DOE F 4650.2
(10-99)
(All Other Editions Are Obsolete)

Department of Energy
Office of Science (SC)

OMB Control No.
1910-1400
(OMB Burden Disclosure
Statement on Back)

Face Page

TITLE OF PROPOSED RESEARCH: Science UltraNet: An Ultra High-Speed Network Test-Bed for Distributed Terascale Computing and Large-Scale Science Applications

1. CATALOG OF FEDERAL DOMESTIC ASSISTANCE #:
81.049

2. CONGRESSIONAL DISTRICT:
Applicant Organization's District: 2nd & 3rd Districts
Project Site's District: 2nd & 3rd Districts

3. I.R.S. ENTITY IDENTIFICATION OR SSN:
621788325

4. AREA OF RESEARCH OR ANNOUNCEMENT TITLE/#:
Notice 03-01
Continuing Solicitation for Office of Science Programs
Focus Element: High Performance Networks

5. HAS THIS RESEARCH PROPOSAL BEEN SUBMITTED TO ANY OTHER FEDERAL AGENCY?
Yes ____ No X

PLEASE LIST: ____

6. DOE/OER PROGRAM STAFF CONTACT (if known):
Thomas D. Ndousse (301) 903-9960

7. TYPE OF APPLICATION:

New	X	Renewal	
Continuation		Revision	
Supplement			

8. ORGANIZATION TYPE:

Local Govt.		State Govt.	
Non-Profit		Hospital	
Indian Tribal Govt.		Individual	
Other	X	Inst. of Higher Educ.	
For-Profit			
Small Business		Disadvan. Business	
Women-Owned		8(a)	

9. CURRENT DOE AWARD # (IF APPLICABLE): ____

10. WILL THIS RESEARCH INVOLVE:
10A Human Subjects No X If yes, ____
Exemption No. ____ **or**
IRB Approval Date ____
Assurance of Compliance No: ____
10B Vertebrate Animals N No If yes, ____
IACUC Approval Date ____
Animal Welfare Assurance No: ____

11. AMOUNT REQUESTED FROM DOE FOR ENTIRE PROJECT PERIOD $ $4,583K

12. DURATION OF ENTIRE PROJECT PERIOD:
8/1/2003 **to** Open
Mo/day/yr. Mo/day/yr.

13. REQUESTED AWARD START DATE
8/1/2003 (Mo/day/yr.)

14. IS APPLICANT DELINQUENT ON ANY FEDERAL DEBT?
Yes (attach an explanation) ____ No X

15. PRINCIPAL INVESTIGATOR/PROGRAM DIRECTOR NAME, TITLE, ADDRESS, AND PHONE NUMBER

Nageswara S. V. Rao
Distinguished R&D Staff
Computer Science and Mathematics Division
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, TN 37831-6355
(865) 574-7517

16. ORGANIZATION'S NAME, ADDRESS AND CERTIFYING REPRESENTATIVE'S NAME, TITLE, AND PHONE NUMBER

Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, TN 37831

Thomas Zacharia
Deputy Associate Laboratory Director
High Performance Computing
(865) 574-4897

SIGNATURE OF PRINCIPAL INVESTIGATOR/
PROGRAM DIRECTOR 7/28/2003
Date

PI/PD ASSURANCE: I agree to accept responsibility for the scientific conduct of the project and to provide the required progress reports if an award is made as a result of this submission. Willful provision of false information is a criminal offense. (U.S. Code, Title 18, Section 1001).

SIGNATURE OF ORGANIZATION'S CERTIFYING
REPRESENTATIVE 7/28/2003
Date

CERTIFICATION & ACCEPTANCE: I certify that the statements herein are true and complete to the best of my knowledge, and accept the obligation to comply with DOE terms and conditions if an award is made as the result of this submission. A willfully false certification is a criminal offense. (U.S. Code, Title 18, Section 1001).

NOTICE FOR HANDLING PROPOSALS
This submission is to be used only for DOE evaluation purposes and this notice shall be affixed to any reproduction or abstract thereof. All Government and non-Government personnel handling this submission shall exercise extreme care to ensure that the information contained herein is not duplicated, used, or disclosed in whole or in part for any purpose other than evaluation without written permission except that if an award is made based on this submission, the terms of the award shall control disclosure and use. This notice does not limit the Government's right to use information contained in the submission if it is obtainable from another source without restriction. This is a Government notice, and shall not itself be construed to impose any liability upon the Government or Government personnel for any disclosure or use of data contained in this submission.

PRIVACY ACT STATEMENT
If applicable, you are requested, in accordance with 5 U.S.C., Sec. 562A, to voluntarily provide your Social Security Number (SSN). However, you will not be denied any right, benefit, or privilege provided by law because of a refusal to disclose your SSN. We request your SSN to aid in accurate identification, referral and review of applications for research/training support for efficient management of Office of Science grant/contract programs.

Science UltraNet: An Ultra High-Speed Network Testbed for Distributed Terascale Computing and Large-Scale Science Applications

Principal Investigators:

Nageswara S. Rao	***ORNL*** – Tel: (865) 574-7517; raons@ornl.gov
William R. Wing	***ORNL***
Tom Dunigan	***ORNL***

Collaborators:

Linda Winkler	***ANL***
Vicky White	***FNAL***
Randall D. Burris	***ORNL***
Micah Beck	***U. Tennessee***
Richard Mount	***SLAC***

Official Signing for Laboratory: *Thomas Zachariah*
Associate Laboratory Director, Computing and Computational Sciences
Phone: (865) 574-4897
Fax: (865) 574-4839
Email: Zachariah@ornl.gov

Submitted to:

Dr. Thomas D. Ndousse
High-Performance Networks Program, Office of Science
U. S. Department of Energy, 19901 Germantown Rd
Germantown, MD 20874-1207; Phone: (301) 903-9960
Email: tndoussee@er.doe.gov

Summary

The next generation of DOE large-scale science facilities, projects, and programs, such as terascale computing facilities, the Spallation Neutron Source, Atlas, and BaBar projects, the terascale supernova initiative, nanoscale materials research, bio-geochemical climate simulations, and genomics, all have requirements that will drive extreme networking. In particular, it is extremely important for the network bandwidths to "match" the data generation speeds of DOE terascale computing facilities, which are expected to exceed 100 Teraflops within the next few years. While some of these projects and programs will merely require petabyte data transfers, others will require distributed collaborative visualization, remote computational steering, and remote instrument control. These latter requirements place very different, possibly conflicting or mutually exclusive, but very stringent demands on the network.

The obvious approach is to take advantage of current optical networking technologies to build a network in which lambdas can be dynamically switched into service as needed. These lambdas will provide dedicated channels for services that must be segregated or provide parallel channels for additional raw bandwidth when needed. The challenge is that such a network has never been built, and the logistical details of the dynamic provisioning, the creation of a network environment in which applications can call for service as needed, and the actual deployment of the sorts of high-performance protocols to be used, all need to be developed.

We propose to build a testbed to prototype and test advanced network technologies and services that harness the abundant bandwidth in optical backbone networks to support network-intensive science applications. The testbed will explore radically new transport protocols and dynamic provisioning methods based on the Generalized Multi-Protocol Lambda Switching technology to extend optical channels directly to science applications. Dynamic provisioning allows quite different network characteristics to be selected and optimized as needed. Dedicated bandwidth and/or segregated transport will be provided on-demand to applications while simultaneously supporting traditional functions. Our testbed provides dynamic combinations of multiple dedicated and production links, switched out of a backbone of multiple OC192 links. It will support the research and development of various networking components, optimizing them for high-performance under real operational conditions with prototypical applications. The technologies and expertise developed using the testbed will be incrementally transitioned and integrated, as they mature, into the applications running on production networks. Our target network-intensive tasks include high throughput data transfers, interactive visualizations, remote steering, control, and coordination over wide-area networks.

Table of Contents

1. DOE Large-Scale Science Networking Requirements

The next generation supercomputers proposed for the DOE terascale computing facilities to support large-scale science computations promise speeds approaching 120 teraflops within the next few years (Cray's Black Widow by 2005) and offer the expectation of petaflop speeds shortly thereafter. They hold an enormous promise for meeting the demands of a number of large-scale scientific computations from fields as diverse as earth science, climate modeling, astrophysics, fusion energy science, molecular dynamics, nanoscale materials science, and genomics. These computations are expected to generate hundreds of petabytes of data at the DOE terascale computing facilities. This data must be transferred, visualized and mined by geographically distributed teams of scientists. It is extremely important for the network bandwidths to "match" these data generation speeds to make cost-effective use of the terascale computing facilities. At the same time, DOE currently operates or is preparing to operate several extremely valuable experimental facilities. These include BaBar at Stanford Linear Accelerator Center (SLAC), Atlas at the Relativistic Heavy Ion Collider (RHIC), and the Spallation Neutron Source (SNS). These experimental facilities will also generate petabyte data streams. The data transfer needs of the high energy community in particular have been documented in the TAN Report [RPT-1]. As the report shows, the ability to remotely perform experiments and then transfer the resulting large data sets so significantly enhances the productivity of the scientists working with these facilities that distributed analysis has been designed into their program plans from the very beginning. As was documented at the DOE High-Performance Networking Planning Workshop that [DOE-A02] data files to be generated by climate modelers (CCSM2) and the genomics program (particularly the Joint Genome Institute (JGI) at Walnut Creek) will soon overwhelm current networks.

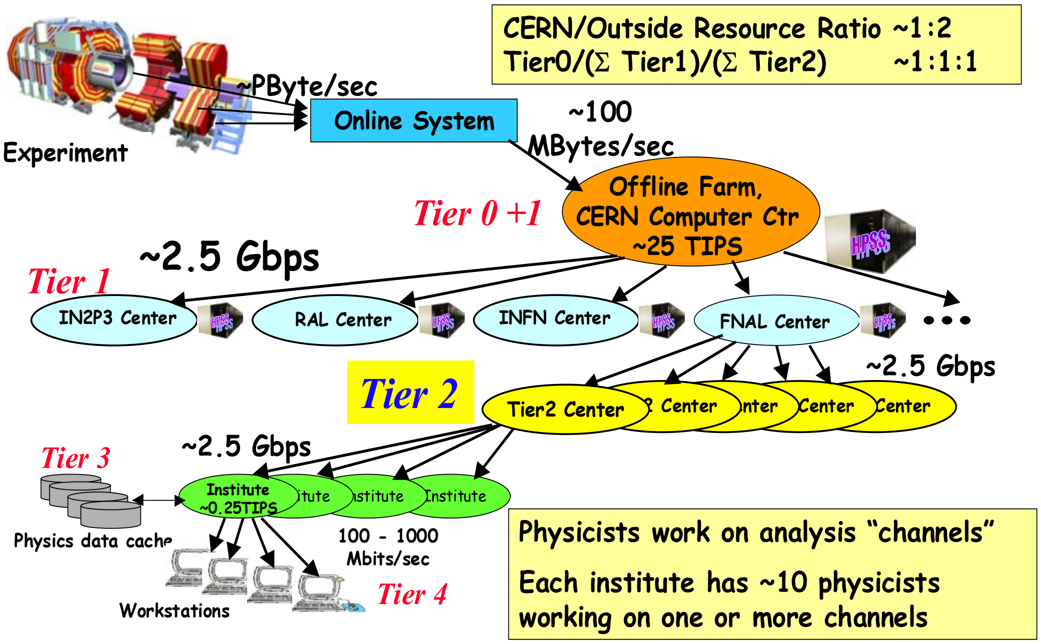

Figure 1. Expected data flow from LHC

The transfer of large data sets is not the only requirement of this next generation of computing and experimental facilities. Other requirements involve distributed collaborative visualization, remote computational steering, and remote instrument control. These requirements place different, possibly mutually exclusive, demands on the network. For example, the collaborative visualization of dynamic objects doesn't always need extraordinary amounts of bandwidth, and often 30-50 Mbps into a site is adequate (because it is an n-squared problem, it does place extraordinary demands on the backbone, but that is a different issue.) However, as we will show, it also requires a stable "low-inertia" message transport, which is incompatible with TCP-based networks. Thus the predicted data volumes for the next generation of DOE projects, together with the control requirements will demand quantum leaps in the functionality of current network infrastructures as well as networking technologies. In particular, the requirements of interactive visualization and remote instrument control add a whole new dimension to the required network capabilities, particularly so if they have to co-exist with high throughput transfers. There is no User-to-Network Interface (UNI) mechanism for even specifying the requirements on jitter, latency, or agility necessary for supporting remote instrument control, remote collaborative visualization, or remote computational steering. This is unfortunate because it is exactly these new capabilities that will be crucial to the next generation of high-impact large-scale science projects. The network metrics associated with these requirements are not well known, although estimates can be inferred. To close a human response-time feedback loop, the time constant for damping the control loop should be roughly 300ms or less. But, as is well known, TCP networks frequently deliver packets that are late by many, sometimes tens of seconds. The jitter due to late packets can and will destabilize a 300ms control loop. Similarly, to support a shared distributed visualization, the network has to have low inertia. The data stream has to "turn on a dime" in order to respond to the interrupt generated by a remote mouse movement. TCP's almost linear congestion response, which can take many seconds to reach full speed after an interruption, simply won't work.

SCIENCE AREAS	**TODAY** END2END THROUGHPUT	**5 YEARS** END2END THROUGHPUT	**5-10 YEARS** END2END THROUGHPUT	**REMARKS:** BASIC RESEARCH, TESTING AND DEPLOYMENT
High Energy Physics	0.5 Gbps E2E	100 Gbps E2e	1 Tbps	high throughput
Climate Data & Computations	0.5 Gbps E2E	160-200 Gbps	Tbps	high throughput
SNS NanoScience	does not exist	1Gbps steady state	Tbps & control channels	remote control & high throughput
Fusion Energy	500MB/min (Burst)	500MB/20sec (burst)	Tbps	time critical transport
Astrophysics	1TB/week	N*N multicast	1TB+ & stable streams	computational steering & collaborations
Genomics Data & Computations	1TB/day	100s users	Tbps & control channels	high throughput & steering

Table 1. Network requirements of DOE large-scale science applications.

High-performance networking capabilities add a whole new dimension to all DOE high-performance computing and experimental user facilities. They eliminate the “single location, single time zone" bottlenecks that otherwise plague these valuable resources. More generally, advances in high-performance networks hold an unprecedented potential for expanding the impact of a number of DOE large-scale science computations and experiments, conducted in a wide spectrum of disciplines. Such networking needs have been identified in DOE workshops and a series of Access Grid meetings [DOE-A02, A03]. Some specific numbers and examples from several disciplines are shown in Table 1 (see appendix for details in two specific areas), but in short, it is expected that speeds of 100 Gbps or more will be required within next five years and Tbps within the decade.

In summary, the challenge is that by 2005 DOE projects will require: 1, network throughput on the order of 160-200 Gbps for transferring high-priority petabyte data files on 24 hour time scales; 2, networks which support TCP-hostile protocols (low-jitter, for remote control, low-inertia for visualization, and high-throughput for petabyte file transfer); 3, roughly 40 Gbps of backbone production capacity for supporting high-impact science programs.

Dynamically provisioning of 3-4 link-layer OC768 circuits that parallel the production backbone can provide these capabilities. For example, a "clear channel" link dedicated to a UDP-based control protocol can deliver the needed capabilities, while at the same time allowing normal production traffic to proceed uninterrupted. The problem is that to deliver on the promise of switched link-layer circuits, the logistics of brokering, switching, and delivering bandwidth to users, services and file systems has to be developed and tested prior to deployment. This proposal describes the UltraNet which constitutes an initial implementation of a testbed network that will support these research, development, and testing activities.

ESnet Characteristics

- **POS/IP-based – Enhance Internet**
- **2.5 Gbps/10 Gbps Backbone**
- **Best Effort traffic – No QoS**
- **500 Mbps Max end-to-end throughput**

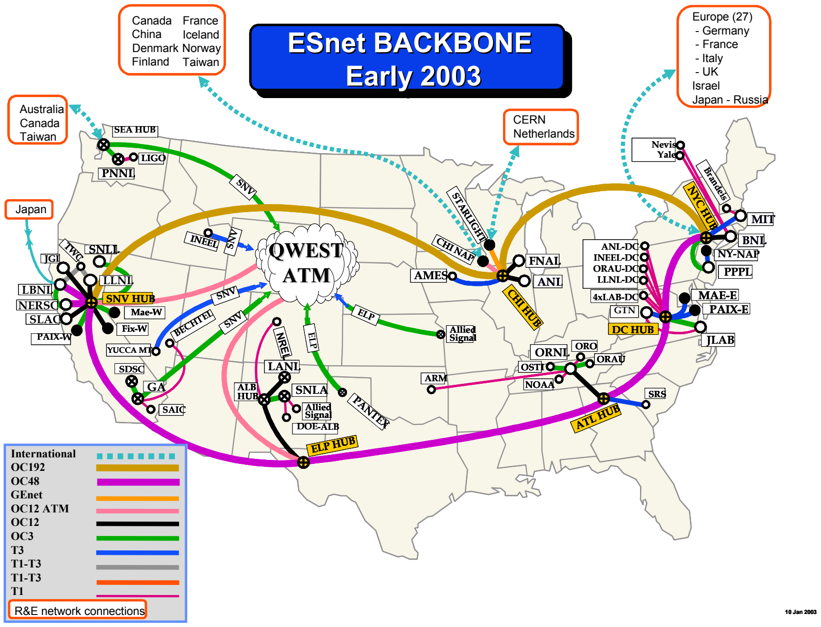

Figure 2. ESnet current backbone map and architecture

2. Current State of Networking

The newer challenges of DOE large-scale science applications require capabilities that far transcend its production network capabilities. Consequently, the next generation network demands are simply beyond the capabilities of ESnet (shown in Figure 2) both in terms of the required large bandwidths and the sophistication of the capabilities. First, there is no provision in ESnet for testing Gbps dedicated cross-country connections with dynamic switching capability. Second, during the technology development process, it is quite possible for various components of the network to be unavailable for production operations; such situations cause undue disruptions for normal ESnet activities. Therefore, as the recent series of DOE Roadmap workshops has indicated, the future of DOE networking lies in creating high-impact and research networks. This proposal can be viewed as the foundational step towards building such a research network envisioned in this roadmap.

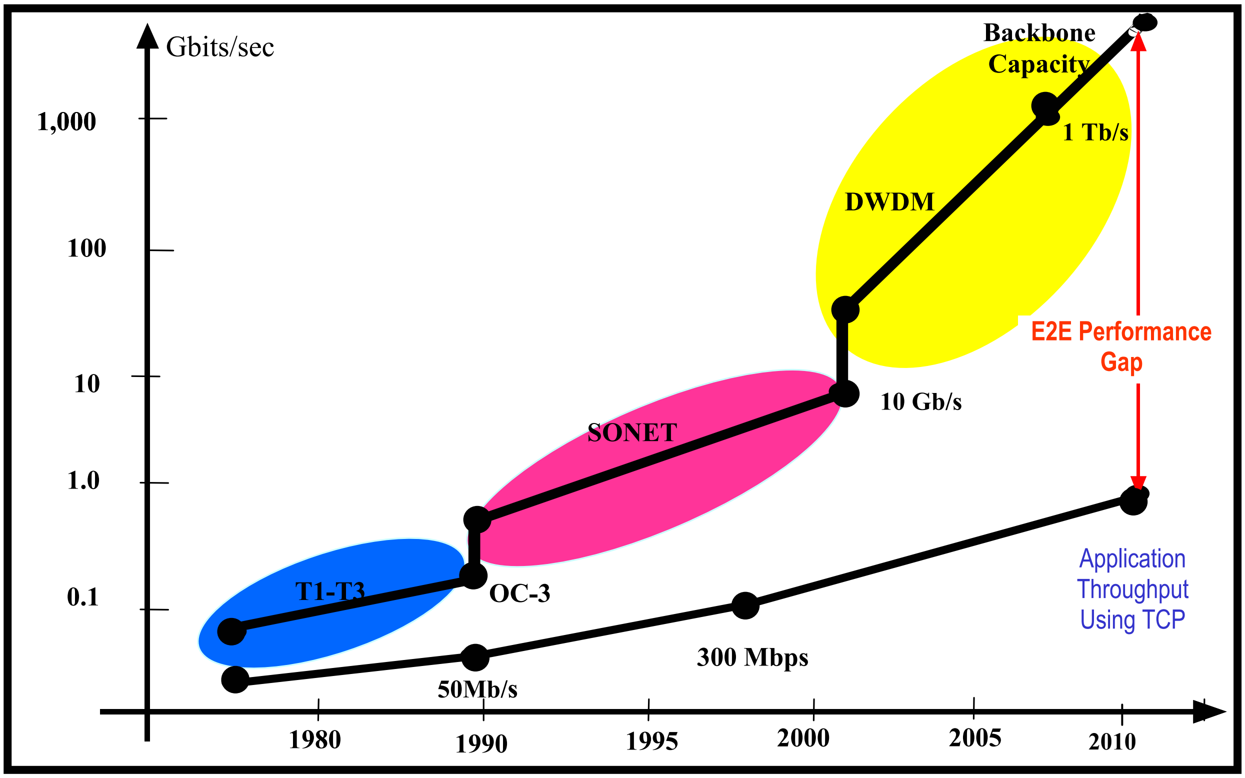

Figure 3. Performance gap between application throughput and optical network speeds

The field of ultra high-speed networking is currently at a critical crossroads with no clear evolutionary path to eliminate the ever-widening performance gap between link speeds and application throughputs indicated in Figure 3. On one hand, optical technologies promise lambda-switched links at Tbps rates but offer no corresponding provisioning and transport technologies to deliver this performance to applications. On the other hand, legacy protocols, including the most widely deployed transport protocols, namely Transmission Control Protocol (TCP), and other network components (that are optimized for low network speeds) cannot easily scale to the unprecedented optical link bandwidths. A comprehensive solution to achieving the end-to-end Tbps application throughputs must address the following end-to-end issues:

a. **Backbone Networks** - Today's core network is a static configuration of several layers of transport network technologies involving IP, ATM, MPLS, SONET, and DWDM technologies shown in Figure 4. Each layer of technology adds a degree of complexity to the overall backbone network, often limiting the capability that can be pushed to the applications. UltraNet will explore innovative scalable architectural options that use a minimum number of layers, primarily options (d) and (e) in Figure 4, that make wavelengths available directly to the applications.

