High Performance Computing: Blue-Gene and Road Runner

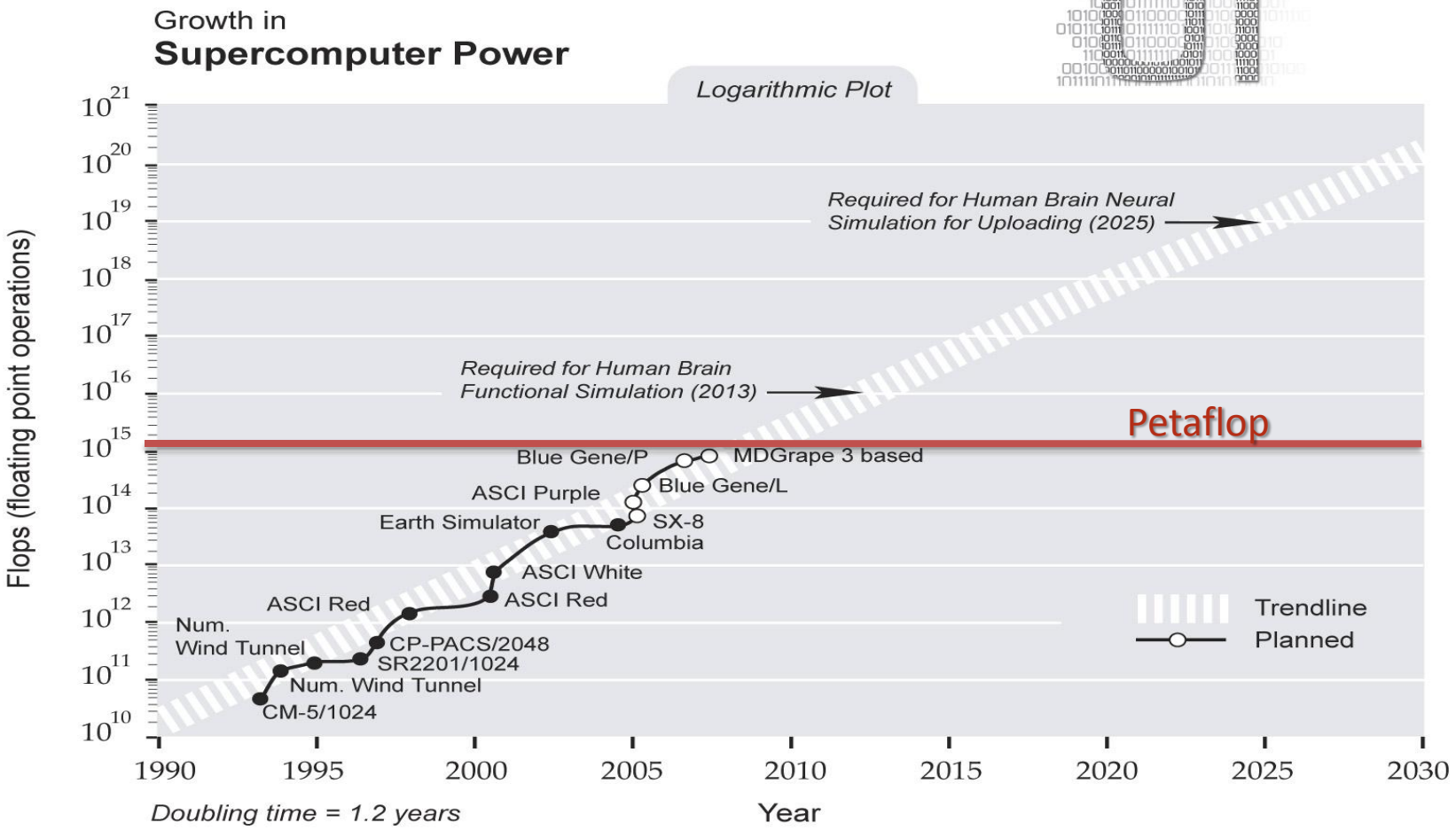
Ravi Patel

HPC General Information

HPC Considerations

- Criterion
  - Performance
    - Speed
    - Power
  - Scalability
    - Number of nodes
    - Latency bottlenecks
  - Reliability (boring)
    - Fault tolerance
    - Fault Isolation

HPC Performance History

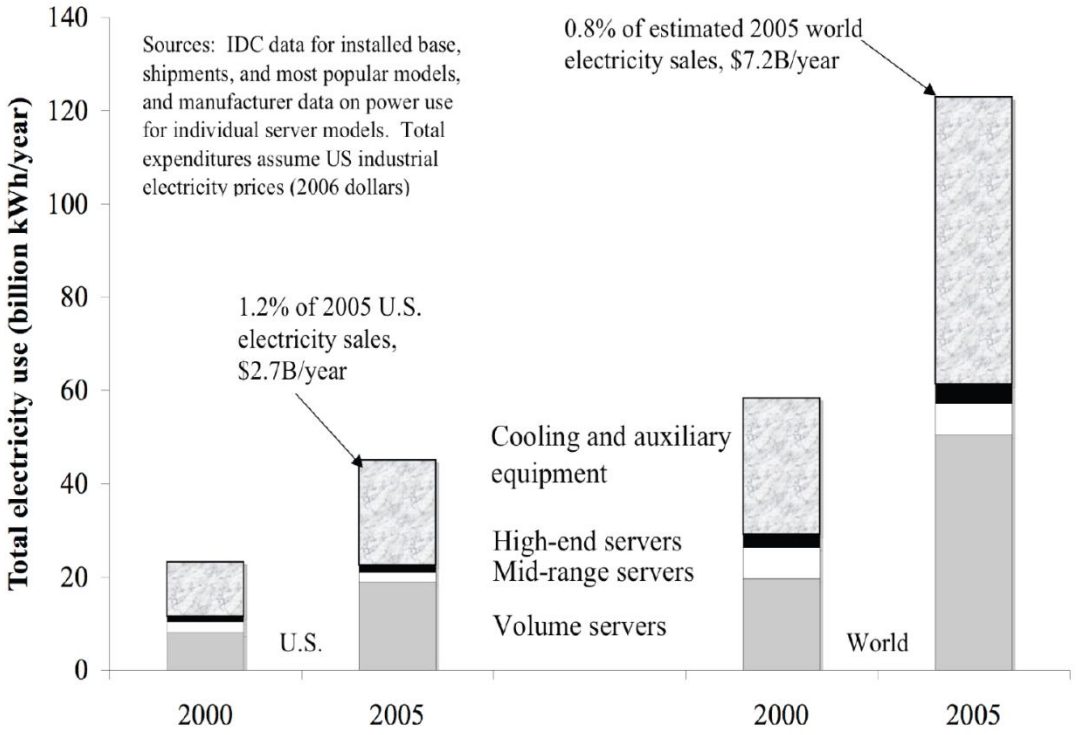


Kurzweil, Ray. "Growth of Supercomputer Power" 2008 .URL: http://www.singularity.com/charts/page71.html

Server Power History

- Acquired data
  - Server power usage
  - Server cooling and auxiliaries
- Unaccounted
  - Data storage power
  - Network equipment
  - Cooling and auxiliaries for each
  - 20 to 40 % of datacenter load

Power Consumption for Servers



J.G. Koomey, "Estimating Total Power Consumption by Servers in the U.S. and the World"; http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf.

