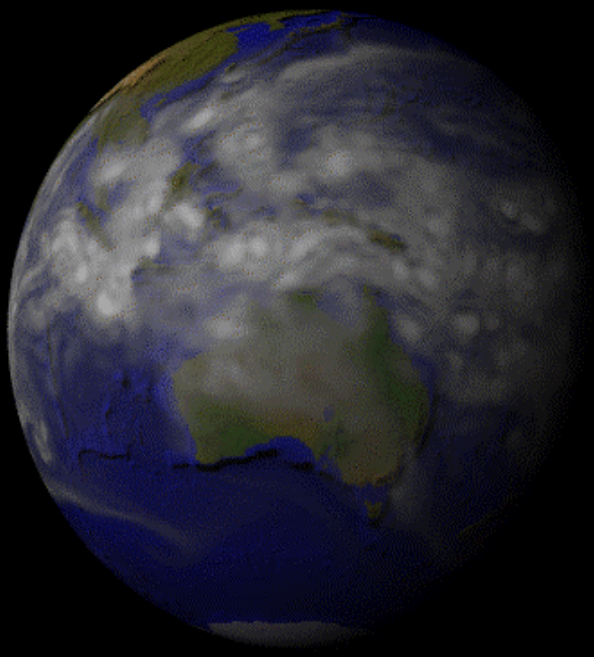
Prototyping an Earth System Grid



DOE Next Generation Internet: Applications, Network Technology, and Network Testbed Partnerships Program

Response to Program Notice 99-09

March 31, 1999

Principal Investigators:

Steve Hammond,	National Center for Atmospheric Research PO Box 3000 Boulder, CO 80307
Dean Williams,	Lawrence Livermore National Laboratory Mail Stop: L-264 7000 East Ave., P.O. Box 808 Livermore, CA 94550

Co-principal Investigators:

Wu Feng,	Los Alamos National Laboratory
Ian Foster,	Argonne National Laboratory
Bill Hibbard,	University of Wisconsin
Carl Kesselman,	USC Information Sciences Institute
Arie Shoshani,	Lawrence Berkeley National Laboratory
Brian Tierney,	Lawrence Berkeley National Laboratory

Table of Contents

Executive Summary

The need to evaluate climate change scenarios under the Kyoto accord makes climate modeling a mission critical application area. DOE's Accelerated Climate Prediction Initiative (ACPI) seeks to address this need through the creation of an advanced climate simulation program, which will accelerate the execution of climate models one hundred-fold by 2005 relative to the execution rate of today's climate models [ACPI98]. High-resolution, long-duration simulations performed under ACPI will produce tens of petabytes of output. The output in turn will be made available to global change impacts researchers nationwide through a network of diagnostic and regional climate centers [ACPI98, GATE99]. These distributed centers, users, models, and data will be connected in a virtual collaborative environment called the Earth System Grid (ESG). The Earth System Grid will provide scientists with virtual proximity to the distributed data and resources comprising this collaborative environment.

Creating an effective and efficient ESG in support of ACPI is challenging at multiple levels, but *above all it is a Next Generation Internet (NGI) problem.* A large community of global change researchers at laboratories and universities around the nation will need to access significant fractions of the data. User requests for data products will be translated into appropriate combinations of accesses to data caches, requests to central data archives, and new large-scale simulations. The effective management of the required data movement operations will tax even the highest performance and most advanced networks.

We propose a research and development project that will take a first step towards the creation of an Earth System Grid. Specifically, we propose to prototype a system that will support:

- The rapid transport of climate data between centers and users in response to user requests. The focus is on end-user accessibility and ease of use, which will be accomplished through both the modification of existing applications and tools and the development of new tools as needed to operate seamlessly in the Earth System Grid.
- Integrated middleware and network mechanisms that broker and manage high-speed, secure, and reliable access to data and other resources in a wide area system.
- A persistent Earth System Grid testbed that provides virtual proximity and demonstrates reliable high-performance data transport across a heterogeneous environment.

In constructing this system, we will build on a substantial base of software and experience that includes parallel climate models, high-performance networking, climate model analysis tools, and advanced networked middleware. We also leverage substantial existing investments in supercomputers, servers, mass storage systems, and high-speed networking.

The proposed research and development activities will be performed by a partnership between four DOE Laboratories (ANL, LANL, LBNL, LLNL), a NSF center (NCAR), and two universities (Wisconsin, USC). This uniquely qualified team--most of whom have worked together closely over many years--includes experts in applications, middleware, and networking. Working together, this team will construct an outstanding driver and showcase for DOE NGI research and networks. In addition, climate research will also be accelerated by making it possible for the research community to readily access distributed computers, simulation models, and data for scientific discovery.

1. Background and Motivation

Climate modeling is unique among the Grand Challenge problems in terms of the length of simulations and volume of model output produced. The reason for this is physical in origin: the global climate-change "signal" or "signature" must be discernable above the "noise" of the natural variability of the climate system. Variability in the atmosphere alone has fairly short periods: weeks to months. However, the ocean is quite different. Paleoclimate data collected from Greenland and Antarctic ice cores tell us that profound fluctuations in climate have occurred on time-scales of tens to hundreds of years; these are generally attributed to shifts in the ocean circulation, especially the Gulf Stream. The ocean also has modes of variability with periods extending to hundreds or thousands of years, so long, in fact, that they have only been explored with rather crude models. As a result, very long-duration runs accompanied by frequent output of very large files are needed to analyze the simulated climate variability, which must then be compared with what is known about the observed variability.

1.1. Data Volume and Transmittal Problems

As we enter the era of teraflop computing systems, our ability to generate model output is in danger of outpacing our ability to archive it and to transport it from site to site. As an example, running a high-resolution ocean model on present-day computers with peak speeds in the 100-gigaflop range would generate a dozen multi-gigabyte files in a few hours at an average rate of about 2 MB/second. Computing a century of simulated time would take more than a month to complete and would produce about 10 TB of output to be archived. Archival systems capable of storing hundreds of terabytes are required to support calculations of this scale on a regular basis. Moving to a one-teraflop system could multiply each of the figures above ten-fold, making petabyte archives essential. The ACPI mentioned above, which is being proposed by DOE, is targeting a 5-teraflop machine for FY00 and a 40-teraflop machine in FY03. Given this magnitude of model output, how can it be made available to the research and climate-impacts communities?

Four modes of distribution can be envisioned that place very different demands on resources at the host site and on the network:

1. Real-time transfer of raw model output as it is being generated by the model.
2. Post-run transfer of raw model output.
3. Post-run transfer of pre-selected reduced datasets.
4. Post-run generation, transfer and/or visualization of user-selected reduced datasets.

Characteristics of these four modes are summarized in Table 1.

Table 1

Characteristics	1. Real-time raw output transfer	2. Post-run raw output transfer	3. Pre-selected data on caches	4. User-selected data processing
Available data	All (at client site)	All (at client site)	Limited	All (at host site)
Transfer rate off-site	> Production[3] rate: Low	Medium[1,2]	High	= Processing rate[4]: Medium[1,2]
Access host archive[1]	No	Yes	No	Yes
Firewall on host[2]	Yes	Yes	No	Yes
Sites served this way	One (doing run)	Any with archiving	All	All
Client archive	Required	Required	Optional	Optional
Processing	At client site	At client site	Pre-processed	On-line at host
Software tools	Client's	Client's	Client's or host's	Provided on host

[1]It is assumed that archival storage means tape storage, not a fast data cache.

[2]It is assumed that only refreshable data caches are exposed and that computers and archival storage devices are located behind firewalls. Performance is degraded if either the source or destination device is behind a firewall.

[3]Production rate is the average rate at which the model generates output data as it runs. The transfer rate must exceed this rate on average; otherwise, the model output will have to go to archival storage or the model run will have to be suspended until the network can catch up.

[4]Processing rate is the average rate at which the datasets needed to fulfill the user's requests can be downloaded from the site's archival storage and processed. Retrieval from archival storage is likely to be the rate-limiting step.

Modes 3 and 4 will be the most commonly used, especially by customers of the regional centers mentioned above. Mode 2 use will be limited to facilities having large archival storage systems. Mode 4, which is a major focus of this proposal, makes all output potentially available to the user with only very modest facilities. Mode 1 is likely to be used only by a person or team that is running its own model at a remote site and wants all the output to be available and stored locally as soon as possible. Consequently, it will probably be the least used mode. Nevertheless, comparing the "low" production rate (defined in footnote 3 of Table 1) with local and remote network technologies provides an interesting perspective, as shown in Table 2.

Table 2

Time frame	1993-97	1999	2000-01
Computer speed, peak	~ 100 GF	1 TF	5 TF
Production rate (MB/s)	~ 2	~ 20	~ 100
Local network	HIPPI-800 (100 MB/s), 100 Mb/s Ethernet, 100 Mb/s FDDI	HIPPI-800, Gigabit Ethernet, 100 Mb/s Ethernet, 100 Mb/s FDDI, Myrinet	HIPPI-6400 (800 MB/s), Gigabit Ethernet, 100 Mb/s Ethernet, 100 Mb/s FDDI, Myrinet
Inter-site network	OC-3	OC-12	OC-12
Peak rate (MB/s)	19	78	78
In practice (MB/s)	~ 0.3-3	~ 50[1]	~ 50[1]

[1]See section 2.4.

As the table indicates, even the seemingly low production rate is still in the range of network peak speeds and may pose problems as computer speeds increase (doubling on average every 1½

years), particularly in view of the fact that networks are shared resources. Thus, there is a time-critical need for the development of the capabilities and functionality proposed here. The more demanding requirements associated with modes 2-4 in Table 1 make this need even more pressing.

The following example illustrates why increased network performance is vital to progress in climate science. An atmospheric scientist recently published several papers analyzing the intraseasonal oscillation among fifteen coarse grid atmospheric general circulation models. The study tested the ability of climate models to reproduce one of the fundamental features of the tropical circulation. The simulations consisted of atmosphere-only 10-year simulations with data sampled at six hour intervals. Models were run at fifteen different institutions and data transferred using a combination of Mode 2 and Mode 4 from Table 1. As a first step, he extracted the wind components, outgoing longwave radiation, latent heat flux, and sea surface temperature. The extraction of the data took several months (part electronic transfer and part mailing data tapes), the calculation of velocity potential took weeks, the regridding, filtering, and the principal component analysis took weeks. Much of the time was consumed by moving data, extracting the desired fields, and organizing the data for analysis. To gain appreciation of the magnitude of future model simulations, imagine that the models were run at a high resolution, the simulations were hundreds of years in duration, and there were scores of ensembles. With existing network and data storage access capabilities, this kind of research will take years.

1.2. A Vision for the Future

We believe that the tremendous scale of future climate datasets, when combined with the ambitious scientific goals of the Accelerated Climate Prediction Initiative [ACPI98,GATE99], requires new approaches to scientific investigation. Because the datasets of interest are too large to be replicated at every site, a new shared infrastructure will have to be created to support effective access to data and computing by a large community. Because the problems to be addressed are so complex, this community will also create increasingly complex multi-institutional collaborations. Hence, we anticipate the creation of an *Earth System Grid*, a new infrastructure designed to support virtual collaboration, distance computing, remote data access, and distributed applications [CS92,FK99a,SWDC97].

The availability of a usable Earth Systems Grid will fundamentally change and enhance the way climate research scientists work together and address major challenges in climate simulation. Scientists will be able to request complex data products via convenient "Grid-enabled" desktop tools. The Grid will determine where these data products should be computed, stored, post-processed and visualized, using the most effective combinations of local and remote resources. Once created, a scientist will be able to discuss new results with colleagues, regardless of their location, and record observations in electronic logbooks.

Many components are necessary to create such a distributed, collaborative infrastructure and they include:

- Teraflop computers to run highly sophisticated atmospheric, ocean, and coupled atmospheric ocean models;
- Data storage facilities that can hold petabytes of raw and processed climate data with fast access times;
- Networking hardware that allow information to be transmitted at high speeds for collaborative audio and visual communication, white board interaction, visualization, and animation; and

- Powerful, cost-effective local and desktop facilities for visualization and data processing.

Tying these diverse components together in an easily interconnected and comprehensible way is the integrating network infrastructure and middleware software: that is, the services such as security and resource discovery that sit between application and network, and that help researchers navigate and manage a secure system. Experience with similar systems [CS92,DS99,FK99a,Jon99,MBMR99,Mes99] suggests that the development and effective use of this network infrastructure and middleware software is a major challenge, which if not addressed has the potential to significantly slow progress towards the realization of the goals of the Accelerated Climate Prediction Initiative.

1.3. Our Proposed Approach

The creation of a complete Earth System Grid is a major undertaking. However, we believe that we can make significant progress in a short time, producing technologies that will be immediately useful to climate scientists and at the same time contributing to the state of the art in networking and middleware. These goals are achievable because we tackle them in a coordinated fashion. There are two well-defined, central problems and we bring to bear a world-class team of researchers and technologists with expertise in climate applications, climate data analysis, advanced middleware, and networks. Briefly, the two problems are the following:

1. *The high-speed movement, caching, and location of large datasets*, as required for the transfer of climate model data from a central computing site to remote archives or users. We propose to focus on the problem of high-speed data transfer in heterogeneous Grid environments. Our proposed work includes the construction of a multi-site database, where the most commonly used data is replicated at multiple sites. A user interface that allows the end-user to transparently access both raw datasets and derived products wherever they reside will also be developed.
2. *Flexible on-demand discovery and scheduling of caches, networks, and computers* in order to meet user-level requests (issued from Grid-aware analysis tools) for data products. Here, the application focus is the efficient decomposition of data analysis tasks across distributed systems; technology developments will include remote execution, security, cost estimation, and resource management mechanisms.

Building on a substantial base of existing code and expertise (DPSS, Globus, STACS, PCMDI* software etc.), we will develop software that incorporates solutions to these problems, and will use this software to deliver solutions of immediate value to climate scientists. These solutions include rapid dissemination of climate data products from central sites and the development of Grid-enabled climate data analysis tools that provide seamless access to remote data and computing. Software will be deployed, evaluated, and demonstrated on a substantial Earth System Grid prototype, comprising computer and storage resources at ANL, LANL, LBNL, LLNL, and NCAR (and later U. of Wisconsin and USC, as prototype university partners) and connected by ESnet, NTON, vBNS, and MREN. Technology innovations resulting from this project will also be propagated to the larger community via the already very successful code distribution systems that are in place for Globus, DPSS, and PCMDI software.

* the Program for Climate Model Diagnosis and Intercomparison located at the Lawrence Livermore National Laboratory in Livermore, California

2. Previous Work and Preliminary Studies

We review briefly our relevant previous work on climate data analysis tools, middleware services, and networking technologies.

2.1. The PCMDI Software System

PCMDI has over the past several years developed a sophisticated suite of climate data analysis tools [W97] designed to simplify the storage, diagnosis, and visualization of climate data. The tools have played and continue to play a crucial role in the development, testing, validation, and intercomparison of global climate models. They are comprised of three parts:

- The Climate Data Analysis Tool (CDAT) [Wet.al.99] manipulates data and provides climate scientists with diagnostic, statistical, and regridding routines;
- The Climate Data Management System (CDMS) [D99] automatically locates and extracts metadata (for example, variables, dimensions, grids, attributes, etc.) from a collection of model runs and analysis files; and
- The Visualization and Computation System (VCS) [W96] displays, animates, and manipulates scientific data.

Each software product is independent and can run as a stand-alone process or together as part of a single process. Scientific routines written in C, C++, or Fortran are easily integrated and controlled with the embedded Python scripting language [LA99]. Python is an open-source object-oriented scripting language and users find it to be simple to use, dynamic, powerful, and flexible. Serving as an excellent "glue" or "steering" language, Python has the power to integrate unrelated software packages under a single system. This modular structure will be important as we extend the framework to provide a Grid-enabled data analysis system.

Our basic approach using Python steering is also being used by the ASCI code "Kull", a new national tokamak modeling effort, a French national computational fluid dynamics code called "eisA", and many other smaller efforts. We therefore are leveraging our infrastructure efforts with a national and international community of contributors. Figure 1 shows a subset of the PCMDI Software System concept.

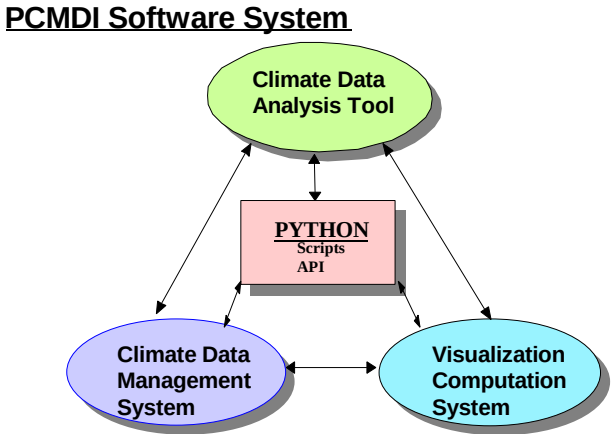


Figure 1. Conceptual view of PCMDI's Software System.

2.2. Storage Management of Very Large Scientific Datasets

LBNL has for a long time been involved in developing data management technology for scientific applications. More recently, two different projects focused on the research and development for scientific applications that generate very large quantities of data. The volumes of data may reach hundreds of terabytes per year that need to be stored on robotic tape systems. In particular, the OPTIMASS project developed the technology to reorganize and access spatio-temporal data from robotic tape systems [CDK+95a, CDK+95b]. Another project, called the HENP-Grand Challenge, developed the technology and software to manage terabytes of High Energy and Nuclear Physics data and their movement to a shared disk cache [SBN+98, SBN+99, BNR+98]. This system, called STACS (Storage Access Coordination System) was integrated into the production analysis system for the STAR and PHENIX experiments at Brookhaven National Laboratory.

STACS [STA99] has three main components that represent its three functions: 1) The Query Estimator (QE) uses the index to determine what files and what chunks in each file (called "events" in physics applications) are needed to satisfy a given range query. 2) The Query Monitor (QM) keeps track of what queries are executing at any time, what files are cached on behalf of each query, what files are not in use but are still in cache, and what files still need to be cached. The Query Monitor consults an additional module, called the Caching Policy module, that determines what file to cache next according to the policies selected by the system administrator. 3) The Cache Manager (CM) is responsible for interfacing to the mass storage system (HPSS) to perform all the actions of staging files to and purging files from the disk cache. In this project, we plan to apply the QM and CM modules in each node on the network. The CM module will be extended so that the Cache Managers communicate with each other to request files that do not have locally, and perform the data transfer.

2.3. Globus Grid Middleware Services

The Argonne/ISI Globus project [FK98,FK99b] has developed a range of basic services designed to support efficient and effective computation in Grid environments [BFKTT99, CFK99, CFNK+98, FFKL+97, FGKT97, FKL+99, FKTT98]. These services have been deployed in a large multi-institutional testbed that spans some 40 sites worldwide and has been used in numerous application projects ranging from collaborative design to distributed supercomputing (e.g., [BCFF+98, LFIB+99]). Services relevant to the current project include:

- The Grid Security Infrastructure (GSI) [FKTT98], which provides public key-based single sign-on, run anywhere capabilities for multi-site environments, supporting proxy credentials, interoperability with local security mechanisms, local control over access, and delegation.
- The Metacomputing Directory Service (MDS) [FFKL+97], which provides a uniform representation of, and access to, information about the structure and state of Grid resources, including computers, networks, and software. Associated discovery mechanisms support the automatic determination of this information.
- Globus resource management services [CFNK+98, CFK99, FKL+99, HJFR98], which provide uniform resource allocation, object creation, computation management, and co-allocation mechanisms for diverse resource types.
- The Global Access to Secondary Storage (GASS) service [BFKTT99], which provides a uniform name space (via URLs) and access mechanisms for files accessed via different protocols and stored in diverse storage system types (HTTP and FTP are currently supported, HPSS and DPSS are under development.

2.4. Networking

2.4.1. Current performance & analysis over ESnet and vBNS from LANL to NCAR.

Because the current locus for large-scale DOE-funded climate simulations is the LANL parallel computing facility, data transfer rates from LANL to NCAR and other sites are of considerable concern. Unfortunately, preliminary tests show that we cannot achieve better than 12 Mb/s for a single TCP stream session between LANL and NCAR despite the use of large 1-MB window sizes, even though the limiting network technology between LANL and NCAR is a 100-Mb/s FDDI network. Thus, the end-user application only sees 12% of the available FDDI bandwidth. This utilization is appalling in comparison to the utilization seen in the China Clipper Project over an ATM-only network (see Section 2.4.2), which is around 75% between LBNL and SLAC. If this proposal is funded, a high priority effort will be made to isolate the networking bottlenecks between LANL and NCAR and to leverage the lessons learned from the China Clipper Project (see below) in order to bring the bandwidth-utilization numbers up to acceptable levels.

Unfortunately, the ESnet and vBNS networks are controlled by independent Internet Service Providers (ISPs), which may make it difficult to isolate network bottlenecks, sources of packet loss, the presence (or absence) of flow- and congestion-control mechanisms, etc. For example, experiments show that cells are being lost between the Cisco 7507 and FORE ASX-1000 switch interfaces used in vBNS nodes. In addition, many packets are being dropped at the Chicago vBNS node due to simple overload. Both of these problems could adversely affect the effective throughput from application to application.

