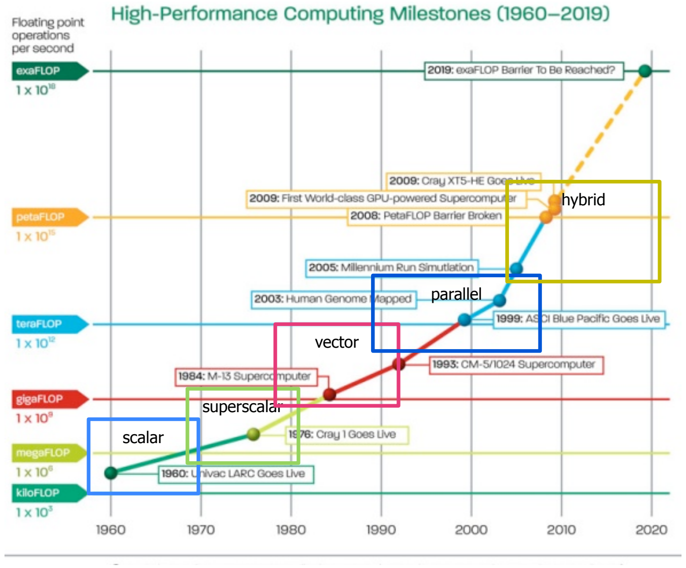
Parallelization

Hardware architecture

Contents



Why Parallel Computing

Primary reasons:

- Save time
- Solve larger problems
- Provide concurrency (do multiple things at the same time)

Classification of HPC hardware

- Architecture
- Memory organization

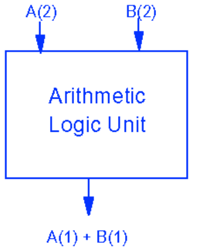
1st Classification: Architecture

- There are several different methods used to classify computers
- No single taxonomy fits all designs
- Flynn's taxonomy uses the relationship of program instructions to program data
  - SISD - Single Instruction, Single Data Stream
  - SIMD - Single Instruction, Multiple Data Stream
  - MISD - Multiple Instruction, Single Data Stream
  - MIMD - Multiple Instruction, Multiple Data Stream

Flynn's Taxonomy

- SISD: single instruction and single data stream: uniprocessor
- SIMD: vector architectures: lower flexibility
- MISD: no commercial multiprocessor: imagine data going through a pipeline of execution engines
- MIMD: most multiprocessors today: easy to construct with off-the-shelf computers, most flexibility

SISD



- One instruction stream
- One data stream
- One instruction issued on each clock cycle
- One instruction executed on single element(s) of data (scalar) at a time
- Traditional 'von Neumann' architecture (remember from introduction)

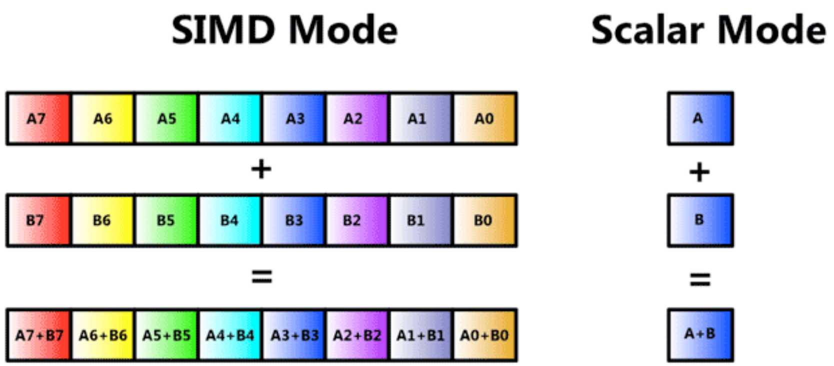
SIMD

- Also von Neumann architectures but more powerful instructions
- Each instruction may operate on more than one data element
- Usually intermediate host executes program logic and broadcasts instructions to other processors
- Synchronous (lockstep)
- Rating how fast these machines can issue instructions is not a good measure of their performance
- Two major types:
  - Vector SIMD
  - Parallel SIMD

Vector SIMD

- Single instruction results in multiple operands being updated
- Scalar processing operates on single data elements. Vector processing operates on whole vectors (groups) of data at a time.
- Examples:
  - SSE instructions
  - NEC SX-9
  - Fujitsu VP
  - Hitachi S820

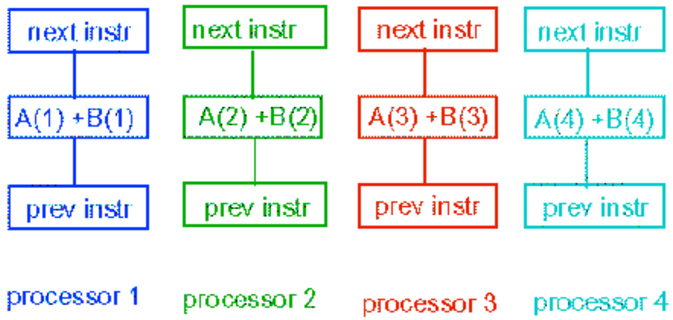
Vector SIMD



Parallel SIMD

- Several processors execute the same instruction in lockstep
- Each processor modifies a different element of data
- Drawback: idle processors
- Advantage: no explicit synchronization required
- Examples
  - GPGPU's
  - Cell

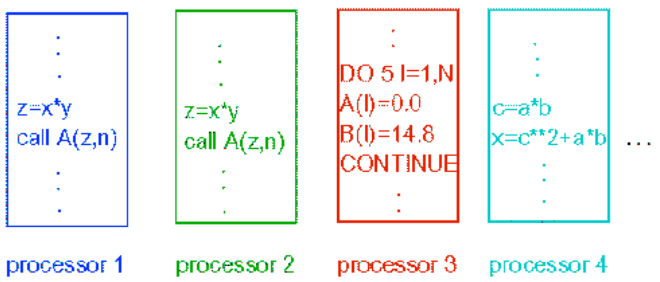
Parallel SIMD

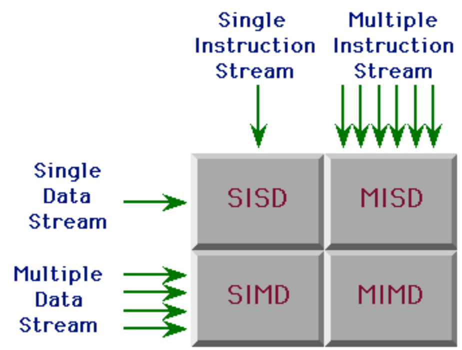


MIMD

- Several processors executing different instructions on different data
- Advantages:
  - different jobs can be performed at a time
  - A better utilization can be achieved
- Drawbacks:
  - Explicit synchronization needed
  - Difficult to program
- Examples:
  - MIMD Accomplished via Parallel SISD machines: all clusters, Cray XE6, IBM Blue Gene, SGI Altix
  - MIMD Accomplished via Parallel SIMD machines: NEC SX-8, Convex(old), Cray X2

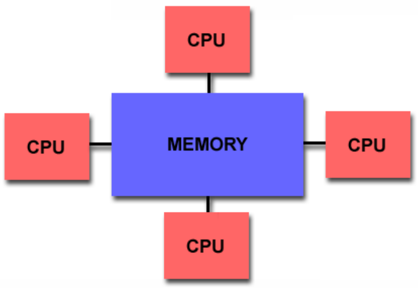
MIMD Model





2nd Classification: Memory organization

- Shared memory (SMP)
  - UMA
  - NUMA
    - CC-NUMA
- Distributed memory

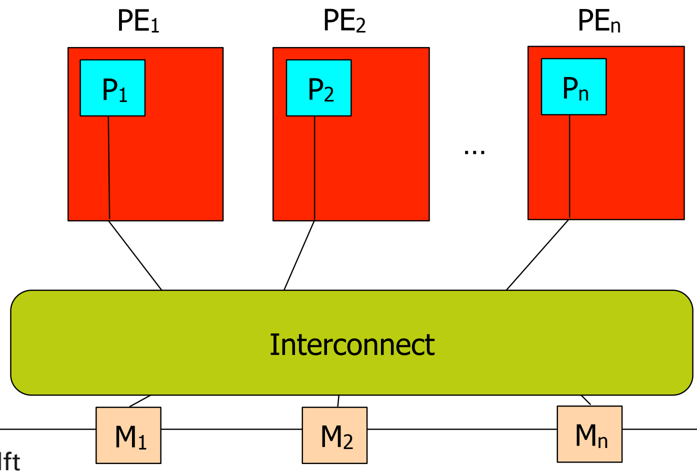


Memory Organization

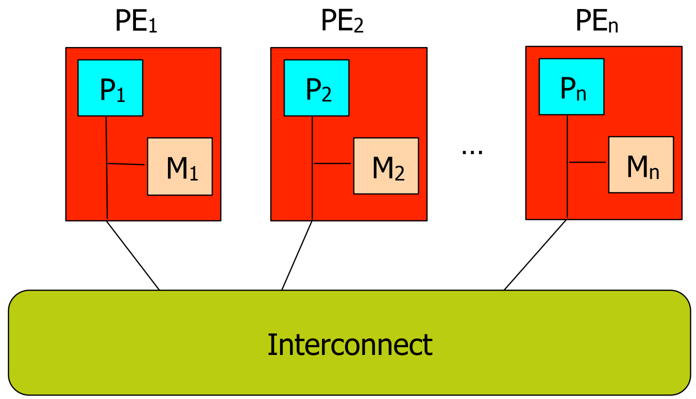
Symmetric shared-memory multiprocessor (SMP)

- Implementations:
  - Multiple processors connected to a single centralized memory – since all processors see the same memory organization -> **uniform** memory access (UMA)
  - Shared-memory because all processors can access the entire memory address space through a tightly interconnect between compute/ memory nodes - **non-uniform** (NUMA)

