
The SciDAC Scientific Data Management Center: Infrastructure and Results

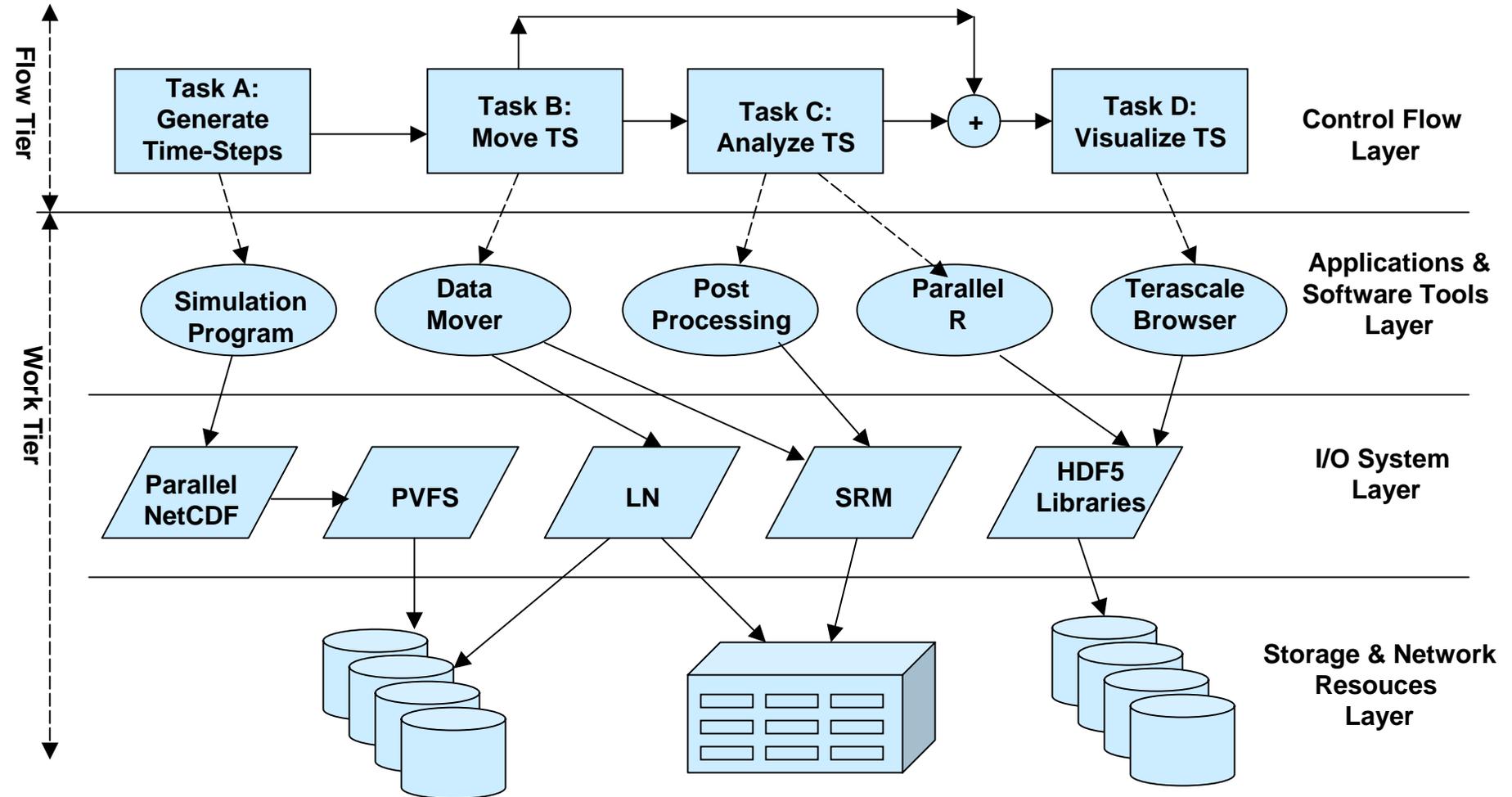
Arie Shoshani

Lawrence Berkeley National Laboratory

SC 2004

November, 2004

A Typical SDM Scenario



Dealing with Large Data volumes: Three Stages of Data Handling

- **Generate data**
 - Run simulations, dump data
 - Capture metadata
- **Post-processing of data**
 - Process all data, produce reduced datasets, summaries
 - Generate indexes
- **Analyze data**
 - Interested in subset of the data, search over large amounts of data, produce relatively little data (for analysis, vis, ...)

Stage \ Data	Input	Output	Computation
Generation	Low	High	High
Post-processing	High	Med-High	High
Analysis	Med-High	Low	Low-Med

I/O is the predicted bottleneck

Balance vs. Bottlenecks				
Processors	Kiloflops	Megaflops	Gigaflops	Teraflops
I/O	Megabytes	Gigabytes	Terabytes	Petabytes
Network/sec	Kilobits	Megabits	Gigabits	Terabits
Memory Size	Kilobytes	Megabytes	Gigabytes	Terabytes
	1970's	1980's	1990's	2000's

Main reason: *data transfer rates to disk and tape devices have not kept pace with computational capacity*

Source: Celeste Matarazzo

Data Generation – technology areas

- **Parallel I/O writes – to disk farms**
 - Does technology scale?
 - Reliability in face of disk failures
 - Dealing with various file formats (NetCDF, HDF, AMR, unstructured meshes, ...)
- **Parallel Archiving – to tertiary storage**
 - Is tape striping cost effective?
 - Reorganizing data before archiving to match predicted access patterns
- **Dynamic monitoring of simulation progress**
 - Tools to automate workflow
- **Compression – is overhead prohibitive?**

Post Processing – technology areas

- **Parallel I/O reads – from disk farms**
 - Data clustering to minimize arm movement
 - Parallel read synchronization
 - Does it scale linearly?
- **Reading from archives – tertiary storage**
 - Minimize tape mounts
 - Does tape striping help?
 - What's after tape? Large disk farm archives?
- **Feed large volumes data into machine**
 - Competes with write I/O
 - Need fat I/O channels, parallel I/O

Analysis – technology areas

- **Dynamic hot clustering of data from archive**
 - Based on repeated use (caching & replication)
 - Takes advantage of data sharing
- **Indexing over data values**
 - **Indexes need to scale:**
 - Linear search over billion of data objects
 - Search over combinations of multiple data measures per mesh point
 - Take advantage of append-only data
 - **Parallel indexing methods**
- **Analysis result streaming**
 - On the fly monitoring and visualization
 - Suspend-Resume capabilities, clean abort

SDM Technology: Layering of Components

**Scientific Process Automation
(Workflow) Layer**



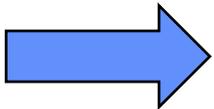
**Data Mining & Analysis
Layer**



**Storage Efficient Access
Layer**



Hardware, OS, and MSS (HPSS)



Data Generation

Scientific Process
Automation Layer

Workflow Design and Execution

Data Mining and
Analysis Layer

Simulation
Run

Storage Efficient
Access Layer

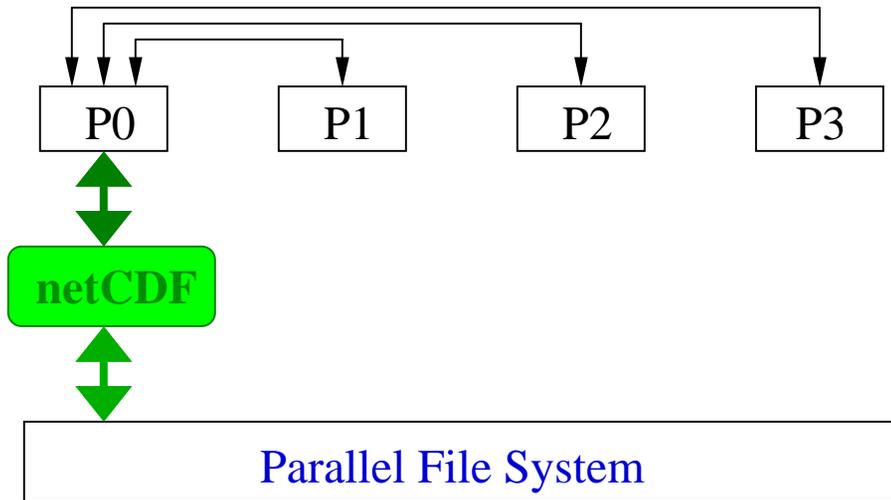
Parallel
netCDF

MPI-IO

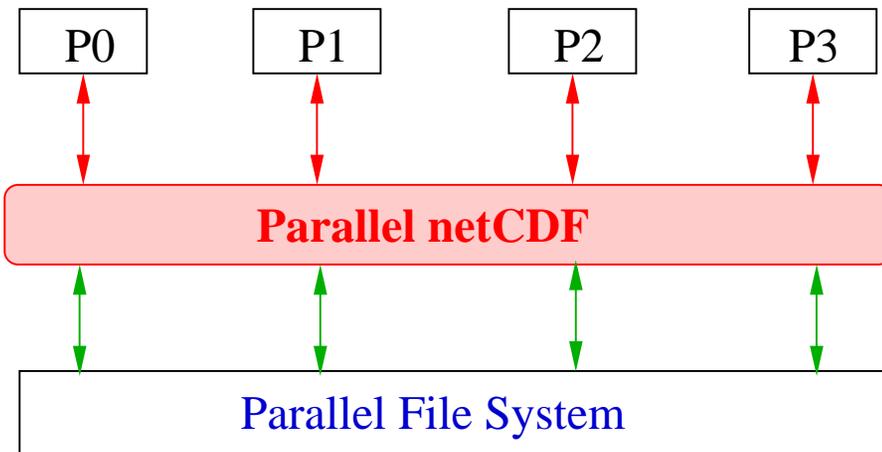
PVFS2

OS, Hardware (Disks, Mass Store)

