High-Performance Computing:
Blue Gene & Cray

Aaron Carpenter

March 28th, 2007

CSC 458

Outline

- HPC
- BlueGene/L
  - Processing module
  - Interconnect
  - Communication primitives
- Cray
- Cell Processor

HPC - Top500 List

Vendors/Architecture Type (Systems)

	Cluster	Constellations	MPP	Total
Appro International	2	-	-	2
Atipa Technology	4	-	-	4
Bull SA	-	1	-	1
California Digital Corporation	1	-	-	1
Cray Inc.	2	-	13	15
DALCO AG Switzerland	1	-	-	1
Dawning	1	-	-	1
Dell	18	-	-	18
Fujitsu	2	3	-	5
Galactic Computing	1	-	-	1
Hewlett-Packard	130	27	-	157
Hitachi	-	-	6	6
Hitachi/Fujitsu	1	-	-	1
HPTi	1	-	-	1
IBM	171	-	66	237
IBM/HP	1	-	-	1
Intel	1	-	-	1
lenovo	1	-	-	1
Linux Networx	7	-	-	7
NEC	-	-	3	3
NEC/Sun	1	-	-	1
Rackable Systems	1	-	-	1
Self-made	5	-	-	5
SGI	-	-	20	20
Sun Microsystems	9	-	-	9
Total	361	31	108	500

X1

Blue Gene/L

[1] TOP500 Report for November 2006 www.top500.org

HPC General

- Architectures:
  - Cluster: Parallel system of independent nodes
    - Each node is a system in itself
      - Constellation: Proc per node > Node per system
        - Uses shared memory model
      - Cluster-NOW: Proc per node < Node per system
        - Uses message passing memory model
  - MPP: Massive Parallel Processing
    - Not independent systems

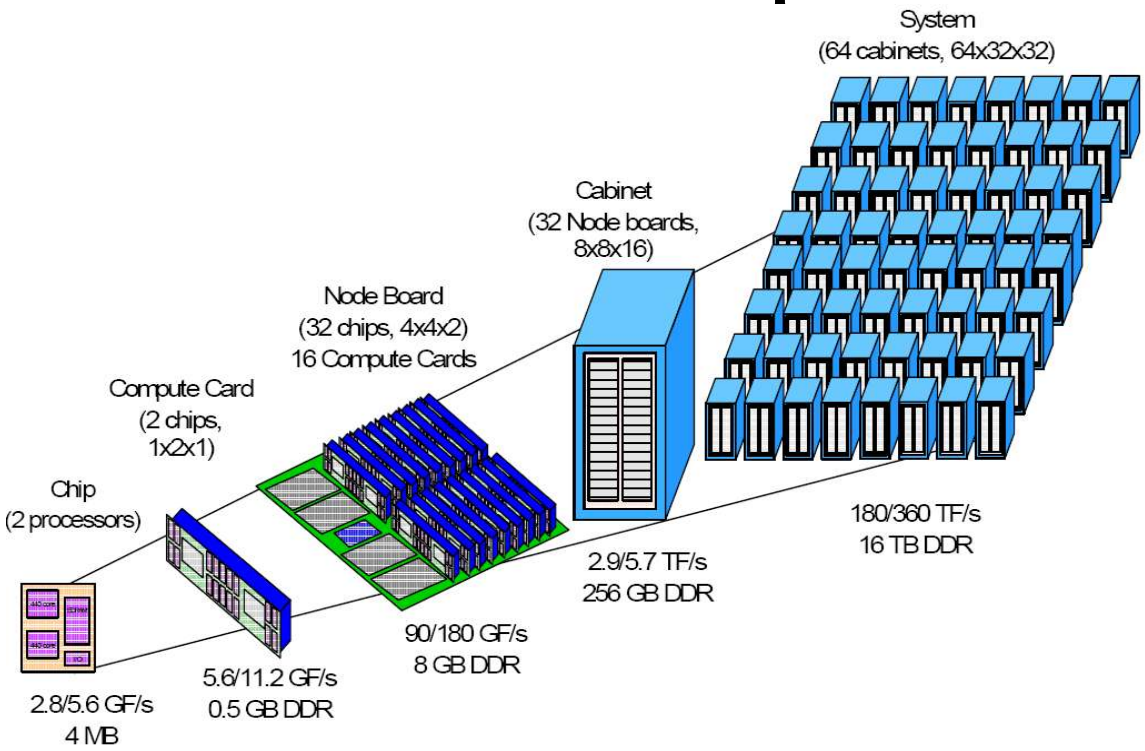
HPC General (cont.)

- Custom vs. Commodity (available for purchase)
  - Cray Red Storm (XT3)
    - Uses AMD for computation and SeaStar chip for communication
  - Blue Gene
    - PowerPC is off the shelf
- Communication
  - High Bandwidth, low contention
  - Low-latency interconnects → Scalability
- Operational division
  - Data processing, control, memory, communication
  - BlueGene/XT3 both specialize communication/data
  - Cell (not really HPC) specialized control/data

Blue Gene/L

- Joint Project: IBM & Lawrence Livermore National Laboratory
  - Funded by DOE ASC (Advanced Simulation and Computing)
- Fastest computer in the world (as of 11/06)
  - 65,536 nodes (2 processors/node, 2 FPUs per processor)
  - 280 TeraFlops (max)

Scaled Up

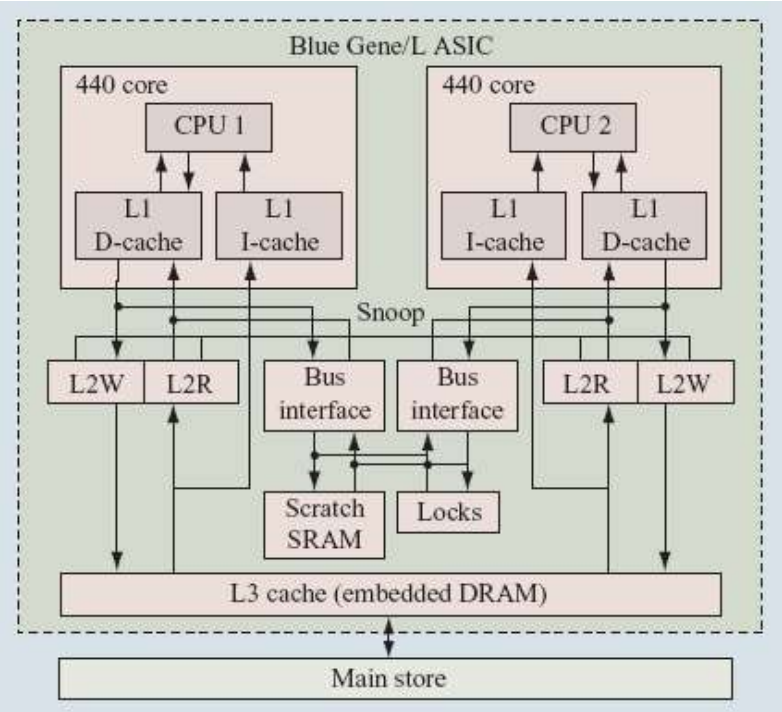


2 chips/card x 16 cards/board x 32 boards/cabinet x 64 cabinets/system
= 65,536 chips/system (2 proc each)

[2] "An Overview of the BlueGene/L Supercomputer," IBM Journal of Research & Development, Vol 49, No 2/3, March/May 2005.

