BIG DATA AND SUPERCOMPUTING: FROM COLLISION TO CONVERGENCE

Arie Shoshani

Lawrence Berkeley National Laboratory

September 4, 2015

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

2012-2017

Arie Shoshani (PI)

Co-Principal Investigators from:

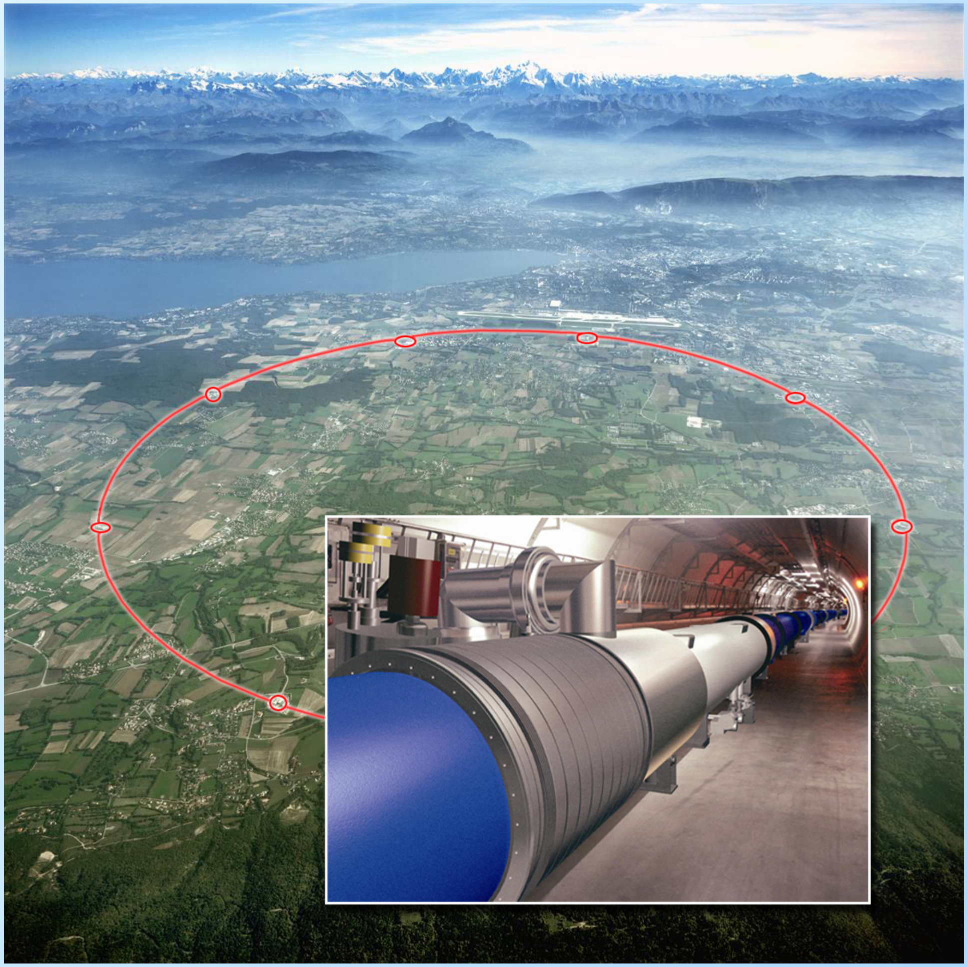
Laboratories	Universities
ANL	GTech
LBNL	NCSU
LLNL	NWU
ORNL	OSU
LANL	UCD
SNL	Rutgers
Kitware (Industry)	UUtah

http://sdav-scidac.org/

Example of Big Data in Science

Large Hadron Collider: to find the "God" particle (Higgs boson)

- sensors capable of 140PB/s
- reduce 99.99% of data by hardware triggers
- Keep 15 PB per year
- 27 km tunnel
- ~10,000 superconducting magnets
- Operating temperature 1.9 Kelvin
- Construction cost: US$9Billion
- Power consumption: ~120 MW



More Big Data at DOE

- Experimental Data
  - Light Sources
  - Nanoscience Centers
  - Neutron Facilities
  - The Joint Genome Institute (JGI)
  - Systems Biology Knowledgebase
  - The Environmental Molecular Sciences Laboratory (EMSL)
- Observational Data
  - Cosmological sky surveys
  - Cosmic microwave background telescopes
  - Neutrino detectors
  - Climate Atmospheric Radiation Measurements (ARM)