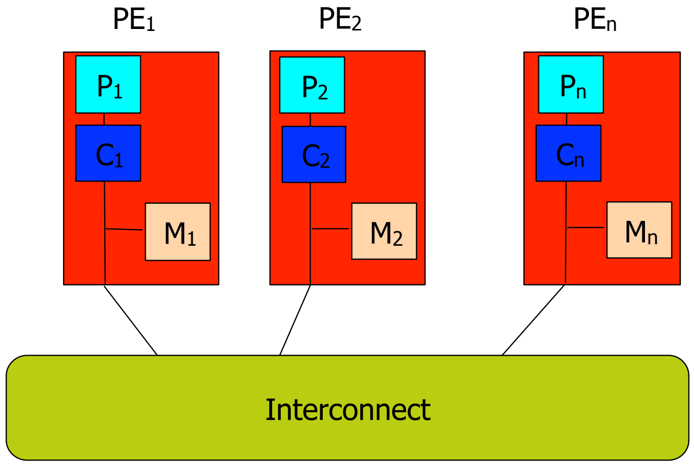
UMA (Uniform Memory Access)



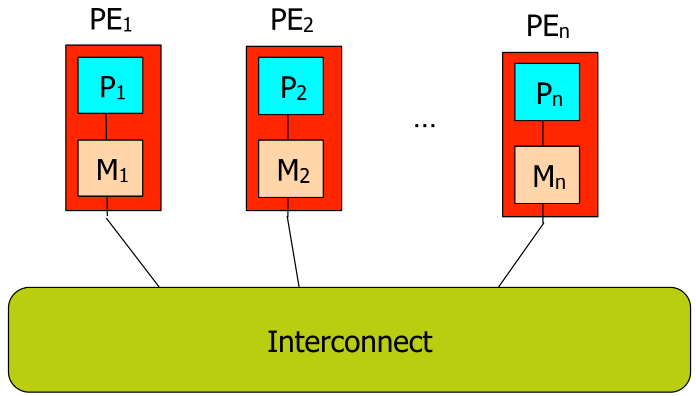
NUMA (Non Uniform Memory Access)



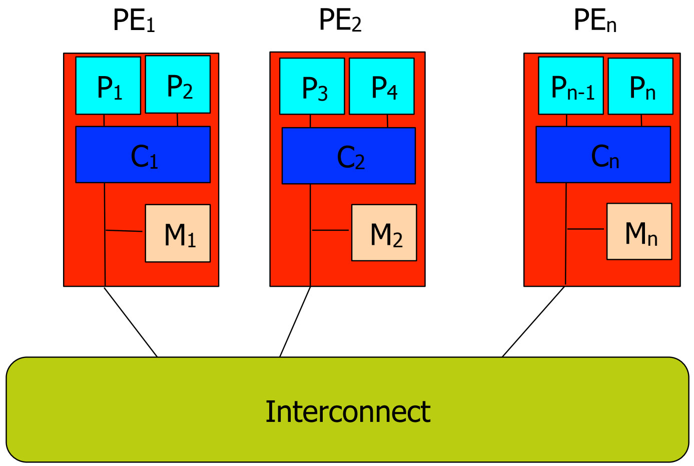
CC-NUMA (Cache Coherent NUMA)



Distributed memory

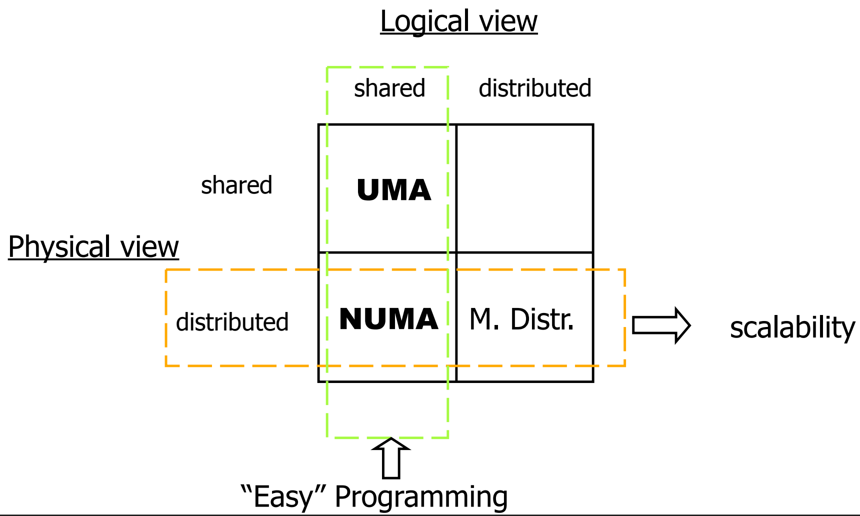


Hybrid Distributed memory



The largest and fastest computers in the world today employ both shared and distributed memory architectures.

Memory architecture



Shared-Memory vs. Distributed-Memory

Shared-memory:
- Well-understood programming model
- Communication is implicit and hardware handles protection
- Hardware-controlled caching
- OpenMP and MPI

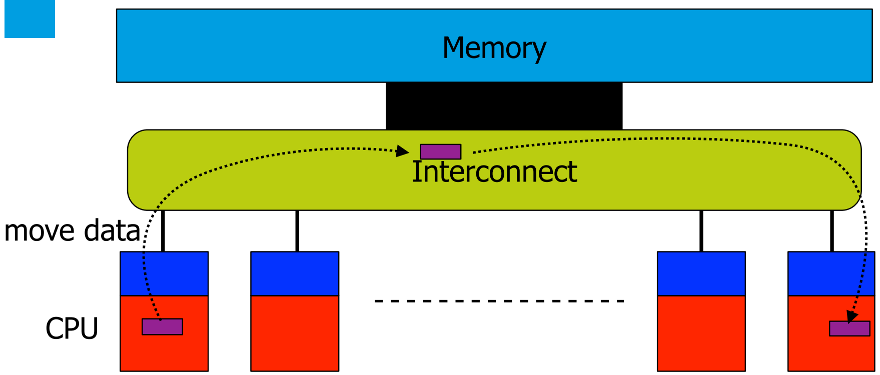
Distributed-memory:
- No cache coherence → simpler hardware
- Explicit communication → easier for the programmer to restructure code
- Sender can initiate data transfer
- MPI, PGAS

Hardware implementation (MIMD)

- Shared memory
- Distributed memory

A Shared Memory Computer

Data movement is transparent to the programmer

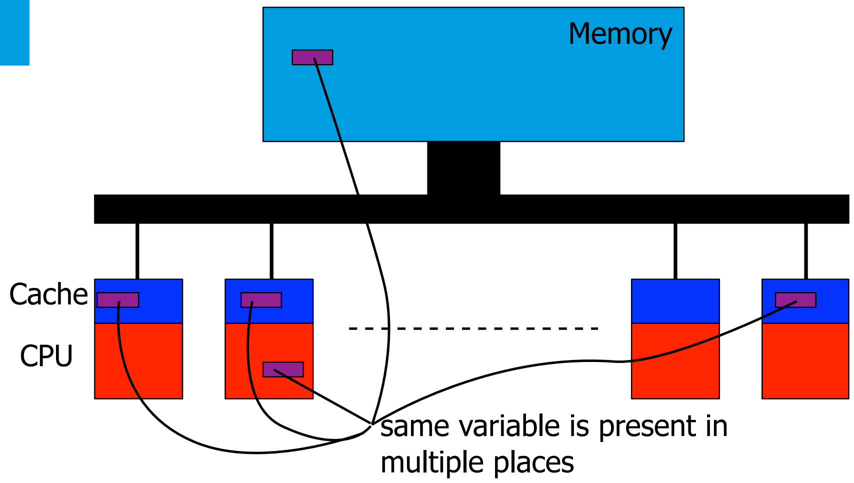


SMP = Symmetric Multi-Processor
Note that the CPU can be a multi-core chip

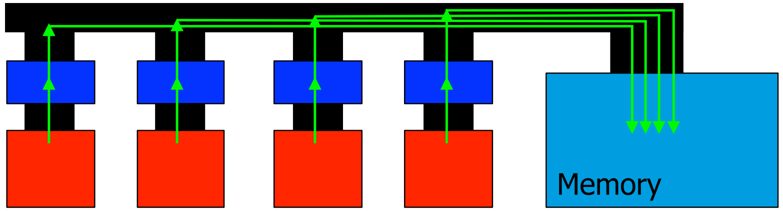
Issues for MIMD distributed Shared Memory

- Memory Access
  - Can reads be simultaneous?
  - How to control multiple writes?
- Synchronization mechanism needed
  - semaphores
  - monitors
- Local caches need to be co-ordinated
  - cache coherency protocols

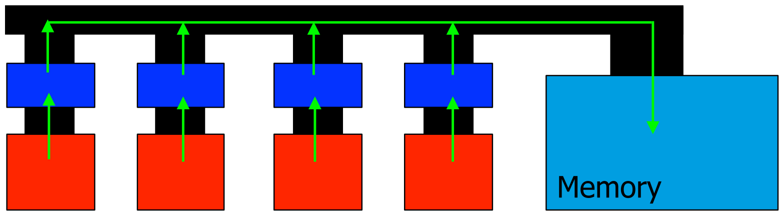
cache coherency ensures that one always gets the right value ... regardless of where the data is



write-through: simple, but wastes memory bandwidth



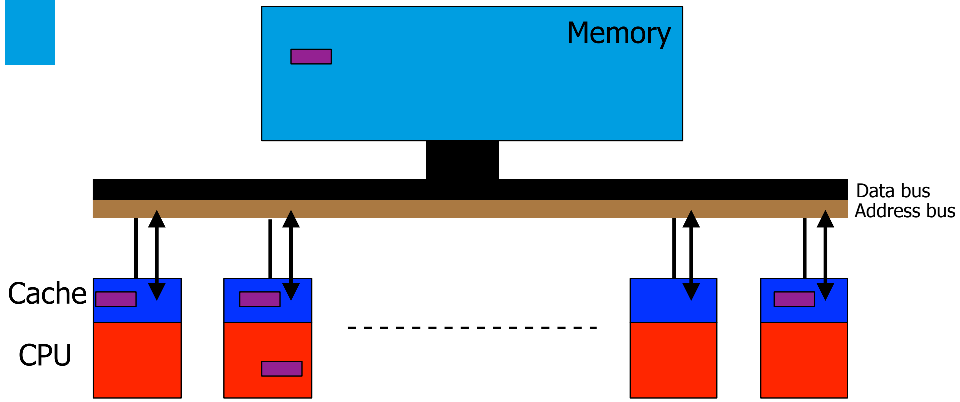
write-back: minimizes bandwidth, takes extra logic



