# Scientific Data Management Center


Presented by

## Nagiza F. Samatova

Oak Ridge National Laboratory

## Arie Shoshani (PI)

Lawrence Berkeley National Laboratory


### Coprincipal 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
- **U. Utah**: Claudio Silva

# Scientific Data Management (SDM) Center

**Lead institution: LBNL**

**PI:** Arie Shoshani

**Laboratories:**

ANL, ORNL, LBNL, LLNL, PNNL

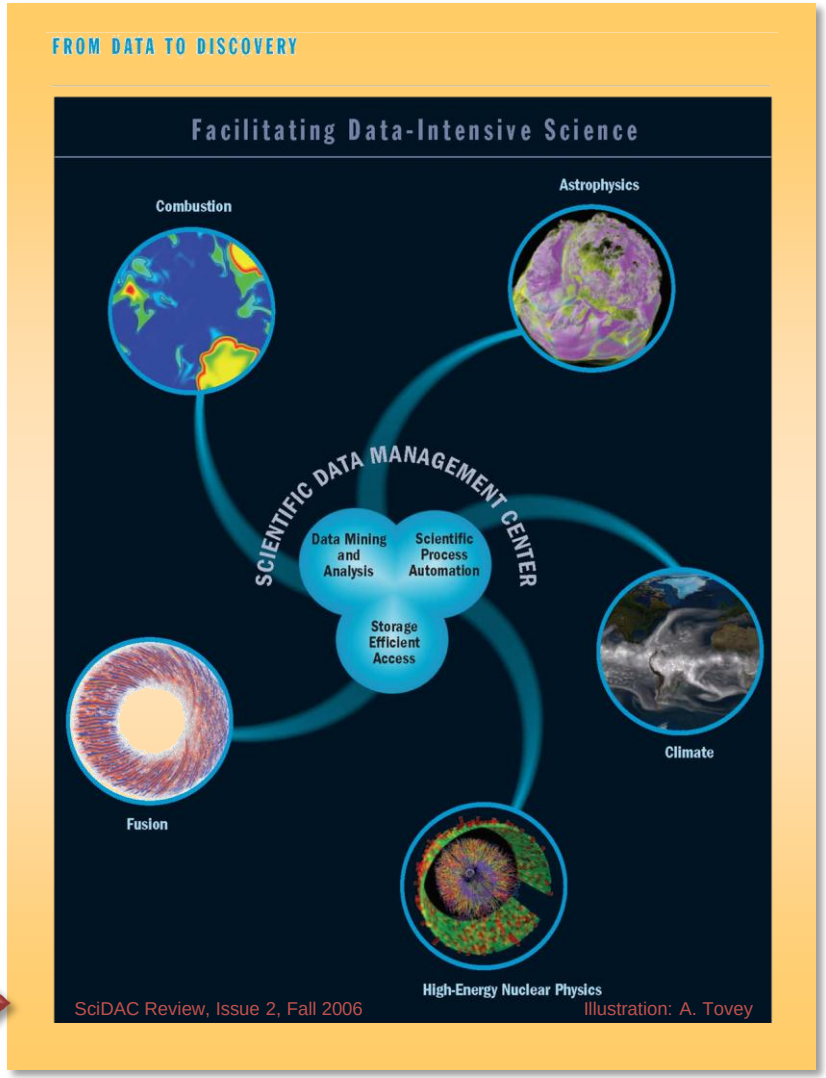
**Universities:**

NCSU, NWU, SDSC, UCD, U. Utah

Established 6 years ago (SciDAC-1)

Successfully recompeted for next 5 years (SciDAC-2)

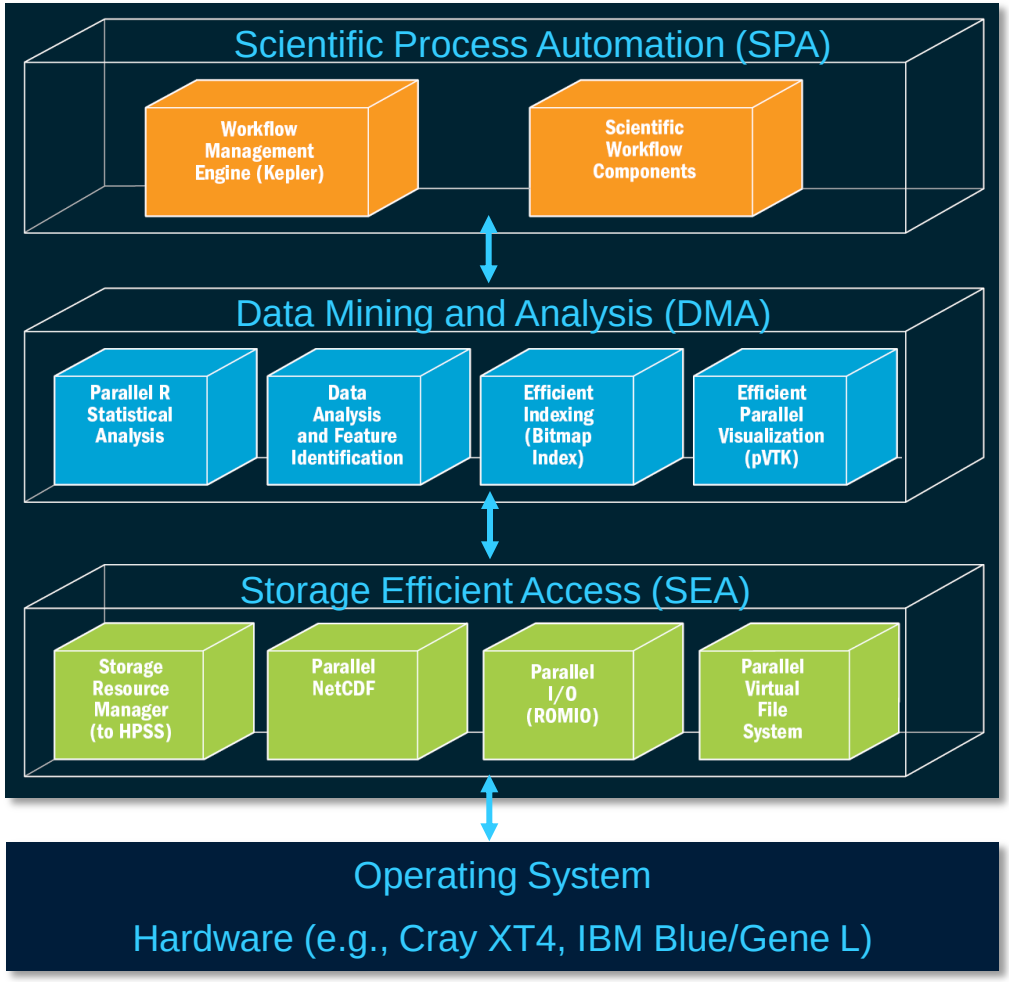
Featured in fall 2006 issue of *SciDAC Review* magazine



# SDM infrastructure
## Uses three-layer organization of technologies

**Goal: Reduce data management overhead**

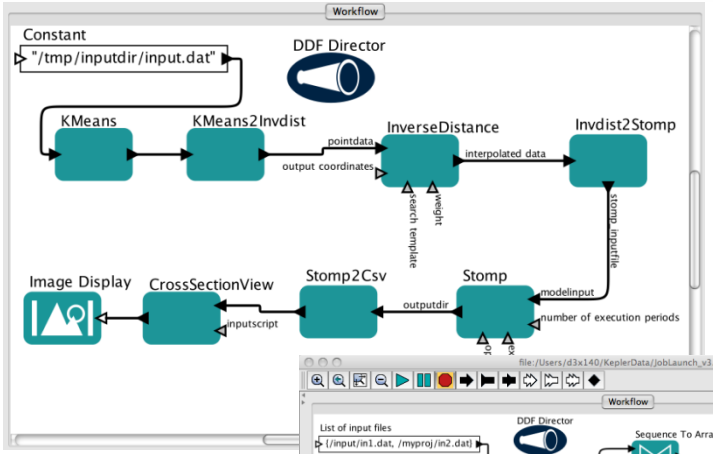
- Integrated approach
  - To provide a scientific workflow capability
  - To support data mining and analysis tools
  - To accelerate storage and access to data
- Benefits scientists by
  - Hiding underlying parallel and indexing technology
  - Permitting assembly of modules using workflow description tool



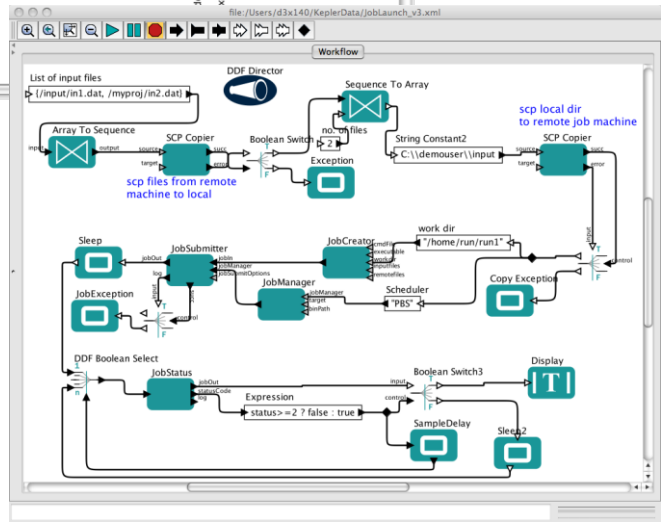
# Subsurface flow and transport modeling workflows

**Working closely with groundwater scientists to develop scientific workflows for executing subsurface flow and transport simulations**

- **Top-level workflow generates a conceptual model, interpolates the conceptual model to an input grid, executes the simulation, and visualizes simulation results**
- **Simulation execution workflow stages input files, submits a groundwater simulation job for execution, and then monitors the status of the job**