Advanced Light Source - ANL

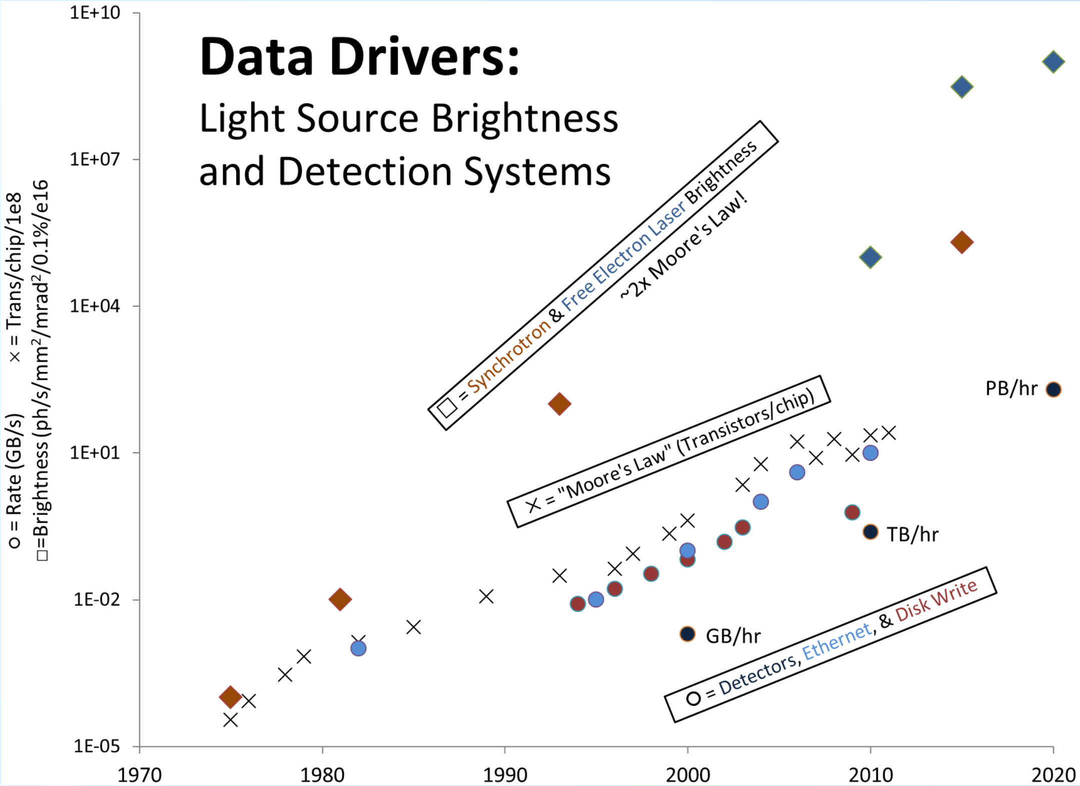
Advanced Photon Source - ANL

Nano facility - BNL

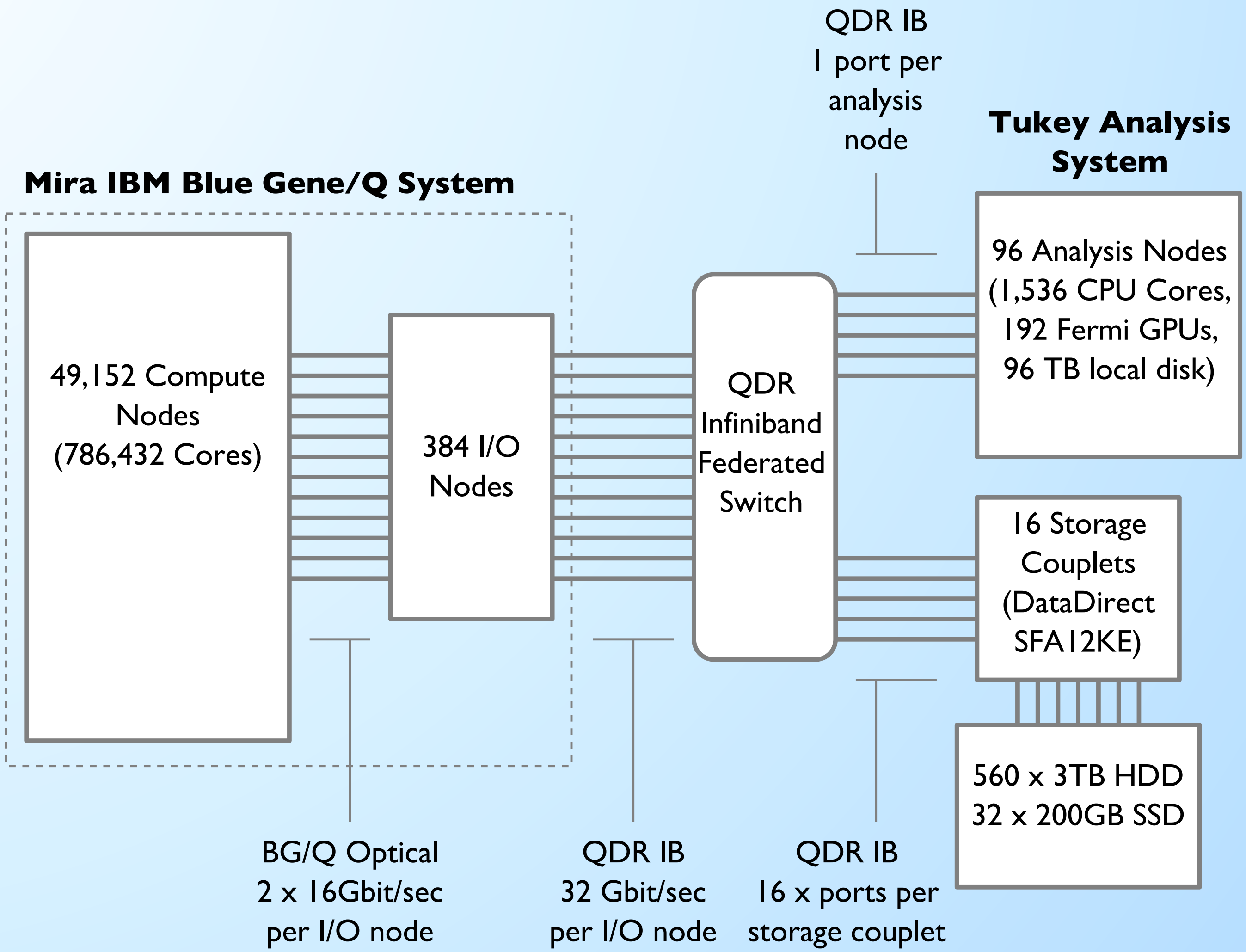
Light Sources – faster than moore's law

- Major light source facilities at DOE labs
  - The National Synchrotron Light Source II (NSLS-II) at BNL
  - The Advanced Photon Source (APS) at ANL
  - The Advanced Light Source (ALS) at LBNL
  - The Linac Coherent Light Source (LCLS) at SLAC
  - The Stanford Synchrotron Radiation Lightsource (SSRL) at SLAC.
- Higher resolution and faster speed

