Potential LHC Networking SC24 NRE Demonstrations

Edoardo Martelli, Carmen Misa Moreira, Joe Mambretti, Bruno Hoeft, Tom Lehman, Shawn McKee, Marian Babik, Vitaliy Kondratenko, Tristan Sullivan, Phil Demar, Syed Asid Shah, et al,

LHCOPN-LHCONE MEETING #52
CITTADELLA UNIVERITARIA
CANTANIA, ITALY
APRIL 9-11, 2024

Planning for SC24 (Atlanta, Georgia)

- **SCinet Sponsored Network Research Exhibition (NRE) Descriptions (June 1, 2024)**
- **NRE Submissions Define Demonstrations**
- **Lead To Assessment of Required Resources, Including WANs, Edge Devices (Enhanced Descriptions Planned for SC24)**
- **Results: Implementation of Services/Resources**
- **Assists With Pre-Conference Staging Facilities**

NREs: Verifying/Authenticating New Advanced Concepts

- Formulating New Architecture, Services, Techniques, Technologies Through Large Scale, WAN Demonstrations
- Proving Concepts With Empirical, Replicatable Experiments
- Creating Prototypes
- Communicating Results To Wide Audiences
- Leveraging Large Scale Testbeds, e.g., SCinet
- Contributing To The Design and Implementation of Testbeds
- Building Blocks for Global Research Platform (GRP)

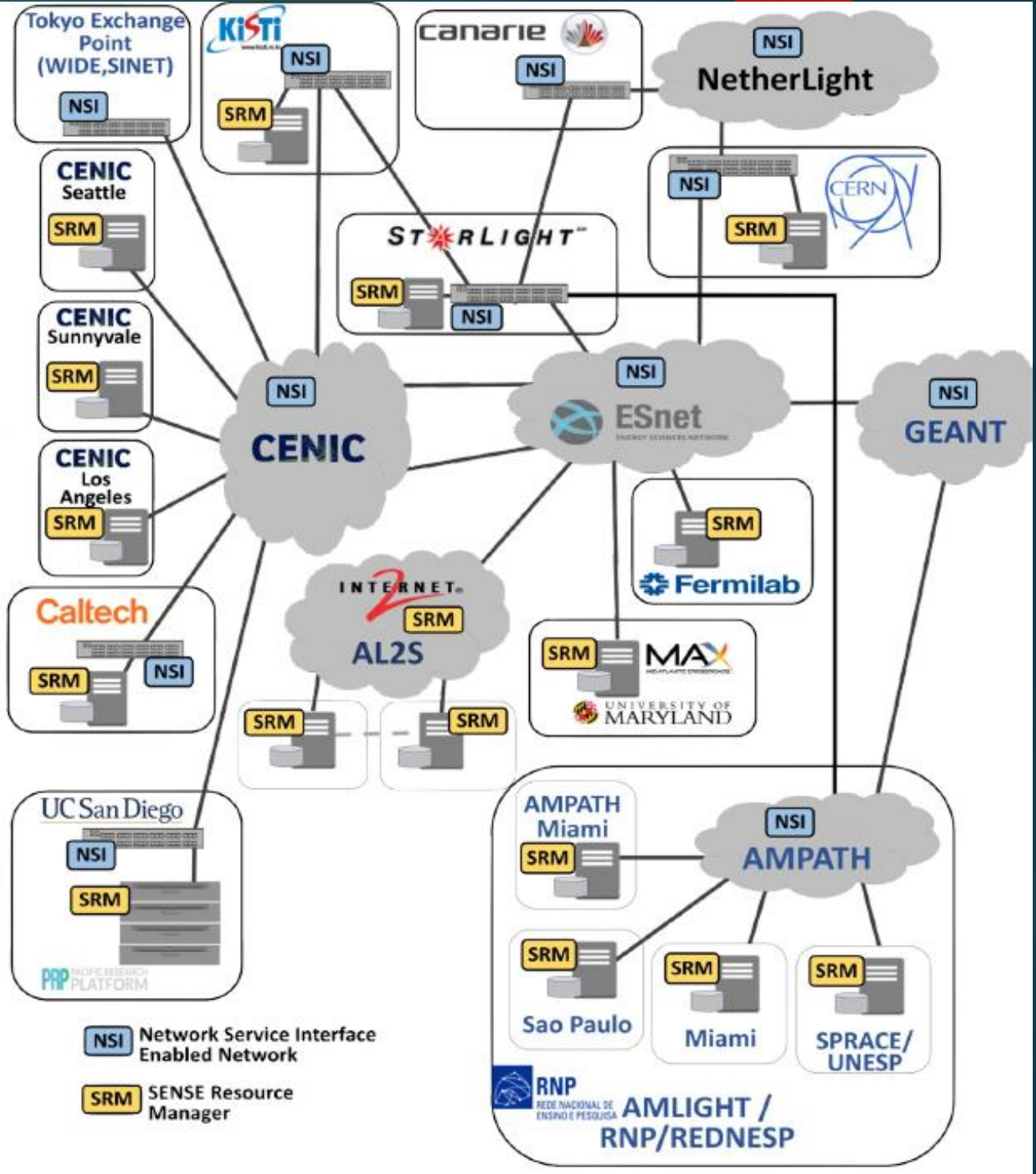
The GRP: A Platform For Global Science

A Next Generation, Software Defined, Globally Distributed, Multi-Domain Computational Science Environment

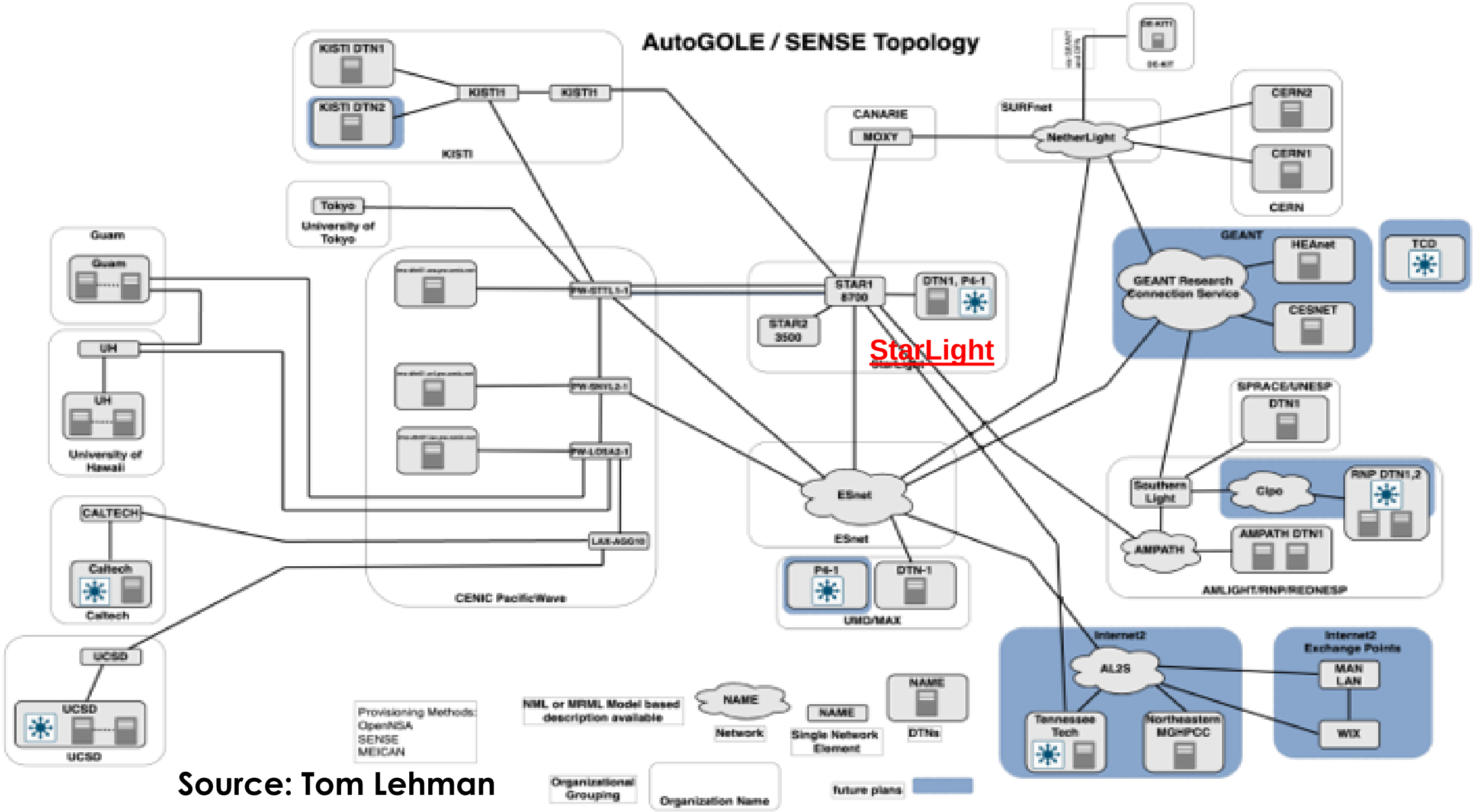
Global Research Platform/AutoGOLE Open R&E Exchanges

SENSE/AutoGole

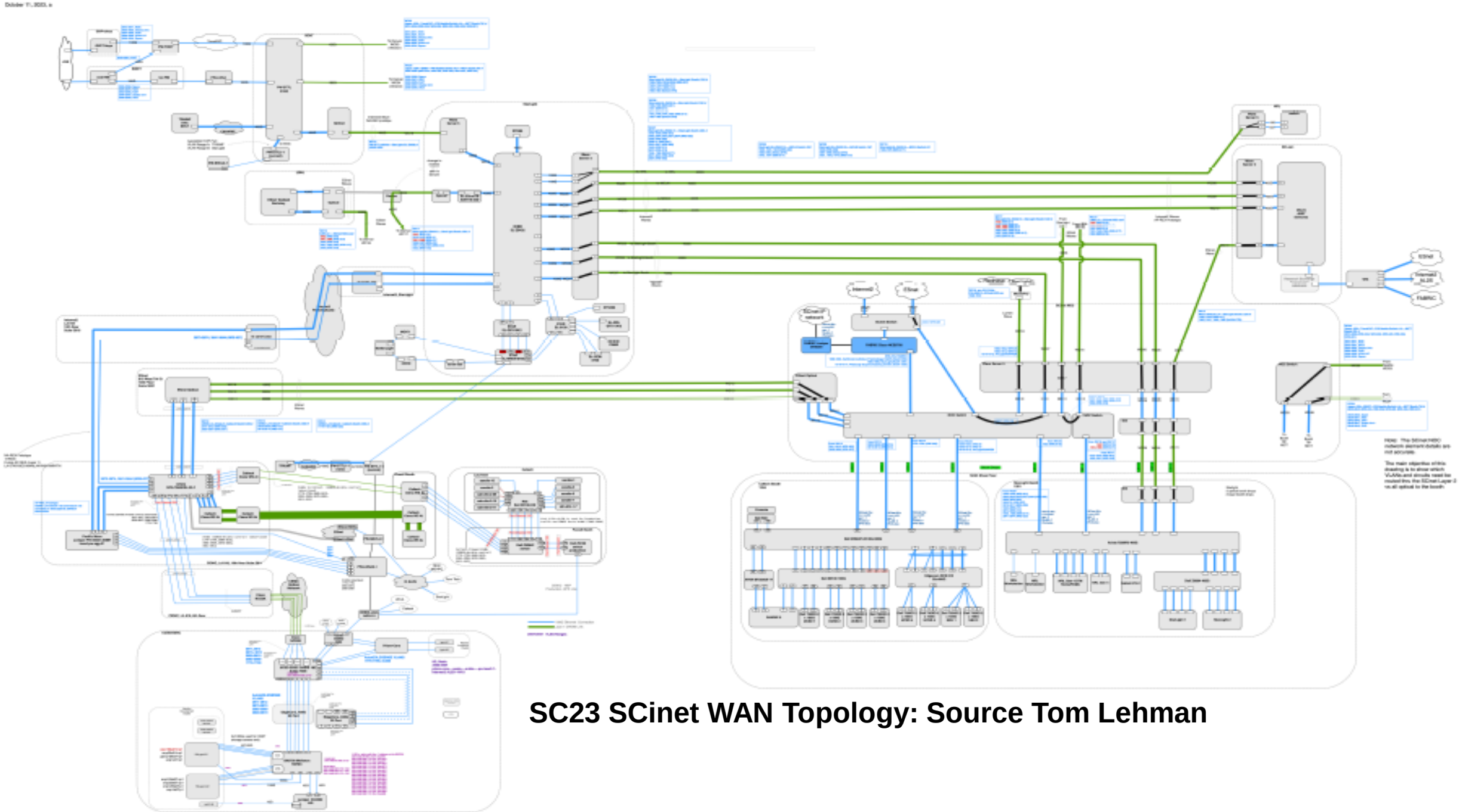
- AutoGOLE, NSI, and SENSE working together provide the mechanisms for complete end-to-end services which includes the network and the attached End Systems (DTNs).