HPC Power

- Study accounts for all servers
  - Not HPC
  - Cheap servers have worse power performance
- Importance of low power design
  - Environmental management costs
    - Limits scalability
    - Limits extensibility

HPC Scaling

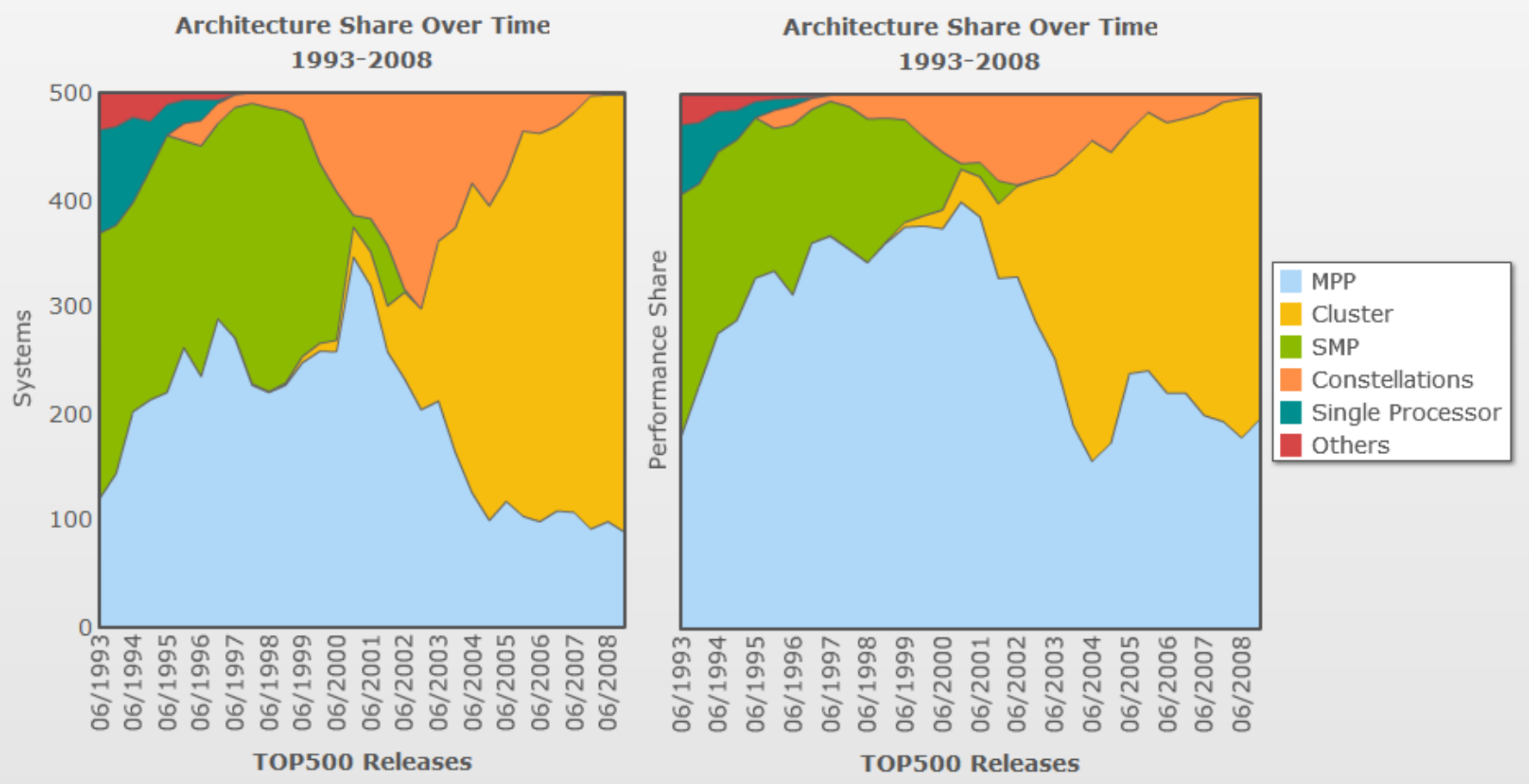
- Weak Scaling:
  - Global Problem size grows with number of nodes
- Strong Scaling:
  - Global Problem size remains fixed with number of nodes

Application scaling behavior	*Scaling limitations*							
	Amdahl	*Problem segmentation limits*	*Surface-to-volume communication dominates*	*Load imbalance*	*Small messages*	*Global communication dominates*	*Memory footprint*	*File I/O*
Strong scaling	✓		✓	✓	✓	✓	✓	✓
Weak scaling	✓	✓		✓		✓	✓	✓

HPC Architecture

- MPP: Massively Parallel Processor
  - Blue Gene and Roadrunner
- Cluster: Group of autonomous systems
  - Rely on general network structure
- SMP: Single Massive System
  - Looks like one giant system
- Constellation: Sun HPC System Architecture

HPC Architecture



[1] TOP500 Supercomputer List – Dongarra, Meuer, et al. – November 2008 URL: http://www.top500.org

HPC History

Table 1. Platform configurations.

Characteristic	SGI Altix	Alpha SC	Earth Simulator	IBM SP4	Cray X1
Processor	Intel Itanium 2	Compaq Alpha EV67	NEC SX-6	IBM Power4	Cray X1
Interconnect	Numalink	Quadrics (Elan3)	Custom crossbar	HPS or SP Switch2	Cray X1
Processor speed (MHz)	1,500	667	500	1,300	800
Memory/node (Gbytes)	512	2	16	32	16
L1 cache size (Kbytes)	32	64	NA	32	16 (scalar)
L2 cache size (Mbytes)	0.256	8	NA	1.5	2 (per MSP)
L3 cache size (Mbytes)	6	NA	NA	128	NA
Processor peak performance (Mflops)	6,000	1,334	8,000	5,200	12,800
Peak memory bandwidth (Gbytes/s)	6.4	5.2	32 (per each processor)	51 (per mutichip modules)	26 (per each MSP)

- Performance
  - Add more processors
    - Use Vector Processors (VLIW)

T.H. Dunigan, Jr., J.S. Vetter et al., Performance Evaluation of the Cray X1 Distributed Shared Memory Architecture, IEEE Micro, 25(1):30-40, 2005.

