SCIENTIFIC DATA MANAGEMENT:
Essential Technology for Data-Intensive Science

Arie Shoshani

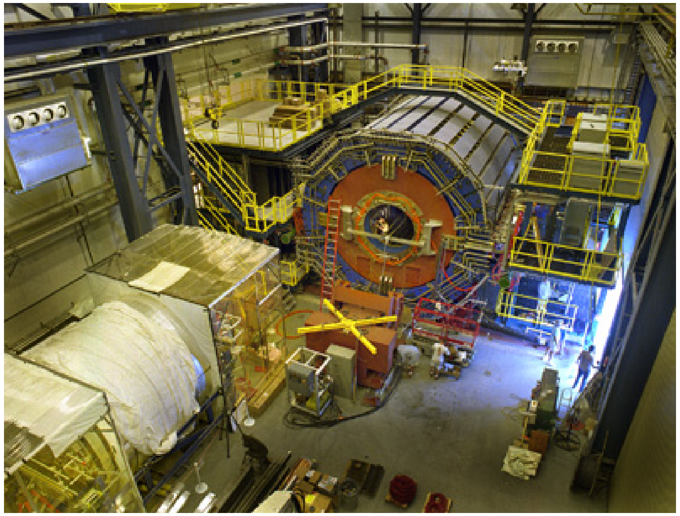
High Performance Computing Research Department
Computational Research Division,
Computing Sciences Directorate.

November, 2006

Motivating Example 1: HENP Data Generation Analysis Phases

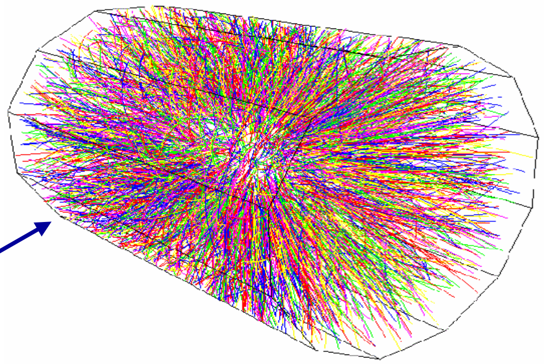
- Optimizing Storage Management and Data Access for High Energy and Nuclear Physics Applications

Experiment	# members /institutions	Date of first data	# events/ year	volume/year-TB
STAR	350/35	2001	10^8-10^9	500
PHENIX	350/35	2001	10^9	600
BABAR	300/30	1999	10^9	80
CLAS	200/40	1997	10^{10}	300
ATLAS	1200/140	2007	10^{10}	5000



STAR: Solenoidal Tracker At RHIC
RHIC: Relativistic Heavy Ion Collider

LHC: Large Hadron Collider
Includes: ATLAS, STAR, …

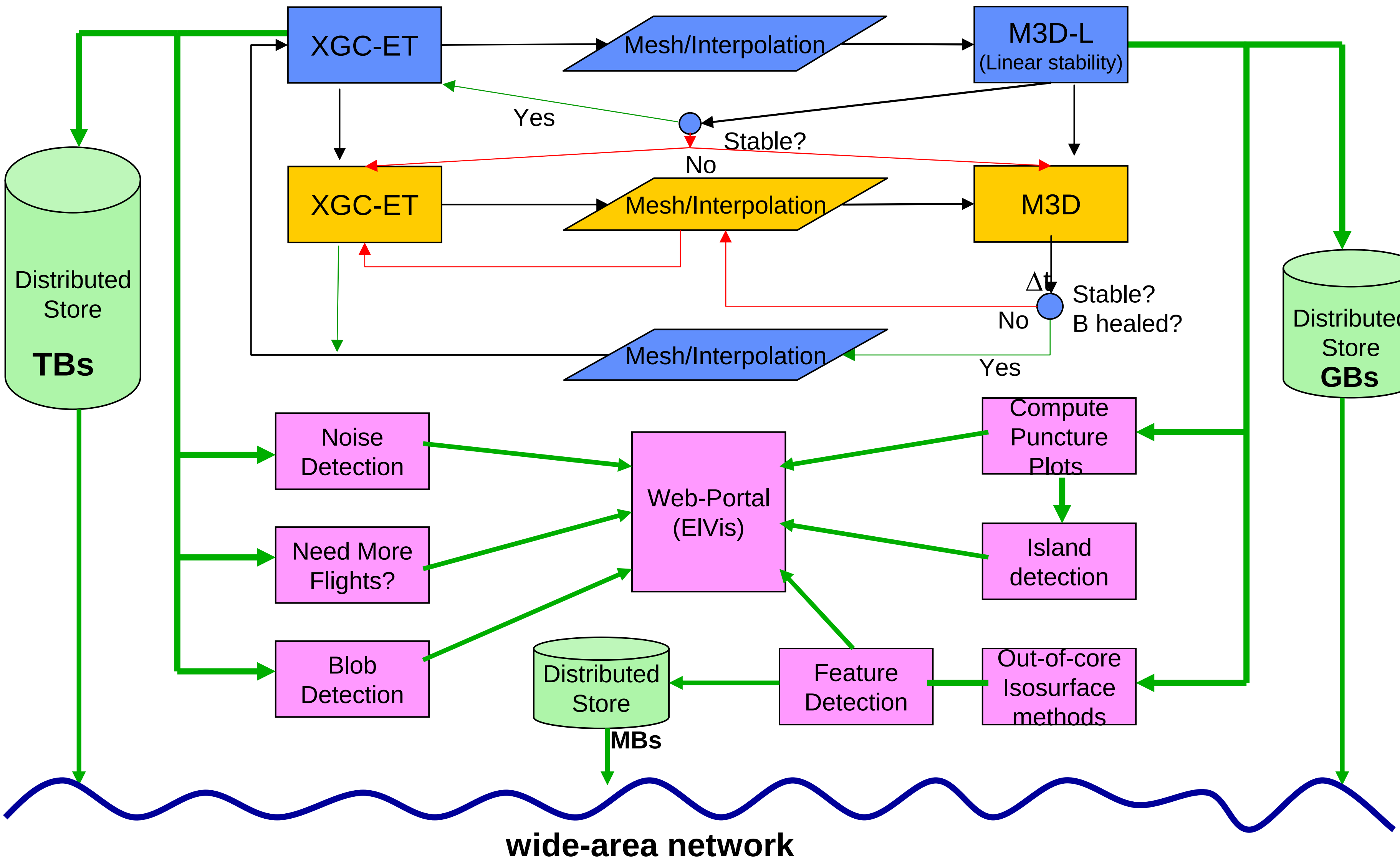


A mockup of
An "event"

Typical Scientific Exploration Process

- Data Generation Phase
  - Generate large amounts of raw data
    - large simulations
    - collect from experiments
- Post-processing Phase
  - analyze data (find particles produced, tracks)
  - generate summary data
    - e.g. momentum, no. of pions, transverse energy
    - Number of properties is large (50-100)
- Analysis Phase
  - use summary data as guide
  - extract subsets from the large dataset
    - Need to access events based on partial properties specification (range queries)
    - e.g. ((0.1 < AVpT < 0.2) ^ (10 < Np < 20)) v (N > 6000)
  - apply analysis code

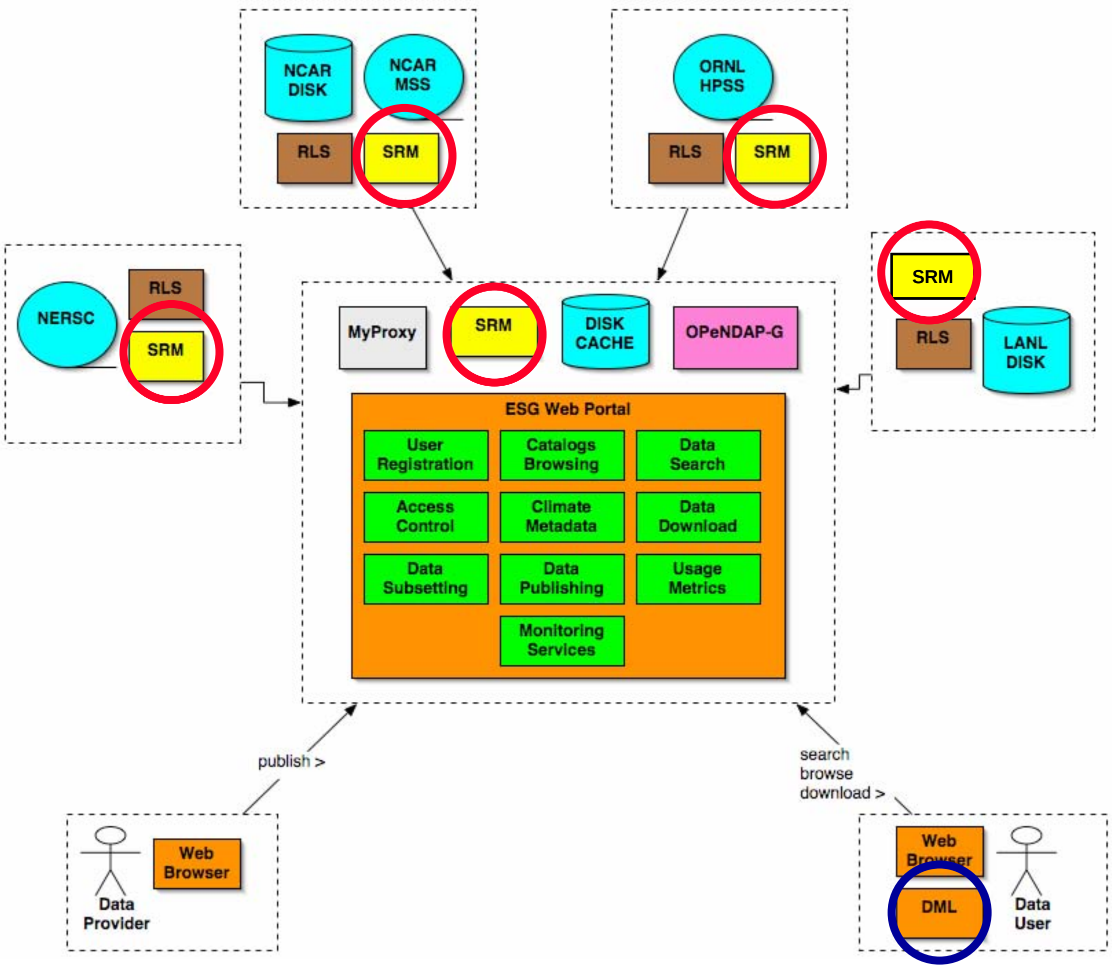
Motivating Example 3: Fusion Coordination between Running Codes



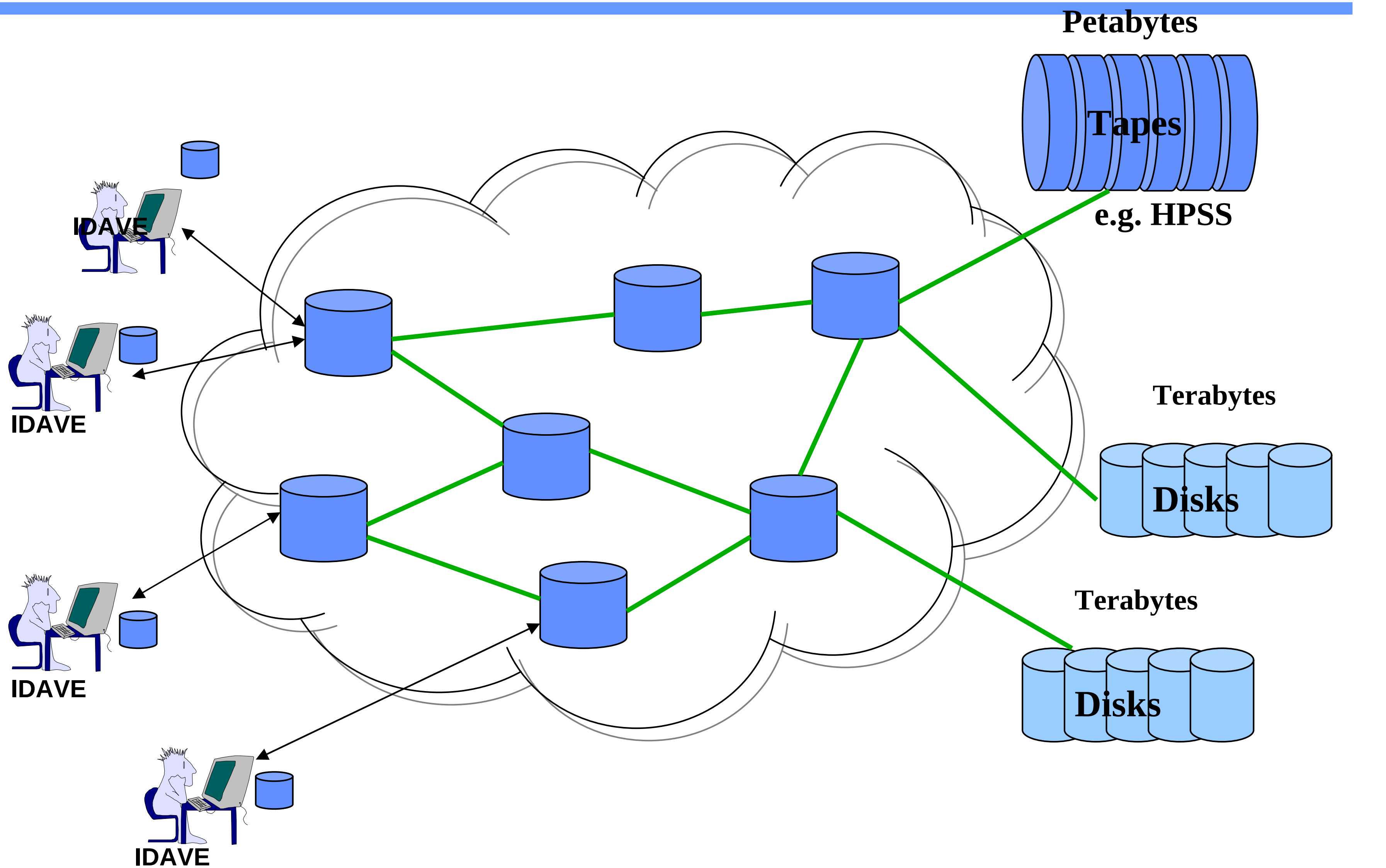
Motivating Example 2: Climate Modeling Large Scale Data Distribution

- Earth System Grid

  Accessing large distributed stores by 1000’s of scientists
- Problems
  - * Different systems
  - * Security procedures
  - * File streaming
  - * Lifetime of request
  - Garbage collection
- Solution
  - Storage Resource Managers (SRMs)



Ubiquitous and Transparent Data Sharing



The Scientific Data Management Center

(Center for Enabling Technologies - CET)

- PI: Arie Shoshani, LBNL
- Annual budget: 3.3 Million
- Established 5 years ago (SciDAC-1)
- Successfully re-competed for the next 5 years (SciDAC-2)
- Featured in next issue of SciDAC magazine
- Laboratories
  - ANL, ORNL, LBNL, LLNL, PNNL
- Universities
  - NCSU, NWU, SDSC, UCD, Uof Utah,

FROM DATA TO DISCOVERY

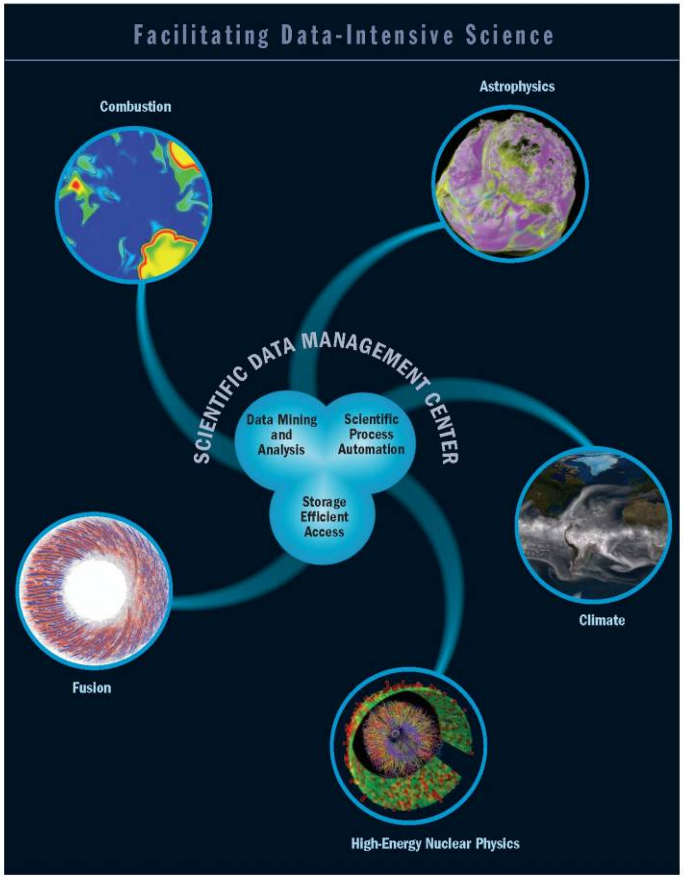
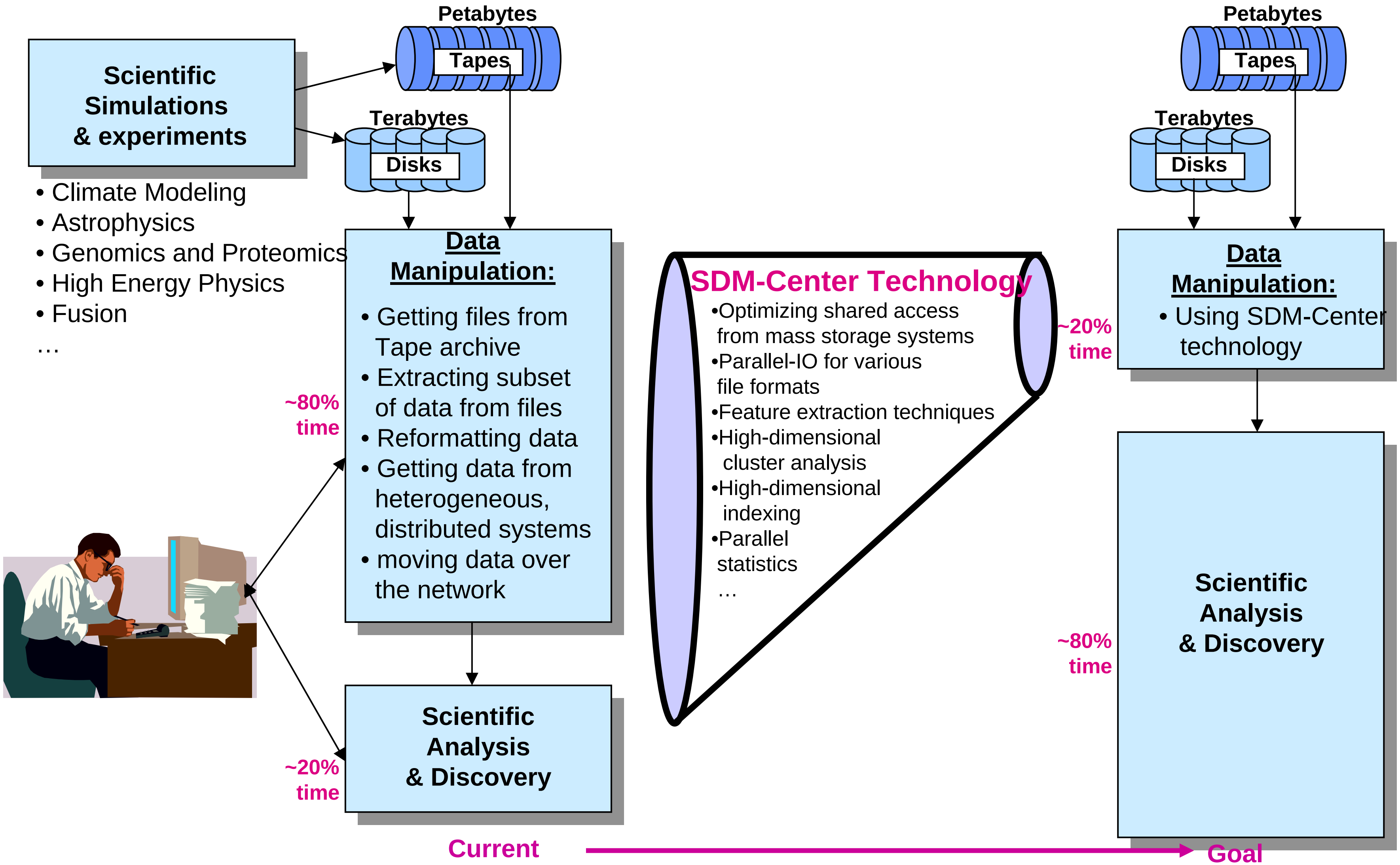
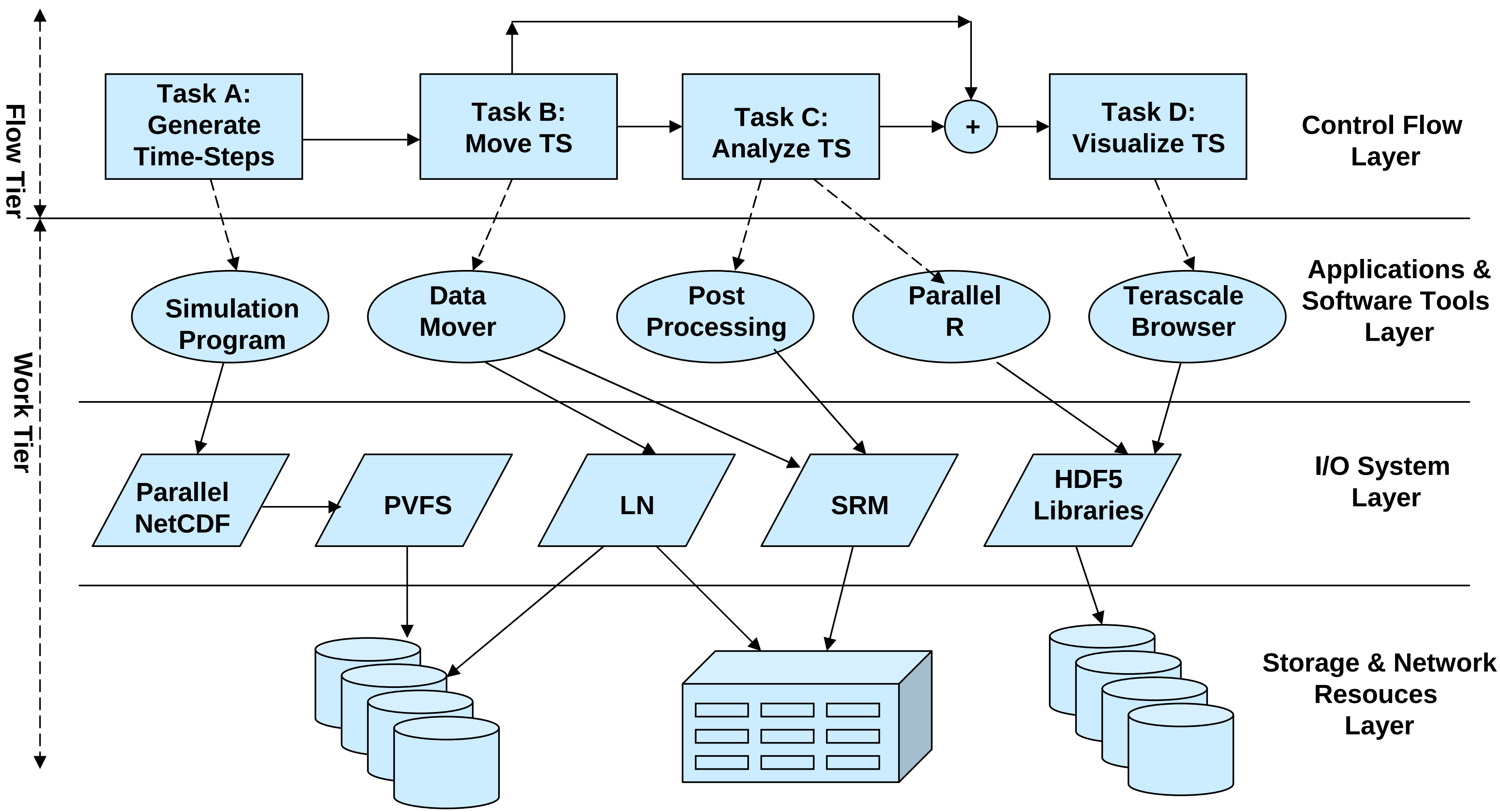


