GLIF 2017 Demonstrations: International Multi-Domain SD-WAN Services

Joe Mambretti Presenting for All Collaborators: Will Black, Pieter de Boer, Jim Chen, Wei-Yu Chen, Buseung Cho, Leon Gommans, John Hess, Joseph Hill, Marc Lyonnais, Gerben van Malenstein, John Macauley, Warrick Mitchell, Chris Myers, Dave Reese, Thomas Tam, J.P.Velders, Migiel de Vos, Kevin Wang, David Whittaker, David Wilde, Rod Wilson, Fei Yeh, Se-Young Yu

LHCOPN LHCONE Networking Meeting
Co-Hosted with HEPiX
KEK
Tsukuba , Japan

SINET 2*10G
SINET 10G
KREONET/KREONet2-KRLight 110G
ASGCNet 10G
TaiwanLight 10G
KREONET/KREONet2-KRLight 12.5G
TransPAC-PacificWave 100G
CSTNET 10G
PacificWave 100G
Pacific Wave-WRN 100G
Internet2 100G
AARNet-SXTransPORT 100G
Internet2-CERNET 10G
TransPAC-PacificWave 100G
CANARIE 100G
CANARIE 100G
Internet2 100G
ESnet 100G/10G
SINET 100G
SINET 100G
TaiwanLight 10G
SingAREN-Internet2 100G
PIREN 100G
UHnet 10G
AARNet-SXTransPORT 100G
PIREN 100G
SingAREN-Internet2 100G
AARNet 2.5G
SINET 10G
TaiwanLight 10G
CSTNET 2*2.5G
SingAREN-TEIN 10G
SINET 2*10G
ASGCNet 10G
CENI 20G
KREONET/KREONet2-KRLight 30G
ANA-300G 100G
ESnet 40G
ESnet 2*100G
NEAAR 100G
ASGCNet 10G
ANA-300G 100G
ANA-300G 100G
ESnet 100G
Informica-RUNNet 100G
ReTN Optical Rail 10G
Informica-RUNNet 2*10G
ReTN Optical Rail 10G
SANReN-GNA Pathfinder 10G
SANReN 50G
CERN-TIFR 10G
AmLight Protect 100G
AmLight Express n*100G
RedCLARA 10G
RedCLARA 2.5G
RedCLARA 2.5G
AmLight Protect 100G
RedCEDIA 10G
RedCLARA 10G
RedCLARA-BELLA-T 10*10G
AmLight Express n*100G
RedCLARA 10G
AmLight Protect 100G
RedCLARA 10G
AmLight Protect 100G
REUNA 1G
REUNA-ESO 10G
EVALSO 10G
RedCLARA-RNP 10G
AmLight Protect 100G
RedCLARA 10G
Innova-Red 10G
SANReN 50G
SANReN-GNA Pathfinder 10G
SANReN 60G
AARNet 2.5G
AARNet n*100G
AARNet n*100G
AARNet 2.5G
AARNet-SXTransPORT 100G
REANNZ 10G
REANNZ 20G
SingAREN-TEIN 10G
NKN 10G

Six Related Demonstrations of Advanced Capabilities

- Research Motivation:

- Software Defined Services To Enhance International Collaborations Across Federated WAN Domains

- A Prototype Of Future Services and Capabilities

Research: Key Issues

- Today Almost All Networks Provide Only “One-Size-Fits-All” Undifferentiated L3 Services
- These Services Are Suboptimal For Many Applications and Services, Especially Those Based On Emerging Capabilities
- Future Networks Will Provide For Multiple Types Of Services Differentiation, e.g. via Slicing To Address These Issues
- These Six Related Demonstration Showcase How Services Based On SD WAN Capabilities Can Make Tomorrow’s Networks Available Today

Challenges & Opportunities

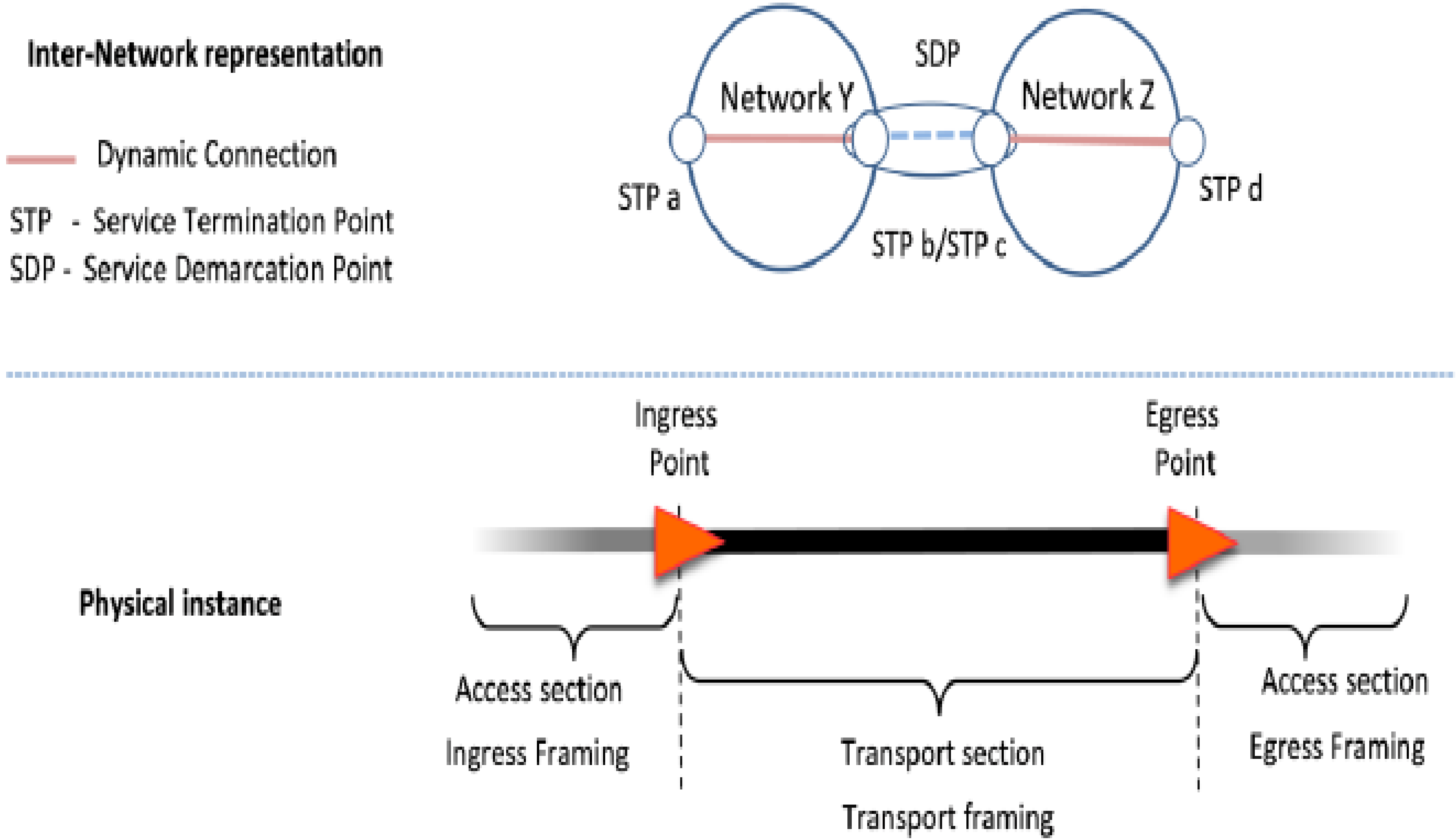
- Challenges: On Today's Networks, Even R&E Networks, It Is Difficult To Transport Extremely Large Files and Collections of Many Files Over WANs, Especially Over Multi-Domains
- Solution: Using SD-WAN Capabilities (Programmable Network Slicing To Segment Network Resources Allows Different Services To Co-Exist Without Interference

Global LambdaGrid Workshop Demonstrations:

