Parallel Computer Architecture II

Stefan Lang

Interdisciplinary Center for Scientific Computing (IWR)
University of Heidelberg
INF 368, Room 532
D-69120 Heidelberg
phone: 06221/54-8264
email: Stefan.Lang@iwr.uni-heidelberg.de

WS 14/15

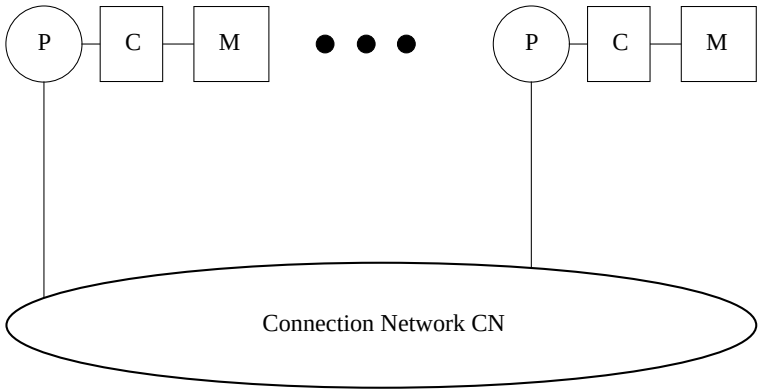
Parallel Computer Architecture II

- Multiprocessor architectures
- Message passing
- Network topologies
- Example architectures
- Routing
- TOP 500
- TOP2 Architectures

Classificaton of MIMD Architectures

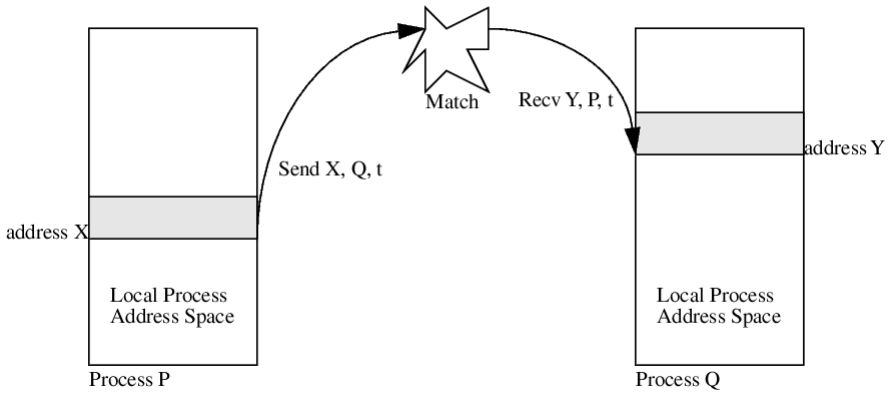
- Physical memory arrangement
  - shared memory
  - distributed memory
- Address space
  - global
  - local
- Programming model
  - shared address space
  - message passing
- Communication structure
  - Memory coupling
  - Message coupling
- Synchronization
  - semaphores
  - barriers
- Latency treatment
  - latency hiding
  - latency minimization

Distributed Memory: MP



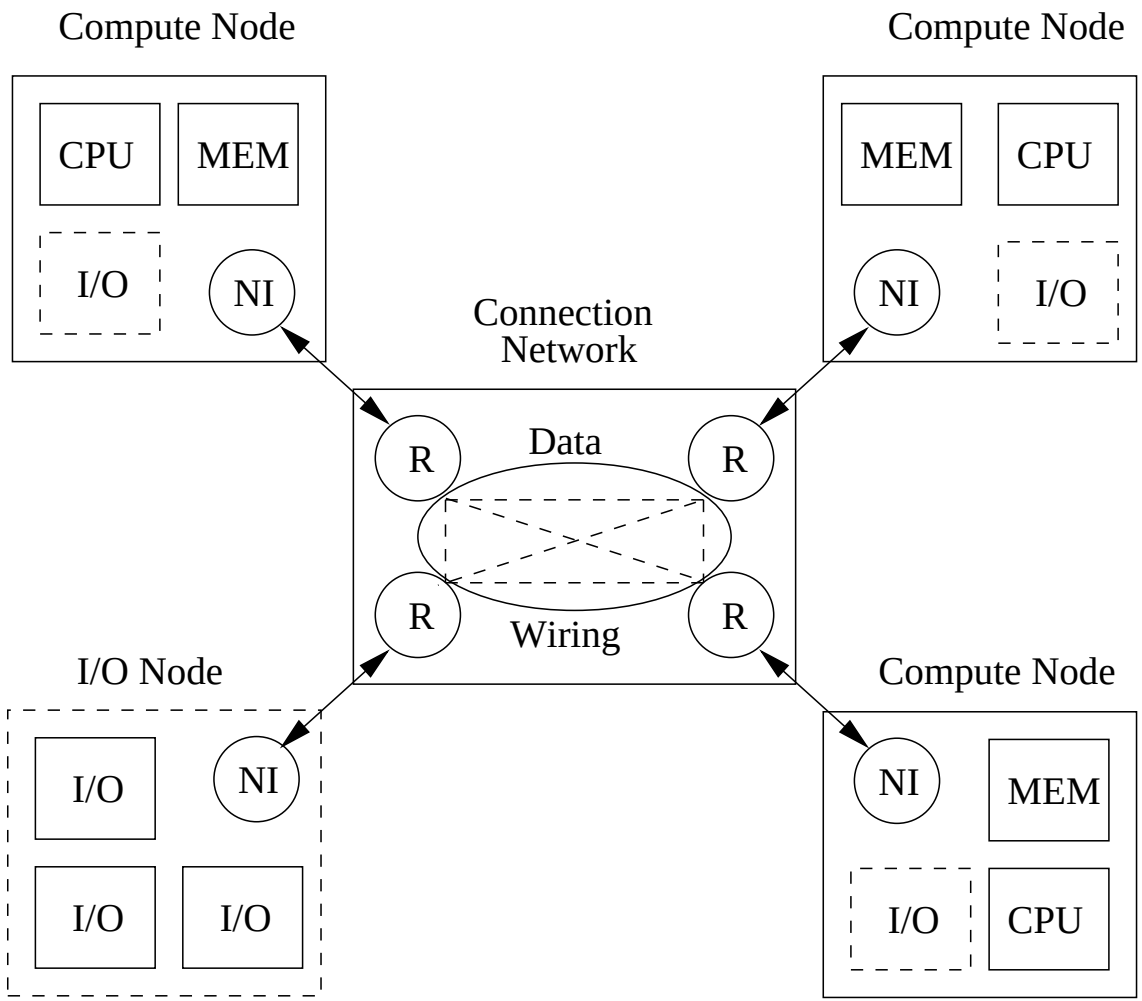
- Multi processors have a *local address space*: Each processor can access only its associated memory.
- Interaction with other processors exclusively by sending of messages.
- Processors, memory and cache are standard components: Full utilization of the price advantage of high quantity of units.
- Connection network ranging from Fast Ethernet to Infiniband.
- Approach with highes scalability: IBM BlueGene > 100 K processors

Distributed Memory: Message Passing



- Processes communicate data between distributed address spaces
- Explicit message passing is necessary
- Send-/Receive operations

A Generic Parallel Computer Architecture



Generic approach of a scalable parallel computer with distributed memory

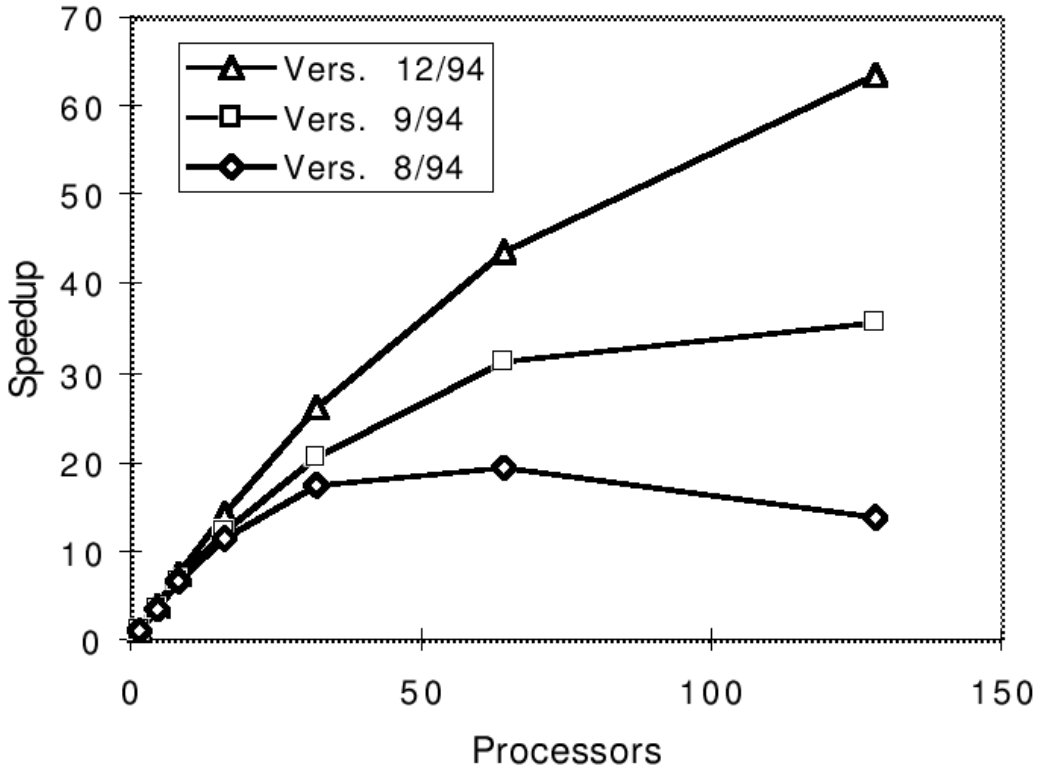
Scalability Parameters

Parameters, that drive the scalability of a computing system:

- Bandwidth [MB/s]
- Latency [μs]
- Costs [$]
- Physical size [m^2,m^3]
- Power consumption [W]
- Fault tolerance / recovery abilities

A truely scalable architecture should avoid any hard limits!

Influence of Parallel Software?

