

## **A Networking Strategy for Peta-Scale Science within the DOE Office of Science**

Over the next 5-10 years, we expect the computation and storage capabilities within the DOE Office of Science to move to the peta scale, with petaflops rates of computation using and producing multi-petabytes of data. For the most demanding applications of importance to the Office of Science, the pursuit of science at this peta scale will still be governed by limitations of hardware and software, not by the requirements of the scientific applications. Because the highest-end resources will continue to be a proverbial bottleneck, efficient use of these resources will remain critical, and this efficiency requirement will drive the high-end needs for networking.

Another requirement driving networking needs is that of secure, convenient, distributed, collaborative access to the high-end resources from a diverse group of scientists using a range of interface devices, from handheld devices to desktops to synchronized distributed display walls. The efficient use of high-end resources is only profitable to the extent that scientists are able to translate that use into scientific discovery.

These various resource requirements, in the context of trends in industry-standard information technology (IT), suggest the following networking strategy for peta-scale science.

- The high-end computation and storage resources of the DOE Office of Science are tightly coupled with ultra-high-bandwidth networking, providing an integrated core infrastructure.
- Users access this core through user-interface (UI) services based on prevailing industry standards, using commodity clients, heavily leveraging the public and commercial IT hardware and software infrastructure.

Targeting a tightly coupled, integrated core maximizes the efficiency of the high-end resources. Targeting commodity clients and standard IT maximizes the flexibility and accessibility of those resources. Further targeting UI serving, as opposed to data or compute serving, decouples the computation and storage scaling requirements of the high-end applications from the scaling requirements of multiple scientists and distributed collaboration.

The following sections describe the requirements of the two networking domains, the peta-scale core and the UI-service domain, in greater detail. This document does not address the extent to which current DOE projects, such as the DOE Science Grid, might meet these requirements. We anticipate discussion of this topic at the High Performance Network Planning Workshop.

### **Peta-Scale Core**

The peta-scale core infrastructure for the DOE Office of Science will likely include a moderate number of large, geographically distributed resources, such as high-performance-computing (HPC) centers, data repositories, data-analysis centers, and experiment facilities. A unifying feature among these resources is data intensity. Large

experiments may produce petabytes of data, petaflops computations may consume and produce petabytes, repositories must store petabytes or more, and these petabytes must be analyzed for the ultimate goal of scientific discovery.

We expect peta-scale resources to be designed with petabytes of high-speed local IO as a working cache for data, but important output must eventually be stored on media that are fault tolerant and externally accessible. Such media currently take the form of site-local archives, often implemented with HPSS. We currently have unmet end-to-end bandwidth requirements for moving data between high-end sites, such as between the teraflops computers in the ORNL CCS and the PCMDI climate archive at NERSC.

To better use future peta-scale resources, and indeed the resources currently available in the DOE Office of Science, we need much higher end-to-end bandwidth between core high-end sites. Ultra-high-bandwidth networking would facilitate critical data movement between the working caches of distributed peta-scale sites:

- from experiments to data-analysis and visualization servers,
- from archives of measurement data to supercomputers running models using those data,
- from supercomputers to archives of computational results,
- from archives of raw results to data-analysis systems enabling scientific discovery, and
- between distributed archives, for fault tolerance and disaster recovery.

The network within the core will need to be able to move petabytes in reasonable time limits, certainly better than the effective bandwidth of physical tapes sent by overnight courier. With the estimate of this requirement as a petabyte in 24 hours, the backbone network will need a capacity of roughly 160 Gb/s (gigabits per second). Saturating a tens-of-gigabits network connection is well beyond the capability of current data-transfer protocols, such as HTTP, SFTP, or Grid FTP. A high-end system today can saturate a 1Gb/s connection with about twenty streams. If systems scale over the next two years to 10 Gb/s interfaces, the scaling behavior of TCP implies that 200 streams may be necessary to saturate that 10 Gb/s. To saturate a 160 Gb/s pipe would then require 3200 streams, which is impractically high.

To provide the necessary core bandwidths, DOE will need to implement improved protocols, such as XCP, Tsunami, *etc.*, and possibly new transport technologies. It is likely that new protocols will not be TCP-friendly and will require a parallel backbone. Such a backbone might serve as a redundant circuit for standard IP traffic, however, in the event of an outage in neighboring networks.

With this strategy, the networking resources connecting the core sites will see a qualitatively different usage pattern than traditional IP networks. Instead of the highly shared, overlapped, and variable volume typical of interactive network use, core networking will need to support large-volume bulk transfers with pre-scheduled, dedicated or quality-guaranteed access to resources. An analogous comparison is interactive use of a PC versus long-running batch computations on a supercomputer.

