

Parallele Rechnerarchitektur II

Stefan Lang

Interdisziplinäres Zentrum für Wissenschaftliches Rechnen
Universität Heidelberg
INF 368, Raum 532
D-69120 Heidelberg
phone: 06221/54-8264
email: Stefan.Lang@iwr.uni-heidelberg.de

WS 11/12

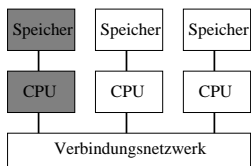


Parallele Rechnerarchitektur II

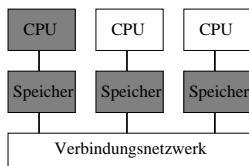
- Multiprozessor Architekturen
- Nachrichtenaustausch
- Netzwerktopologien
- Architekturbeispiele
- Routing
- TOP 500
- TOP2 Architekturen



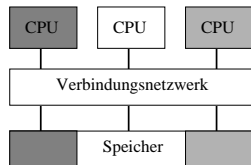
Kommunikationsarchitekturen



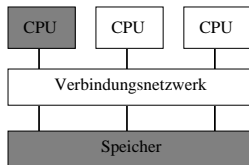
(a) verteilte Speicherorganisation,
lokaler Adressraum



(b) verteilte Speicherorganisation,
globaler Adressraum



(c) zentrale Speicherorganisation,
lokaler Adressraum



(d) zentrale Speicherorganisation,
globaler Adressraum

architecture types distinguished by memory organization



Einteilung von MIMD-Architekturen

- physische Speicheranordnung

- ▶ gemeinsamer Speicher
- ▶ verteilter Speicher

- Adressraum

- ▶ global
- ▶ lokal

- Programmiermodell

- ▶ gemeinsamer Adressraum
- ▶ Nachrichtenaustausch

- Kommunikationsstruktur

- ▶ Speicherkopplung
- ▶ Nachrichtenkopplung

- Synchronisation

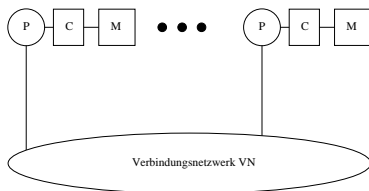
- ▶ Semaphore
- ▶ Barriers

- Latenzbehandlung

- ▶ Latenz verstecken
- ▶ Latenz minimieren



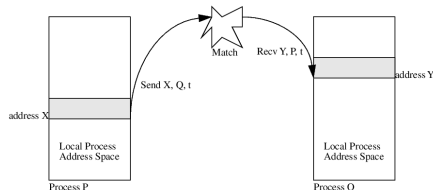
Distributed Memory: MP



- Multiprozessoren haben einen *lokalen Adressraum*: Jeder Prozessor kann nur auf seinen Speicher zugreifen.
- Interaktion mit anderen Prozessoren ausschließlich über das Senden von Nachrichten.
- Prozessoren, Speicher und Cache sind Standardkomponenten: Volle Ausnutzung des Preisvorteils durch hohe Stückzahlen.
- Verbindungsnetzwerk von Fast Ethernet bis Myrinet.
- Ansatz mit der höchsten Skalierbarkeit: IBM BlueGene > 100 K Prozessoren



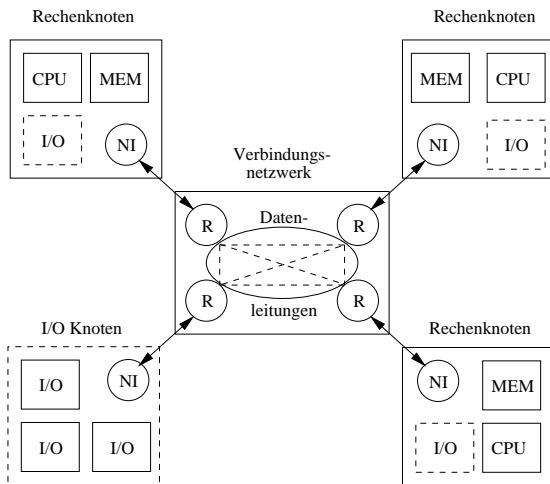
Distributed Memory: Message Passing



- Prozesse kommunizieren Daten zwischen verteilten Adreßräumen
- expliziter Nachrichtenaustausch notwendig
- Sende-/Empfangsoperationen



Eine generische Parallelrechnerarchitektur



Generischer Aufbau eines skalierbaren Parallelrechners mit verteiltem Speicher



Skalierbarkeit: Parameter

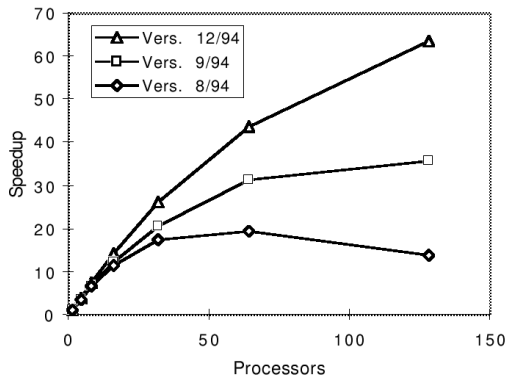
Parameter, welche die Skalierbarkeit eines Systems bestimmen:

- Bandbreite [MB/s]
- Latenzzeit [μ s]
- Kosten [\$]
- Physische Größe [m^2, m^3]

Eine skalierbare Architektur sollte harte Grenzen vermeiden!



Einfluß von Parallelen Architekturen?



from Culler, Singh, Gupta: Parallel Computer Architecture

Eine skalierbare Architektur ist Voraussetzung für skalierbares Rechnen

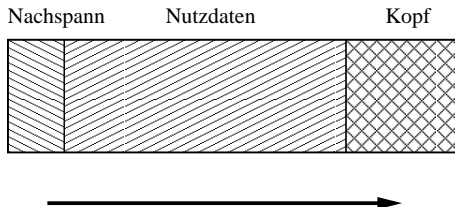


Nachrichtenaustausch

Speicherblock (variabler Länge) soll von einem Speicher zum anderen kopiert werden

Genauer: Vom Adressraum eines Prozesses in den eines anderen (auf einem anderen Prozessor)

Das Verbindungsnetzwerk ist paketorientiert. Jede Nachricht wird in Pakete fester Länge zerlegt (z. B. 32 Byte bis 4 KByte)



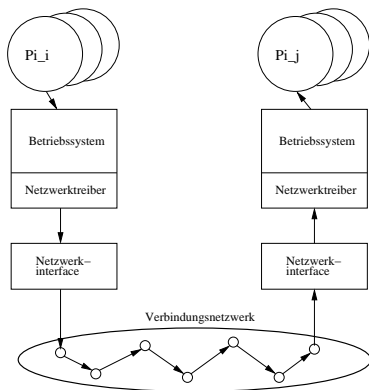
Kopf: Zielprozessor, Nachspann: Prüfsumme

Kommunikationsprotokoll: Bestätigung ob Paket (richtig) angekommen, Flusskontrolle



Nachrichtenaustausch

Schichtenmodell (Hierarchie von Protokollen):



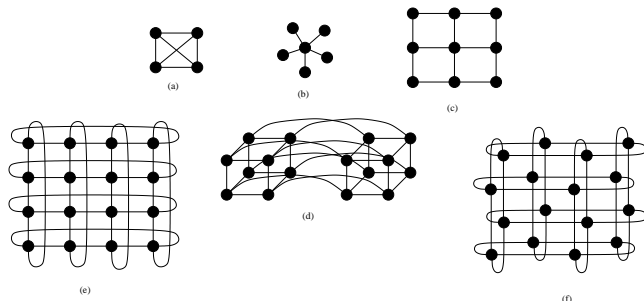
Modell der Übertragungszeit:

