HPC's Pivot to Data

Suzanne Parete-Koon (OLCF)

OAK RIDGE LEADERSHIP COMPUTING FACILITY

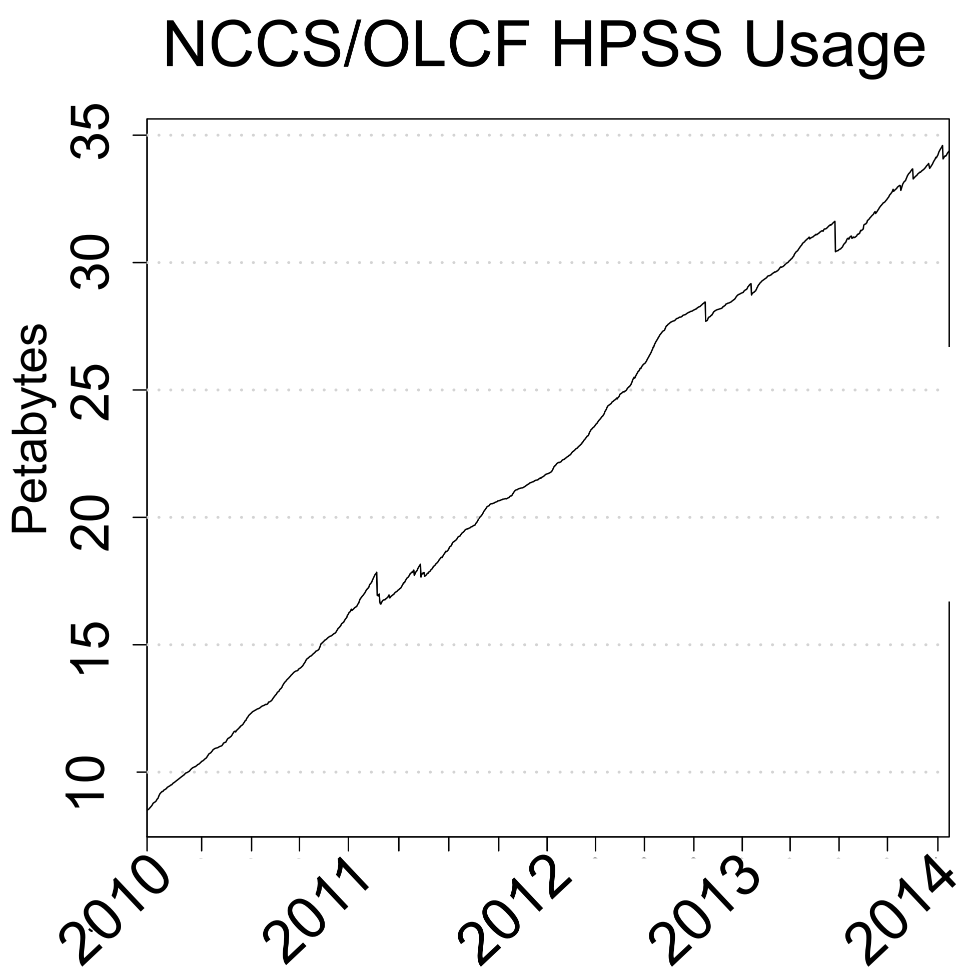
National Energy Research Scientific Computing Center

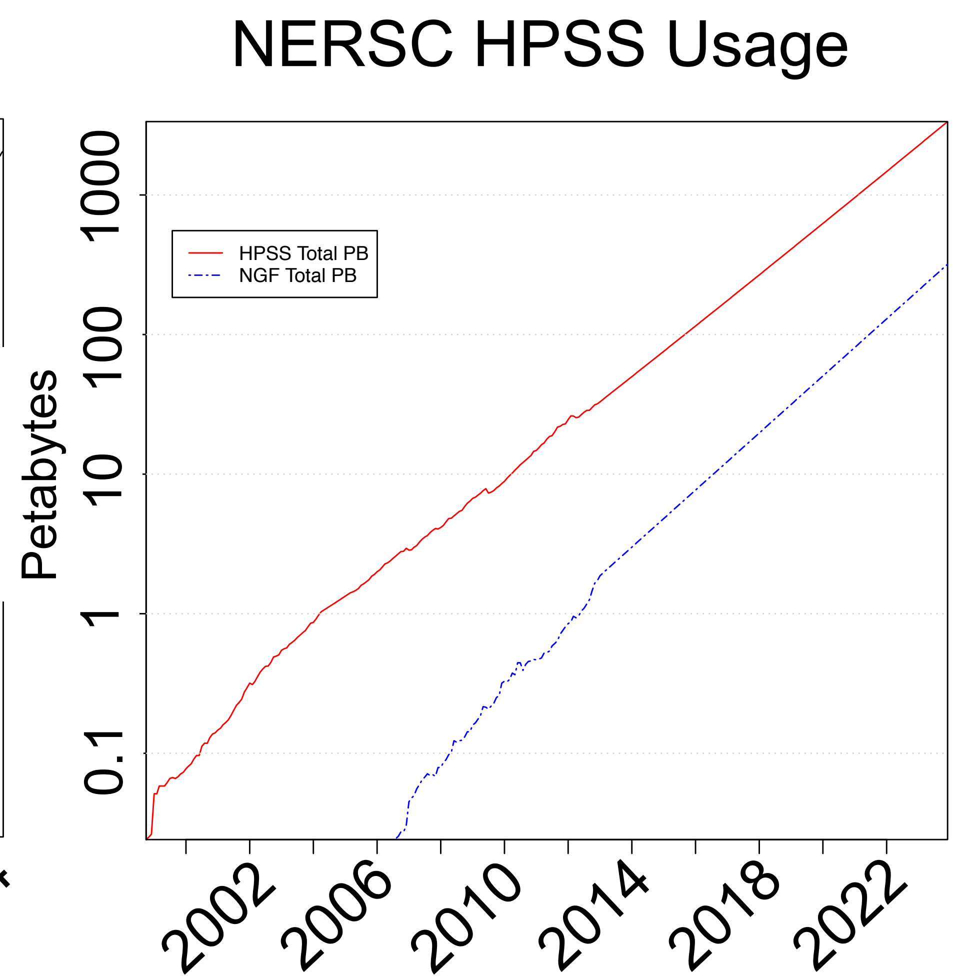
ESnet
Energy Sciences Network

What is HPC's Pivot to Data?

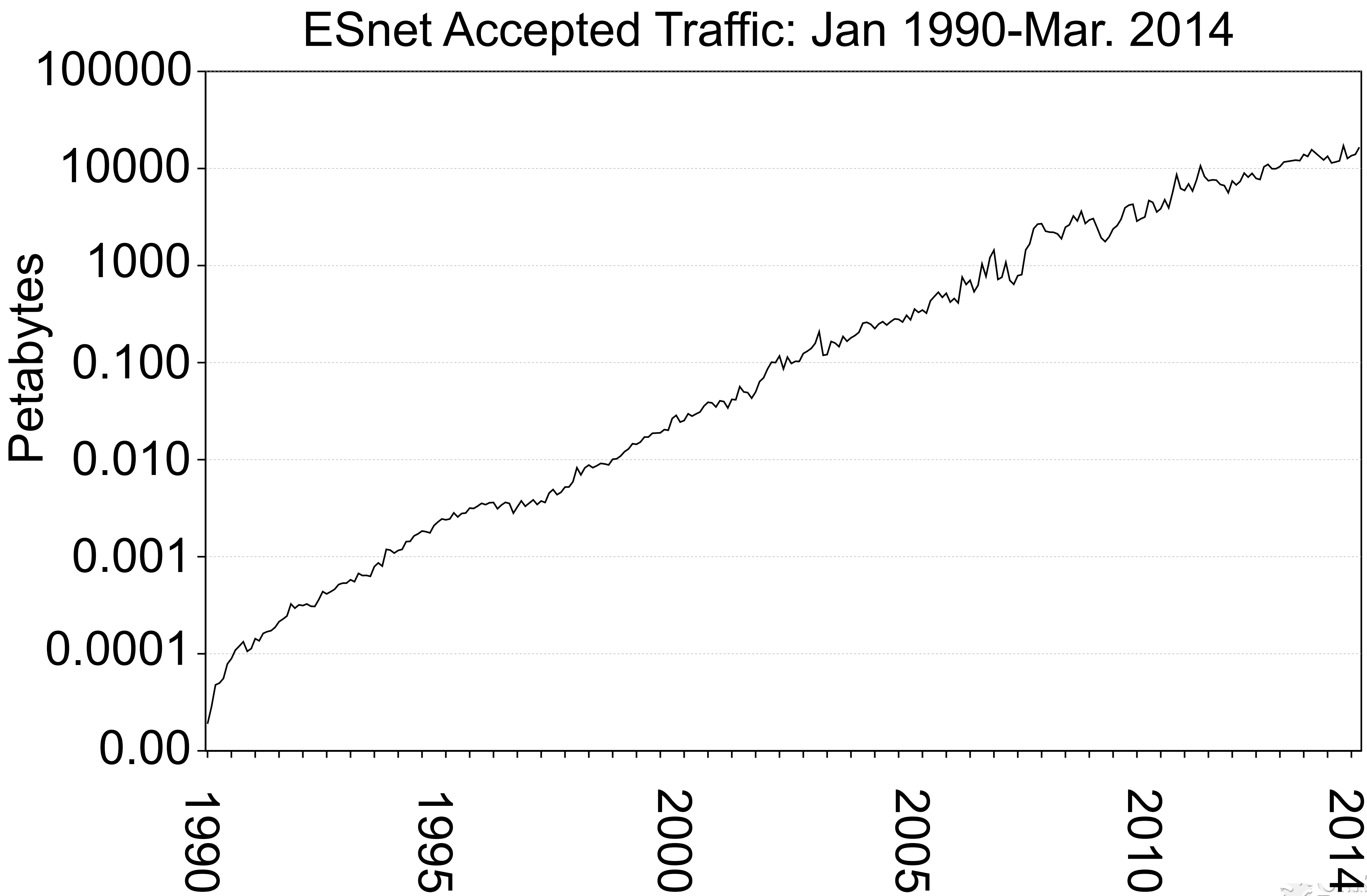
- The boundaries of the scientific computing ecosystem are being pushed beyond single centers providing HPC resources or any single experimental facility.
- Increasing data volume, variety, and velocity require additional resources to foster a productive environment for scientific discovery. This focus on data does not imply a reduction in importance of large-scale simulation capabilities
- Predicting how and why scientific workflows achieve their observed end-to-end performance is a growing challenge for scientists.
- This talk will focus on data movement- because it is an excellent example of a task that requires multi-facility collaboration to enable multi-facility scientific workflows

Data Growth

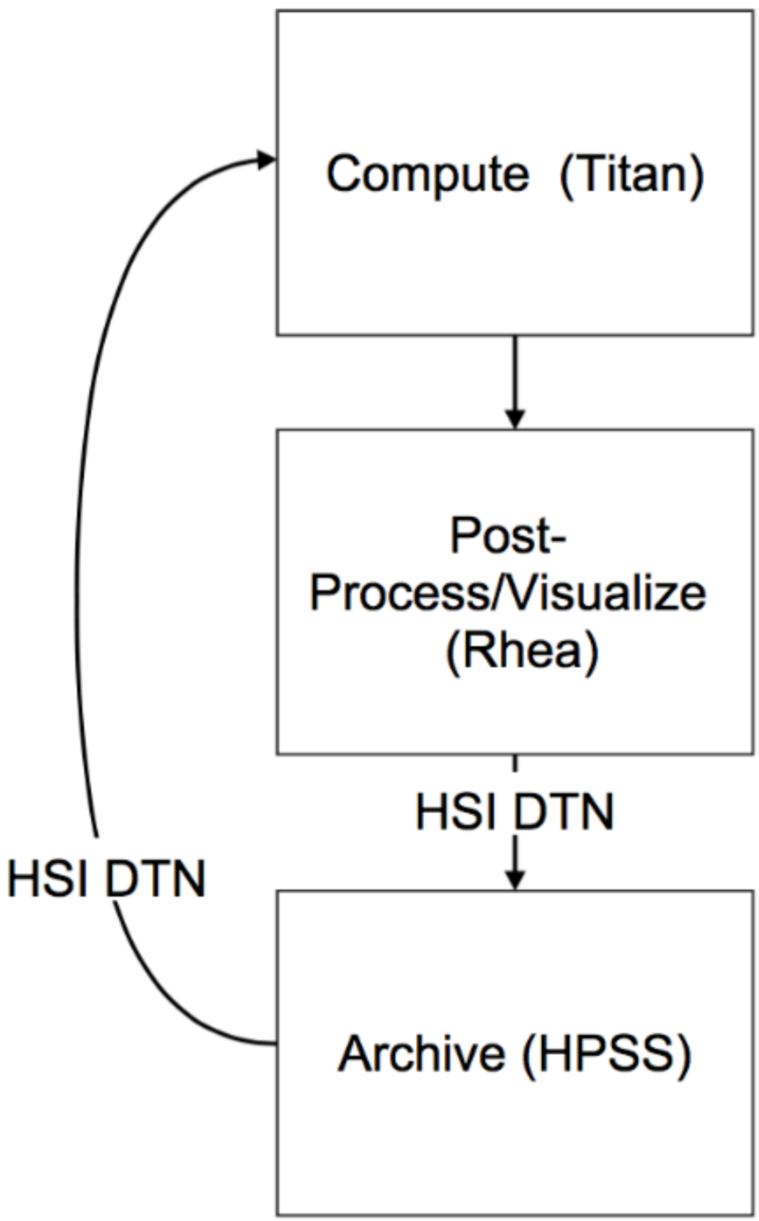

NCCS/OLCF HPSS Usage
Petabytes
10
15
20
25
30
35
2010
2011
2012
2013
2014


NERSC HPSS Usage
Petabytes
0.1
1
10
100
1000
HPSS Total PB
NGF Total PB
2002
2006
2010
2014
2018
2022

Data Movement



Inter-Workflow: One HPC Center



Even projects with workflows designed to utilize only one HPC center need efficient data transfer, because data is typically moved to more permanent storage at the end of a project.

Intra-workflows: Simulation

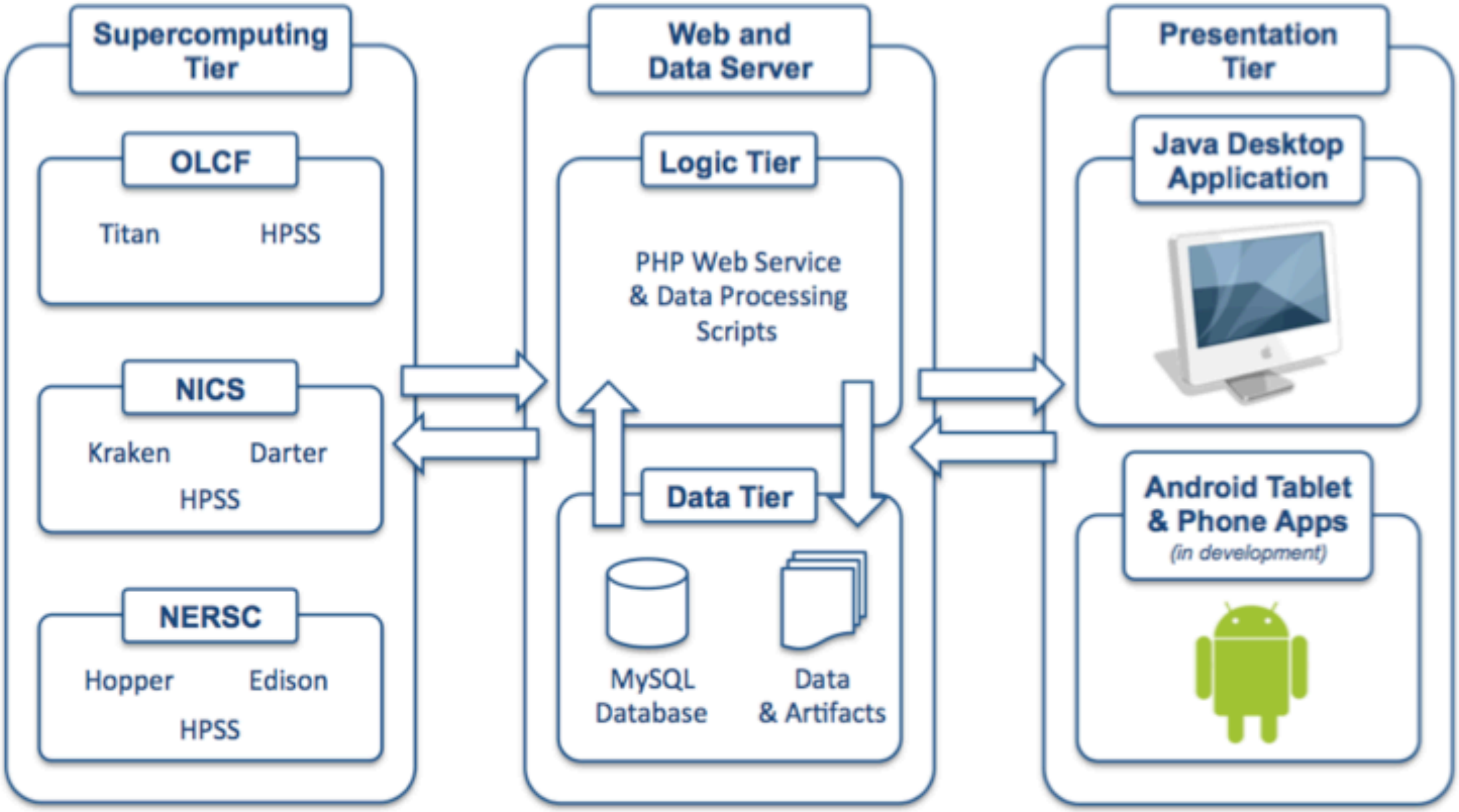
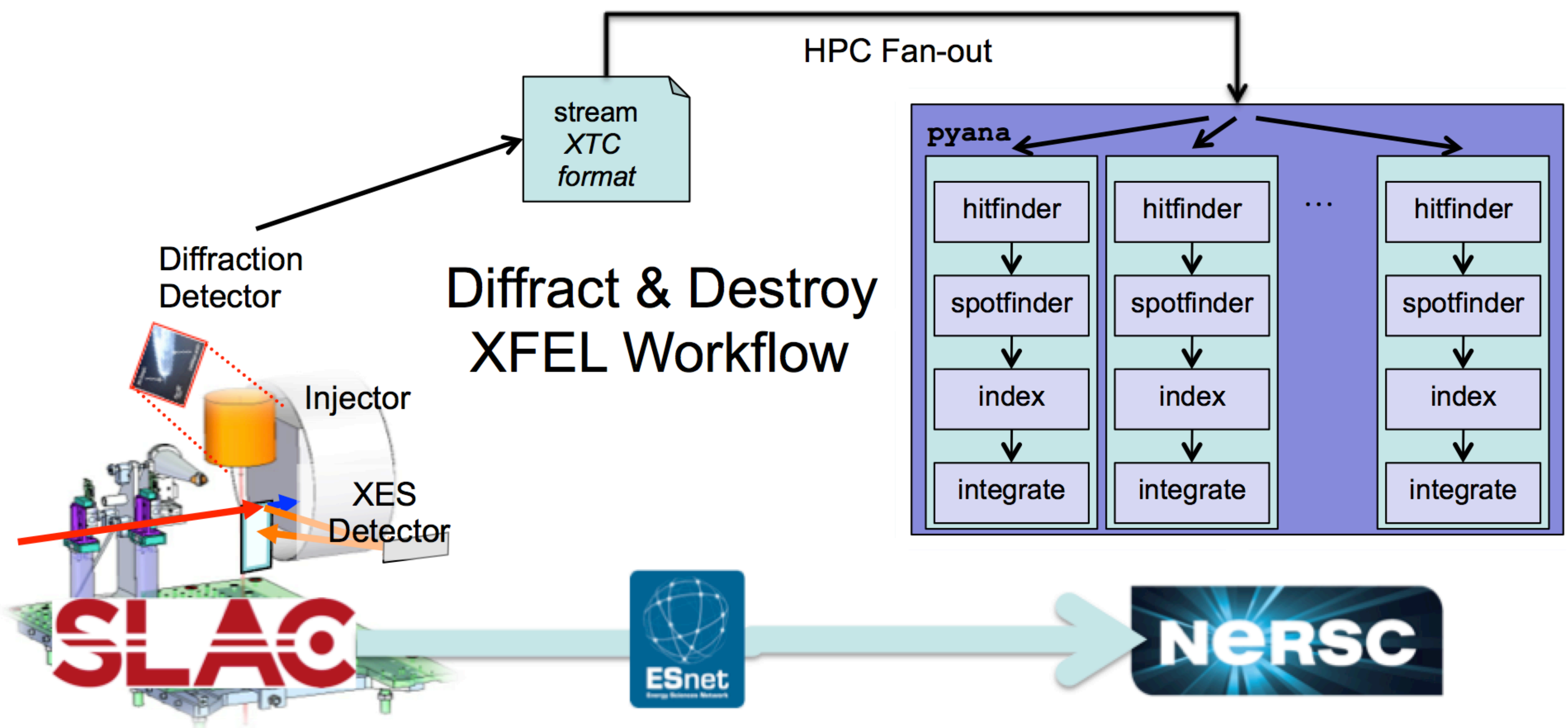


Figure 1: A schematic representation of Bellerophon's n-tier architecture.

Lingerfelt et al 2014, Near real-time data analysis of core-collapse supernova simulations with Bellerophon," in *The International Conference on Computational Science 2015 Proceedings, in press*, 2014.

Intra-workflows : Experiment



Single particle X-ray Crystallography
10E6 diffractive images reconstructed to reproduce angstrom scale 3-D structure.

Enabling Intra-workflow: Remote Transfers

Centers

- How can we enable secure data movement?
- What transfer tools are best for the system?
- How can we efficiently monitor our network?

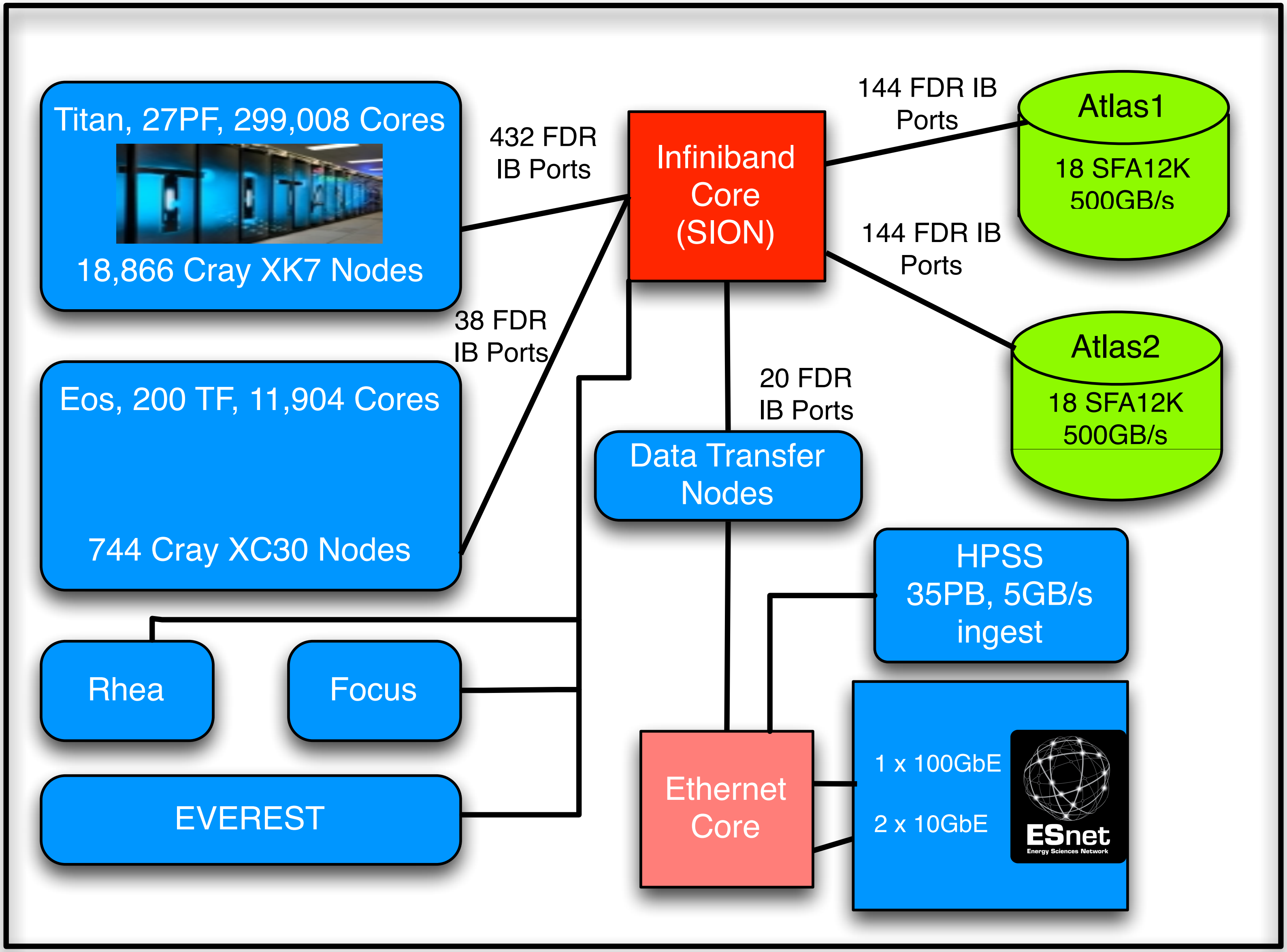
Users

- Can I scrip the transfer? Click and forget ?
- How long will the transfer take?
- What tools do you recommend?

The Science DMZ

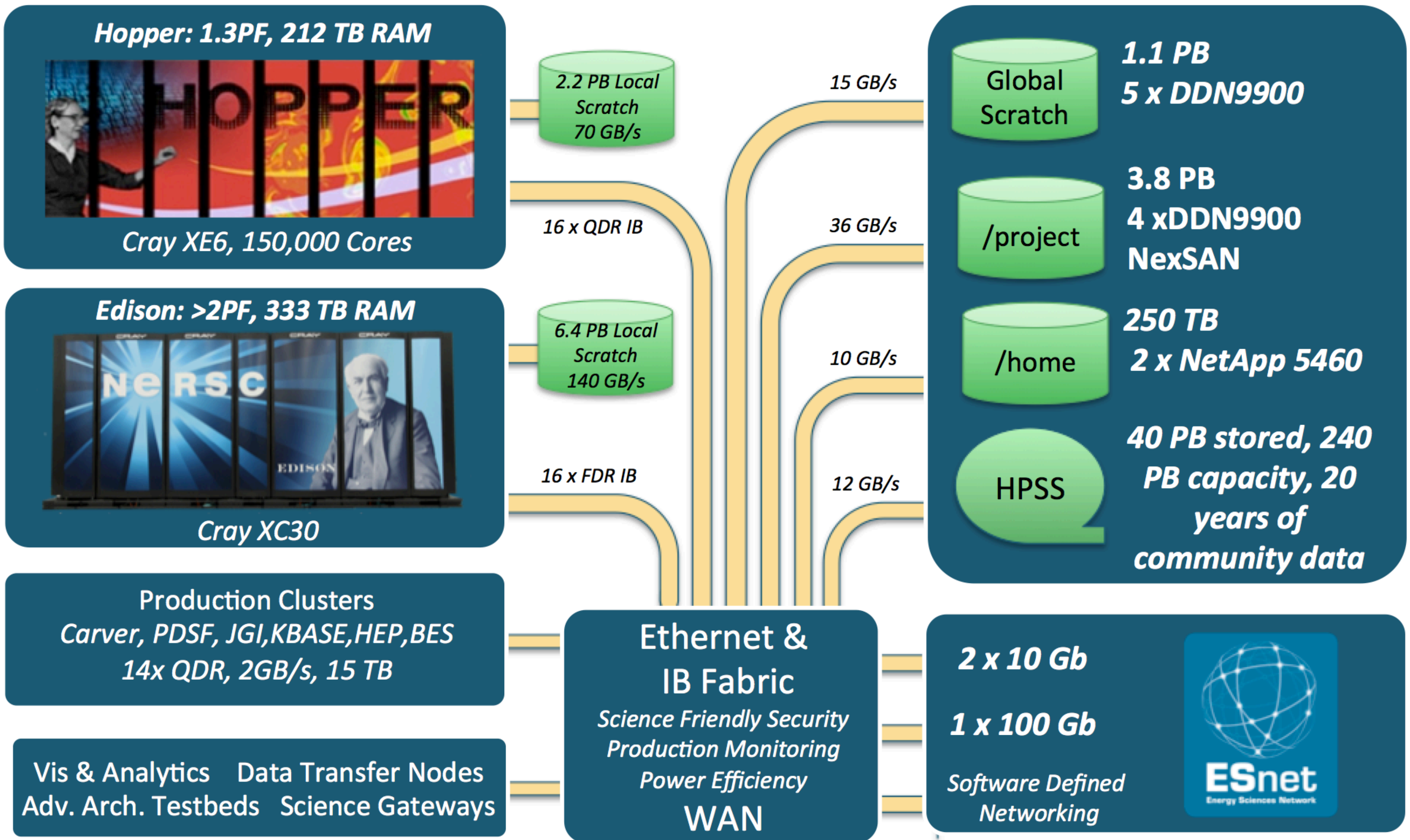
- A network architecture explicitly designed for high-performance applications, where the science network is distinct from the general-purpose network
- The use of dedicated systems for data transfer
- Performance measurement and network testing systems that are regularly used to characterize the network and are available for troubleshooting
- Security policies and enforcement mechanisms that are tailored for high performance science environments

OLCF Infrastructure

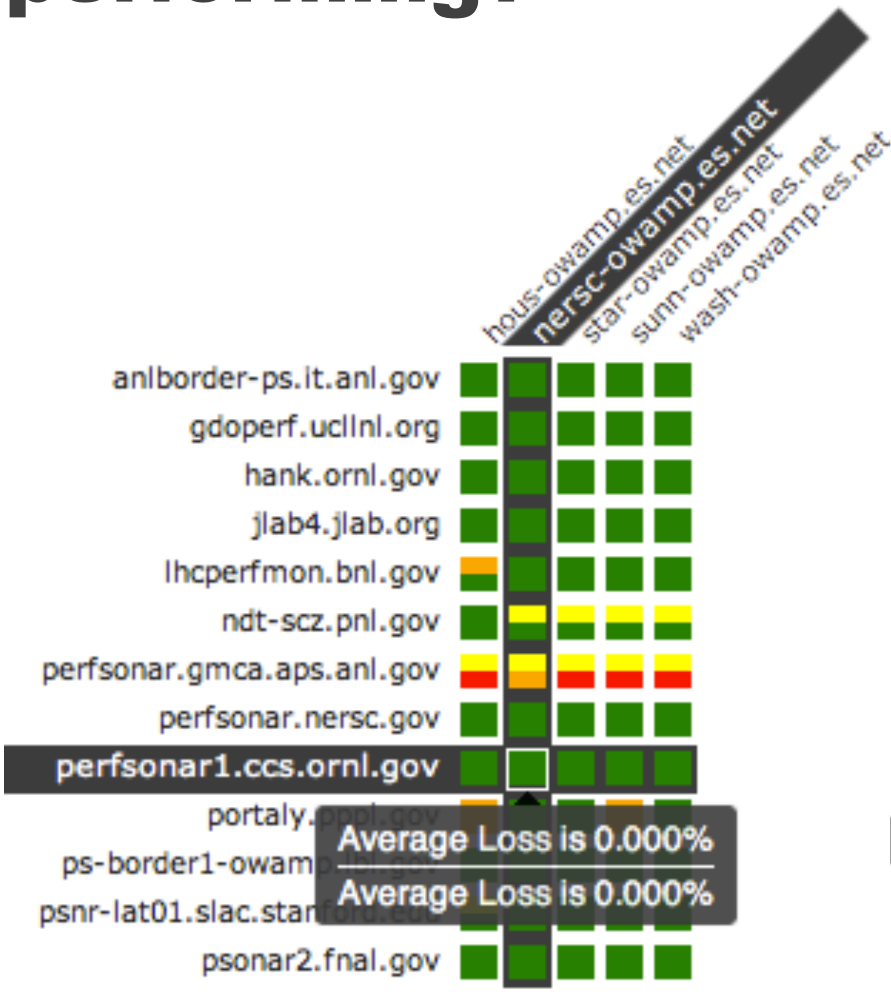


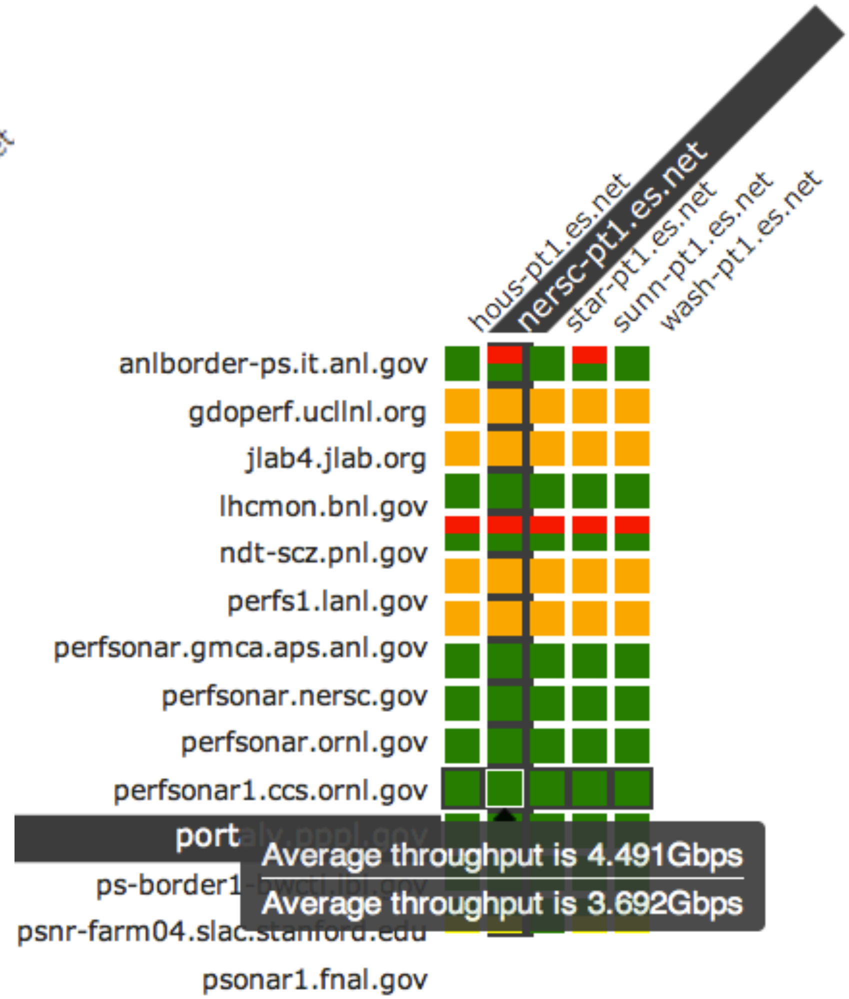
NERSC Infrastructure

NERSC Facility: Computing, Data and Interconnect



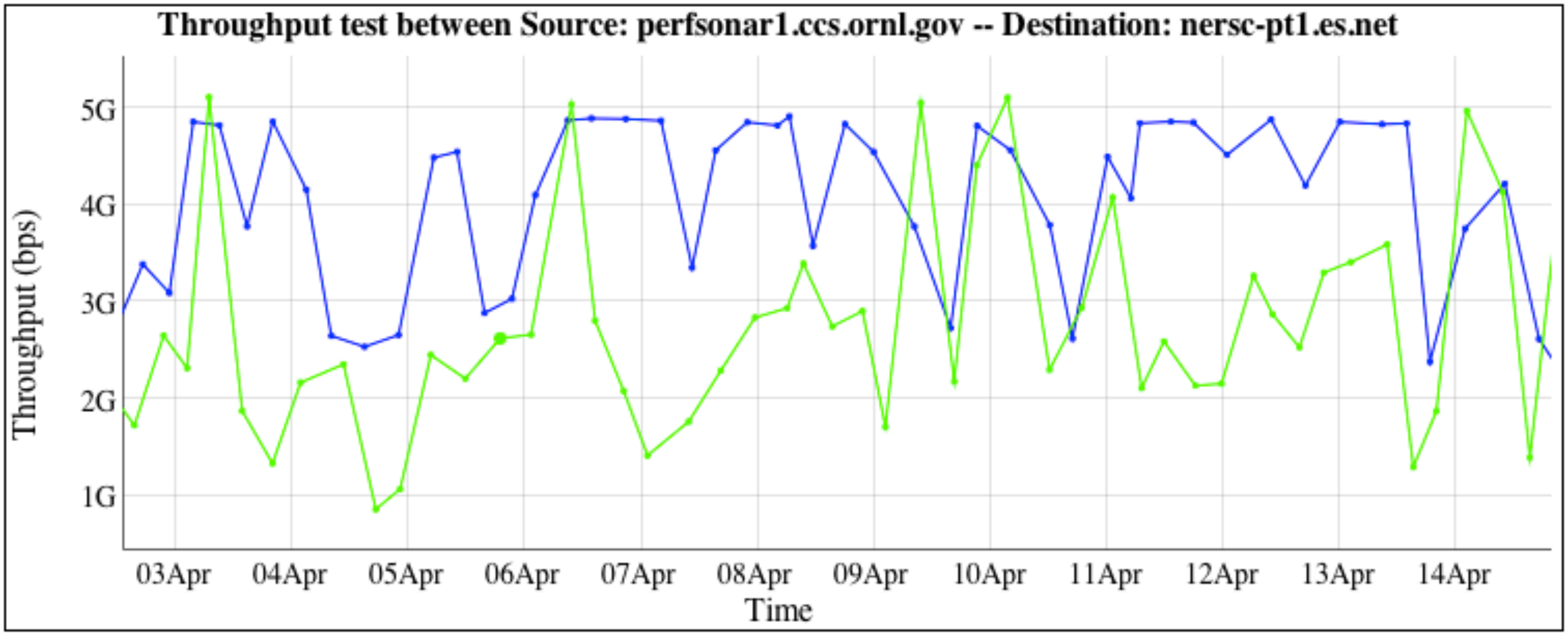
Perfsonar: How well is the transfer performing?



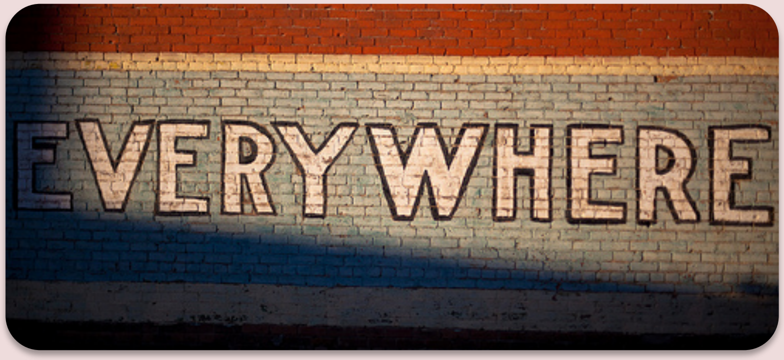


http://fasterdata.es.net/performance- testing/perfsonar/ perfsonar- dashboard/

Perfsonar: How is my transfer performing?



The Pros and Cons of SCP and Rsync



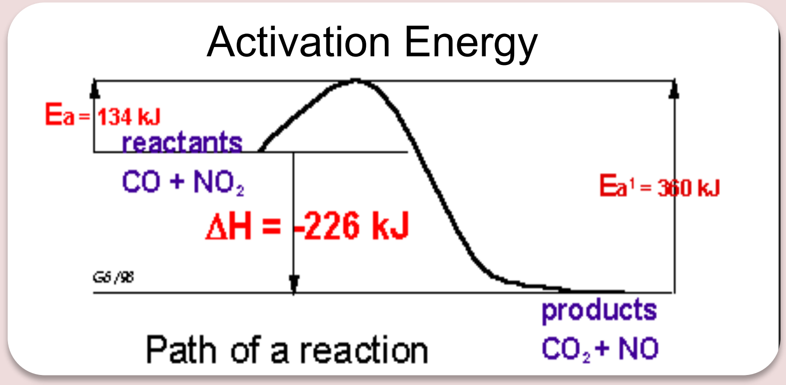
SCP Common on all Unix-like systems

Can be scripted into a workflow when the transfer destination allows for passwordless SSH logins.

Single stream: therefore slow and take poor advantage of WAN.

The common version do not allow much control over buffer size and fault checking.

The Pros and Cons Parallel Streams: GridFTP, globusonline, bbcp



Multi-streams allow fast transfers on WAN.

Many Options for customization.

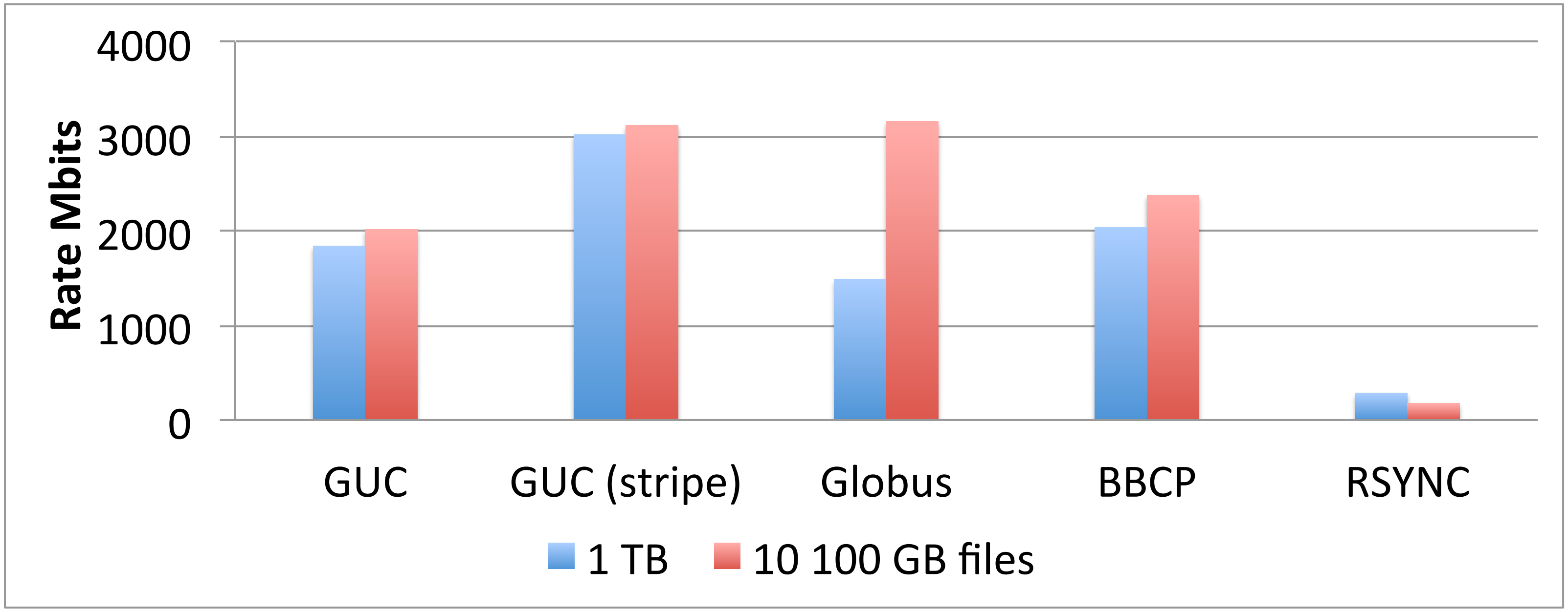
Globus is extremely user-friendly once it is set up.

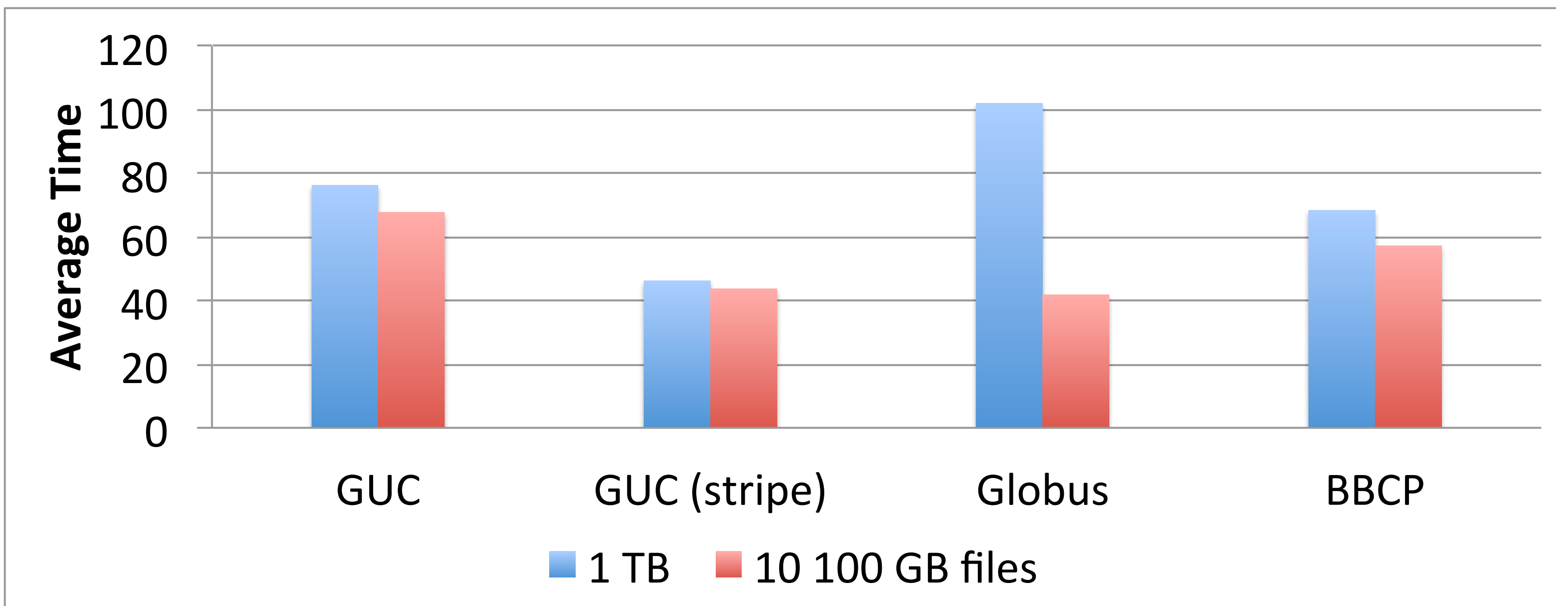
The software must be available at both end of the transfer and the availability varies.

Setup required.

Security policies impact ease of use.

What rates can I expect?





Best tool features for the system?

THROUGHPUT TEST USING GUC

Test	Average Rate Mbit/s	% Network Capacity
OLCF Mem to NERSC Mem	4144	100
NERSC Disk to ORNL Mem	3335	80
OLCF Disk to NERSC Mem	2078	50
1 1TB file	1932	46

THROUGHPUT TEST USING GUC STRIPE: UTILIZING TWO DTNS

Test	Average Rate Mbit/s	% Network Capacity
OLCF Mem to NERSC Mem	8874	100
NERSC Disk to ORNL Mem	5339	60
ORNL Disk to NERSC Mem	4646	52
1 1TB file	2965	33

DISK THROUGHPUT MEASUREMENT WITH FIO

Site	IO Operation	Number of Processes	Average Rate Mbit/s
NERSC	read	1	14859
NERSC	read	2	13193
ORNL	write	1	2113
ORNL	write	2	3684
ORNL	write	8	11642

This filesystem benchmarking with fio shows that the performance improvement from GUC to GUC with striping comes from parallelism at the Lustre client.

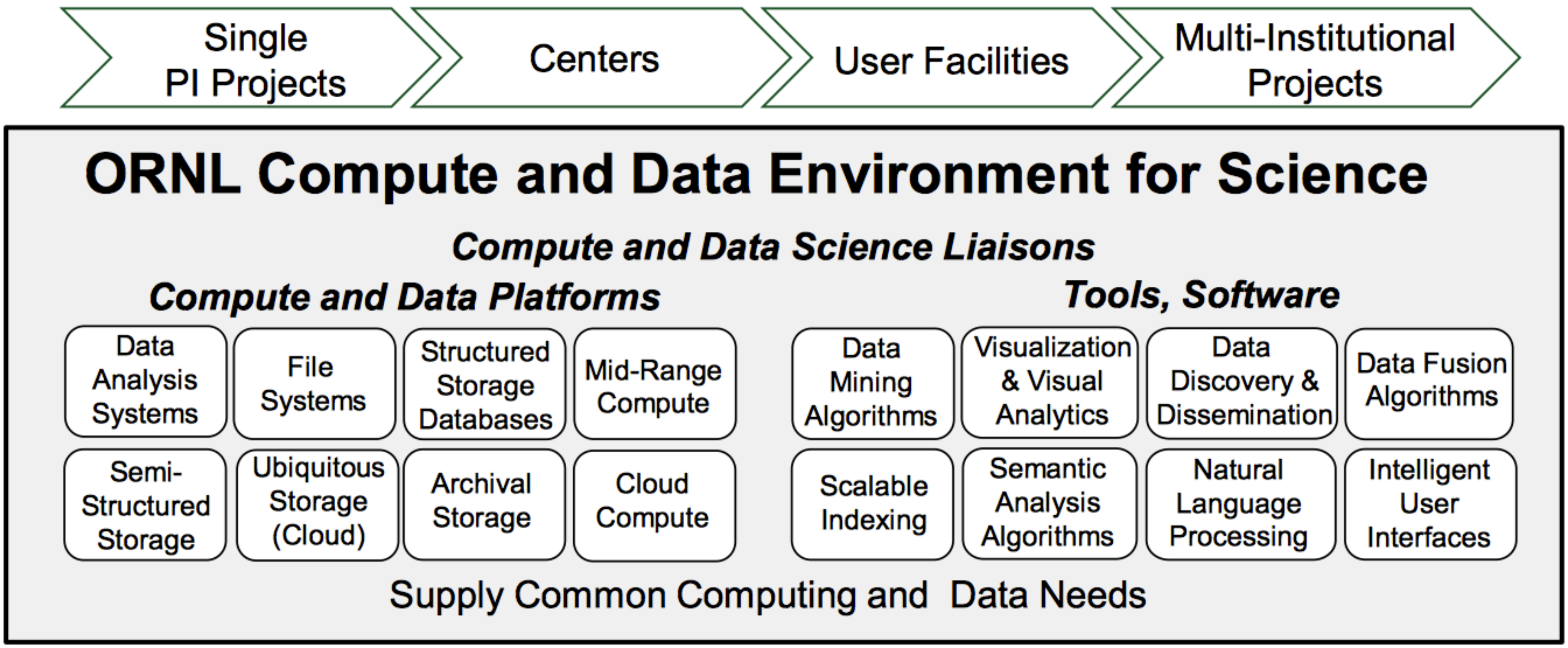
Case Studies

	Year	Methods	Data	Transfer
Combustion research challenge for multi-lab work-flow	2007-2009	dtns/SCP/ Kepler workflow	10TB	OLCF→NERSC
20th Century Climate Reanalysis	2011	HPSS Direct	40 TB	OLCF→NERSC
DNS Combustion	2013	dtns/ Globus	80TB	OLCF→ALCF
LSCS	2013-14	dtns / bbcp Globus	130 TB	SLAC→NERSC

Road Maps: Data-Centric Services

- ORNL is establishing a Compute & Data Environment for Science (CADES). OLCF is a key partner.
- CADES will be a HUB to share data infrastructure and compute & data science capabilities with and among many projects.

Compute and Data Environment for Science targets full range of needs



Conclusion

- Success for workflows with large data movement depends critically on a ”Science DMZ” like approach to connectivity
- HPC centers and their users benefit from collaboration between different centers and facilities and collaboration between centers and projects.
- Computing centers traditionally focused on large-scale simulation are expanding their repertoire to include user-facing data services.
- Data services are built around a long-lived resources, the data itself. Both HPC architecture and allocation of will need to adapt to this longevity.

Data Transfer Working Group

CADES	ESnet	OLCF	NERSC
G. Shipman	E. Dart	B.Caldwell	S. Cannon
	J. Zurawski	J. Hill	D. Hazen
		C. Layton	J. Hick
		S. Parete-Koon	D. Skinner
		D. Pelfrey	
		H. Nam	
		J. Wells	

This manuscript has been authored by an author at Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 with the U.S. Department of Energy. This work used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.