*Top-Level Workflow*




*Simulation Execution Workflow*

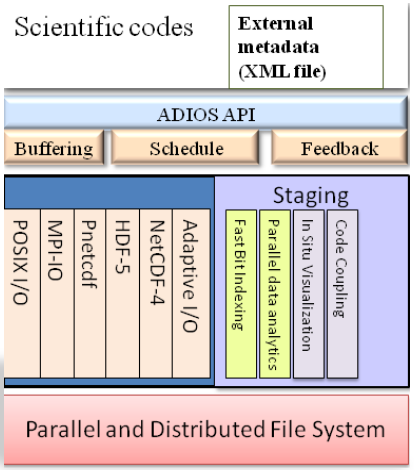
**Contact: Terence Critchlow (terence.critchlow@pnl.gov)**

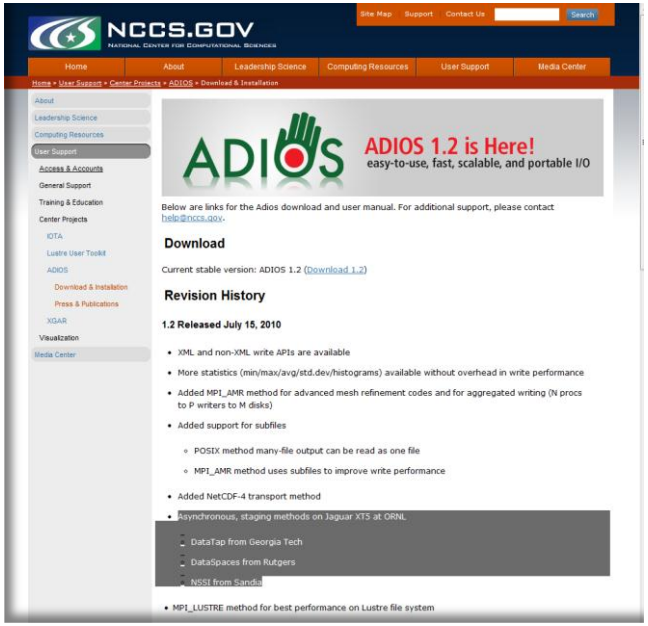
# ADIOS 1.2: Preparing for the Xscale

- Provides portable, fast, scalable, easy-to-use, metadata-rich output
- Near peak performance for reading and writing on Cray XT platform for many large simulations at scale
- Simple API, tested on over 240K cores
- Change I/O method by changing XML file only
- Layered software architecture
  - Allows plug-ins for different I/O implementations
  - Abstracts the API from the method used for I/O
- Open source

  http://www.nccs.gov/user-support/center-projects/adios/
- Provides open source framework for in situ I/O pipelines
- Research methods from many groups

  Examples: **Rutgers:** DataSpaces/DART **Georgia Tech:** DataTap
  **Sandia:** NSSI, Netcdf-4 **ORNL:** MPI_AMR

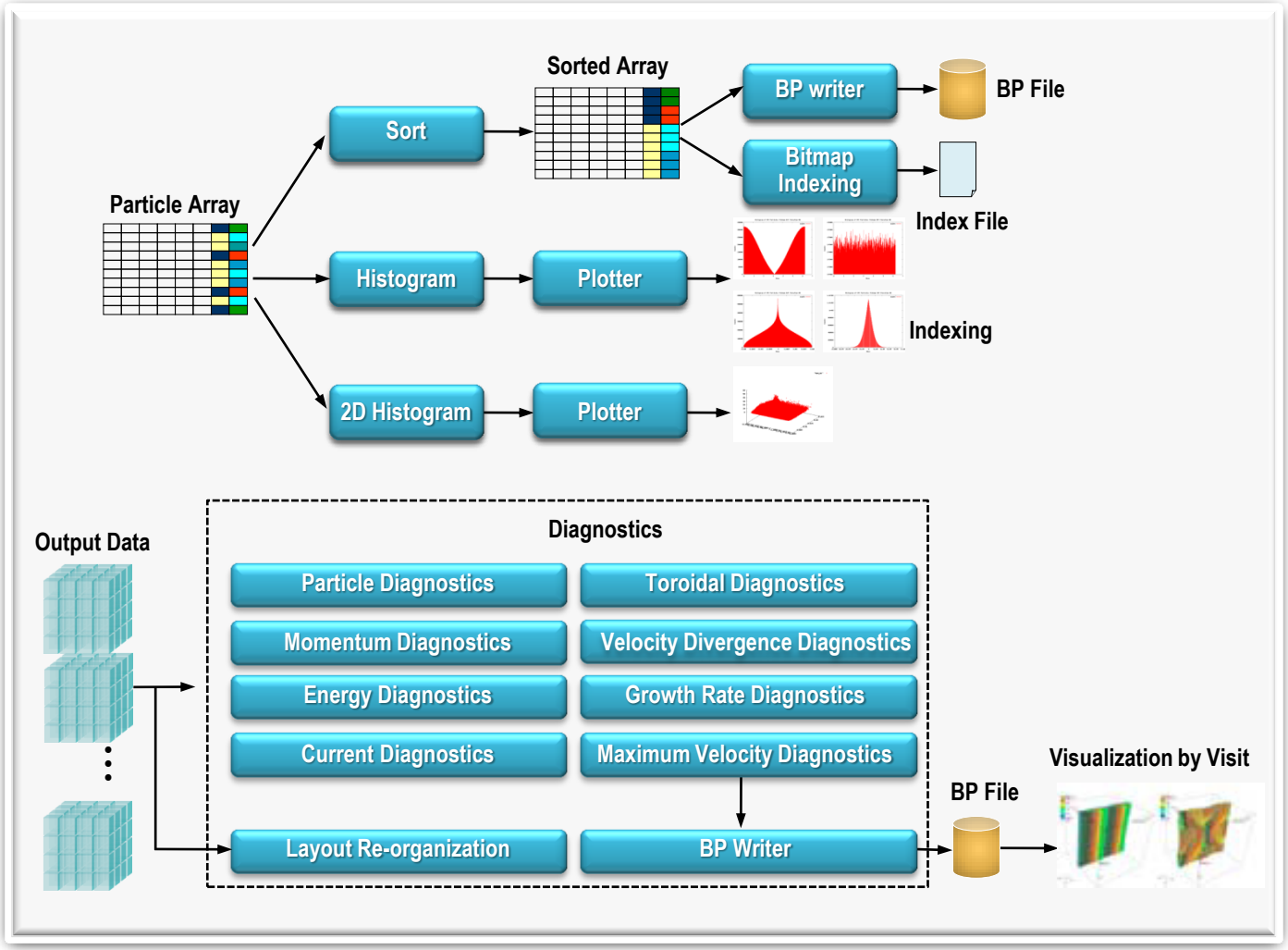


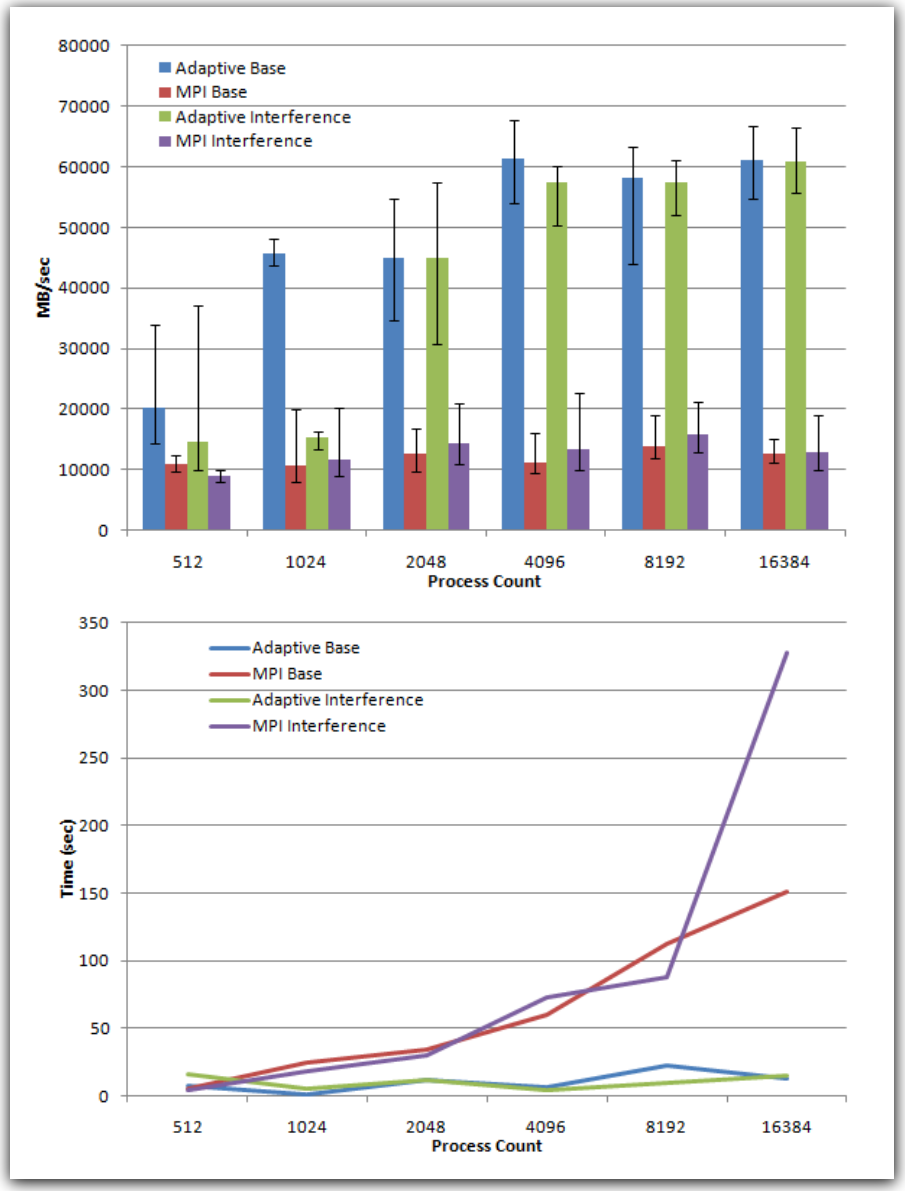

**Contact: Scott Klasky, ORNL (klasky@ornl.gov)**

# ADIOS 1.2: I/O pipelines

- Use the staging nodes and create a workflow in the staging nodes
- Allow us to explore many research aspects
- Improve total simulation time by 2.7% over synchronous writes to disk
  - While getting more science done, by scheduling I/O and not altering code



# ADIOS write performance for XGC1




- Idea is to get a better QoS for I/O, with high performance
- Older I/O methods show performance problems on 16K cores, with a high level of variability (red, purple)
- Best performance on Jaguarpf (OLCF) is ~60 GB/s on a loaded system
- New ADIOS methods shown in green (with outside users on the system) and blue (with no outside users)
  - Show the standard deviation of the write time on 16K sockets = 196608 cores

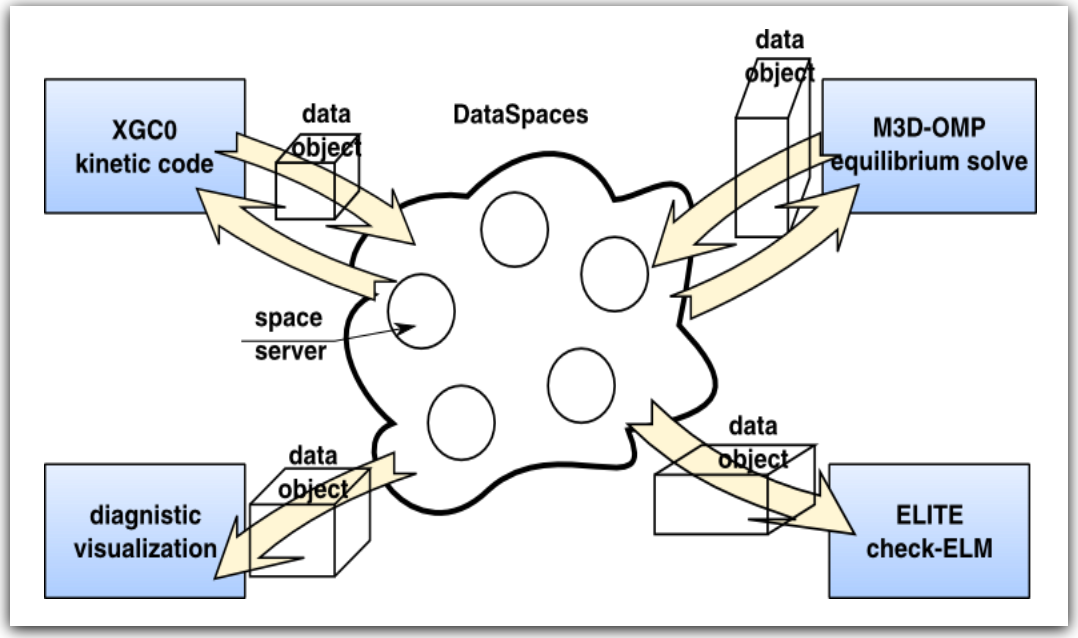


ADIOS

OAK RIDGE National Laboratory

# ADIOS with DataSpaces for in-memory loose code coupling

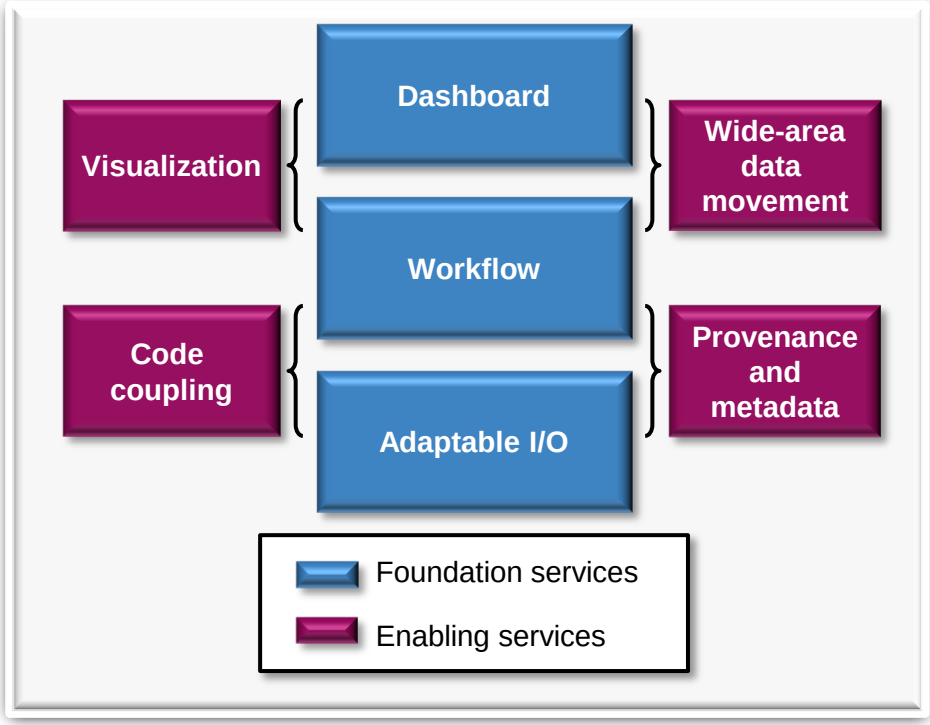
- Semantically specialized virtual shared space
- Constructed on-the-fly on the cloud of staging nodes
  - Indexes data for quick access and retrieval
  - Provides asynchronous coordination and interaction and realizes the shared-space abstraction
- Complements existing interaction/coordination mechanisms
  - In-memory code coupling becomes part of the I/O pipeline




- Supports complex geometry-based queries
- In-space (online) data transformation and manipulations
- Robust decentralized data analysis in-the-space

# Framework for integrated end-to-end SDM technologies

- 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

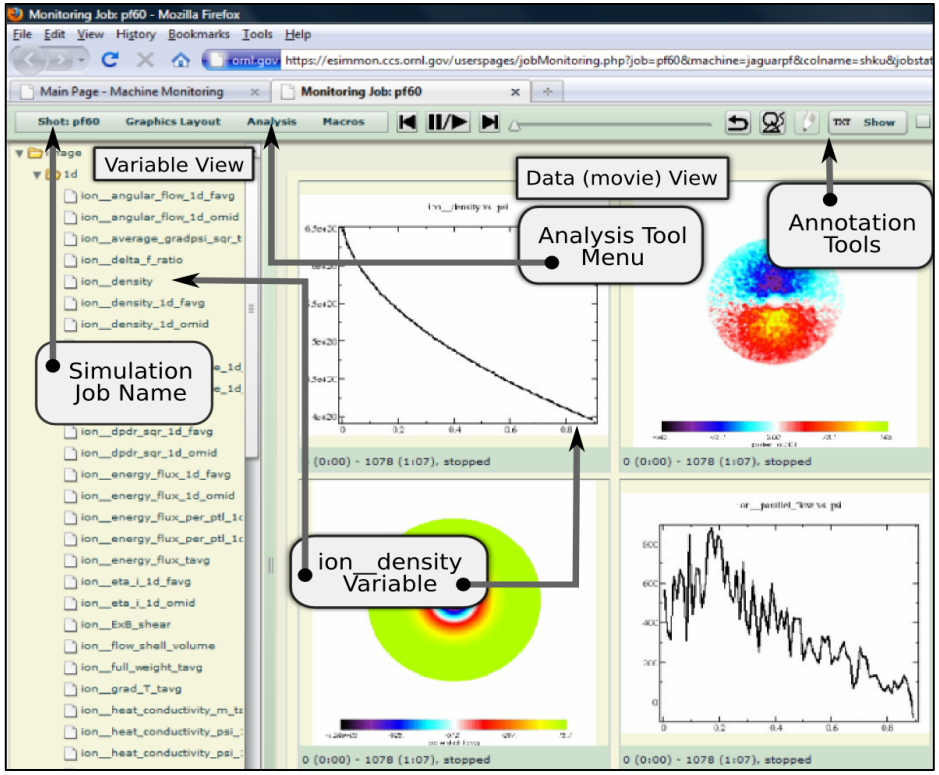
**Benefit:** Automates complex tasks and allows users to interact through simple interfaces that expose physics products remotely over the web

Contact: Scott Klasky, ORNL (klasky@ornl.gov)

# ORNL dashboard

## Some Characteristics

- **Functionality:** e.g., at-a-glance view of the status and health of a simulation. Scientists can monitor a running simulation or explore and manipulate data from past simulations from any web browser
- **Ease of use:** e.g., does not require the end user to be an expert on File Systems or Visualization tools
- **Collaboration:** e.g., allows scientists to collaborate and focus just on the science. The eSimMon dashboard organizes users' and collaborators' simulation runs (shots)
- **Annotation:** e.g., users can annotate shots and search through several archived shots
- **Data reduction and abstraction:** raw data into
- **Analysis:** Provides several types of analysis tools



