4Q FY 09

Lawrence Berkeley National Laboratory Advanced Networking Initiative

Design Document

U.S. DEPARTMENT OF ENERGY
Office of Science

Energy Sciences Network, Lawrence Berkeley National Lab

Table of Contents

Abstract 3
Background and Purpose 3
Project Scope 5
Component-1: 100 Gbps Prototype Network 5
Component-2: BayExpress 6
Component-3: ChiExpress 8
Component-4: NYExpress 9
Component-5: NashExpress 10
Component-6: Testbed 11
Testbed Timeline: 14
Summary: Advanced Networking Initiative Component Goals 14
Component-1 Goal: 14
Component-2 Goals: 14
Component-3 Goals: 14
Component-4 Goals: 15
Component-5 Goals: 15
Component-6 Goals: 15
Overall Project Goal: 15

Abstract

DOE scientists and collaborators worldwide have long required high-bandwidth connections between DOE sites and the university community. The emergence of distributed petascale science, characterized by extreme-scale computing resources and large experiments generating petabytes of data used worldwide, demands that the network be fast enough to meet the needs of this new, highly distributed model. By leveraging the existing knowledge base of the Energy Sciences Network located at LBNL, and funded through the American Recovery and Reinvestment Act of 2009, the Advanced Network Initiative (ANI) will demonstrate a prototype network capable of enhanced support of "big science" while furthering the state of commercial high-speed data communication by accelerating and influencing the roadmap of terabit networking.

ANI will conduct research and development on the challenges of providing 100 Gbps throughput capacity via point-to-point circuits by demonstrating a developmental prototype backbone network on a national fiber footprint and across several metropolitan area networks. This endeavor requires close collaboration between the research and education community and industry to develop, test and deploy new technologies, some pre-standard, in order to validate a prototype infrastructure capable of carrying multiple 100 Gbps circuits.

The Advanced Network Initiative will additionally provide for the creation and deployment of an experimental network research environment that will enable researchers to develop and test new and radical network and service concepts across multiple domains and network layers.

Background and Purpose

Large-scale collaborative science – big facilities, massive amount of data, thousands of collaborators – is the key to success for many of the DOE's Office of Science (SC) research activities. There are some 20 major instruments and facilities currently operated or being built by SC, plus the LHC (CERN, Switzerland) and ITER (France) collaborations. Physical facilities such as DOE's big accelerators (RHIC at Brookhaven, SNS at Oak Ridge) and supercomputer centers (NERSC at Lawrence Berkeley, OLCF at Oak Ridge, and ALCF at Argonne) are typical of the science infrastructure supported by ESnet. ESnet also supports virtual facilities such as the Earth System Grid and the Open Science Grid. The Earth System Grid is a federated network of thousands of global climate modelers whose federated data and simulations represent the majority of the global effort in support of the Intergovernmental Panel on Climate Change. The Open Science Grid is a consortia of 30 communities and 70 sites accelerating large-scale science (including global high energy physics experiments through LHC in CERN) through a consortia-based, persistent infrastructure that supports sharing of distributed computing, data,

middleware, and storage. Distributed computing and storage systems for data analysis, simulations, instrument operation, etc., are becoming commonplace as we move closer to the paradigm of cloud computing where the aspiration of 'the network being the computer' can only become reality if the network is as fast as the computer, if not faster.

This new distributed science environment is very different from that of a few years ago and places substantial new demands on the network. High-speed, highly reliable connectivity between labs and US and partner R&E institutions is required to support the inherently collaborative, global nature of large-scale science. Increasing capacity is needed to accommodate the large and steadily growing data sets that must traverse the network to get from instruments to scientists and to analysis, simulation, and storage facilities. High network reliability is required for interconnecting components of distributed large-scale science computing and data systems and to support various modes of remote instrument operation. New network services are needed to provide bandwidth guarantees for data transfer deadlines, remote data analysis, real-time interaction with instruments, coupled computational simulations, etc.

It is believed that this endeavor, the Advanced Networking Initiative, will provide the foundation for revolutionizing the way scientists communicate and manage data and computing in their day-to-day work environment and could have transformational impacts on commercial network providers as well. This effort will provide a research prototype infrastructure for ESnet as the primary network provider connecting the distributed research elements in DOE. Understanding the technical and operational challenges related to 40 and 100 Gbps technologies is essential because, while the current ESnet4 backbone supports multiple optical circuits to exchange scientific data, forecasts of near-term network demand cannot be met by current state-of-the-art 10 Gbps systems.