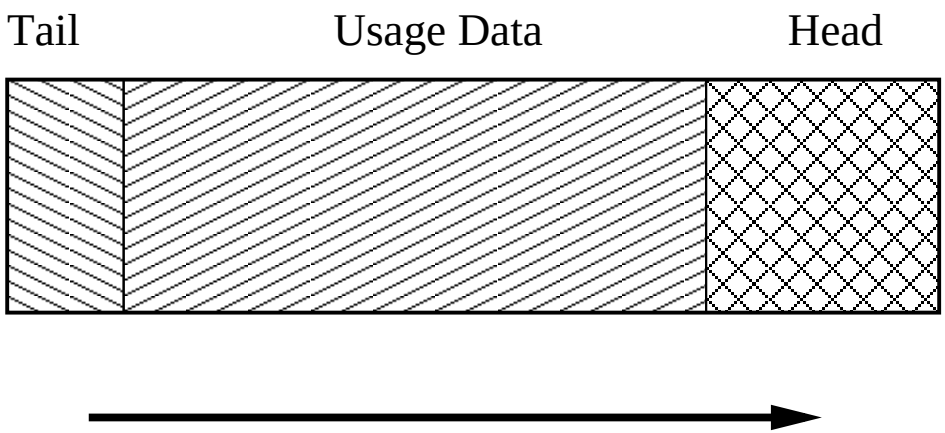


from Culler, Singh, Gupta: Parallel Computer Architecture

Although the parallel architecture scales scalable software is prerequisite for scalable numerical computation.

Message Passing

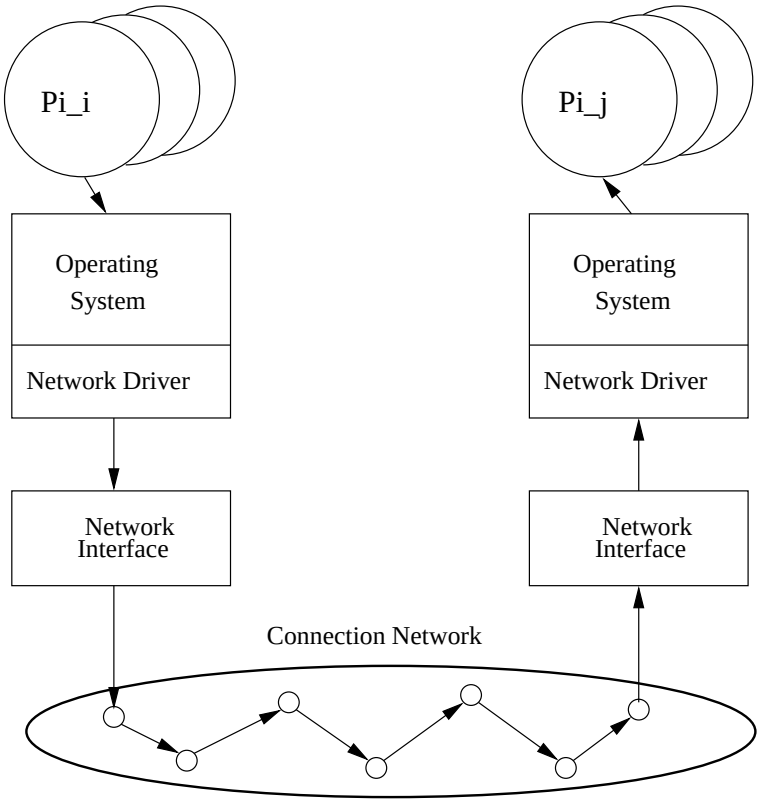
Memory block (of variable length) shall be copied from one memory to another
More precise: From the address space of one process to the address space of another process (running on a different processor)
The connection network is packet oriented. Each message is subdivided in packets of fixed length (e. g. 32 byte to 4 kbyte)



Head: target processor, tail: check sum
Communication protocol: acknowledgment whether packet has arrived valid, flow control

Message Passing

Layer model (Hierarchy of protocols):

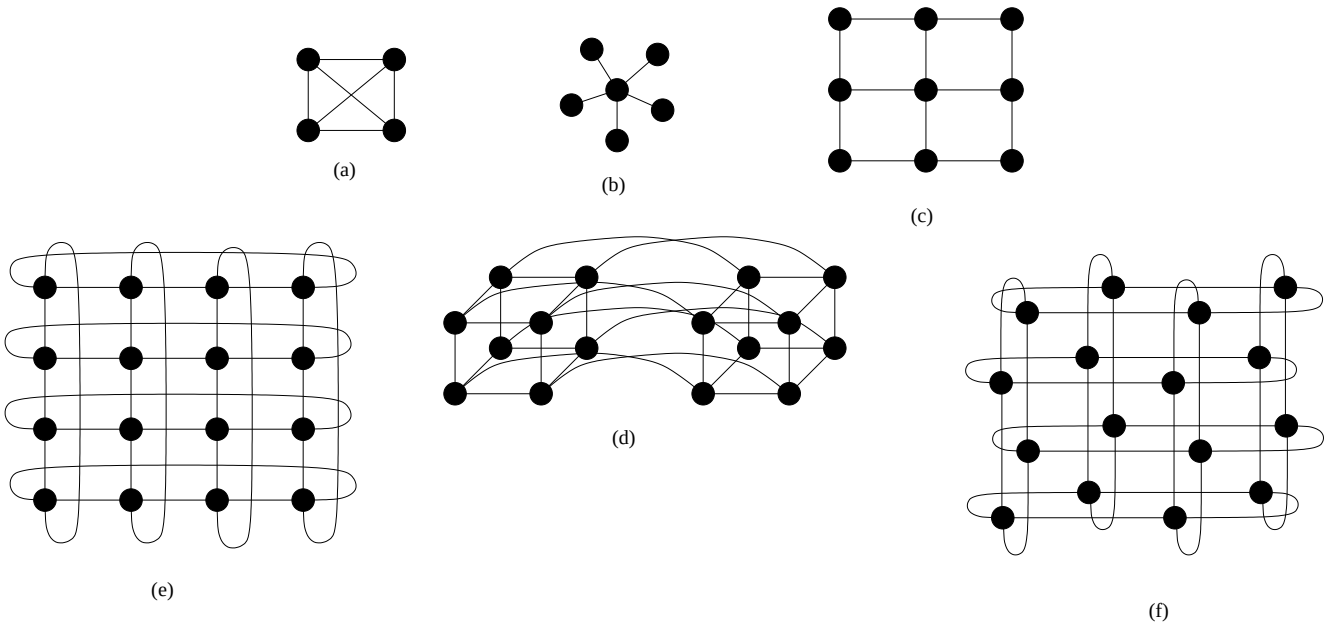


Model of transmission time:

$$t_{mess}(n) = t_s + n * t_b.$$

t_s: setup time (latency), t_b: time per byte, $1/t_b$: bandwidth, dependent on protocol

Network Topology I



(a) full connected, (b) star, (c) array
(d) hypercube, (e) torus, (f) folded torus

- *Hypercube*: of dimension d has 2^d processors. Processor p is connected to q if their binary representation differs *in exactly one bit*.
- Network node: Earlier (before 1990) this was the processor itself, nowadays it is a dedicated communication processor.

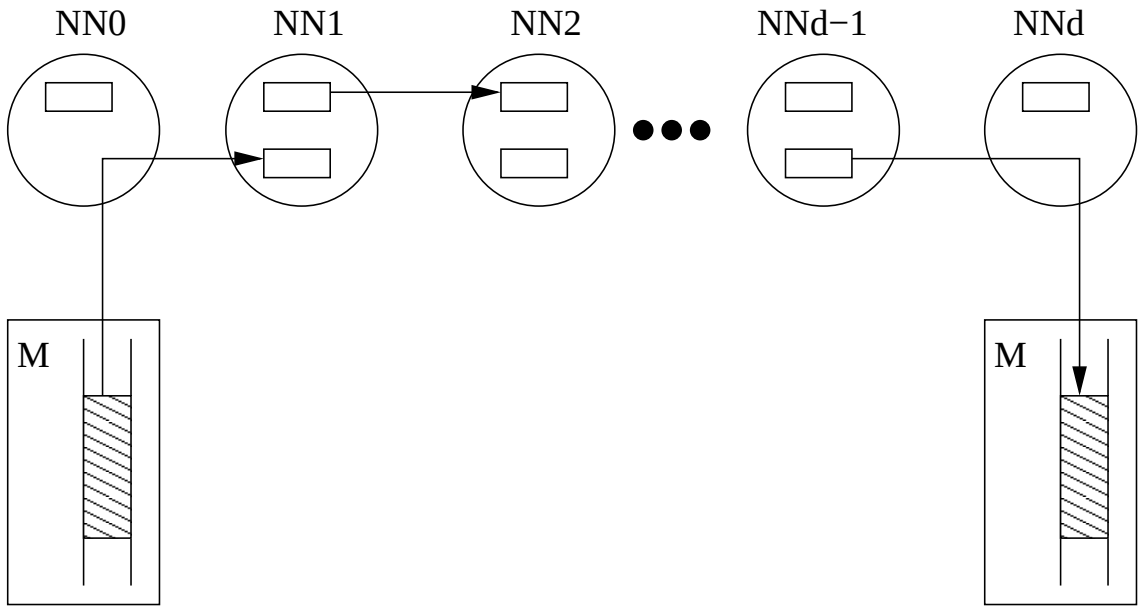
Network Topology II

Reference parameters:

Network topology	Node-degree K	Wire-count L	Dia-meter D	Bisection-bandwidth B	Sym-metry
Full Connectivity	$N-1$	$N(N-1)/2$	1	$(N/2)^2$	yes
Star	$N-1$	$N-1$	2	$\lfloor N/2 \rfloor$	no
2D Grid	4	$2N-2\sqrt{N}$	$2(\sqrt{N}-1)$	$\sqrt{N}$	no
3D Torus	6	$3N$	$3\lfloor \sqrt{N}/2 \rfloor$	$2\sqrt{N}$	yes
Hypercube	$\log_2 N$	$nN \log_2 N$	n	$N/2$	yes
k-ary n-cube ($N = k^n$)	$3N$	$n\lfloor k/2 \rfloor$	nN	$2k^{n-1}$	yes