Tightly coupling peta-scale sites within the DOE Office of Science with such dedicated networking hardware enables improved integration of those resources from the perspective of usage and administration. To effectively use petabytes of data at multiple sites, those data must be accessible across sites, both physically through networking hardware and logically through appropriate shared interfaces.

One useful interface is that of a file system. A logically shared, globally accessible archival file system among the core sites could dramatically simplify data management. The file-system model is not ideal, however. The finest level of data granularity is the file, and the only mechanism for searching and organizing data is by file path. The integration of more general mechanisms for data search and discovery, such as relational-database interfaces, may not only prove useful but essential for managing and utilizing huge data volumes.

We suggest a single logical archive with data and metadata distributed among core sites. Users of other core resources would schedule retrieval of data to the working cache of each resource, perform work, and schedule writing of permanent data back into the archive. Current tape-based archives can provide high bandwidth but suffer from relatively high latencies. The performance profile of ultra-high-bandwidth networking fits this regime well; the latency of tape-based operations hides network latencies. The features desirable in a global DOE archive include the following:

- single, secure sign-on,
- uniform global namespace,
- flexible, location-independent data description and discovery,
- sub-file-level access to data objects,
- support for version control and synchronization,
- automated redundancy for fault tolerance and disaster recovery, and
- automated, performance-optimized selection among redundant copies.

Higher levels of software integration among the DOE sites are required to provide such a global interface to data, and this integration could provide additional benefits. Current HPC sites in the DOE each have their own unique solutions for many infrastructure needs, including user authentication and authorization, batch-system configuration, account administration, allocations and accounting, and file systems. Better integration of the core sites could reduce the current redundancy of effort, providing a more-consistent user experience and better amortizing the costs of the infrastructure maintenance and development. The mission-oriented nature of the DOE and its resources makes such integration feasible. With careful and pragmatic planning, increased integration could improve efficiency and productivity without stifling the creativity and innovation of individual sites.

The strategy of a tightly coupled core providing standard interfaces to commodity clients is designed to improve efficiency, but it is not intended to isolate the DOE core from other peta-scale resources. Collaboration with other agencies will remain important, and the benefits of ultra-high bandwidth among core resources could extend to peta-scale

sites of other agencies, such as the NSF and NASA. Higher software integration would also prove useful but will be constrained by the different requirements of the various agencies in terms of security, administration, and scientific mission.

### **User Interface Services**

Whereas high-end applications may be limited by whatever resources are available, whether at the tera, peta, or exa scales or beyond, user interfaces have more-achievable upper bounds on resource requirements. A desktop computer, for example, is limited to the data rate of keyboard and mouse for input and the screen size times frame rate for output. Unlike for peta-scale computation and storage, scientific discovery has similar requirements for UI technology and infrastructure as the commercial world that drives UI technology and infrastructure.

In the strategy suggested here, the core DOE high-end infrastructure will provide access through standard UI services, leveraging the same software and hardware infrastructure used for public, commercial, and educational purposes. Today such UI services include HTTP, SSH, X-Windows, Java, and streaming video. In 5-10 years, commercial drivers such as E-commerce, videoconferencing, video on demand, and virtual reality may bring the state of practice for secure, distributed UI services up to the current state of the art for high-end visualization. The most efficient and effective course for scientific discovery at the peta scale may be to simply ride this technology curve, targeting industry-standard UI and remote-authentication services. This strategy maximizes the accessibility of the peta-scale core for scientists, who will already have familiarity with these UI services.

### **Conclusions**

The basic strategy proposed here is to enable scientific discovery at the highest levels of computation and data volume by scaling and integrating core DOE systems as much as possible while providing industry-standard user interfaces to those core systems. This strategy is intended to maximize the capability of the core while also maximizing its accessibility to scientists.

The implication of this strategy for networking hardware is that investment should be focused on connecting the core systems for large end-to-end volume and predictable quality of service. As with hardware, investments in networking software should target the core systems, both at low levels, such as enabling peta-scale bulk transfers between systems, and at high levels, such as enabling global data description, discovery, and management.

The implication for user-level software at the peta scale, such as tools and applications, is that the interaction with the “outside world” should be through industry-standard UI services. For example, applications should not rely on the transfer of data files to the user workstation for analysis. At peta scales and beyond, requiring much more out of remote devices than human input and video output is not feasible. Also, given the plethora of possible remote devices, from handheld devices to display walls, and the fairly universal

nature of user-interface requirements, applications and tools should not require unique clients or protocols outside of the core resources.