U.S. NAVAL RESEARCH LABORATORY

Resilient Distributed Processing

SC21 Demonstration
Naval Research Laboratory
Center for Computational Science

November 15-18, 2021

SC21 St. Louis, MO science & beyond.

Basil Decina (basil.decina@nrl.navy.mil)
Linden Mercer (linden.mercer.ctr@nrl.navy.mil)
Dardo Kleiner (dardo.kleiner.ctr@nrl.navy.mil)
Larry O'Ferrall (larry.oferrall@nrl.navy.mil)
Louis Berger (louis.berger.ctr@nrl.navy.mil)
Richard Elliott (richard.elliott.ctr@nrl.navy.mil)

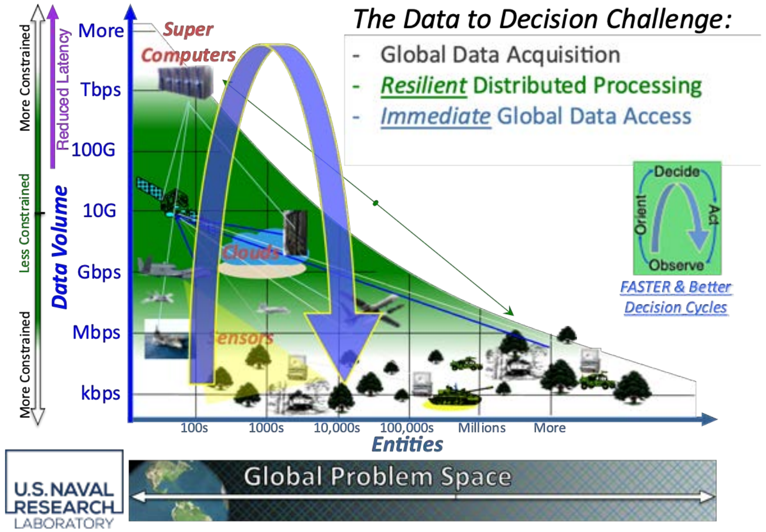
DISTRIBUTION STATEMENT A. Approved for public release. This material is based upon work supported by the Department of Defense, US Naval Research Laboratory.

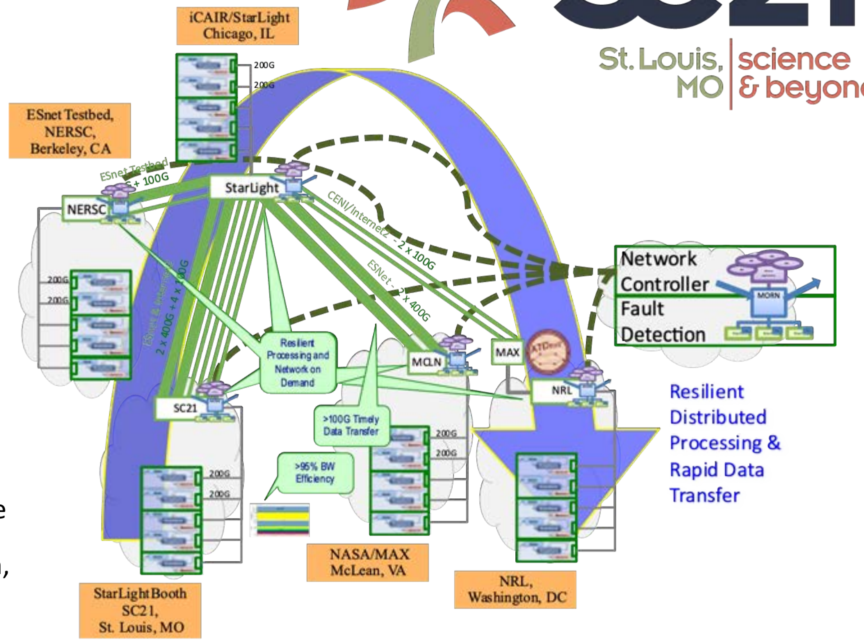
DISTRIBUTION A: Approved for Public Release ; Distribution Unlimited

Resilient Distributed Processing

Naval Research Laboratory
Center for Computational Science
SC21 Demonstrations

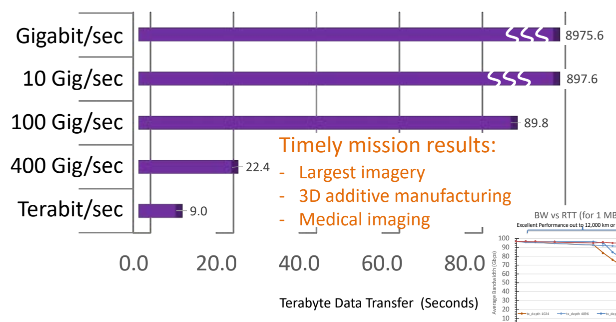
For More Information Contact:
Basil Decina, basil.decina@nrl.navy.mil
Linden Mercer, linden.mercer.ctr@nrl.navy.mil

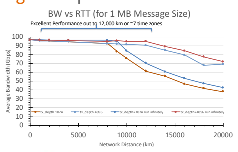




NRL aims to demonstrate:
- Dynamic arrangement and re-arrangement of widely distributed processing of large volumes of data across a set of compute and network resources organized in response to resource availability and changing application demands
- Real-time video processing pipeline will be demonstrated from SC21 to compute and storage assets in Washington, DC; McLean, VA; Chicago, IL; and Berkeley, CA

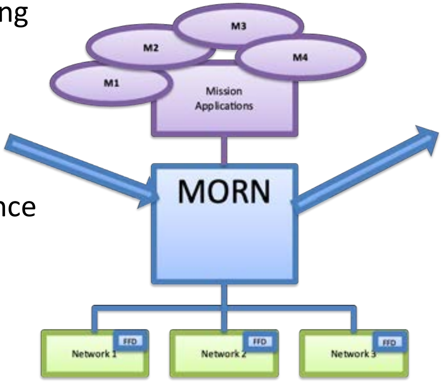
Terabyte Data Movement

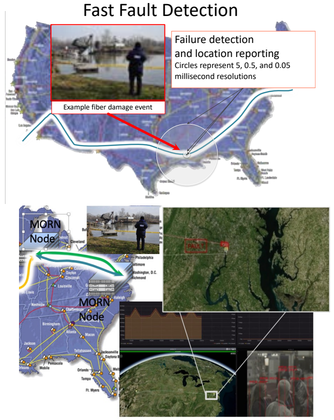




100 Gigabyte = ~3 hours of high quality 4K video (H.265), the best Blu-ray disc, 9 hours of Netflix 4K video

Specific goals:
1. Fast fault detection and location using an active probe.
2. Dynamic shifting of processing and network resources from one location/path/system to another (in response to demand and availability).
3. Leverage RDMA/distance performance for timely Terabyte bulk data transfers (goal < 1 min Tbyte transfer on 400G network).
4. Network data flows protected by IP and Ethernet Traffic Flow Security.



