LASER INTERFEROMETER GRAVITATIONAL WAVE OBSERVATORY
- LIGO –
CALIFORNIA INSTITUTE OF TECHNOLOGY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Document Type *Annual Report*	LIGO- T010166-00-E	30 June 2001

LIGO DCC Archival Copy:

GriPhyN Annual Report for 2000–2001

NSF Grant 0086044

GriPhyn Collaboration
http://www.griphyn.org

California Institute of Technology
LIGO Laboratory - MS 18-34
Pasadena CA 91125
Phone (626) 395-212
Fax (626) 304-9834
E-mail: info@ligo.caltech.edu

Massachusetts Institute of Technology
LIGO Laboratory - MS 16NW-145
Cambridge, MA 01239
Phone (617) 253-4824
Fax (617) 253-7014
E-mail: info@ligo.mit.edu

www: http://www.ligo.caltech.edu/

Final Version
June 30, 2001

GriPhyN Annual Report for 2000–2001

NSF Grant 0086044

1 The GriPhyN Project

1.1 Introduction and goals

GriPhyN is a collaboration of information technology (IT) researchers and experimental physicists who aim to provide the IT advances required to enable Petabyte-scale data intensive science in the 21st century. Driving the project are unprecedented requirements for geographically dispersed extraction of complex scientific information from very large collections of measured data. To meet these requirements, which arise initially from the four physics experiments involved in this project (ATLAS, CMS, LIGO and SDSS) but will also be fundamental to science and commerce in the 21st century, the GriPhyN team will pursue IT advances centered on the creation of *Petascale Virtual Data Grids* (PVDG) that meet the data-intensive computational needs of a diverse community of thousands of scientists spread across the globe.

This report is being written at the start of GriPhyN's ninth month of funding. We have started substantial research IT activities in areas such as data management, query optimization, data grid architecture, as was reported at our all-hands meeting in April and described in Section 2.3 below. We have also performed substantial early experiments, notably a large CMS computation across a Data Grid spanning Caltech, UW, and NCSA. Much initial effort has also been spent on recruiting and hiring students, postdocs, scientists and staff; writing detailed requirements documents; setting formal milestones; implementing a management structure and organizing our collaboration in line with these milestones; and interacting with other Grid projects and the four experiments which compose GriPhyN. During this time, we have had two collaboration wide meetings, at which students and postdocs made significant contributions, as well as numerous smaller meetings for CS research groups and CS-experiment interactions. The report below discusses these topics in more detail.

1.2 Participants and first year funding

A table showing a list of participants and their affiliations is shown in Appendix I. We have expended much effort in finding and hiring personnel, but have experienced a somewhat slower rampup in personnel than we expected when the proposal was submitted, even though we had assumed 50% of our total FTE count in the first year. At the time of this report (June 2001) we have hired or made offers to approximately 80% of our full complement of hires (including people who were promised as matching contributions at Florida, Boston U and Indiana. We have had a particularly difficult time finding suitable candidates for the Project Coordinator position, in spite of a large cross-disciplinary search. Nevertheless, we have identified one and possibly two candidates for the position and we hope to have a Project Coordinator hired sometime in July. On the other hand, we have identified an outstanding candidate for the Outreach Coordinator (Manuela Campanelli) and she will start in Fall, 2001.

Although our hiring has gone somewhat more slowly than expected, we have incurred some additional costs for our External Advisory Committee, travel and web development. We estimate that slightly more than 20% of our funds will not be expended at the end of the first year, though it is hard to be precise more than three months before that date (12 subcontracts need to be tracked and the dates of some hires are not precise). We are setting up procedures to track our total expenses every three months to aid our planning and reporting.

1.3 Management structure

The management of the GriPhyN project has been conducted largely in accordance with the management plan presented in the proposal and in the more detailed *GriPhyN Management Plan*[1] prepared in November, 2000. The organization chart from the latter document is reproduced below. We describe briefly the activity in each of the chart's sections.

- The Project Coordination Group is chaired by the Project Directors Avery and Foster. It meets once or twice each week by telephone conference to review and plan the major activities of the project.
- The Applications Committee has not had formal meetings, but its members, the representatives of the four component physics collaborations, have worked actively with the CS Research Committee to define requirements and goals for integration of GriPhyN middleware into each collaboration's testbed as a function of time.
- The Virtual Data Toolkit Committee is chaired by Miron Livny. It has worked actively with the CS Research Committee to define the first release of the Toolkit as described in Section 2.2.2 below.
- The CS Research Committee, chaired by Carl Kesselman, has met as needed during the past several months to coordinate work on the four areas of CS research.
- The Technical Coordination Committee has worked only informally so far.
- The External Advisory Committee has been formed and has met to review the progress of the GriPhyN project during the general GriPhyN meeting at ISI in April. The EAC report[2] can be accessed at the GriPhyN web site.[3]
- An active search has been underway for the Project Coordinator position. We expect to fill the position by the middle of July 2001.

GriPhyN Management Org Chart

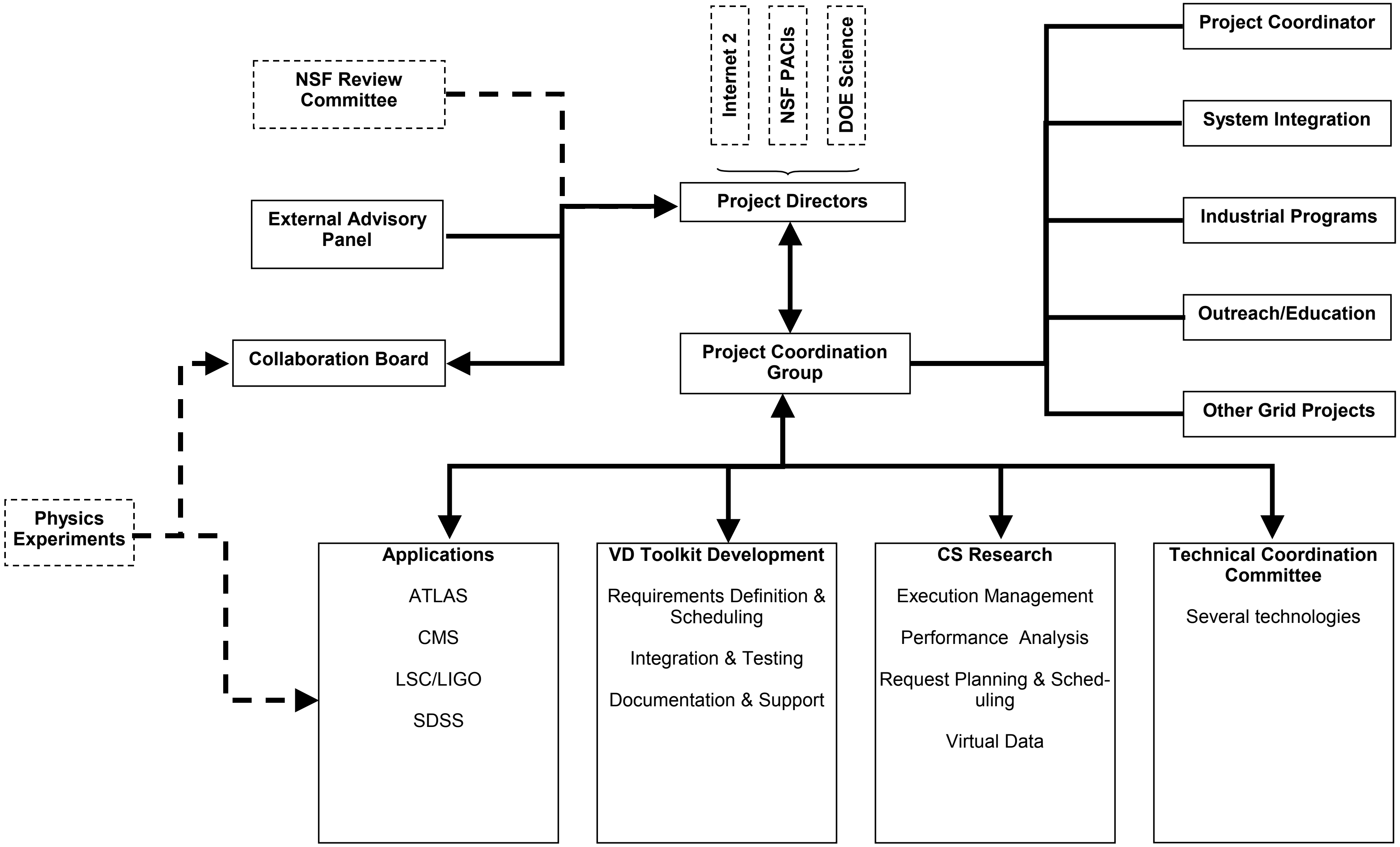


Figure 1: GriPhyN management diagram from Nov. 2000 management document[1]

2 Activities and Findings for 2000–2001

2.1 Creation of a Working Grid

This past spring, a collaborative effort involving physicists and computer scientists from Caltech, the National Center for Supercomputing Applications (NCSA), and the University of Wisconsin-Madison produced a working example of the Grid in action. The details are described in Appendix II, which will appear in NCSA Magazine.

