Big Data

in Scientific Domains

Arie Shoshani

Lawrence Berkeley National Laboratory

NIST Big Data Workshop
June 2012

SDAV
Scalable Data Management, Analysis, and Visualization

Office of Science
U.S. DEPARTMENT OF ENERGY

The Scalable Data-management, Analysis, and Visualization (SDAV) Institute

2012-2017

Arie Shoshani (PI)

Co-Principal Investigators from:

Laboratories	Universities
ANL	GTech
LBNL	NCSU
LLNL	NWU
ORNL	OSU
LANL	UCD
SNL	Rutgers
Kitware (Industry)	UUtah

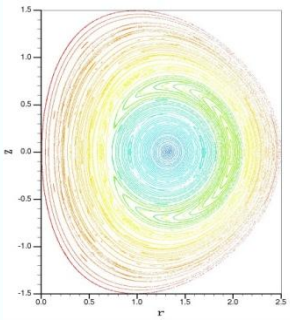
http://sdav-scidac.org/

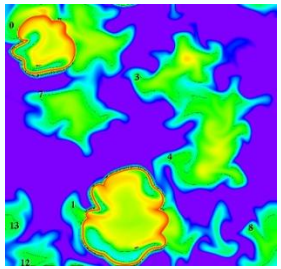
Outline

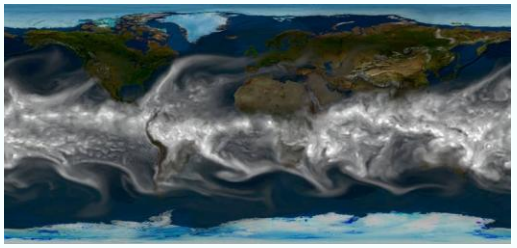
- Emerging challenges with Big Data in scientific domains (5 min)
- Examples of current approaches and solutions (15 min)
- Description of SDAV organization and technologies (5 min)

Scientific Data Management, Analysis, and Visualization

- **Applications examples**
  - Climate modeling
  - Combustion
  - Fusion
  - Astrophysics





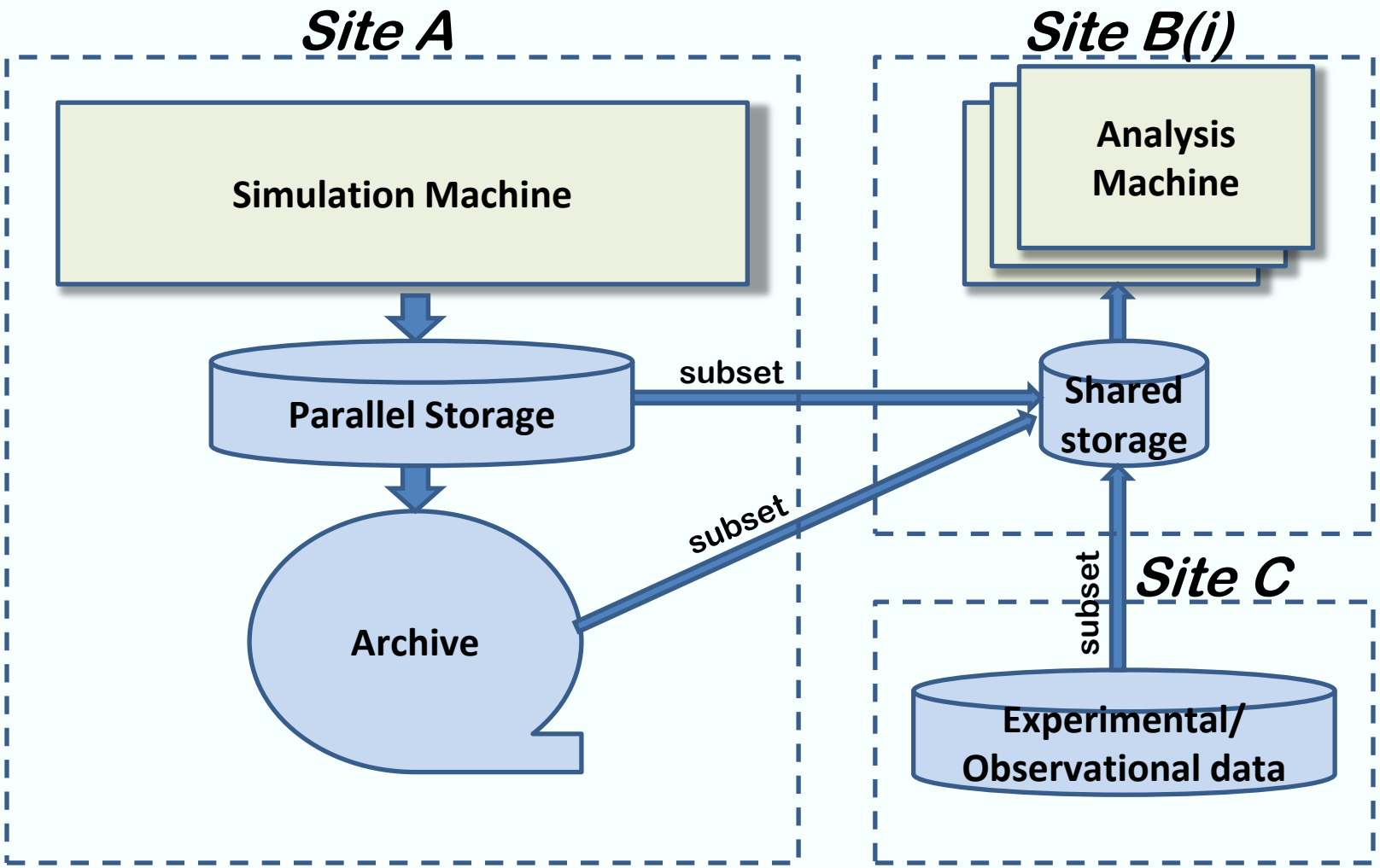


- **Algorithms, techniques, and software**
  - Representing scientific data – data models, metadata
  - Managing I/O – methods for removing I/O bottleneck
  - Accelerating efficiency of access – data structures, indexing
  - Facilitating data analysis – data manipulations for finding patterns and meaning in the data
  - Visual analytics – help understand data visually

A Typical Scientific Investigation Process

- **Current practice – data intensive tasks**
  - Runs large-scale simulations on large supercomputers
  - Dump data on parallel disk systems
  - Export some of the 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 experimental / observational data to analysis sites
  - Perform comparison of experimental/observational to validate simulations
  - Iterate

A typical Scientific Investigation lots of Data Movement (GBs – TBs)

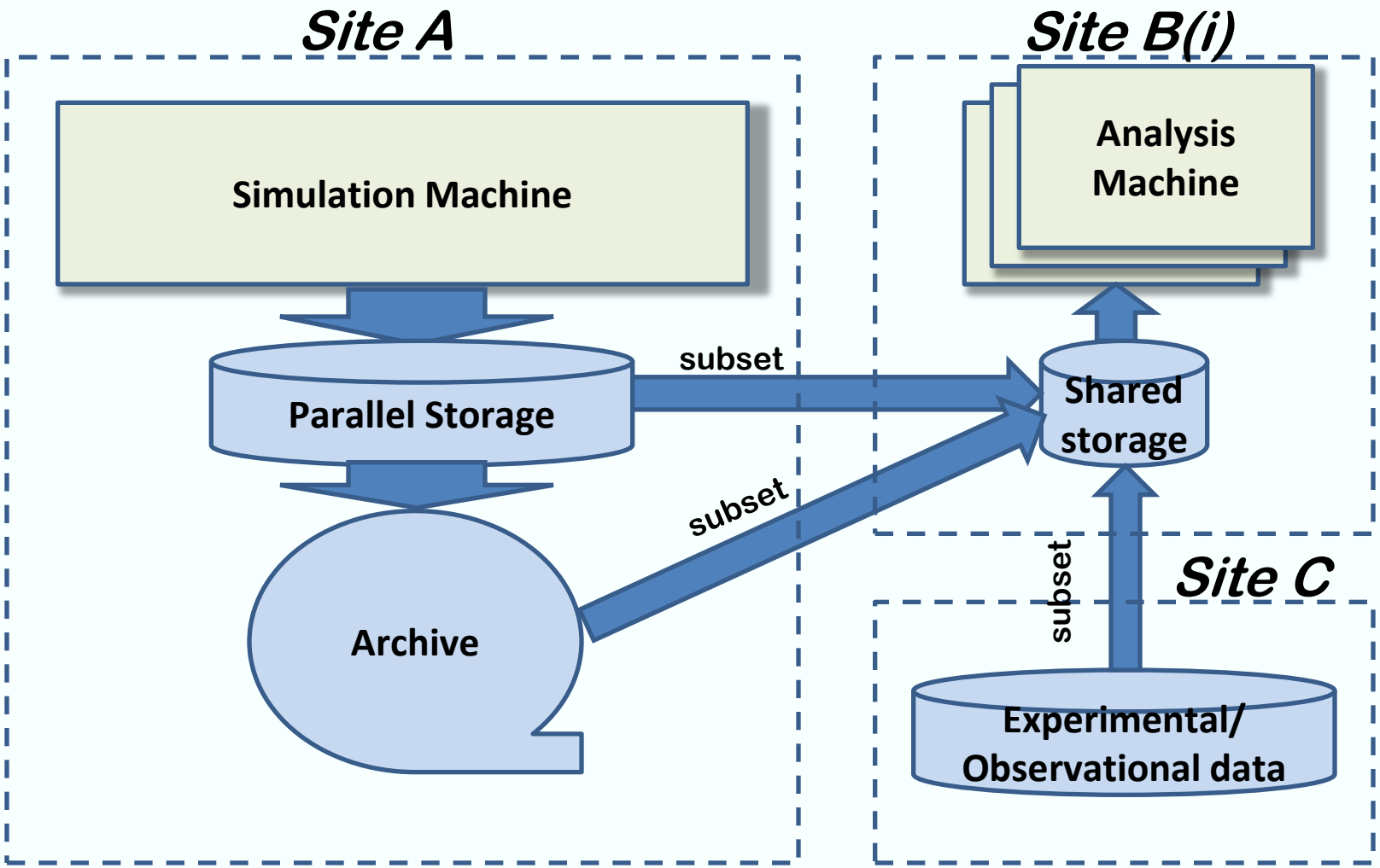


Exascale Systems: Potential Architecture

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)	

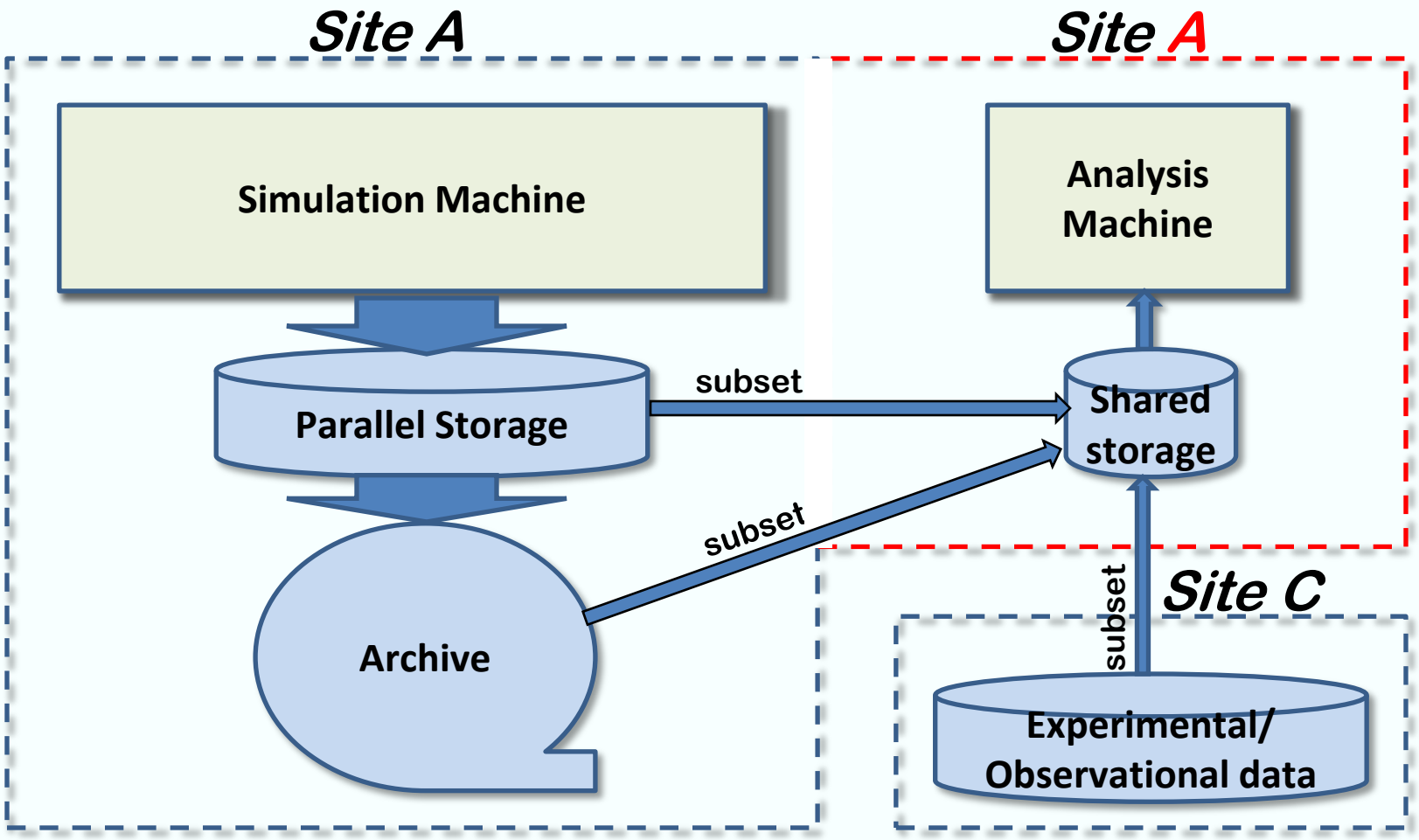
From J. Dongarra, "Impact of Architecture and Technology for Extreme Scale on Software and Algorithm Design," Cross-cutting Technologies for Computing at the Exascale, February 2-5, 2010.

At Exascale (PBs) – data volume challenge



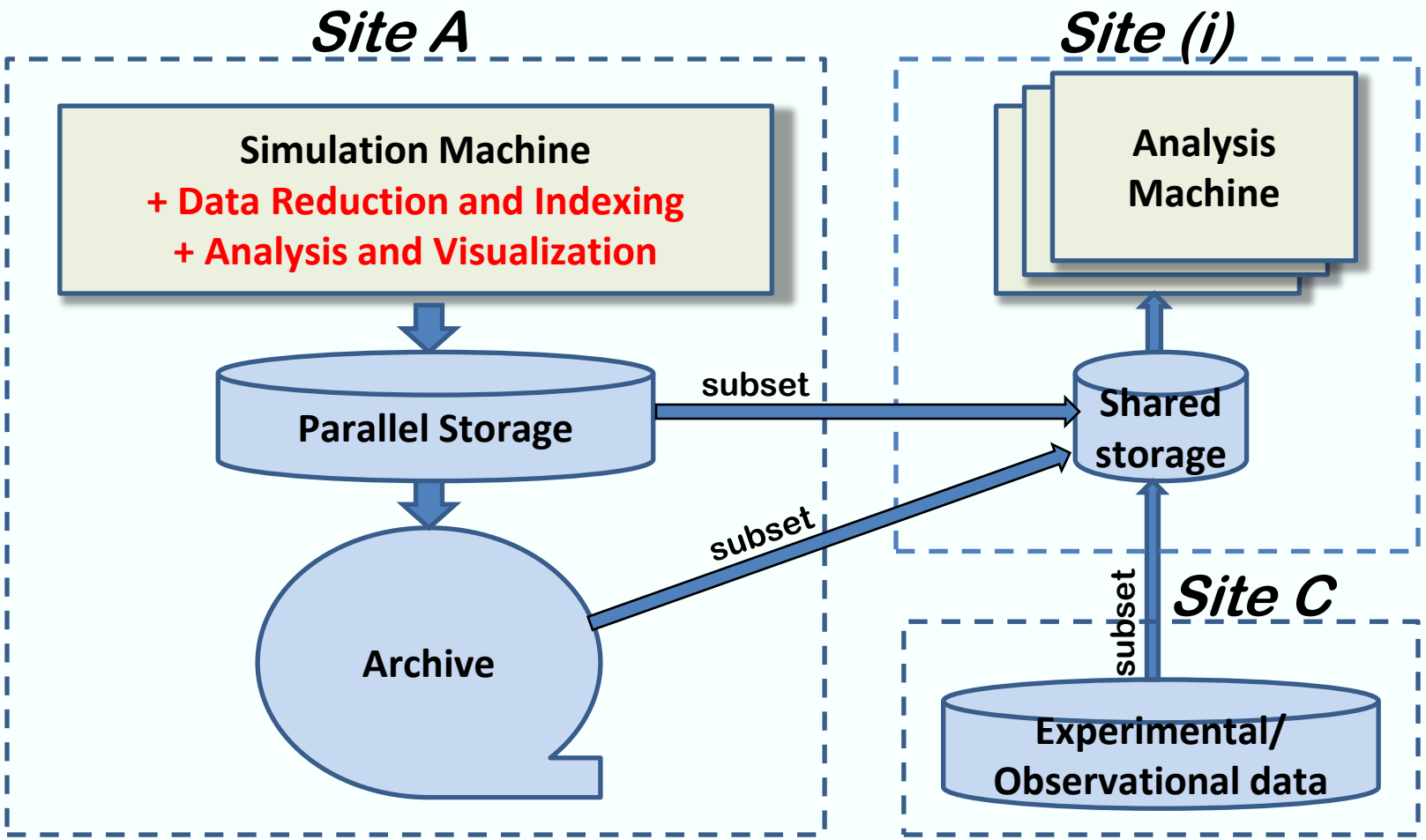
What Can be Done?

Provide analysis machines where data is generated

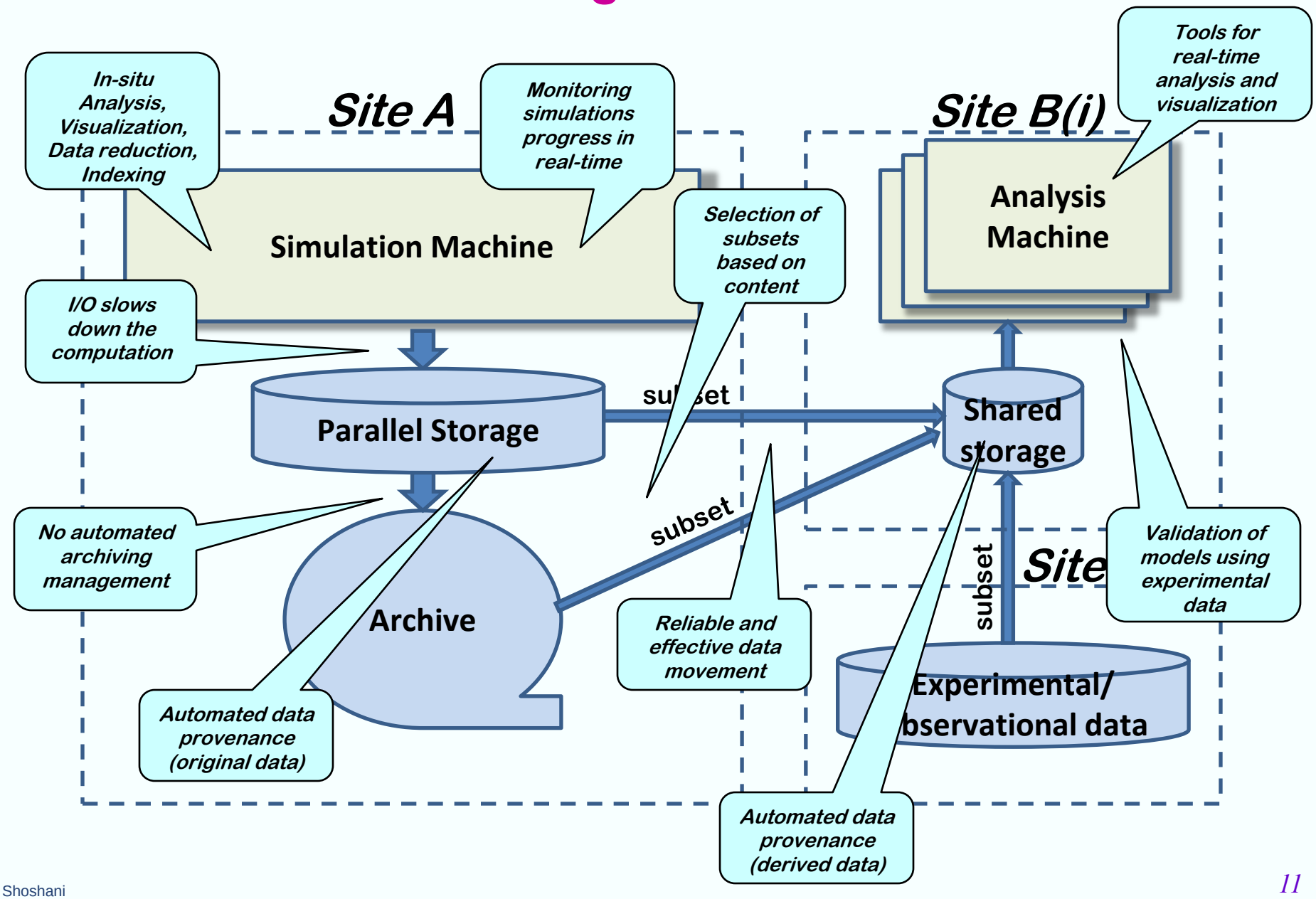


What Else Can be Done?

- Perform some data analysis and visualization on simulation machine (in-situ)
- Reduce Data and prepare data for further analysis



Bottlenecks/Problems for handling Big Data



Some Solutions are Emerging

(1) In-situ Analysis, Visualization, Data reduction, Indexing

(2) Monitoring simulations progress in real-time

(3) I/O slows down the computation

(4) No automated archival management

(5) Selection of subsets based on content

(6) Reliable and effective data movement

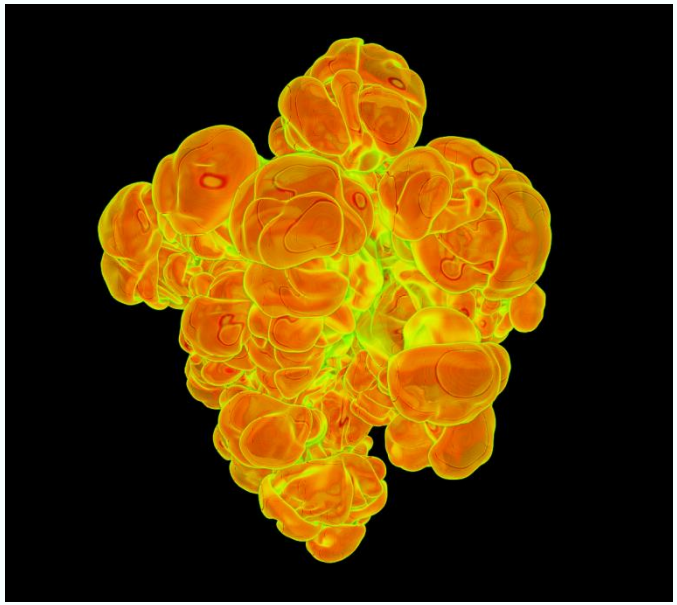
(7) Tools for real-time analysis and visualization

(8) Validation of models using experimental data (some)

(9) Automated data provenance (some)

Example:
Framework for Improving I/O and in situ Visualization

I/O bottlenecks and analysis challenges faced by applications running on leadership systems



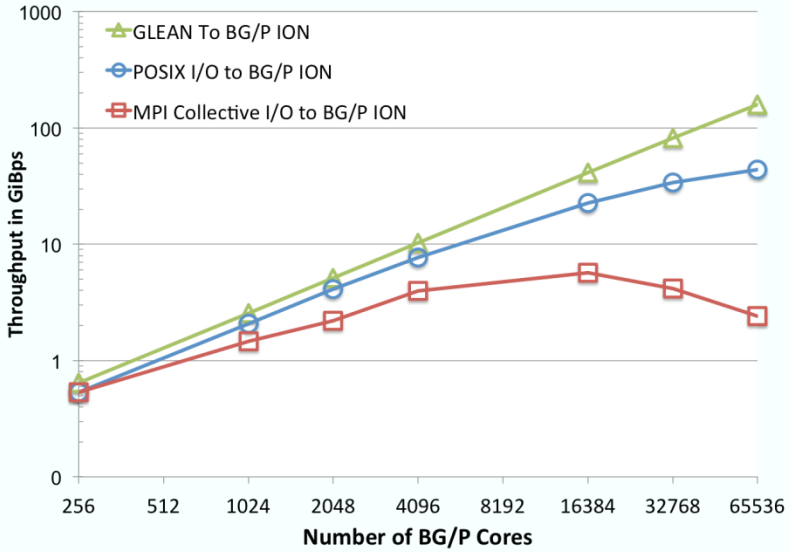
Visualization of Type 1A supernova detonation

- FLASH is multi-scale, multi-physics code used in domains including astrophysics, cosmology and high-energy density physics.
- It uses a block-structured AMR, and at 32K cores, I/O time is about 30% of the entire run achieving a max of 1GB/s out of 35 GB/s on the ALCF Intrepid BG/P system.
- Storage is sometimes referred to as the "black-hole" by FLASH scientists as it significantly impedes the time to glean insights from the simulation.

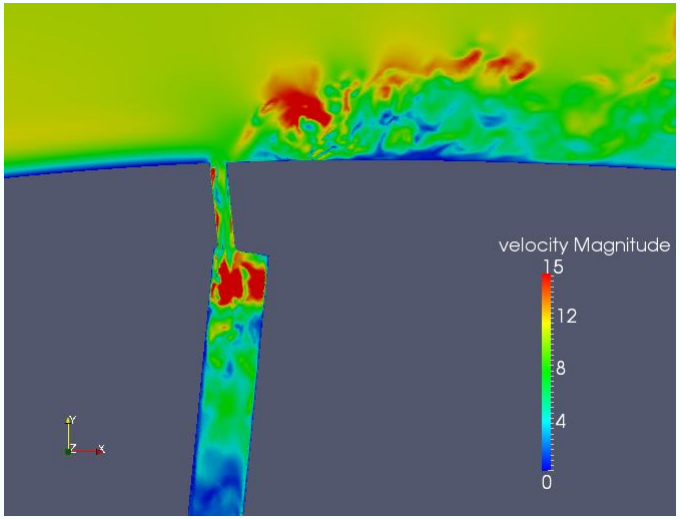
Improving I/O performance and reducing the time-to-discovery is critical to FLASH. Similar challenges faced by several applications running on DOE leadership systems such as PHASTA CDF code.

Highlight: GLEAN

- GLEAN is a flexible and extensible framework to facilitate simulation-time data analysis and I/O acceleration.
- Features include: topology-aware data movement, asynchronous data staging and burst buffering, leverages application data models, scalable analysis algorithms and infrastructure (in situ, co processing, in flight).
- Scaled to entire ALCF infrastructure (160K BG/P Intrepid cores), achieved multi-fold I/O improvement for FLASH, and demonstrated in situ analysis.



Strong scaling performance for 1GB data movement from ALCF Intrepid Blue Gene/P. Strong scaling is critical as we move towards systems with increased core counts.



Co-visualization of a 3.3 billion element PHASTA simulation of an aircraft wing running on 160K cores of ALCF Intrepid Blue Gene/P using ParaView on 100 nodes of ALCF Eureka analysis cluster enabled by GLEAN.

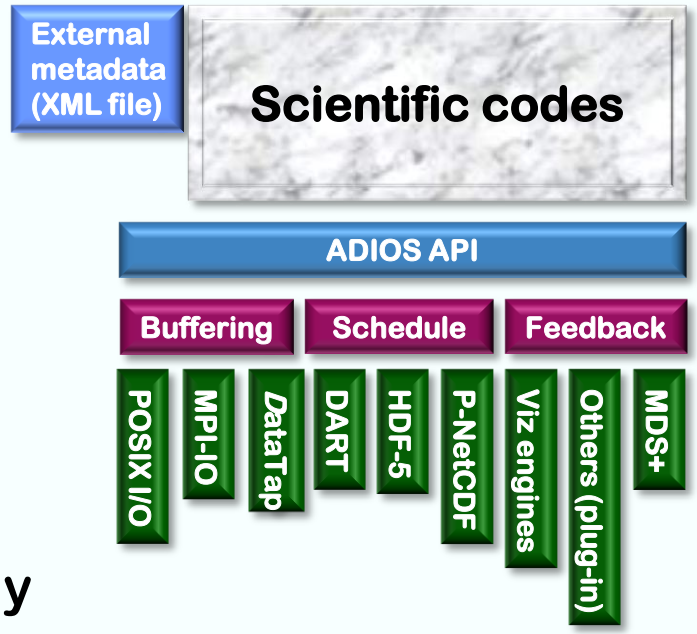
Example:

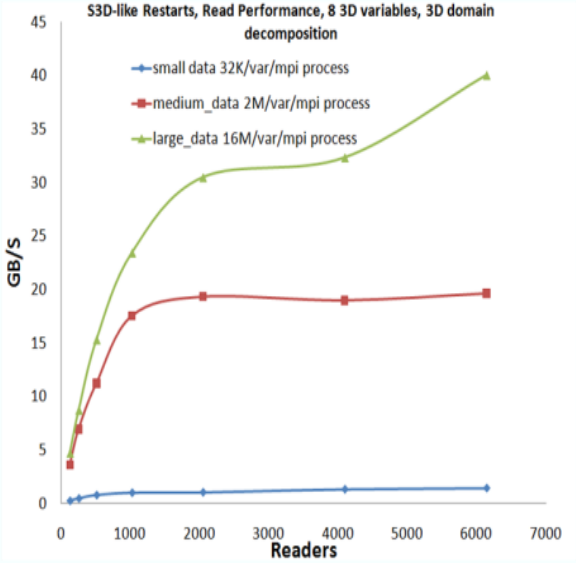
Capturing I/O and

Monitoring simulations progress in real-time

ADIOS: Adaptable I/O System

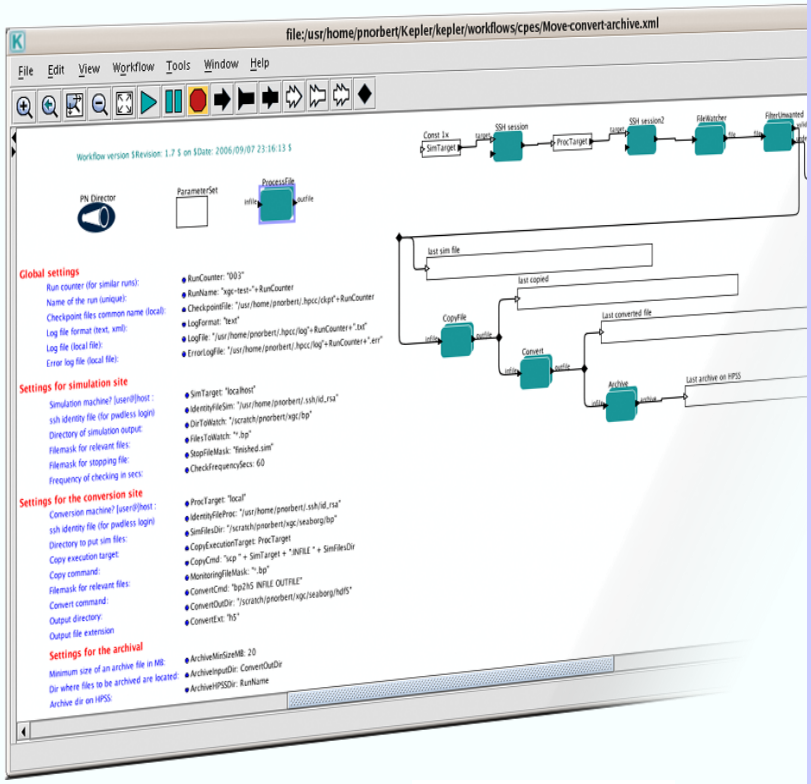
- Overview: service-oriented architecture:
  - Allows plug-ins for different I/O implementations
  - Abstracts the API from the method used for I/O
- Simple API, almost as easy as F90 write statement
- Synchronous and asynchronous transports supported with no code changes
- Change I/O method by changing XML file only
- ADIOS buffers data
- ADIOS allows multiple transport methods per group





