Trans-Atlantic 100 Gbps Service Experiments and Results

Joe Mambretti, Director, (j-mambretti@northwestern.edu)
International Center for Advanced Internet Research (www.icair.org)
Northwestern University
Director, Metropolitan Research and Education Network (www.mren.org)
Co-Director, StarLight, PI-iGENI, PI-OMNINet (www.startap.net/starlight)
Co-PI Chameleon (www.chameleoncloud.org)

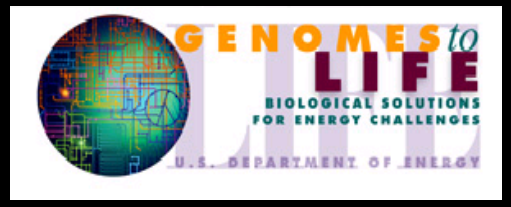
Global LambdaGrid Workshop
Queenstown, New Zealand
September 30 - October 1, 2014

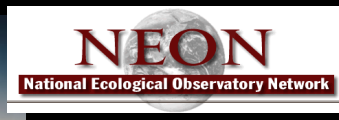
iCAIR

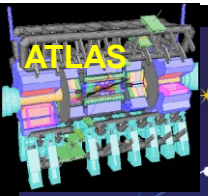
STARLIGHT℠

Context Issues

- Network Services Are Oriented to Supporting Aggregations of 10s of Millions of Individual Streams
- Currently, Network Services, Architecture and Supporting Technologies Are Not Designed and Implemented To Support High Capacity Individual Streams, i.e., Large Scale, High Performance, Wide Area Data Transport at 100 Gbps
- A⇔N*TB/PB of Data ⇔BA
- Such a Service Is Required (It Is Actually Essential) , e.g., To Support Data Intensive Science But Also Other Application Areas
- Such a Transport Service Does Not Exist Except On A Few Specialized Networks (e.g., ESnet) At a Few Specialized Sites
- Two Macro Network NB: Capacity (e.g., 100 Gbps) Does Not Equate With Capability
- Science Themes: Capacity and Programmability



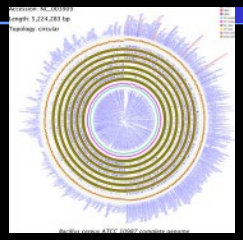




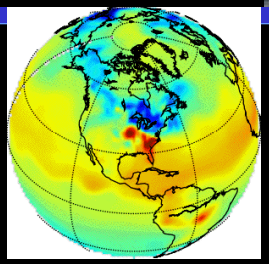
ALMA: Atacama Large Millimeter Array

ANDRILL: Antarctic Geological Drilling
www.andrill.org

BIRN: Biomedical Informatics Research Network
www.nbirn.net



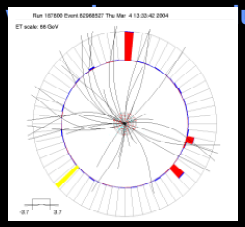
CAMERA metagenomics
camera.calit2.net



Carbon Tracker
www.esrl.noaa.gov/gmd/ccgg/carbontracker

CineGrid
www.cinegrid.org

LHCONE
www.lhcone.net



DØ (DZero)
www-d0.fnal.gov

GEON: Geosciences Network
www.geongrid.org

GLEON: Global Lake Ecological Observatory Network

OOI-CI
ci.oceanobservatories.org

ISS: International Space Station
www.nasa.gov/station

Comprehensive Large-Array Stewardship System
www.class.noaa.gov

LIGO
www.ligo.org

WLCG
lcg.web.cern.ch/LCG/public/

Pacific Rim Applications and Grid Middleware Assembly
www.pragma-grid.net

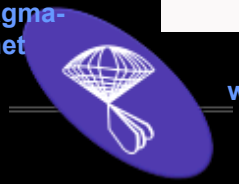
TeraGrid
www.teragrid.org

IVOA: International Virtual Observatory
www.ivoa.net

OSG
www.opensciencegrid.org

Globus Alliance
www.globus.org

SKA
www.skatelescope.org



Sloan Digital Sky Survey
www.sdss.org

XSEDE
www.xsede.org

Compilation By Maxine Brown

STARLIGHT SM

Transporting Data at 100 G (Alex Szalay's Perspective)

- Without 100 Gbps Networks, It Is Challenging To Move Petabytes: Transferring 1PB At an *Effective* 5 Gbps Would Take 18.5 Days
- Such Efficiency Today Is Still Difficult To Achieve
- Moving Data At an Effective 100 Gbps Reduces the Time To One Day
- This Is a Make-Or-Break Difference for PB Science
- Effective Transport of Petabyte-Scale Data Sets Enables Many Research Communities To Solve Cutting Edge Problems, From the Traditional HPC and CFD To Investigations of Internet Connectivity

Issues for Science Data WAN/LAN Transport at 100 G

- Terascale and Petascale Data, Storage Systems and File Systems
- Architecture and Technology For 100 G Edge Nodes (Related to Workflow, Edge Processes, Transport Protocols, Etc.)
- Architecture and Technology for Interfaces To LAN and WAN 100 Gbps Transport
- National and Regional Transport Capacity at 100 G
- Architecture, *Services*, and Technology for 100 Gbps Transport – Ultra High Capacity Single Streams
- Currently, Without Services, The Substitute is Network and Edge Node Science, Prototype Technology, and a Few "Magic Formulas."

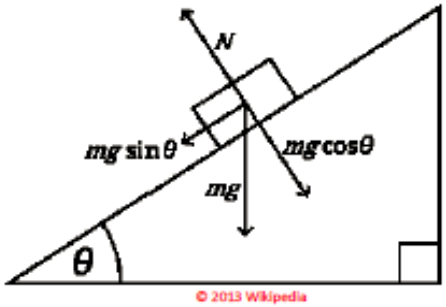
Considerations For Large Scale 100 Gbps Edge Nodes

- High Quality Edge Hosts, Especially Backplanes, Memory, Cores, and I/O Channels
- High Performance Transport Protocols
- High Performance File Systems
- Highly Tuned OSs (Best = A Custom Tuned High Perf. Linux)
- Tuned BIOS
- Tuned I/O Scheduler
- Virtual Memory Subsystem
- SSD Tuning
- High Performance, Tuned RAID Controller
- Highly Tuned High Quality NICs
- Configuration of Processing Cores
- Interrupt Binding
- MTU Configurations (e.g., Minimally Jumbo Frames)
- Large Widow Size (e.g., Set SO_SNDBUF/SO_RCVBUF to ~ 4MB)
- Packet Pacing
- And Much More!

The Science DMZ* in 1 Slide

Consists of **three key components**, all required:

- "Friction free" network path
  - Highly capable network devices (wire-speed, deep queues)
  - Virtual circuit connectivity option
  - Security policy and enforcement specific to science workflows
  - Located at or near site perimeter if possible
- Dedicated, high-performance Data Transfer Nodes (DTNs)
  - Hardware, operating system, libraries all optimized for transfer
  - Includes optimized data transfer tools such as Globus Online and GridFTP
- Performance measurement/test node
  - perfSONAR
- Engagement with end users



globus

perfSONAR

Details at http://fasterdata.es.net/science-dmz/

* ***Science DMZ*** is a trademark of The Energy Sciences Network (ESnet)

ESnet

STARLIGHT SM

High Speed WAN Data Transfers

Goals

- Move bulk data traffic from USLHCNet servers in CERN to storage server in Caltech (Pasadena)
- A Proof of Concept, Designing 100G Data Cache Server in front of large Tier1 and Tier2 centers
- Achieve maximum Network to Disk ratio
- Reduce CPU utilization and application wise network to disk Latency
- An easy setup for others to use as a design guide

CalTech⇔CERN Experiments

Azher Mughal
July 17, 2014

STARLIGHT SM

National Science Foundation Initiatives

- Campus Cyberinfrastructure - Network Infrastructure and Engineering Program (CC-NIE)
- Invests In Improvements and Re-Engineering At Campus Level to leverage Dynamic Network Services To Support Range of Scientific Data Transfers and Movement.
- Program Also Supports Network Integration Activities Tied To Achieving Higher Levels of Performance, Reliability and Predictability for Science Applications and Distributed Research Projects.
- Campus Cyberinfrastructure - Infrastructure, Innovation and Engineering Program (CC*IIE)
- Invests In Improvements and Re-engineering at Campus Level To Support Range of Data Transfers Supporting Computational Science and Computer networks and Systems Research.

Specialized Edge Node Devices

- Customized 40 Gbps – 100 Gbps Edge Node Devices
- PetaTranS Node (iCAIR)
- NASA Goddard HECN Node
- FIONA (UCSD)
- CalTech/CERN
- And Others…

TransLight/StarLight
Petascale Science Prototype Services Facility

- Project Is Funded By the National Science Foundation Funded International Research Network Connections Program (IRNC)
- Goal: Prototyping Trans-Atlantic 100 Gbps Architectures, Services, and Emerging Technologies Among Institutions Connected to NetherLight, StarLight, and Other Participating GOLEs in North America and Europe
- The TransLight/StarLight Consortium Was Awarded a National Science Foundation (NSF) Grant To Establish An Initiative To Design and Implement Network Architectures, Services, Technologies, and Core Capabilities In Support of Big Data Science Over 100 Gbps Trans-Atlantic Paths, Enabling Large-Scale Global Scientific Research Experiments, Instrumentation Access, Collaborations, Data Sharing and High-Resolution Visualization.

StarLight 100 Gbps/Tbps Initiatives

- StarLight Has Established Multiple Initiatives That Are Directed At Creating High Capacity Network Transport Services, Architecture, Technology, and Networks Based on 100 Gbps and Higher Service, Including Investigations of the Potentials of Individual 100 Gbps and Tbps Streams
- Foundation Research Is Based On Earlier Experience Including With Dynamic Lightpath Technologies
- 100 Gbps Capabilities Must Include More Than Capacity (e.g., Programmability, Dynamic Control Over Channel Segments, Customization)

TransLight/StarLight
Petascale Science Prototype Services Facility

- **This Project Is Designing, Implementing, and Experimenting With Prototype Services and Capabilities That Have the Potential to Optimize Advanced Networks for Production Science Research, Particularly for Large-Scale Data Transport, Including Persistent, Ultra-Large-Capacity, Real-Time, Long-Duration Streams. These Experiments Are Being Conducted With Multiple National and International Partners.**
- ***Four Major Themes of This Initiative Are To Provide: (a) Large-Scale Network Capacity, Including Support For Extremely High-Volume Individual Data Streams, (b) Network Services and Resource Programmability For This Capacity, (c) Edge Access To These Capabilities, and (d) Exchange Facilities That Can Support These Services and Capabilities.***

StarLight – "By Researchers For Researchers"

StarLight Is An Experimental And Production Optical Infrastructure Based **Proving Ground for Advanced Network Services** Optimized for High-Performance Data Intensive Applications

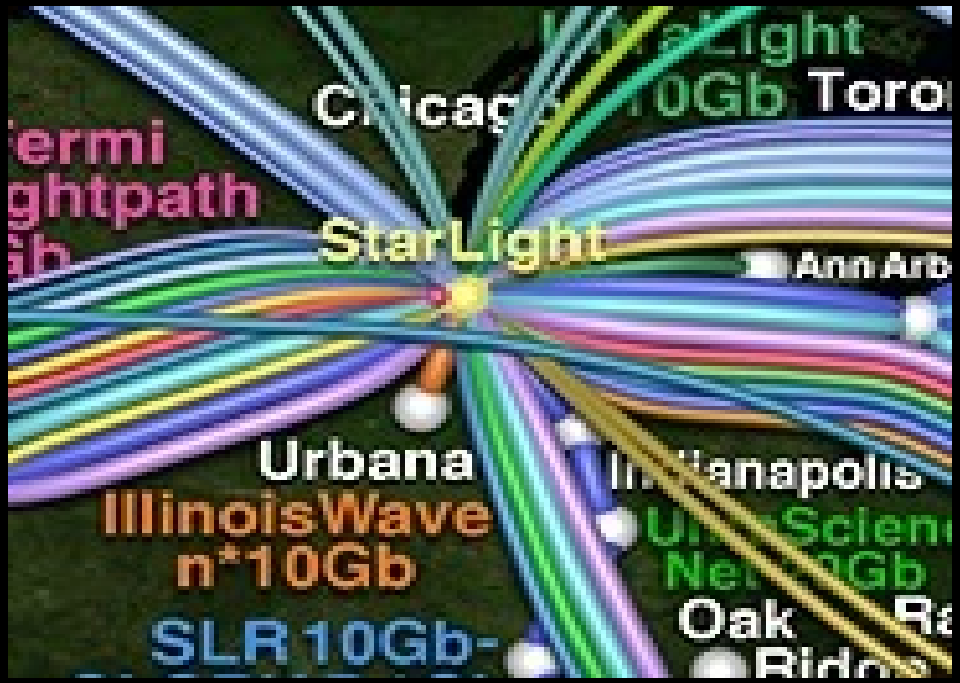
View from StarLight

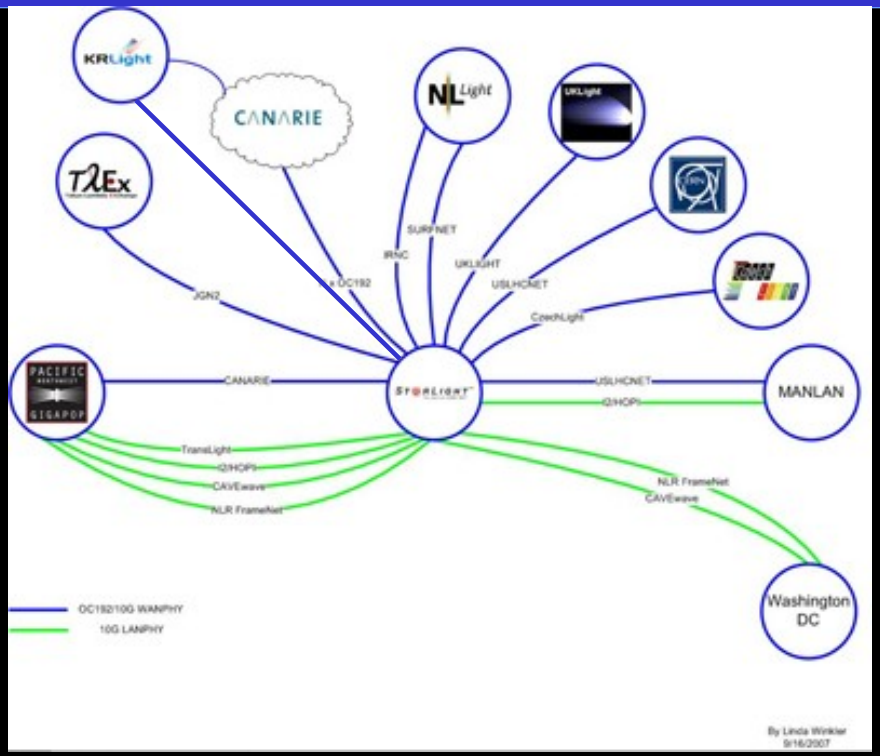
Abbott Hall, Northwestern University's Chicago Campus

STARLIGHT℠

Relatively Current StarLight Infrastructure

Ciena OME 8700
Calient PXC (L1)
Juniper MX 960 (L2/L3)
Many Lambdas & Collaborators
Many 100 G Paths





http://wiki.glif.is/index.php/StarLight

Measurement Servers:
bwctl, owamp, ndt/npad, perfSONAR

Selected PSPSF Applications at 100 G

- SDN, Science DMZ
- WAN/LAN High-Performance Large capacity Flows and File Transfers
- Computational Genomics and Analytics
- Open Science Data Cloud (HPC Clouds)
- DataScope (Computational Astrophysics, e.g., Using SDSS Data)
- LHC Research
- 8K Digital Media Streaming
- SAGE Streaming (Scientific Visualization)
- Network Science Experimental Research (Including GENI/iGENI and International Network Research Testbeds)
- Petascale Computational Science (e.g., Blue Waters, Titan)
- 8k Digital Media/HPDMnet
- CineGrid
- HPC Storage Streaming

