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

Arie Shoshani (LBNL)

Scott Klasky (ORNL)

Rob Ross (ANL)

(and the entire SDM center team)

SciDAC Meeting, July 12-15, 2010

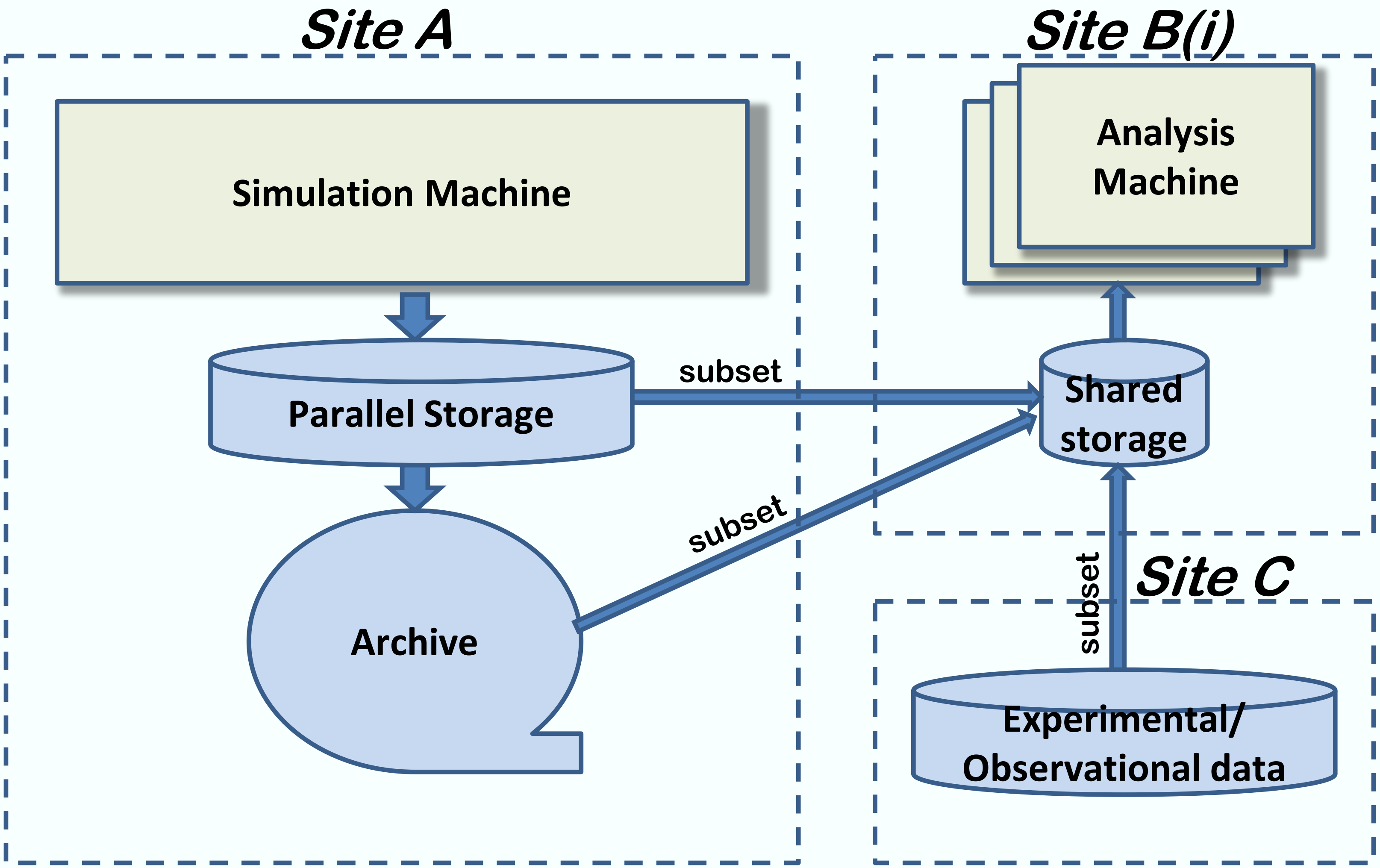
Outline

- Data challenges in the extreme scale
  - The data volume and I/O challenge
  - The data analysis challenge
  - The energy reduction challenge
- Overview of successful technologies in the SDM center
  - High Performance Technologies
  - Usability and effectiveness
  - Enabling Data Understanding
- Implications from SDM center experience
  - Techniques that could be adapted to extreme scale
- Summary of approaches to extreme scale challenges
  - With examples of approaches already being developed

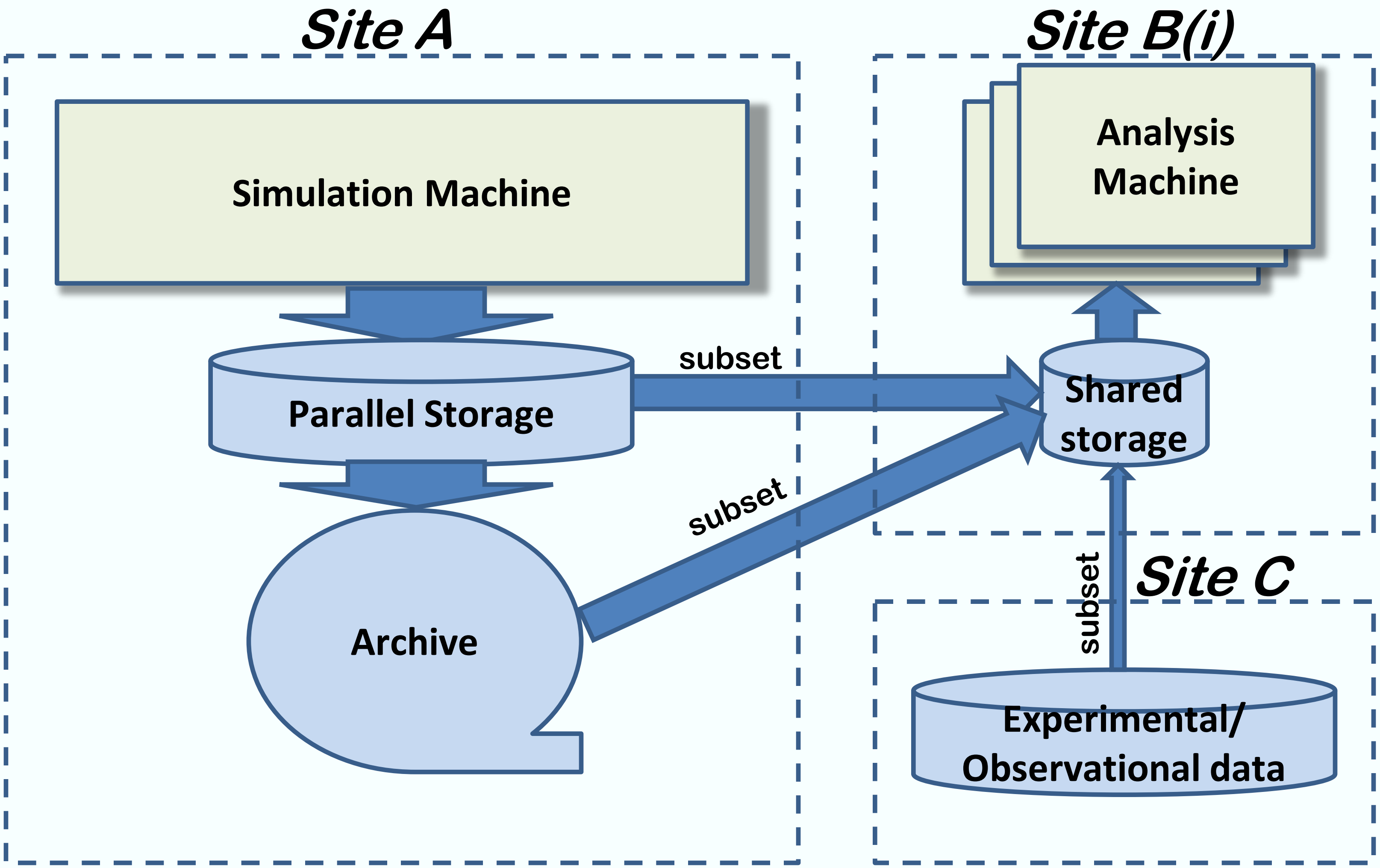
What is Scientific Data Management?

- 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)

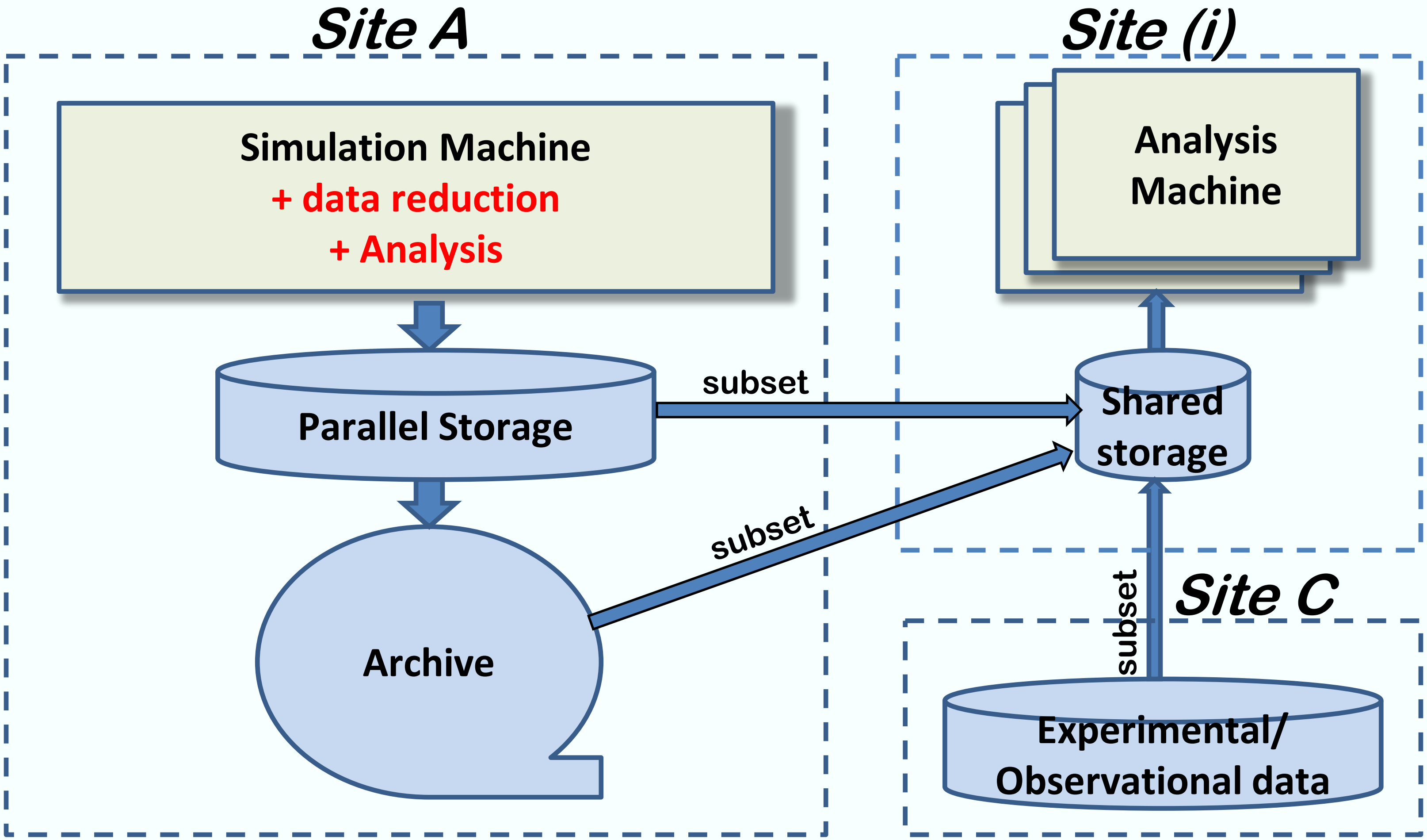


At Exascale (PBs) – data volume challenge



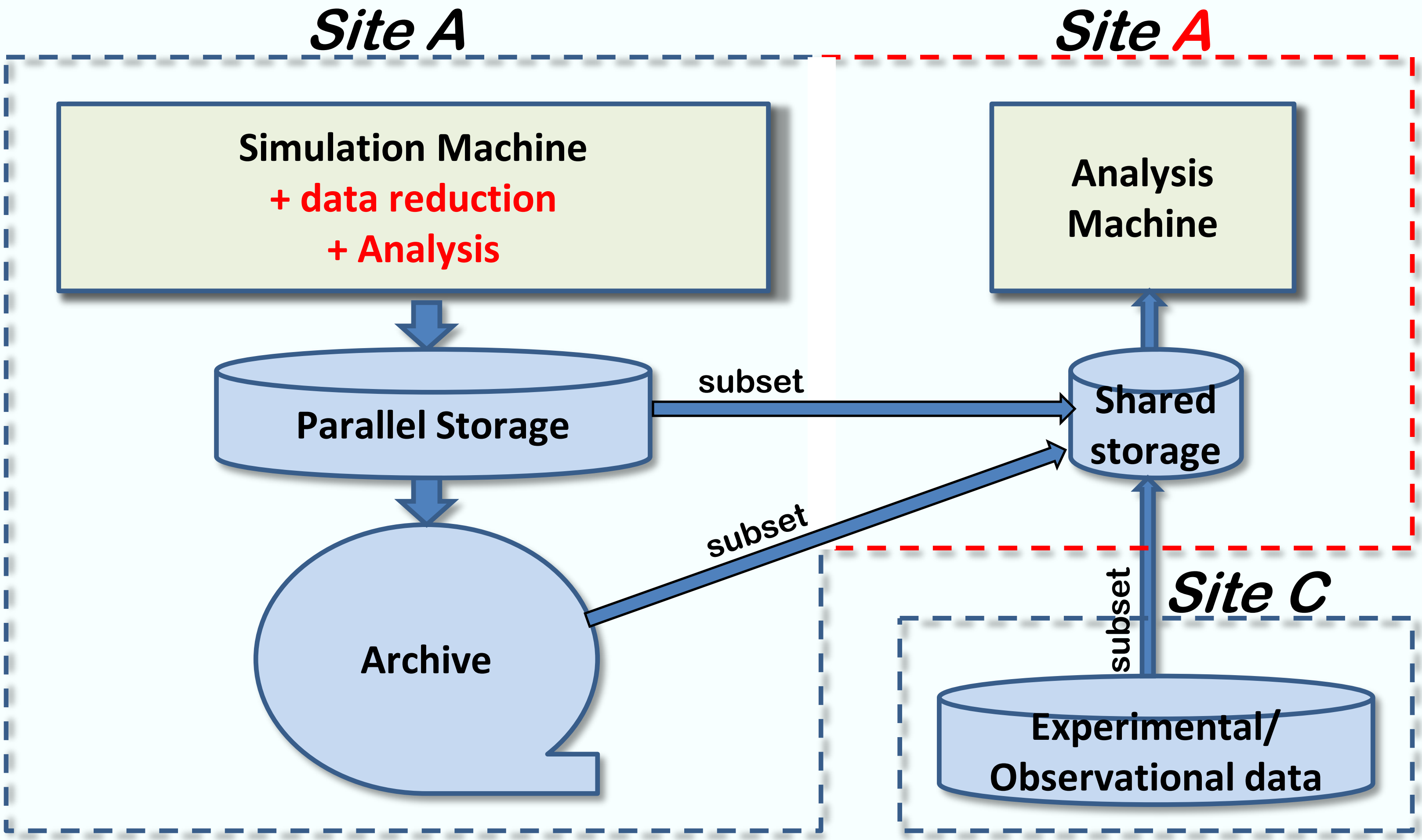
What Can be Done?

