ESnet
ENERGY SCIENCES NETWORK

The Science DMZ: Strategic Future

Eli Dart
Energy Sciences Network (ESnet)
Lawrence Berkeley National Laboratory

TNC 2023
Tirana, Albania
8 June, 2023

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB

Outline

- Science DMZ as a foundation
- New uses: Streaming DTNs
- Security: Science DMZ and Zero Trust

The Science DMZ as a Foundation

- At this point, there are many Science DMZ deployments
- Common use case: DTNs for file transfer
  - This is old hat or "normal" at this point, but it's still important and valuable
  - Routine performance is enabling for science
- Wider deployment comes with network effects
  - Scientists can expect data transfer to work well
  - Good performance is becoming normalized, which means scientists can include it in their designs
- Growth beyond the simple case – more in a moment

ESnet

The Value Of Routine Performance

- It’s important to get to where high performance is normal
- No magic, no arcana, things just normally work – for petabytes of data
- DOE HPC facilities now easily shuffle around hundreds of terabytes
  - Some people have smaller data sets too
  - But the point is that it’s normal and routine
- What follows is one specific example, chosen because of some specific features

ESnet
ENERGY SCIENCES NETWORK

National Energy Research Scientific Computing Center

Interfaces Flow

Website http://www.nersc.gov/

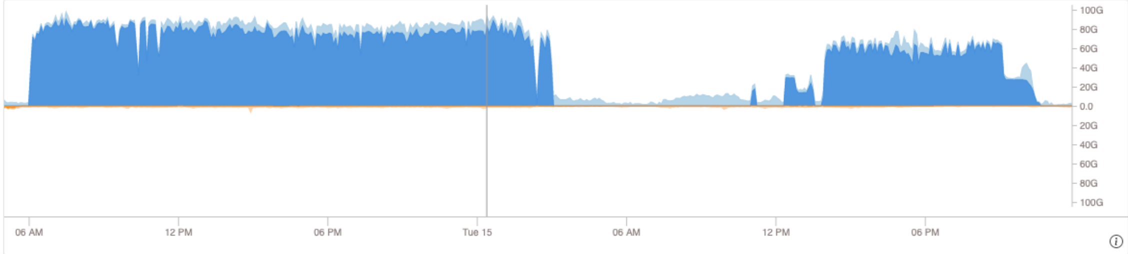
BREAKDOWN Autonomous Systems (origin)

TIME day week month custom

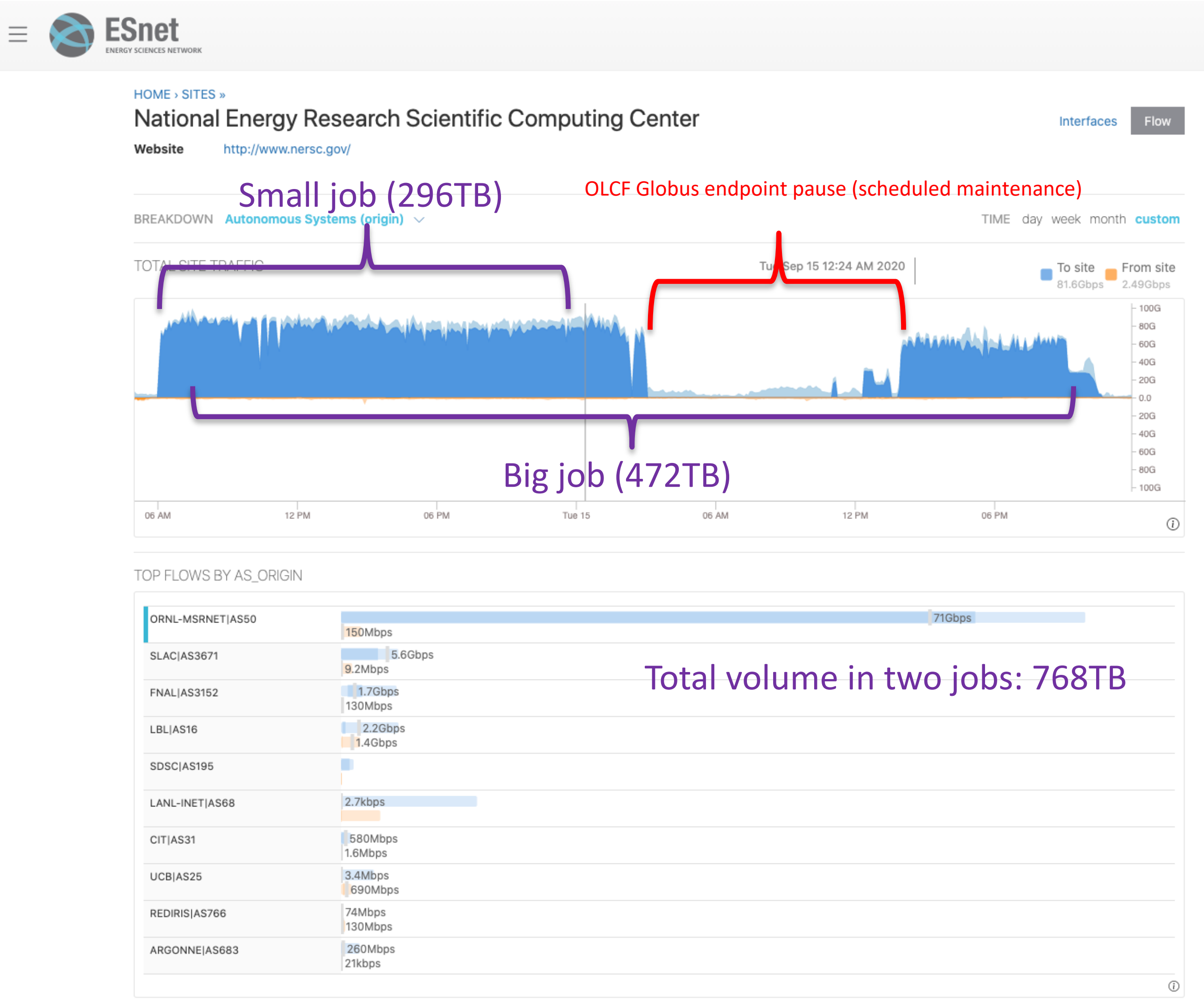
TOTAL SITE TRAFFIC

Tue Sep 15 12:24 AM 2020

To site 81.6Gbps From site 2.49Gbps



TOP FLOWS BY AS_ORIGIN



ESnet
ENERGY SCIENCES NETWORK
HOME › SITES »
National Energy Research Scientific Computing Center
Interfaces
Flow
Website
http://www.nersc.gov/
Small job (296TB)
OLCF Globus endpoint pause (scheduled maintenance)
BREAKDOWN
Autonomous Systems (origin)
TIME
day
week
month
custom
TOTAL SITE TRAFFIC
Tue Sep 15 12:24 AM 2020
To site
81.6Gbps
From site
2.49Gbps
100G
80G
60G
40G
20G
0.0
20G
40G
60G
80G
100G
Big job (472TB)
06 AM
12 PM
06 PM
Tue 15
06 AM
12 PM
06 PM
TOP FLOWS BY AS_ORIGIN
ORNL-MSRNET|AS50
71Gbps
150Mbps
SLAC|AS3671
5.6Gbps
9.2Mbps
Total volume in two jobs: 768TB
FNAL|AS3152
1.7Gbps
130Mbps
LBL|AS16
2.2Gbps
1.4Gbps
SDSC|AS195
LANL-INET|AS68
2.7kbps
CIT|AS31
580Mbps
1.6Mbps
UCB|AS25
3.4Mbps
690Mbps
REDIRIS|AS766
74Mbps
130Mbps
ARGONNE|AS683
260Mbps
21kbps

