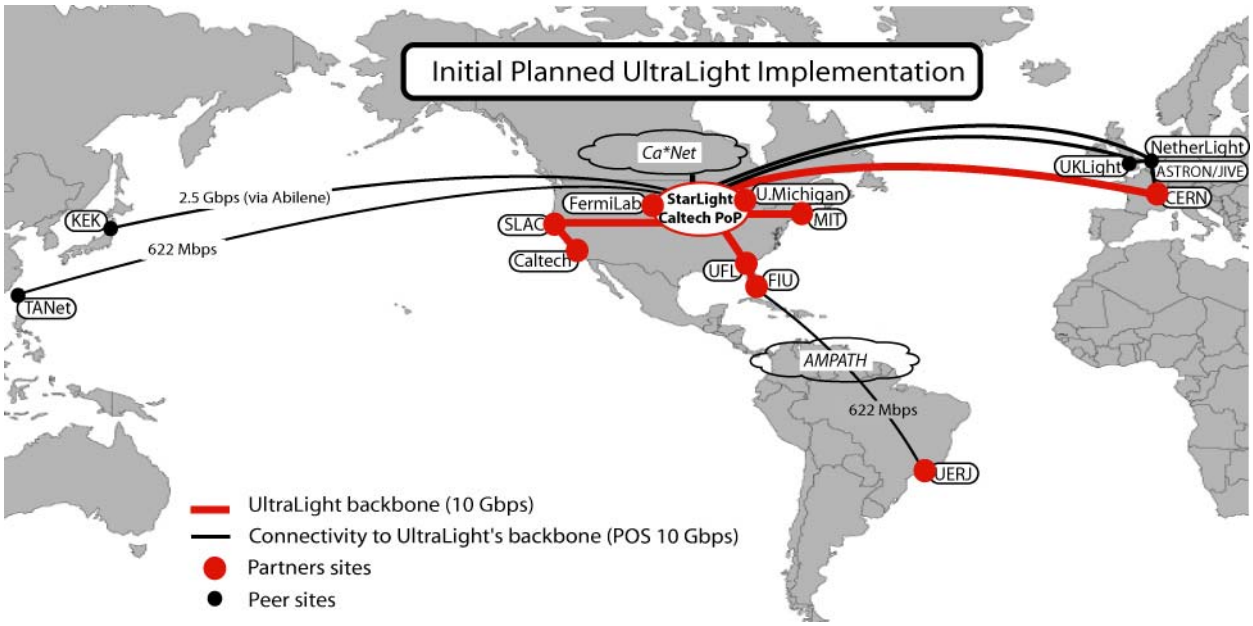
UltraLight: An Ultra-scale Optical Network Laboratory for Next Generation Science

Submitted to the NSF Experimental Infrastructure Network Program

May 8, 2003

Proposal #0335287

Main Partner Sites	Supporting Organizations
California Institute of Technology	DataTAG/CERN
University of Florida	Starlight/Translight
Florida International University	AMPATH
University of Michigan	SURFNet/Netherlight
Haystack Observatory/MIT	HepGridBrazil (UERJ)
Stanford Linear Accelerator Center	CA*NET4
Fermi National Accelerator Laboratory	UKLight
Internet2/UCAID	CENIC
	National Lambda Rail
	Cisco Systems
	Level(3)



Senior Participants

Caltech
- Julian Bunn
- Philippe Galvez
- Iosif Legrand
- Steven Low
- Harvey Newman (PI)
- Sylvain Ravot

University of Florida
- Paul Avery
- Vincent Frouhar
- Alan George
- Chris Griffin
- Jatinder Palta
- David Pokorney
- Sanjay Ranka

Florida International University
- Heidi Alvarez
- Julio Ibarra
- Laird Kramer

UCAID
- Guy Almes
- Steve Corbato

University of Michigan
- Brian Athey
- Thomas Hacker
- Shawn McKee

Haystack/MIT
- David Lapsley
- Alan Whitney

Stanford Linear Accelerator Center
- Les Cottrell

Fermi National Accelerator Laboratory
- Don Petravick
- Victoria White

Table of Contents

B Project Summary

Intellectual Merit: We propose to develop and deploy UltraLight, the first integrated packet switched and circuit switched hybrid experimental research network, to meet the data intensive needs of next generation science, engineering and medical applications, and to drive the development of a new generation of globally Grid-enabled distributed systems. A unique feature of the UltraLight concept is its end-to-end monitored, dynamically provisioned mode of operation, with agent-based services spanning all layers of the system, from the optical cross-connects to the applications. Our experimental network will focus on three "flagship" application areas: (1) particle physics experiments exploring the frontiers of matter and spacetime (LHC), (2) astrophysics projects studying the most distant objects and the early universe (e-VLBI), and (3) medical teams distributing high resolution real-time images. These disciplines collectively present fundamental challenges in distributed terascale data access, processing and analysis that cannot be met by existing network infrastructures.

The UltraLight project will develop a global optical network testbed, and scalable distributed storage and Grid systems, integrating and leveraging the major facilities of LHCNet and DataTAG with transcontinental 10 Gbps wavelengths from National Lambda Rail, in research partnership with Starlight, UCAID, Cisco and Level(3). Additional trans- and intercontinental wavelengths in our partner projects TransLight, Netherlight, UKlight, AMPATH, and CA*Net4 will be used for network experiments on a part-time or scheduled basis. This will give us a core experimental network with principal nodes in California (Caltech and SLAC), Chicago (StarLight and Fermilab), Florida (U. Florida and FlU), Michigan (U. Michigan), Massachusetts (MIT/Haystack), CERN, Amsterdam and the United Kingdom (UC London), with extensions across Abilene, to Taiwan and Japan in Asia, and across AMPATH to South America. The foundation of UltraLight will be a very flexible network fabric, high performance network-resident server clusters, "ultrascale" protocols designed to support stable high performance, MPLS for fair-shared use of networks at multi-Gbps speeds, and intelligent software agents in the MonALISA monitoring framework that will manage network operations and provide an interface between the flagship applications and the ultrascale protocol stack to optimize performance.

UltraLight will operate in a new "hybrid" mode. The principal 10G wavelength interconnecting the core sites will accommodate a diverse traffic mix resulting from the flagship applications, using packet switching. Monitoring agents will handle network flows exceeding the capacity of the core wavelength by dynamically routing the traffic over alternate paths, using the multiply connected topology and the multi-wavelength nature of UltraLight (across the Atlantic and NLR in particular), to transport TByte to PByte datasets while simultaneously delivering multi-Gigabyte images in seconds. A principal aim of UltraLight will be to apply its results on network operation, dynamic provisioning, optimization and management broadly, and to develop standards that will serve many fields.

Broad Impact: Our international optical network testbed will span key core sites in the US and Europe, with a growing set of extensions to Asia and South America. We will deploy "ultrascale" network protocols such as FAST TCP that are capable of stable, fair-shared operation at 1-10 Gbps and eventually higher speeds, and follow a systematic development path addressing issues of protocol fairness and end-to-end performance among disk servers. This will enable us to progressively deploy the results from each development cycle of our R&D across the world's major production networks, including Abilene, National Lambda Rail, and TeraGrid in the US, and GEANT in the EU, as well as major international links serving the scientific communities in South America, Japan, and Korea.

UltraLight will have several long-term achievements: (1) **Enable** a new generation of scientific and medical discoveries by creating a dynamic, agent-driven networking infrastructure supporting distributed data intensive research; (2) **Integrate** our results with emerging data-intensive Grid systems, to drive the next generation of Grid developments, worldwide monitoring systems and new modes of collaborative work; (3) **Collaborate** with groups leading the development of optical switching technologies to understand the relative roles of circuit-switching and packet-switching in support of major scientific applications; and (4) **Apply** our results to a broader set of applications, from new biomedical applications, to high-end real-time visualization of large-scale climatology datasets, bringing the benefits of advanced networks to science and society, while transforming the evolution of the Internet.

Education and Outreach: We will provide students with innovative educational outreach opportunities that utilize the groundbreaking work of the project, and leverage existing programs such as FIU's CHEPREO and CIARA projects to provide recruitment and infrastructure. The UltraLight proposal directly supports E&O activities including development of EIN applications, participation in experiments and deployments of infrastructure, and networking internships at participating institutions, through an allocation of 4% of the overall budget. These opportunities will provide training and inspiration for the next generation of scientists and network engineers.

C Project Description

C.1 UltraLight Project Vision

We propose to develop and deploy UltraLight, the first packet switched and circuit switched hybrid experimental research network, to meet the unprecedented needs of next generation science, engineering and medical applications, and to drive the development of the first of a new generation of globally Grid-enabled data intensive managed distributed systems. Our experimental network will support three "flagship" application drivers: (1) particle physics experiments exploring the frontiers of matter and space-time (LHC), (2) astrophysics projects studying the most distant objects and the early universe (e-VLBI), and (3) medical teams distributing high resolution real-time images. These disciplines collectively pose a broad set of networking requirements and present fundamental challenges in distributed terascale data access, processing and analysis that cannot be met by existing network infrastructures.

UltraLight will establish a global testbed based on a dynamically configured optical network fabric, distributed storage and Grid systems, and a scalable end-to-end monitoring and management system, integrating and leveraging the major facilities of LHCNet [LHC03] and DataTAG [DAT03] with transcontinental 10 Gbps wavelengths from National Lambda Rail [NLR03]. Additional trans- and intercontinental wavelengths in our partner projects TransLight [TRA03], Netherlight [NET03], UKlight [UKL03], AMPATH [AMP01] and CA*Net4 [CAN03] will be used for network experiments on a part-time or scheduled basis, giving us a core experimental network (see Figure 1) with principal nodes in California (Caltech, SLAC), Chicago (StarLight, Fermilab), Florida (U. Florida, Florida International U.), Michigan (U. Michigan), Massachusetts (MIT/Haystack), CERN, Amsterdam and United Kingdom (UC London), partnering closely with advanced production and research networks such as Abilene, TeraGrid and the future GLORIAD (US-Russia-China optical ring) project, as well as the Level(3) MPLS network at multiple locations.

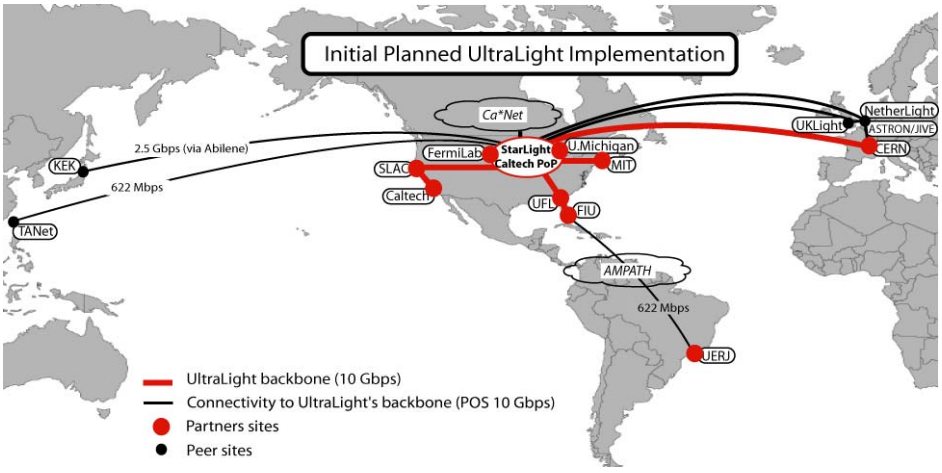


Figure 1: UltraLight Site Diagram

The foundation of UltraLight will be a dynamically configurable network fabric, high performance network-resident server clusters, ultrascale protocols designed to support stable high performance, fair-shared use of network paths with MPLS at multi-Gbps speeds, and intelligent software agents in the MonALISA [MON03a/b] monitoring framework that will manage network operations and provide an interface between the flagship applications and the protocol stack, as well as MPLS, to optimize performance.

UltraLight will operate in a new "hybrid" mode. The majority of network traffic will run over the principal 10G "core" wavelength using packet switching, to accommodate a diverse traffic mix of up to several Gbps from our applications. End-to-end monitoring agents will handle demands for network flows that exceed the capacity available along the main network path dynamically, by (1) using the multiply connected topology and the multi-wavelength nature of the testbed (across the Atlantic and NLR in particular), (2) scheduling additional wavelengths to deliver multi-GByte images in a few seconds, or a Terabyte block in minutes, (3) optimizing the sharing among data streams by tuning the protocol stacks to respond to "requests" from instrumented applications reporting (through agents) the scale and quality of their network requirements, (4) creating guaranteed bandwidth paths across the network fabric using MPLS, and (5) shaping and/or redirecting traffic flow through higher-level agent-based services that optimize system throughput using adaptive algorithms, such as neural networks [SON97, SON01].

The higher level agent services of UltraLight will monitor the overall system state and the evolution of throughput

for the ensemble of data streams. This will allow both the managements of scientific collaborations and network operators to set and adjust the operation of the UltraLight system to match the policies of the collaboration: for the long-term sharing of network resources, and the relative speed (priority) of handling different data streams. By monitoring the clusters and storage systems, and using detailed knowledge of their performance under (CPU, I/O and network) load, in addition to the network performance, realistic real-time estimates of expected performance along any given path will be derived. These estimates will be used to provide decision support, and eventually automated decisions aimed at improving network performance (throughput, response time profiles) according to specified metrics. The performance will be compared to the actual throughput obtained along each path, in order to (1) develop optimal workflow strategies, and (2) trap problems and/or redirect application tasks as needed.

One principal research aim of UltraLight will to make this new mode of managed, policy-driven, optimized end-to-end operation of multi-Gbps global networks widely available and deployable in production networks, to serve the needs of science and society across a broad front.

We will deploy "ultrascale" network protocols such as FAST TCP that are capable of stable, fair-shared operation at 1-10 Gbps and eventually higher speeds, and follow a systematic development path addressing issues of protocol fairness. We will use MPLS and other modes of bandwidth management, along with dynamic adjustments of optical paths and their provisioning, in order to develop the means optimize end-to-end performance among a set of virtualized disk servers, a variety of real-time processes, and other traffic flows. This will give us the necessary scope and the full range of Layer 1, 2, and 3 operational handles in the network, along with an agent hierarchy to integrate the system up to the applications layer, enabling us to develop the first of a new class of global networks.