Figure 10. *Combustion, Astrophysics, Climate, High-Energy Physics, and Fusion are some of the many research applications that have benefited from the creative technologies offered by the SDM Center*

Scientific Data Management Center

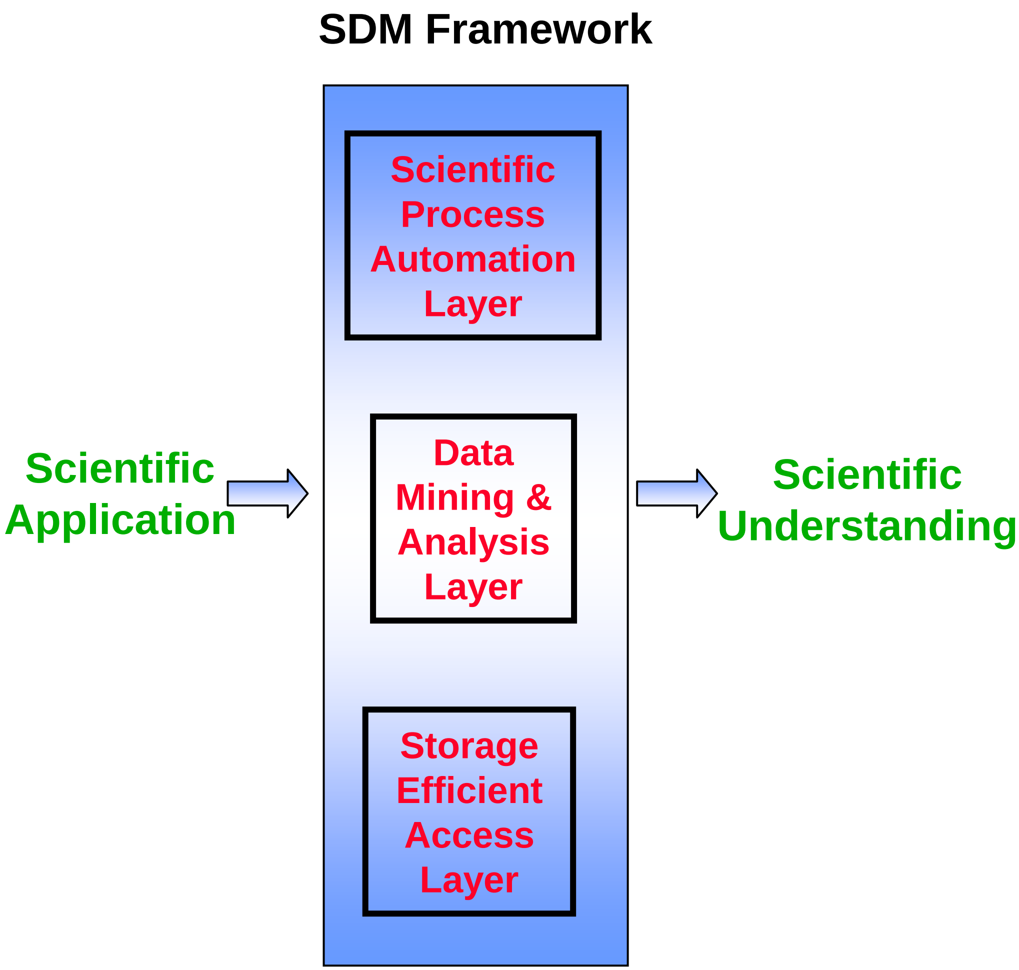


A Typical SDM Scenario

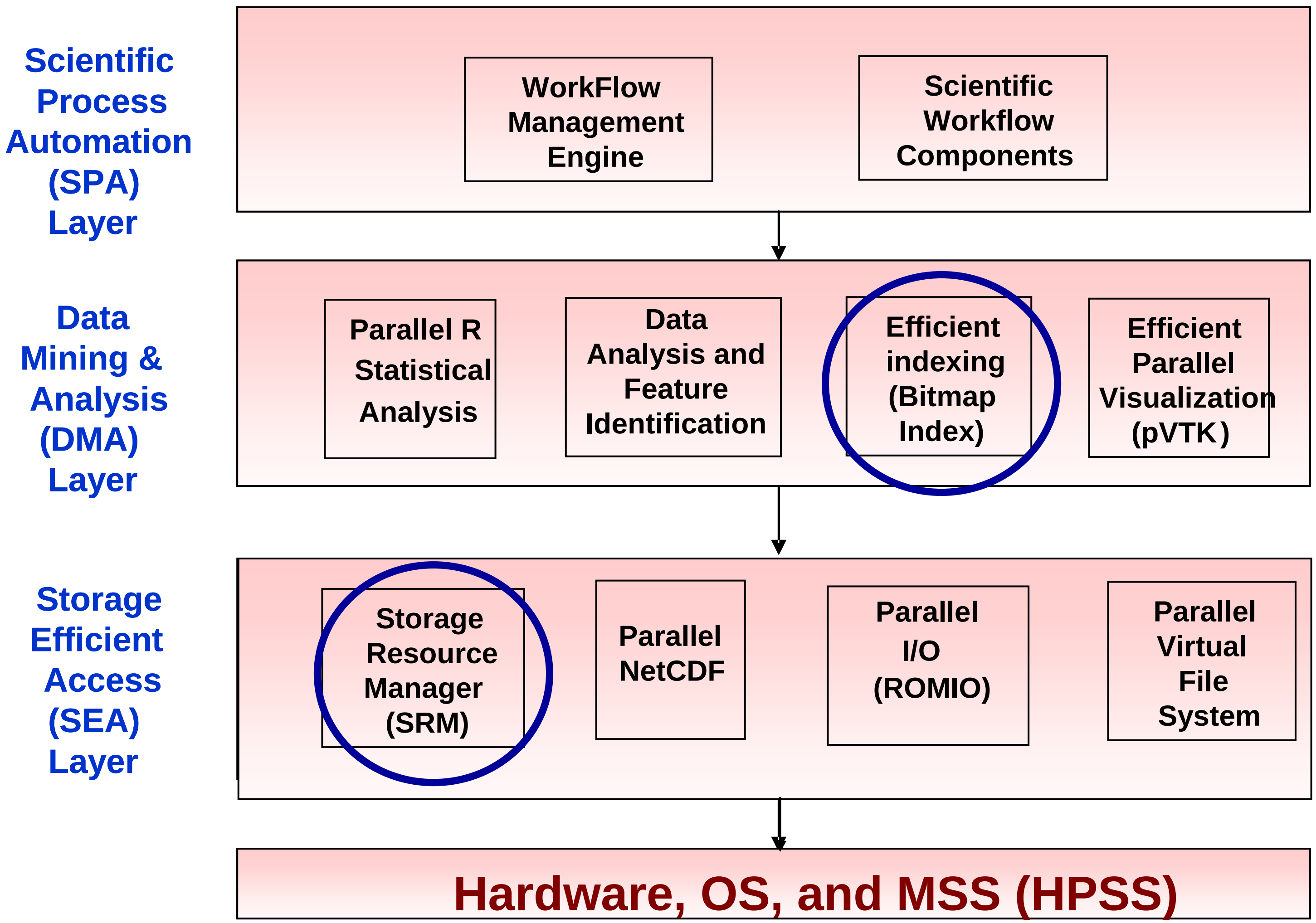


Approach

- Use an integrated framework that:
  - Provides a scientific workflow capability
  - Supports data mining and analysis tools
  - Accelerates storage and access to data
- Simplify data management tasks for the scientist
  - Hide details of underlying parallel and indexing technology
  - Permit assembly of modules using a simple graphical workflow description tool



Technology Details by Layer



SDM Technology: Layering of Components

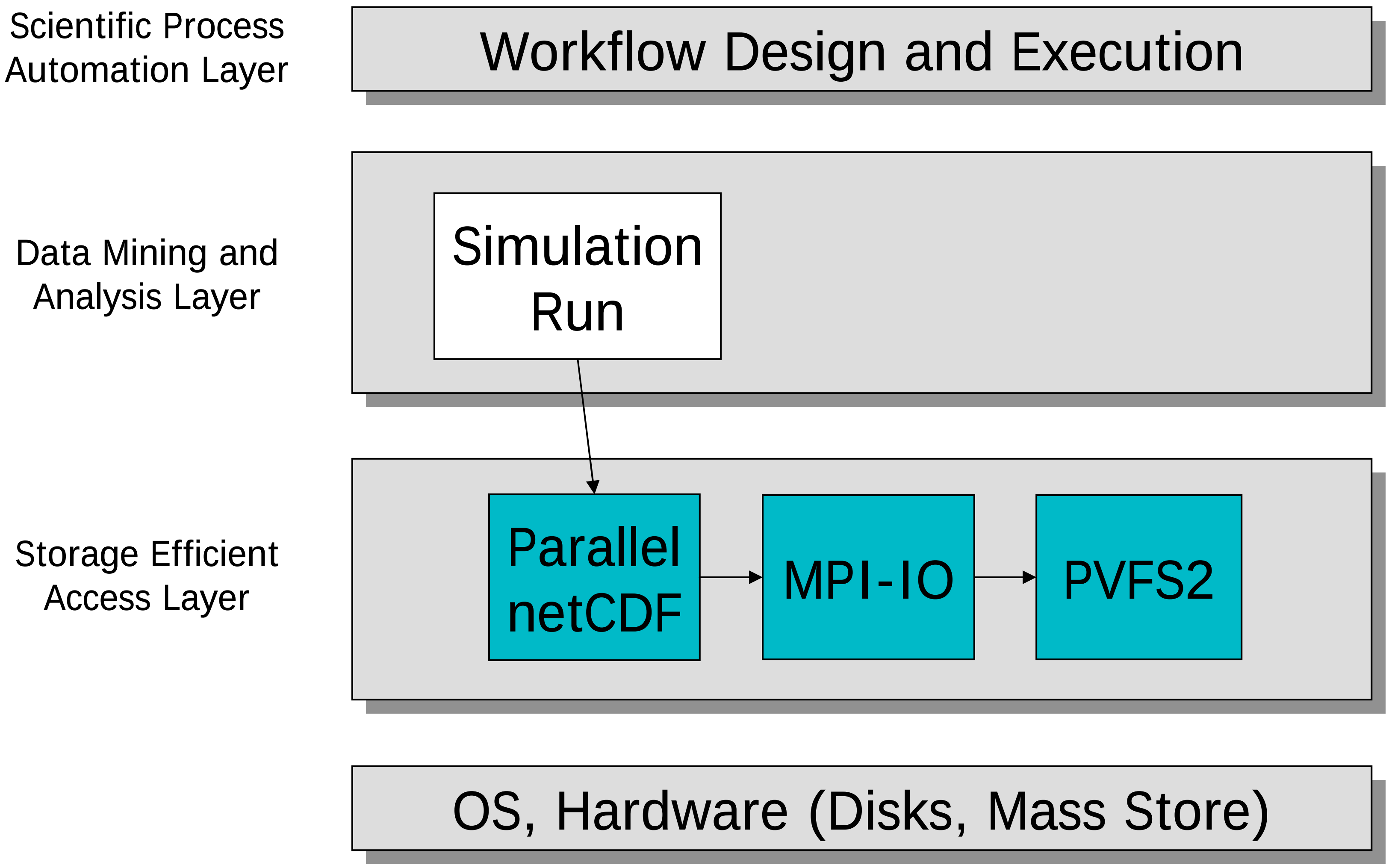
Scientific Process Automation (Workflow) Layer

Data Mining & Analysis Layer

Storage Efficient Access Layer

Hardware, OS, and MSS (HPSS)

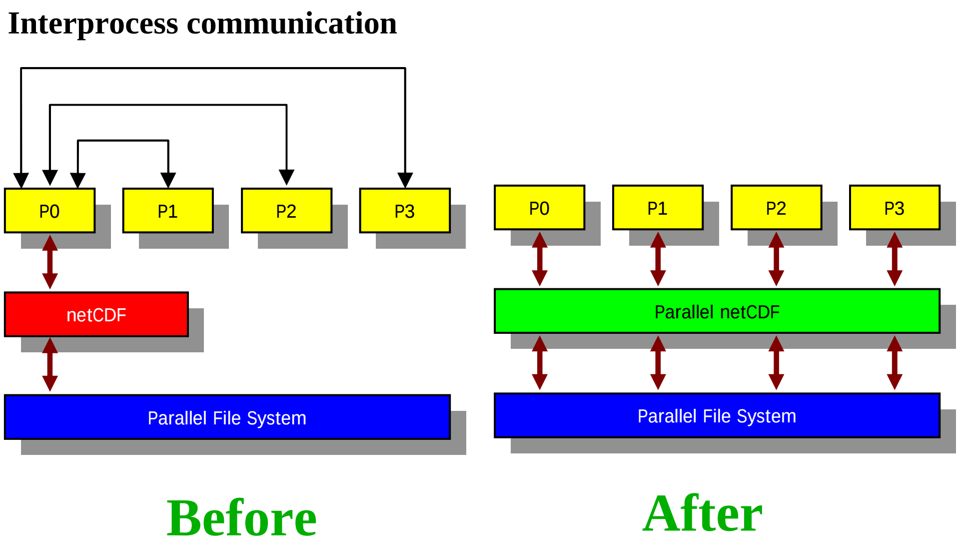
Data Generation

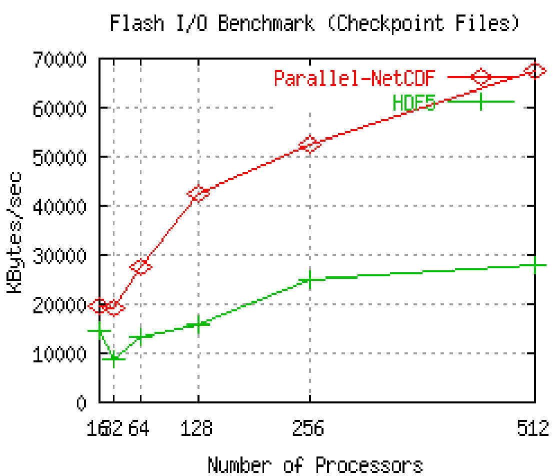


Parallel NetCDF v.s. HDF5 (ANL+NWU)

