The ASCR Joule Metric for Computational Effectiveness

An Update on Recent Activities and Outcomes

D. Kothe, Oak Ridge Leadership Computing Facility, Oak Ridge National Laboratory

K. Roche, Pacific Northwest National Laboratory, University of Washington

U.S. DEPARTMENT OF ENERGY Office of Science

OAK RIDGE National Laboratory

Pacific Northwest NATIONAL LABORATORY

UNIVERSITY OF WASHINGTON 1861

Outline

- ASCR Joule metric for computational effectiveness
- FY09 Joule metric results
  - The benchmarking activity itself engenders performance gains by mutually challenging applications and leadership systems
  - Summary and recommendations
- FY04-FY09 Joule retrospective
- FY10 Joule metric status

The Joule Metric

(SC GG 3.1/2.5.2) Improve computational science capabilities, defined as the average annual percentage increase in the computational effectiveness (either by simulating the same problem in less time or simulating a larger problem in the same time) of a subset of application codes. Efficiency measure: X% (FY09, x=100)

Strong Scaling

"simulating the same problem in less time"

Algorithm, machine strong scaling :

Q4 problem := Q2 problem
Q4 algorithm := Q2 algorithm
Q4 machine ~ k * Q2 machine
Q4 time ~ 1/k * Q2 time

Algorithm enhancements, performance optimizations:

Q4 problem := Q2 problem
Q4 algorithm ~ enhanced Q2 algorithm
Q4 machine := Q2 machine
Q4 time ~ 1/k * Q2 time

*Could consider other variations: algorithm and machine are varied to achieve reduction of compute time

Weak Scaling

"simulating a larger problem in same time"

Algorithm, machine weak scaling (defined as 100%):

Q4 problem ~ k * Q2 problem
Q4 algorithm := Q2 algorithm
Q4 machine ~ k * Q2 machine
Q4 time := Q2 time

Algorithm enhancements, performance optimizations:

Q4 problem ~ k * Q2 problem
Q4 algorithm ~ enhanced Q2 algorithm
Q4 machine := Q2 machine
Q4 time := Q2 time

*Could consider other variations: problem, algorithm and the machine are varied to achieve fixed time assertion

Performance Enhancements
Examples from the machine perspective

Strong Scaling

Machine Events	Q2	Q4
INS	2.1466E+15	2.1130E+15
FLOPS	5.8962E+14	5.8947E+14
PEs	5632	11264
Time[s]	121.252233	57.222988

INS:
2113046508030116 /
2146627269408190 = **0.9843**

FP_OP:
589469277576687 /
589624961638025 = **0.9997**

PEs: 11264 / 5632 = **2**

Time[s]:
57.222988 / 121.252233 = **0.472**

Weak Scaling

Machine Events	Q2	Q4
INS	5.18E+17	1.93E+18
FLOPS	4.63E+17	1.81E+18
PEs	7808	31232
Time[s]	25339	23791

INS: 3.72

FP_OP: 3.92

PEs: 4

Time[s]: .938

NB: .938 * 4 = 3.752

Efficiency

Machine Events	Q2	Q4
INS	3.16E+12	4.37E+11
FLOPS	5.50E+11	5.53E+11
PEs	1	1
L2DCM	823458808	34722900
Time[s]	826.494142	79.414198

INS: 0.1381 (7.239x)

FP_OP: 1.0053 (0.99475x)

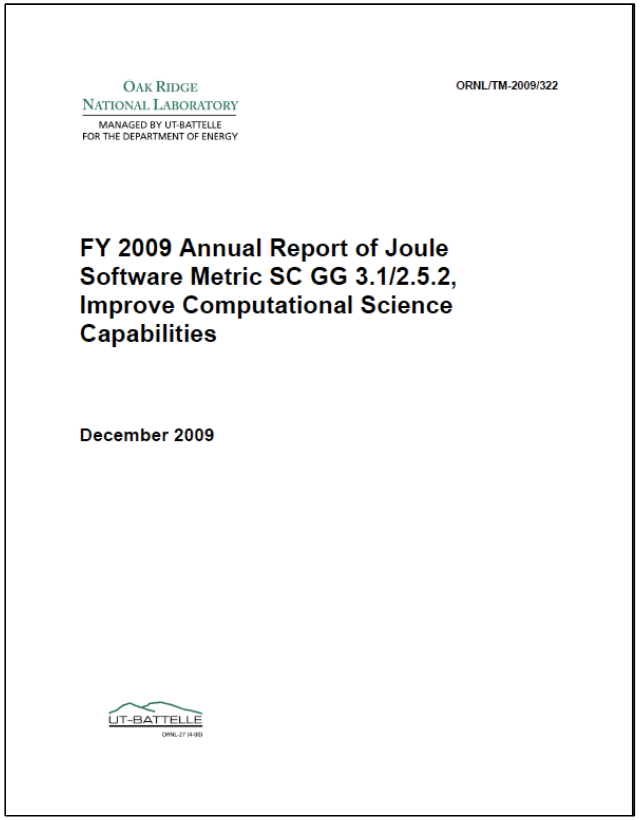
PEs: 1

L2DCM: 0.0422 (23.715x)

Time[s]: 0.0961 (10.407x)

Benchmark Information
Gathered per application

- **Description of Problem Domain, Target Problems**
- **Description of Application Software, Algorithm Implementation**
- **Benchmark Parameters Q2**
  - problem instance
  - build environment, build
  - runtime environment, run script
- **Benchmark Results Q2**
  - performance data
    - wall time
    - machine events
  - simulation results
- **Benchmark Parameters Q4**
  - problem instance
  - build environment, build
  - runtime environment, run script
- **Benchmark Results Q4**
  - performance data
    - wall time
    - machine events
  - simulation results
- **Comparative Analysis of Q2 and Q4 results**
  - description of problem related findings
  - description of software enhancements



OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

ORNL/TM-2009/322

FY 2009 Annual Report of Joule Software Metric SC GG 3.1/2.5.2, Improve Computational Science Capabilities

December 2009

UT-BATTELLE

http://www.nccs.gov/wp-content/media/nccs_reports/FY09Q4-JouleMetric-Report.pdf

U.S. DEPARTMENT OF ENERGY Office of Science
OAK RIDGE National Laboratory
Pacific Northwest NATIONAL LABORATORY

FY09 Joule Metric Results Summary
Applications: RAPTOR, CAM, XGC1, VisIT

Application	VisIt		CAM	XGC1	RAPTOR
Metric	Image construction/display time	Image construction/display time	Simulation time	Grind time and particle rate Time per time step Particles pushed per second	Grind time Time per cell per time step
Problem	Isosurface • 1,024 × 1,024 pixels • Iso @ 0.001, 0.01, 0.1, 1.0, 10.0, 100.0 • Q2 dataset: 103.7M cells, 4,096 cores, 27 groups • Q4 dataset: 321.1M cells, 12,720 cores, 27 groups	Volume render • 1,024 × 1,024 pixels • 2,000 samples per ray • Q2 dataset: 103.7M cells, 4,096 cores, 27 groups • Q4 dataset: 321.1M cells, 12,720 cores, 27 groups	1 simulated month • T341 mesh • 150 sec time step • 26 vertical levels • Spectral Eulerian core	DIII-D experimental tokamak • 13.5B particles • Q2: 4,000 time steps • Q4: 16,000 time steps	DLR-A configuration • 50 time steps • 110 × 40 jet diam in axial and radial directions • Q2: 10,285,056 cells • Q4: 24,261,120 cells
Hardware (cores) **Q2** **Q4**	 4,096 12,720	 4,096 12,720	 8,192 8,192	 29,952 119,808	 47,616 112,320
Time (seconds) **Q2** **Q4**	 0.01778 per contour 0.01686 per contour	 28.729 6.378	 6,481.724 3,241.144	 86,400 75,600	 1,034.0 444.0
Metric target	Q2:Q4 contour time ≥ 1.0	Q2:Q4 time ≥ 3.10	Q2:Q4 time ≥ 2.0	Q2:Q4 grind time ≥ 1.0 Q2:Q4 particle rate ≥ 4.0	Q2:Q4 grind time ≥ 1.0
Metric result	1.05	4.50	2.10	1.14 4.57	2.34

FY09 Joule Metric Results Summary

All applications exceeded target metric *and* expectations

RAPTOR

- DLR-A benchmark, super linear weak scaling
- 2.3588X increased hardware utilization

CAM

- Atmospheric T341 benchmark, super linear strong scaling result
- 2.1X reduction in compute time with fixed hardware utilization

XGC1

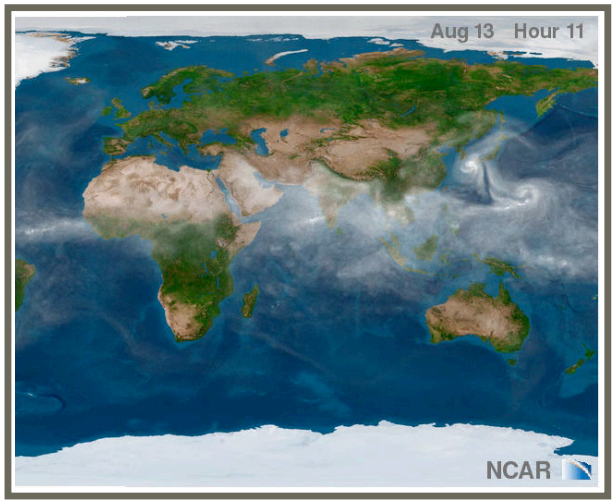
- DIII-D full-f benchmark, super linear weak scaling result
- 4X increased hardware utilization
- New result: steady state equilibrium achieved

VisIt

- Denovo power plant simulation data set, linear / super linear weak scaling
- 3.1X increased hardware utilization for isosurface and volume rendering

FY09 Joule Application: CAM
Community Atmospheric General Circulation Model

- V3.5 is 5th-generation design based on collaborative effort led by NCAR and Atmospheric Model Working Group
  - Physical parameterizations for prognostic cloud moisture, radiative effect of aerosols, long- & short-wave radiation interaction, interfaces with land & ocean
- Predictability on decadal time scales is critical for climate research
  - Requires a numerical model with high spatial resolution (<30 km resolution in longitude & latitude)
  - Global atmospheric and land surface models must be configured with forcing datasets that are best estimates of observed solar variability and greenhouse gas mixing ratios
  - Must adequately represent regional features, e.g., orographic precipitation signal
  - Realistic soil moisture pattern in the land model
  - Realistic representation of both extra-tropical and tropical storms

