END TO END COMPUTING TECHNOLOGY AT ORNL

Center for Gyrokinetic/MHD Hybrid Simulation of Energetic Particle Physics in Toroidal Plasmas (CSEPP)

March 29, 2008

Scott Klasky

R. Barreto, S. Hodson, C. Jin, N. Podhorszki

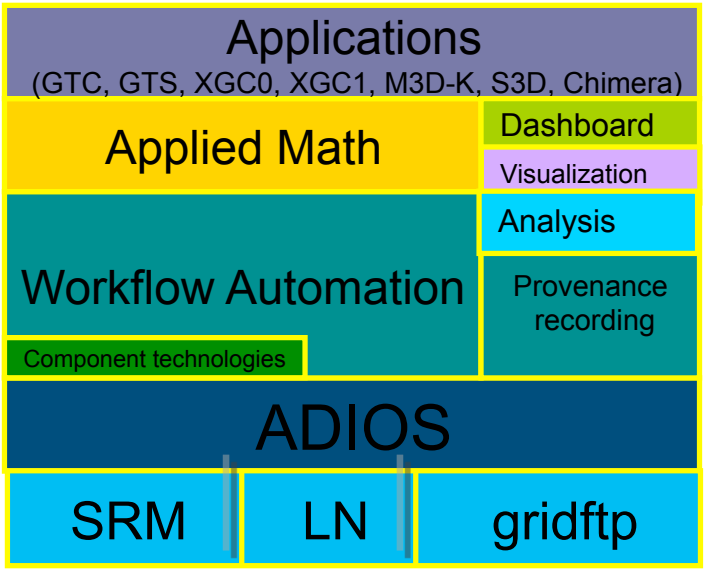
SDM CENTER CSEPP

Office of Science U.S. DEPARTMENT OF ENERGY

OAK RIDGE National Laboratory

END TO END COMPUTING AT ORNL

- Combines
  - Petascale Applications.
  - Petascale I/O techniques.
  - Workflow Automation.
  - Provenance capturing system.
  - Dashboards for real-time monitoring/controlling of simulations.

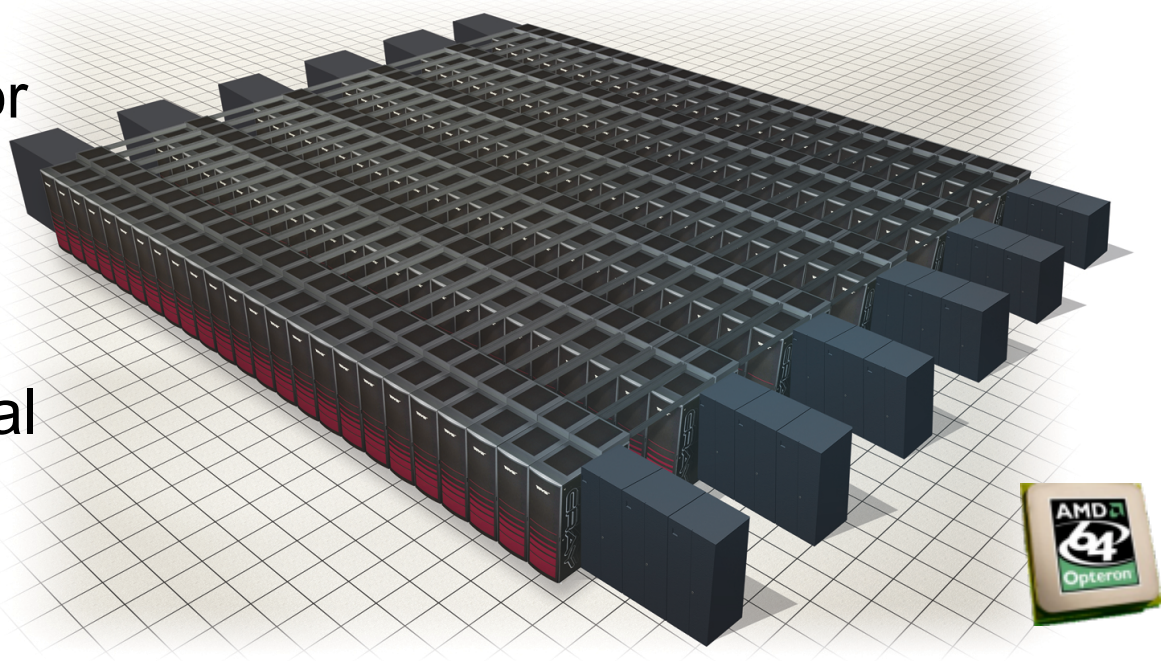


- **Basic Idea**: place highly annotated fast, easy to use I/O methods in the code, which can be monitored, and controlled, have a workflow engine record all of the information, visualize this on a dashboard and allow collaborators easy access to data. Have everything report to a database.
- Remember: It's all about the science!

1 PETAFLOPS SYSTEM - CRAY

- FY 2009: Cray XT

- 1 Petaflops system
- 37 Gigaflops processor
- Over 27K quad-core processors
- 2 GB/core; 223 TB total
- 240 GB/s disk bandwidth
- 7.5 MW system power
- Liquid cooled



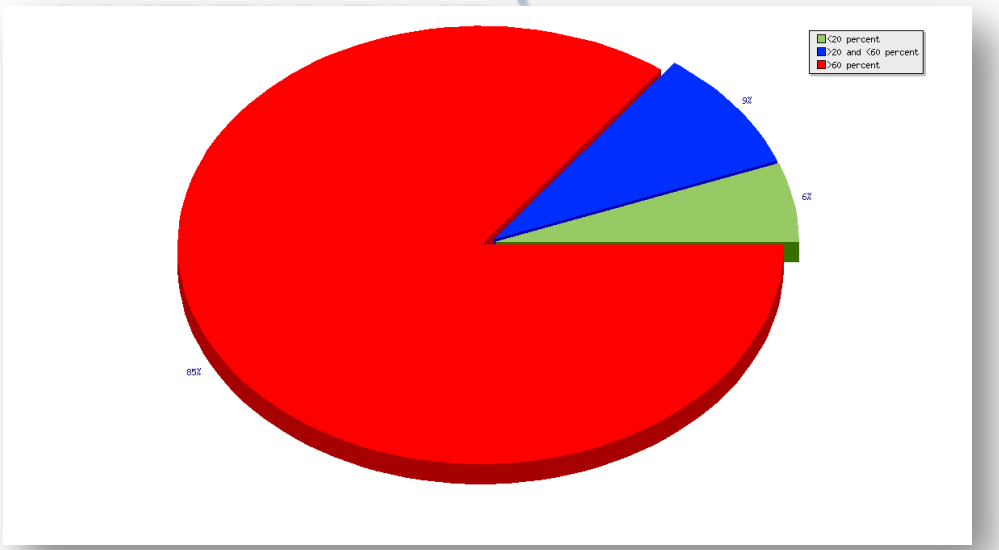
FUSION INCITE 2007: BIGGEST USERS

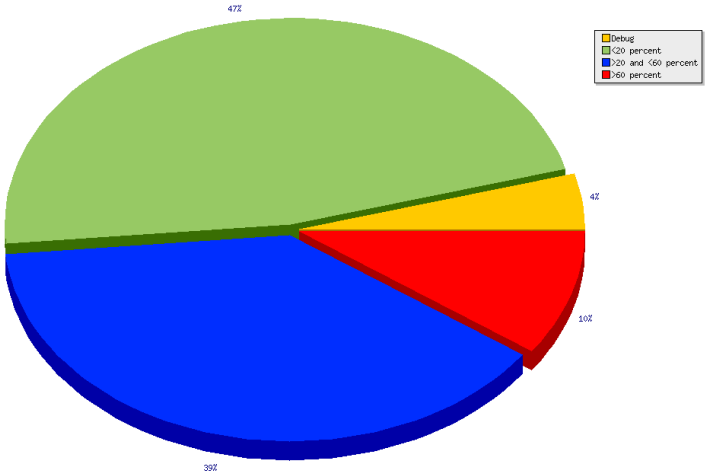
Username	CPU hours (K)	Typical jobs
Ku	2434	4K – 10K
Rewoldt	1682	500-2K
Chen	1417	1K-2K
Xiao	998	8K-10K
Lin	952	6K-8K
Jaeger	896	4K – 6K
Holland	431	1K – 2K
Lang	362	500-1K
Choi	283	1K – 2K
Candy	213	500-1K
Breslau	125	<500

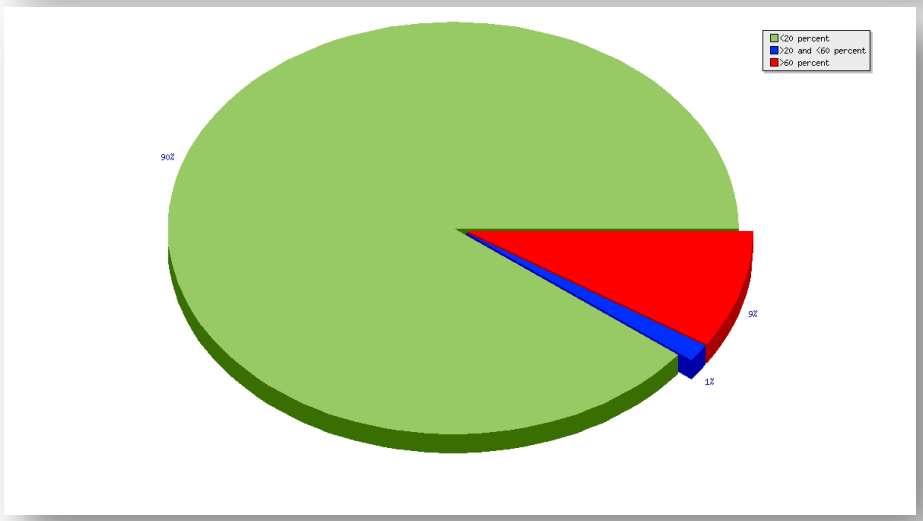
WHAT MAKES USERS ATTRACTABLE.

