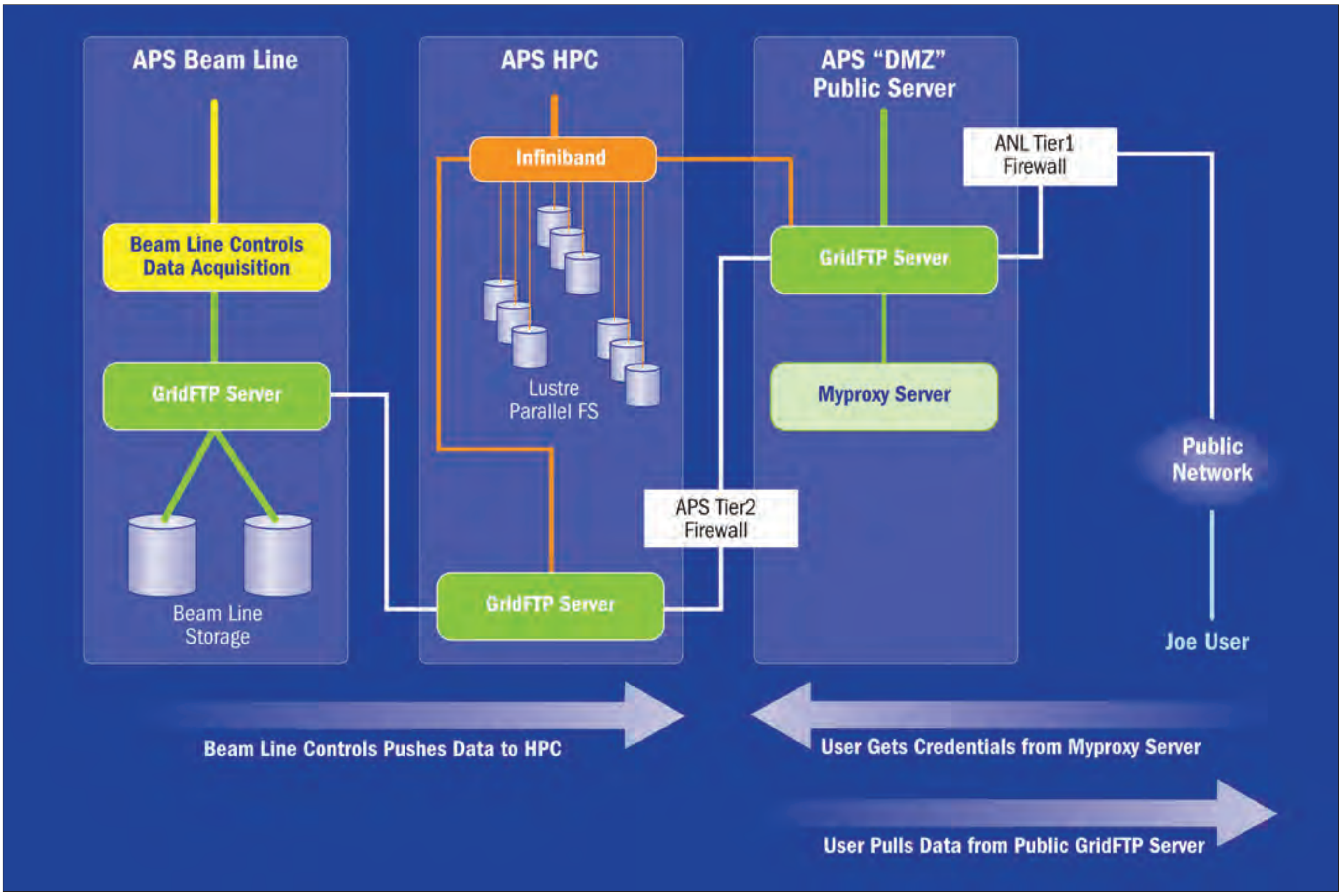


Figure 9. *Automated data movement using GridFTP at APS.*

Source: SciDAC CEDPS Illustration: A. Tovey

reasons the Universe is accelerating (figure 8). To accomplish this goal, researchers will use the Large Synoptic Survey Telescope to survey more than 20,000 square degrees to an unprecedented depth over 10 years, creating a database of more than 100 Pb. The telescope's data torrent presents significant challenges for near-real-time analysis of data streams. Preliminary experiments are using GridFTP for access to its enormous datasets (figure 9). "The project has needed to transfer data between NCSA and SDSC, between NCSA and Fermilab, from dedicated servers to shared supercomputer platforms, from parallel file systems to tape archives. For all of these use cases, GridFTP has provided the flexible and high-performance solution required by the Dark Energy Survey Data Management project," says Dr. Greg Daues, a research programmer at NCSA who works on the Dark Energy Survey project.