source: D. Parkinson, LBNL

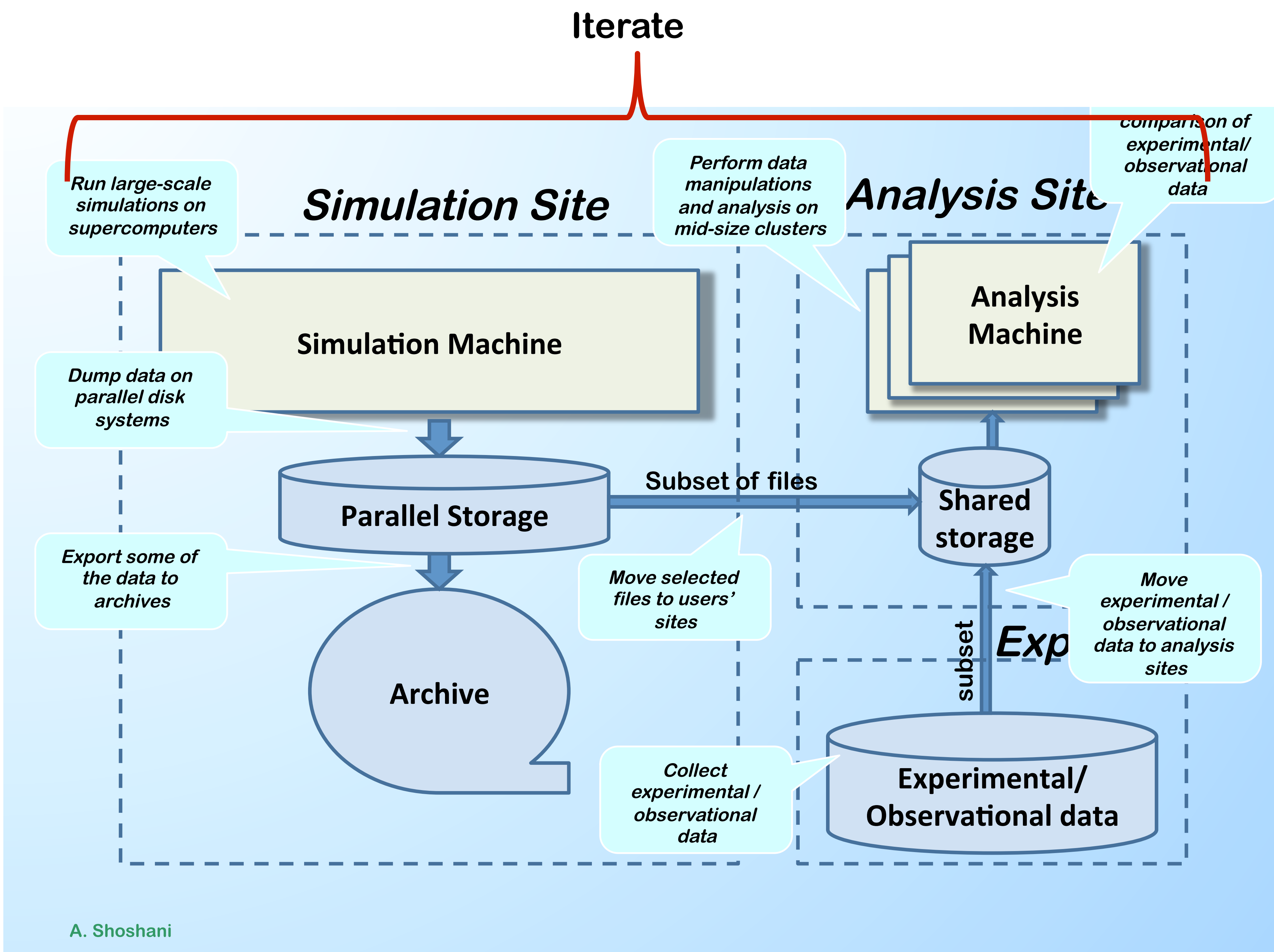


Leadership System Architectures



High-level diagram of 10 Pflop IBM Blue Gene/Q system at Argonne Leadership Computing Facility

