NESE: The North East Storage Exchange

James Cuff[1], John Goodhue[2], Saul Youssef[3], Rajiv Shridhar[4], Ralph Zottola[5], Chris Hill[6], Glenn Bresnahan[3]

Harvard[1], MGHPCC[2], Boston University[3], Northeastern University[4], University of Massachusetts[5], MIT[6]

Abstract

Research progress is increasingly dependent upon the available capacity of storage to flexibly exploit large volumes of digital information. The North East Storage Exchange (NESE) project creates a next-generation storage infrastructure specifically targeted at enabling new levels of collaborative research in projects regularly involving petabytes of information. This storage exchange will integrate with a computational and network infrastructure that links Harvard University, Boston University, the Massachusetts Institute of Technology (MIT), Northeastern University and the University of Massachusetts system. This project contributes to building a national data infrastructure to support advanced research in such priority topics as health care, epidemiology, physics, and earth science, among others. NESE will provide a high capacity, highly networked, secure, cost effective, scalable, and accessible data store that lowers barriers to research, collaboration, and information sharing within and beyond the participating multi-university community. Some examples of NESE projects that will be early users of NESE include one of the four US Tier 2 centers that store and process ATLAS data from the Large Hadron Collider; the Center for Brain Science at Harvard University, which is generating 300 million micron-resolution images to map the billion neurons and synapses that make up a cubic millimeter of the human brain; and MIT collaborations with NASA and DARPA in next generation global ocean modeling and monitoring systems. NESE addresses several critical infrastructural challenges: the creation of a sustainable multi-institutional resource; advancement of methods for data retention, management, and access to sensitive research data; implementation of controls that simplify protection of sensitive data; and building a sustainable, collaborative operating infrastructure to support future research.

Guiding Principles

1. **SECURE**: As consent based research data sets become standard practice (in particular within Health Science), security models are having to catch up with the Data Use Agreements required. From dbGap, to CMS/Medicare, our researchers manage significantly more human and health care subject data than ever before. For societal change to occur, and to produce better outcomes for patients through research and basic science, we need significantly more performant and secure data storage systems. We can't do this alone, or in isolation.
2. **ARCHIVE**: Scientists and researchers discuss data retention, archive and provenance on what seems to be a daily basis. We have multiple solutions to this challenge, but no unified overarching system that we can point to as a "standard". As funding agencies require more sophisticated "Data Management Plans", our research faculty are left with a bewildering array of options, each more confusing than the last. This has to stop.
3. **COST**: Storage is expensive. Many hundreds of millions of dollars are spent annually attempting to solve the challenge of reliable, available storage for science. The potential for economies of scale by collecting and coordinating resources here in what could well be argued as the most research data intensive part of the nation is vast. We are capable, and have proven by MGHPCC that we can do more with less. Much more.
4. **CAPACITY:** We have heard this for many years now - there is quite simply an explosion of data in science, it is not being managed, and this proposal points to both technology and process to be able to manage unlimited capacity requirements.
5. **BANDWIDTH:** Science data requirements demand high performance storage. It is not sufficient to simply provide large capacity, as data access patterns vary dramatically across disciplines, and each NSF directorate. Fortunately, "object stores" (the technology we will deploy as part of NESE), are inherently designed to scale out for both speed and capacity.