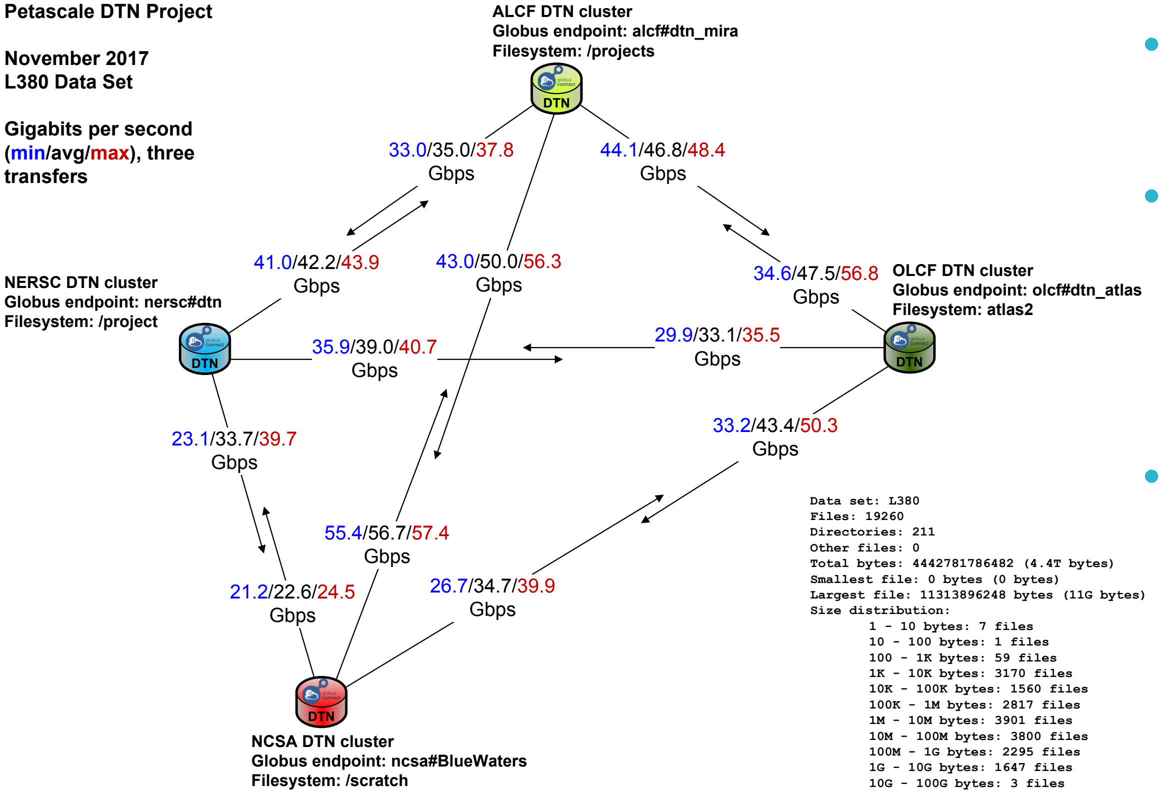
Key Points (1)

- Two Globus transfers from OLCF to NERSC were started by the same user (presumably for the same project – DESI – which is dark energy/cosmology) within 40 minutes of each other.
  – One transfer was ~350k files, ~296TB
  – Other transfer was ~1.8M files, ~472TB
  – Data transfer rate peaks over 80Gbps
  – Total transfer volume ~768TB
  – Total wall clock time ~39 hours
- Current tools (Filesystems, DTNs, Globus) handle petascale data sets easily

Key Points (2)

- Sophisticated tools bring multiple benefits
  - Both transfers experienced random data corruption (multiple checksum failures) which was automatically corrected by Globus (data set is ¾ of a petabyte….a lot can happen with that much data)
    - User had to expend zero effort to find and fix the corrupted files
  - OLCF endpoint paused for a workday for normal scheduled maintenance, and transfer resumed after maintenance concluded
    - OLCF staff did not have to worry about the user when planning maintenance
    - User had to expend zero effort to work around HPC facility maintenance
  - Easy to use tools with automated fault recovery reduce human effort
- Large scale data transfer is a key enabler for scientific productivity
  - Benefits of large-scale INCITE allocation brought back to collaborators
- Normal, routine operations – no magic

Petascale DTN Project – Durable Gains

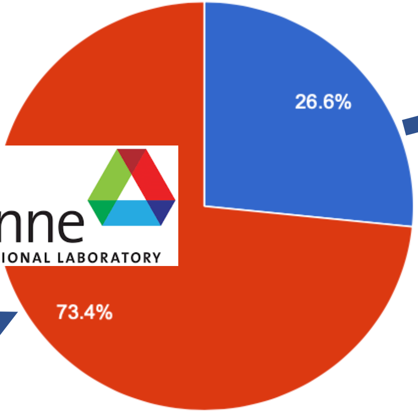


- Current performance 2x faster than 2017 numbers
- HPC facilities continue to maintain and improve capability because it is valuable to users
- Users now view this as normal, and can build new workflows on top of it

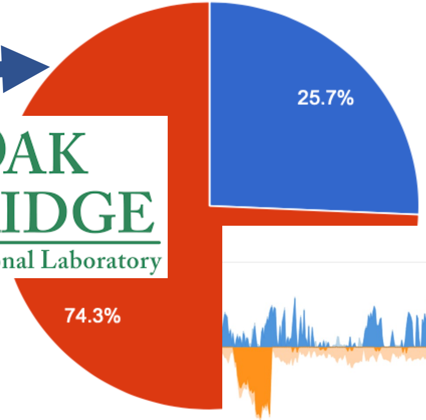
Challenge: Replicate 7+ petabytes of climate data to ANL and ORNL

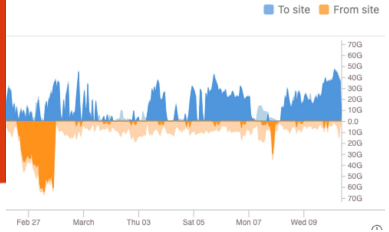
Solution: Use Globus to copy data over ESnet

1.5 GB/s



4 to 6 GB/s





Replication to ALCF

ACTIVE, PAUSED and the latest SUCCEEDED transfers

No	Datasets	From	Requested	Completed	Status	Directories	Files	Bytes Transferred	Faults	Rate
1	/css03_data/CMIP6/CMIP/MOHC/HadGEM3-GC31-LL/historical	LLNL	2022-03-10 13:19:03		ACTIVE (20%)	6125	6515	6138646980430	0	832 MB/s
2	/css03_data/CMIP6/CMIP/MIROC/MIROC-ES2L/historical	LLNL	2022-03-10 05:35:04		ACTIVE (79%)	37994	409095	24611252181300	12	699 MB/s
3	/css03_data/CMIP6/CMIP/MOHC/HadGEM3-GC31-LL/amip	LLNL	2022-03-10 12:12:03	2022-03-10 13:18:06	SUCCEEDED	3908	1892	3091419704055	0	780 MB/s
4	/css03_data/CMIP6/CMIP/MOHC/HadGEM3-GC31-LL/abrupt-4xCO2	LLNL	2022-03-10 11:40:03	2022-03-10 12:11:57	SUCCEEDED	1121	953	1559216858805	0	814 MB/s

Replication to ORNL

ACTIVE, PAUSED and the latest SUCCEEDED transfers

No	Datasets	From	Requested	Completed	Status	Directories	Files	Bytes Transferred	Faults	Rate
1	/css03_data/CMIP6/CMIP/MIROC/MIROC-ES2L/esm-piControl	ALCF	2022-03-10 15:14:03		ACTIVE (25%)	1236	40039	1407934487539	0	2.93 GB/s
2	/css03_data/CMIP6/CMIP/IPSL/IPSL-CM6A-LR/historical	ALCF	2022-03-09 22:02:03		ACTIVE (77%)	73193	36610	129503497305534	1	2.08 GB/s
3	/css03_data/CMIP6/CMIP/MIROC/MIROC-ES2L/esm-hist	ALCF	2022-03-10 14:51:04	2022-03-10 15:13:24	SUCCEEDED	3706	39663	2973432261868	0	2.22 GB/s
4	/css03_data/CMIP6/CMIP/MIROC/MIROC-ES2L/amip	ALCF	2022-03-10 14:47:03	2022-03-10 14:50:22	SUCCEEDED	3126	12284	446324011629	0	2.25 GB/s

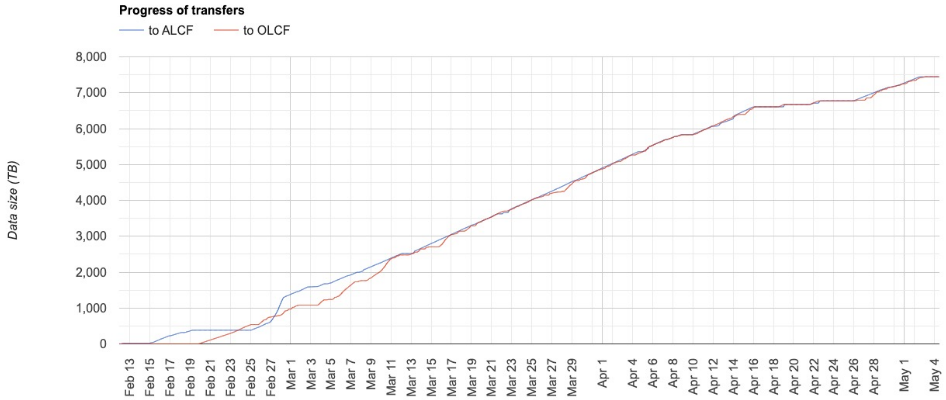
https://dashboard.globus.org/esgf

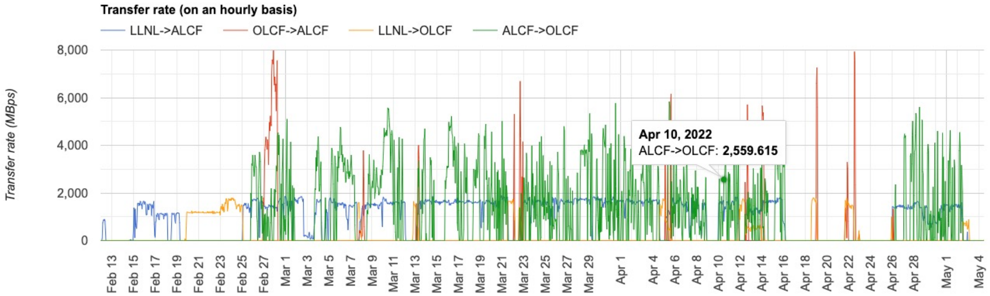
Slide credit: Ian Foster, ANL

As of March 10, 2022

Climate Data Replication

- Transfers mostly LLNL→ANL→ORNL except during maintenance windows
- Overall end2end performance limted on LLNL side
  – Storage system config
  – Keep that one working, sync between the other two
- This is a good reason to deploy capable DTNs and storage systems
  – Technical capabilities define the boundaries of the possible
  – If your users can't reason about something, they won't ask you for it

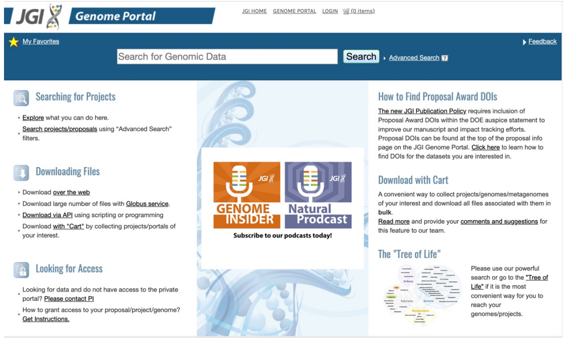


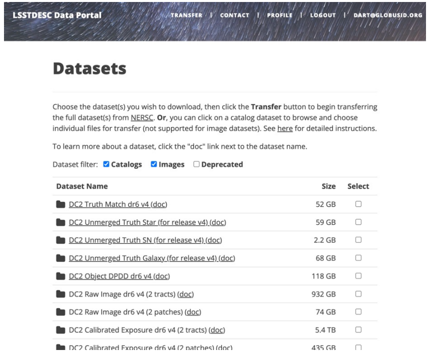


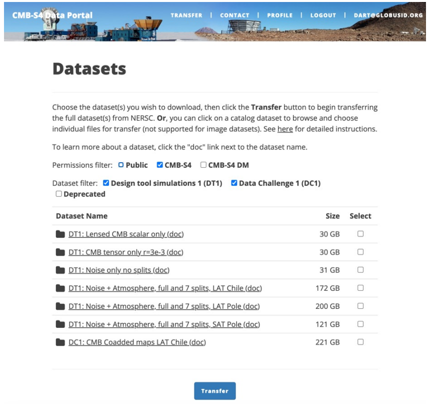
https://dashboard.globus.org/esgf

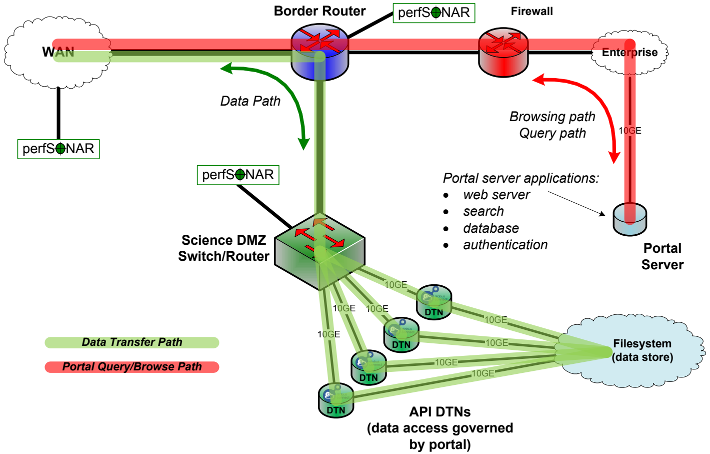
ESnet

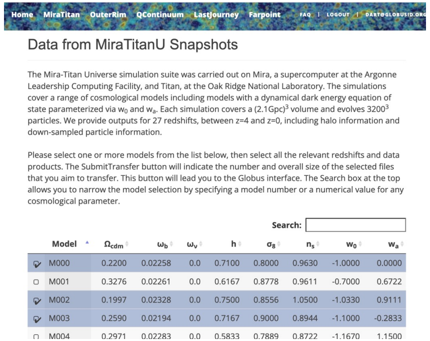
Data Portals Increasing Adoption









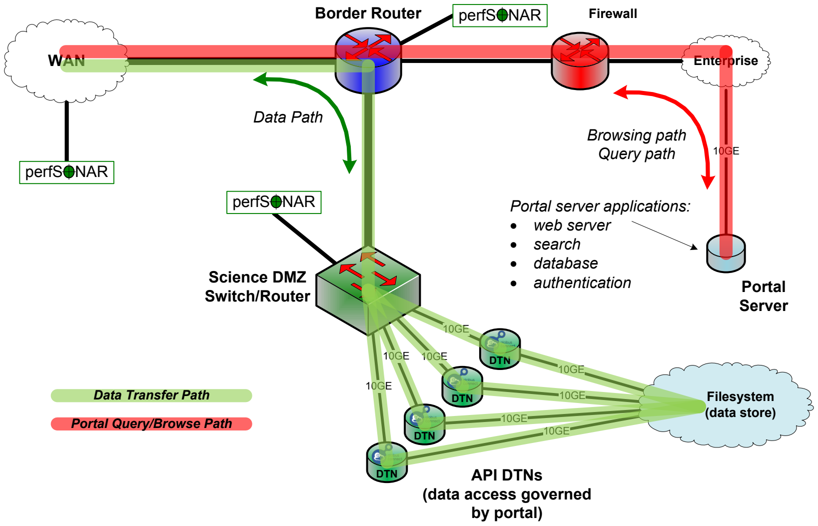


ESnet

https://peerj.com/articles/cs-144/

Data Portals Increasing Adoption

- Data publication and access built on the Science DMZ foundation
- Multiple fields: Astronomy, Cosmology, Genomics, Climate
- Science DMZ and DTNs support data search, publication, access
- Typically classical file-based access, but important to scientific communities for navigating large repositories



https://peerj.com/articles/cs-144/

Integrating Many Science DMZs

- Instead of building one Science DMZ for one service, some communities have built one service from many Science DMZs
- Use the DMZ and a DTN as a building block
  - Many DTNs working together result in a distributed capability
  - Higher-layer software stack may be more than just data transfer
- Example: PRP and NRP
  - DTN + GPUs
  - Data-centric distributed computing capability