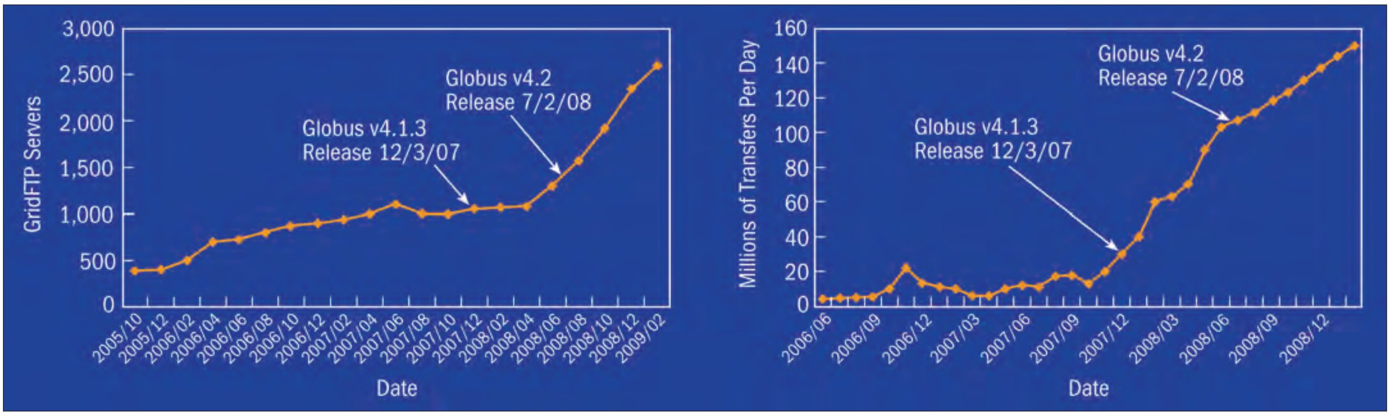
CEDPS enhancements to Globus GridFTP seem to be correlated with substantial increases in both the number of GridFTP servers reporting using GridFTP's optional usage reporting feature and the number of reported downloads (figure 10, p28).

CEDPS also supports work on a second implementation of GridFTP that was incorporated in the dCache distributed storage system used in high-energy physics experiments. One important enhancement to dCache that GridFTP introduced was the support for checksum calculations — an important capability given that the 16-bit Transmission Control Protocol checksum is widely viewed as inadequate given the large amounts of data transferred over modern networks. This feature detects an average of 40 errors per million transfers on data transferred by the D0 experiment at Fermilab and an unrevealed (but we expect similar) number in the more than 10 TB of data transferred per day for the European Organization for Nuclear Research's Compact Muon Solenoid experiment.

Raising the Level of Discourse

Rapid, secure, and reliable movement of individual files (and sets of files) is a necessary capability for many DOE science projects. However, it is not in itself sufficient for many purposes. In particular, many users ask for more sophisticated management functionality, so that they can, for example, request that "only files that have changed, in this directory, should be replicated." CEDPS is developing a variety of higher-level tools and services to meet this sort of demand.

Rapid, secure, and reliable movement of files is a necessary capability for many DOE science projects. CEDPS is developing a variety of higher-level tools and services to meet this sort of demand.



Source: SciDAC CEDPS Illustration: A. Tovey

Figure 10. *Opt-out usage reporting integrated into GridFTP provides partial data on usage. Shown here (left) is the number of unique GridFTP servers with usage reporting enabled, and (right) the number of transfers per day reported by those servers, over a roughly three-year period.*

Data mirroring refers to the synchronization of source and destination directories. A mirroring operation involves first the evaluation of the state of these directories followed by the updating of the destination to bring it into synchrony. The evaluation of the state of synchrony of the directories may include comparing file names, sizes, modification timestamps, and/or checksums. Ideally, a data mirroring tool copies only files or portions of files that do not currently exist or that require updating at the destination site.

Data mirroring may be required for a variety of reasons. For example, we may want to increase availability in case of hardware failures, network outages, or even natural disasters. Alternatively, we may wish to improve performance by allowing a dataset to be accessed from the best of several locations or simultaneously from multiple locations. Probably the most common reason is a desire to have a copy of essential datasets at home institutions. We may also want to replicate data to a storage system near a specific computational resource to improve performance.