- 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



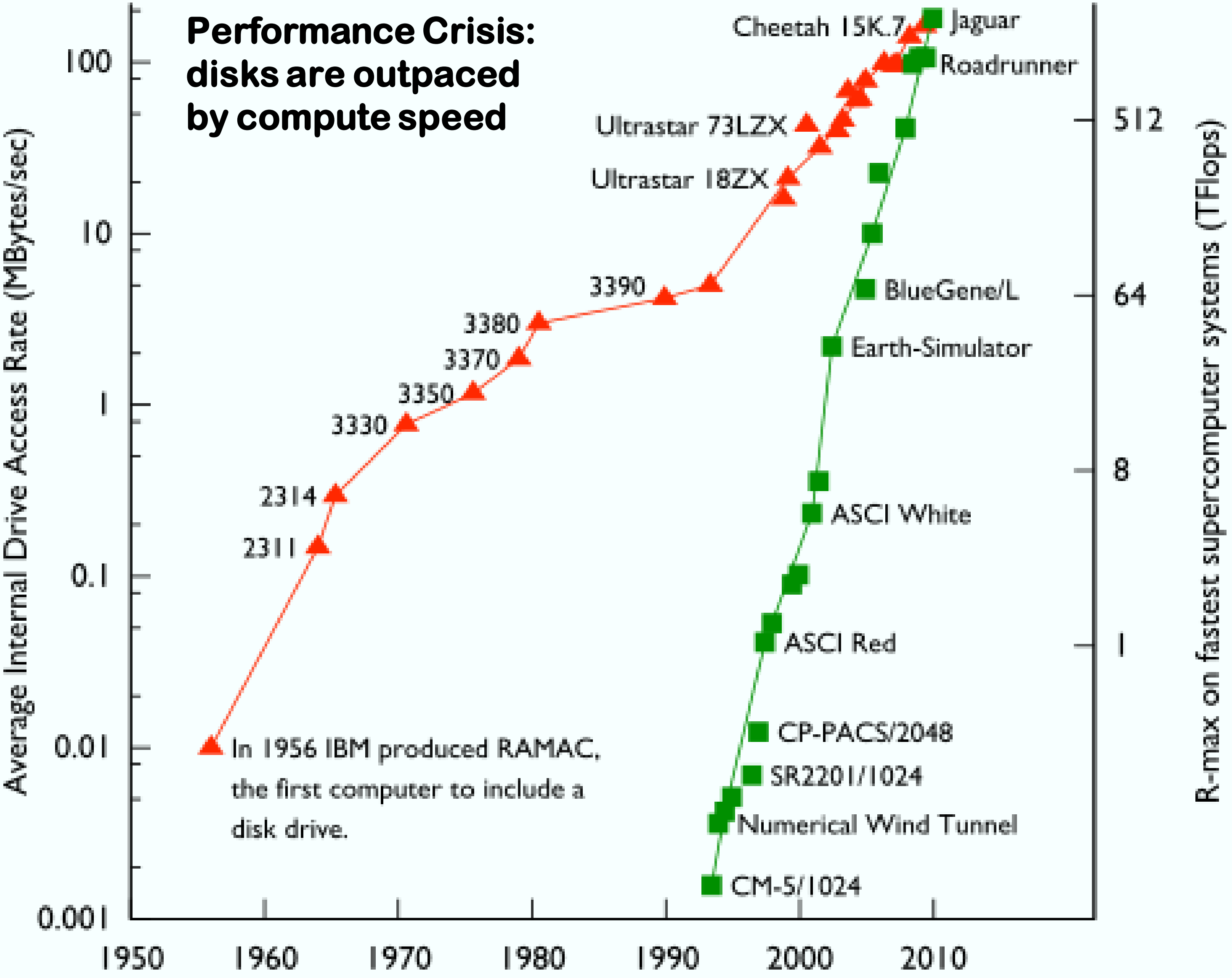
Data Analysis

- 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 mid-size 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


Performance Crisis: disks are outpaced by compute speed

In 1956 IBM produced RAMAC, the first computer to include a disk drive.

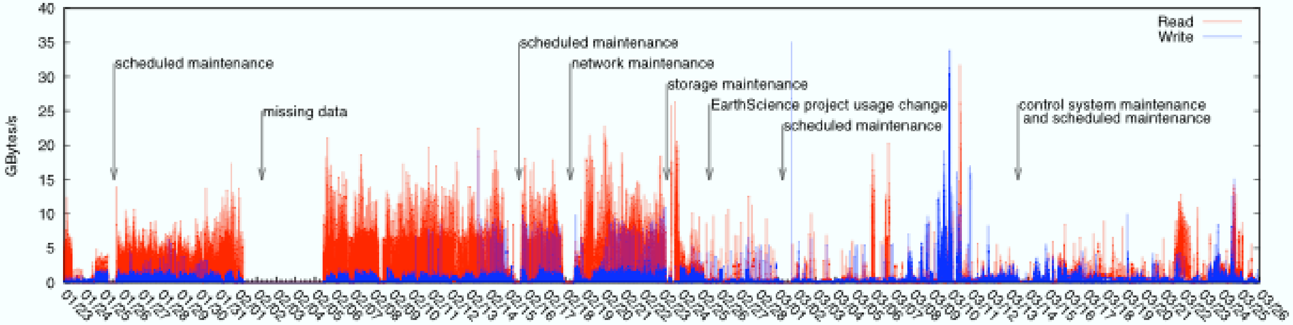
Exascale Systems: Potential Architecture

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.

Asynchronous Data Staging

Bursty I/O patterns result in periods of low activity, time we could be use to move data asynchronously



- Technique now used: move data asynchronously
  - reduce peak I/O demands from applications
  - E.g., by asynchronously staging checkpoints to storage, computation could continue while data moves.
- Questions remain in integrating into exascale systems:
  - Where do we store the data before we have a chance to move it?
  - What software layer is responsible for data movement?
  - How do we drive asynchronous data movement in exascale systems?
- Ideas for the exascale
  - Perform additional tasks during I/O at dedicated “staging nodes” or “I/O nodes”
  - e.g. statistical summaries, other data reduction, index building, generate plots, …

The energy challenge

- Data movement within the exascale machine
  - Future systems are designed to have low-power chips, but …
  - 70% of the exascale system power consumption will be memory and data movement*
    - On node: 5 levels memory (registers, L1-L3 cache, main memory) – 10’s of cycles access time
    - Energy cost of data movement to memory = ~200 X relative to L1
    - Between nodes: 100’s of cycles
    - Hard drive: 10,000’s cycles
  - Ideas for exascale
    - Minimize data movement between nodes
    - Take advantage of NVRAM (SSD) to minimize I/O
    - Redesign analysis codes (not only simulation codes) to take advantage of L1 – needs programming support

* From A. Geist, “Paving the Roadmap to EXASCALE,” SciDAC Review, NUMBER 16 Special Issue 2010.

The energy challenge

- Data movement on analysis machines
  - Assuming a reasonable number of analysis (cloud) machines (e.g. 20)
  - Then, having a large number of spinning disks wastes energy
    - 10 watts – energy cost of a disk idle spinning
    - 20 sites x 100K disks = 20 MV (same as big machine)
  - Ideas for exascale
    - Intelligent spin down of disks based on access patterns
    - Take advantage of NVRAM
    - Continue to use tape for deep archive

* From A. Geist, “Paving the Roadmap to EXASCALE,” SciDAC Review, NUMBER 16 Special Issue 2010.

Overview of successful technologies in the SDM center

http://sdmcenter.lbl.gov

Arie Shoshani (PI)

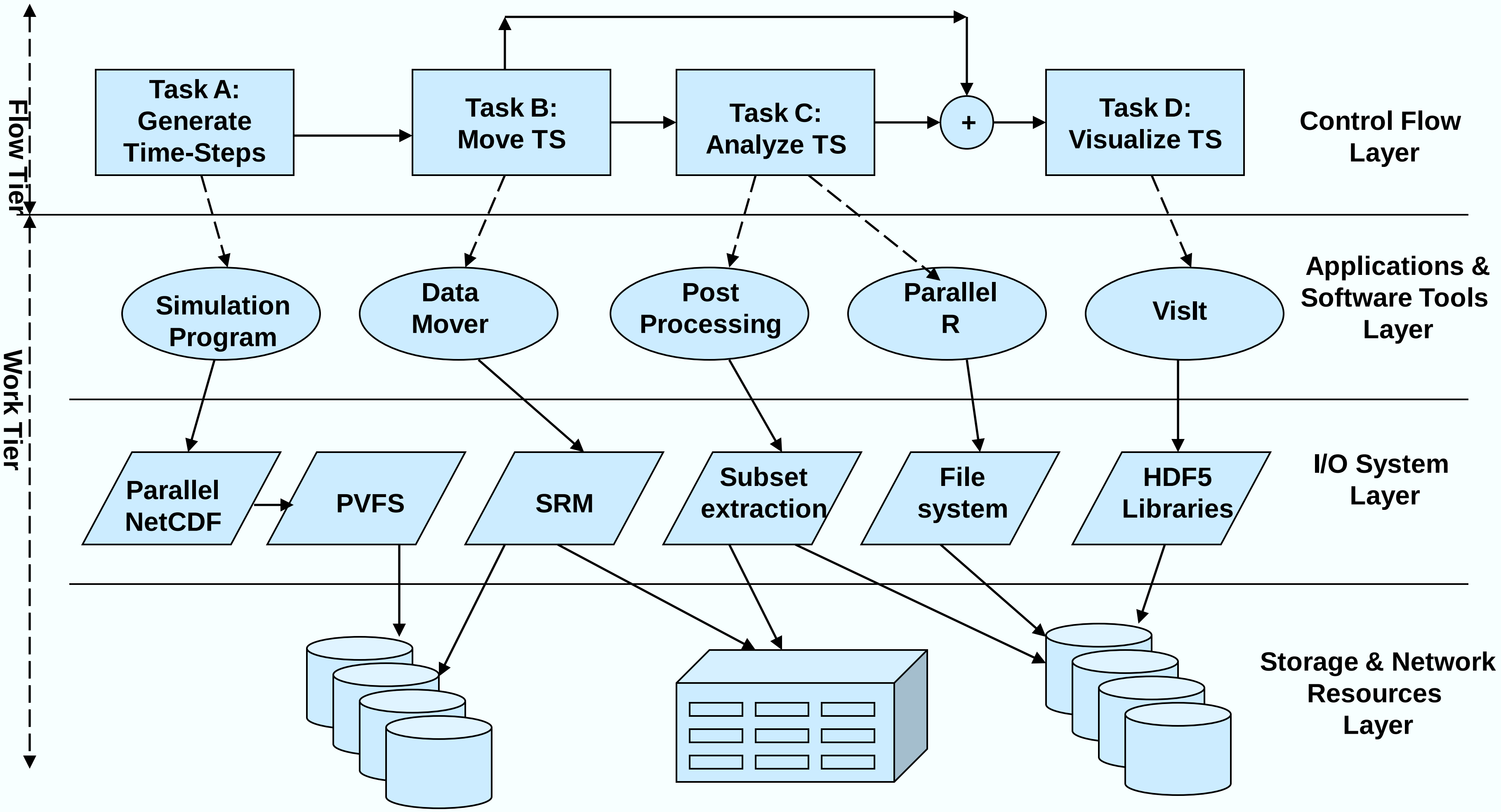
Co-Principal Investigators

DOE Laboratories		Universities	
ANL:	Rob Ross	NCSU:	Mladen Vouk
LBNL:	Doron Rotem	NWU:	Alok Choudhary
LLNL:	Chandrika Kamath	UCD:	Bertram Ludaescher
ORNL:	Nagiza Samatova	SDSC:	Ilkay Altintas
PNNL:	Terence Critchlow	UUtah:	Claudio Silva

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**

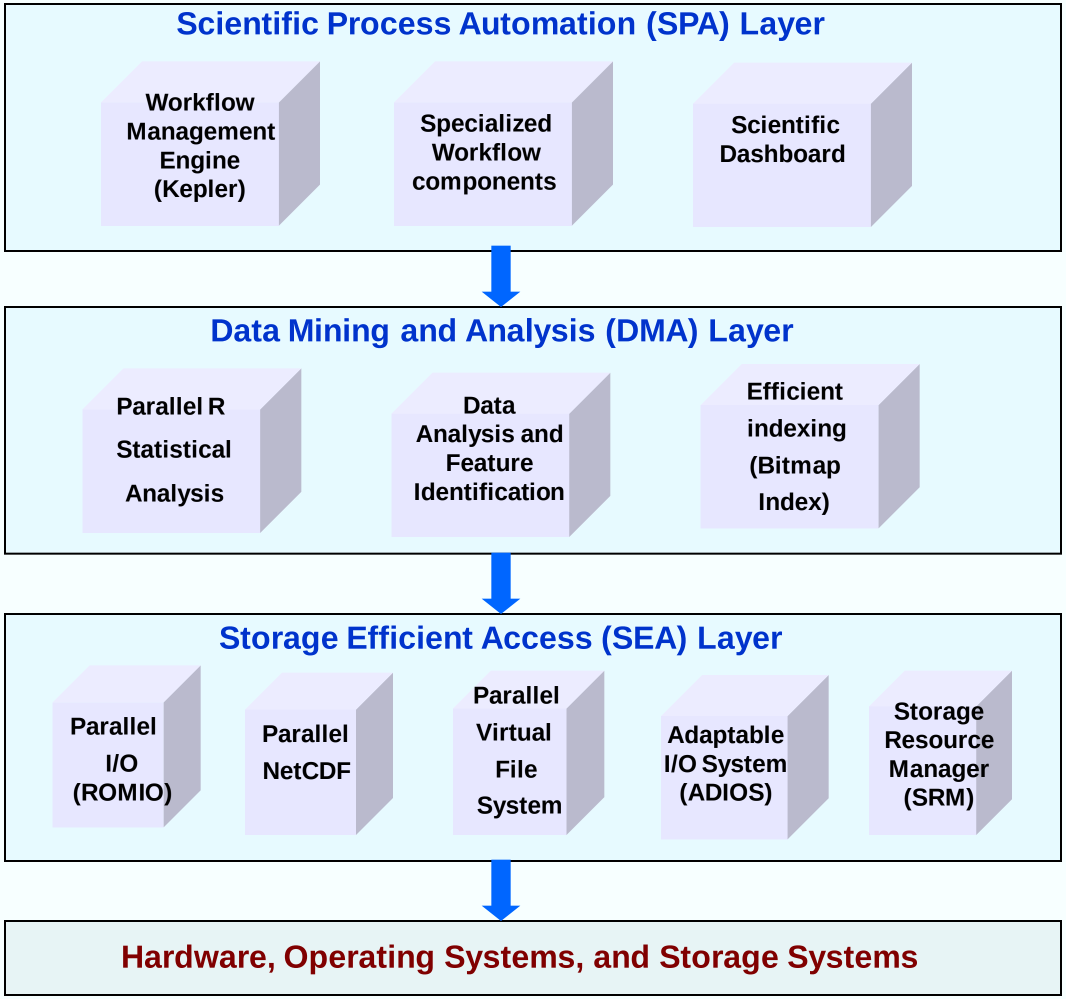
A motivating SDM Scenario (dynamic monitoring)



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


Scientific Process Automation (SPA) Layer
Workflow Management Engine (Kepler)
Specialized Workflow components
Scientific Dashboard
Data Mining and Analysis (DMA) Layer
Parallel R Statistical Analysis
Data Analysis and Feature Identification
Efficient indexing (Bitmap Index)
Storage Efficient Access (SEA) Layer
Parallel I/O (ROMIO)
Parallel NetCDF
Parallel Virtual File System
Adaptable I/O System (ADIOS)
Storage Resource Manager (SRM)
Hardware, Operating Systems, and Storage Systems

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
- Establish dialog with scientists
  - partner with scientists,
  - education (students, scientists)

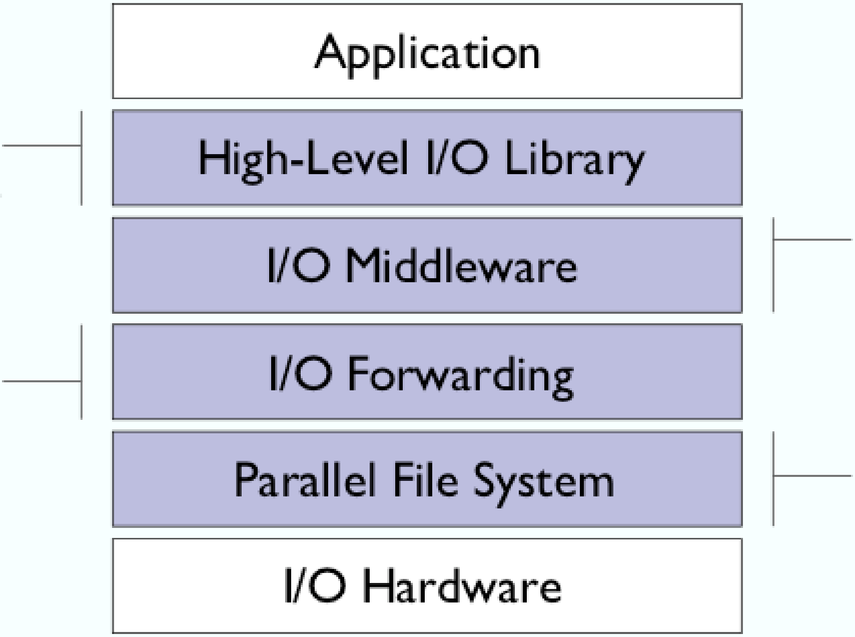
Results

✓ High Performance Technologies

Usability and effectiveness

Enabling Data Understanding

The I/O Software Stack



High-Level I/O Library
maps application abstractions onto storage abstractions and provides data portability.