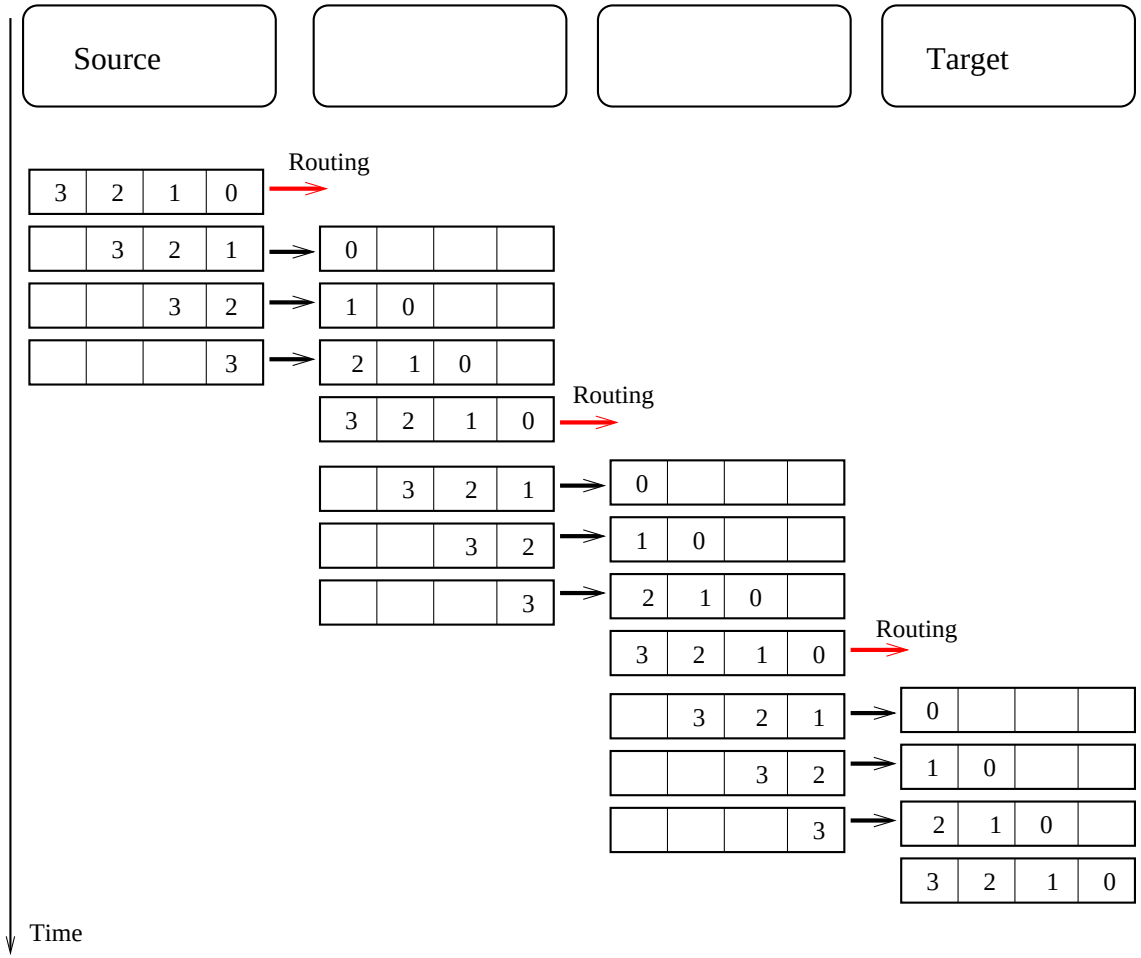
Store & Forward Routing

Store-and-forward routing: Message of length n is subdivided into packets of length N. Pipelining on packet level: Packet is stored completely in the network node.



Store & Forward Routing

Transmission of a packet:



Store & Forward Routing

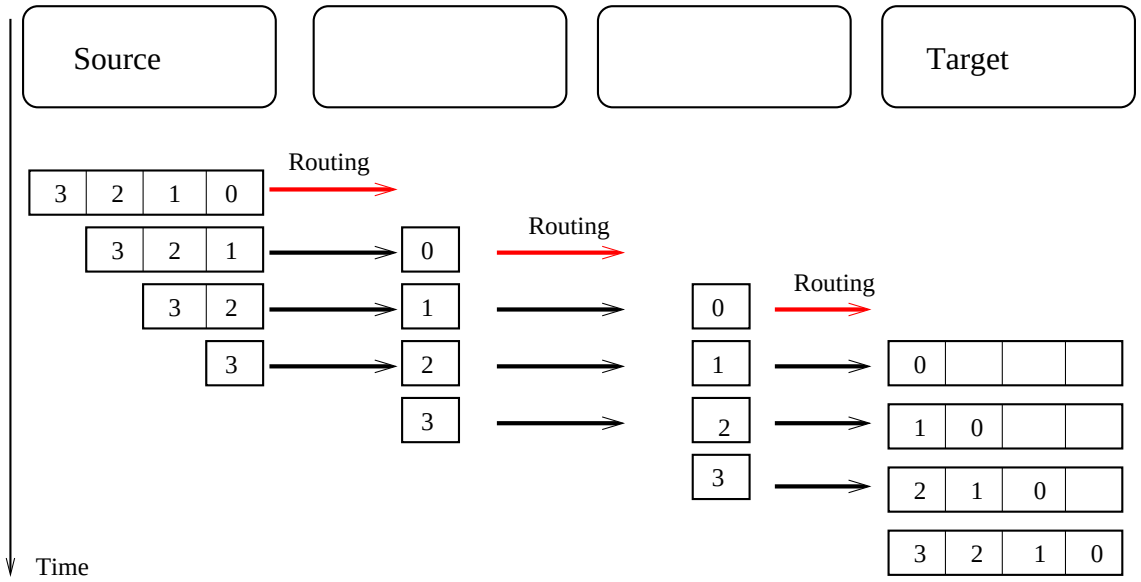
Run time:

$$t_{SF}(n, N, d) = t_s + d(t_h + Nt_b) + \left(\frac{n}{N} - 1\right)(t_h + Nt_b)$$
$$= t_s + t_h\left(d + \frac{n}{N} - 1\right) + t_b\left(n + N(d-1)\right).$$

- t_s: time, that is needed on source and target computer until the network is instructed with the message transmission, respectively until the receiving process is informed. This is the software share of the protocol.
- t_h: time that is necessary to transmit the first byte of a message from a network node to another one (node latency, hop–time).
- t_b: time for transmission of a byte from network node to network node
- d: Hops to reach the target node.

Cut-Through Routing

Cut-through routing or *wormhole routing*: Packets are not buffered, each word (so called *flit*) is routed immediately to the next network node.
Transmission of a packet:



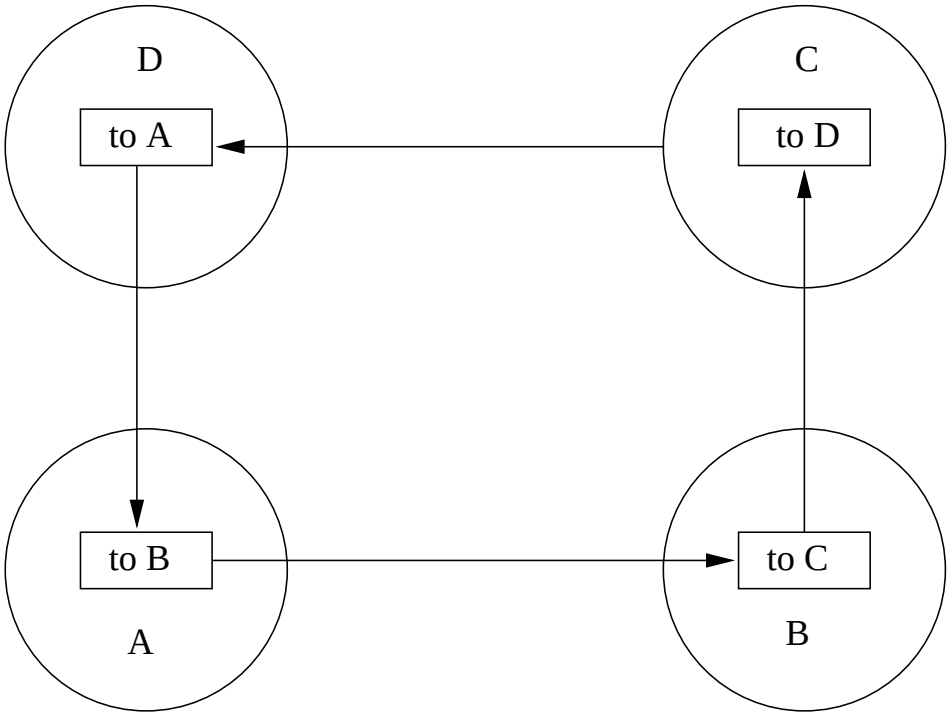
Run time:

$$t_{CT}(n, N, d) = t_s + t_h d + t_b n$$

Time for short message ($n = N$): $t_{CT} = t_s + dt_h + Nt_b$. Because of $dt_h \ll t_s$ (Hardware!) nearly distance independent