Cache Coherence Protocols

- Directory-based: A single location (directory) keeps track of the sharing status of a block of memory
- Snooping: Every cache block is accompanied by the sharing status of that block – all cache controllers monitor the shared bus so they can update the sharing status of the block, if necessary
  - Write-invalidate: a processor gains exclusive access of a block before writing by invalidating all other copies
  - Write-update: when a processor writes, it updates other shared copies of that block

Cache Coherence - Snooping



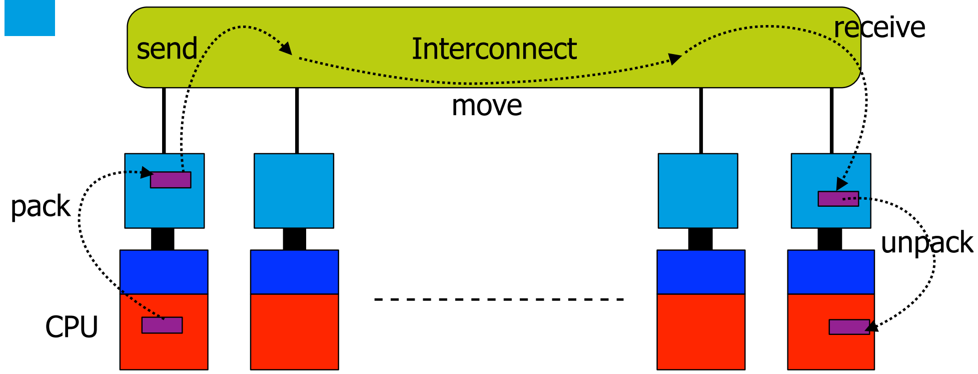
With a snooping protocol, ALL address traffic on the bus is monitored by ALL processors

MIMD-Distributed Memory

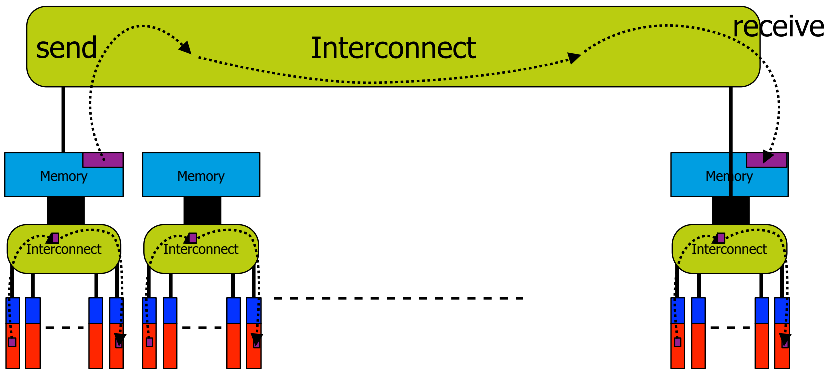
- Connection Network
  - fast
  - high bandwidth
  - scalable
- Communications
  - explicit message passing
  - parallel languages
    - Unified Parallel C, Co-Array Fortran, HPF
  - libraries for sequential languages
    - MPI, PVM, Java with CSP

A Distributed Memory Computer

The system is programmed using message passing



Hybrid: MIMD with shared memory nodes



And now imagine a multi-core chip at the lowest level.

Interconnection Network

- Speed and Bandwidth are critical
- Low cost networks
  - local area network (ethernet, token ring)
- High Speed Networks
  - The heart of a MIMD-DM Parallel Machine

Issues for Networks

- Total Bandwidth
  amount of data which can be moved from somewhere to somewhere per unit time
- Link Bandwidth
  amount of data which can be moved along one link per unit time
- Message Latency
  time from start of sending a message until it is received
- Bisection Bandwidth
  amount of data which can move from one half of network to the other per unit time for worst case split of network

Design Characteristics of a Network

Design Characteristics of a Network

- Topology (how things are connected):
  - Crossbar, ring, 2-D and 3-D meshes or torus, hypercube, tree, butterfly, ...
- Routing algorithm (path used):
  - Example in 2D torus: all east-west then all north-south
- Switching strategy:
  - Circuit switching: full path reserved for entire message, like the telephone.
  - Packet switching: message broken into separately-routed packets, like the post office.
- Flow control (what if there is congestion):
  - Stall, store data in buffers, re-route data to other nodes, tell source node to temporarily halt, discard, ...

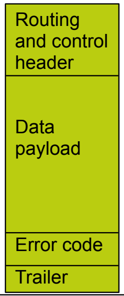
Performance Properties of a Network: Latency

- Latency: delay between send and receive times
  - Latency tends to vary widely across architectures
  - Vendors often report hardware latencies (wire time)
  - Application programmers care about software latencies (user program to user program)
- Latency is important for programs with many small messages

Performance Properties of a Network: Bandwidth

- The bandwidth of a link = w * 1/t
  - w is the number of wires
  - t is the time per bit

- Bandwidth typically in GigaBytes (GB), i.e., 8* 2^{20} bits
- Effective bandwidth is usually lower than physical link bandwidth due to packet overhead.



- Bandwidth is important for applications with mostly large messages

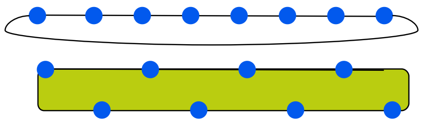
Common Network Topologies

Linear and Ring Topologies

- Linear array

  - Diameter = n-1; average distance ~n/3.
  - Bisection bandwidth = 1 (in units of link bandwidth).

- Torus or Ring

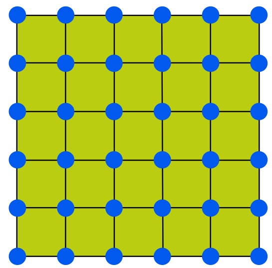


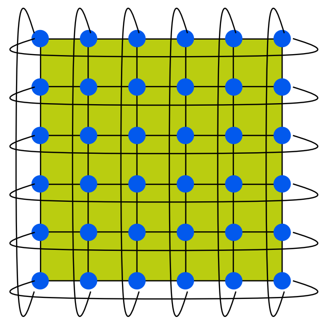
  - Diameter = n/2; average distance ~ n/4.
  - Bisection bandwidth = 2.
  - Natural for algorithms that work with 1D arrays.

Meshes and Tori

Two dimensional mesh
- **Diameter = 2 * (sqrt(n) – 1)**
- **Bisection bandwidth = sqrt(n)**

Two dimensional torus
- **Diameter = sqrt(n)**
- **Bisection bandwidth = 2* sqrt(n)**



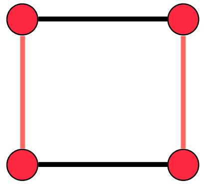


- **Generalises to higher dimensions (Cray XT5 used 3D Torus).**
- **Natural for algorithms that work with 2D and/or 3D arrays.**

Hypercube

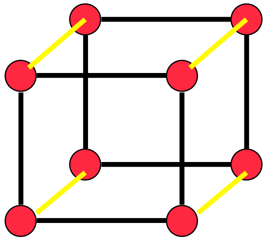
One Dimensional

Hypercube



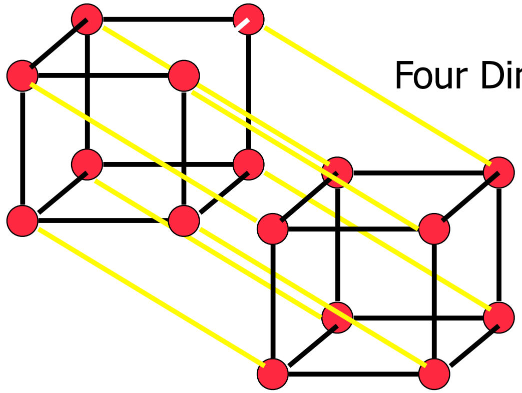
Two Dimensional

Hypercube



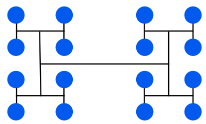
Three Dimensional

Hypercube

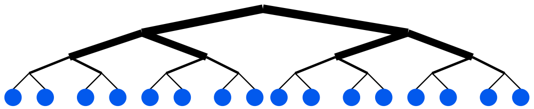


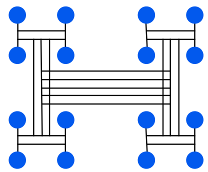
Four Dimensional

Trees



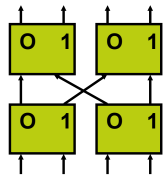
- Diameter = log n.
- Bisection bandwidth = 1.
- Easy layout as planar graph.
- Many tree algorithms (e.g., summation).
- Fat trees avoid bisection bandwidth problem:
  - More (or wider) links near top.



