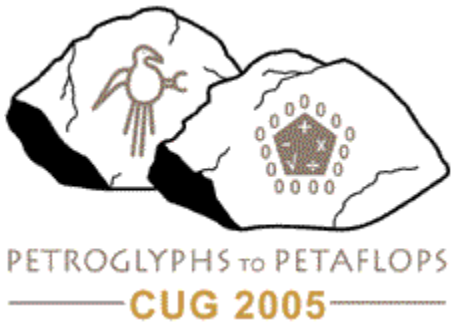
Early Applications Experience on the Cray XT3

Nick Nystrom, Shawn Brown, Jeff Gardner, Roberto Gomez, Junwoo Lim, David O'Neal, Richard Raymond, R. Reddy, John Urbanic, and Yang Wang

May 17, 2005



Outline

- Architecture and environment
- User engagement
- Preliminary performance
- Scientific progress

PITTSBURGH SUPERCOMPUTING CENTER

PITTSBURGH SUPERCOMPUTING CENTER

Application-Enabling XT3 Architecture

- 3D torus direct-connected processor (DCP) architecture
- high bandwidth, low latency interconnect with embedded communications processing and routing
- OS system designed to support scalable applications
- integrated RAS system provides high reliability, allowing full applications to run to completion
- high-speed, global I/O
- standards-based programming environment
  - Fortran, C, C++, MPI
  - Convenient development on smaller systems

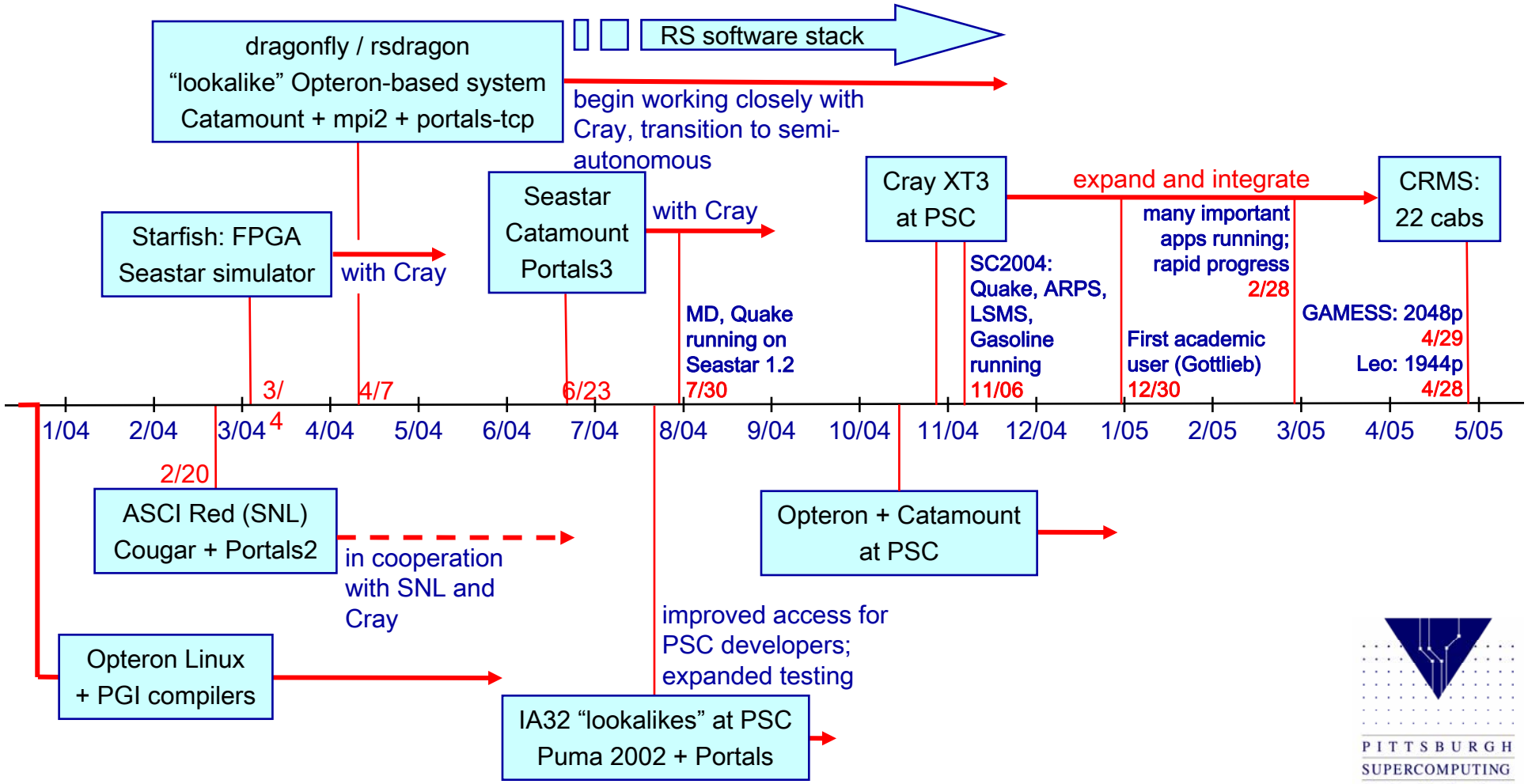
PITTSBURGH SUPERCOMPUTING CENTER

Toward a New Capability Resource

- Engage leaders of all aspects of high-end computing
  - Computational science
  - Performance modeling and evaluation
  - Computer science
  - Participation by diverse NSF directorates, universities, and labs
- Pursue scalable applications for all problem sizes, ranging from unprecedented through moderate
  - Superior, cost-effective computational performance coupled with a low-latency, high-bandwidth interconnect, a parallel filesystem, and TeraGrid connectivity enables capability computing
  - Excellent interprocessor bandwidth is critical when the amount of communication relative to computation is high
- Work closely with Cray Inc.
  - extending communications, I/O, scheduling, applications

