Simulation Driven I/O and the Exascale - an LLNL perspective

Presented to MSST 2010 Conference

Mark Seager

Lawrence Livermore National Laboratory

5 May 2010

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 LLNL-PRES-TBD

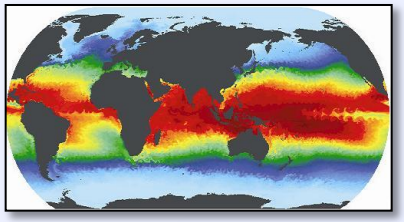
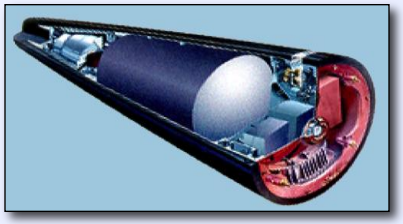
Talk Overview

- DOE Exascale Mission Drivers
- Implications for Storage
- Propose a New Exascale Storage Co-Design Concept

Artist's rendition of Sequoia

DOE mission imperatives require simulation and analysis for policy and decision making

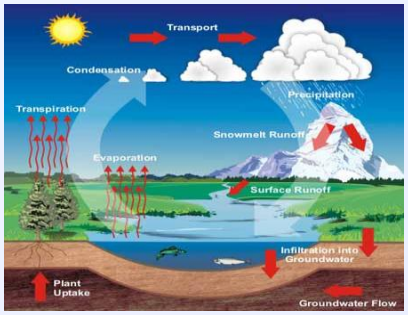
- *Climate Change*: Understanding and mitigating the effects of global warming
  - Sea level rise
  - Severe weather
  - Regional climate change
  - Geologic carbon sequestration
- *National Nuclear Security*: Maintaining a safe, secure and reliable nuclear stockpile
  - Stockpile certification
  - Predictive scientific challenges
  - Real-time evaluation of urban nuclear detonation
- *Energy*: Reducing U.S. reliance on foreign energy sources and reducing the carbon footprint of energy production
  - Reducing time and cost of reactor design and deployment
  - Improving the efficiency of combustion energy sources


Accomplishing these missions requires exascale resources.

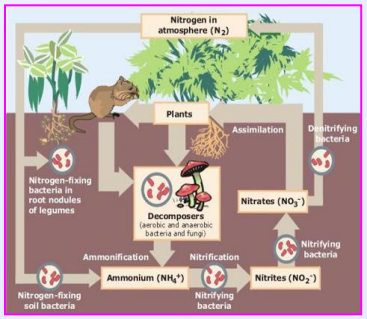
Potential Impact of Exascale Computing on Climate Assessments

- Ocean and Ice
  - Improved modeling of land ice to reduce uncertainties in predictions for sea level rise and changes in ocean circulation.
- Hydrological cycle
  - Reduce uncertainties in prediction of cloud radiative effects and precipitation due to climate change.
- Extreme Weather
  - Use multi-scale Earth System Models (ESMs) to understand statistics of cyclones, blizzards, meso-scale storms, heat waves, droughts, and frost.
- Carbon, Methane and Nitrogen cycles
  - Use ESMs with improved chemical and biological process models to reduce uncertainties in predictions of the evolution



Hydrological Cycle
NWS

Greenland Ice
NASA

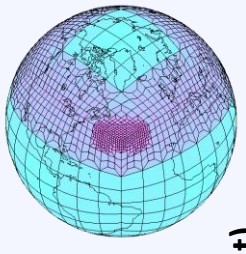


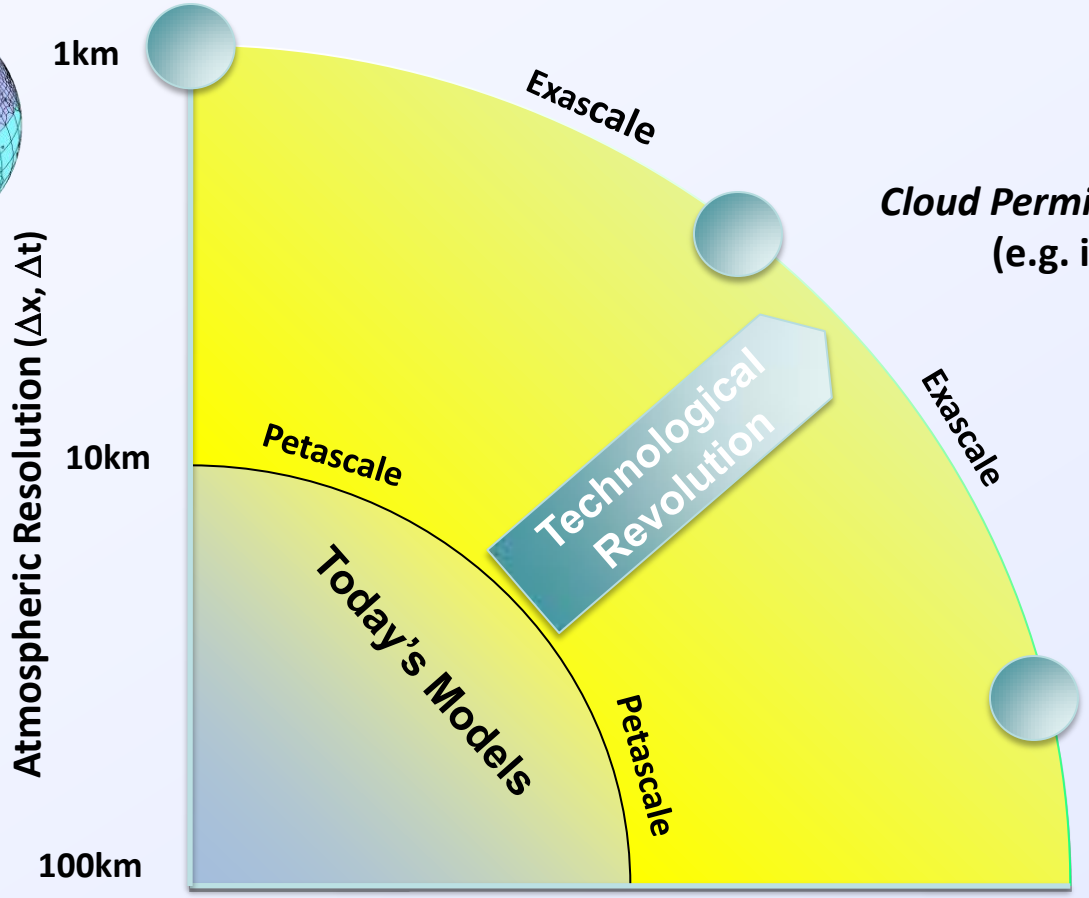
Nitrogen Cycle
EPA

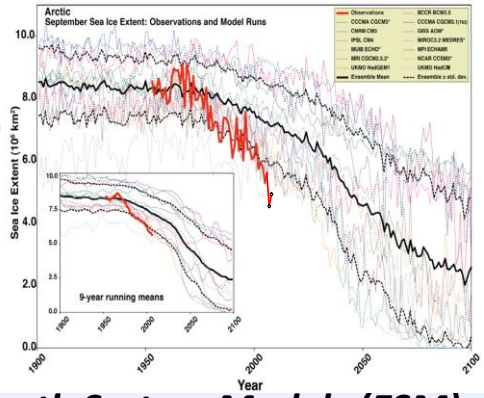
Hurricane Floyd
NOAA

"Given these drivers … it is clear that exascale computers and ultra fast networks, data systems and computational infrastructure will be required by 2020." ***Challenges in Climate Change Science and the Role of Computing at Extreme Scale, November, 2008***

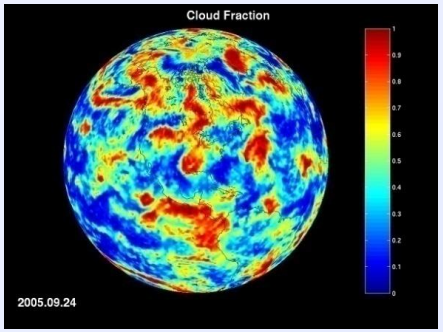
Global Cloud Resolving
(e.g. reduce main climate change uncertainties (clouds)



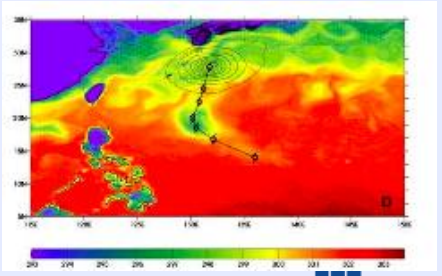




Cloud Permitting Earth System Models (ESM)
(e.g. improved sea-level rise forecasts)



Massive ESM ensembles
(e.g. decadal forecasts, extreme weather)

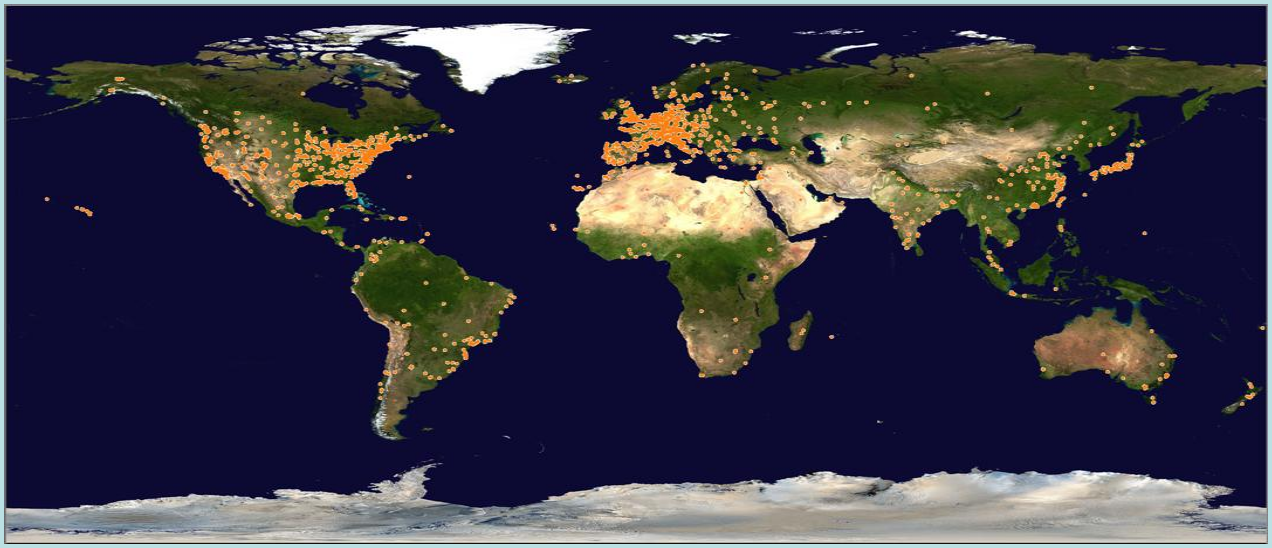


Science Gateway

Earth System Grid
Center for Enabling Technologies

Agencies

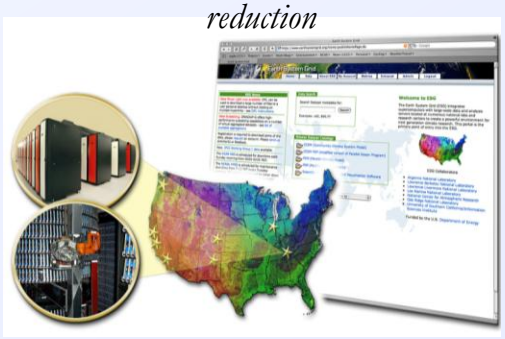
- **Registered sites**



ESG-CET enables Scientific Discovery in Climate Science by providing an international community of over 16,000 registered users with climate simulation data, climate models, analysis and visualization tools, and enabling technologies for a distributed, global science enterprise

ESG turns climate science data into community resources

Data warehouse, search and discovery, access, and reduction



Data used in hundreds of scientific papers

Much of which provided a basis for the 4th Assessment Report of the IPCC

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE

The Nobel Peace Prize 2007

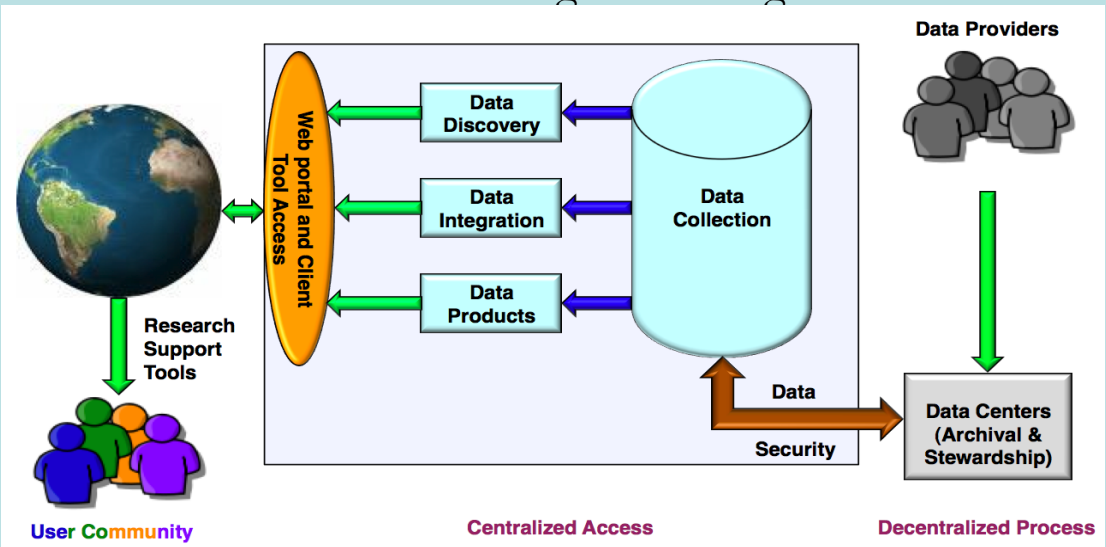
Intergovernmental Panel on Climate Change (IPCC)

Enabling Technologies

Earth System Grid
Center for Enabling Technologies

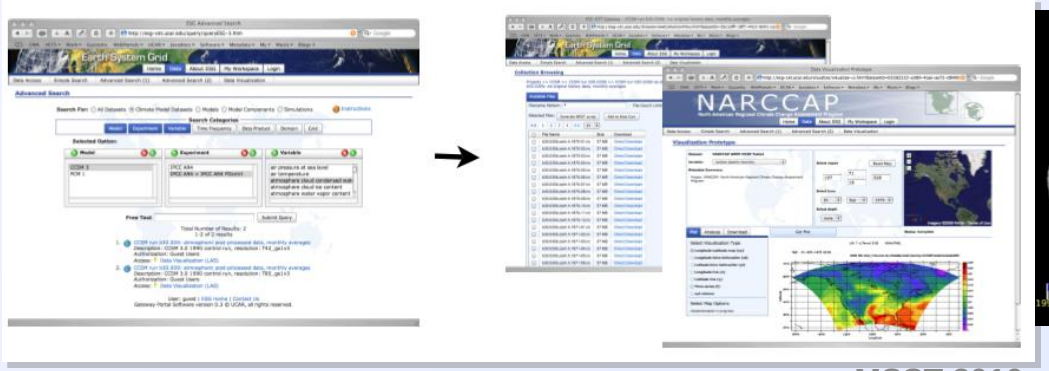
Technologies

User Community: Scientist, policymakers, economists, health officials, etc.

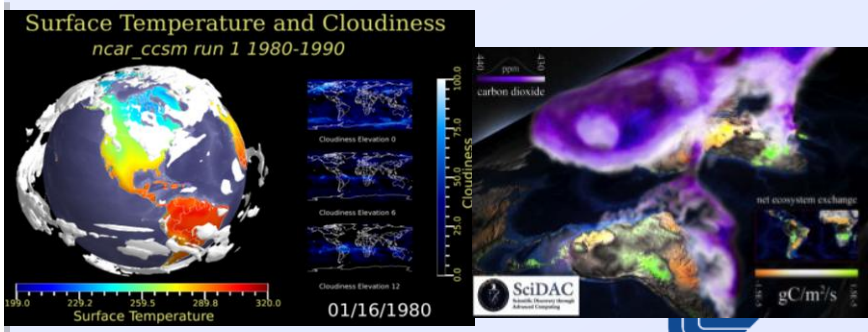


Climate change is not only a scientific challenge of the first order but also a major technological challenge. PCMDI is working to integrate a far larger number of distributed data providers, high-bandwidth wide-area networks, and remote computers in a highly collaborative problem-solving environment.

Research Support Tools (e.g., Web Applications)



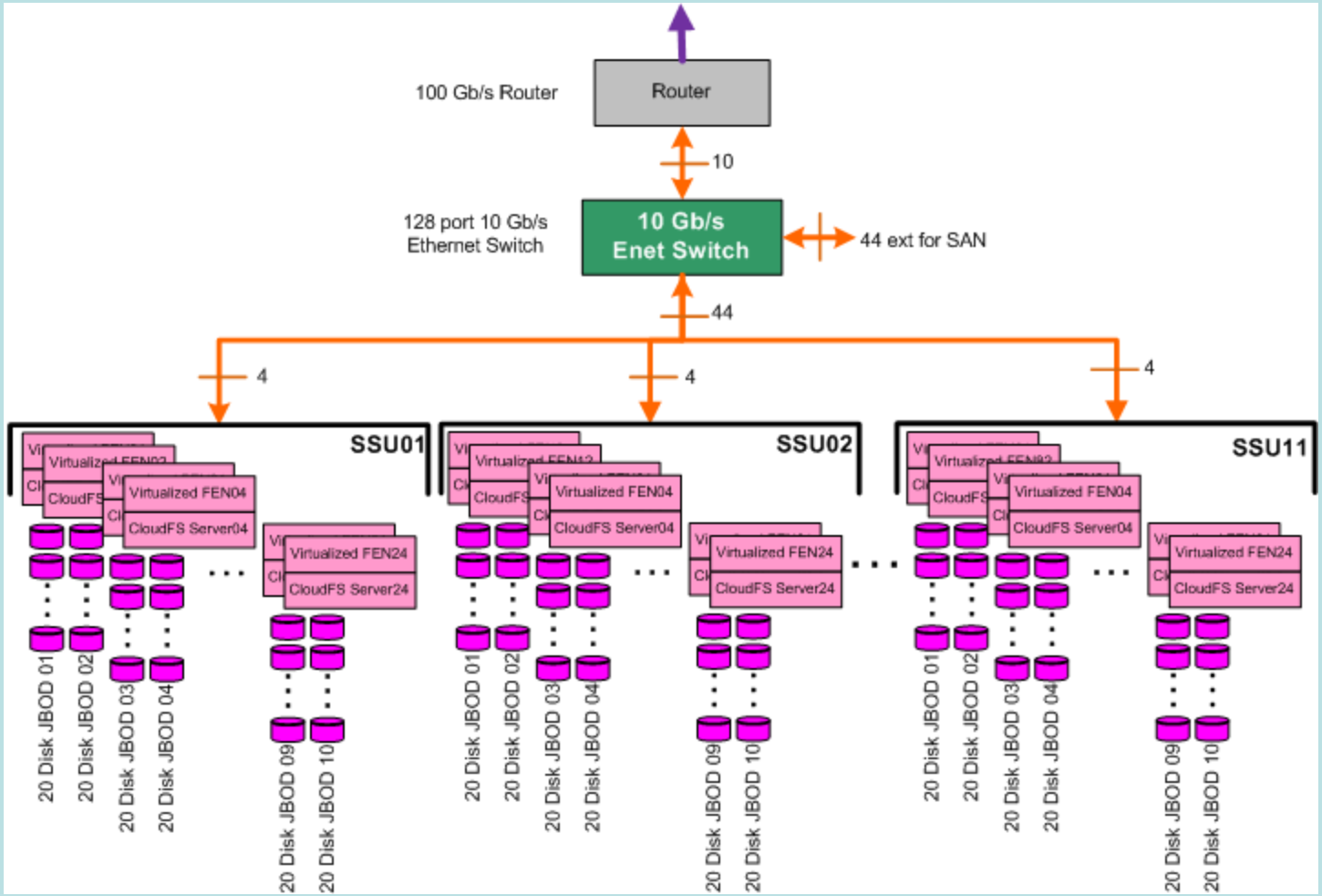
Data Products (e.g., Analysis and Visualization)



Cloud Green Data Oasis Hardware

Earth System Grid
Center for Enabling Technologies

- 264 virtualized CloudFS servers
- 10.6 PB raw disk
- 10 GB/s IO BW to ESNet 100Gb/s Router
- 44 GB/s to SAN



Challenges of an Ensemble of Calculations

How do I run an ensemble of calculations???

Problem: Different machines, architectures, command sets, etc.

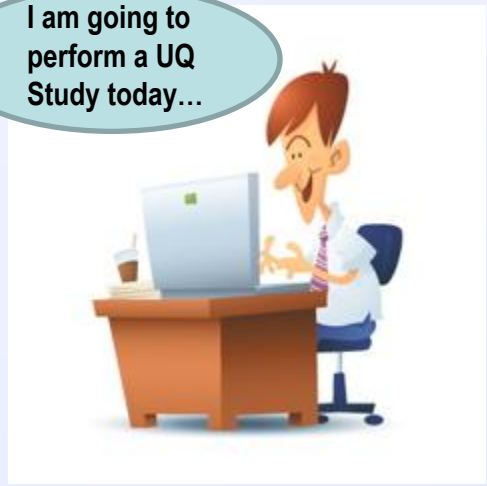
Want to shield the user from these differences and yet allow the user to utilize the different capabilities offered by the resources.

What variables are intended to be sampled investigated???

Problem: Curse of Dimensionality (2^n)

A set of sampling techniques is needed to mitigate the problem

I am going to perform a UQ Study today…

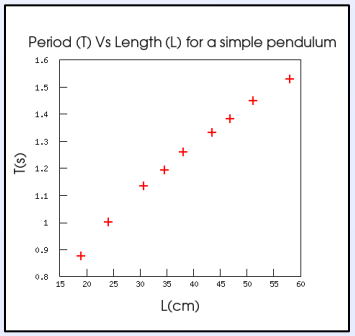


Thousands of simulations with terabytes of data produced

What is the status???

Data Overload

Need status/analysis tools that efficiently describes to the user the status of their simulation



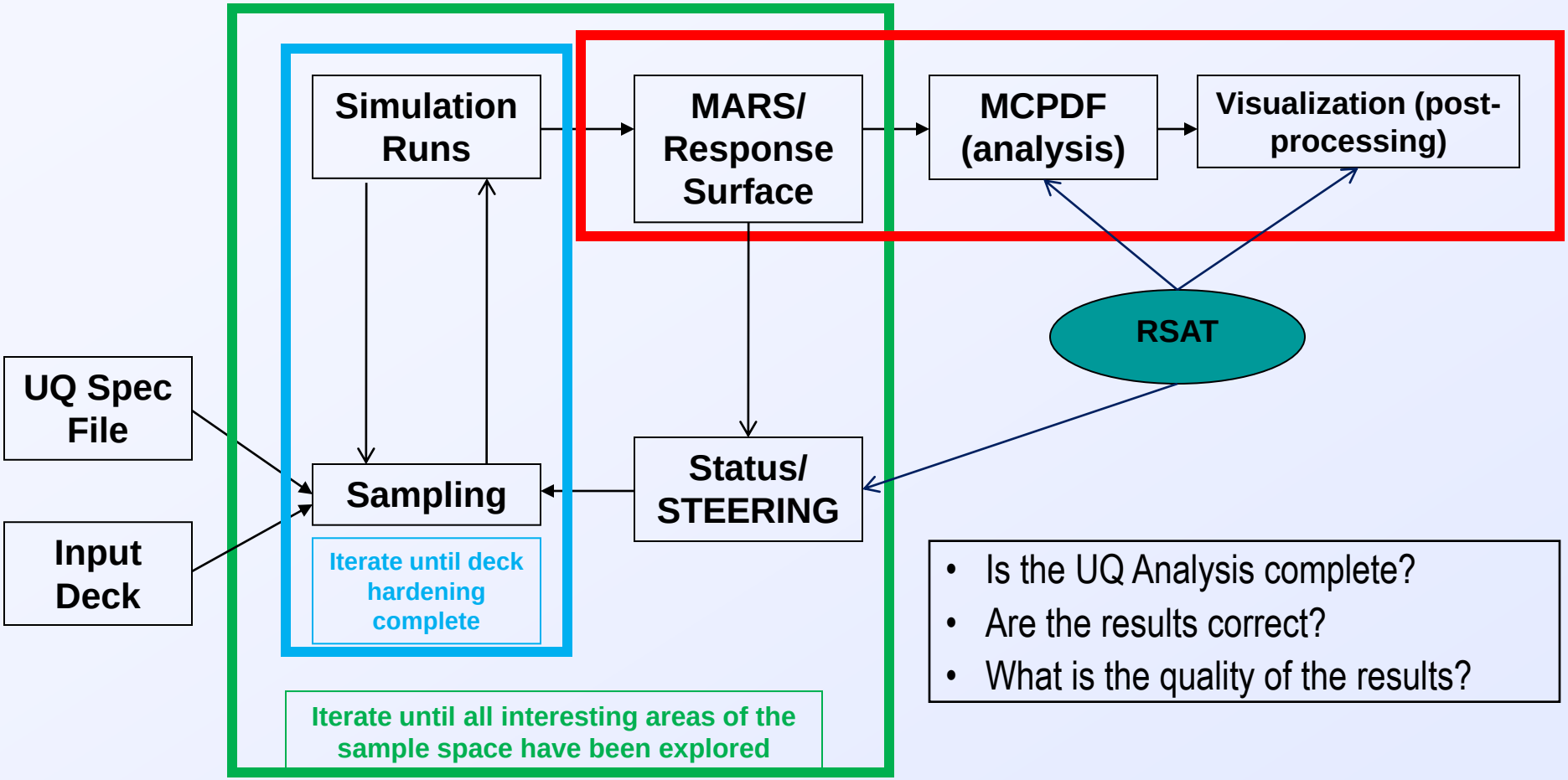
Does I need another set of simulations or am I finished?

If not, where in the sample space should I focus my next set of simulations?

Other Challenges:

1) Visualization of Large amounts of data
2) Others???

Overview of the UQ Pipeline (AX V&V Methodology)



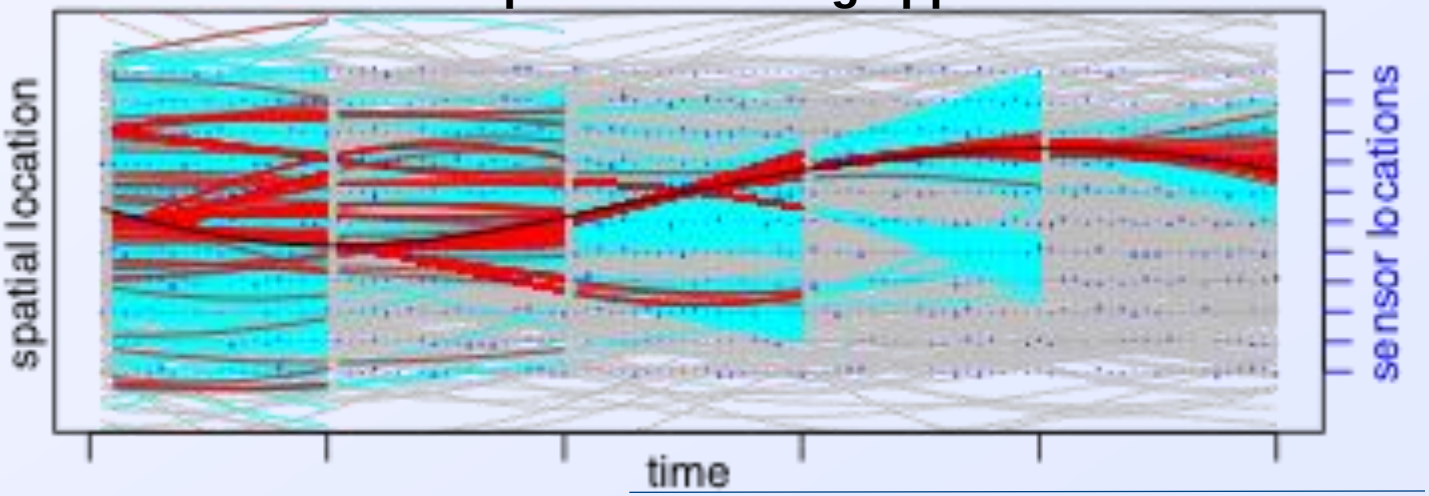
Uncertainty quantification will require tightly coupled resources

capacity → **capability**

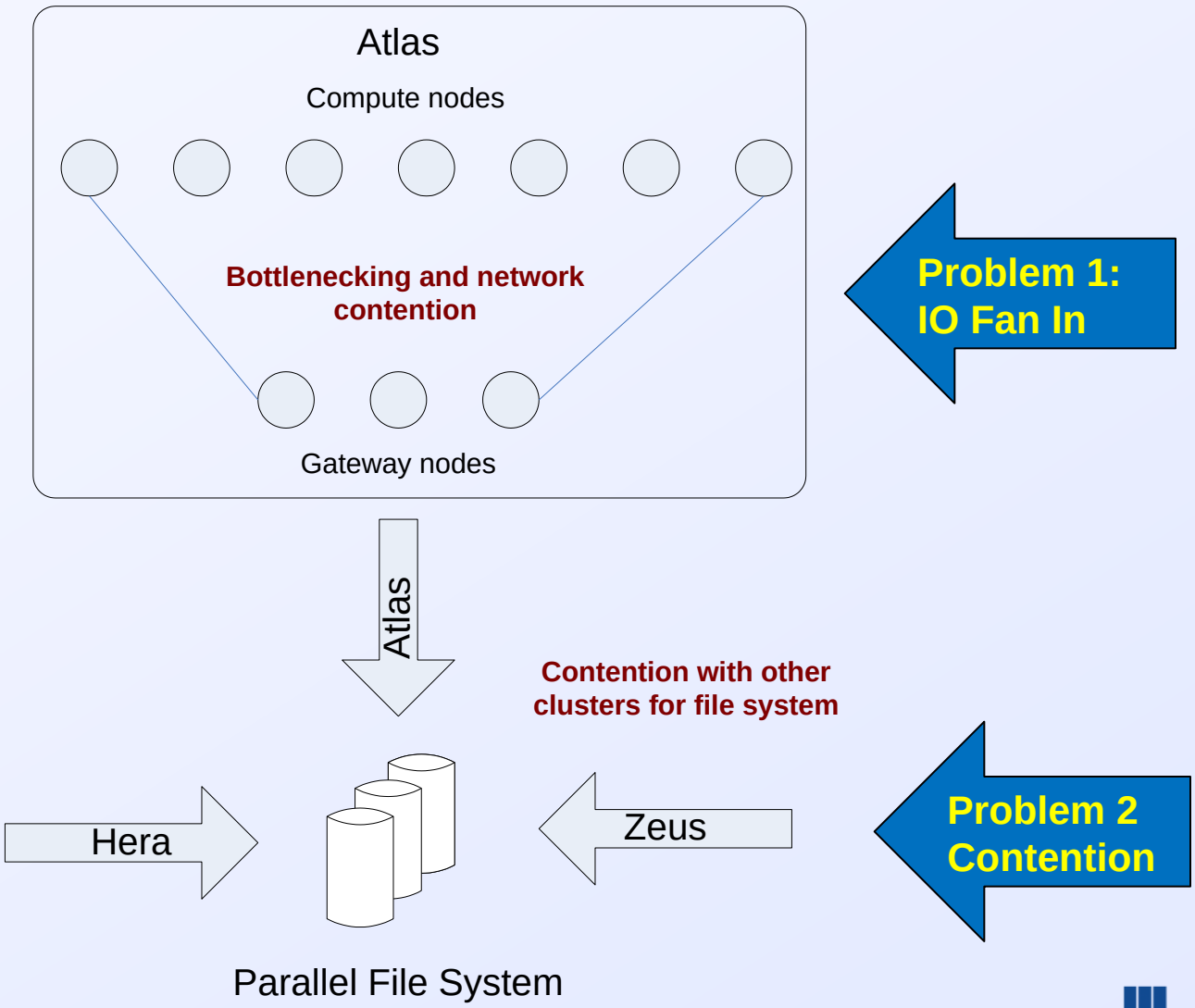
- $10^{3\text{-}8}$ runs
- Response surface
- Posterior exploration
- Finding least favorable priors
- Bounds on functionals

- 10 runs
- Adjoint enabled forward models
- Extract data from forward model
- Local approximations, response surface, filtering

Example of a filtering approach



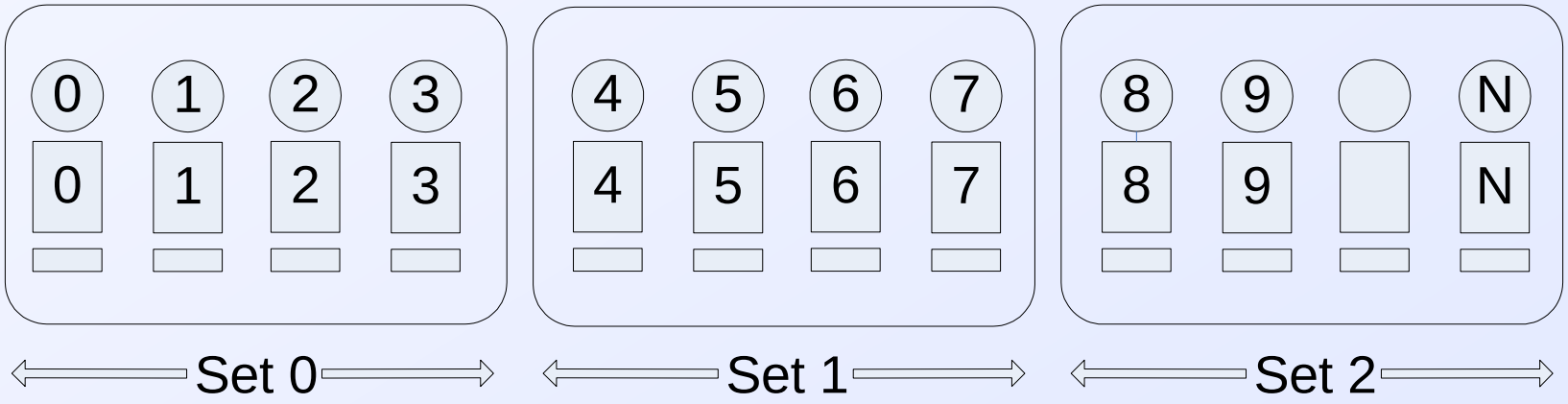
By writing to local file systems, Scalable Ceckpoint/Restart (SCR) Library Avoids Two Problems



From Adam Moody

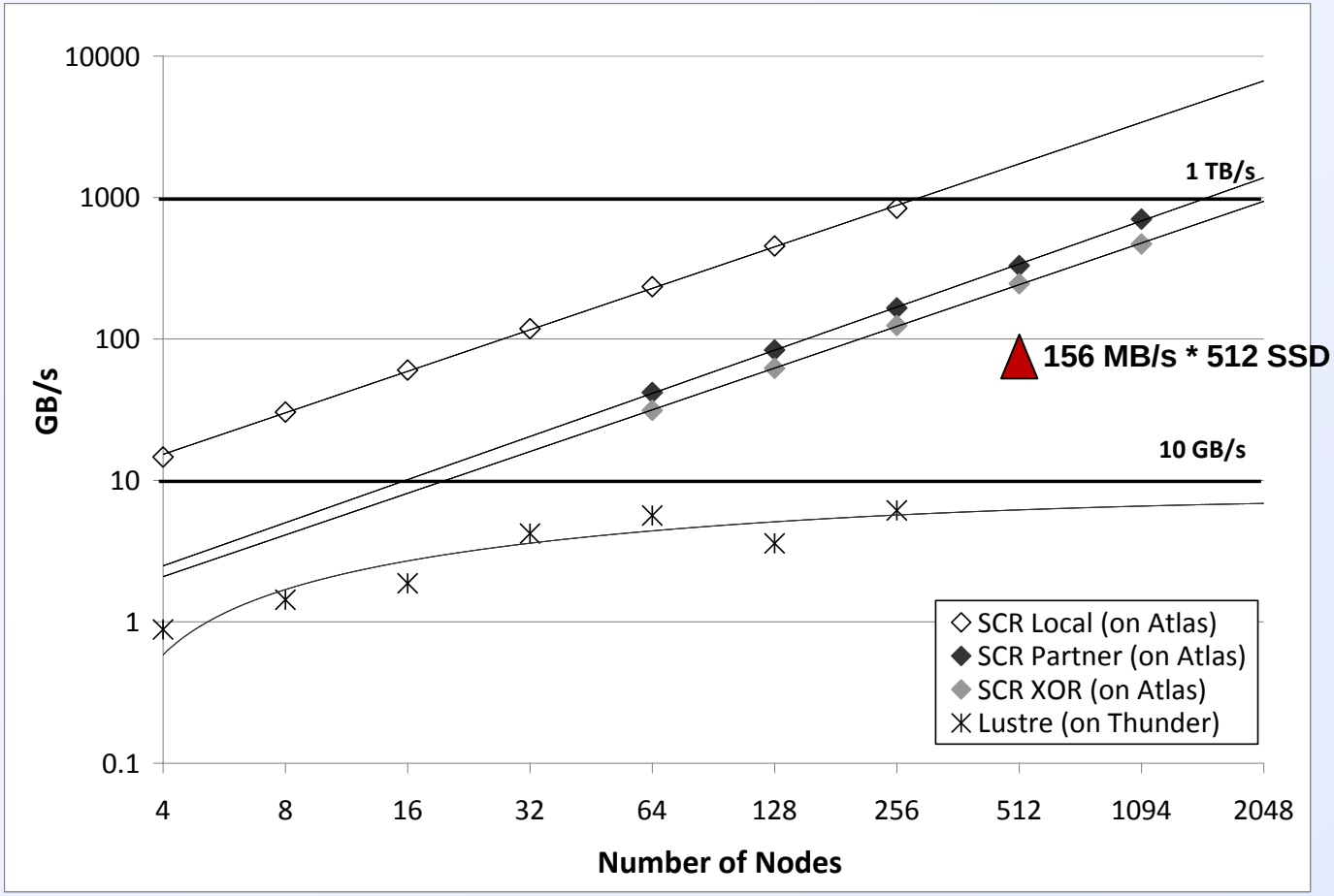
SCR utilizes a sophisticated XOR redundancy to distribute data and reduce overheads

- Break nodes for job into smaller sets, and execute XOR reduce scatter within each set.
- Can withstand multiple failures so long as two nodes in the same set do not fail simultaneously.



From Adam Moody

Benchmark checkpoint times to RAM disk and local SSD provide scalable bandwidth to applications

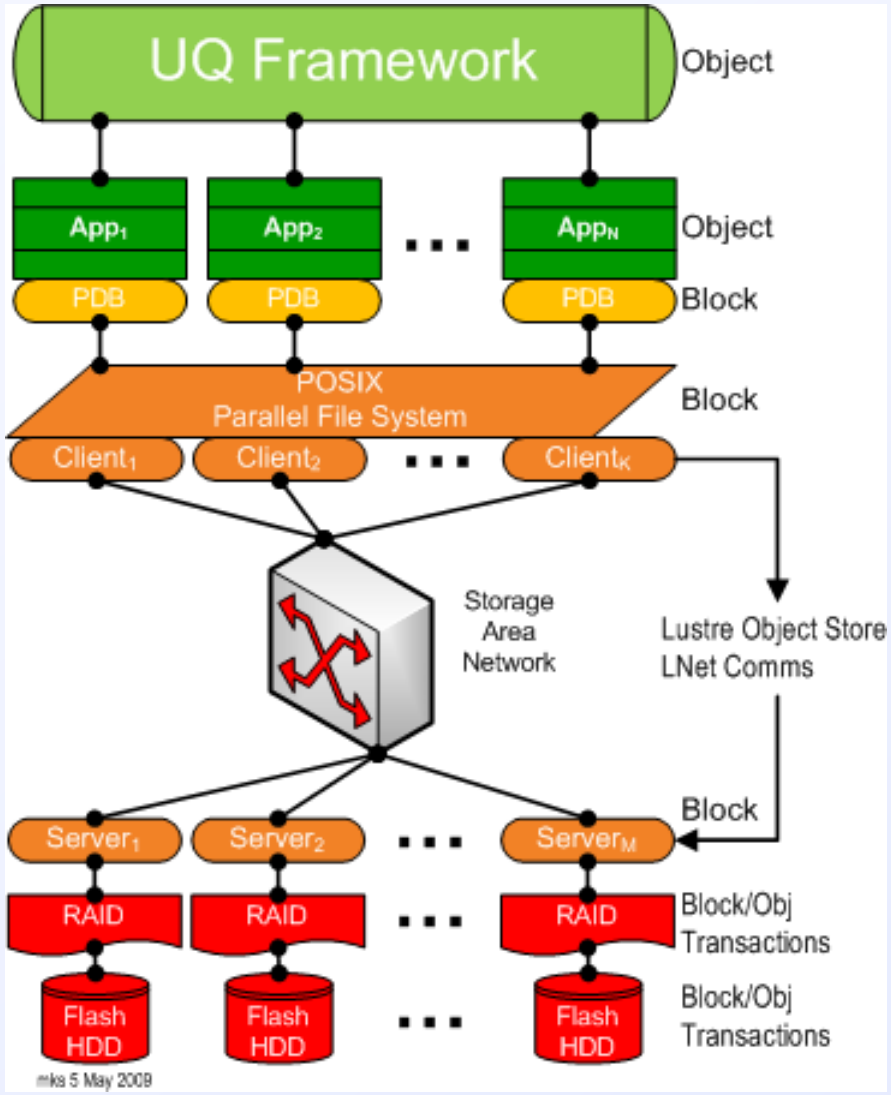


From Adam Moody

Exascale Mission Drivers Will Require Rethinking Data Storage Management

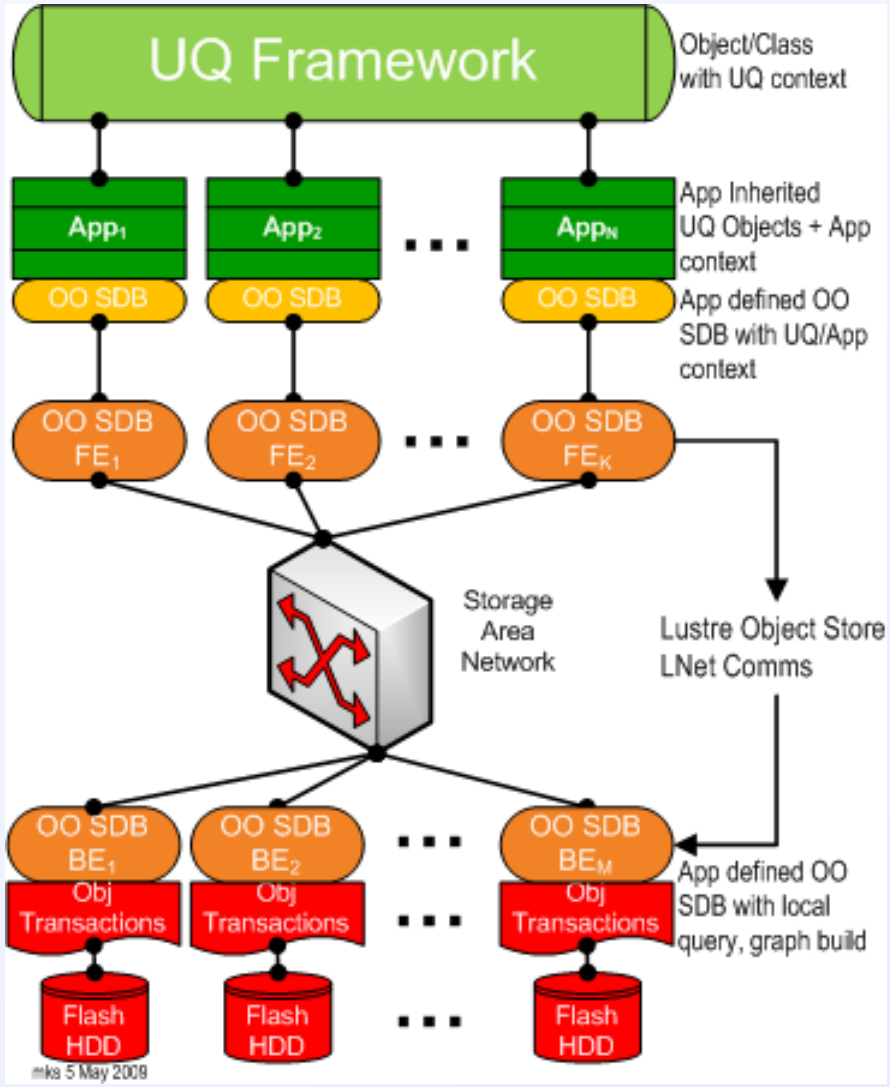
- Climate Model Analysis
  - Requires 10s of codes, 100s of runs, 1000s of model parameter variations
  - O(10s) PB at petascale, O(10s) EB at exascale
  - Metadata and Data provenance are critical
  - Long lived data and post processing essential
- UQ Analysis
  - Multiple code runs, curse of dimensionality
  - Requires capacity (big data from 10^{3-8} runs and capability (big data per run)
  - Metadata and data provenance are critical
  - Short to medium term data life, dynamic post processing guides future runs
- Resiliency
  - Higher bandwidth better
  - Short data life
  - Minimize data movement

Current data storage hierarchies don't support applications driven storage model



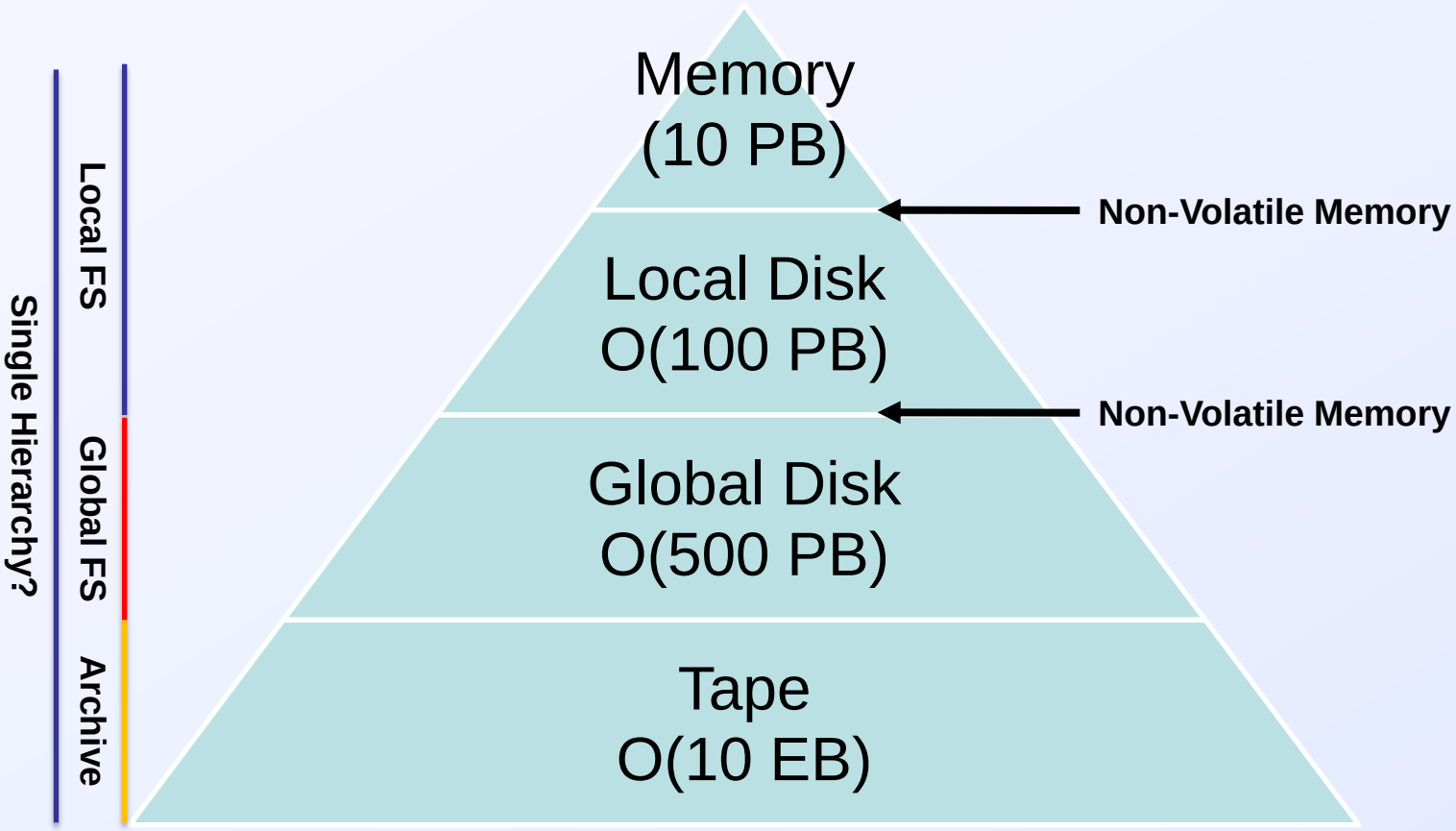
- File system block interface breaks object model
- Based on 1960's technologies and techniques
- Data and storage differentiated
- User access is shell+ls
- User metadata is lost

How about an applications driven object oriented data storage heirarchy



- UQ, Applications define objects
- Storage of objects is abstracted
- Includes remote method invocation for user computations near the data
- Access transformed from shell+ls ➔ Python
- Metadata is accreted during object creation and IO

Exascale Storage Hierarchy May Drive Different Choices



Summary

- Exascale Mission Drivers Span the Space from Weapons Science to Climate
  - Ensemble calculations and UQ analysis will become the norm
  - The complexity of data will grow enormously
  - Role of metadata and data provenance will become mission critical
  - Traditional storage hierarchies need major revision
- Propose a Two Pronged Approach for Exascale Storage Co-Design
  - Evolutionary approach
  - Revolutionary approach based on a blank sheet of paper and 50 years of “lessons learned”
  - Bridge from one to the other