筑波大学計算科学研究センターにおける PCクラスタのあゆみ

朴 泰祐

筑波大学計算科学研究センター・センター長

筑波大学計算科学研究センター

- **Center for Computational Sciences (CCS)**
- 広範な計算科学分野の研究者と高性能計算工学分野の研究者が常駐する研究センター
  - 計算科学分野
    - 素粒子物理，宇宙物理，物性物理，地球環境，生命情報
  - 計算機工学分野
    - 高性能計算システム，グリッド，大規模データベース，マルチメディア
- 応用側のニーズとシステム側のシーズの融合
  - アプリケーションの研究者とシステムの研究者の日常的な共同研究
  - このようなセンターは日本中にほとんどない
- 実応用に即した高性能計算に関する研究の日常的推進
- 研究に必要な計算機を、ただ買ってくるのではなく自ら設計し作り上げる（メーカーとの共同研究開発で）

PACSとは？

- 筑波大学においてアプリケーション指向で開発・導入を続けてきた並列計算機・スーパーコンピュータのプロジェクト名
  - 1992年の計算物理学研究センター（CCP）設立 ⇒ CP-PACS (PACS-VI), MPP, 筑波大+日立
  - 2004年の計算科学研究センター（CCS）への改組
    ⇒ PACS-CS (PACS-VII), HA-PACS (PACS-VIII), COMA (PACS-XI), Cygnus (PACS-X)
- 星野力教授＋川合敏雄氏（当時日立）による「マイクロプロセッサの並列処理で核炉心シミュレーションができないか」の発想から8bit マイクロプロセッサMC6800を9台，2次元トーラス結合したPACS-9が開発された（1978年）
- PACS = Processor Array for Continuum Simulation
  ⇒ 後に PAX (Processor Array eXperiment) (PAX-128から)
  ⇒ さらにCP-PACS = Computational Physics by Parallel Array Computer System
  ⇒ さらに PACS-CS = Parallel Advanced Computer System for Computational Sciences
  ⇒ さらに HA-PACS = Highly Accelerated Parallel Advanced system for Computational Sciences
- HA-PACSの後，「PACSなんとか」という命名を廃止，ローマ数字の番号を補助名とし，システム名は独立に付けるようになった（システムニックネームは付ける）

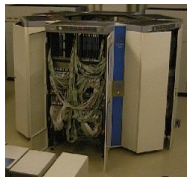
筑波大学におけるPACS (PAX)シリーズ開発の歴史

- 1977: research started by T. Hoshino and T. Kawai
- 1978: PACS-9 (with 9 nodes) completed
- 1996: CP-PACS, the first vendor-made supercomputer at CCS, ranked as #1 in TOP500

