Scientific Data Management Challenges in Extreme Scale Systems

Arie Shoshani

Lawrence Berkeley National Laboratory

Salishan Meeting, April 27-29, 2010

Outline

- Overview of successful technologies in the SDM center
  - High Performance Technologies
  - Usability and effectiveness
  - Enabling Data Understanding
- Data challenges in the extreme scale
  - Minimize volume of data to be stored
  - Prepare data for analysis before storing
  - Reduce energy
- Implications from SDM center experience
  - Techniques that could be adapted to extreme scale

The Scientific Data Management Center

http://sdmcenter.lbl.gov

Arie Shoshani (PI)

Lawrence Berkeley National Laboratory

Co-Principal Investigators

DOE Laboratories

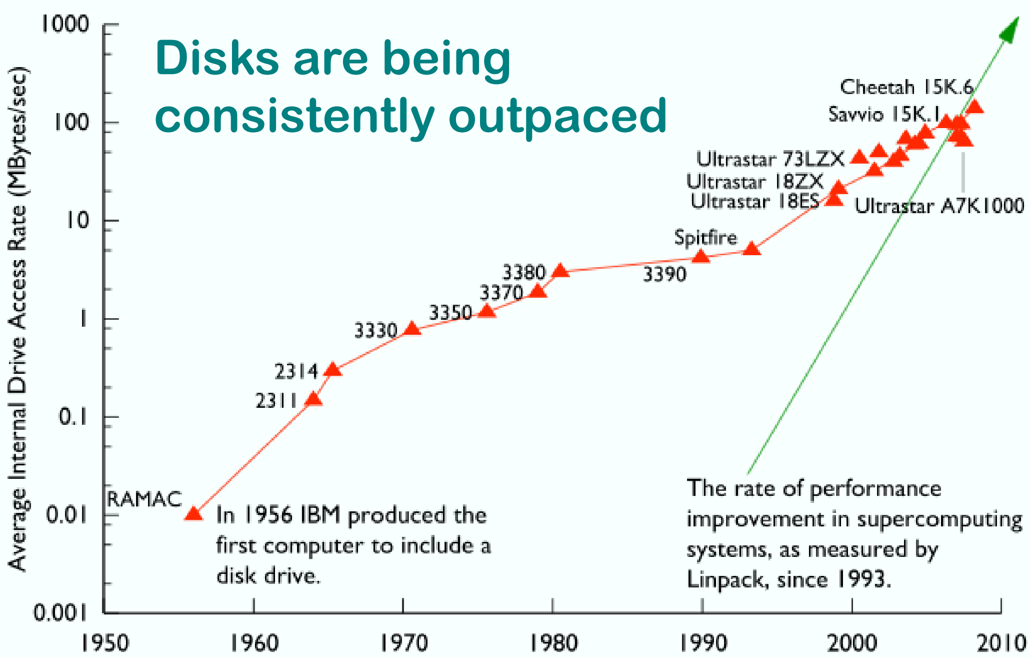
- ANL: Rob Ross
- LBNL: Doron Rotem
- LLNL: Chandrika Kamath
- ORNL: Nagiza Samatova
- PNNL: Terence Critchlow

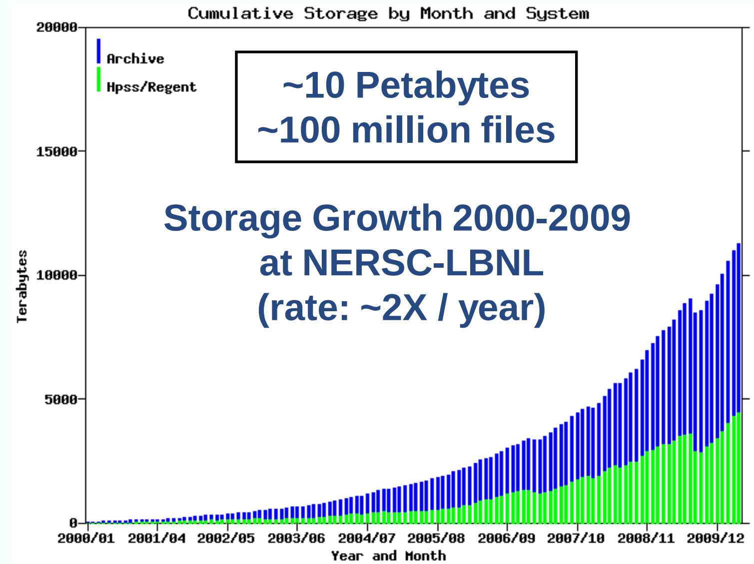
Universities

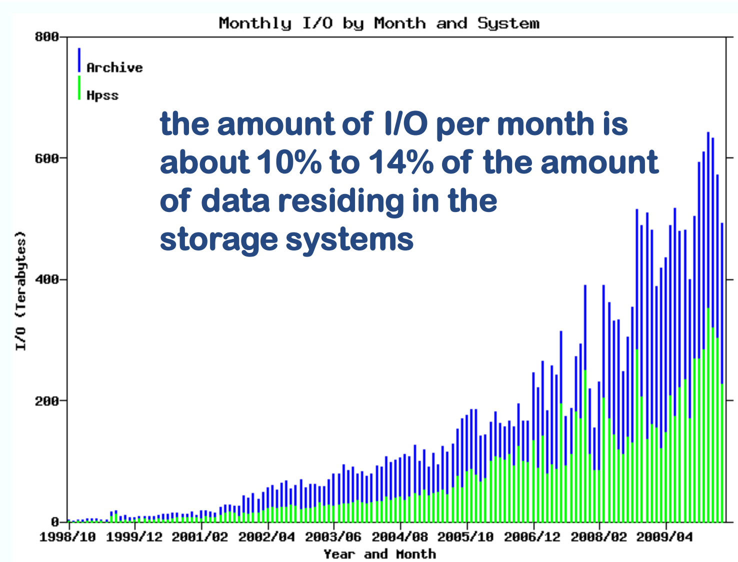
- NCSU: Mladen Vouk
- NWU: Alok Choudhary
- UCD: Bertram Ludaescher
- SDSC: Ilkay Altintas
- UUtah: Claudio Silva

Storage and I/O Trends

Scientific data management is a collection of methods, algorithms and software that enables efficient capturing, storing, moving, and analysis of scientific data.



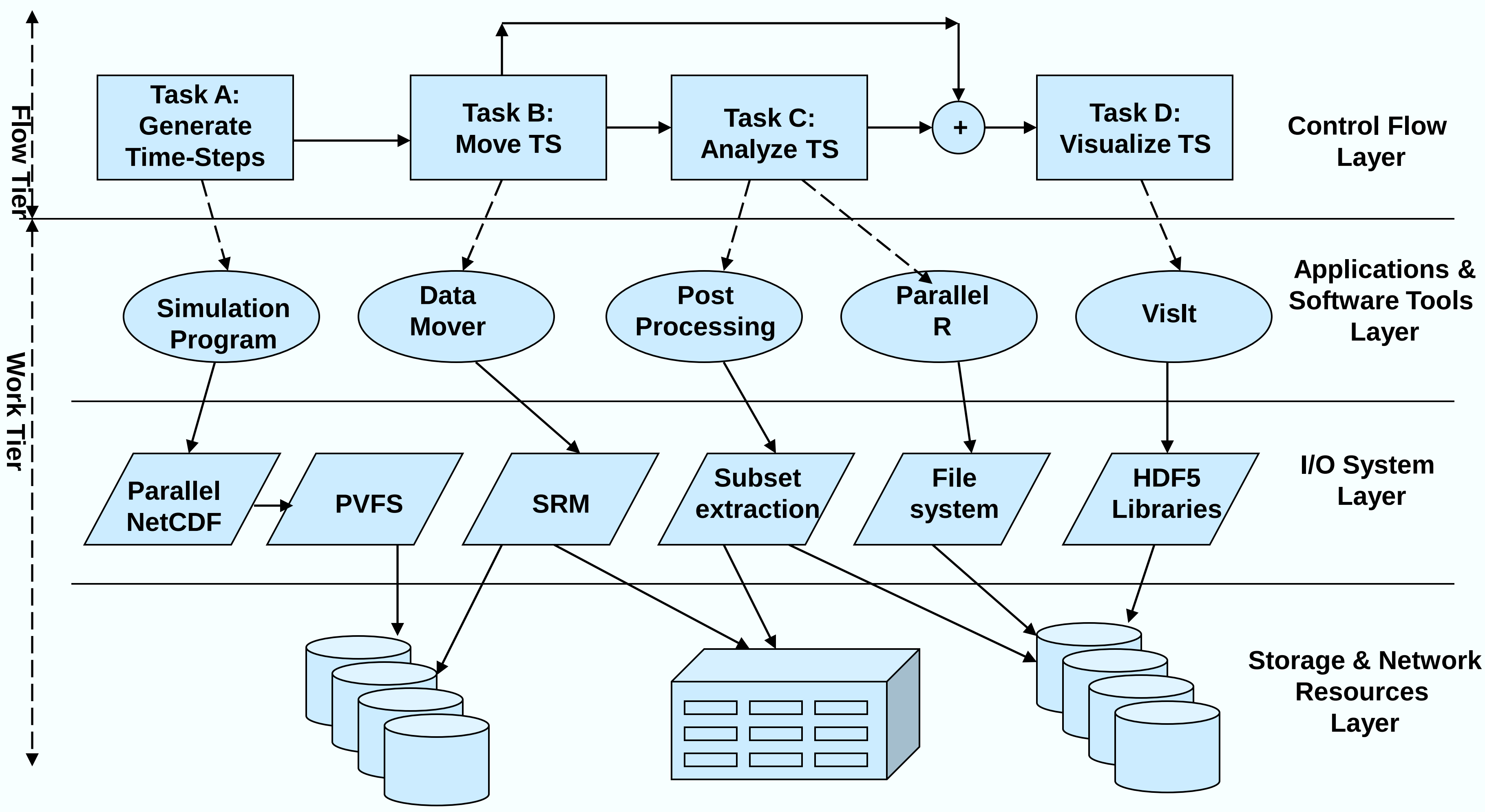




Problems and Goals

- **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**

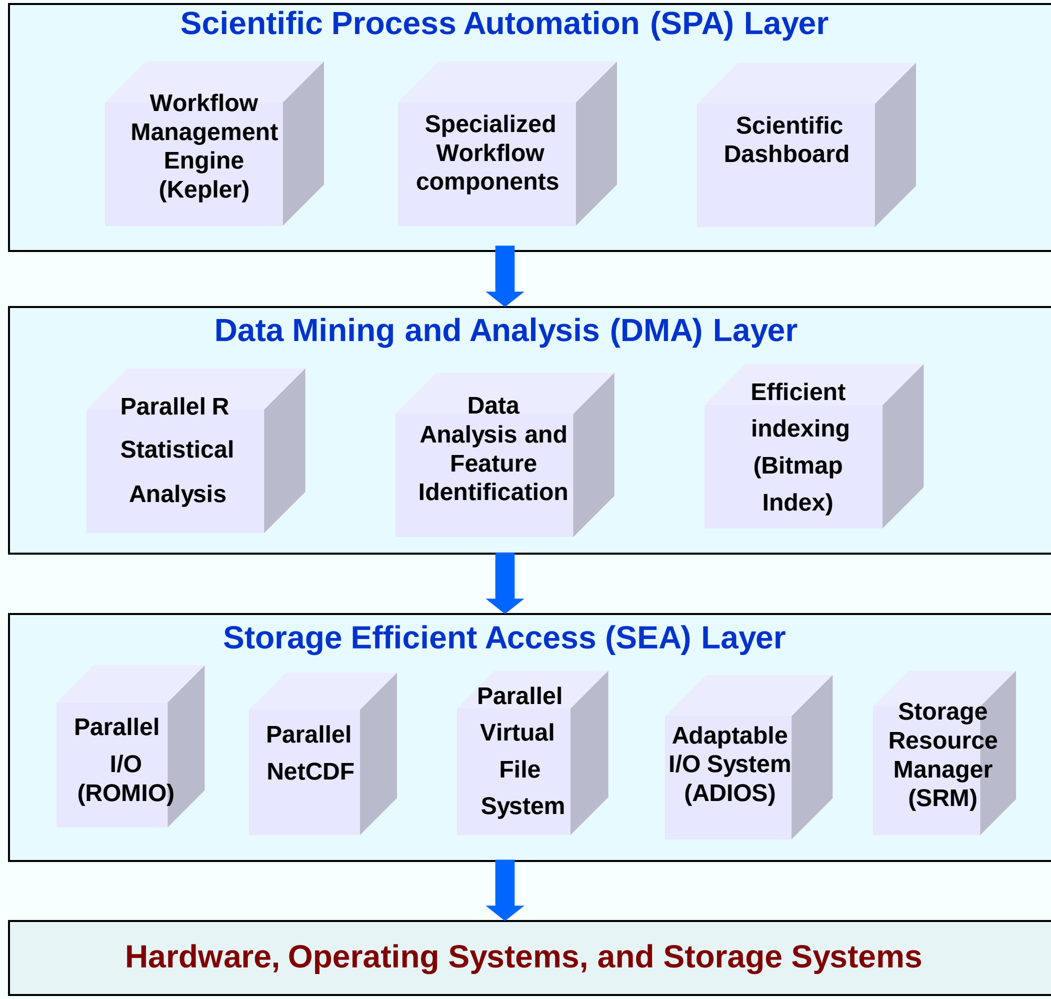
A motivating SDM Scenario (dynamic monitoring)



Organization of the center:
based on three-layer organization of technologies

Integrated approach:

- To provide a scientific workflow and dashboard capability
- To support data mining and analysis tools
- To accelerate storage and access to data



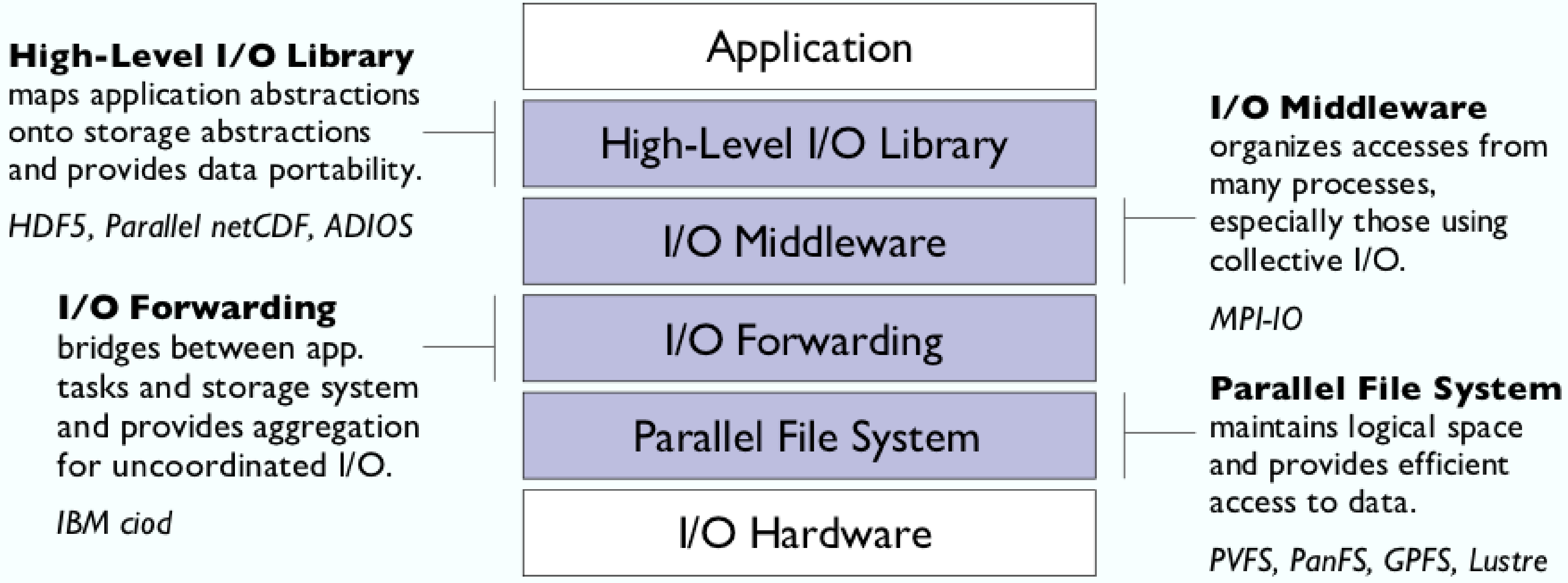
Focus of SDM center

- **high performance**
  - fast, scalable
  - Parallel I/O, parallel file systems
  - Indexing, data movement
- **Usability and effectiveness**
  - Easy-to-use tools and interfaces
  - Use of workflow, dashboards
  - end-to-end use (data and metadata)
- **Enabling data understanding**
  - Parallelize analysis tools
  - Streamline use of analysis tools
  - Real-time data search tools
- **Sustainability**
  - robustness
  - Productize software
  - work with vendors, computing centers
- **Establish dialog with scientists**
  - partner with scientists,
  - education (students, scientists)

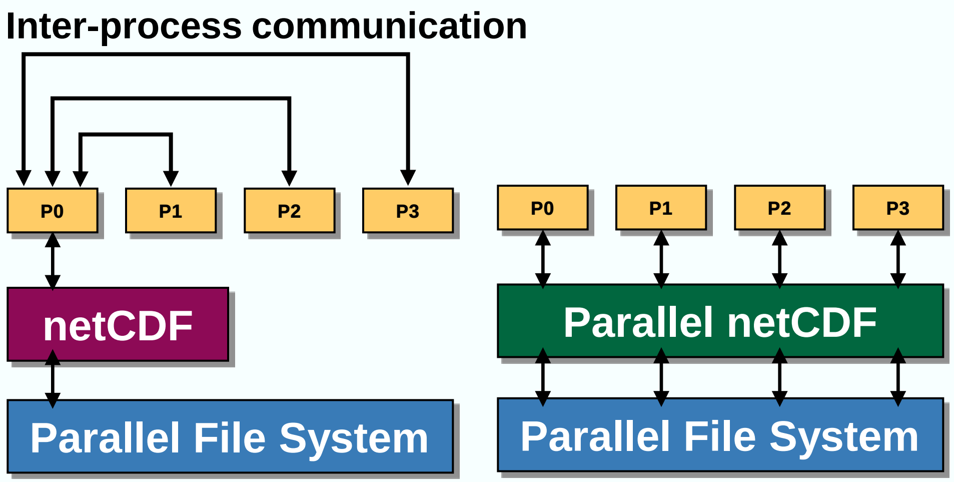
Results

- ✓ **High Performance Technologies**
- **Usability and effectiveness**
- **Enabling Data Understanding**

The I/O Software Stack



Speeding data transfer with PnetCDF

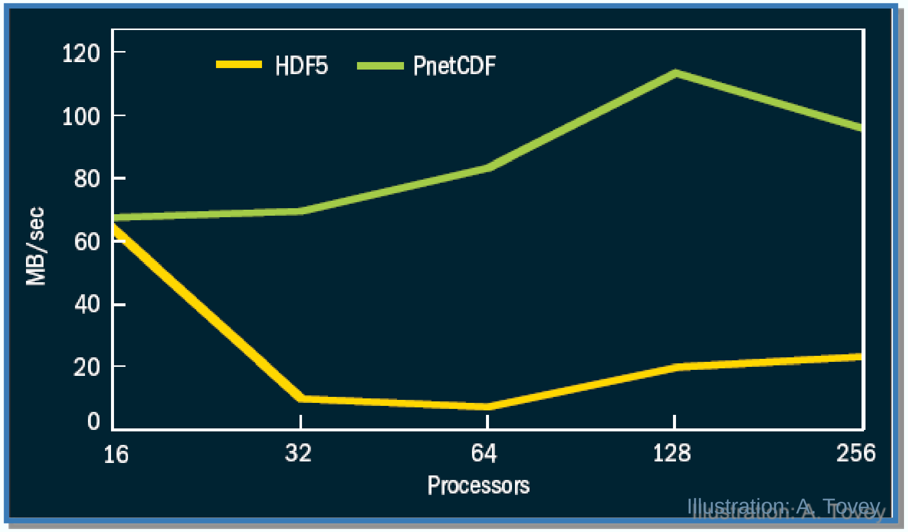


Enables high performance parallel I/O to netCDF data sets

Achieves up to 10-fold performance improvement over HDF5

Early performance testing showed PnetCDF outperformed HDF5 for some critical access patterns.

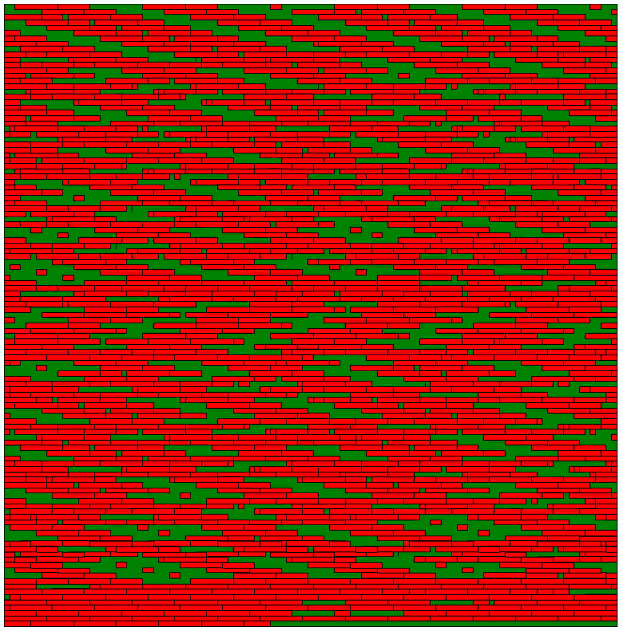
The HDF5 team has responded by improving their code for these patterns, and now these teams actively collaborate to better understand application needs and system characteristics, leading to I/O performance gains in both libraries.



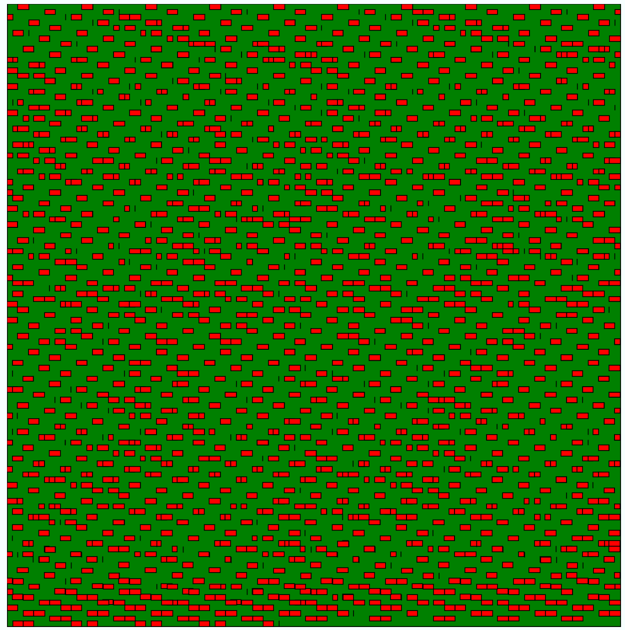
Contacts: Rob Ross, ANL, Alok Choudhari, NWU

Visualizing and Tuning I/O Access

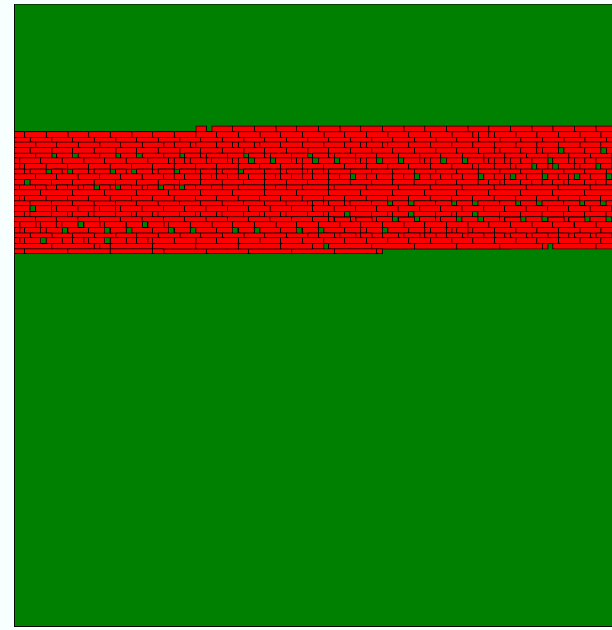
This view shows the entire 28 Gbyte dataset as a 2D array of blocks, for three separate runs. Renderer is visualizing one variable out of five. Red blocks were accessed. Access times in parenthesis.



Original Pattern



MPI-IO Tuning



PnetCDF Enhancements

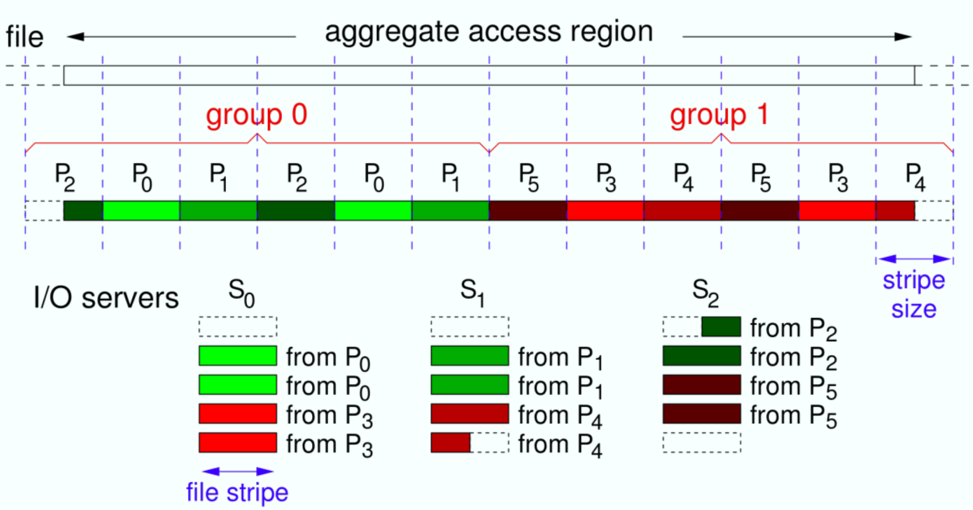
Data is stored in the netCDF “record” format, where variables are interleaved in file (36.0 sec). Adjusting MPI-IO parameters (right) resulted in significant I/O reduction (18.9 sec).

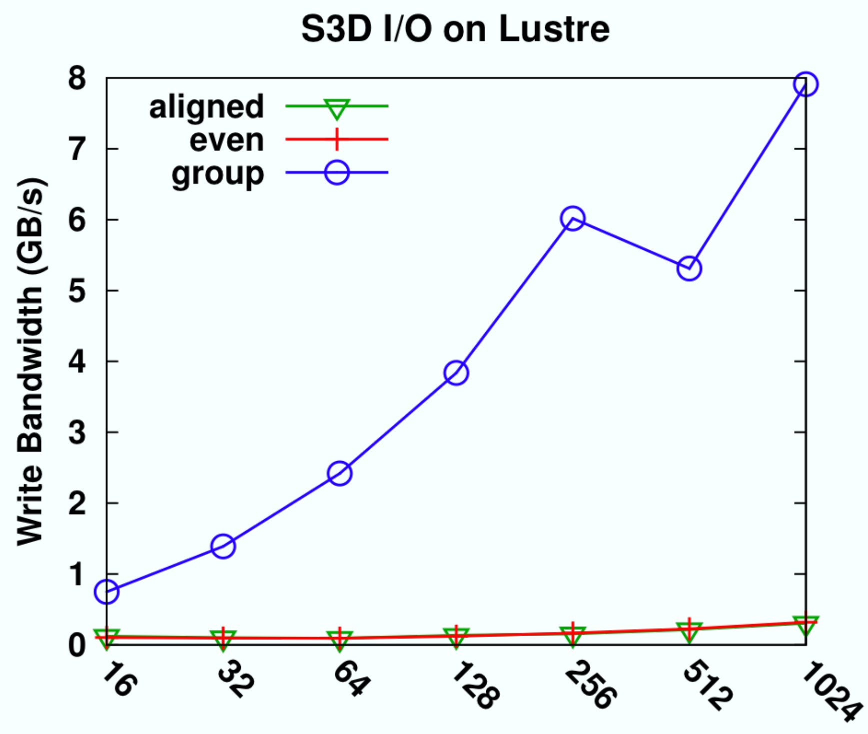
New PnetCDF large variable support stores data contiguously (13.1 sec).

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.

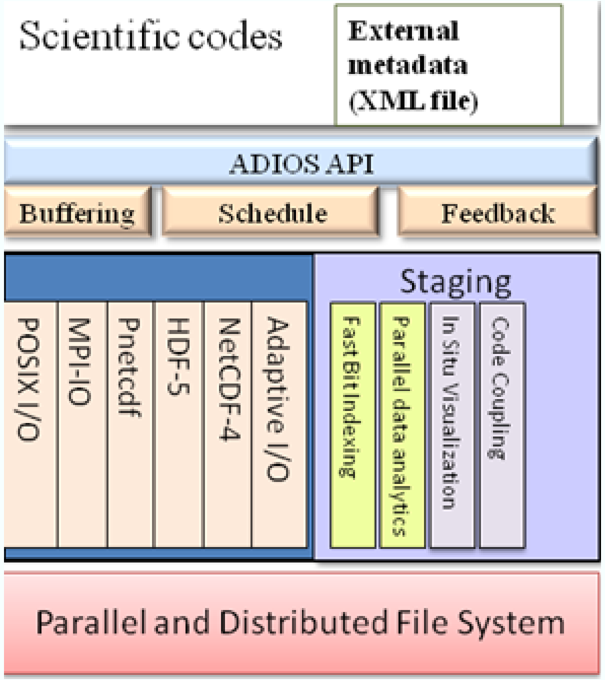




ADaptable IO System (ADIOS)

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

- Provides 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
  - Abstracts the API from the method used for I/O Operate across multiple HPC architectures and parallel 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)
  - Facilitates switching underlying file formats to reach performance goals
- Compensate for inefficiencies in the current I/O infrastructures



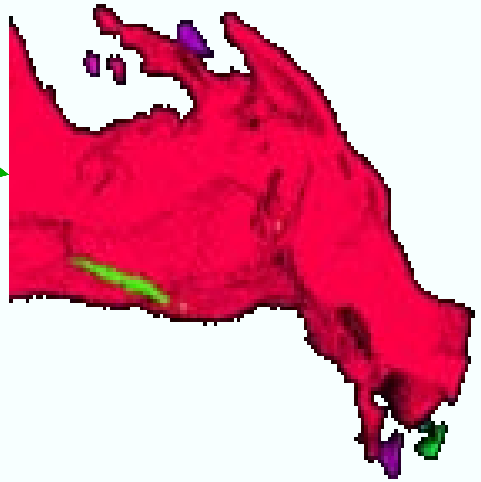
Searching Problems in Data Intensive Sciences

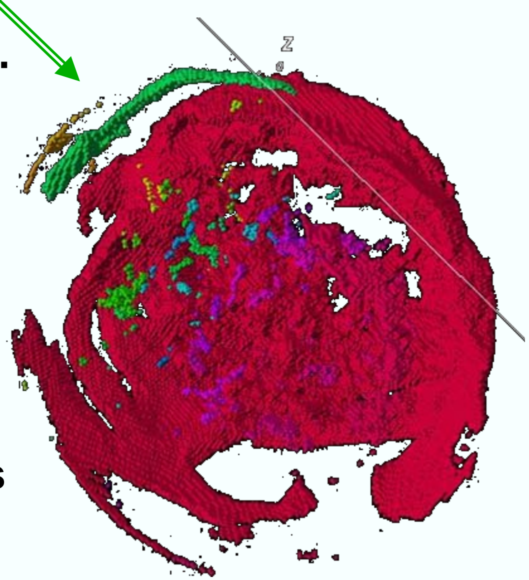
- Find the HEP collision events with the most distinct signature of Quark Gluon Plasma
- Find the ignition kernels in a combustion simulation
- Track a layer of exploding supernova

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



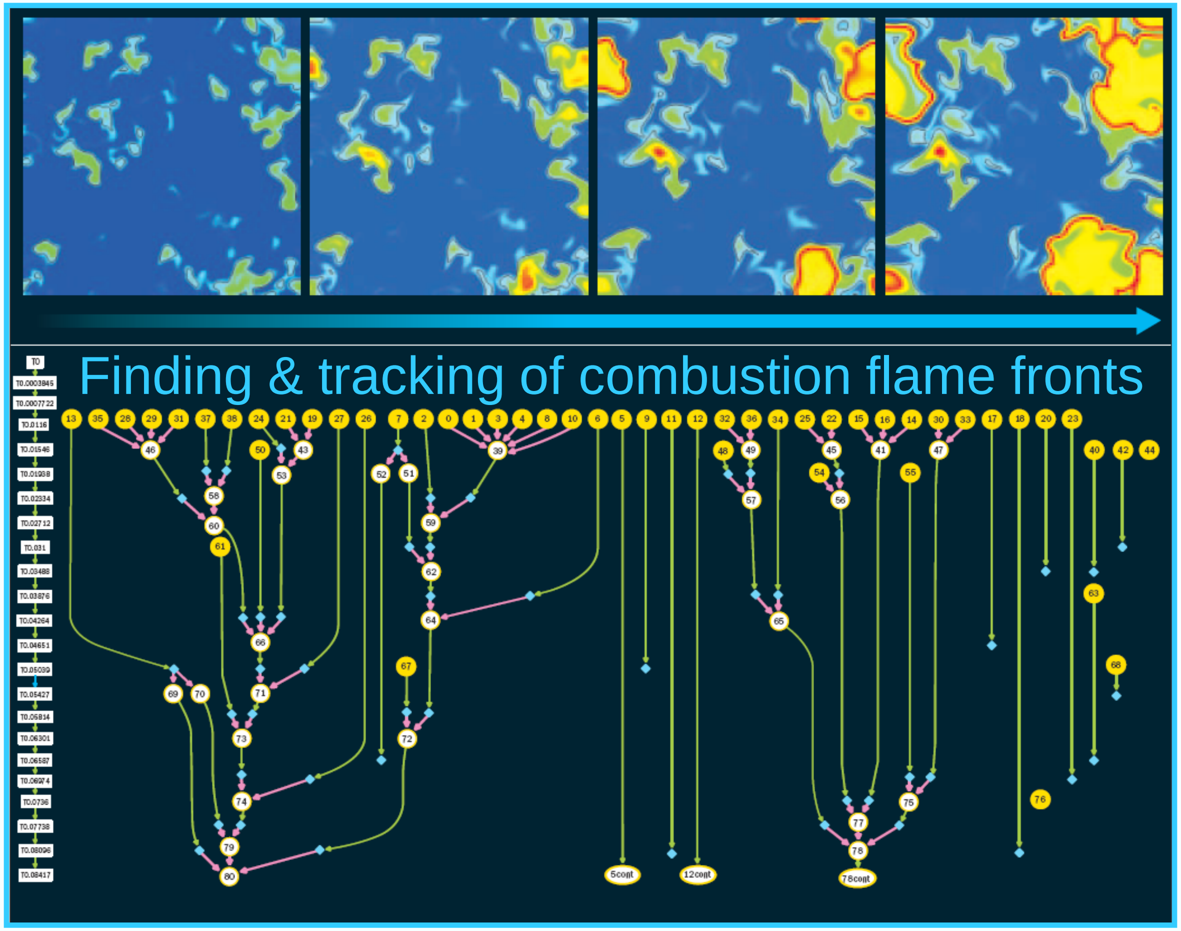


FastBit: accelerating analysis of very large datasets

- 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

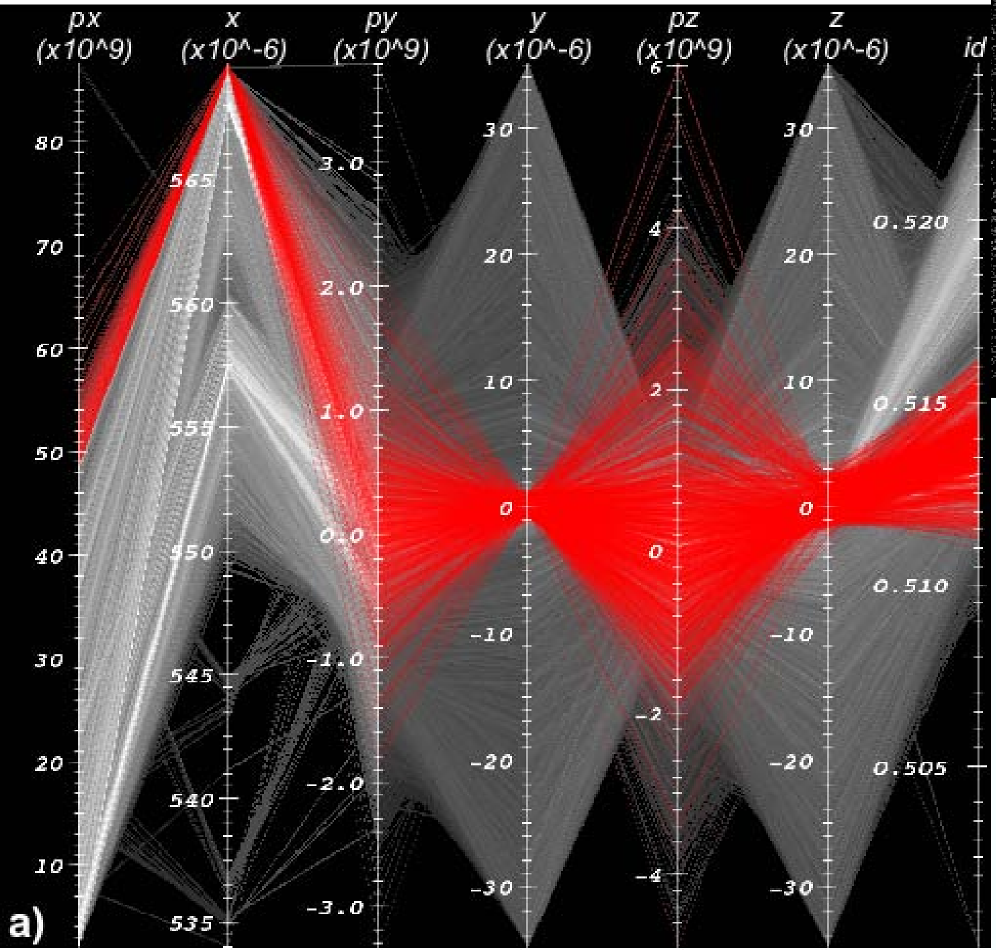
Connect neighboring cells into regions

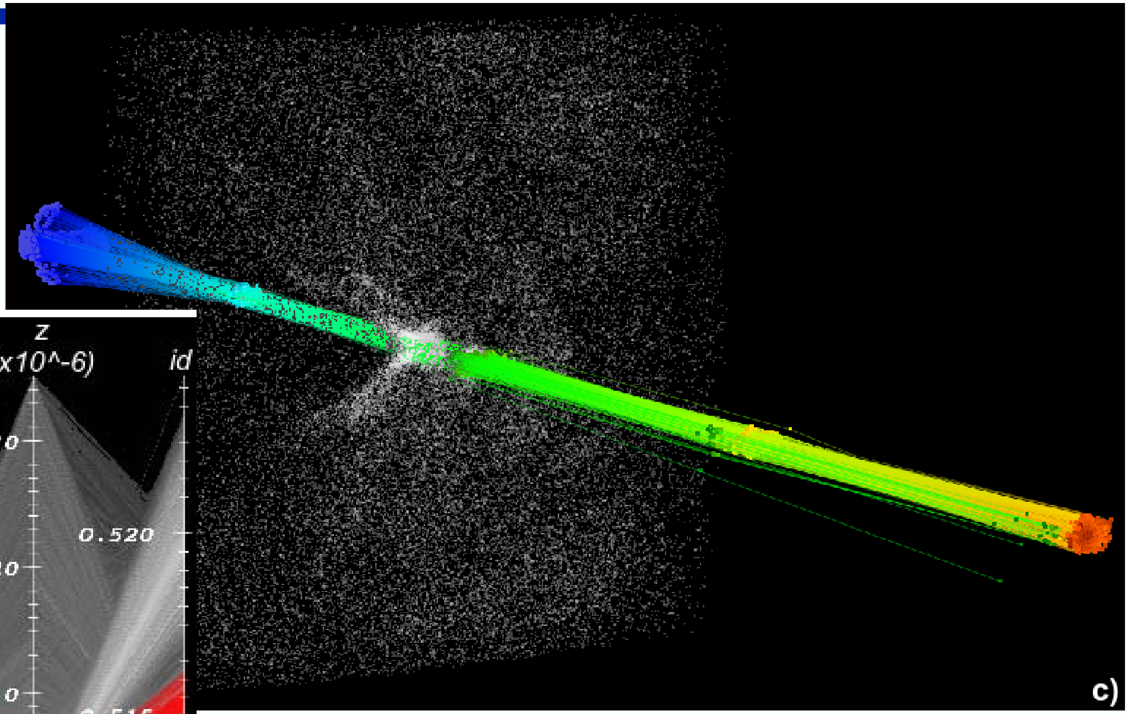
Region tracking

Track the evolution of the features through time

3D Analysis Examples

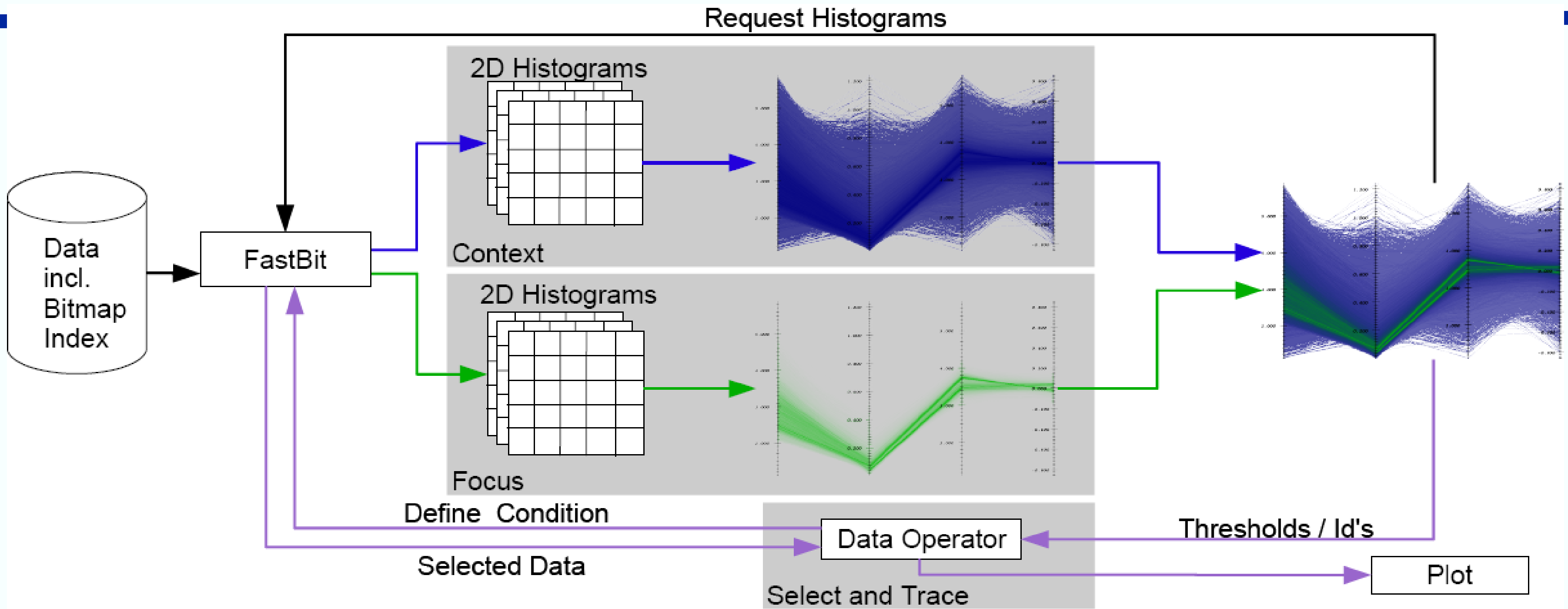
Selecting particles using parallel coordinate display





Trace selected particles

Query-Driven Visualization



- Collaboration between SDM and VIS 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)**

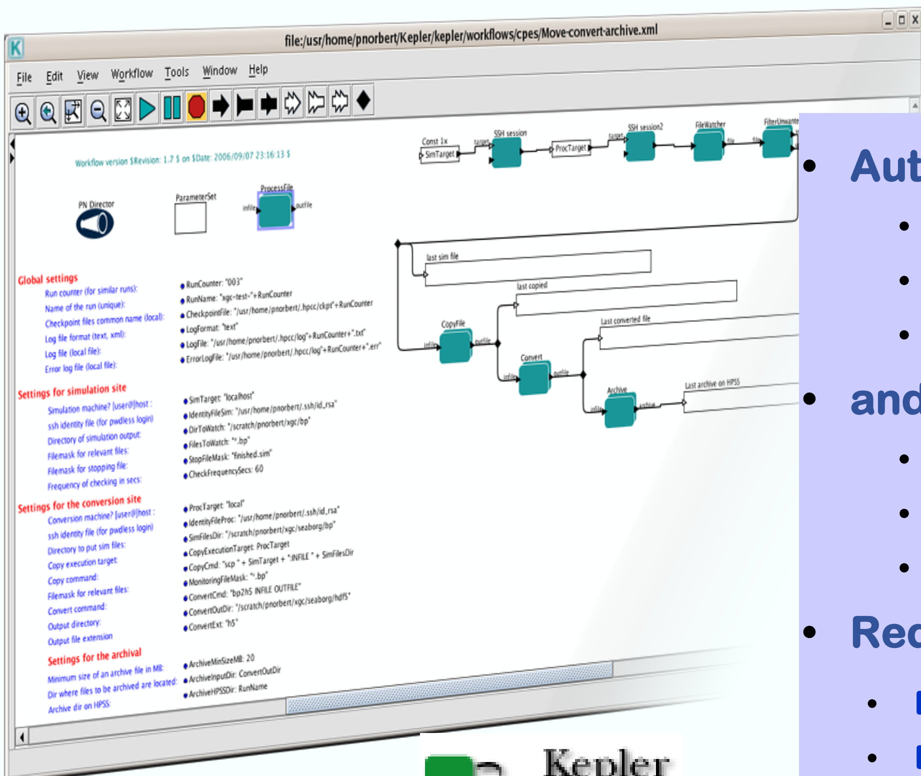
Results

High Performance Technologies

✓ Usability and effectiveness

Enabling Data Understanding

Workflow automation requirements in Fusion Plasma Edge Simulation

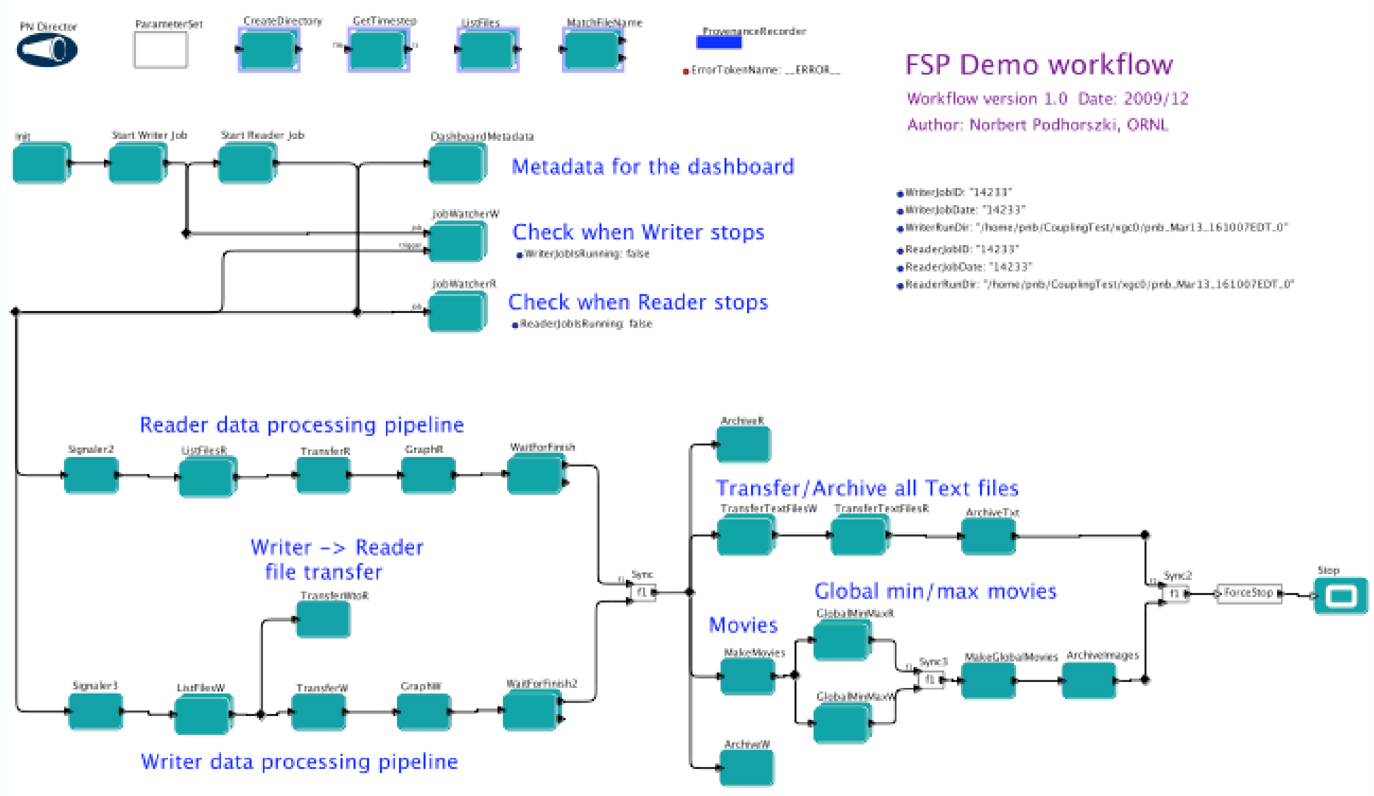


- **Automate the monitoring pipeline**
  - **transfer of simulation output to remote machine**
  - **execution of conversion routines,**
  - **image creation, data archiving**
- **and the code coupling pipeline**
  - **Run simulation on a large supercomputer**
  - **check linear stability on another machine**
  - **Re-run simulation if needed**
- **Requirements for Petascale computing**
  - **Easy to use**
  - **Dashboard front-end**
  - **Dynamic monitoring**
  - **Parallel processing**
  - **Robustness**
  - **Configurability**

Kepler

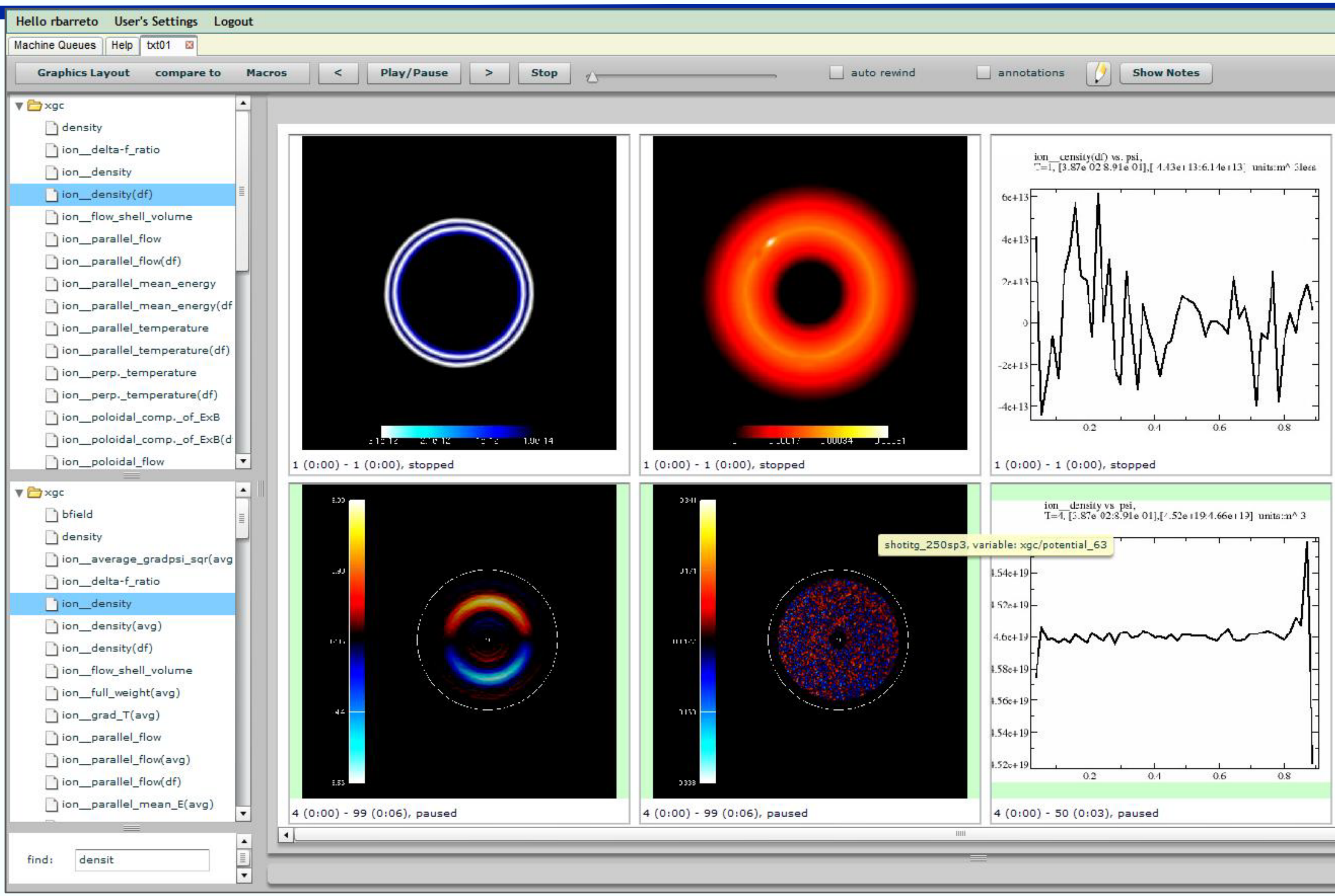
Contact: Scott Klasky, et. al, ORNL

The Kepler Workflow Engine



- **Kepler is a workflow execution system based on Ptolemy (open source from UCB)**
- **SDM center work is in the development of components for scientific applications (called actors)**

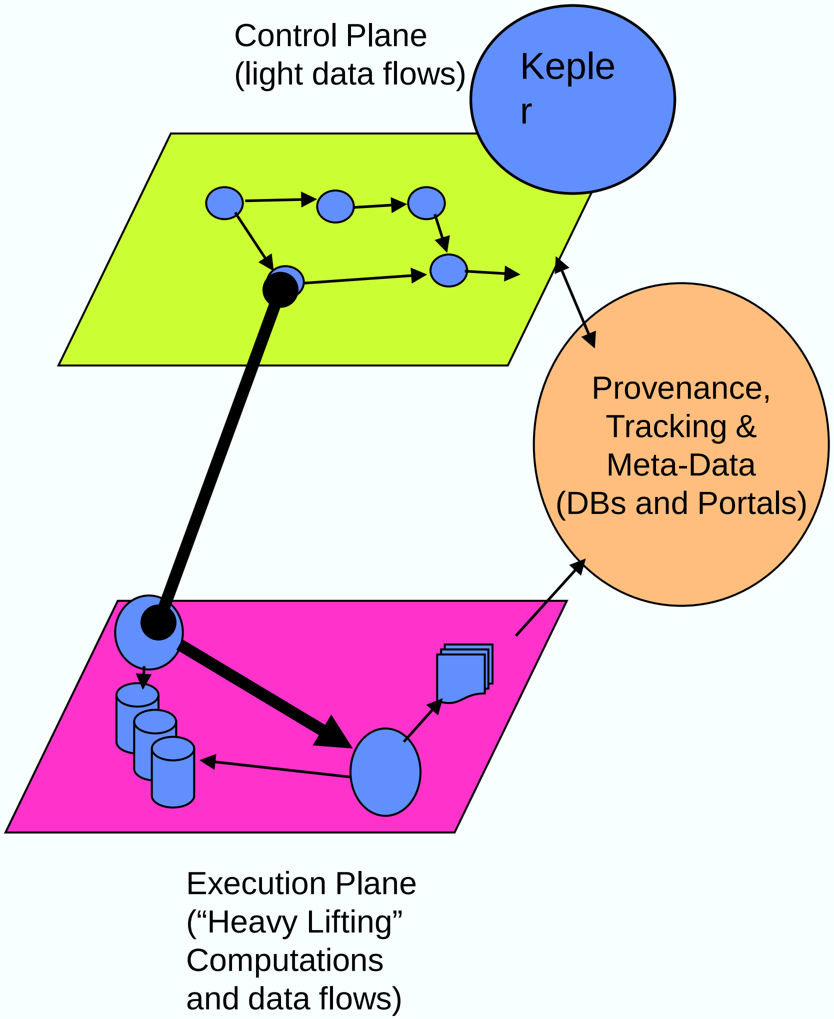
Real-time visualization and analysis capabilities on dashboard



visualize and compare shots

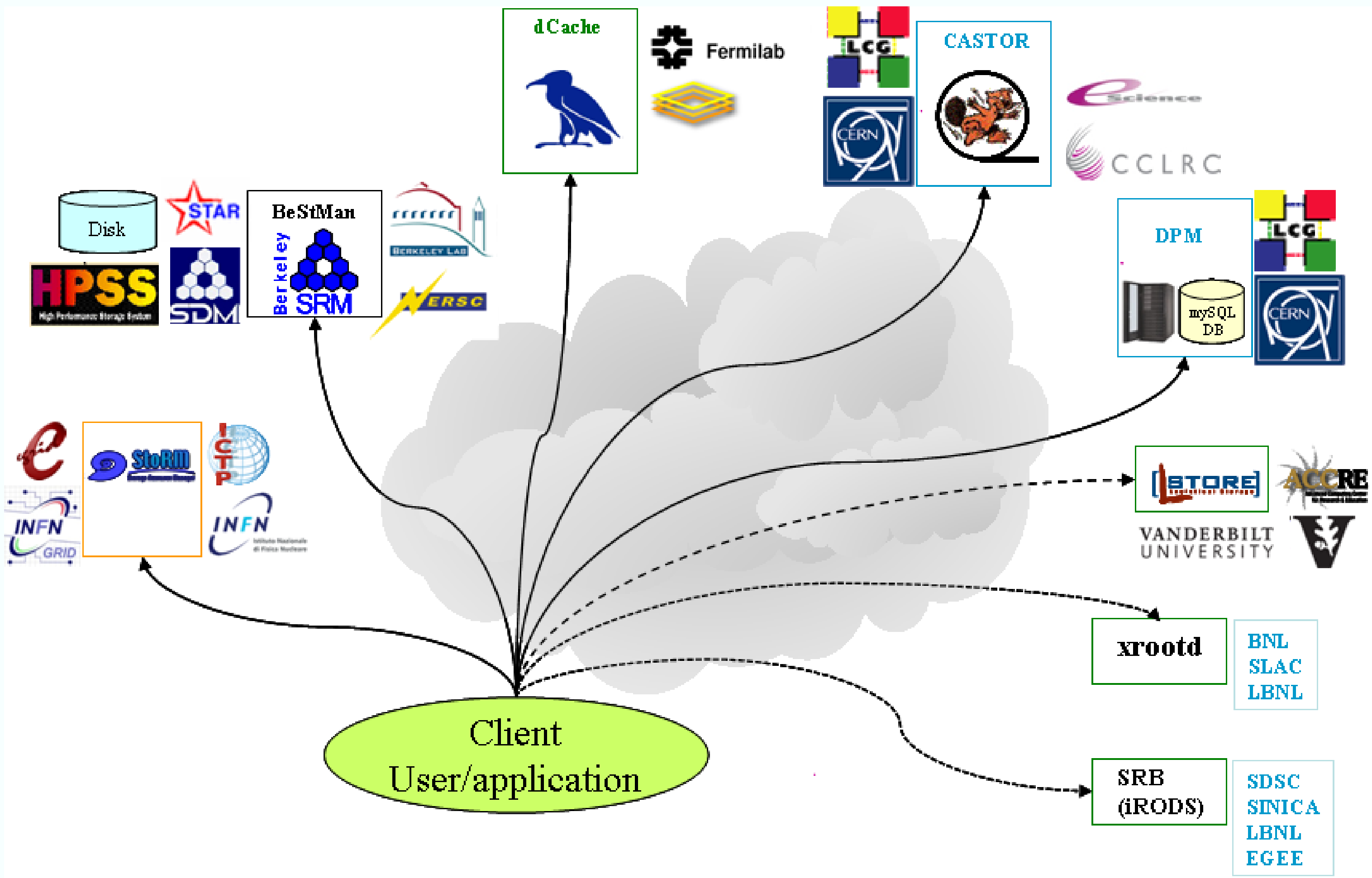
Capturing Provenance in Workflow Framework

- ***Process* provenance**
  - **the steps performed in the workflow, the progress through the workflow control flow, etc.**
- ***Data* provenance**
  - **history and lineage of each data item associated with the actual simulation (inputs, outputs, intermediate states, etc.)**
- ***Workflow* provenance**
  - **history of the workflow evolution and structure**
- ***System* provenance**
  - **Machine and environment information**
  - **compilation history of the codes**
  - **information about the libraries**
  - **source code**
  - **run-time environment settings**

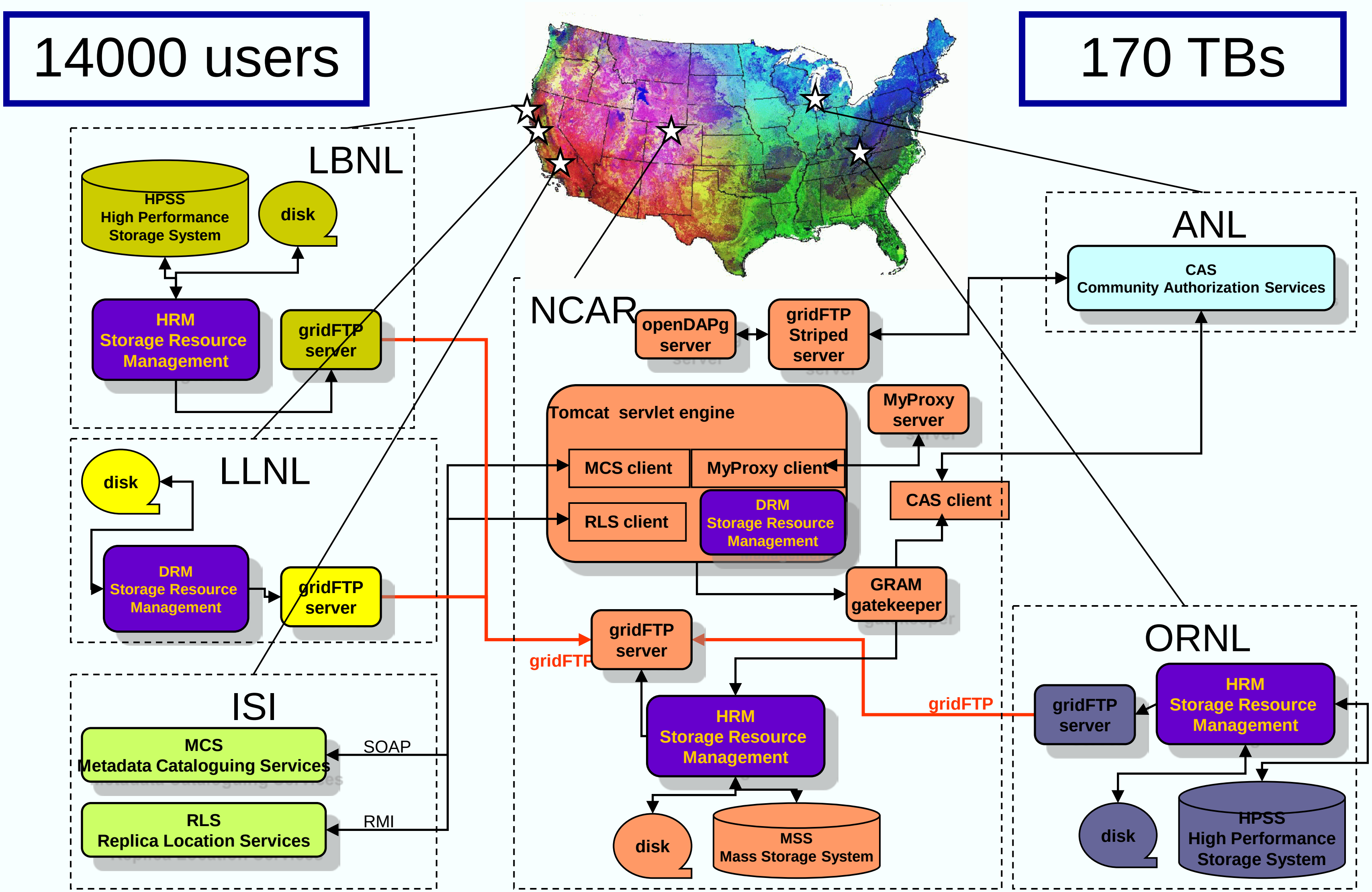


SDM Contact: Mladen Vouk, NCSU

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

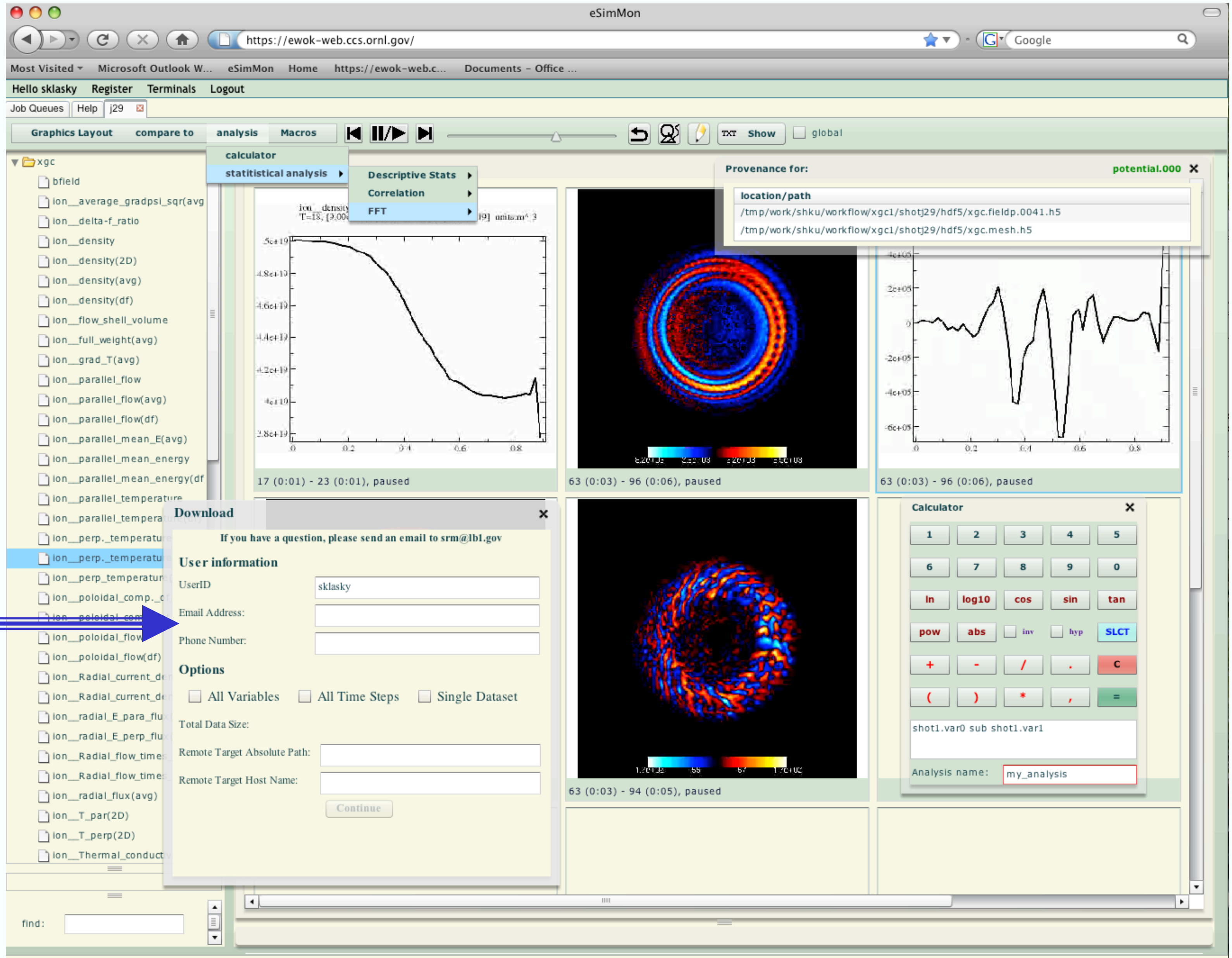


SRM use in Earth Science Grid



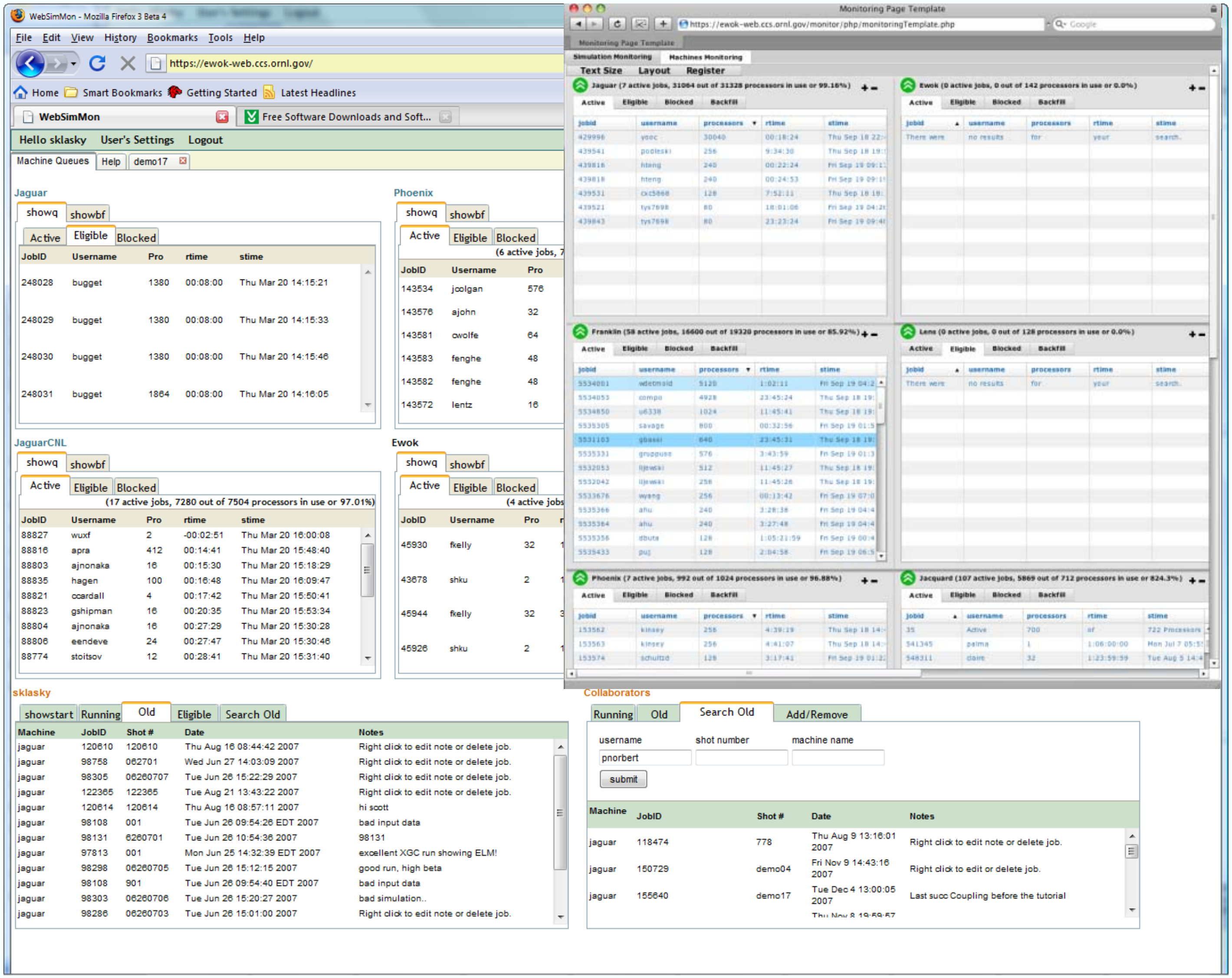
SDM Contact: A. Sim, A. Shoshani, LBNL

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



Download window

Dashboard is used for job launching and real-time machine monitoring



- Allow for secure logins with OTP.
- Allow for job submission.
- Allow for killing jobs.
- Search old jobs.
- See collaborators jobs.

Results

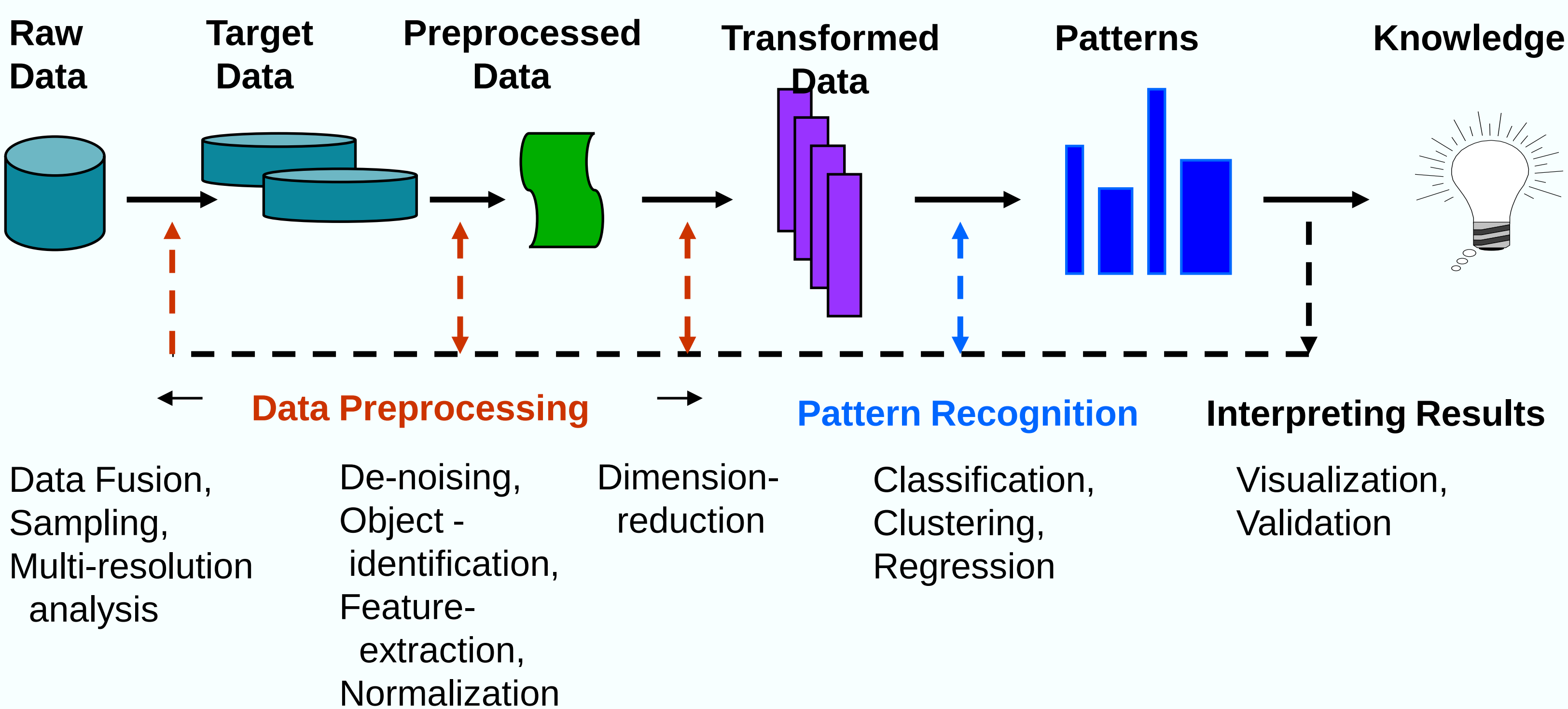
High Performance Technologies

Usability and effectiveness

✓ Enabling Data Understanding

Scientific data understanding: from Terabytes to a Megabytes

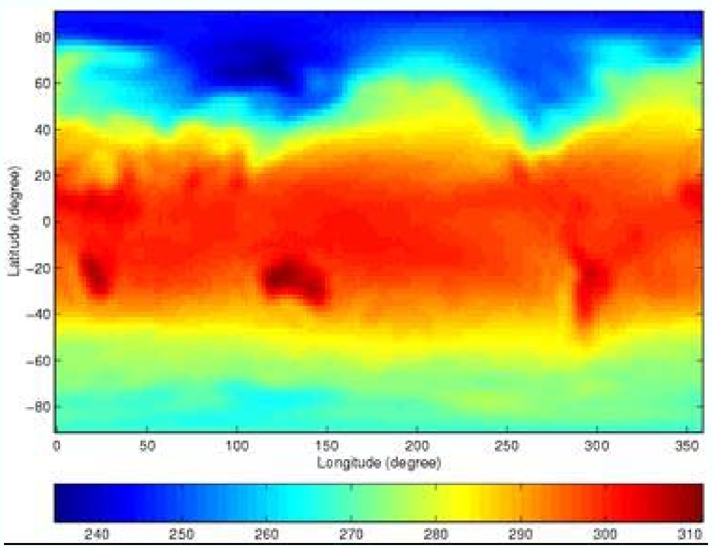
- **Goal: solving the problem of data overload**
  - **Use scientific data mining techniques to analyze data from various applications**
  - **Techniques borrowed from image and video processing, machine learning, statistics, pattern recognition, …**

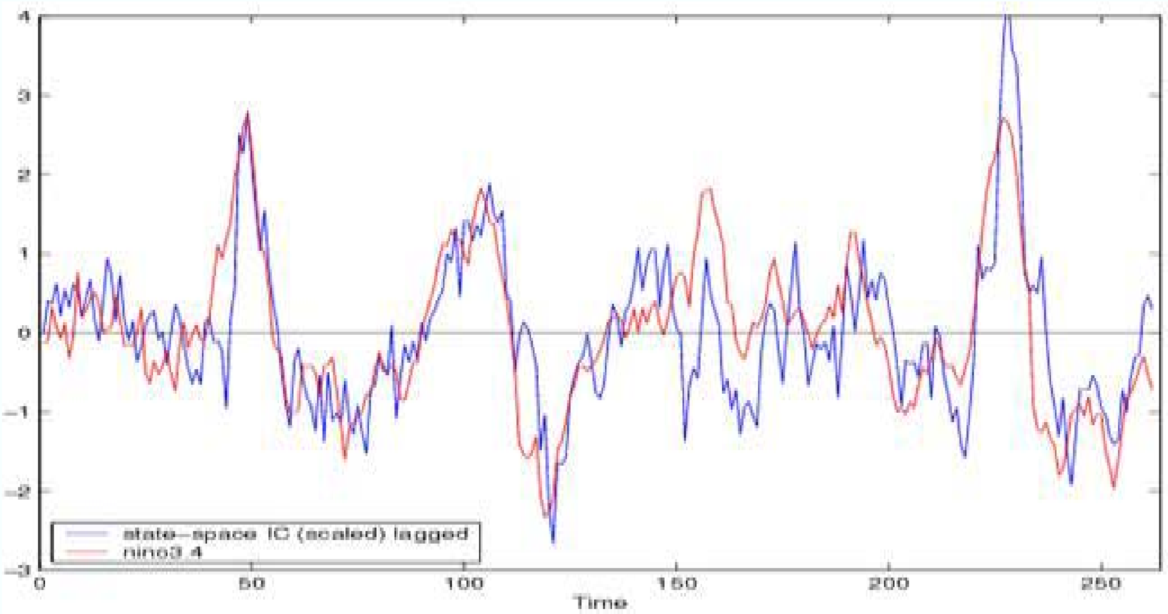


An iterative and interactive process

Separating signals in climate data

- Independent component analysis was used to separate El Niño and volcano signals in climate simulations
- Showed that the technique can be used to enable better comparisons of simulations

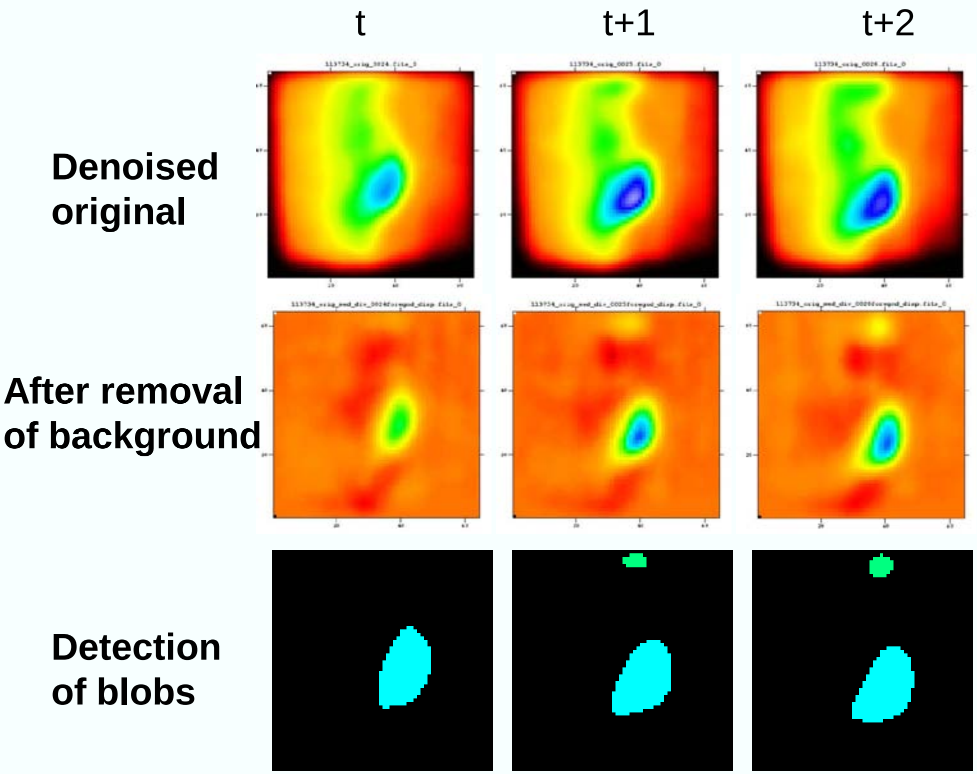




Collaboration with Ben Santer (LLNL)

Tracking blobs in fusion plasma

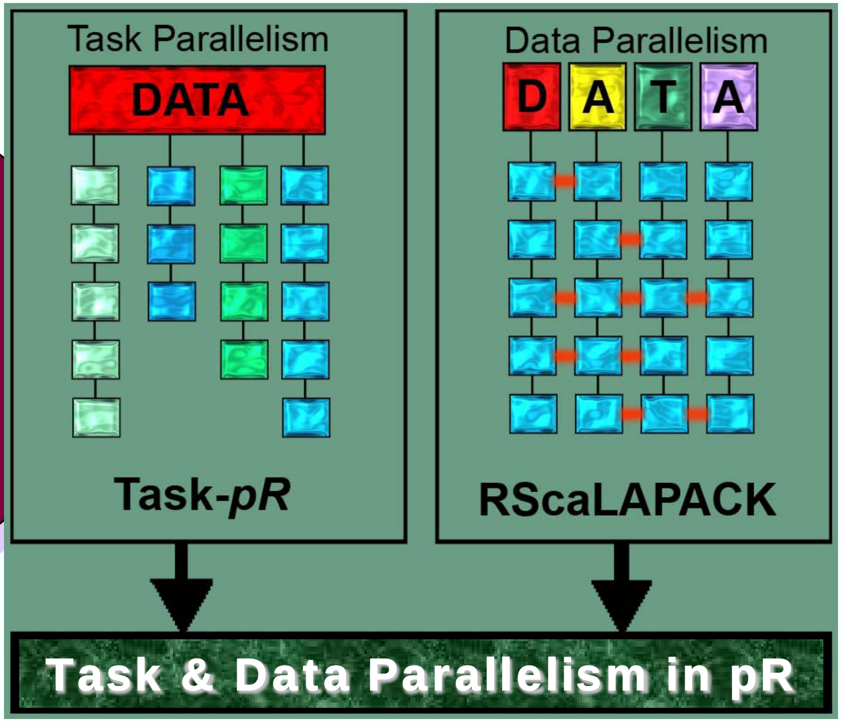
- **Using image and video processing techniques to identify and track blobs in experimental data from NSTX to validate and refine theories of edge turbulence**



Collaboration with S. Zweben, R. Maqueda, and D. Stotler (PPPL)

Task and Data Parallelism in *pR*

Task Parallelism **Data Parallelism**



<u>Goal</u>: Parallel R (*pR*) aims:

(1) to automatically detect and execute *task-parallel* analyses;

(2) to easily plug-in *data-parallel* MPI-based C/Fortran codes

(3) to retain high-level of *interactivity, productivity* and *abstraction*

Task-parallel analyses:

- Likelihood Maximization
- Re-sampling schemes: Bootstrap, Jackknife
- Markov Chain Monte Carlo (MCMC)
- Animations

Data-parallel analyses:

- *k*-means clustering
- Principal Component Analysis
- Hierarchical clustering
- Distance matrix, histogram, etc.

ProRata use in OBER Projects

TOOLbox

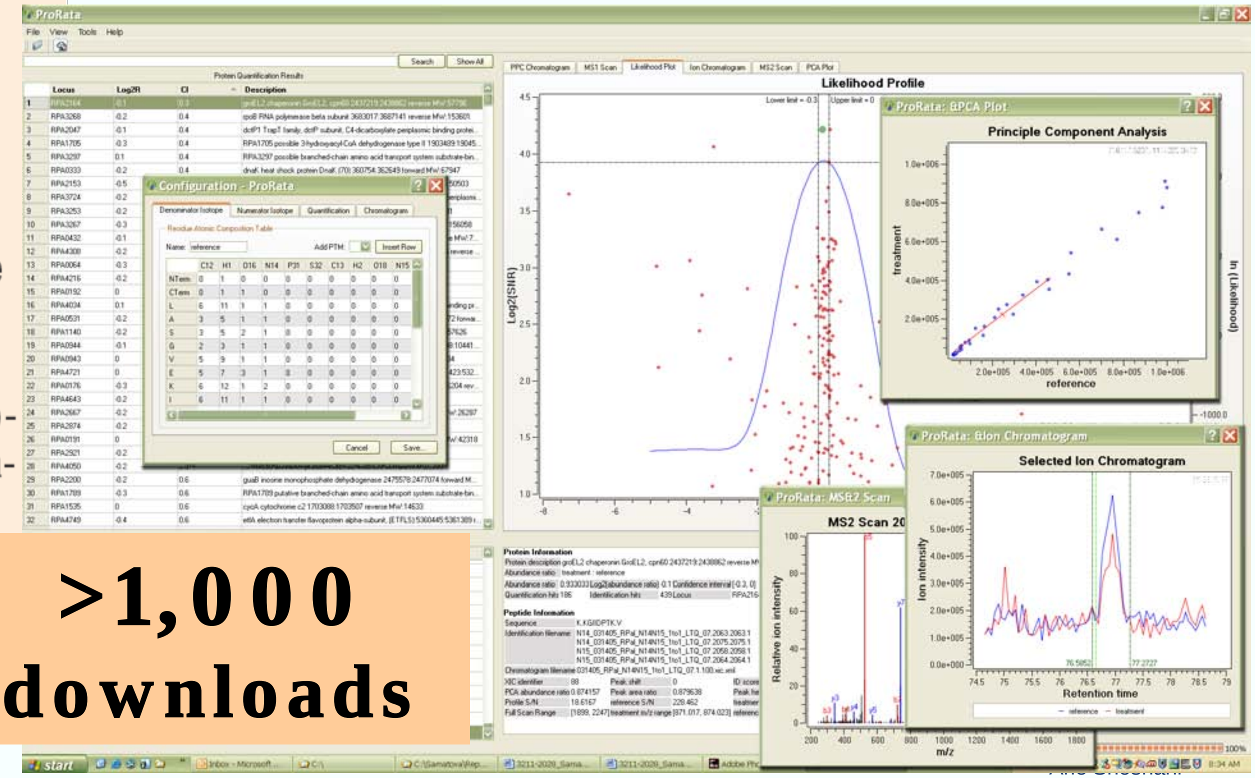
ProRata

To detect quantitative protein differences between two conditions, researchers often perform stable-isotope labeling. The algorithms that currently are applied to the data, however, can make incorrect assignments because the S/N typically is low in these experiments. Also, the programs do not assess bias and variability. So, Nagiza Samatova, Robert Hettich, and colleagues at the Oak Ridge National Laboratory and the University of Tennessee developed a program called ProRata that improves identifications. The program can be applied to data obtained from various stable-isotope labeling techniques.

J. of Proteome Research
Vol. 5, No. 11, 2006

DOE OBER Projects Using ProRata:

- *Jill Banfield, Bob Hettich*: Acid Mine Drainage
- *Michelle Buchanan*: CMCS Center
- *Steve Brown, Jonathan Mielenz*: BESC BioEnergy
- *Carol Harwood, Bob Hettich*: MCP *R. palustris*



>1,000 downloads

SDM center collaboration with applications

Application Domains	Workflow Technology (Kepler)	Metadata And provenance	Data Movement and storage	Indexing (FastBit)	Parallel I/O (pNetCDF, etc.)	Parallel Statistics (pR, …)	Feature extraction	Adaptive I/O System (ADIOS)
Climate Modeling (Drake)	workflow				pNetCDF	pMatlab		
Astrophysics (Blondin)	data movement	dashboard						
Combustion (Jackie Chen)	data movement	distributed analysis	DataMover-Lite	flame front	Global Access	pMatlab	tranient events	monitoring
Combustion (Bell)			DataMover-Lite					
Fusion (PPPL)							poincare plots	
Fusion (CPES)	data-move, code-couple	Dashboard	DataMover-Lite	Toroidal meshes		pR	Blob tracking	M-to-M coupling
Materials - QBOX (Galli)					XML			
High Energy Physics	Lattice-QCD		SRM, DataMover	event finding				
Groundwater Modeling	identified 4-5 workflows							
Accelarator Science (Ryne)					MPIO-SRM			
SNS	workflow	Data Entry tool (DEB)						
Biology	ScalaBlast					ProRata		
Climate Cloud modeling (Randall)					pNetCDF			
Data-to-Model Coversion (Kotamathi)								
Biology (H2)								
Fusion (RF) (Bachelor)							poincare plots	
Subsurface Modeling (Lichtner)						Over AMR		
Flow with strong shocks (Lele)						conditional statistics		
Fusion (extended MHD) (Jardin)								
Nanoscience (Rack)						pMatlab		

Ongoing collaborations   In progress   interest expressed

Data challenges in the extreme scale

- Minimize volume of data to be stored
- Prepare data for analysis before storing
- Reduce energy

Approaches to the data challenges in the extreme scale

- Minimize volume of data to be stored
  - In-situ analysis
    - Use extra cores, I/O nodes, and staging nodes
  - Summarize data in-situ => parallel statistics
    - Many statistical functions can be parallelized
    - Need algorithms for piece-wise statistical computation
    - Take advantage of multi-cores and GPU technologies
  - Avoid getting data to disk for intermediate data
    - E.g. Parameter setup, Validation, and Uncertainty Quantification (UQ) requires many runs, but only summaries of each run is needed
    - Move summaries to staging nodes for guiding next step in parameter choices
  - Perform monitoring of simulation progress in Situ
    - Requires in-memory support of workflows

Approaches to the data challenges in the extreme scale

- Prepare data for analysis before storing
  - Index generation
    - Use index methods that take advantage of invariant data
    - Implement indexes that can be assembled piece-wise in parallel
    - Take advantage of extra cores for index pieces, and staging nodes for assembling the pieces
  - Include pre-computed statistics with files
    - e.g. nim/max for verifying correctness or finding outliers
    - e.g. averages can be used to find trends
  - Perform data transposition in-situ before storing
    - Organize multi-variable spatio-temporal data by variable over time
    - e.g. reorganize climate time-step data by pressure, temperature, etc. over years
  - Chunk data according to multi-dimentional access patterns
    - General algorithms to co-locate data that is accessed together
  - Pre-compute multi-level summaries
    - e.g. monthly means in climate

Approaches to the data challenges in the extreme scale

- Reduce energy (1)
  - Store intermittent data on large SSDs
    - Checkpoint data => remove previous checkpoints data ASAP
    - Monitoring data => make SSD storage visible to external tools (web)
  - Store longer term data on tape
    - Tape is still cost effective, and will continue to be so for awhile
    - Requires no/little energy to store
    - Only 10-15% usually accessed, but data is important to keep
    - Use data-time-stamping scheme for automatic removal of un-needed data (consult with owner or administrator)
  - When data is stored on disk for further analysis
    - Bring from tape as needed
    - Power down disks when possible based on access patterns
    - Use multi-speed disk
    - Use SSDs to front disks to hold "hot files"

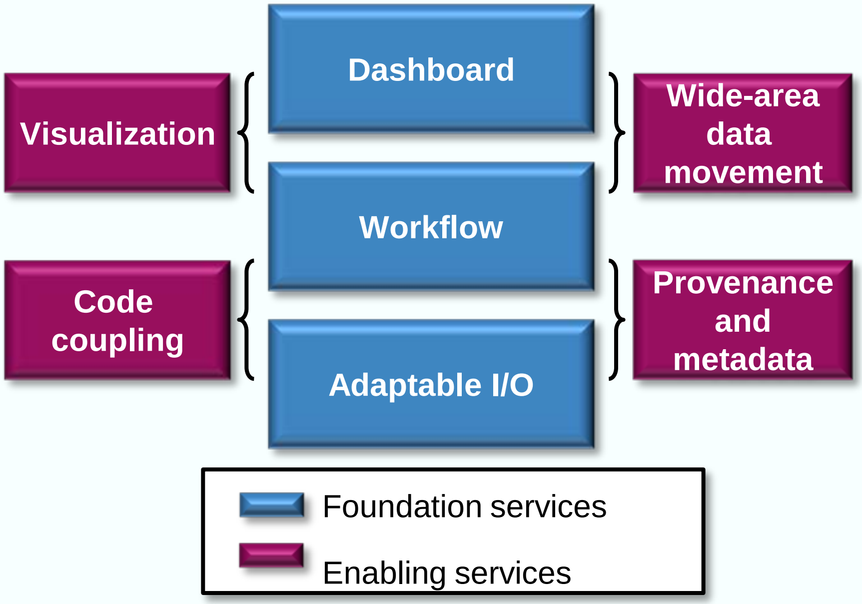
Approaches to the data challenges in the extreme scale

- Reduce energy (2)
  - Minimize data movement in exascale machine
    - Take advantage of multi-core , localize data access
    - E.g. When creating an index, use extra core to generate partial indexes, then combine results
    - Collect multiple “writes” and write them asynchronously
    - Perform code-coupling in memory
  - Maximize sharing of data through SSDs/shared storage
    - Manage shared multi-user access to same data object – avoid replication
    - E.g. A group of users access the same data for analysis (HEP, Climate)
  - Minimize data access after it is stored – take advantage of indexing
    - Indexing pinpoints the data that needs to be accessed
    - Indexing can include statistics – e.g. count, min/max per index-bin to generate histograms
    - Use index to perform region-growing and …
    - to perform region tracking (e.g. front propagation)

Implications from SDM center experience: integration of tools

FIESTA: Framework for Integrated End-to-end SDM Technologies and Applications

- Adaptable I/O
- Workflows
- Dashboard
- Provenance
- Code coupling
- WAN data movement
- Visualization

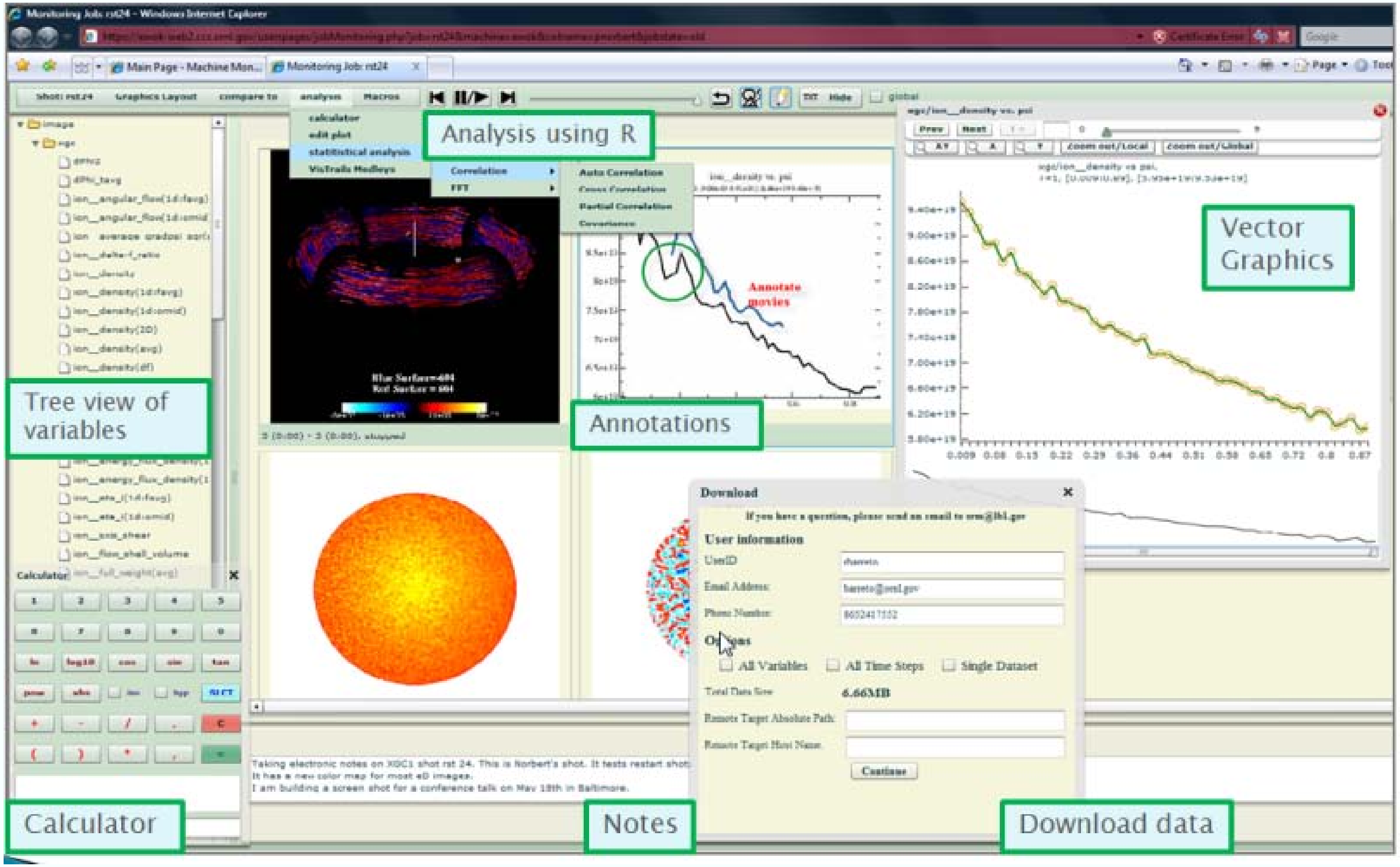


Approach: Place highly annotated, fast, easy-to-use I/O methods in the code, which can be monitored and controlled; have a workflow engine record all of the information; visualize this on a dashboard; move desired data to the user's site; and have everything reported to a database.

Benefit: automate complex tasks, and allow users to interact through simple interfaces that expose physics products remotely over the web.

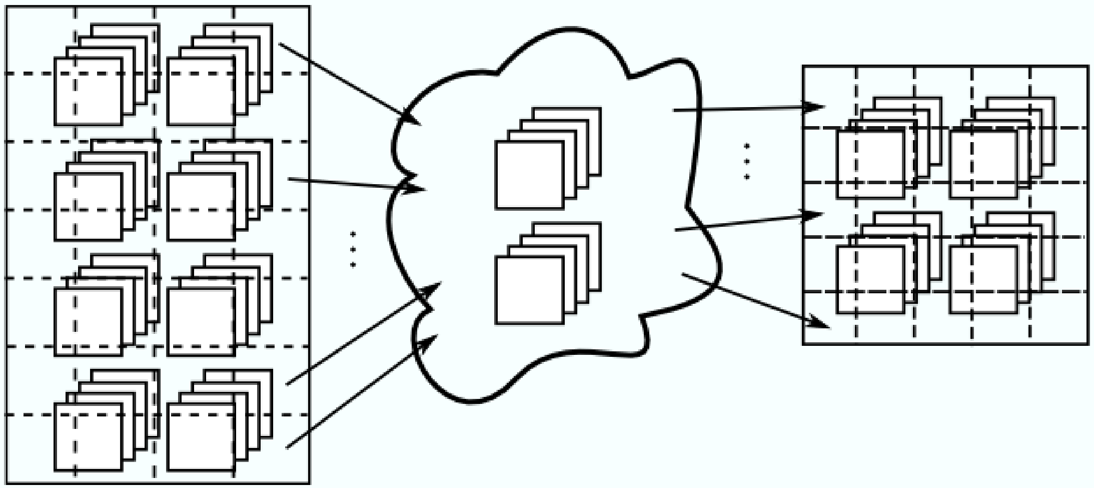
Implications from SDM center experience: web interfaces for users

eSimMon: dashboard for collaborative data management, analysis, and visualization



Two Fusion codes Coupled Using DataSpaces

- **The simulations exchange multi-dimensional data arrays (e.g., 2D)**
  - **Domain discretization is different for the two applications**
  - **Data redistribution is transparent and implicit through the space**
- **The simulations have different interaction patterns**
  - **e.g., one-to-many, many-to-many, many-to-one**



2D Array of Doubles Distributed on 32 Processors

DataSpaces

2D Array of Doubles Distributed on 6 Processors

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, eSiMon 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**
  - **Deep data analysis requires effective data access and transfer to user's facilities**
    - **In center: DataMover**
  - **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)**

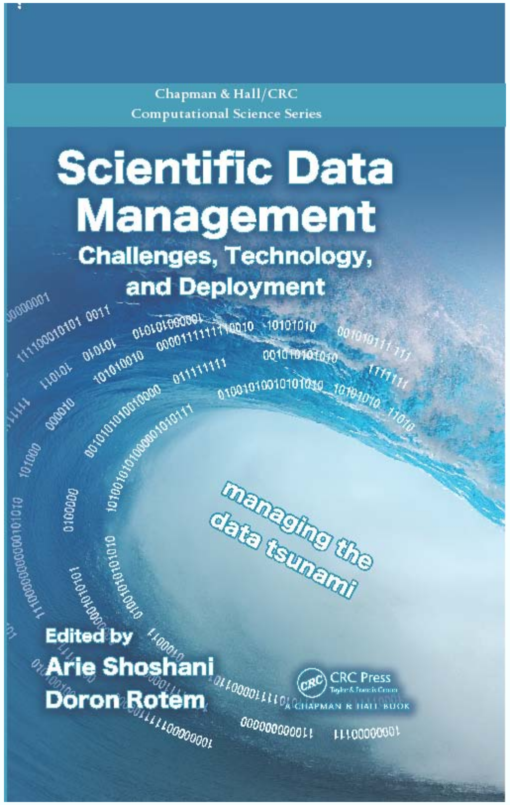
Analysis at extreme scale: Data-Side Analysis Facility

- It is becoming impractical to move large parts of simulation data to end user facilities
  - "Near data" could be a high capacity wide-area network (100 Gbps)
  - On-the-fly processing capabilities – as data is generated
- Data-side analysis facility (exascale workshops)
  - Have an analysis cluster near the data generation site
  - Have parallel analysis and visualization tools available on facility
  - Have workflow tools to compose "analysis pipelines" by users
  - Reuse previously composed pipelines
  - Package specialized components (e.g. Poincare plot analysis)
- Use dynamically or as post-processing
  - Invoke as part of end-to-end framework
  - Use provenance store to track results

SDM Book – December 2009

New book edited and many chapters written by SDM Center members (Arie Shoshani and Doron Rotem, editors)

- **Scientific Data Management: Challenges, Technology, and Deployment**
- **Chapman & Hall/CRC**



Book Organization

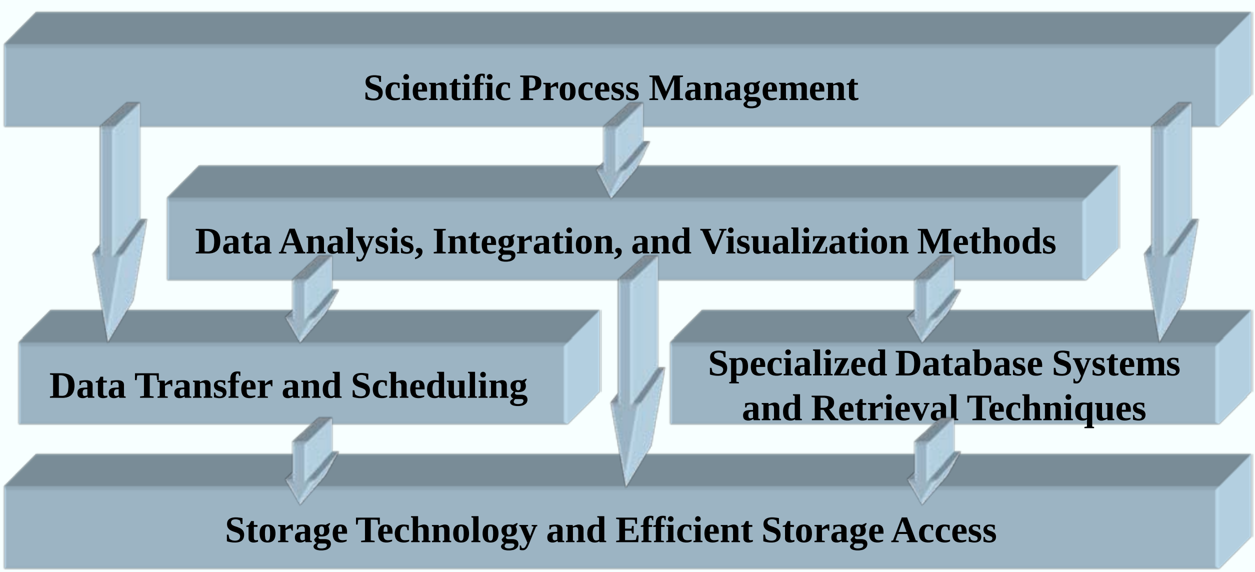


Table-of-Contents

The END