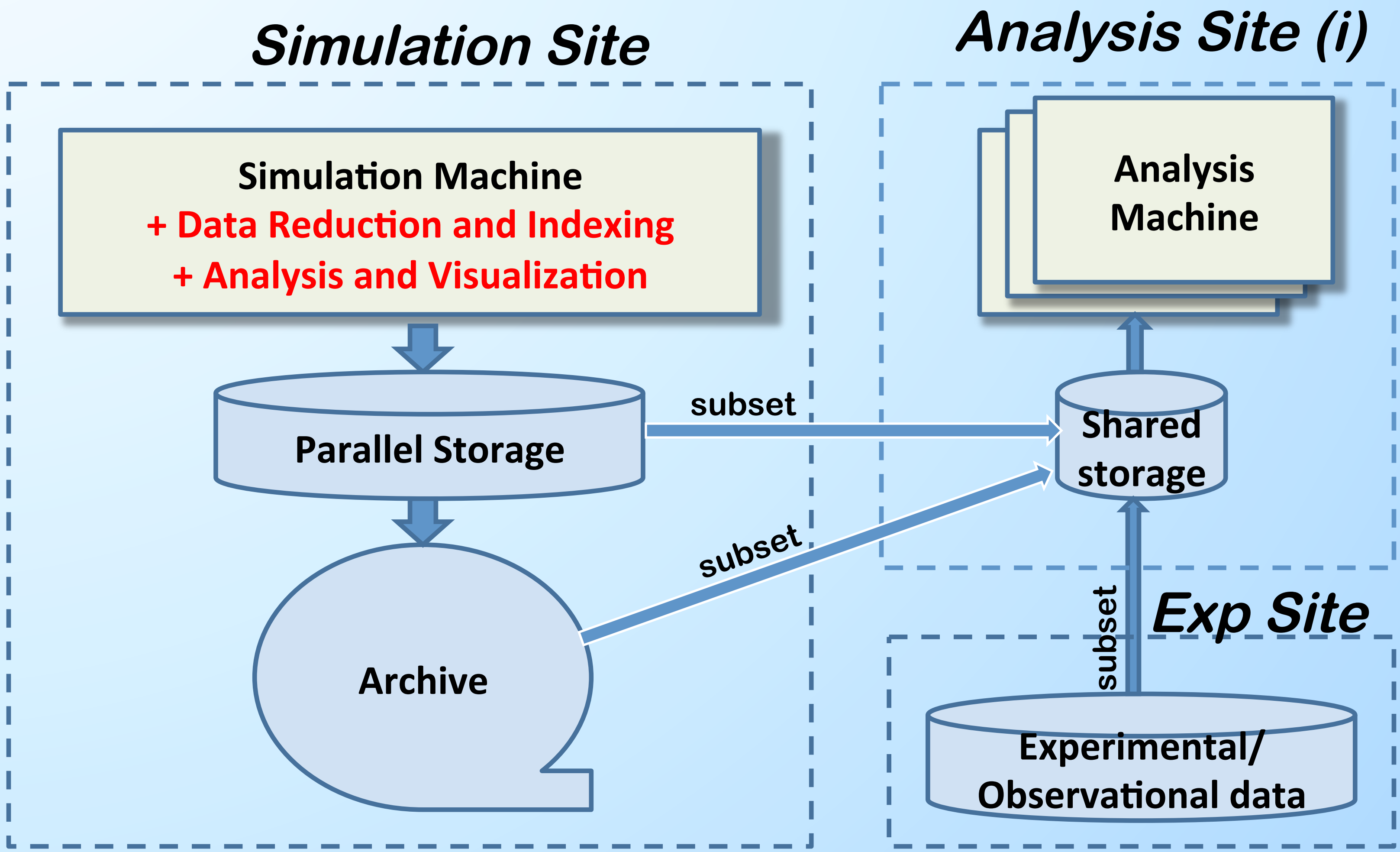
Courtesy: Rob Ross, ANL



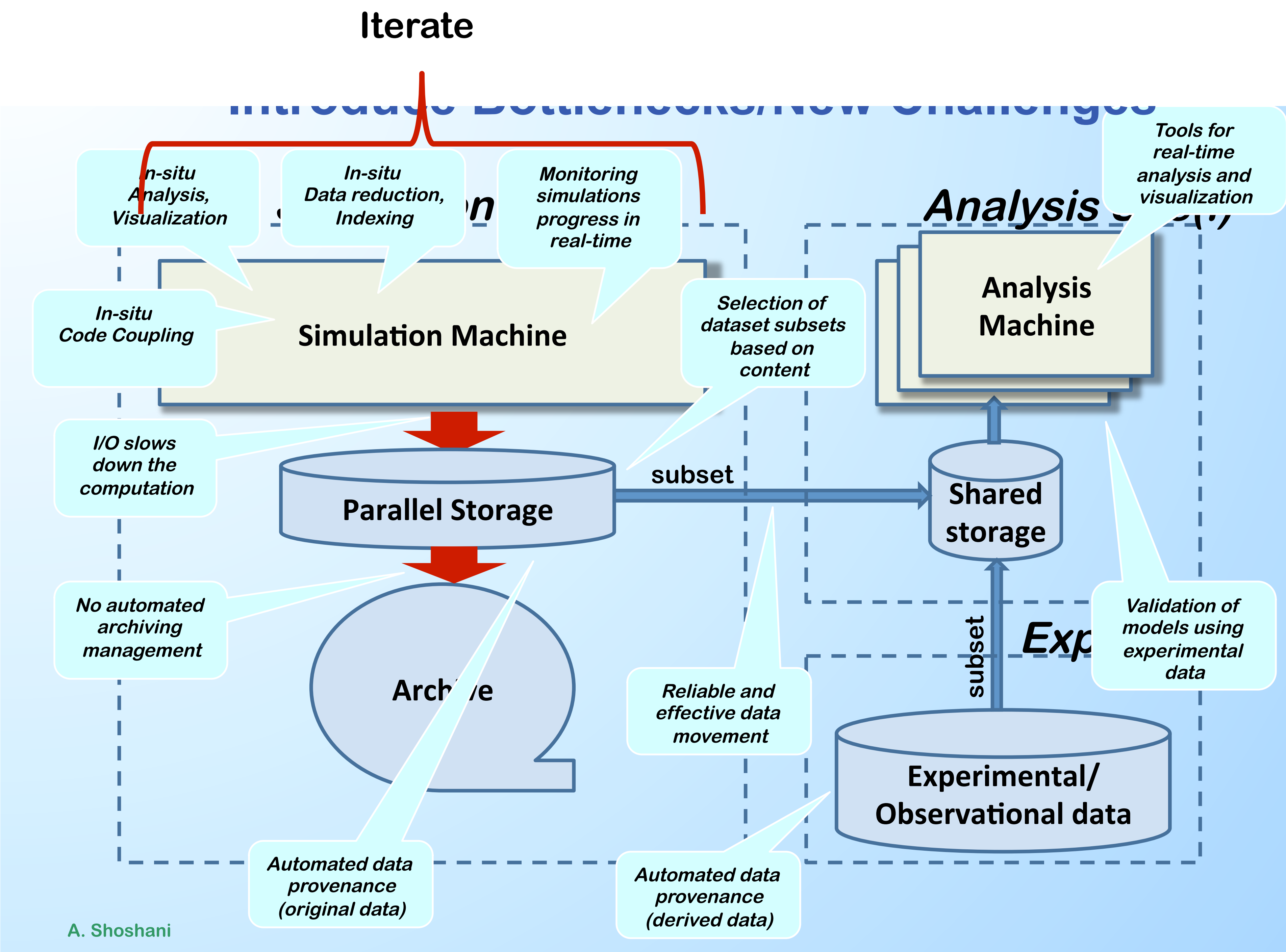
Iterate
Simulation Site
Analysis Site
Run large-scale simulations on supercomputers
Perform data manipulations and analysis on mid-size clusters
comparison of experimental/ observational data
Simulation Machine
Analysis Machine
Dump data on parallel disk systems
Parallel Storage
Subset of files
Shared storage
Export some of the data to archives
Move selected files to users' sites
Move experimental / observational data to analysis sites
subset
Exp
Archive
Collect experimental / observational data
Experimental/ Observational data
A. Shoshani

