*A Proposal Submitted to the National Science Foundation 2004 ITR Program*
*NSF 04-012; February 25, 2004*

# GECSR:
# A Global Grid-Enabled Collaboratory for Scientific Research


Julian Bunn (co-PI), Philippe Galvez
California Institute of Technology

Daniel Elvira, Daniel Green, John Wormersley
Fermi National Accelerator Laboratory

Heidi Alvarez
Florida International University

Deborah A. Agarwal
Lawrence Berkeley National Laboratory

Raymond Brock
Michigan State University

Yasar Onel, Shaowen Wang
University of Iowa

Nicholas Hadley, Shuichi Kunori (co-PI), Andris Skuja
University of Maryland

Eric Dey, Tom Finholt(co-PI), Steven Goldfarb, Homer A. Neal(PI), Charles Severance
University of Michigan

Kaushik De
University of Texas at Arlington

# Project Summary

## ITR: A Global Grid-Enabled Collaboratory for Scientific Research

The High Energy and Nuclear Physics (HENP) communities face unprecedented challenges as they seek to effectively execute the next generation of experiments involving hundreds to thousands of collaborators around the world. The very success of these experiments will critically depend on sustaining vigorous geographically distributed collaborations that are coherent, efficient and effective while accessing, processing, and sharing Petabyte-scale data. This has never been done before and there are serious concerns that the science emerging from these projects will be limited by the absence of robust tools to make such large-scale, highly interactive collaborations possible. We seek to address these challenges by developing and deploying a collaborative environment, called the ***Grid-Enabled Collaboratory for Scientific Research*** (GECSR), that both integrates existing technologies and develops innovative new technologies that will be readily adopted by the target communities.

GECSR will combine the best open-source applications from within the HENP communities (and from allied Grid and National Middleware efforts) within a common portal interface. To ensure that the capabilities of the GECSR meet the needs of target users, we will follow an iterative development model that in each cycle will include systematic assessment of user requirements, tool development and deployment consistent with requirements, and evaluation to determine whether deployed tools will satisfy requirements. A distinguishing feature of this approach is that the combination of expertise in physics, computer science, and social science will greatly enhance the chances of both technological and sociological success. Institutions in our proposal have established track records of international leadership in each of these areas.

The recent report of the NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure identifies five key service categories that will provide a foundation for the comprehensive knowledge environments that will enable individuals, teams and organizations to revolutionize scientific practice. The HENP community, working with computer science and communities in astrophysics and other disciplines, has addressed four of these services, based largely on common underlying middleware, in the widespread deployment of grid-enabled high performance computing resources, and through several data grid projects. The proposed Grid-Enabled Collaboratory for Scientific Research will provide the collaboration services, the fifth of these service elements of cyberinfrastructure, that is required for the HENP community to fully realize a functionally complete environment, with the potential to significantly transform the conduct of research.

**Broader Impact**

This proposal is distinguished by a tight integration between **1)** the science of collaboratories**, 2)** a globally scalable working environment built on the foundation of a powerful fully functional set of working collaborative tools, **3)** an agent-based monitoring and decision-support system that will allow collaborating scientists to perform data intensive analysis tasks efficiently, and **4)** an education and outreach agenda that is inclusive of minorities, and reaches high-school physics programs through vehicles such as QuarkNet and CHEPREO and success models such as ThinkQuest. Assessment of the methodology of scientific collaborations and the iterative evaluation of the tools by a team independent of the developers will be a critical element ensuring the success of the proposed work and insuring its value beyond HENP.

The broader implications of this proposal, and its mission to grant full partnership to groups in all world regions, are driven by **(1)** the right of groups and governments that have contributed to the global Collaborations to share and collaborate in the data analysis, and thus in the process of search and discovery that is at the root of all basic research. **(2)** the need to justify global projects, and global Collaborations as a valid and vital means of conducting future leading-edge scientific research and other endeavors, without exploitation of the poorer nations and world regions and **(3)** the need to develop effective means of collaboration and cooperation among sub-groups with vastly different academic and cultural backgrounds, and especially to allow effective collaboration by university-based students in the native cultural milieu of their home countries. Six percent of this proposal's budget is allocated to Education and Outreach that specifically targets minority populations: this should produce a sea change in how high energy physics experimentation, and the necessary global collaboration, impacts the quality of science education and addresses the needs of underserved populations across the US, and around the world.

# ITR: A Global Grid-Enabled Collaboratory for Scientific Research

The High Energy and Nuclear Physics (HENP) communities face unprecedented challenges as they seek to effectively execute the next generation of experiments involving hundreds to thousands of collaborators around the world. The very success of these experiments will critically depend on sustaining vigorous geographically distributed collaborations that are coherent, efficient and effective while accessing, processing, and sharing Petabyte-scale data. This has never been done before and there are serious concerns that the science emerging from these projects will be limited by the absence of robust tools to make such large-scale, highly interactive collaborations possible. We seek to address these challenges by developing and deploying a collaborative environment that both integrates existing technologies and develops innovative new technologies that will be readily adopted by the target communities. That is the purpose of the proposal presented herein.

The HENP communities have a strong history of innovative use of information technology to support collaborative activity, as indicated by HENP leadership in development of the world's most advanced Grid-based systems [1,2,3,4], in deployment of scalable videoconferencing tools [5], and in the use of Grid-enabled analysis environments [6]. However, despite this important progress, there is no unifying collaborative framework that links these key capabilities together. Recent work has shown that heterogeneous collaborative tools – i.e., where physicists must shift from one application to another as they perform different tasks – can introduce new sources of complexity that impede rather than aid collaborations [7]. We will seek to avoid this problem by adopting a holistic design approach that involves representatives from the end-user community in all stages of the research, development and deployment processes.

We propose to produce an integrated collaborative environment, called the ***Grid-Enabled Collaboratory for Scientific Research*** (GECSR), that will combine the best open-source applications from within the HENP communities (and from allied Grid and National Middleware efforts) within a common portal interface. To ensure that the capabilities of the GECSR meet the needs of target users, we will follow an iterative development model that in each cycle will include systematic assessment of user requirements, tool development and deployment consistent with requirements, and evaluation to determine whether deployed tools will satisfy requirements. A distinguishing feature of this approach is that the combination of expertise in physics, computer science, and social science will greatly enhance the chances of both technological and sociological success. Institutions in our proposal have established track records of international leadership in each of these areas.

The significance of the GECSR effort is underlined by the recent report of the NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure[8], which identified five key service categories that form the foundation for comprehensive knowledge environments to enable individuals, teams and organizations to revolutionize scientific practice. These include: high performance computation services; data, information knowledge management services; observation, measurement, fabrication services; interfaces, visualization services; and collaboration services. The HENP communities have generated coordinated activity within the first four of these categories– primarily through: the use of common underlying middleware; widespread deployment of grid-enabled high performance computing resources; data grid projects for management of data, information and knowledge [9]; instruments that can be monitored remotely through eLogs and other interfaces; and development of the Grid-enabled Analysis Environment [6]. In contrast, progress within the collaboration services category has not been at a comparable level. As a result, there is some risk that the important work done in the other areas may be diminished if physicists lack highly usable and standard collaborative interfaces that tie together the disparate computational, data, and instrument resources. The proposed GECSR fills this gap by creating a coherent community-sanctioned activity within the collaboration services category that is equal in caliber and level of coordination to ongoing efforts within the other four categories.

## *HENP's Collaboration Challenges*

The major HENP experiments of the next twenty years will break new ground in our understanding of the fundamental interactions, structures and symmetries that govern the nature of matter and space-time in our universe. Among the principal goals at the high energy frontier are to find the mechanism responsible for mass in the universe, to discover the "Higgs" particles associated with mass generation, to understand the unification of the fundamental forces of nature, to search for new symmetries and extra dimensions, and to explore the fundamental mechanism that led to the predominance of matter over antimatter in the observable cosmos.

The largest collaborations today, such as CMS[10] and ATLAS[11] who are building experiments for CERN's[12] Large Hadron Collider (LHC; [13] ) program, each encompass of the order of 2000 physicists from 150 institutions

in more than 30 countries. Each of these collaborations include 300-400 physicists in the US, from more than 30 universities as well as the major US HEP laboratories. The LHC experiments are expected to begin collecting data in 2007, with a high level of collaborative activity required up to that date and well beyond. The current generation of experiments now in operation and taking data at SLAC[14] and Fermilab (D0[15] and CDF[16]) are similar in scale to the US contingent of the next-generation experiments. Each of these experiments faces unprecedented challenges in terms of:

- The global extent and multi-level organization of the physics collaborations, leading to the need to collaborate and share data-intensive work in fundamentally new ways.
- The data-intensiveness of the work, where the data volume to be processed, distributed, accessed and analyzed by a major experiment are in the Petabyte ($10^{15}$ Bytes) range now, and are expected to rise to the Exabyte ($10^{18}$ Bytes) range within the next ten years.
- The complexity of the data, particularly at the LHC where the physics discovery potential is related to the very high intensity (luminosity) as well as the high energy of the collisions, such that ~20 interactions accompany the particle interaction of interest

The new paradigm of "Grids" and grid-computing[17] is thought to hold the key to meeting the collaboration, computing and data-management needs of HENP and there are significant efforts underway to explore and develop the grid toolkits and middleware that will be required for success in HENP. The complex problems of connecting and enabling resources like networks, computers and storage certainly justify the attention and support from the funding agencies and the HENP communities. However, these purely technical issues are equaled in complexity by the socio-technical problems of effectively and efficiently connecting and enabling physicists to do their science using emerging cyberinfrastructure – and these topics have not received as much attention from HENP physicists or the agencies. There is considerable risk in allowing this discrepancy to persist and the primary goal of this proposal is to redress this situation.

A particular problem for the organization and conduct of cyberinfrastructure-based physics is that past practice and experience provides only limited guidance for what should be done in the future. Up to the present, HENP physicists have performed experiments and analyses in tightly coupled cooperating groups. Day-to-day research was typically conducted in smaller teams (1-5 physicists) that work closely together and then share their results within a larger overarching collection of teams for verification and further analysis. The scale of these enterprises, combined with frequent geographic concentration of collaborators, meant that communication could be organized primarily through informal and formal face-to-face contact (e.g., hallway encounters or lab meetings). By contrast, the worldwide scope and large size of newer HENP collaborations, such as CMS and ATLAS, make dependence on face-to-face communication unrealistic. The geographic dispersion of these next generation collaborations highlights the need for novel tools and work practices. For example, while experiments may run twenty-four hours a day, physicists are most likely to be active during working hours in their local time zones. Working together under these circumstances will require collaboration tools that support both synchronous and asynchronous activity.

To summarize, we argue that a focused effort to develop a robust collaborative environment for large-scale HEPN experiments is critical, because physicists will receive their primary exposure to the other components of cyberinfrastructure via collaboration services – and if these services are deficient, then the potential benefit of the other capabilities will be reduced. The GECSR, then, is an effort to create a collaboration services element that is equal to the outstanding work already accomplished and underway in other physics cyberinfrastructure projects.

# Approach and Architecture

To ensure the highest probability of meeting user needs, we will employ an iterative development strategy that combines assessment, development, deployment, and evaluation within each release cycle [18,19,20,21,22]. Our goal is the creation of a common portal interface that provides a standard look and feel for uniform access to a diverse array of the best open source applications from within the HENP communities, and from related Grid and NMI efforts. The basis for the GECSR portal interface will be the Open Grid Computing Environment (OGCE) [23]. The advantage of the OGCE is that it provides a common context to plug-in specialized capabilities – such as features specific to the HENP communities – alongside a standard set of collaboration tools and services (e.g., announcement capabilities, persistent chat, shared calendars, role-based access control, and threaded discussions). At the University of Michigan, a variant of the OGCE is currently in use as an enterprise-wide course management

system with approximately 30,000 users per term. In the OGCE, the common context to add specialized features is achieved through the use of open-source components and standards, such as JSR 168, a java standard that defines elements called portlets for delivery of applications through portals. An advantage of the portlet approach, as implemented in OGCE, is that portlets are decomposed into their presentation components supported by persistent services accessed via a standardized API. In addition, the OGCE optionally extends login authentication to use the Grid [24] as an authentication provider, such that a proxy credential can also be stored by OGCE to allow portlets to perform Grid operations.

Users will organize activity in the GECSR portal environment through *sessions*, or combinations of people, data, and communication that are tracked by the system. For example, in the case of a research group with ongoing meetings a new session will begin by calling up archived data, metadata, and tools that the group has used in the past. A key feature of any session is that resources available to users will be consistent with preferences and access/authorization levels of everyone involved in the session. For instance, a session will not employ a given capability, such as audio recording, unless a group decision rule is satisfied (e.g., majority approval or unanimous approval). Sessions will consist of both synchronous and asynchronous activity. In the synchronous setting, participants in a session will be able to access a wide array of communication capabilities, including video and audio conferencing, supplemented by the ability to view and control common applications, such as text editors, data analysis tools, or graphics viewers. In the asynchronous setting, users will be able to access previous sessions and to receive notifications when aspects of a session have changed, such as the addition of new or revised data. To aid retrieval of session content, the GECSR will use automatic archiving, where all versions of artifacts used in a session, such as documents, will be captured and cataloged.

GECSR sessions will draw on the features of leading collaboration tools – many developed by or with the help of personnel involved with this proposal. A key motivation for selection of these tools, rather than comparable commercial products, is that the applications proposed for the GECSR are open source and in the public domain. This distinction is critical because the costs of commercial collaboration tools -- such as WebEx, Microsoft Office Live Meeting (formerly Placeware), or Xerox Docushare – represent a significant barrier to widespread adoption and use. Further, customization of these off-the-shelf products is unlikely due to proprietary control of the underlying software. Therefore, for synchronous audio and video communication, the GECSR will employ both the Virtual Rooms Videoconferencing System (VRVS) [5] and the Access Grid [25]. VRVS, installed on 26,000 machines with over 8200 registered users in 99 countries, provides a worldwide multi-point connection service that allows users to work together with different protocols (SIP, H.323, Mbone, Access Grid, MPEG2, and MPEG4), operating systems, and applications. VRVS users are growing at a rate of roughly one new user per hour. The Access Grid is a specific implementation of high-quality audio hardware with large video displays that allows natural conversations in group-to-group interactions. VRVS interoperates with the Access Grid and provides scalable access to an AG along with interfacing of H.323 hardware. For sharing and viewing common applications in real time, GECSR will use the Virtual Network Computer (VNC) system. For secure presence awareness and instant messaging, the GECSR will use the Pervasive Collaborative Computing Environment (PCCE) [26], which includes Jabber eXtensible Messaging and Presence protocols as well as collaborative computational workflow tools. For asynchronous playback of audio, video, and PowerPoint slides, the GECSR will use the Synco-mat application developed for the Web Lecture Archive Project [27]. For access to stored data, the GECSR will use a peer-to-peer file sharing system based on a reliable and secure group communication protocol [28]. Finally, to automatically configure resources within sessions, the GECSR will use the Monitoring Agents in A Large Integrated Services Architecture (MonALISA) system [29], a monitoring framework and multithreaded, auto-discovering services architecture.

We plan to integrate the tools described above into a unified system with the following novel features:

- ***Persistent Collaboration***. Object persistence is a critical attribute of a collaborative environment. Without the ability to archive and make available the documents, whiteboards, conversations and interactions that occur in the course of an ongoing collaboration we lose one of the most transforming aspects collaborative environments could provide. Persistent documents and interactions in collaborative sessions create a form of group memory, keeping otherwise transient events or results available, not only for those who participated in the original session, but any others who may need to find such results;
- ***Language of access***. We need to provide the users of the GECSR with a language of access to allow each user to define the circumstances under which various collaborators, groups or agents can interrupt them and at what level. The language should be rich enough to allow definition of automated ways of responding

under specified circumstances and must encompass dynamic definition of new groups or interactions. This language should also extend to allowing authorization decisions for persistent objects and the context sensitive use of collaborative tools, as well as use of tools consistent with group preferences and norms (i.e., as expressed through group decision rules). Such a capability is required to assure the users that they will have the flexibility to map social and organizational policies from existing collaborations to GECSR-enabled collaborations.