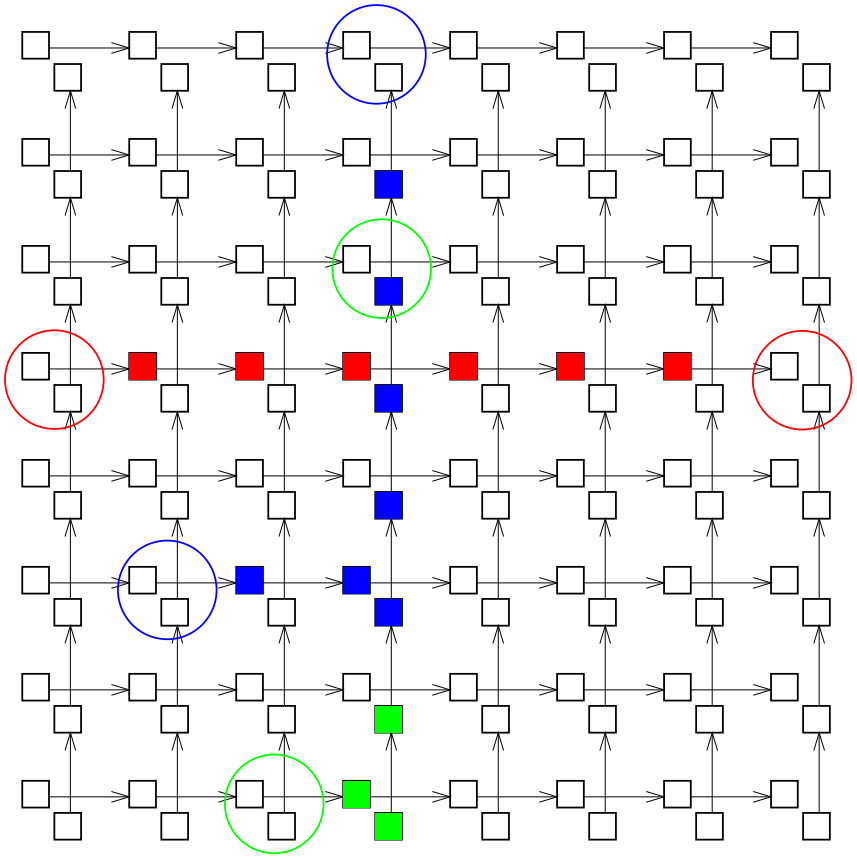
Deadlock

In packet transmiting network the danger of a *store-and-forward deadlock* exists:



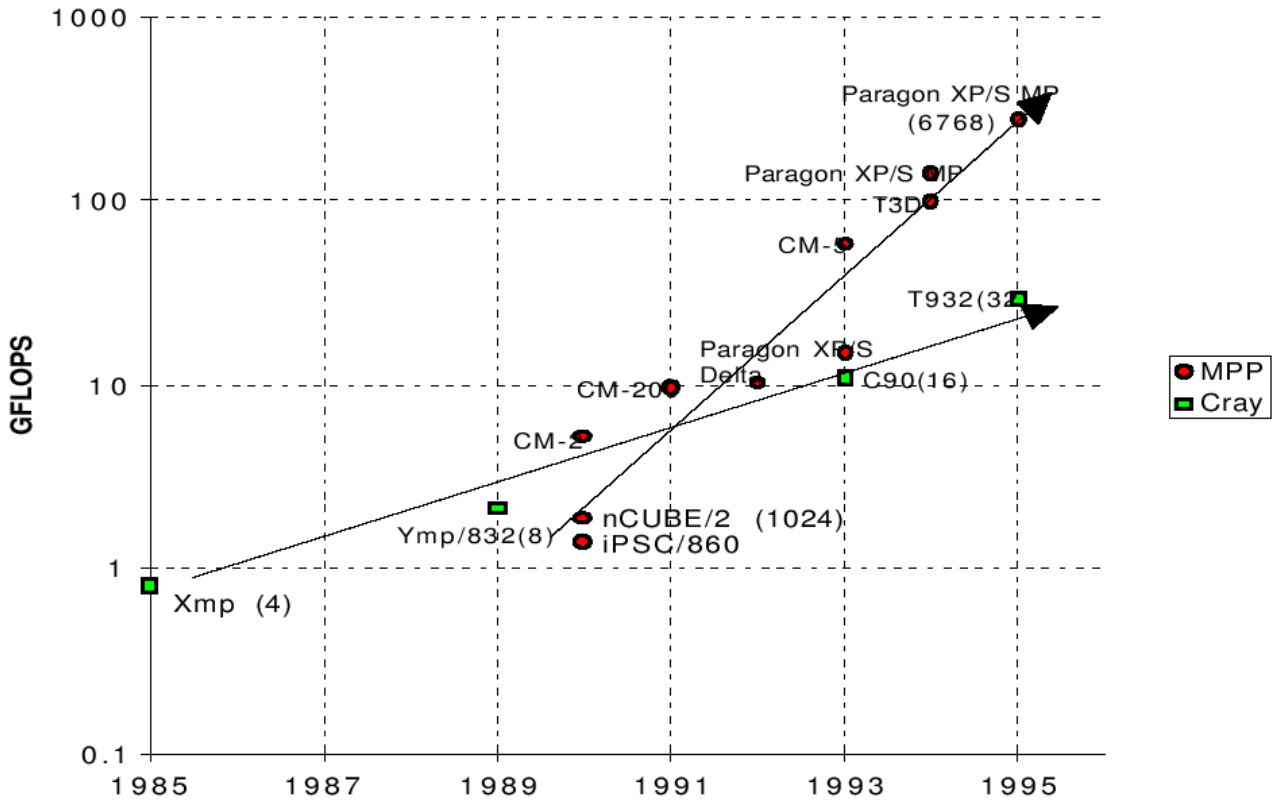
Deadlock

Together with cut-through routing:



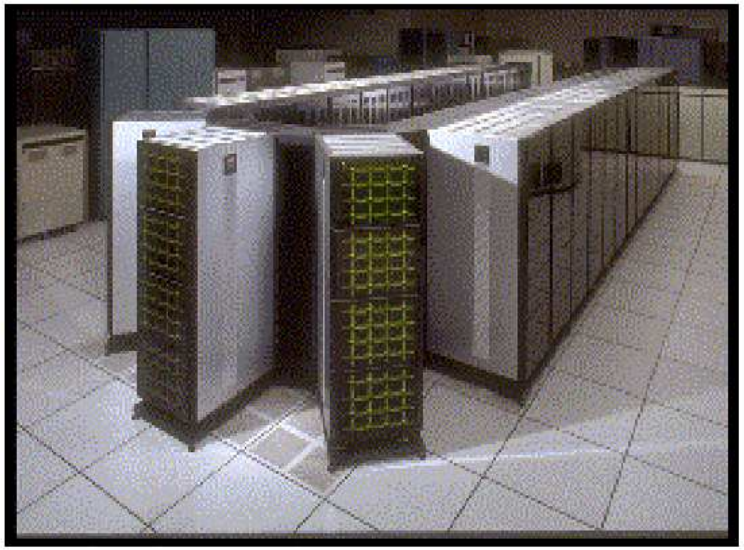
Deadlock free „dimension order routing“.
Example 2D grid: Partition network into $+x$, $-x$, $+y$ and $-y$ networks each with individual buffers. Message is routed first in the sender row, then in the receiver column.

Multi-Processor Performance

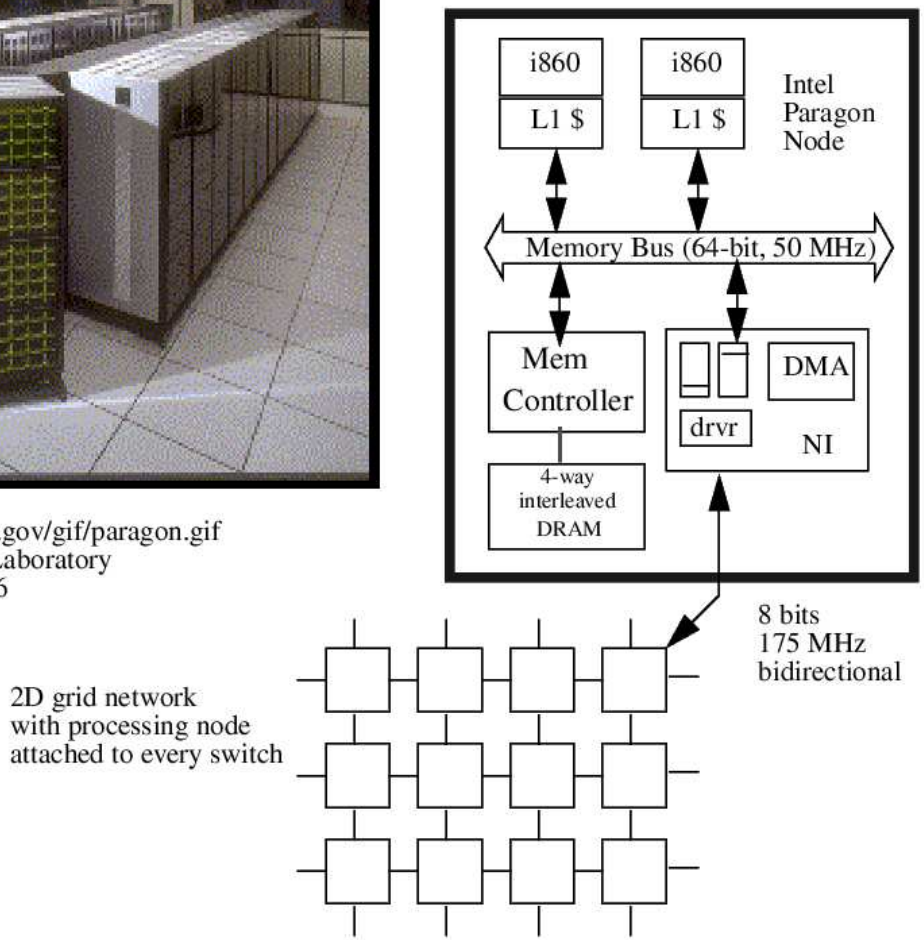


from Culler, Singh, Gupta: Parallel Computer Architecture

Message Passing Architectures I



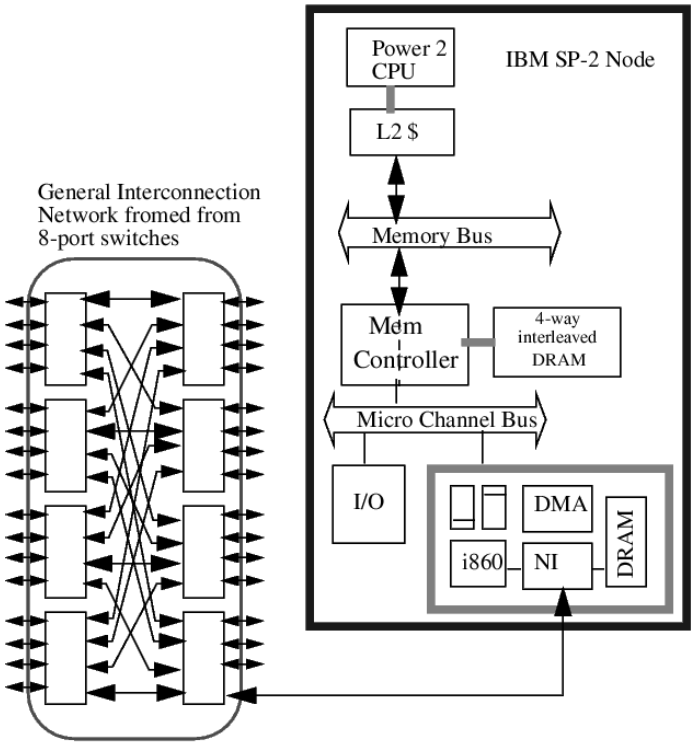
Source: http://www.cs.sandia.gov/gif/paragon.gif
courtesy of Sandia National Laboratory
1824 nodes configured as a 16
high by 114 wide array



Intel Paragon:

- First machine with parallel Unix
- Process migration, gang scheduling

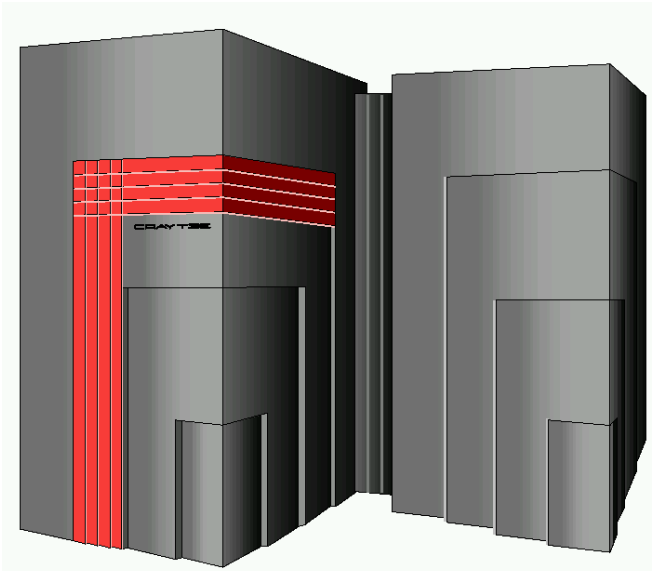
Message Passing Architectures II

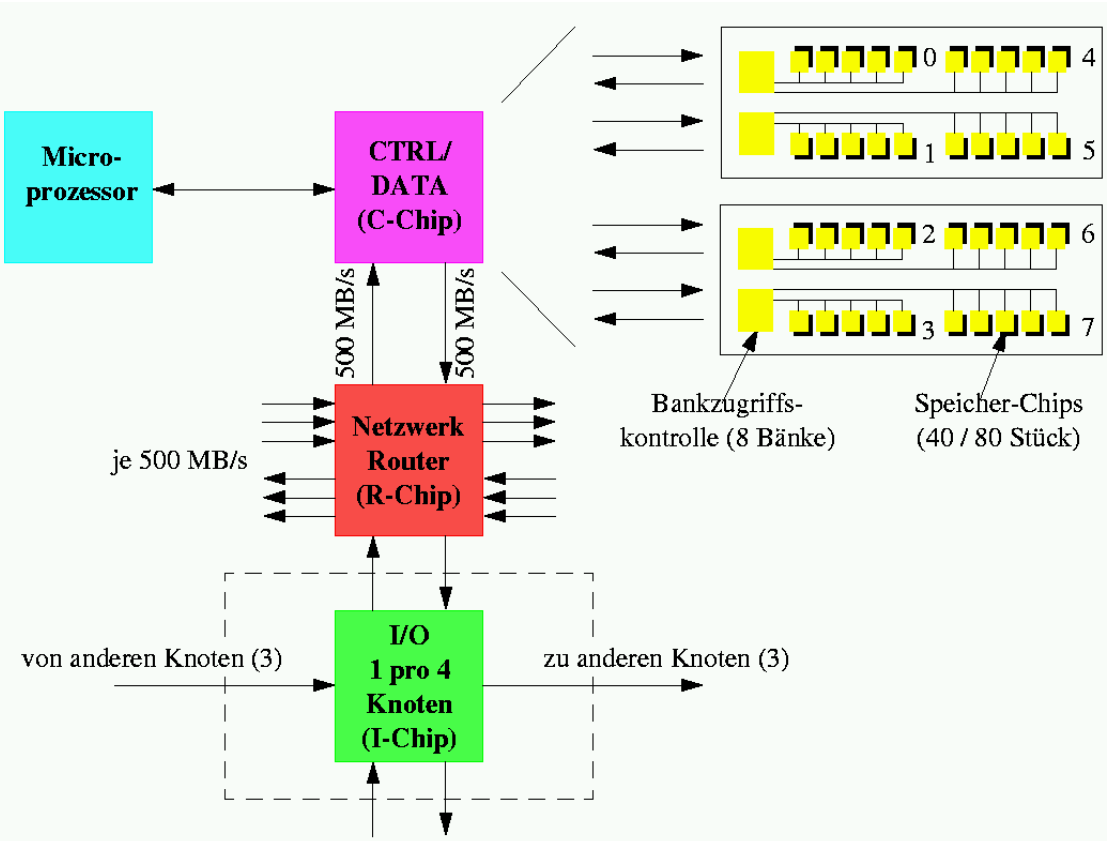


IBM SP2:

- Compute nodes are RS 6000 workstations
- Switching network

Message Passing Architectures III





Cray T3E:

- high package density
- a single system wide clock
- virtual shared memory

Top500

Top500 Benchmark:

- LINPACK benchmark is used for evaluation of the systems
- Benchmark performance does not reflect the overall performance of the system
- Benchmark indicates performance during solution of dense linear equation sytem
- Very regular problem: high acchievable performance (near peak performance)

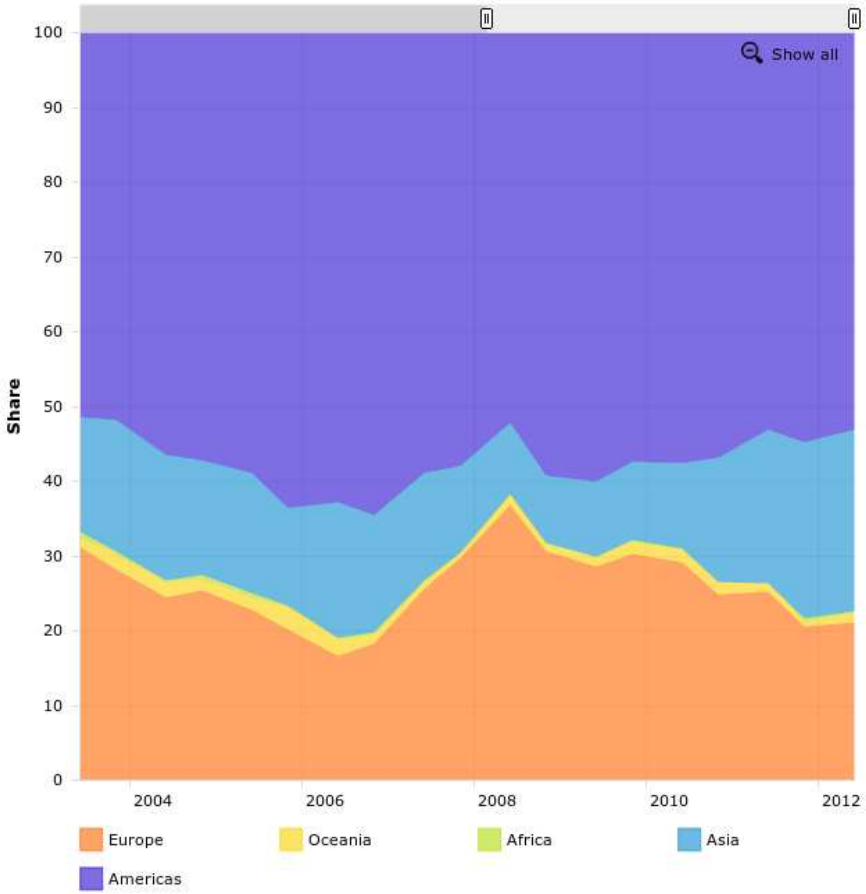
Top 10 of Top500

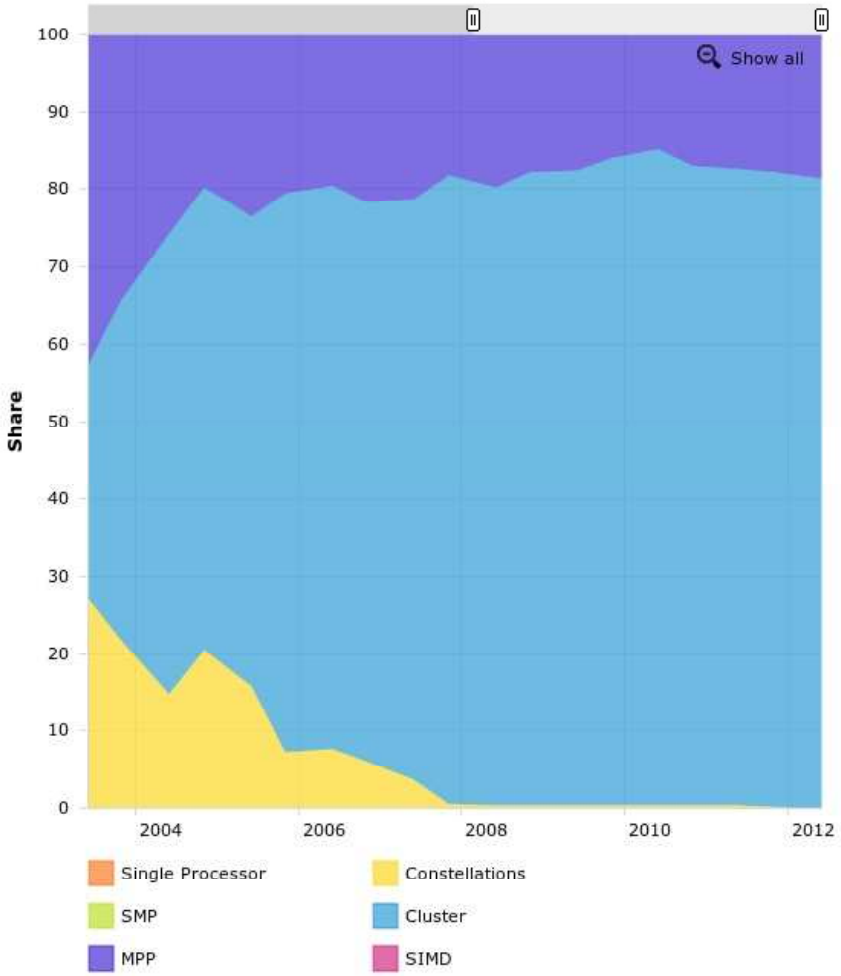
Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National University of Defense Technology China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31 S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0 GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945
6	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462,462	5,168.1	8,520.1	4,510
7	Forschungszentrum Juelich (FZJ) Germany	**JUQUEEN** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	458,752	5,008.9	5,872.0	2,301
8	DOE/NNSA/LLNL United States	**Vulcan** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393,216	4,293.3	5,033.2	1,972
9	Leibniz Rechenzentrum Germany	**SuperMUC** - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR IBM	147,456	2,897.0	3,185.1	3,423
10	National Supercomputing Center in Tianjin China	**Tianhe-1A** - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 NUDT	186,368	2,566.0	4,701.0	4,040