- International Multi-Domain Provisioning Using AutoGOLE Based Network Service Interface (NSI 2.0)
- Using RNP MEICAN Tools for NSI Provisioning
- Large Scale Airline Data Transport Over SD-WANs Using NSI and DTNs
- Large Scale Science Data Transport Over SD-WANs Using NSI and DTNs
- SDX Interdomain Interoperability At L3
- Transferring Large Files E2E Across WANs Enabled By SD-WANs and SDXs

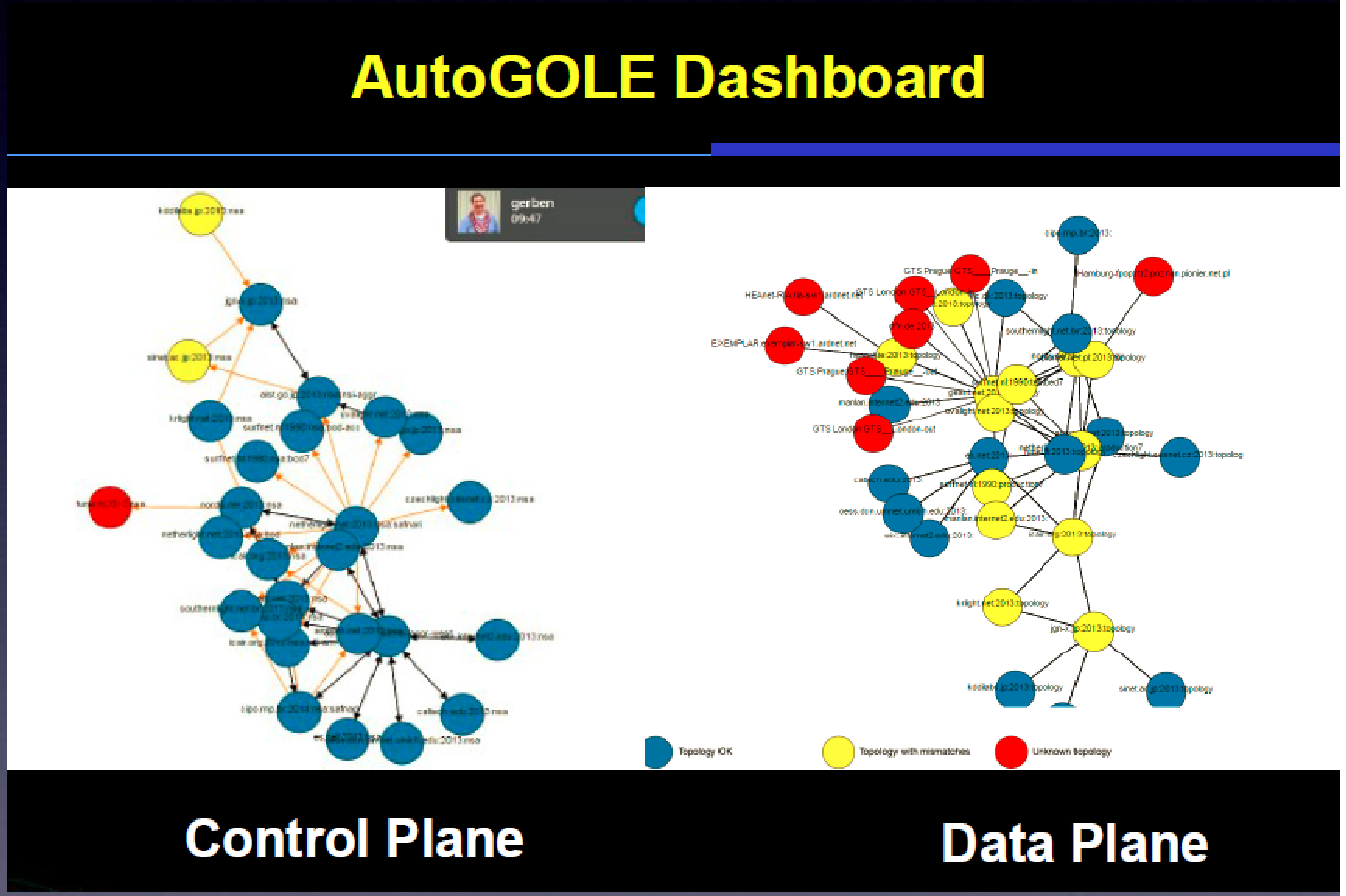
Demonstration 1: International Multi-Domain Provisioning Using AutoGOLE Based Network Service Interface (NSI 2.0)

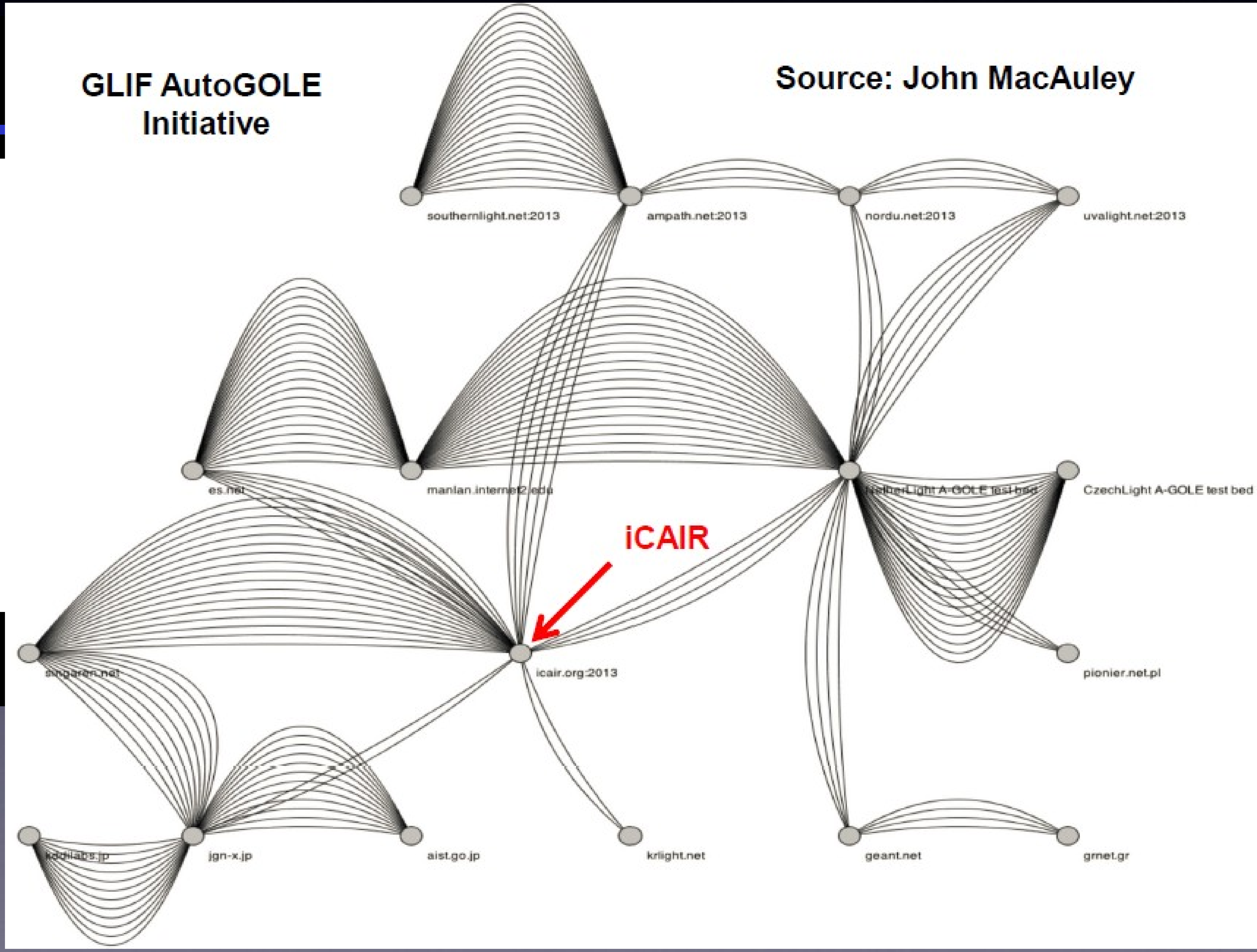
- Network Service Interface (NSI 2.0)
- An Architectural Standard Developed By the Open Grid Forum (OGF)
- OGF Pioneered Programmable Networking (Initially Termed “Grid Networking”)
- Techniques That Made Networks ‘First Class Citizens” in Grid Environments – Programmable With Grid Middleware
- Currently Being Placed Into Production By R&E Networks Around the World


