



Argonne
NATIONAL
LABORATORY

... for a brighter future



U.S. Department
of Energy

UChicago ►
Argonne_{LLC}



A U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC

Early Applications Scaling on the ALCF BG/P

Raymond Loy

*Applications Performance Engineering and Data
Analytics (APEDA)*

Argonne Leadership Computing Facility

And

ALCF APEDA

- Kalyan Kumaran
- Vitali Morozov
- James Osborn
- Katherine Riley

- Paul Fischer (MCS)
- Dinesh Kaushik (MCS)

With help from

- Sameer Kumar (IBM)
- Carlos Sosa (IBM)

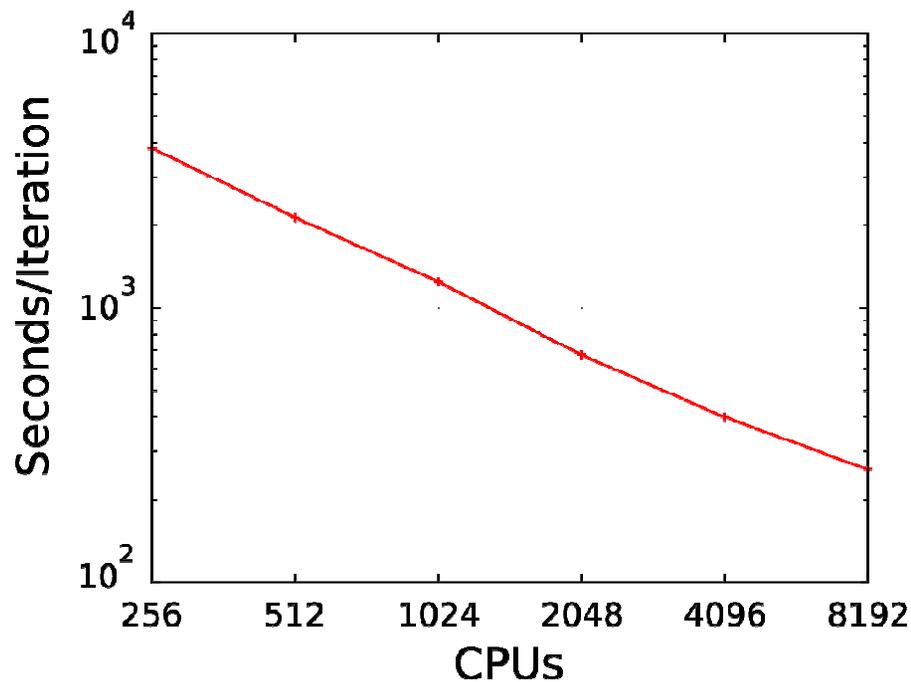
Early Science Applications for ALCF

- We worked with DOE Office of Advanced Scientific Computing Research to Pre-select applications for Early Science runs on the BG/P
- These codes and science problems provide a broad set of problems spanning scientific areas, algorithms and requirements to help insure the ALCF is broadly capable of providing “science useful cycles” from day one
- All of these codes run on BG/L at scale
- Codes, algorithms and science problems
 - QBOX - density functional theory, nanoscience property design
 - MILC - lattice gauge theory,
 - FLASH - CFD and multi-physics, Type 1a Supernova offset bubble
 - NAMD - classical MD, Ion channel conformational change
 - NEK5000 - CFD, reactor core design

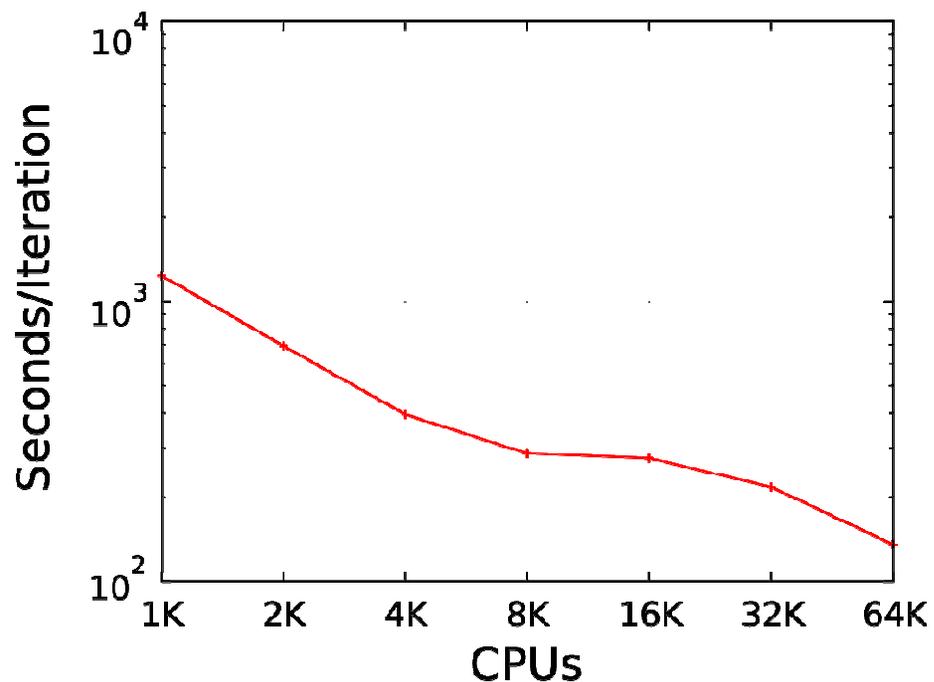
QBOX

- First-principles molecular dynamics
- C++/MPI implementation for massively parallel computers.
- Used for studying liquids, semiconductor nanostructures and materials science under extreme conditions.
- Ported to BlueGene/L. Achieved a performance of 207 Tflops and won the Gordon Bell award in 2006.

QBOX Scaling



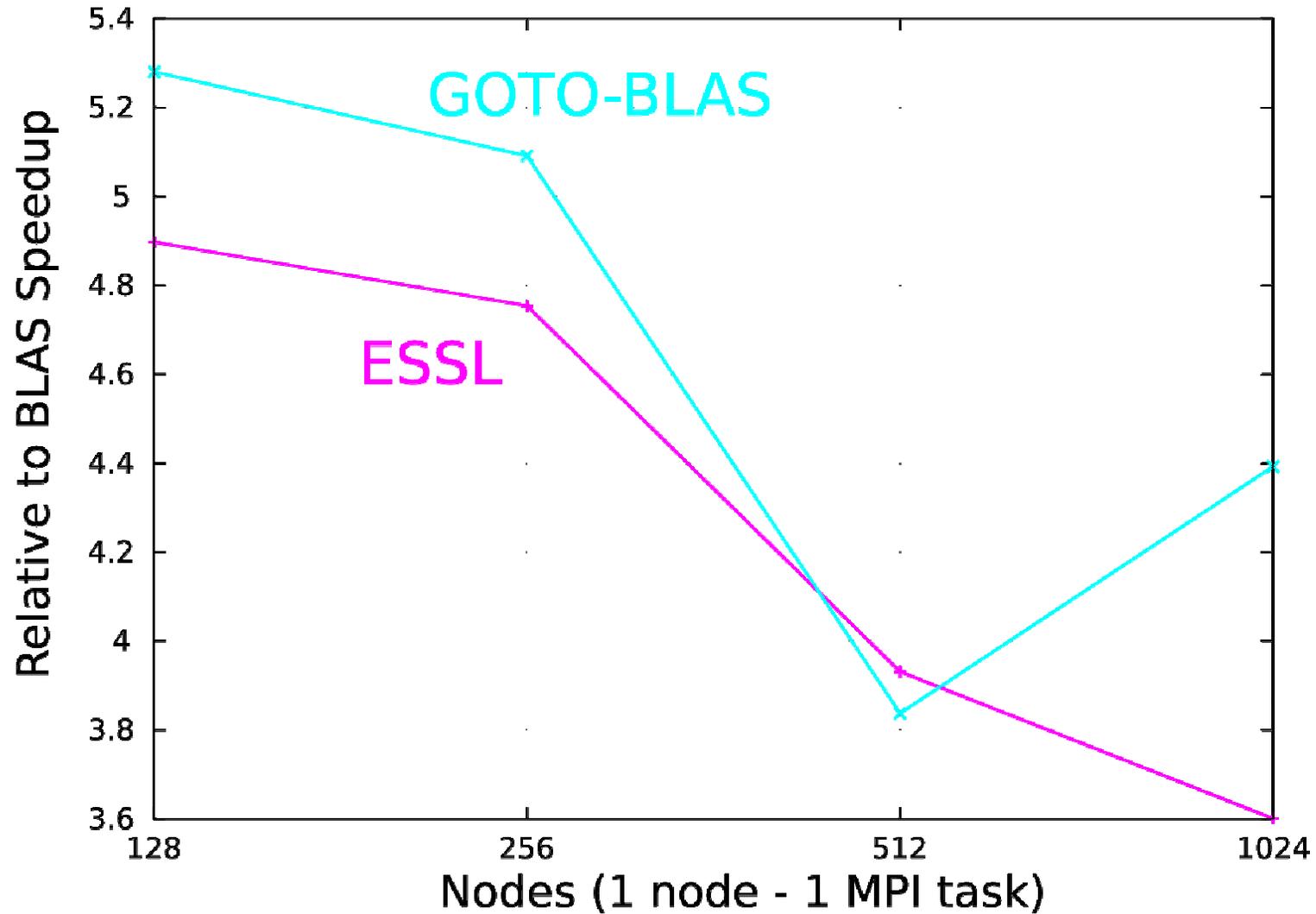
Test 1
Basis size: 180K
Ecut: 27Ry
#SCF steps: 30



