筑波大学計算科学研究センター
CCS HPCセミナー
「並列システム」

小林　諒平
kobayashi@cs.tsukuba.ac.jp
筑波大学大学院システム情報工学研究科
計算科学研究センター

「並列システム」内容

- 並列計算機アーキテクチャ
  - 分散メモリ, 共有メモリ, SMP, NUMA...
- 並列処理ネットワーク
- 実システムの紹介

並列処理システム

- 並列計算機 (並列処理システム) も計算機であり，プロセッサ (CPU)，メモリ (memory)，入出力装置 (I/O) 等の構成要素を持つ点はPC, スマホ, ゲーム機と同じ
- 複数のプロセッサ間を結合するハードウェアの仕組み
- 並列にプログラムを実行するソフトウェアの仕組み
- システム規模は2プロセッサから1000万超プロセッサまで
  - 2~8プロセッサ：現在ではsingle chipで実現（multi-core CPU）
  - 数十プロセッサ：研究室レベルのクラスタ，共有メモリシステム
  - 数百プロセッサ：センター運用クラスタ，小型MPP (Massively Parallel Processor)
  - 数千プロセッサ~：MPP
    (multi-core CPUの台頭によりクラスタでも数万プロセッサが実現可能に)
- スパコン富岳は 76,000,000 CPU cores のシステム

HPCは物量 (= 数の暴力) でゴリ押しがキモなのですね！

並列処理システムの要素

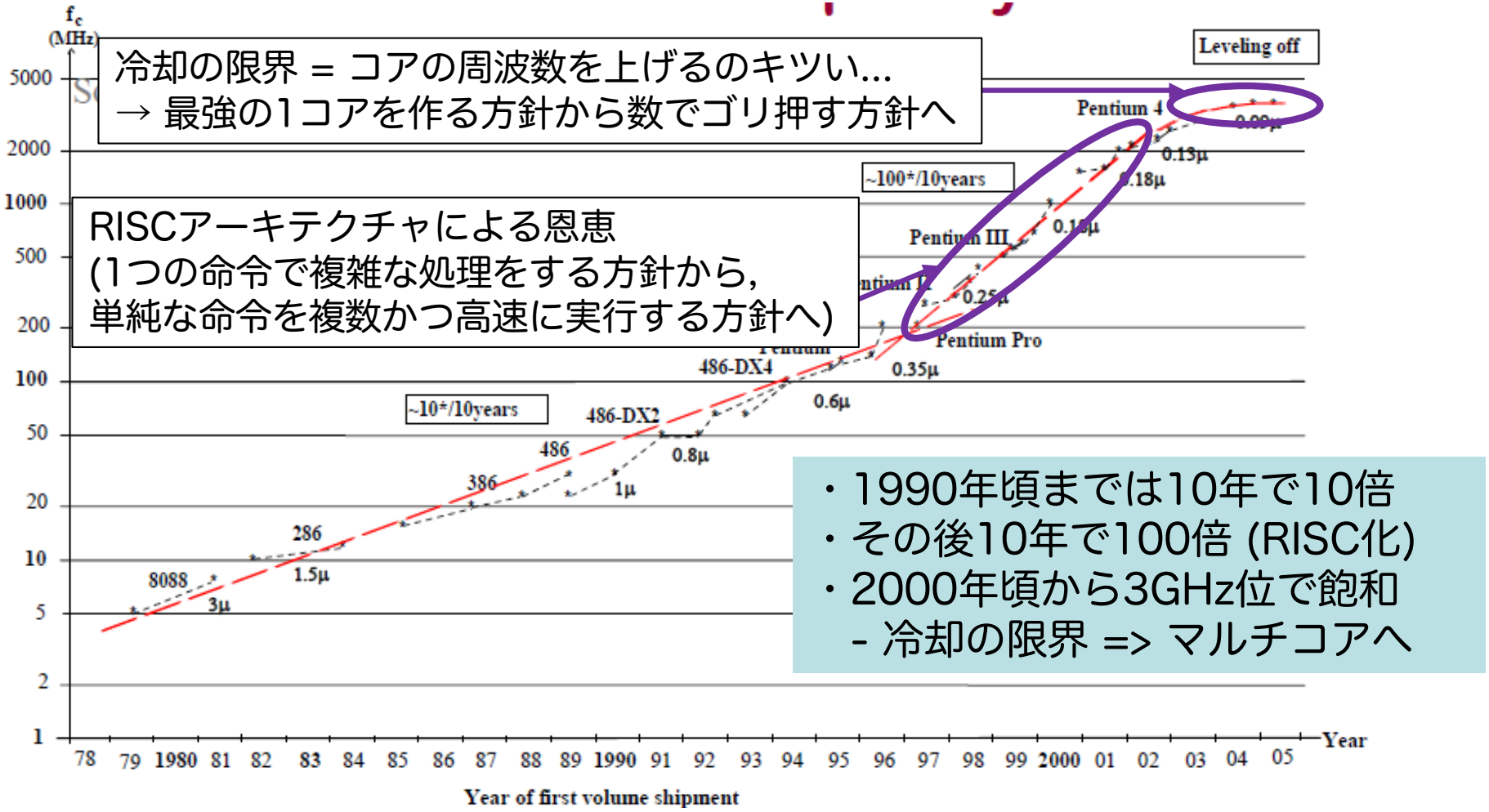
- 一般的なPCと異なる部分
  - 何らかのプロセッサ間結合ネットワーク（相互結合網：Interconnection Network）
  - プロセッサ間通信の結果のデータをメモリに保持する機構
    - 分散メモリ型マシン：相互結合網からのデータ（メッセージ）内容をメモリに保持
    - 共有メモリ型マシン：メモリ自体が逐次システムと異なり、並列プロセッサ間で共有される
      ⇒特殊なハードウェアが必要
  - 特別な同期機構を持つ場合もある
  - その他の周辺装置
    - システム全体を１つにまとめる管理機構
    - 並列プロセッサから共有可能なファイルシステム

計算機の進歩
~プロセッサの観点から~

- 個々のプロセッサ
  - ベクトルプロセッサ　⇒　20年前のスーパーコンピュータではこのタイプが多かった
    - 一つのプロセッサで行列演算等を高速処理可能
  - スカラープロセッサ：
    x86 (IA32), Power, Arm

- 近年におけるプロセッサの動向
  - マルチコアが標準になってきた
    - Intel や AMD のプロセッサ ⇒ 8 ~ 16 コア (x86)
  - メニーコア (8 ~ 512コア) も登場！
    - IBM Cell Broadband Engine (8 コア)
    - ClearSpeed (96 core×2)
    - GPU (NVIDIA K20X 896 DP unit)
    - Intel Xeon Phi (72 core)
    - Fujitsu A64FX (スパコン富岳に搭載されている48コアCPU)

スカラプロセッサのクロック周波数



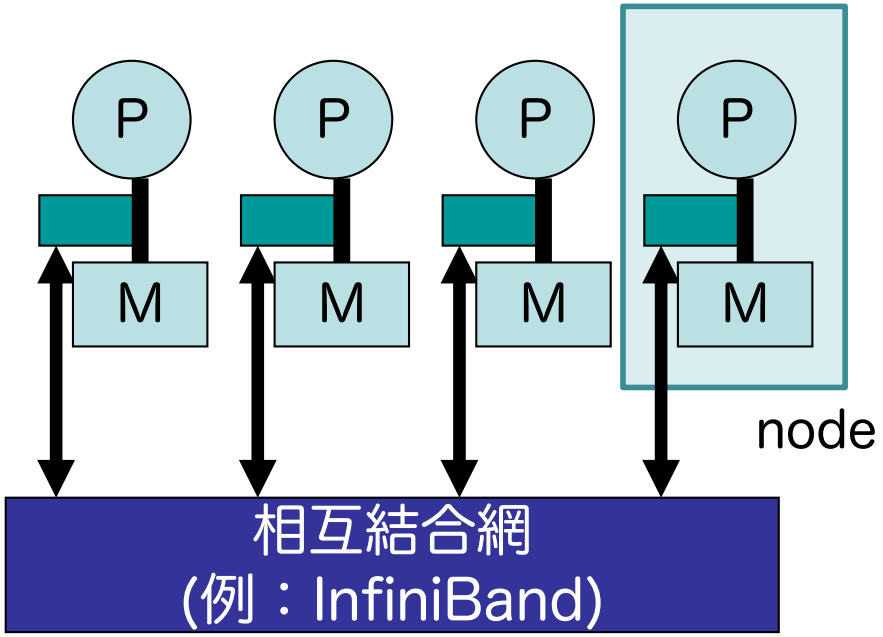
並列計算機の変遷

- 科学技術計算向けベクトルプロセッサ
  - 一種の並列計算機とみなせる（パイプライン並列）
  - ベクトルプロセッサを複数持つ並列ベクトルが登場
- スカラプロセッサをベースにした並列計算機
  - ～100プロセッサ程度の共有メモリマシン（SGI等）
  - quad-core CPUの登場により、8 core程度であればデスクトップPCでも共有メモリ並列システムとなる
  - MPP (Massively Parallel Processor)：1980年代後半から多数登場⇒一部を除き消滅しつつある (富岳はMPPシステム)
- クラスタ型計算機の登場
  - 以前は NOW (Network of Workstation), COW (Cluster of Workstation) 等と呼ばれていた
  - Linux PC を用いたものが現在の主流（Linuxがオープンシステムであるため、MPI等の並列化ツールも充実）
  - 数千プロセッサ規模のものが多数構築されている

並列計算機アーキテクチャ

- 分散メモリ型システム
  - 各プロセッサは独自のメモリを持ち，ネットワークを用いたメッセージパッシングによってデータ交換を行う
  - 各プロセッサのメモリは他のプロセッサから直接アクセス不可
- 共有メモリ型システム
  - 並列プロセッサ間で物理的に共有される共有メモリを持ち，各プロセッサが普通のロード/ストア命令を発効してデータの読み書きを行う
  - アーキテクチャのタイプとしてSMPとNUMAがある
- ハイブリッド型システム
  - 共有メモリシステムを分散メモリ型に統合したもの
  - 最近のマルチコアCPUの影響からこれが普通になってきた
  - ノード内にアクセラレータを接続したヘテロジニアス構成をとる計算機も増えてきた

分散メモリ型並列計算機



任意のプロセッサ間で
メッセージを送受信

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

●CPUとメモリという一つの計算機システム（ノード）が、ネットワークで結合されているシステム

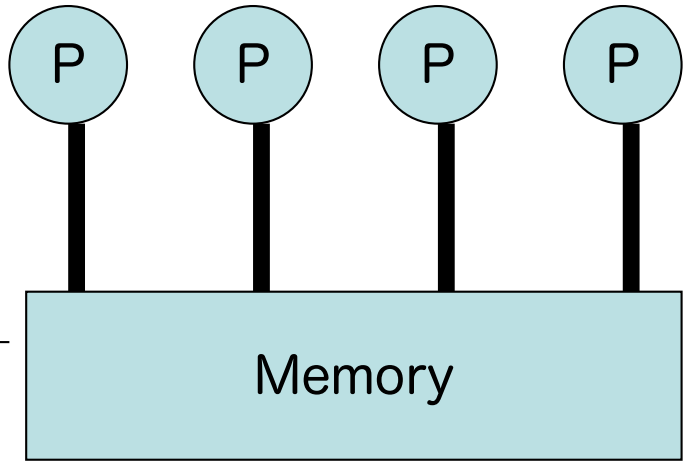
●それぞれの計算機で実行されているプログラムはネットワークを通じて、データ（メッセージ）を交換し、動作する

●比較的簡単に構築可能・拡張性(scalability) が高い
- ◆超並列計算機 (MPP：Massively Parallel Processing)
- ◆クラスタ型計算機

分散メモリ型並列計算機の特徴

- 基本的にCPU+memory(+I/O)という逐次計算機構成を何らかのネットワーク（専用 or 汎用）で結合しているため、ハードウェア的にシンプル
- プログラム上からの明示的なmessage passingで通信を行うためユーザプログラミングは面倒
  - MPI (Message Passing Interface)のような標準的なツールが提供されている
  - ソフトウェア分散共有メモリによる簡便なアプリケーション記述の試みも
  - domain decompositionのような単純なデータ並列や、master/worker型の処理は比較的容易に記述可能
- システム性能は個々のプロセッサ／メモリの他、相互結合網の性能によって大きく左右される
- 1980年代後半からMPPの典型的な実装として登場、現在はPCクラスタの基本的アーキテクチャとなっている