Inter-Network representation
Dynamic Connection
STP - Service Termination Point
SDP - Service Demarcation Point
Network Y
SDP
Network Z
STP a
STP d
STP b/STP c
Ingress Point
Egress Point
Physical instance
Access section
Ingress Framing
Transport section
Transport framing
Access section
Egress Framing


KRLight
KISTI
JGN
KDDI Labs
SINET
AIST
PacificWave
ESnet
Caltech
StarLight
MAN LAN
AMPATH
GÉANT
NetherLight
Pionier
UvA
CERN
CzechLight
SouthernLight
RNP
Automated GOLE Fabric

AutoGOLE Fabric: Another View



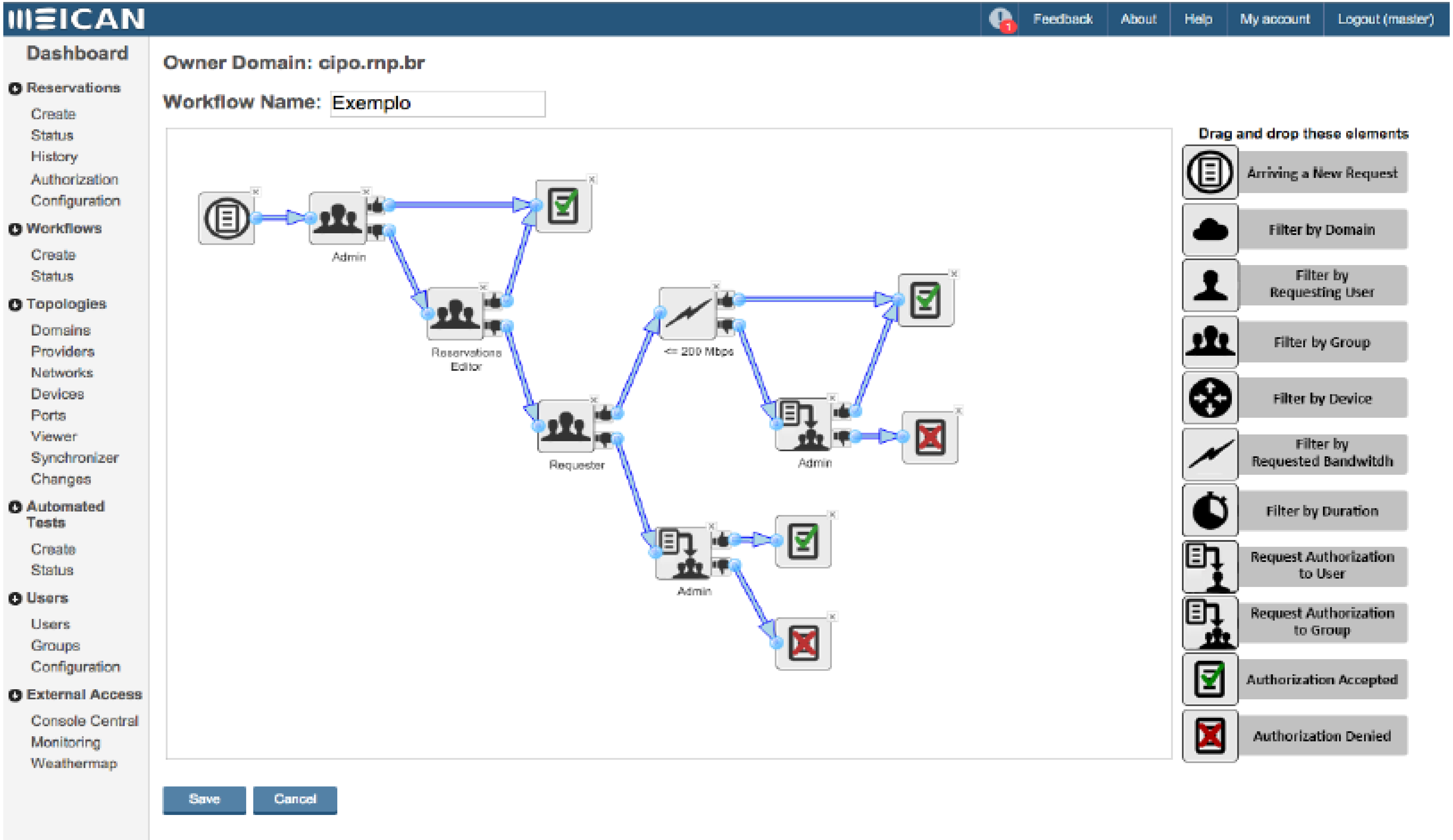


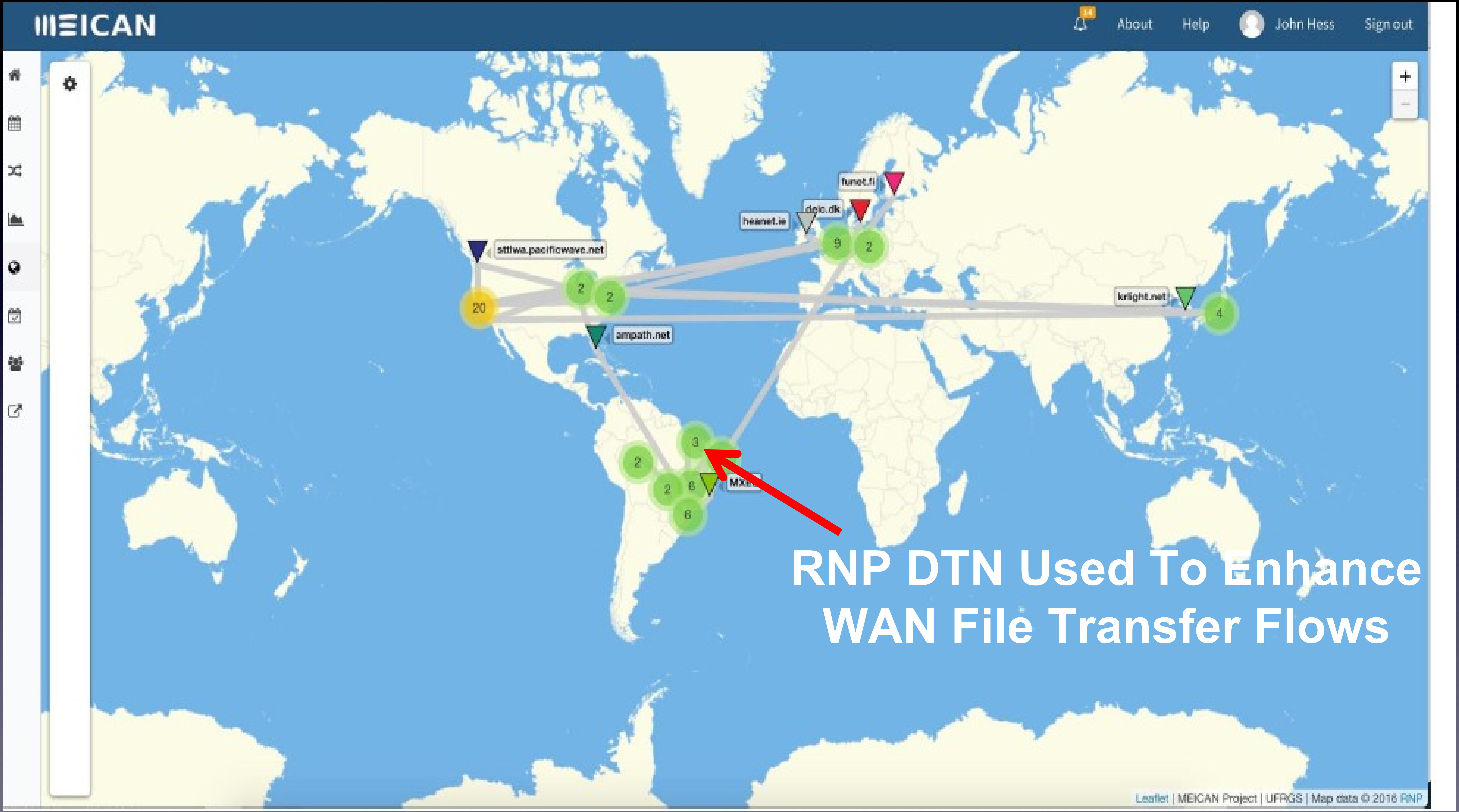
Demonstration 2: Using MEICAN Tools For International Multi-Domain Multi-Domain Provisioning

- MEICAN = Management Environment of Interdomain Circuits for Advanced Networks
- Web Application That Enables Users To Request VCs Between Well Defined Endpoints
- Implementation Depends On Operation Policies And Authorization Located In the Intermediate Domains That Connect Source and Destination Endpoints.
- MEICAN Uses Business Process Management (BPM) Concepts for Managing the VCs Establishment Process, Since VC Requested By Enduser to Network Devices Configurations.
- Main Contribution => Providing Dynamic Authorization Strategies Composed for Policies and Support.