CEDPS has developed a data mirroring tool called globus-url-sync to provide secure and efficient data mirroring for high-bandwidth and large data environments. Popular UNIX mirroring tools, such as rsync, were originally developed for low-bandwidth environments. Thus, these tools incur high CPU costs, because they use aggressive checksumming to minimize data transport. In addition, they do not support high-performance data protocols such as GridFTP and provide only limited data transfer security. For these reasons, the tools are not often used in scientific environments, particularly when large datasets are involved. The globus-url-sync tool extends the functionality of the popular globus-url-copy command line tool to mirror a source and destination directory, transmitting only modified files and using GridFTP as the underlying transfer protocol.

The globus-url-sync tool has been designed in collaboration with Stephen Miller's Neutron Scattering Science Division of the Spallation Neutron Source at ORNL. Their data mirroring requirements include sharing within their collaboration and also with the APS group at ANL. We are in the process of deploying this mirroring tool within the ORNL environment and performing initial data mirroring operations.

DataKoa (Further Reading, p33) is a hosted data movement service — that is, a service operated by a third party to which a user can hand off the problem of managing data movement among two or more locations in a distributed system. CEDPS researchers believe that a hosted service can achieve better reliability and a more rapid response to problems than a service operated by an individual user on, for example, a desktop.

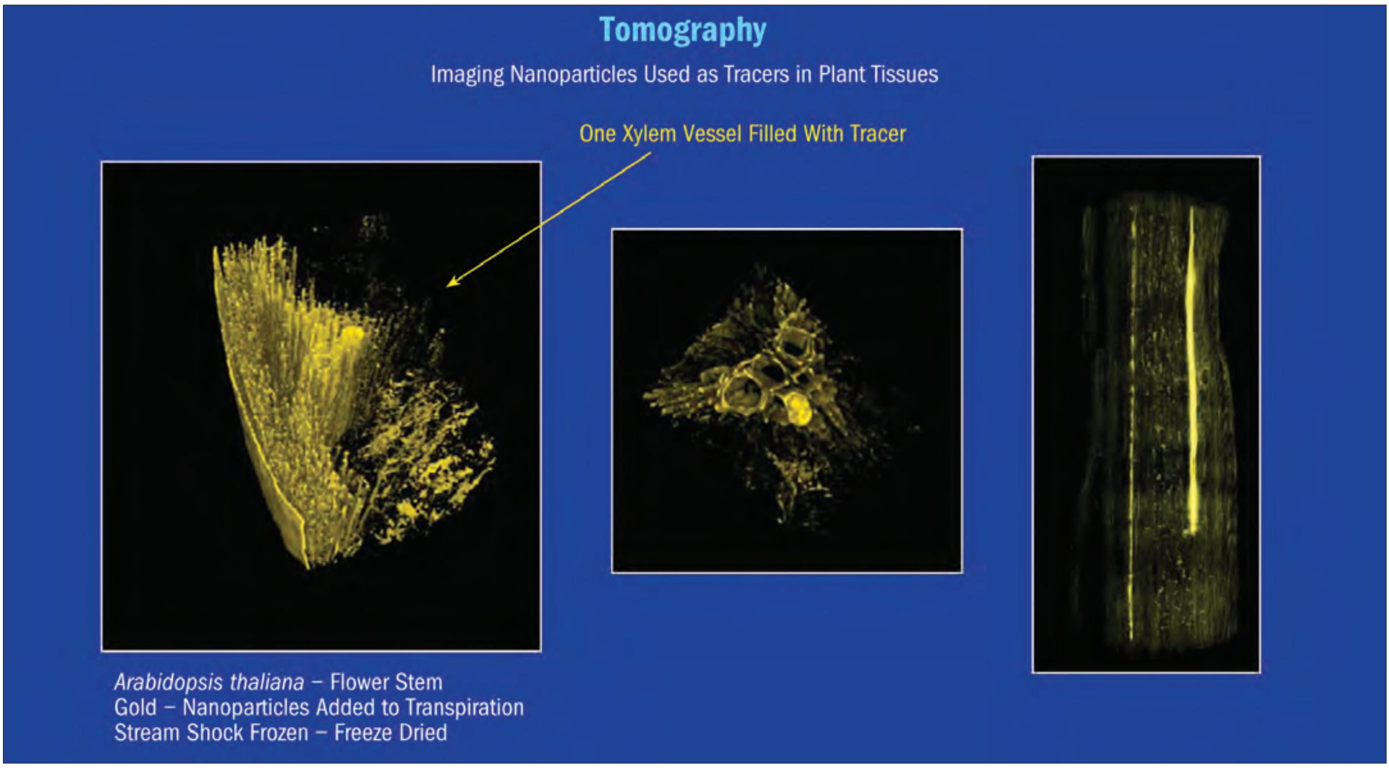
CEDPS researchers believe that a hosted service can achieve better reliability and a more rapid response to problems than a service operated by an individual user on, for example, a desktop.