2015 Vision: The Pacific Research Platform Will Connect Science DMZs Creating a Regional End-to-End Science-Driven Community Cyberinfrastructure

NSF CC*DNI Grant
$7.3M 10/2015-10/2020

PACIFIC RESEARCH PLATFORM

Source: John Hess, CENIC

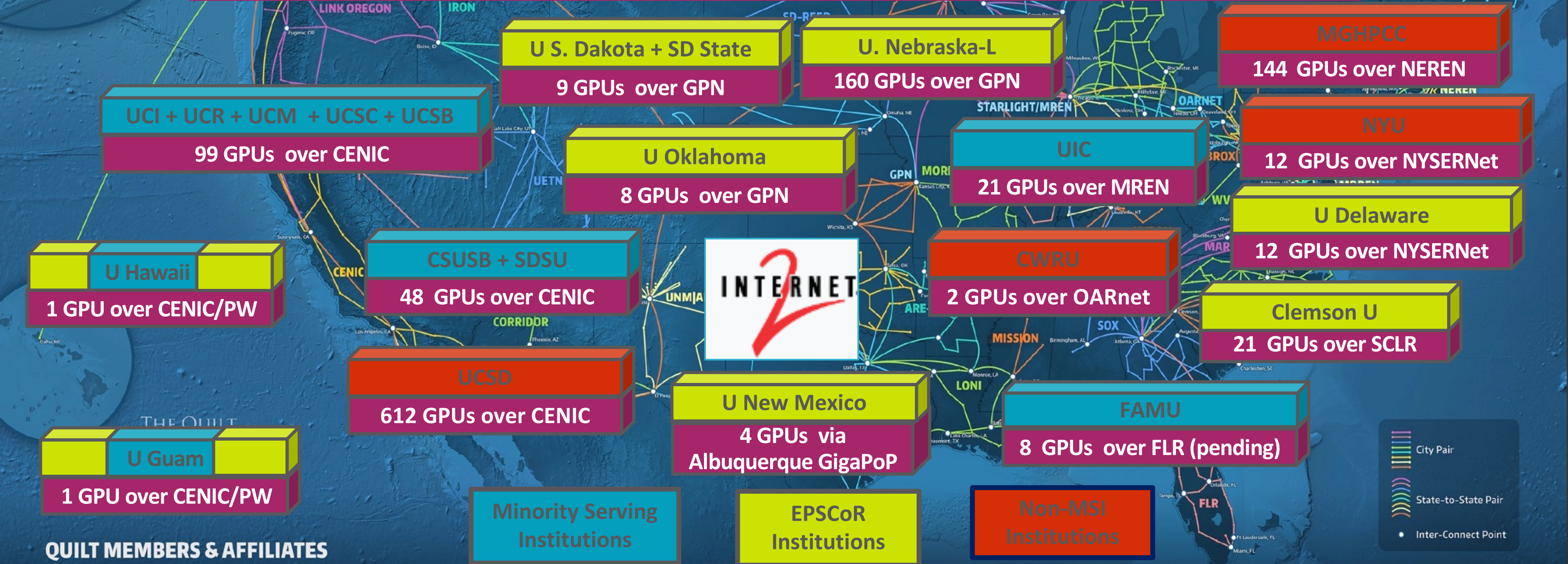
PACIFIC RESEARCH PLATFORM

CITRIS AND THE BANATAO INSTITUTE

SDSC

Nautilus ~1,100 GPUs Distributed over US Networks—Fall 2022

(New: NSF PNRP's 352 GPUs, and ARO/DURIP's 144 GPUs and SDSU's 4 DGX A100s)

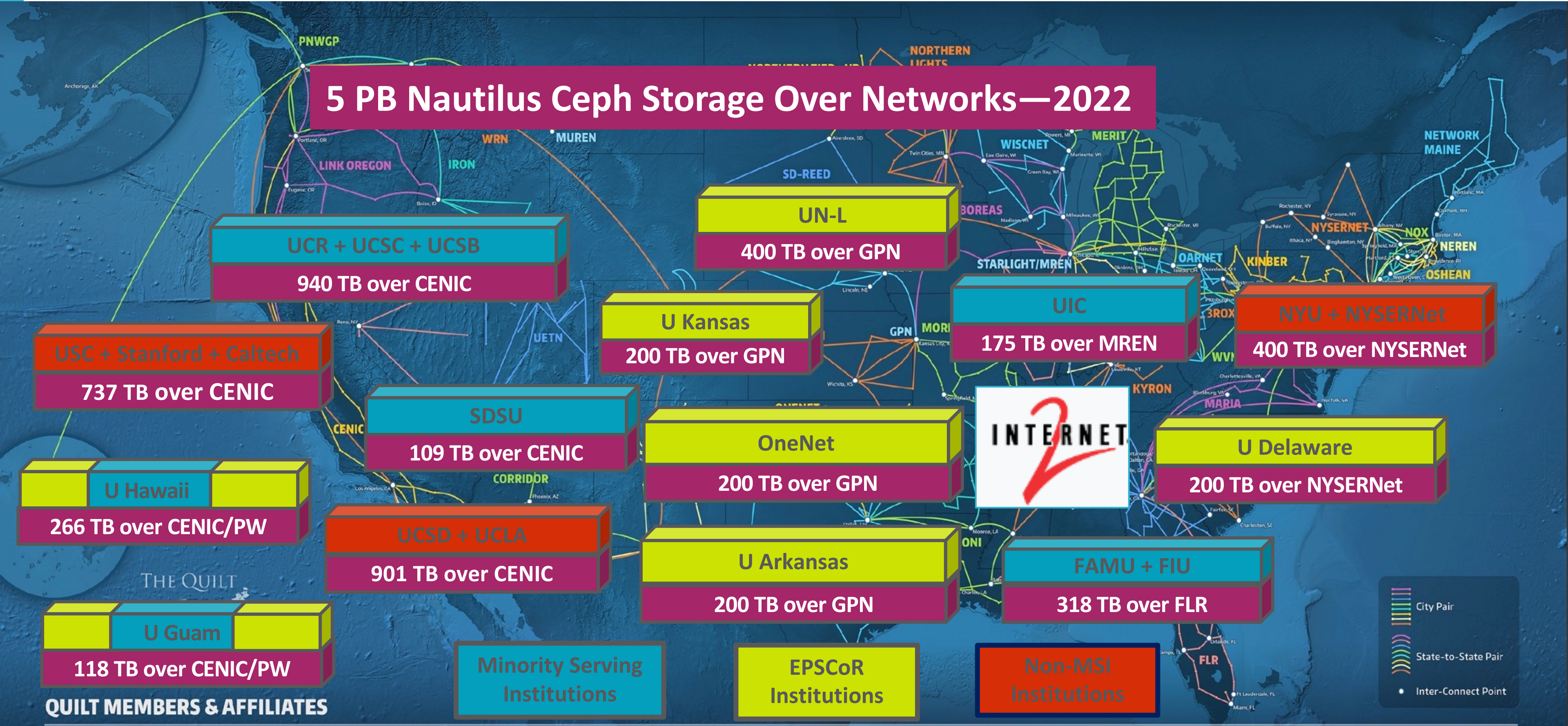


QUILT MEMBERS & AFFILIATES

Slide Credit:

Tom DeFanti, UCSD

5 PB Nautilus Ceph Storage Over Networks—2022



UCR + UCSC + UCSB
940 TB over CENIC

UN-L
400 TB over GPN

USC + Stanford + Caltech
737 TB over CENIC

U Kansas
200 TB over GPN

UIC
175 TB over MREN

NYU + NYSERNet
400 TB over NYSERNet

SDSU
109 TB over CENIC

OneNet
200 TB over GPN

INTERNET2

U Delaware
200 TB over NYSERNet

U Hawaii
266 TB over CENIC/PW

UCSD + UCLA
901 TB over CENIC

U Arkansas
200 TB over GPN

FAMU + FIU
318 TB over FLR

U Guam
118 TB over CENIC/PW

Minority Serving Institutions

EPSCoR Institutions

Non-MSI Institutions

QUILT MEMBERS & AFFILIATES

3ROX · ARE-ON · Light · CAAREN · CEN · CENIC · FRGP · GPN · I-Light · Indiana GigaPOP · Illinois Century Network · KANREN · KINBER · KyRON · LEARN · LINK OREGON
LONI · MaRIA · MAX · MCNC · MDREN · merit · MOREnet · MREN · NCREN · NETWORKMAINE · Nevada Net · NJEdge
Northern Lights · nysernet · OARnet · OneNet · OSHEAN · Pacific Northwest GigaPOP · SoX · Sun Corridor Network · uen · WiscNet · WVNET

Slide Credit:

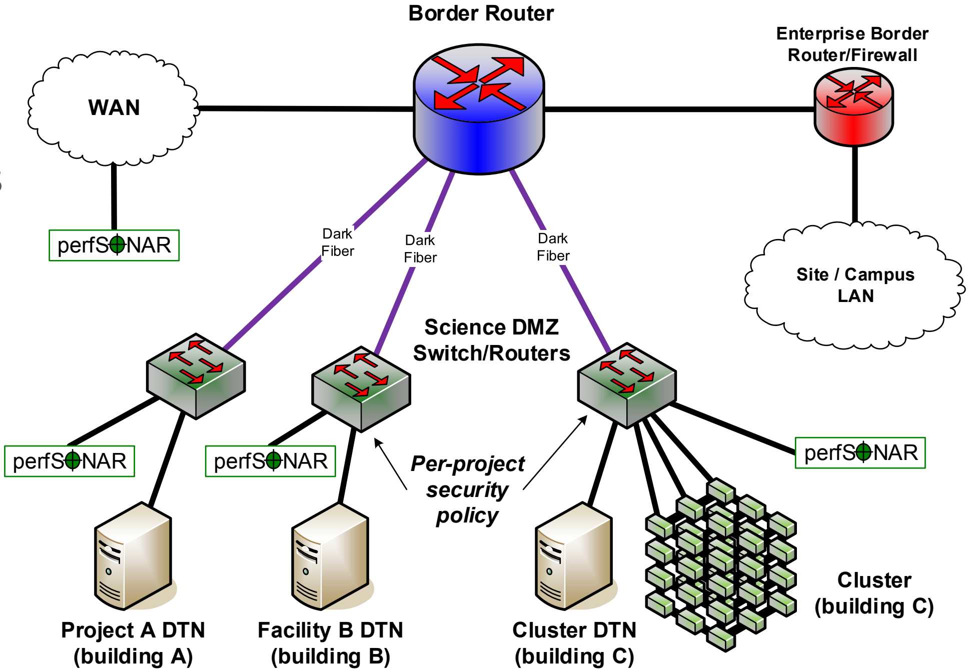
Tom DeFanti, UCSD

Streaming DTNs: Near Real Time Data Processing

- File movement is great for many applications, but not all
  - Fast feedback for experiment guidance
  - Integration of detectors and computing
- In some cases, filesystem I/O is too slow
- In other cases, file semantics aren't a good fit
  - Message passing is more appropriate in some workflows
  - Files may be present in the workflow, but only after a streaming pipeline

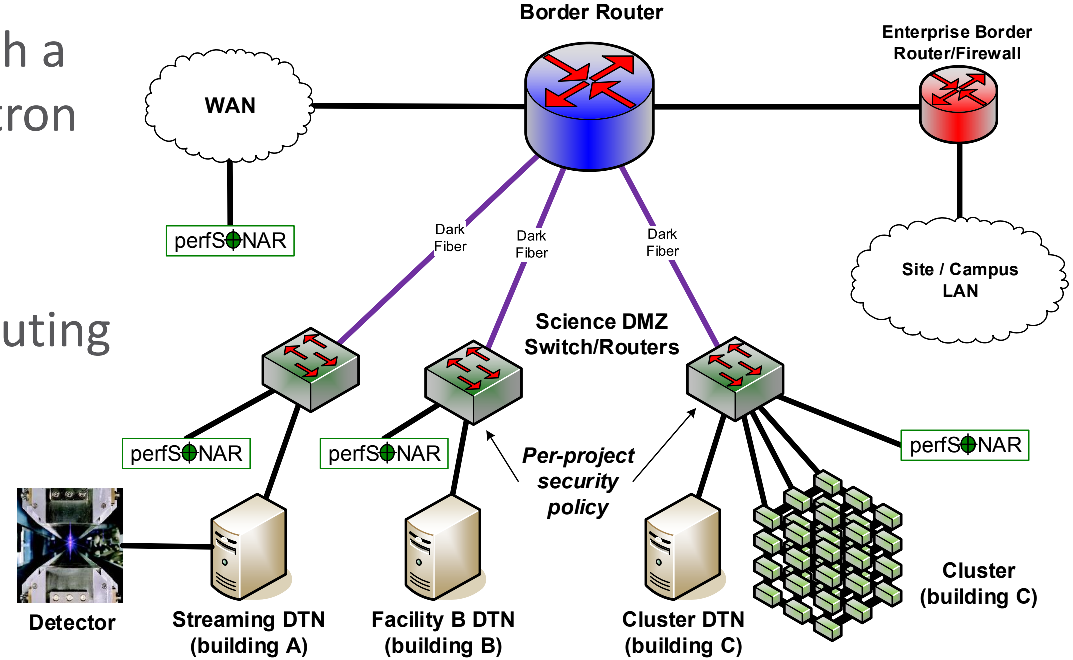
Streaming DTN – Just Another Flavor

- In many cases, a DTN is deployed to support file transfer
  - Large data sets between filesystems
  - Well-understood workflow
- Can have many DTNs in a single Science DMZ, or multiple Science DMZs as appropriate
- Let's put Project A at an experimental facility



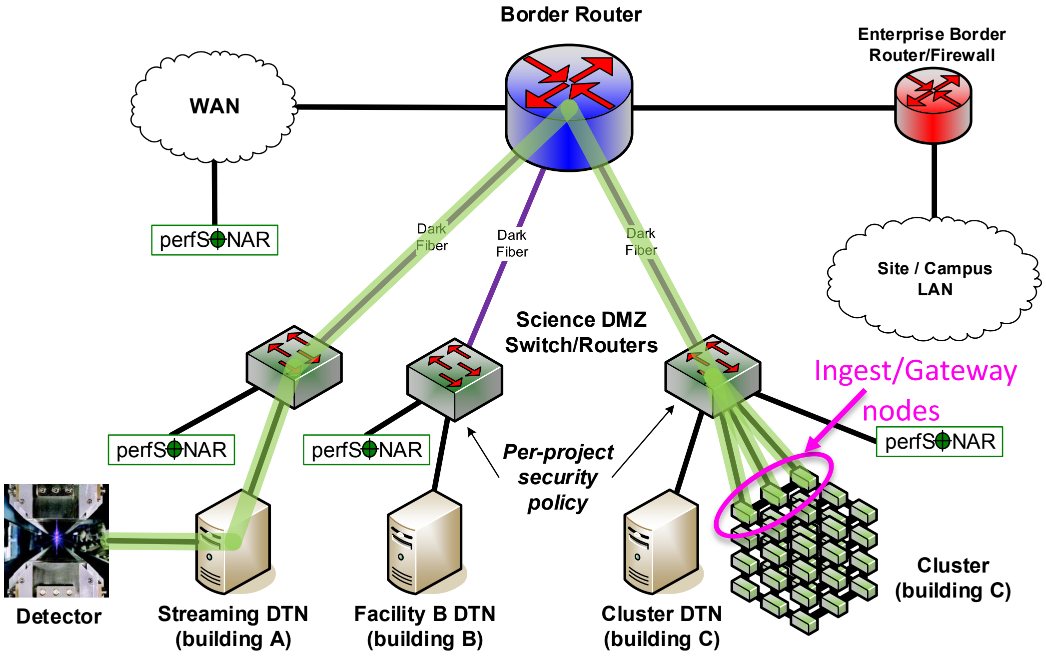
Common Case – DAQ for a Detector

- Project A is an experiment with a high-speed detector (synchrotron light source, Cryo-EM, etc.)
- Rapid analysis for experiment feedback requires more computing than can be deployed at the experiment
- Stream the data to remote compute resource