Source: Tom Lehman

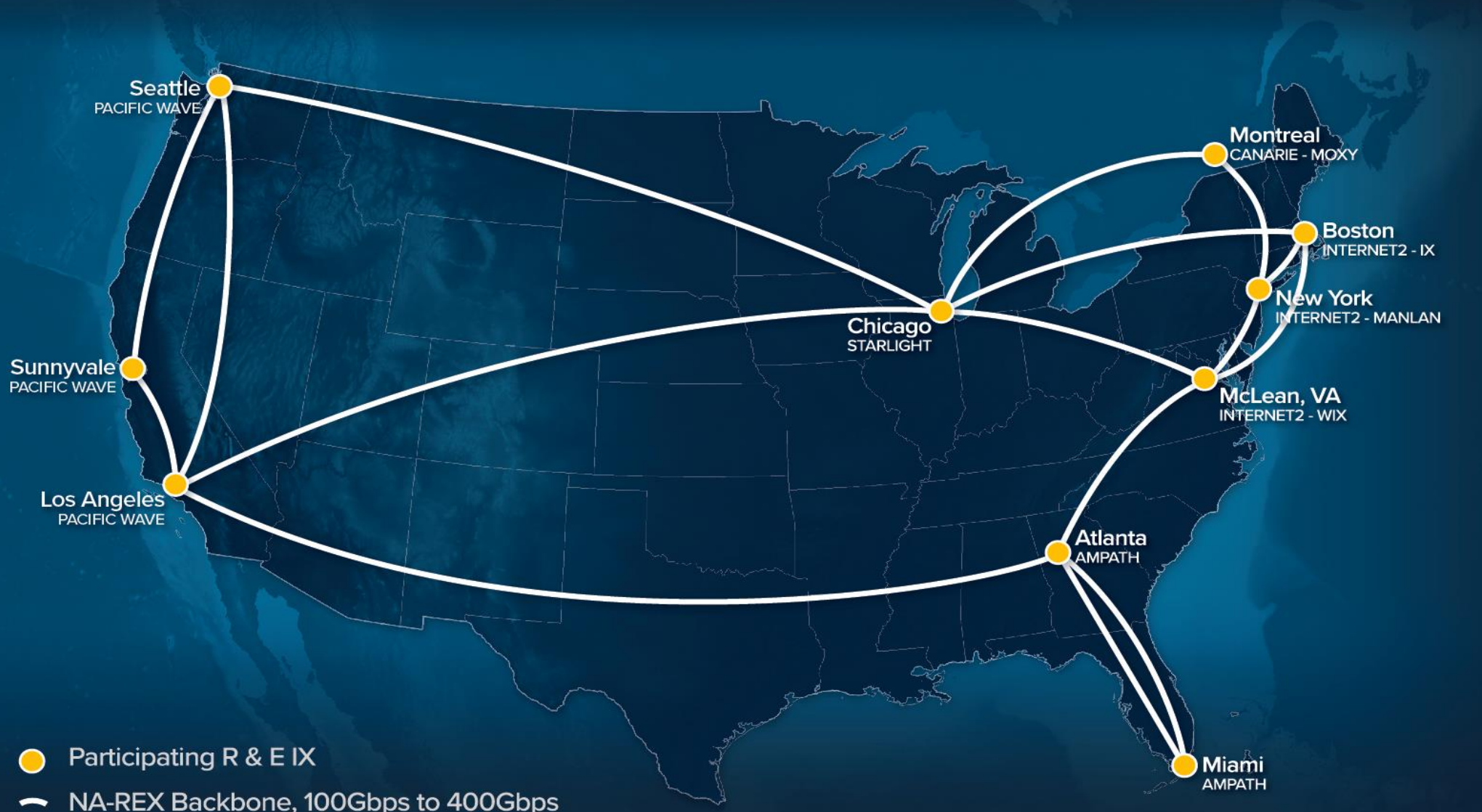


Source: Tom Lehman



SC23 SCinet WAN Topology: Source Tom Lehman

NA-REX North America Research & Education Exchange Collaboration



AMPATH PATHWAY WITH THE AMERICAS

canarie

ESnet ENERGY SCIENCES NETWORK

INTERNET2

International Networks at Indiana University

STARLIGHT The Optical STAR TAP

CENIC

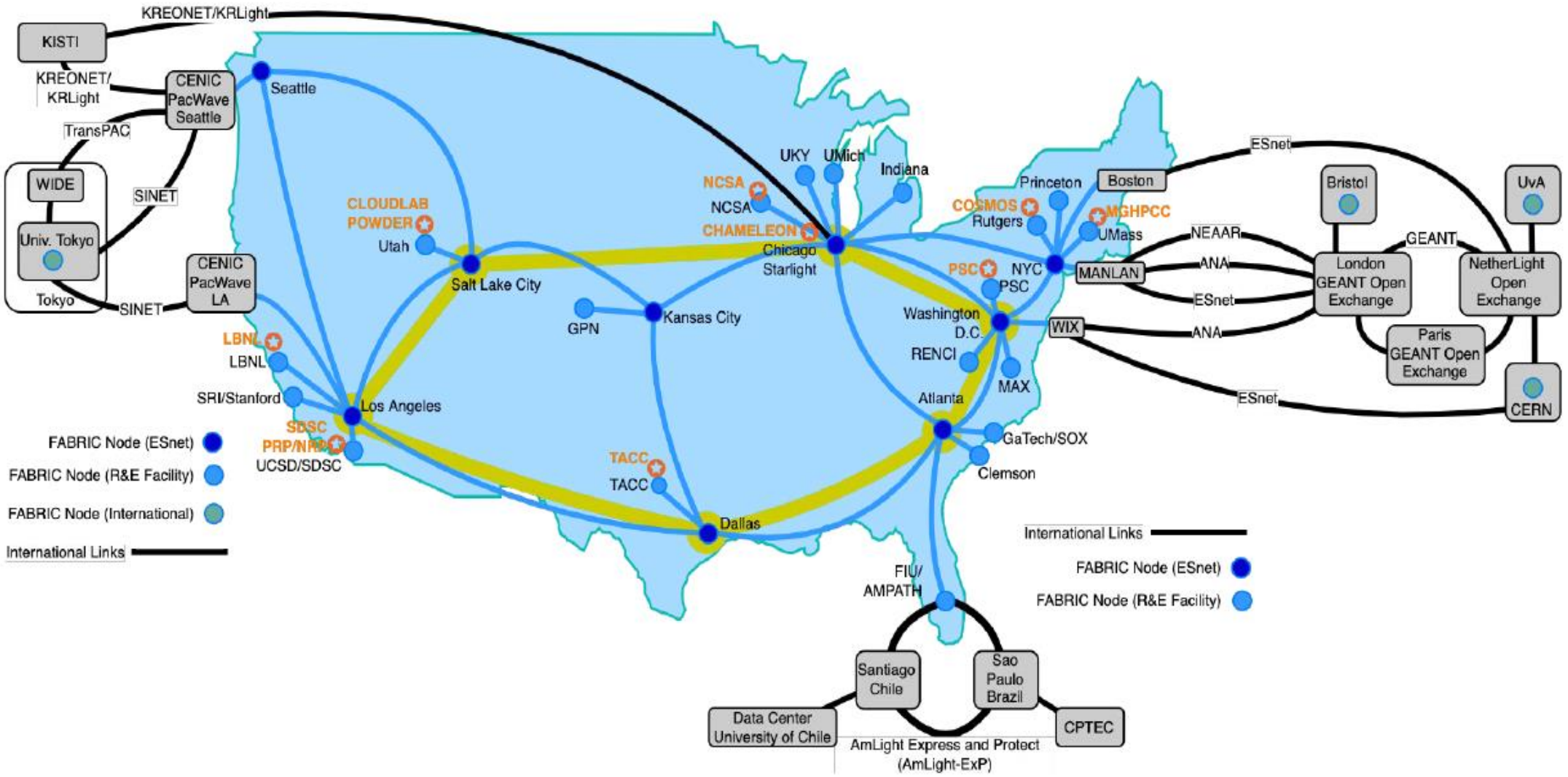
PACIFIC NORTHWEST GIGAPOP

PACIFIC WAVE

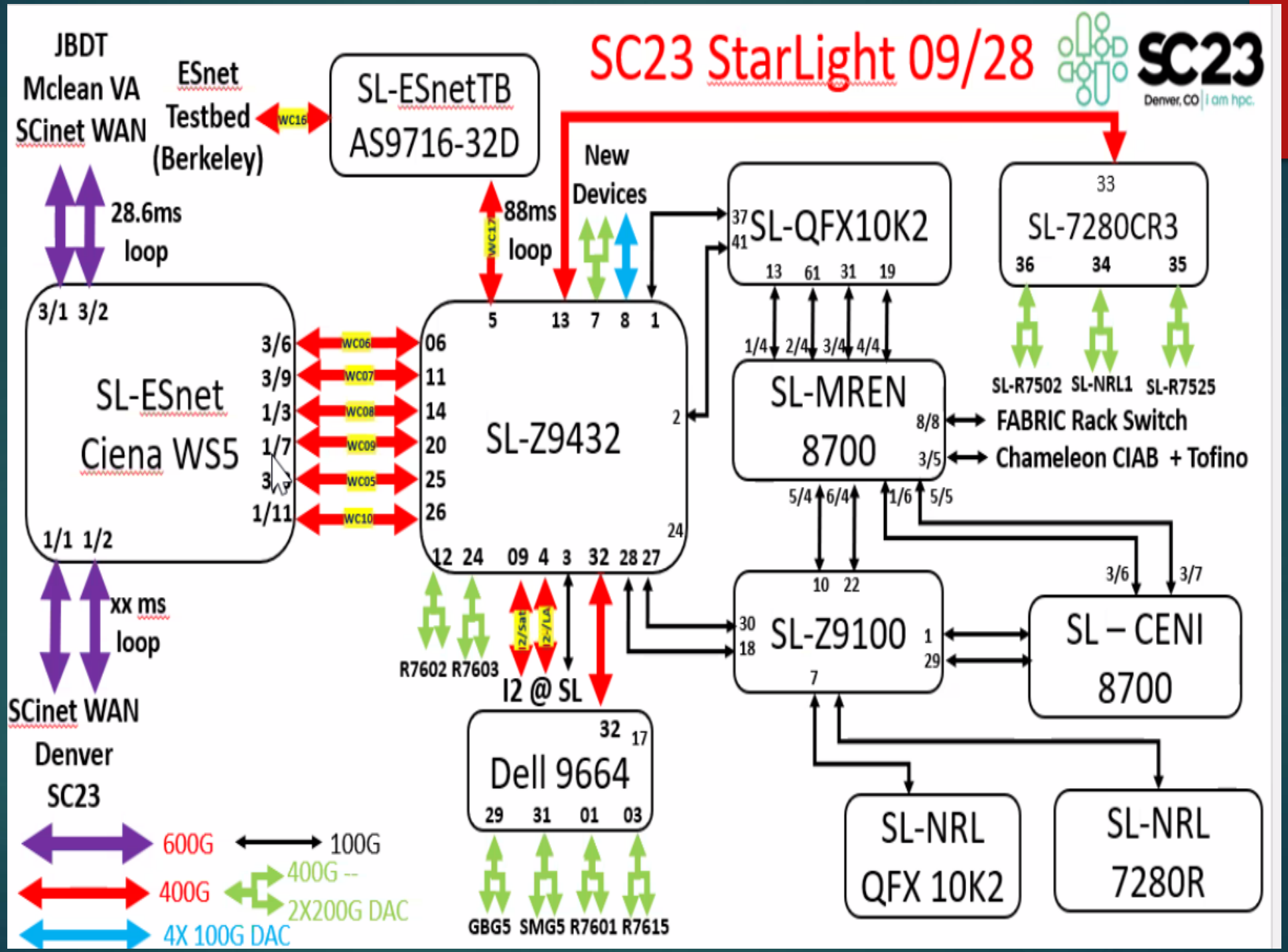
November 2023

FABRIC Testbed (+FAB)