PITTSBURGH SUPERCOMPUTING CENTER

PSC's Cray XT3 Timeline



PITTSBURGH SUPERCOMPUTING CENTER

PSC System Enhancements to Support Computational Scientists (1)

- Push the dev harness from 2 to 4 to 8 to 10 to 22 cabinets → scale applications to full configuration
- Kerberos logins → streamline system access, interface with TeraGrid
- Lustre → improve I/O reliability, transition to parallel filesystem
- Automated booting scripts to interact with scheduler to reboot as needed under the dev harness → improve availability
- xtconsole logging collection script and tool to search logs for specific events related to time span and/or integer nids under CRMS → provide complete diagnostics to software developers
- 10 GbE working in SIO node;
- InfiniBand working in SIO node
- SCSI working SIO node, allowing direct attachment of storage shelves
- Demonstration of GridFTP on the NSF TeraGrid
- System monitors to visually scan and interrogate system activity

PITTSBURGH SUPERCOMPUTING CENTER

PSC System Enhancements to Support Computational Scientists (2)

- Adapt Torque to the XT3 dev harness
  - Added capability to schedule to multiple harness systems containing various cabinet count
  - → management of multiple jobs and users without need to specify nodelists
- Designed preliminary custom scheduler for XT3, cf. Simon on LeMieux
  - Drain whole system
  - Drain harness system (set of cabs)
  - Backfill (using aggressive backfilling)
  - Nidmasks (let users target specific nids)
  - Lustre switch (to let users target jobs to harness system with Lustre)
- Setup post-job node checking via ping_node on dev harness system
- Setup pre-job node checking via ping_list (fast/wide ping_node)

PITTSBURGH SUPERCOMPUTING CENTER

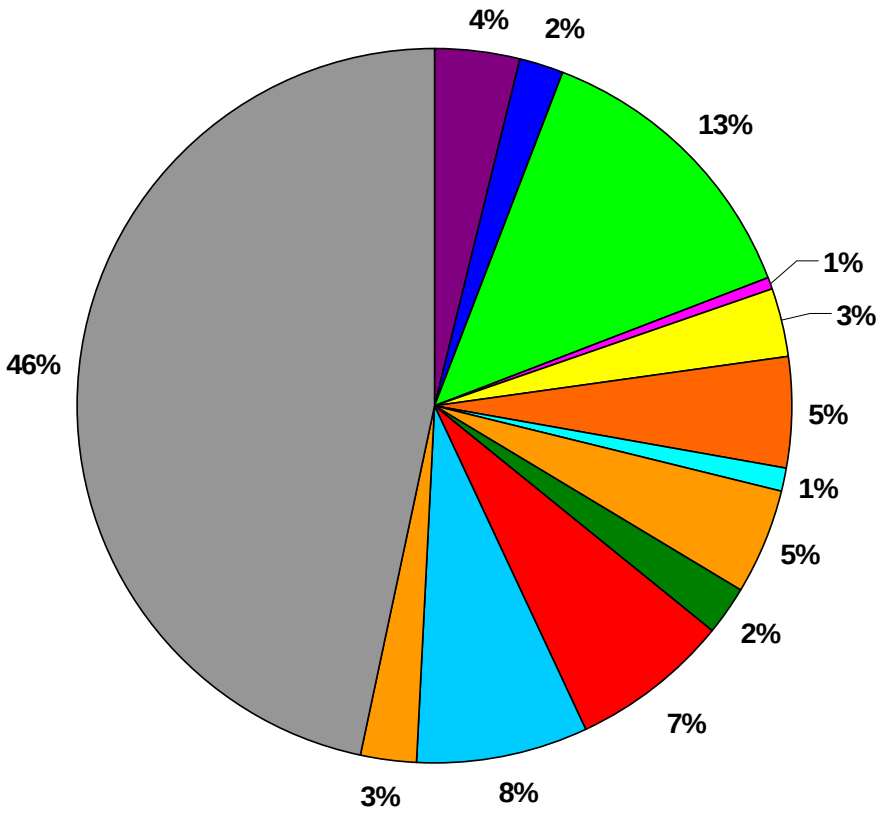
PSC System Enhancements to Support Computational Scientists (3)

- PD: native Portals code for pairwise and broadcast/gather communications
- PDIO: I/O library with file semantics, bypassing yodio
- "share" tester that yods more than one process to the same nid to test yod and Portals communications within a processor (Portals, excluding routing).
- fwstat driver script that can automatically add problem nodes to the PBS "bad NIDS list", based on the machine's response to the fwstat probe
- benchmarks/diagnostics that test portals bandwidth and routing
- utility codes, e.g. nid_parse
- RCA (Resiliency Communication Agent) test codes to generate, receive, and monitor RCA events

PITTSBURGH SUPERCOMPUTING CENTER

Early XT3 Users

Normalized usage on PSC Platforms Feb04-Jan05



- The users with which PSC is working constitute 54% of normalized usage on PSC platforms over the last year.
- Examples:

biophysics	25,563,887
QCD	14,831,603
astrophysics	13,402,569
turbulence	9,513,116
solar astrophysics	9,136,520
storm simulation	7,243,520
CFD	5,787,768
geodynamo	4,829,447
MD	4,266,049
Total	191,928,345