Scalable Services

Moving data to computation is not always feasible — and may be expected to be less feasible in the future, as data volumes continue to grow. Thus, we seek methods for enabling remote access to code and easily moving computation to remote computers. Two major initiatives in the CEDPS Services area address these two challenges.

Service-Oriented Science Tools

CEDPS has worked closely with researchers at Ohio State University and elsewhere to produce, deploy, and evaluate tools for creating, deploying, publishing information about, discovering, invoking, and composing web services — services that encapsulate some specific data or software of interest to remote users. In particular, CEDPS has developed tools for wrapping science applications as services — the grid Resource Allocation and Virtualization Environment (gRAVI) — and has worked with various DOE teams to apply these tools in different settings.


Tomography
Imaging Nanoparticles Used as Tracers in Plant Tissues
One Xylem Vessel Filled With Tracer
Arabidopsis thaliana – Flower Stem
Gold – Nanoparticles Added to Transpiration
Stream Shock Frozen – Freeze Dried

Source: SciDAC CEDPS Illustration: A. Tovey

Figure 11. *A three-dimensional density map of a sample taken using gRAVI.*

The CEDPS team has worked with the team at Ohio State University in developing Introduce, which provides one-stop shopping for building a fully-functional service from scratch. Introduce aims to reduce the service development and deployment effort by hiding low-level details of the Globus Toolkit and to enable the implementation of strongly typed Grid services. Introduce also hides the complexity in developing and deploying secure web services. Introduce is built using a plug-in architecture that gRAVI utilizes to automatically generate Grid services around applications. The end user can start with an application or an executable and create a fully functional, secure Grid service around that application without writing a single line of code. Also, gRAVI annotates the services generated with appropriate metadata that could help scientists discover these services and use them in their workflows. In addition, gRAVI generates code to register the generated service to a centralized registry with appropriate metadata that would make it easier for users to discover the service. Introduce and gRAVI together provide the full spectrum of tools that are required to create, secure, deploy, and register application services.

The CEDPS Services recently completed a full implementation of rapid creation and deployment of application services using gRAVI. APS is an early adopter of gRAVI in the area of high-throughput tomography. Tomography at APS uses high-intensity X-rays to create many projections through a sample across a range of angles. These projected images contain information about how the X-rays were absorbed through various paths in the sample. A numerical processing technique can then be used on the projections to develop a three-dimensional density map of the sample known as a reconstruction (figure 11). This density map is a nondestructive look at the interior structures of the sample. Tomography, therefore, is a powerful tool to nondestructively look at the detailed structure within an optically opaque sample. Tomography is used to study a wide array of samples from many disciplines, such as materials science, biology, and medicine.

High-throughput tomography requires that a number of computational resources be available to efficiently process raw data into high-quality reconstructions. At APS, a parallel processing cluster-coupled with a multi-terabyte disk array has been employed to process data at rates comparable to acquisition rates. Additional beamlines have since developed the ability to perform tomography experiments. The challenge at APS has thus become how to provide support for the same high-performance computational systems without necessarily duplicating the costs. A collaborative effort with the CEDPS team led to the development of an implementation that may answer this challenge. The group at APS used gRAVI to wrap analytical routines as secure services and created workflows that provide secure local and remote access to locate sample data within the system, browse the data, and create and

A collaborative effort with the CEDPS team led to the development of an implementation that may answer this challenge ...

A Cloud for STAR

The advantages of cloud computing were dramatically illustrated in March 2009 by researchers working on the STAR nuclear physics experiment at Brookhaven National Laboratory's Relativistic Heavy-Ion Collider. The STAR scientists had a late-coming simulation request to produce results for the Quark Matter conference, but all the computational resources were either committed to other tasks or did not support the environment needed for STAR. Fortunately, working with the Nimbus team, the STAR researchers were able to dynamically provision virtual clusters and run the additional computations just in time.

This most recent achievement is a culmination of a few years of close collaboration between the Nimbus team and the STAR experiment. Nimbus is an open source cloud computing infrastructure consisting of tools that allow users to deploy virtual machines (VM) on resources in a manner similar to Amazon's EC2 (Workspace Service) and then combine them into "turnkey" virtual clusters, sharing security and configuration context, that can be used as deployment platforms for scientific computations (Context Broker).

The STAR team recognized the benefits of virtualization early on: it allows scientists to configure a VM image exactly to their needs and have a fully validated experimental software stack ready for use. A virtual cluster composed of hundreds of such images can be deployed on remote resources in minutes. In contrast, Grid resources available at sites not expressly dedicated to STAR can take months to have their configuration adapted to support the STAR environment. Furthermore, since software provisioning and updates of remote sites are hard to track, the scientists cannot always be sure that their data productions are consistent.