renci



FABRIC Topology - with FAB Sites


SC23 StarLight 09/28
SC23
Denver, CO | i am hpc.
JBDT
Mclean VA
SCinet WAN
ESnet
Testbed
(Berkeley)
WC16
SL-ESnetTB
AS9716-32D
28.6ms
loop
88ms
loop
WC17
New
Devices
SL-QFX10K2
37
41
13 61 31 19
SL-7280CR3
33
36 34 35
SL-R7502 SL-NRL1 SL-R7525
SL-MREN
8700
8/8
3/5
FABRIC Rack Switch
Chameleon CIAB + Tofino
1/4 2/4 3/4 4/4
5/4 6/4
1/6 5/5
3/1 3/2
SL-ESnet
Ciena WS5
1/1 1/2
3/6
3/9
1/3
1/7
3
1/11
WC06
WC07
WC08
WC09
WC05
WC10
SL-Z9432
5 13 7 8 1
06
11
14
20
25
26
2
24
12 24 09 4 3 32 28 27
R7602 R7603
I2/Sat
I2-/LA
I2 @ SL
xx ms
loop
SCinet WAN
Denver
SC23
SL-Z9100
10 22
30
18
1
29
7
3/6
3/7
SL – CENI
8700
Dell 9664
32 17
29 31 01 03
GBG5 SMG5 R7601 R7615
SL-NRL
QFX 10K2
SL-NRL
7280R
600G
400G
4X 100G DAC
100G
400G --
2X200G DAC

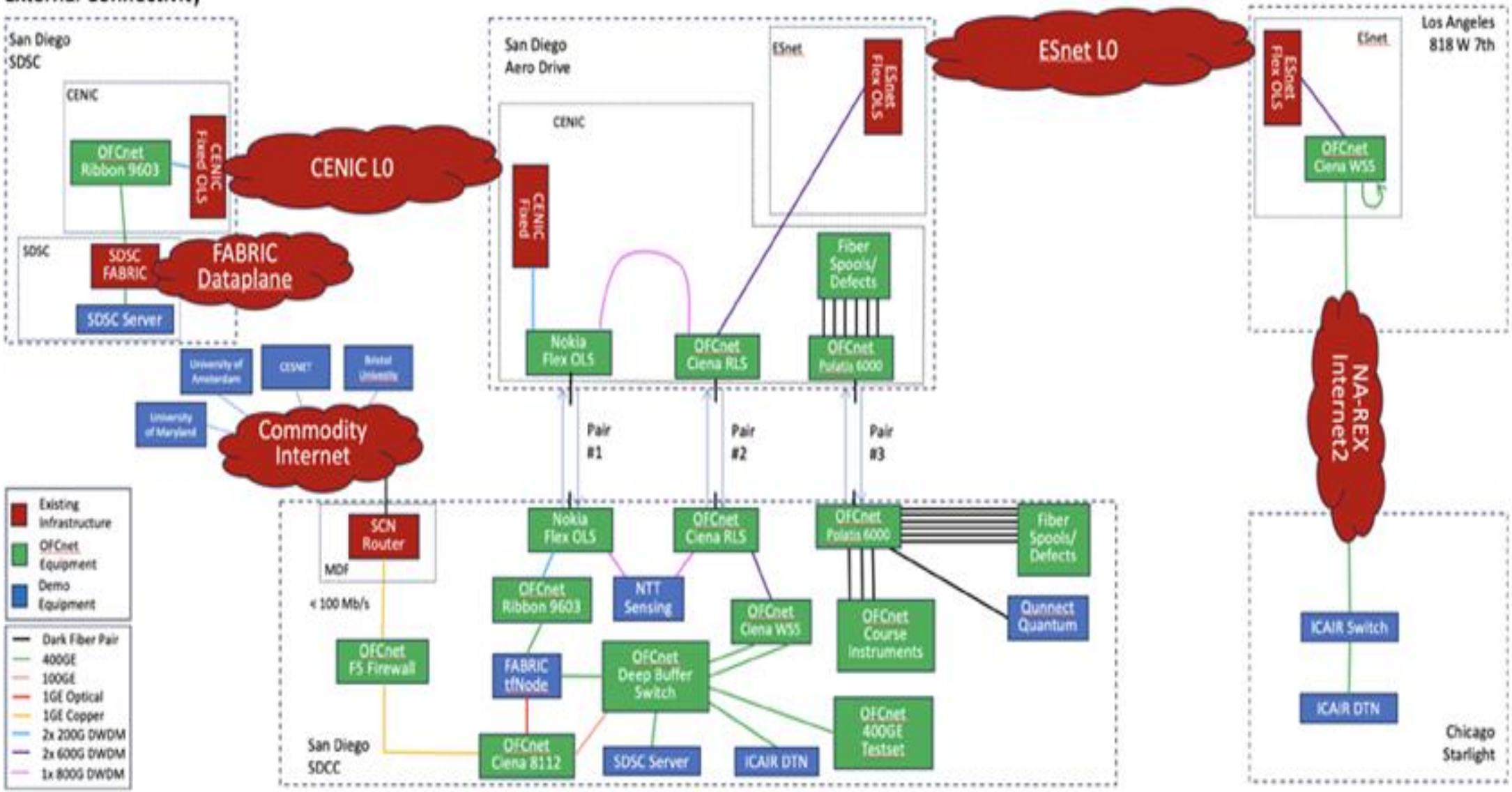
SENSE provisioning system

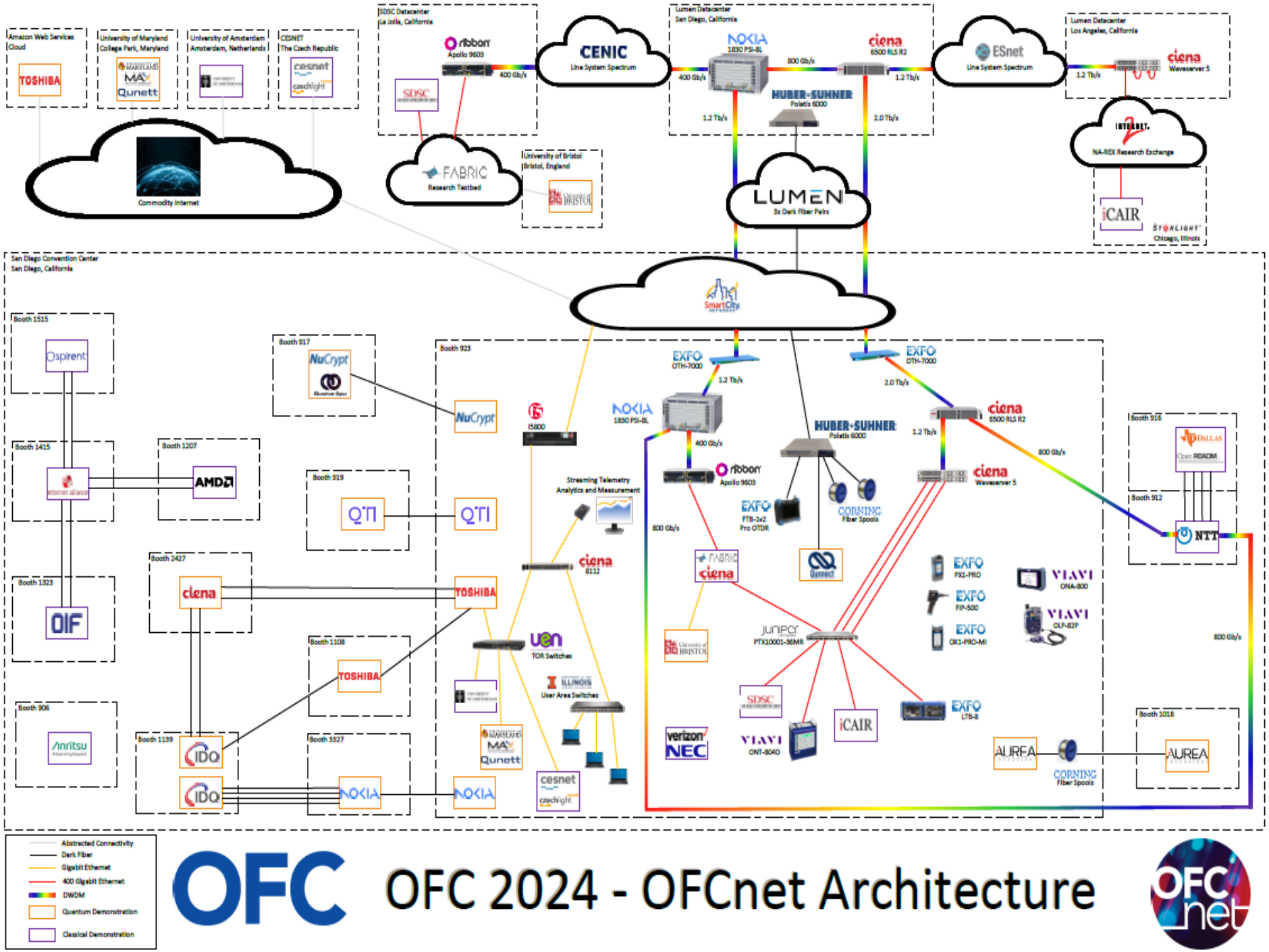
SENSE (SDN for E2E Networked Science at the Exascale): provision system that dynamically builds end-to-end virtual guaranteed networks across administrative domains without manual intervention.

- Provisioning automation: bring-up and management of services without human involvement.
- Multi-domain: multiple administrative domains, independent policies and AUP (Acceptable Use Policy).
- Resource orchestration: allocation and reservation of resources including compute, storage and network.
- End-to-end: DTN NIC to DTN NIC, across Science DMZ (Demilitarized zone), WANs, Open exchange points...

