NAMD on BlueWaters

Presented by: Eric Bohm
Team: Eric Bohm, Chao Mei, Osman Sarood, David Kunzman, Yanhua, Sun, Jim Phillips, John Stone, LV Kale

NSF/NCSA Blue Waters Project

- Sustained Petaflops system funded by NSF to be ready in 2011.
  - System expected to exceed 300,000 processor cores.
- NSF Acceptance test: 100 million atom Bar Domain simulation using NAMD.
- NAMD PRAC The Computational Microscope
  - Systems from 10 to 100 million atoms
- A recently submitted PRAC from an independent group wishes to use NAMD
  - 1 Billion atoms!

NAMD

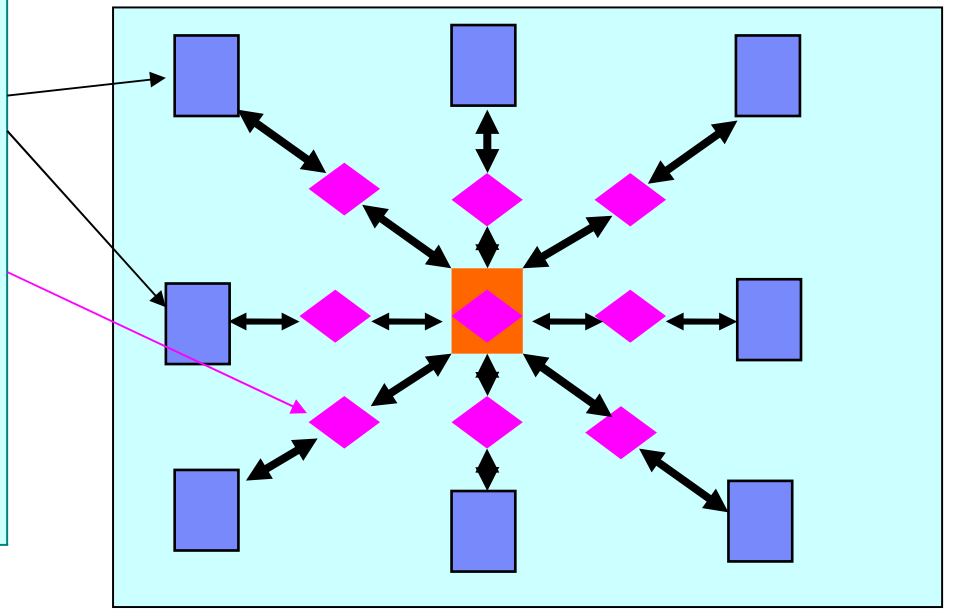
- Molecular Dynamics simulation of biological systems
- Uses the Charm++ idea:
  - Decompose the computation into a large number of objects
  - Have an Intelligent Run-time system (of Charm++) assign objects to processors for dynamic load balancing

Hybrid of spatial and force decomposition:

•Spatial decomposition of atoms into cubes (called patches)

•For every pair of interacting patches, create one object for calculating electrostatic interactions

•Recent: Blue Matter, Desmond, etc. use this idea in some form



BW Challenges and Opportunities

- Support systems >= 100 Million atoms
- Performance requirements for 100 Million atom
- Scale to over 300,000 cores
- Power 7 Hardware
  - PPC architecture
  - Wide node at least 32 cores with 128 HT threads
- BlueWaters Torrent interconnect
- Doing research under NDA

BlueWaters Architecture

- IBM Power7
- Peak Perf ~10 PF
- Sustained ~1 PF
- 300,000+ cores
- 1.2+ PB Memory
- 18+ PB Disc
- 8 cores/chip
- 4 chips/MCM
- 8 MCMs/Drawer
- 4 Drawers/SuperNode
- 1024 cores/SuperNode
- Linux OS

Power 7

- 64-bit PowerPC
- 3.7-4Ghz
- Up to 8 FLOPs/cycle
- 4-way SMT
- 128 byte cache lines
- 32 KB L1
- 256 KB l2
- 4 MB local in shared 32 MB L3 cache

- 2 fixed point, 2 load store
- 1 VMX
- 1 decimal FP
- 2 VSX
  - 4 FLOPs/cycle
- 6-wide in-order
- 8-wide out-of-order
- 12 data streams prefetch

Hub Chip Module

- Connects 8 QCMs via L-local (copper)
  - 24 GB/s
- Connects 4 P7-IH drawers L-remote (optical)
  - 6 GB/s
- Connects up to 512 SuperNodes D (optical)
  - 10 GB/s

Availability

- NCSA has BlueDrop machine
  - Linux
  - IBM 780 (MR) POWER7 3.8 Ghz
  - Login node 2x8 core processors
  - Compute note 4x8 core in 2 enclosures
- BlueBioU
  - Linux
  - 18 IBM 750 (HV32) nodes 3.55 Ghz
  - Infiniband 4x DDR (Galaxy)