Streaming Data Path

- High performance data path requires consistent behavior, low packet loss, etc.
- Perfect fit for Science DMZ performance engineering
- Stream to local or remote computing depending on application
  - (local shown here because it all fits in one diagram)



Stream Processing in Light Source Facilities

A Science Driver

Experiment

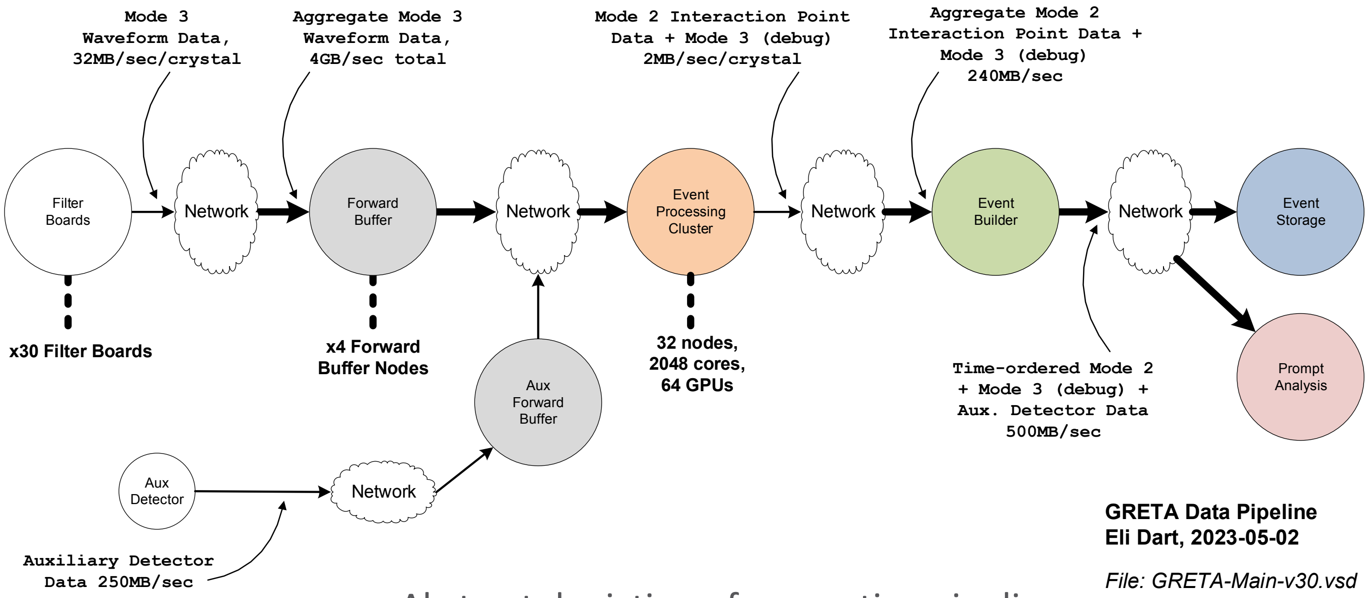
Data Analysis

Transfer data

Decision/
Steer

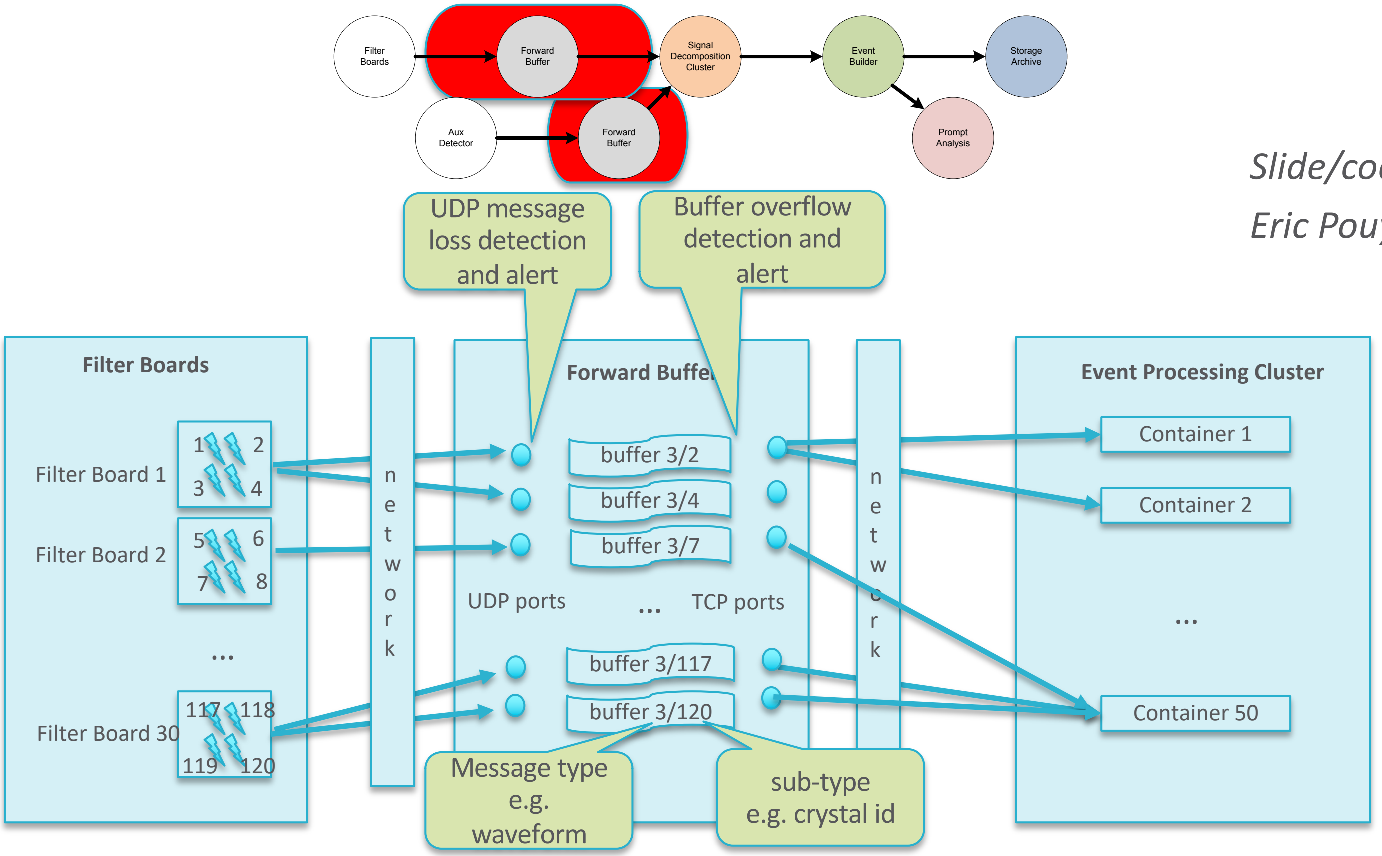
Slide Credit: Raj Kettimuthu, ANL – see next talk!

GRETA Computing/Data Pipeline



- Abstract depiction of computing pipeline, including device counts and data rates
- Computing pipeline serves as a platform for data processing, from detector electronics on the left to visualization and storage on the right

Forward Buffer Architecture



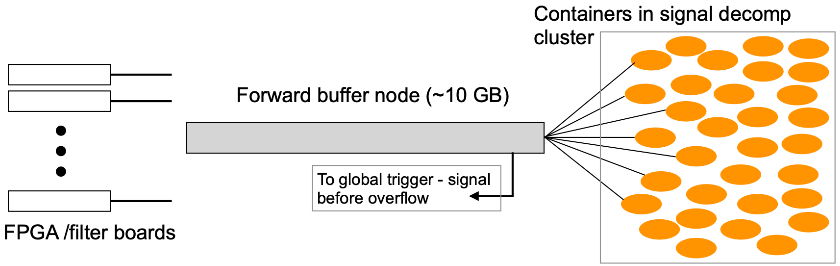
Slide/code credit:

Eric Pouyoul, ESnet

ESnet

Forward Buffer Design Pattern

- Generalized forward buffer design under development
- Use GRETA pipeline as a starting point, remove GRETA-specific aspects
- Goal: re-usable code to couple UDP detector readout to self scheduling inline computing



Forward buffer performance: see *Simple and Scalable Streaming: The GRETA Data Pipeline, EPJ Web of Conferences 251, 04018 (2021)*

Message size	Maximum message rate (Single stream)	Maximum streams at 20k messages / sec
100 bytes	≥ 400k*	50
500 bytes	≥ 400k*	50
1 kB	270k	50
5 kB	210k	50
8 kB	172k	50
8.8 kB	170k	50
9.2 kB	110k	20
20 kB	80k	18
40 kB	50k	16
64 kB	28k	*

Streaming DTNs

- There are more applications of streaming DTNs
  - FPGA-based devices to distribute detector output to computing (EJFAT)
  - More to come, I'm sure
- Need Science DMZ capabilities
  - High bandwidth
  - Specific security
- Difficult to deploy and operate in an environment without Science DMZ

Science DMZ Security

- Goal – disentangle security policy and enforcement for science flows from security for business systems
- Rationale
  - Science data traffic is simple from a security perspective
  - Narrow application set on Science DMZ
  - Traditional perimeter security is a poor fit for high performance flows
- Separation allows each to be optimized
  - Key point: the Science DMZ is an example of segmentation for a specific purpose, and with specific security policy
  - Separate Science DMZ from the perimeter for security and performance reasons

Zero Trust – The Demise Of The Perimeter

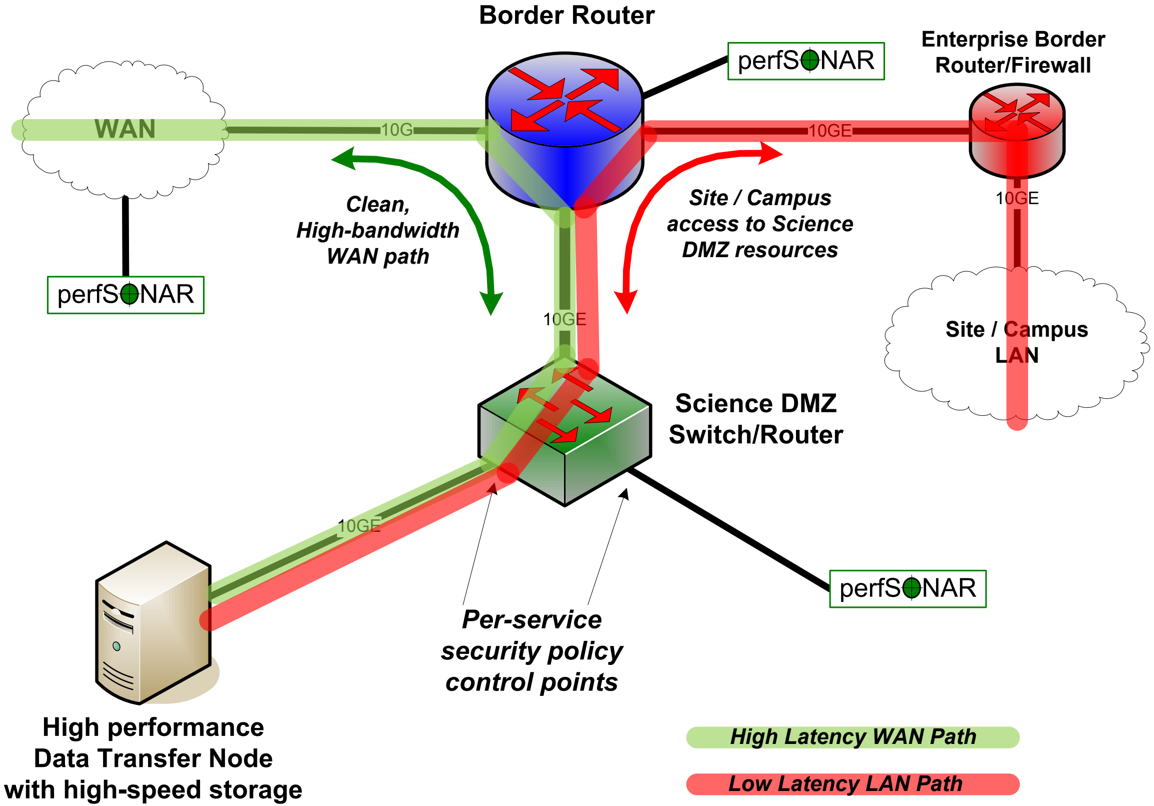
- No implied trust based on network location
  - “Network location” means IP address and physical topology
  - ”Inside the firewall” does not confer additional trust
- Access control is as granular as possible
  - “Bulk” security at the perimeter → per-service policy + filter
  - Policy enforcement close (in topology) to running service
  - Focus on authn, authz, reducing zones of implied trust
- What does this mean for Science DMZ?