1978
1st gen: PACS-9

1980
2nd gen. PACS-32

1989
5th gen, QCDPAX



1996
6th gen: CP-PACS
Ranked #1 in TOP500

2006
7th gen: PACS-CS

2012~2013
8th gen: GPU cluster HA-PACS

2014
9th gen: COMA

2019
10th gen: Cygnus

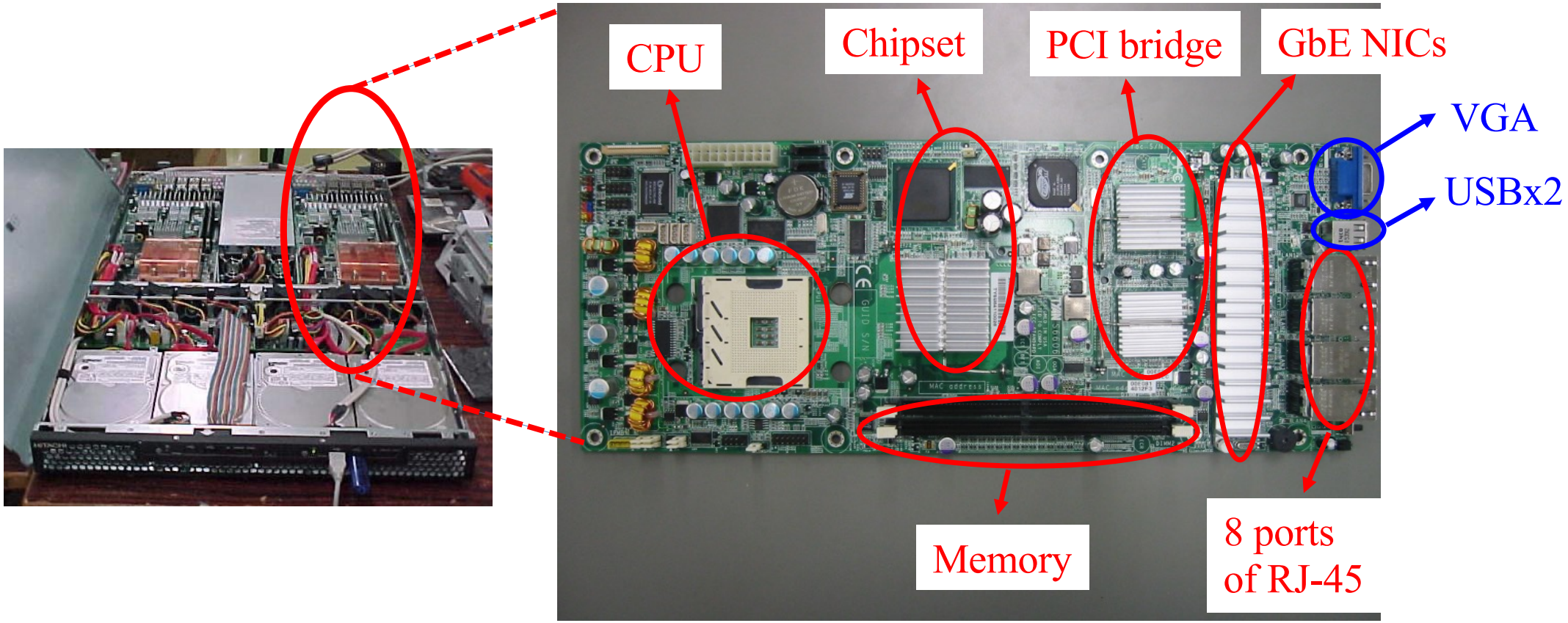
Year	Name	Performance
1978	PACS-9	7 KFLOPS
1980	PACS-32	500 KFLOPS
1983	PAX-128	4 MFLOPS
1984	PAX-32J	3 MFLOPS
1989	QCDPAX	14 GFLOPS
1996	CP-PACS	614 GFLOPS
2006	PACS-CS	14.3 TFLOPS
2012~13	HA-PACS	1.166 PFLOPS
2014	COMA (PACS-IX)	1.001 PFLOPS
2019	Cygnus (PACS-X)	2.5 PFLOPS

- *co-design* by computer scientists and computational scientists toward “practically high speed computer”
- Application-driven development
- Sustainable development experience

変革１：PCクラスタの導入～PACS-CS (2006)

- 1996/11にTOP500 #1となったCP-PACS (HPL: 360GFLOPS)の後，しばらく大型システム予算がつかずCP-PACSを10年近く運用
  ⇒#1になったマシンは退役までTOP500を落ちることはなかった
- 概算要求ベースの研究システムをようやく導入⇒実装をどうする？
  - もはやCP-PACS型のMPP開発は大学予算では不可能
  - PCクラスタで検討，ただし
    - 通常のクラスタではCPU性能：メモリバンド幅：ネットワークバンド幅の比率が悪い
    - CS (Computational Science)のためには総ピーク性能追求よりもバンド幅追求で実行効率重視
    - 相互結合網は？（Myrinetなど？）
  - 安価なGbEthernetを束ね，latencyよりもbandwidthを稼ぐ戦略
    - Single core CPU x 1, DDR memory, GbE x 8 (6 for computation, 2 for management)
  - 6本のGbEで3次元のHXB (Hyper Crossbar Network)をソフトウェ実装 (PM-Ethernet for 3D HXB)

Motherboard+chassis



ソフトウェア：SCoreの導入⇒2560 nodeシステムが完成

- CP-PACSで培った3-D HXB (Hyper Crossbar)ネットワークの資産（アルゴリズム上のルーティングなど）を活用
- ネットワークはEthernetなのでソフトウェアでのルーティングを行う
  - SCoreを導入し，ネットワークドライバであるPMを3D-HXB用に構築
- ハードウェアは日立，ソフトウェアは富士通という画期的なタッグ
  - SCoreとPM-Ethernetの実装技術は富士通がリードしていた
  - 筐体設計は日立，3D-HXBの経験も豊富
- 3-DHXBはメッシュ・トーラスに近い特性を持っていて，ノード間のジッターに弱かった
- SCoreの導入を機にPCクラスタコンソーシアムに加入