- ***Human-System-Human as well as Human-Human interactions for collaborative work.*** A customizable dynamic environment that archives interactions and allows dynamic creation of groups could quickly produce more data and meta-data than a typical user can process. Providing agent-assisted collaboration as an integral part of the GECSR will be critical to giving users the capabilities they will need to find information, join groups and interact with colleagues. In addition, monitoring tools can aid collaborations by automatically optimizing communication channels, such as in videoconferences. Finally, for data-intensive transactions, agents can help users determine optimal strategies for moving and viewing data.

# The importance of assessment and evaluation

Because collaborative tools intersect with individual, group, and organizational work practices, a critical goal for successful development and deployment is identification and resolution of key socio-technical problems. Preliminary studies of the ATLAS Collaboration [7], for example, show that dissatisfaction with currently deployed technologies reflects the mismatch between expectations and performance, such as the difference between the ease of face-to-face meetings compared to videoconferences. GECSR's effectiveness depends on its ability to provide capabilities essential to the HENP community. To guide and improve the alignment of performance with expectations, the GECSR team will include researchers with expertise in human-computer interaction (HCI) who can distill common themes from diverse and complicated social and technical feedback. Understanding these social and technical themes will produce two critical benefits for the GECSR project. First, the HCI analyses will help identify the functionality the various GECSR systems must provide to meet the requirements of the HENP community. Second, the HCI analyses will play a role in the adoption of GECSR by providing project visibility to target users very early in the design process, managing expectations, and keeping the end users engaged. The HCI effort falls into two categories of activity: assessment and evaluation. Assessment activity can be thought of as elicitation of requirements from physicists. Evaluation activity can be thought of as measurement of satisfaction with deployed systems.

To meet the goals of the assessment effort, we will use multiple, complementary methods. Specifically, we propose to combine a community-based survey with selected intense observation of physicists at work. Surveys are a well-understood mechanism for efficiently obtaining data about preferences in a larger population through extrapolation from a relatively small number of randomly-selected respondents. The sample for the GECSR user survey will be a randomly stratified selection of physicists, students and technicians involved in the ATLAS, CMS, D0, and CDF collaborations. Items included in the questionnaire will cover: a) communication and collaboration (e.g., frequency and mode of communication by location of collaborators; size and geographic distribution of participants within primary collaborations); b) barriers and challenges to communication and collaboration (e.g., coordination loss, delays, miscommunication, information overload); c) practice (e.g., primary field of research; allocation of time across research activities); d) expectations for the GECSR (e.g., new opportunities for collaboration; better access to experimental data; new ways to incorporate research data in the classroom); e) priorities for GECSR capabilities (e.g., access to data repositories; remote participation in meetings and experiments; collaborative data visualization); and f) policies for data and instrument use (e.g., safety; data access; intellectual property). The aim of the GECSR survey activity will be to produce snapshots of needs and attitudes within the HENP community of users at the start, mid-point, and end of the GECSR project.

In contrast to the broad perspective provided by the survey data, we propose to periodically observe activity and interview participants at 3-5 sites from U.S. institutions involved in the CMS and ATLAS collaborations. Data collected within these settings will help illuminate variation in user needs across different social and institutional arrangements (e.g., a large research university versus a small liberal arts college). Specifically, systematic observation is often the best way to detect the causes of low adoption of deployed technologies – such as Ruhleder and Star's [30] analysis of Worm Community System use by *c. elegans* researchers (WCS) or Orlikowski's [31] analysis of Lotus Notes use by employees of a consulting firm. In both cases, technically elegant tools failed

because key features did not match critical features of work organization, such as criteria for promotion or rewards systems. The aim of the observational and interview activity will be to generate descriptions of processes that either enhance or impede adoption of GECSR applications. As with the survey effort, these data will be used to shape deployment and development priorities – and serve as input to the evaluation effort.

In the iterative development scheme proposed for GECSR, the evaluation component provides closure on each release cycle – and provides input for the next cycle – by offering feedback on user satisfaction and performance related to GECSR use. We propose to measure user satisfaction in terms of subjective reports and objective indicators. We plan to obtain subjective reports of user satisfaction through questionnaire items included on the second and third wave administrations of the community survey (described previously). These items will focus on: a) interface design (e.g., clarity of interface elements; function of icons); b) usability (e.g., do actions have expected consequences; are there predictable failure modes; is system help sufficient); c) functionality (e.g., does an application do what a user wants; are there missing capabilities); d) graphical layout (e.g., do plots show standard units; do applications consume appropriate screen real estate); and e) performance (e.g., did the application update promptly; did data transfer reliably). Objective indicators of satisfaction will include usage logs, where use levels are assumed to relate positively with satisfaction.

In terms of performance related to GECSR use, we propose to use a variety of measures. First, because a target of GECSR development is collaboration, we will focus on the structure of collegial networks before and after GECSR introduction as a marker of GECSR impact on the extent and diversity of network ties. That is, Newman et al. [32] have developed a number of techniques for analyzing and visualizing graphs of extensive networks of association among physicists (e.g., using co-authorship). We will employ similar techniques to characterize networks among the population of scientists within the ATLAS, CMS, D0, and CDF collaborations. For example, a critical question will be the degree to which GECSR use results in more diverse collaborations, both measured in terms of geographic dispersion and in terms of research focus. In addition, we will be interested in whether GECSR use provides a greater capacity for collaboration, as measured by the number and size of primary collaborations. Second, because we expect GECSR use to enhance the significance of collaborations, we will be examining subjective and objective measures of collaboration impact related to use. For instance, subjective measures might include questionnaire items about the self-reported importance of a collaboration, the likelihood of future collaborations with the same collaborators, and the quality of the collaboration experience (e.g., trustworthiness of collaborators). Objective measures of impact might include the output from a collaboration, such as the number of papers and reports – and the visibility of this output. Finally, because one of the anticipated benefits of collaboratory-based science is a leveling of access to experts, instruments, and data – we will be examining whether the GECSR expands participation in physics research. That is, beyond the focal audience of users – the physicists in the large collaborations – we will look for evidence that GECSR tools create new opportunities for engagement by faculty at non-PhD institutions, by undergraduate, by K-12 students, and by the general public. Specifically, we plan to produce case descriptions of novel forms of involvement by new players that the GECSR may enable. The results of this evaluation activity will be shared with the larger scientific community and will inform the development of future tools, environments and collaborations in other science, engineering and education domains.

# Education and Outreach (E&O)

The Education and Outreach (E&O) program associated with this proposal will enable a collaborative learning community engaging traditional and non-traditional students in our studies of particle physics at the energy frontier. In the collaborative learning environments that will result from the development components of the proposal, students will become empowered to do science projects and/or help develop the learning environment itself under the supervision of teachers, and education researchers. We will deploy collaboratory tools into high schools and college classrooms to disseminate high-energy physics knowledge, enhance physics classroom activities, and explore how the tools adapt to the school community.

Florida International University (FIU) will introduce the tools to local high schools and evaluate their impact. FIU, in collaboration with partners at Florida State University (FSU), the University of Florida (UF), and the California Institute of Technology (Caltech), in cooperation with the NSF, are creating and operating an interregional Grid-enabled Center for High-Energy Physics Research and Educational Outreach (CHEPREO) at FIU, encompassing an integrated program of research, network infrastructure development, and education and outreach at one of the largest minority schools in the US [33]. CHEPREO is extending FIU's existing research activities at Jefferson National Laboratory to the long-term high-energy physics research program at the CMS experiment at CERN, create a robust

outreach activity based on CMS research, develop an advanced networking and Grid computing infrastructure that will draw in new collaborators from South America, and enhance science and math education in South Florida for underserved minority students through pedagogic enhancements and teacher training led by a Physics Learning Center. Through GESCR support, FIU will also lead a coordinated program for E&O by creating a community of practice among the various collaborating institutions involved in this proposal, enabled through GESCR tools to insure the integration of E&O project goals: to engage minorities; improve physics classroom curriculum directed by teacher/student contribution and response; add a new dimension to existing programs (QuarkNet, ThinkQuest); and employ education research techniques to investigate these approaches.

QuarkNet-affiliated [34] teachers will utilize the collaboratory tools through summer high-energy research programs. QuarkNet is a successful national program of particle physics education and outreach, supported by NSF that partners high school teachers with physicist mentors. Teachers work as researchers and collaborators in forefront physics projects. The program, now in its fifth year, has over 50 participating university and national laboratory centers, five of which are represented in this proposal. They have well established relationships between teachers and physicists and have reached tens of thousands of high school students through the programs. The University of Texas at Arlington (UTA) and the University of Iowa (UI) are among the twelve original QuarkNet Centers funded by NSF since 1999. FIU is a new QuarkNet center. The groups also cover significant geographic regions and have the potential to engage underrepresented groups in HENP. The E&O team envisages a competition, modeled after the ThinkQuest [35] concept of forming 5 person teams with a coach to develop a project to be shared. While ThinkQuest projects are website lessons, we would develop live lesson plans that would involve remote teams linked together into the collaborative learning environment. Moreover teams could and should consist of geographical separated members with like interests. By providing access to our facilities and research activities, we will build personal fulfillment and science excitement in students as well as open their eyes to new career opportunities. Teachers will be reinvigorated through greater exposure to pedagogical innovations that will translate into successful physics classroom experiences.

Our first goal is to disseminate the collaboratory tools into schools located near our partners. The tools will provide a 'real window' into the exciting world of particle physics at remote accelerator laboratories for many K-12 students and teachers. For education and outreach involving K-12 teachers, we will draw from the pool of active and highly regarded QuarkNet Teachers affiliated with our participating groups' QuarkNet Centers . Through the collaboratory tools, students and teachers will interact with physicists at universities and laboratories to bring high energy physics into the classroom. Classroom-classroom interactions will also be a vital part of the proposal to foster teacher mentoring as well as to build a community among students. We will also provide internships for students to actively work on developing, prototyping, testing and using the collaborative tools. The Education and Outreach Area Coordinator will establish a group that will discuss communication and development issues related to education. These discussions will be aimed at the goals of developing collaborative learning sessions referencing both QuarkNet and ThinkQuest team lesson development model, with a new dimension of live interactive participation, and collaborative class and research projects, among undergraduates and high school students.

Prairie View A&M (PVAMU) is the second oldest university in the state of Texas. Blacks constituted 92% of the undegraduate enrollment of 7255 students at PVAMU in Fall 2002. UTA will partner with PVAMU to concentrate on minority recruiting programs. Since we are delivering a new collaborative environment in this proposal, it would be appropriate to focus on how this new paradigm can help traditionally under-represented students in physics. PVAMU played a major role in developing graphics software for BaBar. PVAMU is in the process of putting together a large donated grid computing cluster. The collaborative tools to be developed through this proposal, with active student involvement at PVAMU, can open doors for traditionally underrepresented students to participate in the greatest scientific discoveries at U.S. and international laboratories.

In the area of education research, we will develop direct and collaborative links with colleagues engaged in studying the effectiveness of collaboratory tools from an educational perspective at the undergraduate level as well as in the education of scientists joining large-scale research projects. To do this, we provide support for a graduate student doing dissertation research in the School of Education at Michigan on the effectiveness of information delivery using the WLAP technology. Undergraduates participating in a summer Research Experience for Undergraduates (REU) at CERN and general physics undergraduates will be used as subjects in research on design principles for the WLAP technology and in an attempt to study the role of individual difference in the design of instructional materials for same [36,37]. The results of this research will be immediately applied to the research and development efforts of this proposal that relate to human-computer interface and will be incorporated into the other aspects of this educational and outreach proposal. Through this research, physicists will gain insight from professional educators

into the nature and effectiveness of information delivery using the WLAP technology. Simultaneously, educators will gain insight into the nature and operation of science experimentation in large collaborations, which may spawn further research. The team at the University of Michigan will also carefully examine the transferability of the knowledge gained in the development of the GECSR Project to the development of college-level courses. By drawing upon the 300 + high energy physics lectures already recorded by the proposal participants from the CERN Summer Student Program and interfacing these through the OGCE portal we plan to develop a full senior-level course in high energy physics and to evaluate and track the performance of students who enroll in it.

The University of Iowa team will leverage four existing programs to conduct education and outreach for this proposal: Iowa QuarkNet, Science Education Center of School of Education, GROW (Grid Research and education group @ IoWa), and the Studio of Academic Technologies for Education.

A special effort will be undertaken by Florida International University in developing a "Grids for Kids" Collaboration on a physics grid tailored for high school implementation which will enable grades 9–12 students to interact in a virtual global village where hands-on implementation and use can promote both understanding and spark an interest to continue studies in physics and technology at the undergraduate level. A small grid array (1-5 servers) in a high school classroom may easily stimulate interest by placing the tools to analyze a simple result set to modeling complex physical behavior. Other related projects are planned at the K-12 level and for teachers. Teachers lesson plans now can include more meaningful applications of theory. Experiments can include larger data sets; more complex data and their results could be analyzed at a rate much higher than could be accomplished with current technology. An environment where teachers can submit a data set to run on a physics cluster with Gigaflops of computing power and also monitor the job as it runs; all this from a remote location, can be an exhilarating experience for kids of any background. Grid physics for kids would immensely help the entry level knowledge of university students; wherever this initiative is implanted.

# Summary of Coordination Plan

A significant challenge in the proposed GECSR project involves successful coordination of diverse participants representing multiple institutions and different disciplinary backgrounds. For example, a recent analysis of 62 completed projects from the NSF Knowledge and Distributed Intelligence (KDI) program, conducted by Cummings and Kiesler [38,39] found varying outcomes in geographically distributed multi-disciplinary teams. Specifically, performance did not vary by multidisciplinarity – but was strongly influenced by geographic dispersion. In particular, dispersed projects that did better employed a number of common management practices – including holding a kickoff workshop or conference, conducting regular audio conferences, and exchanging email regularly. However, workshops and sustained communication were costly in terms of resources and attention of project leaders. Analysis of responses from the PIs in the KDI projects showed the need for tools and practices that could accomplish the following: a) manage and track the trajectory of tasks over time; b) reduce information overload; c) support spontaneous conversation – including awareness of availability for conversation at a distance; d) support simultaneous decision making; and e) support scheduling of and participation in presentations and meetings across distance.

In recognition of the risks posed by a team distributed across eight or more institutions we will take explicit steps to address the pitfalls identified in the KDI study (in addition to regular use of standard modes of communication, such as the phone, email and etc.). Note that the short description here is expanded in the separate three-page Coordination Plan that accompanies this submission. First, project activity will be lead and directed by a management committee consisting of the four Co-PIs and chaired by Dr. Neal. The management committee will meet three times a month, with the default mode being via videoconference – using technology under development in the scope of the GECSR project (e.g., VRVS and Access Grid). Once a month the committee will hold an augmented management meeting where leaders of the core technology will join the four Co-PIs. Second, the project will use the calendar tool within the OGCE to maintain an updated roadmap of project deliverables and milestones. In addition, a current Gannt chart will be posted in the resources area of the project OGCE workspace for quick review of schedule expectations and contingencies. Third, to reduce information overload, the project will use the resources area within the OGCE project space to post and review critical documents and URLs, with notification levels that individuals can adjust (e.g., frequency and type of notifications). Rigorous use of the resources area should also reduce the phenomenon of "death by email attachment" that plagues many distributed projects. Fourth, to provide presence awareness and opportunities for spontaneous interaction, project members will adopt the Jabber-based instant messaging application being developed for GECSR – and also use the built-in presence awareness and

chat capabilities available within OGCE. Fifth, to support distributed presentations, the project will use tools in the Access Grid distribution, such as *dppt*, to view PowerPoint slides across multiple locations. Finally, to jumpstart creation and maintenance of trust and cohesion within the project team, the project will hold an annual workshop where participants can gather and discuss issues face-to-face – with the initial meeting held as a kickoff event shortly after notification of funding. In addition, the project leadership will identify opportunities for sub-sets of the project personnel to meet, such as combining project-related meetings with larger meetings (e.g., ATLAS week).