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

t_s : setup-Zeit (latency), t_b : Zeit pro Byte, $1/t_b$: Bandbreite, abh. von Protokoll



Netzwerktopologien I



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

- *Hypercube*: der Dimension d hat 2^d Prozessoren. Prozessor p ist mit q verbunden wenn sich deren Binärdarstellungen *in genau einem Bit* unterscheiden.
- **Netzwerkknoten**: Früher (vor 1990) war das der Prozessor selbst, heute sind es dedizierte Kommunikationsprozessoren



Netzwerktopologien II

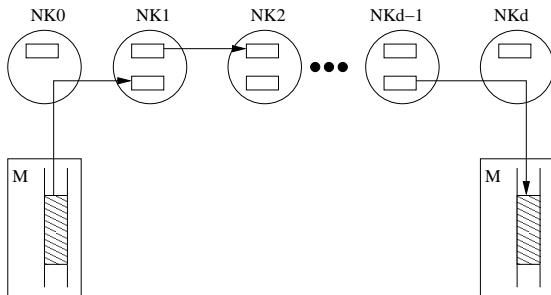
Kennzahlen:

Netzwerktopologie	Knoten- grad K	Leitungs- anzahl L	Durch- messer D	Bisektions- bandbreite B	Sym- metrie
Volle Konnektivität	$N - 1$	$N(N - 1)/2$	1	$(N/2)^2$	Ja
Stern	$N - 1$	$N - 1$	2	$\lfloor N/2 \rfloor$	nein
2D-Gitter	4	$2N - 2\sqrt{N}$	$2(\sqrt{N} - 1)$	\sqrt{N}	nein
3D-Torus	6	$3N$	$3\lfloor \sqrt{N}/2 \rfloor$	$2\sqrt{N}$	ja
Hypercube	$\log_2 N$	$nN \log_2 N$	n	$N/2$	ja
k-ärer n-cube ($N = k^n$)	$3N$	$n\lfloor k/2 \rfloor$	nN	$2k^{n-1}$	ja



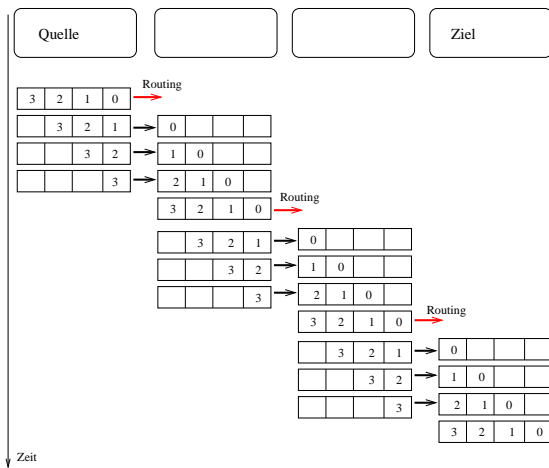
Store & Forward Routing

Store-and-forward routing: Nachricht der Länge n wird in Pakete der Länge N zerlegt. Pipelining auf Paketebene: Paket wird aber vollständig im NK gespeichert



Store & Forward Routing

Übertragung eines Paketes:



Store & Forward Routing

Laufzeit:

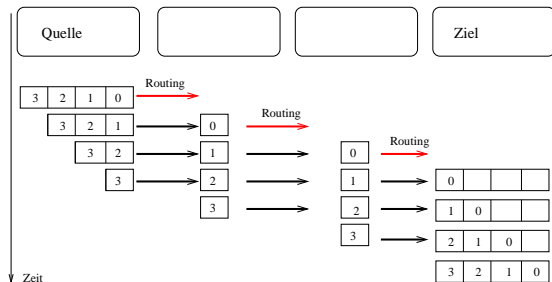
$$\begin{aligned}t_{SF}(n, N, d) &= t_s + d(t_h + Nt_b) + \left(\frac{n}{N} - 1\right) (t_h + Nt_b) \\ &= t_s + t_h \left(d + \frac{n}{N} - 1\right) + t_b (n + N(d - 1)).\end{aligned}$$

- t_s : Zeit, die auf Quell- und Zielrechner vergeht bis das Netzwerk mit der Nachrichtenübertragung beauftragt wird, bzw. bis der empfangende Prozess benachrichtigt wird. Dies ist der Softwareanteil des Protokolls.
- t_h : Zeit die benötigt wird um das erste Byte einer Nachricht von einem Netzwerkknoten zum anderen zu übertragen (*engl.* node latency, hop-time).
- t_b : Zeit für die Übertragung eines Byte von Netzwerkknoten zu Netzwerkknoten.
- d : Hops bis zum Ziel.



Cut-Through Routing

Cut-through routing oder *wormhole routing*: Pakete werden nicht zwischengespeichert, jedes Wort (sog. *flit*) wird sofort an nächsten Netzwerkknoten weitergeleitet
Übertragung eines Paketes:



Laufzeit:

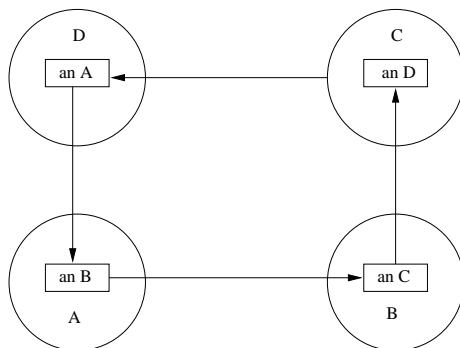
$$t_{CT}(n, N, d) = t_s + t_h d + t_b n$$

Zeit für kurze Nachricht ($n = N$): $t_{CT} = t_s + dt_h + Nt_b$. Wegen $dt_h \ll t_s$ (Hardware!) quasi entfernungsunabhängig



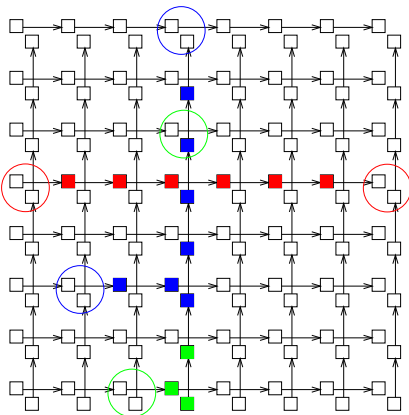
Deadlock

In paketvermittelnden Netzwerken besteht die Gefahr des *store-and-forward deadlock*:



Deadlock

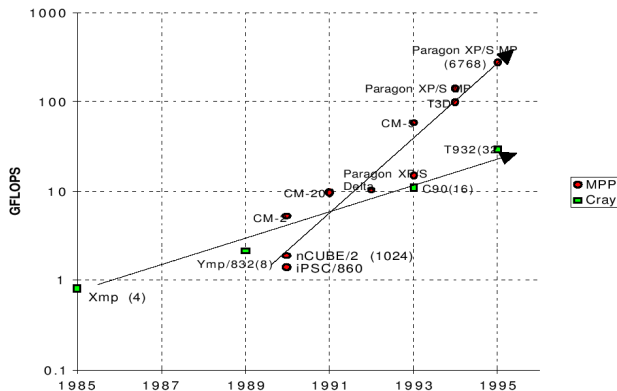
Zusammen mit cut-through routing:



Verklemmungsfreies „dimension routing“. Beispiel 2D-Gitter: Zerlege Netzwerk in $+x$, $-x$, $+y$ und $-y$ Netzwerke mit jeweils eigenen Puffern. Nachricht läuft erst in Zeile, dann in Spalte.



