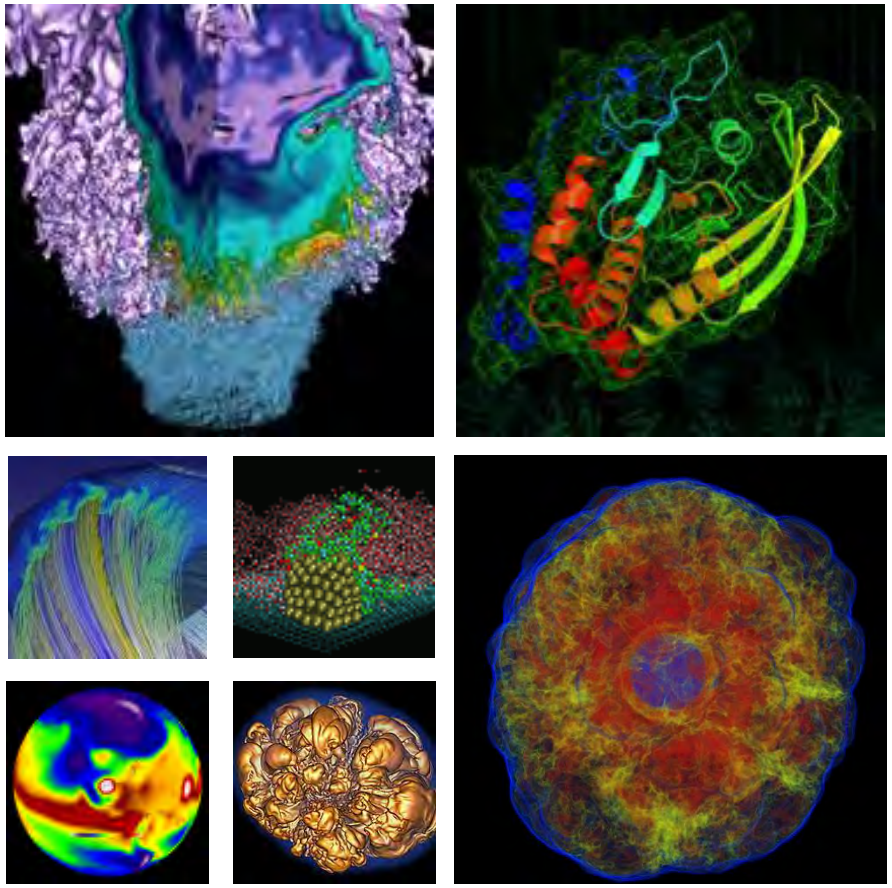
Superfacility:
How new workflows in the DOE Office of Science are influencing storage system requirements

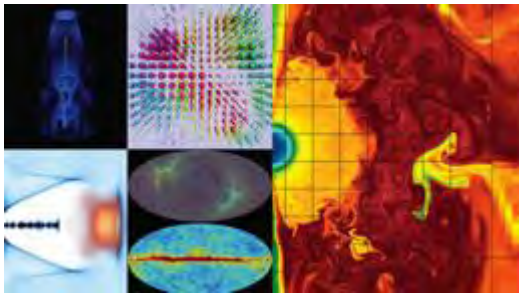


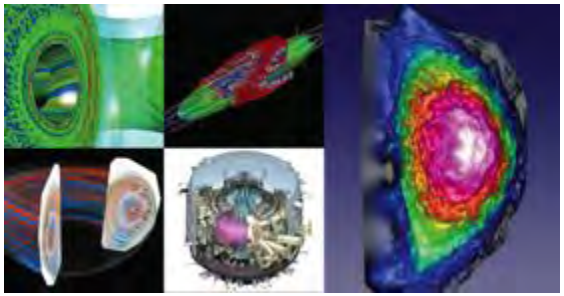
Katie Antypas
Department Head Scientific Computing and Data Services

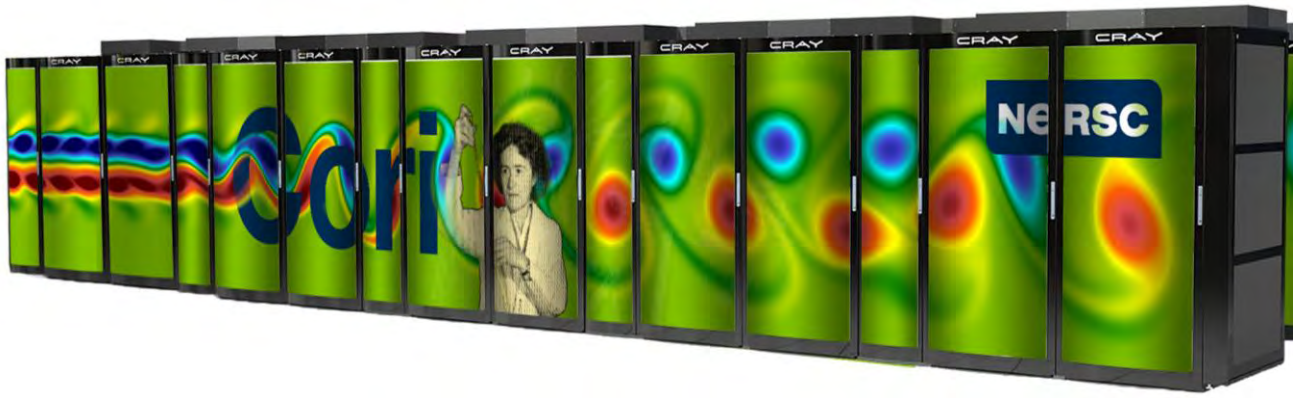
May 3, 2016

NERSC is the mission HPC computing center for the DOE Office of Science

- **NERSC deploys advanced HPC and data systems for the broad Office of Science community**
- **NERSC staff provide advanced application and system performance expertise to users**
- **Approximately 6000 users and 750 projects**

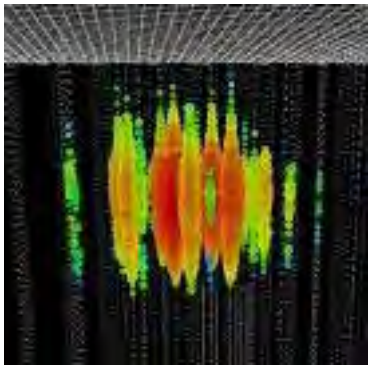






BERKELEY LAB

NERSC has been supporting data intensive science for a long time



Ice Cube
Neutrinos

Planck Satellite
Cosmic Microwave Background
Radiation

Alice
Large Hadron
Collider

Dayabay
Neutrinos

Joint Genome
Institute
Bioinformatics

Atlas
Large Hadron
Collider

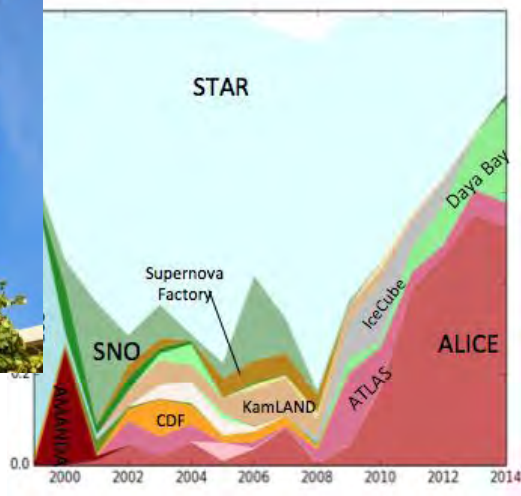
Historically NERSC has deployed separate Compute Intensive and Data Intensive Systems

Compute Intensive

Data Intensive

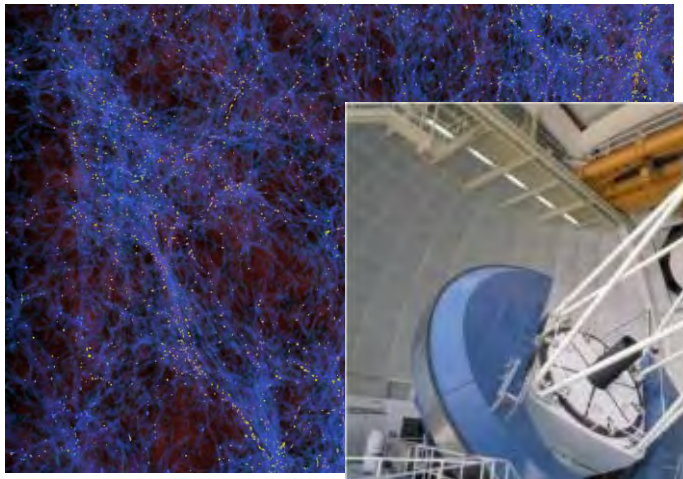
Carver

Genepool

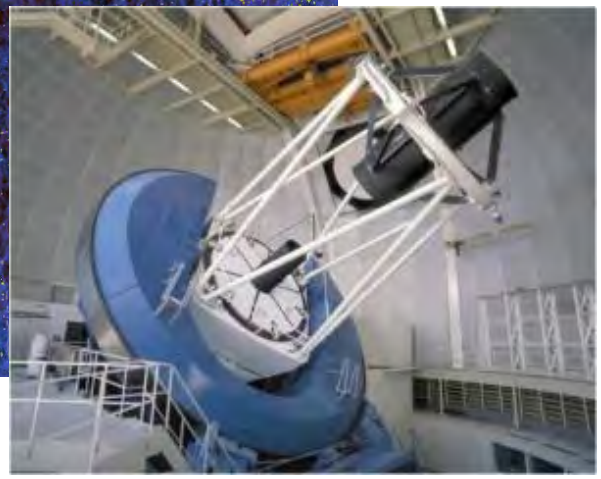


PDSF

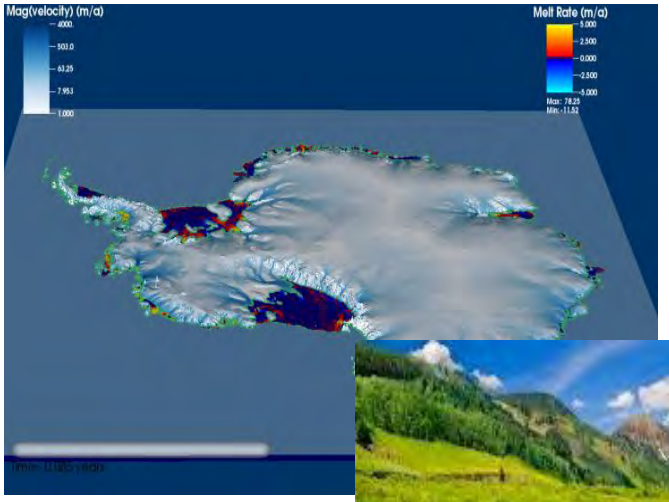
What has changed? Coupling of experiments with large scale simulations



Nyx simulation of Lyman alpha forest



Kitt Peak National Observatory's Mayall 4-meter telescope, planned site of the DESI experiment



New climate modeling methods, produce new understanding of ice

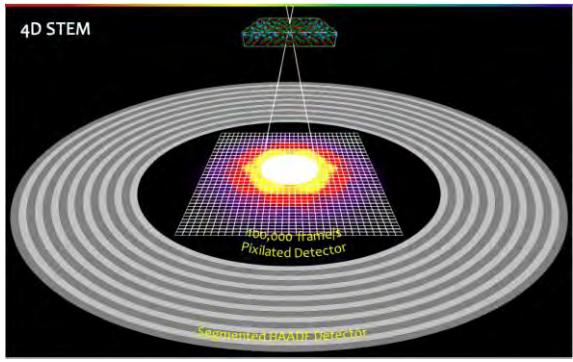
Genomes to watersheds

data rates and new sensing capabilities

LCLS
Light Source

Advanced Lightsource Upgrade

Environmental sensors

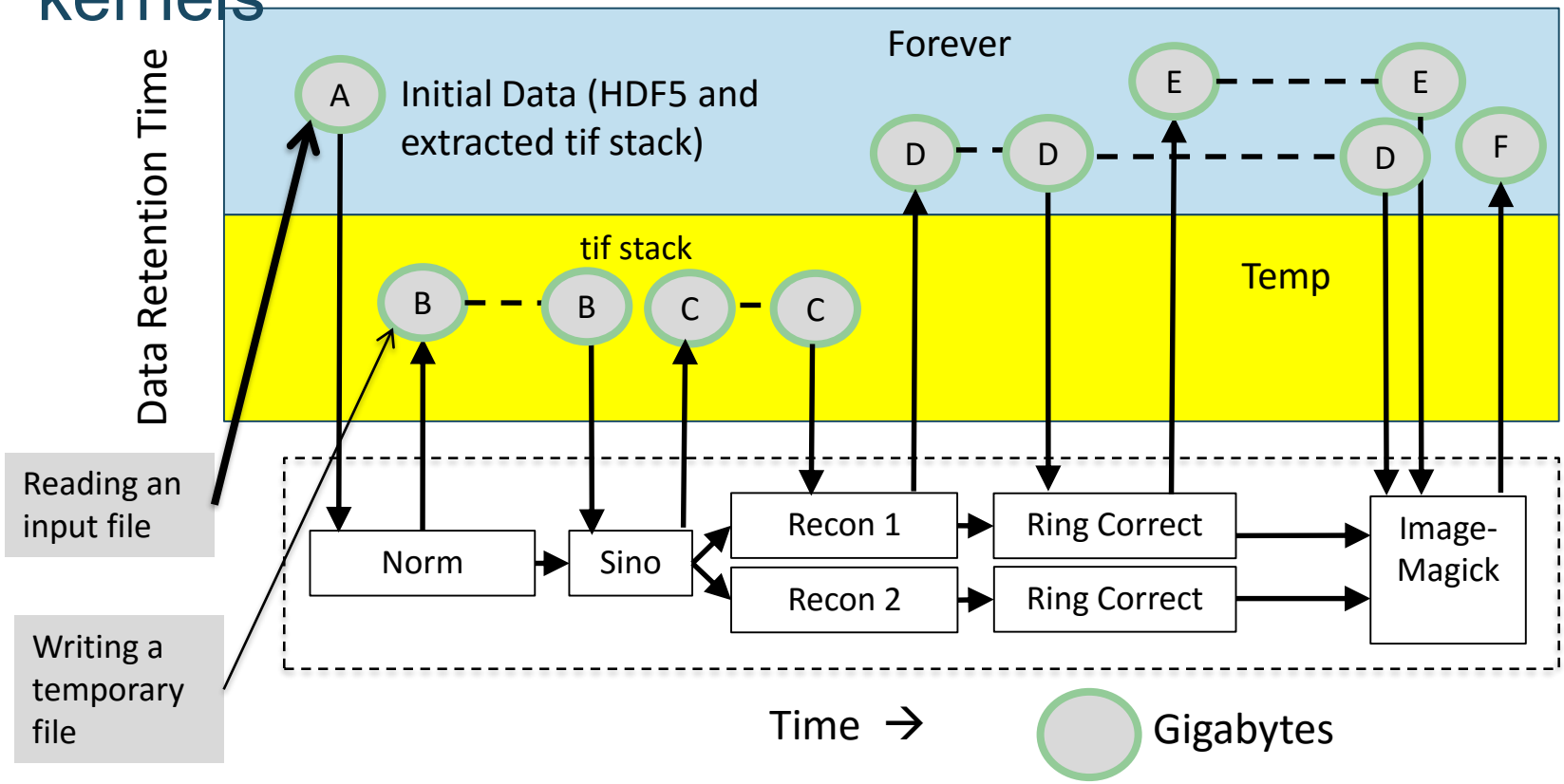


Next generation electron microscope

Sequencers that fit into the palm of your hand

- In the next 5 years, data rates will be approaching Tb/sec for many instruments
- Infeasible to put a supercomputer at the site of every data generator

Optimizing workflows becomes as important as optimizing computational kernels



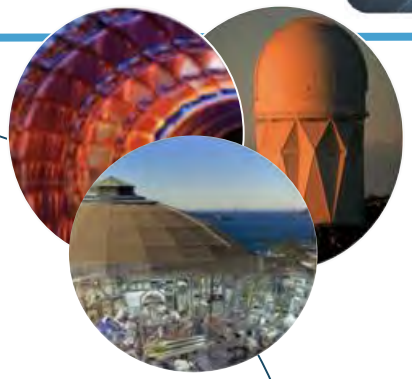
Web Thumbnails and tifs packed as HDF5 for Visit Vis.

- **This workflow consists of many dependent tasks which read and write files**
  - Files are either discarded (in yellow layer - bottom) or saved forever (in blue layer - top)
- **Helps us understand how the scientist wants to use storage**

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

NERSC

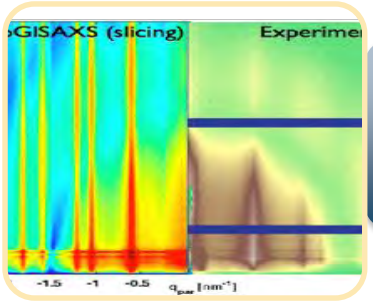
Experimental Facilities



Real-time analysis and Data management

New mathematical models

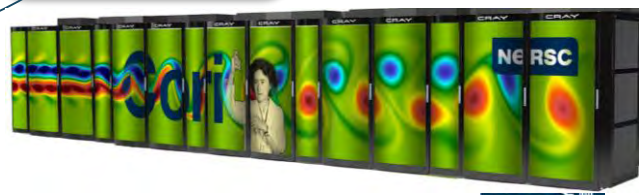
ESnet
ENERGY SCIENCES NETWORK



Fast Implementations on latest computers

Network for Big Data Science

Unified Computing Facilities



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Some thoughts on how storage requirements will be influenced by experimental data

- **Seamless data movement and management from experiment through memory/storage hierarchy will require more coordinated software stacks, data models and metadata**
- **The same data will need to be accessed by different users and groups during a workflow**
- **Components of workflows outside a compute system, (web gateways and databases), will need equal access to data and storage**

Some thoughts on how storage requirements will be influenced by experimental data

- **Scheduling will need to expand to more than just compute -- to include storage, bandwidth and experiment allowing guaranteed QoS**
- **Analyzing streaming data will require high bandwidth networking to storage and compute nodes**
- **Authentication and identity management across facilities and storage systems will need to be robust and coordinated**