HA-PACS (2560 node), 2006
Hitachi + Fujitsu
16x16x10 3D-HXB with PM/SCore
Peak 14.3 TFLOPS
HPL 10.35 TFLOPS (#34 in TOP500, 2006/06)

変革２：マルチコアCPU時代～T2K-Tsukuba

- 筑波大・東大・京大のアライアンス (T2K) に基づく共通基本仕様のクラスタの構築
  - AMD Opteron (quad-core)の全面採用
  - InfiniBand DDR x 4chan. によるFull Bisection Bandwidth Interconnect
- PACSシリーズには含まれていない
  - PACS (PAX) は研究開発予算によるシステム開発・実装・運用
  - T2K-TsukubaはVPP-5000で運用されていたスーパーコンピュータ・レンタル予算の系列（元々は学内センターだった学術情報メディアセンターのスーパーコンピュータ予算をCCSに移行）
    ⇒ Oakforest-PACSも同様
- 筑波大として初めて外資系システムを導入
  - Appro + Cray Japan
  - Mellanox, DDN

T2K-Tsukuba (640 node), 2008
Appro X3-Server, Opteron quad-core x 4/node
Peak 95.3 TFLOPS
HPL 76.46 TFLOPS (#20 in TOP500, 2008/06)

変革３：GPUの導入～HA-PACS

- 本格的なGPU時代に突入，演算性能重視＋アクセラレータ対応コード開発
- HA-PACS = Highly Accelerated PACS (PACS-VIII)
  - PCI gen3 x 40 lane を搭載したIntel SandyBridge x 2でNVIDIA M2090 (Fermi) GPUx4をフルバンド幅でサポート
  - InfiniBand QDR x 2
- その後拡張部分としてTCA (Tightly Coupled Accelerator)というFPGAを追加したHA-PACS/TCAを導入，Green500で3位を獲得（NVIDIA Kepler K40導入）
  - 3.5 GFLOPS/W (HPL ratio 76%)
  - IvyBridge x 2
  - TCA (Altera Stratix-V) interconnect for GPU direct comm.

HA-PACS (268 node), 2012
Appro, Intel SDB, NVIDIA M2090
Peak 802TFLOPS
HPL 421.6 TFLOPS (#41 in TOP500, 2012/06)

変革４：many-core導入～COMA, Oakforest-PACS

- COMA (Cluster Of Many-core Architecture, PACS-IX)
  - Intel KNC (Knights Corner)
  - Xeon + KNC
  - InfiniBand FDR
- Oakforest-PACS (JCAHPC with 東大)
  - Intel KNL (Knights Landing)
  - Intel OmniPath (OPA) x 4
  - Linux + McKernel
  - MCDRAM
    - cache mode | flat mode

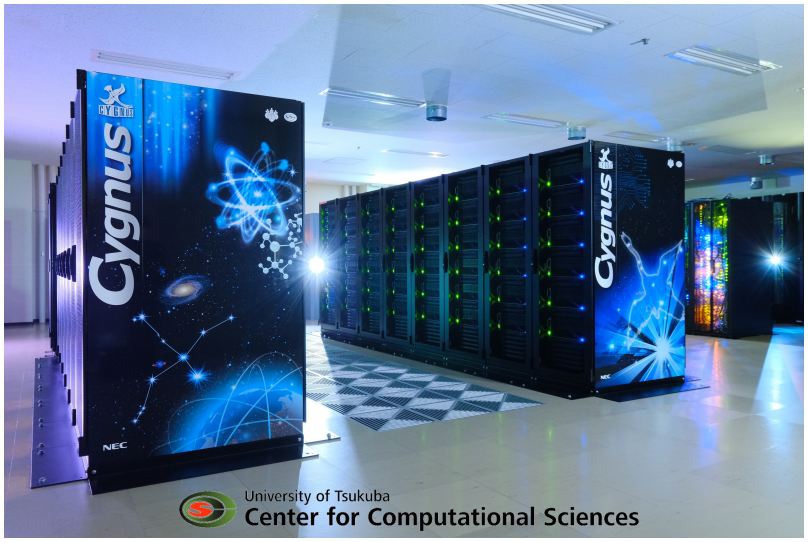
COMA (393 nodes), 2014
Cray, Intel KNC+IB FDR
Peak 1001 TFLOPS
HPL 746 TFLOPS
(#51 in TOP500 2014/06)