The work proposed here will not only push equipment vendors and telecommunications service providers toward the next generation of US networking technology, it will provide leadership for other research and education network organizations that support the institutions that collaborate with the DOE Labs in order to accomplish the large-scale science mission of the Office of Science. This work is motivated by the fact that the projected total bandwidth capacity required by the Office of Science in 2004 led to today's ESnet4 designed capacity. However, even though the ESnet4 design included what seemed at the time to be a comfortable capacity margin, the needs of the SC community as determined by workshops in 2007-2008 have grown faster than was projected in 2004, and will substantially exceed the ESnet4 capacity based on the current 10 Gbps optical channels. In fact, the needs are growing so fast that ESnet engineers believe that it will not be possible to satisfy them on the number of 10 Gbps optical channels that are currently available to ESnet on the national fiber footprint of the ESnet-Internet2 partnership. This motivates the need to increase the capacity of the fiber

while maintaining spectral efficiency by transitioning to greater than 10 Gbps optical channels.

Project Scope

The ANI Prototype Demonstration Network effort is divided into six distinct tasks, or components. Each component is a standalone undertaking, dealing with the specific challenges associated with demonstrating a 100 Gbps capability to a given location, or, in the case of the testbed, creating a facility to carry out research on optical control protocols. Component 1, the prototype demonstration backbone, will validate the concept of a wide area network with 100 Gbps capabilities and provide better understanding of its operational challenges. Component 2, BayExpress, will extend this capability to the demark of the National Energy Research Science Computer Center (NERSC) from the ESnet hub in Sunnyvale by solving the issues related to moderate size metropolitan network deployments. Component 3, ChiExpress, connects the ESnet point-of-presence in Chicago to the Argonne Leadership Computing Facility (ALCF) demark over a larger metropolitan network. NYExpress, Component 4, is designed to extend 100 Gbps capabilities from the demonstration backbone to the multi-agency peering point in New York City and the issues related to operating a high-speed peering facility. The demark to Oak Ridge Leadership Computing Facility (OCLF) at Oakridge National Laboratory is tied to the prototype backbone via Component 5, NashExpress. Finally, Component 6 will create an experimental network research environment at sufficient scale to test experimental approaches to next generation networks. LBNL will attempt to secure dark fiber to accomplish the goal for each listed component, where procuring such an asset is feasible and fiscally practical.

The core scope of the ANI project is described in the preceding paragraph; there are however, stretch goals for the project envisioned to be applicable to meeting the longer-term needs of the Office of Science mission. These goals will be attained, if possible, through the opportunistic acquisition and realization of strategic networking resources.

Component-1: 100 Gbps Prototype Network

The objective of Component-1 (prototype demonstration backbone) is to conduct research and development to demonstrate a prototype national network that will span at least four (4) distinct geographic regions, covering the three (3) major ASCR computing facilities and the New York multi-agency peering point which provides transatlantic Research and Education (R&E) connectivity.

Separately, DOE has funded Magellan, an associated DOE computing project, at $33M that will utilize the 100 Gbps prototype network infrastructure. This research and development effort to establish a nationwide scientific mid-range distributed computing and data analysis testbed will have at least two sites (NERSC / LBNL and

ALCF / ANL) with multiple 10's of teraflops and multiple petabytes of storage, as well as appropriate cloud software tuned for moderate concurrency.

The 100 Gbps prototype network will leverage computational and filesystem resources deployed by the Magellan Project as the endpoints for demonstrating the efficacy of 100 Gbps networks. The Magellan resources are expected to have filesystem bandwidth in excess of 200 Gbps and are an ideal demonstration platform for end-to-end performance.

The topology for the 100 Gbps prototype network, overlaid on the existing ESnet4 network, is shown below.

Figure 1 ANI Prototype Network Backbone

While the objectives of Component 1 may be satisfied by procuring 100 Gbps capacity lit service from the ultimately chosen vendor, it has been made clear by the ANI review panel that the selection criteria should place a high priority on securing dark fiber assets. The LBNL best value source selection procurement methodology will be employed on Component 1, as well as all other components, to insure the short and longer term impacts of this investment are maximized.

Component-2: BayExpress

The objective of Component-2 (BayExpress) is to conduct research and development on the challenges of providing 100 Gbps throughput capacity via point-to-point 10/40/100 Gbps circuits from the border of the Office of Science ASCR computing facility at NERSC to the 100 Gbps capable transport prototype backbone network node in the Sunnyvale, CA region. Today, NERSC connectivity is

supplied by a commercial carrier utilizing a "lit service" over a dark fiber ring at 2 x 10 Gbps utilizing last-generation metropolitan optical transport hardware. NERSC and the Sunnyvale hub are part of the Bay Area Metropolitan Network (BAYMAN), which ties together 7 ESnet sites in a ring topology. Prior to the Advanced Networking Initiative, the commercial carrier had no intention of upgrading the optical transmission equipment to 40 and 100 Gbps technologies despite ESnet's expressed interest in moving to those speeds. Challenges to upgrading the BayExpress to 100 Gbps capabilities are considerable. LBNL will need to either work with a commercial service provider on their first deployment of 100 Gbps transport technologies or discover new fiber/carrier resources and appropriate 100 Gbps capable metro transport equipment to light the path between NERSC and the prototype network backbone node. No matter which course LBNL pursues, it will be faced with the unfamiliar challenges of deploying 40 and 100 Gbps optical technologies over different fiber types (ITU-T G.652, G.655, etc.) in the field and how their chromatic and polarization mode dispersion characteristics and deployment mode (buried or aerial) affect the transmitted signal. Also, significant compatibility testing and follow on work with various vendors will be necessary as it is likely both the transport services and routing/switching vendors will employ pre-IEEE standards technologies. If the vendor selected remains the current vendor using the current fiber path, it is imperative that the new prototype service be constructed in such a way as minimize disruptions to the current production traffic while building out the 100 Gbps transport capability.