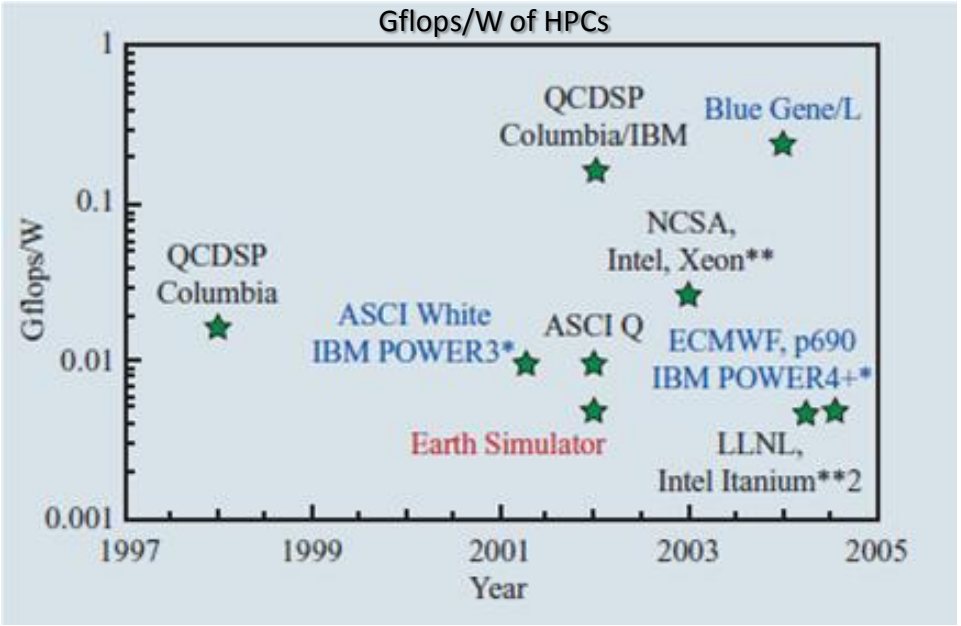
HPC History

- Vector Processors:
  - Secondary specialized floating point processors
- Interconnect:
  - Shifted from crossbar designs to meshes and trees
    - Crossbars have better bisection bandwidth
    - More complex

Blue Gene/L

Blue Gene/L

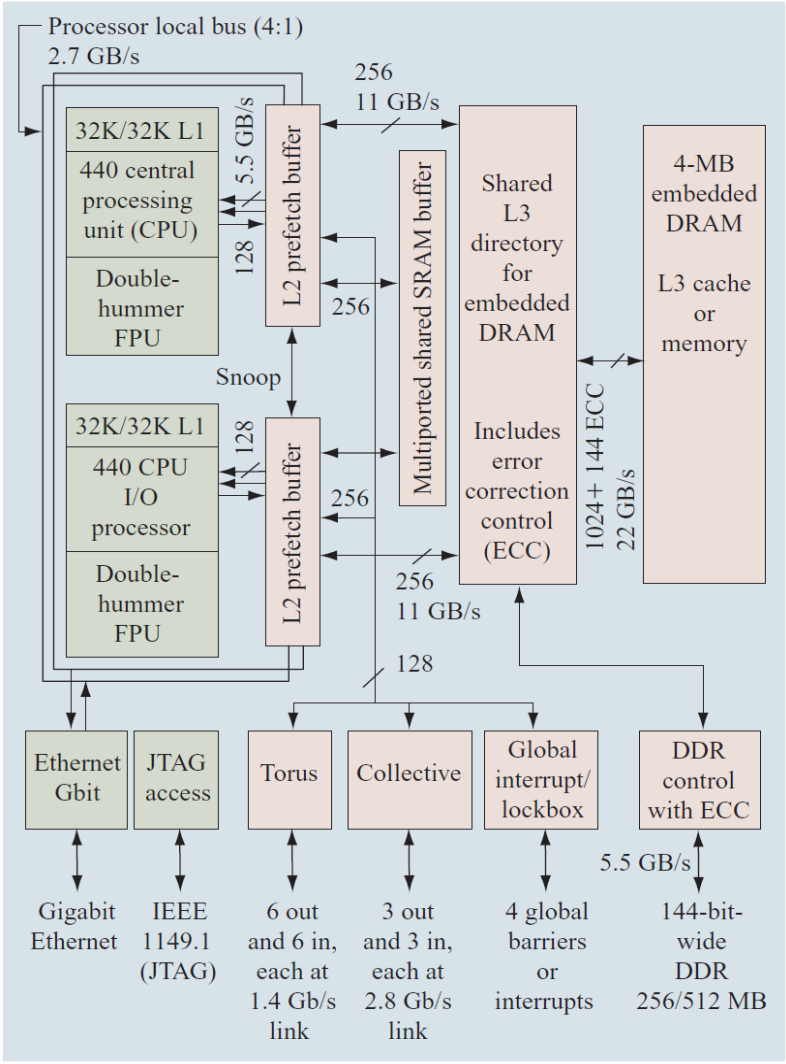
- Design Goals:
  - Low power
  - High performance from scalability
    - 65,000+ nodes
    - Parallelism over speed
  - Small footprint



A. Gara, M. A. Blumrich, D. Chen, G. L.-T. Chiu, P. Coteus, M. E. Giampapa, R. A. Haring, P. Heidelberger, D. Hoenicke, G. V. Kopcsay, T. A. Liebsch, M. Ohmacht, B. D. Steinmacher-Burow, T. Takken, and P. Vranas. Overview of the Blue Gene/L System Architecture. IBM Journal of Research and Development, 49(2/3):195–212, 2005.