Oakforest-PACS (8208 nodes), 2016
Fujitsu, Intel KNL+OPA
Peak 25 PFLOPS
HPL 13.5 PFLOPS
(#6 in TOP500 2016/11)

変革５：複数種類のアクセラレータ

- Cygnus (PACS-X)
- GPUだけでは最終的に加速できないコード部分（アムダール則的に）にFPGAを投入
  ⇒ Multi-Hybrid Accelerating Supercomputer
- CHARM concept (Cooperative Heterogeneous Acceleration with Reconfigurable Muilti-devices)
- NVIDIA Tesla V100 x4 + Intel Stratix10 FPGA (32 nodeのみ)
- 宇宙物理コードにおいてGPUのみの場合に対し最大17倍の高速化

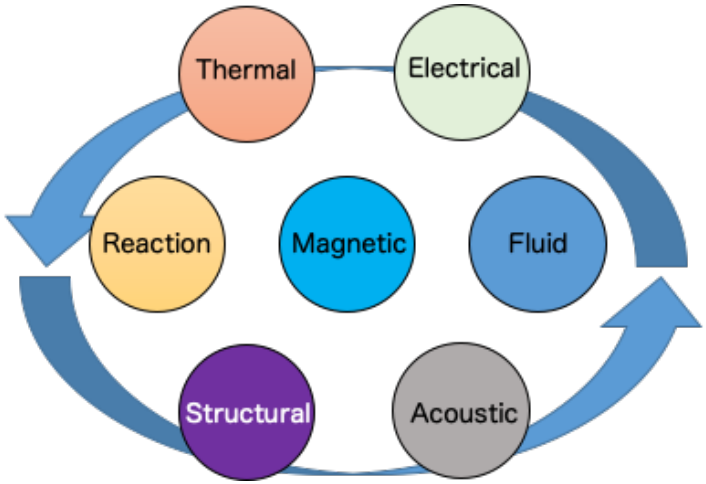
The world first practical supercomputer with Multi-Hybrid (GPU + FPGA) Accelerating Architecture: 320 GPUs (V100) + 64 FPGAs (Stratix10) in 80 nodes



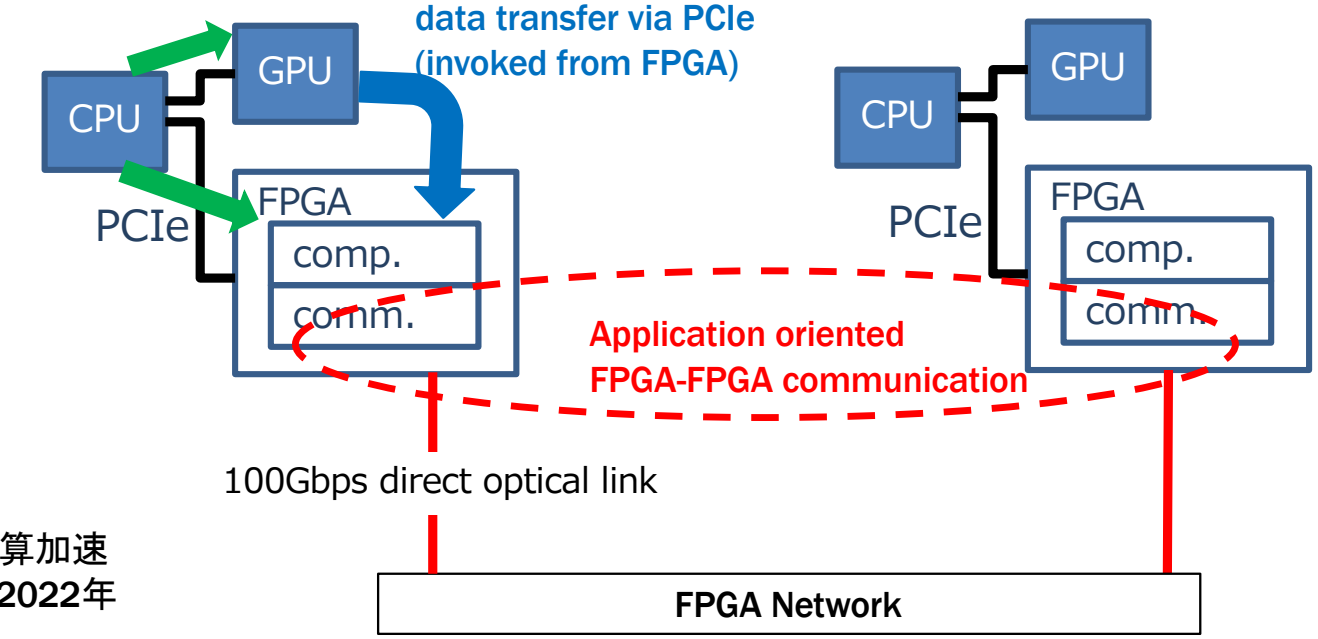
Cygnus (81 nodes) + 4 nodes, 2019 (+2020)
NEC, NVIDIA V100 x4 + Intel Stratix10 x 2 (32 nodes)
Peak 2.5 PFLOPS
HPL 1.58 PFLOPS (#264 in TOP500 2019/06)

CHARM: Cooperative Heterogeneous Acceleration with Reconfigurable Multi-devices

multi-physics/multi-scale complicated problem

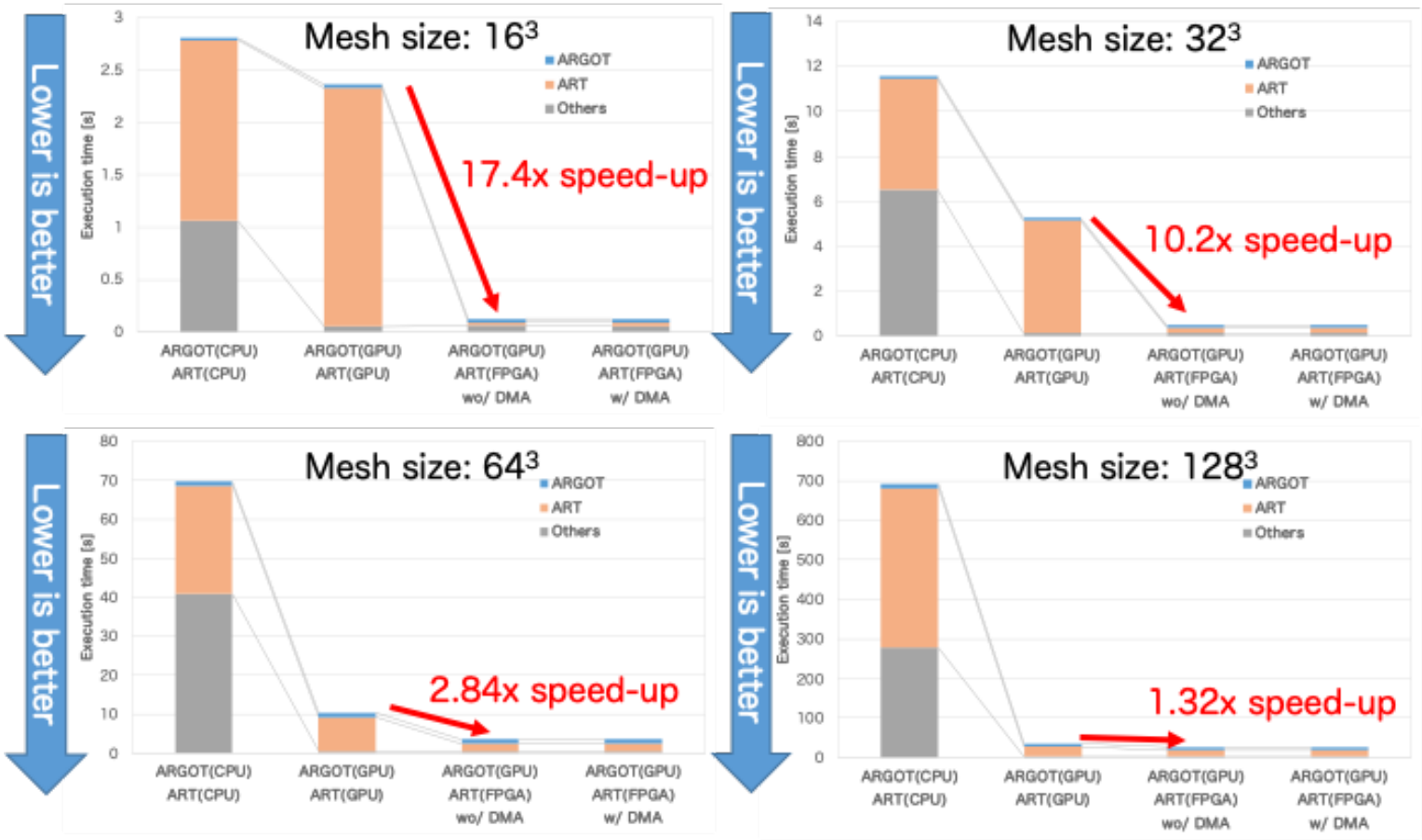


Basic cluster with GPUs (by InfiniBand)



（参考）朴泰祐, “oneAPIによるFPGAとGPUの複合演算加速アプリケーション実行”, Intel FPGA Technology Day, 2022年12月9日

GPU-only vs GPU-FPGA coworking on astrophysics simulation ARGOT



Single node GPU (V100) + FPGA (Stratix10) evaluation on Cygnus
