ESnet

ENERGY SCIENCES NETWORK

The Data Mobility Exhibition

Eli Dart, Science Engagement

Energy Sciences Network (ESnet)

Lawrence Berkeley National Laboratory

CC* PI meeting

Minneapolis, MN

September 24, 2019

U.S. DEPARTMENT OF ENERGY

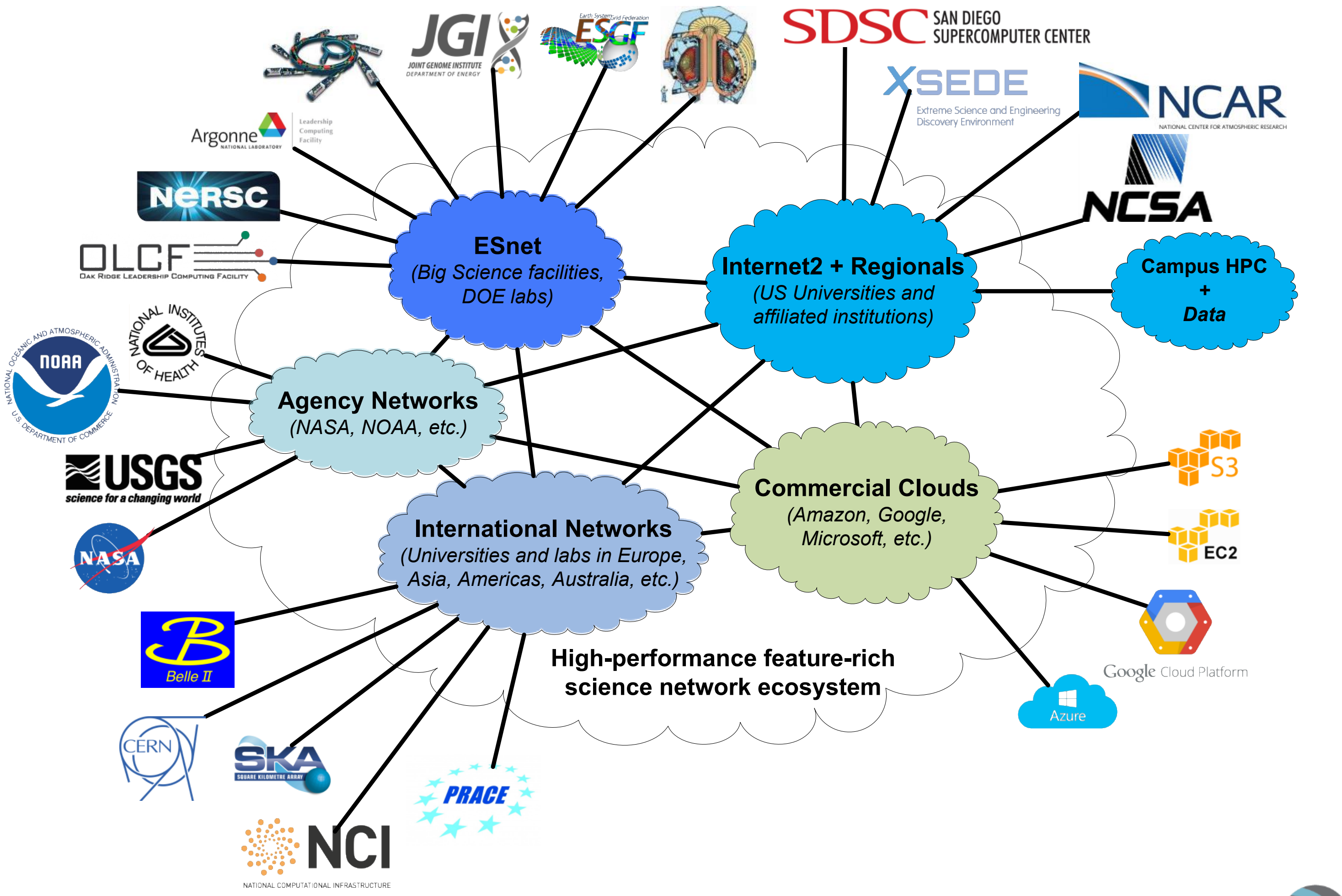
Office of Science

BERKELEY LAB

Our Community Has Made Great Progress

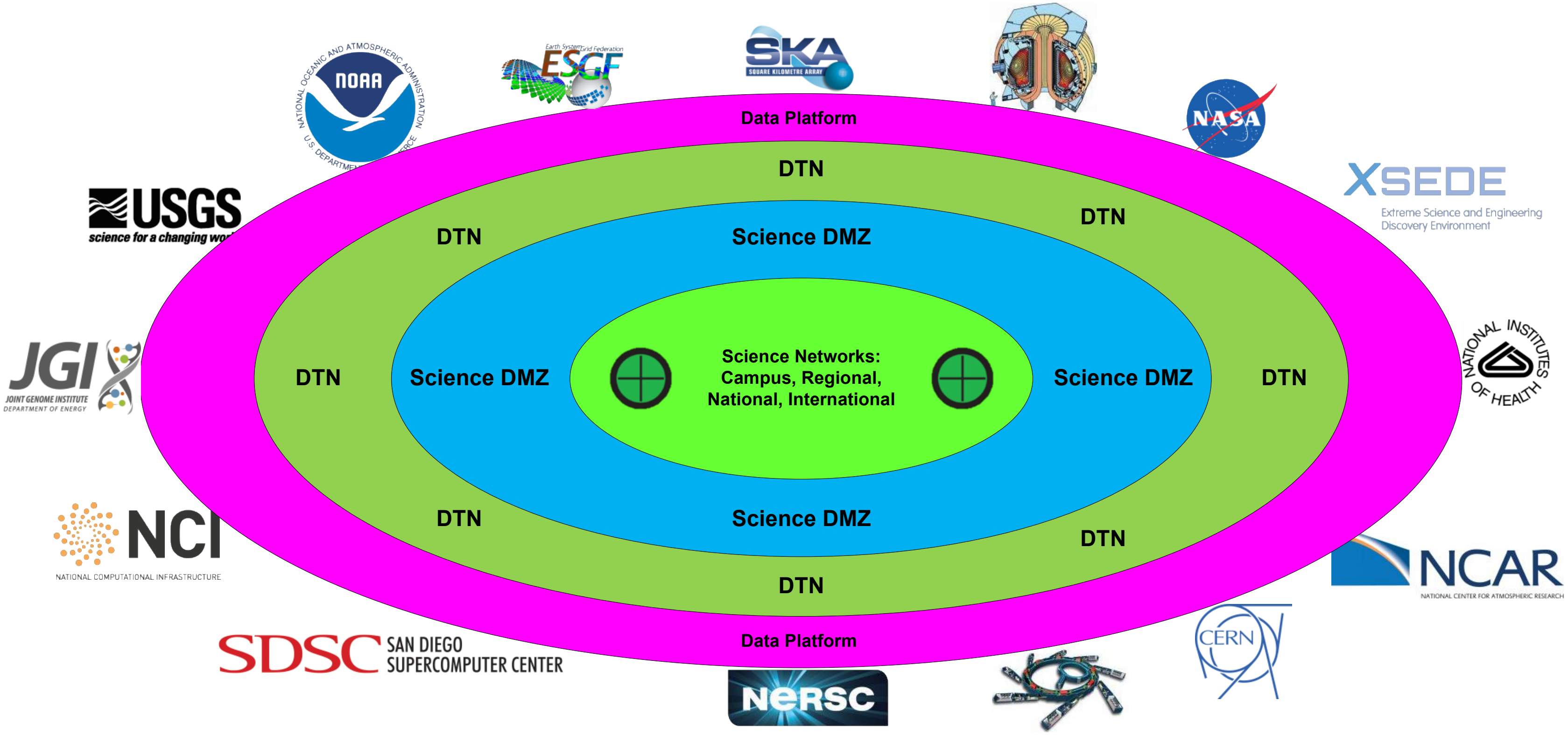
- Over the past 8-10 years or so, we have made great strides forward
- Science networks are big, fast, and clean
- Science networks are instrumented for performance
- The Science DMZ model is widely deployed
- DTNs in the Science DMZs
- Data orchestration platforms running on the DTNs

Data Ecosystem - Abstract Network Diagram


ESnet (Big Science facilities, DOE labs)
Internet2 + Regionals (US Universities and affiliated institutions)
Campus HPC + Data
Agency Networks (NASA, NOAA, etc.)
Commercial Clouds (Amazon, Google, Microsoft, etc.)
International Networks (Universities and labs in Europe, Asia, Americas, Australia, etc.)
High-performance feature-rich science network ecosystem
JGI Joint Genome Institute Department of Energy
ESGF Earth System Grid Federation
SDSC San Diego Supercomputer Center
XSEDE Extreme Science and Engineering Discovery Environment
NCAR National Center for Atmospheric Research
NCSA
Argonne National Laboratory Leadership Computing Facility
NERSC
OLCF Oak Ridge Leadership Computing Facility
National Institutes of Health
NOAA National Oceanic and Atmospheric Administration U.S. Department of Commerce
USGS science for a changing world
NASA
Belle II
CERN
SKA Square Kilometre Array
NCI National Computational Infrastructure
PRACE
S3
EC2
Google Cloud Platform
Azure

ESnet

Data Ecosystem – Non-Network View



What Remains To Be Done?

- We’re getting close to completing a transformation of Cyberinfrastructure
- Three major tasks remain
  - Data orchestration across Science DMZs
    - This includes test, verification, and performance engineering
    - Partially complete today
  - Upgrading the data portals - Modern Research Data Portal
    - This has begun – lots left to do
  - Onboarding scientists and collaborations
    - Science Engagement
    - We understand it, but we need to scale it
- Remember – this has to be useful to scientists, so it has to work for them

Data Movement Exhibition

- Current and previous CC* Awardees, along with the greater R&E community, are encouraged to participate
- Will be highlighted at the 2019 and 2020 CC* PI Meetings
- Using reference data sets, and existing campus CI components, participants will work on scientific data movement capabilities:
  - Download/transfer data
  - Measure performance
  - Potentially improve
- This event has begun (August 2019), and will extend for a full year (through the CC* 2020 PI meeting)

Data Movement Exhibition

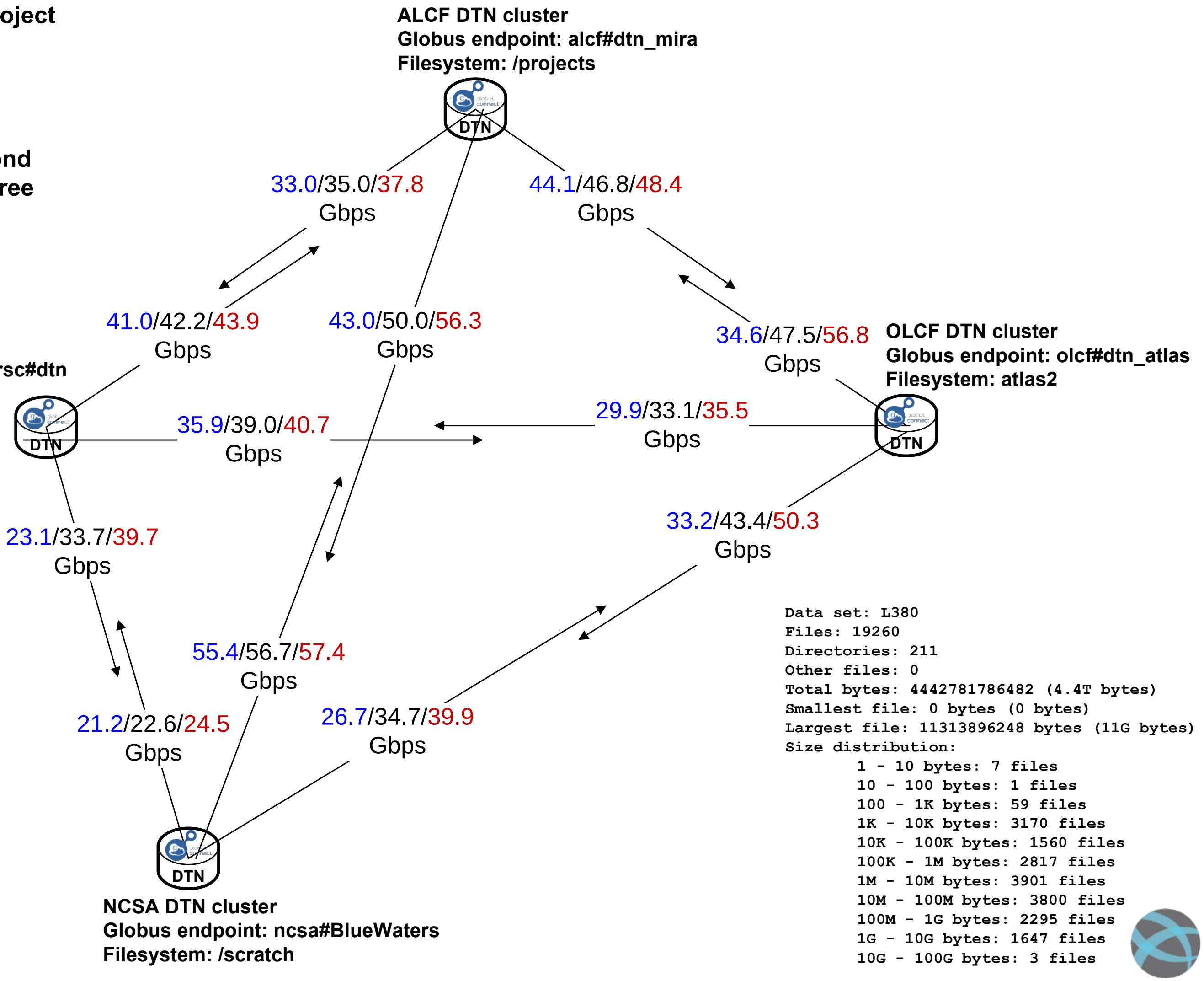
- Basic Idea:
  - Create a brief (1-2 page) description of the network and data architecture for your campus environment
  - Prepare a local data transfer machine, and get Globus working
    - Hopefully this is a production DTN in your Science DMZ
    - Even better: front-end for campus HPC or other production CI
    - Even better: DTNs for a data portal – can we download from you?
  - Download the reference data sets (specific instructions forthcoming)
  - Share the description, and your results (specific instructions forthcoming)
  - For those that want to accelerate their results, 1:1 assistance, via the Engagement and Performance Operations Center (EPOC), is available: epoc@iu.edu

National Facility Endpoints (Petascale DTN)

Petascale DTN Project

November 2017
L380 Data Set

Gigabits per second (min/avg/max), three transfers



ALCF DTN cluster
Globus endpoint: alcf#dtn_mira
Filesystem: /projects

NERSC DTN cluster
Globus endpoint: nersc#dtn
Filesystem: /project

OLCF DTN cluster
Globus endpoint: olcf#dtn_atlas
Filesystem: atlas2

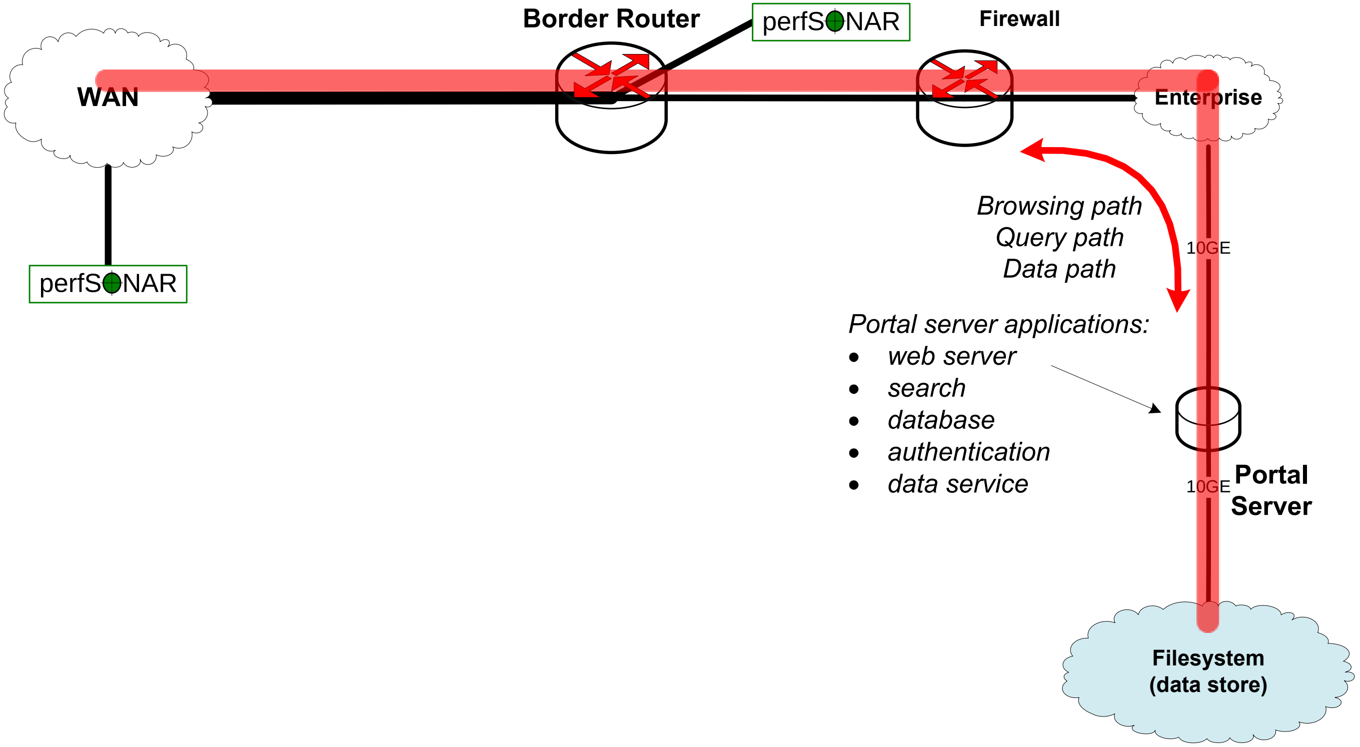
NCSA DTN cluster
Globus endpoint: ncsa#BlueWaters
Filesystem: /scratch

```
Data set: L380
Files: 19260
Directories: 211
Other files: 0
Total bytes: 4442781786482 (4.4T bytes)
Smallest file: 0 bytes (0 bytes)
Largest file: 11313896248 bytes (11G bytes)
Size distribution:
        1 - 10 bytes: 7 files
        10 - 100 bytes: 1 files
        100 - 1K bytes: 59 files
        1K - 10K bytes: 3170 files
        10K - 100K bytes: 1560 files
        100K - 1M bytes: 2817 files
        1M - 10M bytes: 3901 files
        10M - 100M bytes: 3800 files
        100M - 1G bytes: 2295 files
        1G - 10G bytes: 1647 files
        10G - 100G bytes: 3 files
```