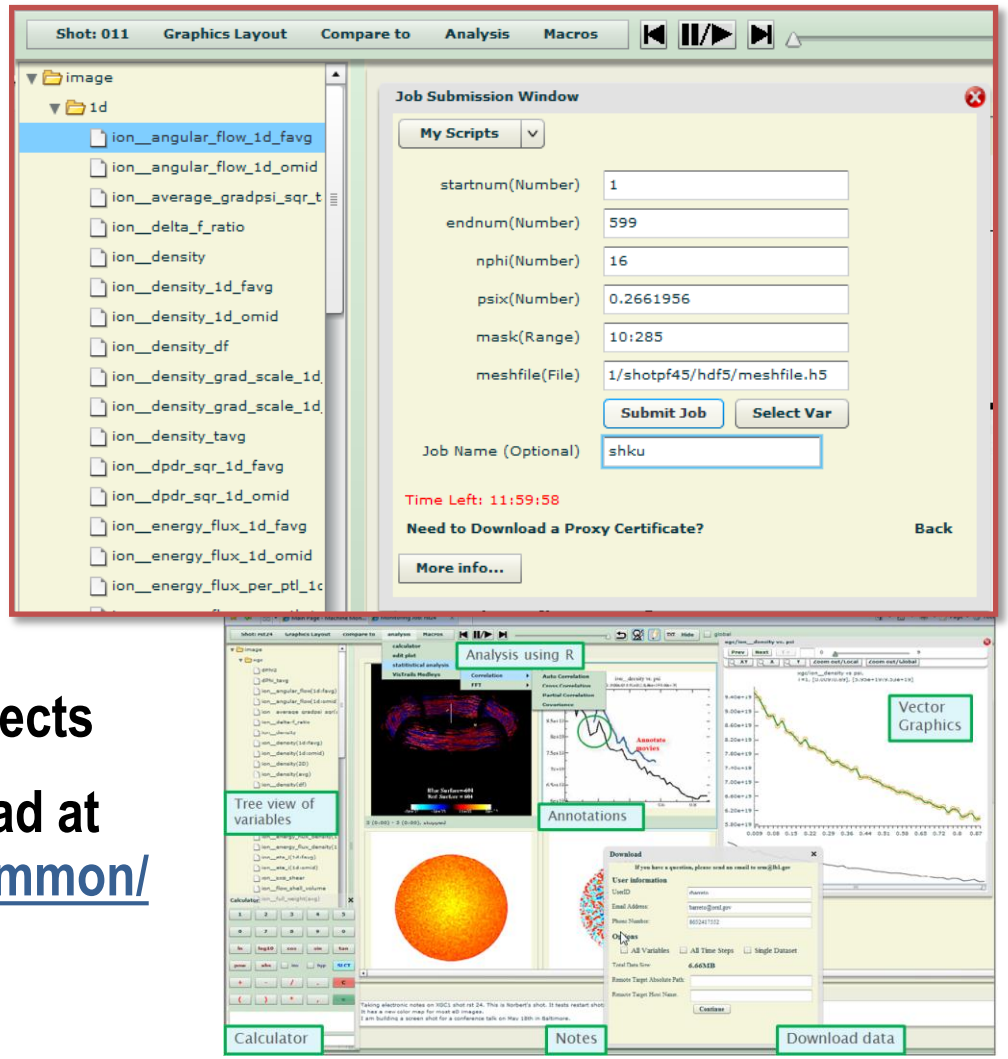

***Motivation for eSimMon***

Live version of the dashboard is at ORNL but requires an NCCS account for access. Therefore, we created eSimMon dashboard for other users

# eSimMon 1.0

- Creation of a data analysis facility
- Provenance information tracked and displayed on the SDM dashboard
- Matlab jobs allow users to analyze data
- Nov. 10 open source release
- Analysis and Visualizations results stored in MySQL database
- Used in CPES, GTC, GTS, S3D projects
- Dashboard 1.0 is ready for download at [http://users.nccs.gov/~rbarreto/esimmon/](http://users.nccs.gov/~rbarreto/esimmon/)

Matlab jobs launched from eSimMon



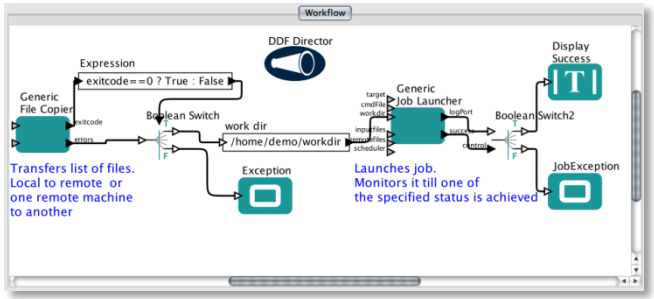

**Contact: Scott Klasky, ORNL (klasky@ornl.gov)**

# Scientist-centered workflow abstractions

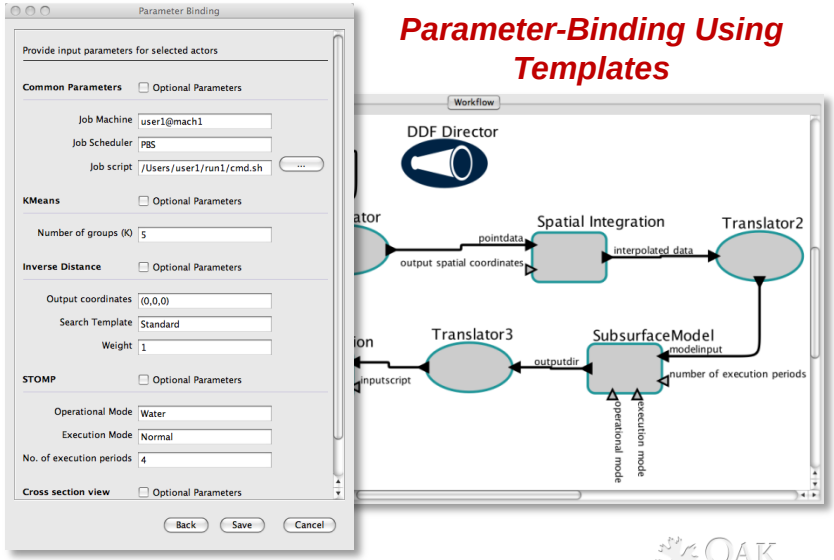
**Research and development of workflow abstractions to directly support domain scientists in workflow design and execution**

- Generic actors that encapsulate multiple functions or protocols into more general versions
- Abstract workflows and components to which scientists may annotate to collaborate on tasks and share knowledge
- Parameter-binding workflow templates that will collect and bind workflow and actor parameters at design or run time
- Actor-binding workflow templates that will bind functionality to abstract actors at design time
- Context-awareness mechanisms that can collect context information from the computational and workflow environment and provide to adaptive actors and workflows

