

# Trends in Storage and File Systems

Shane Canon

Data System Group Leader

Lawrence Berkeley National Laboratory

ICAP09

September 2, 2009





NATIONAL ENERGY RESEARCH  
SCIENTIFIC COMPUTING CENTER

# Outline

- Trends in Storage
- Trends in (HPC) File Systems
- Trends in IO Middleware
- Impact on Applications



# Enterprise Storage Trends

According to **COMPUTERWORLD**

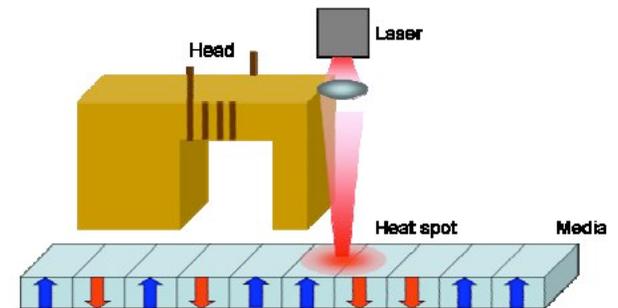
- Virtualization
- **10G versus Fibre Channel**
- Disk versus Tape
- Deduplication
- **Flash-based Storage**

And

- Magnetic Disk capacity continue to grow

# Magnetic Storage Trends

- Heated Assisted Magnetic Recording (HAMR) and Bit-patterned media could help push magnetic media to 50 Tb/in<sup>2</sup>.
- **Bad News:** Capacity scales with aerial density, BW scales with  $\sqrt{\cdot}$ .
- Also, SAS replacing FC for Enterprise market



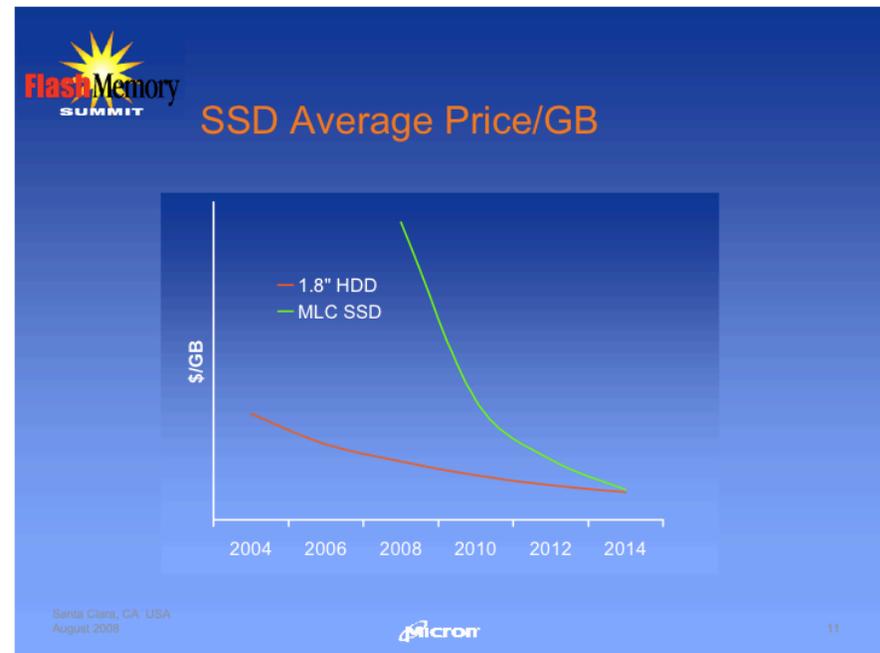
# Converged Fabrics

- Converged Enhanced Ethernet enables SAN and IP over a single fabric
  - Adds flow control and congestion management to Ethernet
  - Makes Ethernet a “loss-less” fabric
- InfiniBand also supports these capabilities
  - InfiniBand SRP for storage
  - Native RDMA support
- Even if a system doesn't use a converged fabric this trends impacts the evolution of SANS and Networks



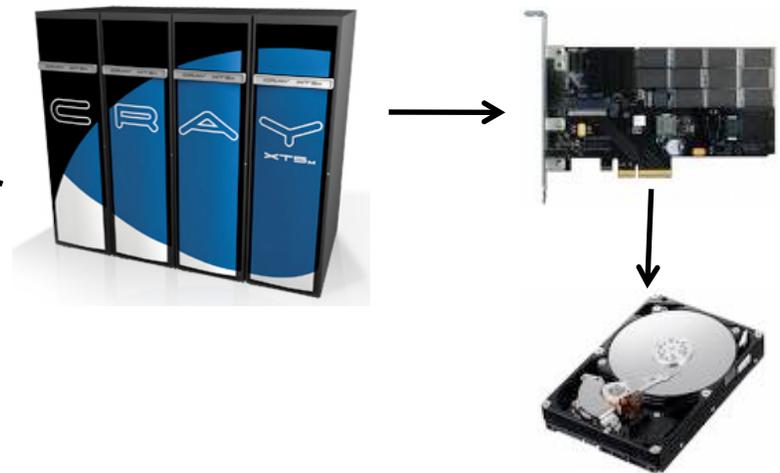
# Flash Technology Trends

- Solid state storage predicted to match disk storage in \$/GB by 2014 timeframe (but possibly much later)
- However, impact could be felt sooner. (Will R&D investment levels in magnetic media continue if SSDs take over consumer market?)
- \$/IOPS is competitive, especially for ~1 TB- solutions
- Phase Change Memory could replace NAND (faster and more reliable)
- Probe memory could enter in commodity marketplace (higher bit density, higher latency)



# Flash Storage in an Exascale Architecture

- Flash integrated on systems – Accelerate Checkpoints, Extend main memory, replace local disk
- Flash to accelerate Metadata
- Flash in Storage Arrays at the interconnect edge – First level cache to deal with I/O burst and reorder data. Lower power than 100,000 (or more) spindles



# Challenges for HPC I/O beyond Petascale

- Data Integrity
  - Bit Error rates haven't improved for HDD (1 in  $10^{14}$  or  $10^{15}$  bits)
  - Data integrity checks on read are becoming a must and will only increase in importance
  - Exascale will require end-to-end checks
- Capacity continues to outpace bandwidth for Magnetic storage
  - Flash Storage may help
  - Better Middleware can help to efficiently utilize the bandwidth



NATIONAL ENERGY RESEARCH  
SCIENTIFIC COMPUTING CENTER

# HPC File Systems Trends

- Improvements will be required in the file systems to fully leverage new technology
  - Hierarchy to leverage solid state storage
  - Metadata code enhancements to take advantage of faster devices
- Improved Data Protection
  - RAID at the file level
  - End-to-End protections



NATIONAL ENERGY RESEARCH  
SCIENTIFIC COMPUTING CENTER

# Middleware

- Improvements in parallel I/O middleware to address
  - Scaling (limited writers, file per process)
  - Idiosyncrasies of the file system (block alignment, block size)
- Goal: Ease of use
  - Separate logical data layout from physical layout
  - database users don't worry about how a record is stored, why should we?)



NATIONAL ENERGY RESEARCH  
SCIENTIFIC COMPUTING CENTER

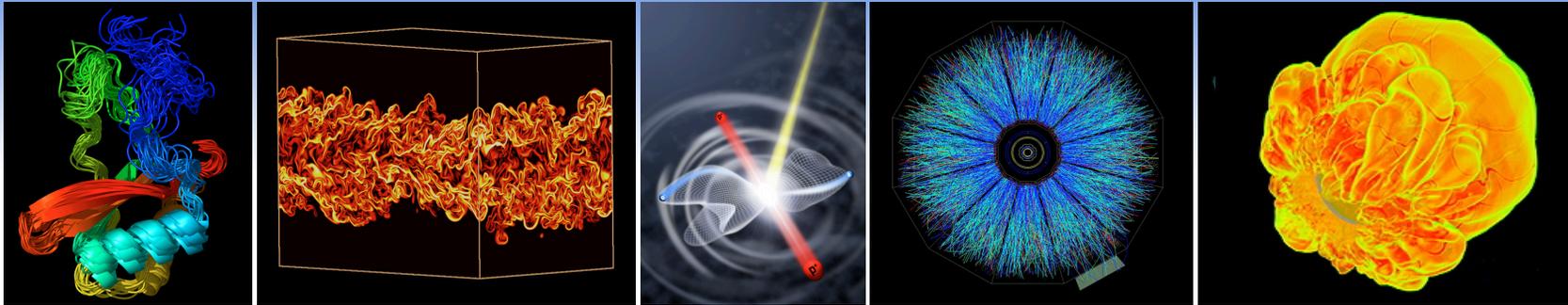
# MapReduce and Hadoop

- Entirely new models like MapReduce could dramatically change how data intensive work is carried out.
  - Designed to scale
  - Move the work to the data
  - Fault Tolerance built into the application framework



# What's this mean for applications

- Local Solid State Storage could dramatically alter Checkpoint methods
- SSS will make general I/O perform closer to best case (but introduce its own set of issues).
- File systems will (hopefully) preserve data integrity (but likely still get blamed when it fails)
- Middleware will hide the complexity and simplify getting good performance (except for the cases where it doesn't)

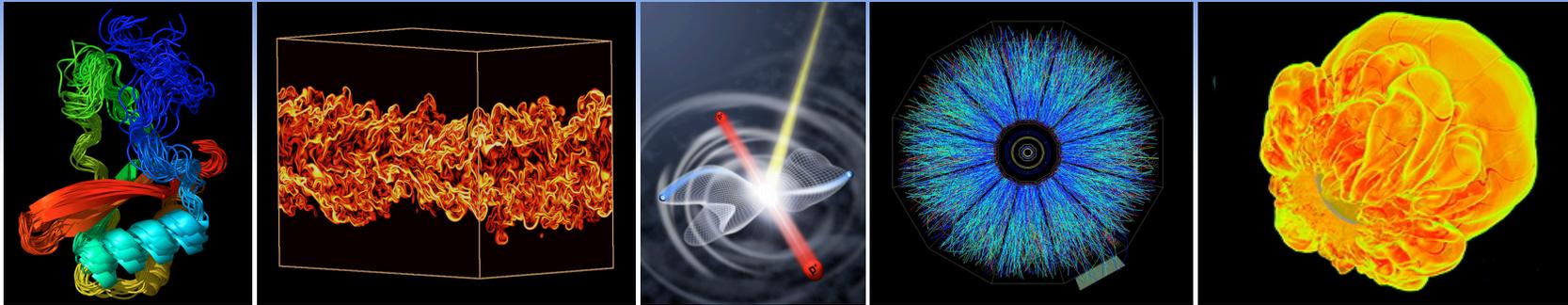


# Thank You

## Contact Info:

Shane Canon

Canon at nersc dot gov



# Bonus - Flash



NATIONAL ENERGY RESEARCH  
SCIENTIFIC COMPUTING CENTER

# Not without Challenges

## Reliability Overhead

- Where remapping logic is completed (hardware, software)
- Wear issues (10k cycles for MLC, 100k cycles for SLC)

## Erasure Overhead

- Writes require erase (slow). Erases must be done in blocks. This leads to trouble with non-sequential I/O as the card fills up. Different cards use different grooming techniques (some CPU intensive - 60% for one card)

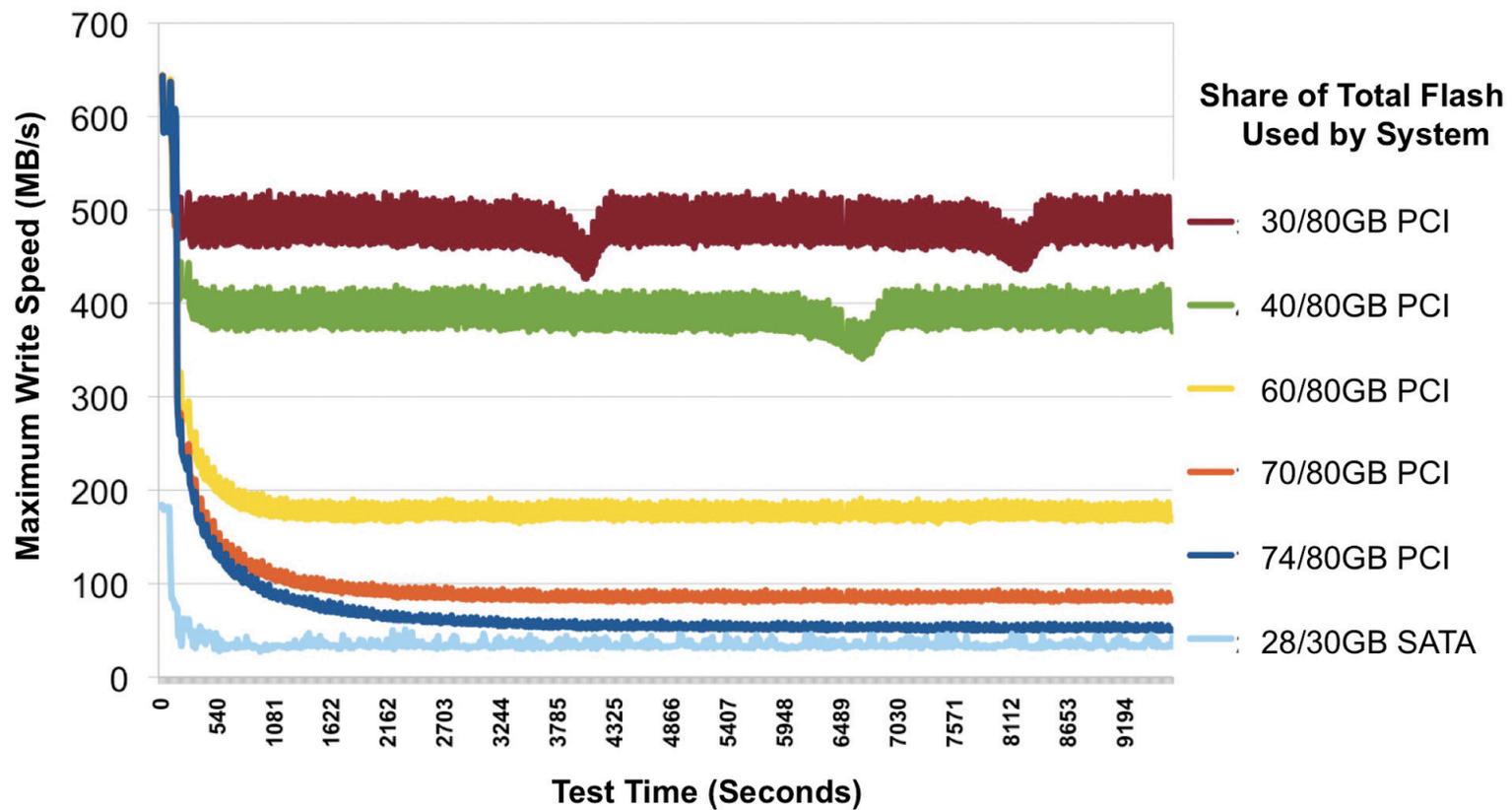
## NAND SSDs don't act like disks or RAM

- Decades of software development to deal with idiosyncrasies of disk. Similar investments required for solid state storage
- Chip failure requires card replacement (implies mirroring across cards)

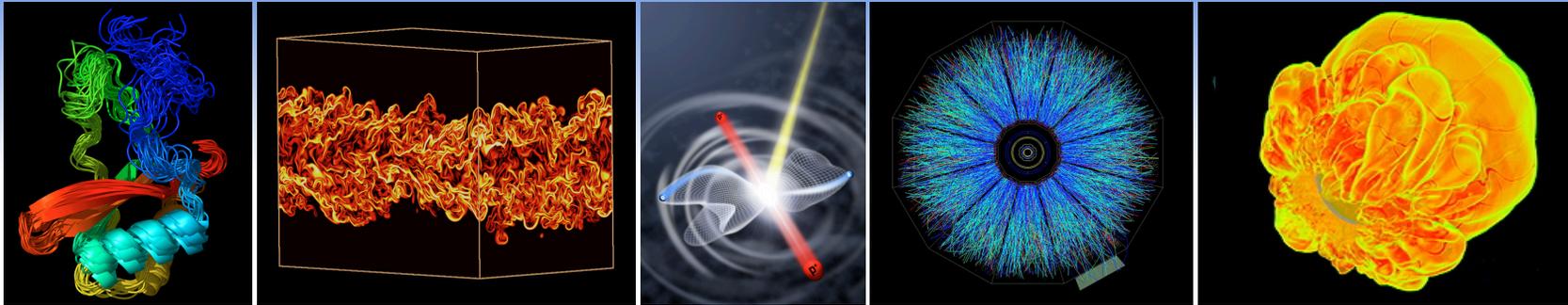
## Platform Support

- Linux and Windows

# Flash Behavior



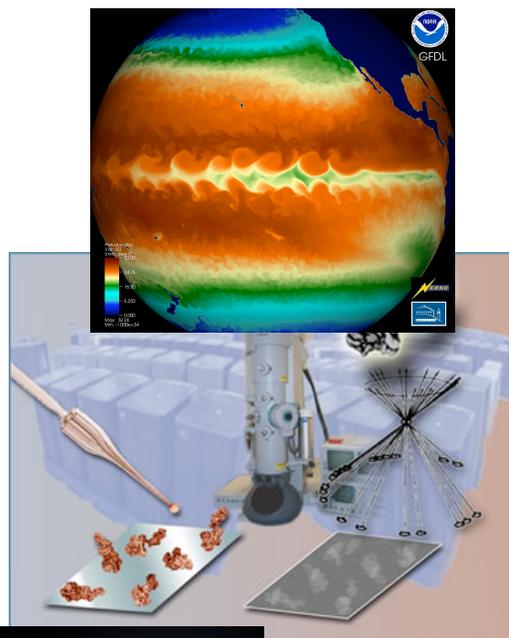
*Courtesy of SNIA SSS Initiative Whitepaper*



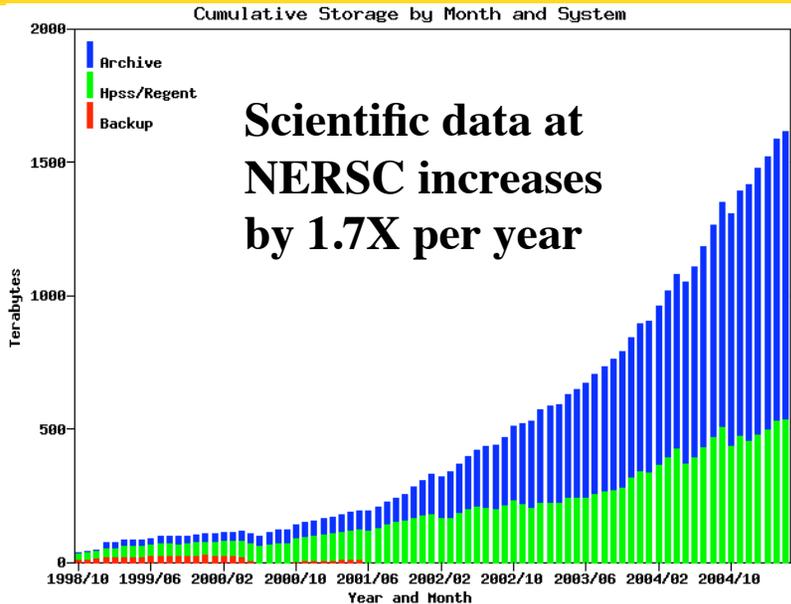
# NERSC Data

# Data Needs Continue to Grow

- Scientific data sets are growing exponentially
  - Simulation systems and some experimental and observational devices grow in capability with Moore's Law
- Petabyte (PB) data sets will soon be common:
  - *Climate modeling*: estimates of the next IPCC data is in 10s of petabytes
  - *Genome*: JGI alone will have .5 petabyte of data this year and double each year
  - *Particle physics*: LHC are projected to produce 16 petabytes of data per year
  - *Astrophysics*: JDEM alone will produce .7 petabytes/year
- We will soon have more data than we can effectively store and analyze
- Question: What are your data set sizes (active disk vs. archive), bandwidths?

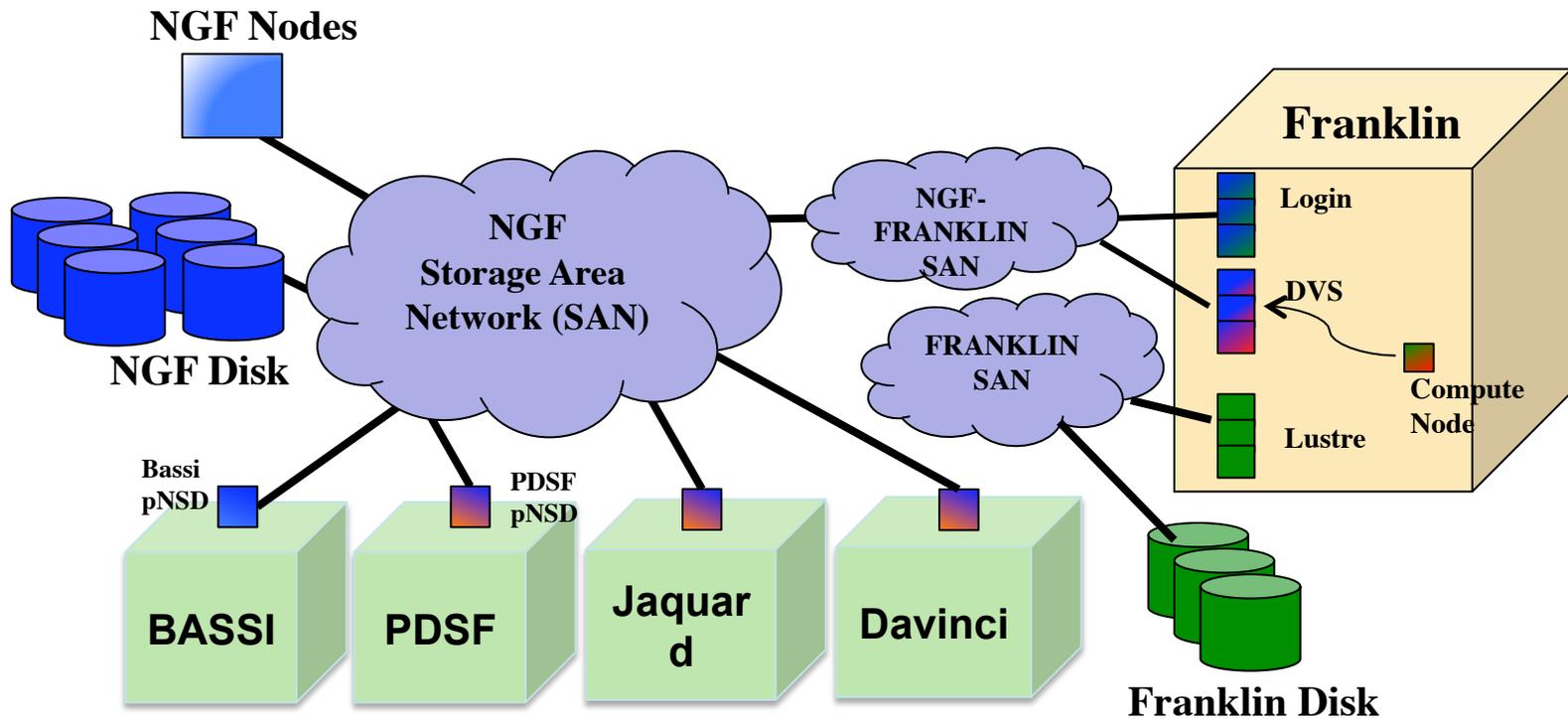


# Tape Archives: Green Storage



- Tape archives are important to efficient science
  - 2-3 orders of magnitude less power than disk
  - Requires specialized staff and major capital investment
  - NERSC participates in development (HPSS consortium)
- Questions: What are your data sets sizes and growth rates?

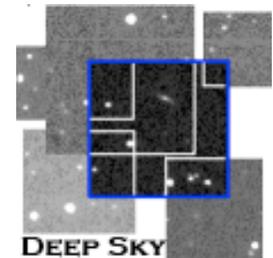
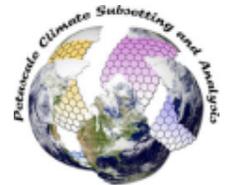
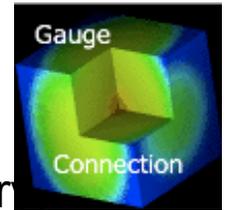
# NERSC Global File system (NGF)



- **A facility-wide, high performance, parallel file system**
  - Uses IBM’s GPFS technology for scalable high performance
  - Makes users more productive
  - **Questions: How large is your “working set”? Is it shared community-wide?**

# Science Gateways

- Create scientific communities around data sets
  - Models for sharing vs. privacy differ across communities
  - Accessible by broad community for exploration, scientific discovery, and validation of results
  - Value of data also varies: observations may be irreplaceable
- A *science gateway* is a set of hardware and software that provides data/services remotely
  - Deep Sky – “Google-Maps” of astronomical image data
    - Discovered 36 supernovae in 6 nights during the PTF Survey
    - 15 collaborators worldwide worked for 24 hours non-stop
  - GCRM – Interactive subselection of climate data
  - Gauge Connection – Access QCD Lattice data sets
  - Planck Portal – Access to Planck Data
- Building blocks for science on the web
  - Remote data analysis, databases, job submission



# Data Acquisition and Visualization Pipeline

