Superfacility
Next step in data science discovery

Shane Canon, LBNL/NERSC

MAGIC
October 7, 2015

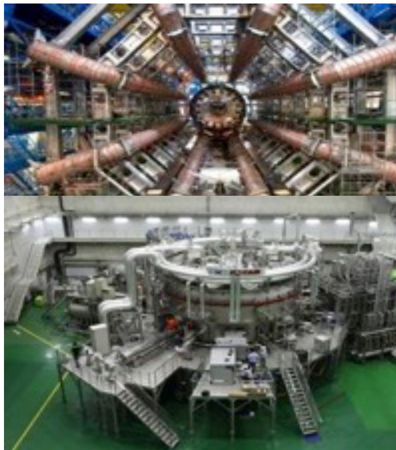
Acknowledgements

Slides courtesy of:

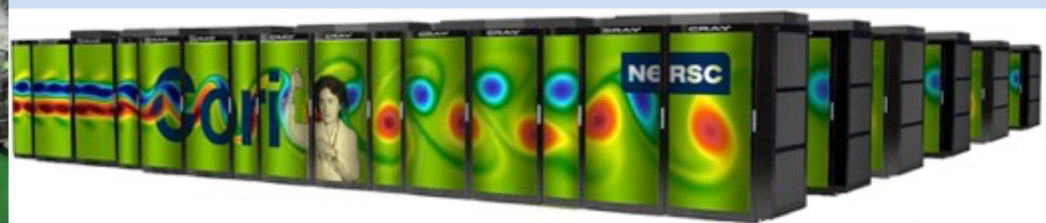
Katie Antypas (NERSC) and

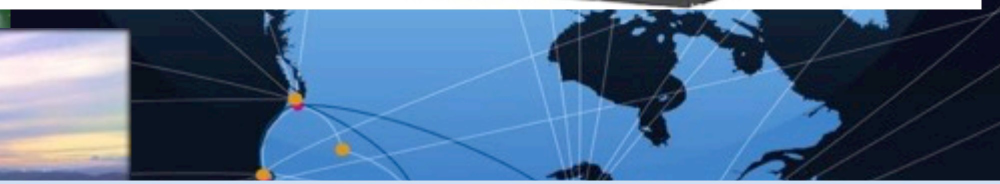
Inder Monga (ESnet)

Shifter: Doug Jacobsen (NERSC)



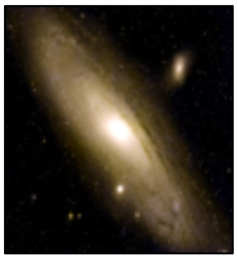
Experimental and observational science is at crossroads



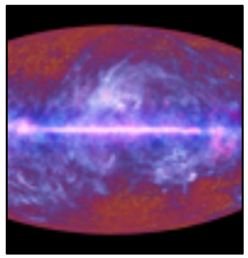


- Data volumes are increasing faster than Moore's Law
- Facility data exceeds local computing and networking capabilities
- Infeasible to put a supercomputing center at every experimental facility

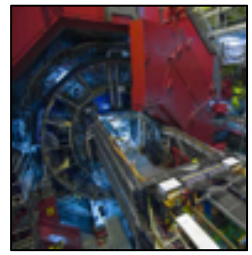
Berkeley Lab has a history of serving large data-rich collaborations



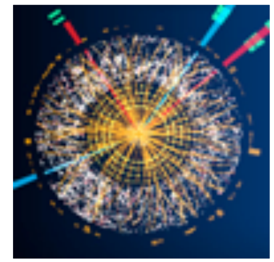
Palomar Transient
Factory
Supernova



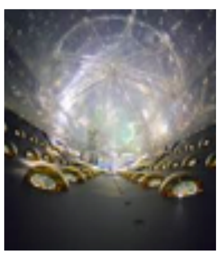
Planck Satellite
Cosmic Microwave
Background Radiation



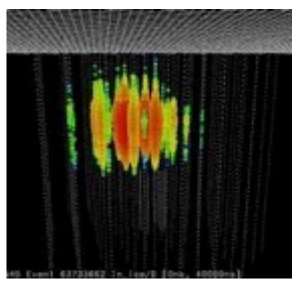
ALICE
Large Hadron Collider



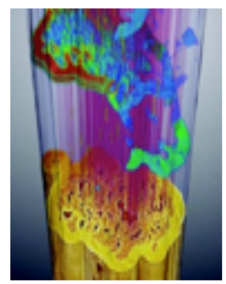
ATLAS
Large Hadron Collider



Daya Bay
Neutrinos



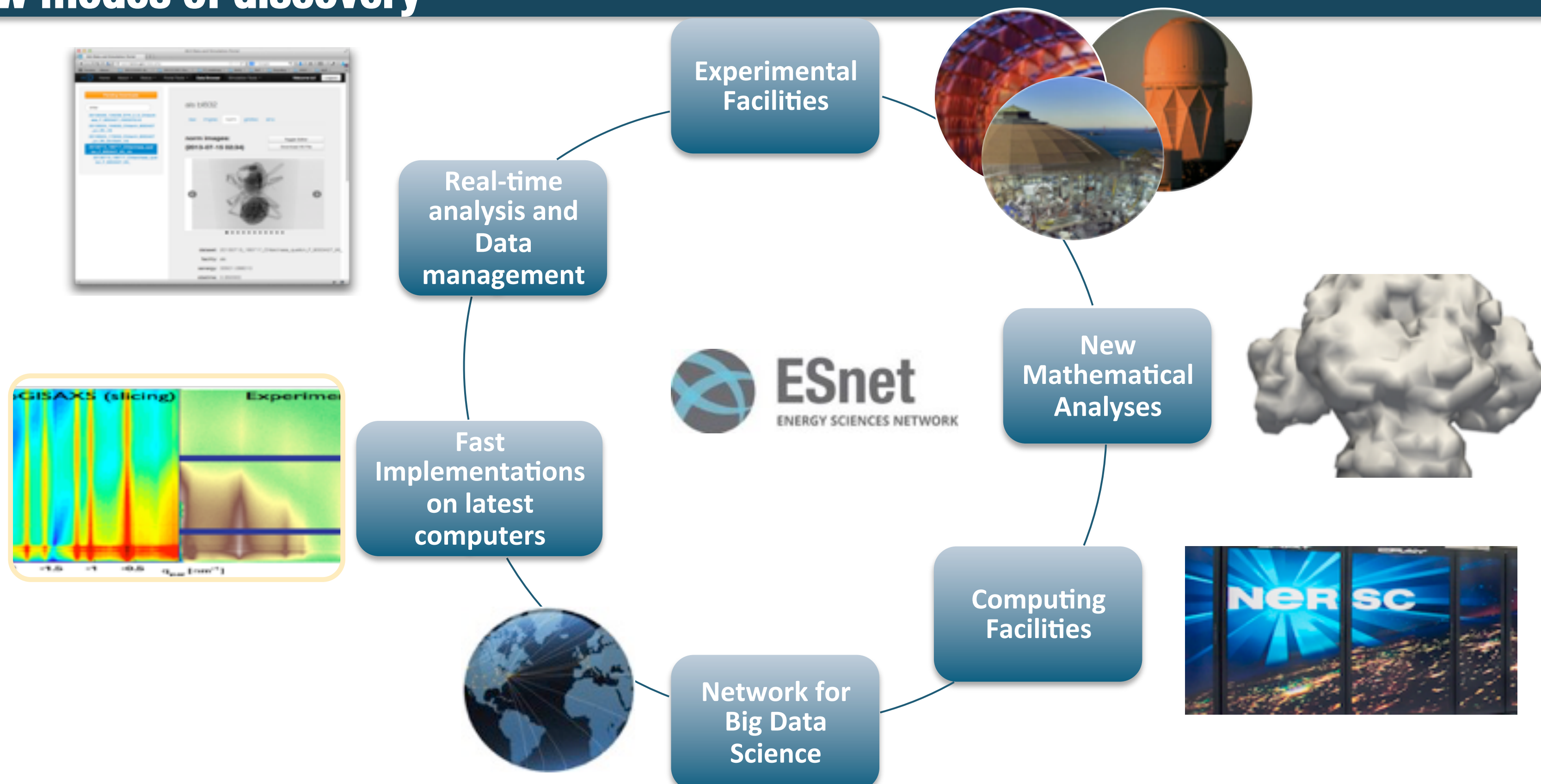
Ice Cube
Neutrino Detector



ALS/LCLS
Light Sources

Joint Genome Institute
Bioinformatics

Superfacility Vision: A network of connected facilities, software and expertise to enable new modes of discovery

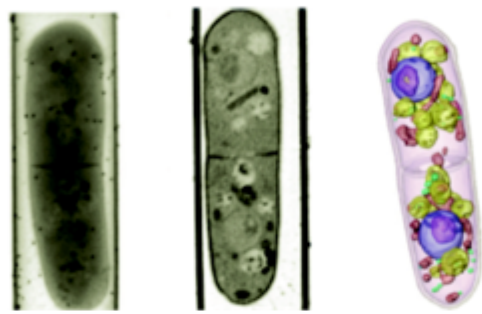


Superfacilities can transform scientific discovery

Palomar Transient Factory

Enabling new capabilities

ALS tomography beamline



Coupling data analysis and simulation

Science Gateways

Sharing datasets more widely

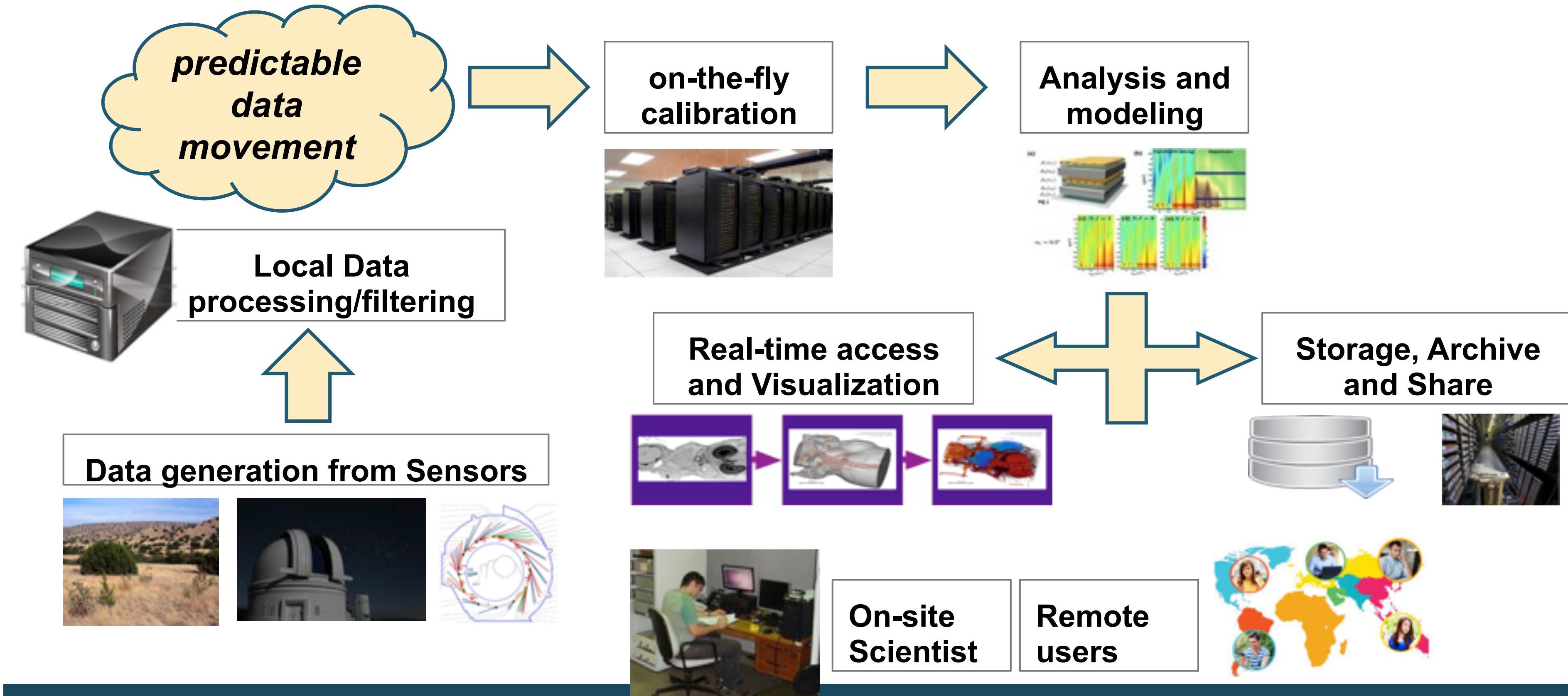
All too common process of discovery

Superfacility Prototype : Tight coupling of facilities with the network speeds up discovery

- Real-time analysis of ‘slot-die’ technique for printing organic photovoltaics
- Using ALS + NERSC (SPOT Suite for reduction, remeshing, analysis)
- OLCF (HipGISAXS running on Titan w/ 8000 GPUs).
- Results presented at March 2015 meeting of American Physical Society by Alex Hexemer.
- Additional DOE contributions: GLOBUS (ANL), CAMERA (Berkeley Lab)

‘Eliminate boundaries between the Scientist and the world’s best Algorithms running on the best architecture for that code’ – Craig Tull

Based on our experiences supporting science from experimental facilities, we see a Common Design Pattern emerging



But, there are challenges to achieving our superfacility vision.

1. Unified computing architecture
2. Predictable, programmable networks
3. Workflows for seamless data movement
4. Productive user environment for data analysis

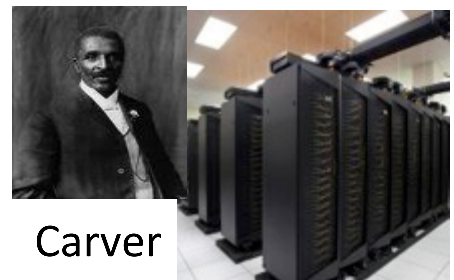
Goal: Create and deploy a superfacility architecture that is applicable to multiple disciplines

Challenge #1: A unified architecture for data analysis and simulation and modeling

NERSC has deployed separate systems for simulation and data analysis for many years

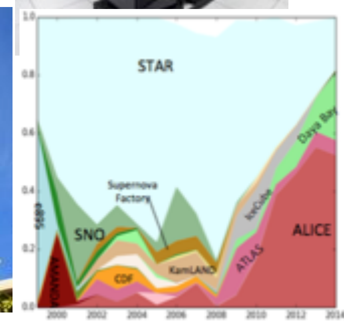
Simulation and Modeling Systems

Data Analysis Systems



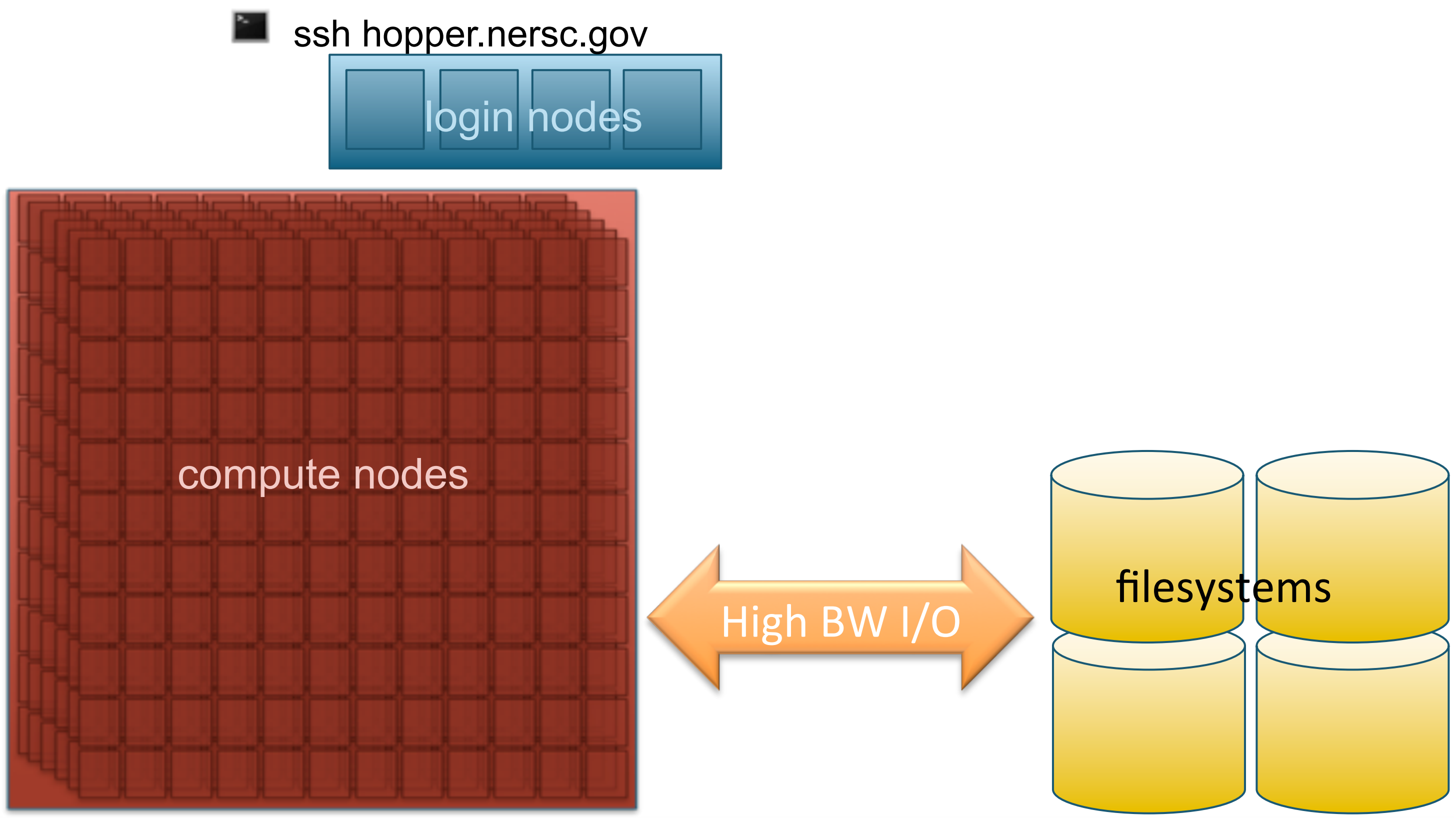
Carver

Genepool

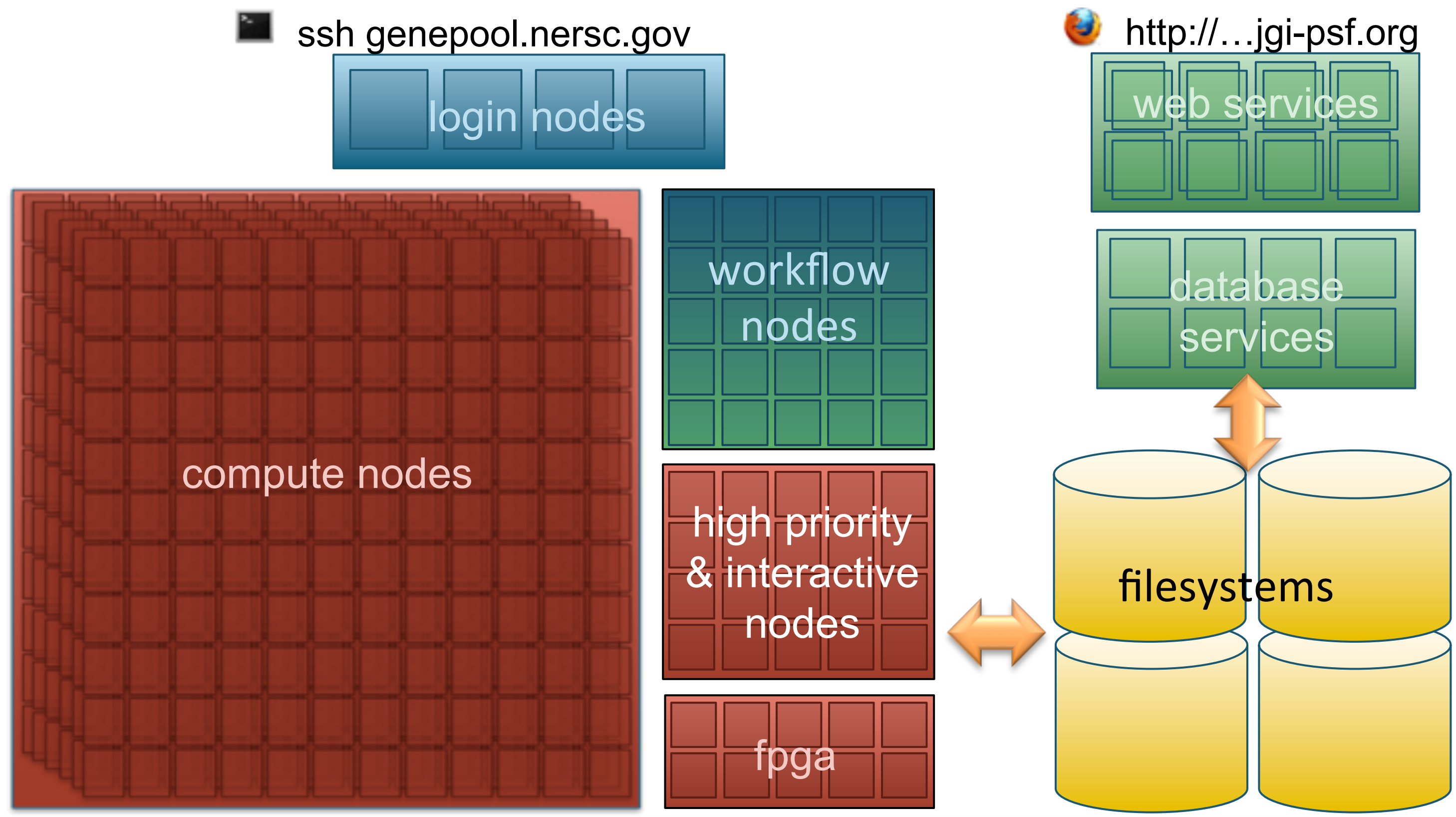


PDSF

Traditional Supercomputer



Joint Genome Institute's compute cluster at NERSC



But how different really are the compute and data intensive platforms?

Policy

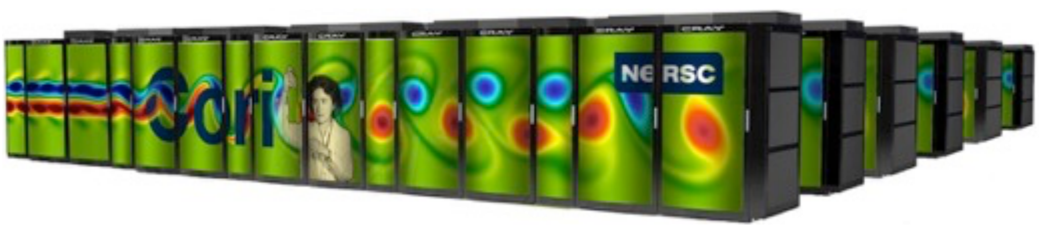
- Fast turn around time. Jobs start shortly after submitted
- Can run large number of throughput jobs

Software/Configuration

- Support for complex workflows
- Streaming data from external databases and data sources
- Easy to customize user environment

Hardware

- Local disk for fast I/O
- Some systems (not all) have larger memory nodes
- Support for advanced workflows

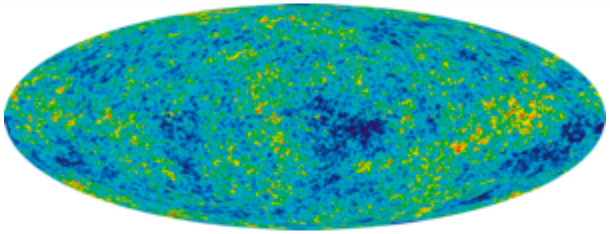


Popular features of a data intensive system can be supported on Cori

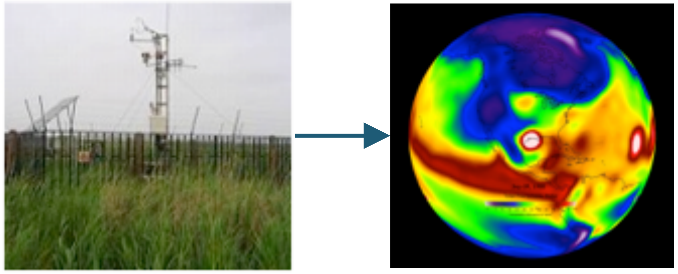
Cori – A Unified System for Big Data and HPC

- Data Partition with Traditional Xeon Processors and larger memory (128Gb)
- HPC Partition with Xeon Phi (KNL) Processors
- Common High-Bandwidth Interconnect
- Common Access to NVRAM Burst Buffer and High-Bandwidth Parallel File System
- Ability to support custom User Defined Images
- Advanced Gateways

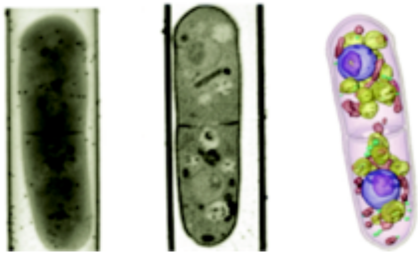
A unified system enables the coupling of experimental data with simulation and modeling



Enable tight coupling between modeling & simulation and experimental data in cosmology



Drive simulations with experimental data from environmental sensors



Compare theory to experiment with reconstructed data sets

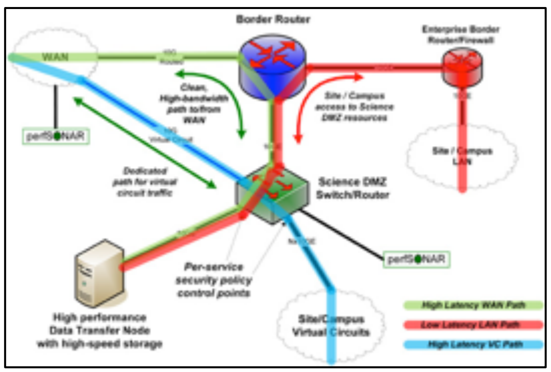
Challenge #2: A predictable and programmable network environment supporting science applications

Science DMZ, a network design pattern, improves the baseline end-to-end performance through ongoing global adoption

Science DMZ, facilitating great end-to-end network hygiene

- “Friction free” network path
- Dedicated, high-performance Data Transfer Nodes (DTNs)
- Performance measurement/test node

A **prerequisite** for any superfacility architecture

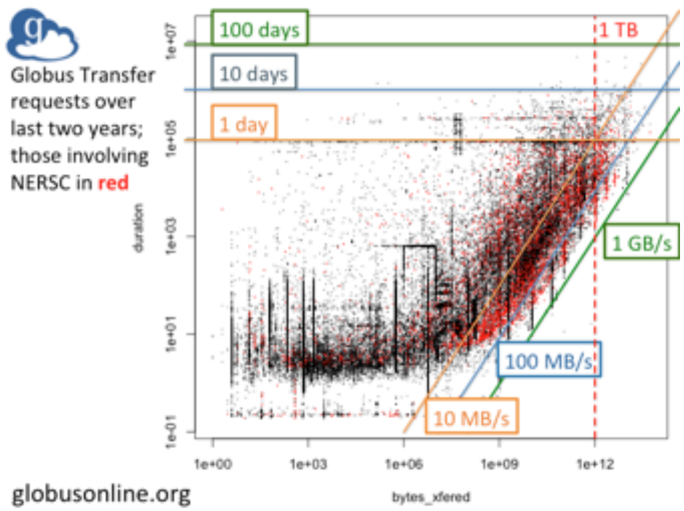


Science DMZ design pattern

$80M+ funding to implement Science DMZ design pattern in Universities

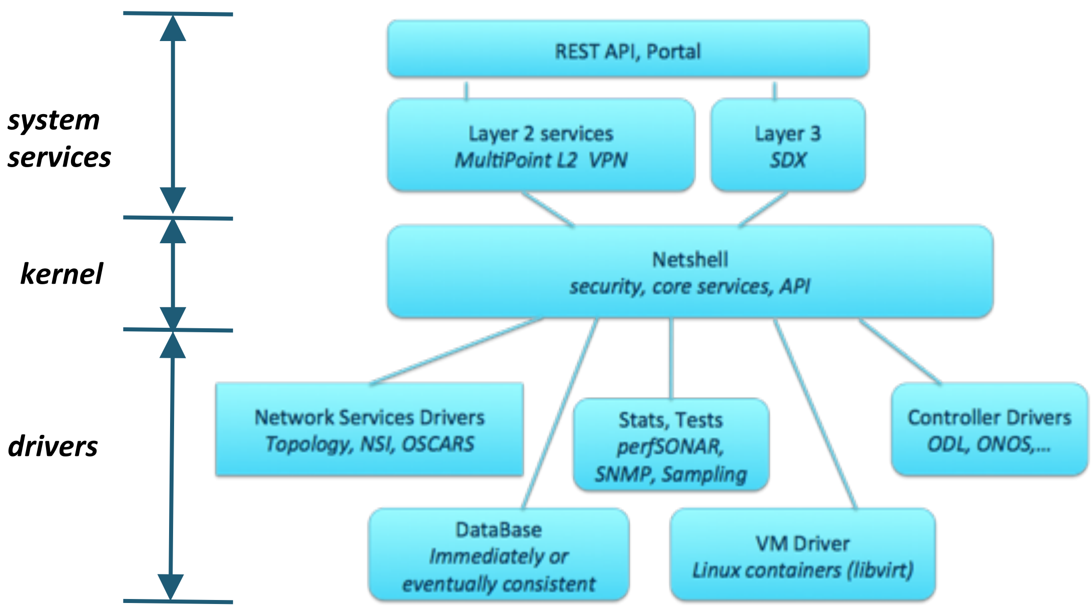
But still limited predictability of end-to-end network data transfers

Transfers over the network are not predictable

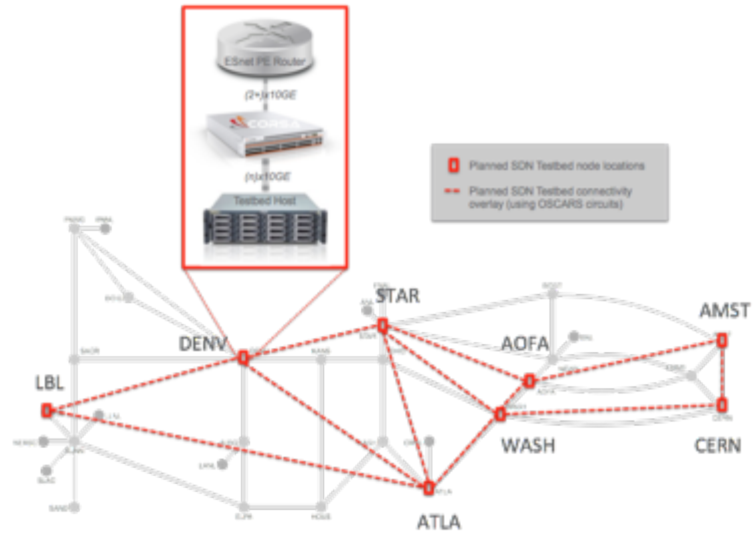


Superfacility workflows depend heavily on data movement infrastructure

ESnet Network Operating system (LDRD)



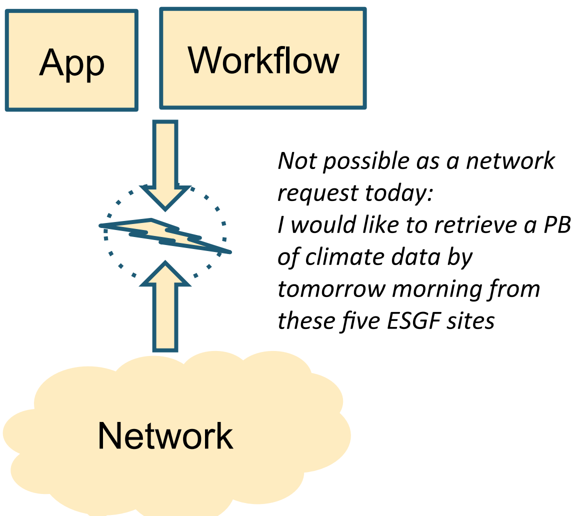
ESnet Network Operating System (ENOS)



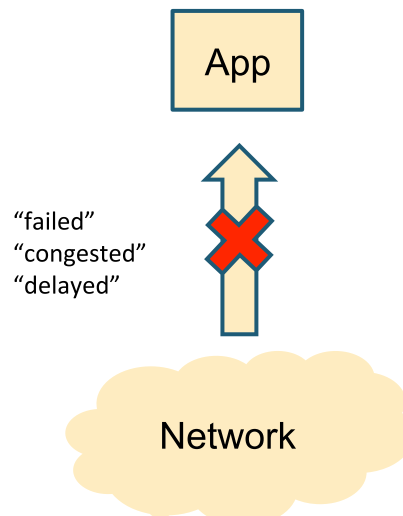
SDN Testbed

Constrained application-network interaction prevents benefits of automation, orchestration, optimization

High-level network abstractions for applications-network to have an 'intent' based conversation missing

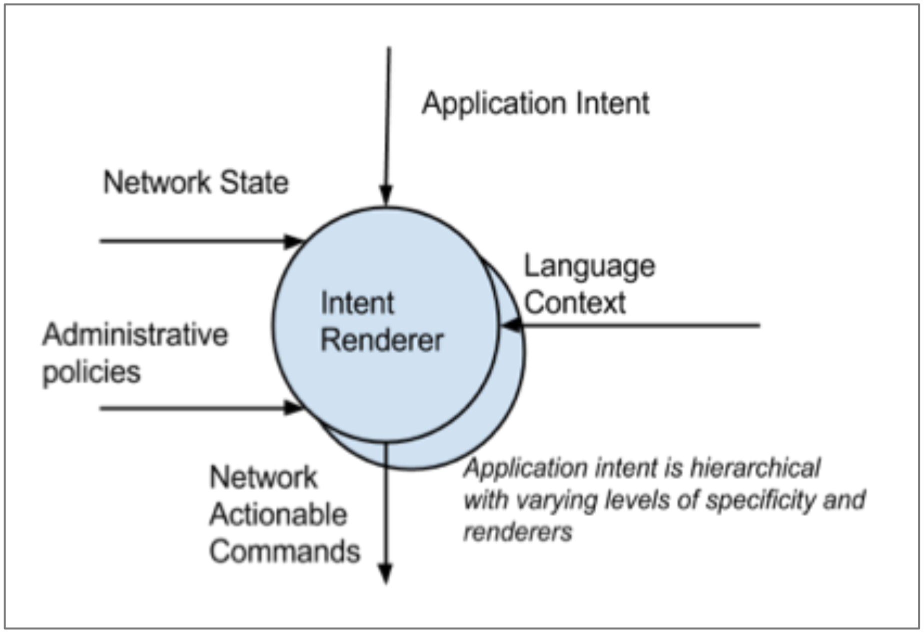


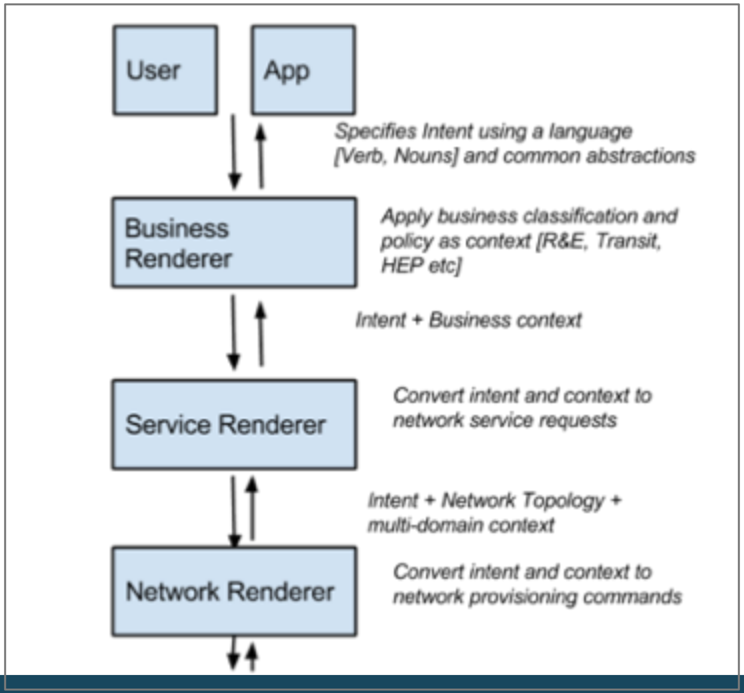
No abstractions for networks to feedback 'service state' to the application



Application Intent Interfaces

Enables applications to express ***what*** (descriptive) they require from the network without constraining ***how*** (prescriptive) the service should be delivered.



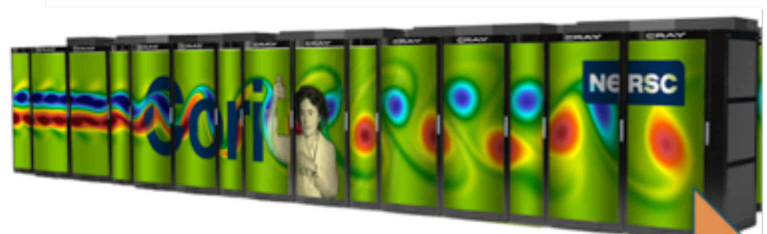


Challenge #3: Supporting workflows that allow seamless data movement from experiment, to analysis and data curation

Enable faster data pipeline from experiment to computation

Experiments (Data movers) → WAN (ESnet) → LBL (Campus) → NERSC (Facility) → Cori (Interconnect)

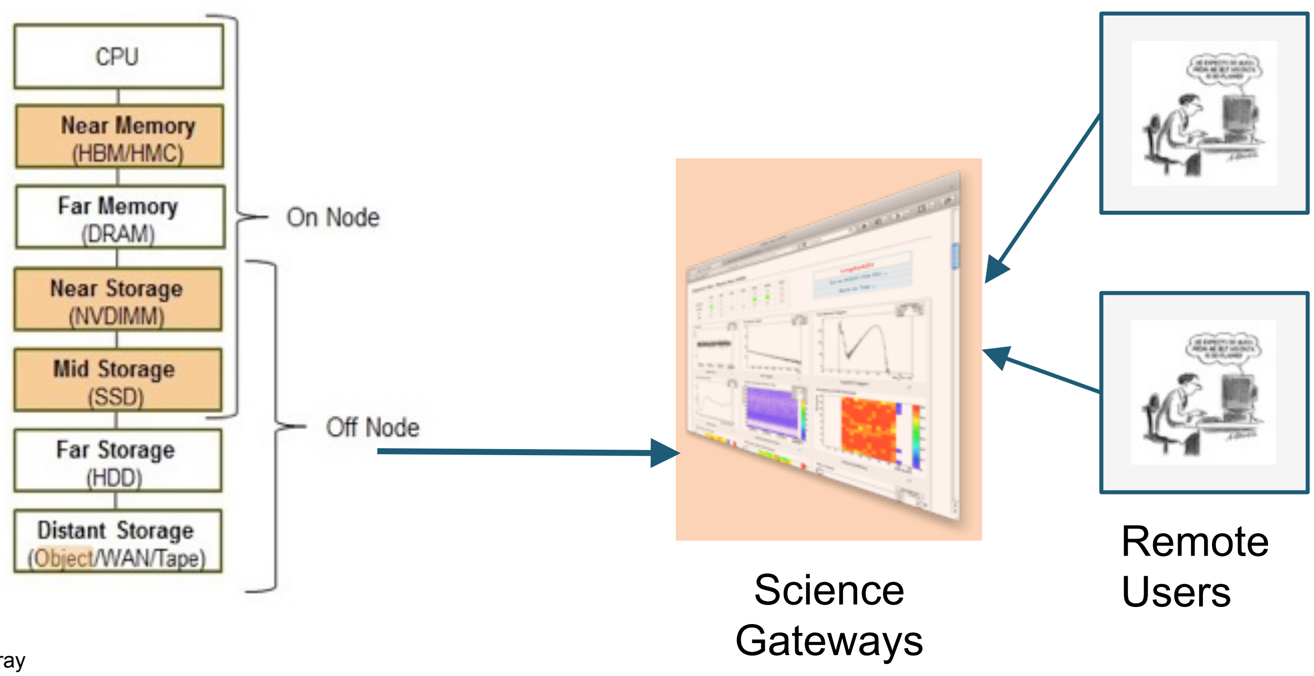
Program data streams / flows directly to the Burst Buffer and compute nodes



Enable 100Gb+ Instrument to Cori

New techniques for moving and managing data through the complex memory and storage hierarchy

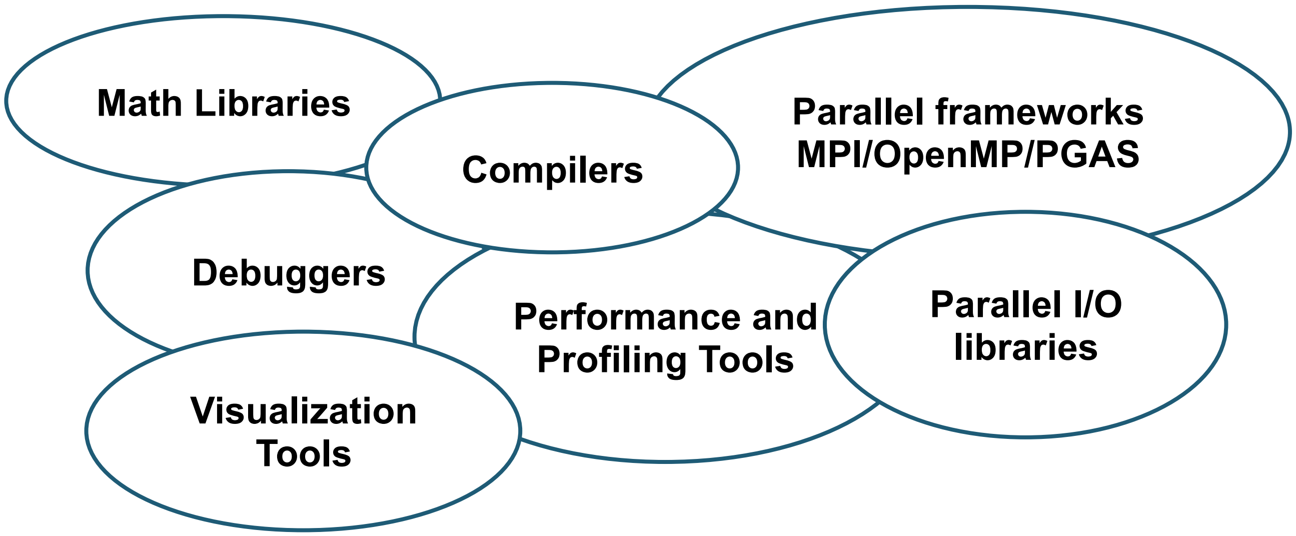
Future memory and storage hierarchy



Graphic from Cray

Challenge #4: Creating a productive software environment for data analysis

Today, the software ecosystem on HPC platforms is optimized for large scale simulation and modeling



Software for large scale data analysis on HPC platforms is limited and when available, often low performing.

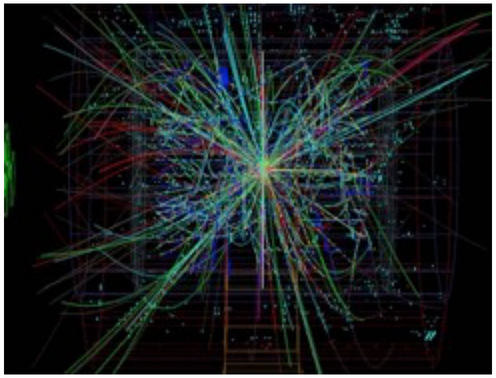
The ecosystem needed to support experimental science will need to be much richer

Capability	**Eco-system**
Data Processing	Workflow tools, processing frameworks
Data Analytics/Visualization	New scalable, high performance algorithms and methods
Data Access	Databases, science gateways
Data Transfer	Fast data transfer software between sites and transparent data movement within systems
Storage/Management	Portable data formats and I/O libraries

Our approach is to develop and support a scalable, high performance ecosystem for large scale data analysis

Shifter brings user defined images to supercomputers

- **Shifter, a container for HPC, allows users to bring a customized OS environment and software stack to an HPC system.**
- **Use cases**
  - High energy physics collaborations that require validated software stacks
  - Cosmology and bioinformatics applications with many 3rd party dependencies
  - Light source applications that with complicated software stacks that need to run at multiple sites



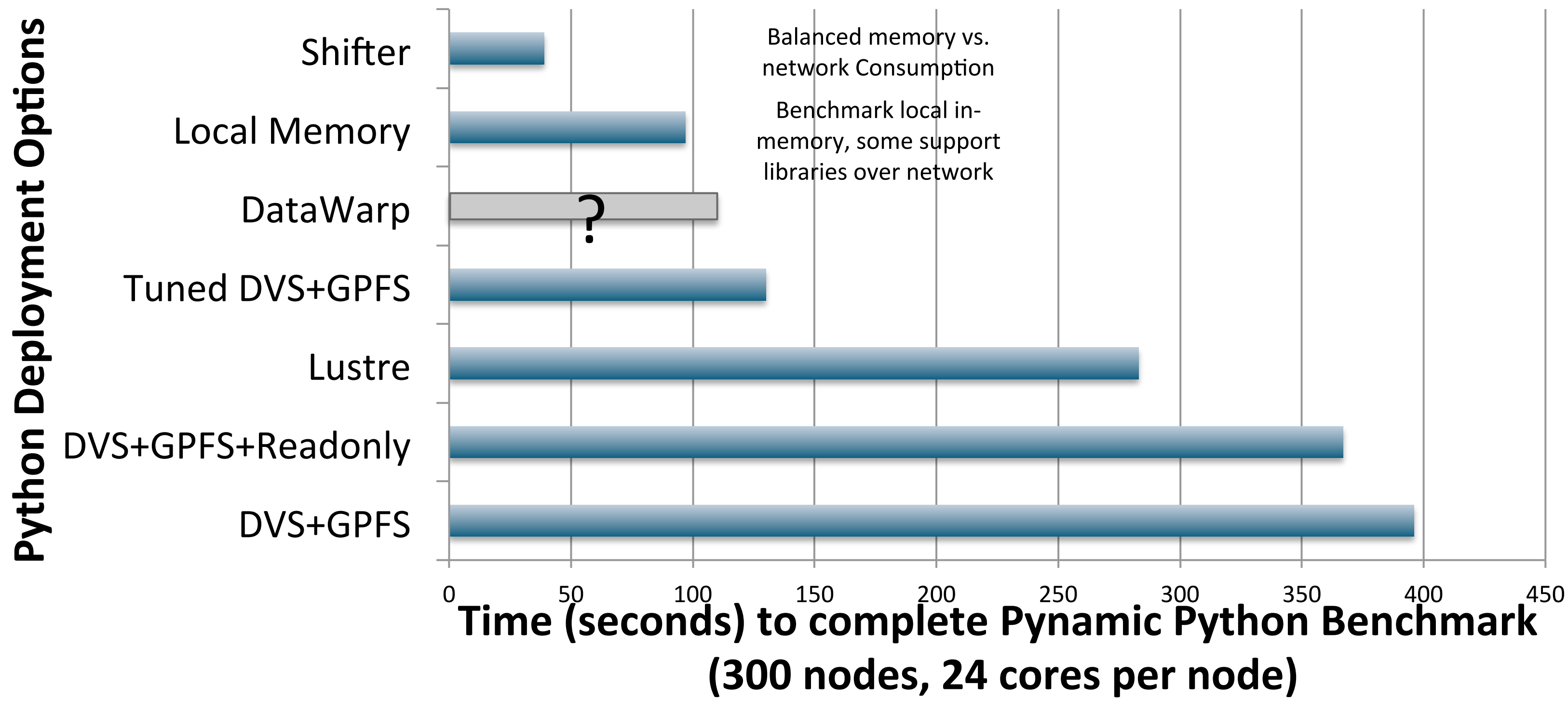
User Defined Images/Containers in HPC

- **Data Intensive computing often require complex software stacks**
- **Efficiently supporting "big software" in HPC environments offers many challenges**
- **shifter – First production containers in HPC**
  - NERSC R&D effort, in collaboration with Cray, to support User-defined, user-provided Application images
  - "Docker-like" functionality on the Cray
  - Efficient job-start & Native application performance

User-Defined Images

- **User-Defined Images (UDI): A software framework which enables users to accompany applications with portable, customized OS environments**
  – e.g., include ubuntu base system with Application built for ubuntu (or debian, centos, etc)
- **A UDI framework:**
  – Enables the HPC Platforms to run more applications
  – Increases flexibility for users
  – Facilitates reproducible results
  – Provide rich, portable environments without bloating the base system
  – **Presents HPC platform with a generic interface to the user application**

Shifter Delivers Performance – Pynamic



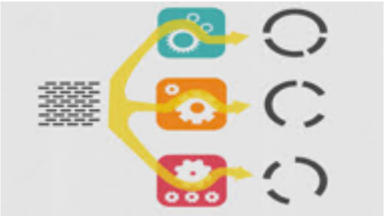
Where are we now?

- **An early version of Shifter is deployed on Edison. Early users are already reporting successes!**
- **Shifter is fully integrated with batch system, users can load a container on many nodes at job-start time. Full access to parallel FS and High Speed Interconnect (MPI)**
- **Shifter and NERSC were recently featured in HPC Wire. Many other sites have expressed interest**
- **Our early users:**

Light Sources
Structural Biology
(early production)

LHC – Nuclear Physics
(testing)

Cosmology
(testing)



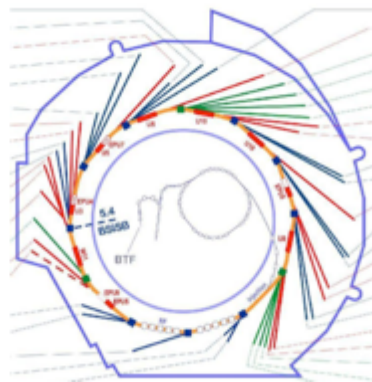
nucleotid.es
Genome Assembly
(proof of concept)

Conclusions

DOE user facilities will need to evolve dramatically

- New focus on time-sensitive applications, rather than system utilization
- Science Engagement i.e. meaningful interaction with scientists will become even more important
- Definition of 'users' will need to change for example to include another user facility
- Coordinated user support across facilities will need to be scaled and responsibilities articulated
- Orchestration across facilities will need to be coordinated at multiple levels, from resources to outages.

Superfacilities will transform Experimental and Observational Science



Cosmology

Merging large scale simulation and data analysis

Light Sources

Time sensitive processing and seamless data movement

Environment

Integrating remote sensing data, across multiple scales, into theory and models

Questions