StarLight 100 Gbps/Tbps Initiatives

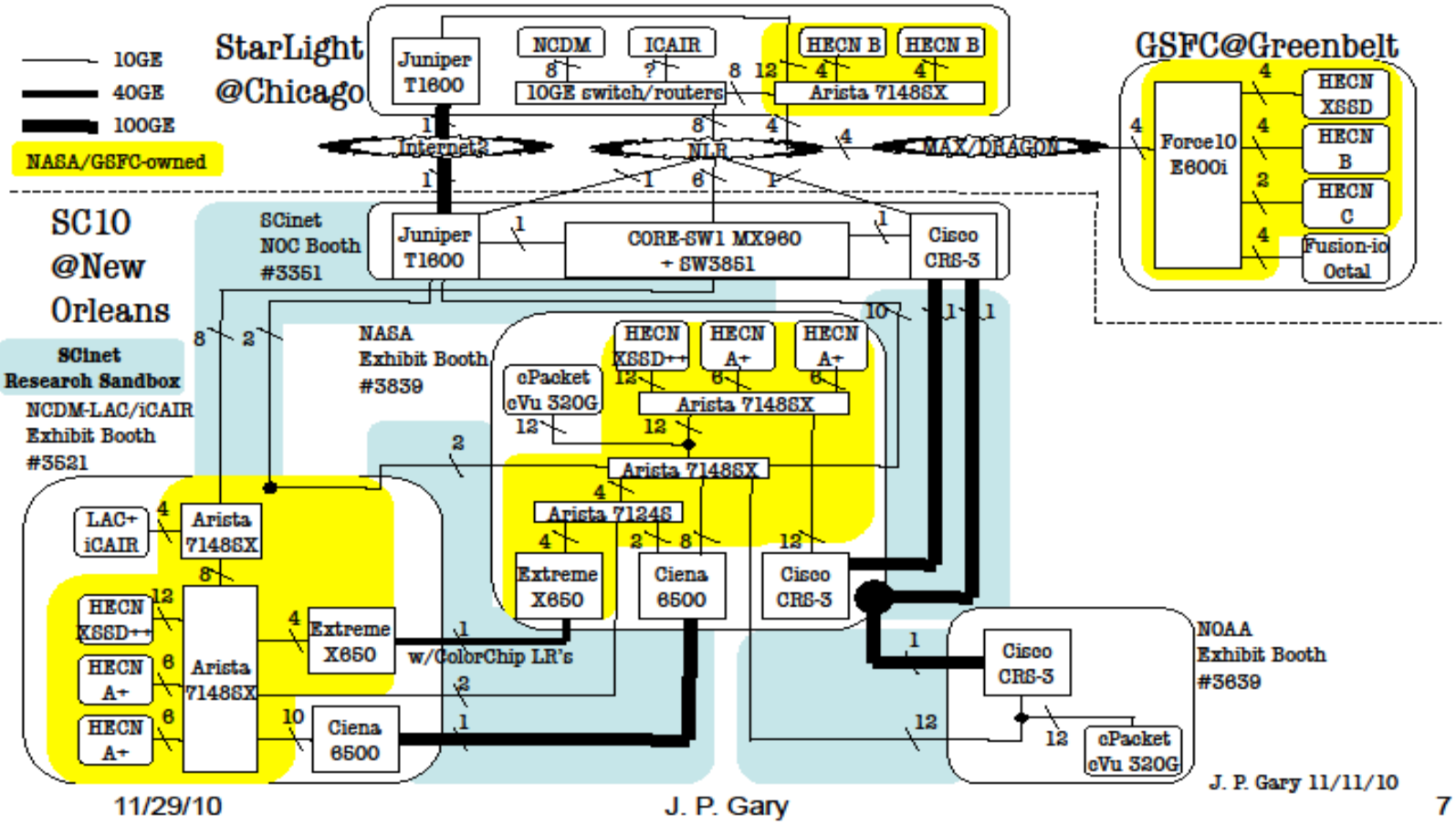
- StarLight Has Established Several Initiatives That Are Directed At Creating Networking Services, Architecture, Technology, and Networks Based on 100 Gbps and Higher Service, Including Tbps
- Foundation Research Is Based On Earlier Experience Including With Dynamic Lightpath Technologies
- 100 Gbps = More Than Capacity (e.g., Dynamic Control Over Channel Segments, Customization)

StarWave: A Multi-100 Gbps Exchange Facility

- StarWave, An Advanced Multi-100 Gbps Exchange Facility and Services Implemented Within the StarLight International/National Communications Exchange Facility
- StarWave Was Implemented In 2011 To Provide Services To Support Large Scale Data Intensive Science Research Initiatives
- StarWave Was Developed With Support from an NSF Advanced Research Infrastructure Award
- StarWave Supported Multiple SC13 Demonstrations
- 10 Separate Sets of 100 Gbps Demonstrations Are Planned for SC14 in New Orleans in November 2014

Using 100G Network Technology in Support of Petascale Science

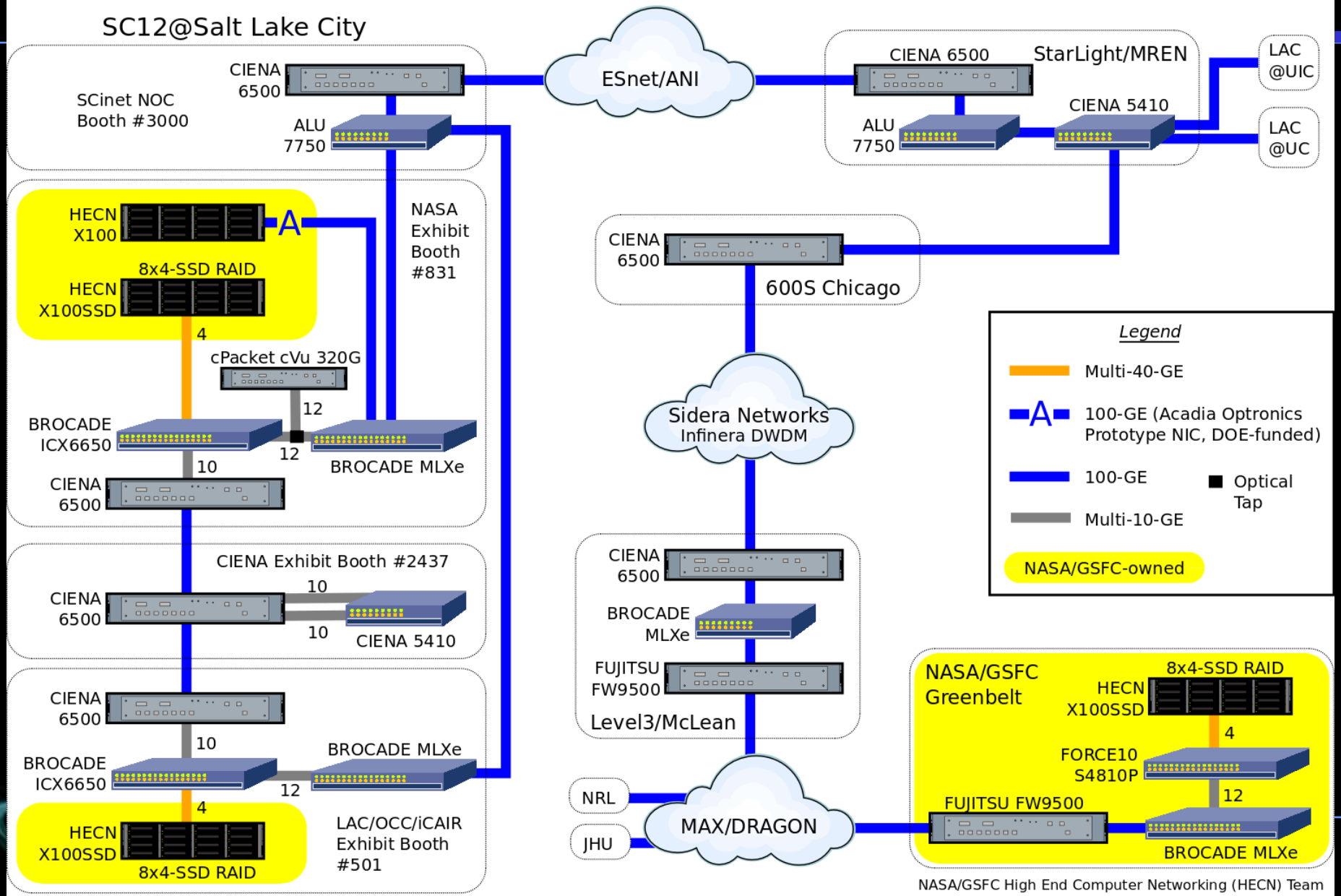
A Collaborative Initiative Among NASA, NLR, NOAA, Northwestern/iCAIR, SCinet & UIC/LAC
Also Using Internet2's Multi-Vendor 100GigE Infrastructure Between StarLight and SC10



STARLIGHT SM

Evaluations/Demonstrations of 100 Gbps Disk-to-Disk File Transfer Performance using OpenFlow Across LANs & WANs

An SC12 Collaborative Initiative Among NASA and Several Partners



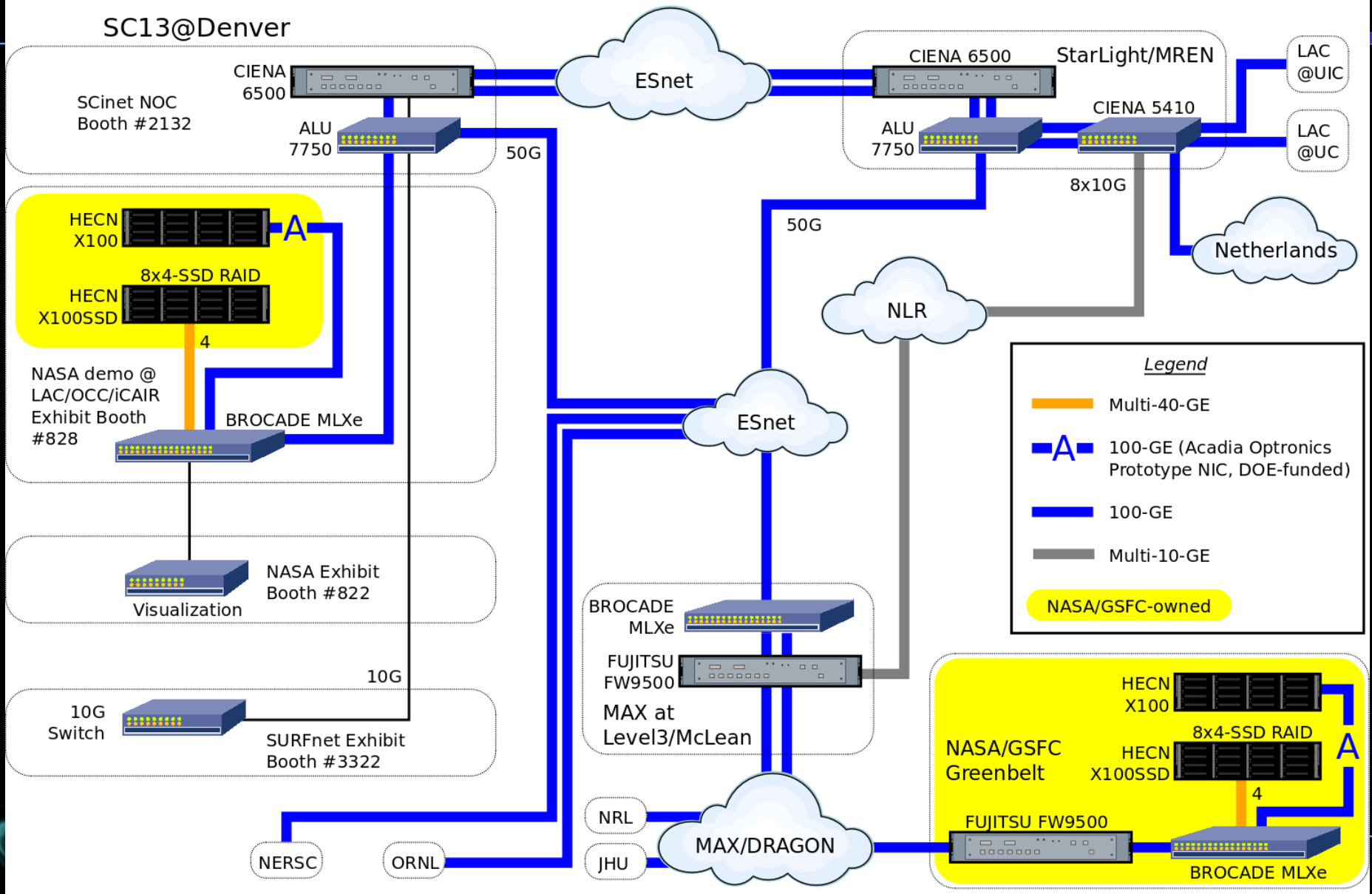
NASA/GSFC High End Computer Networking (HECN) Team
Diagram by Bill Fink / Paul Lang - 9/27/2012

SC13 100 Gbps NRE Demonstrations

- 100 Gbps Networks for Next Generation Petascale Science Research and Discovery
- Data-Scope at 100 Gbps Across National Data-Intensive Computational Science Testbeds
- Next Generation Genome Sequencing Data at 100 Gbps
- The Global Environment for Network Innovations (GENI) at 100 Gbps
- HPC Open Science Data Cloud (OSDC) for Data Intensive Research at 100 Gbps
- Using Remote I/O for Large-Scale Long Distance Data Storage and Processing
- ExoGENI @ 40G: Dynamic Monitoring and Adaptation of Data Driven Scientific Workflows Using Federated Cloud Infrastructure

Evaluations/Demonstrations of 100 Gbps Disk-to-Disk WAN File Transfer Performance

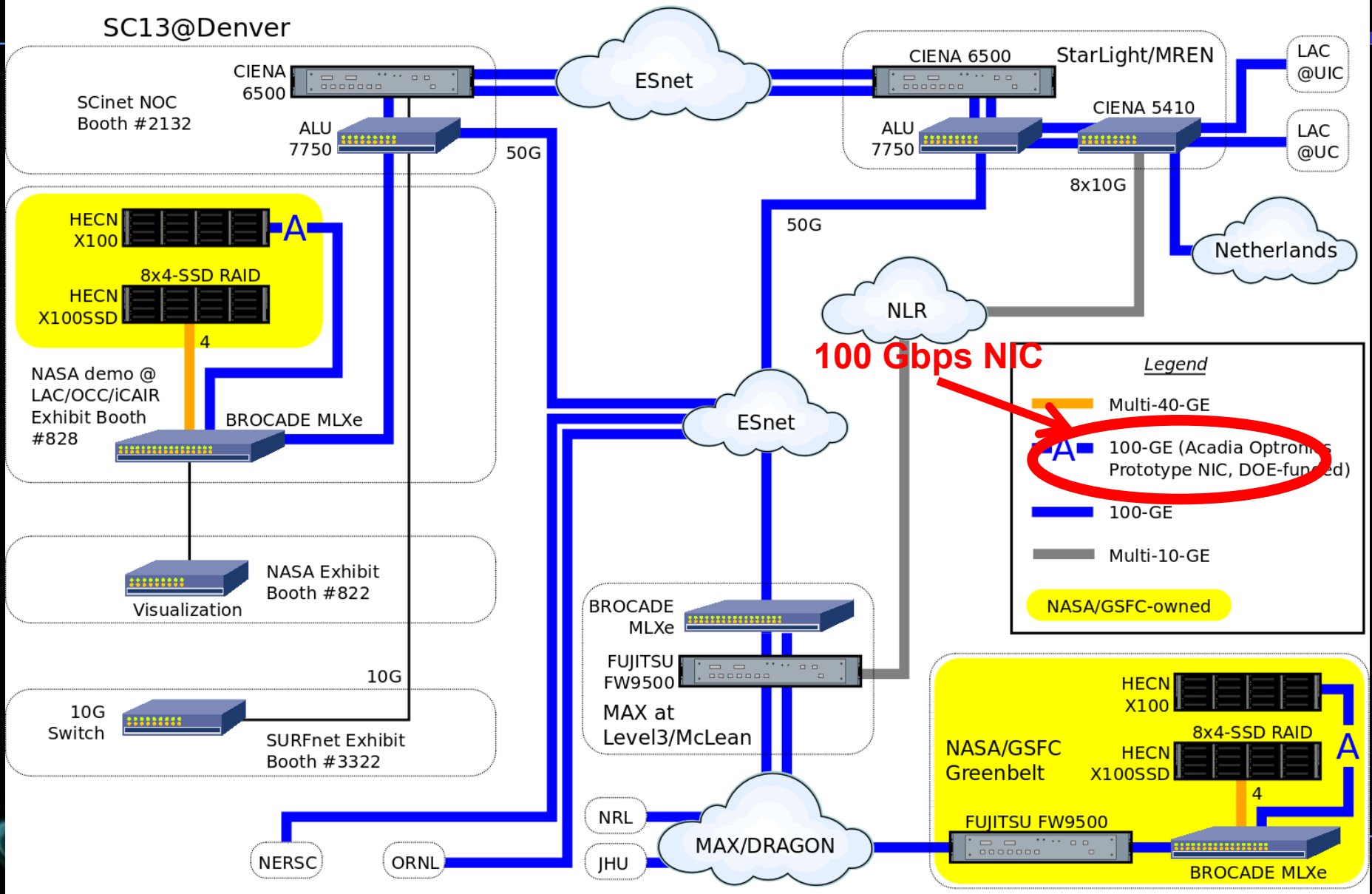
An SC13 Collaborative Initiative Among NASA and Several Partners



