Open Storage Network

Alexander Szalay, Bloomberg Distinguished Professor, the Alumni Centennial Professor of Astronomy, and Professor, Department of Computer Science, Johns Hopkins University. Director, Institute for Data Intensive Science, Fellow of the American Academy of Arts and Sciences, PI Open Storage Network Project

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)
Director, StarLight, PI StarLight IRNC SDX,Co-PI Chameleon, PI-iGENI, PI-OMNINet (www.startap.net/starlight)

Global LamdaGrid Workshop 2018
Co-Located With NORDUNET Conference 22018
Helsingør, Denmark
September 20-21, 2018

iCAIR

STARLIGHT℠

GriPhyN

Data Intensive Science

NEPTUNE

A Fiber-Optic Telescope to Inner Space

NEES

GENOMES to LIFE

BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES

NG Digital Sky Survey

NEON

National Ecological Observatory Network

ATLAS

ALMA

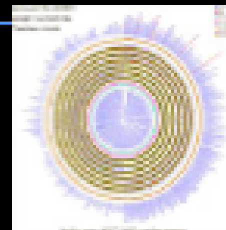
ALMA: Atacama Large Millimeter Array

ANDRILL: Antarctic Geological Drilling

www.andrill.org

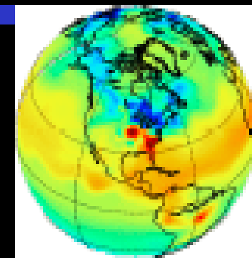
BIRN: Biomedical Informatics Research Network

www.nbirn.net



CAMERA metagenomics

camera.calit2.net



Carbon Tracker

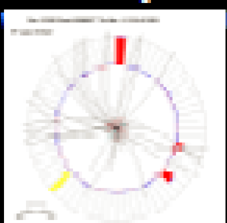
www.esrl.noaa.gov/gmd/ccgg/carbontracker

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

cl.oceanobservatories.org

ISS: International Space Station

www.nasa.gov/station

CLASS

Comprehensive Large-Array Stewardship System

www.class.noaa.gov

IVOA: International Virtual Observatory

www.ivoa.net

Two neutron stars orbit each other

LIGO

www.ligo.org

WLCG

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

PRAGMA

Pacific Rim Applications and Grid Middleware Assembly

www.pragma-grid.net

TeraGrid

www.teragrid.org

Open Science Grid

OSG

www.opensciencegrid.org

the globus alliance

Globus Alliance

www.globus.org

SKA

www.skatelescope.org

Sloan Digital Sky Survey

www.sdss.org

XSEDE

www.xsede.org

LIGO

Compilation By Maxine Brown

STARLIGHT℠

Large Scale Data Intensive Science Motivates the Creation of Next Generation Communications

- **Large Scale, Data (and Compute) Intensive Sciences Encounter Technology Challenges Many Years Before Other Domains**
- **Resolving These Issues Creates Solutions That Later Migrate To Other Domains**
- **30+ Year History of Communication Innovations Has Been Driven Primarily By Data and Compute Intensive Sciences**
- **Best Window To the Future = Examining Requirements of Data and Compute Intensive Science Research**
- **Science Is Transitioning From Using Only Two Classic Building Blocks, Theory and Experimentation To Also Utilizing a Third – Modeling and Simulation – With Massive Amounts of Data**
- **Petabytes, Exabytes, Zettabytes**
- **For Communications, Data Volume Capacity Not Only Issue, But a Major Issue**

Petascale Computational Science

For Decades, Computational Science Has Driven Network Innovation Today – Petascale Computational Science

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

National Center for Supercomputing Applications, UIUC

XSEDE

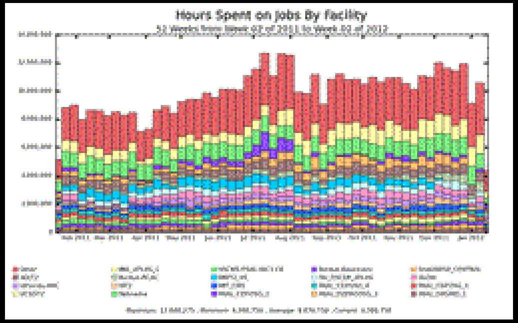
- Extreme Science and Engineering Discovery Environment (XSEDE)
- Goal: Create a Distributed Computational Science Infrastructure to Enable Distributed Data Sharing and High-Speed Computing for Data Analysis and Numerical Simulations
- Builds on Prior Distributed TeraGrid