NAMD on BW

- Use SMT=4 effectively
- Use Power7 effectively
  - Shared memory topology
  - Prefetch
  - Loop unrolling
  - SIMD VSX
- Use Torrent effectively
  - LAPI/XMI

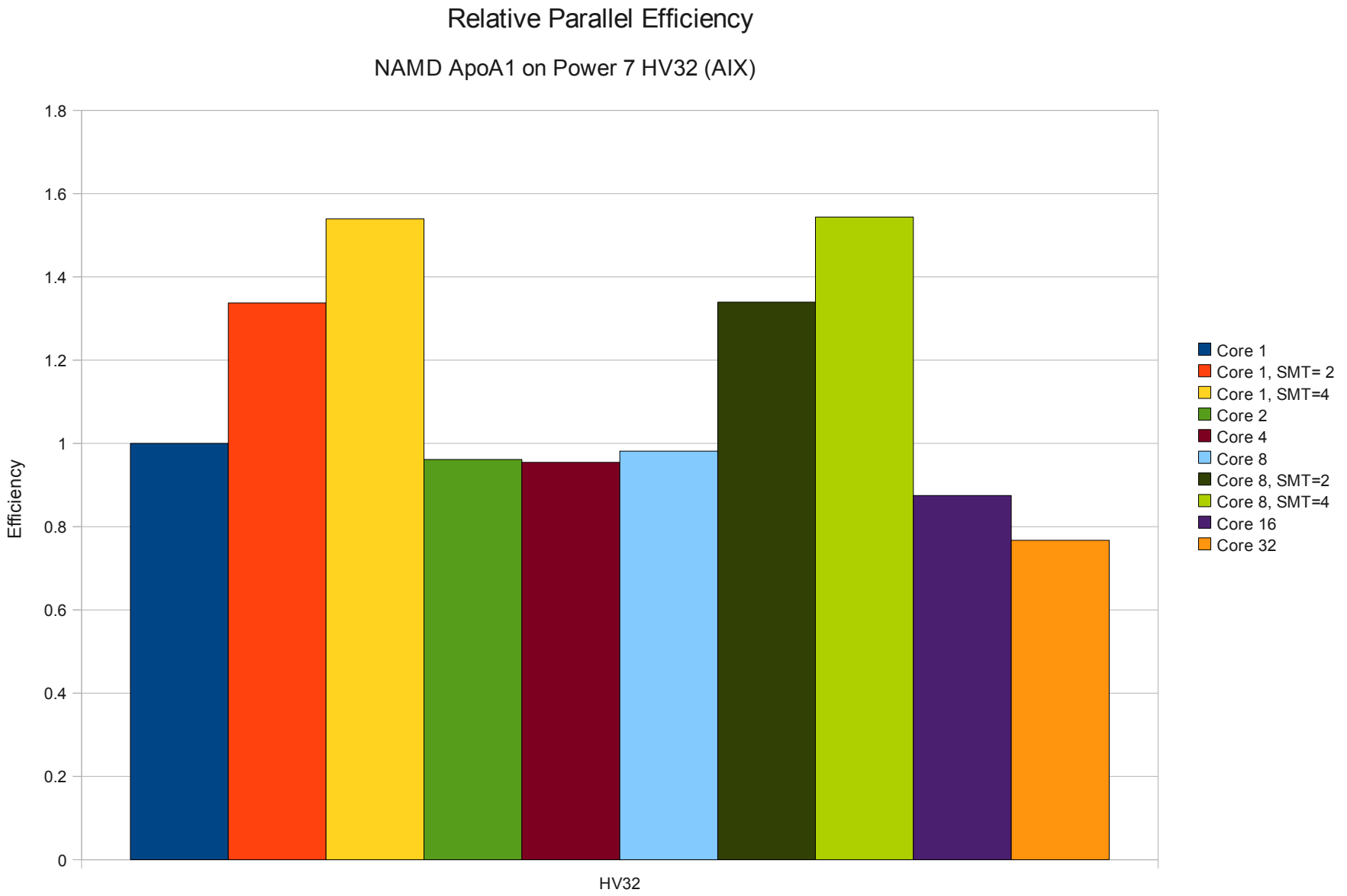
Petascale Scalability Concerns

- Centralized load balancer - solved
- IO
  - Unscalable file formats - solved
  - input read at startup - solved
  - Sequential output – in progress
- Fine grain overhead – in progress
- Non-bonded multicasts – being studied
- Particle Mesh Ewald
  - Largest grid target <= 1024
  - Communication overhead primary issue
  - Considering Multilevel Summation alternative

NAMD and SMT=4

- P7 hardware threads are prioritized
  - 0,1 highest
  - 2,3 lowest
- Charm runtime measure processor performance
  - Load balancer operates accordingly
- NAMD on SMT=4 35% faster than SMT=1
  - No new code required!
- At the limit it requires 4x more decomposition

NAMD on Power7 HV 32 AIX



SIMD -> VSX

- VSX adds double precision support to VMX
- SSE2 already in use in 2 NAMD functions
- MD-SIMD implementation of nonbonded MD benchmark available from Kunzman
- Translate SSE to VSX
- Add VSX support to MD-SIMD

MD-SIMD performance

Support for Large Molecular Systems

- New Compressed PSF file format
  - Supports >100 million atoms
  - Supports parallel startup
  - Support MEM_OPT molecule representation
- MEM_OPT molecule format reduces data replication through signatures
- Parallelize reading of input at startup
  - Cannot support legacy PDB format
  - Use binary coordinates format
- Changes in VMD courtesy John Stone

Parallel Startup

Table 1: Parallel Startup for 10 Million water on BlueGene/P

Nodes	Start (sec)	Memory(MB)
1	NA	4484.55 *
8	446.499	865.117
16	424.765	456.487
32	420.492	258.023
64	435.366	235.949
128	227.018	222.219
256	122.296	218.285
512	73.2571	218.449
1024	76.1005	214.758

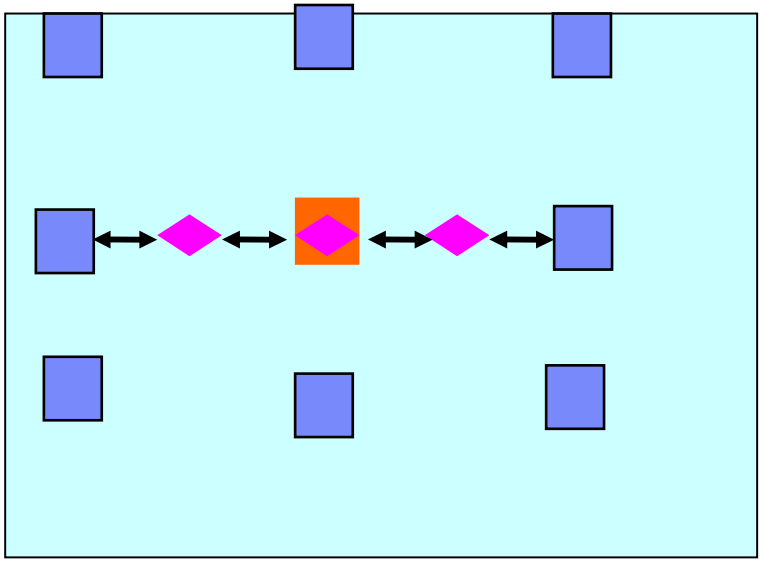
Table : Parallel Startup 116 Million BAR domain on Abe

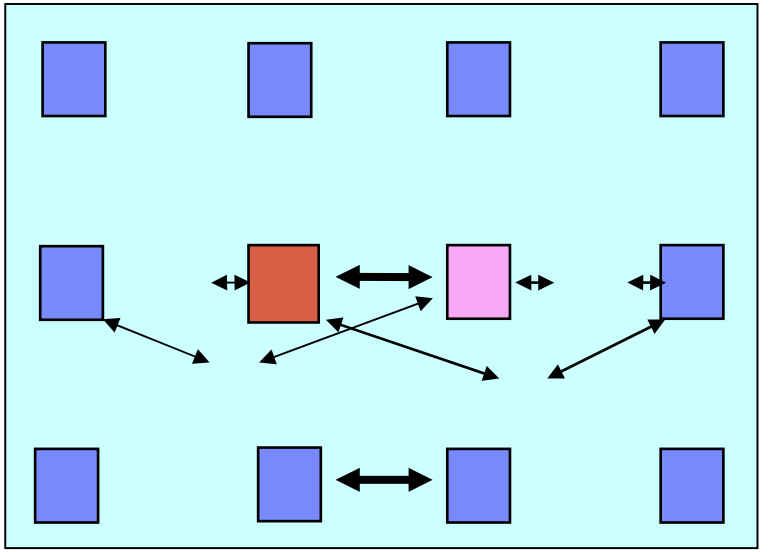
Nodes	Start (sec)	Memory (MB)
1	3075.6 *	75457.7 *
50	**340.361**	1008
80	**322.165**	908
120	**323.561**	710

Fine grain overhead

- End user targets are all fixed size problems
- Strong scaling performance dominates
  - Maximize number of nanoseconds/day of simulation
- Non-bonded cutoff distance determines patch size
  - Patch can be subdivided along x, y, z dimensions
    - 2 away X, 2-away XY, 2 away XYZ
      - Theoretically K-away...

1-away vs 2-away X





Fine-grain overhead reduction

- Distant computes have little or no interaction
  - Long diagonal opposites of 2-awayXYZ mostly outside of cutoff
- Optimizations
  - Don't migrate tiny computes
  - Sort pairlists to truncate computation
  - Increase margin and do not create redundant compute objects
- Slight (<5%) reduction in step time

Future work

- Integrate parallel output into CVS NAMD
- Consolidate small compute objects
- Leverage native communication API
- Particle Mesh Ewald improve/replace
- Parallel I/O optimization study on multiple platforms
- High (>16k) scaling study on multiple platforms