ESnet

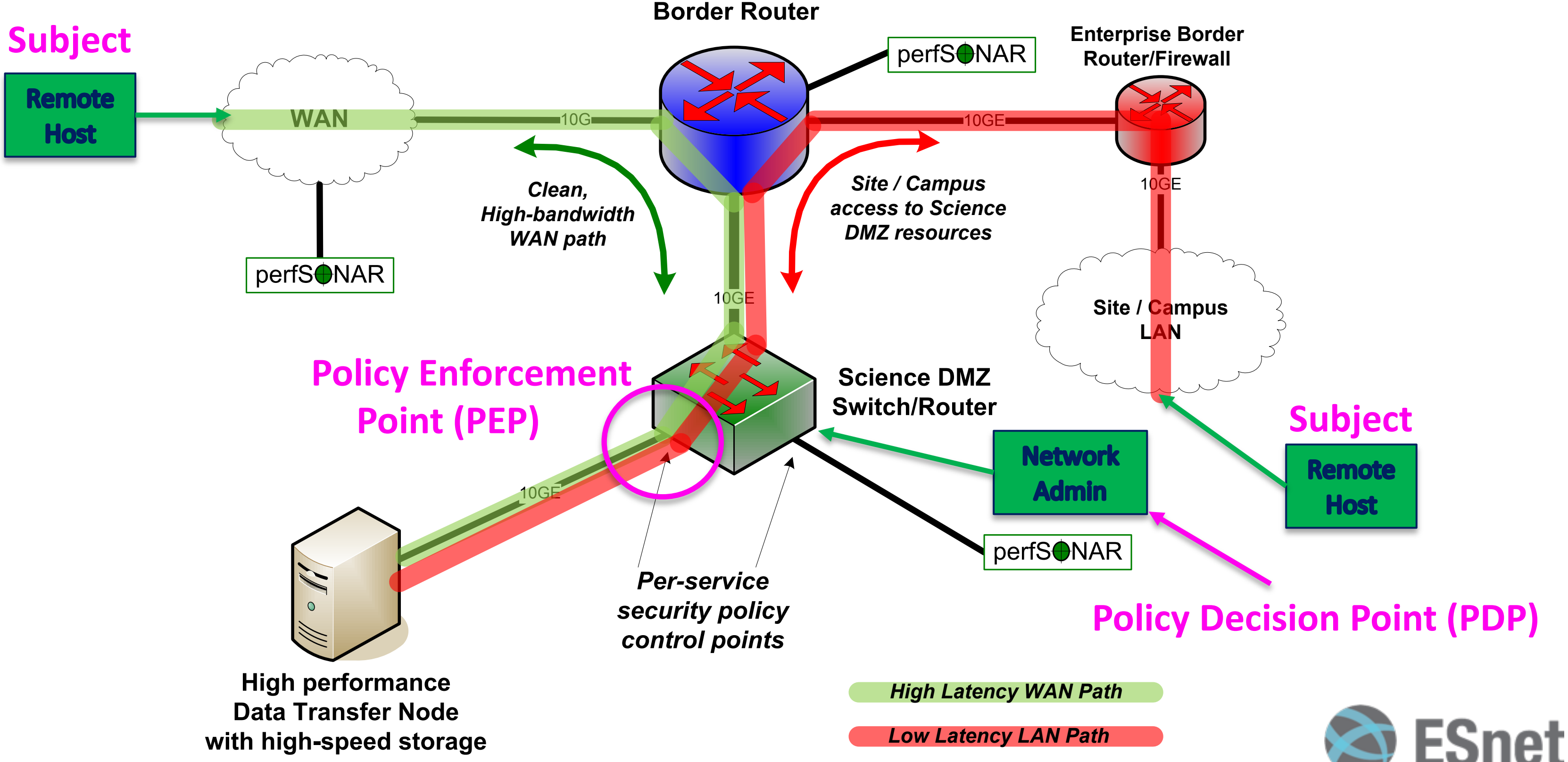
Zero Trust Elements

- Multiple information sources
  - NIST 800-207
  - CISA Zero Trust Maturity Model
  - Others
- A Subject accesses a Resource
  - Permission granted (or not) by Policy Decision Point (PDP)
    - Policy Engine
    - Policy Administrator
  - Policy is enforced by Policy Enforcement Point (PEP)
    - Path between Subject and PEP is untrusted
    - Path between PEP and Resource is trusted
- Micro-segmentation, per-session authentication are key ideas

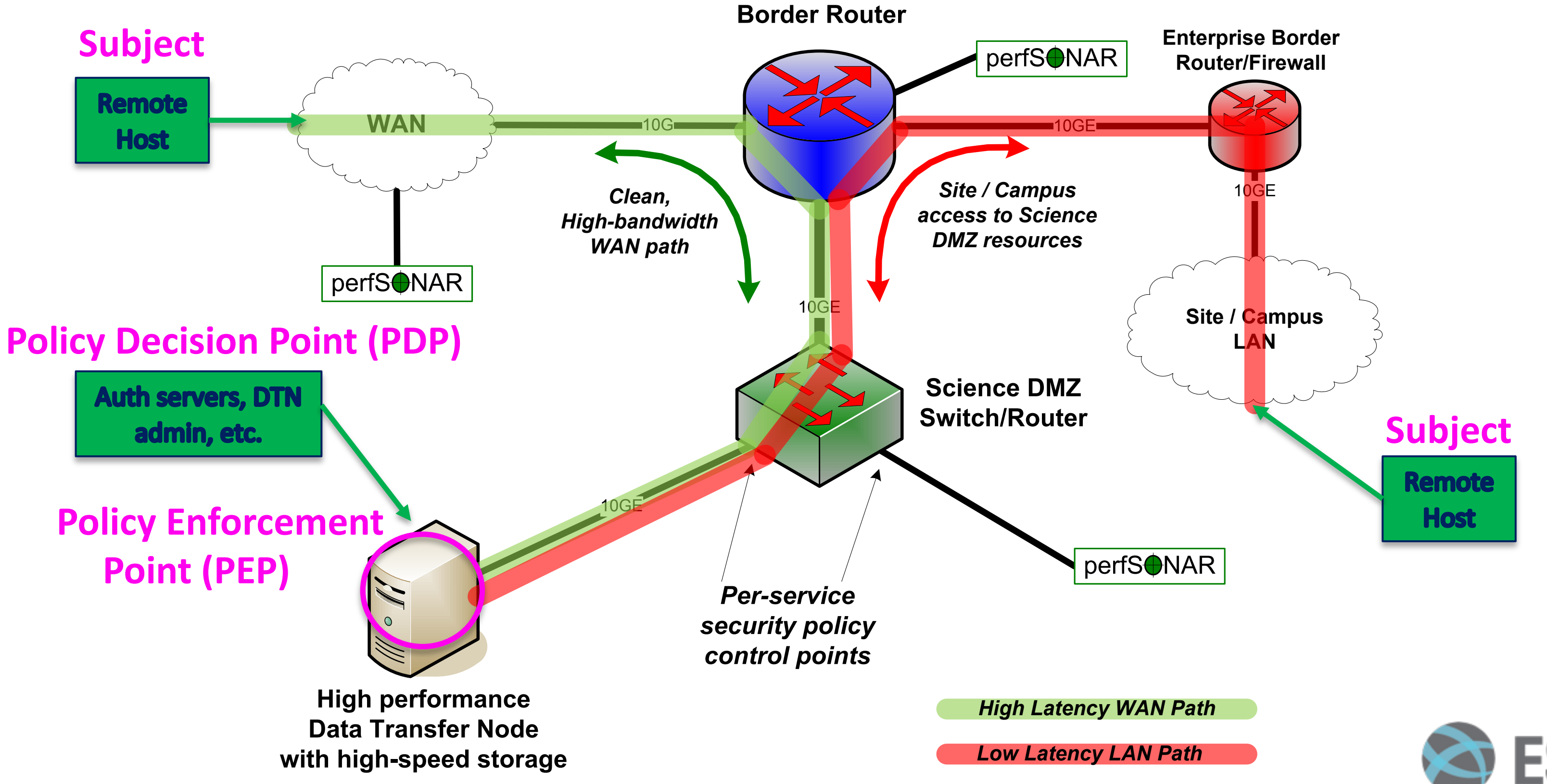
Science DMZ With Zero Trust Labeling



Science DMZ With Zero Trust Labeling – Network Layer



Science DMZ With Zero Trust Labeling – Application Layer



Zero Trust – Segmentation Matters

- The Science DMZ model already brings security policy enforcement close to the service
  - For example, we put DTN filters on the Science DMZ switch or router if possible, not at the perimeter
  - Segmentation by function is relevant to both security and performance
  - Micro-segmentation is a good fit for Science DMZ (and we do this already)
- Network layer filters are still relevant, even as more focus is given to the application
  - Layer 3 filter is a network-layer policy decision on what remote hosts to serve
  - Layer 4 filter is a network-layer policy decision on permitted DTN services
  - Most of Zero Trust works at the application layer (users, roles, data, etc)

Zero Trust and Science DMZ

- Transition to Zero Trust will affect DTN applications more than DMZ networks
  - Science DMZ itself is important for performance and security
  - Application stack needs modern authn, authz
    - Affects application choices and configuration
    - Performance requirements are unaffected by auth mechanisms
- Topology is important for performance
  - Data plane capabilities, reduction of complexity, etc.
  - Important to retain data plane improvements as we improve application and data security using Zero Trust
- Science DMZ is inherently Zero Trust compatible
- Iterative deployment/improvement is still key

ESnet

Science DMZ – Moving Forward

- The Science DMZ is part of the data ecosystem
- Solid foundation for building higher-level services and capabilities
  - Many communities have made it their own
  - Keep what you need, add new things, whatever works for the science
- File-based workflows continue to be important
  - Data transfer
  - Data portals
- Streaming is a new and growing area
- Zero Trust is changing security – Science DMZ is a good fit

ESnet
ENERGY SCIENCES NETWORK

Thanks!

Energy Sciences Network (ESnet)

Lawrence Berkeley National Laboratory

http://fasterdata.es.net/

http://my.es.net/

http://www.es.net/

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB