SCinet DTN-as-a-Service Framework

Se-Young Yu, Jim Chen, Joe Mambretti, Fei Yeh, Xiao Wang, Anna Giannakou, Eric Pouyoul, Marc Lyonnais

Data Intensive Science Trends

- Large scale, data (and compute) intensive sciences encounter technology challenges before other domains
- In addition to the network performance, data movement performance is critical to science research infrastructure
- Automated frameworks are required to setup, optimize and analyze the performance of data movement
- Resources must be shared over a network with flexible and interactive workflow management
- DTN-as-a-Service provides a framework to integrate all above

3 years of SCinet X-NET + NRE project

- SC17 : Data Transfer Node Service in SCinet
- SC18: SCinet Multi 100G Data Transfer Node for Multi-Tenant Production Environment
- **SC19: Toward SCinet DTN-as-a-Service**
- Plan to establish as a standard SCinet service for SC20 and beyond

iCAIR

STARLIGHT SM SDX

Any machine can be a DTN

If it has:

1. Clean network connection
2. High-performance storage to match network throughput

We will:

1. **Optimize** hardware, OS and software
2. Select and integrate file transfer protocol
3. Map DTN environment with **science workflow**
4. Virtualize it

iCAIR

STARLIGHT SDX

Data Transfer Nodes Challenges

Virtualization - There are already virtualized environments and orchestrators but configuring containers and VMs with network, storage and CPU is difficult

Workflow mapping - Mapping transfer workflows to the science workflow

Performance tuning - tuning and testing DTN instances

Evaluation - monitoring and reconfiguring parameters

iCAIR

STARLIGHT SM SDX

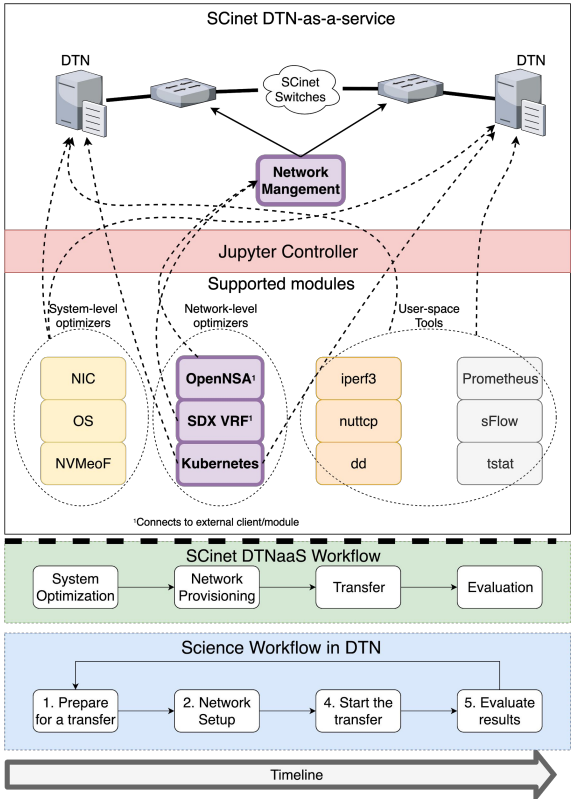
SC19 SCinet DTN-as-a-Service Framework

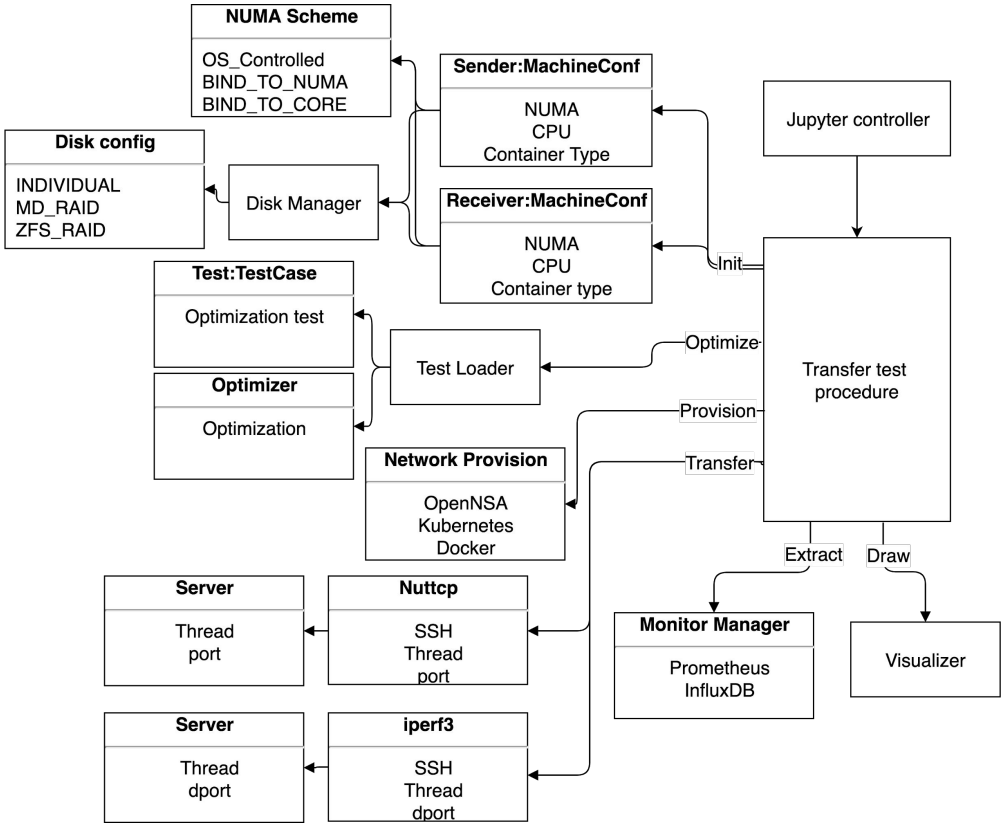
- Provides **Data Transfer Node software and hardware platform** as prototype service to support SC19 SCinet community before and during the SC conference
- Supports testing, demonstration, experimentation, evaluation and other SC and SCinet related activities, especially those for **data intensive science**
- For SC19, new prototype services include: **kubernetes, NVMeoF** and **400G LAN/WAN** experiments
- Provides **workflow** to support integrating new network technology and new data movement technology with minimum operational impact

