High Performance Computing

Index

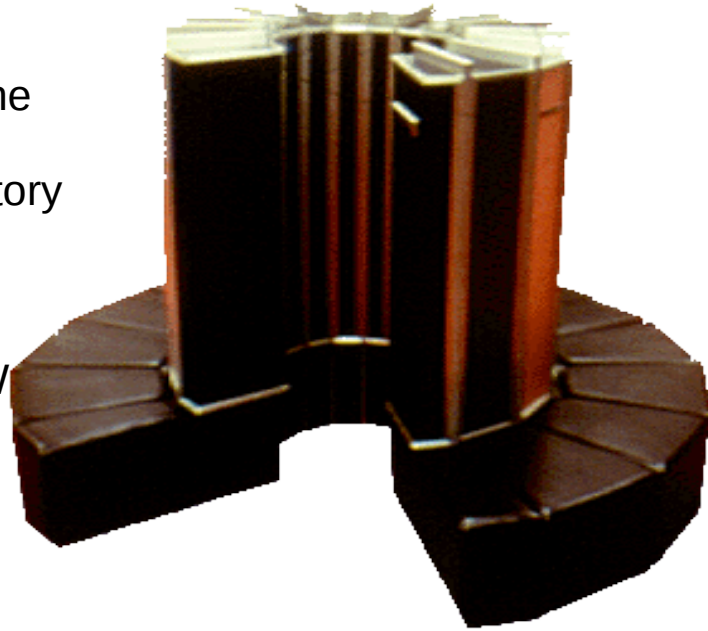
- History
- Performance Statistics
- Classfications
- Interconnect Networks
- Blue Gene
  - Blue Gene L
  - Blue Gene P
  - Blue Gene Q
- Titan Cray XK7
- Comparison

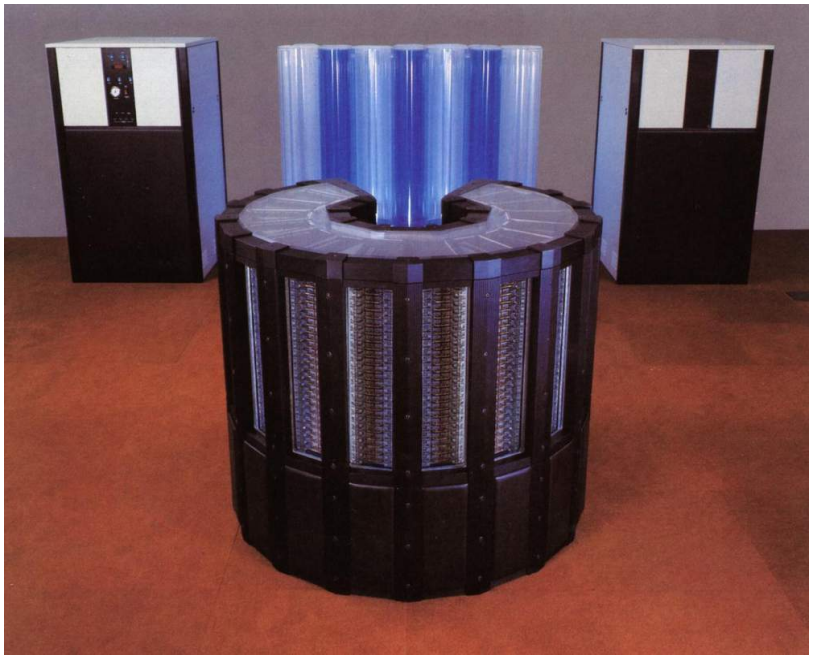
History

History

- 1960 : CDC 1604, first first commercially successful transistorized super-computer executing 100,000 OPS

- 1976: Cray 1, one of the most successful supercomputers in history operating at 80MHz providing a peak performance of 250MFLOPS at 115kW





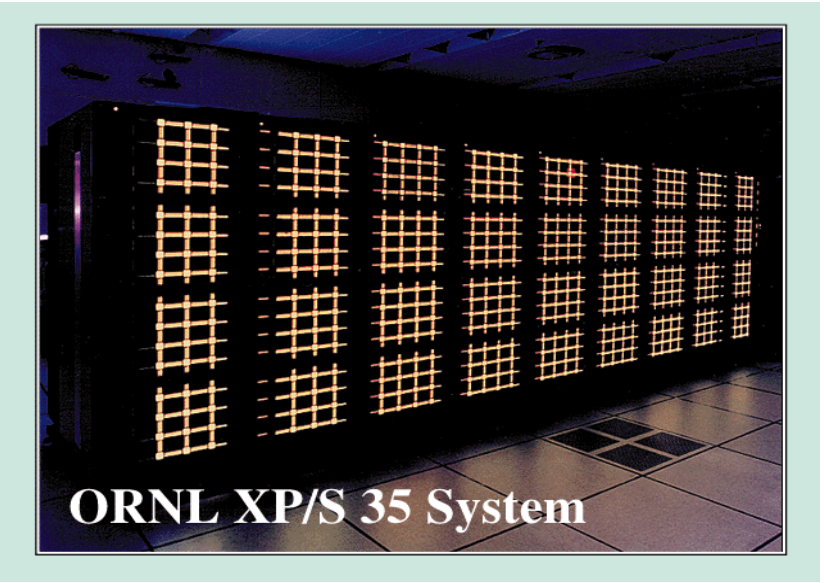
- 1985: Cray-2, a supercomputer consisting of four vector processors with a peak performance of 1.9 GFLOPS at 115-120kW became the fastest supercomputer until 1990s

http://en.wikipedia.org

History

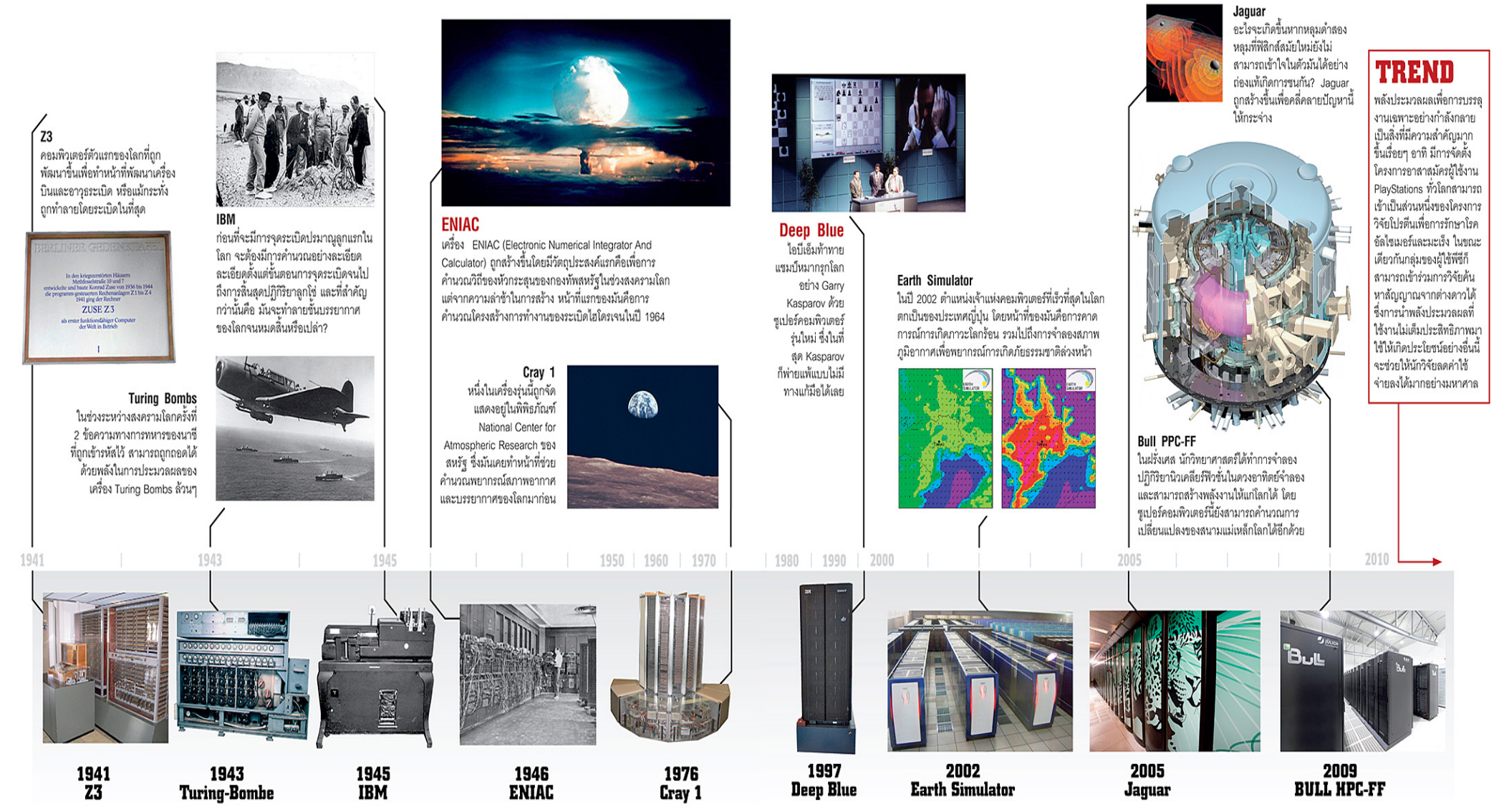
- 1994: Fujitsu's Numerical Wind Tunnel supercomputer used 166 vector processors to gain the top spot by attaining a peak speed of 1.7 gigaflops per processor.
- 1996: The Hitachi SR2201 obtained a peak performance of 600 gigaflops by using 2048 processors connected via a fast three dimensional crossbar network.
- 2000: ASCI White, a computer cluster based on IBM's commercial RS/6000 SP computer (512x16) 8,192 processors with a theoretical processing speed of 12.3 teraflops at 6MW
- 2004: Earth Simulator supercomputer built by NEC at the Japan Agency for Marine-Earth Science and Technology (JAMSTEC) reached 131 teraflops, using 640 nodes, each with eight proprietary vector processing chips
- 2007 : Blue Gene/P installation, JUGENE, with 16 racks (16,384 nodes, 65,536 processors) was running at Forschungszentrum Jülich in Germany with a performance of 167 TFLOPS at 7-8MW. This configuration incorporated new air-to-water heat exchangers between the racks, reducing the cooling cost substantially.

