Scalable, Automated Parallel Performance Analysis with TAU, PerfDMF and PerfExplorer

Kevin A. Huck, Allen D. Malony,
Sameer Shende, Alan Morris
khuck, malony, sameer, amorris@cs.uoregon.edu
http://www.cs.uoregon.edu/research/tau
University of Oregon
Department of Computer and Information Science
Performance Research Laboratory

Performance Research Lab

- Prof. Allen D. Malony
- Sameer Shende
- Alan Morris
- Wyatt Spear
- Scott Biersdorf
- Aroon Nataraj
- Kevin A. Huck
- http://www.cs.uoregon.edu/research/tau/

UNIVERSITY OF OREGON

Tuning and Analysis Utilities

P • R • L Performance Research Lab

Overview

- TAU Introduction (brief)
- PerfDMF Introduction
- PerfExplorer Introduction
- PerfExplorer Ongoing Analysis Examples
- Summary

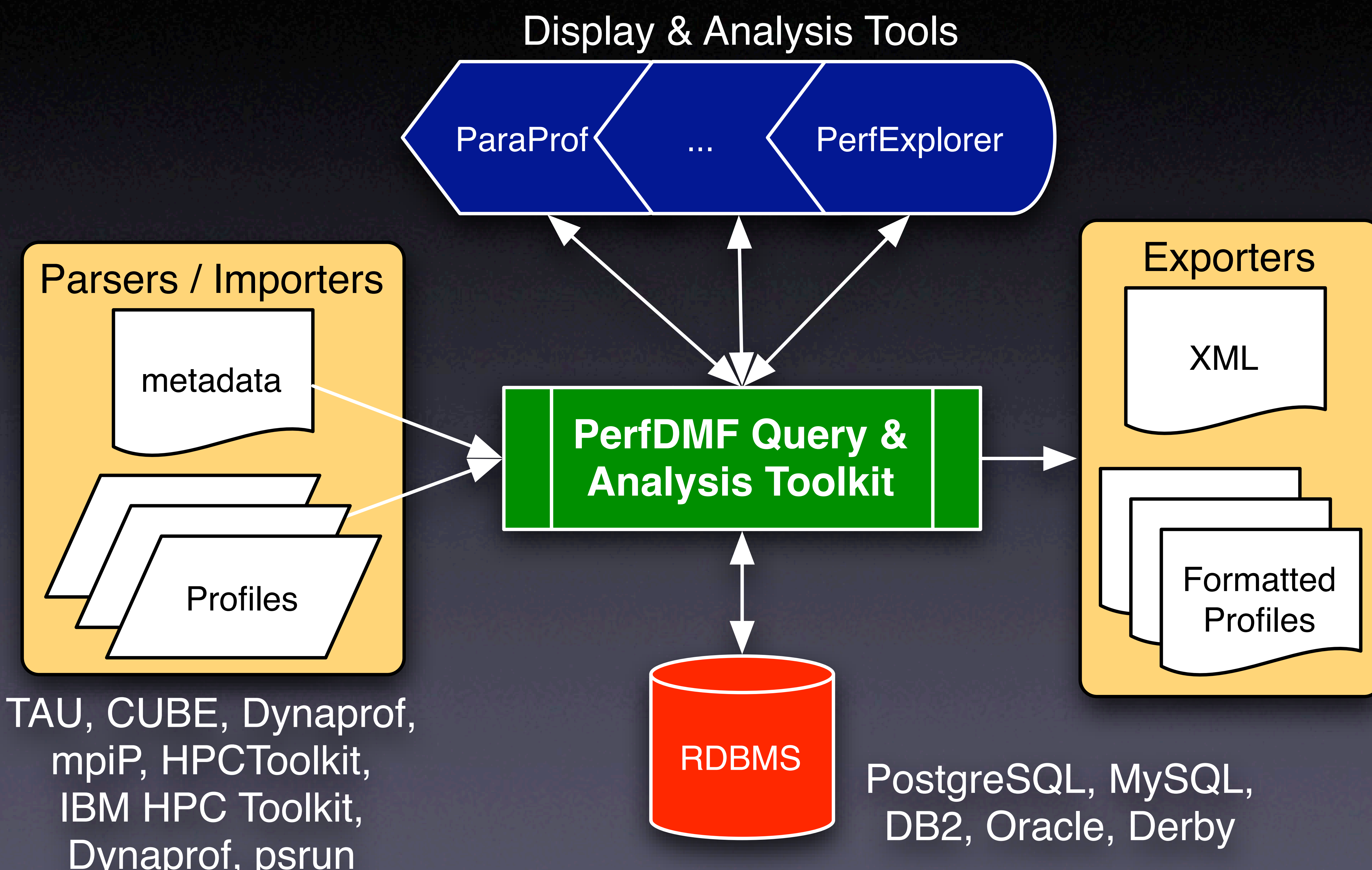
TAU Performance System

- ***T***uning and ***A***nalysis ***U***tilities (15+ year project effort)
- Performance system framework for HPC systems
  - Integrated, scalable, flexible and parallel
- Targets a general complex system computation model
  - Entities: nodes / contexts / threads
  - Multi-level: system / software / parallelism
  - Measurement and analysis abstraction
- Integraged toolkit for performance problem solving
  - Instrumentation, measurement, analysis and visualization
  - Portable performance profiling and tracing utility
  - Performance data management and data mining
- ***Partners***: Forshungszentrum Jülich, LLNL, ANL, LANL

PerfDMF

- Performance Data Management Framework
- Originally designed to address critical TAU requirements
- Broader goal is to provide an open, flexible framework to support common data management tasks
- Extensible toolkit to promote integration and reuse across available performance tools
  - Supported profile formats:
    TAU, CUBE, Dynaprof, HPCToolkit (Rice), HPC Toolkit (IBM), gprof, mpiP, psrun (PerfSuite), OpenSpeedShop (in development), application
  - Supported DBMS:
    PostgreSQL, MySQL, Oracle, DB2, Derby/Cloudscape

PerfDMF Architecture



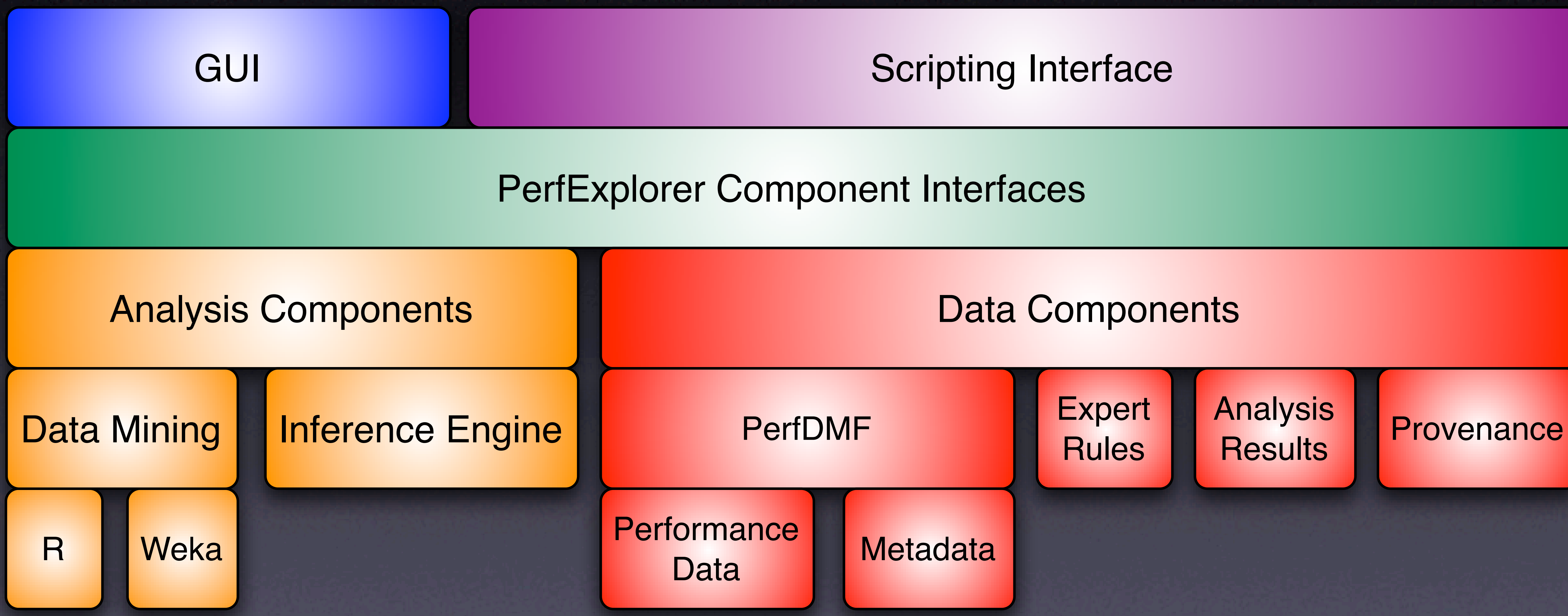
Recent PerfDMF Development

- Integration of XML metadata for each profile
  - Common profile attributes
  - Thread / process specific profile attributes
  - Automatic collection of runtime information
  - Any other data the user wants to collect can be added
    - build information
    - job submission information
  - Two methods for acquiring metadata:
    - TAU_METADATA() call from application
    - Optional XML file added when saving profile to PerfDMF
  - TAU Metadata XML schema is simple, easy to generate from scripting tools (no XML libraries required)

PerfExplorer

- Performance knowledge discovery framework
  - Data mining analysis applied to parallel performance data
    - comparisons, clustering, correlation, dimension reduction, ...
  - Uses the existing TAU infrastructure
    - TAU & other performance profiles, PerfDMF
- Technology Integration
  - Java API and toolkit for portability
  - R-project / Omegahat statistical analysis
  - Weka data mining package
  - JFreeChart for visualization, output (EPS, SVG, PNG)

PerfExplorer Design

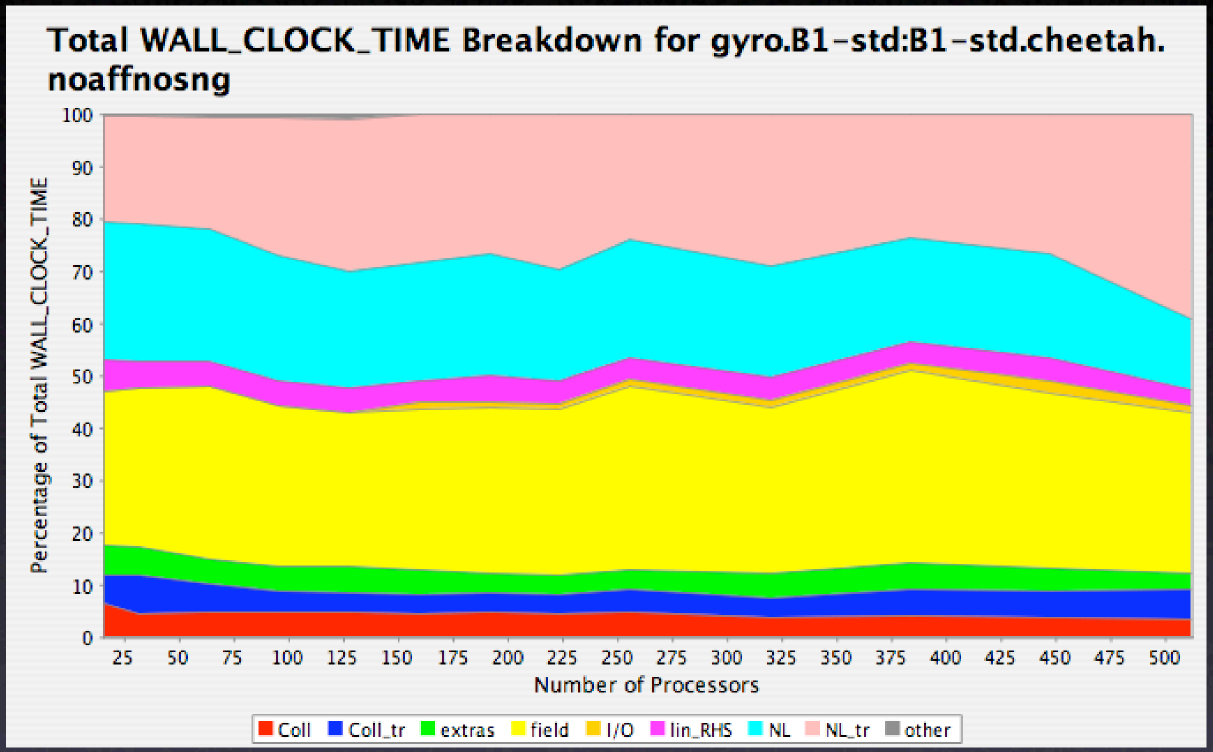


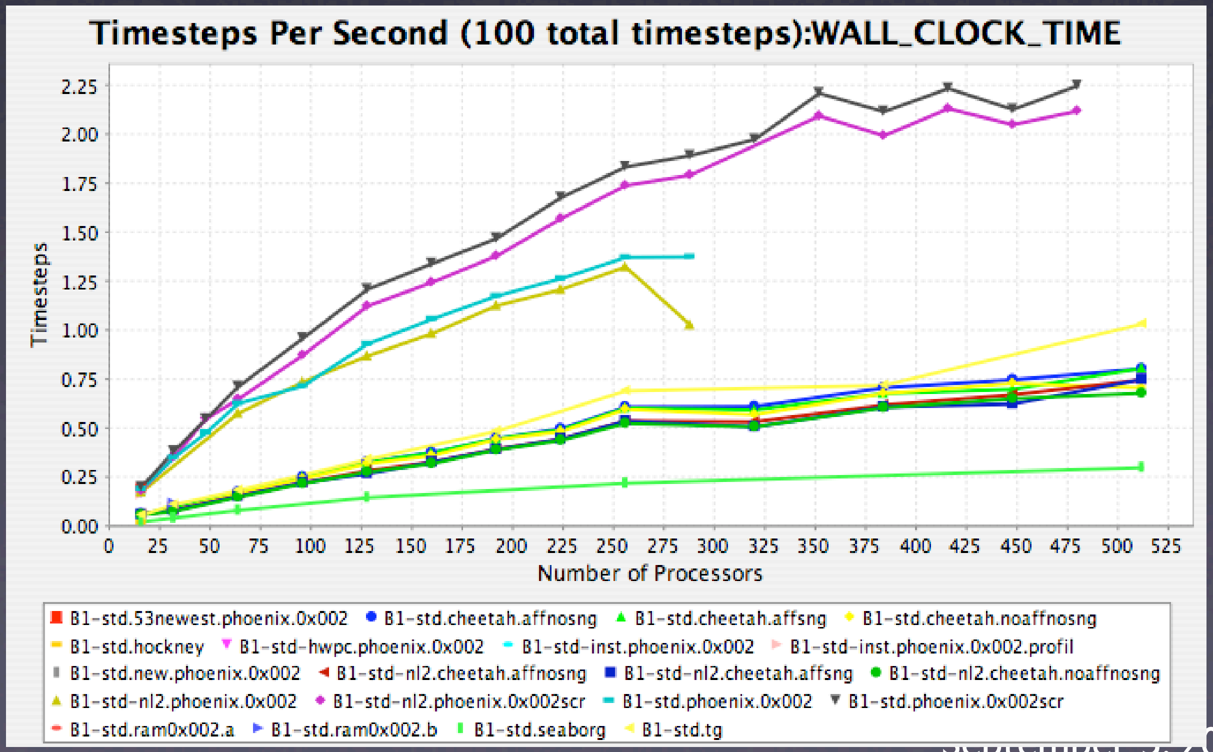
PerfExplorer Analysis Methods

- Data summaries, distributions, scatterplots
- Clustering
  - k-means
  - hierarchical
- Correlation analysis
- Dimension reduction
  - PCA
  - Random Linear Projection
  - Threshholds
- Comparative analysis
- Data management views

Relative Comparisons

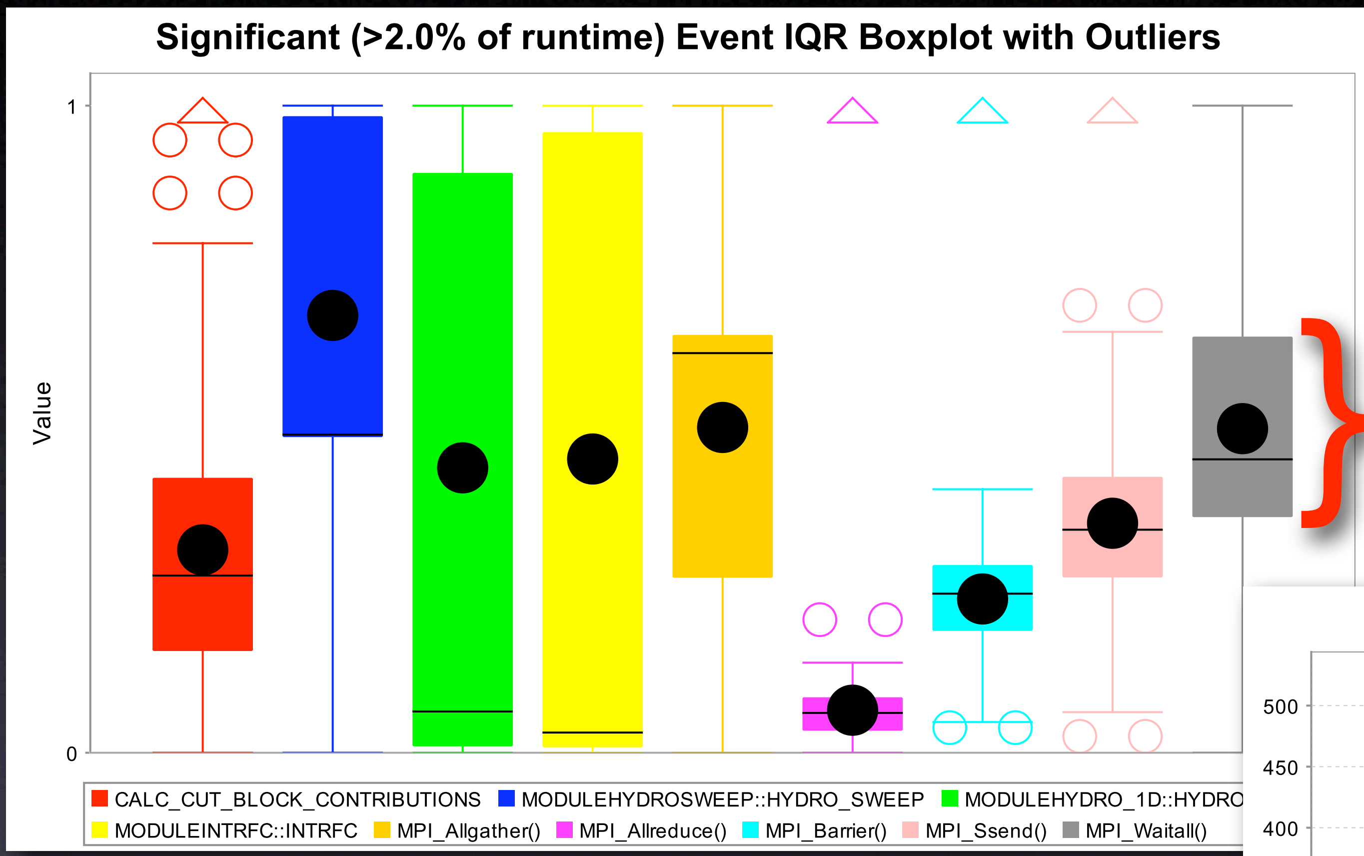
- Total execution time
- Timesteps per second
- Relative efficiency / speedup for total / per event
- Group fraction of total
- Runtime breakdown
- Phase comparisons





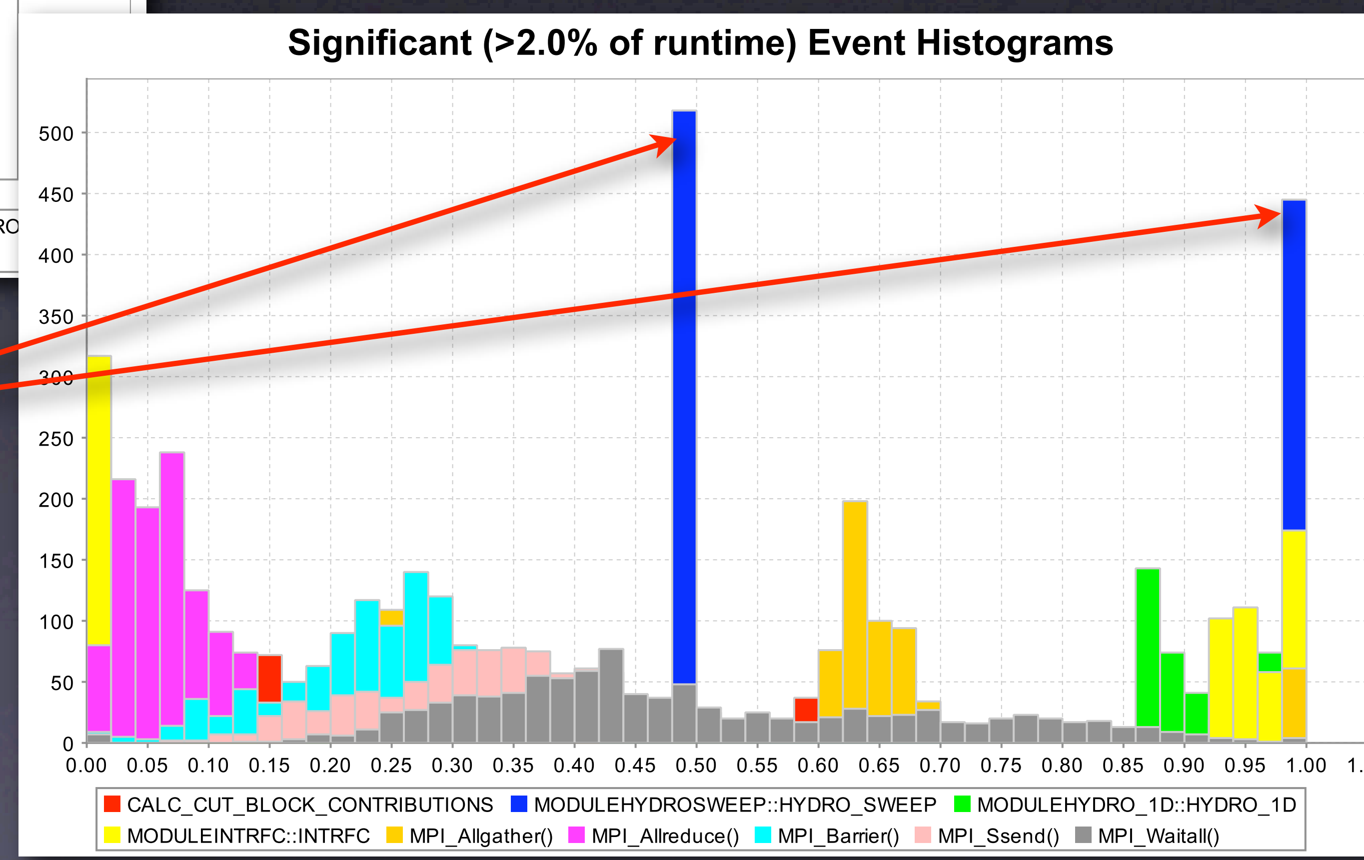
Data: GYRO on various machines

Distribution Visualization



Indicates min, Q1, mean, median, Q3, max, outliers

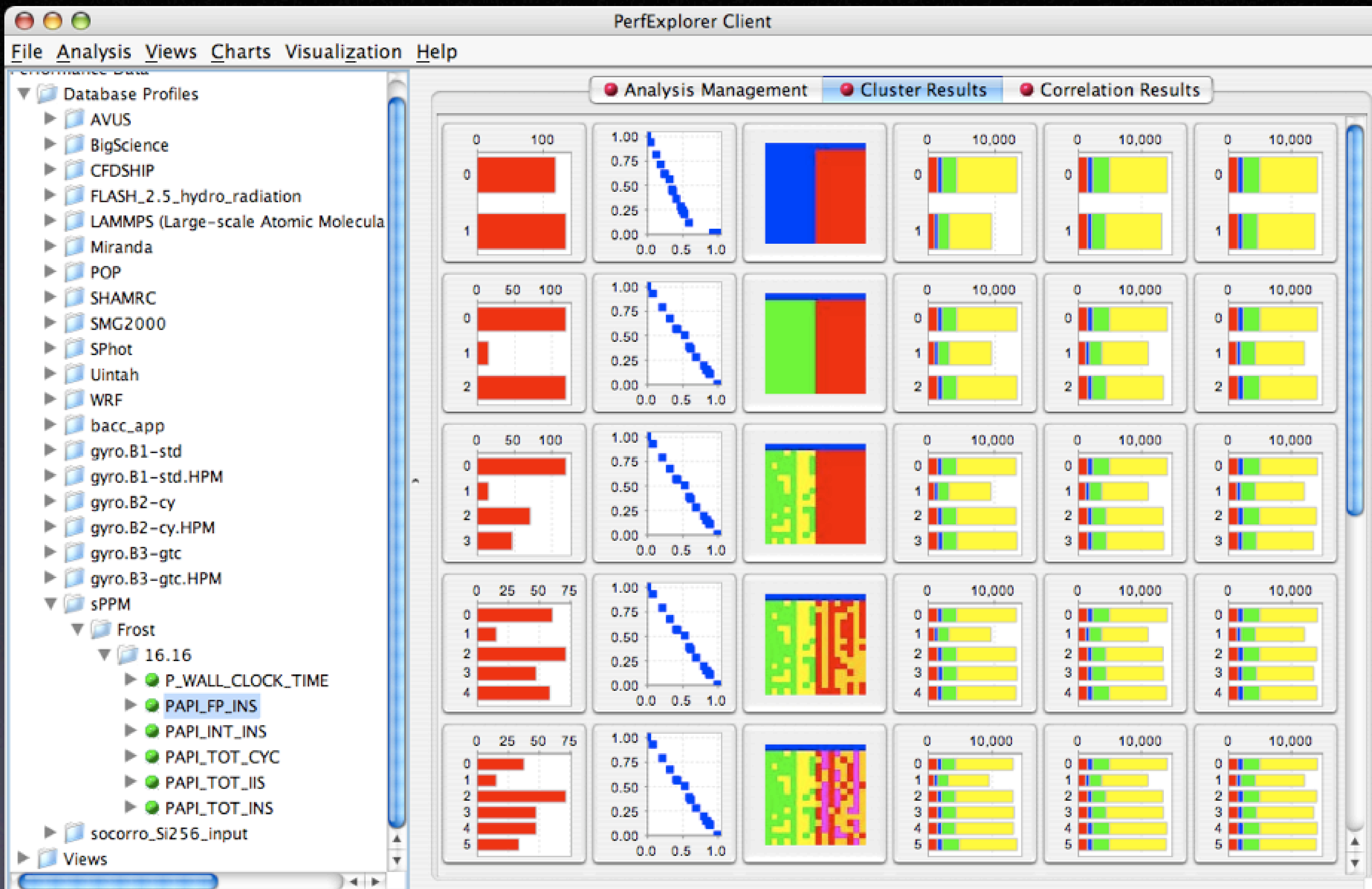
Visualizes Intra-Quartile Range (median 50%)