Parallel Virtual File System:
Enhancements and deployment

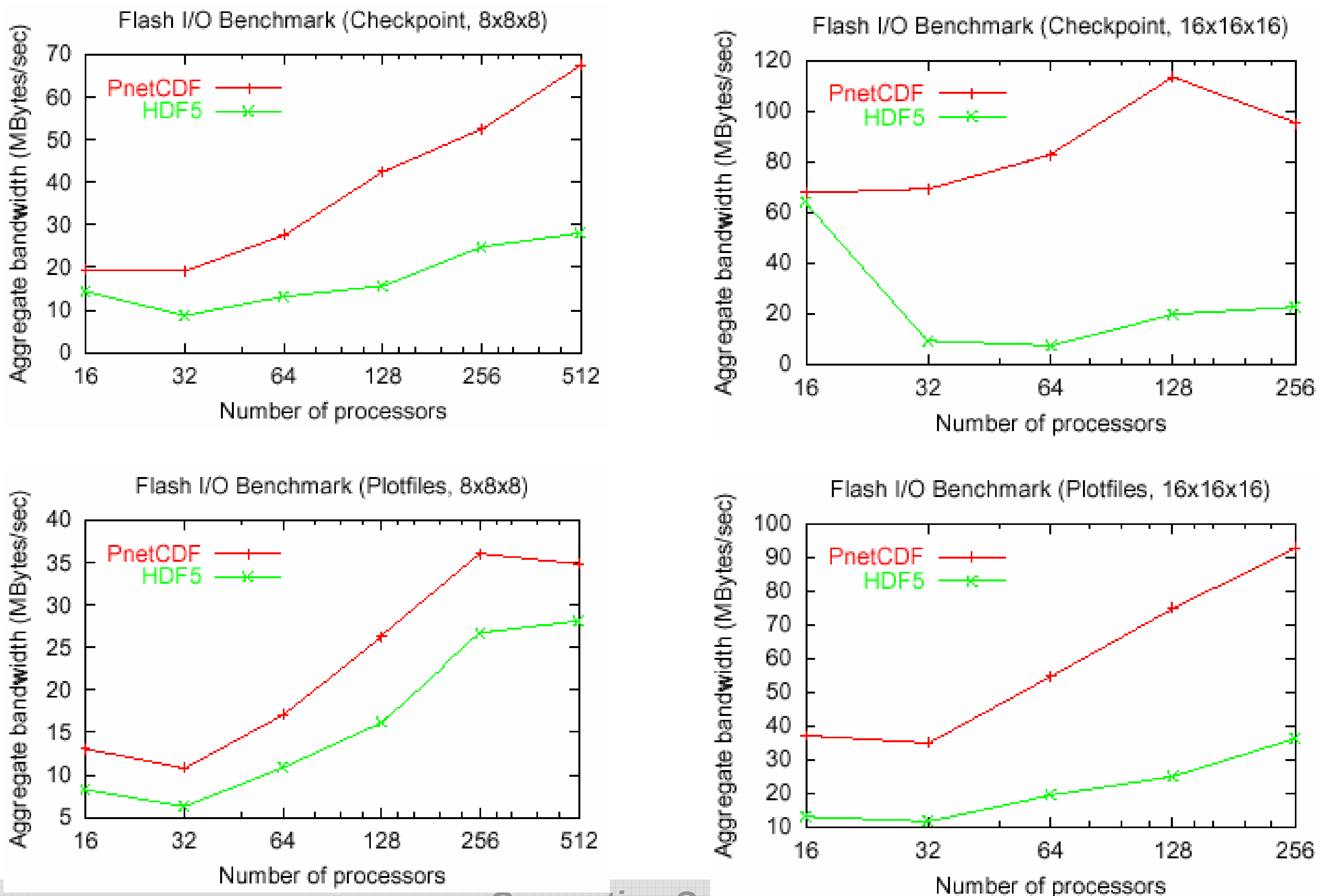
- Developed Parallel netCDF
  - Enables high performance parallel I/O to netCDF datasets
  - Achieves up to 10 fold performance improvement over HDF5
- Enhanced ROMIO:
  - Provides MPI access to PVFS2
  - Advanced parallel file system interfaces for more efficient access
- Developed PVFS2
  - Production use at ANL, Ohio SC, Univ. of Utah HPC center
  - Offered on Dell clusters
  - Being ported to IBM BG/L system
- Deployed an HPSS Storage Resource Manager (SRM) with PVFS
  - Automatic access of HPSS files to PVFS through MPI-IO library
  - SRM is a middleware component



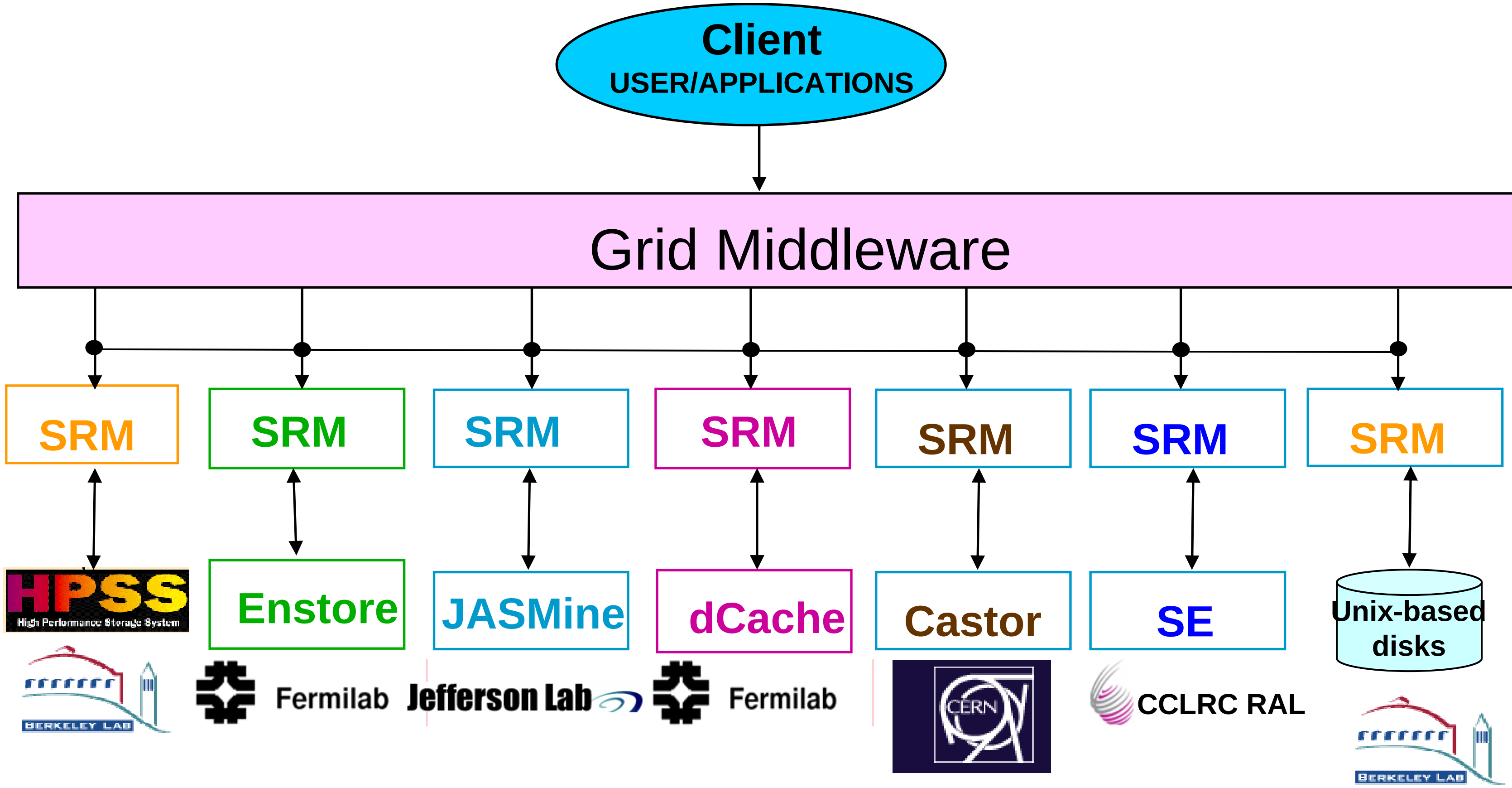


FLASH I/O Benchmark Performance (8x8x8 block sizes)

Parallel netCDF Performance

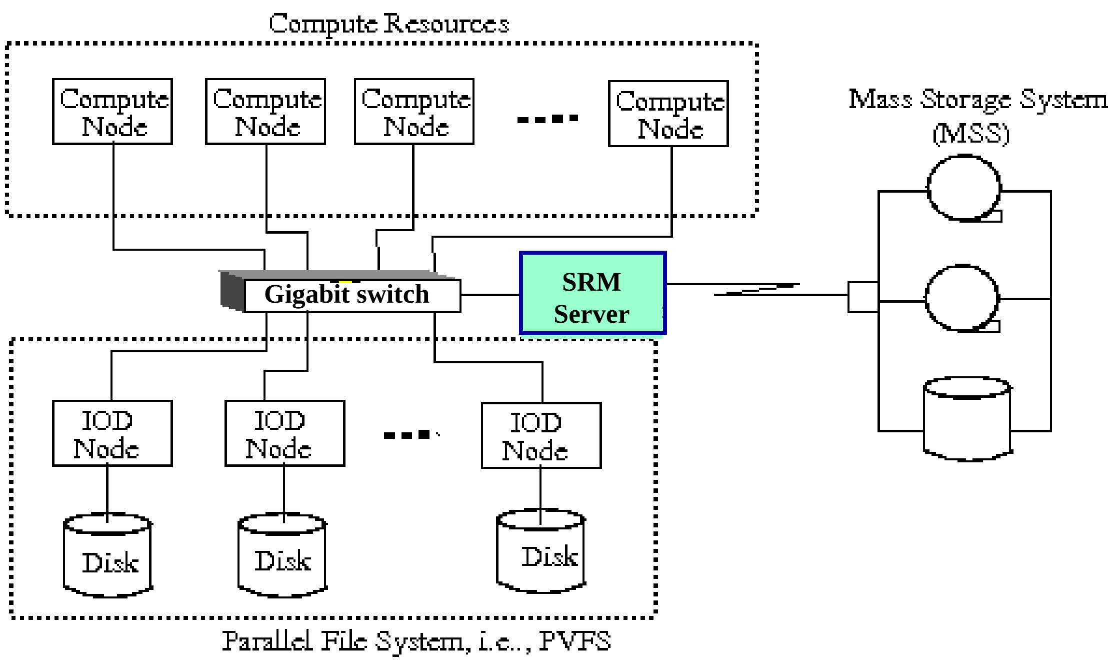


Storage Resource Managers (SRMs)



MPI/IO-SRM (LBNL)*

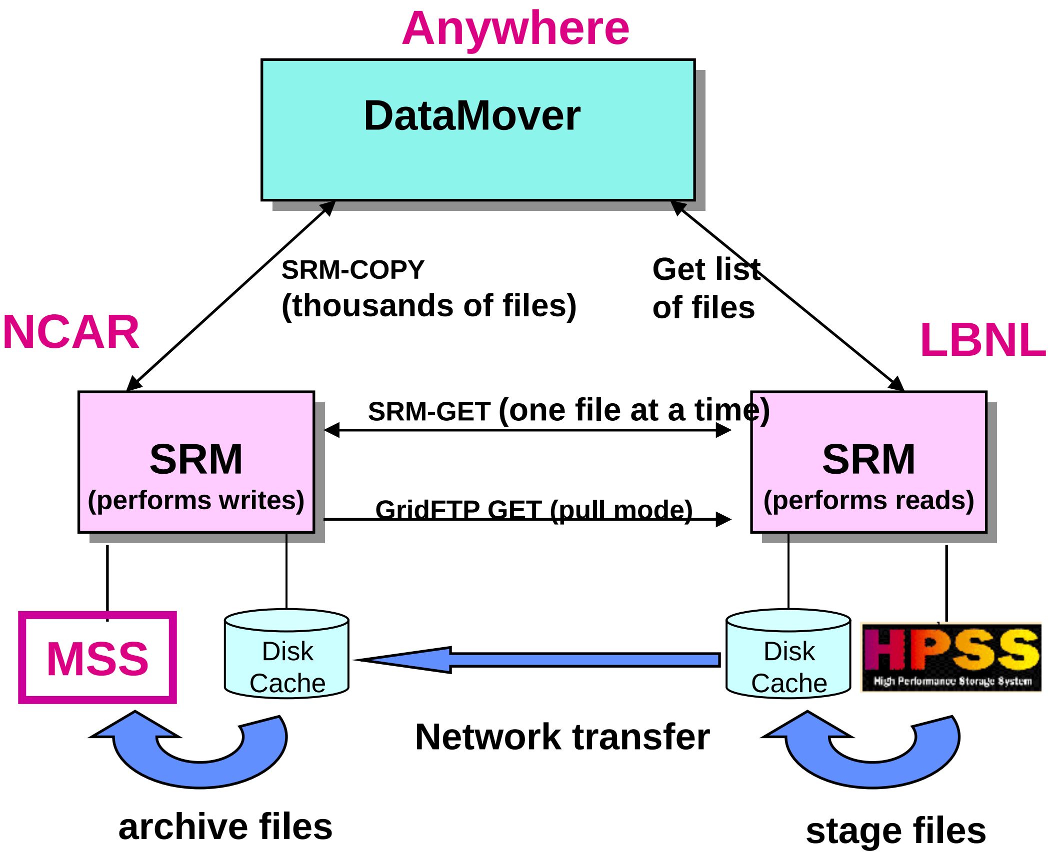
- Collaboration of LBNL and ANL
  - Access through MPI/IO to files that are on mass storage system
  - Developed MPI/IO-SRM libraries to interact with SRM and HPSS: different hint scenarios (set, bundle, ordered-list)



Data Flow Requires Robust Data Movement (LBNL)

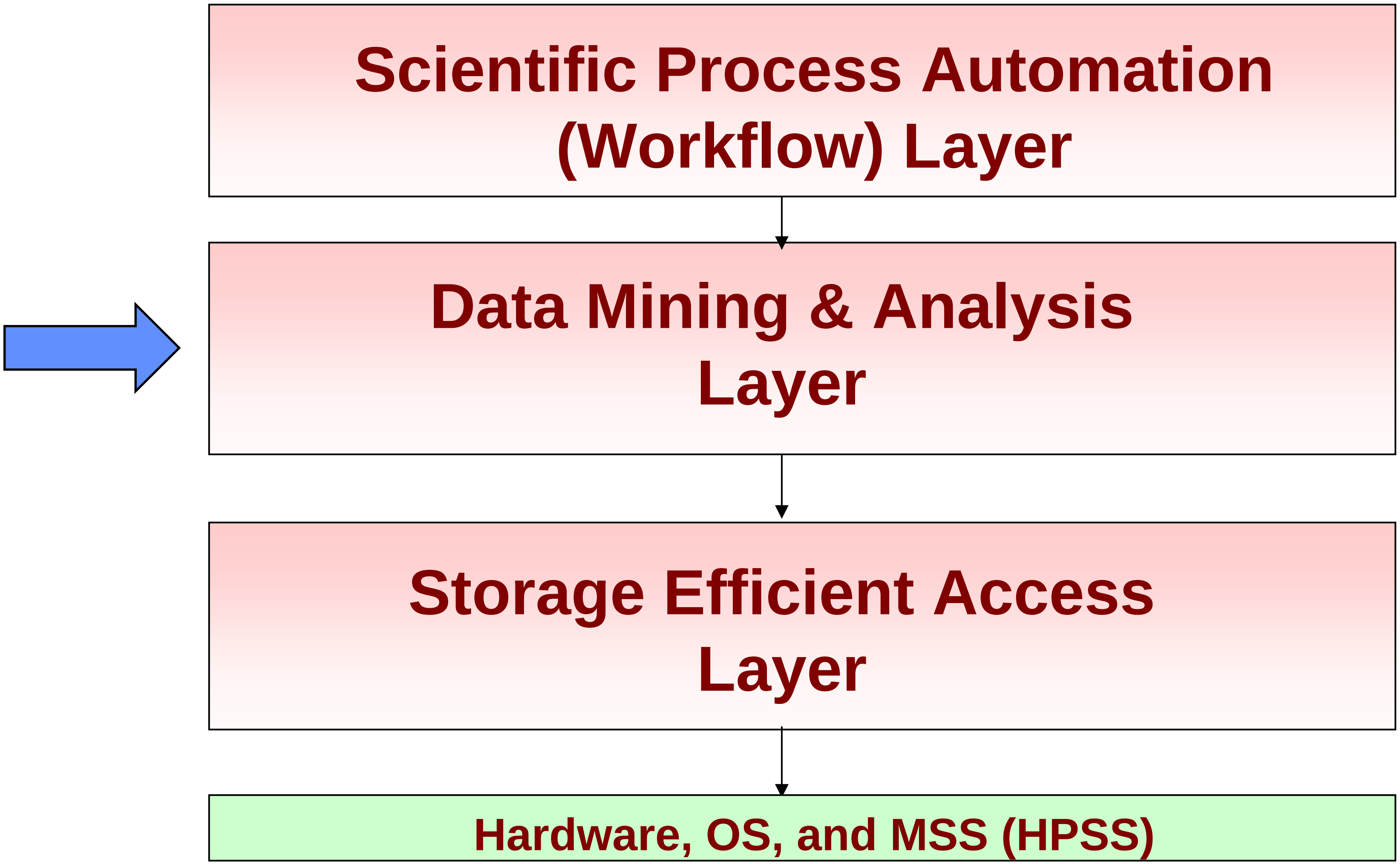
- Problem: move thousands of files robustly
  - Takes many hours
  - Need error recovery
    - Mass storage systems failures
    - Network failures
- Solution: Use Storage Resource Managers (SRMs)

- Problem: too slow
- Solution:
  - in GridFTP
    - Use parallel streams
    - Use large FTP windows
  - Pre-stage files from MSS
  - Use concurrent transfers



* http://www.ppdg.net/docs/oct04/ppdg-star-oct04.doc

Layering of Components



Statistical Computing with *R*

About *R* (http://www.r-project.org/):

- *R* is an Open Source (GPL), most widely used programming environment for statistical analysis and graphics; similar to *S*.
- Provides good support for both users and developers.
- Highly extensible via dynamically loadable add-on packages.
- Originally developed by Robert Gentleman and Ross Ihaka.

```
> library(mva)
> pca <- prcomp(data)
> summary(pca)
```

```
> …
> dyn.load( "foo.so")
> .C( "foobar" )
> dyn.unload( "foo.so" )
```

```
> library (rpvm)
> .PVM.start.pvmd ()
> .PVM.addhosts (...)
> .PVM.config ()
```

Towards Enabling Parallel Computing in *R*:

- ***Rmpi*** (Hao Yu): ***R*** interface to ***LAM-MPI***.
- ***rpvm*** (Na Li and Tony Rossini): ***R*** interface to ***PVM***; requires knowledge of parallel programming.
- ***snow*** (Luke Tierney): general API on top of message passing routines to provide high-level (*parallel apply*) commands; mostly demonstrated for *embarrassingly parallel* applications .

snow API

High Level Routines	
`parLapply`	parallel `lapply`
`parSapply`	parallel `sapply`
`parApply`	parallel `apply`
Basic Routines	
`clusterExport`	export variables to nodes
`clusterCall`	call function on each node
`clusterApply`	apply function to arguments on nodes
`clusterApplyLB`	load balanced `clusterApply`
`clusterEvalQ`	evaluate explicit expression on nodes
`clusterSplit`	split vector into pieces for nodes
Administrative Routines	
`makeCluster`	create a new cluster of nodes
`stopCluster`	shut down a cluster
`clusterSetupSPRNG`	initialize random number streams

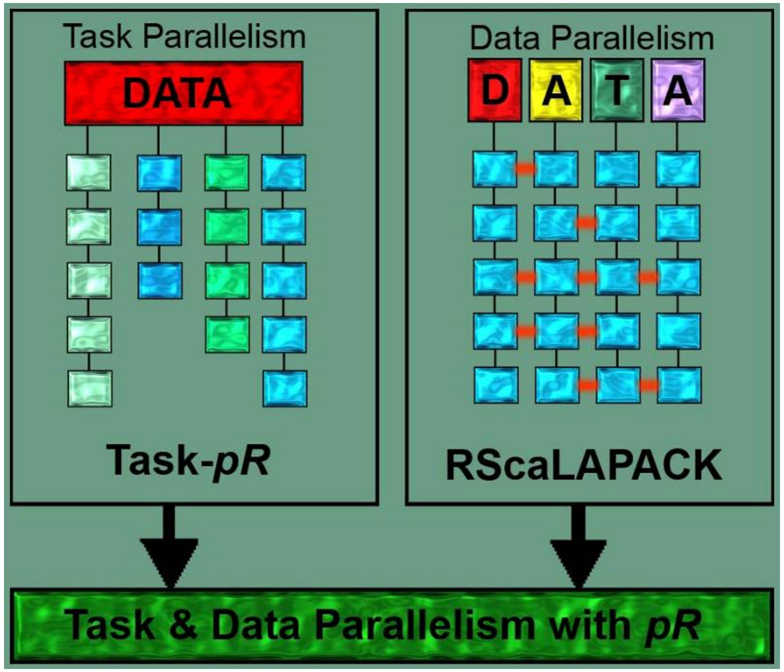
Providing Task and Data Parallelism in *pR*

Task-parallel analyses:

- Likelihood Maximization.
- Re-sampling schemes: Bootstrap, Jackknife, etc.
- Animations
- Markov Chain Monte Carlo (MCMC).
  - – Multiple chains.
  - – Simulated Tempering: running parallel chains at different “temperature“ to improve mixing.