The topology for the 100 Gbps BayExpress, overlaid on the existing ESnet4 BAYMAN, is depicted below. The 100 Gbps circuit between NERSC and the Sunnyvale backbone hub is shown in red.

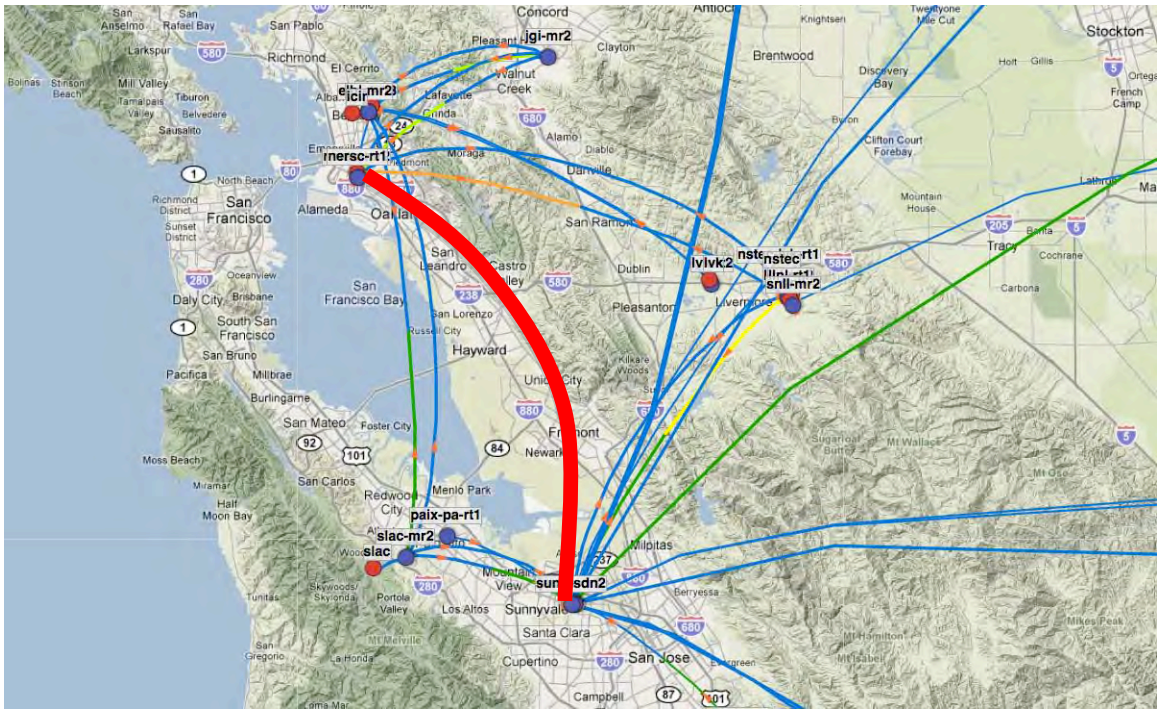

Figure 2 ANI BayExpress Topology

Component-3: ChiExpress

The objective of Component-3 (ChiExpress) is to conduct research and development on the challenges of providing 100 Gbps throughput capacity via point-to-point 10/40/100 Gbps circuits from the border of the Office of Science ASCR computing facility at Argonne National Laboratory to the 100 Gbps capable transport prototype backbone network node in Chicago. Today, ANL and Fermi National Accelerator Laboratory (FNAL) connect to the Chicago hub utilizing optical technologies deployed on dark fiber procured from several sources; the State of Illinois IWIRE project as well as fiber owned by the Laboratories themselves. These services, which ESnet uses as its Chicago Metropolitan Area Network (CHIMAN), are provided through a collaboration between ANL and FNAL. CHIMAN is constructed using multiple 10 Gbps Ethernet point-to-point circuits in a ring topology and lit with lab-owned metropolitan optical transport equipment. Challenges in upgrading to 100 Gbps capabilities for ChiExpress are many. ESnet and the CHIMAN service providers are unfamiliar with the challenges of deploying 40 and 100 Gbps optical technologies over different fiber types (ITU-T G.652, G.655, etc.) in the field and how chromatic and polarization mode dispersion characteristics and deployment mode (buried or aerial) affect the transmitted signal. Also, significant compatibility testing and follow-on work with various vendors will be necessary as it is likely both the transport services and routing/switching vendors will employ pre-IEEE standards technologies. The demonstration network must be built taking care to minimize disrupting the current data capabilities in the Chicago area, particularly the high energy physics community's traffic from the Large Hadron Collider (LHC) transiting the CHIMAN infrastructure to FNAL.

The topology for the 100 Gbps ChiExpress, overlaid on the existing ESnet4 CHIMAN, is depicted below. The 100 Gbps circuit between ANL and the Chicago backbone hub is shown in red.

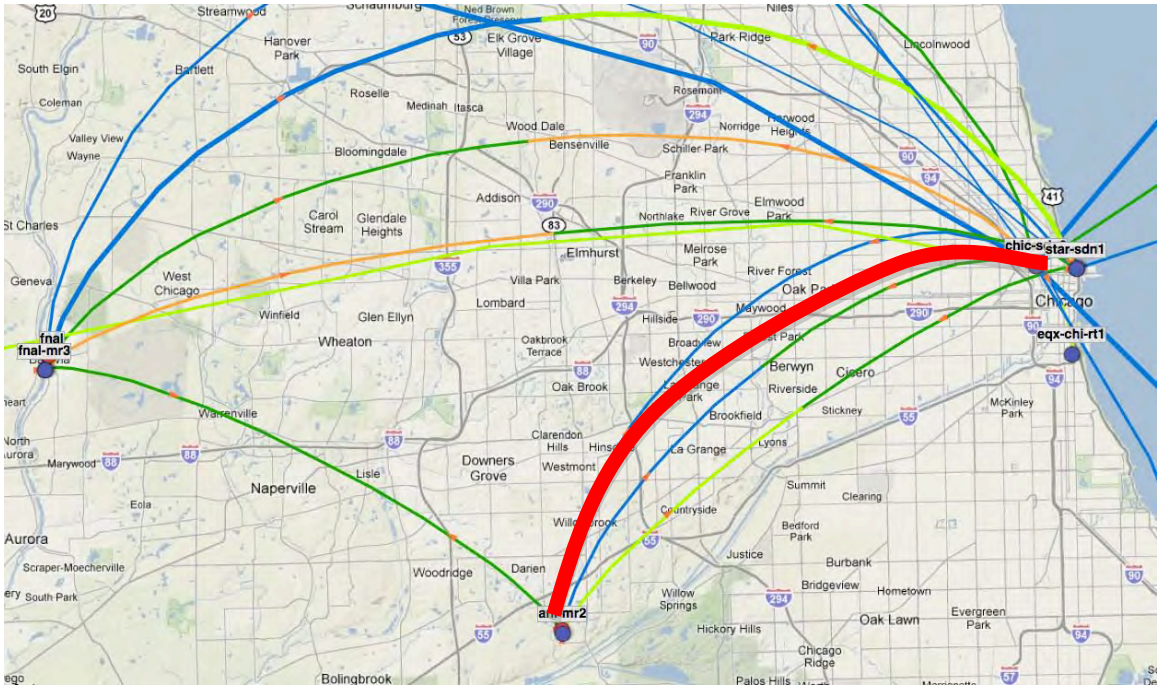

Figure 3 ANI ChiExpress Topology

Component-4: NYExpress

The objective of Component-4 (NYExpress) is to conduct research and development on the challenges of providing 100 Gbps throughput capacity via point-to-point 10/40/100 Gbps circuits from MANLAN, the NY multi-agency peering facility, to the 100 Gbps capable transport prototype backbone network node on the 24th floor, 111 8th Avenue, New York City. Today, ESnet connectivity is supplied by its R&E networking partner Internet2 utilizing dark fiber between the locations lit with Infinera's Digital Transport Node optical equipment and managed by the Indiana University Global Network Operations Center. ESnet also has leased fiber from a commercial provider as part of the Long Island Metropolitan Area Network (LIMAN) project. Challenges in upgrading this infrastructure are considerable. LBNL will either need to procure their own metropolitan and riser fiber within a multi-tenant environment and light it with its own optical transport equipment, or work with Internet2 on their very first deployment of 100 Gbps transport technologies. No matter which course LBNL pursues, it will be faced with the unfamiliar challenges of deploying 40 and 100 Gbps optical technologies in a metropolitan multi-tenant environment with multiple patch panels and cross-connects. As ESnet will be connecting to many international partners at the peering facility, significant testing and follow on work with various vendors will be necessary to ensure compatibility not only with other 100 Gbps transport services and routing/switching vendors

deploying pre-IEEE standards technologies, but also sub-sea cable systems used for trans-Atlantic connectivity. It will be imperative that the new prototype service be provisioned in such a way as to minimize disruptions to the current Large Hadron Collider (LHC) production traffic traversing the LIMAN between the Tier 1 center at Brookhaven National Laboratory (BNL), and CERN and Tier 2 centers located at universities worldwide.

The topology for the 100 Gbps NYExpress, overlaid on the existing ESnet4 LIMAN, is depicted below. The 100 Gbps circuit between MANLAN and the NYC backbone hub is shown in red.

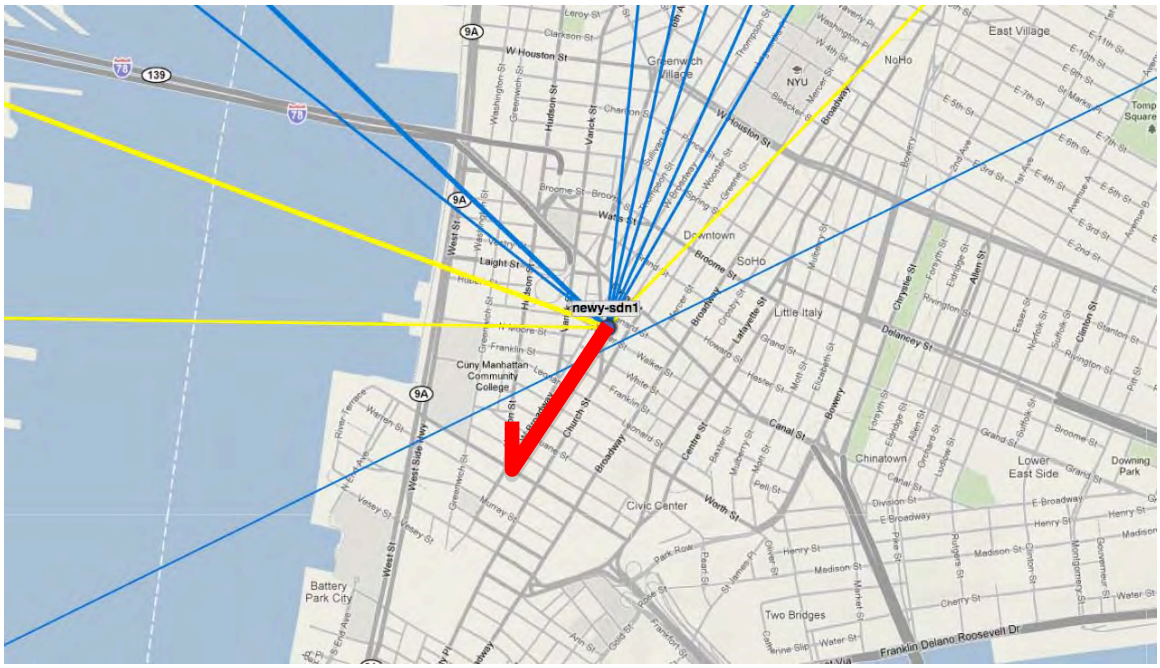

Figure 4 ANI NYExpress Topology

Component-5: NashExpress

The objective of Component-5 (NashExpress) is to conduct research and development on the challenges of providing 100 Gbps throughput capacity via point-to-point 10/40/100 Gbps circuits between the Office of Science ASCR computing facility at Oak Ridge National Laboratory and the 100 Gbps capable transport prototype backbone network node in Nashville, TN. Currently, ORNL has 10G connectivity to ESnet via a combination of manage services from a commercial service provider and dark fiber lit by R&E partners. The commercial carrier manages the ORNL-leased Ciena optical transport equipment on fiber provided by the carrier and the Tennessee Valley Authority. The R&E partner uses a different vendor's optical networking gear to light its leased fiber. Challenges to providing 100 Gbps capable transport services to ORNL are considerable. The existing Ciena equipment will need to be upgraded or replaced in order to be 100 Gbps capable. Also, because the R&E partner provides services to other institutions, ORNL is

pursuing their own dark fiber solution to span the 6 miles between the two facilities in Nashville. If they are not successful in procuring fiber in Nashville, LBNL will need to explore the costs and issues related to connecting ORNL to the ANI backbone node in Chicago over ORNL's FutureNet. No matter which course is pursued, ESnet and its partners will be faced with the challenges of deploying 40 and 100 Gbps optical technologies over different fiber types (ITU-T G.652, G.655, etc.) with varying chromatic and polarization mode dispersion characteristics. Lastly, significant compatibility testing and follow on work with various vendors will be necessary as it is likely the different transport services and routing/switching vendors will employ pre-IEEE standards technologies. It will be imperative that the new prototype service is brought on line with minimal disruption to the current production traffic between ESnet and ORNL.

The topology for the 100 Gbps NashExpress, overlaid on the existing ESnet4 NASHMAN, is depicted below. The 100 Gbps circuit between ORNL and the Nashville backbone hub is shown in red.

