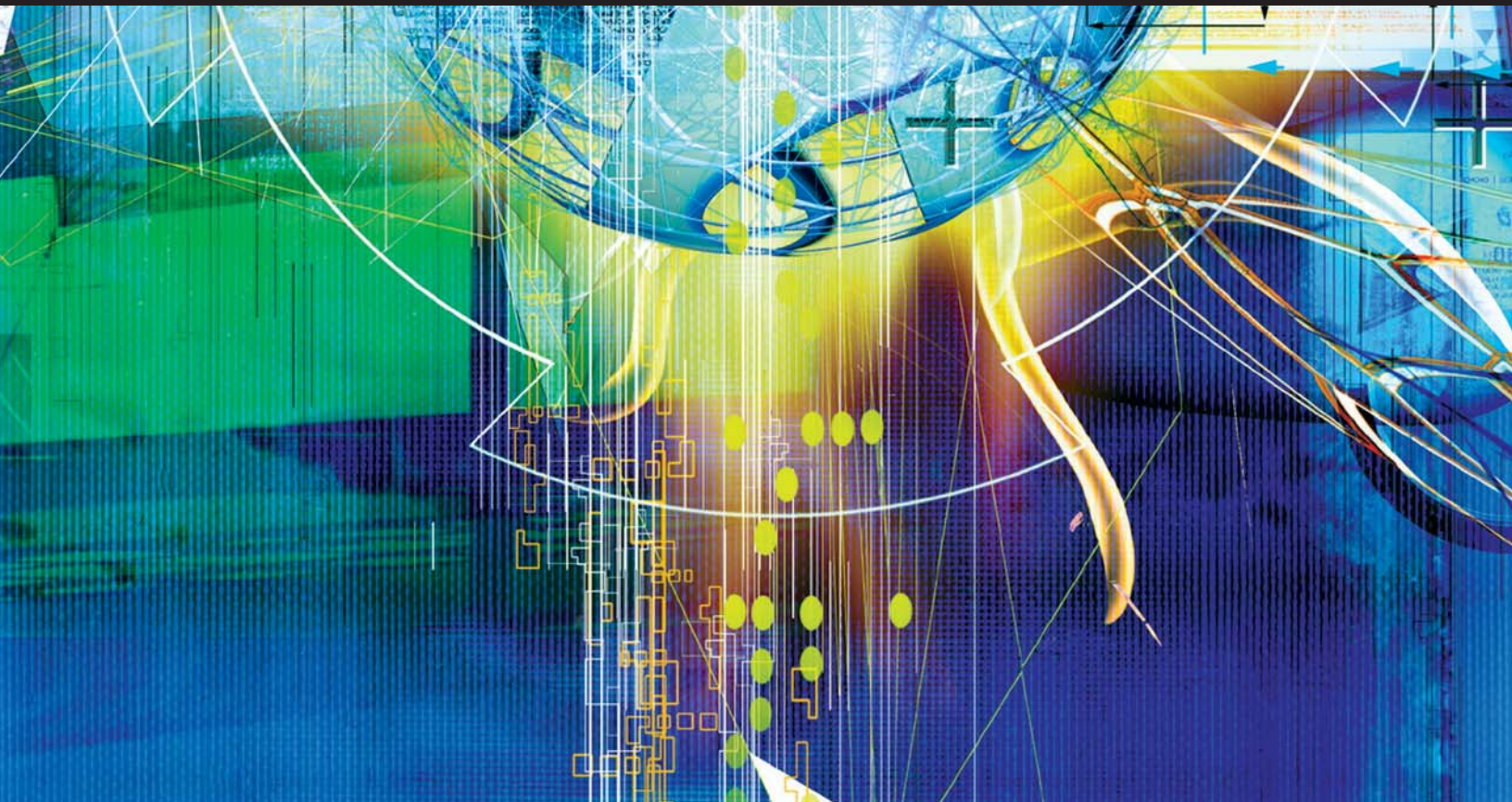




**Trends in
Information Technology Infrastructure
in the Ocean Sciences**



OITI Working Group Members

Tom Powell, University of California, Berkeley (chair)
Mark Abbott, Oregon State University
Scott Doney, Woods Hole Oceanographic Institution
Marjy Friedrichs, Old Dominion University
Dale Haidvogel, Rutgers University
Carl Kesselmann, Information Sciences Institute
Larry Mayer, University of New Hampshire
Reagan Moore, San Diego Supercomputer Center
Ray Najjar, Pennsylvania State University
John Orcutt, Scripps Institution of Oceanography
Rob Pennington, National Center for Supercomputing Applications
Nick Pisis, Oregon State University

Additional Workshop Attendees

Larry Atkinson, Old Dominion University
Peter Cornillon, University of Rhode Island
John Delaney, University of Washington
Eric Itsweire, National Science Foundation
Marlon Lewis, Dalhousie University
Andrew Maffei, Woods Hole Oceanographic Institution
Stephen Meacham, National Science Foundation
Frank Rack, Joint Oceanographic Institutions
Jim Yoder, National Science Foundation

Workshop Speakers

Danny Cohen, Sun Microsystems
Kevin Kalajan, Sun Microsystems
Jason Leigh, University of Illinois, Chicago
Paul Morin, University of Minnesota
Atul Prakash, University of Michigan
Bill Pulleyblank, IBM
Michael Speer, Sun Microsystems
Frank Vernon, Scripps Institution of Oceanography

Preferred Citation

Ocean ITI Working Group. 2004. *Trends in Information Technology Infrastructure in the Ocean Sciences*, 24 pp., www.geo-prose.com/oceans_iti_trends



Trends in
Information Technology Infrastructure
in the Ocean Sciences

Report of the National Science Foundation Ocean Sciences
Workshop on Information Technology Infrastructure

Held May 21-23, 2003
at the Skamania Lodge, Stevenson, WA

Contents

- I. Context** 1
 - A. The First Ocean Information Technology Infrastructure (OITI) Report 1
 - B. Emerging Components of ITI 1
 - C. The Emergence of Cyberinfrastructure at NSF 4

- II. The Information Technology Infrastructure Trends Workshop** 5
 - A. Key Trends 5
 - B. Networks and ITI 7

- III. Next Steps** 19
 - A. Challenges 19
 - B. Pilot Projects and Ocean Information Technology Infrastructure 20
 - C. Some Possible Pilot Project Areas 21
 - D. Challenges in Developing an OITI Management Structure 23

- Acronyms** 24

I. Context

A. THE FIRST OCEAN INFORMATION TECHNOLOGY INFRASTRUCTURE (OITI) REPORT

To assess the ocean sciences community's current and future information technology infrastructure (ITI) needs, the Ocean Information Technology Infrastructure (OITI) Steering Committee, with support from the Office of Naval Research (ONR) and the National Science Foundation (NSF), gathered input from the ocean sciences community, supercomputer centers within the United States, and from ITI experts (OITI Steering Committee, 2002). Particular emphasis was given to the needs of the modeling and data assimilation projects that were being developed through the National Oceanographic Partnership Program (NOPP). The OITI Steering Committee determined that the ocean sciences community was experiencing severe bottlenecks in the availability of high-performance

computing resources (CPU cycles, memory, storage, and network bandwidth). The need for re-coding models for new hardware architectures, data visualization, and community models was also noted to be acute. Lastly, the OITI Steering Committee also noted the extreme shortage of trained technical staff accessible to the ocean sciences community.

The OITI Steering Committee developed four recommendations based on its study and analysis:

1. Improve access to high-performance computational resources across the ocean sciences,
2. Provide technical support for maintenance and upgrade of local ITI resources,
3. Provide model, data, and software curatorship,
4. Facilitate advanced applications programming.

B. EMERGING COMPONENTS OF ITI

Although *An Information Technology Infrastructure Plan to Advance Ocean Sciences* (OITI Steering Committee, 2002) focused primarily on the ocean sciences community's needs for access to high-performance computing (HPC) resources, the report commented that access to local ITI at ocean research institutions is a long-term issue. Given the increasing and continuing trend towards commodity pricing for basic ITI services including computation, storage, and network bandwidth, it has become extremely challenging for research institutions to maintain their ITI capabilities in the face of rapid change. The forces driving this change are primarily the needs for the entertainment and business sectors. Moreover, frequent shifts in hardware and software architectures impose a significant human cost, especially for smaller institutions

and research groups. While the need for HPC will continue, it is also apparent that many HPC services will be provided through local, distributed sources in addition to centralized systems such as HPC centers.

