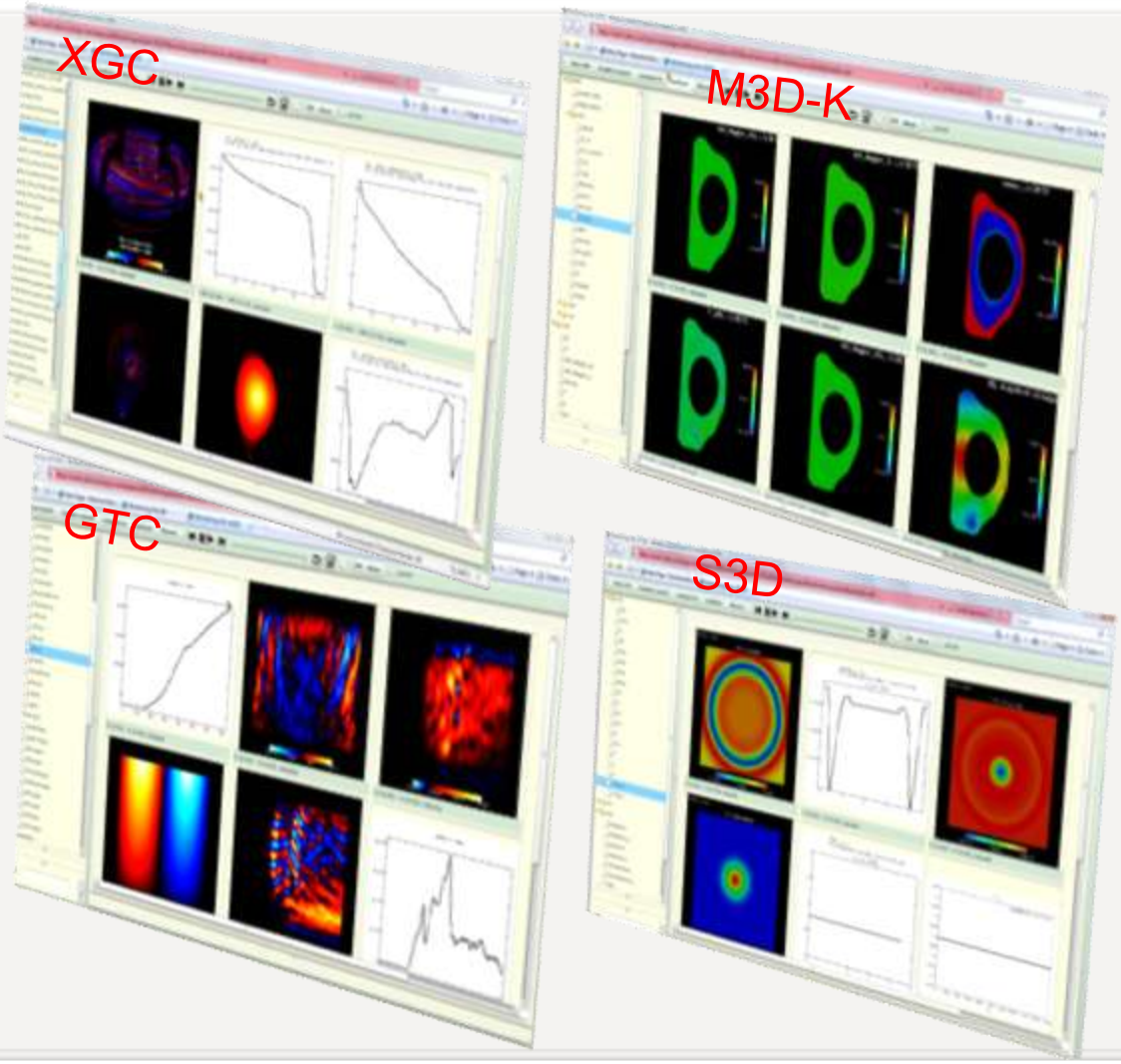
# Petascale Data Management using FIESTA




## Fourth Workshop on Ultrascale Visualization

## 10/28/2009

Scott A. Klasky

klasky@ornl.gov

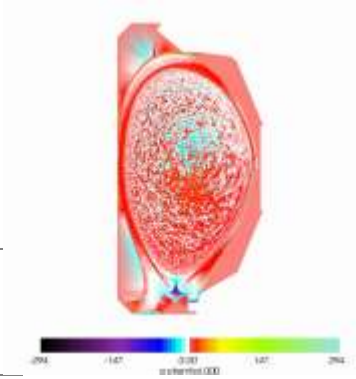
Collaborators from
SDM Center, CPES, GPSC,
GSEP, Sandia, ORNL

H. Abbasi, J. Cummings, C. Docan,, S. Ethier, A Kahn, Q. Liu, J. Lofstead, P. Mouallem, M. Parashar, N. Podhorszki, K. Schwan, A. Shoshani, C. Silva, R. Tchoua, M. Vouk, M. Wolf, F. Zheng,

# Outline

- Application Driven!
- SOA
- FIESTA
- ADIOS
- Workflows
- Dashboards.
- Conclusions

# Our CPES workflow: multi-scale, multi-physics, multi-codes. (2004)



Blob Detection

L-mode and L-H transition (Kinetic-edge)

$\tau_{MHD} >> \tau_{Edge\ Transport}$

TB's of particle data

Profile Data – Every .1 ms (10minutes), ~10mb

Puncture plot

Slow time scale(ms) Pedestal growth (Kinetic-edge)

Check ELM (M3D-edge)

Island detection

No

10 GB

Yes

Yes

Feature Detection

Visualization Post-processing

Particle data

Fast time scale Er & f evolution and transport (Kinetic-edge)

Profile data

Fast time scale Profile evolution by ELM crash (M3D-edge)

In-situ vis

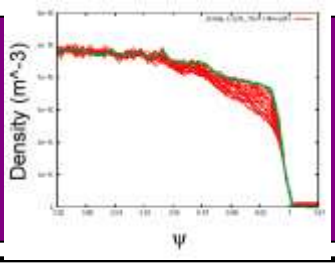
Fluid-data

Er and Closure

L-mode

H-mode

Crash ends. Check L-mode (Kinetic-edge)




Data storage

Collaboration of data

# Goal of our project.

- New scientific insights ar
  fusion c
- Many as
  - Strong c
  - Loose c
- Our app oks at coupling as data
  movem een services.
- High Lev
  - Make it framework.
  - Allow us s for their services.
- We take oach.


SDM CENTER

OAK RIDGE National Laboratory

# Why talk about this at a Visualization meeting?

- All of the data being moved needs to be analyzed/visualized.
- All of the data being coupled needs to be visualized, to aid in debugging the science of complex physical phenomena.
- How do you determine if there is a problem when coupling 10 codes, each running on over 10K cores?
  - Visualization is a key to aid in the development and understanding of the science.

SDM CENTER

OAK RIDGE National Laboratory

# Service Oriented Architecture

- *I/O services:* Efficient Adaptable I/O support is critical
  - This I/O service should be componentized so that it allows codes to switch between and tune different methods easily.
  - For example, a user may easily switch between file-based I/O to memory-base I/O or to switch formats without changing any code, all the while maintaining the same data model when switching I/O components.
- *Code-coupling services:* Code-coupling services includes codes that can run on the same or different machines, codes that are tightly coupled (memory-to-memory), as well as codes that are loosely coupled (exchange of data through files).
- *Automatic online (in-transit) data processing* services:
  - Examples : generating summary statistics, graphs, images, movies, etc. Efficient data movement services for distributing and archiving data are critical.

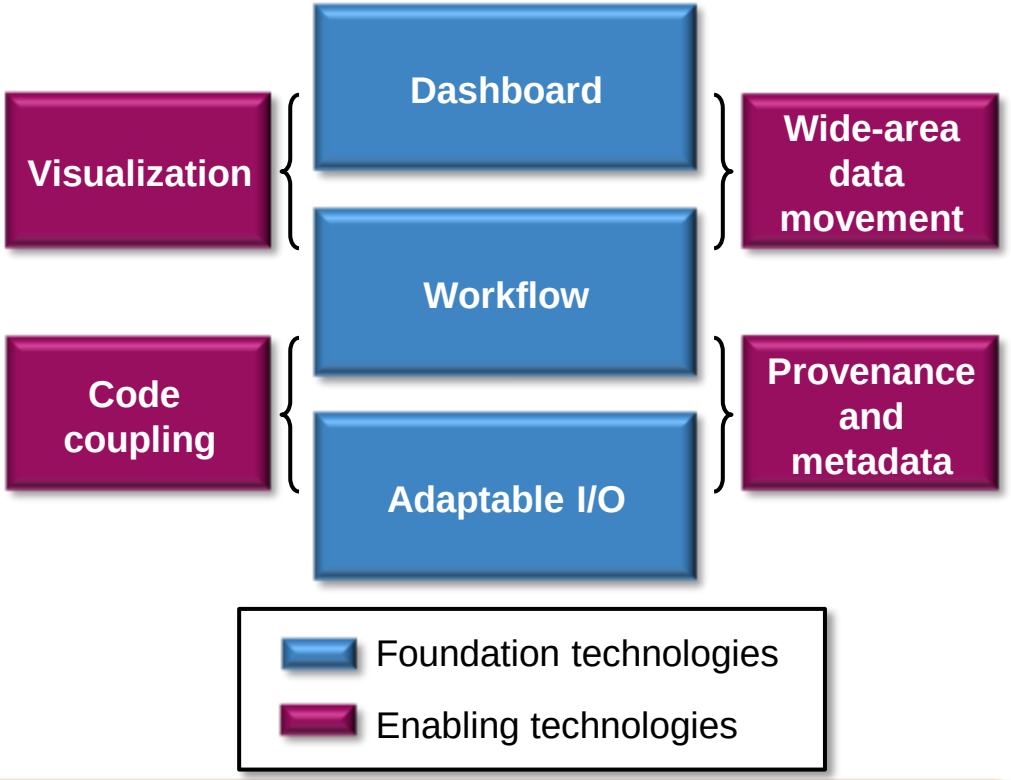
# Service Oriented Architecture-2

- *Workflow engine*: orchestrates the execution of the components within a simulation as well as end-to-end application workflows
- *Monitoring Services*: Dynamic real-time monitoring of large-scale simulations (during execution) is critical
- *Automated collection of provenance*
  - "data provenance"
  - "system provenance
  - "workflow provenance"
  - "performance provenance"
- *Interfaces and portals*:
  - Powerful, but simple to use, user interfaces (e.g., dashboards) provide critical access to the simulation process and the data products for understanding and exploration, as well as management and control

# FIESTA

Framework for Integrated End-to-end SDM Technologies and Applications

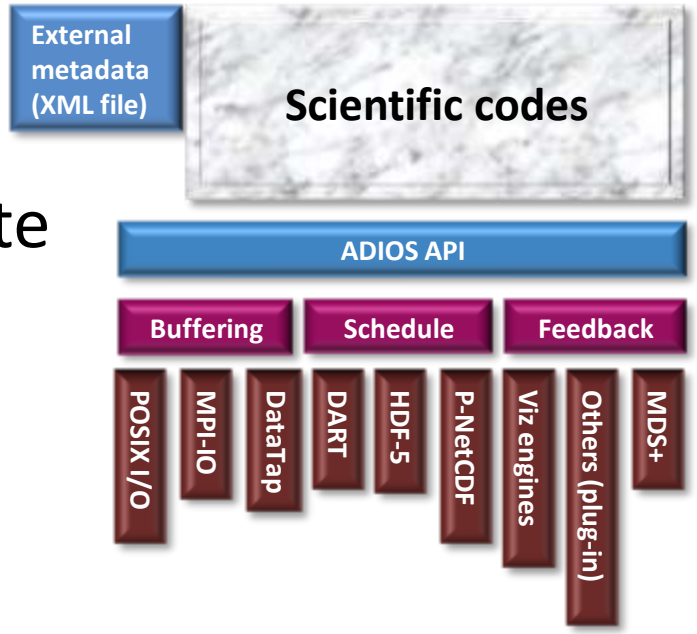
- Adaptable I/O
- Workflows
- Dashboard.
- Provenance.
- Code Coupling.
- WAN data movement.
- Visualization.




**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

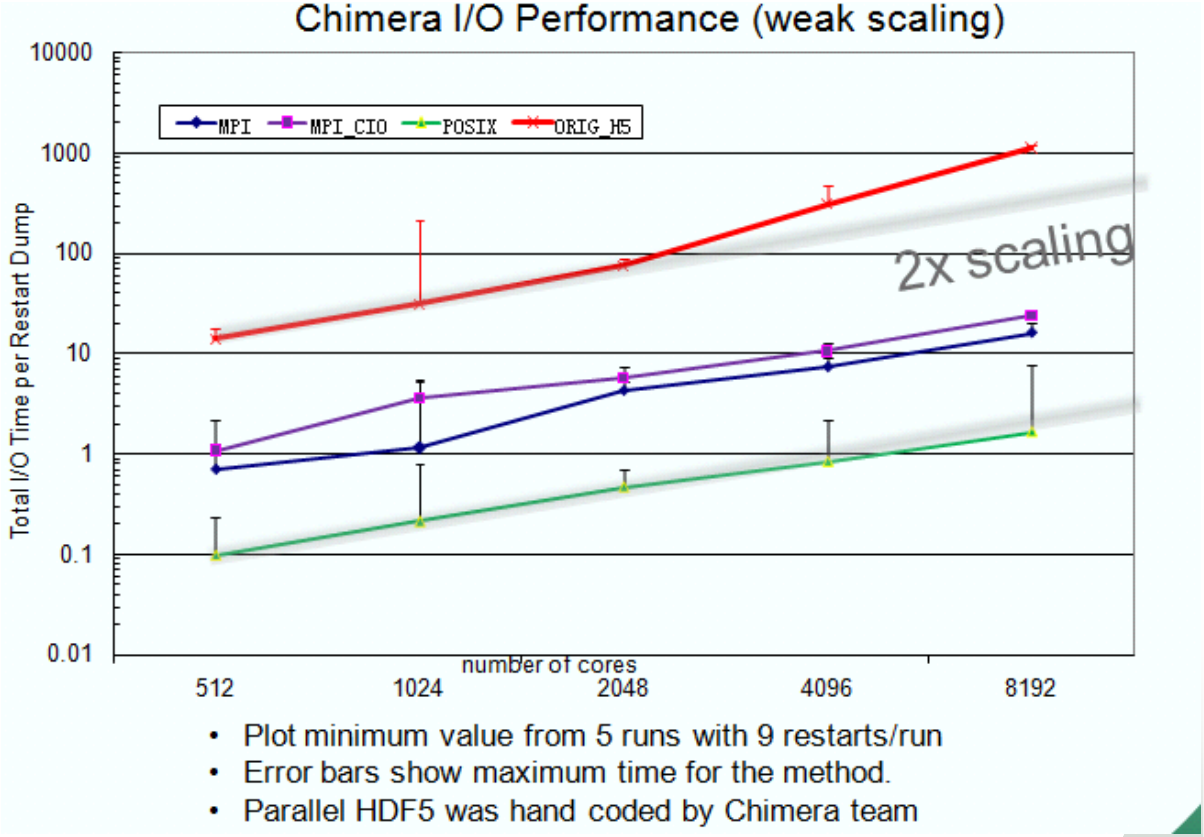
# ADIOS: Adaptable I/O System

- Overview
  - 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



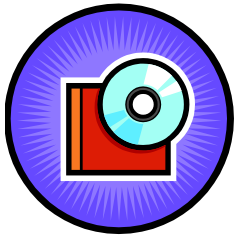
# J. Lofstead, F. Zheng, S. Klasky, K. Schwan, Input/output APIs and Data Organization for High Performance Scientific Computing, PDSW 2008.