Data Volume Explosion: What Can be Done?

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



A. Shoshani

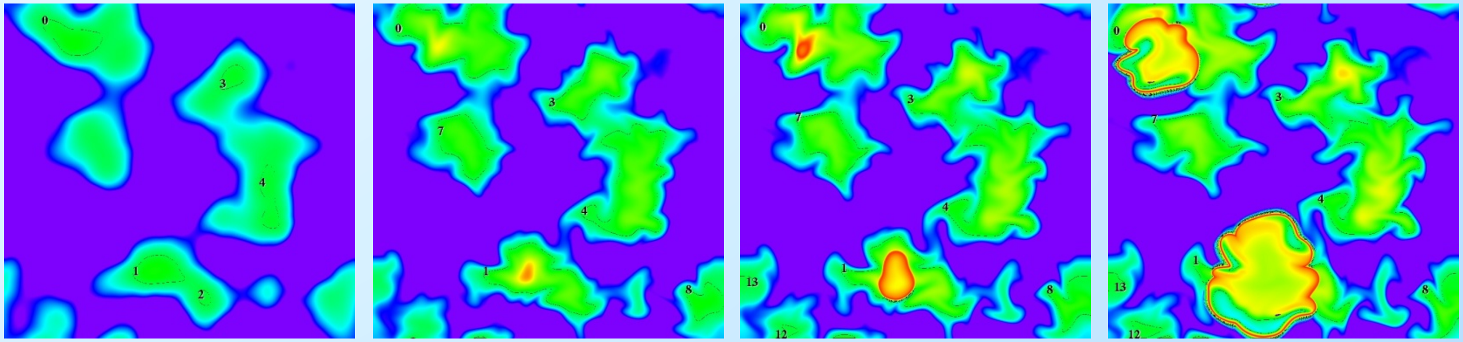

Iterate
In-situ Analysis, Visualization
In-situ Data reduction, Indexing
Monitoring simulations progress in real-time
Tools for real-time analysis and visualization
Analysis
In-situ Code Coupling
Simulation Machine
Selection of dataset subsets based on content
Analysis Machine
I/O slows down the computation
Parallel Storage
subset
Shared storage
No automated archiving management
Archive
Reliable and effective data movement
subset
Validation of models using experimental data
Exp
Experimental/ Observational data
Automated data provenance (original data)
Automated data provenance (derived data)
A. Shoshani

Data Analysis in Scientific Domains

- 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

Example of pattern matching in scientific domains

- Finding & tracking of combustion flame fronts
  - Cell identification: Identify all cells that satisfy conditions: “600 < Temperature < 700 AND HO2-concentr. > 10^{-7}”
  - Region growing: connect neighboring cells into regions
  - Region tracking: track the evolution of the features through time

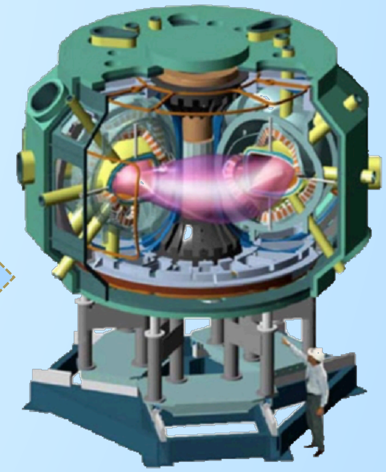


3D simulations: 1000x1000x1000, 1000 time steps = 10^{12} cells, 20 – 50 variables each cell. 400 TBs dataset.