The STAR scientists started out by developing and deploying their VMs on a small Nimbus cloud configured at the University of Chicago. Once the virtual machines were deployed, they used the Nimbus Context Broker to configure them into Open Science Grid (OSG) clusters where a job could be submitted just as if it were another OSG cluster. However, since STAR production runs require hundreds of nodes, the collaborating teams soon started moving those clusters to commercial infrastructure. The Nimbus Context Broker was adapted to work with EC2, and a gateway was developed to facilitate access. The first STAR production run on EC2 took place in September 2007, and over the following year the STAR scientists in collaboration with the Nimbus team made further progress, evaluating the performance and successfully conducting a few non-critical runs preparing ground for elastic use of EC2 in the latest run.

CEDPS has contributed to the development of the Nimbus virtual machine provisioning software, which in turn has developed a substantial user community within and beyond DOE.

run simple workflow systems to automate processing and delivery of large numbers of samples.

The gRAVI tools have also been adopted enthusiastically at NERSC for rapid virtualization and provisioning of applications. In addition, the tools are being used in the cancer Bioinformatics Grid and the Cardio Vascular Grid Research projects, both sponsored by the National Institutes of Health and the OMII-UK team.

Infrastructure as a Service (IaaS)

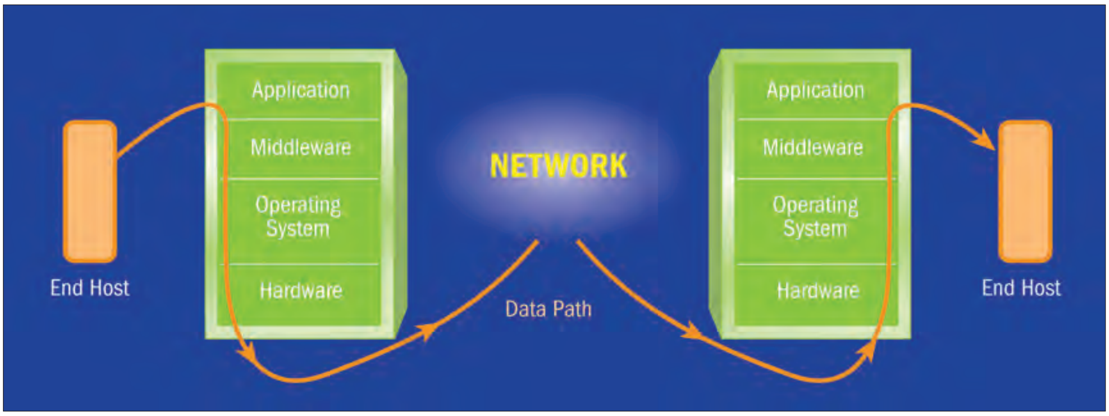
Also known as "cloud," the IaaS has emerged as a useful approach to delivering computing and storage capabilities to users who do not want or cannot afford to operate their own dedicated systems (sidebar "A Cloud for STAR"). Use of virtual machine technologies also addresses code portability issues that often bedevil scientific projects. For example, the Solenoid Tracker at RHIC (STAR) team tells us that it can take several weeks to install and validate on a new computer the more than 200 packages that comprise their data analysis system. With virtual machines, it suffices to download and start up the virtual machine image.

CEDPS has contributed to the development of the Nimbus virtual machine provisioning software, which in turn has developed a substantial user community within and beyond DOE. An early application success was enabling the first production run of the nuclear physics STAR applications on Amazon's EC2 cloud computing infrastructure in September 2007. The deployment of the STAR cluster on EC2 was orchestrated by the Nimbus Context Broker service that enables automatic and secure deployment of turnkey virtual clusters, bridging the gap between functionality provided by EC2 and the end product that scientific communities need to deploy their applications. Scientific production runs require careful and involved environment preparation and reliability; this run was a significant step toward convincing the broad STAR community that real science can be done using cloud computing.

Troubleshooting

You have just created 10 TB of data. You fire up your GridFTP client to move these data to a remote computer for analysis, and you are told it should be done in three hours. It is late in the day, so you head home and plan to start the analysis tomorrow. The next morning, however, you discover that the transfer is only 10% done. Apparently something went wrong or was wrong all along. The cause could be your client settings, your host computer, the network gateway, the network itself, or the receiver's hardware or software. How can you tell where the problem lies, and how to correct it?