Parallel NetCDF v.s. NetCDF (ANL+NWU)



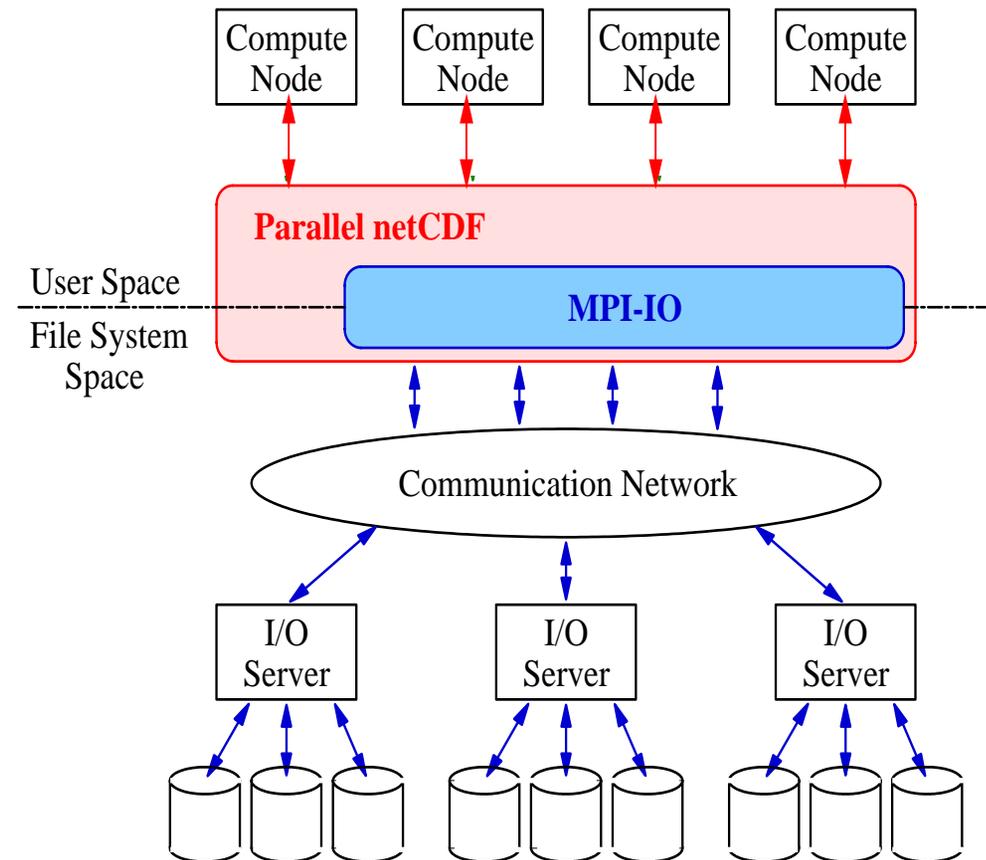
- **Slow and cumbersome**
- **Data shipping**
- **I/O bottleneck**
- **Memory requirement**



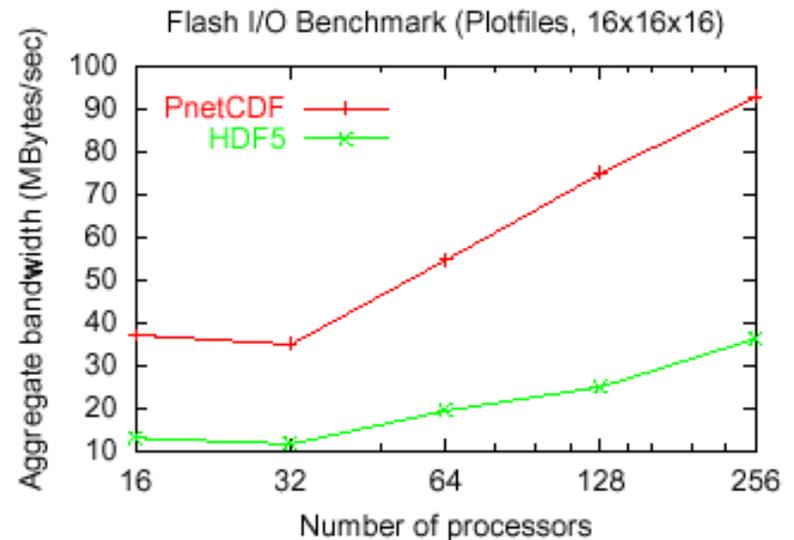
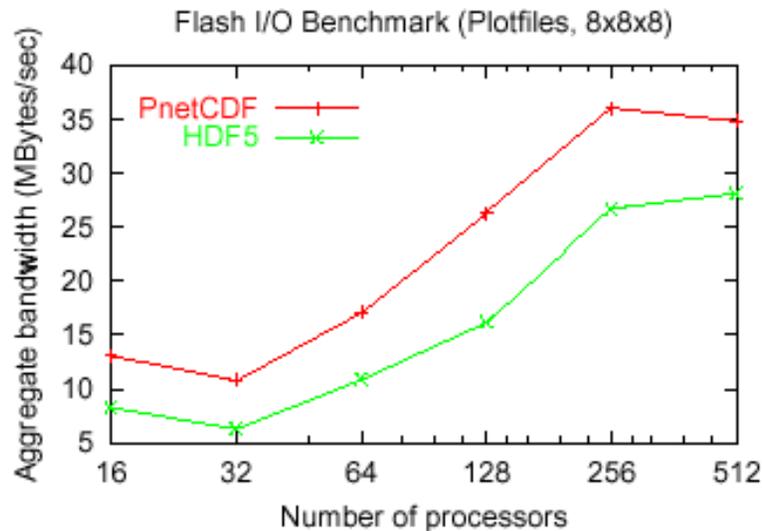
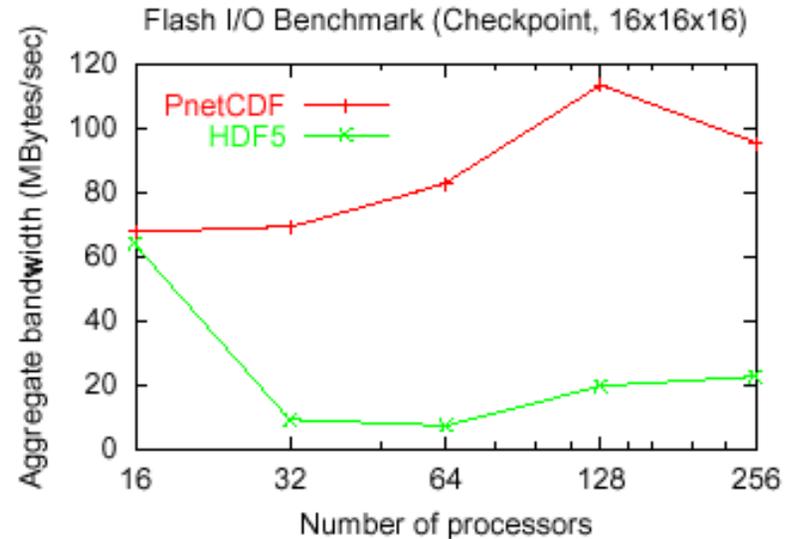
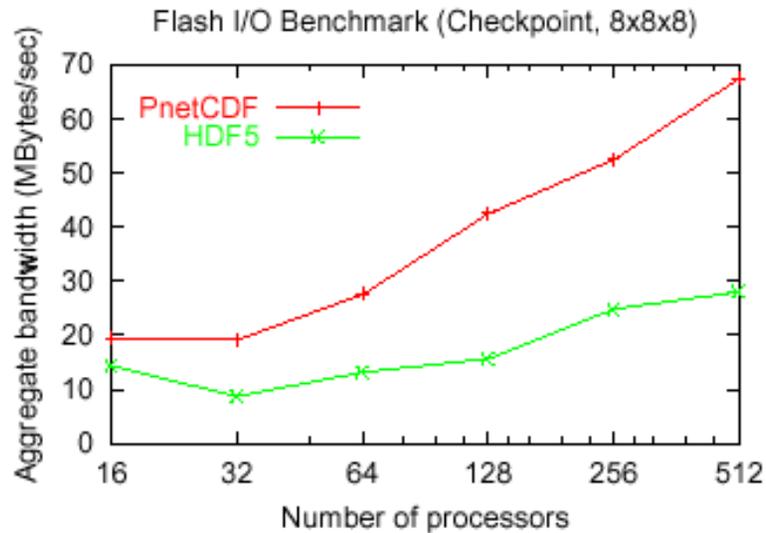
- Programming Convenience
- Perform I/O cooperatively or collectively
- Potential parallel I/O optimizations for better performance