We will progressively deploy the results of each UltraLight development phase, across the world's major production and research networks, including Abilene, National Lambda Rail, and TeraGrid in the US, GEANT in the EU, as well as major international links serving the scientific communities in South America, Japan, and Korea.

Participants and partners include universities in the U.S. and overseas, as well major DOE funded high energy physics laboratories, all with a track record of leadership in the development of wide area networks and Grid systems. We also have established a full partnership with Cisco's research and networking divisions, along with Leve(3) Communications. Our team's aggregate expertise, experience and leadership in networking, computing and scientific research together offer a unique opportunity, to drive the rapid acceptance and deployment of the products of our proposed networking research on production networks. Relevant activities and experience include (1) innovative networking protocol developments that recently set several world records for trans- and intercontinental data transfer (Caltech, CERN, SLAC); (2) development and deployment of high speed national and international networks such as the US to CERN link (Caltech. CERN), Internet2 (UCAID), AMPATH to South America (FIU), UK-Light (UC London), NetherLight (U. Amsterdam), National Lambda Rail; (3) a long-running program for monitoring networks worldwide (SLAC); (3) development of the agent-based MonALISA monitoring framework (Caltech) and the GEMS fault-tolerant monitoring package (UF); (4) creation, operation and development of the Internet2 network connecting 200+ major U.S. research universities (UCAID); (5) development and operation of major facilities for the operation of major high-energy physics experiments (CERN, FNAL, SLAC) that are among the largest users of national and international networks; (6) development of new scientific and medical applications (UF, Caltech, MIT/Haystack, UM); (7) leadership in the development and deployment of international Grid computing infrastructures (UF, Caltech, FNAL); (8) participation in and creation of major education and outreach programs such as CIARA (FIU), and CHEPREO (FIU, Caltech, UF, UERJ/Brazil); (9) development and production of innovative routers, switches, optical multiplexers, system software and facilities for operating global enterprise networks (Cisco Systems); and (10) development, deployment and operation of global optical networks, MPLS-based network provisioning and management facilities, as well as multi-wavelength dark fiber infrastructures in cooperation with the research and academic communities (Level(3) Communications).

C.2 Applications

Three flagship application areas were chosen for this proposal: high-energy and nuclear physics (HENP), Very Long Baseline Interferometry (VLBI) and Radiation Oncology. All require advanced network capabilities not available today in production networks, and each presents unique and difficult requirements in terms of bandwidth, latency, guaranteed throughput and/or low response time, and burst rate capability – all of which must be met by our experimental network infrastructure. In addition, we will work with several physics and astronomy related Grid projects led by UltraLight participants. These projects are described below.

C.2.1 High Energy and Nuclear Physics

Collisions of particles at increasingly large energies have provided rich and often surprising insights into the fundamental particles and their interactions. New discoveries at extremely small distance scales are expected to have profound and even revolutionary effects on our understanding of the unification of forces, the origin and stability of matter, and structures and symmetries that govern the nature of matter and space-time in our universe.

Experimentation at increasing energy scales, increasing sensitivity and the greater complexity of measurements have necessitated a growth in the scale and cost of the detectors, and a corresponding increase in the size and geographic dispersion of scientific collaborations. The largest collaborations today, such as CMS [CMS03] and ATLAS [ATL03] which are building experiments for CERN's Large Hadron Collider (LHC) [CRN03, LHC03] program, each encompass 2,000 physicists from 150 institutions in more than 30 countries. Current experiments taking data at SLAC and Fermilab [FNL01, DZR03, CDF03] are approximately one quarter of this size.

HENP Computing Challenges: Current and next generation HENP experiments face unprecedented challenges in terms of: (1) the *data-intensiveness* of the work, where the data volume to be processed, distributed and analyzed is now in the multi-Petabyte (10^{15} Bytes) range, rising to the Exabyte (10^{18} Bytes) range within a decade; (2) the *complexity* of the data, particularly at the LHC where rare signals must be extracted from potentially overwhelming backgrounds; and (3) the *global extent* and multi-level organization of the collaborations, leading to the need for international teams in these experiments to collaborate and share data-intensive work in fundamentally new ways.

The IT infrastructures being developed for these experiments are globally distributed, both for technical reasons (e.g., to place computational and data resources near to the demand) and for strategic reasons (e.g., to leverage existing technology investments, and/or to raise local or regional funding by involving local communities). The LHC experiments in particular are developing a *highly distributed, Grid based* data analysis infrastructure to meet these challenges that will rely on networks supporting multiple 10 Gbps links initially, rising later into the terabit range.

Network Characterizations of HENP Applications: International HENP collaborations require ultra-fast networks to link their global data and computing resources in order to support a variety of data-intensive activities.

Bulk data transfers: Many petabytes of raw and simulated data must be moved between CERN where the data is collected to many national laboratories and hundreds of universities. Much of this data will exist as copies, and the integrity and identity of the copies must be tracked and stored in metadata databases. The data can be moved by bulk transfer over relatively long periods and can be interrupted by higher-priority network traffic.

Short-term data transfers: Distributed "analysis" teams will rapidly process multi-terabyte sized data sets that generally have to be moved to available clusters at any of the collaboration's computational centers worldwide. As an example, processing 10 terabytes in one hour would require ~20 Gbps of network bandwidth just to transfer the data. A large LHC collaboration could have hundreds of such analysis applications ongoing at any one time.

Collaborative interactions: HENP physicists will increasingly exploit advanced collaborative tools and shared visualization applications to display, manipulate and analyze experimental and simulated results. Some of these applications, especially those that are real-time and interactive, are sensitive to network jitter and loss.

Network Infrastructure Requirements for HENP

1. Ultra-high end-to-end data rates: Very high bandwidth is needed both for long-term bulk data transfers and short-term (few minutes) transactions. The required bandwidth follows the accumulated size of the data collection, which will increase at faster than linear rates. All aspects of the transfer process, including the OS kernels, disk systems, net interface firmware and database software, must be optimized for extremely high throughput.
2. High accuracy: Guarantees are needed to ensure the complete integrity of the data transfer process, even in the face of repeated interruptions by higher priority network traffic.
3. Data copies: If multiple copies of data are requested, the copies can be created anywhere in the network fabric, subject to the requirement that the experiment Grid infrastructure be able to track the location of all the copies.
4. Monitoring: A pervasive monitoring system must track the hundreds of current and pending data transfers that must be scheduled throughout the global Data Grid infrastructure.

C.2.2 Very Long Baseline Interferometry

Very-Long-Baseline Interferometry (VLBI) has been used by radio astronomers for more than 30 years as one of the

most powerful techniques for studying objects in the universe at ultra-high resolutions (far better than optical telescopes), and for measuring earth motions with exquisite precision [WHI03]. The precisely determined locations of distant quasars reveals a wealth of information both on the earth's surface (i.e. tectonic plates) and the internal motions of the Earth system, including interactions with the atmosphere and oceans.

VLBI combines simultaneously acquired data from a global array of up to ~20 radio telescopes to create a single coherent instrument. Traditionally, VLBI data are collected at data rates up to ~1 Gbps of incompressible data on magnetic tapes or disks that are shipped to a central site for correlation processing. This laborious and expensive data-collection and transport process now can be replaced by modern global multi-Gbps networks, enabling real-time data acquisition and analysis at increased speed, resulting in possible new scientific discoveries.

Advanced networks and e-VLBI: The transmission of VLBI data via high-speed network is dubbed 'e-VLBI', the further development of which is one of the main thrusts of this proposal. This mode of data transfer, which can be either real-time or quasi-real-time through buffering, offers several advantages over traditional VLBI:

1. *Higher sensitivity:* The potential to extend e-VLBI to multi-Gbps data rates will allow an increase in the sensitivity of observations. VLBI resolution generally improves as the square root of the data rate.
2. *Faster turnaround*: Traditional VLBI processing takes days or weeks because of tape shipping times. This makes it almost impossible to use 'fast-response' observations of important transient events such as extragalactic supernova or gamma-ray-burst events.
3. *Lower costs*: E-VLBI will eliminate the need for large pools of expensive tapes or disks while at the same time allowing full automation of observations, all towards the goal of lowering cost.
4. *Quick diagnostics and tests*: Some aspects of VLBI equipment are very difficult to test and diagnose without actually taking data with another station and processing it, costing valuable time. E-VLBI can maximize the prospects of fully successful observations on expensive and difficult-to-schedule antenna facilities.
5. *New correlation methods*: E-VLBI data from all antennas are brought to central location for correlation using special-purpose hardware. The aggregate bandwidth to the correlator depends on the number of stations, presenting network scalability problems for large numbers of sites. We will explore 'distributed correlation', where some data are transported to multiple correlators utilizing high-performance computer clusters at universities.

Both faster turnaround and higher sensitivity will open doors to new science while lower costs, easy diagnostics and better correlations lead to more science impact per dollar.

E-VLBI Network Infrastructure Requirements

1. Large bandwidth: Very high rates (multiple Gbps) are fundamental since the angular resolution scales with the inverse square root of the accumulated data.
2. Continuous data stream: E-VLBI requires a continuous real-time data stream from antenna to correlator; limited buffering (a few seconds worth) is available at both the antenna and the correlator. Disk buffering at one or both ends of a baseline can be used as long as total data volume does not exceed the available disk capacity.
3. Transient data stream: Once raw e-VLBI data has been correlated, the data are discarded. Therefore, the raw ultra-high-speed data need only exist for a short amount of time. The correlation results are typically compressed by factor of 10^7 or more compared to the original data flows.
4. Low sensitivity to small data losses: Raw e-VLBI data is fundamentally just noise from the radio objects (it is only the *correlation* with data from other antennas that is important), thus the loss of up to a few percent of the data normally results in only a few percent decrease in signal-to-noise ratio. This characteristic allows e-VLBI data to be transmitted on a less-than-best-effort basis in many cases, so as not to dominate a network channel.

C.2.3 Radiation Oncology

Radiation Oncology is involved with the treatment of cancer. A critical aspect of radiation oncology is managing patient imaging and radiation treatment data and making it available to clinicians and experts spread over geographically distributed regions for input and review. The Resource Center for Emerging Technologies (RCET) system at UF provides the storage and compute resources for the cooperative groups engaged in advanced technology radiation therapy trials that generate voluminous imaging and radiation treatment data [PVD03]. It has four main components: 1) a storage system that provides mechanisms for storing and accessing multimodal data such as imaging data, radiation therapy planning and delivery data; 2) a collaboration system that supports concurrent authoring

and version control and ensuring that any changes are available for rapid review; 3) a network transmission and caching system that encrypts data to maintain patient confidentiality and caches frequently accessed data locally; 4) an AVS based highly functional image processing and visualization client that allows a remote user to perform a number of compute and data intensive operations locally. Additional features such as automatic image attribute analysis and data mining are planned. This repository is mirrored for fault tolerance and high availability.

The amount of data per patient per visit (treatment) can vary from 100 to 500 megabytes. This is because hundreds of slices of CT, MR, PET images, and treatment planning data have to be stored. Correct interpretation and clinical diagnosis require that the imaging data be preserved in its original form. This requires the use of lossless compression techniques during transmission. Additionally, this data has to be encrypted to manage privacy issues and HIPPA guidelines. A clinical trial group such as the Radiation Therapy Oncology Group (RTOG) can handle 2000 patients per year, which translates to over 1 terabyte of data. Other clinical trial groups have similar needs. The advantages of the RCET system over traditional physical film based methods of patient data archival are the following:

1. *Comprehensive Protocol Specific Patient Data*: The system enables clinical trial groups to implement comprehensive clinical protocols by storing multi-modal electronic imaging and multiphase planning data from diverse geographic areas. The information stored can be mined for correlations that impact patient care.
2. *Lower costs*: The system eliminates the need for large pools of expensive film repositories or diskettes. The system is largely automated to minimize the staff required.
3. *Improved Patient Care*: The system allows for a substantially faster turnaround time, and better collaboration among individuals with synergistic expertise in remote and distant areas. It allows presentation of the information in identical format to all the experts; previous studies have shown that this positively impacts the diagnosis.

Impact of Advanced Networks on Radiation Oncology:

1. *Improved Interactivity*: The interactively of the RCET system is severely constrained by the size of the data involved (100 – 500 MB per patient may require hours of transmission time, currently). Ultra-high speed network bandwidths will enable interactive team collaboration that in turn will lead to improved patient care.
2. *Improved Availability and Reliability*: Use of distributed centers reduces the number of concurrent users per center and improves reliability, but substantial bandwidth is required for maintaining consistent replicas of the data.
3. *Lower cost of software development*: Thick clients allow for local data processing to reduce network bandwidth; supporting these clients on multiple systems has huge software development costs. Higher bandwidth allows the use of highly-functional thin clients by moving most tasks to the server and reducing this cost.

TeleRadiation Therapy: We will develop RCET for conducting interactive Teleradiation therapy using Ultralight (the work plan is presented in section C5). The Michigan Center for Biological Information at the University of Michigan [MCBI03] will collaborate with RCET to test this functionality. MCBI will engage the appropriate members of the UM Medical School Department of Radiation Oncology in this effort. The following characteristics of teleradiation therapy present unique opportunities, and challenges in the development of networks for this purpose:

1. Bursty Transfer: Interactivity will require transferring patient data sets (100 to 500MB) in ~one second, and preferably less. For supporting thin clients, the image processing time needs to be included in this time interval.
2. Bulk Transfer: Managing multiple replicated data centers is necessary for higher availability and fault tolerance. This however requires intermittent server reconciliation (synchronization). This reconciliation may require the transfer of several gigabytes of data. However, this transfer can be done in bulk potentially at off-peak hours.
3. High sensitivity to small data losses and patient confidentiality: The nature of the application requires that the data be transferred accurately and securely.
4. Ordering of data within a request is not required. Unlike applications like streaming, this application does not require that data transferred within each request maintain any ordering.

C.2.4 Grid Applications

Senior UltraLight participants also direct major data intensive Grid projects in the U.S. and Europe that will provide additional application opportunities for conducting high-speed data movement and distributed system tests. These experiments will generate and store hundreds of Petabytes of data to be used by thousands of researchers worldwide:

- **Particle Physics Data Grid (DOE)**: PPDG [PPD03] is developing advanced monitoring systems (including MonALISA), high-speed database replication, distributed data analysis and Grid authentication tools for HENP.