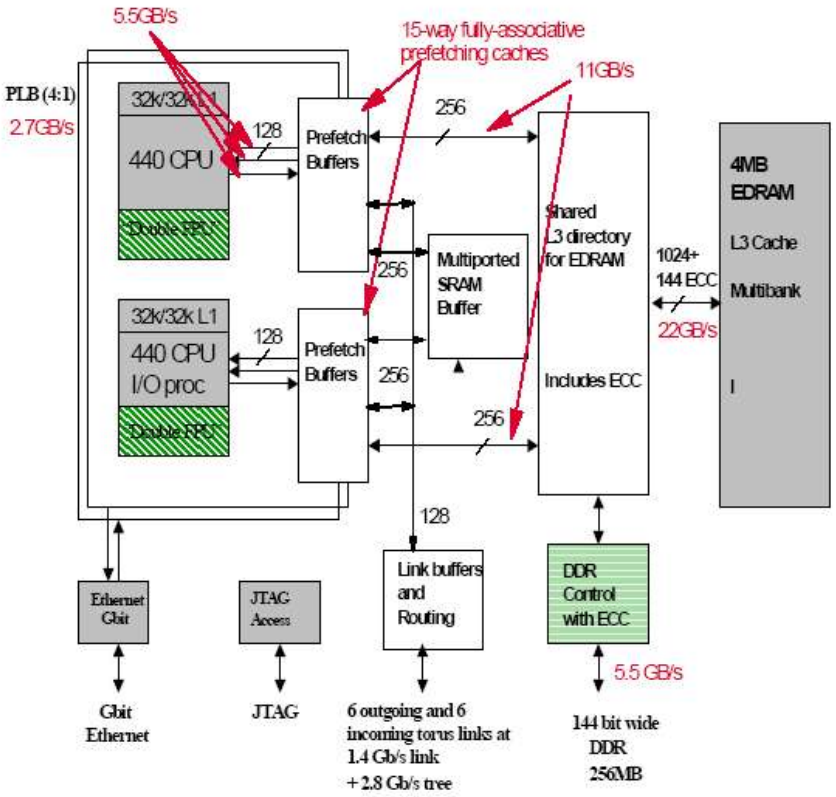
BG/L Processing Unit

- PowerPC 550 FP2 Core (700 MHz)
  - Chosen for its high performance/watt
  - 2 per node
    - How to use them?
  - Holds coherence primitives (invalidation)
    - No inherent coherence
    - Coherence must be explicit in program
- Cache Hierarchy
  - On chip L1 and L2 per processor
  - L2 small (acts as buffer)
  - L3 for dynamic shared memory



[2] "An Overview of the BlueGene/L Supercomputer," IBM Journal of Research & Development, Vol 49, No 2/3, March/May 2005.

BG/L Processing Unit



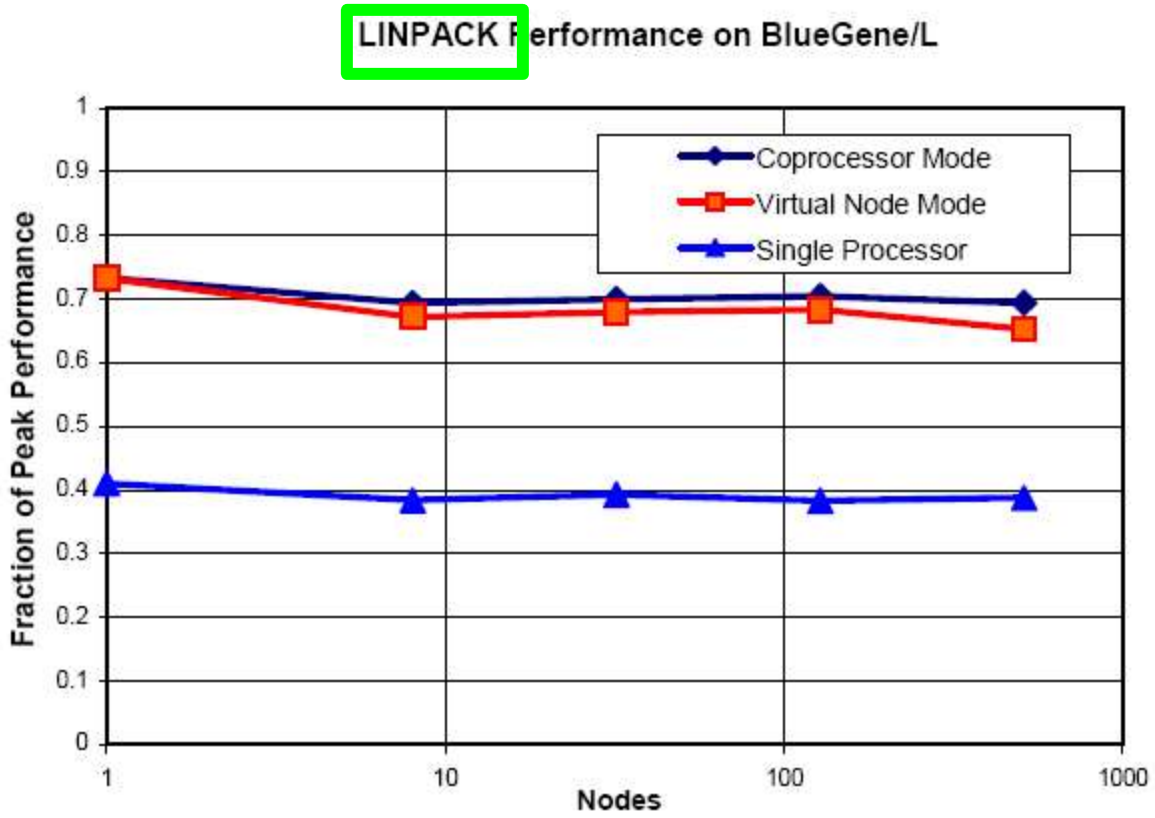
- Each core has 2 FPU
  - Each chip can have 4 flops per cycle
  - =2.8 Gflops/s
- Also provides links to interconnects