NASA/GSFC High End Computer Networking (HECN) Team
Diagram by Bill Fink / Paul Lang - 11/6/2013

Evaluations/Demonstrations of 100 Gbps Disk-to-Disk WAN File Transfer Performance

An SC13 Collaborative Initiative Among NASA and Several Partners

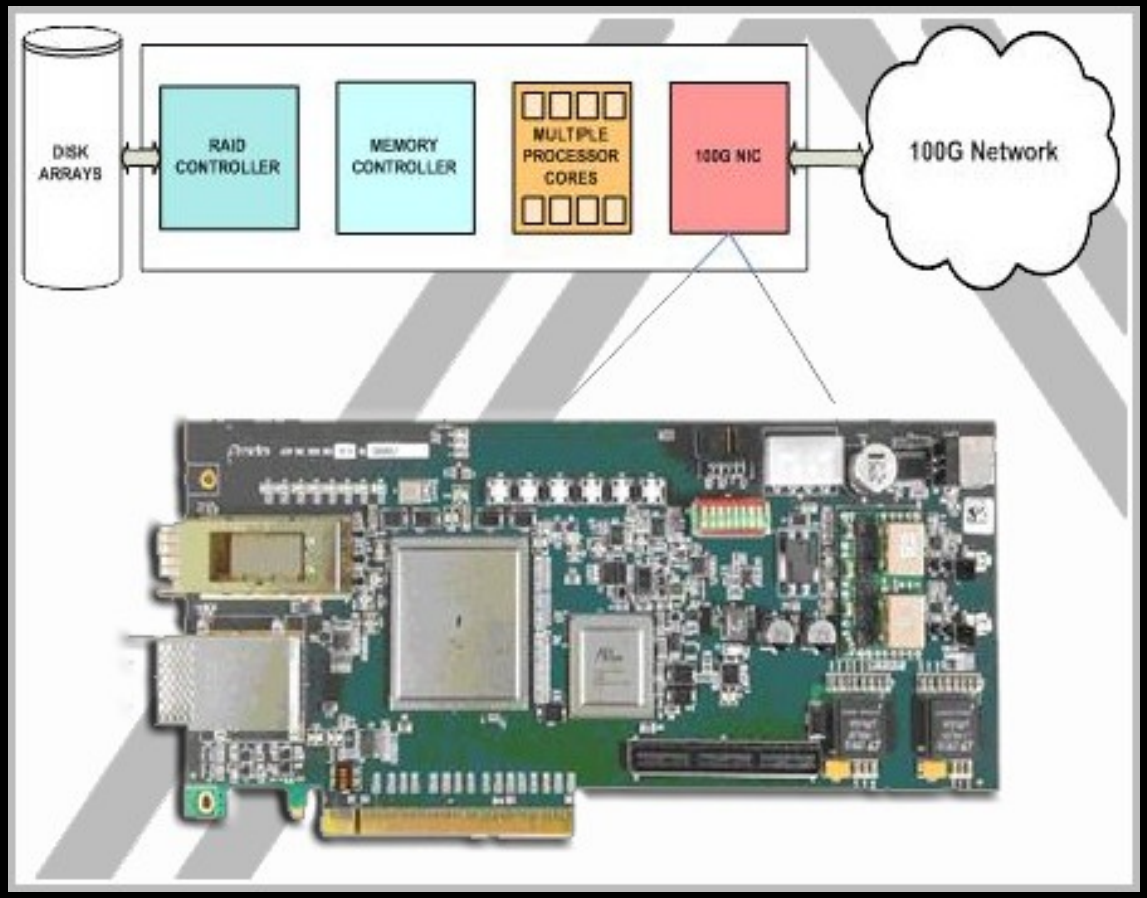


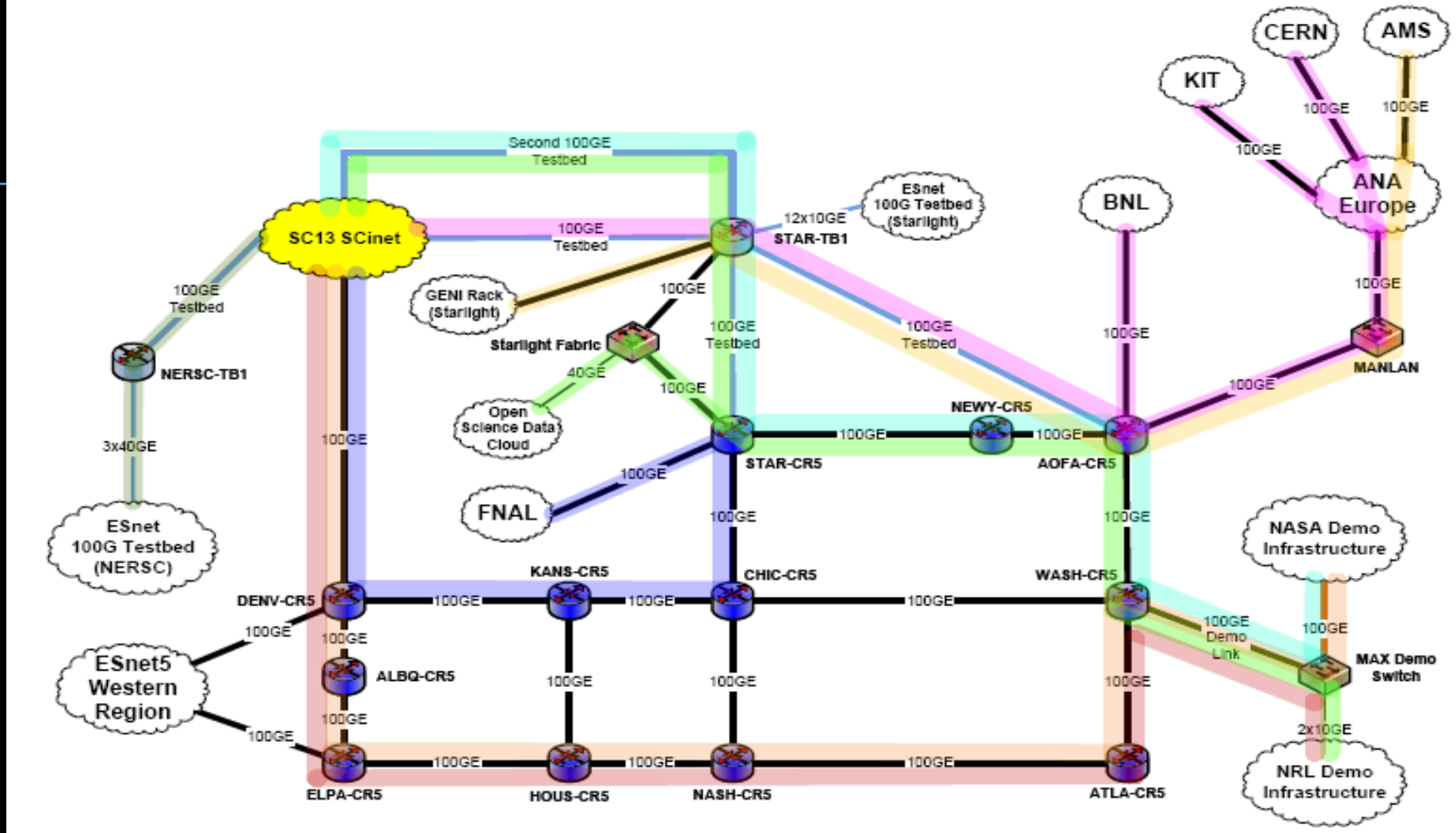
NASA/GSFC High End Computer Networking (HECN) Team
Diagram by Bill Fink / Paul Lang - 11/6/2013

Several SC13 Petascale Science Demonstrations Incorporated 100 Gbps NIC

Acadia Optronics

Development Funded By Department of Energy's Office of Science





NRL demo Northern path (20G)	VLANs 1900, 1901
NRL demo Southern path (20G)	VLANs 1902, 1903
NASA demo – production path (50G)	VLAN 1801
NASA demo – testbed path (50G)	VLAN 1800
OpenFlow/SDN demo – ANA path (100G)	VLANs 1921-1929
Caltech demo – ANA path (100G)	VLANs 2602, 2603, 2606, 2607
Caltech demo – FNAL path (60G)	VLANs 2600, 2601
Caltech demo – NERSC TB path (100G)	VLANs 2604, 2605

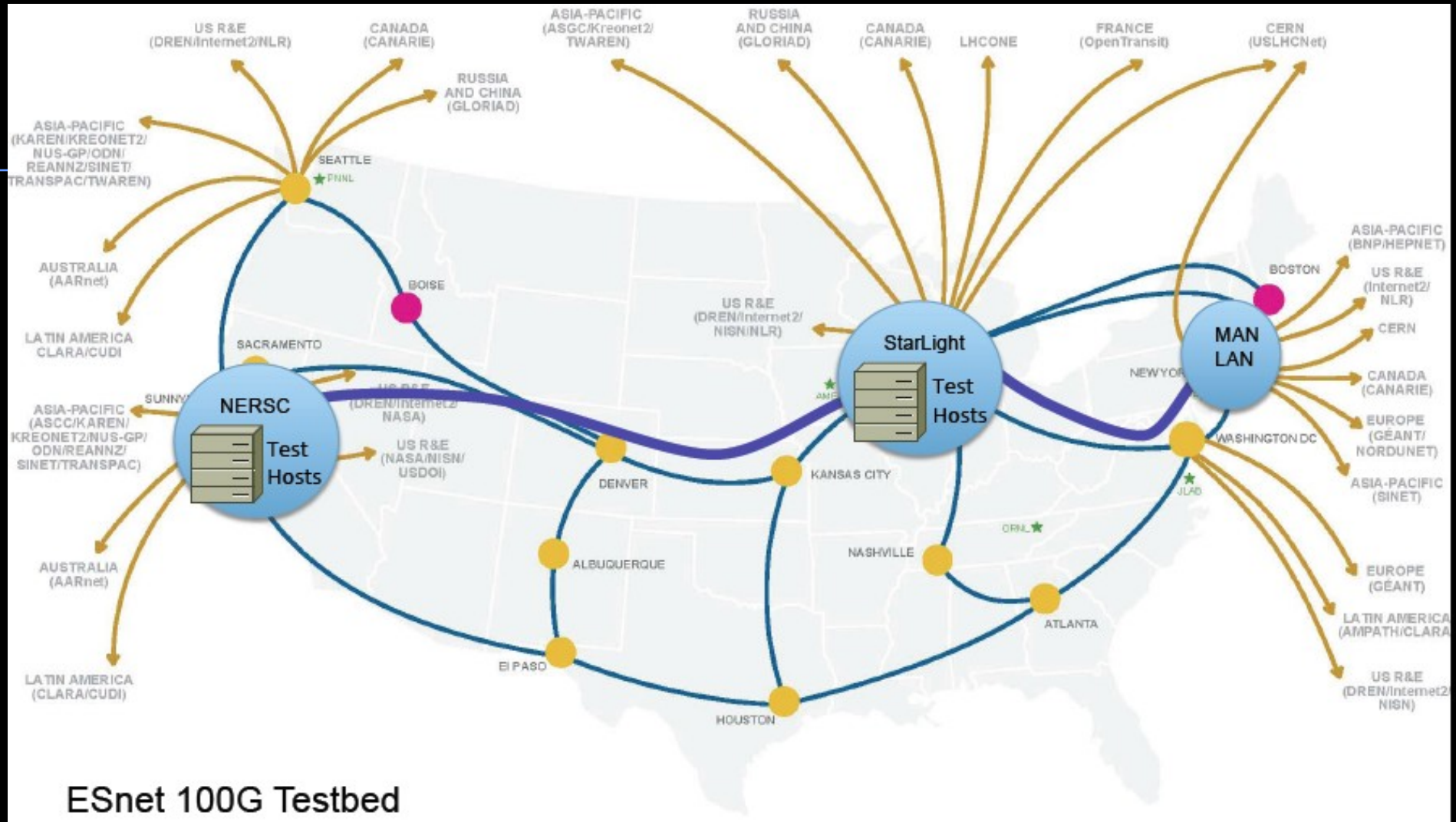
ESnet
Energy Sciences Network

SC13 demos – ESnet5 map

Eli Dart, ESnet 11/14/2013

FILENAME SC13-DEMOS-V24.VSD

STARLIGHT℠



ESnet 100G Testbed

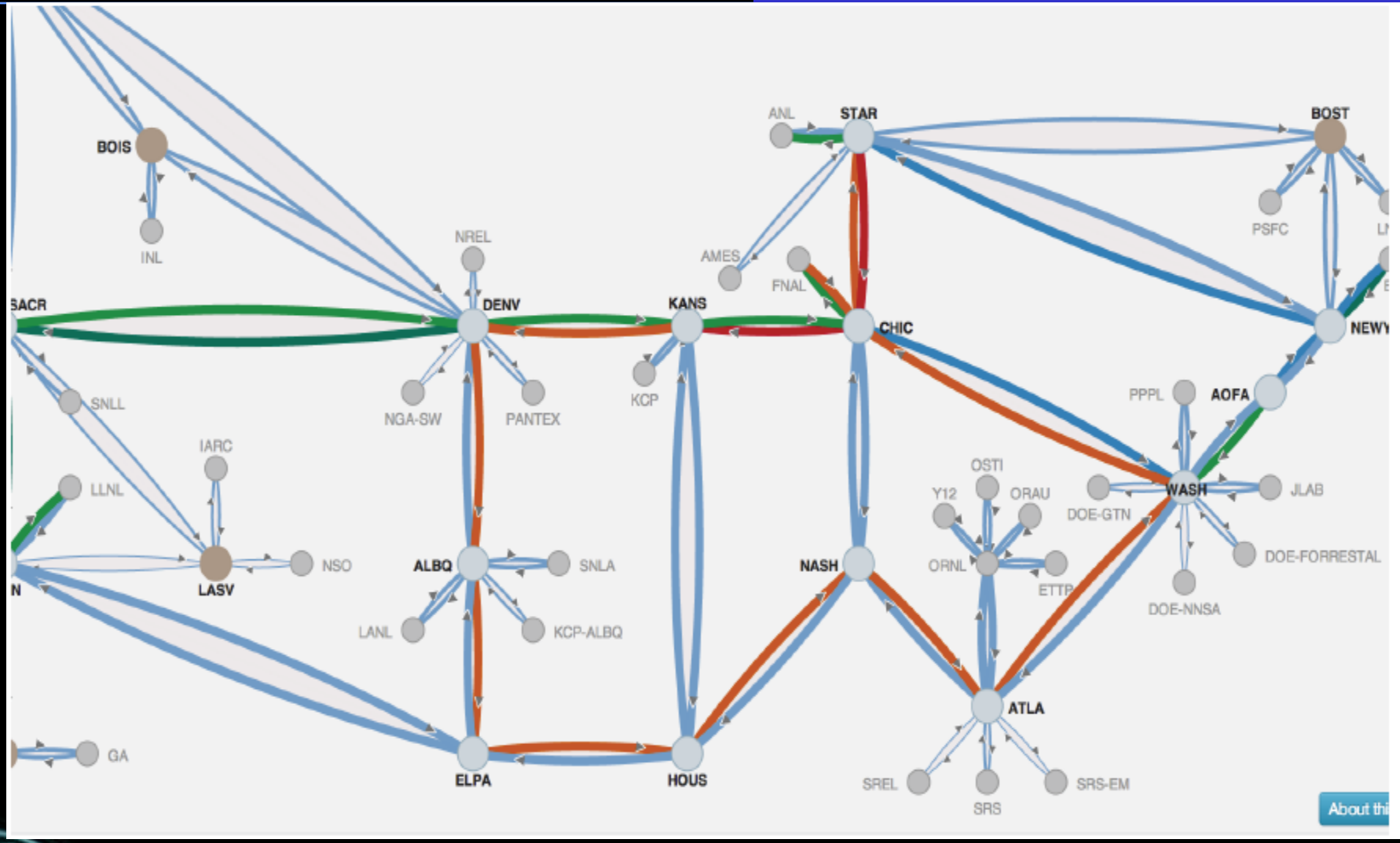
ESnet Energy Sciences Network

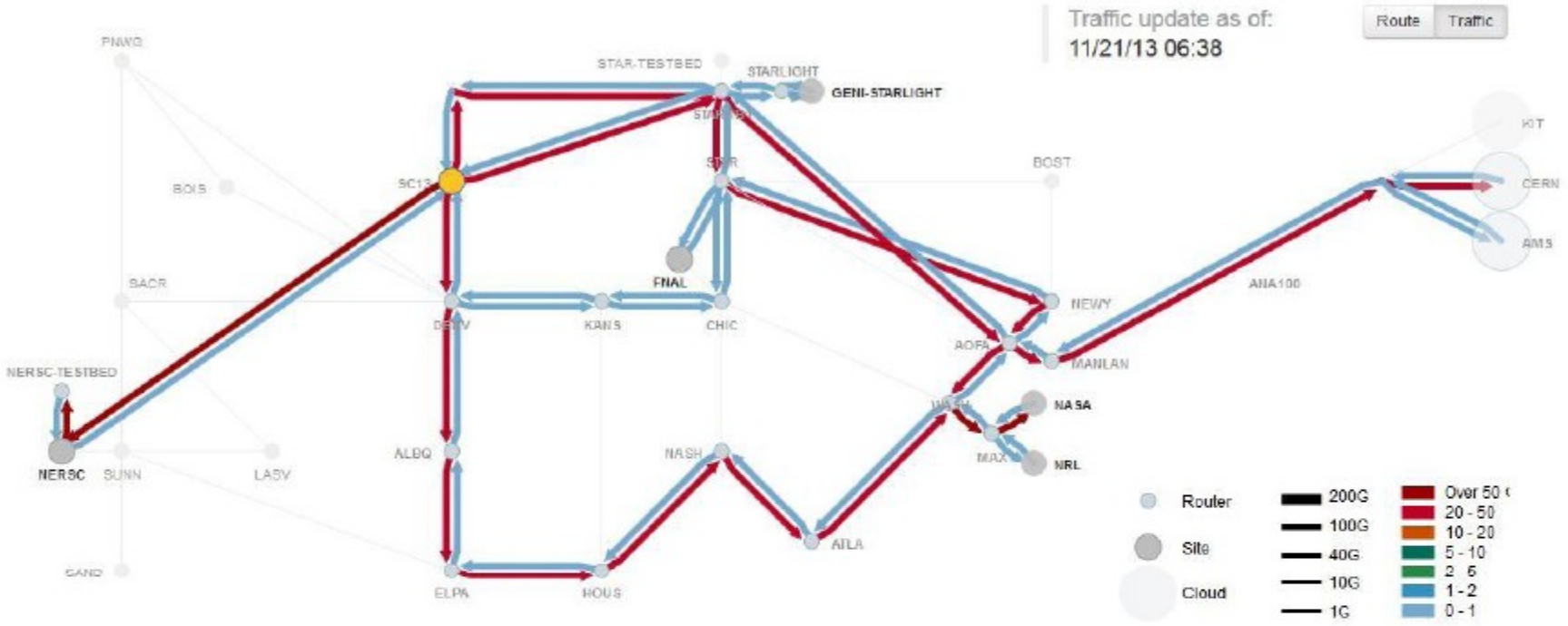
- 100G IP Hubs
- 4x10G IP Hub
- Major R&E and International peering connections

★ Office of Science National Labs

Ames	Ames Laboratory (Ames, IA)
ANL	Argonne National Laboratory (Argonne, IL)
BNL	Brookhaven National Laboratory (Upton, NY)
FNAL	Fermi National Accelerator Laboratory (Batavia, IL)
JLAB	Thomas Jefferson National Accelerator Facility (Newport News, VA)
LBNL	Lawrence Berkeley National Laboratory (Berkeley, CA)
ORNL	Oak Ridge National Laboratory (Oak Ridge, TN)
PNNL	Pacific Northwest National Laboratory (Richland, WA)
PPPL	Princeton Plasma Physics Laboratory (Princeton, NJ)
SLAC	Stanford Linear Accelerator Center (Menlo Park, CA)

Loop Test From/To NASA GSFC On Esnet: Enabled Testing Over Thousands of Miles

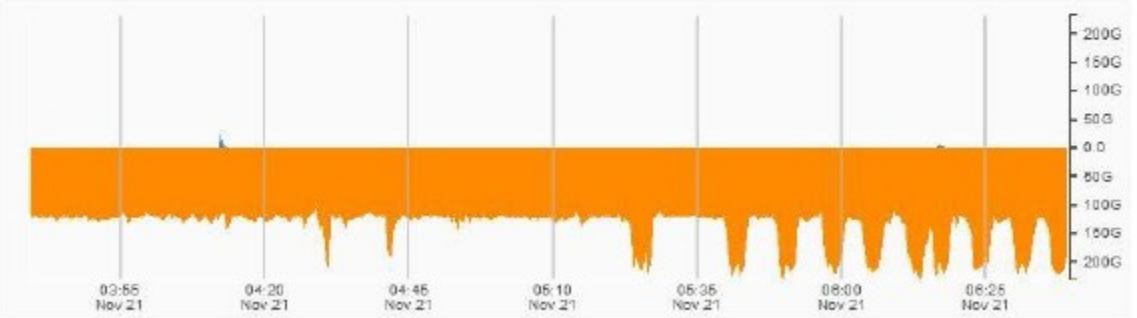




ESnet at SC13

This section of the portal visualizes how ESnet is supporting SC13. ESnet is providing four 100 Gbps circuits to the showfloor, one to the production backbone and three more which connect to the ESnet testbed.

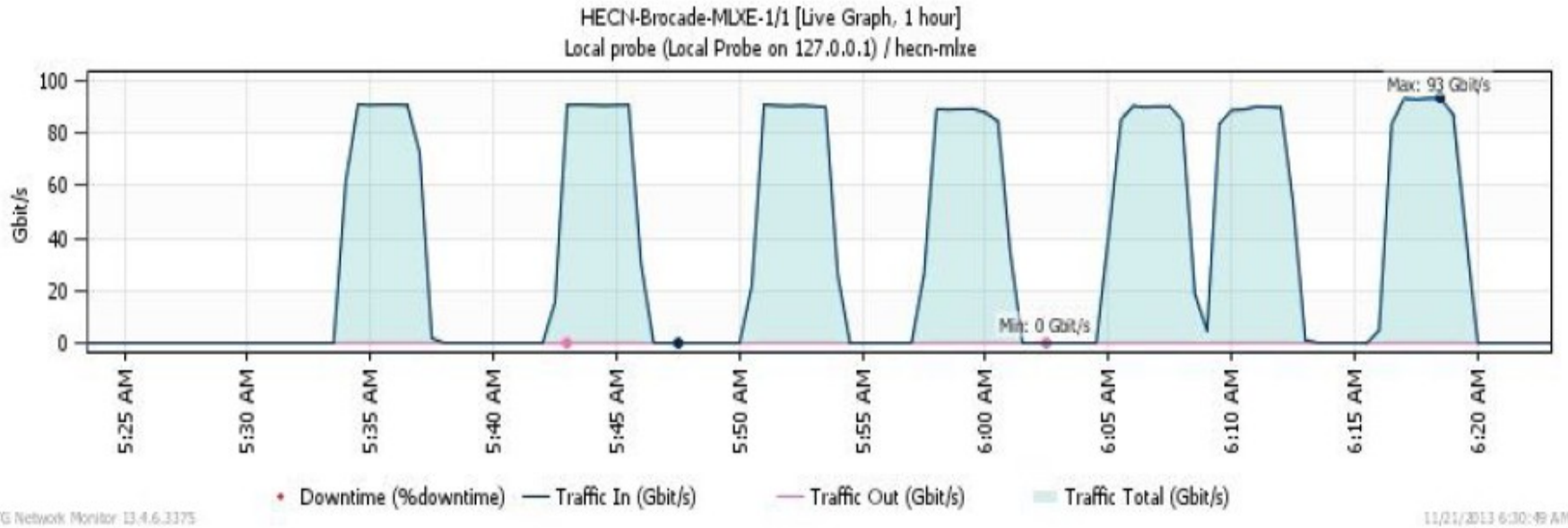
Traffic

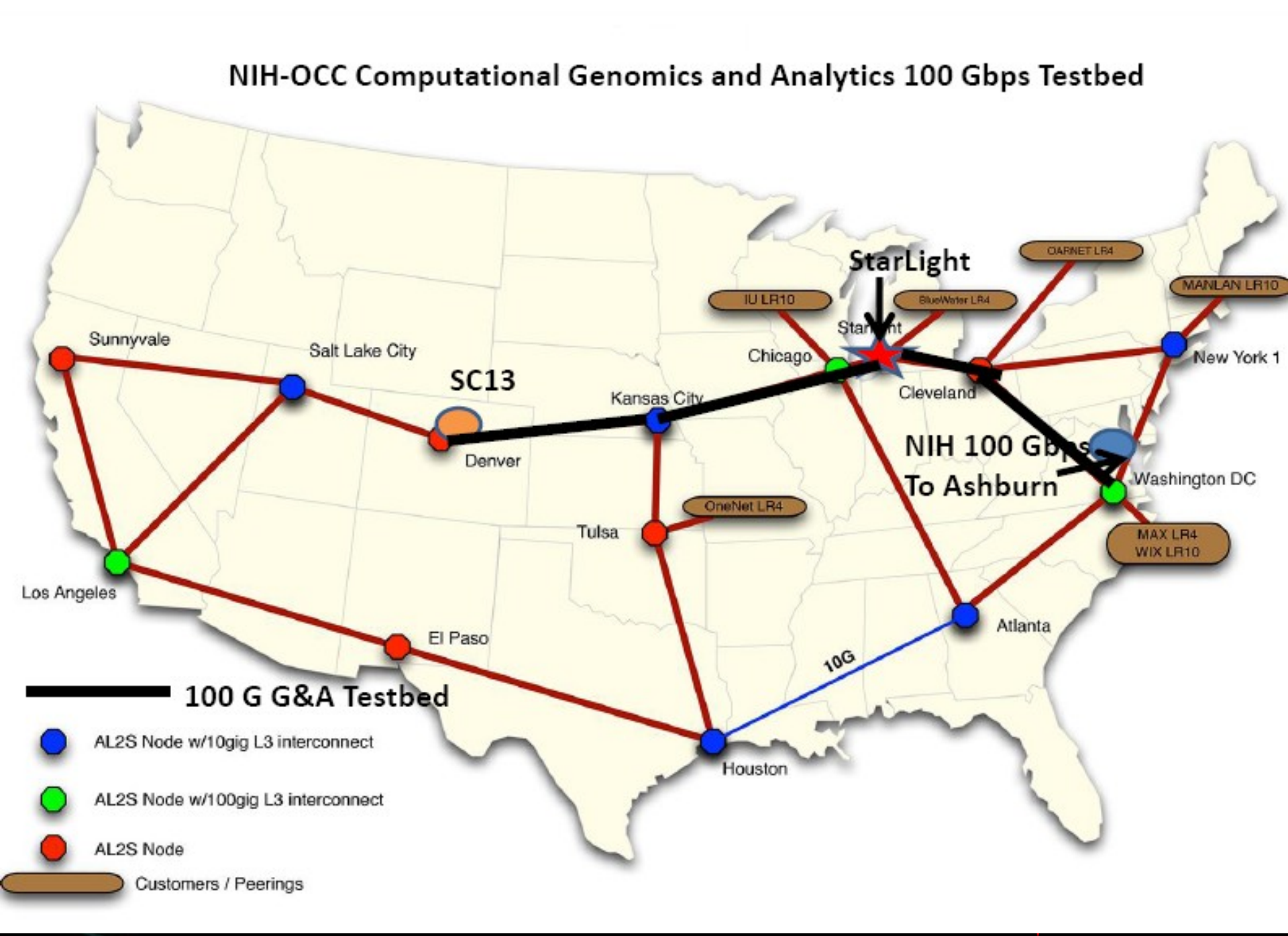


High End Computer Networking (HECN) – GSFC/NASA

Supercomputing 2013

Disk to Disk Data Rates (93Gbps)

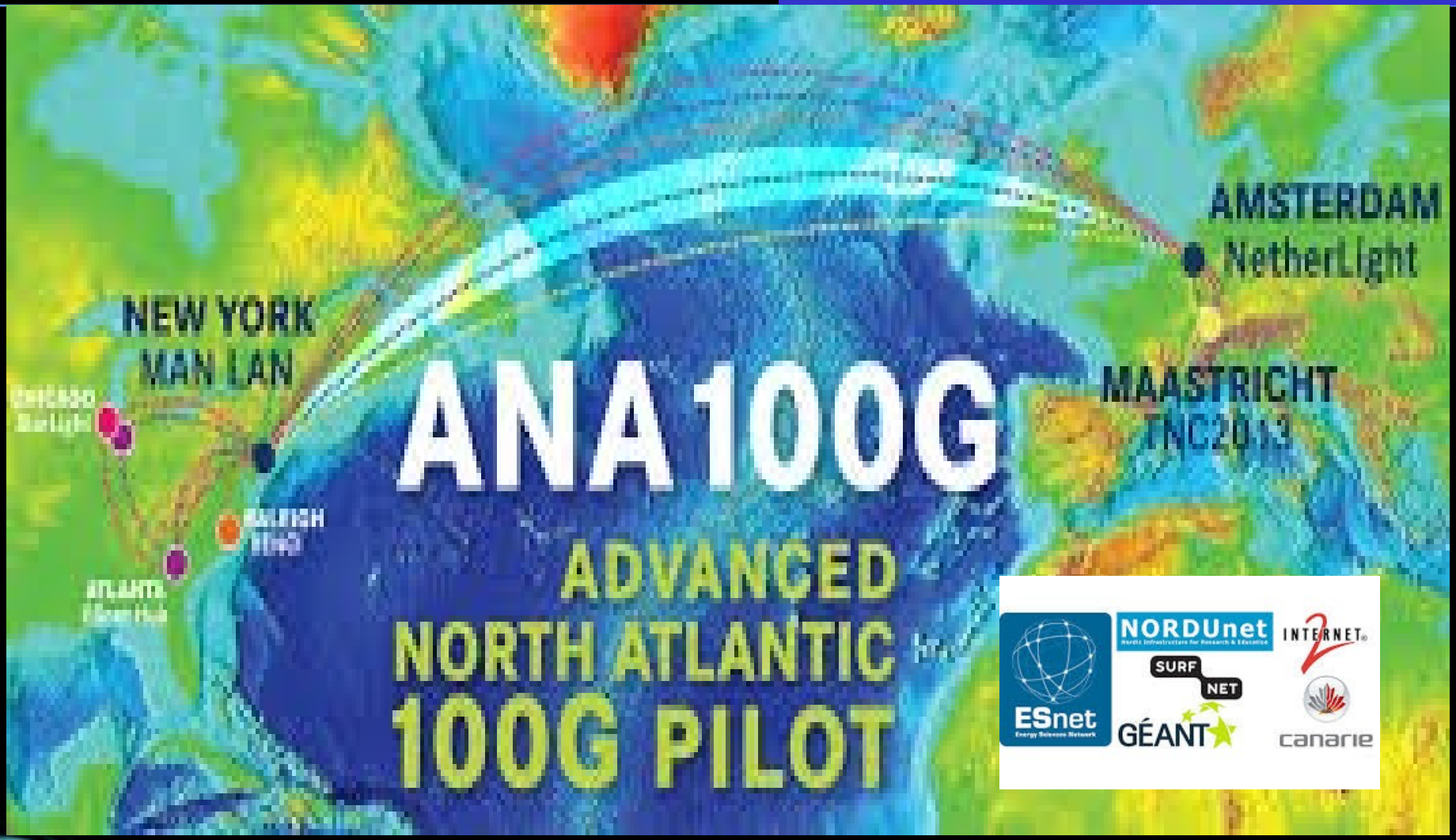



NIH-OCC Computational Genomics and Analytics 100 Gbps Testbed
StarLight
OARNET LR4
MANLAN LR10
IU LR10
BlueWater LR4
Sunnyvale
Salt Lake City
SC13
Chicago
New York 1
Kansas City
Cleveland
NIH 100 Gbps
To Ashburn
Washington DC
Denver
OneNet LR4
Tulsa
MAX LR4
WIX LR10
Los Angeles
Atlanta
El Paso
10G
Houston
100 G G&A Testbed
AL2S Node w/10gig L3 interconnect
AL2S Node w/100gig L3 interconnect
AL2S Node
Customers / Peerings

Selected SC13 Results & Next Step

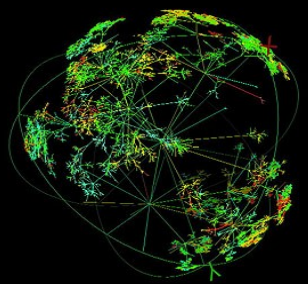
- Demonstrations of 98.7 Gbps Over WAN (Thousands of Miles), (Thanks To ESnet!) Using Multiple High Performance Transport Protocols
- Also 99 Gbps Demonstrated Over LAN, Using Multiple High Performance Transport Protocols
- At SC13 NSAS GSFC Demonstration Set World Record for WAN Disk to Disk Transfer >93 Gbps
- Multiple Other Demonstrations of 40 Gbps-60, 70 Gbps Gbps-80 Gbps Data Transfers
- ***Next Step = International Large Capacity Data Stream Transfers At Near 100 Gbps***

Advanced North Atlantic 100G Pilot



Transporting Big Data Internationally: SDSS Transport and Visualization

Joe Mambretti, Director, (j-mambretti@northwestern.edu)
International Center for Advanced Internet Research (www.icair.org)
Northwestern University
Director, Metropolitan Research and Education Network (www.mren.org)
Co-Director, StarLight, (www.startap.net/starlight)
Cees de Laat, System and Network Engineering Research Group
Informatics Institute
University of Amsterdam
Jim Chen, Fei Yeh
International Center for Advanced Internet Research (www.icair.org)
Northwestern University
Ralph Koning, Poala Grosso, System and Network Engineering Research Group
Informatics Institute
University of Amsterdam
Robert Grossman, Renuka Arya
Laboratory for advanced Computing, Open Cloud Consortium
University of Chicago
Todd Margolis, Associate Scientist, UCSD
Alex Szalay
Alumni Centennial Professor
Department of Physics and Astronomy,
Johns Hopkins University
Architect for the Science Archive of
the Sloan Digital Sky Survey and
Director of the Institute for Data Intensive Engineering and Science



STARLIGHT SM

Multiple HPC Cloud Computing Testbeds Specifically Designed for Science Research

At Scale Experimentation
Integerated With High Performance Networks

SC13 100 Gbps SCINet NRE Demonstrations

- 100 Gbps Networks for Next Generation Petascale Science Research and Discovery
- Data-Scope at 100 Gbps Across National Data-Intensive Computational Science Testbeds
- Next Generation Genome Sequencing Data at 100 Gbps
- The Global Environment for Network Innovations (GENI) at 100 Gbps
- HPC Open Science Data Cloud (OSDC) for Data Intensive Research at 100 Gbps
- Using Remote I/O for Large-Scale Long Distance Data Storage and Processing
- (Planned But Not Executed Because of Fiber Break: International ExoGENI @ 40G: Dynamic Monitoring and Adaptation of Data Driven Scientific Workflows Using Federated Cloud Infrastructure)

Overview of the Alex's Data-Scope

- Data-Scope Is An NSF Funded Novel Instrument Designed To Observe Extremely Large Data Sets
- Data-Scope Was Implemented To Undertake Highly Data-Intensive Analyses Simply Not Possible Anywhere Else Today
  - 6.5 PB of Fast Storage
  - Sequential IO 500Gbytes/sec
  - 10G Interconnect local
  - 100G External Connectivity
  - 96 Servers
- 100 Gbps LAN and WAN Connectivity

Source: Alex Szalay

DataScope Data Sets

Data Sets On the System Include

- Large N-body Simulations From Astrophysics (>1PB)
- Computational Fluid Dynamics (300TB)
- Astronomical Data From the SDSS (400TB)

Soon coming

- Bioinformatics and Neuroscience (2-300TB each)
- Ocean Circulation Models (600TB)

Main Challenge

- How To Bring the Data Sets To the Data-Scope (Transport)
- Most Are Generated Externally, e.g., Teragrid/XSEDE, the Oak Ridge Titan, Telescopes, Sequencers, Etc

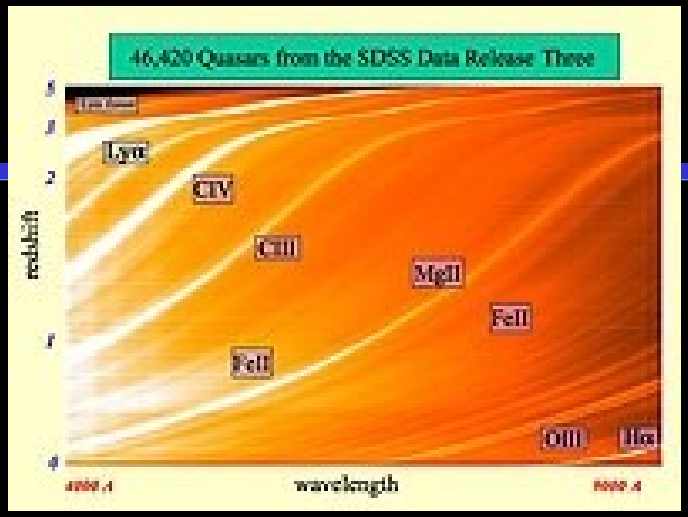
Source: Alex Szalay

Sloan Digital Sky Survey (SDSS)

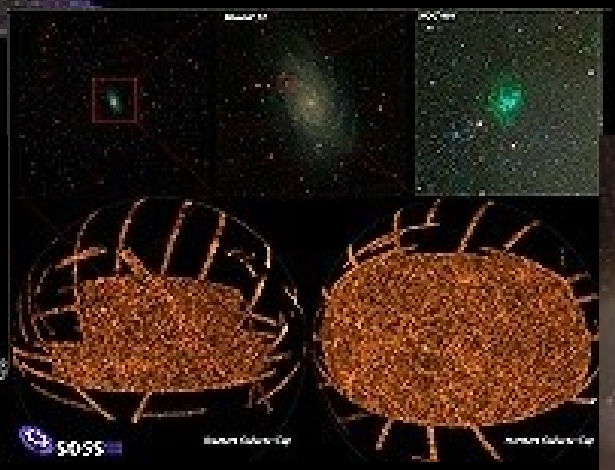
- SDSS Is Officially Described As "One Of the Most Ambitious And Influential Surveys In the History of Astronomy."
- The Project Obtained Deep, Multi-Color Images Covering More Than a Quarter of the Sky and Created 3D Maps Comprised of More Than 930,000 Galaxies and More Than 120,000 Quasars.
- The SDSS Created The First Detailed 3D Map of Much of the Universe
- SDSS Data Have Been Incrementally Released To Scientific Community And the General Public.
- The Current Phase of the SDSS Will Operate Through 2014.
- The Data Are Stored In About 100s of Files and Total Approximately 400 TB
- The Data Are Used for Analysis, To Create Models, and To Create Images
- Transporting This Data Quickly Over Long Distances Would Be Useful
- In 2012, Experiments Undertaken Successfully To Investigate These Issues at 10 Gbps – e.g., Sending SDSS data from JHU to StarLight to ORNL's Jaguar, Building Models and Sending Them Back to JHU (146TB)

SDSS Images

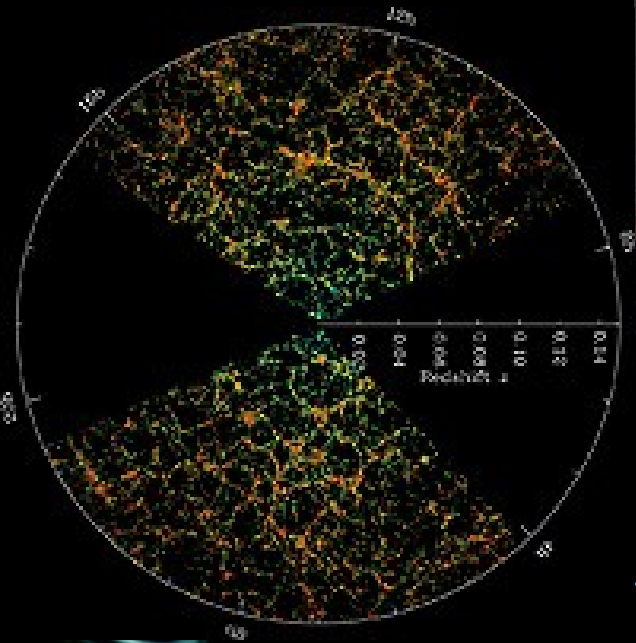
Field of Streams



Quasar Spectra



Composite Survey Image



Galaxy Map

Whirpool Galaxy

STARLIGHT SM

Renuka Arya's (RA's) Summary of Results: UoC=>UvA

- Data Flow: From UoC to UvA
-

  1) Test 1: Single Flow Transfer of Files 1 to 9 Using UDR. Throughput 85 -90 MB/s = 680 - 720 Mbps

- 2) Test 2: Two Flow Transfer of Files 10 -19 & 20 -29 Using UDR. Throughput 170-180 MB/s = 1.36 - 1.44 Gbps

- 3) Test 3: Three Flows Transfer of Files 31-39, 40-49 & 50-59 Using UDR. Throughput 240 - 250 Mb/s = 1.92 - Essentially 2 Gbps

- 4) Test 4: Transferr Of Single File Ssing SCP: 12 -15 MB/s = 96 - 120 Mbps

- 5) Transferred All Remaining Files Using UDR.

RA's Summary of Results: UvA=>UoC

Data Flow : From UvA to UC

1) Test 1 : Transferred a 85GB File Using UDR.
Throughput ~84 MB/s =672 Mbps
2) Test 2: Transferred a 85GB File Using UDPIPE.
Throughput : 1.05 Gbps

NUTTCP: From UvA to UoC (Could Not Be Run in the Reverse Direction)
1) Average Throughput : ~2.2 Gbps, With Zero Retransmissions

RA's Issue 1: udpipe

- udpipe Is a UDT Data Transfer Protocol Based On The Functionality of netcat. There Were Issues Transferring Data From UoC to UvA Using UDPIPE
- Packet Captures (And Other Network Utilities Including iftop, jnettop, vnstat) Show Sender/Receiver Sending and Receiving Data Respectively.
- Utilities Show the Actual Data Begin To Be Transferred and ACKs Returned From Receiver. However, The Application Stoped Writing To the File After Researching a Certain Byte (Varies For Different Files).
- The Application Was Waiting To Read From Socket, Not Obtaining Any Data (After Certain Point Is Reached).
- However, UDPIPE Worked Fine For Other Hosts and Worked In Reverse Direction - From UvA To UoC.

RA's Issue 2: Support for Jumbo Frames

- StarLight and Its Networks Have Supported Jumbo Frames For Many, Many Years.
- However, Not All Sites Support Jumbo Frames.
- The Trace Path Showed That Jumbo Frame Path Support Was Implemented from UvA to UoC But Not From UoC (local) to UvA.

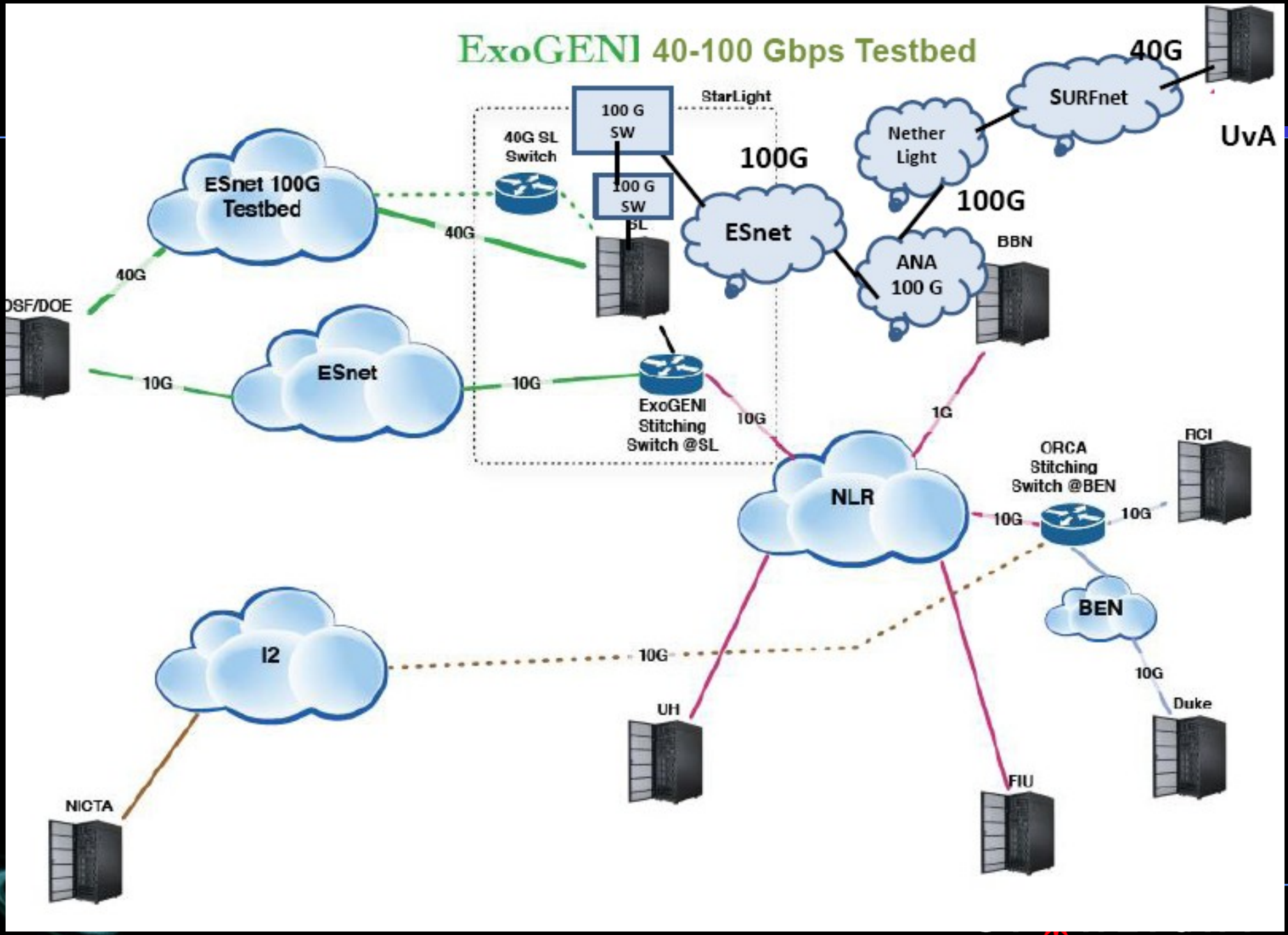
MultiPath TCP Over 100 Gbps

- MPTCP Demonstration Organized for TERENA 2013 Maastricht, Netherlands.
- Lead: Ronald van der Pol, SURFnet With Gerben van Malenstein, Migiel de Vos, (SURFnet), Michael Bredel, Artur Barczyk, Azher Mughal (Caltech), Benno Overeinder (NLnet Labs), Niels van Adrichem (TU Delft), Christoph Paasch (Universite Catholique de Louvain), Joe Mambretti, Jim Chen (iCAIR)
- Demonstration Showed How Multipathing, OpenFlow and Multipath TCP (MPTCP) Can Support Large File Transfers Among Data Centres (Multiple 10s of Gbps).
- Based On Customized Intercontinental Multipathed OpenFlow Testbed Spanning CERN, NetherLight, StarLight and Maastricht.
- Used Multiple Paths Simultaneously
- The Testbed Was Based On OpenFlow Switches and Multiple Link Paths Between Servers. An OpenFlow Application Provisioned Multiple Paths Between Servers and MPTCP Was Used on Servers to Simultaneously Send Traffic Across All Paths
- 20 Gbps Transported

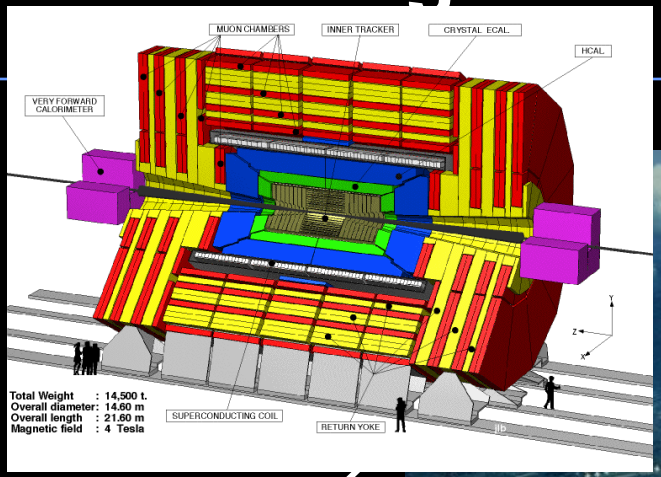
- Path: CERN⇔NetherLight⇔ANA⇔ESnet⇔StarLight

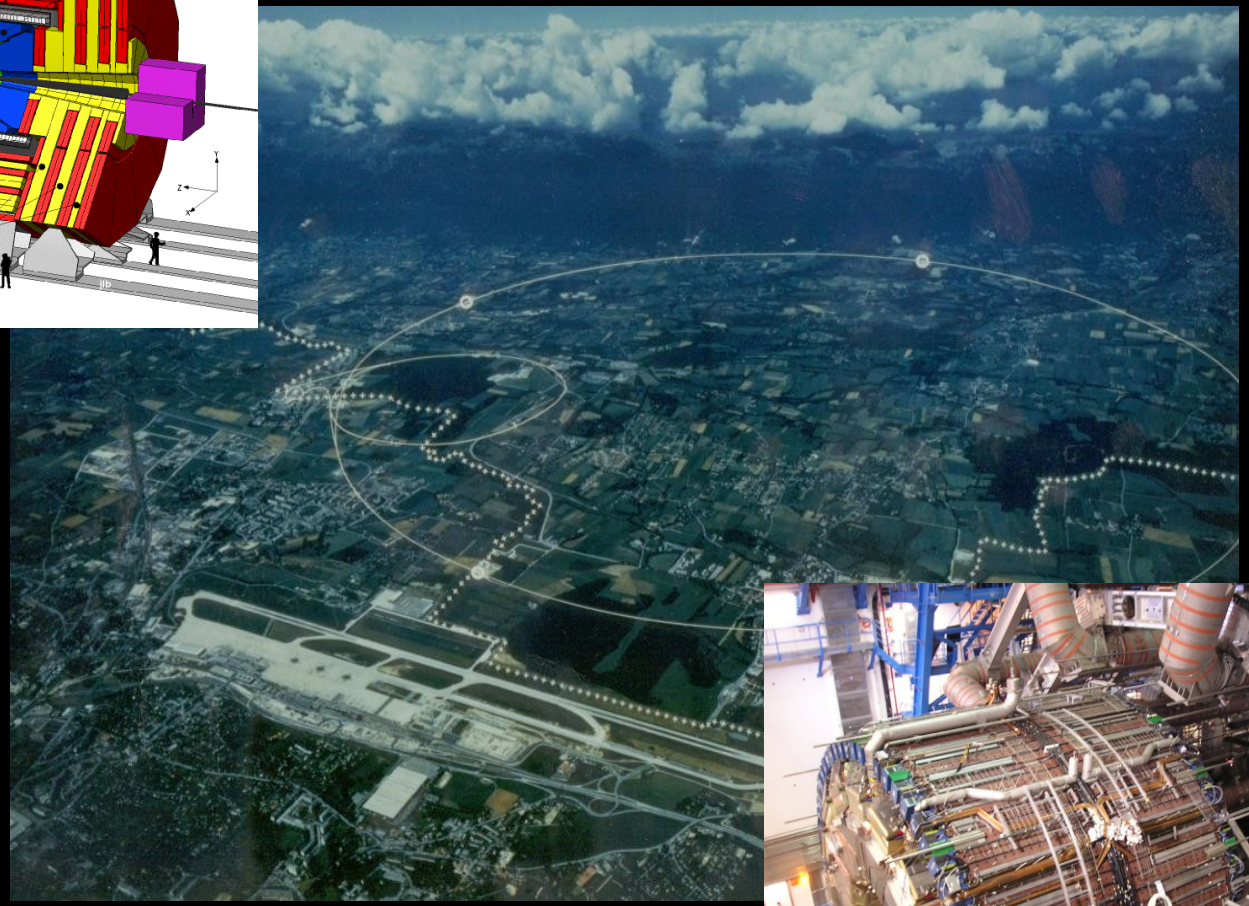
ExoGENI @ 40 G and 100 G

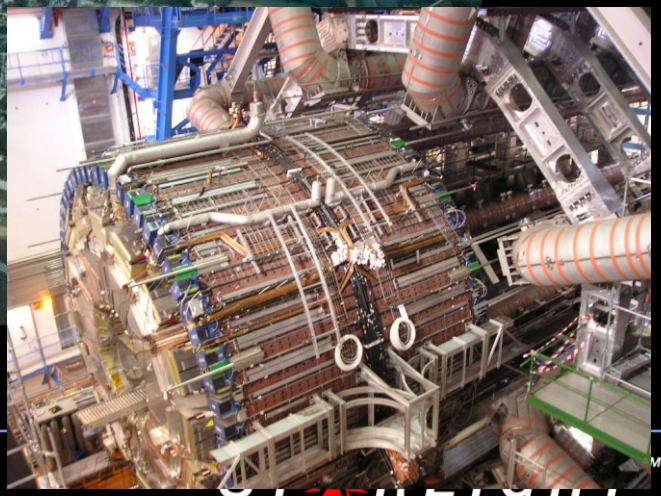
- Multiple Organization Partnership Initiative
- Renaissance Computing Institute, Lawrence Berkley Labs, ESnet, ESnet 100 Gbps Testbed, ExoGENI, Global Environment for Network innovations (GENI), International Center for Advanced Internet Research, University of Amsterdam, StarLight International/National Communications Exchange Facility, Metropolitan Research and Education Network (MREN)
- Integration Of Dynamic Provisioning and 40 G - 100 G Paths
- Based On the Open Resource Control Architecture (ORCA), One of the GENI Control Frameworks
- Demonstrated 38 Gbps Flows From StarLight ExoGENi RackTo University of Amsterdam ExoGENI Rack
- StarLight ExoGENI⇔ESnet⇔ANA⇔NetherLight ⇔SURFnet⇔University of Amsterdam⇔ExoGENI


ExoGENI 40-100 Gbps Testbed
40G
SURFnet
UvA
StarLight
100 G SW
40G SL Switch
100G
Nether Light
100G
ESnet 100G Testbed
40G
100 G SW SL
ESnet
ANA 100 G
BBN
40G
OSF/DOE
ESnet
10G
10G
ExoGENI Stitching Switch @SL
10G
1G
NLR
ORCA Stitching Switch @BEN
RCI
10G
10G
BEN
I2
10G
10G
Duke
UH
FIU
NICTA

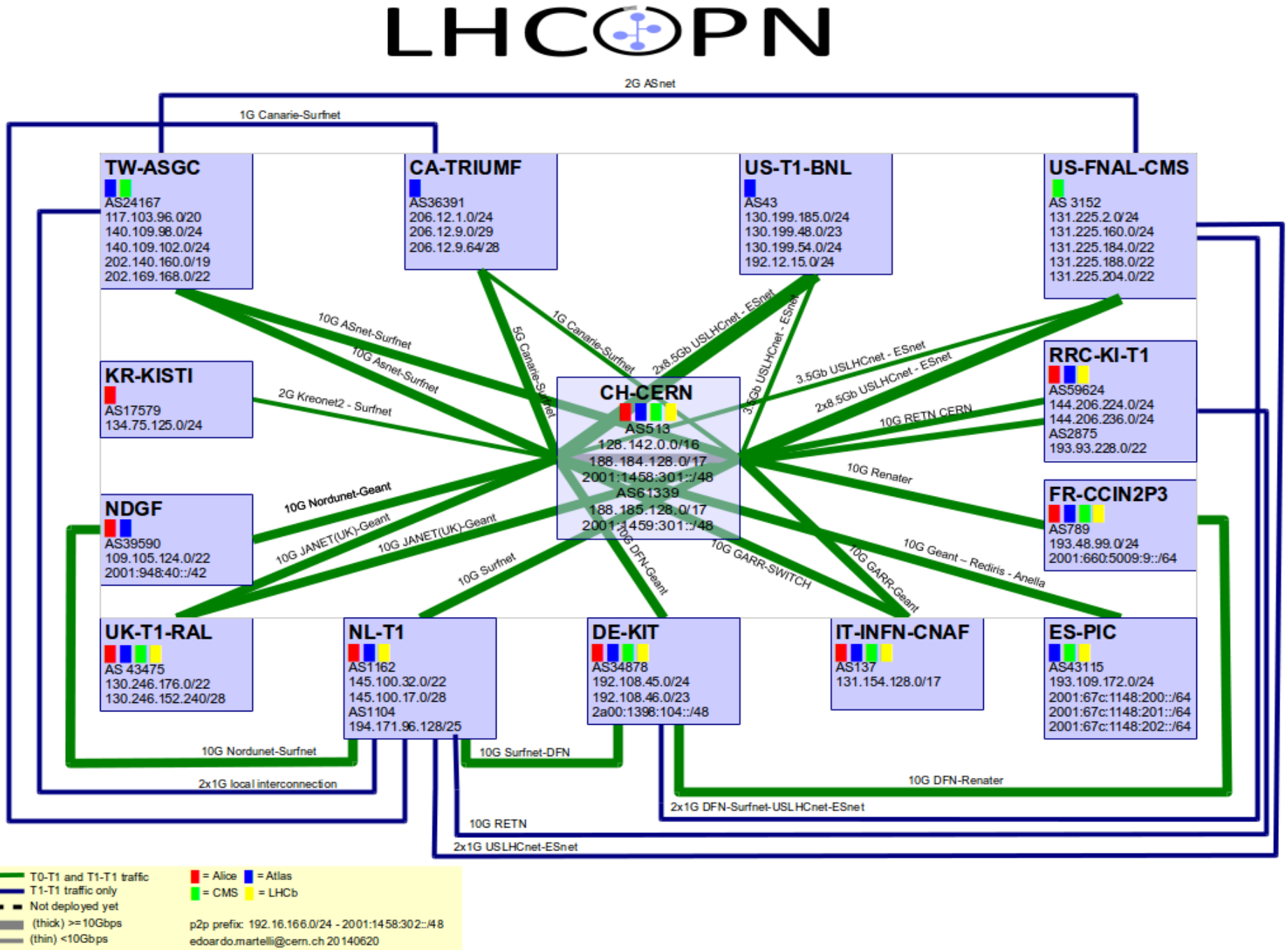
Large Hadron Collider at CERN

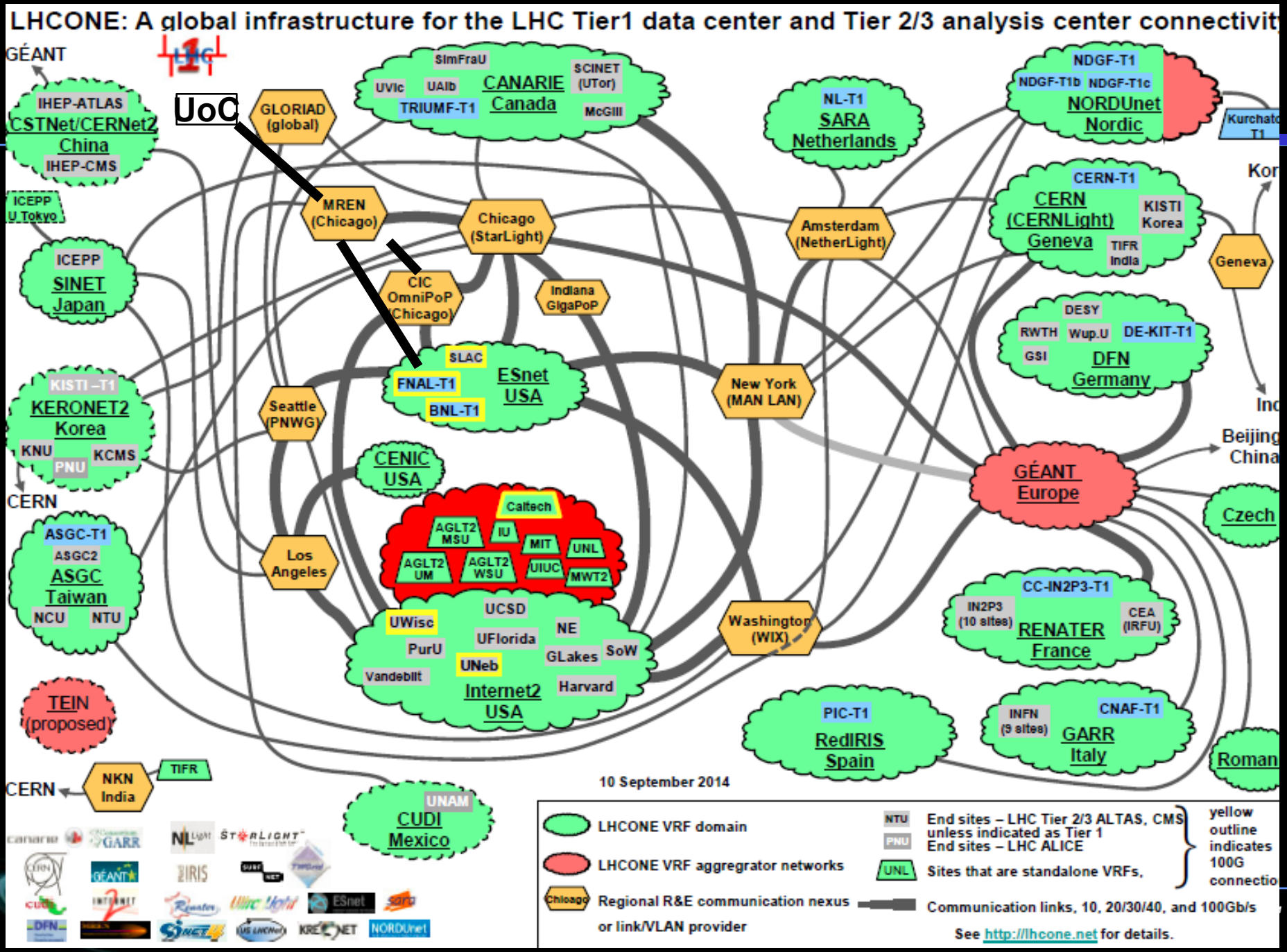






iCAIR

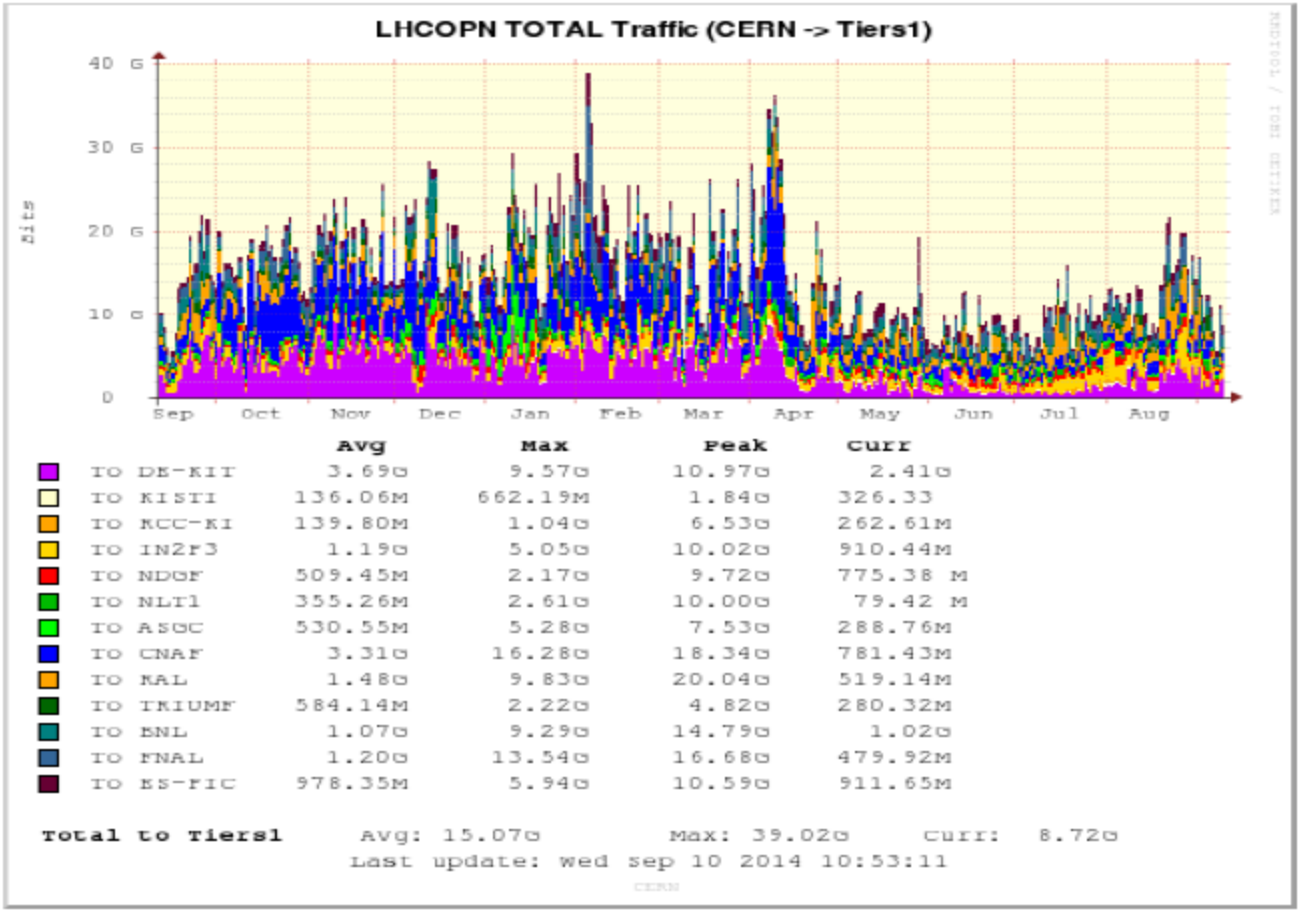

LHCOPN
2G ASnet
1G Canarie-Surfnet
TW-ASGC
AS24167
117.103.96.0/20
140.109.98.0/24
140.109.102.0/24
202.140.160.0/19
202.169.168.0/22
CA-TRIUMF
AS36391
206.12.1.0/24
206.12.9.0/29
206.12.9.64/28
US-T1-BNL
AS43
130.199.185.0/24
130.199.48.0/23
130.199.54.0/24
192.12.15.0/24
US-FNAL-CMS
AS 3152
131.225.2.0/24
131.225.160.0/24
131.225.184.0/22
131.225.188.0/22
131.225.204.0/22
10G ASnet-Surfnet
10G Asnet-Surfnet
5G Canarie-Surfnet
1G Canarie-Surfnet
2x8.5Gb USLHCnet - ESnet
3.5Gb USLHCnet - ESnet
3.5Gb USLHCnet - ESnet
2x8.5Gb USLHCnet - ESnet
KR-KISTI
AS17579
134.75.125.0/24
2G Kreonet2 - Surfnet
CH-CERN
AS513
128.142.0.0/16
188.184.128.0/17
2001:1458:301::/48
AS61339
188.185.128.0/17
2001:1459:301::/48
RRC-KI-T1
AS59624
144.206.224.0/24
144.206.236.0/24
AS2875
193.93.228.0/22
10G RETN CERN
10G Renater
10G Nordunet-Geant
10G JANET(UK)-Geant
10G JANET(UK)-Geant
10G Surfnet
10G DFN-Geant
10G GARR-SWITCH
10G GARR-Geant
10G Geant – Rediris - Anella
NDGF
AS39590
109.105.124.0/22
2001:948:40::/42
FR-CCIN2P3
AS789
193.48.99.0/24
2001:660:5009:9::/64
UK-T1-RAL
AS 43475
130.246.176.0/22
130.246.152.240/28
NL-T1
AS1162
145.100.32.0/22
145.100.17.0/28
AS1104
194.171.96.128/25
DE-KIT
AS34878
192.108.45.0/24
192.108.46.0/23
2a00:1398:104::/48
IT-INFN-CNAF
AS137
131.154.128.0/17
ES-PIC
AS43115
193.109.172.0/24
2001:67c:1148:200::/64
2001:67c:1148:201::/64
2001:67c:1148:202::/64
10G Nordunet-Surfnet
10G Surfnet-DFN
2x1G local interconnection
10G DFN-Renater
2x1G DFN-Surfnet-USLHCnet-ESnet
10G RETN
2x1G USLHCnet-ESnet
T0-T1 and T1-T1 traffic
T1-T1 traffic only
Not deployed yet
(thick) >=10Gbps
(thin) <10Gbps
= Alice
= Atlas
= CMS
= LHCb
p2p prefix: 192.16.166.0/24 - 2001:1458:302::/48
edoardo.martelli@cern.ch 20140620


LHCONE: A global infrastructure for the LHC Tier1 data center and Tier 2/3 analysis center connectivity
GÉANT
IHEP-ATLAS
CSTNet/CERNet2
China
IHEP-CMS
ICEPP
U Tokyo
ICEPP
SINET
Japan
KISTI –T1
KERONET2
Korea
KNU
PNU
KCMS
CERN
ASGC-T1
ASGC2
ASGC
Taiwan
NCU
NTU
TEIN
(proposed)
CERN
NKN
India
TIFR
UoC
GLORIAD
(global)
MREN
(Chicago)
CIC
OmniPoP
Chicago)
Chicago
(StarLight)
Indiana
GigaPoP
SLAC
FNAL-T1
BNL-T1
ESnet
USA
Seattle
(PNWG)
CENIC
USA
Los
Angeles
Caltech
AGLT2
MSU
IU
MIT
UNL
AGLT2
UM
AGLT2
WSU
UIUC
MWT2
UCSD
UWisc
NE
UFlorida
PurU
GLakes
SoW
UNeb
Vandebilt
Harvard
Internet2
USA
UNAM
CUDI
Mexico
SInFraU
UVic
UAlb
CANARIE
Canada
SCINET
(UTor)
TRIUMF-T1
McGill
NL-T1
SARA
Netherlands
Amsterdam
(NetherLight)
New York
(MAN LAN)
Washington
(WIX)
NDGF-T1
NDGF-T1b
NDGF-T1c
NORDUnet
Nordic
CERN-T1
CERN
(CERNLight)
Geneva
KISTI
Korea
TIFR
India
DESY
RWTH
Wup.U
DE-KIT-T1
GSI
DFN
Germany
GÉANT
Europe
CC-IN2P3-T1
IN2P3
(10 sites)
RENATER
France
CEA
(IRFU)
PIC-T1
RedIRIS
Spain
INFN
(9 sites)
CNAF-T1
GARR
Italy
Kurchato
T1
Kor
Geneva
Inc
Beijing
China
Czech
Roman
10 September 2014
LHCONE VRF domain
LHCONE VRF aggregator networks
Chicago
Regional R&E communication nexus
or link/VLAN provider
NTU
PNU
End sites – LHC Tier 2/3 ALTAS, CMS
unless indicated as Tier 1
End sites – LHC ALICE
UNL
Sites that are standalone VRFs,
yellow
outline
indicates
100G
connectio
Communication links, 10, 20/30/40, and 100Gb/s
See http://lhcone.net for details.

LHCONE Trails and Experiments

- Series of Experiments With LHC Traffic Injection Into The ANA Path
- Also, Removing LHC Streams From the ANA
- And Then Replacing The LHC Streams
- Then Repeating the Insertions
- All Processes Were Successful

LHCOPN traffic trend



	Avg	Max	Peak	Curr
TO DE-KIT	3.69G	9.57G	10.97G	2.41G
TO KISTI	136.06M	662.19M	1.84G	326.33
TO KCC-KI	139.80M	1.04G	6.53G	262.61M
TO IN2P3	1.19G	5.05G	10.02G	910.44M
TO NDGF	509.45M	2.17G	9.72G	775.38 M
TO NLT1	355.26M	2.61G	10.00G	79.42 M
TO ASGC	530.55M	5.28G	7.53G	288.76M
TO CNAF	3.31G	16.28G	18.34G	781.43M
TO KAL	1.48G	9.83G	20.04G	519.14M
TO TRIUMF	584.14M	2.22G	4.82G	280.32M
TO BNL	1.07G	9.29G	14.79G	1.02G
TO FNAL	1.20G	13.54G	16.68G	479.92M
TO ES-PIC	978.35M	5.94G	10.59G	911.65M

Total to Tiers1 Avg: 15.07G Max: 39.02G Curr: 8.72G

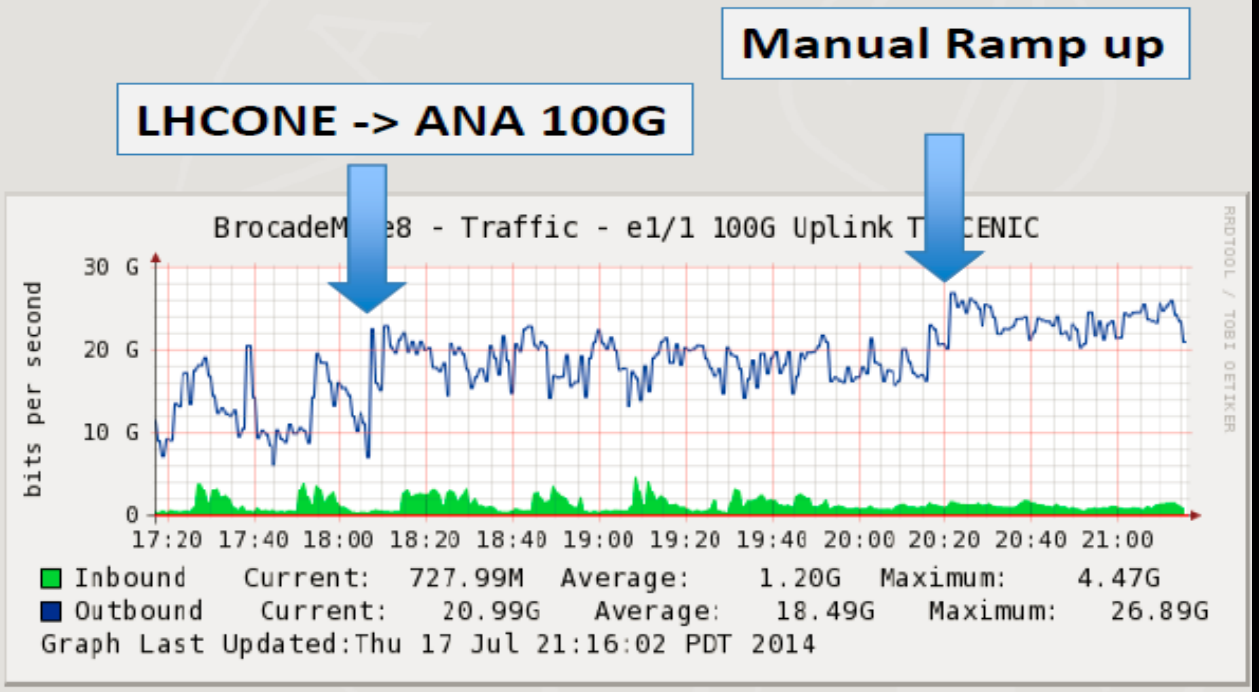
Last update: Wed Sep 10 2014 10:53:11

STARLIGHT SM

LHCONE Trans-Atlantic Testing

Capacity vs Application Optimization

- Increases transfers observed during the LHCONE ANA integration.
- Manual ramp up was just a test because remote PhEDEx sites were not subscribing enough transfers to FTS links (e.g. Caltech - CNAF)



FIONA (Flash I/O Network Appliance)

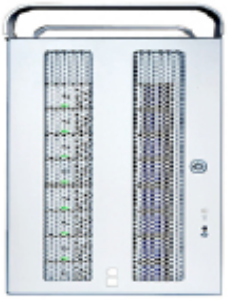
- FIONA: Innovative Network Appliance for Big Data
- Flash I/O Network Appliance (FIONA), Engineered For Large Data Flows.
- Designed By Qualcomm Institute's Tom DeFanti and SDSC's Phil Papadopoulos (UC San Diego's Research Cyberinfrastructure (RCI) Group)
- FIONA: Low-Cost, Flash Memory-Based Data Server Appliance That Researchers Can Install In Their Labs As a "Big Data Hub," Interfacing Data-Generating Scientific Instruments and High-Speed Optical Networks.
- Can Also Be Used As Big Data Analysis Workstation Driving Interactive Big Data Screens (HD Single Screens To Ultra Resolution Tiled Display Walls).

FIONA

- FIONA: Inexpensive, Low-power, Low-noise Appliance Comprised of Flash Drives, Local Disk Drives, High Performance Graphic Processing Units (GPUs).
- FIONA Appliance Directly Connects To 40Gbps Networks and Drive Display Walls Directly.
- FIONA Can Drive Displays, Fetch Data Over the Network, and Cache Necessary Data Locally.
- FIONA Manages Data Transport and Local Caching, and Works Seamlessly with Rendering Algorithms for Visualization
- Experiments Undertaken With Network Distances

FIONA Building Blocks

FIONA Building Blocks



Desktop Form Factor
With Single Socket Server $3K

Total Cost - $6K - $7K

1X - RAID Controller LSI HBA
SAS
~ $300 each

1 X - Dual 10GbE
Myricom, Inc or
Dual Mellanox 40GbE
~$800 each

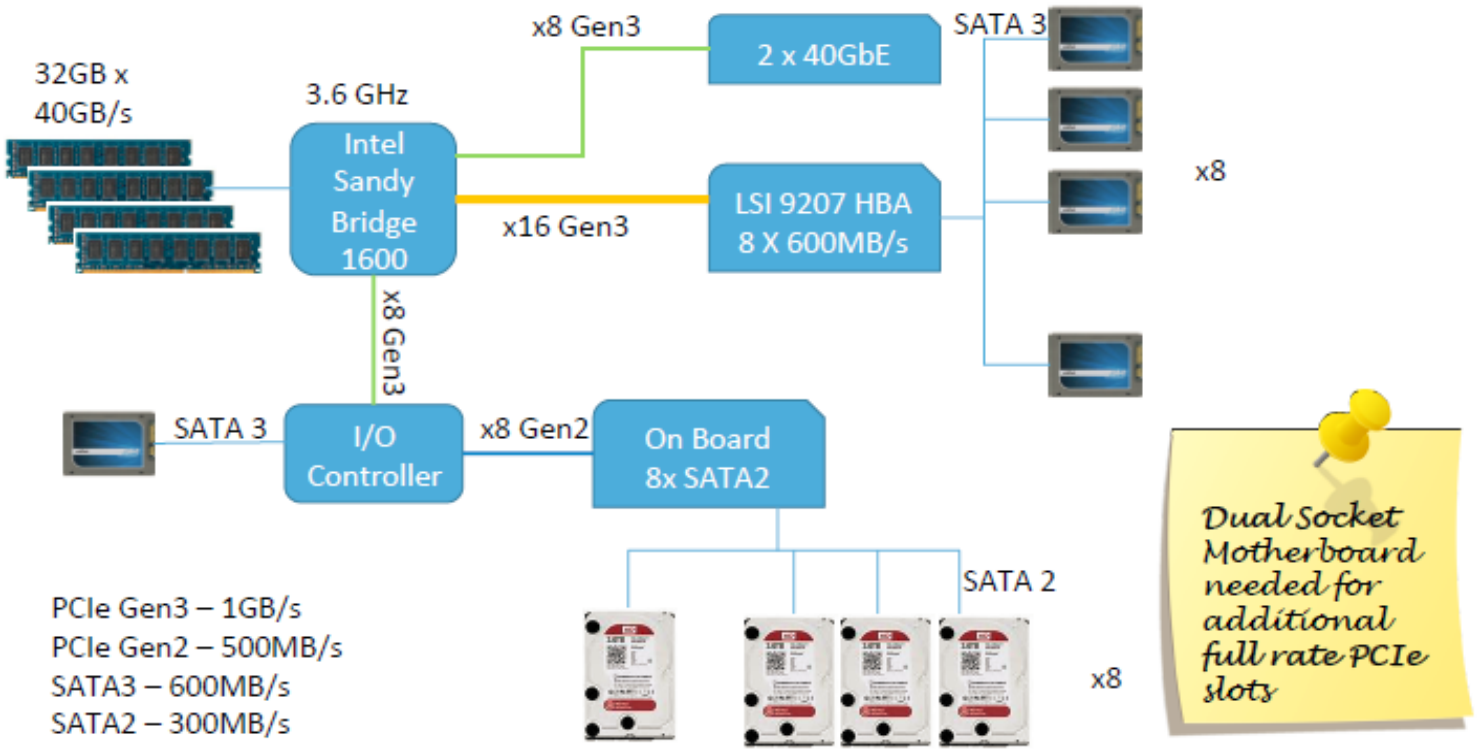
8 X 3TB SATA Drive
Hitachi/Seagate
$150.00 each

8 x 240GB SSD
$180.00 each
500+MB/sec

Source: UCSD

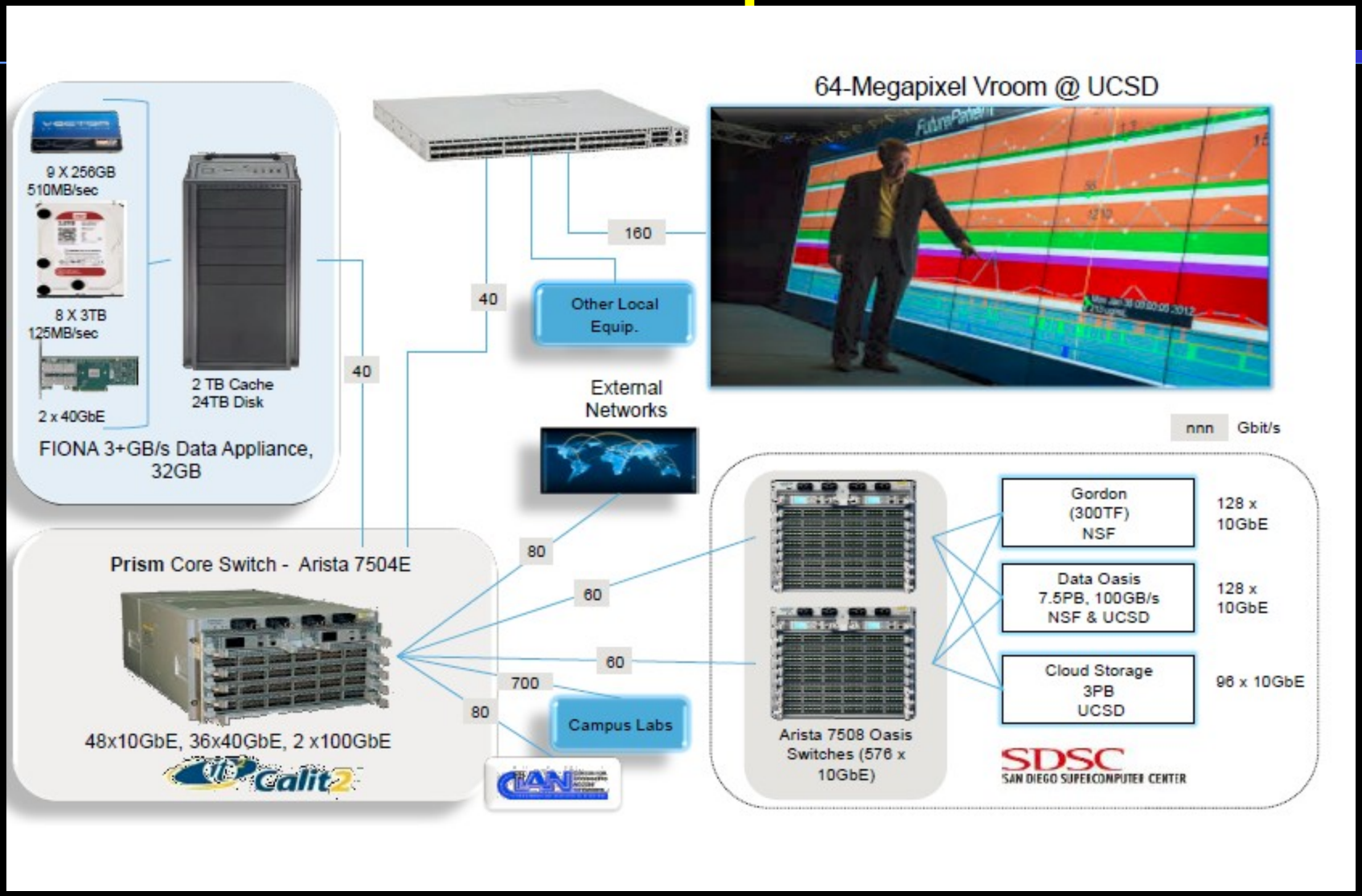
FIONA Speeds and Feeds

FIONA Internal Speeds and Feeds (some channels close to saturation)



Source: UCSD

FIONA Components

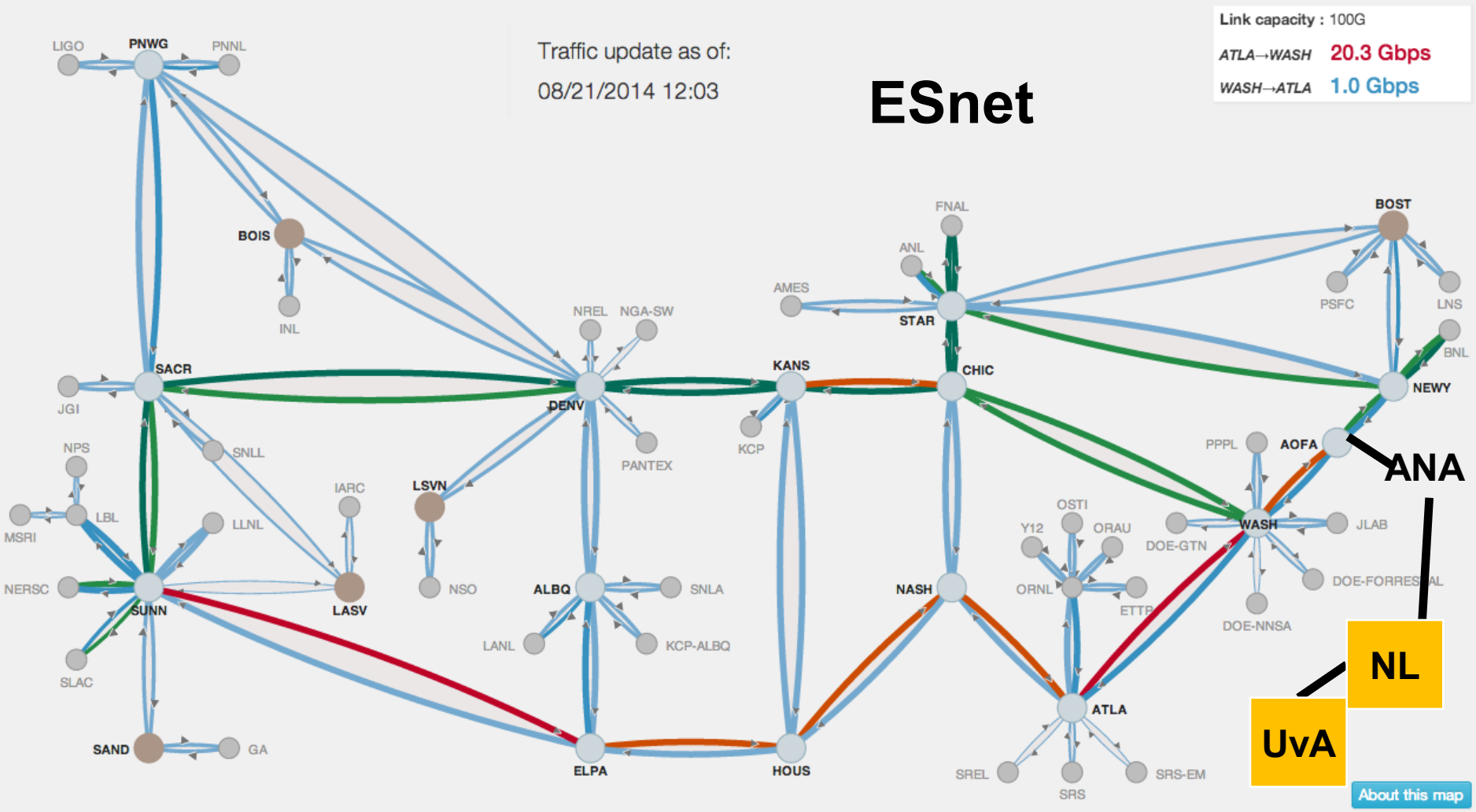


Source: UCSD

FIONA and TransLight/StarLight/NetherLight/UvA

- FIONA Can be Implemented Over WANs, Including Internationally (Ref ANA Experiments)
- Experiments Conducted Via the TransLight/StarLight Petascale Science Services Prototype Facility Initiative
- Funded By the National Science Foundation
- Testbed Established On the ESnet National 100 Gbps Production Network
- FIONA⇔CalIT2⇔UCSD⇔CENIC⇔Sunnyvale⇔ESnet⇔ANA⇔NetherLight ⇔SURFnet⇔University of Amsterdam⇔FIONA
- E2E Experiments Over WANs (Across the World)
- Ref: Topology

FIONA National and International Testbed



E⇔E ~ 34 Gbps

STARLIGHT℠

Summary of Multiple ANA Experimental Results

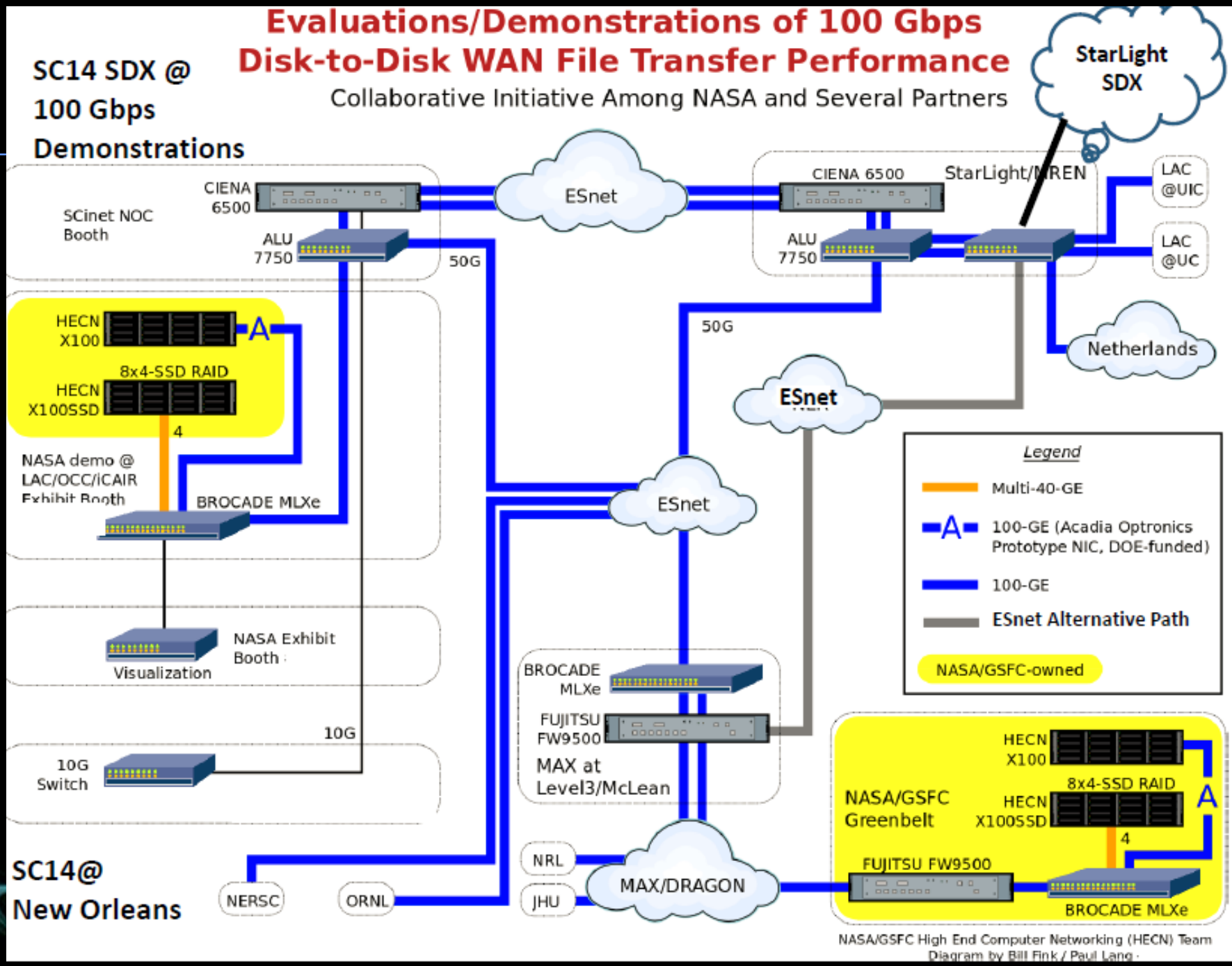
- ***High Tuned Components at All Levels Is Essential***
- High Capacity E2E Paths Are Also Essential
- Demonstrated E⇔E 98.6+ Gbps Sustained Individual Flows Over WANs For Long Periods With No Packet Loss Memory⇔Memory Within US
- Demonstrated E⇔E 93.4+ Gbps Sustained Individual Flows Over WANs For Long Periods With No Packet Loss Disk⇔Disk Within US (e.g., GSFC⇔SC13)
- Demonstrated Multiple Applications Streaming Internationally at 30 Gbps - 38 Gbps Memory⇔Memory, Disk⇔Disk Over ANA
- Only Real Constraint Is Capacity At Edge Device
- Many Additional Experiments/Demonstrations Are Planned, e.g., At SC14

Plans For SC14 In New Orleans

- 10 Separate Sets of 100 Gbps Demonstrations Are Being Planned
- Several Incorporated Into Software Defined Networking Exchanges (SDXs)
- Several Incorporating Trans-Oceanic Paths
- Ref: Topology

Evaluations/Demonstrations of 100 Gbps Disk-to-Disk WAN File Transfer Performance

Collaborative Initiative Among NASA and Several Partners



NASA/GSFC High End Computer Networking (HECN) Team
Diagram by Bill Fink / Paul Lang

Prototype SDX BioInformatics Exchange

- Demonstration Showcase Designed for SC14 in New Orleans, Louisiana
- Builds On Previous Initiatives Developing Services for Computational BioInformatics and Computational Genomics at 100 Gbps
- Previous Demonstrations At Several Conferences, Including SC13
- The SDX BX Is Veing Designed Specifically for BioInformatics/Computational Genomic Worksflows
- Appears As a Private Exchange for BioInformatics Research Communities

Ciena's OPⁿ research network testbed

Research-On-Demand 100 G Testbed (RODnet)

ciena

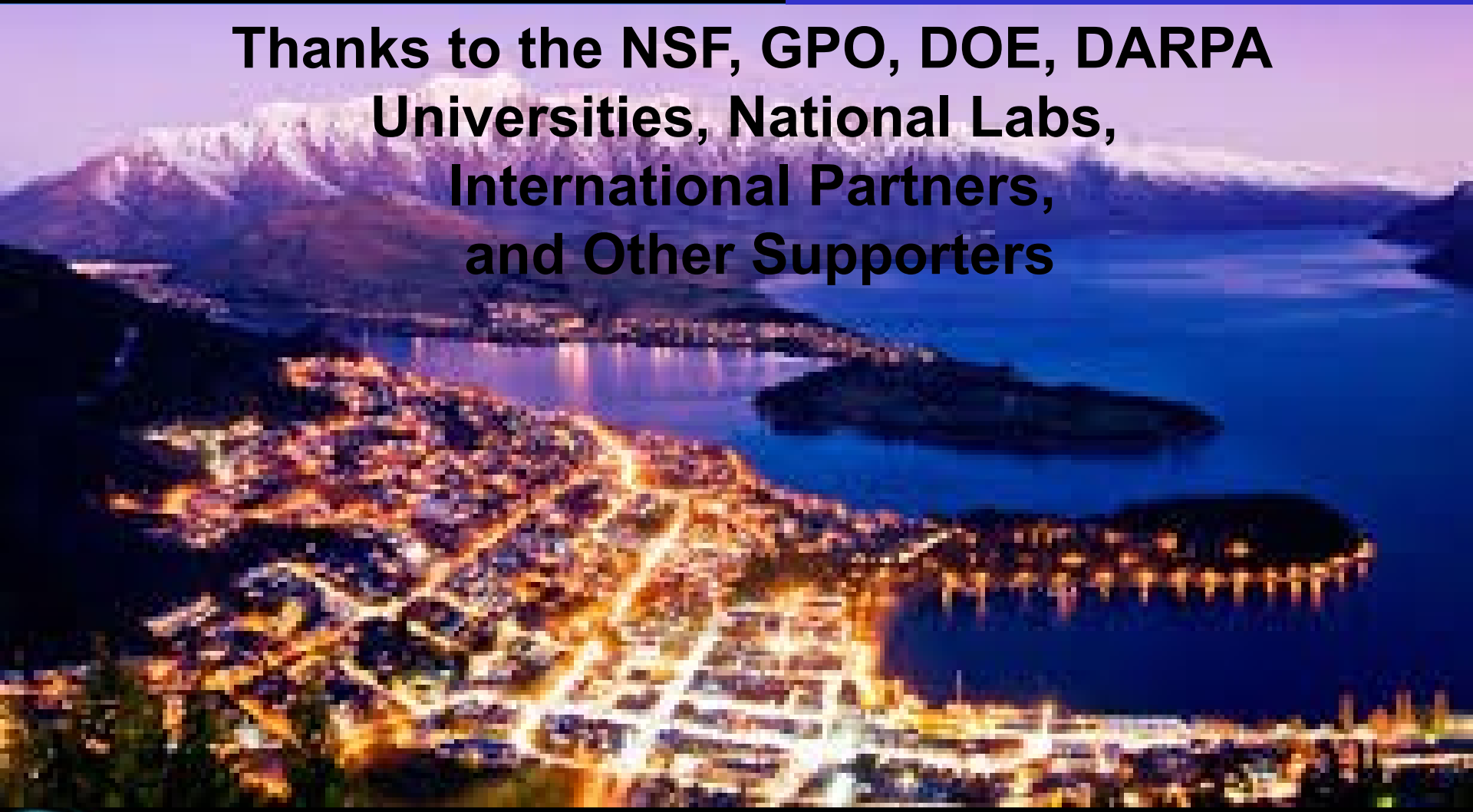
STARLIGHT℠

Summary

- The Architecture and Technology Components Required for Supporting High Capacity Individual Streams, i.e., Large Scale, High Performance, Wide Area Data Transport at 100 Gbps **Exist** And They Have Been Demonstrated At Several Forums, Primarily At SC Conferences
- However, These Are Not Yet Integrated Into a Readily Available *Service*
- Various Projects Today Are Attempting To Develop Such Services – Nationally and Internationally, e.g., Esnet and the Petascale Science Prototype Services Facility

StarLight/StarWave/GLIF Continually Progressing Forward!

www.startap.net/starlight



Thanks to the NSF, GPO, DOE, DARPA Universities, National Labs, International Partners, and Other Supporters