The National Science Foundation convened a workshop on "Trends in Information Technology Infrastructure in the Ocean Sciences" to examine trends in networking and their impacts on the ocean research community. The focus of the workshop, held May 21-23, 2003, was on both the services that would be enabled through ubiquitous networks and the technical and human infrastructure that would be needed to develop and support these services and capabilities.

The oceans sciences community is planning many ambitious long-term observing programs such as the Ocean Research Interactive Observatory Networks (ORION), the Integrated Ocean Observing System (IOOS), as well large field and modeling programs including Climate Variability and Predictability (CLIVAR), Global Ocean Data

Adequate ITI capabilities at each local research institution are essential for designing and implementing new scientific programs and new observing systems that are needed to address complex, interdisciplinary problems in ocean research.

Assimilation Experiment (GODAE), and Ocean Carbon and Climate Change (OCCC). In terms of ITI, there are hardware and software requirements in, for example, computational power, data storage and access, and visualization, such as, “We recommend...initial purchase...[of] capacity equivalent to roughly one-third of a leading terascale machine” (OITI Steering Committee, 2002). These items can be discussed in terms of design or implementation requirements such as, “the system needs to provide 1 terabyte (TB) of on-line storage.” There is a higher level of requirements known as “functional requirements.” These requirements describe the intended behavior of the system in terms of the services, tasks, and functions that it will perform. For these emerging programs, there is a new set of functional requirements that is increasingly important beyond those that we usually describe in terms of hardware and software implementation.

These functional requirements include network and data security (in terms of data integrity, access, and intellectual property), real-time, high-availability (24 hours, 7 day/week), and intelligent methods of sorting and seeking relationships within data and information. As ocean scientists tackle more complex and interdisciplinary problems, the requirement for remote collaboration becomes critical. Moreover, the need to accommodate rapid technological change should be viewed as a functional requirement.

Meeting these new functional requirements imposes serious challenges in a widely distributed ITI architecture. With technical evolution occurring more rapidly at the fringes of the network rather than at centralized HPC facilities, the acquisition of new ITI may follow different paths than what we have relied upon in the past. Nearly 20 years ago, the first UNIX workstations from Sun Microsystems were beginning to appear on the desks of the research community. IBM personal computers and Apple Macintoshes were becoming more prevalent. The rapid changes in price and performance since then are well known. What is perhaps less understood are the implications of this technology for research organizations and for individual scientists. We have moved from a centralized, mainframe computer world where advances in technology were largely driven by the needs of the scientific and technical communities to a distributed computer world that focuses on commercial needs and is driven by mass-market forces.

The pace of change has quickened with significantly shorter product life cycles. The scientific community was used to communicating directly with industry representatives who understood both the corporate product line and the needs of the

customer. With the decline of the mainframe as a scientific computing tool, obtaining and using high-performance ITI has become much more complex for the scientific user. The mainframe vendor provided a complete, end-to-end solution. This is now a task for the user, although one now has the freedom to assemble a cost-effective solution from various hardware and software components. Moreover, planning for frequent infusions of ITI is daunting.

ITI has expanded beyond the traditional computer/disk drive/output device to include laboratory instrumentation, environmental sensors, and even building-control systems. TCP/IP protocols are now used for many devices essential for ocean research, both directly and indirectly. Ocean instrumentation and observing systems are moving towards more widespread, adaptable, always-connected architectures using satellite, cellular, Wi-Fi and other communication methods. Ensuring links among these sensor/instrumentation networks and the storage/computation networks will be an important task in the coming decade.

The net result of these changes is that while researchers may be developing and maintaining their own information technology environments, they need more-sophisticated technical support than in the past. For example, a scientist can assemble an ITI system with capabilities that once were the exclusive domain of supercomputer centers. Although these changes to a distributed IT environment have made it easier to customize systems for a particular analysis or observation problem, there are hidden costs. More time is spent researching, developing, and maintaining these systems, and increased technical skills require increased salaries. The centralized computer center recouped these costs

by charging user fees; the distributed system hides these costs in researchers' salaries and potentially lower productivity on the part of programmers and other technical support staff who are now engaged in system design and maintenance rather than data analysis and modeling. Moreover, the "centralized" services that support a distributed ITI such as network fabric, switches, and disk arrays, must still be funded and maintained, which at most academic research institutions is primarily through individual research grants. The consequence is that these essential services are often severely underfunded.

Attendees at the Workshop on Information Technology Trends in the Ocean Sciences recommend moving towards the next stage of ITI, emphasizing local, distributed computing needs, and integrating these with the earlier OITI Steering Committee recommendations. Moreover, and more importantly, the time is ripe to begin implementing integrated plans for ITI in support of ocean research. This workshop report focuses on the needs of individual researchers and research institutions, which are distinct from needs best addressed at larger HPC centers. In addition, the report focuses upon imaginative solutions that take into account the link between resources at local facilities and those at remote sites.

Adequate ITI capabilities at each local research institution are essential for designing and implementing new scientific programs and new observing systems that are needed to address complex, interdisciplinary problems in ocean research. New ITI capabilities will support more-effective data management and dissemination, advanced analyses of complex data sets, and communication with the public and policymakers.

C. THE EMERGENCE OF CYBERINFRASTRUCTURE AT NSF

In 2003, the NSF released its report on “cyber-infrastructure,” which “refers to an infrastructure based upon computer, information and communication technology (increasingly) required for discovery, dissemination, and preservation of knowledge” (NSF Blue Ribbon Advisory Panel on Cyberinfrastructure, 2003). This NSF report identified cyberinfrastructure (CI) as a “middle layer” that would act as a first-class tool to enable a new level of science. Using global connectivity, CI will support next-generation HPC, instruments for observation and characterization (including organization, activation, and manipulation, such as laboratory instrumentation), knowledge management, collaboration services, and visualization services. The CI report recommended that NSF invest one billion dollars per year in new funds in research and development, provisioning of operational services, support for domain-specific CI activities, and education and engagement of the broader community. Thus, CI is not merely a research and development activity for computer and information science; it also must move forward as a partnership between the domain science communities and the computer/information science communities.

In the wake of the NSF CI report, the three divisions of NSF’s Geosciences Directorate (Ocean Sciences, Atmospheric Sciences, and Earth Sciences) have each established CI working groups, and each has pursued independent, but complementary paths to meet the CI needs of their communities. Integrating the recommendations of these three working groups is a continuous process. One of the first steps toward integration was sharing ideas at a workshop on Cyberinfrastructure for Environmental Research and Education (NCAR, 2003). This workshop identified several challenges and opportunities, many of which were similar to

those identified in the report of the OITI Steering Committee (2002): (1) providing integrated and balanced solutions, with attention to “last mile” needs remains challenging in a diverse, distributed scientific environment, (2) encouraging more-effective collaboration across disciplines and especially between computer scientists and environmental scientists, and (3) balancing innovation with the more mundane requirements of providing IT services, which continues to keep computer scientists and environmental scientists from working together effectively. There are also programmatic challenges as different federal agencies pursue different components of an overall CI without adequate coordination. Long-term, sustained support for CI does not fall within the traditional research-grant focus nor the one-time infusions of infrastructure associated with the Major Research Equipment Facilities Construction (MREFC) program at NSF.

Specific, near-term opportunities were identified for Environmental CI, including:

- Development and deployment of more-effective and economical collaboration tools,
- Improvement in the accessibility and utility of computational tools and interactive capabilities of data systems, data archives, and repositories,
- Model simulation,
- Continued improvement in scalable computing capabilities and access to such capabilities, and
- CI community building.

The Environmental CI workshop participants made two recommendations: (1) NSF should establish a cross-cutting program to support the development and deployment of CI for environmental research and education, and (2) NSF should ensure that this new program is coordinated with relevant activities at other federal agencies.

