The Use of e-Science Grids to Support ORION

By Larry Smarr
Director, California Institute for Telecommunications and Information Technologies
Harry E. Gruber Professor of Computer Science and Engineering, UC San Diego
lsmarr@ucsd.edu
Draft 12/7/2003

e-Science Requires Cyberinfrastructure

We are at a point of major change in the way in which data-intensive science in carried out. Modern sensors of all kind are capable of gathering data at an ever growing rate. Whereas these scientific instruments used to store the data they produced locally and only later would some "cleaned up" version of the date be placed in a networked data archive, much more frequently today's instruments are sending their data and instrument profiles directly into networked federated repositories available to anyone on the World Wide Web. In the near future, this will evolve into a distributed Data Grids managed by middleware sitting over the networked instruments, computers, and data storage systems. This allows user-defined software to be written which can automatically retrieve linked data sets generated by multiple instruments worldwide.

More actively, a user can reach out through the net to control in real time a remote instrument, a process called telescience [see https://telescience.ucsd.edu/]. Essentially, any Internet connected remote instrument can be used by a distant user. Or in response to an unexpected event (e.g. a forest fire, a supernova, an earthquake) an intelligent constellation of instruments can be taken out of their normal monitoring mode and collectively aimed at the event to fuse their data types into a real-time multimodal observation of the event [e.g. NASA's sensorweb concept-see http://ams.confex.com/ams/FIRE2003/techprogram/paper_75080.htm].

Since many high performance computers are attached to the same Internet, one can easily imagine simulation codes which could use intelligent agents to seek out the data needed to initiate a run. Or an observer could use such a simulation code to perform data assimilation to smooth out an irregularly spaced dataset. That simulation can be used to determine where more data would most reduce the uncertainty in the overall physical system being studied. Ultimately, the code could use a sensorweb to reach out in real-time to the instruments themselves to get the data needed directly rather than go to aging data archives [esto.nasa.gov:8080/conferences/igarss-2002/ 02Presnt/02060820.ppt].

Essentially, these emerging e-Science systems blend the world of data archiving and model simulations into an integrated knowledge system which is constantly being improved by the addition of new data. I believe this is the vision which ORION should aspire to.

In such a world of distributed cyber-infrastructure the separate processes of data gathering, archiving, assimilation, and mining can be interconnected with data simulation and remote instrument control, all through the use of specialized Grid middleware. Such a transformation creates a "digitally enabled science." Essentially all science is carried out in an integrated software system whose endpoints are the scientific instruments, the storage devices contains the data, the computers which compute on the data and the human interface to this global digital system. This is being extended to include real time collaboration between the humans studying the data and the scientific literature which can be hyperlinked into the data.

More and more shared science "observatories" are being designed in this networked digital fashion, although none of them yet have achieved the high vision described above. Among them are large-scale experiments in the U.S., funded by NSF, NASA, NIH, DOE, as well as international science and engineering projects funded by the UK, EU, Japan and others. In fact, the systematic use of a "Grid" architecture is more advanced in the UK and in the EU than in the US. [see e.g. http://web.datagrid.cnr.it/GridMediaRepository/europe_exceeds_US.htm].

[For details on these project see:

www.geongrid.org
www.griphyn.org
www.neesgrid.org
www.nbirn.net

www.nsf.gov/bio/neon
http://spsosun.gsfc.nasa.gov/eosinfo]

One of the most complete overviews of the potential for such distributed science information systems is a report created for NSF by a blue-ribbon panel called the Atkins Report: [NSF Report on Revolutionizing Science and Engineering through Cyber-Infrastructure (Atkins Report) **www.communitytechnology.org/nsf_ci_report/**]