Data-parallel analyses:

- *k*-means clustering
- Principal Component Analysis (PCA)
- Hierarchical (model-based) clustering
- Distance matrix, histogram, etc. computations




```
:::::::
fileList<-list.files(pattern=”*.nc);
for (i in 1:length(fileList)) {
  matrix [i] ← readNcFile (fileList[i]);
  pca [i] ← prcomp (matrix [i])
}
::::::::::::
```

R

```
:::::::
fileList<-list.files(pattern=”*.nc”);
PE (  for (i in 1:length(fileList)) {
        matrix [i] ← readNcFile (fileList[i]);
        pca [i] ← sla.prcomp (matrix [i])
} )
::::::::::::::
```

pR

Parallel R (pR) Distribution

Releases History:

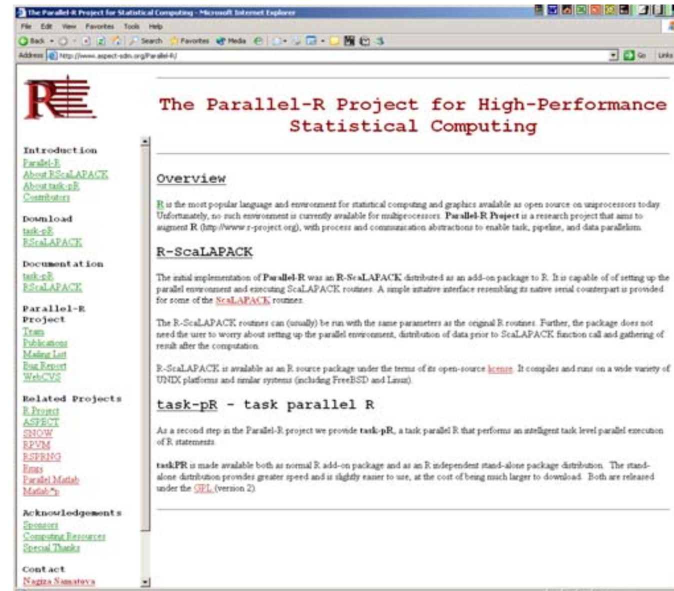
- ***pR*** enables both *data and task parallelism* (includes task-pR and RScaLAPACK) (version 1.8.1)

- ***RScaLAPACK*** provides ***R*** interface to ***ScaLAPACK*** with its scalability in terms of problem size and number of processors using *data parallelism* (release 0.5.1)

- ***task-pR*** achieves parallelism by performing *out-of-order execution of tasks*. With its intelligent scheduling mechanism it attains significant gain in execution times (release 0.2.7)

- ***pMatrix*** provides a parallel platform to perform *major matrix operations in parallel* using ScaLAPACK and PBLAS Level II & III routines

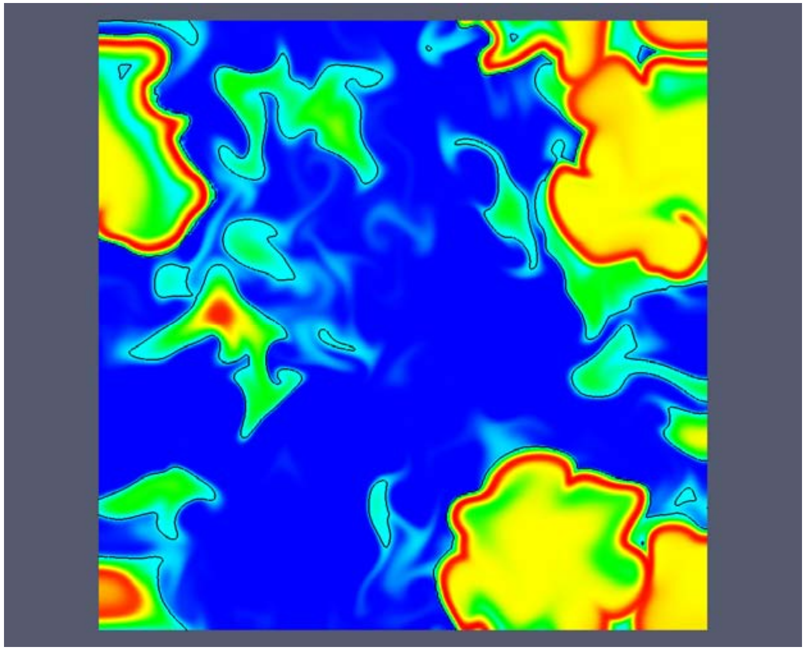
http://www.ASPECT-SDM.org/Parallel-R



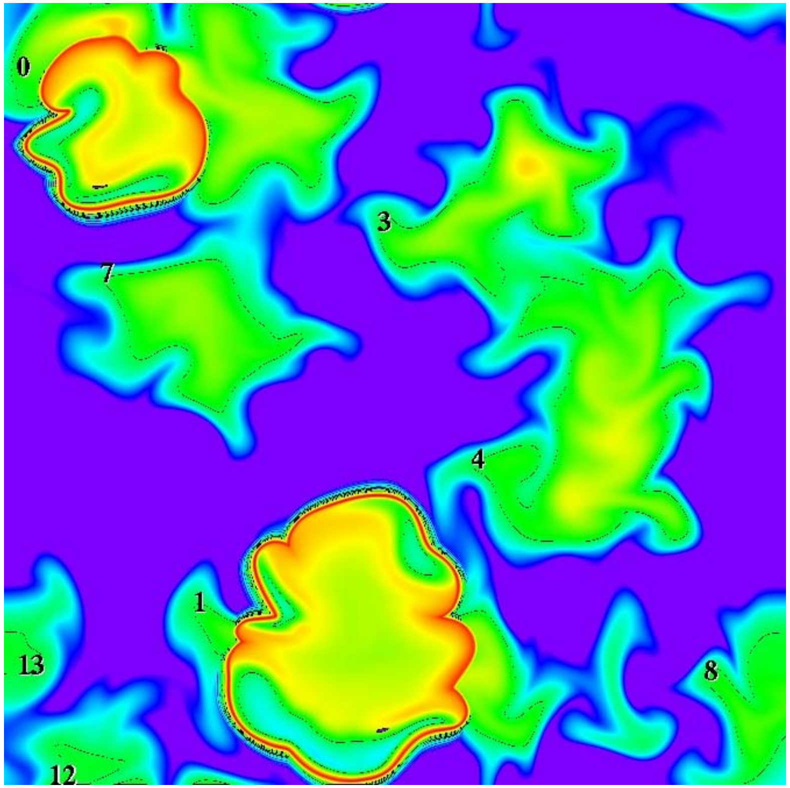
<u>Also</u>: Available for download from ***R***'s **CRAN** web site (www.R-Project.org) with **37 mirror sites in 20 countries**

FastBit – Bitmap Indexing Technology (LBNL)

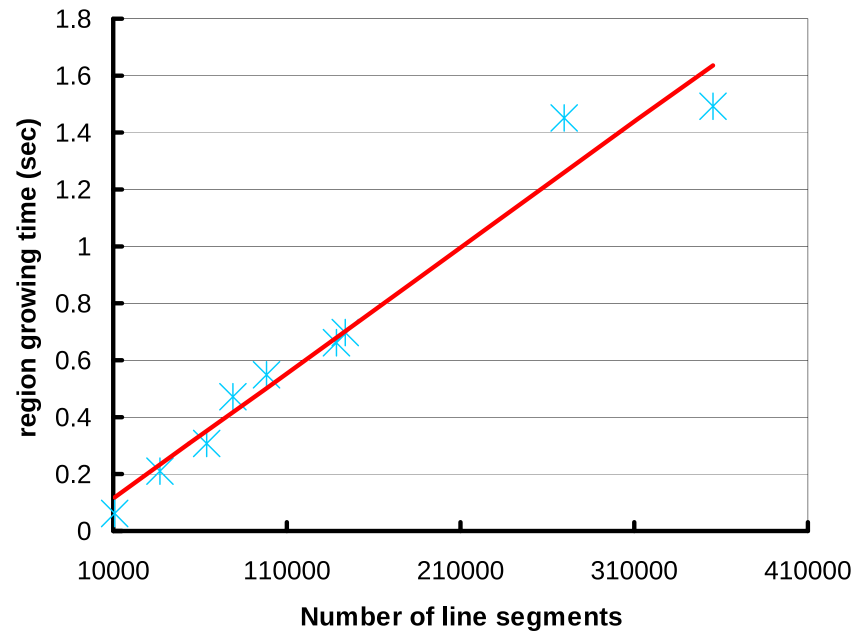
- Astrophysics or Combustion simulations (Spatio-temporal datasets)
  - 1000x1000x1000 mesh with 10-100s of variables per cell over 1000s of time steps - 10^{13} - 10^{14} data values
- This is an image of a <u>*single*</u> variable
- What's needed is search over <u>*multiple*</u> variables, such as
- Temperature > 1000
  AND pressure > 10^6
  AND HO_2 > 10^{-7} AND HO_2 > 10^{-6}
- Combining multiple single-variable indexes - a problem
- Solution: specialized bitmap indexes (logical operation on resulting bitmaps)
- Search time proportional to number of hits



FastBit-Based Multi-Attribute Region Finding is Theoretically Optimal



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

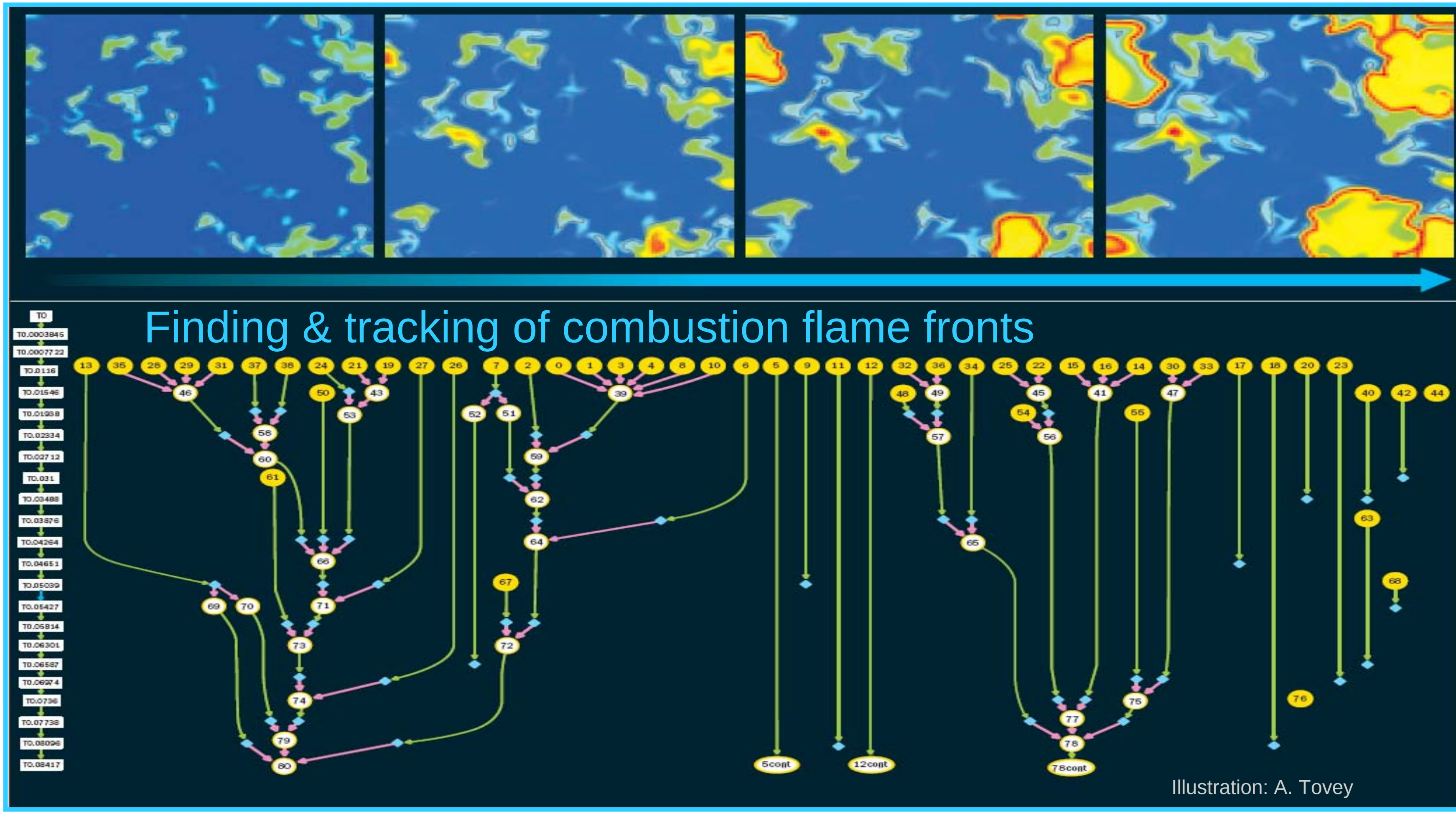


Time required to identify regions in 3D Supernova simulation (LBNL)

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

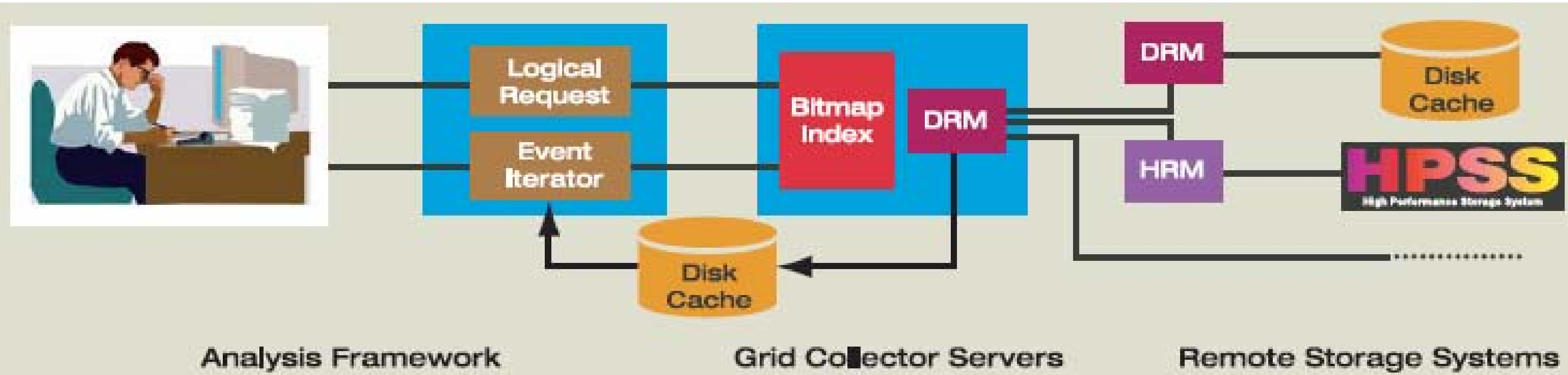
Searching for regions that satisfy particular criteria with FastBit

Plotting progression of flames can be performed in real time



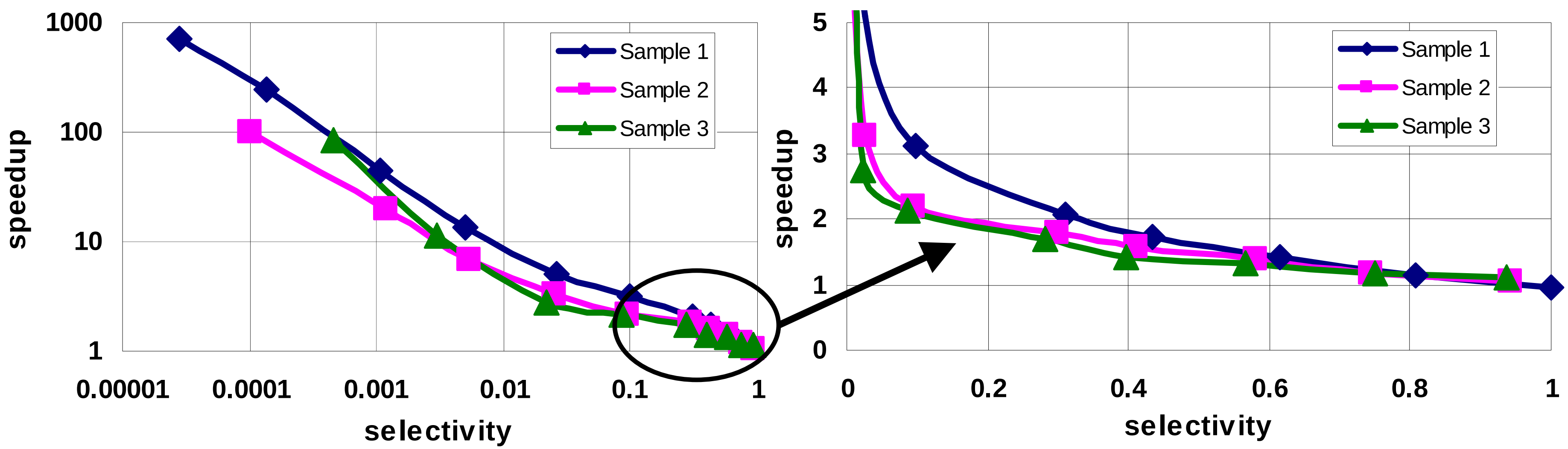
Ollaborators: SNL: Drs. J. Chen, W. Doyle, NCSU: Dr. T. Echekki

Grid Collector



- Searching over billions of objects with hundreds of descriptors each:
  - search over descriptors of billions of HEP (collision) events
  - Data volume Terabytes-Petabytes
- Benefits of the Grid Collector:
  - transparent object access
  - Selection of objects based on their attribute values
  - Improvement of analysis system’s throughput
  - Interactive analysis of data distributed on the Grid
  - Best Paper Award (ISC’05)

Grid Collector Speeds up Analyses



- Test machine: 2.8 GHz Xeon, 27 MB/s read speed
- When searching for rare events, say, selecting one event out of 1000, using GC is <u>20 to 50</u> times faster
- Using GC to read 1/2 of events, speedup > 1.5, 1/10 events, speed up > 2.

Grid Collector Success Stories

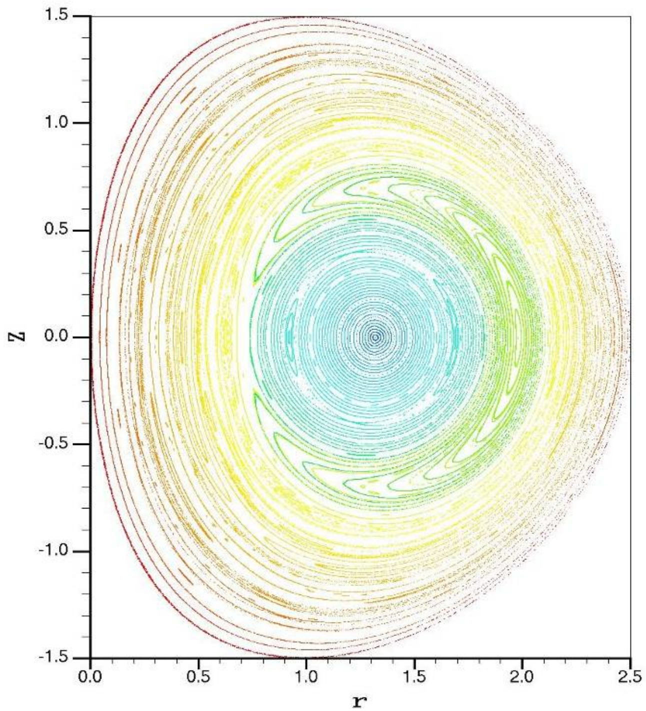
- Searching for anti-^{3}He
- Lee Barnby, Birmingham
- Previous studies identified collision events that possibly contain anti-^{3}He, need further analysis

- Searching for strangelets
- Aihong Tang, BNL
- Previous studies identified events that behave close to strangelets, need further investigation

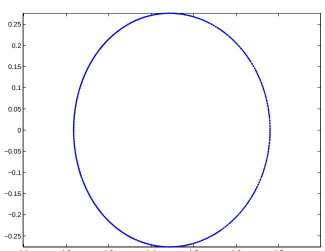
- Without Grid Collector, one has to retrieve every file from mass storage systems and scan them for the wanted events – may take weeks or months
- With Grid Collector, both jobs completed within a day

Piecewise Polynomial Models for Classification of Puncture (Poincaré) plots

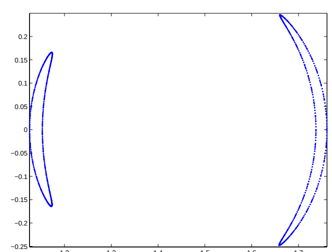
- Classify each of the nodes: quasiperiodic, islands, separatrix
- Connections between the nodes
- Want accurate and robust classification, valid when few points in each node