Open Science Grid: Selected Investigations

DNA Modeling

Gravity Wave Modeling

Nutrino Studies

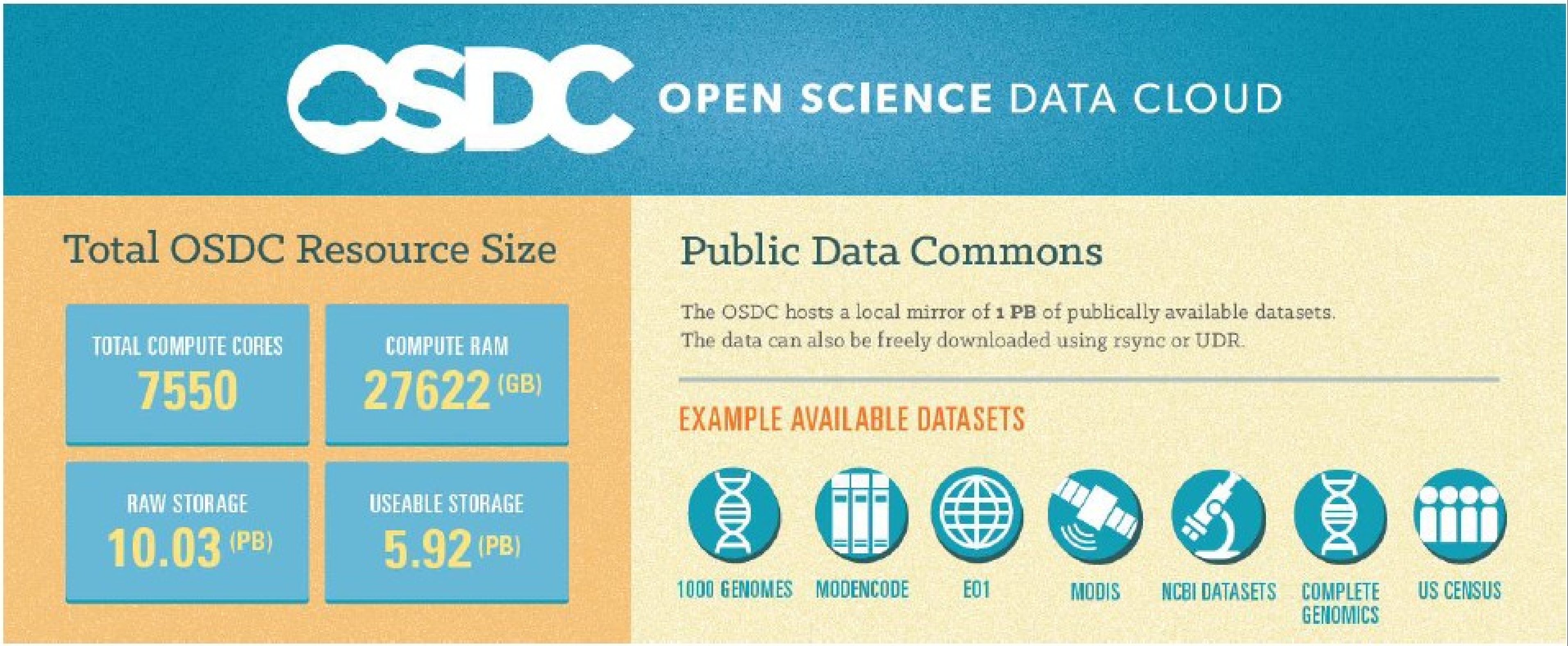


Usage

Open Science Grid

This Distributed Facility Supports Many Sciences

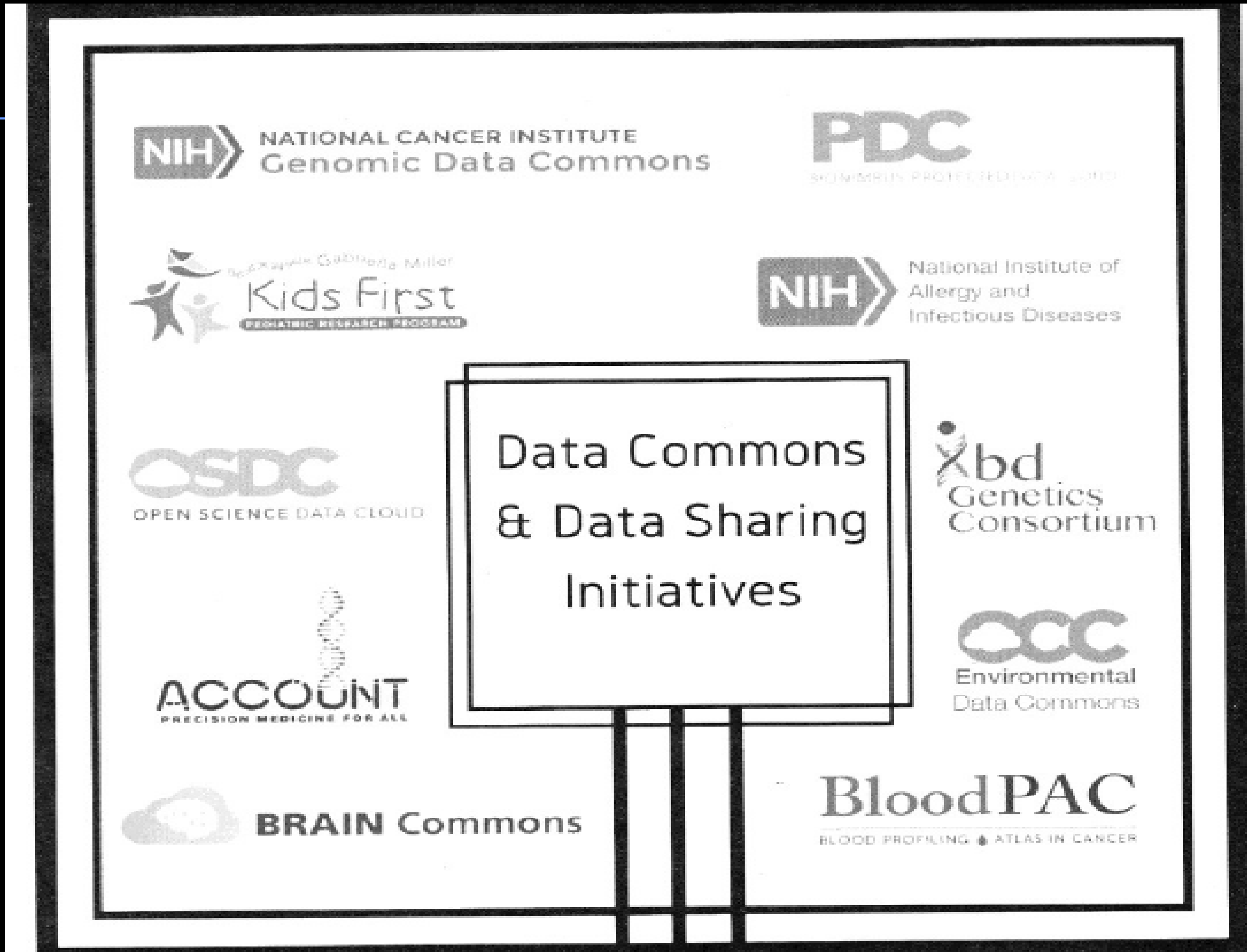
The Open Science Data Cloud (OSDC) is an **open-source**, **cloud-based** infrastructure that allows scientists to manage, share, and analyze medium to large size scientific datasets.

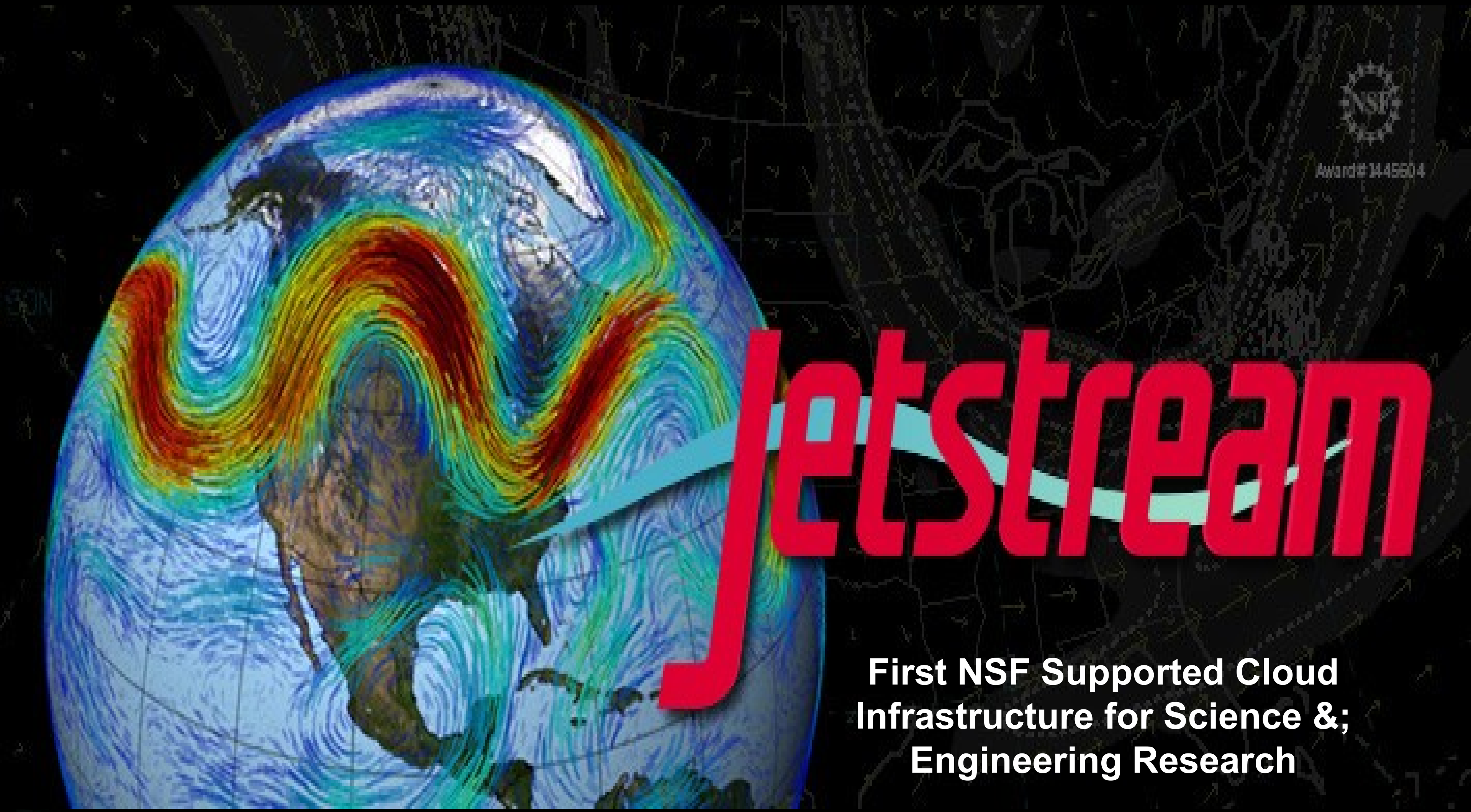


Application for resources available to anyone doing scientific research:

Open Commons Consortium

www.opensciencedatacloud.org


NIH
NATIONAL CANCER INSTITUTE
Genomic Data Commons
PDC
Kids First
PEDIATRIC RESEARCH PROGRAM
NIH
National Institute of Allergy and Infectious Diseases
OSDC
OPEN SCIENCE DATA CLOUD
Data Commons & Data Sharing Initiatives
bd Genetics Consortium
ACCOUNT
PRECISION MEDICINE FOR ALL
Environmental Data Commons
BRAIN Commons
BloodPAC
BLOOD PROFILING ATLAS IN CANCER



First NSF Supported Cloud Infrastructure for Science &; Engineering Research

Summit At Oak Ridge National Laboratory – A Step To A21 Exescale Computer at Argonne National Laboratory (2021)

STARLIGHT℠

HEP = Staggering Amounts of Data

BaBar 0.3 PetaByte/year (2001)

CDF or D0 Run II

0.5 PetaByte/year (2003)

LHC Mock Data Challenge
1 PetaByte/year (~2005)

CMS or ATLAS

2 PetaBytes/year (~2008)

KTeV 50 TeraBytes /year (1999)

SLD 3 TB /year (1998)

Run I (CDF or D0) 20 TB /year (1995)

L3 5 TB /year (1993)

E791 50 TB /year (1991)

EMC 400GB /year (1981)

In 1977 the Upsilon (bottom quark) was discovered at Fermilab by experiment E288 led by now Nobel laureate Leon Lederman

The experiment took about 1 million events and recorded the raw data on ~ 300 magnetic tapes for about 6 GB of raw data

