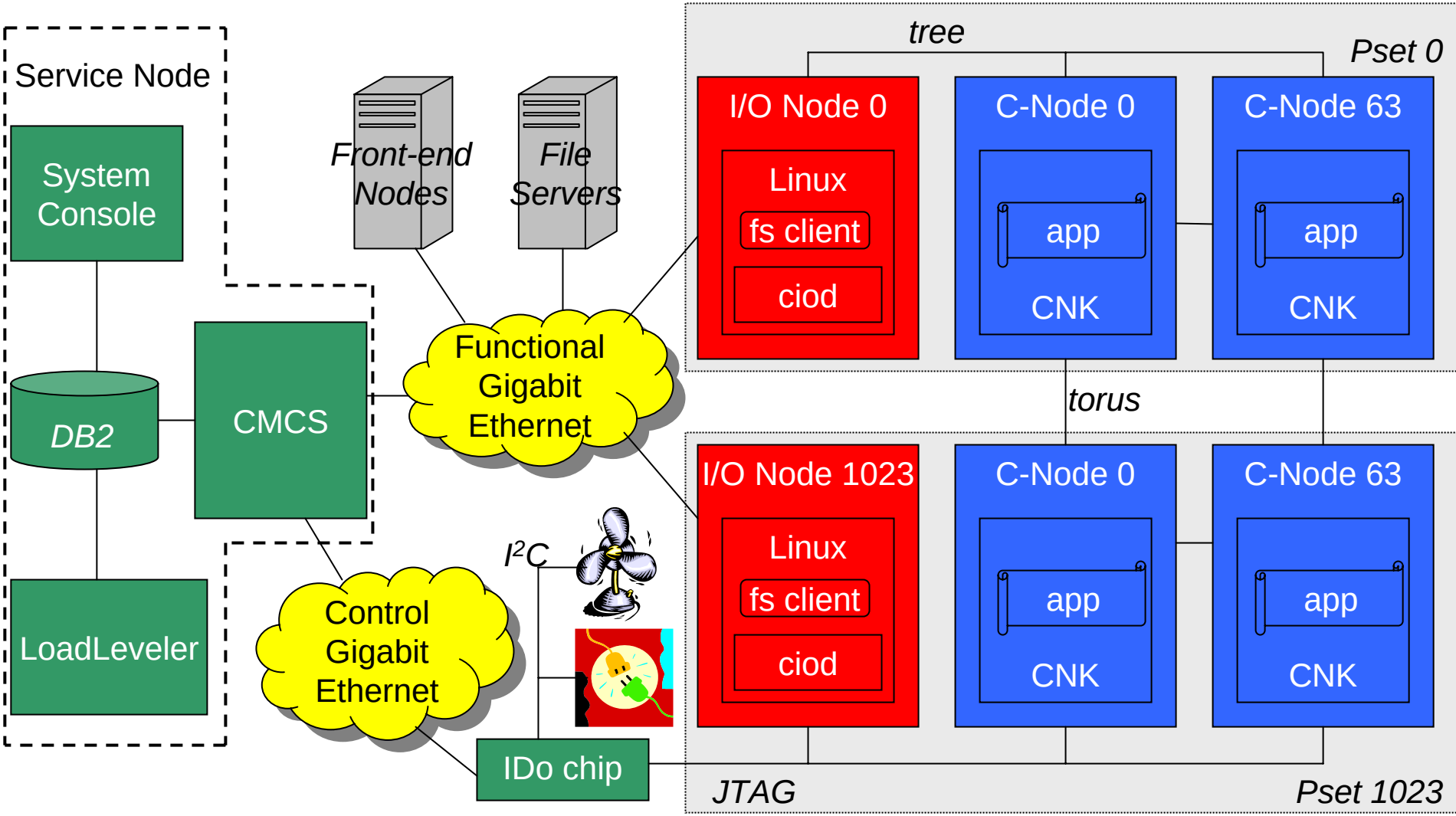
IBM Systems and Technology Group

Blue Gene/L Programming Environment

Manish Gupta
IBM Research

© 2005 IBM Corporation

Blue Gene/L System Architecture



© 2005 IBM Corporation

HPC Tools Software for Blue Gene

IBM Software Stack

- XL Compilers
  - Externals preserved
  - New options to optimize for specific Blue Gene functions
- LoadLeveler
  - Same externals for job submission and system query functions
  - Backfill scheduling to achieve maximum system utilization
- GPFS
  - Provides high performance file access, as in current pSeries and xSeries clusters
  - Runs on IO nodes and disk servers
- ESSL/MASSV
  - Optimization library and intrinsics for better application performance
  - Serial Static Library supporting 32-bit applications
  - Callable from FORTRAN, C, and C++
- MPI
  Message passing library, based on MPICH2, tuned for Blue Gene architecture

