BlueGene/L Supercomputer

George Chiu
IBM Research

The Blue Gene Project

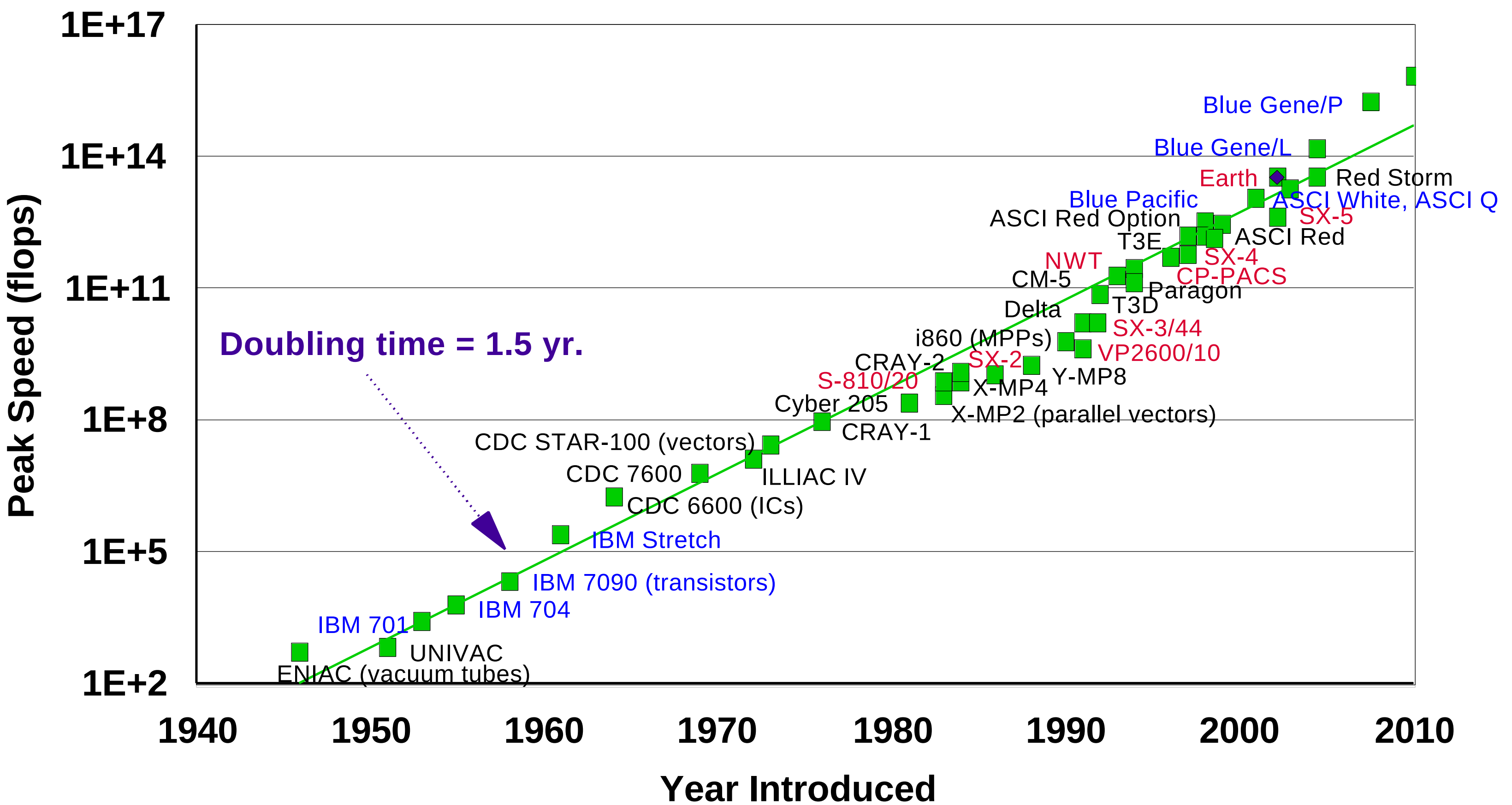
- In December 1999, IBM Research announced a 5 year, $100M US, effort to build a petaflop supercomputer to attack problems such as protein folding.
- The Blue Gene project has two primary goals:
  - Advance the state of the art in computer design and software for extremely large scale systems.
  - Advance the state of the art of biomolecular simulation.

- In November 2001, an R & D partnership with Lawrence Livermore National Laboratory was announced.
- In November 2002, a contract with Lawrence Livermore National Laboratory was signed to build two systems: ASCI Purple and BlueGene/L.