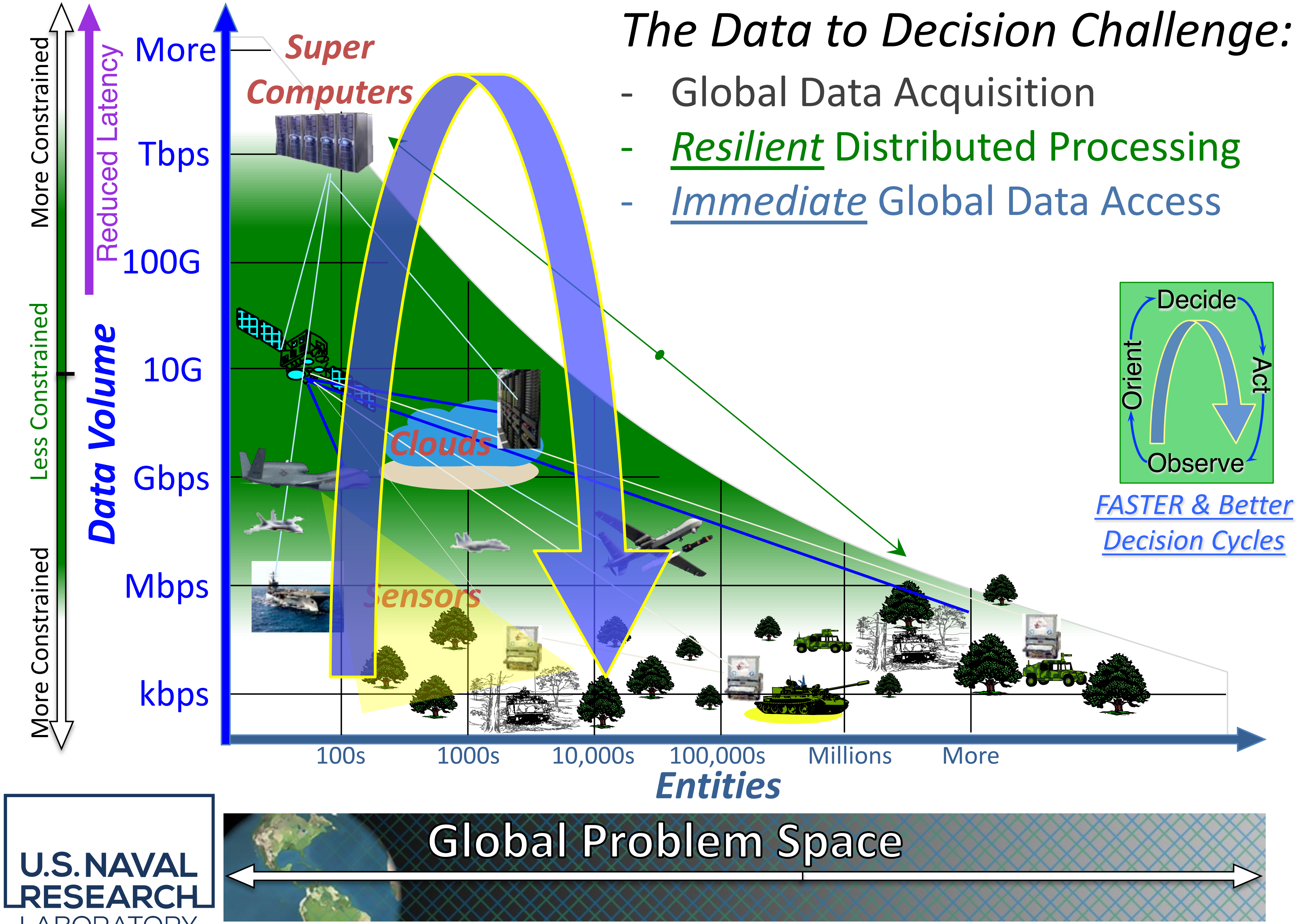


Interconnected and interlocking problems demand a resilient distributed DoD, high performance, low latency data sharing infrastructure

ATDnet ADVANCED TECHNOLOGY DEMONSTRATION NETWORK — DoD HPC DREN — STARLIGHT — iCAIR — NASA — MAX MID-ATLANTIC CROSSROADS — INTERNET2 — ESnet ENERGY SCIENCES NETWORK — ciena — ARISTA — U.S. NAVAL RESEARCH LABORATORY

The Data to Decision Challenge:

- Global Data Acquisition
- <u>*Resilient*</u> Distributed Processing
- <u>*Immediate*</u> Global Data Access



FASTER & Better Decision Cycles

U.S. NAVAL RESEARCH LABORATORY

Naval Research Laboratory Center for Computational Science "SC21" Demonstration

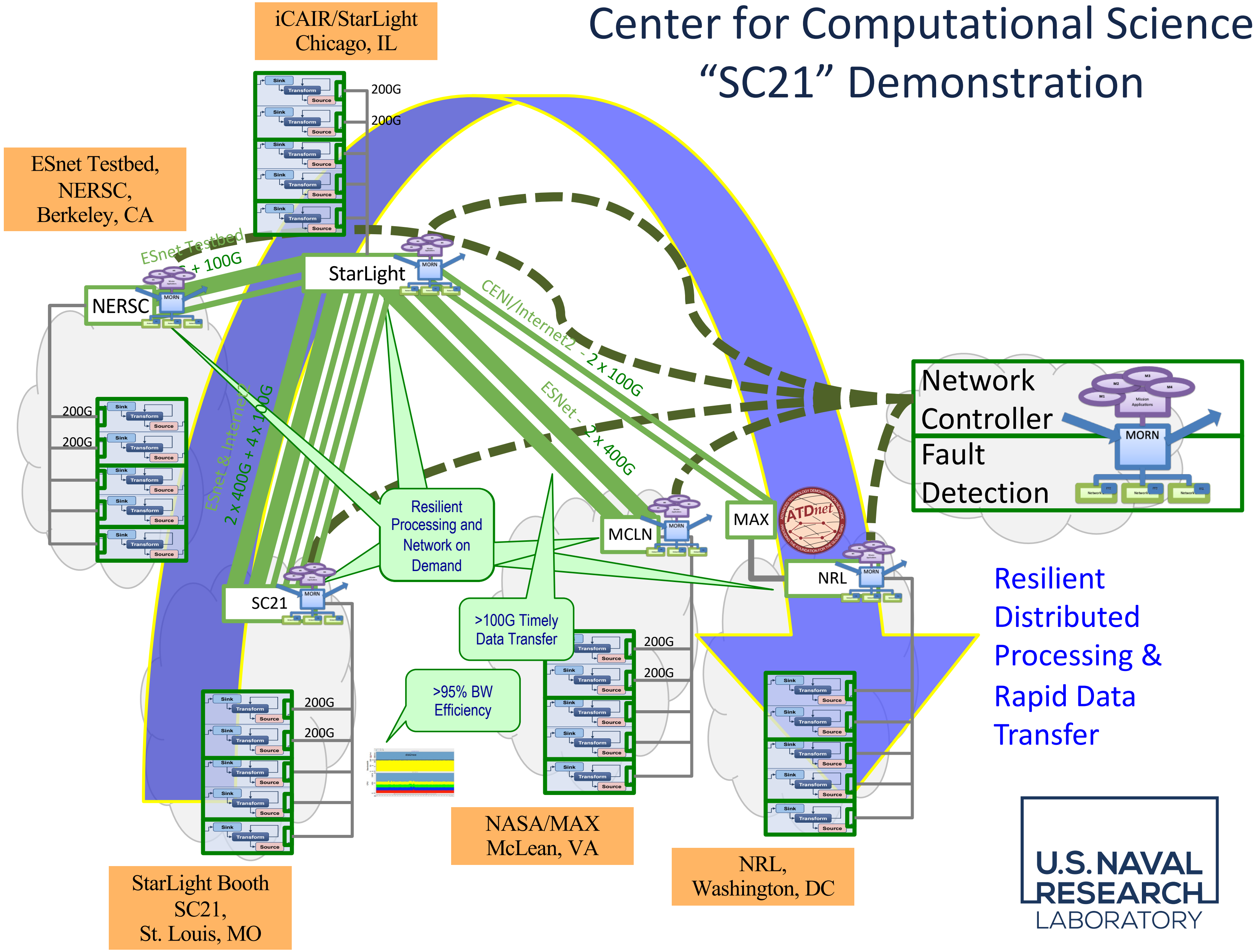
NRL aims to demonstrate:

- Dynamic arrangement and re-arrangement of widely distributed processing of large volumes of data across a set of compute and network resources organized in response to resource availability and changing application demands
- Real-time video processing pipeline will be demonstrated from SC21 to compute and storage assets in Washington, DC; McLean, VA; Chicago, IL; and Berkeley, CA

Specific goals:

1. Fast fault detection and location using an active probe.
2. Dynamic shifting of processing and network resources from one location/path/system to another (in response to demand and availability).
3. Leverage RDMA/distance performance for timely Terabyte bulk data transfers (goal < 1 min Tbyte transfer on 400G network).
4. Network data flows protected by IP and Ethernet Traffic Flow Security.

