The Blue Gene/L Architecture

George Chiu

Lawrence Livermore National Laboratory

Blue Gene/L (BGW) at T.J. Watson Research

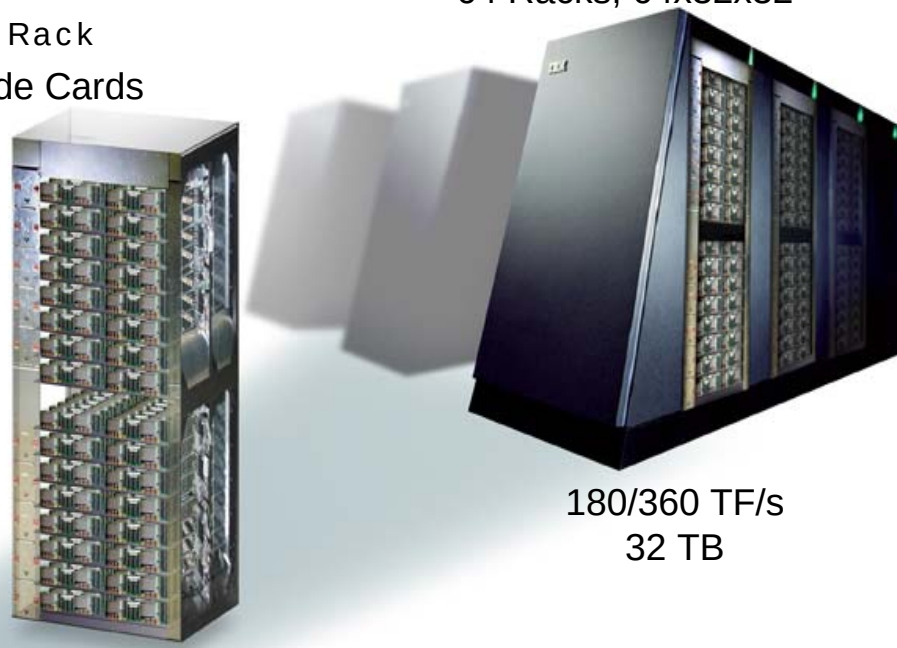
BlueGene/L System Buildup

System

64 Racks, 64x32x32

Rack

32 Node Cards

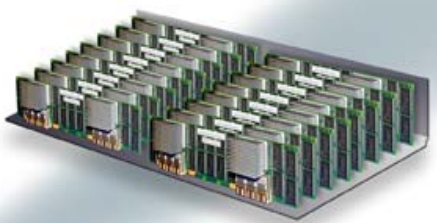


180/360 TF/s
32 TB

2.8/5.6 TF/s
512 GB

Node Card

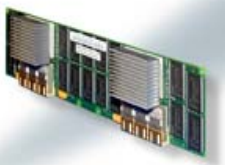
(32 chips 4x4x2)
16 compute, 0-2 IO cards