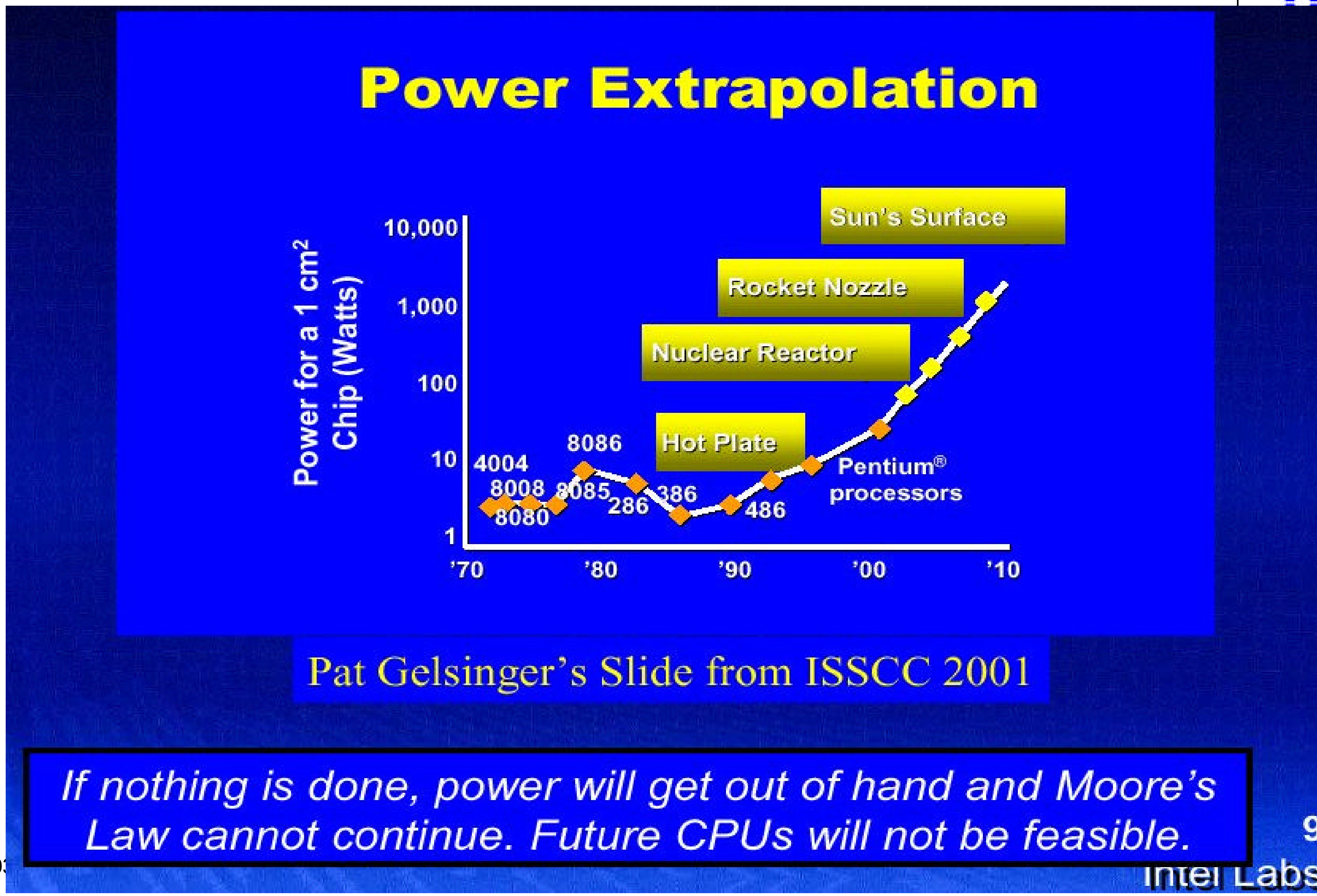
- In October 2003, a 512-way prototype achieves 1,413 GFlop/s on Linpack and permits entry onto TOP500.

- Deliverable: a 180/360 TF/s BlueGene/L system to Lawrence Livermore National Laboratory in 1H2005

Supercomputer Peak Performance



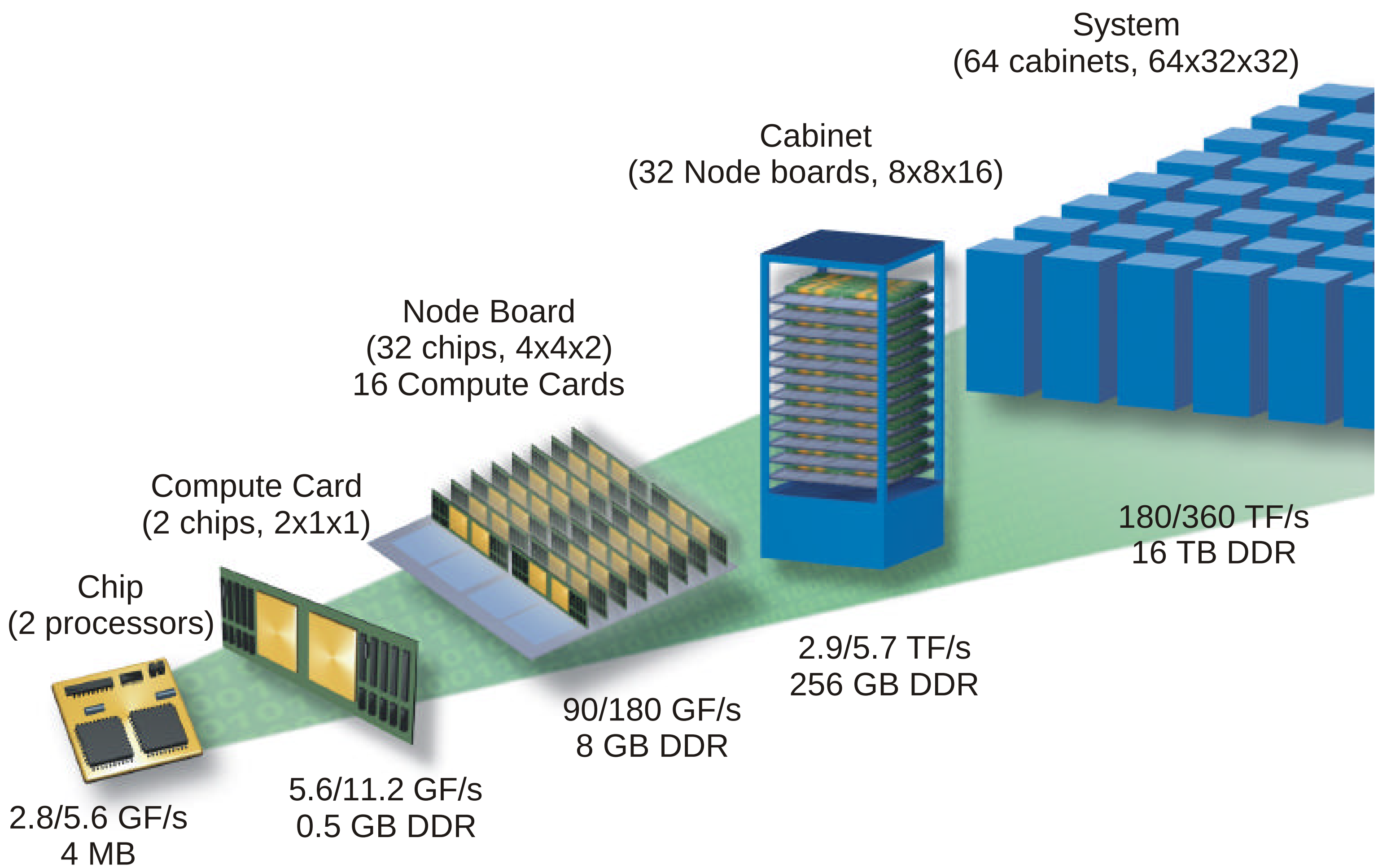
Microprocessor Power Density Growth



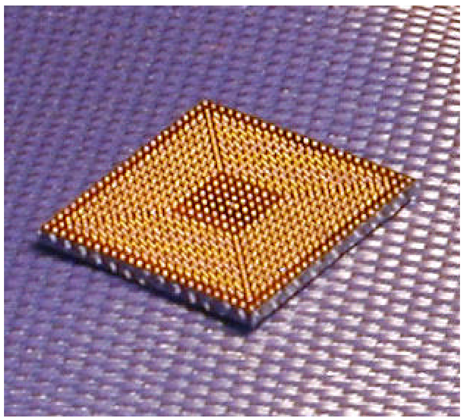
What is the Blue Gene/L Project?

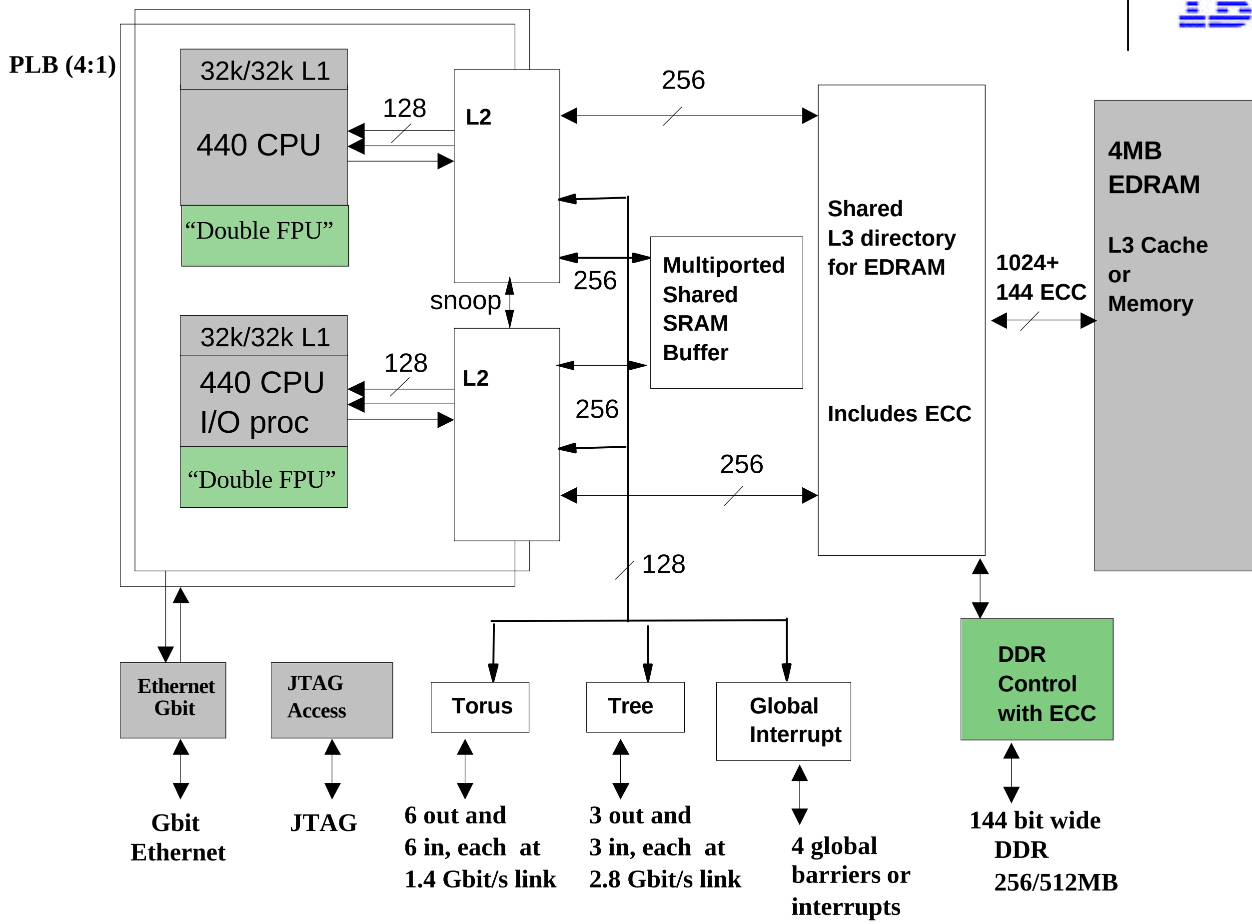
- Partnership between IBM, ASCI-Trilab and Universities to develop and build a 180-360 TFlops/s computer.
- Focus is on numerically intensive scientific problems.
- A 64k node highly integrated supercomputer based on system-on-a-chip technology. Two ASICs were designed:
  - BlueGene/L Compute (BLC)
  - BlueGene/L Link (BLL)
- LLNL contributions
  - Validation and optimization of architecture based on real applications.
  - Researchers are accustomed to "new architectures" and will work hard to adapt to constraints. Assists us in the investigation of the reach of this machine
- Grand challenge science requirements stress I/O, memory (bandwidth, size and latency), and processing power.

BlueGene/L


System
(64 cabinets, 64x32x32)
Cabinet
(32 Node boards, 8x8x16)
Node Board
(32 chips, 4x4x2)
16 Compute Cards
Compute Card
(2 chips, 2x1x1)
Chip
(2 processors)
180/360 TF/s
16 TB DDR
2.9/5.7 TF/s
256 GB DDR
90/180 GF/s
8 GB DDR
5.6/11.2 GF/s
0.5 GB DDR
2.8/5.6 GF/s
4 MB

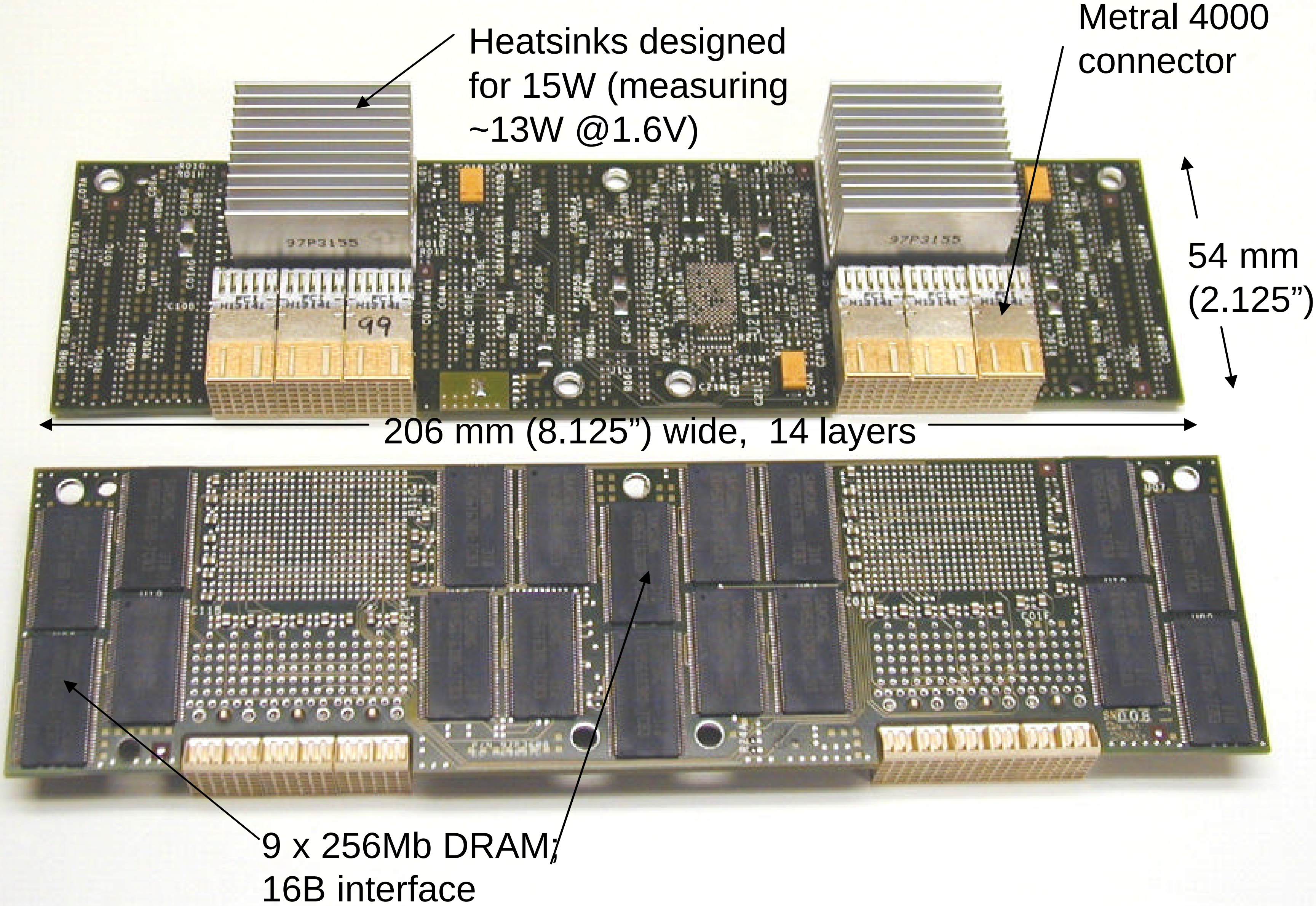
BlueGene/L Compute ASIC

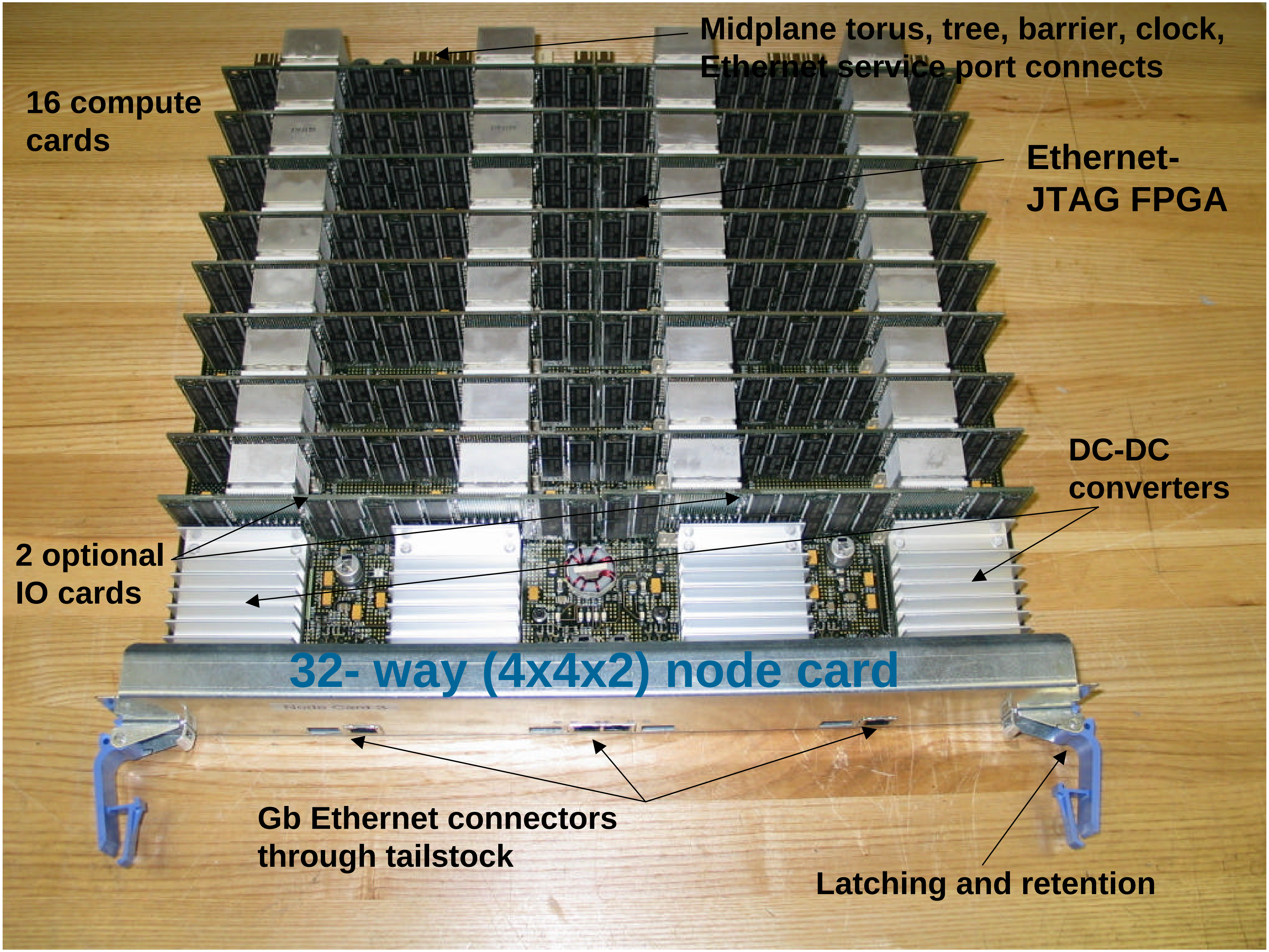


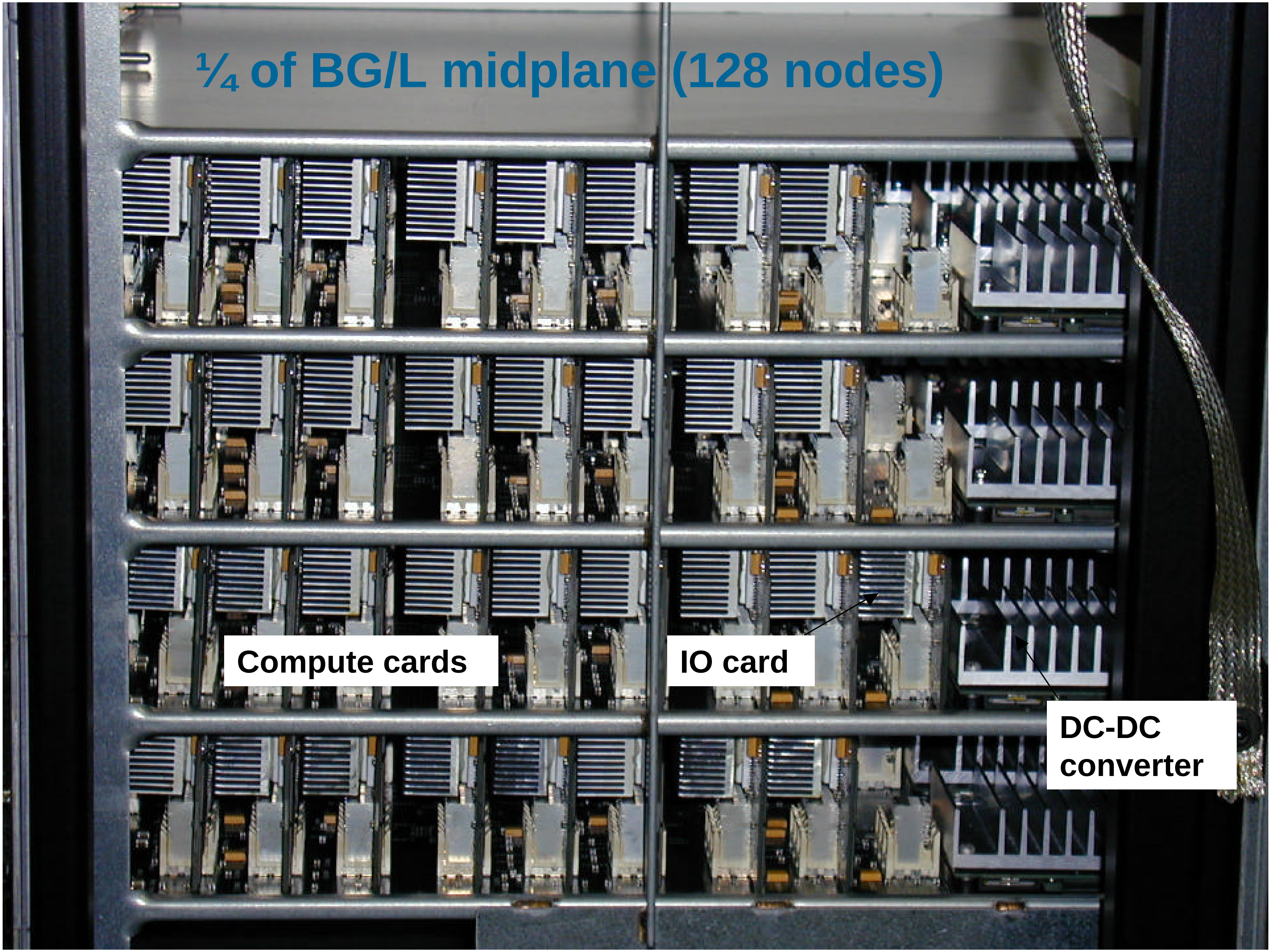


- IBM CU-11, 0.13 µm
- 11 x 11 mm die size
- 25 x 32 mm CBGA
- 474 pins, 328 signal
- 1.5/2.5 Volt

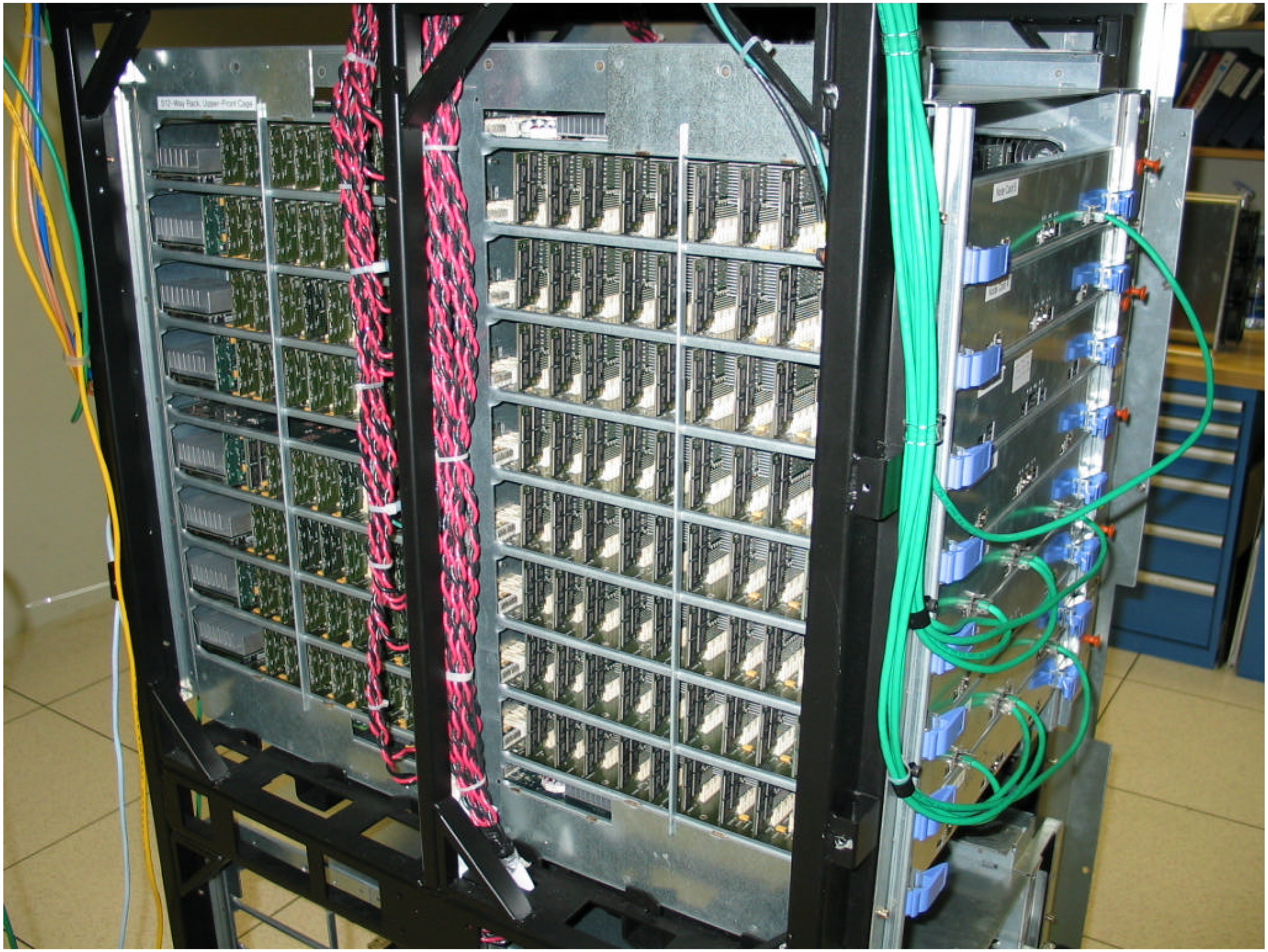
Dual Node Compute Card



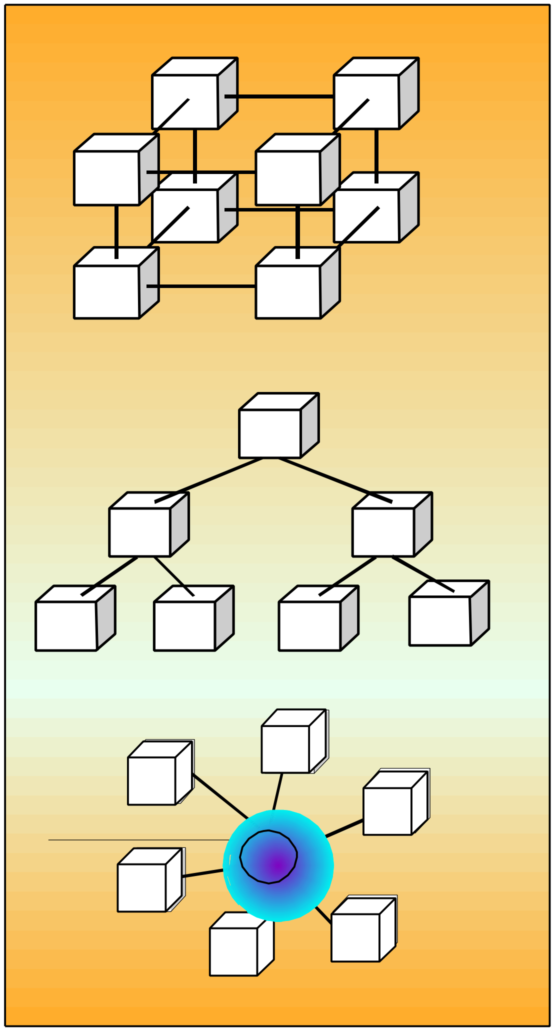

Midplane torus, tree, barrier, clock,
Ethernet service port connects
16 compute
cards
Ethernet-
JTAG FPGA
DC-DC
converters
2 optional
IO cards
32- way (4x4x2) node card
Gb Ethernet connectors
through tailstock
Latching and retention


¼ of BG/L midplane (128 nodes)
Compute cards
IO card
DC-DC converter

512 Way BG/L Prototype



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)
- Communications backbone for computations
- 0.7/1.4 Tb/s bisection bandwidth, 67TB/s total bandwidth

Global Tree

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

Ethernet

- Incorporated into every node ASIC
- Active in the I/O nodes (1:64)
- All external comm. (file I/O, control, user interaction, etc.)

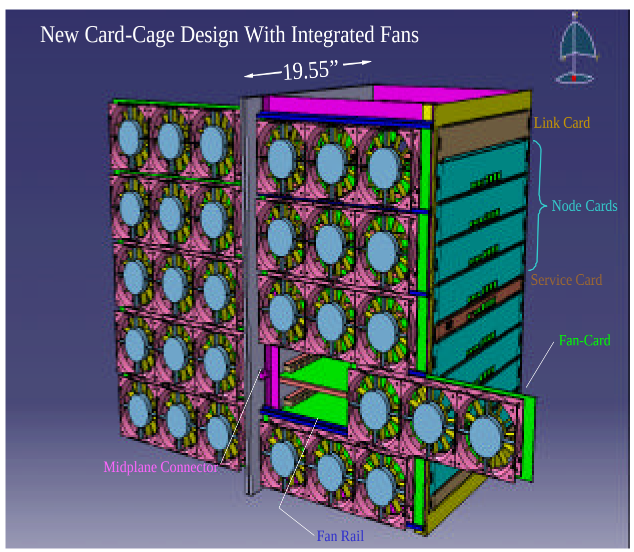
Rack, without cables

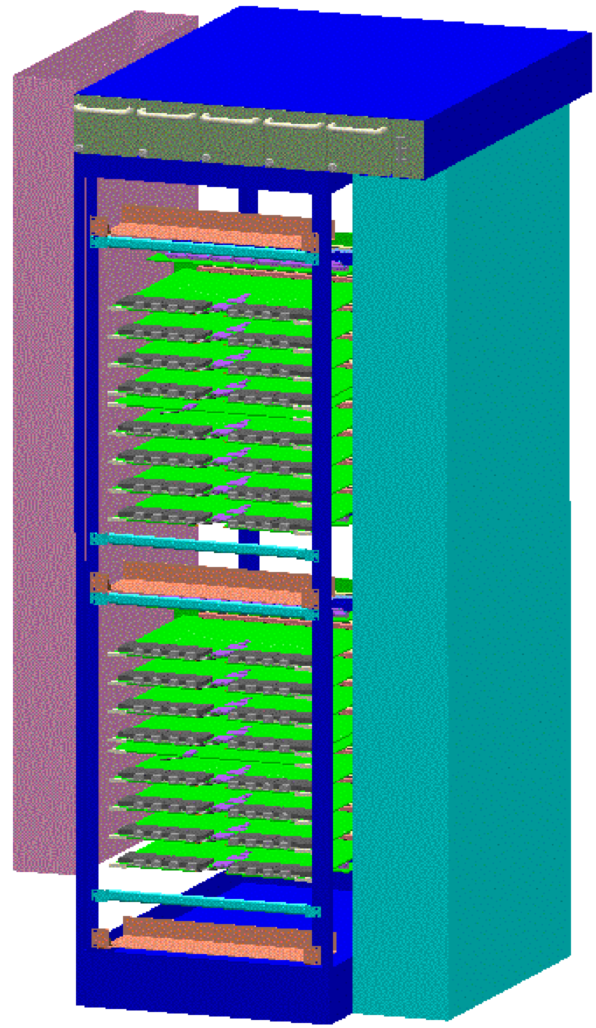
- 1024 compute nodes
  - 1024 compute proc
  - 1024 communication proc
  - 256 GB DRAM
  - 2.8TF peak

- 16 I/O nodes
  - 8 GB DRAM
  - 16 Gb Ethernet channels

- ~20KW, air cooled
  - redundant power
  - redundant fans

- ~36inch wide,
~36 inch deep,
~80 inch high (40-42U)





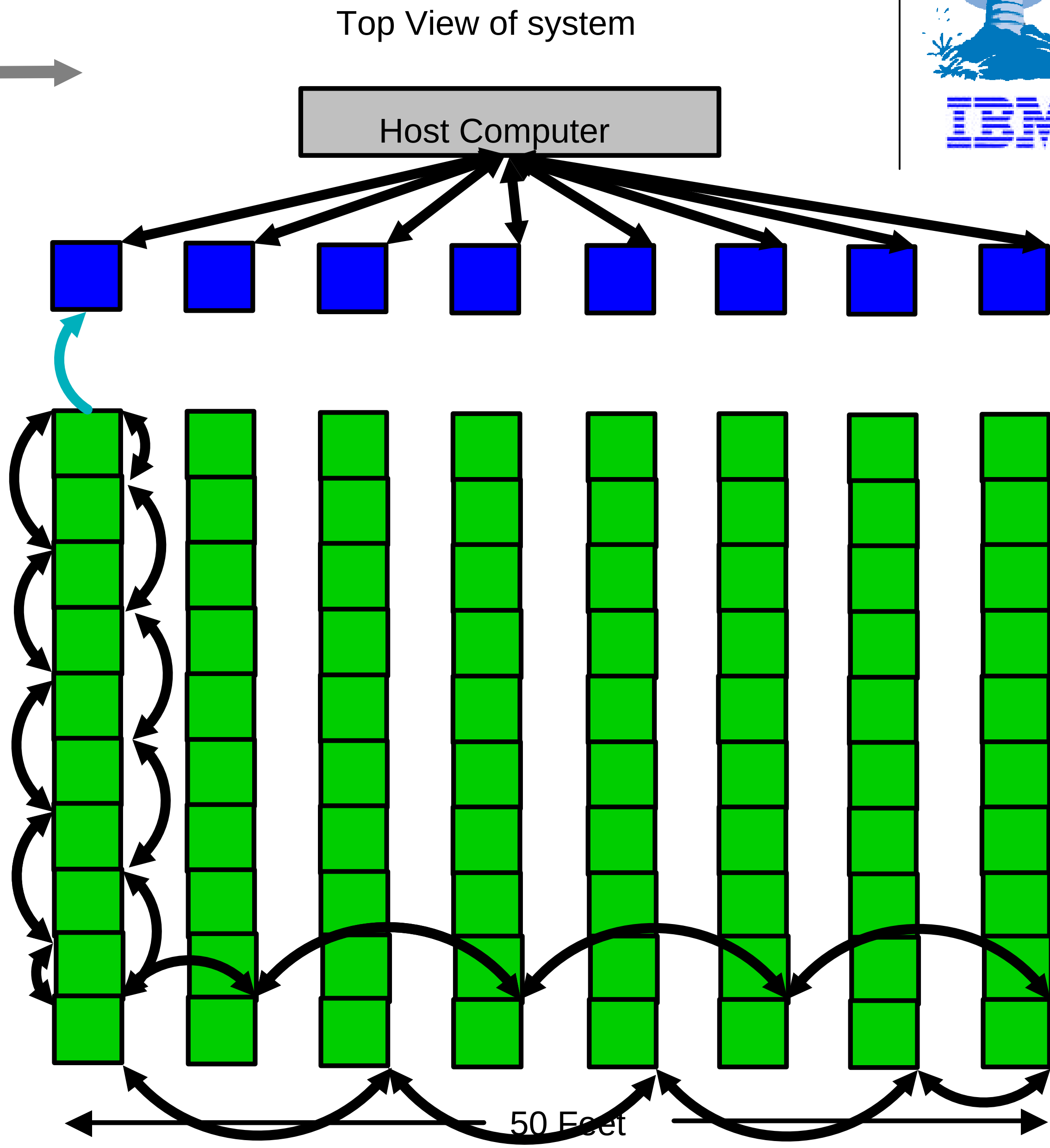
System Organization (conceptual)

Host System:
- Diagnostics, Booting, Archive
- Application dependent requirements

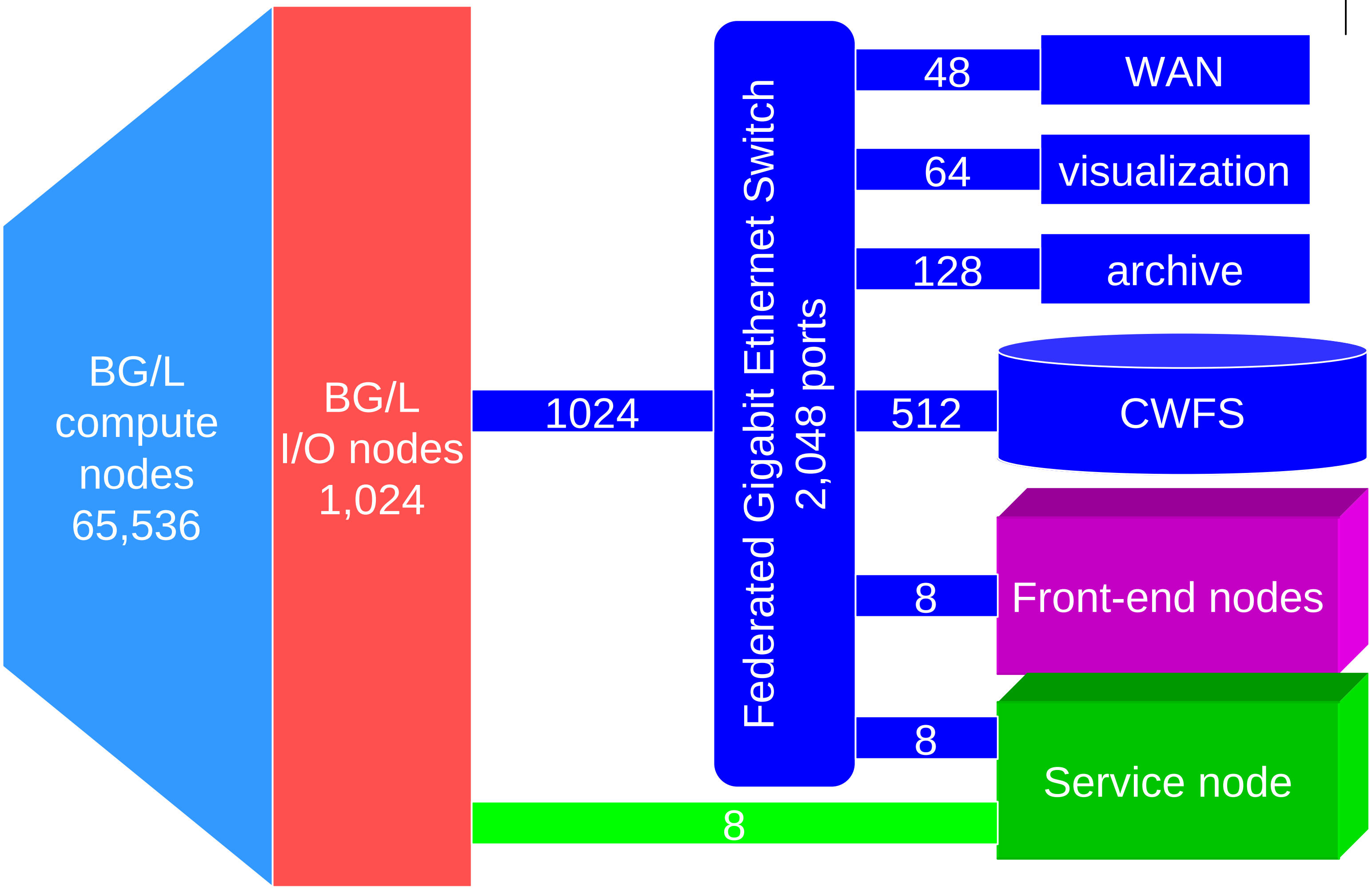
File Server Array
- ~ 500 RAID PC servers
- Gb Ethernet and/or Infiniband
- Application dependent requirements

BlueLight Processing Nodes
- 81920 Nodes

  Two major partitions
  - 65536 Nodes Production Platform (256 TFlops Peak)
  - 16384 Nodes partitioned into code development Platforms



Complete BlueGene/L System at LLNL



Software Design Summary

- **Familiar software development environment and programming models**
  - Fortran, C, C++ with MPI
    - Full language support
    - Automatic SIMD FPU exploitation
  - Linux development environment
    - User interacts with system through FE nodes running Linux – compilation, job submission, debugging
    - Compute Node Kernel provides look and feel of a Linux environment – POSIX system calls (with restrictions)
  - Tools – support for debuggers, hardware performance monitors, trace based visualization

- Scalability to *O(100,000)* processors

Summary of performance results

- DGEMM:
  - 92.3% of dual core peak on 1 node
  - Observed performance at 500 MHz: 3.7 GFlops
  - Projected performance at 700 MHz: 5.2 GFlops (tested in lab up to 650 MHz)
- LINPACK:
  - 77% of peak on 1 node
  - 69% of peak on 512 nodes (1413 GFlops at 500 MHz)
- sPPM, UMT2000:
  - Single processor performance roughly on par with POWER3 at 375 MHz
  - Tested on up to 128 nodes (also NAS Parallel Benchmarks)
- FFT:
  - Up to 508 MFlops on single processor at 444 MHz (TU Vienna)
  - Pseudo-ops performance (5N log N) @ 700 MHz of 1300 Mflops (65% of peak)
- STREAM – impressive results even at 444 MHz:
  - Tuned: Copy: 2.4 GB/s, Scale: 2.1 GB/s, Add: 1.8 GB/s, Triad: 1.9 GB/s
  - Standard: Copy: 1.2 GB/s, Scale: 1.1 GB/s, Add: 1.2 GB/s, Triad: 1.2 GB/s
  - At 700 MHz: Would beat STREAM numbers for most high end microprocessors
- MPI:
  - Latency – < 4000 cycles (5.5 μs at 700 MHz)
  - Bandwidth – full link bandwidth demonstrated on up to 6 links

Applications

- BG/L is a general purpose technical supercomputer
- N-body simulation
  - molecular dynamics (classical and quantum)
  - plasma physics
  - stellar dynamics for star clusters, galaxies
- Complex multiphysics code
  - Computational Fluid Dynamics (weather, climate, sPPM...)
  - Accretion
  - Raleigh-Jeans instability
  - planetary formation and evolution
  - radiative transport
  - Magnetohydrodynamics
- Modeling thermonuclear events in/on astrophysical objects
  - neutron stars
  - white dwarfs
  - supernovae
- Radiotelescope
- FFT

Summary

- Embedded technology promises to be an efficient path toward building massively parallel computers optimized at the system level.

- Cost/performance is ~20x better than standard methods to get to TFlops.

- Low Power is critical to achieving a dense, simple, inexpensive packaging solution.

- Blue Gene/L will have a scientific reach far beyond existing limits for a large class of important scientific problems.

- Blue Gene/L will give insight into possible future product directions.

- Blue Gene/L hardware will be quite flexible. A mature, sophisticated software environment needs to be developed to really determine the reach (both scientific and commercial) of this architecture.