- GOOD Science.
- History of running large simulations and publishing results.
- Can describe all of the time they will use.
  - 8M hours running this 4 computational experiments.
  - 2M hours running another experiment.
  - 100K hours for debugging.
- Good eye on I/O and data management techniques.

WE TRACK THE USERS. (RED + BLUE GOOD, GREEN BAD)







SO WHO GOT THE INCITE COMPUTER TIME AT ORNL?

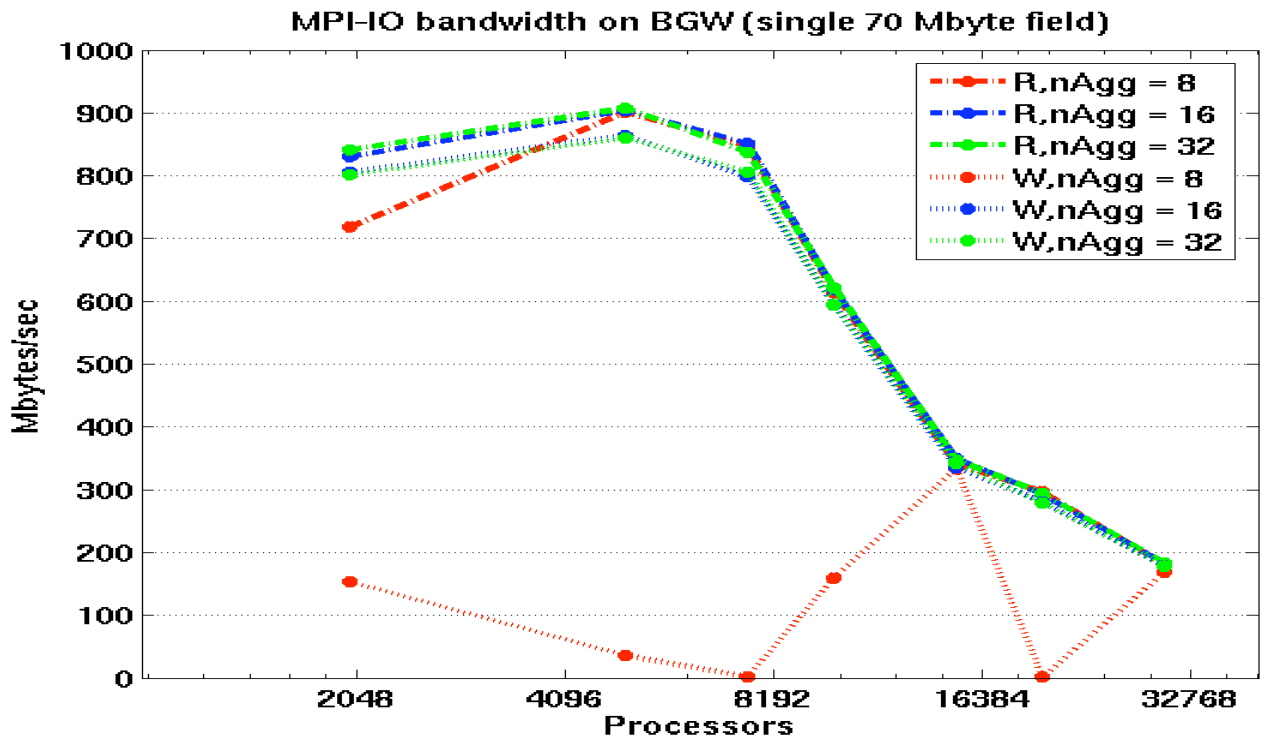
Jackie Chen	Combustion	18M hours
Tony Mezzacappa	Supernova	16M hours
Warren Washington	Climate	16M hours
Robert Harrison	Chemistry	10M hours
Thomas Schulthess	Materials	10M hours.
Jinui Yang	Materials	10M hours
Patrick Diamond	Fusion	8 M hours
David Dean	Nuclear	7.5 M hours
Robert Sugar	QCD	7.1M hours
The rest		35 M hours
TOTAL		142 M hours

CSEPP BEST WAY TO GET TIME.

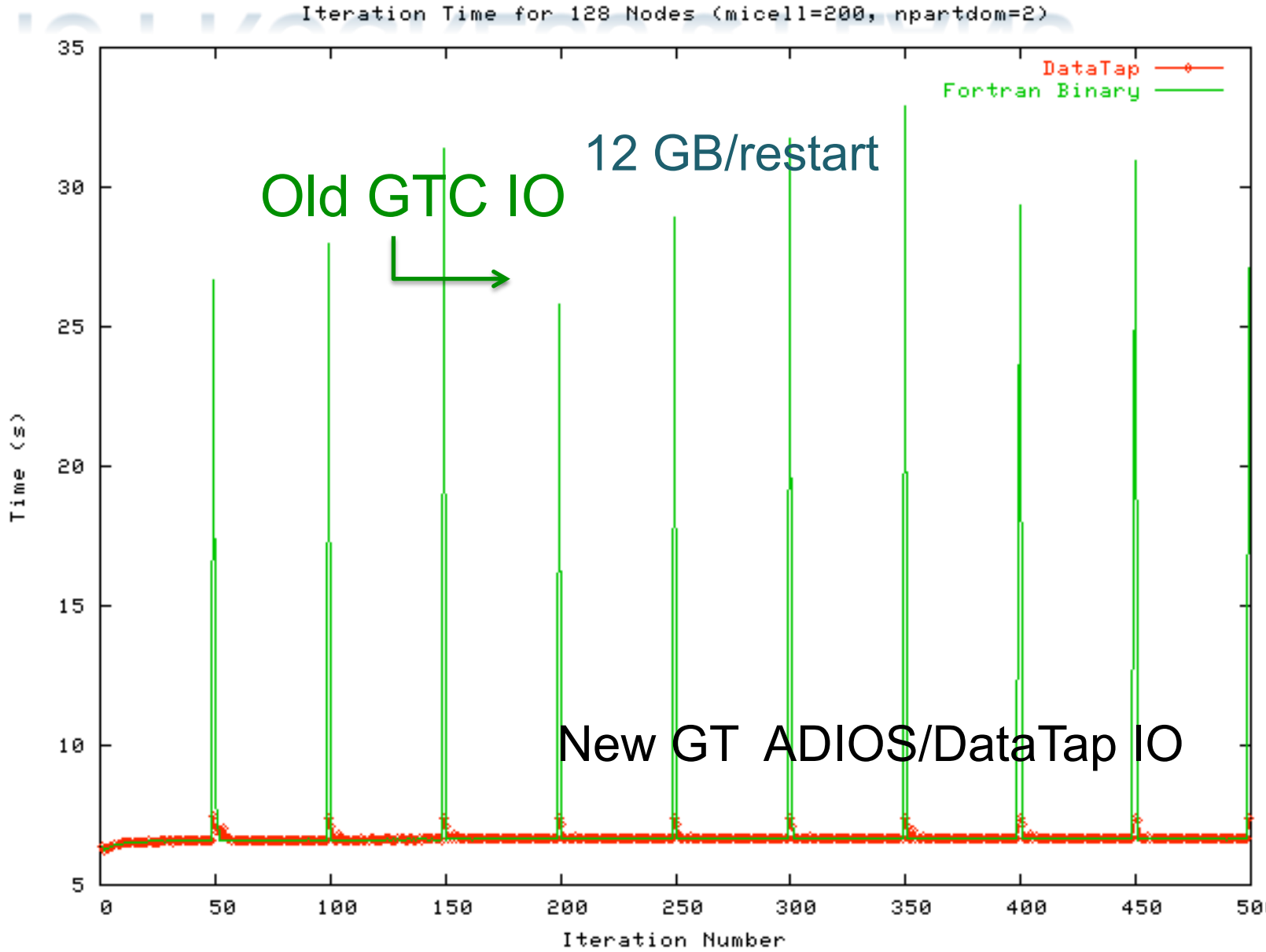
- Team up with others!
  - You need a better track record.
  - Imperative to show a scientific result this year from NERSC run at over 5K processors.
  - 4 camps, Gyro, GTC, AORSA, XGC1.
  - Publish the simulation results and acknowledge the large run.

ORNL END TO END TECHNOLOGIES

- IO is a major problem on supercomputers.
- Want metadata rich output, but fast on all platforms!



IO PROGRESS & PLANS



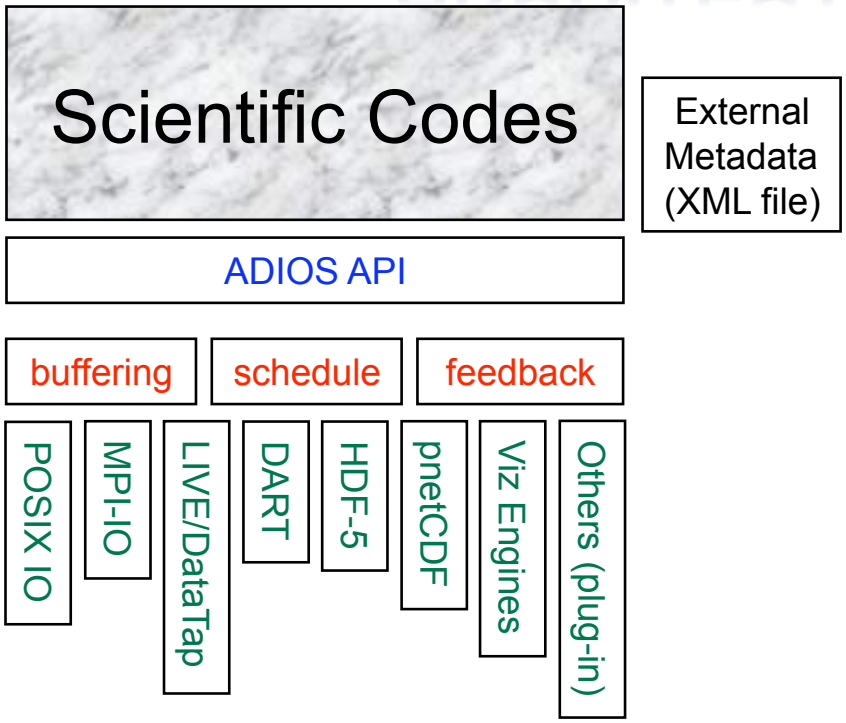
ADAPTABLE IO SYSTEM (ADIOS)