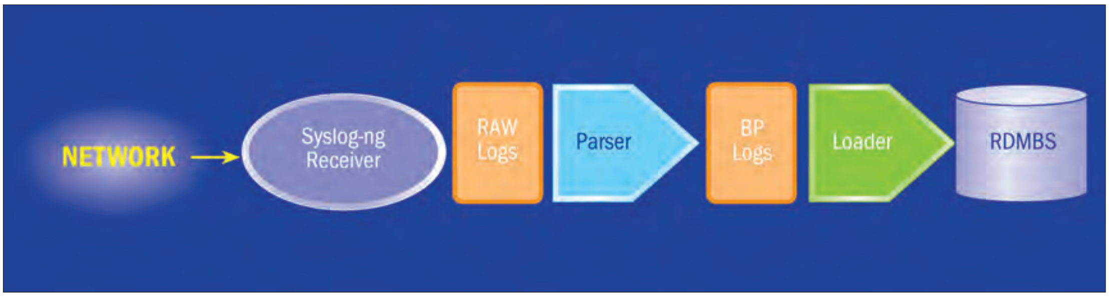
This vignette captures the essence of the problem that CEDPS troubleshooting researchers seek to solve. It is a difficult problem because wide area networks inevitably span administrative domains. Sending data involves engaging a multiscale complex system, just as driving your car involves engag-



Source: SciDAC CEDPS Illustration: A. Tovey

Figure 12. *The CEDPS troubleshooting group focuses on all aspects of the data problem, since virtually any component can impact successful data transfer.*

CEDPS troubleshooting focuses on the end-to-end problem — that is, on correlating all available performance information from the application layer, middleware, host, and network.



Source: SciDAC CEDPS Illustration: A. Tovey

Figure 13. *Framework developed by CEDPS and incorporated into the NetLogger project.*

ing the complex internal machinery of a modern automobile. The big difference between the network and the automobile, from the perspectives of complexity and reliability, is that the set of components that make up your car have been designed and tested to work together (you hope!) by one manufacturer, using one set of specifications, quality control procedures, and so on. An end-to-end data path, on the other hand, must mesh together applications, middleware components, and networks that are developed and operated independently.

CEDPS troubleshooting focuses on the end-to-end problem — that is, on correlating all available performance information from the application layer, middleware, host, and network (figure 12). In order to perform these correlations, we have developed a framework for collecting monitoring data from a distributed system, then normalizing this information into a common data model and loading it into a relational database (figure 13). This framework has been incorporated into a software project that predates CEDPS, called NetLogger, and thus it is called the NetLogger Pipeline. To collect logs, we employ the widely used open-source software *syslog-ng*. To normalize and load the performance information into a database, we use a lightweight set of Python modules.

The flexibility of the NetLogger Pipeline is evidenced by the variety of applications to which it has been applied, including:

- On the NERSC Parallel Distributed Systems Facility (PDSF), analysis of data from STAR Berkley Storage Manager (BeStMan) data transfers have revealed unexpected network performance, which has triggered PDSF upgrades and configuration changes.

- The NERSC Project Accounts team is using the NetLogger Pipeline to normalize the logs and log database to perform traceability analysis.

- The Pegasus team at USC/ISI uses the NetLogger Pipeline for large CyberShake workflows. Special analysis tools were able to efficiently analyze execution logs of earthquake science workflows consisting of upwards of a million tasks (figure 14, p32).

- The Tech-X STAR job submission portal uses the NetLogger Pipeline database to drill down to site-specific information for a STAR portal job. A prototype of this functionality was demonstrated at SC08.

Storage Resource Managers (SRMs) are middleware components that provide a common access interface, dynamic space allocation, and file management for shared distributed storage systems. A natural use of SRMs is the coordination of large-volume data streaming between

Source: SciDAC CEDPS Illustration: A. Tovey

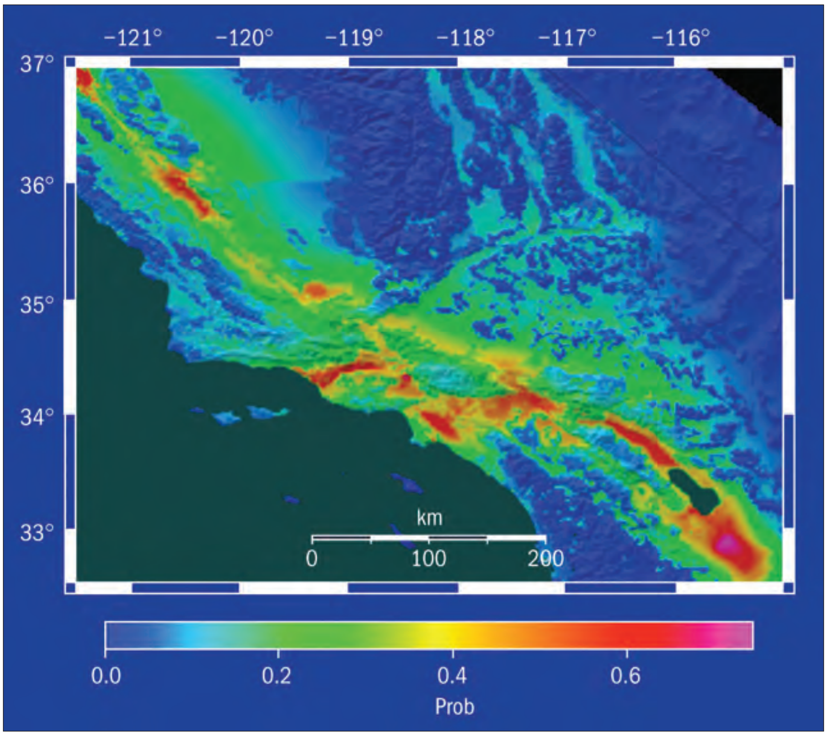


Figure 14. *Earthquake movement from the San Joaquin Valley, California to Mexico, across the Los Angeles basin, moments after a simulated rupture. Such simulations can generate up to 40 TB of data.*

Source: SciDAC CEDPS Illustration: A. Tovey

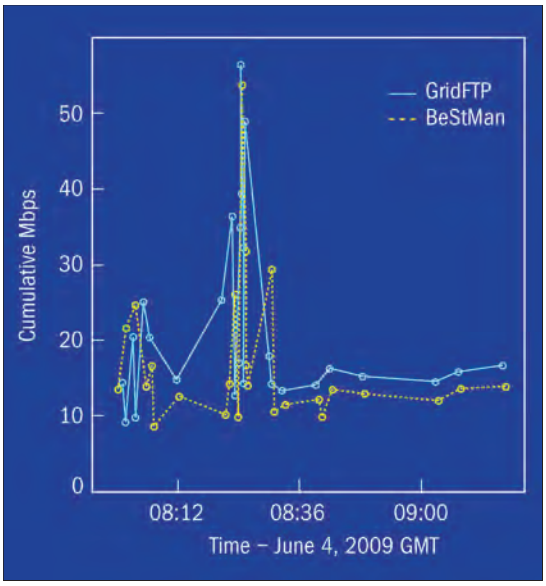


Figure 15. *Throughput of BeStMan client at NERSC communicating with GridFTP servers at Brookhaven National Laboratory, as seen from both ends of the connection.*

sites. Optimizing the performance of SRMs over high-speed network interfaces is an important challenge that impacts the work of science research such as the STAR and Earth System Grid projects. The most recent version of an SRM, developed at Lawrence Berkley National Laboratory (LBNL) and currently deployed in several projects by the SciDAC Storage Data Management Center, is called BeStMan. When managing multiple files, BeStMan can take advantage of the available network bandwidth by scheduling multiple concurrent file transfers. CEDPS and storage Data Management researchers have been working together to understand and optimize the performance of BeStMan.

The STAR experiment uses BeStMan to move datasets between Brookhaven National Laboratory (BNL) and the PDSF analysis cluster at NERSC. In this case, BeStMan communicates with GridFTP servers at BNL. Using CEDPS tools, we can view the transfer performance as seen by each endpoint (figure 15). The GridFTP and BeStMan estimates of bandwidth roughly track each other, although there are some discrepancies because each component records the start and end time of a transfer slightly differently.

More important is that initial analyses of the throughput between these sites revealed that the data transfer rate was far below what anyone expected. Note in figure 15 that the upper limit is around 50 megabits per second (Mbps); independent tests could achieve upwards of 500 Mbps! (For more details, see "PDSF–BNL Bandwidth" in Further Reading).

This result triggered a flurry of activity that resulted in three discoveries:

- The network interface controller in the end-host at PDSF was 10 times slower than the cross-country link from California to New York. This problem was easily fixed but is a classic example of the configuration problems often encountered in the "last mile" of the network.

- The hardware at Brookhaven was older and slower than required. Large data transfers require a lot of memory, because the data nodes need to remember all the data in transit in case some of it is lost. The longer and "fatter" (that is, faster) the network is between the two endpoints, the more memory is required. Again, this problem was easily fixed once recognized.

- Even after end-host upgrades and under the best possible circumstances, the transfer rates are asymmetrical. Data flow roughly twice as fast from BNL to PDSF than the other way around. The reasons for this difference are still being investigated.