Test 2
Basis size: 1.14M
Ecut: 90Ry
#SCF steps: 5

SiO₂ cluster of 1536 atoms

Dense Linear Algebra - Key to QBOX Performance



QBOX Conclusions

- QBOX scales to 32 racks (dual mode) on BG/P
- The main loop of QBOX is dense linear algebra, and therefore, significant gain can be reached from using optimized BLAS/LAPACK routines, provided by Goto-BLAS and/or ESSL
- Mapping QBOX MPI-tasks is very important on Blue Gene/P
 - 512 vn XYZT 768 s ; TXYZ 1140 s
 - 1K vn XYZT 578 s ; TXYZ 401 s
- IO scheme can be a limiting factor
 - Improved MPI-IO based IO-scheme under development

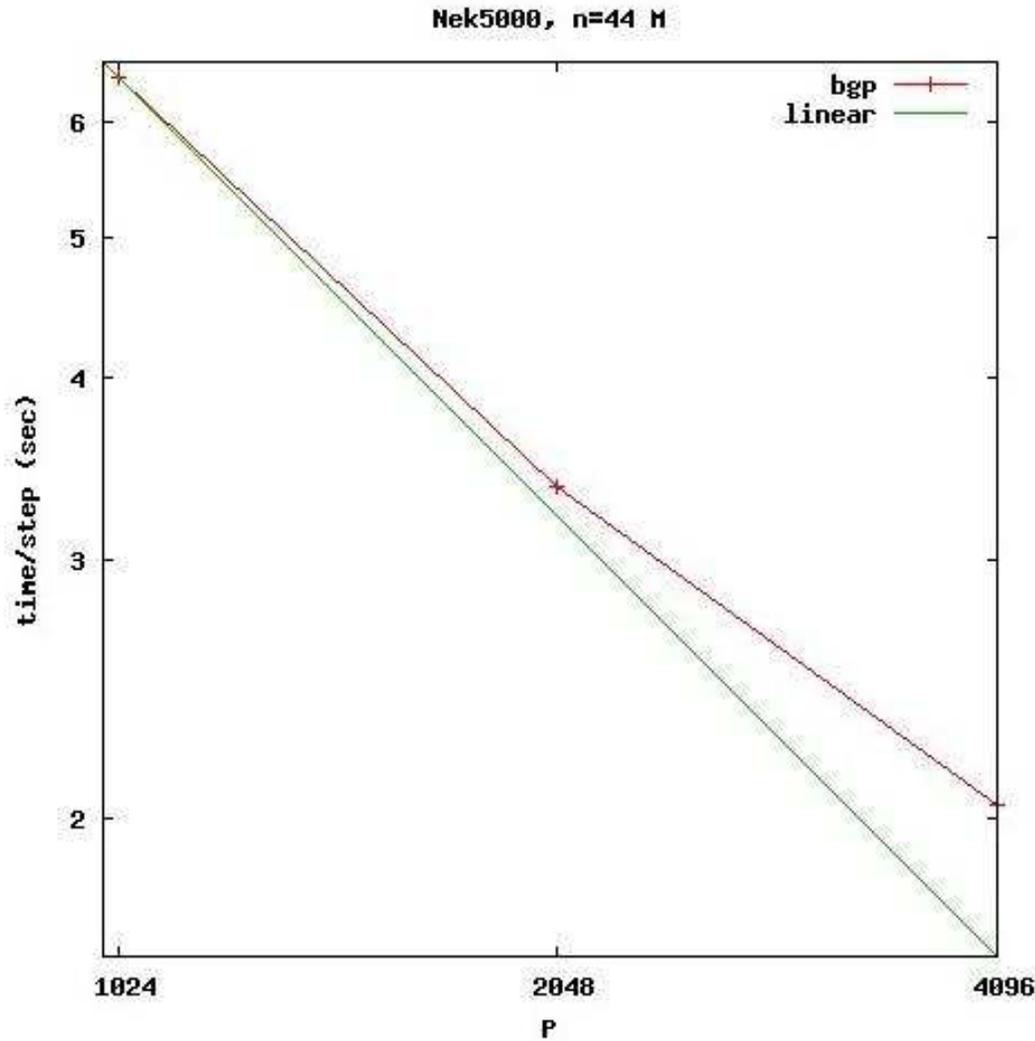
NEK5000

- Spectral element code for the simulation of unsteady incompressible fluid flow, heat transfer, and MHD in general three-dimensional domains.
- Code is written in Fortran & C and uses MPI.
- Scales to 32,000 processors on BG/L
- Won Gordon Bell Prize in 1999.

Performance of NEK5000 on BG/P

- Three benchmark cases
 - Small: 8100 elements
 - Medium: 25,920 elements
 - Large: 132,192 elements
- BG/P vs BG/L at Watson (256 cores)
 - Factor of 3.10 per node using 2 MPI processes per node
- BG/P (write through) vs. BG/L (Write Back)
 - Clock ratio of 1.21
 - Medium: factor of 1.13 on 2K cores, and 1.42 on 8K cores
 - Large: factor of 1.65 on 8K cores
- Have run BG/P acceptance test on 131K processors (32 racks)

NEK5000 Scaling



Reactor Cooling Benchmark

Strong Scaling:

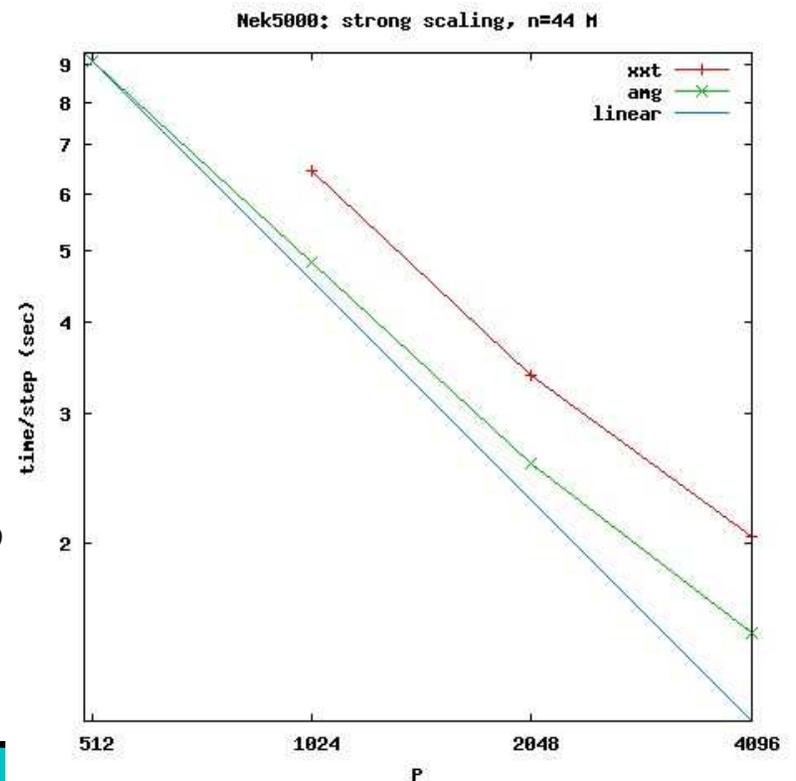
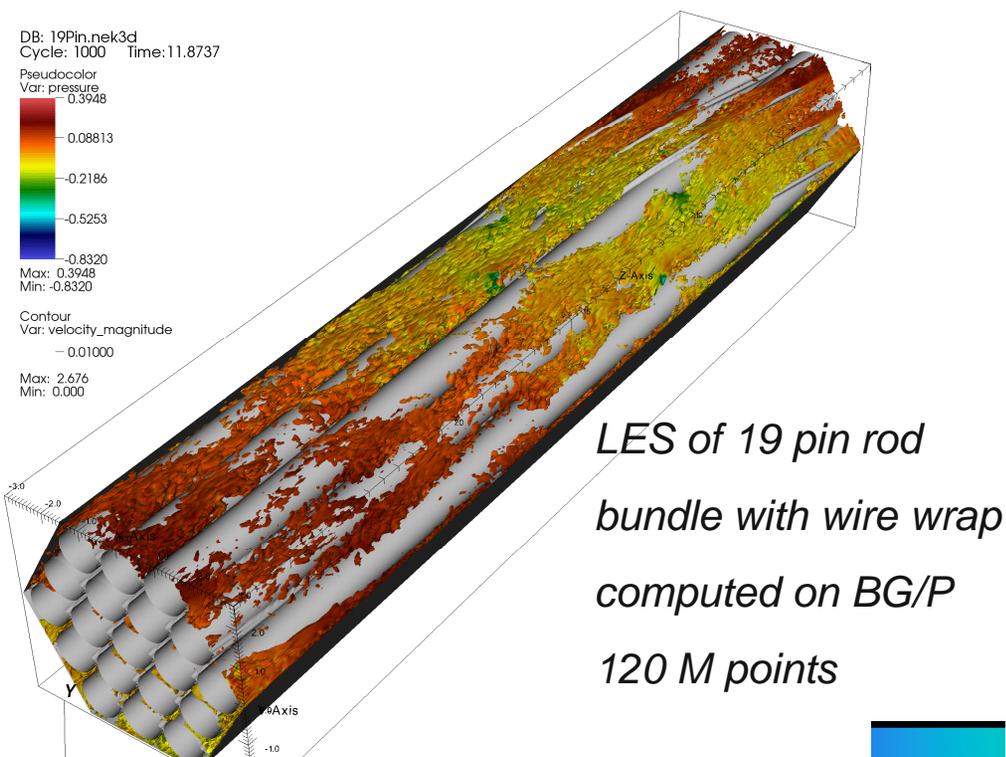
@ P=4096:

- ~10,000 points/proc
- Parallel efficiency: $\eta \sim 0.79$

LES Analysis of coolant in fast reactor subassemblies: 19 pin case

2008 INCITE award to analyze 19- and 37-pin configurations.

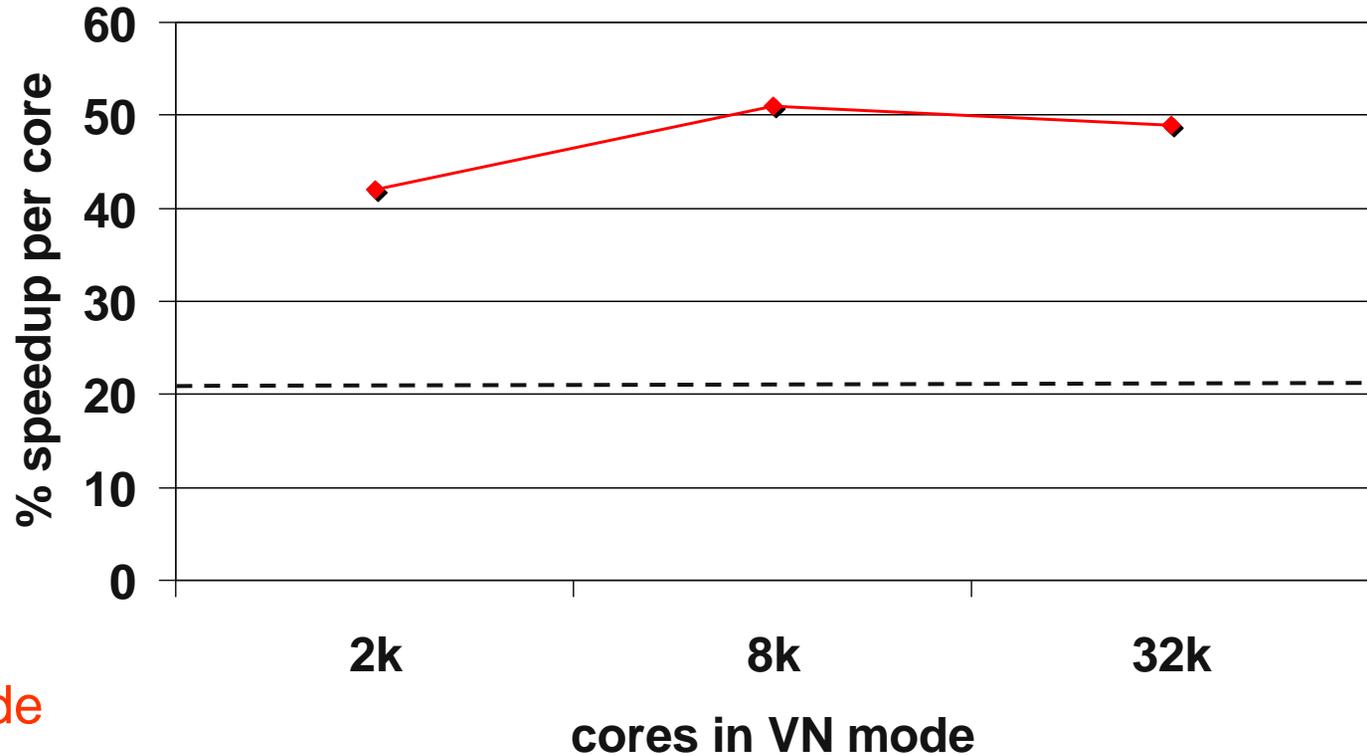
- Analysis of wall effect is important.
- Have simulated full transition for 19-pin configuration using 4096 processors of BG/P:
~800,000 CPU hours, 120 M points (360,000 elements of order $N=7$)
- New AMG-based coarse grid solver provides ~30% savings; tuning remains...



MILC

- QCD simulations
 - generate QCD field configurations through Monte Carlo process
 - requires solving Dirac equation – large sparse linear system
 - mostly nearest-neighbor communications on 4D torus
- MILC collaboration code
 - developed over many years; contains many applications
 - mostly in C with some assembly using MPI for communications
 - has been ported and run on most major computing architectures
- porting to BG/L
 - part of SciDAC project
 - improved domain specific linear algebra library to take advantage of BG/L double hummer FPU
 - wrote native communications library bypassing MPI to reduce latency

MILC Comparative Analysis of BG/P vs. BG/L



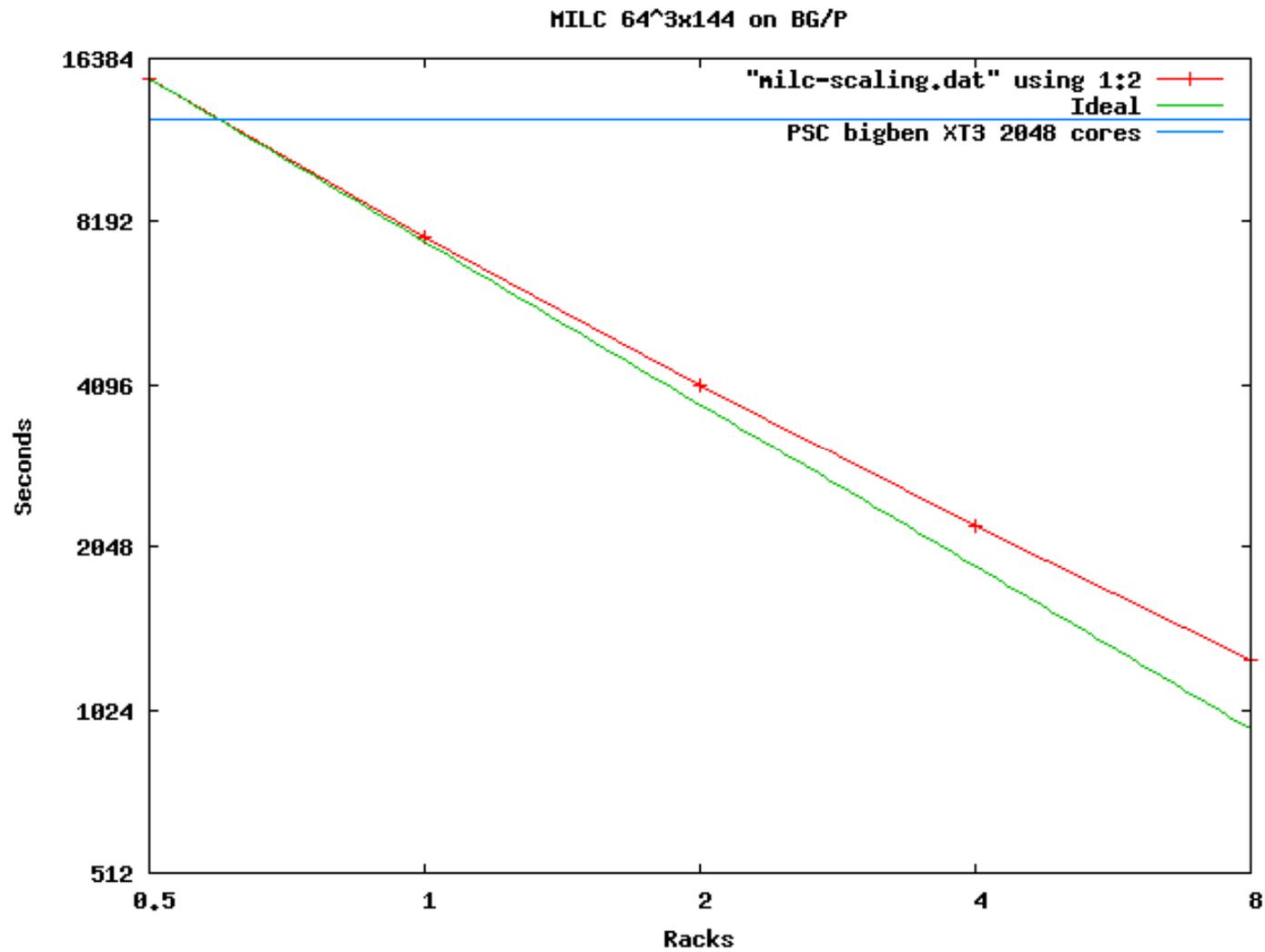
Virtual Node mode

2 MPI threads per node in BG/L
4 MPI threads per node in BG/P

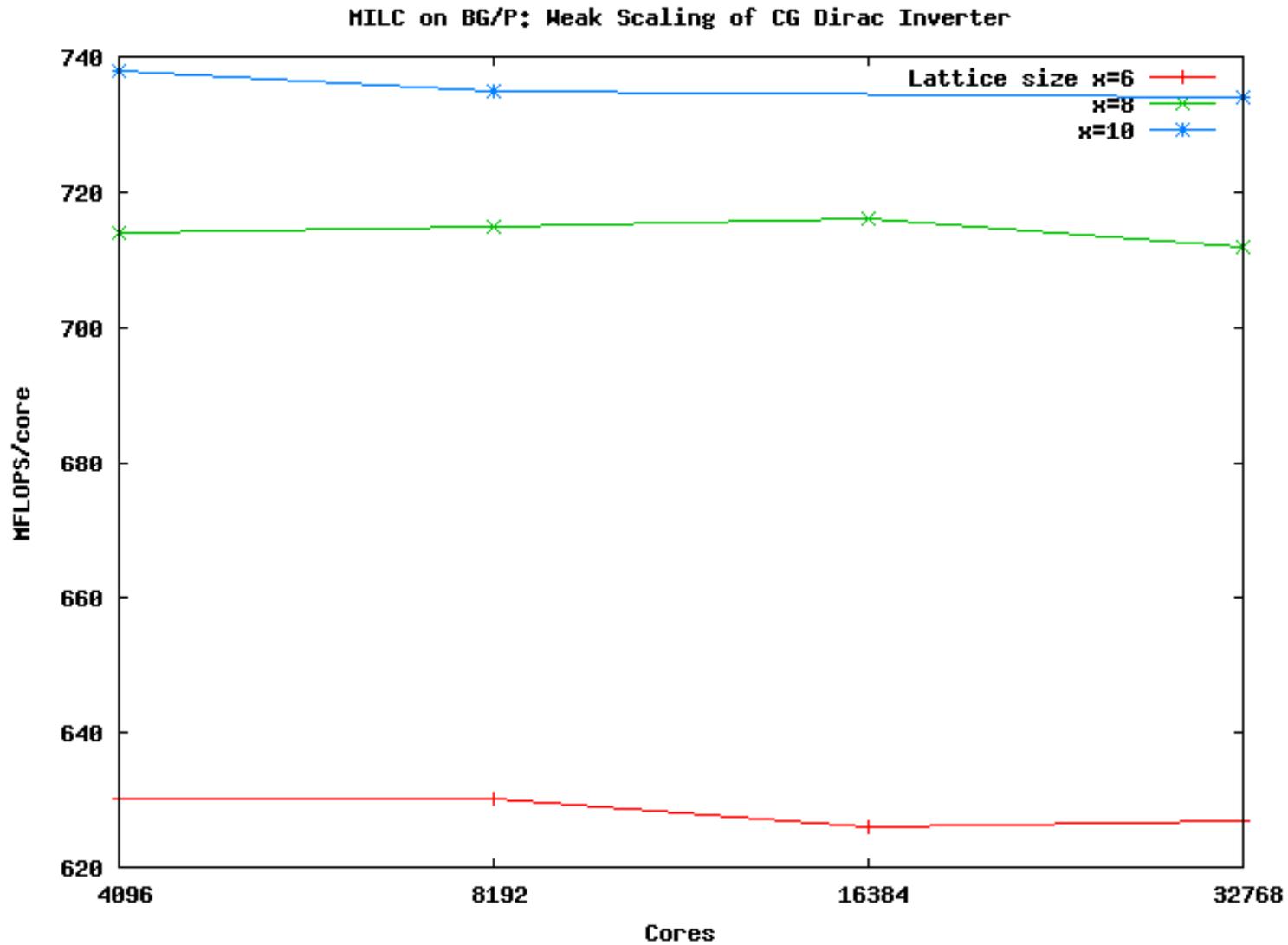
Comparison for equal number of cores
(2x as many BG/L nodes vs. BG/P)
21% due to frequency increase

	lattice sizes
2k cores:	64 x 64 x 64 x 32
8k cores:	64 x 64 x 64 x 128
32k cores:	128 x 128 x 128 x 64

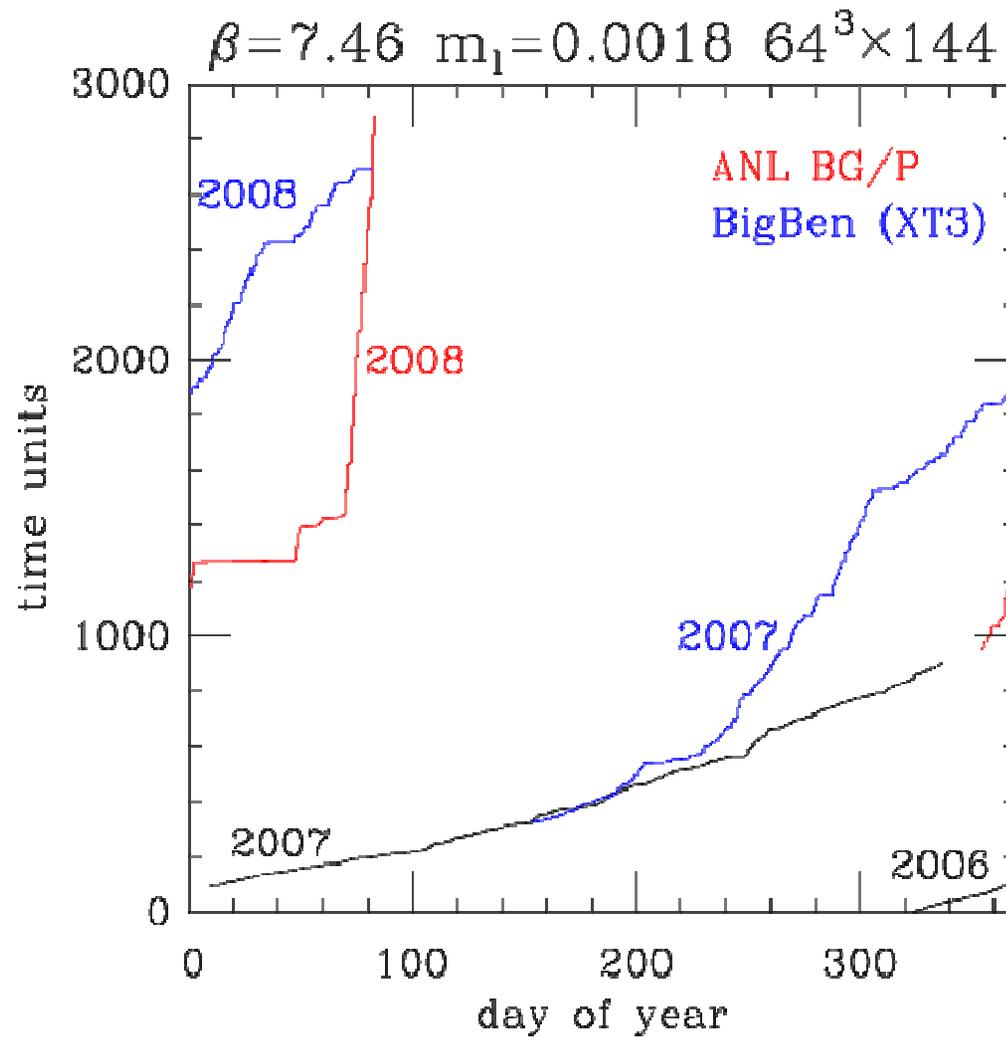
MILC: Science run on BG/P – strong scaling



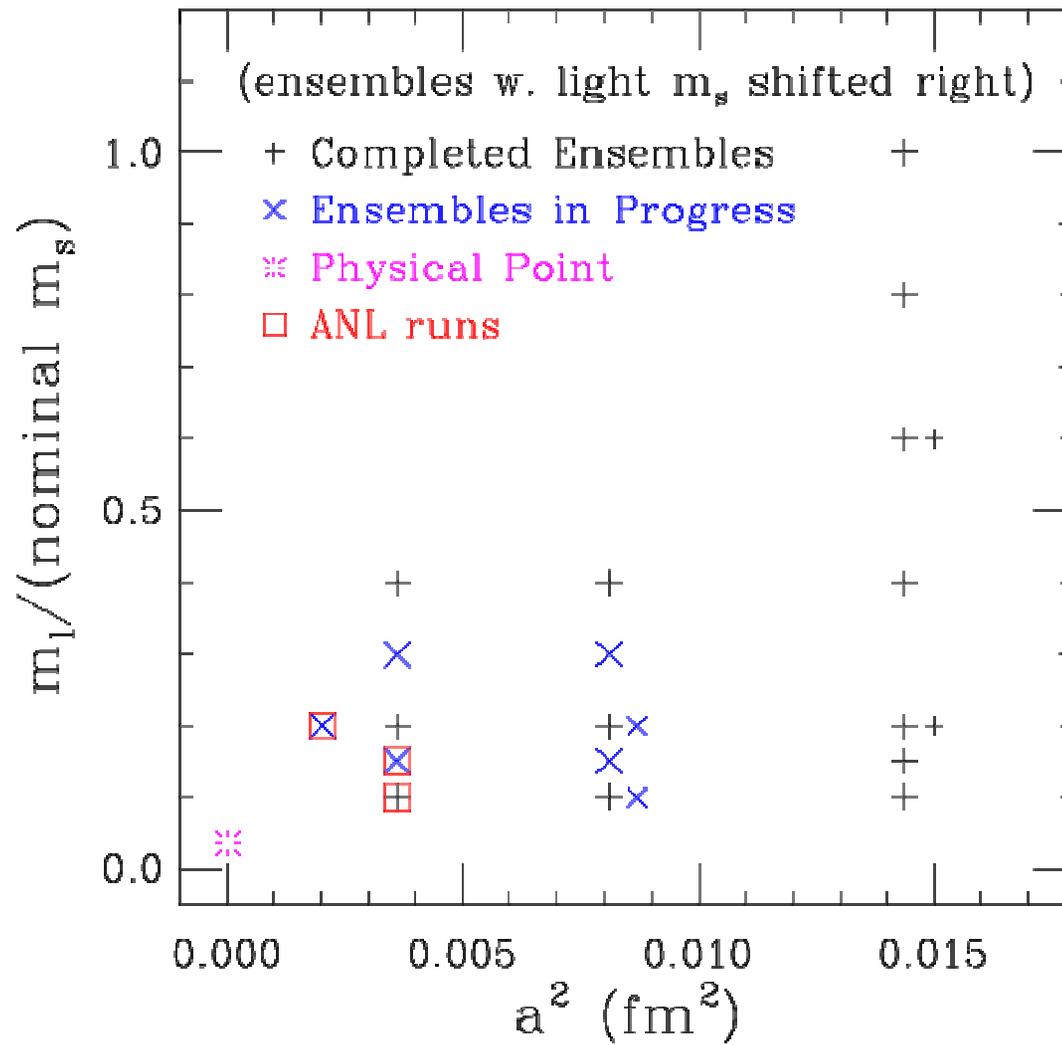
MILC on BG/P



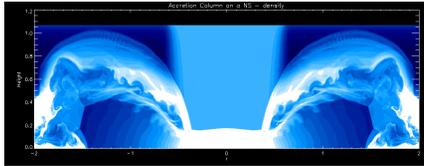
MILC: Lattice generation history



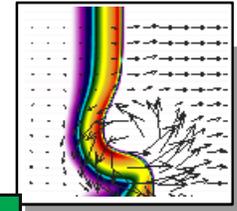
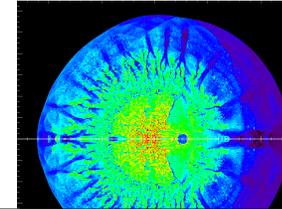
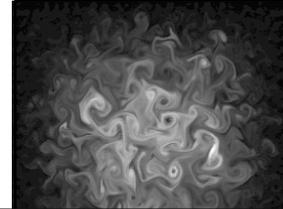
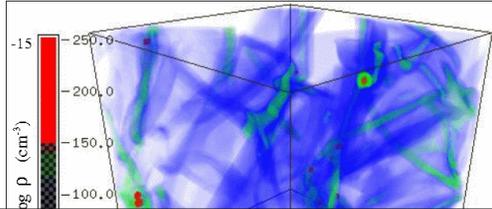
MILC on BG/P



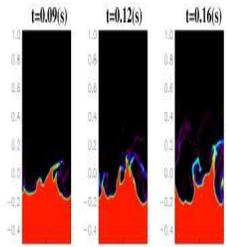
FLASH



Shortly: Relativistic accretion onto NS

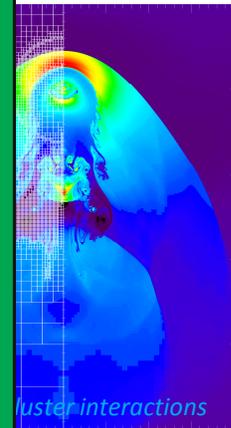


vortex interactions

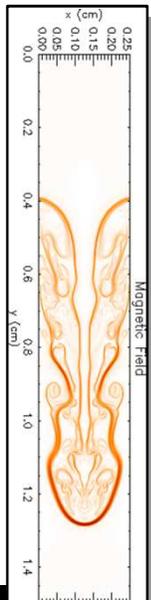


Wave breaking on w

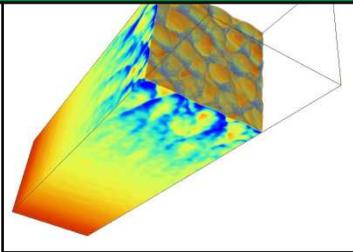
- The FLASH code
1. Parallel, adaptive-mesh refinement (AMR) code
 2. Block structured AMR; a block is the unit of computation
 3. Designed for compressible reactive flows
 4. Can solve a broad range of (astro)physical problems
 5. Portable: runs on many massively-parallel systems
 6. Scales and performs well
 7. Fully modular and extensible: components can be combined to create many different applications



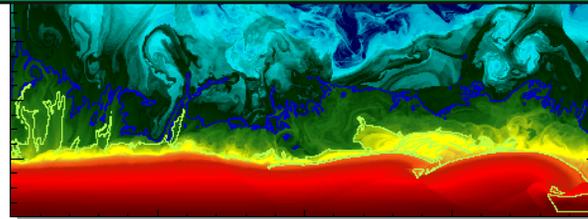
cluster interactions



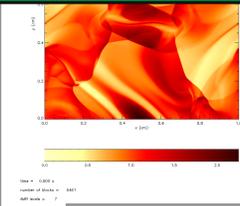
M



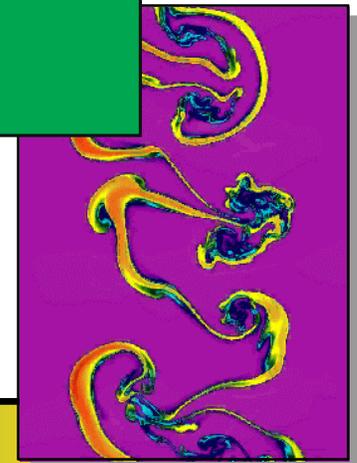
Cellular detonation



Helium burning on neutron stars



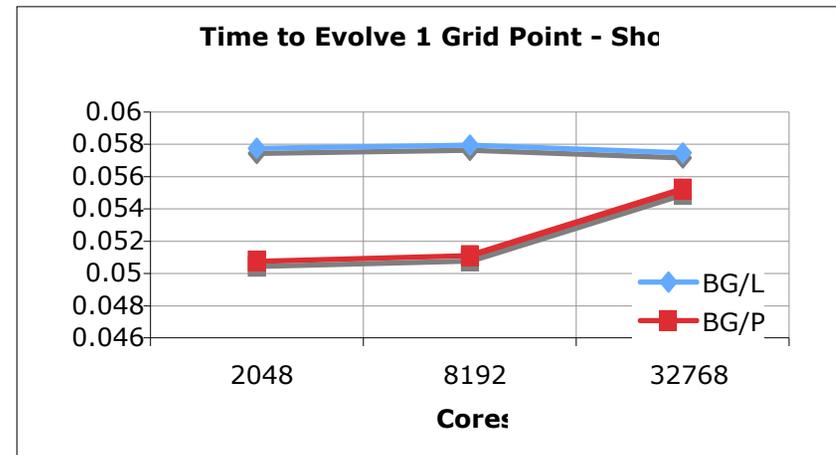
Orzag/Tang MHD vortex



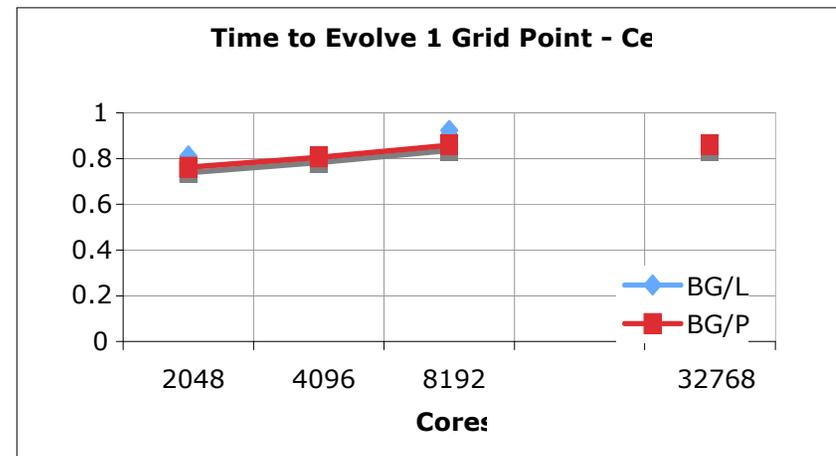
Richtmyer-Meshkov instability

FLASH Performance Tests on BG/L and BG/P

- FLASH running on BG/P since first day of application access to early hardware
- Simulations
 - Shock Cylinder
 - *Uniform Grid*
 - *Compressible hydrodynamics*
 - Cellular
 - *Block structured adaptive grid*
 - *Compressible, reactive flow*
- vn mode
 - Single core performance on P comparable better than L
 - Includes strong and weak scaling
- Dual Mode
 - Indications of significant core speedup for cellular



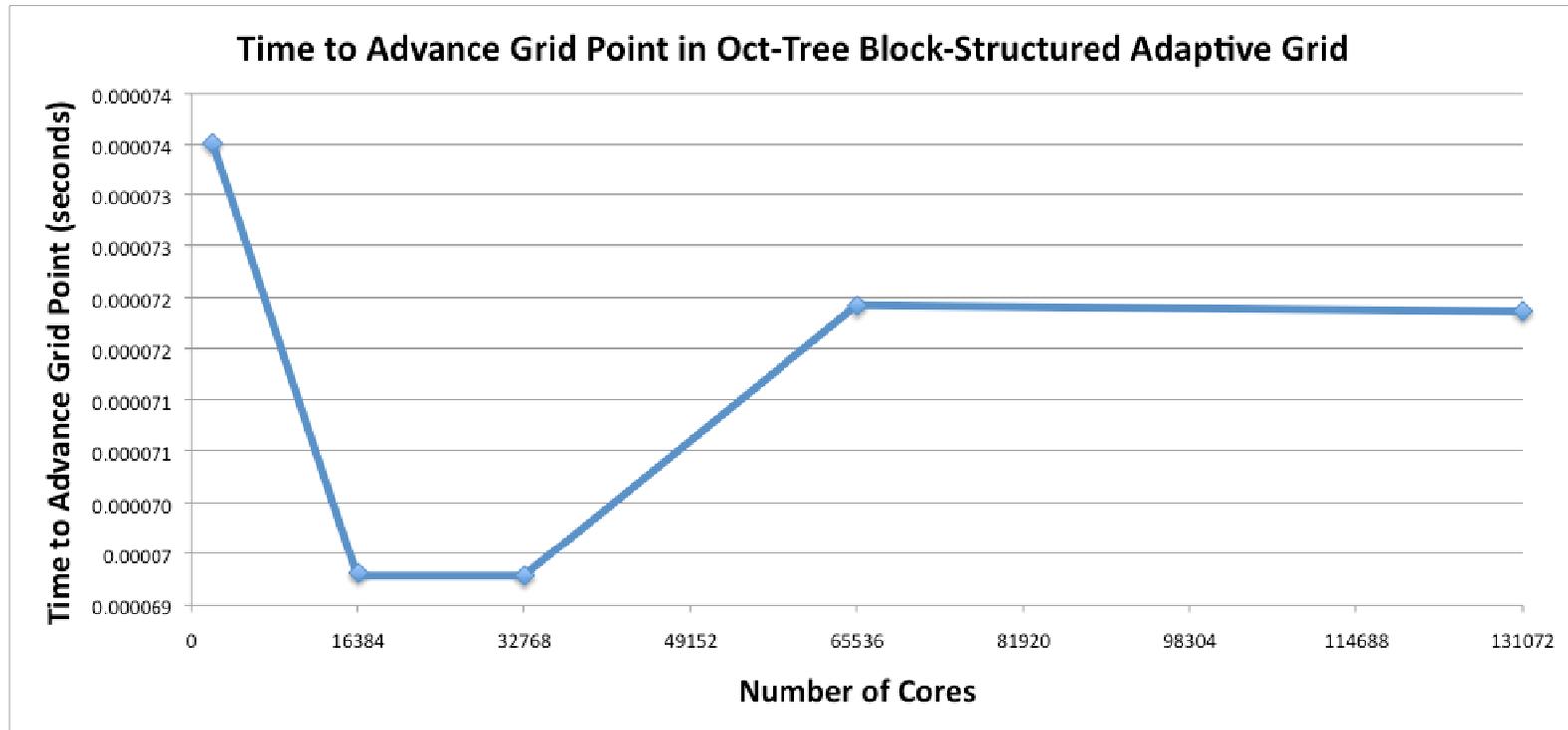
vn: BG/P from 14%-5% speedup



vn: BG/P comparable

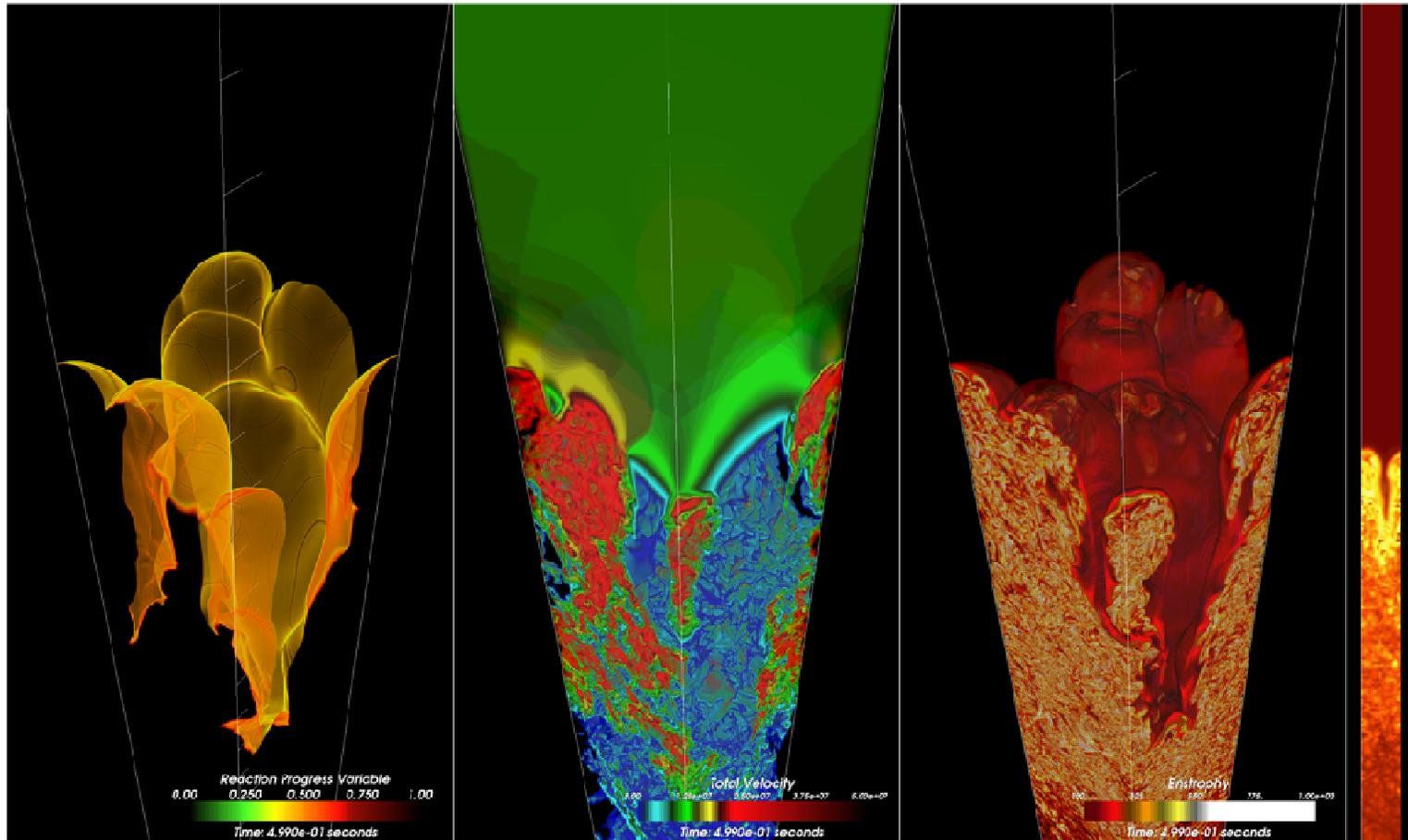
FLASH3 White Dwarf Problem on Blue Gene/P

Weak scaling



- 3-D Adaptive Grid with multiple physics
 - Explicit Hydro, Nuclear Burning, Gravity

FLASH: Turbulence-driven nuclear burning (rtflame)



Flash Conclusions

- Per-cpu optimizations most important
 - -O4 a big boost
- Using mass/massv and enforcing alignment helpful
 - Kernels improved 20%

NAMD

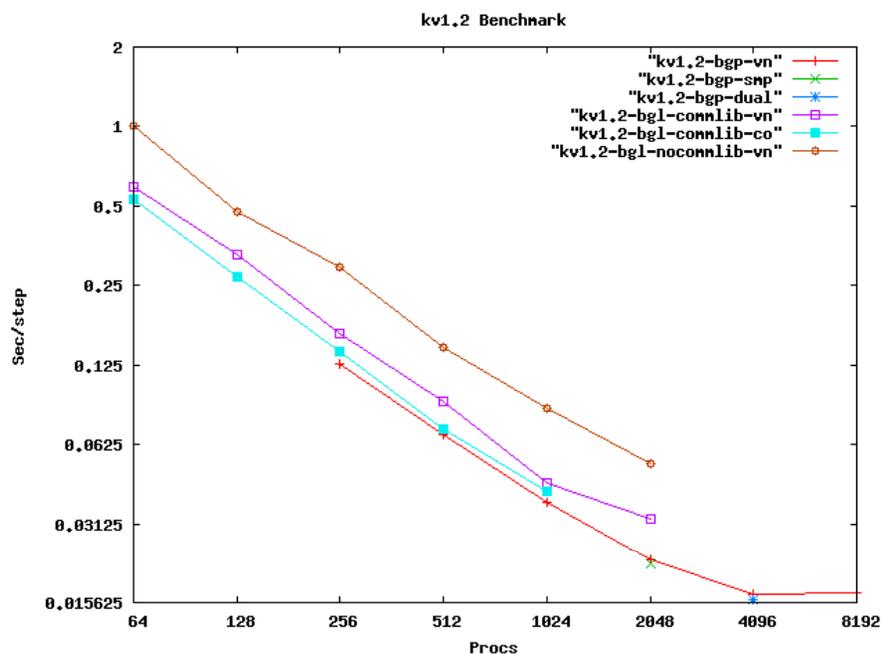
- Parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems
- Implemented in C/C++ using Charm++ parallel objects
- Gordon Bell award in 2002
- BlueGene/L performance: 1 rack 393GF, 4 racks 1.2TF

NAMD on BlueGene/P

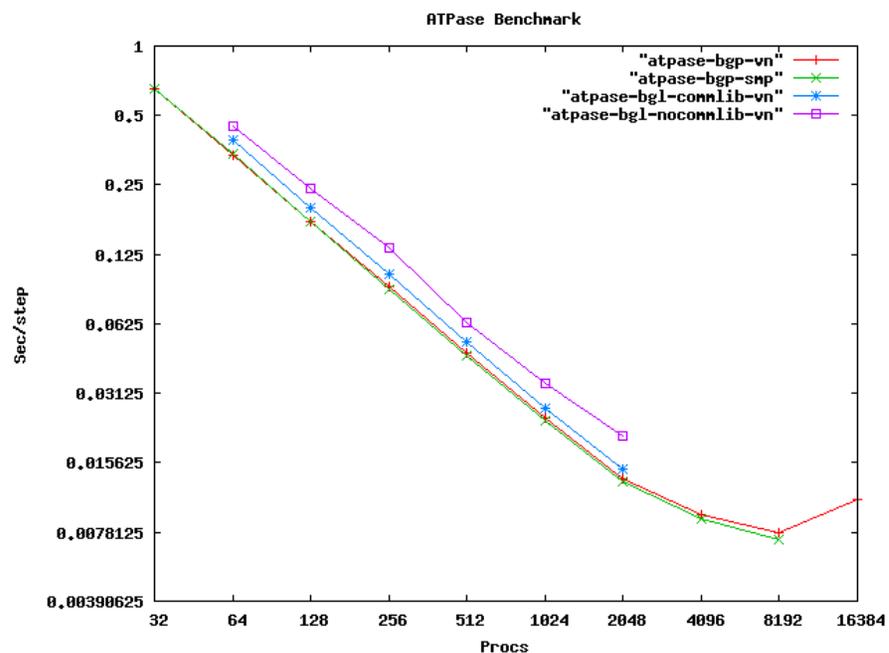
- Input data sets

- Apoa1 protein (90K atoms)
- kv1.2 voltage gated potassium ion (352K atoms)
- ATPase enzyme (327K atoms)

NAMD Scaling



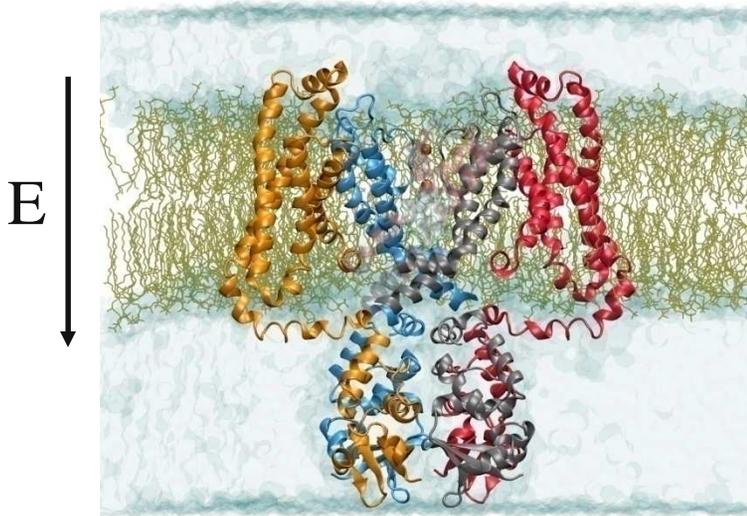
kv1.2 Benchmark (352K atoms)



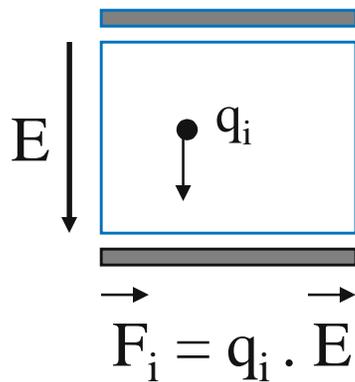
ATPase Benchmark (327K atoms)

Achieves 15-20% gain over BG/L customized ("conmlib") version.

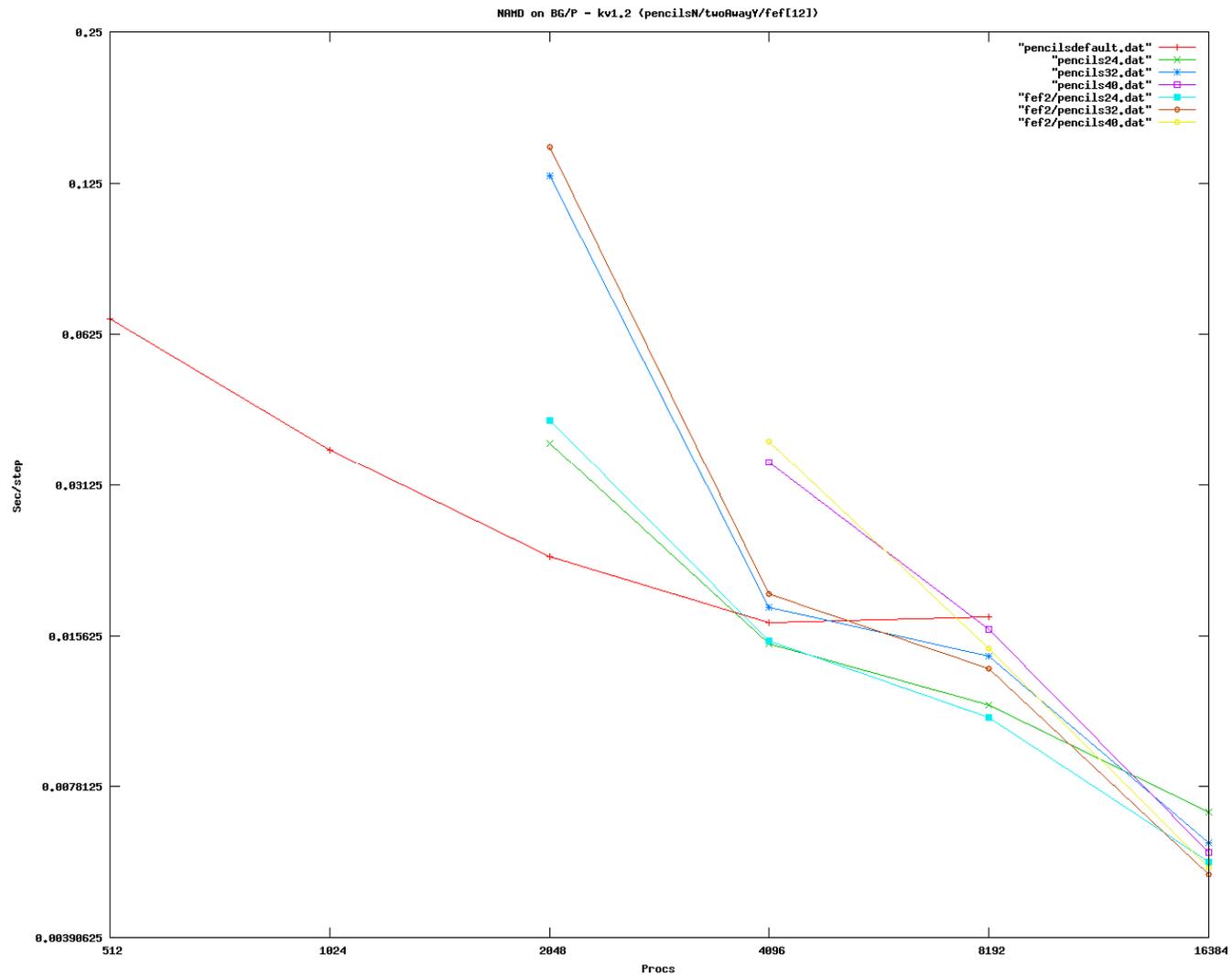
Structural Refinement (Molecular Dynamics)



- Open and closed state models of kv1.2 are equilibrated in DPPC bilayer
- 350K atoms (full-length channel)
- 100K atoms (voltage-sensor domain)
- constant pressure and temperature
- NAMD
- Voltage bias: $\pm 500\text{mV}$, $\pm 250\text{mV}$
- 250ns (full-length channel), 450ns (VSD)



NAMD PMEPencils + twoAway + fef2



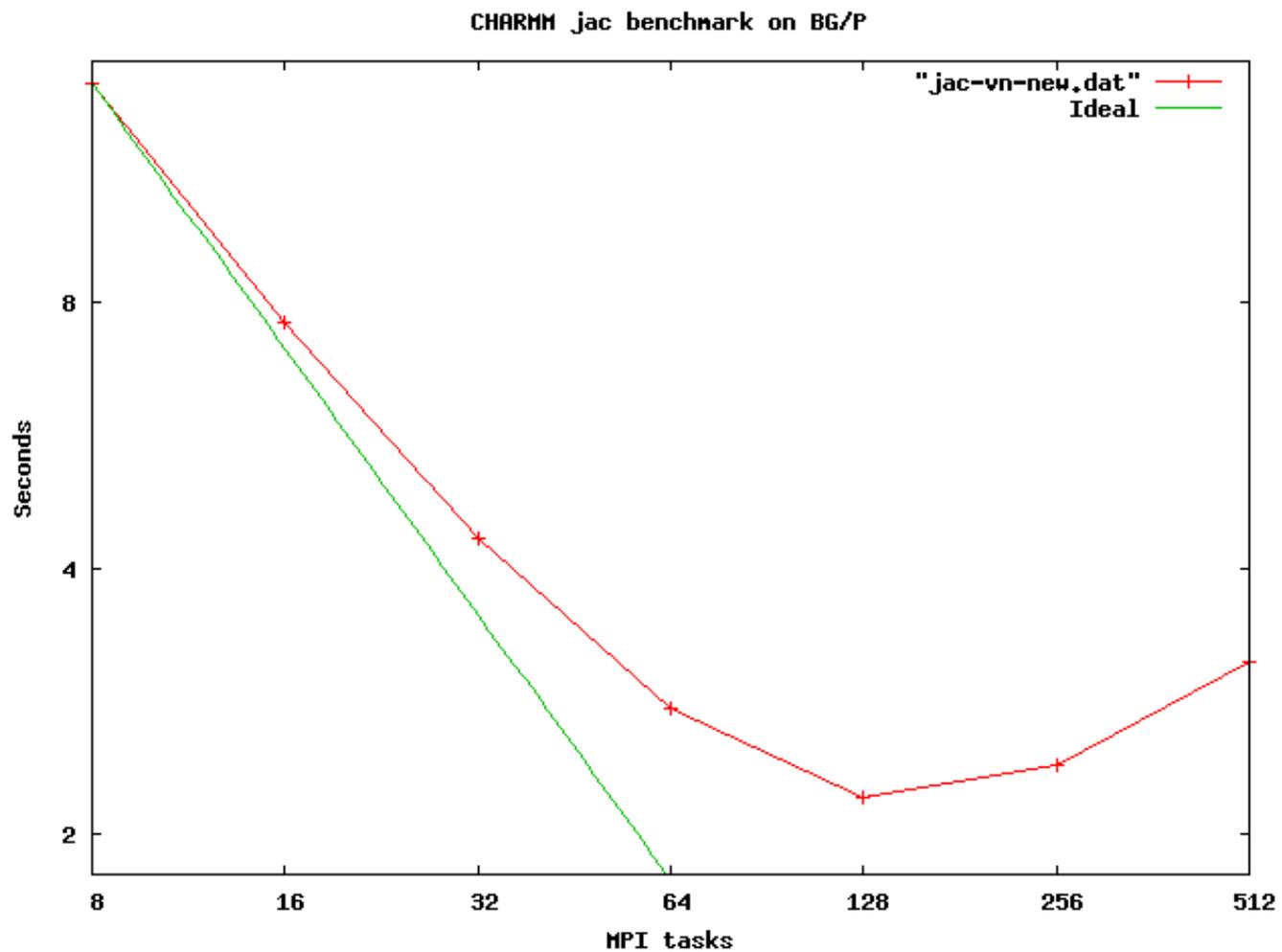
NAMD Conclusions

- Performance exceeds BG/L
 - *But does not need platform-specific enhancement*
- *The right parameter set is key to NAMD performance*

CHARMM

CHARMM (Chemistry at HARvard Macromolecular Mechanics)

CHARMM – jac benchmark (23.5k atoms)



CHARMM – coping with (lack of) scaling

- Multiple runs needed!
 - Replica exchange
 - Free energy calculations
- HTC mode?
 - Need hybrid mode
- Multi-CHARMM ?
 - In progress
- Static memory
 - LARGE mode is max in vn

Conclusions

- Applications surpass BG/L performance
 - Many scale well to large rack counts
- Flash and MILC instrumental in detecting HW issues during acceptance
- Optimizations
 - Serial
 - Libraries
- Communications better than BG/L even without platform-specific code
- Parameter tuning
- IO