II. The Information Technology Infrastructure Trends Workshop

Building on the Environmental CI workshop report (NCAR, 2003) and the earlier work by the OITI Steering Committee (OITI Steering Committee, 2002), the Ocean ITI working group was charged by NSF to identify technological trends in the areas of HPC, data grids, sensor webs, collaboration, and visualization. These topics were

identified by the Environmental CI report as the primary capabilities that would be provided by the CI “middle layer.” The ultimate goal of the May 2003 workshop was to develop recommendations to NSF’s Division of Ocean Sciences that could be used to develop specific CI initiatives, with specific emphasis on distributed ITI capabilities.

A. KEY TRENDS

1. Changes in Workflow. The extensive availability of broadband networking is fundamentally changing how the ocean sciences community interacts with its observing systems, and analyzes, visualizes, and manages its data sets. Workflow structure is changing, focusing more on near-real-time analyses and highly distributed collaboration and computation, rather than cruises that may be analyzed over a period of years by a small group of single-discipline oceanographers who may be located at one institution. As noted earlier, the IT industry is now focusing more on “commodity” technologies that are highly distributed and highly connected. With TB storage systems costing less than \$1300 each, the need for highly centralized systems with TB storage capacities diminishes. Similar advances in desktop computation capabilities and other IT components are occurring at the edges of the Internet in the office, laboratory, and home. With high-speed, broadband connections there is the potential to provide a centralized view of this highly distributed system. Moreover, networks are now pervasive and components of the ITI that were formerly isolated (especially observing systems) are now becoming globally connected

to the Internet. Thus, we move data and knowledge through a highly distributed and rapidly changing network, rather than from static data storage systems and centralized computation systems. Although everyone is aware of Moore’s Law in which the density of transistors on a chip doubles approximately every two years, the doubling times of other critical parameters are even shorter. For example, processing speeds double every 1.6 years, storage densities every 12 months, and networks’ speeds every 9 months. Four years from now, chip densities will have increased by a factor of 4, processing speeds by 5.6, storage by 16, and network speed by 40. The 1 TB disk that costs \$1300 today will cost only \$81, and today’s bleeding edge network at 10 Gbps will be operating at 400 Gbps.

2. Distributed Computation. Regional networks have traditionally provided access to distant HPC resources, but the fundamental model is identical to the job submission/computer center model of 30 years ago where a single computing center provided all the tools and jobs were queued and executed according to a local priority system. However, new approaches such as the computational grid rely on

the integration of distributed HPC assets. For example, one center may provide a particular type of computational engine, another a massive data store, and another a data-rendering and visualization engine. Grid technologies provide a unified system view of these resources that are transparent to the user. In contrast, some distributed computation technologies rely primarily on distributed and underutilized computational resources to partition one task across many machines. SETI@Home is one example of a distributed computation system, but most HPC applications require a uniform operating environment in both hardware and software. However, for most compute-intensive ocean applications, simply using “unused cycles” in the manner of SETI@Home is not realistic.

3. Data Storage and Access. The notion of central data repositories has shifted in response to changes in network capacity and local storage technologies. Central archives such as the National Oceanographic Data Center (NODC) have struggled with the increasing volume and variety of oceanographic data. Moreover, the National Research Council (NRC) Committee on Data Management and Computation (CODMAC) reports from the early 1980s emphasized the importance of maintaining close links between active scientific investigations and data archives to ensure that data remained relevant to contemporary science. These forces have resulted in a more distributed and ad hoc approach to data management. Investigators are making data sets available over the Internet through Web-based interfaces to local data holdings. However, it is often difficult to locate such data holdings without personally knowing the investigator. Technologies such as the Distributed Ocean Data System (DODS); its

successor, OpenDAP (Open Data Access Protocol); and the Storage Resource Broker (SRB) have provided standardized methods to publish and access such local data holdings and data stored in centralized archives. New methods, such as XML (Extensible Markup Language) and data ontologies, will provide more capabilities for researchers to organize and publish their data holdings and for locating and accessing other researchers’ data. However, long-term storage and data curation will continue to require a more centralized approach as they exceed the capability of a single investigator. An emerging approach has been the development of intermediate-size, special-focus data centers such as the Joint Global Ocean Flux Study (JGOFS) data system that focus on the needs of a specific science program. This approach maintains the connection between the data sets and science knowledge in an effective manner. No matter what approach is followed, the costs of effective management are substantial.

4. Data Analysis and Visualization. Data visualization and analysis continue to be challenging as the volume and complexity of observations and model output increases. Interactive dynamical analysis, data mining, and semantic-based analysis tools must be developed for ocean sciences applications. To solve scientific problems requires an interdisciplinary approach and, therefore, multiple investigators must be able to share data and insights. Analysis and visualization tools must preserve the provenance and lineage of data sets and analysis products. Simply producing compelling graphics and animations is not sufficient. As with HPC and data storage, many of the advanced capabilities have required specialized tools and technical staff, but increasingly such IT is available at the

local level. In part because of digital games, advanced computer graphics and rendering tools have now moved into commodity machines, many of which have graphics processors with capabilities that once were the exclusive domain of HPC machines. Moreover, higher-bandwidth networks now make it possible to share and create complex visualizations within a real-time, collaborative environment. Thus far, such capabilities are restricted to networked games, but it is easy to imagine scientific applications being developed in the near future.

5. Smart Sensors. Continued miniaturization of IT technology has now led to a wide range of embedded devices used as environmental sensors. Lightweight operating systems (such as TinyOS and Linux) and hardware-based software stacks, such as http and TCP/IP, have enabled the development of “smart” sensors that, with sufficient network connectivity, can now form “sensor webs.” These webs can be fully integrated with the data analysis and delivery system, thus bridging the existing gap where data are collected in isolation for later analysis and archiving. Although the ocean presents a

significant challenge for communication and networking technologies, some components of future ocean observatories (e.g., Internet links to research vessels [HiSeasNet], deep-sea cables, buoys, shore-based sensors) will be able to be connected to the Internet.

6. Collaboration Tools. The ocean sciences are inherently interdisciplinary. The increasing complexity of scientific questions and data sets requires more-sophisticated collaboration tools. Moreover, the scientific expertise required for addressing these scientific problems is distributed among many institutions. Effective, network-based collaboration tools will be an important component of the scientific process. However, present tools are relatively limited in their capabilities, and most network-based collaboration relies simply on email and ftp.

To address these six issues, the Oceans ITI Trends workshop organized a series of presentations by experts from both industry and academia. These presentations were used as the basis for a set of technical reports that are available on the Web (www.geo-prose.com/oceans_iti_trends/).

B. NETWORKS AND ITI

The OITI report (OITI Steering Committee, 2002) recommended several actions regarding access and use of information technology. Most of these recommendations focused on the need to provide high-end resources to address computationally intensive research problems. These went beyond the need for flops to include issues such as technical support, training, and visualization. However, advances in network capability as well their increasing pervasiveness imply that focusing

on local institutional needs is as important as the need for centralized resources. It is not an either/or situation: a balance must be struck between IT capabilities at central nodes and those at the fringes of the network.

Networked capabilities for HPC, data management, visualization, environmental sensing, and collaboration require significant investment at local institutions, all the way down to the desktops,

laptops, and other personal devices. The workshop hosted presentations on: (1) networks, (2) massively distributed high-performance computing, (3) data grids and related issues, (4) wireless sensor grids, (5) distributed visualization, and, (6) distributed collaboration. Each workshop topic demonstrated a form of distributed data manipulation.

1. Grid Computing

Computation is becoming widely recognized as a third component of scientific research, complementing experimentation and theory. As computational models become increasingly sophisticated, they require significantly more powerful platforms. Two factors that continue to drive this need are the desire to deal with multiscale phenomena (such as mesoscale weather models) and the requirement to manage huge amounts of data. Parallelism may provide the needed level of computation, relying on vector architectures, symmetric multiprocessors (SMP), or message-passing architectures. In the context of networked applications, “grid” architectures are being used where the problem can be parsed among several compute nodes distributed over high-speed networks, or where various compute functions (e.g., storage, computation, and visualization) are distributed to different locations. Grid computing uses sophisticated interconnection technology to apply distributed processors to such a problem. It makes use of a relatively modest number of large computer systems, linked by TCP/IP, to assemble sufficient compute resources. Note that for many emerging networks such as the National Lambda Rail (NLR), TCP/IP is far too slow to manage high-speed data transfer.

Grid computing provides a user with an integrated view of many computing systems. A grid consists of a number of separate computing systems, called

nodes, together with an interconnected network that allows communication among these systems. A user on the grid will typically access the grid at some node and then run an application on a node, or set of nodes, on the grid. The choice of node(s) on which to run applications will depend in part on the characteristics of the nodes themselves (processor capability, memory, available capacity), in part on the data available at the node, in part on the communication capacity among nodes, and in part on access controls that form part of a grid infrastructure.

One advantage of a grid is that it permits high-level integration of a broad variety of systems. In principle, a user can select a node appropriate for the application as well as arrange for distributing load across a system. The grid also has the potential to make more-efficient use of resources by allowing them to be used where needed, independent of low-level demand on the node containing the resources. Indeed, one of the early motivators for development of grid computing was the desire to “scavenge” the unused compute cycles on the thousands of PCs and workstations that spend most of their time sitting idle.

A more important use of the grid model may be in sharing data across multiple nodes. For example, a user may have a real-time satellite feed of weather data at one grid node, sensor data at other nodes, and a need to run a model for ocean/atmosphere prediction at another node that has high computation capability.

Because of their distributed nature as well as the high latency experienced by communication among nodes, grids are normally more suitable for very coarse-grained, parallel applications that do not require extensive communication among nodes during execution of application components.

However, they do provide good support for on-demand computing and for autonomic computing (this refers to the ability of a machine to self-manage much of its operation with minimal human intervention)—enabling applications to deal with capacity disruptions and problems at grid nodes.

Grid computing represents a fourth step in the evolution of the growth of Internet-based computer systems. In the first stage, the Internet was intended to provide efficient connectivity between a terminal and a variety of applications running on a (possibly remote) node. In the second stage, it provided the foundation for e-mail—point-to-point communication between two people. In the third stage, we saw the evolution of the World Wide Web, permitting rapid exchange of information among members of a global community. Grid computing extends this communications system to allow applications to be run across a variety of nodes.

2. Data Grids

The ocean sciences community requires a software infrastructure that is able to manage the heterogeneity inherent within data sources, storage repositories, and access mechanisms. Example data management systems, currently in use within NSF, NASA, DOE, and NIH projects, individually address many of the challenges of the ocean sciences environment. By integrating appropriate capabilities from the extant systems, a viable infrastructure can be assembled to support data sharing (data grids), data publication (digital libraries), and data preservation (persistent archives).

The ocean sciences community works with real-time data from sensors, simulation output, observational data collections, and historical data archives.

Data are transmitted over wireless networks at low to moderate bandwidths, over telephone lines and the Internet, and over optical networks at 10 gigabyte per second speeds. The size of data sets, and the rates at which data are moved vary widely. Collection types range from lists of point data, measured in megabytes, to sets of image data measured in tens to hundreds of terabytes. The access mechanisms range from portals for Web-based interaction, to application programmer interfaces (APIs) that directly manipulate data-set structures, to digital libraries that support sophisticated query and discovery. Data are stored on small caches in remote sensors, on-board ships with satellite communication, on land-based disk caches, and in deep archives. The challenge of managing ocean sciences data lies in the remarkably broad range of types of data sources, storage repositories, and access mechanisms.

Data grids provide both the storage repository abstractions (an abstraction is a general concept formed by extracting common features from specific examples or instances) needed to manage interactions with heterogeneous storage systems, and the access abstractions needed to support the desired access mechanisms. Data grids form the core data management technology used to support scientific disciplines, providing the ability to share data across administration domains. Data grids may be used as part of digital libraries, through the addition of services to manipulate information. Data grids may also be used to support persistent archives through the addition of services to manage technology evolution.

Within the ocean sciences community, there are multiple existing efforts for the management of data, upon which appropriate ocean sciences cyber-infrastructure can be assembled. The projects cover

the range from data collection to data organization and storage, to data access. Example technologies include:

- Data acquisition – Object Ring Buffer (ORB). ORBs manage access to real-time sensor data, supporting queries on recent observations. They interface to a variety of sensor types, managing access latencies, network losses, and disparate communication bandwidths.
- Collection building – Storage Resource Broker (SRB) metadata catalog. Descriptive metadata assigned to each sensor observation or simulation output are managed in metadata catalogs. The metadata are organized in collection hierarchies that can be queried and browsed.
- Collection federation – SRB data grid. The integration of ORBs and the SRB creates a Virtual Object Ring Buffer or VORB, enabling access to sensor data stored in multiple ORBs. The SRB data grid is also used to integrate disk caches with archives, enabling access to the most recent observations and archived data through the same access mechanisms. .
- Data access – OpenDAP. Access mechanisms for data retrieval vary from bit-based access through UNIX file system semantics supported by the SRB, to structure-based access through the OpenDAP protocol. OpenDAP supports direct extraction of variables from files, making it possible to request a named data variable. OpenDAP provides a way to describe the structure of data sets, and a way to apply semantic labels to the structures.
- Data analysis – Grid Portals. Management of application codes that support data analysis and manipulation is typically controlled through portals that link simulation output and input files with access to the remote collections. Por-

als can also be integrated with grid services for the remote execution of jobs on systems such as the NSF Teragrid.

- Data analysis – visualization systems. The most challenging data sets manipulated within the ocean sciences community are multi-dimensional satellite observations and simulation results of time-evolving 3-D fields. It is possible for simulation output file sets to be larger than local disk. The 3-D rendering of large data sets is then done through the paging of data from a remote archive, for example, using a 3-D visualization toolkits developed at the San Diego Supercomputer Center (SDSC), the Electronic Visualization Laboratory at the University of Illinois at Chicago, and the Visualization Center at Scripps Institution of Oceanography.

Examples of the piecewise integration of the above systems exist. A VORB is being used within the NSF RoadNet project at Scripps Institution of Oceanography to support real-time sensor data. The metadata catalog is used to manage both descriptive metadata about the sensors, and administrative metadata for data location and access control. The OpenDAP protocol has been integrated with the SRB as part of a NASA data grid for the Data Assimilation Office at NASA Goddard. The system provides access to individual variables through the DODS interface to data stored within the SRB data grid. Grid Portals are used at SDSC to support interactive access to jobs running on the NSF Teragrid. Grid Portals access data on local file systems, as well as data stored within SRB collections. Simulation output files are stored in the SRB along with descriptive metadata. Finally, the 3-D visualization toolkit at SDSC is integrated with the SRB to optimize access to remote data through a data paging system. The 3-D visualizations toolkit has been used to generate fly-throughs of

multi-terabyte simulations that are currently being displayed in the Rose Planetarium at the American Museum of Natural History in New York.

The ocean sciences community needs a data management system that integrates all of the capabilities described above. This is not without substantial computer-science challenges. Development of generic data manipulation systems requires input from the computational science community on the following issues:

- Data formats – generic characterization of structures within data sets. Multiple mechanisms are currently being used to describe and manipulate file structures, including Hierarchical Data Format, OpenDAP, and DataCutter data subsetting filters. A common representation is needed that can be used within any of these approaches to data subsetting.
- Data models – generic description of semantic labels applied to data structures. OpenDAP and the SRB through XML provide a way to apply semantic tags.
- Data dictionary – list of semantic labels used to describe oceanographic quantities. The ocean sciences community needs a set of commonly used terms to describe all variables recorded by sensors and generated by simulations.
- Digital ontology – characterization of the relationships that are applied to data set structures. Examples include mapping of data arrays to coordinate systems, mapping of time stamps to time coordinates, and mapping of coordinate systems to geometries.
- Concept spaces – characterization of the relationships among semantic tags defined in data dictionaries.

- Knowledge management – organization, storage, and manipulation of relationships. An example is support for querying across federated collections, including electronic journals now commonplace in ocean and Earth sciences. Each collection may have a unique set of attributes that have been assigned to the data. The attributes are listed in the data dictionary, and relationships among the attributes are described in the concept space. A knowledge-based query uses the relationships present within the concept space to identify equivalence of terms within the federated collections, to provide a relevant response.

These computer-science issues are highly dependent upon the data models and semantics used by a discipline. Progress in building generic infrastructure for creating and manipulating knowledge requires the active involvement of both ocean and computer scientists. While many of the issues are being explored within the Global Grid Forum and through Digital Library Initiatives, the ocean sciences community must also be involved to ensure that infrastructure relevant to their discipline is created. The true test will be use of data, information, and knowledge management technology in direct support of ocean sciences community data collections.

3. Grid Visualization

The chief requirement for obtaining useful visual displays of data is the collaborative and adaptive real-time fusion, querying, and display of data from a wide range of data sources, such as distributed data stores and real-time remote sensors. This requirement will have the greatest impact in the ocean sciences in the 21st Century. This capability

must interoperate on a variety of display systems, including Web browsers, immersive stereoscopic displays, and scalable, high-resolution displays.

Visual displays of data are most powerful when they show multiple data sets simultaneously. For example, to study ocean circulation, data from a large number of sources must be integrated. In an ideal world, this would be on a single display. Bathymetry from databases and shipboard sonar, wind and temperature data from buoys, and sea surface topography and wave height all must be available and integrated on-demand to provide the most complete picture of ocean state as possible. Visual data fusion also allows us to leverage data sources intended for other disciplines (e.g., geology, atmospheric science). Earth and ocean sciences are unique because they are integrative and multi-scale; there are environmental and natural hazards applications that can significantly benefit from these enhanced visual displays.

As visual data fusion requires rapid access of data from a variety of data sources and sensors from around the world, high-performance international networking is required. Although Internet2 already provides a significant amount of bandwidth to universities in the United States, it is already saturated. Researchers trying to download data have to compete against students who are downloading anything from music to cookie recipes. This will continue to be the case even when more bandwidth is appropriated by Internet2. A different strategy is needed. Oceanographers need to partner with NSF programs that are developing extremely high bandwidth, dedicated, and perhaps even router-less, all-optical scientific networks to ensure that there is uninterrupted bandwidth available. This is

especially important if the measurements are used in real-time emergency response applications such as in the prediction of tsunamis.

Currently, it is difficult for scientists to use the same piece of visualization software across the wide variety of platforms they may want to access, for example, laptops, immersive displays (GeoWalls), and Scalable High Resolution Displays (SHRDs). Every available visualization tool is targeted at a specific computing architecture, and either performs poorly on other platforms, or is not reusable at all. Future generation visualization systems need to be smarter and more flexible. They need to be aware of the underlying networking, data, compute, and rendering services that are available on the Grid to create an optimum visualization package for the viewer. One example of multiscale visualization is the Visualization Center at Scripps' visual objects, built on Interactive Visualization Systems' Fledermaus, which can be displayed on all of these system effectively.,

SHRDs are needed because Earth's resolution is not 1280 X 1024. Data sets are being acquired at higher and higher resolution and they need to be displayed in their full glory. We know from moving from typewriters to word processors that everything does not need to be printed. We can now collect bathymetry data at a resolution of a few centimeters. There is no way to view all the data from a survey on a screen and there is no way that we can print it all. We need to develop scalable resolution displays that can fit in between these extremes.

Currently there is a significant lack of good visualization tools for SHRDs. Tools are available for computer scientists to develop visualization applications, but there are few, if any, visualiza-

tion applications that truly harness the capabilities of SHRDs for specific application domains such as oceanography. Furthermore, commercial tools are few, and the few that can drive SHRDs are not able to manage very large data sets. Research should be conducted to take advantage of trends in the commodity market to develop lower-cost SHRDs for oceanography. Furthermore, there should be research into developing stereoscopic versions of scalable displays because much data in the ocean sciences are inherently 3-D and time-dependent. Stereopsis has already been shown to be invaluable in the geoscience community with the adoption of 200 GeoWalls in a period of two years. The same is expected to be true in the ocean sciences. Current, leading-edge research in visualization uses clusters with commercial, commodity graphics cards to drive display walls (Geowall2) with resolution of 50 megapixels and more. Limitations remain in reliance on 32-bit processors with limited memory addressing. Shared memory systems with TB memory are becoming available to approach scaling consistent with Earth observations and models.

There needs to be aggressive introduction of modern visualization and computing systems into the ocean sciences curriculum, especially at introductory levels. Today's undergraduates are already comfortable using computers to access data from the Web. Both students and researchers are likely to spend more time analyzing their data on computers than going on ocean research cruises. Both communities, therefore, need to be shown the additional benefits that are afforded by the use of grid visualization, data, and computing services. However, introduction at the undergraduate level is crucial to ensure adoption at higher levels.

4. Sensor Grids

Data grids, today's cutting-edge technologies for information management and data manipulation, are distributed networks that integrate data and computational resources. Existing and evolving grid projects deal with static data files, and until the genesis of the ROADNet program, no one had attempted, from a "grid" perspective, to tackle the unique challenges associated with the management and manipulation of continuous data streaming from myriad sensors. The goal of the ROADNet program is to develop an integrated, seamless, and transparent environmental information network that will deliver geophysical, oceanographic, hydrological, ecological, and physical data to a variety of end users in real-time.

Considering that a vast majority of environmental data collected today is not being captured in a way that makes it available to multiple users, it is a challenge to create a real-time (or near-real-time) data management system. ROADNet was designed to meet the following requirements:

- Capture, process, control for quality, and integrate real-time data streaming from different sources—collected for different purposes, on different temporal and spatial scales, and measured by different methods,
- Make heterogeneities in platforms, physical location and naming of resources, data formats and data models, supported programming interfaces, and query languages transparent to the user,
- Adapt to new and changing user requirements for data and data products,
- Dynamically reconfigure with the addition or removal of observational equipment and/or scientific instrumentation,

- Provide Internet access to integrated data collections along with visualization, data mining, data discovery, analysis, and modeling capabilities, and
- Build on or extend current data collection initiatives (e.g., digital libraries), promoting broad-based user access and long-term data stewardship.

The foundation of the ROADNet data telemetry and distribution system comprises middleware that enables users to access real-time data. The middleware, designed for data management computers, has the following attributes:

- Server-client approach for managing the ring buffer,
- Real-time data processing,
- Multiple simultaneous read and/or write clients;
- Clients can be anywhere that is network accessible,
- Data packets may be of any size, format, information content,
- Highly robust and error-free,
- Adaptive to data bandwidth availability,
- Receive and process data from all types of instruments,
- Exchange data with other data servers,
- Reliably transfer near real-time data over wired, wireless, and satellite TCP/IP connections,
- Prioritize data sets to be telemetered,
- Optimize the use of available telemetry bandwidth,
- Locally archive data,
- Recover data from local archives after telemetry outages,
- Provide network security shielding instruments from unauthorized manipulation, unauthorized addition of sensors, and access control, where needed, to data,

- Update metadata in real-time for automatic processing and for data archives
 - o All site characteristics
 - Latitude, longitude, elevation, etc.
 - o All instrument characteristics
 - Detailed responses
 - Orientation angles and precise location relative to reference point
 - o Ancillary information
 - Data format descriptions
 - Communication characteristics
 - State-of-health parameters
- Adopt a relational database formalism for generalizing the archiving of time series data plus metadata plus all processing results,
- All formatting at application level,
- System should be symmetric with respect to read/write/local/remote access,
- System should be error-free, robust and generally of “commercial grade,”
- Impervious to communications failures,
- Impervious to computer shutdown-startup (generally means non-volatile buffers, auto-reconnects, auto-restarts, etc.),
- Accurate timing provided by the data logger.

