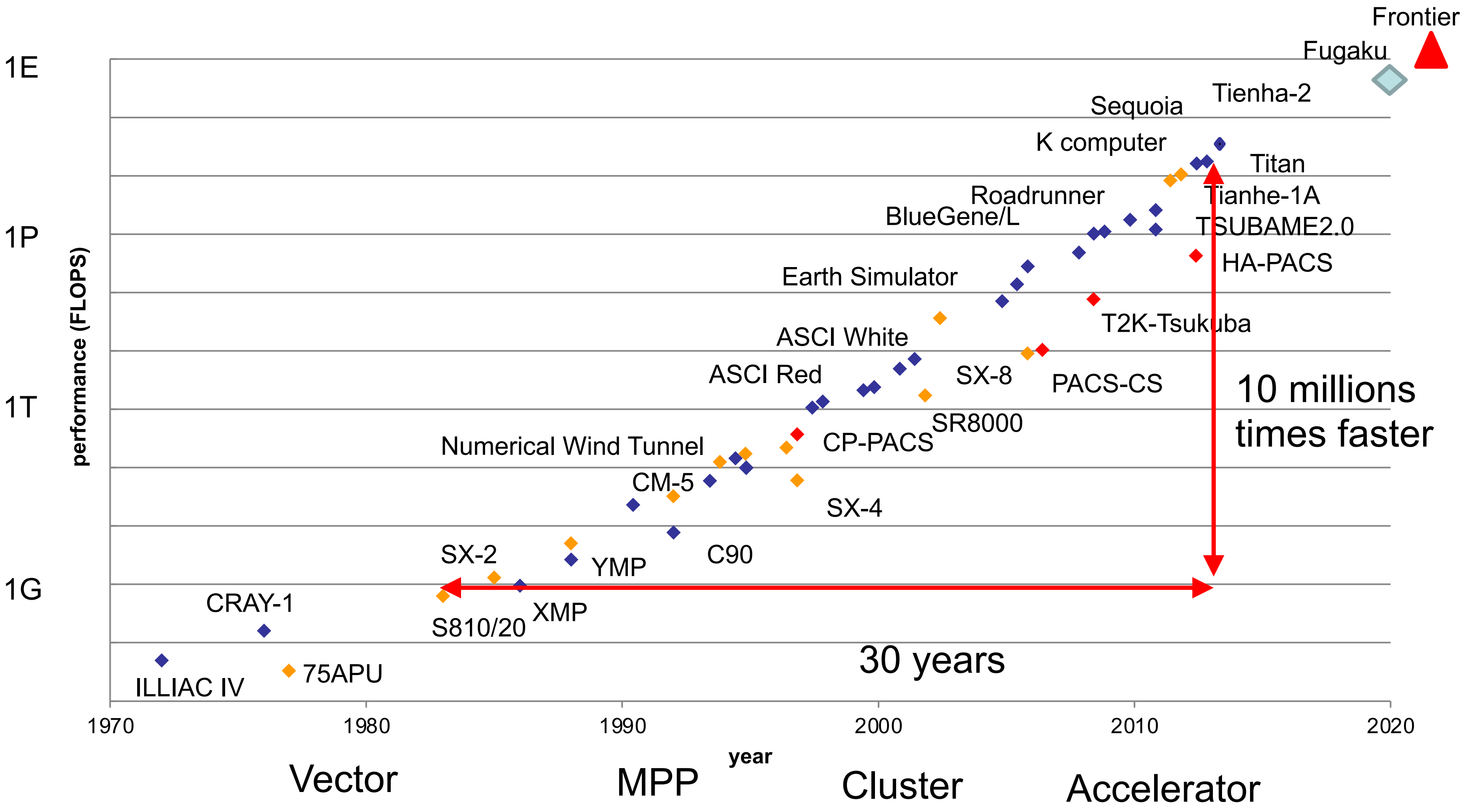
Japan-Korea HPC Winter School & High Performance Parallel Computing Technology for Computational Sciences #2 “Parallel Processing Systems”

Ryohei Kobayashi
kobayashi@cs.tsukuba.ac.jp
Center for Computational Sciences
University of Tsukuba

Contents

- History of parallel systems
- Architecture of parallel systems
- Interconnection Network of parallel systems
- Overview of Supercomputers

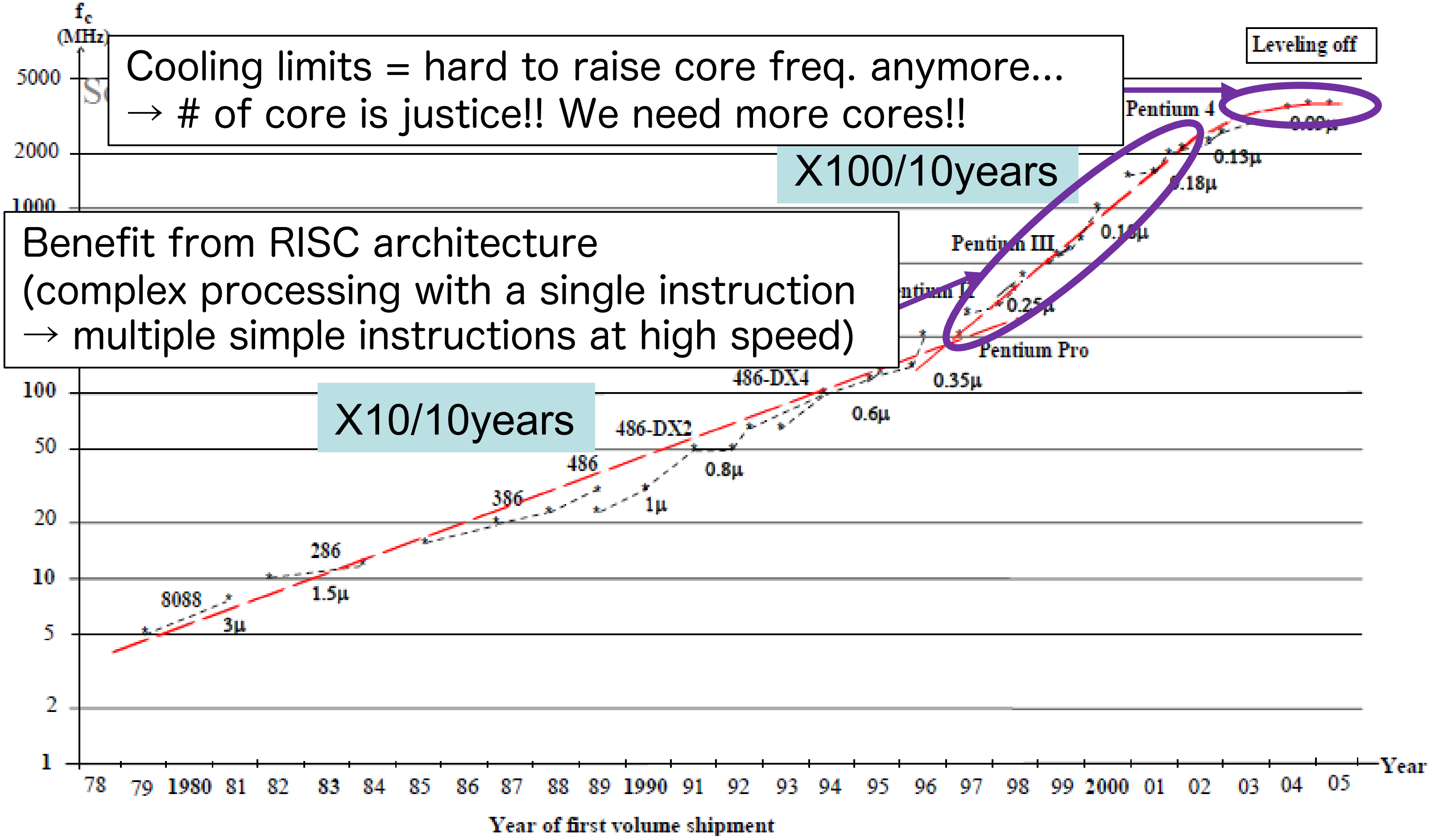
History of supercomputer



Advances in parallel computer -- processor level --

- Vector processor ⇒ many supercomputer used vector processors in 20 years ago
    - Single processor can calculate array processing in high speed
- Scalar processor
  x86 (IA32), Power, Itanium (IA64), Sparc
- Recent trend in processor
    - multi-core processor becomes popular
        - Intel & AMD ⇒ 8 ～12 core (x86)
    - many-core（8～512 cores）processor is available for accelerator
        - IBM Cell Broadband Engine (8 core)
        - ClearSpeed (96 core×2)
        - GPU (NVIDIA K20X 896 DP unit)
        - Intel Xeon Phi (72 core)
        - Fujitsu A64FX (48 cores)

Clock speed of scalar processors



Cooling limits = hard to raise core freq. anymore...
→ # of core is justice!! We need more cores!!

X100/10years

Benefit from RISC architecture
(complex processing with a single instruction
→ multiple simple instructions at high speed)

X10/10years

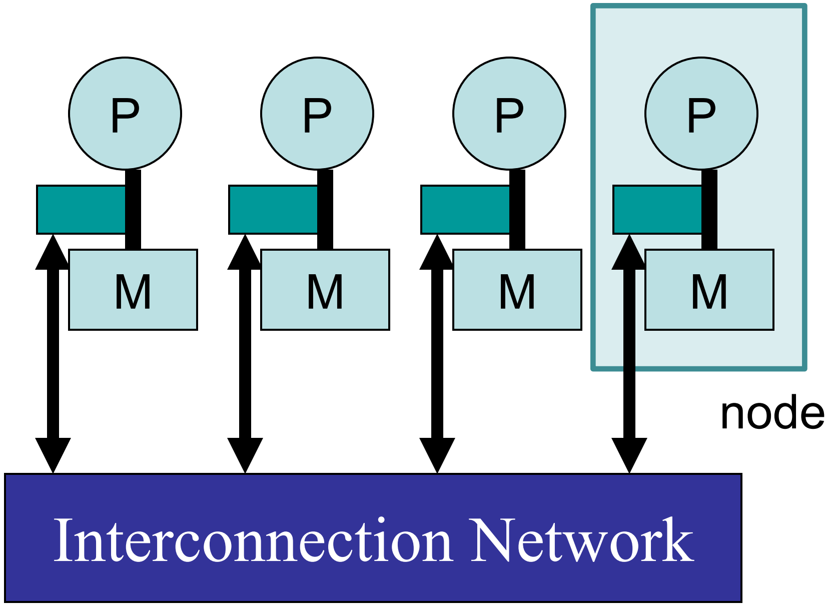
Contents

- History of parallel systems
- Architecture of parallel systems
- Interconnection Network of parallel systems
- Overview of Supercomputers

Memory architecture of parallel systems

- Distributed memory system
  Each processor has own memory that cannot be directly accessed by other processors, and accesses the remote data by message passing using interconnection network.
- Shared memory system
  Each processor has physically shared memory, and accesses the memory by normal load/store instructions.
- Hybrid memory system
  Combination of shared-memory system and distributed-memory system. In a node, it is shared-memory system using multi-core CPU, and between nodes, it is distributed-memory system using interconnection network.

Distributed-memory system (1)



Message passing
between any processors

P ... Processor
M ... Memory
🟩 NIC (network interface controller)

●Nodes are the unit of parallel systems, and each node is a complete computer system with CPU and memory. Nodes are connected by interconnection network via NIC.

●A process runs on each node and communicates data (message) between nodes through network.

●Building system is easy and scalability is high.