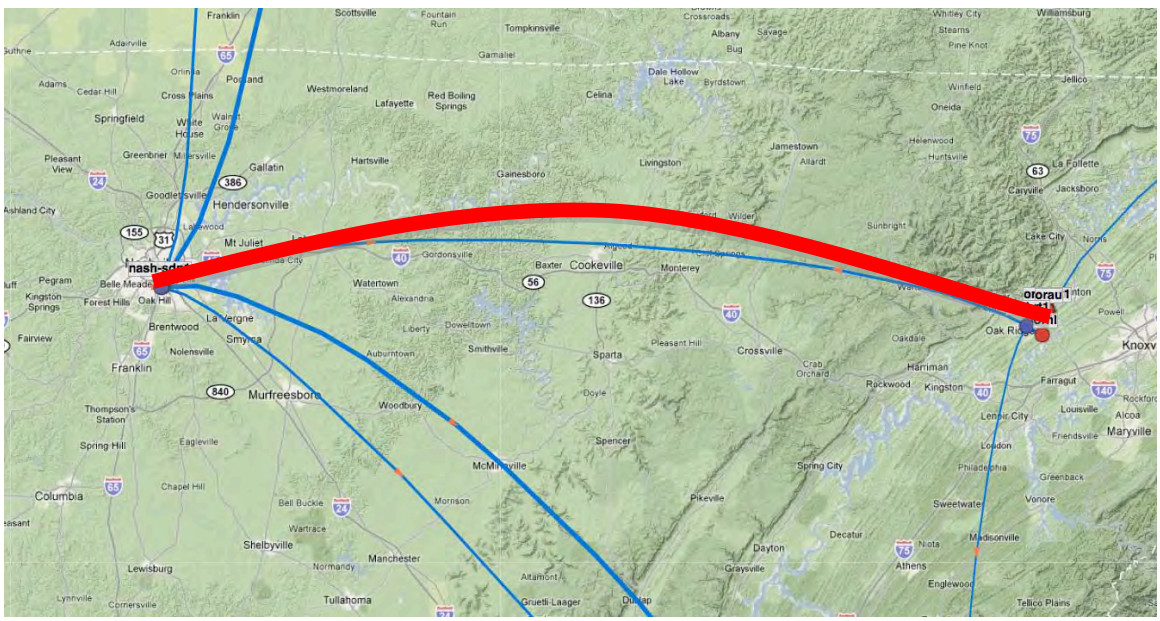

Figure 5 ANI NashExpress Topology

Component-6: Testbed

The objective of Component-6 (Testbed) is to create an experimental network research environment at sufficient scale to usefully test experimental approaches to next generation networks. The proposed research facility will be a breakable network research environment that will enable researchers to develop and test new and radical network concepts. This testbed will consist of advanced network devices and components assembled to give network and middleware researchers an environment in which to experiment with capabilities such as the following:

- Plan to acquire and use dark fiber on a portion of testbed footprint

    - Supporting the ability to do hybrid (layer 0-3) networking, thereby combining the best features of packet and circuit switching into one integrated infrastructure
- Use virtual machine technology to support protocol and middleware research, including research that spans boundaries with Magellan (mid-range computing and data analysis testbed at NERSC/LBNL and ALCF/ANL.
- Detailed monitoring of the testbed will be made available to the community
    - Researchers will have access to all possible monitoring data from the network devices
- Multi-layer dynamic network technologies support for advanced services such as secure end-to-end on-demand bandwidth and circuits over Ethernet, SONET, and optical transport network technologies
- Terabit networking component testing – via partnerships and joint-development with telecommunication vendors and ISPs, to conduct trials on new technologies and concepts critical to DOE but that are not yet commercially available
- Provisioning a community network R&D resource – the experimental facility will be open to researchers to conduct research activities
- Data movement middleware that can scale to speeds of 100 Gbps and beyond
- Ability to test the automatic classification of large bulk data flows and move them to a dedicated virtual circuit.
- Network-aware application testing – provide opportunities for network researchers and application developers such as Grid-based middleware, cyber security services, and so on, to exploit advanced network capabilities in order to enhance end-to-end performance and security.
- Technology transfer to production networks – ESnet, as host of the facility, will develop strategies to move mature technologies from testing mode to production service.

Each node will consist of:

- DWDM device (Layer 0-1)
- Programmable Ethernet Switch (layer 2)
- Standard and programmable Router (layer 3)
- Test and measurement hosts
    - VM based test environments

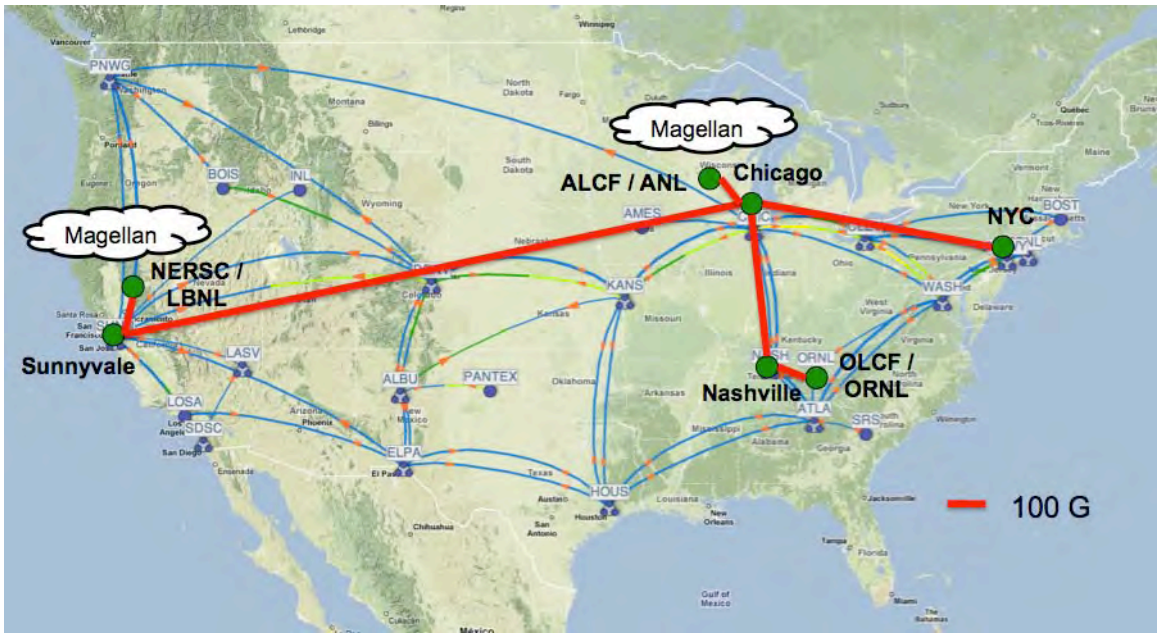

Figure 6 National Testbed Topology

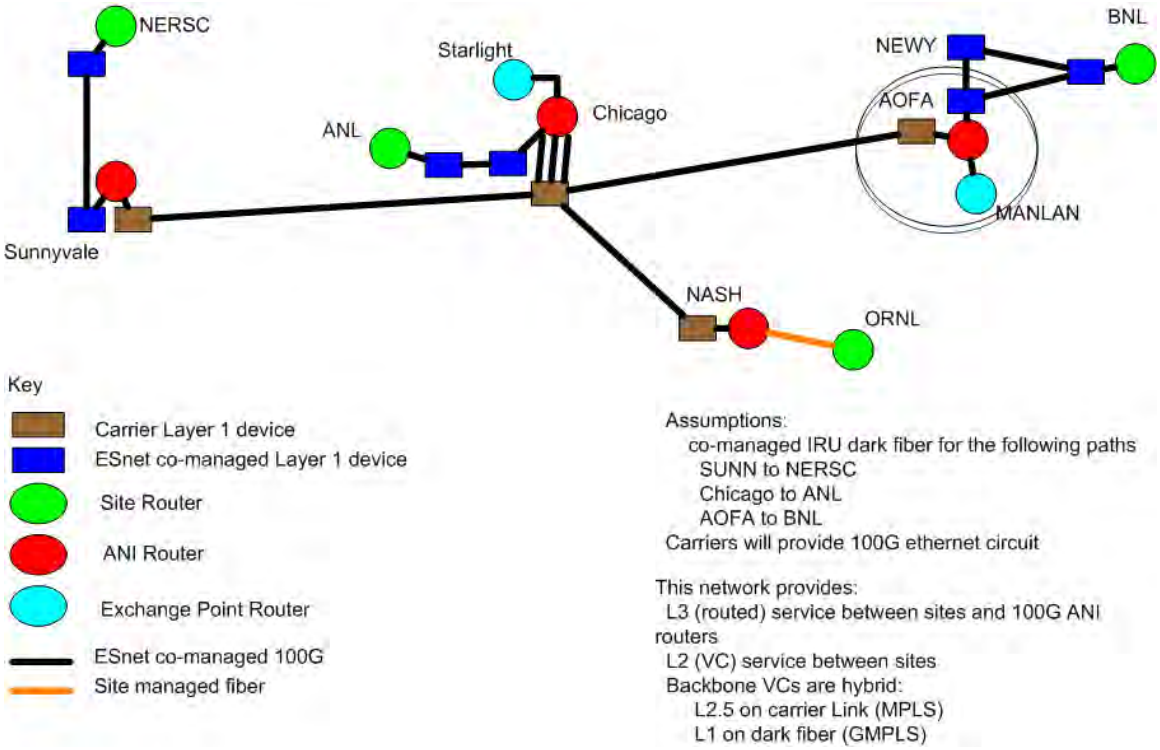

Figure 7 ANI Logical Testbed Topology

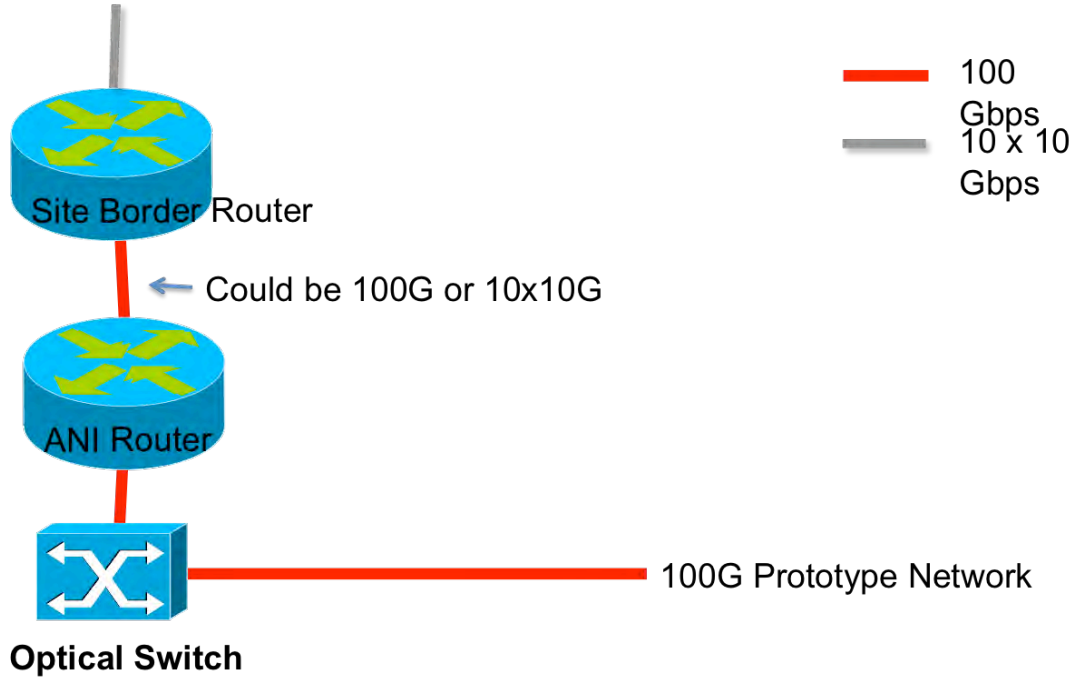

Figure 8 Typical Site Configuration

Testbed Timeline:

February 2010: Initial 'table top' node hardware setup at LBL
September 2010: 10 Gbps table top testbed available to researchers
April 2011: 10/100 Gbps WAN testbed available to researchers
January 2012: full 100 Gbps WAN testbed available to researchers

Summary: Advanced Networking Initiative Component Goals

Component-1 Goal:

- Demonstrate between prototype backbone hub locations a 100 gigabit per second optical transport capability.

Component-2 Goals:

- Provide research and development on technical and operational challenges of deploying a point-to-point prototype infrastructure tying the nationwide ANI 100 Gbps developmental prototype backbone to the BayExpress.
- Demonstrate an aggregate bandwidth capacity of 100 Gpbs between NERSC demark and the 100 Gpbs developmental prototype backbone.

Component-3 Goals:

- Provide research and development on technical and operational challenges of deploying a point-to-point prototype infrastructure tying the nationwide ANI 100 Gbps developmental prototype backbone to the ChiExpress.
- Demonstrate an aggregate bandwidth capacity of 100 Gbps between ANL demark and the 100 Gbps capable developmental prototype backbone.

Component-4 Goals:

- Provide research and development on technical and operational challenges of deploying a point-to-point prototype infrastructure tying the nationwide ANI 100 Gbps demonstration prototype backbone to the NYExpress.