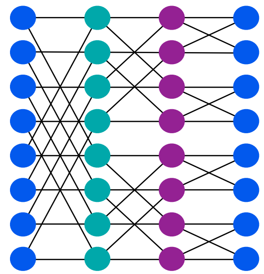


Butterflies with n = (k+1)2^k nodes

- Diameter = 2k.
- Bisection bandwidth = 2^k.
- Cost: lots of wires.
- Used in BBN Butterfly.
- Natural for FFT.



butterfly switch



multistage butterfly network

Topologies in Real Machines

older → newer

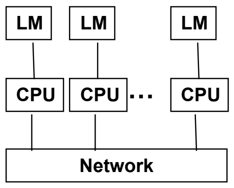
Cray XC30	**Dragonfly**
Cray XE series	**3D Torus**
Blue Gene/L /P	**3D Torus**
SGI Altix	**Fat tree**
Cray X1	**4D Hypercube***
Myricom (Millennium)	**Arbitrary**
Quadrics	**Fat tree**
IBM SP	**Fat tree (approx)**
SGI Origin	**Hypercube**
Intel Paragon (old)	**2D Mesh**

Many of these are approximations: E.g., the X1 is really a "quad bristled hypercube" and some of the fat trees are not as fat as they should be at the top

MIMD - clusters

Cluster

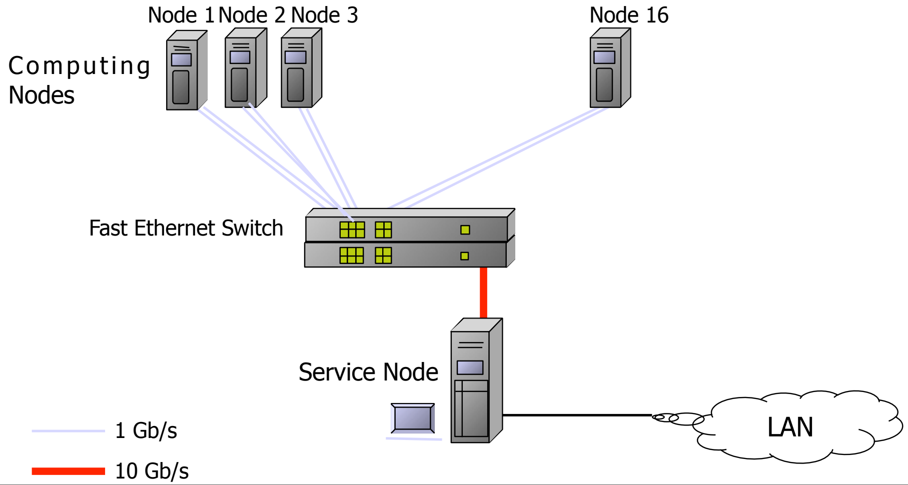
- A cluster is a type of parallel or distributed processing system, which consists of a collection of interconnected stand-alone or complete computers. These computers co-operatively work together as a single, integrated computing resource.



Cluster

Construction of a Beowulf Cluster

Topology



Beyond a Cluster: Grid

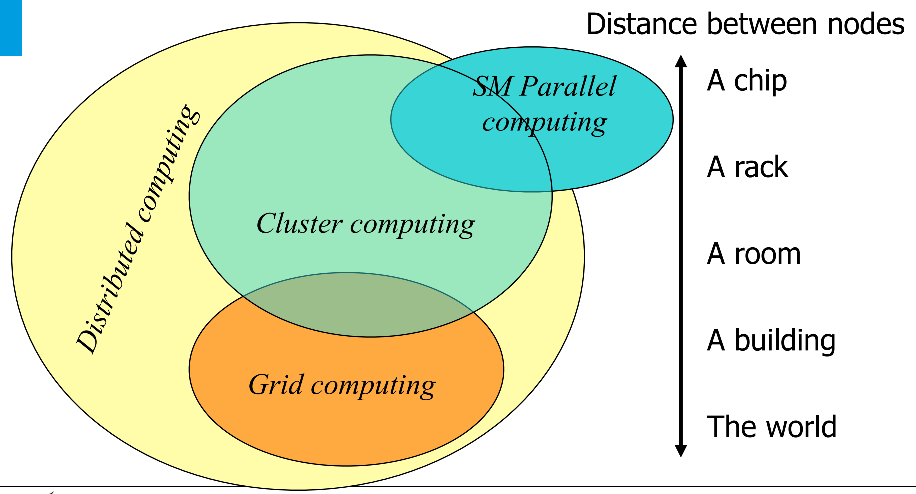
Computational Grids

- A network of geographically distributed resources including computers, peripherals, switches, instruments, and data.
- Each user should have a single login account to access all resources.
- Resources may be owned by diverse organisations.

GRID vs. Cluster

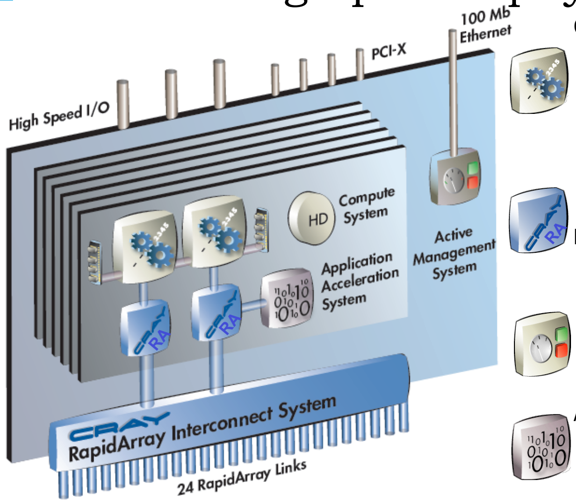
- **Cluster**: Computer network typically dedicated 100 % to execute a specific task
- **GRID**: computer networks distributed planet-wide, that can be shared by the means of resource management software

Cluster computing vs. others



Some HPC specialized hardware

General design philosophy



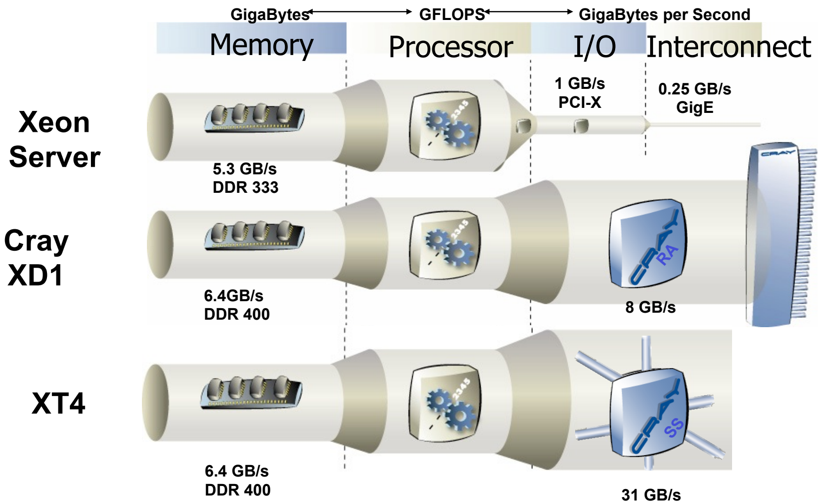
Compute
- 12 AMD Opteron processors 32/64 bit, x86 processors
- High Performance Linux

RapidArray Interconnect
- 12 communications processors
- 1 Tb/s switch fabric

Active Management
- Dedicated processor
- 6 co-processors

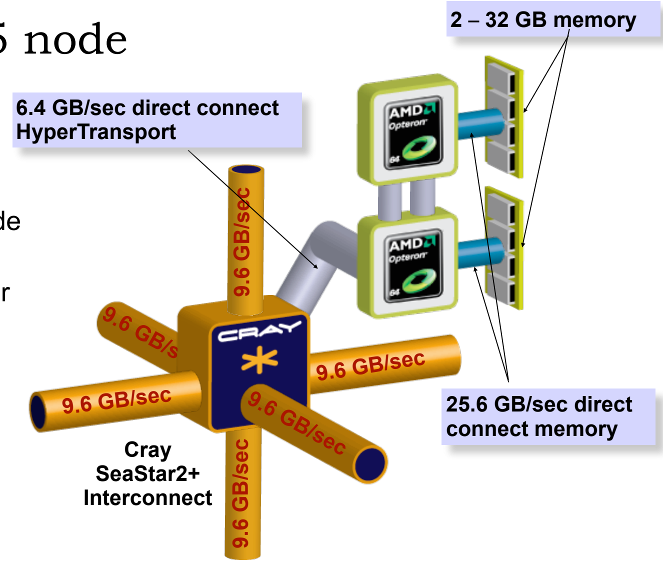
Processors directly connected via integrated switch fabric

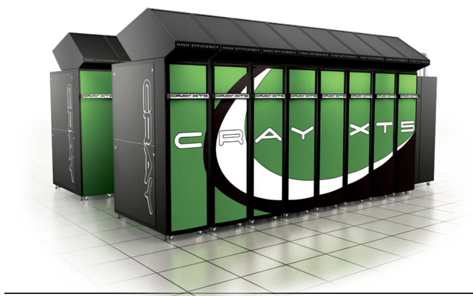
Balanced Interconnect

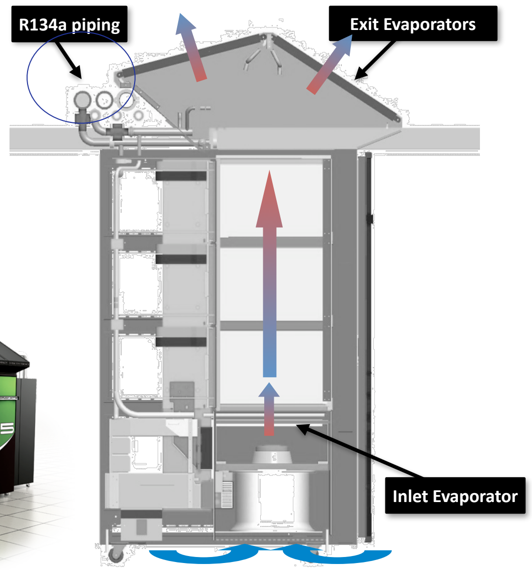


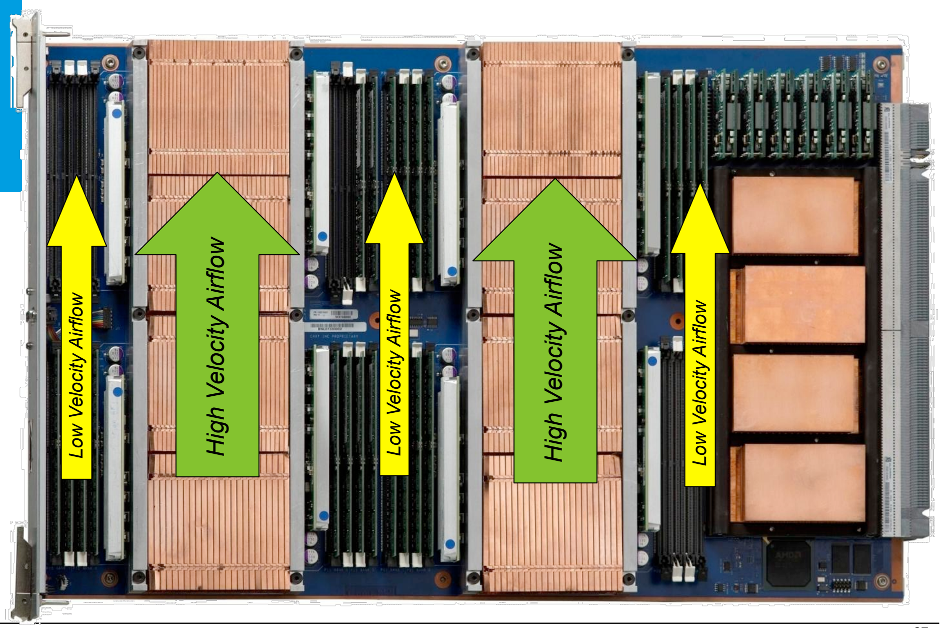
Cray XT5 node

- 8-way SMP
- >70 Gflops per node
- Up to 32 GB of shared memory per node
- OpenMP Support