"Interconnected and interlocking problems" demand a high performance dynamic distributed data centric infrastructure

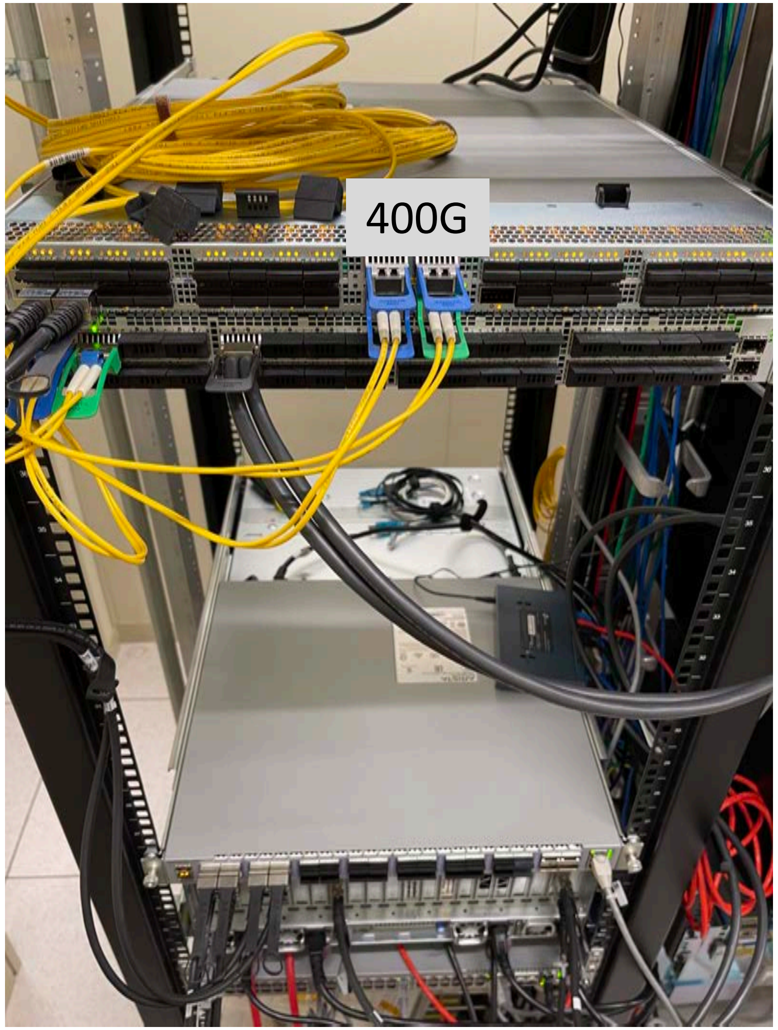


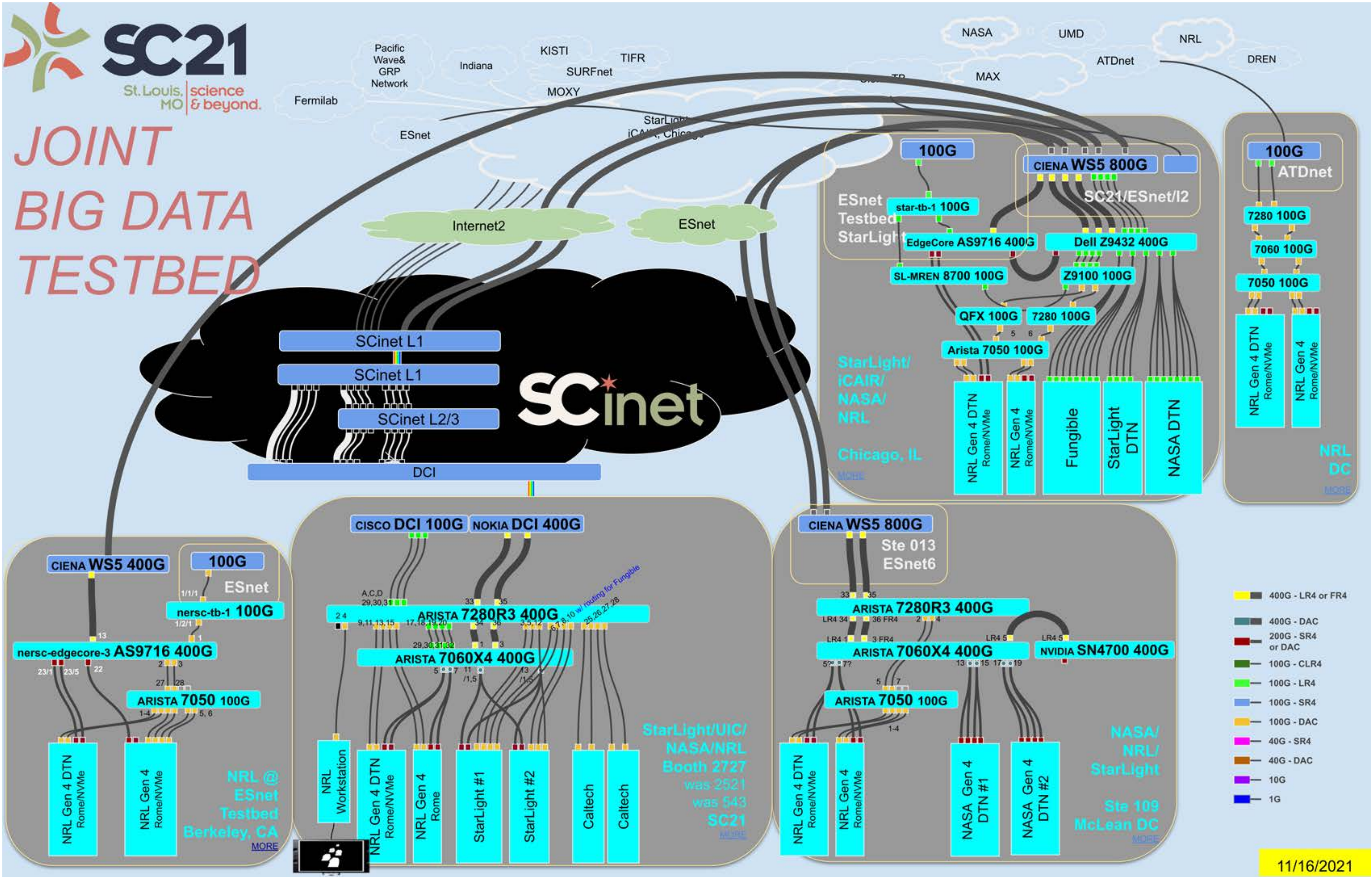
Resilient Distributed Processing & Rapid Data Transfer