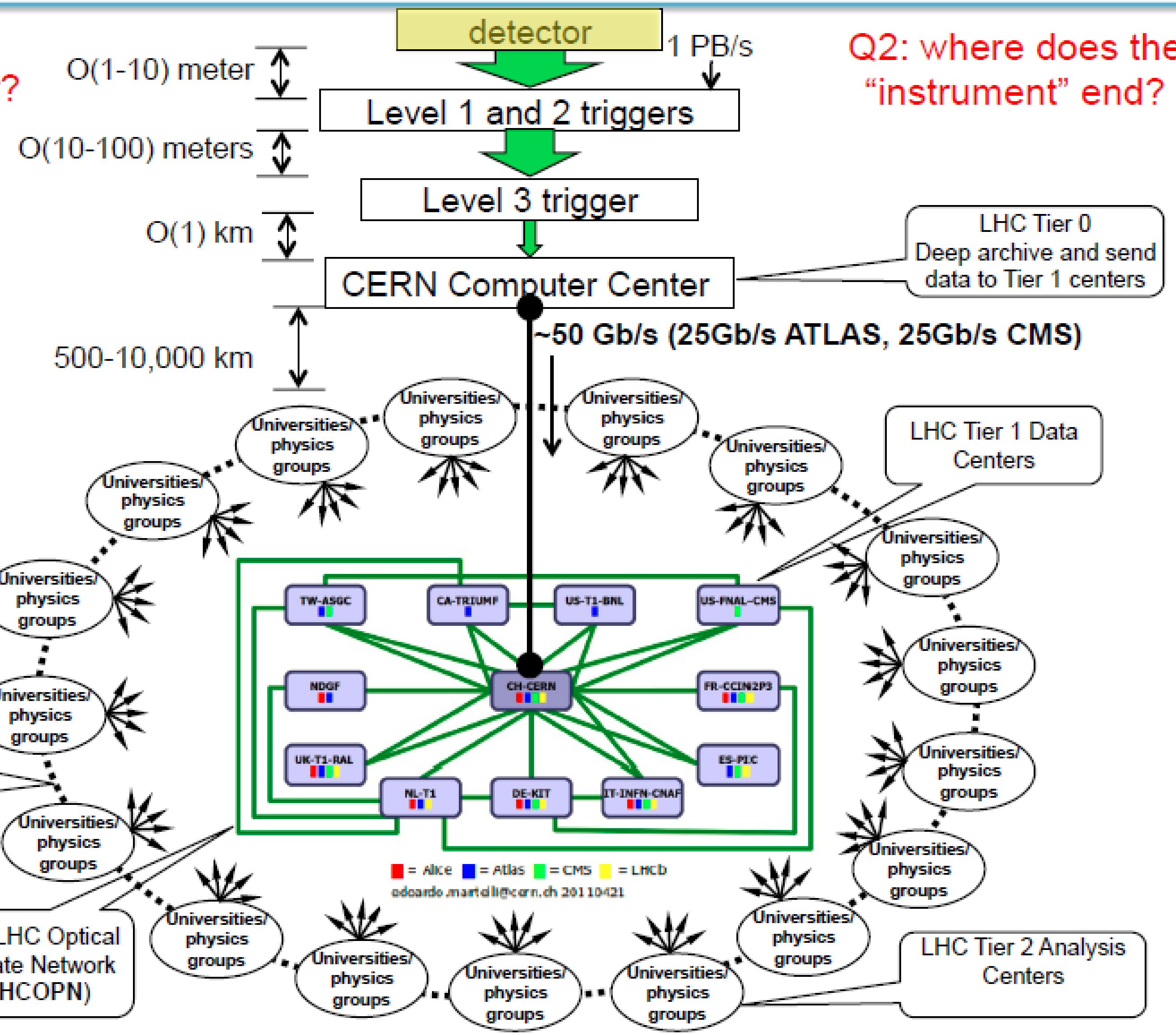
Source: Fermi Lab

Network-Centric View of Large Hadron Collider (@CERN)

Q1: where does "discovery" occur?

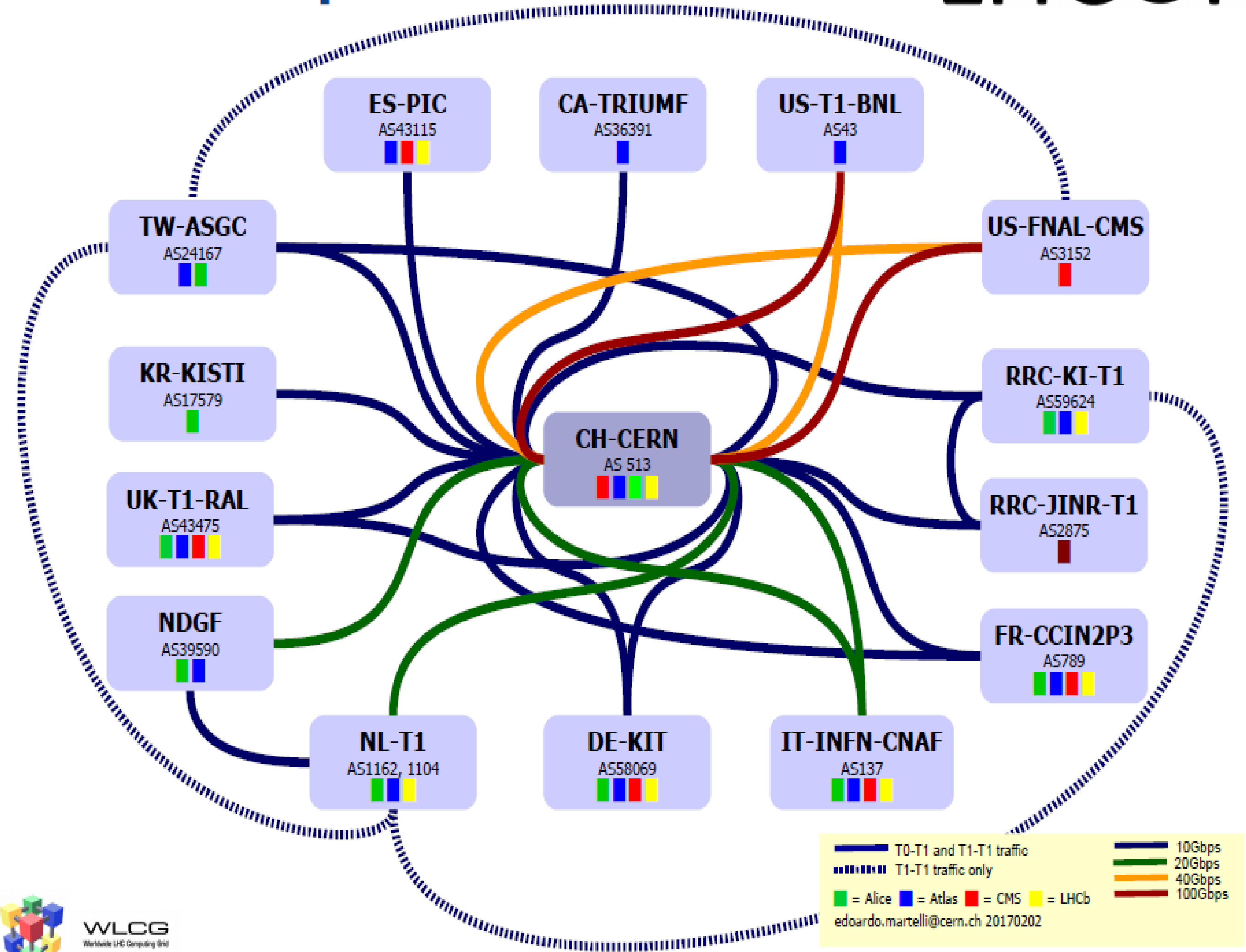
Q2: where does the "instrument" end?

CERN →T1	miles	kms
France	350	565
Italy	570	920
UK	625	1000
Netherlands	625	1000
Germany	700	1185
Spain	850	1400
Nordic	1300	2100
USA – New York	3900	6300
USA - Chicago	4400	7100
Canada – BC	5200	8400
Taiwan	6100	9850



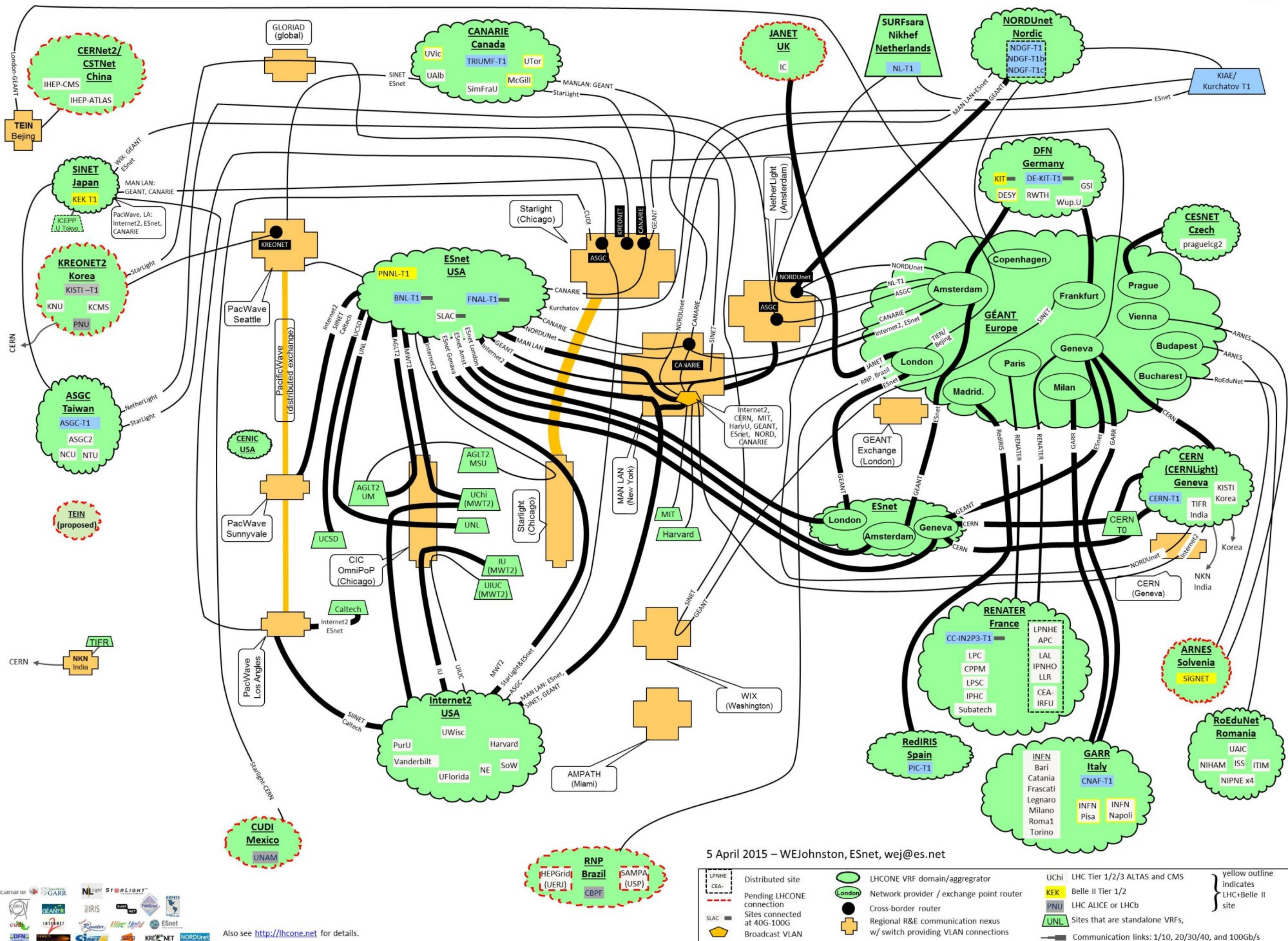
Source: Bill Johnston

LHCOPN map



ES-PIC
AS43115
CA-TRIUMF
AS36391
US-T1-BNL
AS43
TW-ASGC
AS24167
US-FNAL-CMS
AS3152
KR-KISTI
AS17579
RRC-KI-T1
AS59624
CH-CERN
AS 513
UK-T1-RAL
AS43475
RRC-JINR-T1
AS2875
NDGF
AS39590
FR-CCIN2P3
AS789
NL-T1
AS1162, 1104
DE-KIT
AS58069
IT-INFN-CNAF
AS137
T0-T1 and T1-T1 traffic
T1-T1 traffic only
= Alice = Atlas = CMS = LHCb
edoardo.martelli@cern.ch 20170202
10Gbps
20Gbps
40Gbps
100Gbps