- **GriPhyN and iVDGL (NSF)**: GriPhyN [GRI03] is researching and developing a "Petascale" Grid infrastructure for four frontier experiments in high-energy physics [CMS03, ATL03], gravitational wave research and full-sky digital astronomy. iVDGL [IVD03] is deploying a global Grid laboratory where grid tools and distributed applications can be tested at a large scale.
- **Datagrid and DataTAG (EU)**: DataGrid is developing and deploying Grid tools for the LHC experiments, earth satellite observations and bioinformatics. DataTAG [DAT03] is creating a large-scale intercontinental Grid testbed that focuses on advanced networking issues and interoperability between these intercontinental Grid domains, extending the capabilities of each and enhancing worldwide Grid development.

These projects already coordinate closely with one another and have numerous joint projects, on account of common LHC experiments that they serve. We expect to be able to conduct joint ultra high-speed data movement tests with resources provided by these Grid projects while engaging researchers from their constituent scientific communities.

C.3 Experimental Network Services

We propose to develop and deploy the first 10+ Gbps packet switched and circuit switched experimental network and develop a set of services that will support a new generation of globally Grid-enabled data intensive systems. These systems are required for successful enablement of the target applications that require network performance and functionalities far beyond what can be provided by Internet 2 or other production networks today.

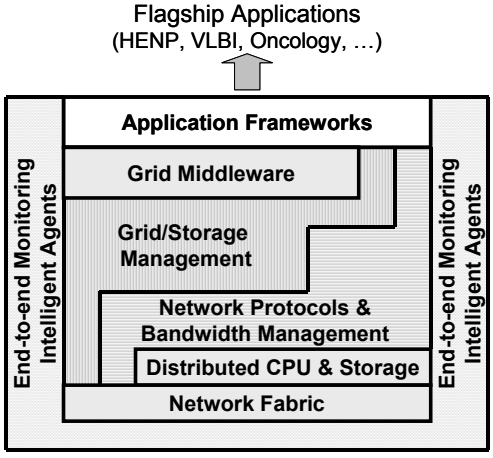


Figure 2: UltraLight Architecture
Hatched areas developed in proposal

Novel methods for transport, storage and monitoring are required to achieve the above objectives in a scalable and robust manner. We will develop these methods with a solid theoretical foundation and validate them through extensive simulations. Mathematical analysis and simulation, while critical in algorithmic development, inevitably ignores many important implementation issues that affect both the performance and the deployment of these methods. The UltraLight testbed, through its close contact with flagship applications, state-of-the-art infrastructure and global reach, will provide a unique and rich environment in which to test, develop and successfully deploy them on a wide scale. The real traffic and usage patterns of the targeted applications will provide an ideal launching platform from which the proposed experimental systems can be deployed in the real world.

The foundation of UltraLight will be a high performance dynamically configurable network fabric, network-resident storage server clusters, ultrascale protocols to support stable fair-shared use of network paths at speeds up to 10 Gbps, real-time, fault tolerant and scalable monitoring systems based on MonALISA [MON03a/b] and GEMS [GEM03a/b] and intelligent software agents that will manage network operations, and provide an interface between the flagship applications and the ultrascale protocol stack to optimize performance (see Figure 2). This infrastructure will leverage existing grid tools, web services, storage software and network protocols. In the following, we briefly describe the key services that we propose to develop, and their relationship to extant software and hardware.

C.3.1 Network Protocols and Bandwidth Management

For all the flagship applications, rapid and reliable data transport, at speeds of 1 to 10+ Gbps is a key enabler. In this project, we will explore three different approaches to bandwidth management - packet switching, MPLS and circuit switching - to satisfy the complementary "QoS" requirements of these applications, while simultaneously exploiting the hybrid nature of the proposed network fabric. The close coupling of applications and infrastructure in UltraLight will allow us to derive realistic requirements that drive protocol development. UltraLight will provide a rich mix of complementary network services, supported by the following protocols:

- Packet switching makes the most efficient use of network bandwidth, requires no network support, but provides the least QoS guarantee. It is ideally suited for serving applications that are bandwidth intensive but delay tolerant, such as HEP applications.
- Circuit switching requires intelligence in the network for connection set up and routing, provides the most reliable QoS guarantee, may be the least bandwidth-efficient if the application data rate varies rapidly. It can be used to support applications with very stringent delay requirements, such as real-time TeleRadiation oncology.

- MPLS [MPA01] is a hybrid between packet switching and circuit switching, both in terms of flexibility and efficiency. MPLS features, such as explicit path selection, bandwidth reservation [MRT01] and differentiated services [MDS02] allow for intelligent control of network resources by the end applications.

We will evaluate the efficacy of these protocols for the three target applications for a variety of workloads – from dedicated bandwidths for each application to bandwidth being shared by all the applications. Each application can have multiple concurrent users and will require variable network resources as described in section C2. A key goal will be to re-examine the current control protocols and resource management schemes and redesign those that do not scale to the ultrascale regime. To this end, we will investigate and develop the following to support our flagship applications:

1. The performance under a variety of workloads of the new generation of transport protocols that target scalable high performance. These include FAST TCP [OFC99, FST03a, UVE02, SVG03, FST03b, TAQ03], GridDT [GridDT], HS TCP [HST03], Scalable TCP [STC02], XCP [XCP02], Tsunami [TSU03], SABUL [SAB03], that attempt to make effective use of high capacity network resources.
2. Integration of the FAST TCP prototype with the proposed hybrid network, to support our applications with optimal combination of throughput and response time, as appropriate. FAST TCP, developed by us and based on a solid theoretical foundation and extensive simulations, aims at high utilization, fairness, stability and robustness in a dynamic sharing environment with arbitrary capacity, delay, routing and load; and we will evaluate it in all these aspects in a variety of working conditions. We will also develop a QoS-enhanced version of FAST that will support different application requirements, by expanding and optimizing the interaction between the transport and application layers. We will interface it with other bandwidth management schemes, including circuit switching and MPLS, and other services such as network monitoring, storage, and admission control.
3. Combining the techniques of FAST TCP and Combined Parallel TCP [HaNo2002] to create a new hybrid protocol that can effectively use bandwidth on underutilized networks without unfairly appropriating bandwidth from competing streams on busy networks.

New protocol developments will also be undertaken to specifically take advantage of the traffic management and QoS features provided by MPLS. The development will also look at the use of application level real-time routing of streams (and redirection) through use of these features. These new protocols will be developed and analyzed using network simulation tools, and will then be implemented on network hosts to support the applications. Finally, they will be integrated across the hybrid network and evaluated through demonstrations and field tests.

An important issue in protocol development for wider deployment is the interface. Our goal will be to develop protocols with simple interfaces (thereby easing application development) and high functionality (thereby providing high performance for a variety of applications). In many cases these are contradictory goals, but striking the right balance will be a key contribution of the proposed research.

C.3.2 Storage and Application Services

While there has been significant recent progress in demonstrating 10+ Gbps transfers across wide-area networks, this has been achieved by using "memory-to-memory" transfers. However, most applications are concerned with moving information from more persistent storage media to remote storage via the network. Future networks need to anticipate this mode of operation and provide support for enabling high bandwidth data transfers.

There are numerous issues to contend with when worrying about storage-to-storage transfers across the proposed networks. Bottlenecks can exist in many places and must be identified and removed. Novel storage architectures and software are required to support high bandwidth data transfers. These issues are significantly different from managing storage on heterogeneous resources [RWM02, BVRS02]. The research conducted on the Internet Backplane Protocol [BBFM01] has addressed some of the issues for storage and access of data for high bandwidth transfers on wide area networks. However, much of the work is focused on replication for streaming applications. To serve the needs of the range of applications described in this proposal, we propose to investigate the following:

- Design and build a high performance disk storage subsystem from commodity parts architected to fully drive a 10 Gbps network from disks, not just from memory. As part of UM-VHP development, we found that for many applications, users are not interested in retrieving all accessible data in one atomic operation. Retrieving the entire dataset may require significant transmission time and, in some cases, may preclude the use of any applications that have real-time constraints (such as Radiation Oncology). Developing a storage strategy that addresses these

access constraints is critical for problems that require low latency access to subsets of large datasets. We will leverage the use of existing distributed storage systems such as SRB [RWM02], NEST [BVRS02], IBP, and parallel I/O file systems [NFSv4] to build a distributed storage system between the Ultralight partners.

- An application library that interfaces the storage layer with the network layer consisting of protocols such as Combined Parallel TCP [HaNo2002], FAST TCP and MPLS to deliver effective and high performance end-to-end transfers. The co-development of the application library and the network protocols will deliver several useful artifacts. First, the application library will allow us to modify the underlying TCP protocol without requiring re-coding of scientific applications to make use of the modifications. Second, we will be able to assess the performance of different TCP methods *in situ* with real applications, with the goal of optimizing end-to-end application performance. Third, we will observe and collect real application access patterns to create accurate network simulation modules for ns2 (this will also be valuable for Grid development). These ns2 simulations will be used to investigate application behavior and interactions with protocols, and assess proposed modifications before implementation. Finally, using the network monitoring approaches (described in the next subsection) along with Web100 and Net100, we will create an integrated end-to-end framework for designing, assessing, and deploying network protocols and physical infrastructure components to deliver high throughput and low latency for scientific applications, and to provide a unified view of component and application performance. We will also provide a set of application examples and templates that demonstrate successful network-aware programming techniques.

C.3.3 Network Monitoring and Simulation

UltraLight requires pervasive end-to-end global resource monitoring that is tailored to the application. The monitor must span, interact with, and support all levels of UltraLight, from the optical fabric to the applications. Global monitoring services must simultaneously provide: (1) real-time monitoring and immediate response to "problem situations" as they occur; and (2) long-term task monitoring in different "work classes," modifying system behaviors (e.g. priority profiles) to enforce policy based resource constraints while maintaining acceptable performance. Application diversity in terms of flow size, duration, "burstiness", accuracy, etc. is a significant obstacle to effective monitoring, since the monitor must react quickly and be lightweight.

We will extend MonALISA (developed at Caltech [MON03a/b]) and GEMS (developed at Florida [GEM03a/b]) to address these issues. MonALISA is based on a highly efficient distributed services architecture that provides a global framework for monitoring and coordination [DSA01, DIG03]. GEMS is a quorum based system for sensor measurement and information dissemination for large sites [GOS99, GOS01a, GOS03] [ANS01]. Our goal is to leverage the strongest features of these tools to build an advanced end-to-end resource monitoring service that is hierarchical, distributed, robust, scalable, and responsive with minimal impact on network performance. We propose to do the following:

- Leverage MonALISA and GEMS to create a new multilevel, end-to-end service for global, regional and local monitoring. A key challenge will be the non-intrusive integration of mechanisms for end-to-end measurements. Sensor data must be efficiently collected and disseminated from host-based and fabric-based sensors.
- Develop sensors for global and local resources congruent with activities on intelligent network agents and storage services. A key challenge is the near-real-time support of global optimization services via intelligent agents that are loosely coupled with the monitoring infrastructure. We will define monitoring hierarchy performance profiles and optimize key attributes (e.g. sensor sampling and quorum patterns) to maximize responsiveness where needed (e.g. short-term flows in HENP, bursty flows for oncology) and reduce overheads when possible (e.g. continuous flows in e-VLBI), while ensuring scalability and non-intrusiveness of sensors. Another challenge will be to develop these sensors for measurements at the high networking speeds expected in UltraLight, where most existing active end-to-end network performance sensors perform poorly. Examples of sensors we expect to extend include active measurement packet-pair techniques such as the SLAC Available Bandwidth Estimator [ABWE].
- We will develop prediction tools that provide both short-term (several minutes into the future) and long-term (up to days in advance) predictions. We will explore how to integrate the IEPM measurements into our infrastructure to provide the strongest features of each service. We will also develop tools that use the predictions to provide anomaly alerts and to automatically make extra measurements to assist in troubleshooting.
- We will develop measurements and prediction APIs based on emerging Global Grid Forum standards for naming hierarchies [NMWG] and schemas to ensure portability between the wide varieties of grid network measurements systems. We will integrate the monitoring system with network/storage services and intelligent agents, and an ex-

tensive series of performance tests will be conducted to analyze and optimize the various sensor, monitor, and agent features and settings for the flagship applications taken alone and together.

- We will develop a monitoring infrastructure for making regularly scheduled active end-to-end network and application performance measurements. This will be based on an extended version of the SLAC developed IEPM-BW [IEPM] toolkit that is deployed at over 50 high-performance HENP/Grid or network sites. This will provide an archival database of regular network and application (e.g. bbftp [BBFTP], GridFTP [GRIDFTP], tsunami [TSUNAMI]) performance measurements between participating sites.

Simulation is a critical tool in this project where fundamental issues related to network, node, storage, and system design will be difficult and costly to address directly on the testbed. UltraLight researchers will leverage tools and models such as MONARC [MON99] from Caltech/CERN, the Caltech FAST protocol model [FST03a, SVG03, FST03b], and optical layer and other related models [GOS99, SIM03] at Florida. Key simulation experiments involving adjacent layers in the performance hierarchy will help researchers quickly identify bottlenecks, compare algorithms and architectures, and assess tradeoffs and the impact of low-level options on the design of experimental network services and the applications they serve. With simulation, if unexpected performance limits emerge in the testbed, researchers can fall back on simulation and inject data into simulation experiments to investigate fundamental issues and determine the best methods by which to address them in achieving optimal end-to-end performance.

C.3.4 Intelligent Network Agents

The network protocols and storage services will provide a flexible infrastructure that can adapt to user and application requirements. The monitoring system will provide real-time fine-grained, and precise data about the performance and usage of the underlying system. We propose to develop a suite of intelligent agents that can use this information for decision-making and optimization of the underlying resources. These intelligent agents will use real-time data provided by the underlying monitoring system and symbiotically work with each other (collectively) to adapt the network to extant grid, web, network and storage services. For example, appropriate action may be taken if a task is not getting what it needs (as communicated to the managing agent from the application), or what it "deserves" (e.g. according to a relative priority or resource-sharing metric). The actions taken could include diverting the end-to-end path, scheduling another wavelength along part or all of the path or stopping/re-queuing another task that is taking too long, and has lower priority.