Parallel NetCDF Library Overview

- User level library
- Accept parallel requests in netCDF I/O patterns
- Parallel I/O through MPI-IO to underlying file system and storage
- Good level of abstraction for portability and optimization opportunities

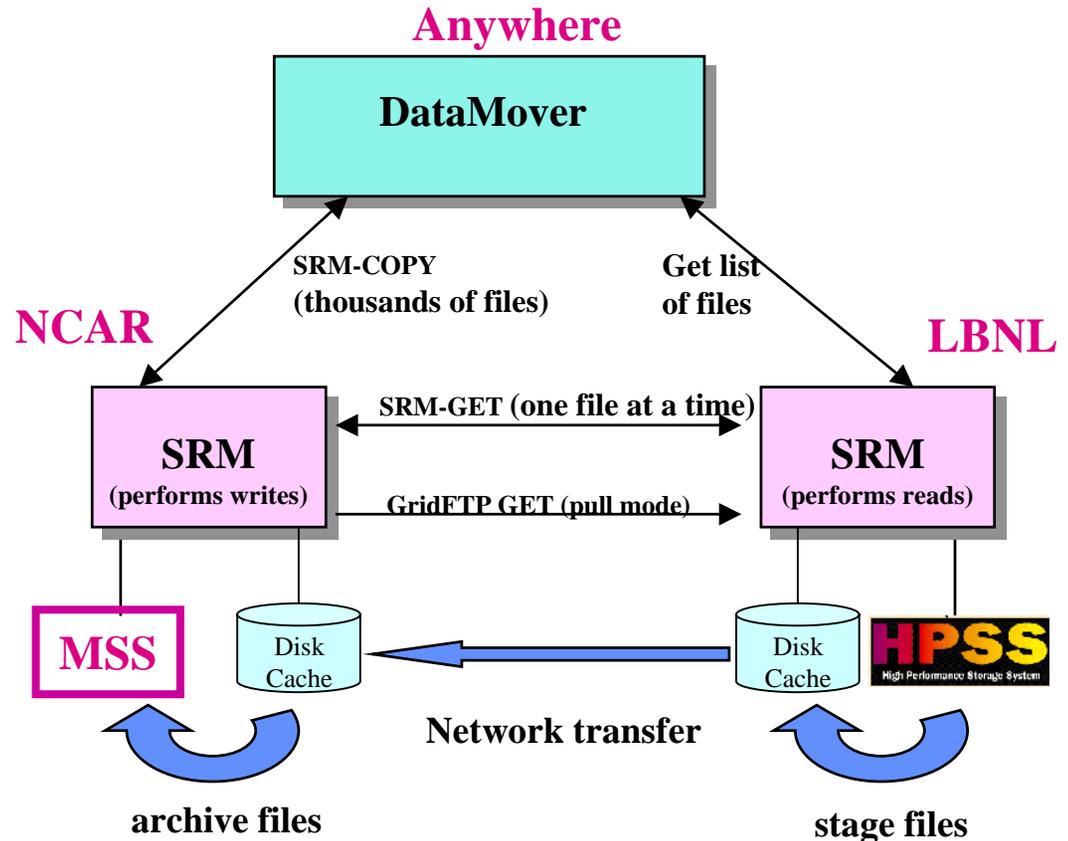


Parallel netCDF Performance

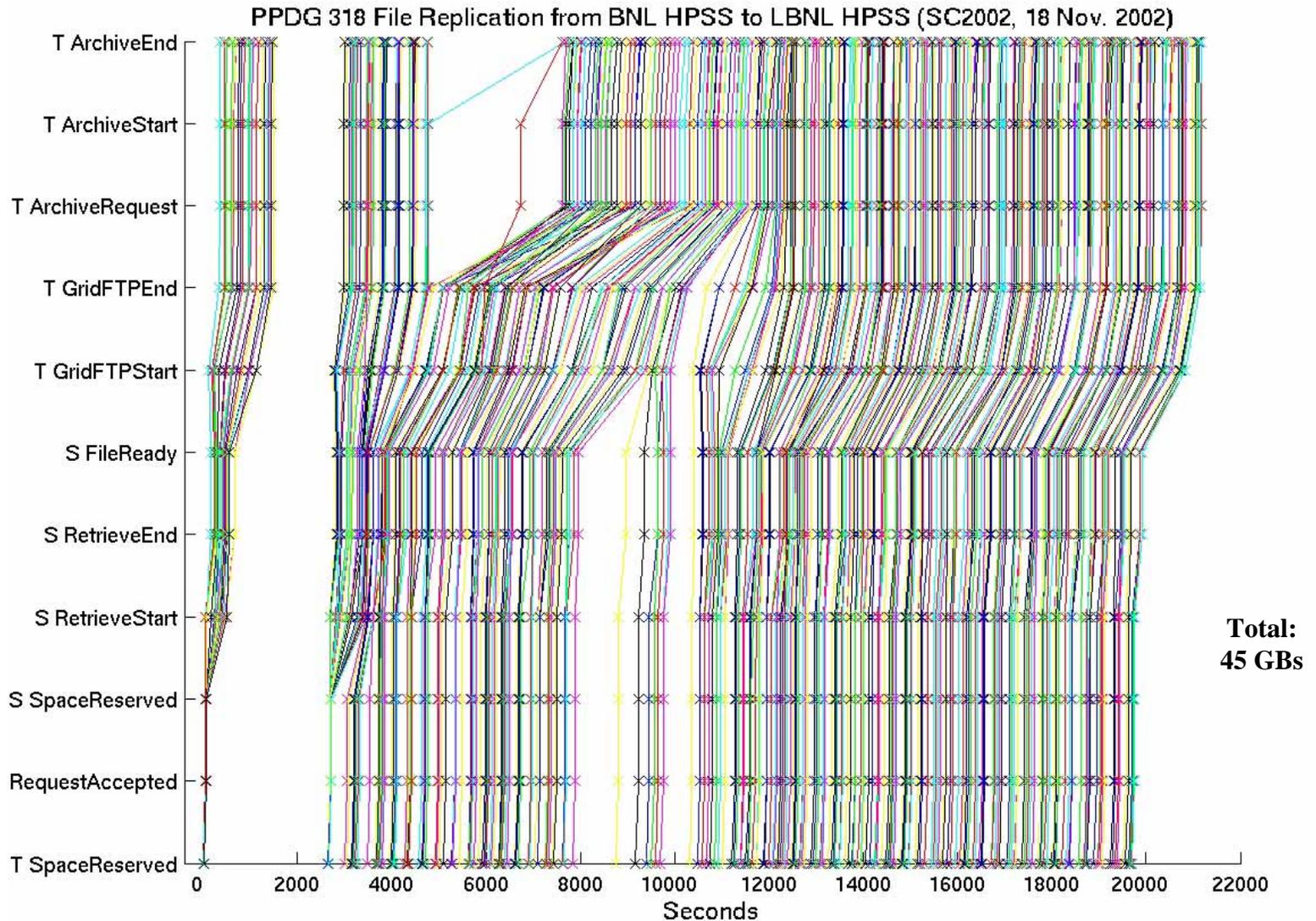


Robust Multi-file Replication

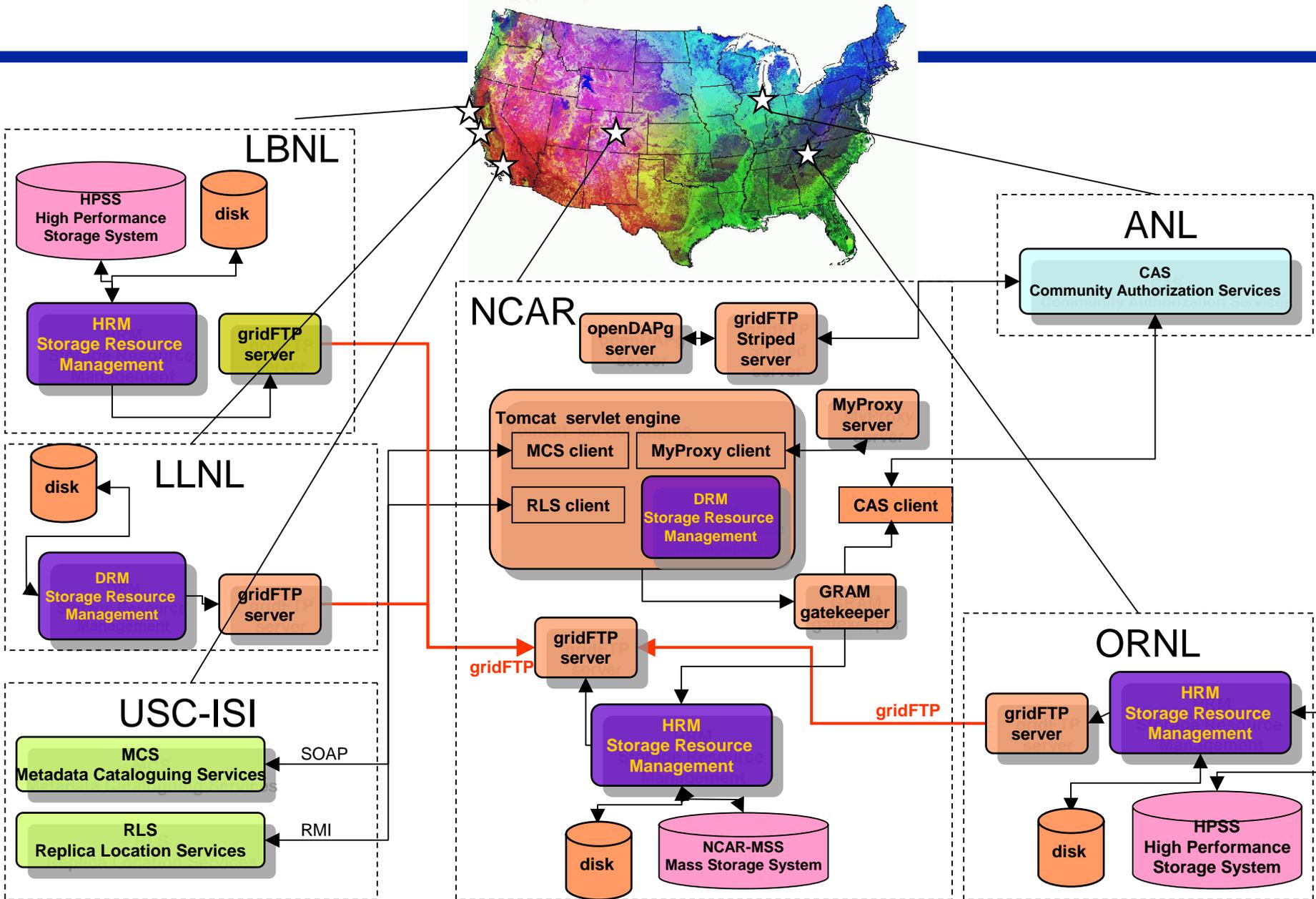
- **Problem:** move thousands of files robustly