From Networked Archives to Grid-Enabled e-Science

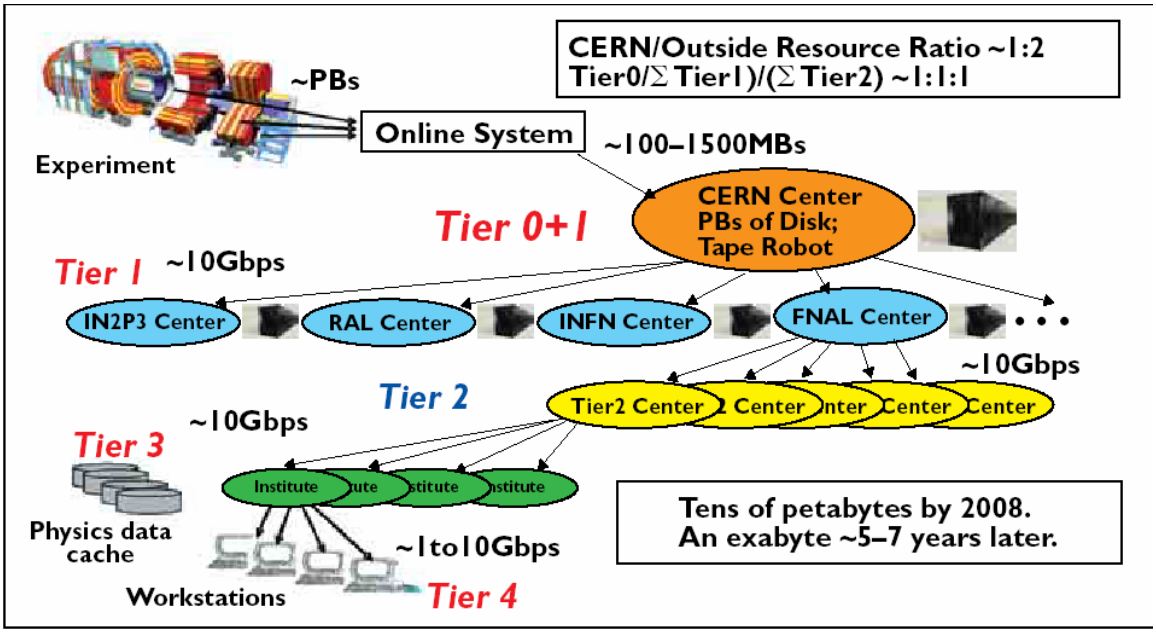
Since Grids have really only emerged over to last five years, existing e-Science systems have been defined and deployed using an evolving set of these concepts embedded in them. Let's examine briefly a few of them.

NASA's EOSDIS [http://spsosun.gsfc.nasa.gov/eosinfo/EOSDIS_Site/index.html], operational since 1994, manages data from NASA's Earth science research satellites and field measurement programs, providing data archiving, distribution, and information management services. It also commands and controls EOS satellites and instruments, and generates useful products from orbital observations. EOSDIS also supports generation of data sets made by assimilation of satellite and observations into global climate models. EOSDIS is a distributed system with many interconnected nodes, including Distributed Active Archive Centers (DAACs) with specific responsibilities for production, archival and distribution of Earth science data products (e.g. physical Oceanography or Snow and Ice). EOSDIS is the largest e-Science system in production, digesting daily over 3 TeraBytes of new data from Earth Systems satellites. In 2003 there were 25 million data products derived to over 2.3 million data request. This system was architected a decade before today's Grid middleware systems were developed, but it is the best example of how a global set of diverse scientific instruments can have a common networked storage system available to a large number of scientific and application researchers.

In contrast, the NSF funded a large Information Technology Research project in 2001 called GriPhyN (www.griphyn.org) to study how to create a modern Grid cyber-infrastructure to support several large-scale physics data projects: the Laser Interferometer Gravitational Wave observatory (LIGO), the Large Hadron Collider (LHC), and the Sloan Sky Survey. These communities are engaged in the collaborative analysis and transformation of large quantities of data over extended time periods. GriPhyN has introduced the abstraction of a petascale "Virtual Data Grid (VDG)," a scalable system for managing, tracing, communicating, and exploring the derivation and analysis of diverse data objects. This allows for tracking "virtual data objects" which are derived from transformations of datasets which were generated by the scientific instruments. VDG expands the data system architecture to an integrated treatment of not only data, but also the computational procedures used to manipulate data and the computations that apply those procedures to data. This integrated treatment is motivated by two observations: first, in many communities, programs and computations are significant resources— sometimes even more important than data; and second, understanding the relationships among these diverse entities is often vital to the execution

and management of user and community activities. GriPhyN terms this integration "a virtual data system," because it allows for the representation and manipulation of data that does not exist, being defined only by computational procedures.

The largest of these data-intensive projects is the LHC, headquartered in CERN in Geneva, Switzerland. Unlike ORION, the LHC has just a few instruments generating data from one site. However, it shares with ORION the need to make this data available to a large number of end users in many countries. Below is a diagram from one of the PIs of GriPhyN, Caltech's Harvey Newman, showing a conceptual diagram of one possible distributed Grid infrastructure being studied for supporting the LHC when it comes online in 2007.

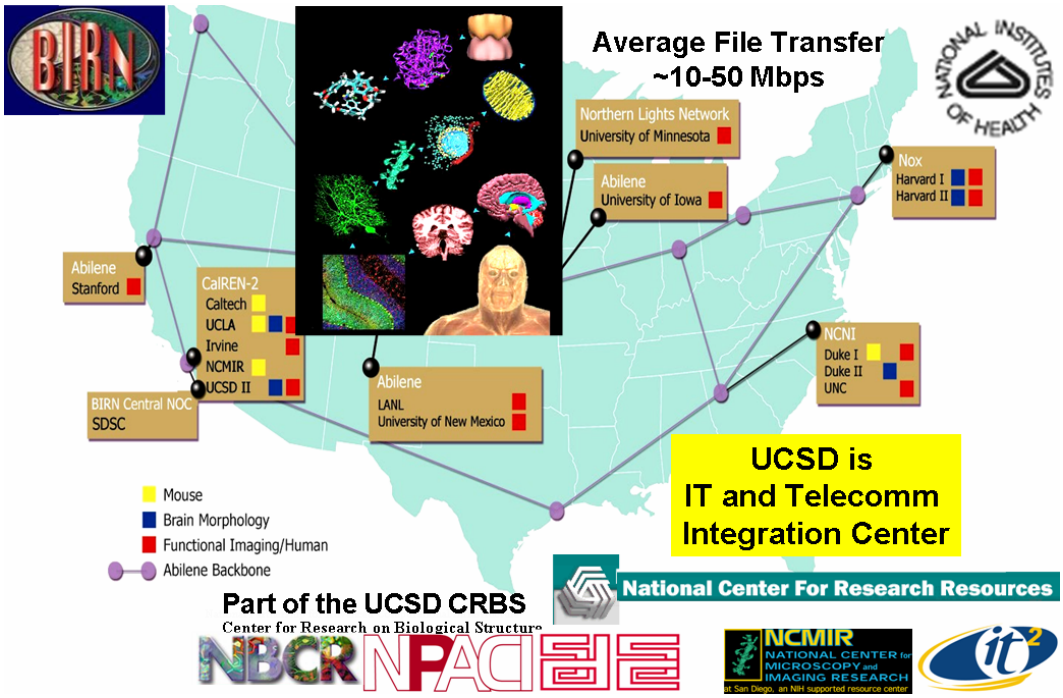


[Figure from **Data-intensive e-science frontier research by** Harvey B. Newman, Mark H. Ellisman, John A. Orcutt, pages: 68 – 77 in Communications of the ACM, Volume 46, Issue 11 (November 2003) Special issue: Blueprint for the future of high-performance networking.]

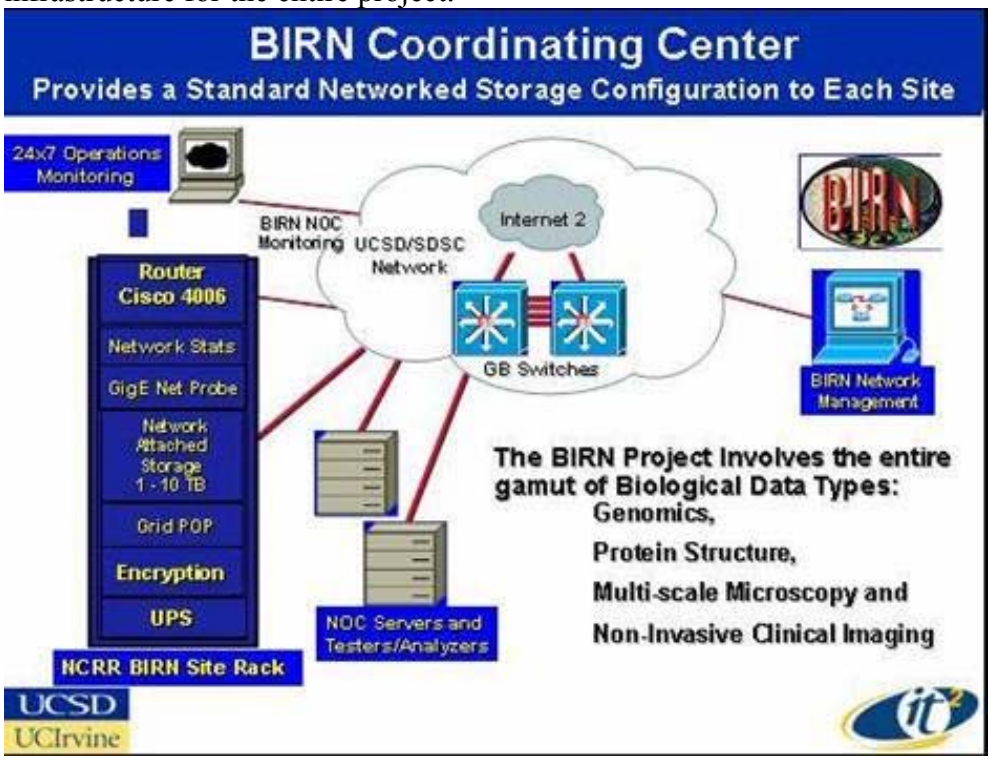
Another Data Grid described in that same CACM article, is the largest one supported by the National Institutes of Health (NIH)--the Biomedical Information Research Network (BIRN-see www.nbirn.net), which feeds output from multiple digital microscopes and 3D imaging devices into a common software federated repository, connected by Internet2, and using a modern Grid middleware infrastructure. The first series of BIRN projects connect multiple university research groups working on multi-scale imaging of the brains of mice and humans.

I believe the BIRN project is a good model for ORION because it is a cyber-infrastructure that connects multiple types of scientific instruments at different sites to a national-scale federated repository using modern data and Grid middleware. The end users are focused on different scientific research topics that draw on multi-scale data objects in the repository. Most of the data objects are currently 2D or 3D images, but

these are becoming linked with other datatypes, such as proteins or gene expression data appropriate to various subareas of the brain.



But perhaps the biggest success feature of BIRN is that the funding agency and the end users supported the notion of a single BIRN Coordinating Center, [see www.calit2.net/birn/features/1-7-03.html] which has responsibility for the entire BIRN Grid: defining its data and software architecture, developing a standard BIRN "rack" which is deployed at each end user lab as the BIRN access device, hiring and training the support people for each BIRN site, and monitoring and tuning the networked infrastructure for the entire project.



This division of labor allows the scientist to simply access the entire Grid system through a local "BIRN Portal" and get on with doing then science while the BIRN CC focuses on supporting a constantly improving cyber-infrastructure dedicated to their discipline.

What is a Grid?

I have used the term "Grid" frequently without really defining it. Here I discuss it a bit more fully and give references to exhaustive discussions of all the subtopics. A "Grid" is a distributed "virtual" collection of networked computing, data, visualization and instrumentation resources that are integrated seamlessly via middleware into a collective whole that provides an integrated capability to end users. Thus, a Grid is a more sophisticated software overlay of traditional Internet connected resources. A Grid may be optimized for computation ("computational Grid"), data storage (Data Grid), or for collaboration (Access Grid). For a fully elaborated version of each see the NSF TeraGrid (www.teragrid.org), the European Union DataGrid Project (http://eu-datagrid.web.cern.ch), or the Access Grid (www.accessgrid.org), respectively.

To be specific about how the Grid adds services over the TCP/IP internet connectivity, let's briefly look at the EU Data Grid project. The software development is spread over five Work Groups:

- Work Scheduling
- Data Management
- Monitoring Services
- Fabric Management
- Storage Management

In addition, there is WG on Integration Testbed and Support and one on the Network itself. On top of this EU Grid infrastructure run very different disciplinary distributed data systems: Particle Physics, Earth Observation, and Biology. The EU Data Grid provides a good model for ORION planners to study in detail as they develop their own Grid infrastructure.

Grid middleware enables disparate resources to appear to the end user to have a coherent collection of capabilities. It implements a common set of services for applications to insulate them from the details of the specific host operating systems. Examples of these services include security, authentication, directories, file transfers, resource managers and brokers, resource discovery, reservation and scheduling services, and shared data services. One of the most successful efforts to build the software tools for developing Grids is the Globus project (see www.globus.org). The Globus developers have also recently worked with the community to develop an open architecture for Grid services (*i.e.*, the Open Grid Services Architecture or OGSA) that provides a framework for developing and deploying scalable Grid capabilities. OGSA has support from major computer vendors, particularly IBM, and is on track to become an international standard for Grid computing.

The systematic development of Grids and their uses to support large scale science project like ORION on have rapidly evolved over the last ten years. Two recent books will provide the interested reader with far more than they will want to know about the subject:

The Grid : Blueprint for a New Computing Infrastructure by Ian Foster (Author), Carl Kesselman (Second Edition) November 2003

Grid Computing: Making the Global Infrastructure a Reality edited by Fran Berman, Geoffrey Fox and Tony Hey. This is a book (over 1000 pages) published March 2003 by Wiley [Table of contents www.grid2002.org]

Today, dozens of Grid projects are being pursued in the US, Europe, and Asia Pacific Regions, and several efforts are underway to link these projects into an international Grid that can serve entire worldwide communities. The Global Grid Forum (www.gridforum.org) is an international organization created to coordinate the development of Grid standards to ensure that future Grids will be as interoperable and "software compatible" as the current Internet. Pacific Rim Applications and Grid Middleware Assembly (PRAGMA) [www.pragma-grid.net] is an open, international initiative to establish sustained collaborations and advance the use of the computational grid among a community of investigators at the leading research institutions around the Pacific Rim.

From Grids to Lambda Grids

One of the remaining problems to solve with traditional Grids is that one is not able to schedule the underlying "best effort" network. However, major technological and cost breakthroughs in networking technology over the past few years have made it possible to send multiple lambdas down a single length of user-owned optical fiber. (A "lambda," in networking parlance, is a fully dedicated wavelength of light in an optical network, capable of bandwidth speeds of 1–10Gbps.) Rather than being a bottleneck, metro and long-haul lambdas at 10Gbps are 100 times faster than 100T-base Fast Ethernet local area networks used by PCs in research laboratories. The exponential growth rate in bandwidth capacity over the past 10 years has surpassed even Moore's Law, due, in part, to the use of parallelism in network architectures.

Exploiting this new predictable lambda network enables a high-performance Grid called the LambdaGrid. In contrast with the existing shared Internet using the TCP/IP protocols, a LambdaGrid treats dedicated lambdas as schedulable resources. Thus, one will be able to reserve dedicated lightpaths among laboratories, data resources, instruments, and user access facilities using a Lambda Grid. Because the network resources are dedicated, one will be able to use specialized protocols not appropriate for the shared Internet. For example, the LambdaGrid will have protocols optimized specifically for moving large data objects over high bandwidth networks. The largest project researching LambdaGrids is the NSF- funded OptIPuter (www.optiputer.net). One aim of the OptIPuter project is to make interactive visualization of remote gigabyte data objects as easy as the Web makes manipulating megabyte-size data objects today.

To understand the quantitative advantage of using lambdas rather than the shared internet, it is instructive to return to NASA's EOSDIS. NASA measures the average file transfer rather between the various DAACs and end user sites. Typical performances are in the 10-50 Mbps range, even though the Internet2 backbone over which the files are

transferred is rated at 10 Gbps. Thus, the user is experiencing less than a percent of the available bandwidth. In contrast, recent OptIPuter experiments by Jason Leigh of UICs Electronic Visualization laboratory [www.evl.uic.edu/cavern/rg/20030817_he] showed that using alternative IP protocol, one can achieve over 93% of a national scale lambda with 10 Gbps bandwidth. This two order of magnitude improvement would have a qualitative effect on the science that an end user of a Data Grid could accomplish.

The NSF-funded OptIPuter research project exploits a new Net-Centric world in which the central architectural element is optical networking, not computers – creating "supernetworks". The OptIPuter is named for its use of Optical networking and Internet Protocol to link PC clusters optimized for computer, storage, and visualization. We think of it as a "virtual" parallel metacomputer in which the individual "processors" are widely distributed clusters; the backbone network is provided by IP delivered over multiple dedicated lambdas (each 1-10 Gbps); and, the "mass storage systems" are large distributed scientific data repositories, fed by scientific instruments as near-real-time peripheral devices or by supercomputer simulations. Individually controlled end-to-end lambdas, connecting clusters at a few key research laboratories to each other, as well as to remote instruments, high end computers, and data storage, can provide deterministic and guaranteed bandwidth to high-end users at multi-gigabit speeds.

The OptIPuter team has two e-science application drivers, the National Institutes of Health's Biomedical Informatics Research Network (described above), headquartered at UCSD's National Center for Microscopy Imaging Research, and the National Science Foundation's EarthScope, with local participants at the UCSD Scripps Institution of Oceanography (SIO), both of which involve many multi-gigabyte-sized individual data objects.

For a recent set of articles on this emerging world of Lambda networking, see the Special Issue on "Blueprint for the future of high-performance networking" of the Communications of the ACM, Volume 46 , Issue 11 (November 2003):

Introduction
Maxine D. Brown
Pages: 30 - 33

TransLight: a global-scale LambdaGrid for e-science
Tom DeFanti, Cees de Laat, Joe Mambretti, Kees Neggers, Bill St. Arnaud
Pages: 34 - 41

Transport protocols for high performance
Aaron Falk, Ted Faber, Joseph Bannister, Andrew Chien, Robert Grossman, Jason Leigh
Pages: 42 - 49

Data integration in a bandwidth-rich world
Ian Foster, Robert L. Grossman
Pages: 50 - 57

The OptIPuter
Larry L. Smarr, Andrew A. Chien, Tom DeFanti, Jason Leigh, Philip M. Papadopoulos
Pages: 58 - 67

Data-intensive e-science frontier research
Harvey B. Newman, Mark H. Ellisman, John A. Orcutt
Pages: 68 - 77

The OptIPuter as a model for an ORION Lambda Grid

Since UCSD's SIO is a key research partner with the OptIPuter, we have begun to focus our research onto how these new techniques could be used to support oceanographic Data Grids. In 2003, we have been rapidly expanding the SIO OptIPuter and customizing it to the need of ocean sciences. Assuming a number of collaborations being explored bear fruit, we may soon have an ideal prototyping laboratory for ORION using the latest in Lambda Grid technologies. This would allow unprecedented opportunities to manage and analyze massive ocean data sets, intimately coupling data with simulations. Furthermore, this performance could be accessed remotely in user's laboratories using standard Linux clusters configured appropriately.

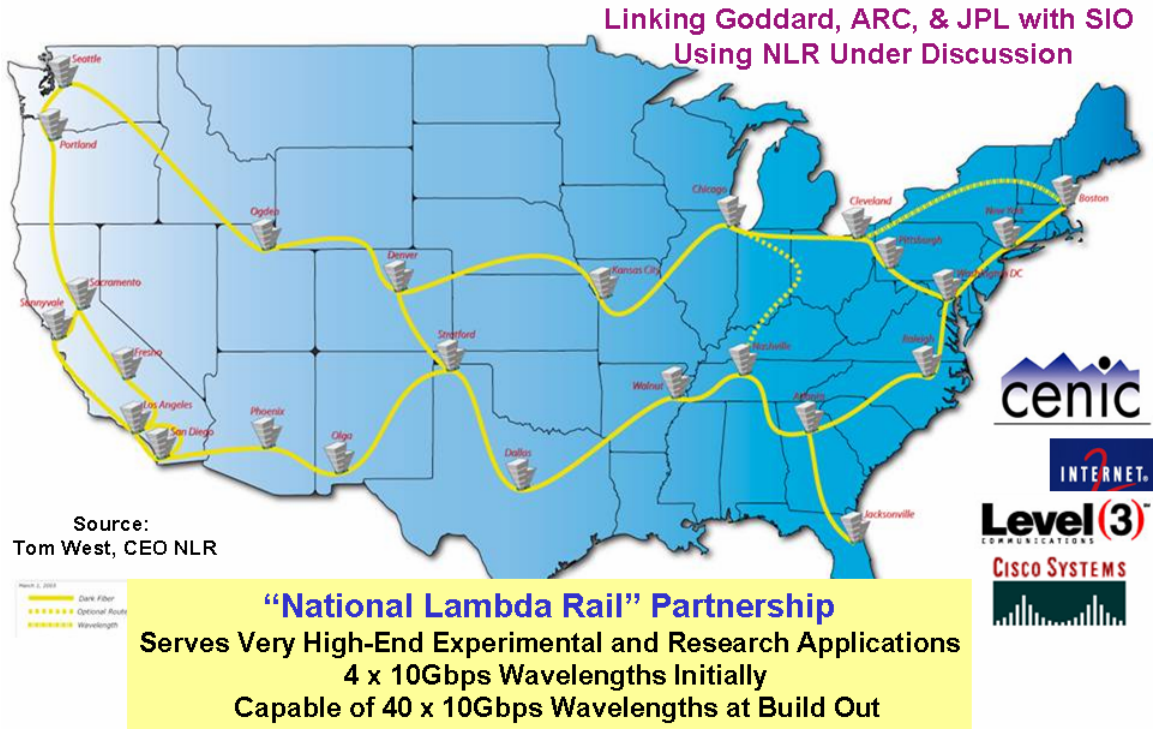
The SIO Center for Observations, Modeling, and Prediction (COMPAS) has a broad mandate to model, understand, and predict physical processes and phenomena in the ocean and atmosphere through numerical investigations in combination with SIO's historic strength in observations. The emphasis on integrating observations with modeling is a key part of our approach. The COMPAS cluster at SIO is a traditional compute-intensive Linux PC cluster using Myrinet interconnect and the NPACI Rocks software stack (http://rocks.npaci.edu/Rocks/). Parallel ocean and atmospheric models run on the cluster using a message passing programming paradigm.