U.S. NAVAL RESEARCH LABORATORY

DISTRIBUTION A: Approved for Public Release; Distribution Unlimited

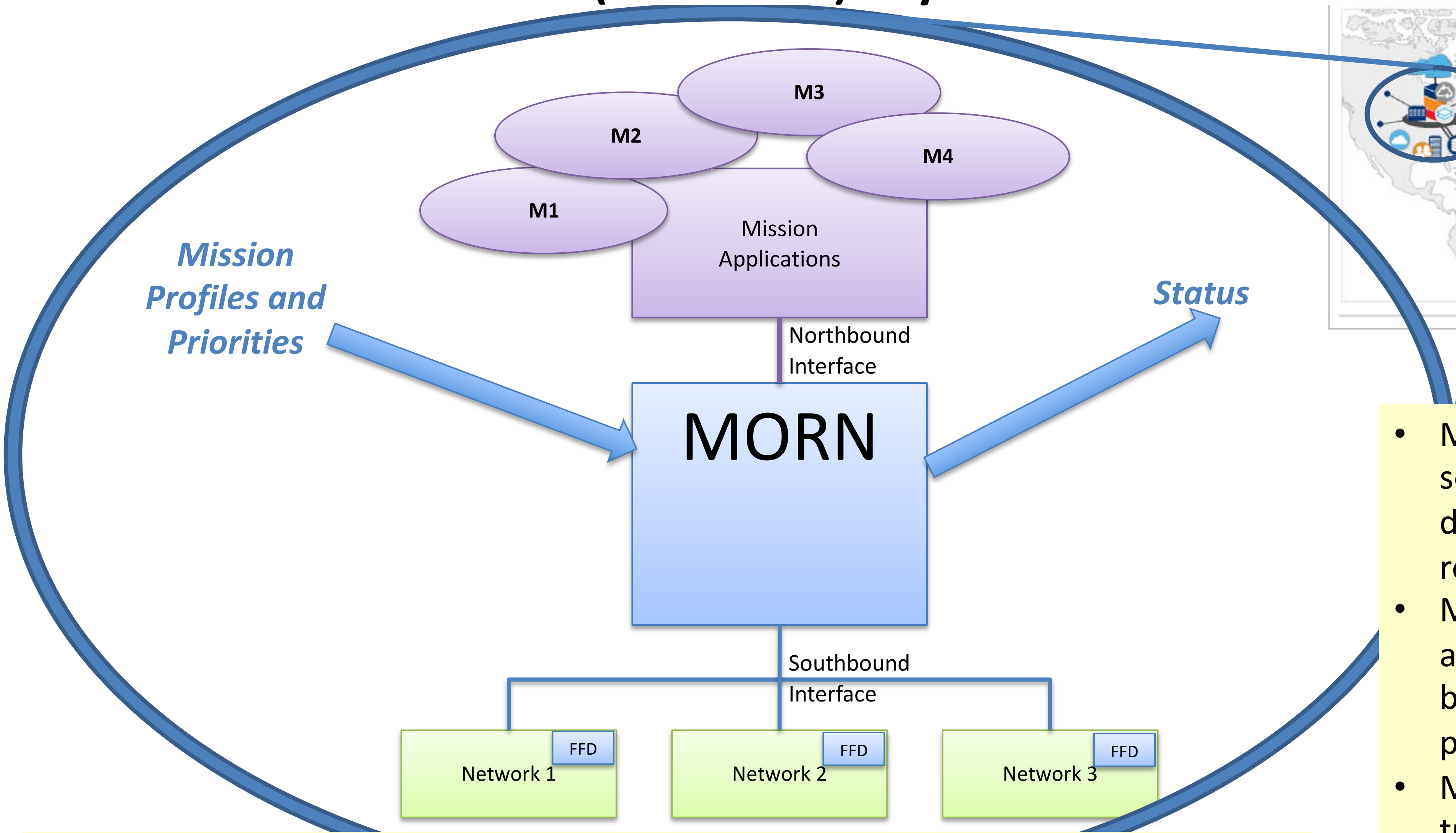
400G SC21 Stand-up

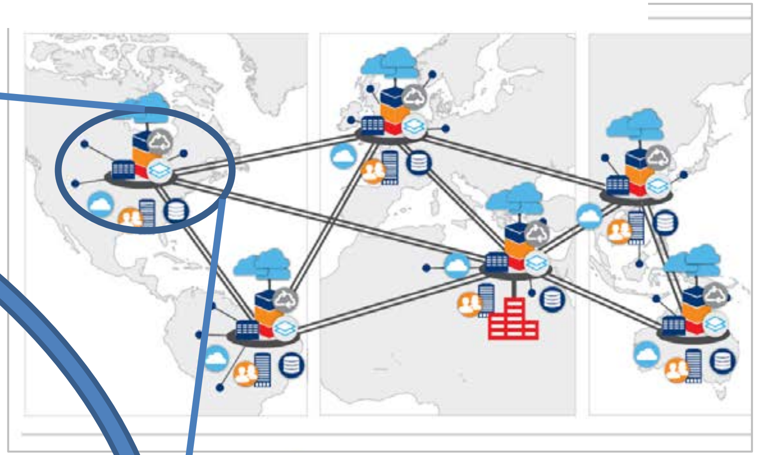




DISTRIBUTION A: Approved for Public Release; Distribution Unlimited

Mission Oriented Reconfigurable Networking (MORN) System View

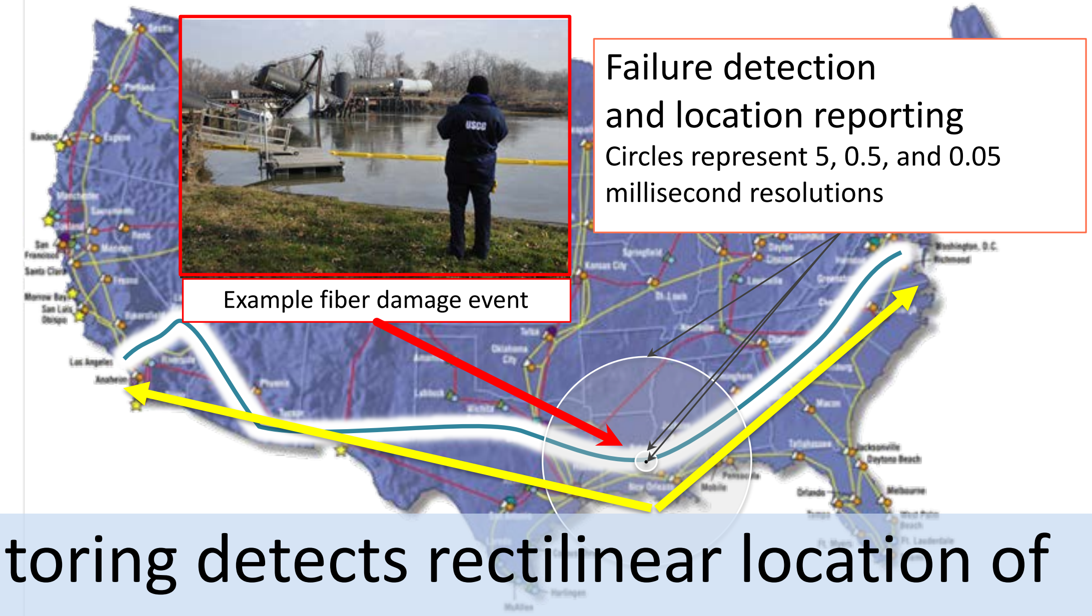




- MORN provides resilient service delivery using failure detection to trigger prioritized restoration
- MORN maps applications to available network resources based on mission policy and priorities
- MORN leverages industry traffic engineering and SDN

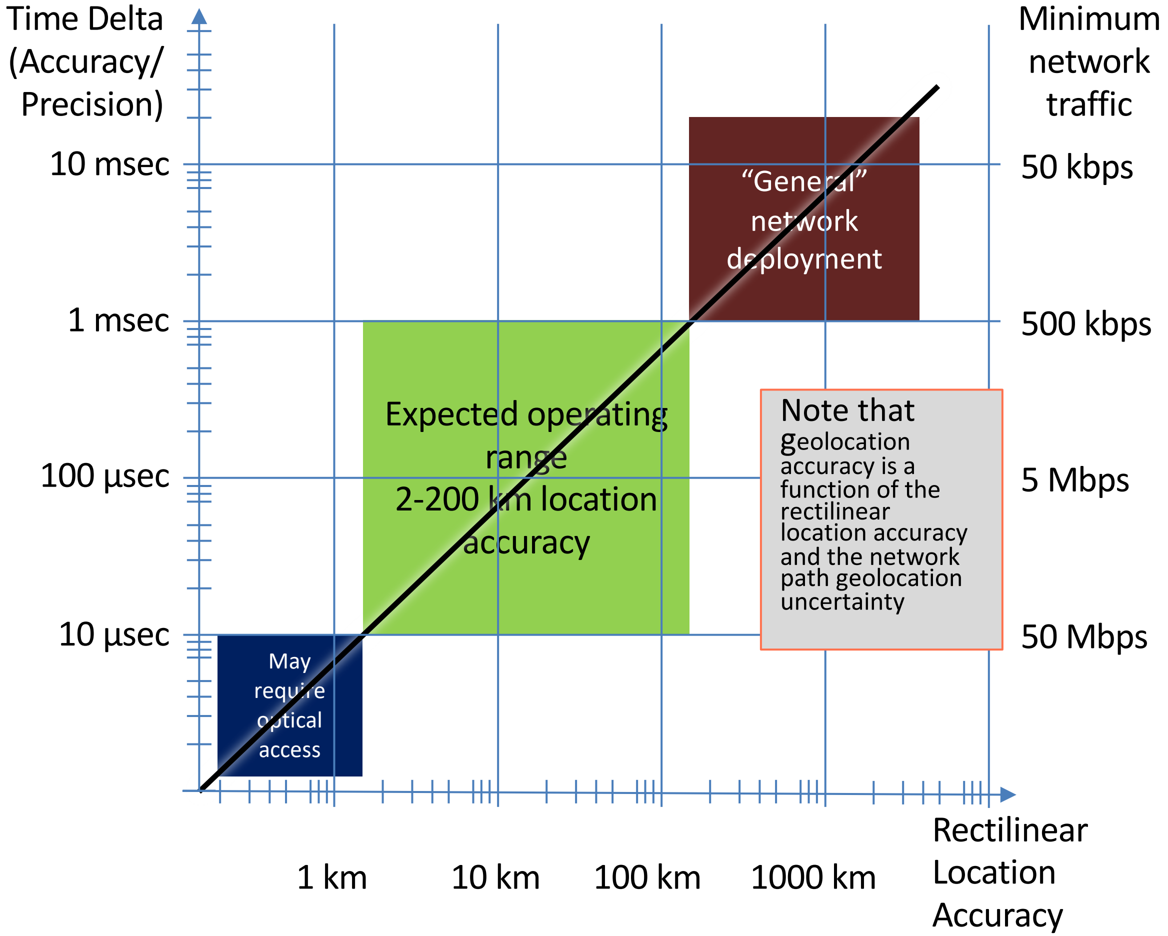
MORN is our approach to optimizing the allocation of resources to meet mission requirements

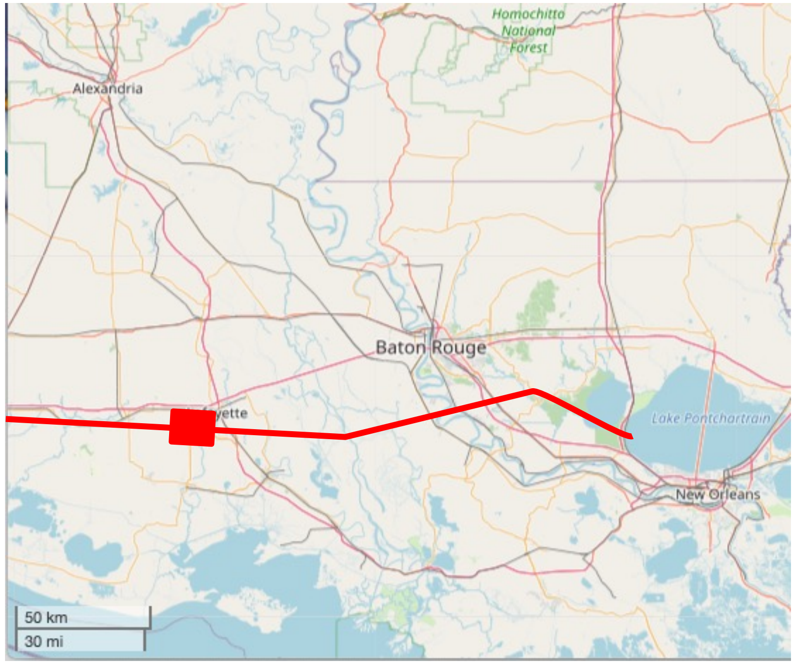
Fast Fault Detection

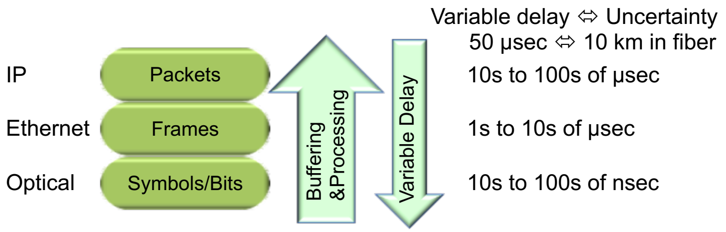


- Synchronized monitoring detects rectilinear location of two-way faults
  - Limited by *time synchronization and network delay variation*
- Knowledge of the path allows mapping of rectilinear location to geographic location
  - Limited by accuracy of *detailed path knowledge*