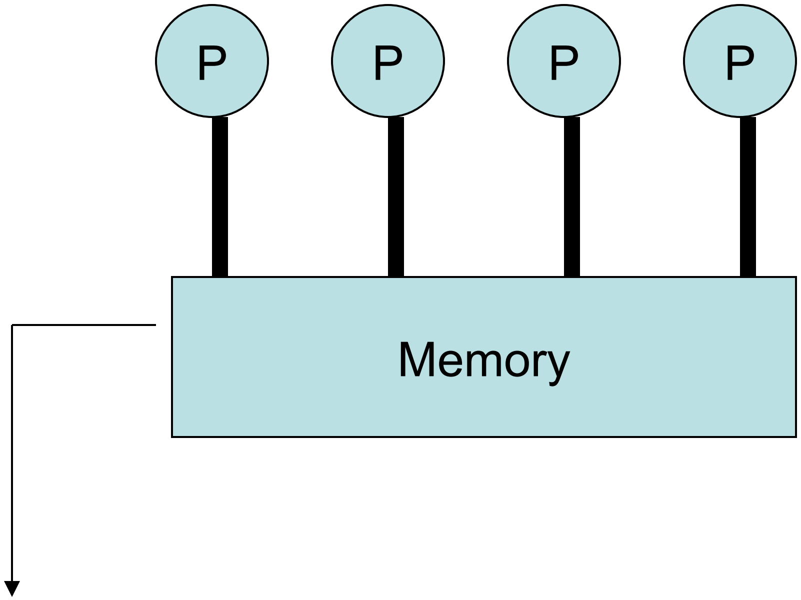
◆MPP：Massively Parallel Processor

◆Cluster computer

Distributed-memory system (2)

- Program on each node should communicate to other nodes explicitly by message passing, user programming is complicated.
  - MPI (Message Passing Interface) is a standard tool to communication
  - Typical style of parallel application is relatively easy to write a program such as data parallel of domain decomposition or master/worker processing
- System performance depends on performance of interconnection network as well as performance of processor and memory.
- Typical implementation of MPP in late 1980, and also basic architecture of current PC cluster

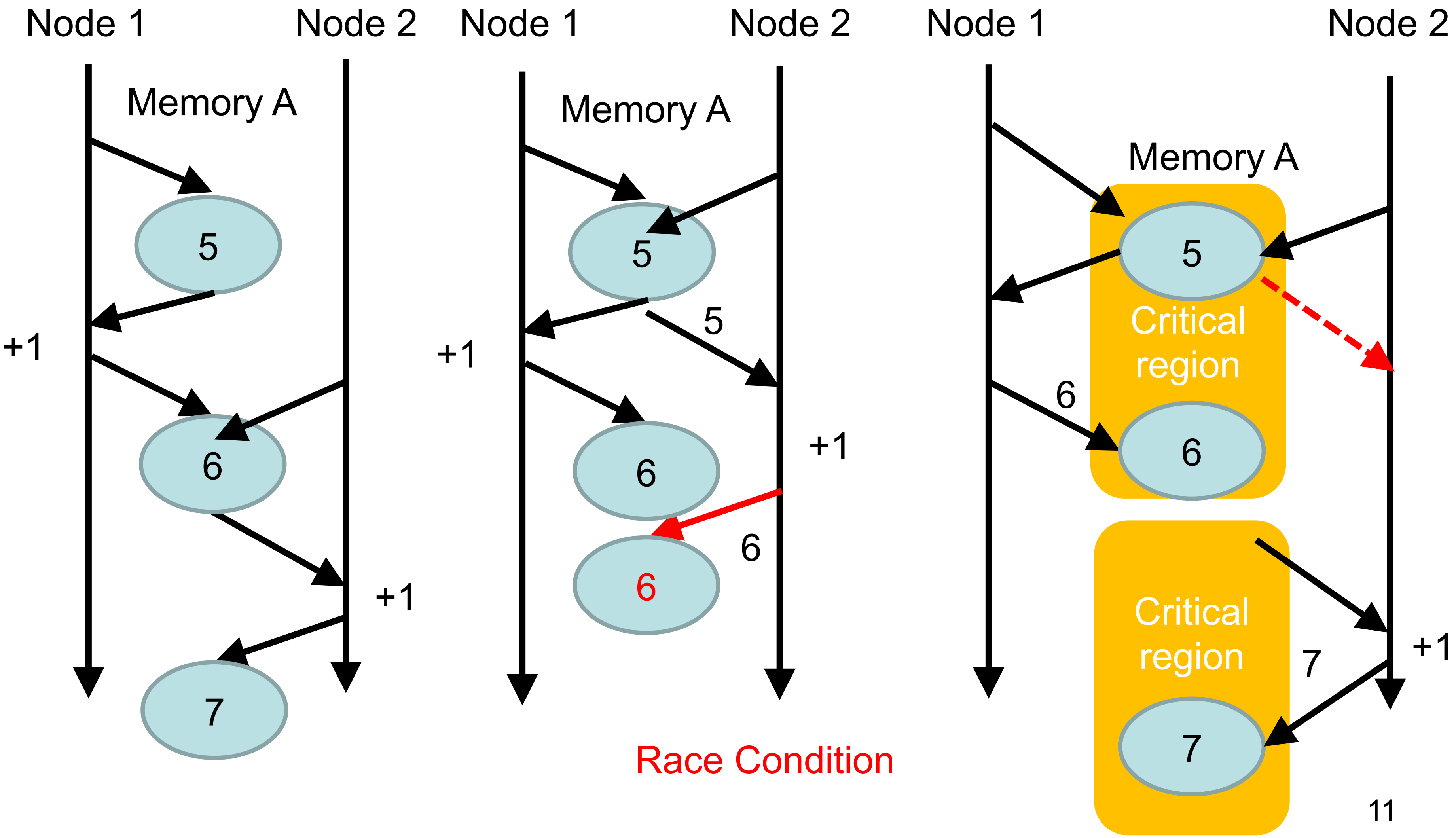
Shared-memory system (1)



Memory should arbitrate simultaneous requests from multiple processors.

- ●Multiple processor access the same memory
- ●Each program (thread) on processor accesses data on memory freely. In other words, it should be care of race condition for multiple updates.
- ●Small to medium scale server
- ●Recent multi-core processor is shared memory in a processor.
- ●Architecture is classified to SMP and NUMA.

Access conflict in shared memory

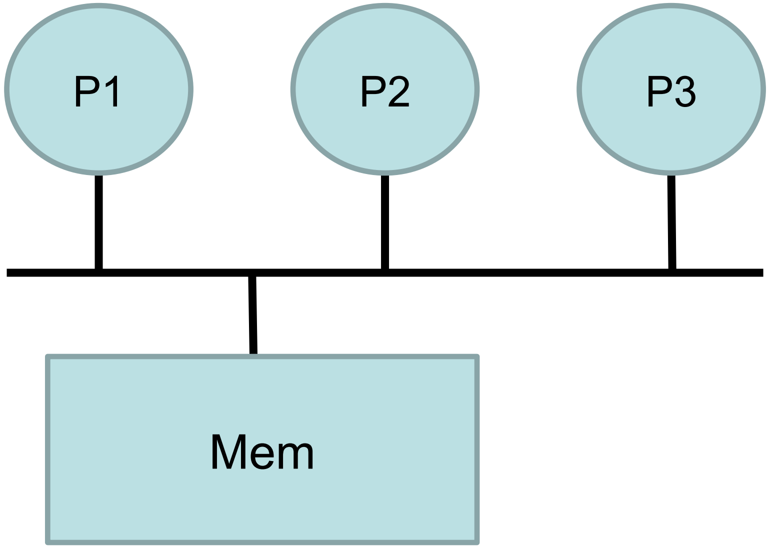


Shared-memory system (2)

- Programming on shared memory is easy for users.
  - Multithread programing model (POSIX thread, etc)
  - Programming tools based on shared memory (OpenMP: directive base programming)
- Shared-memory is simple as model, but hardware will be complicated to realize the model with high performance because “memory” is a quite primitive element of computer.
- Memory access becomes bottleneck when many processor access a location of memory
  - It is difficult to achieve scalability of system (about 100 processors will be a limit)

Architecture of shared memory

- Shared memory bus is the simplest shared memory system, but it cannot achieve scalability.
  - Shared bus was popular in PC cluster.
  - The Bus becomes bottleneck (bus is occupied by a transaction in a time)
  - To reduce the overhead of the bus conflict, each node has coherent cache.



