ProjectNESE

The North East Storage Exchange

Dr. James Cuff

Assistant Dean & Distinguished Engineer for Research Computing
PI for ProjectNESE

Harvard University

7th National Data Service Consortium Workshop
Chicago, Illinois

April 14th 2017

projectnese.org

@projectnese

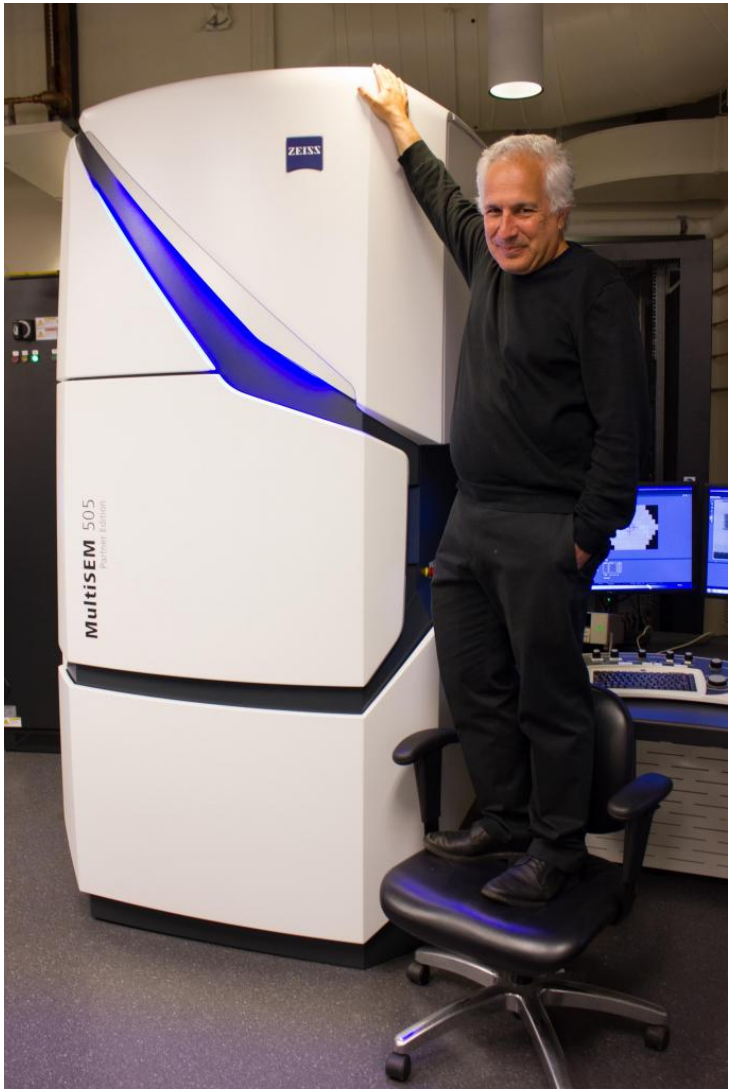


3TB

EVERY

SINGLE

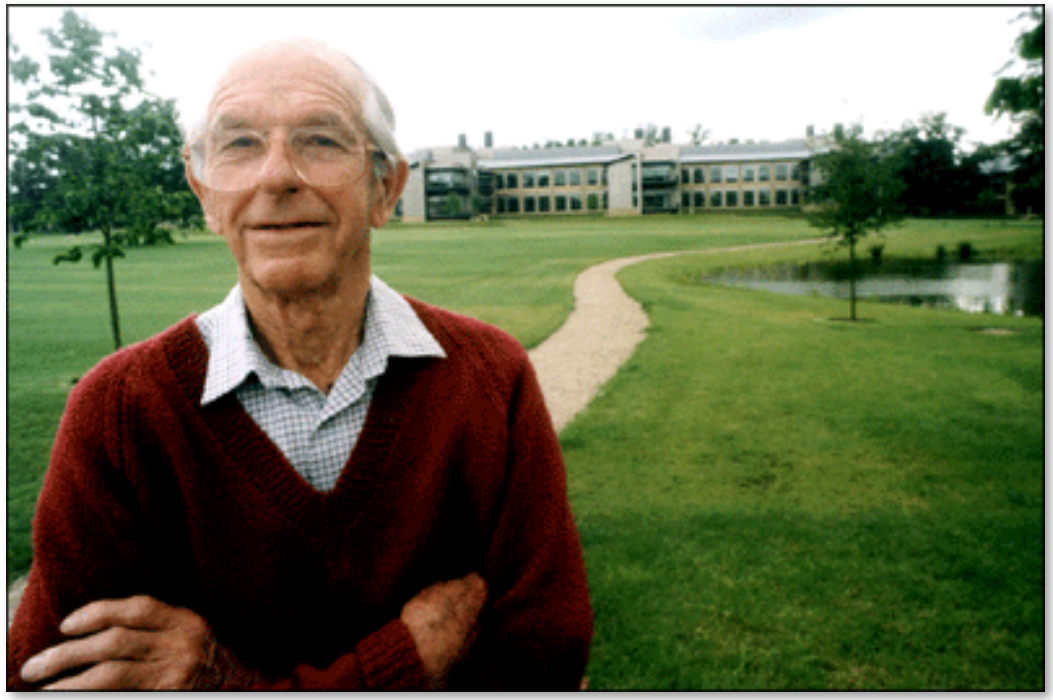
HOUR




2.8P raw…

This **SINGLE** microscope generates an **ENTIRE**

Year 2000 Sanger Institute **EVERY HOUR**!



No Room!

Showing results in **Electronics**. Show instead results in All Departments.

Related Searches: nas, synology.

See Capacity Options

Drobo 5N: Network Attached Storage - 5 bay array with mSATA SSD acceleration - Gigabit Ethernet port (DRDS4A21)
by Drobo

$489.00 ~~$499.00~~ Prime
Get it by **Tomorrow, May 14**

More Buying Choices
$489.00 new (63 offers)
$449.99 used (4 offers)
335

See Capacity Options

Drobo Gen 3: Direct Attached Storage - 4 bay array - USB 3 port. (DDR3A21)
by Drobo

$259.99 Prime
Get it by **Tomorrow, May 14**

More Buying Choices
$259.99 new (57 offers)
120

See Capacity Options

See Capacity Options

Drobo mini: Direct Attached Storage - 4 bay array - USB3 and Thunderbolt ports - Designed for portability. (DR-MINI-1A21)
by Drobo

$623.85 ~~$649.00~~

More Buying Choices
$623.85 new (53 offers)
$433.95 used (5 offers)

FREE Shipping on eligible orders
86

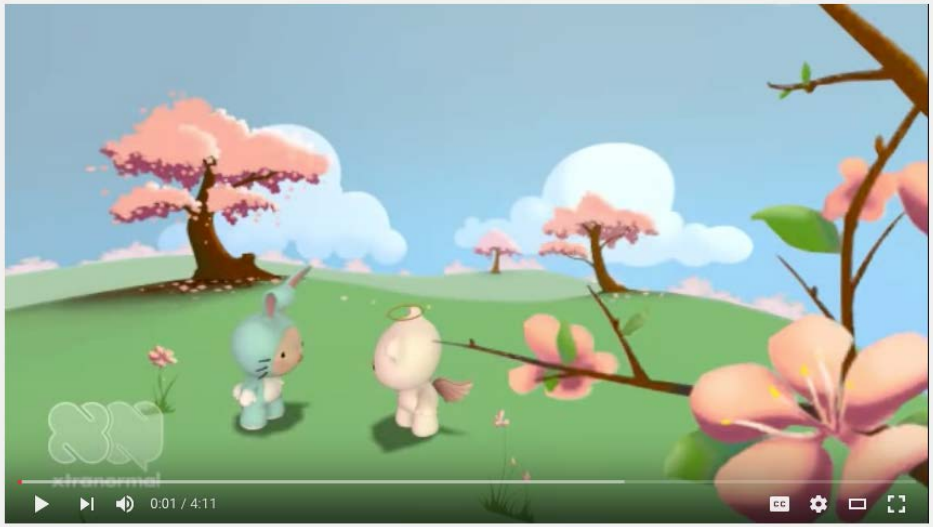
See Size Options

Drobo B810n 8-Drive Network Attached Storage (NAS) Array, Gigabit Ethernet x 2 ports (DR-B810N-5A21)
by Drobo

$1,576.99 ~~$1,599.00~~ Prime
Only 15 left in stock - order soon.