90/180 GF/s
16 GB

Compute Card

2 chips, 1x2x1



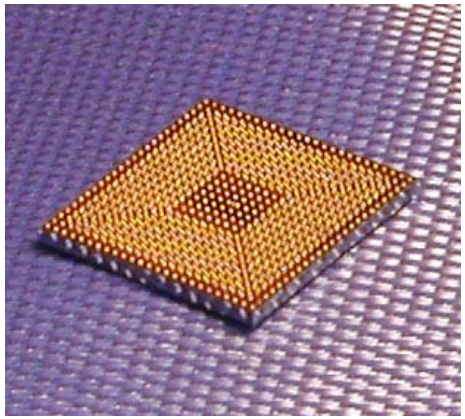
5.6/11.2 GF/s
1.0 GB

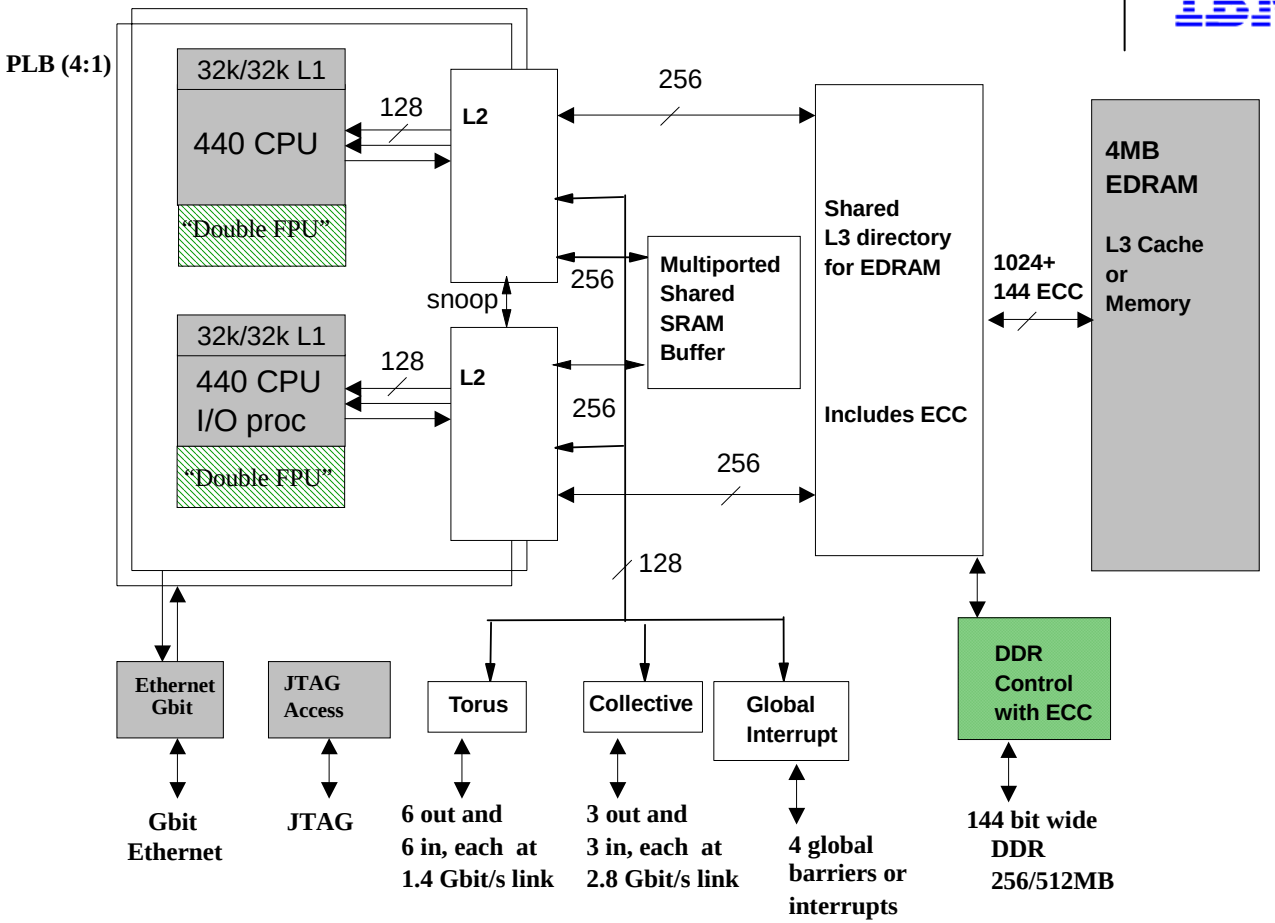
Chip

2 processors

2.8/5.6 GF/s
4 MB

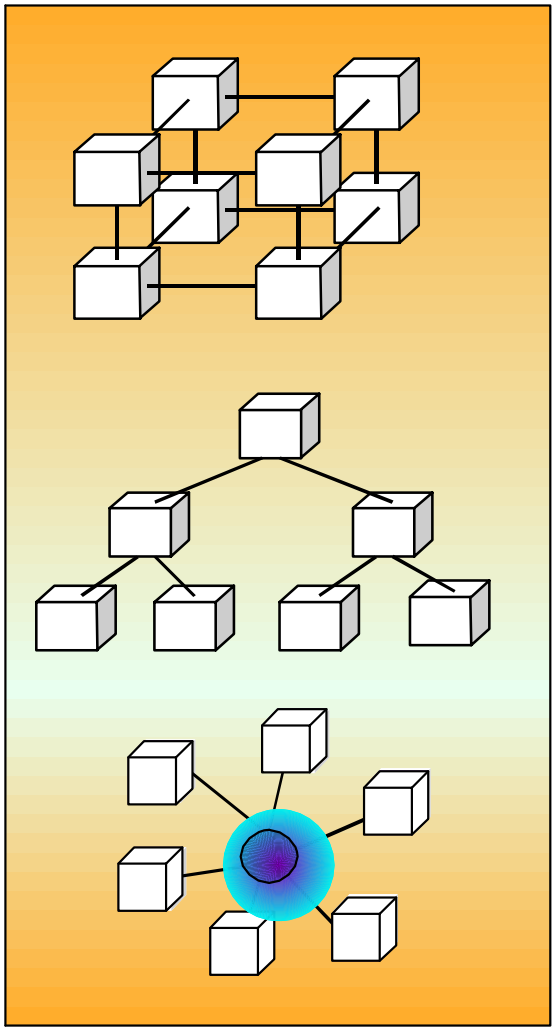
BlueGene/L Compute ASIC





- IBM CU-11, 0.13 µm
- 11 x 11 mm die size
- 25 x 32 mm CBGA
- 474 pins, 328 signal
- 1.5/2.5 Volt

BlueGene/L Interconnection Networks



3 Dimensional Torus

- Interconnects all compute nodes (65,536)
- Virtual cut-through hardware routing
- 1.4Gb/s on all 12 node links (2.1 GB/s per node)
- 1 µs latency between nearest neighbors, 5 µs to the farthest
- MPI: 3.3 µs latency for one hop, 10 µs to the farthest
- Communications backbone for computations
- 0.7/1.4 TB/s bisection bandwidth, 68TB/s total bandwidth

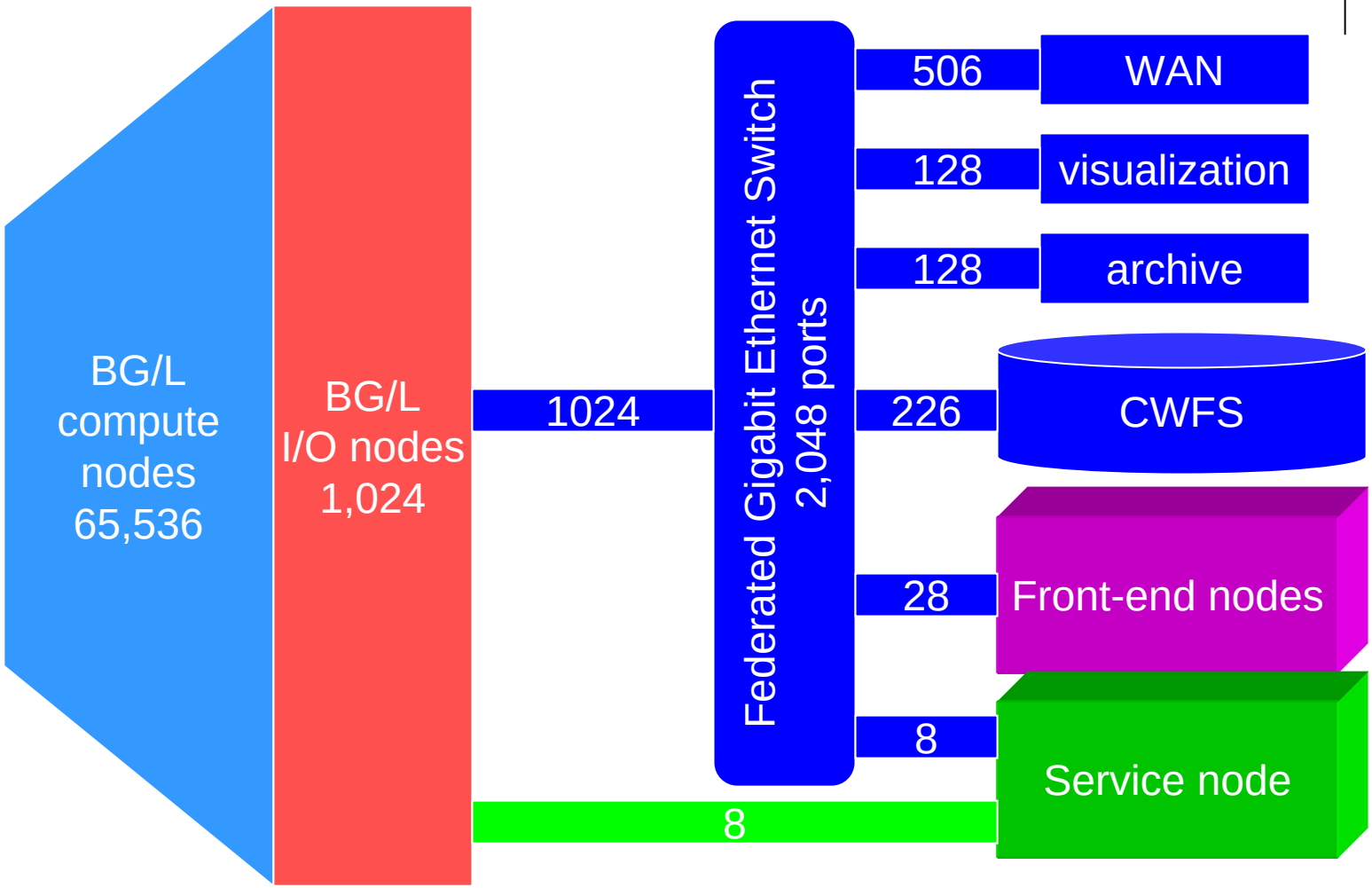
Collective Network

- One-to-all broadcast functionality
- Reduction operations functionality
- 2.8 Gb/s of bandwidth per link
- Latency of one way tree traversal 2.5 µs, MPI 6 µs
- ~23TB/s total binary tree bandwidth (64k machine)
- Interconnects all compute and I/O nodes (1024)

Low Latency Global Barrier and Interrupt

- Latency of one way to reach all 64K nodes 0.65 µs, MPI 1.6 µs

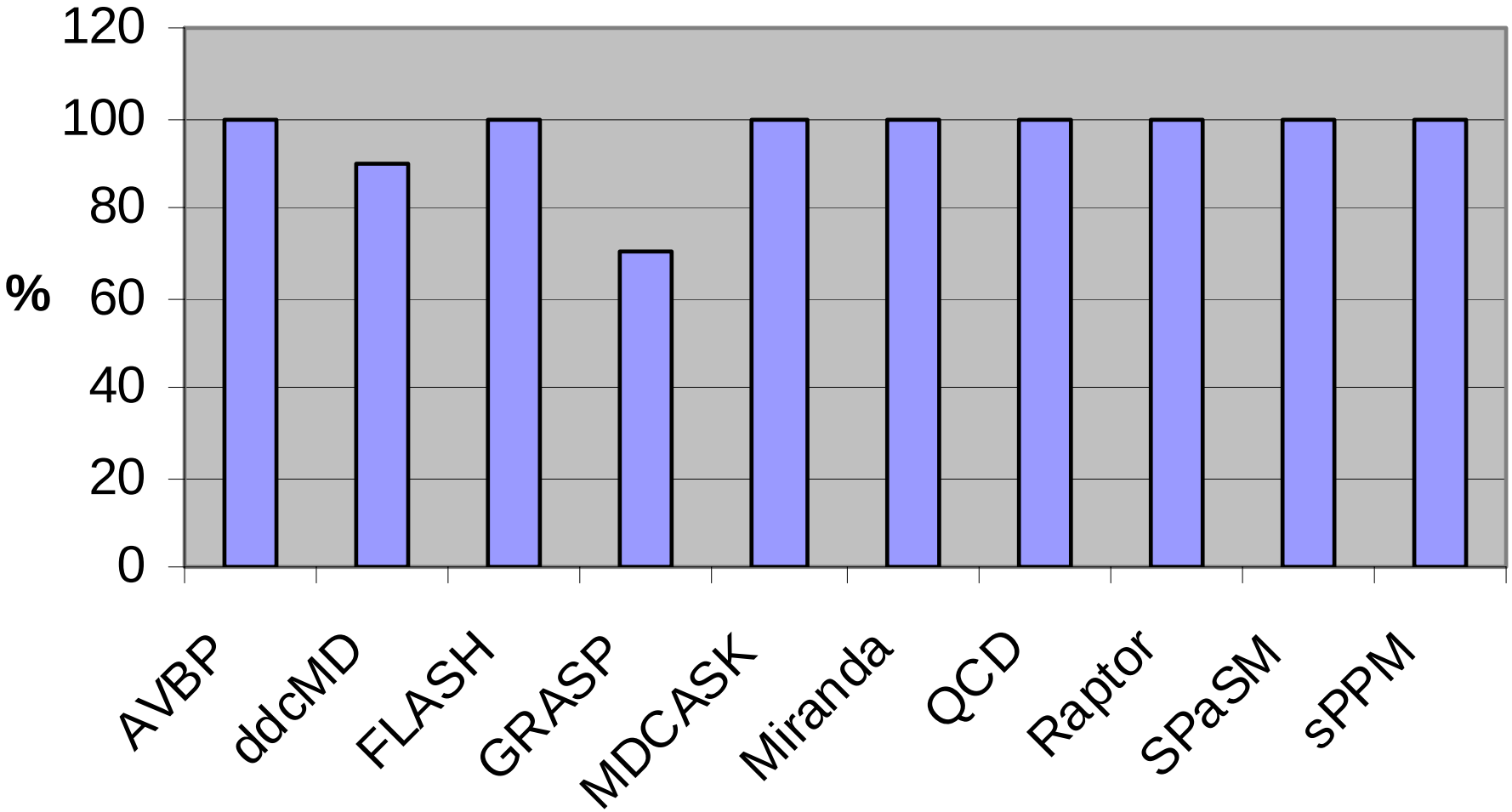
Complete BlueGene/L System at LLNL



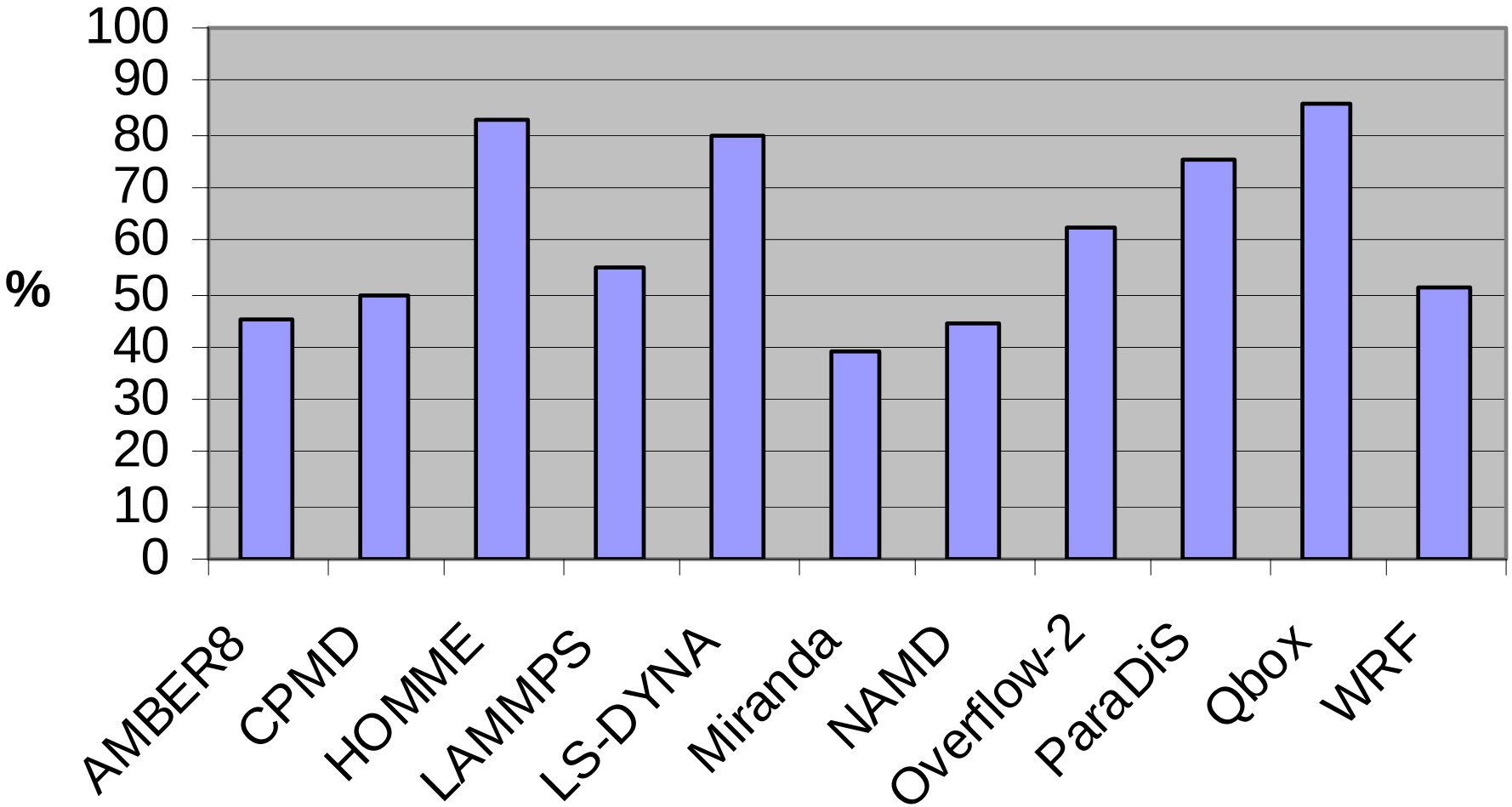
Closing Points

- Blue Gene represents an innovative way to scale to multi-teraflops capability
  - Low power => efficient packaging => floor space consumption
  - Massive scalability
  - Balanced design in terms of memory subsystem and communication subsystem
  - Better than COTS clusters by virtue of density, scalability and innovative interconnect design
  - Bottom up hardware architecture
  - Top Down software design
- Blue Gene is applicable to a wide range of HPC workloads
- Blue Gene R&D continues to protect the investment of users

Weak Scaling Parallel Efficiency



Strong Scaling Parallel Efficiency



Applications

- N-body simulation
  - Classical molecular dynamics: AMBER8, Blue Matter, ddcMD, DL_POLY, GRASP, GROMACS, LAMMPS, LJ, MDCASK, NAMD, PRIME (Schrodinger), SPaSM
  - Quantum chemistry: CHARMm, CPMD, FEQMD, GAMESS-UK, GAMESS-US, Gaussian, NWChem, Qbox,
  - Plasma Physics: TBLE, GTC
  - Stellar dynamics of galaxies: Enzo
- Complex multiphysics code
  - Climatology: CCSM, HOMME
  - Computational Fluid Dynamics: FLUENT, Miranda, Overflow, POP (Ocean), Raptor, SAGE, sPPM, STAR-CD,
  - Astronomy: Accretion, planetary formation and evolution, stellar evolution, FLASH (supernova), Radiotelescope (ASTRON)
  - Electromagnetics: FDTD code
  - Finite element analysis, Car Crash: LS-Dyna, NASTRAN, PAM-Crash (ESI), HPCMW (RIST)
  - Radiative transport: 2-D SPHOT, 3-D UMT2000
  - Neutron transport :Sweep3D
  - Weather: MM5, IFS (ECMWF), WRF
- Life Sciences and Biotechnology: mpiBLAST, Smith-Waterman
- CAD/CAE: AVBP
- Crystallography with X-Ray Diffraction: Shake & Bake
- Drug Screening: OpenEye Scientific Software, Tripos, MOE (CCG: Chemical Computing Group)
- Finance: NIWS (Nissei)
- Nanophysics and Materials: LSMS
- Quantum Chromodynamics: IBM, Boston University, KEK, Jülich