- Combines
  - High performance I/O.
  - In-Situ Visualization.
  - Real-time analytics.
- Collaborating with many institutions (GT, Rutgers, NWU, ANL, CMU, +…)

	GTC	GTC_s	Flash	XGC1	Chimera	S3D	M3D	XGC0
MPI-IO/ORNL Jaguar	25 GBs	22GBs		15 GBs	20 GBs			
Async MPI-IO Jaguar								
DART Jaguar	1.2TB <1							
Datatap/jaguar								
Maviz/jaguar								
Visit/jaguar								
Paraview/jaguar								
Phdf5/jaguar								
Pnetcdf/jaguar								
BGP/IB/GPFS..								

% overhead

ADIOS ARCHITECTURE



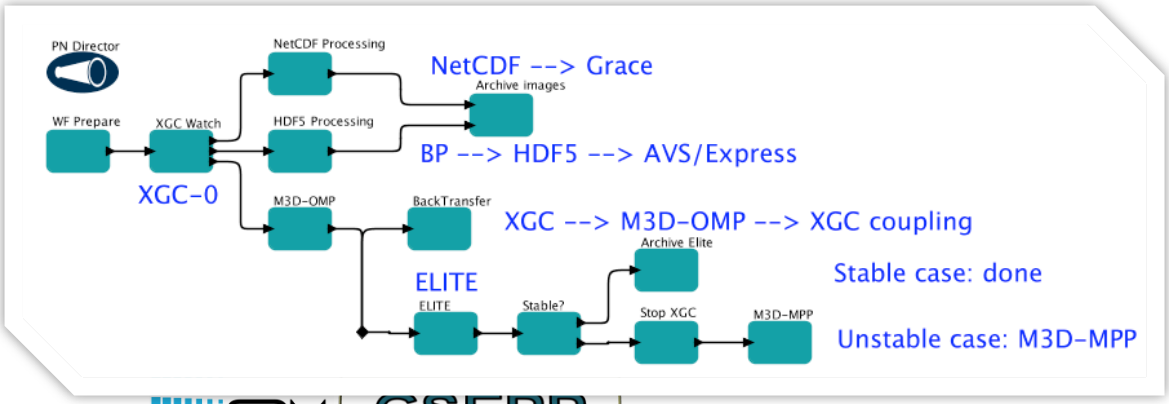
- Simple API for Fortran and C nearly as simple as Fortran standard IO for both read and write
- Both synchronous and asynchronous transports supported without code changes. Change IO method by changing XML file only!
- Free hooks into visualization and workflow systems through the data flows.
- Optimized IO implementations provided for each transport method (e.g., MPI-IO, HDF-5, pnetCDF, etc.)
- Binary, tagged format provided by default (including support for data paths and attributes).

ADIOS PHILOSOPHY (END USER)

- Simple API very similar to standard Fortran or C POSIX IO calls.
  - As close to identical as possible for C and Fortran API
  - get_type, open, read/write, close is the core
  - group_by, set_path, end_iteration, begin/end_computation, init/ finalize are the auxiliaries
- No changes in the API for different transport methods.
- Metadata and configuration defined in an external XML file parsed once on startup.
  - Describe the various IO grouping including attributes and hierarchical path structures for elements as a datatype
  - Define the transport method used for each datatype and give parameters for communication/writing/reading
  - Change on a per element basis what is written
  - Change on a per datatype basis how the IO is handled

KEPLER WORKFLOW AUTOMATION.

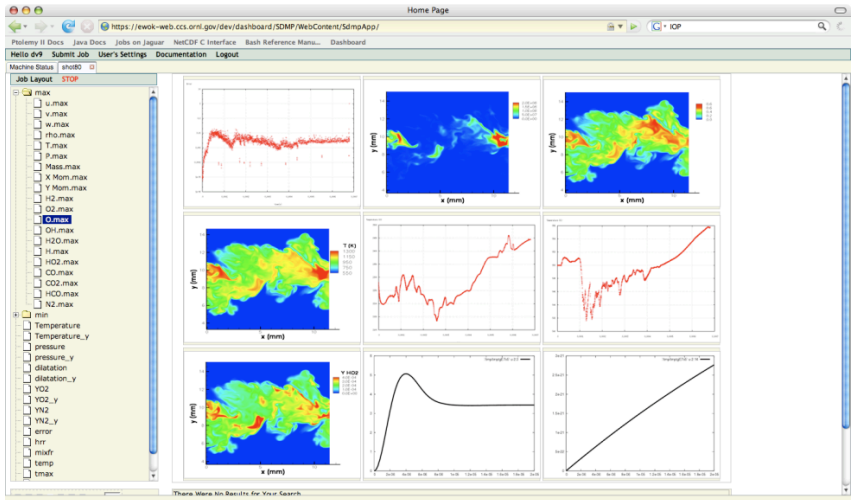
- Application design and actual program are the same
- Parallel execution of independent actors
- Pipeline parallel processing on a stream of data
- Restartable workflow
  - added value by the actors designed for that

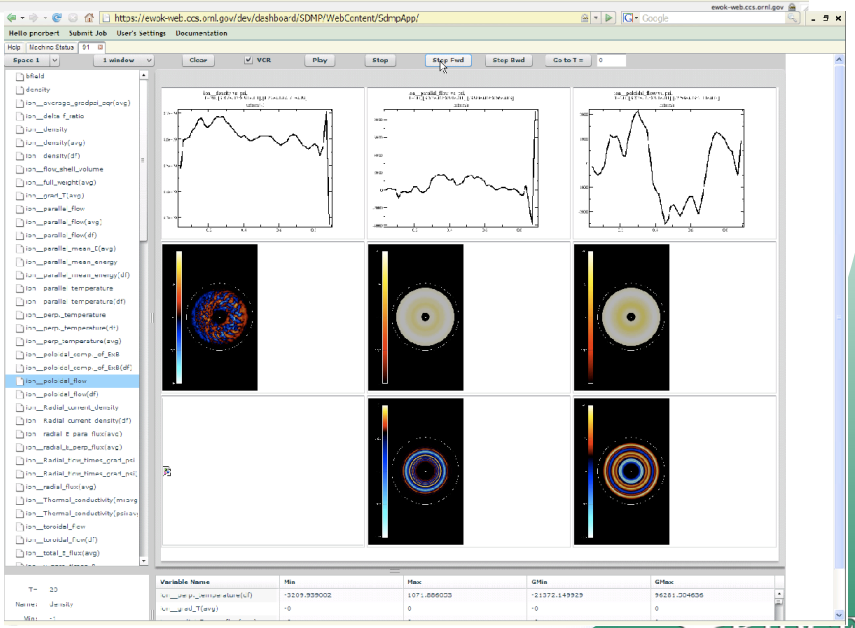


SDM CENTER CSEPP

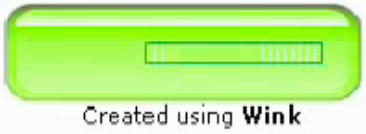
ORNL/CPES/SDM DASHBOARD.

- Designing a basic browser-based visualization tool.
- Local interaction for line graphs, 2D slices, particles
- Server used for data manipulation, extraction
- Server used for more complex visualizations
- “Google maps” for your data.
- Allows for pan and zoom locally
- Asynchronously queries



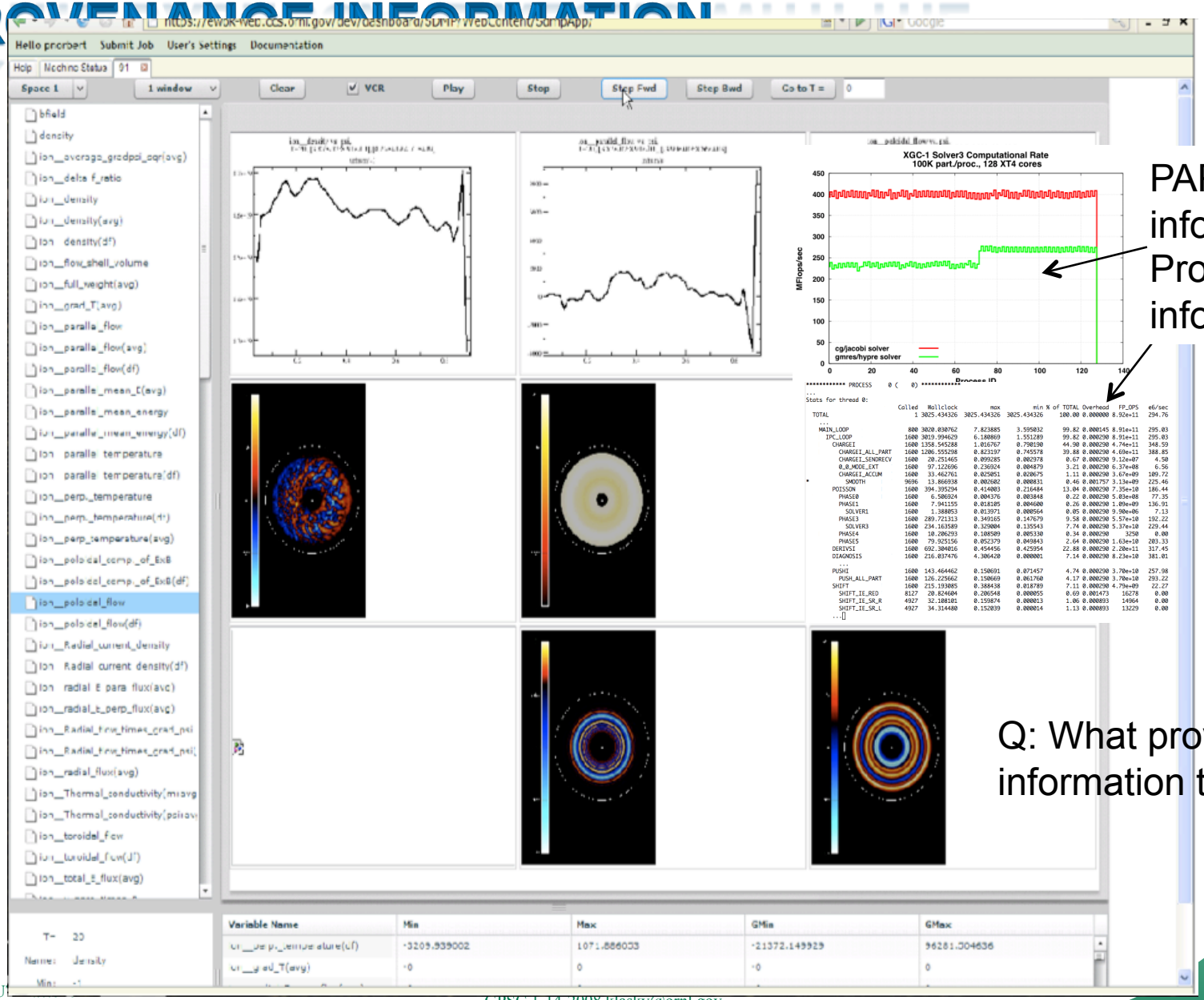


DASHBOARD MOVIE



DASHBOARD POST PROCESSING.

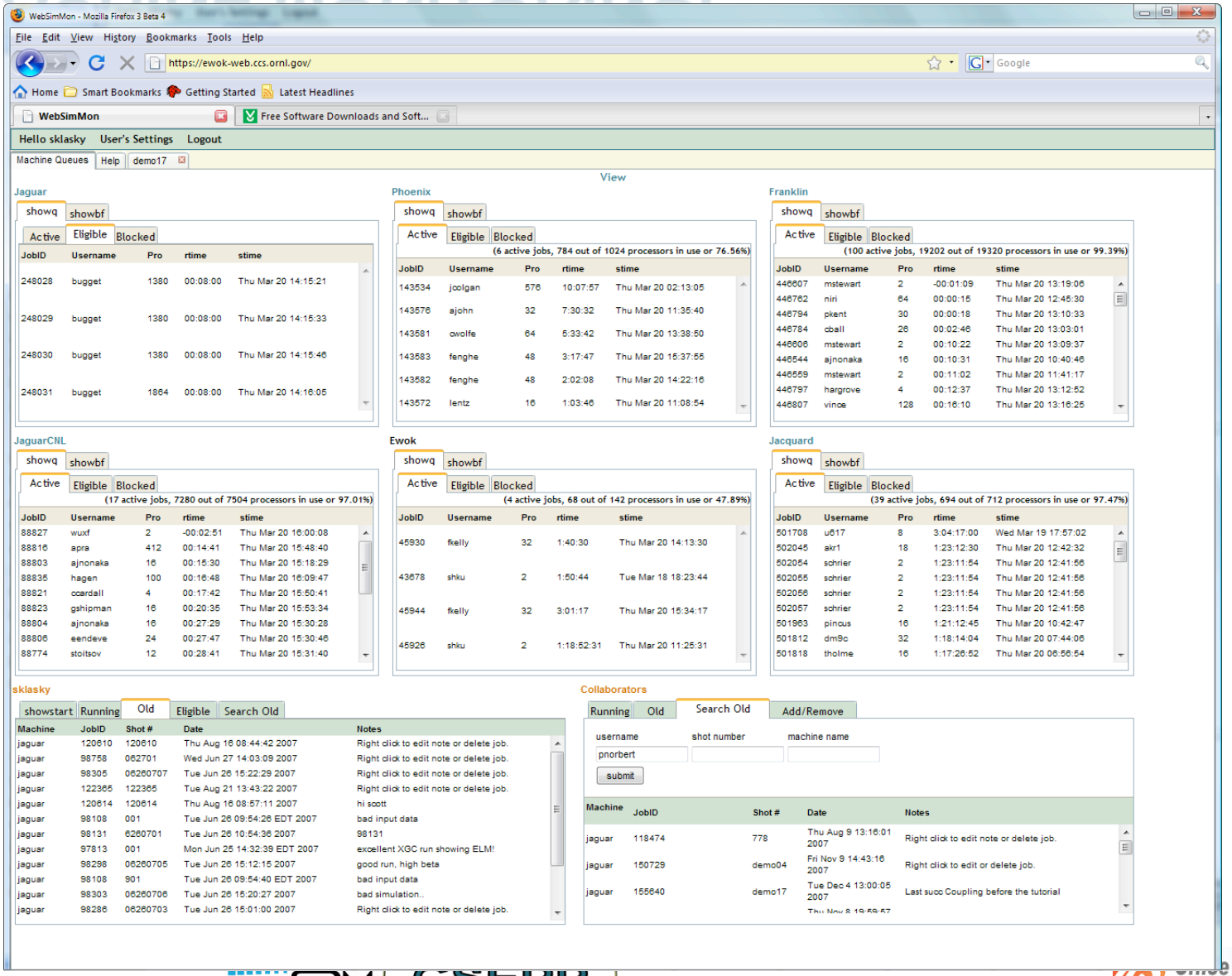
PERFORMANCE CHARACTERIZATION INTEGRATED IN THE DASHBOARD ALONG WITH THE PROVENANCE INFORMATION



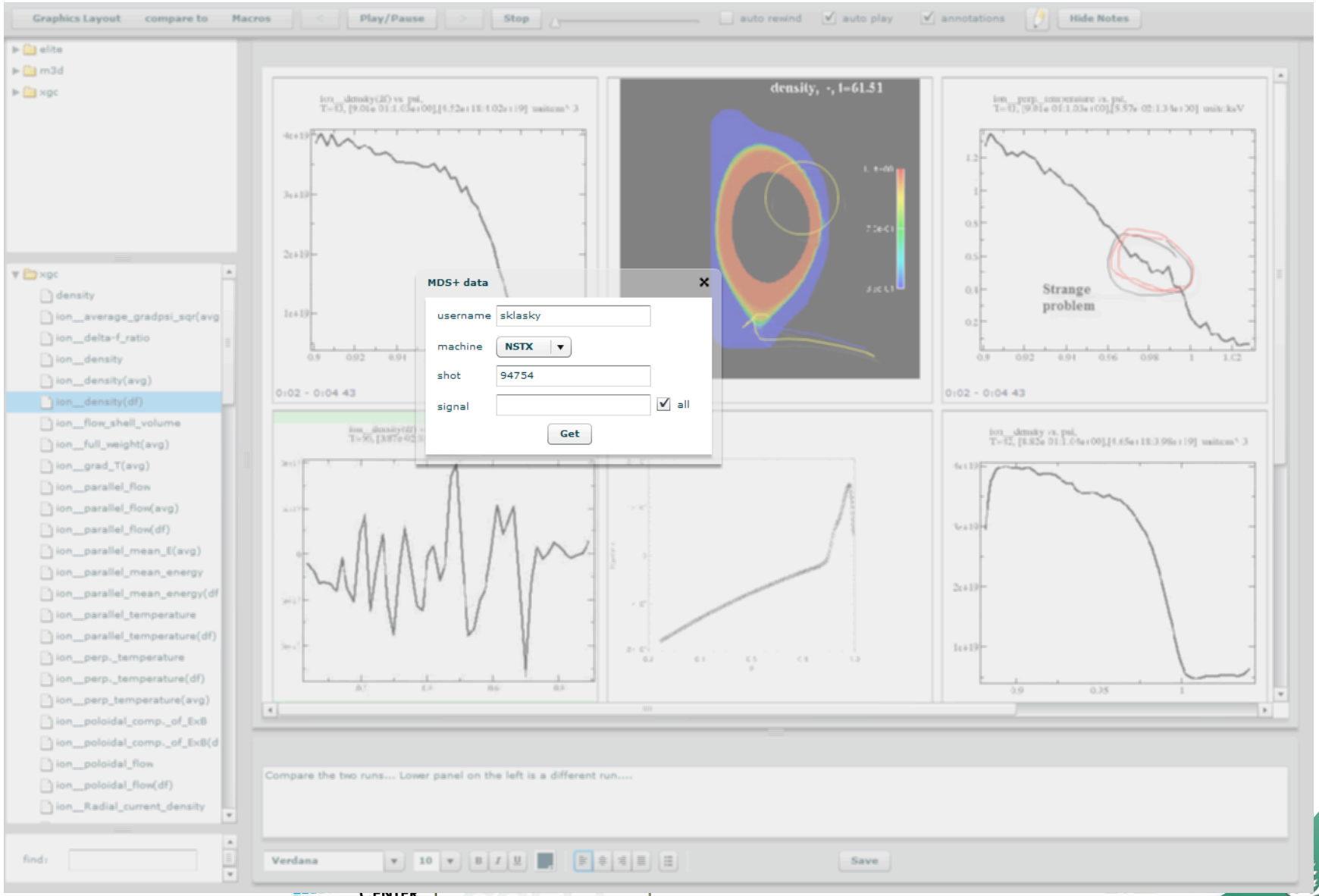
PAPI information
Provenance information

Q: What provenance information to keep?

MACHINE MONITORING.



EVENTUAL HOOKS INTO MDS+ FOR EXPERIMENTAL COMPARISON.



FUTURE WORK FOR YOUR PROJECT

- It will be hard to get an INCITE.
  - Must get good PR!
- We will ADIOS up your code(s).
- We will hook codes with our monitoring workflow.
- Why we want to work with you?
  - We need your feedback on our projects.
  - Helps harden our routines, and extend them.