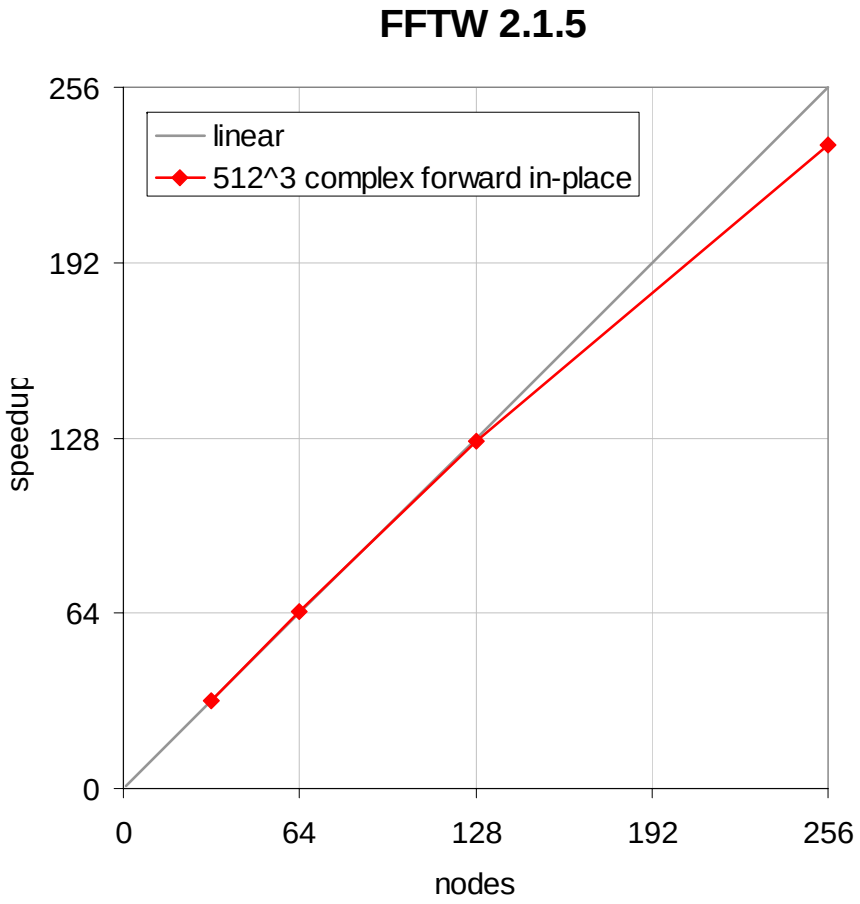
Representative Applications

applications	domain	nodes
GAMESS	Quantum chemistry	2048
STREAM	Memory bandwidth benchmark	2048
PD	Portals Direct	1960
Leo	Numerical relativity	1944
sPPM	Piecewise Parabolic Method bmk	1944
MILC	QCD	1024
Quake	Earthquake modeling	1024
Gasoline	N-body astrophysics	1024
ZEUS-MP	Astrophysics	1000
MPQC	Massively Parallel Quantum Chemistry	941
Dynamo	QM/MM biochemistry	900
HPCC	HPC Challenge benchmarks	900
cpu_burn	System stability	760

application	domain	nodes
LSMS 1.6	Materials science / electronic structure	756
NBP EP	NAS Parallel Benchmarks: EP	600
HPCC	HPC Challenge benchmark	552
TBLC	Turbulent Boundary Layer Code	512
FFTW	Adaptive FFT library	256
NAS: LU, CG, MG, FT		256
VASP	Plane wave DFT	128
AMBER	Molecular dynamics	128
ARPS	Storm modeling	128
CHARM++	Parallel language & runtime sys	128
PMB	Pallas MPI benchmarks	64
svr	Volume Rendering	64
NAMD	Molecular dynamics	16
CHARMM	Molecular dynamics	16

Applications and libraries Benchmarks and kernels Systems and stress tests

Seastar Interconnect Yields High Efficiency for Bandwidth-Intensive Operations



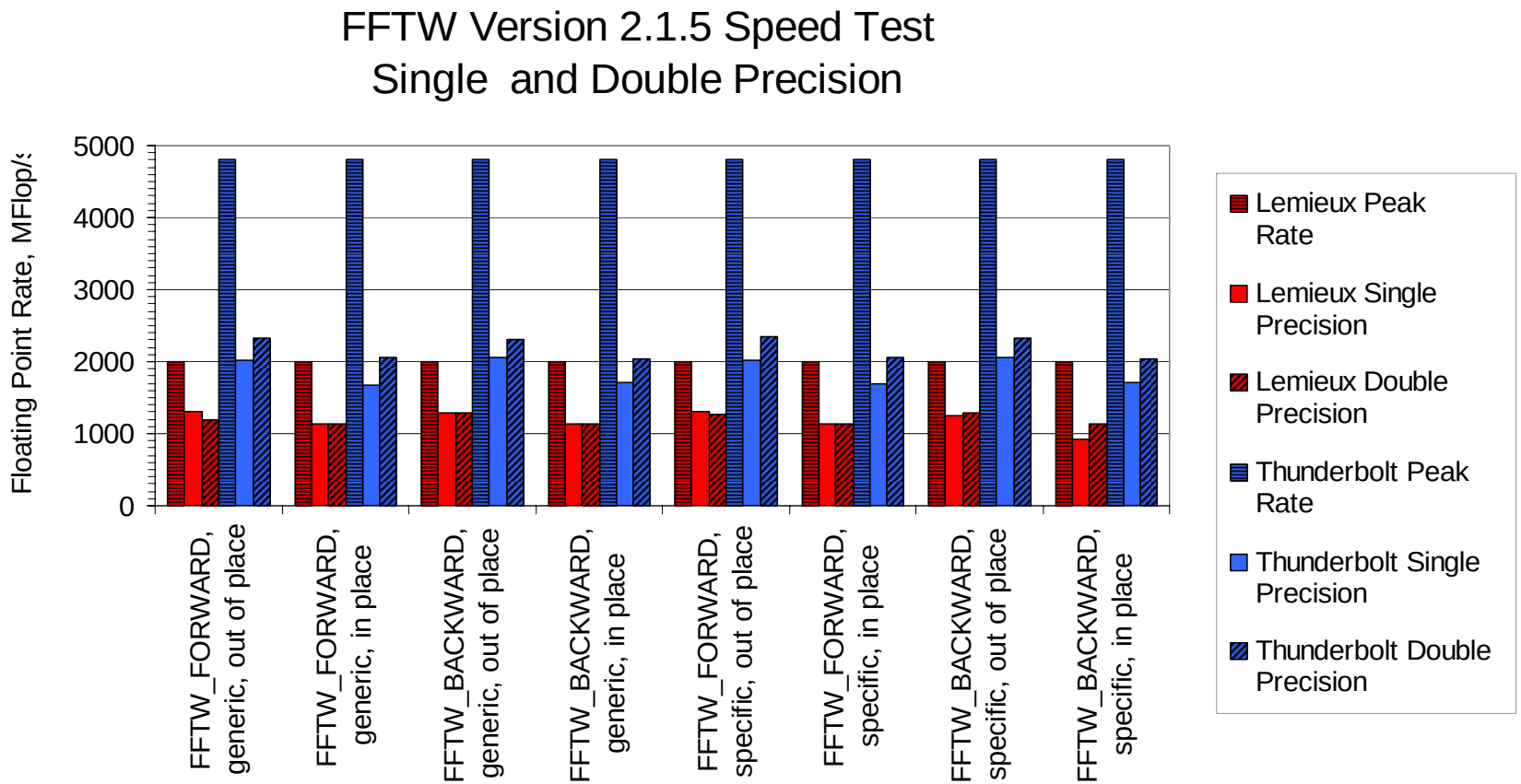
- Favorable scaling on FFTs and other transpose-intensive operations is essential to numerous applications

