Grids, the TeraGrid, and Beyond

To enable resource and data sharing among collaborating groups of science and engineering researchers, the NSF's TeraGrid will connect multiple high-performance computing systems, large-scale scientific data archives, and high-resolution rendering systems via very high-speed networks.

Daniel A. Reed
National Center for Supercomputing Applications

Imagine correlating petabytes of imagery and data from multiple optical, radio, infrared, and x-ray telescopes and their associated data archives to study the initial formation of large-scale structures of matter in the early universe.

Imagine combining real-time Doppler radar data from multiple sites with high-resolution weather models to predict storm formation and movement through individual neighborhoods.

Imagine developing custom drugs tailored to specific genotypes based on statistical analysis of single nucleotide polymorphisms across individuals and their correlation with gene functions.

Although they involve disparate sciences, these three scenarios share a common theme. They depend on joining a new generation of high-resolution scientific instruments, high-performance computing (HPC) systems, and large-scale scientific data archives via high-speed networks and a software infrastructure that enables resource and data sharing by collaborating groups of distributed researchers.

RESEARCH INFRASTRUCTURE

The National Science Foundation's TeraGrid (www.teragrid.org) is a massive research computing infrastructure that will combine five large computing and data management facilities and support many additional academic institutions and research laboratories in just such endeavors. Scheduled for completion in 2003, the TeraGrid is similar in spirit to projects like the United Kingdom's National e-Science grid and other grid initiatives launched by the academic community, government organizations, and many large corporations. All are creating distributed information technology infrastructures to support collaborative computational science and access to distributed resources.

The TeraGrid will be one of the largest grid-based HPC infrastructures ever created. As Figure 1 shows, initially, the TeraGrid will tightly interconnect the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign; the San Diego Supercomputer Center (SDSC) at the University of California, San Diego; the Argonne National Laboratory (ANL) Mathematics and Computer Science Division; and the California Institute of Technology (Caltech) Center for Advanced Computing Research. The Pittsburgh Supercomputing Center (PSC) recently joined the TeraGrid as well.

When operational, the TeraGrid will help researchers solve problems that have long been beyond the capabilities of conventional supercomputers. It will join large-scale computers, scientific instruments, and the massive amounts of data generated in disciplines such as genomics, earthquake studies, cosmology, climate and atmospheric simulations, biology, and high-energy physics.

Just as today's Internet evolved from a few research networks, like the early Arpanet, we expect the TeraGrid to serve as an accretion point for integrated scientific grids. Conjoining discipline-specific, regional, and national grids promises to create a worldwide scientific infrastructure. This active infosphere will enable international, multidisciplinary scientific teams to pose and answer questions by tapping computing and data resources without regard to location. It creates a new paradigm in which scientists can collaborate anytime, anywhere on the world's most pressing scientific questions.

TERAGRID COMPUTING ARCHITECTURE

HPC clusters based on the Linux operating system and other open source software are the Tera-

Published by the IEEE Computer Society
0018-9162/03/$17.00 © 2003 IEEE

Figure 1. TeraGrid partners map. Initially, the TeraGrid will provide an infrastructure that interconnects five large computing and data management facilities and supports additional academic institutions and research facilities.

Grid's fundamental building blocks. Intel's 64-bit Itanium2 processors, commercially introduced in July 2002, power these Linux clusters. The terascale computing capabilities of each cluster accrue by aggregating large numbers of Itanium2 processors to form a single, more powerful parallel system. The "Building the Next-Generation Itanium Processor" sidebar provides information about the development of these processors.

