

# END TO END COMPUTING TECHNOLOGY AT ORNL

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

March 29, 2008

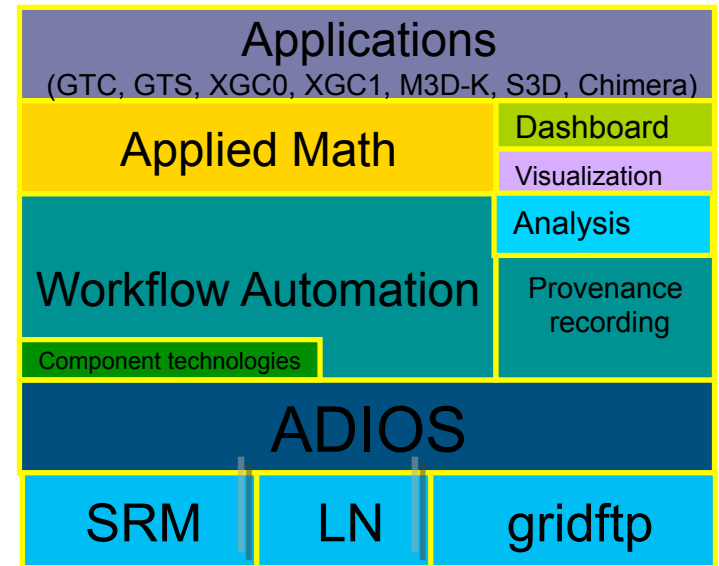
Scott Klasky

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

# 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.

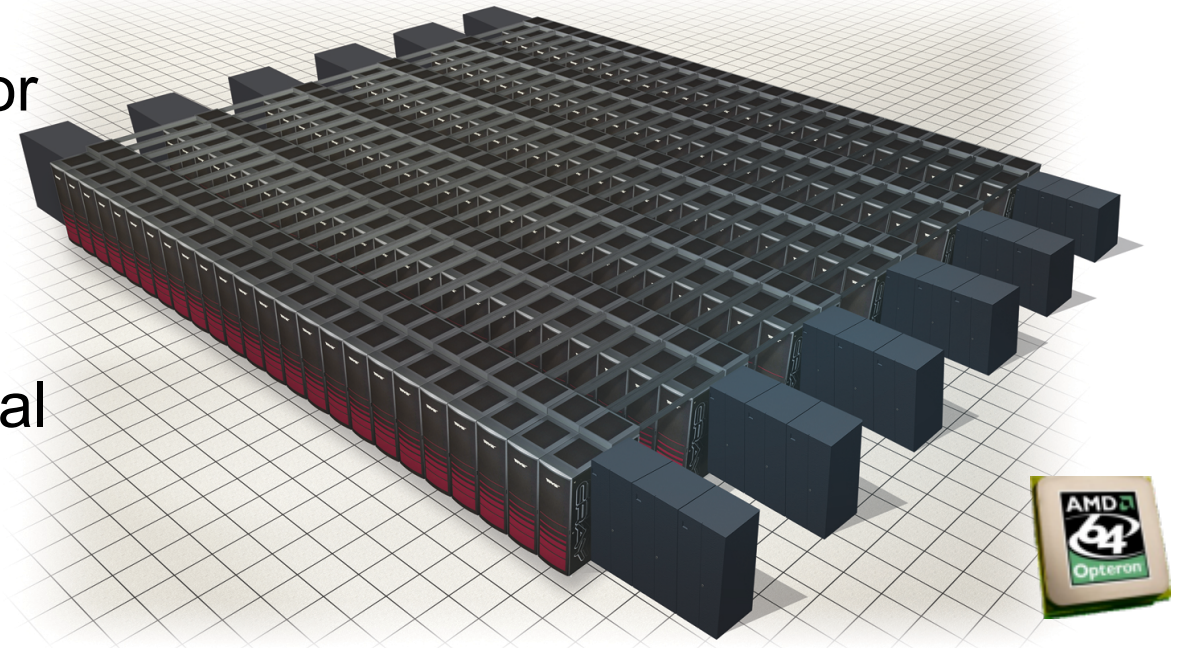


- **Basic Idea:** place **highly annotated fast, easy to use** I/O methods in the code, which can be **monitored**, and **controlled**, have a **workflow** engine record all of the information, **visualize** this on a **dashboard** and allow **collaborators** easy access to data. Have everything report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