iCAIR

STARLIGHT SM SDX

Design and implementation





DTN-as-a-Service Modules

DaaS provides an environment for high-speed transfer

Performance testing tools with an optimized environment

Provides API for storage and NUMA configuration

Incremental support for transfer protocols

NVMe over Fabrics support

Containerized for docker and K8s support

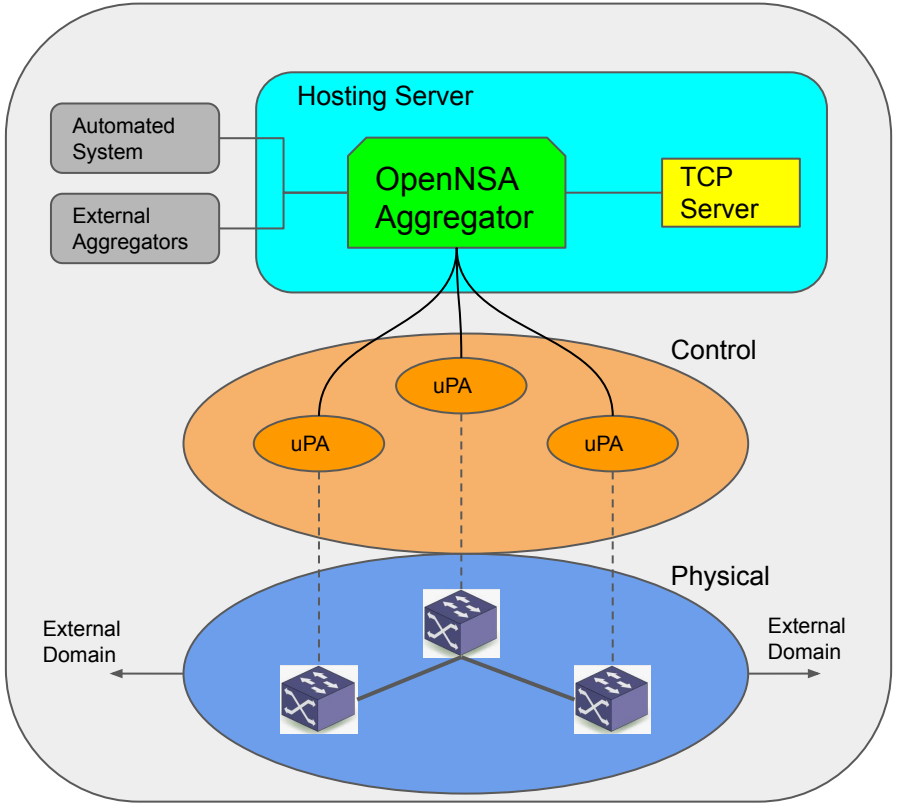
iCAIR

STARLIGHT SM SDX

OpenNSA

The OpenNSA is an open-source implementation of the NSI protocol developed by NORDUnet, GEANT and other contributors.

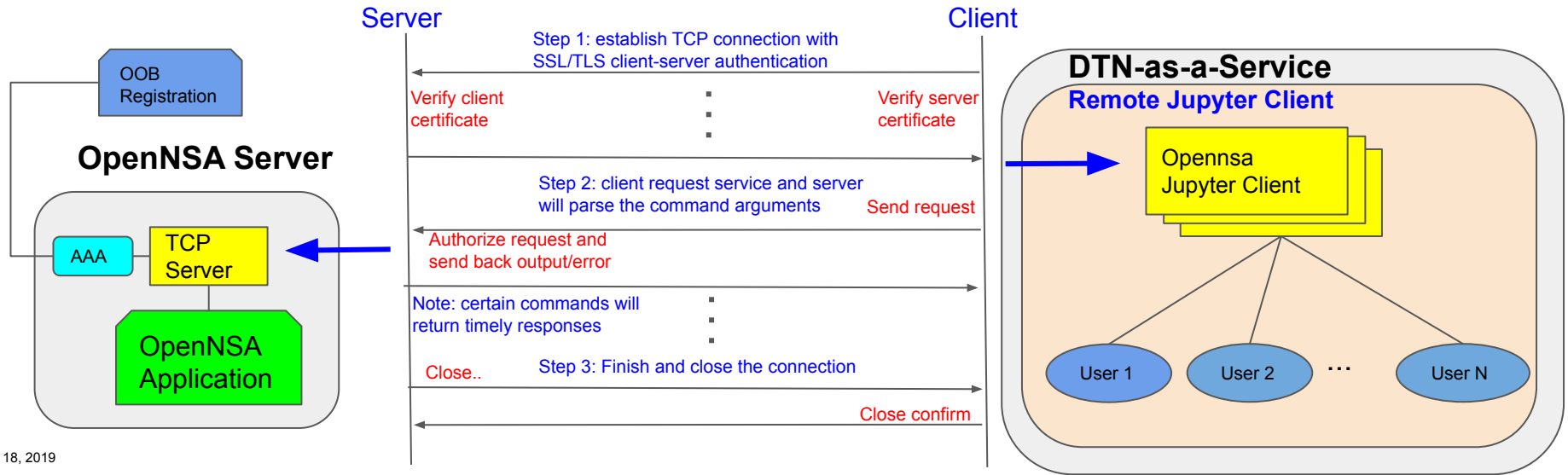
- Manage uPA instances
- Peer with external aggregators.
- Support multi-vendor backends



