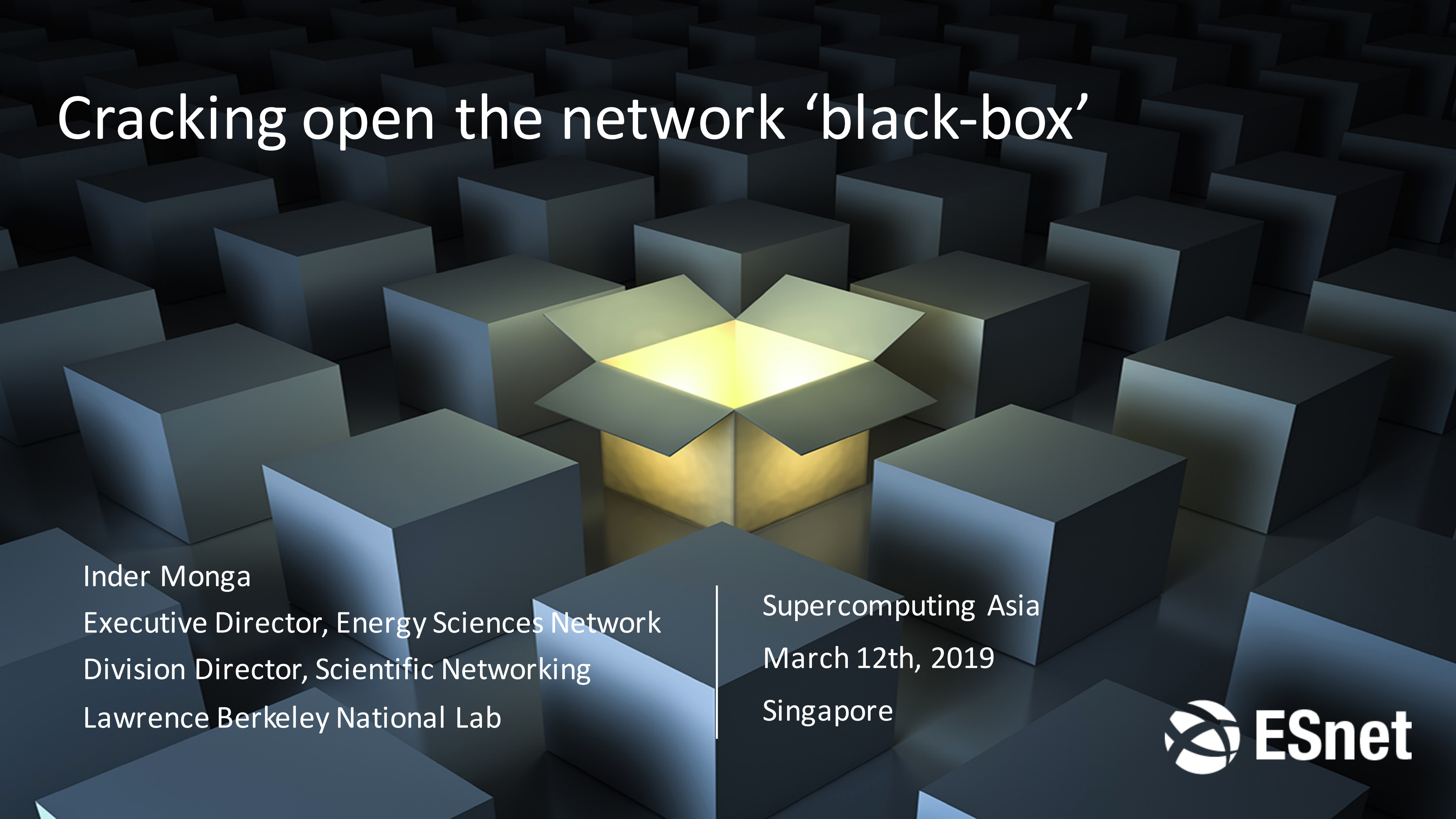
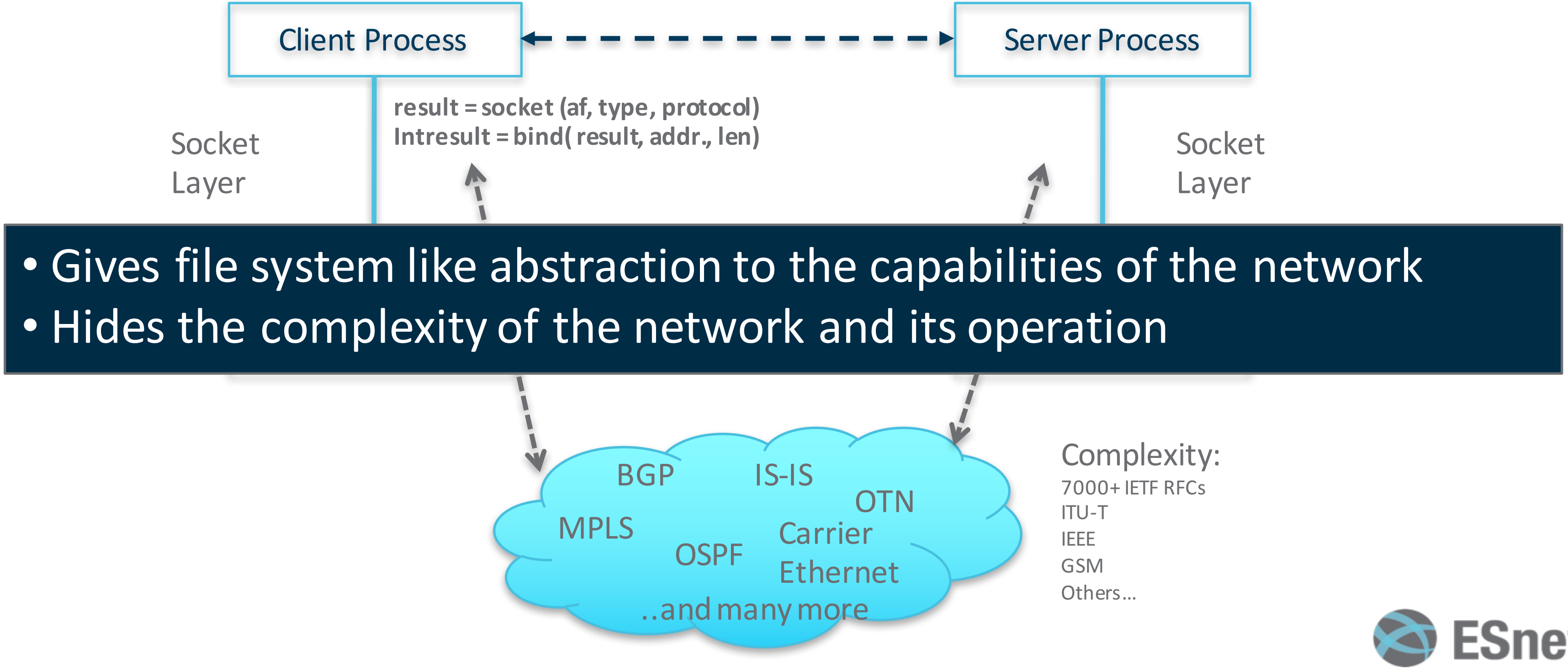


Cracking open the network ‘black-box’

Inder Monga
Executive Director, Energy Sciences Network
Division Director, Scientific Networking
Lawrence Berkeley National Lab

Supercomputing Asia
March 12th, 2019
Singapore

Socket Interface: Most successful data plane abstraction
Forces the network to be a black box to applications



- Gives file system like abstraction to the capabilities of the network
- Hides the complexity of the network and its operation

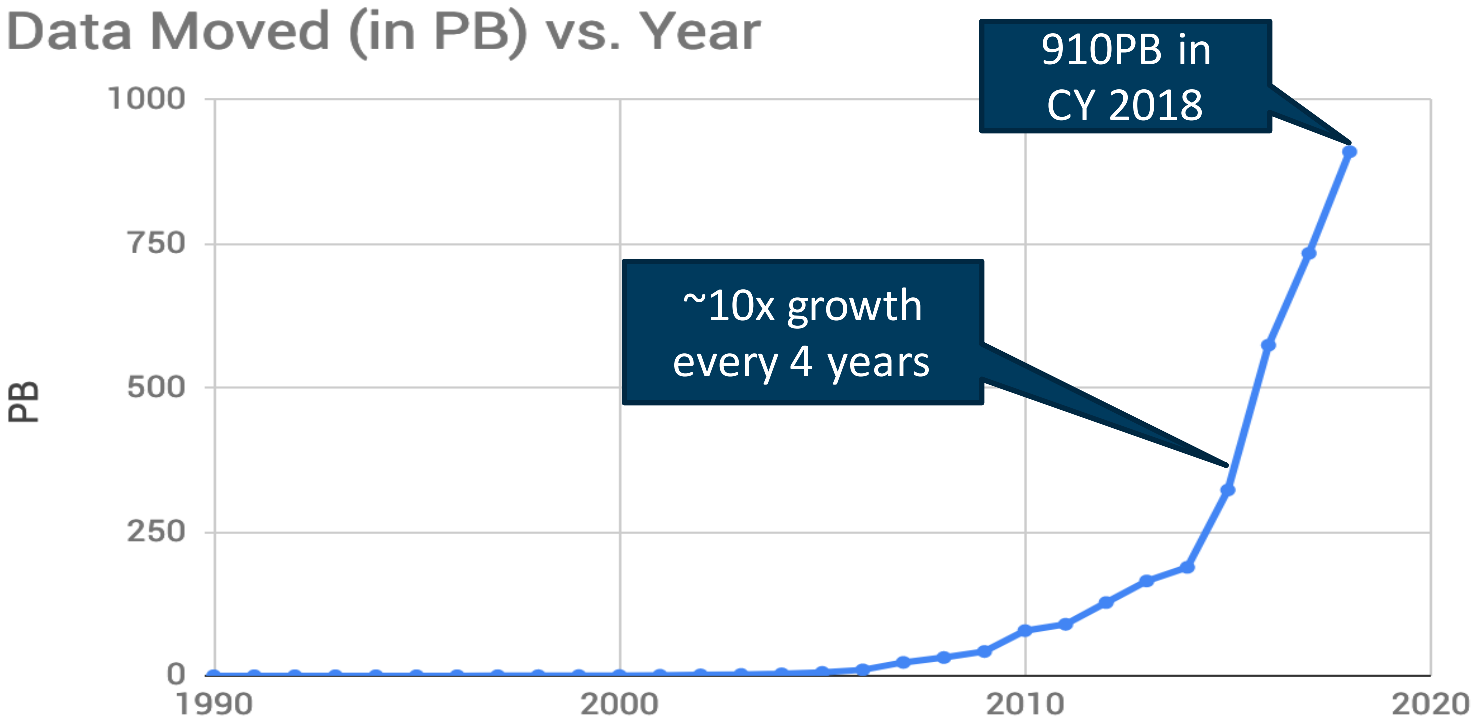
Complexity:
7000+ IETF RFCs
ITU-T
IEEE
GSM
Others...

High-performance science network user facility
Optimized for enabling big-data science

INL
Idaho National Laboratory
Pacific Northwest NATIONAL LABORATORY
AMES LABORATORY
BERKELEY LAB
Jefferson Lab
Thomas Jefferson National Accelerator Facility
OAK RIDGE National Laboratory
NREL
NATIONAL RENEWABLE ENERGY LABORATORY
Argonne NATIONAL LABORATORY
ESnet
SLAC NATIONAL ACCELERATOR LABORATORY
Fermilab
PPPL
PRINCETON PLASMA PHYSICS LABORATORY
Lawrence Livermore National Laboratory
BROOKHAVEN NATIONAL LABORATORY
Sandia National Laboratories
Los Alamos NATIONAL LABORATORY EST. 1943
NETL
NATIONAL ENERGY TECHNOLOGY LABORATORY
SRNL
SAVANNAH RIVER NATIONAL LABORATORY

Provides connectivity to all of the DOE labs, experiment sites, & user facilities (> 34417 users)

An ~exabyte scale network today



910PB in CY 2018

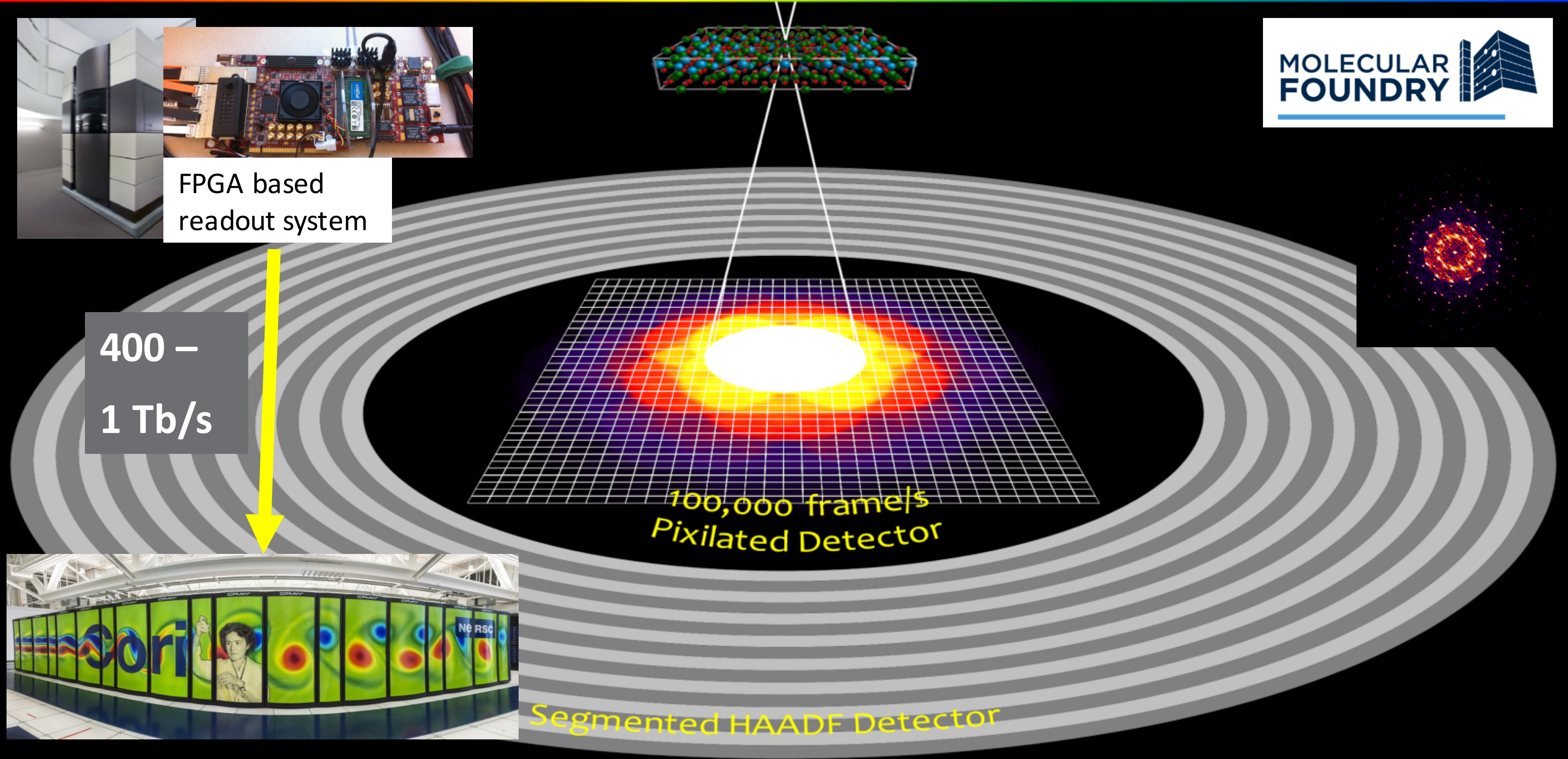
~10x growth every 4 years

exponential traffic growth over past 28 years

measures ingress or egress only, not traffic per link

ESnet

New instruments, more data: NCEM 4D-Stem

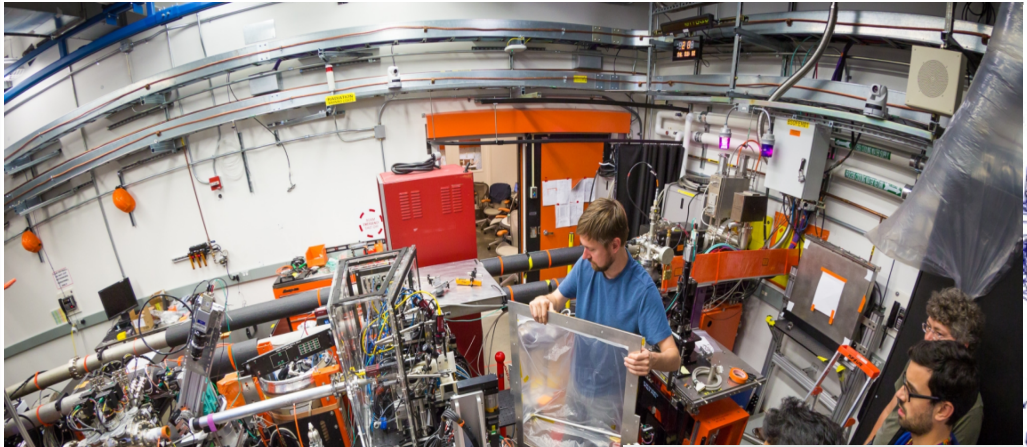


FPGA based readout system

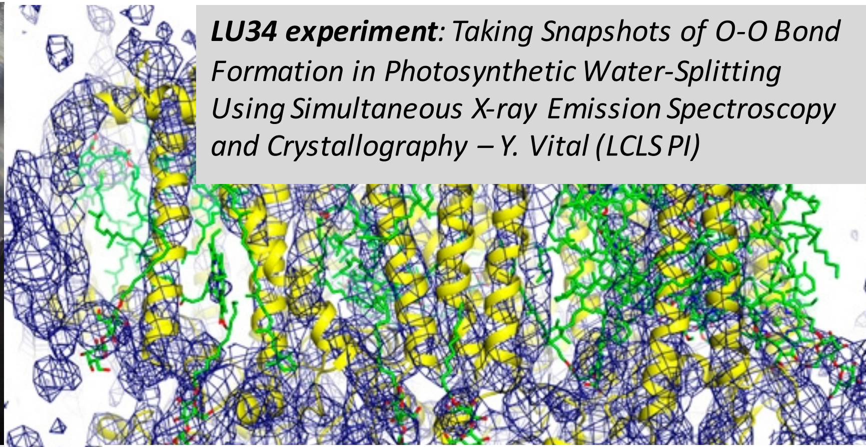
400 –

1 Tb/s

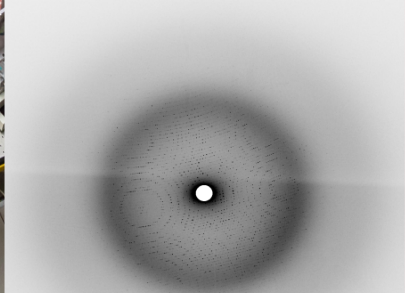
New instruments, more data : LCLS-II



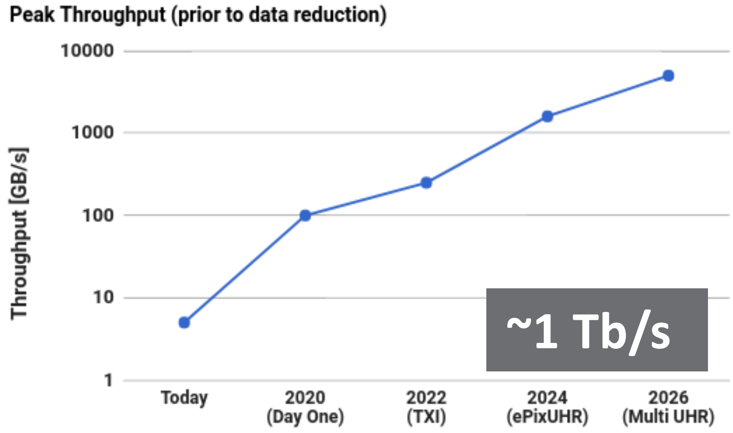
LCLS SLAC NATIONAL ACCELERATOR LABORATORY



***LU34 experiment**: Taking Snapshots of O-O Bond Formation in Photosynthetic Water-Splitting Using Simultaneous X-ray Emission Spectroscopy and Crystallography – Y. Vital (LCLS PI)*

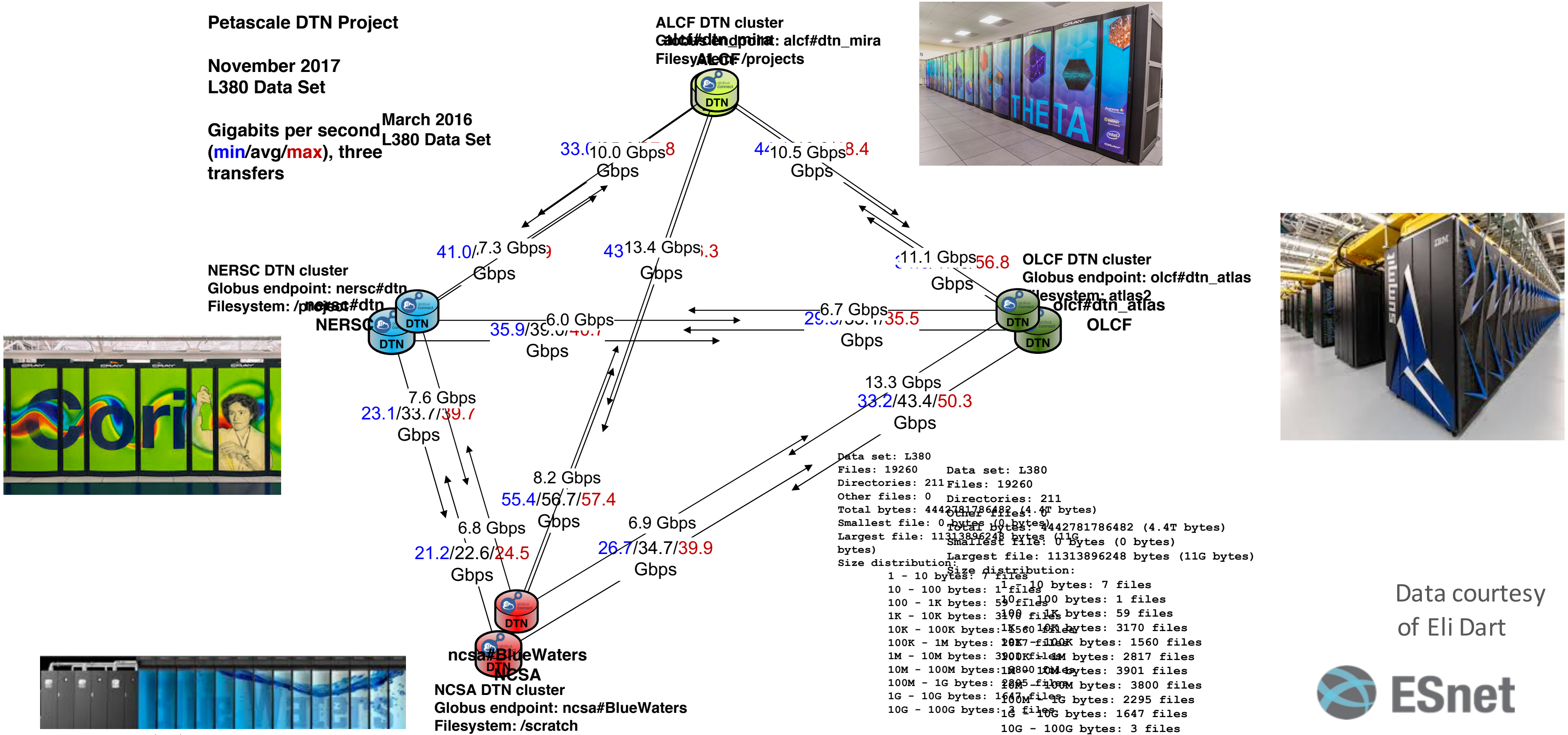


Diffraction pattern from LU34



~1 Tb/s

Science DMZ architecture [ESnet] has been impactful around the world and adopted by {Pacific, National, Asia} Research platforms



Petascale DTN Project

November 2017
L380 Data Set

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

March 2016
L380 Data Set

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

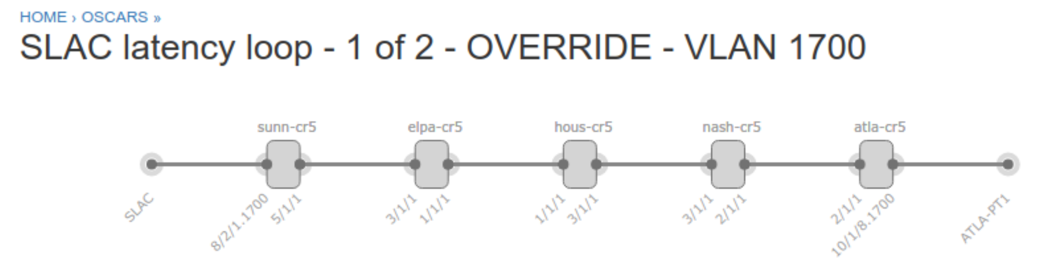
NCSA DTN cluster
Globus endpoint: ncsa#BlueWaters
Filesystem: /scratch

```
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
```

```
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
```

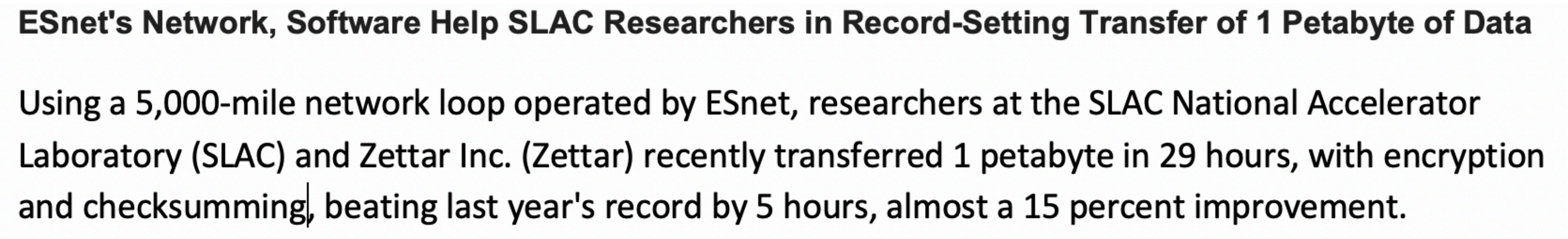
Data courtesy of Eli Dart

Data movement software keeps on improving: from 1 PB/week to 1 PB/day (approx.)



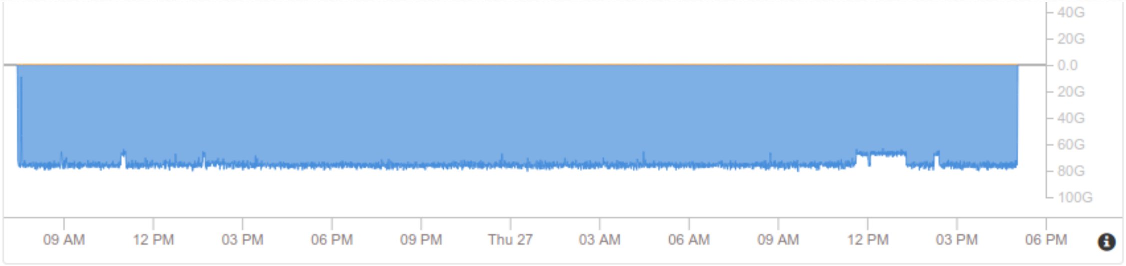
HOME › OSCARS »

SLAC latency loop - 1 of 2 - OVERRIDE - VLAN 1700



ESnet's Network, Software Help SLAC Researchers in Record-Setting Transfer of 1 Petabyte of Data

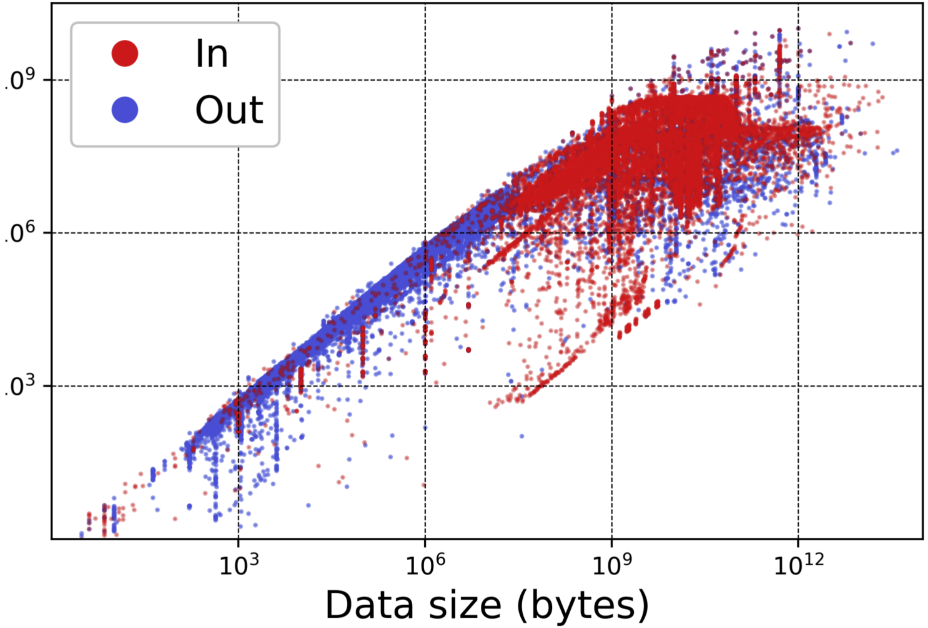
Using a 5,000-mile network loop operated by ESnet, researchers at the SLAC National Accelerator Laboratory (SLAC) and Zettar Inc. (Zettar) recently transferred 1 petabyte in 29 hours, with encryption and checksumming, beating last year's record by 5 hours, almost a 15 percent improvement.



But, from an application perspective, what happens when your network data transfer fails?

ESnet

Even well tuned infrastructure does not get consistent service



(d) Petrel

Nine orders of variability

Chard K, Dart E, Foster I, Shifflett D, Tuecke S, Williams J. (2018) The Modern Research Data Portal: a design pattern for networked, data-intensive science. *PeerJ Computer Science*4:e144 https://doi.org/10.7717/peerj-cs.144

How to ‘crack open the network black box’ without destroying the power of the abstraction?

1. High-precision telemetry
2. Scalable analytics infrastructure
3. Model-based approach to request network services
4. Network prediction using machine learning techniques

1. High-precision telemetry: deep insight into flows

Jupiter with the naked eye

Jupiter Close Up

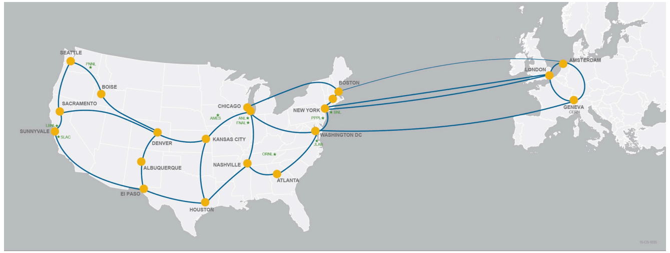
Per flow, high-precision telemetry

- Per packet-metadata tracking (e.g. timestamp, ingress location, etc)
- 10 ns precision in timing

Use high-fidelity data to get better insights and analytics:

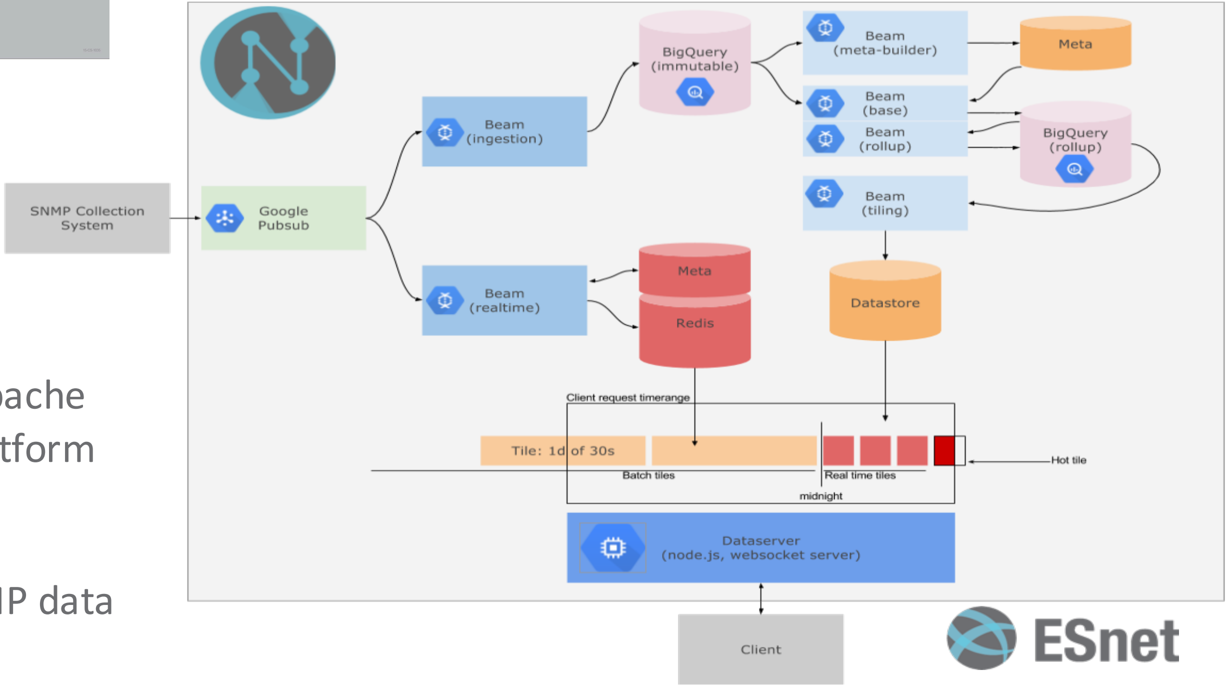
- Packet Microbursts
- Path deviations (RTT and Delay)
- Security / anomaly detection
- Head of Queue Blocking
- Many others…

2. Scalable analytics infrastructure



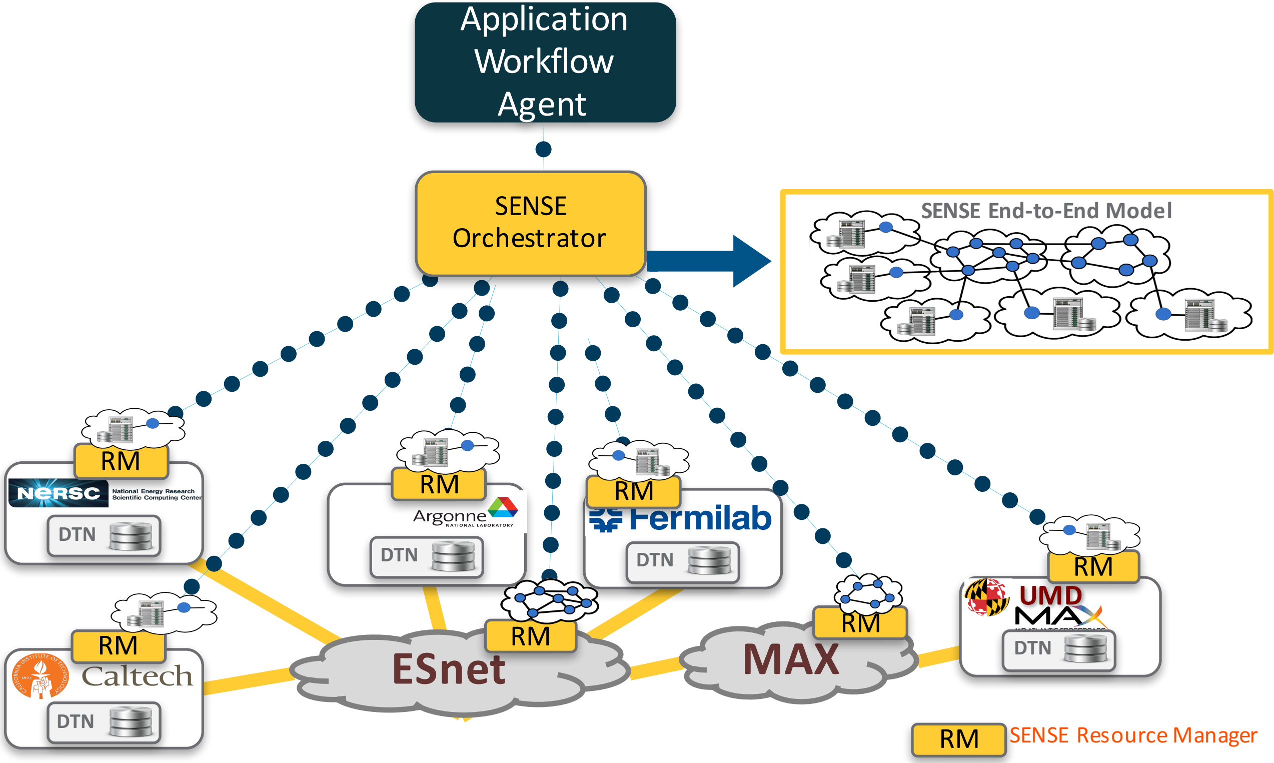
Streams telemetry

On-demand analytics infrastructure



- Real-time telemetry from the network
- **netbeam** platform: Using Apache BEAM and Google Cloud Platform
- Both Batch and Stream processing in parallel
- In production for ESnet SNMP data

3. Model based approach to request network services

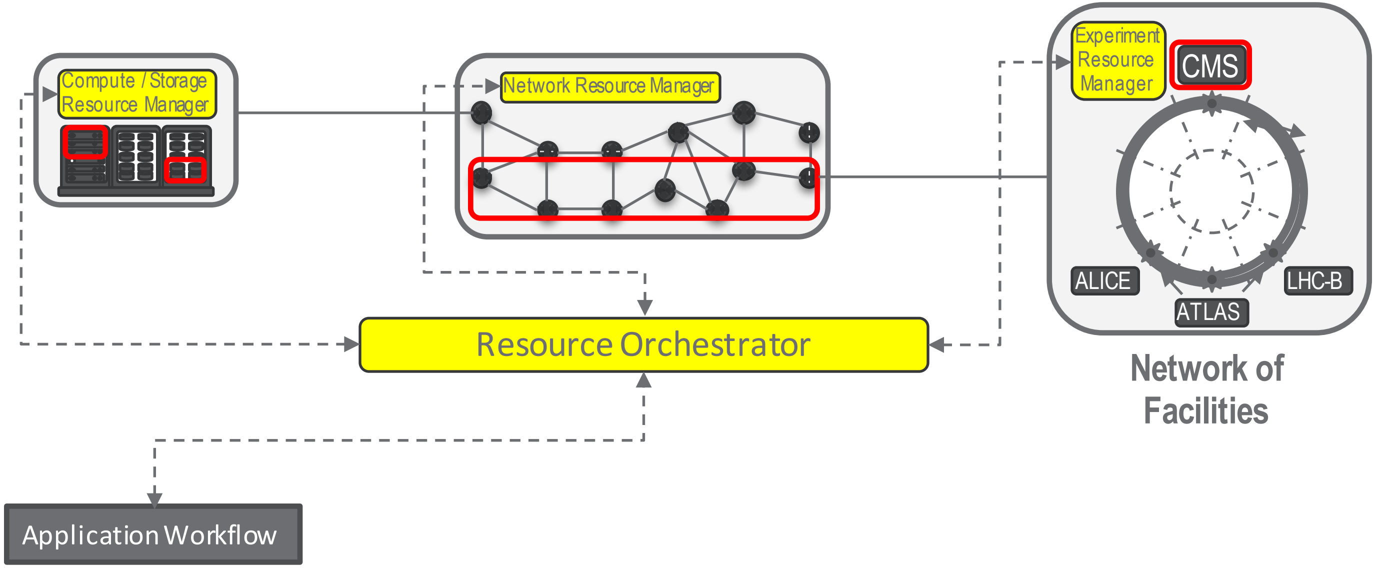

Application Workflow Agent
SENSE Orchestrator
SENSE End-to-End Model
RM
NERSC National Energy Research Scientific Computing Center
DTN
RM
Argonne National Laboratory
DTN
RM
Fermilab
DTN
RM
Caltech
DTN
RM
ESnet
RM
MAX
RM
UMD MAX Mid-Atlantic Crossroads
DTN
RM SENSE Resource Manager

Realtime system based on Resource Manager developed infrastructure and service models

Multi-resource models abstracts the network specifics away and allows for higher-level service request

SENSE project is a research project funded by ASCR, DOE, US

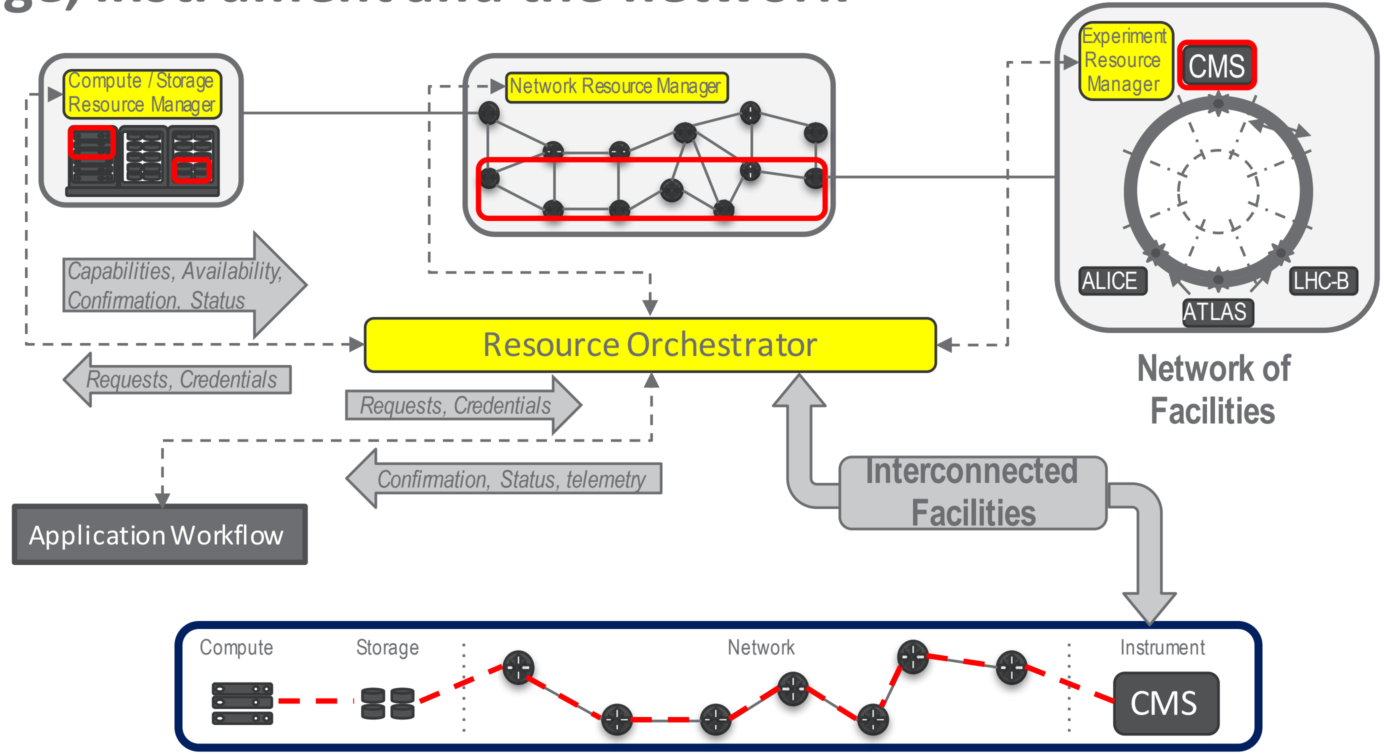
3. Enables an end-to-end service model that includes compute, storage, instrument and the network



Slide courtesy of Chin Guok

ESnet

3. Enables an end-to-end service model that includes compute, storage, instrument and the network



Slide courtesy of Chin Guok

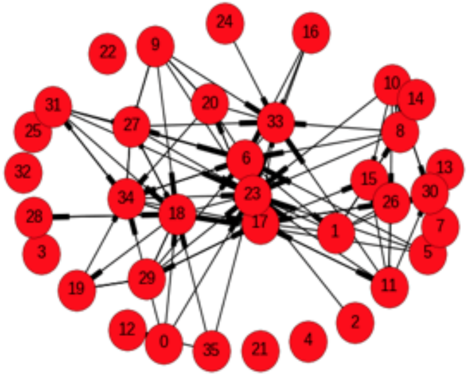
ESnet

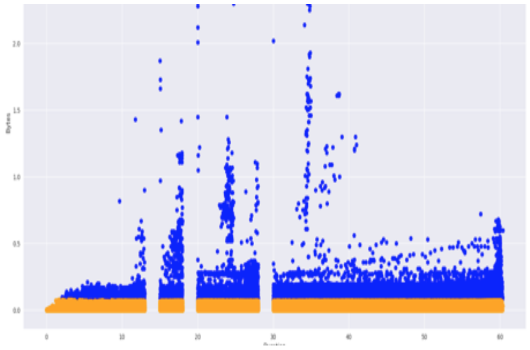
4. Network prediction using Machine Learning techniques

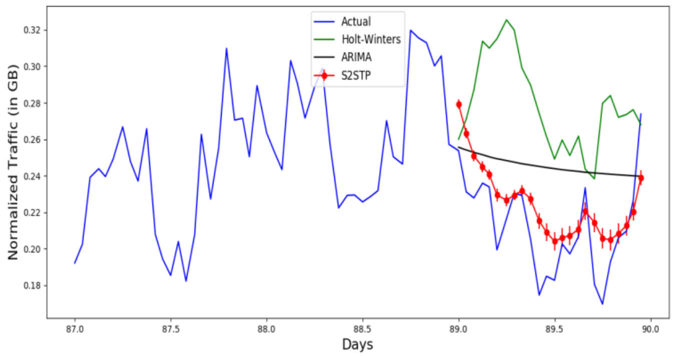
Understanding which sites are busiest at different times

High-Speed classifying of big and small flows to redirect packet routes

Prevent congestion and links failures by anticipating traffic 24 hours in advance





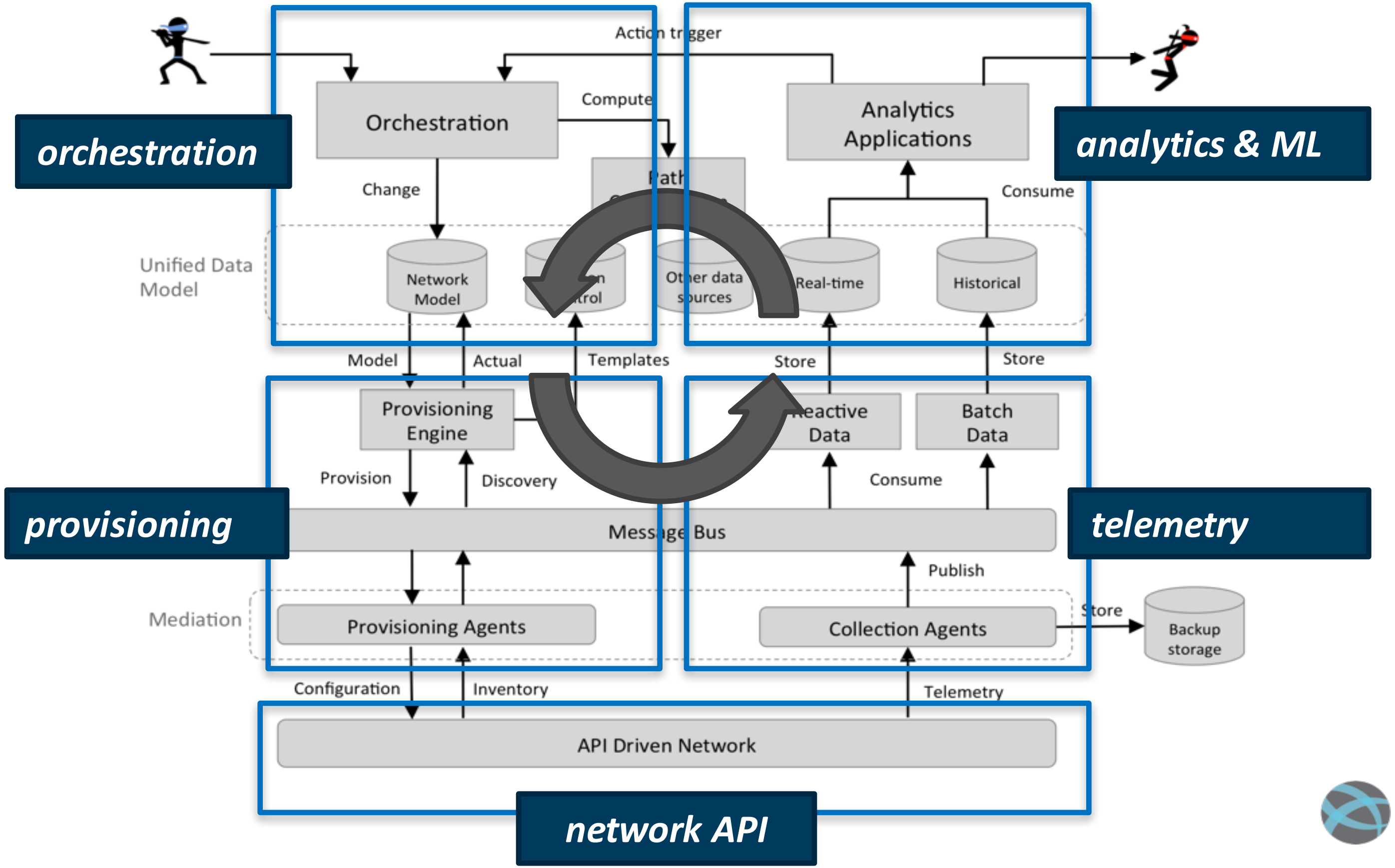


Network traffic prediction will help us bring route and engineer flows appropriately and on-the-fly

Slide courtesy of Mariam Kiran

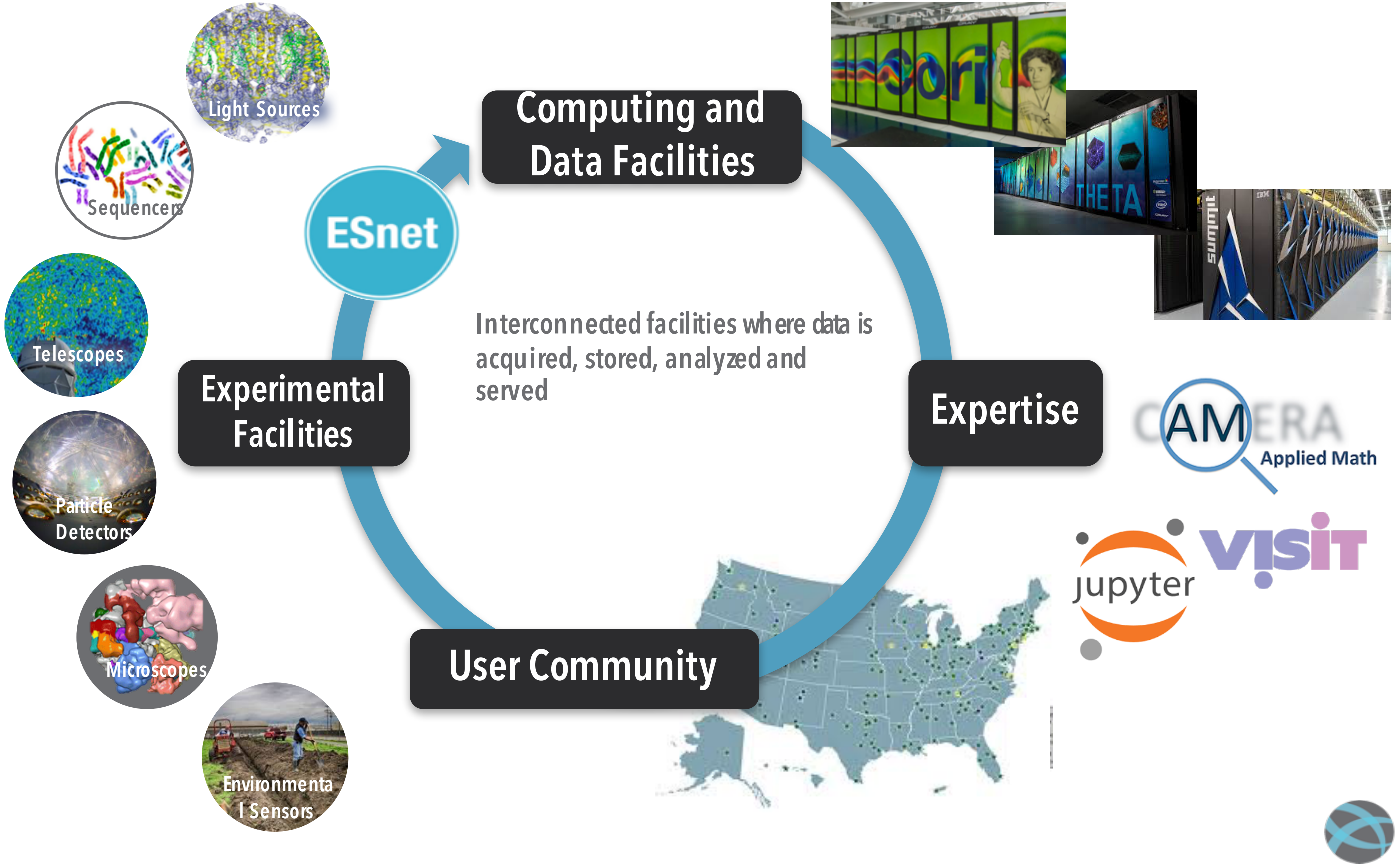
ESnet

ESnet's next-generation (ESnet6) software architecture



ESnet

Superfacility model for productive, reproducible science



Computing and Data Facilities

Expertise

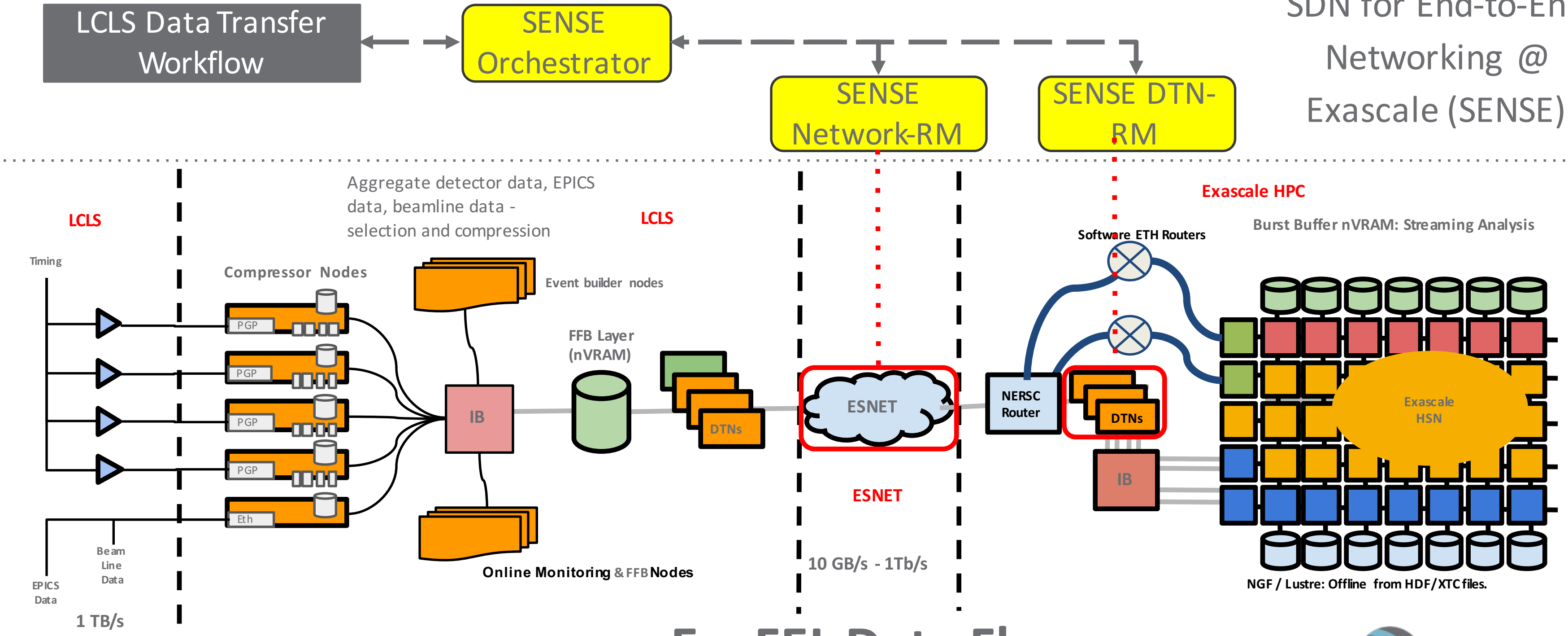
User Community

Experimental Facilities

Interconnected facilities where data is acquired, stored, analyzed and served

ExaFEL: A science example of the Superfacility model

SDN for End-to-End Networking @ Exascale (SENSE)



LCLS Data Transfer Workflow

SENSE Orchestrator

SENSE Network-RM

SENSE DTN-RM

LCLS

Aggregate detector data, EPICS data, beamline data - selection and compression

LCLS

Exascale HPC

Burst Buffer nVRAM: Streaming Analysis

1 TB/s

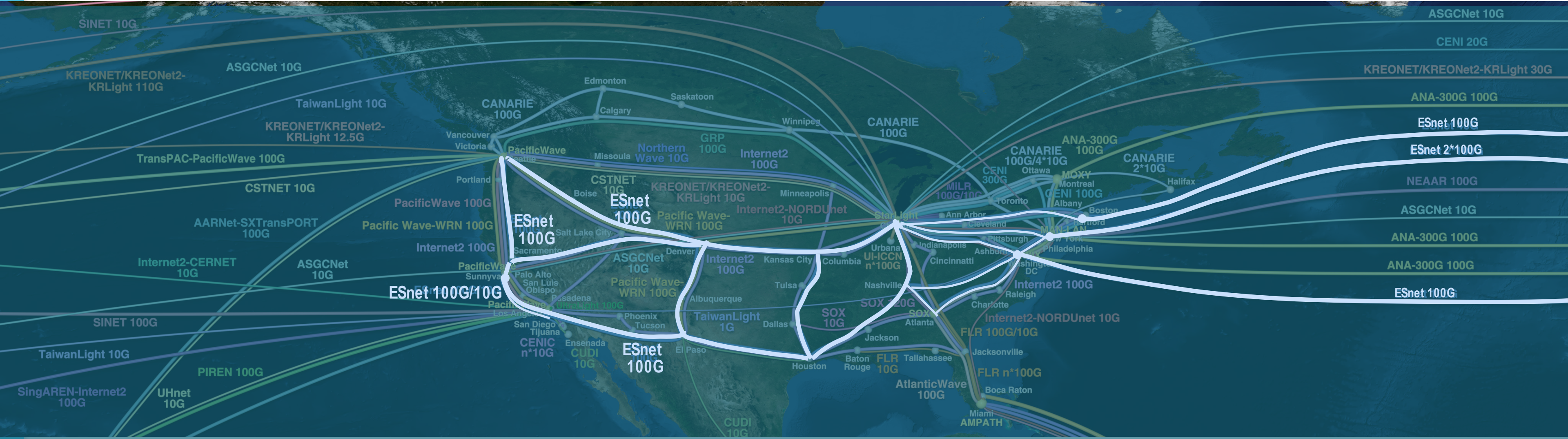
ESNET

10 GB/s - 1Tb/s

NGF / Lustre: Offline from HDF/XTC files.

ExaFEL Data Flow

The circulatory system for DOE collaborative science



- **Networking tailored to meeting science demands**
  - Bandwidth reservations, performance monitoring, Science “DMZ” model for the last mile,..
- An effective **network design and application interaction** is extremely important to meet the needs of next-generation of data science

ESnet

Thank you.

imonga@es.net

The slides represent work done by colleagues from ESnet and Berkeley Lab