- Demonstrate an aggregate bandwidth capacity of 100 Gbps between the MANLAN demark and the 100 Gbps capable demonstration prototype backbone.

Component-5 Goals:

- Provide research and development on technical and operational challenges of deploying a point-to-point prototype infrastructure tying the nationwide ANI 100 Gbps developmental prototype backbone to the NashExpress.
- Demonstrate an aggregate bandwidth capacity of 100 Gbps between the ORNL demark and the 100 Gbps capable developmental prototype backbone.

Component-6 Goals:

- Create an experimental network research testbed environment where researchers can prototype dynamic provisioning of multiple-layers and multiple-domain hybrid packet/circuit-switched networks.
- Provide researchers with a realistic layer 1 and 2 environment that can be flexibly configured for developing and testing layer 1 and layer 2 control plane software that manages the configuration and connectivity of the testbed switching and routing devices.
- Provide researchers with a realistic environment to prototype and test scenarios from Virtual Facilities and Virtual Organizations (e.g. Earth System Grid, Open Science Grid) that reflect near-term requirements for large-scale science. This includes scaling middleware and software stacks to 100 Gbps.

Overall Project Goal:

- Demonstrate end-to-end 100 Gbps capability from an aggregation switch/router directly connected to hosts at one facility's Magellan compute cluster, across the wide area network, to a similar aggregation switch/router located at a second Magellan facility.