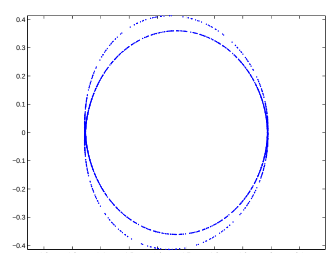
National Compact Stellarator Experiment



Quasiperiodic



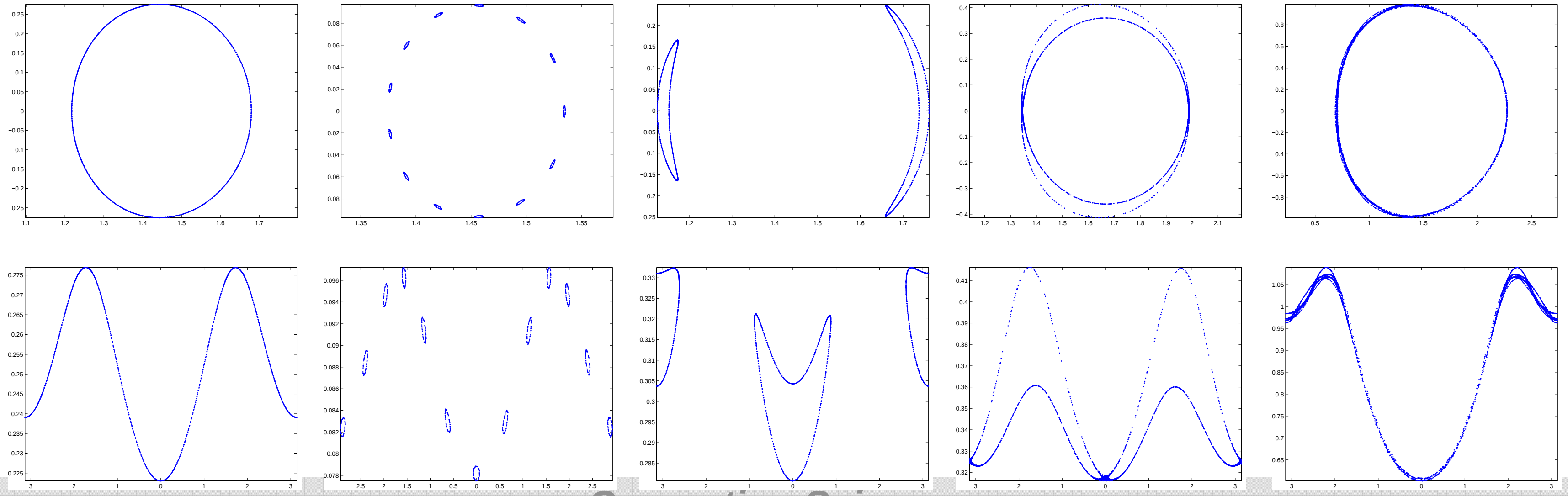
Islands



Separatrix

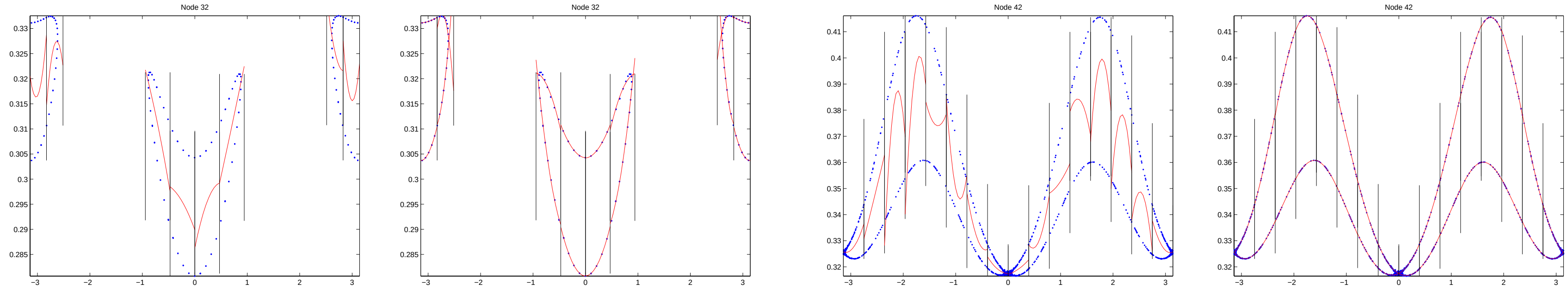
Polar Coordinates

- Transform the (x,y) data to Polar coordinates (r,θ).
- Advantages of polar coordinates:
  - Radial exaggeration reveals some features that are hard to see otherwise.
  - Automatically restricts analysis to radial band with data, ignoring inside and outside.
  - Easy to handle rotational invariance.



Piecewise Polynomial Fitting: Computing polynomials

- In each interval, compute the polynomial coefficients to fit 1 polynomial to the data.
- If the error is high, split the data into an upper and lower group. Fit 2 polynomials to the data, one to each group.

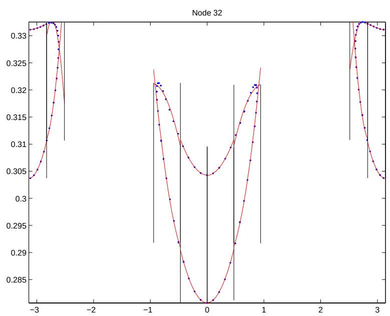


Blue: data.   Red: polynomials.   Black: interval boundaries.

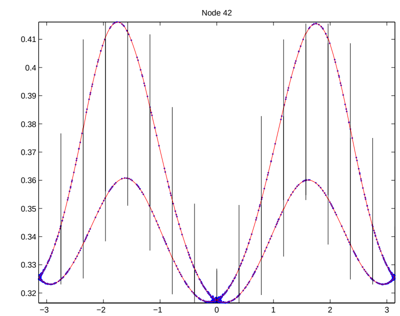
Classification

- The number of polynomials needed to fit the data and the number of gaps gives the information needed to classify the node:

	Number of polynomials	
Gaps	one	two
Zero	Quasiperiodic	Separatrix
> Zero		Islands

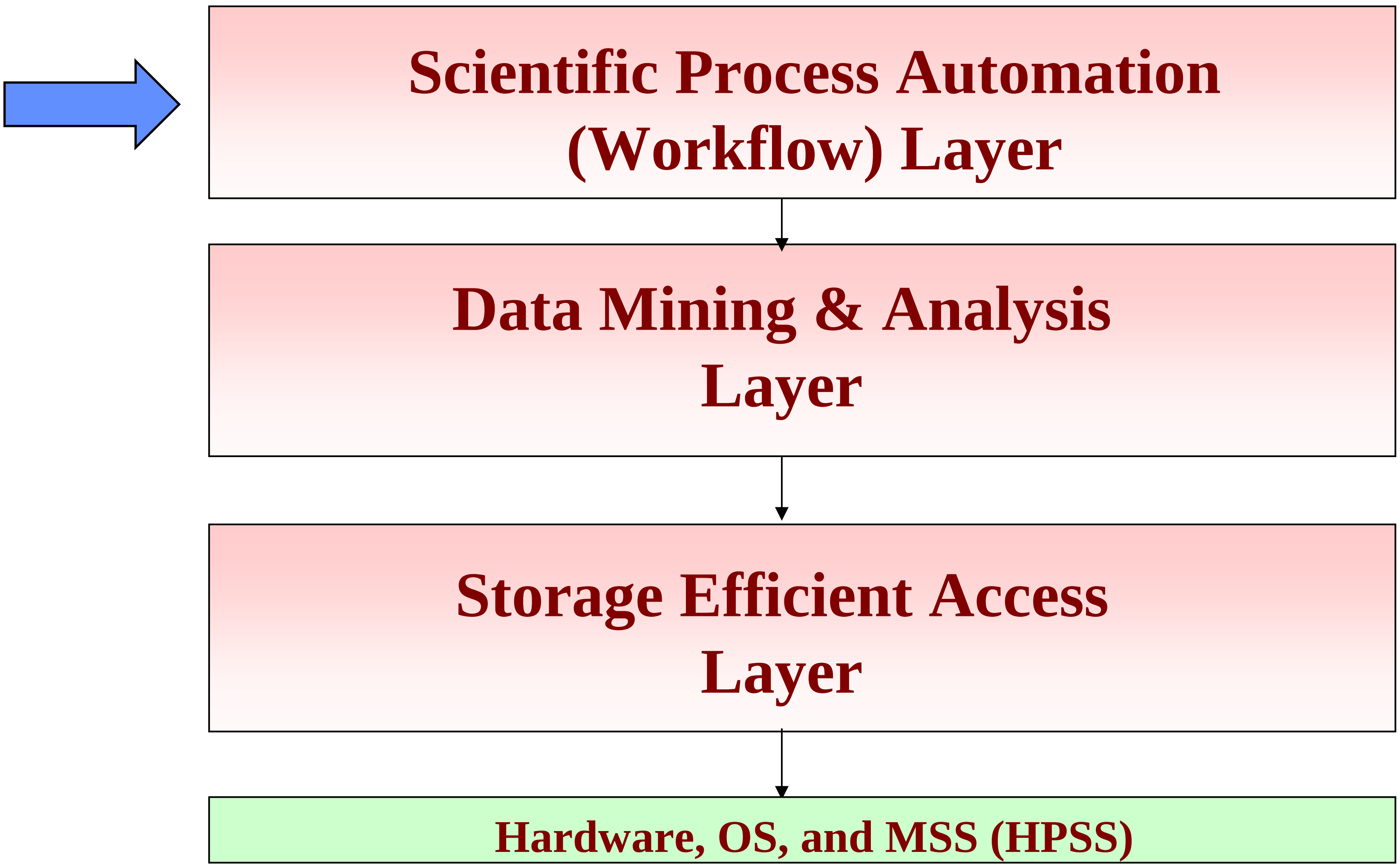


2 Polynomials
2 Gaps
→ Islands

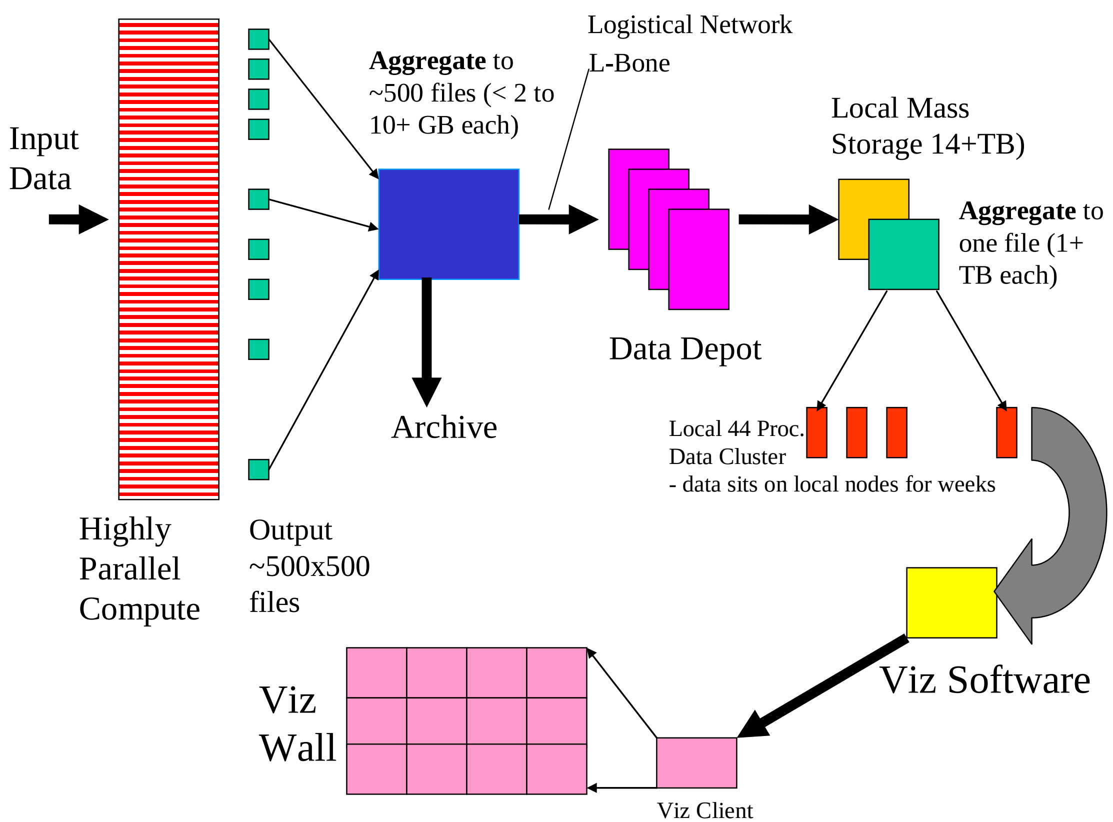


2 Polynomials
0 Gaps
→ Separatrix

Layering of Components



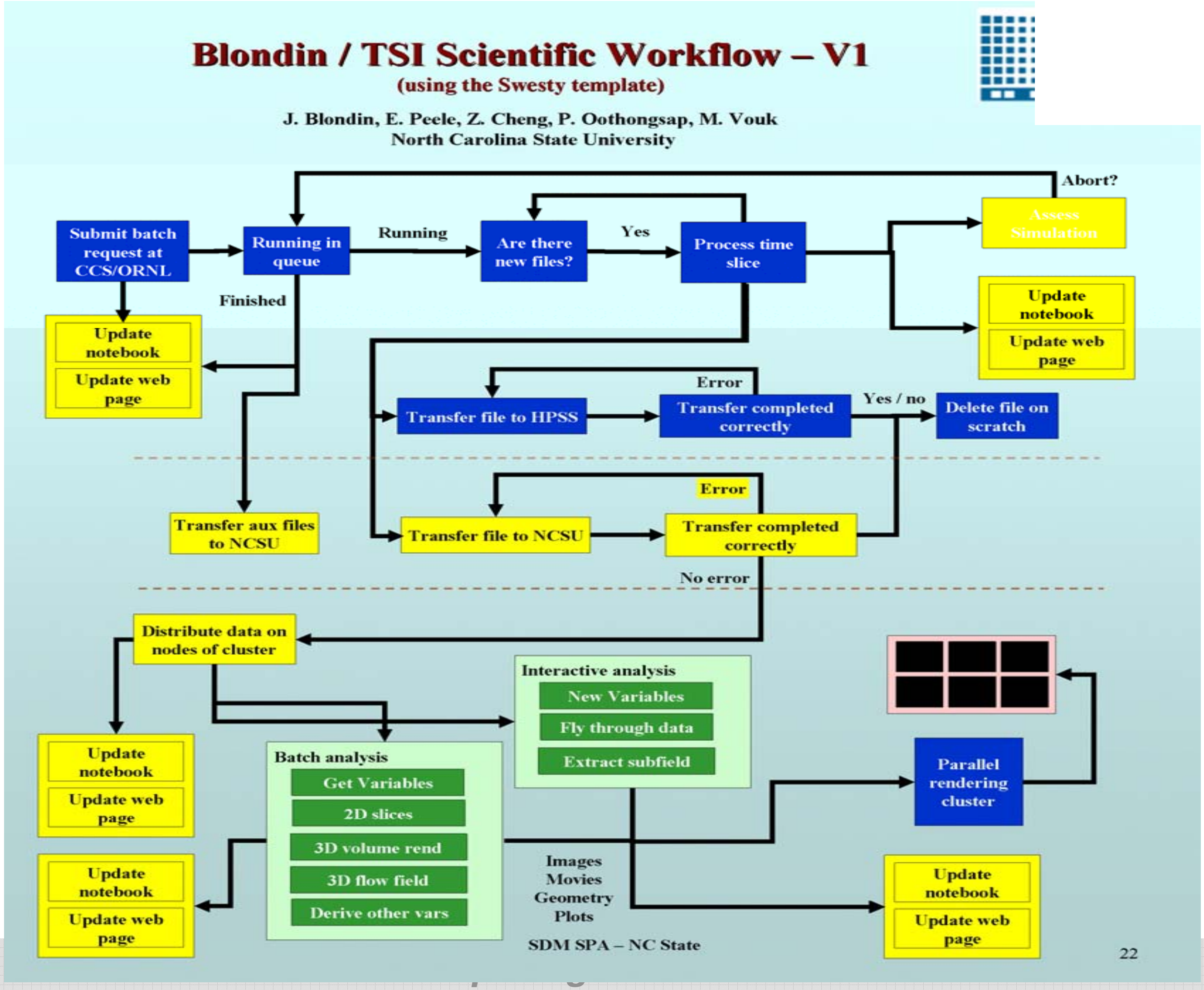
Example Data Flow in TSI



Courtesy: John Blondin

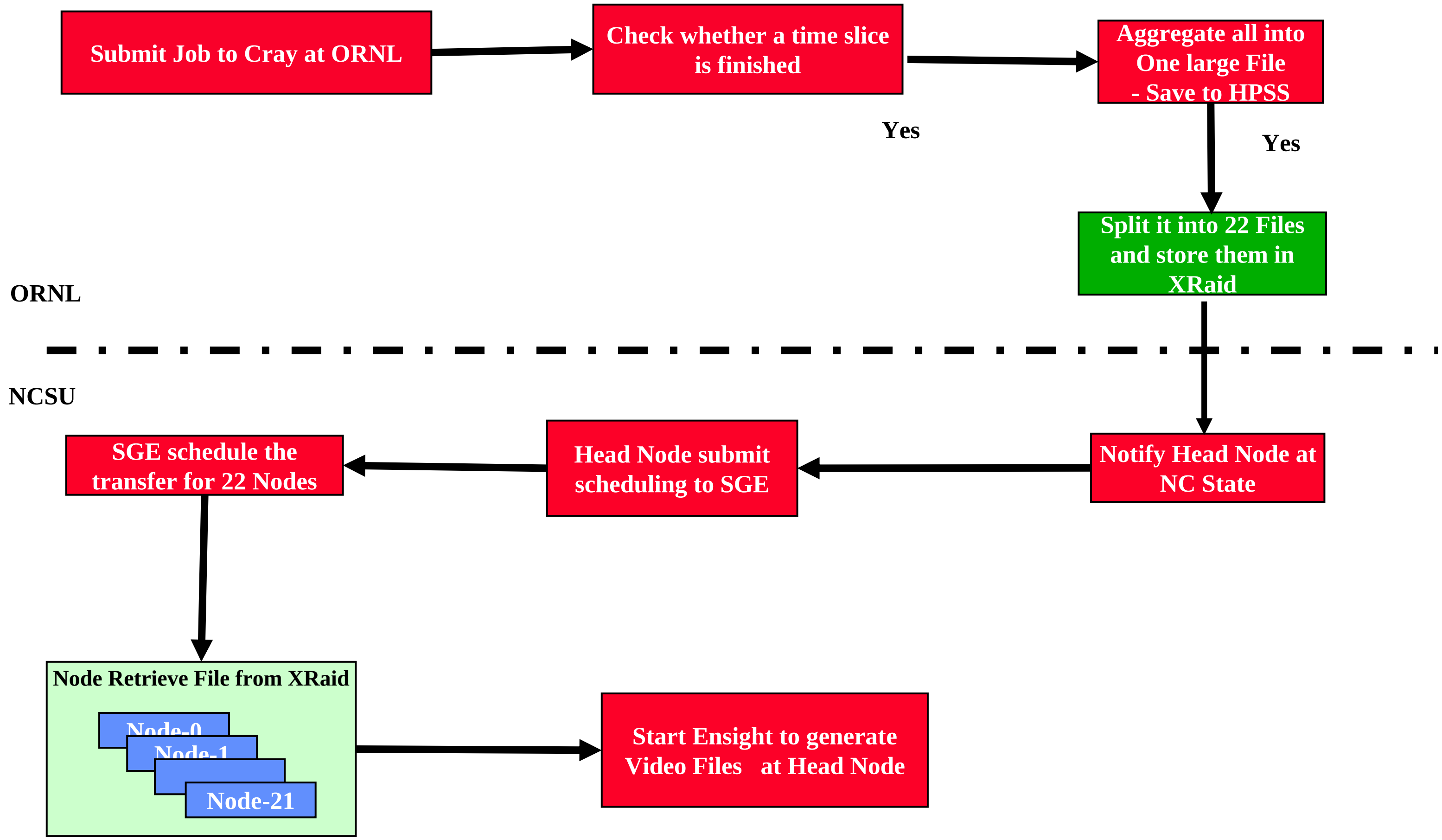
Original TSI Workflow Example with John Blondin, NCSU

