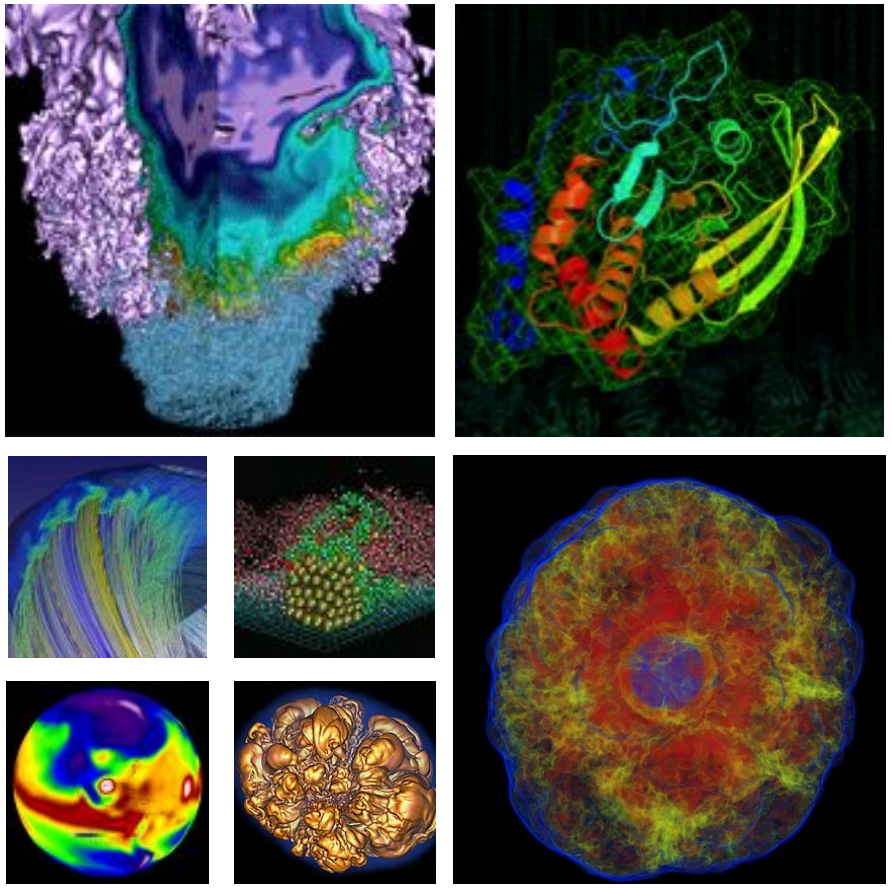
NUG Monthly Telecon

February edition



NERSC

Feb 11th 2016

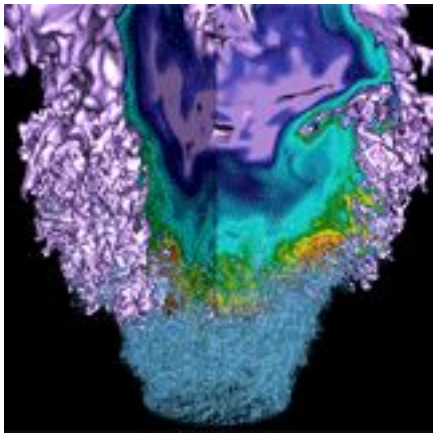
U.S. DEPARTMENT OF ENERGY | Office of Science

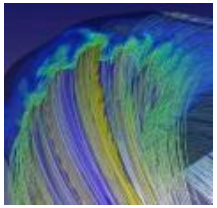
BERKELEY LAB
Lawrence Berkeley National Laboratory

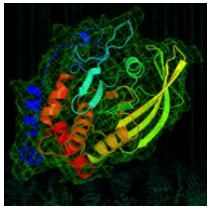
Agenda

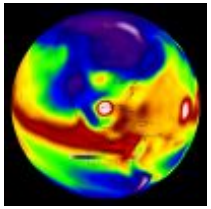
- Systems update
- Mendel Move update
- Cori update
- Edison update
- NUG annual meeting
- Queues and System Usage
- User Survey results
- Data movement and networks mini-tutorial
- Globus at NERSC

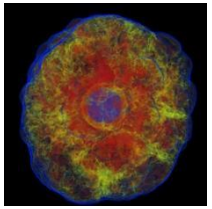
Systems Update

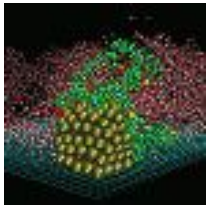












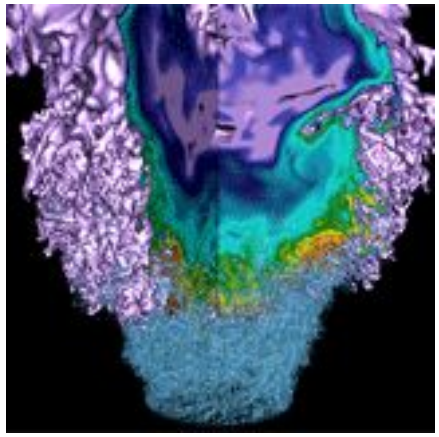
Tina Declerck

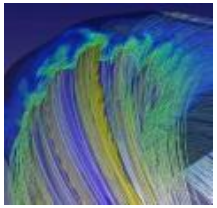
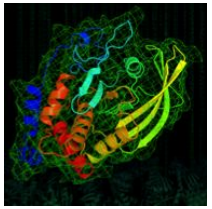
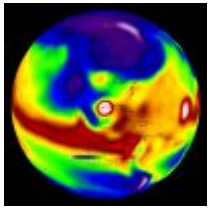
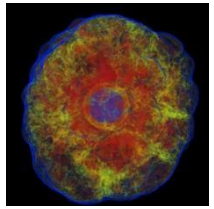
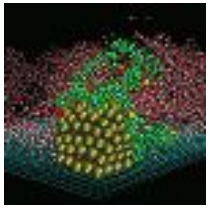
NUG Monthly, 2/11/2016

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Mendel move update



Dan Udwary

NUG Monthly, 2/11/2016

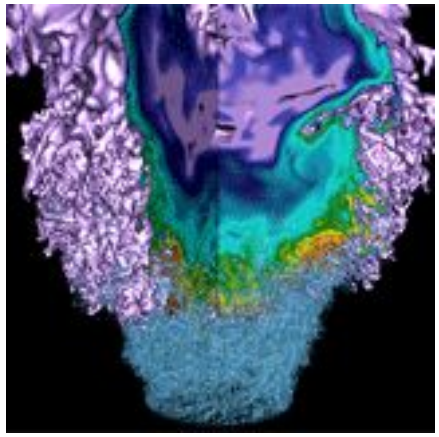
U.S. DEPARTMENT OF ENERGY | Office of Science

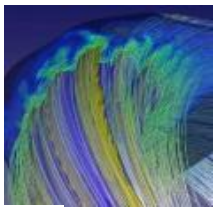
BERKELEY LAB

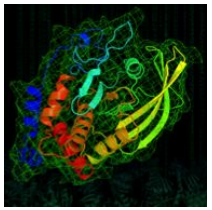
Mendel move update

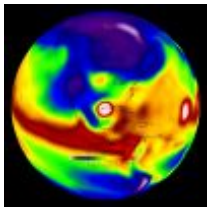
- Cori, Edison and all file systems are now in operation at Wang Hall
- HPSS remains in Oakland, connected to Wang Hall via a 400 GB/sec network link
- A cluster servicing JGI & PDSF is shut down and in process of move from OSF to Wang Hall
- No overall Genepool or PDSF downtime due to move

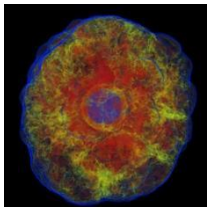
Cori User Environment and Running Jobs Update

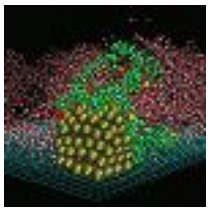












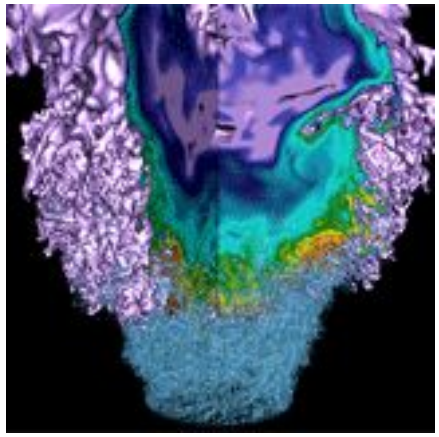
Helen He

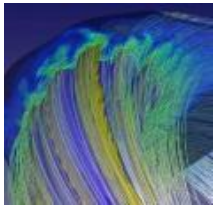
NUG Monthly, 2/11/2016

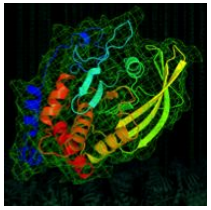
Cori Batch System

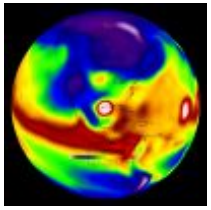
- **Major issue with SLURM batch system: batch jobs lost on Feb 9**
  - Jobs submitted between 11:50 am and 5:10 pm were lost.
  - All running jobs were terminated at 6:02 pm due to the control daemon crash during the recovery effort.
  - Batch system was back to normal as of 8:46 pm.
  - *We are sorry about the loss and inconvenience.*
- **Tips to run your jobs faster: https://www.nersc.gov/users/computational-systems/cori/running-jobs/queues-and-policies/**
- **Hyperthreading is enabled by default.**
  - SLURM sees 64 CPUs per node (each Cori node has 32 physical cores, total of 64 logical cores per node.)
- **Always use "#SBATCH -N" to request number of nodes.**
  - If asking nodes with "#SBATCH -n" (for num_MPI_tasks) only, you may get half the #nodes desired.

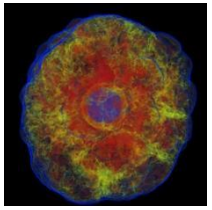
Edison Update

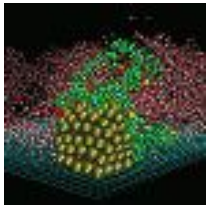










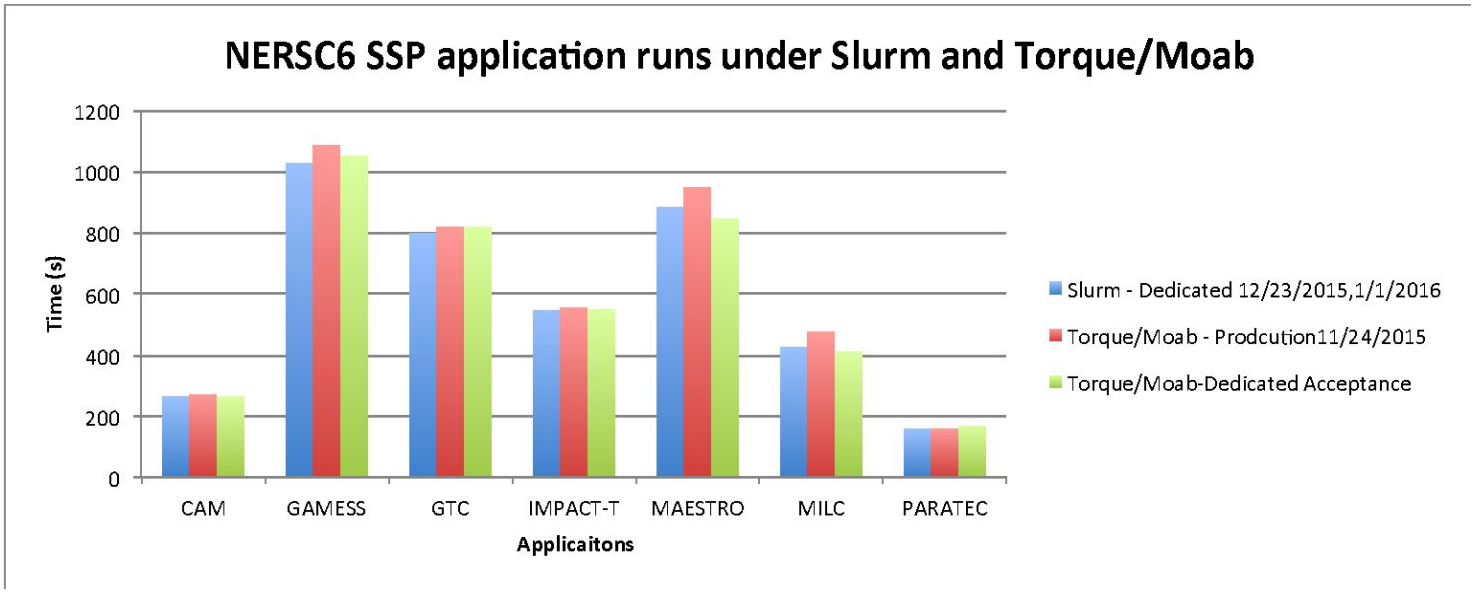


Zhengji Zhao
NERSC User Engagement Group

NUG Monthly Teleconference,
Berkeley CA, Feb 11 2016

Edison post-move news

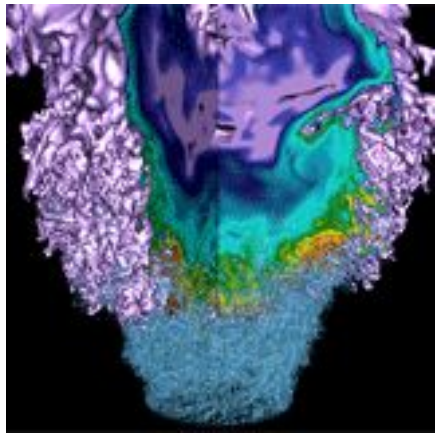
- **Host IP change**
- **NEW SSH authentication mechanism (1/12/2016)**
- **Power outages**

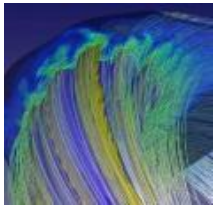


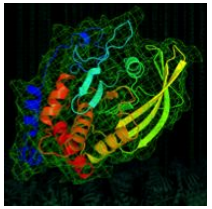
- The SSP benchmark we use to monitor the Edison performance get similar performance before and after the move.
- Please let us know if you observe any performance regression after the Edison move.

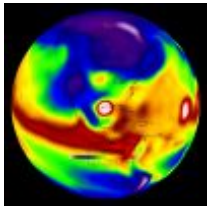
http://www.nersc.gov/users/computational-systems/edison/updates-and-status/timeline-and-updates/

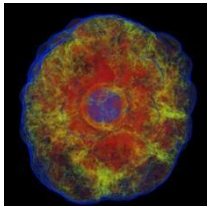
Queues and System Usage

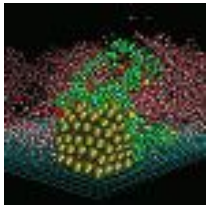








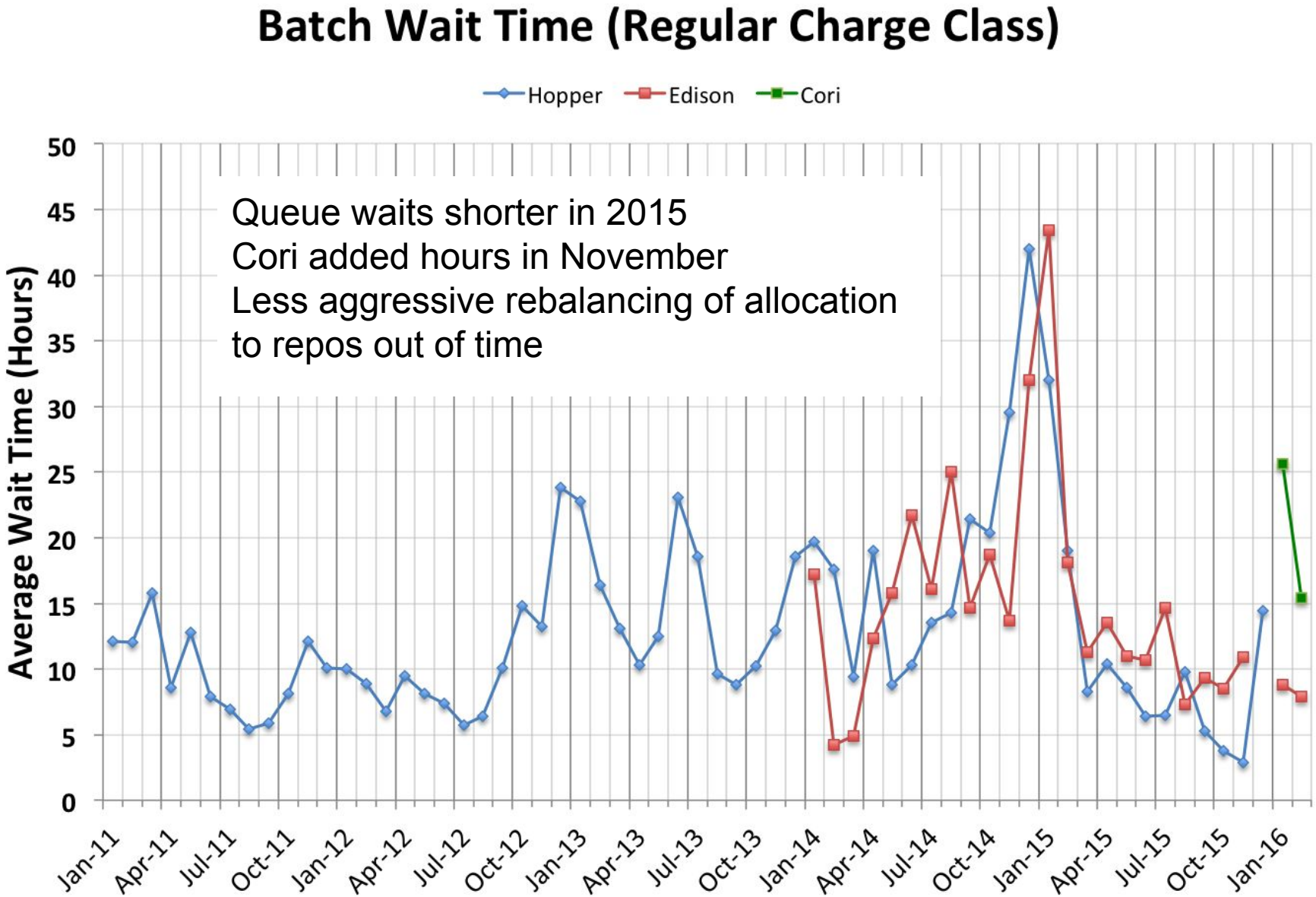




Cori/Edison Queues and Usage

- **Goals**
  - Edison: large compute intensive jobs
  - Cori: general purpose HPC jobs, data-optimized features
- **Simplified queue structure for 2016**
- **Repos out of time get jobs moved to scavenger queue automatically**
  - Minimal transfers to repos that exhaust their allocation
- **NERSC recently implemented SLURM batch scheduler on Cori and Edison**
  - still working out edge cases
- **There have been user questions about job mix and throughput**

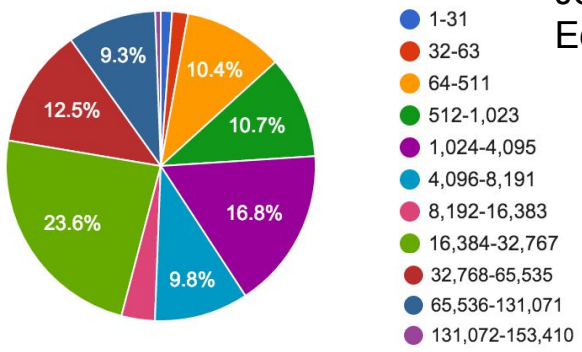
Wait Times


Batch Wait Time (Regular Charge Class)
Hopper
Edison
Cori
Queue waits shorter in 2015
Cori added hours in November
Less aggressive rebalancing of allocation
to repos out of time
Average Wait Time (Hours)
0
5
10
15
20
25
30
35
40
45
50
Jan-11
Apr-11
Jul-11
Oct-11
Jan-12
Apr-12
Jul-12
Oct-12
Jan-13
Apr-13
Jul-13
Oct-13
Jan-14
Apr-14
Jul-14
Oct-14
Jan-15
Apr-15
Jul-15
Oct-15
Jan-16

Job Distribution AY2016

Queue	Edison Hours	Edison N Jobs	Cori Hours	Cori N Jobs
Debug	3.3 %	33.0 %	5.2 %	19.0 %
Premium	1.1 %	1.8 %	7.4 %	1.7 %
Regular	**82.9 %**	**60.1 %**	**83.0 %**	26.7 %
Low	10.8 %	4.7 %	1.7 %	3.5 %
Scavenger*	1.2 %	0.4 %	1.1 %	0.3 %
Shared Node			1.2 %	**46.2 %**

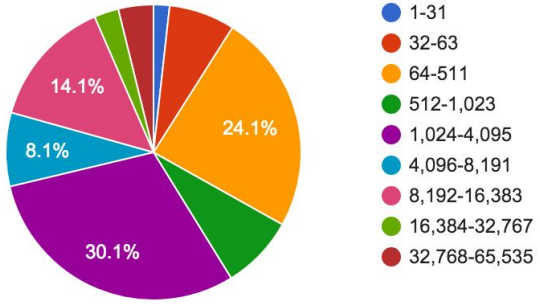
Raw Machine Hours by Cores Used (Percent)



Job Size Distribution

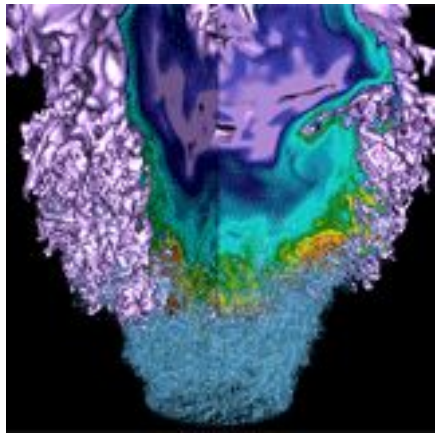
Edison Cori

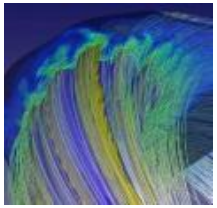
Raw Machine Hours by Cores Used (Percent)

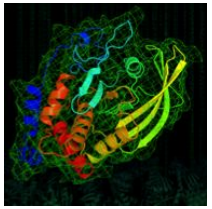


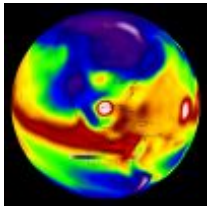
*9 out of 699 repos are out of time

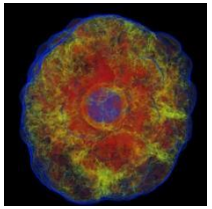
Users Group and Achievement Awards

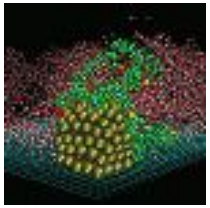












Users Group Meeting 2016

NUG 2016

March 21-24, 2015
Wang Hall/CRT Building
Berkeley Lab

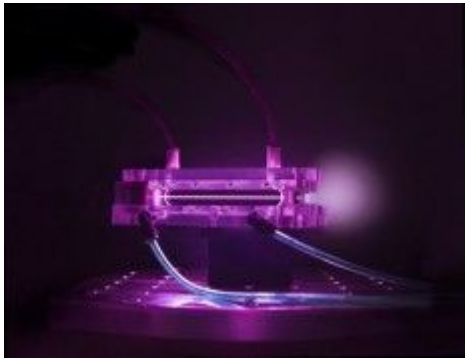
Registration is Open!

http://www.nersc.gov/users/NUG/annual-meetings/nug-2016/

Monday, March 21: New User Training/Data & Vis training; Hack-a-thon
Tuesday, March 22: Science & Technology
Wednesday, March 23: Application Readiness and NESAP
Thursday, March 24: Business Meeting (NERSC Updates and Discussions)

NERSC Science Achievement Awards: Call for Nominations Coming

2015 Recipients



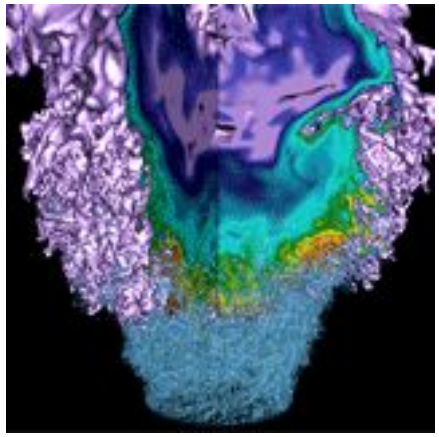
High Impact Scientific Achievement
Berkeley Lab Particle Accelerator Sets World Record, LBNL BELLA team (Carlo Benedetti)

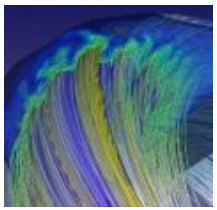
Early Career - Ken Chen, UC Santa Cruz
Current: EACOA Fellow at the National Observatory of Japan

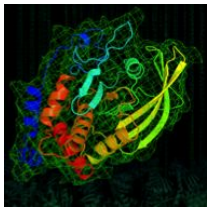
Innovative Use of HPC
SPOT Suite Transforms Beamline Science (Craig Tull)

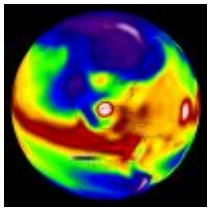
Early Career - Taylor Barnes, Caltech
Current: Rear Admiral Grace Murray Hopper Postdoctoral Fellow at NERSC

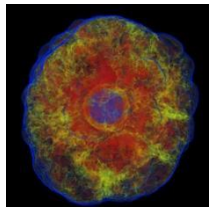
User Survey Results and User Ticket Statistics

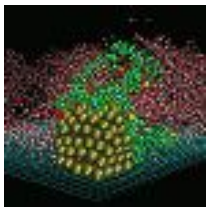




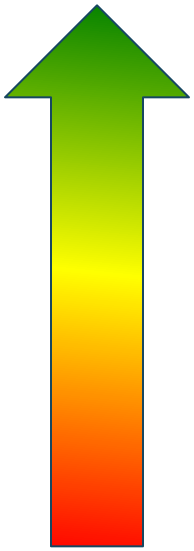








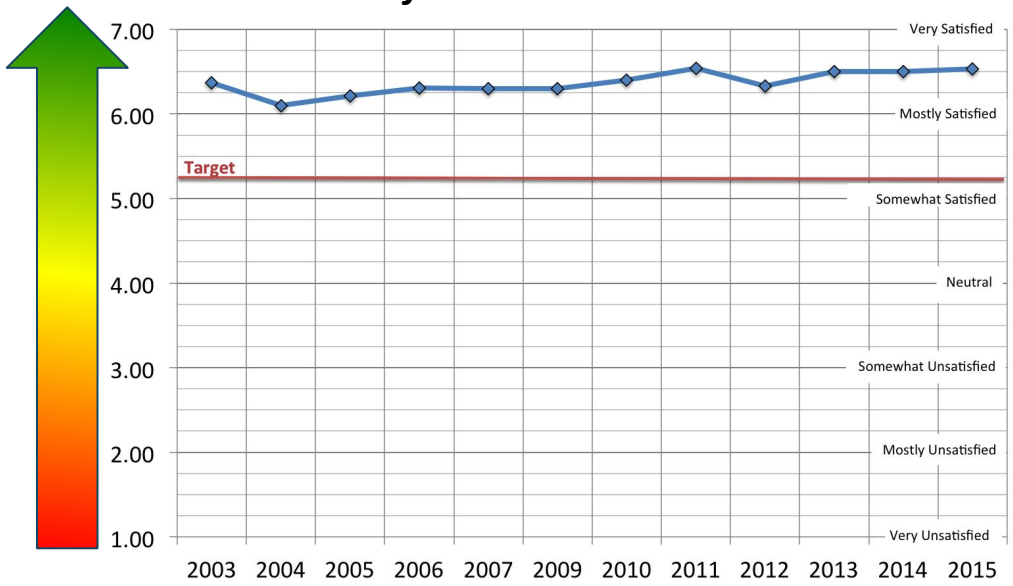
Customer Satisfaction



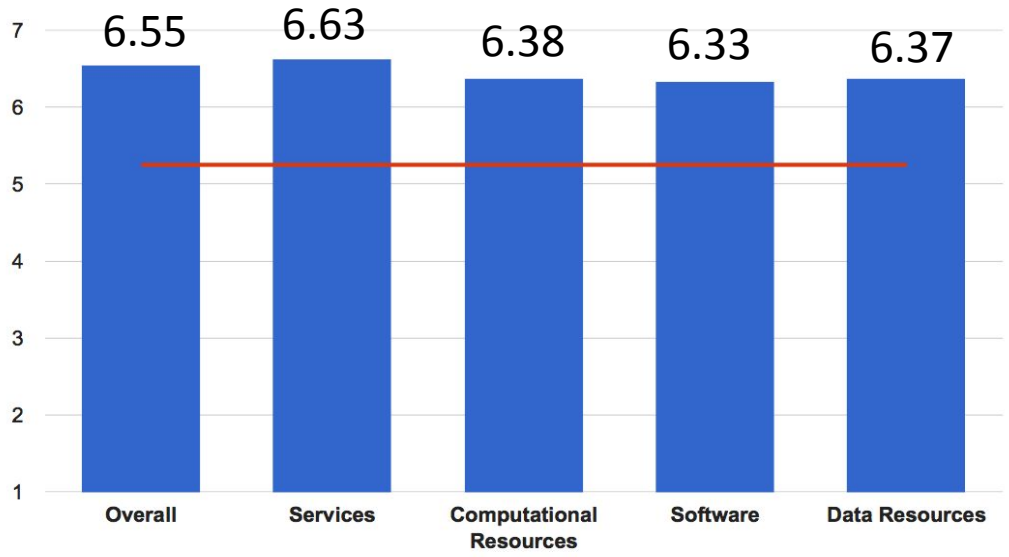
- **Annual NERSC User Survey**
- **Users score NERSC using a 1-7 scale**
  - 7 Very satisfied
  - 6 Mostly satisfied
  - 5 Somewhat satisfied
  - 4 Neutral
  - 3 Somewhat dissatisfied
  - 2 Mostly dissatisfied
  - 1 Very dissatisfied
- **Minimum satisfactory score 5.25**
- **606 survey respondents**
  - Similar response rate as in previous years (~10%)
  - Represents 63% of raw hours used on NERSC systems

User satisfaction across all areas was high

Overall Satisfaction with NERSC 2002-2015



User Satisfaction by Major Category



U.S. DEPARTMENT OF ENERGY | Office of Science

Greatest Areas of Satisfaction and Areas of Improvements

Survey Item	2015 Score
HPSS: Uptime (Availability)	6.74
SERVICES: Account support and passwords	6.72
PROJECT: Reliability	6.71
OVERALL: NERSC Security	6.70
HPSS: Reliability (data integrity)	6.69
PROJECT: Uptime	6.66
CONSULT: Consulting Overall	6.64
WEB: System Status Info	6.64
OVERALL: NERSC Services	6.63
CONSULT: On-line help desk	6.62

Survey Area	Score	Improvement over 2014
NERSC SW: Applications software	6.43	**+0.11**
HOPPER: Overall	6.38	**+0.12**
HOPPER: Batch queue structure	6.17	**+0.20**
DATA: Long-term data retention	6.15	**+0.18**
HOPPER: Batch wait time	5.92	**+0.75**
DATA: Scratch purge policy	5.86	**+0.30**
EDISON: Batch wait time	5.12	**+0.24**

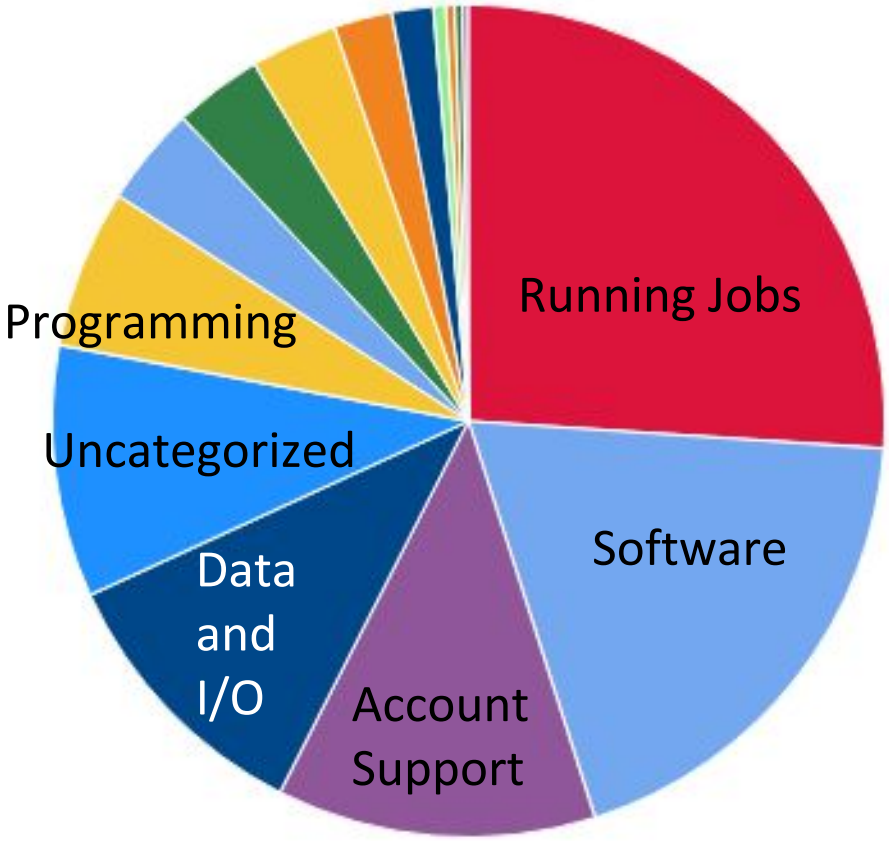
Scores with decreased satisfaction

Survey Area	Score	Decrease from 2014
WEB: NIM web accounting interface	6.40	-0.13
PROJECT: File and Directory Operations	6.32	-0.18
EDISON: Uptime (Availability)	6.27	-0.13
SERVICES: Allocations process	6.23	-0.20
TRAINING: Video Tutorials	5.95	-0.28

Tickets by Category

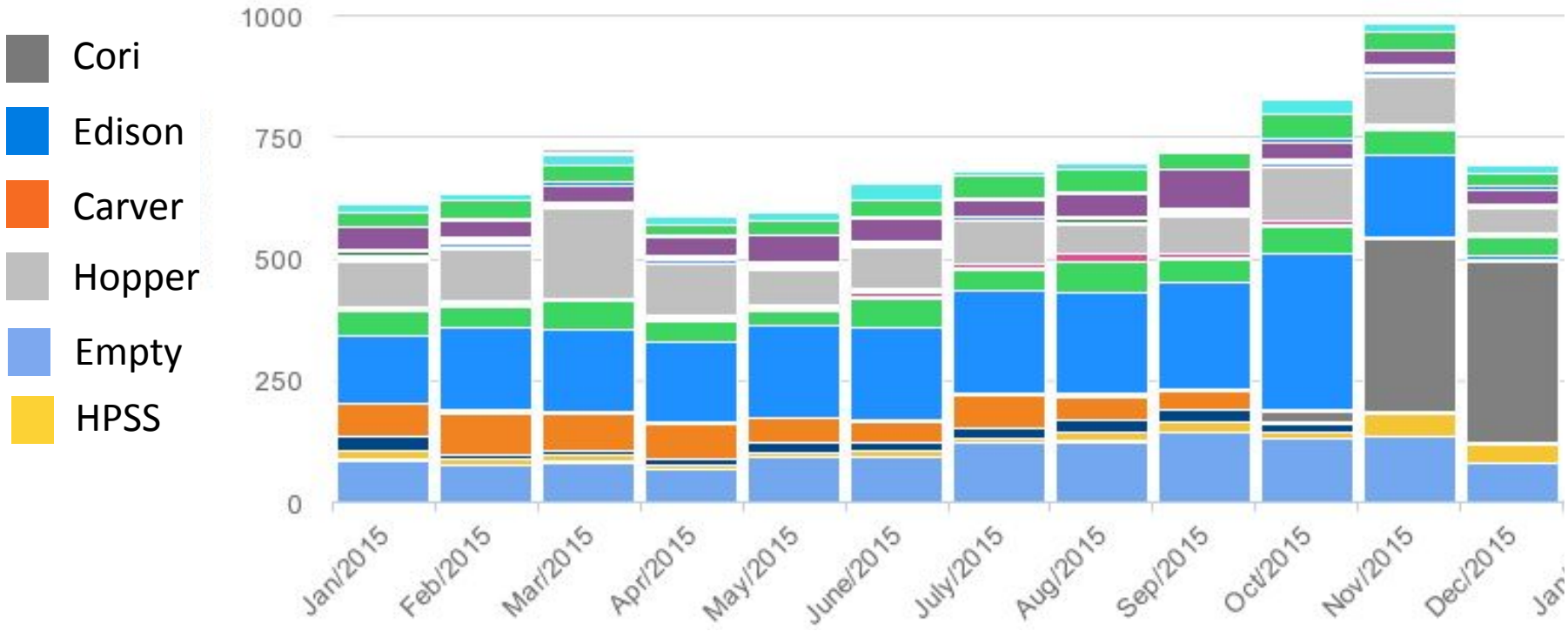
- **8,422 tickets were submitted by NERSC program users in 2015**
- **Compared to 2014 consultants are fielding more questions about performance, profiling, and data and I/O**

2015 Tickets by Category



Tickets by Month and System

Ticket Count by Month and by System



ESnet
ENERGY SCIENCES NETWORK

Moving Data Over Networks

Network-Based Data Transfer at NERSC

Eli Dart, Network Engineer
ESnet Science Engagement
Lawrence Berkeley National Laboratory

NERSC User Group Monthly Telecon
Berkeley, CA
February 11, 2016

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB

Motivation

- Networks are an essential part of data-intensive science
  - Connect data sources to data analysis
  - Connect collaborators to each other
  - Enable machine-consumable interfaces to data and analysis resources (e.g. portals), automation, scale
- Performance is critical
  - Exponential data growth
  - Constant human factors
  - Data movement and data analysis must keep up
- Effective use of wide area (long-haul) networks by scientists has historically been difficult
- Some of this is for your system administrator
  - Point your sysadmin to http://fasterdata.es.net/ for more info
  - Feel free to follow up with me later – engage@es.net

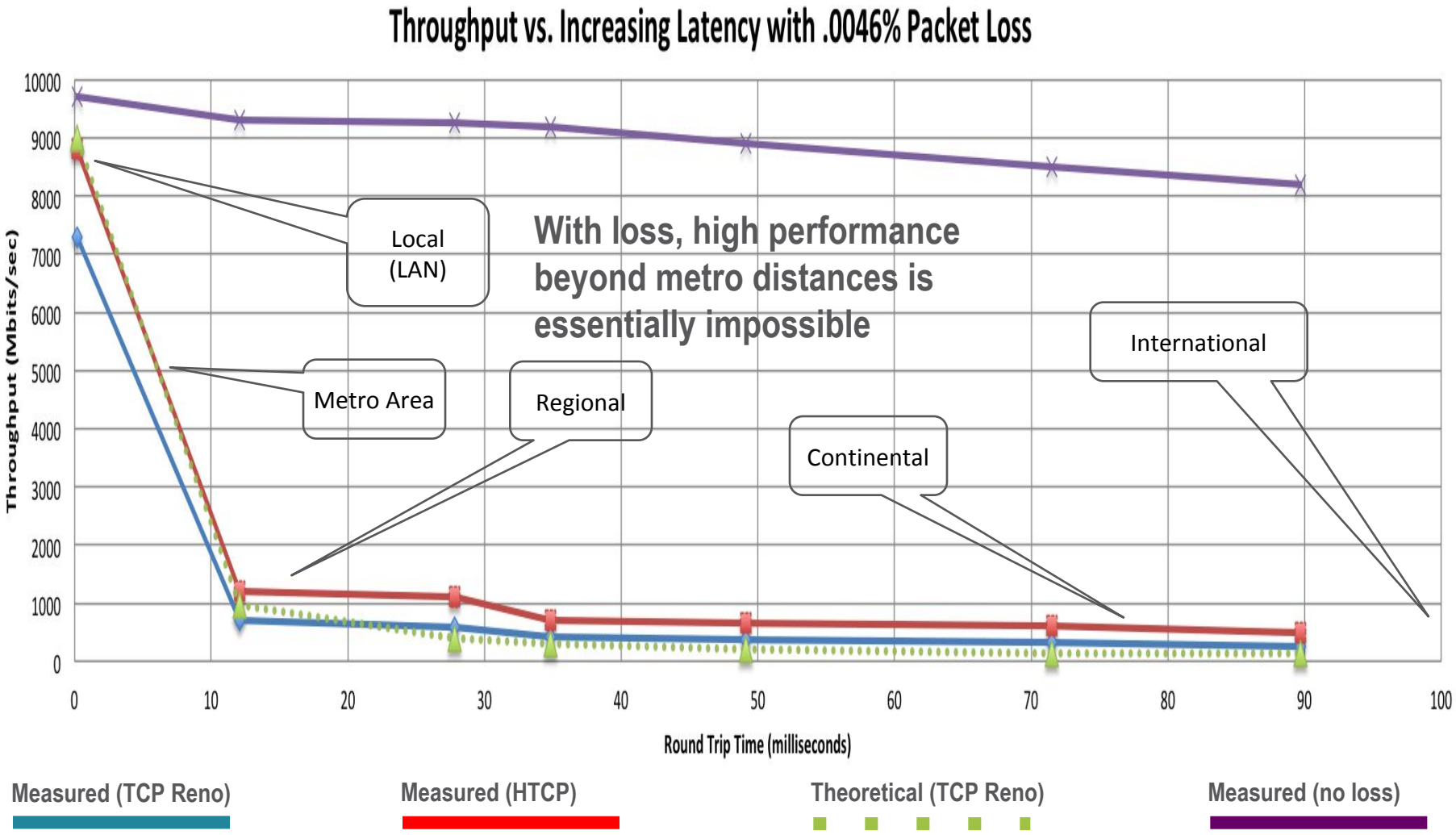
The Central Role of the Network

- The very structure of modern science assumes science networks exist: high performance, feature rich, global scope
- What is "The Network" anyway?
  - "The Network" is the set of devices and applications involved in the use of a remote resource
    - This is not about supercomputer interconnects
    - This is about data flow from experiment to analysis, between facilities, etc.
  - User interfaces for "The Network" – portal, data transfer tool, workflow engine
  - Therefore, servers and applications must also be considered
- What is important? Ordered list:
  1. Correctness
  2. Consistency
  3. Performance

TCP – Ubiquitous and Fragile

- Networks provide connectivity between applications running on hosts
  - From an application's perspective, the interface to "the other end" is a socket
  - Host operating system kernel provides socket interface, kernel implements TCP where the application can't see
  - Communication is between applications – mostly over TCP
- TCP – the fragile workhorse
  - TCP is (for very good reasons) timid – packet loss is interpreted as congestion
  - Like it or not, TCP is used for the vast majority of data transfer applications (more than 95% of ESnet traffic is TCP)
  - Packet loss in conjunction with latency is a performance killer

A small amount of packet loss makes a huge difference in TCP performance



Working With TCP In Practice

- Far easier to support TCP than to fix TCP
  - People have been trying to fix TCP for years – limited success
  - Like it or not we're stuck with TCP in the general case
- Pragmatically speaking, we must accommodate TCP
  - Sufficient bandwidth to avoid congestion
  - Zero packet loss
  - Verifiable infrastructure
    - Networks are complex
    - Must be able to locate problems quickly
    - Small footprint is a huge win – small number of devices so that problem isolation is tractable
- What if I don't use TCP?
  - TCP benefits are significant, but are not the only reason for Science DMZ
  - Architecture, cost, operational benefits

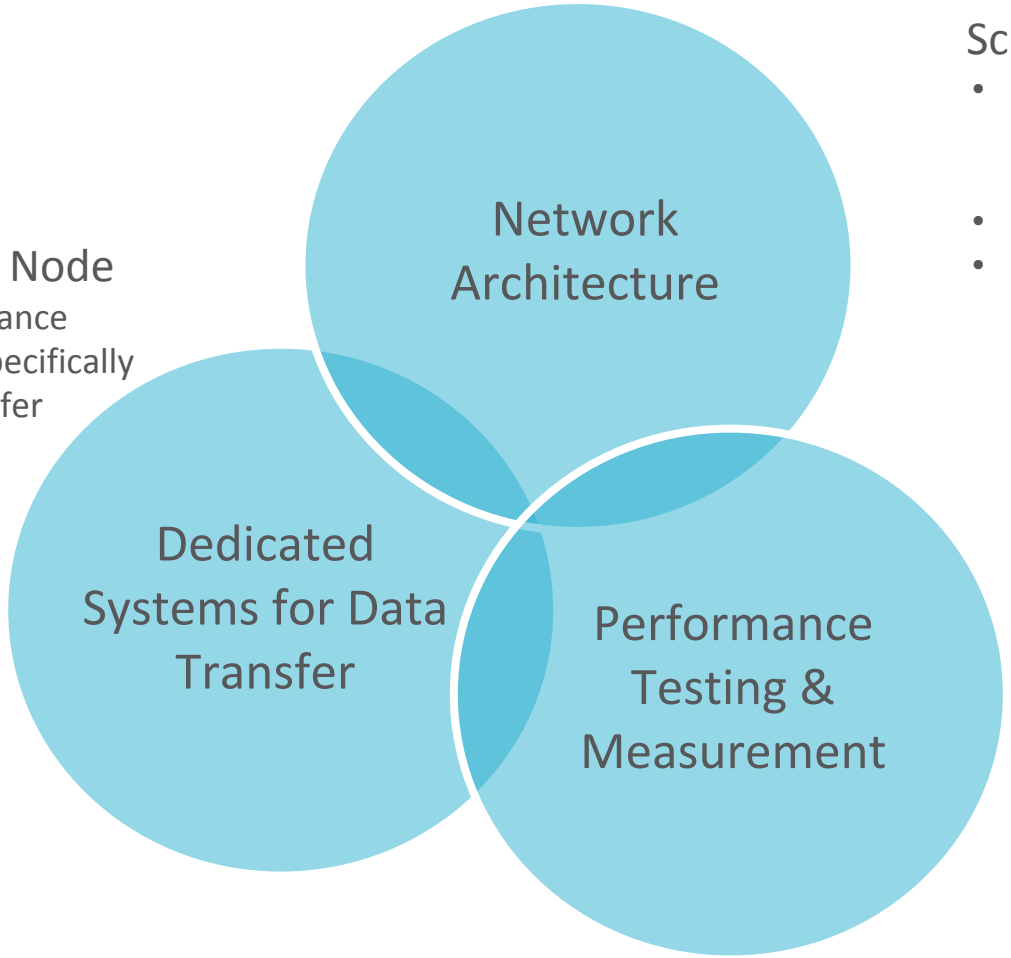
ESnet

Putting A Solution Together

- Effective support for TCP-based data transfer
  - Design for correct, consistent, high-performance operation
  - Design for ease of troubleshooting
- Easy adoption is critical
  - Large laboratories and universities have extensive IT deployments
  - Drastic change is prohibitively difficult
- Cybersecurity – defensible without compromising performance
- Borrow ideas from traditional network security
  - Traditional DMZ
    - Separate enclave at network perimeter ("Demilitarized Zone")
    - Specific location for external-facing services
    - Clean separation from internal network
  - Do the same thing for science – ***Science DMZ***

ESnet

The Science DMZ Design Pattern



Data Transfer Node

- High performance
- Configured specifically for data transfer
- Proper tools

Science DMZ

- Dedicated network location for high-speed data resources
- Appropriate security
- Easy to deploy - no need to redesign the whole network

perfSONAR

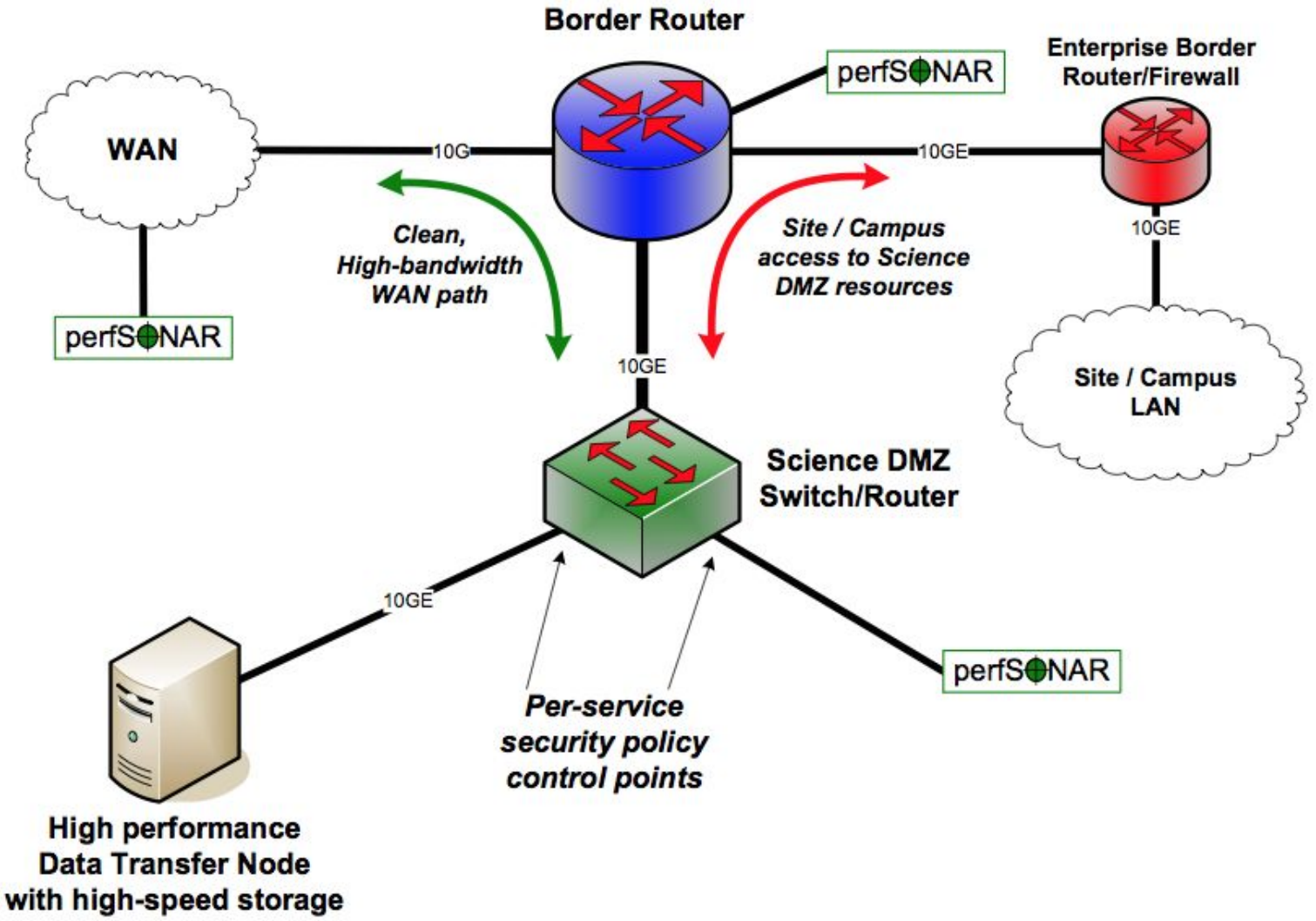
- Enables fault isolation
- Verify correct operation
- Widely deployed in ESnet and other networks, as well as sites and facilities

ESnet

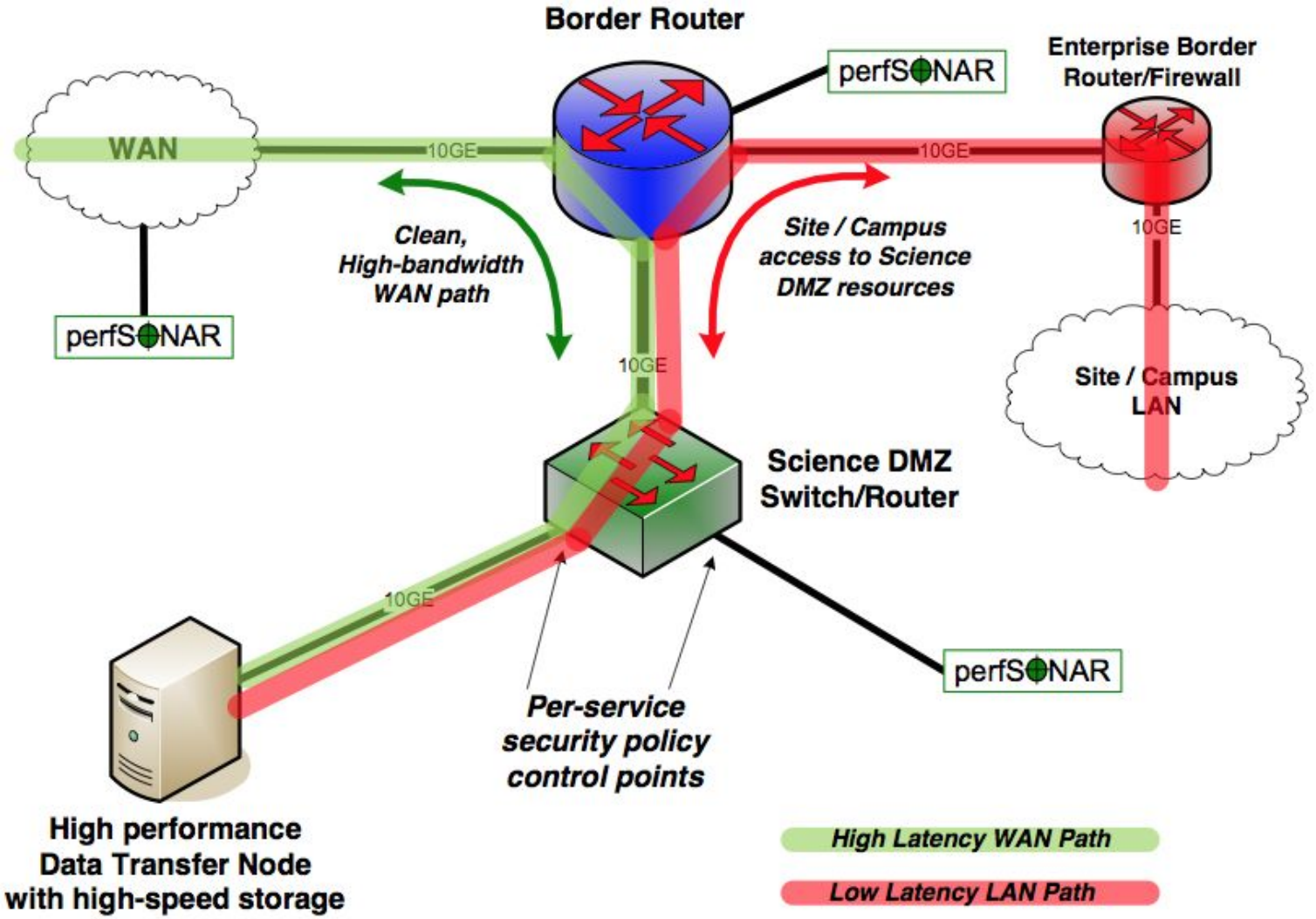
Abstract or Prototype Deployment

- (This section is for your system administrator – send them to me, use engage@es.net)
- Add-on to existing network infrastructure
  - All that is required is a port on the border router
  - Small footprint, pre-production commitment
- Easy to experiment with components and technologies
  - DTN prototyping
  - perfSONAR testing
- Limited scope makes security policy exceptions easy
  - Only allow traffic from partners
  - Add-on to production infrastructure – lower risk than rebuilding existing infrastructure

Science DMZ Design Pattern (Abstract)



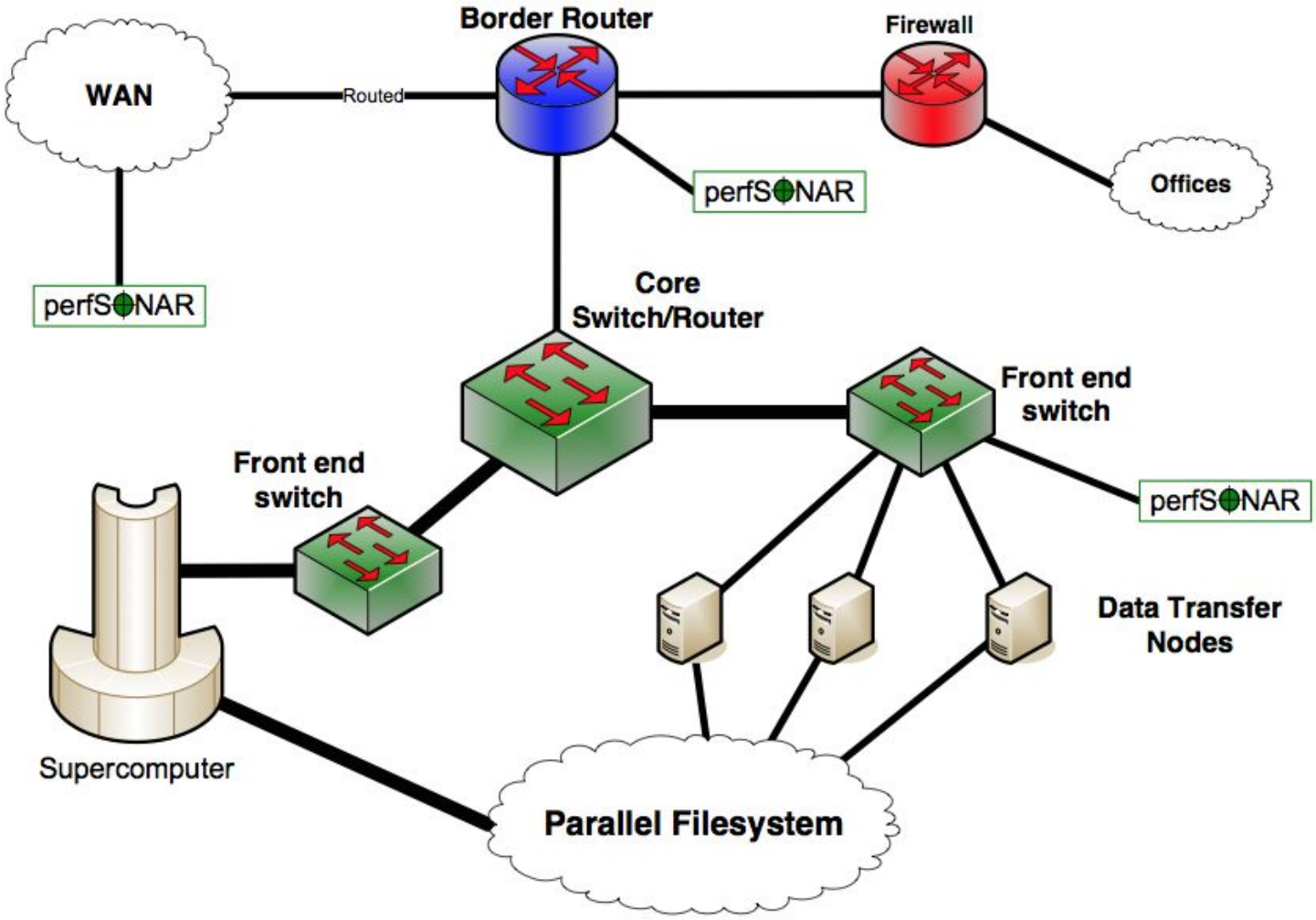
Local And Wide Area Data Flows



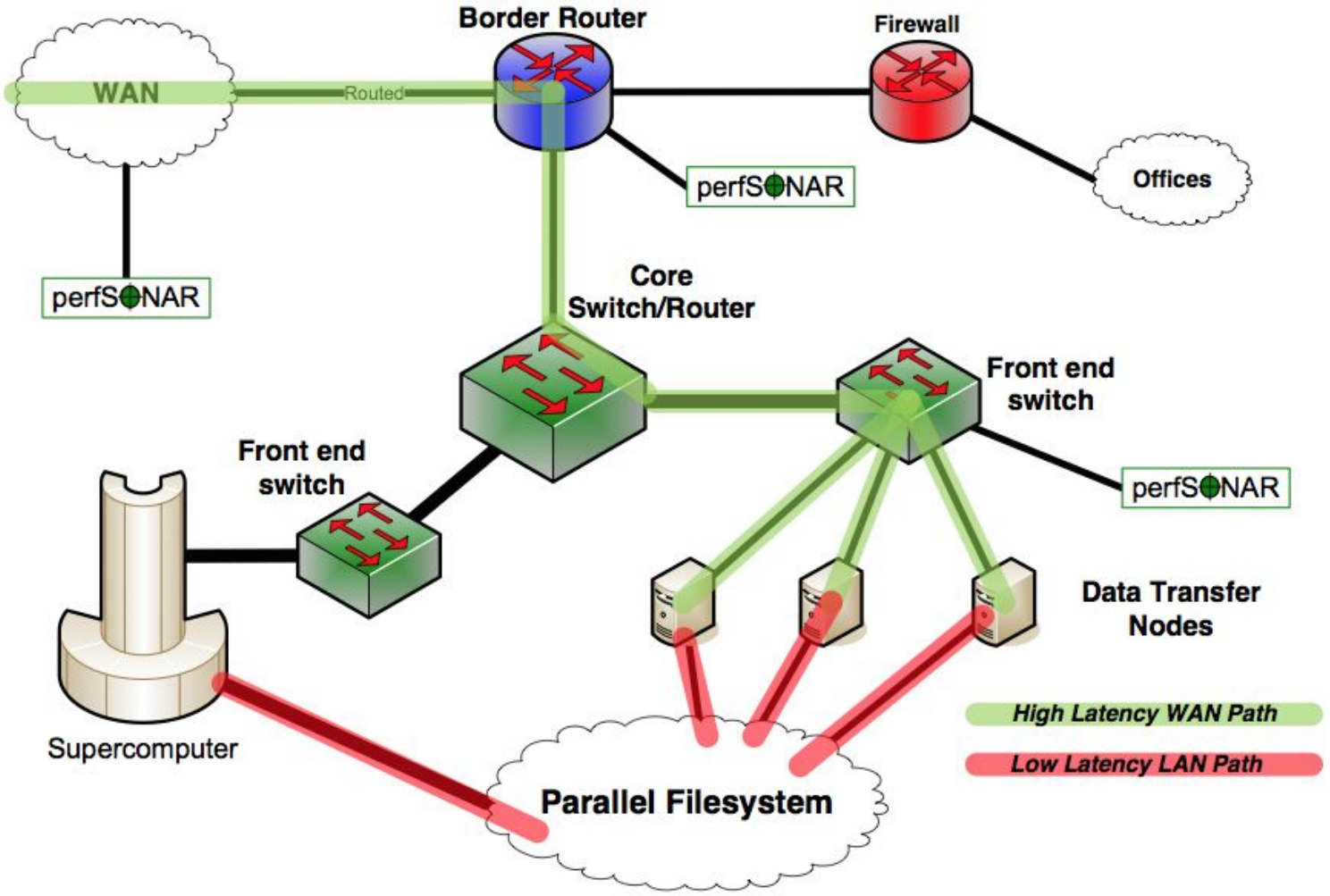
Supercomputer Center Deployment

- High-performance networking is assumed in this environment
  - Data flows between systems, between systems and storage, wide area, etc.
  - Global filesystem often ties resources together
    - Portions of this may not run over Ethernet (e.g. IB)
    - Implications for Data Transfer Nodes
- "Science DMZ" may not look like a discrete entity here
  - By the time you get through interconnecting all the resources, you end up with most of the network in the Science DMZ
  - This is as it should be – the point is appropriate deployment of tools, configuration, policy control, etc.
- Office networks can look like an afterthought, but they aren't
  - Deployed with appropriate security controls
  - Office infrastructure need not be sized for science traffic

HPC Center



HPC Center Data Path



Common Threads

- Two common threads exist in all these (and many other) examples
- Accommodation of TCP
  - Wide area portion of data transfers traverses purpose-built path
  - High performance devices that don't drop packets
- Ability to test and verify
  - When problems arise (and they always will), they can be solved if the infrastructure is built correctly
  - Small device count makes it easier to find issues
  - Multiple test and measurement hosts provide multiple views of the data path
    - perfSONAR nodes at the site and in the WAN
    - perfSONAR nodes at the remote site

Dedicated Systems – Data Transfer Node

- The DTN is dedicated to data transfer
- Set up **specifically** for high-performance data movement
  - System internals (BIOS, firmware, interrupts, etc.)
  - Network stack
  - Storage (global filesystem, Fibrechannel, local RAID, etc.)
  - High performance tools
  - No extraneous software
- ***Limitation of scope and function is powerful***
  - No conflicts with configuration for other tasks
  - Small application set makes cybersecurity easier

Data Transfer Tools For DTNs

- Parallelism is important
  - It is often easier to achieve a given performance level with four parallel connections than one connection
  - Several tools offer parallel transfers, including Globus/GridFTP
- Latency interaction is critical
  - Wide area data transfers have much higher latency than LAN transfers
  - Many tools and protocols assume a LAN
- Workflow integration is important
- Key tools: Globus, HPN-SSH

Data Transfer Tool Comparison

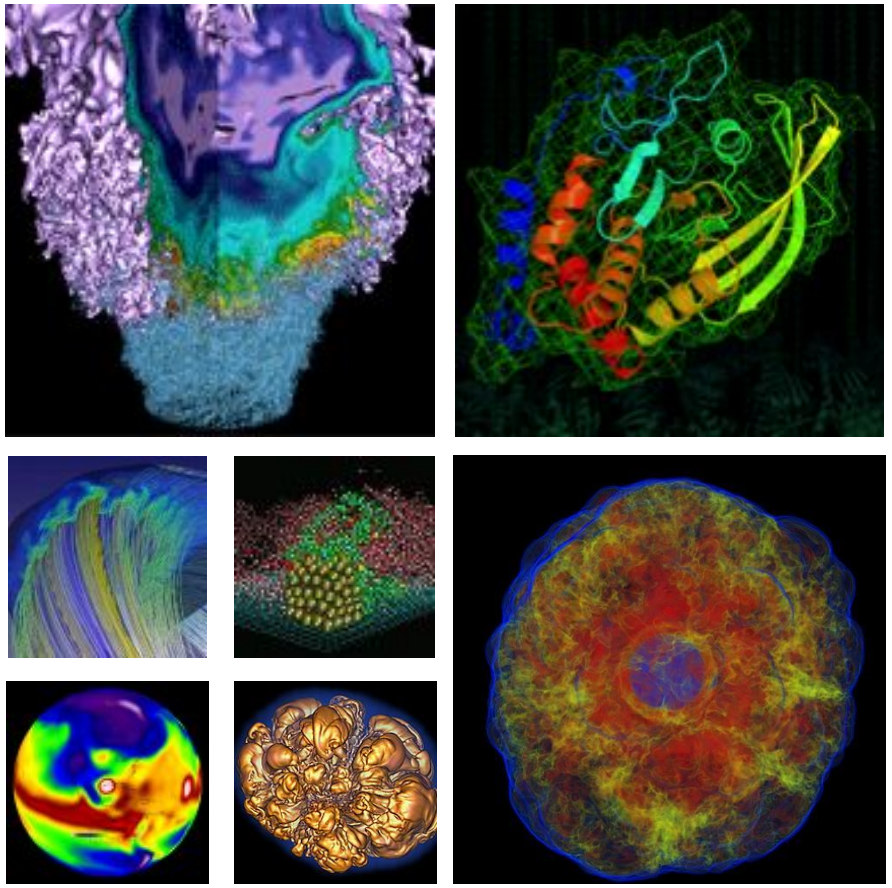
- In addition to the network, using the right data transfer tool is critical
- Data transfer test from Berkeley, CA to Argonne, IL (near Chicago). RTT = 53 ms, network capacity = 10Gbps.

Tool	Throughput
SCP:	140 Mbps
HPN patched SCP:	1.2 Gbps
FTP	1.4 Gbps
GridFTP, 4 streams	5.4 Gbps
GridFTP, 8 streams	6.6 Gbps

- NERSC DTNs have both HPN-SSH and Globus
- Key point – your local DTN and network significantly affect your ability to move data in and out of NERSC

ESnet

Data Transfers at NERSC



Jeff Porter
rjporter@lbl.gov
Data and Analytics Services

NERSC User Group Webinar

February 11, 2016

Dedicated Data Transfer Systems: Data Transfer Nodes

- **Data Transfer Nodes (DTN) are servers dedicated to data transfer**
  - At NERSC there are currently 4 nodes → dtn[01-04].nersc.gov
- **DTN features**
  - High bandwidth network interfaces
  - Access to global NERSC file systems
  - Tuned for efficient data transfers
  - Tuned for transferring large volumes of data between NERSC and other major facilities (ORNL, ANL etc.)
  - Can also move data between NERSC file systems and HPSS
- **Use DTNs to move large volumes of data in & out of NERSC**

Data Transfer Tools

- **Globus.org ("Globus Online")**
  - Reliable transfers of large data sets between sites or systems
- **scp**
  - copying individual files and directories
- **GridFTP (globus-url-copy)**
  - high performance CLI w/ grid certificates
- **hsi/htar**
  - data transfer into & out of HPSS
- **Science Gateways**
  - Community-specific services → will be a topic for future NUG

Globus.org

- **Managed 3rd party transfers**
  - https://www.globus.org/
  - Web based interface
  - REST/API for script access
- **Managed data operations:**
  - Endpoints
    - Sources & Destinations
    - NERSC endpoints:
      - #dtn
      - #cori, #edison, #pdsf, #jgi
  - Transfers
    - Files & directory selection
    - Initiate transfers
  - Activities
    - Status of transfer tasks
- **Web Interface Demo**