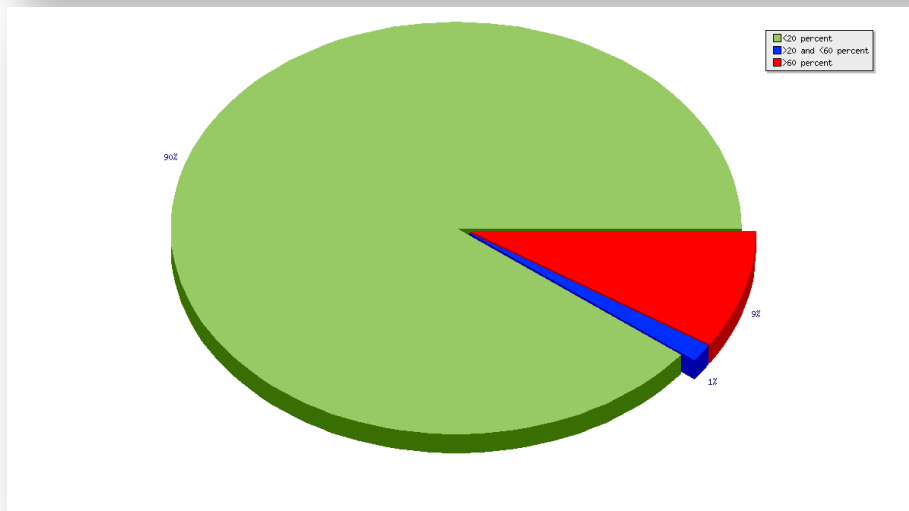
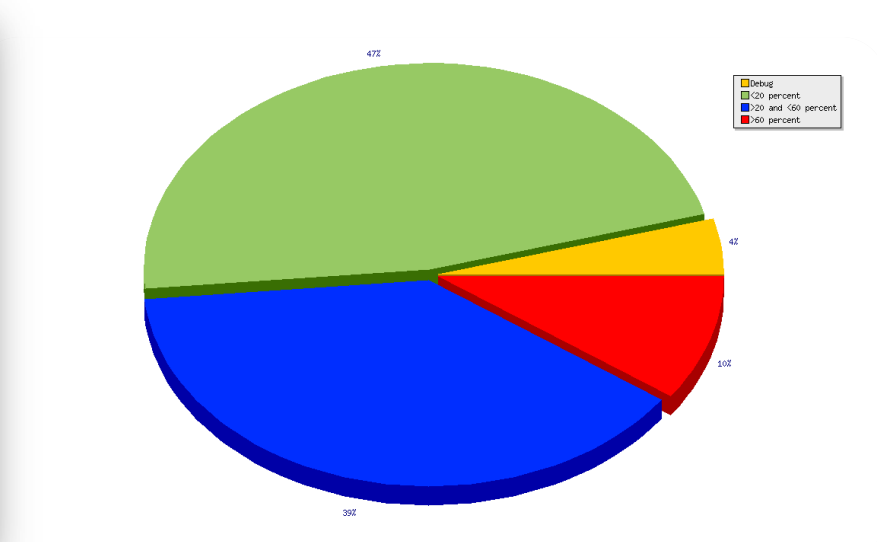
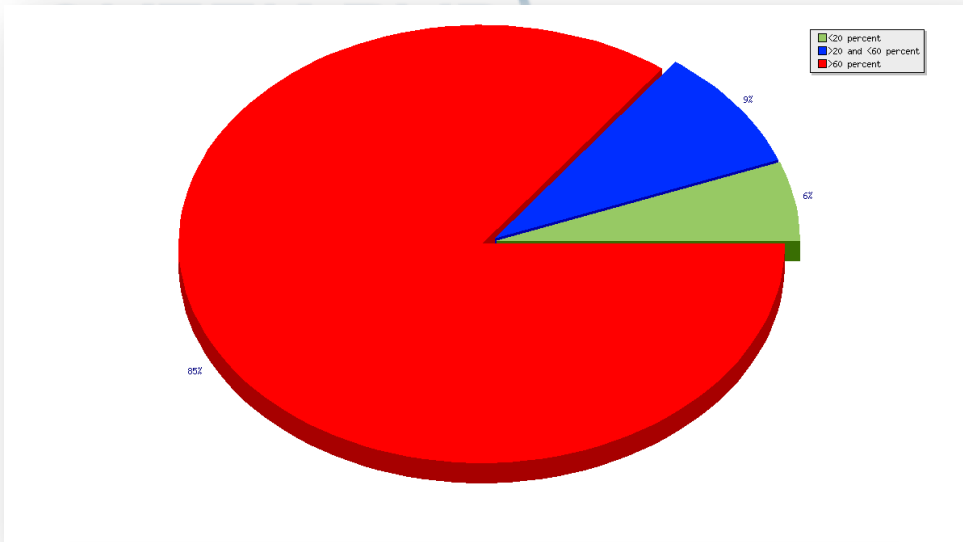
# FUSION INCITE 2007: BIGGEST USERS

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



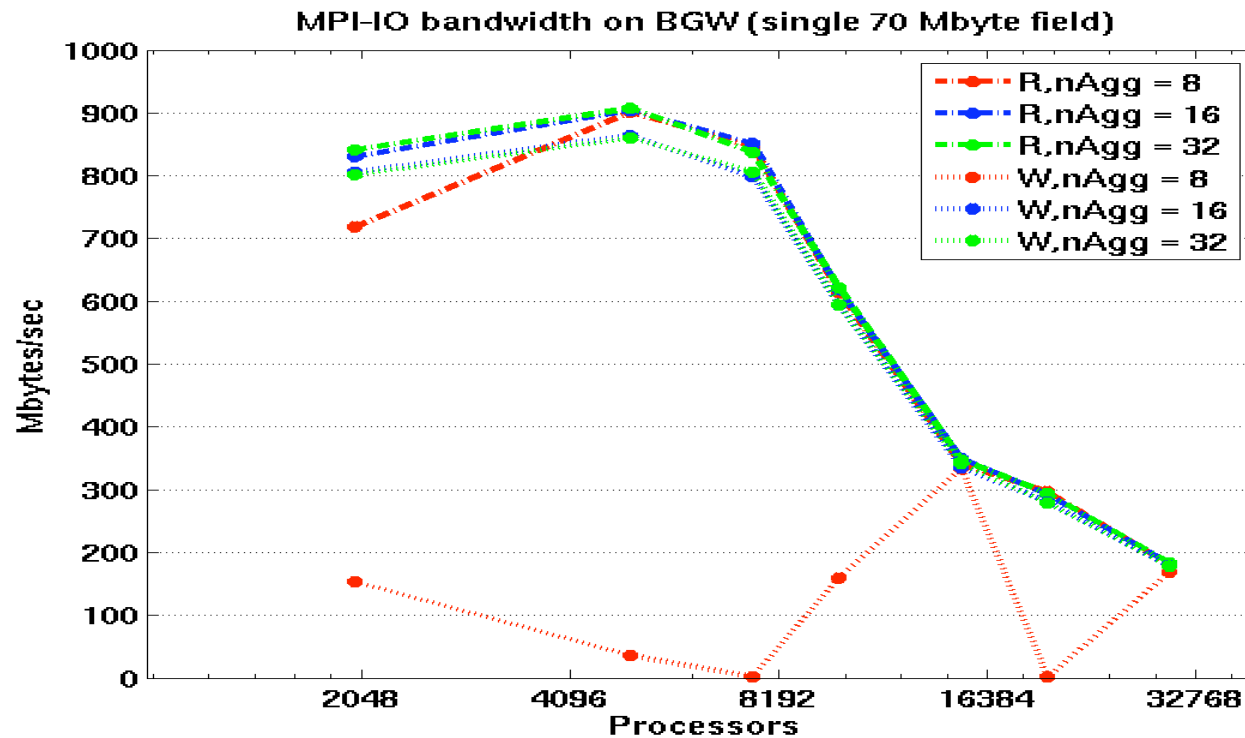
# CSEPP BEST WAY TO GET TIME.

- 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.



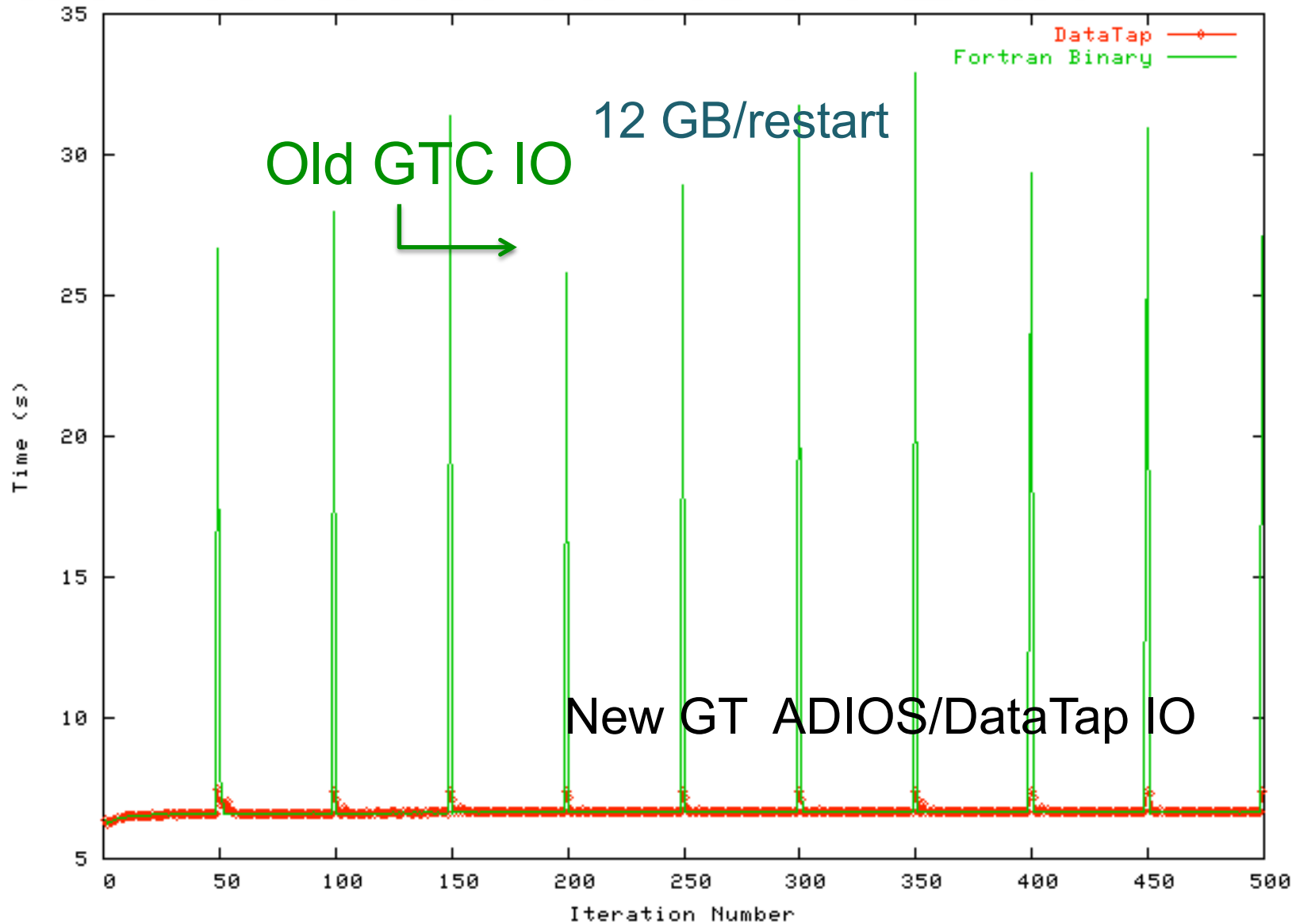
# ORNL END TO END TECHNOLOGIES

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



# IO PROGRESS & PLANS

Iteration Time for 128 Nodes (micell=200, npartdom=2)

