

DataGrid Middleware: Enabling Big Science on Big Data

Co-PIs: Ian Foster, ANL, Carl Kesselman, USC/ISI, Miron Livny, UWis

Senior Personnel: Bill Allcock (ANL), John Bent (UWis), John Bresnahan (ANL), Ann Chervenak (USC/ISI), Joe Link (ANL), Bob Schwartzkopf (USC/ISI), Doug Thain (UWis), Steve Tuecke (ANL)

One of the most demanding and important challenges that we face as we attempt to construct the distributed computing machinery required to support SciDAC goals is the efficient, high-performance, reliable, secure, and policy-aware management of large-scale data movement. This problem is fundamental to application domains as diverse as experimental physics (high energy physics, nuclear physics, light sources), simulation science (climate, computational chemistry, fusion, astrophysics), and large-scale collaboration. In each case, we have highly distributed user communities that require high-speed access to valuable data, whether for visualization or analysis. The quantities of data involved (terabytes to petabytes), the scale of the demand (hundreds or thousands of users, data-intensive analyses, real-time constraints), and the complexity of the infrastructure to be managed (networks, tertiary storage systems, network caches, computers, visualization systems) make the problem extremely challenging.

There is a significant class of scientific problems that require access to tremendous quantities of data, as well as computation. These are referred to as Data Grid problems. Physicists around the world cooperate in the analysis of petabytes of accelerator data. Climate modelers compare massive climate simulation outputs. Output from multi-million dollar online instruments, such as the Advanced Photon Source or earthquake engineering shake tables, must be visualized in real time so that a scientist can adjust the experiment while it is running.

In order for these applications to be feasible, infrastructure must be in place to support efficient, high performance, reliable, secure, and policy aware management of large-scale data movement. The SciDAC DataGrid Middleware project is providing tools in three primary areas in support of this goal. GridFTP, developed primarily at ANL, provides a secure, robust, high performance transport mechanism that is recognized as the def-facto standard for transport on the Grid. The Globus replica tools, developed primarily at ISI, provide tracking of

replicated data sets and provide efficiency in the selection of data sets to access. The Condor team at the University of Wisconsin is providing storage resource management, particularly space reservation tools. Together, these three components provide the basis for Data Grid applications.

Multiple SciDAC projects employ our Data Grid tools. The Earth Science Grid (ESG) and the Particle Physics Data Grid (PPDG) use GridFTP servers to stage input data and move results to mass storage systems. They also employ our first generation replica catalog to determine the best location from which to store and/or retrieve data. The Laser Interferometer Gravitational Wave Observatory (LIGO) project has moved over 50 TB of data and has an RLS with over 3 million logical files and over 30 million physical filenames. The Grid3 project, part of the Grid Physics Network (GriPhyN), moves over 4TB a day.

The Globus Toolkit is seeing widespread community and commercial support. The second generation Replica Location Service

(RLS) was co-developed with the EU DataGrid project, and the UK e-Science center is contributing Data Access and Integration (DAI) code. IBM and Platform Computing offer commercially supported distributions. Companies porting new versions to their platforms include IBM, HP, Hitachi, NEC, and Fujitsu.

Perhaps the ultimate measure of the success of infrastructure is its ubiquity. TCP, IP, and HTTP are successful, not because they are the fastest or most efficient, but because they are standardized, broadly deployed, and enable interoperability. The Data Grid tools are well on their way to following this path. GridFTP is now a Global Grid Forum standard and there is a replication working group working on standards as well.

To build on our early success, we need to continue to expand our functionality and respond to the needs of our user communities. To this end, we have integrated the Community Authorization Service (CAS) (a product of the SciDAC security project) into GridFTP. We have also released an alpha of our completely rewritten GridFTP server, utilizing the new eXtensible Input/Output (XIO) libraries. This will enable much easier integration of GridFTP support into 3rd party applications, greater ease in adding new features, and provide a more stable, maintainable code base. We will also be releasing striping functionality for the first time this summer.

Efficient utilization of the Grid resources available requires that we improve scheduling and coordination. Reservation of resources is critical to make this practical. The Condor team will continue to improve their storage resource management capabilities. We will also be leveraging our previous work on the General purpose Architecture for Reservation and Allocation

(GARA) to incorporate resource reservations into our transfers.

Performance and scalability will also be addressed. At SC2003, we demonstrated GridFTP running over UDT (a non-TCP, reliable UDP transport protocol), by replacing the underlying XIO TCP driver with a UDT driver. This allows us to overcome TCP's limitations in high bandwidth, high latency networks. The Replica Location System (RLS) includes distributed catalogs and a highly tunable index / aggregation system for the catalogs, which results in excellent performance and scalability.

Finally, to leverage these underlying tools, we will need to develop higher-level "collective" services that coordinate scheduling, reservation of bandwidth and storage, high-performance, robust data transport and update of appropriate replica catalogs and indexes.

In order for this work to progress, it is critical that the underlying network infrastructure continue to keep pace with increases in performance. 10 Gigabit Ethernet projects such as I-Wire, the TeraGrid and StarLight are extremely useful. However, 10 Gigabit Ethernet is now available for the desktop. It will soon be possible to utilize the telecoms OC-768 links to achieve 40 Gigabits per second and research labs are reporting 100 Gigabit per second in the lab. With network speeds doubling every 9 months, this will be difficult to manage.

For more information, please contact:

Ian Foster (foster@mcs.anl.gov)

Carl Kesselman (carl@isi.edu)

Miron Livny (livny@cs.wisc.edu)