History



http://en.wikipedia.org

Timeline



http://chip.in.th/public/img/posts/timeline/supercomputer-timeline-full.jpg

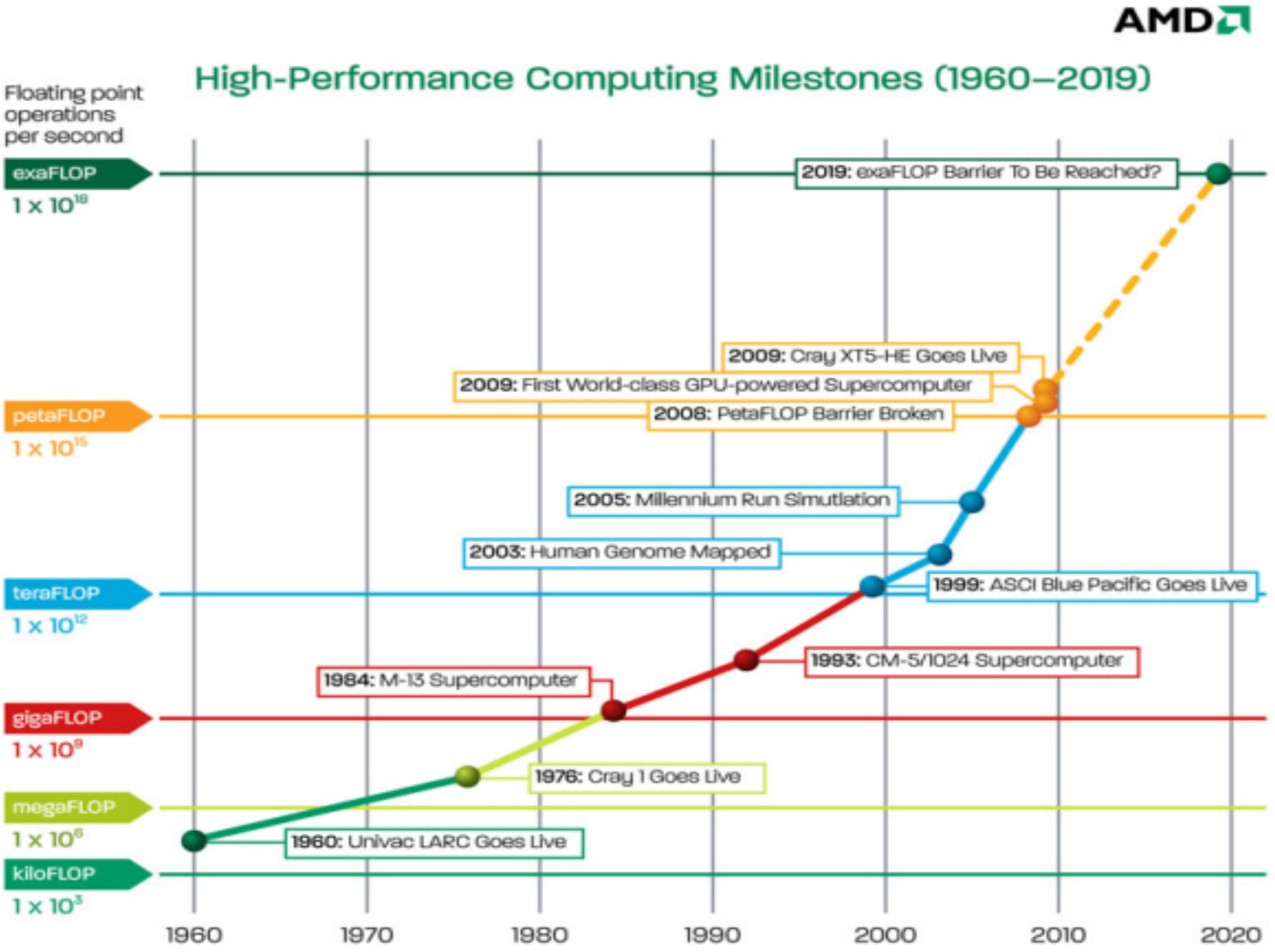
Performance Metrics

- Speed
- Power
- Cost (?)
- Scalability

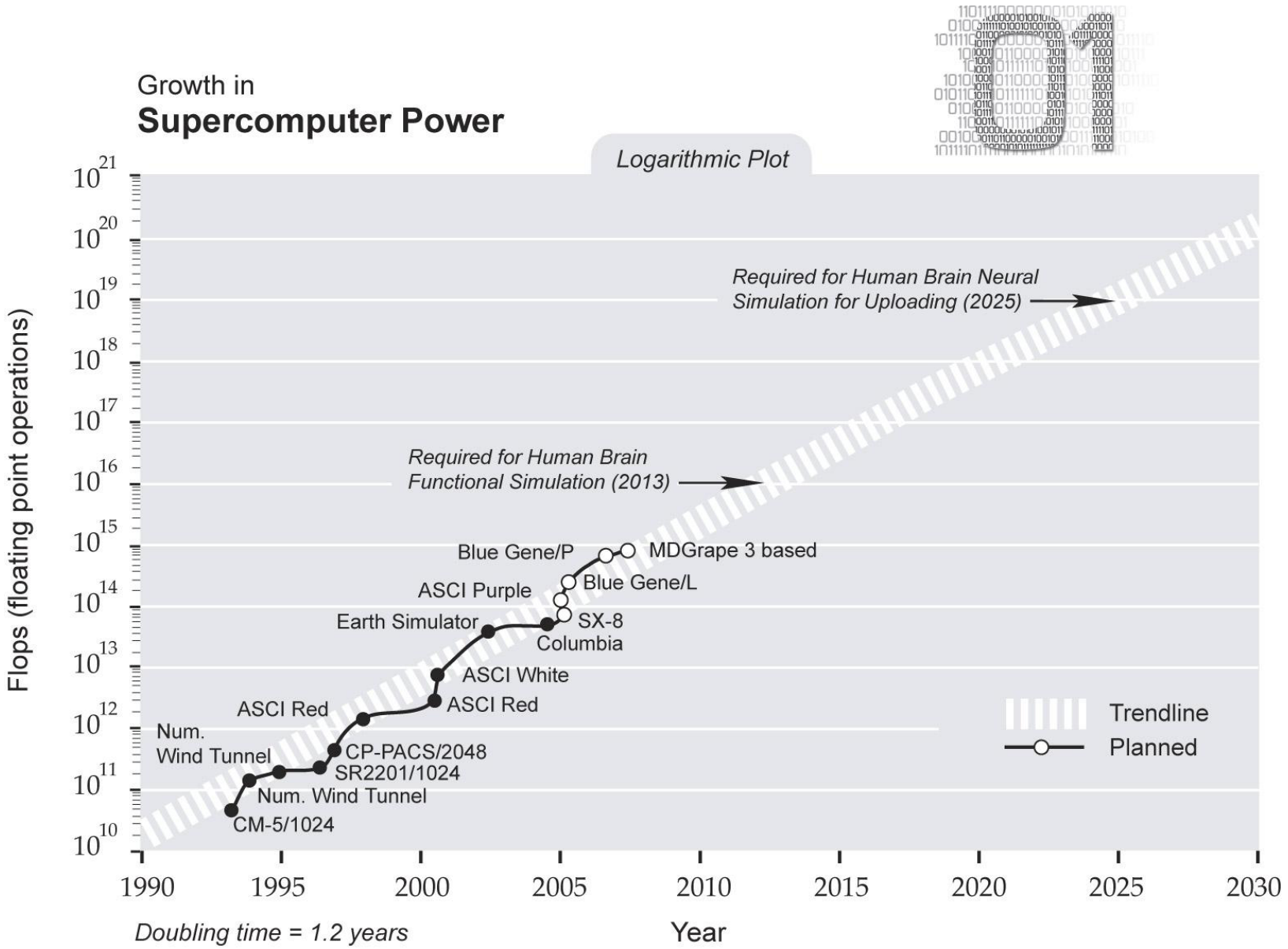
Performance Metics

Speed

Speed



Speed

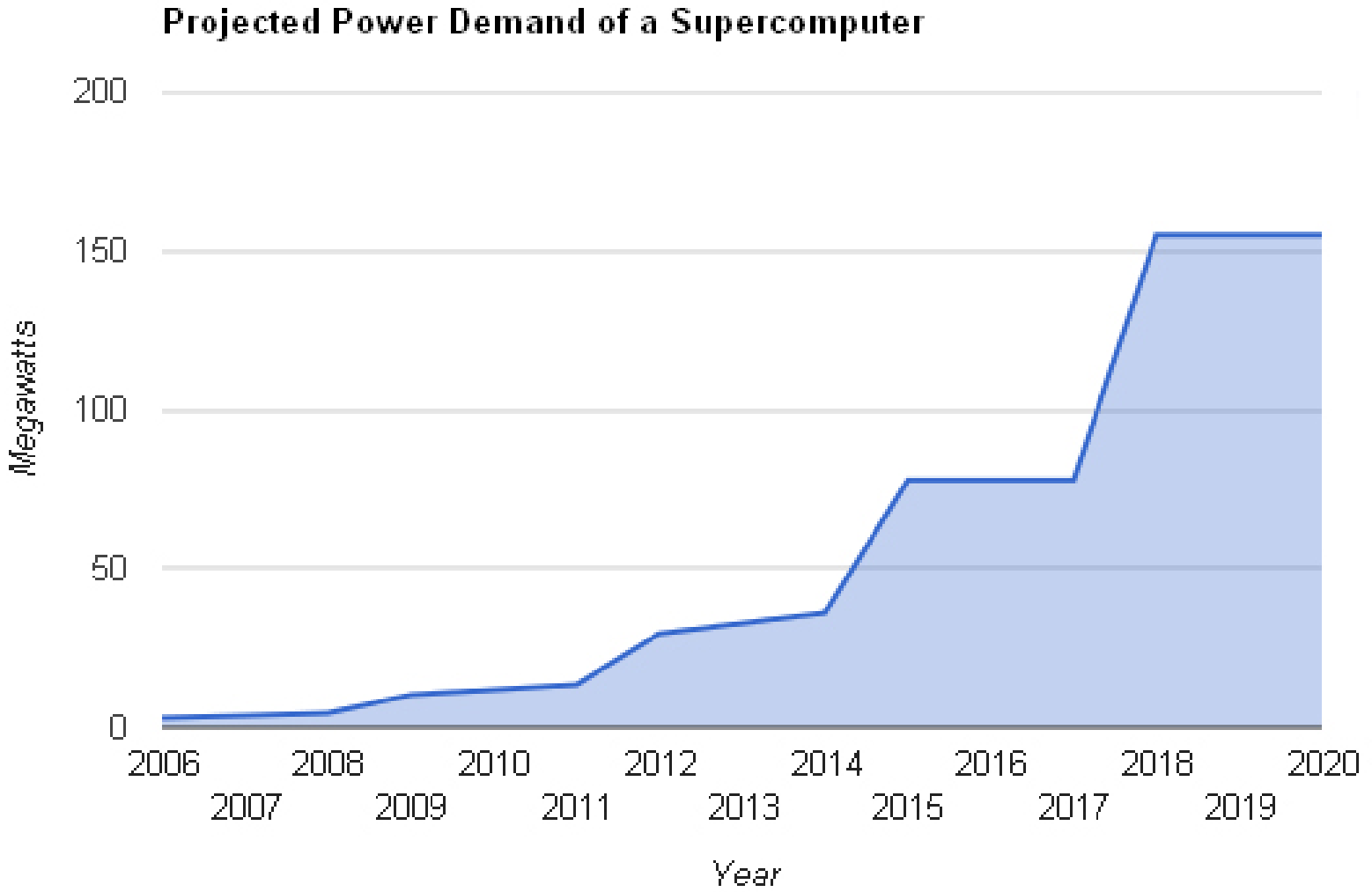


www.singularity.com

Performance Metics

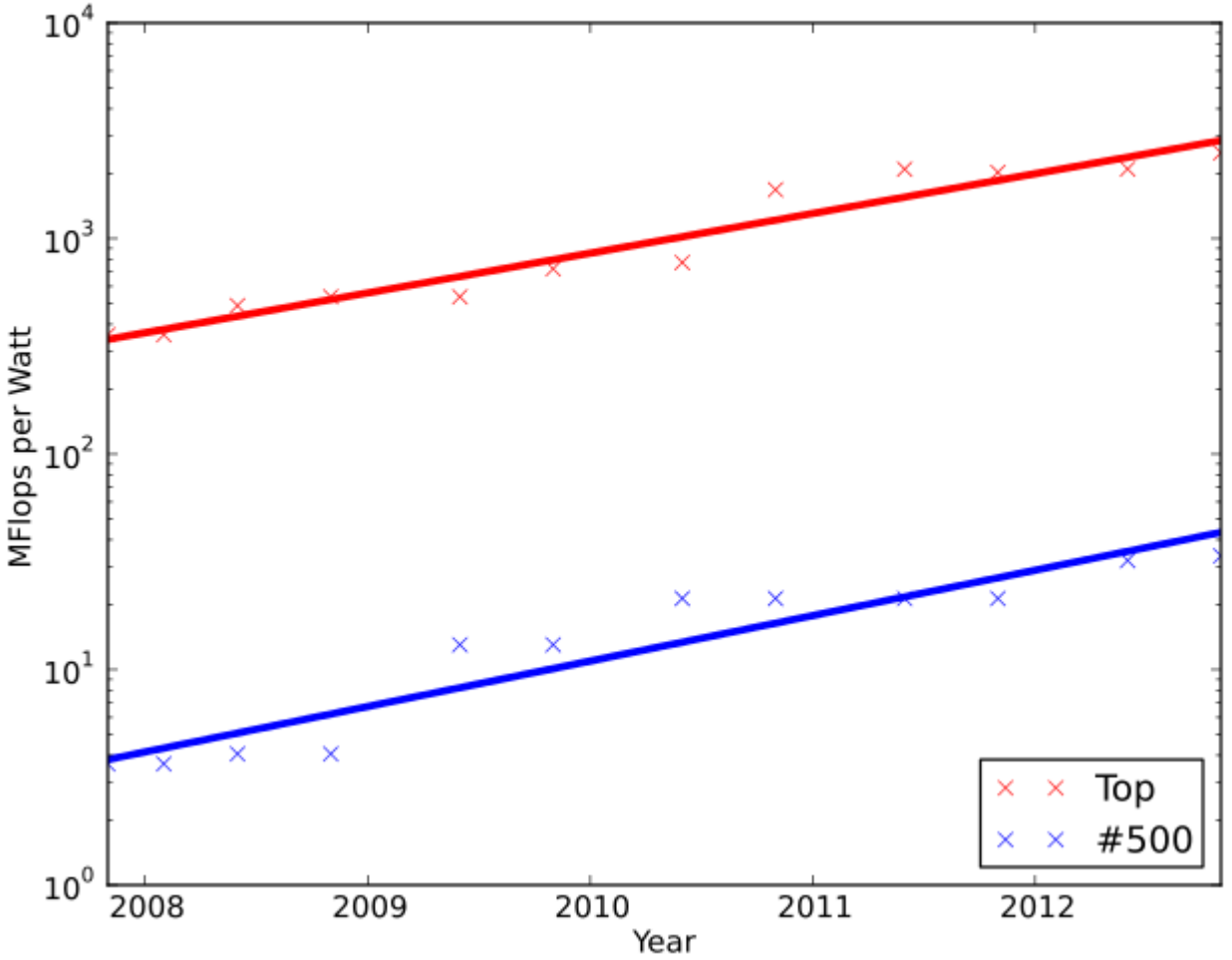
Power

Power



Peter M. Kogge, "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems," Sept. 28, 2008