LHCONE: A global infrastructure for the High Energy Physics (LHC and Belle II) data management

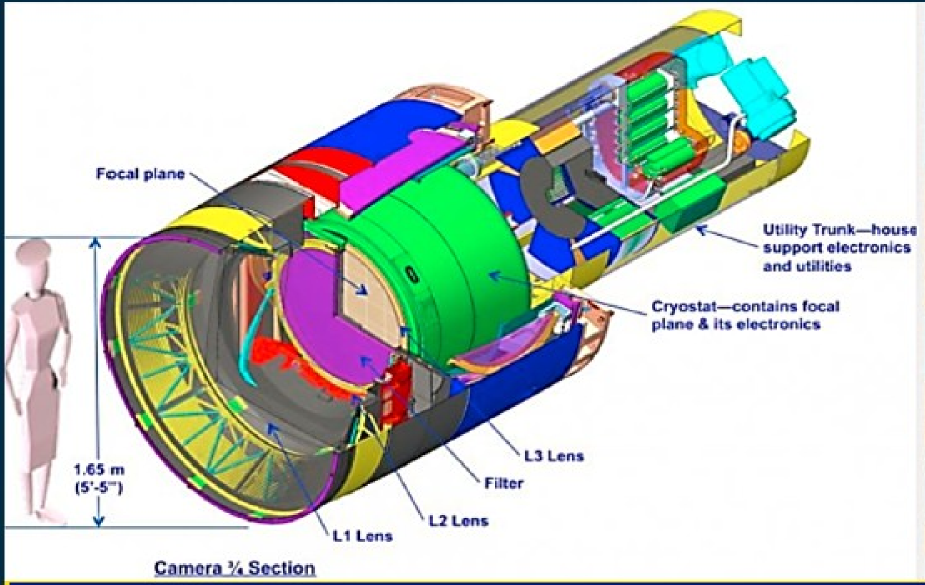


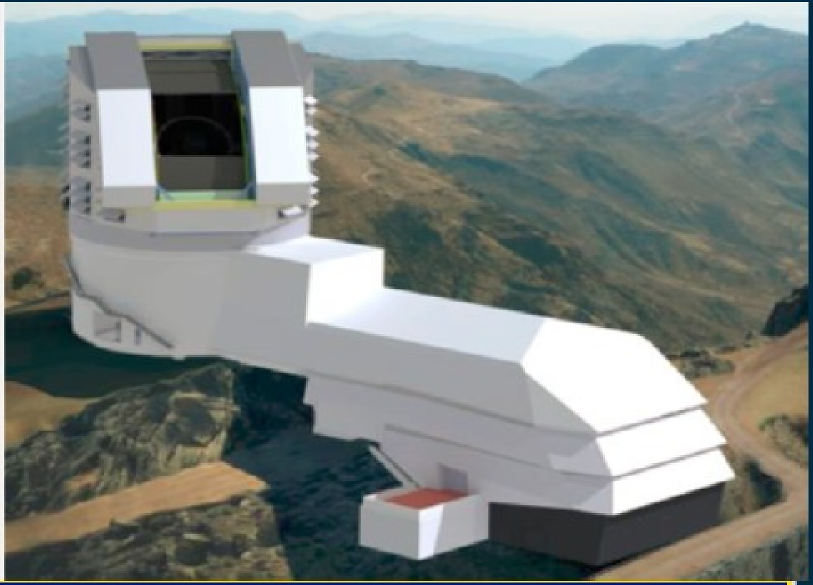
Also see http://lhcone.net for details.

Argonne National Laboratory Advanced Photon Source

Square Kilometer Array

LSST Data Movement
Upcoming challenges for Astronomy





- 3.2 Gigapixel Camera with calibrated exposures at (10 Bytes / pixel)
- Planned Networks: Dedicated 100G for image data, Second 100G for other traffic, and 40G for diverse path
- Lossless compressed Image size = 2.7GB (~5 images transferred in parallel over a 100 Gbps link)
- UDP based custom image transfer protocols

Current Storage Landscape

- Storage Is Isolated, Difficult To Access
  - Multiple Isolated Facility, Campus, instrument Systems
  - Incompatibilities, Inefficiencies
  - Especially Problematic for Petabytes – Exebytes Are Also Required
- Commercial Clouds Not A Solution
  - Cost Model Makes Collaborative Distribution Prohibitive
  - Operations Prevent Detailed Performance Analytics
  - Limited Data Tools

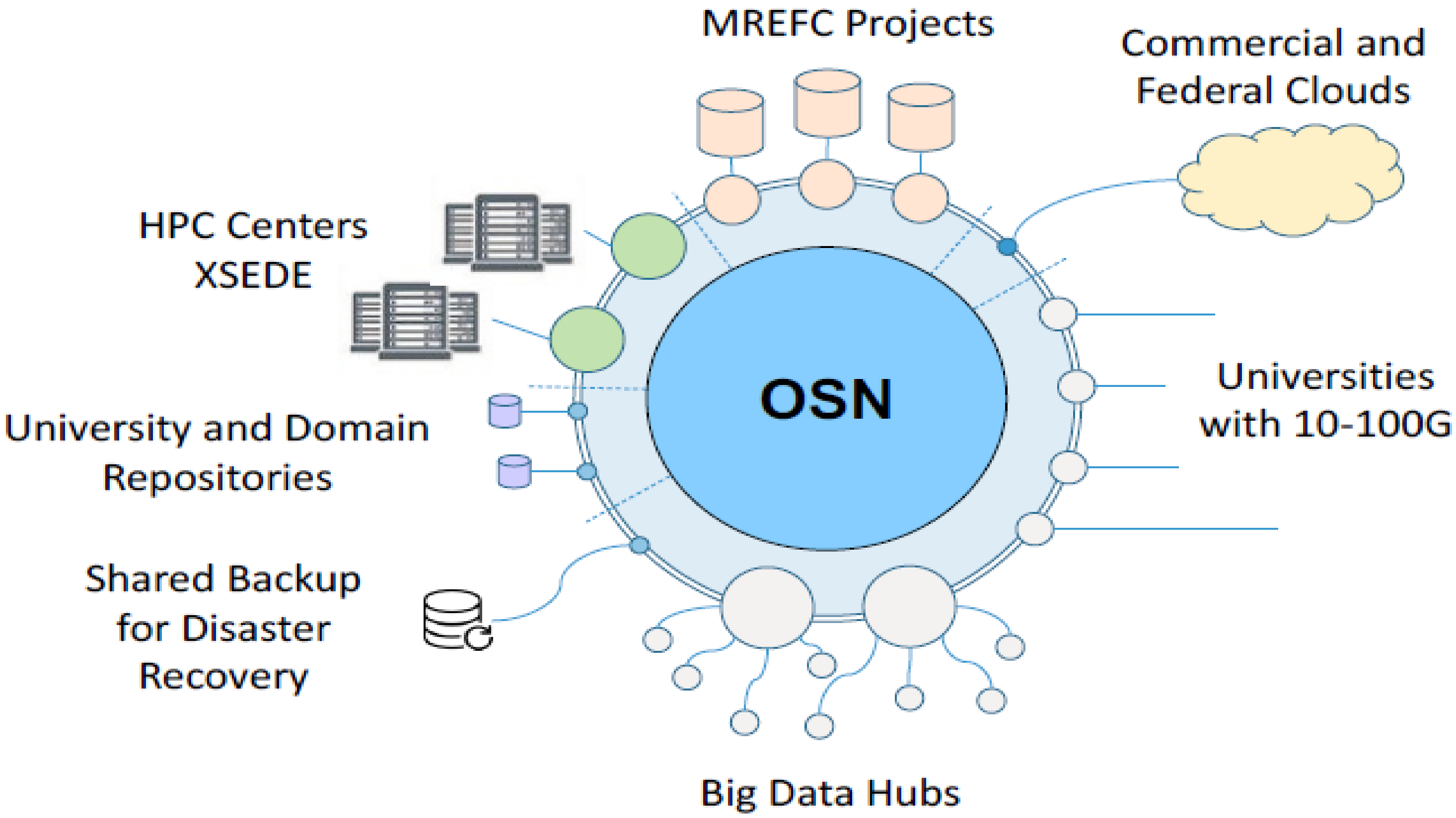
Opportunity: Next Gen Storage For Science

