# Blue Gene/L

- The fastest supercomputer as of 2006
- A typical supercomputer consumes 20 MW of power
  - 20,000 houses
- Goals
  - higher performance with an improved power/performance ratio,
  - scalable system,
  - efficient utilization of distributed memory,
  - for specific classes of applications (large-scale simulation of physical phenomena, real-time data processing, offline data analysis)

# Blue Gene/L

- Design

Low-power → reduced complexity → simplified the design, verification, etc.
→ enabled a dense, efficient packaging design

1024 dual-processor nodes in a single rack of 0.9 (w) x 0.9 (d) x 1.9 (h) m^3;
→ (1) a rack consumes 27.5 kW
→ (2) 85% of inter-node connectivity is contained within the racks
→ reduces connectivity across racks dramatically
→ higher density, higher reliability, better manageability

Scalability, efficient utilization of distributed memory → message passing system

Certain classes of applications → ASIC level design → able to integrate many features of high performance servers

# Blue Gene/L

- Scaling (Application)

"Weak scaling:" the domain for each node is fixed as the number of nodes increases
(Gustafson's law)
Load imbalance and global communication determine scalability
(load imbalance is not a hardware issue)
→ two additional networks
Collective network: low latency global communication
Global barrier network: extremely low latency in barriers and notification through combinational logic

"Strong scaling:" the domain for each node decreases as the number of nodes increases
(Amdahl's law)
"Surface-to-volume" effect → requires efficient boundary information exchange
→ 3-D mesh (torus) topology

# Blue Gene/L

- System Configuration

Nodes
Up to 65,536 nodes (64 racks)
Each node includes 2 IBM PowerPC 440's and 512 MB with 3 levels of caches
512 nodes on a double-sided board ("*midplane*") 8x8x8

Link chip
6 ports
2 ports (in and out) directly connected to nodes in a midplane and 4 ports for reconfigurable connections to nodes in other midplanes
16 unidirectional links (channels) supported by each port
→ Each midplane is a 8x8x8 cube which has 64 nodes on each of the 6 faces
→ 64x6 = 384 face nodes are supported by 24 link chips (384/16=24)

# Blue Gene/L

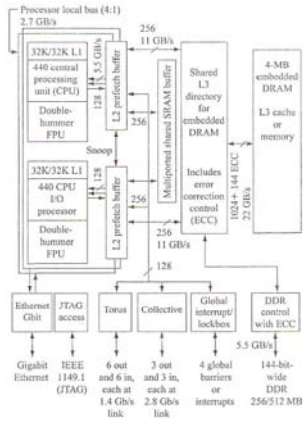



Figure 5

Blue Gene/L compute (BLC) chip architecture. Green shading indicates off-the-shelf cores. ©2002 IEEE. Reprinted with permission from G. Almasi et al., "Cellular Supercomputing with System-on-a-Chip," *Digest of Technical Papers*, 2002 IEEE International Solid-State Circuits Conference.

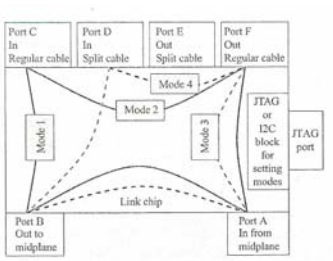



Figure 2

Blue Gene/L link chip switch function. Four different modes of using the link chip.

# Blue Gene/L

- Interconnection Networks

  5 networks

  3-D torus
  - Main interconnection network
  - 8x8x8 fixed 3-D mesh within each midplane
  - Link chips among midplanes
    - → a full configuration of 64 K nodes: 64x32x32 3-D torus
  - Each node has 6 bidirectional near-neighbor links (aggregate BW: 2.1 GB/s)
  - 100 ns latency per hop
  - Worst case latency: 32 + 16 + 16 (network diameter) = 64 hops → 6.4 *u*s

# Blue Gene/L

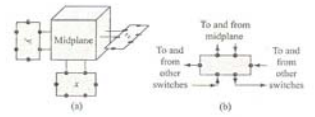



Figure 1

A midplane and its switches: (a) the three switches; (b) switch I/O ports.

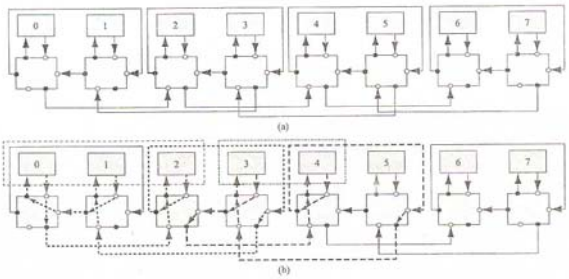



Figure 2

(a) Blue Gene/L *x*-line. (b) Two toroidal partitions in a single *x*-line.

# Blue Gene/L

Collective network
- Global communication such as broadcasting, multi-casting, reduction (summing up local data,e.g.), etc.
- An independent collective network is formed for each partition

Barrier network
- Global “AND” for global barrier
- Global “OR” for global interrupt

Control system network
- Monitor/control temperature sensor, clock tree, fans, power supplies, etc.

Gigabit Ethernet
- In addition to compute nodes, there are I/O nodes with a maximum I/O to compute node ratio of 1:8.
- I/O nodes to file systems

# Blue Gene/L

- Resource allocation (system partitioning)

Trade-off between granularity and manageability
→ partition at the level of midplane

Two-phase partition allocation (given a request)
(1) Search for all 3-D rectangular and contiguous sets of free midplanes that match the request
(2) For each set, search for an available set of free links to connect them in a mesh or torus according to the request

(Link search in a dimension is independent of that in another dimension since the set of links for a dimension is independent of that for others.)

Choose the "best" partition according to the merit criteria

# Blue Gene/L

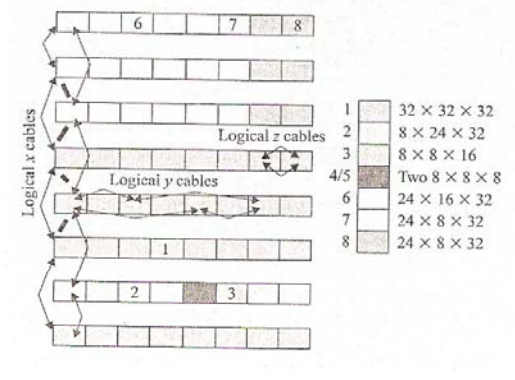



Figure 3

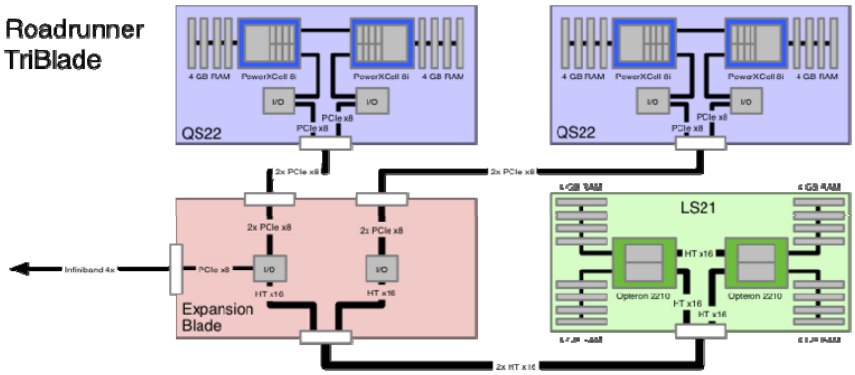
A 64-rack Blue Gene/L space-partitioned for eight users.

# IBM Roadrunner

- The first (parallel) computer system breaking the Petaflop barrier
  - 1.105 petaflops
  - 6,480 tri-blades
    - Each tri-blade: 2 IBM QS22 blade servers (Cell) and 1 IBM LS21 server (AMD Opteron)
    - 9 CPU's/Cell, 2 CPU's/Opteron → 20 CPU's/tri-blade
      → 129,600 CPU's
    - Cell: Numerical and CPU-intensive computation
    - AMD Opteron: Standard processing (file system I/O)
  - 103 tera bytes of memory
  - 2.48 megawatts

# IBM Roadrunner

A tri-blade logically consists of 4 Cell's and 2 AMD Opterons.



# IBM Roadrunner

Roadrunner, tiered architecture