More Buying Choices
$1,576.99 new (24 offers)
4

YouTube Search



a cautionary tale

jamesdotcuff
Subscribe 9

12,257 views

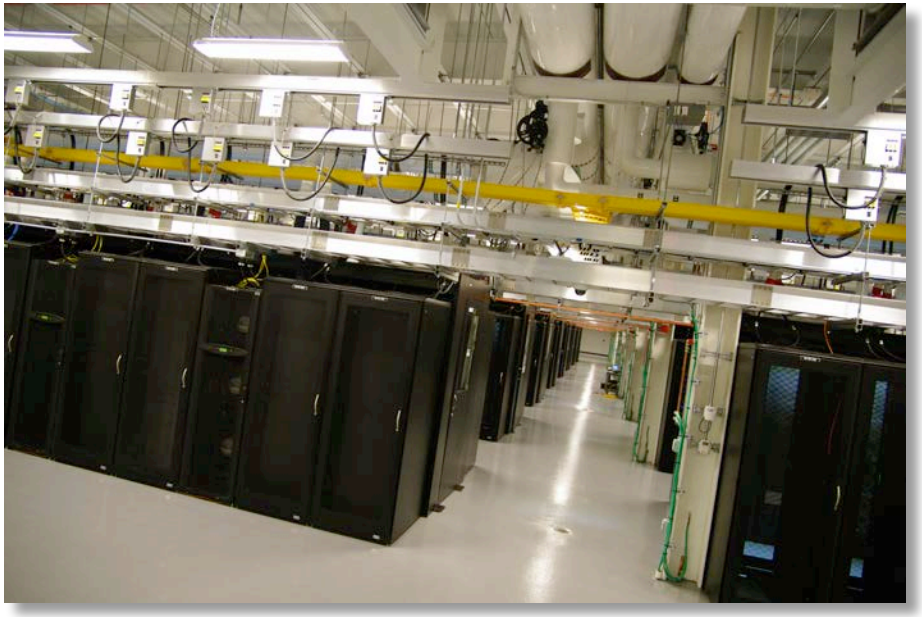
Add to Share More 85 3

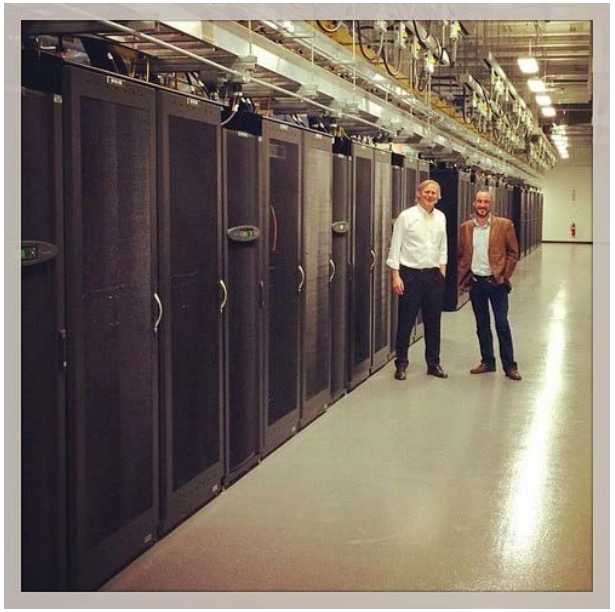
Uploaded on Dec 22, 2011
Storage... storage storage ;-)

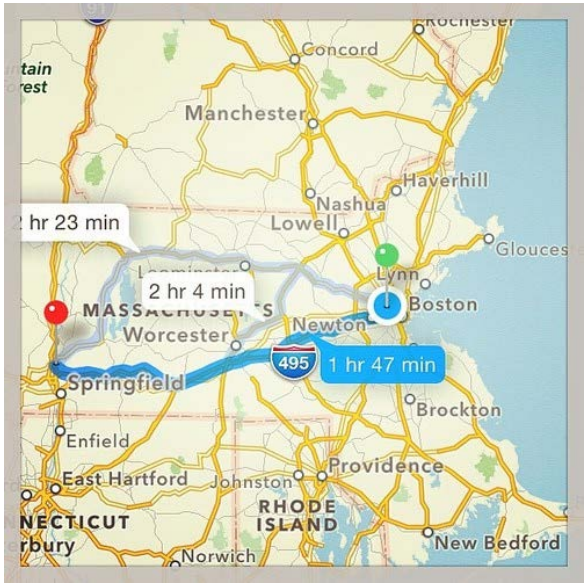
SHOW MORE

https://www.youtube.com/user/jamesdotcuff

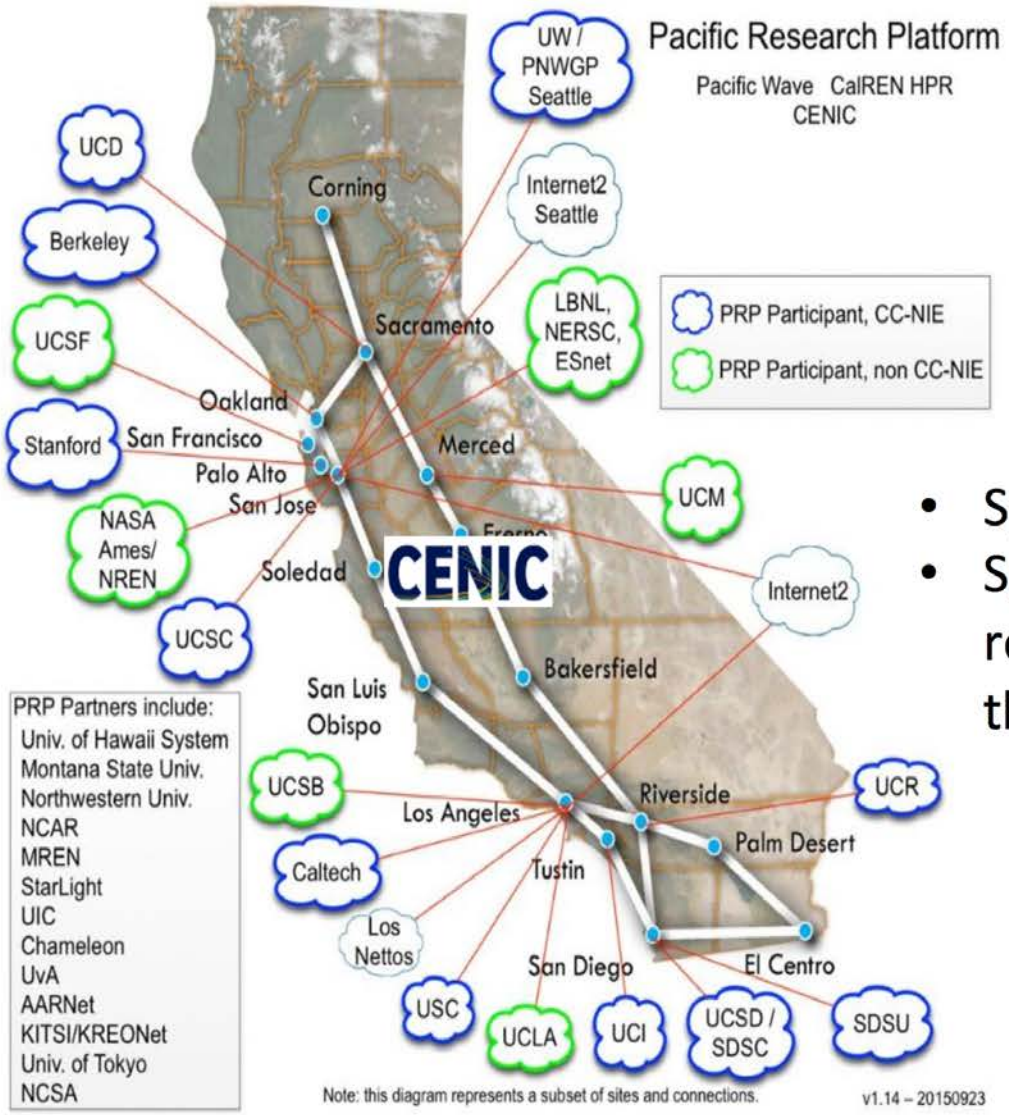
BUILD UPON PRIOR EXCELLENCE!