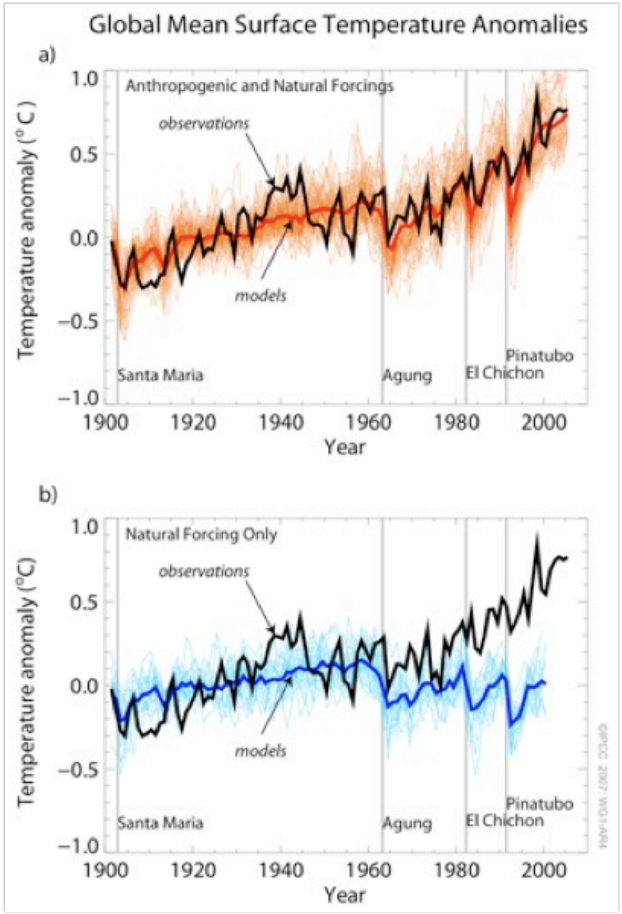


- Scientific demands do not at this time justify moving to a much finer mesh than tested (T341 mesh)
- Must reduce many weeks of computer time currently required to complete multi-decadal high-resolution simulations
  - Allows more realizations and ensembles

U.S. DEPARTMENT OF ENERGY Office of Science · OAK RIDGE National Laboratory · Pacific Northwest NATIONAL LABORATORY

FY09 Joule Application: CAM
Algorithm and Implementation

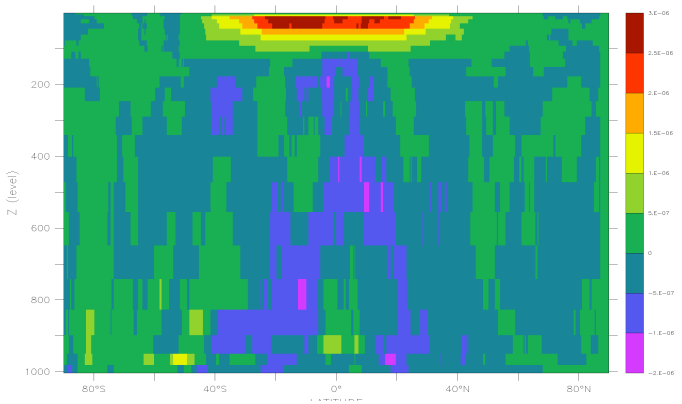
- Characterized by 2 computational phases
  - Resolved dynamics: evolution equations for atmospheric flow
  - Physics: subgrid-scale phenomena such as precipitation processes, clouds, long- and short-wave radiation transfer, and turbulent mixing
- Control moves between dynamics and physics at least once per time step
  - Dynamic-physics coupler moves info between data structures
- Multiple options for dynamics: spectral Eulerian, spectral semi-Lagrangian, finite-volume semi-Lagrangian
  - Spectral and FV dynamics use different grids
- Uses mixed-mode (OpenMP/MPI) parallelism
  - Physical parameterizations only in vertical dimension, all on-processor, and not load balanced



U.S. DEPARTMENT OF ENERGY Office of Science
OAK RIDGE National Laboratory
Pacific Northwest NATIONAL LABORATORY

FY09 Joule Application: CAM
Performance enhancements

- Limited to 512 MPI PEs
  - Requires fine grain parallelism
- Compiler flags and improvements
  - PGI 7.2.3 to PGI 9.0.1
  - Disable '–Kieee' flag
  - Enable '-Mvect=sse' flag
- Run-time configuration flags
  - Optimized number of vertical pencils per chunk
- Modifications to the I/O subsystem
  - Spider enabled 2X for the Q2 I/O



- **Differences in short-wavelength heating rates between 2 T341 CAM runs w/ & w/o volcanic aerosols**
- **Oct 1991 average during Mount Pinatubo eruption**
- **Y axis depicts vertical pressures (mbars)**
- **Red signifies areas where the volcanic CAM run has more heating**

- Different parallel strategy for dynamics and subphysics –requires transformation between data structures each time step
- Physical, chunk is set of vertical pencils (z-comp), each process computes some number of chunks; OMP task parallelism over chunks
- Short wavelength radiation calculation for sunlit chunks (1/2 not lit at any moment in simulation) – load imbalance
- Dycore, discretization is over bands of latitudes to MPI processes and when more than single latitude per process, OMP threads are assigned; work is 1D forward and reverse Fourier transforms

FY09 Joule Application: XGC1

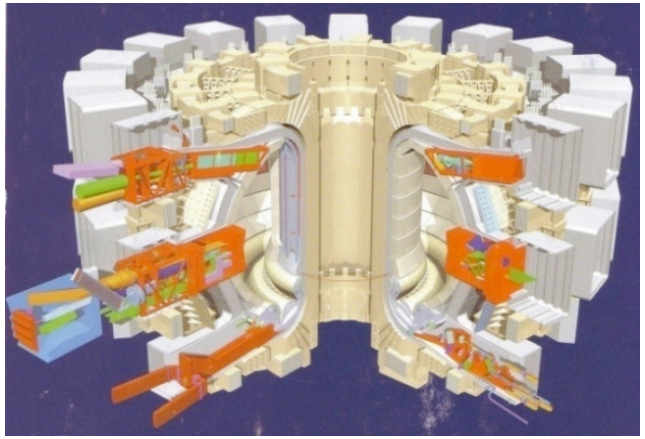
5D Gyrokinetic Full Function Particle-in-Cell Model for Whole Plasma Dynamics in Experimentally Realistic Magnetic Fusion Devices

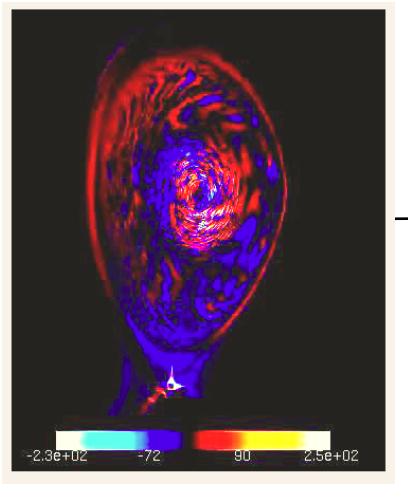
Model

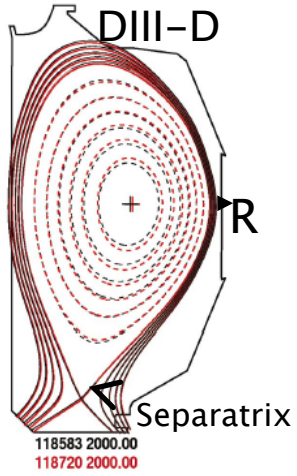
- Gyrokinetic "full-f" PIC model of magnetic fusion plasmas, with inclusion of magnetic separatrix, magnetic X-point, conducting material wall, & momentum/energy conserving Coulomb collisions
- Full-f description allows turbulence and background plasma to interact self-consistently and background plasma to evolve to a self-organized state
- **Focus:** understand and predict plasma transport and profile in the "edge pedestal" around separatrix

Algorithm & implementation

- Fixed unstructured grid following equilibrium magnetic field lines with embedded discrete marker particles representing ions, electrons, and neutral particles
- Marker particles time-advanced with Lagrangian equation of motion (either 4th order PC or 2nd order RK)
- Marker particle charges accumulated on grid, followed by gyrokinetic Poisson solve for electrostatic field
- PETSc for Poisson solve, ADIOS for I/O, Kepler for workflow, Dashboard for monitoring/steering







U.S. DEPARTMENT OF ENERGY Office of Science

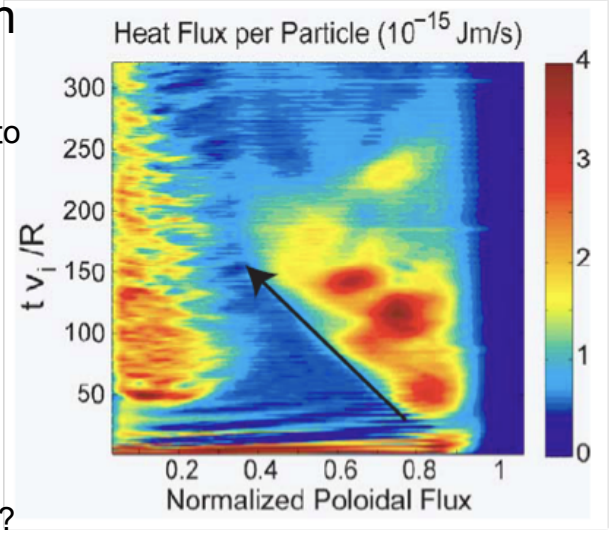
OAK RIDGE National Laboratory

Pacific Northwest NATIONAL LABORATORY

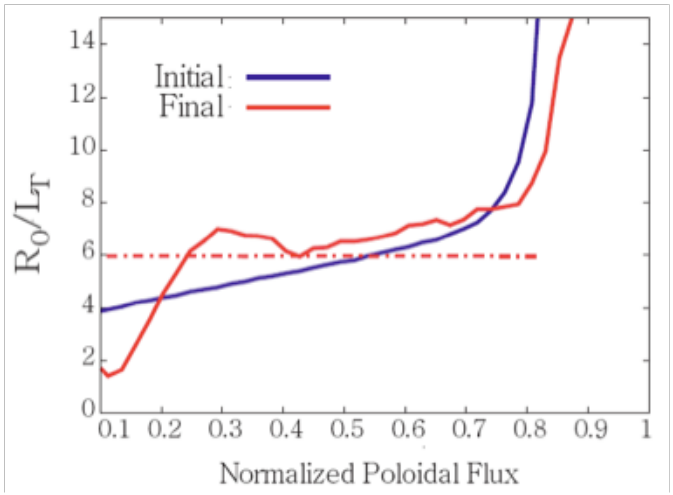
UNIVERSITY OF WASHINGTON 1861

FY09 Joule Application: XGC1

- High-confinement mode ("H-mode") and operation appears to be required for adequate yield ratios (Q>10) in magnetic toroidal fusion plasmas
  - At high enough core heating, plasma can bifurcate from low density/T state @ edge to very high just inside of magnetic separatrix; core temperature then continues to rise without the high T plasma contacting the wall (the "edge pedestal")
  - Core ion T increases in proportion to the edge pedestal T, with its radial slope being "stiff" and independent of the core heating power, entering into the "H-mode" of operation



- Many aspects of the H-mode remain poorly understand over the last 25 years
  - Why does the edge pedestal form this shape? Why is strong core heating necessary? Why is there an instantaneous central T_i and turbulence improvement after H-mode bifurcates? Why is the radial T_i profile stiff?