R. Kobayashi, et. al., “Accelerating Radiative Transfer Simulation with GPU-FPGA Cooperative Computation”, ASAP2020, Jul. 2020.

PCクラスタの導入の最大のメリット＝拡張性+自由度

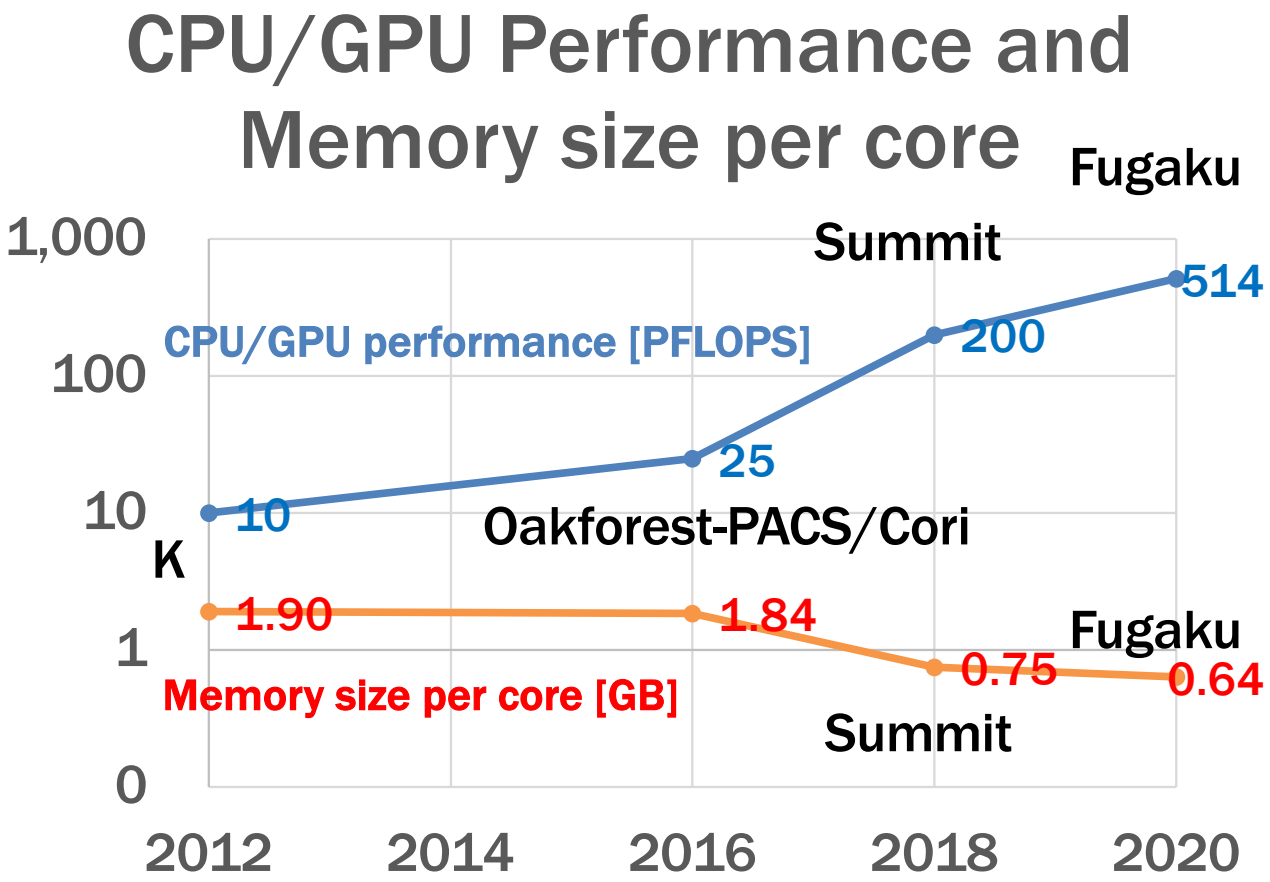
- PACS-CSから4世代に渡るPACSシステム（+T2K+OFP)の演算装置の変遷
  - single-core
  - multi-core
  - GPU
  - many-core
  - GPU+FPGA
- 特にGPU, FPGAの導入はPCクラスタだからこそ可能
- 演算器，ネットワーク，メモリ，ストレージの組み合わせをアンサンブルできることがPCクラスタのメリット
- 筑波大CCSはcodesignの「老舗」
  ⇒ 基本的なクラスタの構成に対し演算加速やネットワーク等のシステム拡張のチャレンジを行ってきた
  ⇒ 拡張容易性が大事