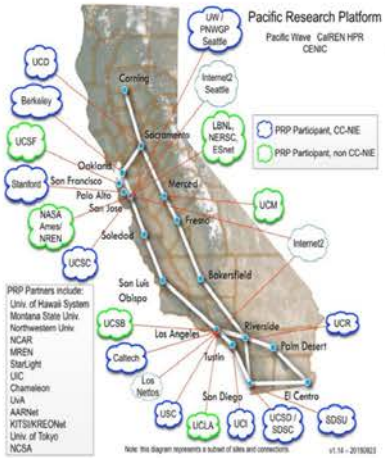



Concord
Manchester
Haverhill
Nashua
Lowell
hr 23 min
Glouces
Lynn
2 hr 4 min
Boston
Worcester
Newton
495
1 hr 47 min
Springfield
Brockton
Enfield
Providence
East Hartford
Johnston
RHODE ISLAND
New Bedford
Norwich

The Pacific Research Platform



- Spans the pacific coast
- Spans top universities and research institutions in the U.S. and the world

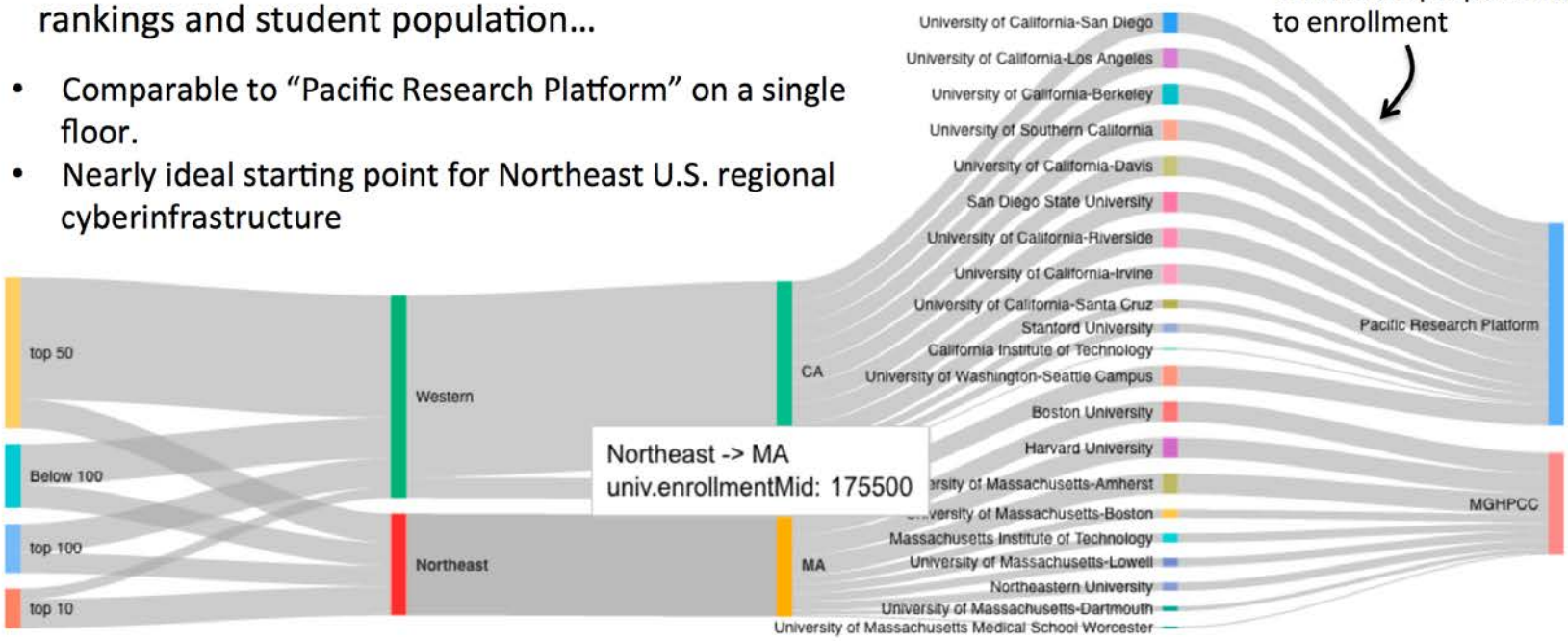


Imagine shrinking PRP down to the size of a building and compare with MGHPCC

Even though we are comparable to PRP in university rankings and student population…

- Comparable to "Pacific Research Platform" on a single floor.
- Nearly ideal starting point for Northeast U.S. regional cyberinfrastructure

Widths are proportional to enrollment



National Science Foundation
WHERE DISCOVERIES BEGIN

SEARCH

HOME | RESEARCH AREAS | FUNDING | AWARDS | DOCUMENT LIBRARY | NEWS | ABOUT NSF

Awards

- Search Awards
- Recent Awards
- Presidential and Honorary Awards
- About Awards

How to Manage Your Award

- Grant Policy Manual
- Grant General Conditions
- Cooperative Agreement Conditions
- Special Conditions
- Federal Demonstration Partnership
- Policy Office Website