Blue Gene: Overview: Node

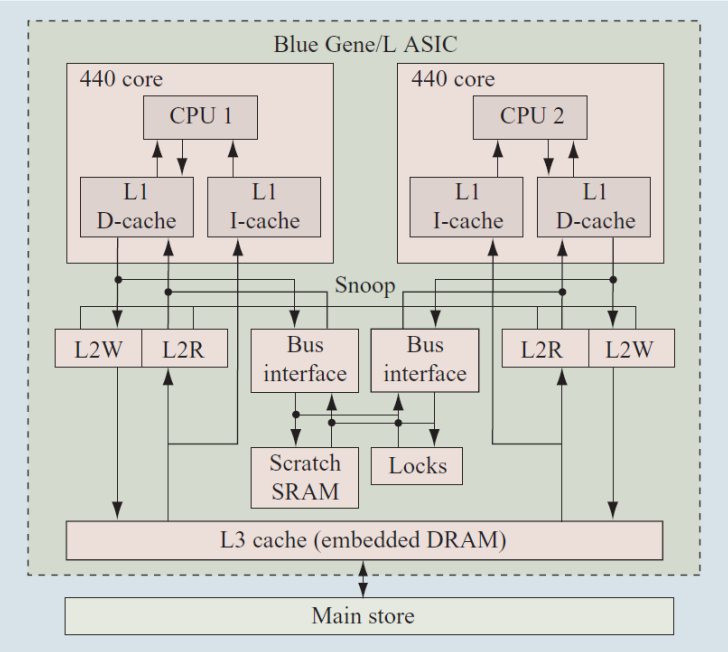
- ASIC
  - Dual Core PPC 440
    - 700 MHz clock
  - 4 MB Embedded DRAM
  - Contains all functions
  - 1 Watt operation
  - 4 Floating point units
    - Separate Instructions
- 512 MB memory
- Two node types
  - Compute or I/O



Blue Gene: Memory Hierarchy

- Slower Processor leads to flatter memory hierarchy

Attribute	*L1*	*L2*	*L3 embedded DRAM*	*Scratch SRAM*	*Main memory*
Size	32 KiB (I) 32 KiB (D) per processor	2 KiB per processor	2 banks of 2 MiB/bank = 4 MiB total shared by both processors	16 KiB shared by both processors	512 MiB shared by both processors
Latency (pclk)	3	11	28/36/40 (hit/miss precharged/missed busy)	15	86 (L3 cache enabled)
Sustained bandwidth: random quad load access (B/pclk)	NA	NA	1.8/1.2 (hit/miss)	2.0	0.8/0.5 (single/dual processor)
Sustained bandwidth: sequential access (B/pclk)	16.0	5.3	5.3/5.3 (hit/miss)	5.3	5.1/3.4 (single/dual processor)
Line width (B)	32	128	128		
Number of lines	1,024	16	32,768		
Coherent	No	Yes (weakly)	Yes	Yes (weakly)	Yes
Associativity	64 way	Fully associative	8 way/bank 2 banks	NA	NA

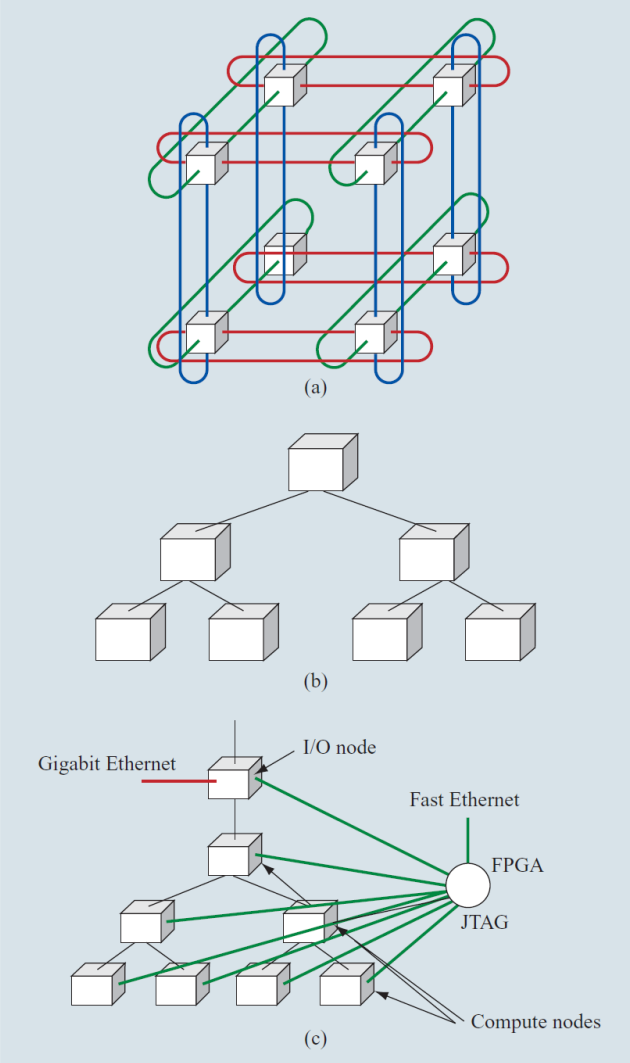


Blue Gene: Coherence

- Software managed coherence
  - Communication Coprocessor Mode
    - One core acts to control messaging
  - Virtual Node Mode
    - Divides memory in half
    - Allows read only access to other processor
- L2, L3 are sequentially consistent
- Caches are not inclusive

Blue Gene: Interconnect

- Main 3D Torus
  - Point to point/multicast
  - 100 ns hop (worst case 64)
- Collective network
  - Global Reduction operations
  - 5 µs
- Barrier Network
  - Global barriers/interrupts
  - 1.5 µs
- Service Networks
  - Ethernet
  - JTAG
  - Interfaced through I/O Nodes



Misc. System Characteristics

- Programming Environment
  - C/C++, Fortran
  - Optimized MPI
- Reliability
  - Simple components reduce complexity
  - Multilevel Redundancy
  - ECC memory (DDSC)
  - CRC protected transfers

Roadrunner

Roadrunner

- 3,060 compute nodes
  - 12,240 IBM PowerXCell 8i
  - 12,240 AMD Opteron cores
- LINPAC Performance:
  - 1.38 Pflops/s peak (1.026 sustained)
  - 536 Mflops/(s-Watt)