# Timeline and Milestones

## *First Year: "Integrating Collaborative Tools"*

The goal of the first year will be successful integration of core collaborative tools within the OGCE portal environment. Specifically, at the end of Year 1, users will be able to access key capabilities to support GECSR sessions within a common interface, including: videoconferencing, presence awareness and messaging, control and viewing of shared applications, and a shared file system. Each tool identified for integration into OGCE will be enhanced as needed to provide security, persistence, and compatibility with access and monitoring capabilities (e.g., MonALISA).

In terms of synchronous collaborative tool development, key deliverables in Year 1 will be:

- Adaptation of VRVS to work within the OGCE environment. Specifically, this will involve development of management tools within OGCE that allow users to create groups and associate these groups with GECSR sessions. The management tools will also allow users to associate VRVS resources with GECSR sessions, such as live or archived videoconferences. For example, users will have the option to convert a transient GECSR session into a persistent session, where the GECSR system will automatically apply archiving (e.g., indexed MPEG audio/video streams) to communication in persistent sessions – subject to the preferences of the collaborators; (CalTech)
- Adaptation of VNC to work within the OGCE environment. Specifically, this will involve development of OGCE tools that allow users associated with a GECSR session to control and view shared applications. Early development of specialized modules, for example java applets, optimized for quality of service under dynamically changing bandwidths will be explored and developed; (CalTech, Michigan)
- Adaptation of MonALISA to work within the OGCE environment. The MonALISA system provides a distributed monitoring and control service [40] for collaborative applications. MonALISA is based on a scalable Dynamic Distributed Services Architecture [41] and uses a multi-threaded engine for asynchronous data collection and interaction with other services and agents. Activity related to MonALISA in Year 1 will involve development of OGCE tools that allow users to access output of MonALISA processes, such as monitoring of hosts where VRVS reflectors are running (load, CPU, IO traffic), monitoring the status of VRVS reflectors (number of VRVS rooms, clients, peers, traffic), providing remote administrative control of reflectors using a X.509 certificate based SSL connection, providing real time measurements for the quality of connectivity between selected possible peers (ABPing measures RTT, lost packages, the jitters using UDP and computes the quality of the connection every 2 s); (CalTech, Michigan)
- Adaptation of the PCCE to work within the OGCE environment. The PCCE software provides secure presence and messaging that provides both synchronous and asynchronous communication. The PCCE software is based on the XMPP standards[42] and leverages the Jabber open source software developments[43]; (LBNL,UTA)

In terms of asynchronous collaborative tool development, key deliverables in Year 1 will be:

- Adaptation of reliable and secure peer-to-peer information sharing within the OGCE environment. The scalable and secure peer-to-peer information sharing tool enables groups to securely share information among collaborators from the information's natural storage location (e.g. files and databases)[28]; (LBNL);
- Adaptation of automatic archiving systems within the OGCE environment; (All)

In terms of deployment, key deliverables in Year 1 will be:

- Use of GECSR by the physics participants in this proposal to perform physics analysis, detector commissioning and detector operation; (All, Maryland leads)
- Tighter integration of the legacy experiment specific code with the GESCR environment, for example, integrating PCCE and OGCE infrastructure into data production and analysis tools; (All, Michigan leads)

In terms of assessment and evaluation, key deliverables in Year 1 will be:

- Selection of the community sample, administration of the baseline community survey, analysis of the survey data, and submission of a report based on the survey results; (Michigan)
- Selection of the sites for in-depth observation and interviews, completion of the baseline visits to these sites, and submission of a report based on the results of the observations and interviews; (Michigan)
- Analysis of feedback from physics participants in the GECSR project about the performance and usability of the GECSR environment; (Michigan)

## *Second Year: "Introducing Dynamic Collaboration Capability"*

The focus of the development effort in Year 2 will shift from the integration effort of the Year 1 release to the development of new tools that leverage capabilities that will exist in GECSR after the initial release and add completely new capabilities to GECSR. In particular, a key goal for Year 2 will be development of a "language of access" that when combined with other monitoring capabilities in GECSR will assist users in discovering and joining interesting collaborations. All of the OGCE services will be adapted to also support WSDL bindings using the WS-Resource Framework (WS-RF) [44]. This will allow collaborative capabilities to be added to any application that is capable of acting as WS-RF client.

In terms of synchronous collaborative tool development, key deliverables for Year 2 will be:

- Dynamic optimization of connectivity. Using MonALISA, we will provide a dynamic distributed decision unit to compute a global Minimum Spanning Tree (MST) to optimize connectivity and to correctly solve the connectivity when reflectors are down or unreachable due to network problems. This is done using a set of agents deployed to all the monitoring services to collect the connectivity information and to dynamically compute the MST; (CalTech)
- Implementation of a language of access. The GECSR will have a language of access, using authorization tools [45] that allow association of ad-hoc groups of physicists with GECSR sessions based on individual X.509 certificates. Prior work has produced Web-based tools that implement much of the language of access functionality. The new work in Year 2 will integrate these existing tools within the OGCE environment; (CalTech)
- Implementation of location-aware monitoring. We will use a geographic-information-systems-based Grid information broker (GIS-GIB) [46] to provide physicists with a location-aware mechanism for monitoring GECSR-based collaborations, discovering interesting collaborations, and analyzing the spatial and temporal characteristics of collaborations. GIS-GIB will help physicists dynamically identify where and when interesting collaborations are taking place. In addition, the GIS-GIB can be used to organize and analyze collaborative efforts that are geographically distributed and where the level of activity varies over time; (Iowa)
- Implementation of collaborative, grid-enabled dialogs. Collaborative work-sharing will be enhanced by decision support from the monitoring and other Grid-management agents, who will interact with the collaborators in a given session to determine their needs, match the needs to Grid system conditions, and begin a simple dialog in case there are difficult choices to be made (e.g., in terms of moving large volumes of data); (CalTech, UTA)

In terms of asynchronous collaborative tool development, key deliverables for Year 2 will be:

- Adaptation of HEPBook [47] to work within the OGCE environment. HEPBook provides a graphical user interface for entry of multi-media content, extensive search capabilities, access control, authorization, and protocols to connect to remote repositories. In addition, we will provide an application programmer interface to allow users to define new HEPBook entry types and to define new software agents (e.g., for

managing HEPBook configuration). In particular, users will be able to use languages other than Java and the Java Expert System Shell (JESS). The basis for the agent platform will be JBOSS and Java Enterprise Servers; (Maryland)

- Adaptation of Quadtree-based Replica Management Service to work within the OGCE environment. An adaptive quadtree approach is proposed to support replica management service in a dynamic data sharing system within a GECSR session or across sessions. It is anticipated that Grid-based collaborative LHC analysis will require frequent data replication and distribution based on dynamic data access patterns, continuing generation of new data, and diverse scopes of human-to-human collaborations. Current Grid middleware (e.g., GDMP) does not support regional and global optimization of dynamically distributing, replicating, and sharing large LHC datasets on the Grid. A preliminary study [48] has demonstrated that a quadtree-based algorithm can adaptively decompose and share datasets on the Grid based on their characteristics such as frequency of data access. The quadtree approach will be integrated with GECSR-based group communication protocols to provide optimal solutions to large-scale data allocation and replication within GECSR sessions; (Iowa)
- Document versioning and annotation capabilities associated with the peer-to-peer information sharing tool. This capability will provide an initial course-grained collaborative document editing facility. (LBNL)

In terms of deployment, key deliverables in Year 2 will be:

- Release of the GECSR environment for use by a limited number of external user groups in the HENP community (e.g., including but not limited to groups associated with CDF and D0 at the Tevatron and with ATLAS and CMS at the LHC). These groups will be selected to cover a representative distribution of potential GECSR users. Candidate groups include two physics object groups from CMS (each group has 30+ members focused on development of algorithms for physics object reconstruction and implementation of these algorithms in software), sub-groups of the CMS object groups (2-5 members focused on specific tasks), the Hadron calorimeter group (this group has around the clock activity associated with beam test and detector commissioning at CERN – there are already plans to perform remote tasks from the US), and the challenge driven data production groups in the LHC experiments. (Maryland)

In terms of assessment and evaluation, key deliverables in Year 2 will be:

- Analysis of feedback from test users in the external groups about the performance and usability of the GECSR environment. (Michigan)

In terms of support of users, key deliverables in Year 2 will be:

- Establishment of a single integrated GECSR help desk and support function. (Maryland)
- Enhancements to deployed tools based on prioritization by the project management team using the assessment and evaluation information from the first year. (All)

## *Third Year: "Deploying a production environment"*

The focus of activity in Year 3 will shift to deployment of GECSR within the larger HENP community.

In terms of asynchronous collaborative tool development, key deliverables in Year 3 will be:

- Implementation of a prototype tool for collaborative LaTeX document preparation. Physics collaborations often revolve around preparation of LaTex documents and mechanisms for marking up, versioning, and commenting on these documents are a critical need. Work on this capability will begin in year two and a first prototype capability will be available in year 3. (LBNL)

In terms of deployment, key deliverables in Year 3 will be:

- Deployment of the GECSR environment for the QuarkNet [46], CHEPREO[47] and PVAMU outreach participants. (FIU, UTA)
- Deployment of prototype collaborative editing capabilities to a select HENP group for testing and feedback. (Maryland, MSU)

In terms of assessment and evaluation, key deliverables in Year 3 will be:

- Administration of the mid-point community survey, analysis of the survey data, and submission of a report based on the survey results. (Michigan)
- Completion of the mid-point visits to the in-depth sites, and submission of a report based on the results of the observations and interviews. (Michigan)

In terms of support of users, key deliverables in Year 3 will be:

- Set up and opening of GECSR support office at Fermilab. We propose to have one dedicated computer professional for deployment and support. Since this person needs to support a large number of internal and external user groups, the deployment and support activity will reside at Fermilab and utilize the infrastructure of the US CMS Tier-1 center for the most efficient deployment and support of the user community. In addition, the person will work closely with computing and software support groups in each HEP collaboration to provide additional support to users. The deployment and support person will also collect bug reports on the deployed GECSR environment, and pass this feedback to the development team for further improvement of GECSR. (Maryland)
- Enhancements to deployed tools and capabilities based on prioritization by the management team using the assessment and evaluation from the second year. (All)

### *Fourth Year: "Operational use of the GECSR"*

The focus of activity in Year 4 will shift to operational use of GECSR within the larger HENP community.

In terms of asynchronous collaborative tool development, key deliverables in Year 4 will be:

- Collaborative LaTeX document editing capability. This capability will provide users with rudimentary change tracking and commenting facilities that are integrated into LaTeX. (LBNL)

In terms of deployment, key deliverables in Year 4 will be:

- Widespread adoption and use of GECSR within the HENP community. Physicists in this collaboration will work closely with the management group of each target HEP collaboration for wide adoption and use of the GECSR environment. Some members in this proposal are from such management group. D. Green is the US-CMS Research Program manager and also a member of the international CMS management board. J. Womersley is a co-spokesperson of the D0 collaboration. We have already started discussing with managers of other HEP collaborations. (All, Maryland leads)

In terms of assessment and evaluation, key deliverables in Year 4 will be:

- Administration of the final community survey, analysis of the survey data, and submission of a report based on the survey results. (Michigan)
- Completion of the final visits to the in-depth sites, and submission of a report based on the results of the observations and interviews. (Michigan)

In terms of support of users, key deliverables in Year 4 will be:

- Continued operation of the GECSR support office at Fermilab. (Maryland)
- Enhancements to deployed tools and capabilities based on prioritization by the management team using the assessment and evaluation from the second year. (All)

# Broader Impacts

Though HENP will be the targeted community for this proposal, the tools, systems, and the integrated working environment, proposed here will be designed and constructed to be broadly applicable, so that GECSR will contribute to improved modes of collaborative work in many fields of science and engineering, education, and homeland security. Specifically, there are three target areas where the GECSR experience strongly overlaps with

related ongoing efforts in other fields and will have immediate impact. First, integration of videoconferencing capabilities within the OGCE framework will deliver video applications to other communities using OGCE, such as the earthquake engineers via NEESgrid, who currently rely on costly proprietary unicast technologies (e.g., H.323 devices) [49]. Second, GECSR applications will be compatible with the Sakai environment, since OGCE and Sakai share the same base software. Sakai is a recently announced Mellon Foundation effort involving Michigan, MIT, Stanford, Indiana and the uPortal consortium that will build an open source course management system. Assuming widespread adoption of Sakai, any institution using Sakai will also be able to incorporate GECSR features, allowing a seamless integration of research and classroom tools [50]. Third, the NIH has created eight regional centers for biodefense, called Regional Centers of Excellence, that will require tools for secure collaboration and data movement across geographically distributed labs. At least one of these centers, the Great Lakes RCE, has already adopted an OGCE-based collaboration environment [51] – and the NIH has announced a competition to construct cyberinfrastructure to join all eight centers which is likely to expand the use of OGCE beyond the Great Lakes center. Therefore, while GECSR development is imperative for the health and success of current and future global physics collaborations, output from the GECSR project will also be vital for the success of other high visibility cyberinfrastructure initiatives.

In terms of the impact on physicists, the collaborations enabled and fostered by the GECSR, with the support of governments and funding agencies around the world, will start to fulfill the vision that scientists from all regions are allowed, encouraged, and given the means to function as full partners in the scientific process of search and discovery**.** This proposal is an initial attempt to realize worldwide participation in the next generation of HENP experiments. True global collaboration in the data-intensive analysis that upcoming experiments will require creates a demand for a new generation of grid-enabled collaborative tools that can both access resources distributed around the world and knit together geographically dispersed collaborators.

The broader implications of this proposal, and its mission to grant full partnership to groups in all world regions are driven by (1) The right of groups and governments that have contributed to the global collaborations to share and collaborate in the data analysis, and thus in the process of search and discovery that is at the root of all basic research. (2) The need to justify global projects, and global collaborations as a valid and vital means of conducting future leading-edge scientific research and other endeavors, without exploitation of the poorer nations and world regions and (3) The need to develop effective means of collaboration and cooperation among sub-groups with vastly different academic and cultural backgrounds, and especially to allow effective collaboration by university-based students from their home countries.

# Background of our Groups

This proposal brings together several groups with complementary and extensive knowledge and experience in several areas of collaborative scientific research:

**Caltech:** Caltech's expertise is in videoconferencing software, virtual rooms, Virtual Room Videoconferencing System, agent based monitoring, virtual organization management, high performance wide area network implementation, management and operations, Grid systems and Grid-Enabled Analysis development, self-learning distributed systems optimization. The Caltech group, led by the US CMS Collaboration Board Chair, originated and has had responsibility for transatlantic networking for HEP since 1982. It currently co-manages the operation and development of the "LHCNet" links between CERN and the Starlight in Chicago [52]. It led the MONARC project [53] that developed the worldwide distributed computing model for LHC data analysis, and originated the Data Grid hierarchy concept that has been adopted by the LHC experiments. It shares leadership in the Particle Physics Data Grid [2] and iVDGL[54]. It developed VRVS [5] which went into production in 1997 and currently runs on over 20,000 hosts..