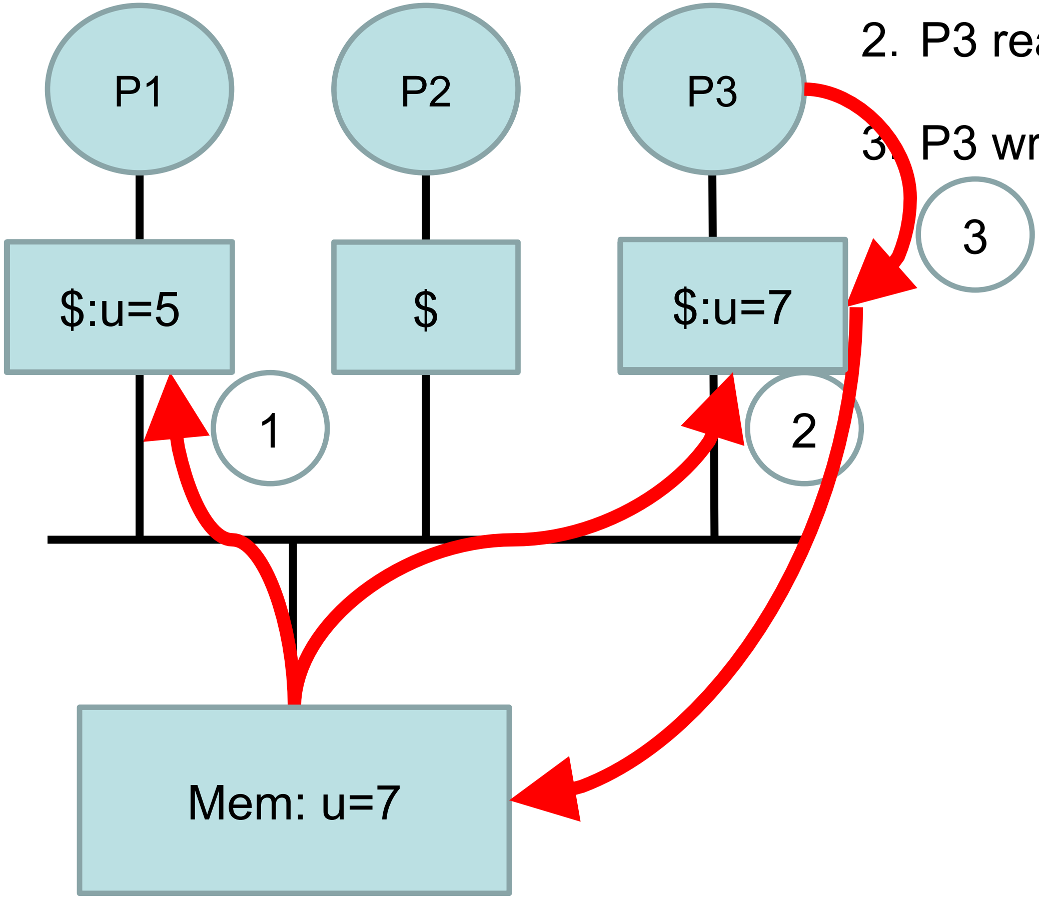
Non-coherent cache

1. P1 read Mem:u

4. P1 re-read Mem:u

2. P3 read Mem:u

3. P3 write Mem:u



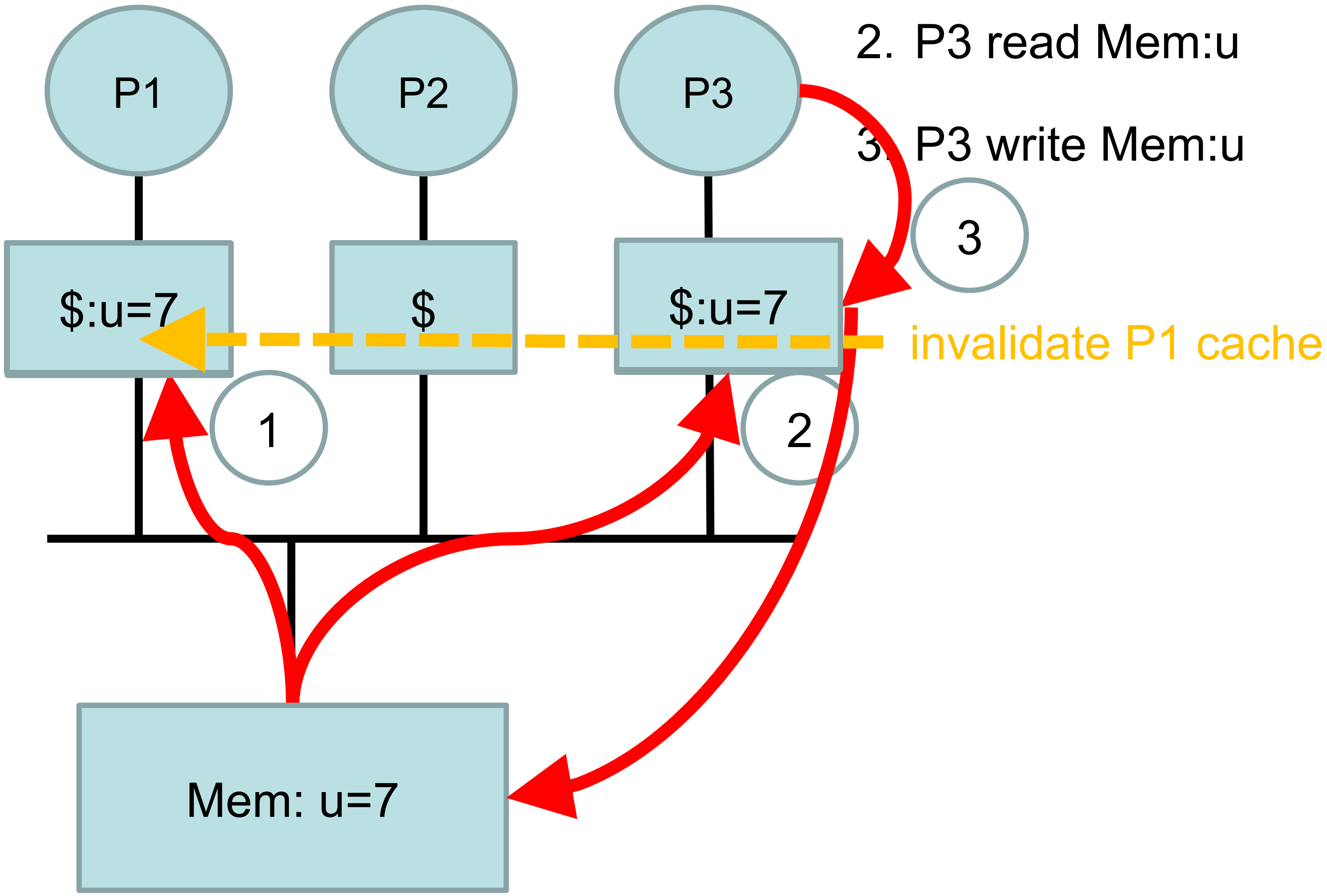
Read old value u=5
from cache

Coherent cache

1. P1 read Mem:u

4. P1 reread Mem:u

Read correct value from memory



2. P3 read Mem:u

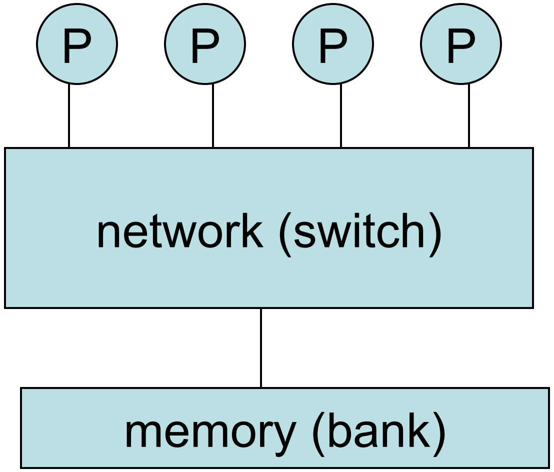
3. P3 write Mem:u

Architecture of shared memory

- How to avoid access conflict to shared memory
  - memory bank: address blocks are distributed to memory modules
  - split transaction: request and reply for memory are separated.
  - crossbar network：processors and memories are connected by switch not a bus.
  - coherent cache： each processor has own cache. If other processor update the memory, cache will be updated automatically.
  - NUMA (Non-Uniformed Memory Access)：memory modules are distributed, and the distance to each memory is different.

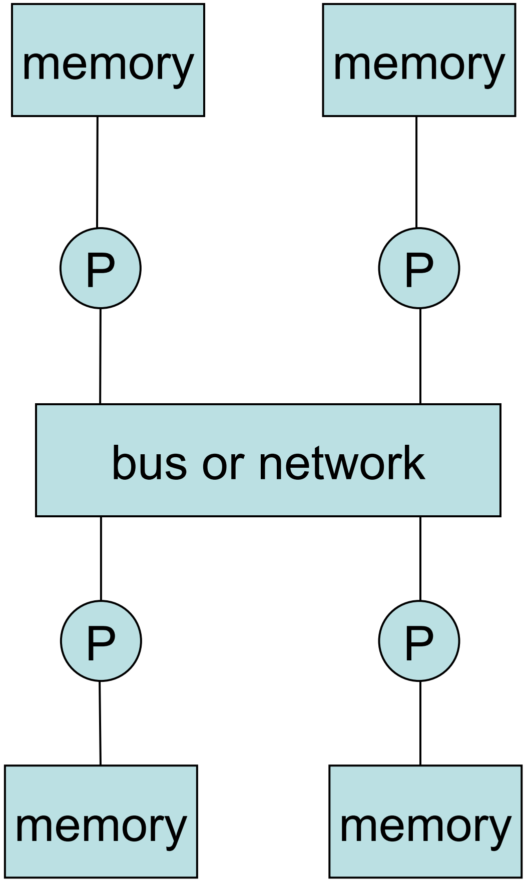
Symmetric Multi-Processor (SMP)

- The distance to any memory from a processor is same
- Shared bus or switches connect multiple processors and memory modules evenly.
- Multiprocessor node with previous Intel processor was SMP
- Large scale SMP system: Fujitsu HPC2500 and Hitachi SR16000
- Coherent cache is usually used
- Processor don’t have to consider the data location
- When memory access is concentrated, the performance is reduced

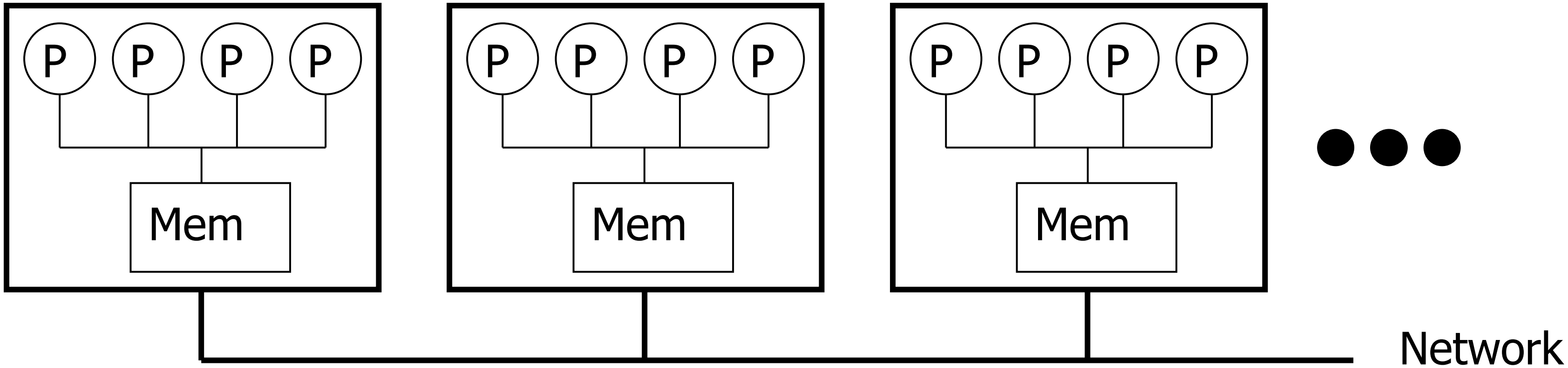


Non-Uniformed Memory Access (NUMA)

- A processor has own memory (local memory)
- Processor can access the memory of other processor (remote memory) via shared bus or network between processors
- It needs excessive time to access the remote memory（non-symmetric）
- AMD used NUMA from Opteron in 2003. Intel also uses NUMA from Nehalem in 2008.
- Large scale NUMA system: SGI Origin, Altix series.
- If data is distributed in each memory, and processor accesses to local memory, the memory performance will be increased (memory affinity)



Hybrid memory system



- Combination of shared memory and distributed memory
- Node itself is a shared memory multiprocessor system (SMP or NUMA).
- Each node is connected to other nodes with network, and access the remote memory with distributed memory.
- Hybrid system becomes popular because CPU becomes multi-core processor where each core is a shared memory architecture.

Parallel system with Accelerator

- Each node includes not only CPU but also accelerator that is a hardware to accelerate arithmetic operations.
  - GPU (Graphic Processing Unit)
    recently called GPGPU (General Purpose GPU), available general programming
  - FPGA (Field Programmable Gate Array)
    Reconfigurable hardware for specific purpose
  - General accelerator
    ClearSpeed, etc. (obsolete!)
  - Processor itself is hybrid architecture with fat core and thin core
    CBE (Cell Broadband Engine) ⇒ LANL Roadrunner
    (obsolete!)

MultiCore, ManyCore, GPU

	Multi core	Many core	GPU
Ex.	Intel Xeon Platinum 8280	Intel Xeon Phi 7250P	NVIDIA Tesla V100
# of cores	28	72	5120
Freq.	2.7 GHz	1.2GHz	1.53 GHz
Perf.	605 GFLOPS	1.1 TFLOPS	7.8 TFLOPS
Memory Capacity	1 TB	16 GB	6 GB
Memory Bandwidth	141 GB/s	450 GB/s	900 GB/s
Power	205 W	200 W	300 W
Program	C, OpenMP, MPI	C, OpenMP, MPI	CUDA　(C, Fortran)
Exec. Model	MIMD	MIMD	STMD

Contents

- History of parallel systems
- Architecture of parallel systems
- Interconnection Network of parallel systems
- Overview of Supercomputers

Interconnection Network

- Aim
  - Explicit data exchange on distributed memory architecture
  - Transfer data and control message on shared memory architecture
- Classification
  - static (direct) / dynamic (indirect)
  - diameter (distance)
  - degree (number of links)
- Performance metric
  - throughput
  - latency

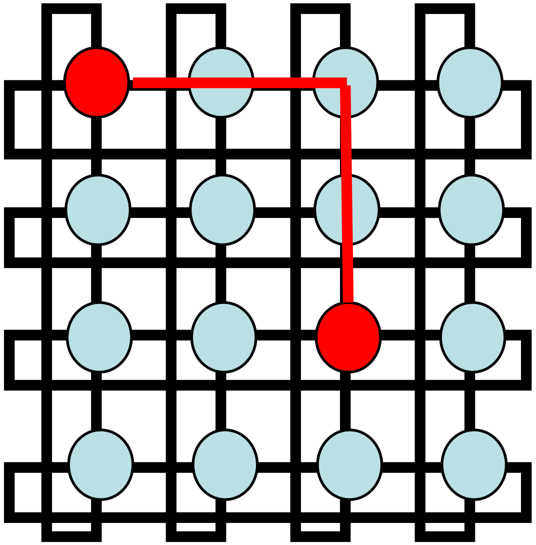
Direct network

- All network nodes have processor (or memory) with multiple links connected to other nodes.
- In other words, direct connection between nodes, and no switches.
- Messages are routed on nodes.
- Typical topology of direct network
  - 2-D/3-D Mesh/Torus
  - Hypercube
  - Direct Tree

Examples of Direct network

Mesh/Torus (k-ary n-cube)

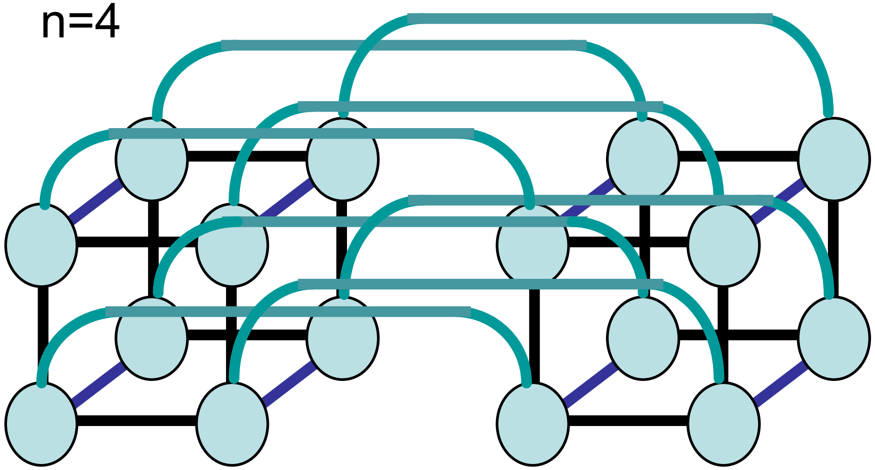
4 x 4 2D torus



Cost: N (=k^n)
Diameter: n(k-1) in mesh
nk/2 in torus

Hypercube (n-cube)

2x2x2x2

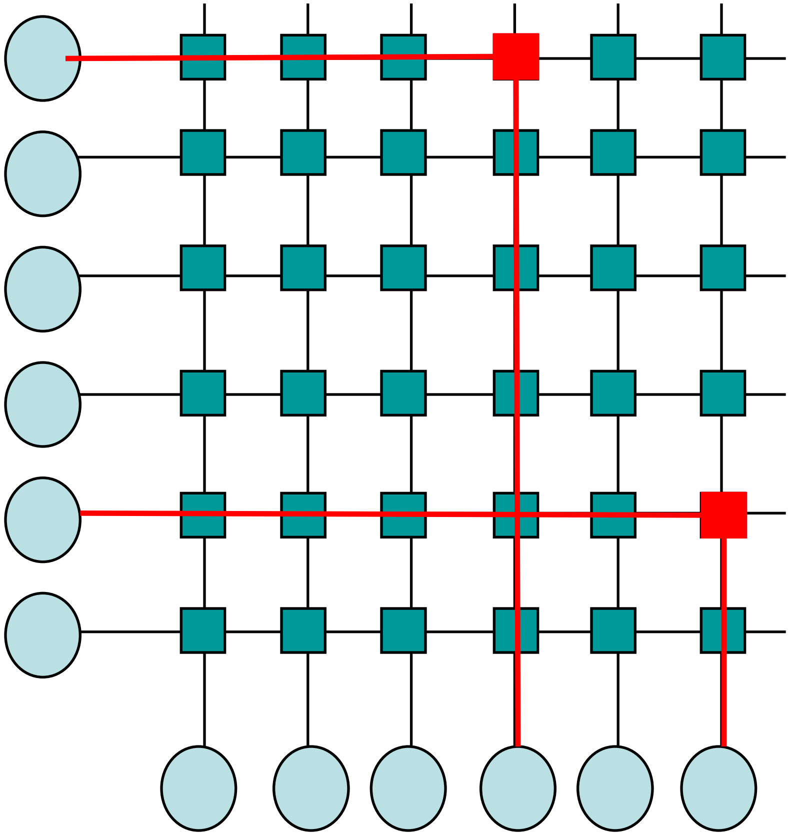


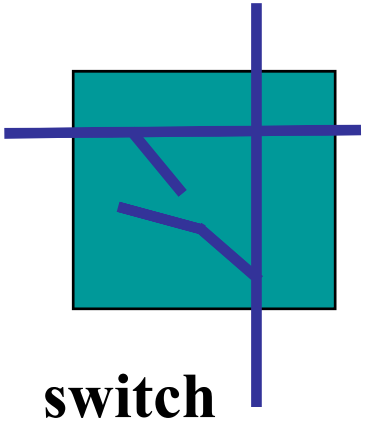
Cost: N (=2^n)
Diameter: n

Indirect network

- Each node (processor) has a link, or multiple links.
- The link connects to switch that connects to other switches.
- Messages are routed on switches.
- No direct connection between processors
- Typical topology of indirect network
  - Crossbar
  - MIN (Multistage Interconnection Network)
  - HXB (Hyper-Crossbar)
  - Tree (Indirect)
  - Fat Tree

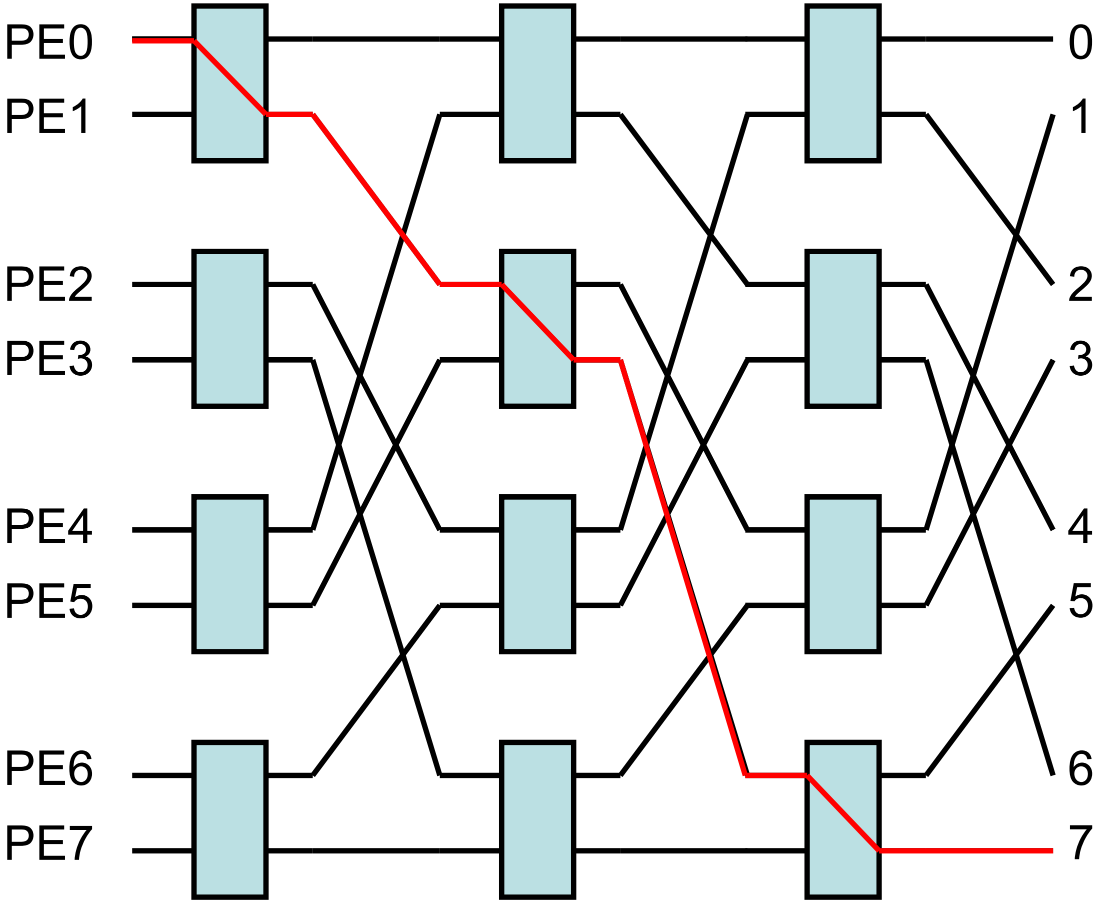
Crossbar



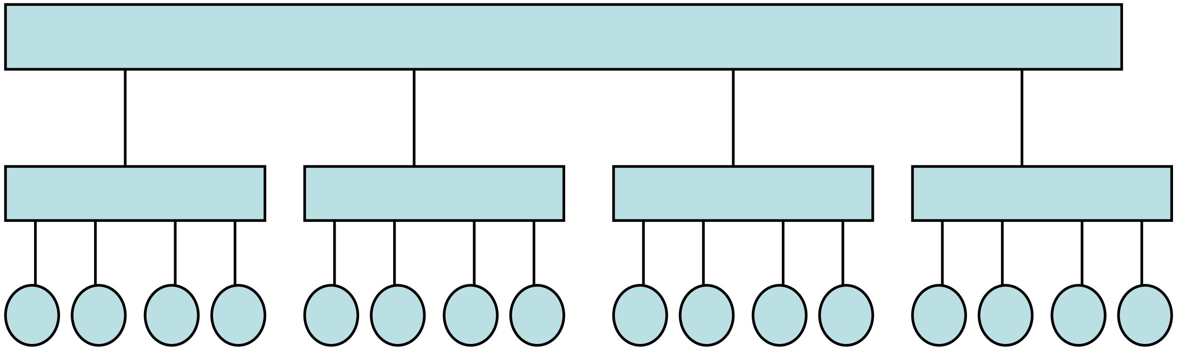


Cost: N^2
Diameter: 1

MIN (Multi-stage Interconnection Network)



Cost: *N* log*N*
Diameter: log*N*

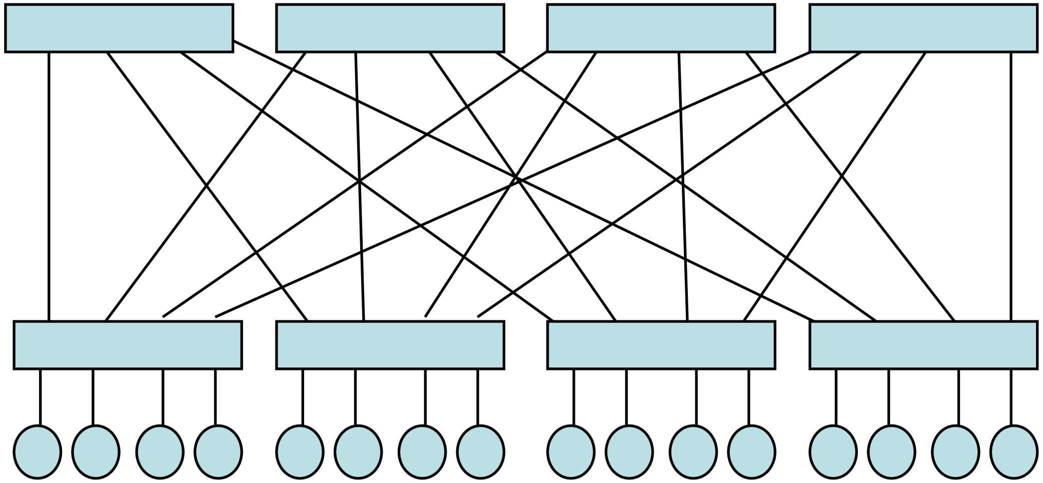


Tree

Diameter: $2\log_k N$

Fat Tree

Diameter: $2\log_k N$



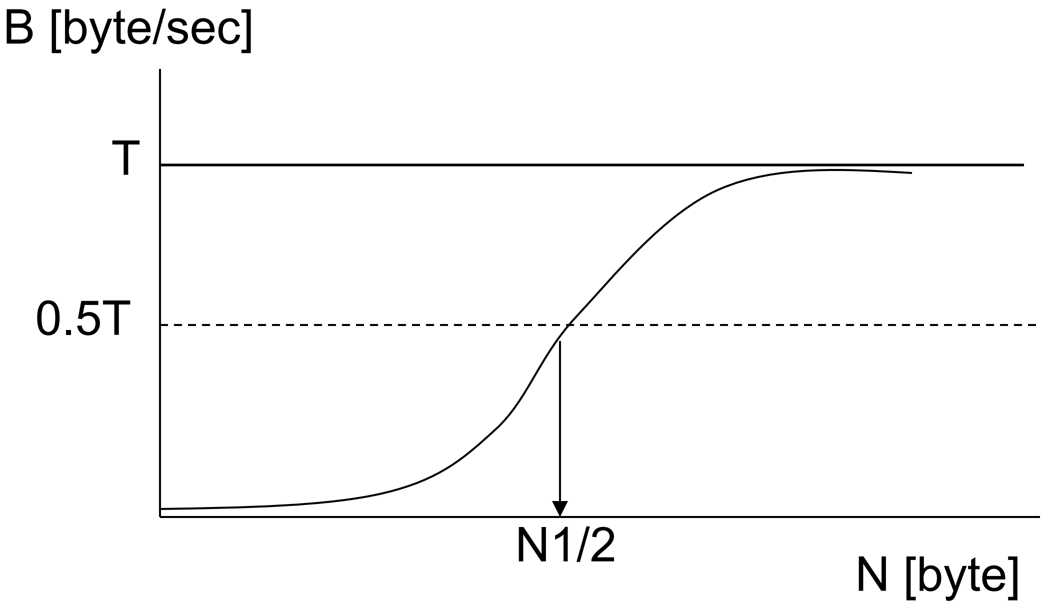
Performance metric of interconnection

- Throughput
  - The size of transferred message in a time unit
  - Unit:[Byte/sec]
    (or bit/sec, where 8bit = 1byte is not always true by encoding such as 8b/10b)
- Latency
  - Narrow： the time from the source sending the beginning of a packet to the destination receiving it. (here this definition used)
  - Wide： the time from the source sending a packet to the destination receiving its whole data. It depends on the size of packet.
  - Unit:[sec]

Performance and message size

- The relation between the size of message N[byte] and the effective bandwidth B[byte/sec] is the following equations and graph where there is no conflict on interconnect network, the network throughput T[byte/sec], latency L[sec], and the total transfer time t [sec] .

$t = L + N/T \qquad B = N / t$



$N_{1/2}$: the message length to achieve the half of theoretical peak performance, and calculated in theory.

$N_{1/2}[byte] = L \times T$

$N_{1/2}$ means that L is dominant if N is less than $N_{1/2}$, and T is dominant if N is more than $N_{1/2}$. Smaller $N_{1/2}$ shows the network has good performance for shorter message.

Contents

- History of parallel systems
- Architecture of parallel systems
- Interconnection Network of parallel systems
- Overview of Supercomputers

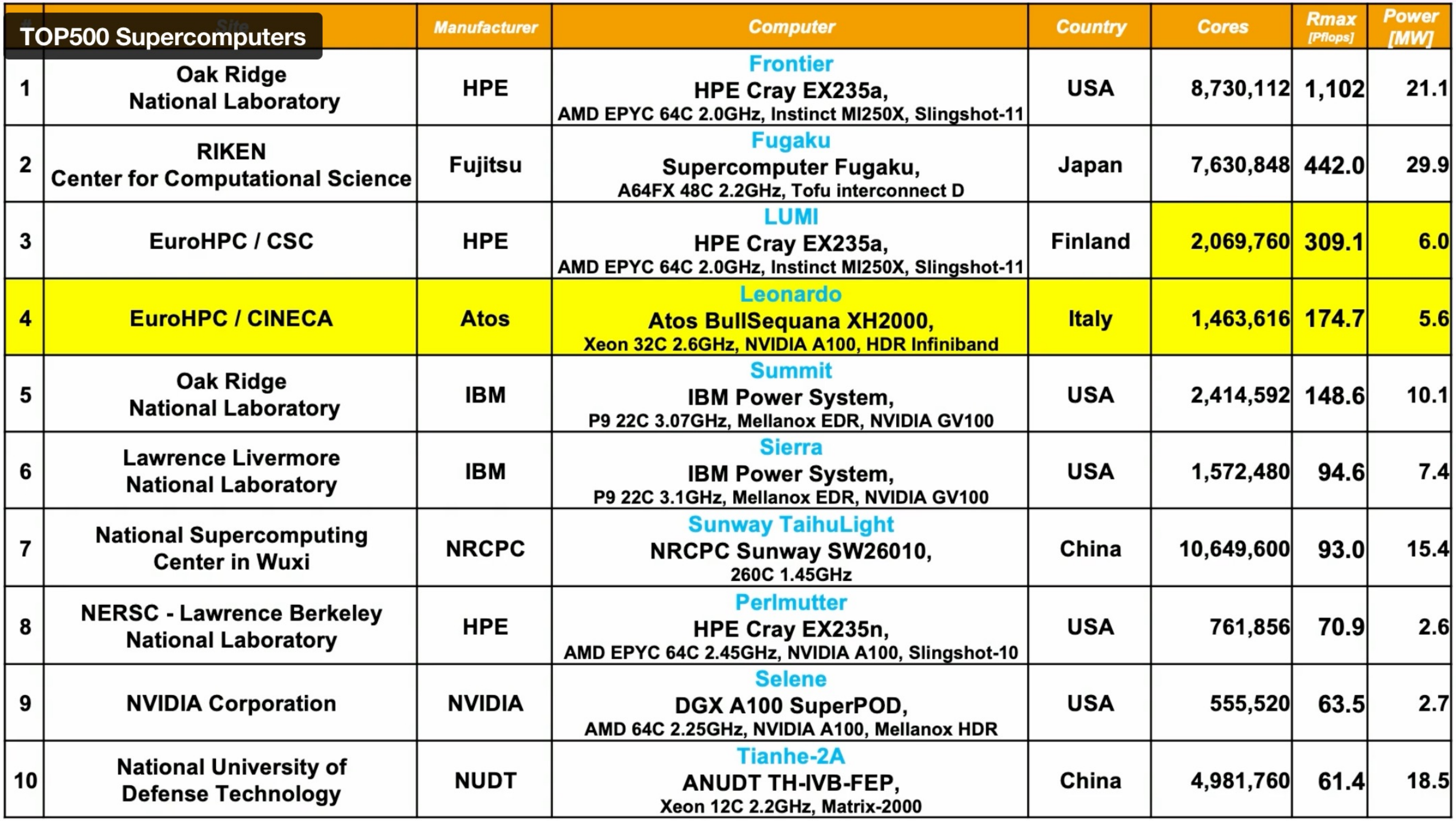
Overview of Supercomputers

- System classification
  - MPP
    - Univ. of Tsukuba/Hitachi CP-PACS (SR2201)
    - RIKEN/Fujitsu Fugaku supercomputer
    - LLNL /IBM BlueGene/Q Sequia
    - ORNL/Cray XK7 Titan
  - Large scale parallel vector computer
    - NEC Earth simulator
  - Scalar parallel computer (including cluster)
    - Univ. of Tsukuba･Hitachi･Fujitsu PACS-CS
    - Univ. of Tsukuba･Tokyo･Kyoto/Appro･Hitachi･Fujitsu T2K
  - Hybrid parallel computer with accelerator
    - LANL/IBM Roadrunner
    - Tokyo Tech/HP (SGI) TSUBAME3.0
    - SYU/NUDT Tianhe-2
    - Univ. of Tsukuba/NEC Cygnus

TOP500 List

- The list ranked supercomputers in the world by their performance on one index based on user submission.
- Index = performance (FLOPS) of LINPACK (solver for a dense system of linear equations by Gaussian elimination)
- update the list half-yearly, every June in ISC and November in SC http://www.top500.org
- Easy to understand because of one index
- Benchmark characteristic
  - The kernel part of Gaussian elimination is implemented by small scale matrix multiplication, and cache architecture can reuse the highly part of data. Good estimation of peak performance.
  - The performance is not highly dependent on the network performance
- Since LINPACK does not require the high memory bandwidth and high network performance, “Is LINPACK suitable for HPC benchmark ?” is discussed, but LINPACK is one of well-known performance index. (HPCC(HPC Challenge Benchmark) is another benchmark for HPC)

TOP500 on Nov. 2022 (www.top500.org)



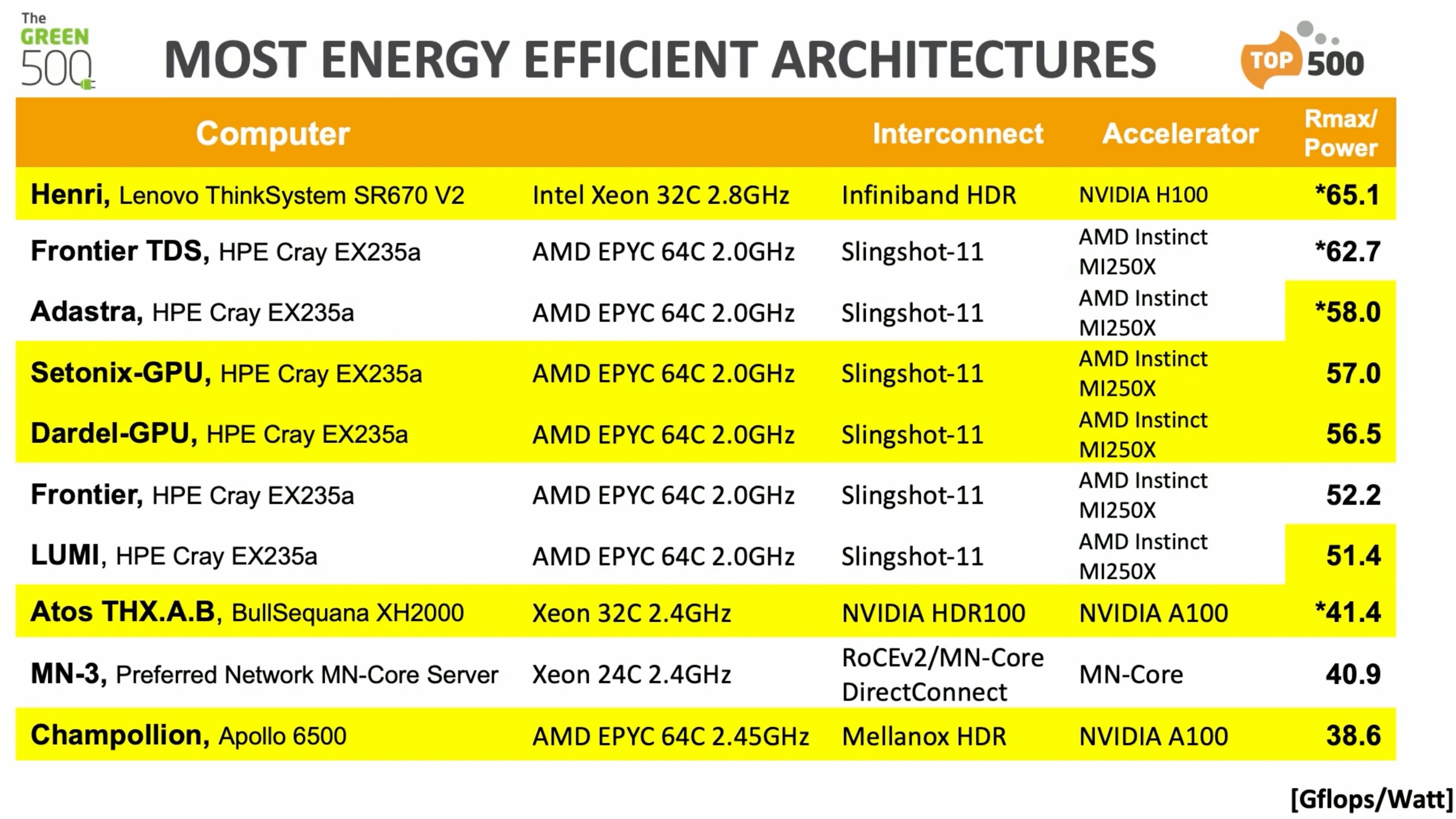
TOP500 Supercomputers

#	Site	Manufacturer	Computer	Country	Cores	Rmax [Pflops]	Power [MW]
1	Oak Ridge National Laboratory	HPE	Frontier HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	USA	8,730,112	1,102	21.1
2	RIKEN Center for Computational Science	Fujitsu	Fugaku Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D	Japan	7,630,848	442.0	29.9
3	EuroHPC / CSC	HPE	LUMI HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	Finland	2,069,760	309.1	6.0
4	EuroHPC / CINECA	Atos	Leonardo Atos BullSequana XH2000, Xeon 32C 2.6GHz, NVIDIA A100, HDR Infiniband	Italy	1,463,616	174.7	5.6
5	Oak Ridge National Laboratory	IBM	Summit IBM Power System, P9 22C 3.07GHz, Mellanox EDR, NVIDIA GV100	USA	2,414,592	148.6	10.1
6	Lawrence Livermore National Laboratory	IBM	Sierra IBM Power System, P9 22C 3.1GHz, Mellanox EDR, NVIDIA GV100	USA	1,572,480	94.6	7.4
7	National Supercomputing Center in Wuxi	NRCPC	Sunway TaihuLight NRCPC Sunway SW26010, 260C 1.45GHz	China	10,649,600	93.0	15.4
8	NERSC - Lawrence Berkeley National Laboratory	HPE	Perlmutter HPE Cray EX235n, AMD EPYC 64C 2.45GHz, NVIDIA A100, Slingshot-10	USA	761,856	70.9	2.6
9	NVIDIA Corporation	NVIDIA	Selene DGX A100 SuperPOD, AMD 64C 2.25GHz, NVIDIA A100, Mellanox HDR	USA	555,520	63.5	2.7
10	National University of Defense Technology	NUDT	Tianhe-2A ANUDT TH-IVB-FEP, Xeon 12C 2.2GHz, Matrix-2000	China	4,981,760	61.4	18.5

Green500

- Ranking by performance per watt (GFLOPS/W) from TOP500 list.
- Power supply becomes a bottleneck for large scale supercomputer, and the index is recently focused.
- update the list half-yearly same as TOP500 list http://www.green500.org/
- The value of TOP500 is used for the performance index, there is same problem as TOP500.
- Entry of Green500 should be in TOP500, but the amount of power supply is very different from 10MW to 30kW. In general, small system is better in the index of performance per watt. So large scale production level system got the special award, or small system not in TOP500 also listed as little Green500.

Green500 on Nov. 2022 (www.green500.org)



The GREEN 500

MOST ENERGY EFFICIENT ARCHITECTURES

TOP 500

Computer		Interconnect	Accelerator	Rmax/ Power
Henri, Lenovo ThinkSystem SR670 V2	Intel Xeon 32C 2.8GHz	Infiniband HDR	NVIDIA H100	*65.1
Frontier TDS, HPE Cray EX235a	AMD EPYC 64C 2.0GHz	Slingshot-11	AMD Instinct MI250X	*62.7
Adastra, HPE Cray EX235a	AMD EPYC 64C 2.0GHz	Slingshot-11	AMD Instinct MI250X	*58.0
Setonix-GPU, HPE Cray EX235a	AMD EPYC 64C 2.0GHz	Slingshot-11	AMD Instinct MI250X	57.0
Dardel-GPU, HPE Cray EX235a	AMD EPYC 64C 2.0GHz	Slingshot-11	AMD Instinct MI250X	56.5
Frontier, HPE Cray EX235a	AMD EPYC 64C 2.0GHz	Slingshot-11	AMD Instinct MI250X	52.2
LUMI, HPE Cray EX235a	AMD EPYC 64C 2.0GHz	Slingshot-11	AMD Instinct MI250X	51.4
Atos THX.A.B, BullSequana XH2000	Xeon 32C 2.4GHz	NVIDIA HDR100	NVIDIA A100	*41.4
MN-3, Preferred Network MN-Core Server	Xeon 24C 2.4GHz	RoCEv2/MN-Core DirectConnect	MN-Core	40.9
Champollion, Apollo 6500	AMD EPYC 64C 2.45GHz	Mellanox HDR	NVIDIA A100	38.6

[Gflops/Watt]

Frontier (ORNL)

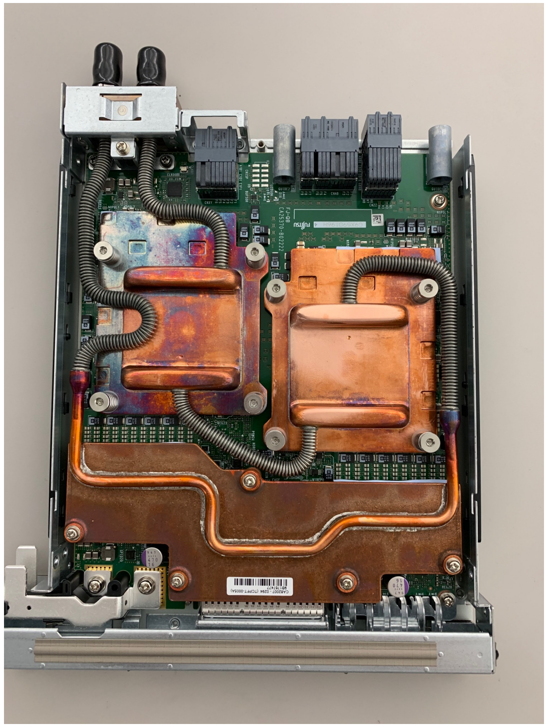
https://www.olcf.ornl.gov/frontier/

TOP500 ranking #1 as of November 2022

- DOE/SC/Oak Ridge National Laboratory
- HPE Cray EX235a with AMD EPYC CPU and AMD Instinct MI250X
- HPE Slingshot-11 interconnect

- 64 cores/CPU
- 9,408 CPU, 37,632 GPU
- Peak 1.7 EFLOPS
  HPL 1.1EFLOPS

Fugaku



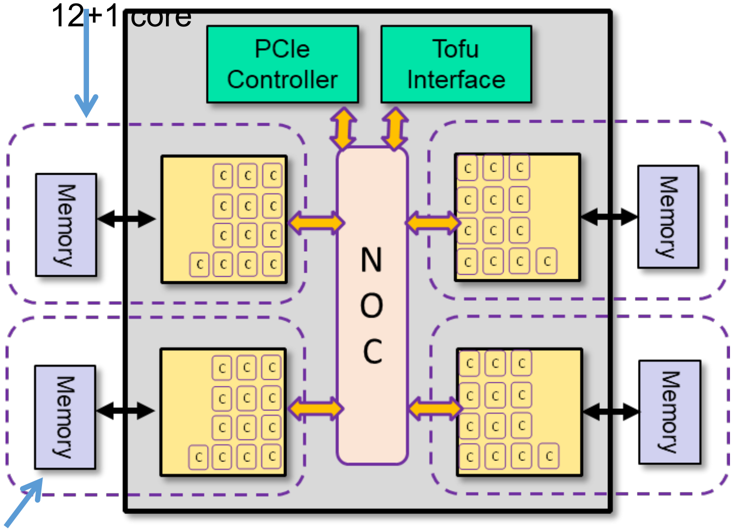
- Linpack 442PFLOPS (Peak 520PFLOPS), #1 in TOP500 on 2020/06, 2020/11, 2021/06, 2021/11
- CPU: Fujitsu original A64FX, 48 core/chip (many-core) x 150k node
- water cooling on CPU chips and interconnection driver chips, single CPU chip/node (in the above photo, two CPU chips share a motherboard but they are separated into two nodes actually)
- Interconnection Network: TOFU-D (6-D Torus, but user can access 3-D Torus)
- Operation: RIKEN R-CCS (Riken Center for Computational Science)
- ARM-base CPU with high performance SIMD instruction set, supporting FP64, FP32 and FP16 (for AI)
- While most of top10 machines are equipped with accelerators (GPU), Fugaku is equipped with general
- purpose many-core CPU for easy programming and application porting

CPU Architecture: A64FX

- **Armv8.2-A (AArch64 only) + SVE (Scalable Vector Extension)**
  - FP64/FP32/FP16 (https://developer.arm.com/products/architecture/a-profile/docs)
- **SVE 512-bit wide SIMD**
- **# of Cores: 48 + (2/4 for OS)**
- Co-design with application developers and high memory bandwidth utilizing on-package stacked memory: **HBM2(32GiB)**
- Leading-edge Si-technology (7nm FinFET), low power logic design **(approx. 15 GF/W (dgemm))**, and power-controlling knobs
- PCIe Gen3 16 lanes
- Peak performance
  - > 2.7 TFLOPS (>90% @ dgemm)
  - Memory B/W 1024GB/s (>80% stream)
  - Byte per Flops: approx. 0.4

◆ "Common" programing model will be to run each MPI process on a NUMA node (CMG) with OpenMP-MPI hybrid programming.
◆ 48 threads OpenMP is also supported.

CMG(Core-Memory-Group): NUMA node



HBM2: 8GiB

TaihuLight（太湖之光, WXSC）

- National Supercomputer Center in Wuxi
- Sunway SW26010 CPU (original)
- InfiniBand FDR
- TOP500#1 2016/6-
- (64 thin core + 1 thick core) * 4 / CPU
- 40960 CPU (10649600 cores)

- Peak 125PFLOPS
HPL 93PFLOPS
- Awarded as 2016 ACM Gordon Bell Prize machine (climate code)

CPU (SW26010) of TaihuLight

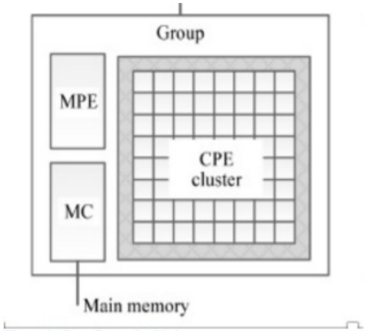


Figure 1: Core Group for Node

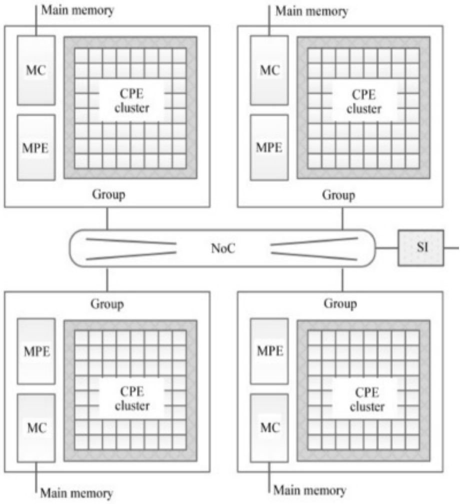
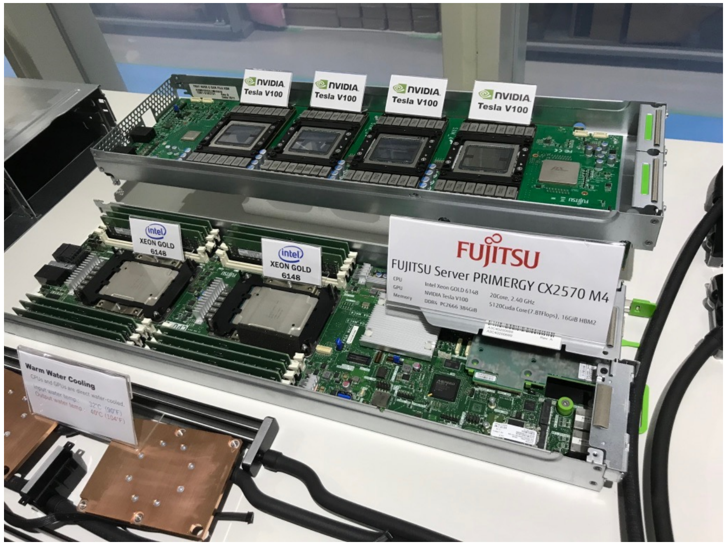


Figure 2: Basic Layout of a Node

- Highly FLOPS-intensive architecture
- 3TFLOPS/chip with 256 thin cores + 4 thick cores
- Each core has very small amount of local memory
- Medium class main memory is shared by 260 cores
- Unbalanced B/F (very weak for memory-intensive applications)
- Interconnection with InfiniBand FDR (7GB/s) is also poor compared with 3TFLOPS CPU performance
- Difficult to tune the performance, but received 2016 ACM Gordon Bell Award

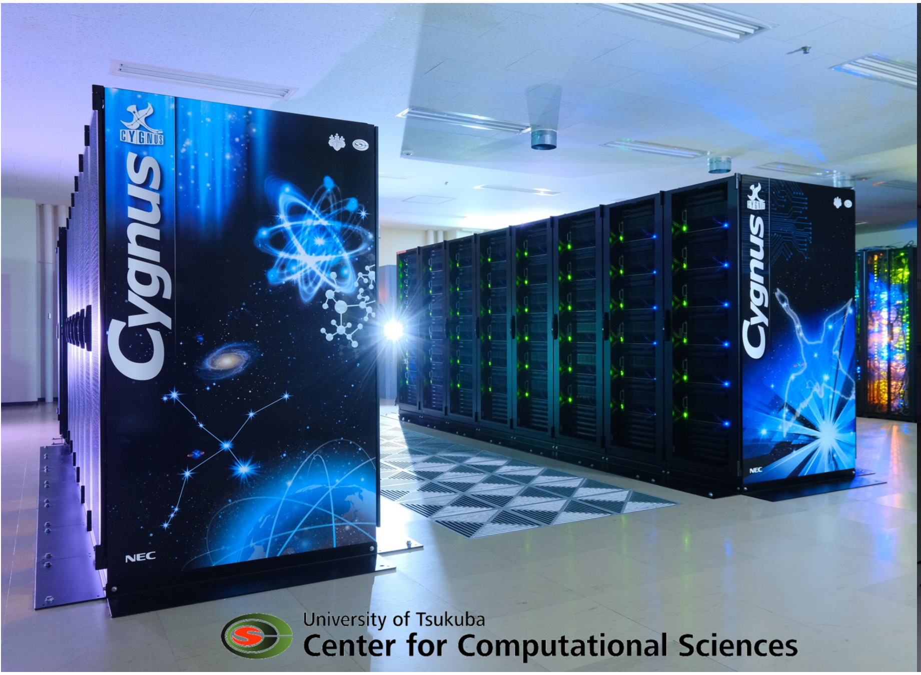
ABCI (AIST)

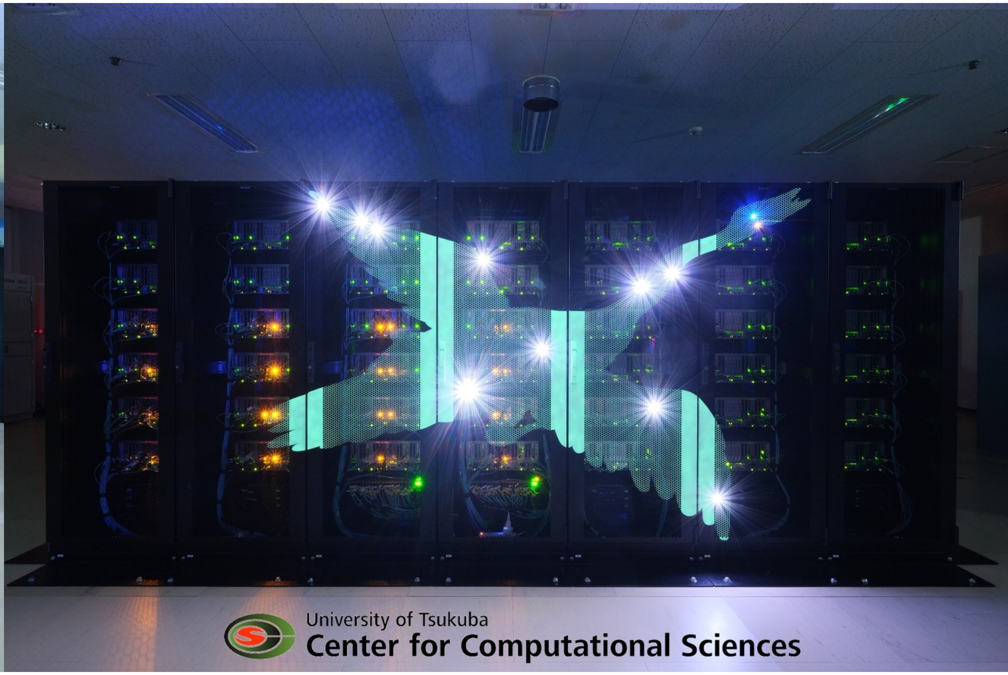


- AIST（産総研）
- Fujitsu PRIMERGY CX2570M4
- Top500 2018/06 #5
  19.9PFLOPS (efficiency 61%),
  43520 CPU core + 4352 GPU,
  3MW, 6.6 GFLOPS/W

- InfiniBand EDRx2 (non-full fat tree)
- CPU Intel Xeon Gold 6148, 2.4GHz
- GPU NVIDIA Tesla V100
- Half precision（16bit）:
  0.5EFLOPS（for Deep Learning）

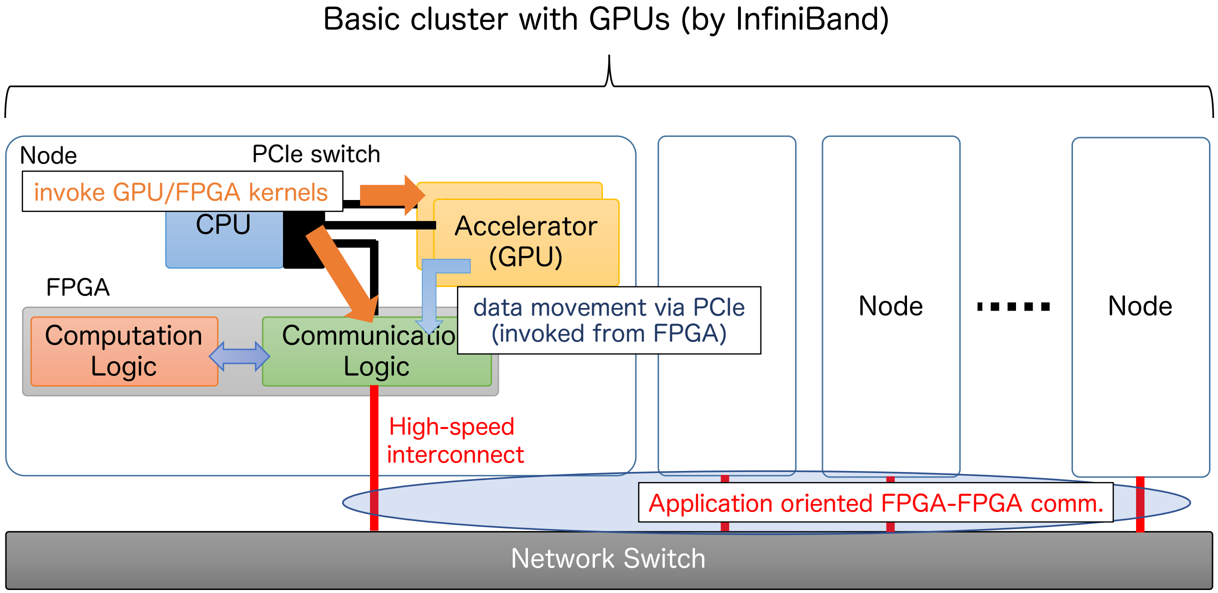
Cygnus (Univ. of Tsukuba)





CHARM: Cooperative Heterogeneous Acceleration with Reconfigurable Multidevices

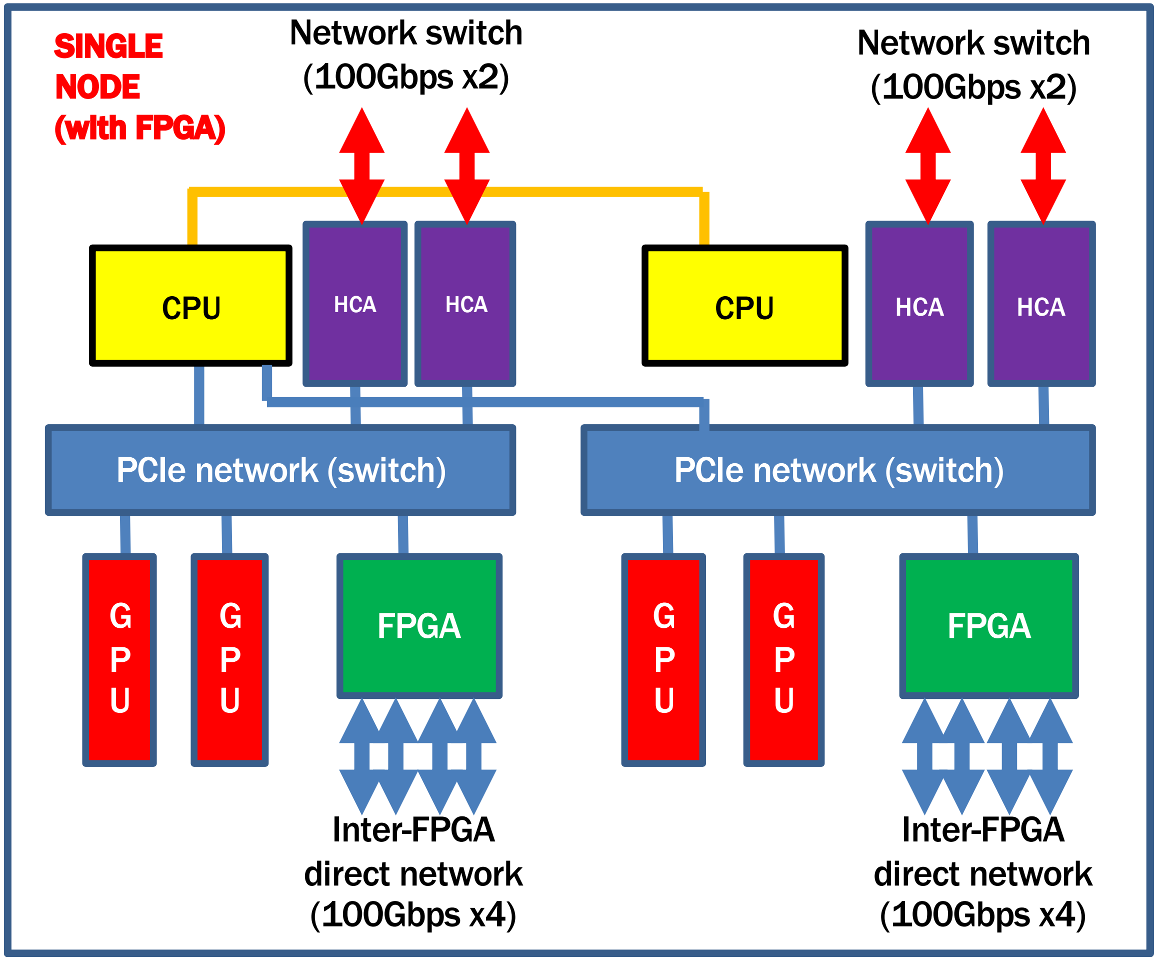
- FPGA can work both for computation and communication in unified manner
- CPU / GPU can request application-specific communication to FPGA



Single node configuration (Albireo)

- Each node is equipped with both IB EDR and FPGA-direct network
- Some nodes are equipped with both FPGAs and GPUs, and other nodes are with GPUs only

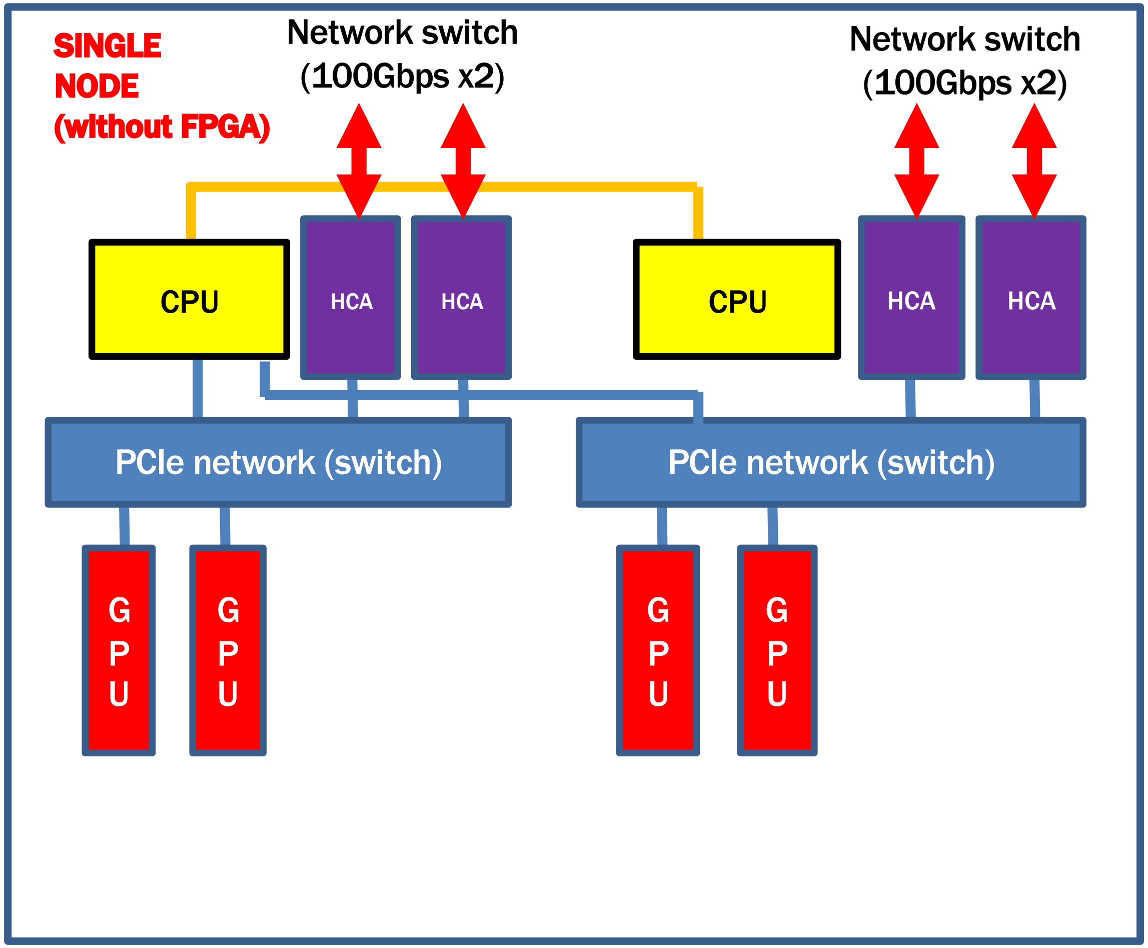
Albireo node


SINGLE NODE (with FPGA)
Network switch (100Gbps x2)
Network switch (100Gbps x2)
CPU
HCA
HCA
CPU
HCA
HCA
PCIe network (switch)
PCIe network (switch)
G P U
G P U
FPGA
G P U
G P U
FPGA
Inter-FPGA direct network (100Gbps x4)
Inter-FPGA direct network (100Gbps x4)

Single node configuration (Deneb)

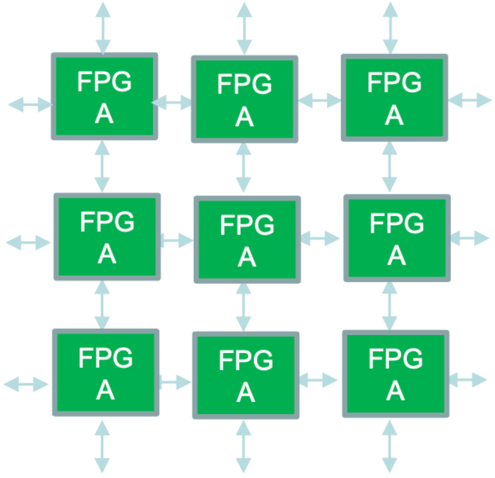
- Each node is equipped with both IB EDR and FPGA-direct network
- Some nodes are equipped with both FPGAs and GPUs, and other nodes are with GPUs only

Deneb node

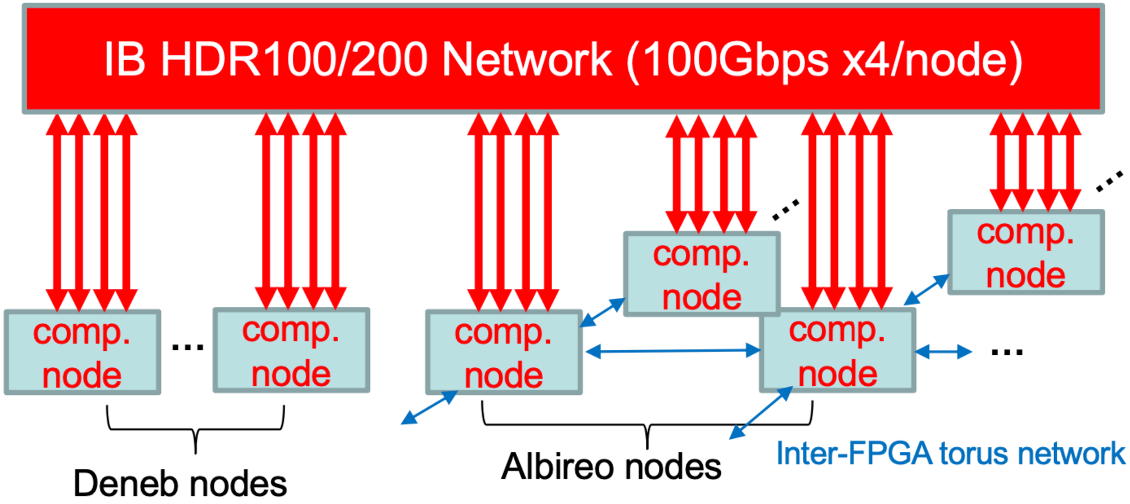

SINGLE NODE (without FPGA)
Network switch (100Gbps x2)
Network switch (100Gbps x2)
CPU
HCA
HCA
CPU
HCA
HCA
PCIe network (switch)
PCIe network (switch)
GPU
GPU
GPU
GPU

Two types of interconnection network

Inter-FPGA direct network
(only for Albireo nodes)



InfiniBand HDR100/200 network for parallel processing communication and shared file system access from all nodes

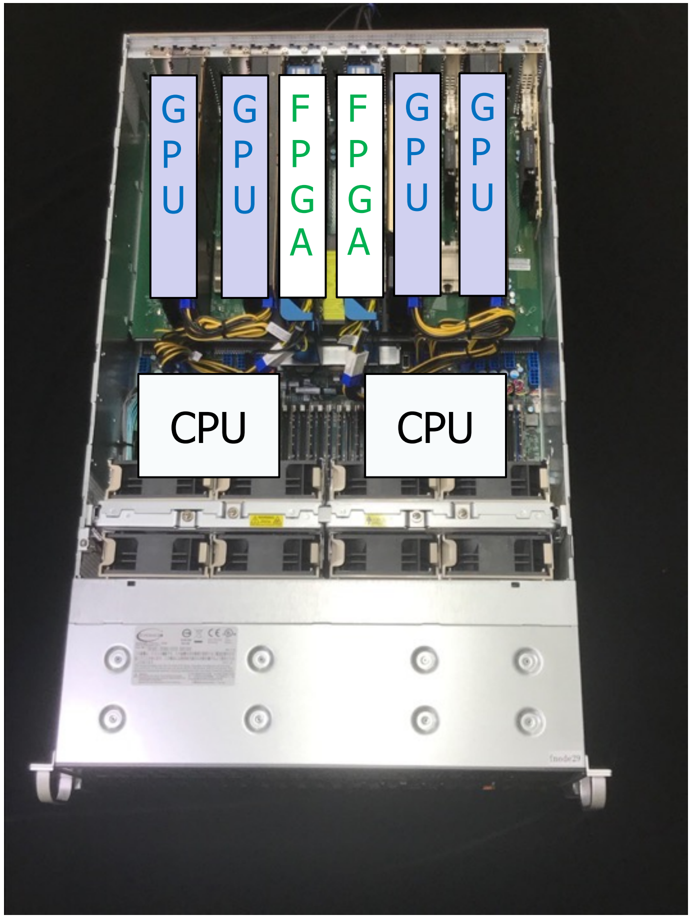


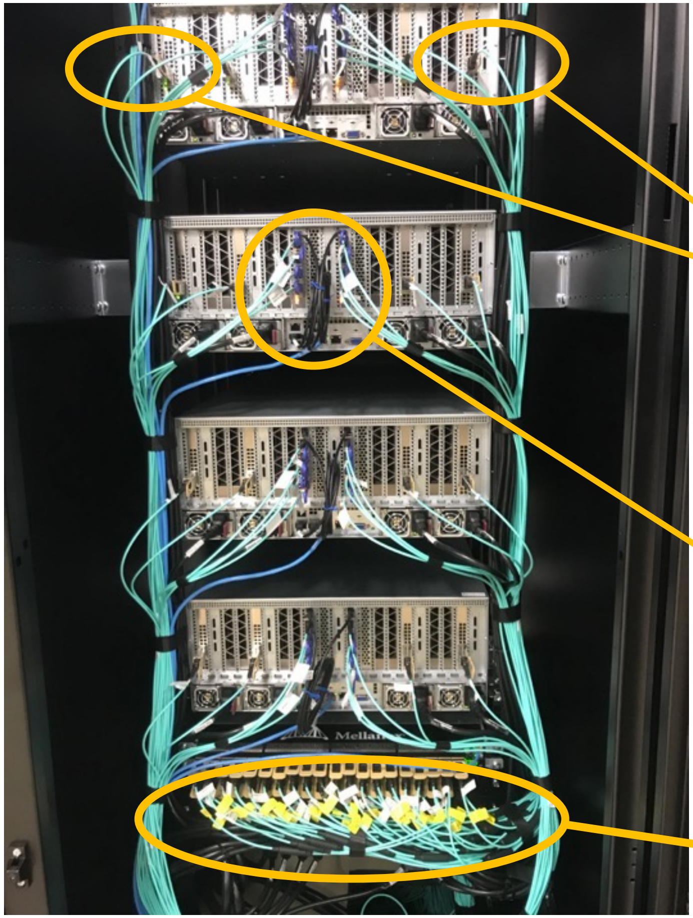
64 of FPGAs on Albireo nodes (2 FPGAS/node) are connected by 8x8 2D torus network without switch

For all computation nodes (Albireo and Deneb) are connected by full-bisection Fat Tree network with 4 channels of InfiniBand HDR100 (combined to HDR200 switch) for parallel processing communication such as MPI, and also used to access to Lustre shared file system.

Specification of Cygnus

Item	Specification
Peak performance	2.43 PFLOPS DP (GPU: 2.27 PFLOPS, CPU: 0.16 PFLOPS, FPGA: 0.6 PFLOPS SP) ⇨ enhanced by mixed precision and variable precision on FPGA
# of nodes	81 (32 Albireo (GPU+FPGA) nodes, 49 Deneb (GPU-only) nodes)
Memory	192 GiB DDR4-2666/node = 256GB/s, 32GiB x 4 for GPU/node = 3.6TB/s
CPU / node	Intel Xeon Gold (SKL) x2 sockets
GPU / node	NVIDIA V100 x4 (PCIe)
FPGA / node	Intel Stratix10 x2 (each with 100Gbps x4 links/FPGA and x8 links/node)
Global File System	Lustre, RAID6, 2.5 PB
Interconnection Network	Mellanox InfiniBand HDR100 x4 (two cables of HDR200 / node) 4 TB/s aggregated bandwidth
Programming Language	CPU: C, C++, Fortran, OpenMP, GPU: OpenACC, CUDA FPGA: OpenCL, Verilog HDL
System Vendor	NEC





IB HDR100 x4
⇨ HDR200 x2

100Gbps x4
FPGA optical
network

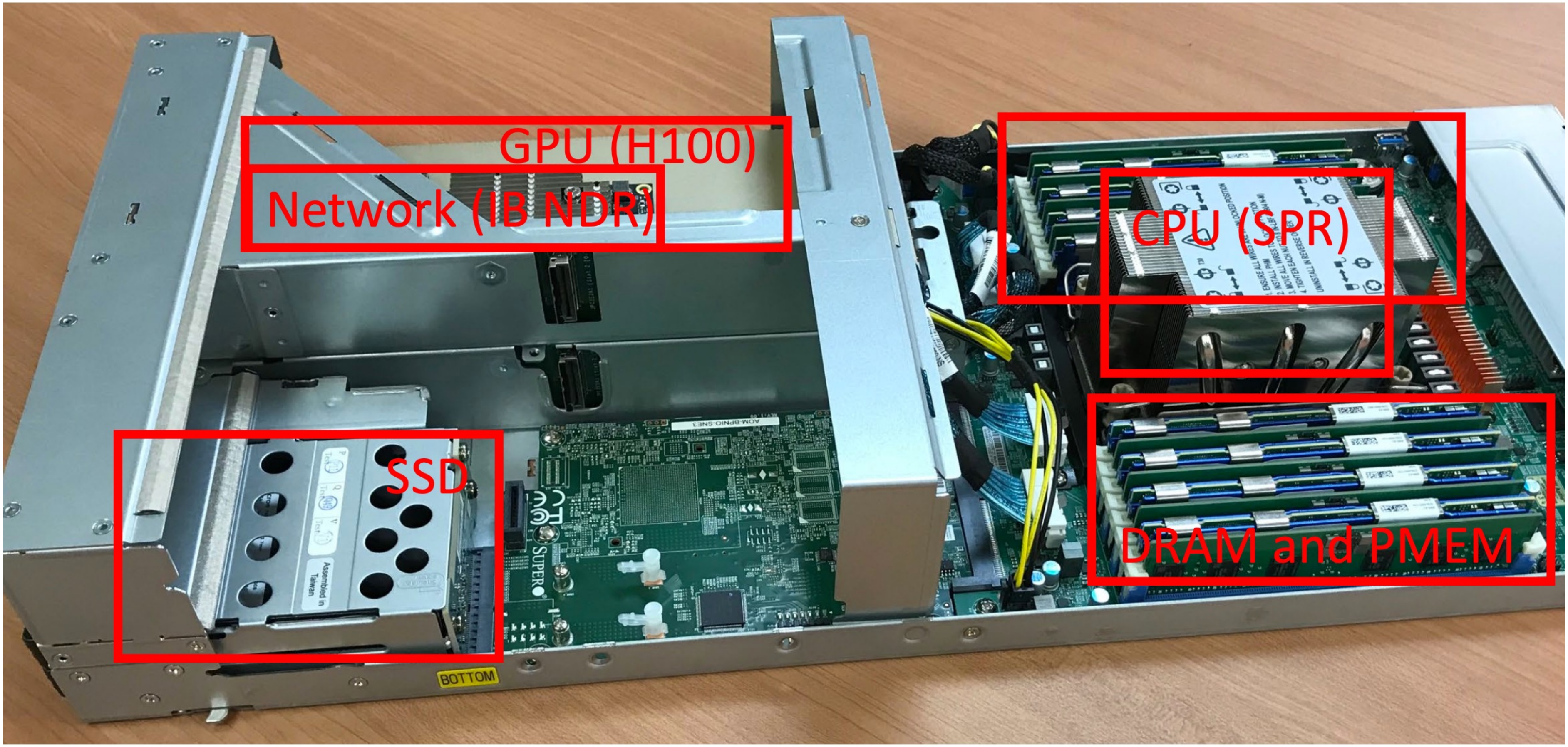
IB HDR200
switch (for
full-bisection
Fat-Tree)

Pegasus (Univ. of Tsukuba)

- CCS's brand new supercomputer
- Official operation start: 2023/04

Item	Specification
Peak performance	6.5 PFLOPS DP
# of nodes	120
Total Memory size	255 TiByte (15 TiByte DDR5 + 240 TiByte Persistent Memory)
Parallel file System	7.1PB DDN EXAScaler (40 GB/s)
Interconnection Network	Full bisection fat-tree network interconnected by the NVIDIA Quantum-2 InfiniBand platform
System Vendor	NEC

Compute Node (Blade)

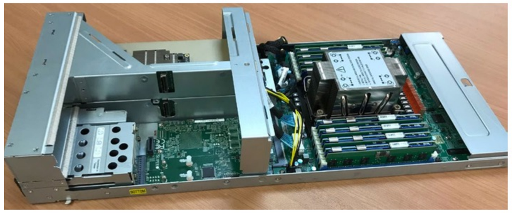


Node Specification

- Compute nodes (NEC LX 102Bk-6) x 120

CPU	Intel Xeon Platinum 8468 (codenamed Sapphire Rapids) 2.1GHz/48c (3.2256 TFlops)
GPU	NVIDIA H100 Tensor Core GPU with PCIe (51 TFlops in FP64 Tensor Core)
Memory	128GiB DDR5-4800 (282 GB/s)
Persistent memory	2TiB Intel Optane persistent memory 300 series
SSD	2 x 3.2TB NVMe SSD (7 GB/s)
Networking	NVIDIA Quantum-2 InfiniBand platform (200 Gbps)

NEC LX B1000E Blade Enclosure



NEC LX 102Bk-6

- Login nodes (NEC LX 124Rk-2) x 3
  - 2 x Intel Xeon Platinum 8468 (2.1GHz/48c)
  - 256 GiB DDR5 Memory, NVMe SSD, InfiniBand x 2, 100GbE

NEC LX 124Rk-2

For more information about Pegasus
→ https://www.ccs.tsukuba.ac.jp/eng/supercomputers/

Trend of parallel system

- Commodity based cluster increase
  - Commodity scalar processor (IA32=x86)
  - Commodity network I/F and switch
    - Ethernet (1Gbps ⇒ 10Gbps)
    - Infiniband (2GB/s ⇒ 8GB/s, the price gradually reduced)
- The balance between processor, memory and communication performance becomes worse.
  - Arithmetic performance increases smoothly with multi-core processor
  - Memory performance (bandwidth) will not increase, and relatively reduced for a core (pin-bottleneck, 3D or wide-I/O memory ?)
  - Communication performance will increase step-wise, but relatively reduce for a core (Ethernet, etc)
  - Processor cost is *O(N),* but network cost is *O(N log N),* so the network cost relatively large
  - It becomes difficult to improve the parallel efficiency in large scale parallel system. Algorithm level improvements are required.
- Cluster with accelerator will increase
  - high performance per cost, performance per watt, …
- Exa FLOPS era has begun

Summary

- Parallel system / architecture
  - The performance of sequential processor (core) has been limited, so total performance will be increased by parallel processing.
  - Scalability with keeping performance is important
  - Distributed memory vs. shared memory
- Interconnection network
  - Scalability is most important
  - MPP had wide variety of implementations (custom network)
  - Current cluster network has scalability with fat-tree topology using commodity network.
  - Two performance metrics: throughput & latency
- Trend and problems for parallel computer
  - The number of core will be increased, to 1 million cores with multicore processor.
  - The balance between processor, memory, network performance will be worse.
  - Node with accelerator is attracted (GPU is the most popular)

Report

Answer the following questions referring TOP500 List on Nov. 2022, about #1, #2 and #3 systems.

1. Explain the peak performance of these systems referring the performance (FLOPS) for each core, each chip, each computation node and entire system. Make the table of these values. For the performance per core, show the evidence as much as possible; ex) processor frequency, floating point operation count per clock, etc. Also show the number of cores per chip and computation node. If a node is heterogeneous (or with some accelerators), explain both for host CPU and accelerators.
2. Show the system power consumption, Linpack execution efficiency (Linpack performance compared with the theoretical peak performance) and performance per watt on Linpack execution. Summarize the trend of power consumption for performance on these systems.
3. Describe the characteristics of these systems on their system architecture such as memory system, interconnection network, operating system, etc.

Show the evidence of documents or URLs if you refer any material in the Internet.