nodes	efficiency
32	1
64	1.01
128	0.990
256	0.918

Thanks to Jim Maltby, Cray Inc.

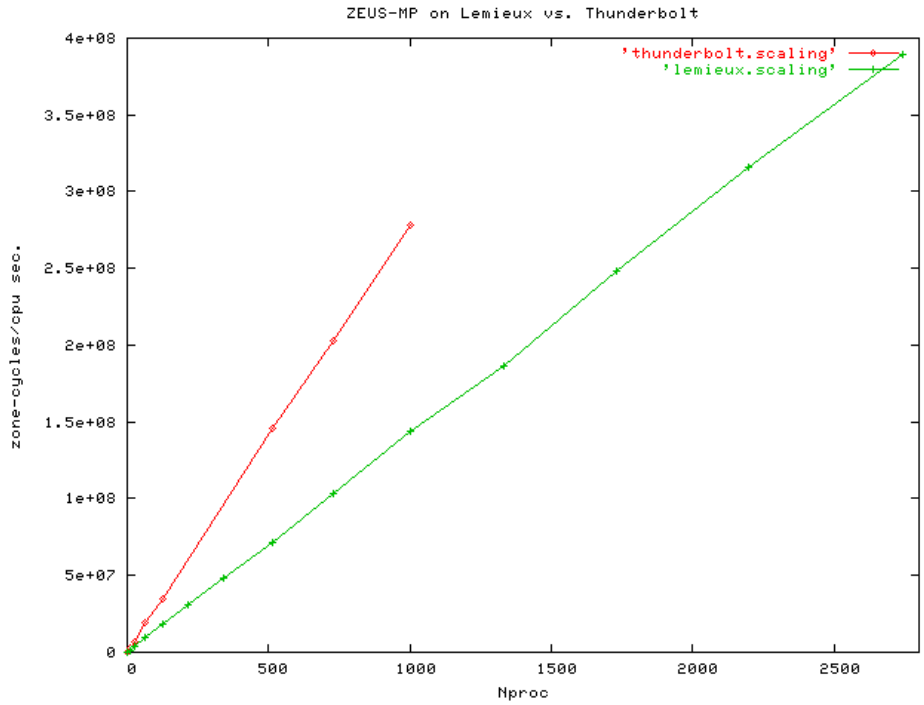
PITTSBURGH SUPERCOMPUTING CENTER

Single-processor FFT performance



ZEUS-MP

- Astrophysical CFD: solves the equations of ideal (non-resistive), non-relativistic, hydrodynamics and magnetohydrodynamics, including externally applied gravitational fields and self-gravity.
- Representative PSC users: Michael Norman, Mordecai-Mark Mac Low
- Figure depicts strong scaling, with performance expressed as number of zone-cycles solved per second.



Performance Metrics

- *The following performance figure are extremely preliminary and will improve as the system matures*
- STREAM
  - Copy : 4.798 GB/s per processor;
  - Scale : 4.783 GB/s per processor;
  - Add : 4.611 GB/s per processor;
  - Triad : 4.638 GB/s per processor;
- HPCC (552 nodes: P=24, Q=23)
  - HPL : 15.3 TFlop/s (76% of theoretical peak for N=100,000)
  - PTRANS : 49.573 GB/s
  - DGEMM : 4.32 GFlop/s
- Pallas MPI Benchmarks
  - ping-pong bandwidth: 1094.12 MB/s

PITTSBURGH SUPERCOMPUTING CENTER

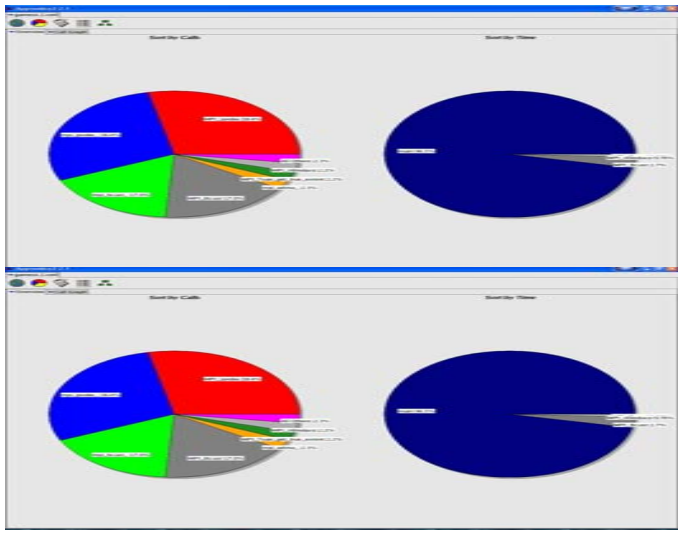
XT3 Programming Environment

- PGI 6.0.1
- Cray performance tools
  - Apprentice^2
  - PAPI 3
  - CrayPat
  - Thanks to Luiz DeRose for facilitating early access to Apprentice2 and related tools on the XT3!
- Etnus TotalView