14/06/2016

Management Environment of Inter-domain Circuits for Advanced Networks





RNP DTN Used To Enhance WAN File Transfer Flows

Demonstration 3: Transferring Large Scale Airline Data E2E Across WANs Using DTNs

- Transporting Large Files And Collections of Many Small Files E2E Across WANs
- Use NSI To Find Potential Path
- Use NSI To Establish Path
- Send Files
- Return Path Resources To Repository

Transferring LargeScale Airline Data E2E Across WANs Using DTNs

v5, 21 SEP 2017

Ingredients
- Using Globus Toolkit (NOT Globus Online)
- Has GridFTP under the hood
- Under Globus license (must be evaluated)
- 40Gbit/s data transfer expected
- VLAN 2038, multipoint/extending
- Including authentication/authorization framework, e.g. SURFconext

Minimal setup
- Data transfer between DTN@UvA to DTN@StarLight at 40G
- Compare this to IPv4 performance Chicago-Amsterdam

Additional features
- Single Sign-On
- Comparison to IPv6
- Auto-deletion of file when transfer completed
- >40Gbps data transfer
- Expanding sites for GLIF and/or SC

Ideas
- Dutch Research LAN Project

Initial SDX DTN Baseline Test Results

Summary of the mem-to-mem transfer

Sender \ Receiver	datanode-Startlight	dtn0.lsanca. pacificwave.net	Fiona-r-uva. vlan7.uvalight.net
dtn0.lsanca.pacificwave.net	48 Gb/s	-	-
fiona-r-uva.vlan7.uvalight.net	36 Gb/s	36 Gb/s	-
dtn1wa.datamovers.aarnet.edu.au	7.3 Gb/s*	7.0 Gb/s *	5.7 Gb/s*

Summary of the disk-to-disk transfer

Sender \ Receiver	datanode-Startlight	dtn0.lsanca. pacificwave.net	Fiona-r-uva. vlan7.uvalight.net
dtn0.lsanca.pacificwave.net	27 Gb/s	-	-
fiona-r-uva.vlan7.uvalight.net	19 Gb/s	9.3 Gb/s	-
dtn1wa.datamovers.aarnet.edu.au	4.6 Gb/s	5.4 Gb/s	2.2 Gb/s

* Peak Utilzation > 90 %, however, the average throughput affected by TCP slow-start

Initial SDX DTN Baseline Test Results and Setting

Summary of the disk IO performance

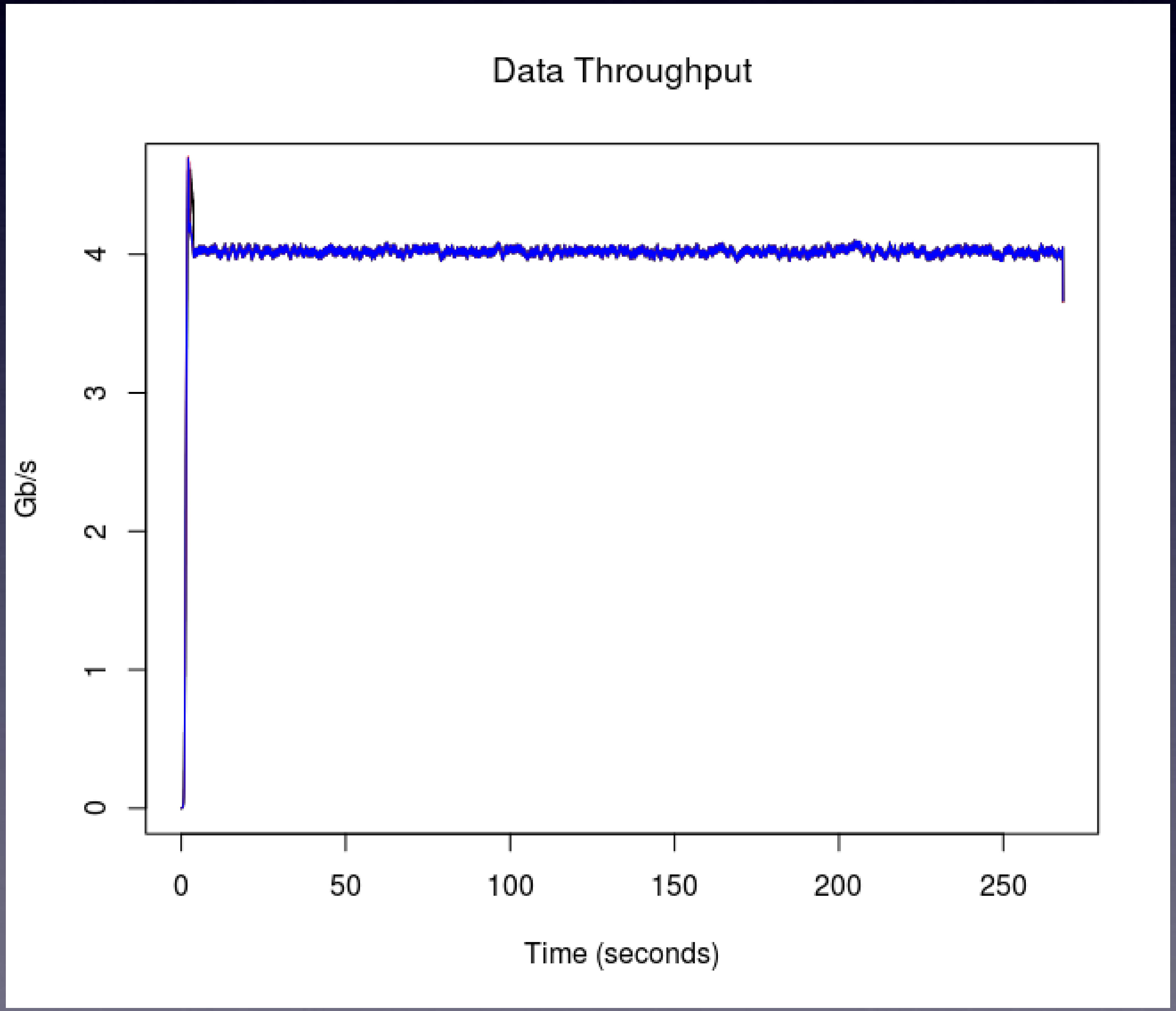
Machine	Disk read	Disk write
dtn0.lsanca.pacificwave.net	2.6 GB/s	1.1 GB/s
fiona-r-uva.vlan7.uvalight.net	3.5 GB/s	3.2 GB/s
dtn1wa.datamovers.aarnet.edu.au	2.6 GB/s	N/A
Starlight-datanode2	N/A	12 GB/s

RTT/CWND used for tests

Sender \ Receiver	datanode-Startlight	dtn0.lsanca.pacificwave.net	Fiona-r-uva.vlan7.uvalight.net
dtn0.lsanca.pacificwave.net	67 ms / 110 MB	-	-
fiona-r-uva.vlan7.uvalight.net	95 ms / 110 MB	162 ms / 193 MB	-
dtn1wa.datamovers.aarnet.edu.au	235 ms / 134 MB	216 ms / 128 MB	330 ms / 150 MB