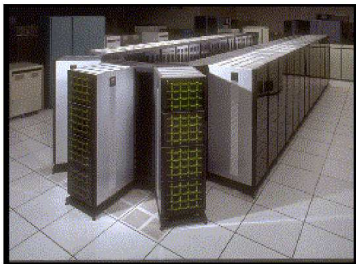
Multi-Processor Performance



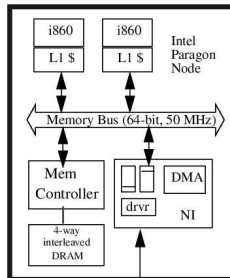
from Culler, Singh, Gupta: Parallel Computer Architecture



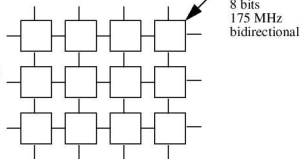
Message Passing Architectures I



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



2D grid network
with processing node
attached to every switch

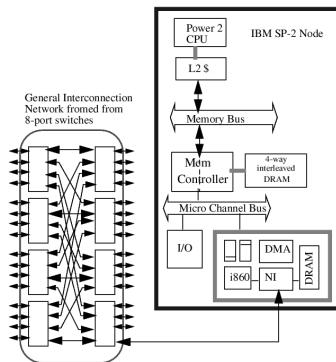


Intel Paragon:

- erste Machine mit parallelem Unix
- Prozeßmigration, Gangscheduling



Message Passing Architectures II

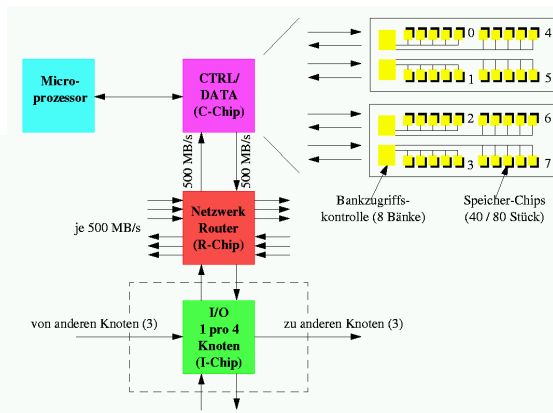
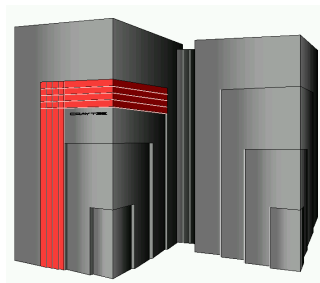


IBM SP2:

- Rechenknoten sind RS 6000 workstations
- Switching Netzwerk



Message Passing Architectures III



Cray T3E:

- hohe Packungsdichte
- ein systemweiter Taktgeber
- virtual shared memory



Top500 Benchmark:

- LINPACK Benchmark wird zur Evaluation der Systeme verwendet
- Benchmarkleistung reflektiert nicht die Gesamtleistung des Systems
- Benchmark zeigt Performanz bei der Lösung von dichtbesetzten linearen GLS
- Sehr reguläres Problem: erzielte Leistung ist sehr hoch (nahe peak performance)



Top 10 of Top500

Rank	Site	Computer/Year Vendor	Cores	R _{max}	R _{peak}	Power
1	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 GHz / Opteron DC 1.8 GHz, Voltaire Infiniband / 2008 IBM	122400	1026.00	1375.78	2345.50
2	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution / 2007 IBM	212992	478.20	596.38	2329.60
3	Argonne National Laboratory United States	Blue Gene/P Solution / 2007 IBM	163840	450.30	557.06	1260.00
4	Texas Advanced Computing Center/Univ. of Texas United States	Ranger - SunBlade x6420, Opteron Quad 2GHz, Infiniband / 2008 Sun Microsystems	62976	326.00	503.81	2000.00
5	DOE/Oak Ridge National Laboratory United States	Jaguar - Cray XT4 QuadCore 2.1 GHz / 2008 Cray Inc.	30976	205.00	260.20	1580.71
6	Forschungszentrum Juelich (FZJ) Germany	JUGENE - Blue Gene/P Solution / 2007 IBM	65536	180.00	222.82	504.00
7	New Mexico Computing Applications Center (NMCAC) United States	Encanto - SGI Altix ICE 8200, Xeon quad core 3.0 GHz / 2007 SGI	14336	133.20	172.03	861.63
8	Computational Research Laboratories, TATA SONS India	EKA - Cluster Platform 3000 BL460c, Xeon 53xx 3GHz, Infiniband / 2008 Hewlett-Packard	14384	132.80	172.61	786.00
9	IDRIS France	Blue Gene/P Solution / 2008 IBM	40960	112.50	139.26	315.00
10	Total Exploration Production France	SGI Altix ICE 8200EX, Xeon quad core 3.0 GHz / 2008 SGI	10240	106.10	122.88	442.00

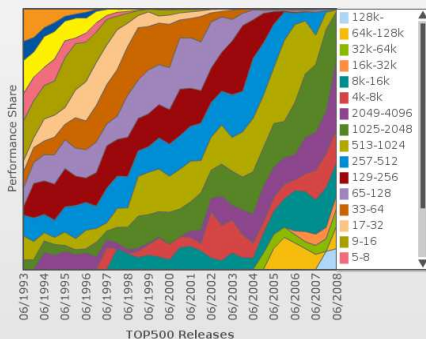
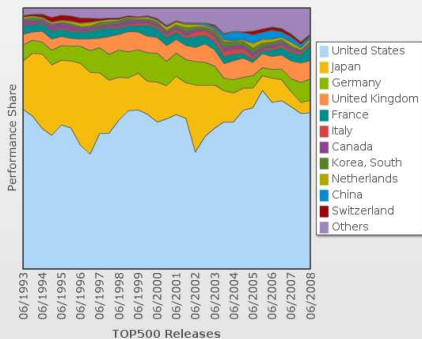


Top500 key facts

- alle Systeme haben mehrfache TeraFlop/s Leistung
- 1. Rechner (roadrunner) hat 1.026 PFlop/s in LINPACK benchmark
- 1. Rechner (roadrunner) ist der Energieeffizienteste der Liste
- Energieverbrauchsdurchschnitt von TOP10 ist 1.32 MW und 248 MFlops/W
- 500. Rechner hat 9.00 TFlop/s in LINPACK benchmark
- akkumulierte Leistung ist 11.7 PFlop/s (4.92 PFlop/s)
- TOP 100 Mindestleistung 12.97 TFlop/s (9.29 TFlop/s)
- 400 (80%) Systeme sind Cluster, Rest MPP
- Prozessortyp: Intel Harpertown 375, IBM Power 68, AMD Opteron 56
- Quad core vorwiegende Chip Architektur (56%)
- InfiniBand Technologie in 120 Systemen
- Skalarprozessor 95.8%, Vectorprozessor 4.2%

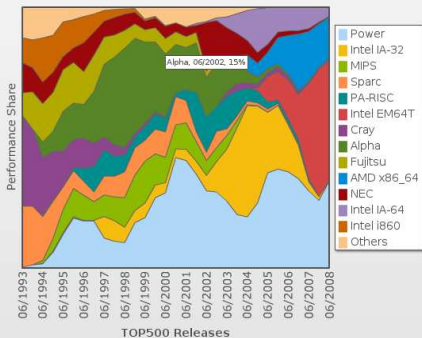


Top500 Country + Processor Count

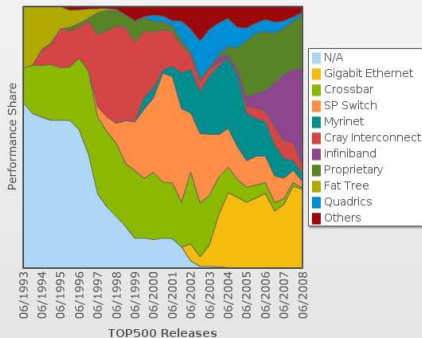


Top500 Processor + Interconnect

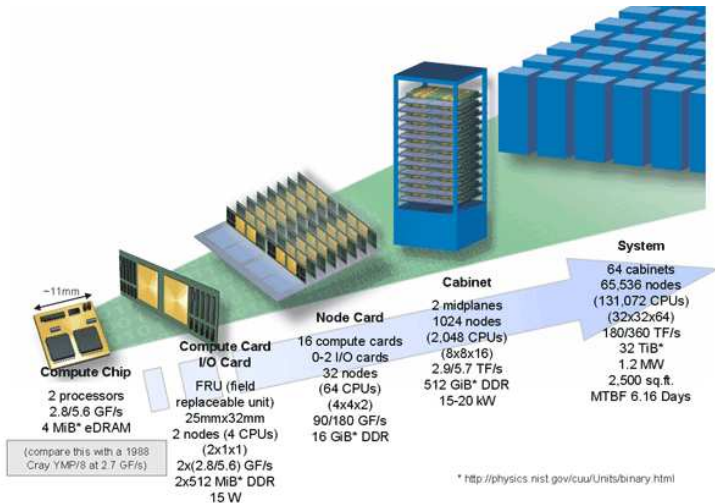
Processor Family Share Over Time
1993-2008



Interconnect Share Over Time
1993-2008



IBM Blue Gene/L Architecture



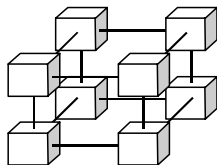
IBM Blue Gene/L Specification

367 teraFLOP/s (in symmetric mode)	
184 teraFLOP/s (in communications co-processor mode).....	Peak computational rate
16 TB ($16 \cdot 2^{40}$ bytes).....	Aggregate memory
400 TB ($400 \cdot 10^{12}$ bytes).....	Aggregate global disk
40 GB/s ($40 \cdot 10^9$ bytes/second).....	Delivered I/O bandwidth to applications
1,024 x 1-Gb/s Ethernet (in 10^9 bits/second).....	External networking
65,536 (131,072).....	Number of nodes (processors)
256 MB ($256 \cdot 2^{20}$ bytes).....	Memory per node
Dual PowerPC 440.....	Microprocessor technology
2 MW ($2 \cdot 10^6$ Watts).....	Power required for computer and cooling
>4,500,000 BTU/hr.....	Heat generated
>5,000.....	Cables in the machine
>12 miles.....	Aggregate cable length



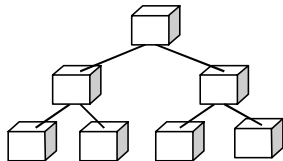
IBM Blue Gene/L Networks

65536 Knoten verbunden über drei integrierte Netzwerke



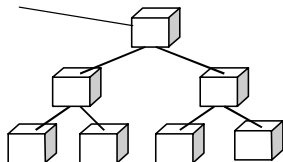
3 Dimensional Torus

- 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

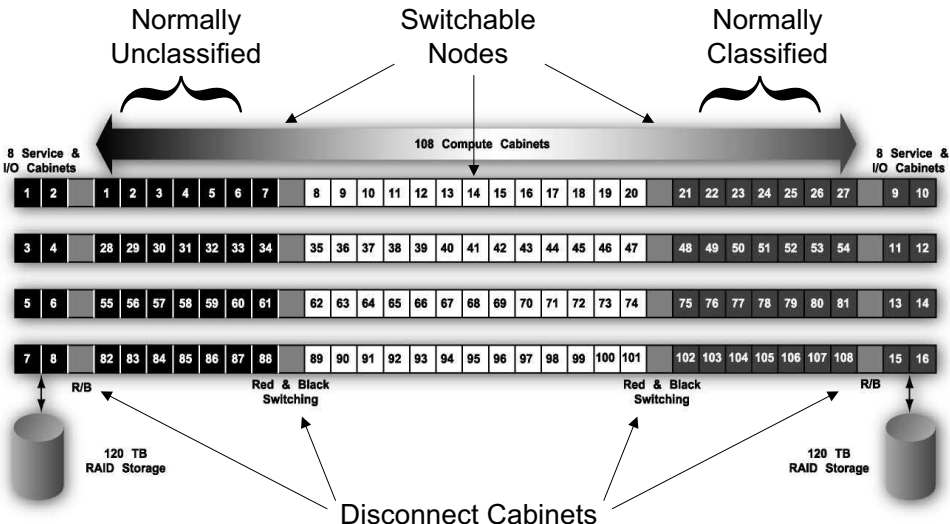
- Incorporated into every node ASIC
- Disk I/O
- Host control, booting and diagnostics



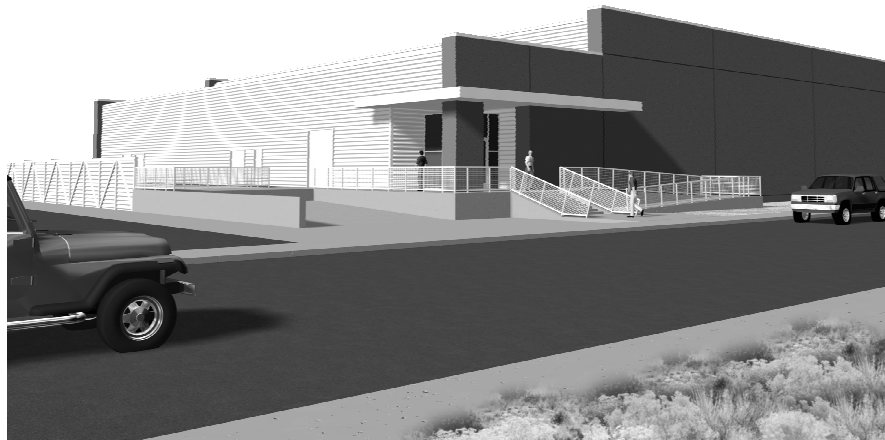
Cray RedStorm



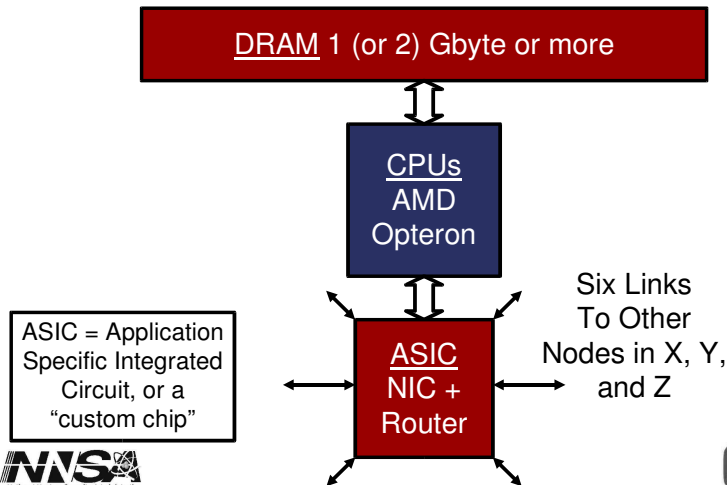
Cray RedStorm Configuration



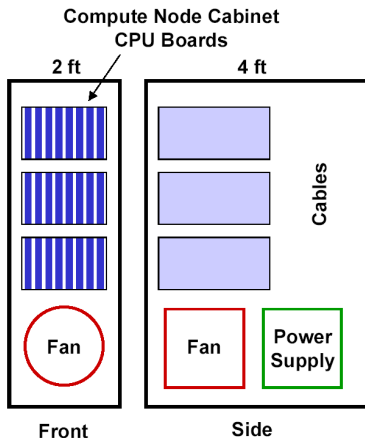
Cray RedStorm Building



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)