- **Large National Distributed Storage System:**
- ***– Perhaps 1-2PB Storage Rack On Each CC* Campus (~200PB)***
- ***– Create Redundant Interconnected Storage Substrate***
- ***Using Industrial Strength Erasure Code Storage***
- **– Provide *High Capcity Aggregate Bandwidth, Easy Data Flow Among Sites (100 Gbps Channels)***
- ***– Potential For Acting As Gateways To Cloud Providers***
- ***– Automatic Compatibility, Simple Standard API (S3)***
- ***– Implement a Set of Simple Policies***
- ***– Enable Sites To Add Additional Storage Locally Funded***
- ***– Provide Variety of Services Built By Communities***

Transformational Impact

- **Potentiual To Totally Change Landscape For Academic Big Data Research (Even At Petabyte Scale)**
- ***– Create Homogeneous, Uniform Storage Tier For Science***
- ***– Liberate Communities To Focus On Science, Discovery, Analytics Collaboration, and Preservation***
- ***– Amplify NSF CC* Investment***
- ***– Very Rapidly Spread Best Practices Nationwide***
- ***– Links to Large Science Instruments, Compute Facilities, Data Facilities (Including Big Data Hubs), Analytics Sites, et al***
- ***– Big Data Projects Can Use It For Data Distribution***
- ***– Small Projects Can Build On Existing Infrastructure***
- ***– Enablibg Whole Ecosystem of Specialized Services to Flourish***
- ***<u>Major Opportunities for Novel Interdisciplinary Research</u>***

Connections



Many Issues

- Architecture
- Technologies
- Services
- Mechanisms For Integration With Instruments, Compute Facilities, Data Hubs, Analytic Centers, etc
- Policies
- Communications/Education
- Security
- Financial Models
- Mechanisms For Innovation and On-Going Extensions and Enhancements
- Governance/Management of Data and Facilities
- Long vs Short Term Repositories
- Etc

Building Blocks

- Scalable element (SE)
  - *500TB of storage+ single server*
  - *Support 40G interface for sequential read/write*
  - *Should saturate 40G for read, about half for write*
- Stack of multiple SEs
  - *Aggregated to 100G on a fast TOR switch, now becoming quite inexpensive (<$20K)*
- These can also exist inside the university firewall
  - *But purchased on local funds, storing local data*
- Software stack to be discussed
  - *ZFS, Ceph, Mero,…*
  - *Integrated with Globus "Lite", with streamlined stack*

Building Blocks 2

- E2E Services (e.g., BigData Express, SENSE)
- APIs
- Workflow Managers (e.g., Jupyter)
- Data Transfer Nodes
- Performance Monitoring/Measurement/Analytics Instrumentation
- Ultra High Performance File Systems
- Piplines Based On Direct Connections Between HP File Systems and High Performance Optical Channels (e.g., Lightpaths)
- Interdomain Services

Management

- Who owns it?
  - *OSN storage should remain in a common namespace*
  - *This would enable uniform policies and interfaces*
- Software management
  - *Central management of software stack (push)*
  - *Central monitoring of system state*
- Hardware management
  - *Local management of disk health*
  - *Universities should provide management personnel*
- Policy management
  - *This is **hard** and requires a lot more discussion*
- Monitoring
  - *Two tier, store all events and logs locally, send only alerts up*
  - *Try to predict disk failures, preventive maintainance*
- Establish metrics for success

Initial OSN Faciltity Sites

- Johns Hopkins University, Baltimore Maryland
- StarLight International/National Communications Exchange Facility, Chicago, Illinois
- University of California At San Diego Supercomputing Center, La Jolla, California
- Future Sites Under Discussion
- International Extensions Possible

StarLight – "By Researchers For Researchers"

StarLight is an experimental optical infrastructure and **proving ground for network services** optimized for high-performance applications

Multiple 10GE+100 Gbps StarWave

Multiple 10GEs Over Optics – World's "Largest" 10G/100G Exchange

First of a Kind Enabling Interoperability At L1, L2, L3

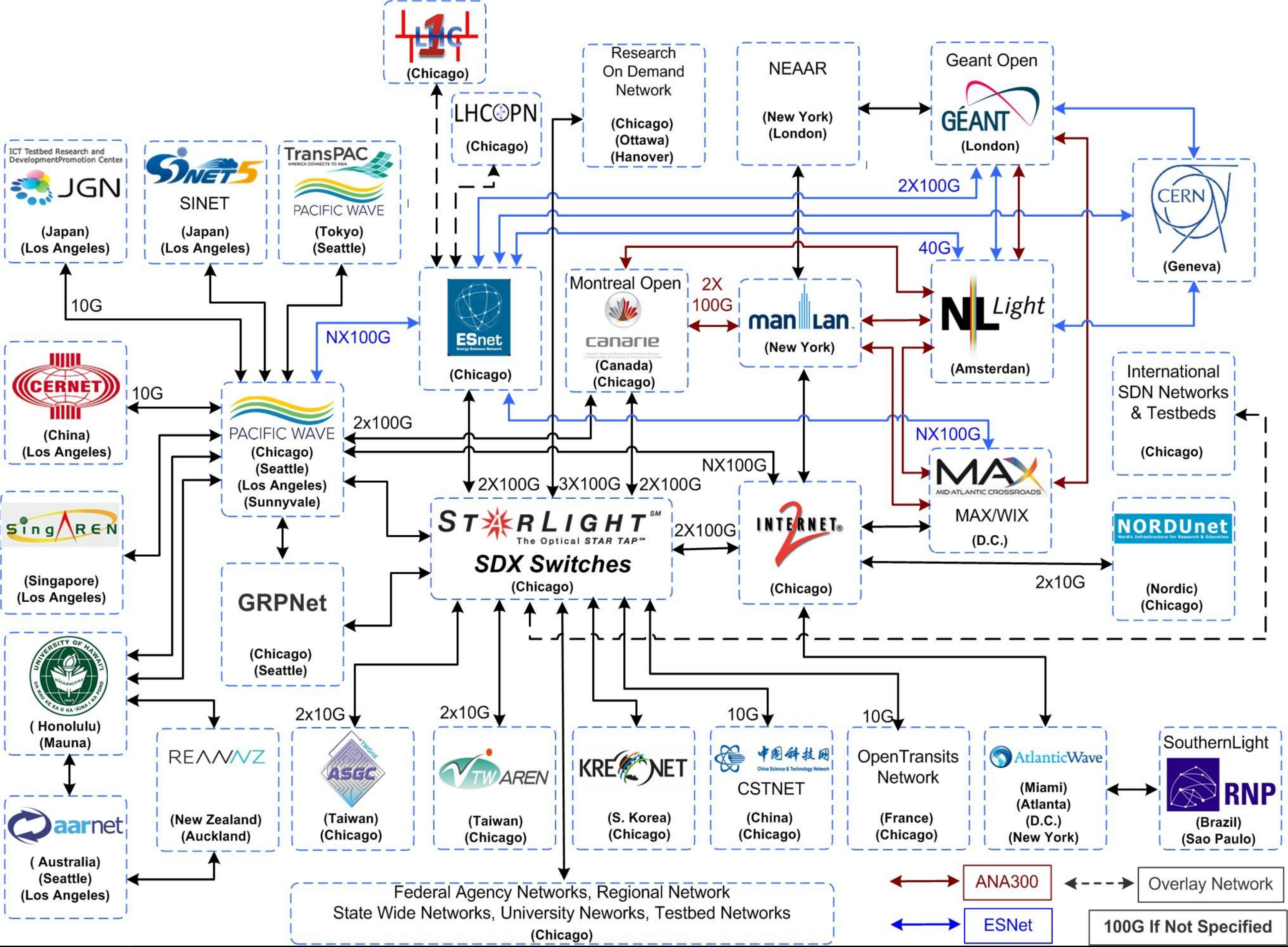
View from StarLight

Abbott Hall, Northwestern University's Chicago Campus

IRNC: RXP: StarLight SDX A Software Defined Networking Exchange for Global Science Research and Education

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)
PI IRNC: RXP: StarLight SDX
Co-PI Tom DeFanti, Research Scientist, (tdefanti@soe.ucsd.edu)
California Institute for Telecommunications and Information Technology (Calit2),
University of California, San Diego
Co-Director, StarLight
Co-PI Maxine Brown, Director, (maxine@uic.edu)
Electronic Visualization Laboratory, University of Illinois at Chicago
Co-Director, StarLight
Jim Chen, Associate Director, International Center for Advanced Internet Research, Northwestern University

National Science Foundation
International Research Network Connections Program

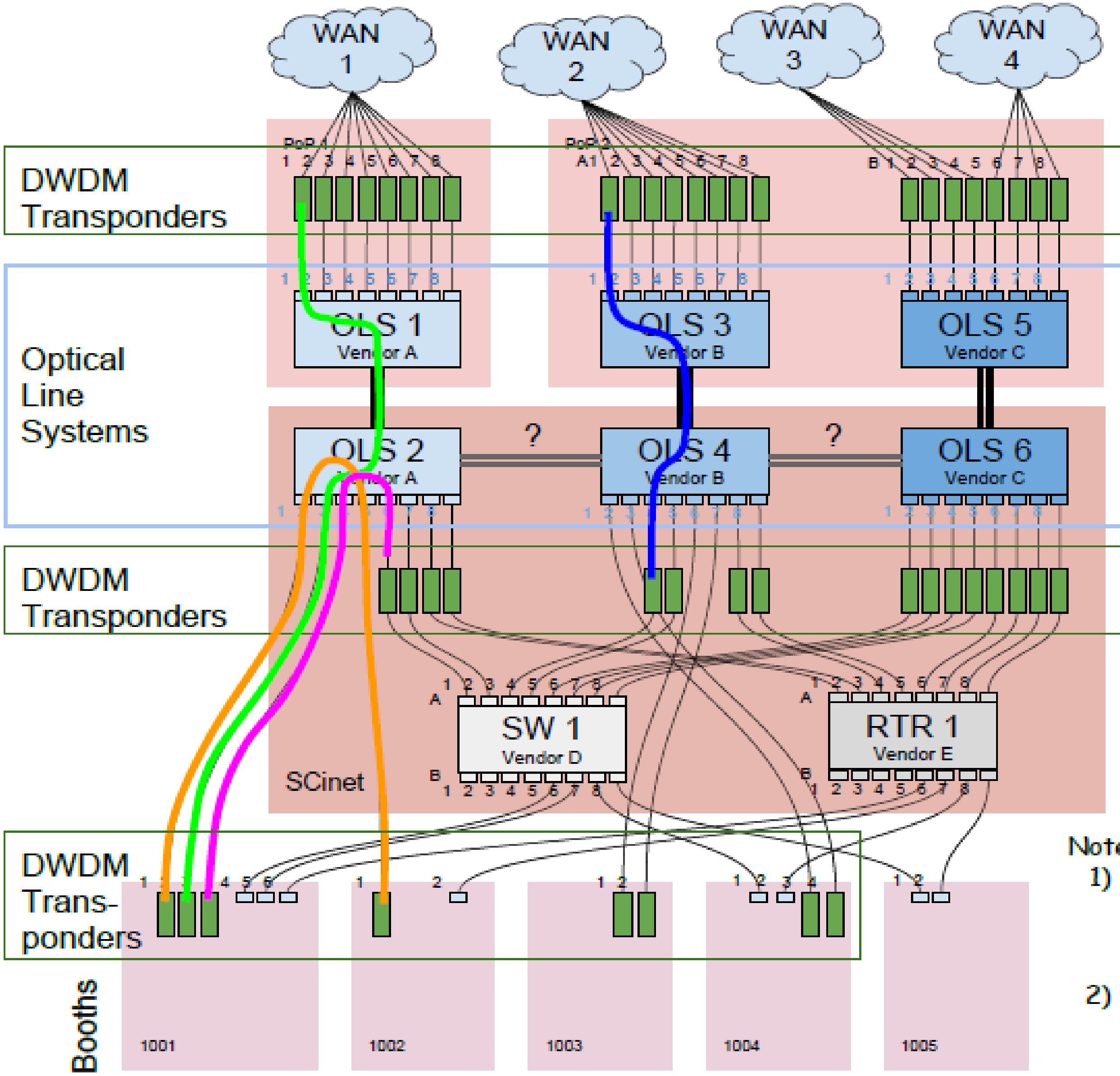


(Chicago)
LHCOPN
(Chicago)
Research
On Demand
Network
(Chicago)
(Ottawa)
(Hanover)
NEAAR
(New York)
(London)
Geant Open
GÉANT
(London)
ICT Testbed Research and DevelopmentPromotion Center
JGN
(Japan)
(Los Angeles)
SINET
(Japan)
(Los Angeles)
TransPAC
AMERICA CONNECTS TO ASIA
PACIFIC WAVE
(Tokyo)
(Seattle)
2X100G
CERN
(Geneva)
40G
10G
ESnet
(Chicago)
Montreal Open
canarie
(Canada)
(Chicago)
2X
100G
man LAN
(New York)
NL Light
(Amsterdan)
NX100G
CERNET
(China)
(Los Angeles)
10G
PACIFIC WAVE
(Chicago)
(Seattle)
(Los Angeles)
(Sunnyvale)
2x100G
International
SDN Networks
& Testbeds
(Chicago)
NX100G
MAX
MID-ATLANTIC CROSSROADS
MAX/WIX
(D.C.)
NX100G
2X100G
3X100G
2X100G
SingAREN
(Singapore)
(Los Angeles)
STARLIGHT
The Optical STAR TAP
SDX Switches
(Chicago)
2X100G
INTERNET2
(Chicago)
NORDUnet
Nordic Infrastructure for Research & Education
(Nordic)
(Chicago)
2x10G
GRPNet
(Chicago)
(Seattle)
(Honolulu)
(Mauna)
2x10G
2x10G
10G
10G
REANNZ
(New Zealand)
(Auckland)
ASGC
(Taiwan)
(Chicago)
TWAREN
(Taiwan)
(Chicago)
KREONET
(S. Korea)
(Chicago)
CSTNET
(China)
(Chicago)
OpenTransits
Network
(France)
(Chicago)
AtlanticWave
(Miami)
(Atlanta)
(D.C.)
(New York)
SouthernLight
RNP
(Brazil)
(Sao Paulo)
aarnet
(Australia)
(Seattle)
(Los Angeles)
Federal Agency Networks, Regional Network
State Wide Networks, University Neworks, Testbed Networks
(Chicago)
ANA300
Overlay Network
ESNet
100G If Not Specified

Global Research Platform: Building On CENIC/Pacific Wave, GLIF and GLIF GOLEs (e.g., StarLight et al)

Current International GRP Partners

A Disaggregated SCinet Optical Layer



Reconfiguration options

A. Booth to booth connections
B. Booth to WAN connections
C. Booth to switch or router connections
D. WAN to switch or router connections

Examples

A. B-B
   a. Booth 1001-1 to 1002-1 via optical layer
   b. Booth 1001-1 to 1004-3 via optical layer (assumes OLS2 to OLS4 path)
B. Booth to WAN
   a. Booth 1001-2 to PoP1-1 via OLS2-2 and OLS1-1
   b. Booth 1001-2 to PoP2-B1 via OLS2-2, OLS4, OLS6 and OLS5-1
C. Booth to switch/router
   a. Booth 1001-3 to SW1-A1
   b. Booth 1003-1 to RTR1-A5 (assumes OLS4 to OLS6 path)
D. WAN to switch/router
   a. PoP2-A1 (WAN2) to SW1-3 via OLS3-1 and OLS4-3
   b. PoP2-A2 (WAN2) to RTR1-3 via OLS3-2 and OLS4-7

Notes

1) Transponders could be from multiple vendors but for near term the links would need to be built with matching transponders.
2) **Controllers and orchestration systems are not shown** but all Tpndr/OLS systems must be connected

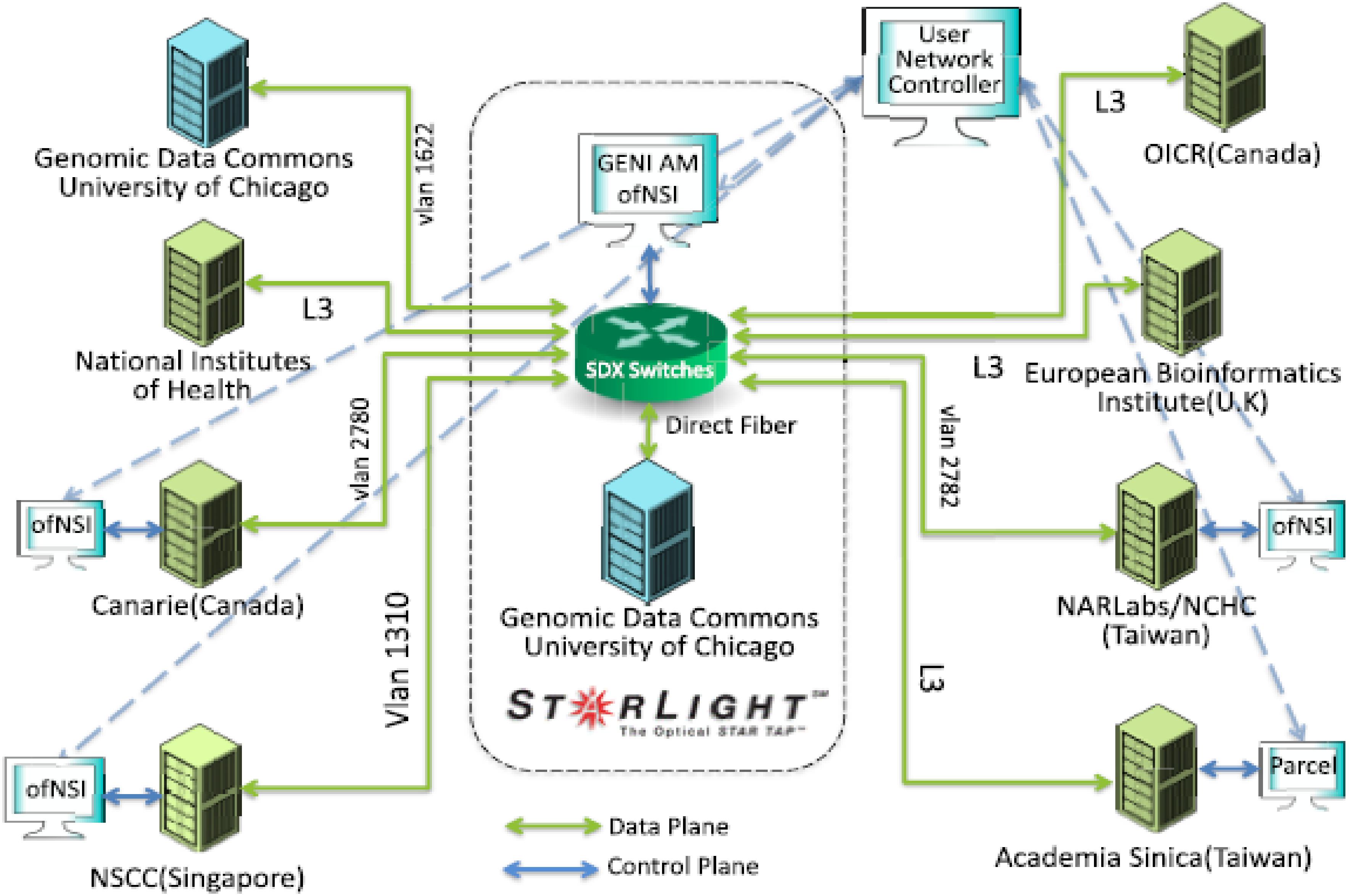
Designing and Deploying A Bioinformatics Software-Defined Network Exchange (SDX): Architecture, Services, Capabilities, and Foundation Technologies

Joe Mambretti, Jim Chen, Fei Yeh
International Center for Advanced Internet Research
Northwestern University
Robert Grossman, Piers Nash, Alison Heath, Renuka Arya, Stuti Agrawal, Zhenyu Zhang
Center for Data Intensive Science
University of Chicago
Chicago, Illinois, USA

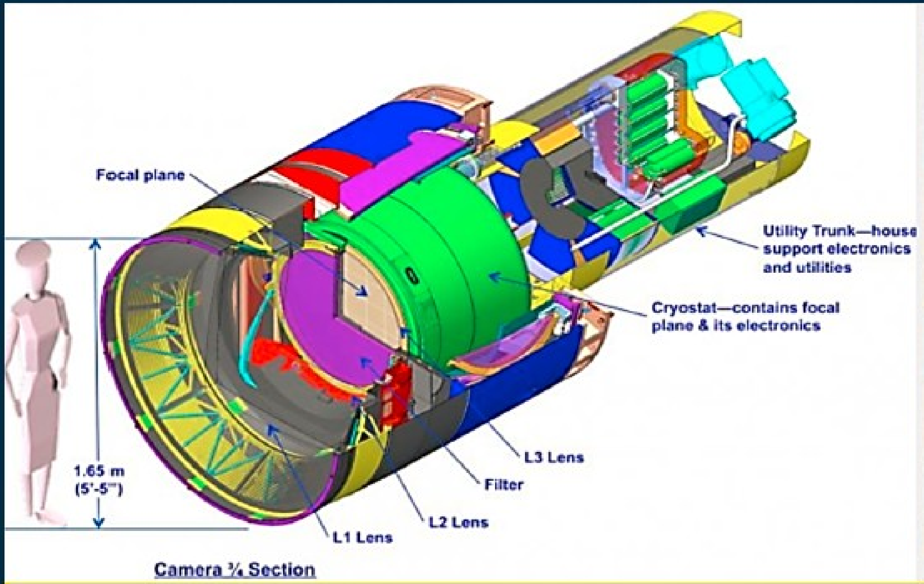
iCAIR

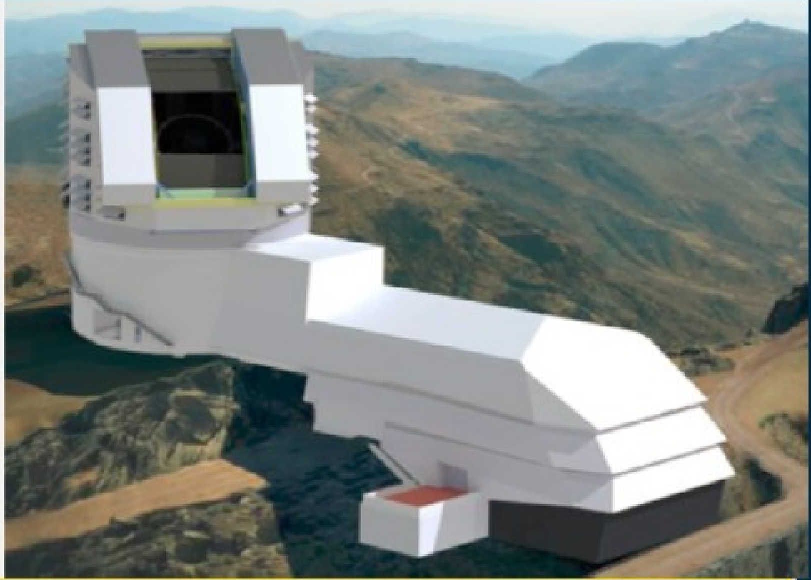
STARLIGHT

2016 Bioinformatics SDXs Network



LSST Data Movement
Upcoming challenges for Astronomy

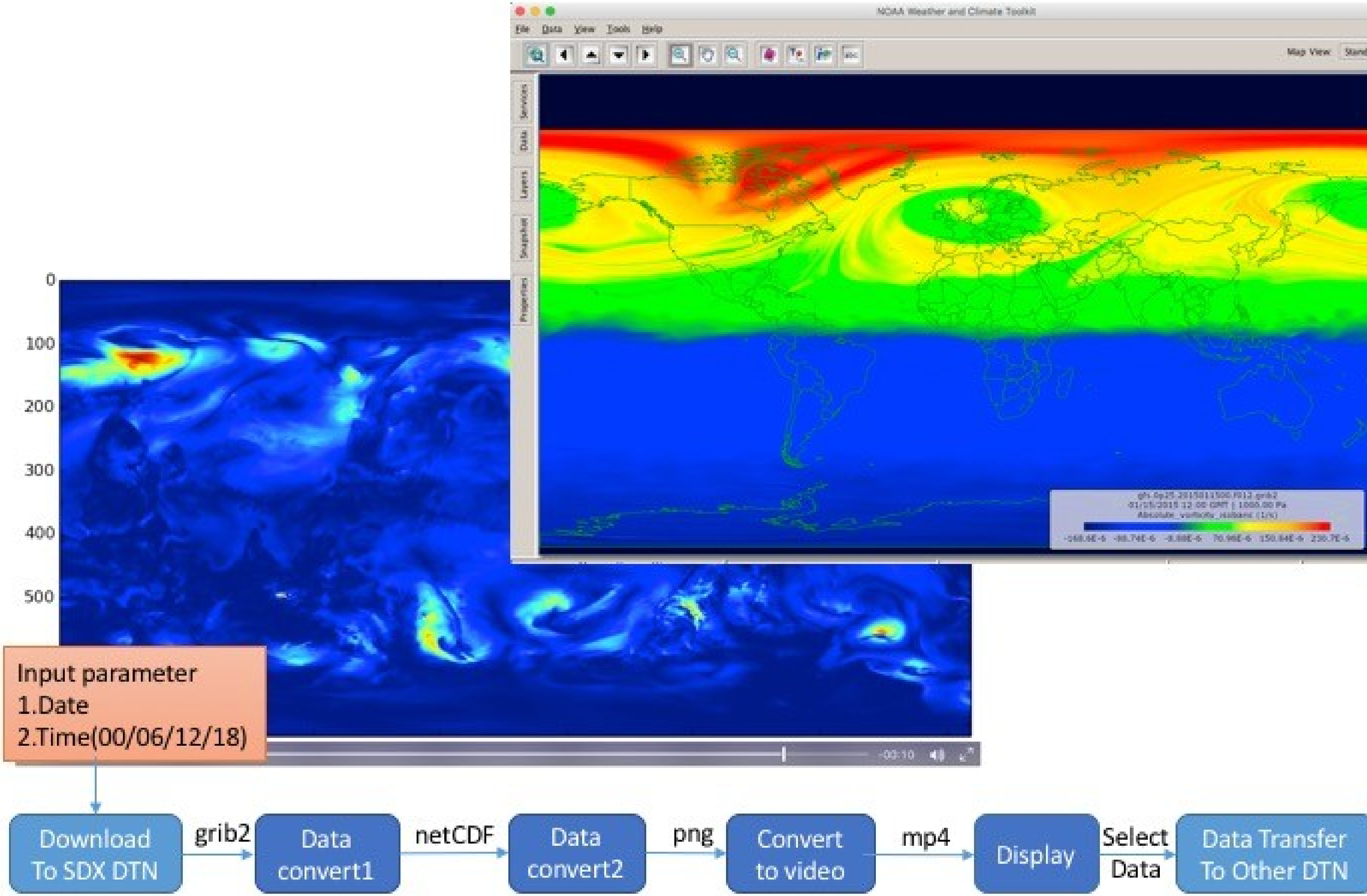




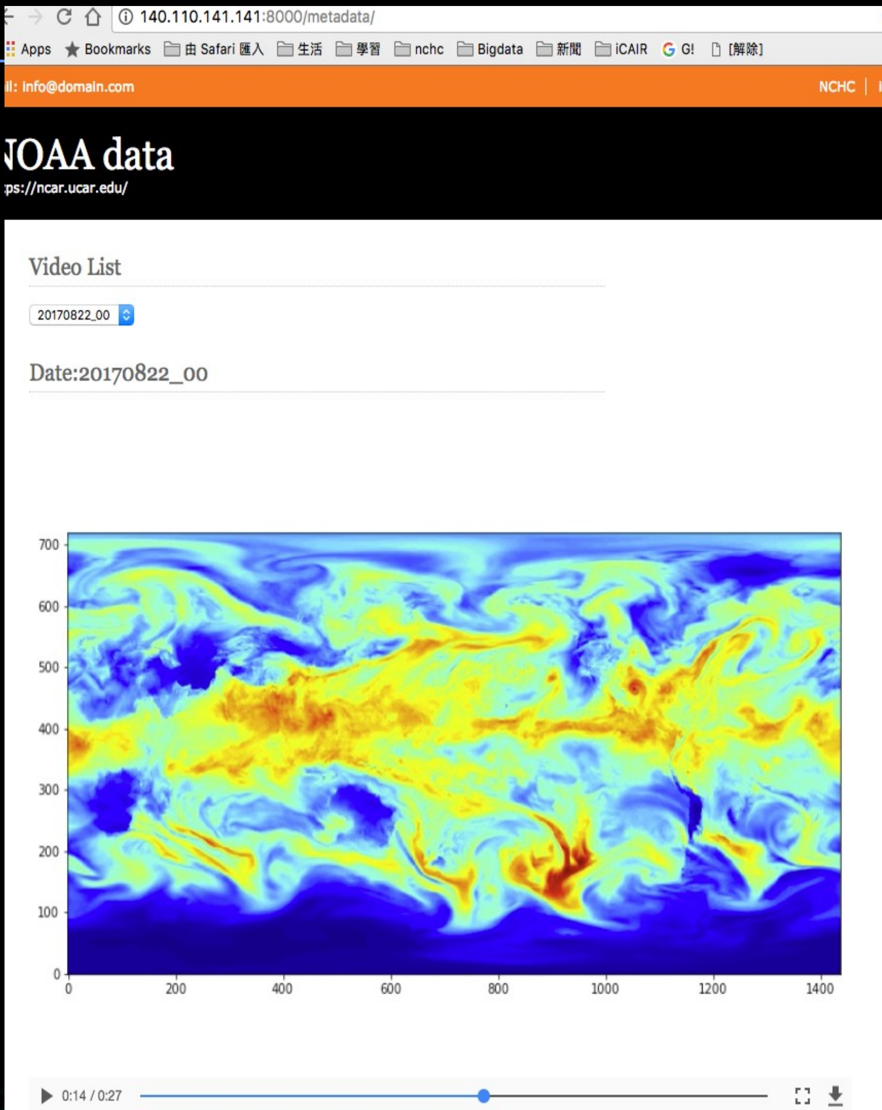
- 3.2 Gigapixel Camera with calibrated exposures at (10 Bytes / pixel)