2.4.2. China Clipper Project

China Clipper is a joint project with ANL, LBNL, and Stanford Linear Accelerator Center (SLAC) which is focused on developing technologies required for widely distributed data-intensive applications. It was applied to the analysis of high-energy physics (HEP) applications. Clipper leveraged a distributed parallel storage system (DPSS), Globus, and ESnet and NTON (OC-12 networks). LBNL has demonstrated a throughput of 57 MB/s from disk storage (four OC-3 servers DPSS, sending to one OC-12 client, tuned with NetLogger) to a remote physics application over a wide area network. This is an order of magnitude faster than has been achievable in other comparable networks, and demonstrates that wide area data-intensive computing is feasible. In aggregate, this was equivalent to moving 4.5 TB/day.

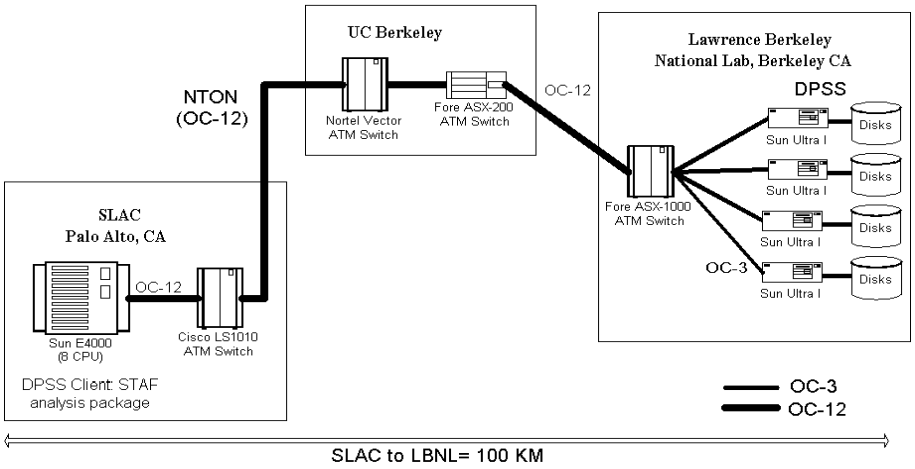


Figure 2. Schematic of China Clipper Project

Recently, LBNL has been testing applications on the new OC-12 connection between LBNL and ANL, an IP-routed high-speed WAN. After a considerable amount of testing and tuning of both the network and the applications, LBNL staff achieved application-to-application throughput of 38 MB/s on this network. Similar types of testing and tuning will be required for any NGI application, and we have developed the necessary tools and skills.

2.4.3. High Performance Network Cache

The China Clipper data architecture uses a high-speed distributed cache as a common element for all of the sources and sinks of data involved in high-performance data systems. This cache-based approach provides standard interfaces to a large, application-oriented, distributed, on-line, transient storage system. Each data source deposits its data in the cache, and each data consumer takes data from the cache, usually writing the processed data back to the cache. The high-speed cache provides a standard high data rate interface for high-speed access by data sources, processing resources, mass storage systems, and user interface elements. It provides the functionality of a single very large, random access, block-oriented I/O device (i.e., a "virtual disk"). It serves to isolate the application from tertiary storage systems and instrument data sources.

Our cache system is called the Distributed-Parallel Storage System [DPSS]. DPSS is a data block server, which provides high-performance data handling and architecture for building high-performance storage systems from low-cost commodity hardware components. This technology has been quite successful in providing an economical, high-performance, widely distributed, and highly scalable architecture for caching large amounts of data that can potentially be used by many different users. Recent work on the DPSS has been to make it "network aware", automatically performing TCP buffer tuning and load balancing based on current network conditions.

2.4.4. Performance Analysis

LBNL has developed a methodology that enables real-time diagnosis of performance problems in complex high-performance distributed systems, called the NetLogger Toolkit [TJC98]. NetLogger includes tools for generating precision event logs that can be used to provide detailed end-to-end application and system level monitoring, and tools for visualizing log data to view the state of the distributed system in real time. NetLogger has proven to be invaluable for diagnosing problems in networks and in distributed systems code. This approach is novel in that it combines network, host, and application-level monitoring, providing a complete view of the entire system. NetLogger monitoring allows us to identify hardware and software problems, and to react dynamically to changes in the system.

3. Technical Approach

We propose to develop a prototype Earth System Grid that will provide scientists with virtual proximity to the distributed data and resources comprising this collaborative environment. Building on a substantial base of existing code and expertise, we will develop a software system that provides rapid dissemination of climate data products from central sites and the development of Grid-enabled climate data analysis tools that provide seamless access to remote data and computing. Software will be deployed, evaluated, and demonstrated on a substantial Earth System Grid prototype, comprising computer and storage resources at ANL, LANL, LBNL, LLNL, and NCAR (and later ISI and Wisconsin, as prototype university partners) and connected by ESnet, NTON, vBNS, and MREN.

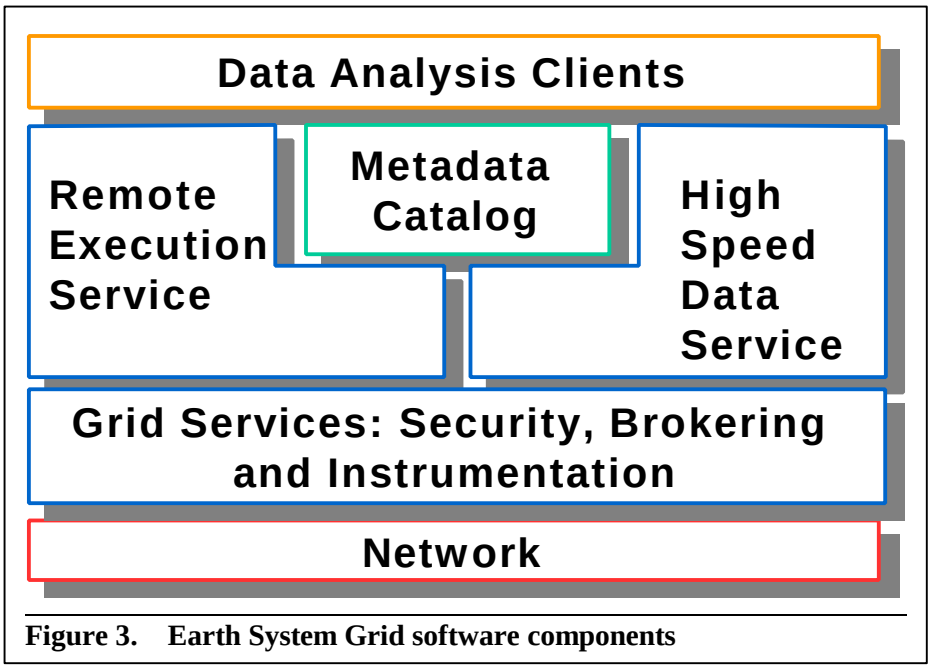


Figure 3. Earth System Grid software components

To this end, we will develop, integrate, deploy and apply six technology components, as illustrated in Figure 3:

1. A set of *Grid-enabled data analysis clients*, which support user-level access to the Earth System Grid and serve as a driver for the various middleware and network services to be developed in the project;
2. A *distributed cache management service*, which supports the replication, discovery, and caching of datasets in the Earth System Grid;
3. A *high-performance data transfer service*, which will be used by both data analysis clients and the data management service for the reliable, high-performance transfer of large datasets;
4. A *remote execution service*, which will be used by data analysis clients to execute analysis components on remote computers;
5. A set of *low-level Grid services*, supporting such common functions as security, brokering, and instrumentation; and
6. A set of *enhanced network services*, which will support high-speed end-to-end data transport, enable network-aware applications, and provide quality of service (QoS) services, including bandwidth reservation, over multiple networks.

The distributed cache management service, high-performance data transfer service, and remote execution service together constitute a set of *intelligent middleware* which will be used by the Grid-enabled data analysis clients to enable flexible user-level access to the Earth System Grid. The low-level Grid services and enhanced network services enable the intelligent middleware to operate effectively in a heterogeneous, high-performance Grid environment.

3.1. Distributed data analysis clients

Most climate data analysis software in use today runs on a user's local machine and local area network and requires no external resource other than what is provided by the client machine and/or the local area network to acquire, process and examine data. Unfortunately, this mode of use leaves the user confined and isolated from other external resources and is certainly not

appropriate for tomorrow's petabyte datasets. Instead, we require distributed or *Grid-enabled* data analysis client software that allows the user to access, compute, and visualize data at a distance. These Grid-enabled data analysis clients will increase the researcher's productivity and organizational value by allowing remote user the ability to access important data boundaries and collaborate with colleagues as if they were next door (see Figure 4). However, substantially new structures and middleware services are required before we can decompose and distribute the various components of the data-access-analysis-visualization "corridor"

In this project, we propose to adopt the PCMDI Software System [W97] (see Figure 5) as our first Grid-enabled data analysis client. We choose this system for two primary reasons. First, its large user base and familiarity to many climate scientists will facilitate adoption by users. Second,. Its modular Python-based structure [LA99] will facilitate integration with our proposed intelligent middleware services. In addition, we note that its structure is common to a number of other systems and that other Python efforts (for example, the "Kull" ACSI code) will be able to leverage off the work proposed here.

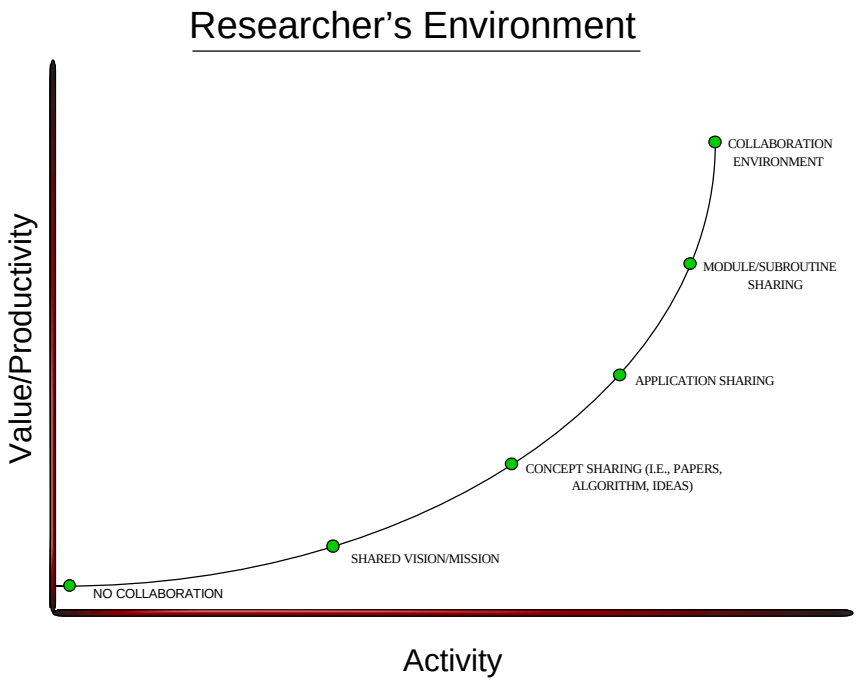


Figure 4. Researcher's Productivity chart.