Other Software Support

- Parallel Debugger
  - Etnus TotalView
- Parallel File Systems
  - Lustre at LLNL, PVFS2 at ANL
- Job Schedulers
  - SLURM at LLNL, Cobalt at ANL
- FFT Library
  - Tuned functions by TU-Vienna
- Performance Tools
  - HPC Toolkit: MP_Profiler, Xprofiler, HPM, PeekPerf
  - Paraver, Kojak, Tau, …

© 2005 IBM Corporation

BG/L Compilers: Initial Experience

- Compiler finding SIMDization opportunities (generating Double Hummer code), but performance improvements not satisfactory
  - Successes
    - 39%-97% performance improvement for HPCC/STREAM benchmark
    - 65% performance improvement for UMT2K
    - Extensive diagnostics reports from compiler – can be used to diagnose problems
  - Challenges
    - Significant performance degradation from SIMDization for most codes and SPECfp2000 benchmarks
- Key to driving work further – single node performance analysis
  - Using profiling tools, identified several individual loops with performance degradation ranging from 8%-80%
    - Further analysis identified several stalls in loads due to microarchitectural constraints – coupling of primary and secondary registers (unexpected), three outstanding loads (known issue)
    - Gone from 8%-80% degradation to 2%-43% improvement in specific loops – up to 10% performance improvement at overall application level
  - In collaboration with customers (LLNL, SDSC), identified several missed opportunities for SIMDization

© 2004 IBM Corporation

Challenges with running/scaling codes on BG/L

- Memory per node too little for some codes!
  - Serial component in algorithm
    - e.g., UMT2K - graph partitioning being done serially
      - Showed up as “insufficient memory” problem as we scaled beyond 2K processors
  - Single processor doing all I/O
    - Common in weather/climate codes
      - Use “fat” nodes with more memory for those I/O operations
      - Right solution – use parallel I/O
  - Read-only tables that require more memory
    - More fundamental problem – workaround to distribute tables introduces unwelcome complexity
    - Sharing a single copy on a multiprocessor node requires extensions to MPI model
- Exit if number of processors exceeds threshold – need to rewrite!

© 2005 IBM Corporation

Other challenges with scaling codes on BG/L

- Load imbalance in application – e.g., ParaDiS
- Tuning Alltoall communication to improve scaling – e.g., Miranda
- Some applications sensitive to mapping to 3D torus toplogy – e.g., QBox, HOMME
- Pacing of messages to avoid network overload – e.g., Raptor

© 2005 IBM Corporation

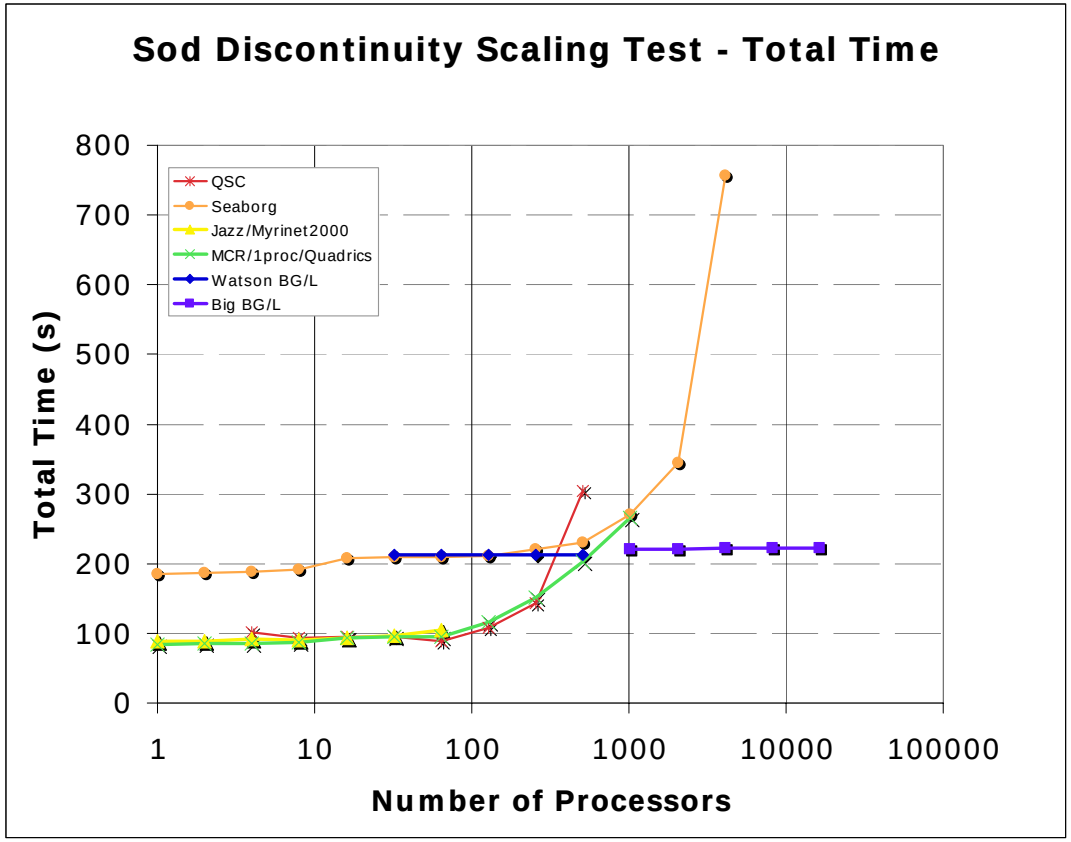
Other challenges with scaling codes on BG/L

- Load imbalance in application – e.g., ParaDiS
- Tuning Alltoall communication to improve scaling – e.g., Miranda
- Some applications sensitive to mapping to 3D torus toplogy – e.g., QBox, HOMME
- Pacing of messages to avoid network overload – e.g., Raptor

Usual experience, excellent scaling with relatively little pain

© 2005 IBM Corporation

FLASH: Astrophysics Code from Argonne National Lab SCALING TO 16x1024 nodes on Blue Gene/L



Big BGL: 16 Racks, coprocessor, 440

Jazz : 350node, 2.4GHz Xeon, ANL

MCR : 1152node,2.4GHz Xeon,LLNL

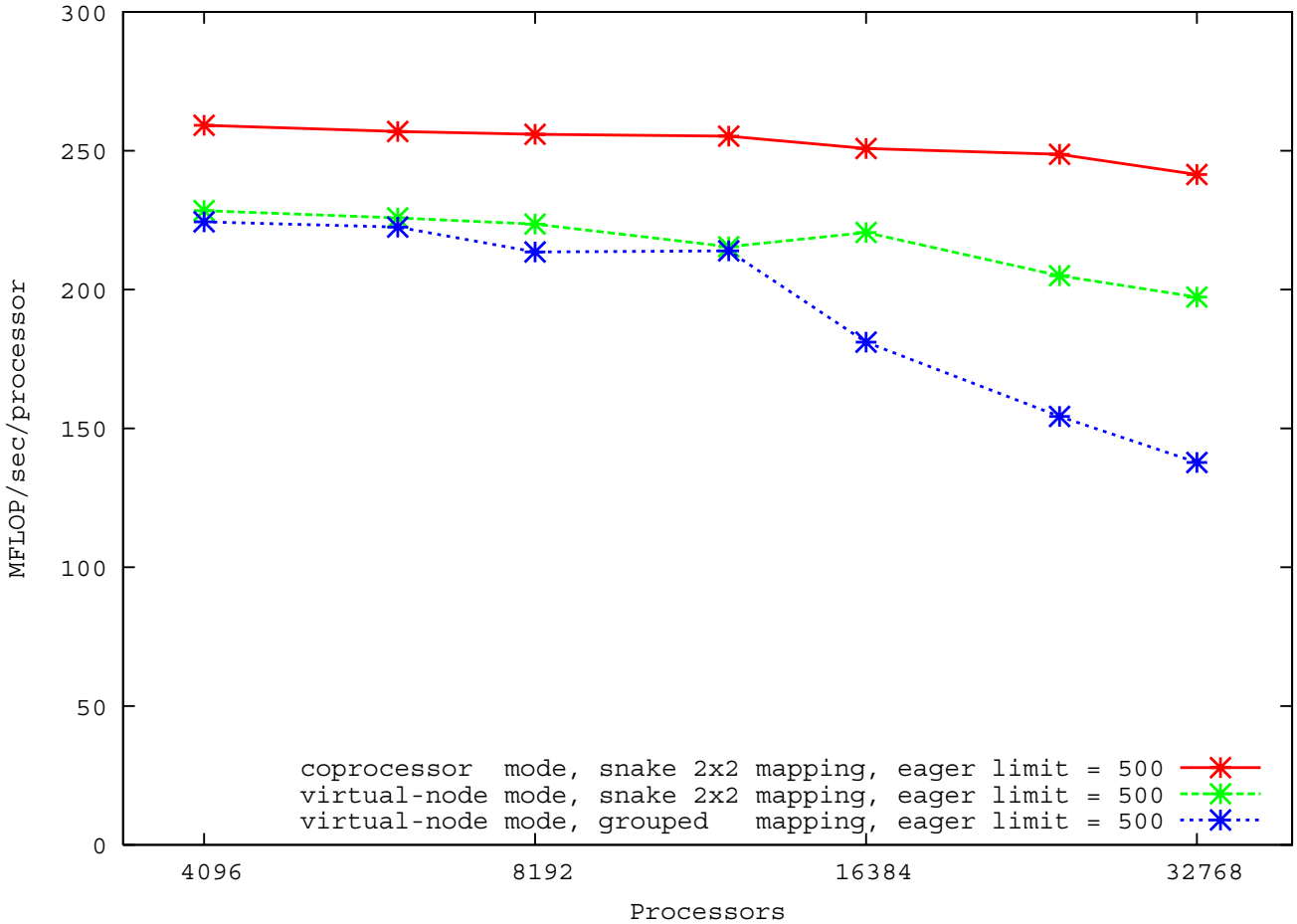
Seaborg: IBM SP, 1.5GF/node NERSC

QSC: 256nodex4way HP Alpha, LLNL

© 2004 IBM Corporation

HOMME (NCAR)

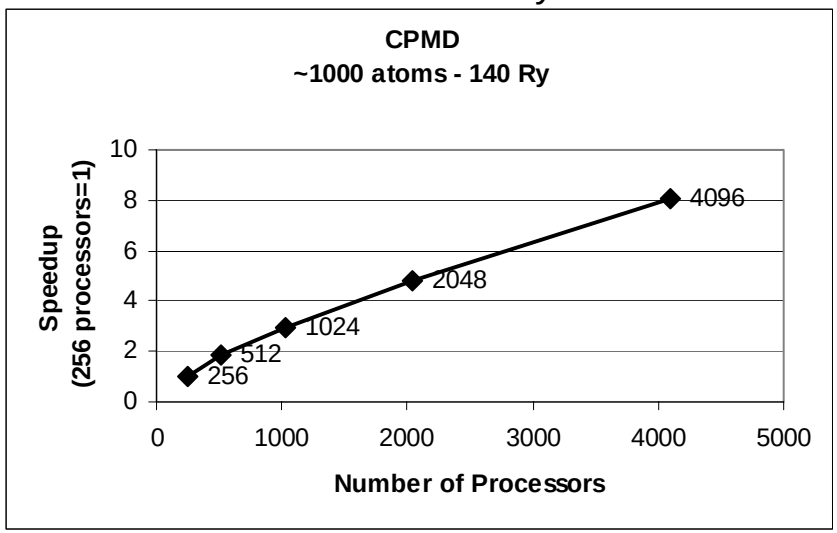
Climate Modeling: moist processes, cloud formation, precipitation, moist convection

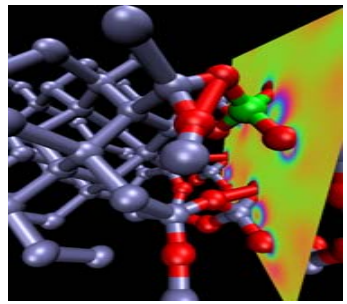


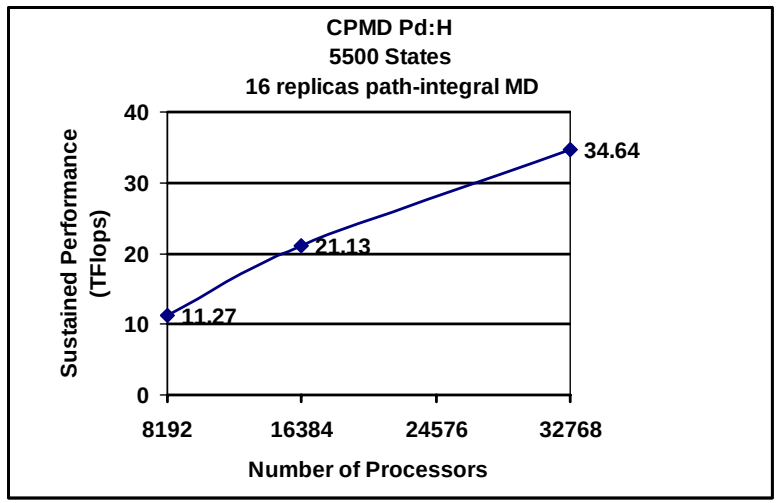
CPMD Performance

Ongoing project : IBM/LLNL PdH Hydrogen Storage
Achieved 34.64 Teraflop/s sustained on BG/L (excellent strong scaling)

Ab-initio Molecular Dynamics





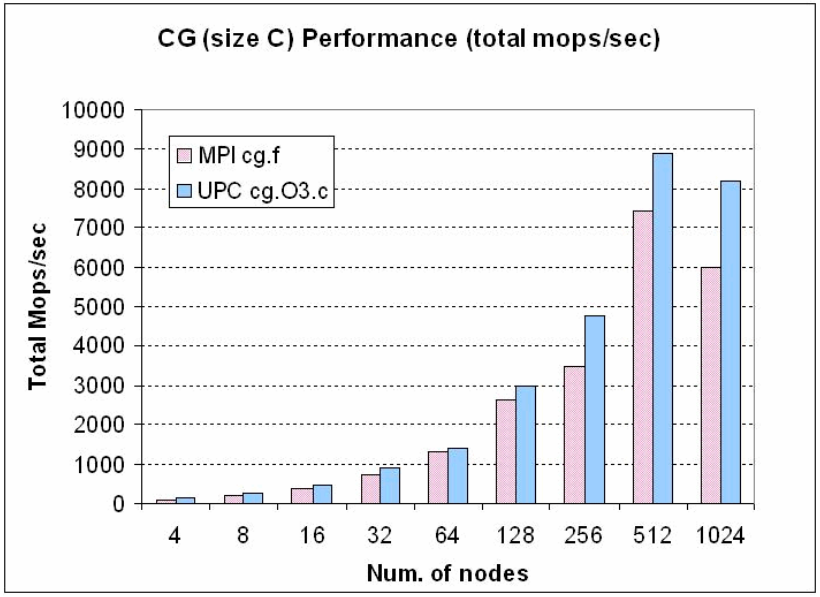


10 sec/step on 2048 BG/L
25 sec/step on 1400 Xeon Cluster at LLNL.

32K procs – 34.64 Teraflop/s

Other APIs being explored

- ARMCI / Global Arrays (GA)
  - **Active message library used by GA for point-2-point communications**
    - GA (and NWChem) also use MPI for collectives, some p2p as well
- NAMD on Charm++/Converse
  - new Converse machine layer using commlib
  - Converse microbenchmark performance improvement:
    - bandwidth increased from 124MB/s (Converse/MPI) to 154MB/s (Converse/commlib)
    - latency down from 10.1us to 6.9us.
  - NAMD scaling to 4K nodes
- UPC (Unified Parallel C)
  - Port of IBM UPC runtime to BG/L
  - Added one-sided communication primitives to commlib: BLRMA_Get, BLRMA_Put (under development)
  - 11.54 GUPS on 32K nodes
  - 47,827 GB/s STREAM triad on 32K nodes



© 2005 IBM Corporation

Conclusions

- Blue Gene/L represents a new level of performance scalability and density for scientific computing
  - Leverages system-on-a-chip technology and powerful interconnects
  - Judicious use of redundancy leads to high reliability system
  - Low power and floorspace requirements
- Blue Gene/L system software stack with Linux-like personality for applications
  - Custom solution (CNK) on compute nodes for highest performance
  - Linux solution on I/O nodes for flexibility and functionality
  - MPI is the default programming model, others are being investigated
  - Complements the hardware to deliver an extremely scalable system – low latency, high bandwidth communication, with negligible noise
- Very encouraging performance results on benchmarks and applications – people beginning to do real science with BG/L
- Continued investment in Blue Gene roadmap

© 2005 IBM Corporation