Top500 key facts

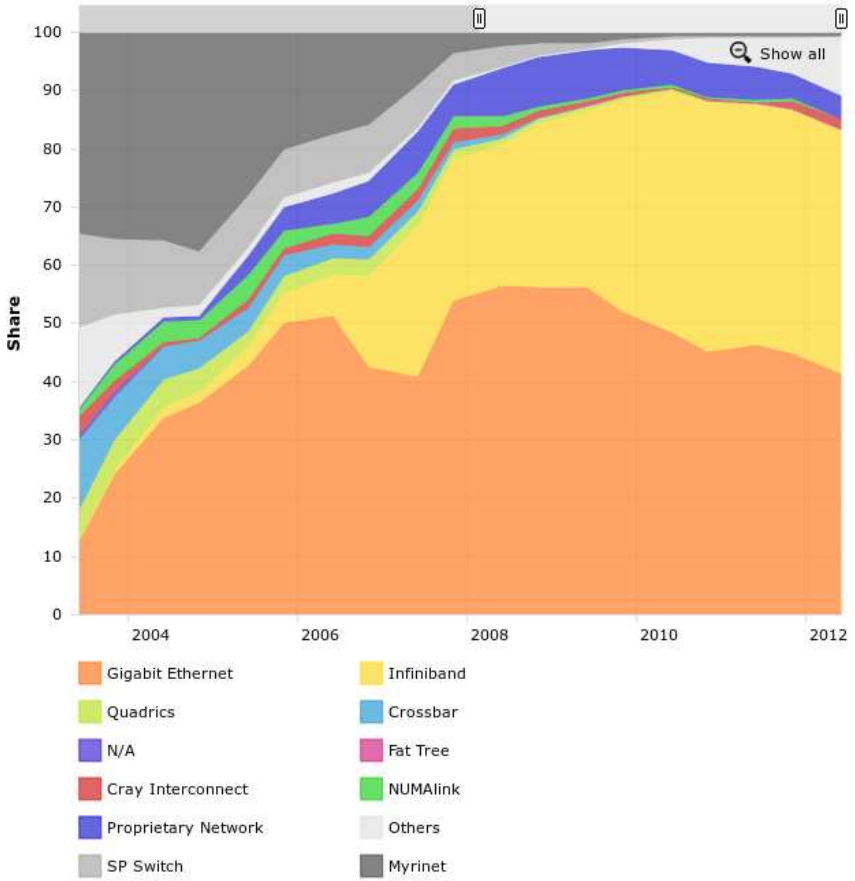
- Entry barrier is a performance of 60.8 TeraFlop/s
- Mean energy consumption of Top10 is 4.09 MW: 0.8-2 GFlops/W
- 40 systems need more than 1 MW
- Accumulated performance is 123.4 PFlops/s (74.2 PFlop/s)
- Top100 minimal performance is 172.6 TFlop/s (115.9 TFlop/s)
- 20 Petaflops systems
- Top500 minial performance is 12.97 TFlop/s (9.29 TFlop/s)
- Processor type: Intel SandyBridge, AMD Opteron, IBM Power 68
- 74.8% of the systems have processors with 6 or more cores
- Infiniband (208) and Gigabit Ethernet (207) networks dominate
- Architecture: 80% Cluster, 20% MPP, 0% SIMD/SMP

Top500 Continent + Archtecture Type

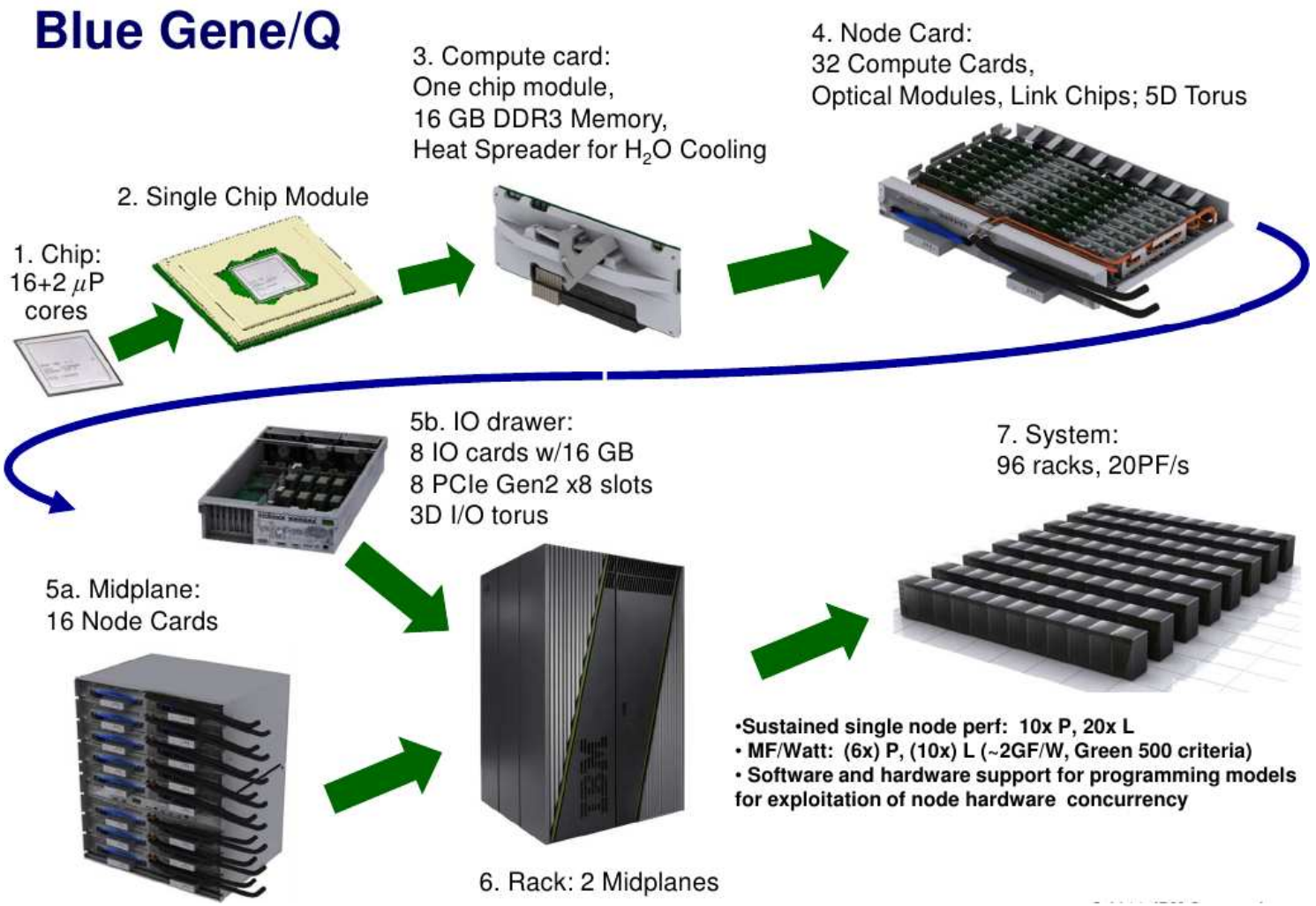




Top500 CoresPerSocket + Interconnect

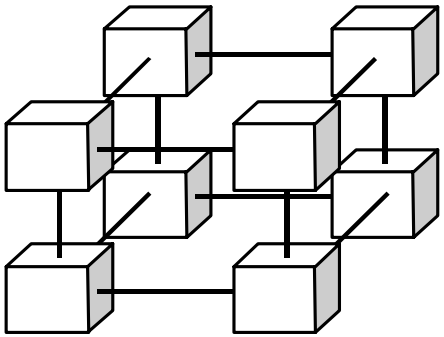


IBM Blue Gene/Q Architecture



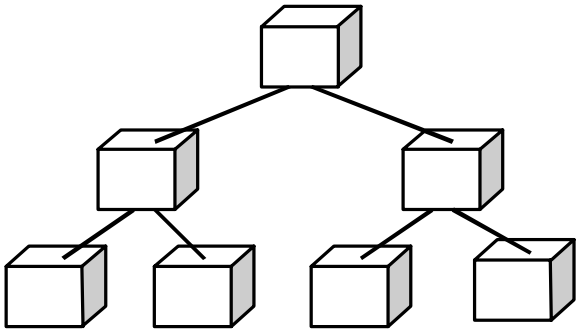
IBM Blue Gene/Q Networks

98304 node are connected by three distinct, integrated networks



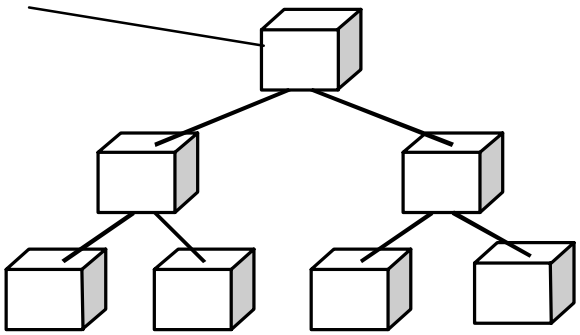
3 Dimensional Torus

- Virtual cut-through hardware routing to maximize efficiency
- 2.8 Gb/s on all 12 node links (total of 4.2 GB/s per node)
- Communication backbone
- 134 TB/s total torus interconnect bandwidth