共有メモリ型計算機



複数のプロセッサからの
同時アクセスを整理する
ことが必要

●複数のCPUが一つのメモリにアクセスするシステム

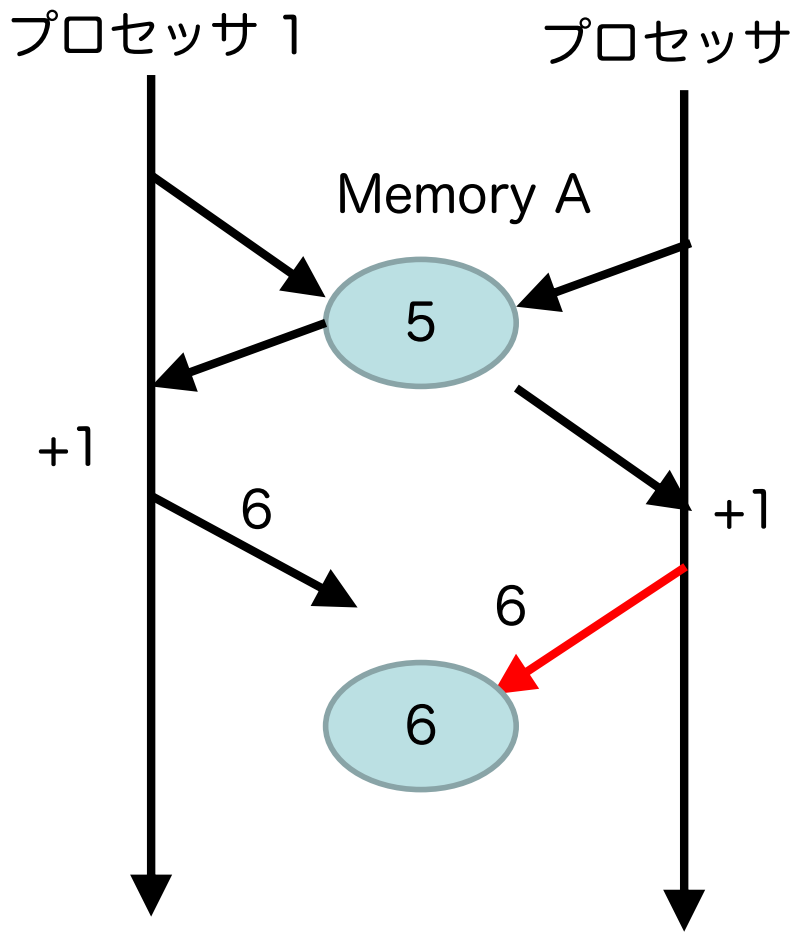
●それぞれのCPUで実行されているプログラム（スレッド）は、メモリ上のデータにお互いにアクセスすることで、データを交換し、動作する

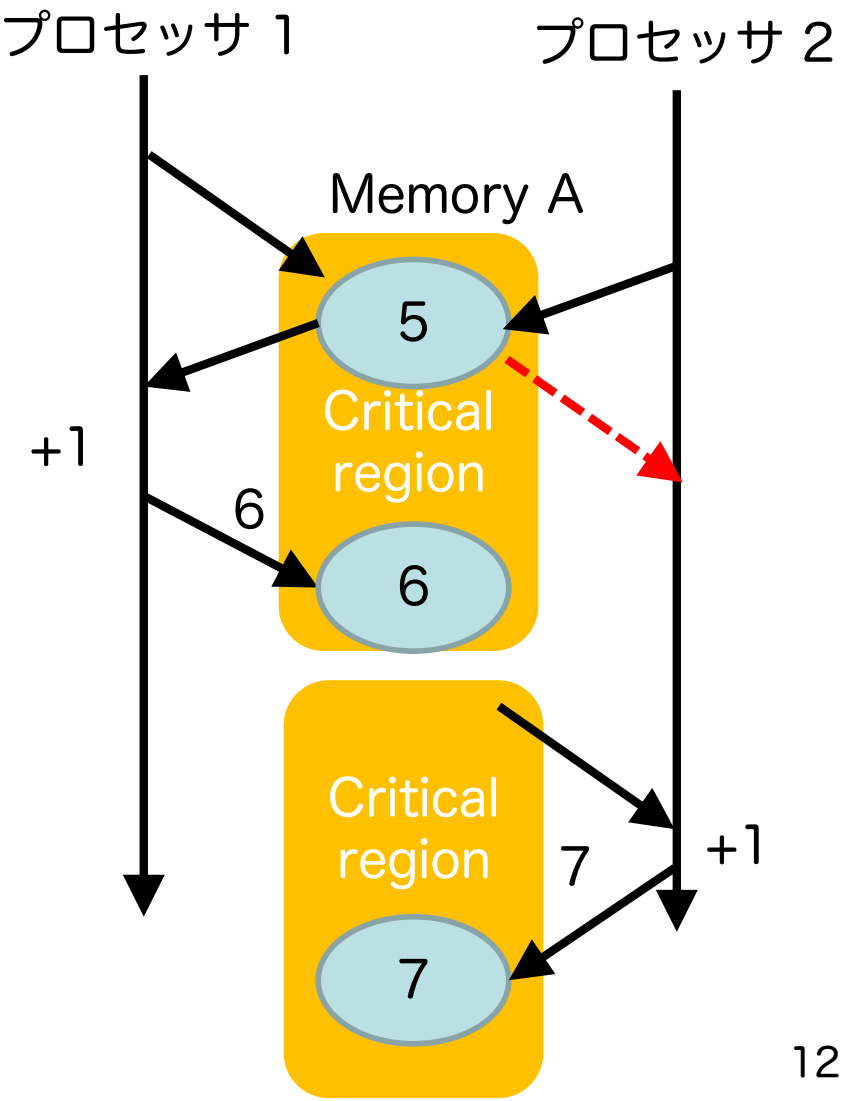
●複数CPUソケットのマザーボードを格納出来る大規模サーバ

●最近はプロセッサ１台が複数のプロセッサコアの共有メモリシステムになっている

●アーキテクチャ的にはさらにSMPとNUMAに分かれる（後述）

共有メモリにおけるアクセス競合





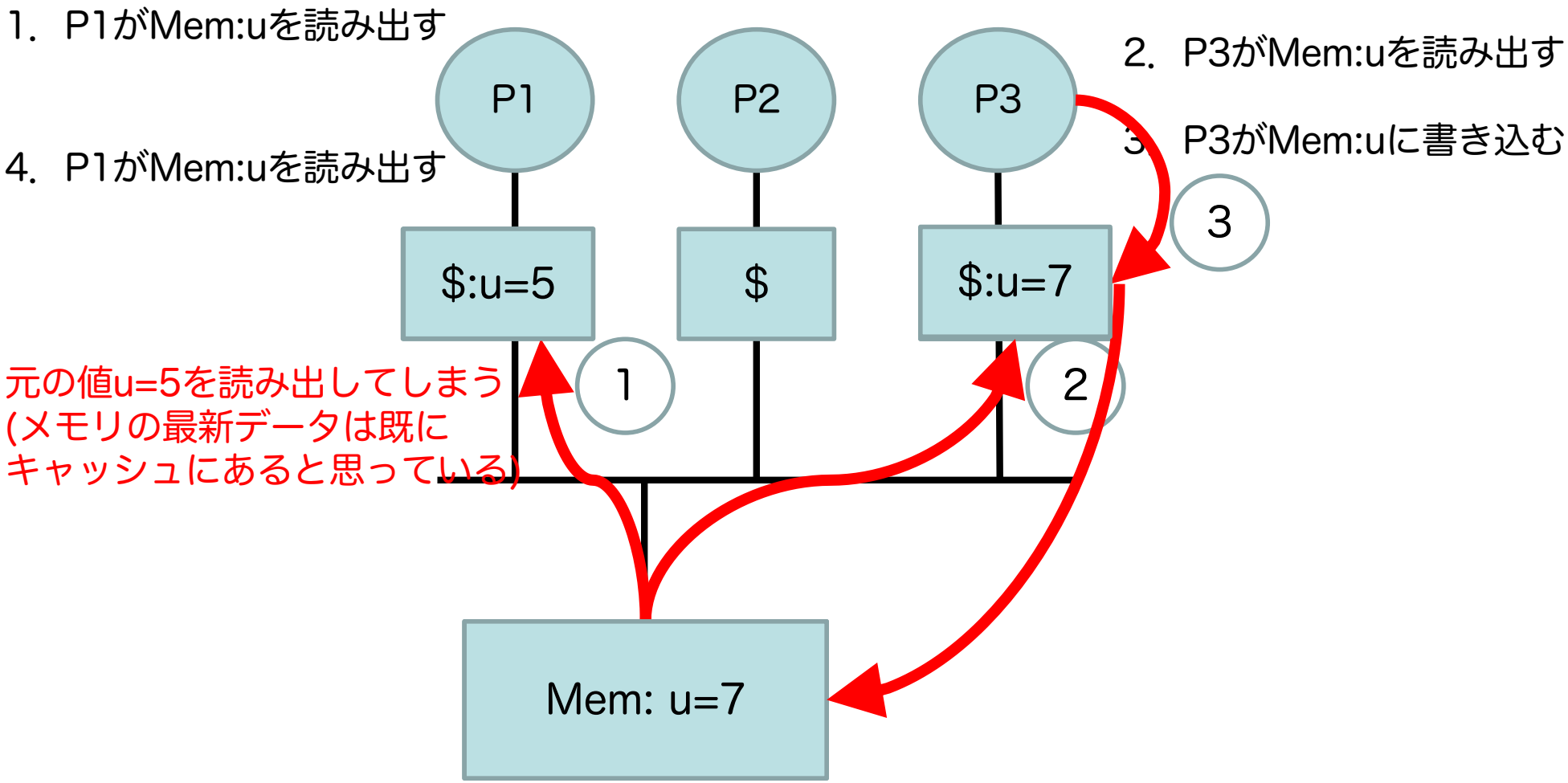
共有メモリ型並列計算機の特徴

- ハードウェアによる共有メモリの提供により、ユーザにとってアプリケーションが非常に書き易い
  - multithreadプログラミング環境（POSIX thread等）
  - 共有メモリを前提とした簡易並列記述システム（標準的なものはOpenMP）
- 「メモリ」という極めてprimitiveな構成要素を共有化しているため、性能を上げるには非常に多くのハードウェア的、アーキテクチャ的工夫が必要
- 多数のプロセッサが１つのメモリ要素をアクセスする状況が簡単に記述でき、極端な性能ボトルネックを生じ易い
  - システムのscalabilityの確保が困難（数百プロセッサが限界）
- 概念的には前頁のような共有バスのイメージだが実際にはscalabilityを確保するためより複雑になっている

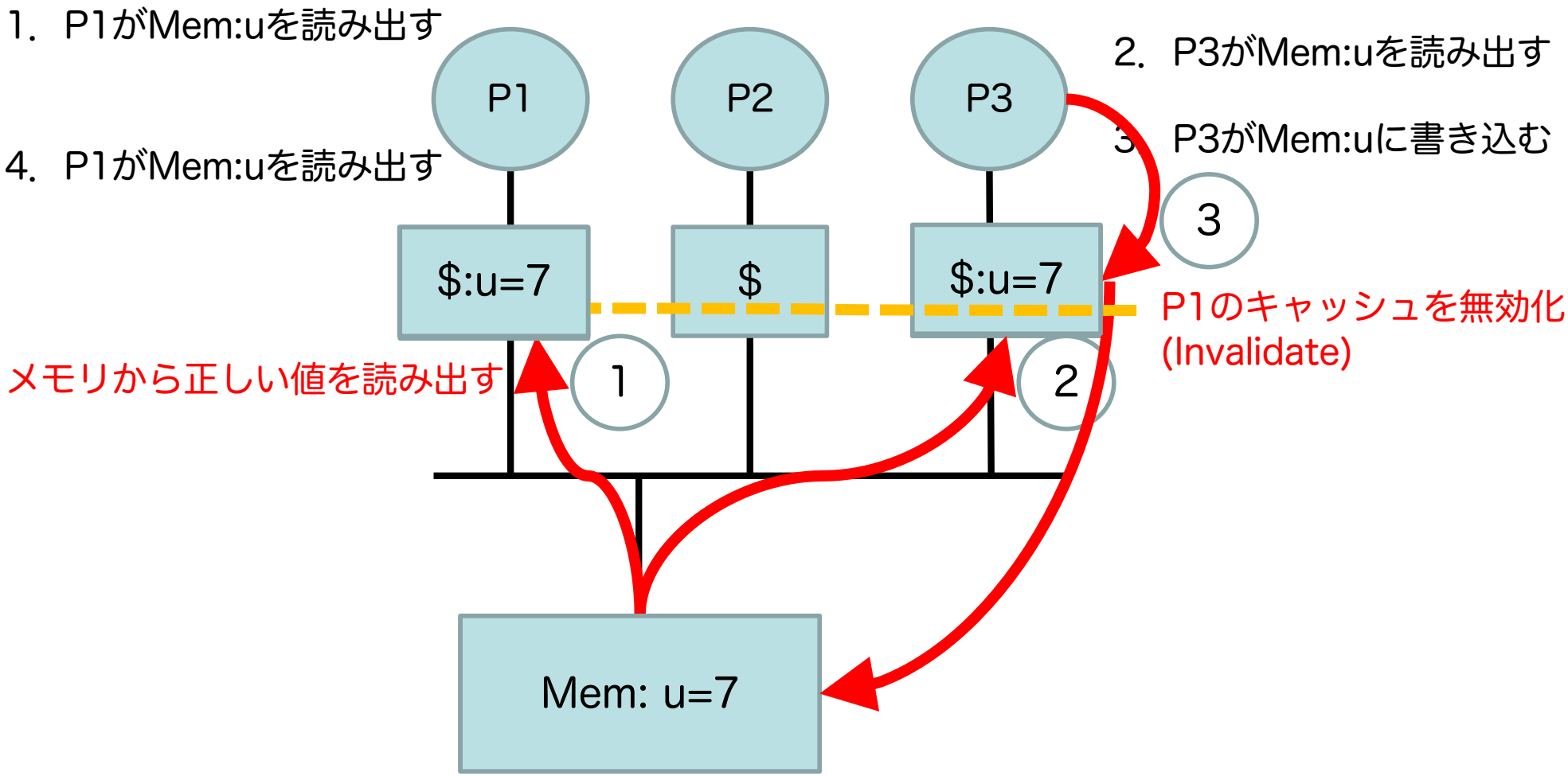
共有メモリ型計算機の構成の詳細

- system scalabilityを確保するため、単純バス構造の共有メモリシステムはもはや存在しない
  - bus bottleneck（busは一時には1つのtransactionで占有されてしまう）
  - 複数busを持つシステムもかつてはあった
- 共有メモリへのアクセス衝突を避けるための工夫
  - memory bank分け：適当なアドレスブロック毎に別のmemory moduleに分散して振り分け
  - crossbar networkの導入：プロセッサとメモリの結合が実際にはスイッチ結合になっている
  - coherent cache：各プロセッサは固有のキャッシュを持ち、普段はそのデータを参照する。他のプロセッサによるデータ更新をキャッチし、うまく自分のキャッシュに反映する。
  - NUMA (Non-Uniformed Memory Access)：物理的にはmemory moduleが分散していて、アドレスに寄るメモリへの距離の差が存在する。coherent cacheと共に用いられるのが普通

コヒーレントでないキャッシュ

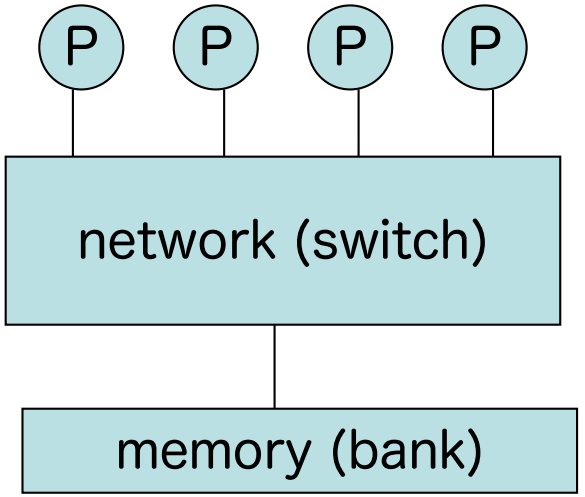


コヒーレントキャッシュ



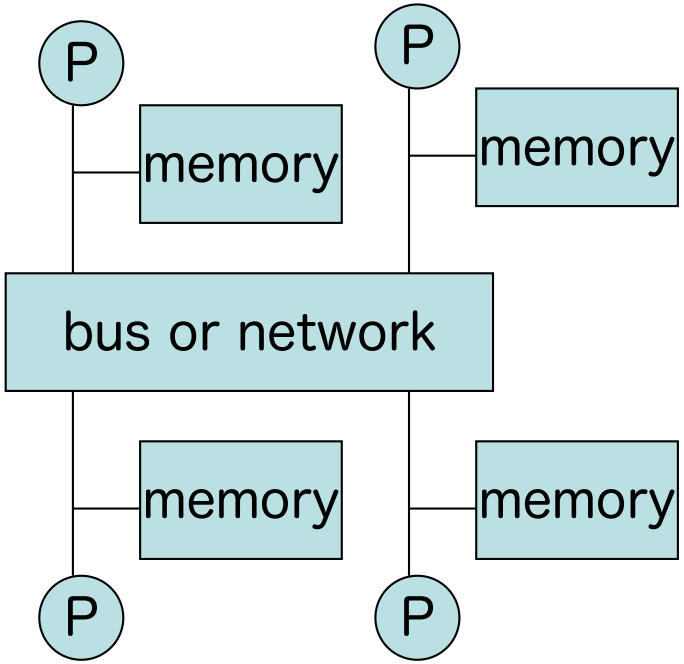
共有メモリアーキテクチャ：SMP

- SMP (Symmetric Multi-Processor)
  - 各プロセッサから見てどのmemory moduleへの距離も等しい
  - 構成としては、複数のプロセッサが共通のバスまたはスイッチを経由して、等しくmemory module（群）に接続されている
  - 大規模システムとしては富士通のHPC2500シリーズ、日立SR16000シリーズ等が該当する
  - coherent cacheとの併用が一般的
  - どのプロセッサからもデータが等距離にあるので偏りを心配しなくてよい
  - トラフィックが集中した場合に性能低下を防げない

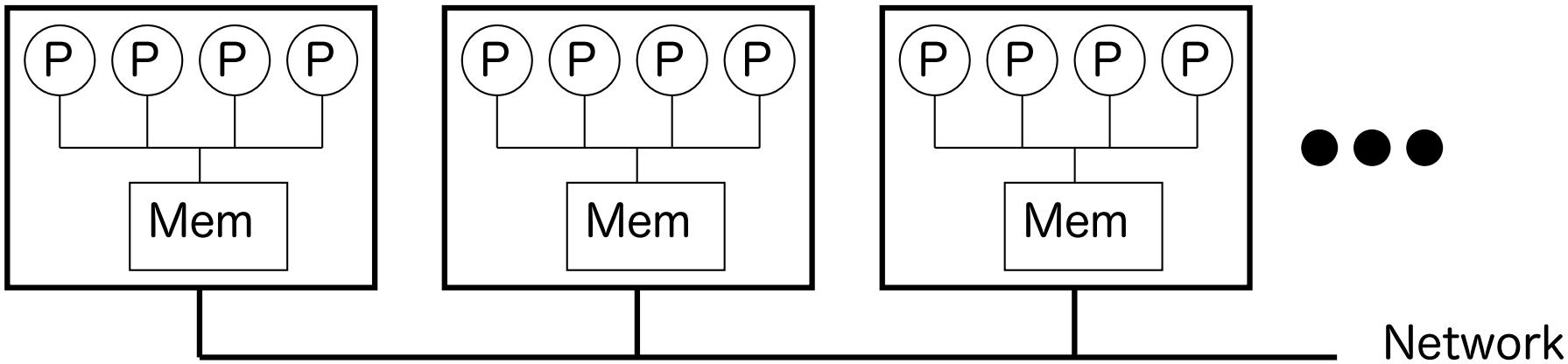


NUMA共有メモリアーキテクチャ

- NUMA (Non-Uniformed Memory Access)
  - CPUに付随して固有のmemory moduleがある
  - 共有バスまたはスイッチを介して、他のCPUのmemory moduleも直接アクセス可能
  - 遠距離memory moduleへのアクセスには時間が余計にかかる（non-symmetric）
  - コモディティスカラプロセッサとしてはAMD (Opteron)が最初に方式を取り入れた
    ⇒ 今では標準になっている
  - 大規模システムとしてはSGI Origin, Altixシリーズ等が該当
    ⇒ もう存在しない（NUMAはCPUボード上でのみ）
  - データをうまく分散し、参照の局所性が生かせれば性能を大幅に向上可能(メモリアフィニティ）
  - 遠距離アクセス時の遅延時間増加に注意



分散／共有メモリ・ハイブリッド



- 共有メモリと分散メモリの組み合わせ
- 分散メモリ型システムの各ノードがそれ自身共有メモリアーキテクチャになっている (SMP or NUMA)
- マイクロプロセッサ自体が1チップで共有メモリ構成（マルチコア）となっていることが大きな要因、近年のマルチコアプロセッサ普及により急激に主流となった

アクセラレータ付並列システム

- 分散メモリ型計算機の各ノードが汎用CPUだけでなく演算性能を加速するハードウェア（アクセラレータ）を伴う
  - GPU (Graphic Processing Unit)
    最近ではGPGPU (General Purpose GPU) と呼ばれ、GPU上で汎用プログラミングも可能に
  - FPGA (Field Programmable Gate Array)
    特殊用途向けに再構成可能なハードウェア
  - 汎用アクセラレータ
    ClearSpeed等
  - プロセッサ自体がヘテロジニアス構成
    CBE (Cell Broadband Engine) ⇒ LANL Roadrunner

マルチコア、メニーコア、GPU

	マルチコア	メニーコア	GPU
例	Intel Xeon Platinum 8280	Intel Xeon Phi 7250P	NVIDIA Tesla V100
コア数	28	72	5120
周波数	2.7 GHz	1.2GHz	1.53 GHz
性能	605 GFLOPS	1.1 TFLOPS	7.8 TFLOPS
メモリ容量	1 TB	16 GB	6 GB
メモリ性能	141 GB/s	450 GB/s	900 GB/s
電力	205 W	200 W	300 W
プログラム	C, OpenMP, MPI	C, OpenMP, MPI	CUDA　(C, Fortran)
モデル	MIMD	MIMD	STMD

「並列システム」内容

- 並列計算機アーキテクチャ
  - 分散メモリ, 共有メモリ, SMP, NUMA...
- 並列処理ネットワーク
- 実システムの紹介

並列処理ネットワーク（相互結合網）

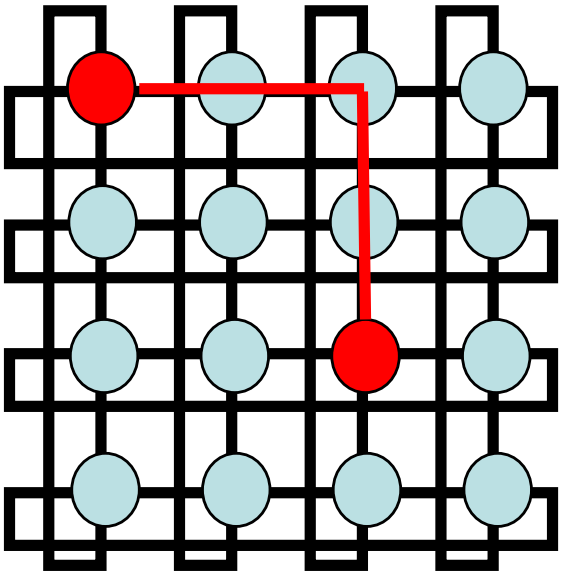
- 役割
  - 分散メモリアーキテクチャに基づく並列計算機における明示的なデータ交換 (例：MPI通信)
  - CC-NUMAアーキテクチャ (Cache Coherent NUMA)に基づく並列計算機におけるデータ及び制御メッセージの転送
- 特性／分類
  - static (direct) / dynamic (indirect)
  - diameter (distance)
  - degree (number of links)
- 性能指標
  - Throughput (単位時間あたりにどのくらいのデータ量を転送できるか)
  - Latency (転送要求を出してからどのくらいの時間でデータがやって来るか)

直接網（静的網）

- ノード（プロセッサ）に数本のリンクを持ち、それらが互いに結合してネットワークを形成
- ノード上でのルーティングが行われるがノード以外のスイッチは持たない
- 代表的な直接網トポロジ
  - 2-D/3-D Mesh/Torus
  - Hypercube
  - Direct Tree

Mesh/Torus (k-ary n-cube)

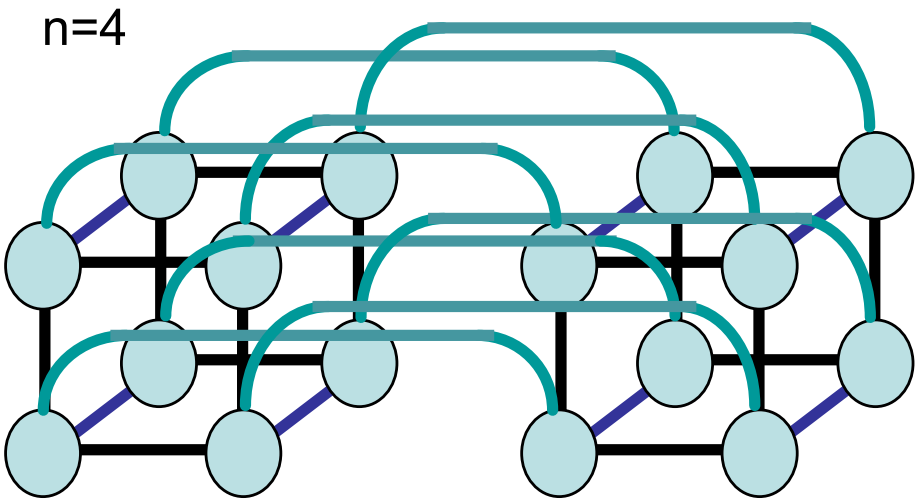
4 x 4 2D torus



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

Hypercube (n-cube)

2x2x2x2

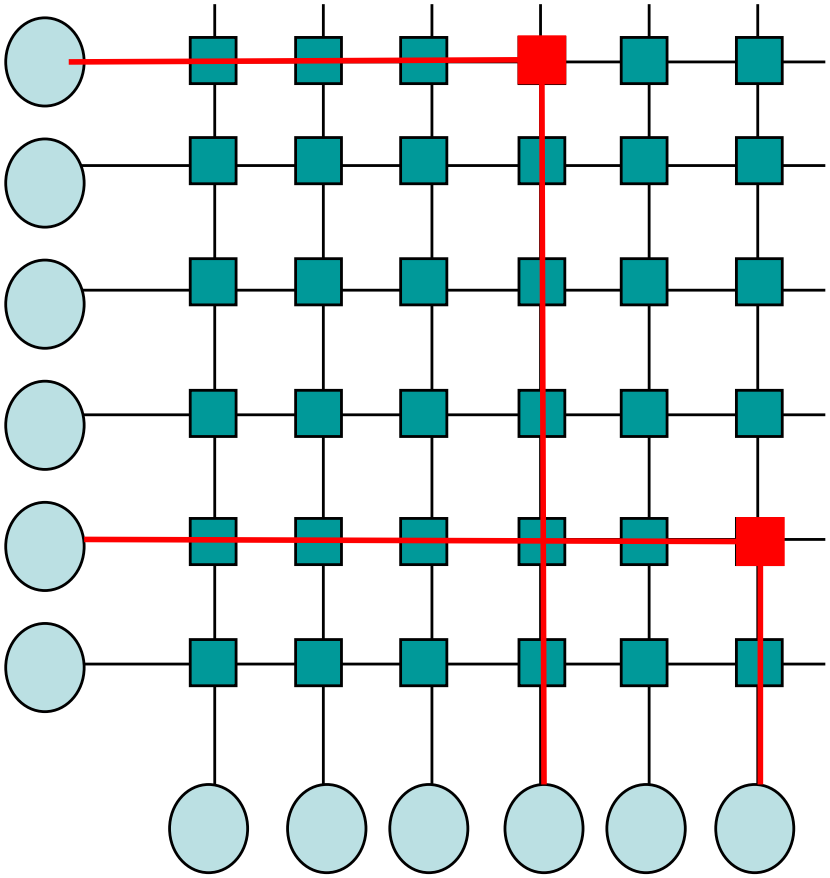


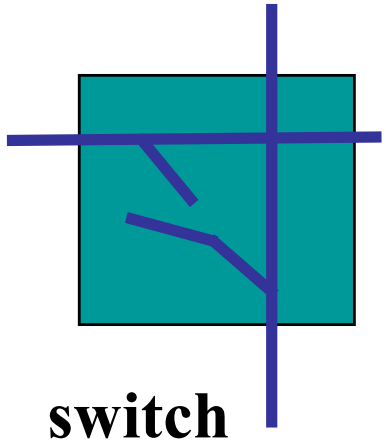
Cost: N (=2^n)
Diameter: n

間接網（動的網）

- ノードからは一般的に1本のリンクのみ（例外あり）
- 各ノードからのリンクを１つ以上のスイッチで結合してネットワークを形成
- スイッチでのルーティングが基本
- 代表的な間接網
  - Crossbar
  - MIN (Multistage Interconnection Network)
  - HXB (Hyper-Crossbar)
  - Tree (Indirect)
  - Fat Tree

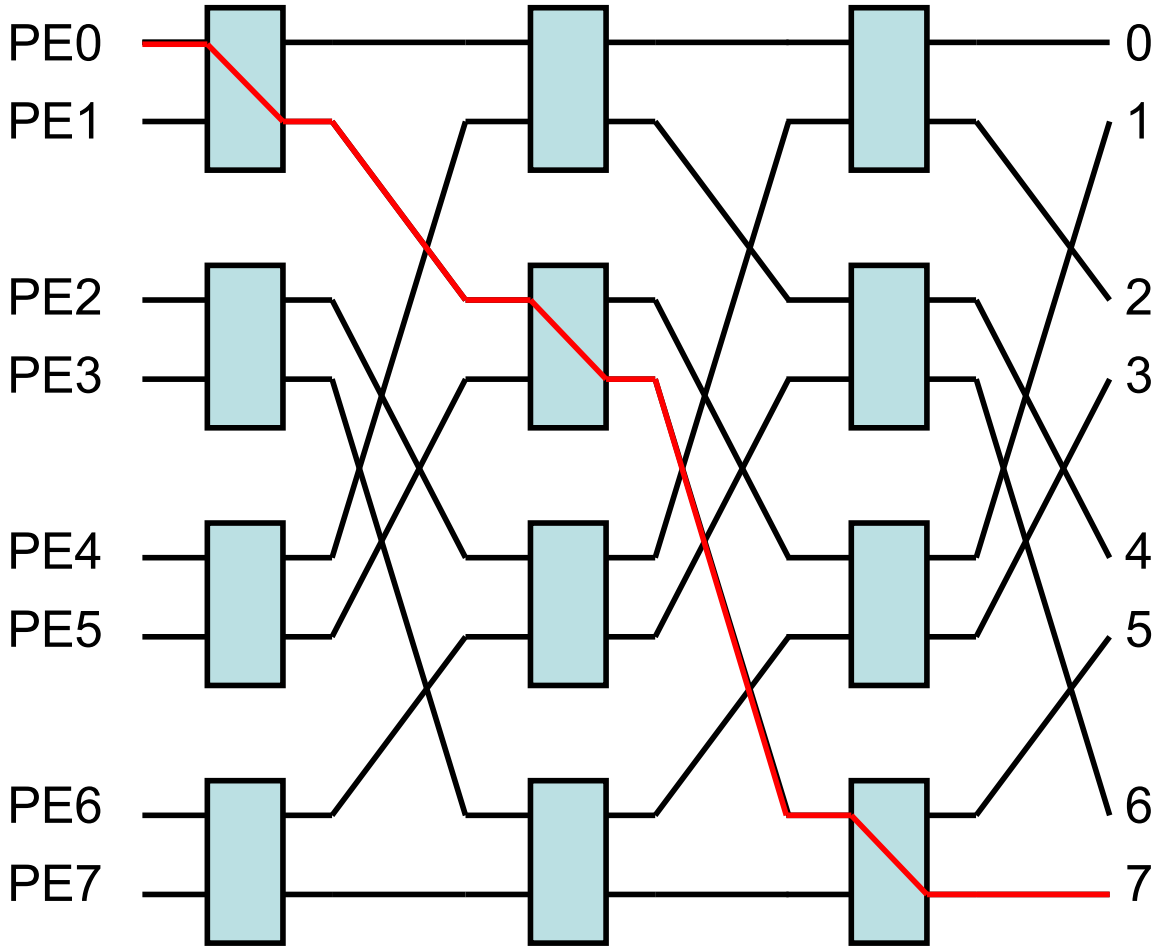
Crossbar



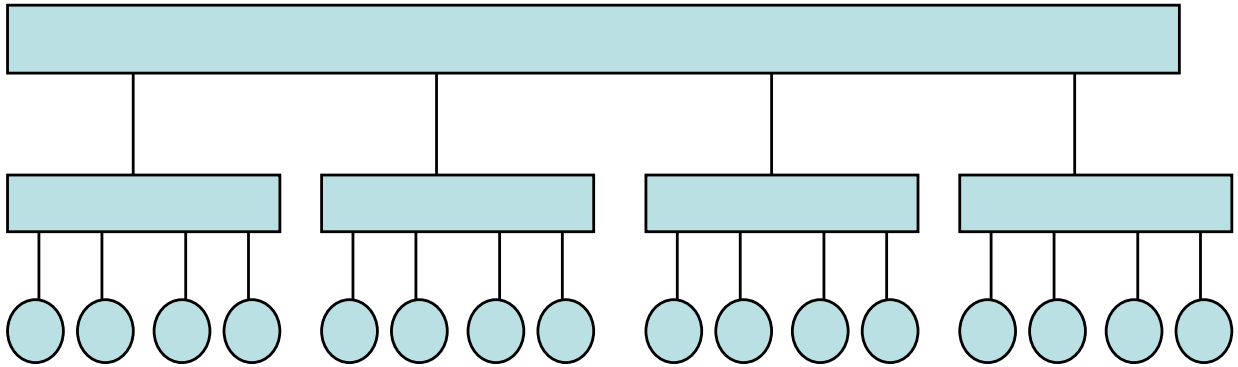


Cost: N^2
Diameter: 1

MIN (Multi-stage Interconnection Network)



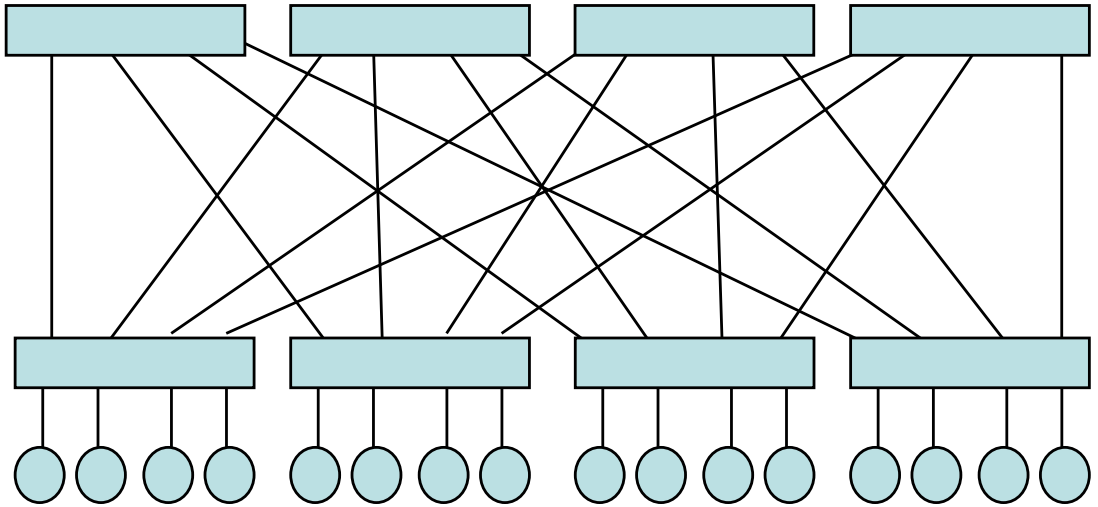
Cost: NlogN
Diameter: logN



Tree

Cost: N/k
Diameter: $2\log_k N$

Fat Tree



Cost: $N/k\log_k N$
Diameter: $2\log_k N$

並列処理ネットワークの性能メトリック

- メトリック = 物差し
  - 何を基準にネットワーク性能の良し悪しを測るか
  - 主な物差しなるのは「スループット」と「遅延時間」

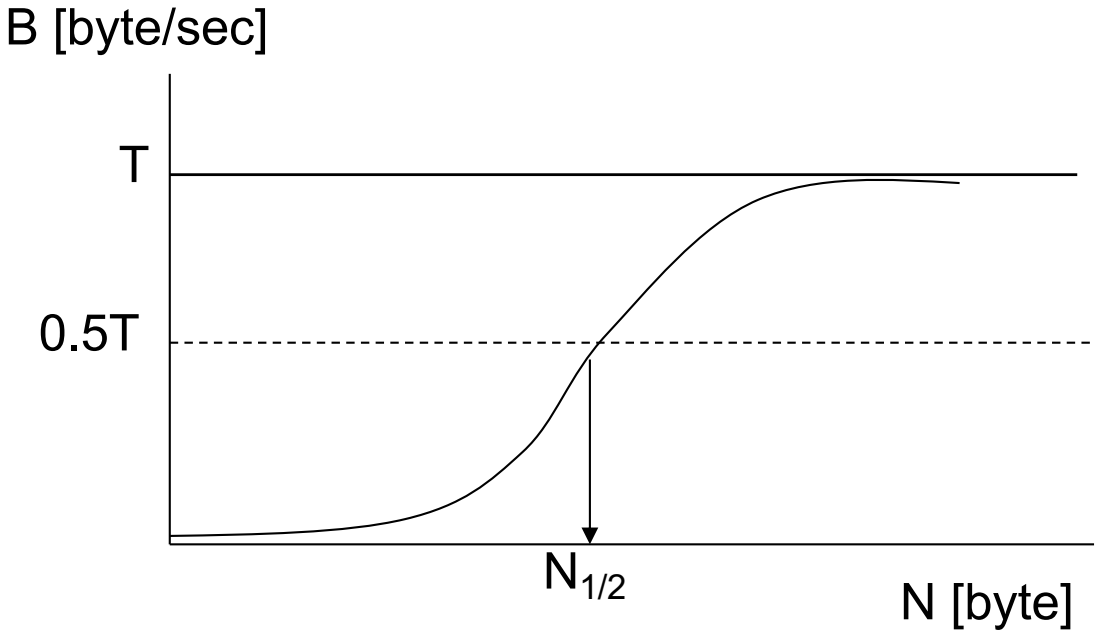
- Throughput（スループット）
  - リンクあるいはネットワーク全体の単位時間当たりのデータ転送性能
  - 単位：[byte/sec]（あるいは [bit/sec] ）

- Latency（遅延時間）
  - 狭義：転送すべきデータの先頭がsourceを出発してからdestinationに到着するまでの時間（ここではこれに従う）
  - 広義：転送すべきデータ全体がsourceを出発してからdestinationに到着するまでの時間
  - 単位：[sec]

ネットワーク転送性能とメッセージ粒度

- ネットワークリンク上で他のメッセージとの衝突がないとする。
  T [byte/sec]のスループットとL [sec]の遅延時間を持つネットワーク上で、*N* [byte]のメッセージを完全に転送し終わるまでの時間*t* [sec]と、有効バンド幅*B* [byte/sec]は以下のようになる

  $t = L + N/T$　　$B = N / t$



ここで、理論ピークバンド幅(T)の半分の0.5Tの性能が出るメッセージ長を$N_{1/2}$ (N-half「半性能長」)と表す。
理論的には

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

となる。
$N_{1/2}$は「この長さ以下ではLがdominantで、この長さ以上ではTがdominantである」ことを表し、これが小さい程、短いメッセージの通信に強いネットワークということになる。

「並列システム」内容

- 並列計算機アーキテクチャ
  - 分散メモリ, 共有メモリ, SMP, NUMA...
- 並列処理ネットワーク
- 実システムの紹介

実際の並列計算機概観

- システムの分類
  - MPP (超並列計算機)
    - 筑波大/日立 CP-PACS (SR2201)
    - LLNL/IBM Sequoia
    - ORNL/Cray Titan
    - 理研/富士通 富岳
  - 大規模並列ベクトル計算機
    - NEC 地球シミュレータ
  - スカラ並列計算機（クラスタを含む）
    - 筑波大/日立/富士通 PACS-CS
    - 筑波大・東大・京大/Appro・日立・富士通 T2K
  - アクセラレータ付ハイブリッド計算機
    - LANL/IBM Roadrunner
    - 東工大/HP (SGI) TSUBAME3.0
    - 筑波大/NEC Cygnus
    - NUDT Tianhe-2

TOP500リスト

- 全世界のスーパーコンピュータ（ただし申請ベース）の性能を１つの尺度で定量化し順位付けを行ったリスト
- 尺度＝Linpack（多次元連立一次方程式のガウスの消去法による直接求解）ベンチマークの性能（FLOPS）
- 毎年6月と11月の２回、リストを更新
  http://www.top500.org
- １つの数値で順位付けするためわかりやすい
- 問題の特徴として
  - ガウスの消去法のカーネル部分は小規模の行列×行列演算の帰着可能、キャッシュアーキテクチャでのデータ再利用性が非常に高い
  - ネットワーク性能は比較的低くても性能に大きく影響しない
- メモリバンド幅やネットワークバンド幅が比較的低くても性能が出るため、「本当にHPCベンチマークとして適当か」という議論はあるが、現時点での性能尺度として最も知られている
  （HPCベンチマークとしてより現実的なものも存在：HPCG等）

Green500

- TOP500の中で、電力あたりの性能（MFLOPS/W)をランク付けしたもの。
- 近年、電力供給が大規模並列計算機のボトルネックの一つといわれており、注目されている。
- 毎年6月と11月の２回、リストを更新
  http://www.green500.org/
- 性能値としてTOP500の値を用いているため、TOP500と同様の問題がある。
- TOP500に入っていることが条件であるが、その電力規模は大きく異なる(10MW – 30kW)。一般的に小電力システムのほうが電力あたりの性能を高めやすいので、単一の指標で良いか議論がある。そのため、大規模運用しているマシンの中で優秀なマシンを特別に表彰したりしている。

Oakforest-PACS (KNL cluster)

- JCAHPC (U. Tsukuba and U. Tokyo)
- Fujitsu PRIMERGY CX1640 M1 cluster
- Intel Xeon Phi (KNL)
- Intel OmniPath Architecture interconnection

- TOP500#6 2016/11
- 68 cores/CPU
- 8208 CPU (nodes)
- Peak 25PFLOPS
  HPL 13.5PFLOPS
  (スパコン京より性能が高い)

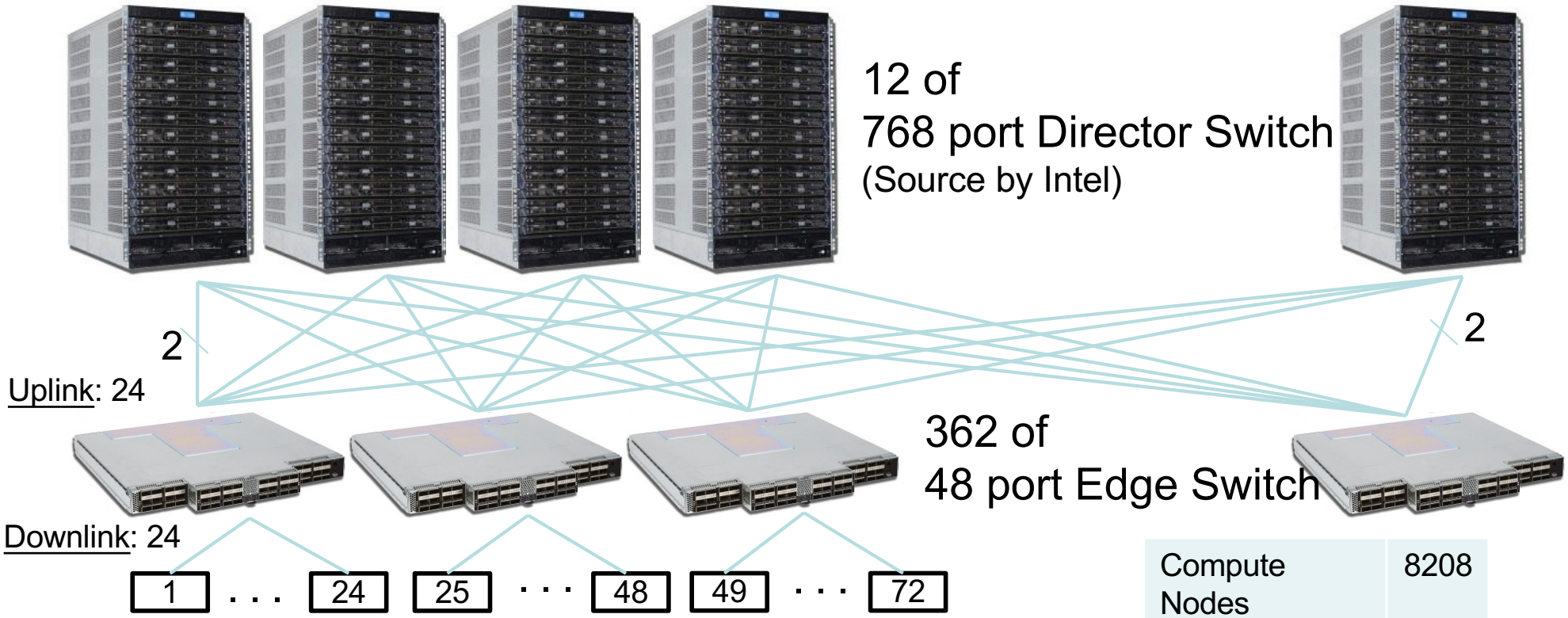
2022 年 3 月末をもってシステムを停止し、すべてのサービスを終了しました

Photo of computation node & chassis

Chassis with 8 nodes, 2U size

Computation node (Fujitsu next generation PRIMERGY)
with single chip Intel Xeon Phi (Knights Landing, 3+TFLOPS)
and Intel Omni-Path Architecture card (100Gbps)

Full Bisection Bandwidth Fat-Tree by Intel Omni-Path Architecture



Firstly, to reduce switches&cables, we considered :
- All the nodes into subgroups are connected with FBB Fat-tree
- Subgroups are connected with each other with >20% of FBB

But, HW quantity is not so different from globally FBB, and globally FBB is preferred for flexible job management.

Compute Nodes	8208
Login Nodes	20
Parallel FS	64
IME	200
Mgmt, etc.	8
Total	**8600**

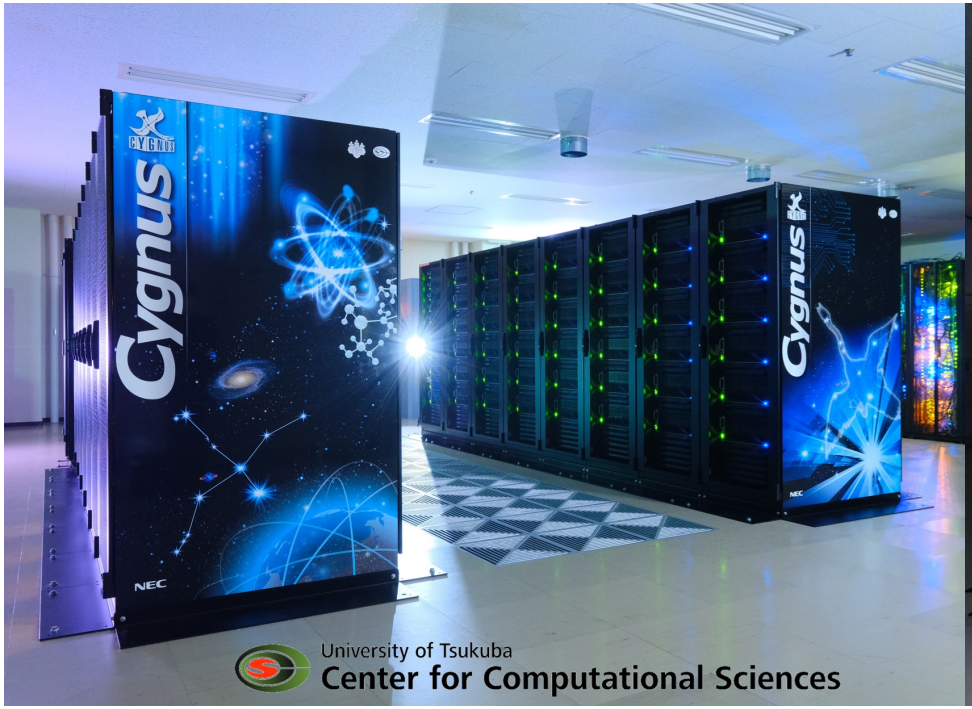
Specification of Oakforest-PACS system

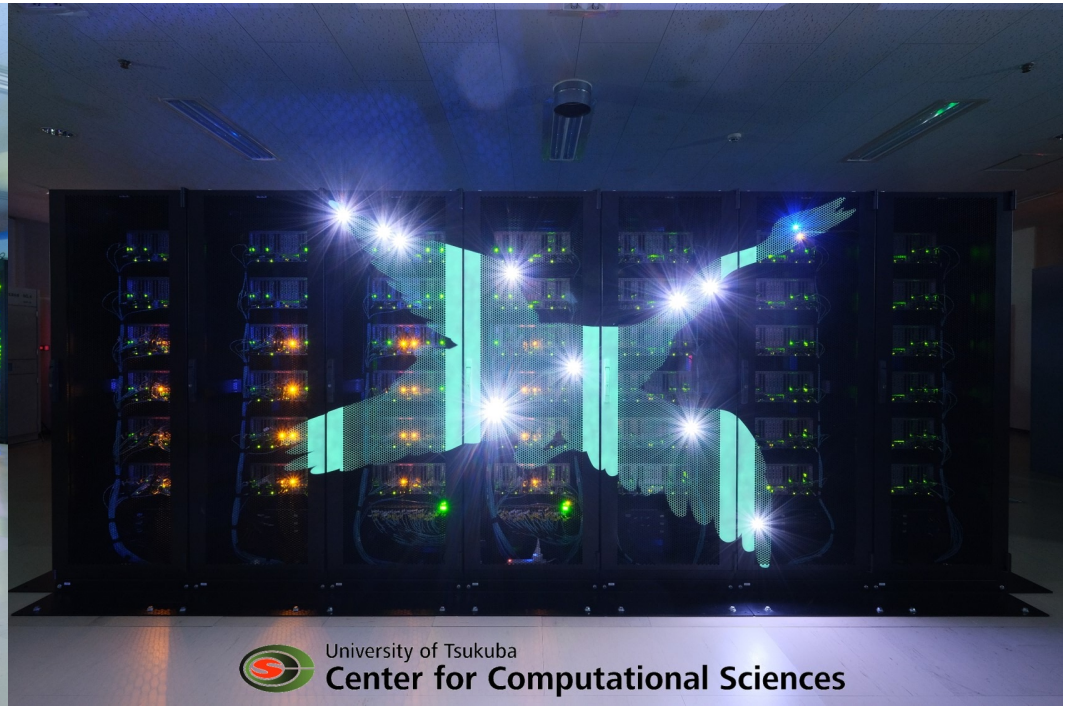
Total peak performance Linpack performance			25 PFLOPS 13.55 PFLOPS (with 8,178 nodes, 556,104 cores)
Total number of compute nodes			8,208
Compute node	Product		Fujitsu PRIMERGY CX1640 M1
	Processor		Intel® Xeon Phi™ 7250（Code name: Knights Landing）, 68 cores, 3 TFLOPS
	Memory	High BW	16 GB, > 400 GB/sec (MCDRAM, effective rate)
		Low BW	96 GB, 115.2 GB/sec (DDR4-2400 x 6ch, peak rate)
Inter-connect	Product		Intel® Omni-Path Architecture
	Link speed		100 Gbps
	Topology		Fat-tree with (completely) full-bisection bandwidth (102.6TB/s)
Login node	Product		Fujitsu PRIMERGY RX2530 M2 server
	# of servers		20
	Processor		Intel Xeon E5-2690v4 (2.6 GHz 14 core x 2 socket)
	Memory		256 GB, 153 GB/sec (DDR4-2400 x 4ch x 2 socket)

Specification of Oakforest-PACS system (I/O)

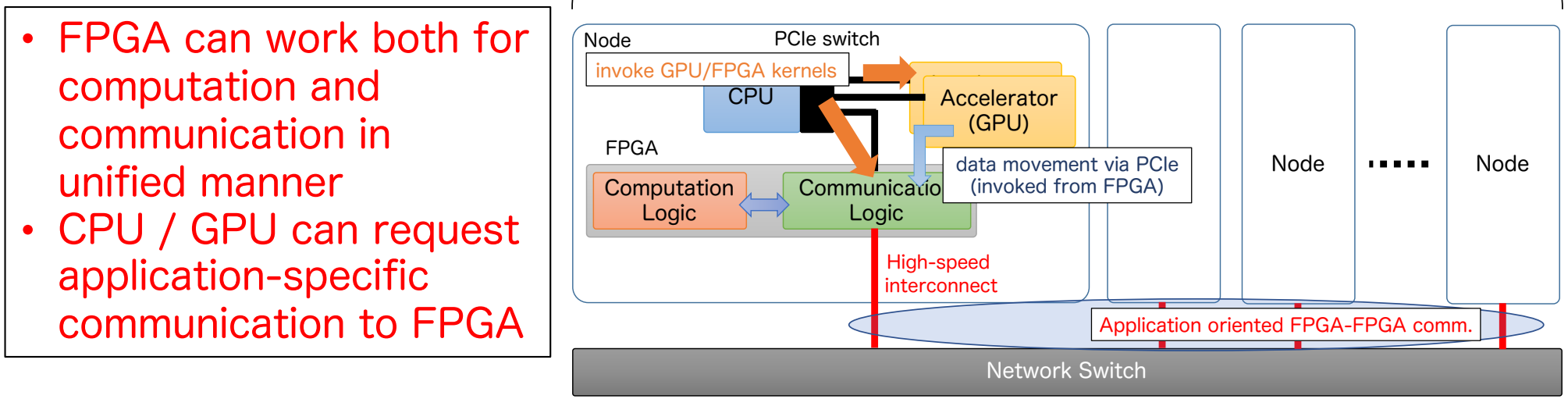
Parallel File System	Type		Lustre File System
	Total Capacity		26.2 PB
	Meta data	Product	DataDirect Networks MDS server + SFA7700X
		# of MDS	4 servers x 3 set
		MDT	7.7 TB (SAS SSD) x 3 set
	Object storage	Product	DataDirect Networks SFA14KE
		# of OSS (Nodes)	10 (20)
		Aggregate BW	500 GB/sec
Fast File Cache System	Type		Burst Buffer, Infinite Memory Engine (by DDN)
	Total capacity		940 TB (NVMe SSD, including parity data by erasure coding)
	Product		DataDirect Networks IME14K
	# of servers (Nodes)		25 (50)
	Aggregate BW		1,560 GB/sec

Cygnus (Univ. of Tsukuba)





CHARM: Cooperative Heterogeneous Acceleration with Reconfigurable Multidevices

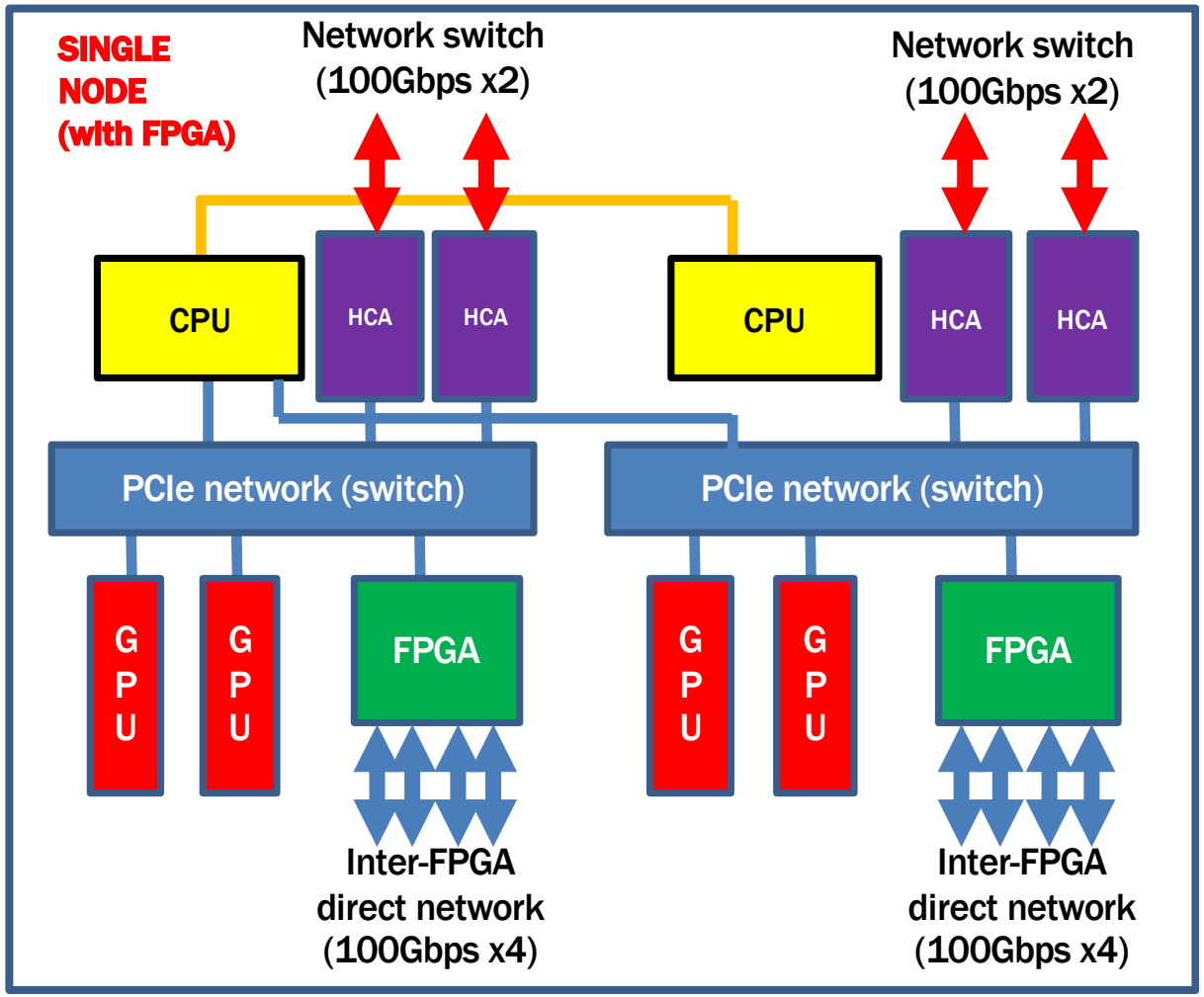


- FPGA can work both for computation and communication in unified manner
- CPU / GPU can request application-specific communication to FPGA

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

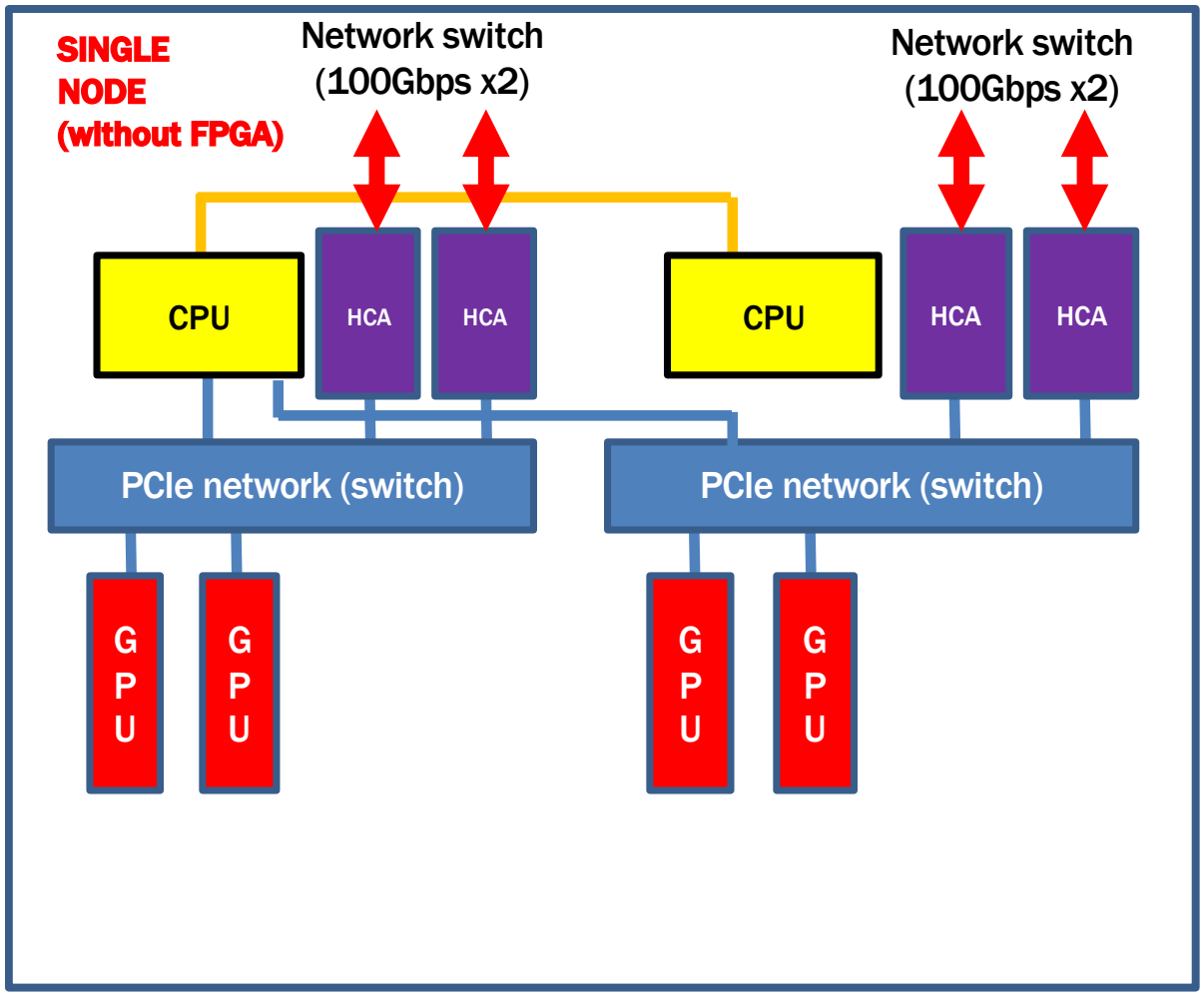
Albireo node



Single node configuration (Deneb)

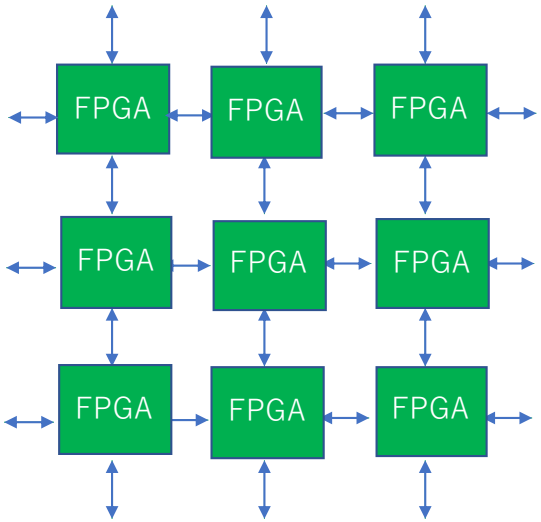
- Each node is equipped with both IB EDR and FPGA-direct network
- Some nodes are equipped with both FPGAs and GPUs, and other nodes are with GPUs only

Deneb node



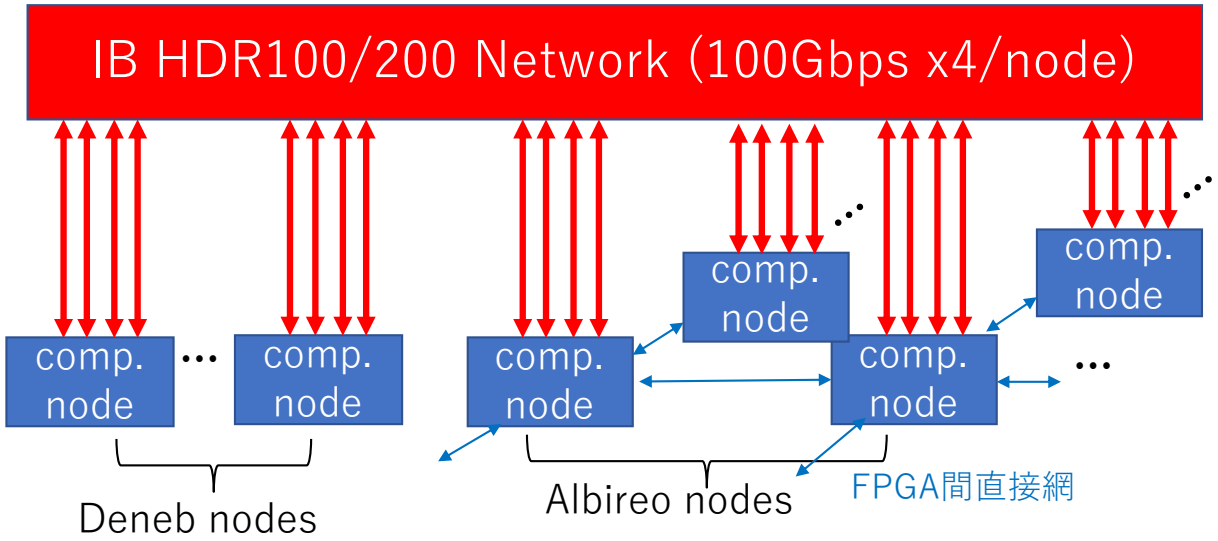
Two types of interconnection network

FPGA間直接網（Albireoのみ）



Albireoノードの64台のFPGA（2 FPGA/node）は2次元トーラスネットワークによりスイッチなしで結合される

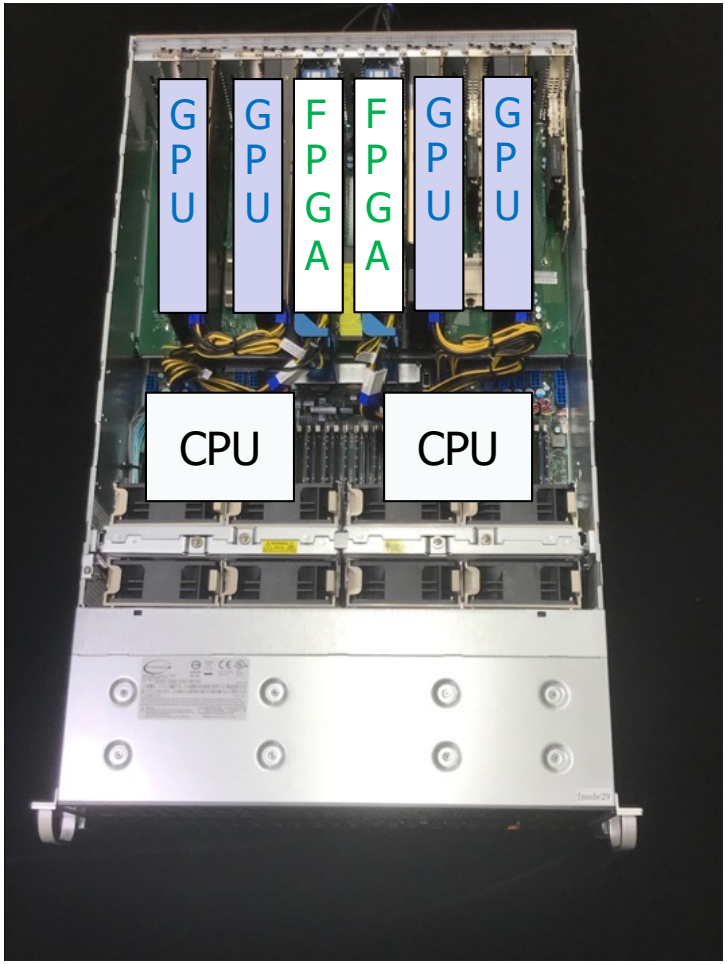
全ノードに共通する並列処理/ストレージアクセス用相互結合網

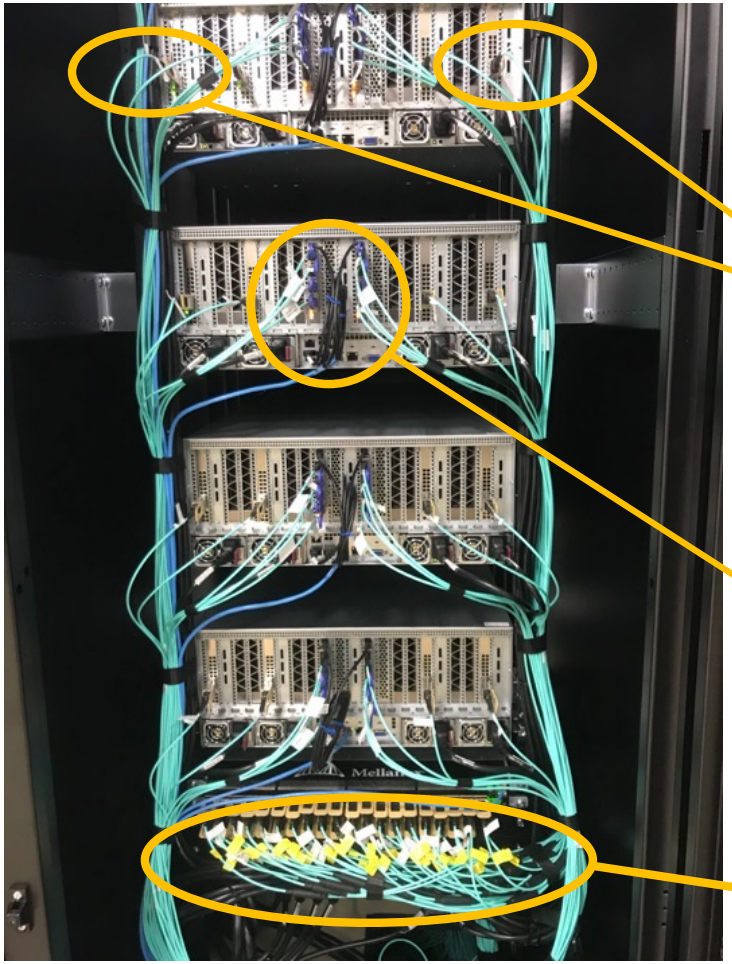


全ノード共通の並列処理ネットワークはInfiniBandによる通常のネットワークで、MPI等による並列処理が可能。CPU及びGPUの通信に用いるが、FPGAからCPUを介しての通信も可能。

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





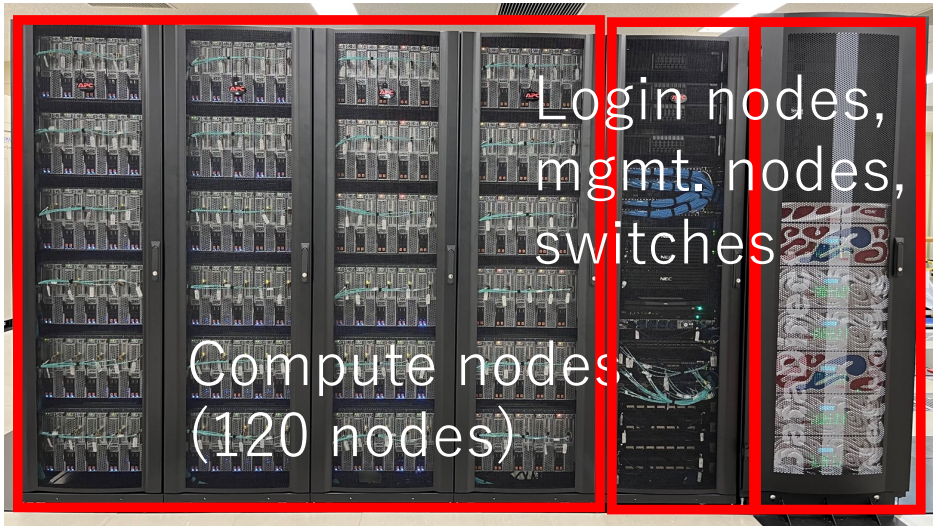
IB HDR100 x4
⇨ HDR200 x2

100Gbps x4
FPGA optical
network

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

Pegasus (Univ. of Tsukuba)

- 4th Gen Intel Xeon SP, NVIDIA H100 Tensor Core PCIe GPU , Intel Optane persistent memoryを搭載し,
  ビッグデータとAIを強力に推進する



Parallel File System

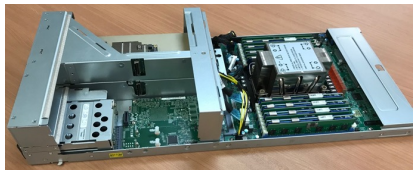
Pegasusの仕様

Pegasusの設計目標

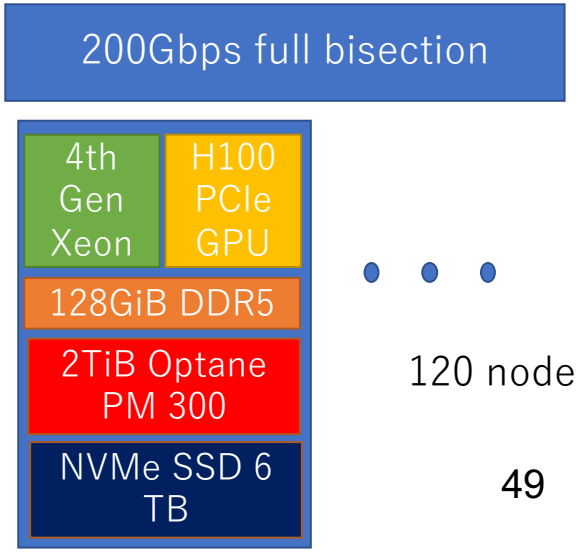
- 大容量メモリでかつ高性能ストレージである不揮発性メモリを活用し，大規模データ分析やビッグデータAIを高速化
- 大規模データ分析やビッグデータAIの新規アプリケーション，システムソフトウェア研究の新分野を開拓

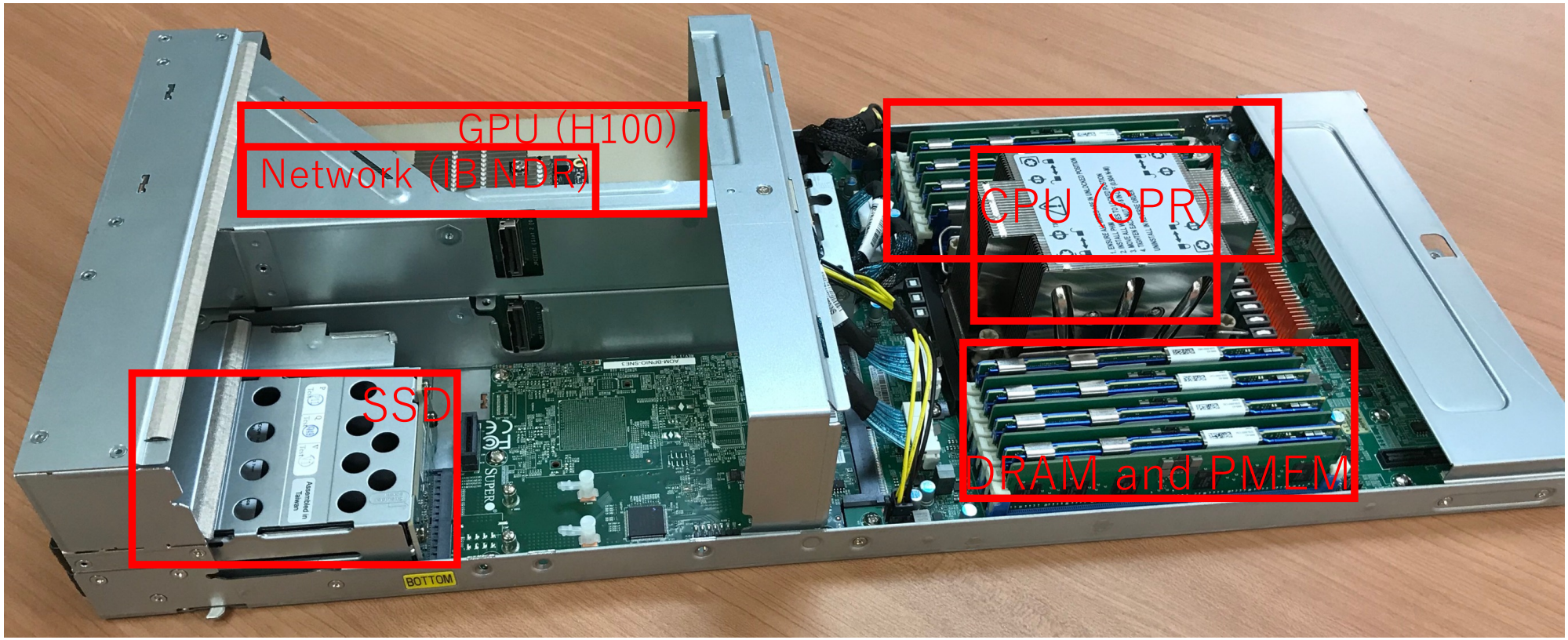
- 2022年第4四半期に導入
- Total Performance
  - 120 nodes, 6.5 PFlops, 240 TiB Pmem
- ノード仕様
  - 3.2 TFlops Intel Xeon Platinum 8468 (codenamed Sapphire Rapids)
  - 51 TFlops NVIDIA H100 PCIe GPU (80GB mem)
  - 128 GiB DDR5 DRAM (282 GB/s)
  - 2 TiB Optane DCPMM 300 series
  - 6 TB NVMe SSD (7 GB/s)
- Interconnection Network
  - NVIDIA Quantum-2 InfiniBand platform NDR200 (200 Gbps) full bisection
- Parallel File System
  - 7.1 PByte DDN EXAScaler (40 GB/s)

NEC LX B1000E Blade Enclosure



NEC LX 102Bk-6




GPU (H100)
Network (IB NDR)
CPU (SPR)
SSD
DRAM and PMEM
BOTTOM

富岳

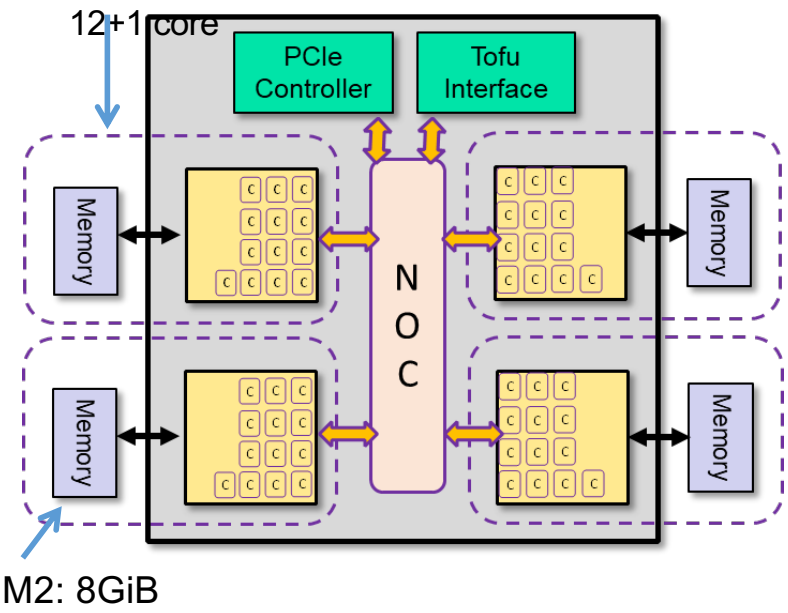
- 理化学研究所・計算科学研究センター
- 独自開発（富士通）・48コア・メニーコアプロセッサ x 15万台
- Linpack 442PFLOPS
- ARMベースのプロセッサを本格的に採用、高性能演算命令追加
- 他の多くのTOP10マシンがアクセラレータ（演算加速装置：GPU等）を使っているが、汎用のプログラムしやすいプロセッサだけで作られている

CPU Architecture: A64FX

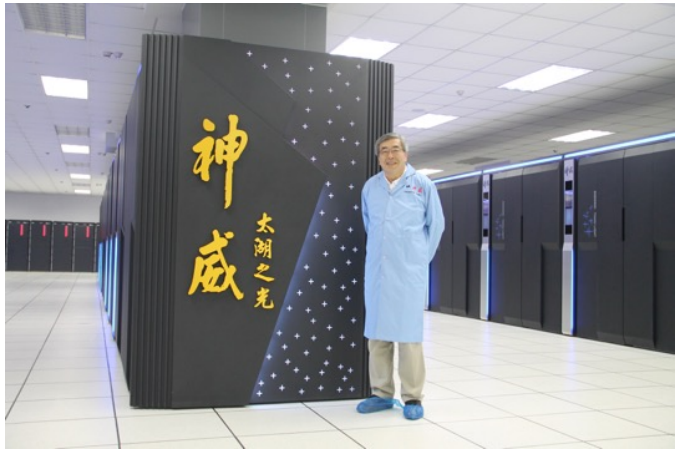
- **Armv8.2-A (AArch64 only) + SVE (Scalable Vector Extension)**
  - FP64/FP32/FP16 (https://developer.arm.com/products/architecture/a-profile/docs)
- **SVE 512-bit wide SIMD**
- **# of Cores: 48 + (2/4 for OS)**
- Co-design with application developers and high memory bandwidth utilizing on-package stacked 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



TaihuLight（太湖之光, Sunway）



- 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

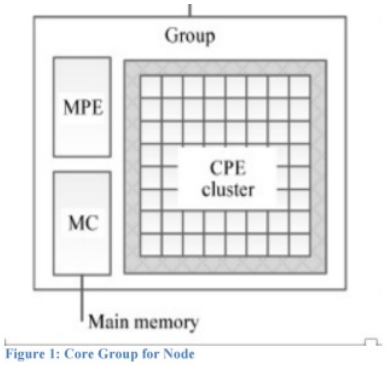


Figure 1: Core Group for Node

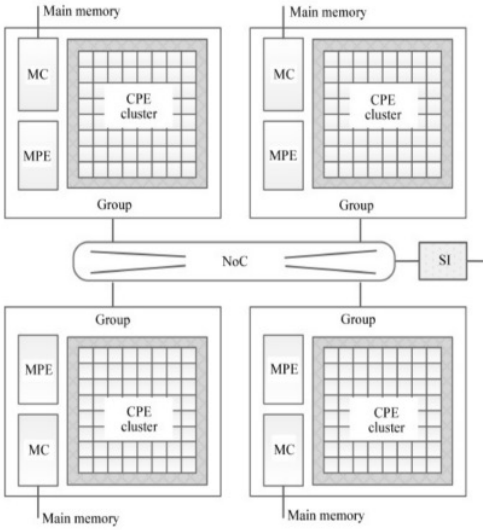


Figure 2: Basic Layout of a Node

- 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

Frontier (ORNL)

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

2023年6月時点でTop500ランキング1位

- 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 1.2EFLOPS

並列処理システムの動向

- MPPは徐々に衰退（特定マシンのみ躍進）
- コモディティ化が進む（クラスタの台頭）
  - コモディティなスカラープロセッサ（64bit IA32=x64）
  - コモディティなネットワークとスイッチ
    - InfiniBand (EDR 100Gbps、高級機器だったが徐々に価格低下）
- 全体的に、演算性能：メモリ性能：通信性能のバランスが悪化
  - 演算性能はプロセッサのmulti-core化等により順調に向上
  - メモリ性能（バンド幅）は相対的に低下（プロセッサが速すぎる）
  - 通信性能は段階的に上がっていく（IB等）
  - プロセッサコストは*O(N)*だがネットワークコストは*O(N log N)*程度なので相対的にシステム価格を圧迫
  - 結果的に並列処理効率を上げるのが難しくなってきている。より一層のアルゴリズム、ソフトウェア上の工夫が必要。
- Exa FLOPS マシンに向けた研究開発も始まっている
  - 2022年6月についにエクサフロップスを達成するマシンが登場！
  - 性能/電力の１００倍程度の向上が必要 (アクセラレータの活用)
  - １０００万並列を効率よく利用できるアルゴリズムの開発

まとめ

- 並列処理システム／アーキテクチャ
  - 逐次プロセッサ（コア）性能の限界により、全体性能は並列処理に頼らざるを得ない
  - 性能を維持しつつ拡張性（scalability）を確保
  - 分散メモリ / 共有メモリ
- 並列処理ネットワーク
  - scalabilityが最も重要
  - 以前はMPP向け、現在はcommodity networkの充実によりfat-treeでかなりの規模が可能
  - ２つの性能メトリック：throughput & latency
- 並列処理システムの実際
  - 2023年6月で，最高Linpack性能約1.2EFLOPS (エクサ時代の到来)
  - 基本は分散メモリシステムだがmulti-coreの一般化によりハイブリッドが基本
  - アクセラレータが注目されている (GPUがポピュラー)

課題

2023/06のTop500で第1位、2位、4位の３つの計算機について以下の問いに答えよ。

1. 各計算機の最大性能について、コアあたり、チップあたり、ノードあたり、システム全体について示せ。コアあたりについては、その根拠を示すこと。ただしノードがヘテロジニアスな構成の場合は各構成要素についても示すこと。
2. Linpack実行時のシステム全体の電力、実行効率などTop500のリストから考察される各計算機の特徴について述べよ。なお、Top500のリストはhttp://www.top500.org/からエクセルの形式でダウンロードできる。
3. 各計算機のメモリシステム、ネットワークの構成について、その特徴を比較して述べよ。