The major components of the TeraGrid will be located at the TeraGrid sites and interconnected by a 40 Gbps wide area network:

Building the Next-Generation Itanium Processor

Intel's Itanium processor family is at the heart of the TeraGrid, expected to be one of the world's largest scientific high-performance computing infrastructures. Gadi Singer, vice president and general manager of Intel's PCA Components Group has this to say about the Itanium processor development process:

> When you link computers to perform a single task, you must look at the processor from the outside in—what it must do—and the inside out—how it does the job.
>
> High-performance computing tasks require scientific computing support on the processor, and that means floating-point calculations and the capability to transfer a lot of information very quickly. Because the data space for most scientific problems is huge, high-performance computing requires a very large address space. It requires 64-bit addressing (rather than conventional 32-bit addressing) and a lot of memory that is accessible near the processor. The Intel Itanium2 processor has 3 Mbytes of on-die cache. It also provides the high levels of RAS—reliability, availability, and serviceability—that high-performance computing needs.
>
> To perform floating-point calculations at high speed, Intel had to position many execution units on the chip. The Itanium2 processor contains six integer execution units plus two floating-point units; with these, it can perform up to six instructions every cycle and up to 20 operations. The Itanium2 processor's bus—the pathway for input and output—can transfer 128 bits at 6.4 Gbps. The architecture also provides parallelism—the ability to activate all these resources simultaneously.
>
> Once we designed these features, there was still the challenge of how to efficiently link different pieces of software so they could operate together smoothly. The operating system, compilers, and other pieces are all part of the Itanium architecture design effort.

A Brief History of High-Performance Computing

The desire for high-speed computing can be traced to the origins of modern computing systems. Beginning with Babbage's analytical engine, a mechanical calculating device, the insatiable appetite for systems that operate faster and that can solve ever-larger problems has been the enabler of new discoveries.

World War II brought the first of the large, electronic calculating systems, based on vacuum tubes. Universities, laboratories, and vendors constructed a series of high-performance systems from the 1950s to the 1970s. Of these, one of the most famous was Illiac IV, developed at the University of Illinois and part of the inspiration for HAL in the movie *2001: A Space Odyssey*.

In 1976, Seymour Cray introduced the Cray 1, a water-cooled vector machine that became synonymous with supercomputing. When IBM introduced the personal computer in the early 1980s, forces for massive change in high-performance computing were unleashed. Moore's law, which states that transistor density—and therefore processing power—approximately doubles every 18 to 24 months, led to the high-performance microprocessors found in today's desktops, workstations, and servers.

As microprocessor performance increased, high-performance cluster computing based on "killer micros" became feasible. Clusters have been around since the early 1980s, but clusters started to get serious attention in 1994 as a result of the Beowulf project. In this project, NASA researchers Thomas Sterling and Don Becker built a cluster from 16 commercial off-the-shelf processors linked by multiple Ethernet channels and supported by Linux.

Within two years, NASA and the Department of Energy had fielded computing clusters with a price tag of under US$50,000 that operated at more than 1 Gflop per second. Today, thousands of clusters based on the Beowulf design are in use in laboratories around the world.

Development of the Intel Itanium architecture has set the stage for the next generation of clustered supercomputing systems. The Itanium2 processor, the second generation in the Itanium processor family, uses 64-bit addressing and support for the high-performance floating-point operations central to scientific computing. The Itanium2 processors are at the heart of the TeraGrid.

- NCSA will house the largest of the TeraGrid computing systems. NCSA's TeraGrid cluster will consist of more than 10 Tflops of computing capacity powered by Intel's Itanium processor family, more than 200 Tbytes of high-performance disk storage, and a large tertiary storage system.
- SDSC will house the TeraGrid's primary data and knowledge management facilities. This will consist of an Itanium cluster with a peak performance of more than 4 Tflops, an IBM Power4 cluster with a peak performance of more than 1 Tflop, 500 Tbytes of disk storage, and a large tertiary storage system. A Sun Microsystems high-end server will provide a gateway to grid-distributed data.
- Argonne will deploy high-resolution rendering and remote visualization capabilities, supported by a 1.25 Tflops Itanium cluster with parallel visualization hardware.
- Caltech will provide online access to large scientific data collections and connect data-intensive applications to the TeraGrid. Caltech will deploy a 0.4 Tflop Itanium cluster and associated secondary and tertiary storage systems.

When the Pittsburgh Supercomputing Center joined the project in August 2002, it added more than 6 Tflops of HP/Compaq clusters and 70 Tbytes of secondary storage to the TeraGrid, complementing the original four sites.

When completed, this US$88 million project will include more than 16 Tflops of Linux cluster computing distributed across the original four TeraGrid sites, more than 6 Tflops of HP/Compaq cluster computing at PSC, and facilities capable of managing and storing more than 800 Tbytes of data on fast disks, with multiple petabytes of tertiary storage. High-resolution visualization and remote rendering systems will be coupled with these distributed data and computing facilities via toolkits for grid computing. These components will be tightly integrated and connected through a network that will initially operate at 40 Gbps, a speed four times faster than today's fastest research network.

TERAGRID PARTNERS AND INFRASTRUCTURE

The TeraGrid sites are working with private sector and consortium collaborators to assemble and deploy the TeraGrid infrastructure:

- IBM is providing cluster integration, storage, and software;
- Intel will provide high-performance, 64-bit Itanium and Itanium2 processors;
- Myricom is the vendor for internal cluster networking;
- Oracle will provide database management and data mining software;
- Qwest will deploy the 40-Gbps network connecting the Illinois and California sites; and
- Sun Microsystems will provide metadata management engines.

Intel's Itanium architecture provides the high performance needed for floating-point intensive scientific calculations. Its scalability will help keep pace with the complex calculations associated with the scientific and engineering research projects that will exploit the TeraGrid infrastructure. The "A Brief History of High-Performance Computing" sidebar provides information about the evolution of high-speed computing.

All of the TeraGrid's components will be integrated using open source software like Linux and grid middleware such as the Globus toolkit.[1] Linux community software tools for cluster management form the software base for clusters. Globus features grid software for intracluster interactions such as security and authentication for remote resource access, resource scheduling and management, distributed data management, and wide-area communication.

RESEARCH OPPORTUNITIES

Scientific and engineering research problems ranging from high-energy physics to geosciences to biology are awaiting TeraGrid deployment. All of these projects share several common characteristics:

- complex, often multidisciplinary computational models that require access to the world's most powerful computing systems;
- remote access to distributed data archives containing observational data from a new generation of high-resolution scientific instruments and distributed sensors;
- real-time access to instruments when time-critical phenomena such as severe storms or supernovas occur; and
- national and international collaborations that require worldwide interactions among individuals and organizations.

A broad range of research projects have similar needs, including environmental modeling, weather prediction and climate change research, astronomy, ecology, and physics. In industry, projects such as supply chain management, distributed decision making, and even the design of advanced heavy equipment engines can benefit from computational grids.

Gravitational wave detection

Within the next year, scientists from the California Institute of Technology's Laser Interferometer Gravitational Wave Observatory (LIGO) will begin collecting more than 100 Tbytes of new data annually to further validate Einstein's theories by attempting to detect gravitational waves. Scientists believe these gravity waves result from interactions between massive objects such as orbiting black holes.

Because it must correlate putative events from two detectors, one in Hanford, Wash., and the other in Livingstone, La., LIGO requires grid infrastructure for data analysis. The "What Is a Grid?" sidebar describes the grid computing infrastructure. LIGO also involves challenging computational searches such as correlating detected gravity waves with astrophysical objects.

Earthquake simulation

Earthquake engineers with the National Science Foundation's George E. Brown Jr. Network for Earthquake Engineering Simulation (NEES) hope to fuse observational data from distributed geophysical sensors and shake tables that provide data on the strength of building materials with simulations of building dynamics and responses to earthquakes and physical attacks (www.nees.org).

The NEESgrid (www.neesgrid.org) will let engineers simulate engineering problems over high-speed networks, adjust simulation parameters in real time, and develop buildings and infrastructures that are better able to withstand earthquakes and

What Is a Grid?

In the world of high-performance computing, a grid is an infrastructure that enables the integrated, collaborative use of high-end computing systems, networks, data archives, and scientific instruments that multiple organizations operate. Grid applications often involve large amounts of data or computing and require secure resource sharing across organizational boundaries—something that today's Internet and Web infrastructures do not easily support.[1]

The grid infrastructure for science and engineering research typically includes the following elements:

- *Smart instruments*. Advanced scientific instruments, such as telescopes, electron microscopes, particle accelerators, and environmental monitors—coupled with remote supercomputers, users, and databases—to enable interactive rather than batch use, online comparisons with previous runs, and collaborative data analysis.
- *Data archives*. Scientific data produced by large-scale computations or obtained from high-resolution scientific instruments. Mining, extracting, and correlating data from distributed archives opens the way for new scientific discoveries.
- *Distributed collaboration*. Shared access to data, computing, and discussion, often via high-bandwidth, multiway audio-video conferencing. The Access Grid uses high-resolution video and multicasting to support distributed meetings.
- *High-performance computing systems*. Increasingly, these systems are based on large numbers of commercial microprocessors connected by high-speed local networks and integrated via Linux and other open source software to function as high-performance computing systems. Laboratory clusters typically contain tens of systems, whereas clustered supercomputing systems contain hundreds or thousands of processors.

The Globus Project (www.globus.org/about/faq/general.html) and TeraGrid (www.teragrid.org) Web sites provide more details about grid computing.

Reference

1. I. Foster and C. Kesselman, eds., *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, 1998.

Deep-sky observations and whole-sky surveys have generated more data in the past 10 years than in the entire history of astronomy.

physical attacks. Such an infrastructure, together with sensors embedded in the building, might have enabled engineers to more quickly access the technical causes for the September 2001 collapse of the World Trade Center towers.

Digital sky surveys

High-resolution charge-coupled devices and the deployment of large-aperture telescopes are generating both deep-sky observations and whole-sky surveys of unprecedented scale. Recent examples include the Two Micron All Sky Survey (2MASS) and the Sloan Digital Sky Survey (SDSS). More data has been captured in the past 10 years from these surveys than in the entire history of astronomy. Correlating such data to identify new phenomena and test theories motivated the recent creation of the United States National Virtual Observatory (NVO) and the European Astrophysical Virtual Observatory.

The NVO (www.us-vo.org) and other virtual observatories will create a distributed information infrastructure for astronomy based on federated databases containing ground- and space-based observational data. A separate institution will manage each principal sky survey, which includes data taken at different wavelengths—optical, infrared, radio, and x-ray—and ranging in size from 3 to 40 Tbytes. The TeraGrid will make it possible to link and analyze surveys and to generate important data points and statistics across the entire visible universe, fundamentally changing astronomy research.

High-energy physics

The Standard Model of Physics is arguably one of the most successful scientific theories in human history. Its predictions have been repeatedly validated experimentally, and it successfully explains electromagnetism and the weak and strong nuclear forces. However, physicists still seek a broader "theory of everything" that integrates gravity and explains the origins of mass. One of the elusive particles being sought is the Higgs boson, believed to be the mechanism via which particles acquire mass.

New accelerators like the Large Hadron Collider (LHC) being constructed at CERN (European Nuclear Research Organization) and its ATLAS (A Toroidal LHC Apparatus) and the Compact Muon Solenoid detectors will generate events at 100 Mbps. An international grid of data archives, with the US anchor at the Fermi National Laboratory, will be created to store and analyze this data during the search for new particles.

Systems biology

Roughly one of every 1,000 nucleotides differs across individual humans. These single nucleotide polymorphisms are responsible for the rich diversity of human characteristics. They also provide the source of variability in our response to viral and bacterial attacks and the efficacy of drugs. Large-scale gene sequencing, microarrays for testing gene expression, and other biological data-capture facilities have presaged a new era in our understanding of such biological processes. Fully understanding gene coding for proteins and protein control of metabolic processes via enzymes is one of the 21st century's great scientific challenges.

Building on this understanding to create models of cells, organs, and organisms that researchers can use to develop more effective drugs or repair genetic defects will require comparative analysis of genomes across species and individuals. It will also require grid-based distributed access to genomic, protein, and metabolic databases and computational modeling of biological processes.

The common theme across these examples is the combination of complex computations; large, distributed data archives and instruments; and collaborative research teams. Developing and supporting such an infrastructure is the TeraGrid team's goal.

USING THE TERAGRID: AN EXAMPLE

Consider the scientific vision with which we began—correlating petabytes of imagery and data from multiple optical, radio, infrared, and x-ray telescopes and their associated data archives to study the initial formation of large-scale structures of matter in the early universe. The NVO and the TeraGrid make this possible. Atop the TeraGrid, at NCSA we deploy copies of radio astronomy databases from the National Radio Astronomy Observatory Very Large Array and the Berkeley-Illinois-Maryland Association Millimeter-Wave Array. Meanwhile, Caltech hosts copies of the SDSS, the Roentgen Satellite X-ray Observatory, and 2MASS. Similarly, SDSC hosts a copy of the Palomar Digital Sky Survey.

A research astronomer in Boston poses a query to construct an unbiased sample of galaxy clusters to test structure formation models. The goal is to compare theoretical models with observed mass and luminosity functions over a cosmologically significant redshift range and study the evolutionary effects. This requires correlation of data from multiple archives, computationally intense clustering, and high-resolution visualization.

Grid software extracts the relevant data from the database archives, then transfers the data to an analysis engine at NCSA, where tasks are launched to

- identify the signatures of galaxy clusters from x-ray emissions of hot gas, and
- convolve optical and infrared data to select and identify clusters, based on the presence of radio source morphology variation in the cosmological microwave background at millimeter wavelengths.

Researchers use the large data sets to construct simulated surveys as functions of galaxy cluster mass, density, redshift, and other properties. They use correlation analysis to compare the results of the different techniques, understand selection effects, and generate predictions of observed sample properties based on various theoretical models.

Concurrently, a remote rendering task is scheduled at Argonne to generate 3D visualizations of the predicted and observed structures and correlation results. Third-party data transfer software such as gridFTP[2] moves the visualizations to the astronomer's desk in Boston for visualization. Just as we use an electrical appliance without knowing the location of generators, transformers, or transmission lines, grid software hides all the details of data staging, computation scheduling, and user-identity authentication on remote systems.

BUILDING THE GRID: A DO-IT-YOURSELF TOOLBOX

Today's Internet began as a few research networks, each dedicated to serving a specific research project, discipline, or community. These networks expanded as new sites joined and interconnected with other networks to enable cross-network sharing. We believe the TeraGrid and other national research grids will play a similar role as accretion points for a new, international network of high-performance computing systems, data archives, collaboration systems, and scientific instruments.

Implicit in such a vision is the need for mechanisms that will let research groups, universities, and institutions quickly deploy their own grid infrastructure for connection to the burgeoning international grid. To simplify creation of clusters, grids, and collaboration and visualization systems, the National Computational Science Alliance and NCSA have developed a set of ready-made "in-a-box" software toolkits. The "NSF PACI Program" sidebar provides more information about the partnerships that have developed these toolkits.

NSF PACI Program

The National Science Foundation PACI Program supports two partnerships: the National Computational Science Alliance (Alliance) and the National Partnership for Advanced Computational Infrastructure (NPACI). Both partnerships operate high-performance computing facilities supporting computational science research by the national research community. Each partnership consists of a leading-edge site, the National Center for Supercomputing Applications (NCSA) in Urbana-Champaign for the Alliance, and the San Diego Supercomputer Center (SDSC) in San Diego for NPACI. The two partnerships also include more than 60 geographically distributed academic and national laboratory partners.

The TeraGrid project complements and builds upon the work of the two partnerships. Launched in 1997, the PACI partnerships—led by NCSA and SDSC—seek to build an advanced computational and information infrastructure for open research in science and engineering. NCSA, SDSC, and their partners have built many of the tools needed to operate grid-based distributed computing and information infrastructures. These tools include middleware such as the Globus toolkit, security systems, interfaces for accessing grid resources, and applications capable of running effectively on distributed systems.