PITTSBURGH SUPERCOMPUTING CENTER

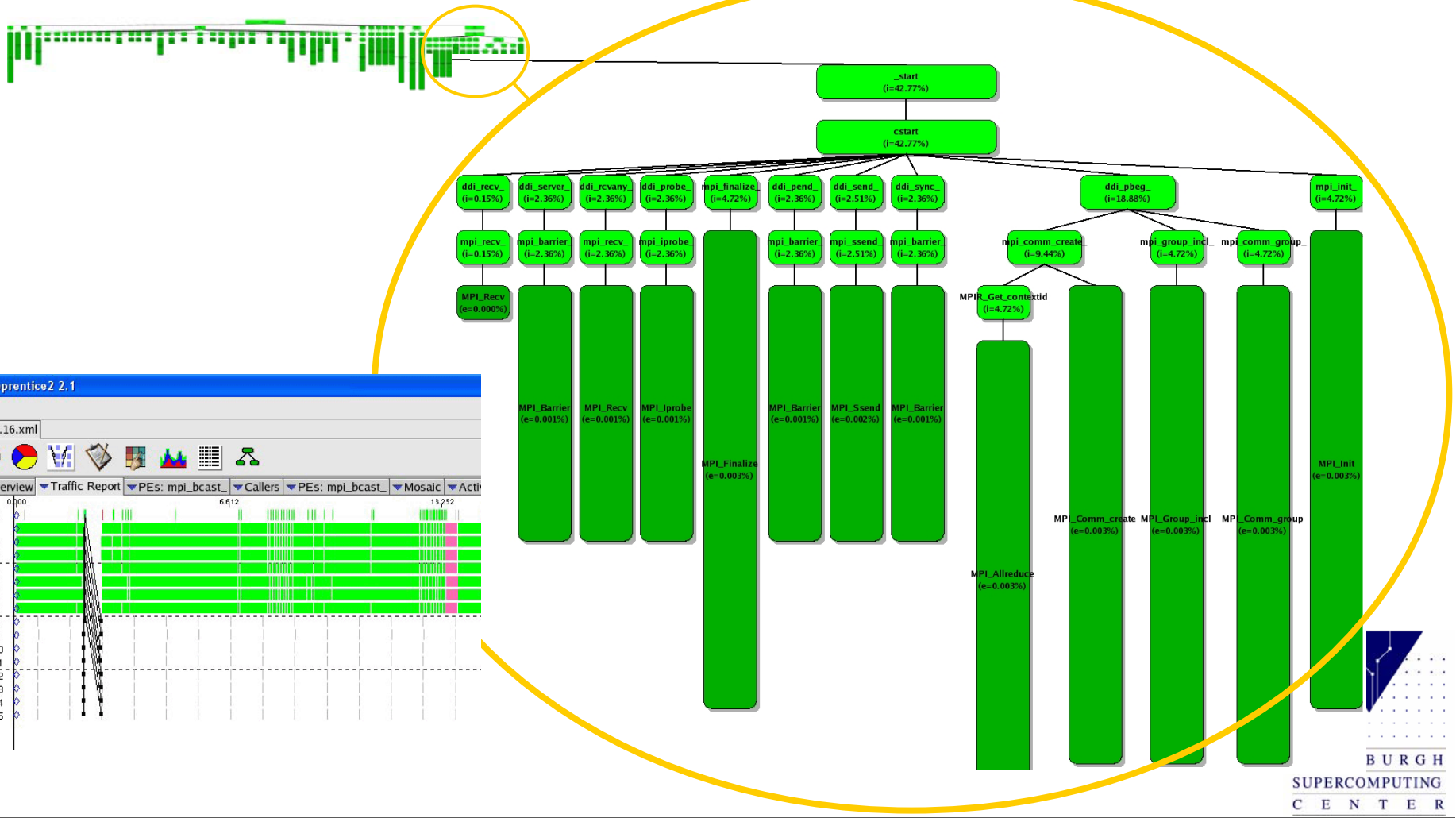
Performance Characterization

- Gaining experience with new Cray performance tools using GAMESS
  - General Atomic and Molecular Electronic Structure System: quantum chemistry
  - DDI layer abstracts communications
  - interesting due to code complexity, research applications, and relevance to other agencies
- Concurrent efforts with TAU (Al Malony, U. of Oregon)
  - numerical relativity, QCD, ...



PITTSBURGH SUPERCOMPUTING CENTER

Example: Call Graph Profile

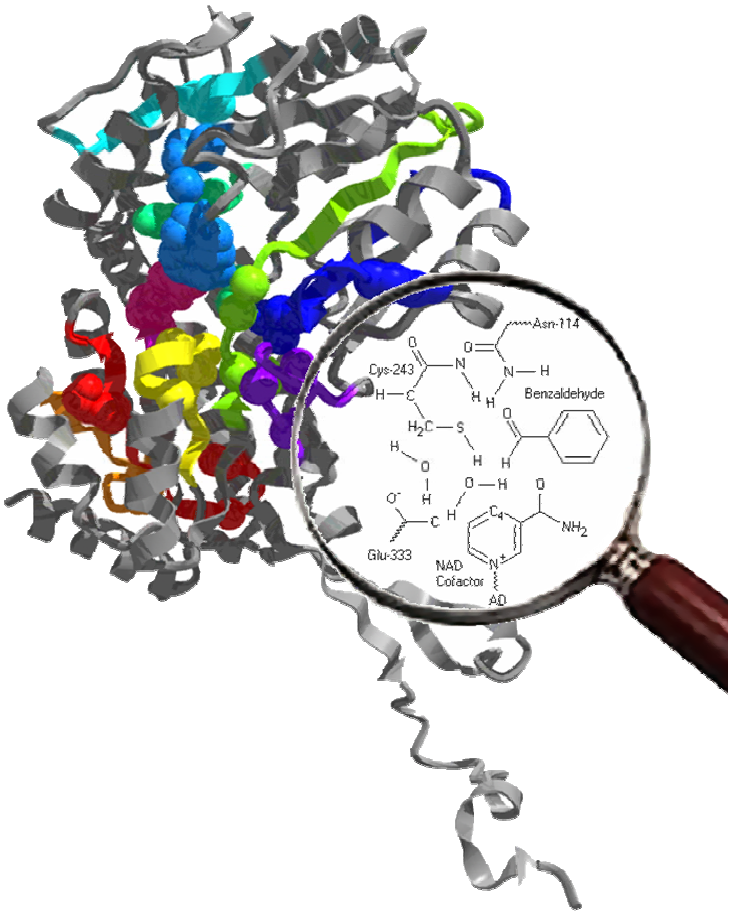


Scientific Targets

- As applications are developed for the Cray XT3, scientific runs have already begun in a “friendly user” period.
- Goals include:
  - Generate scientific results
  - Serve as a workload, facilitating deployment of a productive, stable system when production use commences

PITTSBURGH SUPERCOMPUTING CENTER

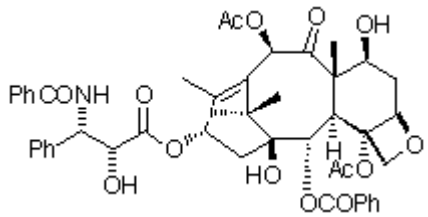
Hybrid Quantum Mechanical/Molecular Mechanical simulations on the Cray XT3 will enable new insight into the mechanisms of enzymes [Dynamo 2.2]



- Path-integral QM/MM simulations reveal the significance of quantum dynamical effects (proton tunneling) in Glutathione S-Transferase (GST) and Aldehyde Dehydrogenase (ALDH)
- Preliminary results support hypothesis that lack of intermediate stabilization is the molecular basis behind two different metabolic diseases, Hyperprolinemia Type II and Sjogren-Larsson Syndrome
- QM/MM simulations on coupled proton transfer-hydride transfer events in ALDH:
  - 36,738 atoms, 10ps, 10^4 timesteps
  - 900 processors: linear scaling
  - 9,450 CPU hours

Developing Faster Integration Schemes for Kohn-Sham Density Functional Theory

- Integrals that arise from the formulation of KS-DFT are of such a complicated nature that it necessary to use numerical quadrature in their solution.
- One method to increase efficiency of the numerical algorithm is to use interpolation from a Cartesian based grid to the angular grids commonly used.
  - Interpolation algorithm implemented in Q-Chem 2.1
- It is critical to analyze the convergence quality of current numerical schemes in order to assess the reliability of the interpolation algorithm
- GAMESS is being used on the Cray XT3 to validate numerical quality of the solution obtained from interpolation
  - 6-31G*/BLYP Taxol; 959 basis functions
  - 708 Cray XT3 nodes
  - Vary angular and radial grid size until convergent energy reached



PITTSBURGH SUPERCOMPUTING CENTER

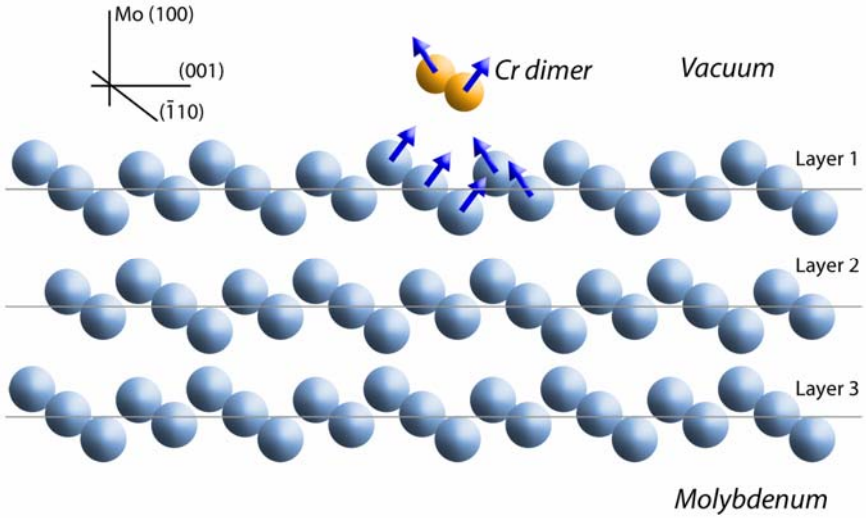
LSMS:
Towards Petacomputing in Nanotechnology

- Locally self-consistent multiple scattering (LSMS) method
  - a first-principles *O*(*N*) scaling technique
  - LSMS achieves 4.65 TFlops on TCS; 1998 Gordon Bell award
- The Cray XT3 and future computing systems will enable realistic quantum mechanical simulation, e.g. study of the dynamics of magnetic switching processes, of real nanostructures.
- Planned calculations will investigate electronic and magnetic structure of a 5nm cube of Fe, which contains approximately 12,000 atoms.
- Yang Wang, PSC
  Malcolm Stocks, D.M.C. Nicholson, and Markus Eisenbach, ORNL
  Aurelian Rusanu and J.S. Faulkner, Florida Atlantic University

PITTSBURGH SUPERCOMPUTING CENTER

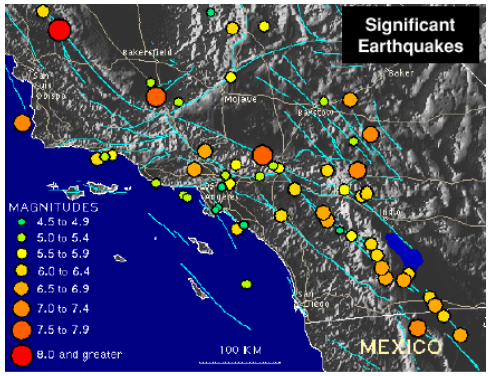
First principles approach to non-collinear magnetic structure of Cr dimers on Mo(110) Surface

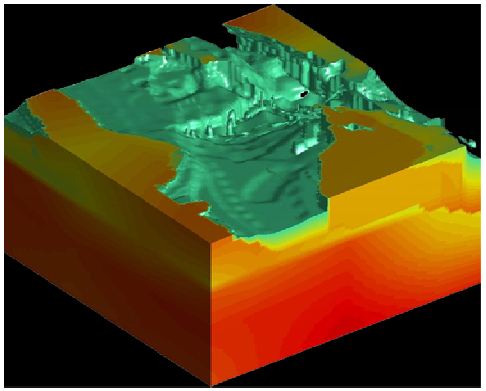
- Small clusters may display innovative properties in ultra small scales, suggesting promising applications.
- Cr has antiferromagnetic ordering in its bulk, described by a spin density wave with a wavelength incommensurate with the lattice constant.
- The locally self-consistent multiple scattering (LSMS) method for spin-dynamics and the full potential linearized augmented plane wave (FLAPW) method for magnetic anisotropy are adopted in the calculations.
- Yang Wang, PSC; Ruqian Wu, University of California, Irvine

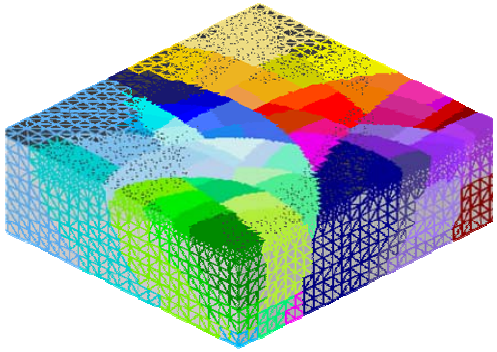


Using electronic structure calculations done with LSMS, Yang and Wu seek to understand interactions between magnetic moments of molybdenum atoms on the (110) surface and the chromium dimer.

Quake







- Complex ground motion simulation:
  - multiple spatial scales: of O(10m-100km); multiple temporal scales: O(0.01-100s); highly irregular basin geometry; highly heterogeneous soil material properties; geology and source parameters observable only indirectly
- Scientific Goals
  - Simulation of a magnitude 7.7 earthquake centered over a 230km portion of the San Andreas fault in southern California
  - 2Hz simulation, using a new adaptive mesh (~10B elements), will afford a 64× larger grid than the SCEC "Terashake" simulation (0.5Hz, 1.8B grid points)
- Scientific motivation: the 2Hz simulation will provide much better resolution and incorporate 4× high frequencies than the SCEC calculation, quantifying the effect of higher frequencies.
- Collaborators
  - Volkan Akcelik, Jacobo Bielak, Ioannis Epanomeritakis, Antonio Fernandez, Omar Ghattas, Eui Joong Kim, Julio Lopez, David O'Hallaron, Tiankai Tu, *Carnegie Mellon University;* George Biros, *University of Pennsylvania;* John Urbanic, *PSC*

AMBER / PMEMD

- Enabling larger molecular dynamics calculations will allow significant advances in biochemistry and structural biology.
- Simulations of 1-20ns, throttled by computational resources, are typical now. Being able to simulate 6ns/day on 128 Cray XT3 processors will allow simulations of 50-100ns to become commonplace.
- Simulation accuracy can be increased through use of polarizable force fields.
- Protein-protein interactions, critical for many biological phenomena, require simulations of sizes rarely attempted today.
- PMEMD XT3 work being done by Robert Duke (UNC-Chapel Hill), John Urbanic (PSC), Jim Maltby (Cray Inc.), and Troy Wymore (PSC).
- PSC users: Brooks, Simmerling, Cheatham, Roitberg

PMEMD: Factor IX (90,906 atoms, constant pressure, 10,000 timesteps)

Cray XT3 nodes	ns/day simulated
16	1.29
32	2.42
64	4.26
96	5.40
128	6.00

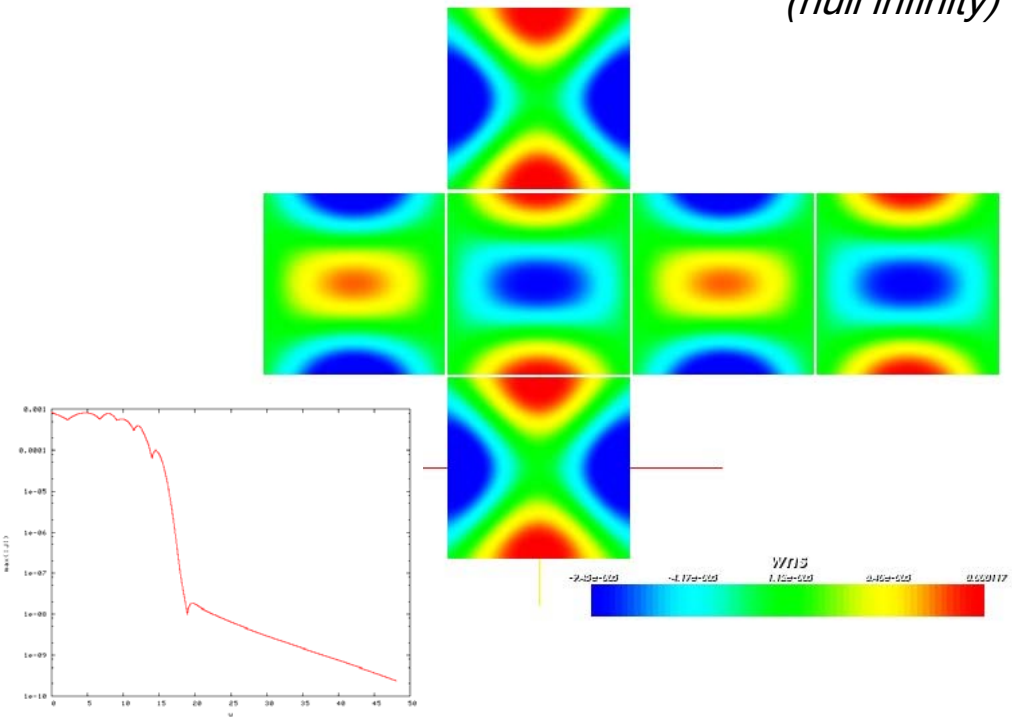
Gasoline

- Multi-platform, massively parallel *N*-body tree code used to simulate a variety of astrophysical processes
  - formation of disk (spiral) galaxies
  - tidal stirring of dwarf galaxies
  - cosmic microwave background & the Sunyaev-Zel'dovich effect
  - gas giant planet formation
  - intracluster light emission
  - galaxy cluster X-ray emissions
- Collaborators
  - George Lake, Tom Quinn, Joachim Stadel, *University of Washington*
  - James Wadsley, *McMaster University*
  - Jeffrey P. Gardner, *Pittsburgh Supercomputing Center*
  - Derek C. Richardson, *University of Maryland*

General Relativity

- Solves Einstein equations in vacuum, modeling single black hole spacetimes
- Code runs on 486 (6×9×9) nodes for 12,288 timesteps (execution time: 3h37m)
- Grid size:
  6×144^2 (angular) $\times$ 512 (radial)
- Norm of metric fields indicates correct execution
- PI: Roberto Gomez, PSC

Metric field W on the outermost sphere (null infinity)



PITTSBURGH SUPERCOMPUTING CENTER

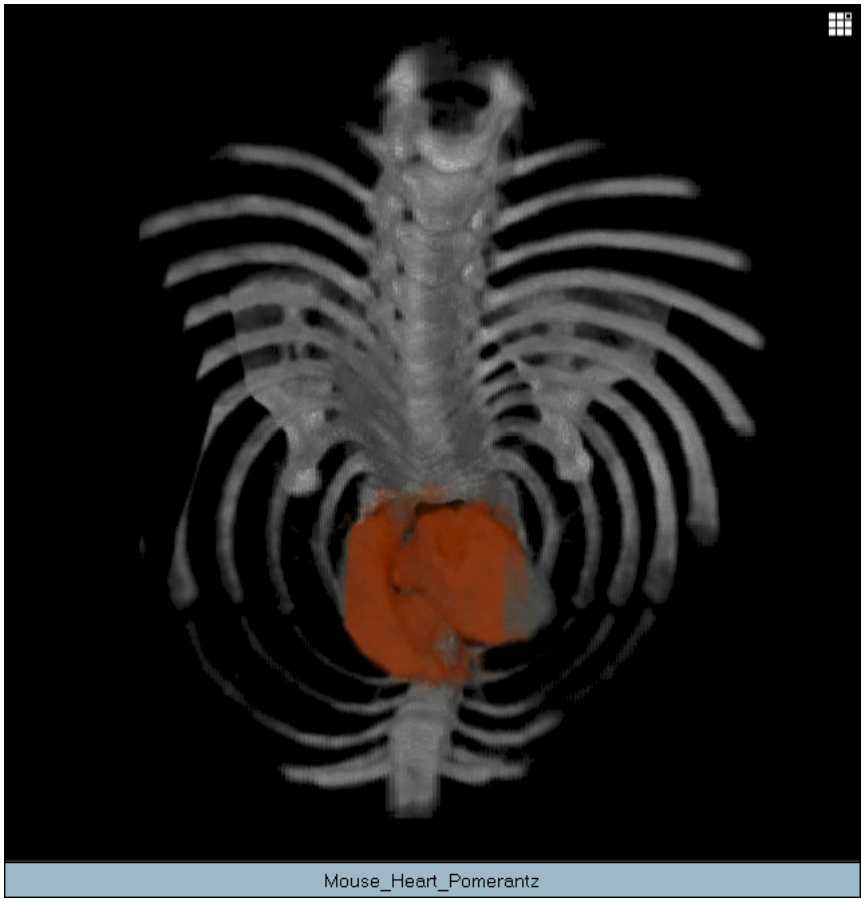
ARPS

- Advanced Regional Prediction System
  - comprehensive regional to storm-scale atmospheric modeling/prediction
  - includes real-time data analysis and assimilation, forward prediction, and post-analysis
  - observations confirm validity of ARPS simulations of intensive convective systems

- Kelvin Droegemeier and Ming Xue, *Center for Analysis and Prediction of Storms, U. of Oklahoma*

Volume Rendering

- 4D volume rendering of beating mouse heart
  Mouse_Heart_Pomerantz.mpg
- CT mouse data (200^3) courtesy of the Duke Center for In-Vivo Microscopy
- Rendered on the Cray XT3
- Rendering and animation:
  Art Wetzel, Stu Pomerantz, Demian Nave (PSC)



Summary

- Progress developing applications and preliminary performance have been highly encouraging
- CRMS is running on our full 22-cabinet configuration.
- We have had “friendly users” on the system since late 2004. Additional friendly users are being gated in regularly in preparation for full production.
- Releases 1.1 and 1.2 will bring increases in performance.
- PSC will continue to add value, with the primary focus of enabling computational science.

PITTSBURGH SUPERCOMPUTING CENTER