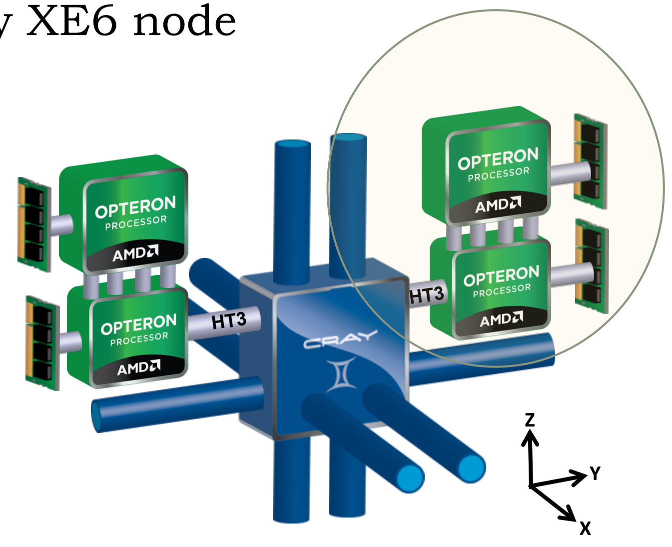


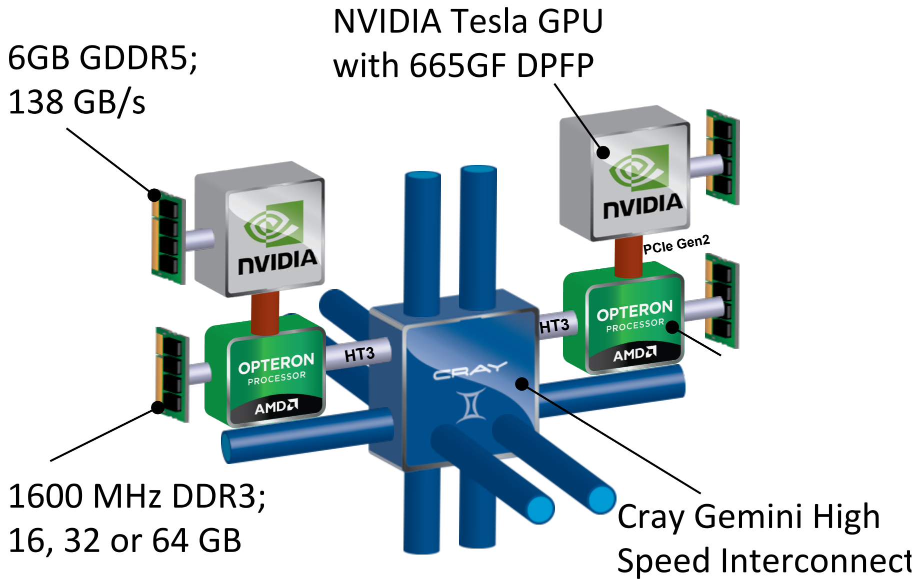




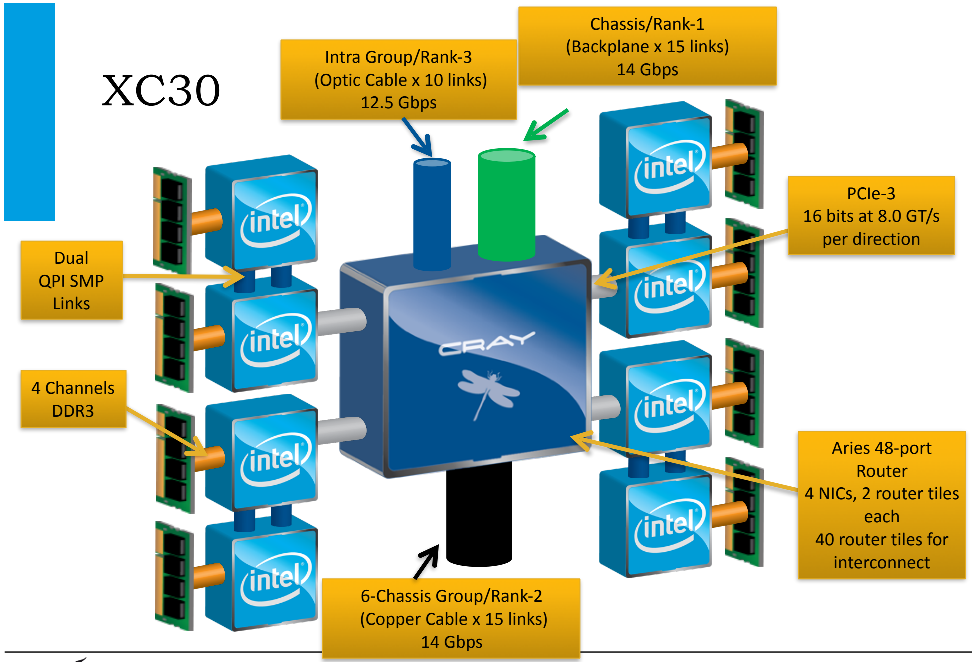


Cray XE6 node

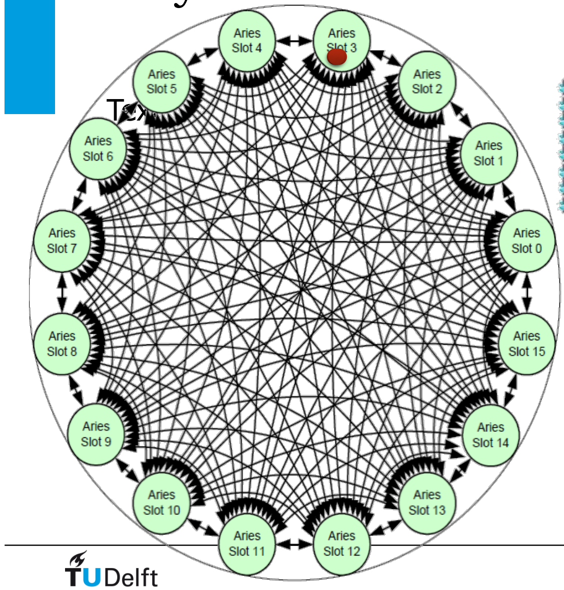


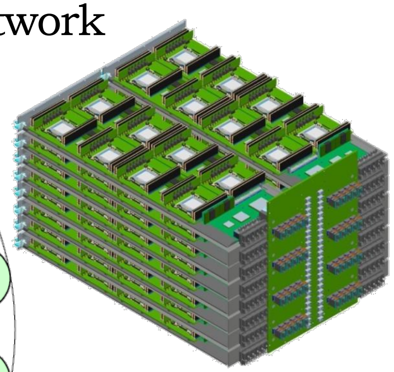


XC30



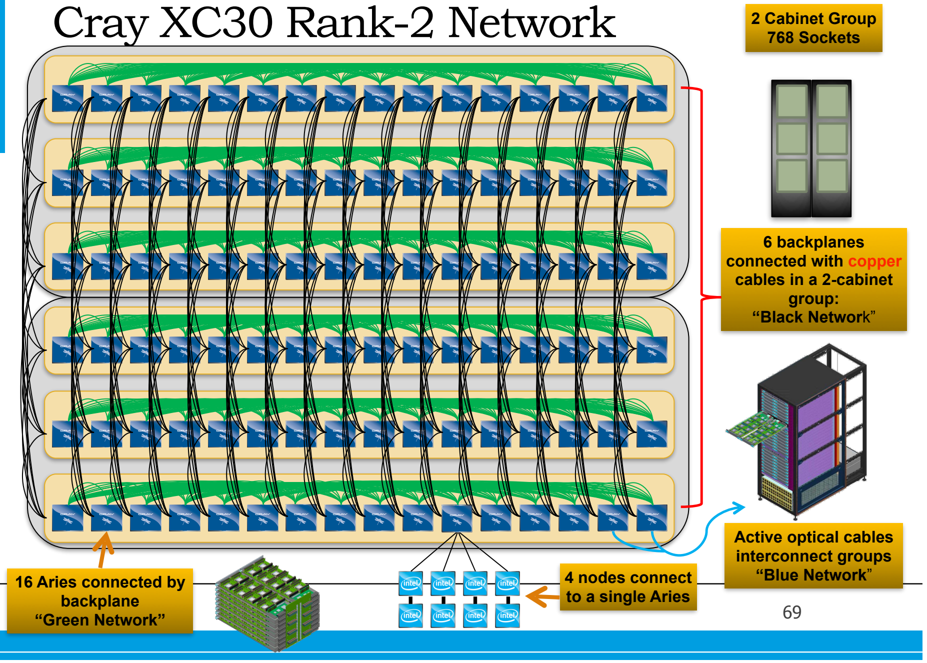
Cray XC30 Rank1 Network



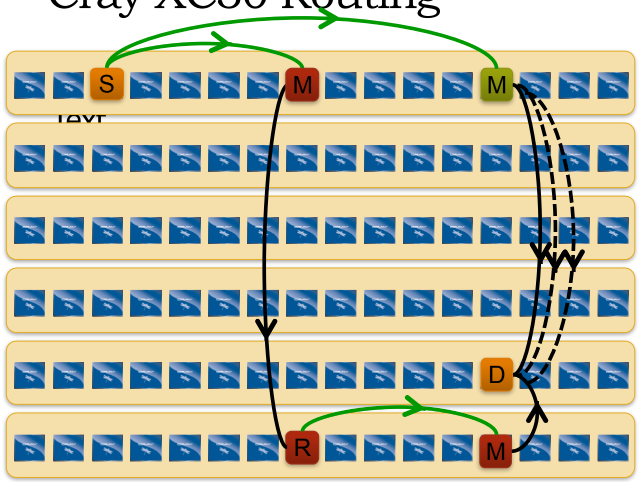


- Chassis with 16 compute blades
- 128 Sockets
- Inter-Aries communication over backplane
- Per-Packet adaptive Routing

Cray XC30 Rank-2 Network



Cray XC30 Routing



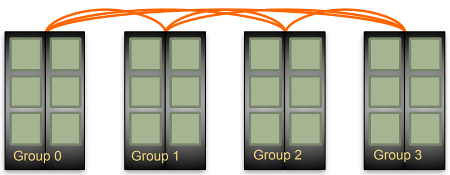
Minimal routes between any two nodes in a group are just two hops

Non-minimal route requires up to four hops.

With adaptive routing we select between minimal and non-minimal paths based on load

The Cray XC30 Class-2 Group has sufficient bandwidth to support full injection rate for all 384 nodes with non-minimal routing

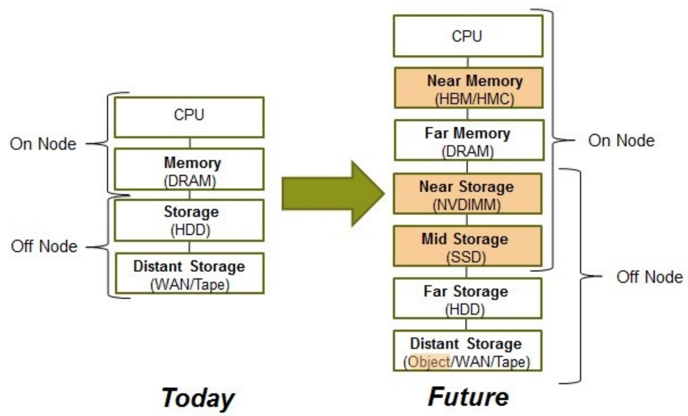
Cray XC30 Network Overview – Rank-3 Network

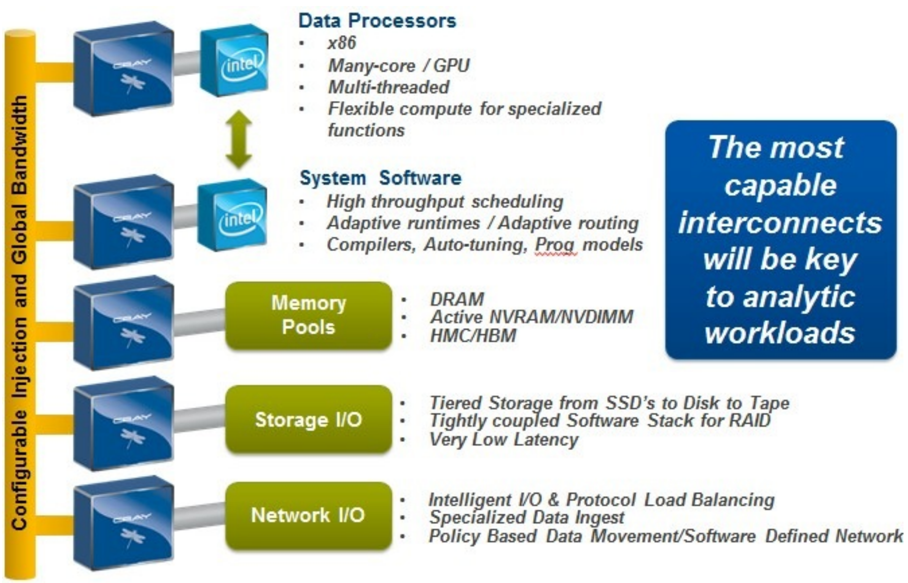


Example: An 4-group system is interconnected with 6 optical "bundles". The "bundles" can be configured between 20 and 80 cables wide

Integration HPC and Big Data

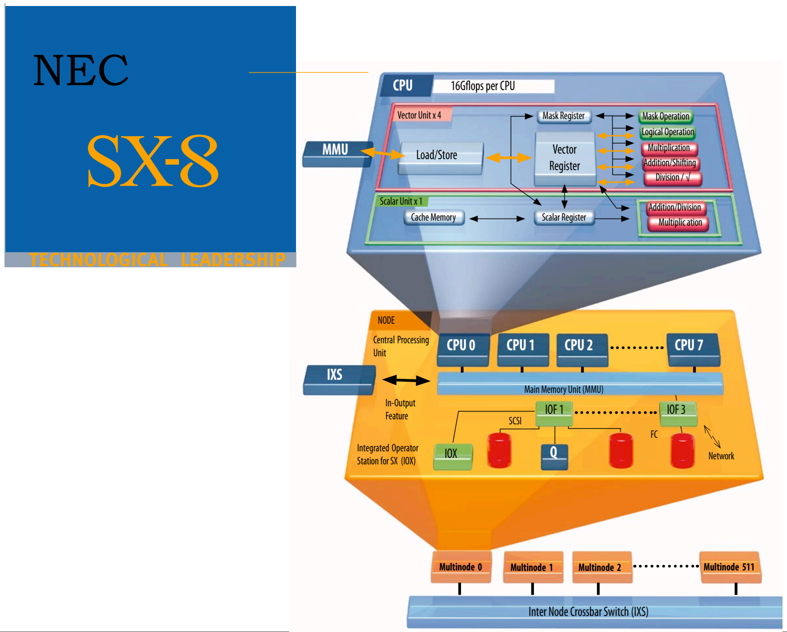
Trends in the Memory/Storage Subsystem



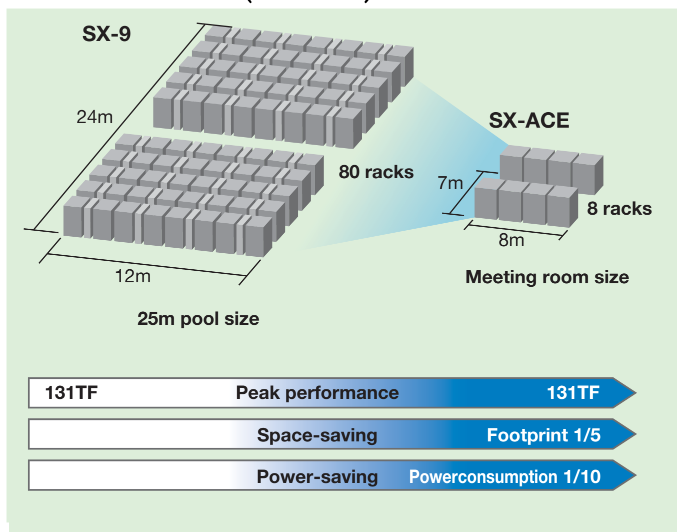


HLRS Stuttgart

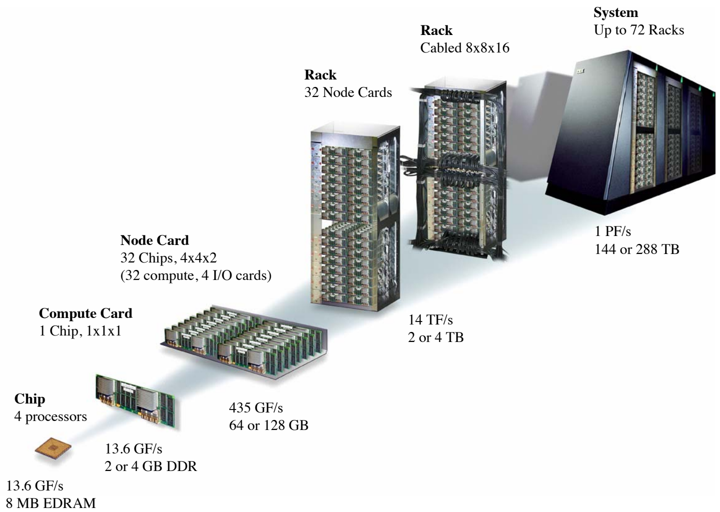
- video of building a computer



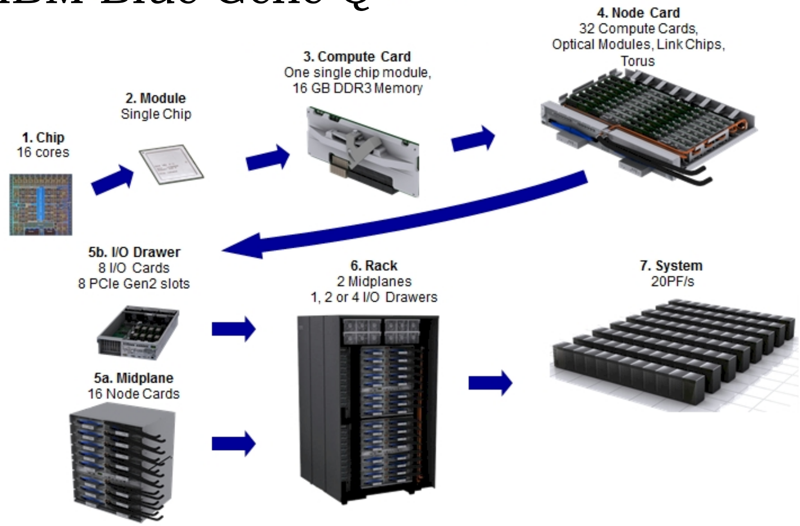
NEC SX-ACE (2014)