次の一手=Cygnus-BD (Big Memory Supercomputer)

- key word: 大規模計算科学と超高速ファイルシステムのためのNVRAMの活用
  - PMEM (Persistent Memory)技術が実用になっている
  - PMEMは memory mode と I/O mode (Intel: app-direct) の両面を持ち混用も可能
  - memory modeではDDRをlast level cacheとしてその向こうにPMEMという使い方もできる
- Extreme Computing + Big Data + AI
  - 宇宙物理，気象などではノード内の演算性能に対しメモリ容量が小さ過ぎる
  - アプリによってはバンド幅よりも容量が大事
  - in-situ的なアプリケーションのカップリング（例：HPC+AI）では local file systemの速度も重要
- AI for HPC (HPC with AI): Cygnus-EC + Cygnus-BD が次の一手
  - Cygnus-EC: Cygnus for Extreme Computing（これまでのCygnusをrename）
  - Cygnus-BD: Cygnus for Big Data（新規調達）

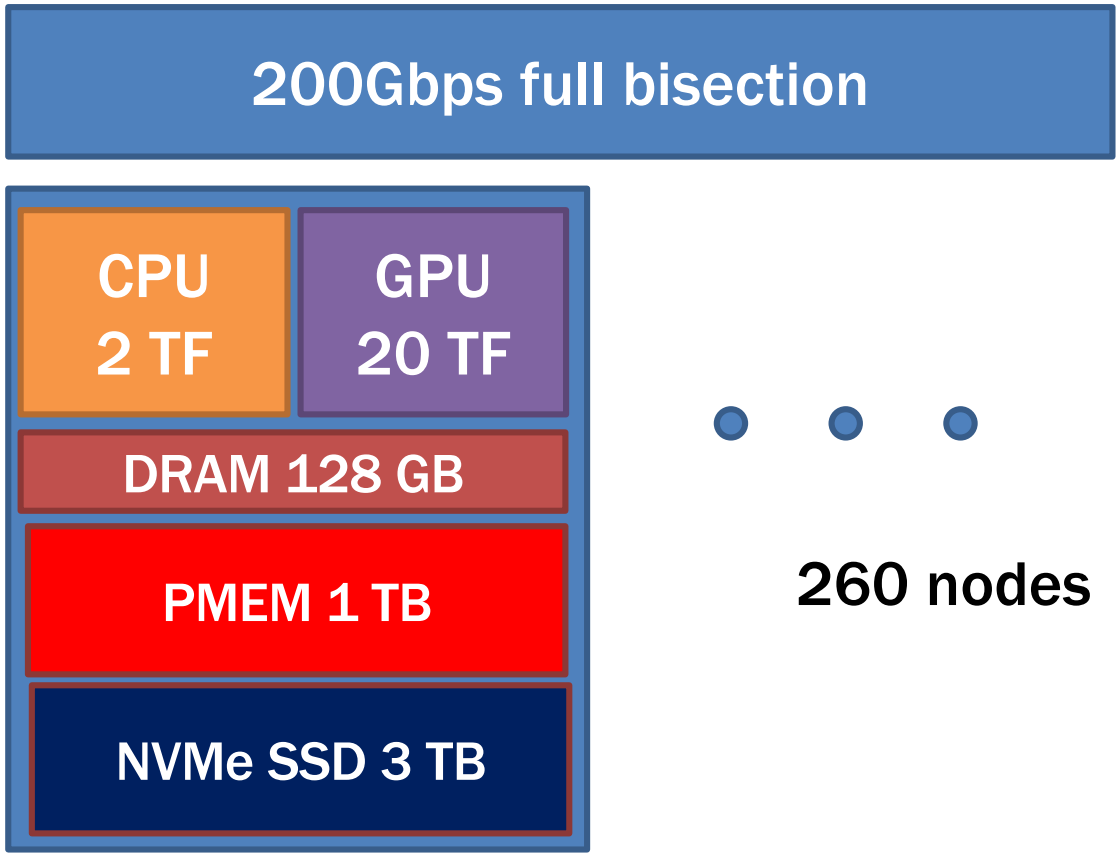
なぜ Big Memory が必要なのか？

- この8年でCPU性能は50xになったがメモリサイズは3.8xしか増えてない
- データサイエンス，AIにとっては深刻な問題
  - メモリサイズとストレージ性能が鍵
- Persistent Memoryの導入
  - memory modeによりメモリサイズ問題を，direct modeによりストレージ性能問題を解決



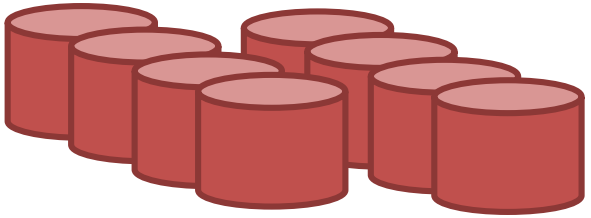
Cygnus-BDのイメージ（仕様変更の可能性あり）

- 2022/08運用開始（目標）
- 総合性能
  - 260 nodes, 5.6 PFlops, 300 TB
- ノード仕様（現状）
  - ~2 TFlops (CPU), ~20 TFlops (GPU)
  - 128 GB DRAM, 1 TB PMEM
  - 3 TB NVMe SSD
- 相互結合網
  - 200 Gbps full bisection
- 共有ファイルシステム（通常）
  - 9 PByte, 80 GB/s



PMEMを用いた ad hoc ファイルシステム

- node-local storageを利用した高速ファイルシステム
- 演算性能とI/O性能のギャップを埋める



- 筑波大CCSにおいて CHFS (Consistent Hash File System) ad hoc file system を開発中
  - metadataを持たず逐次処理部分を排除することで性能とスケーラビリティを向上
  - **CHFSについては2022/01の HPC Asia 2022 でも発表**
    O. Tatebe, et.al., "CHFS: Parallel Consistent Hashing File System for Node-local Persistent Memory", HPC Asia 2022, Jan. 14th, 2022.

おわりに

- 筑波大学CCSでは，その始まりであるCP-PACS (MPP)から，より柔軟性が高く対価格性能比に優れたPCクラスタに舵を切り，15年以上に渡り研究開発を続けてきた
- PCクラスタは構成の柔軟さ（design parameter空間の広さ）に加え，PCIeなどに支えられたアーキテクチャ拡張性が幅広いバリエーションを生んでいる
  ⇒ 我々のcodesigningにとって非常に強力
- single-core ⇒ multi-core ⇒ GPU ⇒ many-core ⇒ FPGA (with GPU) ⇒ PMEM と最先端のテクノロジを駆使して，単なるシステム調達だけでなく研究開発要素を含むスーパーコンピュータの導入をしてきた
- 超並列・大規模システムは電力性能的にMPP実装が避けられないが，小規模～大規模システムでは圧倒的にPCクラスタが有利
- アクセラレータの導入はPCクラスタの機能・性能を飛躍的に高めた
- CCSは今後も先端的スーパーコンピュータの開発を続けていく