Power

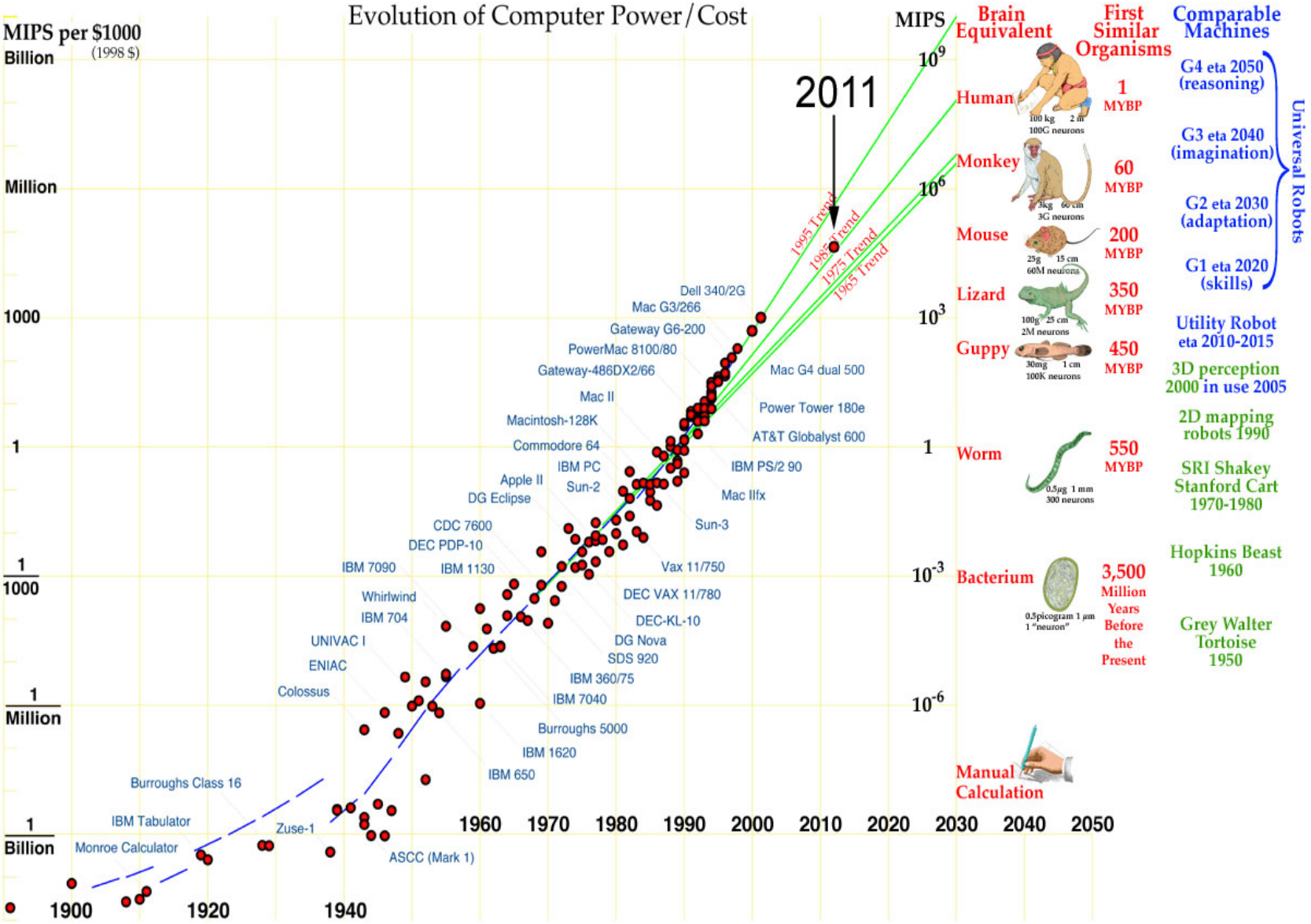


www.top500.org

Performance Metrics

Cost

Cost



"When will computer hardware match the human brain?" , Hans Moravec, Robotics Institute, Carnegie Mellon University, Journal of Evolution and Technology. 1998. Vol. 1

Classification

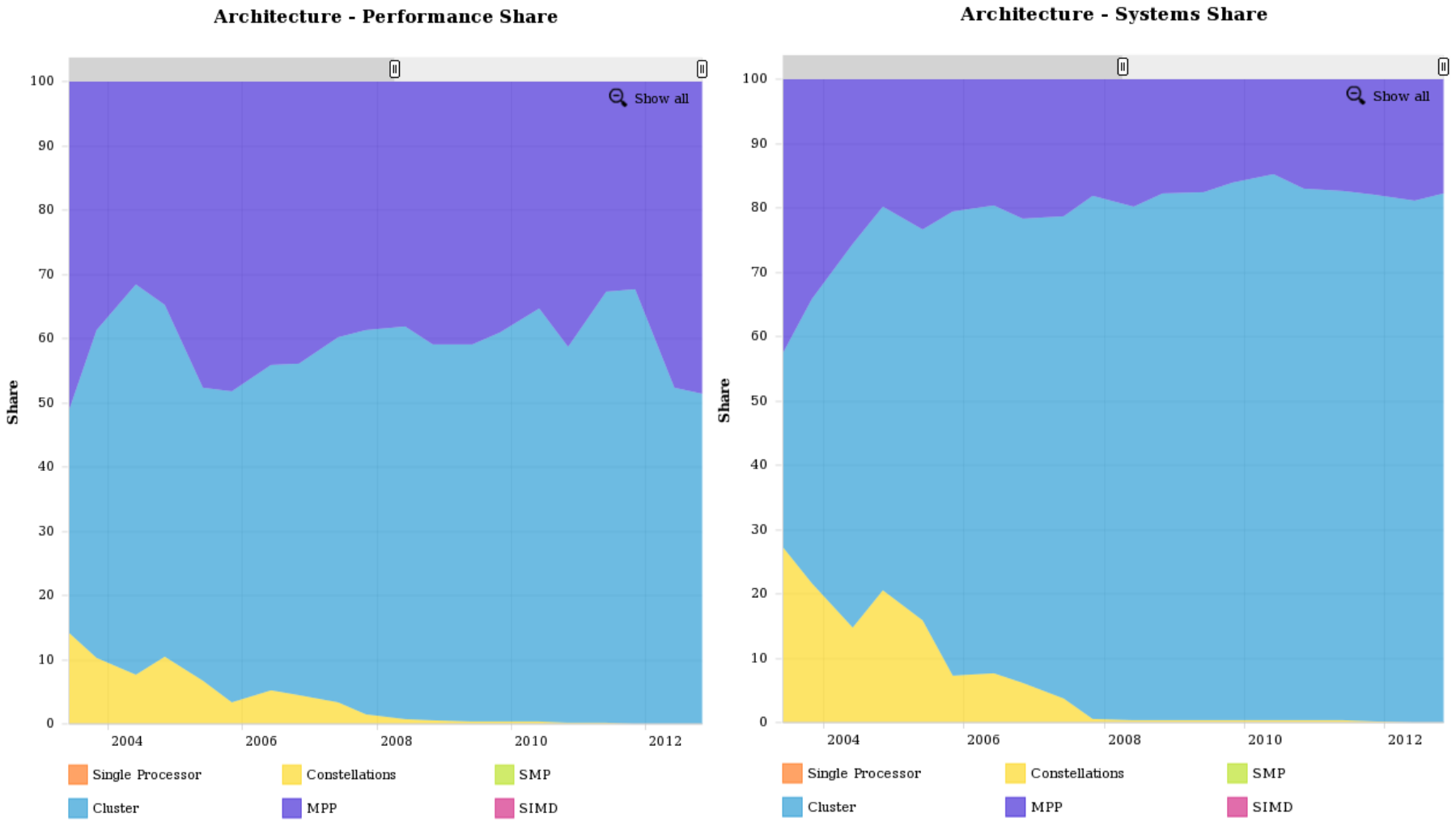
Classification

- MPP (Massively Parallel Processors)
- Clusters
- Constellations
- SMP

Classification

- Symmetric multiprocessing (SMP)
  - A multiprocessor computer hardware architecture where two or more identical processors are connected to a single shared main memory and are controlled by a single OS instance.
- Constellation
  - A cluster of large SMP nodes, where the number of processors per node is greater than the number of nodes.
- Massively Parallel Processing (MPP)
  - An MPP supercomputer usually implies a faster propitiatory very fast interconnect that supports either Distributed Shared Memory or even a Single System Image.
- Cluster
  - A bunch of machines, normally usually Ethernet interconnect (read: network), each running it's own and separate copy of an OS which happen to serve a single purpose.

Classification



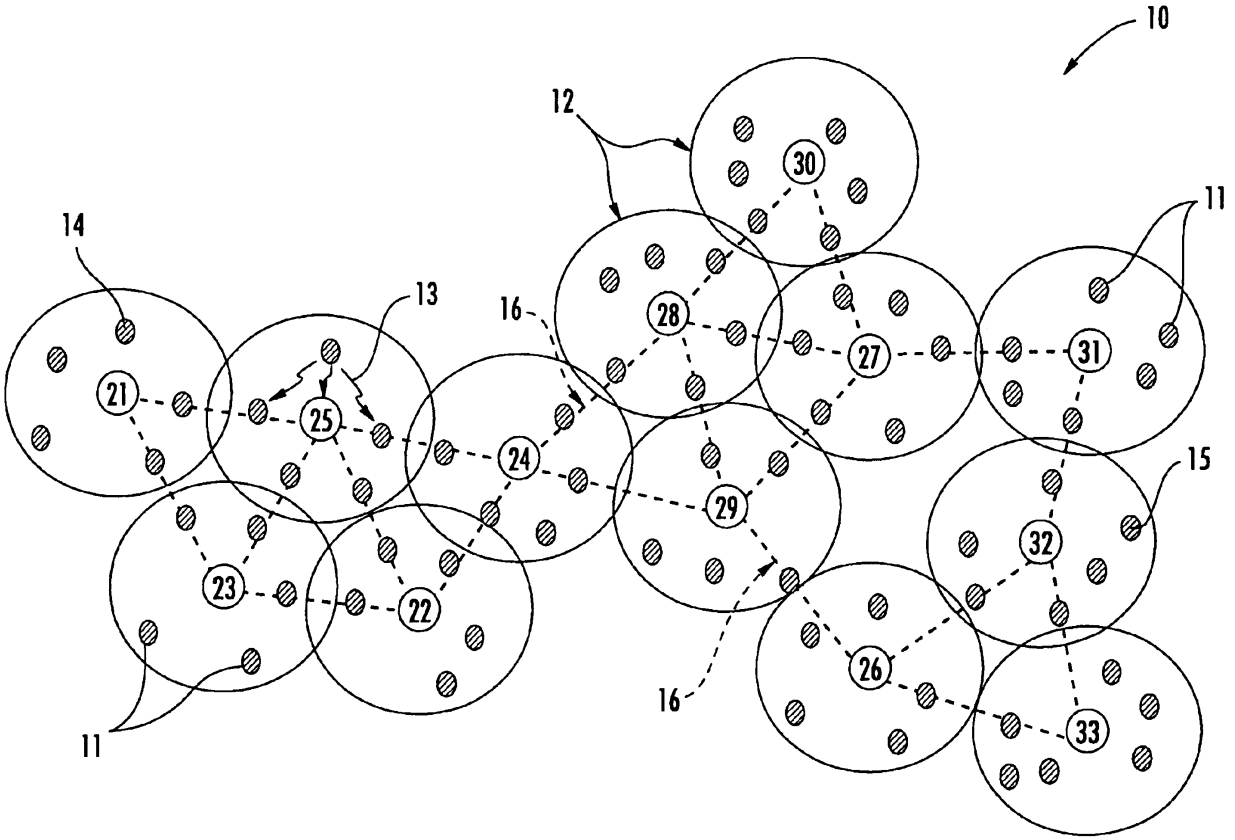
www.top500.org

Networks

Networks

- Mesh Architecture
- Master-Slave Architecture
- Ring Architecture
- Tree Architecture
- All-to-All Architecture

Networks



Networks

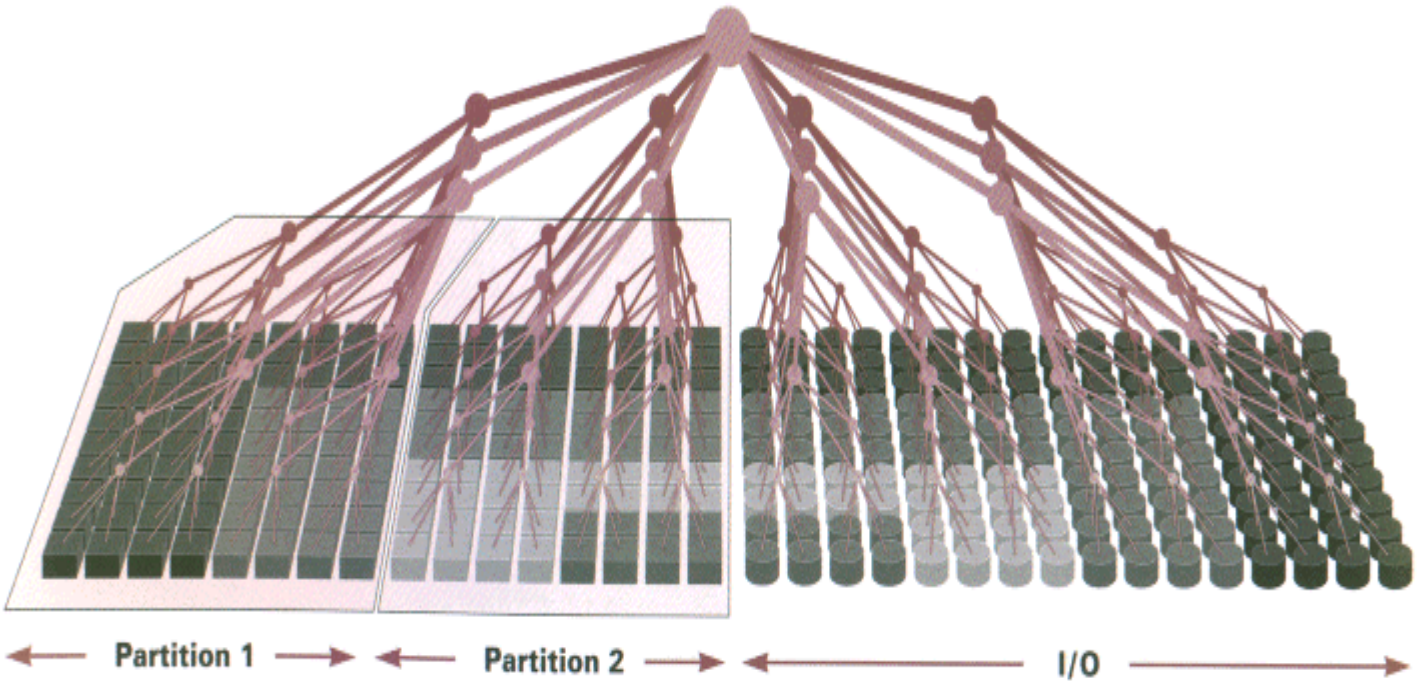
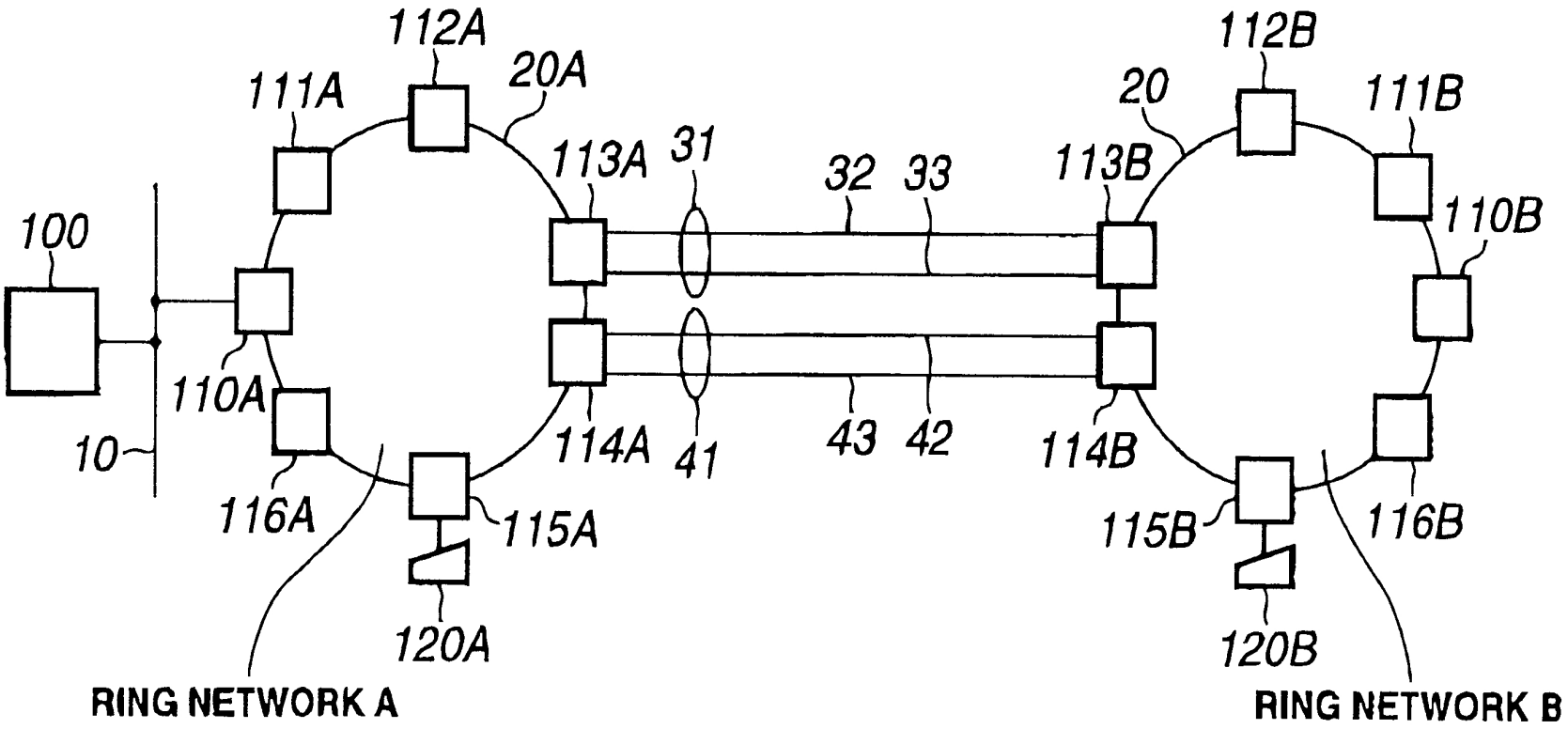
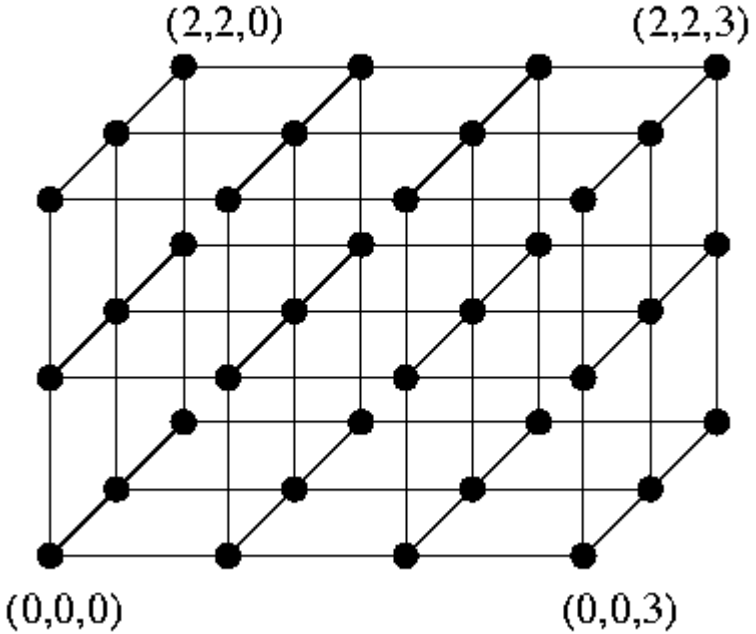


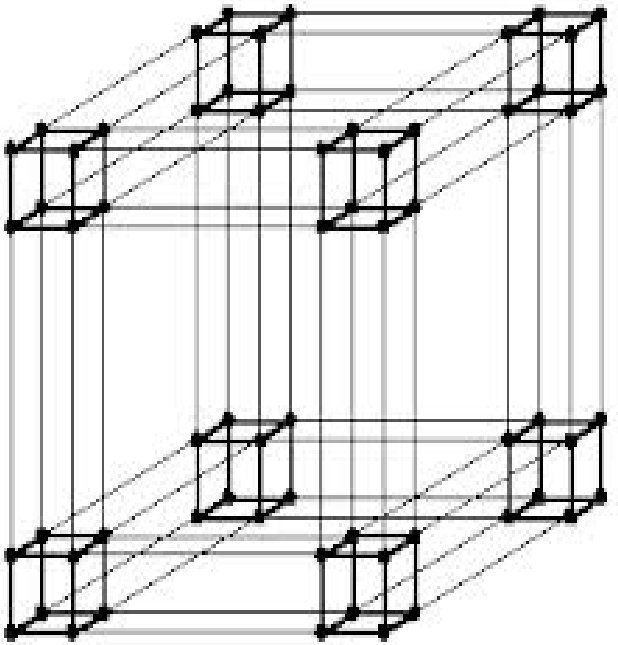
Figure 2

Networks


112A
111A
20A
113A
31
32
33
112B
20
111B
113B
100
110B
110A
10
114A
41
43
42
114B
116A
115A
115B
116B
120A
120B
RING NETWORK A
RING NETWORK B

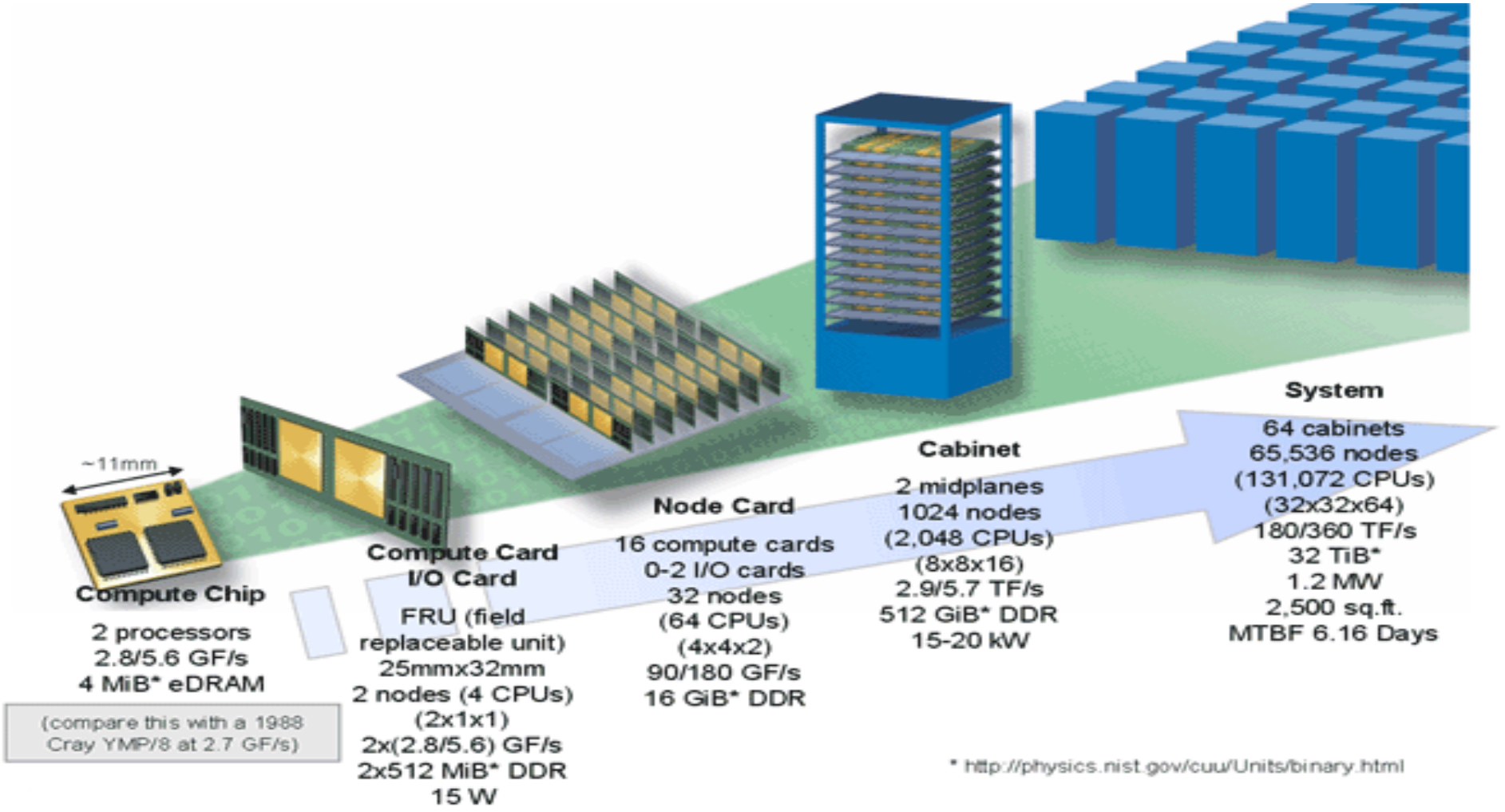
Networks





BlueGene

Blue Gene


~11mm
Compute Chip
2 processors
2.8/5.6 GF/s
4 MiB* eDRAM
(compare this with a 1988 Cray YMP/8 at 2.7 GF/s)
Compute Card
I/O Card
FRU (field replaceable unit)
25mmx32mm
2 nodes (4 CPUs)
(2x1x1)
2x(2.8/5.6) GF/s
2x512 MiB* DDR
15 W
Node Card
16 compute cards
0-2 I/O cards
32 nodes
(64 CPUs)
(4x4x2)
90/180 GF/s
16 GiB* DDR
Cabinet
2 midplanes
1024 nodes
(2,048 CPUs)
(8x8x16)
2.9/5.7 TF/s
512 GiB* DDR
15-20 kW
System
64 cabinets
65,536 nodes
(131,072 CPUs)
(32x32x64)
180/360 TF/s
32 TiB*
1.2 MW
2,500 sq.ft.
MTBF 6.16 Days
* http://physics.nist.gov/cuu/Units/binary.html

Blue Gene

- Blue Gene L
- Blue Gene P
- Blue Gene Q

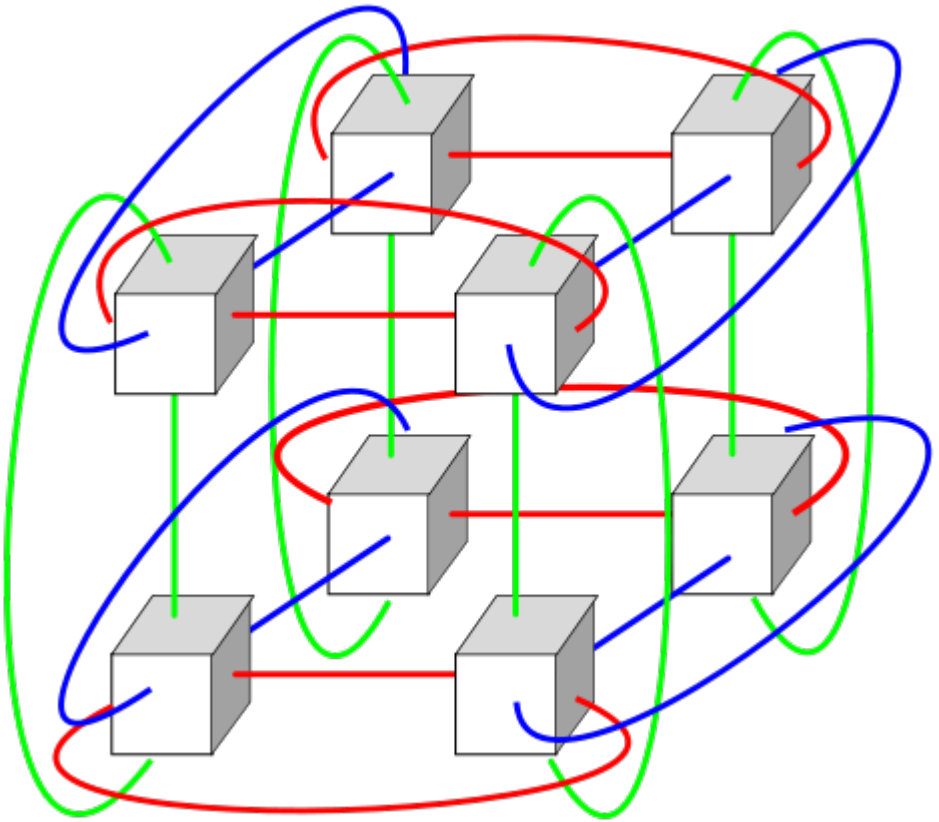
BlueGene L

Blue Gene L

- Unique Architectural Features
  - 3D Taurus Interconnect Network
  - BGL Compute Chip (each rack has 2x2x256)
  - Traded Speed for Power

Blue Gene L

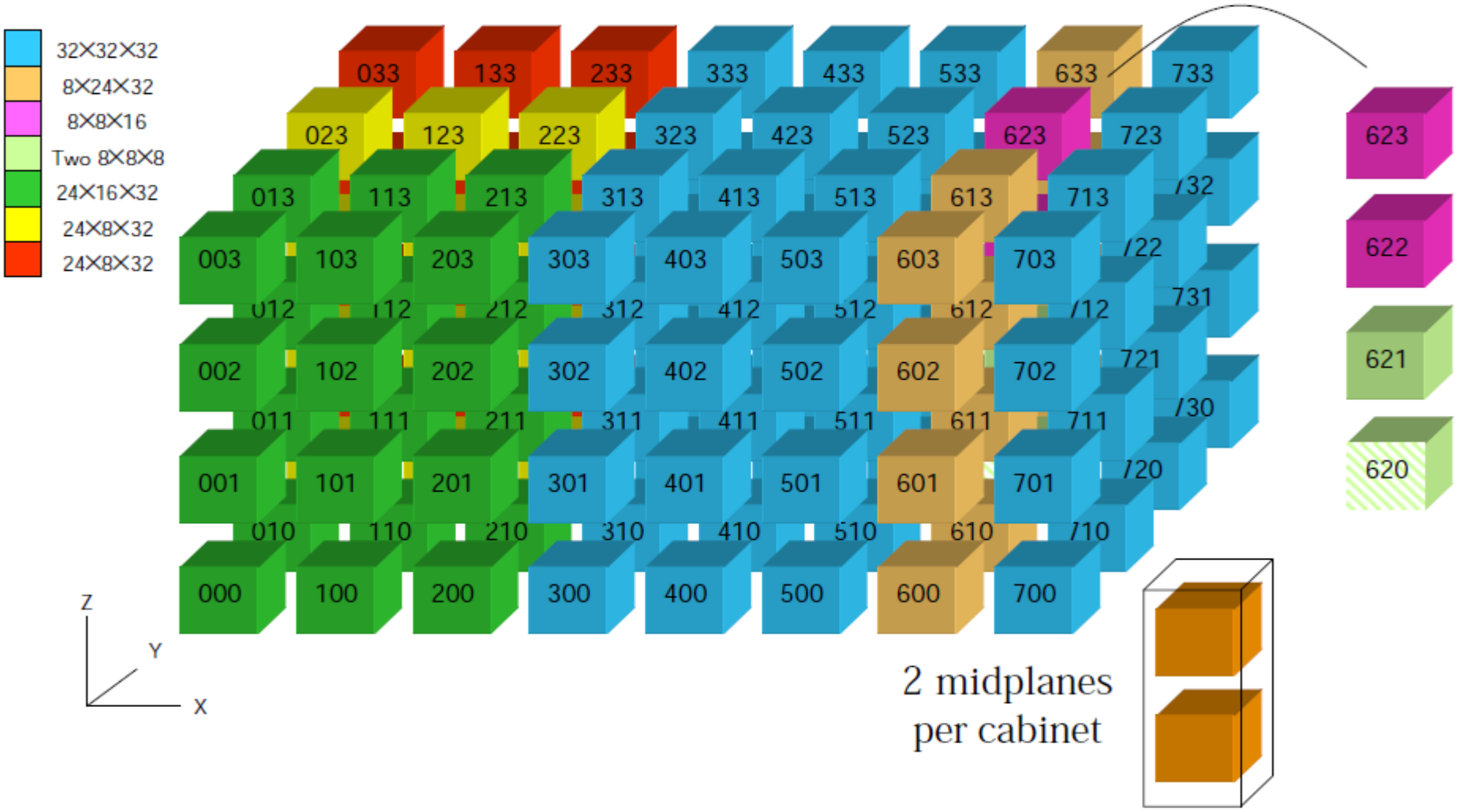
- Taurus Interconnect



http://en.wikipedia.org

Blue Gene L

- Taurus Interconnect



https://asc.llnl.gov/computing_resources/bluegenel/pdf/BGL.pdf

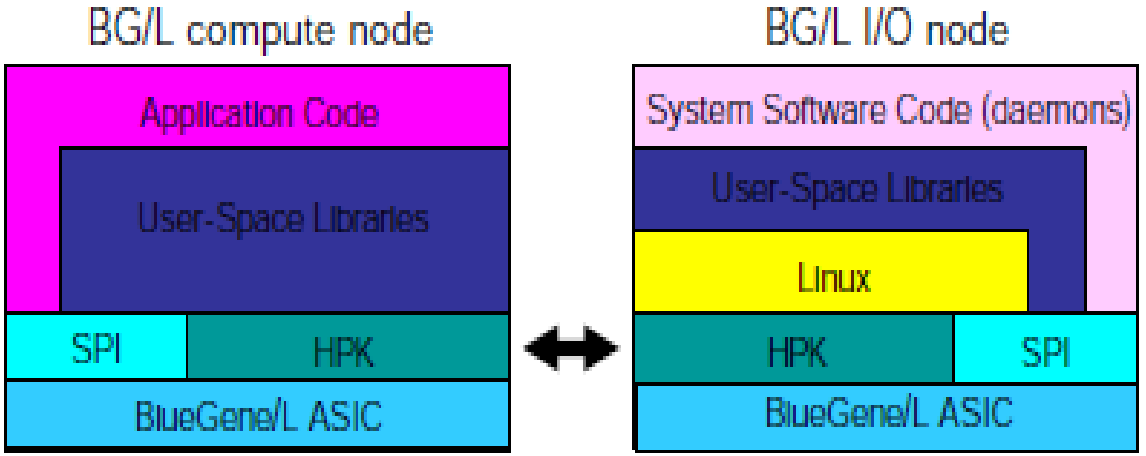
Blue Gene L

- BGL Compute Chip

  (Dual Power PC 440 Cores @ 700 MHz)

  - Features two modes
    - Co-processor mode
    - Virtual-node mode
  - CPUs are not cache coherent

Blue Gene L

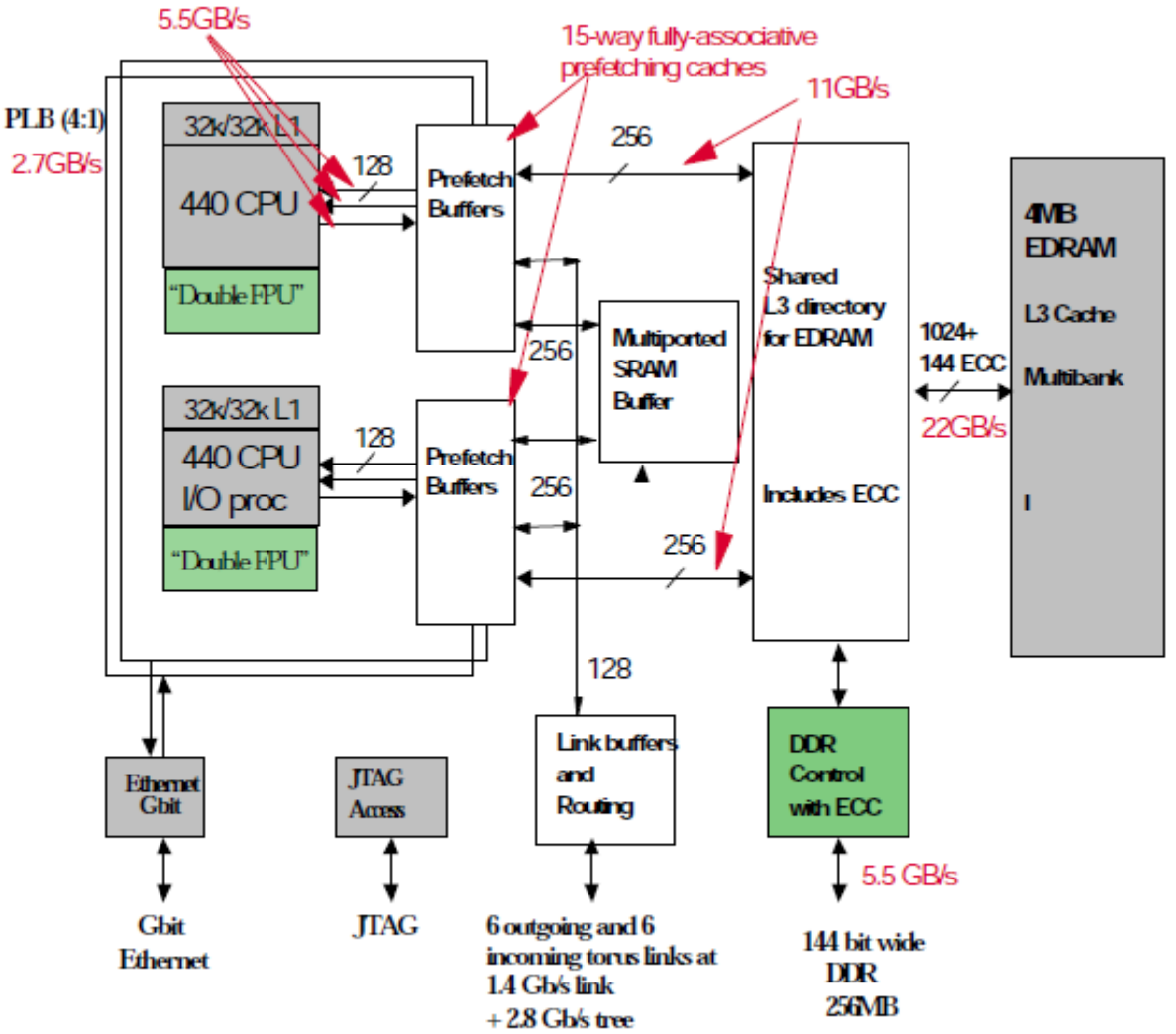


Light-weight kernel AND Linux services

https://asc.llnl.gov/computing_resources/bluegenel/pdf/BGL.pdf

Blue Gene L

- All of this on a 14mm x 14 mm die



https://asc.llnl.gov/computing_resources

Blue Gene L

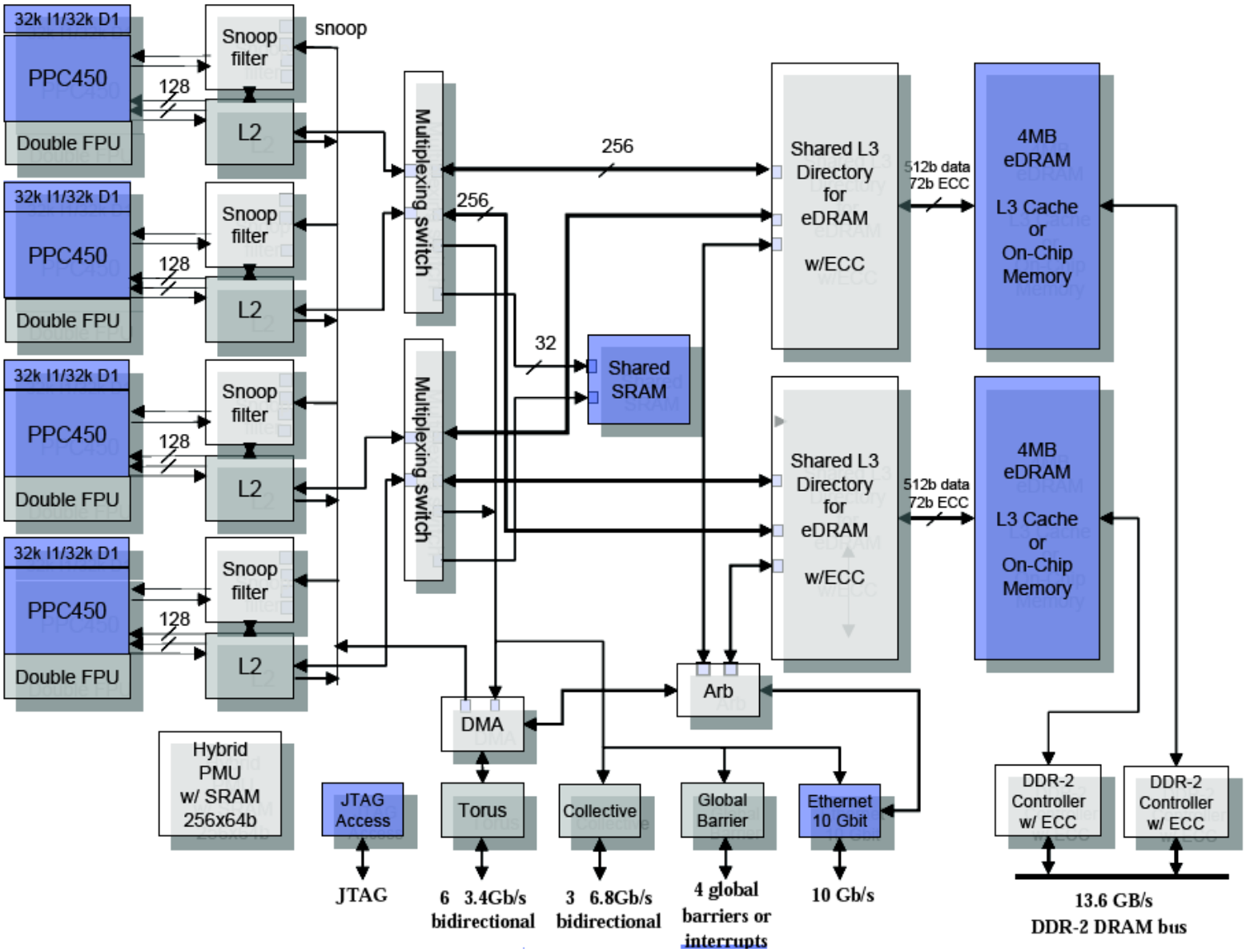
- Used a minimal OS supporting a single user program.
- Only a subset of POSIX calls was supported, and only one process could run at a time on node in co-processor mode—or one process per CPU in virtual mode.
- Programmers needed to implement green threads in order to simulate local concurrency.
- Application development was usually performed in C, C++, or Fortran using MPI for communication.

BlueGene P

Blue Gene P

- A technology evolution of Blue Gene L
  - 4 Power450 Cores @ 850 MHz (4x1024 cores/rack)
  - Cache coherent Cores (Snooping H/W)
  - Can operate as 4 way SMP
  - Taurus interconnect network with twice the bandwidth

Blue Gene P



anitaborg.org/wiki/index.php/File:GHC07-BlueGene_salapura.pdf

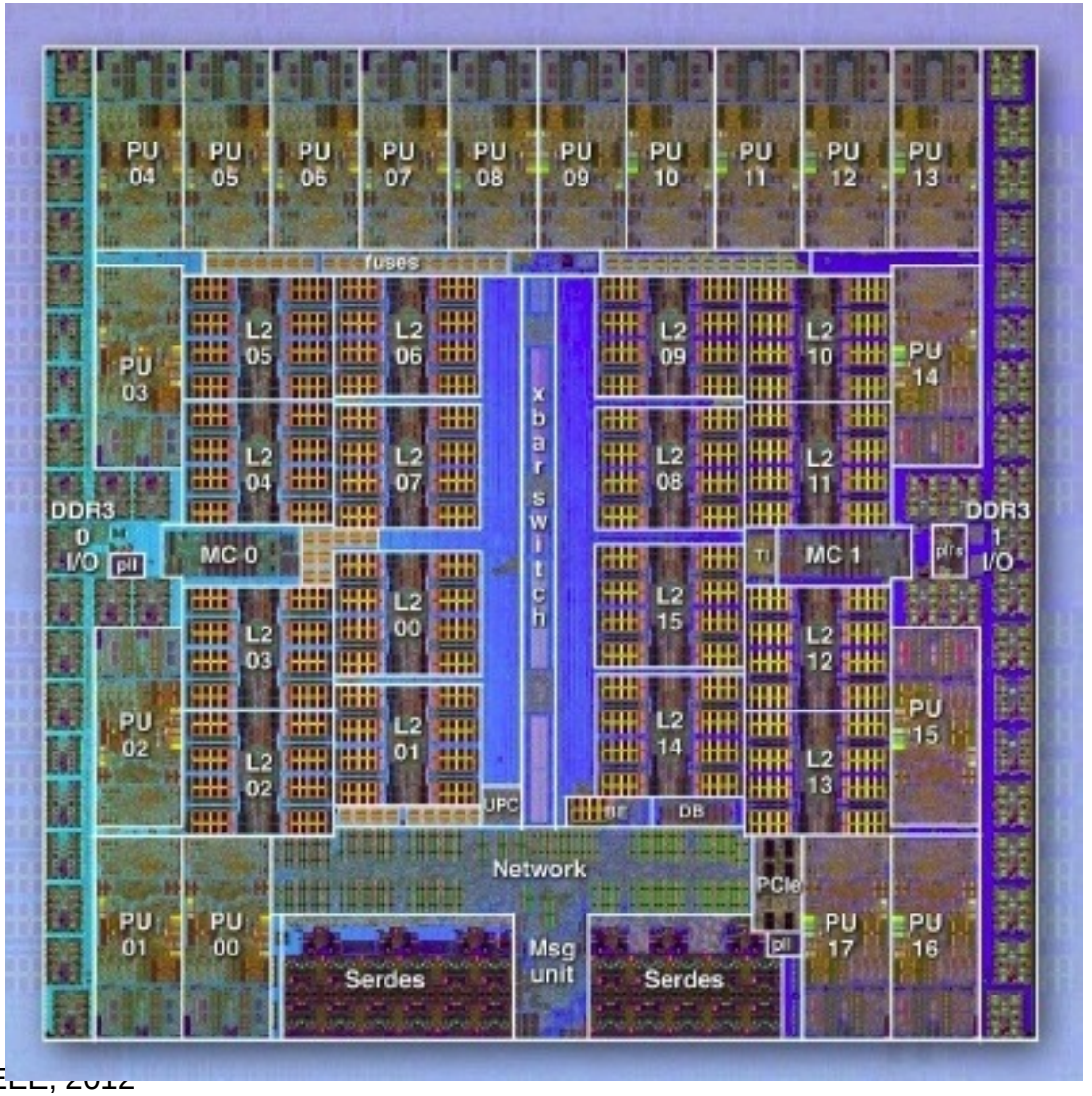
BlueGene Q

Blue Gene Q

- Unique Features
  - PowerPC A2 core (massively parallel)
    - 4 SMT Capability
    - QPU (Quad fp Processing Unit)
    - QPX (Quad Processing eXtension) ISA
    - L1 Prefetch Unit
  - 18 Cores (16 + 1 + 1) 1024 x 16 user cores/rack
  - Crossbar Switch (A network topology in 1996)
  - Transactional Memory (H/W support)
  - 5D Torus Network

Blue Gene Q

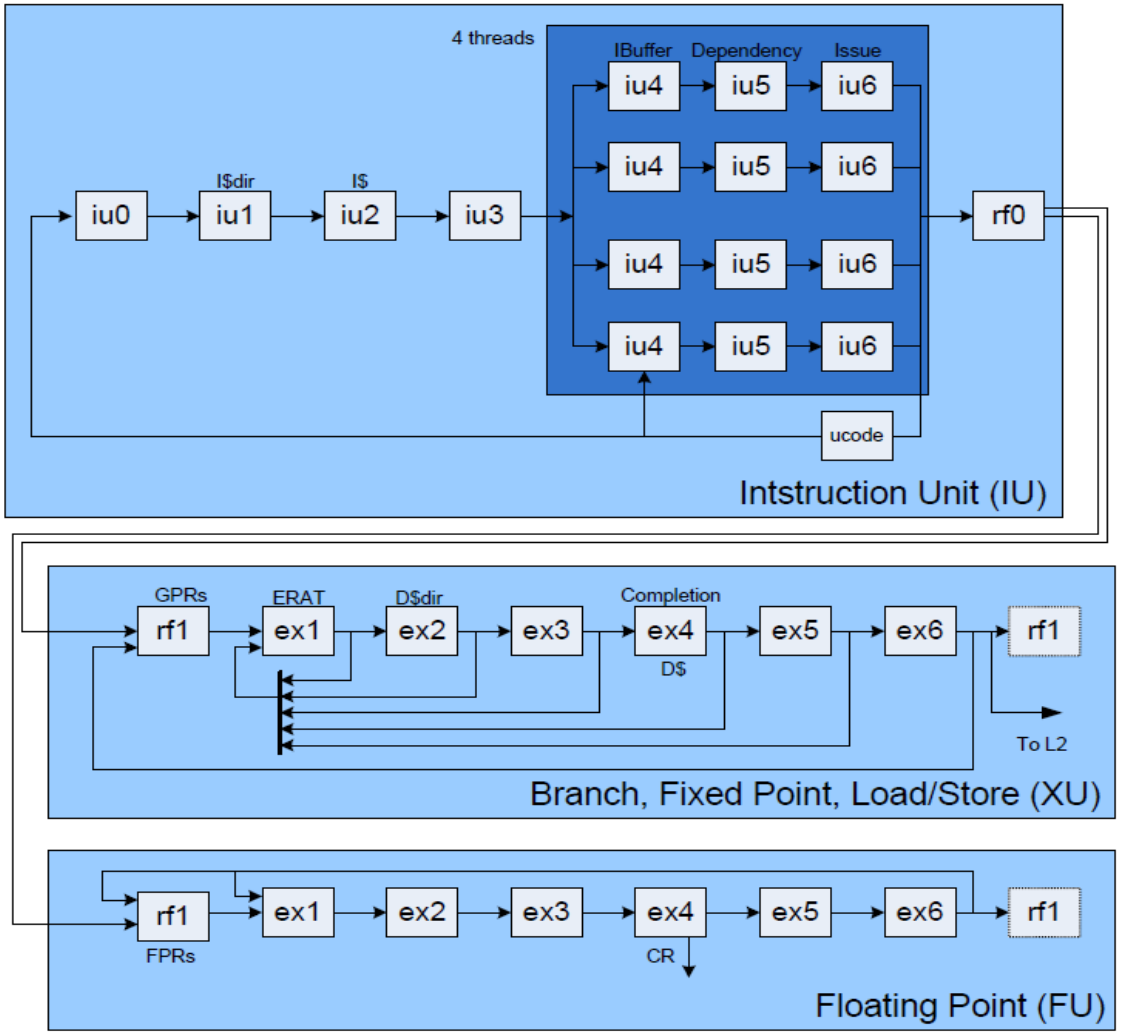
- BGQ Compute Chip
  - 18 Cores
  - Xbar switch



IBM BLUE GENE/Q COMPUTE CHIP, Micro, IEEE, 2012

Blue Gene Q

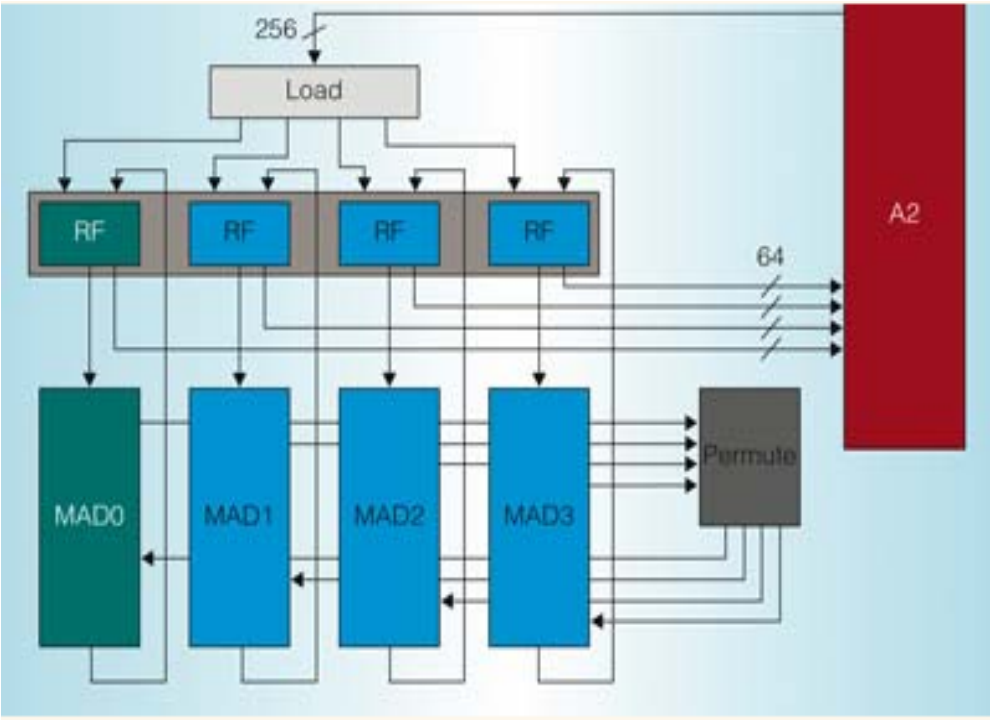
- PowerPC A2 Core



https://wiki.alcf.anl.gov/parts/images/c/cf/A2.pdf

Blue Gene Q

- PowerPC A2 Core
  - Replaces the dual-FU with QPU
    - Multiple Add Dataflow
    - Max 4x2 FLOP



IBM BLUE GENE/Q COMPUTE CHIP, Micro, IEEE, 2012

Blue Gene Q

- Transactional Memory
  - Allows a group of load and store instructions to execute in an atomic way.
  - Why is it needed?

```
def transfer_money (from_, to_, amount_)
    with transaction ():
        from_ -= amount_;
        to_ -= amount_;
```

Blue Gene Q

- Hardware Transactional Memory (HTM)
  - Execution Support
    - Long Running mode
    - Short Running Mode
  - Failures
    - Transactional Conflicts
    - Capacity Overflow
    - Jail Mode Violation (JMV)

Blue Gene Q

L1 prefetcher

- Normal mode: Stream Prefetching
- in response to observed memory traffic, adaptively balances resources to prefetch L2 cache lines (@ 128 B wide)
- from 16 streams x 2 deep through 4 streams x 8 deep

Wake-up unit

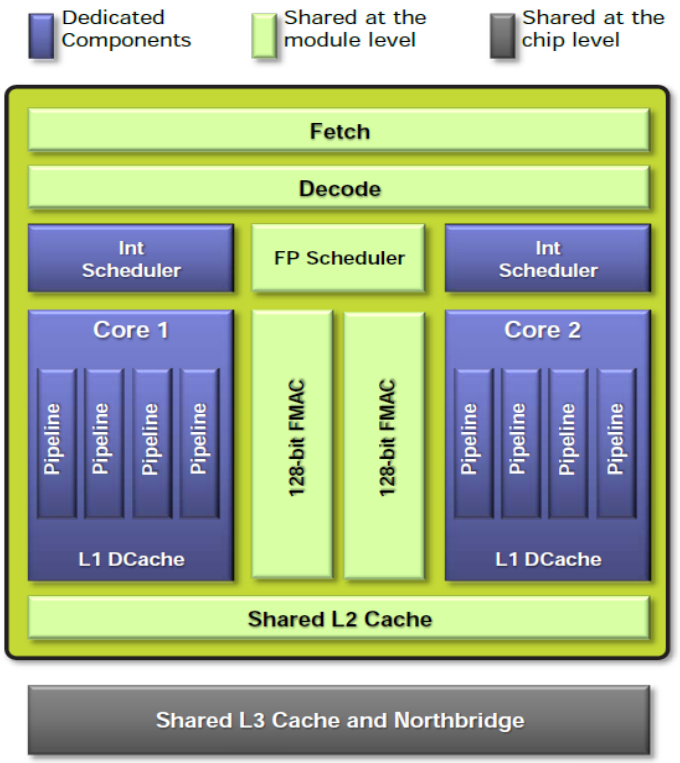
- Will allow SMT threads to be suspended, while waiting for an event
- Lighter weight than wake-up-on-interrupt -- no context switching
- Improves power efficiency and resource utilization

Titan Cray

- An upgrade of Jaguar
- Hybrid CPU/GPU Architecture
- 18,688 Nodes x (16 Core AMD Opteron + 1 Nvidia Tesla)
- Power Usage : 8.2 MW
- Gemini Interconnect Network

Titan Cray

- AMD Opteron



source: AMD

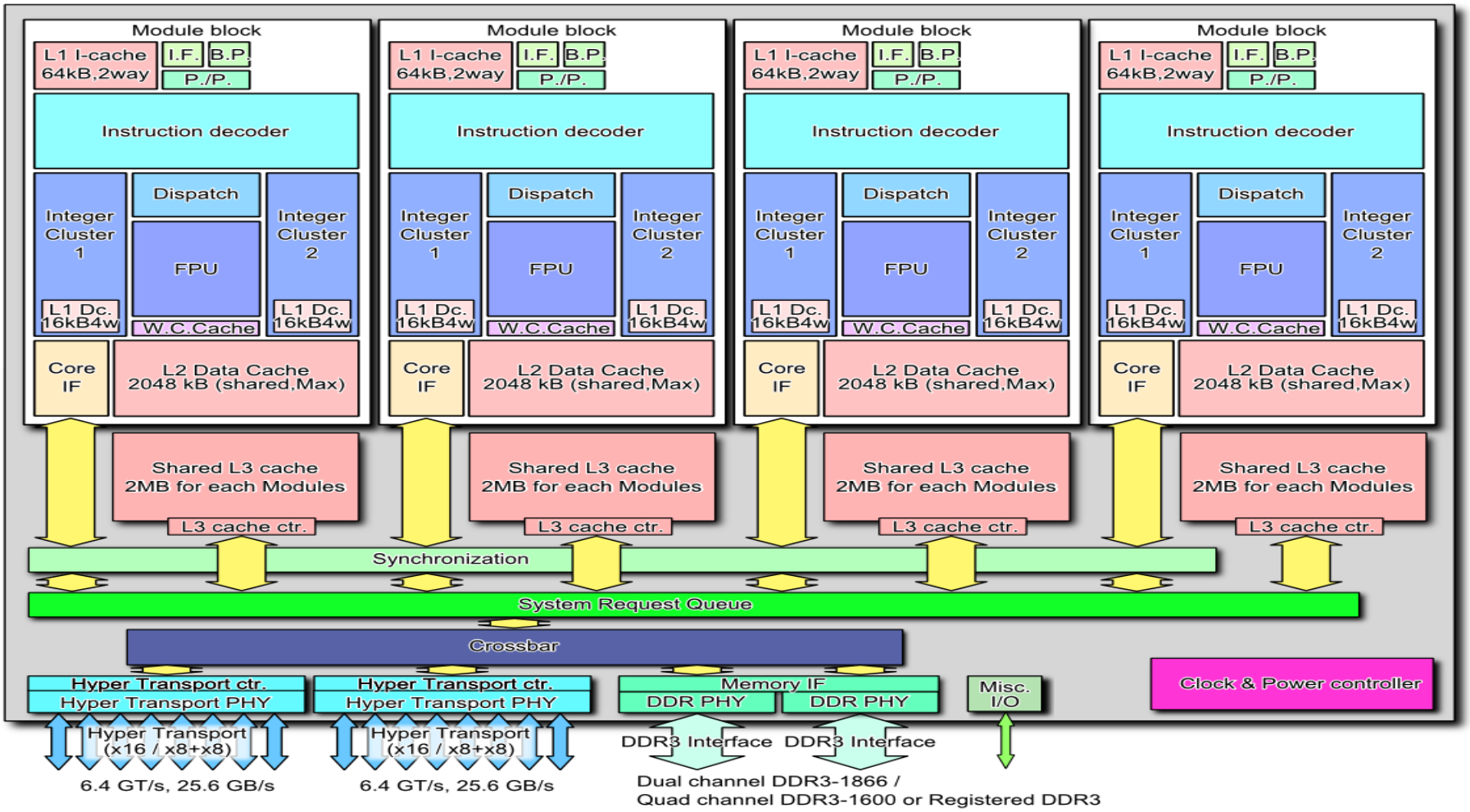
The OS sees the module as 2 cores.
Shared between the cores:
•Instruction Fetch
•Decode
•L1 instruction cache
•L2 cache
•Two 128-bit FMAC floating-point pipelines

Dedicated to each core:
•Integer Scheduler unit
•L1 data cache
•Load Store unit

Chip Level:
•L3 cache
•HyperTransport links

Titan Cray

- Core Architecture (AMD Opteron 6274)



http://en.wikipedia.org

Titan Cray

- NVIDIA Tesla K20X
  - 2, 688 CUDA cores
  - Kepler based architecture
  - Each core @ 732MHz

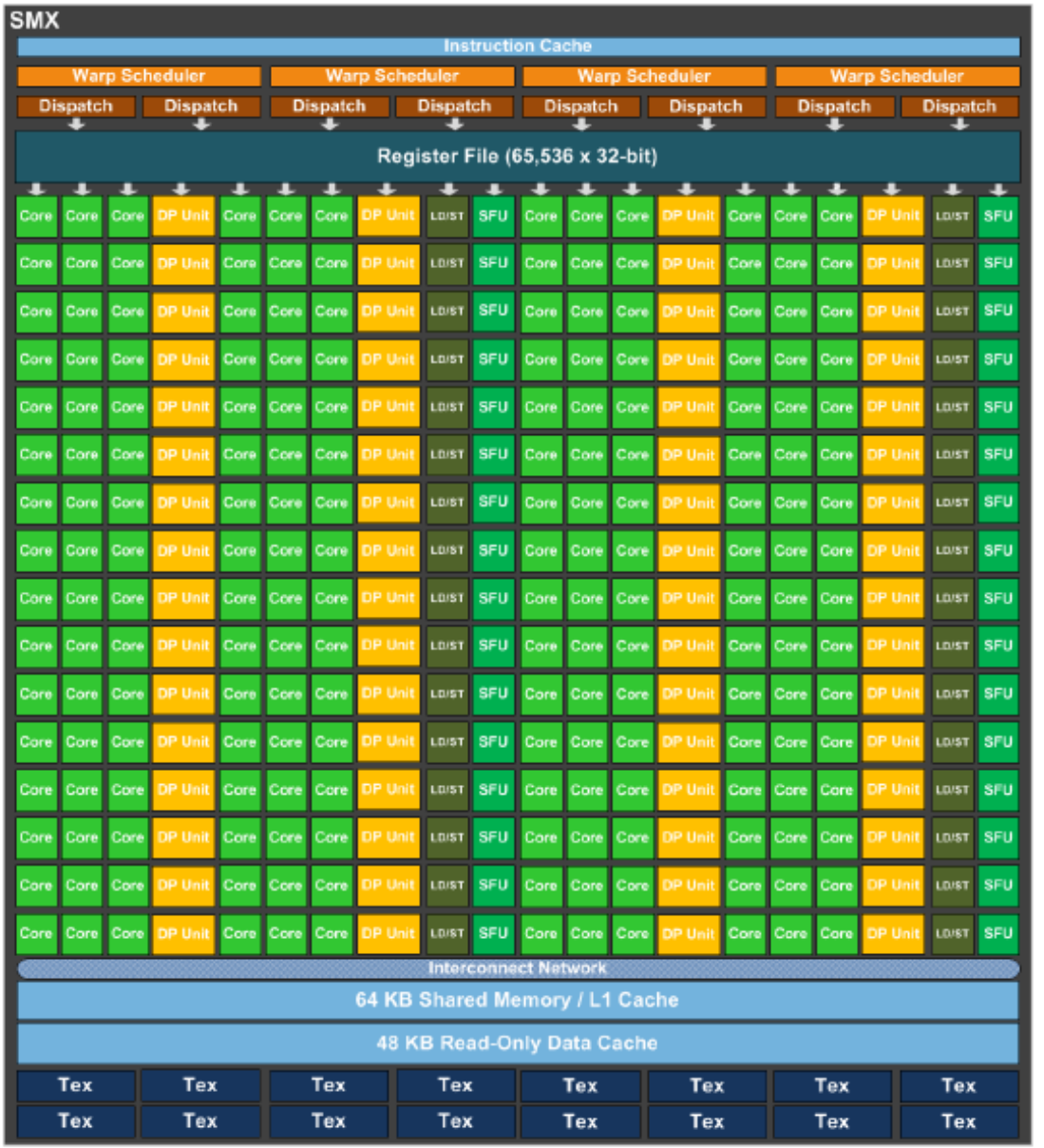
Titan Cray

Kepler GK110 Full chip block diagram

Source: Nvidia

Titan Cray

- Streaming Multiprocessor Architectural Block



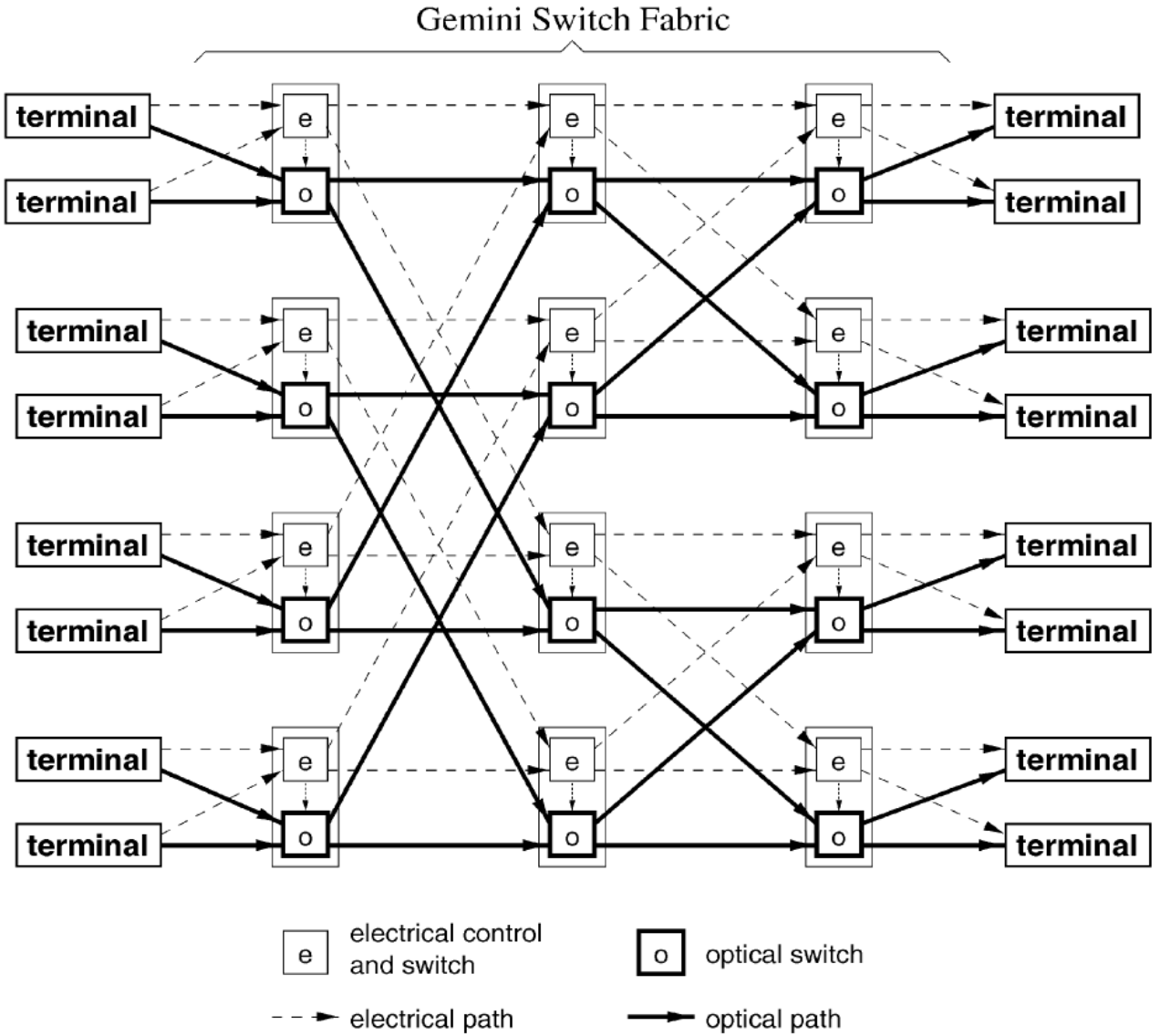
SMX: 192 single-precision CUDA cores, 64 double-precision units, 32 special function units (SFU), and 32 load/store units (LD/ST).

Source: Nvidia

Titan Cray

- Gemini Interconnect Network
  - Optical Network (Speed)
  - Scalable

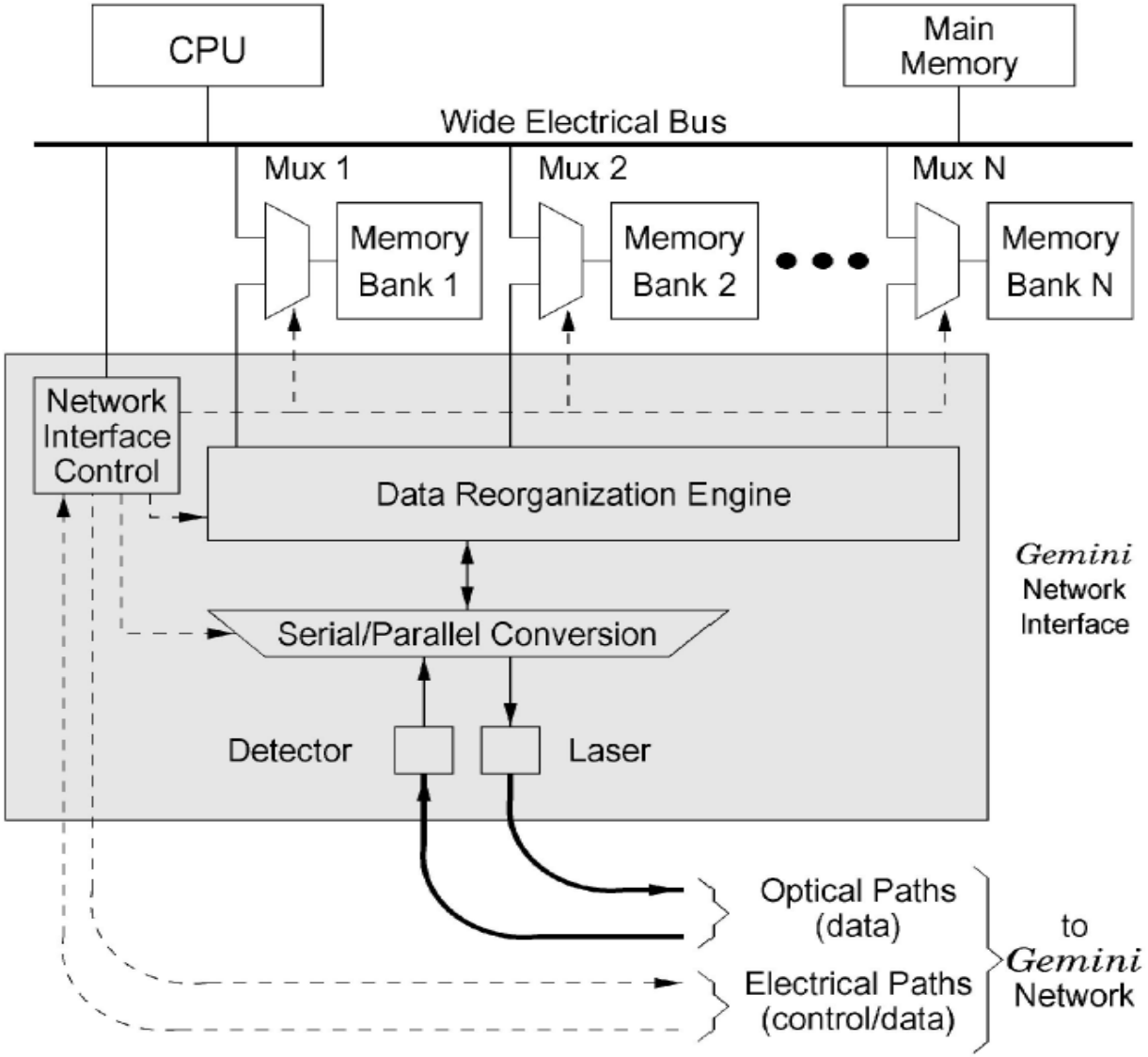
Titan Cray



Ch'ng Shi Baw, Roger D. Chamberlain, Mark A. Franklin, and Michael G.Wrighton, “The Gemini Interconnect: Data Path Measurements and Performance Analysis,” in Proc. of the 6th Int'l Conf. on Parallel Interconnects, October 1999

Titan Cray

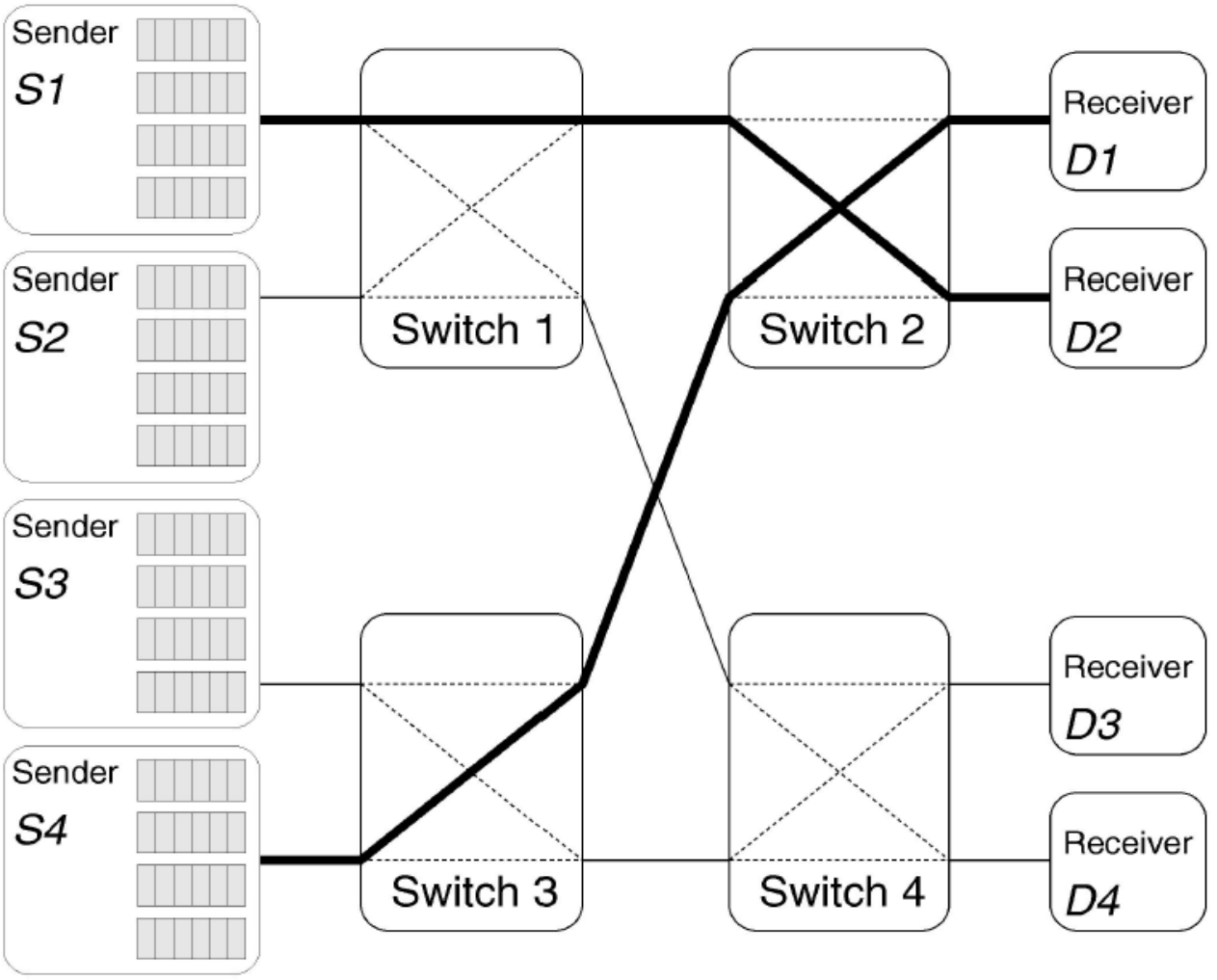
- GEMINI Interconnect Network Interface



Ch'ng Shi Baw, Roger D. Chamberlain, Mark A. Franklin, ar Michael G.Wrighton, “The Gemini Interconnect: Data Path Measurements and Performance Analysis,” in Proc. of the 6th Int'l Conf. on Parallel Interconnects, October 1999

Titan Cray

- GEMINI



Ch'ng Shi Baw, Roger D. Chamberlain, Mark A. Franklin, and Michael G.Wrighton, “The Gemini Interconnect: Data Path Measurements and Performance Analysis,” in Proc. of the 6th Int'l Conf. on Parallel Interconnects, October 1999

Titan Cray

http://en.wikipedia.org

Comparison

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560640	17590.0	27112.5	8209
2	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1572864	16324.8	20132.7	7890
3	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705024	10510.0	11280.4	12660
4	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786432	8162.4	10066.3	3945
5	Forschungszentrum Juelich (FZJ) Germany	**JUQUEEN** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393216	4141.2	5033.2	1970

www.top500.org

References

- IBM BLUE GENE/Q COMPUTE CHIP, Micro, IEEE, 2012
- Ch'ng Shi Baw, Roger D. Chamberlain, Mark A. Franklin, and Michael G.Wrighton, “The Gemini Interconnect: Data Path Measurements and Performance Analysis,” in Proc. of the 6th Int'l Conf. on Parallel Interconnects, October 1999
- Evaluation of Blue Gene/Q Hardware Support for Transactional memories, PACT'12
- https://asc.llnl.gov/computing_resources
- BlueGene/L: the next generation of scalable supercomputer, Lawrence Livermore National Laboratory , P. O. Box 808, Livermore, CA 94551
- www.top500.org
- en.wikipedia.org
- Peter M. Kogge, "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems," Sept. 28, 2008
- www.singularity.com
- “When will computer hardware match the human brain?” , Hans Moravec, Robotics Institute, Carnegie Mellon University, Journal of Evolution and Technology. 1998. Vol. 1