[2] "An Overview of the BlueGene/L Supercomputer," IBM Journal of Research & Development, Vol 49, No 2/3, March/May 2005.

How to use the 2nd processor

- 3 ½ options
  - 1) do nothing, let it sit
    - Wasted, but not resource problems
  - 2) Virtual: split resources in half, use as 2 processors
    - Half the resources for each processor, better parallel capabilities
  - 3) Coprocessor: 1st processor calls 2nd with fork/joins
    - 2nd processor shares resources, 1st processor still does majority of computation
    - 3A) Communication node **
      - 2nd processor handles all outgoing/incoming messages

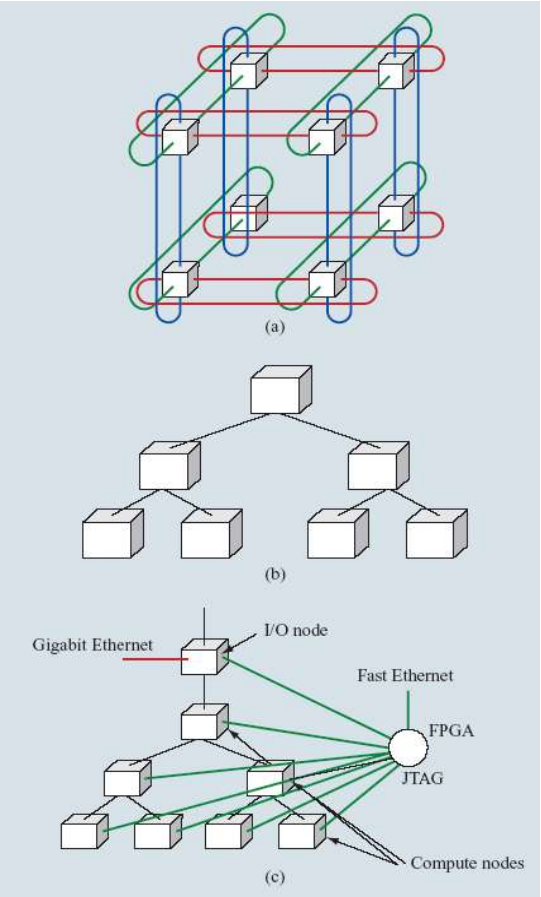
Node Use Comparison



[3] G. Almasi, *et al*, "Unlocking the Performance of the BlueGene/L Supercomputer," SC 2004 conference.

Interconnect

- Bi-directional 3-D Torus network
  - Used for point-to-point
    or multi-cast communication
  - Smallest torus is 512 nodes (smaller becomes cube)
- Binary tree
  - Global communication
  - Separate links from the torus network
- Other
  - Ethernet: connection to file system
  - JTAG: booting and control
  - Global Interrupt: job start, checkpointing, barriers
- Overall
  - Interconnect designed for low-latency
    communication



[2] "An Overview of the BlueGene/L Supercomputer," IBM Journal of Research & Development, Vol 49, No 2/3, March/May 2005.

Cache Coherence

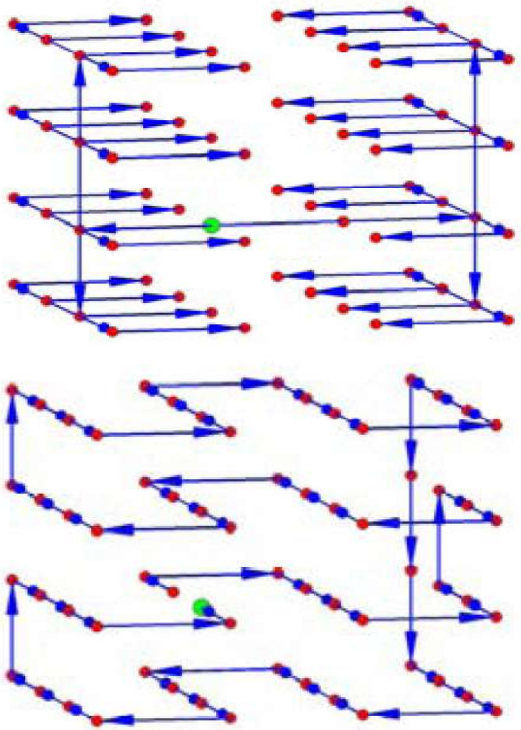
- BG/L has no cache coherence protocol
  - Provides primitives
    - Cache flush & cache invalidate
  - Coherence left up to programmer
    - Choice of node use
    - Programmer uses …

MPI

- Typical MPI would be loaded down by 100,000 processors
  - MPICH2 designed for it
- For send/receive/point to point communication
  - Use typical MPI primitives
- For collectives
  - Using point-to-point for all the nodes would not work
  - Must optimize MPI for the topology/size

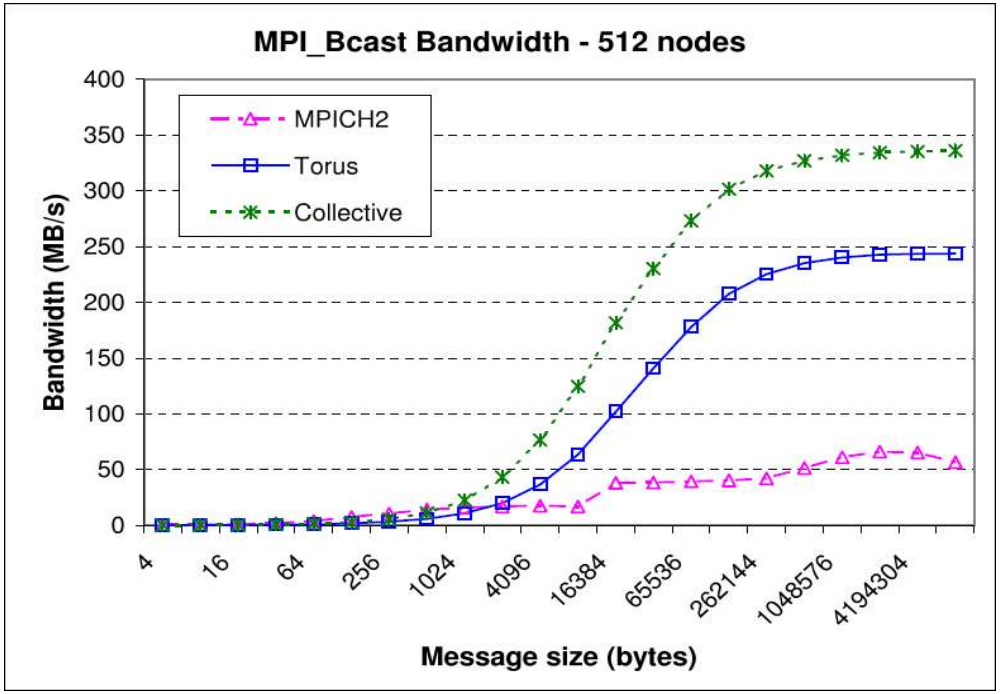
Optimized MPI Collectives

- Want to find the optimized comm. path
  - Bcast: Send to multiple nodes, copying data as necessary
  - Reduce: Only receive data from one neighbor (Hamiltonian path)
- Deposit Bit allows node to record data and then pass it
  - No need to process and resend



[4] G. Almasi, *et al*,"Optimization of MPI Collective Communication on BlueGene/L Systems," ICS 2005, p. 253-262

MPI_Bcast Bandwidth



Collective Network is still faster than Torus

[4] G. Almasi, *et al*,"Optimization of MPI Collective Communication on BlueGene/L Systems," ICS 2005, p. 253-262

BG/L – Parting Thoughts

- Reliability
  - Simple and redundant systems; ECC
  - If node fails, ignore that board
- Highly scalable (65,000+ nodes)
  - Can also partition it for multiple use
- Specialized: Can commit processors to communication/calculation

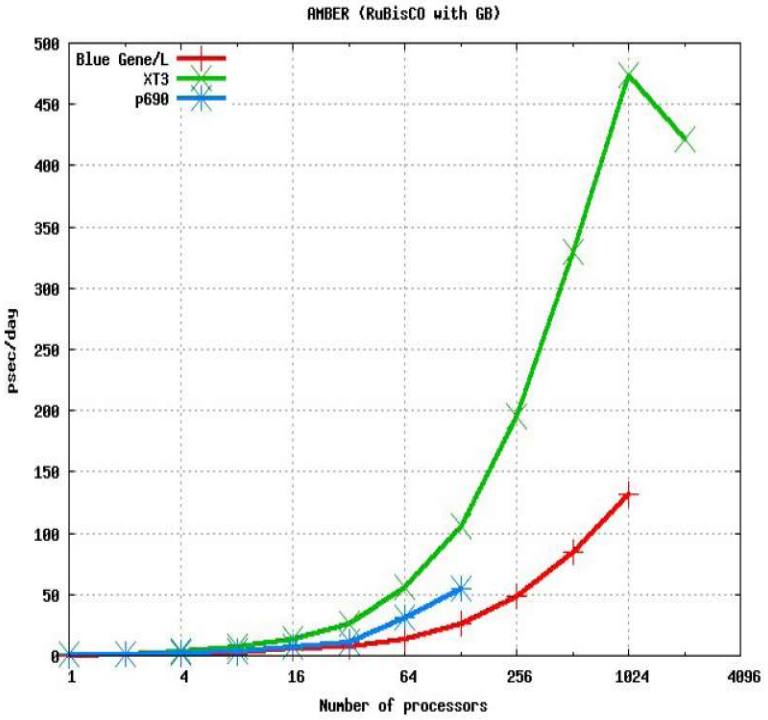
- But Cray is creeping up on them

Cray XT3

- Cray expects to scale further
  - Currently 5-10 thousand processors
- Dedicated:
  - processing node (AMD Opteron-small scale OS)
  - communication node (PowerPC SeaStar)
  - service node (AMD Operton-full Linux)
- 4 compute PEs/blade, 8 blades/chassis, 3 chassis/cabinet = 96 PEs/cabinet
- Nodes interconnected by torus
  - Portals messaging system

Cray XT3/4

- XT3 is climbing the Top500 list, with expected increases as it scales up to larger processor sizes
- #10 on the list
  - Using 1/10th the processors
  - Only ½ as slow

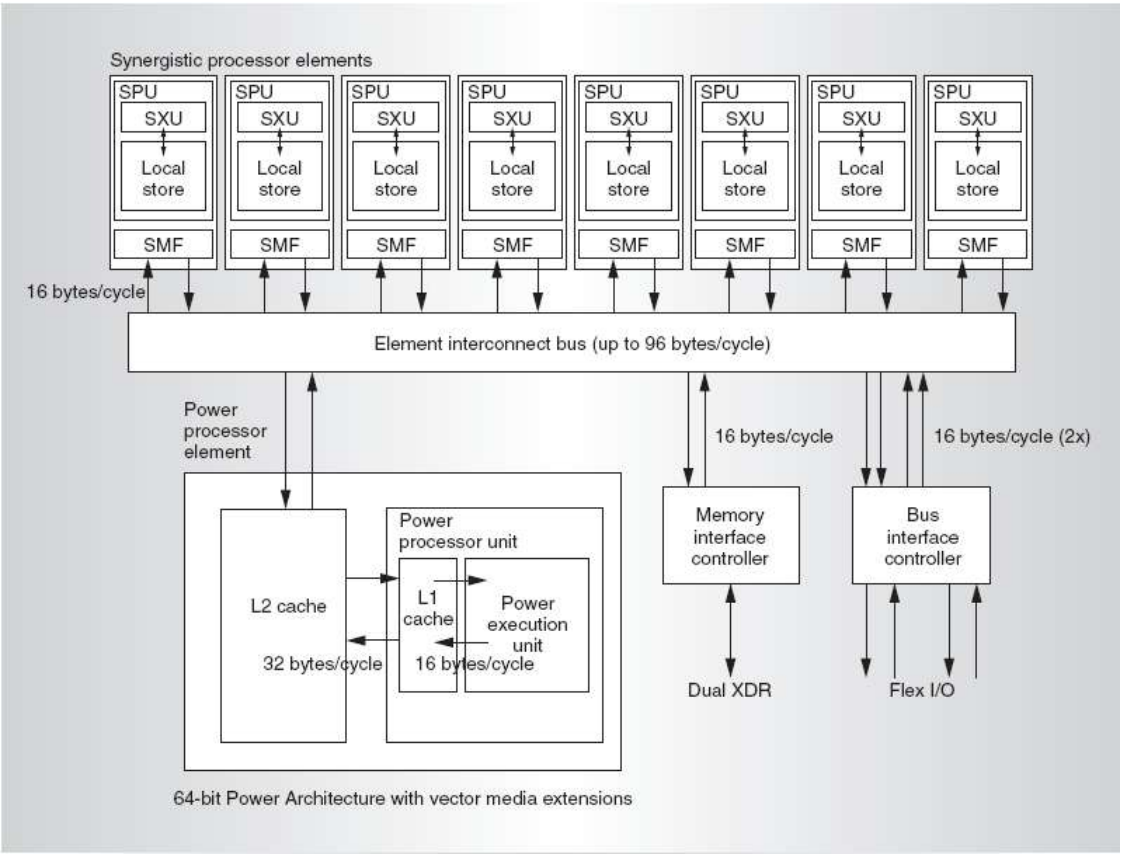


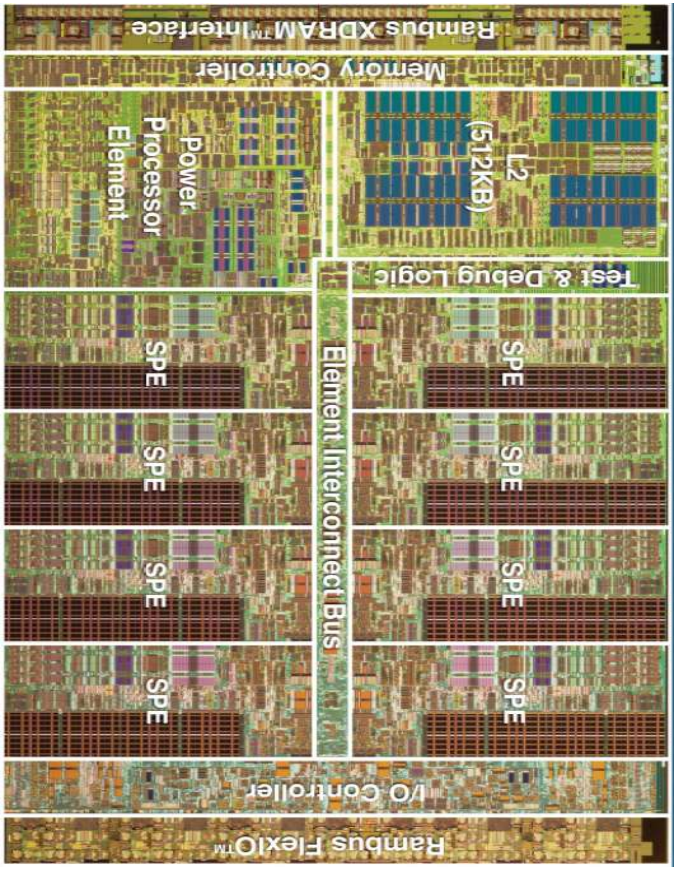
[5] J. Vetter, *et al*,"Early Evaluation of the Cray XT3," IPDPS 2006

IBM Cell Processor

- Not HPC, but demonstrates same principles
- Uses specialized processors
  - PPU: Power PE
    - Control Operations, OS
  - SPU: Synergistic PE
    - Data Processing
- By committing 1 processor to control
  - Can handle plenty of data
  - Most performance for the area
  - PPE is the only hot spot

CELL





[6] M. Gschwind,"Synergistic Processing in Cell's Multicore Architecture," Micro 2006, vol. 26, no. 2, pp 10-24

[7] M. Gschwind,"Chip Multiprocessing and the Cell Broadband Engine," IBM Research Report, Feb 2006

Summary

- HPC principles
  - Communication
    - Fast interconnects
    - Communication nodes
    - Reliable
  - Parallelization
    - Handle the large amount of data
    - Many processors with fast memory access
    - Programs written for lots of parallel accesses
  - Tasks parallelized by specialized hardware
    - Data isn't the only parallel thing

Top500

- New list will be announced at the end of June
  - International Supercomputing Conference
- Odds are BG/L still on top