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



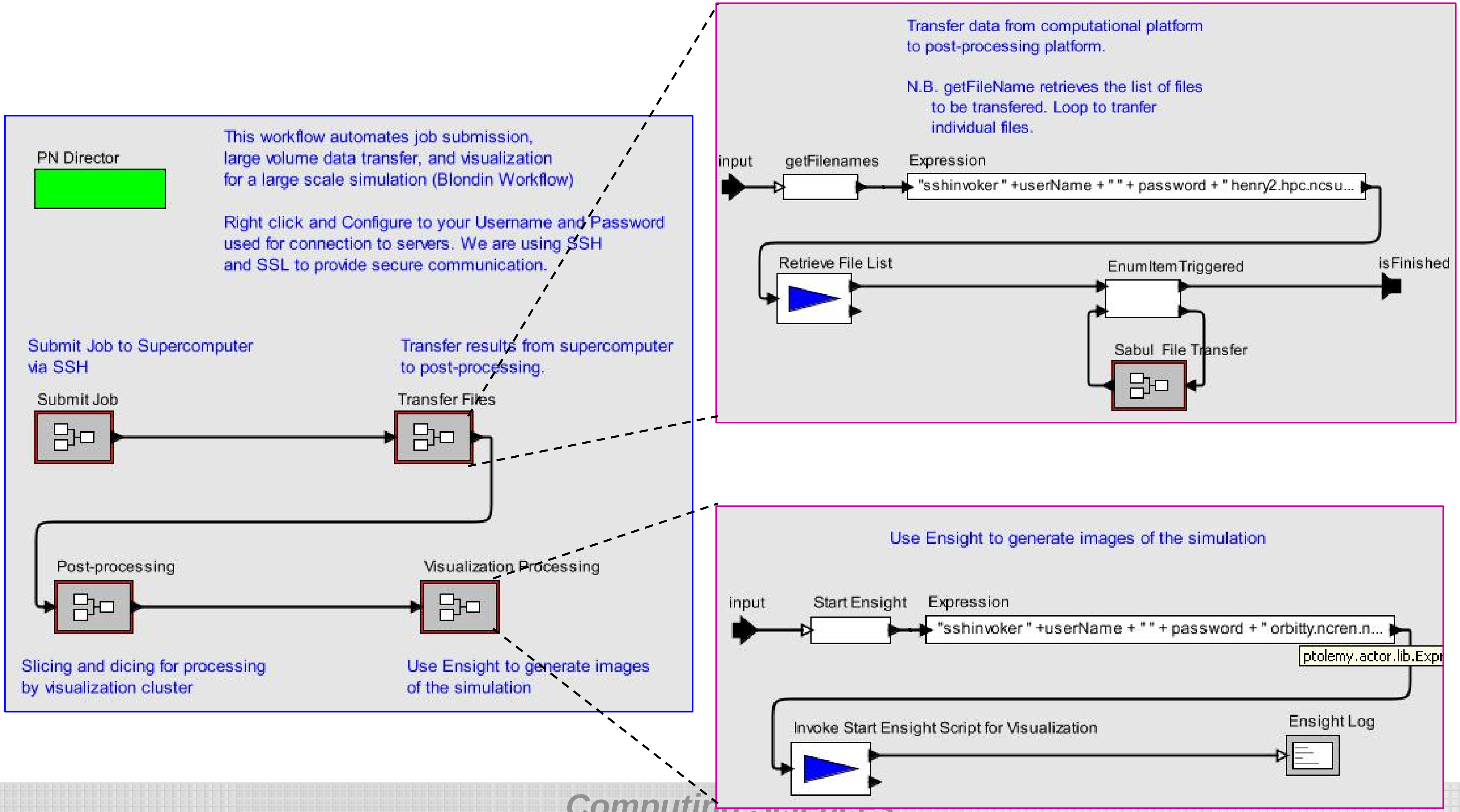
Top level TSI Workflow

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

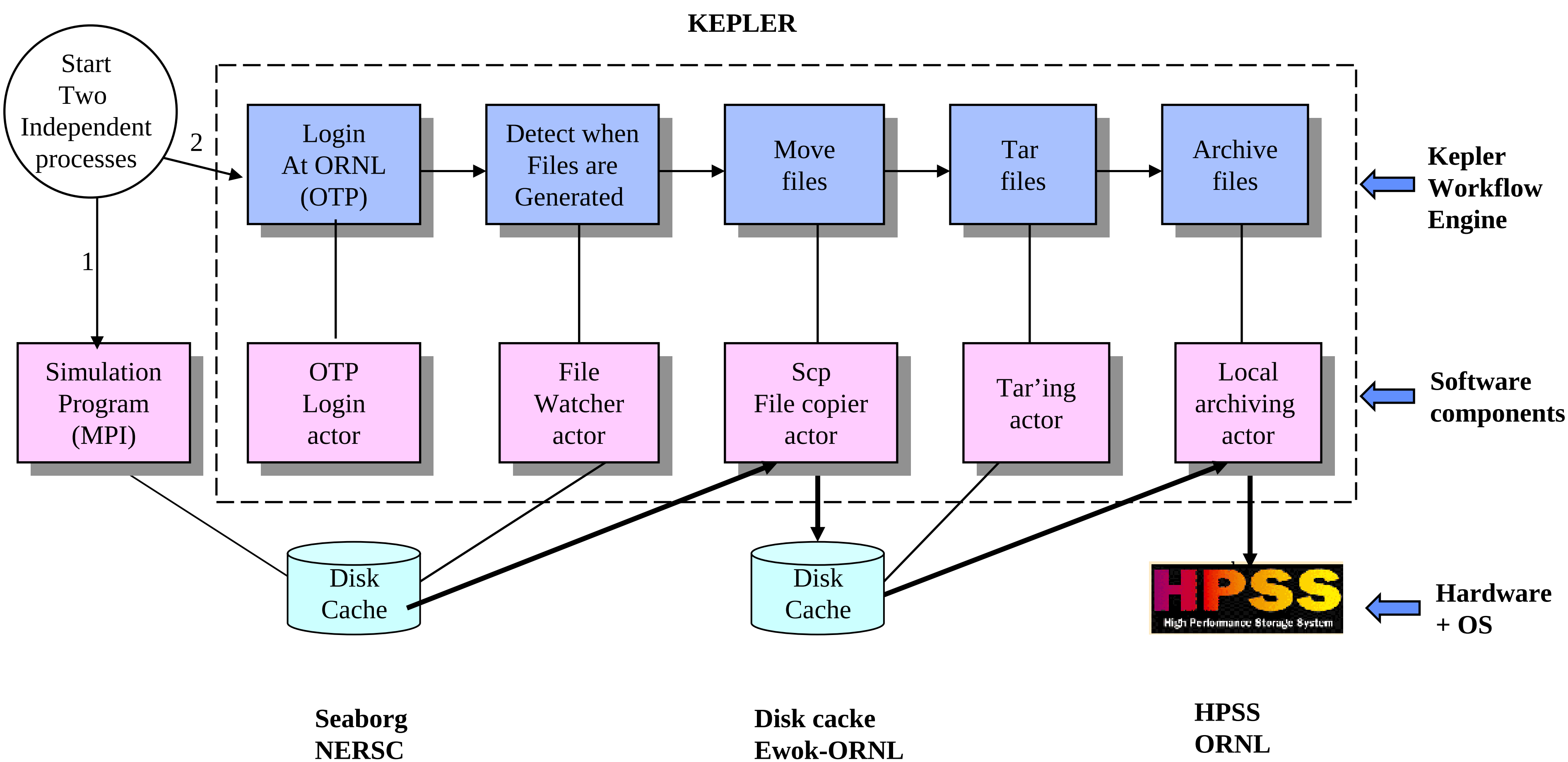


Using the Scientific Workflow Tool (Kepler) Emphasizing Dataflow (SDSC, NCSU, LLNL)

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



New actors in CPES workflow to overcome problems



Re-applying Technology

SDM technology, developed for one application, can be effectively targeted at many other applications …

Technology	Initial Application	New Applications
Parallel NetCDF	Astrophysics	Climate
Parallel VTK	Astrophysics	Climate
Compressed bitmaps	HENP	Combustion, Astrophysics
Storage Resource Managers	HENP	Astrophysics
Feature Selection	Climate	Fusion (exp. & simulation)
Scientific Workflow	Biology	Astrophysics

Broad Impact of the SDM Center…

Astrophysics:

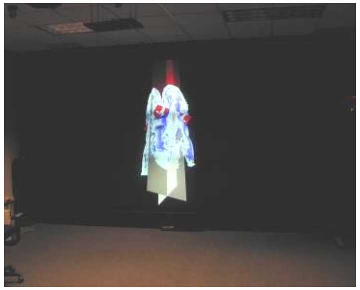
High speed storage technology, parallel NetCDF, parallel VTK, and ASPECT integration software used for Terascale Supernova Initiative (TSI) and FLASH simulations

Tony Mezzacappa – ORNL, Mike Zingale – U of Chicago, Mike Papka – ANL

Scientific Workflow

John Blondin – NCSU

Doug Swesty, Eric Myra – Stony Brook

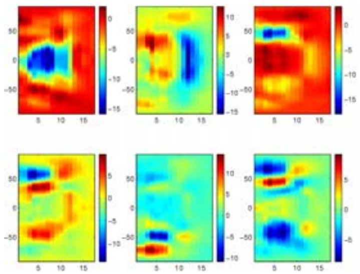


ASCI FLASH – parallel NetCDF

Climate:

High speed storage technology, Parallel NetCDF, and ICA technology used for Climate Modeling projects

Ben Santer – LLNL, John Drake – ORNL, John Michalakes – NCAR

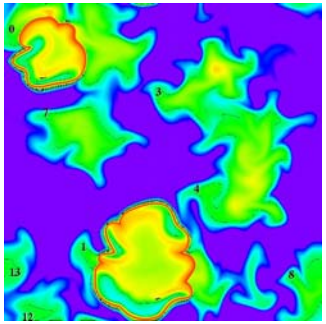


Dimensionality reduction

Combustion:

Compressed Bitmap Indexing used for fast generation of flame regions and tracking their progress over time

Wendy Koegler, Jacqueline Chen – Sandia Lab



Region growing

Broad Impact (cont.)*

Biology:

Kepler workflow system and web-wrapping technology used for executing complex highly repetitive workflow tasks for processing microarray data

Matt Coleman – LLNL

Doug Swesty – U of Brookhaven

High Energy Physics:

Compressed Bitmap Indexing and Storage Resource Managers used for locating desired subsets of data (events) and automatically retrieving data from HPSS

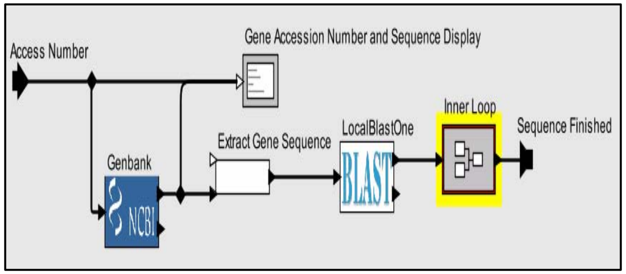
Doug Olson - LBNL, Eric Hjort – LBNL, Jerome Lauret - BNL

Fusion:

A combination of PCA and ICA technology used to identify the key parameters that are relevant to the presence of edge harmonic oscillations in a Tokomak

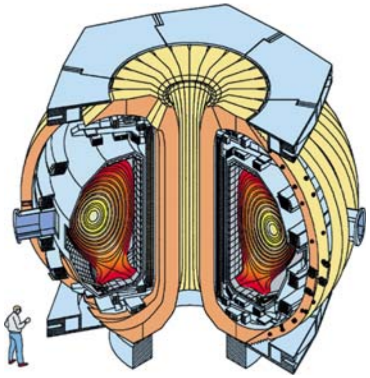
Keith Burrell - General Atomics

Scott Klasky - PPPL



Building a scientific workflow

Dynamic monitoring of HPSS file transfers



Identifying key parameters for the DIII-D Tokamak

Summary

- Scientific Data Management poses many challenges
  - Large volume
    - Indexing, transparent file access, storage management, …
  - Complexity of data model
    - Mesh data, spatio-temporal, sequence data, …
  - Distributed data and file management
    - Keep track of data location and replication, space management, …
  - Metadata
    - automatic capture, data entry, provenance (history)
- Need to support various activities
  - Research, prototype, development, deployment, scaling, maintenance
- The Data Management Group
  - Focuses research program on problems driven by science applications
  - Successful research prototypes hardened and deployed though collaborative projects, including the SDM center
  - Support for deployed software is an important part of our activities
  - Productivity: papers publication, software development and deployment and support, international service activities (standards, conferences, reviews)

Data Management Requirements for Scientific Applications

Fundamental Technology Areas

- From the report from the DOE Office of Science Data-Management Workshops (March – May 2004)

  - Efficient access and queries, data integration
  - Distributed data management, data movement, networks
  - Storage and caching
  - Data analysis, visualization, and integrated environments
  - Metadata, data description, logical organization
  - Workflow, data flow, data transformation

Scientific Workflow Automation

Problem: Unique requirements of scientific WFs

- **Moving large volumes between modules**
  - **Tightly-coupled efficient data movement**
  - **Wide-area loosely coupled data movement**
- **Specification of granularity-based iteration**
  - **e.g. In spatial-temporal simulations – a time step is a "granule"**
- **Support for data transformation**
  - **complex data types (including file formats, e.g. netCDF, HDF)**
- **Dynamic control of workflow by user**
  - **Dynamic user examination of results**

Storage Resource Management

Alex Sim
Junmin Gu
Vijaya Natarajan
Alex Romosan
(Arie Shoshani: Design)

What is SRM?

- Storage Resource Managers (SRM) are middleware components whose function is to provide
  - Dynamic space allocation
  - Dynamic file management in space
- For shared storage components on the WAN

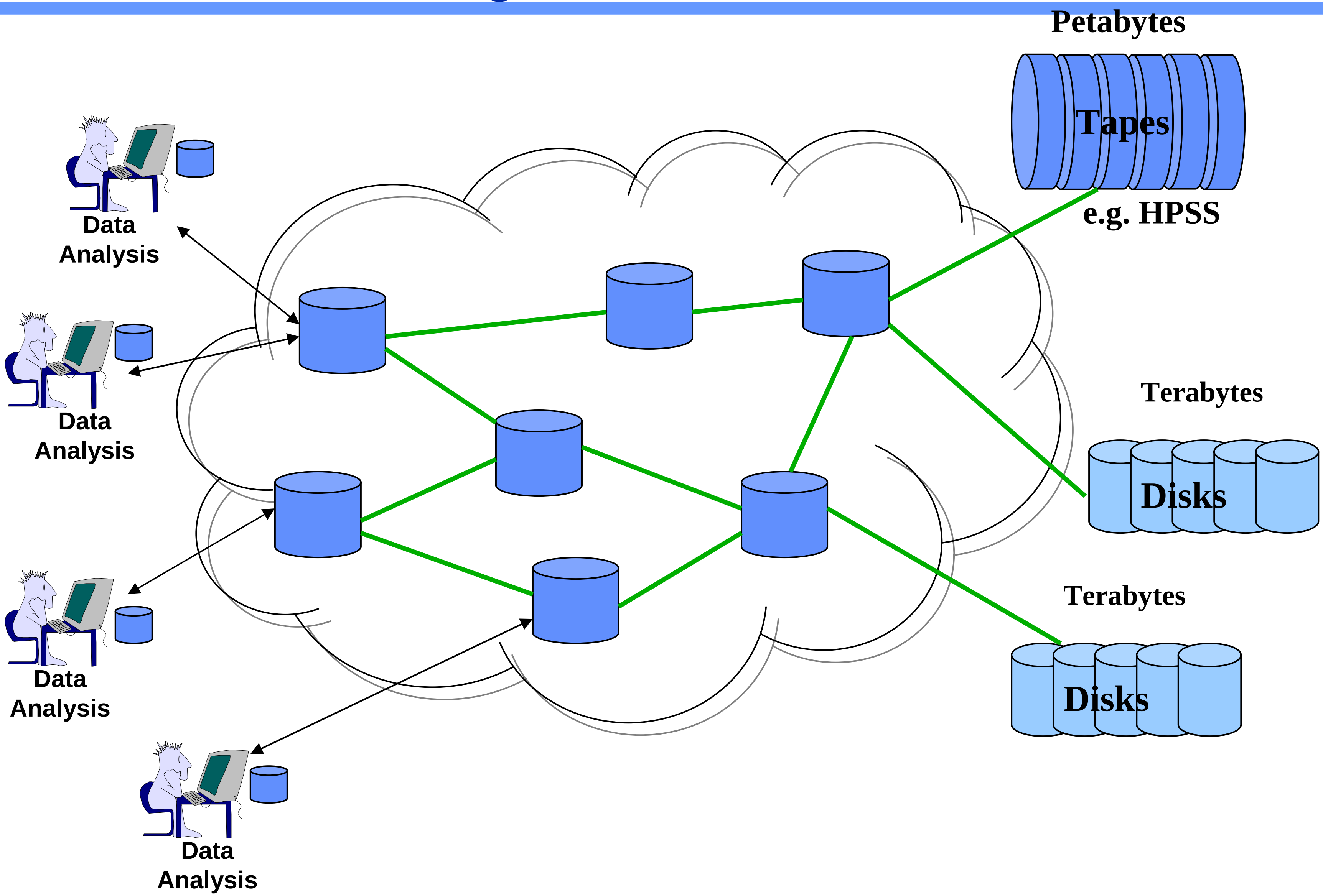
Motivation

- Suppose you want to run a job on your local machine
  - Need to allocate space
  - Need to bring all input files
  - Need to ensure correctness of files transferred
  - Need to monitor and recover from errors
  - What if files don't fit space? Need to manage file streaming
  - Need to remove files to make space for more files
- Now, suppose that the machine and storage space is a shared resource
  - Need to do the above for many users
  - Need to enforce quotas
  - Need to ensure fairness of scheduling users

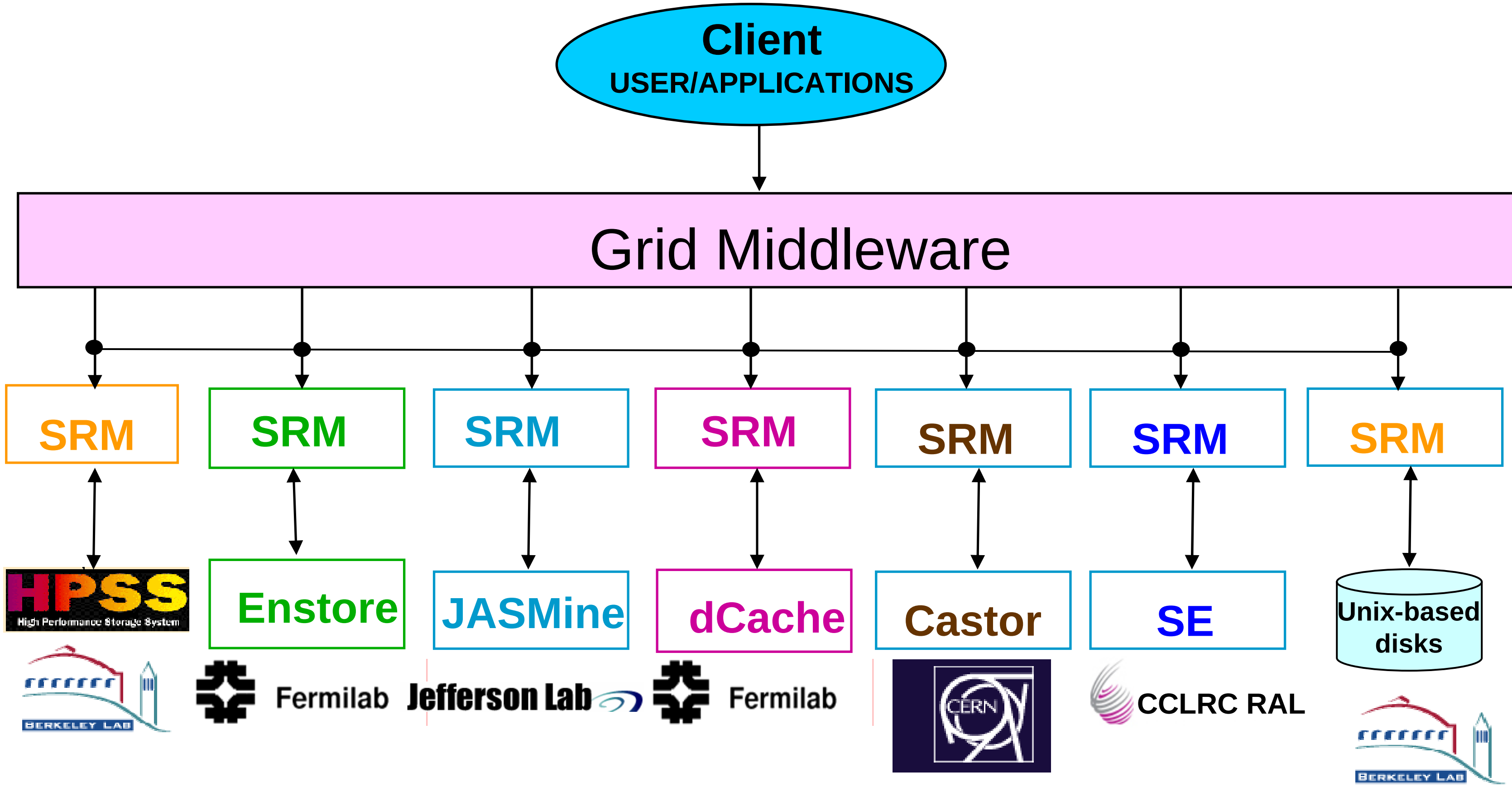
Motivation

- Now, suppose you want to do that on a WAN
  - Need to access a variety of storage systems
  - mostly remote systems, need to have access permission
  - Need to have special software to access mass storage systems
- Now, suppose you want to run distributed jobs on the WAN
  - Need to allocate remote spaces
  - Need to move (stream) files to remote sites
  - Need to manage file outputs and their movement to destination site(s)