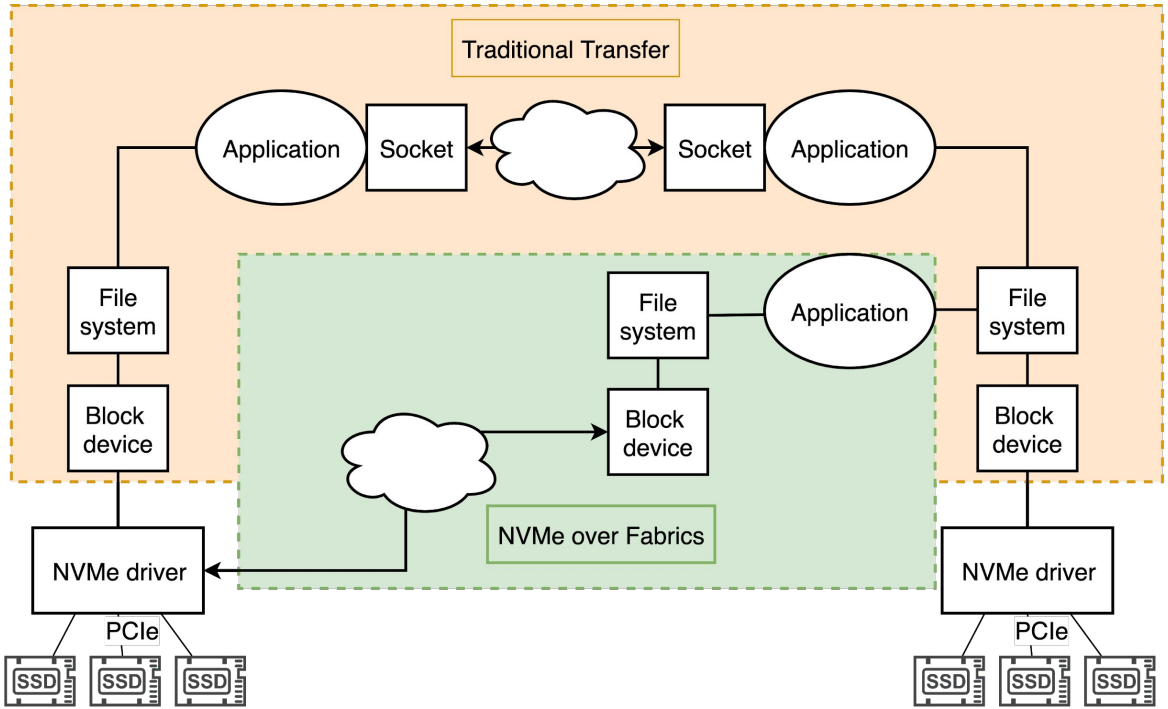
Jupyter Client for NSI OpenNSA Integration

Securely allow users to run NSI OpenNSA services(i.e. dynamically stitch layer 2/3 circuits based on technologies e.g. VLAN.)

- Features: **Authentication, Authorization, Accounting**
- **Authentication**: SSL/TLS operation authenticate both server and client. (certificates need to be exchanged prior unless using public certificates)
- **Authorization**: server will allow access users to request services on certain ports/VLANs based on user identification. I.e. Request command arguments will be parsed and authorized if allowed.
- **Accounting**: for future requirement.
- Additional Feature:Asynchronous(non-blocking operation allows multiple users to request services simultaneously)



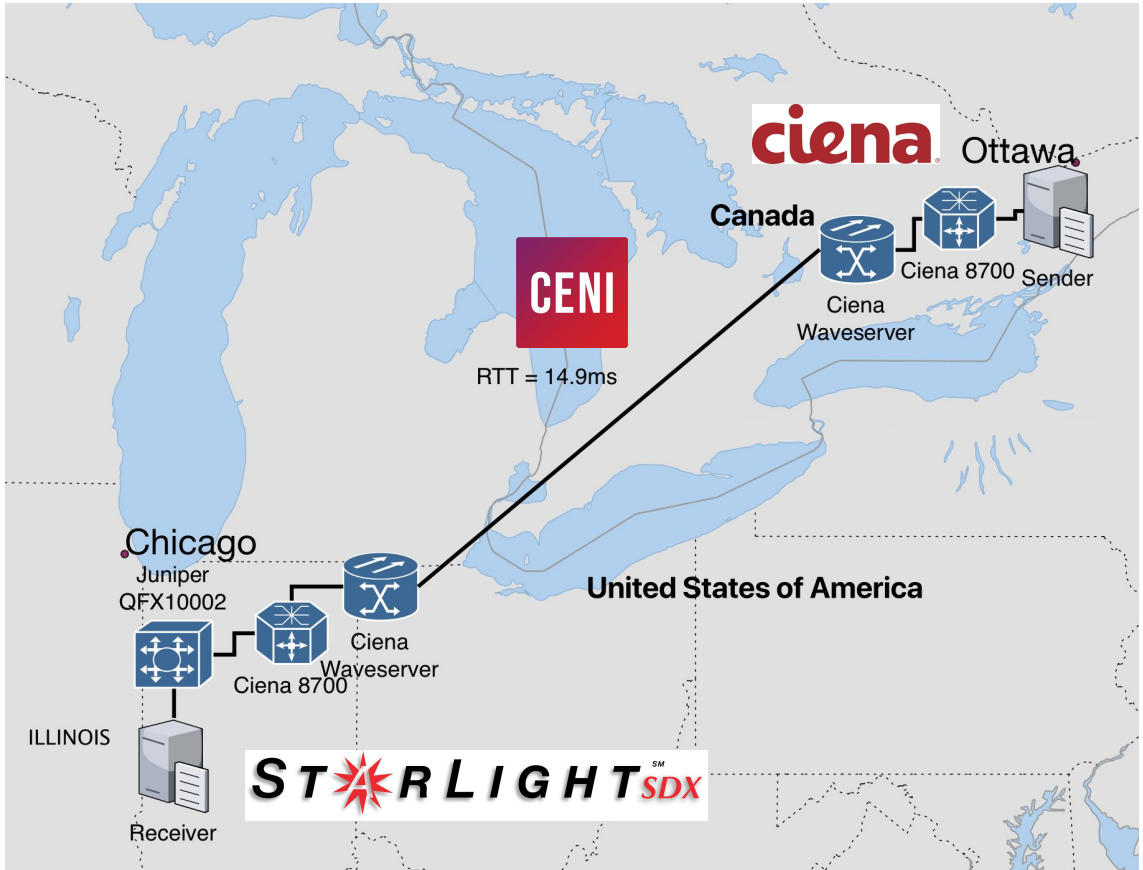
NVMe over Fabrics Overview



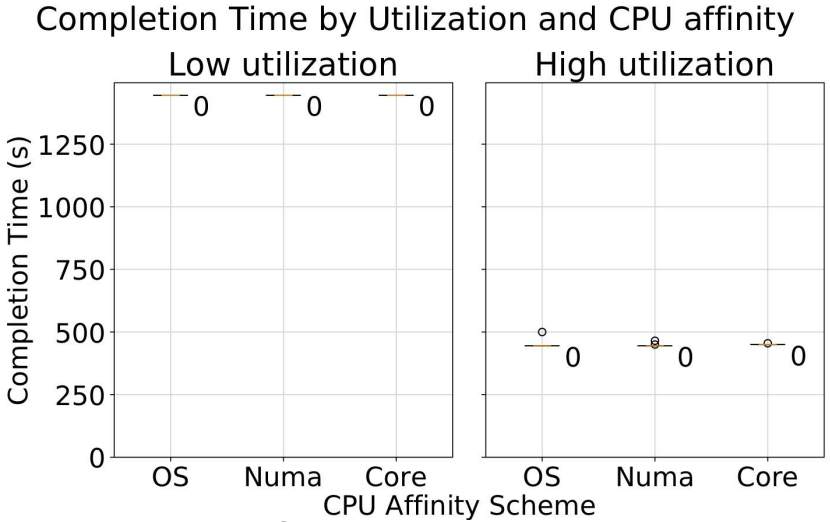
iCAIR

STARLIGHT SM SDX

Evaluation - WAN Transfer

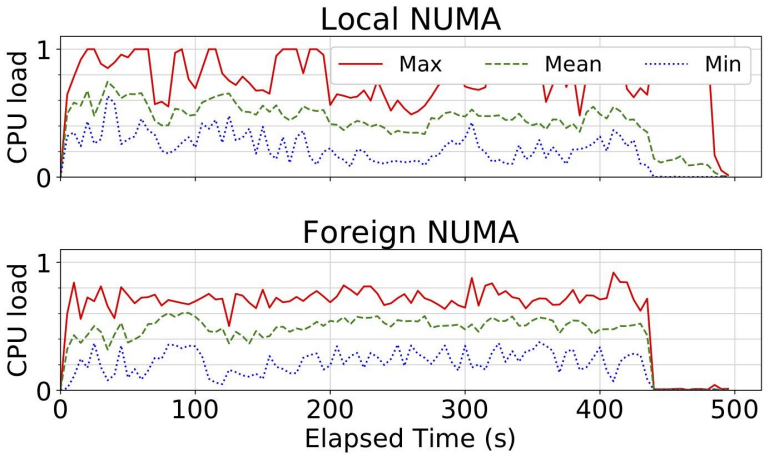


Evaluation - NVMe to NVMe WAN transfer (TCP transfer)



4TB data transfer

Average completion time between 400-1400s

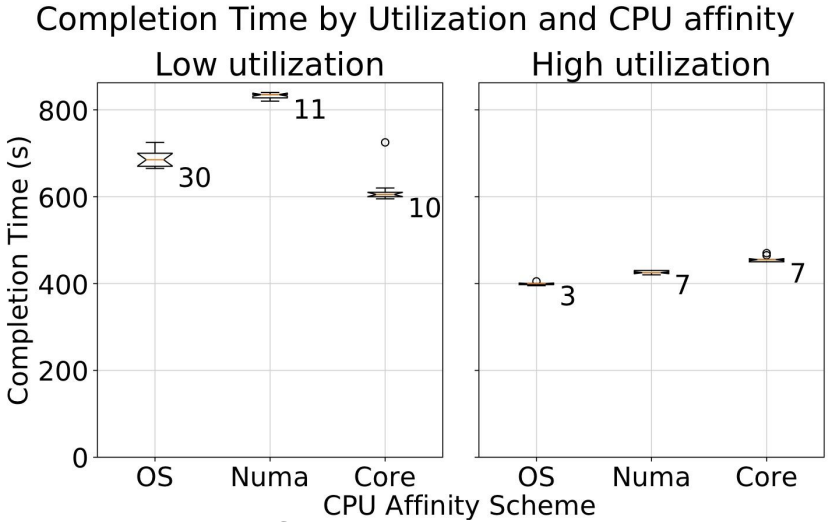


Uneven work distribution between cores

iCAIR

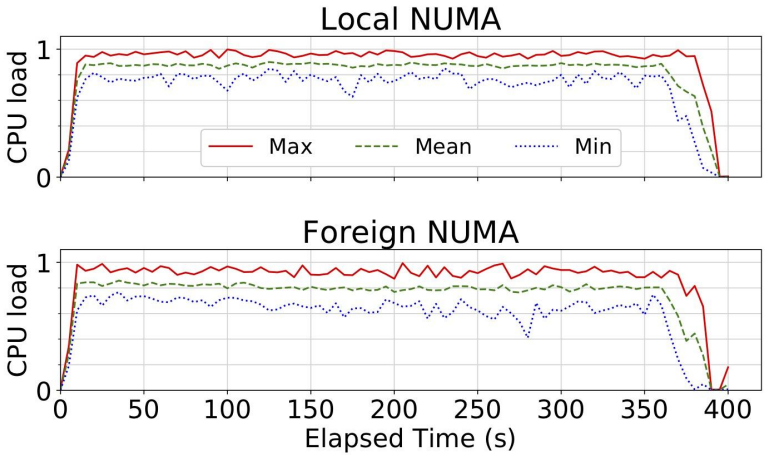
STARLIGHT SDX

Evaluation - NVMe to NVMe WAN transfer (NVMeoF/TCP)



4TB data transfer

Average completion time between 400-800s

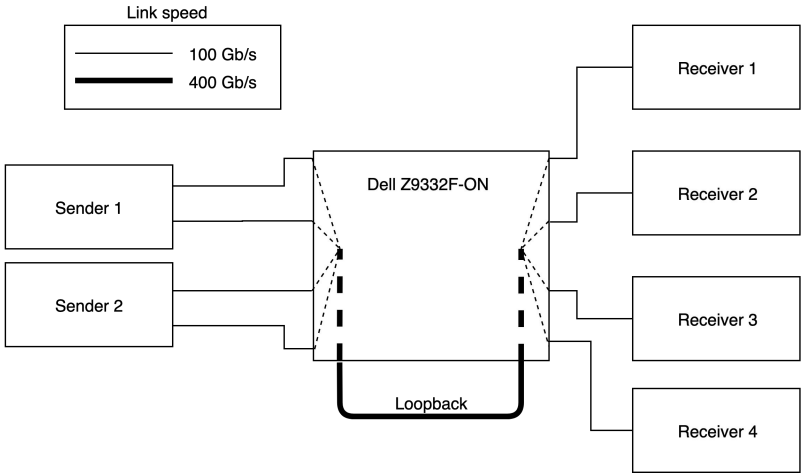


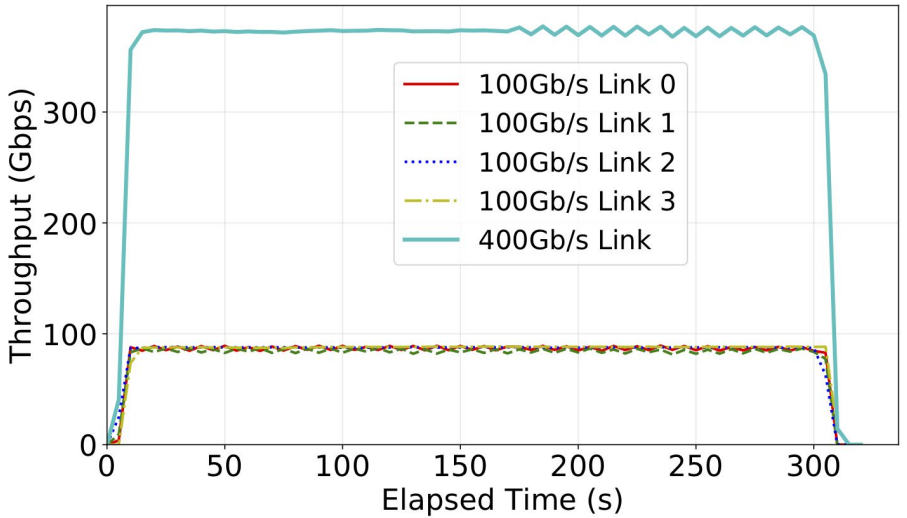
Better work distribution between cores

iCAIR

STARLIGHT SDX

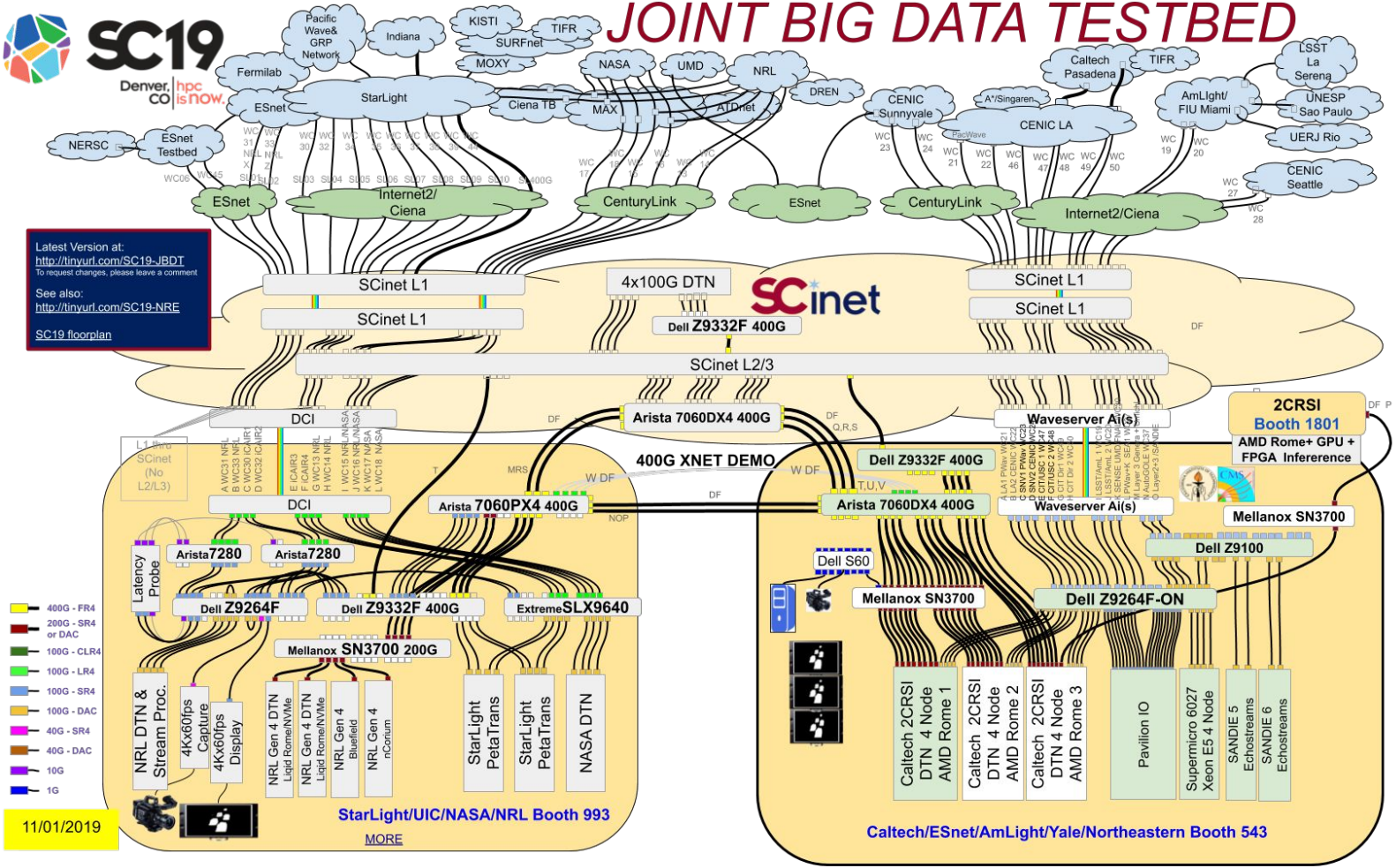
Evaluation - Optimization for 400Gb/s LAN