The PCMDI Software System was designed from the start with a view to taking advantage of future intelligent middleware and providing users with the ability to access remote resources. Utilizing Python's scripting language and API capabilities, we propose to modify the current PCMDI Software System to provide climate research scientists with distributed data management, computing, and visualization from client-side, server-side, or in combination. (see Figure 5.)

Having a well defined intelligent middleware API for the PCMDI Software System will make it easier to incorporate future data analysis clients into this distributed system by either interfacing the data analysis client with Python or by integrating the intelligent middleware API directly into the data analysis client's source code. The first such challenge will be to integrate the VisAD tool into the system. VisAD [VIZAD] is a Java class library for interactive and collaborative visualization and analysis of numerical data. It uses Java3D and Java2D for visualization, Java

RMI for distributed objects and Java JNI for connections to legacy science codes. The system is freely available, including source code and documentation.

Initially, starting with the PCMDI Software System, we will develop a testbed that supports communication with resources at DOE sites and NCAR. Using this testbed, a typical scenario may be to access data at a remote DOE site, post-process the data at another DOE site, and visualize it at NCAR. Another scenario may be to extract and process data locally, but visualize the data remotely. The development of this system will allow the climate community to manage the expected large amount of climate modeling data and process it in a distributed fashion, combining both local and remote testbed facilities into an easy-to-use computational facility.

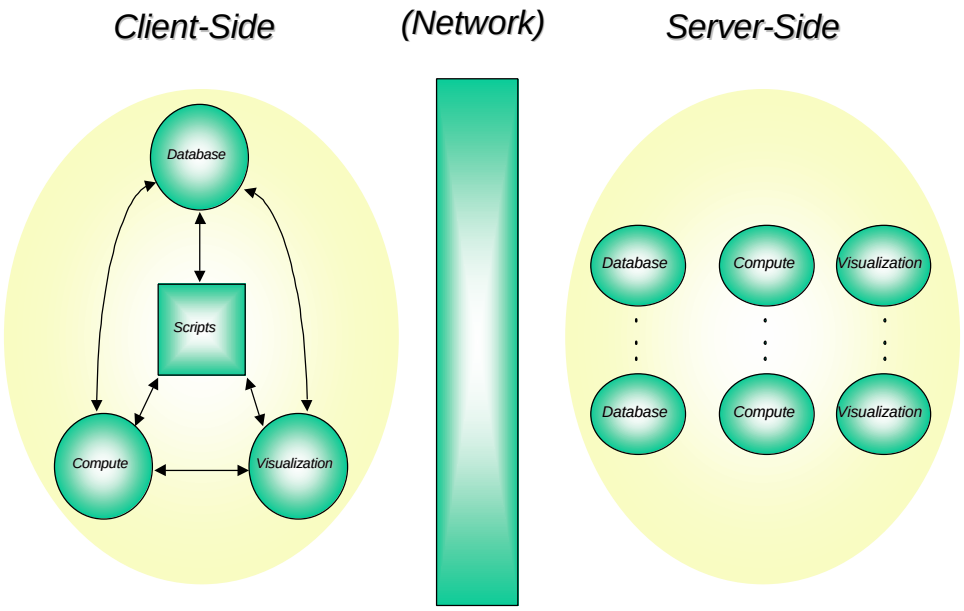


Figure 5. Distributed PCMDI Software System.

Since the characteristics of the available hardware will determine the usefulness of various software approaches, we will strive for flexibility in our approach so that we can adapt our plans to available hardware and software and to adjust actual performance of the system. Our ultimate deliverable is a powerful yet easy to use system that will support climate research scientists in reaching their goals. We believe we can reach this goal in a strongly defined incremental approach that will continuously provide improved tools and an improved computational environment to a widening audience within the climate community. PCMDI has the first part of the infrastructure in beta (that is, the existing PCMDI Software System). The next phase of the development will enable access to remote, and local, database servers. We will then develop the ability to combine remote and local access of the compute and visualization servers and add other data analysis clients to the system. Finally, we will add audio and visual communication and white board interaction among climate researchers.

3.2. Distributed Data Management Service

Applications that require massive amounts of data to be transferred across the network typically have much of the data residing on distributed disk caches and robotic tape systems. This is the case with climate data that are generated by high-resolution simulations. In performing analysis of such data over a network, the end-to-end performance depends a great deal on the efficiency with which data can be accessed data from remote caches and especially from the relatively slow robotic tape systems. Even if the speed of the network is high, the bottleneck is now the speed of reading the data from tapes to a disk cache use by the end-user application. This problem is compounded when there are tens to hundreds of users accessing the data simultaneously.

The typical method for analyzing and viewing climate data is by first generating data products. At PCMDI, the (raw) datasets from 34 different simulation models are stored, and multiple data products are extracted. A data product typically extracts one of 70 variables, such as SST (Sea-Surface Temperature) for a certain spatial region and a certain period of time. There is a "Metadata Catalog" that keeps track of the content of each data product as well as its location. When a request for data is made, the Metadata Catalog is consulted, and if using one of the data products can satisfy the request, the requested data is extracted from that data product. Otherwise, it is extracted from the original dataset, and the new data product can then be added to the system, and the Metadata Catalog gets updated. For the first year of this project, it is expected that all the data will fit into a large cache (2 TB cache will suffice). However, as the resolution of the simulations is increased in later years, it is expected that the original (raw) datasets, as well as many of the data products will reside on tape.

Our goal is to extend this technology to a distributed system of caches and tape systems. In a distributed system the disk cache used by the analysis programs may be at a different physical location than the system that holds the data. In addition, there may be caches of varying sizes at multiple locations. Furthermore, the network speed to the various caches may vary. To address this problem, we plan to use a distributed Metadata Catalog and a distributed "Cache Management" system.

We describe next how this distributed system will work. When an analysis program at some node makes a request, it is passed to a local agent called a "Query Monitor". It consults its local Metadata Catalog, to see what data is in its local cache. For the components that are not in the local cache, the Query Monitor broadcasts the request to all other Metadata Catalogs. Each responds with the information on what they have. The Query Monitor consults the Globus Network Services as to the current state of the network, and makes a decision from where to get the individual pieces of data. It then instructs the local Cache Manager to get the data from the various remote caches. The Cache Manager initiates data transfers from the remote caches in parallel.

The above description is the plan for the first year based on the assumption that the data will reside in disk cache only. For later years, the same design will apply, but an additional component will be added that will be responsible for moving files from tape to disk cache upon request. Cost estimations will then reflect the extra time to access data from tape if necessary.

We base our approach on the experience we have with the Storage Access Coordination System (STACS) that we developed for the GC_HENP project at LBNL, currently being deployed at Brookhaven National Laboratory [SBN98, SBN99, STA99]. We also developed a research prototype for climate modeling (the project was called OPTIMASS (for Optimization of Mass Storage access), which focused on the optimal organization of climate data on tape systems [CDK+95a, CDK+95b]. The main concept is to take advantage of the knowledge of a query specification (consisting of a region in space, and time periods, as well as the desired variables) from which the entire set of data points needed for an analysis was determined. In STACS only a

single tertiary storage system and a single cache was used. To extend this technology to a distributed system with multiple disk caches, and varying network speeds, the following plan is proposed:

- The Metadata Catalog developed at LLNL will be extended to respond to query requests coming from any node on the network. Given such a request, the local Metadata Catalog will respond with what parts of the data requested resides in local cache.
- The STACS Query Monitor module developed at LBNL will be extended to support the coordination of data requests from multiple nodes, by consulting remote Metadata Catalogs.
- The Query Monitor will get network cost estimates from the Globus system.
- The STACS Cache Manager module developed at LBNL will be extended to communicate with remote Cache Managers to achieve parallel data transfers into the local cache.

In future years, we will develop additional features that support caching from tape systems, as well as developing predictive pre-caching techniques. Specifically, the two aspects will be developed:

- Automatic migration from a local tape system to a local cache based on current usage. A simple LRU (Least Recently Used) algorithm will be used, but hints about expected usage will be used as well.
- Automatic migration and replication of "hot" data products according to current use. This will require the use of caching policies that optimize what should reside in disk caches in various nodes of the distributed system.
- A distributed monitoring module will be developed to measure the system performance and identify bottlenecks. The goal is to determine how to balance the distributed system as a whole. An additional goal is to have the data close to the application program, but at the same time have all the resources used efficiently.

Figure 6 shows a schematic for the ESG. In the initial phase, all data will reside on disk, but we

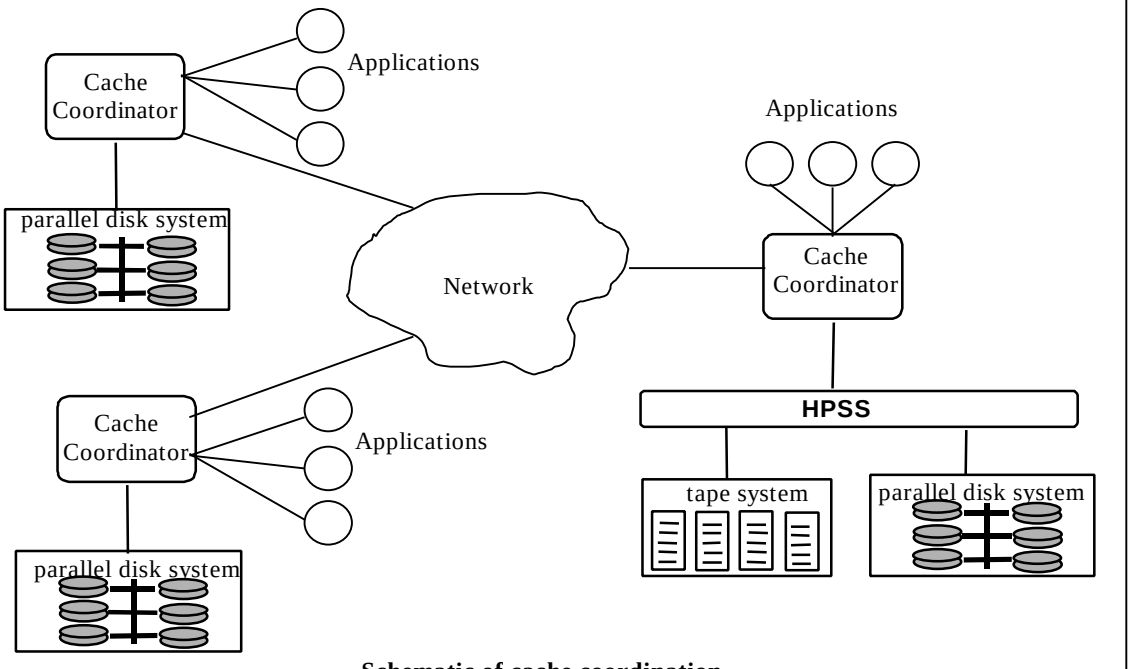


Figure 6. Schematic of data management services in the Earth System Grid.

show in this figure that in the future some of the nodes will have tape storage as well. It shows the distributed Cache Coordination modules. Each Cache Coordination module includes a local Metadata Catalog, a Query Monitor, and a Cache Manager.

3.3. High-Performance Data Transfer Service

As explained above, the high-performance data transfer service that we propose to develop in this project will allow users (or applications such as the distributed cache manager) to express relatively high-level specifications of data transfer operations (eg: file transfer from the DPSS system at LANL to NCAR). This service will hide complicated details from the users and will be responsible for locating, reserving, and configuring appropriate resources so as to ensure required end-to-end quality of service.

The construction of this high-performance data transfer service requires the development and integration of five related technologies:

- Resource estimation techniques able to determine with a fair degree of accuracy the resource requirements of a data transfer or analysis task.
- Discovery mechanisms that determine the current and likely future availability of relevant resources, and relevant properties of those resources.
- Reservation mechanisms that guarantee availability of selected resources
- High performance, network-aware bulk data transfer mechanisms able to configure protocol parameters to achieve high performance
- Instrumentation mechanisms that can determine when desired performance is achieved, and to diagnose problems when it is not.

A variety of these mechanisms exist already, in one form or another. For example, the Globus directory service [FFKL+97] and resource management architecture [CFNK+98, CFK99, FKL+99, HJFR98] provide resource discovery and reservation capabilities, the DPSS client library contains a high performance data transfer capability [DPSS], and the NetLogger Toolkit [TJC98] contains many of the necessary instrumentation mechanisms. We are hopeful that others will be developed in the near future, as a result of existing funding or funding requested from DOE NGI Technology proposals. We see one major challenge in the current proposal as being the *effective integration of these technologies in a realistic (and hence heterogeneous) systems context*. In the following, we touch upon just a few of the major issues that must be addressed in creating the proposed high-performance data transfer service.

3.3.1. High-Speed Transport Issues

The goal of the proposed high-speed remote data access API and client library is to provide applications with high-performance access to a broad variety of remote data systems, including HPSS, DPSS, the SDSC Storage Resource Broker (SRB) [CHAIT98],, and network attached storage. We propose to achieve this goal by building upon the knowledge gained from the DPSS, GASS [BFKTT99], SRB and parallel ftp data access methods.

For example, we will use techniques pioneered within the DPSS client library to obtain high throughput over wide area networks. These techniques include the use of multiple TCP sockets for the data stream, possibly as many as 1 per disk, with each socket managed by a separate thread; the use of large block sizes (at least 64 KB), reading and writing when possible at least 100 blocks at a time; being very careful to set the correct TCP send and receive buffer sizes [SEMK98], as setting them too large or too small adversely affects performance; and being

careful in the implementation to avoid unnecessary data copies, manipulating pointers to data blocks instead. This library will combine these DPSS-inspired techniques with the global naming facility from SRB, security and access control from GSI, data migration from GASS, and so on.

3.3.2. Quality of Service Issues

Achieving predictable performance in such a context is a complex optimization problem, with performance requirements and system characteristics as inputs and resource elections, protocol parameters, and resource reservations as outputs. Specific problems that we need to address include the following:

- How to configure Differentiated Services (diffserv) [ZJN97] configurations and network protocols to achieve good performance for high-bandwidth flows in both homogeneous (all diffserv) and heterogeneous networks
- How to configure disk, computer, and network systems and reservations to meet end-to-end performance requirements [CFK+98, GS99, HJFR98, MIS96, NS96,NCN98]
- How to use historical data concerning resource availability to govern selection and configuration for resources that do not support reservation [DEWITT97,STEEN99,Wol96]

In addition to these issues, resource reservations and protocol parameters required for efficient and effective data transfer are by no means obvious. For example, there is currently very knowledge of how TCP and diffserv interact or of how diffserv and TCP parameters should be selected. Adding in network elements that do not support diffserv, and various end systems, serves to further complicate this situation.

3.4. Remote Execution Service

As explained above, our proposed Grid-enabled data analysis clients will be designed so that the performance of complex analysis tasks can be optimized by selecting an appropriate partitioning of analysis functionality across local and remote systems. For example, an analysis designed to detect anomalous mean values might perform the mean value computation on the remote system, "near" to the data, then perform anomalous value detection and visualization locally.

The remote execution service to be developed in this project is designed to facilitate the development of such partitioned services. In particular, it provides mechanisms for locating appropriate computers and datasets; determining availability and performance characteristics; locating, constructing, or transferring a remote executable; initiating remote computation; and transferring data between components. These mechanisms can to a large extent be defined in terms of Globus security, resource management, computation management, and communication services. However, new work will be required to integrate these components and to address specific concerns that arise in climate systems. For example, the ability to download analysis code to a remote system and to synthesize network bandwidth reservation requests is expected to be important.

3.5. Grid Services and Security Enhancements

The high-speed data transfer and remote execution services described above will be defined in terms of lower-level security, directory, instrumentation, and resource management services. These "Grid services" will be constructed primarily either from existing software (Globus, NetLogger, etc.) or (we hope!) in other proposed DOE NGI projects. However, one specific task is proposed for the current project, namely enhancements to Globus security services.

As noted above, the Globus-based "Grid Security Infrastructure" (GSI) [FKTT98] provides public key-based single sign-on, run anywhere capabilities for multi-site environments, supporting proxy credentials, interoperability with local security mechanisms, local control over access (via access control lists), and delegation. A wide range of GSI-based applications have been developed, ranging from ssh and ftp to the Message Passing Interface (MPI), Condor, and SRB, and numerous institutions world-wide have approved GSI as a basis for resource access. Hence, GSI provides an excellent basis for authentication in the proposed Earth System Grid. However, extensions are required to deal with specific security concerns that arise in some high-security environments (e.g., LANL), such as firewalls and a requirement for tight local control over credentials.

Discussions between Globus project and LANL personnel suggest that a workable approach to these concerns may well be to (a) use RSA smartcards to hold public key credentials and (b) configure firewalls and Globus software to negotiate transit of appropriately authenticated traffic. We propose in this project to investigate the practical utility of this approach, with the goal of enabling secure single sign-on access to resources at all participating institutions. This work will proceed via an initial RSA smartcard pilot activity and associated enhancements to GSI services (similar mechanisms were prototyped by ANL staff in late 1998), followed if successful by the integration of these mechanisms with Grid-enabled analysis clients and broad distribution of smartcard technology.

3.6. Enhanced Network Services

Realizing this vision will require the development of new NGI technologies in addition to the deployment of high-speed networks. Open research issues exist in a number of critical areas, including high-speed end-to-end data transport, network-aware applications, and reservation of bandwidth. All of these issues are complicated by the need of the application to span multiple networks with which are owned and managed independently. We propose work in all of these areas.

We must construct a networking testbed connecting climate researcher to a prototype of the NGI environment in which they will work in the future. We propose to use a combination of NTON, ESnet, vBNS and MREN to facilitate this. NTON will be used to interconnect LBNL to LLNL, and the interconnection of NTON and ESnet is easily accomplished at LBNL. The DOE China Clipper effort has constructed an ESnet networking testbed connecting ANL to LBNL. VBNS will be used to connect NCAR to Esnet, and MREN is separately proposing a University Network Technology Testbed that includes ANL and the University of Wisconsin. We propose to combine these four networks to create a joint DOE-University networking testbed. In the process, we will also learn how to reserve bandwidth, shape traffic, and monitor state on multiple, independently owned and managed networks. This testbed will include a wide range of bandwidth, ranging from OC-48 (NTON) to OC-3 (vBNS and MREN), which will provide an outstanding environment for network research.

We will deploy the NetLogger monitoring tools in this testbed, adding NetLogger monitoring to as many components as possible, including network switches and routers, and application and storage hosts. In addition to this, we will add NetLogger instrumentation to all client and server software. This instrumentation will enable detailed performance analysis and facilitate tuning.

We will characterize and evaluate traffic used by this climate application and determine its effect on the WAN infrastructure. While there have been many isolated and small-scale studies in network traffic characterization [LTWW94, PF95, AW97, CB97], none have been of this magnitude in terms of the volume of data to be moved and the distance over which that data must be moved.

Specifically, we will characterize applications that make heavy use of the LAN, LAN/WAN gateway, and WAN, e.g., climate model data analysis, and then monitor and characterize the network traffic generated by such applications. Based on this characterization, we will propose flow- and congestion-control techniques which will minimize the effects of congestion (i.e., packet/cell loss) and maximize throughput between remote sites.

Within the next two years, parts of the LAN and the SAN (storage area network) to HPSS at LANL will be capable of data transfers up to 6.4 Gb/s. The network research group at LANL will develop the design of a HiPPI-6400/WAN gateway to provide better data off-loading to the WAN. Due to the wide disparity in bandwidth between the LAN (800 Mb/s – 1.2 Gb/s at LANL today, 6.4 Gb/s in 2000) and WAN (155 Mb/s at LANL today, 622 Mb/s in 2000), we will develop appropriate techniques to stripe multiple WAN channels across a LAN channel.

LANL will explore the use of an OS-bypass protocol such as Scheduled Transfer (ST) [PL99, ST], developed at Los Alamos National Laboratory and currently being standardized by ANSI. This would eliminate the interrupt handling that would otherwise have to occur with an OS-based network protocol such as TCP/IP. This approach eliminates interrupts altogether by bypassing the operating system and writing into the receiving host's application space directly. While the ST protocol was designed with the LAN in mind, it has the *potential* to be able to scale over the WAN. Issues related to adapting ST to work in the wide area network will be explored.

Vern Paxson of LBNL, in a paper at SIGCOMM '97, measured that the Internet corrupts 1 out of every 5000 packets [Pax97]. He goes on to say:

> "A corruption rate of 1 packet in 5,000 is certainly not negligible, because TCP protects its data with a 16-bit checksum. Consequently, on average one bad packet out of 65,536 will be erroneously accepted by the receiving TCP, resulting in *undetected data corruption*. If the 1 in 5000 rate is correct, then about one in every 300 million Internet packets is accepted with corruption – certainly many each day. In this case, we argue that TCP's 16-bit checksum is no longer adequate."

For large data transfers, an IP packet typically contains 64 Kbytes of data. So if one in every 300 million packets is corrupt, then for every 19.2 Terabytes transferred, there will be one undetected error. Since we are now talking about single data sets that are a terabyte, these errors can no longer be ignored. We propose to set up an experiment to verify that these errors actually occur, and if so, we will add support to the *High Performance Data Transfer Service* proposed above to correct this. Eventually the TCP protocol will be modified to use a 32-bit checksum, but until then users of extremely large data sets will need to provide their own data integrity checking.