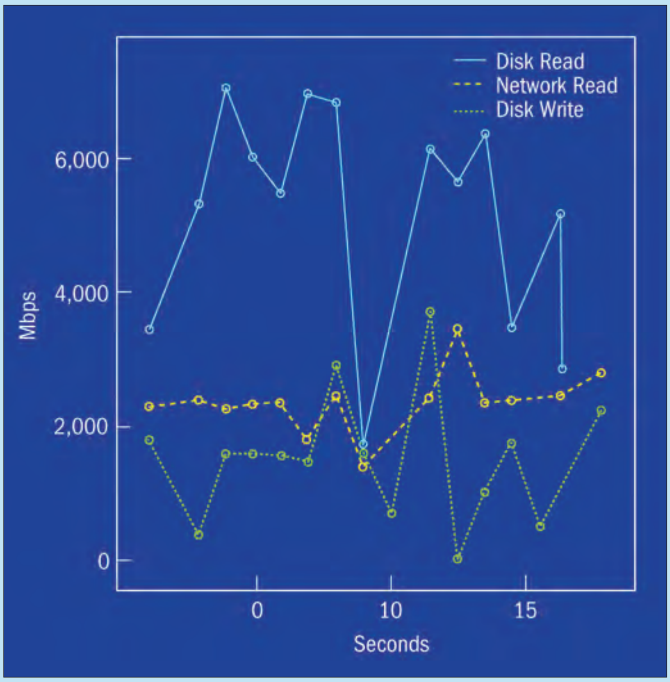
This application also reveals two problems with the current infrastructure. The first is that the monitoring is not truly end-to-end. This arrangement is essentially a political, and not a technical, problem — establishing agreement on

Bottleneck Detection: Disk or Network?

As wide-area networks get faster, they are beginning to outpace disks. Until recently, even a single cheap disk could keep up with a very fast and very expensive wide-area network. Now, with 10 Gbps networks deployed and 100 Gbps around the corner, a single disk may not be anywhere near enough. Therefore it is important when optimizing the performance of a data transfer to know which part of the path – the sender disk, the network, or the receiver disk – is the bottleneck. And it is also important to know, particularly for transfers that take several hours, where this bottleneck is before the transfer completes.

To address this problem, the CEDPS project has added features to the NetLogger Toolkit that allow it to efficiently and transparently summarize the current throughput of a data transfer that has had only a few lines of instrumentation added to its code. The frequency of summarization, from one summary per run, to one summary every *N* seconds, to a detailed log of every single read and write operation, can be controlled at run time. The instrumentation has been added to the GridFTP server software so a running server can show whether the disk or network is a bottleneck for every transfer. An option has also been added to the client program so the user can get a bottleneck analysis report.

An example of one-second summaries on a short local transfer (figure 16) shows the write to disk as usually, but not always, the bottleneck.



Source: SciDAC CEDPS Illustration: A. Tovey

Figure 16. *An example Netlogger summary report of disk and network throughput for a short local GridFTP transfer.*

what to monitor, and how, between many distinct hierarchies of command and control. The second is that configuration of data movement endpoints is difficult because there are many different endpoints being used for different projects; what is needed is a single managed set of hardware resources that is dedicated to data transfers. More detail on how CEDPS and other researchers are cooperating to solve both these problems is provided in the sidebar "Bottleneck Detection: Disk or Network?"

The Next Steps

Science is an end-to-end problem: even the most sophisticated numerical simulation and the most advanced experiment become valuable only when the data that they generate is translated into insight. Thus, methods for reliable, secure, and rapid data movement — or for avoiding data movement altogether by enabling effective server-side analysis — are essential elements of a petascale computing solution.

The SciDAC CEDPS is dedicated to providing such methods and ensuring they are applied effectively within DOE science projects. The project's participants have worked on these problems for many years, both individually and (in many cases) together. What distinguishes CEDPS from past efforts is the focus on scaling to the petascale and, in the process, addressing performance and reliability problems that arise in the context of challenging DOE applications.

A current focus of CEDPS work is demonstrating and evaluating the capabilities of CEDPS tools at substantial scale. To that end, we are defining a set of data challenges, in which we seek to move large quantities of data over wide area networks in an efficient, reliable, and hands-off manner. The first such data challenge will involve one million files of average size 10 MB, a second challenge will involve 10 million files, and a third will feature 100 million files. We encourage readers with interesting distributed data problems to contact us. ●

The SciDAC CEDPS is dedicated to providing methods for reliable, secure, and rapid data movement, and ensuring these methods are applied effectively within DOE science projects.

Contributors Dr. Ian Foster, ANL; Dr. Ann Chervenak, USC/ISI; Dr. Dan Gunter, LBNL; Dr. Kate Keahey, ANL; Ravi Madduri, ANL; and Raj Kettimuthu, ANL

Further Reading

DataKoa
http://www.datakoa.org

PFSF–BNL Bandwidth
http://www.cedps.net/index.php/PDSF_-_BNL_bandwidth_measurements