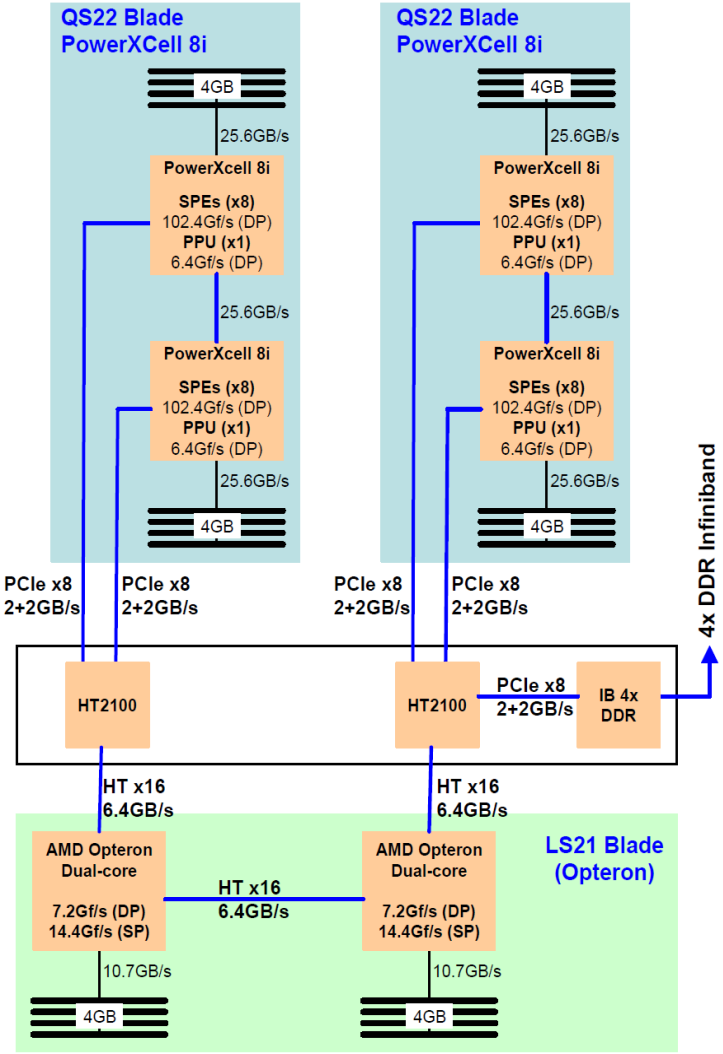
Kevin J. Barker , Kei Davis , Adolfy Hoisie , Darren J. Kerbyson , Mike Lang , Scott Pakin , Jose C. Sancho, Entering the petaflop era: the architecture and performance of Roadrunner, Proceedings of the 2008 ACM/IEEE conference on Supercomputing, November 15-21, 2008, Austin, Texas [doi>10.1145/1413370.1413372]

Roadrunner: Design Goals

- Paradigms:
  - Heterogeneous multiprocessing
    - Use accelerators for computation
  - Modes of operation
    - Support unmodified code from HPC environments
    - Accelerate performance hotspots
      - Accelerator model (push computation down)
    - Run all computational instructions on accelerators
      - SPE-Centric Model (push communication up)

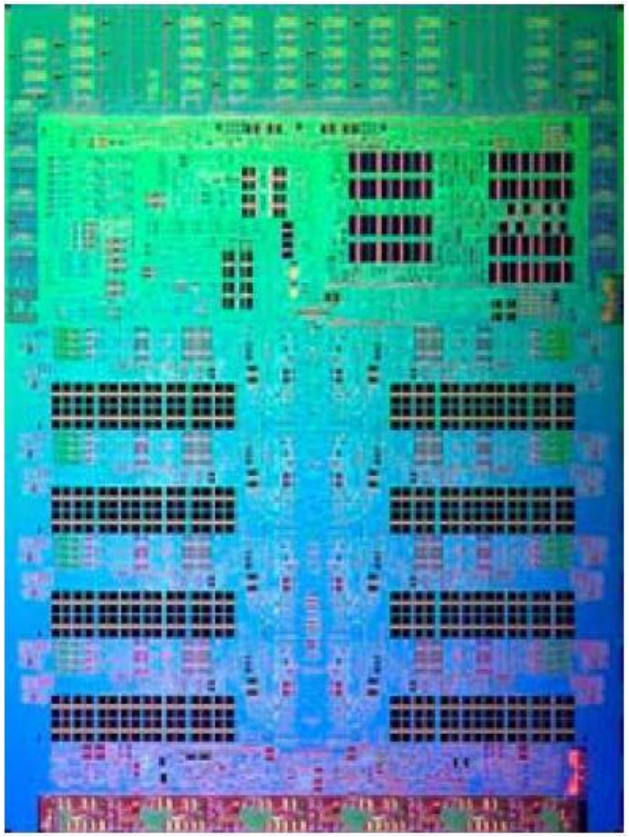
Roadrunner: Node

- Blades
  - Contain processor
  - 4 GB of memory per core
- Triblade (system node)
  - One Dual-Core Opteron (1.8 GHz) blade
  - Two Dual PowerXCell (3.2 GHz) blades



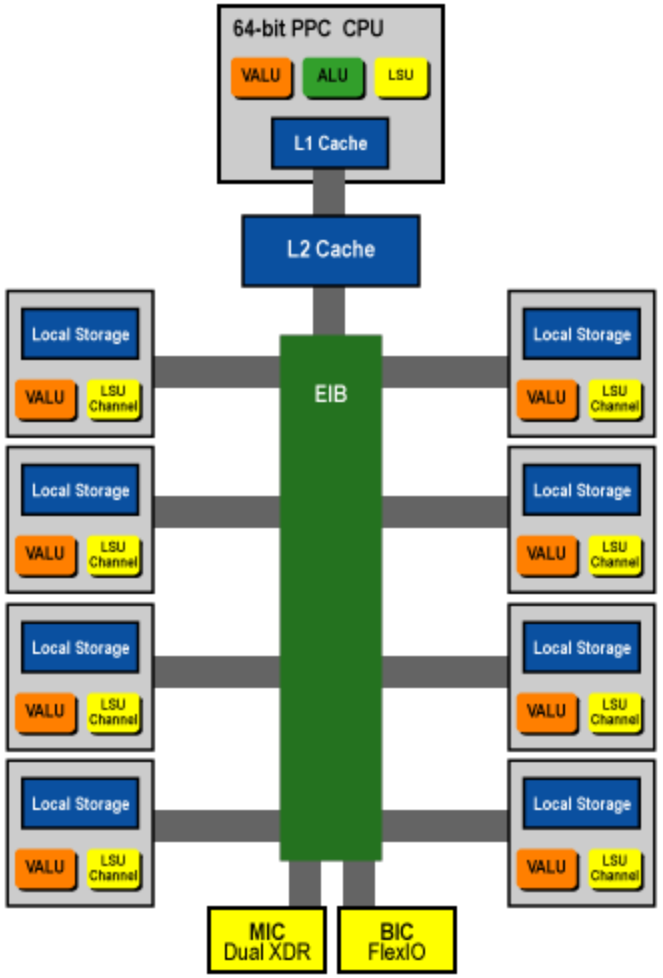
Roadrunner: PowerXCell 8i

- Cell Broadband Engine based
  - Single PPC Core
    - Handles general communication functions
  - 8 Synergistic Processing Elements
    - Specialized for floating point computation
  - Improvements over Cell
    - Improved DP floating point (108.8 Gflop/s)
    - DDR 2 support (up to 32 GB)
    - 3.2 GHz clock frequency



Roadrunner: PowerXCell Cont'd

- SPE
  - SIMD unit that can issue 4 DP operations per cycle
  - 256 KB Local store Memory
    - Only addressable memory
    - DMA transfers to access main memory
- Interconnected by Element Interconnection Bus (EIB)
  - 1.6 GHz
  - Handles 3 operations per concurrently



The CELL Architecture

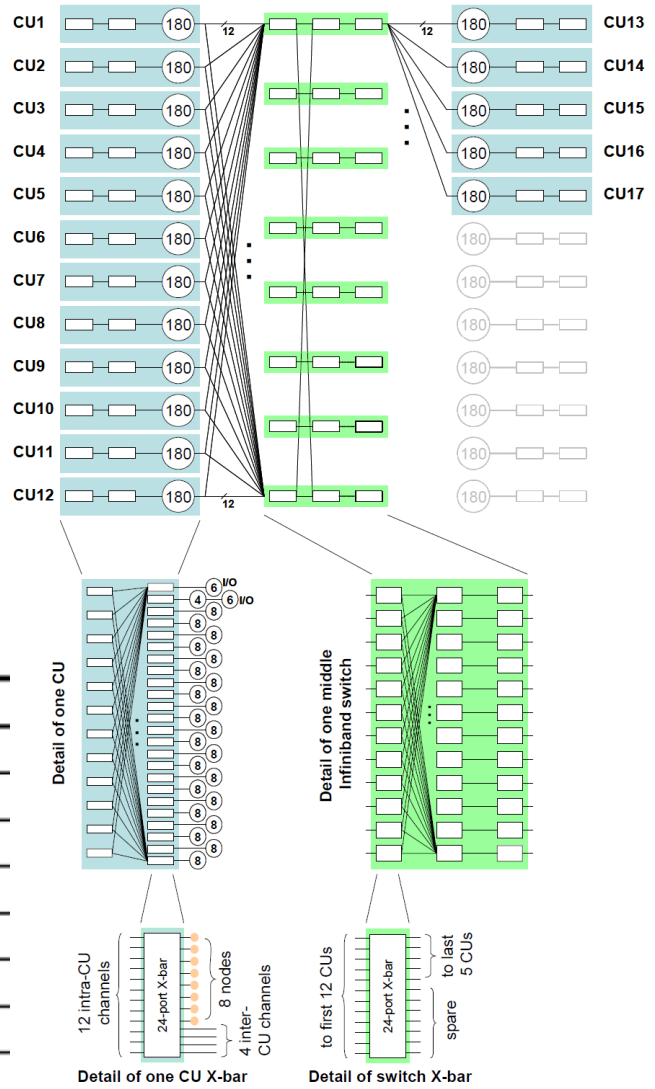
Road Runner: Compute Unit (Level 0 Network Hierarchy)

- 180 triblades
- 12 I/O nodes
- 288 port switch
  - 192 intraCU connections
  - 96 interCU connections
  - 2 GBps bandwidth each

Roadrunner: Network Topology

- Fat-Tree hierarchy of crossbars
  - Level 1: 3 crossbar plains
- Leads to predictable and small latencies

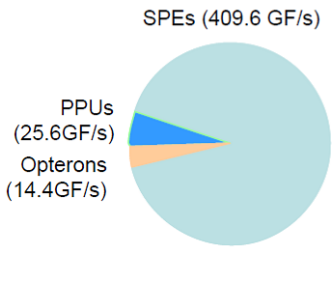
Destination node	No. of destinations	Hop count
Self	1	0
Within same crossbar	7	1
Within same CU	172	3
In CUs 2-12, same crossbar	88	3
In CUs 2-12, different crossbar	1892	5
In CUs 13-17, same crossbar	40	5
In CUs 13-17, different crossbar	860	7
Total	3060	5.38 (average)



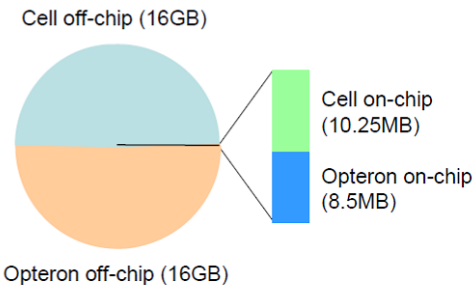
Roadrunner: Summary of Performance Characteristics

- Things to note:
  - Still Top50 network if all PowerXCell blades not used
  - PowerXCell is primary Computational engine

System		
CU count	17	
Node count	3,060	
Peak Performance (DP)	1.38 Pflops/s	
(SP)	2.91 Pflops/s	
Connected Unit (CU)		
Node count	180	
Peak performance / CU (DP)	80.9 Tflops	
(SP)	171.1 Tflops	
Compute Node (triblade)	*1x Opteron blade*	*2x Cell blades*
Processor count	2	4
Processor-core count	4	4 PPEs, 32 SPEs
Clock Speed	1.8 GHz	3.2 GHz
Peak-performance/node (DP)	14.4 Gflops/s	435.2 Gflops/s
(SP)	28.8 Gflops.s	921.6 Gflops/s
Memory per processor	4 GB (667MHz DDR2)	4 GB (800MHz DDR2)



(a) Peak processing rate (DP)



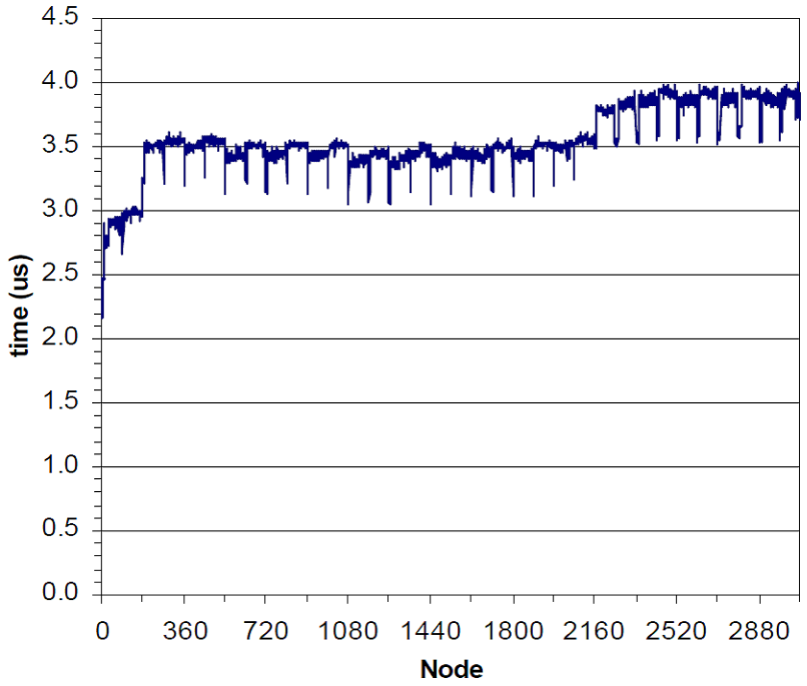
(b) Memory capacity

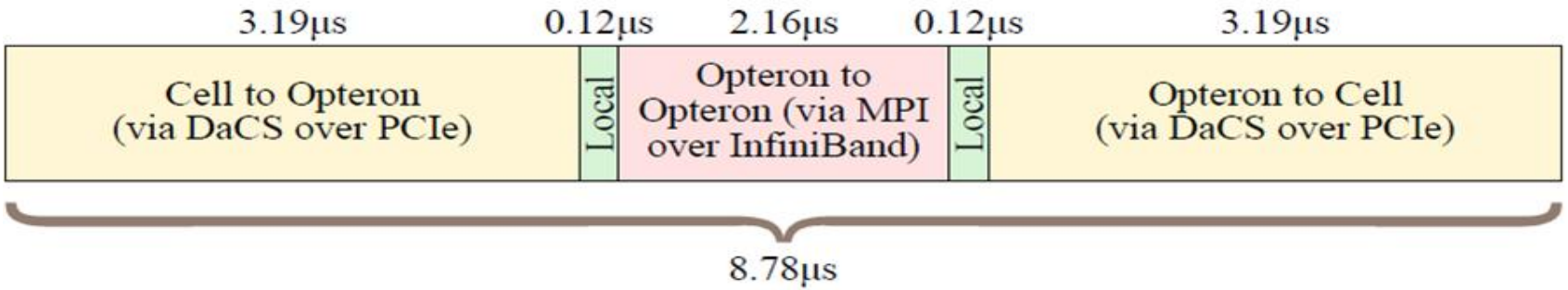
Roadrunner

Observed Performance

Roadrunner: Memory Performance

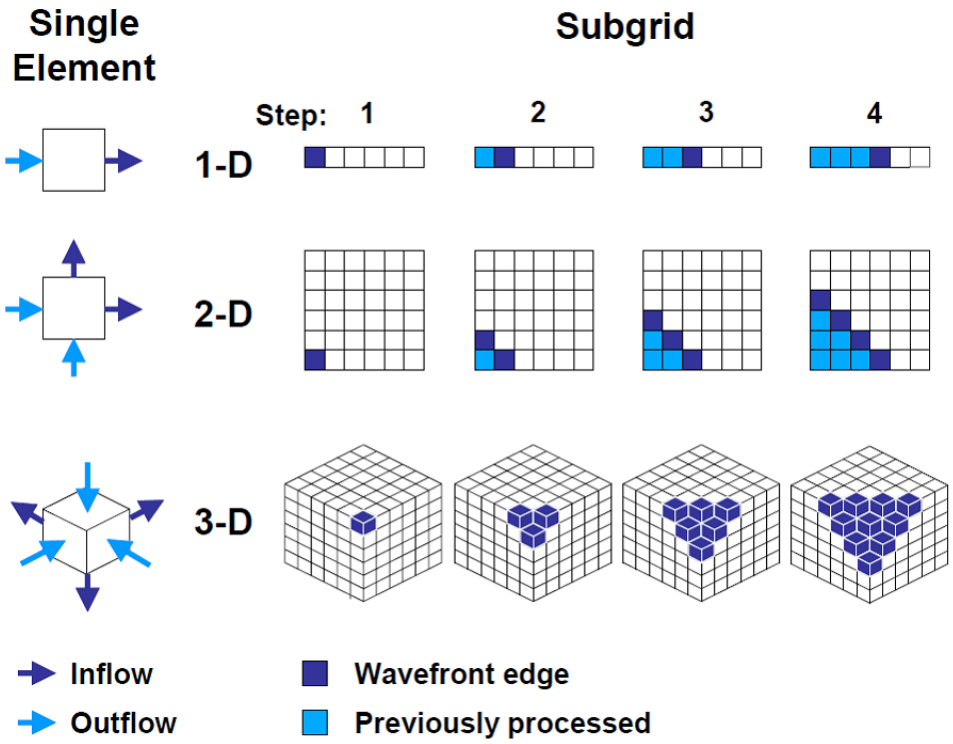
- Ping pong test
  - Observed latency matches
  - Dips caused by intra-crossbar messages every 90 nodes





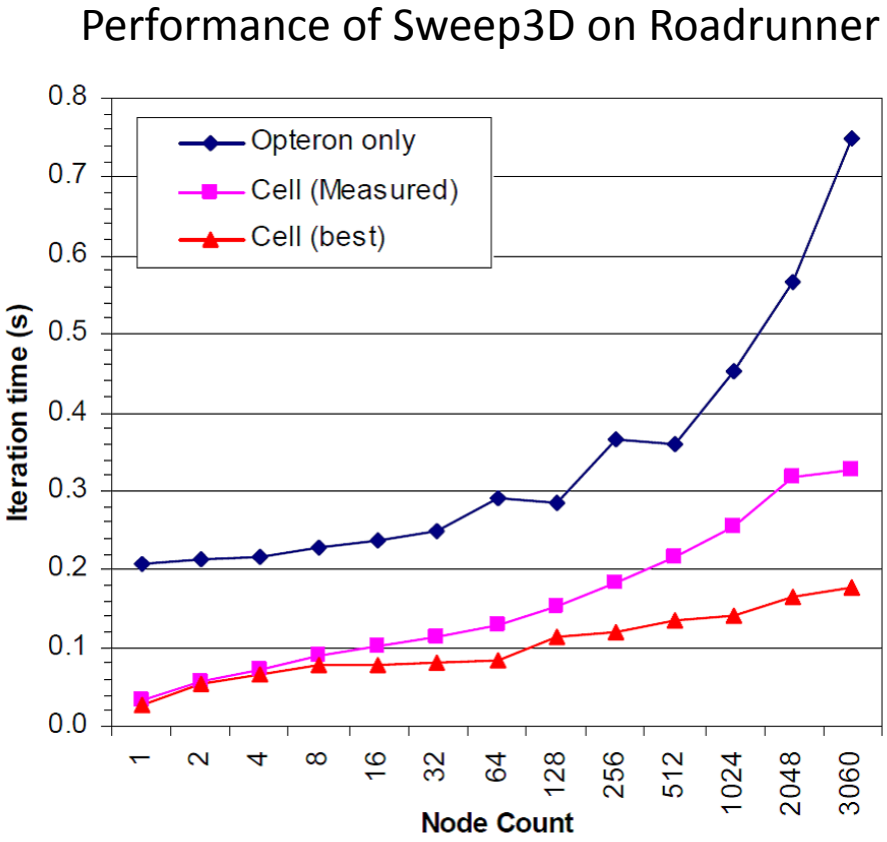
Roadrunner: Application Performance

- Sweep3D
  - Neutron transfer problem
  - Weak scaling
  - SPE-Centric Programming Mode
    - 5x5x400 unit



Performance Cont'd

- Notable:
  - Predicted best is based on making improvements to DaCS communication library



Misc System Characteristics

- MPI based programming
  - Cell Messaging Library (subset)
    - Supports RPC operations which SPEs cannot do (malloc)
- Coherence not a problem
  - Individual PPC/Opteron cores take care of themselves
    - SPEs cannot be directly addressed

Roadrunner: Summary

- Operating paradigms:
  - Accelerator
  - SPE-centric
  - Simple operation
    - Discussion: allows for quicker porting
- Fat Tree like network
  - Tree of crossbars

- It’s good to be the king

Top500 [1]

Rank	Site	Computer
1	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz , Voltaire Infiniband IBM
2	Oak Ridge National Laboratory United States	Jaguar - Cray XT5 QC 2.3 GHz Cray Inc.
3	NASA/Ames Research Center/NAS United States	Pleiades - SGI Altix ICE 8200EX, Xeon QC 3.0/2.66 GHz SGI
4	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution IBM
5	Argonne National Laboratory United States	Blue Gene/P Solution IBM
6	Texas Advanced Computing Center/Univ. of Texas United States	Ranger - SunBlade x6420, Opteron QC 2.3 Ghz, Infiniband Sun Microsystems
7	NERSC/LBNL United States	Franklin - Cray XT4 QuadCore 2.3 GHz Cray Inc.
8	Oak Ridge National Laboratory United States	Jaguar - Cray XT4 QuadCore 2.1 GHz Cray Inc.
9	NNSA/Sandia National Laboratories United States	Red Storm - Sandia/ Cray Red Storm, XT3/4, 2.4/2.2 GHz dual/quad core Cray Inc.
10	Shanghai Supercomputer Center China	Dawning 5000A - Dawning 5000A, QC Opteron 1.9 Ghz, Infiniband, Windows HPC 2008 Dawning

[1] TOP500 Supercomputer List – Dongarra, Meuer, et al. – November 2008 URL: http://www.top500.org

Green500 [2]

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)	TOP500 Rank*
1	536.24	Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw	BladeCenter QS22 Cluster, PowerXCell 8i 4.0 Ghz, Infiniband	34.63	220
2	530.33	Repsol YPF	BladeCenter QS22 Cluster, PowerXCell 8i 3.2 Ghz, Infiniband	26.38	429
2	530.33	Repsol YPF	BladeCenter QS22 Cluster, PowerXCell 8i 3.2 Ghz, Infiniband	26.38	430
2	530.33	Repsol YPF	BladeCenter QS22 Cluster, PowerXCell 8i 3.2 Ghz, Infiniband	26.38	431
5	458.33	DOE/NNSA/LANL	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz , Infiniband	138	41
5	458.33	IBM Poughkeepsie Benchmarking Center	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz , Infiniband	138	42
7	444.94	DOE/NNSA/LANL	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz , Voltaire Infiniband	2483.47	1
8	371.67	ASTRON/University Groningen	Blue Gene/P Solution	94.5	75
9	371.67	IBM - Rochester	Blue Gene/P Solution	126	56
9	371.67	RZG/Max-Planck-Gesellschaft MPI/IPP	Blue Gene/P Solution	126	57

- PowerXCell and Blue Gene are dominant.
- Average Power Top10: [1]
  - June 2008:
    - 1.32 MWatt
    - 248 Mflop/(s-Watt)
  - November 2008:
    - 1.08 MWatt
    - 193 Mflop/(s-Watt)

[1] TOP500 Supercomputer List – Dongarra, Meuer, et al. – November 2008 URL: http://www.top500.org
[2] Green500 Supercomputer List – November 2008 URL: http://www.green500.org

Comparison

Blue Gene\L

- 65,536 Processors
- MPI
  - Specialized for implemented networks
- 360 teraflops peak
- 3D torus, binary trees
- Unit node: 2 cores

- Can it run Crysis
  - no

Roadrunner

- 122,400 cores (if SPE = core)
- MPI
  - Specialized for SPEs
- 1.38 petaflops peak
- Fat tree of crossbars
- Unit node: 20 cores
  - EIB interconnect is faster

- Can it run Crysis
  - Maybe, probably not though

Parting Thoughts

- General Trends:
  - Specialization of processing units
    - More accelerator based architectures
  - Move functionality on chip.
    - Reduction in power
    - Faster networks
  - Coherence is managed with the help of higher level abstractions
  - More cores and memory
    - 48 bits of address space => 32 TB

References

- TOP500 Supercomputer List – Dongarra, Meuer, et al. – November 2008 URL: http://www.top500.org
- Green500 Supercomputer List – November 2008 URL: http://www.green500.org
- T.H. Dunigan, Jr., J.S. Vetter et al., Performance Evaluation of the Cray X1 Distributed Shared Memory Architecture, IEEE Micro, 25(1):30-40, 2005.
- A. Gara, M. A. Blumrich, D. Chen, G. L.-T. Chiu, P. Coteus, M. E. Giampapa, R. A. Haring, P. Heidelberger, D. Hoenicke, G. V. Kopcsay, T. A. Liebsch, M. Ohmacht, B. D. Steinmacher-Burow, T. Takken, and P. Vranas. Overview of the Blue Gene/L System Architecture. IBM
- Journal of Research and Development, 49(2/3):195–212, 2005.
- Kevin J. Barker , Kei Davis , Adolfy Hoisie , Darren J. Kerbyson , Mike Lang , Scott Pakin , Jose C. Sancho, Entering the petaflop era: the architecture and performance of
- Roadrunner, Proceedings of the 2008 ACM/IEEE conference on Supercomputing, November 15-21, 2008, Austin, Texas [doi>10.1145/1413370.1413372]
- Kurzweil, Ray. “Growth of Supercomputer Power” 2008 .URL: http://www.singularity.com/charts/page71.html
- J.G. Koomey, “Estimating Total Power Consumption by Servers in the U.S. and the World”; http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf.