Scientific Data Management: Challenges and Approaches in the Extreme Scale Era

Arie Shoshani (LBNL)

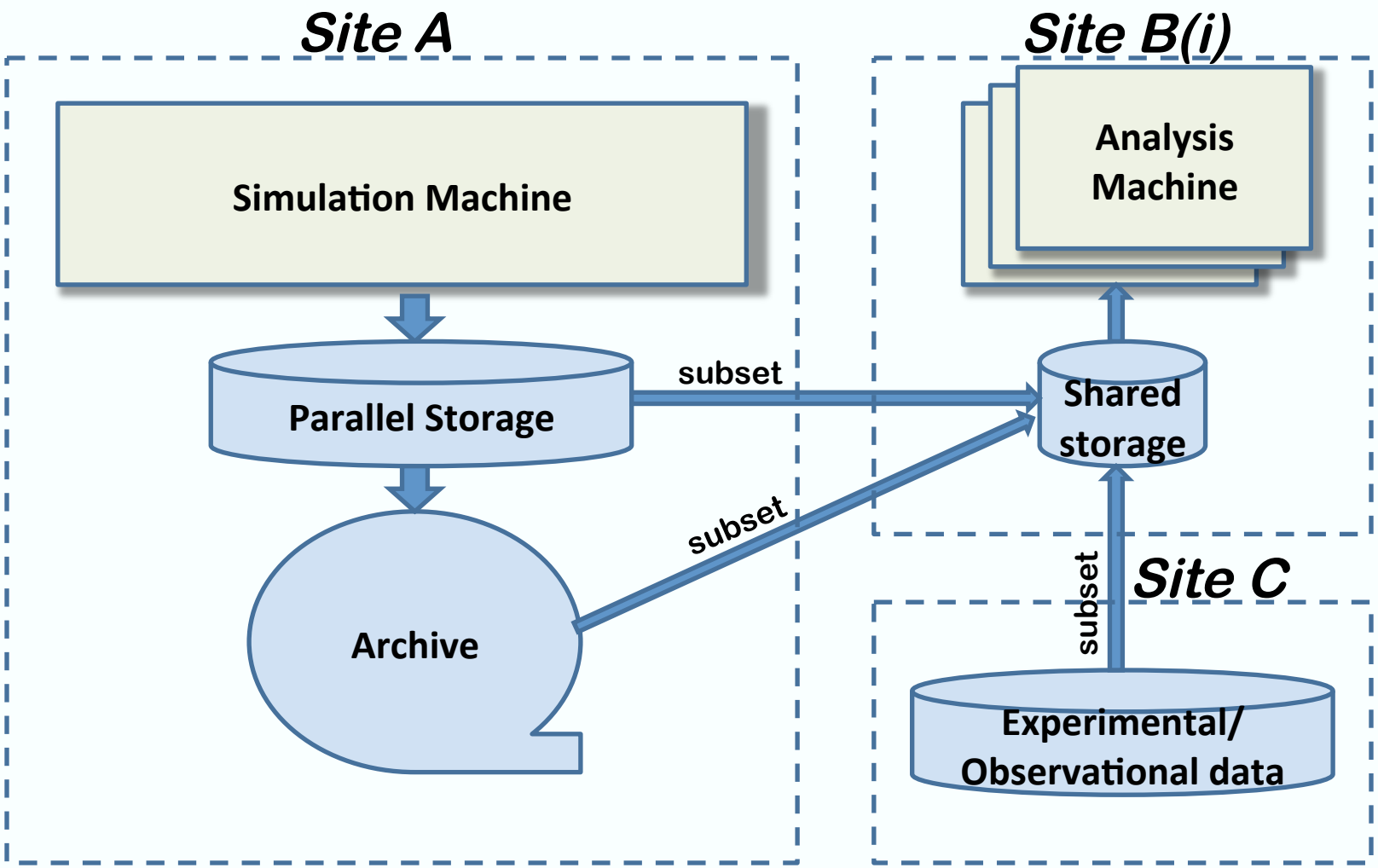
(and the entire SDM center team)

NERSC Workshop, January 5-6, 2010

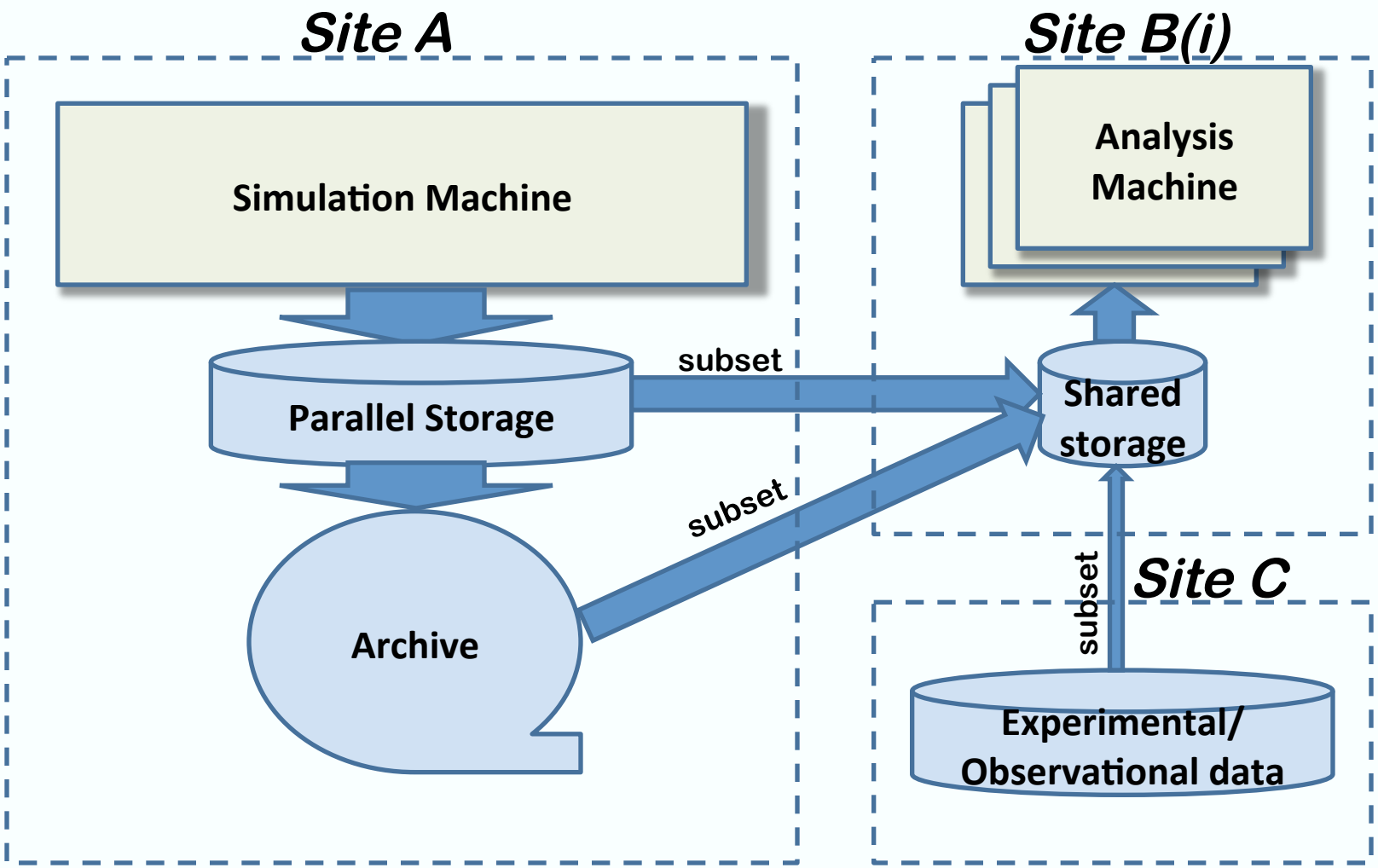
What is Scientific Data Management?

- Algorithms, techniques, and software
  - Representing scientific data – data models, metadata
  - Managing I/O – methods for removing I/O bottleneck
  - Accelerating efficiency of access – data structures, indexing
  - Facilitating data analysis – data manipulations for finding meaning in the data
- Current practice – data intensive tasks
  - Runs large-scale simulations on large supercomputers
  - Dump data on parallel disk systems
  - Export data to archives
  - Move data to users' sites – usually selected subsets
  - Perform data manipulations and analysis on mid-size clusters
  - Collect experimental / observational data
  - Move to analysis sites
  - Perform comparison of experimental/observational to validate simulation data

Lots of Data Movement (GBs – TBs)

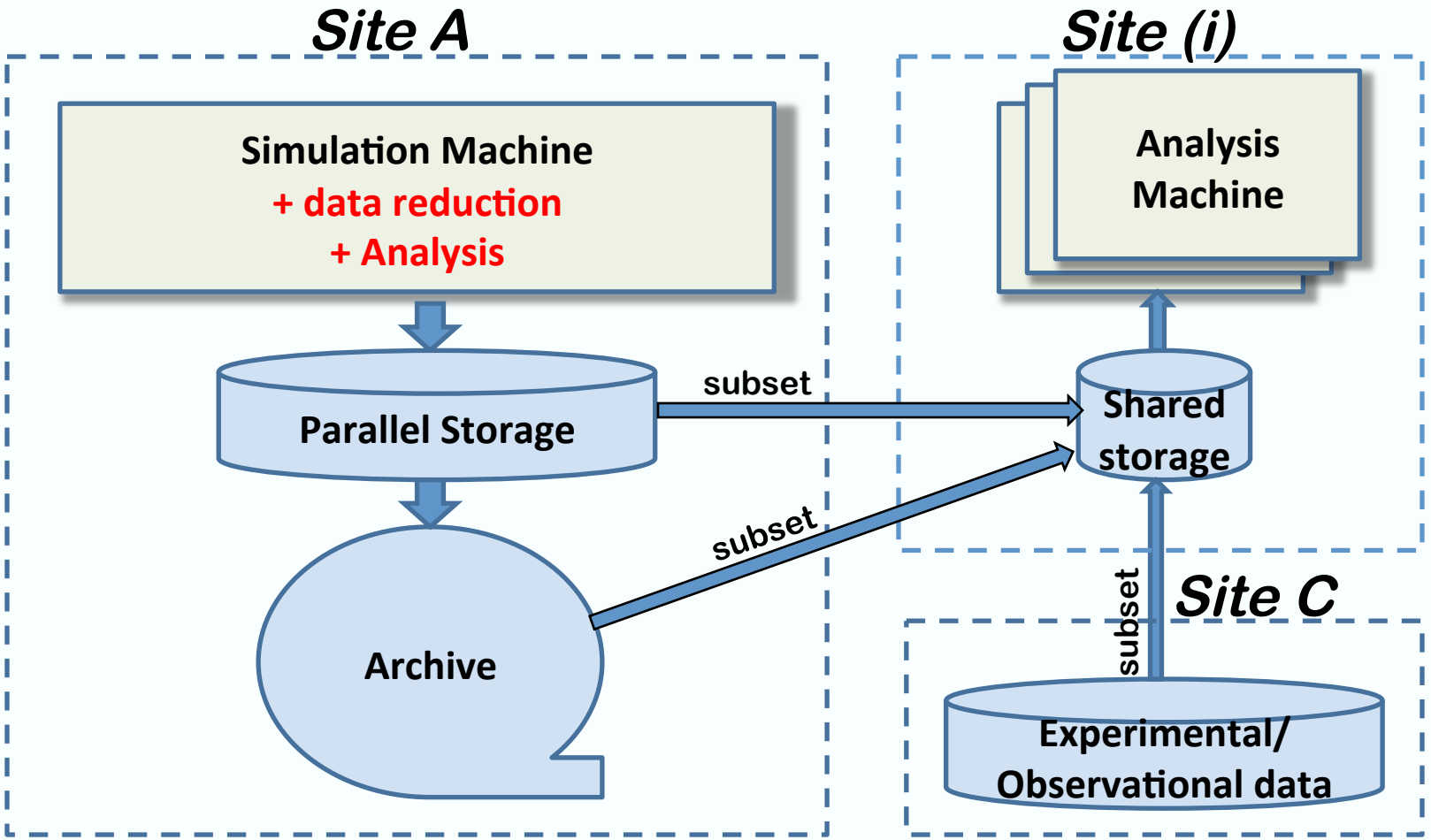


At Exascale (PBs) – data volume challenge



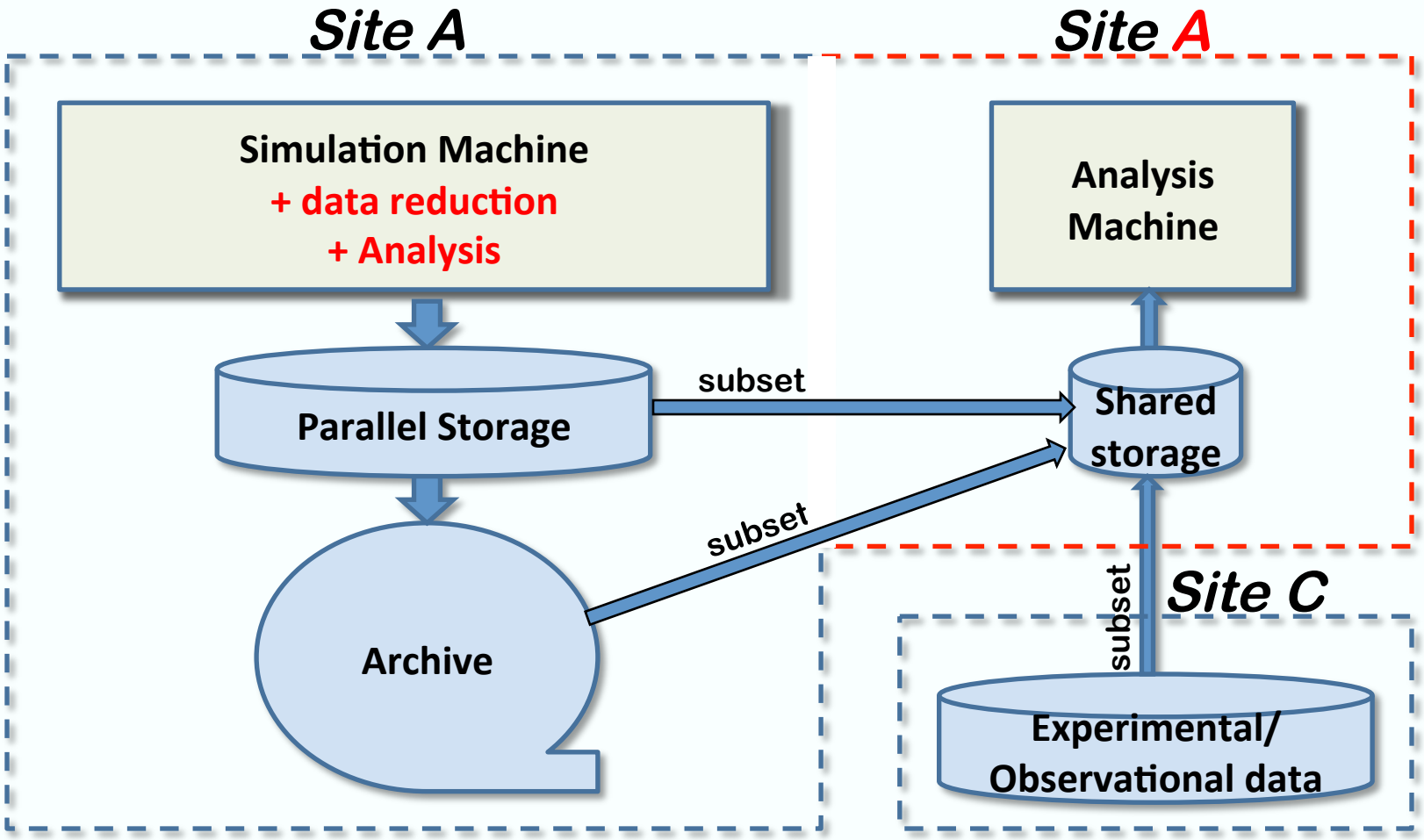
What Can be Done?

- Perform some data analysis on exascale machine
- Reduce and prepare data for further analysis



What Else Can be Done?

- Provide analysis machines where data is generated



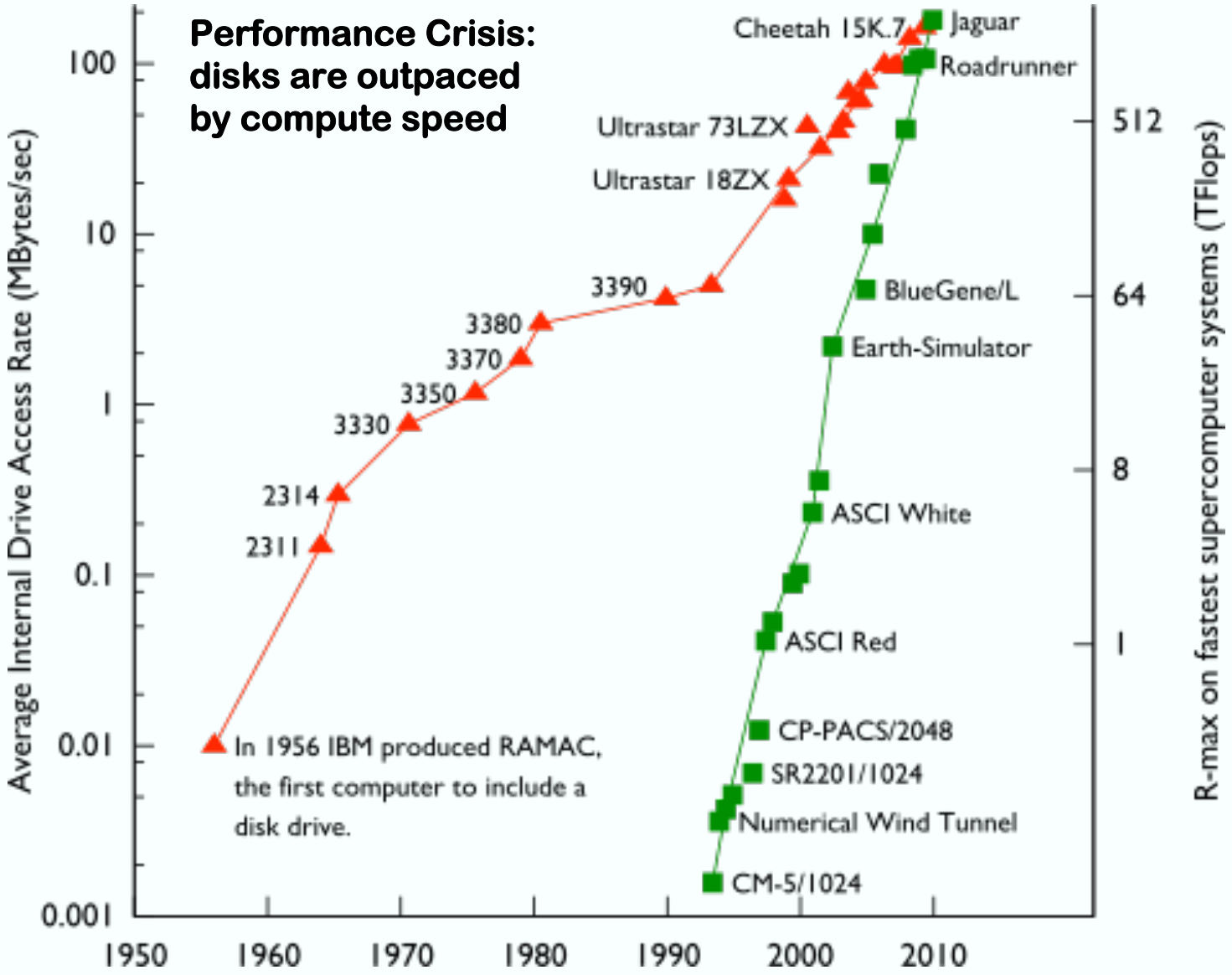
Data Analysis

- Two fundamental aspects
  - Pattern matching : Perform analysis tasks for finding known or expected patterns
  - Pattern discovery: Iterative exploratory analysis processes of looking for unknown patterns or features in the data
- Ideas for the exascale
  - Perform pattern matching tasks in the exascale machine
    - "In situ" analysis
  - Prepare data for pattern discovery on the exascale machine, and perform analysis on analysis machine
    - "In-transit" data preparation
    - "Off-line" data analysis

The Data Analysis Challenge

- Data exploration algorithms
  - Dimensionality reduction, decision trees, …
- Data mining algorithms
  - Graph analysis, clustering, classification, anomaly detection, …
- Data manipulation methods
  - Indexing, data transformation, data transposition, data compression, statistical summarization, …
- Preparing data for visualization
  - Generating graphs, contour plots, parallel-coordinate plots, 3D rotation, movies, …
- Tracking the analysis process
  - Workflow management, tracking cyclical activity, …
- Challenge: to make such tasks work on exascale machines

The I/O Challenge

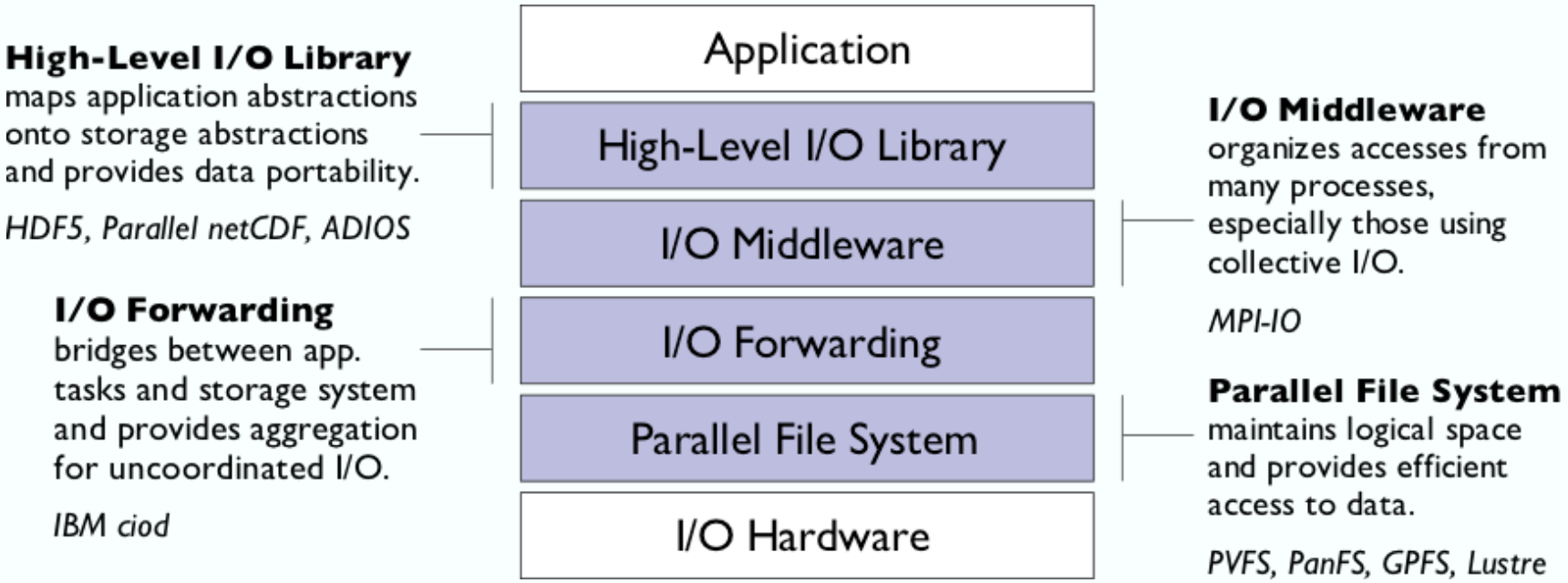


Exascale Systems: Potential Architecture

Systems	2009	2018	Difference
System Peak	2 Pflop/sec	1 Eflop/sec	O(1000)
Power	6 Mwatt	20 Mwatt	
System Memory	0.3 Pbytes	32-64 Pbytes	O(100)
Node Compute	125 Gflop/sec	1-15 Tflop/sec	O(10-100)
Node Memory BW	25 Gbytes/sec	2-4 Tbytes/sec	O(100)
Node Concurrency	12	O(1-10K)	O(100-1000)
Total Node Interconnect BW	3.5 Gbytes/sec	200-400 Gbytes/sec	O(100)
System Size (Nodes)	18,700	O(100,000-1M)	O(10-100)
Total Concurrency	225,000	O(1 billion)	O(10,000)
Storage	**15 Pbytes**	**500-1000 Pbytes**	**O(10-100)**
I/O	**0.2 Tbytes/sec**	**60 Tbytes/sec**	**O(100)**
MTTI	Days	O(1 day)	

From J. Dongarra, "Impact of Architecture and Technology for Extreme Scale on Software and Algorithm Design," Cross-cutting Technologies for Computing at the Exascale, February 2-5, 2010.

The I/O Software Stack



Overview of successful technologies in the SDM center

http://sdmcenter.lbl.gov

Arie Shoshani (PI)

Co-Principal Investigators

DOE Laboratories

ANL: Rob Ross
LBNL: Doron Rotem
LLNL: Chandrika Kamath
ORNL: Nagiza Samatova
PNNL: Terence Critchlow

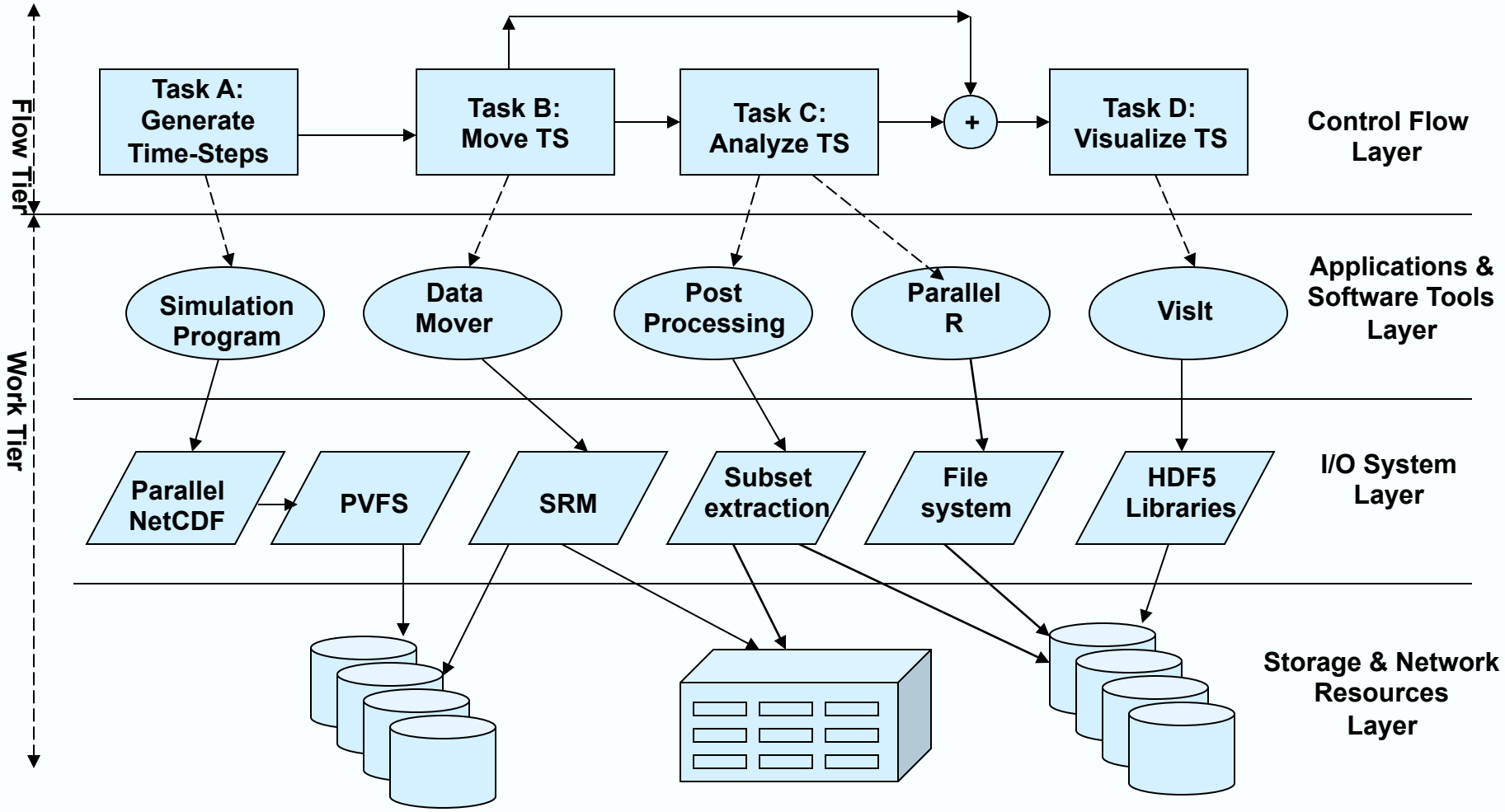
Universities

NCSU: Mladen Vouk
NWU: Alok Choudhary
UCD: Bertram Ludaescher
SDSC: Ilkay Altintas
UUtah: Claudio Silva

SDM Problems and Goals

- **Why is Managing Scientific Data Important for Scientific Investigations?**
  - **Sheer volume and increasing complexity of data being collected are already interfering with the scientific investigation process**
  - **Managing the data by scientists greatly wastes scientists effective time in performing their applications work**
  - **Data I/O, storage, transfer, and archiving often conflict with effectively using computational resources**
  - **Effectively managing, and analyzing this data and associated metadata requires a comprehensive, end-to-end approach that encompasses all of the stages from the initial data acquisition to the final analysis of the data**

A motivating SDM Scenario (dynamic monitoring)



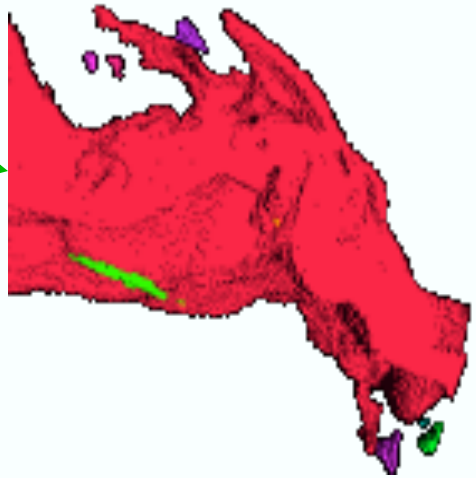
Searching Problems in Data Intensive Sciences

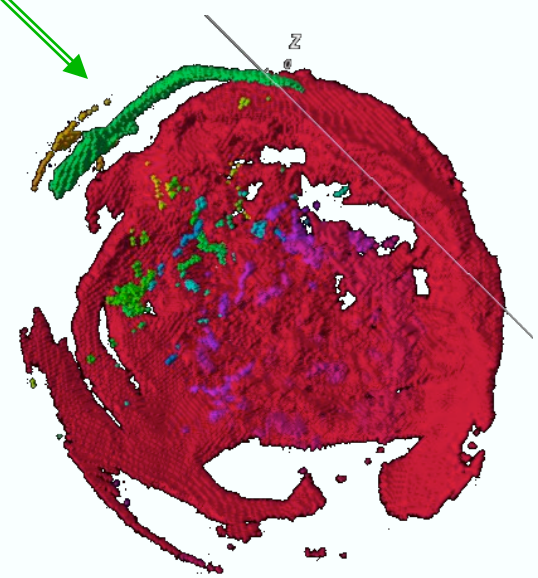
- Find the HEP collision events with the most distinct signature of Quark Gluon Plasma
- Find the ignition kernels in a combustion simulation
- Track a layer of exploding supernova

These are not typical database searches:

- Large high-dimensional data sets (1000 time steps X 1000 X 1000 X 1000 cells X 100 variables)
- No modification of individual records during queries, i.e., append-only data
- M-Dim queries: 500 < Temp < 1000 && CH3 > 10^{-4} && …
- Large answers (hit thousands or millions of records)
- Seek collective features such as regions of interest, histograms, etc.

- Other application domains:
  - real-time analysis of network intrusion attacks
  - fast tracking of combustion flame fronts over time
  - accelerating molecular docking in biology applications
  - query-driven visualization



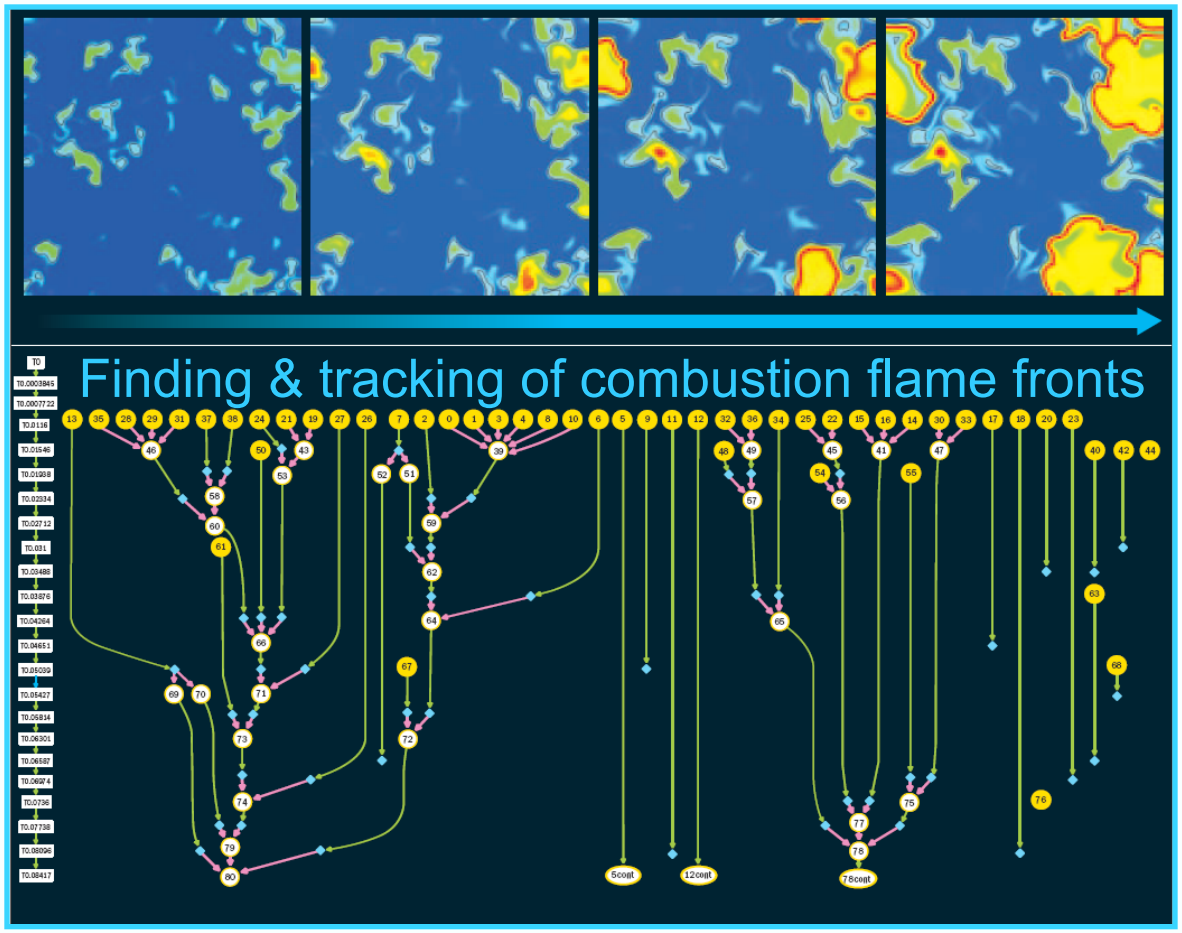


FastBit: accelerating analysis of very large datasets

- Most data analysis algorithm cannot handle a whole dataset
  - Therefore, most data analysis tasks are performed on a subset of the data
  - Need: very fast indexing for real-time analysis
- FastBit is an extremely efficient compressed bitmap indexing technology
  - Indexes and stores each column separately
  - Uses a compute-friendly compression techniques (patent 2006)
  - Improves search speed by 10x – 100x than best known bitmap indexing methods
  - Excels for high-dimensional data
  - Can search billion data values in seconds
- **Size: FastBit indexes are modest in size compared to well-known database indexes**
  - **On average about 1/3 of data volume compared to 3-4 times in common indexes (e.g. B-trees)**

Flame Front Tracking with FastBit

Flame front identification can be specified as a query, efficiently executed for multiple timesteps with FastBit.



Cell identification

Identify all cells that satisfy user specified conditions: "600 < Temperature < 700 AND HO_2concentr. > 10^{-7}"

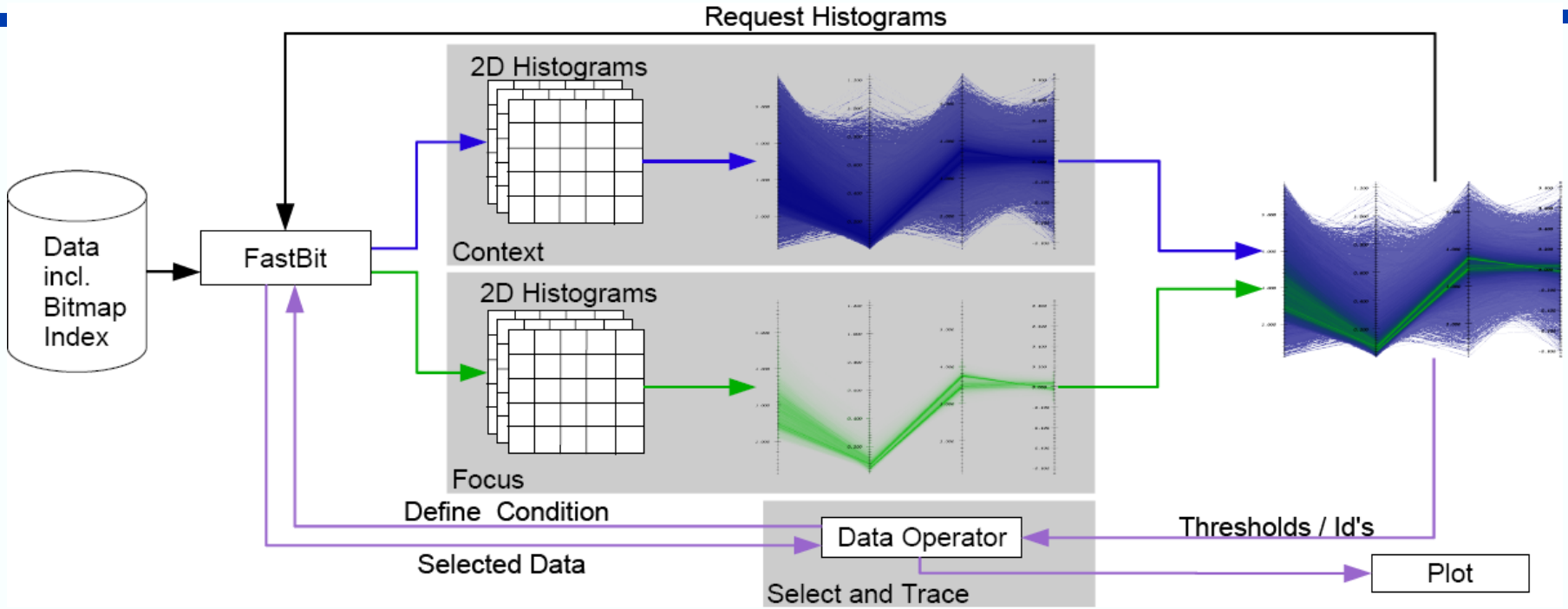
Region growing

Connect neighboring cells into regions

Region tracking

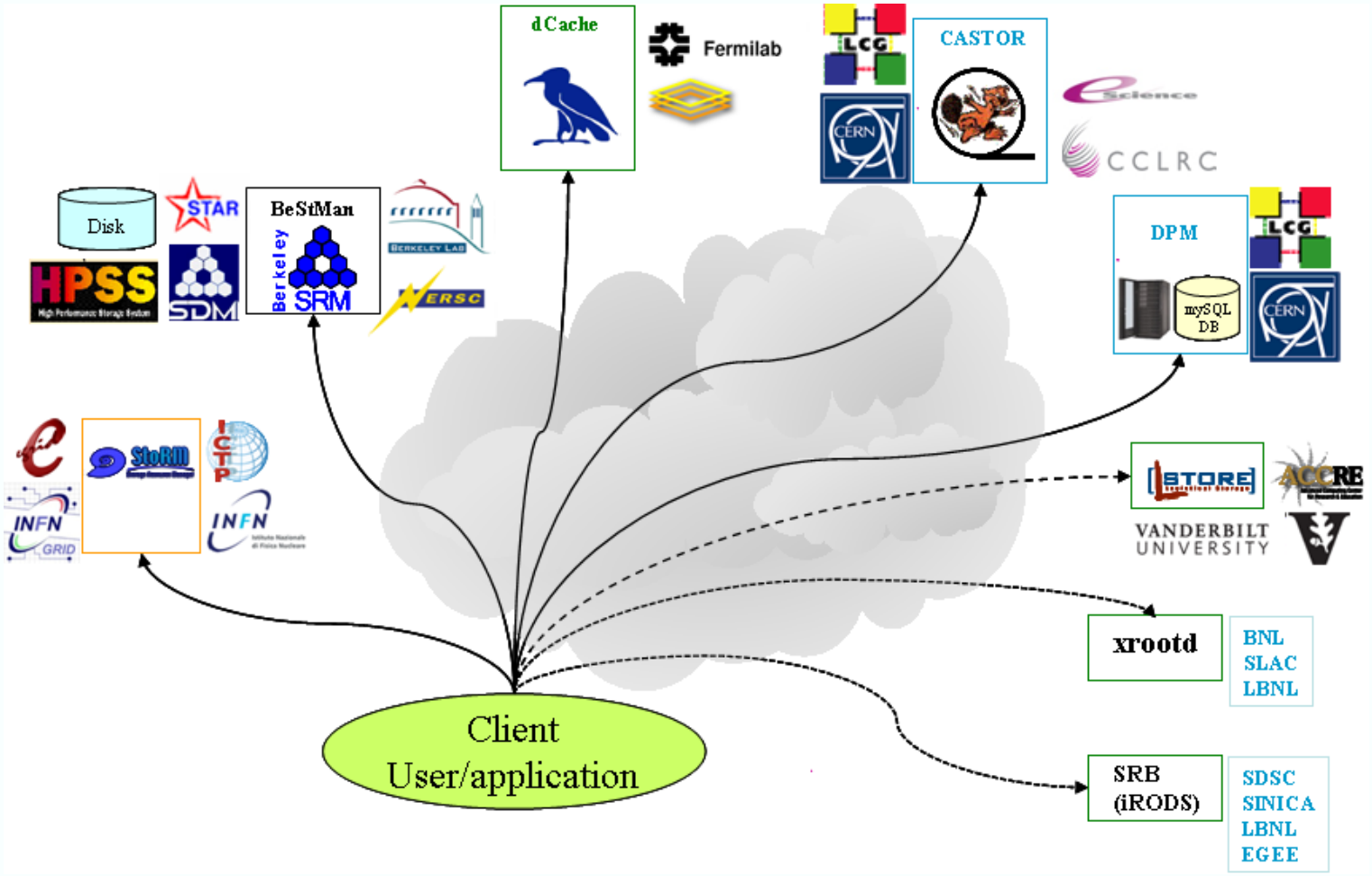
Track the evolution of the features through time

Query-Driven Visualization

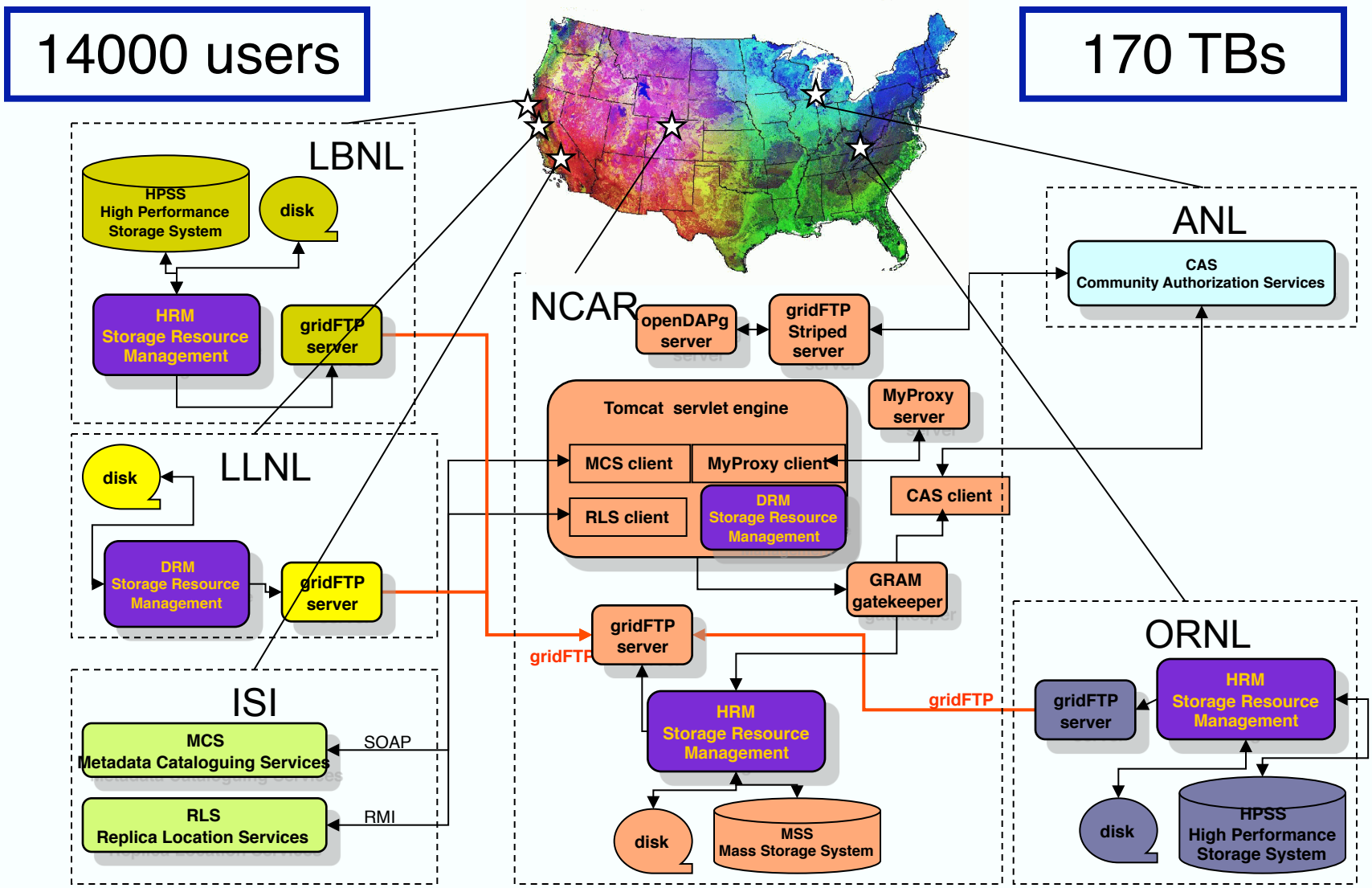


- **Collaboration between SDM and VACET centers**
  - **Use FastBit indexes to efficiently select the most interesting data for visualization**
- **Above example: laser wakefield accelerator simulation**
  - **VORPAL produces 2D and 3D simulations of particles in laser wakefield**
  - **Finding and tracking particles with large momentum is key to design the accelerator**
  - **Brute-force algorithm is quadratic (taking 5 minutes on 0.5 mil particles), FastBit time is linear in the number of results (takes 0.3 s, 1000 X speedup)**

Storage Resource Managers (SRMs): Middleware for storage interoperability and data movement



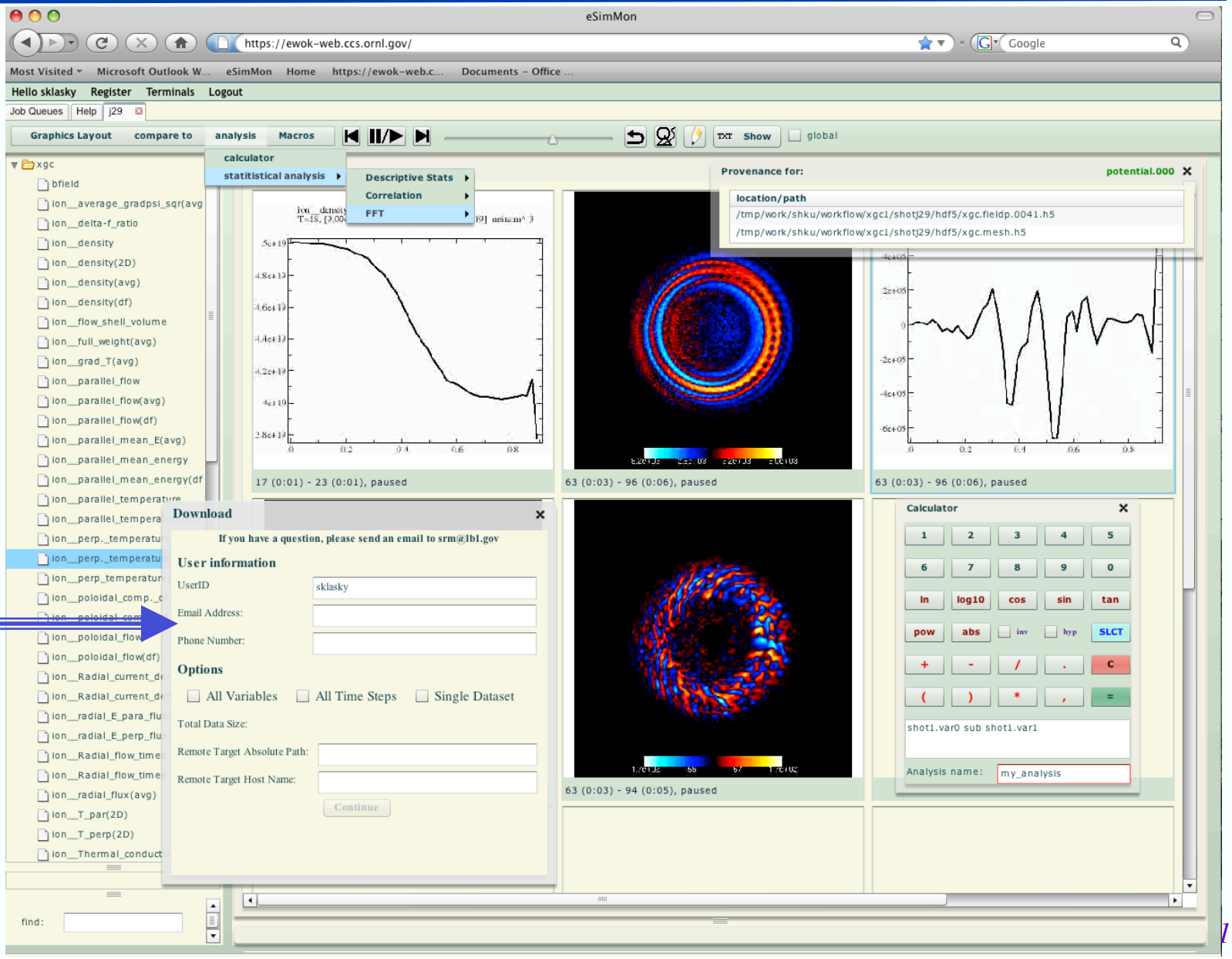
SRM use in Earth Science Grid



SDM Contact: A. Sim, A. Shoshani, LBNL

Dashboard uses provenance for finding location of files and automatic download with SRM

Download window



Implications from SDM center experience (1)

- What was successful and we believe can be done in-situ
  - monitoring of simulations
    - Use of workflow automation, and dashboard technologies
    - In center – Kepler workflow, dashboard
  - Provenance generation
    - Can be captured while data is generated by using workflow
    - In Center – provenance recorder in Kepler
  - Index generation
    - Extremely effective for post analysis, and in-situ product generation
    - In center: FastBit index
  - Summarization and various statistics
    - Shown that many functions can be parallelized
    - In center: Parallel -R

Implications from SDM center experience (2)

- What was successful and we believe can be done in-situ
  - Asynchronous I/O, and combining multiple I/O writes
    - In center: MPI I/O, PnetCDF, ADIOS (already in-situ)
  - Allow statistics to be added into files
    - Permits statistics to be carried with file
    - In center: ADIOS uses extendable BP (binary packed) file format
  - In-memory code coupling
    - To allow multiple codes to couple while running on different partitions of the machine
    - In center: Data Spaces (invoked by ADIOS)
- Exploratory data analysis (pattern discovery) requires effective data access and transfer to user's facilities
  - Will continue to need WAN robust, efficient, data movement
    - Need end-to-end bandwidth reservation (including network and storage)
    - In center: DataMover

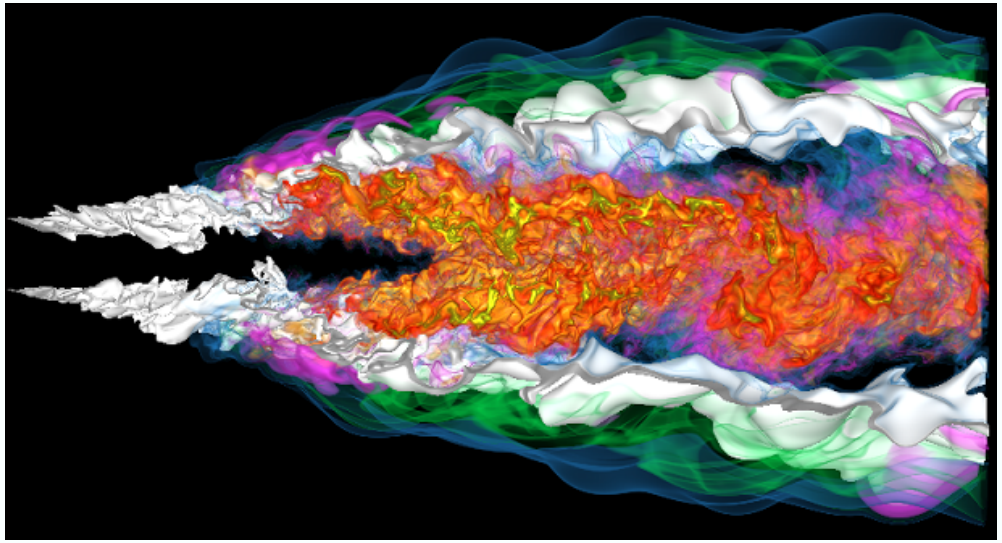
Example of In-Situ Analysis and Data Reduction

In situ analysis incorporates analysis routines into the simulation code. This technique allows analysis routines to operate on data while it is still in memory, potentially significantly reducing the I/O demands.

One way to take advantage of in situ techniques is to perform initial analysis for the purposes of data reduction. With help from the application scientist to identify features of interest, we can compress data of less interest to the scientist, reducing I/O demands during simulation and further analysis steps.

The feature of interest in this case is the mixture fraction with an iso value of 0.2 (white surface). Colored regions are a volume rendering of the HO2 variable (data courtesy J. Chen (SNL)).

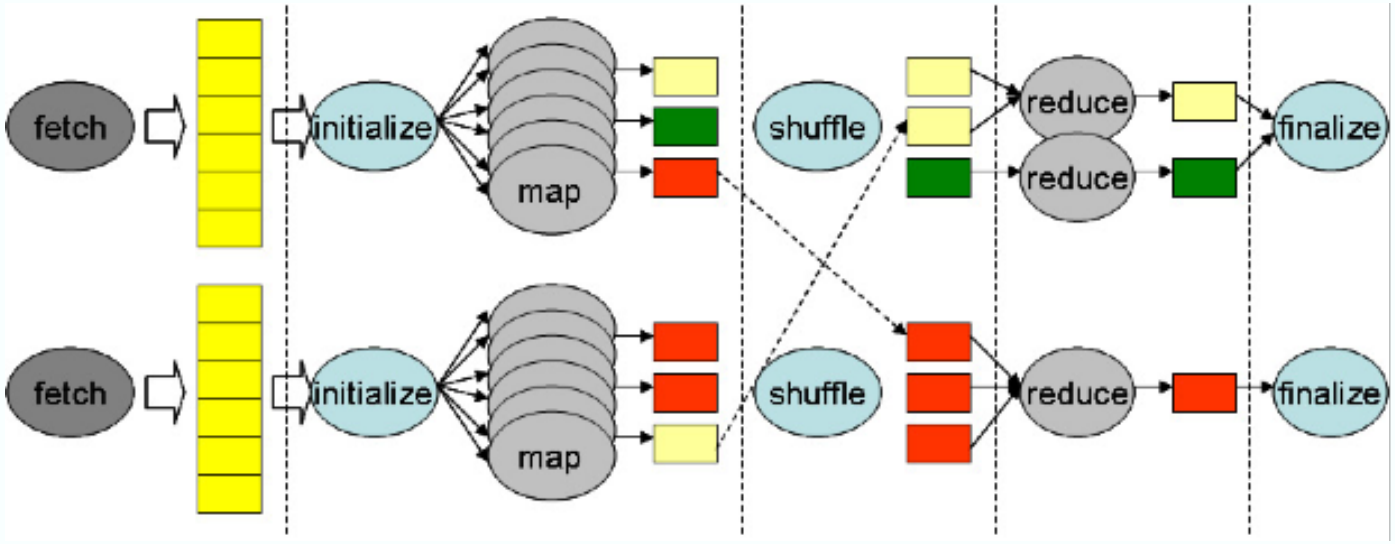
By compressing data more aggressively the further it is from this surface, we can attain a compression ratio of 20-30x while still retaining full fidelity in the vicinity of the surface.



C. Wang, H. Yu, and K.-L. Ma, "Application-driven compression for visualizing large-scale time-varying volume data", IEEE Computer Graphics and Applications, 2009.

An example of stream processing (pipeline) in the staging area

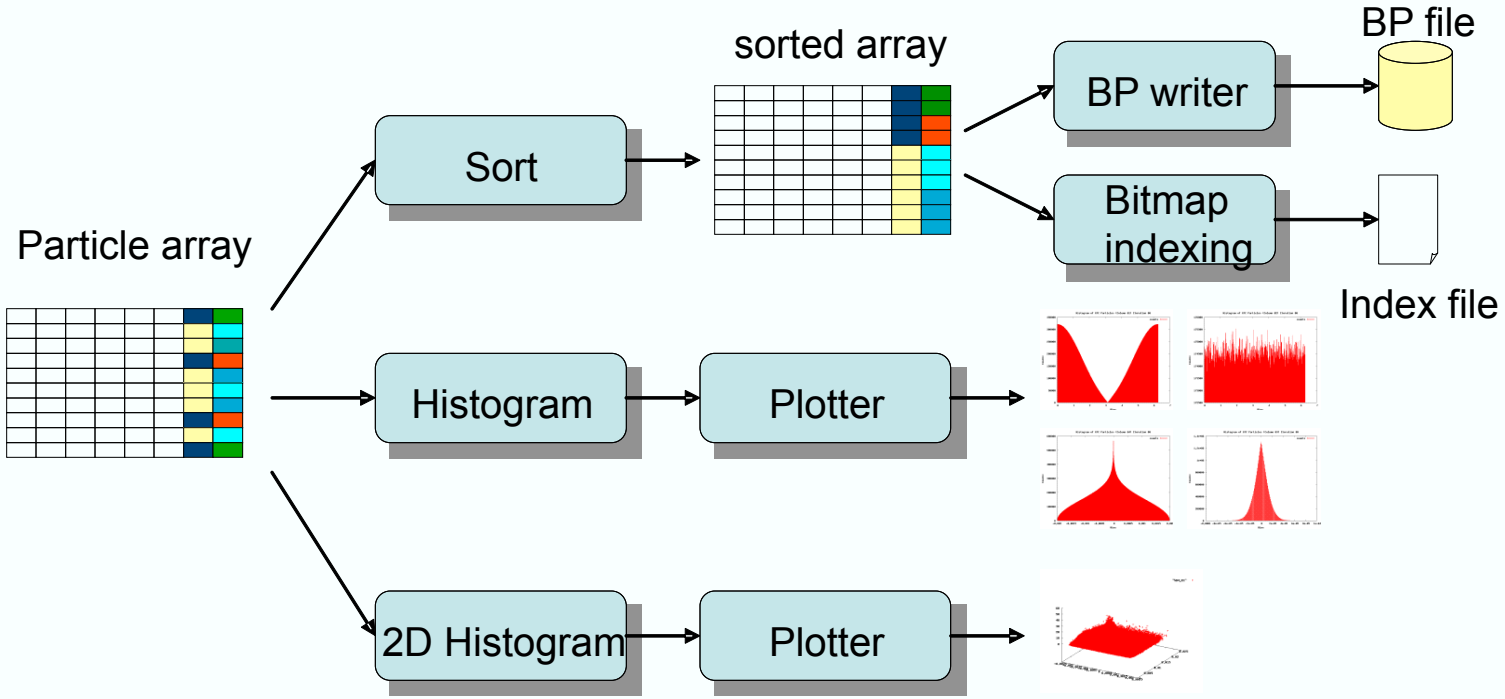
- Similar to Map-Reduce in style, but
- customized data shuffling and synchronization methods performed with highly-optimized MPI codes



From F. Zheng, H. Abbasi, C. Docan, J. Lofstead, S. Klasky, Q. Liu, M. Parashar, N. Podhorszki, K. Schwan, M. Wolf, "PreDatA - Preparatory Data Analytics on Peta-Scale Machines", IPDPS 2010.

Example of Tasks Performed at Staging Nodes

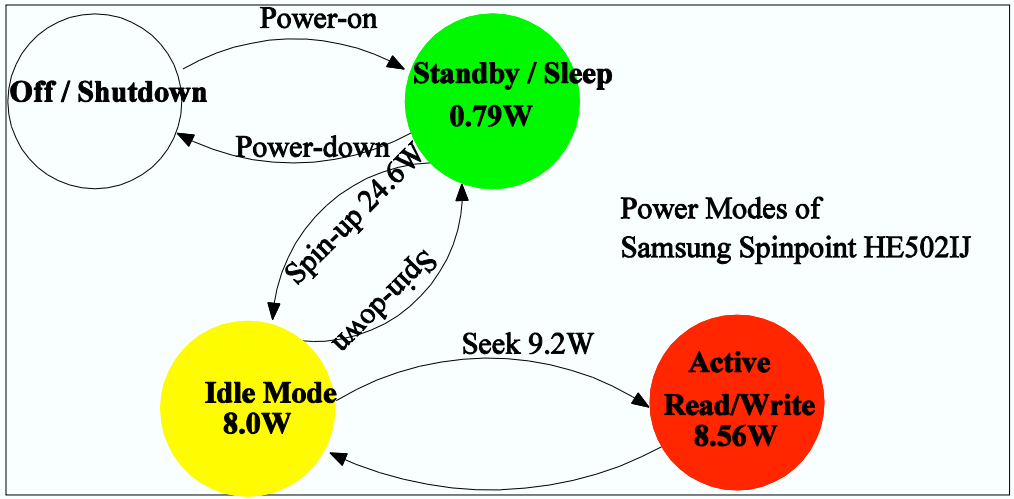
- Use the staging nodes and create a workflow in the staging nodes
- Allow the ability to generate online insights into the 260GB data being output from 16,384 compute cores in 40 seconds
- Prepare data and indexes for exploratory analysis (external to exascale machine)



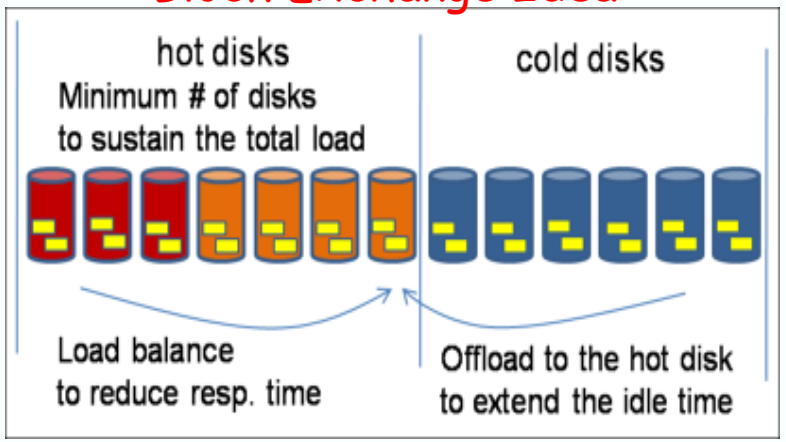
From F. Zheng, H. Abbasi, C. Docan, J. Lofstead, S. Klasky, Q. Liu, M. Parashar, N. Podhorszki, K. Schwan, M. Wolf, "PreDatA - Preparatory Data Analytics on Peta-Scale Machines", IPDPS 2010.

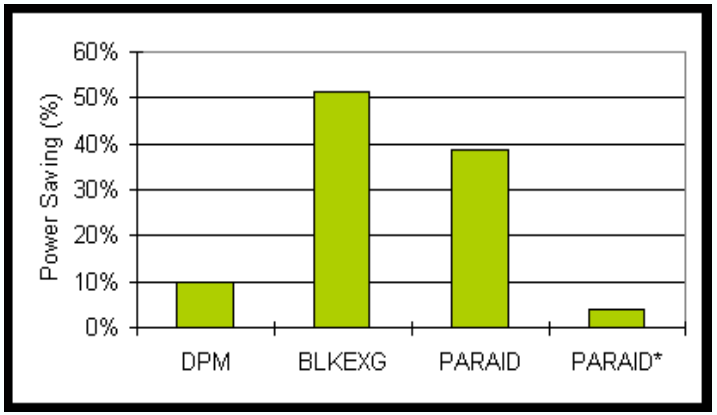
Dealing with Idle disks: Up to 50% power savings has been observed

EPA report states that power and cooling for storage represents 40% of total data centers expense

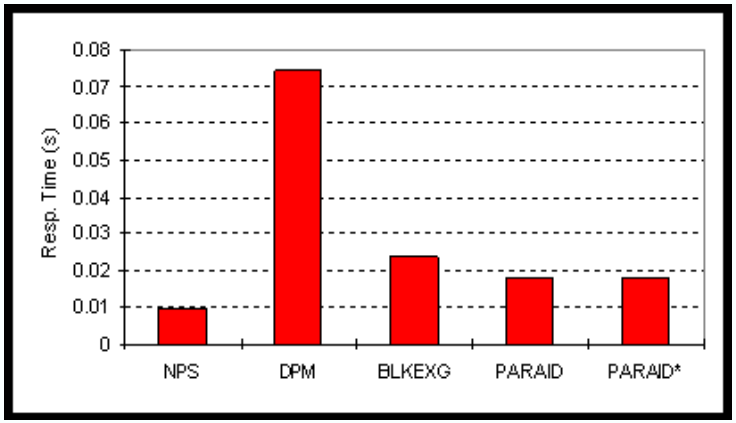


Block Exchange Idea





power



Response time

Use Case: Climate Data

- Run a large-scale simulation at NERSC
  - Produce 1-10 TBs now, 10-100 TBs in the future
  - Move to disk systems and archival storage
- In-situ data manipulations
  - Transposition, summarization, indexing
- Support requests for subsets of the data to be moved to other sites (e.g. ESG)
  - Need metadata management and search capabilities
  - Reliable data movers
- Support for analysis at NERSC
  - Move data to large or mid-size machines after data is stored
  - Support analysis/indexing tools
  - Support workflow and display tools (e.g. dashboards)
  - Model verification: compare to observational data

The END