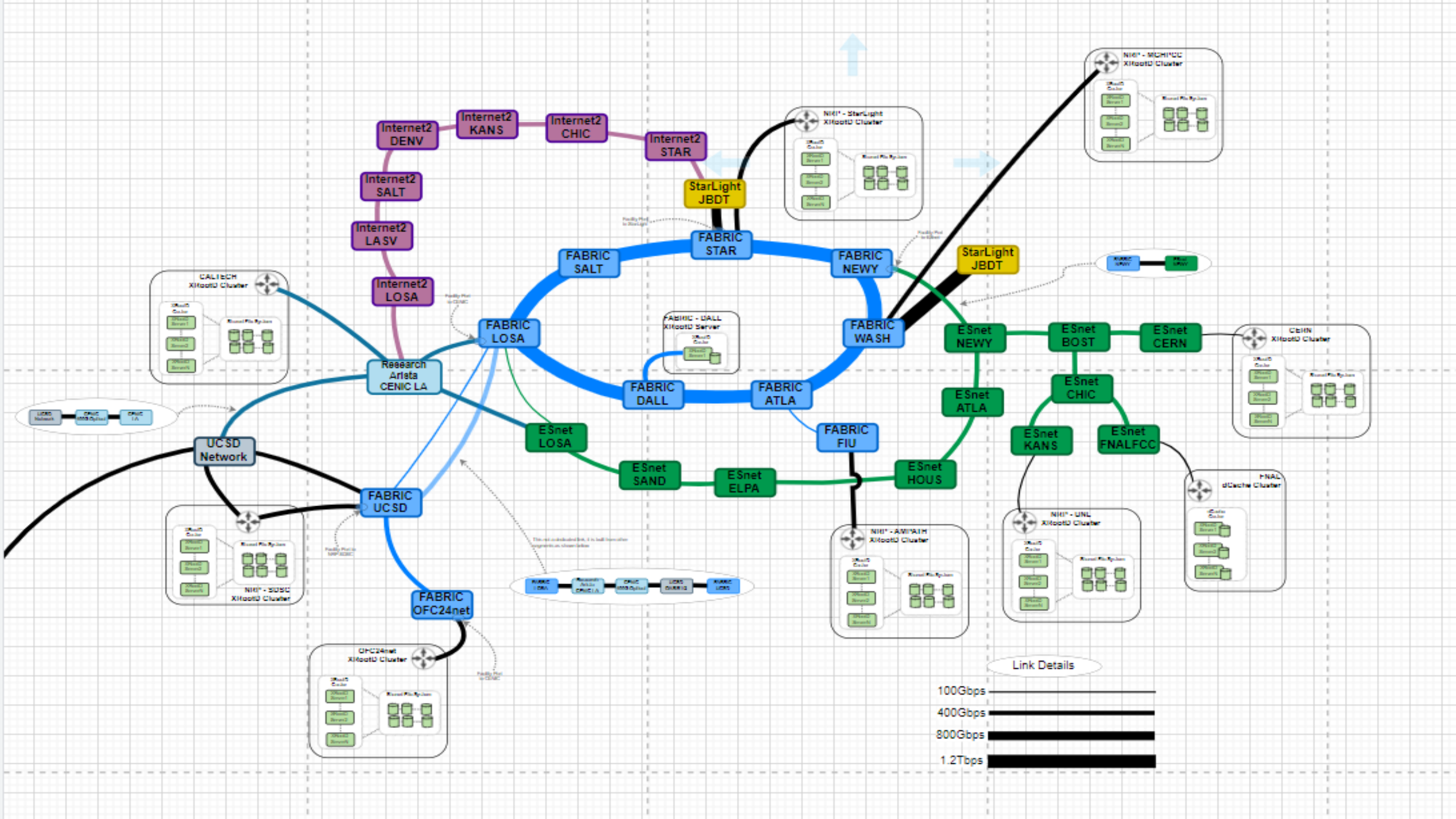
ESnet
ENERGY SCIENCES NETWORK

External Connectivity

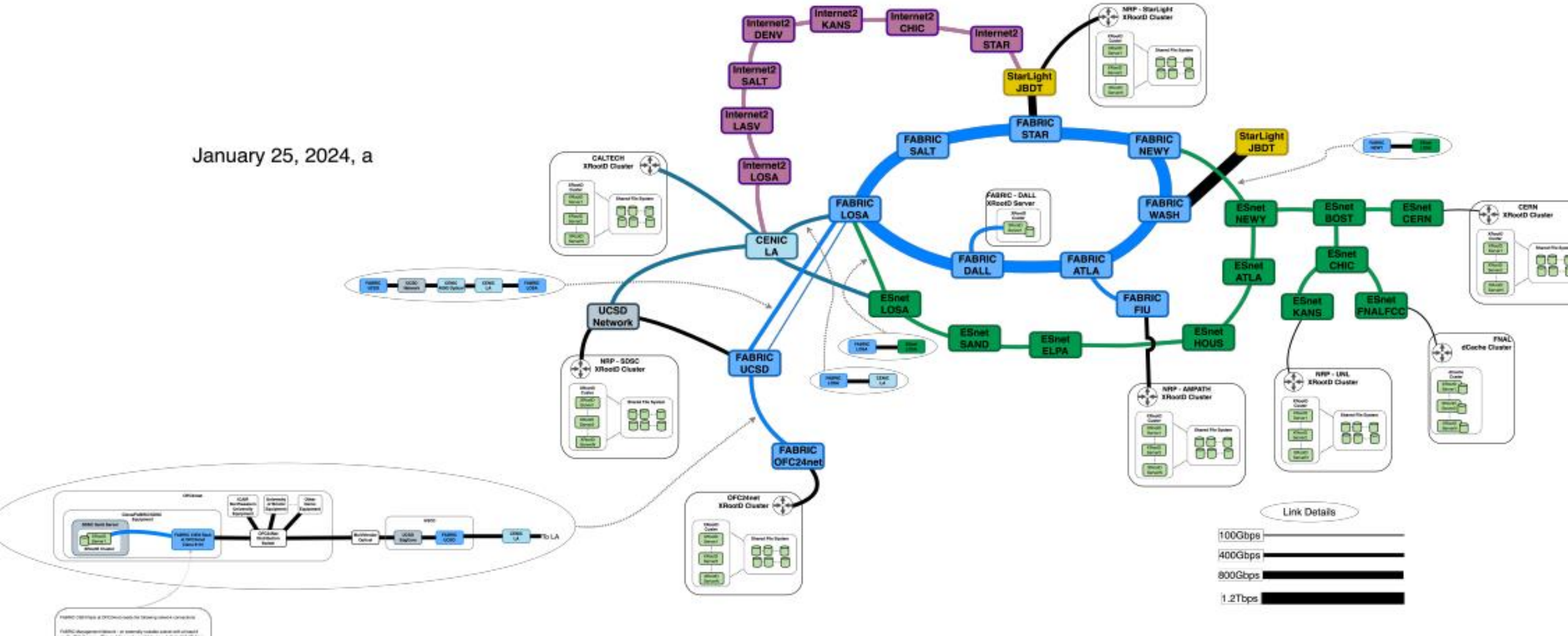


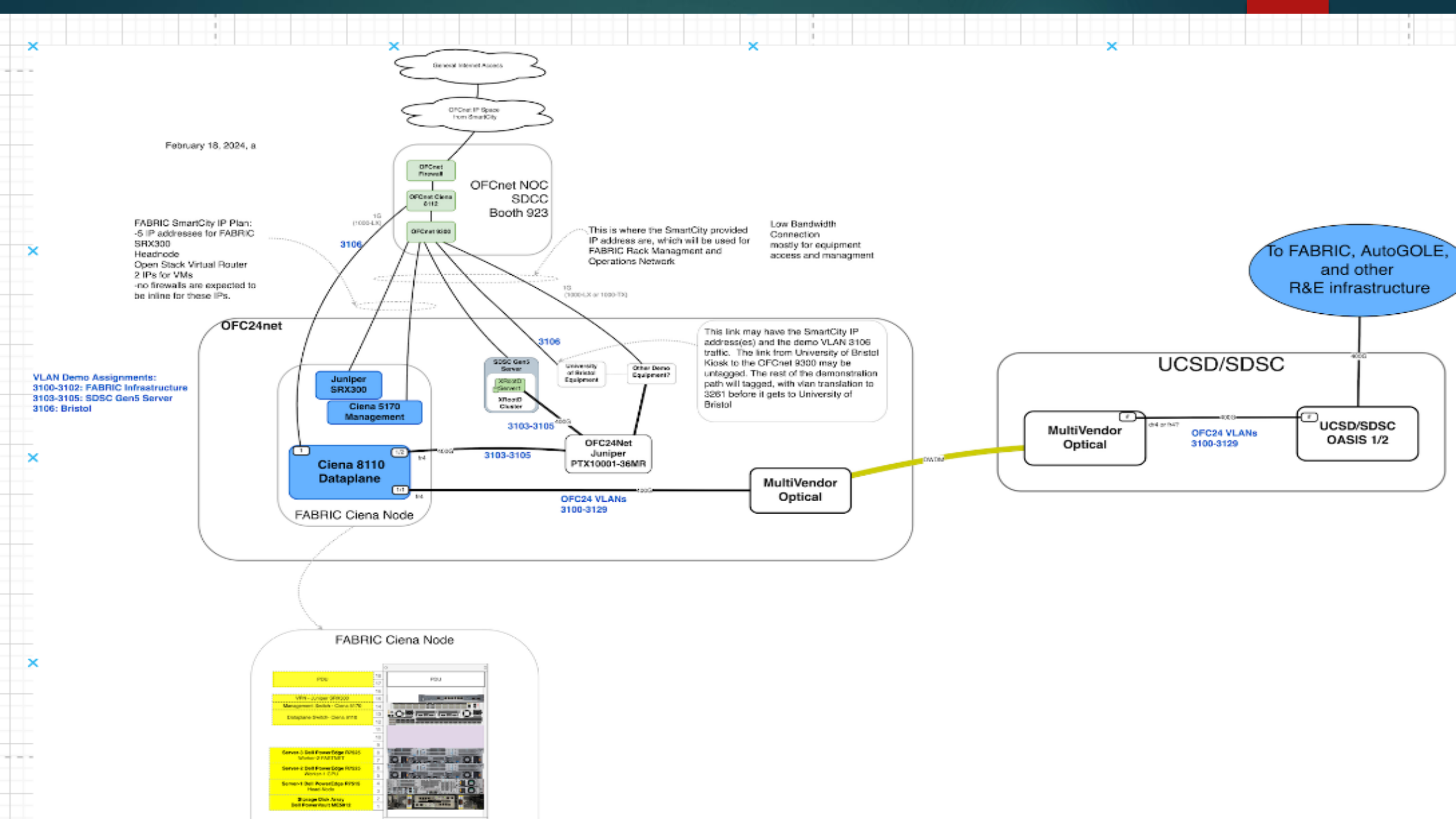


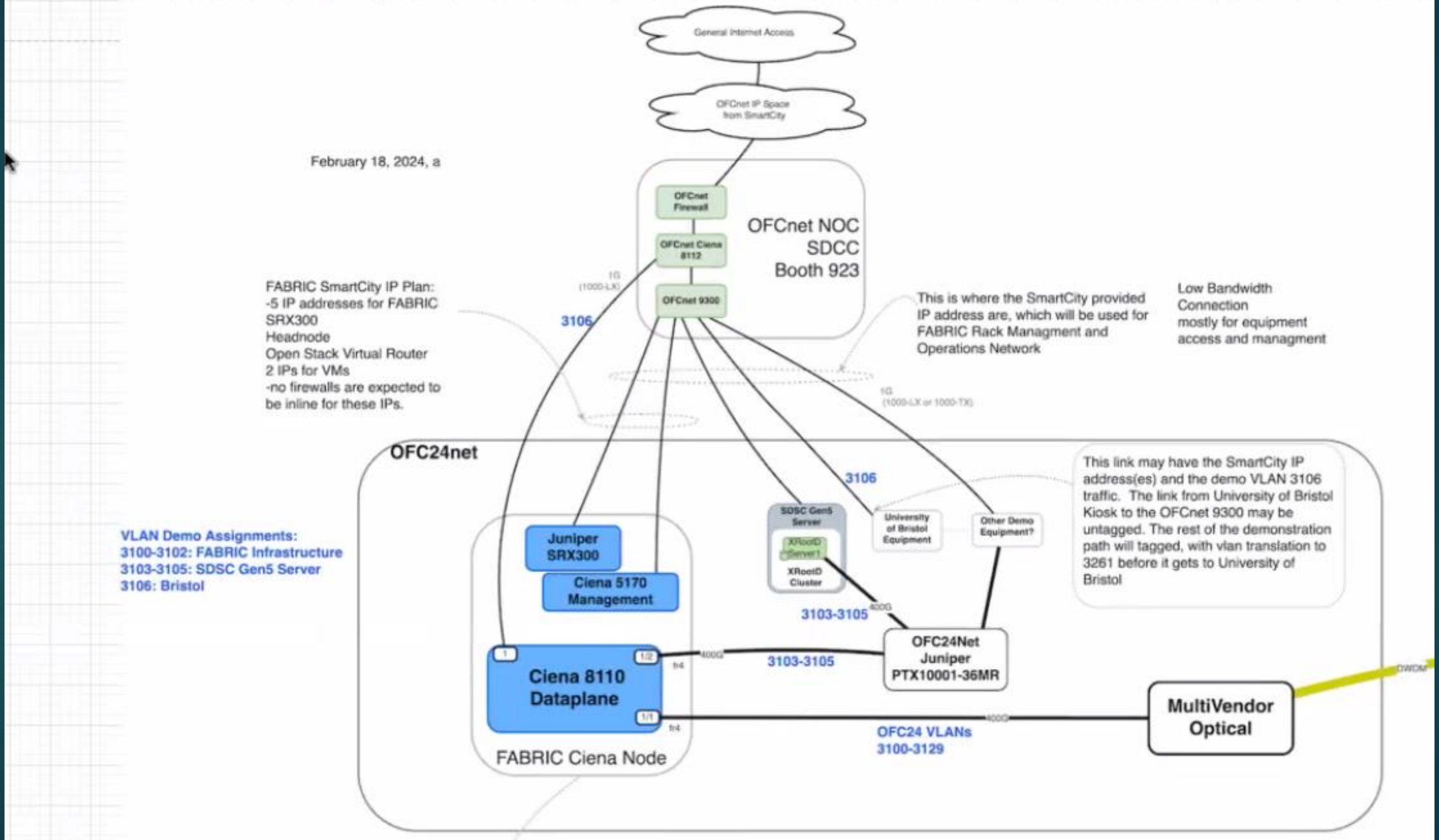
OFC 2024 - OFCnet Architecture


Internet2 DENV
Internet2 KANS
Internet2 CHIC
Internet2 STAR
Internet2 SALT
Internet2 LASV
Internet2 LOSA
StarLight JBDT
FABRIC STAR
FABRIC SALT
FABRIC NEWY
FABRIC LOSA
FABRIC WASH
FABRIC DALL
FABRIC ATLA
FABRIC FIU
FABRIC UCSD
FABRIC OFC24net
Research Arista CENIC LA
UCSD Network
ESnet LOSA
ESnet SAND
ESnet ELPA
ESnet HOUS
ESnet ATLA
ESnet NEWY
ESnet BOST
ESnet CERN
ESnet CHIC
ESnet KANS
ESnet FNALFCC
StarLight JBDT
CALTECH XRootD Cluster
NRP - SDSC XRootD Cluster
OFC24net XRootD Cluster
NRP - StarLight XRootD Cluster
NRP - MICHIGAN XRootD Cluster
NRP - AMPATH XRootD Cluster
NRP - UNL XRootD Cluster
CERN XRootD Cluster
FNAL dCache Cluster
FABRIC - DALL XRootD Server
Link Details
100Gbps
400Gbps
800Gbps
1.2Tbps

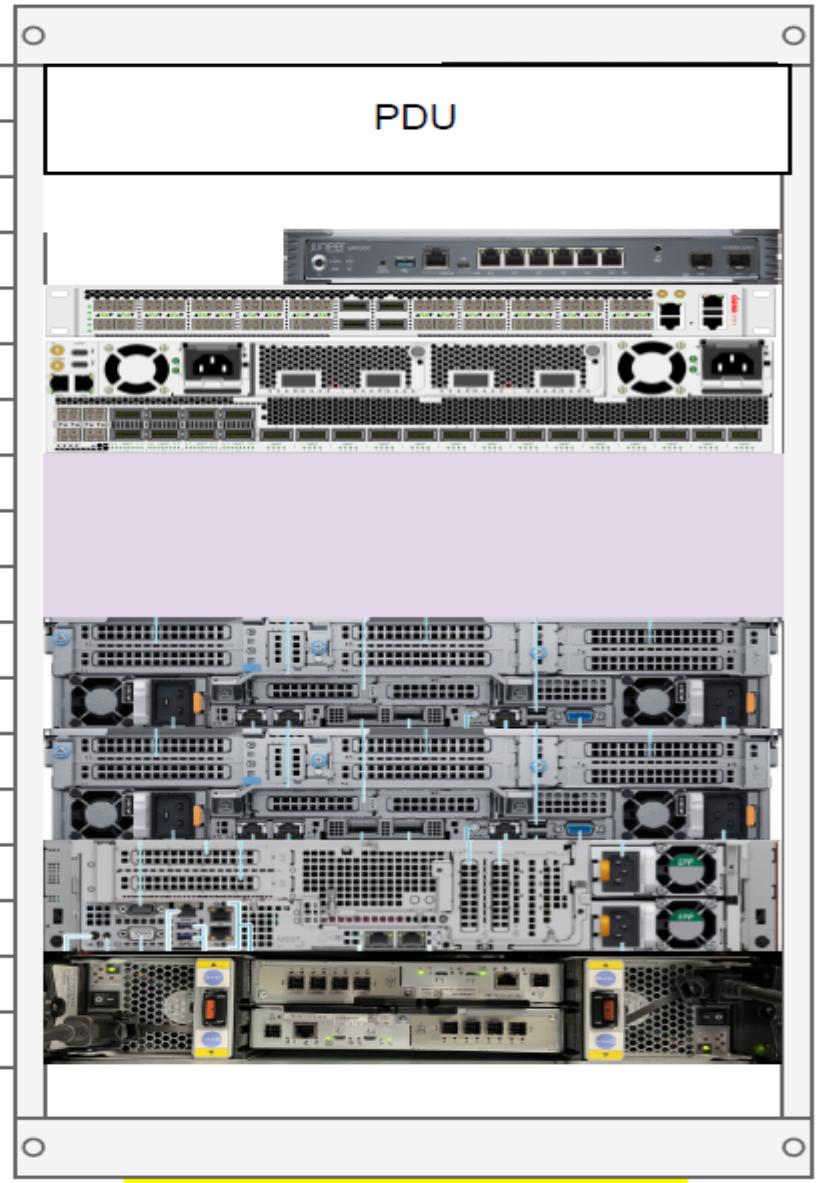
January 25, 2024, a




General Internet Access
OFCnet IP Space from SmartCity
February 18, 2024, a
OFCnet Firewall
OFCnet Ciena 8112
OFCnet 9300
OFCnet NOC
SDCC
Booth 923
FABRIC SmartCity IP Plan:
-5 IP addresses for FABRIC SRX300
Headnode
Open Stack Virtual Router
2 IPs for VMs
-no firewalls are expected to be inline for these IPs.
3106
This is where the SmartCity provided IP address are, which will be used for FABRIC Rack Managment and Operations Network
Low Bandwidth Connection
mostly for equipment access and managment
To FABRIC, AutoGOLE, and other R&E infrastructure
OFC24net
VLAN Demo Assignments:
3100-3102: FABRIC Infrastructure
3103-3105: SDSC Gen5 Server
3106: Bristol
Juniper SRX300
Ciena 5170 Management
SDSC Gen5 Server
University of Bristol Equipment
Other Demo Equipment?
3106
3103-3105
This link may have the SmartCity IP address(es) and the demo VLAN 3106 traffic. The link from University of Bristol Kiosk to the OFCnet 9300 may be untagged. The rest of the demonstration path will tagged, with vlan translation to 3261 before it gets to University of Bristol
Ciena 8110 Dataplane
FABRIC Ciena Node
3103-3105
OFC24Net Juniper PTX10001-36MR
OFC24 VLANs 3100-3129
MultiVendor Optical
UCSD/SDSC
MultiVendor Optical
OFC24 VLANs 3100-3129
UCSD/SDSC OASIS 1/2
FABRIC Ciena Node


General Internet Access
OFCnet IP Space from SmartCity
February 18, 2024, a
OFCnet Firewall
OFCnet Ciena 8112
OFCnet 9300
OFCnet NOC
SDCC
Booth 923
1G (1000-LX)
3106
FABRIC SmartCity IP Plan:
-5 IP addresses for FABRIC SRX300
Headnode
Open Stack Virtual Router
2 IPs for VMs
-no firewalls are expected to be inline for these IPs.
This is where the SmartCity provided IP address are, which will be used for FABRIC Rack Managment and Operations Network
Low Bandwidth Connection
mostly for equipment access and managment
1G (1000-LX or 1000-TX)
OFC24net
3106
This link may have the SmartCity IP address(es) and the demo VLAN 3106 traffic. The link from University of Bristol Kiosk to the OFCnet 9300 may be untagged. The rest of the demonstration path will tagged, with vlan translation to 3261 before it gets to University of Bristol
VLAN Demo Assignments:
3100-3102: FABRIC Infrastructure
3103-3105: SDSC Gen5 Server
3106: Bristol
Juniper SRX300
Ciena 5170 Management
SDSC Gen5 Server
XRootD Server1
XRootD Cluster
University of Bristol Equipment
Other Demo Equipment?
3103-3105
400G
OFC24Net Juniper PTX10001-36MR
Ciena 8110 Dataplane
1/2
1/1
1
fr4
400G
3103-3105
FABRIC Ciena Node
OFC24 VLANs
3100-3129
MultiVendor Optical
DWDM

	PDU	18
		17
		16
Serial: CV0923AN0416	VPN - Juniper SRX300	15
Serial:	Management Switch - Ciena 5170	14
Serial:	Dataplane Switch- Ciena 8110	13
		12
		11
		10
		9
Service Tag: J1FB0R3	**Server-3 Dell PowerEdge R7525** Worker-2 FASTNET	8
		7
Service Tag: DKRT7Y3	**Server-2 Dell PowerEdge R7525** Worker-1 GPU	6
		5
Service Tag: 7MLG7Y3	**Server-1 Dell PowerEdge R7515** Head-Node	4
		3
Service Tag: 51MQ7Y3	**Storage Disk Array** **Dell PowerVault ME5012**	2
		1



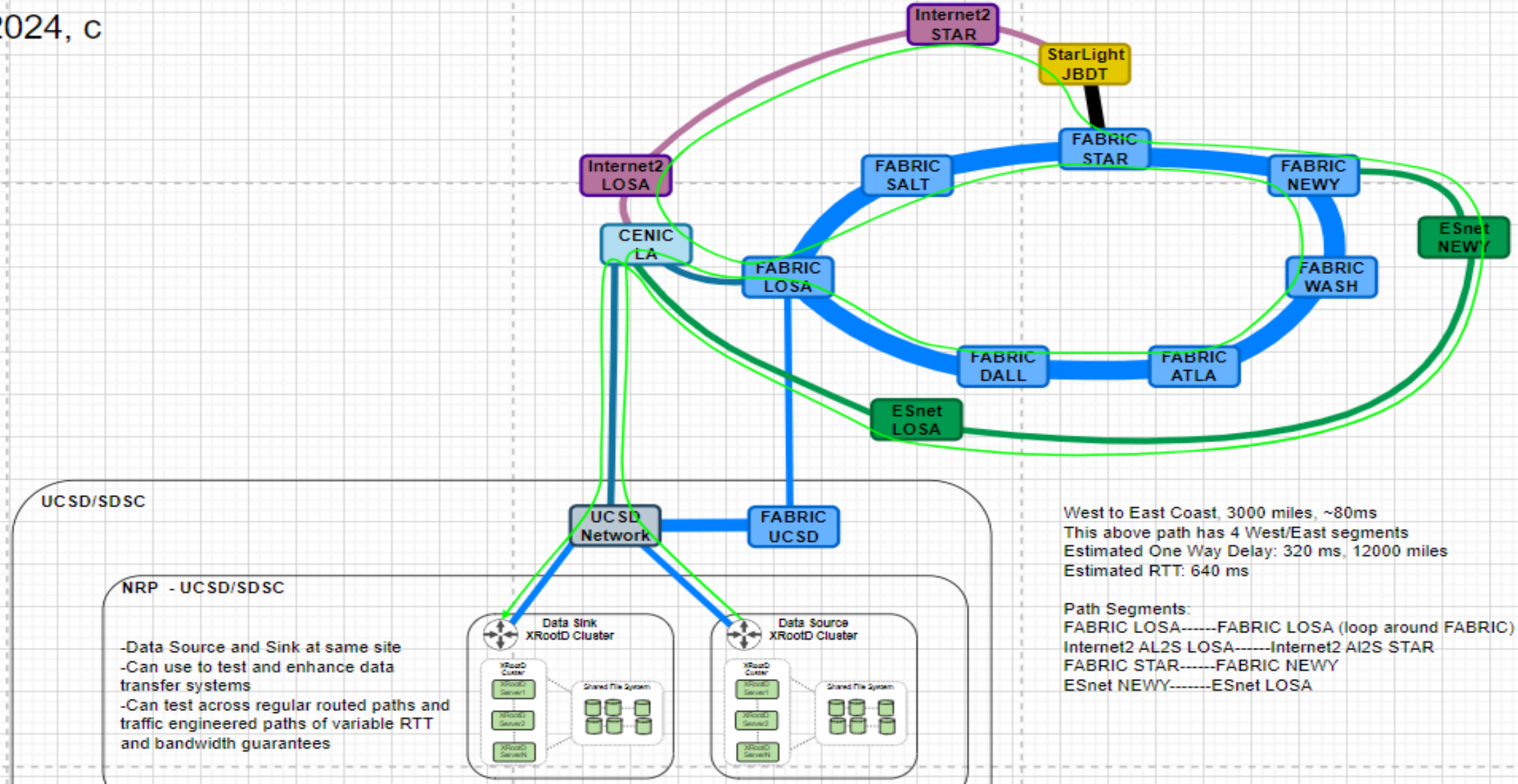
PDU-1 Serial:

PDU-2 Serial:

Rack Enclosure
PCC-18U30
(See Note-1)

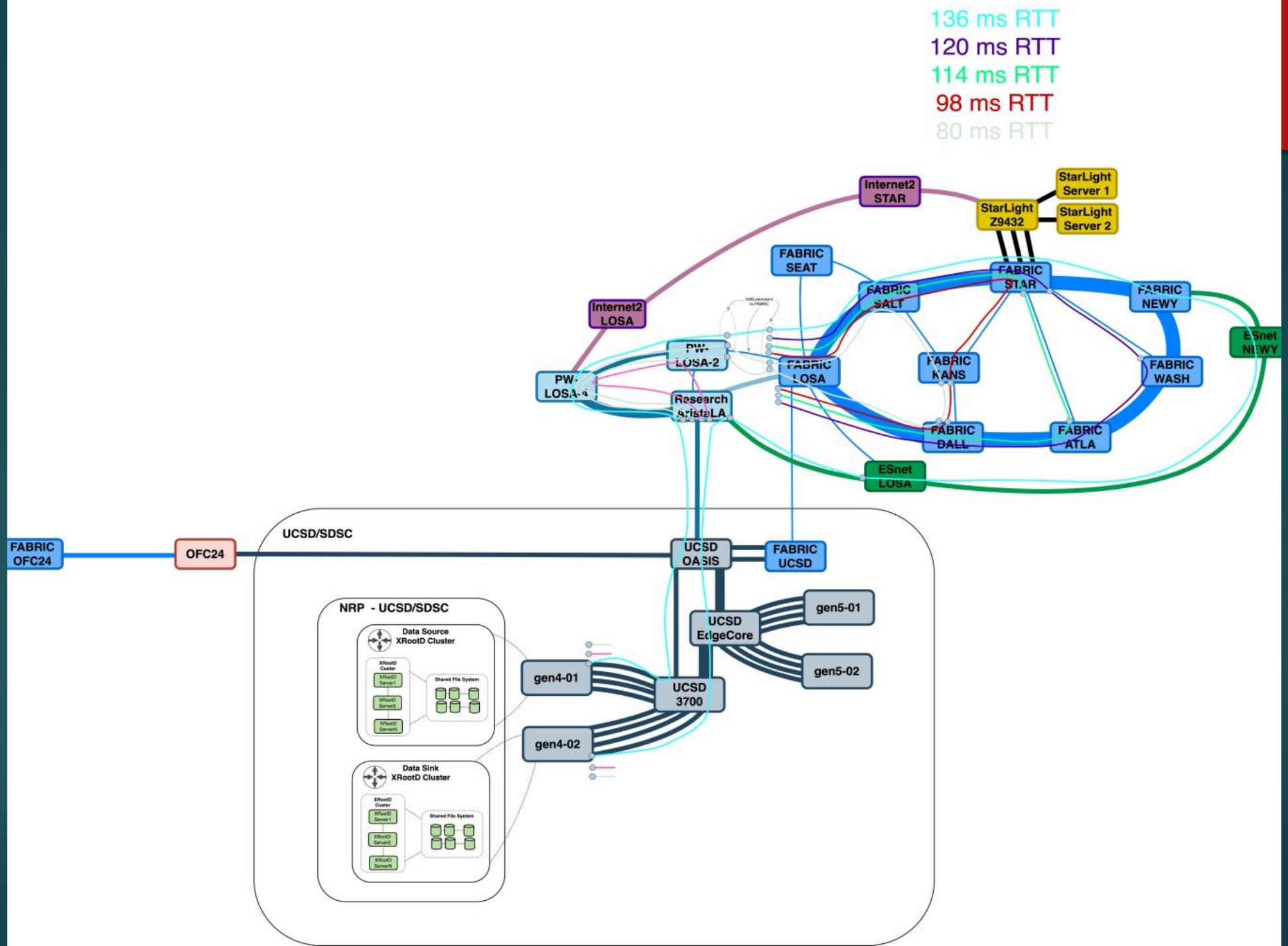
y 25, 2024, c

Example Topology for Testing, Research, Development Data Management and Movement Systems



West to East Coast, 3000 miles, ~80ms
This above path has 4 West/East segments
Estimated One Way Delay: 320 ms, 12000 miles
Estimated RTT: 640 ms

Path Segments:
FABRIC LOSA------FABRIC LOSA (loop around FABRIC)
Internet2 AL2S LOSA------Internet2 AI2S STAR
FABRIC STAR------FABRIC NEWY
ESnet NEWY-------ESnet LOSA

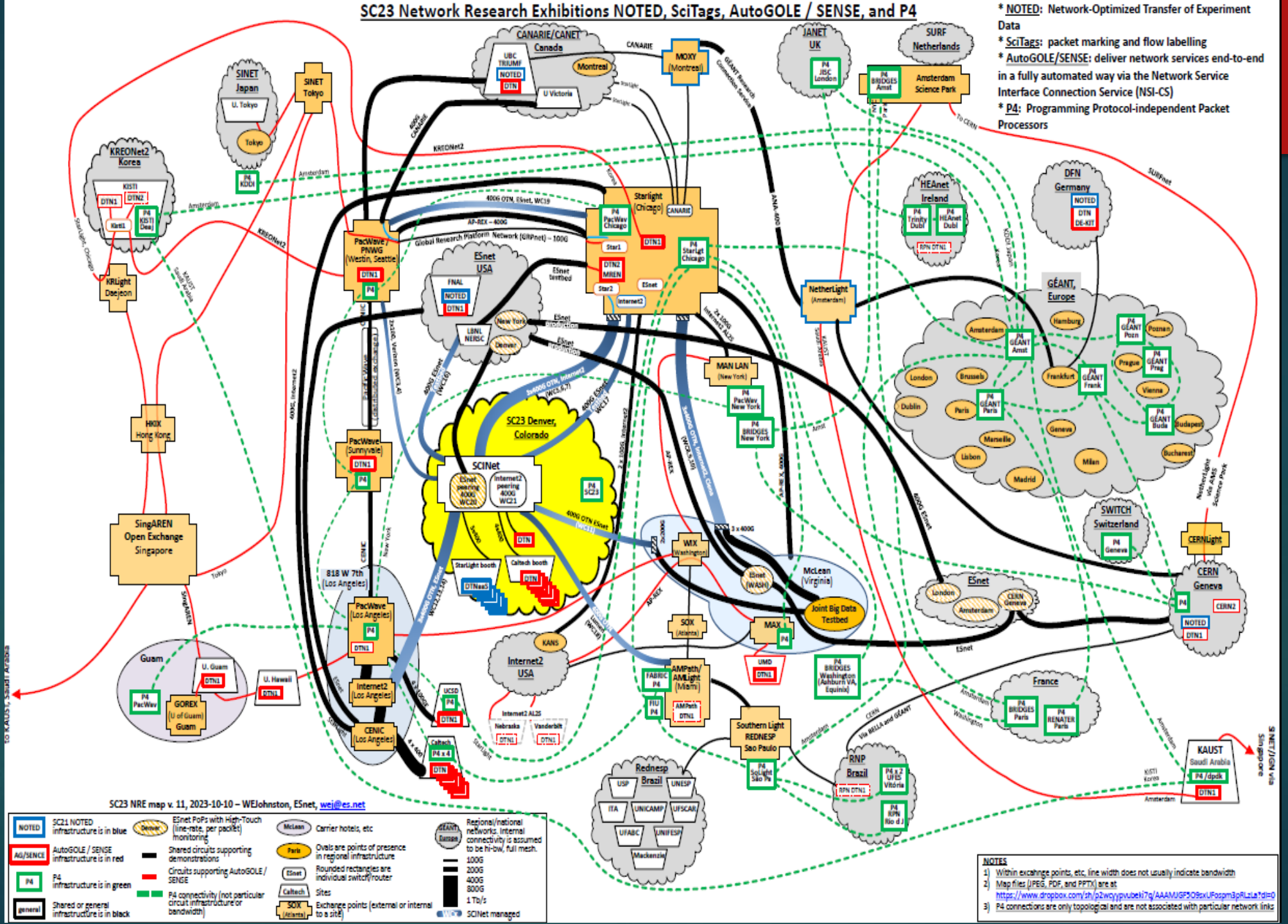

136 ms RTT
120 ms RTT
114 ms RTT
98 ms RTT
80 ms RTT
Internet2 STAR
StarLight Server 1
StarLight Z9432
StarLight Server 2
FABRIC SEAT
FABRIC STAR
FABRIC NEWY
Internet2 LOSA
FABRIC WASH
ESnet NEWY
FABRIC KANS
FABRIC LOSA
FABRIC DALL
FABRIC ATLA
ESnet LOSA
FABRIC OFC24
OFC24
UCSD/SDSC
UCSD OASIS
FABRIC UCSD
NRP - UCSD/SDSC
Data Source XRootD Cluster
Data Sink XRootD Cluster
UCSD EdgeCore
gen5-01
gen5-02
gen4-01
gen4-02
UCSD 3700

Example SC24 SCinet Network Research Exhibitions

- Global Research Platform (GRP)
- SDX 1.2 Tbps WAN Services
- SDX E2E 400 Gbps WAN Services
- 400 Gbps DTNs & Smart NICs
- Network Optimized Transport for Experimental Data (NOTED) – With AI/ML Driven WAN Network Orchestration
- SDX International Testbed Integration
- StarLight SDX for Petascale Science
- DTN-as-a-Service For Data Intensive Science
- P4 Integration With Kubernetes
- PetaTrans Services Based on NVMe-Over-Fabric
- NASA Goddard Space Flight Center HP WAN Transport Services
- Resilient Distributed Processing & Rapid Data Transfer
- PRP/NRP Demonstrations
- Open Science Grid Demonstrations
- N-DISE Named Data Networking for Data Intensive Science
- Orchestration With Packet Marking (SciTags)
- Data Tsunami

SC23 Network Research Exhibitions NOTED, SciTags, AutoGOLE / SENSE, and P4

* NOTED: Network-Optimized Transfer of Experiment Data
* SciTags: packet marking and flow labelling
* AutoGOLE/SENSE: deliver network services end-to-end in a fully automated way via the Network Service Interface Connection Service (NSI-CS)
* P4: Programming Protocol-independent Packet Processors