HDF5, Parallel netCDF, ADIOS

I/O Forwarding
bridges between app. tasks and storage system and provides aggregation for uncoordinated I/O.

IBM ciod

I/O Middleware
organizes accesses from many processes, especially those using collective I/O.

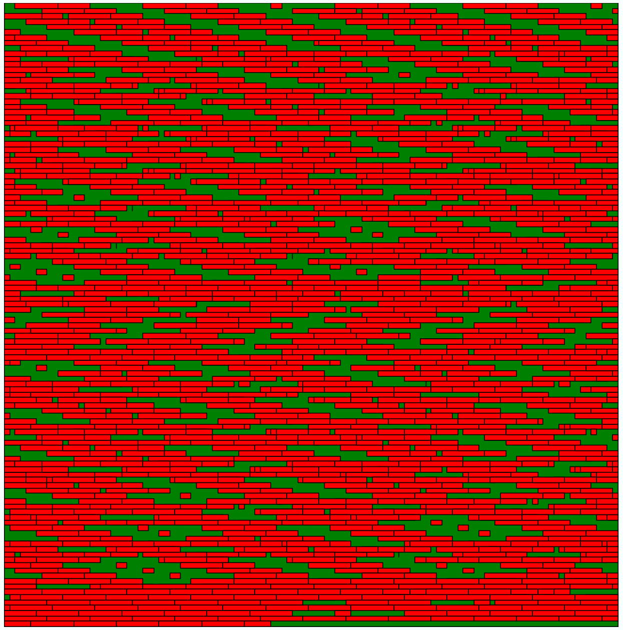
MPI-IO

Parallel File System
maintains logical space and provides efficient access to data.

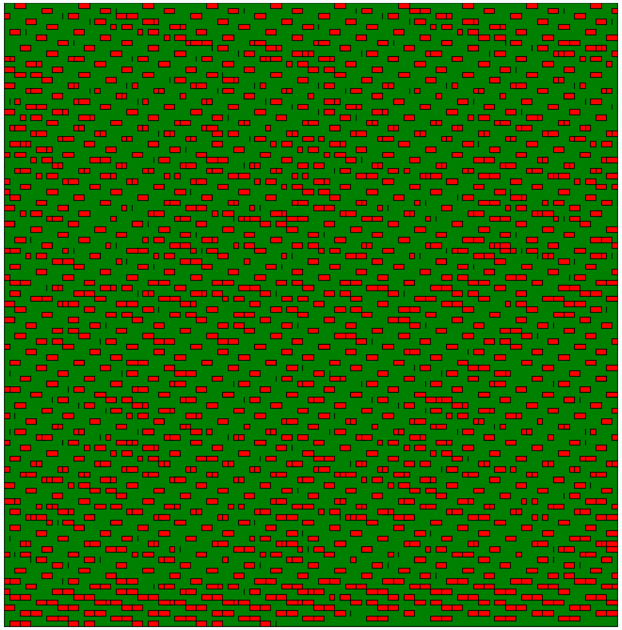
PVFS, PanFS, GPFS, Lustre

Visualizing and Tuning I/O Access

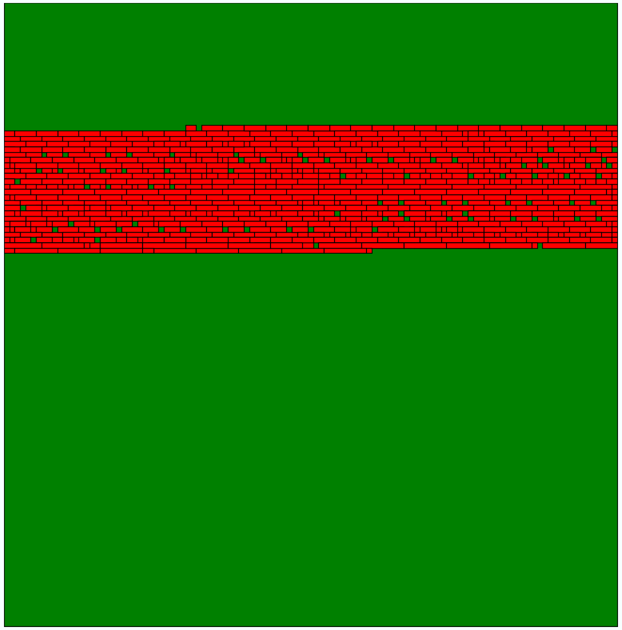
This view shows the entire 28 Gbyte dataset as a 2D array of blocks, for three separate runs. Renderer is visualizing one variable out of five. Red blocks were accessed. Access times in parenthesis.



Original Pattern



MPI-IO Tuning



PnetCDF Enhancements

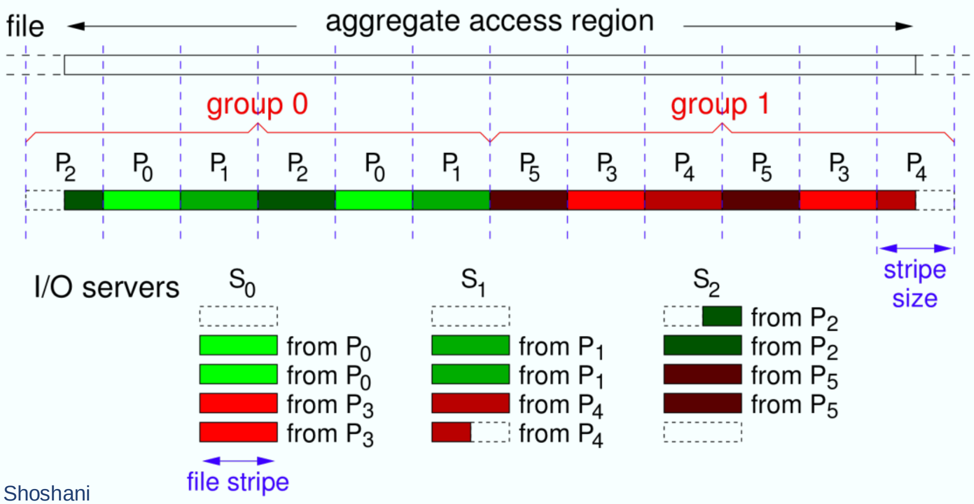
Data is stored in the netCDF “record” format, where variables are interleaved in file (36.0 sec). Adjusting MPI-IO parameters (right) resulted in significant I/O reduction (18.9 sec).

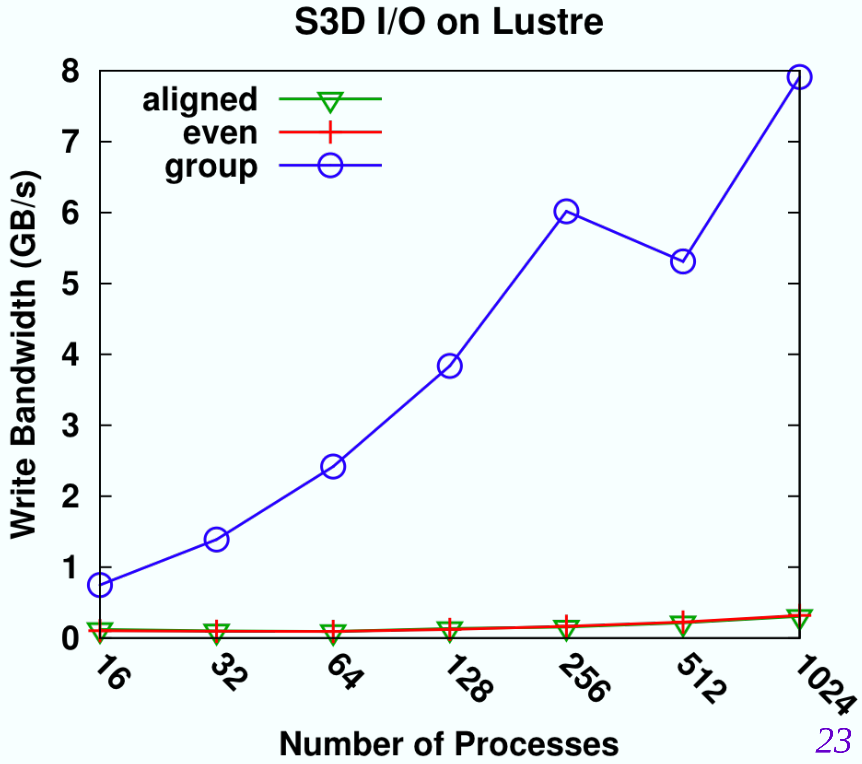
New PnetCDF large variable support stores data contiguously (13.1 sec).

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.

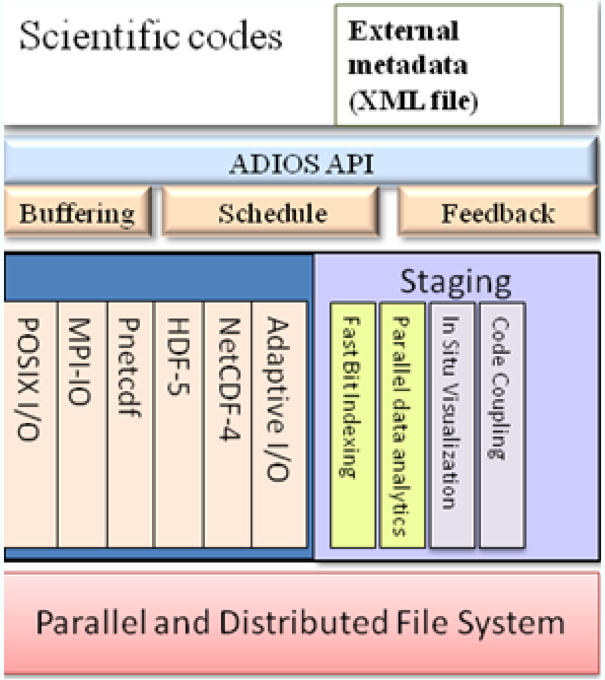




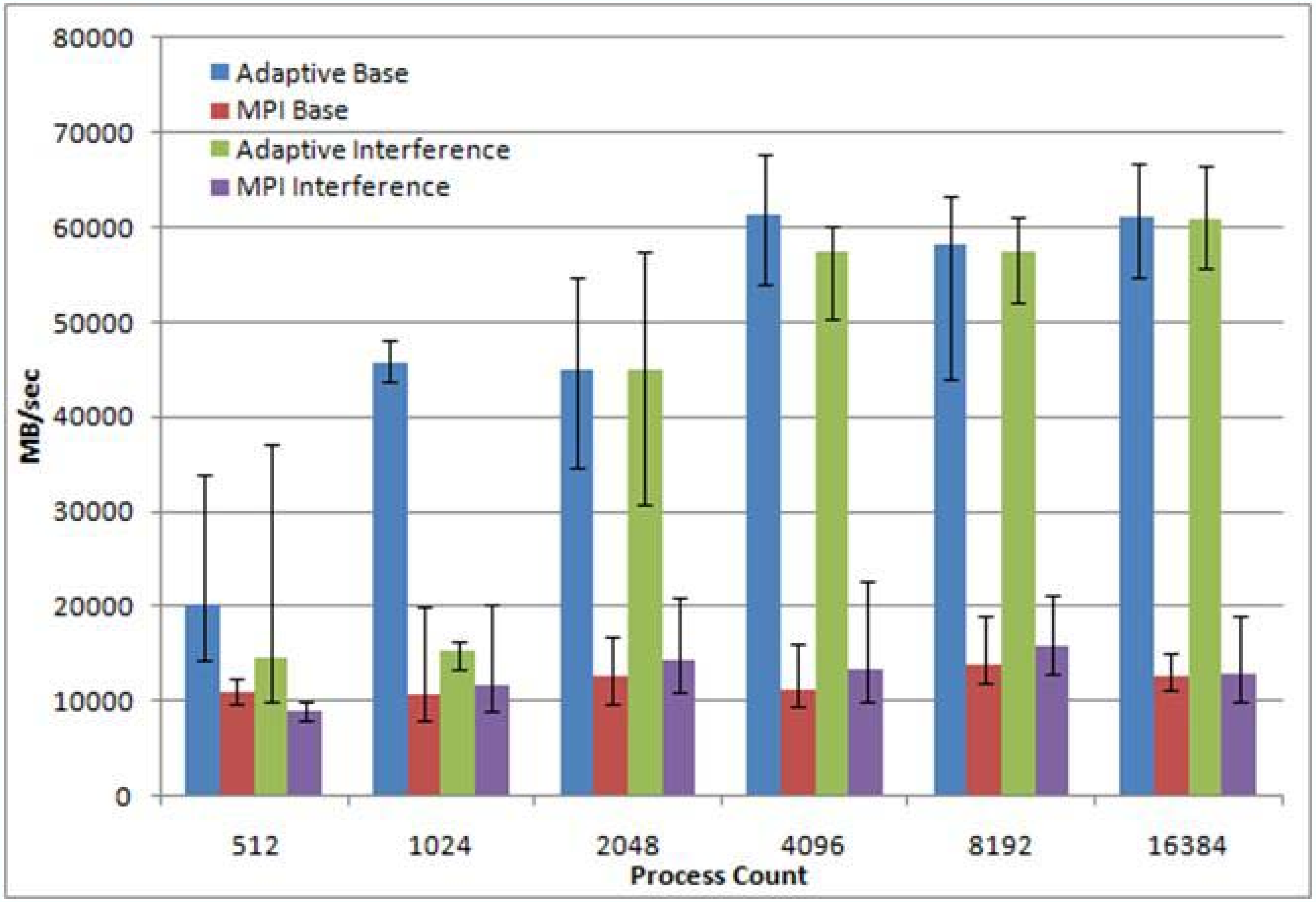
ADaptable IO System (ADIOS)

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

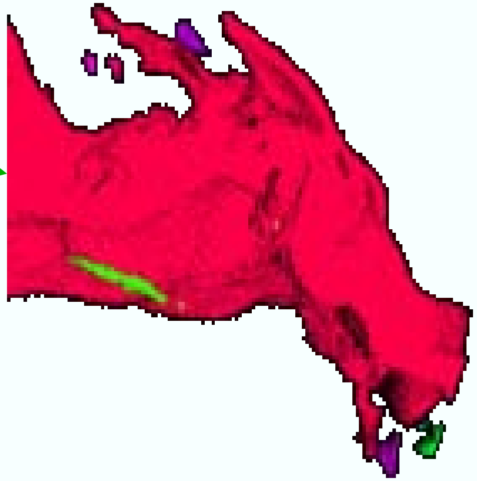


Performance of I/O from the Pixe3D (MHD) fusion code using ADIOS



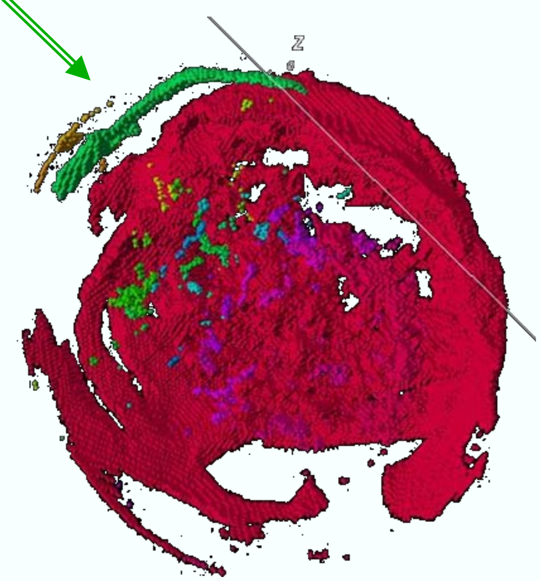
Searching Problems in Data Intensive Sciences

- 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.



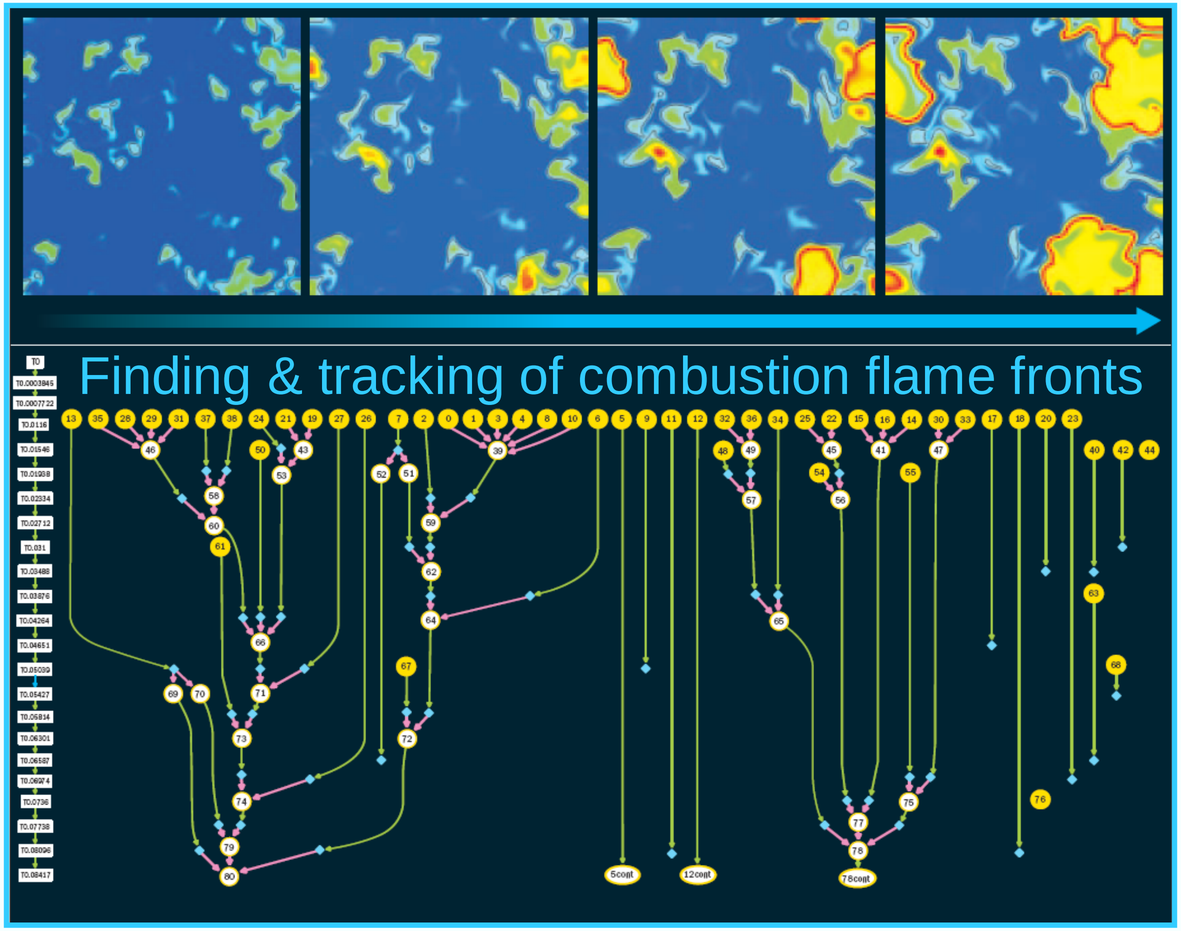
- Other application domains:
  - real-time analysis of network intrusion attacks
  - fast tracking of combustion flame fronts over time
  - accelerating molecular docking in biology applications
  - query-driven visualization

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}”

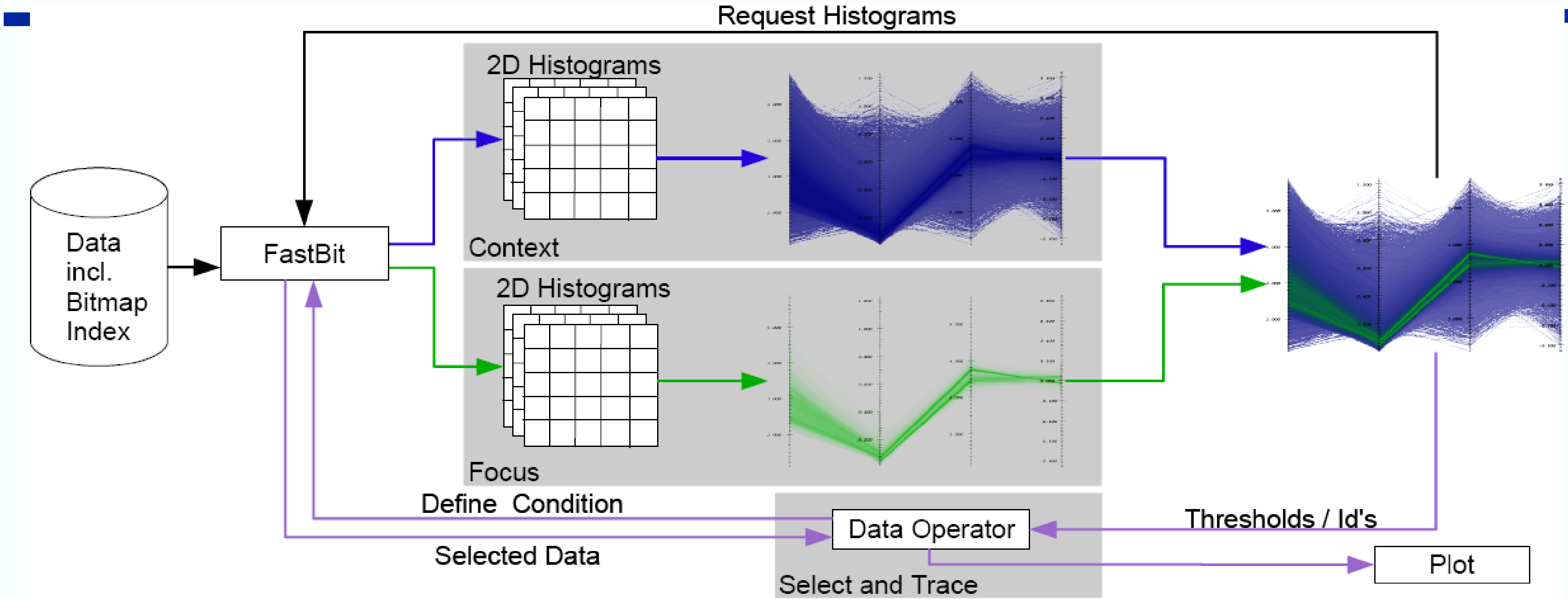
Region growing

Connect neighboring cells into regions

Region tracking

Track the evolution of the features through time

Query-Driven Visualization



- **Collaboration between SDM and 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)**

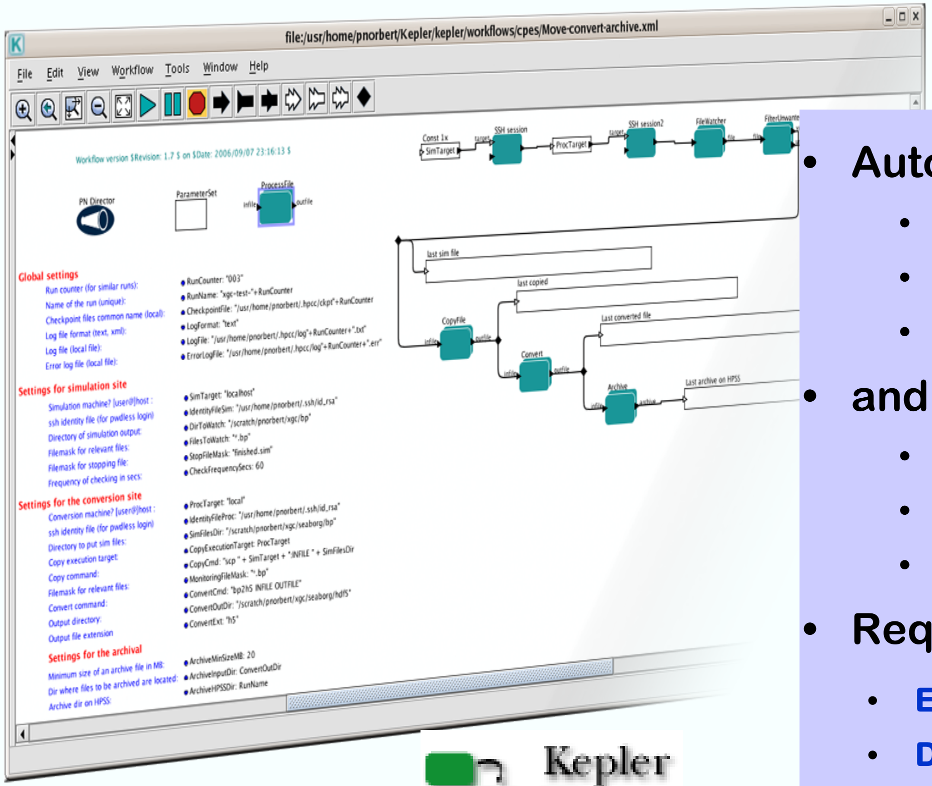
Results

High Performance Technologies

✓ Usability and effectiveness

Enabling Data Understanding

Workflow automation requirements in Fusion Plasma Edge Simulation



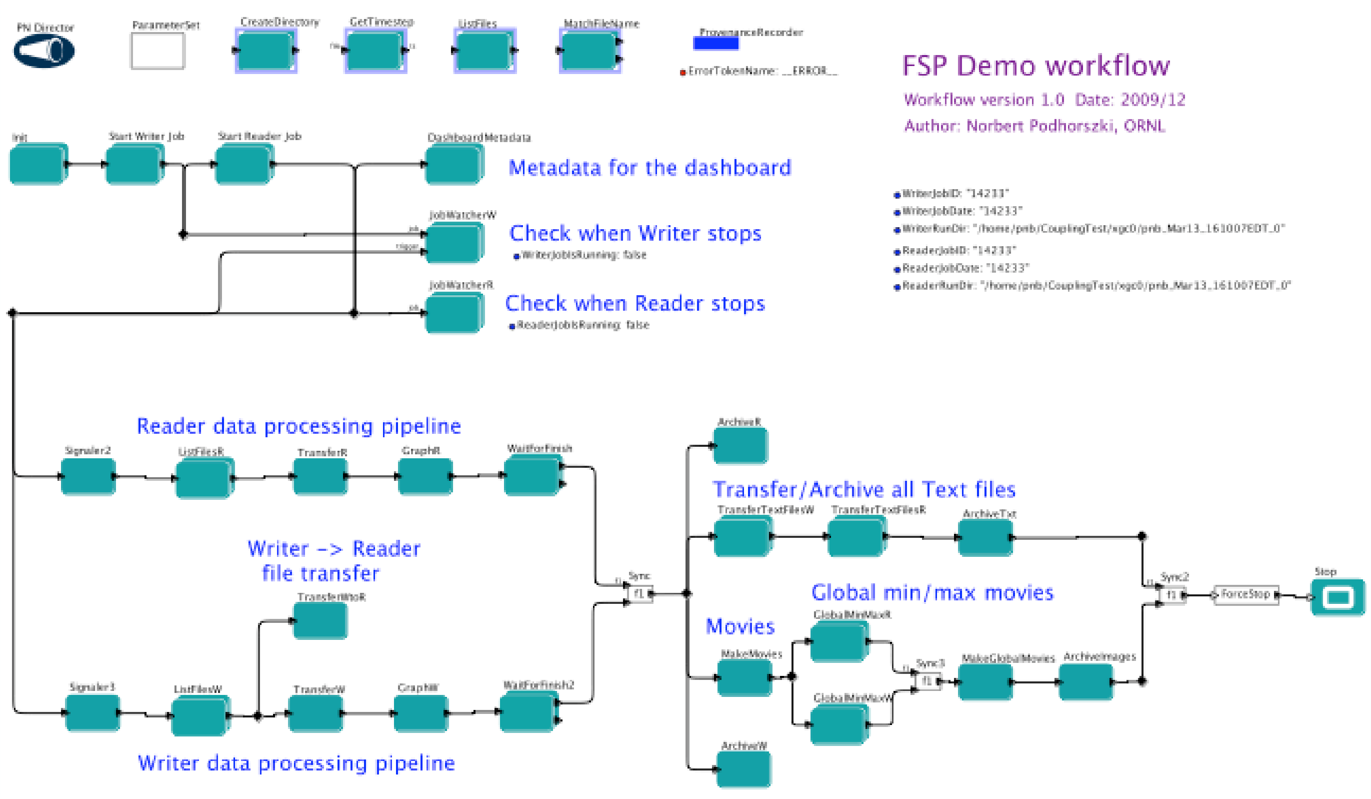
Workflow version $Revision: 1.7 $ on $Date: 2006/09/07 23:16:13 $

Global settings	
Run counter (for similar runs):	RunCounter: "003"
Name of the run (unique):	RunName: "xgc-test-"+RunCounter
Checkpoint files common name (local):	CheckpointFile: "/usr/home/pnorbert/.hpcc/ckpt"+RunCounter
Log file format (text, xml):	LogFormat: "text"
Log file (local file):	LogFile: "/usr/home/pnorbert/.hpcc/log"+RunCounter+".txt"
Error log file (local file):	ErrorLogFile: "/usr/home/pnorbert/.hpcc/log"+RunCounter+".err"
Settings for simulation site	
Simulation machine? [user@]host :	SimTarget: "localhost"
ssh identity file (for pwdless login)	IdentityFileSim: "/usr/home/pnorbert/.ssh/id_rsa"
Directory of simulation output:	DirToWatch: "/scratch/pnorbert/xgc/bp"
Filemask for relevant files:	FilesToWatch: "*.bp"
Filemask for stopping file:	StopFileMask: "finished.sim"
Frequency of checking in secs:	CheckFrequencySecs: 60
Settings for the conversion site	
Conversion machine? [user@]host :	ProcTarget: "local"
ssh identity file (for pwdless login)	IdentityFileProc: "/usr/home/pnorbert/.ssh/id_rsa"
Directory to put sim files:	SimFilesDir: "/scratch/pnorbert/xgc/seaborg/bp"
Copy execution target:	CopyExecutionTarget: ProcTarget
Copy command:	CopyCmd: "scp " + SimTarget + ":INFILE " + SimFilesDir
Filemask for relevant files:	MonitoringFileMask: "*.bp"
Convert command:	ConvertCmd: "bp2h5 INFILE OUTFILE"
Output directory:	ConvertOutDir: "/scratch/pnorbert/xgc/seaborg/hdf5"
Output file extension	ConvertExt: "h5"
Settings for the archival	
Minimum size of an archive file in MB:	ArchiveMinSizeMB: 20
Dir where files to be archived are located:	ArchiveInputDir: ConvertOutDir
Archive dir on HPSS:	ArchiveHPSSDir: RunName

- 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

Contact: Scott Klasky, et. al, ORNL

The Kepler Workflow Engine



FSP Demo workflow

Workflow version 1.0 Date: 2009/12

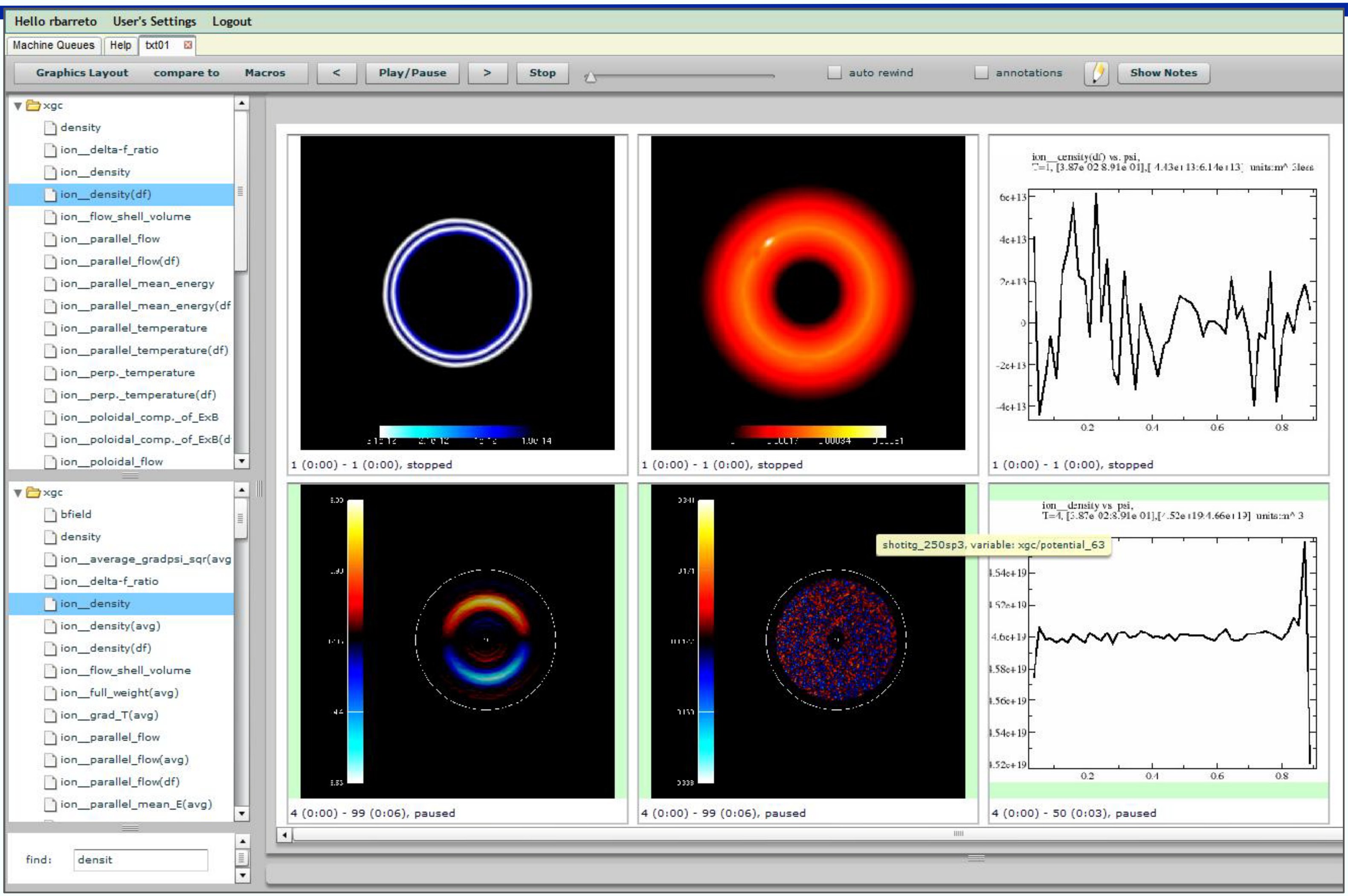
Author: Norbert Podhorszki, ORNL

- WriterJobID: "14233"
- WriterJobDate: "14233"
- WriterRunDir: "/home/pnb/CouplingTest/xgc0/pnb_Mar13_161007EDT_0"
- ReaderJobID: "14233"
- ReaderJobDate: "14233"
- ReaderRunDir: "/home/pnb/CouplingTest/xgc0/pnb_Mar13_161007EDT_0"

- ErrorTokenName: __ERROR__
- WriterJobsRunning: false
- ReaderJobsRunning: false

- 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)

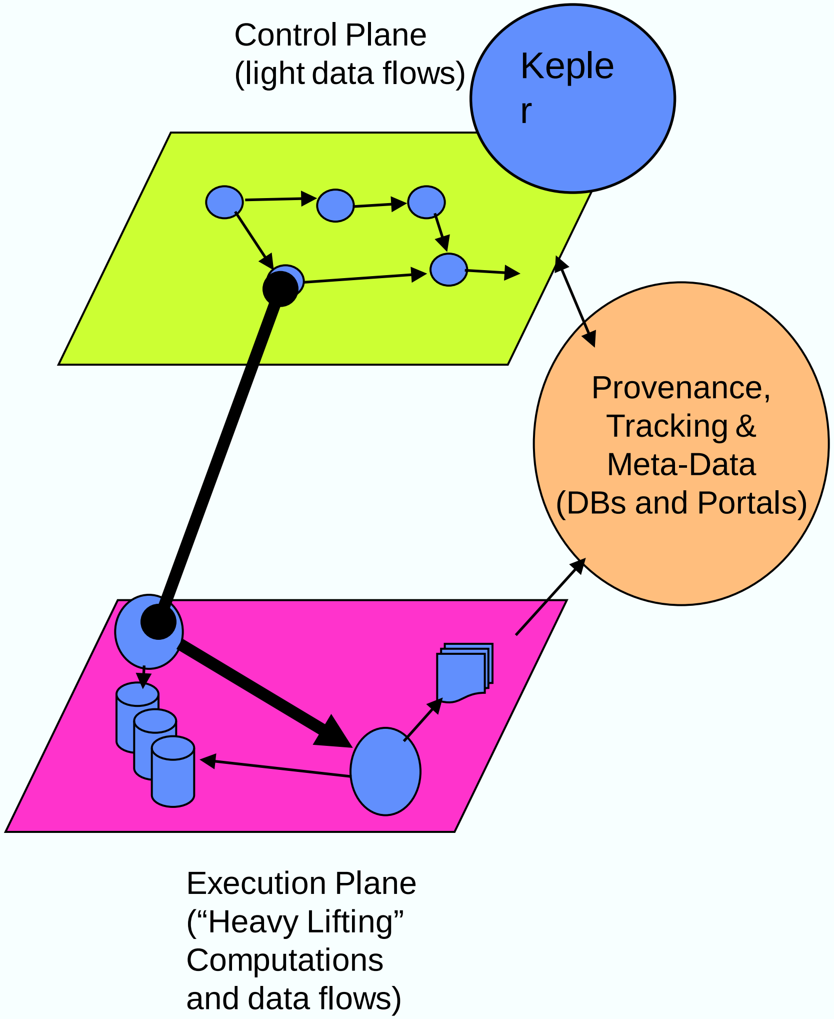
Real-time visualization and analysis capabilities on dashboard



visualize and compare shots

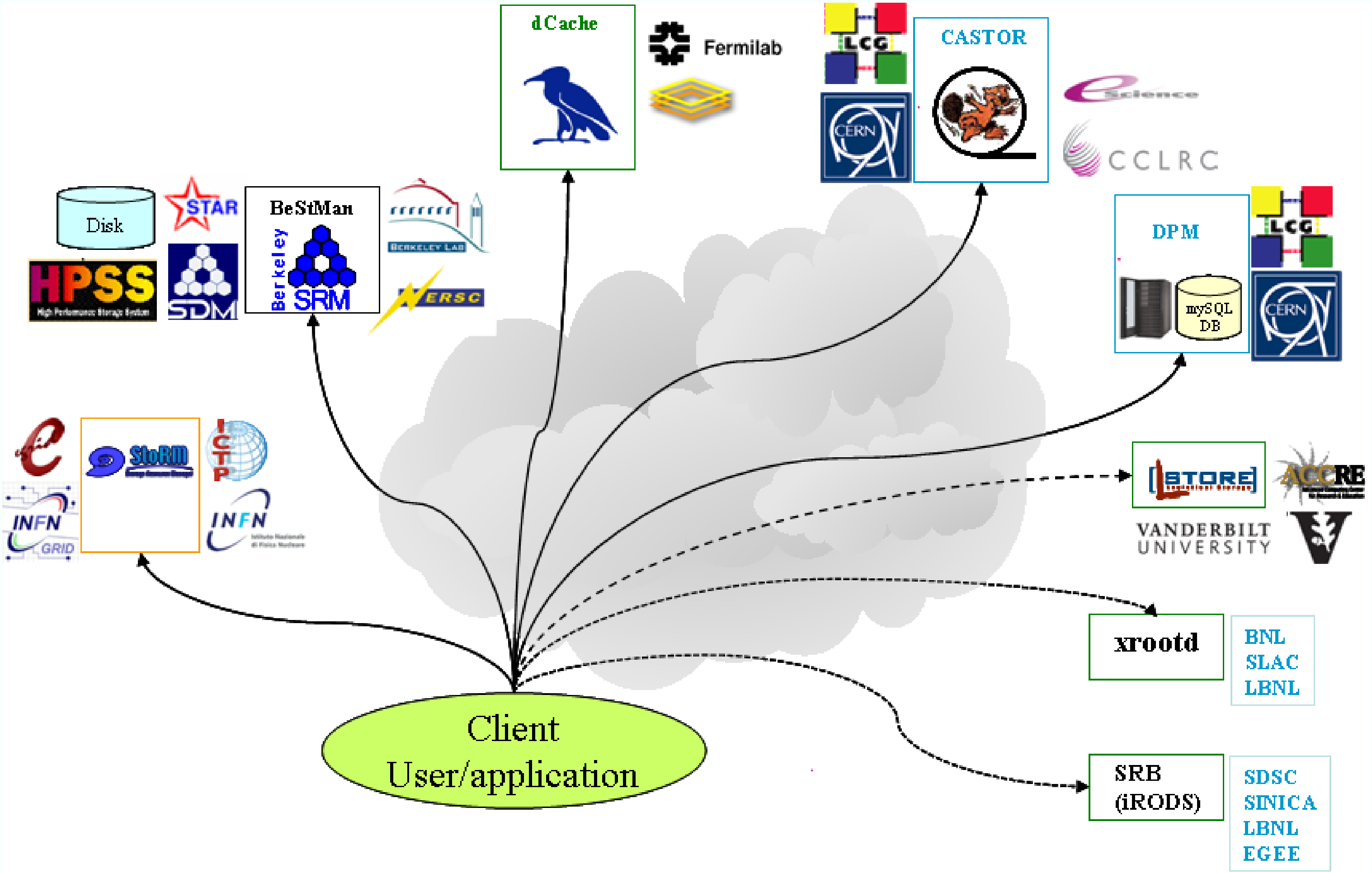
Capturing Provenance in Workflow Framework

- ***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

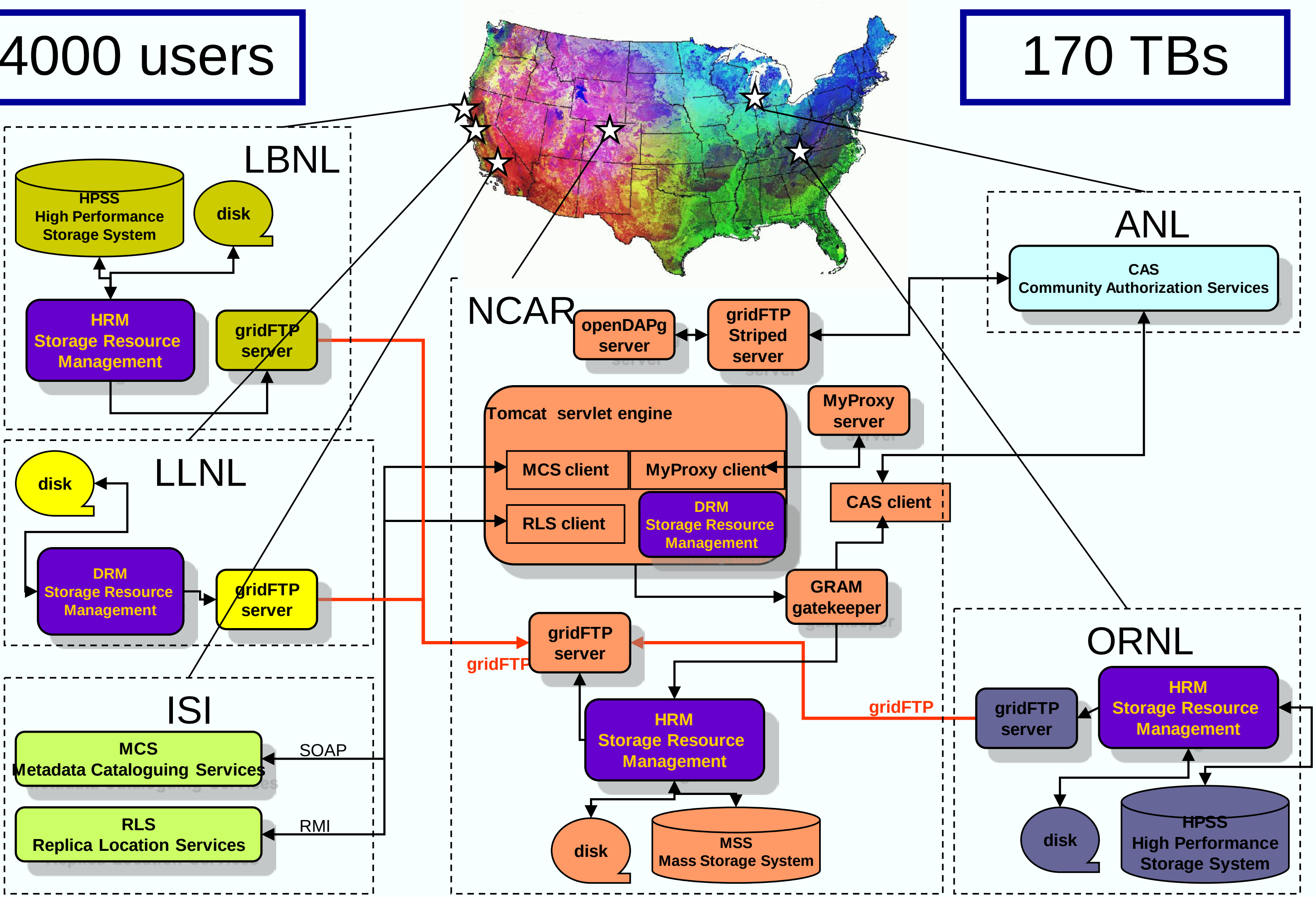
Storage Resource Managers (SRMs): Middleware for storage interoperability and data movement



SRM use in Earth Science Grid

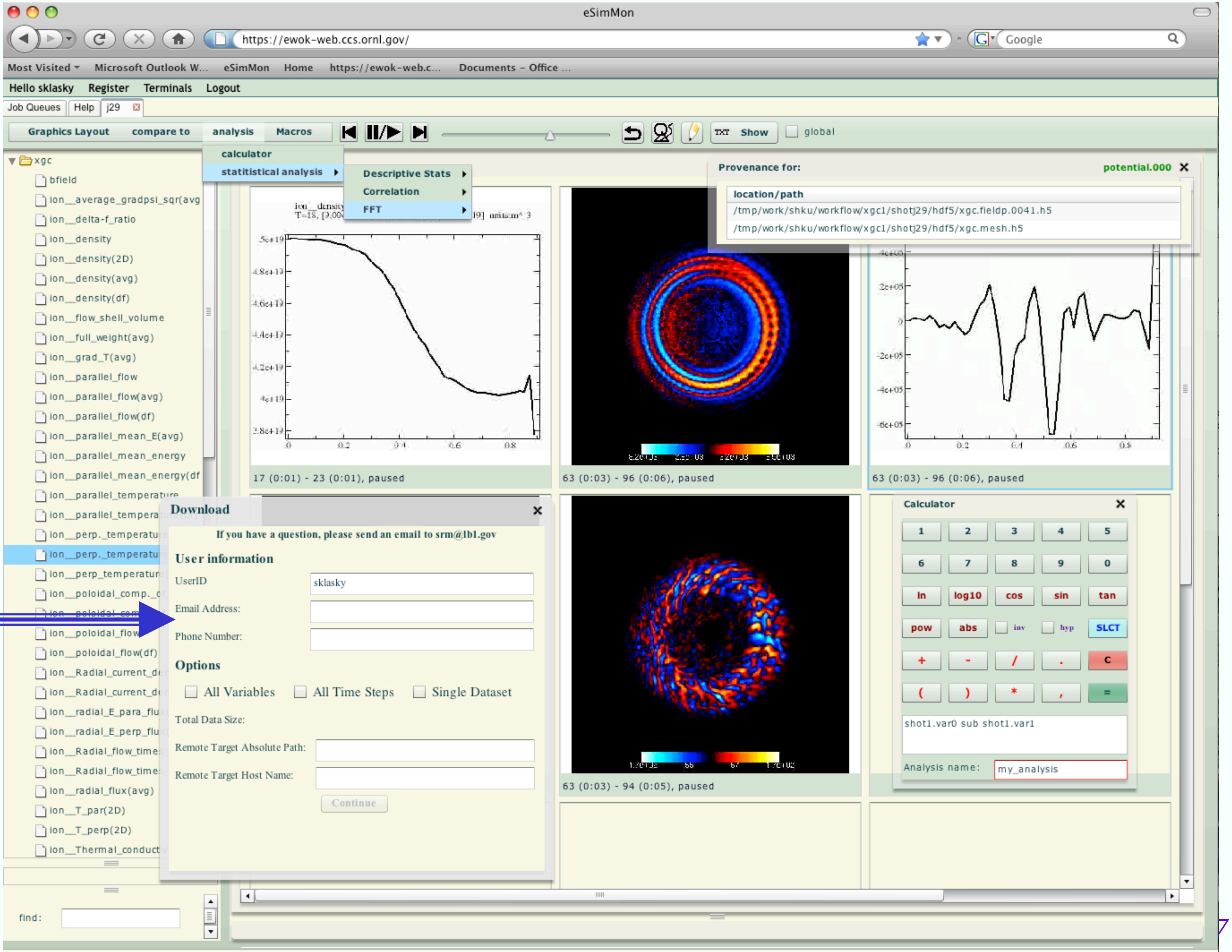
14000 users

170 TBs



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

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



Provenance for: potential.000

location/path
/tmp/work/shku/workflow/xgc1/shotJ29/hdf5/xgc.fieldp.0041.h5
/tmp/work/shku/workflow/xgc1/shotJ29/hdf5/xgc.mesh.h5

17 (0:01) - 23 (0:01), paused

63 (0:03) - 96 (0:06), paused

63 (0:03) - 96 (0:06), paused

63 (0:03) - 94 (0:05), paused

Download

If you have a question, please send an email to srm@lbl.gov

User information

UserID: sklasky

Email Address:

Phone Number:

Options

□ All Variables □ All Time Steps □ Single Dataset

Total Data Size:

Remote Target Absolute Path:

Remote Target Host Name:

Continue

shot1.var0 sub shot1.var1

Analysis name: my_analysis

Download window

Results

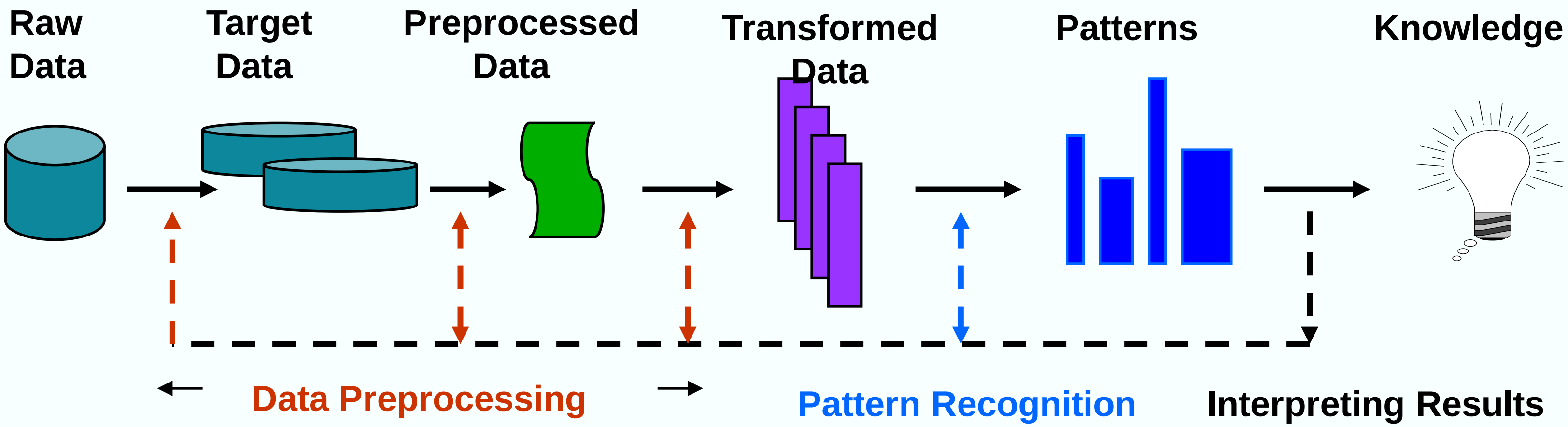
High Performance Technologies

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, …



Raw Data

Target Data

Preprocessed Data

Transformed Data

Patterns

Knowledge

Data Preprocessing

Pattern Recognition

Interpreting Results

Data Fusion, Sampling, Multi-resolution analysis

De-noising, Object - identification, Feature- extraction, Normalization

Dimension- reduction

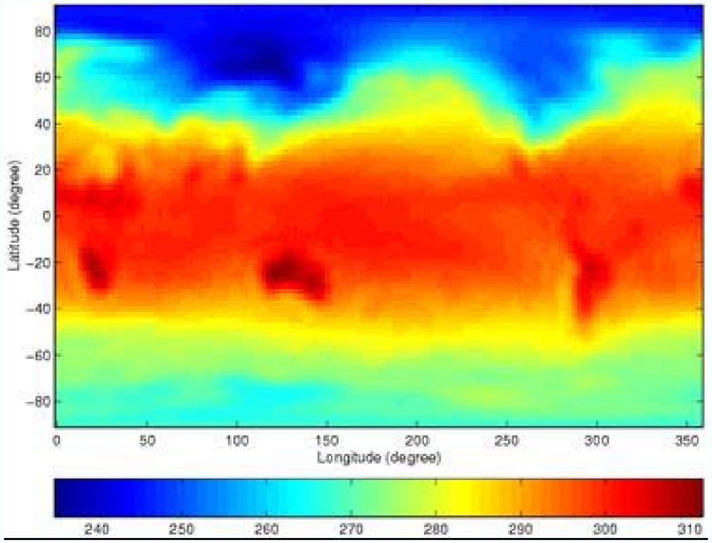
Classification, Clustering, Regression

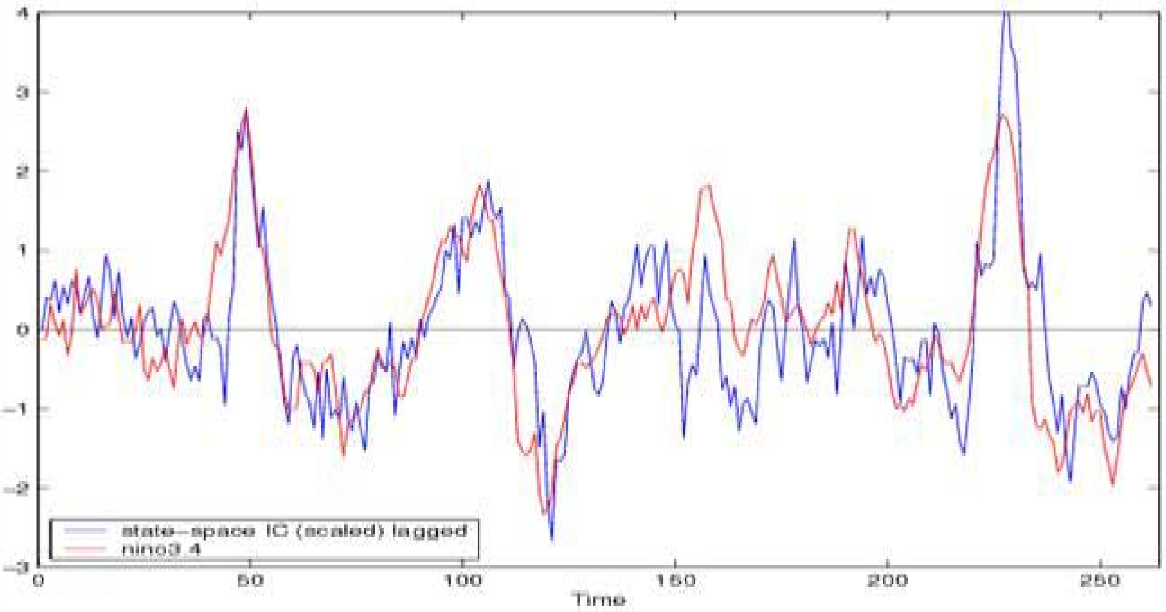
Visualization, Validation

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

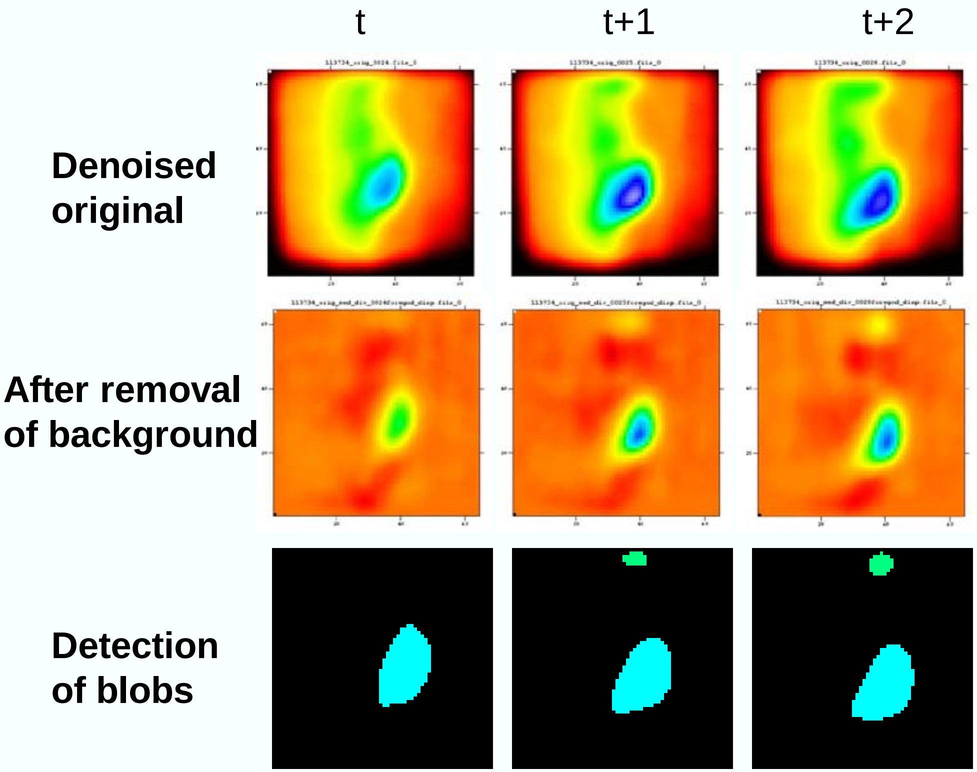




Collaboration with Ben Santer (LLNL)

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



t

t+1

t+2

Denoised original

After removal of background

Detection of blobs

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

pR: Provides Simple, Efficient MPI C/Fortran Access to R Statistical Computing Environment

From R end-user's perspective:

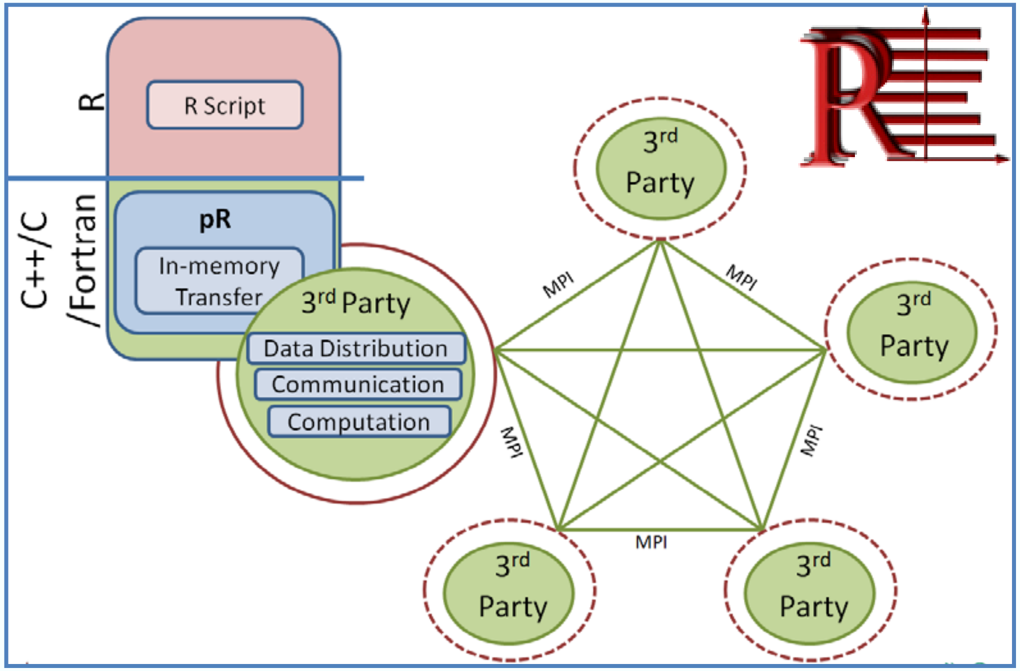
- Require ***NO*** (very trivial) changes to serial R code
- Yet deliver HPC performance

From HPC developer's perspective:

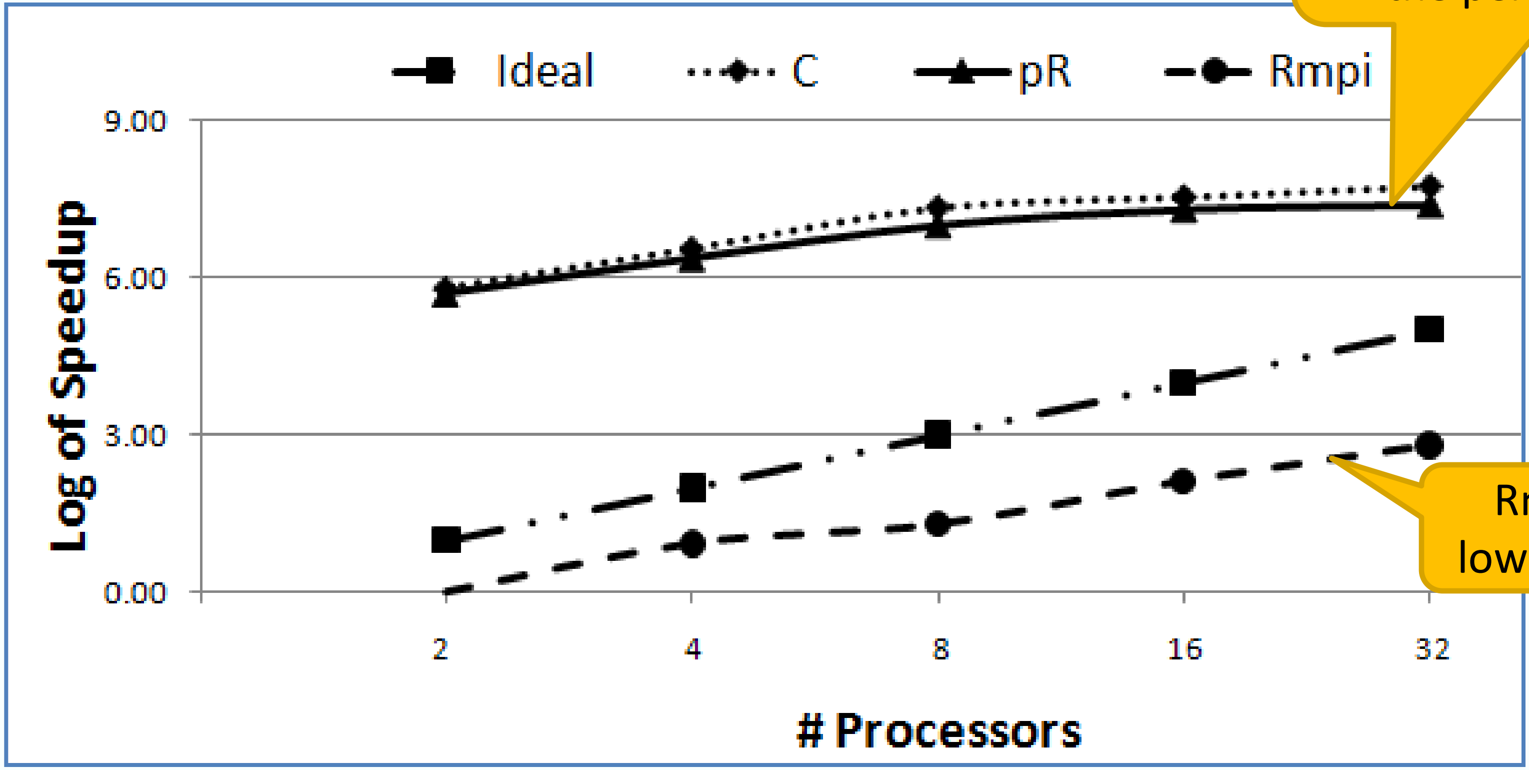
- Provide ***native*** to HPC developer interface to R internals
- With NO (constantly small) overhead

From pR implementation viewpoint:

- Tightly-coupled R interface between R and MPI C/Fortran code
- Bidirectional translation of data objects
- Direct memory access to R objects
- Compared to parallel C: induces negligible (constant) overhead



pR Offers Parallel Scripting Computing with the Same Performance as Parallel Compiled Codes



pR introduces minimal overhead & closely mirrors the performance of C

Rmpi speedup is lower than the ideal

Using a matrix-multiplication testcase on 4096 x 4096 matrices, and comparing against a serial R implementation.

Implications from SDM center experience (1)

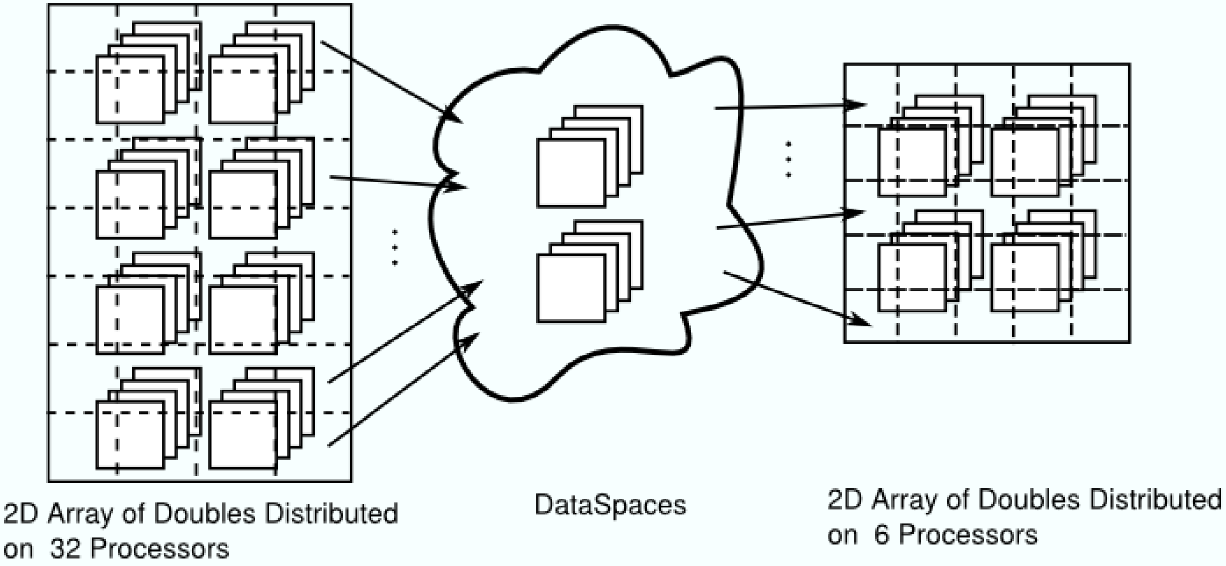
- What was successful and we believe can be done in-situ
  - monitoring of simulations
    - Use of workflow automation, and dashboard technologies
    - In center – Kepler, eSiMon dashboard
  - Provenance generation
    - Can be captured while data is generated by using workflow
    - In Center – provenance 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
  - 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

Implications from SDM center experience: Coupling codes in-memory

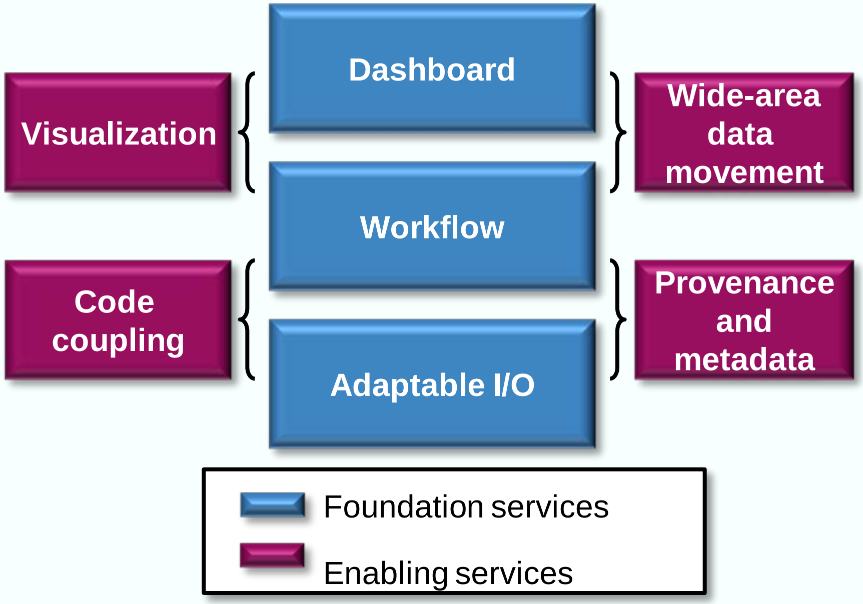
- The simulations exchange multi-dimensional data arrays (e.g., 2D)
  - Domain discretization is different for the two applications
  - Data redistribution is transparent and implicit through the space
- The simulations have different interaction patterns
  - e.g., one-to-many, many-to-many, many-to-one


2D Array of Doubles Distributed on 32 Processors
DataSpaces
2D Array of Doubles Distributed on 6 Processors

Implications from SDM center experience: integration of tools

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

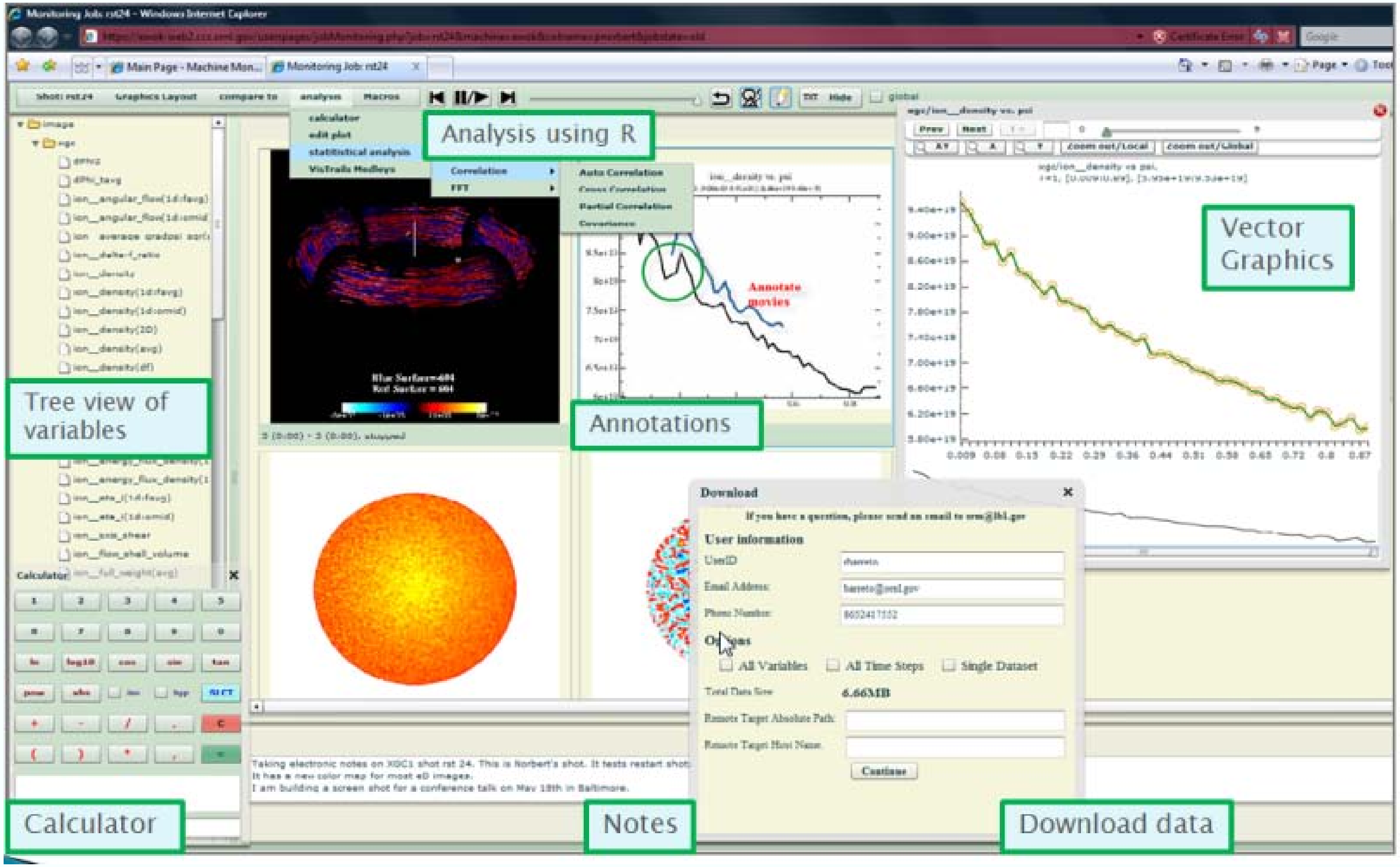
- Adaptable I/O
- Workflows
- Dashboard
- Provenance
- Code coupling
- WAN data movement
- Visualization



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



Taking electronic notes on XGC1 shot rst 24. This is Norbert's shot. It tests restart shot.
It has a new color map for most eD images.
I am building a screen shot for a conference talk on May 18th in Baltimore.

Download

If you have a question, please send an email to srm@lbl.gov

User information

UserID	rbarreto
Email Address:	barreto@ornl.gov
Phone Number:	8602417332

Options

□ All Variables □ All Time Steps □ Single Dataset

Total Data Size **6.66MB**

Remote Target Absolute Path:

Remote Target Host Name:

Continue

Data challenges area in the extreme scale

- The data volume and I/O challenge
- The data analysis challenge
- The energy reduction challenge

Approaches to the Data Volume and I/O Challenge

- Minimize volume of data to be stored
  - In-situ analysis
    - Use extra cores, GPU, 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 (pipelines)
- Reduction of data = reduction in I/O

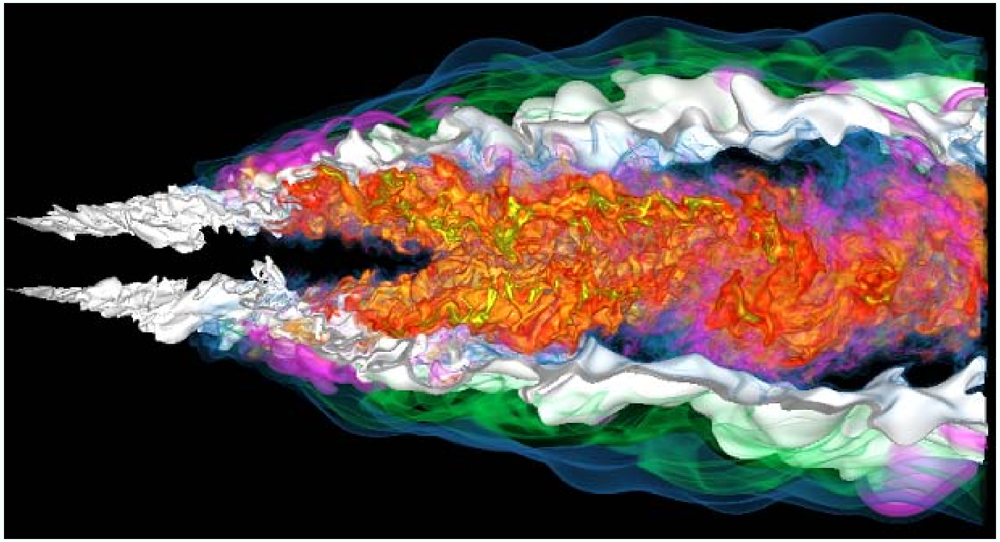
Example of In-Situ Analysis and Data Reduction

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)).

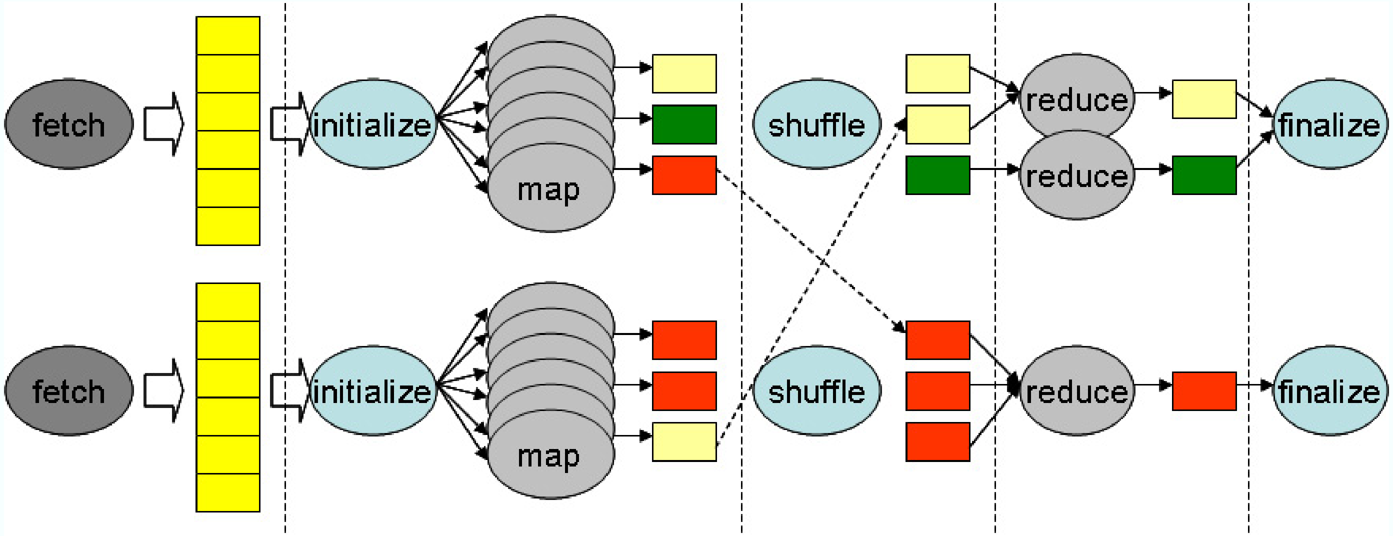
By compressing data more aggressively the further it is from this surface, we can attain a compression ratio of 20-30x while still retaining full fidelity in the vicinity of the surface.



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



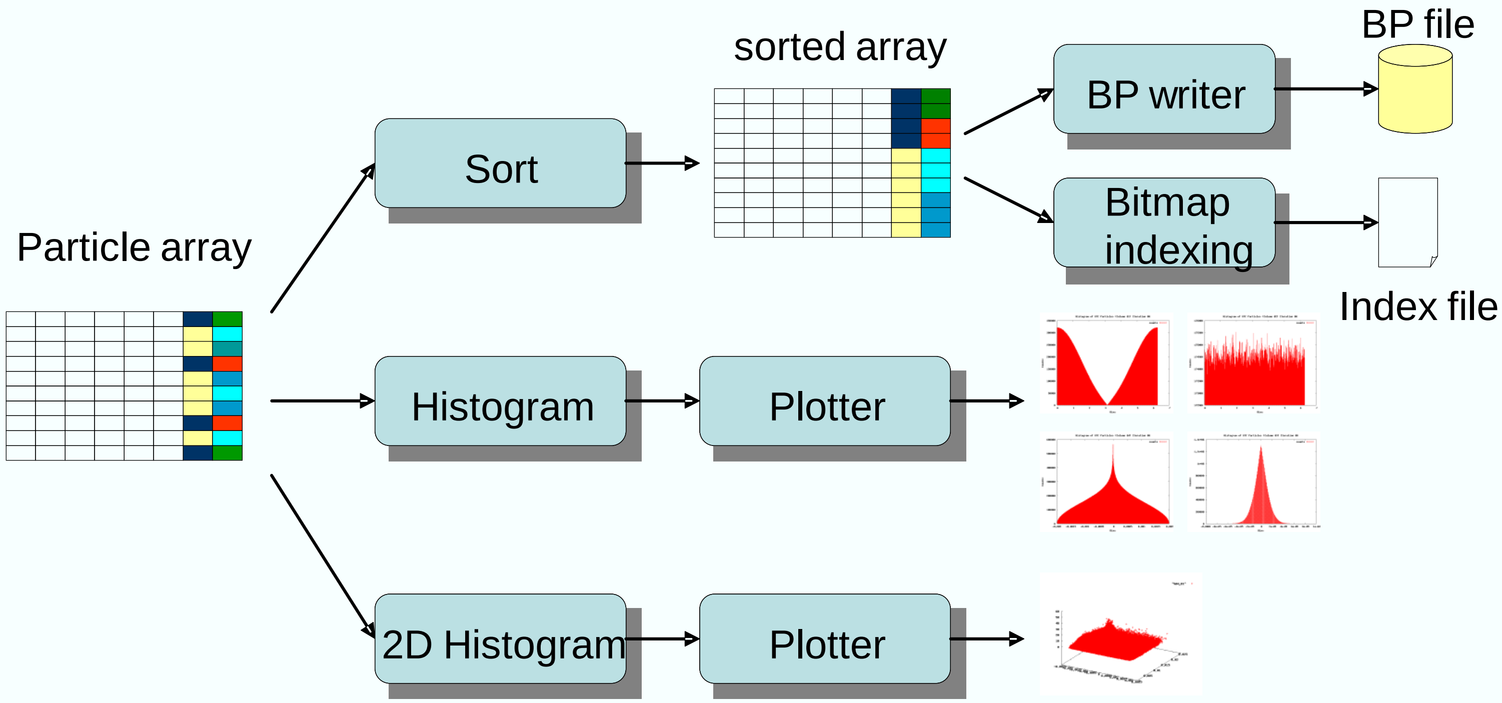
From F. Zheng, H. Abbasi, C. Docan, J. Lofstead, S. Klasky, Q. Liu, M. Parashar, N. Podhorszki, K. Schwan, M. Wolf, “PreDatA - Preparatory Data Analytics on Peta-Scale Machines”, IPDPS 2010.

Approaches to the Data Analysis Challenge

- 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-dimensional access patterns
        - General algorithms to co-locate data that is accessed together
    - Pre-compute multi-level summaries
        - e.g. monthly means in climate

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.

Approaches to the Energy Reduction Challenge (1)

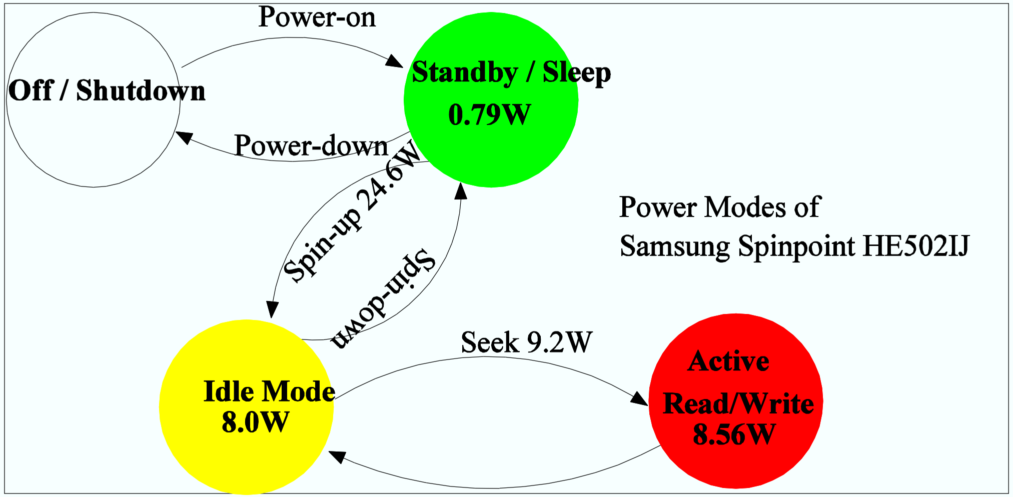
- Store intermittent data on large NVRAM
  - Checkpoint data => remove previous checkpoints data ASAP
  - Monitoring data => make NVRAM storage visible to external tools (web)
- Store longer term data on tape
  - Tape is still cost effective, and will continue to be so for awhile
  - Requires no/little 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 NVRAM to front disks to hold “hot files”

Approaches to the Energy Challenge (2)

- Minimize data movement in exascale machine
  - Take advantage of multi-core , localize data access
  - E.g. When creating an index, use extra core(s) to generate partial indexes, then combine results
  - Collect multiple “writes” and write them asynchronously
  - Perform code-coupling in memory
- Maximize sharing of data through NVRAM/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)

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



Off / Shutdown

Power-on

Standby / Sleep
0.79W

Power-down

Spin-up 24.6W

Spin-down

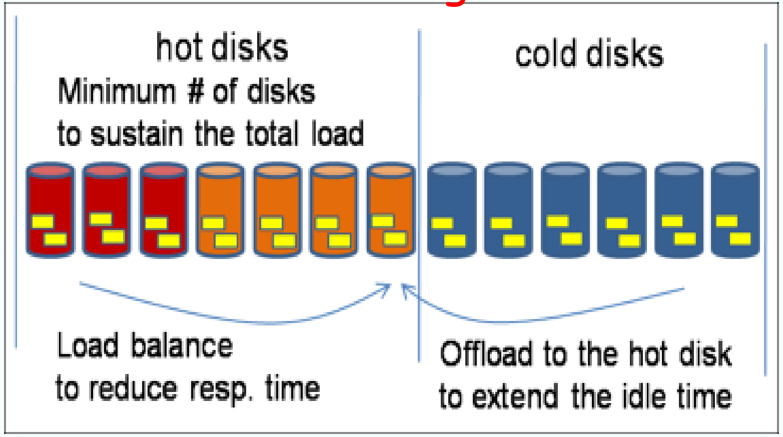
Idle Mode
8.0W

Seek 9.2W

Active
Read/Write
8.56W

Power Modes of
Samsung Spinpoint HE502IJ

Block Exchange Idea

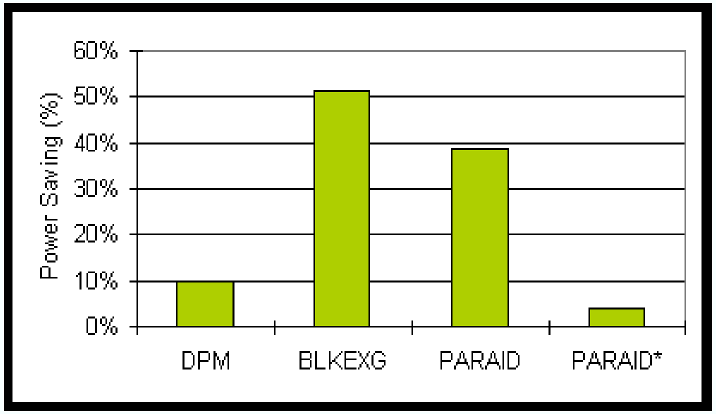


hot disks
Minimum # of disks
to sustain the total load

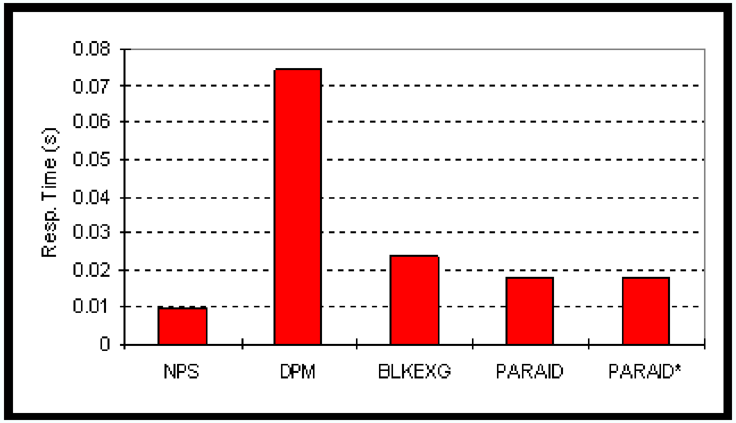
cold disks

Load balance
to reduce resp. time

Offload to the hot disk
to extend the idle time



power



Response time

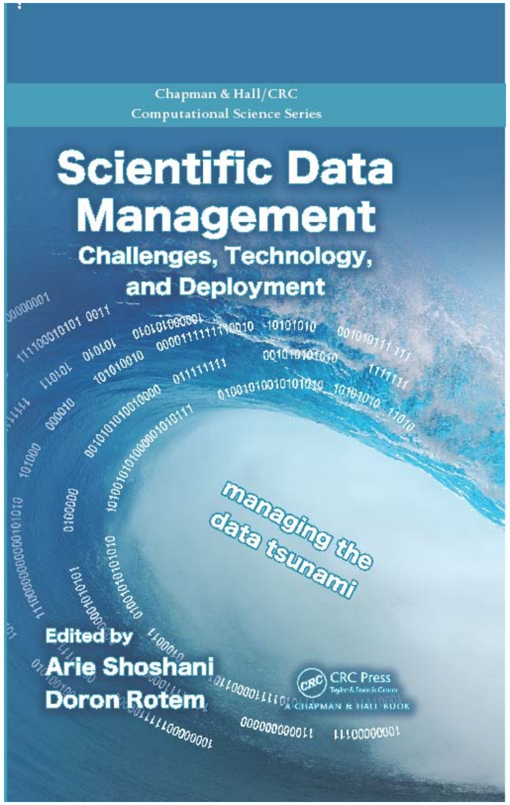
Analysis at extreme scale: Data-Side Analysis Facility

- 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

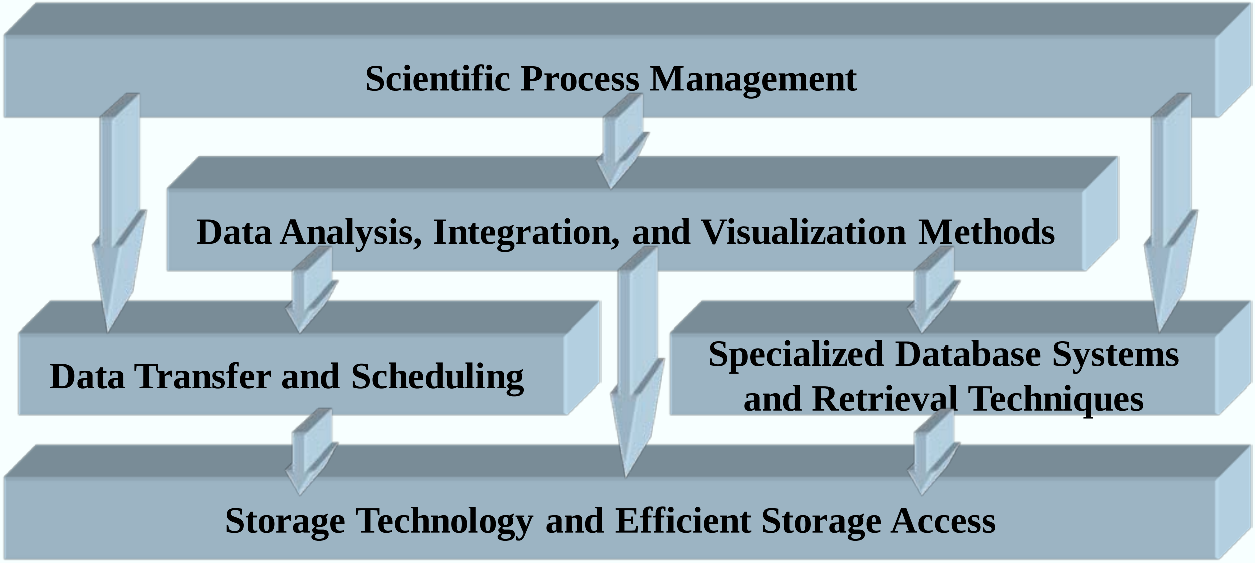
SDM Book – December 2009

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**


Chapman & Hall/CRC
Computational Science Series
Scientific Data Management
Challenges, Technology, and Deployment
managing the data tsunami
Edited by
Arie Shoshani
Doron Rotem
CRC Press
A CHAPMAN & HALL BOOK

Book Organization


Scientific Process Management
Data Analysis, Integration, and Visualization Methods
Data Transfer and Scheduling
Specialized Database Systems and Retrieval Techniques
Storage Technology and Efficient Storage Access

The END