The data management system that provides these functions creates a real-time data grid that is reliable, flexible, and enables users to access real-time data streams. The ROADNet data grid will address system and data interoperability issues, but it does not yet address semantic interoperability and information integration issues (i.e., techniques that move beyond simple distributed access to data files). In addition, ROADNet does not include features for persistent archives and for user tools and interfaces. Commercial software packages can provide many of these capabilities, and they should be leveraged to minimize development time and costs. A fundamental principle is that instruments

and sensors being sampled adhere to TCP/IP standards. This is, today, generally not the case in sensor networks, but is essential for the suites of sensors whose numbers will likely grow geometrically in the coming years.

In practice, sensor network data collections:

- Are designed to meet the specific needs of an experiment or agency,
- Are created by unique experimental platforms and methods,
- Make measurements at different temporal and spatial scales,
- Employ different data (and metadata) formats and management systems,
- Are highly specialized and customized for particular applications, and
- Lack well-documented and generalized abstractions for representing data and processing results.

Until recently, these inherent characteristics have had a negative impact on the broader utility, and subsequent longevity, of many scientific data collections. Thankfully, rapid developments in information management technologies are, for the first time, enabling us to move beyond these inherent constraints and to discard old communications, networking, and data management paradigms. The need for an interoperable infrastructure that enables new science based on distributed computing, data sharing, and information integration is driving many national-scale projects in several disciplines. Furthermore, the exponential growth in storage density and network speed with doubling times of significantly less than a year has eliminated the need for centralized data archives.

Several key factors are driving new technologies for managing scientific data collections:

- The development of an appropriate information model (XML) for describing scientific data,
- The ability to organize collections dynamically using XML Document Type Definitions (DTDs),
- The use of Extensible Stylesheet Language (XSL) style sheets to create interfaces to collections that can be tailored to the requirements of separate user communities (or domains), and
- The development of interoperable systems that support the federation of data collections.

With these new technologies, access and management of scientific data collections becomes a matter of manipulating an associated information model. This is an active area of research in computer science. The ocean sciences community's data management toolkit is being expanded by community-led efforts to develop digital library services, parallel compute platforms, distributed computing environments, and persistent archives. There is a strong synergy among these efforts; each requires an ability to manage knowledge, information, and data objects. This synergy has had a significant but unforeseen consequence: although the requirements driving infrastructure development within specific communities are different, development efforts have converged on a uniform architecture. This architecture provides a suite of functionalities seamlessly assembled to form a "grid."

End-to-end systems should provide support for:

- Information Discovery – ability to query across multiple information repositories to identify data of interest,
- Data Handling – ability to read data from a remote site for use within an application; in fact, the specific location should be immaterial and transparent,

- Remote Processing – ability to filter or subset data before transmission over the network
- Data ingestion or publication – ability to add data to collections for use by other researchers, and
- Analysis – ability to use data in scientific simulations, for data mining, or for creation of new data collections.

Technology research and development should focus on:

- Network design flexibility that accommodates new data sources, ever-increasing data rates and volumes, changes in Internet performance, and variations in user demands and display devices,
- Real-time data delivery and quality control to facilitate personalized, seamless, and transparent access to sensor data streaming from the field and in archival data collections,
- System design that maximizes network reliability and configurability, enabling system components to be reconfigured as determined by shifting priorities for data capture based on real-time environmental events and triggers (i.e., earthquakes or oil spills),
- Integration and dissemination of information for time-critical analysis facilitated by XML-mediated query processing,
- Continuous archives of raw, processed, and analyzed data,
- Access control systems that rank user priority at the network (bandwidth consumption) level and authorize data delivery at the Internet interface, and
- Exploitation of the bidirectional communications in the Internet to allow interactive instrument and network control.

The proposed architecture for ocean sciences is based on a grid-type infrastructure design adapted for real-time data collection and integration. The core of the system comprises:

- A data-handling system that will enable access to data repositories and fast data caches across a distributed network connecting heterogeneous storage systems. The data caches will be used for data staging of real-time data that will be moved after preliminary processing to near-line storage. Managing real-time data accumulation and management across distributed caches and archival storage systems is a key IT research goal.
- An information discovery system that will integrate multiple metadata repositories and enable users to discover data based on data characteristics instead of location. The information discovery system of the knowledge network will manage multiple levels of metadata, from IT-centric metadata to discipline-standardized metadata to sensor/application-level metadata. Metadata integration at multiple levels and across disciplines, from ocean sciences metadata to geodetic metadata to seismic metadata, will be a primary research goal.
- An integrated execution system or scientific work flow that will provide operations on data and data streams at multiple locations in the data management corridor, including near-sensor operations such as data validation to data storage sites; operations such as subsetting, metadata extraction and reformatting; and near-application operations including data layouts and presentation. Extraction of metadata from real-time data flow, as well as metadata fusion across multiple sensor data, is an essential research goal.

The key feature is the development and integration of these components to generate a higher-order, multidisciplinary, combined functionality that can be used for large-scale data processing and sharing.

5. Networked Collaboration

The construction of group collaboration systems has been pursued in both the commercial and academic sectors through the use of both proprietary and open-source tools. For example, the University of Michigan has developed such systems to support scientific research, including NSF-funded collaborative projects in the space sciences. Partly as a result of experiences from few successes and many failures of these collaborative projects, several successful spin-off projects have resulted, including the University of Michigan Coursetools (<http://coursetools.ummu.umich.edu>), which is now used by a significant fraction of University of Michigan instructors to put their course materials online and support class discussions, and UM.Worktools (<http://worktools.si.umich.edu>), which is used to support distribution and archiving of material related to various committees and groups at the university.

Collaboration is natural for oceanographers. During an oceanographic cruise, it is common for a wide variety of scientists to be involved in all the different aspects of data collection and interpretation. For example, in core interpretation, a single core will be analyzed for chemistry, stratigraphy, sedimentology, microfossil species, and other core attributes. Interpretation can progress much more smoothly if each scientist's interpretation, smear slides, well logs, and handwritten notes can be made and stored in such a way that they can be called up on demand. All of this can be extended to remote, Internet-based collaboration when ashore.

Collaboration systems for desktop document sharing is mature, but systems for scientific collaboration employing large data sets and high-resolution, real-time graphics are not. There is currently no widespread market for tools of this kind—consequently, companies have little incentive to develop such technologies for the scientific community. AccessGrid is the only currently widely adopted infrastructure for scientific collaboration, and the development of future science-oriented collaboration tools should attempt to adopt the AG2.0 framework. Tool development should follow the guidelines being formalized in the Advanced Collaborative Environments Research Group in the Global Grid Forum.

Below is a list of lessons learned from our past experience with collaboration tools and future research challenges.

1. Research creating user-level network overlays over available topology is an important issue in building scalable and high-performance collaboration systems. Firewalls and NATs, for example, often hinder use of collaboration tools. Simplified ways must be found for developing high-performance collaboration tools and making them available for use over such networks.
2. New collaboration paradigms need to be explored. Examples of popular paradigms are email, instant messaging, and peer-to-peer infrastructures. But each has weaknesses and strengths. Examples of potential extensions of these systems to make them more powerful from the collaboration perspective are email services that include file sharing and “presence awareness” (that is, knowing who is on line), instant messaging services that can include attachments and support for “disconnected” users, and peer-to-peer infrastructure for conferencing and better security.

3. Security policy management needs to be simplified in group collaboration systems.
4. Security will need to be implemented at the endpoints and made largely transparent. Collaboration systems make it difficult to manage trust and security at the servers alone. Infrastructure will need to ensure privacy during collaborations.
5. Distinguishing good data from bad data is an important research challenge in open collaboration systems. Collaboration systems generally make it easy for people to contribute data, but not to delete data. Even if deleting data were simple, few incentives exist for people to delete any data and often there is no reliable indicator of ownership. People will need more control of their service subscriptions (e.g., research in content-based, publish-subscribe systems becomes relevant).
6. User education and administrative support can be crucial.
7. For success, think big, and support many communities. If collaboration tools are developed only for a specific community (e.g., ocean sciences) and an attempt is not made to extend them to other communities for a wider adoption, the effort is less likely to be successful.
8. Understand user requirements and show clear benefits to BOTH individuals and groups. Unless individuals see benefits from the tools, such as improving their personal productivity, they are less likely to use group tools, or contribute data to them, just for the benefit of others. For example, CourseTools benefited instructors by speeding up creation of a protected course Web site; class lists were automatically updated with information from the registrar's database. This tool thus became widely adopted.
9. If adoption is the goal, it is often best to stick to stable, widely available technology. This is particularly true of collaboration software because if even a fraction of people don't participate, the group may choose to give up on the collaboration tools.
10. Technology adoption challenge is often underestimated. This is particularly true of collaboration technology because it can cause a cultural shift, may cause paradigm shifts in the way people do things, and often requires mass adoption to succeed. Experiences from the Upper Atmosphere Research Collaboratory (UARC) and the Space Physics and Aeronomy Research Collaboratory (SPARC) collaboratory projects, as well as CourseTools and WorkTools projects to highlight the technology adoption aspects—some of these tools turned into niche tools while others continue to be heavily used at the university. The differences in where these tools are now used are probably not a result of technology differences—"marketing" and the nature of the audience targeted by the tools probably played a significant role as well.