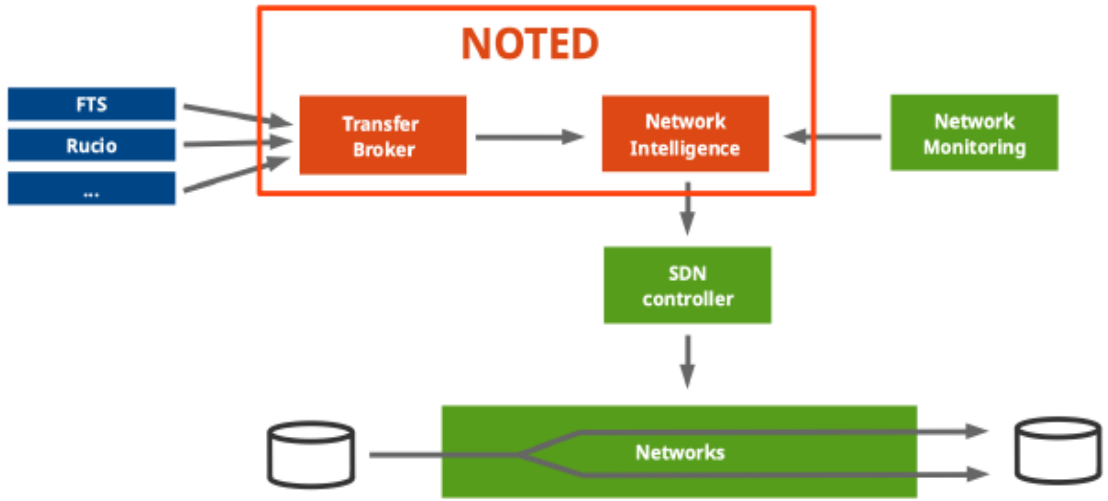


SC23 NRE map v. 11, 2023-10-10 – WEJohnston, ESnet, wej@es.net

NOTED — SC21 NOTED infrastructure is in blue
AG/SENCE — AutoGOLE / SENSE infrastructure is in red
P4 — P4 infrastructure is in green
general — Shared or general infrastructure is in black
Denver — ESnet PoPs with High-Touch (line-rate, per packet) monitoring
Shared circuits supporting demonstrations
Circuits supporting AutoGOLE / SENSE
P4 connectivity (not particular circuit infrastructure or bandwidth)
McLean — Carrier hotels, etc
Paris — Ovals are points of presence in regional infrastructure
ESnet — Rounded rectangles are individual switch/router
Caltech — Sites
SOX (Atlanta) — Exchange points (external or internal to a site)
GÉANT Europe — Regional/national networks. Internal connectivity is assumed to be hi-bw, full mesh.
100G
200G
400G
800G
1 Tb/s
SCINet managed

NOTES
1) Within exchange points, etc, line width does not usually indicate bandwidth
2) Map files (JPEG, PDF, and PPTX) are at https://www.dropbox.com/sh/p2wcyypvubeki7q/AAAMJGP5O8xUFospm3pRLzLaTdl=0
3) P4 connections are only topological and are not associated with particular network links

SKELETON AND ELEMENTS OF NOTED



FTS (File Transfer Service):

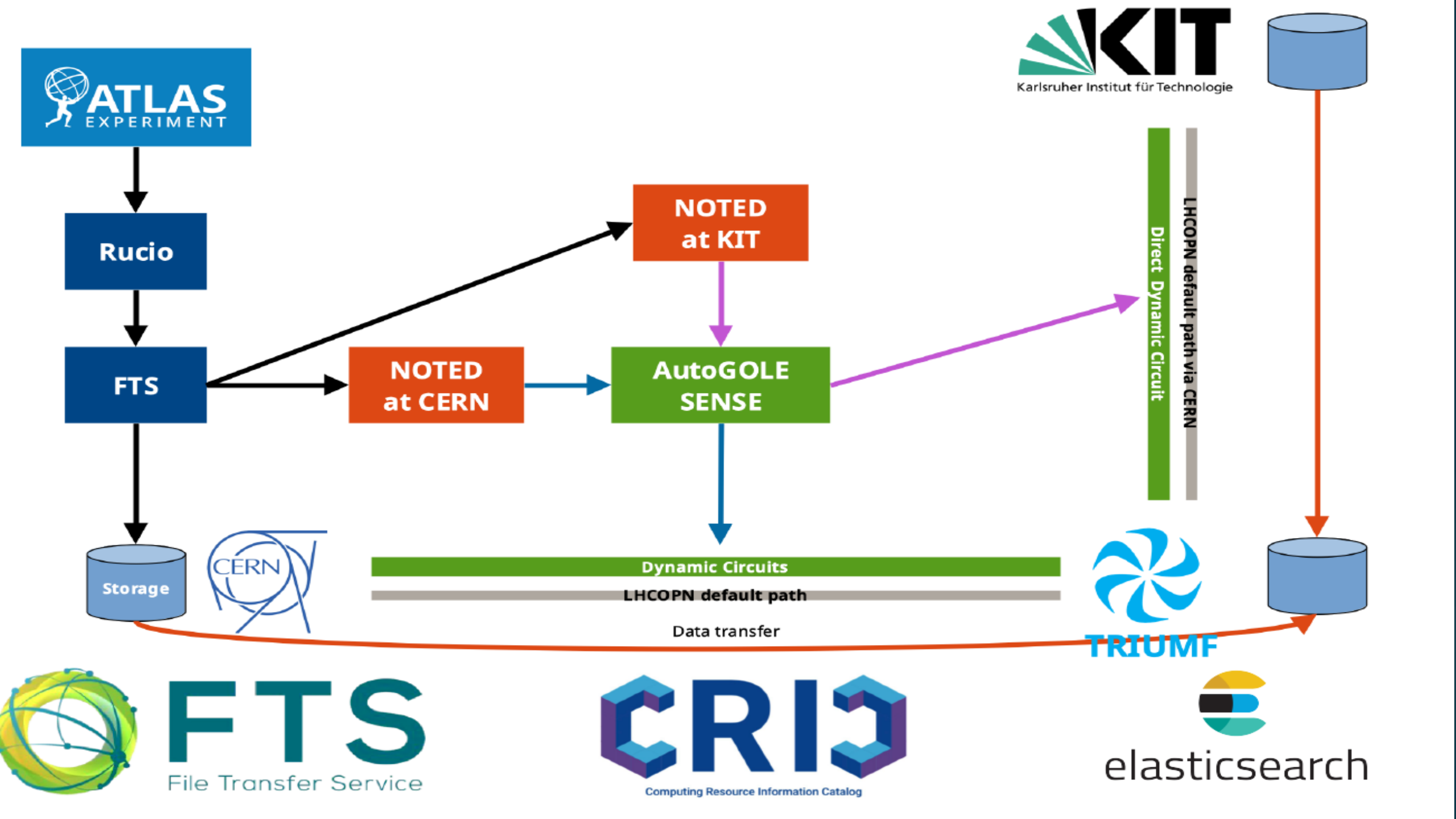
- Inspect and analyse data transfers to estimate if an action can be applied to optimise the network utilization → get on-going and queued transfers.

CRIC (Computing Resource Information Catalog):

- Enrichment to get an overview and knowledge of the network topology → get IPv4/IPv6 addresses, endpoints, rcsite and federation.

FLOWCHART AND DATASET STRUCTURE

- Input parameters: configuration given by the user
  - In noted/config/config.yaml → define a list of {src_rcsite, dst_rcsite}, maximum and minimum throughput threshold, SENSE/AutoGOLE VLANs UUID and user-defined email notification among others.
- Enrich NOTED with the topology of the network:
  - Query CRIC database → get endpoints that could be involved in the data transfers for the given {src_rcsite, dst_rcsite} pairs.
- Analyse on-going and upcoming data transfers:
  - Query FTS recursively → get on-going data transfers for each set of source and destination endpoints.
  - The total utilization of the network is the sum of on-going and upcoming individual data transfers for each source and destination endpoints for the given {src_rcsite, dst_rcsite} pairs.
- Network decision:
  - If NOTED interprets that the link will be congested → provides a dynamic circuit via SENSE/AutoGOLE.
  - If NOTED interprets that the link will not be be congested anymore → cancel the dynamic circuit and the traffic is routed back.


ATLAS EXPERIMENT
Rucio
FTS
NOTED at KIT
NOTED at CERN
AutoGOLE SENSE
KIT
Karlsruher Institut für Technologie
Direct Dynamic Circuit
LHCOPN default path via CERN
Storage
CERN
Dynamic Circuits
LHCOPN default path
Data transfer
TRIUMF
FTS
File Transfer Service
CRIC
Computing Resource Information Catalog
elasticsearch

Scitags Initiative

Leads= Shawn McKee, Marian Babik

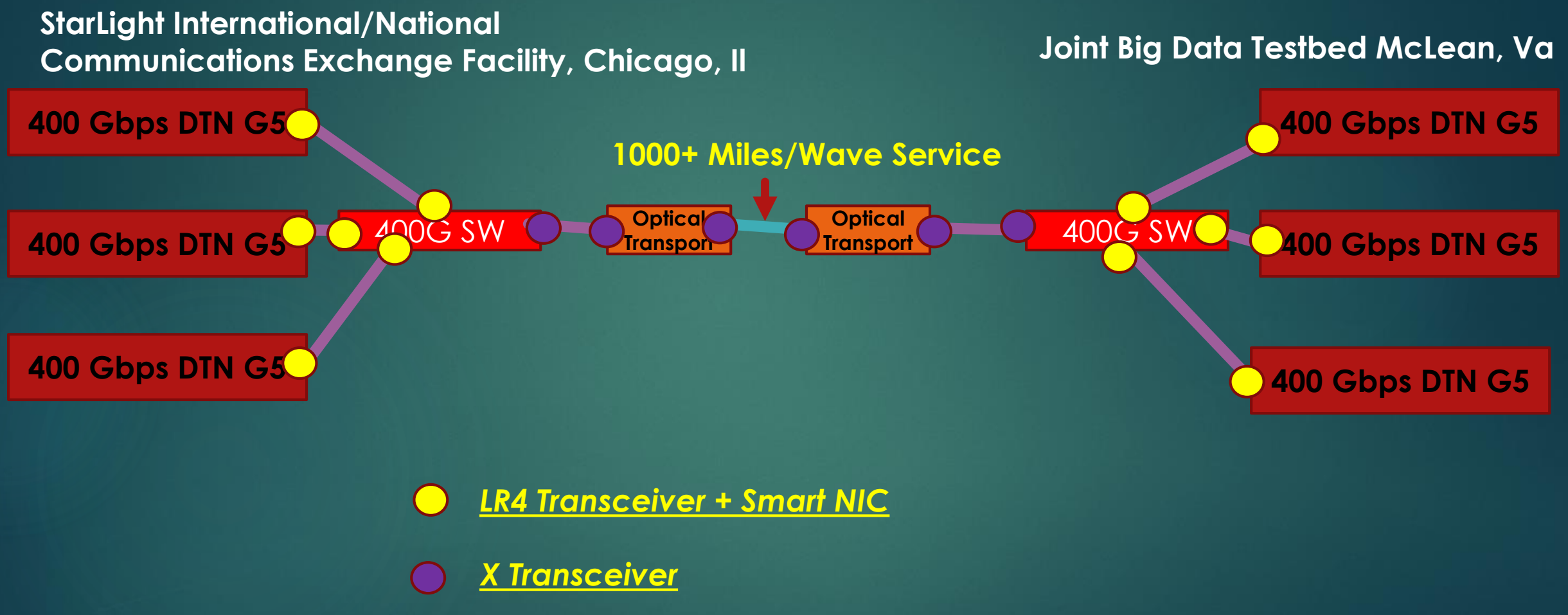
- Scientific Network Tags (scitags) is an initiative promoting identification of the science domains and their high-level activities at the network level.

- Enable tracking and correlation of our transfers with Research and Education Network Providers (R&Es) network flow monitoring
- Experiments can better understand how their network flows perform along the path
  - Improve visibility into how network flows perform (per activity) within R&E segments
  - Get insights into how experiment is using the networks, get additional data from R&Es on behaviour of our transfers (traffic, paths, etc.)
- Sites can get visibility into how different network flows perform
  - Network monitoring per flow (with experiment/activity information)
    - E.g. RTT, retransmits, segment size, congestion window, etc. all per flow

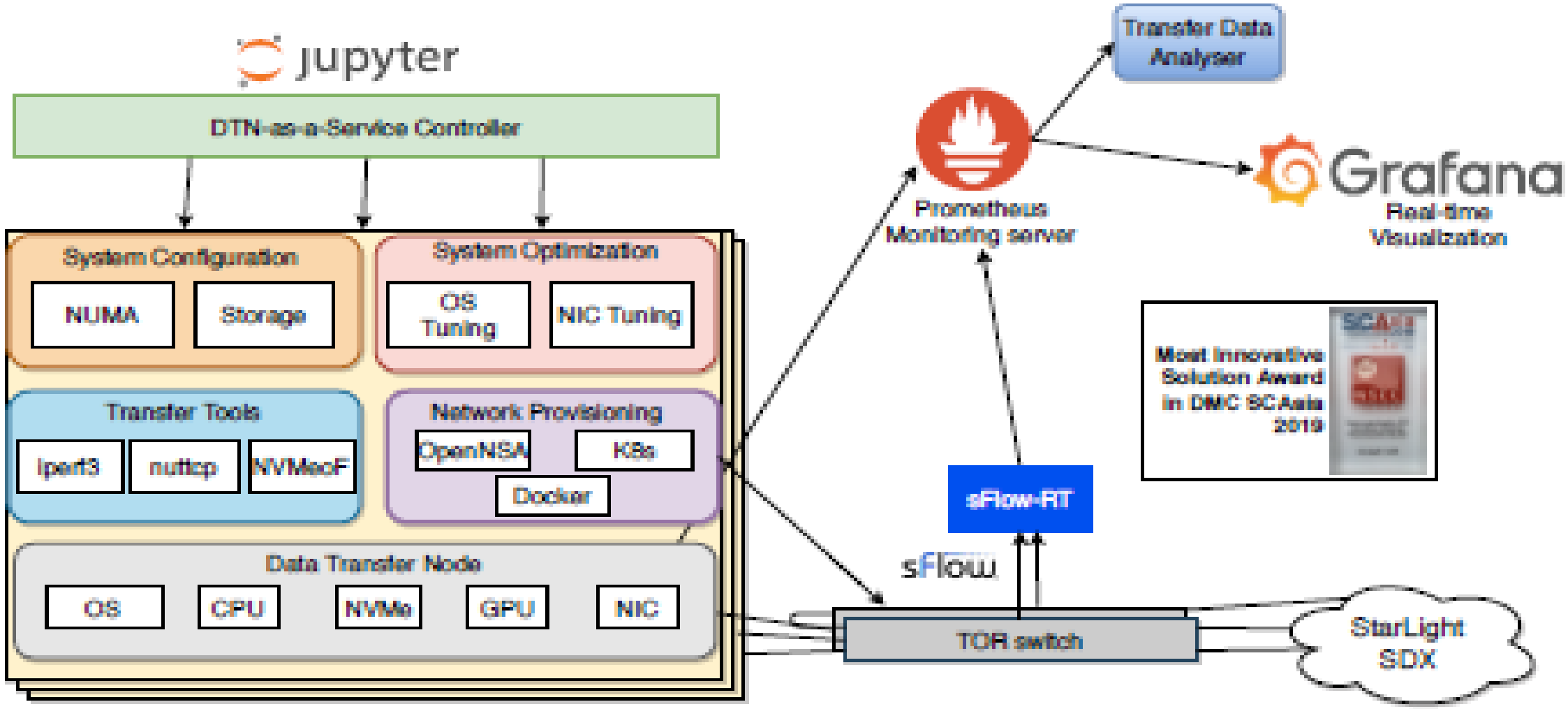
SC24 Packet/Flow Marking NRE

- **Concept: The Goals of the SC24 Packet and Flow Marking NRE Demonstrations Will Build On the SC23 Demonstrations To Showcase The Capabilities of The Scitags Architecture And Methods For Optimizing Data Intensive Science**
- **Five Demonstrations Will Be Staged**
  - **IPv6 Packet Marking With eBPF-TC (100 Gbps)**
  - **XRootD Packet Marking with Flowd+eBPF-TC**
  - **Accounting For Flow Labeled Packets Using a P4 Programmable Switch**
  - **Measurements via Esnet High-Touch Processes**
  - **Scitags Integration With DTN-as-a-Service.**
- **Participants:**
  - **CERN, University of Victoria, KIT, ESnet, StarLight, CANARIE, Fermi National Accelerator Laboratory, SCInet, Digital Alliance, etc**

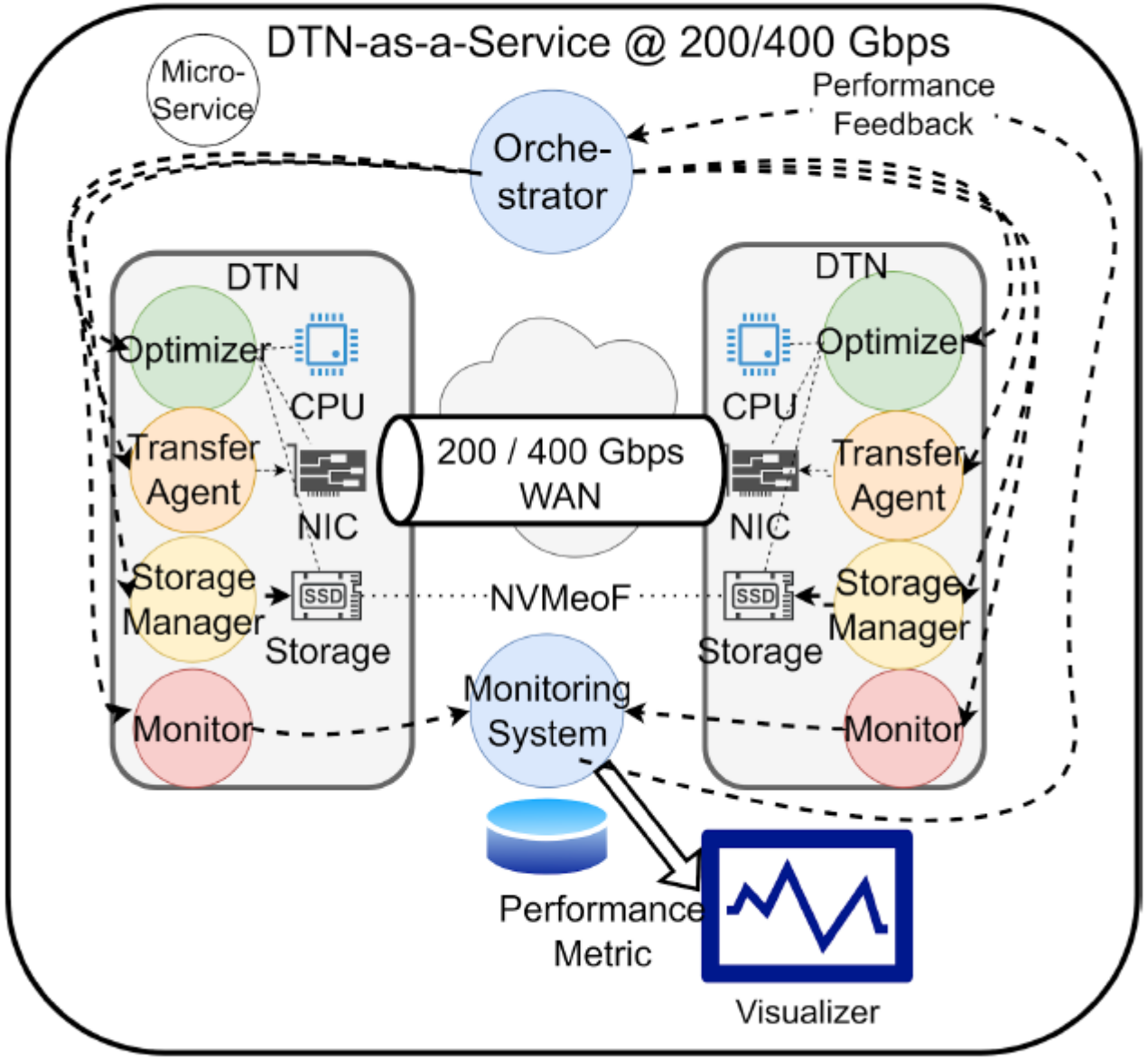
1.2 Tbps WAN Service Prototype for Data Intensive Science



STARLIGHT SDX DTN-as-a-Service



200/400 Gbps DTN-as-a-Service in High-Performance Research Platform



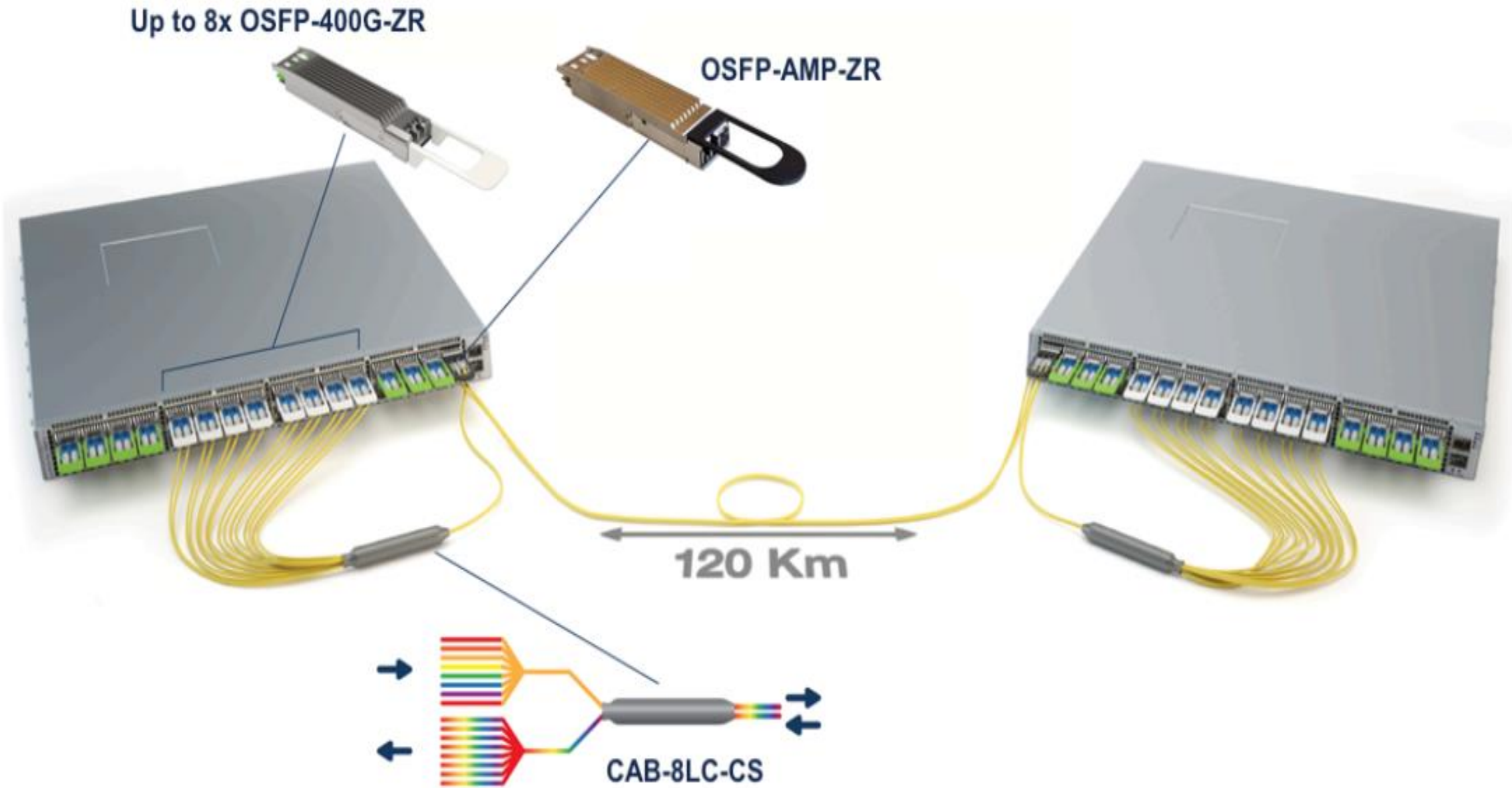
- 200/400 Gbps end-to-end high-performance data transfer over WAN
- DTN-as-a-Service with microservice architecture, optimizing and transferring using containers
- NVMeoF with streaming support
- Performance monitoring and visualization using opensource platforms (Prometheus, Grafana, and sFlow)

iCAIR

STARLIGHT SDX

Peripheral Component Interconnect Express (PCIe)

- PCIe 5.0 server with 32 Lanes = 128GB/s bandwidth 32 GHz
- 3.94 Gbps/Lane
- For Optimization Requires G5 Components
- Foundation For Non-Volatile Memory Express (NVMe) Interfaces, (e.g., SSDs)
- PCIe 6.0 = 2025-2026


Up to 8x OSFP-400G-ZR
OSFP-AMP-ZR
120 Km
CAB-8LC-CS

WAN Optical Transport

- Ciena WaveServer 100GbE and 400 GbE (Soon 800 GbE) applications, scaling capabilities to 12.8 Tb/s client and 12.8 Tb/s line capacity in 2RU

Thanks!

- Questions?