- Introduce ADIOS.
- GTC: 60% raw BW.
  - Used multiple files, by allowing scientist to choose the best number of files based on their physics.
- Chimera 1000x better





- Z Lin, Y Xiao, I Holod, W Zhang, W Deng, S Klasky, J Lofstead, C Kamath, and N Wichmann , *Advanced simulation of electron heat transport in fusion plasmas*, SciDAC 2009, *Journal of Physics: Conference Series* **180**, 012059 (2009).
- *C S Chang et al.,* Whole-volume integrated gyrokinetic simulation of plasma turbulence in realistic diverted-tokamak geometry , SciDAC 2009, *Journal of Physics: Conference Series* **180**,

**ADIOS 1.0: SC 2009 Release**
**http://www.nccs.gov/user-support/adios**

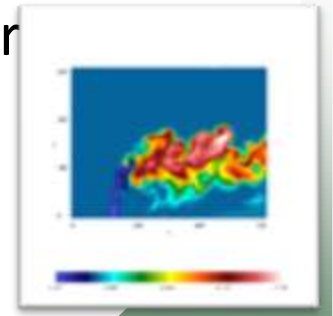
SDM CENTER

OAK RIDGE National Laboratory

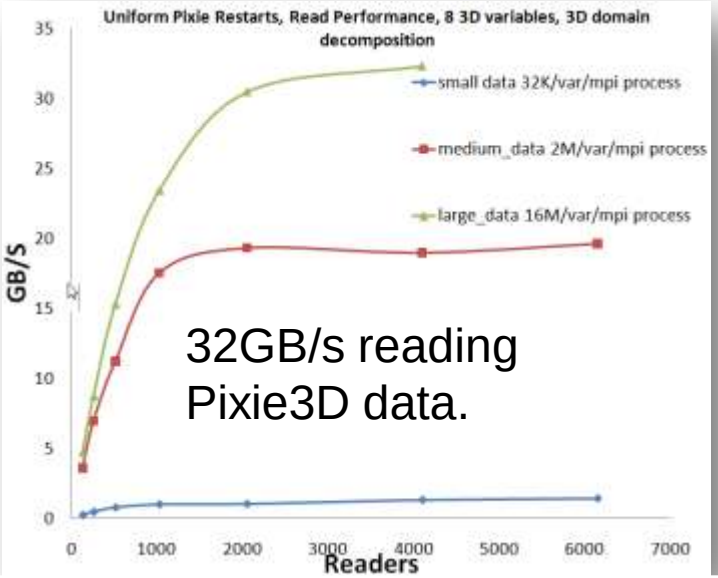
# BP File Format.

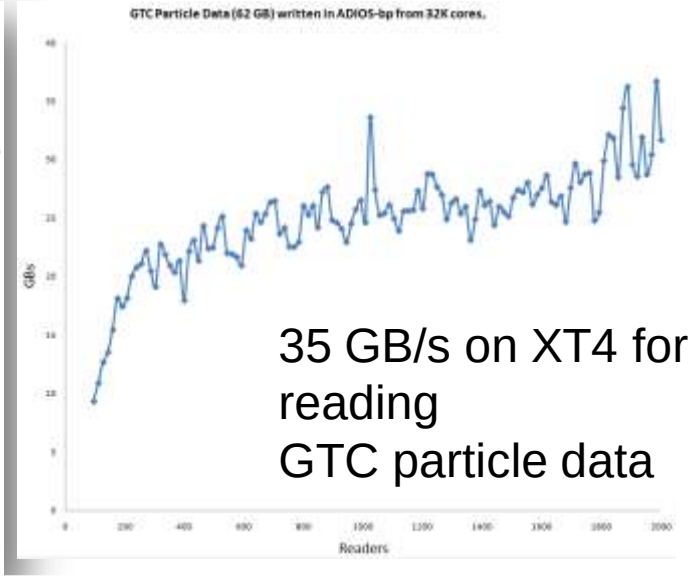
| Process Group 1 | Process Group 2 | ... | Process Group n | Process Group Index | Vars Index | Attributes Index | Index Offsets and Version # |
|---|---|---|---|---|---|---|---|

- J. Lofstead, F. Zheng, S. Klasky, K. Schwan, "Adaptable Metadata Rich IO Methods for Portable High Performance IO", IPDPS 2009, IEEE Computer Society Press 2009.
- Necessary to have a hierarchical view of the data (like HDF5).
- Necessary to have ways to easy extend arrays without moving data.
- Bp is metadata rich.
- Tested at scale (140K processors for XGC-1) with over in a single file for the 2009 Joule run.
- Used by many codes for restarts and analysis output



# But BP is "write-optimized". Let's try reading in BP files from 3D and 1D domain decompositions from physics runs



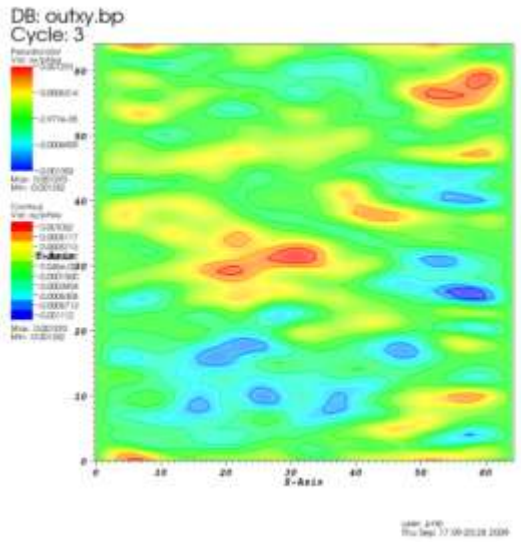



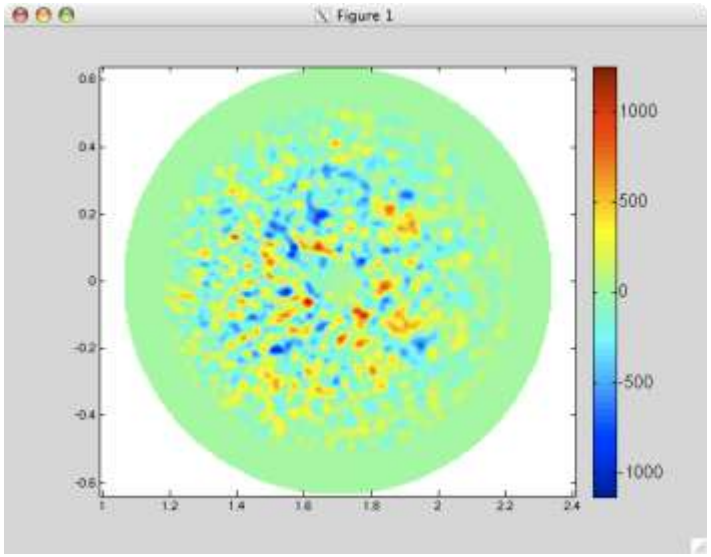

- M. Polte, J. Lofstead, J. Bent, G. Gibson, S. Klasky, Q. Liu, M. Parashar, N. Podhorszki, K. Schwan, M. Wingate, M. Wolf, "… And eat it too: High read performance in write-optimized HPC I/O middleware file formats", PDSW 2009
- Reading of data from the same number of processors, more processors, or less processors results in high performance reading.
  - Compared to other parallel file formats, ADIOS-BP at least as fast when reading in all of the variables from a file.

# bpls (can extract any portion of data).

- $ time /ccs/proj/e2e/pnorbert/ADIOS/ADIOS/trunk/utils/bpls/bpls -l record.bp -v
  of groups: 1
  of variables: 32
  of attributes: 0
  time steps: 10 starting from 1
  **file size: 162 GB**
  bp version: 1
  Group record:
  double /time {10} = 0.003 / 0.03
  integer /itime {10} = 3 / 30
  double /dt {10} = 0.001 / 0.001
  integer /nvar scalar = 8
  integer /dimensions/nxd+2 scalar = 1026
  integer /dimensions/nyd+2 scalar = 514
  integer /dimensions/nzd+2 scalar = 514
  double /var/v1 {10, 514, 514, 1026} = 1 / 1
  **double /var/v2 {10, 514, 514, 1026} = -2.07946e-06 / 3.43263e-08**
  double /var/v3 {10, 514, 514, 1026} = -1.17581e-10 / 1.24015e-10
  double /var/v4 {10, 514, 514, 1026} = -3.65092e-13 / 3.65092e-13
  double /var/v5 {10, 514, 514, 1026} = -7.95953e-11 / 7.95953e-11
  double /var/v6 {10, 514, 514, 1026} = -0.184178 / 0.0123478
  double /var/v7 {10, 514, 514, 1026} = -0.000488281 / 0.984914
  double /var/v8 {10, 514, 514, 1026} = 0 / 0
  byte /name/v1_name {20} = 32 / 111
  byte /name/v2_name {20} = 32 / 94
  byte /name/v3_name {20} = 32 / 94
  byte /name/v4_name {20} = 32 / 94
  byte /name/v5_name {20} = 32 / 94
  byte /name/v6_name {20} = 32 / 94
  byte /name/v7_name {20} = 32 / 94
  byte /name/v8_name {20} = 32 / 101
  integer /bconds {48} = -4 / 7
  **real 0m2.091s**

# ADIOS BP Visit & Matlab Readers



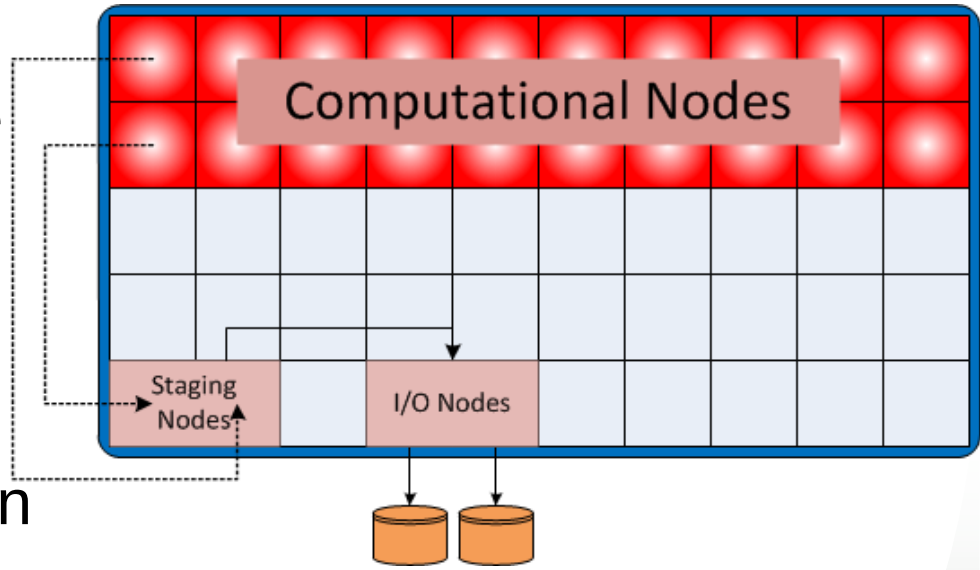

- **rz=adiosread(meshfile,'/coordinates/values');**
- **var=adiosread(pfile,'pot3d','/node_data[1]/values');**
- Visit BP reader is parallel.
- Gives us a 2x speedup for XGC data over other popular file formats.

Matlab jobs will also be allowed to be submitted from the dashboard.

# Use staging nodes to focus on buffered asynchronous I/O to focus on aggregates not peaks

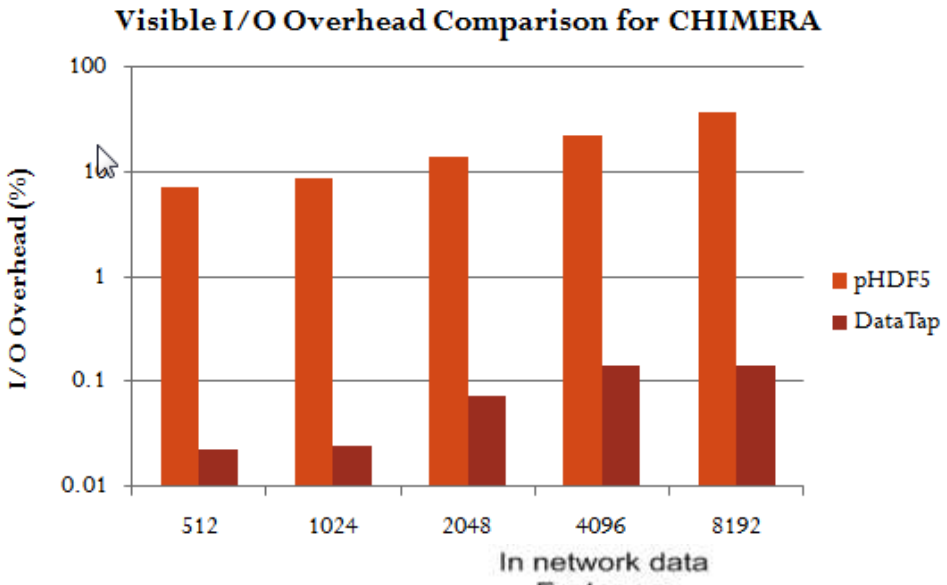
- Abbasi, H., Wolf, M., Eisenhauer, G., Klasky, S., Schwan, K., and Zheng, F. 2009. DataStager: scalable data staging services for petascale applications. In *Proceedings of the 18th ACM international Symposium on High Performance Distributed Computing* HPDC '09.

- Reduces performance linkage between I/O subsystem and application
- Enables optimizations based on dynamic number of writers
- High bandwidth data extraction from application

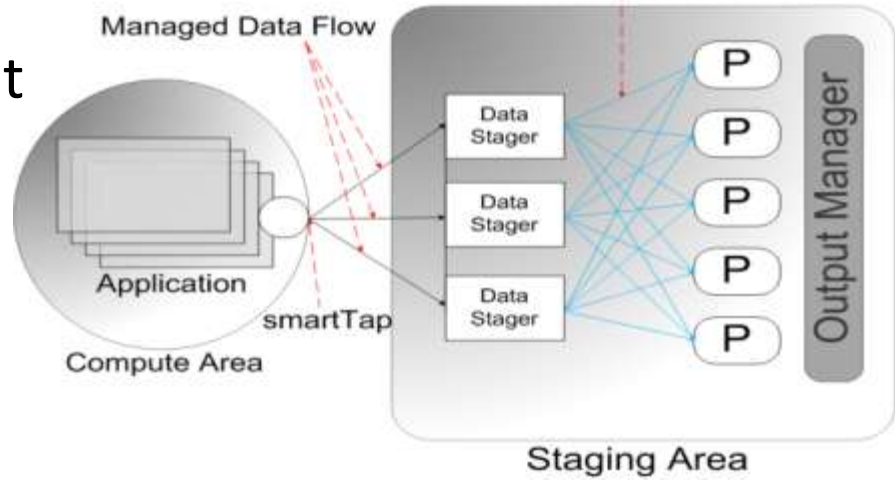


# But can we do more than just I/O?

- H. Abbasi, J. Lofstead, F. Zheng, S. Klasky, K. Schwan and M. Wolf, "Extending I/O through High Performance Data Services", Cluster Computing 2009, New Orleans, LA, August 2009.

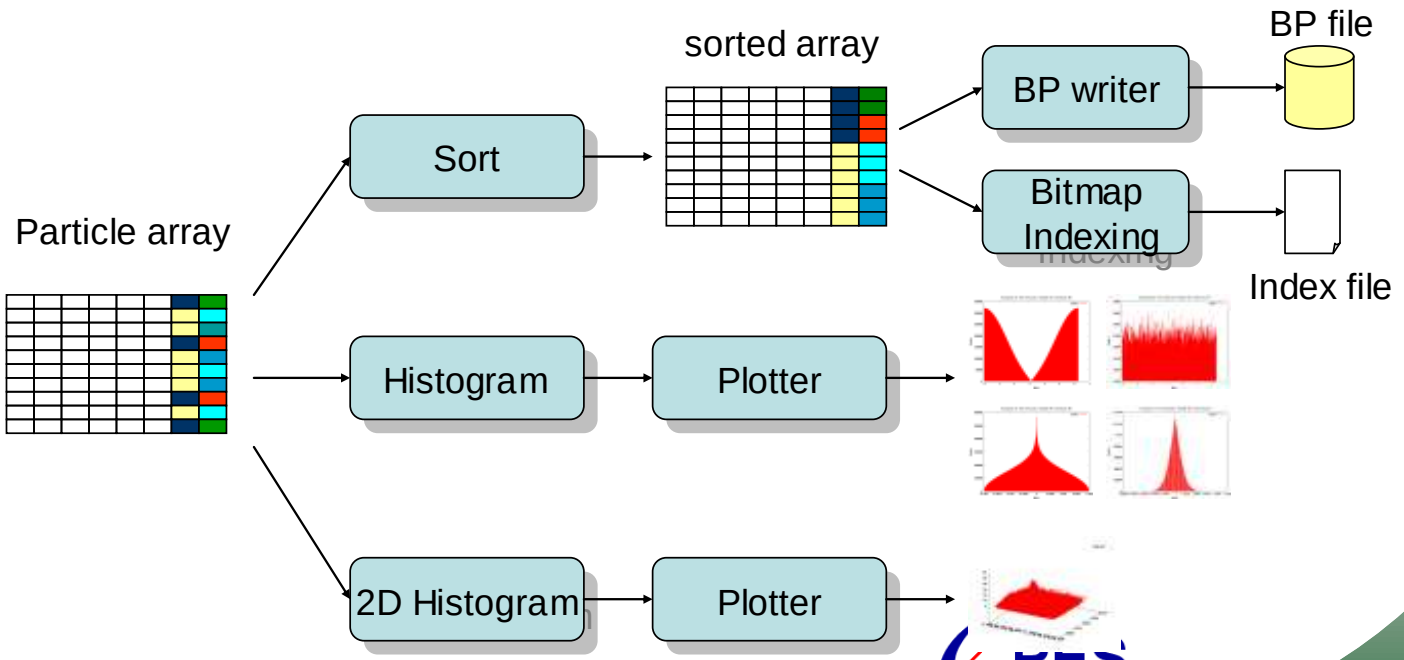



- **Approach**
  - Encode structure information with extracted data
  - C-on-Demand used for in-flight data filtering
  - RDMA based request-read protocol



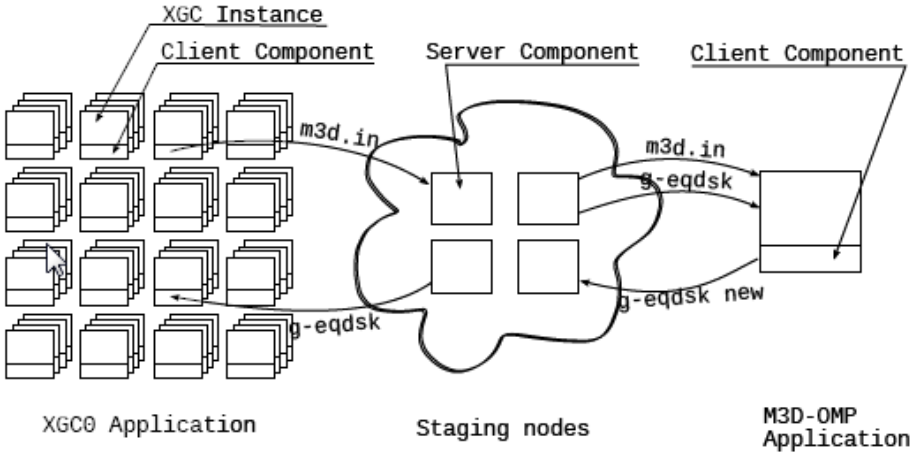
# But can we do more?

- F. Zheng, H. Abbasi, C. Docan, J. Lofstead, S. Klasky, Q. Liu, M. Parashar, N. Podhorszki, K. Schwan, M. Wolf, "PreDatA - Preparatory Data Analytics on Peta-Scale Machines", submitted to IPDPS 2010.
- Use the staging nodes and create a workflow in the staging nodes.
- Allows us to explore many research aspects.
- improve total simulation time by 2.7%
- Allow the ability to generate online insights into the 260GB data being output from 16,384 compute cores in 40 seconds.



# But what about in-memory code coupling?

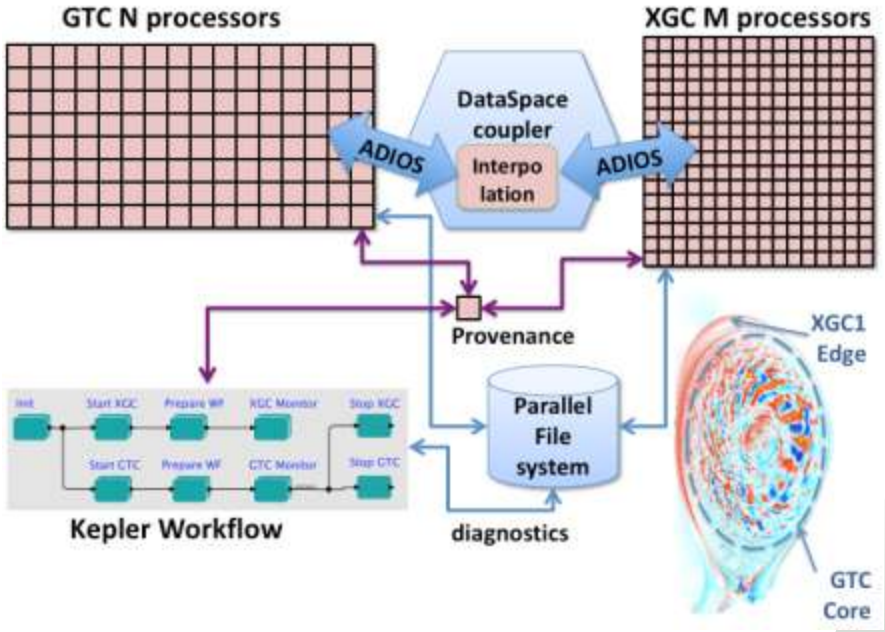
- Docan, Parashar, Cummings, Podhorszki, Klasky, "Experiments with Memory-to-Memory Coupling on Fusion Simulations", Rutgers Technical Report
- *Cummings, Klasky, Podhorszki, Barreto, Lofstead, Schwan, Docan, Parashar, Sim, Shoshani, "EFFIS: and End-to-end Framework for Fusion Integrated Simulation", to appear in PDP 2010,* http://www.pdp2010.org/.
- Key IDEAS:
  - Code coupling becomes part of the I/O pipeline for in-memory code coupling.
  - Uses a shared space approach to couple codes, based on tuple spaces.
  - Using ADIOS, we can switch from a file-based coupled system to a memory code coupling scheme.



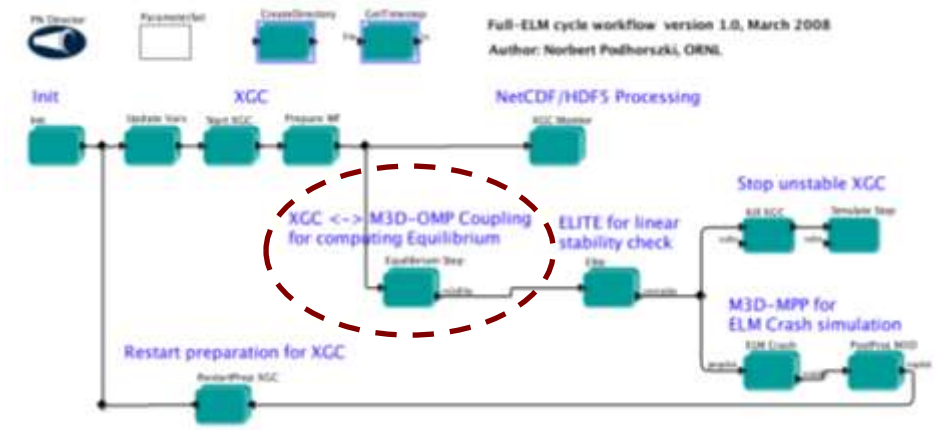
# But we want 1 workflow for all couplings!

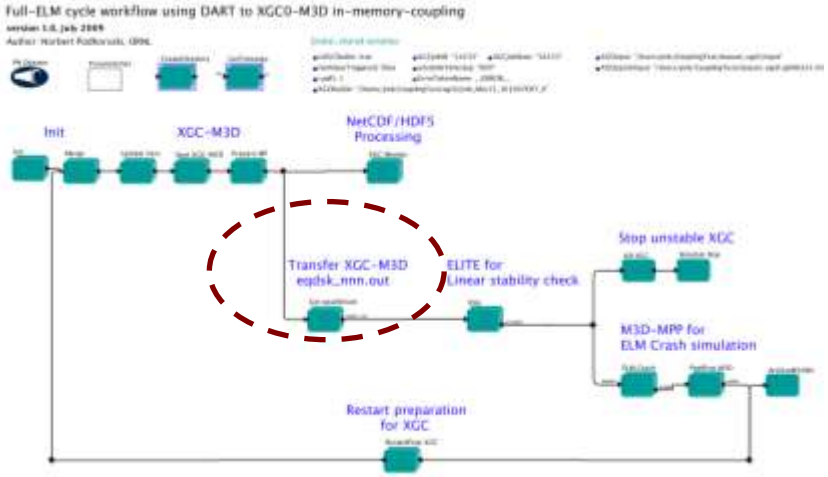
- N. Podhorszki, S. Klasky, Q. Liu, C. Docan, M. Parashar, H. Abbasi, J. Lofstead, K. Schwan, M. Wolf, F. Zheng, J. Cummings, "Plasma fusion code coupling using scalable I/O services and scientific workflows", accepted to works09.

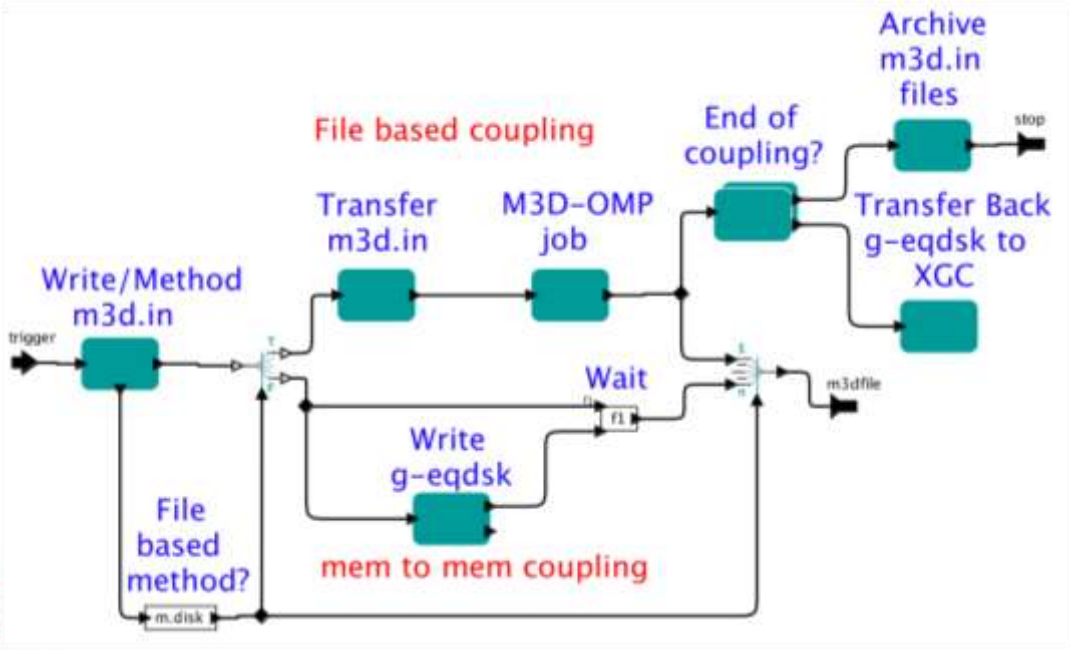
- Key idea.
  - Use 1 workflow for both in-memory coupling and file-based coupling.
  - Provenance is the key using ADIOS for the coupling middleware with Kepler as the director.



# Coupling workflow (file-based – memory based, both)



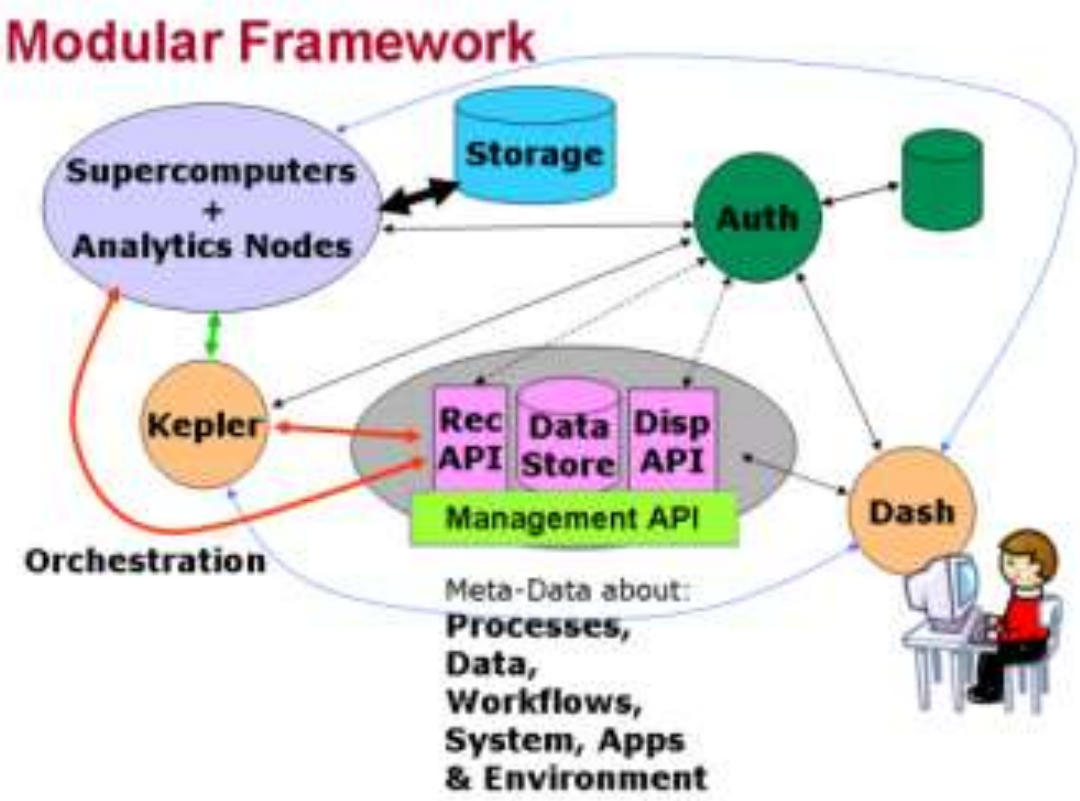




- Actor "Write/Method m3d.in"
  - fire once on trigger; output whenever m3d.in file is written
  - also tell which transport method was used
- Actor "Write g-eqdsk"
  - fire on trigger; output when g-eqdsk file is written

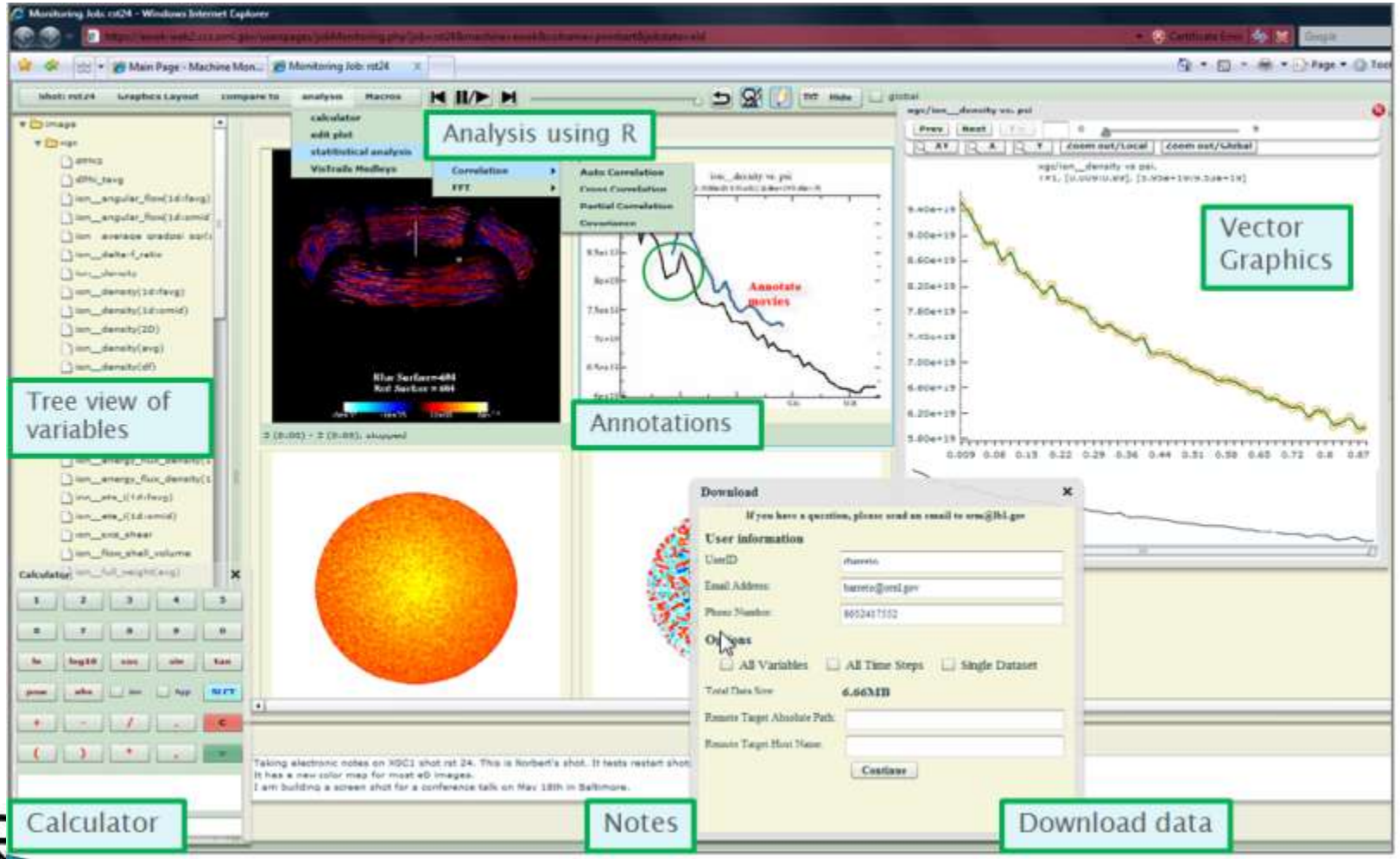
# Provenance is key

- P. Mouallem, M. Vouk, S. Klasky, N. Podhorszki and R. Barreto: “Tracking Files Using the Kepler Provenance Framework”
  Proceedings of 21st International Conference on Scientific and Statistical Database Management, SSDBM’09.
- We have an automated provenance capturing system for all types of provenance, and use this in our dashboard/portal environment.



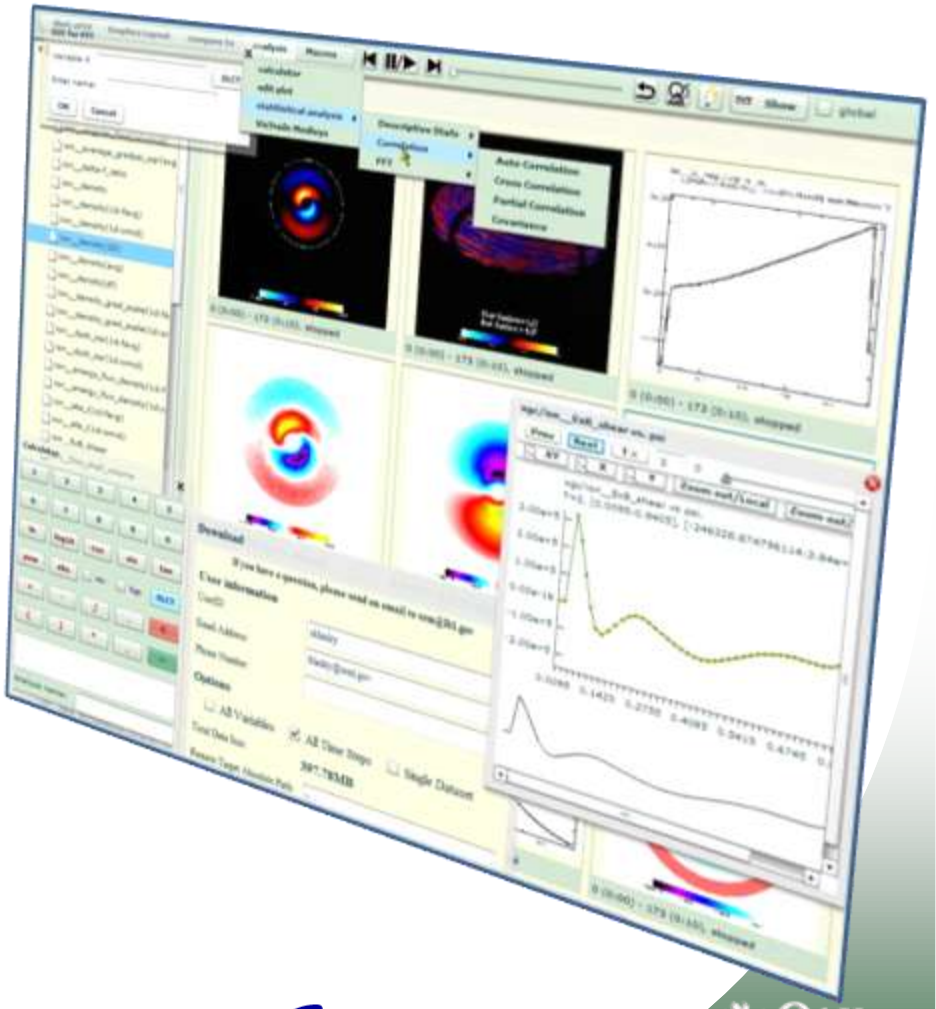
# But where do we see the data? What about collaboration?

- R. Barreto, S. Klasky, N. Podhorszki, P. Mouallem, M. Vouk: "Collaboration Portal for Petascale Simulations", 2009 International Symposium on Collaborative Technologies and Systems, (CTS 2009).



# Dashboard Features

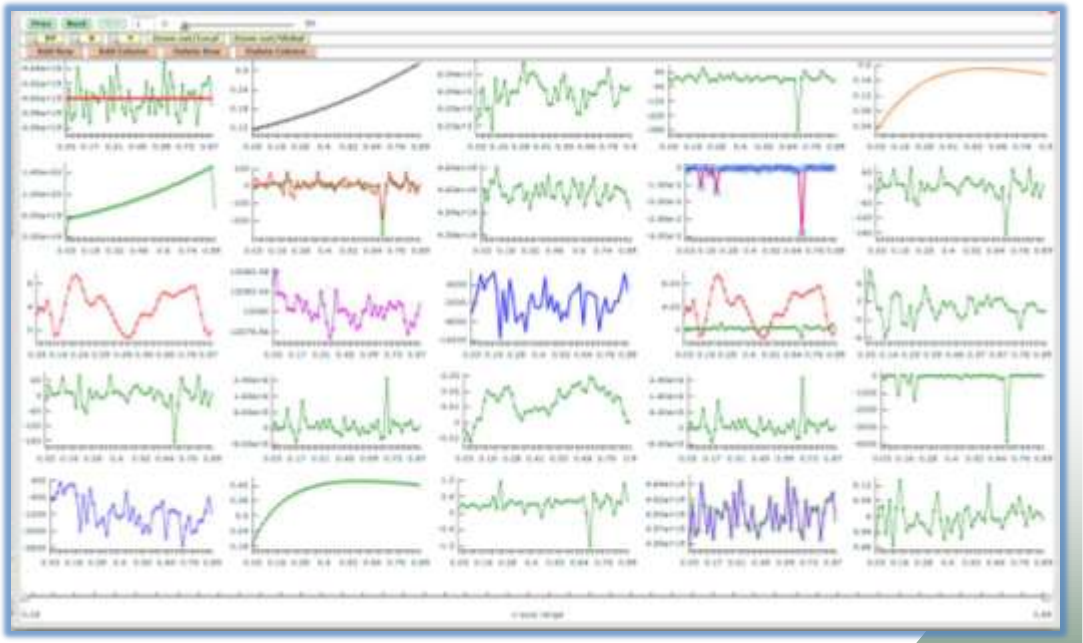
- Machine Monitoring
- Simulation Monitoring
- Collaboration
- Analysis
  - Calculator
  - Statistical Analysis
  - Vector Graphics
  - Users' scripts



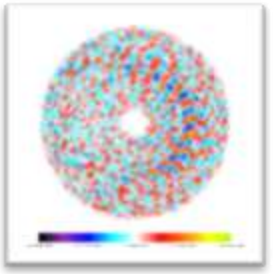
# Vector Graphics

- Workflow produces **default** movies
- The goal with vector graphics is to allow scientists to interact with the data
  - Modify axis and time range
  - Focus on specific areas (zoom, pan)
  - Customize plot (colors, labels etc.)
  - …

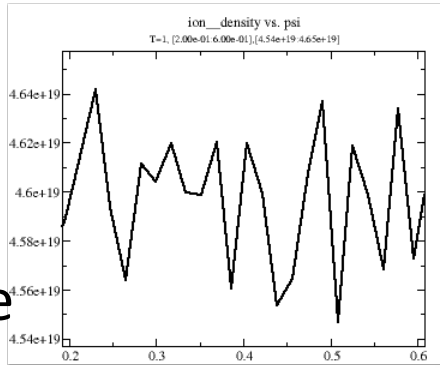
- Ultimately goal is to allow users to generate new publication quality customized plots

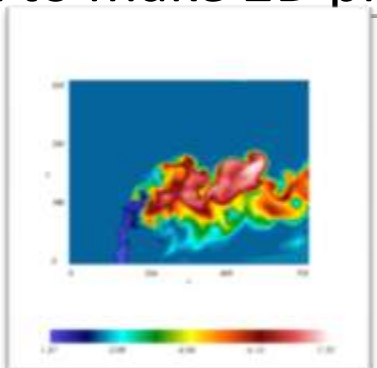


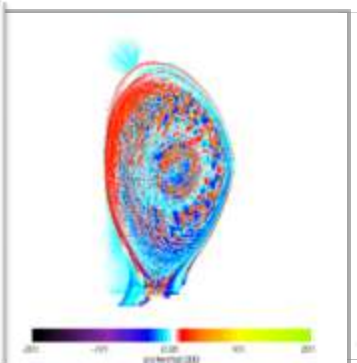
# Need better viz tasks for the workflow for the dashboard: The plotter.

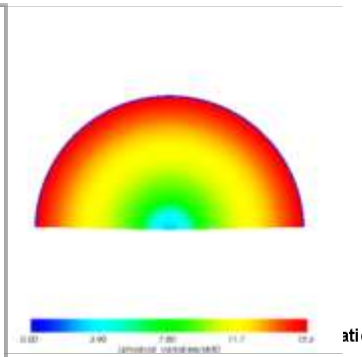


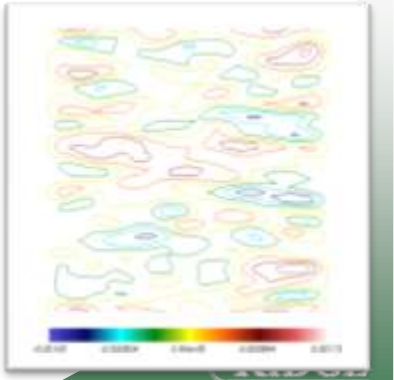
- Goal: solve all viz tasks of our workflows with one tool
  - this is where we spent most of the time when developing new workflows
- a recipe collection
- reads ADIOS BP/NetCDF/HDF5 arrays
  - any slice from any multi-dimensional array
- Use xmgrace to make X-Y plots
- x and y array can come from different files
- loop over a dimension to make many plots at once
- regexp for plotting many variables at once
- additional ‘t’ array to index the time loop for XGC restart support
- use VTK to make 2D plots









# But what about 3D viz in workflows for the dashboard?

E. Santos, J. Tierny, A. Khan, B. Grimm, L. Lins, J. Freire, V. Pascucci, C. Silva, S. Klasky, R. Barreto, N. Podhorszki, Enabling Advanced Visualization Tools in a Web-Based Simulation Monitoring System, accepted escience 2009.

- We propose the use of a summary structure, called contour tree
  - Captures the topological structure of a scalar field
  - guides the user in identifying useful isosurfaces.
- We have designed an interface which has been integrated with FIESTA, that allows users to interact with and explore multiple isosurfaces.

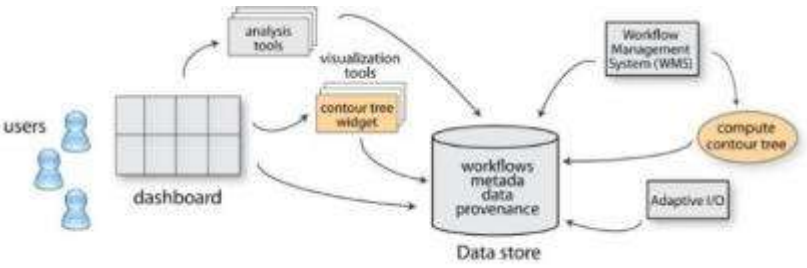
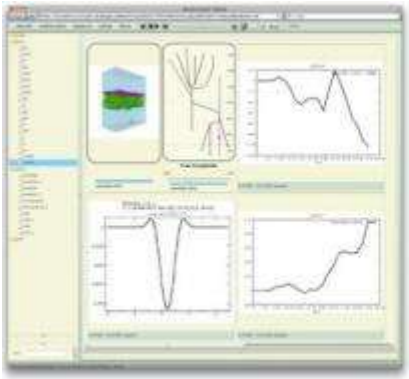
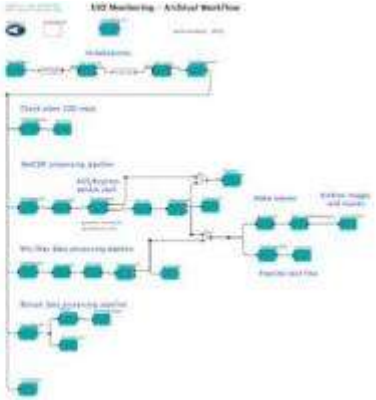



Figure 1. eSimMon system architecture.





# Conclusions

- FIESTA embraces a SOA architecture.
  - Services allow users to exchange components easily.
- FIESTA structure and features driven by technical requirements of application community
  - Extreme code scalability
  - Efficient and flexible I/O system
  - Independent code development and testing
  - Adaptable code coupling with minimal changes
  - Collaboration of theorists and experimentalists
- FIESTA approach should also benefit broader multi-physics simulation community.
- Embrace petascale computing .
- ADIOS y'all