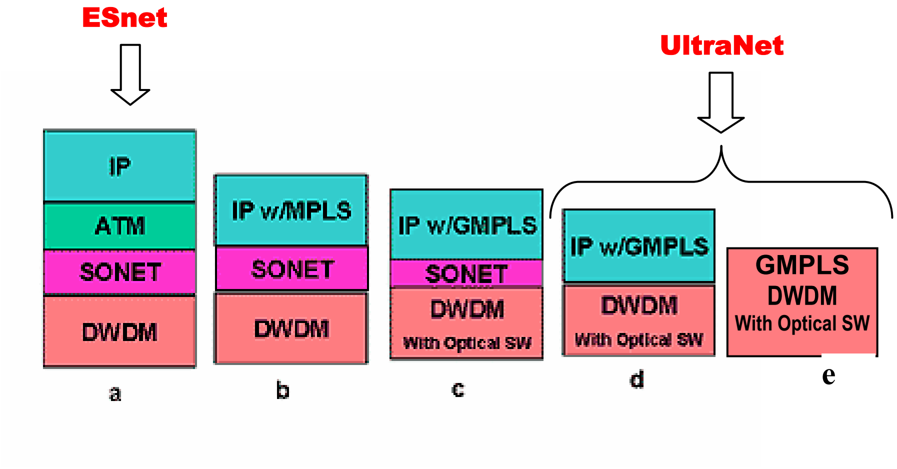

Figure 4. Structure of network layers.

b. **Transport Protocols** - . TCP was designed and optimized for low-speed data transfers over congested IP-based networks. It has many appealing features, such as its ability to deal with congestion and attain certain levels of fairness. These features have contributed to its longevity in different networking environments including wireless, satellite, and optical networks. However, its effectiveness in ultra high-speed networks based on the emerging all-optical networks is being seriously questioned, especially in the transfer of petabytes data over inter-continental distances. Today's IP networks expect that TCP will provide reliable transport mechanism for higher layer services. Exploring radical enhancements to TCP or more general alternatives to it, did not result in the major (for example 1000x) performance improvements needed to harness the abundant bandwidth in the optical core network. UltraNet will provide a rich environment to explore high-performance transport protocols that will achieve throughputs of the order of available capacity in the optical core networks.

c. **Traffic Engineering** – In IP best-effort networks, it is very challenging to deal with the congestion in an effective and fair manner. This problem is further exacerbated by IP routing which has convergence problems in large networks. Attempts to address these issues have generated a wide range of network technologies, notably Quality of Service (QoS) and recently Multi-Protocol Label Switching (MPLS). The corresponding IP traffic engineering methods can potentially steer traffic away from the congested parts of the network. MPLS has recently been extended to IP-based DWDM networks to take advantage of the optical bandwidths to address congestion problem in the IP layer. Unfortunately, the required advanced traffic engineering methods have not been widely deployed in operational networks because they involve complex inter-domain signaling and costing. UltraNet will provide an excellent environment to prototype the needed practical traffic engineering methods within the context of DOE networking environments.

d. **Host Systems** – The components of an end system, such as the transport protocol stack, Network Interface Cards (NICs), operating System, I/O sub systems, and science application modules, all affect the end-to-end network performance. They have many design and architectural aspects that significantly limit the abundant optical capacity from reaching the science applications. Typically, these end system components have been designed and optimized for low speed Internet applications and do not easily scale to accommodate the ultra high-speed links. UltraNet will offer a rich research environment to design and prototype alternatives that can eliminate the host bottlenecks.

e. **Deployable QoS**– Differentiated services in packet networks constitute an area that has received significant attention in the research communities during the last decade. Yet today, it is still unavailable in most production networks, including ESnet. Many science applications with real-time constraints will depend on this critical technology or effective alternatives to it. Differentiated traffic at the packet level is a difficult capability to achieve, especially at ultra high-speeds. There are also several organizational and economic factors that have prevented the large-scale deployment of such technologies. UltraNet will develop alternative solutions to the QoS problem by developing both coarse-grain and fine-grain QoS mechanisms using a combination of dynamic provisioning and GMPLS to provide differentiated services at the wavelength and sub-wavelength levels, respectively.

This proposal constitutes the foundation of a comprehensive effort to develop the infrastructure and networking technologies required to support the needs of DOE large-scale science applications. There have been two major shortcomings in the previous efforts to develop high-performance network capabilities.

- First, there have been no testbeds that provide adequate operating conditions in terms of bandwidths, distances and traffic levels. Historically, methods based on simulations and small-scale testbeds often resulted in technologies which fell short of the needs.
- Second, the adoption of research tools by the users has been highly limited due to the lack of a natural transition path. Tools developed by network researchers often require a significant amount of integration before they can be used by non-experts, which often results in under deployment in the field.

This proposal addresses both these issues by providing a high-performance development and testing environment as well as a smooth transition path for the technologies from research stages all the way to applications running on production networks.

3. Office of Science Networking Roadmap

In anticipation of major scientific computations and experiments requiring unprecedented network capabilities (beyond the projected capabilities of ESnet), the Office of Science conducted a series of workshops to identify the needs and plans for the next generation wide-area networks to meet the demands of DOE large-scale science applications. The DOE planning workshop [DOE-A02] of August 2002 identified a number of science areas with high-performance networking needs. The follow-up workshop of April 2003 [DOE-A03] focused on the specific areas of network provisioning and transport to address the DOE large-scale science networking needs. In the DOE Science Network Workshop [DOE-AJ03.1], a roadmap to 2008 has been laid out to enhance the current production infrastructure to meet these next generation demands both in terms of production, high-impact and research networks. Again, at DOE Science Computing Conference [DOE-J03.2] the highly demanding nature of the high-performance networking needs for large-scale science applications has been asserted. A recurring theme among the conclusions of these workshops has been the need for a wide-area network testbed with adequate capabilities and support environments.

This proposal is a concrete step towards the development of the testbed that was deemed essential to meeting the challenges of these large-scale science applications. A roadmap of network infrastructure has been formulated as a result of extensive discussions at these workshops, which consists of three networks, namely the production network, a high-impact science network and a research network as shown in Figure 5. The production network provides the traditional network services that are akin to the Internet. The high-impact science network provides the infrastructure for applications requiring high bandwidth, low latency and jitter. The research network enables the development of the required network technologies on wide-area networks running actual applications. The main idea is to provide natural transition paths for network technologies from the development stages to testing then to production networks. The UltraNet proposed in this proposal has been designed as a foundation to meet the requirements of the research network identified by the Office of Science networking roadmap.

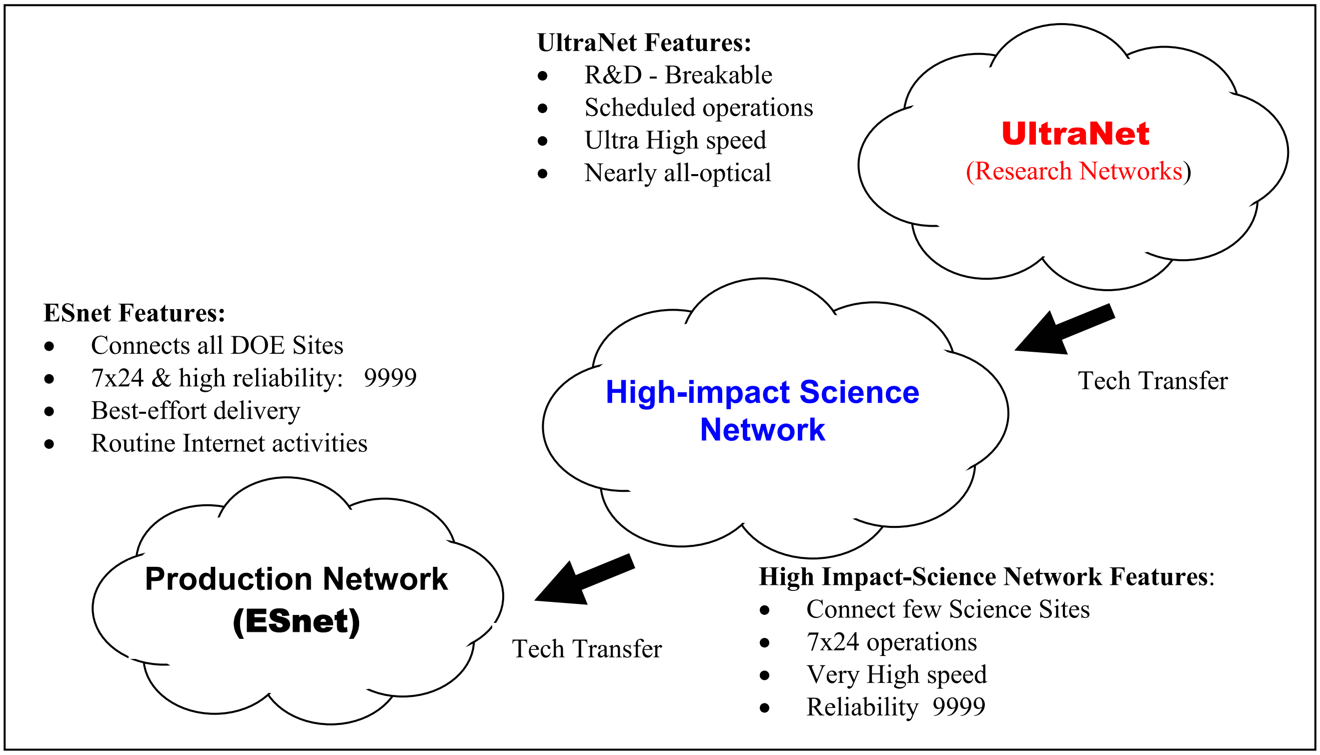

Figure 5. Defining features of production network, high-impact network, and research UltraNet

These DOE workshops and the subsequent plans have been cognizant of and are complementary to related efforts from other federal agencies. Analogous activities but typically with a different focus have been taking place within the National Science Foundation (NSF) over the past few years. The 2001 workshop [NSF-01] on e-Science grand challenges identified the cyber-infrastructure requirements, which included networking technologies, to address the nation's science and engineering needs. The scope of this workshop was broader than this proposal both in terms of class of applications as well as infrastructure areas. The two workshops on testbeds [NSF-02.2] and infostructure [NSF-02.3] specifically dealt with developing networks with capabilities beyond the current ones. Both these workshops focused on technical issues that are broad but not specific enough to encompass DOE large-scale science needs. Several of the high-performance network capabilities could be enabled by optical networking technologies, and the NSF workshop [NSF-02.1] on this topic is narrower in terms of the technologies considered but is broader in terms of the network capabilities.

4. The Proposed Science UltraNet

The networking research areas to be investigated for DOE large-scale science projects are quite specific but wide-ranging, from TCP and non-TCP transport methods capable of sustained throughput at 40Gbps over long-haul networks to cyber security methods capable of operating at extremely high speeds using dedicated paths. Due to the extreme networking requirements, the research and development of the needed capabilities are beyond the scope of existing testbeds, analytical methods and simulation tools. Indeed, several of the required advanced capabilities can only be adequately developed on powerful research networks. The main challenge in building the testbed is to provide realistic test conditions together with robust and stable infrastructure support, which includes long-haul high bandwidth circuits with flexible provisioning and development environments. We note, however, that not all general network research areas absolutely require UltraNet, for example, the development of a new science of network transport based on non-linear stochastic systems. On the other hand, not every UltraNet activity is a basic network R&D activity, for example, interfacing a latest-release application software with transport middleware.

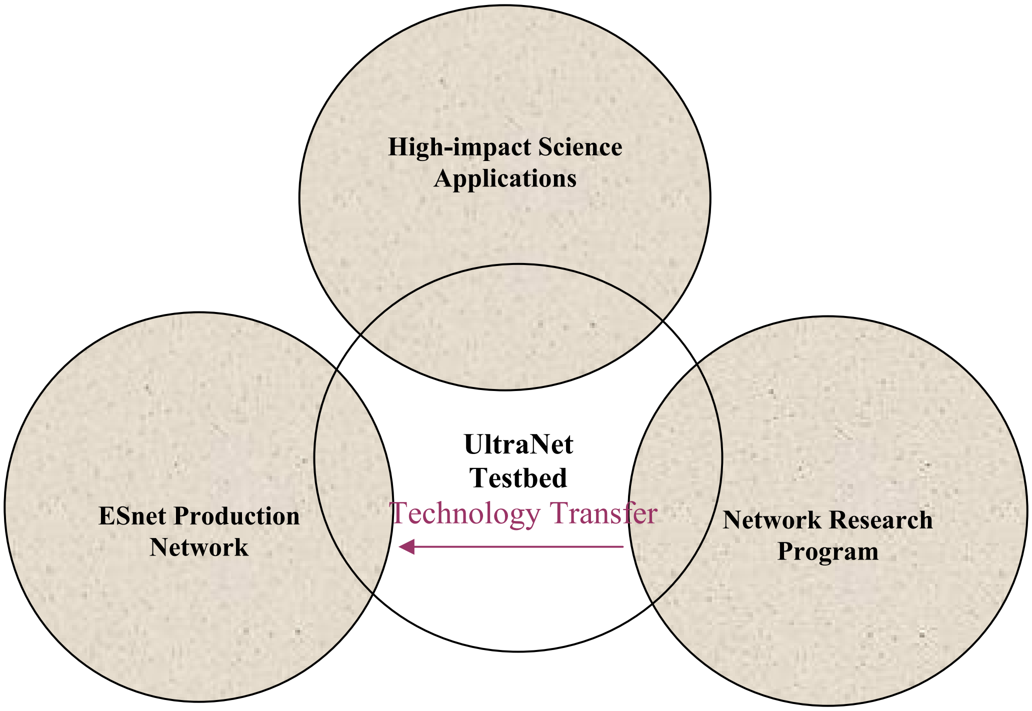

Figure 6: Operational Space of UltraNet in the context of other MICS activities

One of the major objectives of the UltraNet is to develop and test on-demand dynamic provisioning network technologies. Since the Office of Science research environment consists of applications with different network requirements, a single network infrastructure that meets different and sometime conflicting network requirements is a challenging task. We address this problem with a single agile network infrastructure that can be dynamically reconfigured to different network capabilities to match various large-scale science applications. As a consequence, UltraNet will be configured to operate in two basic operating modes, namely packet and circuit switching, as well as their hybrid combination modes. In each mode, the technologies for wavelength scheduling, sharing, and reservation will be developed.

The proposed UltraNet is an integrated network testbed that can deliver production and advanced networking services to high-impact science applications, and experimental network services to support network research and development activities. The research network will be used to design, implement, and test advanced capabilities, and therefore, will need adequate hardware and software components, as well as a flexible development environment in order to ensure their smooth transitions to production networks. Relevant research topics include reliable transport protocols for high sustainable throughput, OS bypass mechanisms to eliminate memory bottlenecks, cyber security operations at 40Gbps or higher rates, guaranteed services to deliver real time capabilities, and end-to-end monitoring and diagnosis to optimize the usage of network-wide resources. Figure 6 illustrates the relationship of UltraNet to the production network, the network research program, and the fact that it will be a vehicle for transferring technology from the research programs to applications running on the production network.

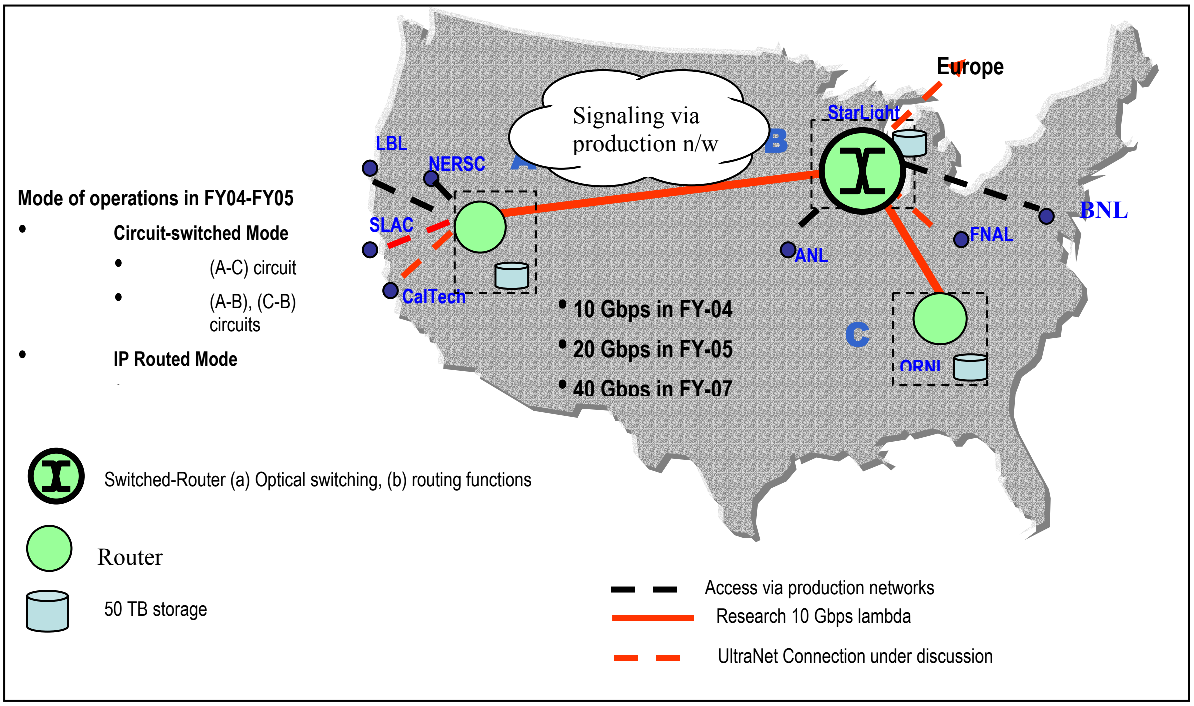

Figure 7. Testbed progression from 10 Gbps to 40 Gbps .

4.1. Phases of Deployment

UltraNet will be deployed in three phases. The first phase, shown in Figure 7, consists of a 10 Gbps DWDM backbone network linking two major national and international POPs, StarLight in Chicago and Sunnyvale in California. The Chicago pop will facilitate the connectivity to Argonne National Laboratory (ANL), Fermi National Laboratory (FNL), Brookhaven National Laboratory (BNL)and the international connectivity to CERN. The Sunnyvale POP will facilitate the connectivity to Lawrence Berkeley National Laboratory (LBNL), Stanford Linear Accelerator Center, and National Energy Supercomputer Center (NERSC). The planned connectivity to the individual national laboratories will be their responsibility. The POPs will be equipped with 50 TB disk storage systems to support the testing of ultra high-speed transport protocols. The initial deployment will cover a distance of 4,500 miles and will therefore include the major physical characteristics of a national scale network, a feature that will be critical in testing the behavior of ultra high-speed transport protocols over long distances.

4.1.1. Phase I - Initial Deployment of 10 Gbps Provisioned Network

During the first year of operation, the research network will be implemented as a pair of OC192 links, one from ORNL to Starlight in Chicago and one from Starlight to Sunnyvale, CA as in Figure 7. We deploy an optical switch at Chicago and SONET Add-Drop Multiplexers (ADM's) at all three sites to perform a number of switching functions. To facilitate testing of the circuit-switching environment as early as possible, these links will be explicitly configured as OC192, rather than OC192c links. This will enable them to be treated as a single OC192c link or as four parallel OC48 links when desired. This trick is enabled by the fact that router interfaces that are able to terminate *either* an OC192 or an OC192c link.

The optical switch provides us the ability to implement three OC192 connections, namely, ORNL-Chicago, Chicago-Sunnyvale, and ORNL-Sunnyvale. In addition, we can also test on-demand provisioning at much finer granularity. A Programmable ADM (PADM), also referred to as a Multi Service Provisioning Platform (MSPP), can easily break out and switch four OC48 links or a larger number of lower bandwidth links such as gigabit Ethernet or fiber channel, a function we expect to use to provide direct switched access to storage. A PADM also provides a User-to-Network Interface that allows this switching to be performed under software control. Thus, for example, the switch in Chicago could be configured to pass two of the links straight through from Sunnyvale to ORNL, and break out the other two links to router interfaces at either Argonne or Fermi Laboratory. This would create a triangular network that could be used to test express or dynamic routing. Alternatively, the entire bandwidth could be switched on-demand and made available to an application at SLAC for moving data to CERN or Fermi Laboratory.

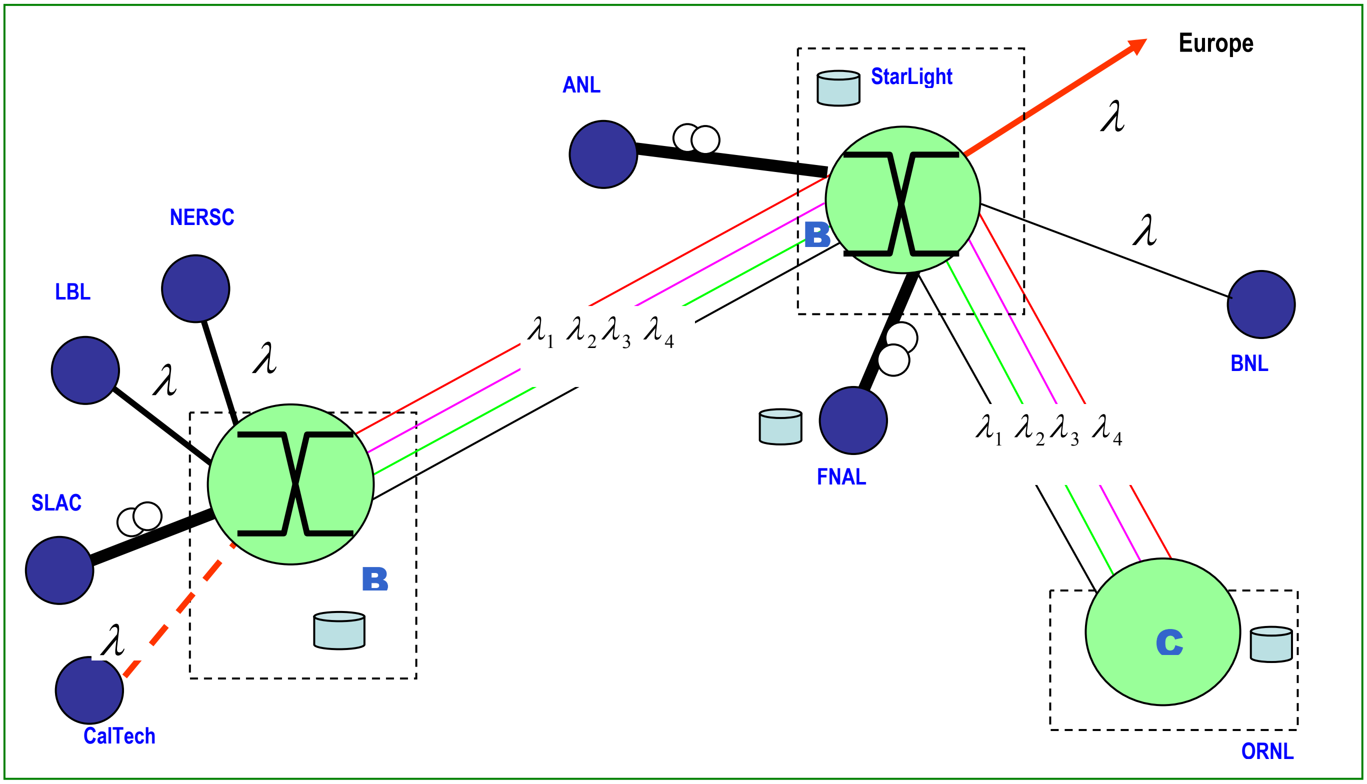
Europe
ANL
StarLight
λ
NERSC
B
λ
BNL
LBL
λ
λ
$\lambda_1\ \lambda_2\lambda_3\ \lambda_4$
BNL
SLAC
FNAL
$\lambda_1\ \lambda_2\lambda_3\ \lambda_4$
B
λ
C
CalTech
ORNL

Figure 8. Configuration of UltraNet in Phase III

4.1.2. Phases II and III - 40 Gbps Integrated UltraNet Testbed

UltraNet will be upgraded to support multi-lambda operation in subsequent phases. Based on the availability of funds, UltraNet will be upgraded to operate at 20Gbps and 40 Gbps in Phase II in FY05 and in Phase III in FY07, respectively., In the final phase, UltraNet, will consist of OC-768 (40 Gbps) lambda-switched shown in Figure 8. The major differences between phase I and the later phases are the increased capacity, connectivity to the major sites, and the incorporation of all-optical networking technologies. It is expected that SLAC, ANL, and Fermi National Labs will have dark fiber links to Sunnyvale and StarLight respectively. This will give these Labs the ability to connect to Sunnyvale and StarLight with several wavelengths. Other Labs, LBL, LBL, NERCS, BNL will be connected to the testbed at 10 Gbps.

Some additional capabilities and features of the network are only hinted at by the drawing in Figures 7 and 8. For example, with the switches providing isolation between the provisioned channels, we can segregate TCP-hostile protocols from TCP traffic by using different dedicated lambdas or using sub-lambda provisioning within a single lambda. As another example, it could provide a direct, dedicated high-bandwidth visualization channel between NERSC and ORNL.

UltraNet is conceived to be an agile infrastructure that can be dynamically reconfigured to produce different network scenarios needed to test different advanced networking technologies and concepts. The network will operate in two distinct switching modes, each corresponding to a transport mode and the corresponding underling layer-two technologies, and also their hybrid combinations. These include:

- **Packet-Switched Operations:** Using UltraNet the packet switched traffic can be supported in a straight forward manner by connecting hosts with GigE NICs directly into a PADM. In this mode the GigE host traffic will be aggregated and sent out on OC192 or OC48 output line cards.

- **Circuit-Switched Operations:** Although Figure 7 shows single connections on each switch, this is my no means the limit**.** By providing more (as many as four) local ports on each switch, we will enable switch selection of the host or service to be connected to the wide area part of UltraNet.

- **Hybrid Operations:** In Phase I, by utilizing one through three OC48 connections out of OC192 in circuit switched mode and others in packet-switched mode, we will realize hybrid connections. In phase II, entire lambdas can be dedicated to individual types of traffic as shown for examples in Figure 9.

- **StarLight:** We expect to locate the Chicago switch at Starlight (710 North Lakeshore Drive in downtown Chicago). This is a carrier-neutral meet-me location at which connections are available to ANL, Europe (CERN) and (soon) to FNL.

- **Transport Protocol Multiplexing:** Using the provisioned circuits, protocols will be multiplexed across various hybrid connections as well as on single connections**.**

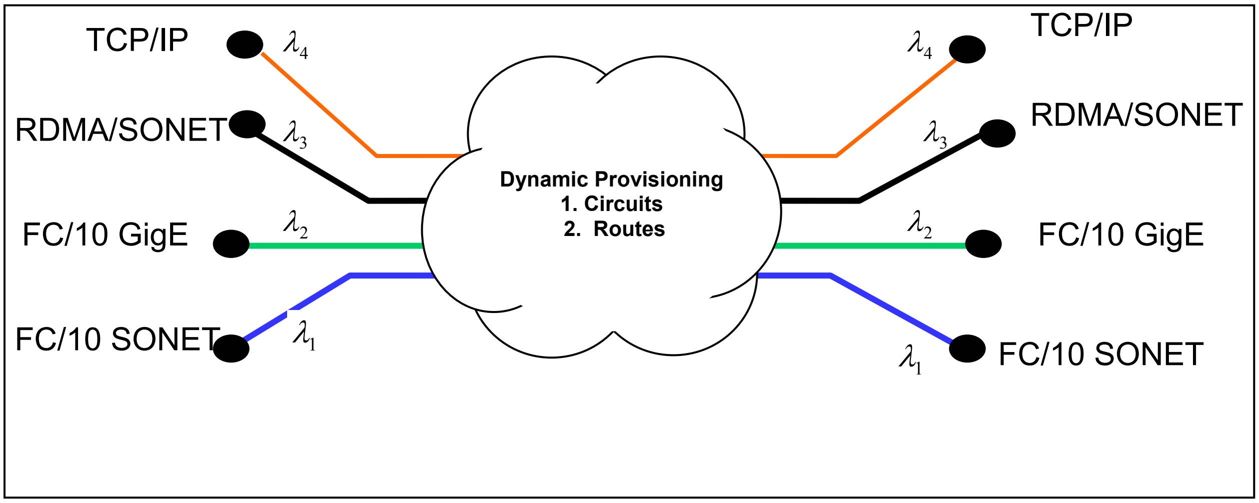

*Figure 9. Multiple traffic types on different wavelengths***.**

In the later phases of UltraNet there will be multiple wavelengths between ORNL and Chicago, and Chicago and Sunnyvale. With several laboratories connected to Sunnyvale and StarLight with one or more wavelengths, most of the dynamic provisioning services between them can be accomplished through optical switching. As an example, all four wavelengths can be used for bulk data transfers between Sunnyvale and ORNL by placing the Chicago switch in the pass-through configuration. Then a high-priority interactive visualization data stream from ORNL to SLAC (for example) could use the entire bandwidth of the switched Sunnyvale-ORNL link, while bulk transfers could use the other three wavelengths. Such transitions can be achieved on-demand by signaling the switches at Chicago and Sunnyvale and the ADM at ORNL. More generally, multiple dedicated wavelengths can be provisioned between various sites on-demand in groups to support various types of traffic as shown in Figure 10. Here we configure the specialized connections of the network into two example cases (visualization and ultra high-speed data transfer) in addition to routine production use. With four lambdas available, production capacity can assume “ownership” of all four as shown in mode one (this is the mode the National Science Foundation’s Distributed Terascale Facility will operate in). But under special circumstances, one or two lambdas could be assigned to applications that need to use TCP-hostile protocols such as those based on UDP. These configurations provide coarse-level QoS where the channels are provides at the resolution of single wavelength.

By combining wavelength-level optical switching with PADMs at the host nodes, we propose to achieve much finer QoS at the levels of sub-wavelength. In these modes, PADMs can partition the outgoing OC192 bandwidth into channels of much smaller resolutions such as GigE, OC48, or much smaller bandwidth. Furthermore, such provisioning can be accomplished for on-demand provisioning of channels at sub-second timeframes. Accomplishing on-demand provisioning of dedicated connections at sub-wavelength level requires capabilities that are unprecedented in current IP networks. First, the signaling infrastructure needed for setting up and tearing done the required connections must be established as an integral part of the network. We propose to utilize the production network for this purpose. Second, we propose to develop a scheduling system to: (a) keep track of the available link bandwidths, and (b) receive and grant the connection requests from the applications. To minimize the cost of development, much of the bandwidth brokering software would be adapted from the DCS scheduling software developed under DOE projects.

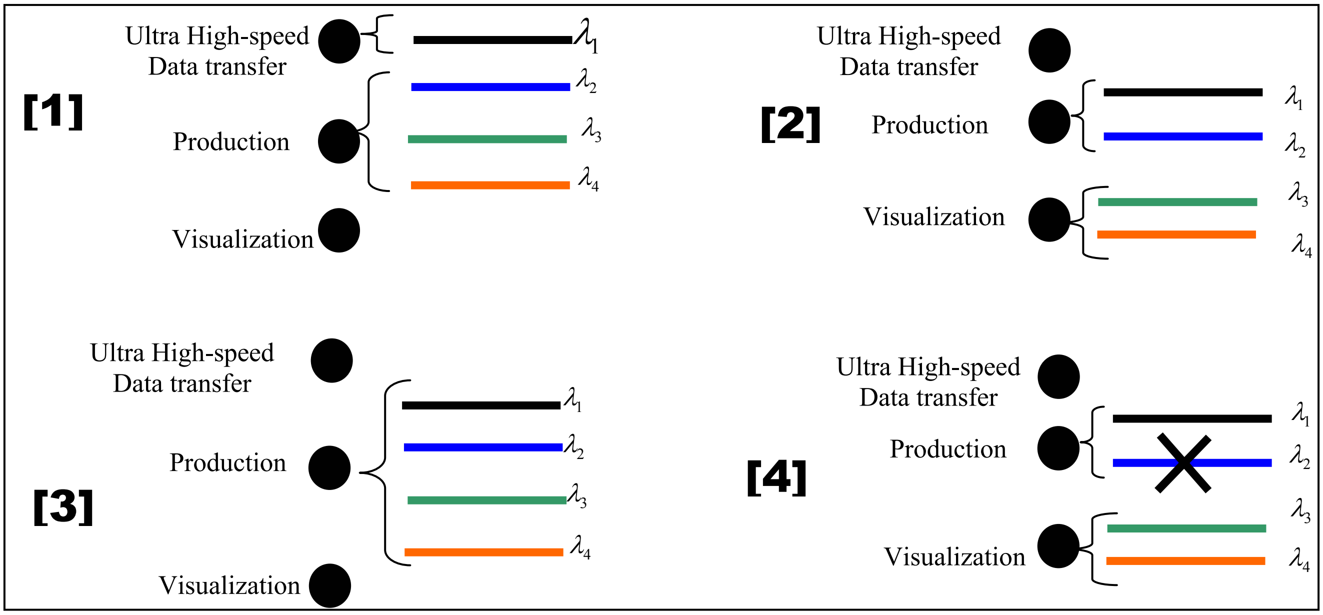

Figure 10. Using dynamic provisioning to realize coarse-grain QOS services.

Although the cost of link-level access to DWDM bandwidth is becoming significantly cheaper, it is still a non-trivial expense. For example, to build a single, unprotected, 10G circuit from the West coast to Chicago using the most aggressive public pricing today would cost ~$1M/yr by utilizing existing companies. Building four parallel OC48 circuits (same total bandwidth) would cost ~$2.5M/yr. With the planned deployment of nationwide optical network infrastructures, the current expectation is that much cheaper prices could be possible for the 10Gbps connections. We propose to pursue the most cost-effective options based on the available service providers at various stages of this project.

4.2. Relationship to other MICS Programs

The UltraNet will be an important vehicle for the development of network research and middleware technologies for DOE large-scale science applications. While it interfaces with various other programs from DOE Division of Mathematics, Information and Computer Science (MICS), it plays an extremely important role that is distinct from the network research, middleware, applications and ESnet programs.

- **UltraNet and Network/Middleware Research Programs:** UltraNet provides a test and development environment for various network research and middleware projects to account for real high-performance networking environments. Note however that there are other areas in these programs that are not necessarily tied to UltraNet such as the development of a science of network transport.
- **UltraNet and Applications:** Several science applications will be executed on UltraNet by utilizing various research modules during the development. Indeed, the network and middleware technologies will be developed within the context of applications. As the technologies mature after an extensive testing phase, they will be transitioned to applications running under production network environments.

- **UltraNet and ESnet:** The UltraNet is separate from the current ESnet. In the road-map for the next generation of ESnet, a research network is planned to be an integral part of the infrastructure [DOE-J03.1]. UltraNet will be a foundation to that research network which will be built by augmenting the former with additional links, routers, switches and end hosts.
- **UltraNet and Office of Science Networking Roadmap:** A research network testbed has been identified as an essential component of Office of Science Networking roadmap needed to meet DOE science demands [DOE-J03.2]. UltraNet has been designed to match these requirements and to be a strong contributor to this roadmap**.**

4.3. Network Operations and Management

The research network will be governed by a committee in deciding the access policies for various institutions and projects at a high-level as well as for various users and applications on a daily operational level. In addition, there are lower level allocation issues due to the on-demand provisioning aspects as well as the possibility of network becoming unavailable for certain durations as result of experimentation and testing. Note that applications could request dedicated circuits or stable but shared connections at certain times. On the other hand, certain network research projects could push the network limits, possibly crashing the routers and/or hosts. These tasks will be scheduled on a demand basis. Furthermore, this committee will also setup and decide the security issues. The connections to other networks, such as EU and NSF networks, will also be decided by this committee. The operational policies will be implemented by a scheduling system that will provide on-demand allocation of circuits, and will also activate the signaling modules to setup and tear down the required paths. These policies will be updated in view of completed and upcoming projects.

5. Collaborations and Applications

The proposed testbed will be flexible to easily "plug-in" the data and middleware from SciDAC and other DOE large-scale science applications to asses the performance improvements offered by various ultra high-speed transport methods. For initial experimentation, we will utilize data from SciDAC Supernova Terascale Initiative in collaboration with Mezzacappa (PI of TSI). Computations on ORNL computers will be used to generate tera-scale data, which will be used for storage, remote visualization and computational steering. The logistical networking framework of Beck (PI on SciDAC and NSF projects on logistical networks) will be utilized for efficiently realizing the needed data transfers by suitably buffering the data at the Chicago data depot. This initial data depot will be augmented with others if Beck wishes to add them. Climate data produced at ORNL will also be utilized for experimentation in collaboration with Burris (in charge of ORNL storage activities). We wish to emphasize that the transport methods will be general and will be equally applicable to other large-scale science applications from HENP and molecular dynamics computations. We will be working closely with various SciDAC and other applications communities with ultra high-speed networking needs to identify the network configurations that meet these requirements.

The transport component of this project significantly leverages a number of ongoing projects at ORNL funded by DOE High-Performance Networking Program, NSF Advanced Infrastructure Initiative and DARPA Network Modeling and Simulation Program. The net100 package, including buffer turning and parallel-TCP modules from the first project will be utilized in developing the transport methods. The remote visualization methods developed under the NSF project will be extended to the proposed high-performance network configuration. The throughput stabilization methods used to control mobile robots remotely under the DARPA project will be extended to support computational steering and remote interactive visualization. The methods from NSF and DARPA projects are basically targeted to Internet environments, and they must be suitably extended or re-designed to suit ultra high-speed networks needed for DOE large-scale science applications.

6. Statement of Work

The main components of this proposal include (a) establishing the testbed in its various configurations, (b) developing the dynamic provisioning capability, (c) adapting various transport and middleware technologies, and (d) testing applications.

Year 1 Goal - To Deploy and configure a 10 Gbps SONET/DWDM channel between ORNL and StarLight in Chicago and between StarLight and Sunnyvale California. Several research projects yet to be funded in FY04 will use the UltraNet for R&R activities

I. Testbed Milestones:

▪ Initiate procurement of ORNL Chicago Link	Sept	2003
▪ Place order for circuit from ORNL to Chicago:	Nov	2003
▪ Purchase needed equipment:	Nov	2003
▪ Install equipment and initiate testing	Dec	2003
▪ Install logistical networking depot	Jan	2004
▪ Install lab-based circuit switches	Jan	2004
▪ Integrate software with lab-based set-up	July	2004

Year-one costs:

▪ Contract for OC192 Circuit ORNL-Chicago	$600k
▪ Network components	$200k
▪ 1 FTE	$200k

II. Engineering, Research and Applications Goals

1. Transport Protocol Testing (with a goal of demonstrating a 10TB transfer)
2. Dynamic Channel Provisioning (demonstrate coarse-grained QoS)
3. Application and Middleware Research Activities (demonstrate stable remote visualization)

Year2: Goal -_Establish Phase II of the testbed with active optical and an additional lambda. Test the dynamic provisioning technologies locally with router-switch configurations; test transport and middleware under routed configurations that simulate various connectivity structures to support collection of lambda and sub-lambda circuits.

I. Testbed Milestones:

▪ Procure second lambda	Oct 2004
▪ Place order for second through fourth all-optical switches	Dec 2004
▪ Procure additional network components for west-coast connection	Aug 2004
▪ Demonstrate lambda-agile switching	Feb 2005

Year-two costs:

▪ Contract for second 10 Gbps Circuit	$600k
▪ Network components	850k
▪ 1 FTE	300k

II. Engineering, Research and Applications Goals

- Transport Protocol Testing (demonstrate 20 TB transfer over 10 Gbps in 3,000 miles)
- Dynamic Channel Provisioning (demonstrate channel scheduling at network level through switched paths)
- Applications Activities (demonstrate synchronizing climate data)

Year 3: Goal - Install additional lambdas and integrate the testbed fully with production traffic; implement the dynamic provisioning on the testbed; test transport and middleware under the dynamic provisioning over the testbed; test two large-scale science applications over the testbed as well as the combination of testbed and production network.

I. Testbed Milestones:

▪ Integrate automated call setup with PADM and Optical Switch	Sep 2005
▪ Migrate Router-based switching to Circuit-based	Dec 2005
▪ Migrate demonstration applications to circuit switched operation	Dec 2005
▪ Initiate procurement of additional lambdas	Dec 2005

Year-three costs:

▪ Additional lambdas	$1,200k
▪ Network components	250k
▪ 1 FTE	300k

II. Engineering, Research and Application Goals

- Transport Protocol Testing (demo 40 TB transfer over 30+ Gbps in 3,000 miles)
- Dynamic Channel Provisioning (demo channel scheduling for concurrent data and control streams)
- Application Activities (demo load-balanced transfers between ORNL and Sunnyvale)

7.0 Budget

The cost of this project is $1071K in the first year, $1753K in the second year, and $1758K in the third year (for FY2003 through 2005). Cost for each year includes the OC192 links from ORNL to Chicago and also Chicago to Sunnyvale including fees to the service providers and the associated routers and switches. In each year, one post-doctoral fellow at full-time and two staff members at half-time will be supported at ORNL. One staff member will be in charge of the testbed hardware setup and configuration, and the other will be involved in software development for dynamic provisioning and adaptation and testing of various transport modules, middleware and application modules. The post-doctoral fellow will be the involved in both aspects and will closely work with both staff members.

DOE F 4620.1
(04-93)
All Other Editions Are Obsolete

U.S. Department of Energy

Budget Page

(See reverse for Instructions)

OMB Control No.
1910-1400
OMB Burden Disclosure
Statement on Reverse

ORGANIZATION Oak Ridge National Laboratory (Year 1)				**Budget Page No:**	Year 1
PRINCIPAL INVESTIGATOR/PROJECT DIRECTOR				Requested Duration:	12 (Months)
A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title; A.6. show number in brackets)	DOE Funded Person-mos.			Funds Requested by Applpicant	Funds Granted by DOE
	CAL	ACAD	SUMR		
1. Wing, William	3.3				
2. Rao, Nageswara	3.3				
3.					
4.					
5.					
6. ( ) OTHERS (LIST INDIVIDUALLY ON BUDGET EXPLANATION PAGE)					
7. ( ) TOTAL SENIOR PERSONNEL (1-6)	6.5			$82,594	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. ( ) POST DOCTORAL ASSOCIATES					
2. ( ) OTHER PROFESSIONAL (TECHNICIAN, PROGRAMMER, ETC.)					
3. ( ) GRADUATE STUDENTS					
4. ( ) UNDERGRADUATE STUDENTS					
5. ( ) SECRETARIAL - CLERICAL					
6. ( ) OTHER					
TOTAL SALARIES AND WAGES (A+B)				$82,594	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				$28,990	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A+B+C)				$111,584	
D. PERMANENT EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM.)					
TOTAL PERMANENT EQUIPMENT				$220,000	
E. TRAVEL 1. DOMESTIC (INCL. CANADA AND U.S. POSSESSIONS)				$18,000	
2. FOREIGN					
TOTAL TRAVEL				$18,000	
F. TRAINEE/PARTICIPANT COSTS					
1. STIPENDS (Itemize levels, types + totals on budget justification page)					
2. TUITION & FEES					
3. TRAINEE TRAVEL					
4. OTHER (fully explain on justification page)					
TOTAL PARTICIPANTS ( ) TOTAL COST					
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES					
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION					
3. CONSULTANT SERVICES					
4. COMPUTER (ADPE) SERVICES					
5. SUBCONTRACTS Includes ESnet PI				$604,200	
6. OTHER-Oranization burden costs associated with personnel				$29,340	
TOTAL OTHER DIRECT COSTS				$633,540	
H. TOTAL DIRECT COSTS (A THROUGH G)				$983,124	
I. INDIRECT COSTS (SPECIFY RATE AND BASE)					
TOTAL INDIRECT COSTS				$88,008	
J. TOTAL DIRECT AND INDIRECT COSTS (H+I)				$1,071,132	
K. AMOUNT OF ANY REQUIRED COST SHARING FROM NON-FEDERAL SOURCES					
L. TOTAL COST OF PROJECT (J+K)				$1,071,132	

DOE F 4620.1
(04-93)
All Other Editions Are Obsolete

U.S. Department of Energy

Budget Page

(See reverse for Instructions)

OMB Control No.
1910-1400
OMB Burden Disclosure
Statement on Reverse

ORGANIZATION
Oak Ridge National Laboratory (Year 2)

Budget Page No: Year 2

PRINCIPAL INVESTIGATOR/PROJECT DIRECTOR

Requested Duration: 12 (Months)

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates	DOE Funded				
(List each separately with title; A.6. show number in brackets)	Person-mos.			Funds Requested	Funds Granted
	CAL	ACAD	SUMR	by Applpicant	by DOE
1. Wing, William	4.0				
2. Rao, Nageswara	4.0				
3.					
4.					
5.					
6. ( ) OTHERS (LIST INDIVIDUALLY ON BUDGET EXPLANATION PAGE)					
7. ( ) TOTAL SENIOR PERSONNEL (1-6)	8.0			$105,899	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. ( ) POST DOCTORAL ASSOCIATES					
2. ( ) OTHER PROFESSIONAL (TECHNICIAN, PROGRAMMER, ETC.)					
3. ( ) GRADUATE STUDENTS					
4. ( ) UNDERGRADUATE STUDENTS					
5. ( ) SECRETARIAL - CLERICAL					
6. ( ) OTHER					
TOTAL SALARIES AND WAGES (A+B)				$105,899	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				$37,171	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A+B+C)				$143,070	
D. PERMANENT EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM.)					
TOTAL PERMANENT EQUIPMENT				$825,000	
E. TRAVEL 1. DOMESTIC (INCL. CANADA AND U.S. POSSESSIONS)				$10,000	
2. FOREIGN					
TOTAL TRAVEL				$10,000	
F. TRAINEE/PARTICIPANT COSTS					
1. STIPENDS (Itemize levels, types + totals on budget justification page)					
2. TUITION & FEES					
3. TRAINEE TRAVEL					
4. OTHER (fully explain on justification page)					
TOTAL PARTICIPANTS ( ) TOTAL COST					
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES					
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION					
3. CONSULTANT SERVICES					
4. COMPUTER (ADPE) SERVICES					
5. SUBCONTRACTS Includes ESnet PI				$604,200	
6. OTHER-Oranization burden costs associated with personnel				$36,000	
TOTAL OTHER DIRECT COSTS				$640,200	
H. TOTAL DIRECT COSTS (A THROUGH G)				$1,618,270	
I. INDIRECT COSTS (SPECIFY RATE AND BASE)					
TOTAL INDIRECT COSTS				$135,226	
J. TOTAL DIRECT AND INDIRECT COSTS (H+I)				$1,753,496	
K. AMOUNT OF ANY REQUIRED COST SHARING FROM NON-FEDERAL SOURCES					
L. TOTAL COST OF PROJECT (J+K)				$1,753,496	

DOE F 4620.1

(04-93)

All Other Editions Are Obsolete

U.S. Department of Energy

Budget Page

(See reverse for Instructions)

OMB Control No.
1910-1400
OMB Burden Disclosure
Statement on Reverse

ORGANIZATION Oak Ridge National Laboratory (Year 3)				**Budget Page No:**	Year 3
PRINCIPAL INVESTIGATOR/PROJECT DIRECTOR				Requested Duration:	12 (Months)
A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title; A.6. show number in brackets)	DOE Funded Person-mos.			Funds Requested	Funds Granted
	CAL	ACAD	SUMR	by Applpicant	by DOE
1. Wing, William	3.3				
2. Rao, Nageswara	3.3				
3.					
4.					
5.					
6. ( ) OTHERS (LIST INDIVIDUALLY ON BUDGET EXPLANATION PAGE)					
7. ( ) TOTAL SENIOR PERSONNEL (1-6)	6.7			$92,662	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. ( ) POST DOCTORAL ASSOCIATES					
2. ( ) OTHER PROFESSIONAL (TECHNICIAN, PROGRAMMER, ETC.)					
3. ( ) GRADUATE STUDENTS					
4. ( ) UNDERGRADUATE STUDENTS					
5. ( ) SECRETARIAL - CLERICAL					
6. ( ) OTHER					
TOTAL SALARIES AND WAGES (A+B)				$92,662	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				$32,525	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A+B+C)				$125,187	
D. PERMANENT EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM.)					
TOTAL PERMANENT EQUIPMENT				$275,000	
E. TRAVEL 1. DOMESTIC (INCL. CANADA AND U.S. POSSESSIONS)				$10,000	
2. FOREIGN					
TOTAL TRAVEL				$10,000	
F. TRAINEE/PARTICIPANT COSTS					
1. STIPENDS (Itemize levels, types + totals on budget justification page)					
2. TUITION & FEES					
3. TRAINEE TRAVEL					
4. OTHER (fully explain on justification page)					
TOTAL PARTICIPANTS ( ) TOTAL COST					
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES					
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION					
3. CONSULTANT SERVICES					
4. COMPUTER (ADPE) SERVICES					
5. SUBCONTRACTS Includes ESnet PI				$1,208,400	
6. OTHER-Oranization burden costs associated with personnel				$31,000	
TOTAL OTHER DIRECT COSTS				$1,239,400	
H. TOTAL DIRECT COSTS (A THROUGH G)				$1,649,587	
I. INDIRECT COSTS (SPECIFY RATE AND BASE)					
TOTAL INDIRECT COSTS				$108,959	
J. TOTAL DIRECT AND INDIRECT COSTS (H+I)				$1,758,546	
K. AMOUNT OF ANY REQUIRED COST SHARING FROM NON-FEDERAL SOURCES					
L. TOTAL COST OF PROJECT (J+K)				$1,758,546	

DOE F 4620.1

(04-93)

All Other Editions Are Obsolete

U.S. Department of Energy

Budget Page

(See reverse for Instructions)

OMB Control No.
1910-1400
OMB Burden Disclosure
Statement on Reverse

ORGANIZATION
Oak Ridge National Laboratory [Summary]

Budget Page No: Summary

PRINCIPAL INVESTIGATOR/PROJECT DIRECTOR

Requested Duration: 36 (Months)

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title; A.6. show number in brackets)	DOE Funded Person-mos.			Funds Requested	Funds Granted
	CAL	ACAD	SUMR	by Applpicant	by DOE
1. Wing, William	10.6				
2. Rao, Nageswara	10.6				
3.					
4.					
5.					
6. ( ) OTHERS (LIST INDIVIDUALLY ON BUDGET EXPLANATION PAGE)					
7. ( ) TOTAL SENIOR PERSONNEL (1-6)	21.2			$281,155	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. ( ) POST DOCTORAL ASSOCIATES					
2. ( ) OTHER PROFESSIONAL (TECHNICIAN, PROGRAMMER, ETC.)					
3. ( ) GRADUATE STUDENTS					
4. ( ) UNDERGRADUATE STUDENTS					
5. ( ) SECRETARIAL - CLERICAL					
6. ( ) OTHER					
TOTAL SALARIES AND WAGES (A+B)				$281,155	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				$98,686	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A+B+C)				$379,841	
D. PERMANENT EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM.)					
TOTAL PERMANENT EQUIPMENT				$1,320,000	
E. TRAVEL 1. DOMESTIC (INCL. CANADA AND U.S. POSSESSIONS)				$38,000	
2. FOREIGN					
TOTAL TRAVEL				$38,000	
F. TRAINEE/PARTICIPANT COSTS					
1. STIPENDS (Itemize levels, types + totals on budget justification page)					
2. TUITION & FEES					
3. TRAINEE TRAVEL					
4. OTHER (fully explain on justification page)					
TOTAL PARTICIPANTS ( ) TOTAL COST					
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES					
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION					
3. CONSULTANT SERVICES					
4. COMPUTER (ADPE) SERVICES					
5. SUBCONTRACTS Includes ESnet PI				$2,416,800	
6. OTHER-Oranization burden costs associated with personnel				$96,340	
TOTAL OTHER DIRECT COSTS				$2,513,140	
H. TOTAL DIRECT COSTS (A THROUGH G)				$4,250,981	
I. INDIRECT COSTS (SPECIFY RATE AND BASE)					
TOTAL INDIRECT COSTS				$332,193	
J. TOTAL DIRECT AND INDIRECT COSTS (H+I)				$4,583,174	
K. AMOUNT OF ANY REQUIRED COST SHARING FROM NON-FEDERAL SOURCES					
L. TOTAL COST OF PROJECT (J+K)				$4,583,174	

8.0 Appendix A: Examples from DOE Large-Scale Science Applications

The following projections of network requirements by Climate and High-Energy Physics are quoted from materials found in [DOEWorkshop-1]. They represent a small fraction of the material presented at the workshop, but are quoted here to give a flavor of the requirements projected for DOE networking in 3-5 year time frame.

8.1 Climate: The Next Five Years

Over the next five years, climate models will see an even greater increase in complexity than that seen in the last ten years. Influences on climate will no longer be approximated by essentially fixed quantities, but will become interactive components in and of themselves. The North American Carbon Project (NACP), which endeavors to fully simulate the carbon cycle, is an example. Increases in resolution, both spatially and temporally, are in the plans for the next two to three years. The atmospheric component of the coupled system will have a horizontal resolution of approximately 150 km and 30 levels. A plan is being finalized for model simulations that will create about 30 terabytes of data in the next 18 months, which is double the rate of model data generation of the PCM. These much finer resolution models, as well as the distributed nature of computing resources, will demand much greater bandwidth and robustness from computer networks than is presently available. These studies will be driven by the need to determine future climate at both local and regional scales as well as changes in climate extremes - droughts, floods, severe storm events, and other phenomena. Climate models will also incorporate the vastly increased volume of observational data now available (and that available in the future), both for hindcasting and intercomparison purposes. The end result is that instead of tens of terabytes of data per model instantiation, hundreds of terabytes to a few petabytes (10^{15}) of data will be stored at multiple computing sites, to be analyzed by climate scientists worldwide. The Earth System Grid and its descendents will be fully utilized to disseminate model data and for scientific analysis.

8.2 High Energy:

In order to build a " survivable" , flexible distributed system, much larger bandwidths are required, so that the typical transactions, drawing 1 to 10 Terabyte and eventually 100 Terabyte subsamples from the multi-petabyte data stores, can be completed in 1 to 10 minutes. Completing these transactions in a few minutes (rather than hours) is necessary to avoid the inherently fragile state that would result if hundreds to thousands of requests were left pending for long periods, and to avoid the bottleneck that would result from tens and then hundreds of such "data-intensive" requests per day (each still representing a very small fraction of the stored data). It is important to note that transactions on this scale correspond to data throughputs across networks of 10 Gbps to 1 Tbps for 10 minute transactions, and up to 10 Tbps (more than the current capacity of a fully instrumented fiber circa 2002) for 1 minute transactions.

In order to fully understand the potential of these applications to overwhelm future planned networks, we note that the binary (compacted) data stored is pre-filtered by a factor of 10^6 to 10^7 by the "Online System" (a large cluster of hundreds to thousands of CPUs that filter the data in real time). This real-time filtering, though traditional, runs a certain risk of throwing away data from subtly new interactions that do not pre-conceived existing or hypothesized theories. The basic problem is to find new interactions from the particle collisions, down to the level of a few interactions per year out of 10^{16} produced. A direct attack on this data analysis and reduction problem, analyzing every event in some depth, is beyond the current and foreseen states of network and computing technologies.

9.0 References[1]

[RPT-1] Report of the Transatlantic Networking Committee, Donno et. al. <http;//www.usatlas.bnl.gov/computing/mgmt/lhccp/henpnet/TAN-report-v7.4a.doc>

[Allcock 2002] B. Allcock et al, Data management and transfers in high-performance computational grid environments, *Parallel Computing*, May 2002, pp. 749-771.

[DOE-J03.2] DOE Science Computing Conference: The Future of High Performance Computing and Communications, June 19-20, 2003, http://www.doe-sci-comp.info

[DOE-J03.1] DOE Science Network Workshop: Roadmap to 2008, June 3-5, 2003, http://www.hep.anl.gov/may/ScienceNetworkingWorkshop

[DOE-A03] DOE Workshop on Ultra High-Speed Transport Protocols and Network Provisioning for Large-Science Applications, April 10-11, 2003, http://www.csm.ornl.gov/ghpn/wk2003

[DOE-A02] High-Performance Network Planning Workshop, August 13-15, 2002, Report: High-Performance Networks for High-Impact Science, http://DOECollaboratory.pnl.gov/meetings/hpnpw

[Dunigan et al 2002] T. Dunigan, M. Mathis and B. Tierney, A TCP Tuning Daemon, **Supercomputing 2002**, July 2002.

[Foster et al] I. Foster et al, Grid services for distributed systems integration, *IEEE Computer*, June 2002, pp. 37-46.

[Mukherjee 1997] B. Mukherjee, Optical Communications Networks, McGraw-Hill, 1997.

[net100] Net100 project, www.net100.org

[NSF-02.1] NSF Workshop on Ultra-High Capacity Optical Communications and Networking, October 21-22, 2002

[NSF-02.2] NSF Workshop on Network Research Testbeds, October 17-18, 2002, http://gaia.cs.umass.edu/testbed_workshop

[NSF-02.3] NSF ANIR Workshop on Experimental Infostructure Networks, May 20-21, 2002, http://www.calit2.net/events/2002/nsf/index.html

[NSF-01]NSF CISE Grand Challenges in e-Science Work, December 5-6, 2001, http://www.evl.uic.edu/activity/NSF/index.html

Ramaswami and Sivarajan 2002] R. Ramaswami and K. N. Sivarajan, Optical Networks: A Practical Perspective, Morgan Kaufman Pub., 2002.

[Rao et al 2003] N. S.V. Rao, Q. Wu, and S.S. Iyengar, Stabilizing network transport, ORNL manuscript, 2003.

[Rao and Chua 2002] N.S.V. Rao and L. O. Chua, On dynamics of network transport protocols, *Proc. of Workshop on Signal Processing, Communications, Chaos and Systems*, 2002.

[1] Due to the large number of networking, middleware and application areas (such as optical networks, switching and routing, transport, ftp, remote visualization, supernova computation and spallation neutron source) covered by this proposal, this list of references is only a small *ad-hoc* sampling of the literature chosen to highlight the presentation and is far from being complete.

[Rao and Feng 2002] N.S.V.Rao and W.C.Feng, Performance trade-offs of TCP adaptation methods, *Proc. Int. Conf. Networking*, 2002.

[Rao 2002] N. S. V. Rao, NetLets for end-to-end delay minimization in distributed computing over Internet using two-Paths, *International Journal of High Performance Computing Applications*, 2002, vol. 16, no. 3, 2002.

[Reed 2003] D. A. Reed, Grids, the TeraGrid, and beyond, *IEEE Computer*, January 2003, pp. 62-68.

[TSI] Terascale Supernova Initiative, http://www.phy.ornl.gov/tsi

10. Biographies of Principal Investigators

Nageswara S. Rao

Education

PhD	1988	Computer Science	Louisiana State University
ME	1984	Computer Science	Indian Institute of Science, Bangalore
BE	1982	Electronics Engineering	Regional Engineering College, Warangal, India

Work Experience

1. Distinguished Research Staff (2001-present), Senior Research Staff Member (1997-2001), Research Staff Member (1993-1997), Intelligent and Emerging Computational Systems Section, Computer Science and Mathematics Division, Oak Ridge National Laboratory..
2. Assistant Professor, Department of Computer Science, Old Dominion University, Norfolk, VA 23529- 0162, 1988 – 1993; Adjunct Associate Professor, 1993 - present.
3. Research and Teaching Assistant, Department of Computer Science, Louisiana State University, Baton Rouge, LA, 1985 - 1988.
4. Research Assistant, School of Automation, Indian Institute of Science, Bangalore, India, 1984 - 1985.

Technical Activities

He published more than 180 research papers in a number of technical areas including the following.

1. Measurement-based methods end-to-end performance assurance over wide-area networks: He rigorously showed that measurement-based methods can be used to provide probabilistic end-to-end delay guarantees in wide-area networks. He also implemented this method over the Internet and showed that significant reductions in the end-to-end delay can be realized using these methods. This work is currently funded by NSF, DARPA and DOE.
2. Wireless networks: He is a co-inventor of the connectivity-through-time protocol that enables communication in adhoc wireless dynamic networks with no infrastructure. He carried out a complete implementation of this system using IEEE 802.11 cards. Here, the dynamic movements of nodes are utilized to deliver messages in an adhoc network. A US patent is pending.
3. Sensor Fusion: Developed novel fusion methods for combining information from multiple sources using measurements without any knowledge of error distributions. He proposed fusers that are guaranteed to perform at least as good as best subset of sources. This work is funded by DOE, NSF and ONR. He organized the first international workshop in this area in 1996, which was subsequently developed into two international conferences.
4. Robot Navigation: He was the first to formulate and solve the terrain model acquisition problem by mobile robots using visibility graph, Voronoi diagram and trapezoidal dual graphs. He developed n-connectivity methods for model acquisition by robot teams. This work is funded by NSF (Research Initiation Award in 1991) and DOE.
5. Fault Diagnosis: He developed diagnosis algorithms for fault propagation graphs. This work is now utilized by others for network intrusion detection and diagnosis of optical networks. This work is funded by NSF.

Publications Related to the Proposal

1. N.S.V. Rao, Multiple paths for end-to-end delay minimization in distributed computing over Internet, Proc. Supercomputing 2001: Conf. on High-Performance Computing and Networking.
2. N.S.V. Rao, NetLets for end-to-end delay minimization in distributed computing over Internet using two-Paths, International Journal of High Performance Computing Applications, 2002, in press.
3. N.S.V. Rao, S.G. Batsell, QoS routing via multiple paths using bandwidth reservation, IEEE INFOCOM'98, The Conference on Computer Communications, 1998.
4. N.S.V. Rao, S.G. Batsell, Algorithm for minimum end-to-end delay paths, *IEEE Communications Letters*, vol. 1, no. 5, 1997, pp. 152-154.
5. N.S.V. Rao, K. Maly, S. Olariu, S. Dharanikota, L. Zhang, D. Game, Average waiting time profiles of uniform distributed queue dual bus system model, *IEEE Transactions on Parallel and Distributed Systems*, vol. 6, no. 10, 1995, pp. 1068-1084.

Other Significant Publications

1. N.S.V. Rao, On fusers that perform better than best sensor, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 904-909, 2001.
2. N.S.V. Rao, V. Protopopescu, On PAC learning of functions with smoothness properties using feedforward sigmoidal networks, vol. 84, no. 10, *Proceedings of IEEE*, 1996, pp.1562-1569.
3. N.S.V. Rao, V. Protopopescu, R.C. Mann, E.M. Oblow, S.S. Iyengar, Learning algorithms for feedforward networks based on finite samples, *IEEE Transactions on Neural Networks*, vol. 7, No. 4, 1996, pp. 926-940.
4. N.S.V. Rao, Distributed decision fusion using empirical estimation, *IEEE Transactions on Aerospace and Electronic Systems*, vol. 33, no. 4, 1997, pp.1106-1114.
5. N.S.V. Rao, Computational complexity issues in operative diagnosis of graph-based systems, *IEEE Transactions on Computer*s, vol. 42, no. 4, 1993, pp. 447-457.

T. H. Dunigan

Education

PhD	1976	Computer Science	University of North Carolina
MS	1972	Computer Science	University of North Carolina
BS	1970	Physics and Mathematics	Duke University

Work Experience:

1976 – Present; Senior Research Staff, Oak Ridge National Laboratory, Computer Science and Mathematics Division
1999 – Present; Computer Science Department, University of Tennessee

Contributions:

Dr. Dunigan has active research interests in intrusion detection systems, network performance, and in the performance characterization and analysis of parallel computers and their communication subsystems. For the last 14 years, he has led efforts at ORNL for evaluating early releases of parallel computing systems, including the Compaq AlphaServer SC, Intel iPSC/1, iPSC/2, iPSC/860, and Paragon, and the SRC 6, Kendall Square, and Chen shared-memory multiprocessors. He has been actively involved in the Internet since helping bring the Arpanet/Internet to ORNL in the early 80s. Recent research activities include using Network traffic flow characteristics to detect intrusions and investigating modifications to the TCP protocol to improve high-latency, high-bandwidth bulk transfers. As a collaborating scientist with the University of Tennessee, Dr. Dunigan has taught graduate-level Computer Science courses in networks and in computer security.

Selected Publications:

1 Flow Characterization for Intrusion Detection, T. Dunigan and G. Ostrouchov, ORNL-TM, December, 2000.

2 Backtracking Spoofed Packets, ORNL-TM, November, 2000.

3 Intrusion Detection and Intrusion Prevention on a Large Network: A Case Study, T. Dunigan and G. Hinkel, *USENIX Workshop on Intrusion Detection*, April, 1999.

4 The Locally Self-Consistent Multiple Scattering Method in a Geographically Distributed Linked MPP Environment, T. Sheehan, W. Shelton, T. Pratt, P. Papadopoulos, P. LoCascio, and T. Dunigan, *Parallel Computing*, v24, November, 1998.

5 Message-passing Performance of Various Computers, ORNL/TM-13006, with Dongarra, 1996, *Concurrency: Practice & Experience*, v9(10), October 1997.

6 Group Key Management, ORNL/TM-13470, with C. Cao, 1997.

7 Poisson Type Models and Descriptive Statistics of Information Flows, ORNL/TM-13468, with Downing, Fedorov, Batsell, 1997.

8 Performance of ATM/OC-12 on the Intel Paragon, ORNL/TM-13239, 1996.

9 Secure PVM, ORNL/TM-13203, with N. Venugopal, 1996.

10 PVM and IP Multicast, ORNL/TM-13030, with K. Hall, 1996.

11 Hypercube Simulation on a Local Area Network, ORNL/TM-10685, 1988.

12 A Message-passing Multiprocessor Simulator, ORNL/TM-9966, 1986.

13 Denelcor HEP Multiprocessor Simulator, ORNL/TM-9971, 1986.

W. R. Wing

Education

PhD	1972	Physics	University of Iowa, Iowa City, IA
MA	1968	Physics	University of Iowa, Iowa City, IA
BA	1965	Physics	University of Iowa, Iowa City, IA

Work Experience:

1999 – Present; Senior Research Staff Member - Networking Research Group, Computer Science and Mathematics Division, Oak Ridge National Laboratory
1991 – 1999; Senior Research Staff Member – Computing, Information, and Networking Division, Office of Computing and Network Management, Office of Laboratory Computing, Oak Ridge National Laboratory
1972 – 1991; Research Staff Member – Fusion Energy Division, Oak Ridge National Laboratory

Research Interests:

Network Monitoring and Instrumentation, Network Simulation, and Protocol Development

Contributions:

Bill Wing joined the Fusion Energy Division immediately after completing his PhD. He developed and applied a wide variety of diagnostic instruments for characterizing the fusion plasmas in experimental devices there. He received a patent for one of these, a Gigacycle Correlator. While there, he started using early laboratory-scale computers (PDP-8, PDP-12, and a PDP-10) for data acquisition and analysis. He led the in-house programming group responsible for writing data acquisition software and developed an integrated data acquisition system that spread throughout the fusion community. His interest in computerized analysis and modeling led to an interest in networking (ORNL's Fusion Energy Division was one of the first backbone nodes on the Magnetic Fusion Energy Network, which linked Fusion sites to the MFE computer center at Livermore). In 1991, he moved from the Fusion Energy Division to the Office of Laboratory Computing to help improve ORNL's position in the high-performance computing and networking community. He chairs the Department of Energy's ESnet Site Coordinating Committee, as well as serving as ORNL's representative on it. He serves on the ESnet Steering Committee, was chair of the SCinet committee for SC99 in Portland, and again for SC01 in Denver.

Selected Publications:

1. Internet Monitoring in the Energy Research Community, with Cottrell et. al. IEEE Network Transactions, special issue on the Internet 1997.
2. Data Acquisition in Support of Physics, Chapter in "Basic and Advanced Diagnostic Techniques for Fusion Plasmas" Published by Commission of the European Communities Directorate General XII – Fusion Programme, 1049 Brussels, Belgium - 1986
3. Soft X-ray Techniques, Chapter in "Course on Plasma Diagnostics and Data Acquisition" Eubank and Sindoni Editors, Published by C. N. R. – Euratom – 1975
4. Configuration Control Experiments Using Long-Pulse ECH Discharges in the ATF Torsatron, 18th Conf. On Controlled Fusion and Plasma Physics, Berlin, 1991
5. Transport Sutdies Using Modulation of Dimensionless Parameters in the Advanced Toroidal Facility; with Murakami, M.; Int. Conf. On Plasma Physics (EPS) Innsbruck, Austria, 1992
6. Energy Confinement Scaling in ATF; Bull. Am. Phys. Soc. 1989
7. Fluctuations and Stability in the ATF Torsatron, with Harris, J.H.;v. 2, pp. 677-84, Proc. 13th Int. Conf. on Plasma Physics and Controlled Nuclear Fusion Research, Washington, DC, Oct.1-6,1990, Nuclear Fusion Suppl., IAEA, 1990

8. Power and Particle Balance Studies Using an Instrumented Limiter System on ATF, with Uckan, T.Bull. Am. Phys. Soc.v.35 1990pp.2063
9. Biasing Experiments on the ATF Torsatron,; withUckan, T.;18th Eur. Phys. Soc. Conf., Div. Plasma Physics, Berlin, June 3-7,1991
10. Configuration Control, Fluctuations, and Transport in Low-Collisionality Plasmas in the ATF Torsatron,; with Harris, J.H.;18th Conf. on Controlled Fusion and Plasma Physics, Berlin, June 3-7,1991
11. Pellet Injection into ATF Plasmas,;with Wilgen, J.B.Bull. Am. Phys. Soc.v.34 1989pp.1947
12. Operating Window Enhancement of the ATF Torsatron,; with Glowienka, J.C.Bull. Am. Phys. Soc.v.34 1989pp.1946
13. Boundary Radiation and Plasma Stability in Currentless Confinement Devices,; with Isler, R.C.;v. 2, Proc. 9th Int. Conf. on Plasma Surface Interactions, Satellite Meet. on the Production and Stability of Cold Edge Layers for Fusion Reactors, Cadarache, France, May 28-30,1990 (1990)
14. The ATF Diamagnetic Diagnostic, Status and Results,;Wing, W.R.;8th Top. Conf. on High-Temperature Plasma Diagnostics, Hyannis, MA, May 6-10,1990