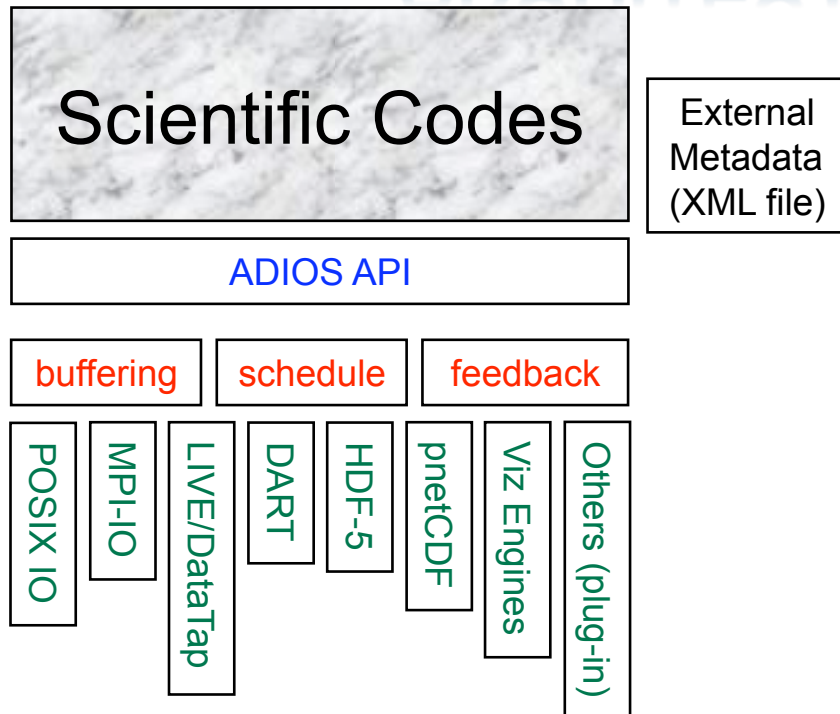


# ADAPTABLE IO SYSTEM (ADIOS)

- 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			% overhead					
DART Jaguar	1.2TB <1							
Datatap/jaguar								
Maviz/jaguar								
Visit/jaguar								
Paraview/jaguar								
Phdf5/jaguar								
Pnetcdf/jaguar								
BGP/IB/GPFS..								

# ADIOS ARCHITECTURE



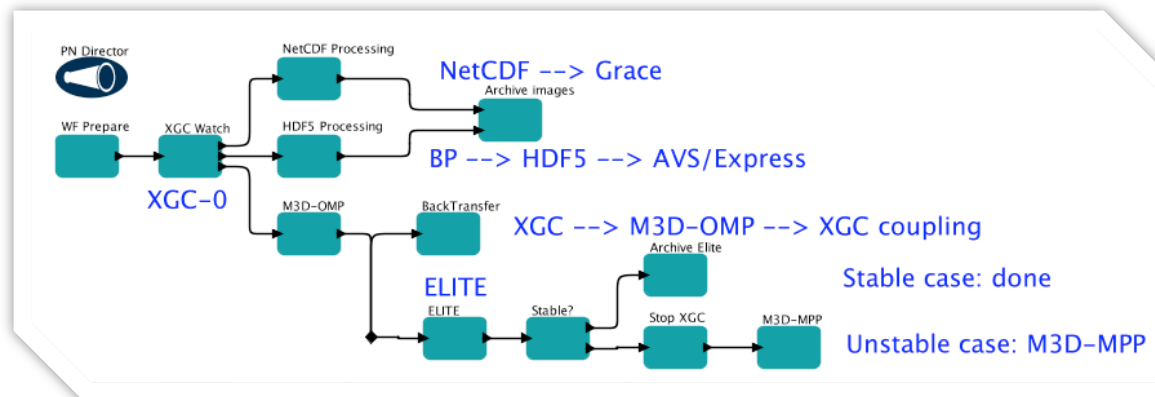
- Simple API for Fortran and C nearly as simple as Fortran standard IO for both read and write
- Both synchronous and asynchronous transports supported without code changes. Change IO method by changing XML file only!
- Free hooks into visualization and workflow systems through the data flows.
- Optimized 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