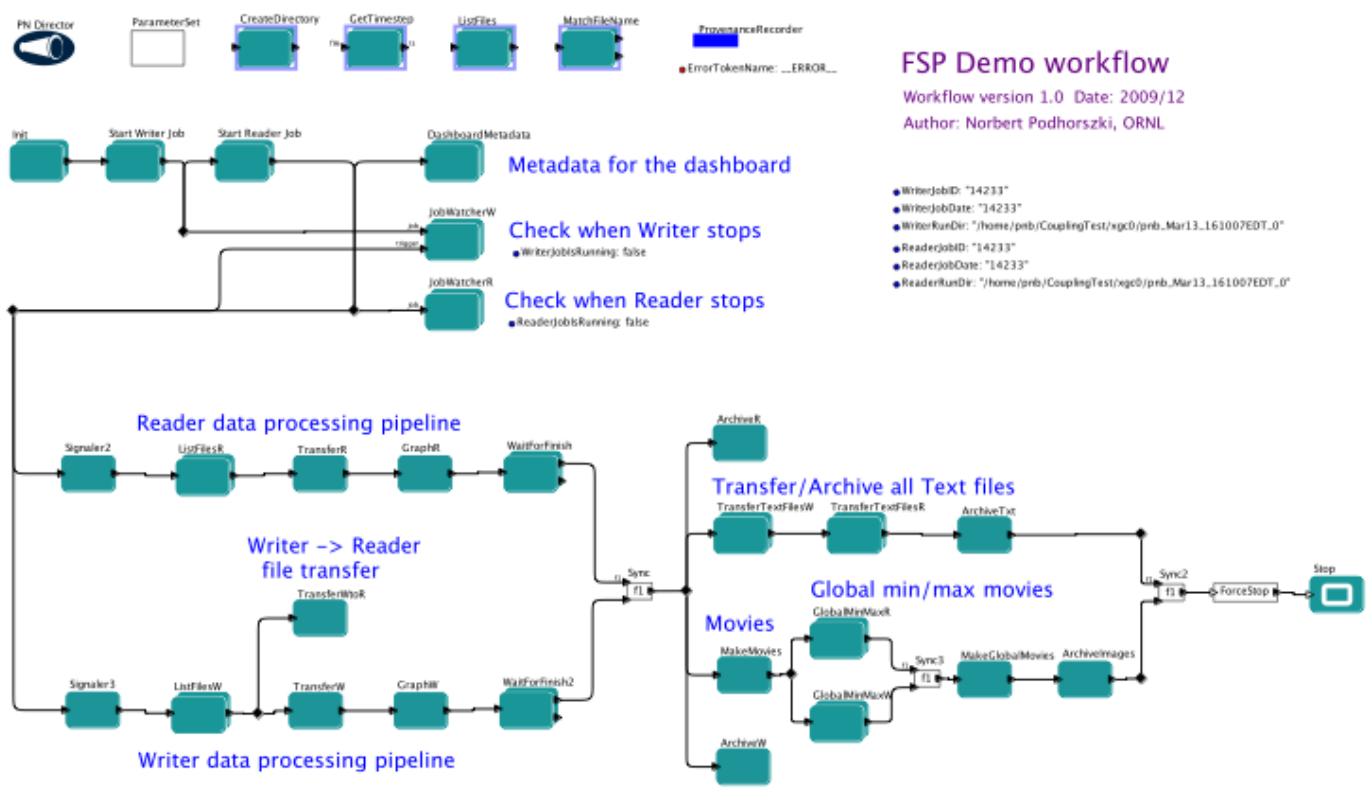
Monitoring simulations progress in real-time (Initially inspired a Fusion project)

Monitoring saves cycles



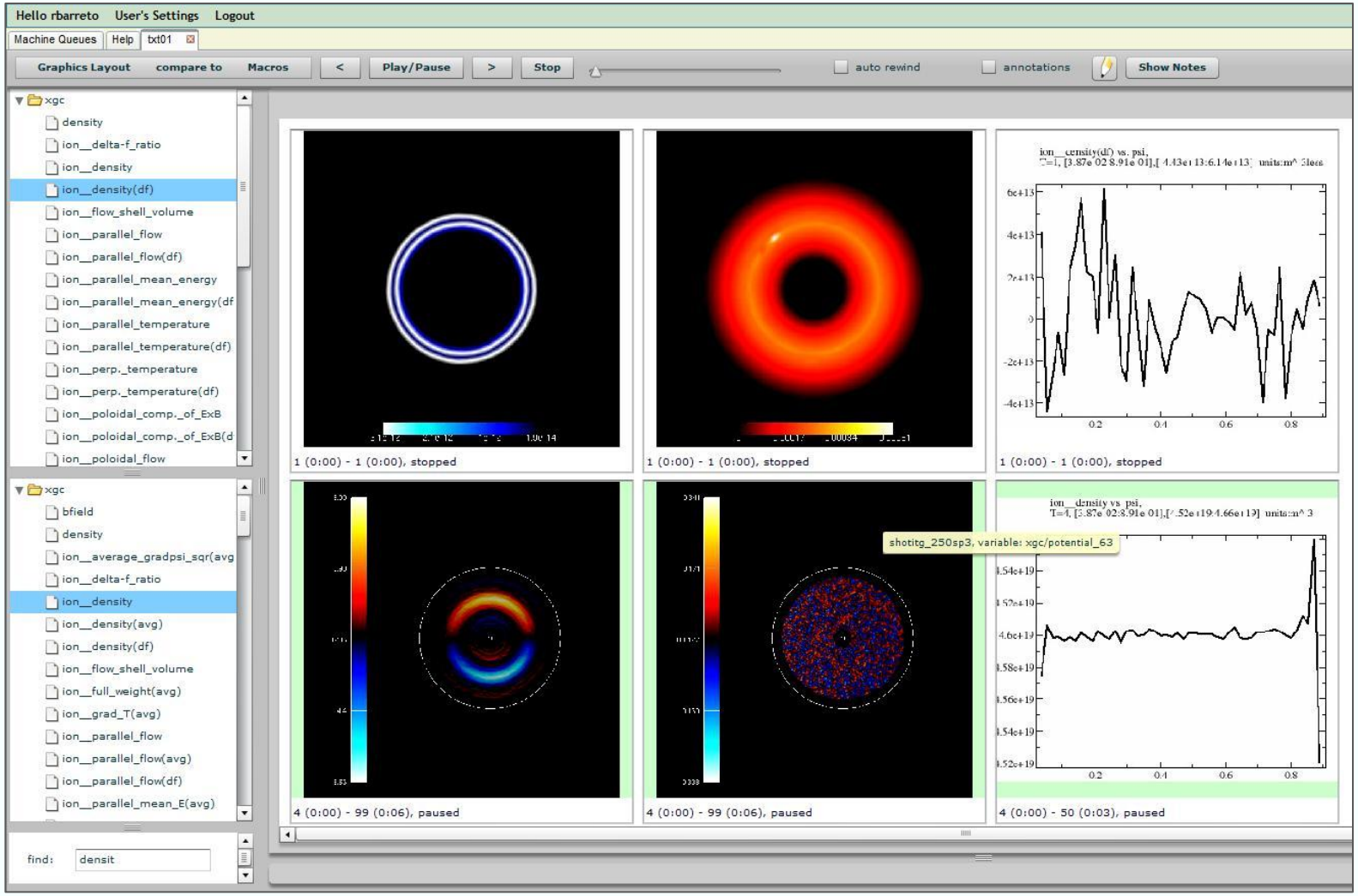
- **Automate the monitoring pipeline**
  Use workflow
  - Monitoring of large-scale simulations must be dynamic
  - transfer of simulation output to remote machines
  - execution of conversion routines
  - image creation, data archiving
- **Requirements for ease-of-use**
  Use dashboard
  - A way to view monitoring information from anywhere
  - A way to see graphs, images, and movies
  - A way to see provenance and move data to user's site
  - A way to perform statistical analysis, compare graphs, etc.

The Kepler Workflow Engine



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

Real-time visualization and analysis capabilities on dashboard



visualize and compare shots

Example:
In Situ Code Coupling
and in Situ Visualization

In-situ Execution of Coupled Scientific Workflow

Driving Application Scenarios:

- Online and In-situ data analytics
  - E.g. visualization, feature tracking

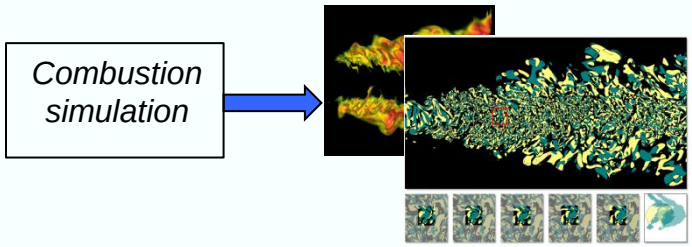


Fig. Tracking vortical structure in turbulent combustion flow field (Figure credit: Hongfeng Yu, Sandia Lab)

- Integrated and coupled multi-physics simulation
  - E.g. integrated climate modeling, fusion simulation, subsurface modeling, material science workflows

Motivations:

- *Emerging scientific workflows are composed of heterogeneous coupled component applications that interact and exchange significant volumes of data at runtime*
- *On-chip data sharing is much cheaper than off-chip data transfers*
  - *Large volumes of data movement over communication fabrics → contention, latency and energy consumption*
- *High-end systems have increased cores count (per compute node)*
  - *E.g., Cray XK6 (Titan) -- AMD 16-core per processors; IBM BG/Q (Mira and Sequoia) --17-core per processor*

Problems with Traditional Approach for Data Sharing:

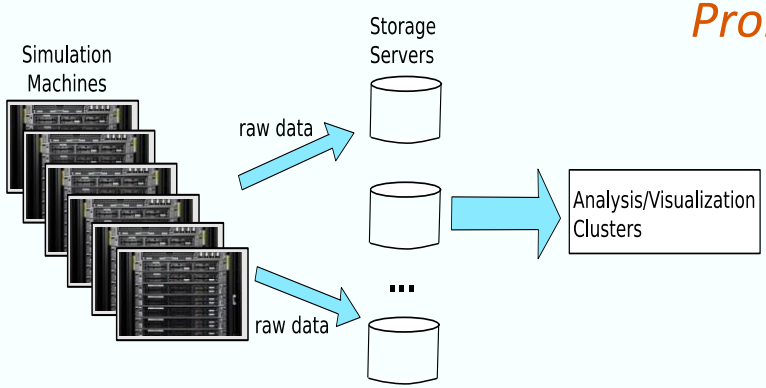


Fig. Illustration of traditional post-processing data processing/analysis pipeline (based on disk IO)

Disk based

- Increasing performance gap between computation and disk IO speeds, IO becomes the bottleneck

Coupler approach (hub-and-spoke)

- Couplers can become the bottleneck limiting scalability
- Larger data sharing latency, data is moved twice

Large volumes of network data movement →
Increasing costs in terms of time and energy!

In-situ Execution of Coupled Scientific Workflow

Technology:

Enable in-situ execution of coupled scientific workflow

- Coupled simulation / data analytics /data processing workflow composed as a DAG of tasks
- **DataSpaces** based shared space programming abstraction to express coordination and data sharing
- Elastic DHT spans in-situ staging cores to keep track of data locations and provide metadata lookup service
- Data-centric task mapping to increase intra-node data sharing and reuse
- HybridDART – dynamically select between RDMA-enabled network or shared memory for data transport

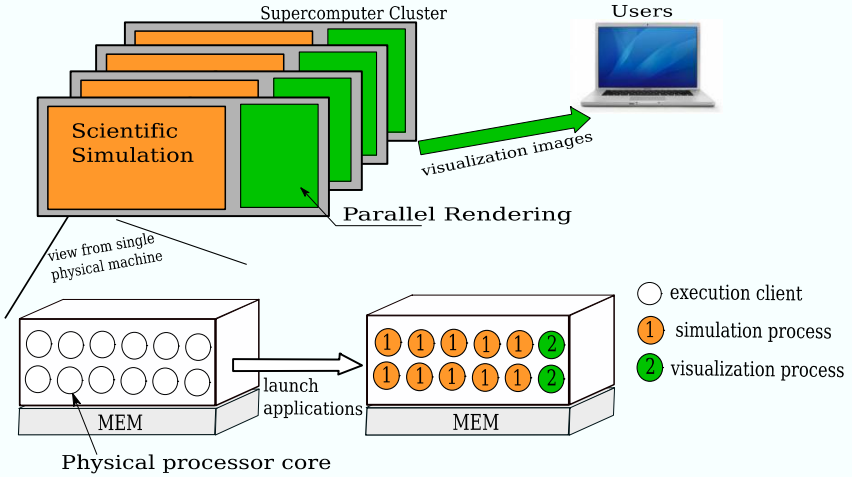


Fig. In-situ execution of simulation and visualization processes on a multi-core platform

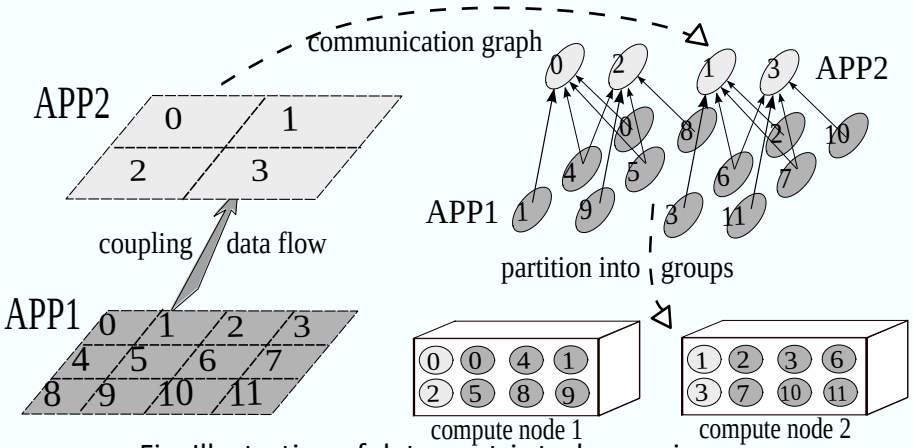


Fig. Illustration of data-centric task mapping

Results:

Two workflow scenarios evaluated on Cray XT5

- Significant saving in the amount of data transferred over the network by co-locating data producers and consumers on the same processor
- Data transfer time (and energy) decreased as much of the coupled data is retrieved directly from on-processor shared memory

References

1. F. Zhang, C. Docan, M. Parashar, S. Klasky, N. Podhorszki, H. Abbasi: Enabling In-situ Execution of Coupled Scientific Workflow on Multi-core Platform. IPDPS'2012

Example:

Client –Server Remote Visualization

Visualization of Laser Back Scatter

Application:

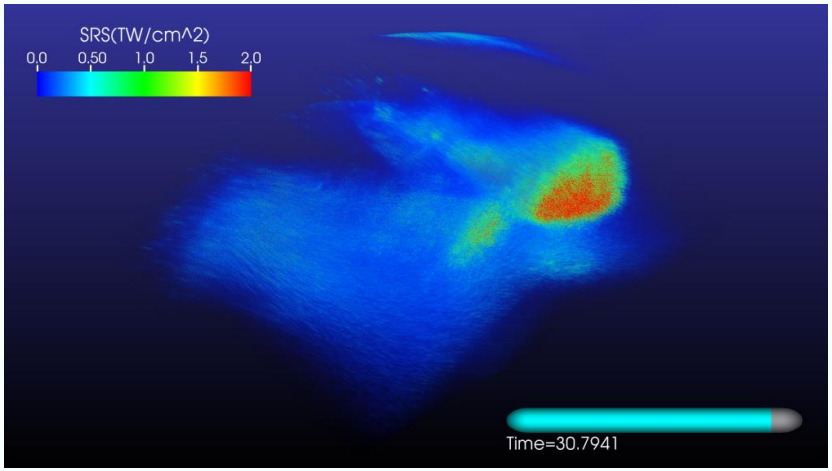
- laser-induced fusion at LLNL
- Laser back scatter modeling by Dr. Steve Langer

Simulation Goal: Understand Laser Back Scatter

- Two laser beams impacting a deuterium and tritium target
- Determine amount and direction of back scattered energy

Requirement: Visualizing the time dependent behavior

- Visualizing input energy and back scatter energy over time
  - Understand how the back scatter is formed
  - Understand the orientation and intensity of the back scattered energy



Challenges

- *Extreme data size generated by high resolution simulation (220 billion cells)*
- *Correlating multiple, complex, 3d phenomenon over time*

A burst of back scatter energy traveling from left to right

Side-by-Side Volume Visualizations of Time Dependent Behavior

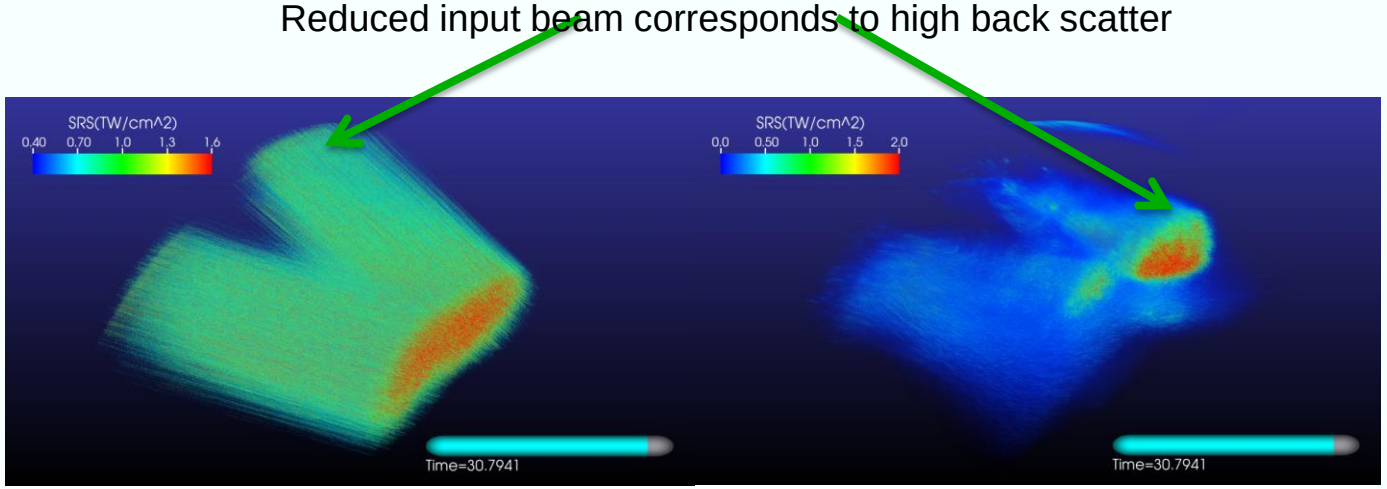
Technology

- Ability to view data using a client/server architecture using **VisIT**
- Ability to quickly generate side-by-side animations of key physics quantities to understand time dependent behavior
- Automated movie generation makes it easy for the user to see time evolution of data with different views and transfer functions.

Result/Impact

- Client/server architecture allows the user to view his data on his desktop without moving the data
- Ability to interactively set transfer function to bring out features of interest at key points in the simulation
- This is the first time scientists could see how the backscatter was forming. Orthogonal slices was inadequate, since this was a 3d phenomenon. This helps them design the target and laser pulse timing.

Reduced input beam corresponds to high back scatter



Input beam

Back scatter

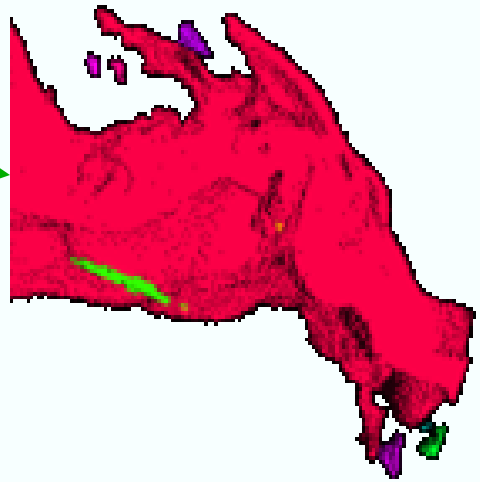
Example:

Indexing to Select Subsets Based on Content

facilitates interactive visualization

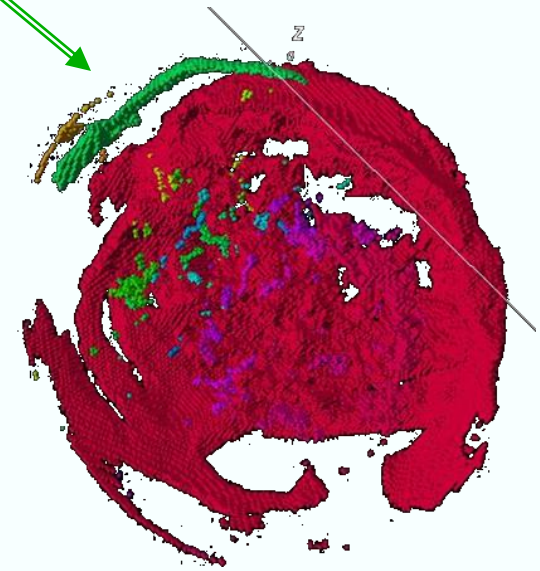
Selection of subsets based on content

- 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) – each time step can have 100 billion data values
- 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.



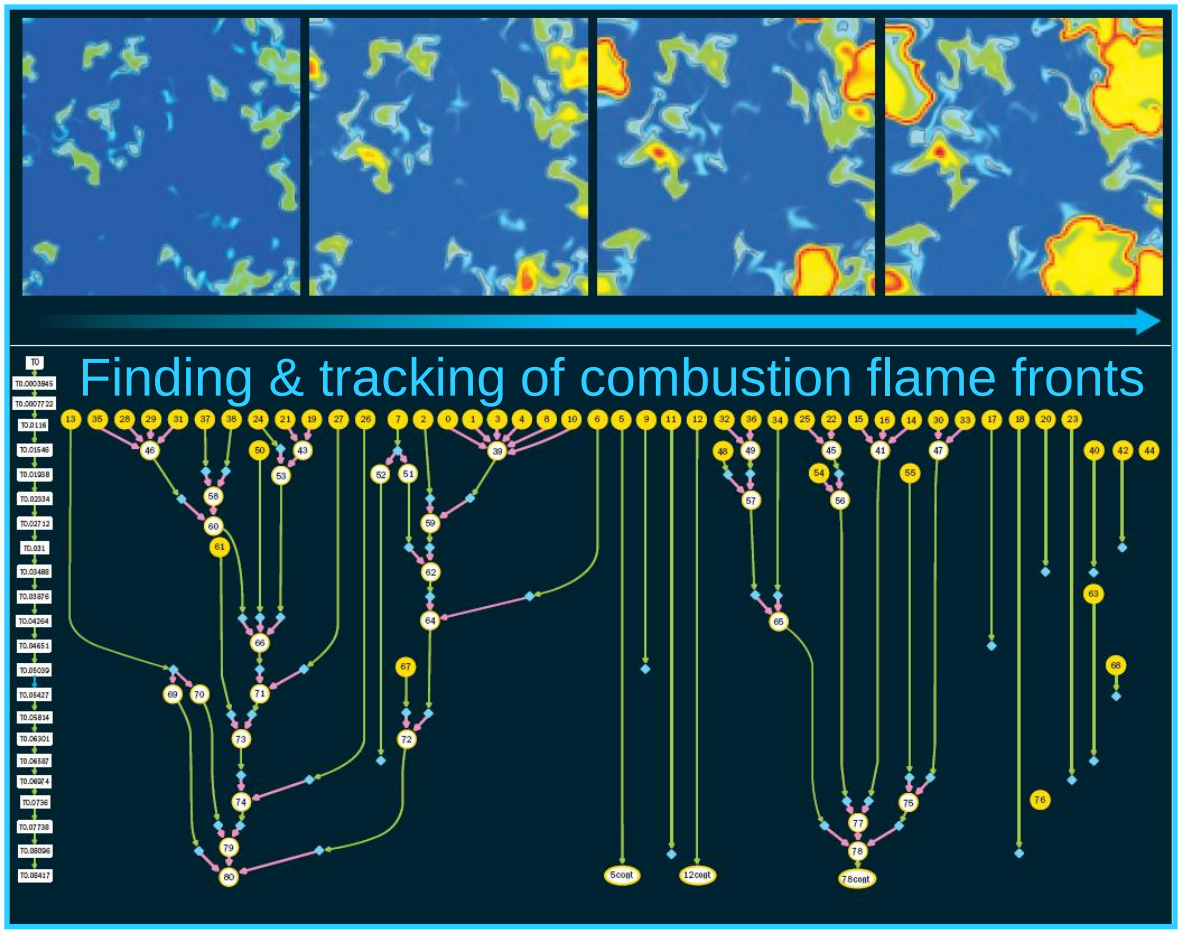
- Other application domains:
  - real-time analysis of network intrusion attacks
  - fast tracking of combustion flame fronts over time
  - accelerating molecular docking in biology applications
  - query-driven visualization

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}"

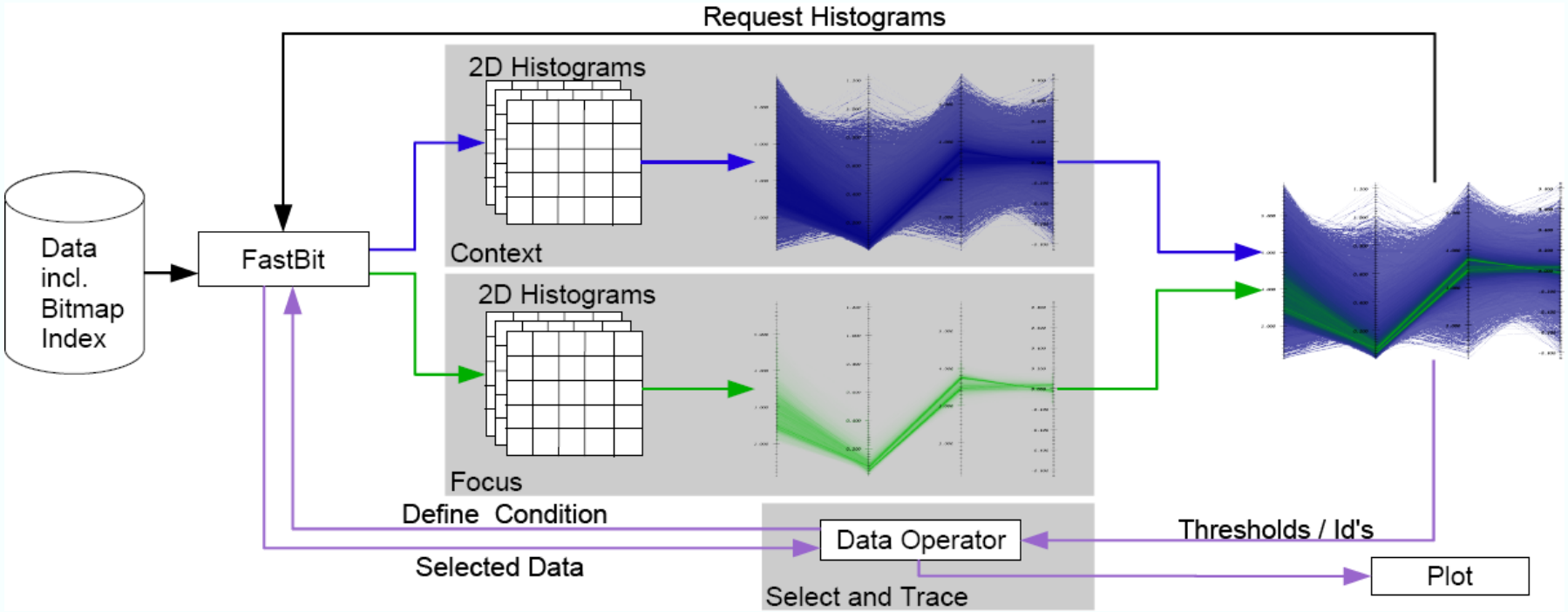
Region growing

Connect neighboring cells into regions

Region tracking

Track the evolution of the features through time

Query-Driven Visualization



- Collaboration between data management and visualization technologies
  - 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)

Example:

In Situ Data Reduction

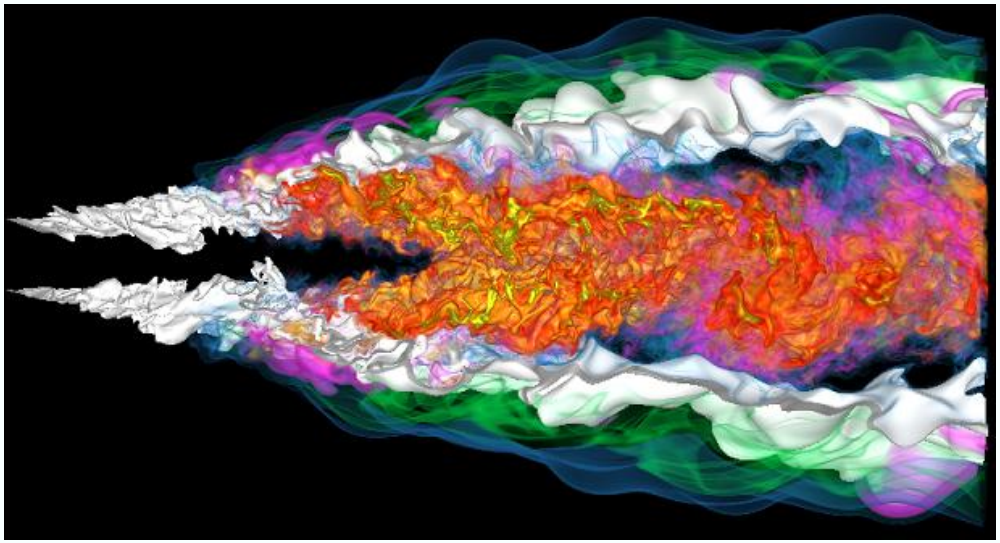
Promising Ideas for larger scale data: and User-Assisted Data Reduction

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.



Contact: Kwan-Liu Ma

UCDAVIS UNIVERSITY OF CALIFORNIA

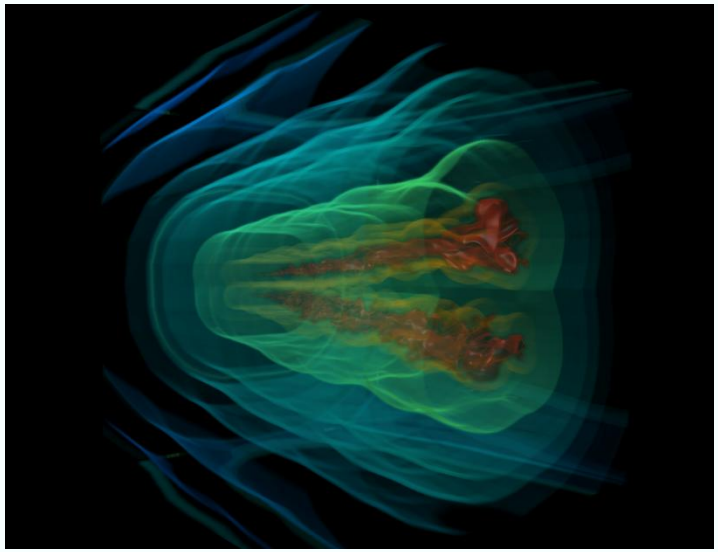
Parallel Distance Field Computing

Technology

- Distance field computing is fundamental to many data analysis and visualization applications.
- This project realizes a highly scalable parallel implementation to support in situ processing and data reduction.
- The design is general to handle a variety of data
- The product is a standalone library to be easily adopted for different settings.

Applications & Impact

- Distance fields can be used as *importance* fields to guide rendering, data compression, sampling, and feature-based optimizations.
- The resulting technology will benefit many SciDAC applications from combustion, fusion, to climate and astrophysics simulations.
- This work will motivate others to develop novel visualization and data reduction methods using distance fields.



Depiction of a distance field computed from a feature surface in data generated by a combustion simulation.

Results

- A use case on a turbulent combustion simulation shows over 80% data reduction while revealing previously hidden flow features.
- A parallel spatial data structure has been designed to accelerate the computation.
- **Tests on the parallel implementation show the data that must be exchanged is under 0.01% of the total data, and the cost to exchange the data is under 0.2% of the overall time.**

Contact: Kwan-Liu Ma

UC DAVIS UNIVERSITY OF CALIFORNIA

Example:

Time-varying Feature Tracking

Analysis and Visualization of Madden-Julian Oscillation (MJO)

Application:

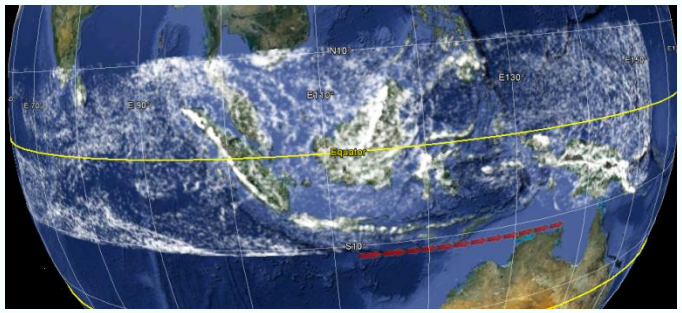
- Climate modeling by Dr. Ruby Leung and Dr. Samson Hagos at PNNL

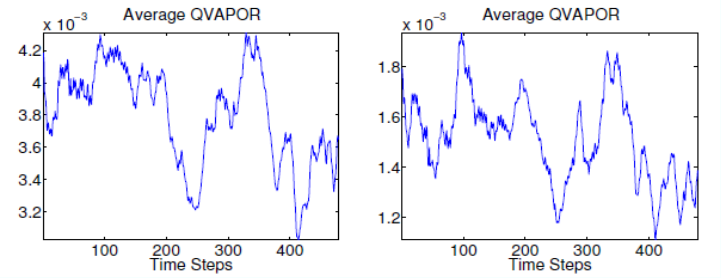
Simulation Goal: Understand MJO

- A complex cloud system in multiple scales over the Indian and Pacific oceans
- An important weather phenomenon related to the tropical intra-seasonal change

Requirement: Time-varying Feature Tracking

- Visualizing the movement of MJO-related quantities (eg. cloud population and precipitation)
  - Understand how MJO is formed
  - Understand how MJO is related to different physical quantities
- Feature identification to filter the major cloud movement of MJO





Challenges

- Large data size generated by high resolution simulations
- Various temporal scales of MJO
  - make traditional techniques that perform per time step analysis and visualization ineffective
  - require global analysis over time other than local feature tracking

(Upper left): Cloud rendering on a virtual globe
(Lower left): The average water vapor mixing rate, indicating the existence of two MJO cycles

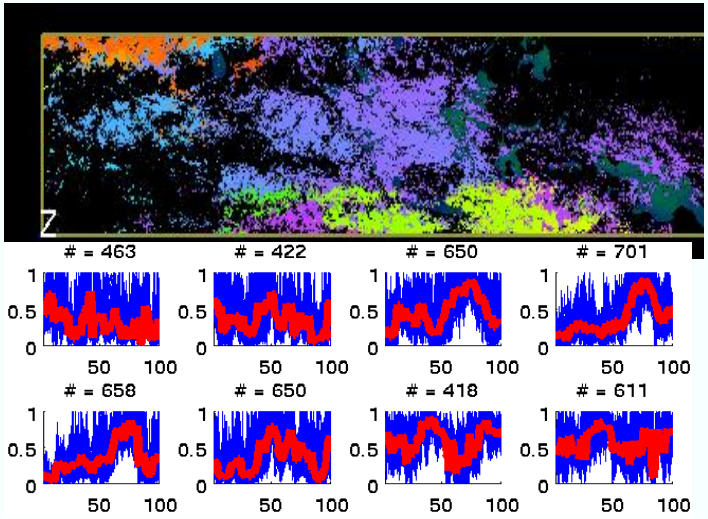
Time-Varying Data Analysis with Time Activity Curves

Technology

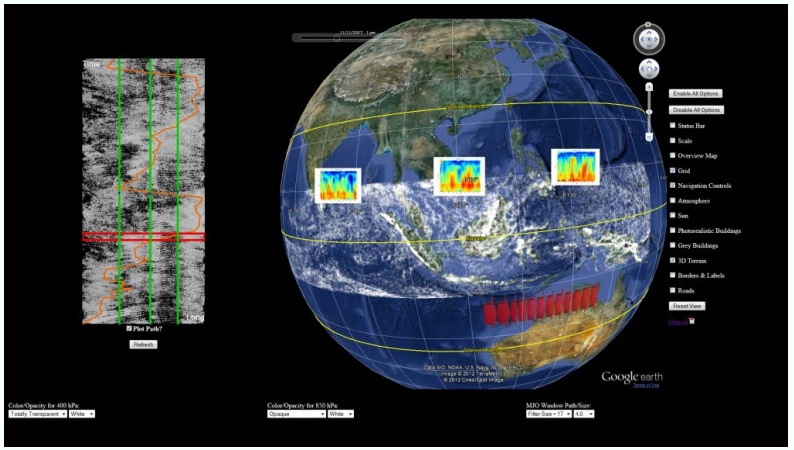
- Robust global time-varying feature descriptors
  - Model the time-varying feature based on the change of value over time at each location
- Visualize complex time-varying phenomena with detailed multivariate temporal trend analysis

Result/Impact

- Flexible interface to allow scientists to interact with data at their desktop
- Replace tedious viewing of animations with more efficient visualization of spatio-temporal features in still images
- Assist in summarizing and verifying major temporal trends such as MJO in extreme scale data sets



(Top): TAC-based (Time Activity Curve) region clustering
(Bottom): Clustering TACs from all 2D location.



Linked views between TAC-based overview and virtual globe.
(Top Left): TACs per longitude as Hovmoller diagram.
(Top Right) Cloud animation with Feature Tracking

Example:

parallelization of large-scale 3D movies

Analysis and Visualization of Magnetic Reconnection

Application:

- Simulation of magnetic reconnection by Bill Daughton (LANL) and Homa Karimabadi (UCSD)

Simulation Goal:

- To understand the 3D evolution of tearing modes – a type of plasma instability that spontaneously produces magnetic reconnection while giving rise to topological changes in the magnetic field.

Requirement: Remote Interactive Analysis

- Interactive 3D visualization of simulation data
  - Particle data
  - Mesh data
- Comparison with theoretical expectations
- Rapid exploration due to limited availability of supercomputer to run large simulations

Challenges

- Large data size generated by high resolution simulation
  - Simulation on 98304 cores
  - 6.4 billion cells
  - 1.5 trillion particles
  - 57 TB data
- Only remote access to the supercomputer
- Lack of dedicated visualization resources

Los Alamos NATIONAL LABORATORY EST. 1943

Kitware

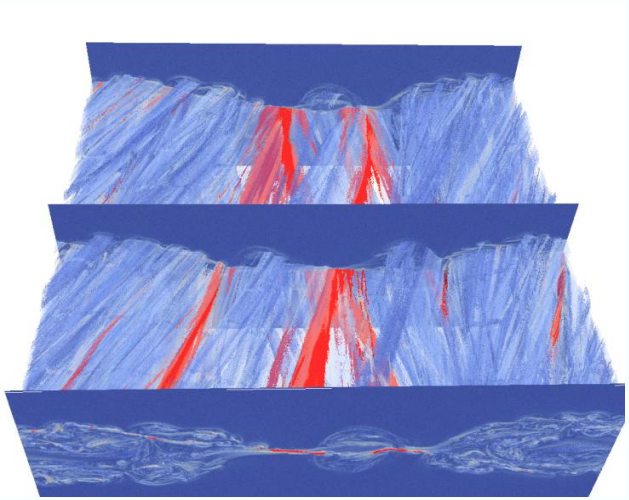
Remote Visualization with ParaView

Technology

- **ParaView**: general purpose data analysis and visualization tool focused on large data
- Client/server architecture for remote visualization on supercomputers
- Designed to be extensible: I/O routines and domain-specific algorithms developed by science teams

Result/Impact

- Allowed scientists to remotely analyze and visualize their data when it is not possible to copy locally
- Allowed scientists "to rapidly explore the grid data to understand the 3D evolution of magnetic reconnection"
- As expected, a spectrum of tearing instabilities develops which interact, forming new current sheets and triggering secondary tearing instabilities



Isosurface of particle density colored by current density

Los Alamos NATIONAL LABORATORY EST. 1943

Kitware


SDAV
Scalable Data Management, Analysis, and Visualization Institute
DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA
• In-situ operation • Visualization • Flexible Analysis • Advanced Architectures • Scientific Softwar
Lead institution: Lawrence Berkeley National Laboratory Berkeley CA, 94720

Goal

- **The goal of SDAV is twofold:**
  - **to actively work with application teams to assist them in achieving breakthrough science;**
  - **to provide technical solutions in the data management, analysis, and visualization regimes that are broadly used by the computational science community.**

Accomplishing Our Goal

- **Community Engagement.**
  - We will actively engage application teams running on leading DOE computing systems, our sibling Institutes, and DOE computing facility personnel over the lifetime of the Institute.
- **Technology Deployment.**
  - We will work with application scientists so that they can use state of the art tools and techniques to support their needs in data management, analysis, and visualization tasks.
- **Research Integration.**
  - We will incorporate ASCR basic research results into our portfolio and develop new technologies as needed to meet the needs of application scientists over the next five years.
- **Software Support.**
  - Just as ongoing communication ensures the relevance of our work, quality software deployment, maintenance, and support ensure the success of our tools.

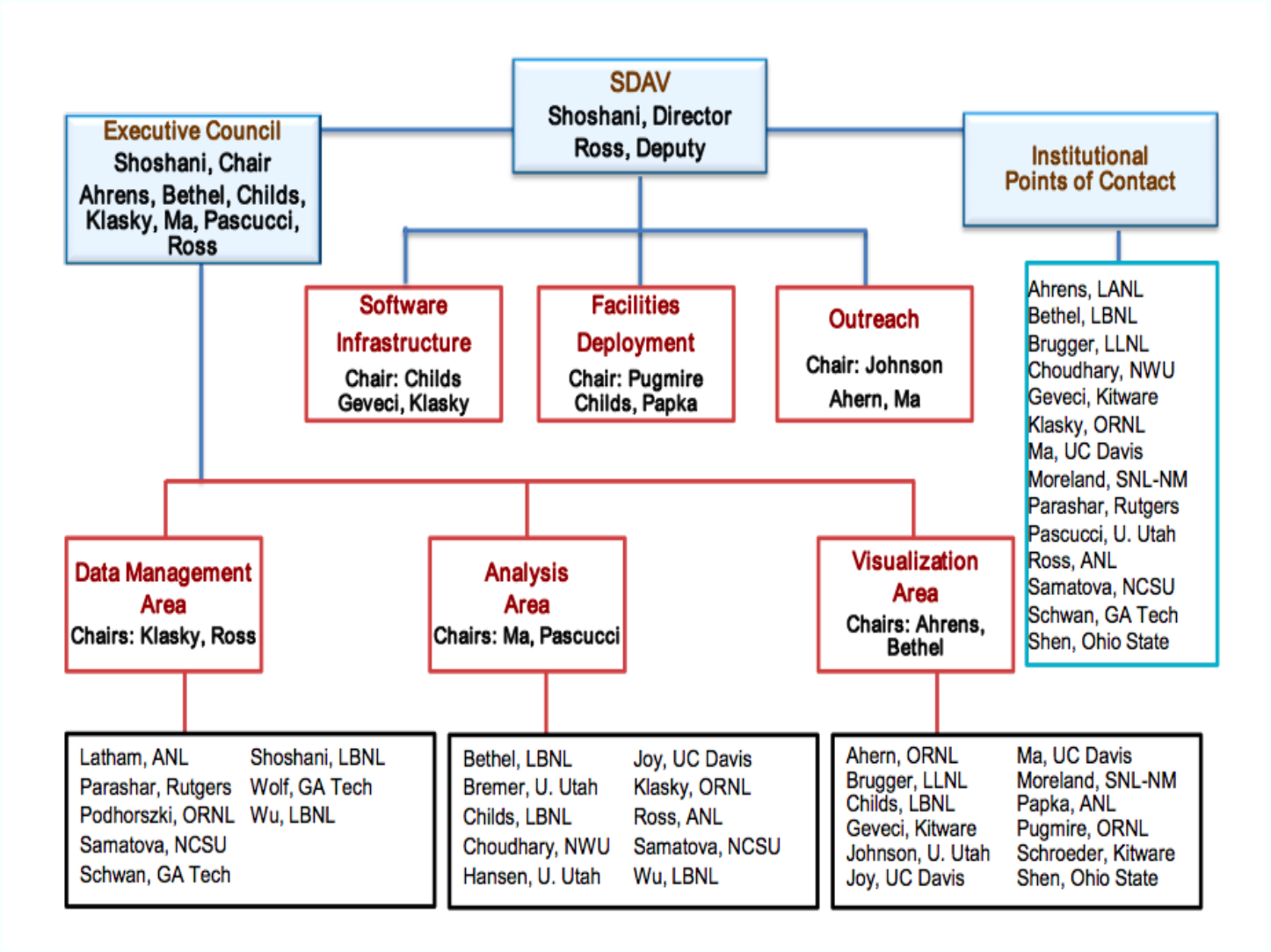
Architecture-Awareness Plan

- **Argonne Leadership Computing Facility.**
  - The ALCF's main computational resource is *Intrepid,* an IBM Blue Gene/P system with 40,960 quad-core compute nodes (163,840 processing cores) and 80 terabytes of memory. Its peak performance is 557 teraflops. *Mira* is the next generation Leadership Computing resource to be deployed at the ALCF in 2013. Mira is an IBM Blue Gene/Q system with 48K 16-core compute nodes and 768 terabytes of memory. Its peak performance will be 10 petaflops.

- **National Energy Research Scientific Computing Center.**
  - NERSC's main computational resource is *Hop- per,* a Cray XE6 system with 6,384 compute nodes with 24 processors per node (153,216 processing cores), 217 TB of memory. Its peak performance is 1.28 petaflops.
  -

- **Oak Ridge Leadership Computing Facility.**
  - The OLCF's main computational resource is *Jaguar,* a Cray XT5 system with 18,688 dual-socket, hex-core SMP nodes (12 cores per node) with 300 terabytes of mem- ory. Its peak performance is 2.33 petaflops. *Titan* is the next generation Leadership Computing resource to be deployed at the OLCF in 2012 - 2013. Titan will couple a traditional multi-core processor with a GPGPU processor optimized for multi-threaded performance and low power consumption. *Titan* peak performance will be 20 petaflops.

SDAV Institute Management Structure



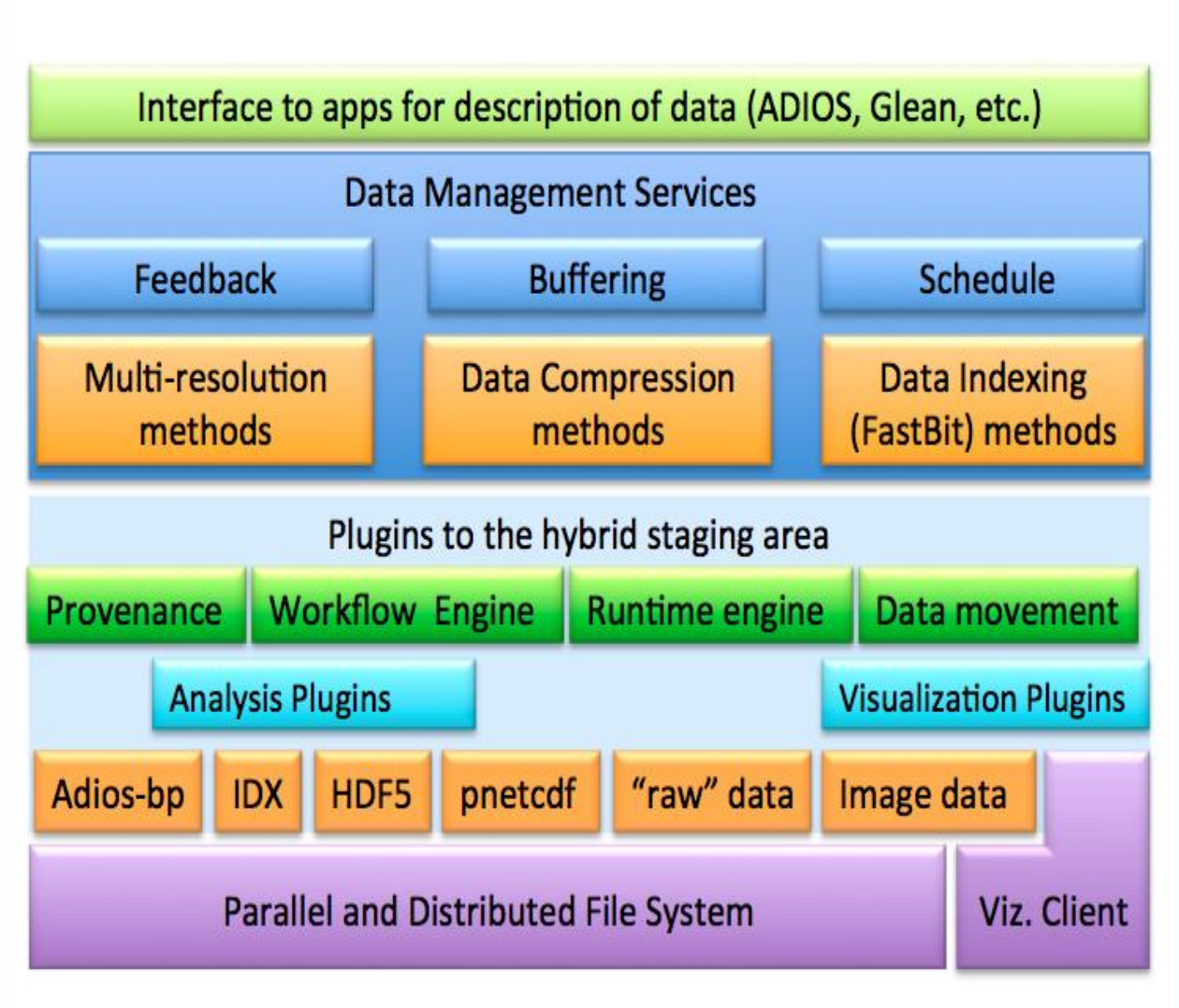
Application Needs and Use Cases

- According to our recent survey of application scientists, we observe that their greatest concerns when they scale their code are how to efficiently:
  - Write and subsequently analyze data
  - Compute and visualize data products
  - Perform remote visualization at the LCFs
  - Perform comparative/ensemble visualization
  - Perform I/O properly on LCFs as well as smaller clusters
  - Utilize in situ data processing

Applications and SDAV Technology Usage

Application	Code	Contact	2011/12 LCF Allocation	Institute technologies used
Astrophysics	Chimera	T. Mezzacappa	60	ADIOS, VisIt, Ultravis-V
Astrophysics	Flash	D. Lamb	80	PnetCDF, GLEAN, ROMIO, VisIt, VTK
Astrophysics	Maestro	J. Bell	50	VisIt
Astrophysics	Enzo	M. Norman	35	ParaView, VisIt
Astrophysics	HACC	S. Habib	150	ParaView, ROMIO, Ultravis-P
Biological	Nektar	G. Karniadakis	50	ParaView
Climate	POP	P. Jones	110	PnetCDF, ParaView, ROMIO
Combustion	S3D	J. Chen	60	ADIOS, Dataspaces, Ultravis-V, Ultravis-P, ViSUS IDX, Topologika
Combustion	Boxlib	J. Bell	40	VisIt, ADIOS , Topologika
Combustion	Nek5000	C. Frouzakis	150	VisIt
Fusion	GTC	Z. Lin	35	ADIOS, DataTap, FastBit, Ultravis-V
Fusion	XGC	C. S. Chang	50	ADIOS, Dataspaces, FastBit, Ultravis-V, VTK
Fusion	GTC-P	W. Tang	58	ADIOS, Ultravis-V, Ultravis-P
Plasma	VPIC	B. Daughton	30	PnetCDF, ParaView, ROMIO
Nuclear	Nek5000	P. Fischer	25	ROMIO, VisIt

SDAV Software Stack



Data Management Tools and Capabilities

- In Situ Processing and Code Coupling
  - ADIOS, Glean
- Indexing
  - FastBit
- In Situ Data Compression
  - ISABELLA
- Parallel I/O and File Formats
  - PnetCDF, BP-files, HDF5

Data Analysis Tools and Capabilities

- **Statistical and Data Mining Techniques**
  - NU-Minebench
- **Importance-Driven Analysis Techniques**
  - Domain-Knowledge Directed
  - Geometry Based
  - Information Theory Based
- **Topological Methods**
  - In Situ Topology
  - Feature-Based Analysis
  - High-Frequency Analysis and Tracking
  - Feature-Based Data Reduction
- **Vector Field Analysis**

Visualization

- Visit and ParaView
- VTK-m framework
- Flow Visualization Methods
- Rendering
- Ensembles, Uncertainty, and Higher-Dimensional Methods

The END

Summary of Plan for Integration of Visualization Accelerator Technology

Scientists

use general-purpose visualization tools such as:

these tools are built on the:

Visualization ToolKit Library

which will integrate these accelerator technology R&D efforts:

DAX

Focus: Visualization Developers

EAVL

Focus: Advanced Data Models

PISTON

Focus: Portability

VTK-Threading

Focus: Evolutionary performance improvement