Global Tree

- One-to-all or all-all broadcast functionality
- Arithmetic operations implemented in tree
- ~1.4 GB/s of bandwidth from any node to all other nodes
- Latency of tree traversal less than 1usec

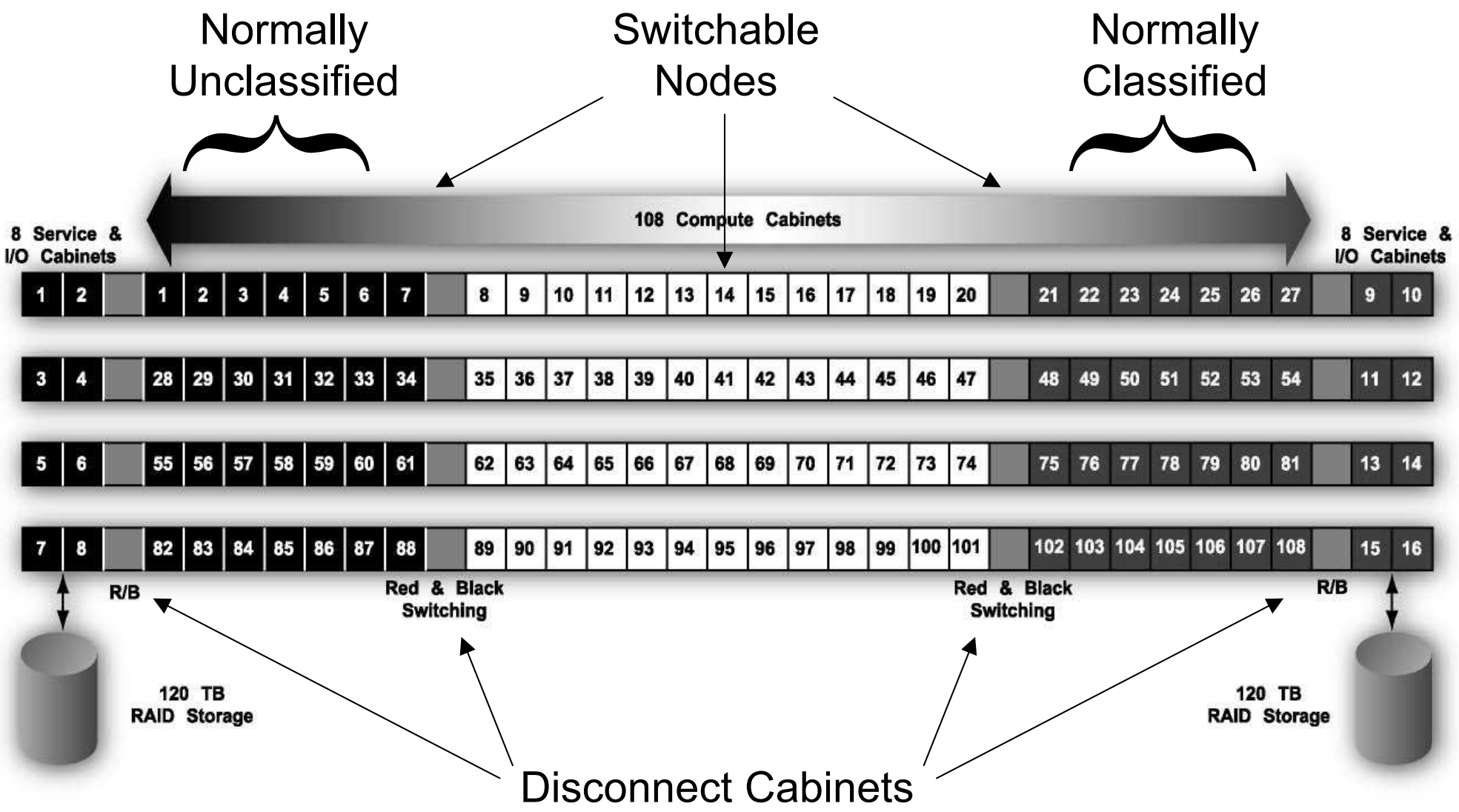


Ethernet

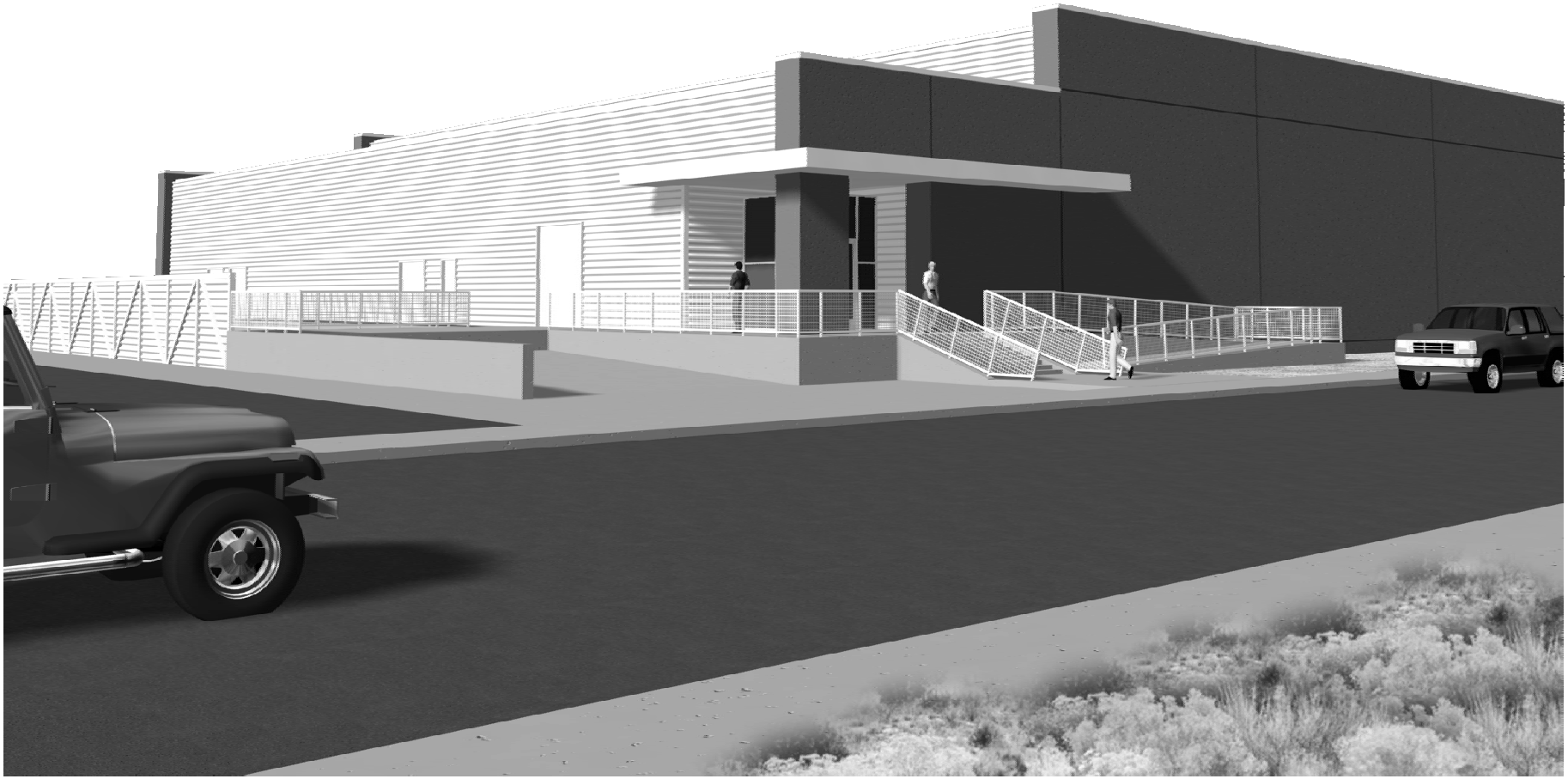
- Incorporated into every node ASIC
- Disk I/O
- Host control, booting and diagnostics

Cray RedStorm

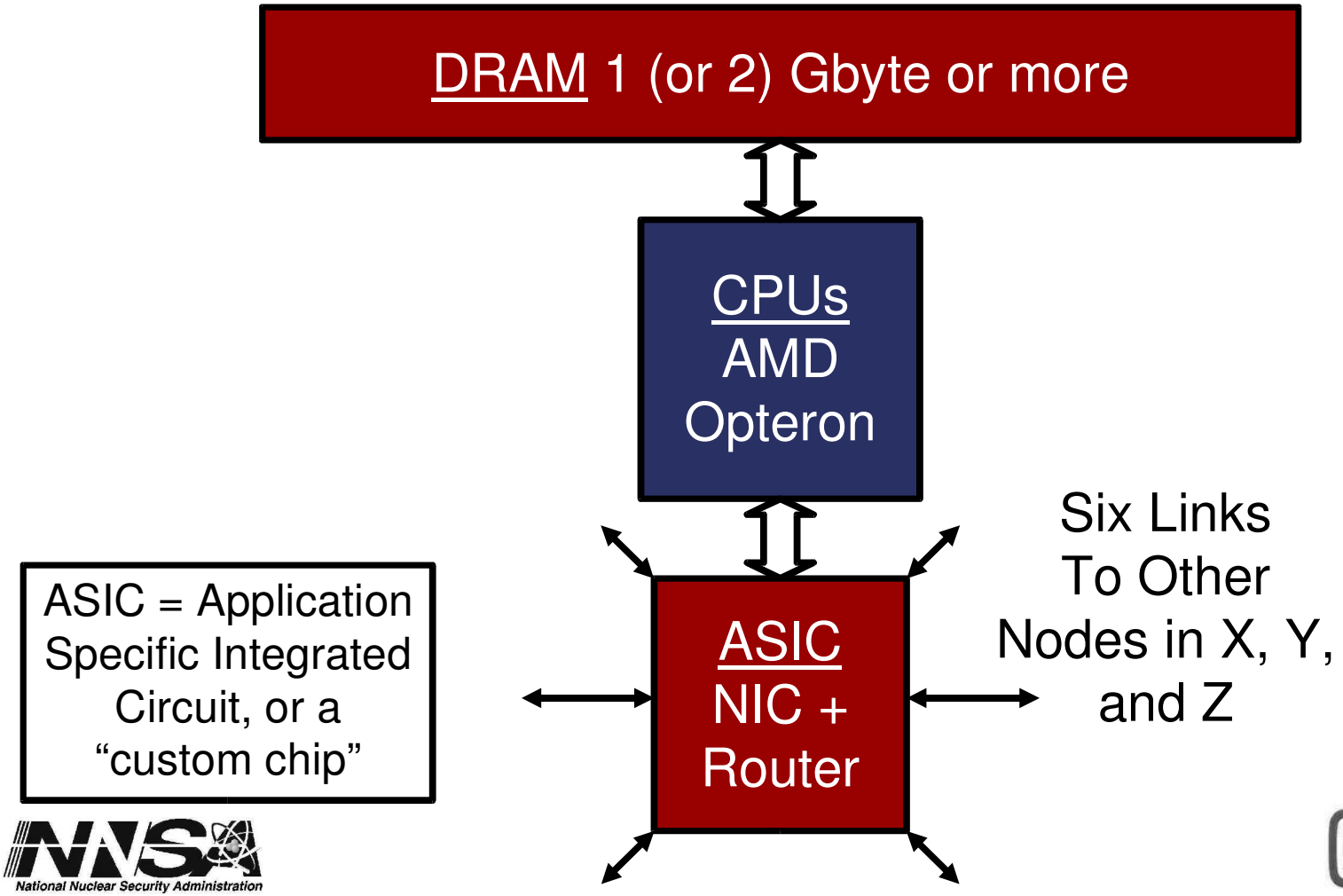
Cray RedStorm Configuration



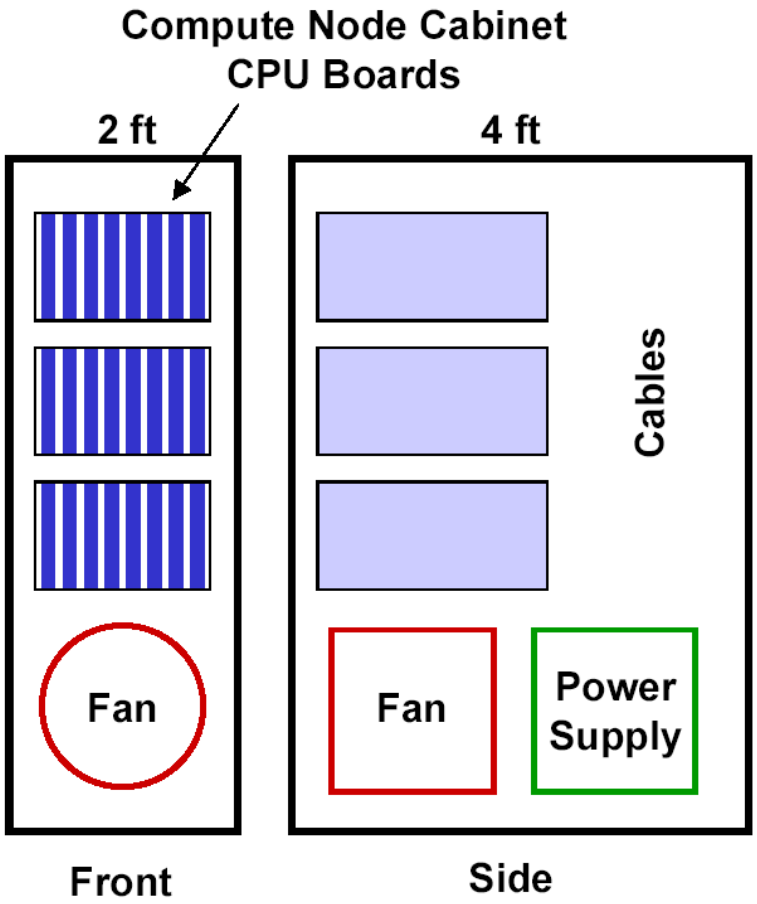
Cray RedStorm Building



Cray RedStorm Compute Node



Cray RedStorm Cabinet



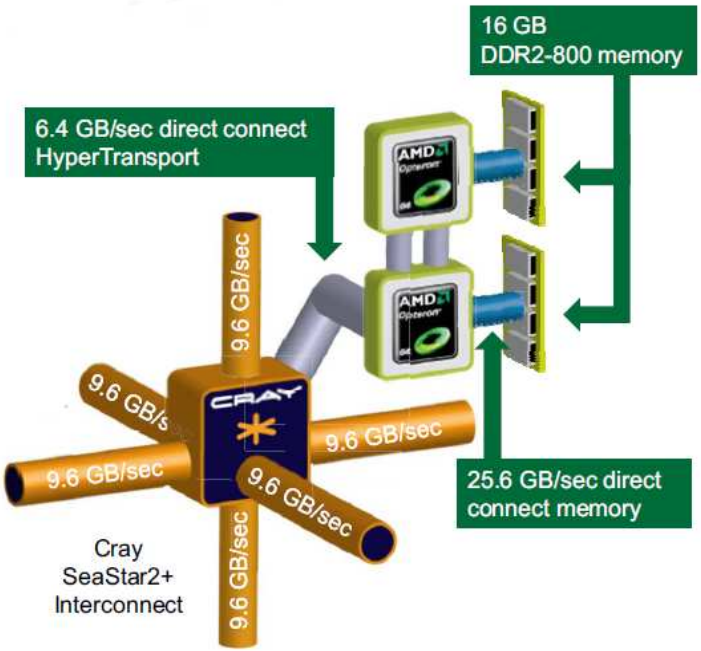
- Compute Node Cabinet
  - 3 Card Cages per Cabinet
  - 8 Boards per Card Cage
  - 4 Processors per Board
  - 4 NIC/Router Chips per Board
  - N + 1 Power Supplies
  - Passive Backplane
- Service and I/O Node Cabinet
  - 2 Card Cages per Cabinet
  - 8 Boards per Card Cage
  - 2 Processors per Board
  - 4 NIC/Router Chips per Board
  - Dual PCI-X for each processor
  - N + 1 Power Supplies
  - Passive Backplane

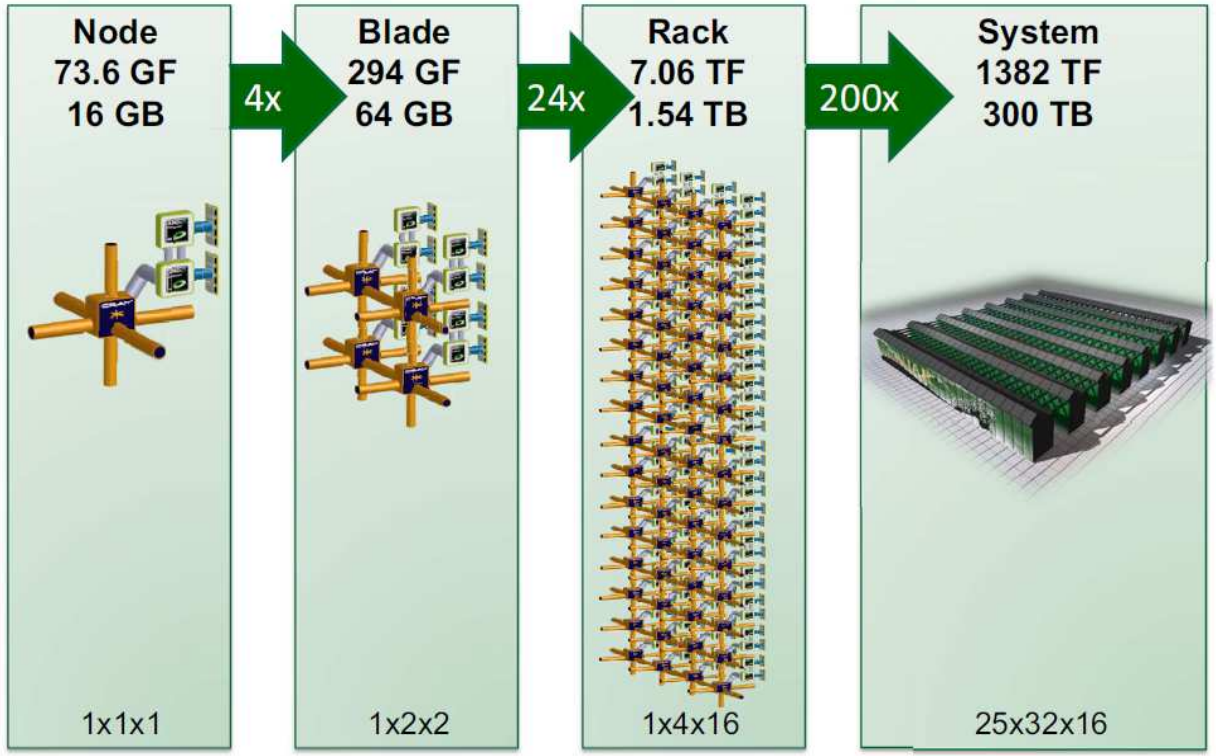
Blue Gene L vs Red Storm

BGL 360 TF version, Red Storm 100 TF version

	Blue Gene L	Red Storm	
Node speed	5.6 GF	5.6 GF	(1x)
Node memory	.25 - .5 GB	2 (1-8 GB)	(4x)
Network latency	7 us	2 us	(2/7x)
Network link bw	0.28 GB/s	6.0 GB/s	(22x)
BW Bytes/Flops	0.05	1.1	(22x)
Bi-Section B/F	0.0016	0.038	(24x)
#nodes/problem	40,000	10,000	(1/4x)

Cray XT-5 Jaguar Architecture





Cray XT-5 Jaguar IO-Configuration