Award Abstract #1640831

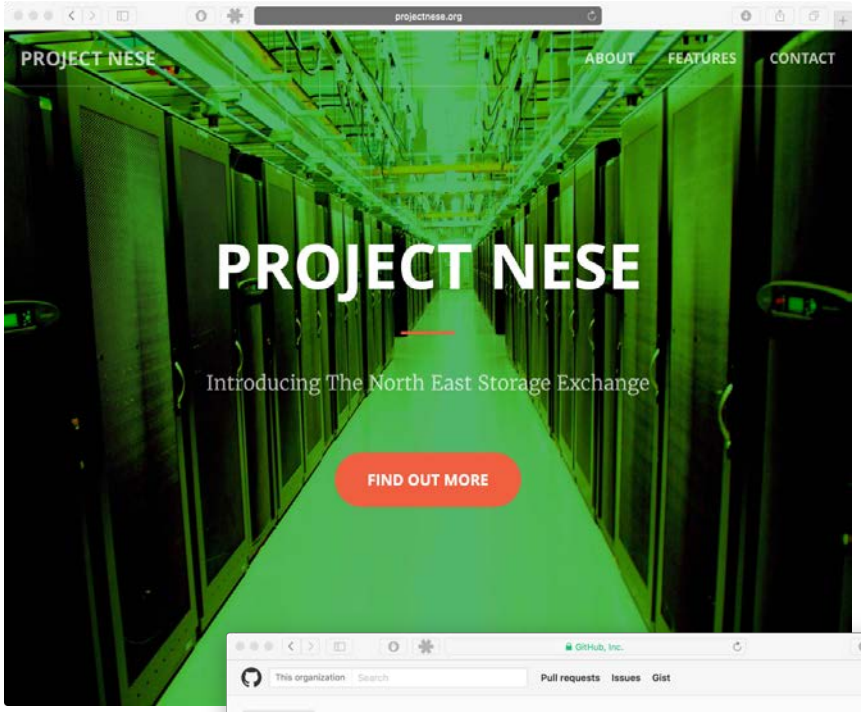
CIF21 DIBBs: EI: North East Storage Exchange

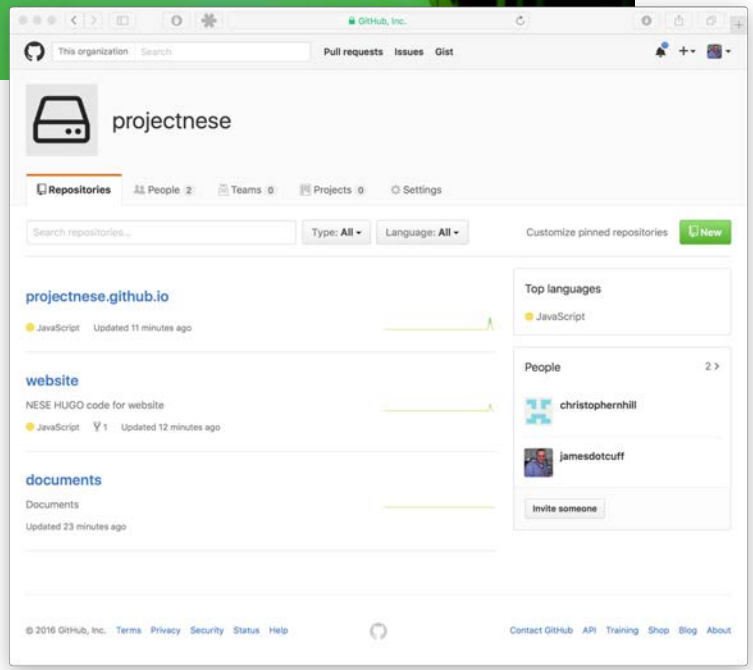
NSF Org:	ACI Div Of Advanced Cyberinfrastructure
Initial Amendment Date:	August 1, 2016
Latest Amendment Date:	August 1, 2016
Award Number:	1640831
Award Instrument:	Standard Grant
Program Manager:	Amy Walton ACI Div Of Advanced Cyberinfrastructure CSE Direct For Computer & Info Scie & Enginr
Start Date:	November 1, 2016
End Date:	October 31, 2021 (Estimated)
Awarded Amount to Date:	$3,973,963.00
Investigator(s):	James Cuff james_cuff@harvard.edu (Principal Investigator) John Goodhue (Co-Principal Investigator) Saul Youssef (Co-Principal Investigator) RAJIV SHRIDHAR (Co-Principal Investigator) Ralph Zottola (Co-Principal Investigator)
Sponsor:	Harvard University 1033 MASSACHUSETTS AVE Cambridge, MA 02138-5366 (617)495-5501
NSF Program(s):	ADVANCES IN BIO INFORMATICS, ETF, DATANET
Program Reference Code(s):	7433, 8048, 8089, 8091
Program Element Code(s):	1165, 7476, 7726

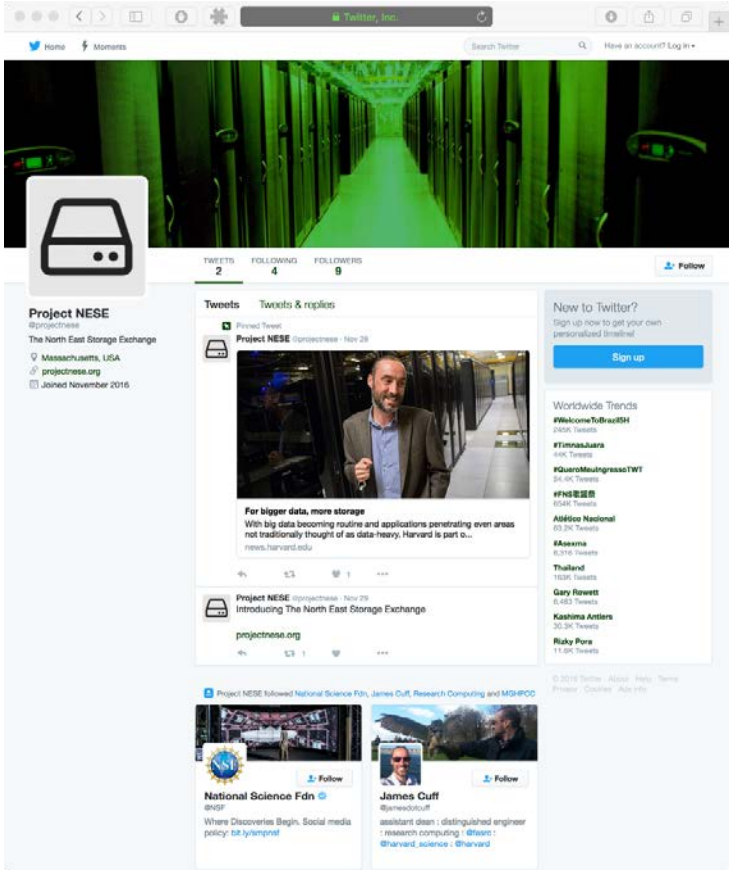
NESE will provide a high capacity, highly networked, secure, cost effective, scalable, and accessible data store that lowers barriers to research, collaboration, and information sharing within and beyond the participating multi-university community.

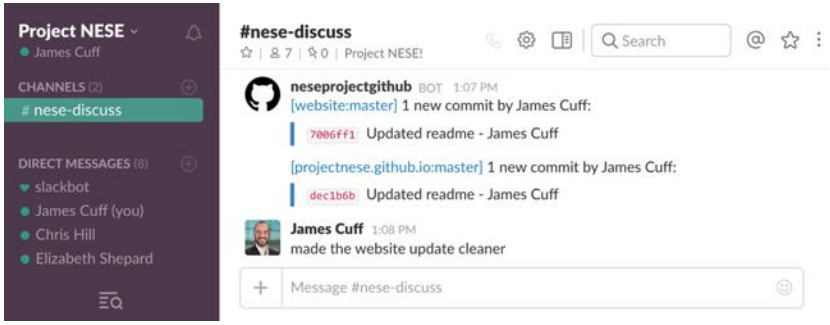
FIVE key areas:

- **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.
- **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.
- **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.
- **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.
- **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.


projectnese.org
PROJECT NESE
ABOUT
FEATURES
CONTACT
PROJECT NESE
Introducing The North East Storage Exchange
FIND OUT MORE


GitHub, Inc.
This organization
Search
Pull requests
Issues
Gist
projectnese
Repositories
People 2
Teams 0
Projects 0
Settings
Search repositories...
Type: All
Language: All
Customize pinned repositories
New
projectnese.github.io
JavaScript
Updated 11 minutes ago
website
NESE HUGO code for website
JavaScript
Updated 12 minutes ago
documents
Documents
Updated 23 minutes ago
Top languages
JavaScript
People
2
christophernhill
jamesdotcuff
Invite someone
© 2016 GitHub, Inc.
Terms
Privacy
Security
Status
Help
Contact GitHub
API
Training
Shop
Blog
About


Twitter, Inc.
Home
Moments
Search Twitter
Have an account? Log in
TWEETS 2
FOLLOWING 4
FOLLOWERS 9
Follow
Project NESE
@projectnese
The North East Storage Exchange
Massachusetts, USA
projectnese.org
Joined November 2016
Tweets
Tweets & replies
Pinned Tweet
Project NESE @projectnese · Nov 29
For bigger data, more storage
With big data becoming routine and applications penetrating even areas not traditionally thought of as data-heavy, Harvard is part o...
news.harvard.edu
Project NESE @projectnese · Nov 29
Introducing The North East Storage Exchange
projectnese.org
Project NESE followed National Science Fdn, James Cuff, Research Computing and MGHPCC
National Science Fdn
@NSF
Where Discoveries Begin. Social media policy: bit.ly/simpnsf
James Cuff
@jamesdotcuff
assistant dean : distinguished engineer : research computing : @fasrc : @harvard_science : @harvard
New to Twitter?
Sign up now to get your own personalized timeline!
Sign up
Worldwide Trends
#WelcomeToBrazilSH
#TimnasJuara
#QueroMeuIngressoTWT
#FNS歌謡祭
Atlético Nacional
#Asexma
Thailand
Gary Rowett
Kashima Antlers
Ricky Pora


Project NESE
James Cuff
CHANNELS (2)
nese-discuss
DIRECT MESSAGES (8)
slackbot
James Cuff (you)
Chris Hill
Elizabeth Shepard
#nese-discuss
7 | 0 | Project NESE!
Search
neseprojectgithub BOT 1:07 PM
[website:master] 1 new commit by James Cuff:
7006ff1 Updated readme - James Cuff
[projectnese.github.io:master] 1 new commit by James Cuff:
dec1b6b Updated readme - James Cuff
James Cuff 1:08 PM
made the website update cleaner
Message #nese-discuss

Meet the team

- James Cuff (Harvard)
- Saul Youssef (BU)
- John Goodhue (MGHPCC)
- Rajiv Shridhar (Northeastern)
- Ralph Zottola (University of Massachusetts)

- Glenn Bresnahan (BU)
- Scott Yockel (Harvard)
- Jim Culbert (MGHPCC)
- Chris Hill (MIT)

- And….

Wes Dillingham – Senior Cyberinfrastructure Storage Engineer, ProjectNESE

Experience

Cloud Infrastructure Engineer

January '15 – Present **Harvard University - Research Computing • Cambridge, MA**

Infrastructure / Systems engineer on a team of three with the primary focus of developing a private cloud (OpenNebula) platform for internal (co-workers) and external (faculty) consumption. Was personally the service owner, domain and organization expert in Ceph, our backend distributed storage platform for our virtual infrastructure - where we utilized CephFS, RBD, RADOS, and CRUSH in production. Wrote Puppet manifests to provision and maintain Ceph in a geographically distributed and redundant architecture. Implemented community gathered web application for monitoring Ceph clusters, deployed diamond collectors across the cluster to gather individual and cluster wide metrics. Wrote custom Nagios checks for monitoring Ceph / OpenNebula clusters. Implemented automatic incremental backup procedure for RBD devices. Had secondary responsibilities of supporting Harvard's largest High Performance compute cluster (Odyssey) for faculty research consumption.

Linux & Unix Systems Administrator

January '13 – January '15 **Channing Laboratory of Brigham and Women's Hospital • Boston, MA**

Systems administrator / engineer on a team developing a hybrid, public/private, cloud architecture for researchers investigating the genetic basis of complex diseases. Leveraging Eucalyptus, Amazon Elastic Compute Cloud, GlusterFS, Chef configuration management, STAR:Cluster, Grid engine, and lots of glue in the middle. Served as the main developer of our chef configuration repository and primary administrator of our Eucalyptus private cloud infrastructure. Worked in a "DevOps" capacity with a python developer building the web app our researchers use to provision our hybrid cloud. Managed organization-wide GitHub enterprise appliance. Assisted researchers with the usage of our grid engine infrastructure and served as technical support for Linux/Unix tickets. Interfaced with other internal systems and networking teams.

Junior Linux Systems Administrator

October '11 – December '12 **Tufts University • Medford, MA**

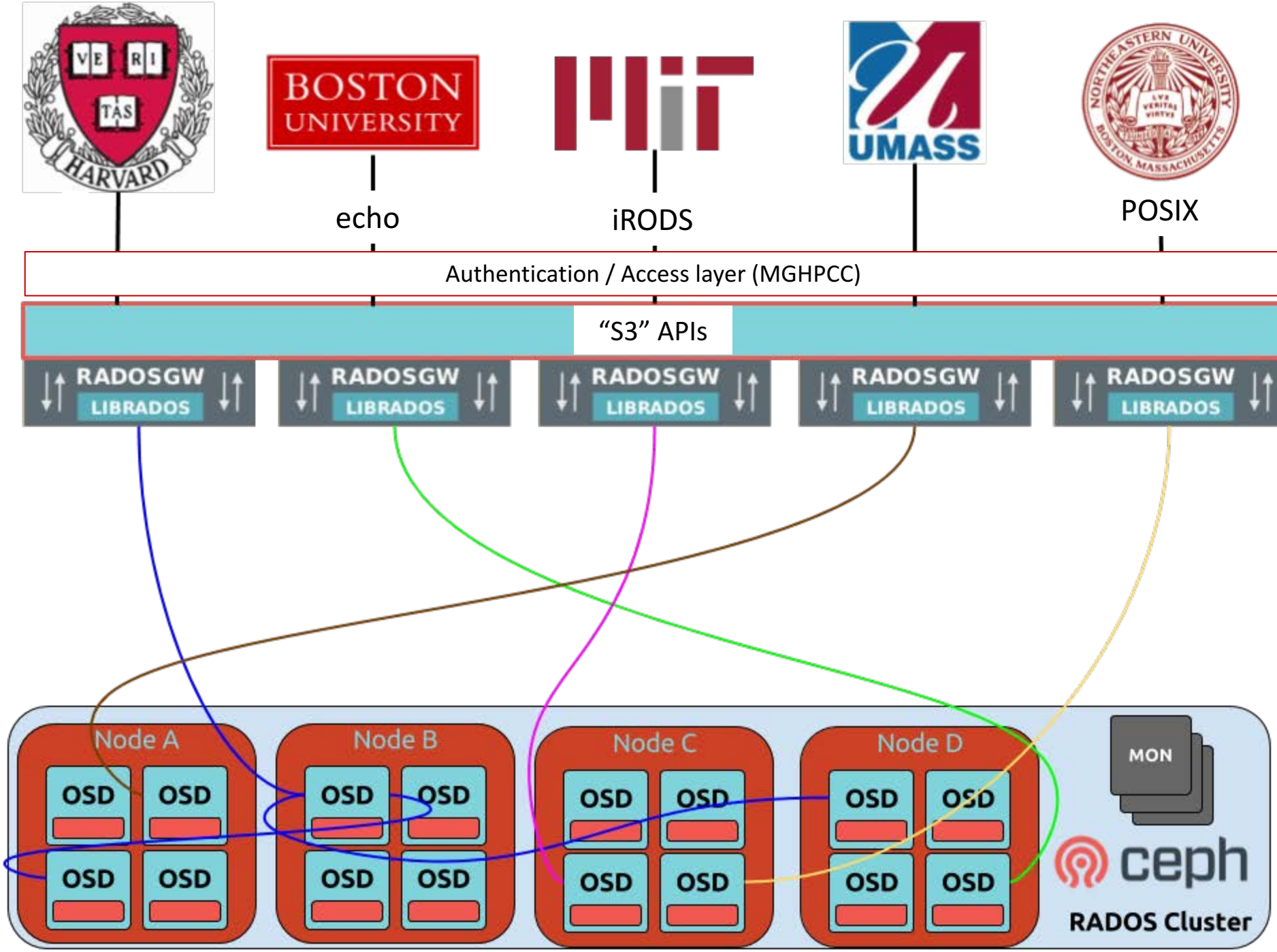
Assisted faculty and graduate students in the Arts, Sciences, & Engineering departments utilize on-campus Linux resources, including a high performance compute cluster. Administered Linux workstations and servers. Wrote scripts to automate processes. Built and administered 40 TB mdadm raid array. Setup drupal and wordpress instances. Configured kerberos for usage with government run systems. Configured NAS devices to automount to computer cluster. Maintained a multi-tiered system security policy on all administered machines.

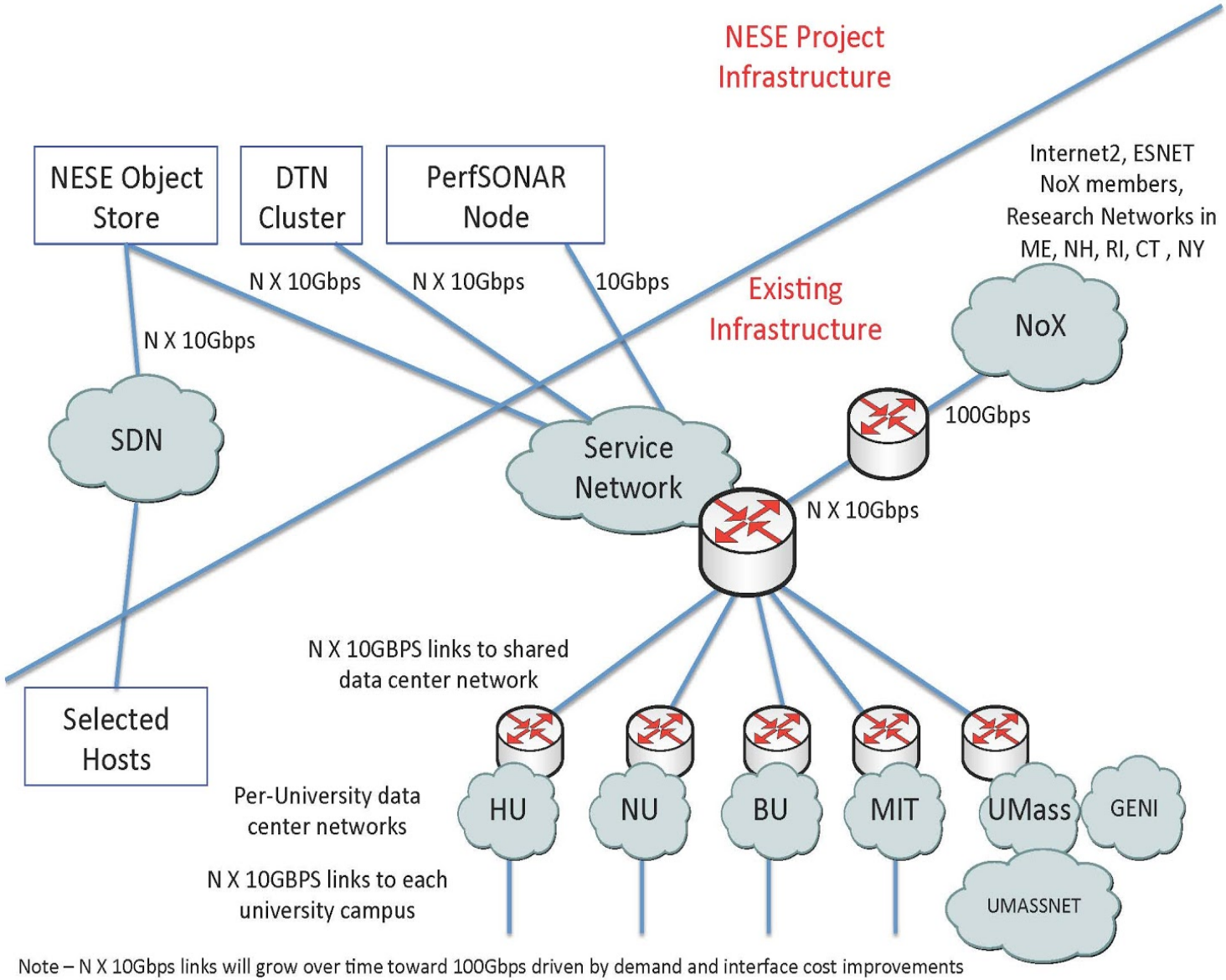
Information Technology Intern

June '10 – March '11 **Savant Systems • Hyannis, MA**

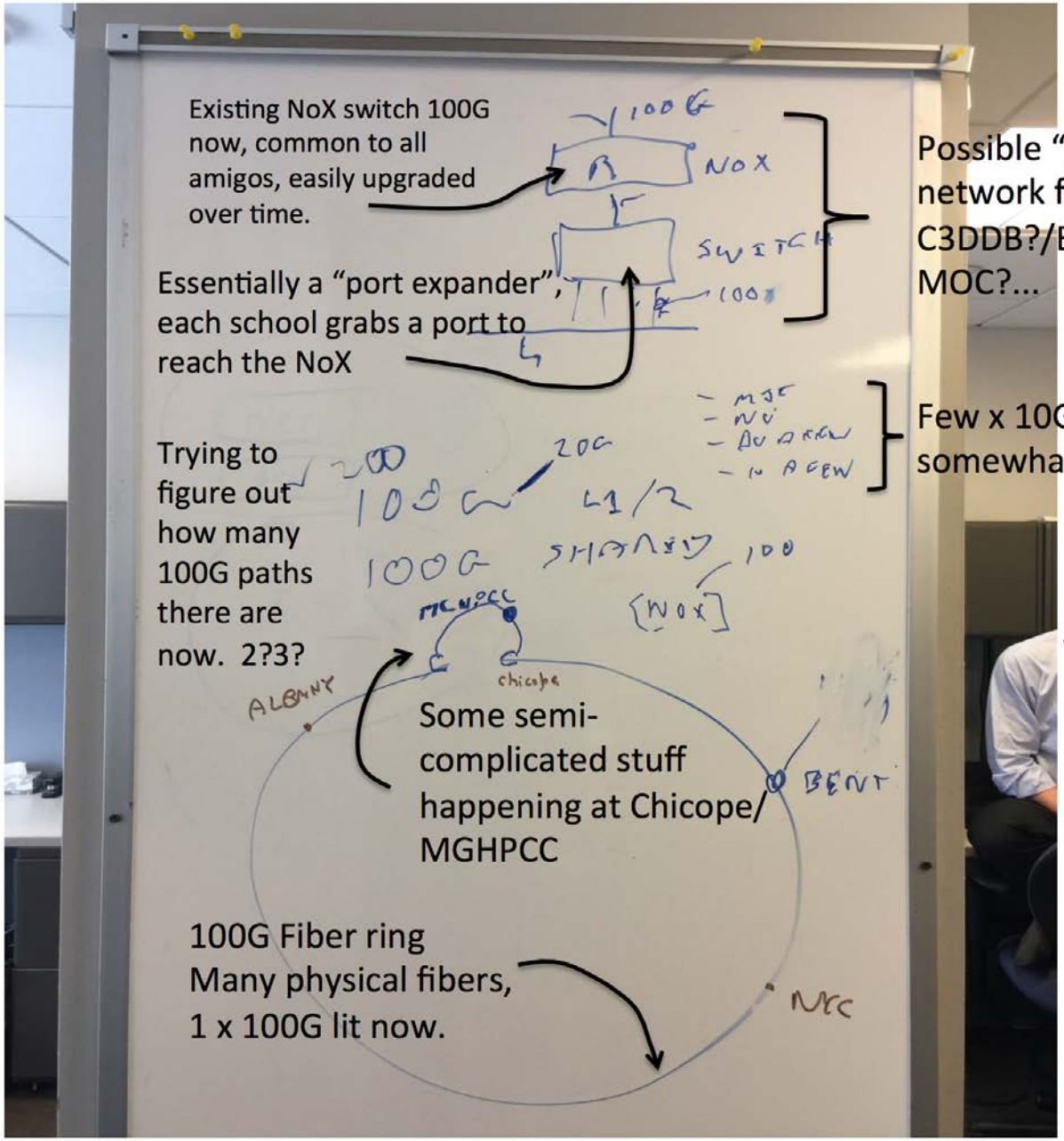
Interfaced with company employees to solve wide-ranging computing and telephony problems and ensured they were seen through to successful resolution. Used scripting skills to automate once timely manual processes.

Organization	Primary Role
Harvard	Build, install and operate CEPH object store hardware, software, and monitoring. DTN cluster configuration and deployment, Globus endpoint management, devops and systems engineering, cluster high availability with required network configuration and security of science DMZ. Assistance to Harvard research groups testing or adopting NESE storage.
Boston University	Planning, network configuration, testing and migration of NET2 storage to NESE. Integration of NESE storage into NET2 operations. File system interfaces to NESE. Assistance for BU research groups testing or adopting NESE storage. Federation with external CEPH clusters.
MIT	iRods overlay to NESE object store. Demonstration and evaluation using 4PiB heterogeneous ocean data. Support of iRods + NESE application to separately funded combined altimetry and ocean color research. Development of general cookbooks illustrating use of iRods and NESE for open data sharing and discovery science activities.
Northeastern	POSIX file and block storage presentations of NESE object store. Evaluation of cost metering and allocation to researchers / projects. Help Northeastern researchers evaluate and use NESE object store.
University of Massachusetts	Policy, Standards, Cybersecurity and Security Operations Center. The resource will join a team that provides 24x7 monitoring, alerting and escalation; ensuring incidents are detected, investigated, communicated, and reported. Assistance to UMass research groups testing or adopting the NESE object store.
MGHPCC	Federated authentication and access control; SDN access, operation of the data center that houses the hardware; planning for long term sustainability; physical security for sensitive data.


HARVARD
BOSTON UNIVERSITY
echo
iRODS
UMASS
POSIX
Authentication / Access layer (MGHPCC)
"S3" APIs
RADOSGW
LIBRADOS
RADOSGW
LIBRADOS
RADOSGW
LIBRADOS
RADOSGW
LIBRADOS
RADOSGW
LIBRADOS
Node A
OSD
OSD
OSD
OSD
Node B
OSD
OSD
OSD
OSD
Node C
OSD
OSD
OSD
OSD
Node D
OSD
OSD
OSD
OSD
MON
ceph
RADOS Cluster



Note – N X 10Gbps links will grow over time toward 100Gbps driven by demand and interface cost improvements
Project funds will ensure that NESE access speeds keep up with demand.
Other MGHPCC activity will drive improvements to existing infrastructure, benefiting NESE as well



Existing NoX switch 100G now, common to all amigos, easily upgraded over time.

Essentially a "port expander", each school grabs a port to reach the NoX

Possible "DMZ", common network for NESE/NET2/ C3DDB?/Engage?/ MOC?...

Few x 10G links, somewhat historical

Trying to figure out how many 100G paths there are now. 2?3?

Some semi-complicated stuff happening at Chicope/ MGHPCC

100G Fiber ring
Many physical fibers,
1 x 100G lit now.

John, taking a break ☺

It’s all about the building blocks!

Funding

- About Funding
- Browse Funding Opportunities A-Z
- Due Dates
- Find Funding
- Merit Review
- Policies and Procedures
- Preparing Proposals
- Recent Opportunities
- Small Business
- Transformative Research

Home › Funding

Email Print Share

Directorate for Computer & Information Science & Engineering

Data Infrastructure Building Blocks (DIBBs)

CONTACTS

Name	Email	Phone	Room
Amy Walton	DIBBsQueries@nsf.gov	(703) 292-8970	
Robert Chadduck	DIBBsQueries@nsf.gov	(703) 292-8970	
Anita Nikolich	DIBBsQueries@nsf.gov	(703) 292-8970	
Peter H. McCartney	DIBBsQueries@nsf.gov	(703) 292-8470	
Sylvia Spengler	DIBBsQueries@nsf.gov	(703) 292-8930	
John C. Cherniavsky	DIBBsQueries@nsf.gov	(703) 292-5136	
Ronald Joslin	DIBBsQueries@nsf.gov	(703) 292-7030	
Joanne D. Culbertson	DIBBsQueries@nsf.gov	(703) 292-4602	
Eva Zanzerkia	DIBBsQueries@nsf.gov	(703) 292-8556	
Lin He	DIBBsQueries@nsf.gov	(703) 292-4956	
Daryl W. Hess	DIBBsQueries@nsf.gov	(703) 292-4942	
Bogdan Mihaila	DIBBsQueries@nsf.gov	(703) 292-8235	
Cheryl L. Eavey	DIBBsQueries@nsf.gov	(703) 292-7269	
Seta Bogosyan	DIBBsQueries@nsf.gov	(703) 292-4766	

PROGRAM GUIDELINES

Solicitation 17-500

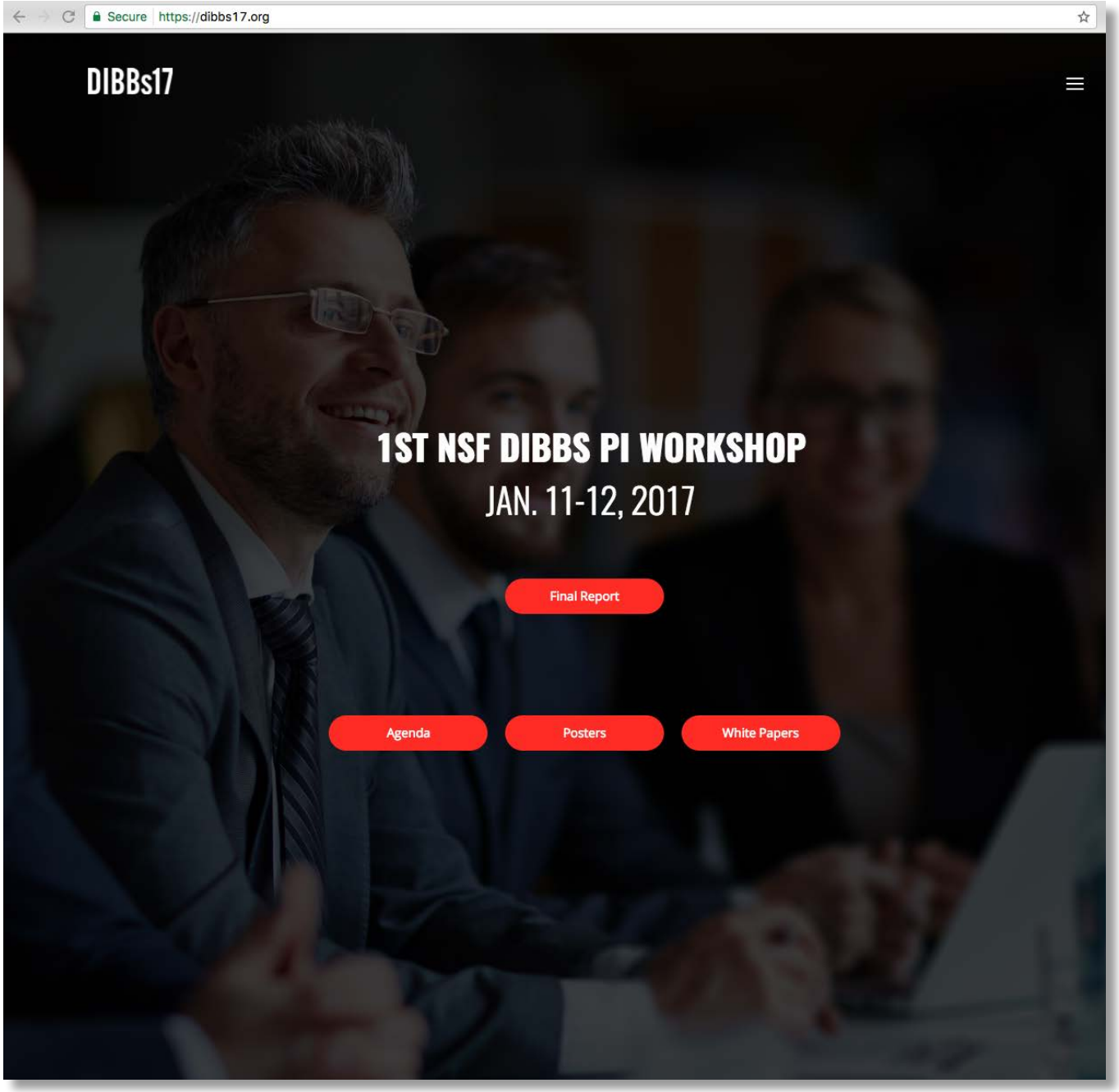
Important Information for Proposers

A revised version of the *NSF Proposal & Award Policies & Procedures Guide* (PAPPG) (NSF 17-1), is effective for proposals submitted, or due, on or after January 30, 2017. Please be advised that, depending on the specified due date, the guidelines contained in NSF 17-1 may apply to proposals submitted in response to this funding opportunity.

SYNOPSIS

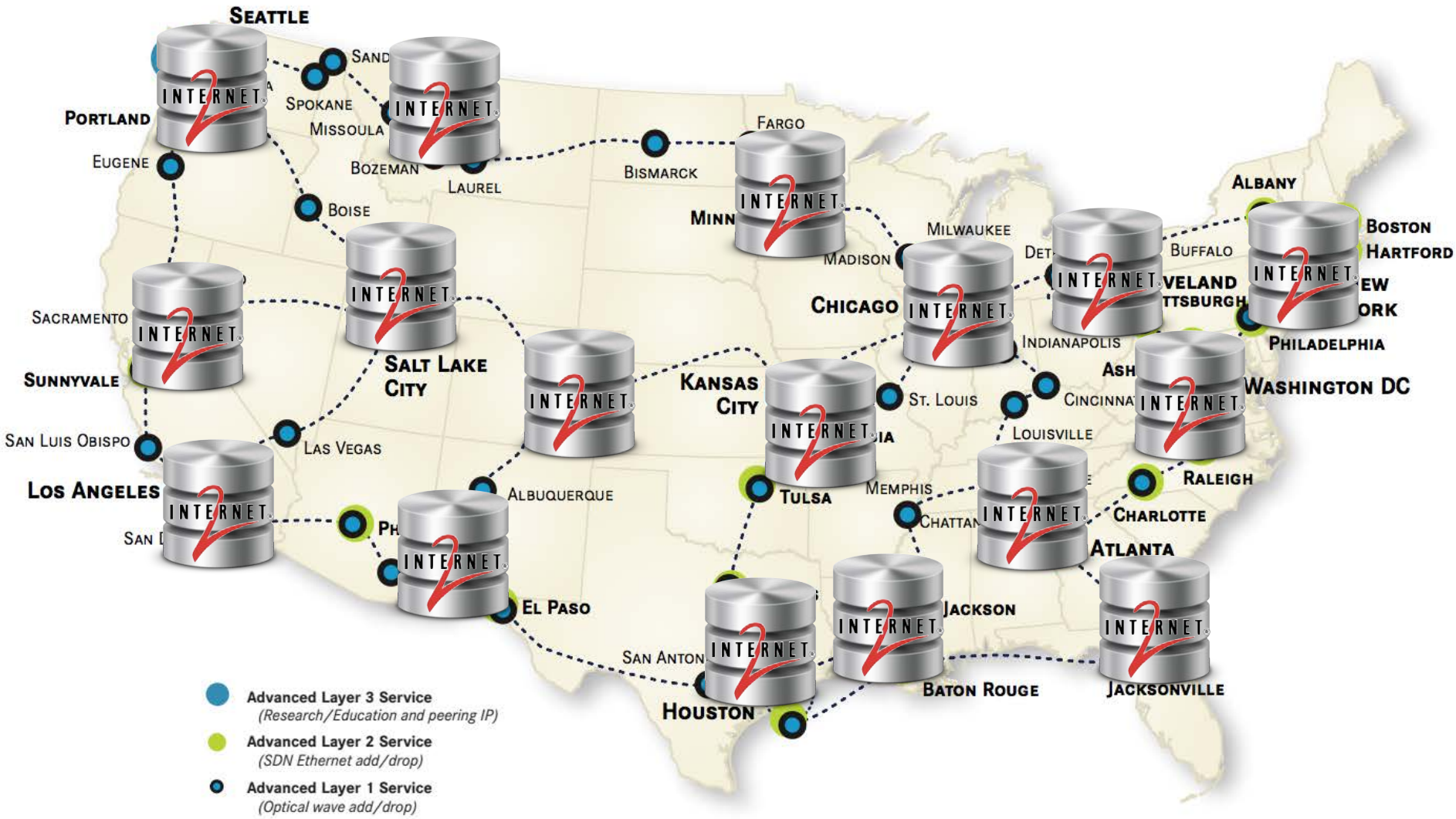
The NSF vision for a Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21) considers an integrated, scalable, and sustainable cyberinfrastructure to be crucial for innovation in science and engineering (see www.nsf.gov/cif21). The Data Infrastructure Building Blocks (DIBBs) program is an integral part of CIF21. The DIBBs program encourages development of robust and shared data-centric cyberinfrastructure capabilities, to accelerate interdisciplinary and collaborative research in areas of inquiry stimulated by data.

DIBBs investments enable new data-focused services, capabilities, and resources to advance scientific discoveries, collaborations, and innovations. The investments are expected to build upon, integrate with, and contribute to existing community cyberinfrastructure, serving as


Secure https://dibbs17.org
DIBBs17
1ST NSF DIBBS PI WORKSHOP
JAN. 11-12, 2017
Final Report
Agenda
Posters
White Papers

Internet2 DataPOP Infrastructure Topology

October 2014



web: projectnese.org

twitter: @projectnese