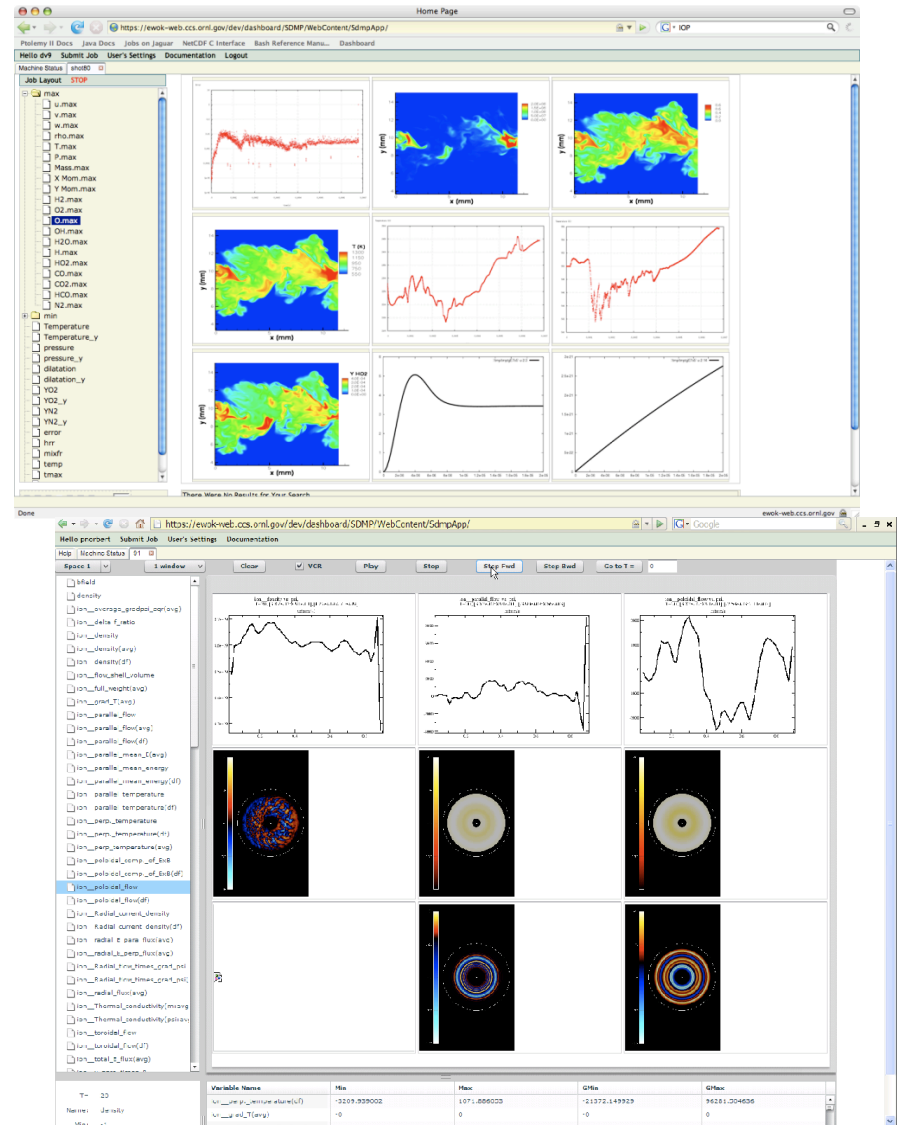
# KEPLER WORKFLOW AUTOMATION.

- 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



# 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



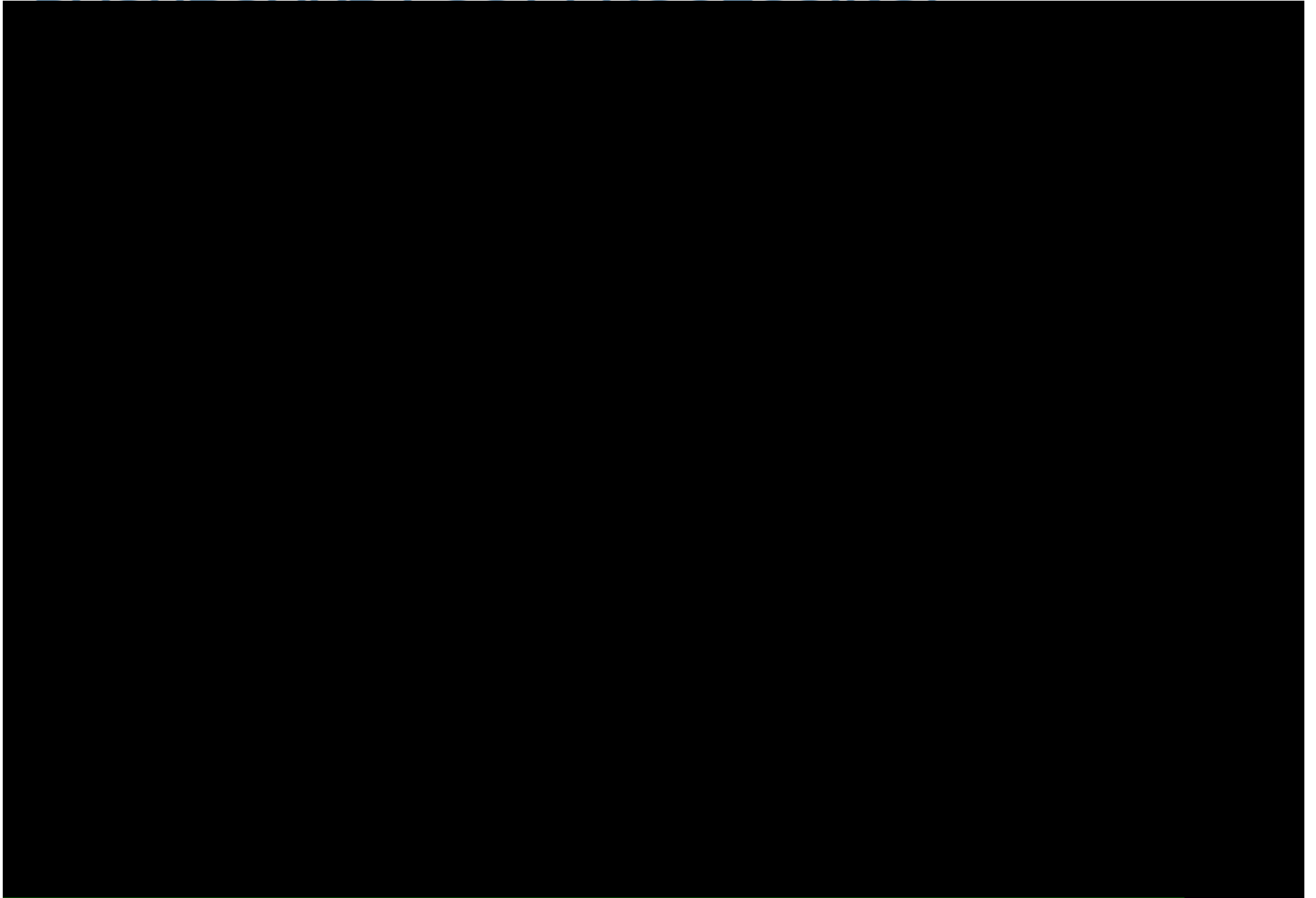


# DASHBOARD MOVIE

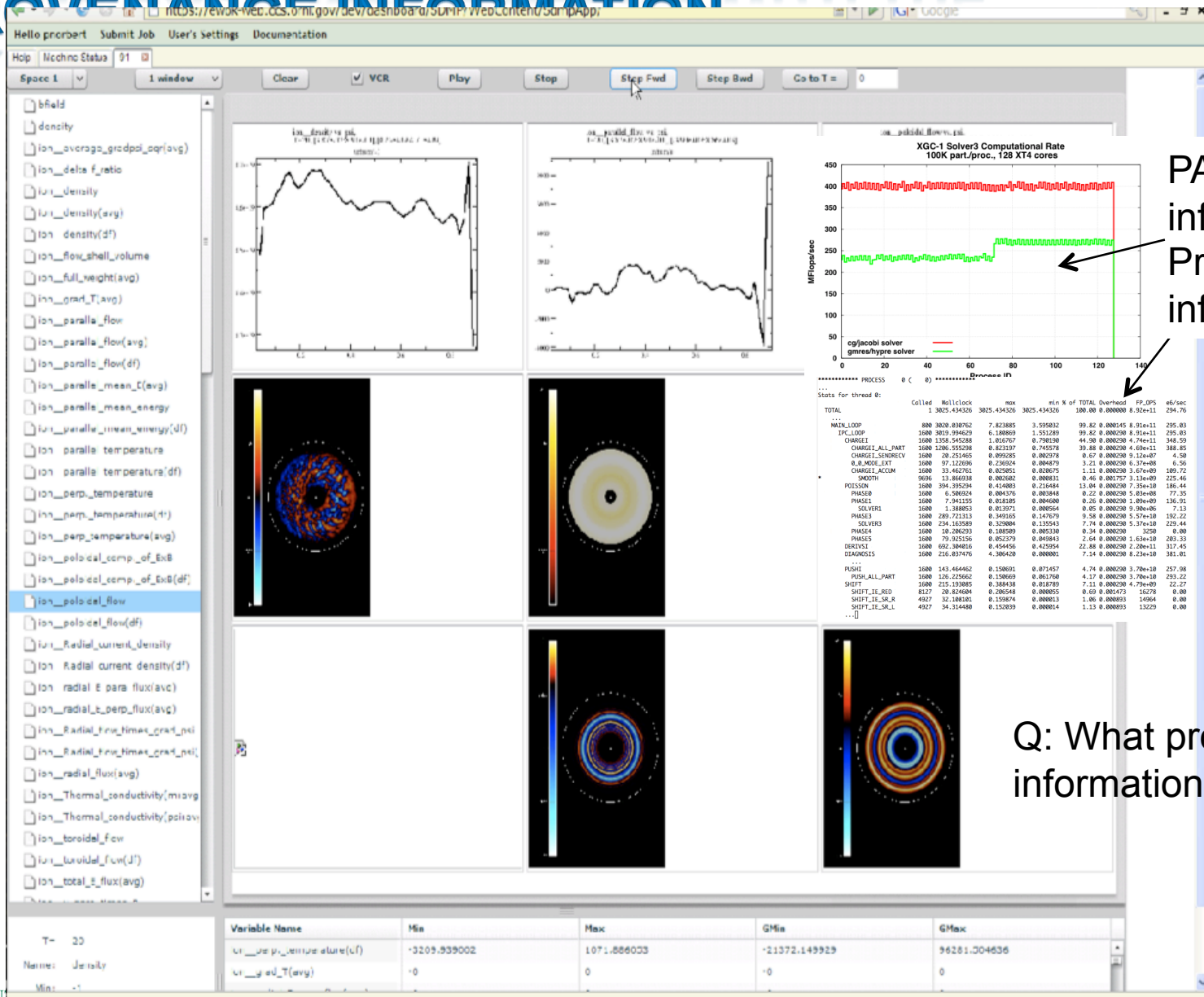


Created using **Wink**

# 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.

WebSimMon - Mozilla Firefox 3 Beta 4  
 https://ewok-web.ccs.ornl.gov/

Hello sklasky User's Settings Logout

Machine Queues Help demo17

View

**Jaguar**

showq showbf

Active	Eligible	Blocked		
JobID	Username	Pro	rtime	stime
248028	bugget	1380	00:08:00	Thu Mar 20 14:15:21
248029	bugget	1380	00:08:00	Thu Mar 20 14:15:33
248030	bugget	1380	00:08:00	Thu Mar 20 14:15:46
248031	bugget	1864	00:08:00	Thu Mar 20 14:16:05

**Phoenix**

showq showbf

(6 active jobs, 784 out of 1024 processors in use or 76.56%)

Active	Eligible	Blocked		
JobID	Username	Pro	rtime	stime
143534	joolgan	576	10:07:57	Thu Mar 20 02:13:05
143576	ajohn	32	7:30:32	Thu Mar 20 11:35:40
143581	owolfe	64	5:33:42	Thu Mar 20 13:38:50
143583	fenghe	48	3:17:47	Thu Mar 20 15:37:55
143582	fenghe	48	2:02:08	Thu Mar 20 14:22:16
143572	lentz	16	1:03:46	Thu Mar 20 11:08:54

**Franklin**

showq showbf

(100 active jobs, 19202 out of 19320 processors in use or 99.39%)

Active	Eligible	Blocked		
JobID	Username	Pro	rtime	stime
446907	mstewart	2	-00:01:09	Thu Mar 20 13:19:06
446762	niri	64	00:00:15	Thu Mar 20 12:45:30
446794	pkent	30	00:00:18	Thu Mar 20 13:10:33
446784	cball	26	00:02:46	Thu Mar 20 13:03:01
446806	mstewart	2	00:10:22	Thu Mar 20 13:09:37
446544	ajnonaka	16	00:10:31	Thu Mar 20 10:40:46
446559	mstewart	2	00:11:02	Thu Mar 20 11:41:17
446797	hargrove	4	00:12:37	Thu Mar 20 13:12:52
446807	vince	128	00:16:10	Thu Mar 20 13:16:25

**JaguarCNL**

showq showbf

(17 active jobs, 7280 out of 7504 processors in use or 97.01%)

Active	Eligible	Blocked		
JobID	Username	Pro	rtime	stime
88827	wuxf	2	-00:02:51	Thu Mar 20 16:00:08
88816	apra	412	00:14:41	Thu Mar 20 15:48:40
88803	ajnonaka	16	00:15:30	Thu Mar 20 15:18:29
88835	hagen	100	00:16:48	Thu Mar 20 16:09:47
88821	coardall	4	00:17:42	Thu Mar 20 15:50:41
88823	gshipman	16	00:20:35	Thu Mar 20 15:53:34
88804	ajnonaka	16	00:27:29	Thu Mar 20 15:30:28
88806	eendev	24	00:27:47	Thu Mar 20 15:30:46
88774	stoitsov	12	00:28:41	Thu Mar 20 15:31:40

**Ewok**

showq showbf

(4 active jobs, 68 out of 142 processors in use or 47.89%)

Active	Eligible	Blocked		
JobID	Username	Pro	rtime	stime
46930	fkelly	32	1:40:30	Thu Mar 20 14:13:30
43678	shku	2	1:50:44	Tue Mar 18 18:23:44
46944	fkelly	32	3:01:17	Thu Mar 20 15:34:17
45926	shku	2	1:18:52:31	Thu Mar 20 11:25:31

**Jacquard**

showq showbf

(39 active jobs, 694 out of 712 processors in use or 97.47%)

Active	Eligible	Blocked		
JobID	Username	Pro	rtime	stime
501708	u617	8	3:04:17:00	Wed Mar 19 17:57:02
502045	akr1	18	1:23:12:30	Thu Mar 20 12:42:32
502054	schrier	2	1:23:11:54	Thu Mar 20 12:41:56
502055	schrier	2	1:23:11:54	Thu Mar 20 12:41:56
502056	schrier	2	1:23:11:54	Thu Mar 20 12:41:56
502057	schrier	2	1:23:11:54	Thu Mar 20 12:41:56
501963	pinous	16	1:21:12:45	Thu Mar 20 10:42:47
501812	dm9c	32	1:18:14:04	Thu Mar 20 07:44:06
501818	tholme	16	1:17:26:52	Thu Mar 20 06:56:54

**sklasky**

showstart Running Old Eligible Search Old

Machine	JobID	Shot #	Date	Notes
jaguar	120610	120610	Thu Aug 16 08:44:42 2007	Right click to edit note or delete job.
jaguar	98758	062701	Wed Jun 27 14:03:09 2007	Right click to edit note or delete job.
jaguar	98305	06260707	Tue Jun 26 15:22:29 2007	Right click to edit note or delete job.
jaguar	122365	122365	Tue Aug 21 13:43:22 2007	Right click to edit note or delete job.
jaguar	120614	120614	Thu Aug 16 08:57:11 2007	hi scott
jaguar	98108	001	Tue Jun 26 09:54:26 EDT 2007	bad input data
jaguar	98131	6260701	Tue Jun 26 10:54:36 2007	98131
jaguar	97813	001	Mon Jun 25 14:32:39 EDT 2007	excellent XGC run showing ELM!
jaguar	98298	06260705	Tue Jun 26 15:12:15 2007	good run, high beta
jaguar	98108	901	Tue Jun 26 09:54:40 EDT 2007	bad input data
jaguar	98303	06260706	Tue Jun 26 15:20:27 2007	bad simulation..
jaguar	98286	06260703	Tue Jun 26 15:01:00 2007	Right click to edit note or delete job.

**Collaborators**

Running Old Search Old Add/Remove

username shot number machine name

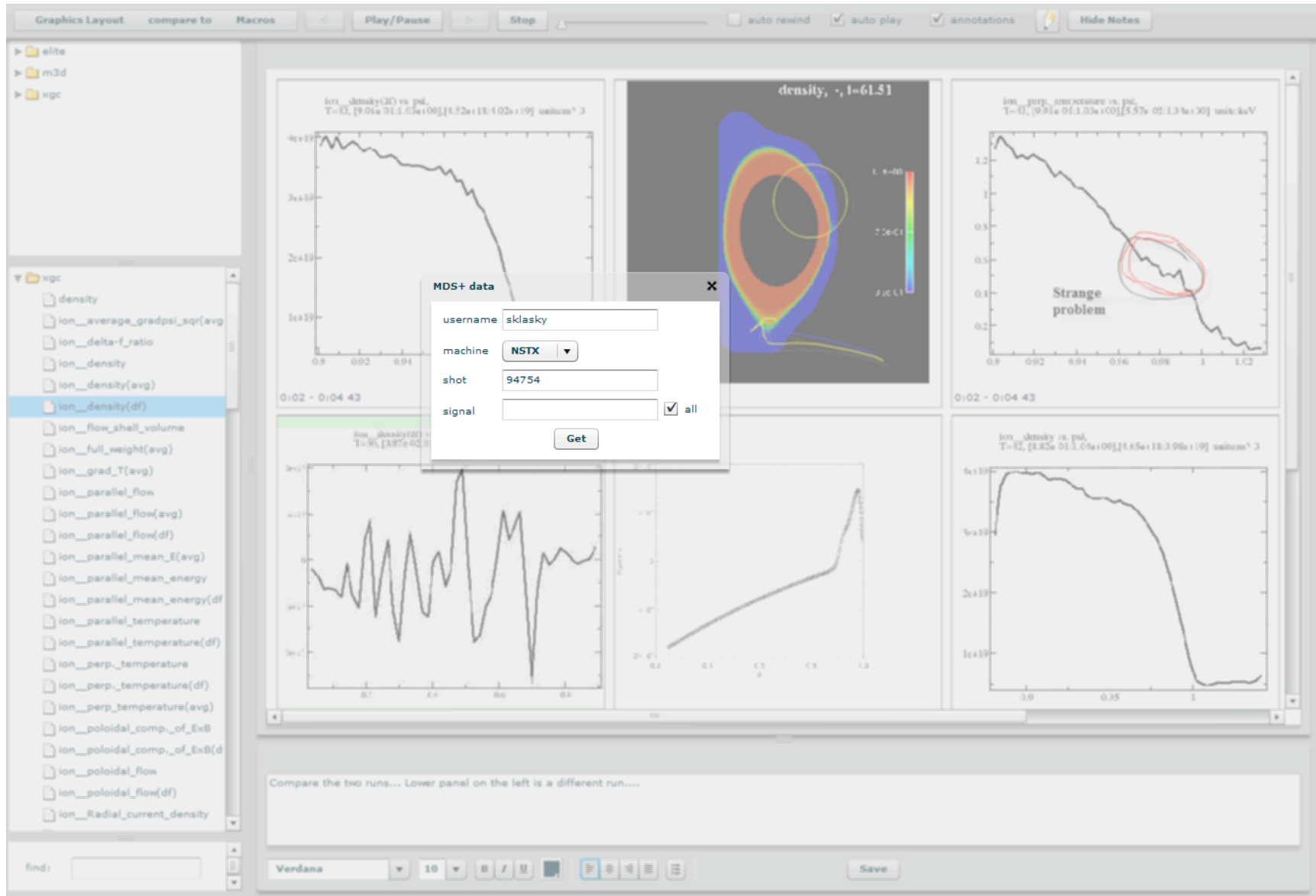
pnorbert

submit

Machine	JobID	Shot #	Date	Notes
jaguar	118474	778	Thu Aug 9 13:16:01 2007	Right click to edit note or delete job.
jaguar	150729	demo04	Fri Nov 9 14:43:16 2007	Right click to edit or delete job.
jaguar	155640	demo17	Tue Dec 4 13:00:05 2007	Last suoc Coupling before the tutorial



# 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.