The following are the agents that we propose to develop:

- Bandwidth Scheduling and Admission Control for a multi-user, multi-application differential QoS environment – We will develop intelligent agents that meet the following requirements. First, dynamic and decentralized bandwidth scheduling algorithms are needed, at burst, flow, and lambda levels that optimize bandwidth reservations with respect to application requirements and network contention. Second, to provide dedicated bandwidth at any of these levels, admission control is necessary to prevent over-subscription. Third, the physical resources will be shared by both circuit-switched and packet-switched traffic; optimal and adaptive resource allocation between these two types of traffic and the interaction between their respective QoS mechanisms require careful design.
- Smart memory and storage buffer management - Supporting a multi-user, multi-application environment will require careful allocation of memory and storage for effectively utilizing the large bandwidth. Novel mechanisms will be developed for managing application memory and disk buffers. Per-user or application-based dedicated buffering schemes have the advantage of predictability but are generally very conservative from a global perspective. Global buffering schemes have the disadvantage of starving an application, which can have a dramatic impact on the quality of service that can be achieved. We will develop (algorithmic and/or heuristic self-learning) methods that provide the right tradeoffs and can effectively utilize user and application profiles and tune them based on real-world traffic and usage patterns.
- User and application profiling – Users and applications are interested in accessing particular data sets, monitoring views etc. Providing a global service that facilitates the discovery of all relevant data of interest using data mining techniques in an accurate and timely manner is an important need. We will use profiles as a universal representation of interest and develop techniques to learn profiles.
- Intelligent caching, pre-fetching and replication based on user and application profiles – For many applications, users access the same data many times or in regular patterns systematically (e.g. a summary file may be updated when new data is added every night). The data access profiles (user-defined or system-learned) can be used to cache, pre-fetch and replicate data to provide better performance to users and applications while balancing the

load on the underlying system (e.g. by using the network bandwidth during off-peak hours). Latency-Recency profiles have been successfully used for reducing bandwidth requirements [BR02].

- Automatic tuning of underlying system parameters – The performance of most of the services can be modeled by a few key parameters. The underlying monitoring systems will collect data for their performance under different system conditions (e.g. network and storage load) and underlying parameter settings. This information will be effectively utilized to develop self-tuning services that work effectively under many system conditions.

C.4 Network Development and Deployment Plan

UltraLight will build a cost-effective and scalable network infrastructure providing a community-driven packet-switched and circuit-switched, end-to-end managed, global optical network testbed, offering a variety of network services to support a large and diverse traffic mix of applications. Our network paradigm is to "*Packet-switch where we can, Circuit-switch where we should, Redirect where we must"*, maximizing the efficiency of the network infrastructure and the heterogeneous traffic flows. We will interconnect *Partner Sites* in California (Caltech and SLAC), Florida (UFL and FIU), Michigan (UM), Massachusetts (MIT/Haystack), CERN, with the *Peer Sites'* project resources of Starlight [STL01], Fermilab [FNL01], AMPATH [AMP01], NetherLight [NET03], UKLight [UKL03] and CA*Net4 [CAN03], and the major facilities of LHCNet [LHC03] and DataTAG [DAT03], to build a community- and application-driven intercontinental testbed to conduct experiments on a part-time and/or scheduled basis.

C.4.1 Network Architecture – the UltraLight Fabric

UltraLight Fabric: UltraLight will provide an optical-switching fabric for Partner Sites to interconnect to each other or to Peer Sites' resources in a flexible and optimized fashion, to support application-driven requirements (see Figure 1). Partner Sites already, or soon will, connect to Peer Sites' project resources depending on their ongoing research and application development. We will thus build UltraLight at an incremental cost to large existing investments. Peer Sites will provide optical switching and network services at Layers 1, 2 and 3. They can also house application elements, such as distributed storage servers, Grid systems, application servers, etc.

Peer sites are the StarLight, Caltech/CACR and TeraGrid, NetherLight, SLAC, AMPATH, FNAL, CERN, TransLight, UKLight and CA*Net4. Several Peer Sites will primarily provide optical switching services – these are designated *Optical Exchange Points* (OXPs). The term Optical Exchange Point is used in a generic fashion to describe sites, typically located in carrier neutral facilities, that house *Optical Cross-Connects* (OXCs) that will provide ingress and egress ports for the provisioning of *lightpaths* [BSA01,LAA01] (bi-directional end-to-end connections with effective guaranteed bandwidth). OXP-type Peer Sites are the NLR Los Angeles and Sunnyvale PoPs in California, the NLR Jacksonville PoP in Florida, as well as the StarLight in Chicago.

The core of the UltraLight fabric will consist of interconnected OXCs that support end-to-end networking of lightpaths. OXCs will be partitioned, assigning groups of ports to project Partners and Peers to configure and administer based on application requirements, giving control of the optical fabric to the users. Optical sub-networks, linking OXCs, will be created for specific application communities and allow direct peering between institutions and researchers, for an application domain, such as the HENP community. The optical sub-networks can be partitioned into different administrative domains, where intelligent network agents (described in Section 3) can control the various functions of the partition. As a result, the partitioned optical sub-networks allow the user and the application to drive the network topology (to date the network topology drives the application) [BSA01, RAJ02].

Underlying Technology: UltraLight has DWDM wavelengths interconnecting Partner Sites, Peer Sites and OXPs, providing scalable capacity and flexible provisioning of lightpaths. Partner and Peer networks will attach to the optical core switching fabric over switched lightpaths (Figure 2 in the Facilities section). As an experimental infrastructure, various transport technologies, including 10GbE LAN-PHY (where available), SONET/SDH and MPLS will be tested to measure their effectiveness under varying traffic mixes. Effective transport should optimize the cost of data multiplexing, as well as data switching over a wide range of traffic volumes [BAN01]. Most international circuits are using SDH as a transport technology. A goal of this testbed is to test the 10GbE WAN-PHY protocol on international circuits, effectively extending Ethernet to all endpoints. This is an ambitious goal given that there will be a mix of vendor equipment at different sites, increasing the likelihood of interoperability issues. The figure in the facilities section shows the initial physical connectivity of the Partner Sites to each Peer Site through StarLight.

In this configuration, a Layer2 (L2) fabric will be established for connectivity between Partner and Peer sites, as in a distributed exchange point. Using 802.1q VLANs, end-to-end connectivity can be provisioned for end-to-end traffic

flows, as well as to establish peering relationships between each of the Partner and Peer sites. MPLS Label Switched Paths (LSPs) will involve MPLS-capable routers in the core (an MPLS architecture diagram and description are in the Facilities and Other Resources section). Since it cannot be assumed that MPLS will work at all layers, MPLS tests might be limited to specific UltraLight sub-networks. The availability of a L2 fabric will remove the dependencies on any intermediate nodes, and will ease compatibility issues. Initially, Partner sites will peer with each other using the BGP inter-domain routing protocol; however, this configuration might not be optimal for an application driven network (e.g., BGP uses static link costs and requires manual policy configuration). Virtual routers (a mechanism to create overlay networks on a single infrastructure) will be investigated as a way to shift control from the network operators to the intelligent network agents. Authentication, Authorization and Accounting (AAA) mechanisms will be implemented to validate intelligent network agent requests, such as dynamic provisioning of QoS lightpaths through multiple administrative domains [LAA02]. There will be a unique opportunity to collaborate with leading researchers at the University of Amsterdam, SURFnet, CERN, GGF, IRTF and other groups.

C.4.2 UltraLight Infrastructure Development Plan

The underlying technology will support transitioning and integration between packet-switching and circuit-switching modes. Packet-switched services will be offered using Ethernet and MPLS; circuit-switched services will be offered using MPLS and IP over the optical transport network. The switching function in an OXC is controlled by appropriately configuring the cross-connect fabric. Cross-connect tables are normally configured manually to switch lightpaths between ingress and egress ports, as well as to partition the table into different application domains. Automated establishment of lightpaths involves setting up the cross-connect table entries in appropriate OXCs in a coordinated manner, such that the desired physical path is realized. Standards-based signaling and routing protocols will be used to test end-to-end provisioning and restoration of lightpaths across multiple administrative domains. The initial testbed will be built preserving the existing administrative domains of Partner and Peer sites. However, creating "optical level" application communities will involve partitioning OXCs, modifying administrative domains to overlay defined application partitions and provisioning an IP transport service across the optical fabric to logically interconnect application communities. Thus integration and management of Partner and Peer site networks to create virtual networks of application driven communities will be an essential component of the infrastructure development plan.

We will explore the use of GMPLS [GMP03] in some areas to allow the integrated and flexible provisioning across network devices capable of time, wavelength, spatial and packet switching. The use of GMPLS will demonstrate the next generation of integrated network management that will allow enterprises and service providers alike to efficiently manage the various layers of their network [GMP01]. It will also allow us to explore enhanced routing algorithms such as optical BGP, OSPF-TE and IS-IS TE. Combined with the application driven approach and the intelligent monitoring and management system, GMPLS will provide a powerful hook into the UltraLight network fabric that will enable optimal use of network resources with minimal effort on the part of the network operators.

Agents: Using signaling protocols at the physical and network layers, intelligent network agents, working with the signaling and control planes, will transition between packet- and circuit-switched modes. Signaling agents will also do topology discovery and traffic engineering of lightpaths across multiple administrative domains. There is a need to develop mechanisms to establish and operate virtual networks, with predictable services, that are built over multiple administrative domains. Partner and Peer site networks, each with IP routers and switches running intra-domain routing protocols, such as OSPF or IS-IS, and BGP for inter-domain communication, form multiple administrative domains that will need to be integrated to achieve a global, scaleable, application-driven optical network testbed. Signaling and control plane agents provide mechanisms to integrate and manage multiple administrative domains.

Network Deployment Phase 1: Create Network Infrastructure

- Procure, install optical & electronic switches, storage servers to build UltraLight physical infrastructure
- Provision optical transport services from each Partner site to the UltraLight Optical Fabric at StarLight. (Timeframe is subject to availability of NLR and the leveraging of funds to provision optical transport services)
- Build Layer2 Ethernet infrastructure over 10GbE LAN-PHY-10 GbE LAN-PHY, or SONET/SDH transport,
- Build MPLS infrastructure between MPLS-capable sites. Test the set up and tear down of LSPs.
- Establish connectivity and peering relationships to all Partner and Peer sites across the optical infrastructure
- Validate the testbed by testing the set up and tear down of lightpaths between end points. Test transport services
- Integrate network management and monitoring systems with the optical network

- Partition OXCs and configure the network to assign dedicated bandwidth for each application

Network Deployment Phase 2: Research and Testing

- Solicit descriptions of testbeds from UltraLight developers of protocols and intelligent network agents
- Provision additional wavelengths between Partner Sites and Peer Sites (leveraging other projects)
- Using NLR, provision testbeds of additional waves when traffic exceeds 10G. Conduct demos at key national and international conferences.
- Develop and implement processes by which flows transition between packet- and circuit-switched modes
- Test signaling and routing protocols for end-to-end provisioning and restoration of lightpaths across multiple OXCs and multiple administrative domains
- Using IPv4, IPv6 and Ultrascale network protocols, test IP transport for reach, establishment of lightpaths between end points and for persistence
- Test signaling & control plane agents to integrate and manage multiple administrative domains
- Test MPLS' optical bandwidth management and real-time provisioning of optical channels
- Closely work with equipment manufacturers to test the optical infrastructure to ensure that it supports a transport of 10G over a single channel in the core, then 40G and 80G equipment, as it becomes available
- Develop test scenarios to allow applications to effectively manage network infrastructure from end-to-end
- Test the dynamic routing of traffic over multiple lightpaths; Test end-to-end performance of traffic flows across integrated Partner and Peer administrative domains; Test efficiency & effectiveness in the usage of shared 10G wavelength and scheduling of multi-Gbps lightpaths
- Using OOO (e.g., existing Calient) switches at StarLight and NetherLight, conduct lightpath flow tests
- Test integration with Grid systems (both production and development Grids)
- Start the transition to production use of many of the tools, subsystems, and services that were developed

Network Deployment Phase 3: Moving Network Services and Tools to Production

- Study feasibility of deployment of GMPLS and, based on existing standards (OIF-UNI), define a "User to Network Interface (UNI)" to allow a user to provision a light path through the optical transport network
- Develop and test policies for dynamic provisioning of optical paths between routers, including access control, accounting and security issues
- Develop a plan for the transition of experimental network services and tools to production services

C.5 Program of Work

Our proposed research of key EIN Services was described in Section C3. Here we briefly describe the key milestones that we will meet and the program of work to achieve our objectives. The work will be conducted in three phases that will last 18 months, 18 months and 24 months respectively.

C.5.1 Phase I: Implementation of network, equipment and initial services (18 months)

1. Network Infrastructure Deployment of networking and storage infrastructure (Described in Section C.4).
2. Development of a small subset of important application features on existing protocols, storage systems and monitoring services assuming dedicated bandwidth for each application. We will demonstrate these implementations at key national and international conferences. In particular, we will perform these activities:
   - **E-VLBI**: Initial Integration with grid systems including monitoring and end-to-end services. Connect Haystack Observatory to UltraLight (via Bossnet and Abilene) at OC-48. Create plan for connecting U.S. VLBI telescopes for e-VLBI work. Preliminary e-VLBI experiments utilizing BOSSnet.
   - **Oncology**: Development of RCET SOANS system using FAST Protocol for multiple concurrent users assuming dedicated bandwidth. Demonstration of TeleRadiation Therapy for single remote expert interaction.
   - **HENP**: Deploy grid toolkits on UltraLight connected nodes. Integrate collaborative toolkits with UltraLight infrastructure. Prototype and test connection between UltraLight monitoring infrastructure, intelligent agents and HENP grid scheduling support. Utilize HEP data transport applications to test UltraLight performance.
3. Develop novel ENS services including algorithmic development and simulation. Alpha version testing of unit level deployment of ENS services on UltraLight infrastructure and demonstration of better performance on small number of key application features. In particular, we will perform these activities:
   - **Protocols**: Integration and evaluation of FAST TCP and other variants with flagship applications. New protocol development for MPLS networks. Integrate transport protocols with flagship applications. Report on

protocol testing across hybrid network.

- **Storage and Application Services:** Build DSS nodes; Report on the use of DSS and Parallel I/O filesystems on 10 Gb/sec network; Develop application library to collect application access and performance statistics.
- **Monitoring and Simulation:** Investigation, development, evaluation of end-to-end performance monitoring framework. Integration of tools & models to build simulation testbed for network fabric. Demonstrate and report on design and performance of global monitor across hybrid network and clusters.
- **Agents**: Development of Agents for Bandwidth Scheduling, smart memory and buffer management and testing on limited data access patterns. Prototype testing on a limited number of application kernels

C.5.2 Phase 2: Integration (18 months)

1. Significant upgrade of the equipment near the end of this phase (Described in section C4).
2. Development of each application using prototype ENS services assuming dedicated bandwidth as well as shared bandwidth environments. Each application will be demonstrated to successfully exploit the integrated ENS services. We will also demonstrate impact on each of the application areas due to the use of a multi-Gbps wide area network. In particular, we will perform these activities:
   - **E-VLBI**: Testing of e-VLBI, including national and international telescopes. Work with NSF and U.S. telescope owners/ institutions to implement high-speed connectivity.
   - **Oncology**: Development of RCET system for multiple users assuming bandwidth shared with other applications. Demonstration of TeleRadiation Therapy for interactive diagnosis of a single treatment plan with multiple experts each using thin clients.
   - **HENP**: Integrate HENP grid-toolkits with agents and monitoring infrastructure. Deploy collaborative tools for UltraLight and integrate both to enable UltraLight work and to test interplay with UltraLight infrastructure. Optimize network storage systems for use in Ultra-scale networks. Initial work on production quality HENP applications layered on UltraLight aware grid-toolkits**.**
3. Refinement of each of the ENS services with an integrated view on multiple applications and other ENS services. We will develop Beta version software of the ENS services for distribution and use to a number of other applications
   - **Protocols**: Integration & evaluation of MPLS protocols with flagship applications on testbed. Integrating MPLS protocols with flagship applications. Demonstration of transport protocols across hybrid network.
   - **Storage and Application Services:** Integration of Combined Parallel TCP and Fast TCP; Build ns2 simulator modules for applications based on measurements; Assess effects of network protocols and infrastructure on application in simulation; Integrate measurement, simulation, and protocol components into MonALISA.
   - **Monitoring and Simulation:** Investigation, development, integration, & evaluation of non-intrusive sensors & functions for intelligent agent and application access. Extension of simulation testbed with network & storage protocol modules. Demonstrate and report on design and initial performance testing of global monitor with preliminary agents.
   - **Agents**: Develop data mining based agents for application & user profiling, & use for intelligent pre-fetching and replication. Integrated testing of multiple agents on a mixture of workloads from different applications.
4. Successful transfer of technology ideas in important standards such as IETF and GGF to develop standards for high bandwidth communication and monitoring. Dissemination of beta code to non-flagship applications.
5. Large amounts of data will be collected and stored for different ENS services, applications under different mixture of workloads. We will develop an UltraLight benchmark (ala Spec Benchmark for CPUs) containing core communication and data access patterns for a range of high bandwidth applications.

C.5.3 Phase 3: Transition to Production (24 months)

1. Development of production-ready versions of these applications. Demonstration of significant impact on the target applications. In particular, we will do the following activities:
   - **E-VLBI**: Expand UltraLight testing, adding more stations and extending bandwidth. Deploy UltraLight protocols for e-VLBI to production networks. Study feasibility of distributed processing of e-VLBI data.
   - **Oncology**: Development and Demonstration of the RCET system with enhanced segmentation and processing capabilities at the server supporting multiple treatment plans with multiple experts.
   - **HENP**: Fully deployed HENP-UltraLight Grid scheduling and management system (UGSM) utilizing intel-

ligent agents and monitoring. HENP data transport built on agent layer that selects optimal protocols & infrastructure for transfers. Collaborative tools integrated with agents and monitoring data.

2. Refinement of ENS services for production environment. In particular, we will do the following activities:
   - **Protocols**: Integration of all protocols with other services. Software integrating existing transport protocols with other network services. Extensive testing of protocols across hybrid network.
   - **Storage and Application Services**: Refinement of storage and application services based on experiments conducted in earlier phases; Deploy prototype integrated storage and measurement system.
   - **Monitoring and Simulation**: Full integration & performance evaluation of global monitor with agents & flagship apps. Analysis & optimization of sensor and data dissemination profiles. Simulative experiments for alternatives. Demonstrate effectiveness of monitored flagship applications alone and in combination on network.
   - **Agents**: Automatic fine-tuning of system parameters. Demonstration in production environment.
3. Near-production quality ENS software and/or technology transfer to industry. Continue work with the IETF and GGF to develop standards

C.5.4 Project Management

Leadership: The PI, Co-PIs and other senior personnel have extensive experience with the management and successful execution of national-level scientific, networking and Grid projects, including H. Newman (chair of the Standing Committee on Inter-Regional Connectivity [SCI03] chair of the International Committee on Future Accelerators [ICF03], chair of the US-CMS collaboration, PI for the Particle Physics Data Grid), Avery (Director of the multi-disciplinary GriPhyN and iVDGL Grid projects), Ibarra (PI of AMPATH). Their leadership in these projects also provides unique opportunities to exploit their considerable personnel and IT resources for UltraLight's benefit.

Management: The management team will consist of the PI as Director, the Co-PIs, and the Project and Technical Coordinators. The Project and Technical Coordinators will be appointed by the PI from the project personnel. The management team will manage the project, as a Steering Group, with input from the following bodies: (1) International Advisory Committee, international experts from the research, academic and vendor communities, which will review project progress and advise the management team; (2) Applications Group, representing end users and applications served by the UltraLight network, which will report regularly to the management team.

C.6 Broad Impact of This Work

The UltraLight project will build the world's first multilevel, end-to-end transcontinental and transoceanic network that will support multiple switching (circuit switched, packet switched and MPLS/GMPLS) paradigms. This along with local computing and storage systems will form a distributed computing laboratory that will enable the prototyping of ultra large-scale, geographically distributed and high bandwidth applications. We will move this work from the laboratory to production, taking advantage of our national and international partners to ensure relevance. The following paragraphs briefly describe the impact of this project on science and society.

Networking: Development and testing of new network protocols, routing mechanisms and bandwidth management for supporting end-to-end high bandwidth applications will lead to implementation data, practical know-how and networking innovations that will form the basis of next generation wide area national and transoceanic networks.

Target Applications: UltraLight will revolutionize the way business is conducted today in the target applications (a) HENP: Providing hundreds of scientists fast, flexible and easy access to vast distributed data archives; (b) VLBI: attainment of ultra-high resolution images at higher speed and lower cost; (c) Radiation Oncology: allow remote experts to work interactively with local experts to substantially improve patient care and treatment outcomes; (d) Grid projects: provide new tools for ultra-speed data transfers for several new applications in a variety of situations

Ultra Large Scale Applications: The diverse networking requirements of our flagship applications ensure our project's impact on many disciplines and activities – from global scientific and engineering collaborations, to distributed groups of clinicians, to multinational corporations with data-intensive business processes (e.g. data transfers between head office and offshore call centers) and defense agencies.

Grid and Web Services: The project will be pivotal in the development of OGSA and web standards for modeling and managing the communication requirements and end-to-end real-time monitoring needs of large scale distributed organizations. This will enhance other federally supported grid projects such as GriPhyN, iVDGL, and PPDG.

Education and Outreach: The Educational Outreach program will be based at Florida International University

(FIU), leveraging its CHEPREO [CHE03] and CIARA [CIA03] activities to provide students with opportunities in networking research. Our E&O program has several innovative and unique dimensions: (1) integration of students in the core research and application integration activities at all participating universities; (2) utilization of the UltraLight testbed to carry out student-defined networking projects; (3) opportunities for FIU students (especially minorities) to participate in project activities, using the CIARA model; (4) UltraLight involvement through the Research Experiences for Undergraduate program, graduate programs such as CIARA, and teacher programs such as Research Experiences for Teachers and QuarkNet [QNE1]; (5) exploitation of UltraLight's international reach to allow US students to participate in research experiences at laboratories / accelerators located outside US borders, in some cases without having to leave their home universities. In general, UltraLight will provide US students with the opportunity to work internationally and be involved in scientific, cultural, and educational exchanges on a global basis enabled by the experimental infrastructure proposed. This should increase student retention in the sciences and engineering, increase rates of discovery by acclimatizing students to global collaboration early on in their careers, and foster trust relationships among diverse cultures of science and engineering students [DES01].

The E&O opportunities will extend from the networking research across the key application areas of HENP, e-VLBI, and Radiation Oncology, directly involving the project participants and the research that it enables. Networking opportunities range from testing components and small test stand assembly to collating network statistics for research. Applications area participants will be first adopters of the technology, benefiting the project through testing while providing participants with access to cutting edge science. The proposal sets aside 4% of the overall budget for these activities. The E&O budget provides for graduate fellowships at Caltech ($15k), Michigan ($15k), MIT ($10k), U of Florida: ($25k), FIU: ($25k). The majority of the support will be used for student stipends. Recruitment, infrastructure support, and other incidental expenses will be leveraged off projects including CHEPREO, CIARA, GriPhyN, iVDGL and PPDG. A three-day workshop will be held during the first year to organize the efforts, with substantial participation starting by year-end.

D References

[ABWE] Jiri Navratil, Les Cottrell, "A Practical Approach to Available Bandwidth Estimation", published at PAM 2003, April 2003, San Diego. http://moat.nlanr.net/PAM2003/PAM2003papers/3781.pdf.

[AMP01] AMPATH homepage: http://www.ampath.fiu.edu/

[ANS01] E. Hernandez, M. Chidester, and A. George, "Adaptive Sampling for Network Management," http://www.hcs.ufl.edu/pubs/JNSM2000.pdf, *Journal of Network and Systems Management*, Vol. 9, No. 4, Dec. 2001, pp. 409-434.

[ATL03] ATLAS experiment home page, http://atlas.web.cern.ch/Atlas/Welcome.html.

[BAN01] Ayan Banerjee, "Generalized Multiprotocol Label Switching: An Overview of Routing and Management Enhancements", http://www.calient.net/files/GMPLS.pdf, IEEE Communications Magazine, January 2001

[BBFM01] Bassi, A., Beck, M., Fagg, G., Moore, T., Plank, J., Swany, M.,Wolski, R. The Internet BackPlane Protocol: A Study in Resource Sharing. In the proceedings of the second IEEE/ACM International Symposium on
Cluster Computing and the Grid (CCGRID 2002), Berlin, Germany, May 21-24, 2002.

[BBFTP] "Large files transfer protocol", http://doc.in2p3.fr/bbftp/.

[BR02] Bright L., Raschid L, "Using Latency-Recency Profiles for Data Delivery on the Web," Proceedings of the Conference on Very Large Data Bases (VLDB), 2002.

[BSA01] Bill St. Arnaud, "Proposed CA*net4 Network Design and Research Program", http://www.canarie.ca/canet4/library/c4design/canet4_design_document.pdf, March 22, 2001.

[BVRS02] John Bent, Venkateshwaran Venkataramani, Nick LeRoy, Alain Roy, Joseph Stanley, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau, Miron Livny, Flexibility, Manageability, and Performance in a Grid Storage Appliance, The 11th International Symposium on High Performance Distributed Computing (HPDC-11) Edinburgh, Scotland, July 24-26, 2002.

[CAN03] "CA*net4", http://www.canarie.ca/canet4/

[CDF03] CDF experiment home page, http://www-cdf.fnal.gov/.

[CHE01] An Inter-Regional Grid-Enabled High Energy Physics Research and Educational Outreach Center (CHEPREO) at Florida International University (FIU) Proposal PHY-0312038 submitted to NSF, well reviewed, awaiting award (2003).

[CIA01] Center for Internet Augmented Research and Assessment (CIARA) at Florida International University (FIU). Proposal to NSF (2003).

[CMS03] The CMS home page, http://cmsinfo.cern.ch/Welcome.html/.

[CRN03] CERN home page, http://www.cern.ch/.

[DAT03] The DataTAG home page, http://datatag.web.cern.ch/datatag/.

[DES01] DeSanctis et al., "Building a Global Learning Community," Communications of the ACM. Dec. 2001/Vol.44. No. 12. Pages 80 82

[DIG03] Julian Bunn and Harvey Newman, "Data Intensive Grids for High Energy Physics," in Grid Computing: Making the Global Infrastructure a Reality, edited by Fran Berman, Geoffrey Fox and Tony Hey, March 2003 by Wiley.

[DSA01] Harvey B. Newman, Iosif C. Legrand, and Julian J. Bunn, "A Distributed Agent-based Architecture for Dynamic Services," http://clegrand.home.cern.ch/clegrand/CHEP01/chep01_10-010.pdf, CHEP - 2001, Beijing, Sept 2001.

[DZR03] D0 experiment home page, http://www-d0.fnal.gov/

[FNL01] Fermi National Labs homepage, http://www.fnal.gov/

[FST03a] S. H. Low, "Duality Model of TCP and Queue Management Algorithms",

http://netlab.caltech.edu/pub/papers/duality.ps, to appear IEEE/ACM Trans. on Networking, October 2003

[FST03b] C. Jin, D. Wei, S. H. Low, G. Buhrmaster, J. Bunn, D. H. Choe, R., L. A. Cottrell, J. C. Doyle, H. Newman, F. Paganini, S. Ravot, S. Singh, "FAST Kernel: Background Theory and Experimental Results", http://netlab.caltech.edu/pub/papers/pfldnet.pdf, Presented at the First International Workshop on Protocols for Fast Long-Distance Networks, February 3-4, 2003, CERN, Geneva, Switzerland.

[GELW00] "Strategic Directions Moving the Decimal Point: An Introduction to 10 Gigabit Ethernet," http://newsroom.cisco.com/dlls/innovators/metro_networking/10gig_wp1.pdf

[GEM03a] P. Raman, A. George, M. Radlinski, and R. Subramaniyan, "GEMS: Gossip-Enabled Monitoring Service for Heterogeneous Distributed Systems," http://www.hcs.ufl.edu/pubs/GEMS2002.pdf, submitted to *Journal of Network and Systems Management*.

[GEM03b] GEMS web site, http://www.hcs.ufl.edu/prj/ftgroup/teamHome.php.

[GMP01] Banerjee, A., Kompella, K., Rekhter Y., et al, "Generalized Multiprotocol Label Switching: An Overview of Routing and Management Enhancements", IEEE Communications Magazine, January 2001.

[GMP03] Eric Mannie et al, "Generalized Multi-Protocol Label Switching Architecture (IETF Draft)," http://www.ietf.org/internet-drafts/draft-ietf-ccamp-gmpls-architecture-06.txt.

[GMPLS00] "Advanced Developments in Integrating IP Infrastructures and Optical Transport Networks," http://www.cisco.com/networkers/nw01/pres/preso/Opticaltechnologies/OPT-411final1.pdf

[GOS01a] S. Ranganathan, A. George, R. Todd, and M. Chidester, "Gossip-Style Failure Detection and Distributed Consensus for Scalable Heterogeneous Clusters," http://www.hcs.ufl.edu/pubs/CC2000.pdf, *Cluster Computing*, Vol. 4, No. 3, July 2001, pp. 197-209.

[GOS01b] D. Collins, A. George, and R. Quander, "Achieving Scalable Cluster System Analysis and Management with a Gossip-based Network Service," http://www.hcs.ufl.edu/pubs/LCN2001.pdf, *Proc. IEEE Conference on Local Computer Networks (LCN)*, Tampa, FL, November 14-16, 2001.

[GOS03] K. Sistla, A. George, and R. Todd, "Experimental Analysis of a Gossip-based Service for Scalable, Distributed Failure Detection and Consensus," http://www.hcs.ufl.edu/pubs/GOSSIP2001.pdf, *Cluster Computing*, Vol. 6, No. 3, 2003, pp. 237-251.

[GOS99] M. Burns, A. George, and B. Wallace, "Simulative Performance Analysis of Gossip Failure Detection for Scalable Distributed Systems," http://www.hcs.ufl.edu/pubs/CC1999.pdf, *Cluster Computing*, Vol. 2, No. 3, 1999, pp. 207-217.

[GRI03] GriPhyN home page, http://www.griphyn.org/.

[GridDT] Sylvain Ravot, GridDT, Presented at the First International Workshop on Protocols for Fast Long-Distance Networks, February 3-4, 2003, CERN, Geneva, Switzerland. http://datatag.web.cern.ch/datatag/pfldnet2003/slides/ravot.ppt.

[GridFTP] "The GridFTP Protocol and Software", http://www.globus.org/datagrid/gridftp.html.

[HaNo2002] Hacker, T., Noble, B., Athey, B., "The Effects of Systemic Packet Loss on Aggregate TCP Flows," Proceedings of SuperComputing 2002, November, 2002, Baltimore, MD.

[HST03] Sally Floyd, "HSTCP: HighSpeed TCP for large congestion windows," Internet draft draft-floyd-tcp-highspeed-02.txt, work in progress, http://www.icir.org/floyd/hstcp.html, February 2003.

[IEPM] "Experiences and Results from a New High Performance Network and Application Monitoring Toolkit", Les Cottrell, Connie Logg, I-Heng Mei, published at PSAM 2003, April 2003, San Diego, http://moat.nlanr.net/PAM2003/PAM2003papers/3768.pdf.

[IPERF] "Measuring end-to-end bandwidth with Iperf using Web100", http://moat.nlanr.net/PAM2003/PAM2003papers/3801.pdf.

[IVD03] International Virtual Data Grid Laboratory home page, http://www.ivdgl.org/.

[LAA01] Cees de Laat, et. Al., "The Rationale of the Current Optical Networking Initiatives," http://carol.wins.uva.nl/~delaat/techrep-2003-1-optical.pdf, Technical Report 2003.

[LAA02] Gommons, L., de Laat, C., Oudenaarde, A. T., "Authorization of a QoS Path based on Generic AAA," http://carol.wins.uva.nl/~delaat/techrep-2003-3-aaa.pdf.

[LAN-PHY] "Frequently Asked Questions," http://www.10gea.org/10GEA_FAQ_0602.pdf.

[LHC03] The Large Hadron Collider home page, http://lhc-new-homepage.web.cern.ch/lhc-new-homepage/.

[MCBI03] The Michigan Center for Biological Information home page, http://www.ctaalliance.org/MCBI/index.html

[MDS02] Le Faucheur, F. et al, "MPLS Support of Differentiated Services (RFC3270)," http://www.ietf.org/rfc/rfc3270.txt.

[MLD03] MLDesign Technologies web site, http://www.mldesigner.com/.

[MON03a] MonALISA web page, http://monalisa.cern.ch/MONALISA/.

[MON03b] Recent developments in MonALISA, http://monalisa.cacr.caltech.edu/developments/.

[MON99] L. Barone et al. (The MONARC Architecture Group), Regional Centers for LHC Computing, MONARC-DOCS 1999-03, http://monarc.web.cern.ch/MONARC/docs/monarc_docs/1999-03.html (1999).

[MPA01] Rosen, E., Viswanathan, A. and Callon R., "Multiprotocol Label Switching (RFC3031)," http://www.ietf.org/rfc/rfc3031.txt.

[MRT01] Awduche, D. et al, "RSVP-TE: Extensions to RSVP for LSP Tunnels (RFC3209)," http://www.ietf.org/rfc/rfc3209.txt.

[MST03] The Minimum Spanning Tree applied to VRVS, http://monalisa.cacr.caltech.edu/.

[NET03] "NetherLight", http://carol.wins.uva.nl/~delaat/optical/.

[NETFLOW] "Cisco IOS Netflow", http://www.cisco.com/warp/public/732/Tech/nmp/netflow/index.shtml.

[NLR03] "National Light Rail Initiative" http://www.cwru.edu/its/strategic/national_light_rail.htm

[NFSv4] "Center for Information Technology Integration, University of Michigan," http://www.citi.umich.edu/projects/asci/index.html.

[NMWG] Bruce Lowekamp, Brian Tierney, Les Cottrell, Richard Hughes-Jones, Thilo Kielman, Martin Swany, "A Hierarchy of Network Performance Characteristics for Grid Applications and Services", available http://www-didc.lbl.gov/NMWG/docs/measurements.pdf.

[NS03] Network simulation ns-2 web site, Information Sciences Institute, University of Southern California, http://www.isi.edu/nsnam/ns/.

[OFC99] S. H. Low and D. E. Lapsley, "Optimization Flow Control, I: Basic Algorithm and Convergence" http://netlab.caltech.edu/FAST/papers/ofc1_ToN.pdf, IEEE/ACM Transactions on Networking, 7(6):861-75, Dec. 1999.

[PPD03] The Particle Physics Data Grid (2003), http://www.ppdg.net/.

[PVD03] J.R. Palta, V.A. Frouhar, and J.F. Dempsey, "Web-based submission, archive, and review of radiotherapy data for clinical quality assurance: a new paradigm," Accepted for publication in the International Journal of

Radiation Oncology, Biology, Physics. March 2003.

[QNE1] Details of the QuarkNet program may be found at the website: http://quarknet.fnal.gov/See also: K. Riesselmann, "Weaving the QuarkNet", FermiNews (July 21, 2000) and M. Bardeen, R.M. Barnett, K. Cecire, T. Jordan, "Caught in the QuarkNet", CERN Courier, Vol. 40, No. 1 (Jan-Feb 2000).

[RAJ02] Bala Rajagopalan, et al, "IP over Optical Networks: A Framework," http://www.ietf.org/proceedings/02mar/I-D/draft-ietf-ipo-framework-01.txt IETF Drafts, Expires August 22, 2002.

[RWM02] Arcot Rajasekar, Michael Wan and Reagan Moore, MySRB & SRB - Components of a Data Grid. The 11th International Symposium on High Performance Distributed Computing (HPDC-11) Edin-

burgh, Scotland, July 24-26, 2002

[SAB03] SABUL homepage: http://www.dataspaceweb.net/sabul-faq-03.htm.

[SIM03] A. George, I. Troxel, J Han, N. Dilakanont, J. Wills, and T. McCaskey, "Virtual Prototyping of High-Performance Optical Networks for Advanced Avionics Systems," http://www.hcs.ufl.edu/prj/opngroup/RockwellMtg11Feb03.ppt, Advanced Networking Meeting, Boeing Corp., St. Louis, MO, February 11, 2003.

[SOM01] Ed. K. Obermayer and T.J. Sejnowski, "Self Organizing Mao Formation", MIT Press, 2001.

[SON01] Harvey B. Newman, Iosif C. Legrand, A Self Organizing Neural Network for Job Scheduling in Distributed Systems, CMS NOTE 2001/009, January 8, 2001, http://clegrand.home.cern.ch/clegrand/SONN/note01_009.pdf

[SON97] B. Fritzke, "A self-organizing network that can follow non-stationary distributions,"

Proc. of the International Conference on Artificial Neural Networks 97, Springer, 1997

[STC02] Tom Kelly, "Scalable TCP: Improving performance in highspeed wide area networks". Submitted for publication, http://www-lce.eng.cam.ac.uk/~ctk21/scalable/, December 2002.

[STL01] Tom DeFanti, StarLight, http://www.startap.net/starlight/.

[SVG03] D. H. Choe and S. H. Low, "Stabilized Vegas," http://netlab.caltech.edu/pub/papers/svegas-infocom03.pdf, Proceedings of IEEE Infocom, San Francisco, April 2003.

[TAQ03] F. Paganini, Z. Wang, S. H. Low and J. C. Doyle, "A new TCP/AQM for stable operation in fast networks," http://netlab.caltech.edu/pub/papers/fast-infocom03.pdf, Proceedings of IEEE Infocom, San Francisco, April 2003.

[TRA03] "TransLight", http://www.internet2.edu/presentations/20030409-TransLight-DeFanti.ppt

[TSU03] Tsunami home page, http://www.indiana.edu/~anml/anmlresearch.html.

[TSUNAMI] Available at http://www.ncne.nlanr.net/training/techs/2002/0728/presentations/200207-wallace1_files/v3_document.htm.

[UKL03] "UKLight", http://www.cs.ucl.ac.uk/research/uklight/

[UVE02] S. H. Low, Larry Peterson and Limin Wang, "Understanding Vegas: A Duality Model", http://netlab.caltech.edu/FAST/papers/vegas.pdf, Journal of ACM, 49(2):207-235, March 2002

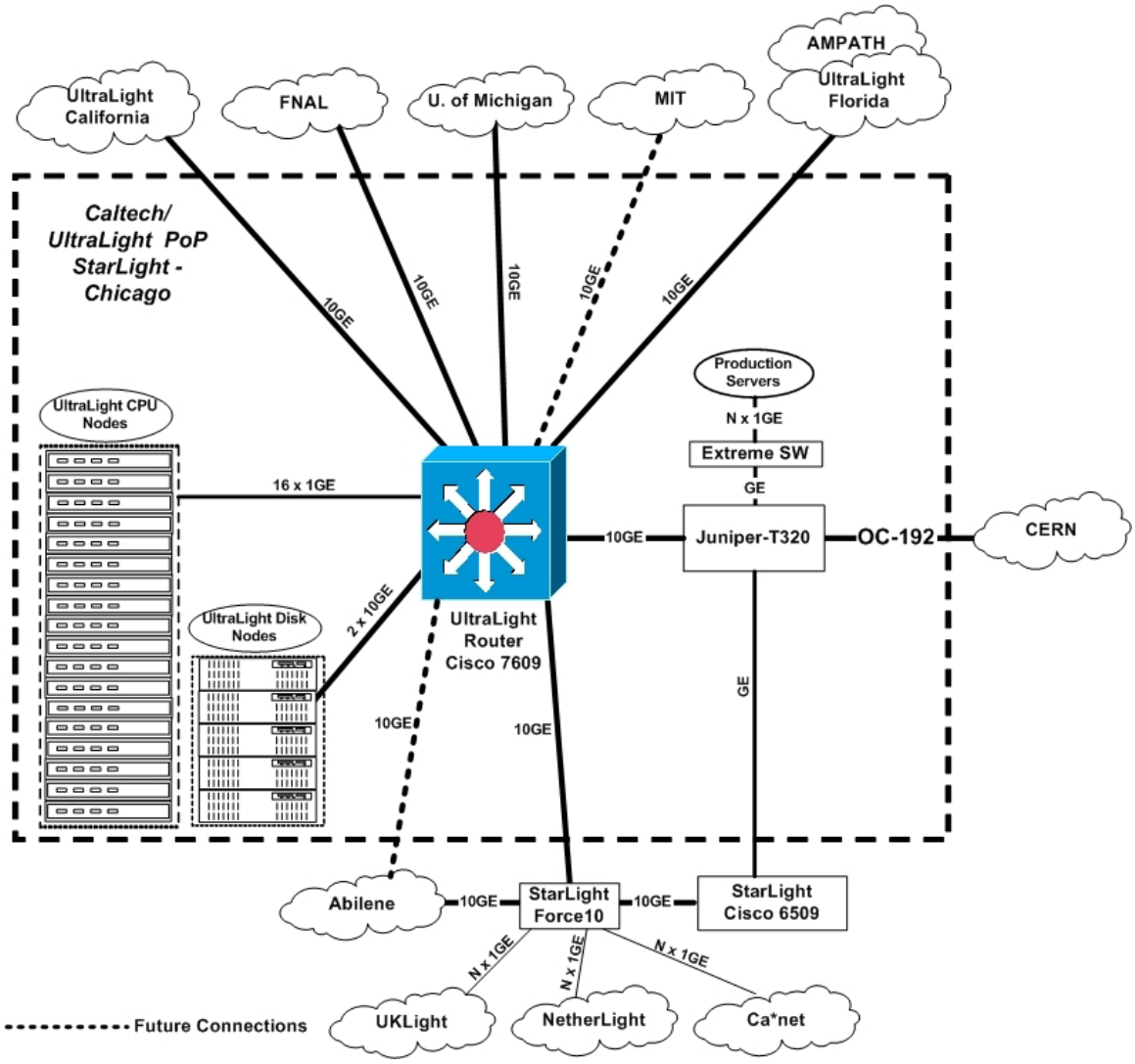
[WEB100] "Web100", http://www.web100.org/.

[WHI03] Alan Whitney et al, "The Gbps e-VLBI Demonstration Project". ftp://web.haystack.edu/pub/e-vlbi/demo_report.pdf.

[XCP02] Dina Katabi, Mark Handley, and Charlie Rohrs, "Congestion Control for High Bandwidth-Delay Product Networks." http://www.ana.lcs.mit.edu/dina/XCP/p301-katabi.ps, In the proceedings on ACM Sigcomm 2002.

E Facilities and Leveraged Equipment: UltraLight Optical Switching Fabric

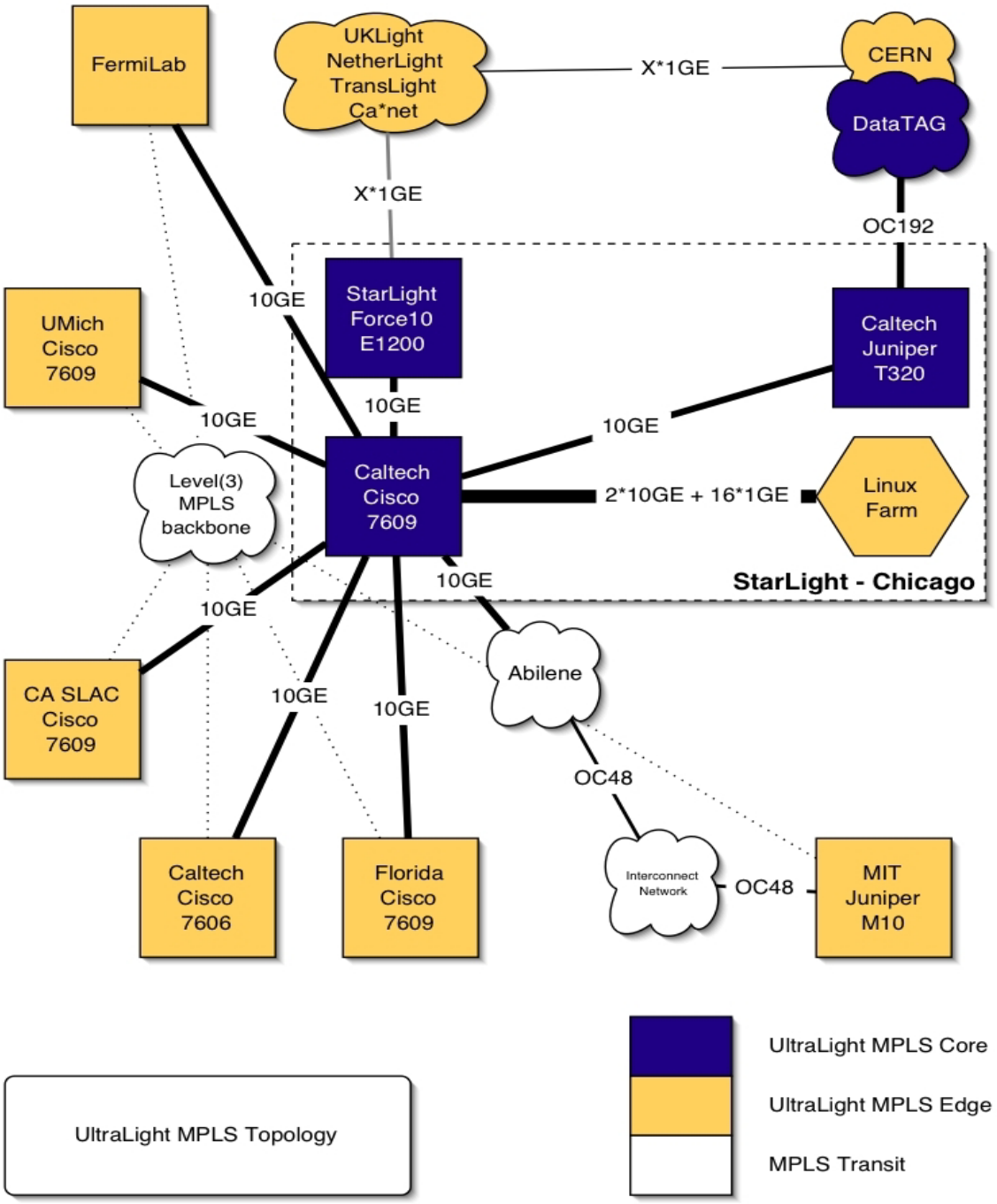
Figure 3: UltraLight Optical Switching Fabric



E.1 UltraLight MPLS Network Architecture

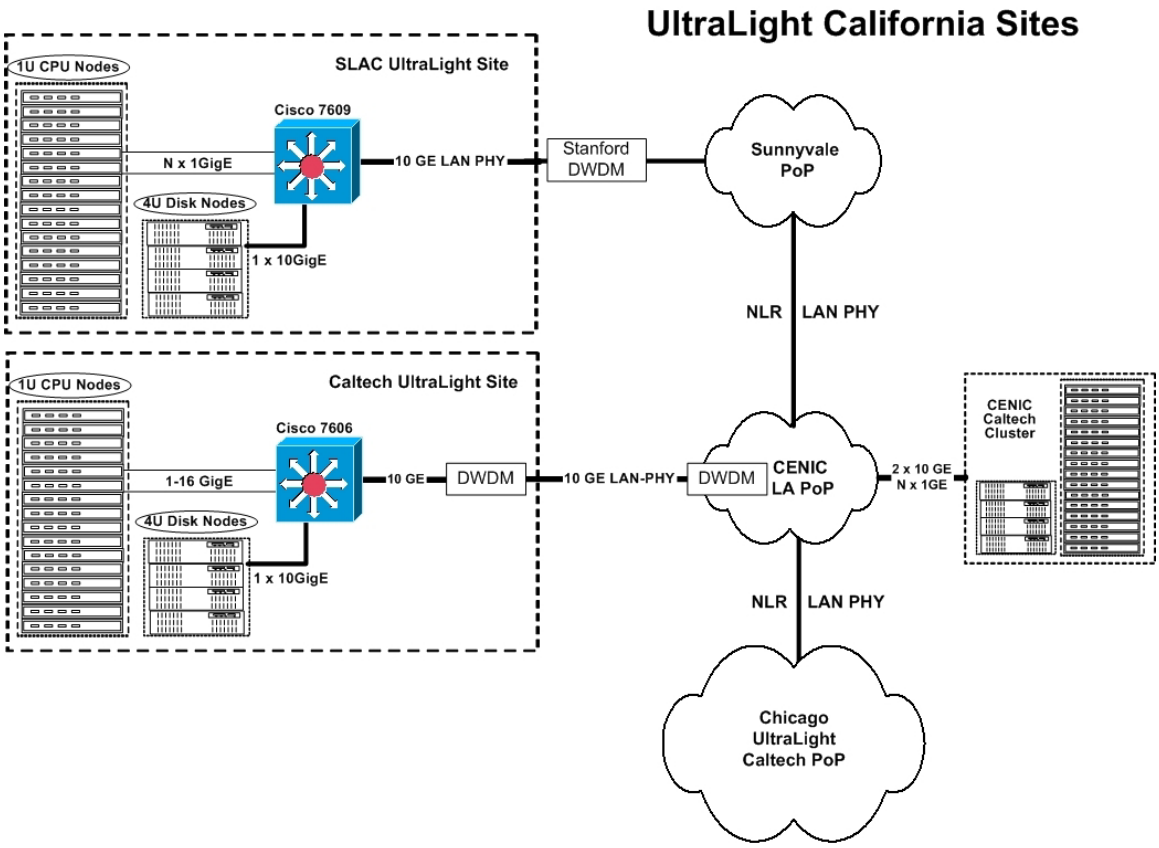
The MPLS network architecture for UltraLight will consist of a core of Cisco and Juniper Label Switch Routers (see Figure 4 below). These routers will interconnect the various edge sites where the Label Edge Routers will reside. Physically, the network will be connected in a star topology. It will provide basic transport services with and without bandwidth reservation, differentiated services support, and more advanced services such as Virtual Private Networks and Virtual Private LAN services. The UltraLight MPLS network will peer with other MPLS networks and use techniques such as priority queuing and shaping to interwork with those networks and provide end to end MPLS services for sites not directly connected into the UltraLight MPLS core. The UltraLight MPLS network will be integrated closely with the autonomous agents that make up the intelligent monitoring and management infrastructure.

Figure 4: UltraLight MPLS Network



E.2 Caltech

Figure 5 California Network and Storage Servers Site Diagram



E.2.1 Leveraged Facilities at Caltech

The following lists detail Caltech's leveraged network facilities:

Caltech StarLight Network Equipment

1. 1 Cisco 7606:Catalyst 6000 SU22/MSFC2 SERVICE PROVIDER W/VIP (supervisor) 2-port OC-12/STM-4 SONET/SDH OSM, SM-IR, with 4 GE Catalyst 6500 Switch Fabric Module (WS-C6500-SFM)
2. 1 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 10 Gigabit Ethernet Module with 1310nm long haul OIM and DFC card 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).
3. 1 Cisco 2950 24 10/100 ports + 2*1000BASE-SX ports Role: Fast Ethernet switch for production with 2 Gbps uplinks.
4. 1 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).
5. 1 Juniper M10 1 port 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.

6. 1 Extreme Summit 5i GbE Gigabit Ethernet switch with 16 ports Role: Network elements interconnection at Gbps speed.
7. 1 Cisco 7204VXR Two-port T3 serial PA enhanced (PA-2T3+) Gigabit Ethernet Port Adapter (PA-GE) Role: Old router for backup and tests.
8. 1 Alcatel 1670 (belong to CERN) 1*OC-48 port. 2 GBE ports Role: Element of the multi-platforms testbed. SONET multiplexer.
9. 1 Alcatel 7770 (belong to CERN) 2 port OC-48 8 port OC-12 8 port GBE Role: Element of the multi-platforms testbed (Datatag project).

E.3 Internet2

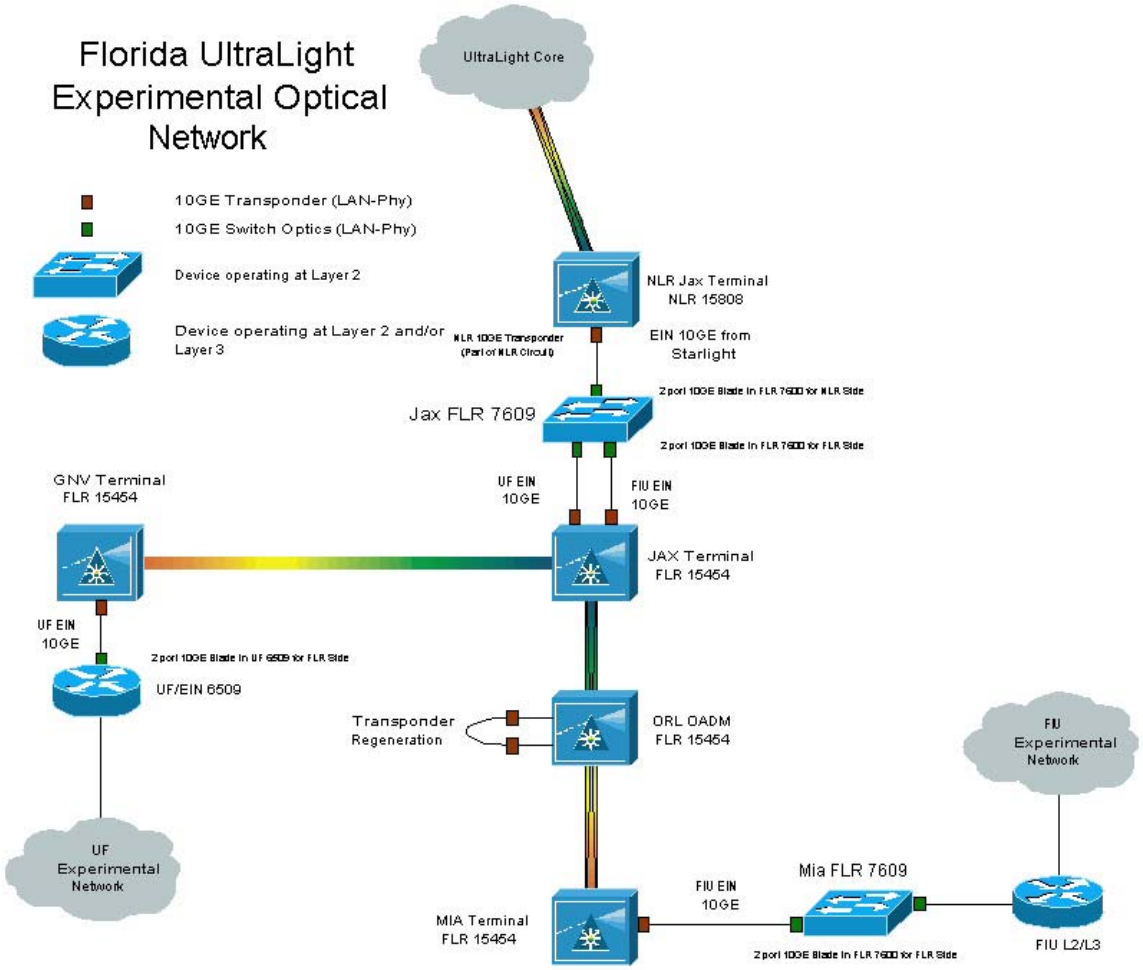
E.3.1 Internet2's contributions will encompass several resources.

1. A special experimentally focused interconnection in Chicago of UltraLight to Internet2's 10 Gb/s Abilene backbone network. At a minimum, this will be done using Abilene's existing 10 Gb/s connection to the StarLight switch. If a separate fiber pair is made available, this will be done with a new dedicated 10 Gb/s connection to the UltraLight switch.
2. Leveraging Internet2's role as a founding member of the NLR effort, engineering resources will be made available to help with the design and engineering of UltraLight's NLR-based facilities.
3. The one-way delay measurement technology as well as other techniques developed in the Abilene Observatory project will be made available for us in the performance measurement activities of UltraLight.
4. Leveraging Internet2's capabilities in end-to-end performance, engineering resources will be made available to study the performance achieved for aggressive applications over the UltraLight facility, and to compare it with the performance achieved over the Internet2 infrastructure.
5. Internet2 will collaborate in the consideration of specific experimentally focused MPLS tunnels between designated sites on the UltraLight facility and on the Internet2 infrastructure, both to broaden the reach of UltraLight's experimental facility and to test the relative efficacy of the MPLS-based techniques developed as part of the UltraLight project.

More generally, Internet2 will engage with UltraLight in understanding how to support UltraLight applications most effectively, both in the current Internet2 production infrastructure, in the proposed UltraLight experimental infrastructure, and in future forms of Internet2's production infrastructure.

E.4 University of Florida

Figure 6: Florida UltraLight Optical Network



The University of Florida and Florida International University will use Florida's emergent Florida Lambda Rail (FLR) optical network to create an experimental network to connect to UltraLight. FLR connects to National Lambda Rail (NLR) in Jacksonville. UFL and FIU each will provision 10GbE LAN-PHY wavelengths to the optical cross-connect (OXC) in Jacksonville, from where UFL and FIU will share another 10GbE LAN-PHY wavelength across NLR will connect Florida's UltraLight network to the UltraLight optical core.