- First attempt to study the nonlocal H-mode coupling physics between the edge and core turbulence in a realistic DIII-D tokamak geometry
  - **Initial stage:** turbulence intensity propagation from edge to core, as a result of nonlocal interaction between edge and core. Initial turbulence intensity is strong and bursty. Plasma conditions not yet close to experimental state. **(Q2)**
  - **Final stage:** plasma in self-organized quasi steady-state, allowing probing of unexplained experimental H-mode phenomena **(Q4)**

U.S. DEPARTMENT OF ENERGY Office of Science · OAK RIDGE National Laboratory · Pacific Northwest NATIONAL LABORATORY · UNIVERSITY OF WASHINGTON

FY09 Joule Application: XGC1
Performance enhancements

- Solving gyrokinetic Poisson equation requires interpolating charges to grid points
- Solutions have to be interpolated back to particle positions to time evolve according to eqns of motion
- B field is evaluated employing spatial splines at each spatial position
  - Precompute and store spline coefficients -search instead of recompute
  - Used common partial results in the computation of derivatives significantly decreasing the number of required floating operations per time step
- Improve MPI communication in Poisson solution
- Improve MPI communication in the reassignment of particles to processes
- OMP parallelism was implemented allowing the use of 1/4 as many MPI processes

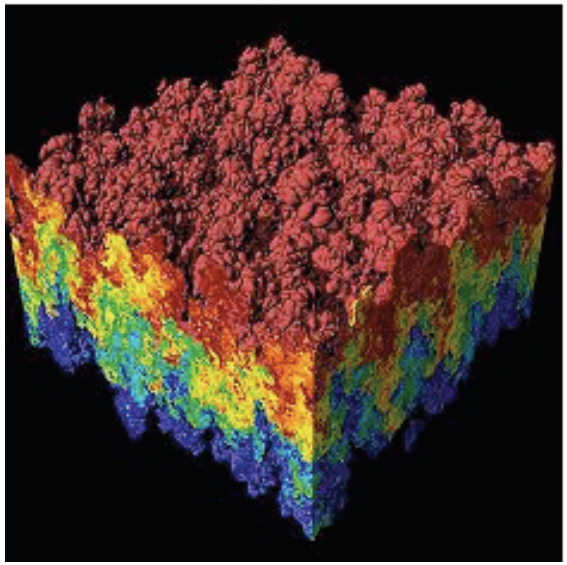
FY09 Joule Application: VisIt

Interactive parallel visualization and graphical analysis tool

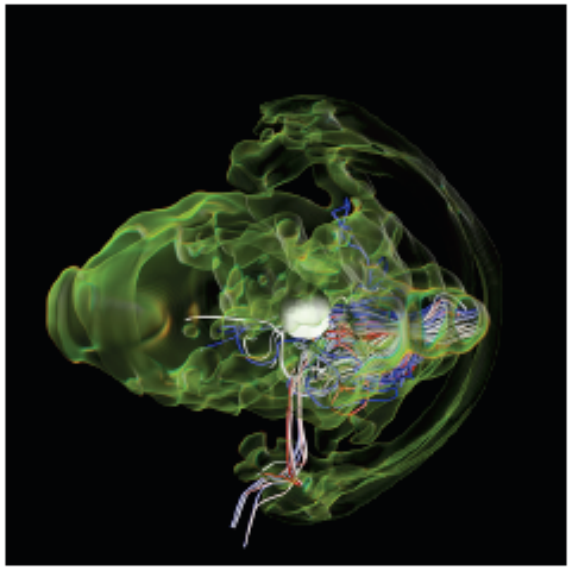
data exploration quantitative analysis comparative analysis

visual debugging communication of results remote visualization

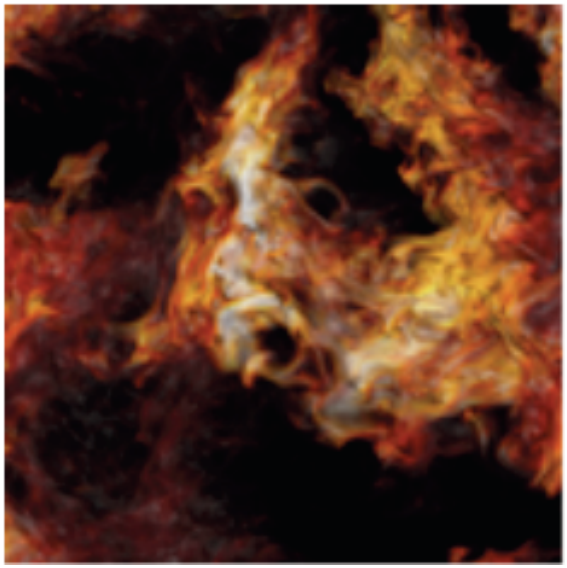
Isosurface Extraction



Streamlining



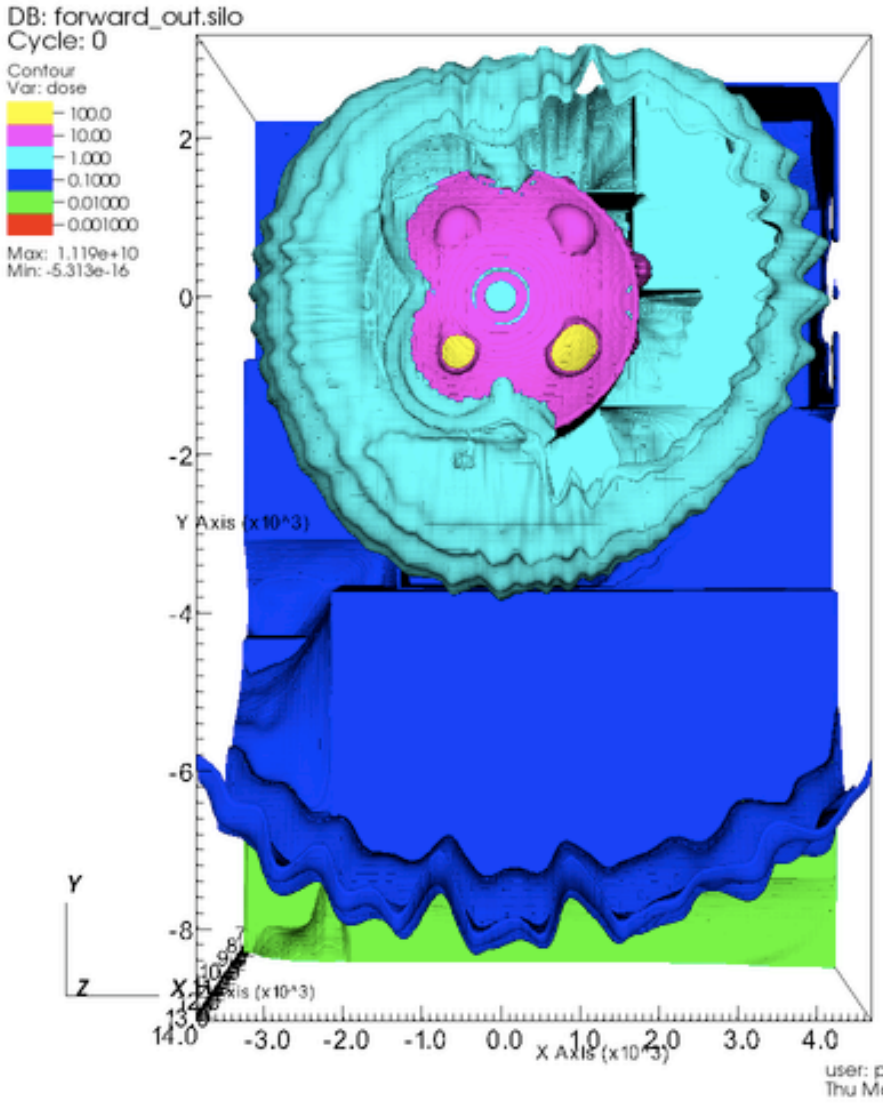
Volume Rendering



* large data strategy is to use distributed memory data parallelism

* each process creates an identical data flow network and works in MIMD model

FY09 Joule Application: VisIt
Isosurface benchmark results



Isosurface extraction :

extract the three dimensional points in a volume with a specific value and connect them with a continuous surface

Contours at isovalues :
0.001, 0.01, 0.1, 1.0, 10.0, 100.0

Resolution : 1024 x 1024 pixels

***27 energy level flux values used by VisIt to compute the dose variable scalar field**

U.S. DEPARTMENT OF ENERGY Office of Science
OAK RIDGE National Laboratory
Pacific Northwest NATIONAL LABORATORY

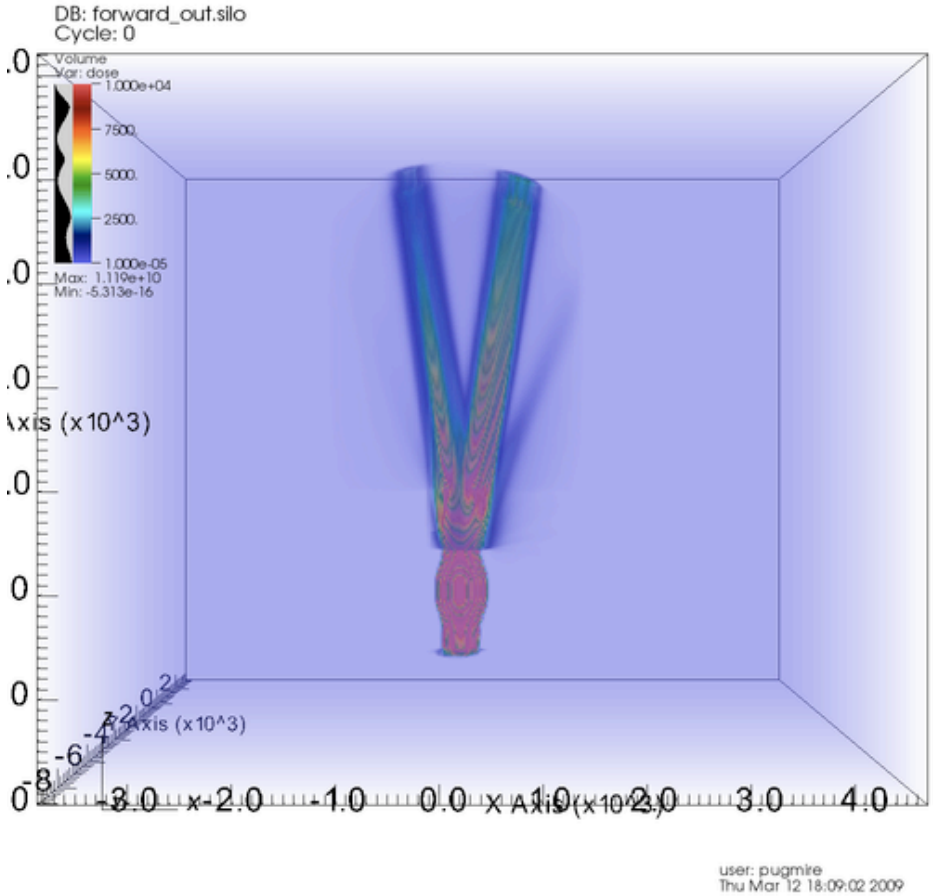
FY09 Joule Application: VisIt
Volume render benchmark results

produces image from a scalar field in a 3D data set

transfer function maps 3D sample points to a color and opacity in the final image

a ray is cast (1 per pixel) into the data volume where a number of uniform samples of the scalar field are taken

each data sample along the ray is assigned a color and opacity from the transfer function