 - Takes many hours
 - Need error recovery
 - Mass storage systems failures
 - Network failures
- **Solution:** Use Storage Resource Managers (SRMs)
- **Problem:** too slow
- **Solution:**
 - Use parallel streams
 - Use concurrent transfers
 - Use large FTP windows
 - Pre-stage files from MSS



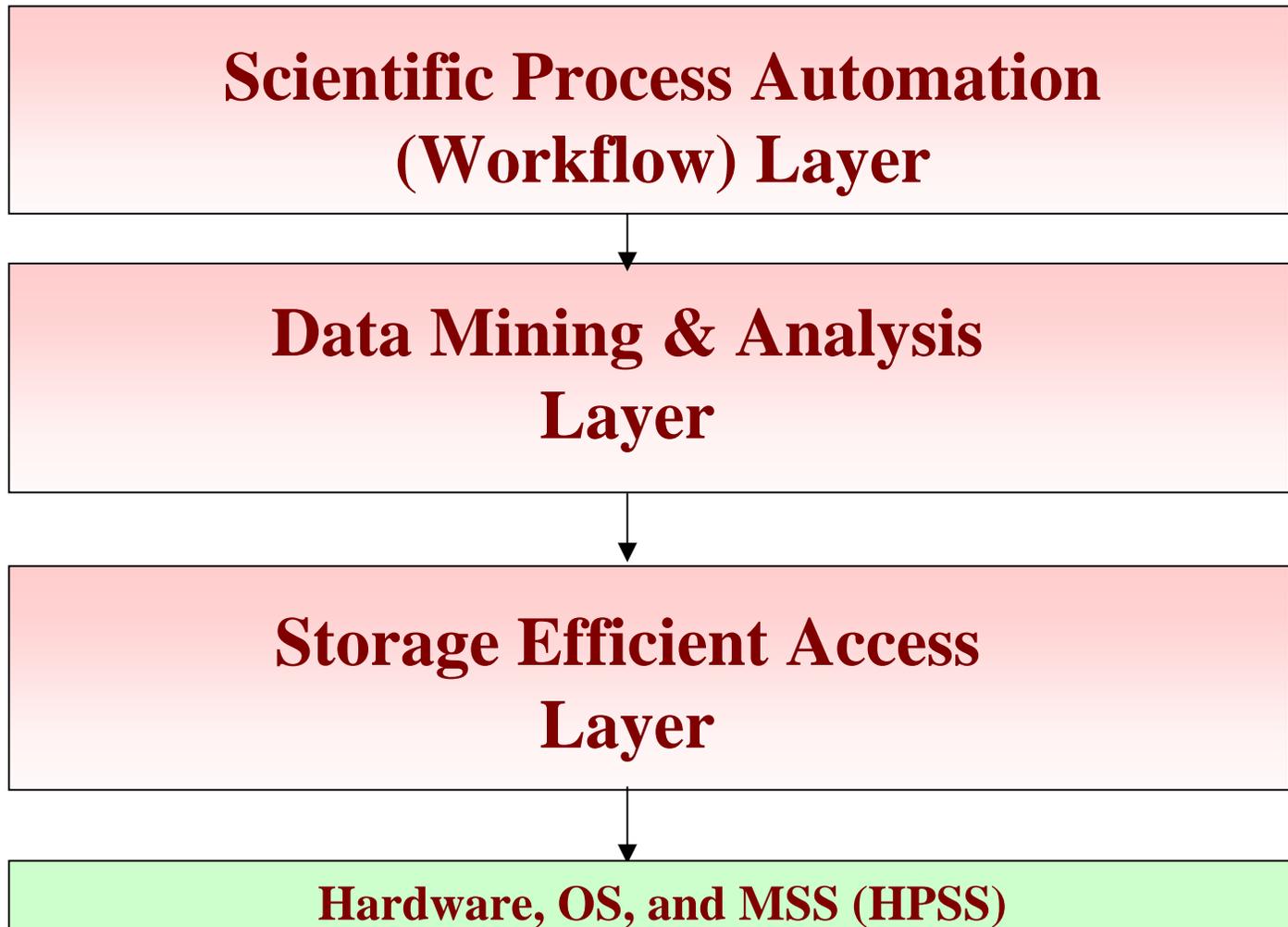
File tracking shows recovery from transient failures



Earth Science Grid



Layering of Components

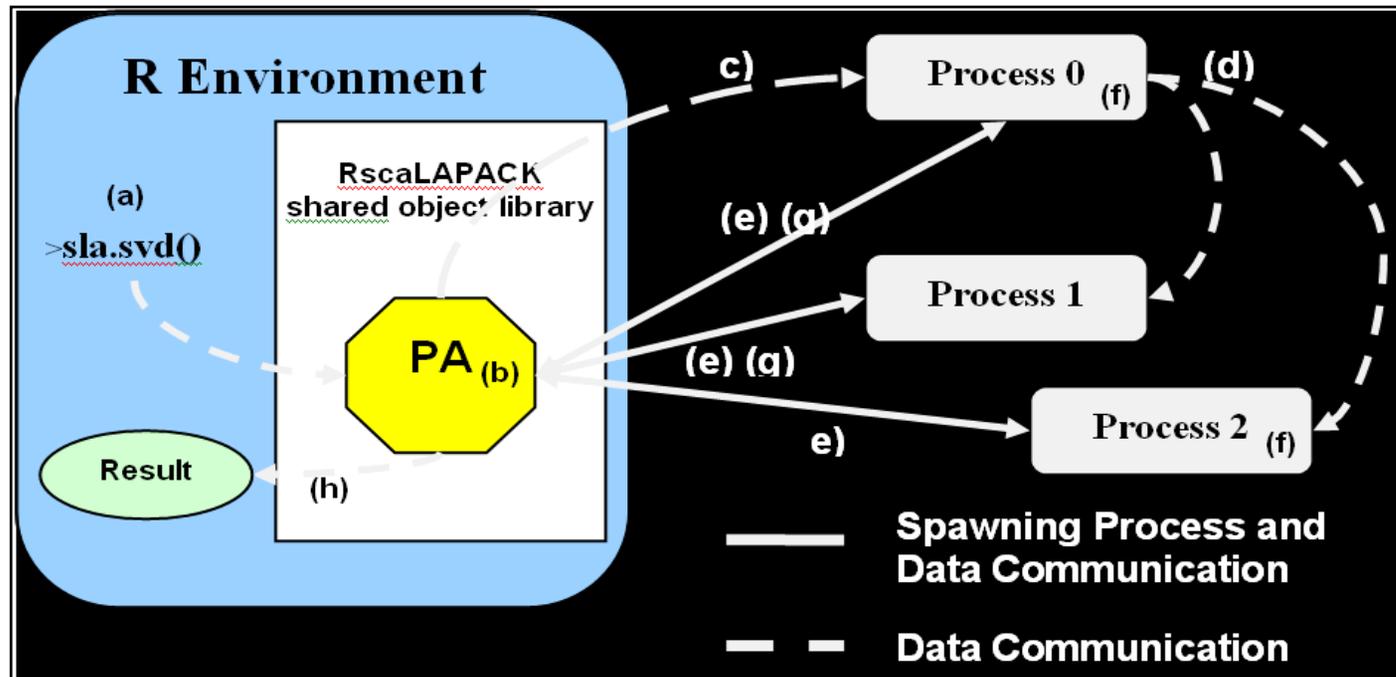


RScalAPACK (ORNL)