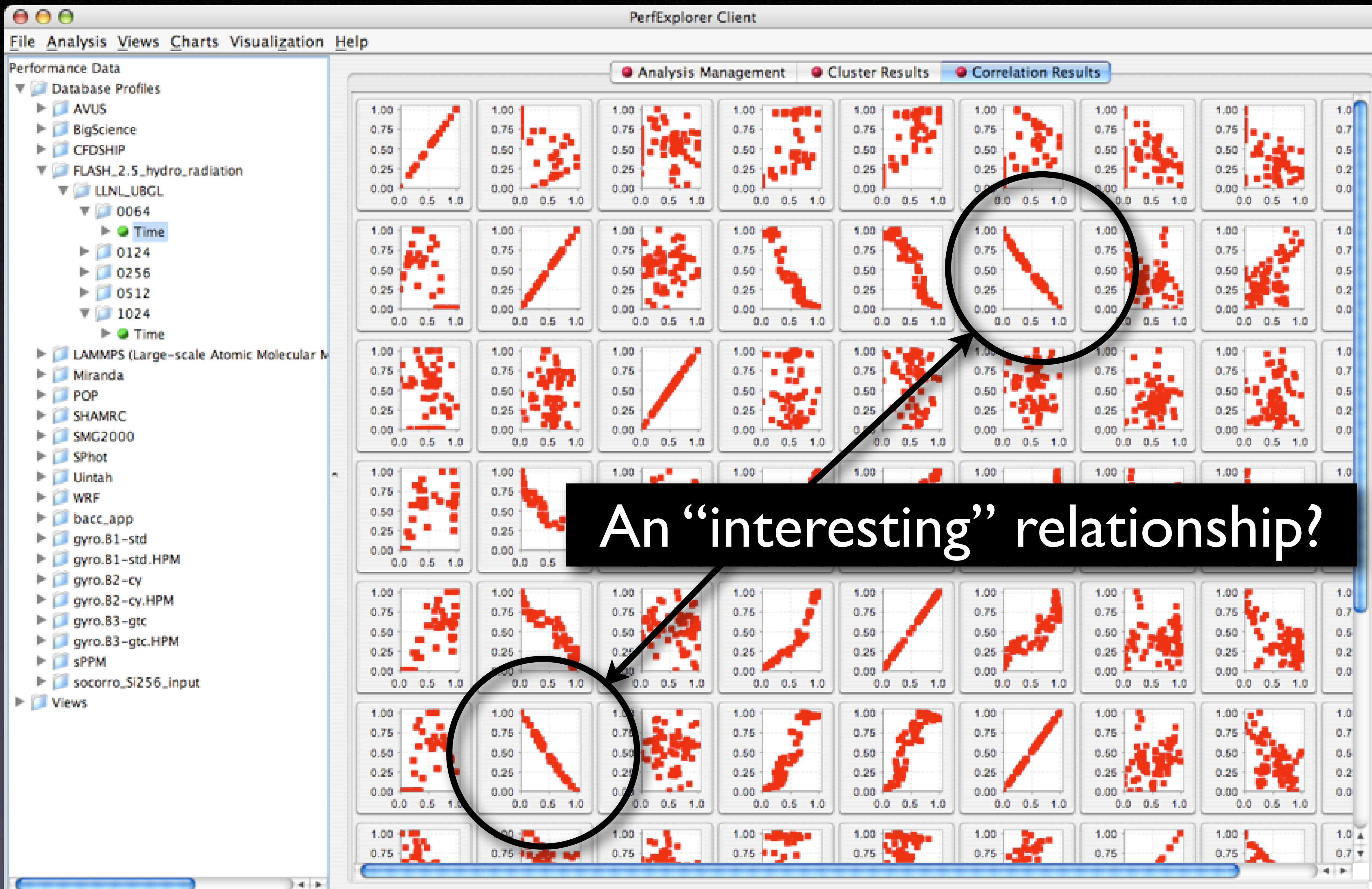
Observation of multi-modal distributions (clusters)

Data: FLASH on BG/L (LLNL)

Cluster Analysis



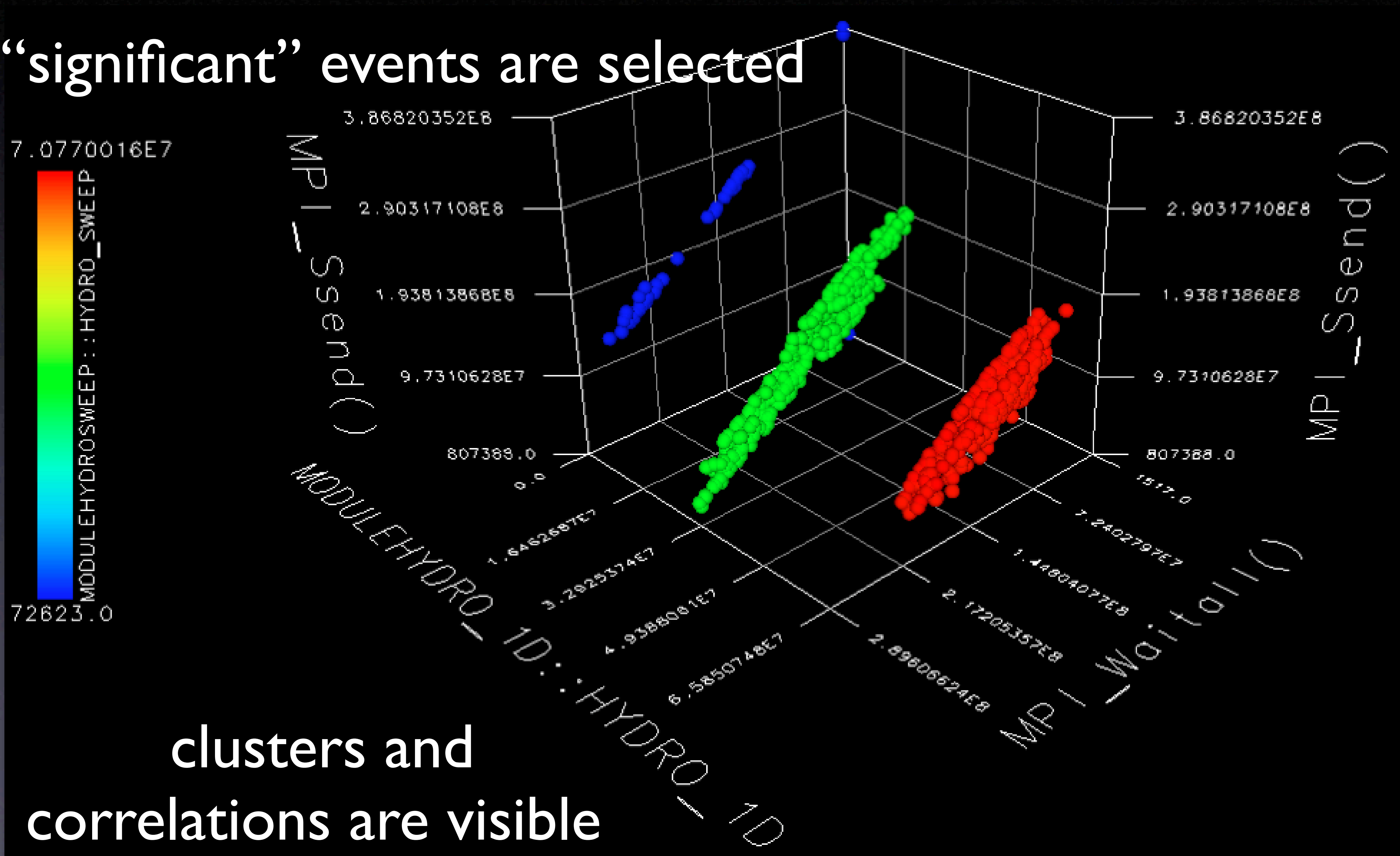
Correlation Analysis



Data: FLASH on BGL(LLNL)

4-D Visualization

4 “significant” events are selected



clusters and correlations are visible

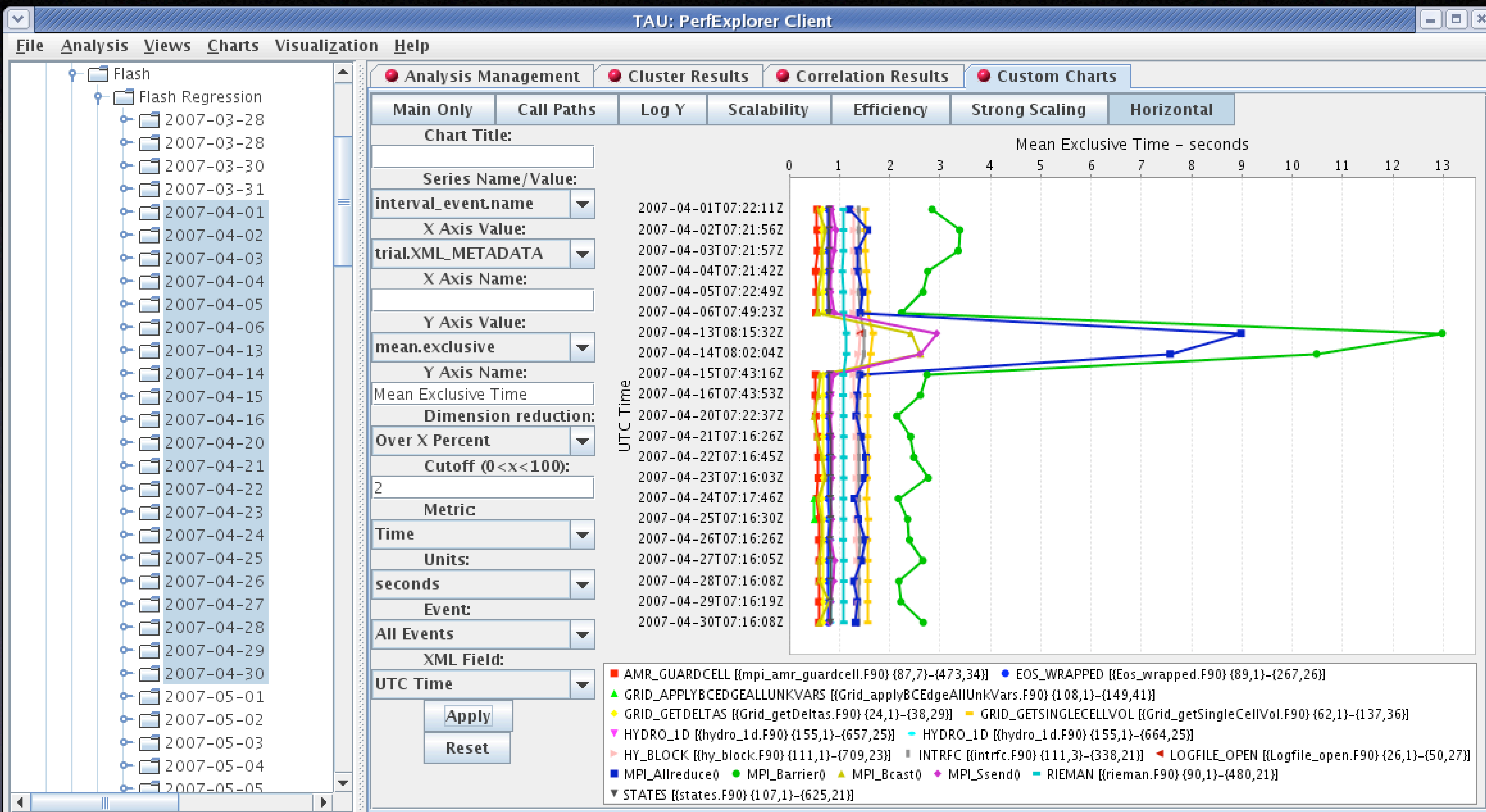
Data: FLASH on BG/L (LLNL)

New PerfExplorer Features:

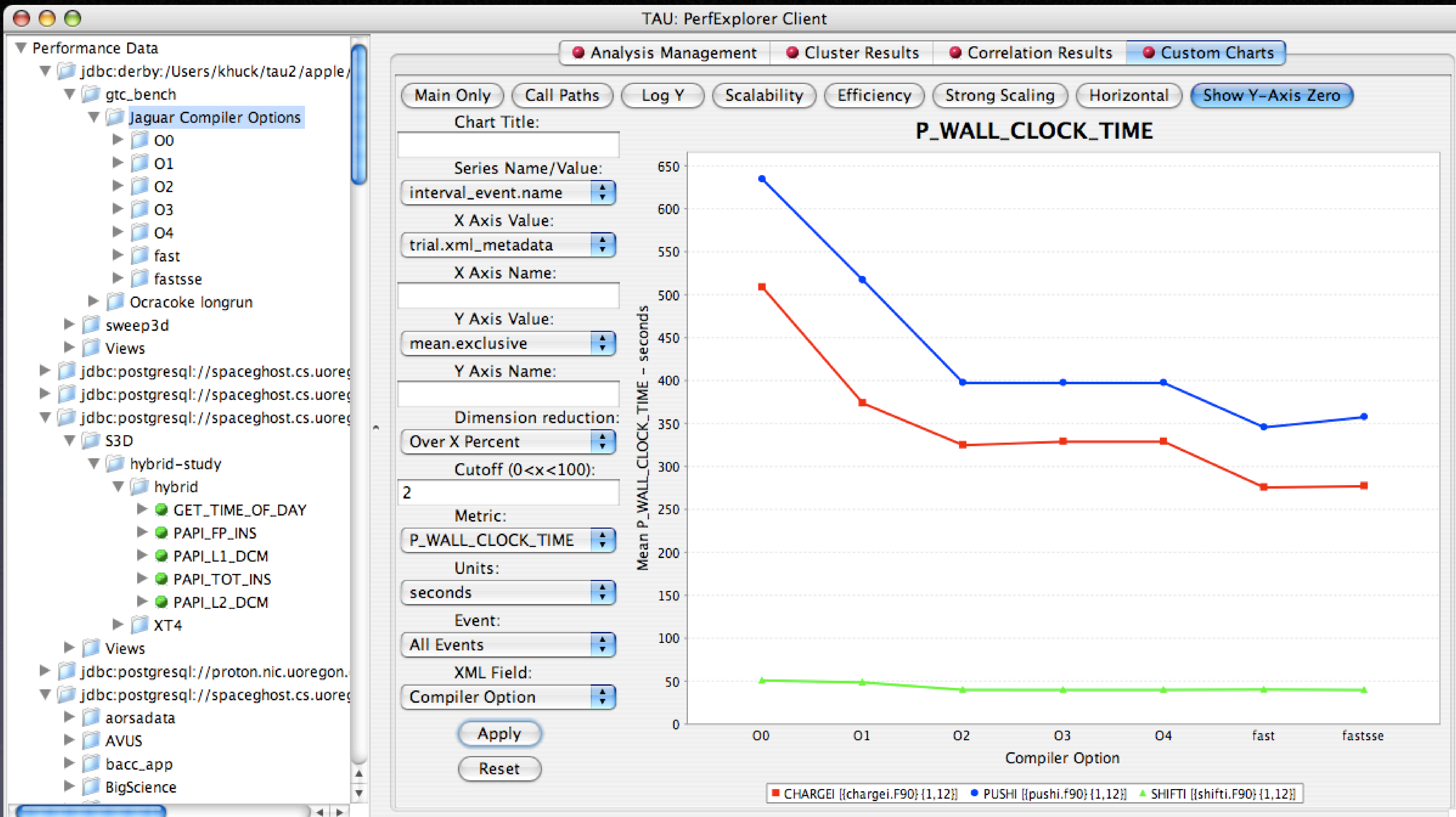
- Additional sophisticated metadata support
- Scripting using Jython*
- Inference engine to “infer” causes of performance phenomena from expert rules
- Redesigned component-based analysis
- Persistence of intermediate results
- Provenance

* or any other scripting language

Example: Regression Testing



Example: Parametric Study



PerfExplorer Example: S3D

- Compressible Navier-Stokes solver coupled with an integrator for detailed chemistry (CHEMKIN-compatible)
- Models turbulent reacting flow in combustion science
- MPI-based parallel computing implementation
- http://www.scidac.gov/BES/BES_TSTC/reports/TSTC2003Annual.html