E.4.1 Leveraged Facilities at UFL

The University of Florida computing equipment is configured as a prototype Tier2 site as part of the 5 Tier global computing infrastructure for the CMS experiment at the LHC. It includes many rack-mounted servers and several TeraBytes of RAID storage. The system is intended for use as a general purpose computing environment for LHC physicists. Other tasks include large scale production of Monte Carlo simulations, high speed network transfers of object collections for analysis, and general prototyping and development efforts in the context of the work on the International Virtual Data Grid Laboratory (iVDLG), which is setting up a global grid testbed. The Tier2 includes a fully up to date software environment with the latest versions of operating systems, firmware and Grid software.

E.5 Florida International University

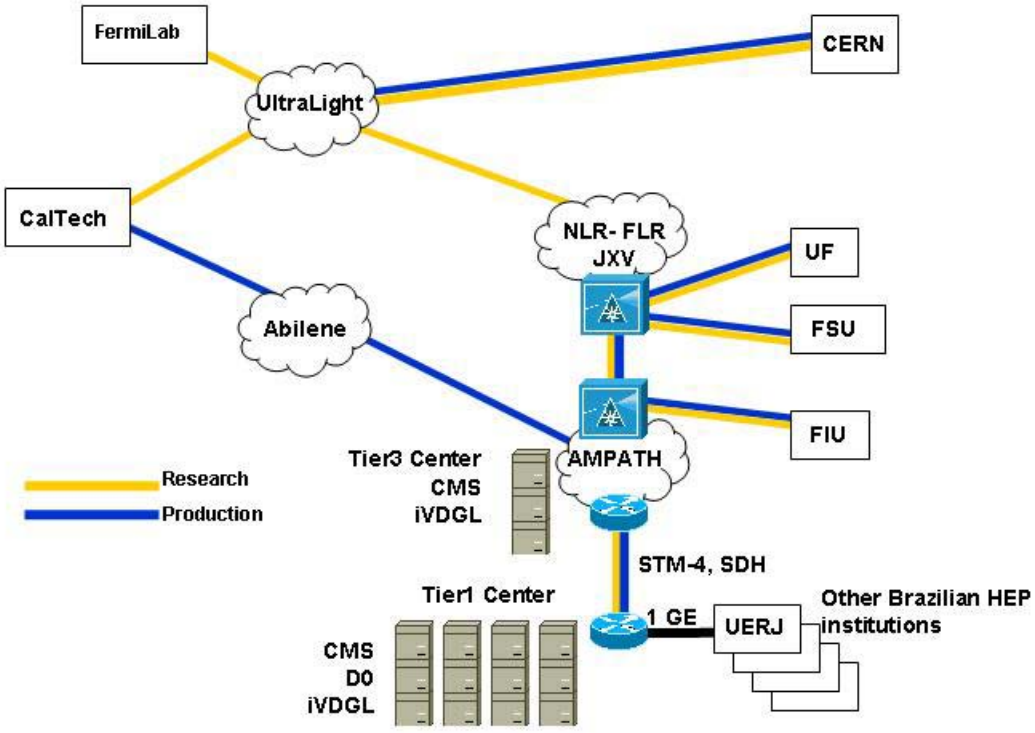
E.5.1 FIU to UltraLight Connectivity Description

FIU will connect to UltraLight through a 10GbE LAN-PHY wavelength interconnecting Miami to Jacksonville, then sharing a 10GbE LAN-PHY wavelength to the UltraLight optical core with UFL (see Figure 6 above). The connection in Miami will be from the NAP Of The Americas, where the AMPATH PoP is located. AMPATH serves as the international exchange point for research and education networks in the Americas.

E.5.2 Leveraged Facilities at FIU

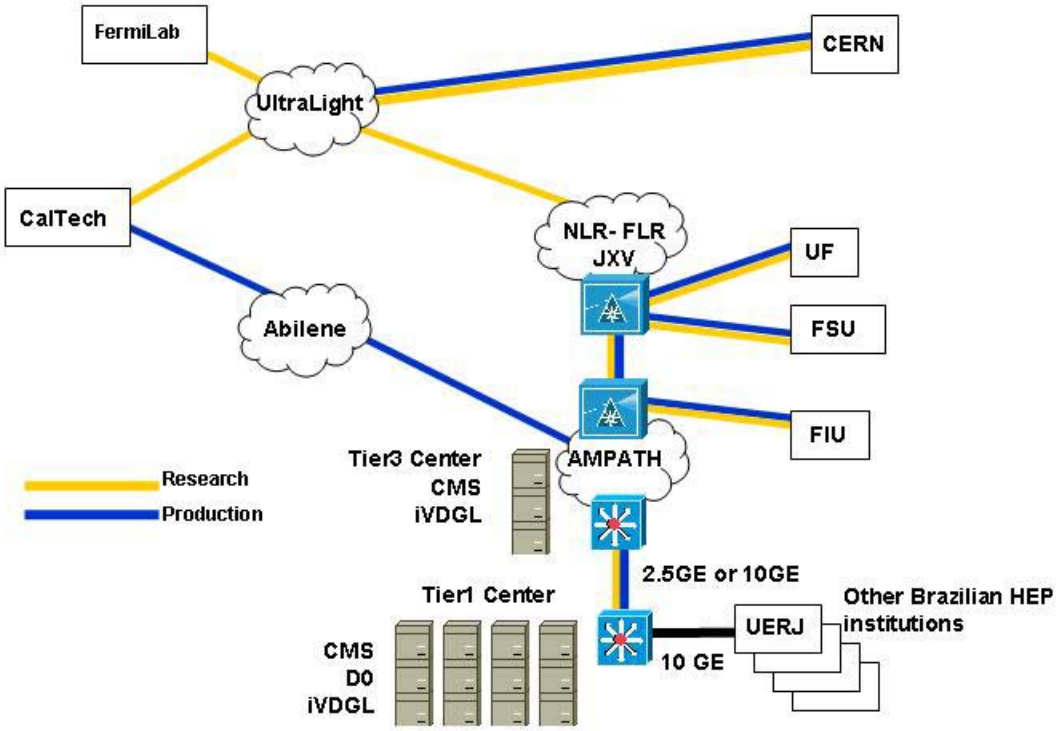
FIU is the lead institution, in collaboration with Caltech, University of Florida and Florida State University, proposing to develop an inter-regional Center for High-Energy Physics Research Education and Outreach (CHEPREO). CHEPREO is a 5-year program, that in year 1, would establish an STM-4 (622 Mbps) 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, UltraLight could leverage the availability of an experimental network to South America. Likewise, the Tier1 facility in Rio and Brazil's HENP community can access UltraLight for connectivity to the Global Grid community. Figure 7 shows how the emergent Grid Physics Tier1 facility at the State University of Rio de Janeiro (UERJ) would be connected to Miami and UltraLight.

Figure 7: UltraLight International Connection to Brazil (years 1-3)



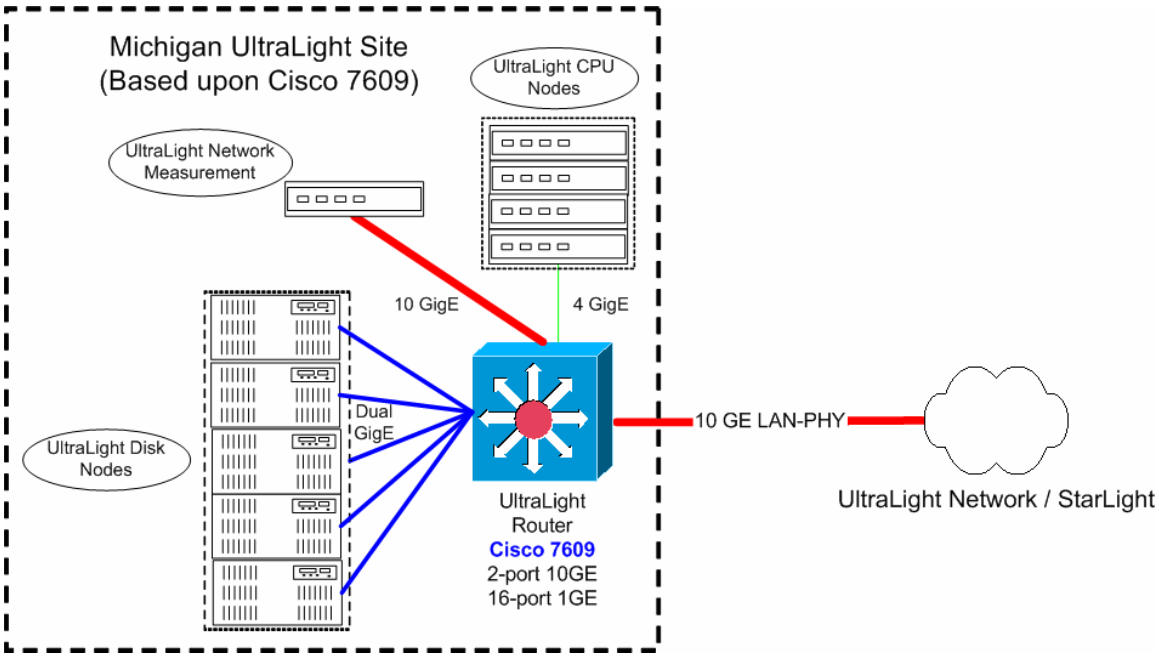
Leveraging the experimental network in Florida, and then transiting NLR, Brazil and South America would be able to participate in UltraLight. By year 3 or possibly sooner, the load on the inter-regional STM-4 circuit is expected to reach capacity. As soon as possible, this circuit should be upgraded to a 2.5G or 10G wavelength and a Layer2 connection extended to South America, as is to other UltraLight Peer Sites. The following figure shows L2-L3 equipment by which South America can connect to the UltraLight optical fabric.

Figure 8: UltraLight International Connection to Brazil (years 4-5)



E.6 University of Michigan

Figure 9: University of Michigan Network Site Diagram



E.6.1 University of Michigan to UltraLight Connectivity Description

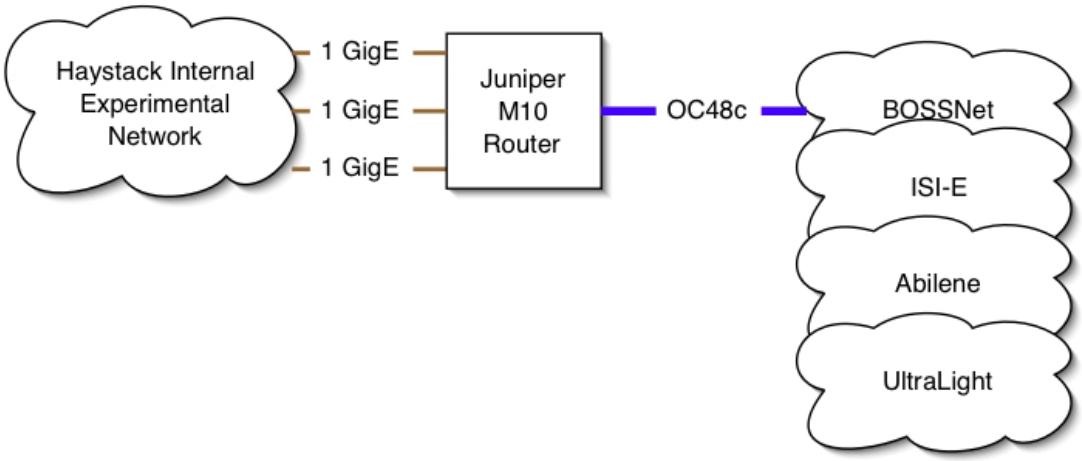
As shown in the above figure, Michigan will have a 10 GE LAN-PHY connection to UltraLight via a University of Michigan owned fiber from Ann Arbor to StarLight in Chicago. We intend to populate the Michigan UltraLight PoP with three kinds of equipment:

1. Network storage servers based upon commodity components, capable of driving the full 10 Gigabits/sec of UltraNet bandwidth via numerous Gigabit Ethernet connections operating in parallel
2. A small number of CPU nodes connected at 1 Gigabit.
3. A network monitoring and testing station connected at 10 Gigabits

Michigan will provide the wave transport for the UltraLight 10GE connection.

E.7 MIT/Haystack

Figure 10 MIT/Haystack Network Site Diagram



E.7.1 MIT/Haystack to UltraLight Connectivity Description

Option A

MIT/Haystack is currently planning to connect into StarLight via Abilene. We will add an extra OC48c wavelength to our existing fiber connection into ISI-E. This path goes via GLOWnet and then BOSSnet into ISI-E in Arlington, VA. We plan to terminate there on a Juniper M40 router which would then connect directly into Abilene. This is subject to approval. Abilene would provide an IP-based Layer3 Transport service to StarLight.

Option B

Our other alternative is to connect directly in to the Level 3 POP at 300 Bent St, Cambridge. In order to get to this POP, we would be required to add another wavelength to our existing connection, strip it off and re-direct through MIT campus to the Level 3 POP. From there we would connect directly over an NLR wavelength to Chicago.

E.7.2 Leveraged facilities at Haystack Observatory

The following existing facilities will be leveraged in the execution of the work under this proposal:

1. Mark 4 VLBI correlator facility, supported by NASA and NSF
2. Juniper M10 router; on loan from Caltech
3. Mark 5 e-VLBI data systems – supported by NASA, NSF and DARPA (http://web.haystack.mit.edu/mark5/Mark5.htm)
4. Glownet (fiber connection from Haystack Observatory to MIT Lincoln Lab) –supported by MIT Lincoln Laboratory, including various routers and switches (ftp://web.haystack.edu/pub/e-vlbi/demo_report.pdf)
5. Bossnet (fiber connection from MIT Lincoln Lab to ISI-E) – supported by MIT Lincoln Laboratory, including various routers and switches (http://www.ll.mit.edu/AdvancedNetworks/bossnet.html)
6. Facilities of ISI-E – supported by Information Sciences Institute of USC, including routers and switches (http://www.east.isi.edu/)
7. Facilities of MAX – supported by the Mid-Atlantic Crossroads consortium, including routers and switches (http://www.maxgigapop.net/)