- Planned Networks: Dedicated 100G for image data, Second 100G for other traffic, and 40G for diverse path
- Lossless compressed Image size = 2.7GB (~5 images transferred in parallel over a 100 Gbps link)
- UDP based custom image transfer protocols

STARLIGHT℠

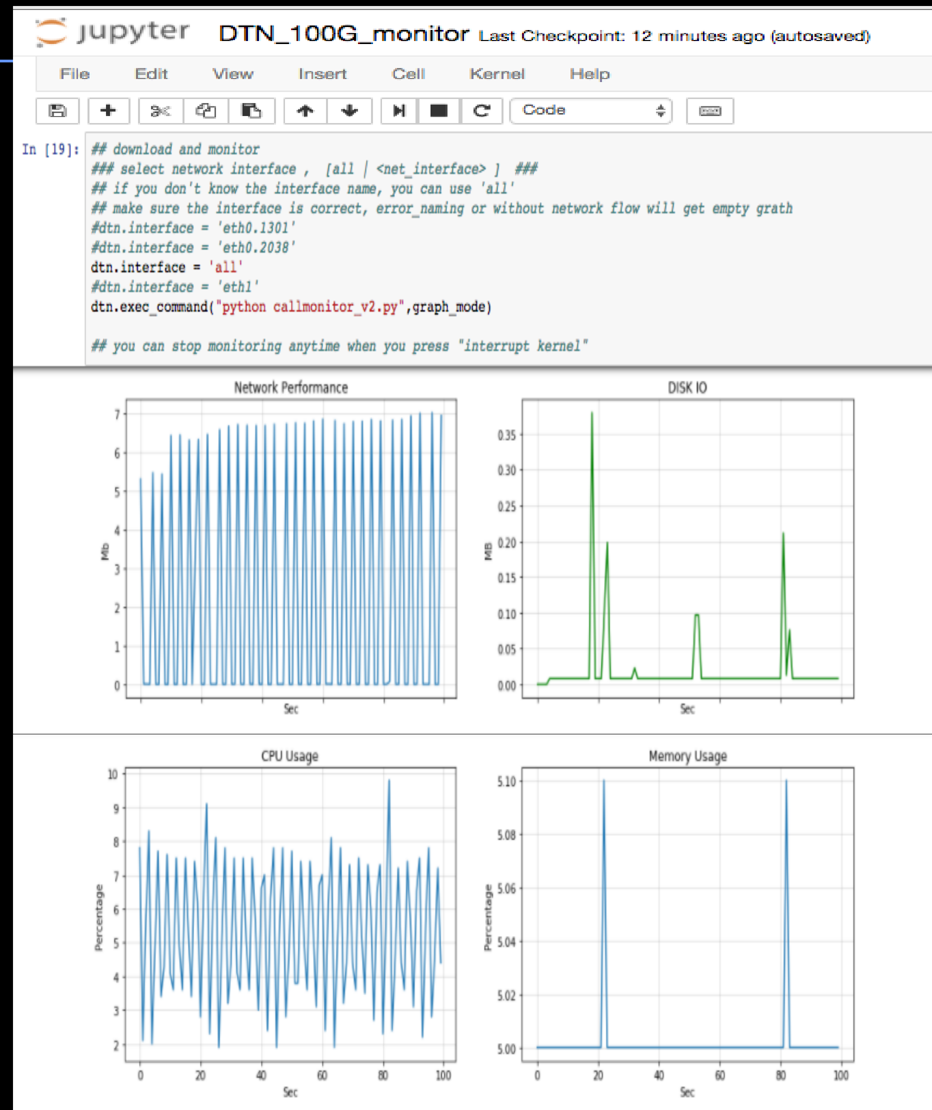
StarLight SDX Geoscience Research Workflow



GeoScience SDX DTN Service Prototype

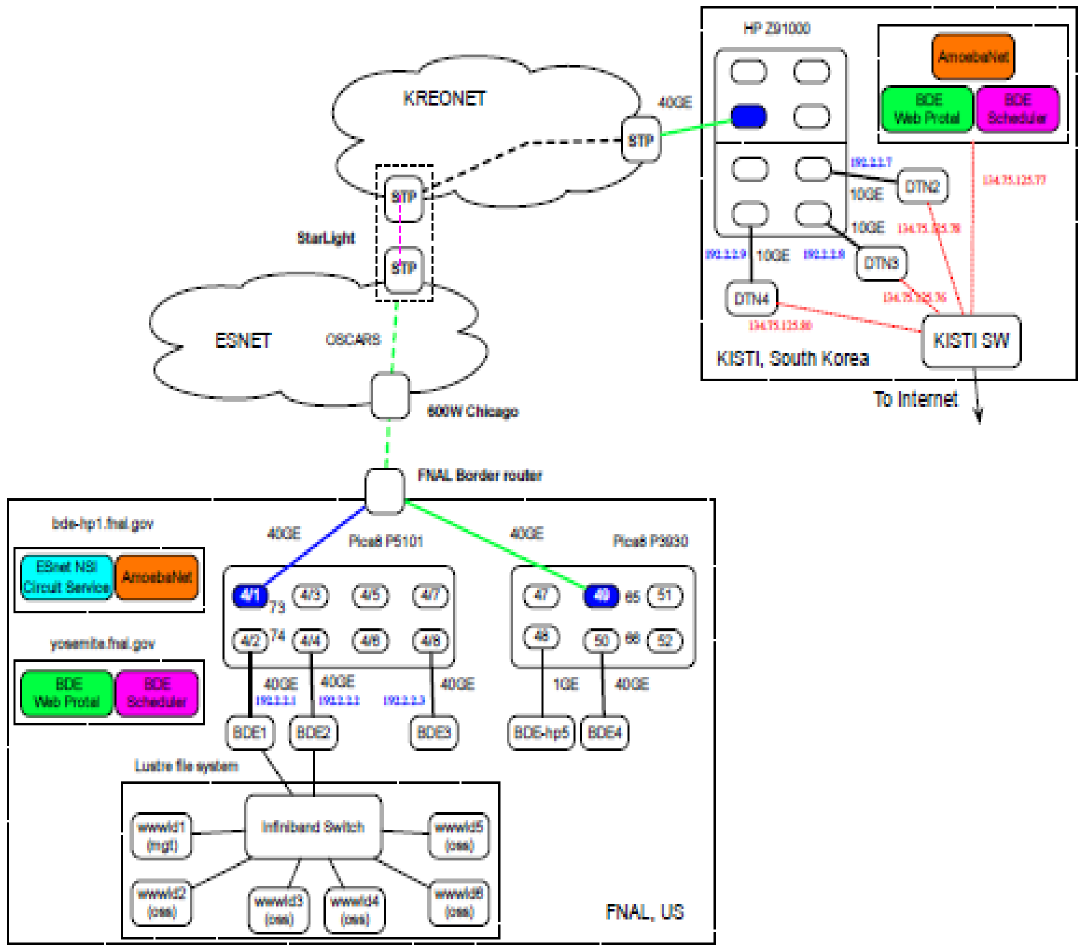


File Screening workflow



File Transfer Workflow

A Cross-Pacific SDN Testbed



www.startap.net/starlight

**Thanks to the NSF, DOE, DARPA,
NIH, USGS, NASA,
Universities, National Labs,
International Partners,
and Other Supporters**