IBM Blue Gene P

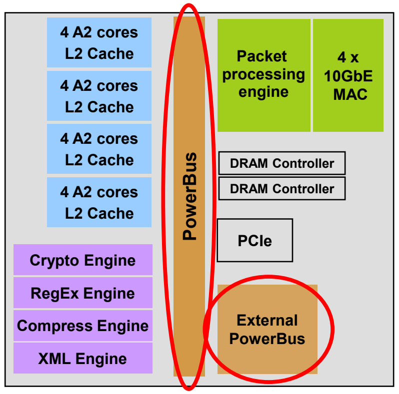


IBM Blue Gene Q

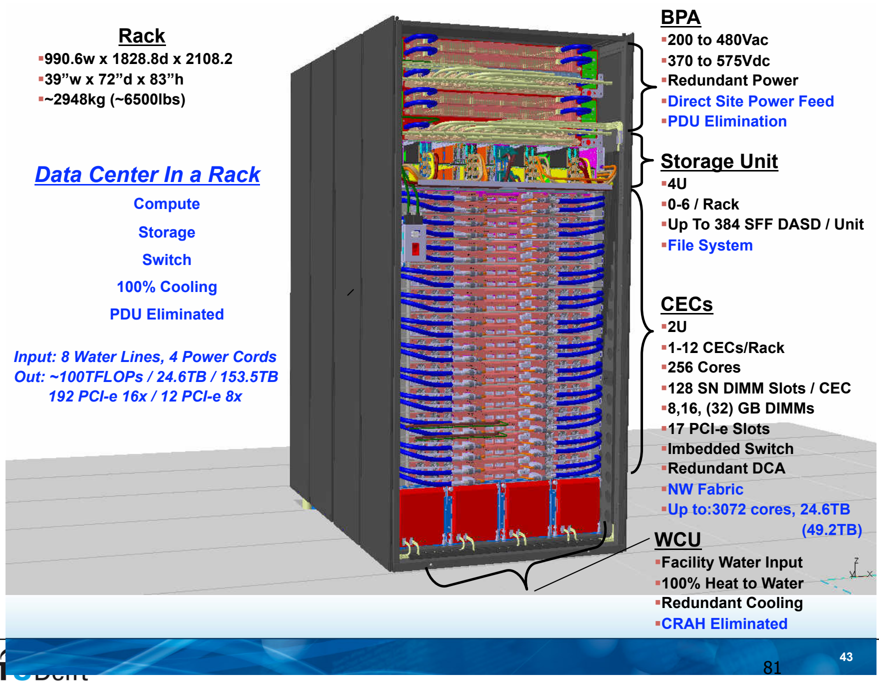


Interconnect Architecture

- *"All Peers"* architecture
  - Accel. and I/O are *first class citizens*
- Proven Power-Bus architecture
  - Independent CMD Network (one/cycle)
  - Two north, two south 16B data busses
  - ECC protected data paths
- 64 Byte Cache Line
- Cache Injection
  - Packets flow to / from Caches
  - New PBus commands
- 1.75 GHz operation
  - Asynchronous connection to AT Nodes and accelerators via PBICs
  - Synchronous connection to DRAM controllers
  - Three 4B 2.5 GHz EI3 external links (1,2, or 4 chip systems)



IBM Power 7 based


Rack
▪990.6w x 1828.8d x 2108.2
▪39"w x 72"d x 83"h
▪~2948kg (~6500lbs)
Data Center In a Rack
Compute
Storage
Switch
100% Cooling
PDU Eliminated
Input: 8 Water Lines, 4 Power Cords
Out: ~100TFLOPs / 24.6TB / 153.5TB
192 PCI-e 16x / 12 PCI-e 8x
BPA
▪200 to 480Vac
▪370 to 575Vdc
▪Redundant Power
▪Direct Site Power Feed
▪PDU Elimination
Storage Unit
▪4U
▪0-6 / Rack
▪Up To 384 SFF DASD / Unit
▪File System
CECs
▪2U
▪1-12 CECs/Rack
▪256 Cores
▪128 SN DIMM Slots / CEC
▪8,16, (32) GB DIMMs
▪17 PCI-e Slots
▪Imbedded Switch
▪Redundant DCA
▪NW Fabric
▪Up to:3072 cores, 24.6TB (49.2TB)
WCU
▪Facility Water Input
▪100% Heat to Water
▪Redundant Cooling
▪CRAH Eliminated
43

Bull


Solution for Peta Scalability
bullx
bullx R servers
bullx B servers
bullx S servers
Water cooling
Storage
Network
bullx Cluster Suite

SGI

The SGI Altix Ultraviolet (UV) System

Evolution

from

ccNUMA Shared Memory (SGI Origin)

to

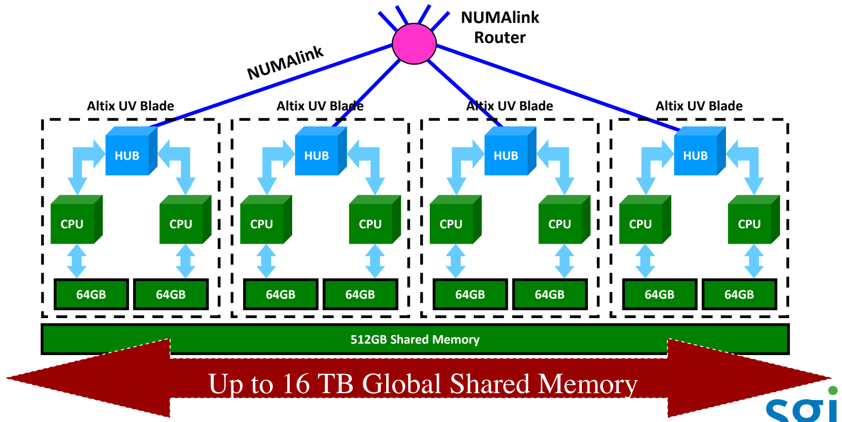
Partitioned Globally Addressable Shared Memory (SGI Altix 4700)

to

HW Accelerated Partitioned Globally Addressable System (SGI Altix UV)

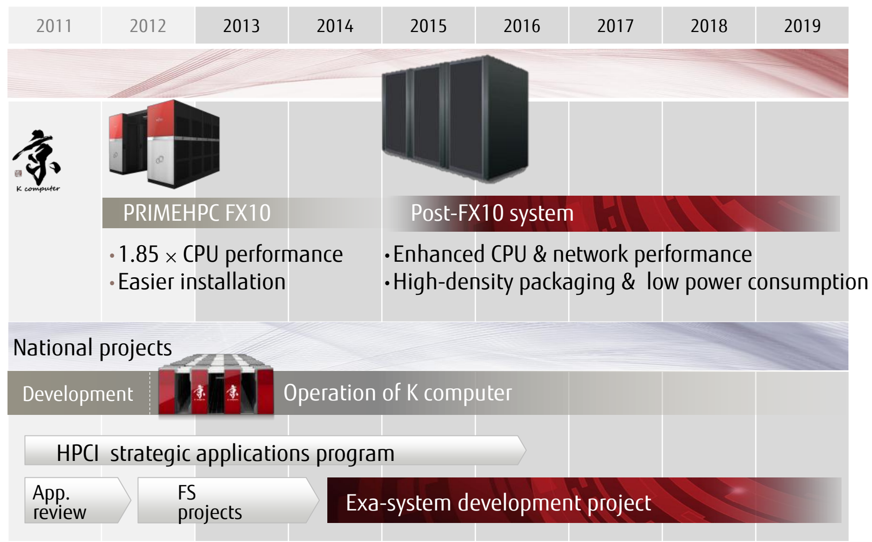
Globally Shared Memory System

- NUMAlink® 5 is the glue of Altix® UV 100/1000



Fujitsu

Road to exascale computing

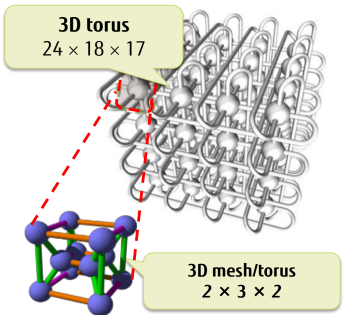


Interconnect of K computer: Tofu (torus fusion)

FUJITSU

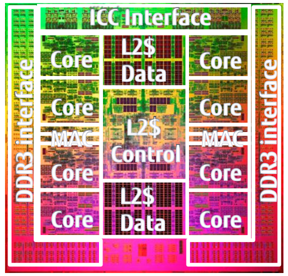
Technology	65 nm
DMA Engine	Send × 4 + recv. × 4
Link BW	5 + 5 GB/s × 10 ports
PCIe	16-lane Gen2
# of transistors	200 M

- 6D mesh/torus direct network
  K: (24 × 18 × 17) × (2 × 3 × 2)
  Low ave. hops and high bisectional BW
- Virtual 3D torus topology for apps.
- Hardware collective comm. support
- Congestion control by inserting GAPs



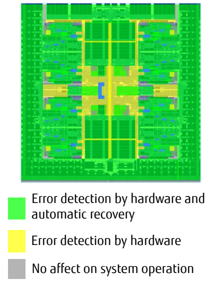
CPU chip of K computer: SPARC64 VIIIfx

FUJITSU



Technology	45 nm
Performance	128 GFLOPS
Memory bandwidth	64 GB/s
Power consumption	58 W
# of transistors	760 M

- Eight core out-of-order super scalar CPU
- HPC-ACE instruction set extension
- VISIMPACT hybrid execution model support
- Low power consumption design
- Highly reliable design inherited from mainframe

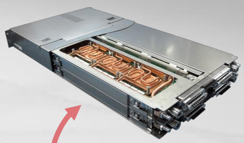


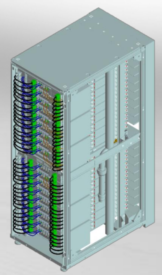
Feature and Configuration of Post-FX10

FUJITSU

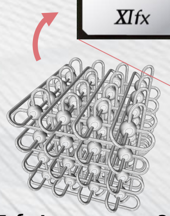
Fujitsu designed SPARC64™ XIfx
- 1TF~(DP)/2TF~(SP)
- 32 + 2 core CPU
- HPC-ACE2 support
- Tofu2 integrated