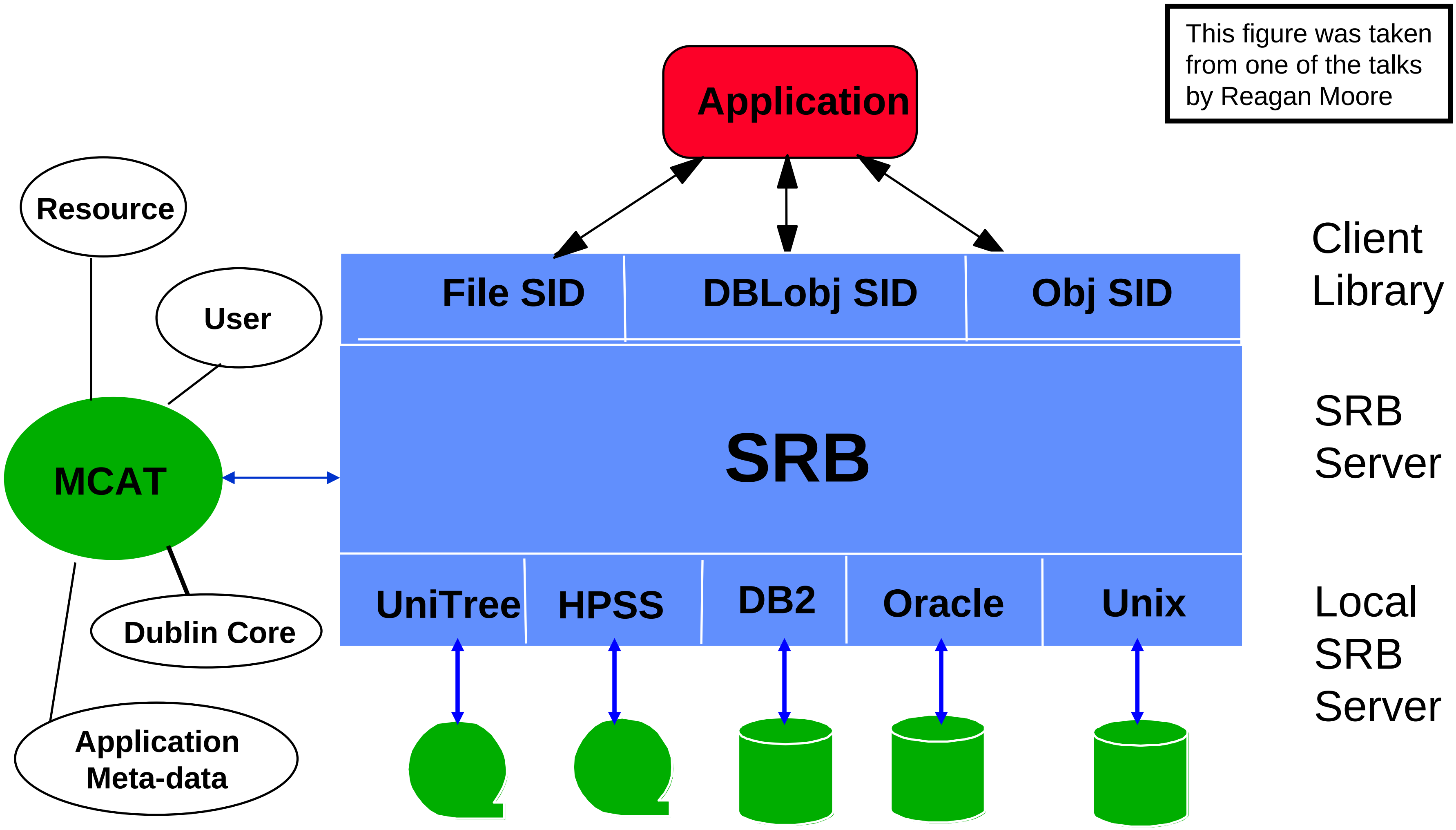
Ubiquitous and Transparent Data Access and Data Sharing



Interoperability of SRMs



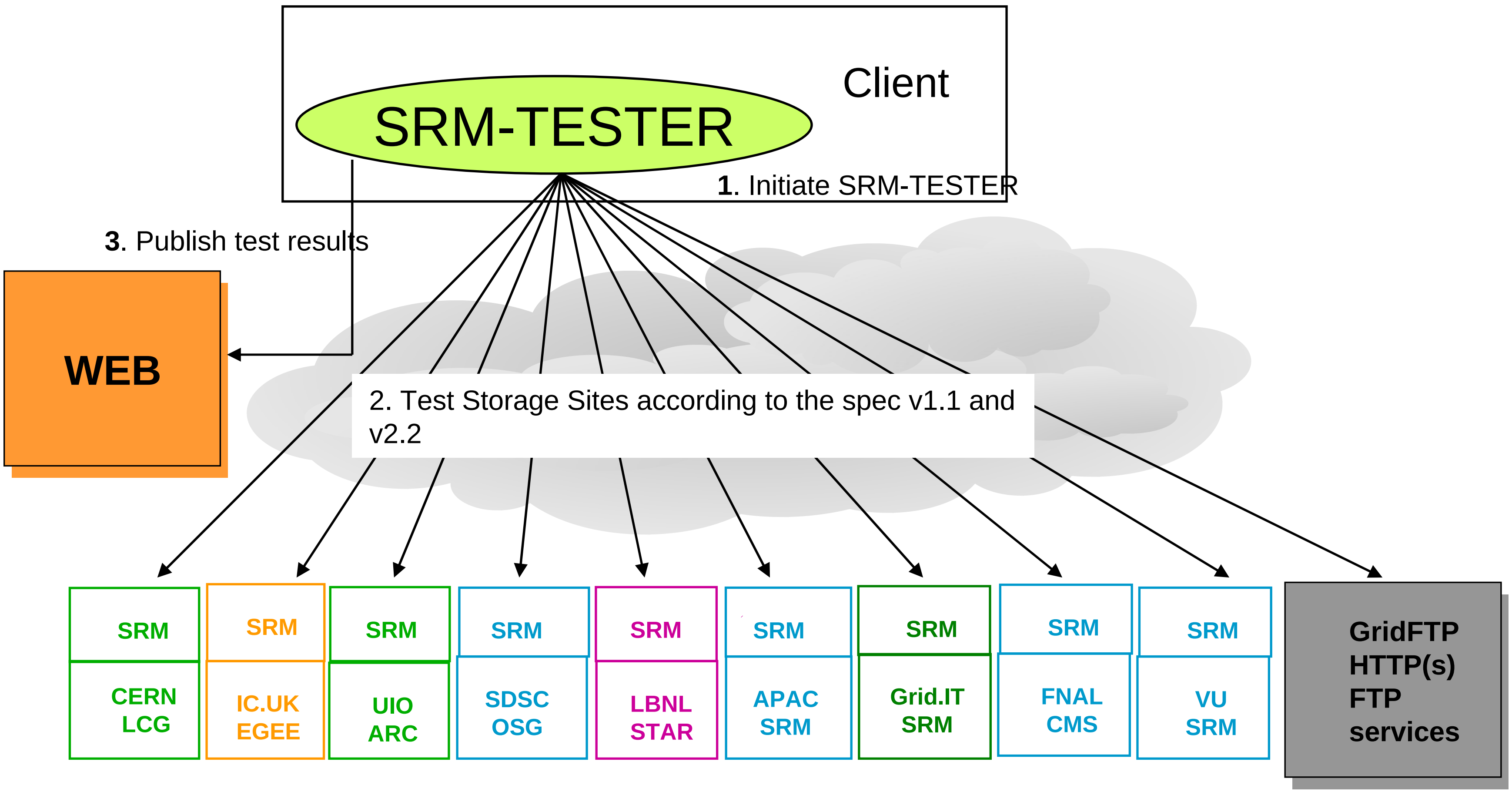
SDSC Storage Resource Broker – Grid Middleware



SRM vs. SRB

- Storage Resource Broker (SRB)
  - Very successful product from SDSC, has long history
  - Is a centralized solution where all requests go to a central server that includes a metadata catalog (MCAT)
  - Developed by a single institution
- Storage Resource Management (SRM)
  - Based on open standard
  - Developed by multiple institutions for their storage systems
  - Designed for interoperation of heterogeneous storage systems
- Features of SRM that SRB does not deal with
  - Managing storage space *dynamically* based client's request
  - Managing content of space based on *lifetime* controlled by client
  - Support for file streaming by *pinning and releasing* files
- Features that SRB that SRM does not deal with
  - Logical organization of file *collections*
  - *Metadata* of files and collections (usage attributes, lineage, etc.)
  - *Mapping* logical file namespace into physical locations of the files.
- Several projects now ask for an SRM interface to SRB and vice versa
  - In OGF: activity to bridge these technologies

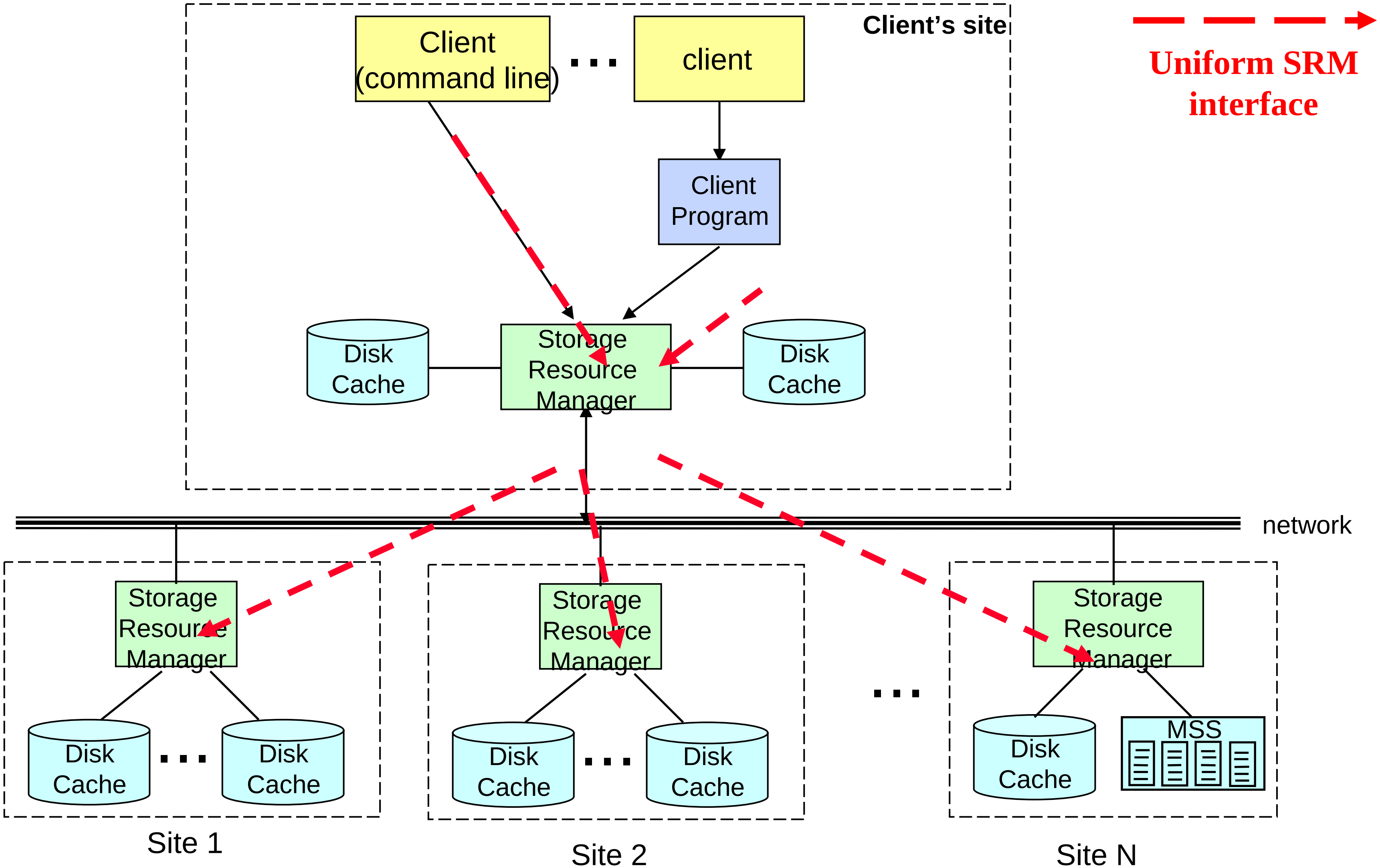
GGF GIN-Data SRM inter-op testing
(GGF: Global Grid Forum, GIN: Grid Interoperability Now)



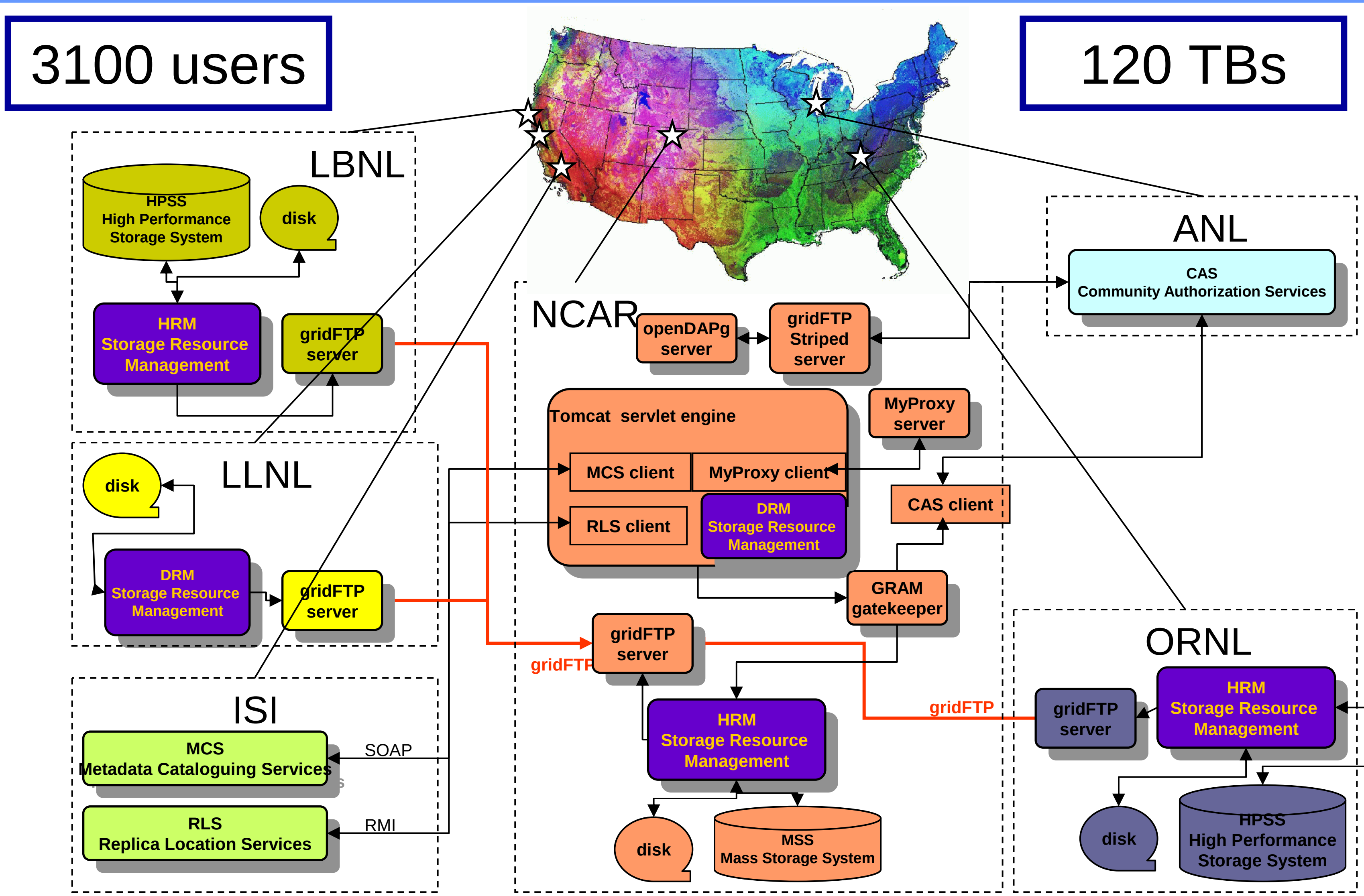
Testing Operations Results

	ping	put	get	Advisory delete	Copy (SRMs)	Copy (gsiftp)
ARC (UIO.NO)	**pass**	fail	**pass**	fail	pass	fail
EGEE (IC.UK)	**pass**	**pass**	**pass**	**pass**	**pass**	**pass**
CMS (FNAL.GOV)	**pass**	**pass**	**pass**	**pass**	**pass**	**pass**
LCG/EGEE (CERN)	**pass**	**pass**	**pass**	**pass**	N.A.	N.A.
OSG (SDSC)	**pass**	**pass**	**pass**	**pass**	pass	fail
STAR (LBNL)	**pass**	**pass**	**pass**	**pass**	**pass**	**pass**

Peer-to-Peer Uniform Interface



Earth Science Grid Analysis Environment



History and Partners in SRM Collaboration

- 5 year of Storage Resource (SRM) Management activity
- Experience with SRM system implementations
  - Mass Storage Systems:
    - HPSS (*LBNL, ORNL, BNL, all using LBNL-SRM*),
    - MSS (NCAR - *Using LBNL-SRM adaptation),*
    - Enstore (Fermi), JasMINE (Jlab), Castor (CERN), Castor (RAL) …
  - Disk systems:
    - *DRM(LBNL)*, jSRM (Jlab), DPM (CERN), universities …
  - Combination systems:
    - dCache(Fermi) – sophisticated multi-storage system
    - L-Store (U Vanderbilt) based on Logistical Networking (*Using LBNL-SRM adaptation*)
    - StoRM – to parallel file systems (ICTP, Trieste, Italy)

Standards for Storage Resource Management

- Main concepts
  - Allocate spaces
  - Get/put files from/into spaces
  - Pin files for a lifetime
  - Release files and spaces
  - Get files into spaces from remote sites
  - Manage directory structures in spaces
  - SRMs communicate as peer-to-peer
  - Negotiate transfer protocols