The Caltech CMS group in HEP has worked in close collaboration with the Center for Advanced Computing Research (CACR) since 1996. This collaboration began with the GIOD (Globally Interconnected Object Databases) [55] project funded by Hewlett Packard, and the first accurate evaluation of the backgrounds to Higgs particle searches using the diphoton signature, using pre-TeraGrid prototypes that took advantage of the large-scale computational resources at CACR and other NPACI sites. This collaboration led to the development of the first prototype Tier2 center at CACR and SDSC, by the CMS groups at Caltech and UCSD. Caltech and UCSD are two of the four TeraGrid [56] sites (along with Argonne and NCSA).

Caltech, and the CMS group in particular, has been very active in fostering collaborations for science and education on a global scale. The group has provided assistance in the design and implementation of Tier2 centers in

developing countries since 1999, including India, Pakistan, China, Korea and Brazil among others. Recent events organized by this group include a lecture series on Grids, Networks and national IT infrastructure at the renowned Nathiagali Summer College in Islamabad Pakistan (founded by Nobel Laureate Abdus Salam in 1974), a session on Grids at UERJ in Rio (and made available to the world scientific community using VRVS and AMPATH) in February 2002, and a session on the "Role of New Technologies in Developing the Information Society" in Bucharest in November 2002 [57]. ]. Recently the Inter-regional Grid Enabled Center for High Energy Physics Research and Educational Outreach (CHEPREO) at FIU was established through the auspices of the NSF, and provides an ideal testbed environment for the deployment, assessment, and evaluation of collaborative tools for radical improvement in science and math education

**Fermilab:** Fermilab is the US center of CMS for detector construction, commissioning and operations, and physics analysis. It is also the US CMS Tier-1 center for the computing. It is also one of the US institutions working on the LHC machine construction and operation. Many of the Fermilab team members are also members of CDF or D0. Their participation in this proposal is as a user group: to help define requirements and evaluate the prototype collaboratory environment. In addition, the members from the Computing Division (CD) will coordinate work on this proposal and other computing projects at Fermilab, which include the US CMS computing project and the HEPbook development project in CD.

**Florida International University** is ranked by the Carnegie Foundation for the Advancement of Teaching as a Doctoral/Research University-Extensive, the highest ranking in its classification system. FIU has more than 32,686 students, 1,100 full-time faculty, and 90,000 alumni, making it the largest university in South Florida and placing it among the nation's 25 largest colleges and universities. It is unique in that it has the highest proportion of international students and faculty of any major university in the country. Its mission includes being the principal educational and research interface between the State universities and South and Central America and the Caribbean. AMPATH provides opportunities for Hispanic minority students to work on the project in technology and administrative positions. FIU is one of the largest minority universities in the continental US and the largest Hispanic serving minority university with over 32,000 students, (55% Hispanic and 20% African American). The FIU nuclear physics group was established in 1995 when 8 positions (6 experimental and 2 theoretical) were created in partnership with the Thomas Jefferson National Accelerator Facility. Members of this group have gone on to collaborate in the CMS experiment through the NSF supported CHEPREO project.

**University of Iowa:** The University of Iowa team members will bring their expertise in the following aspects into the execution of the proposed GECSR project: (1) Grid research, education, and outreach in HENP through the collaboration between GROW (Grid Research and educatiOn group @ IoWa) and university CMS community; (2) CMS collaboration experiences in forward calorimeter construction and data analysis; (3) collaborative visualization (e.g., NSF funded LIVE (Laboratory for Immersive Visualization for the Environment) [58]); (4) geographical-information-systems-based Grid monitoring; (5) an iVDGL member institution participating in the development and operation of USCMS Grid testbeds; and (6) an intra-campus research Grid - HawkGrid supported by its operation center.

**Lawrence Berkeley National Laboratory:** The members of the Lawrence Berkeley National Laboratory team bring to the proposal extensive expertise in physics collaborations and nine years of development, deployment and use of a broad range of collaborative tools including a Pervasive Collaborative Computing Environment[26], secure and reliable group communication protocols [59], secure messaging and presence, Grid computing, secure peer-to-peer shared file systems[28], security models appropriate to collaborative activities[26], and the Access Grid videoconferencing systems.

**Maryland:** The members of the University of Maryland team have extensive experience in coordinating many groups with a large number of people both in CMS and D0. They worked as coordinators of the CMS hadron calorimeter construction project, the CMS Jet and Missing Et physics group, the D0 physics group and its subgroups (top, new phenomena), the D0 Computing group, the D0 simulation group, etc. The Maryland team brings to the proposal close ties to the user community. Also, the team takes a responsibility for deploying the developed collaboratory environment and user support..

**University of Michigan:** The University of Michigan is an international leader in the design and assessment of collaboratories, in the development of open-source environments for Web-based classroom and research collaboration, and in the capture and storage of lecture and presentation content for replay over the Web. In terms of collaboratories, faculty at the School of Information (SI) developed and deployed the NSF-funded Space Physics and Aeronomy Research Collaboratory (SPARC) in 1993, which remains the world's oldest, continuously operating

collaboratory. Over the past decade, SI faculty have led several other important NSF and NIH sponsored collaboratory projects, including the currently funded George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES)[60] – where Dr. Finholt is a co-PI on NEESgrid [49], the grid-based collaboratory component of NEES. In addition, SI faculty are leading an effort to develop general principles for collaboratory design and use through the NSF ITR-funded Science of Collaboratories[61] project, where Dr. Finholt is also a co-PI. In terms of environments for Web-based collaboration, faculty and staff at the University of Michigan's Duderstadt Center (formerly the Media Union) are leading the effort to build an open source course management system, called Sakai [50]. The Sakai project, carried out with MIT, Stanford, Indiana, and the uPortal consortium, was just launched with a major gift from the Mellon Foundation. Dr. Severance is leading the effort to produce grid-enabled versions of Sakai, such as for the NEESgrid and OGCE projects. Finally, in terms of capture and replay of lecture content over the Web, faculty and staff at the MU and the Department of Physics produced the first applications for archiving and replaying CERN seminars and tutorials. Specifically, Professor Neal, Dr. McKee, and Dr. Severance have directed and shaped the Web Lecture Archive Project [27], which in an expanded form will provide a key mechanism for education and outreach in this proposal. Overarching these independent activities is the Michigan Grid Research and Infrastructure Development (MGRID) center[62] – an interdisciplinary effort funded by the University of Michigan administration to develop an internal grid testbed (mgrid.org). MGRID center staff will work with key projects, such as the research proposed here, to enhance evolution of data and computing grid applications at Michigan and elsewhere. Drs. Neal and Finholt are appointed to the Executive Board of MGRID, and Dr. Severance is a key member of the MGRID technical advisory team.

**University of Texas at Arlington:** Close interactions between the HEP researchers and CS researchers at UTA have led to recent acquisitions from the NSF MRI program and joint supervision of graduate students. The HEP group at UTA played a leading role in developing tools (Grid Application Toolkit GRAT[63]) for large scale Monte Carlo data generation using Data Grid middleware like Globus [64] and Condor [65]. Thousands of cpu days are made available for specific studies over short periods of 1-2 weeks at 3 national laboratories and 7 universities in the U.S. The collaboratory tools proposed here will allow collaborations to share and manage the task of data production and analysis. The CS researchers at UTA have extensive experience in multimedia (author of the MtreC MPEG4 Toolkit [66]) and database research. UTA serves a large minority population in the Dallas-Ft. Worth metroplex.

# Results from Prior NSF Support

**Homer Neal:** University of Michigan NSF Award DUE #0333580 ( 10/1/032003 – 9/30/2005; $249,998). PI- " Web Lecture Capture Project" The project is designed to develop a fast and efficient process for capturing and publishing content-rich lectures on the web in the setting of large, multiple-parallel session conferences and meetings using the WLAP technology. Significant milestones have been achieved, including a recent demonstration at the American Physical Society Meeting. University of Michigan NSF Award PHY #0139604 (10/1/2001 – 9/30/2004, $310,000). (PI) --Research Experience for Undergraduates Program at the European Organization for Nuclear Research (CERN). This program provides the opportunity for U.S. undergraduates to spend summers at CERN carrying out research and participating in the prestigious CERN summer student program. A number of students in this program would be expected to work on the GECSR project proposed herein.

**Julian Bunn:** Johns Hopkins University Subaward Agreement No.08002-48195 under NSF Grant No. PHY-9980044 (9/15/1999 - 8/31/2003; $1,051,771), "Accessing Large Data Archives in Astronomy and Particle Physics" Results: The sheer amount of data that are used in the fields of astronomy and particle physics has forced both communities to move from flat files to highly organized databases. This project contributed to 1) The organization of data for efficient access, 2) Storage of data in widely distributed locations, 3) Efficient handling of entire major archives.

**Tom Finholt:** CMS-0117853 (8/01/2001 – 9/30/2004; $1,814,026), "NEESgrid (Network for Earthquake Engineering Simulation): A Distributed Virtual Laboratory For Advanced Earthquake Experimentation and Simulation". Results: This project will contribute to the system integration effort for the George E. Brown, Jr. Network for Earthquake Engineering Simulation. It will define the user requirements for the NEESgrid and contribute several software components including collaboration tools through the CompreHensive CollaborativE Framework (CHEF). Representative Papers: Severance, C.R. (2003). Integrating the grid into CHEF. NEESgrid

Technical Report TR-2003-1 and Finholt, T.A., Wierba, E.E., Birnholtz, J.P. & Hofer, E. (2002). NEESgrid user requirements document. NEESgrid Technical Report.

**Shuichi Kunori:** No prior support

# Conclusion: Transition to a New Era of Collaboration in Scientific Research and Education

We are proposing to launch a next-generation collaborative framework for scientific research, harnessing multi-institutional expertise, and existing and new tools across a broad range of technologies, to begin the transition to a new era of collaboration in the sciences. Just as there is an ongoing transition to grid computing, enabling transparent access between people and computing resources, we foresee a transition to collaborative environments that enable new modes of transparent, persistent and spontaneous interaction between scientists located at sites around the globe. The collaboratory we envision will provide secure, intelligent, easy to use collaborative tools – delivered via a uniform portal interface -- integrated with discipline-specific tools for each scientific field. This makes it a key required element for the success of today's global scientific collaborations, and gives it the potential of becoming a cornerstone of more tightly knit worldwide collaborations of the future. Extension of the GECSR system and concepts to the formation of Grid-enabled Learning Environments will provide educators and students alike with richer and easier modes of access and interaction with scientists located in all world regions; making them, from their classrooms and homes, part of the discovery process of frontier science. The proposal's research outcomes, from the development of the collaboratory to a deeper understanding of effective modes of global communication, could have a profound and broad impact on research and education, as well as the way multi-institution organizations operate, by enhancing our ability to interact and share work collectively, across global networks.

# Management Coordination Plan

**Specific roles of project personnel**

The **Project Director** will be Dr. Homer Neal (University of Michigan), who is the director of the UM-ATLAS Collaboratory Project, the Samuel A. Goudsmit Professor of Physics, Interim President Emeritus, Vice President Emeritus for Research, and former chair of physics from 1987 to 1993 at Michigan. Dr. Neal's research area is experimental high energy physics and he is currently conducting his research at CERN, where his research group is part of the ATLAS Experiment. His technical research expertise includes the design of particle detectors, particle event reconstruction and analysis, large-scale database management, and particle physics phenomenology. He has led many experiments that have elucidated the nature of spin effects in high energy particle interactions, including proton-proton elastic scattering, electron-positron scattering and in various inclusive hadronic reactions. In the ATLAS experiment, for which he is the Michigan Institutional Representative, Neal is involved in developing the computing infrastructure required to carry out the planned physics analysis. This includes work on high speed networking between CERN and the US, quality of service protocols, and the development of collaborative tools. He is also involved with the Michigan effort to construct muon chambers for the ATLAS forward muon spectrometer.

The **Project Coordinator** will be Dr. Shuichi Kunori (University of Maryland), who has long experience in leadership of physics analysis groups. The Project Coordinator has overall responsibility for project coordination and deliverables, as well as for calling meetings, writing reports, dissemination, and the use of the delivered tools in the physics collaborations, as well as working closely with the Technical Coordinator on their joint responsibilities. The Project Coordinator will report to the Steering Group for the Management Team.

The **Technical Coordinator** will be Dr. Julian Bunn (California Institute of Technology), who has significant technical expertise in grids, collaborative systems, and HEP computing. The Technical Coordinator will have overall responsibility for the day-to-day coordination of the technical work of the project, for identifying needed inter-team coordination and work, and for working closely with the Project Coordinator on their joint responsibilities.

The **Assessment and Evaluation Coordinator** will be Dr. Thomas Finholt (University of Michigan), who has lead assessment and evaluation of several major collaboratory projects including NEESgrid, the collaboratory element of the George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES). The Assessment and Evaluation Coordinator will design, administer, and analyze the community survey and design, conduct, and analyze the in-depth site visits. In addition, the Assessment and Evaluation Coordinator will work with the rest of the project management to insure that development and deployment activities are consistent with documented user needs.

The **Education and Outreach Coordinator** will be Dr. Heidi Alvarez (Florida International University), who has led development of the Center for High-Energy Physics Research and Educational Outreach (CHEPREO) at FIU. The Education and Outreach Coordinator will establish and lead the group that will design and implement the GECSR E&O activity. These activities will be aimed at the goals of developing collaborative learning sessions, and collaborative class and research projects, among undergraduates and high school students.

**How the project will be managed**

This project will involve coordination of development and deployment activity across distributed institutions. In addition, this project will be part of the LHC research program, where thousands of physicists are engaged in an ongoing managed international collaborative effort, in which the core of the project team is already embedded as collaborators in the LHC physics experiments. Therefore, the project will employ multiple management structures to meet the need for internal organization and the need for external coordination with the larger LHC research community.

In terms of internal organization, the project will be led by Dr. Neal and a **management team** consisting of the coordinators named previously: Bunn, Finholt, Kunori, and Alvarez. The management team will be responsible for running the project, will employ the specific coordination mechanisms described in the next section, will track the progress of the project, and will administer funds. The experience of this team will enable us to meet the significant management challenges of the project: the broad scope of the problem to be addressed; the diverse multi-disciplinary teams; and the diverse external constituencies including national labs and international organizations. In addition, the project will employ an **augmented management team**, consisting of the core project leadership and the leads within each of the technical development areas.

In terms of external coordination, the project will form a **GECSR user group**, with the responsibility for assisting the assessment and evaluation effort in collection of user requirements and of user experiences with the GECSR tools. The GECSR user group activity will be led by Dr. Raymond Brock (Michigan State University). Through the GECSR user group, interested and early-adopter members of the research community can discuss and give feedback on issues and suggestions for: more effective modes of communication during work sessions; required features or new paradigms; and future development work. A GECSR Web site will be maintained to keep the scientific community up to date with development plans and issues, and to encourage online feedback.

Finally, to determine broad strategic directions and to stay current with the physics community and relevant technology communities, the GECSR project will form an **external advisory board** (EAB). The EAB will engage both the physics user community and the information technology research community to advise the GECSR management team. Members will include people from the physics experiments (e.g., US ATLAS and US CMS research program managers or designees), from Grid projects (e.g., GrPhyN), computer science researchers, and other information technology experts. The EAB will meet once a year.

**Specific coordination mechanisms**

A significant challenge in the proposed GECSR project involves successful coordination of diverse participants representing multiple institutions and different disciplinary backgrounds. For example, a recent NSF report about 62 completed projects from the Knowledge and Distributed Intelligence (KDI) program (Cummings & Kiesler, 2003), found varying outcomes in geographically distributed multi-disciplinary teams. Specifically, performance did not vary by multidisciplinarity – but was strongly influenced by geographic dispersion. In particular, dispersed projects that did better employed a number of common management practices – including holding a kickoff workshop or conference, conducting regular audio conferences, and exchanging email regularly. However, workshops and sustained communication were costly in terms of resources and attention of project leaders. Analysis of responses from the PIs in the KDI projects showed the need for tools and practices that could accomplish the following: a) manage and track the trajectory of tasks over time; b) reduce information overload; c) support spontaneous conversation – including awareness of availability for conversation at a distance; d) support simultaneous decision making; and e) support scheduling of and participation in presentations and meetings across distance.

In recognition of the risks posed by a team distributed across eight or more institutions we will take explicit steps to address the pitfalls identified in the KDI study (in addition to regular use of standard modes of communication, such as the phone, email and etc.). First, the management team will meet three times a month, with the default mode being via videoconference – using technology under development in the scope of the GECSR project (e.g., VRVS and Access Grid). Second, the management team will hold a monthly meeting of the augmented management team. Third, the project will use the calendar tool within the OGCE to maintain an updated roadmap of project deliverables and milestones. In addition, a current Gannt chart will be posted in the resources area of the project OGCE workspace for quick review of schedule expectations and contingencies. Fourth, to reduce information overload, the project will use the resources area within the OGCE project space to post and review critical documents and URLs, with notification levels that individuals can adjust (e.g., frequency and type of notifications). Rigorous use of the resources area should also reduce the phenomenon of "death by email attachment" that plagues many distributed projects. Fifth, to provide presence awareness and opportunities for spontaneous interaction, project members will adopt the Jabber-based instant messaging application developed for GECSR – and also use the built-in presence awareness and chat capabilities within OGCE. Sixth, to support distributed presentations, the project will use tools in the Access Grid distribution, such as *dppt*, to view PowerPoint slides across multiple locations. Finally, to jumpstart creation and maintenance of trust and cohesion within the project team, the project will hold an annual workshop where participants can gather and discuss issues face-to-face – with the initial meeting held as a kickoff event shortly after notification of funding. In addition, the project leadership will identify opportunities for sub-sets of the project personnel to meet, such as combining project-related meetings with larger meetings (e.g., ATLAS week).

**Specific budget items to support coordination**

The most significant budget item in support of coordination within the GECSR project is funding for development and deployment of the GECSR tools themselves – which in addition to be used by physics researchers to do their work will also be used by the members of the GECSR team to execute the GECSR project. Another significant budget items is funding for travel, which will support an annual face-to-face meeting of the augmented management team and an annual meeting of the EAB.

## References


1. Grid Physics Network (2002). GriPhyN – Grid Physics Network. http://www.griphyn.org.

2. Particle Physics Data Grid (2002). Particle Physics Data Grid. http://www.ppdg.net.

3. DataGrid Project (2002). The DataGrid Project. http://eu-datagrid.web.cern.ch/eu-datagrid/.

4. DataTAG (2002). DataTAG Home Page. http://datatag.web.cern.ch/datatag/.

5. VRVS (2003). Virtual Rooms Videoconferencing System. http://www.vrvs.org.

6. Bunn, J.J. (2002). Grid enabled analysis tools. http://pcbunn.cithep.caltech.edu/GAE/GAE.htm.

7. "Opportunities for Use and Development of Collaborative Tools in ATLAS"(2003). *Goldfarb et al.* (*CERN ATL-GEN-2003-002)*

8. Atkins, D.E., Droegemeier, K.K., Feldman, S.I., Garcia-Molina, H., Klein, M.L., Messina, P., Messerschmitt, D.G., Ostriker, J.P., & Wright, M.H. (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure*. http://www.communitytechnology.org/nsf_ci_report/.

9. Avery, P. (2002). Data Grids: A New Computational Infrastructure for Data Intensive Science. *Phil. Trans. R. Soc. London A*, 360, 1191-1209.

10. Compact Muon Solenoid Collaboration (2002). Welcome to the Compact Muon Solenoid Outreach Activities. http://cmsdoc.cern.ch/cms/outreach/html/index.shtml.

11. CERN (1999). Welcome to ATLAS. http://atlas.web.cern.ch/Atlas/Welcome.html.

12. CERN (2002). CERN – Public home. http://public.web.cern.ch/public/.

13. CERN (2003). LHC – The Large Hadron Collider Home Page. http://lhc-new-homepage.web.cern.ch/lhc-new-homepage/.

14. Stanford Linear Accelerator Center (2002). BaBar Public Web Home Page. http://www-public.slac.stanford.edu/babar/.

15. Fermi National Accelerator Lab (2003). The DZero Experiment. http://www-d0.fnal.gov/.

16. Fermi National Accelerator Lab (2003). The Collider Detector at Fermilab. http://www-cdf.fnal.gov/.

17. Foster, I., Kesselman, C. & Tuecke, S. (2001). The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International J. Supercomputer Applications*, 15(3).

18. Gould, J.D., Boies, S.J., Levy, S., Richards, J.T., & Schoonard, J. (1987). The 1984 Olympic Messaging System: A test of behavioral principles of system design. *Communications of the ACM*, *30*(9): 758-69

19. Nielsen, J. (1993). *Usability engineering*. San Diego, CA: Academic Press.

20. Olson, G. M., and Olson, J. S. (1991) User centered design of collaboration technology. *Journal of Organizational Computing,* 1, 61-83.gould

21. Brooks, F.P. (1995). *The mythical man month, Anniversary edition*. New York: Addison-Wesley.

22. Foster, I., Kesselman, C., Nick, J. & Tuecke, S. (2002). The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Open Grid Service Infrastructure WG, *Global Grid Forum*, June 22.

23. The Open Grid Computing Environment. http://www.ogce.org .

24. Basney, J. (2002). MyProxy. http://www.ncsa.uiuc.edu/Divisions/ACES/MyProxy.

25. Access Grid (2003). Access Grid. http://www-fp.mcs.anl.gov/fl/accessgrid/.

26. D. Agarwal, C. McParland, and M. Perry, "Supporting Collaborative Computing and Interaction," *Proceedings of the Grace Hopper Celebration of Women in Computing 2002 Conference*, October 9-12, 2002, Vancouver, Canada.

27. Web Lecture Archive Project (2002). Web Lecture Archive Project. http://www.wlap.org/.

28. Lawrence Berkeley National Laboratory Collaboration Technologies Group (2002). A Scalable and Secure Peer-to-Peer Information Sharing Tool. http://www-itg.lbl.gov/P2P/file-share/.

29. MonALISA web page : http://monalisa.cacr.caltech.edu/

30. Star, L.S., & Ruhleder, K. (1994). Steps towards an ecology of infrastructure: Complex problems in design and acces for large-scale collaborative systems. *Proceedings of the 1994 conference on computer-supported cooperative work* (pp. 253-264). New York: ACM Press.

31. Orlikowski, W. (1992). Learning from Notes: Organizational issues in groupware implementation. *Proceedings of the 1992 conference on computer-supported cooperative work* (pp. 362-369). New York: ACM Press.

32. Newman, M.E.J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Science*, *98*, 404-409.

33. An Inter-Regional Grid Enabled Center for High Energy Physics Research and Educational Outreach (CHEPREO) at Florida International University.

34. QuarkNet (2002) http://quarknet.fnal.gov/

35. ThinkQuest is a program under Advanced Networks & Services Inc http://www.thinkquest.org/programs.htm

36. Mayer, R.E. (2001). Multimedia learning. New York: Cambridge University Press.

37. Kalyuga, S., Chandler, P., & Sweller, J. (2002). Incorporating learner experience into the design of multimedia instruction. Journal of Educational Psychology, 92(1), 126-136.

38. Cummings, J., & Kiesler, S. (2003, December). Coordination and success in multidisciplinary scientific collaborations. Presented at the International Conference on Information Systems 2003, December 14-17, Seattle, WA.

39. Cummings, J., & Kiesler, S. (2003). KDI Initiative: Multidisciplinary scientific collaborations. Arlington, VA: National Science Foundation.

40. H.B. Newman, I.C. Legrand, P.Galvez, R. Voicu, C. Cirstoiu**,** MonALISA: A Distributed Monitoring Service Architecture CHEP 2003, La Jola, California, March 2003

41. Harvey B. Newman, Iosif C. Legrand, Julian J. Bunn A Distributed Agent-based Architecture for Dynamic Services CHEP - 2001, Beijing, Sept 2001

42. The eXtensible Messaging and Presence Protocols working group of the IETF. Documents available at: http://www.ietf.org/html.charters/xmpp-charter.html

43. Jabber Software Foundation: http://www.jabber.org/

44. WS-Resource Framework. http://www-fp.globus.org/wsrf/default.asp

45. Clarens (2003). Grid-enabled Web Services Framework for Collaborative Distributed Analysis, http://clarens.sourcefoge.net.

46. Wang, S., A. Padmanabhan, Y. Liu, R. Briggs, J. Ni, T. He, B. M. Knosp, and Y. Onel (2003). A Multi-agent System Framework for End-user Level Grid Monitoring Using Geographical Information Systems (MAGGIS): Architecture and Implementation. *Proceedings of Lecture Notes in Computer Science*, Accepted.

47. Roediger, G. (2003). Extending the Control Room Logbook to the HEPBook – A personal Collaboratory Notebook. http://cdinternal.fnal.gov/CHEP/chep2003/Roediger.pdf.

48. Wang, S., and M. P. Armstrong (2003). A Quadtree Approach to Domain Decomposition for Spatial Interpolation in Grid Computing Environments. *Parallel Computing*, 29(10): 1481-1504.

49. NEESgrid (2002). NEESgrid :: Virtual Collaboratory for Earthquake Engineering. http://www.neesgrid.org/.

50. Sakai Project. http://www.sakaiproject.org.

51. Great Lakes Regional Center of Excellence. http://

52. HENP Networking Meeting Agenda (2001). http://lexus.physics.indiana.edu/griphyn/henpnet.html.

53. CERN (2000). The MONARC Project. http://monarc.web.cern.ch/MONARC.

54. iVDGL (2003). International Virtual Data Grid Laboratory. http://www.ivdgl.org/index.php.

55. Bunn, J.J. (2002). The GIOD Project. http://pcbunn.cithep.caltech.edu/.

56. TeraGrid (2002). Welcome to TeraGrid. http://www.teragrid.org/.

57. This is part of the Pan-European Ministerial Conference (http://www.wsis-romania.ro/ ), in preparation for the World Summit on the Information Society in Geneva (2003) and Tunis (2005).

58. LIVE (Laboratory for Immersive Visualization for the Environment) http://www.casa.ucl.ac.uk/virgin/cat/project25.htm

59. K. Berket, D. A. Agarwal, O.Chevassut, "A Practical Approach to theInterGroup Protocols", *Future Generation Computer Systems*, volume 18, number 5 (April 2002), pp. 709-719.

60. NEES Consortium (2003). Network for Earthquake Engineering Simulation. http://www.nees.org.

61. Science of Collaboratories Project (2003). Science of Collaboratories Home. http://www.scienceofcollaboratories.org.

62. MGRID (Michigan Grid Research and Infrastructure Development) http://www.mgrid.umich.edu

63. Grid Application Toolkit http://heppc1.uta.edu/atlas/software/user-tools.html

64. The Globus Project (2003) http://www.globus.org

65. Condor High Thoughput Computing (2003) http://www.cs.wisc.edu/condor/

66. MtrecC MPEG4 toolkit

# Facilities and Equipment for
# A Global Grid-Enabled Collaboratory for Scientific Research
# Proposal # 6347522

## *Leveraged Facilities at Caltech*

The Caltech Tier2 center, located at the Center for Advanced Computing Research (CACR), basically comprises four different clusters, namely the Caltech Production Cluster (PG), the Caltech Integrated Grid Testbed Cluster (IGT), the Caltech Developmental Grid testbed Cluster (DGT) and the Caltech Grid3 Cluster (Caltech-Grid3). The PG is used for full fledged CMS Monte-Carlo productions on a 24x7 basis, the IGT for intermediate production and data analysis, the DGT for development and tests of grid software and the Grid3 cluster for interoperability exercises and demonstrations among various experiments. . The main objective of the Tier2 center is to develop an infrastructure for research and development efforts in grid computing, high performance networking and data analysis needed for the highly data intensive LHC physics programme. The center is currently a fully participating component of various grid projects that include GriPhyN, iVDGL, PPDG and Grid3.

Here follows a detailed list of the equipment and facilities involved:

Caltech Production Grid Cluster (PG):

1. ACME 6012PE (Dual Intel P4 Xeon 2.4GHz) 12
   1U Rack-mounted server, 1GB PC 2100 ECC SDRAM Reg. memory, 1 10/100, 1 Gigabit ethernet and 1 Syskonnect Gigabit card, ATI Rage XL PCI graphic controller, 1 Maxtor IDE 80 GB 7200 RPM drive

2. Supermicro SC812I-400Ccase, 1U rackmount server 20
   on Supermicro X5DPE-G2 motherboard, dual Intel 2.8 GHz CPU, 1GB PC 2100 ECC Registerd DDR memory, 80 GB Maxtor IDE drive, dual Intel copper gigabit ports, 1 1.445 MB Floppy drive, 1 slim cdrom drive.

3. 4U Rack-mounted Disk Server based on Supermicro 2
   P4DPE-G2 motherboard with dual Intel P4 Xeon 2.4 GHz processors, Intel E7500 chipset, 2 GB of PC2100 ECC Registered DDR memory, Syskonnect gigabit card, ATI Rage XL PCI graphic controller, 2 3ware 7500-8 RAID controllers, 16 Western Digital IDE disk drives for RAID and 1 for system

4. Dell Powerconnect 5224 Rackmounted managed 2
   configurable network switch with 24 10/100/1000 Mbps (RJ-45 connector) and 4 SPF fiber ports

Caltech Integrated Grid Testbed Cluster (IGT):

5. Dell PowerEdge 4400 Rack-mounted server 1
dual Intel Xeon PIII 1 GHz processors
2 GB PC133 SDRAM main memory, 1 Intel
10/100 ethernet and 2 Syskonnect cards,
ATI 3D Rage graphics controller, Adaptec
SCSI 39160 and AHA-2940U2 cards, 7 SCSI
internal hard drives

6. 2U Rack-mounted compute nodes with dual 20
800MHz Pentium III processors, 512 MB memory
10/100 Ethernet, 2x36 GB disk drives

7. 4U Winchester FlashDisk OpenRAID rack-mounted 3
SCSI RAID storage with total capacity of 2.5 TB

8. Dell Powerconnect 5224 Rackmounted managed 1
configurable network switch with 24 10/100/1000
Mbps (RJ-45 connector) and 4 SPF fiber ports

9. APC Smart-UPS Model: 3000 UPS power supply unit 1

Caltech Developmental Grid Testbed Cluster (DGT):

10. A.Serv 1U-A1210 Rack Server, Dual AMD Athlon 1
1900+ processors on Tyan S2462NG K7 Thunder
motherboard, 1 GB PC2100 DDR ECC Register
memory, 1 Intel Pro/1000T Gigabit and one
3Com 10/100 ethernet ports

11. A.Serv 2U-2200 Rack Server, Dual AMD Athlon 3
1900+ processors on Tyan S2466N Tiger MPX
motherboard, 512 MB PC2100 DDR ECC Registered
memory, 1 Intel Pro/1000T Gigabit and one
3Com 10/100 ethernet ports, ATI XPERT PCI XL
video chipset

12. 1U Athlon based NAS storage server 1

13. Asante IntraCore 35160T gigabit switches 2

Caltech Grid3 Cluster (Caltech-Grid3):

14. 3U Disk Server on Supermicro X5DPE-G2 1
motherboard with dual Intel 2.40 Ghz CPUs,
2 GB ECC Registered memory, 1 3ware SATA RAID
controller, 8 SATA disk drives, 1 1.44 MB floppy
drive and 1 slim CDROM drive, dual Intel copper
gigabit ports

15. Supermicro SC812I-400C case, 1U rackmount server 4
on Supermicro X5DPE-G2 motherboard, dual Intel
2.4 GHz CPU, 1GB PC 2100 ECC Registerd DDR
memory, 80 GB Maxtor IDE drive, dual Intel
copper gigabit ports, 1 1.445 MB Floppy drive,

1 slim cdrom drive.

16. Dell Powerconnect 5224 Rackmounted managed 1
configurable network switch with 24 10/100/1000
Mbps (RJ-45 connector) and 4 SPF fiber ports

Network Servers at Caltech:

17. 1U rackmount server on Supermicro X5DPE-G2 1
motherboard, dual Intel 2.8 GHz CPU, 1GB
PC 2100 ECC Registerd DDR memory, 80 GB Maxtor
IDE drive, dual Intel copper gigabit ports,
1 1.445 MB Floppy drive, 1 slim cdrom drive.

18. ACME 6012PE (Dual Intel P4 Xeon 2.4GHz) 1
1U Rack-mounted server, 1GB PC2100 ECC
SDRAM Reg. memory, 1 10/100, 1 Gigabit
ethernet and 1 Syskonnect Gigabit card,
ATI Rage XL PCI graphic controller,
1 Maxtor IDE 80 GB 7200 RPM drive

Monalisa Servers at Caltech:

19. 1U rackmount server on Supermicro X5DPE-G2
motherboard, dual Intel 2.8 GHz CPU, 1GB 1
PC 2100 ECC Registerd DDR memory, 80 GB Maxtor
IDE drive, dual Intel copper gigabit ports,
1 1.445 MB Floppy drive, 1 slim CDROM drive.

20. 2U Rack-mounted node with dual Intel 800 Mhz 1
Pentium III processors, 512 MB memory
10/100 Ethernet, 2x36 GB disk drives

Network equipment at Starlight (Chicago):
Juniper T320
T320 Flexible PIC Concentrator (accepts PC series PICs)
1-port 10GBASE-LR Interface
T320 Flexible PIC Concentrator (accepts M160-FPC2 PICs)
2port Gigabit Ethernet PIC, SX Optics
Role: Production router connected to the transatlantic OC192 circuit

OSR-Cisco 7609:
Catalyst 6000 SU22/MSFC2 SERVICE PROVIDER W/VIP (supervisor)
1-port OC-48/STM-16 SONET/SDH OSM, SM-SR, with 4 GE
4-port Gigabit Ethernet Optical Services Module, GBIC
Catalyst 6500 16-port GigE module, fabric enable
Catalyst 6500 Switch Fabric Module (WS-C6500-SFM)
Role: Element of the multi-platforms testbed (Datatag project).

Cisco 7609:
Catalyst 6500 / Cisco 7600 Supervisor 720 Fabric MSFC3
6500 Series 4 Port 10GbE Module (4 XENPAC)

Catalyst 6500 10 Gigabit Ethernet Module with 1310nm long haul OIM and DFC card
Cat6500 16-port GigE mod, 2 fab I/F, (SX GBICs)
4000W DC Powr Sup for CISCO7609/13, Cat6509/13 chassis
Role: Element of the multi-platforms testbed (Datatag project).

Cisco 2950
24 10/100 ports + 2*1000BASE-SX ports
Role: Fast Ethernet switch for production with 2 Gbps uplinks.

Cisco 7507
One-port ATM enhanced OC-3c/STM1 Multimode PA (PA-A3-OC3MM)
One-port Fast Ethernet 100BaseTx PA (PA-FE-TX)
Two-port T3 serial PA enhanced (PA-2T3+)
One-port Packet/SONET OC-3c/STM1 Singlemode (PA-POS-SM)
Gigabit Ethernet Interface Processor, enhanced (GEIP+)
One-port Packet/SONET OC-3c/STM1 Singlemode (PA-POS-SM)
Role: Old router for backup and tests (IPv6 and new IOS software release tests).

Juniper M10
1port SONET/SDH OC48 STM16 SM, Short Reach w/Eje
2 ports PE-1GE-SX-B (2*1 port Gigabit Ethernet PIC, SX Optics, with PIC ejector)
Role: Element of the multi-platforms testbed (Datatag project). In particular, it is dedicated to level 2 services.

Extreme Summit 5i GbE
Gigabit Ethernet switch with 16 ports
Role: Network elements interconnection at Gbps speed.

Cisco 7204VXR
Two-port T3 serial PA enhanced (PA-2T3+)
Gigabit Ethernet Port Adapter (PA-GE)
Role: Old router for backup and tests.

Network equipment at Los Angeles (CENIC PoP)
Cisco 7606:
Catalyst 6500 / Cisco 7600 Supervisor 720 Fabric MSFC3
6500 Series 4 Port 10GbE Module (4 XENPAC) (part of the Ultralight proposal)
Role: Provide connection to research backbone.

Network equipment at Caltech (campus)
Cisco 7606:
Catalyst 6500 / Cisco 7600 Supervisor 720 Fabric MSFC3
6500 Series 4 Port 10GbE Module (4 XENPAC)
Catalyst 6500 10 Gigabit Ethernet Module with 1310nm long haul OIM and DFC card
Cat6500 16-port GigE mod, 2 fab I/F, (SX GBICs)
Role: Network element of the Grid Clusters.

**Caltech VRVS System**

1. 1 VRVS.ORG web server CPU Dual Pentium III (Coppermine) 1GHZ RAM 512M HD ~32G
2. 1 VRVS 2.5 demo and development server CPU Single Pentium III (Coppermine) 845MHZ RAM 512M HD ~33G
3. 1 VRVS 3.0 web server CPU Dual Intel(R) XEON(TM) CPU 2.20GHz RAM 2G HD ~65G
4. 1 VRVS MPEG2 MCU and web server CPU Single Pentium III (Coppermine) 800MHZ RAM 256M HD ~4.3G
5. 1 VRVS CALTECH reflector CPU Single Pentium III (Coppermine) 700MHZ RAM 256M HD ~16G
6. 1 VRVS 3.0 development server CPU Dual Intel(R) XEON(TM) CPU 2.40GHz RAM 1G HD ~65G
7. 1 VRVS 3.0 CALTECH StarLight reflector CPU Dual Intel(R) XEON(TM) CPU 2.40GHz RAM 1G HD ~65G

**Caltech MonaLisa Monitoring System**

1. 1 Locate at CACR: 1U rackmount servers based on Supermicro X5DPE-G2 motherboard, Dual Intel Xeon 2.8 GHz CPUs with 512K cache and 533 MHz FSB, 1GB PC2100 ECC Reg. memory, onboard Intel dual 82546EB gigabit Ethernet controller, Maxtor 80GB 7200 RPM hard drive, onboard ATIRAGE XL 8MB PCI graphics controller, slim FDD and CDROM drives.
2. 1 Locate at Chicago: 1U rackmount servers based on Supermicro X5DPE-G2 motherboard, Dual Intel Xeon 2.8 GHz CPUs with 512K cache and 533 MHz FSB, 1GB PC2100 ECC Reg. memory, onboard Intel dual 82546EB gigabit Ethernet controller, Maxtor 80GB 7200 RPM hard drive, onboard ATIRAGE XL 8MB PCI graphics controller, slim FDD and CDROM drives.

**Caltech TeraGrid Network Equipment at CACR**

1. 1 ONI Systems Online Metro DWDM
2. 1 Juniper T640: 3 STM-64/OC-192 SONET SMF-SR-2 3 Ethernet 10GBASE-LR
3. 1 Force 10 E1200:2 LC-ED-10GEL-2Y 10GBASE-LR module 1 LC-ED-RPM management module 6 LC-ED-1GE-24P 24 port GbE module

**Caltech TeraGrid Network Equipment at CENIC**

1. 1 ONI Systems Online Metro DWDM
2. 1 Juniper T640:3 STM-64/OC-192 SONET SMF-SR-2 3 Ethernet 10GBASE-LR

**Caltech servers at CERN (GVA), Starlight (CHI) and Los-Angeles (LA)**

Network servers at Starlight (Chicago)
6 4U Dual Xeon Nodes:

- Dual Intel® Xeon™ processors , 2.20GHz with 512K L2 cache
- SuperMicro P4DP8-G2Motherboard
- Intel E700 chipset

- 1 GB RAM, PC2100 ECC Reg. DDR,
- Onboard Adaptec AIC-7902 dual channel Ultra320 SCSI controller
- 2*Seagate 36.7GB SCSI 80pin 10KRPM Ultra 160
- On board Intel 82546EB dual port Gigabit Ethernet controller
- 2*SysKonnect Gigabit Ethernet card SK-9843 SK-NET GE SX
- 4U Rack-mounted server
- 420W

3 4U Dual Xeon Nodes:

- Dual Intel Xeon processors, 3GHz with 512K L2 cache
- SuperMicro X5DPE-G2 Motherboard
- 2 GB RAM, PC2100 DDR ECC Registered
- 2*3ware 7500-8 or 7506-8 RAID controllers
- 16 WD IDE disk drives for RAID and 1 for system
- 2 Intel 82550 fast Ethernet
- 1 Intel 10GE LR PCI Card
- 4U Rack-mounted server

Network servers at Los-Angeles (CENIC PoP)
1 4U Dual Opteron Node:

- Dual processor AMD Opteron 244 - 1.8Ghz
- 1 Maxtor 80Gb IDE Boot Drive and 16x160GB Seagate 150 7200RPM SATA Drives
- 2 3ware 8500-8 SATA IDE Raid Controllers
- 2 onboard Broadcom BCM5704 Gigabit Ethernet Ports
- 4U Rack-mounted server

2 4U Dual Xeon Nodes:

- Dual Intel Xeon processors, 3GHz with 512K L2 cache
- SuperMicro X5DPE-G2 Motherboard
- 2 GB RAM, PC2100 DDR ECC Registered
- 2*3ware 7500-8 or 7506-8 RAID controllers
- 16 WD IDE disk drives for RAID and 1 for system
- 2 Intel 82550 fast Ethernet
- 1 Intel 10GE LR PCI Card
- 4U Rack-mounted server

1 1U Dual Xeon 3.20GHz Node:

- Dual Intel Xeon processors, 3.06GHz with 1MB L3 cache
- SuperMicro X5DPE-G2 Motherboard
- 2 GB RAM, PC2100 DDR ECC Registered
- 120G Maxtor 7200rpm HD
- 1 Intel 10GE LR PCI Card
- 1U Rack-mounted server

9 1U Dual Xeon 2.40GHz Nodes:

- Dual Intel Xeon processors, 2.40GHz with 512K L3 cache
- SuperMicro P4DPR-I Motherboard
- 1 GB PC2100 DDR ECC Registered
- 2 Intel 82550 fast Ethernet on board
- Hard Drive : 80GB IDE, Maxtor, 7200 RPM
- 1*SysKonnect Gigabit Ethernet card SK-9843 SK-NET GE SX

Network servers at CERN:
2 4U Dual Xeon Nodes:

- Dual Intel® Xeon™ processors , 2.20GHz with 512K L2 cache
- SuperMicro P4DP8-G2Motherboard
- Intel E700 chipset
- 1 GB RAM, PC2100 ECC Reg. DDR
- Onboard Adaptec AIC-7902 dual channel Ultra320 SCSI controller
- Hard drive : Seagate 36.7GB SCSI 80pin 10KRPM Ultra 160
- On board Intel 82546EB dual port Gigabit Ethernet controller
- 2*SysKonnect Gigabit Ethernet card SK-9843 SK-NET GE SX
- 4U Rack-mounted server
- 420W

3 4U Dual Xeon Nodes:

- Dual Intel® Xeon™ processors , 3.02 GHz with 512K L2 cache
- SuperMicro P4DPE-G2 Motherboard:
- 2 GB RAM, PC2100 ECC Reg. DDR
- 2 3ware 7500-8 RAID controllers
- 16 WesternDigital IDE disk drives for RAID and 1 for system
- 2 Intel 82550 fast Ethernet
- 1*SysKonnect Gigabit Ethernet card SK-9843 SK-NET GE SX
- 10 GE Intel card.
- 4U Rack-mounted server
- 480W to run 600W to spin up

3 4U Dual Xeon Nodes:

- Dual Intel® Xeon™ processors , 2.40GHz with 512K L2 cache
- SuperMicro P4DP8-G2Motherboard
- Intel E700 chipset
- 2 GB RAM, PC2100 ECC Reg. DDR,
- Hard drive: 1 x 140 GB - Maxtor ATA-133
- On board Intel 82546EB dual port Gigabit Ethernet controller
- 2*SysKonnect Gigabit Ethernet card SK-9843 SK-NET GE SX
- 4U Rack-mounted server
- 420W

**THE UNIVERSITY OF IOWA** has a proof-of-concept Compact Muon Solenoid (CMS) Grid Tier-2 center. Hardware resources include a specialized Grid system for Large Hadron Collider Computing Grid (LCG)-0 testbed and three Linux clusters that are connected through The University of Iowa campus broadband fiber network. The first cluster is dedicated as a computing farm with 16 dual-PIII computing nodes. The second cluster is dedicated as both a storage system (0.6Tbyte) and a computing farm (3 dual-Athlon 2200+ computing nodes), which will be upgraded to serve as an Integration Grid Testbed (DGT) and a Production Grid (PG) for CMS. The third cluster is currently serving as a CMS Development Grid Testbed (DGT), which includes 20 Pentium4 CPUs. Software installed on these clusters includes basic Grid middleware and CMS tools for data simulation and analysis.

**LAWRENCE BERKELEY NATIONAL LABORATORY's** NERSC Division has established various computing and storage systems for research and development, in addition to desktop workstations running a variety of operating systems (e.g. linux, windows, solaris, freeBSD, and mac OSX) used by members of Computational Research Division's (CRD) Distributed Systems Department.

An IBM RS/6000 SP, makes up the heart of NERSC's computer hardware capability. NERSC's 3,328-processor IBM RS/6000 SP has 208 16-CPU POWER3+ SMP nodes with a peak performance of 5 teraflop/s, making it one of the most powerful unclassified supercomputer in the world. Each node has a common pool of between 16 and 64 gigabytes of memory, and the system has 20 terabytes of disk space.

Additional capabilities are provided by two special-purpose servers: a cluster of four PCs for numerical and statistical processing, and a dedicated Silicon Graphics computer for scientific visualization from remote locations.

NERSC has two Access Grid nodes, and is also equipped for VRVS.

NERSC's research in data-intensive computing is grounded in their operation of a major production facility, the PDSF (Parallel Distributed Systems Facility). The PDSF is a 390-processor Linux cluster used by large-scale high energy and nuclear physics investigations for detector simulation, data analysis, and software development. The PDSF's 48 disk vaults provide a total 35 TB of data storage.

NERSC's research into cluster architectures, the PC Cluster Project, is focused on two systems, on which they are developing the software infrastructure needed to use commodity hardware for high performance computing:

1) the 36-node PC Cluster Project Testbed, which is available to NERSC users for trial use
2) the 12-node Alpha "Babel" cluster, which is being used for Modular Virtual Interface Architecture (M-VIA) development and Berkeley Lab collaborations.

Access to NERSC from anywhere in the U.S. or the world is available through ESnet, which provides OC-12 bandwidth to NERSC and Lawrence Berkeley National Laboratory.

**FLORIDA INTERNATIONAL UNIVERSITY** hosts a VRVS reflector in the NAP of the Americas, which serves the AMPATH project participants and could be used for this project. Additionally, FIU operates the AMPATH project that has as resources 10 DS-3 (45 Mbps) circuits to connect research and education networks in Latin America and the Caribbean to US and non-US Research and Education networks. The AMPATH network and International Exchange Point will enable wide testing of collaboratory tools.

**The Inter-regional Center for High-Energy Physics Research Education and Outreach (CHEPREO)** at FIU: FIU is the lead institution, in collaboration with Caltech, University of Florida and Florida State University, proposing to develop an inter-regional Center for High-Energy Physics Research Education and Outreach (CHEPREO). CHEPREO is a 5-year NSF funded program (MPS 0312038), that in year 1, would establish an STM-4 (622 Mbps) SDH-based transport service between Miami and Rio. CHEPREO would also establish a Tier3 CMS Grid facility in Miami for FIU. The STM-4 circuit would be used for research and experimental networking activities, and production. Through the CHEPREO program, AMPATH will leverage the availability of an experimental network to South America. Likewise, the Tier1 facility in Rio and Brazil's HENP community will access the Global Grid community and participate in data-intensive Grid collaborations. Figure 2 shows how the emergent Grid Physics Tier1 facility at the State University of Rio de Janeiro (UERJ) would be connected to Miami.

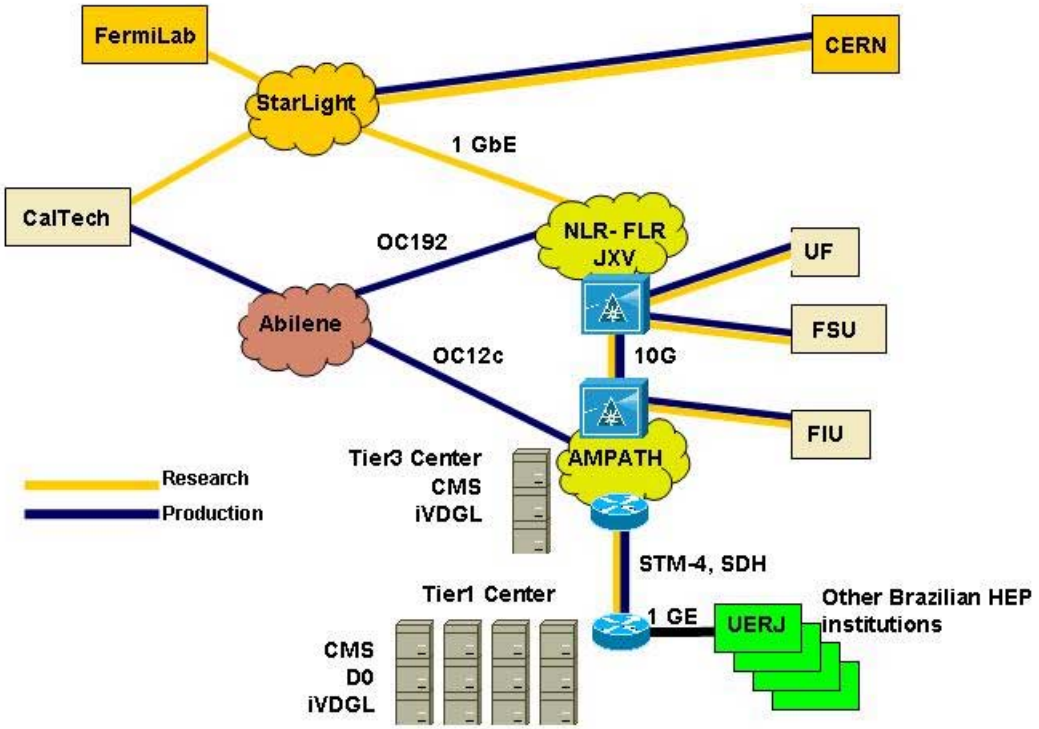



**Figure 2: International Connection to Brazil (years 1 and 2)**

By year 3 or possibly sooner, the load on the inter-regional STM-4 circuit is expected to reach capacity. As soon as possible, this circuit should be upgraded to a 2.5G or 10G wavelength and a Layer2 connection extended to South America, as is to other International Peer Sites. The following figure shows Layers 2 and 3 equipment, by which South America can connect to the AMPATH International Exchange Point.

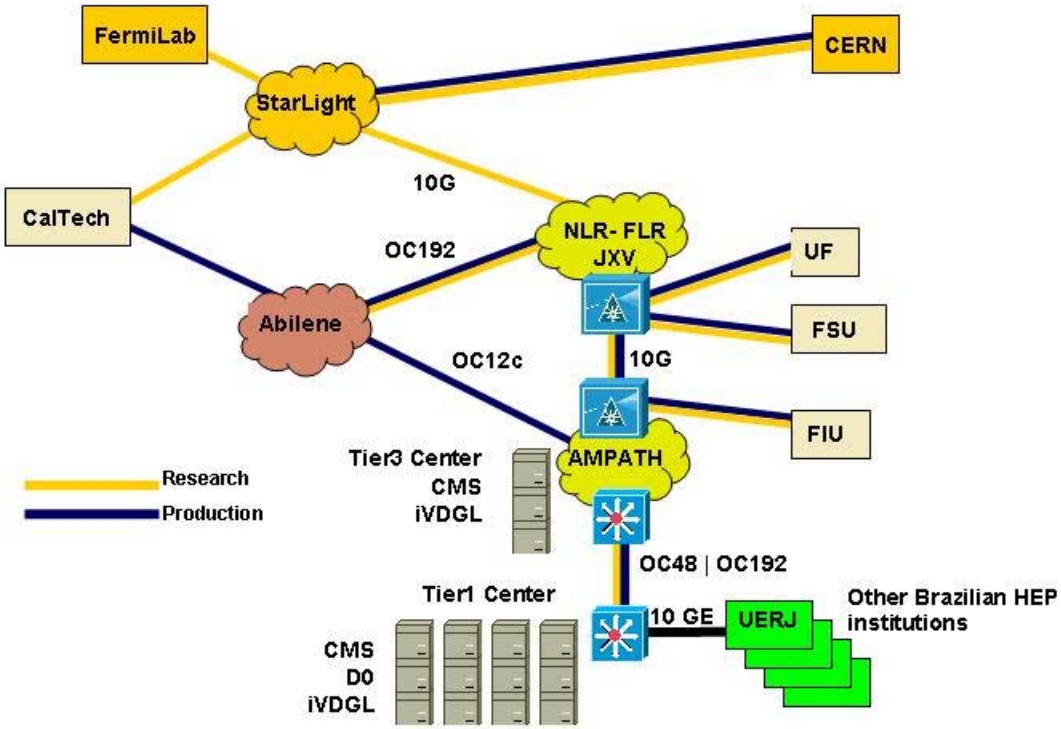



**Figure 3: International Connection to Brazil (years 3-5)**

**UNIVERSITY OF MICHIGAN** hosts one of eight US ATLAS GRID testbed sites. We plan to install and test the collaboratory environment on some of our gigabit connected, dual processor systems that make up the testbed (those associated with and accessible by members of our group). We have 12 dual cpu systems, with memory ranging from 256 Mbytes up to 1 Gigabyte per system. Our storage systems are a combination of SCSI and IDE RAID 5 configurations totaling over 4.5 terabytes of storage. We plan to also test the environment on our desktop systems for our collaborators. These are typically linux or Windows systems with Pentium III or Pentium 4 processors and are in offices in Physics, the School of Information and the Media Union.

## Michigan US ATLAS Grid testbed

Additionally we will involve the MGRID campus infrastructure in our deployment. MGRID (mgrid.umich.edu) shares a campus-wide gigabit test/development network, soon to be upgraded to a 10-gigabit backbone. There are numerous sites participating in MGRID; the US ATLAS grid testbed is one. Other MGRID sites include NPACI in Engineering, the Center for Advanced Computing, the Visible Human Project, the Michigan Center for Biological Information, the Mental Health Research Institute, the Center for Information Technology Integration, and the School of Information.

Each site has significant computational and storage resources and has expressed an interest, through MGRID, of participating in testing and early adoption of our collaborative environment.

**THE UNIVERSITY OF TEXAS AT ARLINGTON** has access to numerous computing resources on campus, some dedicated and some shared clusters, which are used for grid computing in High Energy Physics. The newest resource is the Distributed and Parallel Computing Cluster (DPCC) which contains more than 80 dual processor worker nodes, funded by a the NSF-MRI program. This cluster is part of the U.S. ATLAS computing testbed (and Grid3), and the Southern Analysis Region for D0 (D0-SAR). It is also used for Computer Science research into Data Mining, Parallel Database Development, Network simulations, and Video processing. The cluster contains 81 dual-Xeon compute nodes, 2 head nodes (one of which is dedicated to grid computing), 50 TB of available network disk storage and a 1 gigabit network switching infrastructure.

All of the DPCC compute nodes are dual Xeon based systems with 2 gigabytes of RAM, an IDE based hard drive, dual 1 Gbs network interfaces and running the Linux operating system. Thirty two of the nodes operate at 2.4 GHz and use 60 GB disk drives while the remaining 49 operate at 2.6GHz and are equipped with 80 GB disk drives. The 50 terabytes of storage are found on 11 RAID units for disk storage. Ten of the units are Linux based systems with 4.5 TB of IDE-based disk space served to the cluster through NFS. The last unit is a 5.2 TB unit utilizing Fiber Channel disks and Storage Area Network. Three file servers are also attached to the SAN and implement a Distributed file system to cooperatively manage the storage. These three servers, in turn, offer NFS services to the cluster.

The network infrastructure for the cluster is based on 1-Gbs copper interconnects. The core of the network is the Foundry Networks' FastIron 800 Layer 2/3 switch. Directly attached to the switch are the head nodes, the distributed file system server nodes, 32 compute nodes and five of the IDE based RAID units. The remaining equipment is organized into five groups each with a 1-Gbs switch connected to the core switch.

HEP also has access to a university wide High Performance Computing center with 114 nodes and 6TB storage. This system was effectively used to produce events for the ATLAS Data Challenge 1, using the GRAT grid software. Finally, we have a dedicated Linux farm with about 50 processors, which is used to generate Monte Carlo Data for the D0 experiment at Fermilab. All of these resources will provide a valuable testbed for the GESCR proposal.

THE UNIVERSITY OF IOWA

OFFICE OF THE VICE PRESIDENT FOR RESEARCH

201 Gilmore Hall
Iowa City, Iowa 52242-1320
319-335-2119 *Fax* 319-335-2104

February 20, 2004

Homer A. Neal
University of Michigan Physics Department
500 East University
Ann Arbor, MI, 48109-1120

Dear Dr. Neal:

I am writing in support of the NSF ITR proposal for an integrated collaboration environment, known as "A Global Grid-Enabled Collaboratory for Scientific Research (GECSR)."

NSF's investments in national infrastructure have made high speed networks, high-density disk storage and high-performance computational clusters accessible to many researchers, but the ability to use these systems in a truly collaborative way is hampered by the lack of tools. This proposal addresses these deficiencies by developing new tools to allow physicists to collaborate in a flexible, dynamic, and distributed environment.

The University of Iowa has a combination of people, infrastructure and need that makes our contribution to this proposal important. The Grid Research and educatiOn group at IoWa (GROW) is leading our efforts to bring grid computing and related collaborative technologies to Iowa and surrounding regions. Past collaborations between GROW and high-energy physicists on campus provide a solid basis for the work outlined in the current ITR proposal. University of Iowa membership in Internet2 enhances our ability to be integrated into the Grid infrastructure of this proposal. Iowa's location, away from major population centers, makes the development of the tools being proposed essential to our continued involvement in multi-institutional research projects.

I am excited by this proposal and Iowa's involvement in it. It has my enthusiastic recommendation and support.

Sincerely,

William F. Decker
Interim Vice President for Research

/jrk

# Fermilab

Lothar A. T. Bauerdick
Project Manager U.S. CMS Software & Computing
Fermi National Accelerator Laboratory
P.O.Box 500 • MS234
Batavia, IL • 60510 • Phone 630-840-6804 • Fax -2783

February 18, 2004

To Whom It May Concern:

**Re: Proposal "ITR: Global Grid-Enabled Collaboratory for Scientific Research"**

As leader of the U.S. CMS Software and Computing project I wish to express my support for the ITR proposal to the NSF by Prof. Homer Neal of U.Michigan and a group of computer scientists and physicists at several Universities, collaborating on the CMS and Atlas experiments at the Large Hadron Collider. They propose to develop new technologies and build an operational Grid-Enabled Collaboratory for scientific research.

The LHC at CERN in Switzerland will provide one of the most exciting, challenging, and relevant pieces of science over the next 20 years. U.S. scientists are eagerly looking forward to harvest the scientific yield of many years of preparations and a half billion dollar U.S. investment, when in a few years the LHC starts to collide particles. The LHC's scientific and technical challenges are daunting, but as importantly there will be large cultural and organizational issues to solve when several thousand scientists around the world will start to collaborate on a single, multi-purpose facility with a very broad scientific program.

The GEC SR project sets out to provide an environment for collaboration between science communities. This could very well be the cornerstone for the success of the LHC scientific program, specifically for the U.S. university groups that are separated from the experiment by an ocean and six to nine time zones. Easing participation of U.S. scientists in the LHC research program by enabling collaboration over the distance will open unique new opportunities for discovery to smaller and remote groups, whose intellectual engagement in a "Big Science" project is much desired.

The GEC SR proposal is very well setup to achieve these goals and thus have an important impact. It is a strong partnership between application and computing researchers bringing different skills and expertise. An iterative approach with both a strong research component and a strong deployment and integration component will allow them to make substantial progress over the four years of the project.

The proposal goals and approaches are fully aligned with the mission of the US CMS Software and Computing project, which is part of the U.S. LHC Research Program that the NSF is funding jointly with the DOE. The U.S. CMS project provides the software and computing resources to the U.S. physicists and plans to setup a large "Tier-1" computing center at Fermilab and several "Tier-2" regional centers at U.S. institutions. These regional centers will become an integral part of the "LHC Computing Grid", spanning over the whole globe, and allowing scientists at Universities and labs to fully participate in the research program, without necessarily having to travel to CERN.

U.S. CMS physicists are already now particularly reliant on videoconferencing for meetings and participation in the daily activities of the experiment. Very soon U.S. CMS will start to operate a remote "shift taking" facility in the U.S., including in-depth monitoring of the detector and the data acquisition system, some of which is starting now with test beams and detector commissioning.

All these plans count on the availability of a high quality remote collaborative environment that this project is setting out to deliver. These new capabilities clearly have the potential to revolutionize the way we will do science in the LHC era.

Thus the U.S. CMS Software and Computing project is going to work very closely with the GEC SR project and will support it in any way necessary to reach its goals.

I therefore endorse the proposal in the strongest possible terms.

Sincerely,

Lothar A T Bauerdick
Fermilab
Project Manager U.S. CMS Software and Computing

California Institute for Telecommunications and Information Technology
A UCSD/UCI PARTNERSHIP
www.calit2.net

February 19, 2004

Dr. Homer Neal
Director, UM-ATLAS Collaboratory Project
Professor, Physics Department
Samuel A. Goudsmit Professorship of Physics
University of Michigan
2477 Randall Lab 1120
Ann Arbor, MI 48109-1120

Dear Dr. Neal:

This letter is in support of the ITR proposal entitled: "Global Grid-Enabled Collaboratory for Scientific Research (GECSR)." This proposal has a strong multi-disciplinary and multi-institutional team assembled to build a Grid system to enable a new level and new modes of scientific collaboration on a global scale. It builds on many proven tools, and integrates them into a working system using a proven framework (CHEF) that will be tested out by several disciplines. One of the strengths of this proposal is that a number of your team are drawn from leading researchers from the Grid and network application worlds that have interacted tightly with the computer science and engineering leaders of the Grid infrastructure movement.

One of the attractive features of this proposal is that it builds on work ongoing with the NSF funded Network for Earthquake Engineering Simulation (NEES) and its supporting NEESgrid. However, it then enlarges the set of tools and blends them with those already proven out by the High Energy and Nuclear Physics community. It will also employ a systematic, iterative evaluation process, including ongoing feedback from early adopters in the user community throughout the development phase, based on the expertise and methodology These factors give the proposal lower risk, and increase the likelihood that it will yield a system that users adopt. The lack of involving users early in the design of collaborative systems is one of the failure modes that has dogged such efforts over the last decade, such as two I oversaw, NCSA Collage and NCSA Habanero, while I was director of the National Center for Supercomputing Applications.