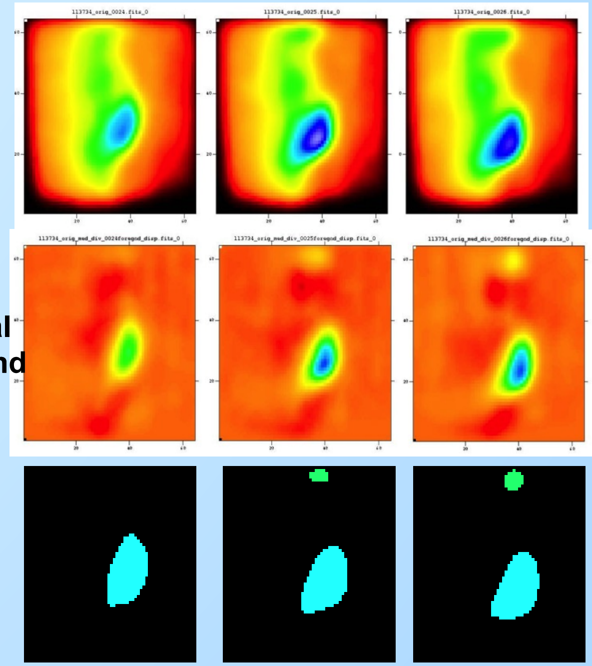
Example of pattern discovery in scientific domains

Feature Extraction and Tracking in Fusion Images

- National Spherical Torus Experiment (NSTX) – in PPPL (Tokomak)
- Images of experiment contain “**blobs**”
- A blob is a coherent structure in the image that carries heat and energy from the center of the torus to the wall



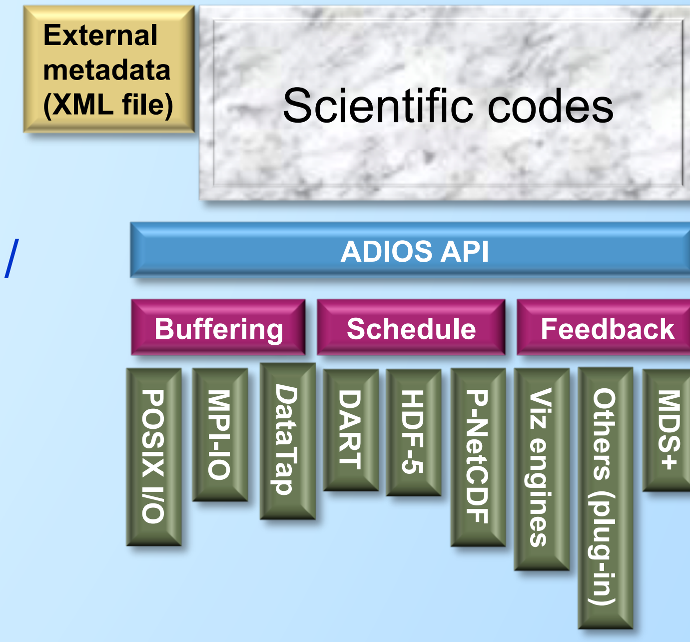
- Analysis required
  - Remove the quiescent background intensity
  - Remove ambient background intensity
  - Use image processing techniques to identify and track blobs over time

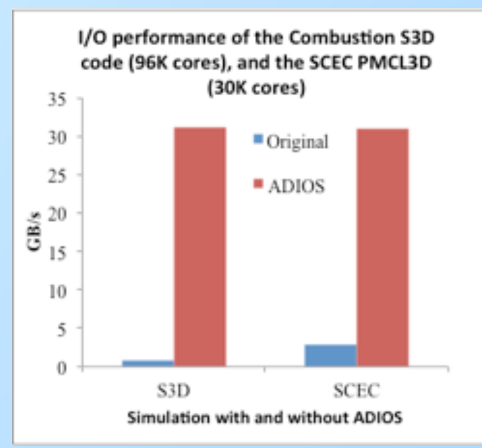


Contact: Chandrika Kamath, LLNL

ADIOS: Adaptable I/O System (ORNL)

- 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 in situ optimization of I/O (e.g. Aggregated read/write)



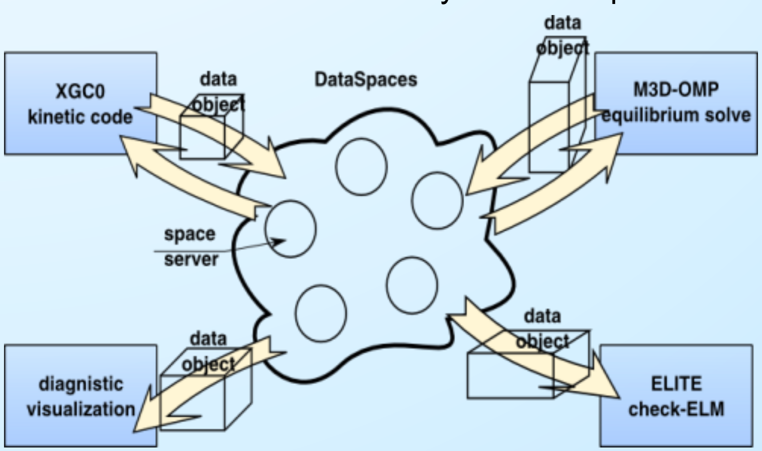


Contact: Scott Klasky, ORNL

What else can be done when capturing I/O

Code coupling with DataSpaces

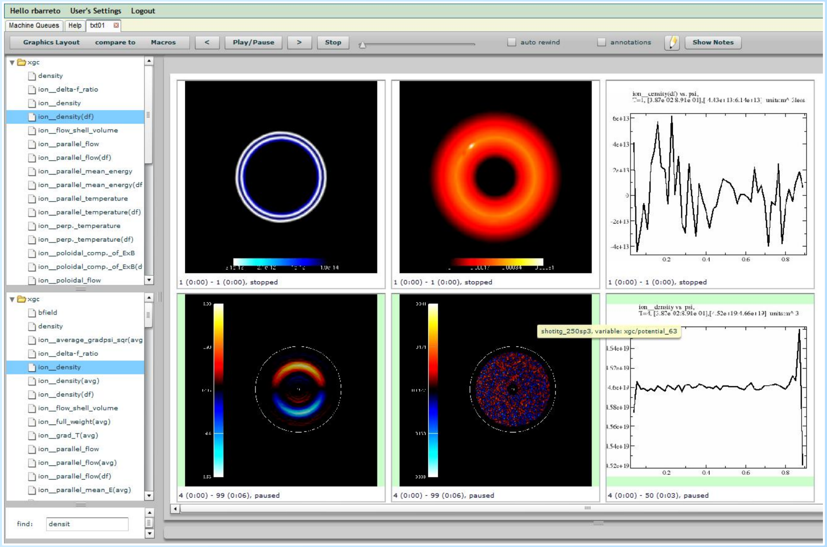
- Virtual shared space
- Constructed on-the-fly on staging nodes
    - Indexes data for quick access and retrieval
    - Provides asynchronous coordination and interaction and realizes the shared-space abstraction
- In-space (online) data transformation and manipulations
- Decentralized data analysis in-the-space



- **Code Coupling: get, put, publish, subscribe service**
- **In transit data movement and workflow management**

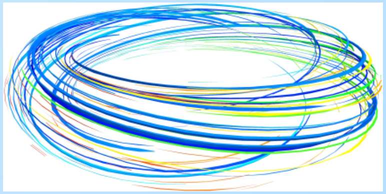
Real-time monitoring on a Dashboard

- Time-step data is captured and processed by a workflow system
- Movies generated by a workflow
- Interactive vector plotting



Imbed FastBit indexing to find regions of interest

- Set of regions with high electromagnetic potential in a torus modeled by GTC
- Achieved speed up of 500-900 fold on magnetic coordinates



Contact: Scott Klasky

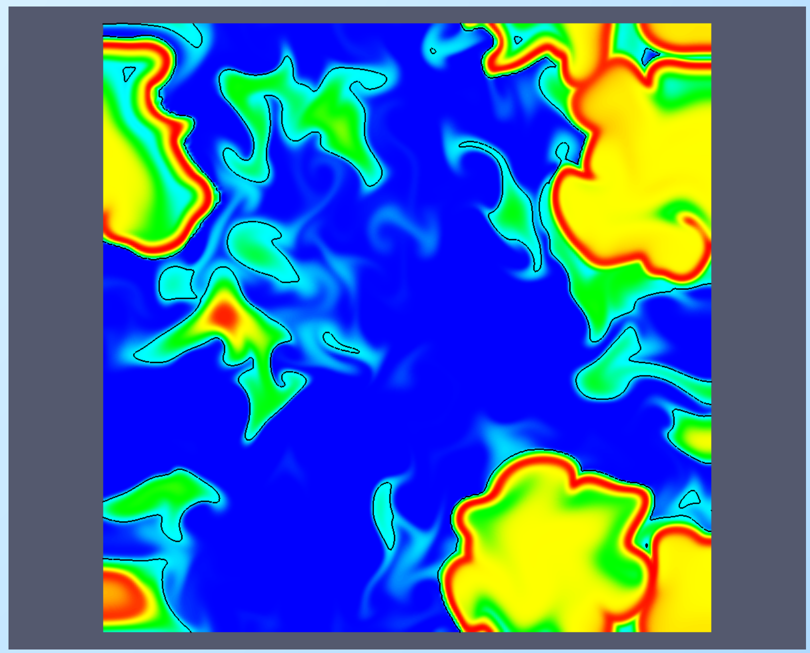
The need for selection of subsets based on content

Combustion simulation: 1000x1000x1000 mesh with 100s of chemical species over 1000s of time steps – 10^{14} data values

- This is an image of a single variable (temperature)
- What's needed is search over multiple variables, such as:

  Temperature > 1000
  AND pressure > 106
  AND HO2 > 10^{-7} AND HO2 > 10^{-6}

- Challenges
  - Multi-variable queries from a subset of variables
  - Search over numerical values
  - Identify large number of regions



A. Shoshani

FastBit properties – highly efficient and compact

Main idea:

- Invented specialized compression methods (was patented) that:
  - Can perform logical operations directly on compressed bitmaps
  - Excels in support of multi-variable queries
  - Can partition and merge bitmaps without decompression – essential for parallelization of indexes

FastBit takes advantage of append only data to achieve:

- Search speed by 10x – 100x than best known bitmap indexing methods
- On average about 1/3 of data volume compared to 2-3 times in common indexes because of compression method
- Proven to be theoretically optimal – data search time is proportional to size of the result

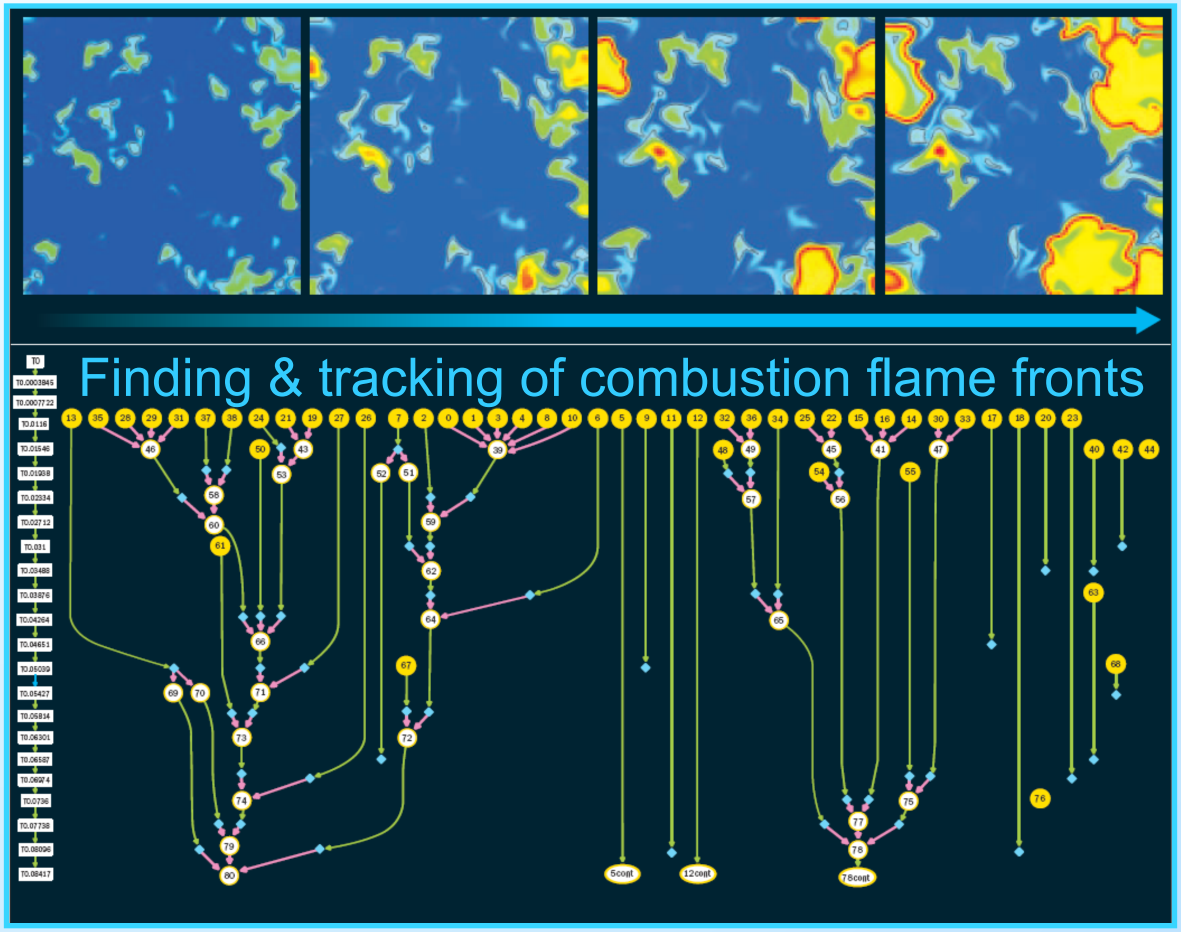
Usage

- In multiple scientific application in DOE
- Embedded into in situ frameworks
- Thousands of downloads around the world (open source under source forge), including commercial companies

Contacts: John Wu, Arie Shoshani

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

Contacts: John Wu, Arie Shoshani

Query-Driven Visualization: Laser Plasma Accelerator Modeling

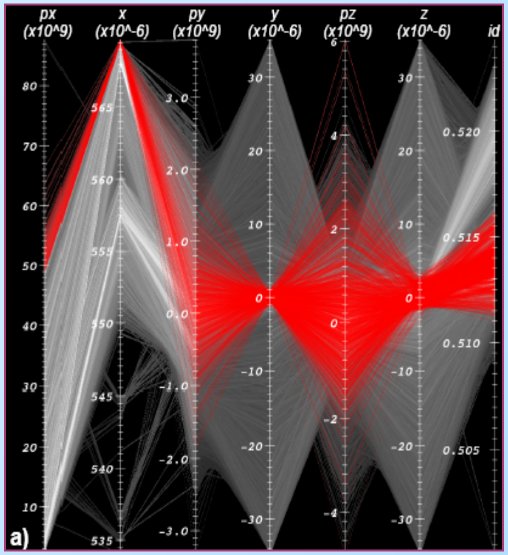
Beam Selection

Parallel coordinates view of *t* = 12

- Grey particles represent initial selection ($px > 2*10^9$)
- Red particles represent "focus particles" in first wake period following pulse ($px > 4.856*10^{10}$) && ($x > 5.649*10^{-4}$)

Volume rendering of plasma density with focus particles included in red (*t* = 12)

- Helps locate beam within wake



Impact

- Enable for the first time detailed analysis of massive particle dataset of unprecedented size in real-time
- 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)

Contacts: Wes Bethel, John Wu

Some panel questions and some answers

- How should HPC change to meet its Big Data needs?
  - HPC already has to deal with Big Data!
  - In-situ data processing, analysis and visualization is essential
- What are the architecture tradeoffs: filesystems vs. object stores?
  - Most scientific data is now in filesytems (Luster, GPFS)
  - Challenge – how to bridge the gap
- What are the workflow requirements
  - In-memory workflow is essential
  - Automated metadata and provenance is essential
- Can the HPC and Big Data communities learn from each other?
  - Yes, especially for analysis tasks