

Center for Enabling Distributed Petascale Science

A Department of Energy SciDAC Center for Enabling Technology

Project Partners: Argonne National Laboratory, Fermi National Laboratory, Lawrence Berkeley National Laboratory, University of Southern California, University of Wisconsin

CEDPS — <http://www.cedps.net>

DOE computational and experimental facilities will soon be producing petabytes of data per year, in fields as diverse as astrophysics, biology, chemistry, combustion, fusion, high energy physics, nanoscience, and nuclear physics. Application communities—often large and distributed—must be able to access this data so they can translate it into knowledge. Thus, we must move data to where it is needed—and/or enable analysis to occur near the data. Each task is challenging in a petascale environment, because of the need to coordinate numerous shared resources, including CPUs, storage, and networks.

The **Center for Enabling Distributed Petascale Science** will address these challenges by designing, developing, deploying, and evaluating new tools for **data placement, scalable services, and troubleshooting.**

Recruiting Partners

If you would like to work with CEDPS, contact Jennifer Schopf, jms@mcs.anl.gov

(1) Data Placement

Current high-performance data transfer mechanisms are not sufficient to enable scalable, communitywide analysis, for example:

- A team running the CCSM climate simulation code wants to publish its output data for community access. The team must both transfer the output data to an HPSS archive and register each file in a metadata catalog.
- Scientists run a combustion simulation at NERSC producing 100 TB of data. They then want to explore that data using visualization tools and replicate the data at five sites.
- Data produced by the CMS experiment at the LHC (at 400 MB/s) must be delivered to a Tier-1 site where it is further processed and distributed among 25 Tier-2 sites.

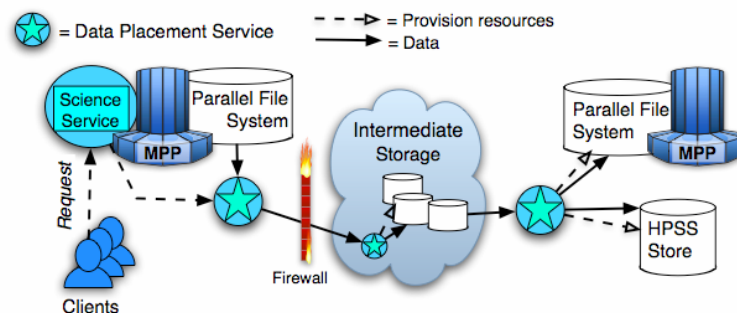
Such scenarios frequently involve several tasks:

- Data movement over high-speed long-haul networks from a diversity of sources and sinks, including parallel file systems, running programs, and hierarchical storage.
- Coordinated data movement across sources, destinations, and intermediate locations and among multiple users and applications.

- Use of techniques such as storage reservation, data replication, online monitoring, and operation retry.
- Need for predictable and coordinated scheduling in spite of varying load and competing use of space and bandwidth.

These elements require the coordinated orchestration of data across the entire set of community resources and therefore go well beyond our current data transfer and storage resource management capabilities.

CEDPS will develop tools and techniques for reliable, high-performance, secure, and policy-driven placement of data. To this end, we will construct a Managed Object Placement Service (MOPS)—a significant enhancement to today's GridFTP and other coordinated tools—that allows for management of the space, bandwidth, connections, and other resources needed to transfer data to and/or from a storage system, and policy-driven data placement tools that build on MOPS mechanisms.



(2) Scalable Science Services

Driven by the need to deliver analysis to expanding nonexpert user communities and a desire to incorporate analysis components as black boxes into complex analysis chains, we see increasing adoption of an approach in which programs are packaged as scalable science services that process (potentially many) requests to access and analyze data subsets via well-defined service interfaces.

Scenarios include the following:

- The fusion code TRANSP, deployed as a service by the fusion community, is accessed by tens to hundreds of remote users. Supporting such levels of access requires tools to manage the allocation of resources for execution, so that all users get acceptable service.
- The PUMA biology data resource must perform millions of BLAST, Blocks, and other computations to integrate new data into their database, which is memory- and CPU-intensive and requires many hundreds of CPUs for days.
- The Earth System Grid, which already supports downloads of terabytes of data per day by a community of thousands, plans to expand to include computational functions such as reductions. A single user request may

require large amounts of computation and data access—and dozens of requests can be active at one time.

These scenarios involve the following issues:

- Need to integrate existing code into a services framework, to enable sharing of the code across community members, composition of analysis capability into end-to-end analysis chains, and the isolation of clients from details concerning the location and implementation of analysis functions.
- Ability to dynamically and reliably configure and deploy new analysis services and to dynamically vary the computational resources used by analysis services, in response to changing community load.
- Ability to manage how services are used by communities on a user-by-user and request-by-request basis and to monitor for performance degradation at that level.

To address these concerns, CEDPS will develop scalable science service tools, including service implementation, end-user client and management interfaces, and mechanisms for the dynamic acquisition of computing, storage, and other resources in response to changing load.

(3) Troubleshooting

Performance and reliability requirements for petascale science are predicated on our ability to monitor, collect, and respond to information about the individual and collective behavior of data and services under dynamic service environments and load. Experience with current DOE distributed system deployments has shown that understanding behavior is a fundamental requirement, not just a desirable enhancement.

For example, according to the Grid2003 “Lessons Learned,” approximately 30% of all Grid jobs fail for unknown reasons and, at present, diagnostic tools are lacking. Analysis of periodic benchmark runs on the GrADS testbed showed that 6% of application failures were due to GridFTP misconfiguration and 20% to NFS problems at one site. Middleware may also mask performance

faults, when applications produce correct results but experience degradation in performance.

Application scientists and middleware developers need to collect and analyze detailed background monitoring and event-driven data to troubleshoot their applications. They must be able to detect errors at run time, analyze baseline performance shifts, and receive warnings about failures. Thus, the ability to instrument, collect, access, and analyze system and service activities for troubleshooting is the third cross cutting issue of CEDPS. While provisioning and deployment must be integrated into interfaces and implementations of our data management and scalable science service frameworks, troubleshooting is unique in that it requires additional services and tools that must exist outside the scope of the core service definition.

For more information: www.cedps.net