Science Data Portals

- Large repositories of scientific data
  - Climate data
  - Sky surveys (astronomy, cosmology)
  - Many others
  - Data search, browsing, access
- Make the data easily accessible on HPC platforms
  - Supercomputers
  - Campus HPC
  - Clouds
- This will feed the rising capabilities of AI/ML and other data analytics

Legacy Portal Design

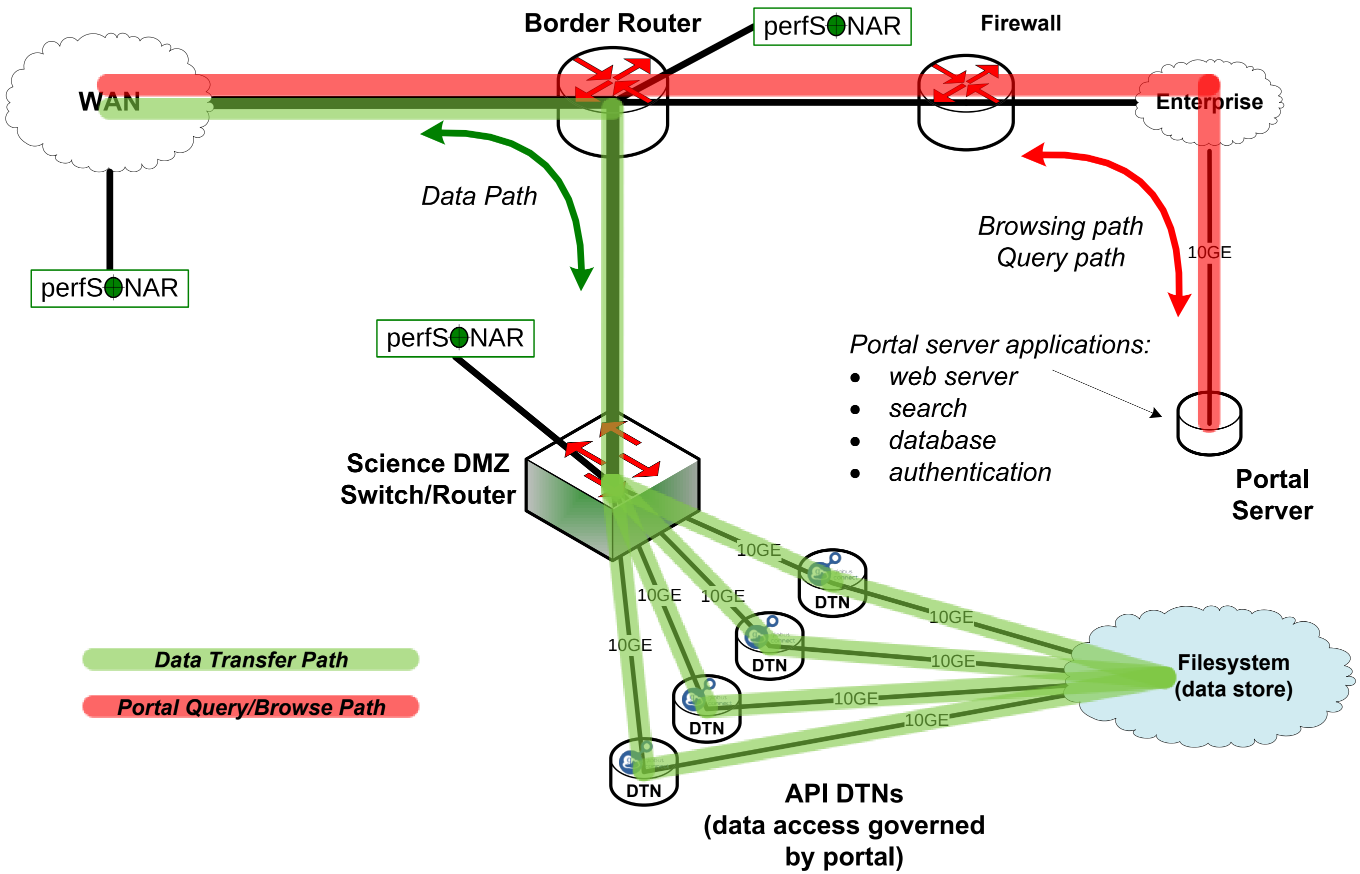


Browsing path
Query path
Data path

Portal server applications:

- *web server*
- *search*
- *database*
- *authentication*
- *data service*

Modern Research Data Portal Leverages Science DMZ

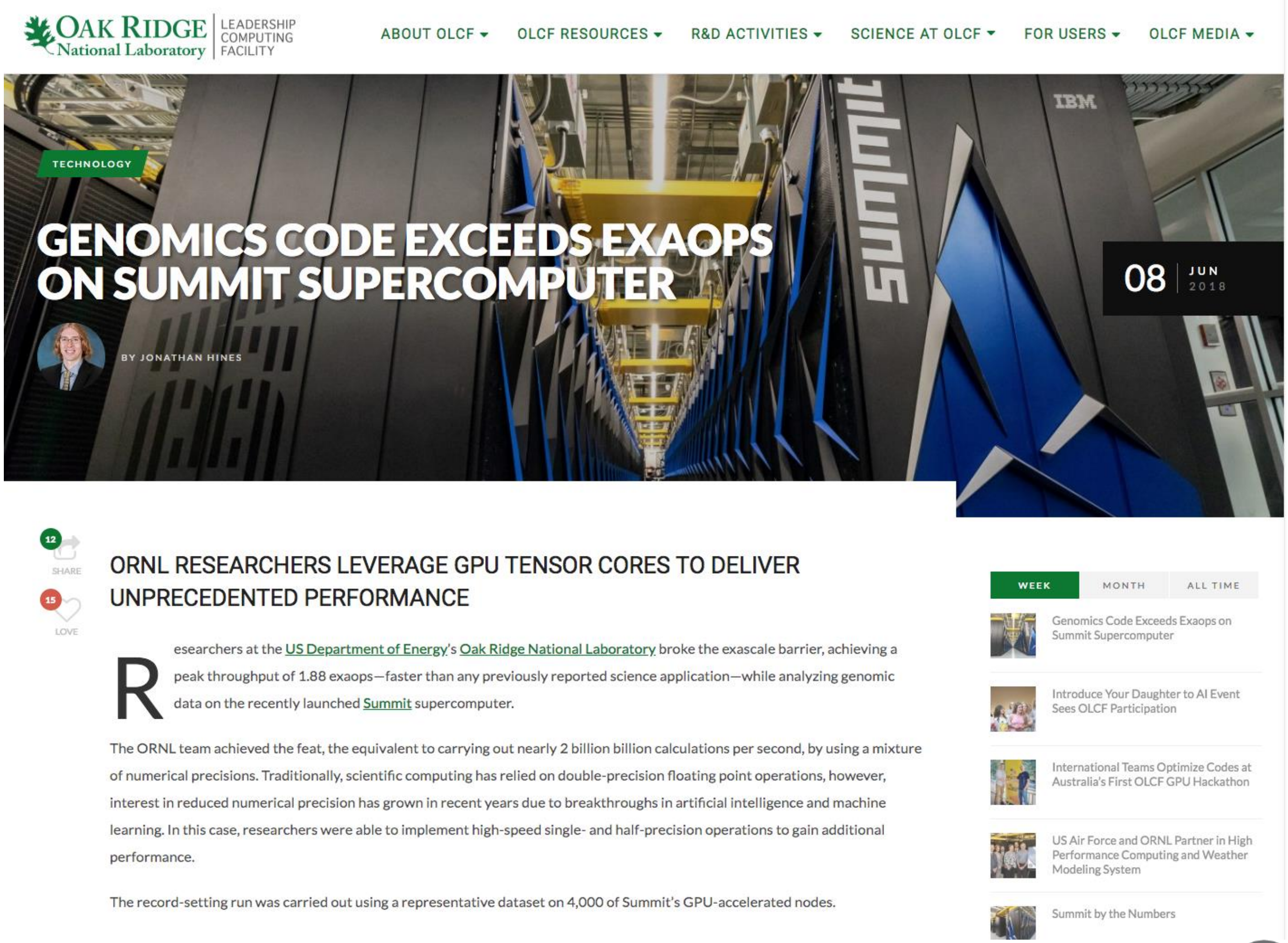


Portal server applications:

- *web server*
- *search*
- *database*
- *authentication*

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

Science at Scale: Genomics (June 2018)



GENOMICS CODE EXCEEDS EXAOPS ON SUMMIT SUPERCOMPUTER

BY JONATHAN HINES

08 JUN 2018

ORNL RESEARCHERS LEVERAGE GPU TENSOR CORES TO DELIVER UNPRECEDENTED PERFORMANCE

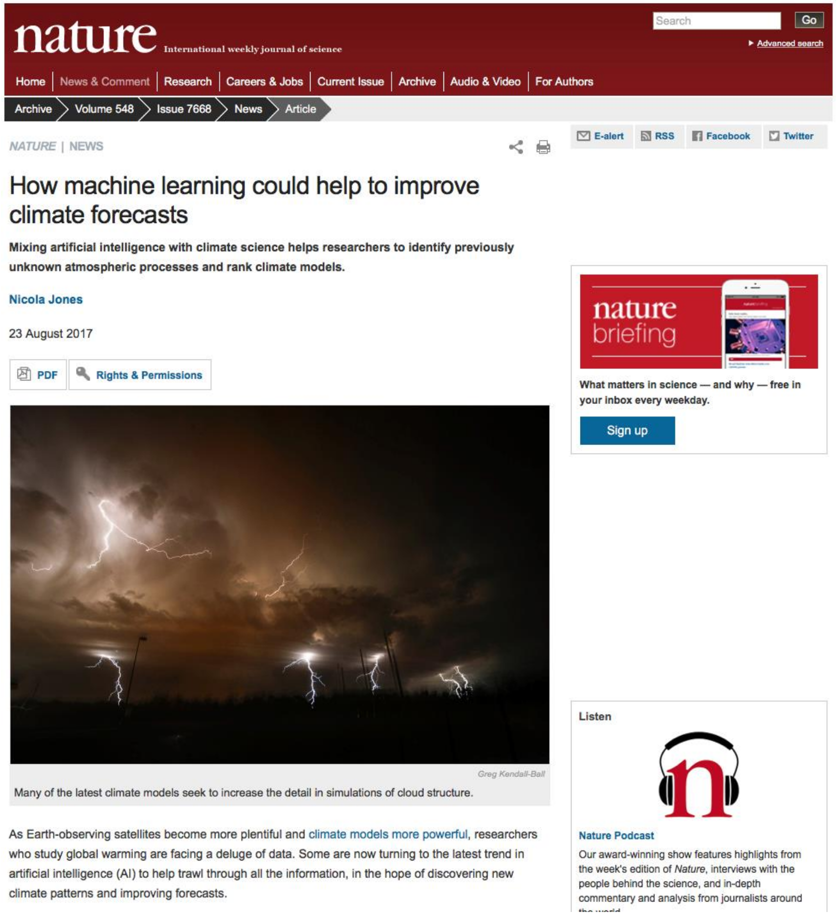
Researchers at the US Department of Energy's Oak Ridge National Laboratory broke the exascale barrier, achieving a peak throughput of 1.88 exaops—faster than any previously reported science application—while analyzing genomic data on the recently launched Summit supercomputer.

The ORNL team achieved the feat, the equivalent to carrying out nearly 2 billion billion calculations per second, by using a mixture of numerical precisions. Traditionally, scientific computing has relied on double-precision floating point operations, however, interest in reduced numerical precision has grown in recent years due to breakthroughs in artificial intelligence and machine learning. In this case, researchers were able to implement high-speed single- and half-precision operations to gain additional performance.

The record-setting run was carried out using a representative dataset on 4,000 of Summit's GPU-accelerated nodes.

- Genomics Code Exceeds Exaops on Summit Supercomputer
- Introduce Your Daughter to AI Event Sees OLCF Participation
- International Teams Optimize Codes at Australia's First OLCF GPU Hackathon
- US Air Force and ORNL Partner in High Performance Computing and Weather Modeling System
- Summit by the Numbers

Science at Scale: Climate (August 2017)



NATURE | NEWS

How machine learning could help to improve climate forecasts

Mixing artificial intelligence with climate science helps researchers to identify previously unknown atmospheric processes and rank climate models.

Nicola Jones

23 August 2017

Greg Kendall-Ball

Many of the latest climate models seek to increase the detail in simulations of cloud structure.

As Earth-observing satellites become more plentiful and climate models more powerful, researchers who study global warming are facing a deluge of data. Some are now turning to the latest trend in artificial intelligence (AI) to help trawl through all the information, in the hope of discovering new climate patterns and improving forecasts.

nature briefing

What matters in science — and why — free in your inbox every weekday.

Listen

Nature Podcast

Our award-winning show features highlights from the week's edition of *Nature*, interviews with the people behind the science, and in-depth commentary and analysis from journalists around the world.

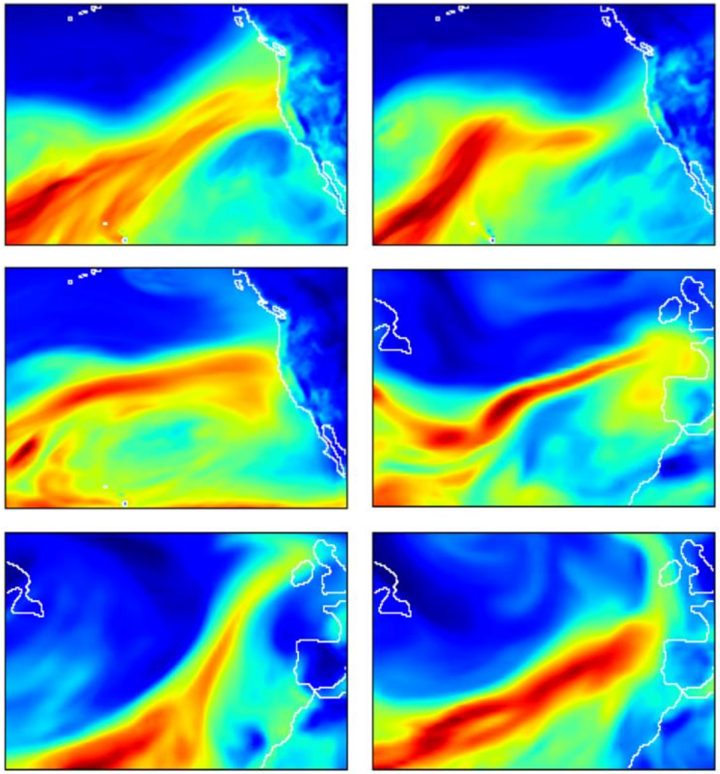


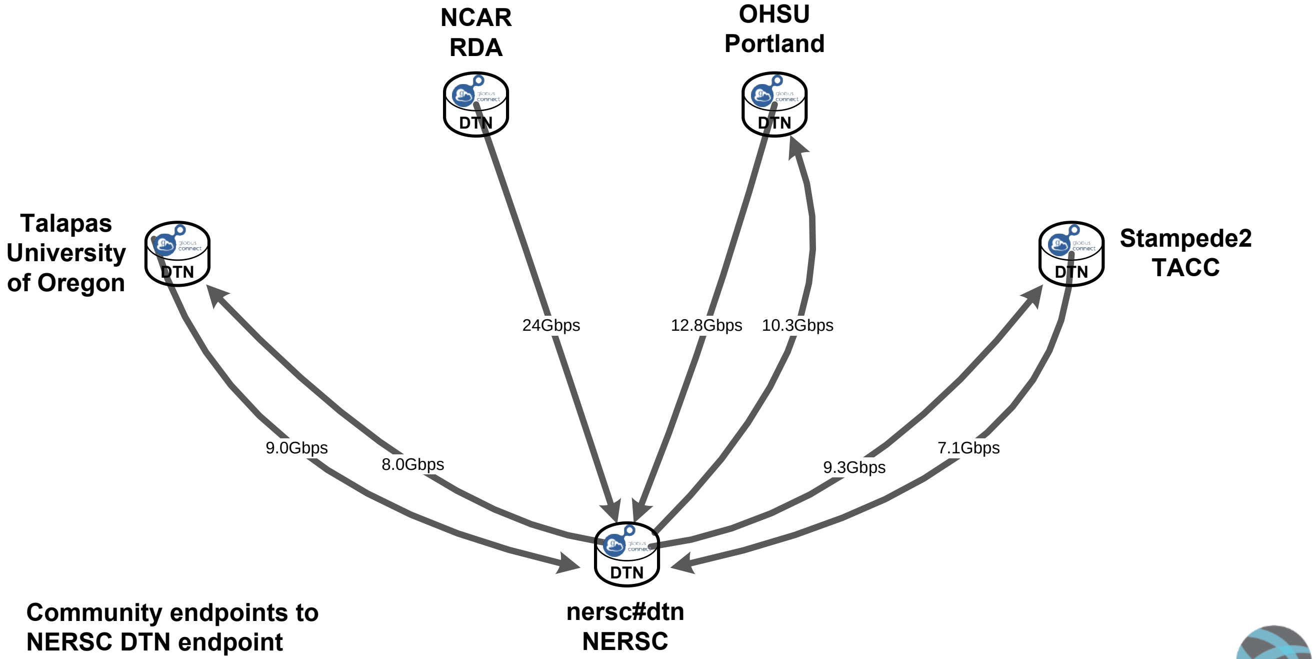
Figure 3: Sample images of atmospheric rivers correctly classified (true positive) by our deep CNN model. Figure shows total column water vapor (color map) and land sea boundary (solid line).

They Can Use All The Data

- Groups like these need large data sets
- Much of the data in many fields is behind legacy portals
  - Significant human effort to retrieve what scientists need
  - Legacy systems perform poorly, especially at scale
- Legacy data portals are a product of their time
  - Remember: these were designed to serve small data to small systems
  - We now live in the future from the perspective of those designs
  - Current systems far exceed the capabilities available 15 years ago
  - From the perspective of today’s systems, legacy portals are products of a bygone past
- It is now perfectly reasonable for a scientist to want all the data
  - Machine learning + HPC
  - But this only works if the scientists can get to the data at scale

Community DTNs

- Initial tests with some others in the community
- University, Data Portal, and Supercomputer Center endpoints
- We need to do a lot more of this



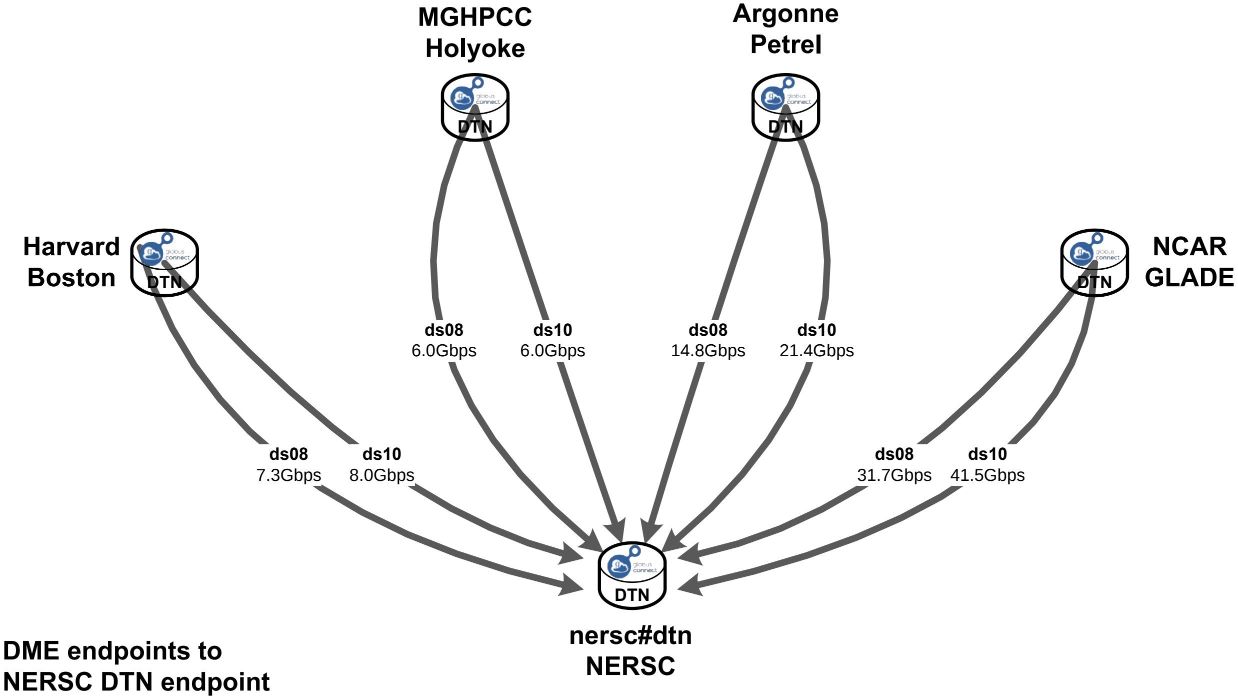
Community endpoints to
NERSC DTN endpoint

Data Mobility Exhibition

- Expansion of the Petascale DTN effort
- Goal is to dramatically increase the number of endpoints known to interoperate well and at high performance
- Ensure campuses can easily exchange data with national facilities
- This is the Data Mobility Exhibition
  - http://fasterdata.es.net/performance-testing/2019-2020-data-mobility-workshop-and-exhibition/2019-2020-data-mobility-exhibition/
  - Data sets and endpoints are described here:
    - https://www.globusworld.org/tour/data-mobility-exhibition
- Why the term "Exhibition?"
  - The point is to interoperate, not to compete
  - Everyone has different mission, constraints, funding, requirements
  - But everyone has users, and they need this stuff to work well

DME Endpoints – Initial Work

- Initial tests with Data Mobility Exhibition endpoints
- Let’s expand this – we can work together to achieve something amazing



MGHPCC
Holyoke

Argonne
Petrel

Harvard
Boston

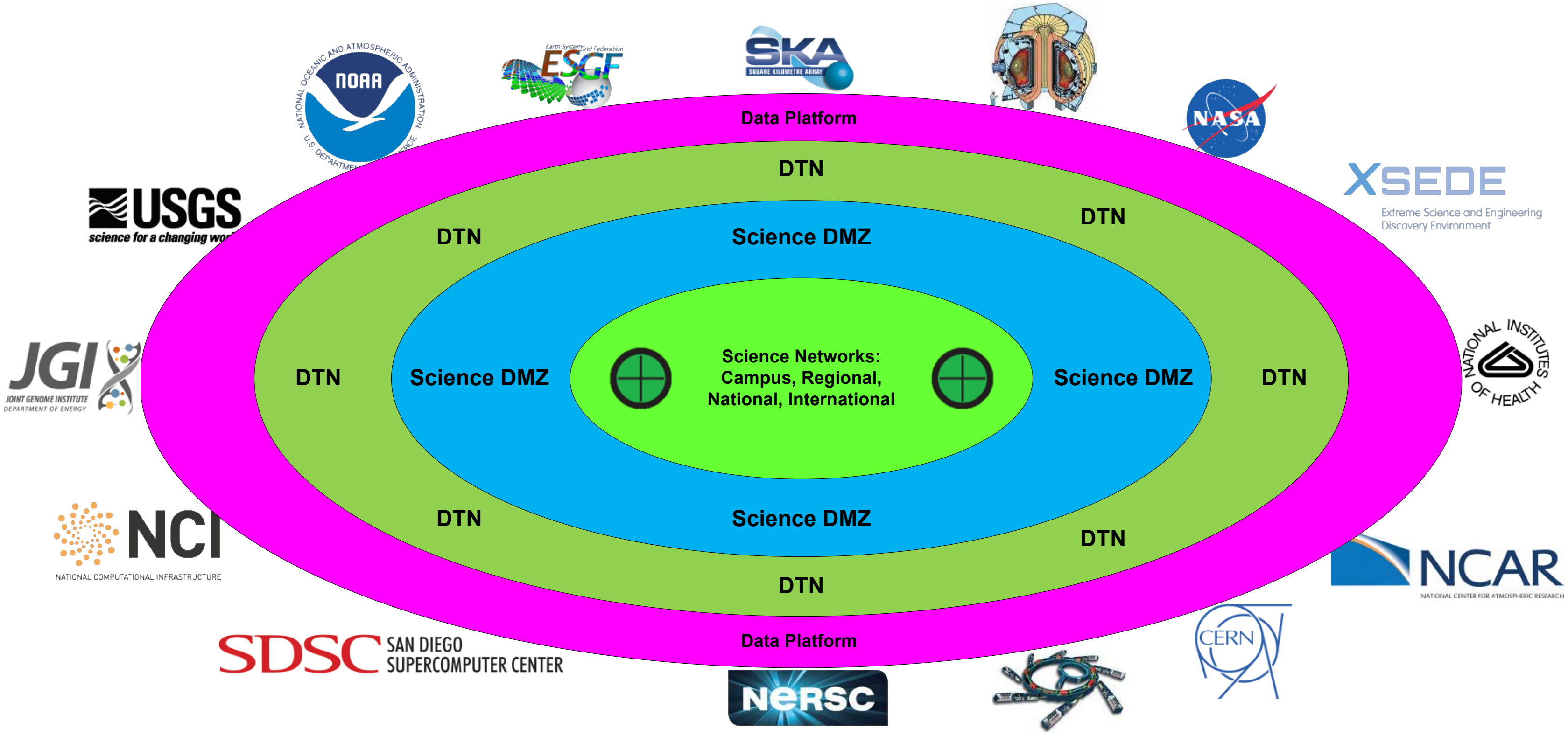
NCAR
GLADE

nersc#dtn
NERSC

**DME endpoints to
NERSC DTN endpoint**

**ds08 data set: 1TB (~30k files, 1MB to 10GB each)
ds10 data set: 1TB (100 files, 10GB each)**

Vision – Interoperable Computing And Data


Data Platform
DTN
DTN
DTN
Science DMZ
DTN
Science DMZ
Science Networks: Campus, Regional, National, International
Science DMZ
DTN
DTN
Science DMZ
DTN
DTN
Data Platform
NOAA
ESGF
SKA SQUARE KILOMETRE ARRAY
NASA
XSEDE Extreme Science and Engineering Discovery Environment
USGS science for a changing world
JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY
NATIONAL INSTITUTES OF HEALTH
NCI NATIONAL COMPUTATIONAL INFRASTRUCTURE
NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
SDSC SAN DIEGO SUPERCOMPUTER CENTER
NERSC
CERN

A Call To Action

- We are almost there!
- By ensuring high-performance interoperability, we prove that our systems are well-integrated and ready for use by the scientific community
- In aggregate, these systems will form a data infrastructure across US R&E
- Please participate in the data mobility exhibition
  - http://fasterdata.es.net/performance-testing/2019-2020-data-mobility-workshop-and-exhibition/2019-2020-data-mobility-exhibition/
- `We will come back in a year with results

ESnet
ENERGY SCIENCES NETWORK

Thanks!

Eli Dart dart@es.net

Energy Sciences Network (ESnet)

Lawrence Berkeley National Laboratory

engage@es.net

http://my.es.net/

http://www.es.net/

http://fasterdata.es.net/

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB