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

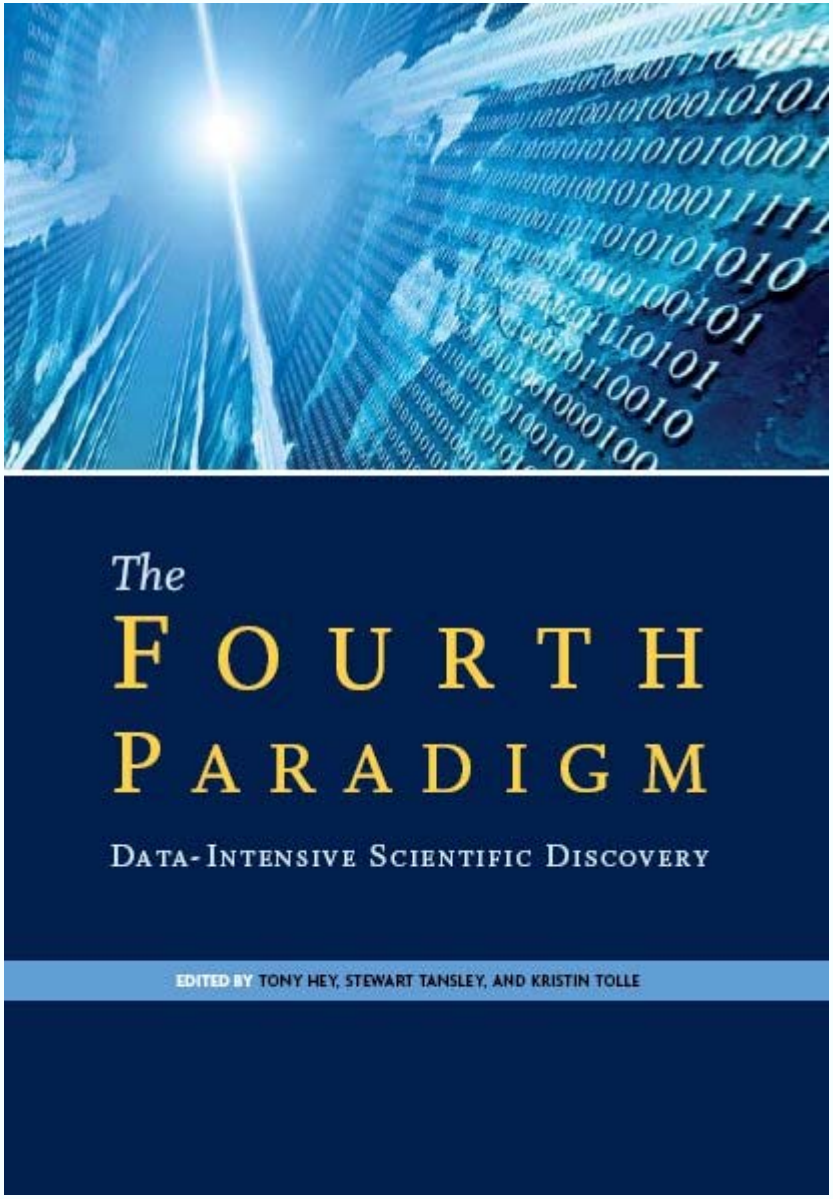
Arie Shoshani, John Wu (LBNL)

and the SDM center

October 26, 2010

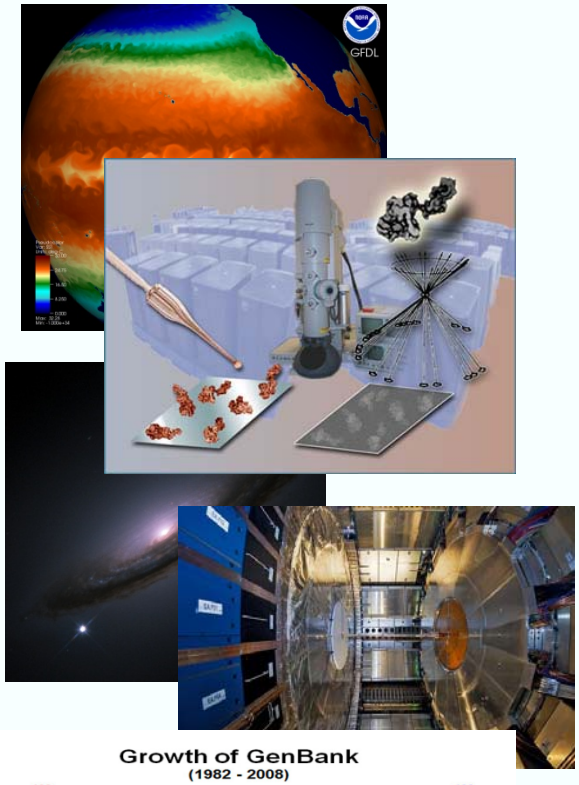
The 4th Paradigm?

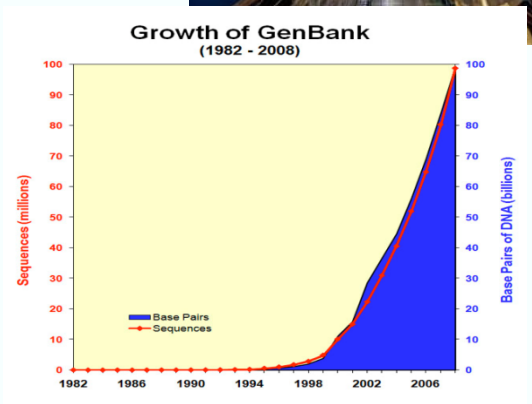
- Maybe
- The reality is that data is becoming a much larger part of science and human society in general



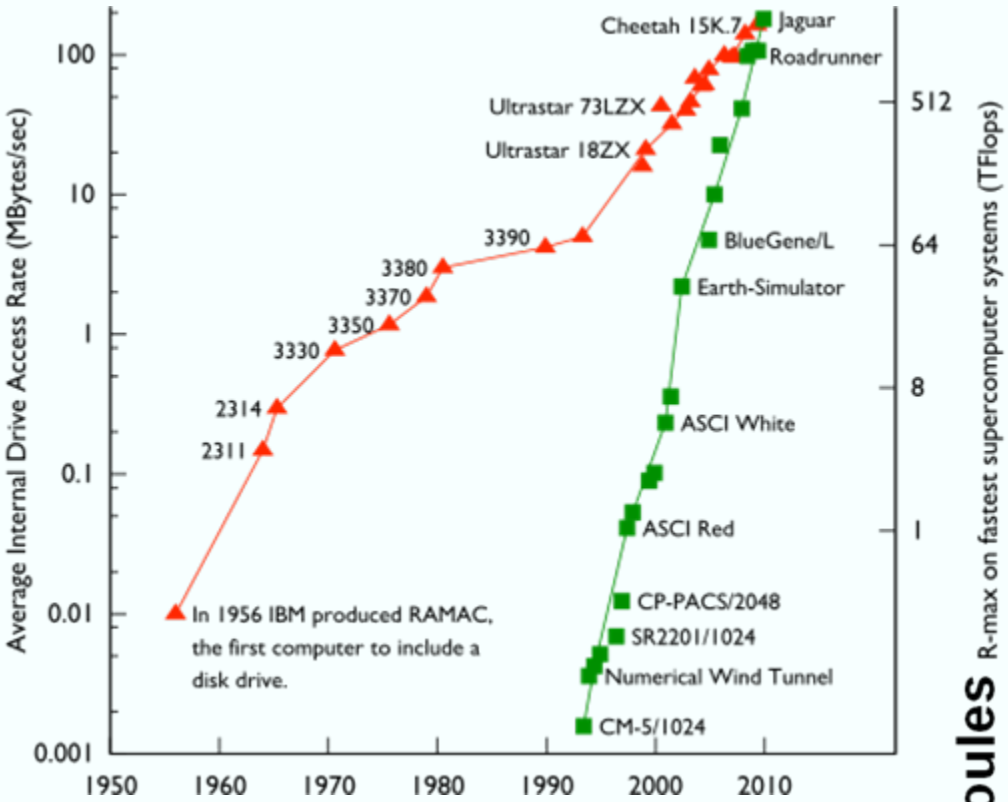
Sciences Becoming Data Intensive

- Existing ability to generate science data is already challenging our ability to store, analyze, and archive.
  - Some observational devices grow in capability with Moore's Law.
  - Data sizes are growing exponentially.
- Petabyte (PB) data sets are becoming common:
  - *Climate:* next IPCC estimates 10s of PBs
  - *Genome:* JGI alone will have about 1 PB this year and double each year
  - *Particle physics*: LHC projects 16 PB / yr
  - *Astrophysics*: LSST, others, estimate 5 PB / yr
- Exascale HPC simulations will lead to exascale datasets



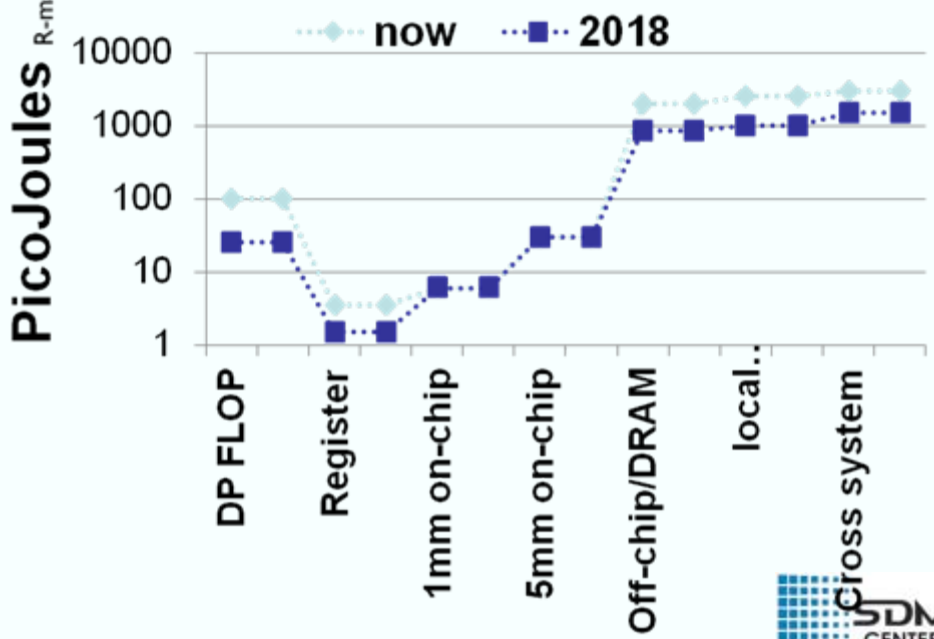


I/O Performance Challenges



Performance Crisis #1

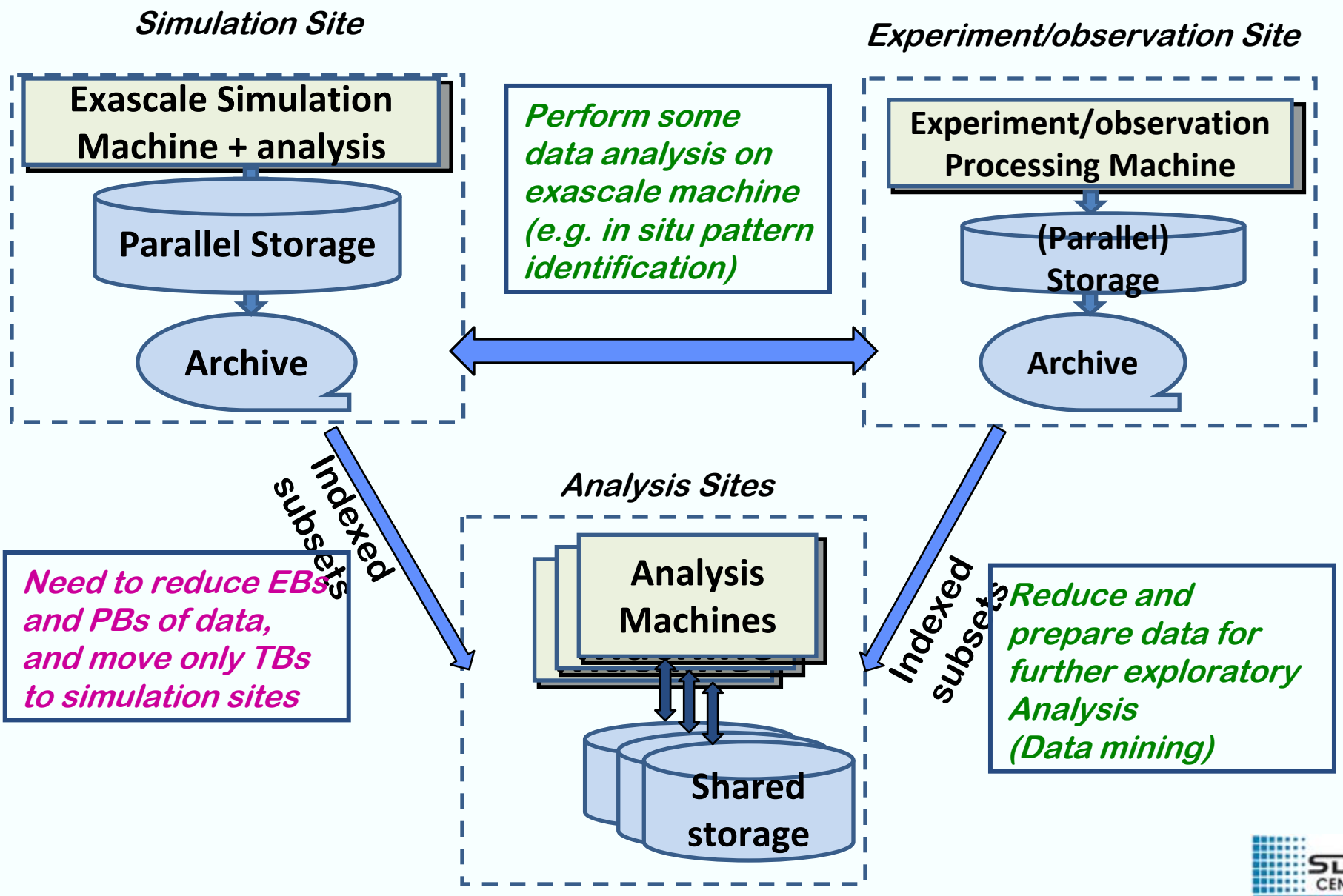
- ***Disk speed lags compute speed.***
- ***To achieve reasonable aggregate bandwidth many spindles needed – 10^3 spindles = 1PB but only ~0.1TB/s !***



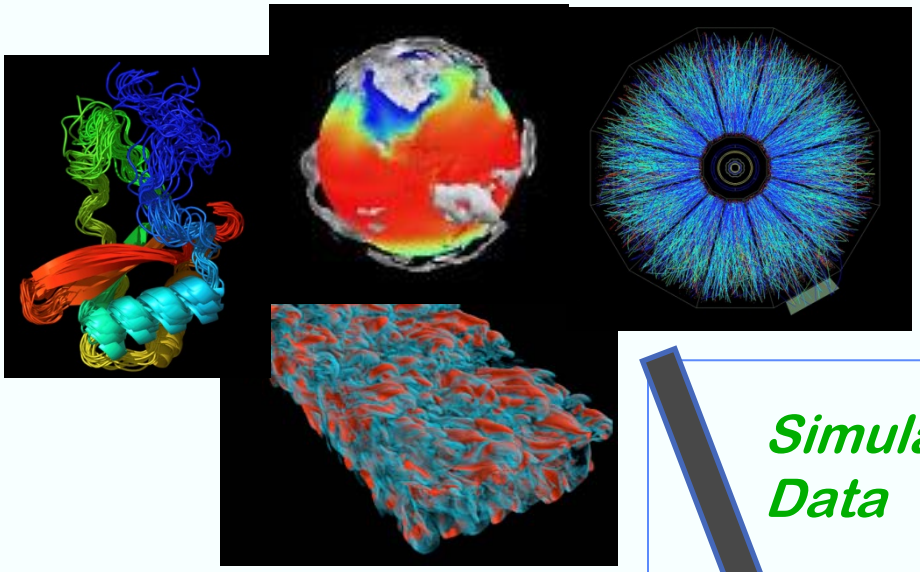
Performance Crisis #2

Moving data on an exascale machine will be expensive – both in terms of electricity & time!

Data Analysis Challenges: Complex Process

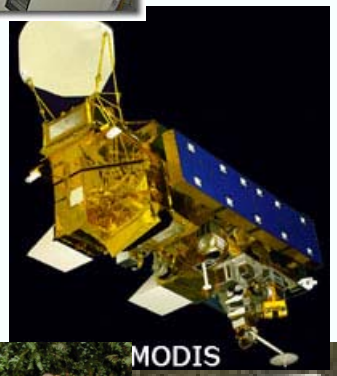


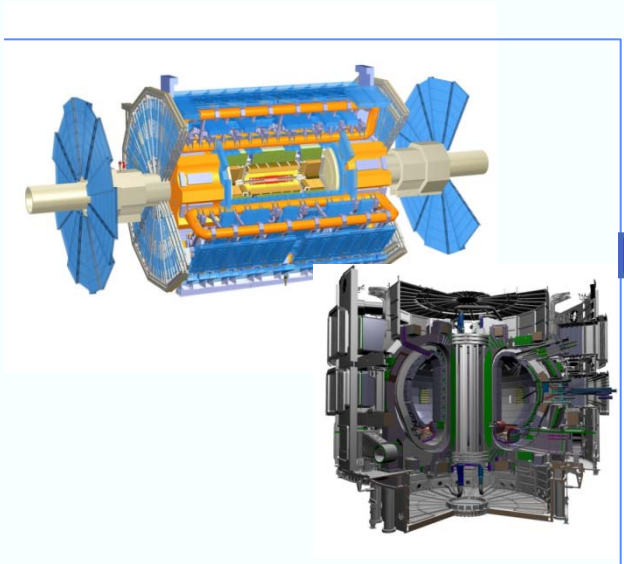
Data Analysis Challenges: Integration
Many sources, Different formats, Varying volumes



Simulation Data

Observation Data





Dataset

Experiment Data

SDM CENTER

Data Analysis Challenges: Verification

- Simulations now being used to make major decisions, not just study phenomena
  - Climate model CCSM 100 year predictions
  - ASCEM – Subsurface simulations for environmental management (DOE-EM)
  - CCSI – Carbon Capture simulation initiative
  - Simulations of combustion for new fuels and burners
- Understanding uncertainty is critical
  - statistical methods for initial data space
  - generates large amounts of data
- Versioning and provenance of results
  - make results traceable and repeatable

SDM CENTER

Data Movement on the Exascale

- Users often need help moving large datasets
  - Five requests recently to move 30+TB sized data sets between NERSC and other facilities.
- Network connections are hampered by firewalls, improperly configurations, and unexpected bottlenecks
  - ~10 MB/sec if it worked
  - More than 30 days to transfer data
- After network tuning, new dedicated tuned systems, file system, client testing, and better management tool (ANL, ORNL and NERSC DTNs), 200+MB/sec
  - Just over 2 days to transfer data
- On the exascale
  - Better management software on intelligent network e.g., ESNet/ANI
  - Move as little data as possible

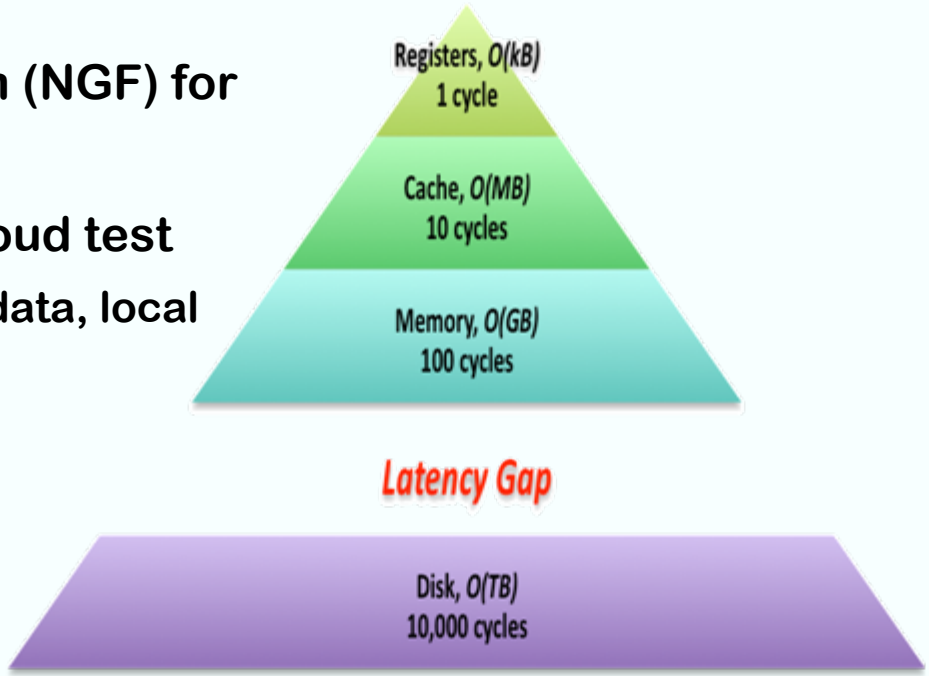
Additional Data Ecosystem Challenges

- Data Analysis
  - Collaborative data validation and verification
  - Data processing to develop an enhanced dataset
  - Support for new analysis capabilities
  - Search based on data properties
- Data Management
  - Ontology development
  - Connecting publications resulting from data analysis to the underlying data
- Tools
  - Generate and process/analyze subsets
  - Visualization support (e.g. data standards…)

SDM CENTER

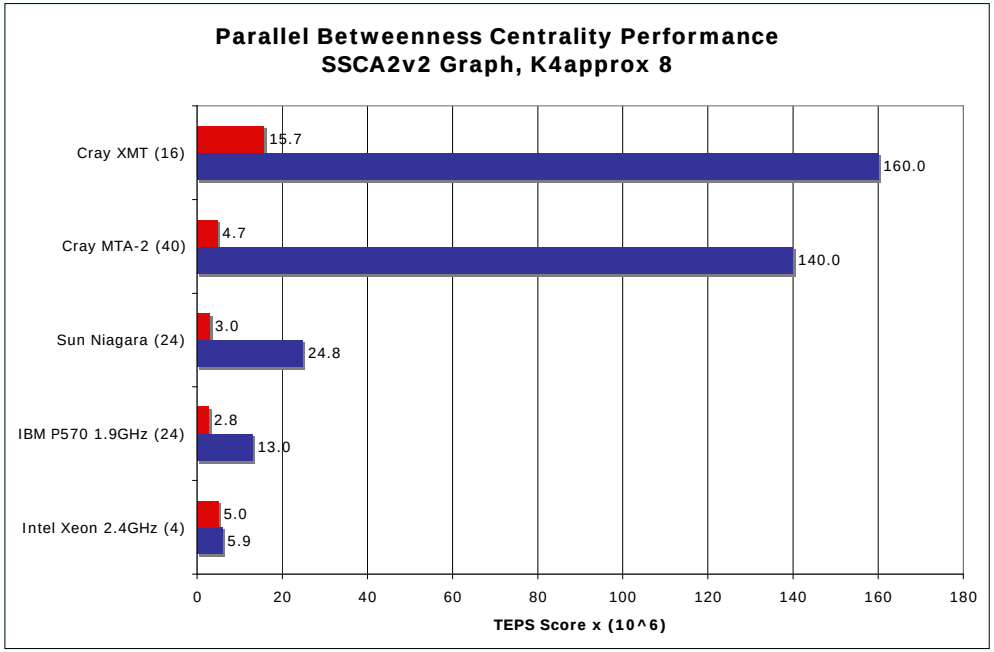
Possible Hardware Solutions: Flash Storage

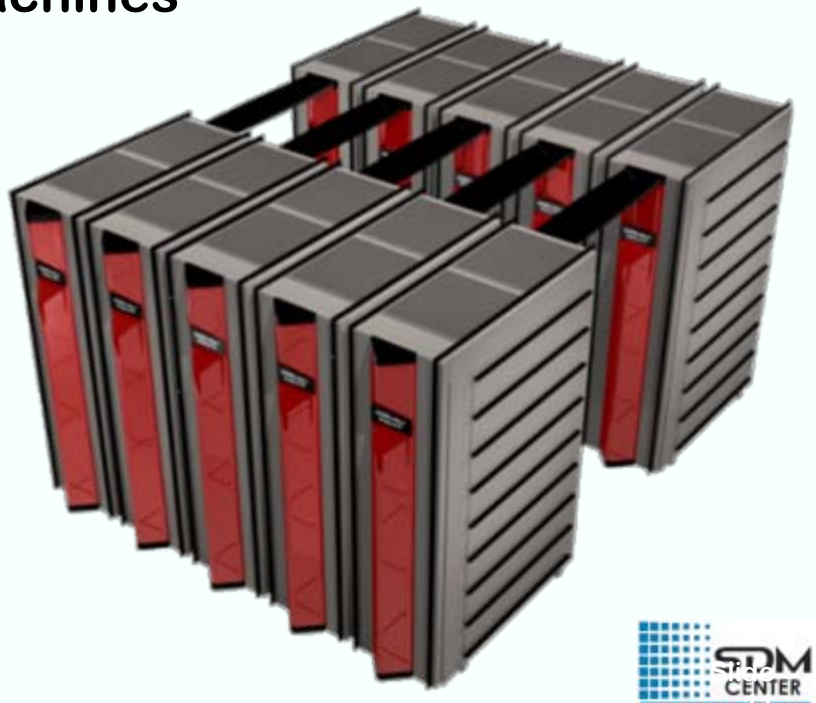
- Memory per Flop decreasing
- Solid State Storage could fill the latency gap
  - High bandwidth, low-latency
- Flash Storage Testbeds
  - ~ 10TB in NERSC Global File system (NGF) for metadata acceleration
  - 16TB as local SSD in "Magellan" cloud test bed for data analytics, local read-only data, local temp storage



Possible Hardware Solutions: Analytics Machines

- Architectures for data analytics
  - With government support, Cray developed the eXtreme MultiThreading (XMT) system and technology to solve unstructured data analysis problems
  - SGI Ultraviolet is latest in a popular line of "distributed shared memory" machines





Possible Software Technologies from the SDM center

http://sdmcenter.lbl.gov

Arie Shoshani (PI)

Co-Principal Investigators

DOE Laboratories

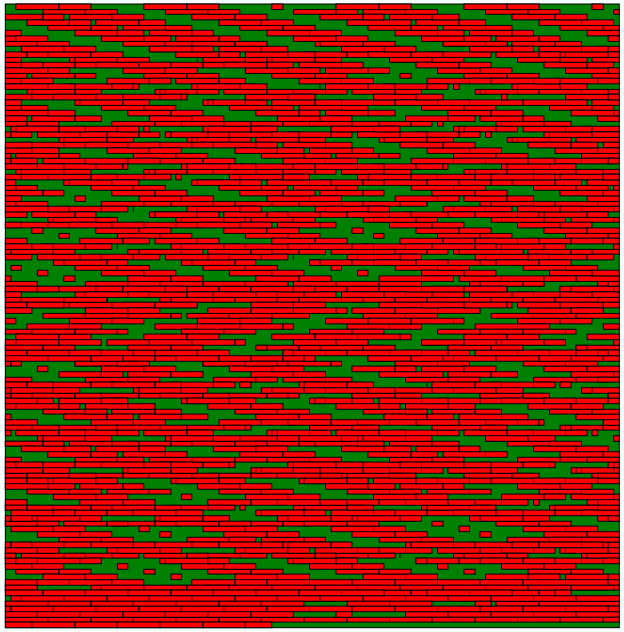
ANL: Rob Ross
LBNL: Doron Rotem
LLNL: Chandrika Kamath
ORNL: Nagiza Samatova
PNNL: Terence Critchlow

Universities

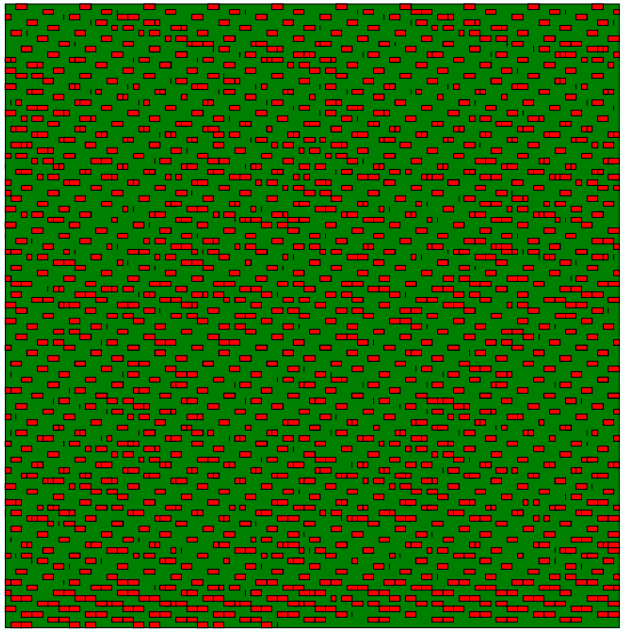
NCSU: Mladen Vouk
NWU: Alok Choudhary
UCD: Bertram Ludaescher
SDSC: Ilkay Altintas
UUtah: Claudio Silva

Visualizing and Tuning I/O Access

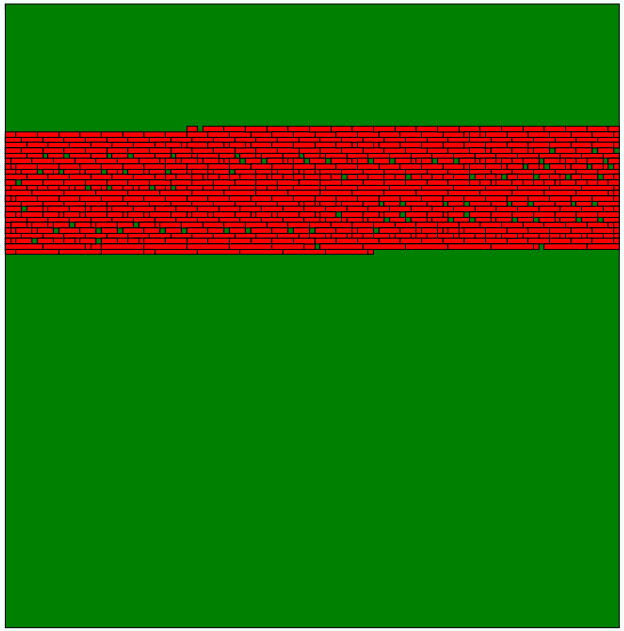
This view shows the entire 28 Gbyte dataset as a 2D array of blocks, for three separate runs. Renderer is visualizing one variable out of five. Red blocks were accessed. Access times in parenthesis.



Original Pattern



MPI-IO Tuning



PnetCDF Enhancements

Data is stored in the netCDF "record" format, where variables are interleaved in file (36.0 sec). Adjusting MPI-IO parameters (right) resulted in significant I/O reduction (18.9 sec).

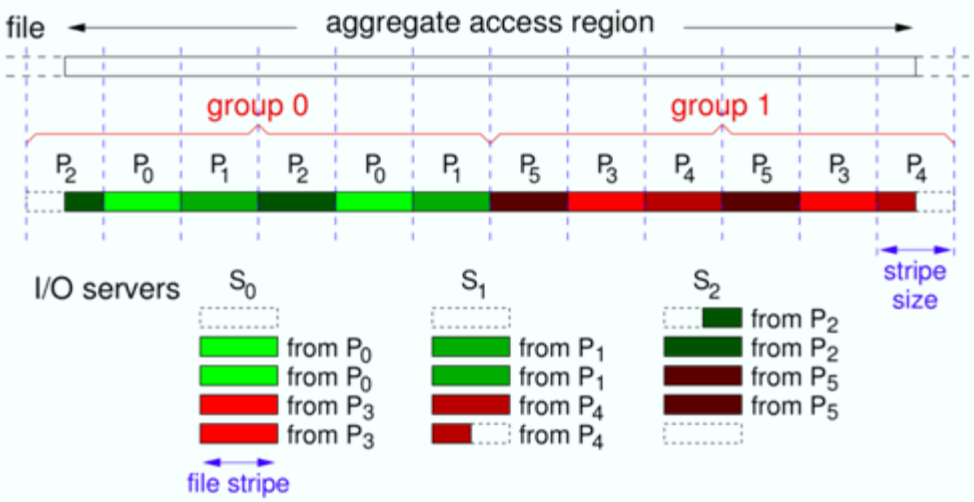
New PnetCDF large variable support stores data contiguously (13.1 sec).

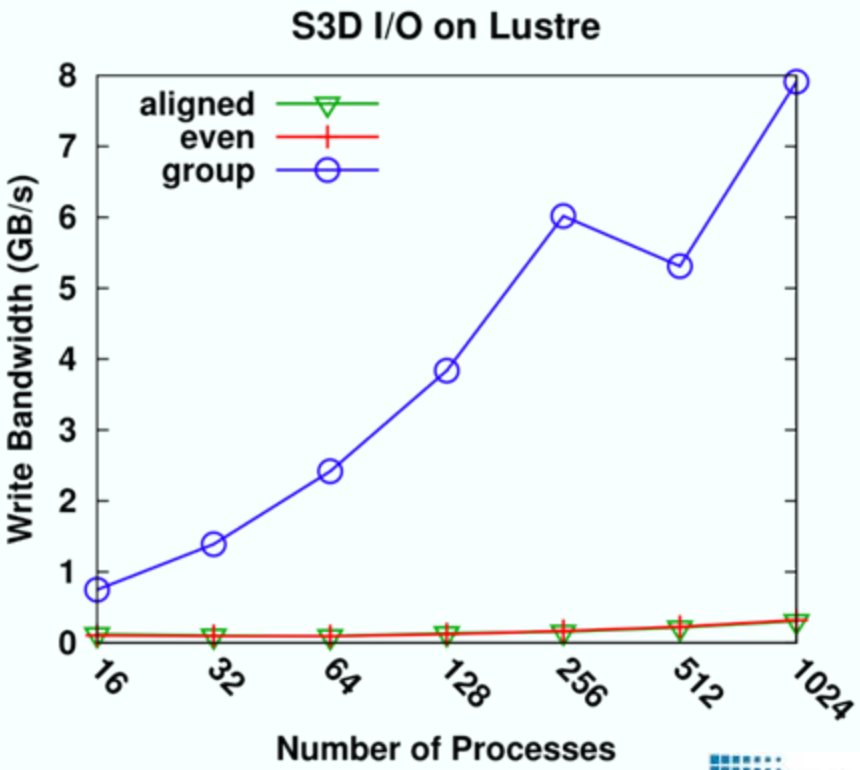
SDM CENTER

Collective I/O and Distributed Locks

Group-cyclic partitioning is an advanced technique for situations where many locks must be obtained during a single I/O operation (e.g. Lustre). Regions of the file are statically assigned to aggregators in a round-robin fashion, and aggregators are placed in groups of N, where N is the number of servers, minimizing number of extent locks requested.

Performance is many times that of "even" partitioning.





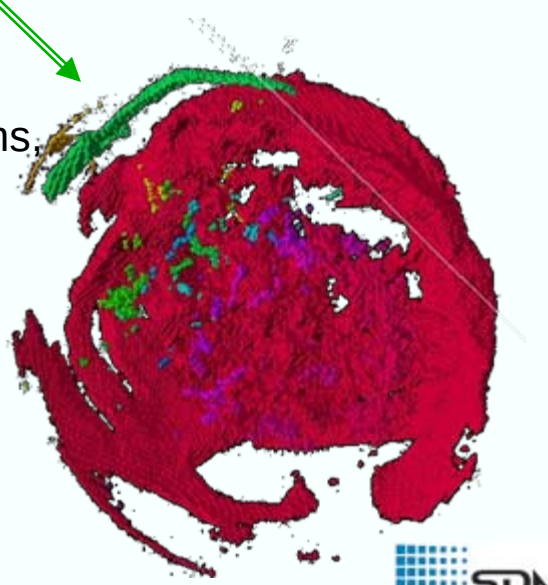
Searching Problems in Data Intensive Sciences

- 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

- FastBit is an extremely efficient compressed bitmap indexing technology
  - Indexes and stores each column separately
  - Uses a compute-friendly compression techniques (patent 2004)
  - 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)

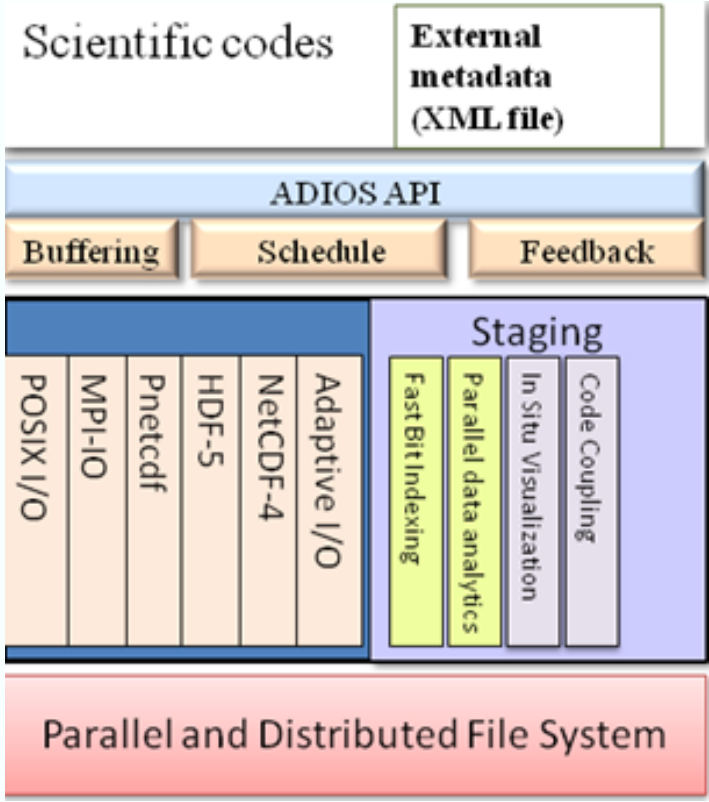
Query-Driven Visualization

- **Collaboration between SDM and VACET centers**
  - **Use FastBit indexes to efficiently select the most interesting data for visualization**
- **Above example: laser wakefield particle 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)**

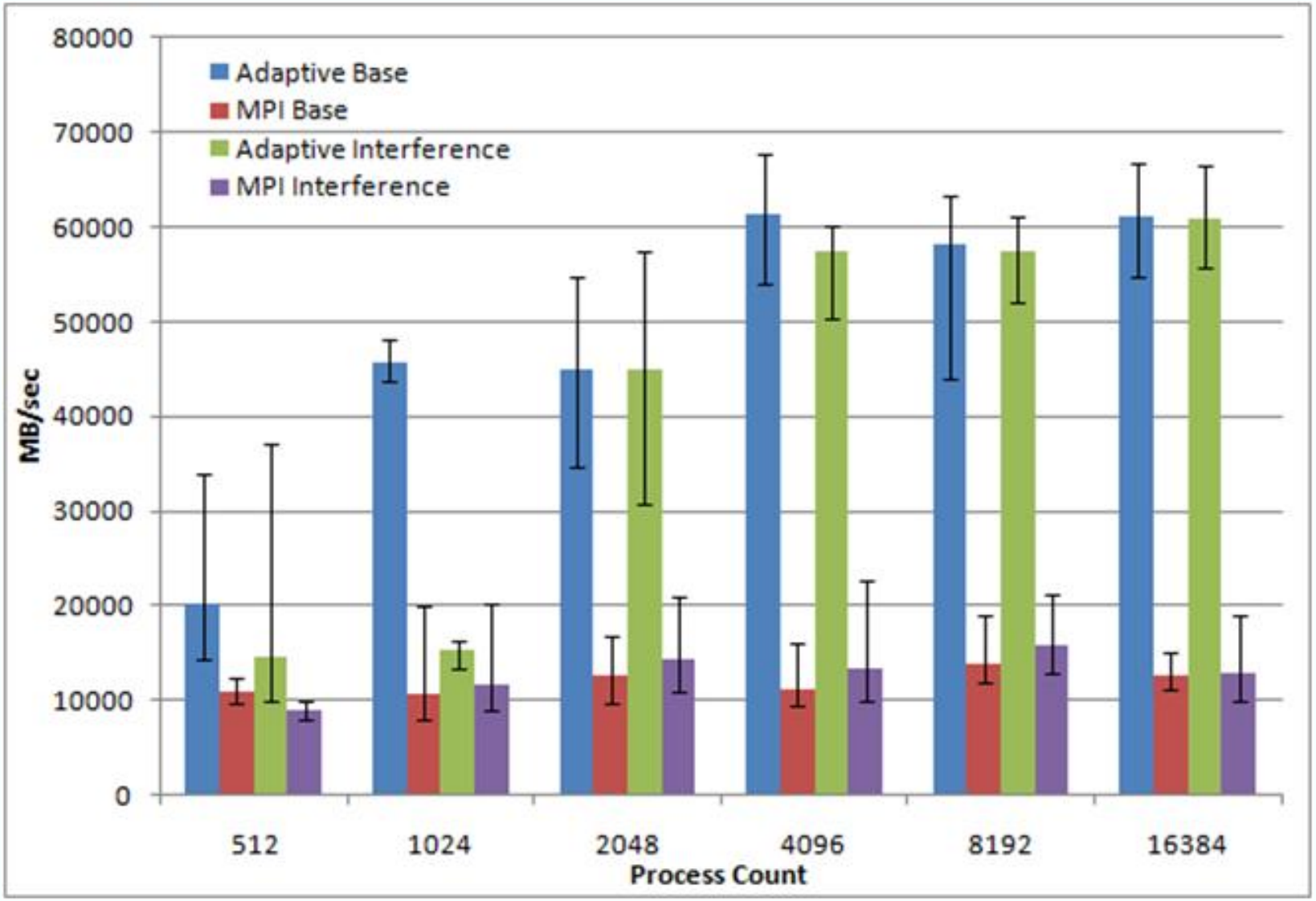
ADaptable IO System (ADIOS)

The goal of ADIOS is to create an easy and efficient I/O interface to hide the details from computational science applications:

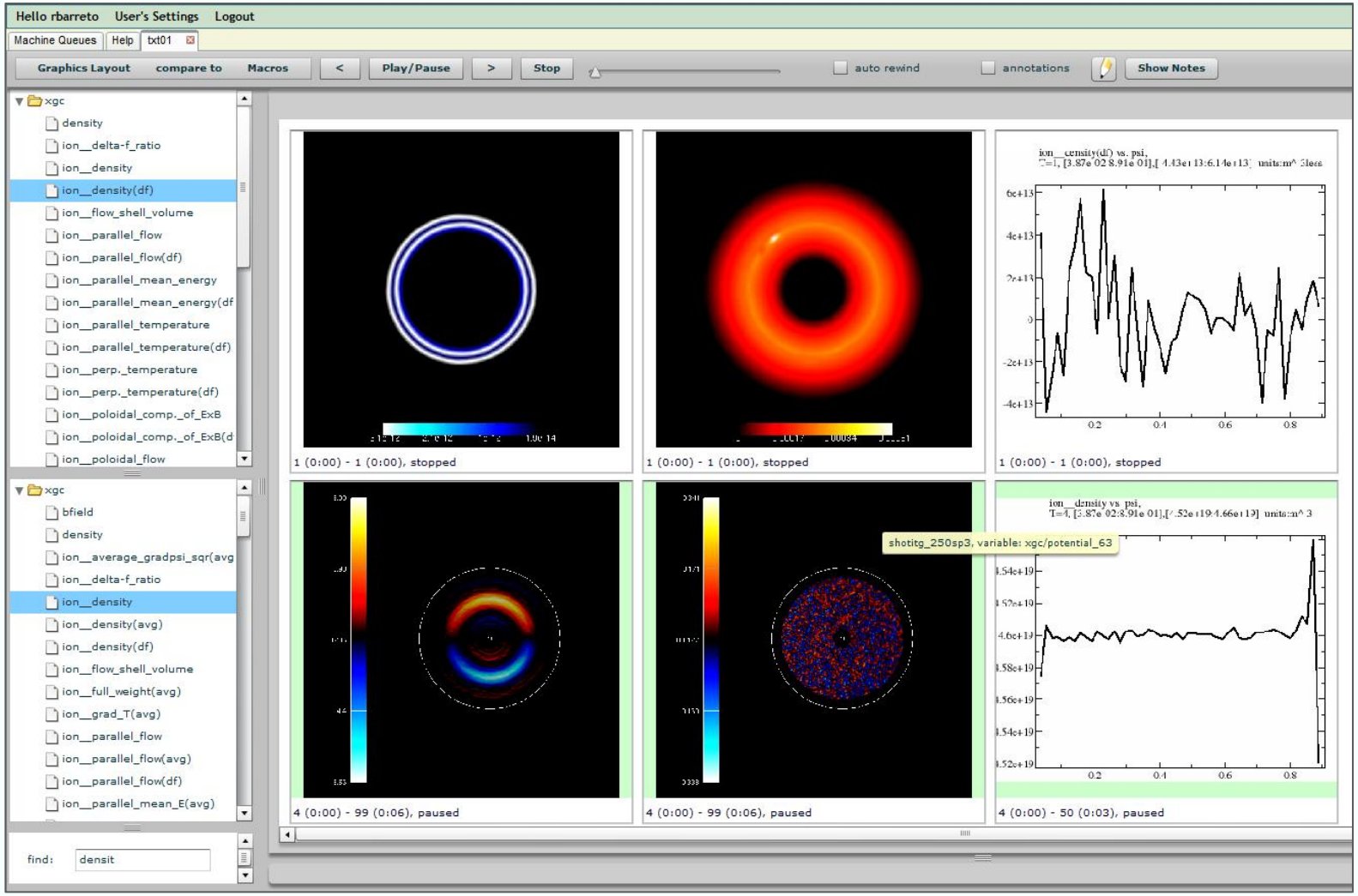
- Provide portable, fast, scalable, easy-to-use, metadata rich output
  - Change I/O method by changing XML file only
  - Allows plug-ins for different I/O implementations
- Operate across multiple architectures and file systems
  - Blue Gene, Cray, IB-based clusters
  - Lustre, PVFS2, GPFS, Panasas, PNFS
- Support many underlying file formats and interfaces
  - MPI-IO, POSIX, HDF5, netCDF, BP (binary-packed)
  - Facilitate switching underlying file formats to reach performance goals
- Compensate for inefficiencies in the current I/O infrastructures



Performance of I/O from the Pixe3D (MHD) fusion code using ADIOS



Real-time Visualization and Analysis with a Dashboard



visualize and compare shots

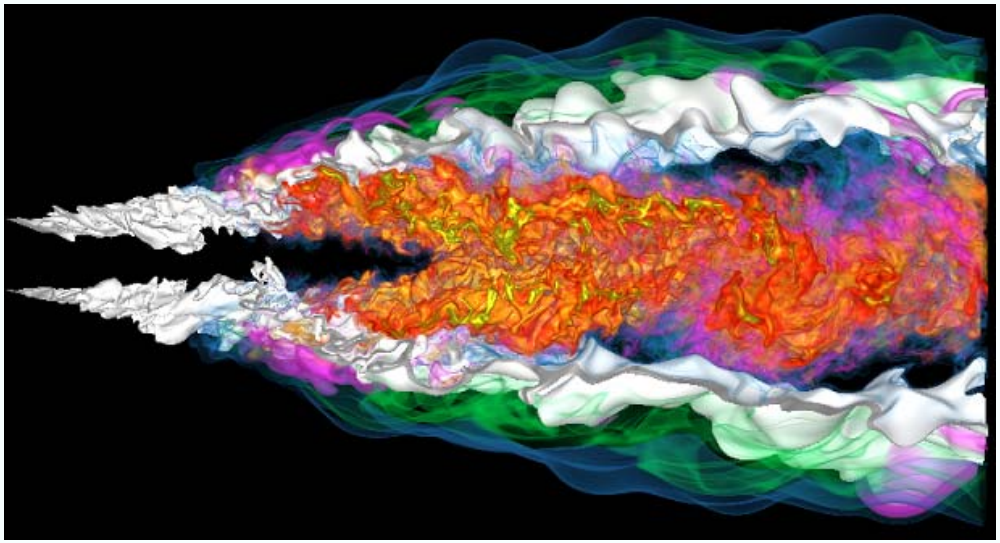
Example of In-Situ Analysis and Data Reduction

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)).

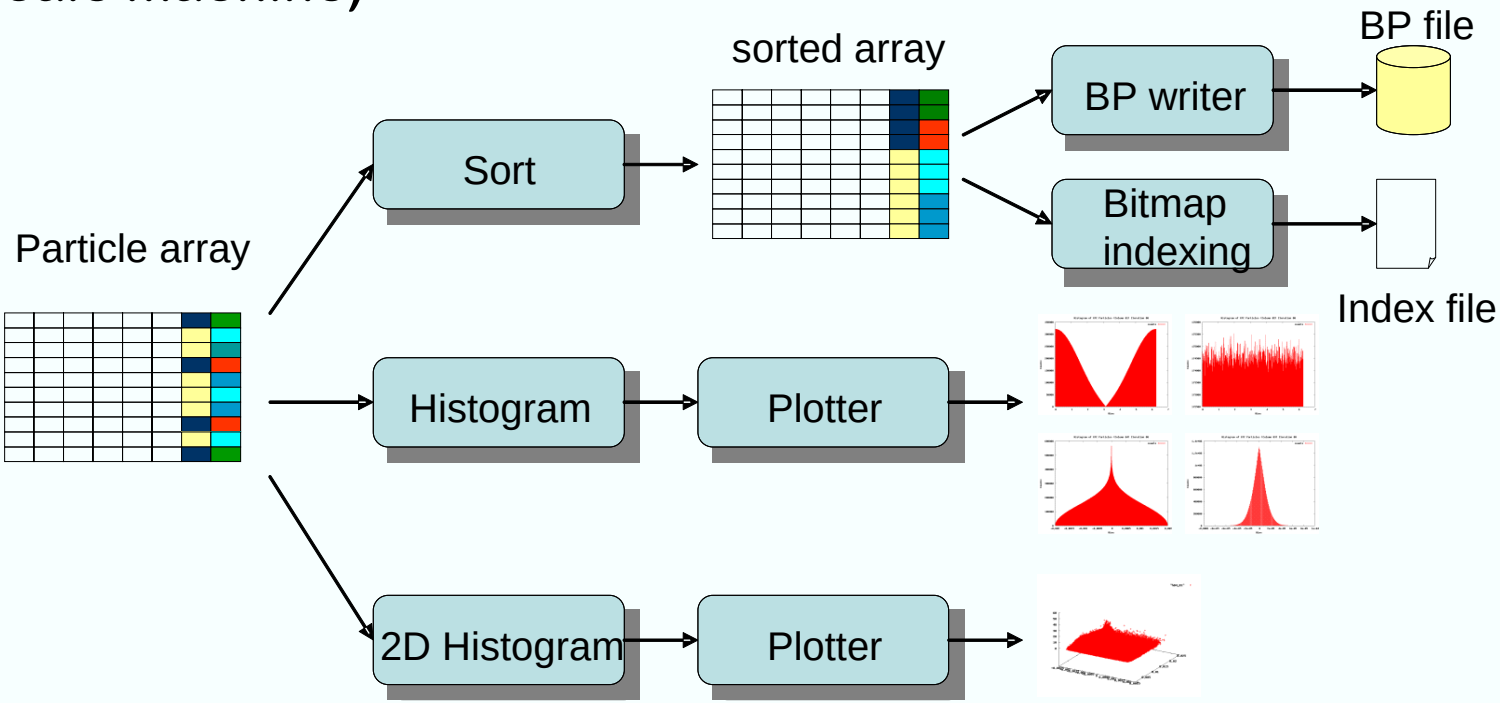
By compressing data more aggressively the further it is from this surface, we can attain a compression ratio of 20-30x while still retaining full fidelity in the vicinity of the surface.



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.

In Situ Analysis through Data Staging

- 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.

Examples of Other Technologies

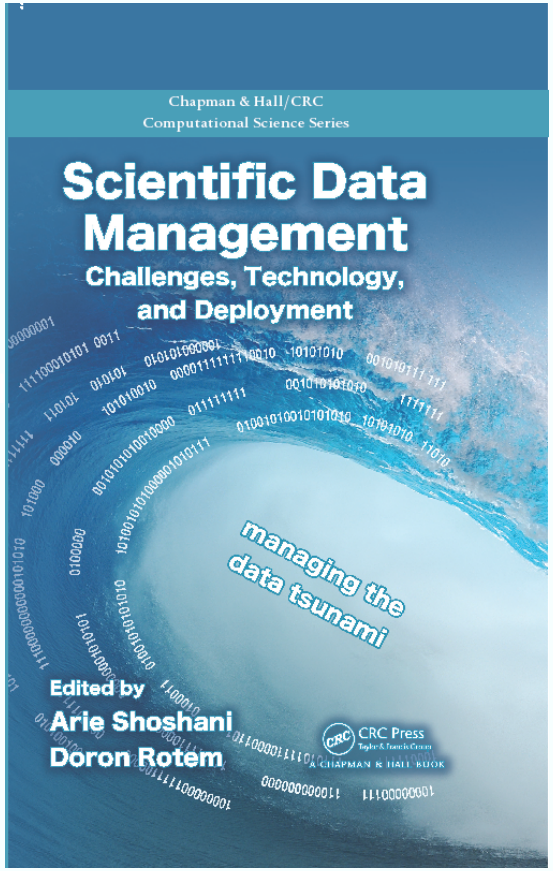
- New computational data processing paradigms
  - Hadoop, MapReduce, etc
  - HDFS
- Data-as-a-Service for end users
  - Enable interactivity
- Provide consistent computational environment
  - Support rare but large computing load
  - Allow use of out of date OS/compilers (avoid re-validation of results)
- …

SDM CENTER

Summary: Combining Software and Hardware into a Data-Side Analysis Facility

- Impractical to move exabytes to end users
  - Minimize data movement
  - Employ specialized data analysis systems
- Data-side analysis facility (exascale workshops)
  - Is near the data generation site
  - Have specialized hardware, e.g., John Hopkins' Amdahl machine
  - Have parallel analysis and visualization tools
  - Have workflow tools to compose "analysis pipelines" by users
  - Reuse previously composed pipelines
  - Have data analysis experts
  - Provide training on data analysis techniques and tools

SUBTITLE HERE



Questions?

http://sdmcenter.lbl.gov