North East Cyberinfrastructure Consortium (NECC) Project Summary

The NSF EPSCoR northeast region (Maine, New Hampshire, Vermont, Rhode Island and Delaware) represents a broad range of cyberinfrastructure assets and needs. The three northernmost states are referred to as a **black hole of connectivity** because ME, NH and VT are not served by facilities-based optical networks in spite of their proximity to backbone connections in surrounding states and Canada. RI, despite its size, has extremely poor connectivity in the southern part of the state. DE is better connected to fiber networks, but requires additional fiber, hardware upgrades and personnel to take maximum advantage of existing and planned collaborations. To develop resilient, high-bandwidth connectivity between research and academic institutions in ME, RI, NH, VT and DE, we formed the North East Cyberinfrastructure Consortium (NECC) in 2006 and initiated two related collaborative efforts to ***identify and promote the shared use of research facilities*** across the region and ***assess and address cyber-infrastructure needs***. This proposal is a result of these efforts. The majority of the proposed budget necessarily will go towards the fiber network so that consortium members can take full advantage of national and international cyber resources for research. However, we are also supporting pilot projects in cyber-enabled research through a virtual organization we have created for the sharing of expertise and facilities in the region, the North East Bioinformatics Collaborative (NEBC). Therefore, while we *do not* propose a single-themed NECC cyber-enabled research project, we present a detailed plan for a fiber network, Data Centers and a highly-collaborative, regional pilot research effort that relies upon a fiber network and regionally-distributed data expertise.

The NEBC is a virtual organization for the support of cyber-enabled research that require data analyses of large datasets. A demonstration project last year used genomics data collected in Maine, stored in Delaware and annotated in Vermont to explore the efficacy of a regional datacenter model. We propose to use the NEBC to facilitate another pilot project that will provide the metagenomes from "NextGen" sequencing of microbial populations in water samples from three stages of cyanobacterial blooms in five lakes in the NE states. The NEBC will work remotely on the resulting large data sets to determine the metagenomes of microbe communities in the blooms. The purposes of this venture, in addition to collecting pilot data for future funding, are to develop the collaborative activities of the NEBC, and, promote the development of protocols at the new Data Centers for the movement, life cycle management, storage and recovery of data that are simultaneously viewed/analyzed/worked on by multiple users across the region. We will encourage further cyber-enabled research on the pollution in the fresh and marine waters of NE through pilot awards.

The NECC states have organized as a region around our water outreach programs for workforce development and broadening participation. This Track-2 proposal has given us the opportunity to create a network of statewide watershed programs so that we can collaborate, exchange students, and work together from remote locations in ways that currently are not possible. The proposed fiber network will enable videoconferencing, communication and cyber-tools training that are all crucial for this effort.

The need for a fiber network for research and education is great in the NE, as indicated by the support of governors, EPSCoR state committees, university leadership, representatives and senators (including the Vice President Elect) (see letters of support in the Supplemental Information). Their support has made it possible to leverage the costs of the fiber network, which otherwise would not be feasible through a mechanism like Track-2.

Intellectual Merit: The consortium has three primary needs to support regional, cyber-enabled research: 1) long-term leases on fiber in specific reaches across the northeast to provide high-speed connectivity with dense-wave division capability; 2) redundant, distributed Data Centers for regional cyber-enabled collaborations; and 3) cyber-knowledgeable personnel to allow researchers to access regional compute, analysis and visualization resources. Much of the physical infrastructure required for the NECC network exists, but there are four key reaches of fiber needed in Maine, New Hampshire, Rhode Island and Vermont. In Maine, two stretches are required to provide a redundant route for national and international connectivity through CANARIE and along the I-95 corridor. A fiber route along the I-89 corridor provides connectivity to Boston for Vermont and New Hampshire to Boston. In Rhode Island, our proposal extends high-speed connectivity to URI. The EPSCoR Committees and EPSCoR PDs of each state have worked to assure that the proposed network fits with the State Science and Technology plans. We have been working with the Northeast Research and Education Network

0918284

(NEREN) to manage the fiber network once it is in place.

Cyber-knowledgeable personnel represent one of the most critical parts of ensuring the long-term success of this project. We have identified, in each NECC state, the critical IT personnel who are needed to implement shared data systems, manage computational resources and provide support to enable regional collaborations. These personnel are in addition to the NEBC members who facilitate data analysis. All the requested personnel are leveraged against state, matching or other grant funds in order to build the collective personnel expertise that we need in the NE.

Two regionally distributed ***Data Centers***, one at the University of Delaware in Newark, DE, the other at the University of Maine at Orono, ME, provide the seats for regional collaborations and are central to the regional network. The Data Centers will provide life cycle management and provenance of large data sets for cyber-enabled discovery that requires highly available mass storage with concomitant compute services and the requisite support infrastructure. The Data Centers will provide consistent and highly-available storage for data that are simultaneously accessed or updated, recovery to a consistent state after hardware, software or user failures and support for efficient *ad hoc* queries. These centers are required to share and analyze large datasets created by next-generation sequencers and the planned cyber-enabled projects. We recognize that the long-term success will depend on the ability to scale rapidly without degradation of performance. Development of working policies that anticipate this need will be critical. The Data Center personnel will develop these policies in conjunction with technical experts from the NECC member institutions. In the event of a catastrophic event (e.g. network outage or fire) in the primary Data Center a transparent failover to the secondary center will provide continued uninterrupted access to data for all researchers. While the ME and DE centers provide physical redundancy, the research and cyber-IT expertise of personnel at the DE and ME datacenters are complementary.

What research will the fiber network, Data Centers and personnel facilitate?

Metagenomics: In this proposal, we enumerate some of the funded research in each NECC member state that would benefit from the new fiber network. However, as a demonstration of our regional organization around collaborative cyber-enabled research, we propose a *pilot* project requiring cyberinfrastructure in an area of common scientific interest and large regional impact. It would be disingenuous at this time to propose a more comprehensive project given our lack of cyber-infrastructure, most of the budget for the Track-2 award is required to establish the fiber network.

We propose a pilot project to determine the metagenomes of cyanobacterial blooms in lakes in the NE. These blooms and subsequent production of toxins by the cyanobacteria impact the health of our residents and animals and the economies of our states, but the cause of the blooms and the triggering of toxin production remain unresolved. We need fundamental new knowledge and an important new insight will be an understanding of the diversity of phytoplankton in stages of the bloom through an unbiased census.

We will use a metagenomics approach to conduct a census and taxonomic study of microbial communities by "NextGen" sequencing of DNA obtained from cyanobacteria in five lakes in VT, NH, ME and RI during three stages of cyanobacterial blooms. Metagenomics provides a composite snapshot of the population and an unprecedented insight into micro-heterogeneity and identification of species that are not expected to be in the sample. Metagenomics allows identification of microorganisms that are not easily cultured, without guessing at which ones are present, such as in sea samples or gut flora. The resulting datasets are large and rich in information about microbial communities. Results from this sequencing project will be stored in a central database. This shared resource, implemented through the NECC Data Centers, will provide live storage and archive services while allowing easy sharing of the data among researchers in the NECC for further analysis. Members of the NEBC and other NECC researchers will analyze the resulting datasets remotely while working together to develop analysis methodologies and best practices. Resulting data will be the basis of publications and proposals for extending the project through extramural funds.

Future projects: We recognize that the NECC states have significant water environmental research that spans streams, regional watersheds, coastal studies and ocean science. Researchers in these areas would benefit from improved cyber-infrastructure. To prepare for future cyber-enabled water research endeavors, we have formed a regional working group tasked with identifying and facilitating collaborative cyber-enabled research. We propose to make four pilot awards to projects that are regional, collaborative and cyber-enabled. The Data Centers and NEBC will provide the expertise that NECC

researchers need to take advantage of the network for water research and is not currently available.