2.2 Meetings

The size, scope and interdisciplinary nature of GriPhyN require building effective communication channels between the members of the collaboration. Early on we decided to use face-to-face meetings to establish and maintain such channels. In early October we had our first general meeting at Argonne National Laboratory[4]. The main objective of this meeting was to introduce the different groups to each other and to agree on a plan for jumpstarting the collaboration. Each of the physics groups gave an overview of the overall scientific mission of the group and the design principals and structure of their software and hardware infrastructure. Carl Kesselman presented a strawman architecture for a Virtual Data Grid and outlined a plan for the first year. A key element of this plan was a series of meetings between each of the four Physics experiments and small groups of Computer Scientists to discuss in details experiment specific technical and conceptual aspects of Virtual Data Grids.

More than a dozen such meetings took place in the following six months. The format, dynamics and frequency of these meetings varied from one experiment to the other. However, they all contributed to a collection of extremely useful documents – use cases and requirements on the physics side and a Virtual Data Grid architecture document on the computer science side. Two meetings in December 2000 – one only of computer scientists and one of representatives from the four experiments and the computer scientists – lead to the formulation of the GriPhyN data grid reference architecture.

Our second "all-hands" meeting[5] was held April 9-11, 2001 at USC/ISI in Marina del Rey. More than 60 members of the collaboration attended the three-day meeting, plus a few visitors from other projects and the External Advisory Committee. The first day was devoted to presentations and discussions on the requirement and architecture documents. Early results from the computer science efforts were reported on the second day. Three planning working groups took place in parallel on the morning of the third day. Each group focused on a key component of Virtual Data Grids. The afternoon was devoted to two invited talks that challenged the attendees with visionary ideas about distributed computing and data management. Details and the contents of almost all talks and supporting reference[6] documents can be found at the meeting web site.[5]

Another Computer Science – application group meeting is planned for August 2-3, 2001 in Chicago. The next all-hands meeting will be held the week of October 15, 2001 in California.

2.3 Testbeds

The activities for developing and deploying testbeds to test components of the GriPhyN architecture have been pursued separately by the different application groups, as discuss in the subsections in Section 2.3. The testbeds for CMS and ATLAS are integrated with the production activities of their international collaborations and are expected to draw on these international facilities for limited GriPhyN tests next year. These application testbeds will be operated independently for the near future. We are exploring the idea of conducting occasional wide scale tests that would require two or more testbeds working coherently, but we anticipate that these kinds of tests would take place after 2002 when more tools are available.

2.4 Research activities

2.4.1 Computer Science

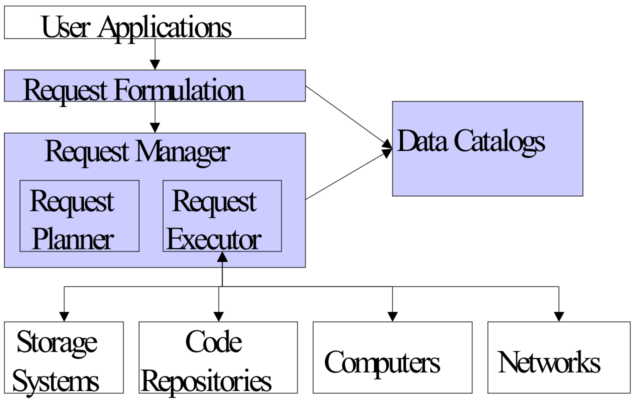
Computer science research during the first nine months of the GriPhyN project has focused on two main activities:

- definition of a baseline virtual data grid architecture
- research on basic virtual data grid technologies, such as scheduling, execution management, and virtual data instantiation.

We review the progress in each of these areas in the following sections.

2.4.1.1 Virtual Data Grid Architecture

Architecture development is a focused activity whose goal is to define a virtual data grid reference architecture that can guide the development of the virtual data toolkit as well as provide a framework for performing initial application experiments. We have produced two main documents: “A Data Grid Reference Architecture” and “Representing Virtual Data: A Catalog Architecture for Location and Materialization Transparency.” Contributions to these documents were made by many of the participating GriyPhyN institutions.



Our initial reference architecture is based on existing data-grid technologies including metadata catalogs, replica catalogs, software catalogs, Grid resource and data management protocols. By heavily leveraging existing data grid components, we anticipate rapid construction and deployment of a virtual data grid testbed, which will enable application experiments. The figure to the left illustrates the main components of the reference data grid architecture. One key element is the specification of a set of catalogs used to determine if virtual data has been materialized and if not, how to locate and invoke the transformations required to generate the requested data. A significant issue yet to be resolved is how to “name” virtual data so that its materialization state can be determined. A number of mined. A number of alternatives are currently being investigated, including the use of application or domain-specific naming strategies that leverage existing naming conventions within an experiment.

2.4.1.2 Research in Virtual Data Technologies

To date, research activities on various specific virtual data technology areas have been largely decoupled. The reference architecture will provide a integrating framework for many of these efforts, and will form the basis of cross-project collaboration. Specifically, during the next year, we anticipate that more mature research activities will be integrated into the overall system architecture and be deployed as part of our testbed environment for evaluation within the context of application experiments. We highlight some of these ongoing research activities below.

Agent based scheduling. Request management in virtual data grid environments is complicated by the complexity of the requests, the potential duration of request execution, and the dynamics of the underlying execution environment, including component failure. For this reason, request planning and execution management for virtual data grids must be very flexible, and capable of preplanning request execution in the case of failure. A promising approach is to exploit Intelligent Agents as a technology for implementing request management. At USC/ISI, we have initiated an investigation of this approach using the Electric Elves software agents as a framework for information gathering, planning and execution monitoring. To date, we have demonstrated the integration of Grid resources into the Elves environment and have used KQML (Knowledge Query and Manipulation Language) based commands to initiate Grid operations. Our next step will be to build a Agent knowledge base that performs intelligent recovery from failed data-movement operations.

Improving Query Performance. Research at UC Berkeley has been focused on developing techniques to reduce latency and improve the interactive performance of access to virtual data products. The work performed to date has focused on two promising directions: 1) Query relaxation and associated caching techniques, and 2) dynamic prefix caching for large files.

Query relaxation is aimed at providing more answers to user queries faster in cases where there are no existing (or readily available) virtual data products that exactly match a given query specification. In this situation, the query relaxation module looks for cached data products that may overlap or be closely related to the requested data. The user is presented with a menu of such products and in indication of the expected time required to obtain them. He/she may then choose to view one of these alternative products while waiting (or perhaps before asking) for the production of a high-latency data product. The ability to relax queries changes the caching problem somewhat, as now individual data products can "cover" multiple queries. Thus, caching policies to maximize query coverage have been developed and an initial simulation study has been performed.

Dynamic prefix caching aims to improve the cache hit rate by caching only the initial portion of large data items. When a request for such an item is received, the initial portion is sent to the user while the remainder of the item is retrieved from the slower location. If the prefix size is sufficiently large, then the latency of accessing the slower location can be completely hidden. This approach, however, may have the side effect of increasing the load on the slower site. Thus, a dynamic approach as been developed, that can balance load vs. cache hit rate by increasing the prefix size for highly popular items. A simulation study of this technique is currently under way.

XML-based data manipulation: The virtual data grid requires mechanisms for the creation of derived data products. At the same time, the virtual data grid requires the ability to query the collections of derived data products. An emerging standard for querying and manipulating XML files is Quilt. A standard version of Quilt is being developed called XQuery. Of great interest is whether XQuery can be extended to support scientific data types. This will make it possible to manipulate an XML encoded digital object directly within the same language that supports information discovery.

At SDSC, Xufei Qian, under the direction of Amarnath Gupta, added data types to XQuery, added overloaded operators to support scientific manipulations, and interfaced the system to the Extensible Scientific Interchange Language (XSIL). XSIL is an XML markup language used to describe LIGO data, developed by Roy Williams (Caltech). The resulting system was used to query and manipulate XSIL encoded data files. In particular, array operations were added to XQuery, including accession, element summation, array summation, and subsequence commands for concatenation. The resulting prototype demonstrated that it is possible to integrate query and manipulation commands for XML encoded digital objects.

Data Management: There are a number of research activities that are investigating various aspects of location transparency, specifically with respect to creating and managing replicas of materialized data. Work at the University of Chicago by Kavitha Ranganathan has used simulation studies to examine alternative replication and caching strategies on application response time and bandwidth requirements. A simulator has been developed that represents data movement within a hierarchically organized data grid, and 5 different replication strategies have been implemented and compared via simulation studies. Complimentary activities at Lawrence Berkeley Laboratory are using analytical and simulation studies to understand the performance of alternative caching strategies specifically targeted towards hierarchal storage managers. Additional simulation studies by Matei Ripeanu at the University of Chicago have examined TCP performance in typical replication scenarios involving parallel bulk transfer of many files. Our next step in the work will be to take the knowledge gained from simulation studies, and instantiate it in request planning strategies that can then be evaluated within the context of specific data access scenarios generated by one or more of the physics experiments.

Flexible Storage Technology: At the core of virtual data grid technology is the need for efficient and manageable physical data storage and movement. Along these lines, at Wisconsin we are currently developing NeST (Network Storage Technology). NeST software transforms a commodity workstation or PC running a commodity operating system into an easy-to-administer and high-performance storage appliance.

To operate effectively in a Grid environment, NeST will support multiple data transfer protocols, enabling integration into local sites with specific protocol preferences (e.g., NFS). However, standard Grid lingua franca such as GridFTP will also be supported. One of the main challenges of NeST is the combination of these protocols into a single, efficient, and yet easy-to-maintain software base. We are also developing the necessary software technology to allow NeST to configure itself to the vagaries of the underlying software and hardware platforms. In particular, NeST can statically choose between multiple concurrency architectures depending on which is best on the given platform, and in the future NeST will be able to select the most appropriate file-access methods given a particular storage platform.

We are currently investigating the following additional research issues within the NeST domain: how to build generic support for third-party transfers and wide-area caching, how to integrate NeST effectively into an opportunistic environment, and how to incorporate resource limitation and management into the NeST framework in a robust and yet flexible manner.

Performance Monitoring and Analysis: The work at Northwestern University has focused on the development of an infrastructure, called Prophesy, to automate the process of generating analytical models. Prophesy automates the following processes: instrumenting of code, recording of performance data into a relational database, and developing analytical models based upon empirical data. Prophesy also accepts data generated by other tools such as FMMPI, gprof, pixie or Rabbit. The data from these tools is placed in the performance database, which includes information about the context in which the data was collected as well as the data. This data is used by the data analysis component to produce an analytical performance model at the level of granularity specified by the user, or answer queries about best implementation of a given function.

We are also developing a performance requirements document for the GriPhyN project. The goal of this document is to identify the performance monitoring and modeling requirements for the four HEP experiments and the CS related activities. Our initial focus is on the ATLAS experiment; future work entails extending the requirements to other three HEP experiments. In terms of requirements, this document attempts to identify (1) what system components and application features require monitoring and recording of performance data and (2) what performance models are needed and the required accuracy. Once these requirements are identified, we can leverage from current work with performance tool development to identify the suite of tools needed.

2.4.2 Virtual Data Toolkit

In preparation for the release of the first version of the Virtual Data Toolkit (VDT) by the end of the year, the Globus and Condor Teams focused on adapting, enhancing and hardening existing software tools. The original timetable and scope of this version were modified to reflect the significant reduction in first year funding allocated to the VDT activities. The VDT work was guided by the requirements and timelines gathered from the experiments and the design and implementation principals established by the GriPhyN architecture document. Catalogs play a key role in this architecture. A number of different data catalogs were prototyped and carefully evaluated and tested. All catalogs use a common software foundation provided by the Globus Toolkit. The performance and robustness of the GridFTP component were also addressed. We expect this component to be the “work horse” of any Data Grid and therefore pay special attention its performance and robustness. The runtime support infrastructure of Condor was enhanced to fully utilize the capabilities of GridFTP and the security infrastructure of Globus. New features were added to the Directed Acyclic Graph Manager (DAGMan) of Condor and extensively tested. The DAGMan provides advanced job control services and is expected to serve as the agent responsible for dynamically materializing Virtual Data. Early versions of VDT components were successfully used for distributed production and reconstruction of simulated events for CMS.

2.4.3 ATLAS Research Activities

ATLAS-GriPhyN activities proceeded along two fronts. First, ATLAS physicist-developers and GriPhyN computer scientists met on two occasions[7] to discuss computing methods in high-energy physics, the architecture and status of the Athena[8,9,10] analysis control framework, and virtual data concepts specific to the Athena framework. A prototype virtual data query use case was developed, and development of associated Athena and ATLAS database software has been incorporated into the ATLAS Software and Computing Project Grid Plan[11]. Second, several more practical projects were undertaken: grid testbed development, prototype Tier 2 center selection and development, grid account registration software, and grid access software.

The testbed development proceeded throughout the year, with a kick-off workshop[12] at the University of Michigan, followed by bi-weekly meetings. There are five university sites participating in the activity: Boston University, Indiana University, University of Michigan, Oklahoma University, and the University of Texas at Arlington; three national laboratories participate: Argonne National Laboratory, Brookhaven National Laboratory, and Lawrence Berkeley National Laboratory. Approximately 15 users are participating in this development. Each site has contributed existing facility resources to be dedicated for this purpose, and site administrators have deployed current releases of grid toolkits including Globus and Condor. Accounts have been created and grid certificates exchanged for the current group of testbed participants. Each site has installed a complete ATLAS software environment, and work is ongoing to develop an ATLAS “kit” for simple distribution and update to future grid nodes. The primary applications to be run on the system in 2001 include fast Monte Carlo simulations executed within the Athena control framework, and full detector simulations using legacy FORTRAN codes. The infrastructure will be used during Mock Data Challenges in 2002 and 2003 to validate the LHC Computing Model[13].

Indiana University and Boston University were selected through a competitive process as the prototype Tier 2 centers for the U.S. ATLAS-GriPhyN project. Development of the first center at Indiana University has begun. A system consisting of a 16-node (32 processors) Linux farm, 500 GB RAID disk, plus access to HPSS hierarchical storage facilities for archival data has been deployed. The system is managed by professional IT staff in a machine room adjacent to the I2/Abilene Network Operations Center (NOC) in Indianapolis. Current releases of grid packages have been installed. A dedicated AFS server ring will be installed this August.

GriPhyN and ATLAS users will use a secure facility to support grid account request and authorization. GRIPE[14], (Grid Registration Infrastructure for Physics Experiments) was implemented by four computer science graduate students at Indiana University under the supervision of Profs. Gardner and Bramley. The system facilitates account creation across all administrative domains. GRIPE users choose trusted sponsors and grid site-specific resources and submit account requests through the system. Site administrators interact with GRIPE to communicate to users and exchange Globus certificates. The system is being tested this summer by the testbed working group, and is the primary mechanism for account creation at the Tier 2 centers.

A web-based portal for ATLAS users is currently under development. GRAPPA[15] (Grid Access Portal for Physics Applications) will provide ATLAS and GriPhyN users a single point of access to grid resources. The system provides a simple utility for job submission to Condor flocks and other available grid resources. The user interface will use script management tools based on Common Component Architecture Toolkit (CCAT) software being developed by Profs. Bramley and Gannon at Indiana University. The first prototypes will be available during Summer 2001.

Summary of meetings held related to ATLAS-GriPhyN project, with strong participation from GriPhyN personnel:

- 2nd International ATLAS Grid Workshop[16], CERN, Geneva, Switzerland, September 29-30, 2000.
- 3rd International ATLAS Grid Workshop[17], CERN, Geneva, Switzerland, January 25, 2001.

- U.S. Large Hadron Collider Common Projects Meeting, CERN, Geneva, Switzerland, January 21, 2001.
- U.S. ATLAS Grid Testbed Meeting, University of Michigan, February 3-4 2001
- HENP Networking Meeting, Indiana University, Indianapolis, June 1-2, 2001

2.4.4 CMS Research Activities

The CMS contingent of the GriPhyN collaboration, which includes researchers from Caltech, Florida, FNAL, UC San Diego and Wisconsin, have been active in several GriPhyN tasks. These activities are described below.

A prototype Tier1 computing center has been installed, configured and deployed at FNAL, as have Tier2 prototypes in California (November 2000) and Florida (June 2001). The California (Caltech/UCSD) Tier2 prototype[18] is fully deployed as a WAN cluster, with one half of the capacity at Caltech, and the other half at UCSD. Both the California and Florida systems are based on many dual slave nodes (Pentium III CPUs) running Linux.

Large scale distributed simulations, facilitated by the use of Globus and Grid middleware, have been completed. These include a prototype simulation run for the NSF Distributed Terascale Facility, and runs using resource allocations on Caltech's HP supercomputers, Wisconsin's Condor farm, and on NCSA's facilities.

Distributed systems modeling, in which the MONARC simulation[19] tool has been employed to model site and workload distributions for various hypothetical and real distributed computing systems. In particular, load-balancing schemes that employ Self Organizing Neural Networks (SONN[20]) were developed and successfully demonstrated. In a related activity, an agent-based task execution scheme has been proposed and is being evaluated. This scheme uses JINI to implement distributed services that are dynamically interconnected on the network, which is used as a fabric on which mobile agents carry a payload of tasks for the system to execute.[21]

An execution service for distributed processors has been developed. The prototype of this service is installed on the California Tier2, and allows the management of large numbers of lengthy tasks whilst conserving network bandwidth, and tolerating partition failures due to, for example, network outages. In particular, the prototype service is being successfully used to control groups of tasks that execute on both the Caltech and UCSD halves of the Tier2.[22]

The Grid Data Management Pilot (GDMP), a Globus-based tool for object database and file replication in the wide area, has been further developed and deployed in successfuly data replication tasks between CERN, FNAL, Caltech and UCSD. GDMP is now using HRM from LBNL for the ENSTORE storage system, and is being integrated with HPSS. Much of this work is being carried out in collaboration with members of the EU-DataGrid project.[23]

Considerable time and thought has been expended on developing a detailed picture of the requirements for Virtual Data in the CMS experiment. This has resulted in a document that describes the current thinking in this area, a sort of handbook for computer scientists who are helping to design and construct the global Grid-based computing system for the experiment.[24]

In the area of physics analysis studies, effort has been focused on the design and characteristics of the event "Tag" data. Each "Tag" event object contains a summary of the salient quantities from each event, and is to be used in rough primary selections over the whole event sample. As such, access to the Tags must be fast and efficient, and there is a trade-off between these requirements and the desire to include as many useful event parameters as possible. One aspect of the Tag studies has been the use of bit-sliced indexes for fast selections. Another has been the comparison of access speeds with the Tags stored in a traditional, versus an Object, database.

Another study in the area of physics analysis has been the development of a prototype remote analysis system for physicists[25]. This prototype is a browser-based system that operates as a thin-client which talks to a server containing the latest versions of the CMS reconstruction and analysis software. By utilizing Java, we have been able to deploy the prototype on "bare-bones" laptops that do not need any CMS software locally installed. The server side uses a mixture of Java and C++ programs.

As part of the continued tracking of network technology, with a view to understanding what can be deployed in the WAN fabric for the experiment, efforts continue to improve monitoring of existing links, and to evaluate new technologies as they appear and fall in price. As an example, we have installed powerful network monitoring systems at CERN, the Chicago PoP, and in Caltech. These are dual processor systems with substantial RAM and equipped with Gbit ethernet interfaces. We have also been investigating maximizing the throughput over the CalREN2 and NTON links between Caltech and SLAC and other WAN links between Caltech and CERN. This involves careful tuning of the TCP parameters, and coordination between all the various network providers on the route that links the end sites.

2.4.5 LIGO Research Activities

The LIGO work has focused on interactions between LIGO at Caltech, LIGO at Milwaukee, the USC Information Sciences Institute (ISI), and Caltech CACR. The initial project was to build a virtual data service specialized for LIGO data. Given a request that is a combination of a time interval and a set of channels, the service checks the Replica Catalog for the data that it needs, or else goes to an archive for raw data. Data in the replica catalog may match the request exactly, or be easily created by using an algorithm from the Transformation Catalog. If not, data is brought from archive and transformed to satisfy the request. The current testbed is installed at ISI, but soon it will connect to the tape archive at Caltech CACR as well as the near-real-time archive at LIGO Hanford. This project will be published in Springer *Lecture Notes in Computer Science* in June 2001.[26]

New work under development concerns data replication: Caltech acts as a Tier 0 center for LIGO, and the UW Milwaukee facility as a Tier 1 center. At Milwaukee, they are using a large amount (6 TB) of cost-effective non-RAID disk storage as a data cache, with a Beowulf system for gravity-wave searches. Globus technology (GridFTP) will be used to automatically rebuild the cache whenever disk failures occur Using Globus and Condor technology including CondorG and GridFTP, data is being automatically replicated from the CACR HPSS archive to Milwaukee. If the files at Milwaukee become damaged or are missing entirely, the scheduled CondorG job automatically detects the problem and connects to the Caltech archive to transfer the necessary data and correct the problem.

We are also using GridFTP in an initial project to search for coincidence events between the US LIGO and French-Italian Virgo gravitational wave observatories. Clearly an apparent astrophysical signal will be far more significant if detected in coincidence between independent experiments. We have begun to set up infrastructure to exchange certain channels from both Hanford and Livingston with Virgo: Gravity-wave, seismometers, and electromagnetic conditions at the observatories. Initially, since neither observatory is yet fully commissioned, we intend to look for coincidence in these latter channels.

2.4.6 SDSS Research Activities

Over the last year the SDSS project has been preparing for its first public data release. This required an extremely concentrated effort to reorganize all of its data products, and to create an environment where every bit of the approximately 3TB of data is fully accessible, and documented. This was achieved by a combination of databases and file-servers. The whole process turned out to be extremely useful in order to understand how can massive amounts of data be delivered for further usage. As discussed in the proposal, we are planning to build several pilot virtual data usage cases – each will depend on an efficient data access layer. This is in place as of June 5.

We have also experimented in comparing Object Oriented and Relational databases, in collaboration with Jim Gray (Microsoft BARC). We have now the same data in two systems, running on identical hardware, one using Objectivity/DB, the other MS SQL Server 2000. Over the next few months we will perform heavy comparisons between the two systems, running identical queries, spanning most of the usage scenarios.

We have done a lot of work in understanding how to speed up sequential IO performance on Intel-based hardware. We have confirmed previous measurements, which have shown that hardware RAID levels off at around 60MB/sec. Using software RAID 0 on DELL and Compaq servers, we have consistently achieved over 250MB/sec read speeds, running database queries. The speed leveled off due to the memory bandwidth of the systems, the rest has still some reserves. Our aim for the Virtual Data experiments was to have a data pump of about 1GB/sec sequential IO, now this will be reachable with about 4 nodes instead of 20.

At FNAL we have made a lot of progress in building the Terabyte Analysis Machine (TAM). The Terabyte Analysis Machine is a Linux cluster of currently 5 machines, connected to a RAID disk via fibre channel. Pilot projects include: using the Grid Data Management Pilot (GDMP) to transfer the Sloan Digital Sky Survey database (SX) among collaborating institutions; (i) testing performance and reliability of the Global File System for concurrent and high data rate accesses to the shared RAID disk and, in particular, its interaction with the Objectivity Database System; (ii) testing farms of IDE disks to provide direct access to a several terabytes single node at low cost, for non I/O bound applications; (iii) developing the Distance Machine framework.

The Distance Machine framework allows specific analysis programs to transparently access large databases, making best use of sophisticated specific algorithms for indexing and partitioning the database. At the current stage of development, the Distance Machine is a library, implemented in C++, which uses Objectivity as database system. To guarantee the scientist flexibility in the choice of data analysis language, a Corba interface based client/server architecture is provided to the user. The transparent access to the large database embodied dataset is granted by embedding the objectivity reference to each data object within a wrapper, a wrapper whose standard array operators are overloaded. A vector of wrappers can be effectively treated as a vector of pointers to memory: this approach not only allows almost immediate extension of most existing standalone analysis applications to large Objectivity managed datasets, but also enables new applications to use the re-indexing and re-partitioning features of the framework to increase the efficiency of the data access.

2.5 Education and Outreach

In September 2000, the administration at UT Brownsville approved the hiring of a full-time tenure-track faculty member to serve as E/O coordinator for GriPhyN. This was a bonus, given that the budget provides funds to support a person at only the post-doc level for this position. The search for the E/O coordinator began in November 2000, and five top candidates from across the world were subsequently interviewed. In March 2001, the position was offered to Manuela Campanelli (then a postdoc at the Albert Einstein Institute in Potsdam, Germany), who will start in earnest as E/O coordinator at the end of August. Her teaching responsibilities at UT Brownsville during the next five years have been limited to allow her to devote the majority of her time to E/O activities for the project. Although the E/O component of GriPhyN has been effectively delayed a year because of the search for a full-time E/O coordinator, we feel that, in the long-run, the project benefits by having such a person in charge of these activities.

Even though Campanelli does not officially start until the end of August, she has already begun making contact with people involved in education and outreach for other major projects--e.g., EOT-PACI, Quarknet, ThinkQuest, the European Data Grid, and SDSS. She attended the Grid-Forum in Amsterdam and the GriPhyN All-Hands Meeting in April. This fall, Campanelli plans to develop a web page for GriPhyN E/O and submit a proposal requesting REU funds to support students during the summer months doing grid related activities. A short planning document, describing these and other planned activities for the next few years, is currently in preparation.

We also point out that UT Brownsville, which is a Hispanic-serving institution, has nearly completed the construction of a 96-node Linux cluster, which will be used primarily for LIGO data analysis. (Funds for this cluster were obtained from a different NSF grant: PHY-9981795.) Thus, another natural near term goal for E/O is to grid-enable this cluster, making it available for use as a testbed for some of the (to-be-developed) GriPhyN virtual data toolkits, and to introduce minority students at UT Brownsville to distributed computing and grid-related technology. An undergraduate student has already begun to learn how to install and run Condor on a single Linux box, in preparation for a full installation on the completed cluster.

We are also beginning to promote Grid applications using virtual data. For example, Roy Williams of Caltech is hosting a workshop[27] June 26 on Grid demonstrations, some of which will demonstrate the power of virtual data in a Grid environment.

2.6 Connection to other efforts and the wider community

2.6.1 Coordination with Other Data Grid Projects

From the start the participants in GriPhyN have recognized the need for close cooperation and collaboration with other data grid projects and with the wider scope of grid computing. Two other projects in particular require close coordination with GriPhyN because of the strong participation in them of experiments at the CERN Large Hadron Collider, as GriPhyN also has with CMS and ATLAS. These projects are the Particle Physics Data Grid in the U.S., funded by DOE, and the EU DataGrid project in Europe, funded by the European Union, and working closely with several national data grid projects in European countries. Many cooperative activities have already begun on several levels, many of them strongly technical, but also at the administrative and management level. For example: Ian Foster and Carl Kesselman are on the Architecture Task Force for the EU Data Grid project; Fabrizio Gagliardi from CERN is on the GriPhyN external advisory board; Foster, Kesselman, and others have visited CERN; and so on.

A strong effort at coordination of the data grid projects has begun at the top level, beginning with a meeting of leaders of all the HEP-related data grid projects before the first Global Grid forum meeting in Amsterdam in March, 2001. In addition to presenting their individual plans and organizations, the participants in that meeting set up a series of working groups to make plans and recommendations for practical coordination of the several projects. Working groups were established on joint testbeds and technical coordination. Strong agreement was reached that the projects should work toward a common or compatible architecture.

A second meeting on data grid coordination will be held in Rome on June 23, 2001. The results of this meeting will be made available on the GriPhyN web site.

2.6.2 Network Coordination

A effort has been started by a GriPhyN applications subgroup (Rob Gardner of Indiana and Harvey Newman of Caltech) with the aim of improving the networking capability for High Energy Physics (HEP) and Nuclear Physics (NP) applications over the next five years. These applications have massive domestic and international networking needs, requiring many 10 Gb/s links within a few years. The CMS and ATLAS subsets of GriPhyN are organizing these meetings, but the results will be applicable to other application groups, including those outside of GriPhyN.

The first meeting[28] took place at Indiana June 1-2, 2001 and was attended by HEP and NP representatives (including many from GriPhyN) as well as US Backbone (BB) and Local Service Infrastructure (LSI) providers. The meeting resulted from an agency-sponsored group of project managers for the U.S. LHC program[29]. Among the goals of the meeting were to provide the backbone network providers (from Internet2 and ESnet) with the maximum amount of information available regarding demands for infrastructure and services from the HENP community, and the timing of this demand; to provide the HENP community with a clear picture of the planned evolution of the backbone networks and the timing of this evolution; to discuss end-to-end performance issues, and to describe the current local site infrastructure for typical large scale data centers and university sites and determine the level of commonality among sites. Two working groups, with strong participation from both networking professionals and physicists, were formed to address requirements and “roadblocks”. An Internet2 working group will be formed, and future meetings of this networking forum will be held, the next to be held at the Fall Internet2 Member meeting, October 1-4, 2001, Austin, Texas.

2.7 External Advisory Committee meeting

The External Advisory Committee (EAC) consists of nine members, with Messina and Reed acting as co-chairs.

Paul Messina	Caltech	Director, Center for Advanced Computational Research
Bill Johnston	LBNL	Director, NERSC
Fabrizio Gagliardi	CERN	Project Leader, EU DataGrid

Jim Gray	Microsoft	Microsoft Research
David Williams	CERN	Former Head, CERN IT Division
Joel Butler	Fermilab	Former Director, Fermilab Computing Division
Dan Reed	NCSA	Director, NCSA Alliance
Fran Berman	SDSC	Director, SDSC
Roscoe Giles	Boston U.	Head of EOT-PACI

The EAC met with GriPhyN on April 12, 2001, immediately following the April 9-11 all-hands meeting. The committee charge, the talks presented at the meeting, and the EAC report are available online[30]. There were a number of detailed recommendations that we are attempting to incorporate in our planning for next year. One recommendation was that the EAC meet at the end of the next all-hands meeting in October, 2001 to provide additional needed support in our first year and to ensure that we maintain good progress at this early stage. We plan to accommodate this request.

3 Products resulting from our work

3.1 Publications

- *Identifying Dynamic Replication Strategies for a High Performance Data Grid*, Kavitha Ranganathan and Ian Foster. *Submitted to Grid Computing Workshop* 2001. http://people.cs.uchicago.edu/~krangana/papers/p_grid_computing.ps.
- *Peer-to-Peer Architecture Case Study: Gnutella*, Matei Ripeanu, *submitted to P2P2001*, 27-29 August 2001, Sweden. http://people.cs.uchicago.edu/~matei/PAPERS/P2P2001.ps.
- *File and Object Replication in Data Grids*, Heinz Stockinger, Asad Samar, Bill Allcock, Ian Foster, Koen Holtman, Brian Tierney, Proc. IEEE Intl. Symp. on High Performance Distributed Computing, IEEE Press, 2001.
- *Condor-G: A Computation Management Agent for Multi-Institutional Grids*, James Frey, Todd Tannenbaum, Miron Livny Ian Foster, Steven Tuecke, Proc. IEEE Intl. Symp. on High Performance Distributed Computing, IEEE Press, 2001. http://www.cs.wisc.edu/condor/doc/condorg-hpdc10.pdf
- *Grid Information Services for Distributed Resource Sharing*, Karl Czajkowski, Steven Fitzgerald, Ian Foster, Carl Kesselman, Proc. IEEE Intl. Symp. on High Performance Distributed Computing, IEEE Press, 2001.
- *A Virtual Data Grid for LIGO*, Ewa Deelman, Carl Kesselman, Roy Williams, Albert Lazzarini, Thomas A. Prince, Joe Romano and Bruce Allen, submitted to Lectr. Notes. Comp. Sci. (to be published), http://www.cacr.caltech.edu/~roy/papers/griphyn_ligo.pdf.
- *The Architecture of Scientific Software*, Boisvert, R., P. Tang, pp. 273-284, "Data Management Systems for Scientific Applications," Kluwer Academic Publishers, 2001.

- *Global Digital Library Development*, Chen, C., pp. 197-204, "Knowledge-based Data Management for Digital Libraries," Tsinghua University Press, 2001.
- *Data and Metadata Collections for Scientific Applications*, Moore R., and A. Rajasekar, European High Performance Computing and Networking (HPCN 2001), Amsterdam, Holland, June 2001.
- *Data Management for Grid Environments*, H. Stockinger, O. Rana, R. Moore, A. Merzky, European High Performance Computing and Networking (HPCN 2001), Amsterdam, Holland, June, 2001.
- *Gathering at the Well: Creating Communities for Grid I/O*, Douglas Thain, John Bent, Andrea Arpaci-Dusseau, Remzi Arpaci-Dusseau, and Miron Livny, to appear in the Proceedings of SC2001, 2001, http://www.cs.wisc.edu/~thain/library/community-sc2001.pdf.
- *Grid Data Management Pilot (GDMP): A Tool for Wide Area Replication*, Asad Samar, Heinz Stockinger, IASTED International Conference on Applied Informatics (AI2001), Innsbruck, Austria, February 2001.
- *A DataGrid Prototype for Distributed Data Production in CMS*, Mehnaz Hafeez, Asad Samar, Heinz Stockinger, to appear in VII International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT2000), October 2000
- *Models for Replica Synchronisation and Consistency in a Data Grid*, Dirk Düllmann, Wolfgang Hoschek, Javier Jean-Martinez, Asad Samar, Heinz Stockinger, Kurt Stockinger, to appear in 10th IEEE Symposium on High Performance and Distributed Computing (HPDC2001) , San Francisco, California, August 7-9, 2001.
- *Data Management in an International Data Grid Project*, Wolfgang Hoschek, Javier Jaen-Martinez, Asad Samar, Heinz Stockinger, Kurt Stockinger, 1st IEEE/ACM International Workshopon Grid Computing (Grid'2000), 17-20 Dec. 2000, Bangalore, India. ("distinguished paper" award)
- *A Distributed Server Architecture for Dynamic Services*, I. LeGrand et al., submitted to CHEP2001, September 2001, http://clegrand.home.cern.ch/clegrand/lia/DistributedServices_D2.pdf.
- *A Self-Organizing Neural Network for Job Scheduling in Distributed Systems*, I. LeGrand et al., ACAT2000 proceedings, http://monarc.web.cern.ch/MONARC/sim_tool/Publish/SONN/AI406_paper.pdf.
- *The MONARC toolset for simulating large network-distributed processing systems*, I. LeGrand et al., in Proceedings of the 2000 Winter Simulation Conference, http://clegrand.home.cern.ch/clegrand/MONARC/WSC/wsc_monarc.pdf.
- *The GriPhyN Project*, The GriPhyN Collaboration, to appear at CHEP2001, September 2001.
- *An International Virtual Data Grid Laboratory*, The iVDGL Collaboration, submitted to CHEP2001, September 2001.
- *Globus Data Grid Tools*, W. Allcock et al., to appear at CHEP2001, September 2001.

- *Dynamic Replication Strategies for a High Performance Data Grid*, Kavitha Ranganathan and Ian Foster, to appear at CHEP2001, September 2001.

3.2 Internal documents

- *GriPhyN Management Plan*[1]
- *Data Grid Reference Architecture*, I. Foster, C. Kesselman[31]
- *Representing Virtual Data: A Catalog Architecture for Location and Materialization Transparency*, Ewa Deelman, Ian Foster, Carl Kesselman, Miron Livny[32]
- *Virtual Data Research challenges*, Ian Foster[33]
- *CMS Virtual Data Requirements*, Koen Holtman et al.[24]
- *LIGO's Virtual Data Requirements*, Roy Williams et al[34]
- SDSS *Virtual Data Requirements*: *A Weak Lensing Map as a Prototype Virtual Data Set*, James Annis, et al.[35]

3.3 Web site

A web site[3] has been developed by a team at Florida and is being hosted at Argonne. The site has a description of recent news items, ads for open positions, an archive for all GriPhyN e-mail lists, links to GriPhyN meetings (including papers presented at the meetings and registration information), pointers to External Advisory Committee meetings, and links to documents and other relevant projects.

The web site has proven to be a very good coordination site for the project, so much so that its deficiencies have become noticeable over the last two months. We plan to make some significant adjustments to the site this summer to (1) add more internal documents and links to other projects, (2) improve navigability among the many pages, (3) add a calendar of meetings and other events and (4) add security so that certain areas can only be accessed by GriPhyN members. We will put together by September, 2001 the first Education/Outreach page, once our E/O Coordinator, Manuela Campanelli, starts her job at Texas, Brownsville. We expect that this initial page will have links to existing projects in which we participate (i.e., QuarkNet, EOT-PACI), but we plan to add much more over the next year.

4 Plans for next year (2001 – 2002)

We are near completion of a detailed project plan for the remainder of the project that reflects all that we have learned during the first 9 months. This project plan states clearly:

- The research goals, responsibilities, milestones, and reporting mechanisms for the various research subprojects.
- The contents and capabilities of the virtual data toolkit, over the course of the project.
- The nature of the joint work and experimental studies to be conducted with the four partner physics experiments, with quantitative goals for size of computations performed.
- The nature of all dependencies between GriPhyN and other projects, including PPDG, European Data Grid, and various other relevant DOE and NSF projects.

Activities and milestones are specified on a 6-monthly timescale through the end of 2002, and on a yearly basis thereafter.

We have referred to future plans throughout the preceding text, but in brief, we will in the next year:

- Instantiate our virtual data grid architecture in a working prototype and evaluate its performance and capabilities in various practical settings.
- Work with each of the four experiments to (a) achieve large-scale application and evaluation of basis data grid concepts (as we have already done with CMS), and (b) explore virtual data grid architecture concepts in practical settings.
- Advance our various research subprojects to the point where their ability to deliver results of direct relevance to our partner applications can be clearly determined.
- Develop and release a second revision of the virtual data toolkit.

Appendix I: Participants in Project (asterix marks institutional contact)

Name	Sex	Position	Duties
Argonne			
Lawrence Price*	M	Division Director	GriPhyN project management, ATLAS requirements, coordination with other data grid projects
Steven Tuecke	M	Software Architect	Grid Architecture
Michael Wilde	M	Technical Staff	Grid Architecture, replication
Veronika Nefedova	F	Technical Staff	Grid Architecture
Edward May	M	Physicist	ATLAS requirements and testbed implementation
Berkeley			
Michael Franklin*	M	Faculty	Research direction
Arie Shoshani	M	Scientist	Research direction
Asha Tarachandani	F	Graduate Student	Research on Caching
David Liu	M	Graduate Student	Research on Querying
Boston University			
James Shank*	M	Faculty	ATLAS applications

Saul Youssef	M	Postdoc	ATLAS applications
Caltech			
Harvey Newman*	M	Physics Faculty	GriPhyN project management, CMS applications leader; Grid and network systems
Julian Bunn	M	Physicist	Computational CMS applications; Grid Scientist and network systems
Takako Hickey	F	Computer Scientist	CMS Grid execution service
Albert Lazzarini	M	Scientist	GriPhyN project management, LIGO applications leader, liaison to LIGO
Roy Williams	M	Scientist	LIGO applications
Renate Bornheim	F	Computational Scientist	CMS systems integrator
Koen Holtman	M	Computer Scientist	CMS applications and software requirements; Object collection prototypes
Iosif Legrand	M	Computer Scientist	Distributed system simulations; Physicist Grid monitoring tools; distributed services architecture
Conrad Steenberg	M	Computer Scientist	Interactive Grid-enabled Physicist Client-server applications
Chicago			
Ian Foster*	M	Faculty	Project co-PI
Von Welch	M	Technical Staff	CS research on security and policy
Jens Voelker	M	Technical Staff	Virtual data toolkit
Kavitha Ranganathan	F	Grad Student	Research on replication strategies
Adriana Iamnitchi	F	Grad Student	Research on resource discovery
Fermilab			
Lothar Bauerdick*	M	Scientist	GriPhyN project management, CMS applications, liaison to CMS
Stephen Kent	M	Scientist	SDSS applications, SDSS requirements
James Annis	M	Postdoc	SDSS applications, SDSS requirements

Florida			
Paul Avery*	M	Faculty	Project PI
Richard Cavanaugh	M	Postdoc	CMS applications, performance monitoring
Dimitri Bourilkov	M	Scientist	CMS applications, simulations
Jorge Rodriguez	M	Scientist	CMS applications
Suchindra Katageri	M	Grad student	Performance monitoring
Harvard			
John Huth*	M	Faculty	GriPhyN project management, ATLAS requirements
Indiana			
Rob Gardner*	M	Faculty	GriPhyN project management, liaison to ATLAS, ATLAS applications; GRAPPA[15], GRIPE[14] development
Dennis Gannon	M	Faculty	Notebook component architecture interfaces to Condor and GRAPPA
Randall Bramley	M	Faculty	Notebook component architecture interfaces for GRAPPA; GRIPE system design and development
Shava Smullen	F	Computer Scientist	DCS components of GRAPPA
Daniel Engh (Sep. 1)	M	Postdoc	Athena/VDT interface
Lisa Ensmen	F	Postdoc	Athena, ATLSIM use cases and interfaces to GRAPPA
Kristy Kallback-Rose	F	IT Staff	GRAPPA project lead, Condor liason
Rob Kopenec	M	Grad student	Athena fast simulation analysis and GRAPPA interfaces
Johns Hopkins			
Alex Szalay*	M	Faculty	GriPhyN project management, liaison to SDSS, database access strategies, sequential I/O measurements, virtual data prototypes,
Northwestern			
Valerie Taylor*	F	Faculty	Performance monitoring

Jennifer Schopf	F	Faculty	Scheduling, resource management, ATLAS applications
Dheeraj Bhardwaj	M	Post-doc	Performance monitoring
U Pennsylvania			
Robert Hollebeek*	M	Faculty	Testbeds
Kevin Sterner	M	Scientist	Testbeds
San Diego			
Keith Marzullo*	M	Faculty	Fault tolerance
Reagan Moore*	M	Scientist	GriPhyN project management, data grids, virtual data
Xufei Quian	F	Grad Student	Virtual data
Chia-yen Shih	F	Grad Student	Fault tolerance
Southern California			
Carl Kesselman*	M	Scientist	GriPhyN project management, Computer Project Management, architecture, LIGO applications
Ewa Deelman	F	Scientist	Architecture, LIGO interaction
Amrish Kaushik	M	Grad Student	Scheduling
Gaurang Mehta	M	Grad Student	LIGO prototype, architecture
Ann Chervenak	F	Scientist	Replica management
Stanford			
Richard Mount*	M	Scientist	Liaison with PPDG, coordination with EU DataGrid
Texas, Brownsville			
Joseph Romano*	M	Faculty	GriPhyN project management, education/outreach, LIGO applications
Manuela Campanelli	F	Faculty	Education/outreach, LIGO applications

Wisconsin, Madison			
Miron Livny*	M	Faculty	GriPhyN project management, Virtual Data Toolkit
A. Arpaci-Dusseau	F	Faculty	Virtual Data Toolkit
R. Arpaci-Dusseau	M	Faculty	Virtual Data Toolkit
John Bent	M	Grad Student	Virtual Data Toolkit
Doug Thain	M	Grad Student	Virtual Data Toolkit
V. Venkataramani	M	Grad Student	Virtual Data Toolkit
Zach Millar	M	Programmer	Virtual Data Toolkit
Wisconsin, Milwaukee			
Bruce Allen*	M	Faculty	LIGO applications
Scott Koranda	M	Assoc Scientist	Data replication for LIGO & grid apps

Appendix II: CMS Information Release About Large-Scale Test

The Grid in Action:
GriPhyN Scientists use Condor and Globus to Simplify Research from Afar

This spring, a collaborative effort involving physicists and computer scientists from Caltech, the National Center for Supercomputing Applications (NCSA), and the University of Wisconsin-Madison produced a working example of the Grid in action.

Vladimir Litvin, a Caltech physicist, used a single application on his desktop at Caltech to submit and manage a distributed computation involving three institutions, two clusters of hundreds of computers at two different locations, a mass storage system, and two different batch scheduling systems—without having to manually log into a single remote system.

The collaboration was conducted by participants in the NSF-funded GriPhyN (“Grid Physics Network”) project, as part of the preparations for the CMS detector at the Large Hadron Collider (LHC) at CERN in Geneva, Switzerland. Expected to come online in 2005, the detector will probe fundamental forces in our Universe and search for the yet-undetected Higgs Boson. Even during this preliminary design phase of the experiment, the computational and data challenges are substantial, and many, many demanding computations are necessary.

Each of Litvin’s runs requires him to schedule hundreds of Monte Carlo detector response simulations to run on the University of Wisconsin’s compute cluster, then transfer each job’s roughly 1 GB of output to the UniTree mass storage system at NCSA, then run hundreds of jobs on

NCSA's new cluster to reconstruct physics from the simulated data, and finally transfer the results back to the UniTree mass storage system.

Last fall, as a first step towards automating this process, researchers from the Condor Team helped Vladimir use the Condor system to automatically manage the simulations at UW-Madison and initiate the data transfers to NCSA. But Litvin had to log into UW-Madison machines to start the process, and he had to log in and manually initiate the reconstruction phase at NCSA once he saw that the simulations at UW and the data transfers to NCSA had completed successfully. The entire process remained unwieldy and error-prone.

This April, using a Grid-enabled version of Condor, called Condor-G, along with the Globus infrastructure installed at all three institutions, Vladimir was able to submit a single "job" to his personal Condor pool at Caltech that marshaled the resources of each institution, monitored the progress of his jobs and data transfers at each, and notified him upon the completion of the entire run.

The Globus middleware deployed across the entire "Alliance Grid" operated by the NCSA-led National Computational Science Alliance provided remote access to computational resources and dependable, automated data transfer capabilities. Condor, layered over Globus, provided strong fault tolerance including checkpointing and migration, and job scheduling across multiple resources, serving as a "personal batch system" for the Grid.

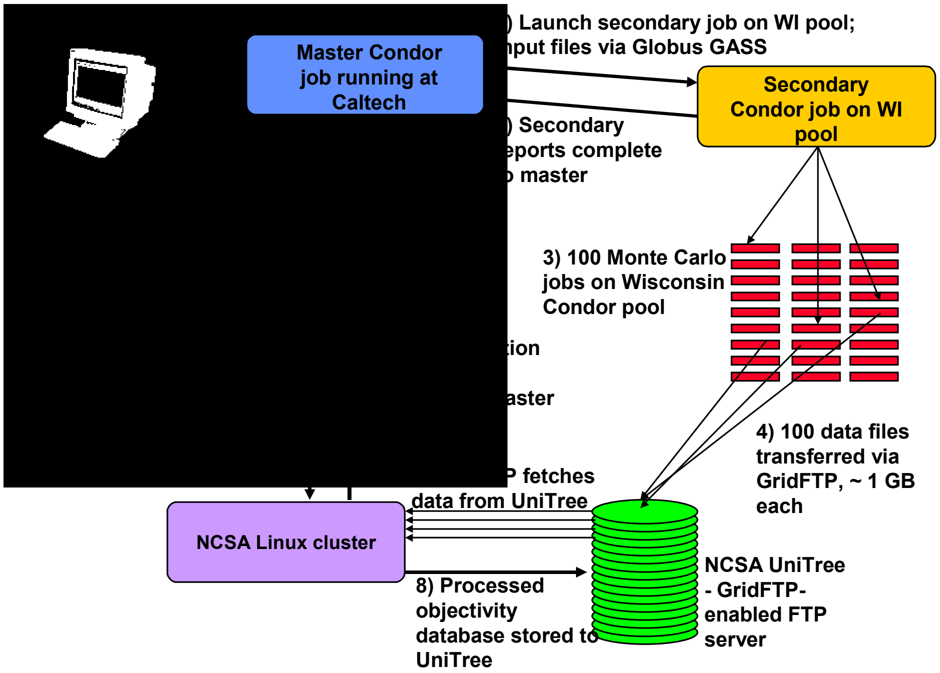
While this success represents a major advance for Litvin and his colleagues, it is just a first step towards the GriPhyN project's eventual goal of creating a national and international "Petascale Virtual Data Grid," that will allow computer, storage, and network resources at hundreds of sites to be applied to the analysis of data from CMS and other physics experiments—or from earthquake engineering, bioinformatics, and other disciplines. GriPhyN was awarded five year's funding in September 2000 to address this goal, and this early success demonstrates that it is already starting to put some of the necessary technologies in place.

For More Information

CMS: http://cmsinfo.cern.ch
Condor: http://www.cs.wisc.edu/condor
Condor-G: http://www.cs.wisc.edu/condor/doc/condorg-hpdc10.pdf
Globus: http://www.globus.org
GriPhyN: http://www.griphyn.org

Further Details

The following figure illustrates the resources and activities involved in the computation.


[redacted]

Master Condor job running at Caltech
) Launch secondary job on WI pool;
nput files via Globus GASS
Secondary Condor job on WI pool
) Secondary
eports complete
o master
3) 100 Monte Carlo jobs on Wisconsin Condor pool
ion
aster
4) 100 data files transferred via GridFTP, ~ 1 GB each
P fetches
data from UniTree
NCSA Linux cluster
8) Processed objectivity database stored to UniTree
NCSA UniTree - GridFTP-enabled FTP server

Each CMSIM simulation job is really two jobs; two executables are required to run. For each CMSIM job, then, there is an ‘A’ job and a ‘B’ job. Further, for each of the CMSIM jobs there is a ‘pre’ script and a ‘post’ script which is run. A Condor scheduling component called DAGMan is used to automate the process of running the pre script, then the A job, then the B job, and finally the post script for each of the CMSIM jobs. Condor DAGMan runs on the cluster front-end node at Wisconsin (beak.cs.wisc.edu) under the Condor “scheduler” universe. This means the DAGMan executable managing the jobs executes on beak, as do the pre and post scripts spawned by DAGMan.

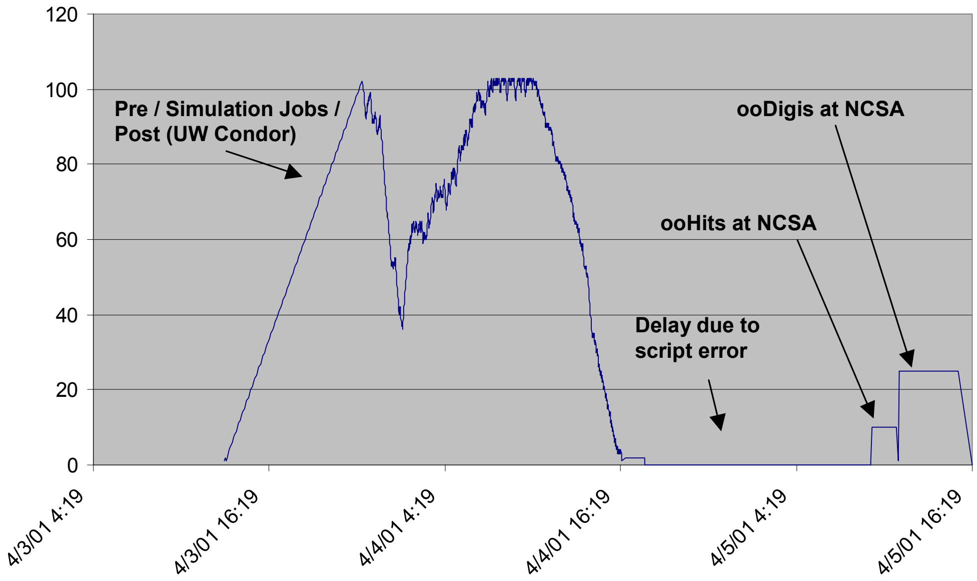
To start the Condor DAGMan executable on beak, Vladimir used Condor-G and a second Condor DAGMan running on his machine at Caltech. The DAGMan running at Caltech submitted a job to that local Condor-G and specified the Condor “globus” universe, which communicates with a special Globus gatekeeper on beak. This special gatekeeper on beak allows the executable to be run in the Condor scheduler universe on beak. The executable specified in the job submitted to the local Condor-G at Caltech is simply the Condor DAGMan job which needs to run on beak.

The DAGMan that runs on beak and which manages the CMSIM jobs requires a DAGMan DAG specification file, along with all the pre and post scripts and the Condor submit-description files for each individual CMSIM job. This collection of input files is sent to Wisconsin automatically by the Condor-G at Caltech using the Globus GASS service. The bottom line is that there is no need to connect to Wisconsin and start the Condor run from there. Condor-G and Condor DAGMan from Caltech manage the entire process.

Each post script run as part of each CMSIM job at Wisconsin uses GridFTP to transfer the generated data to the NCSA UniTree. Authentication is automatic, since the credentials are passed from Caltech to Wisconsin via Condor-G and the Globus gatekeeper on beak.

When the last CMSIM job has finished at Wisconsin and the last data file has been moved to the NCSA UniTree, the DAGMan running at Wisconsin exits and the DAGMan running at Caltech starts a globusrun to initiate the processing at NCSA. The globusrun is sent to the gatekeeper on posic, the experimental Linux cluster at NCSA. The job that runs at NCSA will run in the "dtf" queue which has been setup on posic so that the job actually runs on the new 1GHz nodes that are part of the new NCSA platinum Linux cluster which will soon go into production. The last part of the job running at NCSA again uses GridFTP to transfer the output to NCSA UniTree.

The following figure shows the number of jobs (both compute and data transfer) running at any one time during the computation. The delay in the middle is due to a system misconfiguration; when this was corrected, the entire computation could be resumed, illustrating the fault tolerance of the Condor-G framework.



References

[1] GriPhyN Management Plan v1.3, November 2000, see http://www.phys.ufl.edu/~avery/griphyn/manage/manage_v1.3/.

[2] The GriPhyN External Advisory Committee report is at http://www.griphyn.org/meeting/.

[3] GriPhyN web site, http://www.griphyn.org/.

[4] Web site for October 2000 all-hands meeting, http://www.griphyn.org/news/meetings/2000oct02.html.

[5] Web site for April 2001 all-hands meeting, http://www.griphyn.org/news/meetings/agenda.html.

[6] Supporting documents for April 9-11, 2001 meeting, http://www.griphyn.org/news/meetings/2001apr09.html.

[7] ATLAS-GriPhyN Applications Meeting, Argonne National Laboratory, December 15, 2000; GriPhyN All-Hands Meeting, April 9-11

[8] ATHENA Architecture and Framework, ATLAS Collaboration, http://atlas.web.cern.ch/Atlas/GROUPS/SOFTWARE/OO/architecture/

[9] Gaudi Architecture and Framework, LHCb Collaboration, http://lhcb-comp.web.cern.ch/lhcb-comp/Components/html/GaudiMain.html

[10] LHCb Collaboration, http://lhcb.web.cern.ch/lhcb/

[11] Grid Computing in U.S. ATLAS: http://atlassw1.phy.bnl.gov/Planning/usgridPlanning.html

[12] U.S. ATLAS Grid Testbed Meeting, University of Michigan, February 3-4 2001; http://www-personal.umich.edu/~myers/gridmtg/

[13] Report of the Steering Group of the LHC Computing Review (Hoffman Panel), S. Bethe (Chair), CERN/RRB-D 2001-3, February 22, 2001.

[14] GRIPE: http://lexus.physics.indiana.edu/griphyn/gripe/gripe.html

[15] GRAPPA: http://lexus.physics.indiana.edu/griphyn/grappa/index.html

[16] http://documents.cern.ch/age?a00380

[17] http://documents.cern.ch/age?a0153

[18] Description of Caltech/UCSD Tier2 site, http://pcbunn.cacr.caltech.edu/Tier2/Tier2_Overall_JJB.htm.

[19] MONARC home page, http://monarc.web.cern.ch/MONARC/.

[20] Self Organizing Neural Networks, http://monarc.web.cern.ch/MONARC/sim_tool/Publish/SONN.

[21] JINI distributed services, http://pcbunn.cacr.caltech.edu/uscmssw/DistributedServices_D2.pdf.

[22] Execution service activity, http://pcbunn.cacr.caltech.edu/uscmssw/TakakoExecService.pdf.

[23] EU DataGrid project, http://cmsdoc.cern.ch/cms/grid/.

[24] CMS requirements document, http://kholtman.home.cern.ch/kholtman/tmp/cmsreqsv9.pdf.

[25] http://pcbunn.cacr.caltech.edu/uscmssw/RemoteAnalysisConrad.ppt.

[26] *A Virtual Data Grid for LIGO*, Ewa Deelman, Carl Kesselman, Roy Williams, Albert Lazzarini, Thomas A. Prince, Joe Romano and Bruce Allen, Lecture Notes. Comp. Sci. (to be published), also http://www.cacr.caltech.edu/~roy/papers/griphyn_ligo.pdf.

[27] Workshop on Demonstrations of the Grid, part of the High Performance Computing and Networking conference http://www.cacr.caltech.edu/~roy/hpcn/grid.html, Amsterdam, NL, June 26.

[28] June 1-2, 2001 networking meeting at Indiana, http://www.transpac.org/hep-np_meeting_06-01/index.html.

[29] U.S. Large Hadron Collider Common Projects Meeting, CERN, Geneva, Switzerland, January 21, 2001

[30] EAC meeting April 12, 2001, http://www.griphyn.org/news/meetings/2001apr12.html.

[31] Data Grid Reference Architecture, http://www.griphyn.org/news/meetings/2001apr09/foster_dgra_jan01.doc.

[32] Representing Virtual Data, http://www.griphyn.org/news/meetings/2001apr09/foster_representing_virtual_data_jan01.doc.

[33] Virtual Data Research challenges, http://www.griphyn.org/news/meetings/2001apr09/foster_research_challenges_jan01.doc

[34] LIGO Virtual Data Requirements, http://www.griphyn.org/news/meetings/2001apr09/williams_ligo_requirements.html.

[35] SDSS Virtual Data Requirements, http://www.griphyn.org/news/meetings/2001apr09/annis_sdss_requirements.html.