In summary, basic information technology is fast becoming a commodity. In research, we thus need to place more emphasis on issues such as security, reliability, and making it easier for people to adapt the technology. This is particularly pertinent for collaborative infrastructures.

III. Next Steps

A. CHALLENGES

Both the OITI report (OITI Steering Committee, 2002) and the NSF/Geosciences Cyberinfrastructure report (NSF Blue Ribbon Advisory Panel on Cyberinfrastructure, 2003) highlighted many of the same issues regarding the importance of CI for the “domain sciences” (i.e., non-computer sciences), as well as the challenges in moving forward. CI, if properly implemented, will enable a new level of cross-disciplinary science and knowledge delivery. Moreover, the rapid pace of technological change has made CI, which was once the domain of supercomputer centers, within the reach of the individual scientist.

The nature of the scientific questions in ocean sciences today requires a more wide-ranging approach that does not lend itself to the traditional pipeline processing model where it is relatively easy to allocate specific functions to specific hardware and software components. In the past, we could collect our data (often recorded on paper charts or notebooks) at rates that were easily manageable by a single individual. Our analysis tools were generally home-grown (and written in FORTRAN). We published only in peer-reviewed journals, sometimes years after the data were collected. We are now asking complex, multidisciplinary questions, and the answers need to be reconstituted in a manner to serve a much broader community than simply our disciplinary peers. The volume and complexity of the data sets go far beyond the capacity of a single researcher to manage either from a data management or a data interpretation perspective. The entire process of collection, analysis, and publication is dispersed, complex, and often performed in near real time. With the advent of ocean observatories this trend will accelerate. Thus, our approach to CI must be much more dynamic and iterative; there are no more “point solutions.”

Many of today’s issues, ranging from climate-change research to homeland security, cross traditional disciplinary boundaries. They require synthesis of information from many sources, and they must adapt in response to changing requirements and policies. The information systems supporting these complex, multidisciplinary efforts must also be adaptable. However, our information models are generally static, pipelined, and product-focused.

Modern CI is technically challenging as the line between hardware and software functionality blurs, and the role of the CPU expands beyond simply providing floating-point operations. We are effectively building complex “ecosystems” of CI, and as with all systems, there is emergent and often unexpected behavior. The ocean sciences community is struggling to develop flexible information systems, and sometimes lapses into the old pipeline model with defined, static workflows. Such an approach leads to data and knowledge silos, which are not suited to modern cross-disciplinary research. For example, optimizing a compute system to run a particular benchmark may not be effective for the ocean sciences community where the models and the underlying hardware architecture change frequently. Our workflows must be understood and be able to respond to changing science requirements, and we need solid hardware, libraries, and compilers that are well-integrated to provide robust systems.

The concept of dynamic workflows that go all the way from data collection, analysis, assimilation, and publication requires a new class of software that seeks to capture the basic essence of the scientific structure and dynamics. Moreover, the sensor networks themselves can be dynamically reconfigured

and adapted to the outputs of assimilation models. Developing the necessary software frameworks for such a vision will require that we identify object attributes and their relationships. In this case, sensors, analysis tools, and other components of the workflow are conceived as objects that have defined interfaces and methods.

An ontology is one approach for developing the underlying conceptual models and the relationships among data objects. Furthermore, in the context of a knowledge management framework, new rule sets can be defined and extended for use by computational processes. Such an approach would begin to bridge the gap between the physical world (e.g., sensor networks associated with ocean observatories) and the digital world (e.g., shoreside analyses and publications). Methods such as ontologies and the semantic web will accomplish this bridging through the encapsulation of knowledge in software.

Along with the technical challenges of hardware and software, there are challenges with the human resource side of information infrastructure. Building and retaining a large pool of technically capable people is difficult for oceanographic institutions

and programs. We do not need, necessarily, more programmers; we need individuals familiar with ocean sciences and with modern IT. Such individuals are increasingly difficult to locate and retain, given the soft-money nature of ocean sciences.

Information systems for the 21st century will require a systems approach to both hardware and software; our traditional focus on CPU performance, while continuing to be important, is not the central focus. Issues of distribution and scalability will become dominant. Our approach will need to be based on a firm understanding of the processes involved, instead of the tiered architecture of compute, storage, and networking. The challenge is that this introduces a higher level of development complexity; typical hardware and software vendors cannot work at this level of integration, which means that the scientist must be able to define requirements and identify strategies. Thus, the widespread proliferation of networks presents both an opportunity and challenge to the ocean sciences community. Technical issues such as authentication and security clearly require considerable thought and development, but there are science issues such as data provenance that will require the community's involvement as well.

B. PILOT PROJECTS AND OCEAN INFORMATION TECHNOLOGY INFRASTRUCTURE

A common theme that emerged during the oceans ITI workshop was the imperative for cyberinfrastructure projects that focused on domain-specific applications of new IT technology. Both the academic and private-sector computer scientists emphasized the need to develop new IT capabilities in the context of real-world problems. This is consistent with the NSF cyberinfrastructure report (NSF Blue Ribbon Advisory Panel on Cyberinfra-

structure, 2003) and with the results of the NCAR Environmental Research and Education workshop on cyberinfrastructure (NCAR, 2003). The level of discussion and interaction between computer scientists and oceanographers at the oceans ITI workshop demonstrated that such partnerships can be an effective means to achieve the goals of the NSF cyberinfrastructure program.

An initiative in ocean information technology infrastructure would be a timely investment that positions the ocean sciences community to confront the impending data deluge. The amount of sensor data and simulation data that is expected in the next five years will dwarf all current systems in size and diversity. The ocean sciences community needs to act now to develop the necessary data management infrastructure. The challenges that must be met include not only data management and preservation, but also data analysis to generate new knowledge. Information management technologies will be part of the emerging cyberinfrastructure. The generation of information and knowledge will require the ability to analyze entire data collections. The implication is that ocean information technology will need to be linked to the NSF Teragrid to support comprehensive analyses of ocean data and numerical models.

Two types of projects emerged during the workshop discussions: precursor projects and pilot projects. Precursor projects are those where substantial progress can be made over a two- to three-year period on a community-wide issue. For example, definition of semantics for oceanographic data could be one such precursor project. Pilot projects are based on more-exploratory concepts where the

research agendas of both computer science and oceanography will be advanced. The workshop established the following criteria for selecting such precursor and pilot projects:

- Precursor projects should achieve demonstrable success in two to three years,
- Pilot projects should significantly advance both ocean and computer sciences,
- Both types of projects should address important community-wide problems,
- Both should build bridges between ocean sciences and computer sciences,
- Both should justify increasing financial support over time,
- Both should have a quick, substantial buy-in from the ocean sciences community.

Pilot and precursor projects should test each component of the information management infrastructure. The projects will be implemented most rapidly by building upon existing systems that already support data collections for Earth systems science projects. Many of the pilot projects should be demonstrations of either the integration of appropriate systems to provide better management and access, or the application of an existing system to a new collection to show generic utility.

C. SOME POSSIBLE PILOT PROJECT AREAS

There are numerous possible pilot projects that will demonstrate either the integration or application of emerging information technologies for use by ocean scientists.