I am now director of the California Institute for Telecommunications and Information Technology [Cal-(IT)$^2$], a joint venture between two campuses, UC Irvine and UC San Diego. The mission of Cal-(IT)$^2$ is to team faculty, students, and research professionals with leading California telecommunications, computer, software, and applications companies in order to advance knowledge and create a stronger California economy. Cal-(IT)$^2$ encourages the development of "living laboratories" for researchers to investigate how trends in IT and telecommunications over the next decade will transform educational practices, environmental science and civil infrastructure, intelligent transportation, digitally enabled genomic medicine, and the new media arts.

LARRY SMARR
Director

University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0405

TEL: (858) 822-4284
FAX: (858) 822-3906
EMAIL: lsmarr@ucsd.edu

California Institute for Telecommunications and Information Technology
A UCSD/UCI PARTNERSHIP
www.calit2.net

I will be following the results of this research very closely, because it may well produce a system that we adopt in support of our multi-institutional projects.

Sincerely,

Larry Smarr

Larry Smarr
Harry E. Gruber Professor, Department of Computer Science and Engineering, UCSD
Director, California Institute for Telecommunications and Information Technology

LARRY SMARR
Director

University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0405

TEL: (858) 822-4284
FAX: (858) 822-3906

EMAIL: lsmarr@ucsd.edu

February 17, 2004
Dr. Homer Neal
Department of Physics
University of Michigan
Ann Arbor, Michigan 48109

Dear Homer:

I am writing in support of your NSF proposal for an integrated collaboration environment, known as "A Global Grid-Enabled Collaboratory for Scientific Research (GEC SR)." Emerging approaches to computational science require real-time computing support, large-scale analysis and management of terabytes of observational data. Concomitantly, they require mechanisms for physically and temporally distributed groups, often containing thousands of collaborators, to jointly coordinate data analysis and experimentation. I believe the proposed work will help simplify distributed collaboration and interaction, a critical component of national and international scientific teams.

Traditional collaboration and teleconference systems impose a simplistic and constraining model of distributed interaction that is not well suited to the dynamic, and often informal information sharing and discussion central to scientific research. Your proposal addresses precisely these issues, promising to create a flexible collaboration infrastructure that is tested and validated on an international scale.

I am also delighted to see that the proposal includes a well-balanced mix of application and computing researchers who all have an outstanding track record of collaboration and software development. The grid infrastructure research and developments of the GriPhyN and iVDGL projects have had profound impact on the international physics community and its ATLAS and CMS projects.

This work is also reflective of NCSA and the Alliance's broad vision of a distributed Grid infrastructure based on commodity hardware and widely distributed community software, together with scalable distributed collaboration systems. Based on this vision, as part of the NSF TeraGrid, NCSA will deploy over 10 teraflops of Linux cluster computing capability and ~500 TB of storage, connected by a 40 Gb/s transcontinental network to additional TeraGrid resources and enabling distributed data management, computation, and visualization. We look forward to working you on high-energy physics data analysis atop the TeraGrid.

In addition, the Alliance Scientific Workspaces of the Future (SWOF) expedition is developing a collaboration infrastructure that combines the next generation Access Grid with distributed information sharing, visualization and collaboration support. This expedition is in the same spirit as your proposed work, and I also look forward to fruitful collaborations between the GEC SR team and the Alliance.

In summary, I believe the vision outlined in this proposal exemplifies the future of scientific computing: distributed, data driven, and community based. It has my enthusiastic recommendation and support.

Sincerely,

Robert L. Pennington
Interim Director, National Center for Supercomputing Applications

# California Institute of Technology

Dr. Daniel I. Meiron
Applied Mathematics 217-50
Firestone Laboratory
Pasadena, CA 91125

Tel: (626) 395-4563
Fax: (626) 578-0124
e-mail: dim@its.caltech.edu

February 17, 2004

Prof. Homer A. Neal
University of Michigan

Dear Prof. Neal:

I am writing to add the support of CACR for the proposed work described in your revised proposal entitled "A Global Grid-Enabled Collaboratory for Scientific Research"

I am currently serving as the Acting Director of Caltech's Center for Advanced Computing Research. As you know CACR is dedicated to the pursuit of excellence in the field of high-performance computing, communication, and data engineering. Major activities include carrying out large-scale scientific and engineering applications on parallel supercomputers and coordinating collaborative research projects on high-speed network technologies, distributed computing and database methodologies, and related topics. The proposed work is in perfect alignment with the CACR mission.

We have enjoyed a strong and successful partnership between Caltech's HEP and CACR since 1996, on a variety of projects related to Petascale database access and analysis, Grid systems, and most recently high performance networks. The development of a globally scalable Grid-enabled Collaboratory for scientific work is a natural and important next step that could be of substantial mutual benefit as regards CACR's participation in the Teragrid project. CACR stands ready to aid this work in a variety of ways such as support for high performance computational facilities and networking. If I can be of any assistance in this regard please do not hesitate to contact me.

Sincerely,

Dan Meiron
Acting Director, CACR, Caltech

pervasivetechnologylabs
AT INDIANA UNIVERSITY

February 18, 2004

Dear Homer:

It is my pleasure to endorse the GECSR project submitted to the NSF ITR program. Such collaborative capability is essential to realize the promise of Grids which up to now have concentrated on largely asynchronous sharing of computer and data resources. It is a central part of the idea that Grids support all aspects of a virtual organization including people as well as electronic resources. The lead institutions are highly qualified: CHEF from Michigan and VRVS from Caltech are leading projects in their areas of collaboration. Caltech through CMS has played a major role in projects like GriPhyN, iVDGL, PPDG and recent work on the development of a Grid-enabled Analysis Environment using Web services. These have pioneered critical demonstrations that Grids are important. I am already working with CHEF as part of the MSF Middleware Initiative and I would be happy to further work with your team and contribute ideas and technologies from our Global MMCS (Multi Media Collaboration system) project with Beihang University which focuses on Collaborative Web Services and Collaboration as a Web Service.

Sincerely,

Geoffrey C. Fox

Geoffrey C. Fox
Professor of Computer Science, Informatics & Physics
Director, Community Grids Laboratory
Pervasive Technology Labs at Indiana University

*501 North Morton Street, Suite 224*
*Bloomington, Indiana 47404-3730*
*812-856-1242 Fax 812-856-1537*

CANARIE INC.

Dr. Homer A Neal
University of Michigan Physics Department
500 East University
Ann Arbor, MI, 48109-1120

February 17, 2004

Dear Homer:

I am delighted to be able to express my support for the "Global Grid-Enabled Collaboratory for Scientific Research (GEC SR)". Since the early days of such concepts as collaboratories and grids, the high energy physics community has always been seen as leader in advancing these technologies and concepts.

The ability to harness the expertise of such a collection of institutions combined with existing tools, and developmental abilities in collaborative systems and working methods, Grid and distributed software development and optimization, high performance networking, and related areas, and to work with leading computer scientists as well as major HENP laboratories and the physics community, I am confident will ensure the program's success and effectiveness.

I believe GEC SR could have a transformative impact on our research in the fields of High Energy and Nuclear Physics, as well as many other fields within and beyond the bounds of research. The work program outlined in the proposal is indeed essential for the success of our current and next-generation global scientific collaborations. If these developments are applied broadly, they will have significant benefits to research in many fields, to education in the form of new learning environments, and more broadly to enable collaboration and collaborative work across networks on a global scale.

As you know CANARIE and the Canadian research community has extensive interest in the development of these tools as well and we would look to future collaboration with your team upon successful award funding.

Sincerely

Bill St. Arnaud, P.Eng.
Senior Director Network Projects
CANARIE Inc

110, rue O'Connor Street, 4e étage / 4th Floor, Ottawa, Ontario, Canada K1P 5M9
Tel./Tél.: (613) 943-5454 • Fax./Télécopieur: (613) 943-5443 • e-mail: info@canarie.ca • www.canarie.ca
CANADA'S ADVANCED INTERNET DEVELOPMENT ORGANIZATION
LE FER DE LANCE DE L'INTERNET ÉVOLUÉ AU CANADA

Professor John Huth
236 Lyman Laboratory
Harvard University
Cambridge, MA 02138

617 495 8144
617 495 0416 (fax)

huth@physics.harvard.edu

24 Feb, 2004

Professor Homer Neal
Physics Department
500 East University
Ann Arbor, MI 48109-1120

Dear Professor Neal:

As the Associate Project Manager for Physics and Computing for the U.S. ATLAS Project, I want to strongly endorse the ITR proposal entitled "The Global Grid-Enabled Collaboratory for Scientific Research". I also spoke with my executive program-manager, Dr. James Shank, who agrees with me on the importance of this proposal. The effectiveness of global collaborations, such as those engaged in the Large Hadron Collider (LHC) depends on tools to enable remote users to participate in a meaningful way. The research and proposed here directly address this issue.

Historically, collaborators located at the central location of an experiment have had an undeniable advantage over remote collaborators. With new grid technologies becoming available, and high bandwidth links, it is clear that there is a substantial opportunity to empower remote users to become highly effective collaborators. The group participating in this proposal already has substantial experience in collaboratory efforts. I believe that they are in an excellent position to understand the issues in remote collaboration and have a sensible approach to the problem. My collaboration stands to benefit highly, if they realize the goals set forth. I strongly support this proposal.

Sincerely,

John Huth

# UNIVERSITY OF ILLINOIS AT CHICAGO

**Electronic Visualization Laboratory**
**851 South Morgan Street, Room 1120 (MC 152)**
**Chicago, IL 60607-7053**

**February 16, 2004**

Homer A. Neal
University of Michigan
Ann Arbor, MI, 48109-1120

Dear Homer:

I would like you to add my support to your ITR proposal entitled "PIF: Global Grid-Enabled Collaboratory for Scientific Research." It is a well-conceived proposal with excellent researchers and, in my opinion, it will make long-term and significant improvements to collaborative science and science education.

Your proposal aims to develop needed international Internet services towards building a global cyberinfrastructure. Major NSF, European Union (EU) and Asian initiatives support a number of Grid-intensive applications which will require, over the next several years, distributed terascale computing, petabyte storage and multi-gigabit networks. The major collaborations of High Energy and Nuclear Physics (HENP) are foremost examples of these initiatives; indeed, this proposed Global Grid-Enabled Collaboratory project will demonstrably improve the operational and functional capabilities of the international Internet and maintain the momentum of many related efforts for the benefit of the NSF research and education communities in Physics and Computer Science.

Scientist-empowered "LambdaGrids" may become the basis for the wired Internet infrastructure underlying future customer-driven e-science, education, emergency services, health services and commerce, but it is a technology that needs major work to be fully fleshed out. Technical innovation in optical networking is continuing to progress nicely. This is an opportune time for the e-science community to scale up research and development exploiting ways to achieve fully networked applications. Perfected techniques may eventually allow commercial providers to offer such services profitably, which may help revitalize the software and services industries by taking "broadband" into a new mass market. The new forms of persistent and pervasive collaboration you are proposing, coupled to dynamic access and analysis of worldwide-distributed petascale data, will help usher in a new era of such applications.

On behalf of the StarLight optical Internet exchange here in Chicago, I assure you we will help you create collaboration and computing environments with high-performance connectivity. I look forward to your HENP-based next generation information system and infrastructure project drawing in many collaborators to systematically enhance scientific productivity through communication, education and outreach.

Sincerely,


Dr. Thomas A. DeFanti
PI, StarLight
Distinguished Professor of Computer Science
Director, Electronic Visualization Laboratory

**UIC**

Phone (312) 996-3002 • Fax (312) 413-7585

3025 Boardwalk, Suite 100
Ann Arbor, Michigan 48108
734.913.4250 – 734.913.4255 Fax
www.internet2.edu

February 24, 2004

Homer Neal, Ph.D.
Samuel A. Goudsmit Distinguished
University Professor of Physics
2477 Randall Laboratory
University of Michigan
Ann Arbor, MI 48109-1120

Dear Homer:

On behalf of Internet2, I am pleased to write in strong support of the ITR proposal entitled A Global Grid-Enabled Collaboratory for Scientific Research that you are submitting to the National Science Foundation.

This proposal represents an extremely effective collaboration by the University of Michigan, Caltech, the University of Texas, University of Maryland, University of Iowa, Florida International University, Fermilab, the Lawrence Berkeley Laboratory and the University of Oklahoma. Together, this dynamic group will develop a next-generation collaborative system for research, harnessing multi-institutional expertise, technological tools, and developmental capabilities.

The proposal's strengths include its:

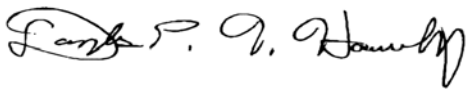
- Effective collaborative infrastructure linking leading computer scientists from cutting-edge research universities, researchers from major HENP laboratories, and the physics research community as a whole;
- Unique combination of joint efforts addressing collaborative systems and working methods, Grid and distributed software development and optimization, and high-performance networking;
- Unprecedented focus on collaborative Grid-enabled and other analysis tools to facilitate data-intensive analysis involving worldwide-distributed resources within a global research community; and
- Enhancement of the cooperative and productive HENP collaborative infrastructure that will likely serve as a model for other multi-organizational efforts.

This proposal represents a partnership between Grid technologies and the HENP model of collaboration. It is likely to create models for many other scientific areas involving current and next-generation data-intensive global collaborations. The proposal's research outcomes are also likely to have a significant impact on education through the creation of new learning

environments, and on other multi-institution organizational infrastructures to enhance investigative interaction across global networks.

We are pleased to offer our support of this initiative.

Sincerely,



Douglas E. Van Houweling
President and CEO, Internet2

Dr. Homer A. Neal
Department of Physics
University of Michigan
Ann Arbor, Michigan 48109
USA

Dear Dr. Neal,

I am writing to provide my endorsement of the proposal "Global Grid-Enabled Collaboratory for Scientific Research (GECSR)" The Large Hadron Collider experiments will require an unprecedented level of collaboration as they pursue fundamental studies of the structure of matter involving thousands of physicists from over 150 institutes around the world. The years of effort required to design, prototype, build, commission and use the particle detectors associated with these experiments involve an enormous number of human interactions at the individual, group, and collaboration-wide levels. Current collaboration tools are not sufficient to meet the needs of this scale of focused human activity, and I am pleased to see that a group of competent scientists are taking on the issue of determining what new tools are needed and how they should be designed and deployed.

I endorse the plan of the proposal for creating a collaborative environment specifically designed to facilitate the work of scientists embarked on a discovery mission in a globally distributed setting. That environment should incorporate the very best technologies available to permit colleagues to easily communicate, to access any needed information, and to enhance interactions in the analysis environment of Grid computing. Proceeding to achieve this goal by utilizing advanced knowledge from the science of collaboratories and tightly integrating the development of the collaborative environment with the desires of the user community, is a laudable approach.

CERN has recognized the importance of collaborative tools and has recently established a "Requirements Technical Advisory Group(RTAG)" to explicitly look into what CERN should do to support the collaboratory tool needs of the four experiments in the LHC. Work of this group should complement the plans articulated in the GECSR proposal where the creation of an overall collaboration environment is envisioned.

I am aware of the work on collaborative tools conducted by the Michigan and Caltech groups and would expect that they and their colleagues have the expertise and institutional support required to successfully carry out the proposed work.

The GECSR project would represent a valuable contribution to the success of large scale collaborative research projects, an area of growing importance to high energy physics and to many other fields of science as well, as the complexity of modern experiments require the involvement of a large number of scientists from many countries around the world.

Sincerely yours,

Alberto Pace
Internet Services Group Leader
Information Technology Department
CERN European Organization for Nuclear Research
CH-1211 Geneva 23, Switzerland