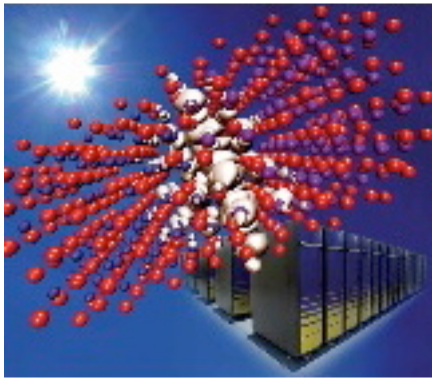
OPENFABRICS ALLIANCE

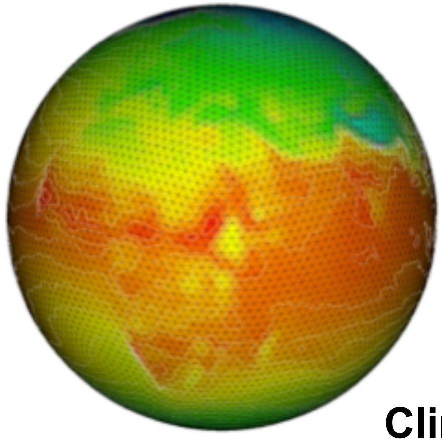
I/O Requirements for Exascale

Author: Jason Hick, NERSC Storage Systems Group Lead, LBNL
Date: 4 April 2011

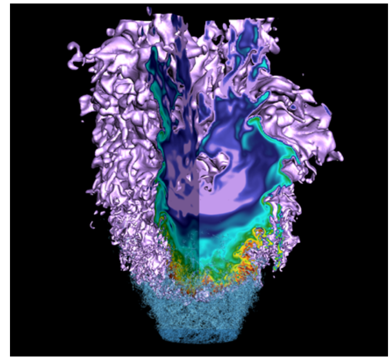
Science is Driving Exascale: Carbon Cycle Research



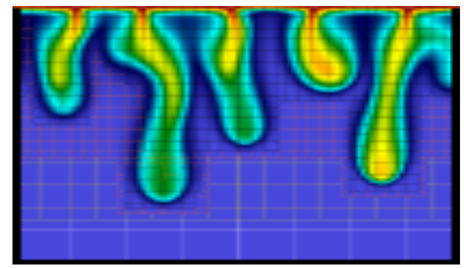
Solar: Materials for solar panels and other applications.



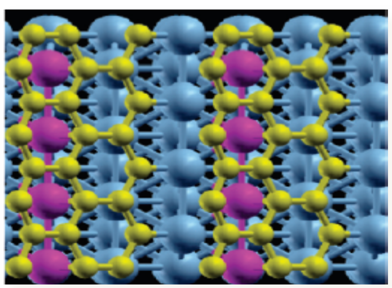
Climate modeling: High resolution, clouds, ice sheet, abrupt change, historical validation.



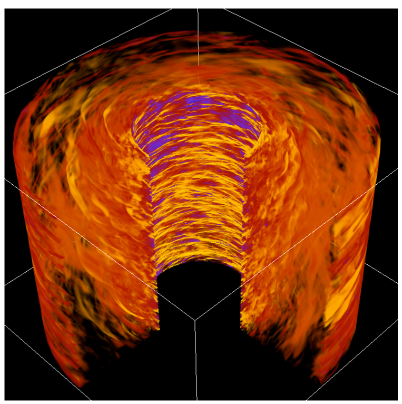
Combustion: New algorithms (AMR) coupled to experiments



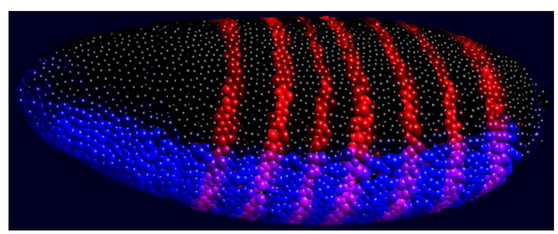
Carbon Capture & Sequestration: Chemistry, dissolution-diffusion-convection processes in aquifers.



Storage, production: Catalysis for fuel cells and batteries



Fusion: Simulations of ITER scale devices

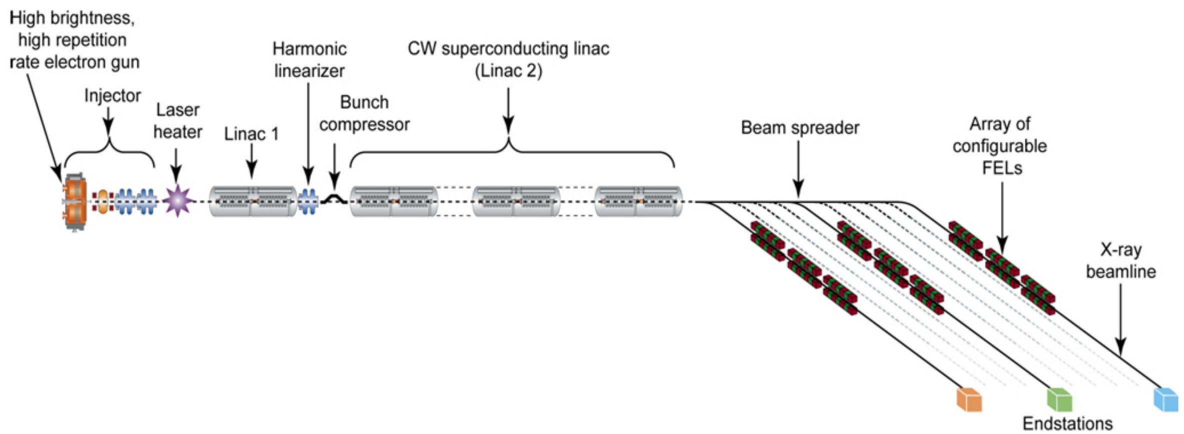


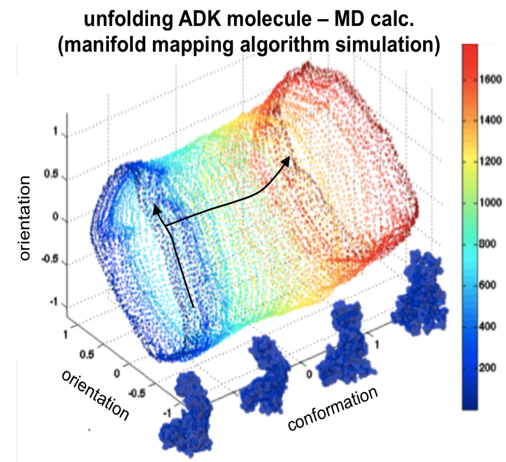
Biology: Data analysis for gene genomics.

Science is Driving Exascale: Nuclear Uncertainty Quantification

- Want to go from an ability to describe natural phenomena with simulations towards a *predictive capability*
  - But nature is messy: need to understand sensitivity to preturbation
  - Numerical simulation answers whether a design is sufficient, but does not quantify the uncertainty of the answer.
  - This is NOT V&*V* *(can only do UQ if you trust your simulation)*
  - Example Application: *rapid qualification of new nuclear power plant design, or many engineering problems*
- Example Approach: *Polynomial Chaos*
  - Run many simulations with input preturbations *(task sched/mgmt)*
  - Statistical summarization across simulation datasets to understand sensitivity to design parameters *(huge data management issues)*
- Requires workflow tools integrated with transport infrastructure
  - Need task farming to prevent batch system from being overwhelmed (need task management & data management)
  - Need coordination with network infrastructure, I/O, and compute
  - *No pretty graphical tools (get over that now!)*

Science is Driving Exascale: Next Generation Light Source





- Computational requirements JUST for orientation reconstruction
  - *Input Data Rate:* 10^5 images/second at 10^6 pixels imaging rate (4TB/sec)
  - 10^5 of images of diffraction patterns representing 2D projection of the sample in random orientation
  - Best available orientation algorithms require ~N^6 flops (N=1000 for NGLS detector)
  - *Total performance required is 10^{18} FLOP/s for pulse rate of 10^5 images/second*
- Similar requirements for shot planning

Both data processing and shot planning will require exascale computing for analysis and terabit networking for data movement

Current Exascale Approaches

- Collaboration and competition
  - DOE NNSA and DOE OS labs collaborations
    - ACES – OLCF/LANL/Sandia
    - ABEL – ALCF/LBNL/LLNL
  - Each aiming for a pre-exascale system (300TF) in 2015 timeframe and exascale system in 2018-2022
- Co-Design
  - Software + Hardware + Applications design collaborations ongoing
- Revolutionary vs. Evolutionary
  - Both approaches are needed due to 100-1000X improvement required in every facet of the system to deliver something useable to science
  - Moving from Petascale to Exascale likely to be as disruptive to users as moving from Vector to Distributed systems

Exascale I/O Approaches

- Collaboration and competition
  - Learn from what I/O systems are working and what aren't at each DOE lab
- Co-Design
  - Data management middleware working with file system/archive developers
- Revolutionary vs. Evolutionary
  - Hardware improvements
    - Need disk spindle reliability improvements
    - Need disk performance improvements
    - Need tape capacity improvements
  - Power efficiency solutions
  - Data management and analysis solutions

IO Requirements Today

- In general, performance needed is achievable
  - Work with users/applications to achieve given hardware/software configuration
- Designs focus on ratios aimed at balancing storage resource capabilities
  - Correlation to amount of memory and network rate
- Time spent ensuring continual data movement up and down the storage hierarchy

Memory and IO

The amount of system memory plays a role in the speed and size of the storage systems at HPC centers

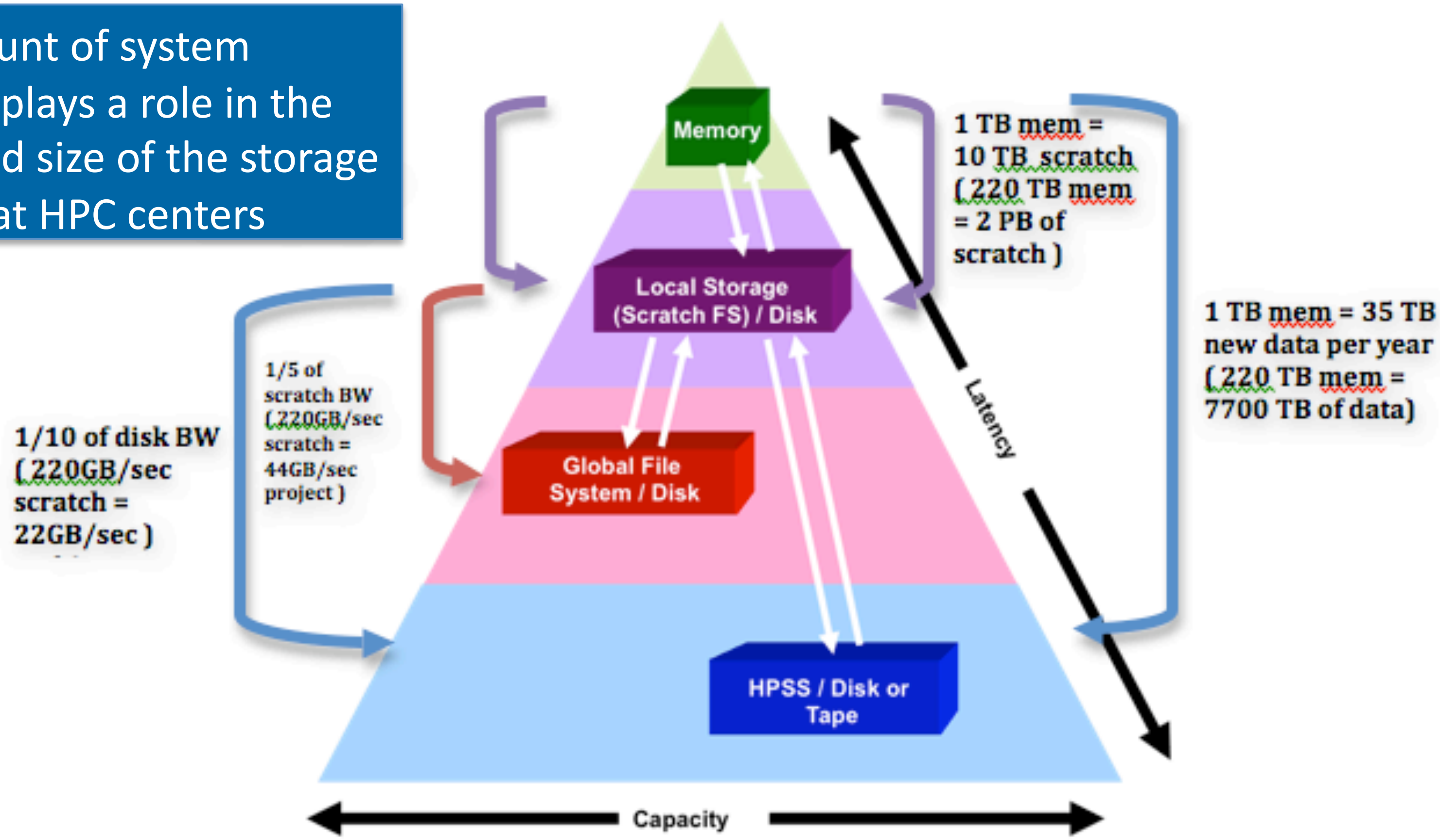
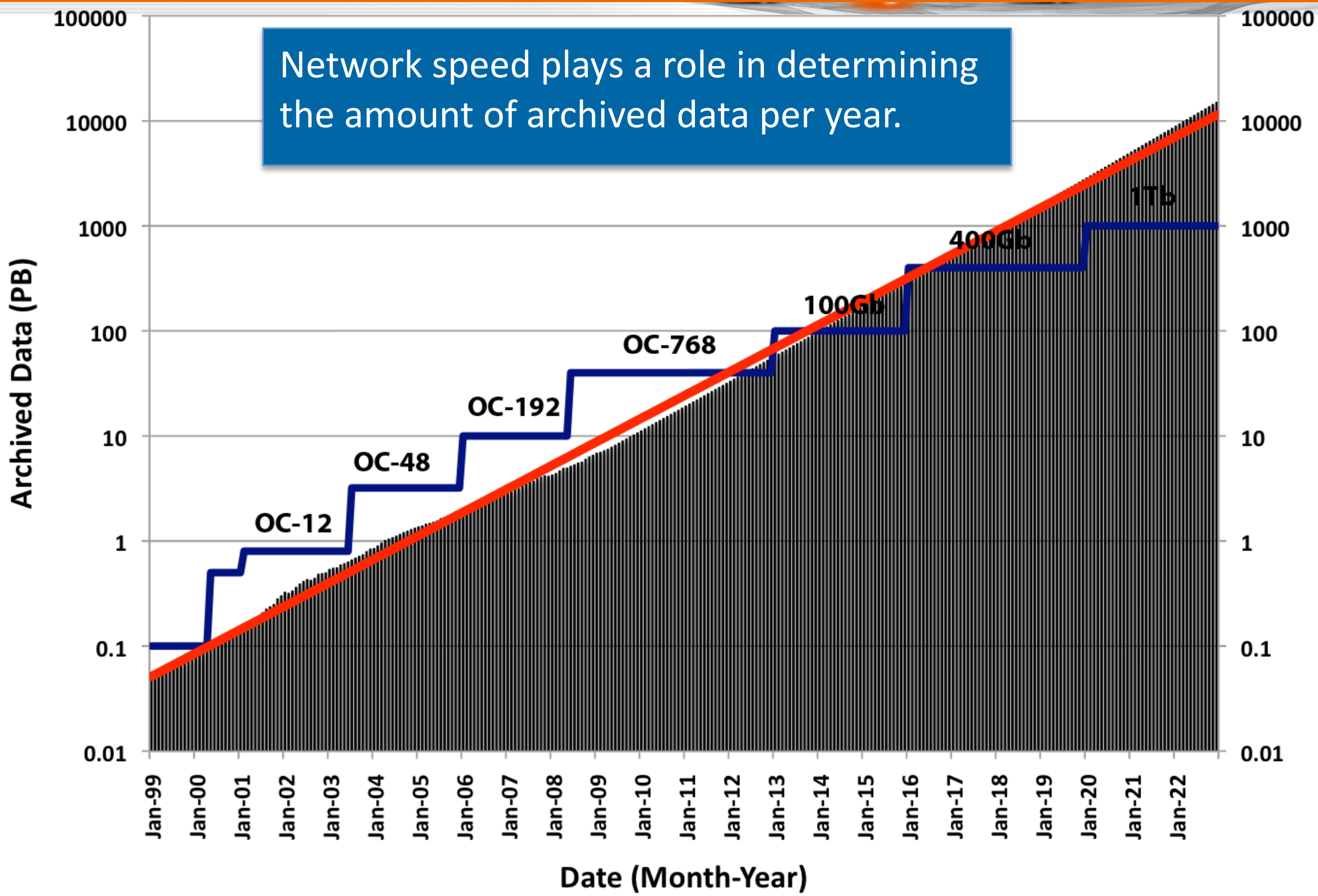


Figure 1. Conventional HPC Storage Planning Guidelines

Network (Ethernet) Rate and Data Stored



The Major System Components of Exascale

- Computational System
  - Motherboards: Heterogeneous
  - Chips: On-board NICs/PCIe
  - Memory: Stacked
- Software: Handled through Co-Design
  - Applications
  - Middleware
  - Compilers
- Networking
  - Interconnect (NDR IB): Between nodes
  - Intra-center resources (100Gb - 400Gb Ethernet): Between systems
  - Inter-center resources (100Gb - 400Gb Ethernet): Between Centers
- IO
  - Off computational system (file system)
  - Long-term storage (archive)
  - WAN data movement (between Centers)

The Major System Components of Exascale

- Computational System
  - Motherboards: Heterogeneous
  - Chips: On-board NICs/PCIe
  - Memory: Stacked
- Software: Handled through Co-Design
  - Applications
  - Middleware
  - Compilers
- Networking
  - Interconnect (NDR IB): Between nodes
  - Intra-center resources (100Gb - 400Gb Ethernet): Between systems
  - Inter-center resources (100Gb - 400Gb Ethernet): Between Centers
- IO
  - Off computational system (file system)
  - Long-term storage (archive)
  - WAN data movement (between Centers)

Exascale I/O: Interconnect Requirements

- Power efficiency gains of 10x over present
  - Optics present on the node possibly on the chip (50% power reduction), especially important for 100Gb+ devices
- Scalability to handle O(100,000) to O(1B) nodes
- Performance improvements
  - 200-400GB/sec inter-node BW
- Resiliency improvements
  - Congestion
- Enable convergence of HPC networks within the center
  - Fiber channel reliability, with IB latency/bandwidth, with ethernet routing/features/manageability

Exascale I/O: File System Requirements

- Usability
  - Features to support data management and data analysis, more than just open/read/write
  - Aid in understanding hardware layout and software configuration to optimize performance
- Power efficiency
  - Enable spin-down of disks, use of flash (4096 byte devices), or other power saving storage
  - If none, expect IO subsystem to require up to 2.5 of 20MW of power
- Resiliency
  - Management/debug features to handle O(20,000) components
  - Software failover, tolerant of errors
  - Software to complement hardware RAID rebuilds/size of disks
- Scalability
  - Need to handle O(20,000) devices and O(100,000-1M) clients
- Performance
  - Target is 1TB/sec
- Metadata
  - Need multiple metadata servers in software
  - Likely using memory for speed-up (FS cache, or DRAM SSD devices)
  - Backups (mostly about a tree-walk) need to be feasible in some number of days
- Cost
  - Need more % of system cost for adequate BW/capacity IO subsystem (high estimate is $60M)

Exascale I/O: Archival Storage Requirements

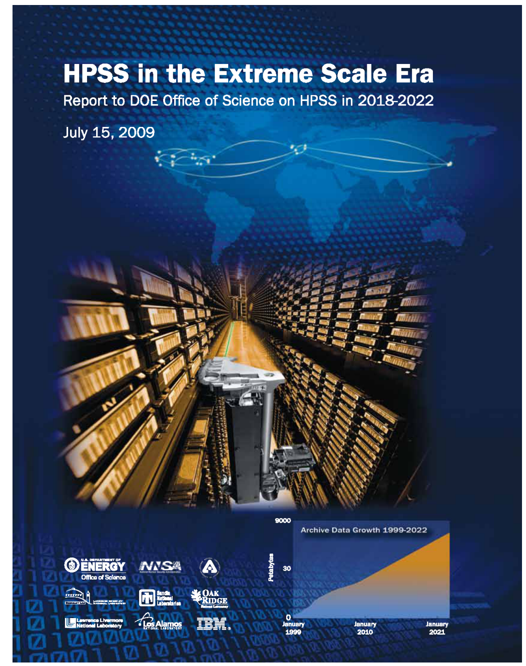
- Usability
  - Features to support data management and data analysis, more than just open/read/write
  - Aid in understanding hardware layout and software configuration to optimize performance
- Power efficiency
  - Enable spin-down of disks, use of flash (4096 byte devices), or other power saving storage
  - If none, expect IO subsystem to require up to 2.5 of 20MW of power
- Resiliency
  - Management/debug features to handle O(20,000) components
  - Software failover, tolerant of errors
  - Software to complement hardware RAID rebuilds/size of disks
- Scalability
  - Need to handle O(20,000) devices and O(100,000-1M) clients
- Performance
  - Target is 1TB/sec
- Metadata
  - Need multiple metadata servers in software
  - Likely using memory for speed-up (FS cache, or DRAM SSD devices)
  - Backups (mostly about a tree-walk) need to be feasible in some number of days
- Cost
  - Need more % of system cost for adequate BW/capacity IO subsystem (high estimate is $60M)

Exascale I/O: WAN Data Movement Requirements

- PB data sets will be common and will need to move between facilities. We are already moving data sets in the 10's of TBs between facilities monthly.
- Human time scales are important
- Mounting of other Center's file systems unlikely to support science
  - Federation of accounting/users (authentication and authorization), very difficult
  - Additional security for devices on someone else's network
  - Changes to enable high-latency operations as the norm
- Explicit data transfers
  - High throughput network configured to optimize data transfers
    - ESnet SDN
  - Software to aid in unattended data movement between facilities
    - Third-party data transfer services GlobusOnline.org
    - Storage resource managers (BeSTMan)
  - Dedicated servers close to site's border with Center's storage resources available to it
    - Data transfer nodes, parallel file systems, archival storage

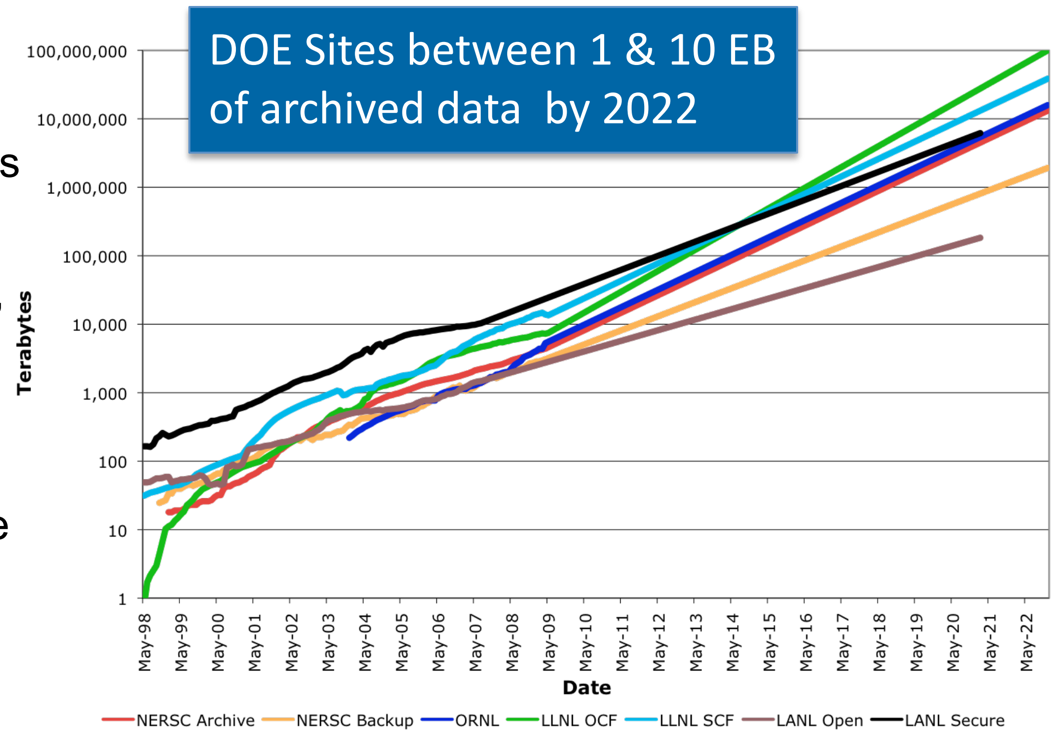
Archival Storage

- Extreme Scale Workshop July 2009
  - "HPSS in the Extreme Scale Era" report
  - Surveyed six DOE sites for data trends and stats
  - Performed a market survey of archival storage software
  - Provided roadmaps for disk & tape through 2022
  - Gathered archival storage requirements from other Exascale reports



Exascale Archival Storage Scalability Requirements

- Storage capacity
  - Annual growth O(10PB)
  - Amount of data stored in single system will be 1-10EB in 1-10B files
- Ingest Bandwidth
  - 10% of Scratch File System speed, O(100GB/s) peak and O(10GB/s) sustained
- Metadata speed
  - PB sized, file operations 10% of file system capabilities
  - Multiple metadata servers (PureScale DB2 interesting)
- Network between systems/storage
  - Network capable of 100GB/s



Exascale Archival Storage Data Management Requirements

- Data discovery
  - Middleware challenge
- Data mining
  - Middleware challenge
- Data set operations
  - GPFS and HPSS have a start on this

Exascale Archival Storage System Management Requirements

- Usability of system management interface
  - Managing O(1,000) software processes in single metadata server
  - Managing multiple metadata servers (like distinct systems)
- Logging subsystem scaling to O(1,000) software processes (100's of threads each) logging in real-time to central source
- Continue scaling real-time monitoring of a very large complex system

Exascale Archive Storage Hardware Requirements

- Affordability at scale
  - O(90,000) tapes with 80TB tape to retain one year of IO to archive from Exascale system. This is $27M in annual tape budget with today's tape cost
- Performance at scale
  - Each tape drive 600MB/s

Final Thoughts

- I/O is a major part of the Exascale system design
- Networking initiatives and research underway
- Co-design proposals being awarded
- Storage requires evolutionary
  - Exascale capable file systems and archival storage to continue improvements
- Revolutionary storage could help with
  - Performance improvements over current rates
  - Reliability improvements over existing systems
  - Power efficiency improvements over existing
  - Moving analysis closer to storage