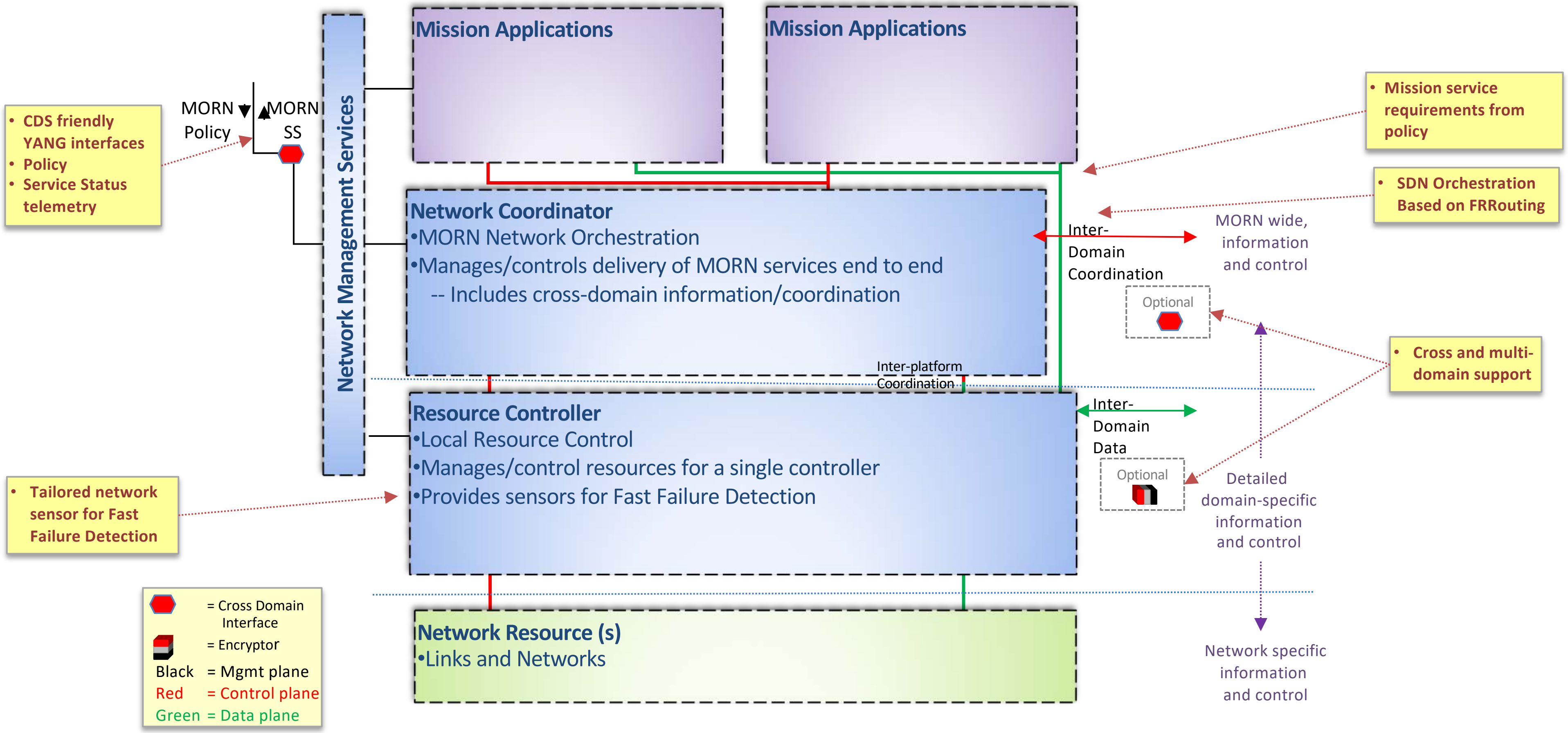
Fault Location Accuracy



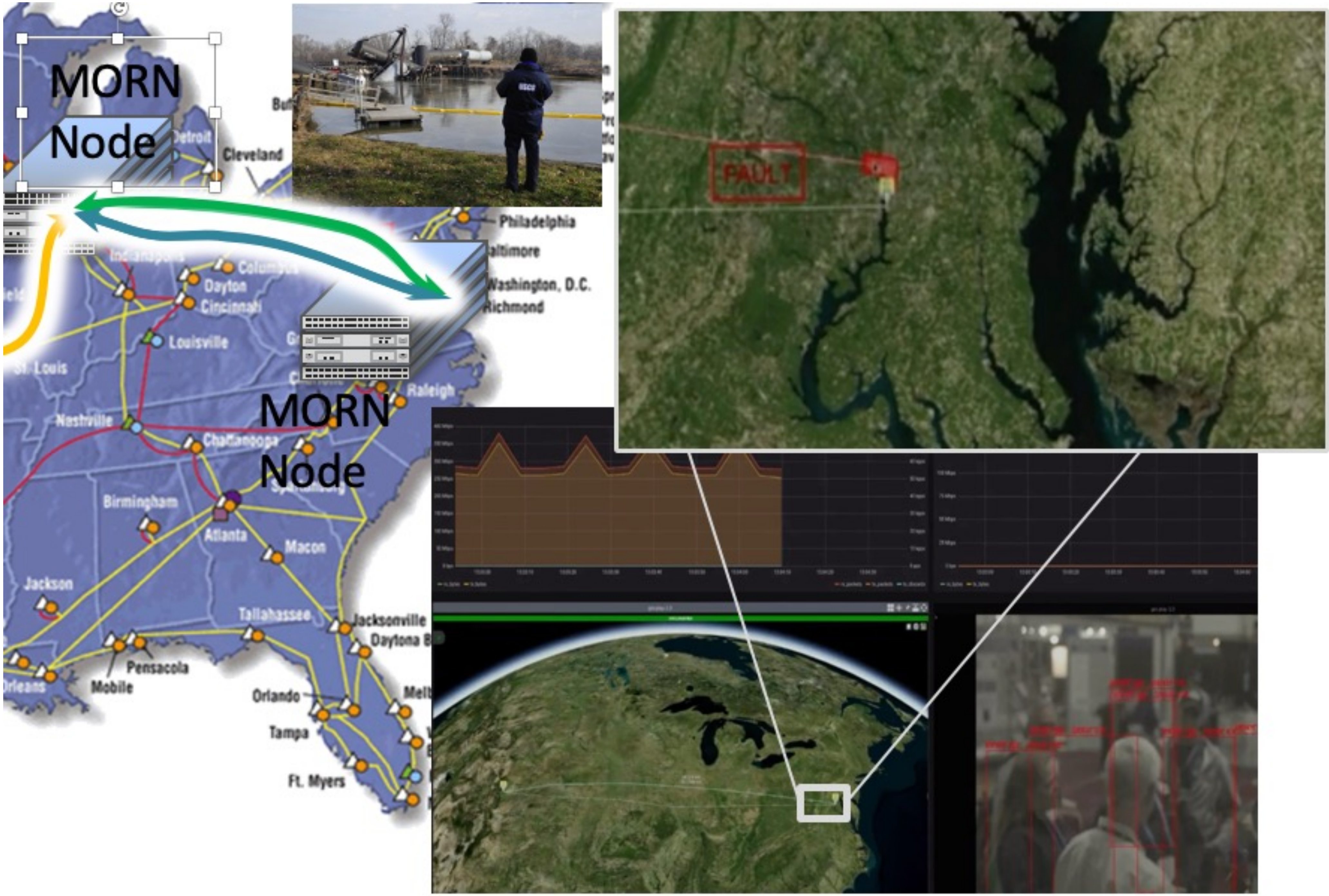




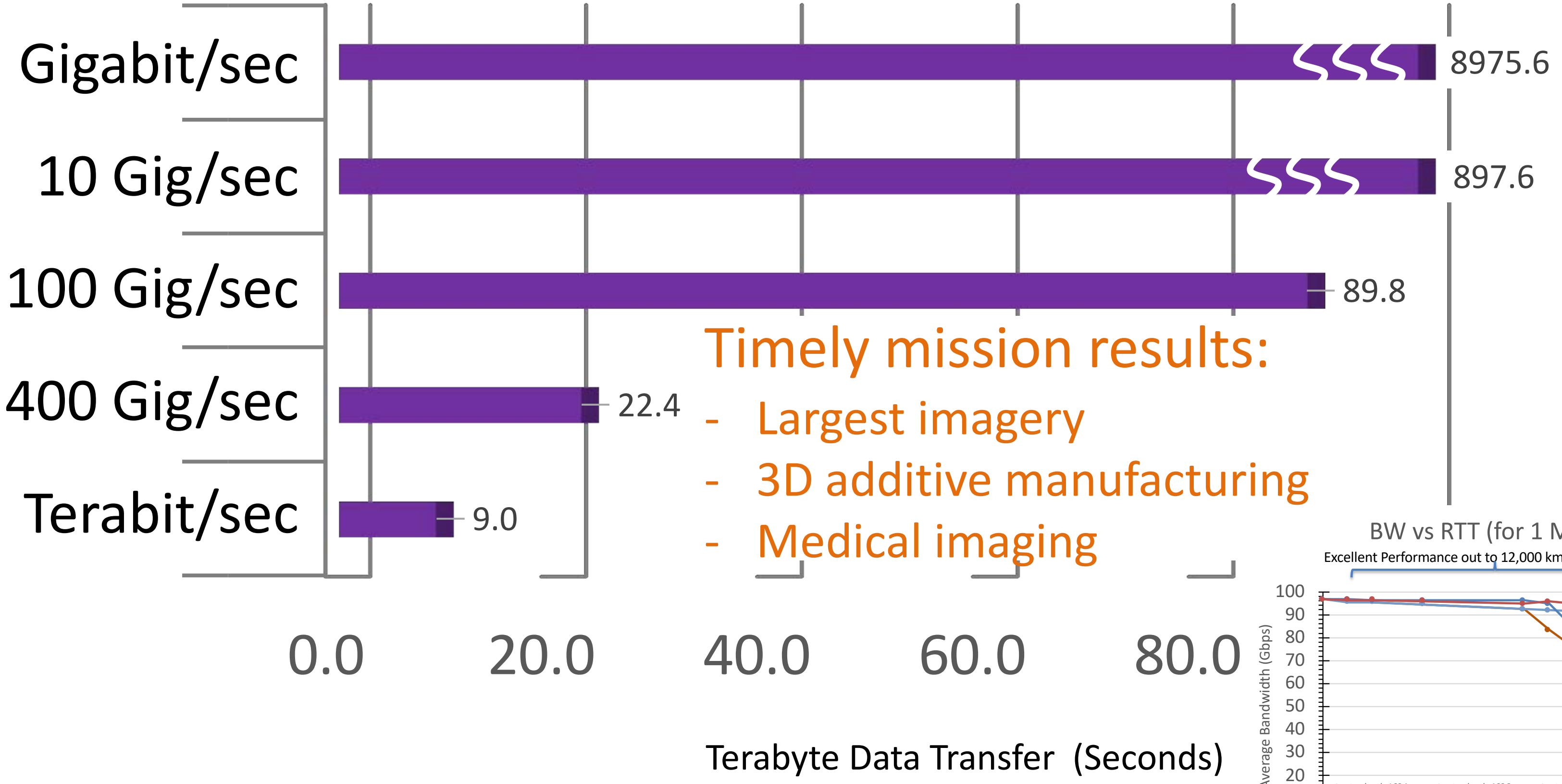
MORN Architecture Components



DISTRIBUTION A: Approved for Public Release; Distribution Unlimited

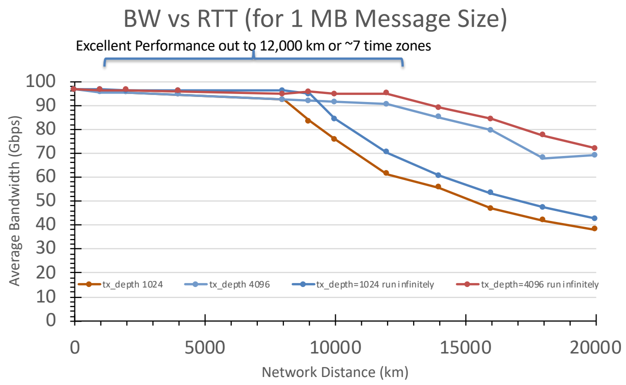

MORN
Node
MORN
Node
Detroit
Cleveland
Philadelphia
Washington, D.C.
Richmond
Columbus
Dayton
Cincinnati
Louisville
St. Louis
Nashville
Chattanooga
Raleigh
Birmingham
Atlanta
Macon
Jackson
Tallahassee
Jacksonville
Pensacola
Mobile
Orlando
Tampa
Ft. Myers
FAULT

Terabyte Data Movement



Timely mission results:

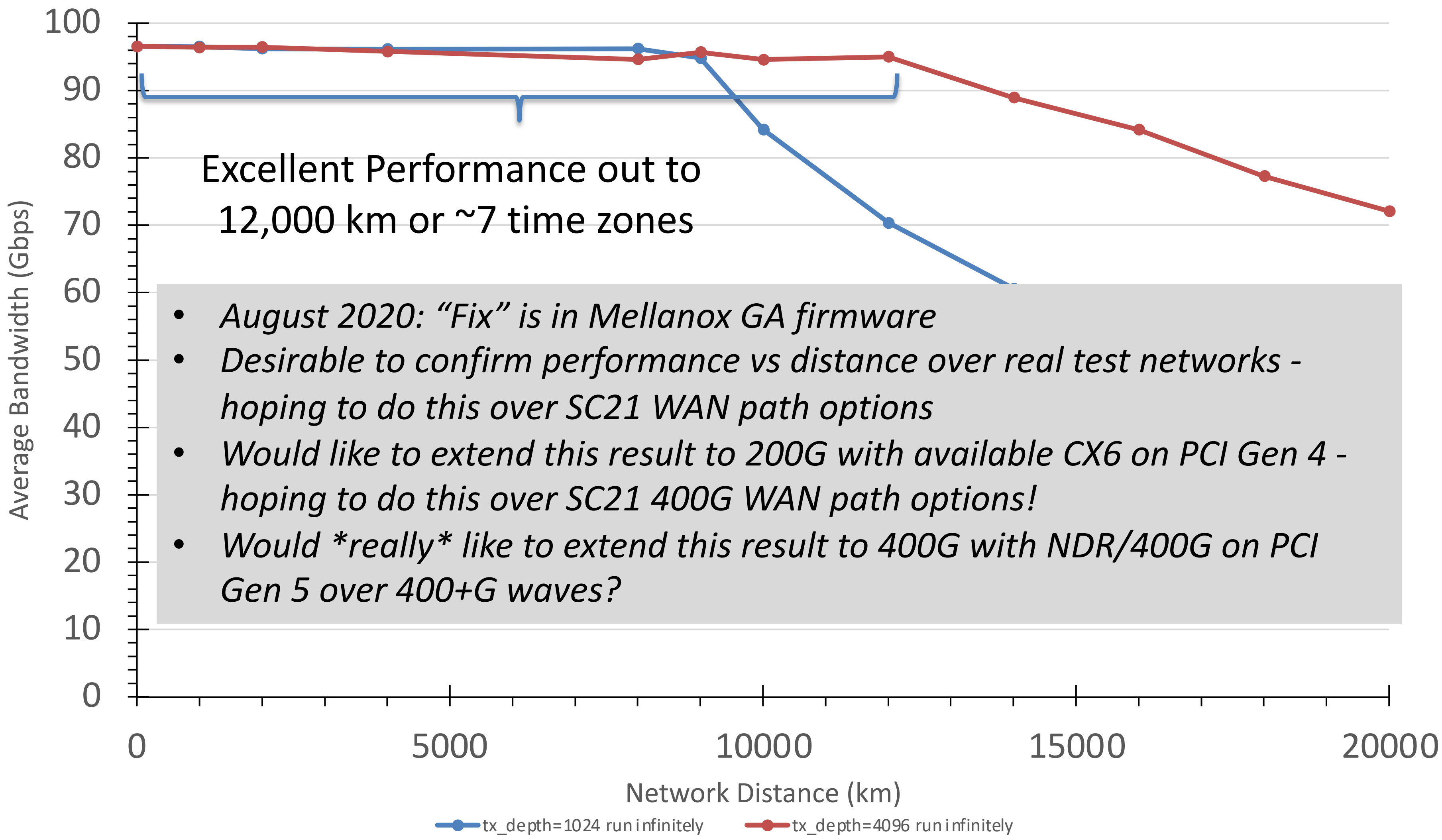
- Largest imagery
- 3D additive manufacturing
- Medical imaging



100 Gigabyte = ~3 hours of high quality 4K video (H.265), the best Blu-ray disc, 9 hours of Netflix 4K video

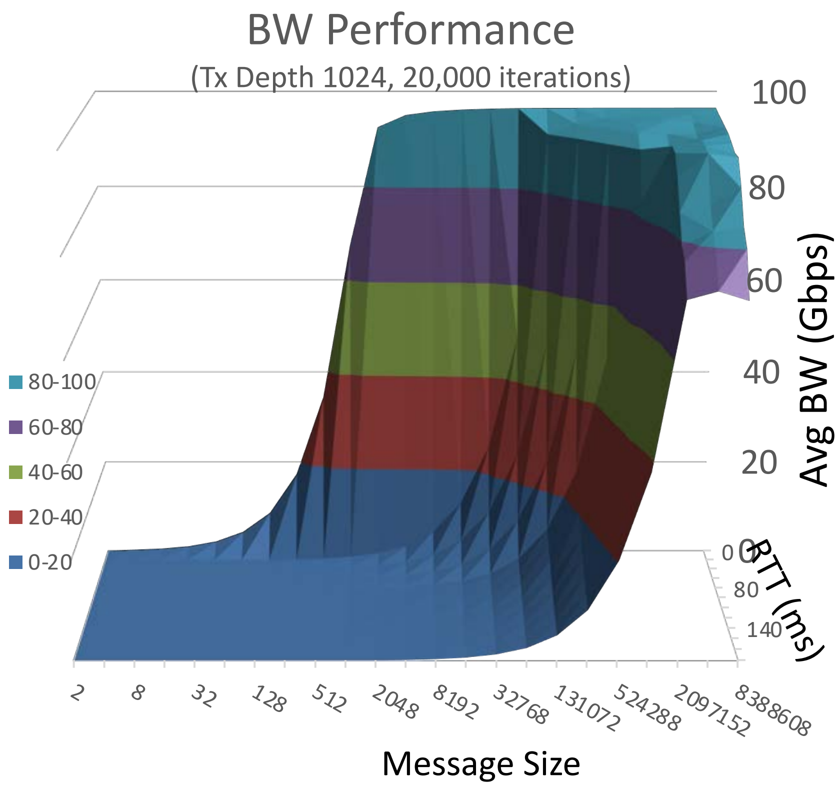
BW vs RTT

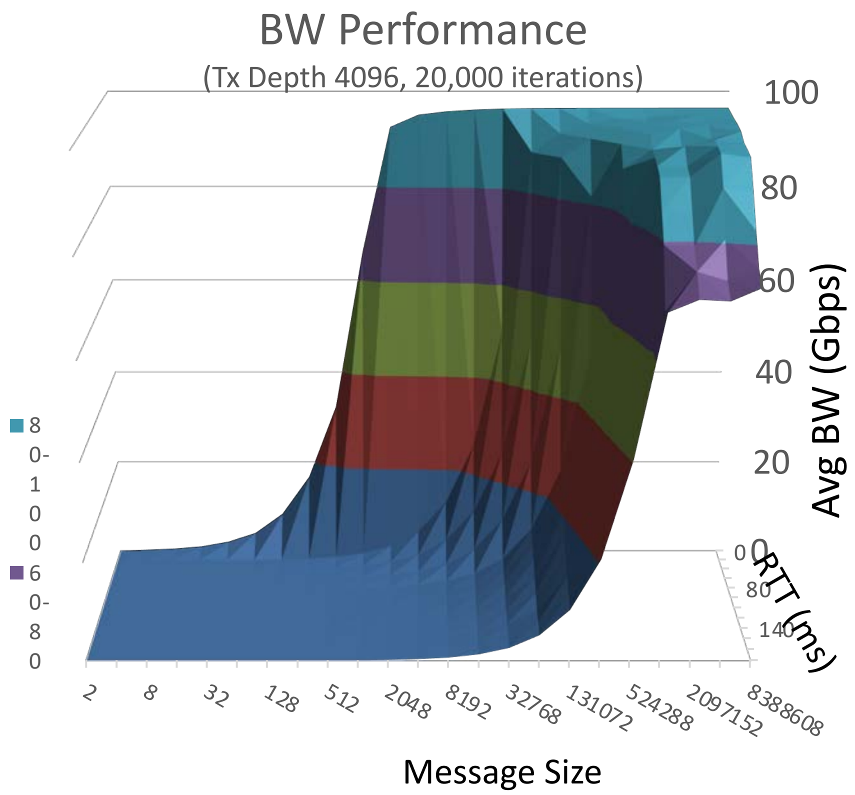
(for 1 MB Message Size)
Most Relevant to Lustre Performance



- *August 2020: “Fix” is in Mellanox GA firmware*
- *Desirable to confirm performance vs distance over real test networks - hoping to do this over SC21 WAN path options*
- *Would like to extend this result to 200G with available CX6 on PCI Gen 4 - hoping to do this over SC21 400G WAN path options!*
- *Would *really* like to extend this result to 400G with NDR/400G on PCI Gen 5 over 400+G waves?*