The first of these, the cluster toolkit, builds on the intense interest among the scientific, commercial, and computing research communities in the evolution and development of cluster computing using industry-standard building blocks and open source software. The cluster toolkit was developed as part of the Open Cluster Group (www.openclustergroup.org), a collaboration between industry and academia that produced the Open Source Cluster Applications Resources software package. SDSC and the National Partnerships for Computational Infrastructure have developed a complementary toolkit called ROCKS (rocks.npaci.edu).

The cluster toolkits let research groups quickly deploy Linux clusters for scientific computing on Intel processors that interoperate with other research clusters and the TeraGrid software environment. The toolkits include step-by-step guidance on cluster software installation and configuration, along with open source numerical and message passing libraries, tools, and monitoring software. Atop this base, researchers can then use Globus, Condor,[3] and other grid tools, packaged via the NSF Middleware Initiative (www.nsf-middleware.org), to configure the cluster as a fully functional node on the grid.

The rapidly increasing size and complexity of scientific imagery has created a need for high-resolution, scalable displays. For example, the charge-coupled-device detectors on today's telescopes produce 8,000 × 8,000 pixel images, with even larger images to come. Human-scale displays let research groups collaboratively examine and correlate data. A scalable display wall, built using the display toolkit, achieves multimegapixel resolution by tiling the output from a collection of lightweight liquid crystal display or digital light processing pro-

jectors into a single image. A Linux cluster, augmented with graphics accelerator cards, drives the projectors.

Finally, the Access Grid is an ensemble of network, computing, and interaction resources that supports group-to-group human interaction across the grid. Originally developed by Argonne National Laboratory and enhanced by the Alliance, Access grid nodes include large-format multimedia displays, presentation and interactive software environments, and interfaces to grid middleware and remote visualization environments.

Grids connect distributed computing systems, data archives, scientific instruments, and collaboration systems. However, their true power lies in reducing or eliminating the barriers of time and space that make it difficult for distributed research groups to collaborate naturally and intuitively.

Generalizing this theme leads to the emergence of a ubiquitous infosphere, which scientists, researchers, engineers, businesses, and the public will use to find and interact with information, technology, and one another without concern for the grid's technical infrastructure. If truly successful, the global grid will become invisible, empowering and enriching the human experience.

With its distributed data archives, support for multidisciplinary teams, and advanced, high-performance computing and communication resources, the TeraGrid is one of the first steps along this path. Not only will it provide the resources needed to tackle today's most demanding scientific and engineering problems, it will also serve as a catalyst in the development of tomorrow's grid-enabled world. ■

Acknowledgments

The TeraGrid and its expansion as the Extended Terascale Facility are the result of the vision, leadership, and dedication of many people. Fran Berman (SDSC), Rick Stevens (Argonne), Paul Messina (Caltech), Michael Levine (PSC), and Ralph Roskies (PSC) are all co-leaders of the effort, in addition to my contributions.

References

1. I. Foster et al., "Grid Services for Distributed System Integration," *Computer*, June 2002, pp. 37-46.
2. B. Allcock et al., "Data Management and Transfer in High-Performance Computational Grid Environments," *Parallel Computing J.*, May 2002, pp. 749-771.
3. J. Basney and M. Livny, "Deploying a High-Throughput Computing Cluster," *High-Performance Cluster Computing*, R. Buyya, ed., Prentice Hall, 1999.

Daniel A. Reed *is a principal investigator for the National Science Foundation TeraGrid and serves as the project's chief architect. He is also director of the National Center for Supercomputing Applications and the National Computational Science Alliance. He holds the Edward William and Jane Marr Gutgsell Professorship in the Department of Computer Science at the University of Illinois at Urbana-Champaign. Contact him at reed@ncsa.uiuc.edu.*