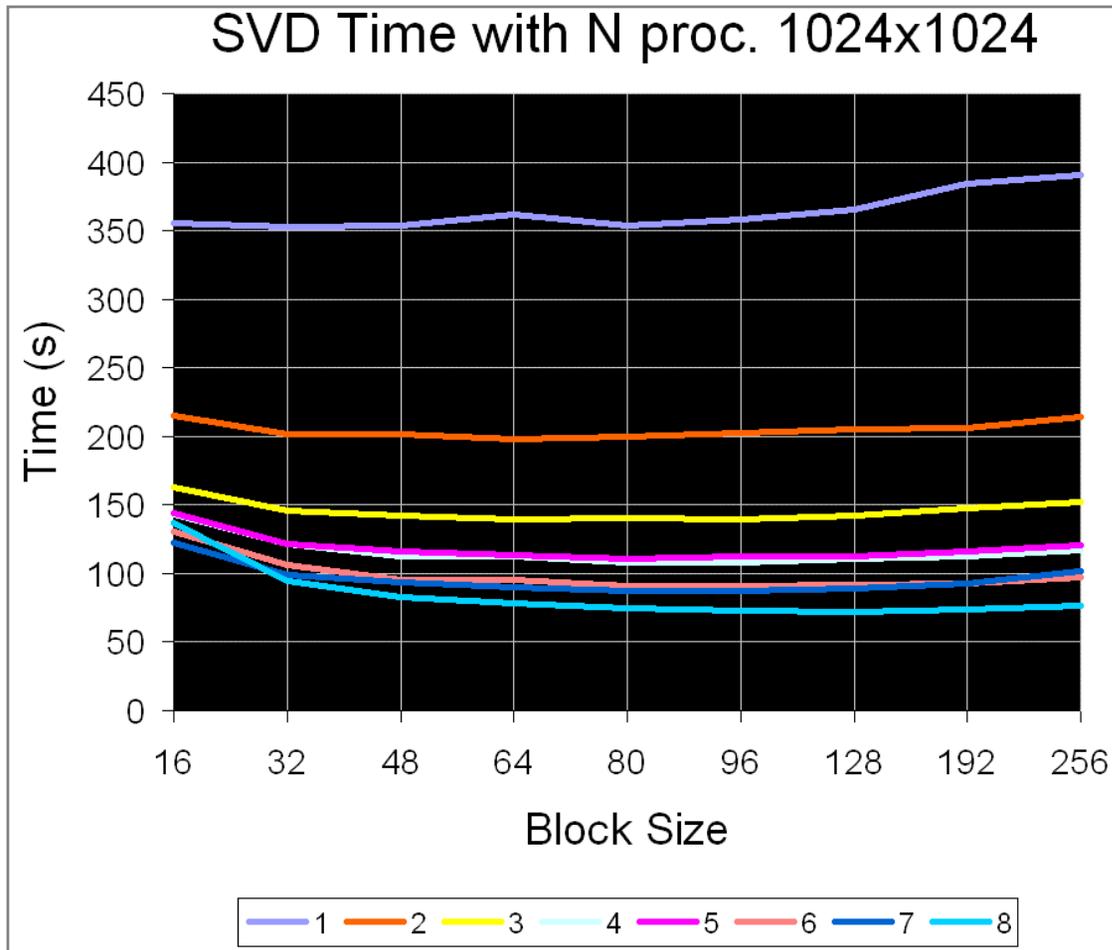
- **Through RScalAPACK we provide a simple intuitive R interfaces to ScaLAPACK routines.**
- **The package does not need the user to worry about setting up the parallel environment and distribution of data prior to ScaLAPACK function call.**
- **RScalAPACK is developed as an add-on package to R.**
- **Significant speed gain is observed in some function execution.**
- **Submitted to CRAN (a network of 27 www sites across 15 countries holding R distribution) in March 2003**

RScaLAPACK Architecture

- (1) Parallel Agent (PA):
Orchestrates entire parallel execution as per user's request.
- (2) Spawned Processes:
Actual parallel execution of the requested function.



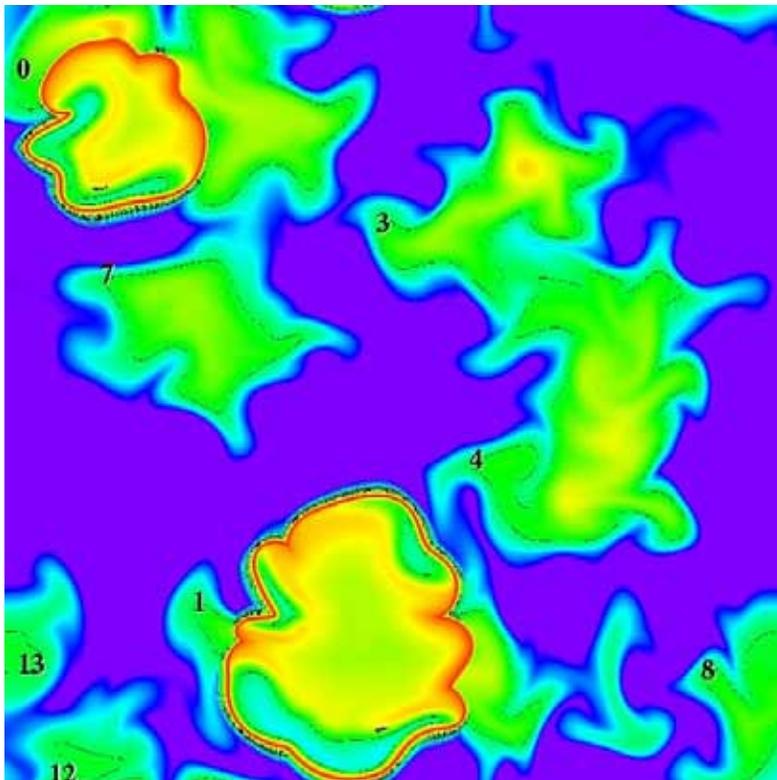
RScaLAPACK Benchmarks



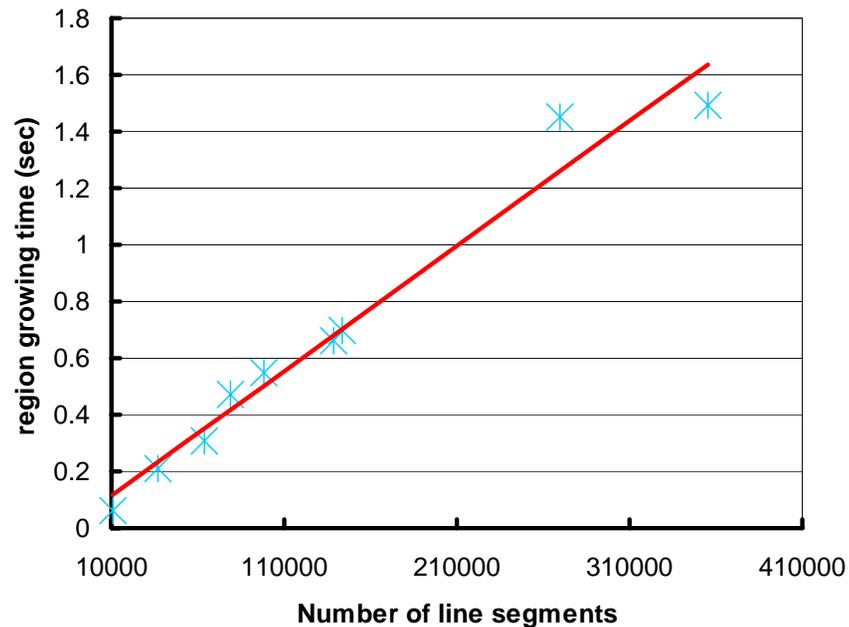
Fastbit – bitmap Indexing Technology (LBNL)

- **Search over large spatio-temporal data**
 - **Combustion simulation: 1000x1000x1000 mesh with 100s of chemical species over 1000s of time steps**
 - **Supernova simulation: 1000x1000x1000 mesh with 10s of variables per cell over 1000s of time steps**
- **Common searches are partial range queries**
 - **Temperature > 1000 AND pressure > 10^6**
 - **$\text{HO}_2 > 10^{-7}$ AND $\text{HO}_2 > 10^{-6}$**
- **Features**
 - **Search time proportional to number of hits**
 - **Index generation linear with data values (require read-once only)**

FastBit-Based Multi-Attribute Region Finding is Theoretically Optimal



Flame Front discovery
(range conditions for multiple measures)
in a combustion simulation (Sandia)

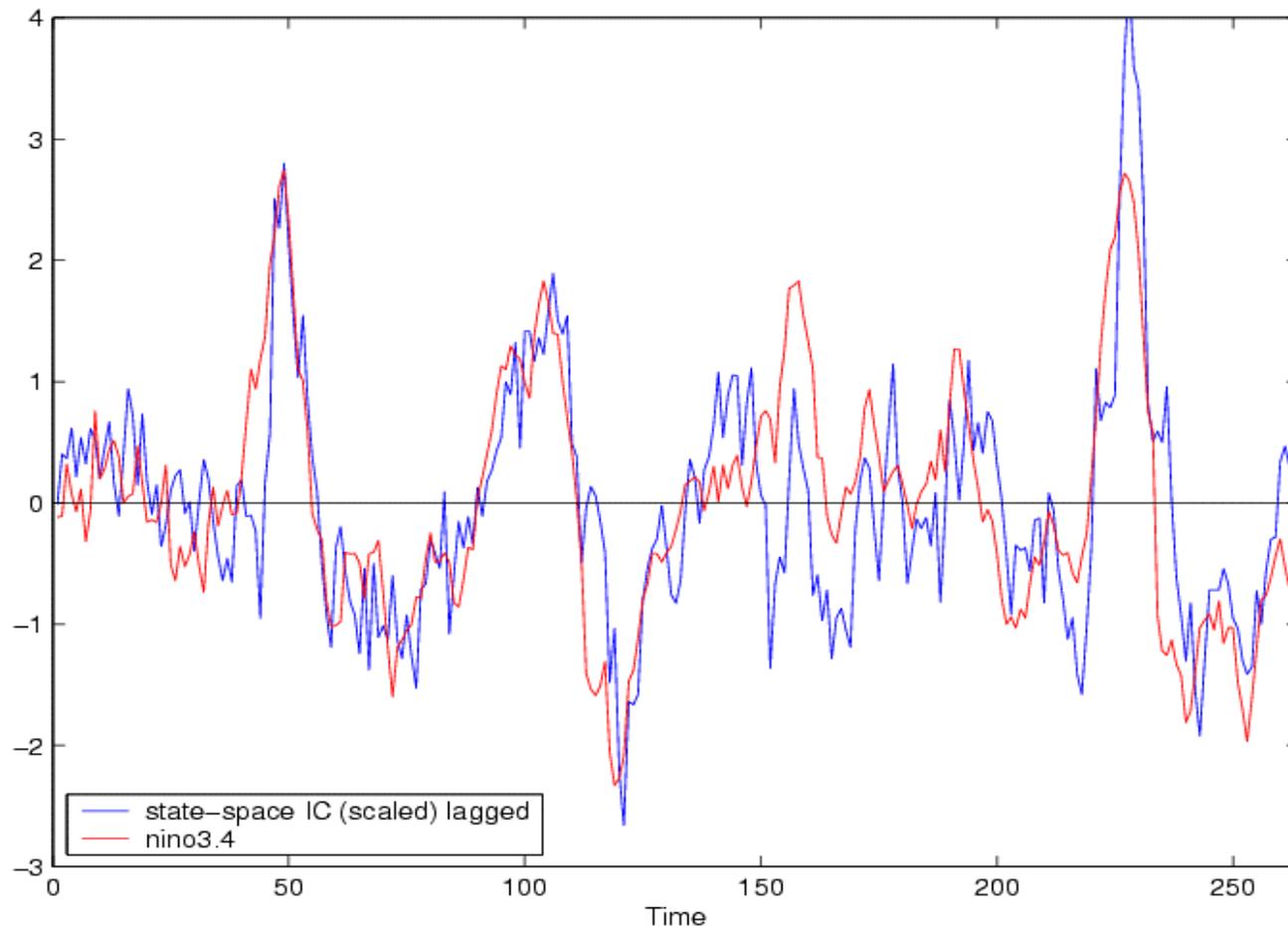


Time required to identify regions in
3D Supernova simulation (LBNL)

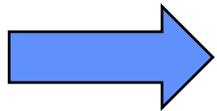
**On 3D data with over 110 million points,
region finding takes less than 2 seconds**

Feature Selection and Extraction: Using Generic Analysis Tools (LLNL)

Comparing Climate simulation to experiment data
Used PCA and ICA technology for accurate Climate signal separation



Layering of Components



**Scientific Process Automation
(Workflow) Layer**



**Data Mining & Analysis
Layer**



**Storage Efficient Access
Layer**



Hardware, OS, and MSS (HPSS)

TSI Workflow Example

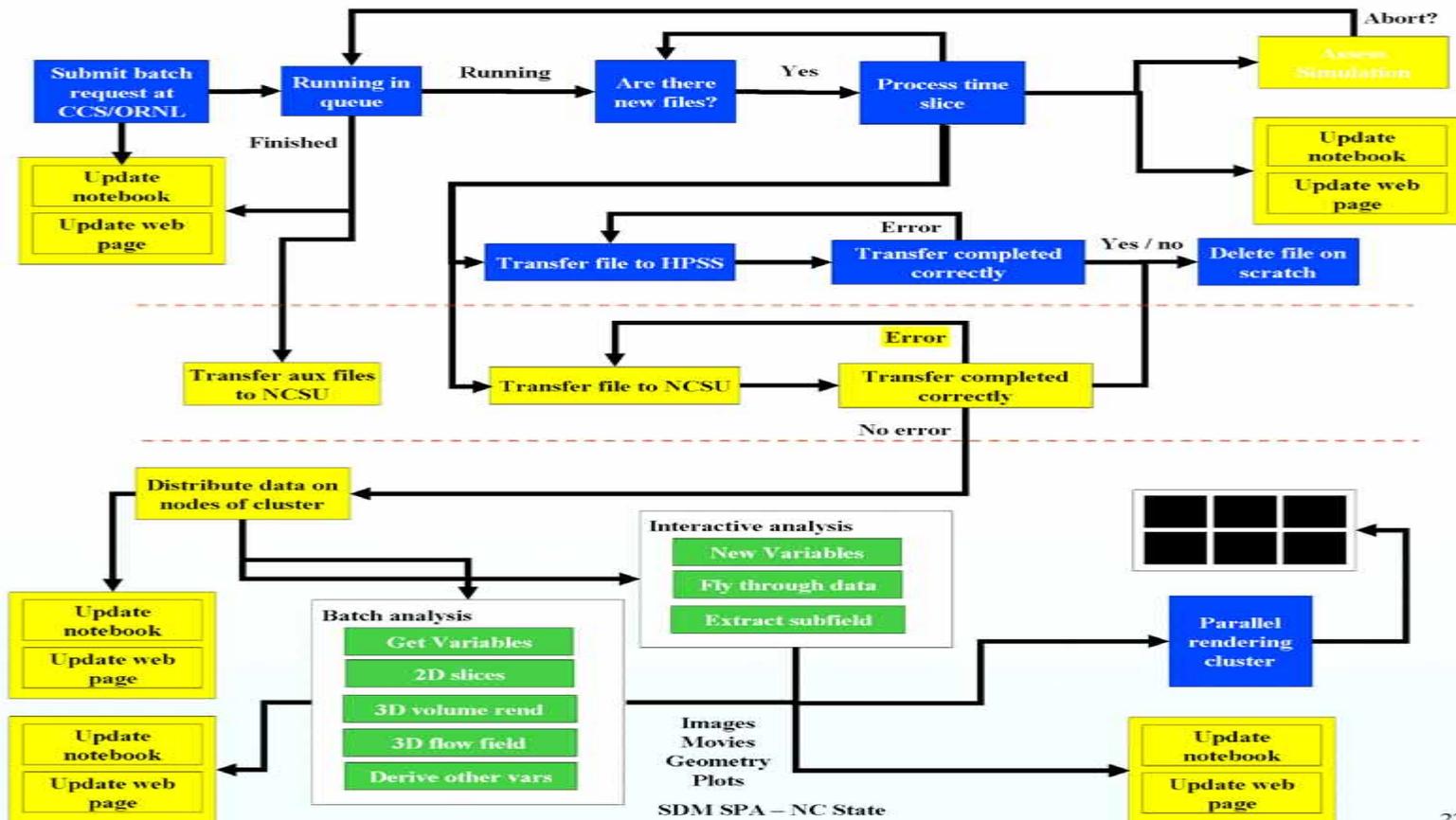
In conjunction with John Blondin, NCSU

Automate data generation, transfer and visualization of a large-scale simulation at ORNL

Blondin / TSI Scientific Workflow – V1 (using the Swesty template)