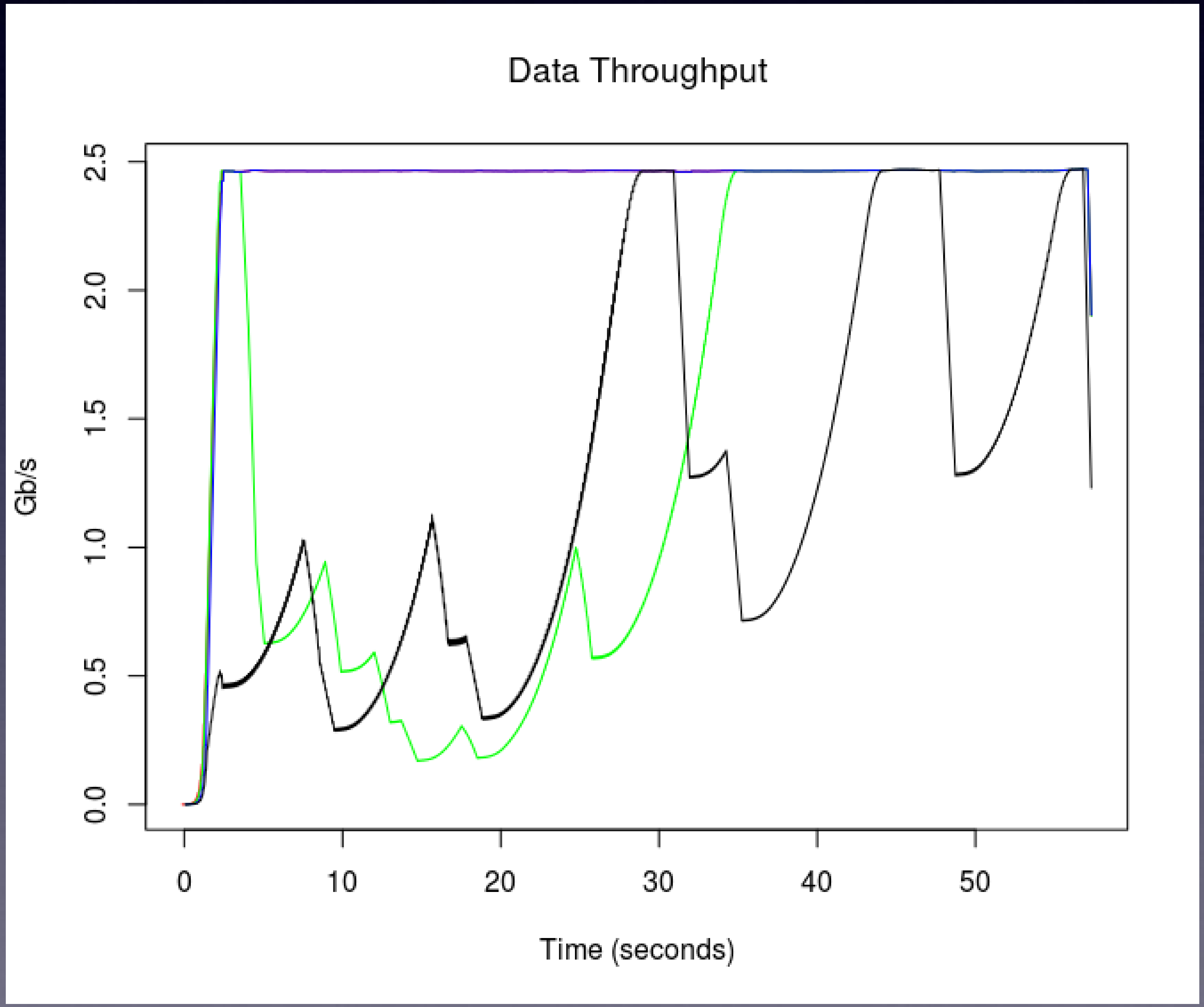
500 GB File Transfer Over Optimized Path

- 500 GB from Amsterdam to Chicago
  - GridFTP With 4 Parallel TCP Streams
  - Dedicated Layer 2 Path
  - 95 ms Round Trip Time
  - No Packet Loss
  - Throughput Bound By CPU



50 GB File Transfer Over Internet

- 50 GB from Amsterdam to Chicago
  - GridFTP With 4 Parallel TCP streams
  - Routed Over Internet
  - 113 ms Round Trip Time
  - Some Packet Loss/Congestion
  - Throughput Bound By Network



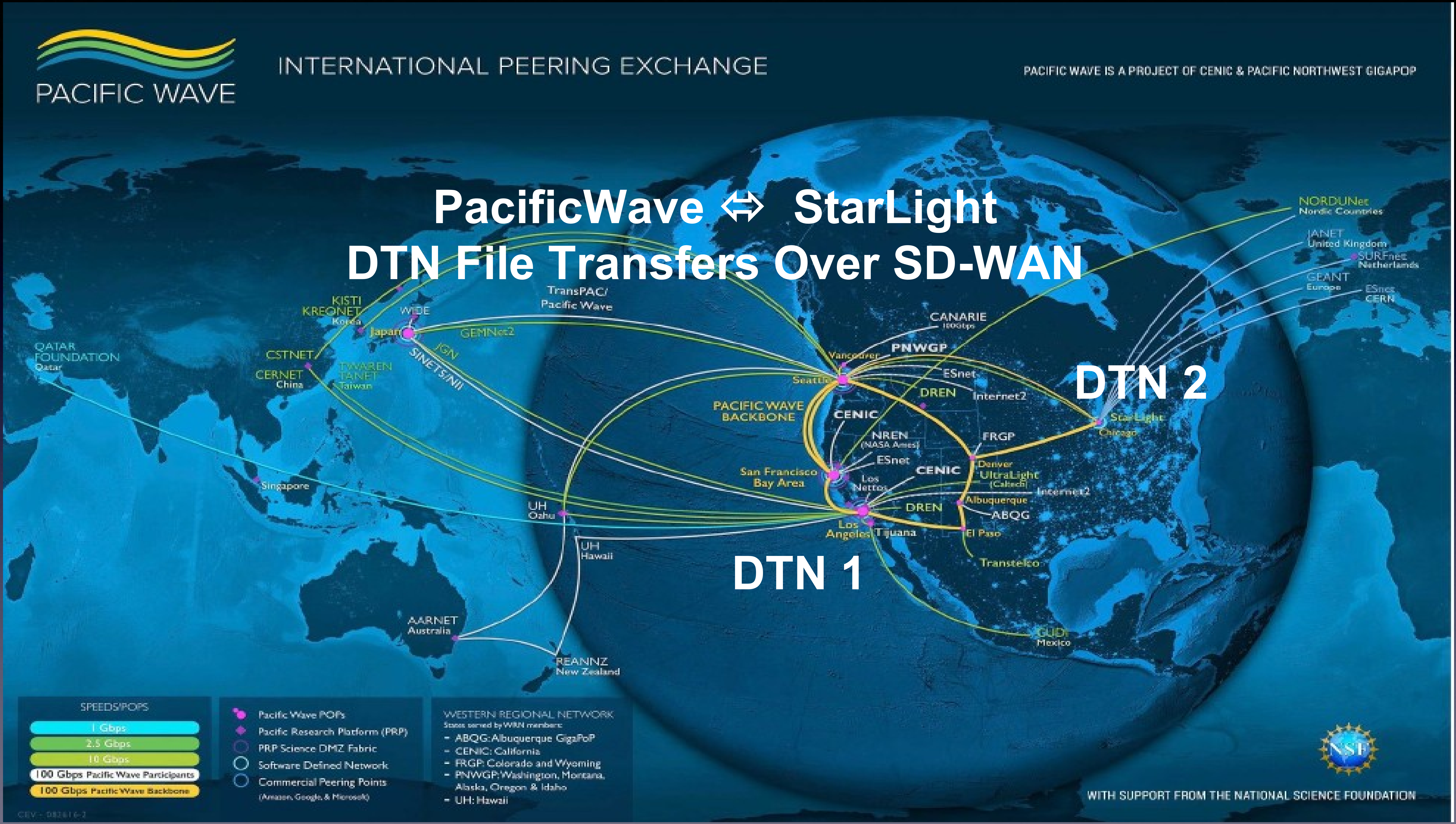
Constraints Affecting Data Transfers

- Input/Output
  - Storage System Needs To Be Able To Fully Process
  - Capabilities Of All Network Devices Along the Path
- CPU
  - Transfer Protocol and Application Used Influences Impact On CPU, e.g. Single/Multi Threaded
  - Storage System May Cause Additional Overhead, e.g. RAID Calculations
- Architectural Constraints
  - How Devices Within a System Are Connected Affects Their Performance When Used In Combination, e.g. NUMA Nodes

Demonstration 4: Transferring Large Scale Science Data E2E Across WANs Using DTNs

- Sending Large Files (OR Collections Of Many Small Files) E2E Across WANs
- Use SDN/SDX To Find Potential Path
- Use SDN/SDX To Establish Path
- Send File(s)
- Return Path Resources To Repository

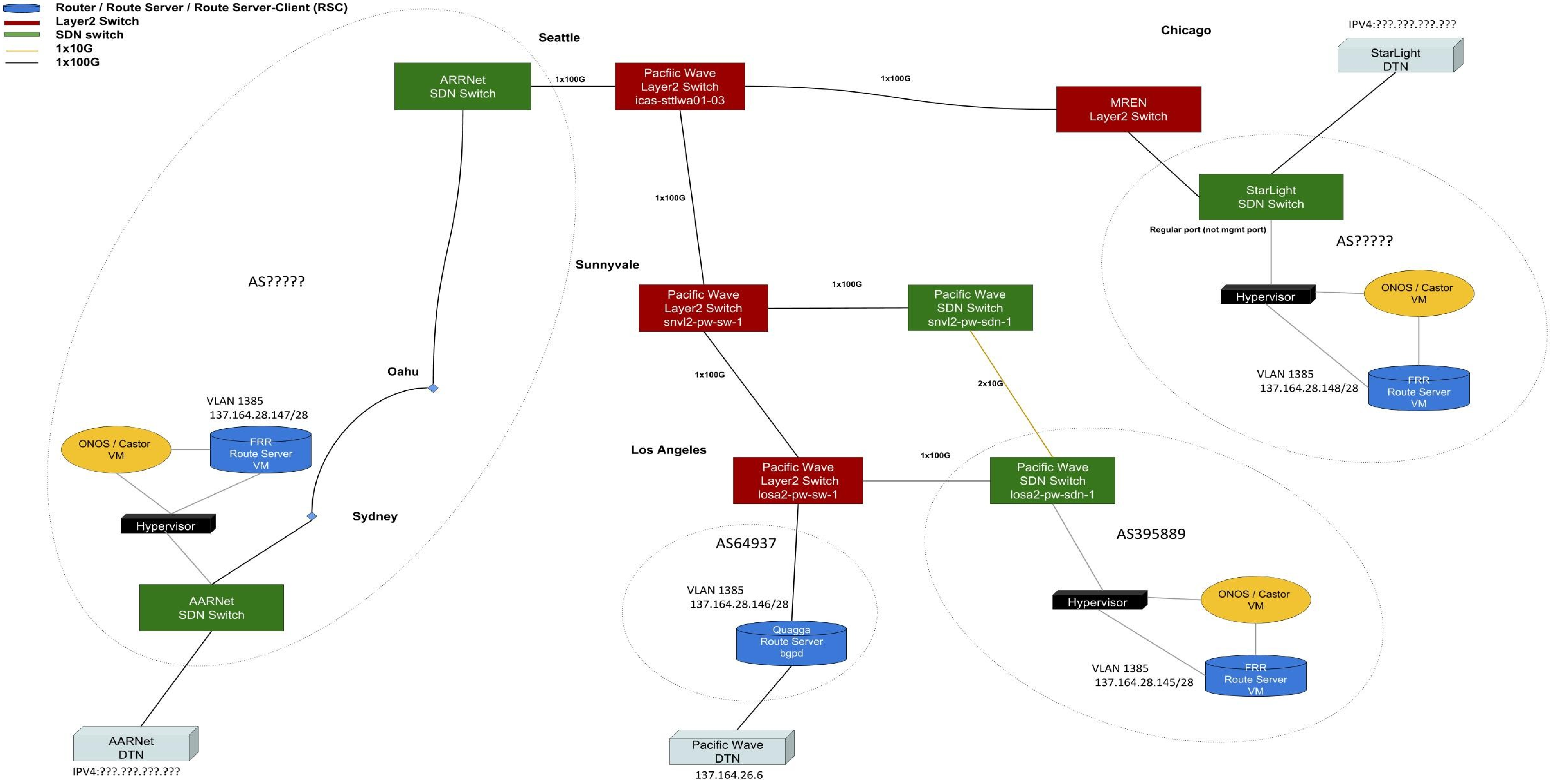
PacificWave ⇔ StarLight
DTN File Transfers Over SD-WAN



Demonstration 5: SDX Software Defined Services

- Creating Services Based On Policy-Driven SDX and SDN Dynamic Provisioning Control.
- Resources at Multiple SDXs
- Resource Requests *and* Orchestration via SDN
- Combining Layer 2 and Layer 3 Resources With SDX Support for BGP Instances

AARNet - Pacific Wave - Starlight
Inter-domain SDX Topology v0.4



NOTE: this diagram represents a subset of sites, devices, and connections

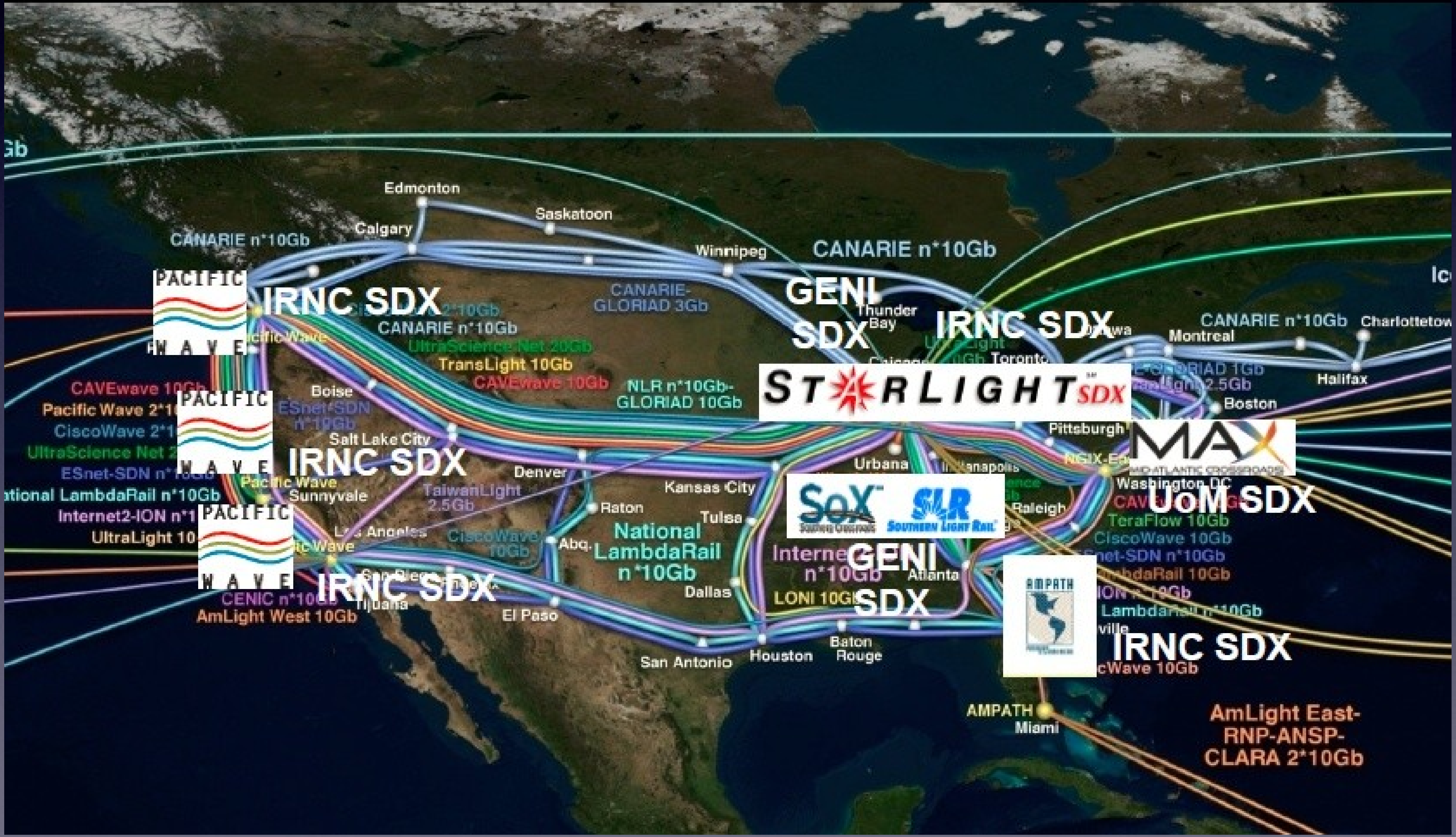
https://docs.google.com/drawings/d/1NIcvdKg8Zy4mFH-uIYmYsJdz5HKpmgjfk1QyBKpntel/edit?usp=sharing
V0.04 20170801

Demonstration 6: Transferring Large Files E2E Across WANs Enabled By SD-WANs and SDXs

- Sending Large Files (OR Collections Of Many Small Files) E2E Across WANs
- Use SDN/SDX To Find Potential Path
- Use SDN/SDX To Establish Path
- Send File(s)
- Return Path Resources To Repository

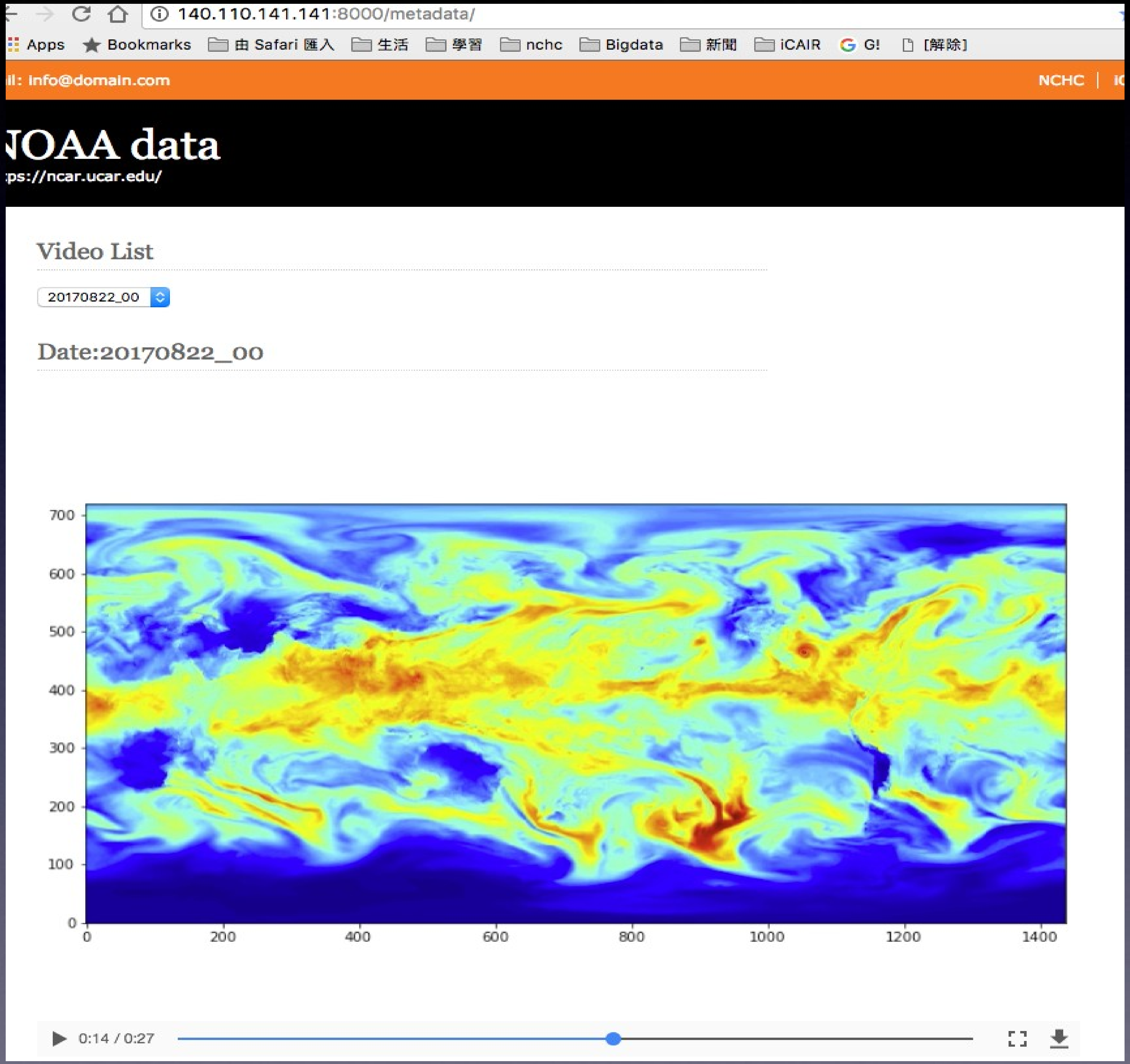
Emerging SDX Fabric

Planned US SDX Interoperable Fabric

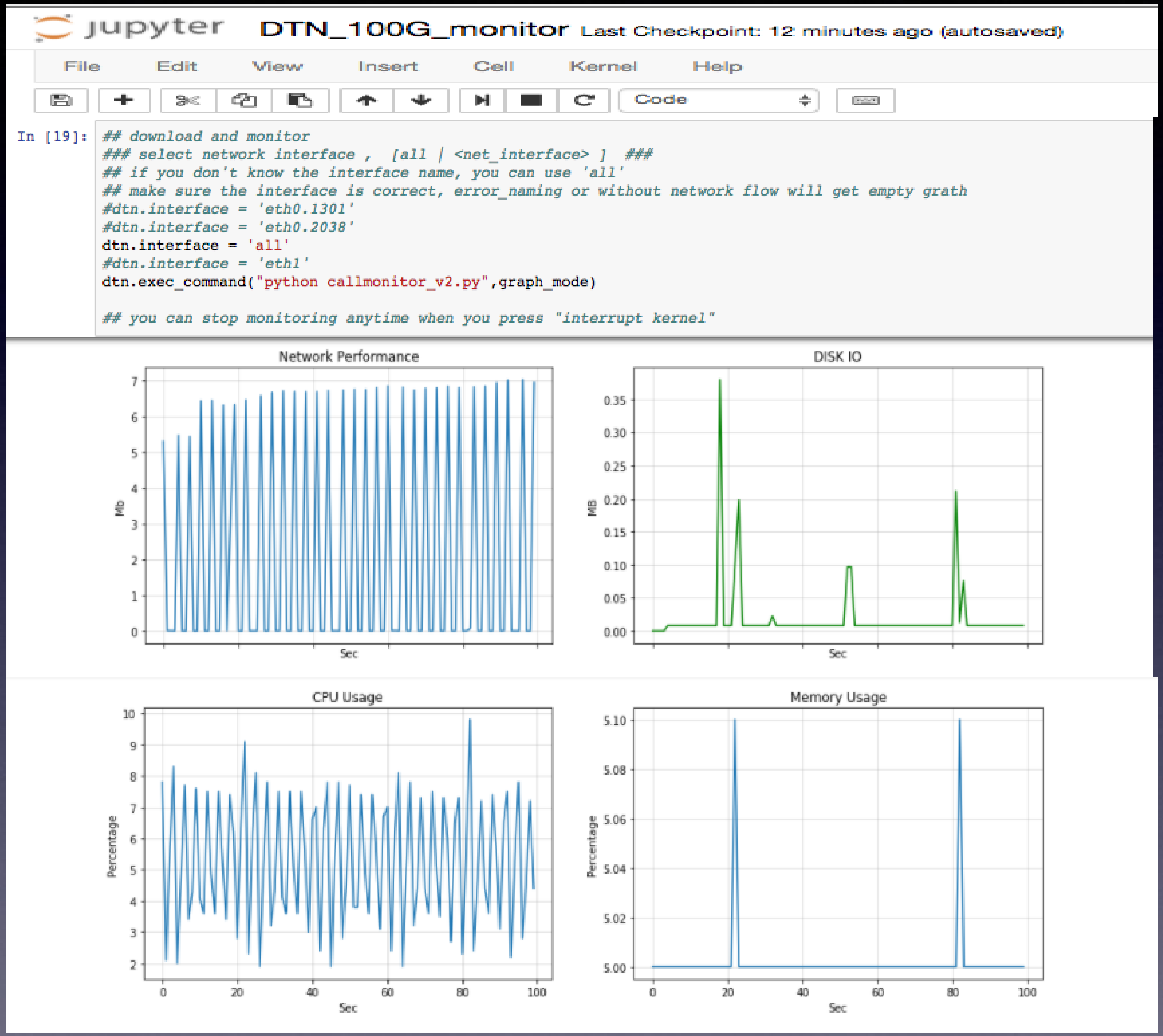


LHC/ATLAS SDX DTN Service Prototype

GeoScience SDX DTN service Prototype

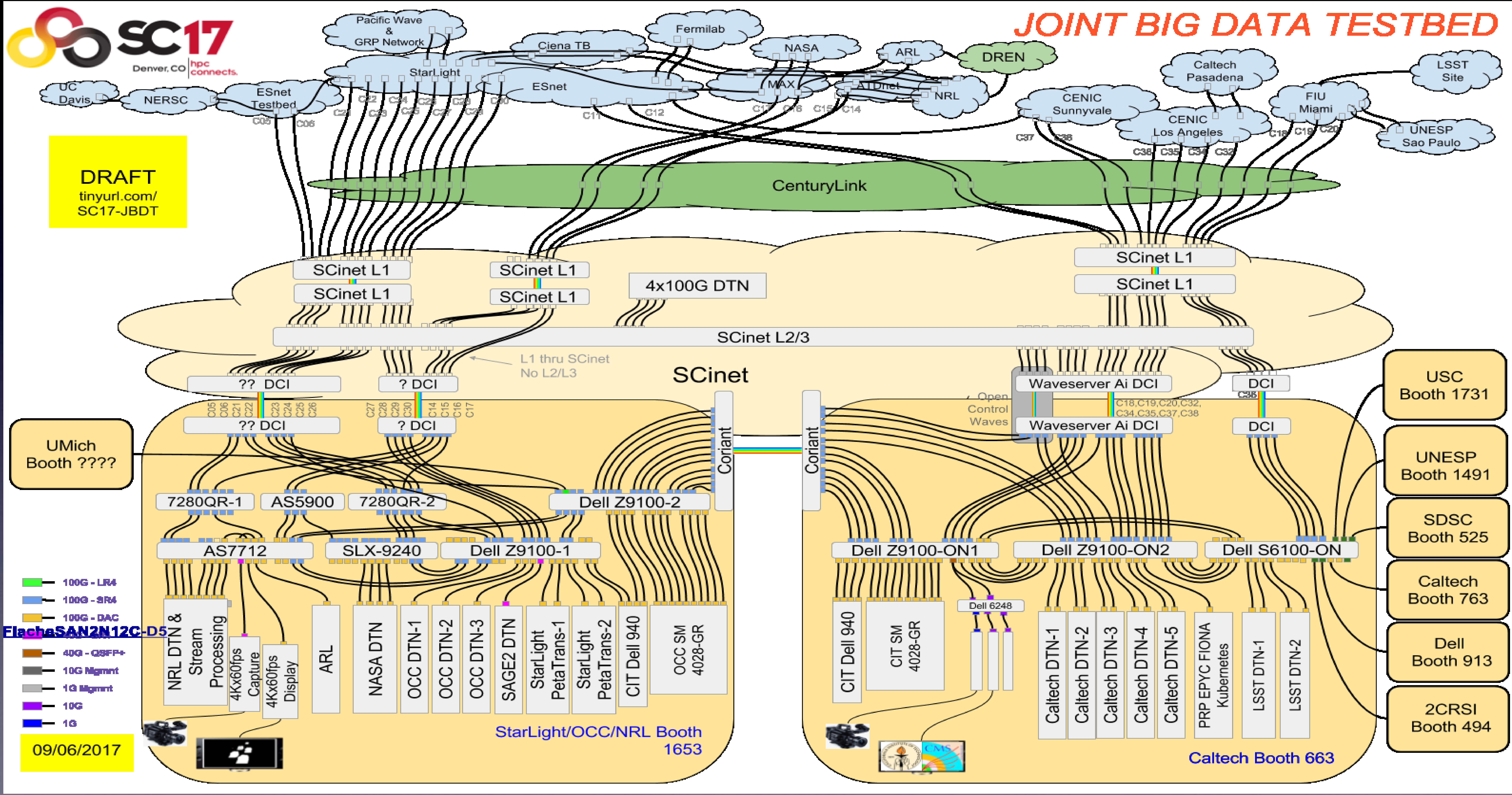


File Screening workflow



File Transfer workflow

Global LambdaGrid Workshop Demonstrations As Prestaging For SC17 – Multiple Large Scale SDN/SDX/SDI International/National Demonstrations



SC17
Denver, CO
hpc connects.
JOINT BIG DATA TESTBED
DRAFT
tinyurl.com/
SC17-JBDT
Pacific Wave & GRP Network
Ciena TB
Fermilab
NASA
ARL
DREN
StarLight
ESnet
MAX
ATDnet
NRL
UC Davis
NERSC
ESnet Testbed
Caltech Pasadena
LSST Site
CENIC Sunnyvale
CENIC Los Angeles
FIU Miami
UNESP Sao Paulo
CenturyLink
SCinet L1
4x100G DTN
SCinet L2/3
SCinet
L1 thru SCinet
No L2/L3
?? DCI
? DCI
Waveserver Ai DCI
DCI
Open Control Waves
C18,C19,C20,C32,
C34,C35,C37,C38
Coriant
UMich
Booth ????
USC
Booth 1731
UNESP
Booth 1491
SDSC
Booth 525
Caltech
Booth 763
Dell
Booth 913
2CRSI
Booth 494
7280QR-1
AS5900
7280QR-2
Dell Z9100-2
AS7712
SLX-9240
Dell Z9100-1
Dell Z9100-ON1
Dell Z9100-ON2
Dell S6100-ON
Dell 6248
NRL DTN & Stream Processing
4Kx60fps Capture
4Kx60fps Display
ARL
NASA DTN
OCC DTN-1
OCC DTN-2
OCC DTN-3
SAGE2 DTN
StarLight PetaTrans-1
StarLight PetaTrans-2
CIT Dell 940
OCC SM 4028-GR
CIT Dell 940
CIT SM 4028-GR
Caltech DTN-1
Caltech DTN-2
Caltech DTN-3
Caltech DTN-4
Caltech DTN-5
PRP EPYC FIONA Kubernetes
LSST DTN-1
LSST DTN-2
StarLight/OCC/NRL Booth
1653
Caltech Booth 663
100G - LR4
100G - SR4
100G - DAC
40G - QSFP+
10G Mgmnt
1G Mgmnt
10G
1G
FlacheSAN2N12C-D5
09/06/2017

Thanks!

- Facilities: SURFnet, NetherLight, KLM-AF Research Network, ANA-300, CANARIE, Montreal Open Exchange (MOXY), SDN Exchanges (SDXs) at the StarLight International/National Communications Exchange Facility in Chicago, Metropolitan Research and Education Network (MREN) in the Midwest, RNP, Pacific Wave, CENIC, Pacific Northwest GigaPOP, ASGC/Taiwan, NCHC/Taiwan, AARnet. RNP SDN (Brazil), KREONET