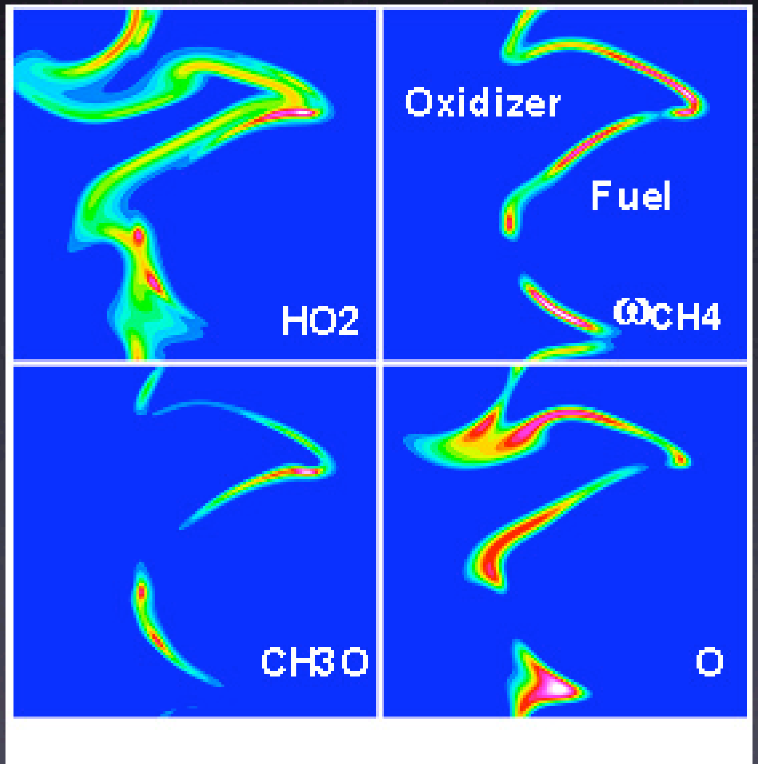
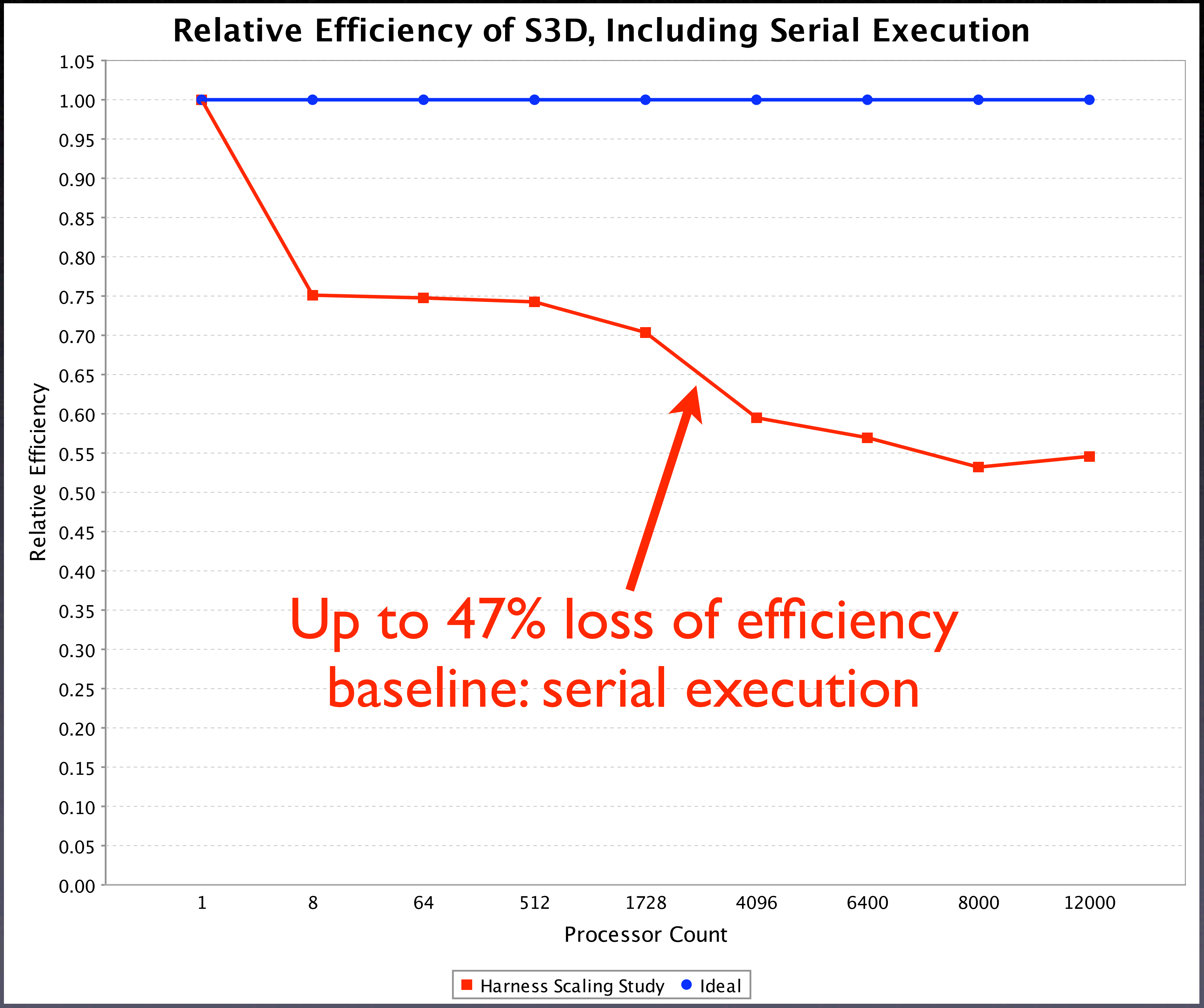
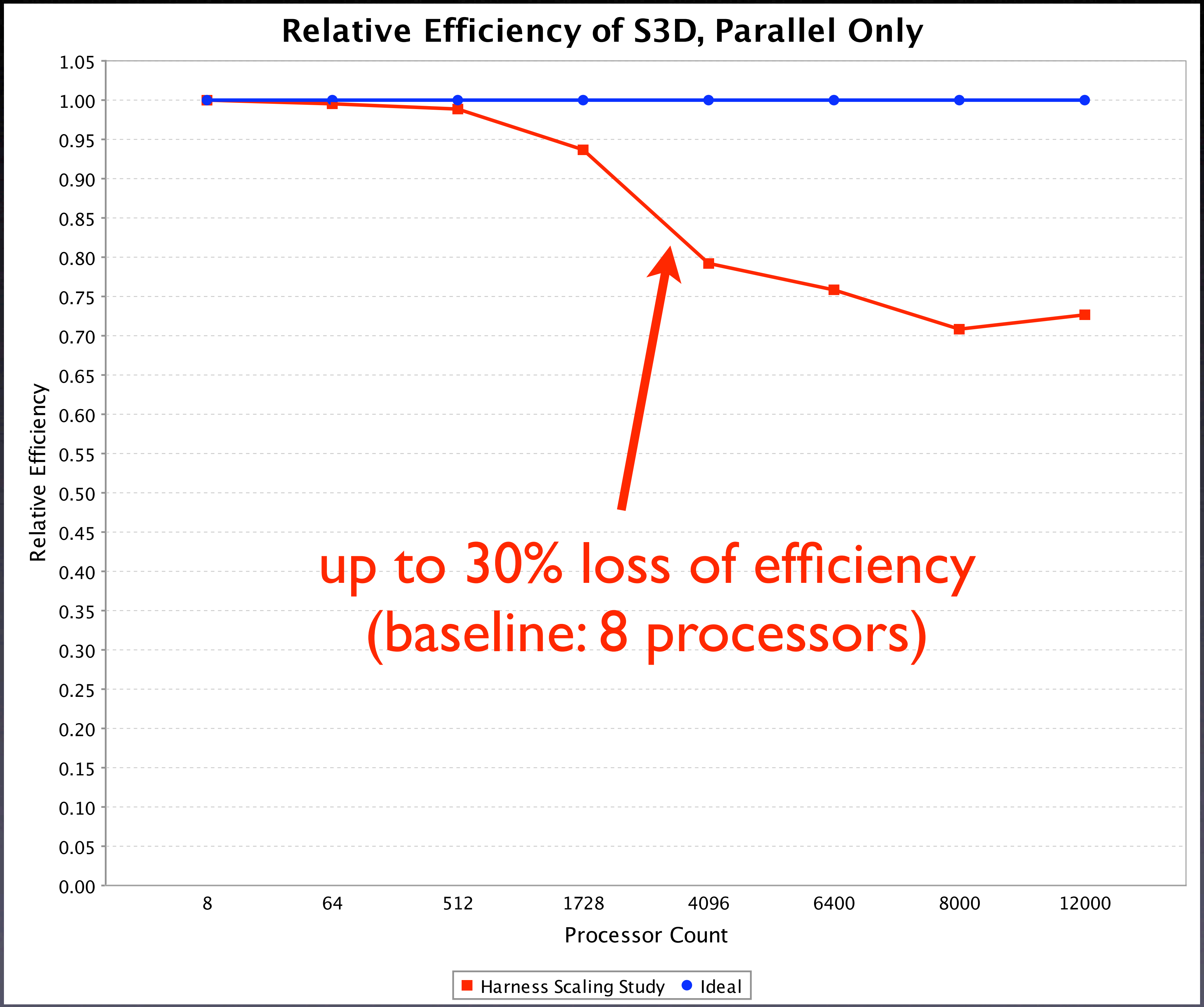


Figure: Two-dimensional DNS of a turbulent methane-air flame using detailed chemistry, GRI-Mech 3.0. Source: http://www.scidac.gov/BES/BES_TSTC/reports/TSTC2003Annual.html

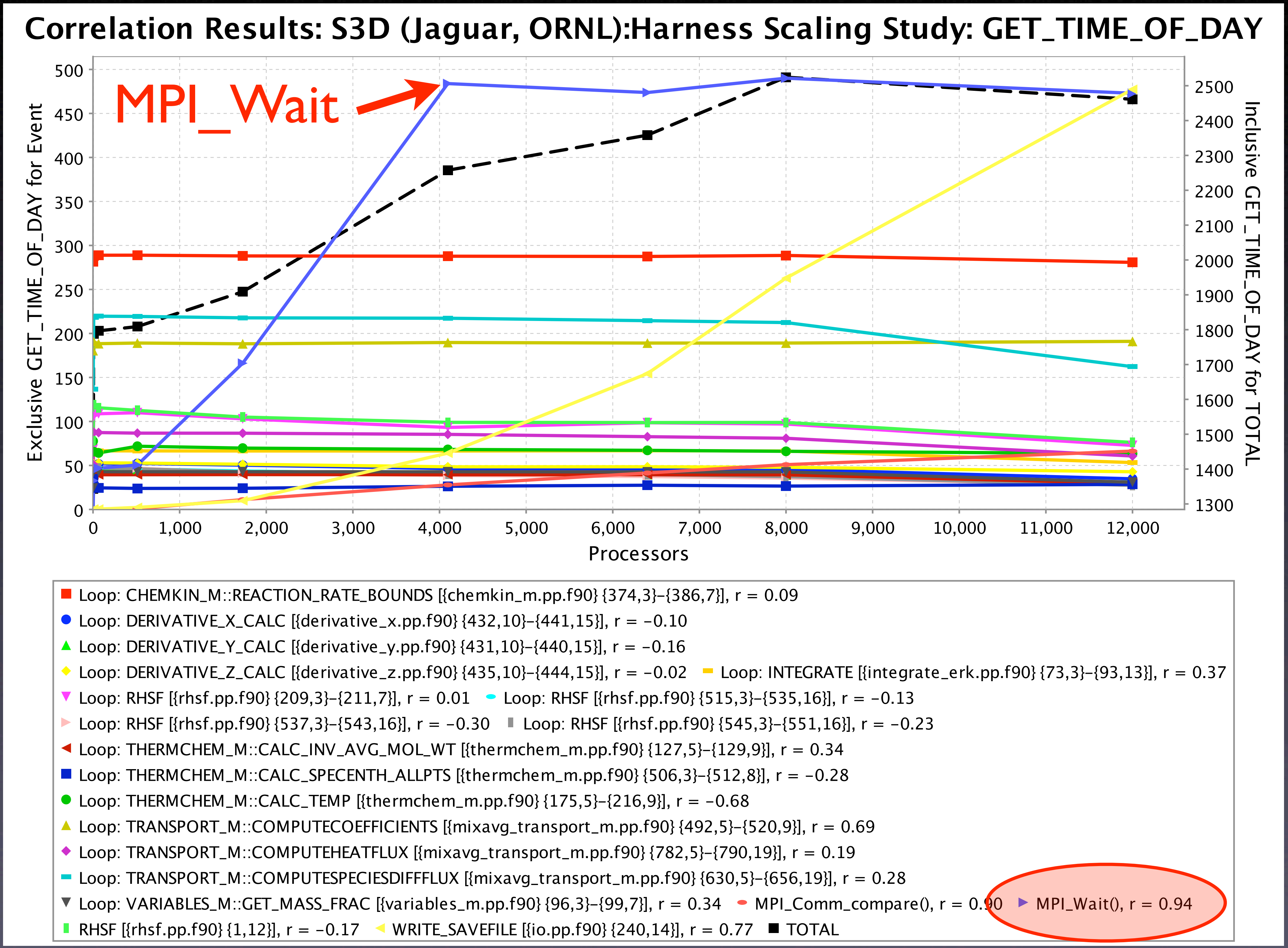
S3D Scaling: Efficiency



S3D Scaling: Parallel Efficiency

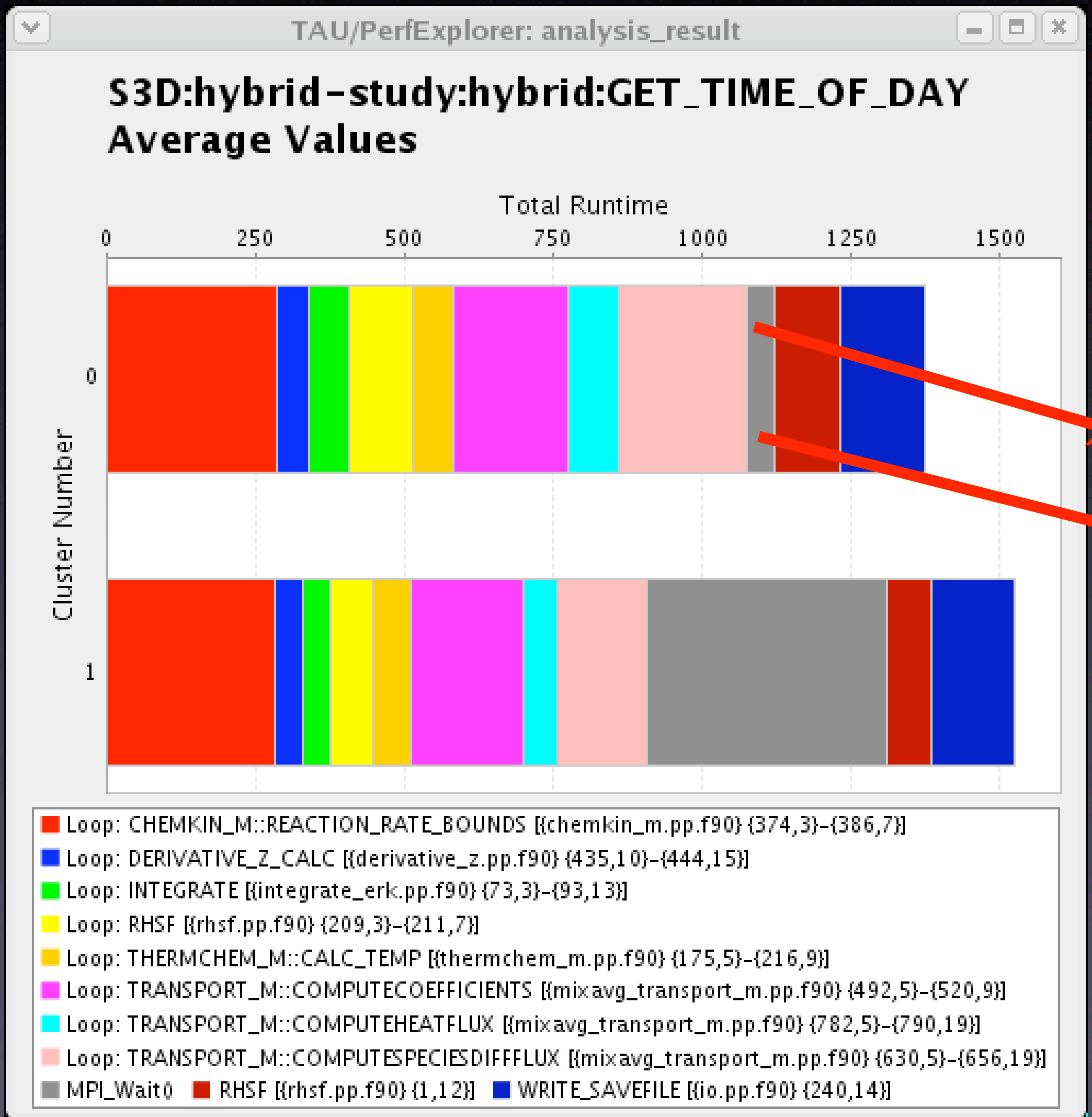


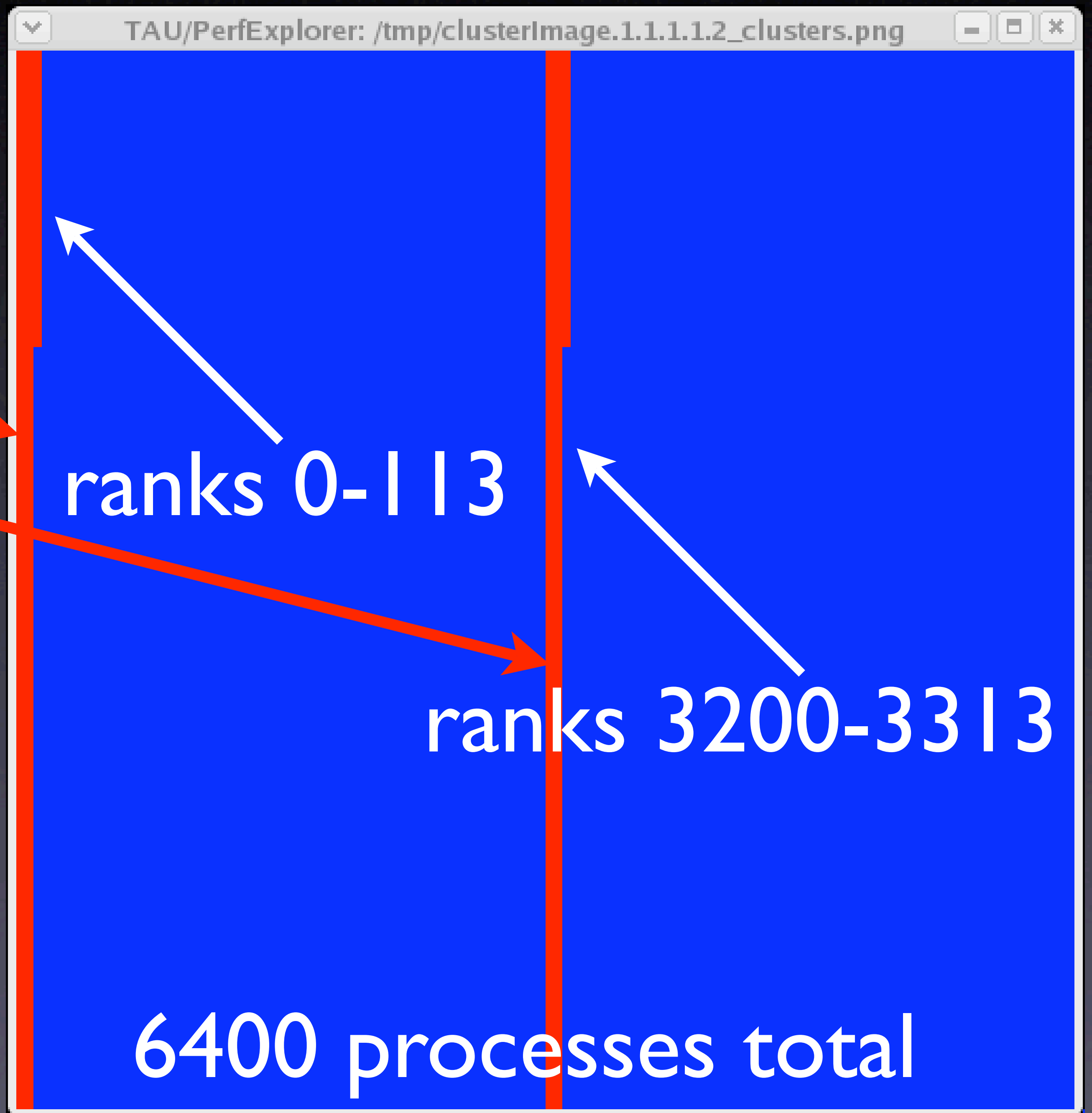
S3D Scaling: Correlation



S3D on Hybrid XT3/XT4

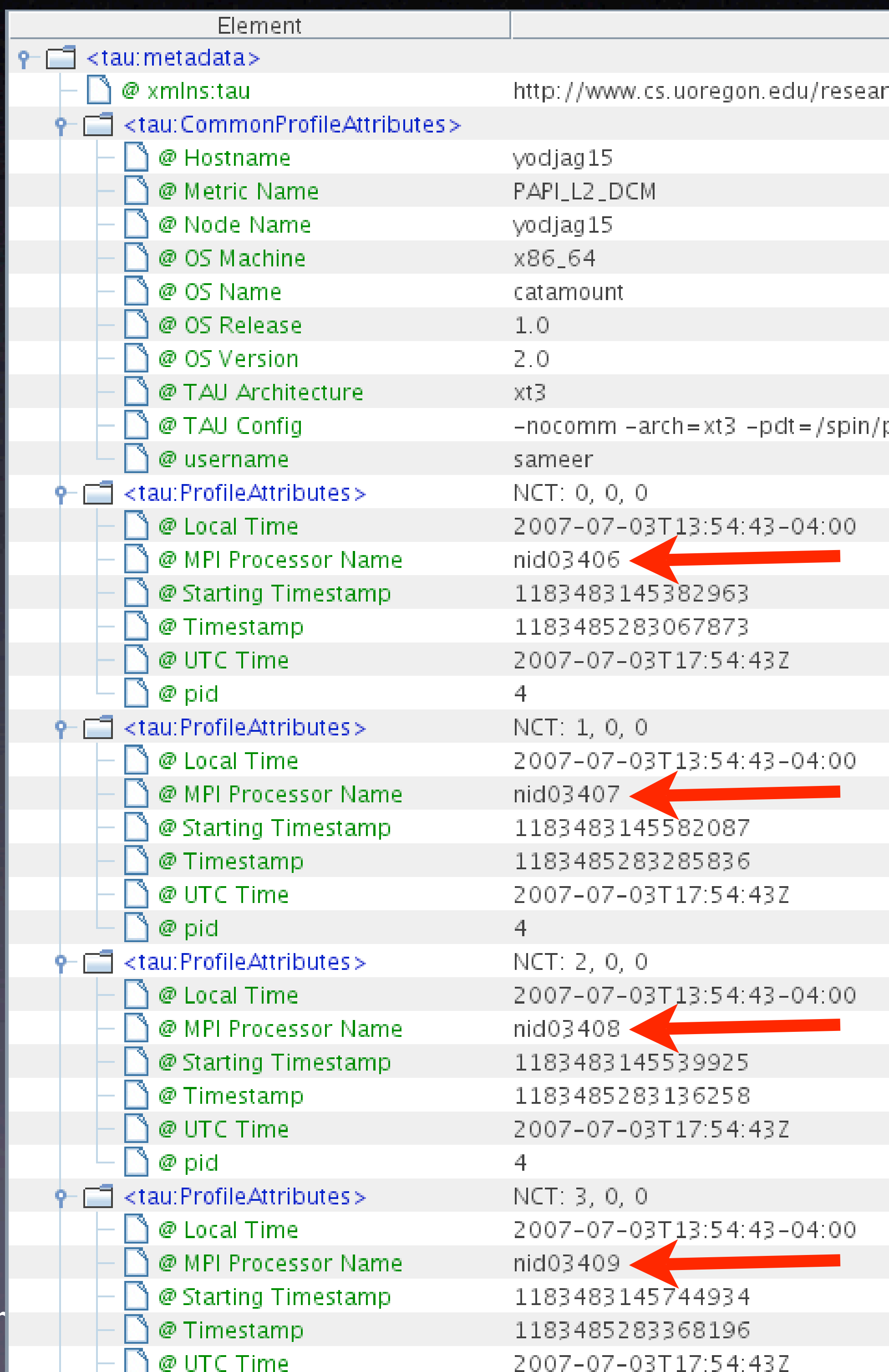
Most nodes are waiting on a few "slower" nodes...





Cluster explanation...

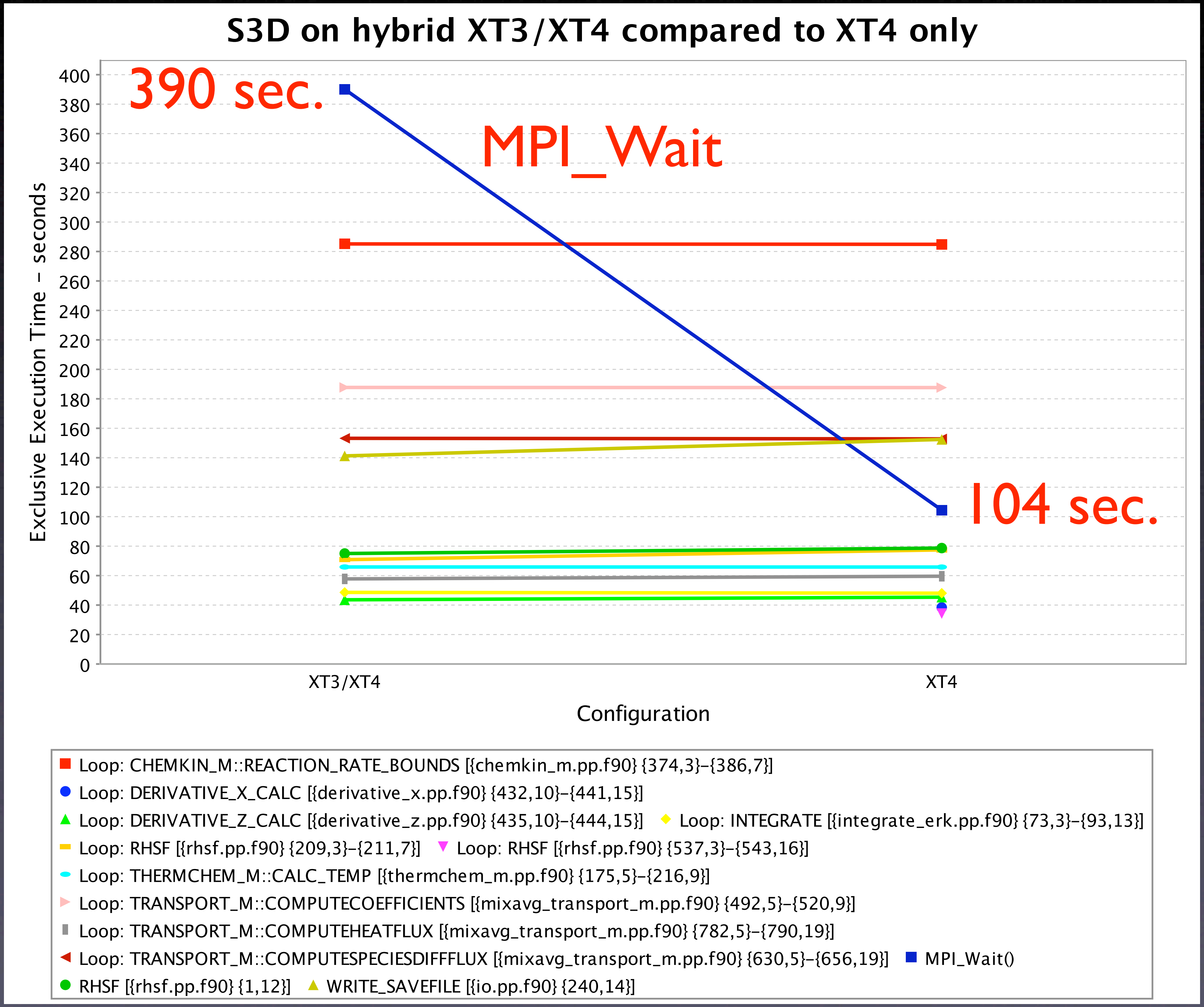
Metadata identifies node names



Element	
<tau:metadata>	
@ xmlns:tau	http://www.cs.uoregon.edu/resear
<tau:CommonProfileAttributes>	
@ Hostname	yodjag15
@ Metric Name	PAPI_L2_DCM
@ Node Name	yodjag15
@ OS Machine	x86_64
@ OS Name	catamount
@ OS Release	1.0
@ OS Version	2.0
@ TAU Architecture	xt3
@ TAU Config	-nocomm -arch=xt3 -pdt=/spin/p
@ username	sameer
<tau:ProfileAttributes>	NCT: 0, 0, 0
@ Local Time	2007-07-03T13:54:43-04:00
@ MPI Processor Name	nid03406
@ Starting Timestamp	1183483145382963
@ Timestamp	1183485283067873
@ UTC Time	2007-07-03T17:54:43Z
@ pid	4
<tau:ProfileAttributes>	NCT: 1, 0, 0
@ Local Time	2007-07-03T13:54:43-04:00
@ MPI Processor Name	nid03407
@ Starting Timestamp	1183483145582087
@ Timestamp	1183485283285836
@ UTC Time	2007-07-03T17:54:43Z
@ pid	4
<tau:ProfileAttributes>	NCT: 2, 0, 0
@ Local Time	2007-07-03T13:54:43-04:00
@ MPI Processor Name	nid03408
@ Starting Timestamp	1183483145539925
@ Timestamp	1183485283136258
@ UTC Time	2007-07-03T17:54:43Z
@ pid	4
<tau:ProfileAttributes>	NCT: 3, 0, 0
@ Local Time	2007-07-03T13:54:43-04:00
@ MPI Processor Name	nid03409
@ Starting Timestamp	1183483145744934
@ Timestamp	1183485283368196
@ UTC Time	2007-07-03T17:54:43Z

- Ranks 0-113 lie on processors 3406-3551
- Ranks 3200-3313 are also on 3406-3551
- 3406-3551 are on the XT3 partition
- XT3 has slower DDR-400 memory (5986 MB/s)
- XT3 has a slower SS1 (1109 MB/s) interconnect
- XT4 partition has faster DDR2-667 memory modules (7147 MB/s) and faster Seastar2 (SS2) (2022 MB/s) interconnect
- Running on XT4 yields 12% improvement
- If XT3/XT4, load balancing will be required

S3D: "Improved" Result



PerfExplorer example: GTC

- GTC: Gyrokinetic Toroidal Code
- Particle-in-cell physics simulation
- CHARGEI: particles apply their charge to the grid cells
- PUSHI: cells update particle locations
- MPI-based parallel implementation

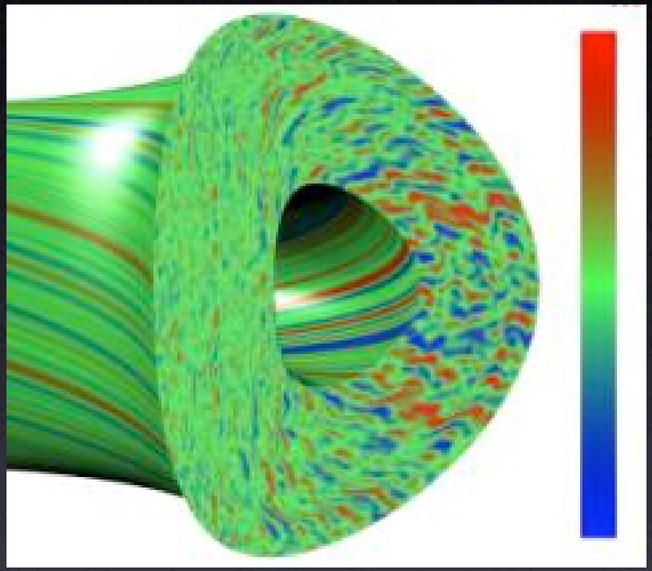
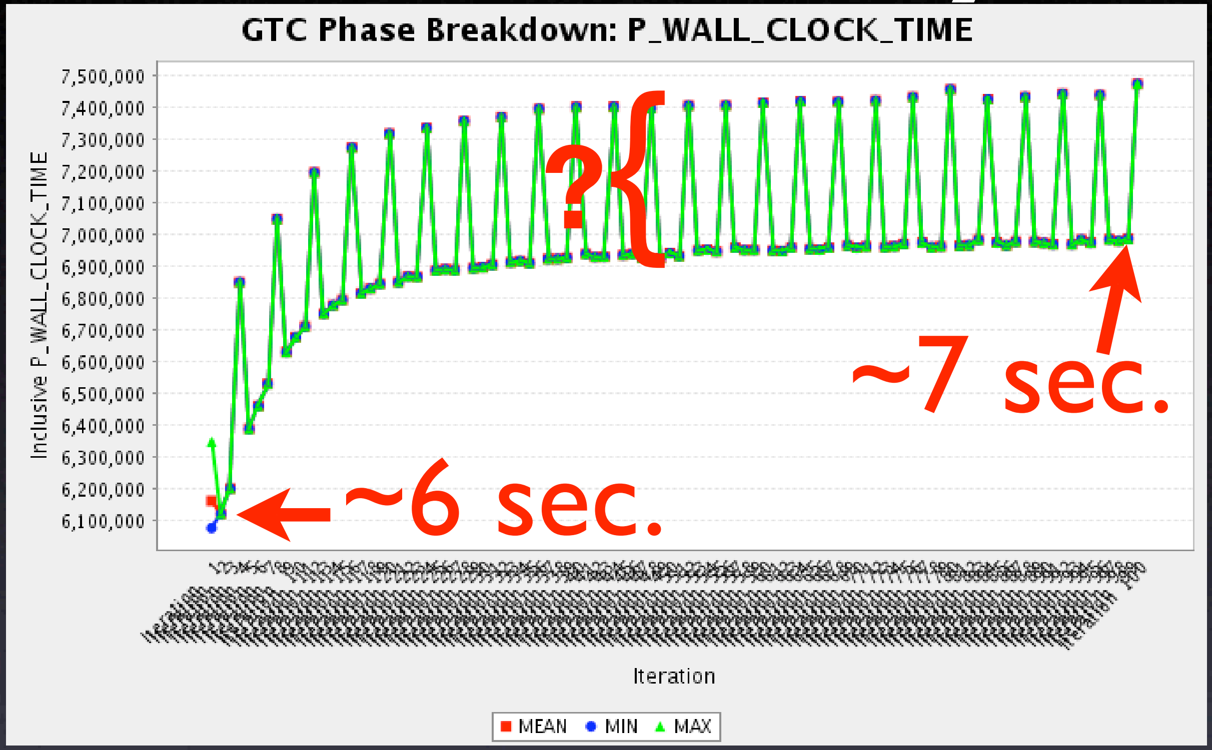
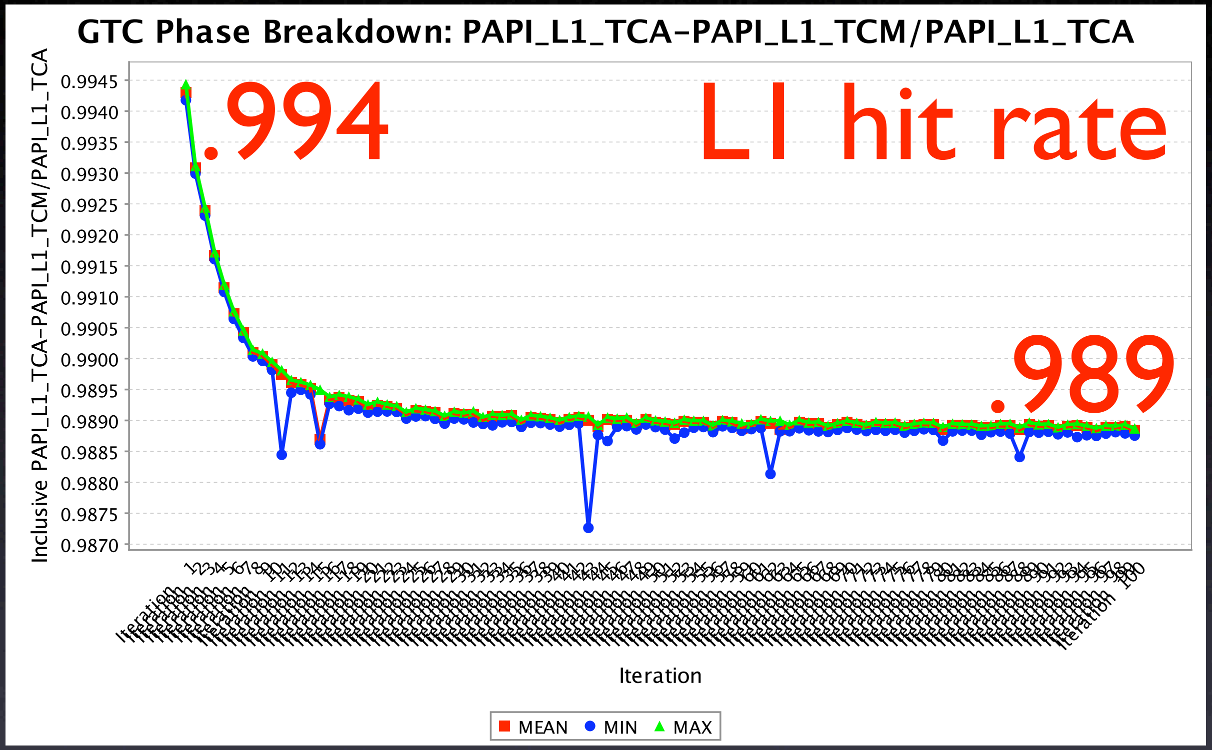
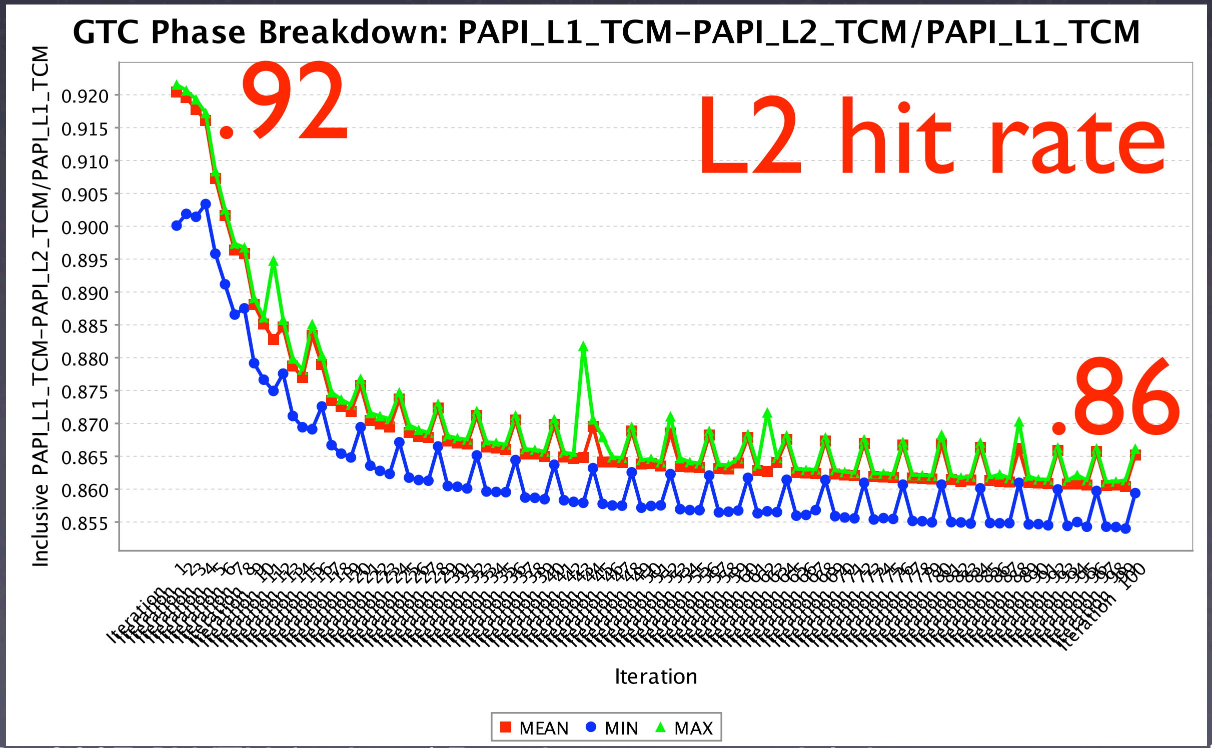


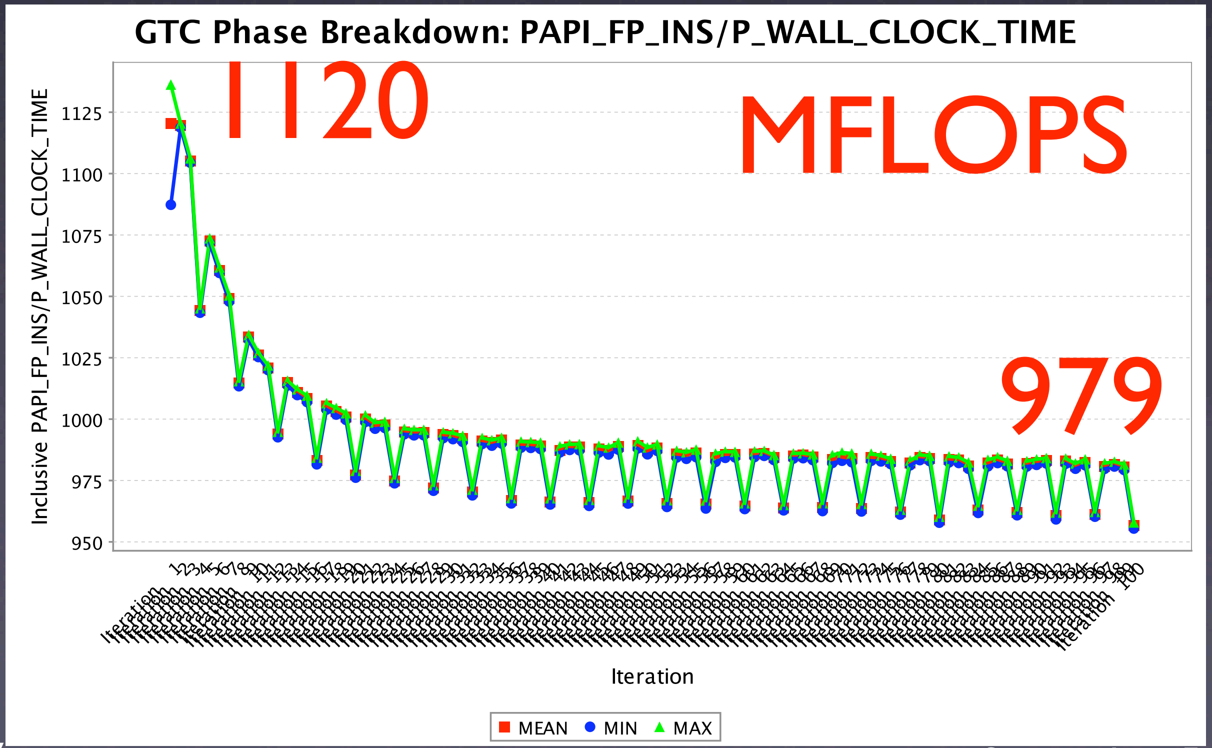
Figure: The (turbulent) electrostatic potential from a GYRO simulation of plasma microturbulence in the DIII-D tokamak.
Source: http://www.scidac.gov/FES/FES_PMP/reports/PMP2004Annual.html

GTC: Dynamic Phases



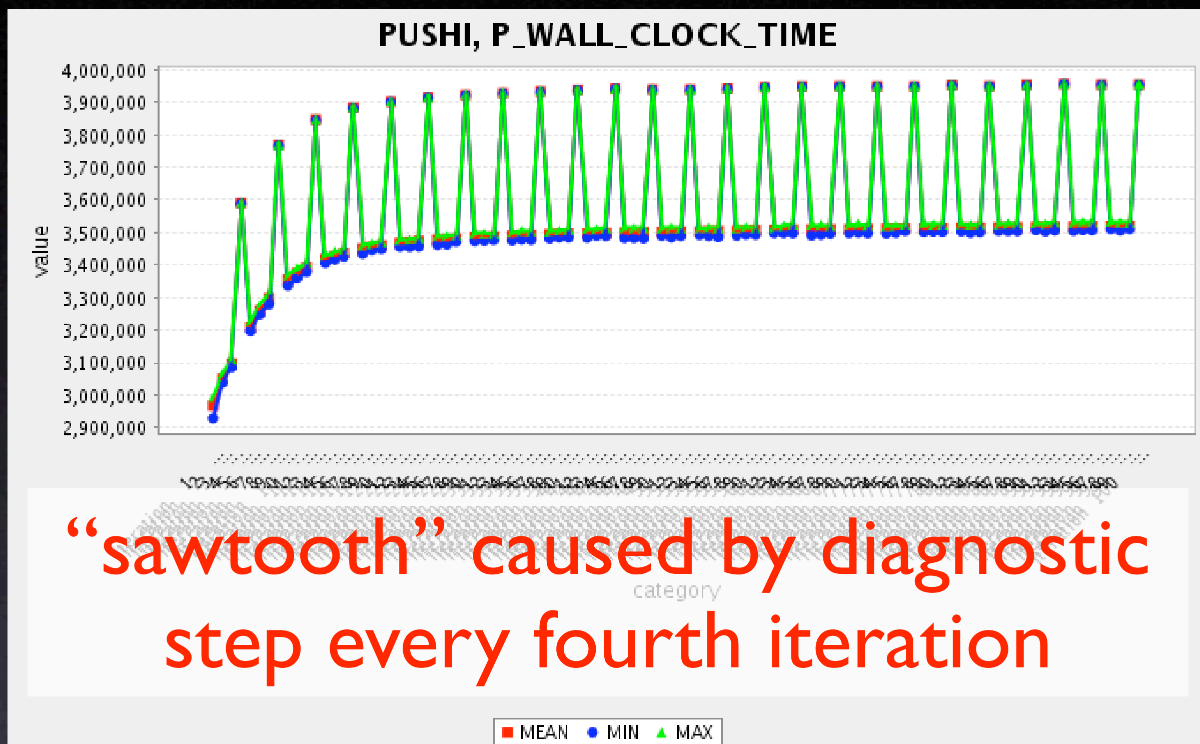


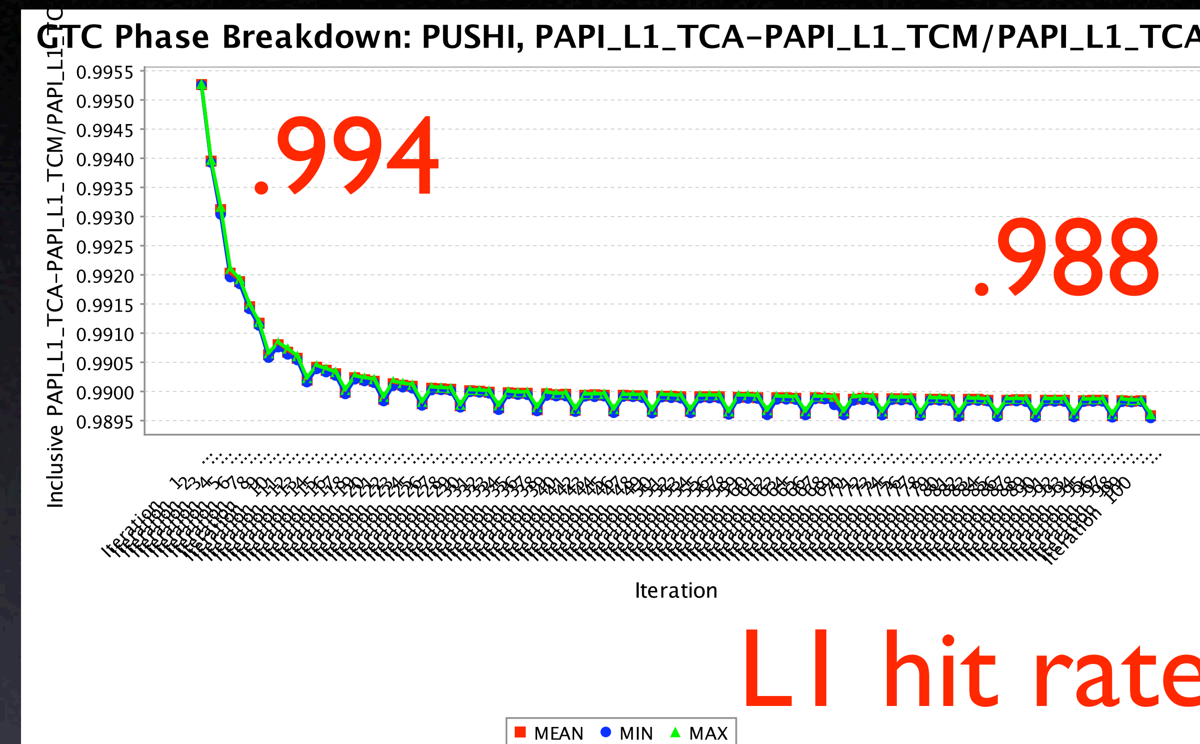


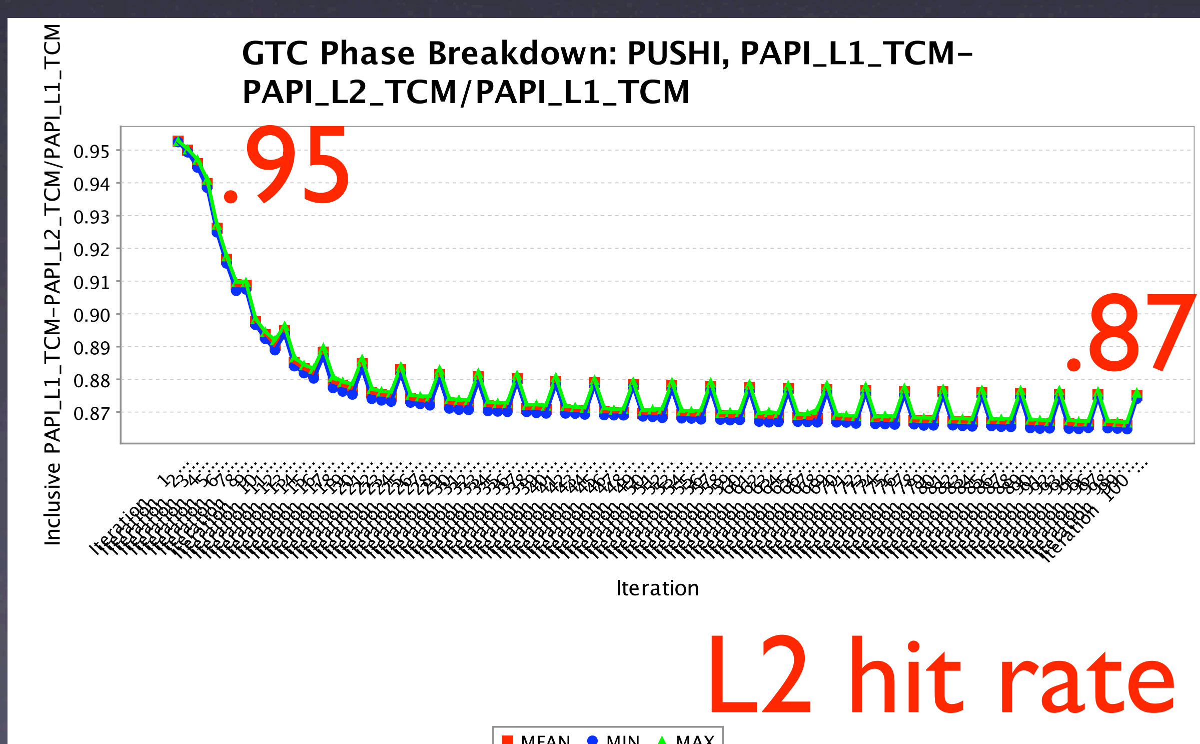


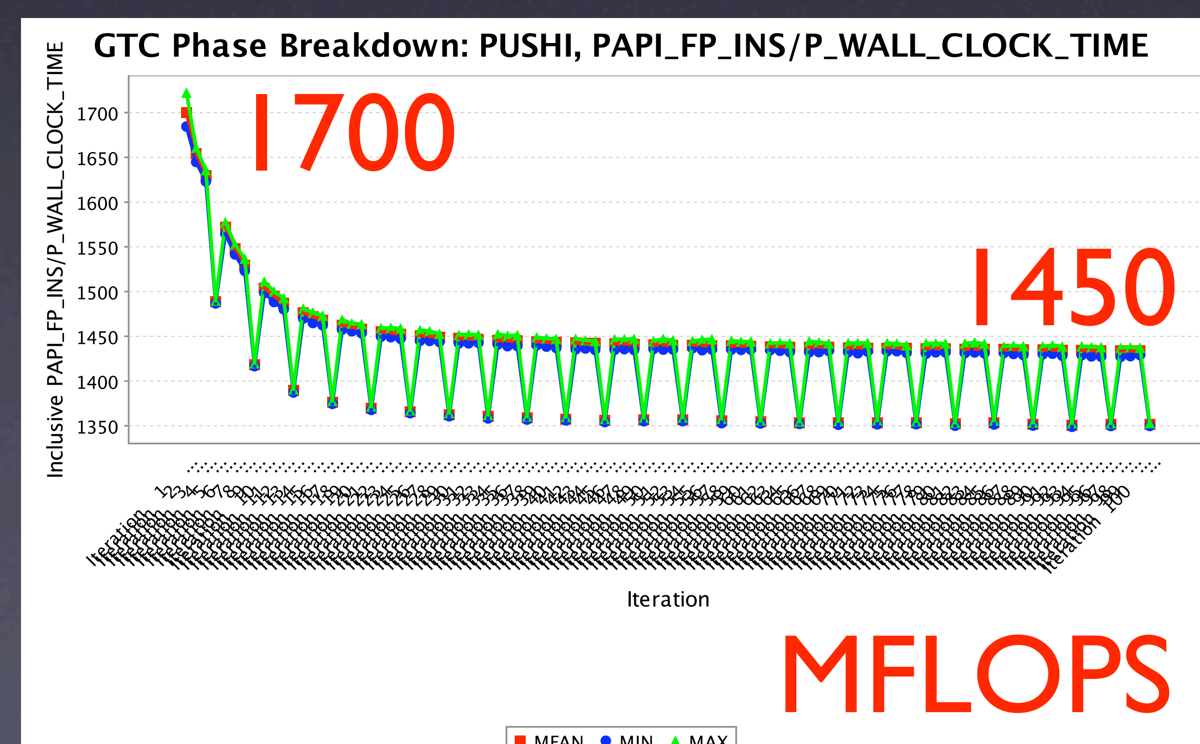
Source: GTC on Cray XT4

GTC: Iteration *i* => PUSHI



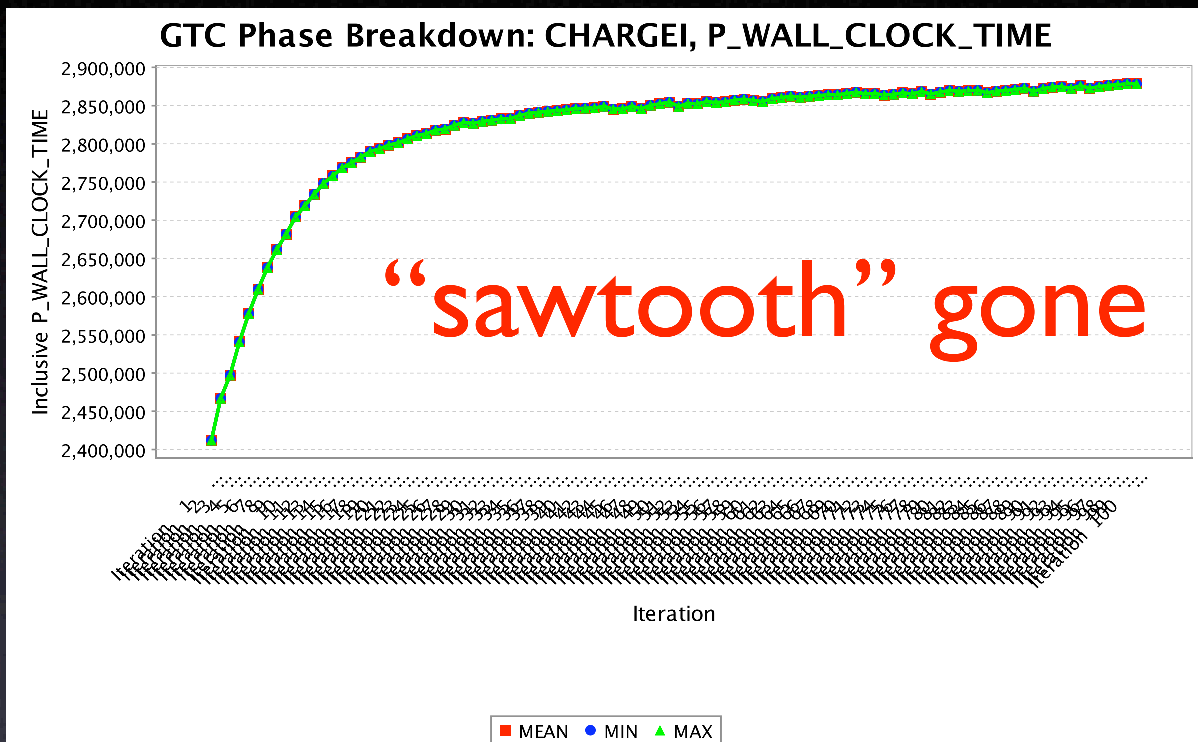


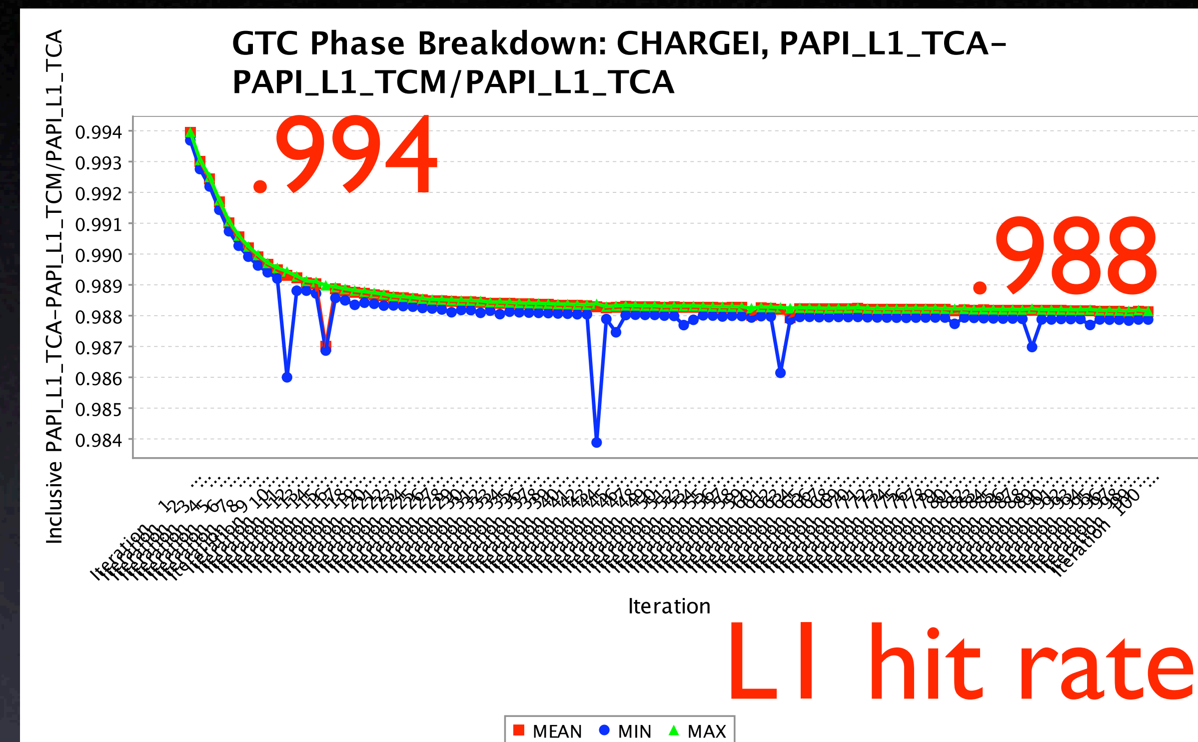


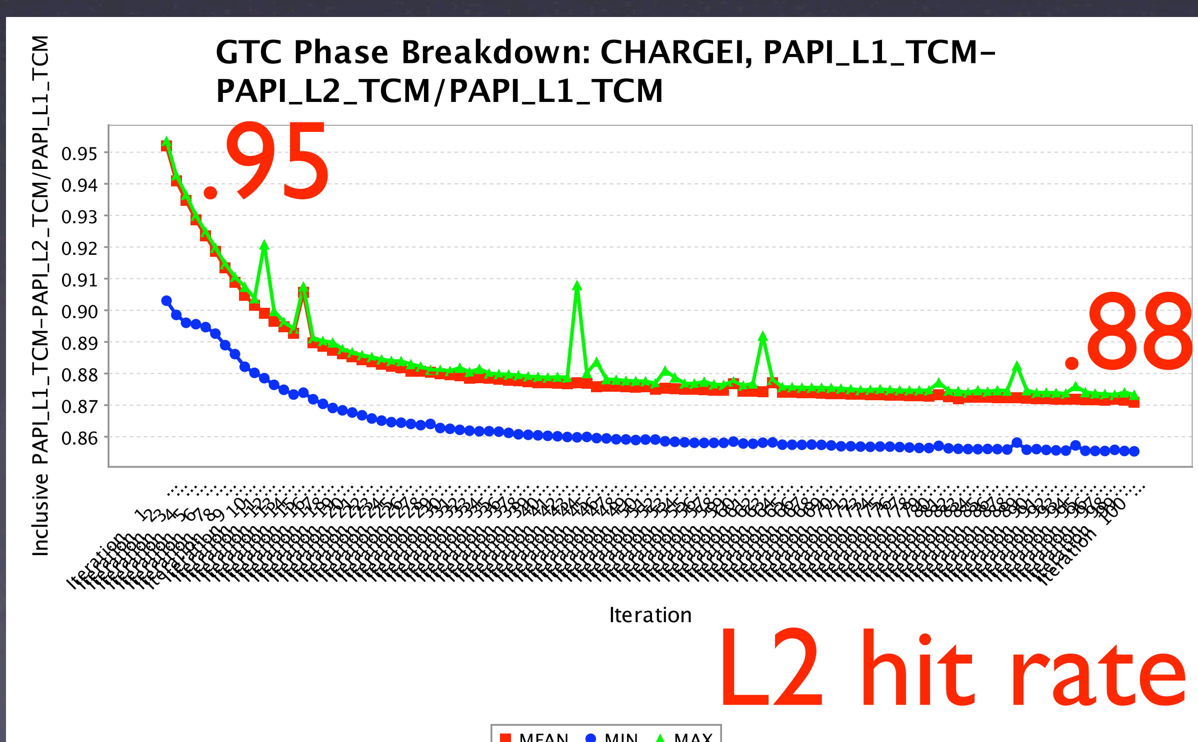


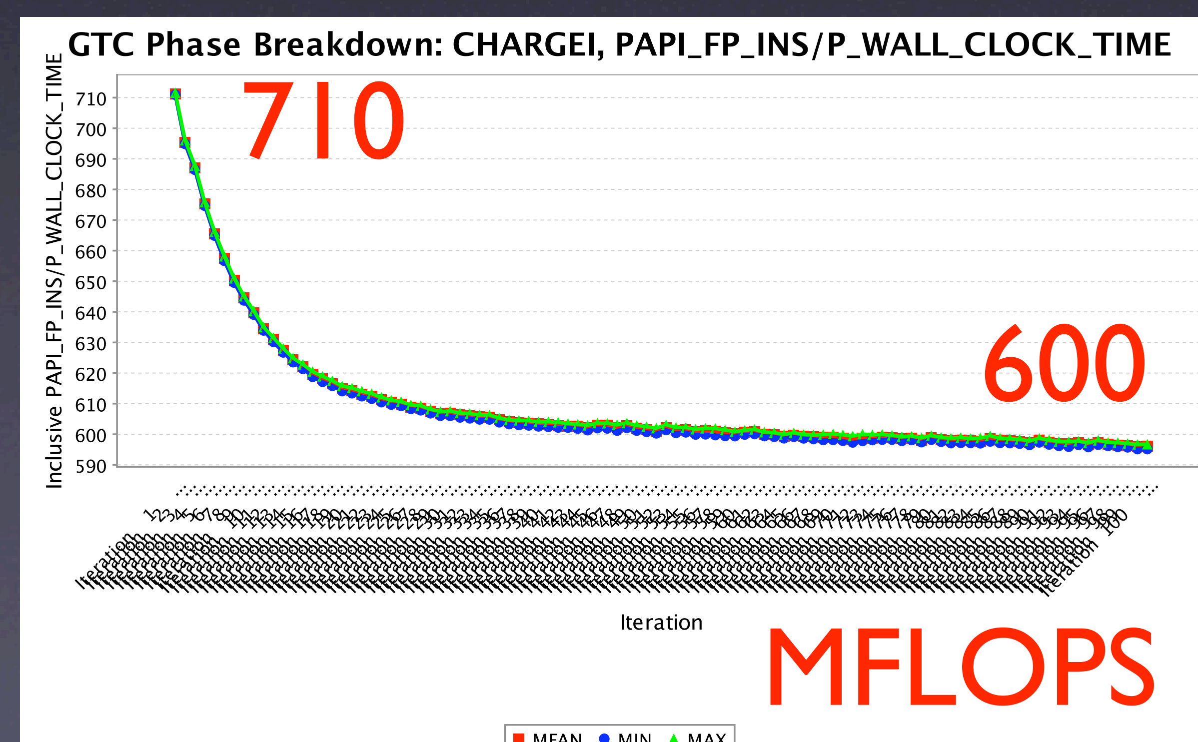
Source: GTC on Cray XT4

GTC: Iteration *i* => CHARGEI



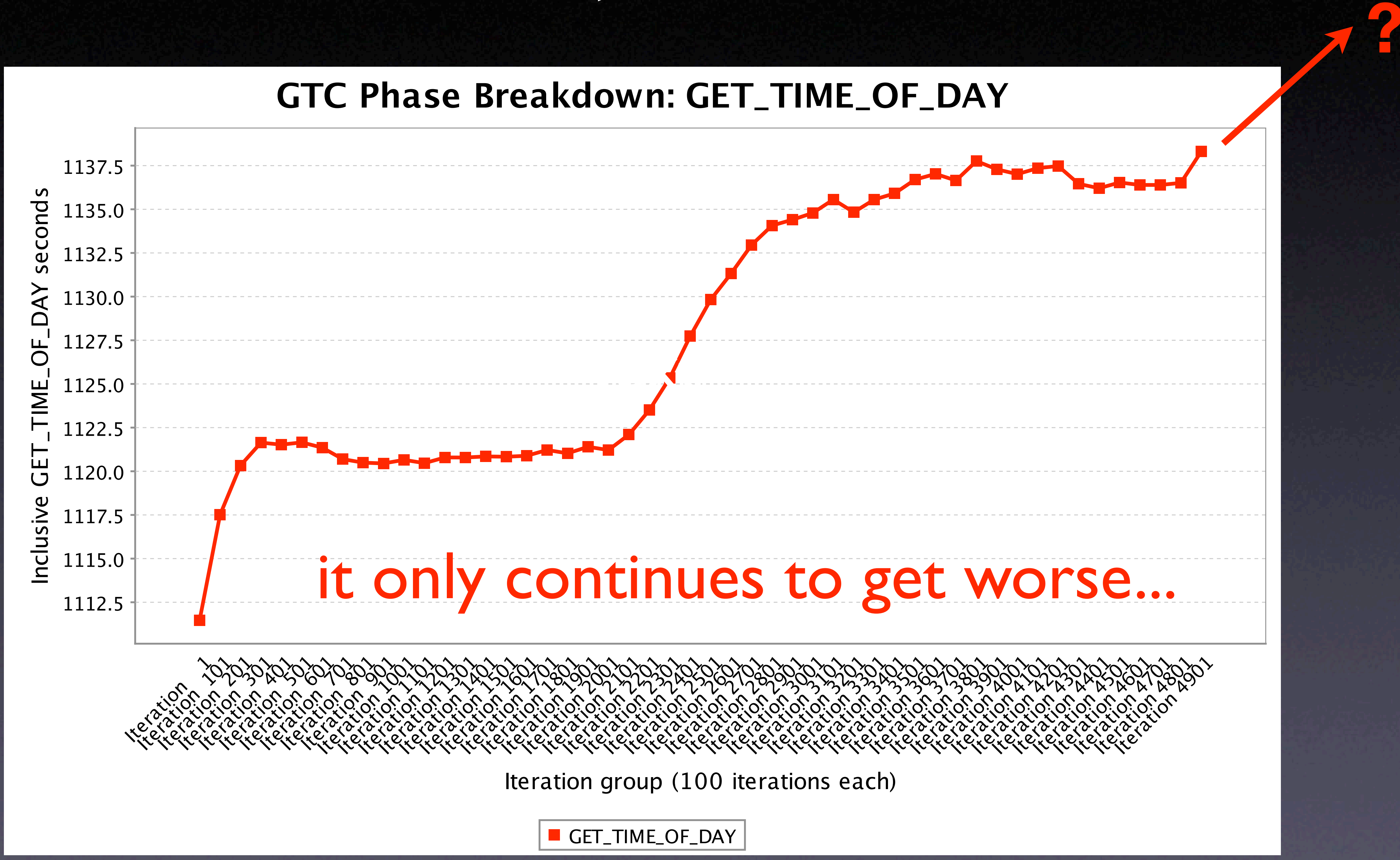






Source: GTC on Cray XT4

GTC: after 5,000 Iterations



Source: GTC on BG/L

GTC Phases: Summary

- CHARGEI, PUSHI events have good spatial locality of particles, but bad temporal locality for cells
- After 100 iterations, each iteration on Cray XT3/XT4 takes ~1 second longer (~12%)
- Full simulation runs for 10,000 timesteps (potential improvement of 2+ hours from 20 hour execution)
- PUSH event calls a diagnostic routine every X timesteps (input variable)
- Analysis is ongoing...

Conclusion

- **TAU:**
  - Portable, configurable, complete instrumentation and measurement of parallel profiles and traces
- **PerfDMF:**
  - Profile management, query and analysis API
  - Supports most commonly used profile formats (if not, we can add it)
- **PerfExplorer:**
  - Robust profile parametric study support
  - In-depth analysis of large scale profiles

Acknowledgments

- US Department of Energy (DOE)
  - Office of Science
    - MICS, Argonne National Lab
  - ASC/NNSA
    - University of Utah ASC/NNSA Level 1
    - ASC/NNSA, Lawrence Livermore National Lab
- US Department of Defense (DoD)
- NSF Software and Tools for High-End Computing
- Forschungszentrum Jülich
- TU Dresden
- Los Alamos National Laboratory
- ParaTools, Inc.
- SciDAC, PERI, ORNL, NERSC, RENCI