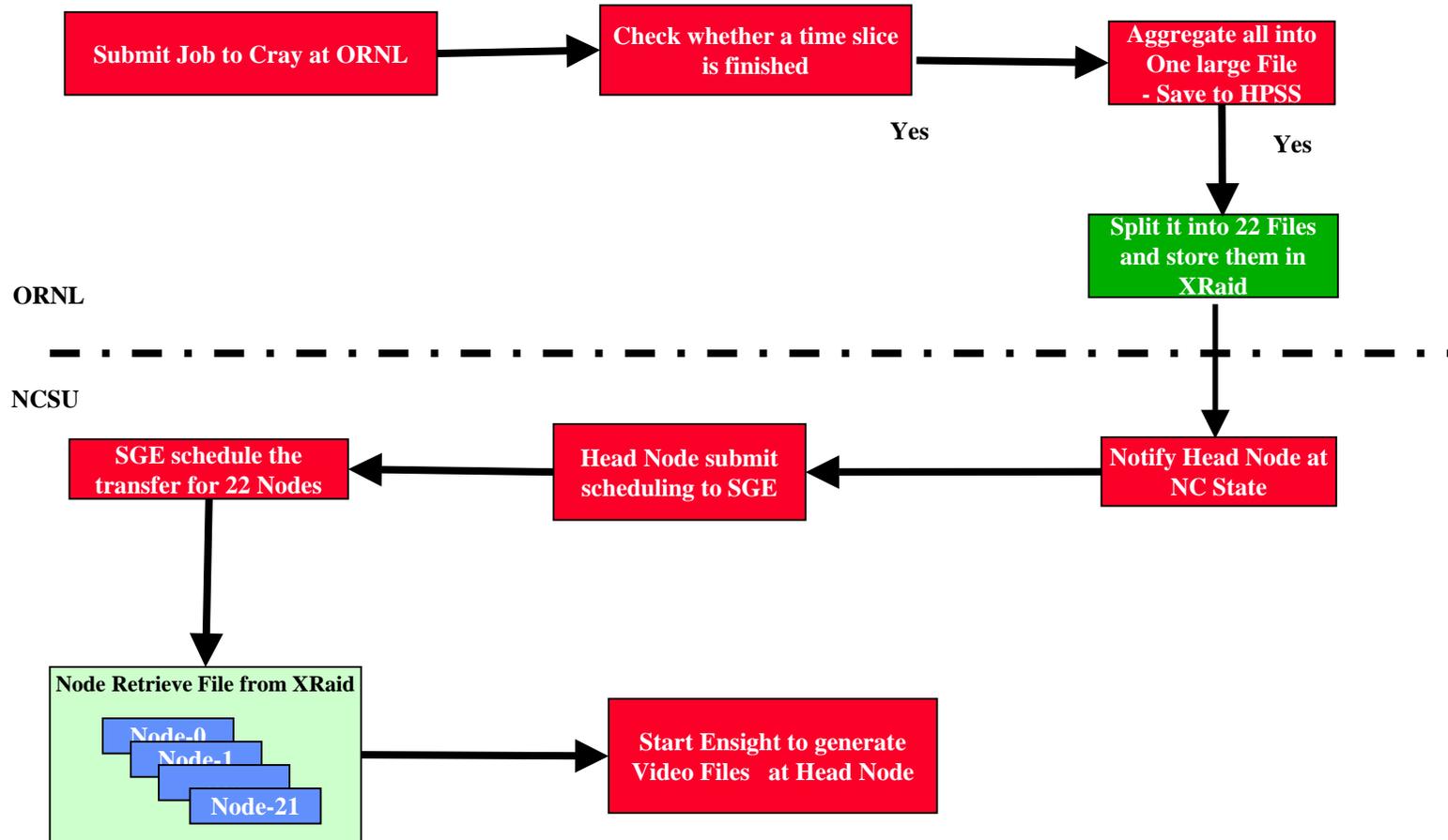
J. Blondin, E. Peele, Z. Cheng, P. Oothongsap, M. Vouk
North Carolina State University



TSI Workflow Example

In conjunction with John Blondin, NCSU

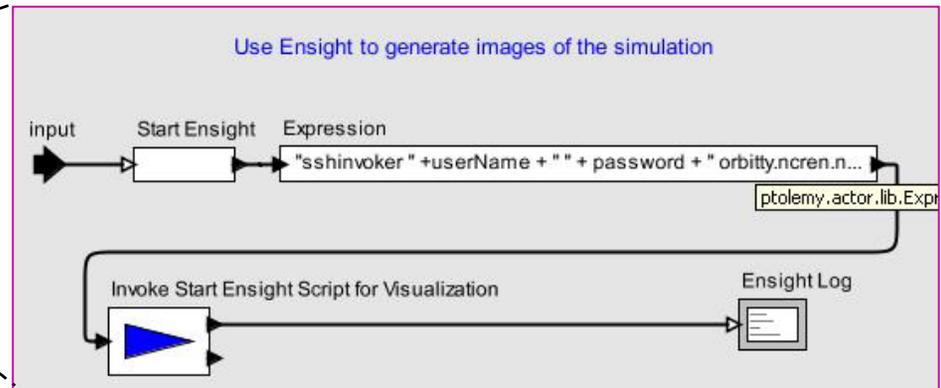
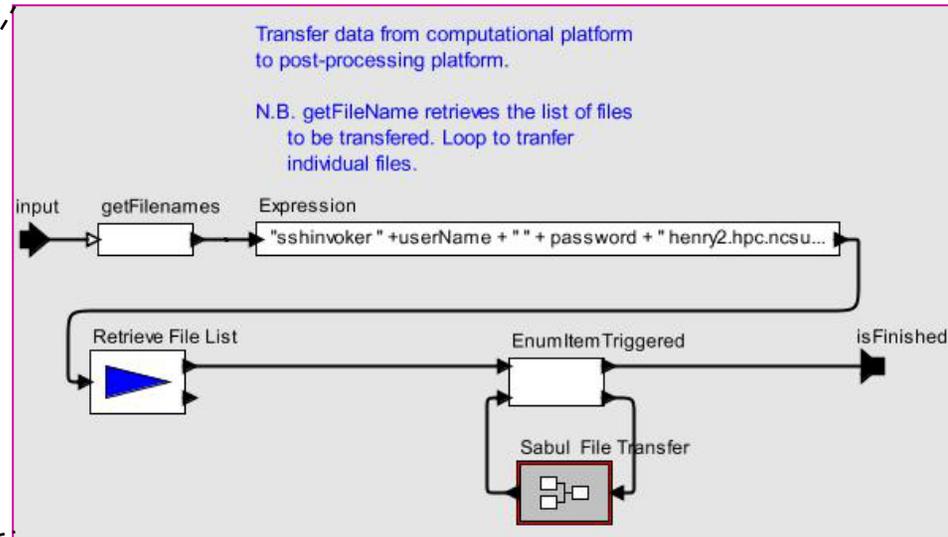
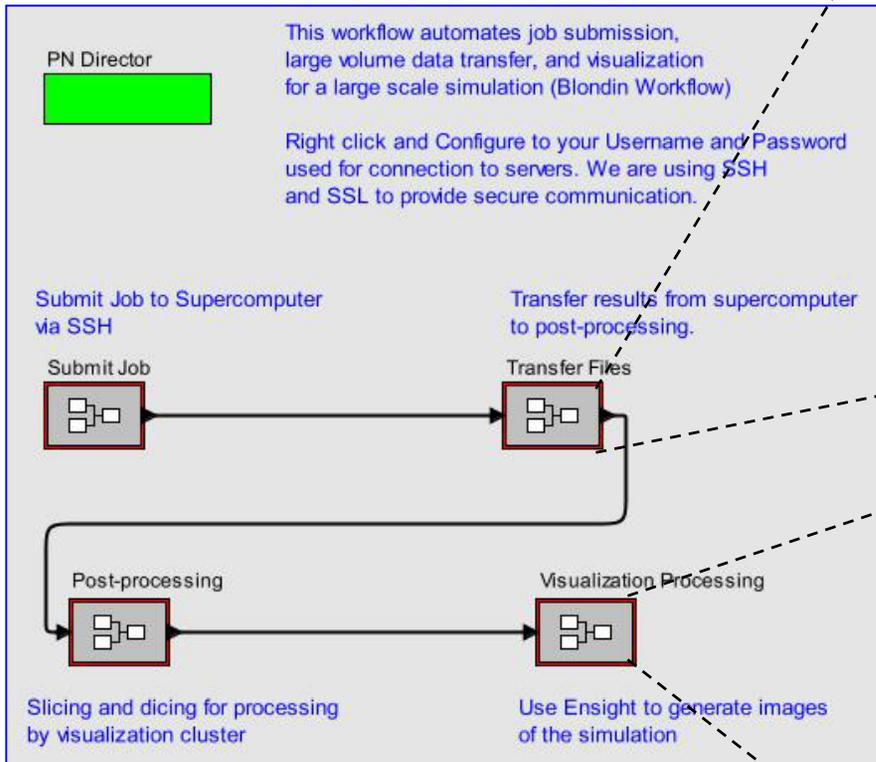
Automate data generation, transfer and visualization of a large-scale simulation at ORNL



Using the Scientific Workflow Tool (Kepler)

Emphasizing Dataflow (SDSC, NCSU, LLNL)

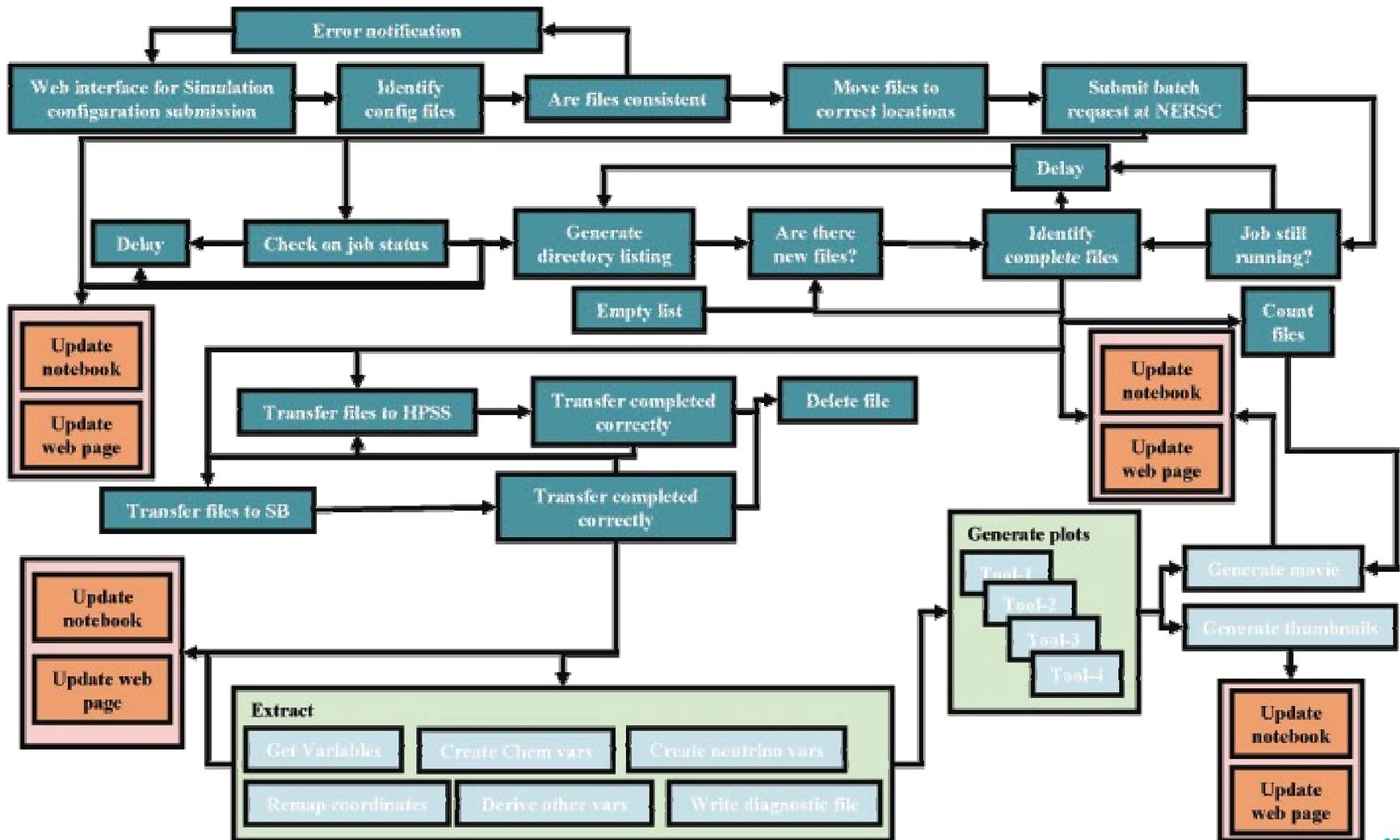
Automate data generation, transfer and visualization of a large-scale simulation at ORNL



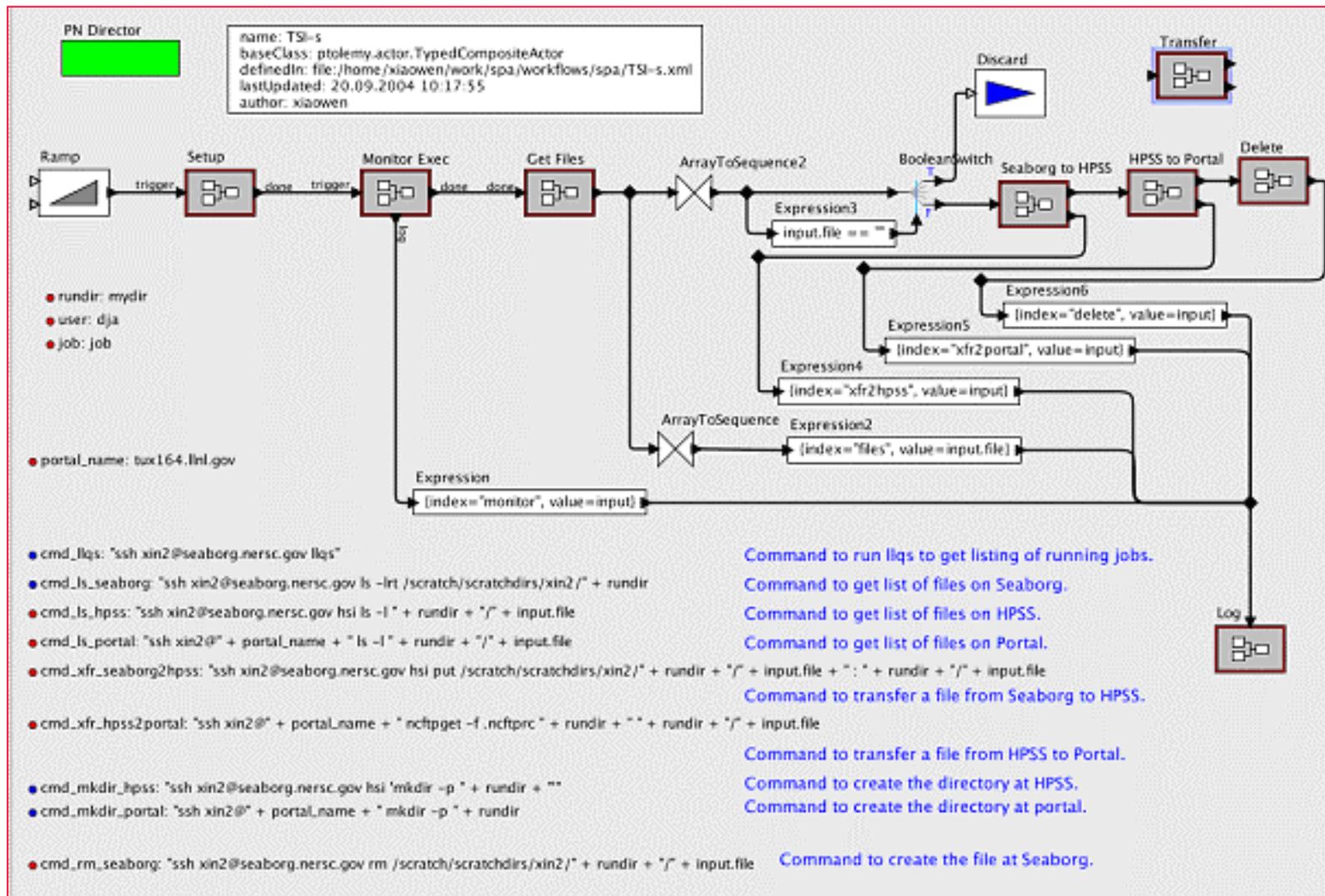
TSI Workflow Example

with Doug Swesty and Eric Myra, Stony Brook

Automate the transfer of large-scale simulation data between NERSC and Stony Brook



Using the Scientific Workflow Tool



Collaborating Application Scientists

- **Matt Coleman - LLNL (Biology)**
- **Tony Mezzacappa – ORNL (Astrophysics)**
- **Ben Santer – LLNL**
- **John Drake - ORNL (Climate)**
- **Doug Olson - LBNL, Wei-Ming Zhang – Kent (HENP)**
- **Wendy Koegler, Jacqueline Chen – Sandia Lab (Combustion)**
- **Mike Papka - ANL (Astrophysics Vis)**
- **Mike Zingale – U of Chicago (Astrophysics)**
- **John Michalakes – NCAR (Climate)**
- **Keith Burrell - General Atomics (Fusion)**

Re-apply technology to new applications

- **Parallel NetCDF**
 - Astrophysics → Climate
- **Parallel VTK**
 - Astrophysics → Climate
- **Compressed bitmaps**
 - HENP → Combustion → Astrophysics
- **Storage Resource Managers (MSS access)**
 - HENP → Climate → Astrophysics
- **Feature Selection**
 - Climate → Fusion
- **Scientific Workflow**
 - Biology → Astrophysics (planned)

Summary

- **Focus: getting technology into the hands of scientists**
- **Integrated framework**
 - **Storage Efficient Access technology**
 - **Data Mining and Analysis tools**
 - **Scientific Process (workflow) Automation**
- **Technology migration to new applications**
- **SDM Framework facilitates integration, generalization, and reuse of SDM technology**