AAMC 2013 Information Technology in Academic Medicine Conference
Vancouver CA  June 5-7, 2013

Michael Sullivan, M.D.
Associate Director, Health Sciences, Internet2

Internet2 Support for Biomedical Research
Overview

Internet2 Research Support
- Community and Network
- Data-intensive Science
- International Collaboration
- Innovation Platform

Big Data Challenges
- Transport
- Security
- Storage and Compute
Internet2 Community

220 Universities
60 Corporations
70 Government agencies
38 Regional and state networks
65 International R&E networks
Advanced 100G Production and Research Network

Internet2 Planned 100 Gigabit Infrastructure Topology (DRAFT)
Draft – Last updated 04/10/11

Internet2 Network by the numbers
10 Juniper T1600 routers
7 Juniper MX960 nodes for TR-CPS
49 custom colocation facilities
250+ amplification racks
15,500 miles of newly acquired dark fiber
2,400 miles of partnered capacity with NTNC
8.8 Tbps of optical capacity
100 Gbps of IP capacity
300+ Ciena ActiveFlex 6500 network elements
Data Tsunami

Physics
Large Hadron Collider

Life Sciences
Magnetic Resonance Imager (MRI)

Image by: CERN
Visualizing Big Data

Physics
LHC – Lead Ion Collision

Life Sciences
MRI – Monkey Brain

Source: CERN (ALICE detector)

Source: Van Wedeen, M.D., Martinos Center and Dept. of Radiology, Massachusetts General Hospital and Harvard University Medical School
Sequencing: Smaller, Faster, Cheaper

Illumina HiSeq 2500/1500

Handheld USB Sequencer

Source: http://www.illumina.com/systems/hiseq_systems/hiseq_2500_1500.ilmn

Image: Oxford Nanopore Technologies
Democratization of Sequencing

2,386 Genome Sequencers Worldwide – 30 May 2013

Source: Map of High-throughput Sequencers
North American Genome Sequencers

998 Sequencers in NA – 30 May 2013

Source: Map of High-throughput Sequencers
Sequencing in Vancouver

13 Sequencers at the Genome Science Center

Source: Map of High-throughput Sequencers
US-based International Exchange Points

US-based Exchange Points
- StarLight, Chicago IL
- MAN LAN, New York NY
- NGIX-East, College Park MD
- AtlanticWave (distributed)
- AMPATH, Miami FL
- PacificWave-S, Los Angeles CA
- PacificWave-N, Seattle WA
Synchronized Genomic Repositories: NCBI, EBI, DDBJ
US – China 10 Gbps Link

Dr. Lin Fang
Fed Ex: 2 days
Internet + FTP: 26 hours
China-US 10G Link: 30 seconds

Dr. Dawei Lin

Sample.fa (24GB)
Innovation Platform

100 GigE Layer 2 Connection

Science DMZ

- SDN Control Server
- Performance Node
- High-Performance Layer 2/3 Switch/Router
- Switches, data stores for data-intensive science
- Traditional Campus Border Router
- Traditional L3 Campus Border Security
- Campus Enterprise Network

For more information, see fasterdata.es.net

Software Defined Networking

- R&E IP
- TR-CPS
- Your Research
- GENI Experiments
- IP Network Layer 3
- Static Layer 2
- Dynamic Layer 2
- GENI
- Traditional Services
- Innovation Services
- Software Defined Networking Substrate
- Optical System
- Dark Fiber

Internet2 innovation backbone delivered as 100G L1

Traditional regional and commodity providers

Traditional L3 Campus Border Security

For more information, see fasterdata.es.net

17 – 6/7/13, © 2012 Internet2
Innovation Platform Pilot Sites

Innovation Campus Pilot Sites
100GE - Science DMZ - Software-Defined Networking
Meeting the Big Data Challenges

**Transport**
- Science DMZ
- PerfSONAR Toolkit
- MaDDash Testing Mesh
- File Transfer Tools

**Security**
- Science DMZ Hardening
- Federated IdM: InCommon and NSTIC

**Storage and Compute**
- Storage and Compute
Challenge #1: Transport

http://fasterdata.es.net/science-dmz/science-dmz-security/
Performance Monitoring

Sample Results: Finding/Fixing soft failures

- Rebooted router with full route table
- Gradual failure of optical line card
MaDDash XSEDE Testing Mesh

<table>
<thead>
<tr>
<th>OWAMP</th>
<th>BWCTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ps.lu.xsede.org</td>
<td>ps.lu.xsede.org</td>
</tr>
<tr>
<td>ps.ncar.xsede.org</td>
<td>ps.ncar.xsede.org</td>
</tr>
<tr>
<td>ps.ncsa.xsede.org</td>
<td>ps.ncsa.xsede.org</td>
</tr>
<tr>
<td>ps.nics.xsede.org</td>
<td>ps.nics.xsede.org</td>
</tr>
<tr>
<td>ps.psc.xsede.org</td>
<td>ps.psc.xsede.org</td>
</tr>
<tr>
<td>ps.purdue.xsede.org</td>
<td>ps.purdue.xsede.org</td>
</tr>
<tr>
<td>ps.sdsc.xsede.org</td>
<td>ps.sdsc.xsede.org</td>
</tr>
<tr>
<td>ps.tacc.xsede.org</td>
<td>ps.tacc.xsede.org</td>
</tr>
</tbody>
</table>

- **OWAMP**
  - Loss is 0
  - Loss is greater than 0
  - Unable to retrieve data
  - Check has not yet run

- **BWCTL**
  - Throughput >= 1000Mbps
  - Throughput >= 100Mbps
  - Throughput < 100Mbps
  - Unable to retrieve data
  - Check has not yet run
File Transfer Tools

**Unix LAN Tools**
- **scp, sftp, rsync** – poor choices for WAN (RTT > 25ms)
- **scp** with HPN patch – better but still has limitations

**TCP – based Open Source**
- **Globus Online** – [http://www.globusonline.org](http://www.globusonline.org)
  - Uses GridFTP with TCP optimizations
  - Friendly GUI, Fire and Forget, Galaxy integration

**UDP – based Commercial**
- **Aspera**: [http://www.asperasoft.com/](http://www.asperasoft.com/)
- **Annai Systems**: [http://www.annaisystems.com](http://www.annaisystems.com)
Tool Speeds

Berkeley, CA ↔ Argonne, IL  RTT=53

Data Transfer Speeds
10Gbps Network

- scp: 0.14
- scp-HPN patched: 1.2
- ftp: 1.4
- GridFTP, 4 streams: 5.4
- GridFTP, 8 streams: 6.6

Gigabits per second
Challenge #2: Security

Hardening the Science DMZ

- ESnet Big Data design pattern
- Internet2 Innovation Platform
- NSF CC-NIE grants
- University of Florida
  - HIPAA alignment
  - Efficient encryption
  - Comprehensive logging
  - Robust authentication

Source: www.securearc.com
Federated Identity Management

Number of Participants

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>10</td>
</tr>
<tr>
<td>2005</td>
<td>10</td>
</tr>
<tr>
<td>2006</td>
<td>20</td>
</tr>
<tr>
<td>2007</td>
<td>30</td>
</tr>
<tr>
<td>2008</td>
<td>40</td>
</tr>
<tr>
<td>2009</td>
<td>60</td>
</tr>
<tr>
<td>2010</td>
<td>100</td>
</tr>
<tr>
<td>2011</td>
<td>150</td>
</tr>
<tr>
<td>2012</td>
<td>350</td>
</tr>
</tbody>
</table>

© 2012 Internet2
NSTIC – National Strategy for Trusted Identities in Cyberspace

- White House initiative administered by NIST
- Goal is to create an “Identity Ecosystem”
- IDEGS – Identity Ecosystem Steering Group
- Five awards for pilots spanning multiple sectors:
  - Resilient Network Systems, AMA, Aetna, ACC, NeHC, ...
  - Criterion Systems, ID/DataWeb, AOL, Experian, Ping Identity, ...
  - Daon, Inc., AARP, PayPal, Purdue, ...
  - American Assoc. of Motor Vehicle Admins, Microsoft, AT&A, etc...
  - Internet2, Carnegie Mellon, Brown, MIT, U. of Texas, U. of Utah...
Challenge #3: Storage and Compute

- **Cloud Computing – many initiatives**
  - Private: NCI bake-off to create Cancer Knowledge Clouds
  - Public/Private: AWS EC2 instances — [100G] — NCBI repository
  - Open Cloud: BioNimbus Protected Data Cloud
  - Proprietary: BGI EasyGenomics Cloud

- **National Cyberinfrastructure**
  - XSEDE
  - Internet2
  - NCGAS
NCI: Cancer Knowledge Cloud - RFI

Summary of Community Input

https://wiki.nci.nih.gov/display/NCIPinput/Summary+of+Input+Request%3A+Computational+Needs+to+Support+Large-Scale+Genomics+Investigations
NCBI: Four Different Approaches

- Reduced Data Size
- Incrementally Transfer Large Files
- High Speed Network Connections
- Cloud Access and Support

Source: Don Preuss, *NCBI Experiences and Big Data Strategy*, presented at 2013 Internet2 Annual Meeting, Arlington, VA
BioNimbus: An Open Cloud with Protected Data

BIONIMBUS PROTECTED DATA CLOUD

Secure cloud services for the scientific community

What is the Bionimbus PDC?

The Bionimbus Protected Data Cloud (PDC) is a collaboration between the Open Science Data Cloud (OSDC) and the IGSB (IGSB), the Center for Research Informatics (CRI), the Institute for Translational Medicine (ITM), and the University of Chicago Comprehensive Cancer Center (UCCCC). The PDC allows users authorized by NIH to compute over human genomic data from dbGaP in a secure compliant fashion. Currently, selected datasets from the The Cancer Genome Atlas (TCGA) are available in the PDC.

How can I get involved?

- Apply for a Bionimbus PDC account and use the Bionimbus PDC to manage, analyze and share your data.
- Partner with us and add your own racks to the Bionimbus PDC (we will manage them for you).
- Help us develop the open source Bionimbus PDC software stack.

You can contact us at info@opencloudconsortium.org.

How do I get started?

First, apply for an account. Once your account is approved, you can login to the console and get started. Support questions can be directed to support@opencloudconsortium.org.

Apply for the PDC Now
Login to the PDC Console

bionimbus.opensciencedatacloud.org
EasyGenomics: BGI’s Cloud Solution

Source: Xu Xing, Managing Big Data: The Genome Center Perspective, presented at Bio-IT World Conference & Expo ‘13, Boston, MA

https://www.easygenomics.com
National Cyberinfrastructure

- **XSEDE**
  - NSF-funded
  - Supercomputers
  - HPC resources
- **Internet2**
  - 220 universities
  - XSEDEnet
- **NCGAS**
  - Indiana University
  - TACC
  - SDSC
  - PSC

Source: https://www.xsede.org/networking
**NCGAS Virtual Instrument**

Networking Issues for Life Sciences Research

Focused Technical Workshop on July 17-18, 2013
Lawrence Berkeley National Laboratory
Berkeley, California

• Building on the success of Joint Techs, meeting will bring together technical experts in a smaller setting with domain scientists.
• Workshop will include a slate of invited speakers and panels.
• Format to encourage lively, interactive discussions with the goal of developing a set of tangible next steps for supporting this data-intensive science community
• Four sub-topic areas: Network Architectures, Workflow Engines, Public and Private Cloud Architectures, and Data Movement Tools
• See: http://events.internet2.edu/2013/ftw-life-sciences/index.cfm
Resources

• The Fourth Paradigm – Data-Intensive Scientific Discovery

• Internet2 Network and Innovation Platform
  – http://www.internet2.edu/network/

• Science DMZ
  – http://fasterdata.es.net/science-dmz/

• perfSONAR
  – http://www.perfsonar.net/

Contact

• Internet2 Research Support Center
  – rs@internet2.edu

• Internet2 Life Sciences – Michael Sullivan, MD, Associate Director
  – msullivan@internet2.edu
Thank You

INTERNET2 SUPPORT FOR BIOMEDICAL RESEARCH

AAMC 2013 Information Technology in Academic Medicine Conference
Vancouver CA  June 5-7, 2013

Michael Sullivan, M.D.
Associate Director, Health Sciences, Internet2