Number of Samples per Ray : 500, 1000, 2000, 4000

Resolution : 1024 x 1024 pixels

Viewpoint : data centered

***27 energy level flux values used by VisIt to compute the dose variable scalar field**

U.S. DEPARTMENT OF ENERGY Office of Science

OAK RIDGE National Laboratory

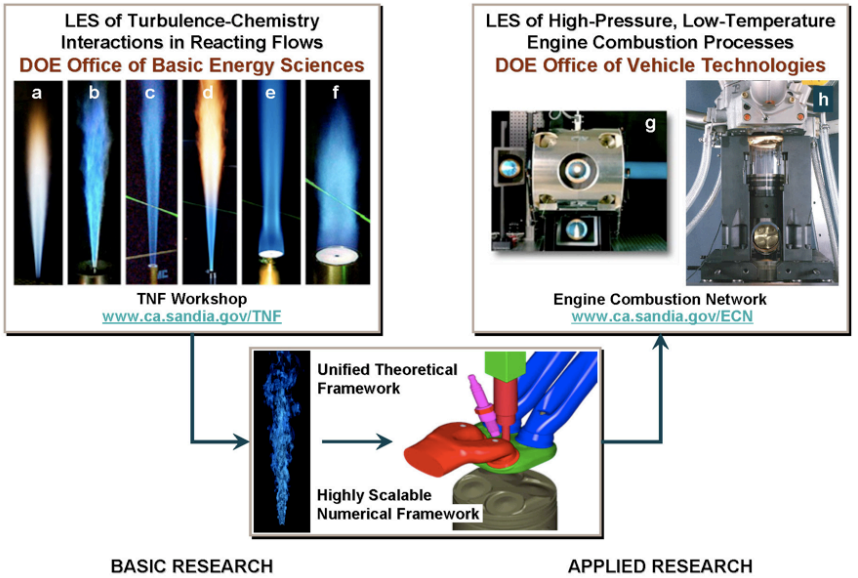
Pacific Northwest NATIONAL LABORATORY

FY09 Joule Application: VisIt
Performance enhancements

- **Isosurface benchmark**
  - No modifications to the isosurface capability were made
- **Volume render benchmark**
  - Corrected communication bottleneck
    - Each process assigned ~same number of mesh cells and generates local samples per ray for the cells it has; followed by assembly of samples along a given ray but over some subset of processors
    - Data was redistributed so that all of the samples along a given ray were located on a single processor … bad all-to-all phase
      - Removal of all-to-all and simple assignment of pixels to processors without concern for the distribution of the sample points
  - Tiling –image is composed of many smaller images
    - Cells per PE fixed prior to volume rendering –a PE may have no work to perform on a given tile and will wait until the next tile is ready … decreasing the parallel efficiency of the algorithm
    - disabled tiling algorithm for parallel environments

FY09 Joule Application: RAPTOR

Large eddy simulation of turbulent, reactive, multiphase flows



Software Implementation

- Distributed multi-block domain decomposition with a generalized connectivity scheme
- Parallelism implemented via MPI and the Single-Program–Multiple-Data model
- Generalized hexahedral cells
- Fully modular, self-contained, and written in ANSI standard Fortran 90
- Extensively validated over last 16 years

Fully coupled conservation equations of mass, momentum, total-energy, and species for a chemically reacting flow system (gas or liquid) in complex geometries

- ✶ Detailed chemistry, thermodynamics, & transport processes at the molecular level and uses detailed chemical mechanisms
- ✶ Generalized subgrid-scale model framework
- ✶ Spray combustion processes and multiphase flows using a Lagrangian-Eulerian formulation

Temporal integration scheme employs an all Mach number formulation using dual-time stepping with generalized preconditioning

- ✶ Fourth-order accurate in time and provides a fully implicit solution using a fully explicit (highly-scalable) multistage scheme in pseudo-time

- Non-dissipative spatial scheme that is discretely conservative, with staggered, finite-volume differencing stencils
  - ✶ Formulated in generalized curvilinear coordinates with a general R-refinement adaptive mesh (AMR) capability.

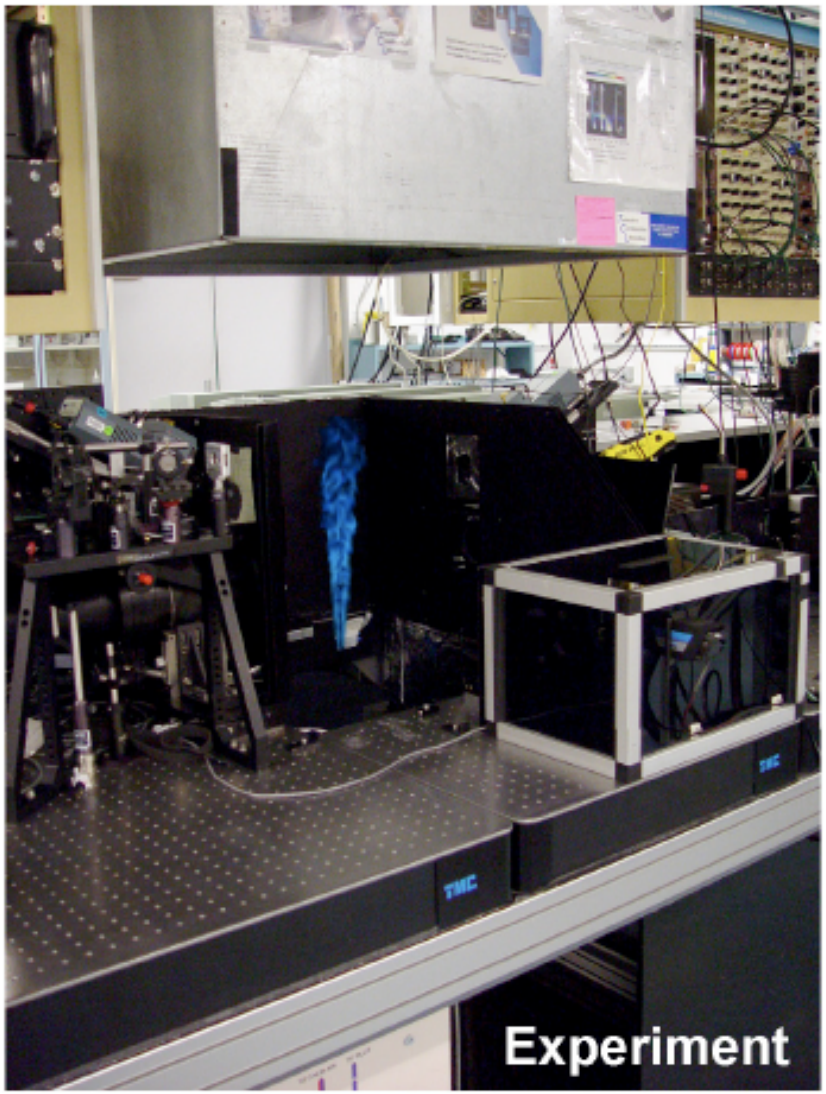
U.S. DEPARTMENT OF ENERGY Office of Science | OAK RIDGE National Laboratory | Pacific Northwest NATIONAL LABORATORY | UNIVERSITY OF WASHINGTON

FY09 Joule Application: RAPTOR

Large eddy simulation of turbulent, reactive, multiphase flows

- How can simulation "bridge the gap" between basic research and conditions of interest in typical applications?
  - Focus: application of LES models to low-temperature, high-pressure IC-engines
  - Establish high-fidelity computational benchmarks that match geometry and operating conditions of key target experiments using a single unified theoretical-numerical framework
  - Establish a scientific foundation for advanced model development
- Understanding and applying Reynolds number (Re) scaling in combustion modeling is crucial for simulation is to affect engine design
  - Focus on flames studied in the Reacting Flow Research Program at SNL – in particular passive scalar mixing – in a baseline flame (DLR-A experiment) configuration
  - Challenge: most data at Re~10^4 or less; IC engines typically run at Re~10^5 or greater
- Can reliable Re scaling relationships for turbulent flame dynamics and scaling mixing processes be devised appropriately?
  - Pushes mesh resolution up hence a weak scale driver
- Perform a series of weak scaling studies to demonstrate effects of increasing Re (starting from 15.2K) on scalar mixing dynamics
  - These benchmarks provide a direct one-to-one correspondence between measured and modeled results at conditions unattainable using DNS - simulations represent the fully coupled dynamic behavior of a reacting flow with detailed chemistry and realistic levels of turbulence.

FY09 Joule Application: RAPTOR
Benchmark motivation



1. study the effects of LES grid resolution on scalar-mixing processes

2. understand the relationship between the grid spacing and the measured turbulence length scales from a companion set of experimental data (DLR-A, shown here)

3. study the effects of increasing jet Reynolds number on the dynamics of turbulent scalar-mixing

DLR-A Flame: Re_d = 15,200

Fuel: 22.1% CH_4, 33.2% H_2, 44.7% N_2

Coflow: 99.2% Air, 0.8% H_2O

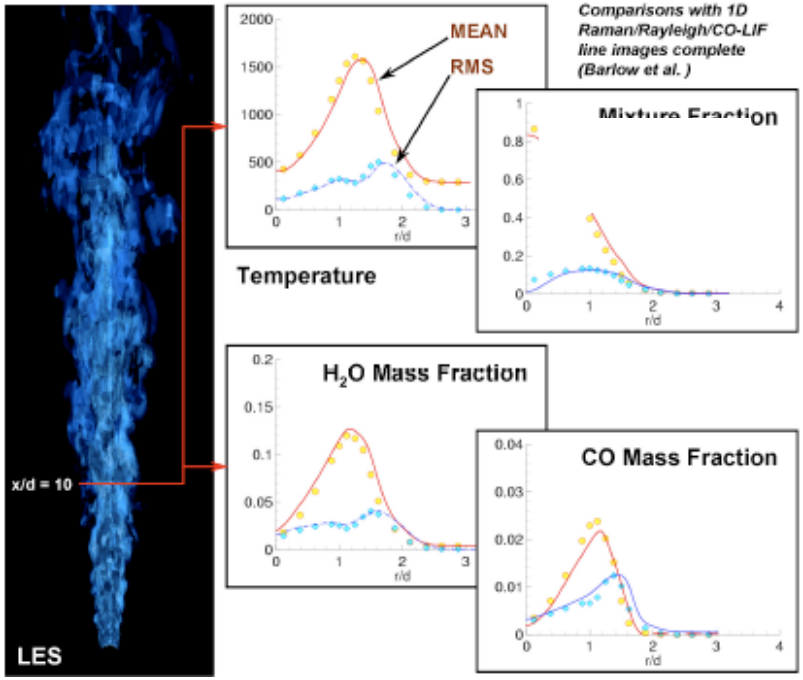
Detailed Chemistry and Transport: 12-Step Mechanism (J.-Y. Chen, UC Berkeley)