RDMA/Distance Reference Data Set





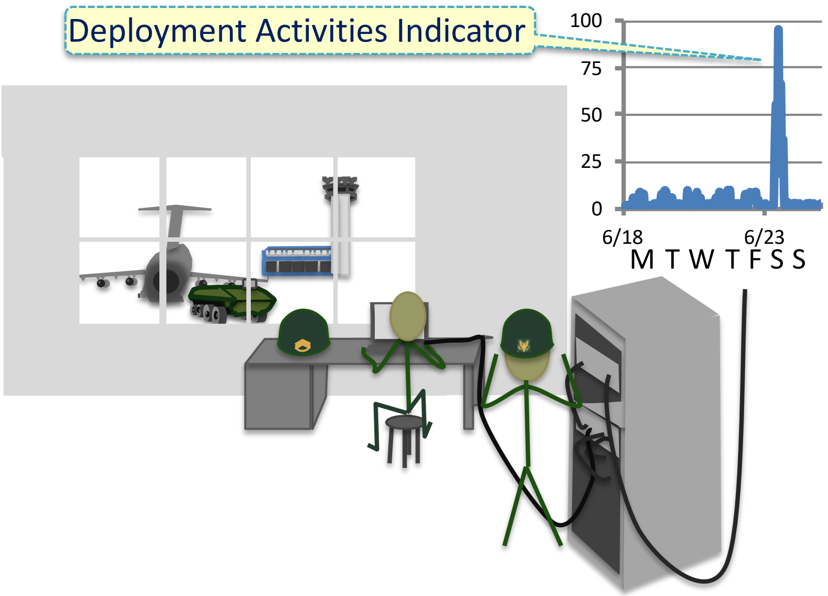
Traffic Flow Security (TFS)

- Full period (bulk) encryption historically used to deliver Traffic Flow Security (TFS) in fixed/wired networks, Network encryption is replacing link encryption in wireline communication systems
- Work underway to update standards
  - IEEE MACsec 802.1 Aedk
    Publication approval process underway

    https://1.ieee802.org/security/802-1aedk/

  - IETF IPsecME WG TFS
    Core protocol accepted, working through process
    Publication approval process requested

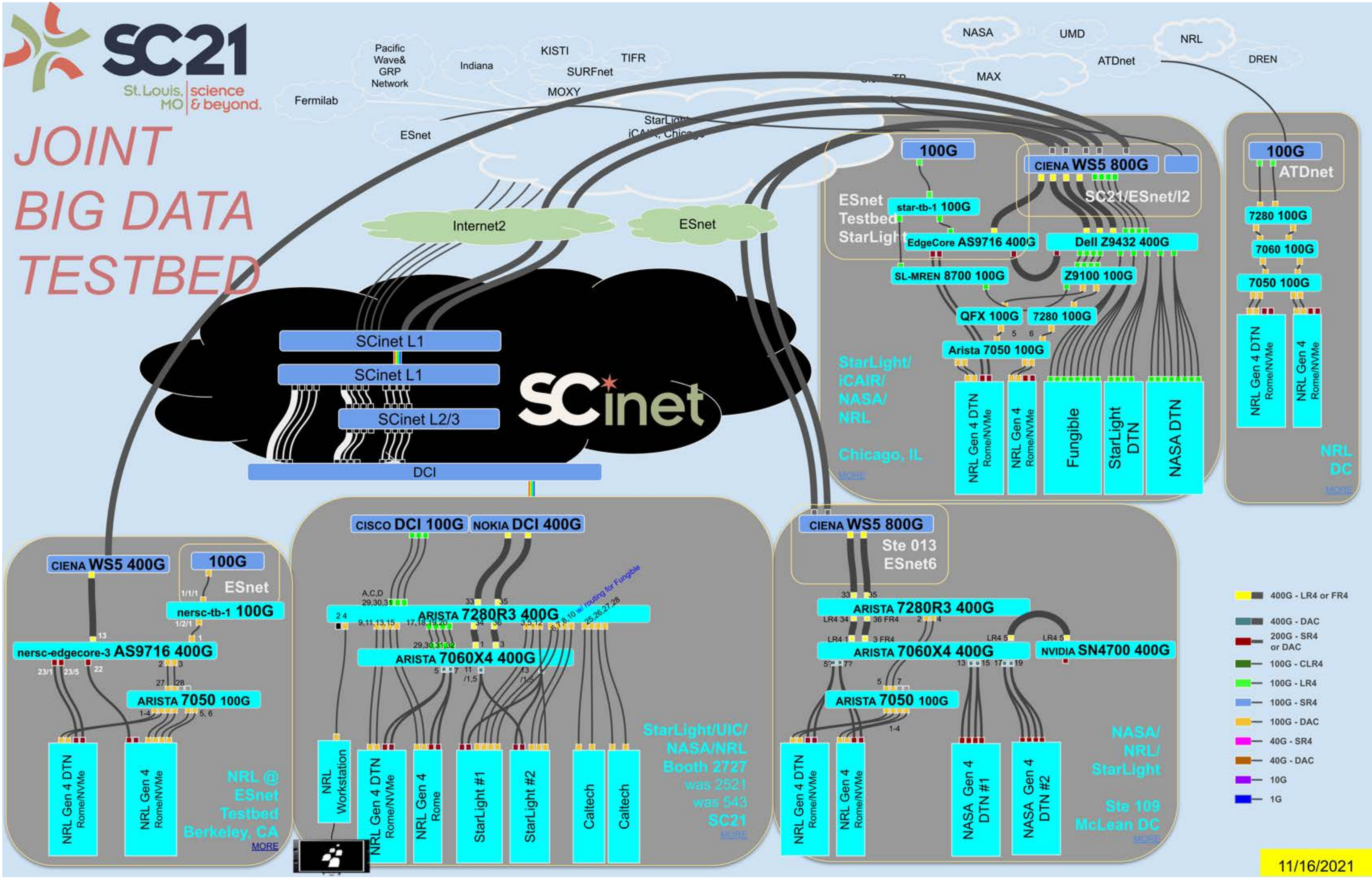
    https://tools.ietf.org/html/draft-ietf-ipsecme-iptfs-03



DISTRIBUTION A: Approved for Public Release; Distribution Unlimited

Terabyte Data Movement

		Now	Next	Soon
Research Network	Network: Systems: Storage: TB Transfers:	2 x 100 Gbps PCIe Gen 4 Memory 75 seconds	400 Gbps 2 x PCIe Gen 4 NVMe 30 seconds	Tbps (3 x 400G?) 3 x PCIe Gen 5 NVMe <10 seconds
Pilot Network	Network: Systems: Storage: TB Transfers:	2 x 100 Gbps 2 x PCIe Gen 3 HD/NVMe 150 seconds	2 x 100 Gbps 4 x PCIe Gen 3 HD/NVMe 55 seconds	400 Gbps 2 x PCIe Gen 4 NVMe 30 seconds
Operational	Network: Systems: Storage: TB Transfers:		2 x 100 Gbps 4 x PCIe Gen 3 HD/NVMe 60 seconds	4 x 100 Gbps PCIe Gen 4 HD?/NVMe 40 seconds



DISTRIBUTION A: Approved for Public Release; Distribution Unlimited

IEEE SC21 Network Research Exhibits Web Abstract

This demonstration will build on our previous demonstrations. We aim to show dynamic arrangement and re-arrangement of widely distributed processing of large volumes of data across a set of compute and network resources organized in response to resource availability and changing application demands. A real-time video processing pipeline will be demonstrated from SC21 to the Naval Research Laboratory assets in Washington, DC, McLean, VA, Chicago, IL, Berkeley, CA and back to SC21. High volume bulk data will be transferred concurrently across the same data paths. A software-controlled network will be assembled using a number of switches and multiple SCinet 100G/400G connections. We plan to show rapid deployment and redeployment, real-time monitoring and QOS management of these application data flows with very different network demands. Technologies we intend to leverage include SDN, RDMA, RoCE, NVMe, GPU acceleration and others.

NRL will have two major thrusts for our SC21 demo/tests: Mission Oriented Reconfigurable Networking (MORN) and rapid terabyte data movement. We have 5 sites (NRL/Washington, DC; McLean, VA; StarLight/Chicago, IL; NERSC/Berkeley CA; SC21/St. Louis, MO) in the 100G network monitored and controlled by MORN and 4 of those sites have 400G wide area network connections supporting fast delivery of critical data sets and massive distributed computing problems.

Specific Goals:

- Fast fault detection and location using an active probe.
- Dynamic shifting of processing and network resources from one location/path/system to another (in response to demand and availability).
- Leverage improved (restored) RDMA/distance performance for timely Terabyte bulk data transfers (goal < 1 min Tbyte transfer on 400G network).
- Network data flows protected by IP and Ethernet Traffic Flow Security

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov