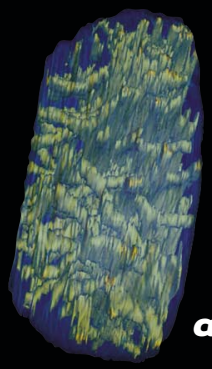
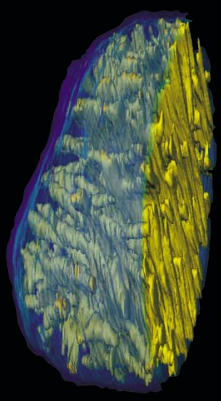
Data-Intensive Computing and Digital Libraries

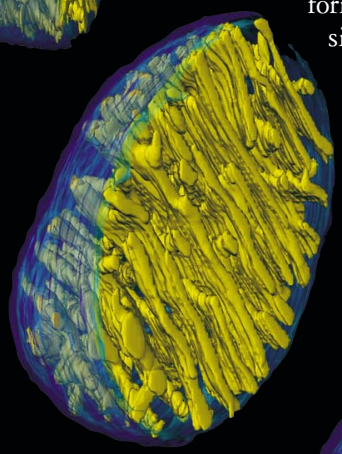
Reagan Moore, Thomas A. Prince, and Mark Ellisman

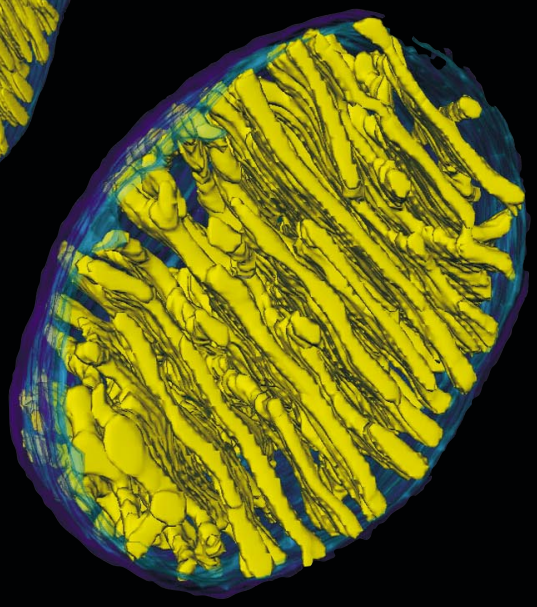


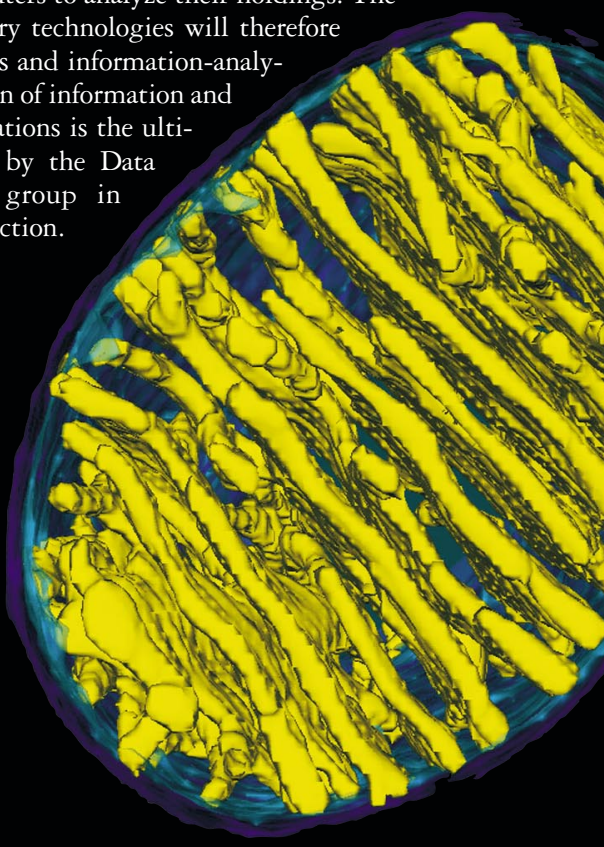
How to automate management of the flood of scientific data being collected in astronomical and neuroscience projects.



COMPUTATIONAL SCIENCE IS EXPANDING TO INCLUDE NOT ONLY ANALYSES BASED ON SIMULATIONS but those requiring manipulation of large data collections. The need for data-intensive computing is being driven by the massive amounts of data now available in various scientific disciplines. As a consequence, supercomputing systems have to incorporate data and information-handling technologies to manage these very large data sets. And as the amount of data in storage environments, such as digital libraries, increases, it will be necessary to use supercomputers to analyze their holdings. The coevolution of supercomputer and digital library technologies will therefore form the basis for future scientific applications and information-analysis activities. For supercomputers, the creation of information and its organization and use in future computations is the ultimate goal. Infrastructure development by the Data Intensive Computing Environments group in NPACI represents a first step in this direction.



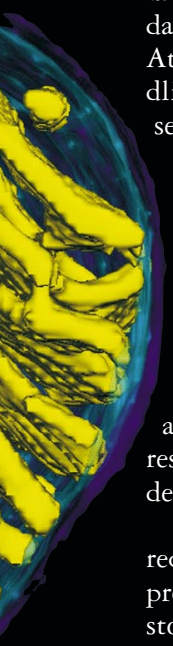




Under the earlier National Science Foundation Supercomputer Center's program, which ended in 1997, the majority of data stored by the centers was generated through numerical simulations. The data collections now being archived will be dominated by, for example, remote observational data obtained from satellites, high-resolution images from magnetic resonance imaging (MRI) systems, and collections of objects from digitized astronomical sky surveys. In effect, the acquisition of data is being automated. For scientific disciplines to survive under the coming data loads, scientific applications have to be able to automatically manage, query, and analyze large data sets.

Applications should be able to query and retrieve new information as part of their normal execution and publish results into discipline-specific data collections as part of their normal output. The result will be a natural extension of the scientific process of analyzing observations to derive information to a process of automated ingestion of prior knowledge as part of the analysis. This concept—we call it "information-based computing"—can lead to rapid growth in the rate of knowledge creation [6, 8].

NPACI is bringing together collaborative research projects in which data collections are distributed across multiple sites. Each site today typically relies on local Unix file systems for storing data, databases for organizing metadata about the data sets, and tape archives for backing up the data. At each site, researchers have three basic data handling needs: identify relevant data sets, move data sets to a computational platform on which they can perform their analyses, and apply analysis algorithms to extract new knowledge.



Supporting these needs is onerous when the data sets consist of hundreds of simulation output files or remote observations. Having to identify data sets through a unique Unix pathname becomes especially difficult when the data being accessed was created by another researcher at another site. Combining data sets from multiple researchers requires a common set of terms for describing data set attributes.

Retrieving data stored at remote sites also requires researchers to understand multiple access protocols and manually initiate data transfer to local storage systems. The amount of data is daunting, and the time needed to move it is increasing to hours of wall-clock time. Increased transmission time is a consequence of the divergence of CPU execution growth rates from network bandwidth growth rates. We are learning to generate data more rapidly than we can move or manage it.

These requirements are illustrated by their application in two disciplines: the Digital Sky Project in astronomy and the sharing of brain images in neuroscience.

Digital Sky project. Large-area digital sky surveys are a recent and exciting development in astronomical research. The combination of the NPACI terabyte/teraflops computational resources with recent large-area surveys in optical, infrared, and radio wavelengths will provide unprecedented computational support for astronomical research. Each survey records information about all objects observable at specific wavelengths of light to the limiting magnitude that can be distinguished from noise. The ability to perform statistical analyses on the entire data collection will help researchers identify all astronomical objects in the observable sky with particular specified properties, dramatically changing the types of investigations that can be conducted.

This project is investigating the technologies needed to integrate access to the multiple sky surveys. Astronomers would have access to a multiwavelength "digital sky" covering a significant fraction of the real sky. A production version of the digital sky and the tools for its exploration could revolutionize multiwavelength astronomical studies, owing to both the sheer increase in the data available and faster and more sophisticated methods for its analysis.

The sky surveys are distributed across multiple sites, including the Digital Palomar Observatory Sky Survey (DPOSS), the 2-Micron All Sky Survey (2-MASS), and the National Radio Astronomy Observatory Very Large Array (NRAO VLA) Sky Survey (NVSS). The project is defining the data manipulation infrastructure needed to integrate these surveys online for the astronomical community, providing simultaneous access to the catalogs and image data, along with enough computing capability to allow detailed correlated studies across the entire data set. This infrastructure will provide

3D reconstruction of dendritic mitochondrion. (Courtesy, Guy Perkins and Terrence Frey, University of California, San Diego, and Mark Ellisman, National Center for Microscopy and Imaging Research at San Diego.)

We are learning to generate data more rapidly than we can move it.

mechanisms for uniform access to the multiple data resources that will be used to store, maintain, and organize the collections.

The number of data objects that can be addressed individually is quite large. For example, data from the DPOSS, 2-MASS, and NVSS surveys is expected to yield on the order of a billion sources, and the image data will comprise several tens of terabytes. Astronomers want to be able to launch sophisticated queries to the integrated catalog that describes the optical, radio, and infrared sources and then perform detailed analysis of the images for morphological and statistical studies of discrete sources and extended structures. Integration of the surveys will require a probabilistic identification mechanism to find the same object within multiple surveys. NPACI facilities will support pattern recognition, automatic searching and cataloging, and computer-assisted data analysis and may require systems to move substantial amounts of data to the computing resources or to maintain replications of the data sets on storage systems connected via high-speed links to the compute platforms.

Significant results have already been obtained from these surveys. For instance, initial small-scale explorations of DPOSS have yielded discoveries of a large number of high-redshift quasars. Having the full data set online along with radio and infrared surveys would increase the scientific potential of this data by orders of magnitude, permitting statistical analyses impossible today. Indeed, the full science potential of the surveys will be realized only with the multiterabytes/teraflops capabilities provided by NPACI facilities.

Mapping the brain. In recent years, an explosion of structural and functional information about the primate (such as human and macaque monkey) brain has become available. In addition, extensive brain mapping data (such as for laboratory rats and crickets) is being accumulated on model systems. Neuroscientists have to deal with staggering amounts of volumetric data (gigabytes in many individual experiments) and diverse multidimensional databases. To maximize the benefits from the accumulating brain-related data and to generalize new domain-specific data-exploration strategies, these data sets should be linked and organized within discipline-wide neuroscience data collections. The federation of multiple-brain-image databases requires an integration project similar to that needed for the Digital Sky project.

When building brain databases or comparing brains, neuroscientists commonly manipulate a small number of very high-resolution images for intra- and intersubject registration, or the mapping of brain structures onto a known coordinate system, so common features can be identified via elastic or fluid deformation (called "warping"). They also commonly conduct serial processing of many single-brain images at lower resolution, such as that required for transforming groups of MRI volumes into a standard space. These operations are accomplished using either a single extremely fast processor or a distributed parallel computing environment. While the deformation of a small number of brains results in new ways to observe anatomical properties, it is only when these techniques are applied to a large number of subjects that statistically significant claims can be made and hypotheses proposed that advance the field.

NPACI researchers have developed methods to address issues of individual variability in brain organization and function, as well as methods allowing comparisons across species (such as from monkey to human and human to monkey). It has become routine to transform one brain to match the shape of another using high-dimensional deformation algorithms. The need to compute, store, and track all these transformations and their associated experimental data deepens the challenges in computational complexity and database management.

These neuroscience brain-image data collections are housed at Washington University in St. Louis, the University of California at Los Angeles, Montana State University in Bozeman, and the University of California, San Diego. The future integrated brain-image collection is expected to aggregate more than a terabyte of data. Access between the sites will be

over the NSF-sponsored very high-speed Backbone Network Service.

Gigabytes of data from individual experiments are being saved locally at each researcher's site. Each brain in an experiment may be described and stored with a variety of data types and formats (such as volumetric structural data, reconstructions of the cortical surface, functional imaging data, and connectional data). These unique attributes need to be accessible for remote identification of a relevant data set. In addition, modeling tools are being developed for analyzing the structure of the brain. The tools being developed for the rat and cricket nervous systems will have to be integrated into the primate database to complement its toolset. This integration effort will in turn require the ability to apply analyses to remote data sets, either by moving the data to the compute platform or by moving the analysis to the data's location.

The data-handling requirements for integrating multiple data collections requires the automation of data set naming and data set access, as well as the ability to publish data, form data collections, and apply analyses to distributed data. The most difficult requirement is to automate these abilities so they can be accessed from a supercomputer application. An information infrastructure will also need to provide application programming interfaces (APIs) for information discovery, data access, and data publication.

Two computer science communities are investigating development of the infrastructure needed for distributed resources: one in metacomputing systems supporting distributed execution of applications and one in digital library systems [1, 5, 9, 10] supporting execution of services against data sets. NPACI is pursuing the integration of these approaches.

Metacomputing systems. Metacomputing systems focus on the ability to distribute application execution across multiple compute platforms (see Figure 1 and A. Grimshaw et al.'s article in this section). This distribution requires an infrastructure that supports scheduling of the compute resources, accounting systems to track CPU usage, and security systems for authenticating users across multiple administrative domains. The systems typically trap references to the local Unix file system and redirect I/O requests to file systems on remote platforms. High performance is achieved through caching data on the local system. Long-term storage is provided by hierarchical storage management systems. Data sets remain local to the researcher; access for other researchers is granted through access control lists.

Digital libraries. Digital library technology focuses on the management of information (see Figure 2). Access is through presentation interfaces for interactive queries issued by a researcher. Information is identified by attributes acquired and organized through a publication infrastructure for ingesting data. Attributes are maintained in databases supporting the processing of queries for data set identification. The data sets are usually stored on the local Unix file system, in an object-relational database or in an object-oriented database. Digital libraries coordinate services that can be used to process the information. The services are executed either locally through Common Gateway Interface

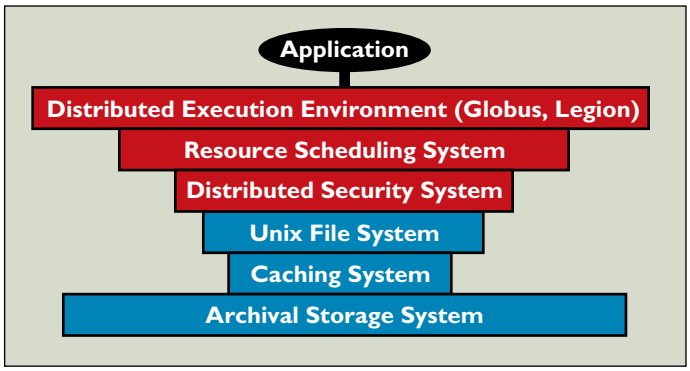


Figure 1. What supercomputer metacomputing environments can do

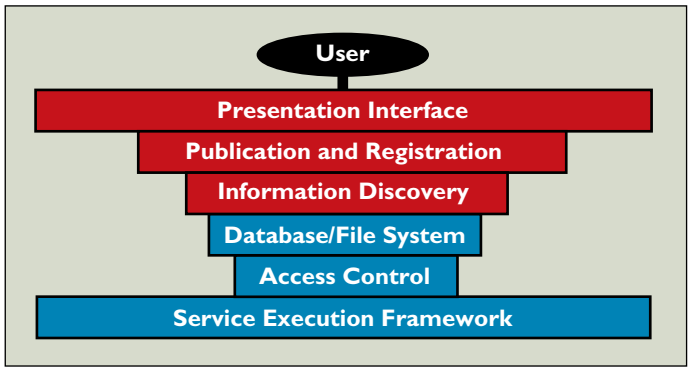


Figure 2. What digital library environments can do

scripts or remotely through Java applets. Access control is assumed to be needed only for the local data sets.

Each system includes capabilities needed to meet the NPACI data-handling requirements. NPACI researchers envision an infrastructure that incorporates:

- Information discovery APIs for attribute-based identification of data;
- Publication APIs for incorporating data sets within collections;

- Data access APIs for accessing distributed data sets;
- A distributed execution environment with support for resource scheduling and security across multiple administrative domains;
- Management of services that can be directly applied to data sets;
- Distributed collections management;
- Distributed data handling; and
- Persistent storage of published data.

Each of these is provided as a service accessible directly from an application (see Figure 3). Metacomputing systems are used to encapsulate distributed resources as an execution service. Digital library technology encapsulates information management and publication capabilities as services. To access distributed data resources, the San Diego Supercomputer Center has developed a storage resource broker (SRB) [2, 3] that provides attribute-based access to remote data sets.

The essential component of the infrastructure is the API provided for each service that is required for automating access to information. Organizing data sets according to attributes maintained in a metadata catalog makes it possible to perform information discovery within distributed environments. Each service can be thought of as an interoperability mechanism. For example, the security service must provide for interoperation between Kerberos, the secure socket layer, and the secure shell environments. The metadata catalog service has to support interoperation among multiple, heterogeneous metadata catalogs that may be distributed. The SRB supports access to a variety of data systems, including archives, databases, file systems, Web sites, and FTP sites; it also supports replicated data sets and will include support for copies maintained in data caches on the network.

The SRB is an example of infrastructure designed to support interoperability and is implemented as servers providing a front end for each data storage resource (see Figure 4). The goal is to allow applications to use a standard data access paradigm to access data stored in heterogeneous storage resources. The access paradigm may require additional information, including data set attributes. By querying the attributes available in a metadata catalog [4, 11], applications can identify relevant data sets. The catalog stores system-level metadata, including audit trails, access-control lists, and location information about the storage system on which the desired data set resides and the definition of the protocol to be used to access the storage system.

The SRB then issues a data request in the appropriate protocol required by the storage system and returns the resulting data to the client application. The interface to each type of storage device is through a set of device-specific driver routines. The net effect is that an application can identify an appropriate data set and then read or write from that data regardless of its location on the network and the type of storage system in which it is stored.

The SRB system supports a variety of computing platforms, including Cray, Digital Equipment, IBM, Silicon Graphics, and Sun Microsystems. It gives access to archival storage systems (such as the IBM

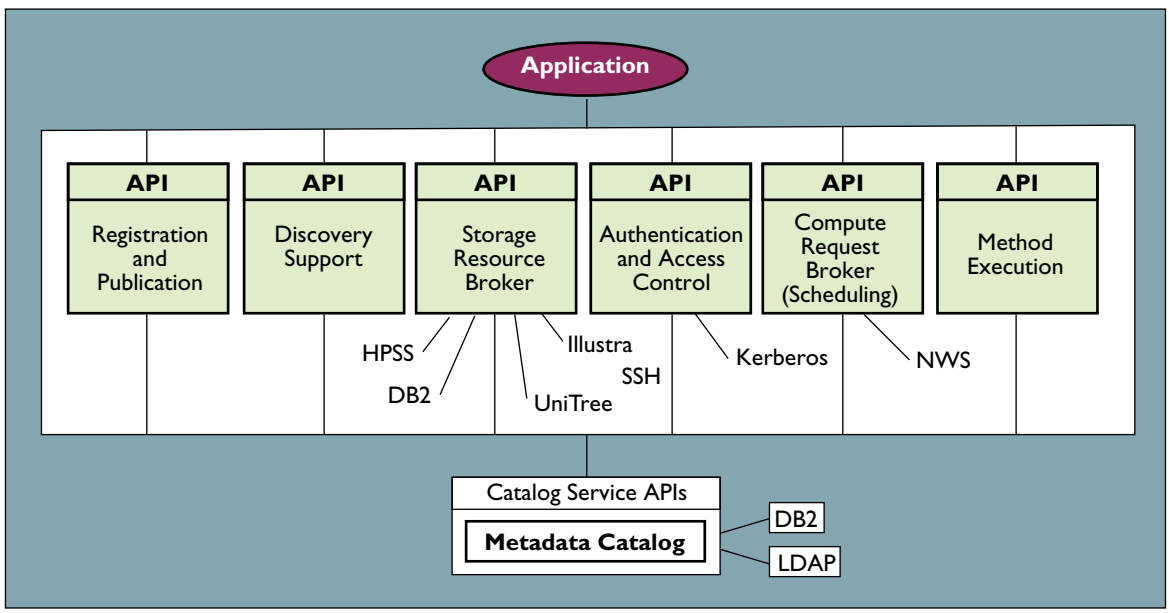


Figure 3. Integration of metacomputing and digital library technology

High-Performance Storage System and UniTree archival storage system), object-relational databases (such as DB2v.2, Illustra, and Oracle), object-oriented databases (such as Objectivity), Unix file systems, and FTP servers. The SRB is being used as the data-handling system for mirrored copies of the University of California, Berkeley Electronic Library [11], the University of California, Santa Barbara Alexandria Digital Library [3], and the University of Michigan Digital Library [5]. The combined system provides digital library services on data sets regardless of whether they are in online or near-line storage, enabling the digital library to manage arbitrarily large amounts of data.

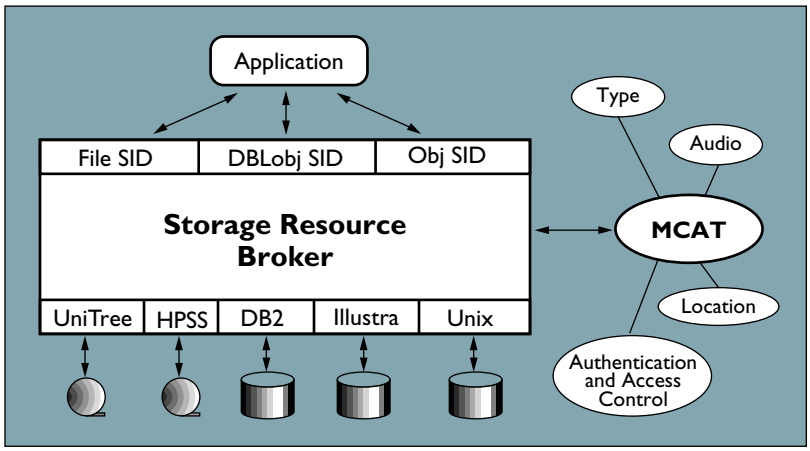


Figure 4. Infrastructure supporting distributed access to data stored in legacy data resources (SID: storage interface driver; MCAT: meta information catalog)

This digital library software infrastructure is being used to establish data collections that support specific disciplines. However, domain expertise is still needed by NPACI computational scientists to define the ontology for organizing information within a discipline, define the attributes to be used as metadata to describe data sets, define the semantic meaning to be associated with attributes, and construct the schema used to organize the attributes. Gaining domain expertise requires working closely with each discipline to understand how it conceptualizes its data space. The goal is to develop a generic digital library infrastructure as the basis for supporting digital data collections regardless of specific discipline.

Integrating the digital library infrastructure with services that are possibly executed in a distributed environment gives rise to the issue of distributed scheduling and the optimal use of distributed resources. For example, should an application-specific algorithm be implemented as a service executed within a digital library, close to where the data resides? Or should the data be moved for analysis to another platform within the metacomputing environment? Each scenario uses different proportions of the resources associated with the data-handling platforms, the network, and the computing platform. The total cost (in time units) of each resource [7] can be calculated based on the following steps:

- Data accessed from the storage device;
- Data analysis at the data-handling platform, characterized by the number of analysis operations per byte of data, thus reducing the size of the data set;
- Data transmission of the reduced data;
- Network transmission time; and
- Reception of the data at the compute platform.

The same cost analysis can be done with all of the data being transmitted to the remote platform for analysis. For example, the resources needed to transmit the reduced data vs. to transmit all the data can be compared by adding the time for each step. It is then possible to determine when the analysis should be executed at the data-handling platform, based on the complexity of the analysis algorithm (the number of operations per byte of data analyzed).

For services exceeding a complexity threshold, the data should always be moved to the remote compute platform for analysis. For simpler computations, even when the available bandwidth is arbitrarily large, the data analysis should always be done at the digital library. Integration of the digital library and the metacomputing system is a way to handle all levels of complexity of analysis.

Scientific Value

Data-intensive computing will aid in the advancement of science by providing the means to manage and manipulate massive data sets. Through digital library technology, it will become possible to integrate scientific data into data collections that can be used as information resources. Through development of information discovery APIs, it will become possible to reference data sets based on their attributes, eliminating the need to know low-level details for identifying each data set. Through development of data-handling systems, applications will be able to directly access data stored within any data collection.

And finally, integrating metacomputing technologies with digital libraries will enable the resulting information-based computing system to support arbitrarily complex analyses of data, providing the core infrastructure for data-intensive computing.

The result can be a natural way to extend the scientific process of analyzing observations to derive information to a process of automated ingestion of prior knowledge as part of the analysis process. C

This work was supported in part by a DARPA grant, Orders No. D007 and D309 issued by ESC/ENS under contracts #F19628-95-C-0194 and #F19628-96-C-0020, and by NSF grant ASC 96-19020.

References

1. Andresen, D., Carver, L., Dolin, R., Fischer, C., Frew, J., Goodchild, M., Ibarra, O., Kothuri, R., Larsgaard, M., Manjunath, B., Nebert, D., Simpson, J., Smith, T., Yang, T., and Zheng, Q. The WWW prototype of the Alexandria Digital Library. In *Proceedings of the ISDL95 International Symposium on Digital Libraries* (Japan, Aug. 22–25, 1995).
2. Baru, C., and Rajasekar, A. A hierarchical access control scheme for digital libraries. In *Proceedings of the 3rd ACM Conference on Digital Libraries DL98* (Pittsburgh, Pa., June 23–25 1998).
3. Baru, C., Wan, M., Rajasekar, A., Schroeder, W., Marciano, R., Moore, R., and Frost, R. Storage Resource Broker, Tech. Rep. (see www.npaci.edu/DICE/).
4. Baru, C. Frost, R., Marciano, R., Moore, R., Rajasekar, A., and Wan, M. Metadata to support information-based computing environments. In *Proceedings of the 2nd IEEE International Conference on MetaData97* (Greenbelt, Md., Sept. 16–17). IEEE Computer Society Press, Piscataway, N.J., 1997.
5. Durfee, E., Kiskis, D., and Birmingham, W. The agent architecture of the University of Michigan Digital Library. In *IEE/British Computer Society Proceedings on Software Engineering 144,* 1 (Special Issue on Intelligent Agents) (Feb. 1997).
6. Foster, I., and Kesselman, C., Eds. *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, San Francisco, 1998.
7. Moore, R. Distributed database performance. Res. Rep. GA-A20776, San Diego Supercomputer Center, 1991. In *Proceedings of the 3rd Gigabit Testbed Workshop* (Jan. 13–15). Corporation for National Research Initiatives, Reston, Va., 1992.
8. Moore, R., Baru, C., Bourne, P., Ellisman, M., Karin, S., Rajasekar, A., and Young, S. Information-based computing networks. In *Proceedings of the Workshop on Research Directions for the Next Generation Internet* (Washington, D.C., 1997).
9. Paepcke, A., Cousins, S., Garcia-Molina, H., Hassan, S., Ketchpel, S., Röscheisen, M., and Winograd, T. Toward interoperability in digital libraries: Overview and selected highlights of the Stanford Digital Library Project. *IEEE Comput.* (May 1996).
10. Phelps, T., and Wilensky, R. Toward active, extensible, networked documents: Multivalent architecture and applications. In *Proceedings of the 1st ACM International Conference on Digital Libraries* (Bethesda, Md., March 20–23). ACM Press, New York, 1996, pp. 100–108.
11. Rajasekar A., Moore, R., Baru, C., Frost, R., Marciano, R., and Wan, M. MDAS: A massive data analysis system. In *Proceedings of Interface97 Symposium* (Houston, May 14–17). Interface Foundation of North America, Fairfax Station, Va., 1997.

Reagan Moore (moore@sdsc.edu) is associate director of the San Diego Supercomputer Center's Enabling Technologies group. **Thomas A. Prince** (prince@caltech.edu) is a professor of physics at the California Institute of Technology in Pasadena. **Mark Ellisman** (mellisma@ucsd.edu) is a professor of neuroscience at the University of California, San Diego, and leader of the NPACI neuroscience thrust area.