Chassis
- 1 CPU/1 node
- 12 nodes/2U Chassis
- Water cooled





Cabinet
- 200~ nodes/cabinet High-density
- 100% water cooled with EXCU (option)



Tofu Interconnect 2
- 12.5 GB/s×2(in/out)/link
- 10 links/node
- Optical technology

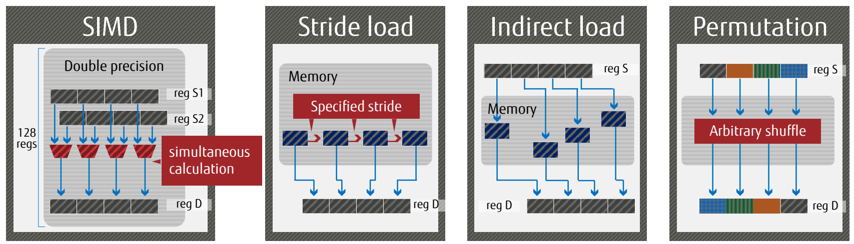
CPU Memory Board
- Three CPUs
- 3 x 8 Micron's HMCs
- 8 Finisar's opt modules, BOA, for inter-chassis connections

2

Flexible SIMD operations

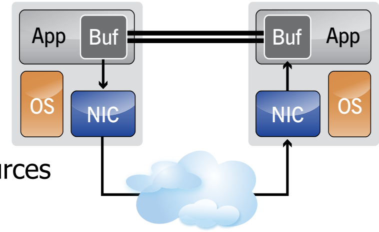
FUJITSU

- **New 256bit wide SIMD functions enable versatile operations**
  - **Four double-precision calculations**
  - **Stride load/store, Indirect (list) load/store, Permutation, Concatenation**



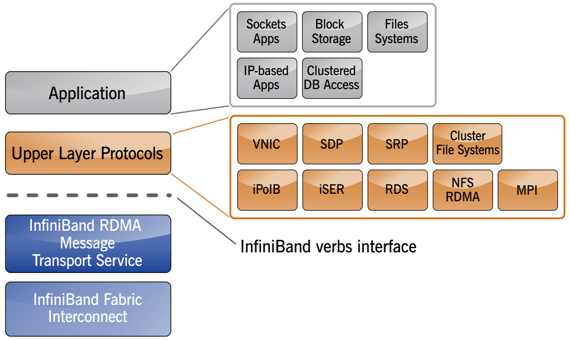
5

Infiniband



- Direct access to communication resources
  - provides a messaging service
  - no need to request the OS
  - directly communicate with another application through devices
- defines an 'API' set of behaviours (verbs)
  - OpenFabrics Alliance (OFA) stack op software for IB
  - The complete set of software components provided by the OpenFabrics Alliance is known as the Open Fabrics Enterprise Distribution – OFED

IB software stack



Benchmarks

- A way to compare the speed of different computer architectures.

Benchmarks

- Performance best determined by running a real application
  - Use programs typical of expected workload
  - Or, typical of expected class of applications
    e.g., compilers/editors, scientific applications, graphics, etc.
- Small benchmarks
  - nice for architects and designers
  - easy to standardize
  - can be abused
- SPEC (System Performance Evaluation Cooperative)
  - companies have agreed on a set of real program and inputs
  - valuable indicator of performance (and compiler technology)
  - can still be abused

Types of Benchmarks

Pros | **Cons**

- Representative

Actual Target Workload

- Very specific.
- Non-portable.
- Complex: Difficult to run, or measure.

- Portable.
- Widely used.
- Measurements useful in reality.

Full Application Benchmarks

- Less representative than actual workload.

- Easy to run, early in the design cycle.

Small "Kernel" Benchmarks

- Easy to "fool" by designing hardware to run them well.

- Identify peak performance and potential bottlenecks.

Microbenchmarks

- Peak performance results may be a long way from real application performance

SPEC: System Performance Evaluation Cooperative

The most popular and industry-standard set of CPU benchmarks

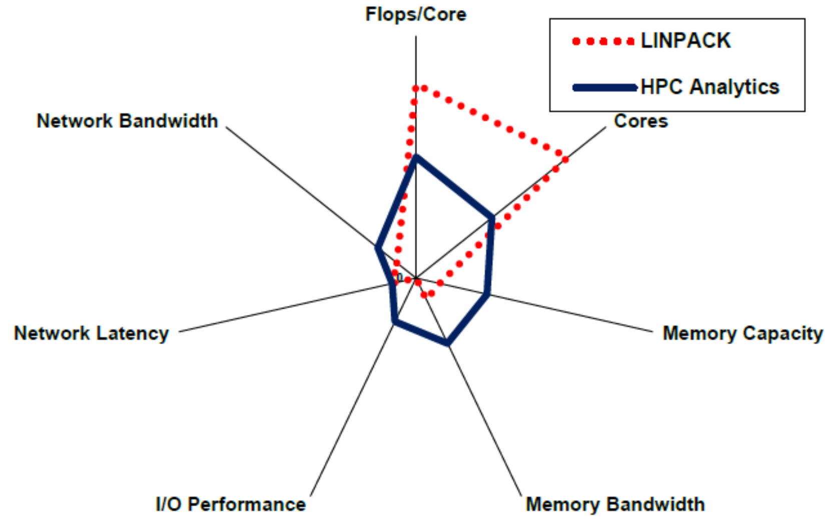
- SPEC CPU2006, combined performance of CPU, memory and compiler:
  - CINT2006 ("SPECint"), testing integer arithmetic, with programs such as compilers, interpreters, word processors, chess programs etc.
  - CFP2006 ("SPECfp"), testing floating point performance, with physical simulations, 3D graphics, image processing, computational chemistry etc.

http://www.spec.org/cpu/
http://www.cpubenchmark.net/

LINPACK N*N

- Customers use TOP500 list as one of the criteria to purchase machines
- TOP500 is based on LINPACK performance
- See http://www.top500.org/

Linpack compute resources



Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron Six Core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.60
2	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband / 2009 IBM	122400	1042.00	1375.78	2345.50
3	National Institute for Computational Sciences/University of Tennessee United States	Kraken XT5 - Cray XT5-HE Opteron Six Core 2.6 GHz / 2009 Cray Inc.	98928	831.70	1028.85	
4	Forschungszentrum Juelich (FZJ) Germany	JUGENE - Blue Gene/P Solution / 2009 IBM	294912	825.50	1002.70	2268.00
5	National SuperComputer Center in Tianjin/NUDT China	Tianhe-1 - NUDT TH-1 Cluster, Xeon E5540/E5450, ATI Radeon HD 4870 2, Infiniband / 2009 NUDT	71680	563.10	1206.19	
6	NASA/Ames Research Center/NAS United States	Pleiades - SGI Altix ICE 8200EX, Xeon QC 3.0 GHz/Nehalem EP 2.93 Ghz / 2009 SGI	56320	544.30	673.26	2348.00
7	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution / 2007 IBM	212992	478.20	596.38	2329.60

November 2009

Rank	Site	Computer
1	National Supercomputing Center in Tianjin China	Tianhe-1A - NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C NUDT
2	DOE/SC/Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz Cray Inc.
3	National Supercomputing Centre in Shenzhen (NSCS) China	Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU Dawning
4	GSIC Center, Tokyo Institute of Technology Japan	TSUBAME 2.0 - HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows NEC/HP
5	DOE/SC/LBNL/NERSC United States	Hopper - Cray XE6 12-core 2.1 GHz Cray Inc.
6	Commissariat a l'Energie Atomique (CEA) France	Tera-100 - Bull bullx super-node S6010/S6030 Bull SA
7	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband IBM
8	National Institute for Computational Sciences/University of Tennessee United States	Kraken XT5 - Cray XT5-HE Opteron 6-core 2.6 GHz Cray Inc.
9	Forschungszentrum Juelich (FZJ) Germany	JUGENE - Blue Gene/P Solution IBM
10	DOE/NNSA/LANL/SNL United States	Cielo - Cray XE6 8-core 2.4 GHz Cray Inc.

TUDelft

November 2010

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu	705024	10510.00	11280.38	12659.9
2	National Supercomputing Center in Tianjin China	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010 NUDT	186368	2566.00	4701.00	4040.0
3	DOE/SC/Oak Ridge National Laboratory United States	Cray XT5-HE Opteron 6-core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.0
4	National Supercomputing Centre in Shenzhen (NSCS) China	Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 / 2010 Dawning	120640	1271.00	2984.30	2580.0
5	GSIC Center, Tokyo Institute of Technology Japan	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows / 2010 NEC/HP	73278	1192.00	2287.63	1398.6
6	DOE/NNSA/LANL/SNL United States	Cray XE6, Opteron 6136 8C 2.40GHz, Custom / 2011 Cray Inc.	142272	1110.00	1365.81	3980.0
7	NASA/Ames Research Center/NAS United States	SGI Altix ICE 8200EX/8400EX, Xeon HT QC 3.0/Xeon 5570/5670 2.93 Ghz, Infiniband / 2011 SGI	111104	1088.00	1315.33	4102.0
8	DOE/SC/LBNL/NERSC United States	Cray XE6, Opteron 6172 12C 2.10GHz, Custom / 2010 Cray Inc.	153408	1054.00	1288.63	2910.0
9	Commissariat a l'Energie Atomique (CEA) France	Bull bullx super-node S6010/S6030 / 2010 Bull	138368	1050.00	1254.55	4590.0

TUDelft

November 2011

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 31S1P 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 2.0GHz, 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

June 2013

RANK	SITE	SYSTEM	CORES	RMAX (TFLOP/S)	RPEAK (TFLOP/S)	POWER (KW)
1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P 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 2.0GHz, 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	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Cray Inc.	115,984	6,271.0	7,788.9	2,325
7	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
8	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
9	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
10	Government United States	Cray CS-Storm, Intel Xeon E5-2660v2 10C 2.2GHz, Infiniband FDR, Nvidia K40 Cray Inc.	72,800	3,577.0	6,131.8	1,499

November 2014

Projected Performance Development