The ocean simulation cluster has been recently linked to an advanced IBM OptIPuter visualization-intensive cluster located at SIO. This ten-node IBM graphics cluster can drive either two IBM T221 "Deep Vision" displays, each of which is capable of displaying 9-million pixels or the Geowall2, a 3x5 tiled PC LCD display—currently capable of visualizing 30-million pixel, expanding to 75-million pixels by the end of 2004. This high-throughput system will radically increase the COMPAS group's ability to visually analyze their simulations. They currently spend a large amount of time moving data, backing it up, and their tools for visualization consist largely of Matlab, Ferret, GMT, and other simple tools. We will use the IBM visualization cluster to explore novel time-varying volume visualization techniques for ocean sciences applications. The clusters with high-performance graphics cards (3D texture mapping for hardware-assisted voxel-based volume rendering) will allow us to use the OptIPuter to simultaneously apply multiple distributed volume rendering algorithms (voxel-, isosurface- and point-cloud-based) to present the same data in multiple ways, which is desirable because different transfer functions highlight different features in a data volume.

A critical component of the OptIPuter model is the requirement for *high-performance data systems* that can both ingest and feed data at multi-Gb/s. The OptIPuter counts on plentiful storage capacity and bandwidth richly interconnecting data into a nearly uniform resource. To achieve its objectives, the current mismatch between native storage data-transfer rates and the optical network must be directly addressed. IBM has just granted the SIO OptIPuter a 48-node 21TB storage-intensive cluster which will become the OptIPuter "capstone" of the existing IBM compute and visualization intensive clusters already at UCSD. We are also asking IBM for the necessary GigE cards to join the SIO ocean modeling cluster to the OptIPuter complex to serve as the compute-intensive node. This will provide a streaming source of 3D ocean simulation data objects which will be stored in the new IBM storage-intensive cluster and visualized on the IBM visualization intensive cluster.

All three types of clusters, compute-intensive, storage-intensive, and visualization-intensive will be joined by dedicated optical fibers routed through the Chiaro Networks [www.chiaro.com] high speed Enstara router. The fully-provisioned Enstara router, which uses novel internal optical phased array switching, provides essentially "unlimited" packet forwarding capability with over 6 Terabits/s of capacity; this means that thousands of 1-GigE interfaces and hundreds of 10-GigE interfaces can be routed at full line speed.

To provide a source of live data streams from multiple instrument types, the OptIPuter project through SIO will explore connecting to ocean observatories off the U.S. west coast such as the Canadian/U.S. Neptune (http://www.neptune.washington.edu/) set of instruments or to the NSF funded prototype Monterey accelerated research system (MARS) hosted by the Monterey Bay Aquarium Research Institute (MBARI) [www.mbari.org/rd/projects/2002/moos/702244_mars.html]. The dedicated optical connection to the Local OptIPuter Lambda Grid complex will be enabled through CENIC (http://www.cenic.org/) or its extension to Washington State called the Pacific Light Rail.



Discussions at high level with NASA are underway to extend the OptIPuter project to link ocean simulations groups running ECCO at SIO, JPL, Ames Research Center, and to Goddard Space Flight Center, using the National Lambda Rail (NLR-see www.nationallambdarail.org) set of lambdas. The NLR is a major initiative of U.S. research universities and private sector technology companies to provide a national scale infrastructure for research and experimentation in networking technologies and applications. Goddard would provide OptIPuter access to Earth Science satellites archives through the EOSDIS, which could be integrated with in situ ocean measurement and simulations.

Thus within 1-2 years a complete national-scale Lambda Grid customized to Ocean Observing could be set up. This would include in situ data streams, earth satellite data sets, ocean assimilations and simulation codes, all supported by massive computing, storage, and visualization capabilities. This possibility will be explored with our colleagues at the ORION workshop in January 2004 in Puerto Rico.