Project Area 1: Integrate modern data-access mechanisms on top of the emerging technologies for management of sensor data. Projects would demonstrate a standard data manipulation

method for selected data set formats for sensor data. The data would be retrieved from the sensor networks, which in turn would access either a system containing the most recent data or an archive. The combined environment would demonstrate the integration of data-collecting technology with digital library technology for access, data grid technology for distributed data storage, and persistent archive technology for long-term preservation.

Project Area 2: Integrate multiple existing collections through a common data federation mechanism. Within the ocean sciences community, multiple data collections are being formed, both on the basis of individual projects, and on the basis of different originating institutions. A demonstration is needed of the ability to provide uniform services across separate collections, as well as the ability to federate disjointed collections into a community library. Possible approaches are to use peer-to-peer federation mechanisms such as those implemented in the SRB, or knowledge-based mediators that impose a domain-specific concept space across the collections. A very aggressive project is to seek the integration of the knowledge-based mediation approach with the peer-to-peer federation technologies, ensuring both semantic and physical access to the data.

Project Area 3: Develop standard approaches for processing systems. A standard approach to the multitude of processing threads is needed to help manage workflow in an increasingly complex computing environment. Workflow management may have special application to HPC, as tools such as multidisciplinary data assimilation, or the development of nested models that involve complex links, are increasingly employed by oceanographers.

Project Area 4: Develop standard services for analyzing and manipulating data. These services could be applied to all elements of a data set. For example, the ability to generate images in the same fashion for all data studies—all data sets—would be a great boon for those attempting to bring research tools into the classroom. Real-time monitoring of remote instruments, or in laboratories on ships, would profit from a standard set of protocols.

Project Area 5: Develop “off-the-shelf” modules that can be easily extended and provide simple-to-use interfaces. Such a “collection in a box” (perhaps “model in a box”, or “data in a box”) or “data bricks” (commodity-based disk caches—also called “grid bricks” or “cyber bricks”) would be widely available and robust, with easily accessible documentation, and formal training of some kind, for the user. Moreover, they would be designed to be used in a parallel environment so that the highest-end computing facilities are not necessary. An example of such a “collection in a box” might be a modest community model, or a data set for a multidisciplinary program in its early stages of development.

Project Area 6: Develop new methods for data preservation, especially for data sets not associated with federal archives. This effort goes well beyond mere archiving; it involves thoughtful replication of the data resources, ready access via online facilities, embedded manipulation tools for ease of use and interpretation, and authentication so that the user may have confidence in the data he or she is accessing. Every large data set in the ocean sciences, existing or conceived, faces these challenges.

Project Area 7: Develop tools to foster knowledge discovery and collaboration. Widely available tools that are interactive, and could be linked readily to specific scientific questions, would be welcomed. There is another area of interest to all, extending well beyond the borders of ocean sciences—“the Web.” Concerns include (but are not limited to): security, access control, service and discovery functions, resource management, authentication, data authenticity, and updating the community on available solutions.

D. CHALLENGES IN DEVELOPING AN OITI MANAGEMENT STRUCTURE

A management structure must be developed for OITI to ensure its success. Although the traditional NSF approach of workshops and community input works for many large programs, there may not be sufficient commonality in the requirements to allow this approach to work. CI is not a ship nor a program. CI is wide-ranging in its capabilities and in its implementation. CI has the potential to be used for a wide range of NSF-funded activities, not just research.

Many IT projects have failed in the past because of an inability to define requirements sufficiently (or correctly), especially when coupled with rapid technology evolution. NASA's EOSDIS is a classic example. As it was being defined, Web access to data was becoming an accepted practice in the scientific community, yet the centralized, large, industry-supported EOSDIS struggled to provide this level of service.

The challenge to NSF is twofold. NSF is comfortable funding large, focused pieces of infrastructure such as HPC centers. It is also comfortable in funding individual investigators. The result may be centers that are struggling to find their place in a commodity IT world coupled with a dispersed and uncoordinated set of local resources provided to individual investigators. Neither strategy leads to a sustainable program in CI. The center approach cannot keep up with the rapid changes at the fringes of system in regards to computation, data management, and data analysis, while the principal investigator approach focuses solely on the IT needs of a specific scientific project.

The University National Oceanographic Laboratory System (UNOLS) model is one possible approach to managing and developing CI resources. It could enhance the capabilities of individual investigators as well as ensure the continuing support of local and regional infrastructure within the national capability for CI. However, particular attention would need to be paid to research and development so that new capabilities can be incorporated into the overall CI. The focus should be on providing services to the geosciences community, where flexibility and evolution is a key requirement. Fundamentally, there are no "solutions" to the community's needs for CI. A research and development program that focuses on the needs of the geosciences community could ensure that new technology is identified and incorporated into CI.

A project office to manage programmatic elements of the NSF/OCE cyberinfrastructure initiative would ensure a successful, community-wide effort. The office would develop a five-year plan for community-wide elements of the program (the plan would be revised at regular intervals), leaving individual projects to the traditional NSF process. The project office might operate under a cooperative agreement, along the lines of the Integrated Ocean Drilling Program, one of a number of possible structures that have proven to be successful in large-program management.

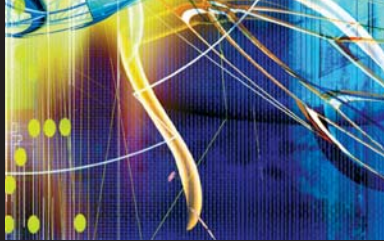
Acronyms

AG2.0.....	AccessGrid 2.0
APIs	Application Programmer Interfaces
CI	Cyberinfrastructure
CLIVAR.....	Climate Variability and Predictability
CODMAC.....	Committee on Data Management and Computation
CPU	Central Processing Unit
DOE	Department of Energy
DODS.....	Distributed Ocean Data System
DTDs.....	Document Type Definitions
EOSDIS.....	Earth Observing Station Data & Information System
ERE	Environmental Research & Education
GODAE	Global Ocean Data Assimilation Experiment
HPC	High-Performance Computing
IBM.....	International Business Machines Corporation
IOOS	Integrated Ocean Observing System
IT	Information Technology
ITI.....	Information Technology Infrastructure
MREFC	Major Research Equipment Facilities Construction
NASA	National Aeronautics and Space Administration
NAT	Network Address Translation
NCAR.....	National Center for Atmospheric Research
NIH.....	National Institutes of Health
NODC	National Oceanographic Data Center
NOPP	National Oceanographic Partnership Program
NRC	National Research Council
NSF.....	National Science Foundation
NSF/OCE.....	NSF Division of Ocean Sciences
OCCC.....	Ocean Carbon and Climate Change
OITI.....	Ocean Information Technology Infrastructure
ONR	Office of Naval Research
OpenDAP	Open Data Access Protocol
ORB	Object Ring Buffer
ORION.....	Ocean Research Interactive Observatory Network
ROADNet.....	Real-time Observatories, Applications, and Data management Network
SDSC	San Diego Supercomputer Center
SETI	Search for Extraterrestrial Intelligence
SHRDs.....	Scalable High Resolution Displays
SPARC.....	Space Physics and Aeronomy Research Collaboratory
SRB	Storage Resource Broker
SMP	Symmetric Multiprocessors
TB	Terabyte
TCP/IP	Transmission Control Protocol/Internet Protocol
UARC	Upper Atmosphere Research Collaboratory
UNOLS	University National Oceanographic Laboratory System
VORB	Virtual Object Ring Buffer
XML	Extensible Markup Language
XSL	Extensible Stylesheet Language
WWW	World Wide Web

References

- OITI Steering Committee. 2002. An Information Technology Infrastructure Plan to Advance Ocean Sciences. 80 pp., www.geo-prose.com/oiti/report.html
- NSF Blue Ribbon Advisory Panel on Cyberinfrastructure. 2003. Revolutionizing Science and Engineering through Cyberinfrastructure. Arlington, VA, National Science Foundation, 84 pp., www.cise.nsf.gov/evnt/reports/toc.cfm
- NCAR. 2003. Cyberinfrastructure for Environmental Research and Education, Boulder, CO, 15 pp., www.ncar.ucar.edu/cyber





www.geo-prose.com/oceans_iti_trends

October 2004