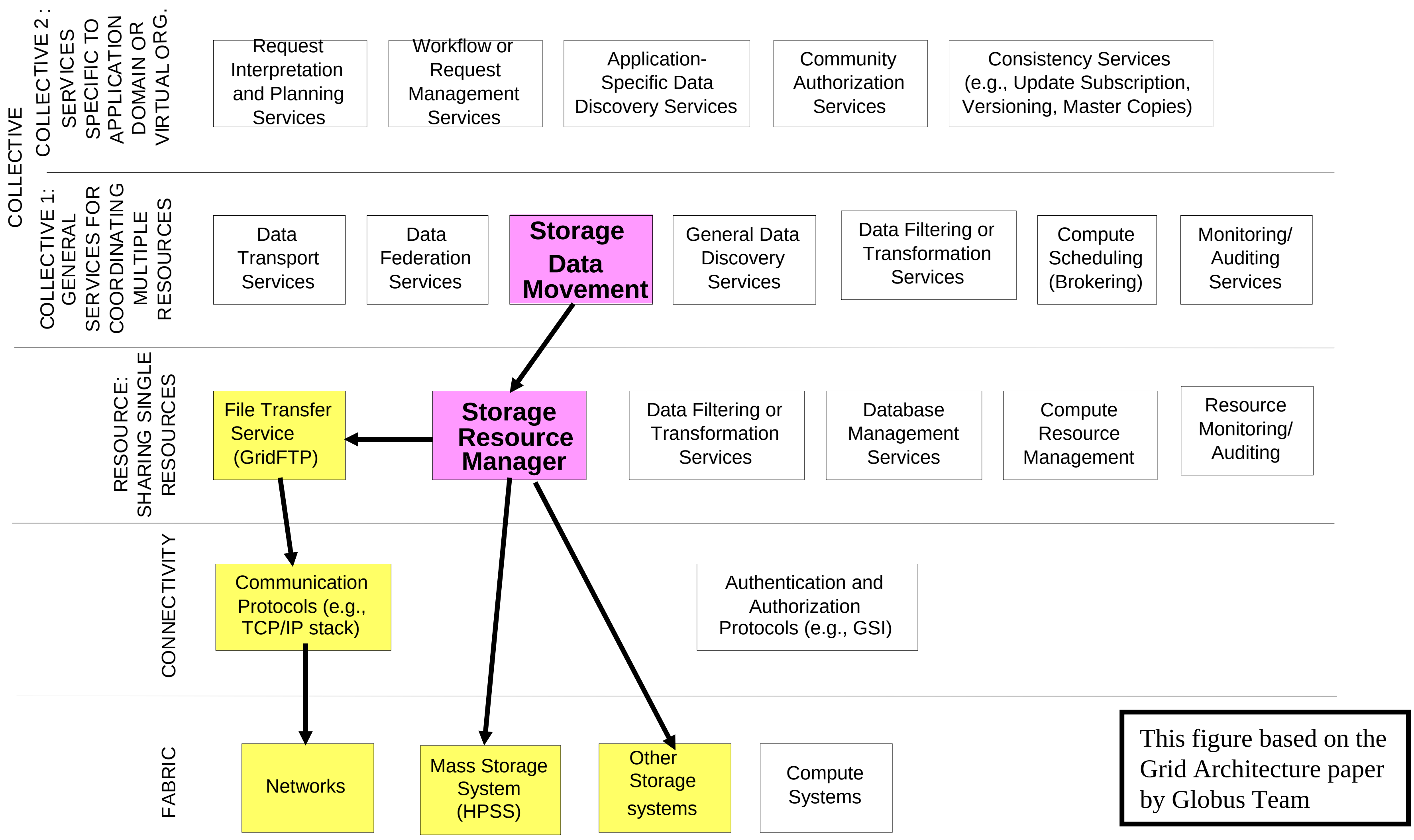
Results

- LBNL leading the ***SRM collaboration***
  - a standard common interface was developed and is evolving
  - Institutions involved: LBNL, FNAL,JTNAF, BNL (US), CERN, RAL (Europe)
  - This standard was now adapted by the international WLCG collaboration.
- SRMs have been used ***in production***
  - several facilities including BNL, NERSC, FNAL, CERN, TJNAF, ORNL and NCAR, and other facilities in Europe and Asia.
- LBNL's SRMs have been used in production over the last few years to support ***intensive robust data movement***
  - between BNL to NERSC for the STAR experiment
  - about 10,000 files (about 1 TB) per week in an automated fashion
- LBNL's SRMs have been used in production in the ***Earth Systems Grid*** (ESG) Project
  - provide transparent access from multiple remote storage systems at NERSC, NCAR, ORNL, LLNL, and LANL, including HPSS and NCAR-MSS.
- LBNL's SRMs have also been used for projects in the ***Scientific Data Management (SDM) Center***
  - *GridCollector*
  - With MPI-IO to transparently access files on remote storage systems

DataMover

Perform “rcp –r directory” on the WAN

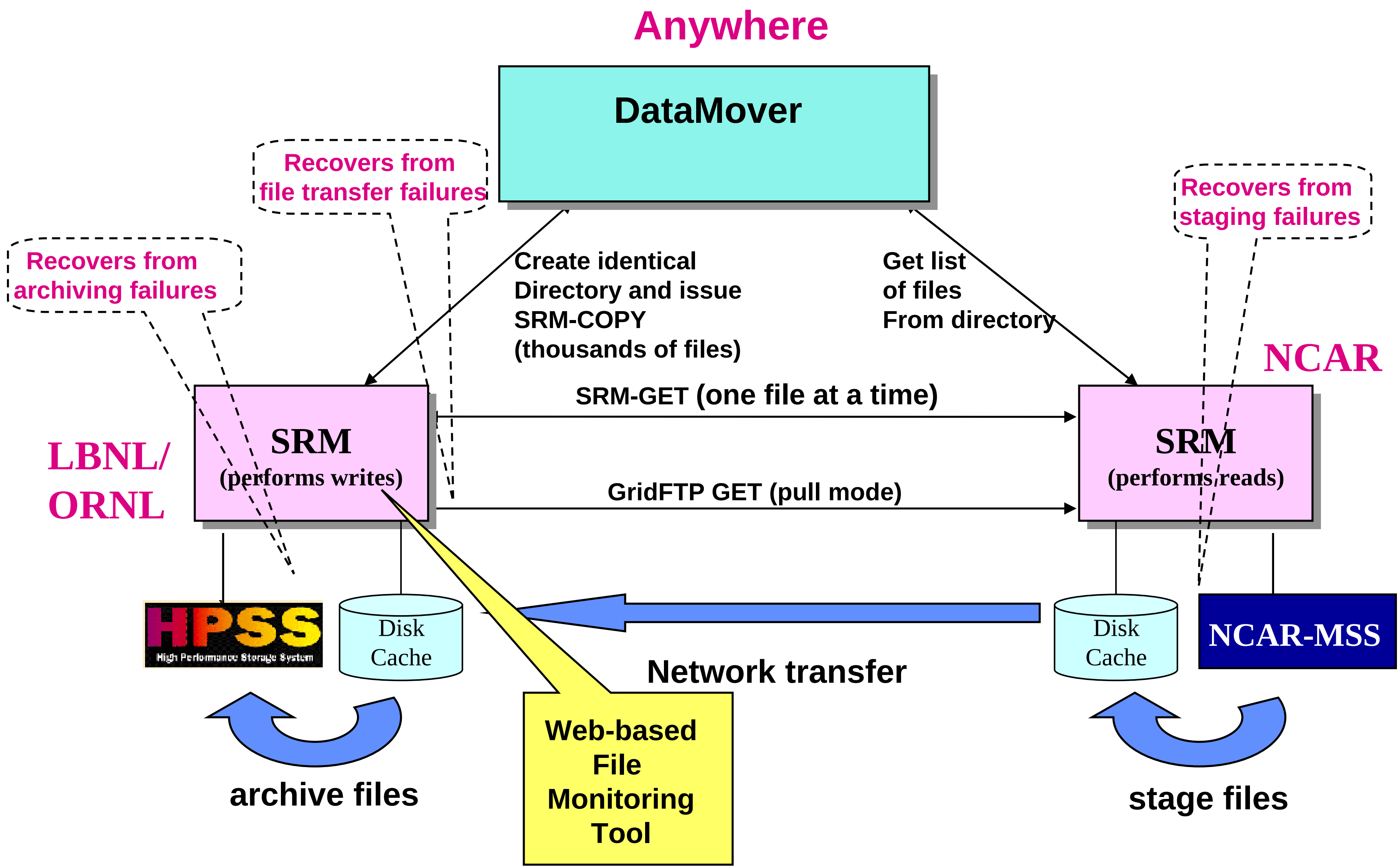
SRMs Supports Data Movement Between Storage Systems



Massive Robust File Replication

- Multi-File Replication – why is it a problem?
  - Tedious task – many files, repetitious
  - Lengthy task – long time, can take hours, even days
  - Error prone – need to monitor transfers
  - Error recovery – need to restart file transfers
  - Stage and archive from MSS – limited concurrency, down time, transient failures
  - Commercial MSS – HPSS at NERSC, ORNL, …,
  - Legacy MSS – MSS at NCAR
  - Independent MSS – Castor (CERN), Enstore (Fermilab), JasMINE (Jlab)

DataMover: SRMs use in ESG for Robust Multi-file replication



Web-Based File Monitoring Tool

Shows:

-Files already transferred

- Files during transfer

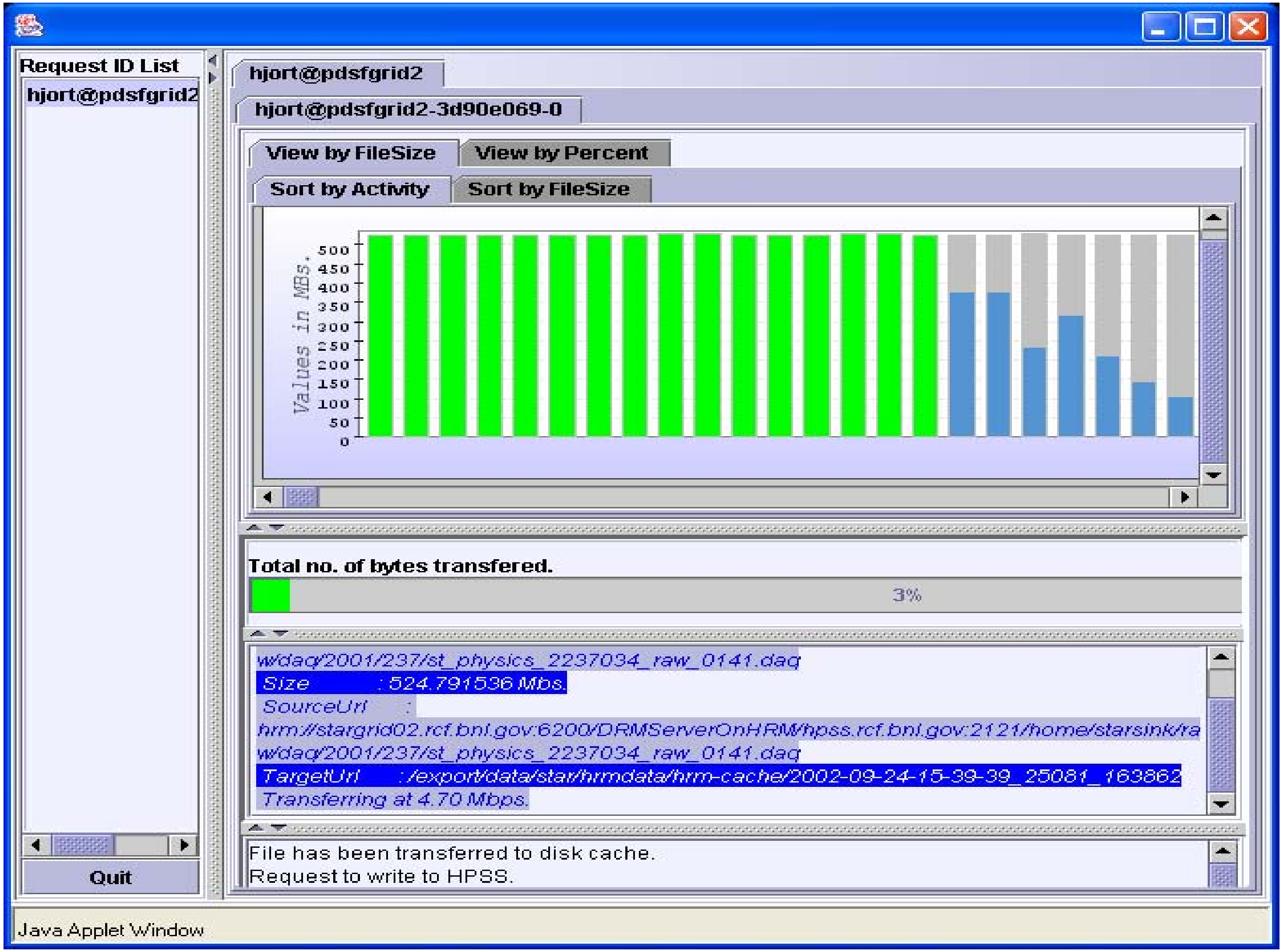
- Files to be transferred

Also shows for each file:

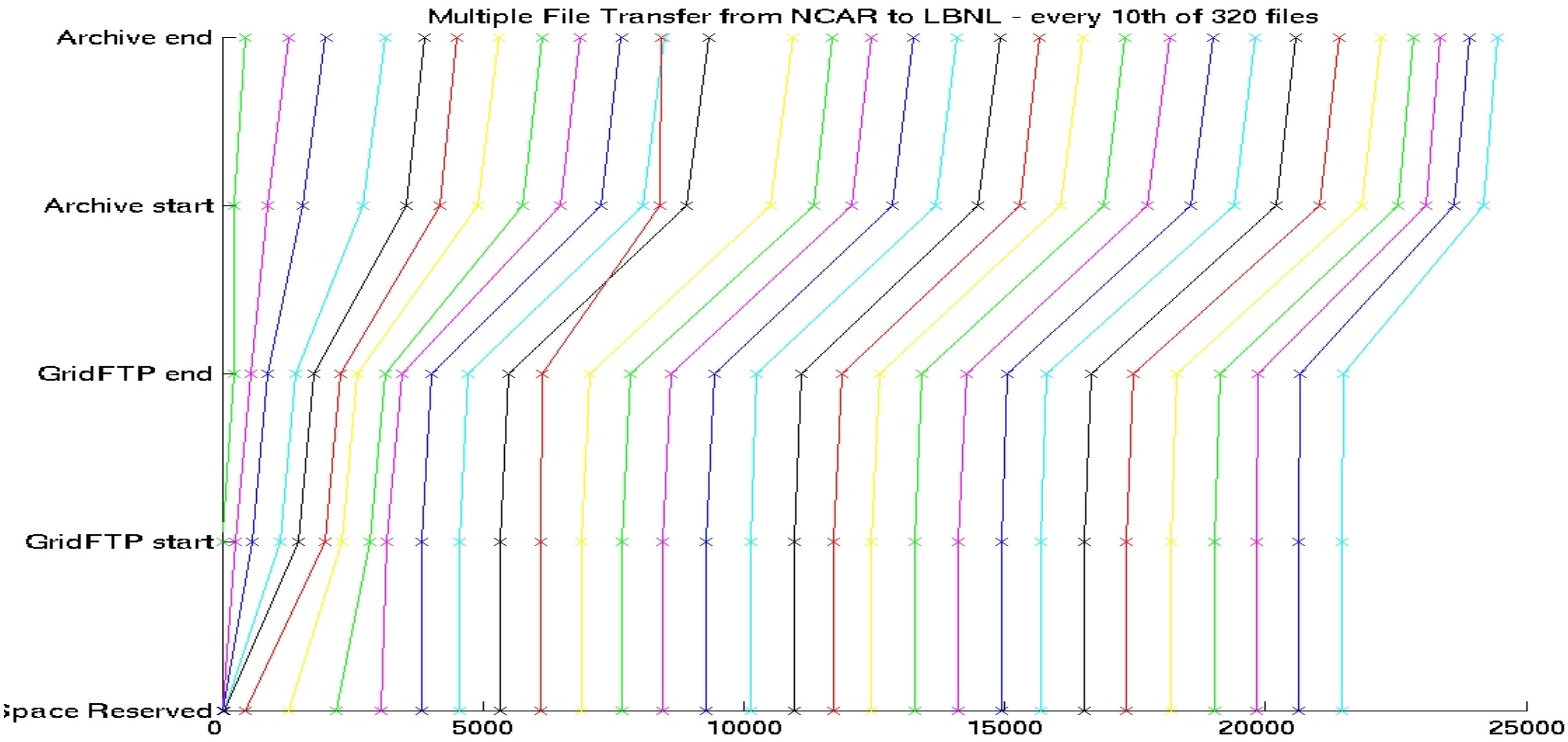
-Source URL

-Target URL

-Transfer rate

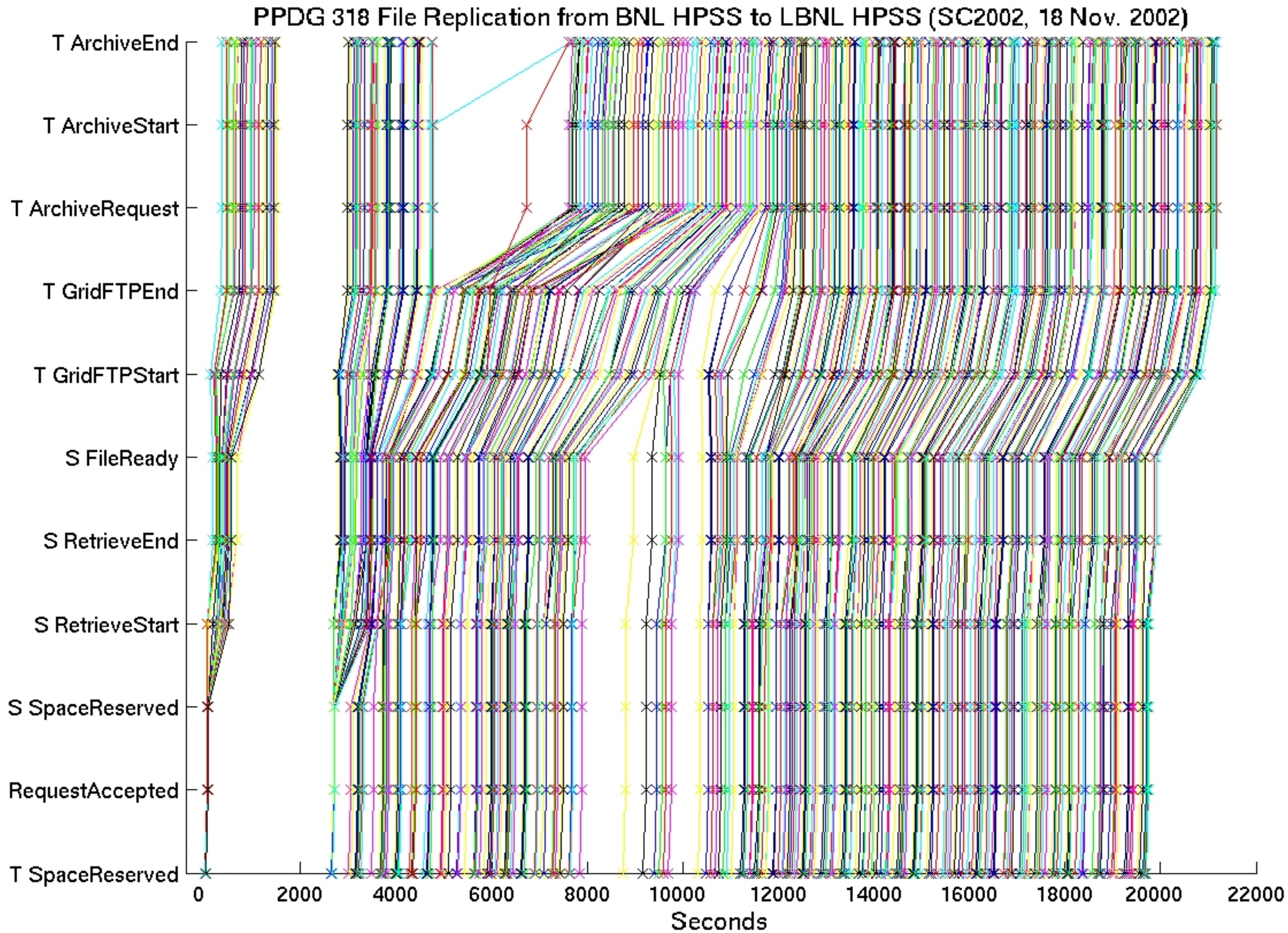


File Tracking Helps to Identify Bottlenecks



Shows that archiving is the bottleneck

File Tracking Shows Recovery From Transient Failures



Total: 45 GBs

Robust Multi-File Replication

- Main results
  - DataMover is being used in production for over three years
  - Moves about 10 TBs a month currently
  - Averages 8 MB/s (64 Mb/sec) over WAN
  - Eliminated person-time to monitor transfer and recover from failures
  - Reduced error rates from about 1% to 0.02% (50 fold reduction)*

* http://www.ppdg.net/docs/oct04/ppdg-star-oct04.doc