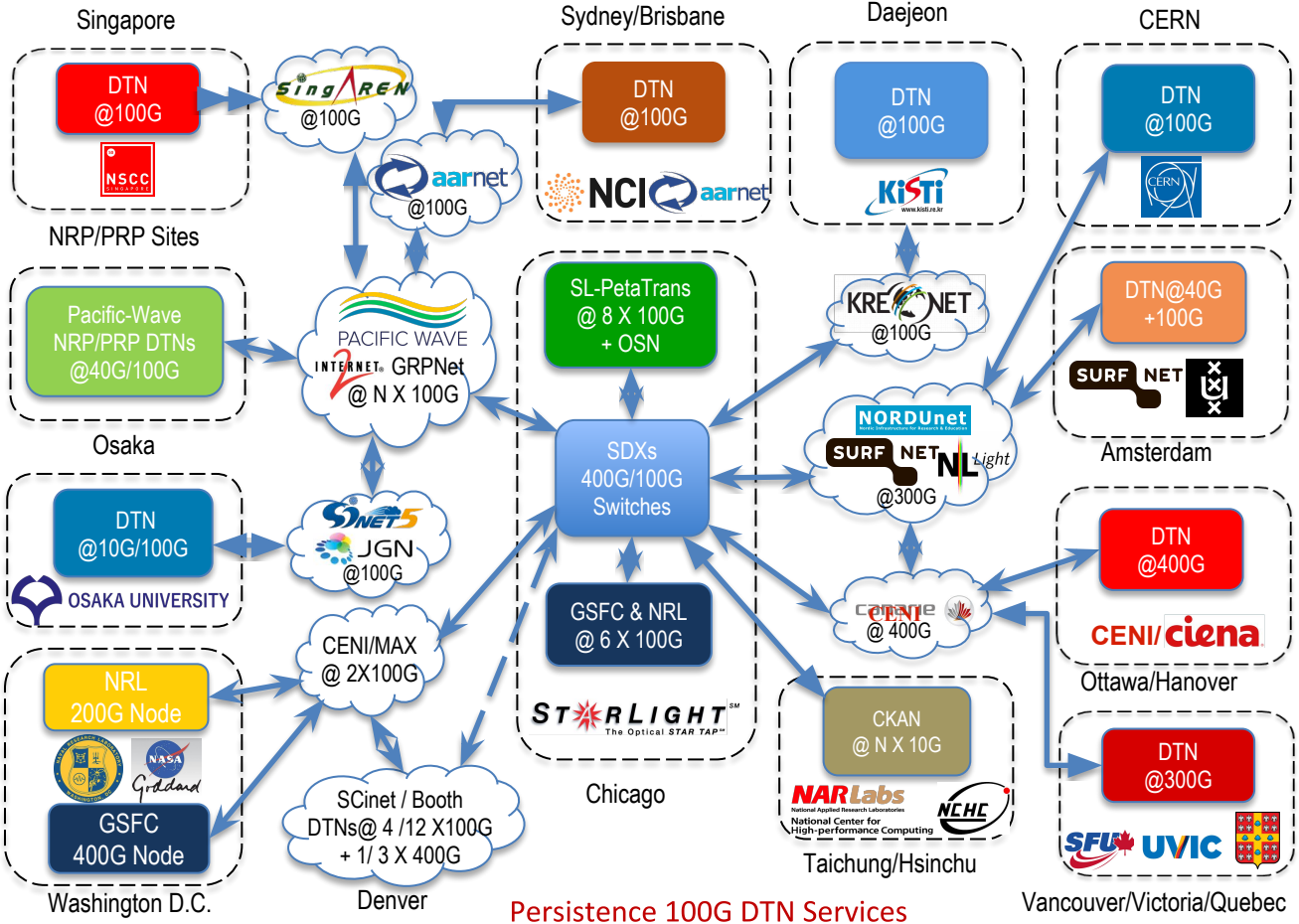


iCAIR

STARLIGHT SDX

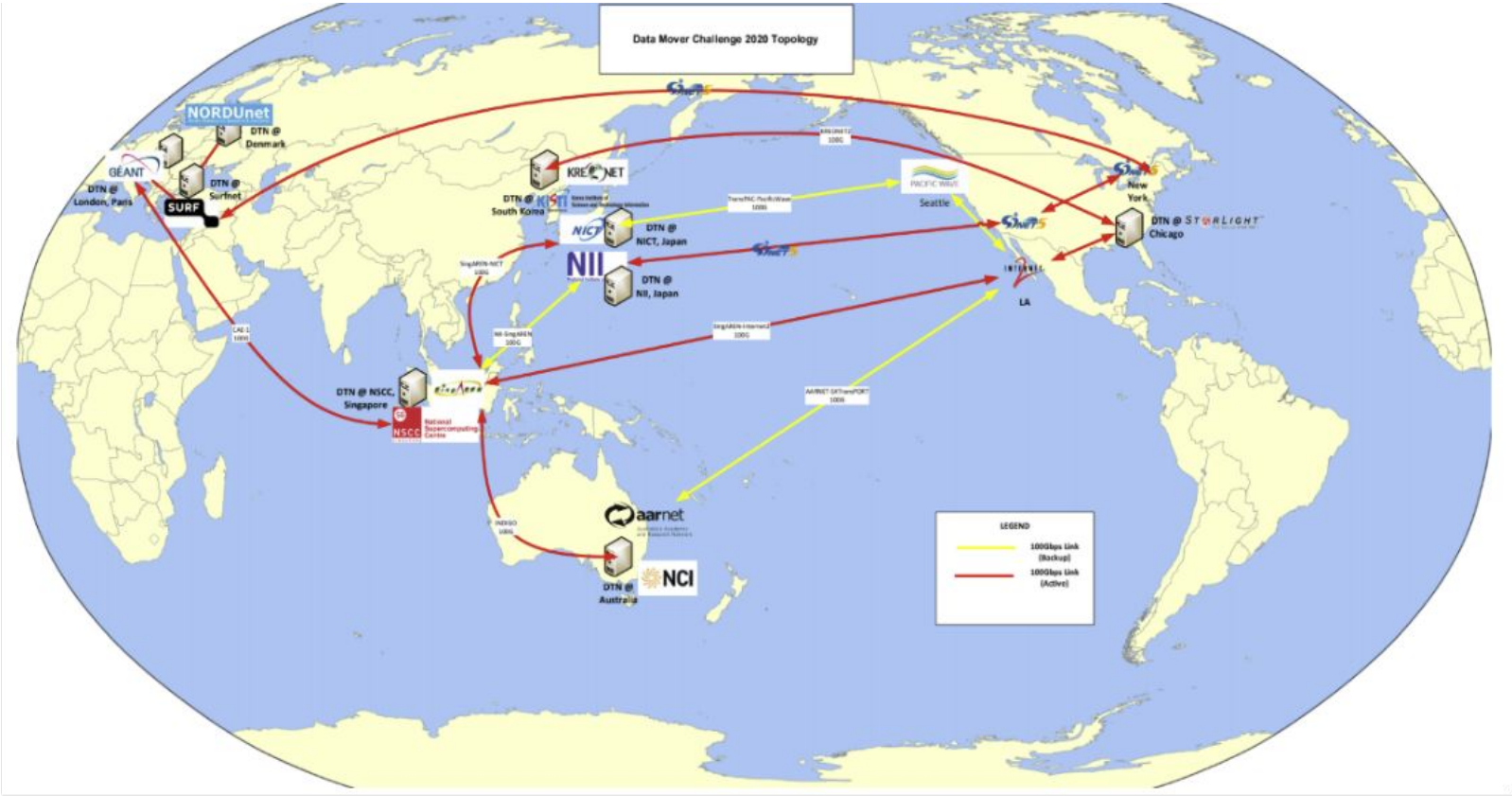

SC19
Denver, CO | hpc is now.
JOINT BIG DATA TESTBED
Pacific Wave& GRP Network
Indiana
KISTI
SURFnet
TIFR
MOXY
Fermilab
ESnet
StarLight
NASA
UMD
NRL
DREN
Ciena TB
MAX
ATDnet
Caltech Pasadena
TIFR
LSST La Serena
CENIC Sunnyvale
A*/Singaren
CENIC LA
AmLIght/ FIU Miami
UNESP Sao Paulo
UERJ Rio
NERSC
ESnet Testbed
PacWave
CENIC Seattle
ESnet
Internet2/ Ciena
CenturyLink
ESnet
CenturyLink
Internet2/Ciena
Latest Version at:
http://tinyurl.com/SC19-JBDT
To request changes, please leave a comment
See also:
http://tinyurl.com/SC19-NRE
SC19 floorplan
SCinet L1
SCinet L1
4x100G DTN
Dell Z9332F 400G
SCinet
SCinet L1
SCinet L1
SCinet L2/3
DCI
DCI
Arista 7060DX4 400G
Waveserver Ai(s)
2CRSI
Booth 1801
AMD Rome+ GPU + FPGA Infererence
400G XNET DEMO
Dell Z9332F 400G
Arista 7060PX4 400G
Arista 7060DX4 400G
Waveserver Ai(s)
Mellanox SN3700
Dell Z9100
Dell S60
Mellanox SN3700
Dell Z9264F-ON
Arista7280
Arista7280
Dell Z9264F
Dell Z9332F 400G
ExtremeSLX9640
Mellanox SN3700 200G
Latency Probe
NRL DTN & Stream Proc.
4Kx60fps Capture
4Kx60fps Display
NRL Gen 4 DTN Liqid Rome/NVMe
NRL Gen 4 DTN Liqid Rome/NVMe
NRL Gen 4 Bluefield
NRL Gen 4 nConum
StarLight PetaTrans
StarLight PetaTrans
NASA DTN
Caltech 2CRSI DTN 4 Node AMD Rome 1
Caltech 2CRSI DTN 4 Node AMD Rome 2
Caltech 2CRSI DTN 4 Node AMD Rome 3
Pavilion IO
Supermicro 6027 Xeon E5 4 Node
SANDIE 5 Echostreams
SANDIE 6 Echostreams
400G - FR4
200G - SR4 or DAC
100G - CLR4
100G - LR4
100G - SR4
100G - DAC
40G - SR4
40G - DAC
10G
1G
11/01/2019
StarLight/UIC/NASA/NRL Booth 993
MORE
Caltech/ESnet/AmLight/Yale/Northeastern Booth 543

PetaTrans: Petascale Sciences Data Transfer



OSG-IRNC DTN Federation

Data Mover Challenge - SC Asia 2020



Conclusion

DTN-as-a-Service provides an environment for high-speed transfer

Flexible and modular design

Support for big science network data transfer

Consistent performance

Harnessing multiple user tools

NVMe over Fabrics and k8s

iCAIR

STARLIGHT SM SDX

Future works

Prototype NVMe over Fabrics WAN Service

Include additional storage system performance tuning and integration

SCinet DTN will graduate from SCinet X-net and move to SCinet DevOps team in SC20

Please visit StarLight booth 993 for DTN-as-a-Service and 400G experiments.

iCAIR

STARLIGHT SDX

Thanks to..

NSF International Research Network Connections (IRNC) Grant, NSF Cloud Grants, and other NSF Grants

All the SCinet teams who make this project possible

EchoStreams, Dell Networking and Server group for equipment and support

All the partners and the participants

iCAIR

STARLIGHT SDX