Broader Impacts: The possibility of a fiber network that would provide adequate bandwidth for videoconferencing has led to our NECC regional organization around ***outreach programs for STEM workforce development and diversity.*** We will create a new ***Watershed Project*** through partnerships among multiple state-based programs for high school and undergraduate students. Students in this project from all the NECC states, NY and Puerto Rico, who otherwise would not even meet, will work together in collaborative watershed research. We will create a large and diverse group of students by including Rhode Island's and Delaware's minority-serving high schools and diverse college populations; Maine's and upper New York State's Native American serving schools; Bronx high schools and Puerto Rico high schools and colleges that have exchange programs with Vermont. Participants will train together in watershed sampling, sample analysis, data base creation and modeling, and use of cyber-tools in watershed science. Following training, teams of high school students and teachers or undergraduates join with state programs to work on watershed science during the summer or through summer and academic year. The participants will reconvene in an NECC state to present their results and draft reports. The entire group will meet through multi-media videoconferencing to learn about databases, modeling, remote sensing and other cyber-based tools, and participate in STEM career opportunity panels. The individual NECC state programs are effective in improving participation in STEM majors and diversity, but with the ***new fiber network and the ability to communicate*** over the new cyber-network, we will be able to make a larger, region-wide effective program, with emphasis on cyber-based communication and research tools that will give students hands-on experience in 21stst century global workforce best practices.

We propose a multi-faceted ***communication plan*** that will spread the word about the importance of our cyber-enabled research to the public through innovative television shows, podcasts and educational materials. We also describe an ***Ambassador Program*** through which we will partner with ***citizen science*** groups to inform the public about the importance of a fiber network to education and science and about the potential impact of the cyber-enabled metagenomics study to the economies of our states. We anticipate that our citizen science partners will help us to make the abstract topic of cyberinfrastructure relevant to the lay public. The academic leadership in our states understands the impact on higher and K-12 education and has provided enthusiastic letters of support.

The fiber network will have an ***enormous economic*** impact on our region, which is in large part why we have the enthusiastic support of governors, senators, the Vice President Elect and EPSCoR state committees. Once fiber is available from Burlington, VT to Boston, MA and from URI to Boston, the vendors that provide the Indefeasible Rights of Usage (IRUs) will be able to make broadband connections more available to academic institutions and the private sector. Moreover, the proposed cyber-enabled research will have positive economic impacts as well. The pilot project on metagenomics of the bacterial communities in blooms in lakes in Vermont, New Hampshire, Maine and Rode Island will contribute to the understanding of the origin of these blooms and their toxins that shut down access to recreational and drinking water sources. Lakes in the NE are extremely important to our economies, with estimates of $1.5B in lake-related revenues to NY, VT and Quebec each year from Lake Champlain; $2B annually from lake recreational revenues to Maine; 14,000 jobs and bring in $1.8B in revenues from boating, fishing, swimming, drinking water and property taxes to New Hampshire. For the Lakes Region of NH alone a perceived loss of clarity or purity would result in a potential loss of $25 million in direct sales, $8.8 million in income and almost 400 jobs.

Finally, the ***private sector*** users of large data sets in our NECC states will benefit immediately from the new fiber network. We describe an example of one such opportunity in Vermont.

Summary: In summary, we present a proposal for a fiber network with distributed Data Centers and cyber-knowledgeable personnel, all of which are highly leveraged in order to maximize the impact of the proposed effort. The proposal comes from a regional group that is organized around sharing resources, expertise and facilities in order to make cyber-enabled collaborative research possible in a sparsely populated region and among non-contiguous states. We propose a regional pilot project that will initiate our use of regional Data Centers, provide a proof-of-concept for cyber-enabled research in the NE and provide important data sets that will be the basis for future funding. The proposed fiber network allows us to design a distributed, collaborative outreach project for STEM workforce development and diversity.