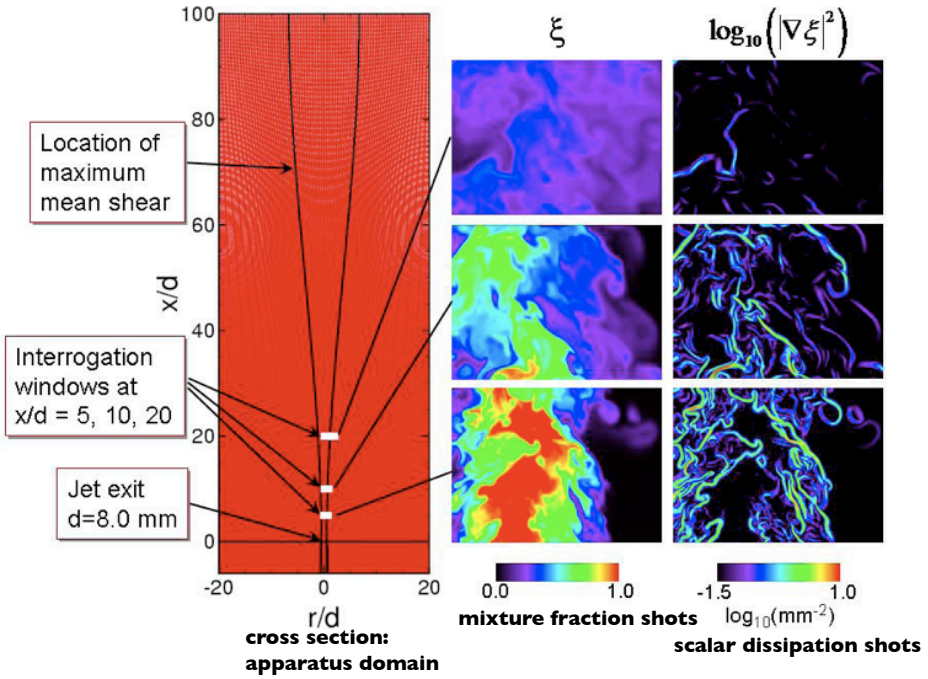
U.S. DEPARTMENT OF ENERGY Office of Science · OAK RIDGE National Laboratory · Pacific Northwest NATIONAL LABORATORY · UNIVERSITY OF WASHINGTON

FY09 Joule Application: RAPTOR
Benchmark configuration

Grid Number	Total Cells	Δt (Re_d = 15,200)
1	1,285,632	1.00 μs
2	10,285,056	0.50 μs
3	82,280,448	0.25 μs

50 physical time steps per grid





Domain: entire burner geometry (inside the jet nozzle and the outer co-flow) + downstream space around burner
Inner nozzle diameter : 8.0 mm
Outer nozzle : surface is tapered to a sharp edge at the burner exit
Specifics: 110 inner jet diameters in the axial direction (88cm) x 40 jet diameters in the radial direction (32 cm)

U.S. DEPARTMENT OF ENERGY Office of Science · OAK RIDGE National Laboratory · Pacific Northwest NATIONAL LABORATORY · UNIVERSITY OF WASHINGTON

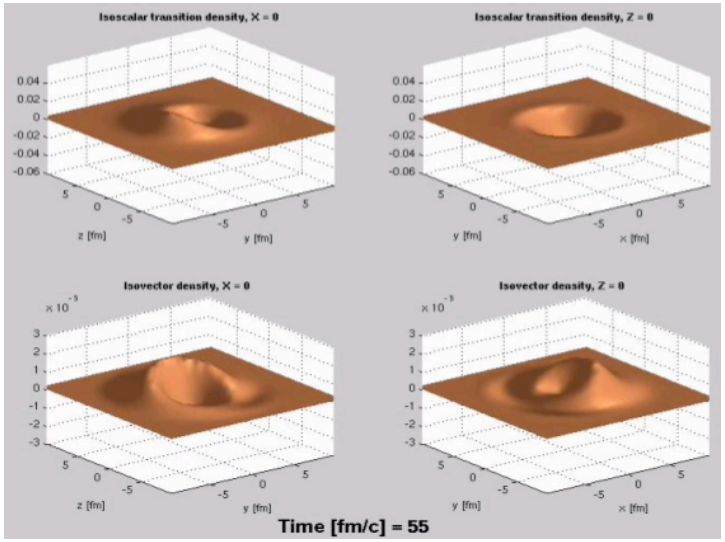
FY09 Joule Application: RAPTOR
Performance enhancements

- **Halo exchanges are nearest neighbor only**
  - ✳ Initial configuration: send/receive calls in pairs corresponding to each neighbor
  - ✳ Fix:

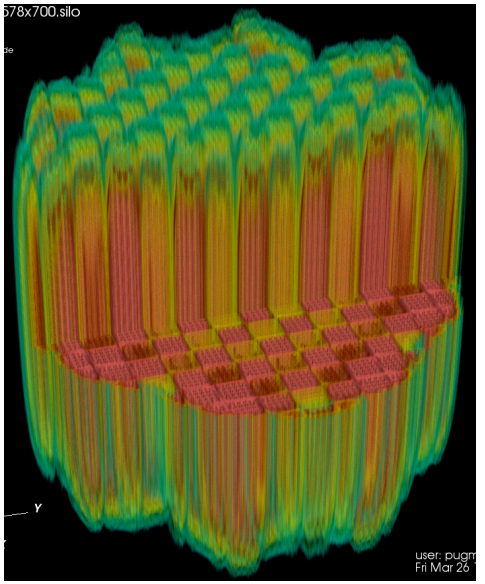
  - prepost all receives as the first operation in the routine (if buffer available)
  - post the sends as soon as the data is available
  - postpone the waits on send operations until the end of the routine. Non-blocking sends and receives are used throughout
  - Interleave computation to give more breathing room for communication
- **Removal of several unnecessary MPI barriers**
- **Convergence of the dual time integrator**
  - ✳ global MPI_allreduce for computing the error norm each iteration

  - use the fact that the number of pseudo-time iterations for convergence does not vary much between consecutive time-steps
  - assign a static variable X to the last pseudo-time step in which convergence was achieved in the previous physical time-step and wait X -1 pseudo-timesteps before computing expensive convergence check

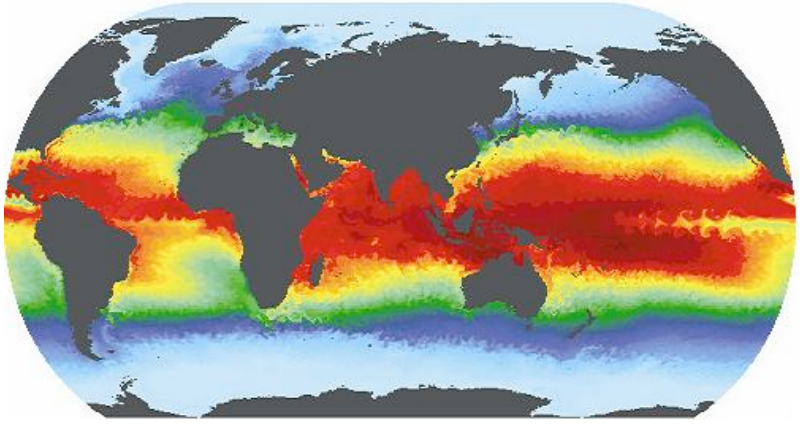
FY10 Metric Applications



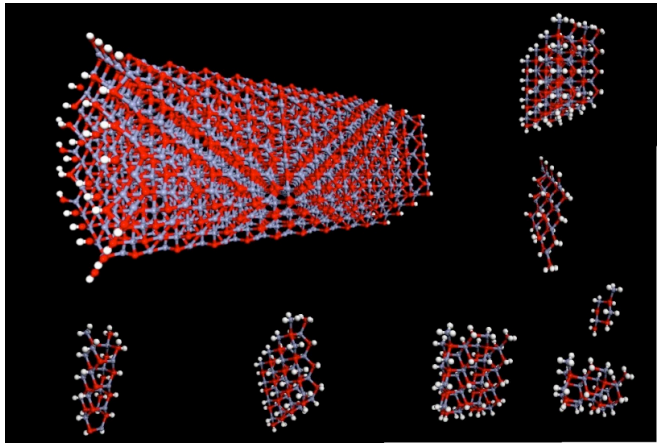
TD-SLDA



Denovo



POP



LS3DF

U.S. DEPARTMENT OF ENERGY
Office of Science

OAK RIDGE National Laboratory

Pacific Northwest NATIONAL LABORATORY

FY10 Joule Metric Status: On Track

Q2 baseline application performance results

Table 1. FY10 Joule software summary of Q2 baseline performance simulations and data

Application	TD_SLDA	POP	LS3DF	Denovo
Metric	Simulation time	Simulation time	Simulation time	Simulation time
Problem	*Vortex study –fermi gas* (self-consistent solver + dynamics) • 5216 particles • 103,917 quasiparticles • 50 x 50 x 100 lattice • 2050 time steps • 25 data io events • 1 check-point, 1 restart *Nuclear study* • Z=74, N=124 • 40 x 40 x 40 lattice • 7466 p-quasiparticle • 8946 n-quasiparticle • 20 time steps • .75fm spacing • 100MeV cutoff	*3 simulated days, ocean-only model* • 0.1-degree tripole global grid (3600×2400) • 42 vertical levels • 10 minute time steps • High-frequency output time slice	*Self-consistent DFT calculation for ZnO nanorod* • 2776 atoms • 24220 valence electrons, d-electrons in valence band • 720 × 300 × 300 numerical grid	*Full Core EDF PWR900 benchmark* • 17x17 fuel assemblies • 17x17 fuel pins per assembly • 2x2 cells per pin cell • 3 fuel enrichments • 45 homogenized pin cell materials per assembly • 135 different pin cell materials • 233,858,800 (578x578x700) cells • 168 angles, 1 moment, 2 energy (fast and thermal) groups • 7.86×10^{10} total unknowns
Hardware (cores) Q2	fg: 7344, 103917 n: 73728, 16412	4800, 9600, 14,400	21,600	17,424
Time (seconds) Q2	fg: 11085.3, 21205.1 n: 6538.5, 2084.4	420, 330, 350	4653	11,260.8
Metric target	Q2:Q4 time ≥ 1.0	Q2:Q4 time ≥ 2.0	Q2:Q4 time ≥ 2.0	Q2:Q4 time ≥ 1.0

Joule Metric Exercise
Summary and recommendations

- **Science drivers**
  - Work with science teams in quantifying a "bottom line number" for science metrics
  - Be cautious about targeting full-system Q4 problems on systems that are new and/or being upgraded
- **Performance data collection**
  - Need a verified standard approach for collecting performance
- **Documenting in detail the methodologies leading to performance gains**
  - Take note of the "do's" and "don't's" for HPC apps performance
  - Performance bottlenecks and solutions are often common and can be shared
- **Bring Joule application teams together as one project team**
  - Hold weekly meetings/telecons, email reflector (joule-metric-apps@email.ornl.gov)
- **Share results and documentation**
  - Drives accountability, competition, and emulation between code teams
- **Provide ample leadership computing resources**
  - That are not implicit in any pre-existing allocation (33M hours used in FY09!)
- **Bring in off-code-team experts (e.g., SciDAC PERI) as part of the Joule effort**

Questions?

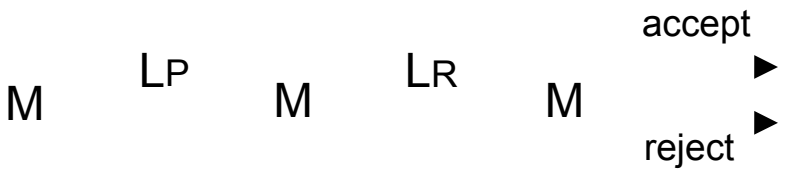
Douglas B. Kothe (kothe@ornl.gov)
Kenneth J. Roche (kenneth.roche@pnl.gov)

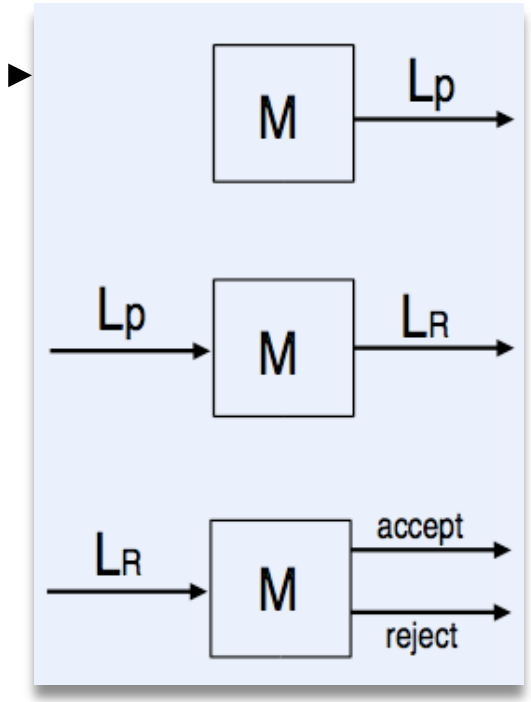
U.S. DEPARTMENT OF ENERGY Office of Science
OAK RIDGE National Laboratory
Pacific Northwest NATIONAL LABORATORY

Supplemental Material

Joule Measures Problem Complexity
By directly monitoring functional unit program counters

- COMPLEXITY
- PROBLEMS
- ALGORITHMS
- MACHINES





Measured time for machine M to generate the language of the problem plus time to generate the language of the result plus the time to accept or reject the language of the result.

Asking questions, solving problems is recursive process

Accepting a result means a related set of conditions is satisfied

$$S = S1 \wedge S2 \wedge \ldots \wedge Sn$$

FY09 Joule Metric Applications

application software	application metric	target platform	problem	joule result
(09)CAM	wall time / simulated year	Cray XT5 8192 pes, 3241.14s	•1 simulated month • t341 mesh •150s time step •26 vertical levels •spectral Eulerian core	strong(k=2)
(09)RAPTOR	wall time / cell / time step	Cray XT5 112,320 pes, 444s	•DLR-A system •50 time steps •110x40 jet diameter in axial, radial comp. •24,261,120 cells	weak(k>2.35)
(09)VisIt	image construction / display time	Cray XT5 12,720 pes, 6.38s	volume render •1024x1024 pixels •2000 samples / ray •321.1M cells •27 groups, 12720 domains	weak(k>3.1)
(09)XGC1	wall time / time step / particle	Cray XT5 119,808 pes, 75,600s	DIII-D experimental tokamak •13.5 billion particle •16000 time steps	weak(k>4)

U.S. DEPARTMENT OF ENERGY Office of Science · OAK RIDGE National Laboratory · Pacific Northwest NATIONAL LABORATORY · UNIVERSITY OF WASHINGTON

FY09 Joule Application: VisIt
Tackles important analysis problems

- Isosurfacing and volume rendering are likely the most common scientific inquiry functions in current end-to-end workflows
- Scientific workflow requirements for Visualization tools such as VisIt include
  - **Scalability:** data sets are either very large (multi TB) or distributed (10^3s of files)
  - **Speed:** scientists need rapid turn-around on the final rendered image for quick "hypothesize-test" cycles
  - **Remote Service:** all inquiries and functions must be able to be delivered remotely (client-server model)
  - **Dataset agnostic:** data coming into and out must be adaptable to a number of domain-centric data types
  - **Complex queries:** combining several data analysis functions for a given query is a common requirement among scientific applications

FY09 Joule Application: VisIt

Isosurface benchmark problem

Denovo: Study the radiation dose concentrations around a reactor core in a nuclear power generating plant

- Computing dose requires combining several data analysis functions for a given query is a common requirement among scientific applications
- Pushes the scalability of VisIt because of the enormous Denovo memory requirements – also a common feature of many domains.
- Trillions of DoF for a code such as Denovo is becoming more frequent (driven by energy groups and directional angles)

steady state Boltzmann transport calculation

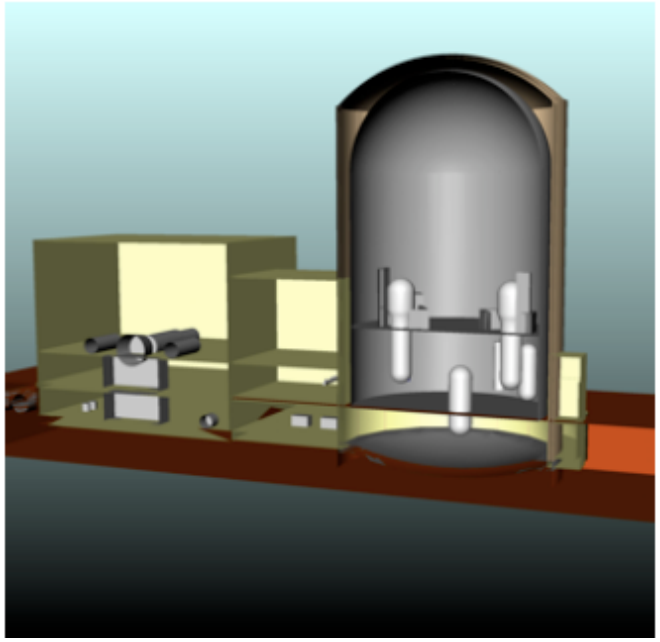
4096 spatial domains

materials: concrete, reactor fuel, steel, reduced density steel, air

mesh size : 456 x 648 x 351

scalar flux values for 27 energy groups per zone

4096 files (1 per domain) totaling 36.23843 GB
-double precision values



FY08 Joule Metric Applications

application software	application metric	target platform	problem	joule result
(08)DCA++	time/disorder configuration	Cray XT5 31272pes, 23791s	256 disorder configs, 150nts in 2D Hubbard Model -impact on Tc	weak(k=4)
(08)GYRO	timesteps /second / process	Cray XT5 24576pes, 152.75s	improve the electron to ion mass ratio (mu=40, 20ts) in magnetically confined tokamak plasma	weak(k>5)
(08)PFLOTRAN	time/dof/PE	Cray XT5 8000pes, 2958.36s	764 x 1414 x 120 grid resolution of reactive transport in Hanford 300	weak(k=2)

FY07 Joule Metric Applications

application software	application metric	target platform	problem	joule result
(07)CHIMERA	Compute time / subcycled hydrodynamic step	Cray XT3 2048 pes, 415 cpu-hours	Post-bounce evolution of 11 solar mass star, s11.2, 100 t steps	weak(k=8)
(07)GTC-S	Number of particles / (compute time / physical time step)	Cray XT4 64, 4096 pes, 575 cpu-hours	PIC microturbulence plasma study in DIIID tokamak exp shot 122338 at 1.6s T, rho profile	weak(k=64)
(07)S3D INCITE	Compute time / dof / physical time step / processor core	Cray XT3 U XT4 14112 pes, 41800 cpu-hours	Premixed methane-air, slot bunsen; non-premixed ethylene-air, planar slot jet	weak(k=1.96)

FY06 Joule Metric Applications

application software	application metric	target platform	problem	joule result
(06)DCA/QMC	Compute time / Green function update/ time slice	Cray X1E 512 pes 6352 cpu-hours	Pairing interaction study of 2d Hubbard Model	performance 74.03%
(06)ENZO	Compute time / processor core / physical time step	IBM Power5 512 pes 6725 cpu-hours	•AMR (4lev) study •High red shift galaxy formation •512^3 grid •512^3 dark matter particles	performance 66.5%
(06)MADNESS	time for projection, compression, reconstruction, multiplication, differentiation	Cray XT3 4096 pes 7430 cpu- hours	Project nuclear potential from 4096 Cu atoms bcc lattice into wavelet basis w/ 1.e-3 precision	performance 77.27%
(06)ScalaBLAST	Compute time / query / processor core	HP Itanium-2 (LP) 1500 pes 45357 cpu-hours	Whole genome sequencing of Sargasso Sea environmental samples vs nr protein data base	new result, sequenced 1.2 million previously unknown proteins

FY05 Joule Metric Applications

application software	application metric	target platform	problem	joule result
(05)AORSA	compute time of FFT, ax=b	Cray X1E 256 pes 533 cpu-hours	Absorption of rf power by non-Maxwellian bulk ion components in NSTX tokamak	performance 55.75%
(05)CCSM	Simulated years / wall clock day	Cray X1E ~ 11904 cpu-hours	CAM3; spectral Eulerian dynamical core study (semi-Lagrangian vs Finite Volume)	performance 53.7%
(05)LAMMPS	Dominated by force computation ala classical pairwise interactions	IBM BG/L	Md simulation of metal island on metal or oxide substrate to study effects of stress on device performance	new result,improved potentials (Yukawa, Morse, Buckingham); resolved dependence of stress in island in island size and adhesion to substrate

FY05 Joule Metric Applications

application software	application metric	target platform	problem	joule result
(05)Omega3P	Compute time / eigenmode / processor core	IBM SP Power3 768 pes , 1753 cpu-hours	HOM study of 9-cell superconducting accelerating cavity in the ILC Tesla Test Facility	performance 81.3%
(05)S3D INCITE	Compute time / grid point / physical time step	Cray X1 256 pe	Non-premixed CO/ H2/N2-air plane jet flame simulation	performance 57.74%
(05)S3D SciDAC	Compute time / grid point / physical time step	HP Itanium-2 (LP) 256 pes , metric only	Fuel spray injection study of effects of droplet size on evolution of carrier gas field features	performance 75%

FY04 Joule Metric Applications

application software	application metric	target platform	problem	joule result
(04)CCSM	Simulated years / wall clock day	IBM p690	T42(2.8d) T85	(Q2) 5 sim yrs / wall clock day, (Q4) > 38 sim yrs / wcd
(04)MILC	Compute time / sparse linear system; compute time / SU(3) matrix vector product	QCDOC	Single mass CG inverter on 128 QCDOC nodes	performance 90%
(04)NSM MC	Compute time / nucleon / shell / sample / imag time step	IBM SP Power3	Mo92, gds / 65536 samples	298MFlops, 74hours, 2048 pes
(04)RMPS	Compute time / Ax=kx solve / pe	IBM SP Power3	Electron impact excitation in DIIID tokamak energy and particle confinement study	Larger inversions , heavier atomic systems (235 level,Ne)
(04)VH-1	Compute time / zone update / processor core	Cray X1	3D SASI, l=1 mode	1,140,000 zone updates / second / pe

Joule Metric Exercise
Benchmark trends FY04-FY10

climate research	4
condensed matter	2
fusion	5
high energy physics	2
nuclear	2
subsurface modeling	1
astrophysics	2
combustion chemistry	4
bioinformatics	1
math, data analytics	2
molecular dynamics, electronic structure	2
nuclear energy	1
Total	28

Cray	X1
	X1E
	XT3
	XT4
4-core	XT5
6-core	XT5
IBM	SP Power3
	P690
	Power5
	BG/L
SGI	Altix
HP Itanium-2	
QCDOC	

U.S. DEPARTMENT OF ENERGY Office of Science
OAK RIDGE National Laboratory
Pacific Northwest NATIONAL LABORATORY

Joule Metric Exercise

Computational costs & benchmark productivity: FY04-FY10

Fiscal Year*	Benchmark CPU-Hours
2005	24,814
2006	211,888
2007	314,459
2008	2,718,788
2009	39,300,189
2010	7,712,255 (Q2) (est > 50M by Q4)

Benchmark Productivity Examples:

- FY08: 2,718,788 CPU-Hours benchmark related work productivity := 271,879 / 2,718,788 ~ **.10%**

- FY09: 39,300,189 CPU-Hours benchmark related work productivity := 3,284,525 / 39,300,189 ~ .083575 or **less than 10%**

- FY10 - Q2 : 7,712,255 CPU-Hours Q2 benchmark related work productivity Q2 := 863,286 / 7,712,255 ~ .111949 or **about 11%**

*FY04 numbers are available but unreliable

Extended Scope of Computational Science Studies

Example Problem: solving algebraically determined systems of linear equations numerically

M L_p ▸ • construct A, b given dim(A) dim(A)=dim(rank(1,A))=dim(rank(2,A))

L_p ▸ M L_R ▸ • compute A=LU, solve Ly=b for y, solve Ux=y for x

L_R ▸ M

accept ▸ • compute $|A|_{inf}$, compute $|x|_2$, compute $|b-Ax|_2$,
approximate ε (machine precision),

reject ▸ form $\kappa = |b-Ax|_2 / (|A|_{inf} * |x|_2 * \varepsilon)$;
if $\kappa \sim$ O(dim(A)) **accept** x, else **reject** x

Q: How do the language of the problem and the accepted result relate to reality?

Requires analysis beyond software analysis above and distinguishes computational science from system and library software development. Takes more time -needs refinement phase of algorithms and metrics.

Trends in Target Computing Platforms

Hex-Core AMD Opteron (TM)	2.6e9 Hz clock	4 FP_OPs / cycle / core 128 bit registers
PEs	18,688 nodes	224,256 cpu-cores (processors)
Memory	16 GB / node 6 MB shared L3 / chip 512 KB L2 / core 64 KB D,I L1 / core	dual socket nodes 800 MHz DDR2 DIMM 25.6 GBps / node memory bw
Network	AMD HT SeaStar2+	3D torus topology 6 switch ports / SeaStar2+ chip 9.6 GBps interconnect bw / port 3.2GBps injection bw
Operating Systems	Cray Linux Environment (CLE) (xt-os2.2.41A)	SuSE Linux on service / io nodes

FY	Aggregrated Cycles	Aggregated Memory	Aggregated FLOPs	Memory/FLOPs
2008	65.7888 THz	61.1875 TB	263.155 TF	0.2556
2009	343.8592 THz	321.057 TB	1.375 PF	0.2567
2010	583.0656 THz	321.057 TB	2.332 PF	0.1513

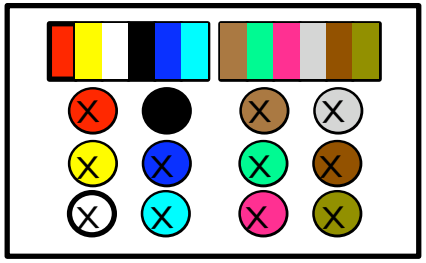
NUMA Node Structure of XT5 --> Hybrid Programming Model

- **MPI processes** spawn lightweight processes
- **OpenMP threads**, #include <omp.h> , omp_set_num_threads();
- **POSIX threads**, #include <pthread.h> , pthread_create();

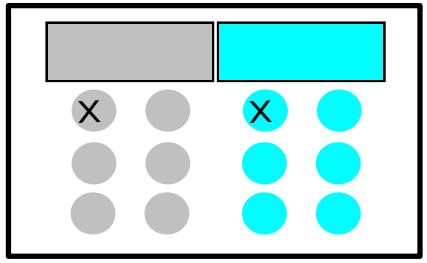
-lsize=12

	MPI	LWP	DRAM
aprun -n <1-12>	1 - 12	1	1.33 * 2^30
aprun -n 2 -sn 2 -S 1 -d 6	2	1 - 6	8 * 2^30
aprun -n 1 -N 1 -d 12	1	1 - 12	16 * 2^30

<-S> * <-d> cannot exceed the maximum number of CPUs per NUMA node



no NUMA, 6 PEs/socket



balanced NUMA, 1 PE / socket

NUMA + memory affinity

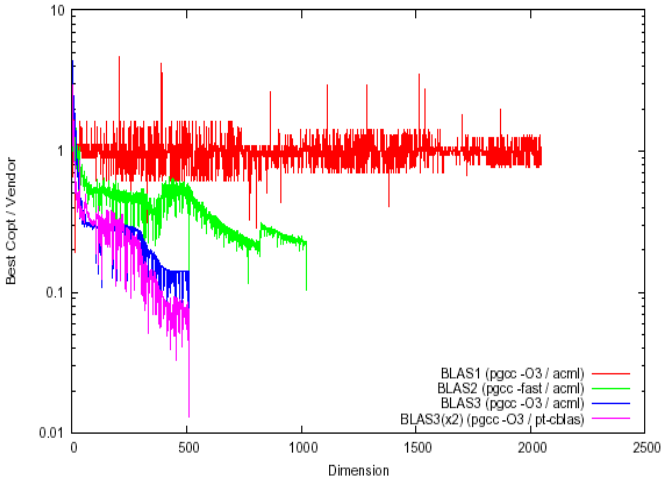
⊗ := 1 MPI process

Memory Wall Remains in Multi-Core

Computation: Theoretical peak: (# cpu cores) * (flops / cycle / core) * (cycles / second)

Memory: Theoretical peak: (bus width) * (bus speed)

BLAS 1: O(n) operations on O(n) operands
BLAS 2: O(n**2) operations on O(n**2) operands
BLAS 3: O(n**3) operations on O(n**2) operands



$y = \alpha x + y$:

3 loads, 1 store (more expensive than FP_OPs by a long shot)
2 floating point operations (maybe 1) on 3 operands

eg, double precision on the FY10 target platform:
(3 operands / 2 flop) * (8 bytes / operand) * 6 core * 4 (flop / cyc /core) * 2.6e9(cyc/sec)
~125 GBps

... We don't have this and to get it is $$$ ... how to achieve **Sustainability??**

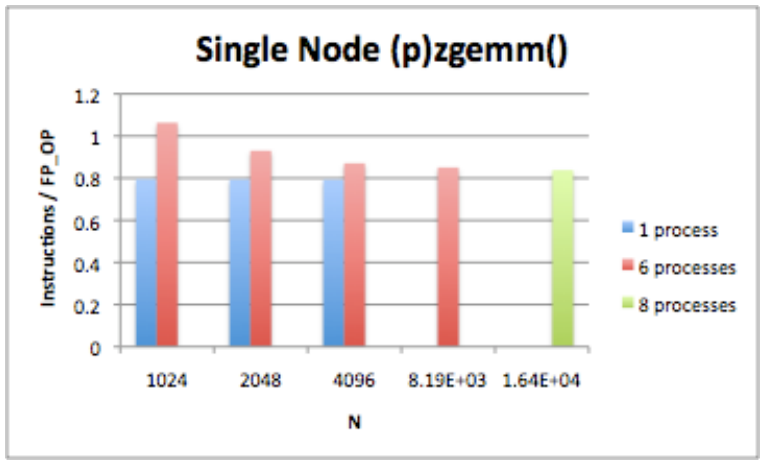
Floating Point Intensity of Joule Benchmark Applications

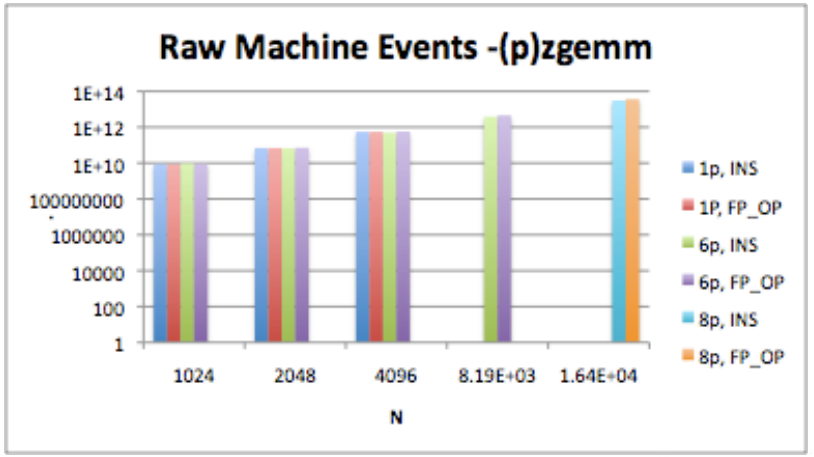
Application	1	2	3	4	5	6	7
Instructions Retired	1.99E+15	8.69E+17	1.86E+19	2.45E+18	1.24E+16	7.26E+16	8.29E+18
Floating Point Ops	3.52E+11	1.27E+15	1.95E+18	2.28E+18	6.16E+15	4.15E+15	3.27E+17
INS / FP_OP	5.64E+03	6.84E+02	9.56	1.08	2.02	17.5	25.3

REFERENCE FLOATING POINT INTENSE PROBLEM :: Dense Matrix Matrix Multiplication

C <--- a A B + b C :: OPERATIONAL COMPLEXITY : A[m,n] , B[n,p] , C[m,p] :: [8mpn + 13mp] FLOP

E.g. m=n=p=1024 ---> 8603566080 FLOP , measure 8639217664





U.S. DEPARTMENT OF ENERGY Office of Science · OAK RIDGE National Laboratory · Pacific Northwest NATIONAL LABORATORY

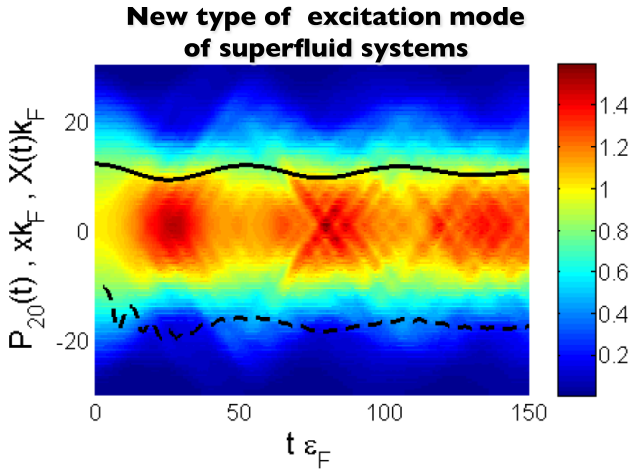
FY10 Joule Application: TD-SLDA

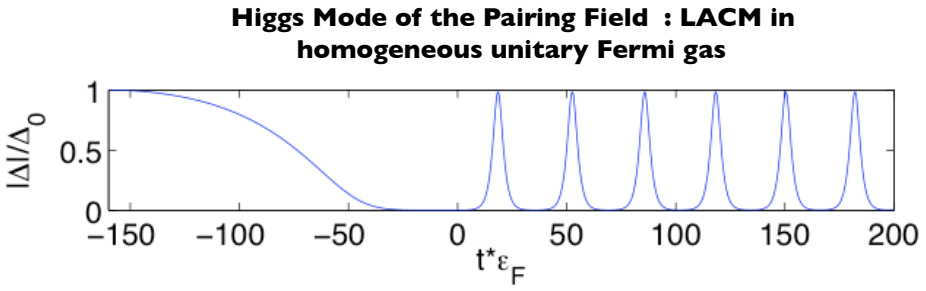
Time-Dependent Superfluid Local Density Approximation extension of Kohn-Sham LDA to time-dependent superfluid / superconducting system

$$i\hbar\partial_t \begin{pmatrix} u_n(\vec{x},t) \\ v_n(\vec{x},t) \end{pmatrix} = \begin{pmatrix} \hat{h}(\vec{x},t) + \hat{V}_{ext}(\vec{x},t) & \hat{\Delta}(\vec{x},t) + \hat{\Delta}_{ext}(\vec{x},t) \\ \hat{\Delta}^\dagger(\vec{x},t) + \hat{\Delta}^\dagger_{ext}(\vec{x},t) & -\hat{h}(\vec{x},t) - \hat{V}_{ext}(\vec{x},t) \end{pmatrix} \begin{pmatrix} u_n(\vec{x},t) \\ v_n(\vec{x},t) \end{pmatrix}$$

- **Model**
  - variations on Hartree-Fock-Bogoliubov
- **Algorithm & implementation**
  - C and Fortran versions w/ MPI + Pthreads
  - Cartesian lattice-based codes
  - Plane wave basis, discrete Fourier transforms (FFTW) for gradients, derivatives
  - Densities are summations of bilinear products of the quasiparticle wave functions
  - Multi-step (predictor-modifier-corrector) time algorithm , O(h^5)
- **Algorithm & S/W performance challenges**
  - superfluid theory is by definition non-local (unlike normal systems -those with no pairing) imposes global update of several observables
  - expensive memory representation -can Pthreads help?

$$Q(\omega) = \sum_\sigma \int Q(\vec{x},\sigma,t)\rho(\vec{x},\sigma,t)e^{i\omega t}d^3xdt$$





FY10 Joule Application: Denovo

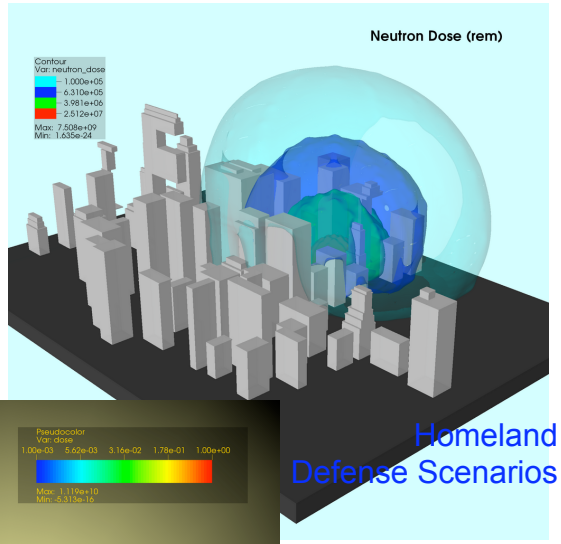
Tom Evans (ORNL)

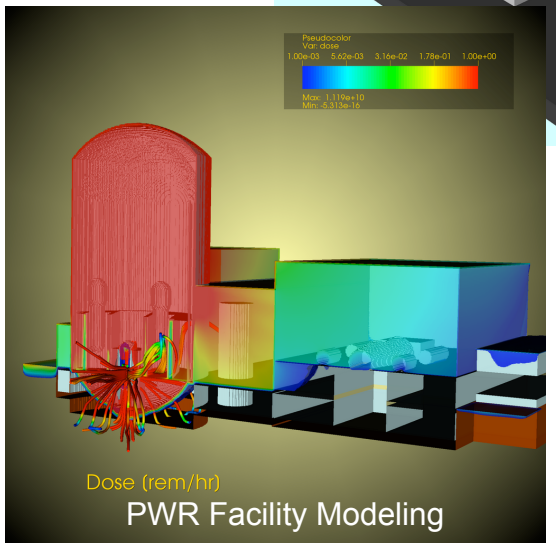
3D Deterministic Radiation Transport for Nuclear Energy, Shielding, Medical Physics, Homeland Defense, and Nuclear Criticality/Safeguards

$$\frac{1}{v}\frac{\partial \psi}{\partial t} + \hat{\Omega}\cdot\nabla\psi + \sigma\psi = \iint \sigma_s(\hat{\Omega}' \rightarrow \hat{\Omega}, E' \rightarrow E)\psi(\mathbf{\Omega}', E')\, d\mathbf{\Omega}' dE' + q(\mathbf{r}, \hat{\Omega}, E, t)$$

$$\psi \equiv \psi(\underbrace{\mathbf{r}}_{3}, \underbrace{\mathbf{\Omega}}_{2}, E, t)$$

- **Model**
  - Forward or adjoint linear Boltzmann transport
- **Algorithm & implementation**
  - Multigroup discrete ordinates (S_N) method with finite element collocation in angle
  - Outer Krylov-based iterative method for within-group scattering
  - OO with C++, C, and Python
- **Algorithm & S/W performance challenges**
  - Managing multiple hierarchical concurrencies: multi-threading angles across cores, decomposing space across nodes, decomposing energy across replicated mesh blocks
  - Hybrid coupling with domain-decomposed Monte Carlo





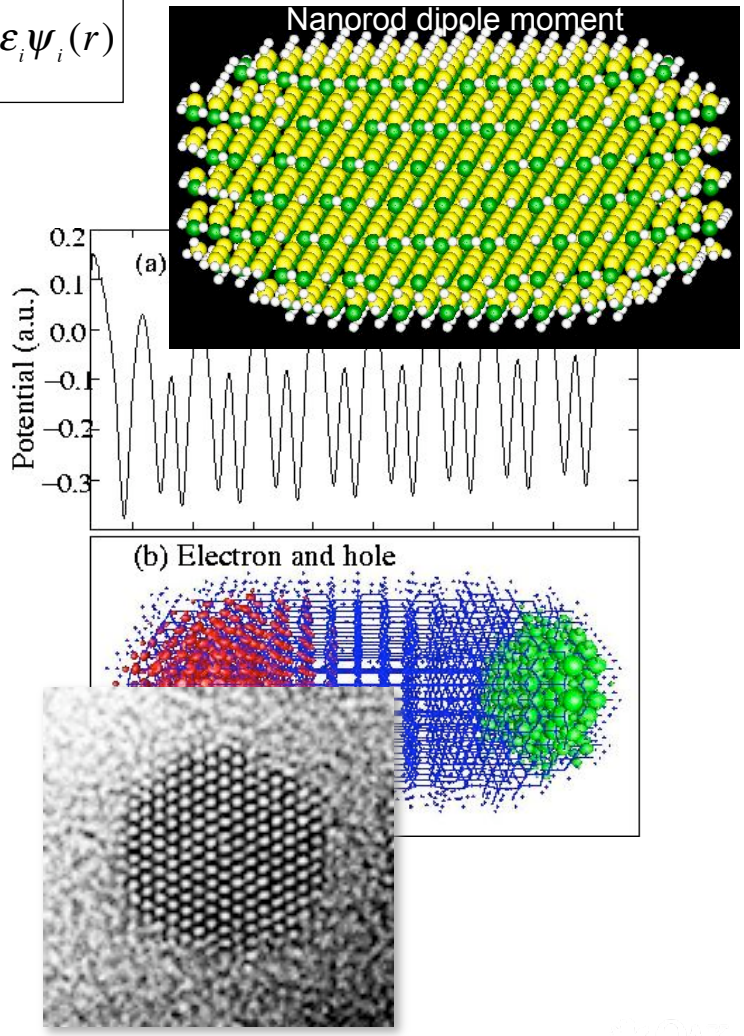
FY10 Joule Application: LS3DF

Lin-Wang Wang (LBNL)

Linearly Scaling 3D Fragment Method for Large Scale Electronic Nanostructure Studies of Solar Cells, Semiconductor Alloys, and Electronic Devices

$$\left[-\frac{1}{2}\nabla^2 + V_{\text{tot}}(r)\right]\psi_i(r) = \varepsilon_i\psi_i(r)$$

- **Model**
  - All-band, plane wave pseudo-potential DFT method for ab initio electronic structure
- **Algorithm & implementation**
  - Divide-and-conquer method that partitions physical space into many overlapping fragments modeled independently on separate processor groups
  - Fragment charge densities patched together with an self-consistent variational Poisson formulation
  - O(N) method able to find solutions for systems with hundreds of thousands of atoms
- **Performance challenges**
  - Multi-level parallelization of the Schrodinger solve for each fragment's wavefunction
  - Specialized matrix-matrix operations



FY10 Joule Application: POP

Phil Jones (LANL)

CCSM Ocean General Circulation Model for National and IPCC Climate Change Assessment and Prediction Studies

- **Model**
  - Circa 1991 ocean general circulation model of Bryan, Cox, Semtner, & Chervin with 3D primitive equations for fluid motion on a sphere under hydrostatic and Boussinesq approximations
  - State of the art physical parameterizations, EOS, & biogeochemistry (carbon/sulfur cycle, carbon sequestration)
- **Algorithm & implementation**
  - Finite difference hydrodynamics with depth as vertical coordinate
  - General horizontal grids (displaced pole, tripole)
  - Fast vertically-uniform barotropic modes integrated implicitly with free surface formulation
  - 3D vertically-varying baroclinic modes integrated explicitly
- **Performance challenges**
  - Unexploited concurrencies, communication latency-dominated strong scaling, scalable barotropic solution techniques