4. Deliverables and Milestones

In this section we list the research outputs for the Earth System Grid project. In addition, we detail a schedule of milestones and deliverables. We propose a three year project.

4.1. Milestones for Project Year 1:

- Creation of prototype high-performance data transfer library.
- Analyzis and characterization of the ESnet to vBNS connection for DOE-NCAR transfers.
- Creation of an online database of archived climate data.
- Establishment of RSA smartcard testbed; investigation of firewall issues.
- Investigation of potential TCP 16-bit checksum problem.
- Quasi-real-time distribution of output from century-long climate simulations to multiple centers.
- As an integrating demonstration, prototype access from analysis clients to remote data.

4.2. Milestones for Project Year 2:

- Remote execution service.
- Location of data replicas via Metadata Catalog.
- Single-sign-on capability for Earth System Grid resources.
- SAN/WAN gateway.
- Prototype tape scheduling service.
- Develop solution to TCP 16-bit checksum problem, if real.
- Diffserv QoS support in high-performance data transfer service.
- Demonstrate grid-enabled analysis clients that can request derived products from dynamically selected remote datasets and computers.
- Prototype visualization of remote datasets on thin clients (VisAD).

4.3. Milestones for Project Year 3

- Scale up prototype system to support real-time distribution of larger datasets and more grid-enabled analysis clients.
- Demonstration of the distributed cache management service, supporting replication, discovery, and caching of datasets.
- Demonstration of the visualization of distributed datasets on thin clients (VisAD).
- Automatic migration from local tape to local disk, automatic distributed cache management.
- Introduction of collaboration technologies.
- Demonstrate Earth Systems Grid Broker that selects distributed data and computing resources on the basis of requirements and current loads, including QoS and bandwidth reservation.

5. Linkages and Technology Transfer

The ambitious goals of this project are possible because we can build on a strong base of software and expertise, in such areas as high-performance networking, advanced networked middleware (e.g., Globus and DPSS), climate data analysis systems, and high-performance storage systems. These connections will continue as we pursue the research goals addressed here, with for example other funding provided by DOE and other agencies being used to perform long-duration climate runs at LANL, further develop PCMDI data analysis systems, and continue Globus development.

More broadly, we view the research described here as forming a key component of a larger activity designed to develop and apply an Integrated Grid Architecture. As discussed at http://www.gridforum.org, this integrated architecture comprises four principal components:

- At the *Application Toolkit* level, toolkits provide more specialized services for various applications classes: e.g., data-intensive, remote visualization, distributed computing, collaboration, problem solving environments.
- At the Grid Services or middleware level, a suite of Grid-aware services implement basic mechanisms such as authentication, authorization, resource location, resource allocation, and event services.
- At the *Grid Fabric* level, primitive mechanisms provide support for high-speed network I/O, differentiated services, instrumentation, etc.
- Finally, specific *Grid-aware applications* are implemented in terms of various Grid Services and Application Toolkit components.

The project proposed here will contribute to this overall architecture at multiple levels. At the fabric and services levels, we will contribute to an understanding of how to support very high bandwidth flows and how to perform resource management to achieve end-to-end performance. At the application toolkit level, we will contribute to the development of distributed computing and data management toolkits. Hence, this work directly supports existing and proposed DOE research activities concerned with distributed computing (e.g., distance corridors, climate modeling, materials science) and complements existing and proposed DOE research activities concerned with networking technology and middleware: in particular, those concerned with resource management and quality of service, instrumentation, and network topology determination. We will also be relying on testbed capabilities provided by ESnet, the sites participating in the Globus Ubiquitous Supercomputing Testbed Organization (GUSTO), and (we hope) the proposed EMERGE testbed.

The proposed research is also highly relevant to the goals of the SSI and ASCI projects. In addition to addressing directly key requirements of the SSI Global Systems component, we will develop technologies highly relevant to ASCI DISCOM and Distance Corridor efforts and to other SSI application areas (e.g., Combustion) with a need for high-bandwidth transport and access to large amounts of data.

Technology transfer from this project to DOE application scientists and further afield will occur via the highly successful PCMDI data analysis clients and Globus software distribution mechanisms; via the various linkages noted above; and via the strong ongoing collaborative links that exist between the participants and other programs, in particular the broad NCAR user community, NSF PACIs, DARPA, and the NASA Information Power Grid program.

6. Team Balance, Qualifications, and Management Plan

Our research team is well qualified to accomplish the proposed work. The team is balanced, with demonstrated expertise in computational climate simulation, data analysis and visualization tools, data management systems, middleware and grid mechanisms, and high performance networking. Furthermore, various subsets of the team have worked together over many years on various other projects: e.g., CHAMMP (ANL/LANL/LLNL/NCAR), DOE2000 (ANL/LANL/LBNL), Grid middleware (ANL/LANL/LBNL), climate simulations (LANL/LLNL/NCAR), Clipper (ANL/LBNL).

Management of the three-year project proposed here will be modeled after the highly successful management strategy for geographically distributed collaborative projects within the DOE Computer Hardware, Advanced Methods, and Model Physics (CHAMMP) program and the current Climate Change and Prediction Program (CCPP). This project involves many of the same principals in these other two DOE-sponsored programs and thus there is an established precedent for collaboration and successful project completion. We plan on managing this project through a combination of frequent emails, conference calls every other week, quarterly meetings, and conducting a major demonstration each year at a conference such as the annual Supercomputing conference or the annual AMS meeting.

Steve Hammond of NCAR and Dean Williams of LLNL will provide overall project management and coordination of application development. Ian Foster at ANL will coordinate middleware development. Wu Feng at LANL and Brian Tierney at LBNL will coordinate the development and operation of the networking efforts for this testbed. Liaison with networking testbeds will be provided by Tierney (NTON, ESnet) and Foster (MREN).

The day-to-day work of the project team will be coordinated in ways that we have found extremely effective in the past, both in the collaborations listed above and in other large projects such as the MAGIC testbed and the Globus project. These techniques include the use of shared code repositories, shared web pages, frequent teleconferences, frequent visits, regular project meetings, and integrated technology demonstrations as a means of assessing progress and forcing integration.

7. References Cited

[ACPI98] "The Accelerated Climate Prediction Initiative: Bringing the Promise of Simulation to the Challenge of Climate Change", Committee Report Comissioned by Ari Patrinos, June 1998. (http://www.epm.ornl.gov/ACPI/Documents/ACPIfinal.html)

[AW97] M. Arlitt and C. Williamson. "Internet Web Servers: Workload Characterization and Performance Implications," IEEE/ACM Transactions on Networking, 5(5):632-645, October 1997.

[BR98] L. Bernardo, H. Nordberg, D. Rotem, and A. Shoshani, Determining the Optimal File Size on Tertiary Storage Systems Based on the Distribution of Query Sizes, Tenth International Conference on Scientific and Statistical Database Management, 1998. (http://www.lbl.gov/~arie/papers/file.size.ssdbm.ps).

[BNR+98] L. Bernardo, H. Nordberg, D. Rotem, and A. Shoshani, Determining the Optimal File Size on Tertiary Storage Systems Based on the Distribution of Query Sizes, Tenth International Conference on Scientific and Statistical Database Management, 1998. (http://www.lbl.gov/~arie/papers/file.size.ssdbm.ps).

[BFKTT99] Joseph Bester, Ian Foster, Carl Kesselman, Jean Tedesco, Steven Tuecke, GASS: A Data Movement and Access Service for Wide Area Computing Systems, Proc. IOPADS'99, ACM Press, 1999.

[BCFF+98] Sharon Brunett, Karl Czajkowski, Steven Fitzgerald, Ian Foster, Andrew Johnson, Carl Kesselman, Jason Leigh, and Steven Tuecke. Application experiences with the Globus toolkit. In Proc. 7th IEEE Symp. on High Performance Distributed Computing, pages 81-89. 1998.

[CB97] M. Crovella and A. Bestavros. "Self-Similarity in WWW Traffic: Evidence and Possible Causes," IEEE/ACM Transactions on Networking, 6(5):835-845, December 1997.

[CDK+95a] L.T. Chen, R. Drach, M. Keating, S. Louis, D. Rotem and A. Shoshani, Efficient Organization and Access of Multi-Dimensional Datasets on Tertiary Storage Systems}, Information Systems Journal, Pergammon Press, April 1995, vol. 20, (no. 2): 155-83. (http://www.lbl.gov/~arie/papers/optimass.Info.Sys.ps).

[CDK+95b] L.T. Chen, R. Drach, M. Keating, S. Louis, D. Rotem, and A. Shoshani, Optimizing Tertiary Storage Organization and Access for Spatio-Temporal Datasets, NASA Goddard Conference on Mass Storage Systems, March 1995. (http://www.lbl.gov/~arie/papers/optimass.goddard.ps).

[CS92] C. Catlett and L. Smarr. Metacomputing. Communications of the ACM, 35(6):44--52, 1992.

[CFK+98] Prashant Chandra, Allan Fisher, Corey Kosak, T.S. Eugene Ng, Peter Steenkiste, Eduardo Takahashi, and Hui Zhang, Darwin: Resource management for value-added customizable network service, In Sixth IEEE International Conference on Network Protocols (ICNP'98), 1998.

[CCLP] China Clipper: http:/www-didc.lbl.gov/Clipper

[CDK95a] L.T. Chen, R. Drach, M. Keating, S. Louis, D. Rotem and A. Shoshani, Efficient Organization and Access of Multi-Dimensional Datasets on Tertiary Storage Systems}, Information Systems Journal, Pergammon Press, April 1995, vol. 20, (no. 2): 155-83. (http://www.lbl.gov/~arie/papers/optimass.Info.Sys.ps).

[CDK95b] L.T. Chen, R. Drach, M. Keating, S. Louis, D. Rotem, and A. Shoshani, Optimizing Tertiary Storage Organization and Access for Spatio-Temporal Datasets, NASA Goddard Conference on Mass Storage Systems, March 1995. (http://www.lbl.gov/~arie/papers/optimass.goddard.ps).

[CFNK+98] K. Czajkowski, I. Foster, N. Karonis, C. Kesselman, S. Martin, W. Smith, and S. Tuecke. A resource management architecture for metacomputing systems. In The 4th Workshop on Job Scheduling Strategies for Parallel Processing, pages 62-82. Springer-Verlag LNCS 1459, 1998.

[CFK99] Czajkowski, I. Foster, and C. Kesselman, Co-Allocation Services for Computational Grids. Proceedings of the IEEE Symposium on High-Performance Distributed Computing, 1999.

[CHAIT98] Chaitanya Baru, Reagan Moore, Arcot Rajasekar, Michael Wan, "The SDSC Storage Resource Broker", Proc. CASCON'98 Conference, Nov.30-Dec.3, 1998, Toronto, Canada. (http://www.npaci.edu/DICE/SRB/)

[DKPS97] M. Degermark, T. Kohler, S. Pink, and O. Schelen. Advance reservations for predictive service in the internet. ACM/Springer Verlag Journal on Multimedia Systems, 5(3), 1997.

[DS99] Tom DeFanti and Rick Stevens, Teleimmersion. In The Grid: Blueprint for a Future Computing Infrastructure, pages 131-155. Morgan Kaufmann Publishers, 1999.

[DEWITT97] DeWitt, T. Gross, T. Lowekamp, B. Miller, N. Steenkiste, P. Subhlok, J. Sutherland, D., "ReMoS: A Resource Monitoring System for Network-Aware Applications" Carnegie Mellon School of Computer Science, CMU-CS-97-194. http://www.cs.cmu.edu/afs/cs/project/cmcl/www/remulac/remos.html

[D99] Robert Drach, "The Climate Database Management System", 1999. (http://www-pcmdi.llnl.gov/software/)

[DPSS] DPSS: http://www-didc.lbl.gov/DPSS/

[FFKL+97] S. Fitzgerald, I. Foster, C. Kesselman, G. von Laszewski, W. Smith, and S. Tuecke, A Directory Service for Configuring High-performance Distributed Computations, Proc. 6th IEEE Symposium on High-Performance Distributed Computing, 365-375, IEEE Press, 1997.

[FGV97] D. Ferrari, A. Gupta, and G. Ventre. Distributed advance reservation of real-time connections. ACM/Springer Verlag Journal on Multimedia Systems, 5(3), 1997.

[FGKT97] I. Foster, J. Geisler, C. Kesselman, S. Tuecke, Managing Multiple Communication Methods in High-performance Networked Computing Systems, Journal of Parallel and Distributed Computing, 40, 35-48, 1997.

[FK98] I. Foster and C. Kesselman, The Globus Project: A Status Report, Proceedings of the Heterogeneous Computing Workshop, IEEE Press, 4-18, 1998.

[FK99a] I. Foster and C. Kesselman, editors. The Grid: Blueprint for a Future Computing Infrastructure. Morgan Kaufmann Publishers, 1999.

[FK99b] I. Foster and C. Kesselman. Globus: A Toolkit-Based Grid Architecture. In The Grid: Blueprint for a Future Computing Infrastructure, pages 259-278. Morgan Kaufmann Publishers, 1999.

[FKTT98] I. Foster and C. Kesselman and G. Tsudik and S. Tuecke, A Security Architecture for Computational Grids, ACM Conference on Computers and Security, 83-91, ACM Press, 1998.

[FKL+99] Ian Foster, Carl Kesselman, Craig Lee, Bob Lindell, Klara Nahrstedt, Alain Roy, and Steven Tuecke, A Distributed Resource Management Architecture that Supports Advance Reservations and Co-Allocation, submitted, 1999.

[GATE99] "Recommended Implementation of the Accelerated Climate Prediction Initiative (ACPI)", Report of the Ad Hoc Inter-agency Committee for ACPI Implementation, W. L. Gates, Chairman, Feburary, 1999.

[GS99] Roch Guerin and Henning Schulzrinne, Network Quality of Service. In The Grid: Blueprint for a Future Computing Infrastructure, pages 479-503. Morgan Kaufmann Publishers, 1999.

[HBD98] A. Hafid, G. Bochmann, and R. Dssouli. A quality of service negotiation approach with future reservations (nafur): A detailed study. Computer Networks and ISDN Systems, 30(8), 1998.

[HJFR98] G. Hoo, W. Johnston, I. Foster, and A. Roy, QoS as middleware: Bandwidth broker system design, Technical report, LBNL, 1999.

[Jon99] William Johnston, Realtime Widely Distributed Instrumentation Systems. In The Grid: Blueprint for a Future Computing Infrastructure, pages 75-103. Morgan Kaufmann Publishers, 1999.

[JLT98] W. Johnston, J. Lee, B. Tierney, and C. Tull, Directions and Issues for High Data Rate Wide Area Network Environments, Proceedings of the Computers in High Energy Physics Conference, August 1998, LBNL-42610.

[LFIB+99] Gregor von Laszewski, Ian Foster, Joseph A. Insley, John Bresnahan, Carl Kesselman Mei Su, Marcus Thiebaux, Mark L. Rivers, Ian McNulty, Brian Tieman, and Steve Wang. Real-time analysis, visualization, and steering of microtomography experiments at photon sources. In Proceedings of the Ninth SIAM Conference on Parallel Processing for Scientific Computing. SIAM, 1999.

[LTWW94] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. "On the Self-Similar Nature of Ethernet Traffic," IEEE/ACM Transactions on Networking, 2(1):1-15, Feburary 1994.

[MIS96] A. Mehra, A. Indiresan, and K. Shin. Structuring communication software for quality-of-service guarantees. In Proc. of 17th Real-Time Systems Symposium, December 1996.

[Mes99] Paul Messina, Distributed Supercomputing Applications. In The Grid: Blueprint for a Future Computing Infrastructure, pages 55-73. Morgan Kaufmann Publishers, 1999.

[MBMRW99] Reagan Moore, Chaitanya Baru, Richard Marciano, Arcot Rajasekar, Michael Wan, Data-Intensive Computing. In The Grid: Blueprint for a Future Computing Infrastructure, pages 105-129. Morgan Kaufmann Publishers, 1999.

[NCN98] K. Nahrstedt, H. Chu, and S. Narayan. QoS-aware resource management for distributed multimedia applications. Journal on High-Speed Networking, IOS Press, December 1998.

[NS96] K. Nahrstedt and J. M. Smith. Design, implementation and experiences of the OMEGA end-point architecture. IEEE JSAC, Special Issue on Distributed Multimedia Systems and Technology, 14(7):1263-1279, September 1996.

[NTON]: http://www.ntonc.org

[PF95] V. Paxson and S. Floyd. "Wide-Area Traffic: The Failure of Poisson Modeling," IEEE/ACM Transactions on Networking, 3(3):226-244, June 1995.

[Pax97] V. Paxson, End-to-End Internet Packet Dynamics, SIGCOMM '97, LBNL-40488 (ftp://ftp.ee.lbl.gov/papers/vp-pkt-dyn-sigcomm97.ps)

[SBN98] A. Shoshani, L. M. Bernardo, H. Nordberg, D. Rotem, and A. Sim, Storage Management for High Energy Physics Applications, Computing in High Energy Physics 1998 (CHEP 98), (http://www.lbl.gov/~arie/papers/proc-CHEP98.ps).

[SBN+98] A. Shoshani, L. M. Bernardo, H. Nordberg, D. Rotem, and A. Sim, Storage Management for High Energy Physics Applications, Computing in High Energy Physics 1998 (CHEP 98), (http://www.lbl.gov/~arie/papers/proc-CHEP98.ps).

[SBN+99] A. Shoshani, L. M. Bernardo, H. Nordberg, D. Rotem, and A. Sim, Storage Management Techniques for Very Large Multidimensional Datasets, February 1999, Submitted for publication.

[SEMKE98] J. Semke, J. Mahdavi, M. Mathis, "Automatic TCP Buffer Tuning", Computer Communication Review, ACM SIGCOMM, volume 28, number 4, Oct. 1998.

[STA99] Storage Access Coordination System (STACS): http://gizmo.lbl.gov/sm/

[STEEN99] P. Steenkiste, "Adaptation Models for Network-Aware Distributed Computations", 3rd Workshop on Communication, Architecture, and Applications for Network-based Parallel Computing (CANPC'99), Orlando, January, 1999.

[SWDC97] R. Stevens, P. Woodward, T. DeFanti, and C. Catlett, From the I-WAY to the National Technology Grid, Communications of the ACM, 40(11):50-61, 1997.

[TJC98] B. Tierney, W. Johnston, B. Crowley, G. Hoo, C. Brooks, D. Gunter, "The NetLogger Methodology for High Performance Distributed Systems Performance Analysis," Seventh IEEE International Symposium on High Performance Distributed Computing, Chicago, Ill., July 28-31, 1998. (http://www-itg.lbl.gov/DPSS/papers.html).

[TJH94] B Tierney, W Johnston, H Herzog, G Hoo, G Jin, J Lee, System Issues in Implementing High Speed Distributed Parallel Storage Systems , Proceedings of the USENIX Symposium on High Speed Networking, Aug. 1994, LBL-35775.

[WM94] Dean N. Williams and Robert L. Mobley, "The PCMDI Visualization and Computation System (VCS): A Workbench for Climate Data Display and Analysis", PCMDI report number 17 and UCRL-ID-116980, 1994. (http://www-pcmdi.llnl.gov/pcmdi/pubs/ab17.html

[Wet.al.99] Dean Williams, Jim Boyle, Clyde Dease, Charles Doutriaux, Robert Drach, Jay Hnilo, Susan Marlais, Charlie O'Connor, Jerry Potter, and Karl Taylor, "A User's Guide to the PCMDI Climate Data Analysis Tool (CDAT): A Workbench for Climate Data Analysis and Display", 1999. (http://www-pcmdi.llnl.gov/software/cdat/)

[WMDP95] Dean N. Williams, Robert L. Mobley, Robert S. Drach, and Thomas J. Phillips, "The Visualization and Computation System (VCS) – Uniquely Versatile Software", DOE Research Summary, No. 33, 1995.

[W96] Dean N. Williams, "The Visualization and Computation System (VCS)", 1996. (http://www-pcmdi.llnl.gov/software/vcs/)

[W97] Dean N. Williams, "The PCMDI Software System: Status and Future Plans", ", PCMDI report number 44 and UCRL-ID-129074, 1997. (http://www-pcmdi.llnl.gov/pcmdi/pubs/ab44.html)

[WS97] L.C. Wolf and R. Steinmetz. Concepts for reservation in advance. Kluwer Journal on Multimedia Tools and Applications, 4(3), May 1997.

[Wol96] R. Wolski, Forecasting Network Performance to Support Dynamic Scheduling Using the Network Weather Service, Proc. 5th IEEE Symposium on High-Performance Distributed Computing, IEEE Press, 1996.

[ZJN97] L. Zhang, V. Jacobson, and K. Nichols, A two-bit differentiated services architecture for the internet. Internet Draft, Internet Engineering Task Force, 1997.

8. ACRONYMS used in this proposal

ACPI	Accelerated Climate Prediction Initiative
ANL	Argonne National Laboratory
API	Applications Program Interface
ASCI	Advanced Scientific Computing Initiative
ATM	Asynchronous Transfer Mode
CCPP	Climate Change & Prediction Program
CDAT	Climate Data Analysis Tool
CDMS	Climate Data Management System
CHAMMP	Computer Hardware, Advanced Math. & Model Physics
CM	Cache Manager
DOE	Department of Energy
DPSS	Distributed Parallel Storage System
ESG	Earth System Grid
ESnet	Energy Sciences Network
GASS	Global Access to Secondary Storage
GATE	GARP Atlantic Tropical Experiment
GSI	Grid Security Infrastructure
GUSTO	Globus Ubiquitous Supercomputing Testbed Organization
HEP	High Energy Physics
HPSS	High Performance Storage System
IP	Internet Protocol
ISP	Internet Service Provider
LANL	Los Alamos National Laboratory
LAN	Local Area Network
LBNL	Lawrence Berkeley National Laboratory
LLNL	Lawrence Livermore National Laboratory
LRU	Least Recently Used
MDS	Metacomputing Directory Service
MPI	Message Passing Interface
MREN	Metropolitan Research & Education Network
MSS	Mass Storage System
MTU	Maximum Transmission Unit
NASA	National Aeronautics & Space Administration
NCAR	National Center for Atmospheric Research
NGI	Next Generation Internet
NLANR	National Laboratory for Applied Network Research
NSF	National Science Foundation
NTON	National Transparent Optical Network
OPTIMASS	Optimization of Mass Storage Access
OS	Operating System
PCMDI	Program for Climate Model Diagnosis & Intercomparison
QE	Query Estimator
QM	Query Monitor
QoS	Quality of Service
SAN	Storage Area Network
SDSC	San Diego Supercomputing Center
SLAC	Stanford Linear Accelerator Center

SRB	Storage Resource Broker
SSI	Strategic Simulation Initiative
SST	Sea Surface Temperature
STACS	Storage Access Coordination System
TCP	Transport Control Protocol
TCP/IP	Transport Control Protocol/Internet Protocol
USC	University of Southern California
vBNS	Very High Performance Backbone Network Service
VCS	Visualization and Computation System
WAN	Wide Area Network

9. Biographies

10. Budget

This proposal requests $2.2M per year for three years. The proposed design, development, and implementation of the prototype Earth System Grid will require considerable collaboration and individual effort. We are requesting funding for personnel sufficient to support this proposed work as well as $85K to enhance the storage system/data cache at LLNL. The details are listed below.

The money requested by University of Wisconsin supports Bill Hibbard and Dave Glowacki is to develop applications of VisAD for remote and collaborative visualization of climate simulations.

At NCAR, the funds requested total 1.5 FTEs for each of three years.
The requested money will be divided as follows:

1. .5 FTE to support a computer system expert to install the Globus software, maintain the utilities as the middleware is enhanced, and make necessary modifications appropriate for NCAR site specifics.
2. A network specialist to coordinate NCAR's networking efforts with the network specialists at LANL and LBNL. To monitor network routes and service and to diagnose any problems that arise. This effort will be front loaded to support the network person .6 FTE, .4 FTE and then .2 FTE over the three years.
3. A computer visualization specialist to collaborate with William Hibbard at U. Wisconsin on development of visualization utilities such as VisAd to deliver animations of very large, distributed climate datasets to light weight clients. This work will be phased in so that the support is .2 FTE, .4 FTE, and .6 FTE for the three years.
5. Finally, .1 FTE each year will be needed to support Steve Hammond as overall project coordinator and applications coordination (both done jointly with Dean Williams).

Materials and supplies expenditures will support project personnel as the prepare scientific papers, posters, and software demonstrations. The travel budget supports close collaboration between NCAR staff and the other sites.

At LBNL, the funds requested total 2.0 FTEs for each of three years.

1. 0.5 FTE: extend the Query Monitor module developed at LBNL to support the coordination of data requests from multiple nodes (see Section 2.2).
2. 0.5 FTE: extend the Cache Manager module developed at LBNL to communicate with remote Cache Managers to achieve parallel data transfers into the local cache (see Section 2.2).
3. 0.5 FTE: work with ANL to develop a "High Performance Data Transfer Service" (see Section 2.3).
4. 0.5 FTE: deploy NetLogger tools, help instrument application, and do performance analysis (see sections 2.3 and 2.6).

At LLNL, the funds requested total 2.0 FTEs for each of three years.

1. 1.0 FTE: coordination of the metadata catalogue development integration with Globus
2. 1.0 FTE: development of distributed analysis client and integration with the remote execution server, metadata catalogue and high speed data service

Upgrade to the existing RAID disk cache will total $85K.

At ANL, funds requested total 2.0 FTEs for each of three years.

The requested funds will be used as follows.

1. 1.8 FTEs will support distributed computing specialists charged with developing the high-performance data service (jointly with LBNL), remote execution service (jointly with ISI), and Grid services (jointly with ISI also).
2. 0.1 FTE per year will support Ian Foster's effort relating to coordination of activities at Argonne and coordination of the overall Earth System Grid middleware effort.
3. 0.1 FTE per year will support Argonne climate scientist John Taylor who will provide consulting assistance on climate-related aspects of the project.

At ISI, the funds requested total approximately 1.15 FTE and one graduate student each year. The requested funds will be used as follows:

1. .5 FTE plus one graduate student will be used to support development of remote execution services and Grid services in collaboration with Argonne National Laboratory.
2. 0.15 FTE per year will support Carl Kesselman's effort in this project. His responsibilities will be directing activities at ISI and interacting with collaborators at Argonne in the overall coordination and development of the Earth System Grid middleware effort.

At LANL, the funds requested total 2.0 FTEs for each of three years.

Year 1.

0.75 FTE: Analyze & characterize ESnet to vBNS connect for DOE-NCAR transfers, investigation of potential TCP 16-bit checksum problem.

0.25 FTE: Network tuning.

0.50 FTE: RSA smartcards & investigation of firewall issues.

0.50 FTE: Install and integrate Globus, security and access issues, MPI-IO, and GASS.

Year 2.

0.75 FTE: LAN/WAN gateway.

0.50 FTE: Traffic shaping of network traffic.

0.25 FTE: Single-sign-on capability to Earth System Grid resources.

0.50 FTE: "Publishing" data sets with GASS, enable parallel data-subsetting jobs via Globus, and linking that to the scheduler. Examine the staging of data both on and off the machine as part of the overall work flow. Diffserv QoS.

Year 3.

0.50 FTE: Completion of network striping work over LAN/WAN gateway. Additional network tuning with respect to these new capabilities.

1.50 FTE: Completion of overall system integration at LANL: middleware, metadata, and networking.

Materials and supplies expenditures will support project personnel as the prepare scientific papers, posters, and software demonstrations. The travel budget supports close collaboration between NCAR staff and the other sites.

Computer service costs and workstations purchases will support computer equipment required for the execution of the proposed tasks.

11. Current and Pending Support

LANL:

W. Feng, *Distance and Distributed Computing*. ASCI/DISCOM Flow Control Program, 1998-present.

W. Feng. *Network Interface Cards as First-Class Citizens: Alleviating the Application-to-Network Bottleneck.* DOE NGI Program LAB 99-08.

W. Feng and I. Philp. Flow and Congestion Control over High-Speed Networks. DOE NGI Program LAB 99-08.

LBNL:

The people that will work on the "Distributed Data Management Service" of this proposal are currently supported by funding from the High Energy Nuclear Grand Challenge (HENP-GC) project funded by the MICS office at DoE. This project will terminate at the end of FY99. Future funding - unknown.

The people that will work on the "High-Speed Transport" of this proposal are currently supported by funding from the DOE China Clipper Project, and DARPA Magic Project (ends this summer), DARPA AMAC Project (only .3 FTE) to Deploy NetLogger Tools on the DARPA SuperNet Testbed.

12. Facilities, Equipment, and Other Resources

There is only $85K in the budget requested for hardware. This will be used data cache upgrades in the LLNL group. The rest of this proposal pays for software/networking research and development. The infrastructure that this project will build upon is listed below:

12.1. Testbed requirements (The infrastructure that we will build upon)

LANL has recently taken delivery of an SGI Origin 2000 system with 2048 250-MHz processors. The system has a peak speed of one-teraflop and 512 GB of memory. Approximately one-third, 640 processors, are devoted to ocean, atmosphere and climate modeling. Funding is provided by the DOE CCPP and HPCC programs. Models running on the system include NCAR's CCM3 atmospheric model at T170 (75 km) resolution, LANL's POP ocean model running at 1/10º (10 km) resolution, and two coupled models from NCAR, the Parallel Climate Model based on CCM3 and POP and the Climate System Model based on CCM3 and the NCOM ocean model. Reliable, high-speed networking from Los Alamos to NCAR and other laboratories is crucial.

Relevant Argonne computing facilities include a 128-processor SGI Origin 2000 and 150-node IBM SP (for a total of around 100 Gigaops); 80 TB-capacity tape archive; multiple TBs of disk (of which we expect to configure at least 1 TB for use as a disk cache for this project); multiple multiprocessor Sun systems, which will be available for networking research purposes; and various production and experimental networking equipment that will support the networking research proposed for this project. The Argonne OC-12 ESnet connection, multiprocessor Sun E4000 system, and networking equipment were used to support the recent 320 Mb/s transfer rate ANL-LBNL networking experiments.

The NCAR scientific computing division operates a supercomputing facility on behalf of the National Science Foundation in support of U.S. academic atmospheric science research. This facility includes a 128 processor SGI Origin 2000, a Cray C9/16, and a group of Cray J90 systems. These systems are devoted to ocean, atmosphere and climate modeling. Funding is provided by the NSF and HPCC programs. In addition, there is a prototype "data park" system being developed to study the analysis of very large datasets stored from long running climate models. The NCAR compute facility has a 150 terabyte mass storage system, which is currently growing at about 4 terabytes per month. Finally, NCAR's network connections include the NSF vBNS which provides reliable, high-speed networking to universities and other laboratories.

LBNL computing facilities include the computational resources of the National Energy Research Scientific Computer Center (NERSC). This includes a massively parallel Cray T3E-900, with 640 application processing elements, each capable of performing 900 MFlops., a cluster of six Cray J90 machines that have a total of 160 vector processors, an 8 processor high performance rendering engine, a 106 TB capacity tape archive, multiple TBs of disk cache. LBNL also has available an ImmersaDesk and a rear projection Wall for display and interaction with semi-immersive visualizations. LBNL and ANL are also connected via an Esnet OC-12. LBNL also has connectivity to NTON, currently at OC-12, and soon to be upgraded to OC-48. LBNL has 2 large Linux PC clusters, the PDSF with 48 nodes, and the Future Technologies group PCP cluster, with 36 nodes. There is also a 300 gigabyte, four server DPSS cache system which is connected to NTON and ESnet and can provide over 450 Mbits/second of cache storage bandwidth.

LLNL/PCMDI facilities include a 1.5TB RAID cache with a SUN 450 2 processor server. Computing facilities include a SUN Ultrasparc 2 with 2 processors and an SGI ONYX. Individual scientists use INTEL Pentium computers running Linux.

At the university of Wisconsin, the facilities available for use in this project include a number of high-performance NT and Sun Solaris workstations and an OC-3 Internet 2 connection.