***Generic Actors***




***Parameter-Binding Using Templates***

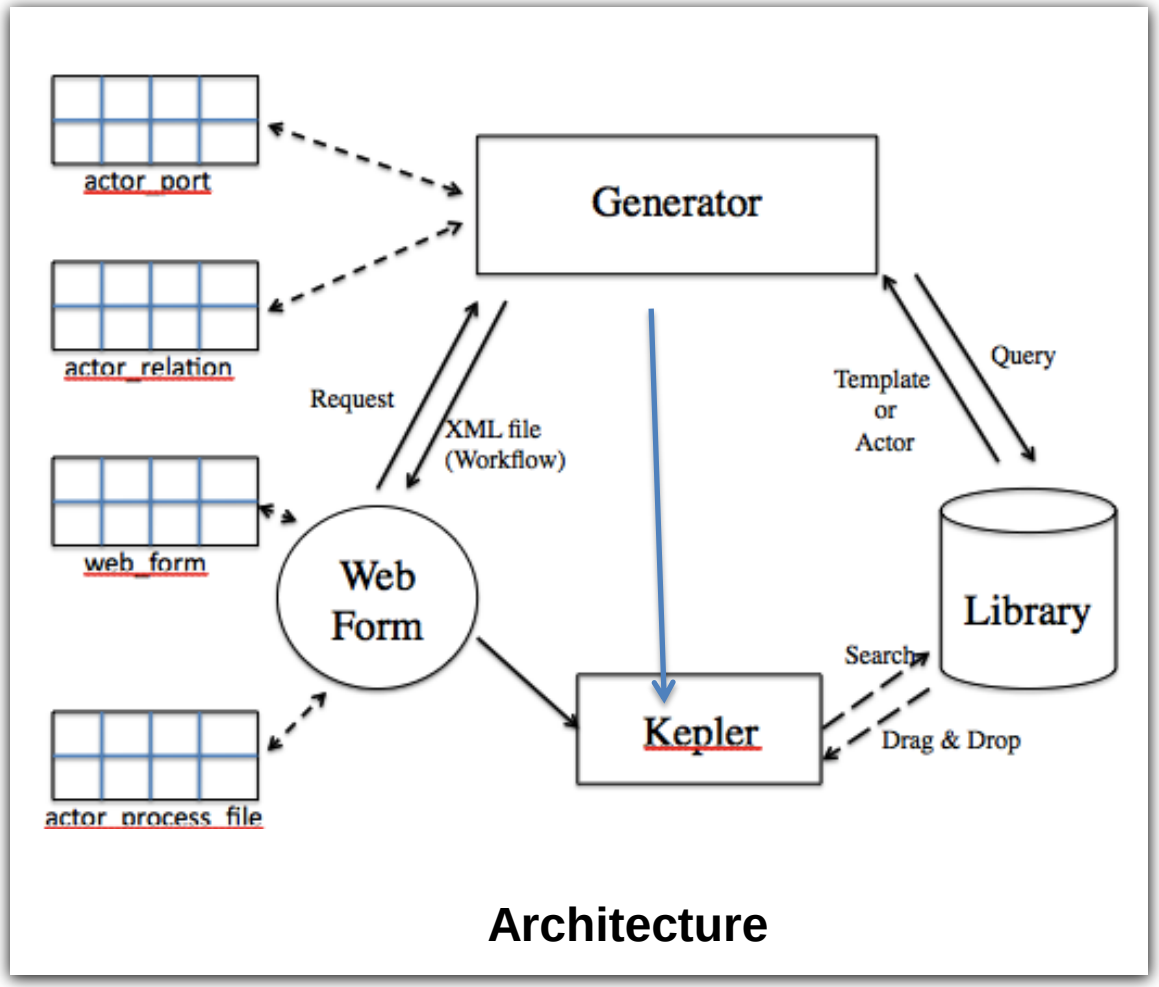



**Contact: Terence Critchlow (terence.critchlow@pnl.gov)**

# Workflow generator wizard

## Abstract and simplify workflow definition

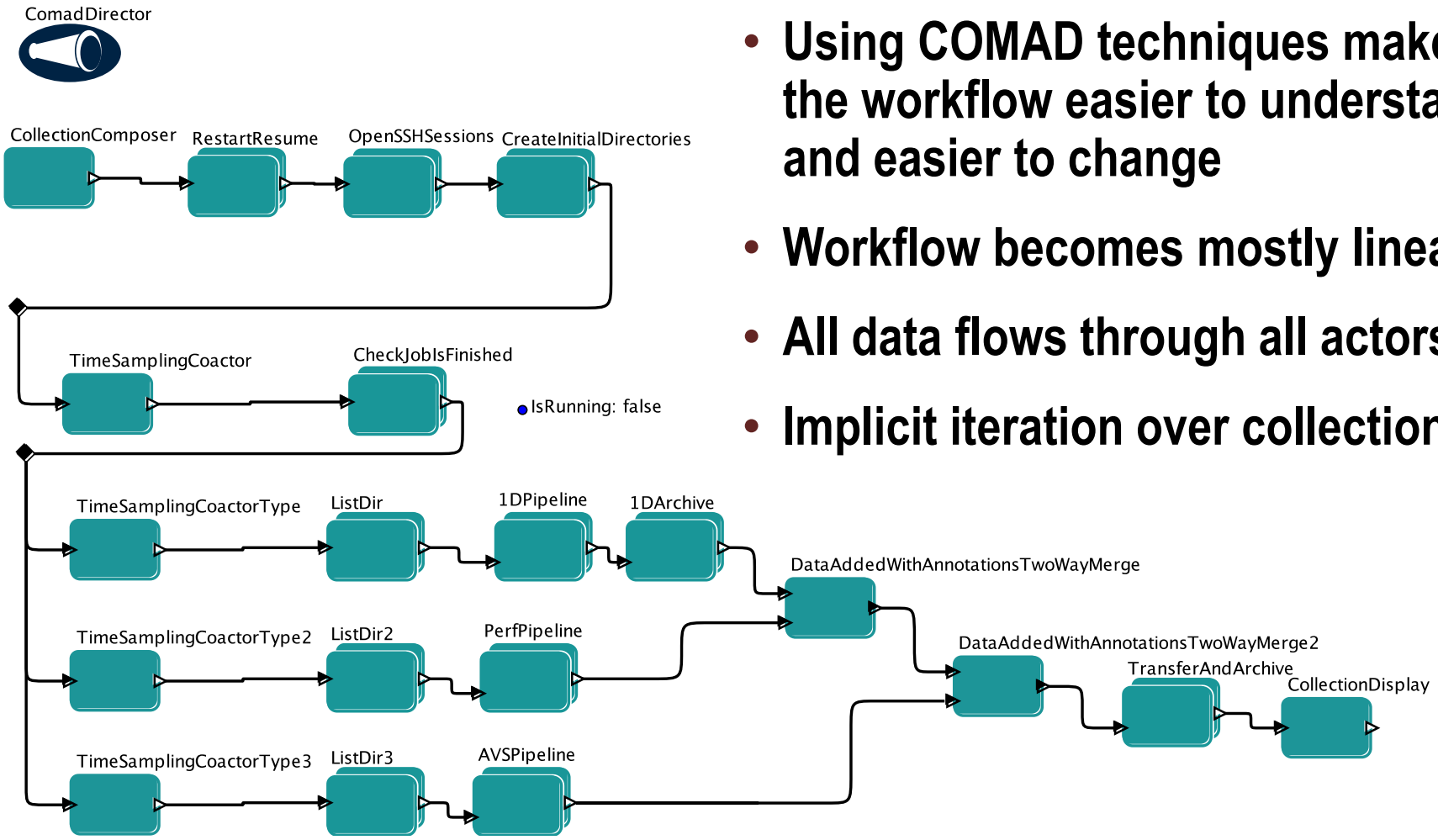
- Web Form
  - User fills in the options on the web form and clicks the submit button
- Library
  - Contains the actors and the templates
- Generator
  - Extracts the information from the web form and uses the library to build the workflow
  - It uses Kepler to view and execute the workflow



Architecture

Contact: Malden Vouk (vouk@ncsu.edu)

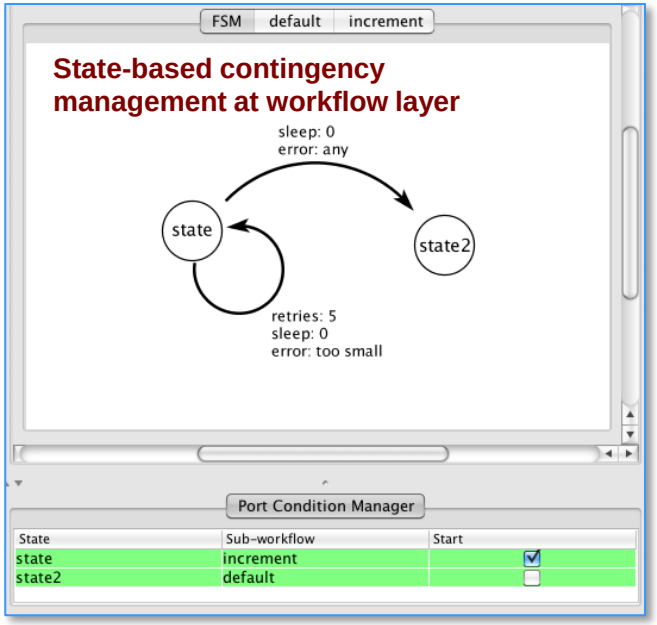
# Use collection-oriented modeling and design techniques



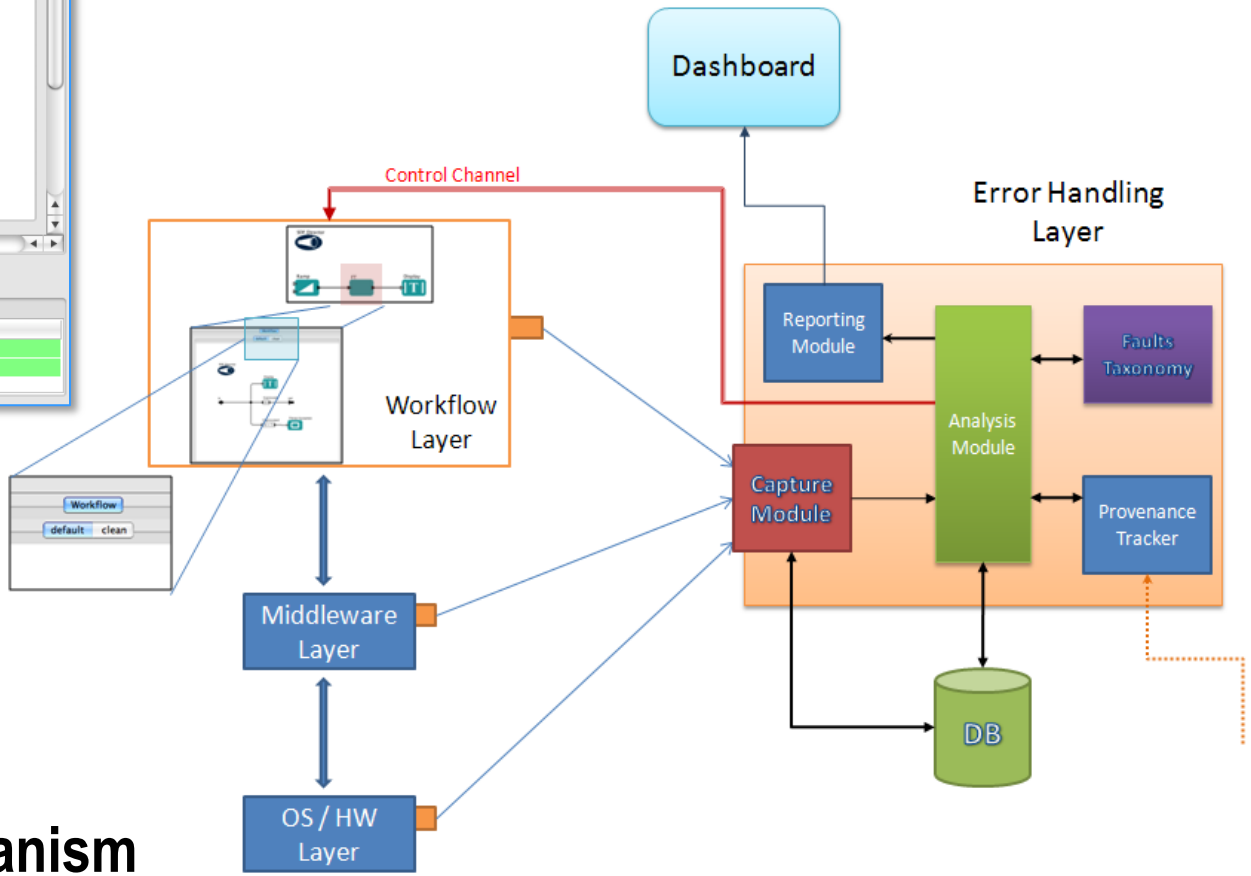

- Using COMAD techniques makes the workflow easier to understand and easier to change
- Workflow becomes mostly linear
- All data flows through all actors
- Implicit iteration over collections

**Contact: Bertram Ludaescher (ludaesch@ucdavis.edu)**

# Fault tolerance framework



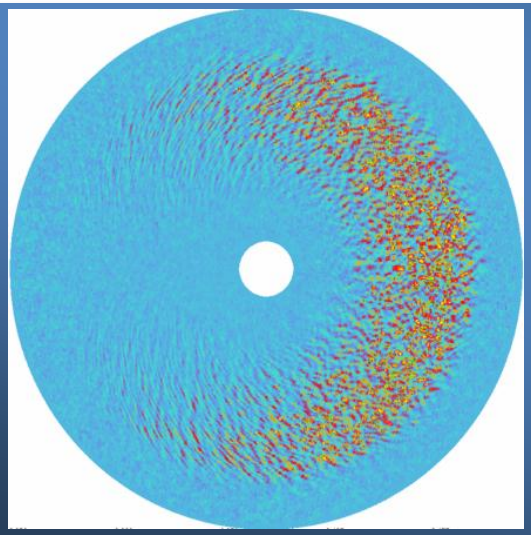

## Enables intelligent error handling within a workflow environment




## F/T framework components:

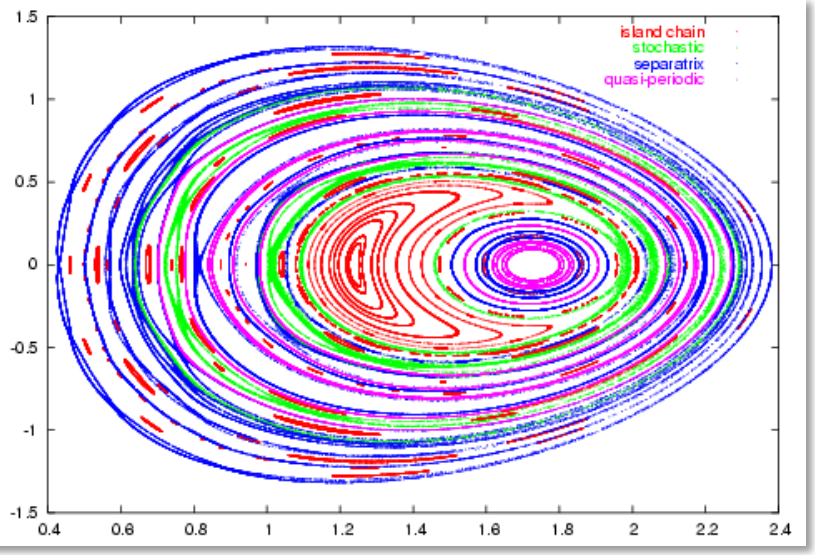
- Contingency actor
- Error handling layer
- Checkpointing mechanism

# Data analysis for fusion plasma



We are using data mining techniques to understand the interactions between coherent structures in the fluid and particle data from the GSEP SciDAC project. Challenges to the analysis include a poor understanding of the underlying physics, the complexity of the three-dimensional structures, and the massive size of the data

**The automatic classification of orbits in Poincaré maps is challenging as a robust set of features has to be extracted to represent the points in two-dimensional space. We have successfully applied data mining techniques to correctly classify the orbits with 96% accuracy**




**Contact: Chandrika Kamath, LLNL (kamath2@llnl.gov)**

# FastBit: Accelerating analysis of very large datasets

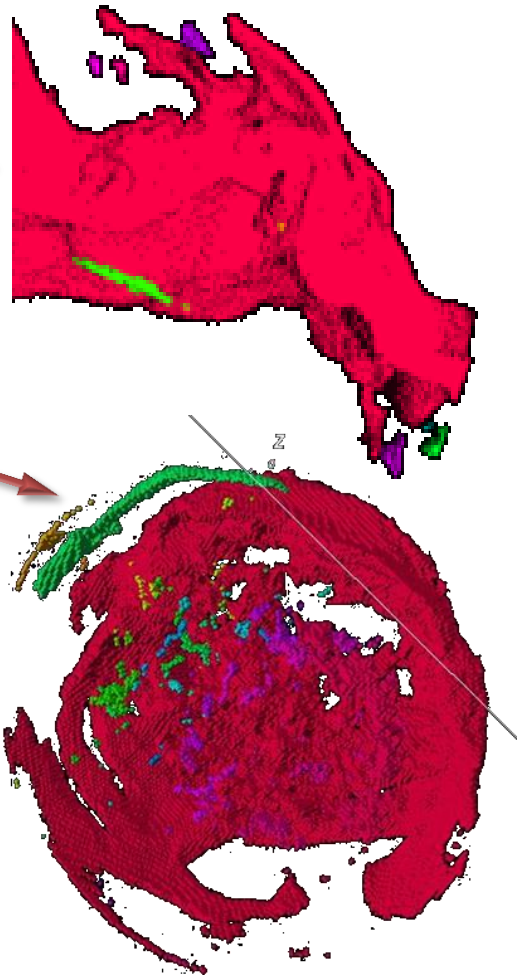
- Most data analysis algorithms cannot handle a whole dataset
  - Therefore, most data analysis tasks are performed on a subset of the data
  - Need: very fast indexing for real-time analysis
- FastBit is an extremely efficient compressed bitmap indexing technology
  - Can search a billion data values in seconds
  - FastBit improves the search speed by 10–100× that of best-known indexing methods
  - Uses a patented compression techniques
- 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)



**Contact: John Wu (kwu@lbl.gov)**

# Searching problems in data-intensive sciences

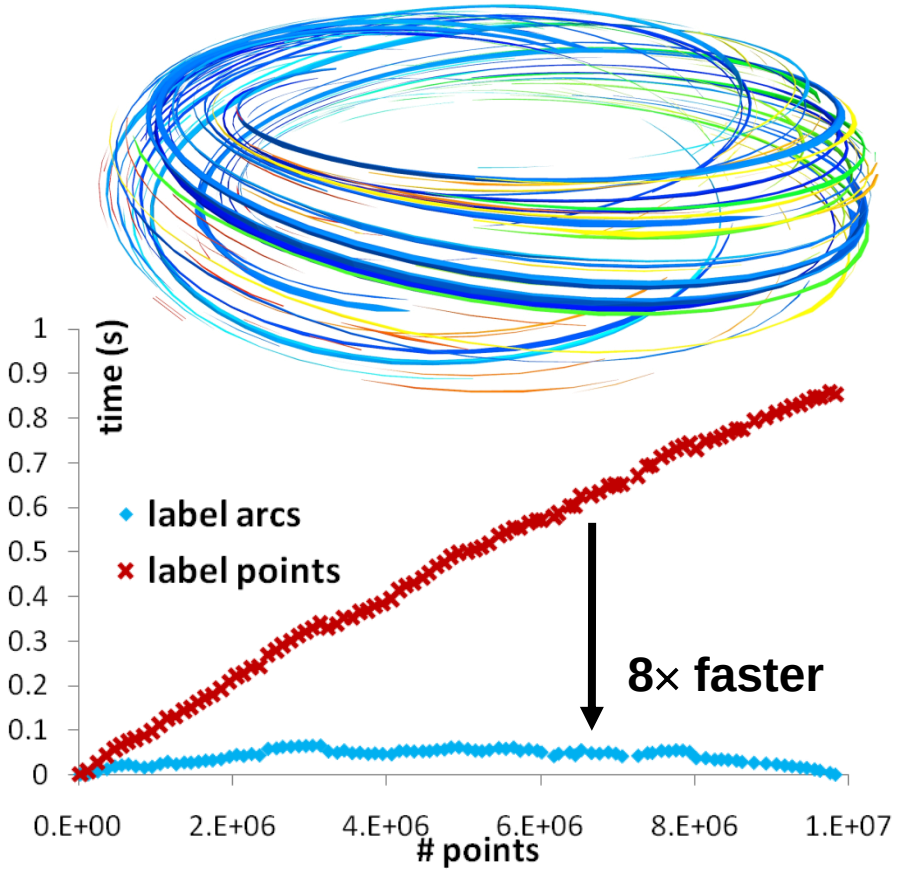
- Examples
  - Find the HEP collision events with the most distinct signature of Quark Gluon Plasma
  - Find the ignition kernels in a combustion simulation
  - Track a layer of exploding supernova
- These are not typical database searches
  - Large high-dimensional data sets (1000 time steps × 1000 × 1000 × 1000 cells × 100 variables)
  - No modification of individual records during queries, i.e., append-only data
  - Complex questions: 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



# Efficient searching algorithms for fusion data

**Using FastBit compressed data structures and an application-specific coordinate system to significantly reduce feature identification time**

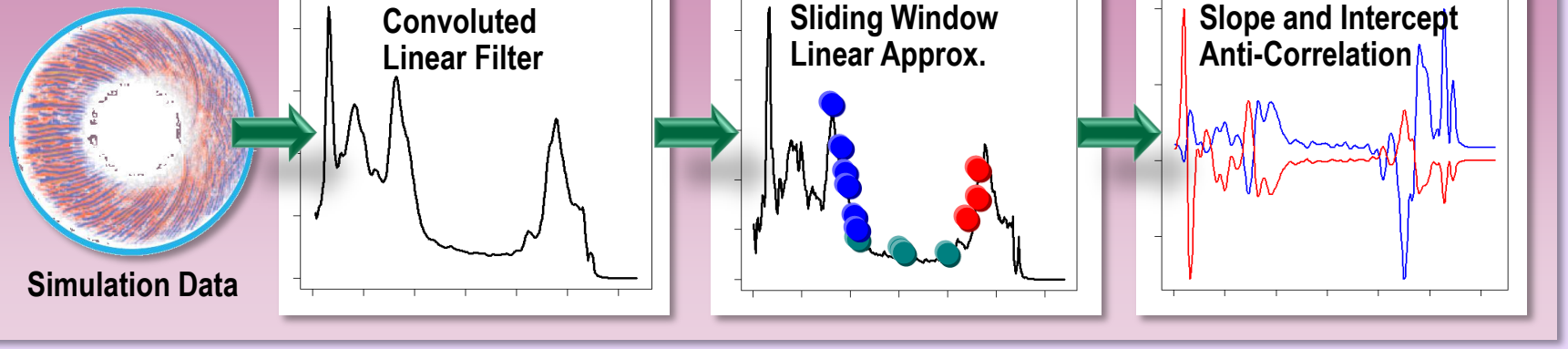
- To study the stability of magnetic confinement of fusion devices, one needs to identify features such as regions with high magnetic potential
- We break the task into two steps: (1) use FastBit index to find all points of interest; (2) assign a unique label to all connected points
- Working with groups of points (arcs) instead of individual points reduces execution time by 8× on average
- Using the magnetic coordinate system to connect the points further reduces the execution time



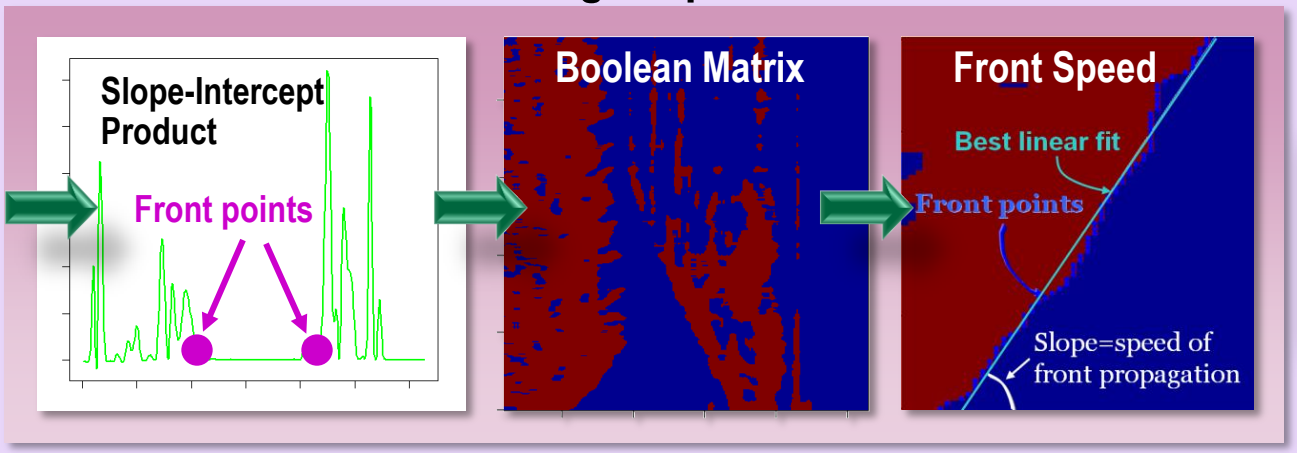

**Contact: John Wu, LBNL (kwu@lbl.gov)**

# Turbulent front detection and tracking
## A multi-step knowledge discovery process

### Data Preprocessing Steps




### Front Detection and Tracking Steps



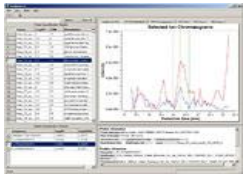

Efficient run is performed by our pRapply parallel system

Contact: Nagiza Samatova (samatovan@ornl.gov)

# Parallel statistics: Open-source software

**pR:** Parallel R for high performance statistical computing
Impact: Pioneered parallel statistical computing with R.
Provides back-end analytical support in Dashboard

**RScaLAPACK**: R library for parallel linear algebra ScaLAPACK
Impact: Distributed through >30 mirror sites across ~20 countries.
Part of different Linux distributions through RPMs



**ProRata:** Quantitative shotgun proteomics GUI software
Impact: Enabled confidence estimation at both peptide and protein abundance levels for the first time in field. >1000 downloads; featured by the *J. of Proteome Research* (2006) and *SciDAC Reviews* (2007); used by DOE OBER Bioenergy and GTL Centers

**mpiBlast-pio:** Efficient parallelization of BLAST sequence search
Impact: Solved I/O bottleneck to scalability to 1000s processors;
Integrated into a popular mpiBLAST software, >1000s downloads

**Contact: Nagiza Samatova (samatovan@ornl.gov)**

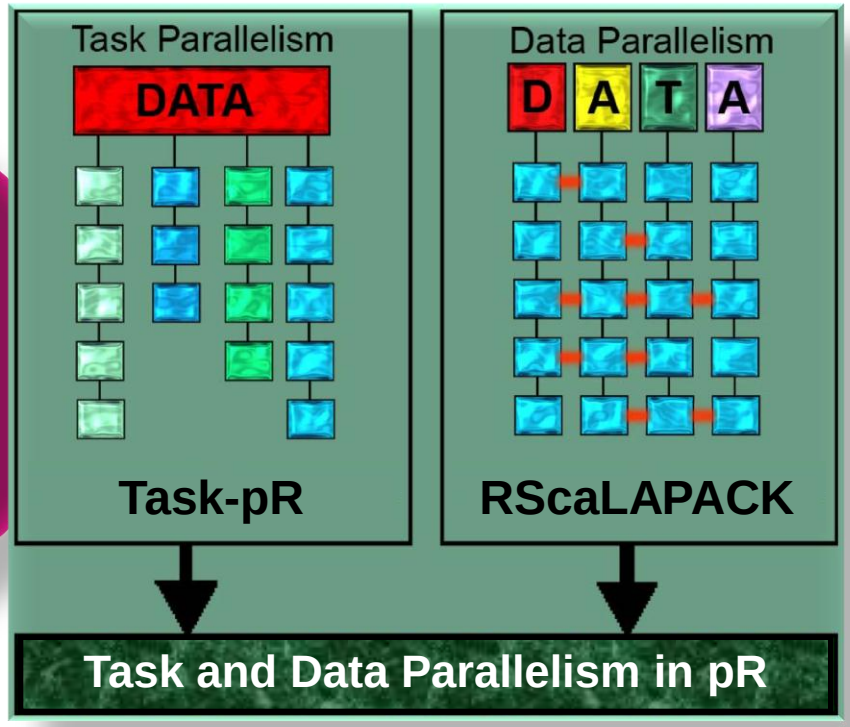
# Task and data parallelism in pR

**Goal: Parallel R (pR) aims to**

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

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

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




**Embarrassingly parallel**

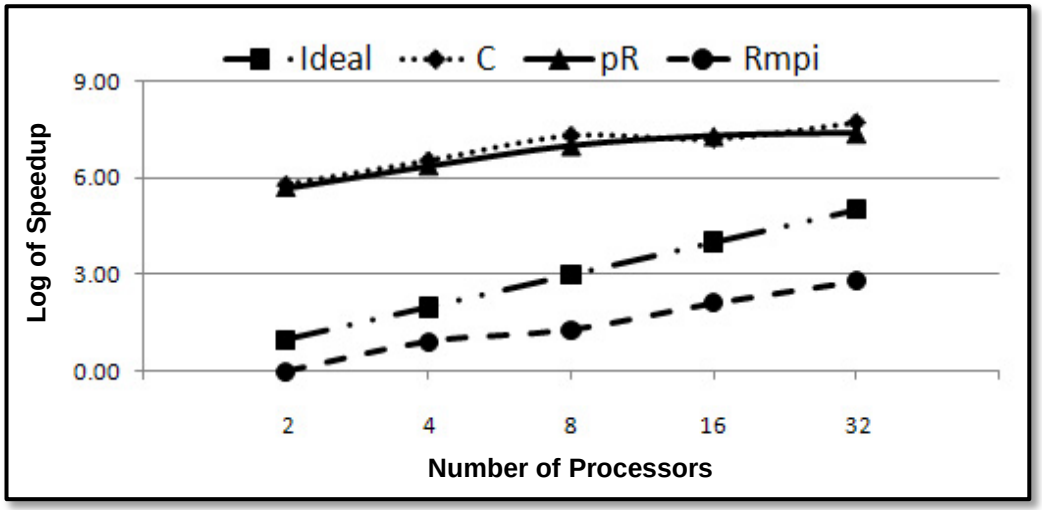
- **Likelihood Maximization**
- **Resampling schemes: Bootstrap, Jackknife**
- **Markov Chain Monte Carlo (MCMC)**
- **Animations**

**Data-parallel**

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

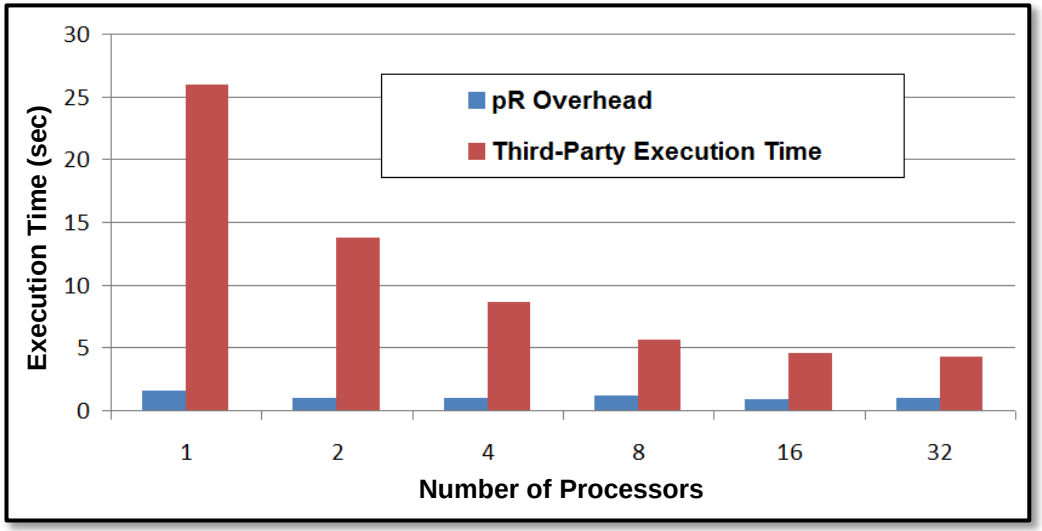
**Contact: Nagiza Samatova (samatovan@ornl.gov)**

# pR performance: Scalability and overhead




## Scalability with number of processors

- Scales the same way as the base MPI C/C++/Fortran code
- Provides super-linear speedup
- Outperforms Rmpi by >30×




## Overhead due to pR

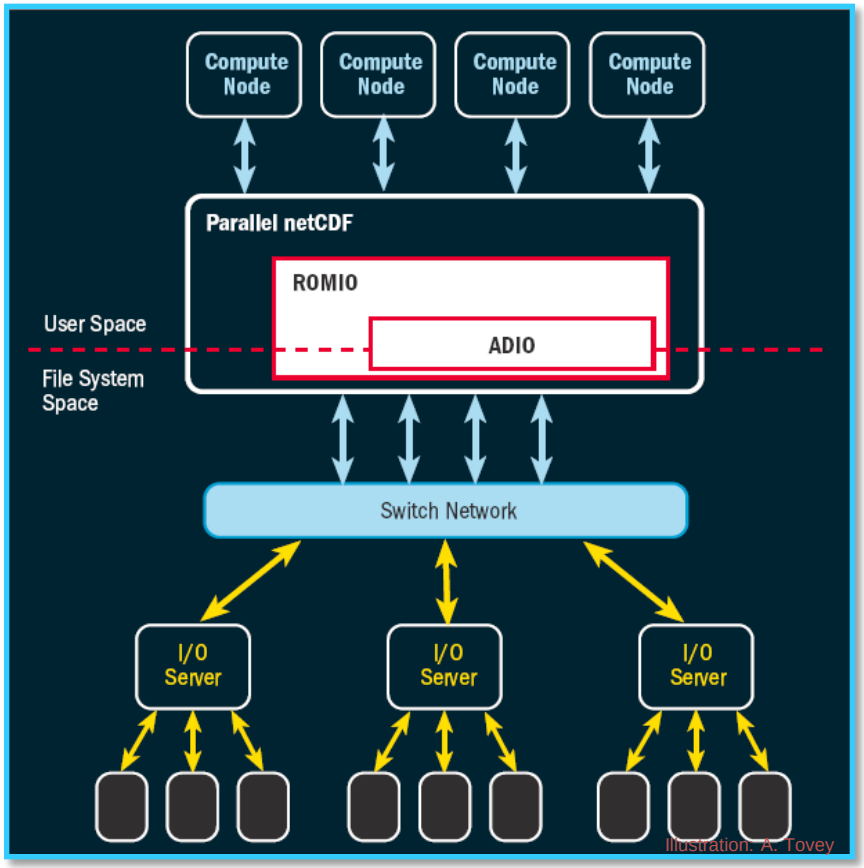
- Introduces negligible overhead
- Improves R performance on a single processor by ~28×

**Contact: Nagiza Samatova (samatovan@ornl.gov)**

# Parallel input/output
## Scaling computational science

**Orchestration of data transfers and speedy analyses depend on efficient systems for storage, access, and movement of data among modules**

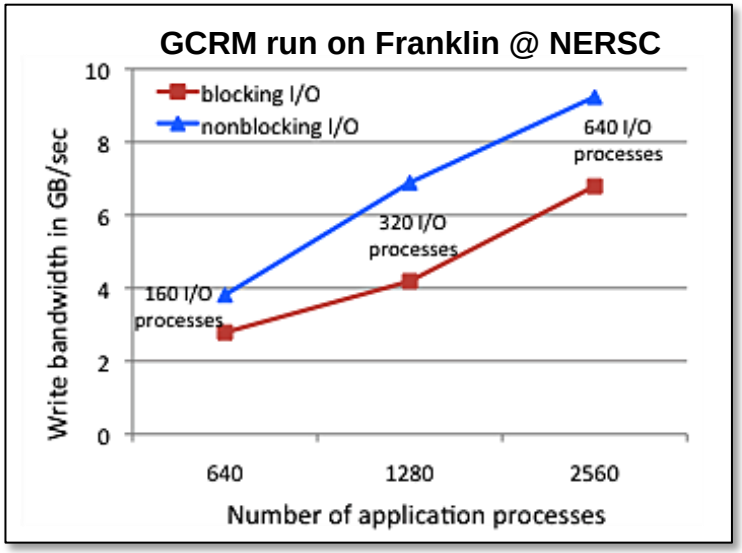
- Multilayer parallel I/O design
  - Supports Parallel-netCDF library built on top of MPI-IO implementation called ROMIO, built in turn on top of Abstract Device Interface for I/O system, used to access parallel storage system
- Benefits to scientists
  - Brings performance, productivity, and portability
  - Improves performance by order of magnitude
  - Operates on any parallel file system (e.g., GPFS, PVFS, PanFS, Lustre)

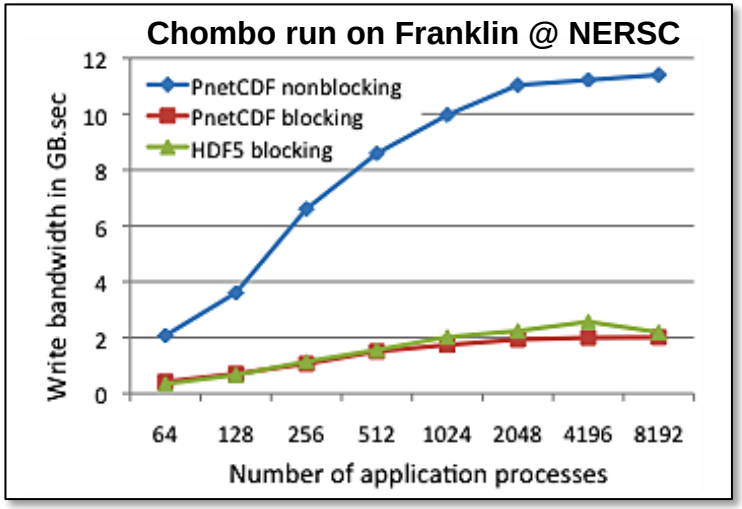



**Contact: Rob Ross, ANL (rross@mcs.anl.gov)**

# Enable I/O request aggregation in PnetCDF nonblocking APIs

- New nonblocking I/O enables request data aggregation
  - Noncontiguous small-sized requests can be aggregated into large, contiguous ones
  - Parallel file systems serve better for large, contiguous requests
  - Combined with MPI collective I/O provides much better bandwidths
- Improve I/O performance
  - GCRM – Global Cloud Resolving Model
  - Chombo – software supporting AMR data structure



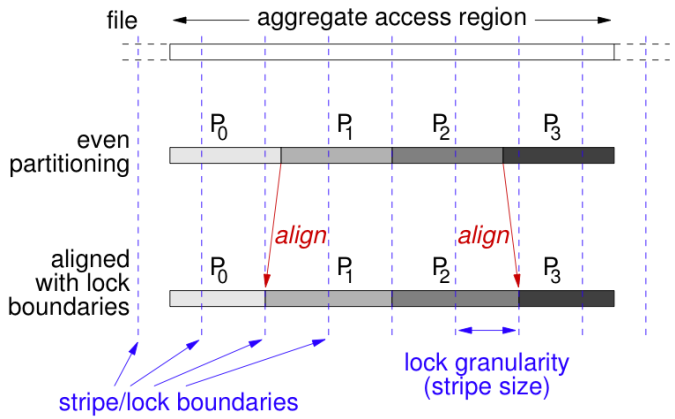



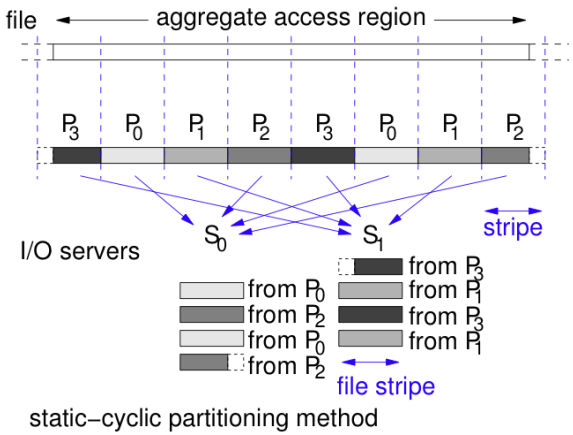

**Example code**

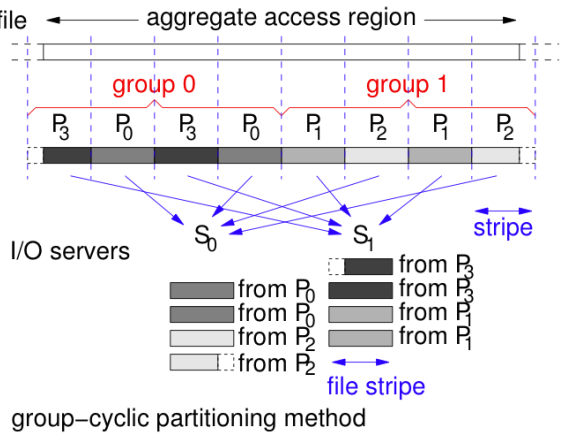
```
nfmpi_iput_vara_real(ncid, var1, . . .)
nfmpi_iput_vara_real(ncid, var2, . . .)
. . .
nfmpi_iput_vara_real(ncid, varN, . . .)

nfmpi_wait_all(ncid, N, reqs, status)
```

**Contact: Wei-keng Liao and Alok Choudhary,**
**NWU, {wkliao, choudhar}@ece.northwestern.edu**

# Enhance MPI collective I/O performance
## Adapting lock-aware file domain partitioning methods



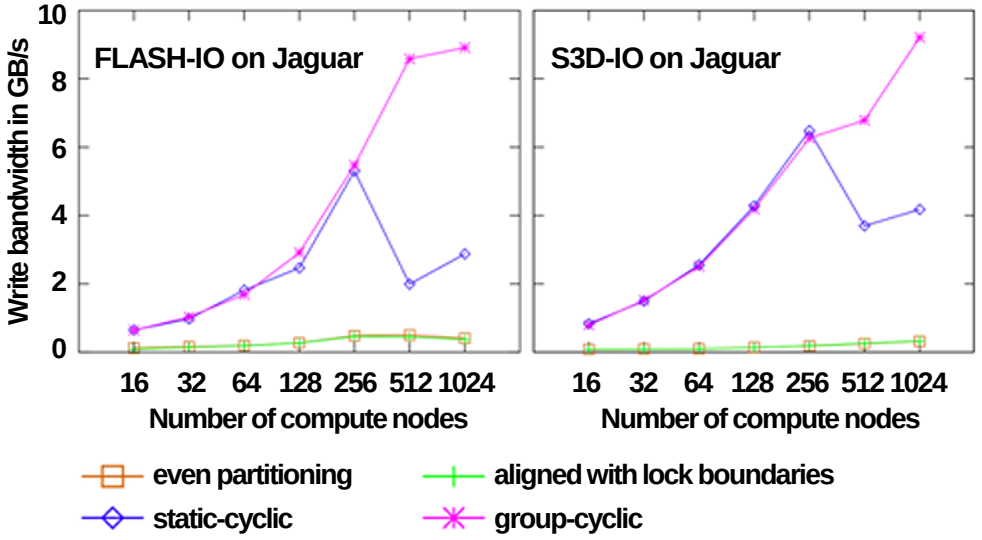




In MPI collective I/O, a file is partitioned into contiguous, nonoverlapping file domains, one for each I/O process

To minimize the lock acquisition cost, MPI-IO must adapt the partitioning method best fit to the underlying file system's locking protocol

- Lock-boundary aligned method
- Static-cyclic method
- Group-cyclic method (best for Lustre)




Contact: Wei-keng Liao and Alok Choudhary, NWU, {wkliao, choudhar}@ece.northwestern.edu

# Contacts


## Arie Shoshani
Principal Investigator
Lawrence Berkeley National Laboratory
shoshani@lbl.gov

## Nagiza Samatova
Data Mining and Analysis Area Leader
Oak Ridge National Laboratory
samatovan@ornl.gov

## Terence Critchlow
Scientific Process Automation Area Leader
Pacific Northwest National Laboratory
terence.critchlow@pnl.gov

## Rob Ross
Storage Efficient Access Area Leader
Argonne National Laboratory
rross@mcs.anl