筑波大学の取り組み

朴　泰祐
計算科学研究センター
筑波大学
taisuke@cs.tsukuba.ac.jp

アウトライン

- 筑波大学計算科学研究センター紹介
- 計算科学研究センターにおけるスパコンの歩み
- 現在のリソースと今後の展開
- 最先端共同HPC基盤施設に向けて

筑波大学計算科学研究センター紹介

筑波大学・東京大学・JCAHPC

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，生命科学
  - 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希少
- 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

計算科学研究センター（続き）

- 応用分野
  - **PP:** 素粒子物理研究部門
  - **ANP:** 宇宙・原子核物理研究部門
  - **QCM:** 量子物性研究部門
  - **LS:** 生命科学研究部門
  - **GES:** 地球環境研究部門
- システム分野
  - **HPC:** 高性能計算システム研究部門
  - **CI:** 計算情報学研究部門
- 一般的な「計算サービス／リソースセンター」ではない
  - 「顔・アプリケーション」の見えるユーザが対象
  - 計算科学に特化した大規模アプリケーション実行環境を全国共同利用施設として提供

計算科学研究センターにおけるスパコン開発・導入の歩み

CCSにおけるスパコン開発・運用の２つの流れ

- 開発系システム（PACS/PAXシリーズ）
  - センター独自の mission oriented なシステムを設計・開発し、「使い易い汎用システム」よりも「特定用途向けのチャレンジシステム」を目指す
  - （原則）計算機レンタル予算ではなく概算要求ベースの独自予算により研究を推進

- 計算リソースサービス系システム（VPP⇒T2K）
  - 「使い易い汎用システム」を市場から調達。「市場」はマシン本体だけでなくCPU、ネットワーク、メモリ等の部品を含む（クラスタ技術）

CCSにおける計算機導入の歴史

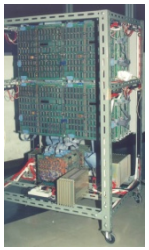
- 独自開発の先進的計算システム～PACS/PAXシリーズ～
  - mission oriented な計算と専門的プログラミングを対象とした（準）専用システムの開発
  - PACS (PAX) series
  - 新プログラム（CP-PACS）、特別教育研究経費（PACS-CS）、特別経費（HA-PACS）、科研特別推進（FIRST）等、文科省を中心とする個別予算ベース
- 一般のスパコン利用を対象とした汎用システム
  - 市場ベース（マシン自体だけでなく個々の部品・要素技術）のシステム
  - 筑波大学学術情報メディアセンターのスパコンサービスをCCSに移行
  - VPP500, VPP5000, T2K-Tsukuba

超並列計算機PAX(PACS)の開発の歴史

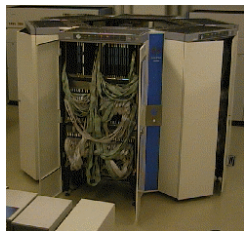
- 1977年に研究開始（星野・川合）
- 1978年に第一号機が完成
- 1996年のCP-PACSはTOP500第一位
- 2012年のHA-PACSは第8号機

1978
第1号機PACS-9

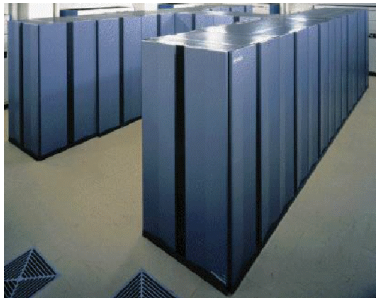
1980
第2号機PAXS-32



1989
第5号機QCDPAX



1996
世界最高速を達成した第6号機CP-PACS



2006
PACS-CS

完成年	名称	計算速度
1978年	PACS-9	7KFLOPS
1980年	PAXS-32	500KFLOPS
1983年	PAX-128	4MFLOPS
1984年	PAX-32J	3MFLOPS
1989年	QCDPAX	14GFLOPS
1996年	CP-PACS	614GFLOPS
2006年	PACS-CS	14.3TFLOPS
2012年	HA-PACS	800TFLOPS

- 計算科学者＋計算機工学者の共同開発による「実用的スパコン」
- Application-drivenな開発
- 持続的な開発による経験の蓄積

CP-PACS (1996年)

- 筑波大学計算物理学研究センター（CCSの前身）
- 筑波大学＋日立
- 1996年3月に1024PU完成
  ⇒ 10月に2048PUに
- 大学主導計算機として世界最高速となった貴重な例
- 計算物理学のための計算機
- ソフトウェアベクトル処理のために強化されたプロセッサ
- 2048 CPU, 614GFLOPS
- Linpack性能368GFLOPS
  1996.11 TOP500で１位

FIRST: 宇宙物理学のための「ヘテロクラスタ」(2005年)

256（16×16）nodes
512 CPU ＋
256 Blade-GRAPE

Total Performance = 38.5 Tflops
Host 3.5 Tflops
Blade-GRAPE 35 Tflops

Total Memory = 1.6TB

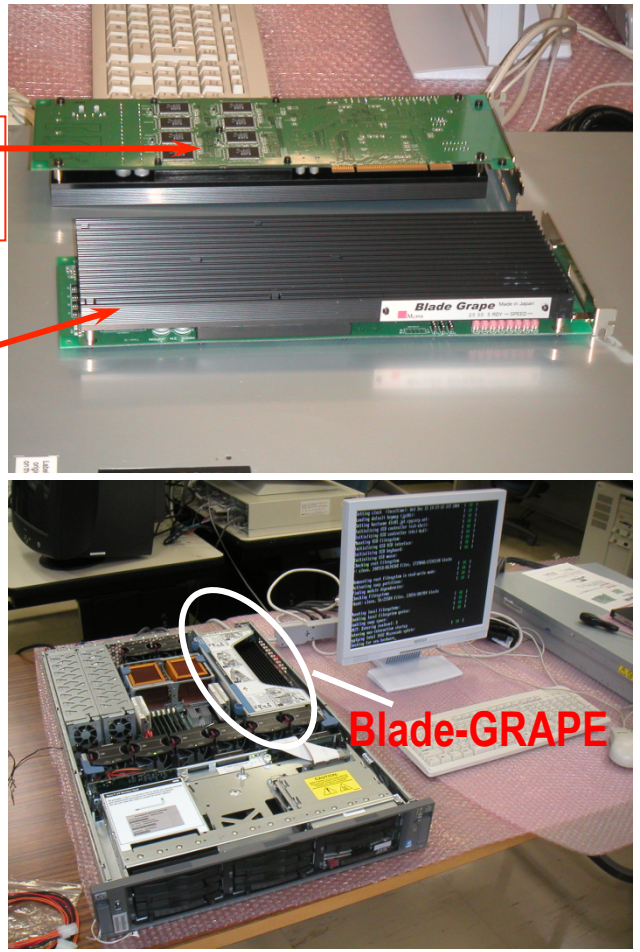
Total storage = 22TB (Gfarm)

- 全ノードにBlade Grape (重力計算アクセラレータ) を搭載した heterogeneous PC cluster
- ClearSpeed等が出始めた頃の heterogeneous computing の走り

Blade GRAPE

FIRSTクラスタのための重力計算アクセラレータ（ノード組み込み型）

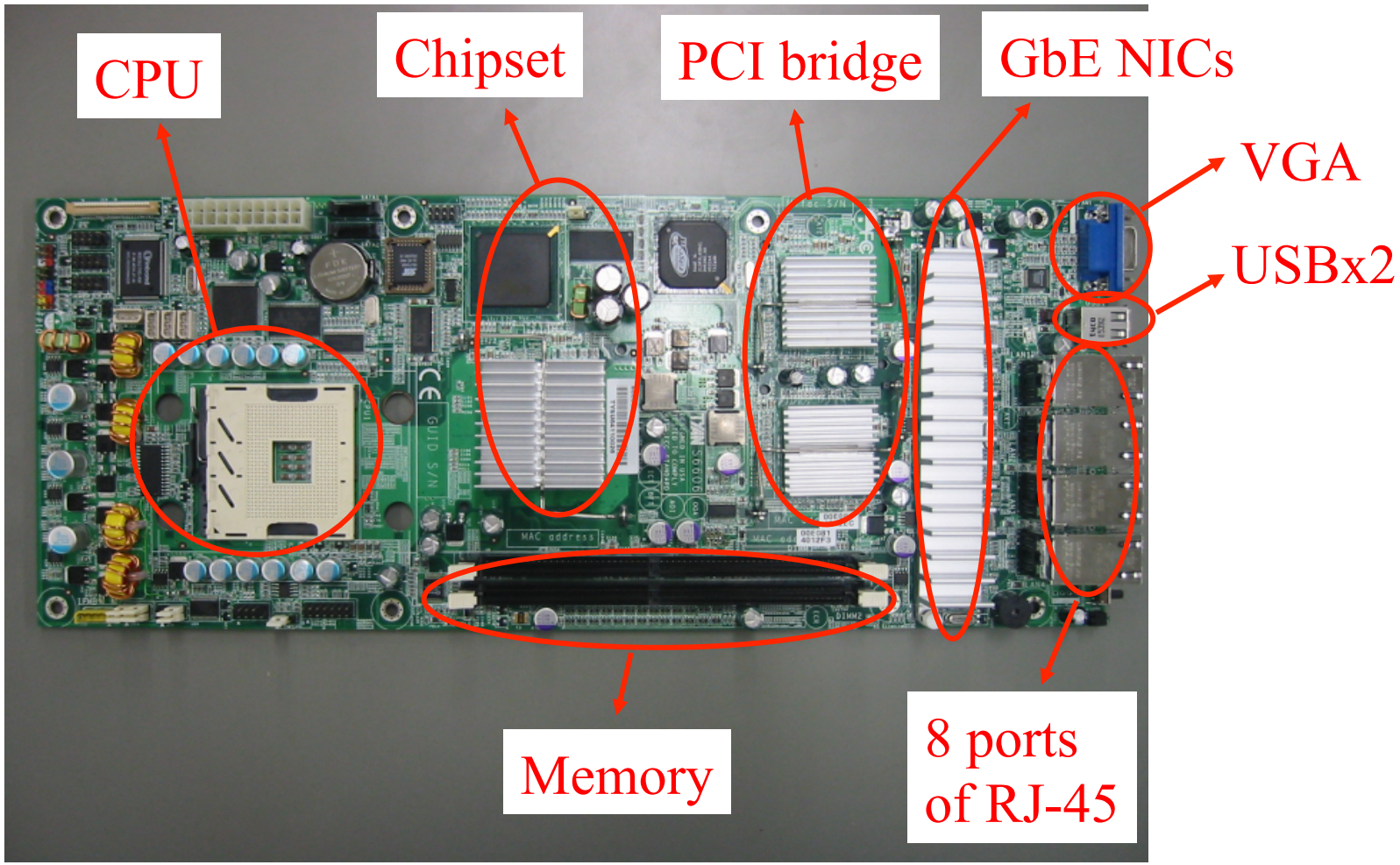
- Full size PCI card requiring 2-PCI slots
- 10 layers in a board
- 4 GRAPE6 chips = **136.8GFLOPS**
- electric power of 54W
- memory of 16MB (260K particles)
- Implemented by Hamamatsu Metrics Co.



PACS-CS: 「バンド幅重視」Cluster (2007年)

- PACS-CSの設計コンセプト
  - バンド幅を重視した設計 (memory & network)
  - コモディティ技術に基づくMPP-like system
- PACS-CSの特徴
  - 16x16x10=2560 node
    ⇒ 14.4 TFLOPS (peak)
    10.6 TFLOPS (Linpack)
    #34 in TOP500 (June 2006)
  - single core, single CPU / node
  - CPU FLOPS：Memory B/S：Network B/S
    = 5.6 GFLOPS：6.4 GB/s：0.75GB/s
  - ハイパクロスバ網により様々な物理ドメイン形状を計算対象に
  - GbEthernet x 6 を３次元展開したネットワーク

PACS-CS マザーボード写真



現在のリソース・研究と今後の展開

T2K-Tsukuba: 大規模汎用クラスタ (2008年)

Open Supercomputer T2K
Open Supercomputer Alliance

#20 at TOP500 on June 2008 (Linpack: 76.46 TFLOPS)
⇒ H26.2末に運用終了予定

計算ノードとファイルサーバ

Open Supercomputer T2K
Open Supercomputer Alliance

Computation node（70racks）
(Appro XtremeServer-X3)

648 node (quad-core x 4socket / node)
Opteron “Barcelona” 8356 CPU
2.3GHz x 4FLOP/c x 4core x 4socket
= 147.2 GFLOPS / node
= 95.3 TFLOPS / system
20.8 TB memory / system

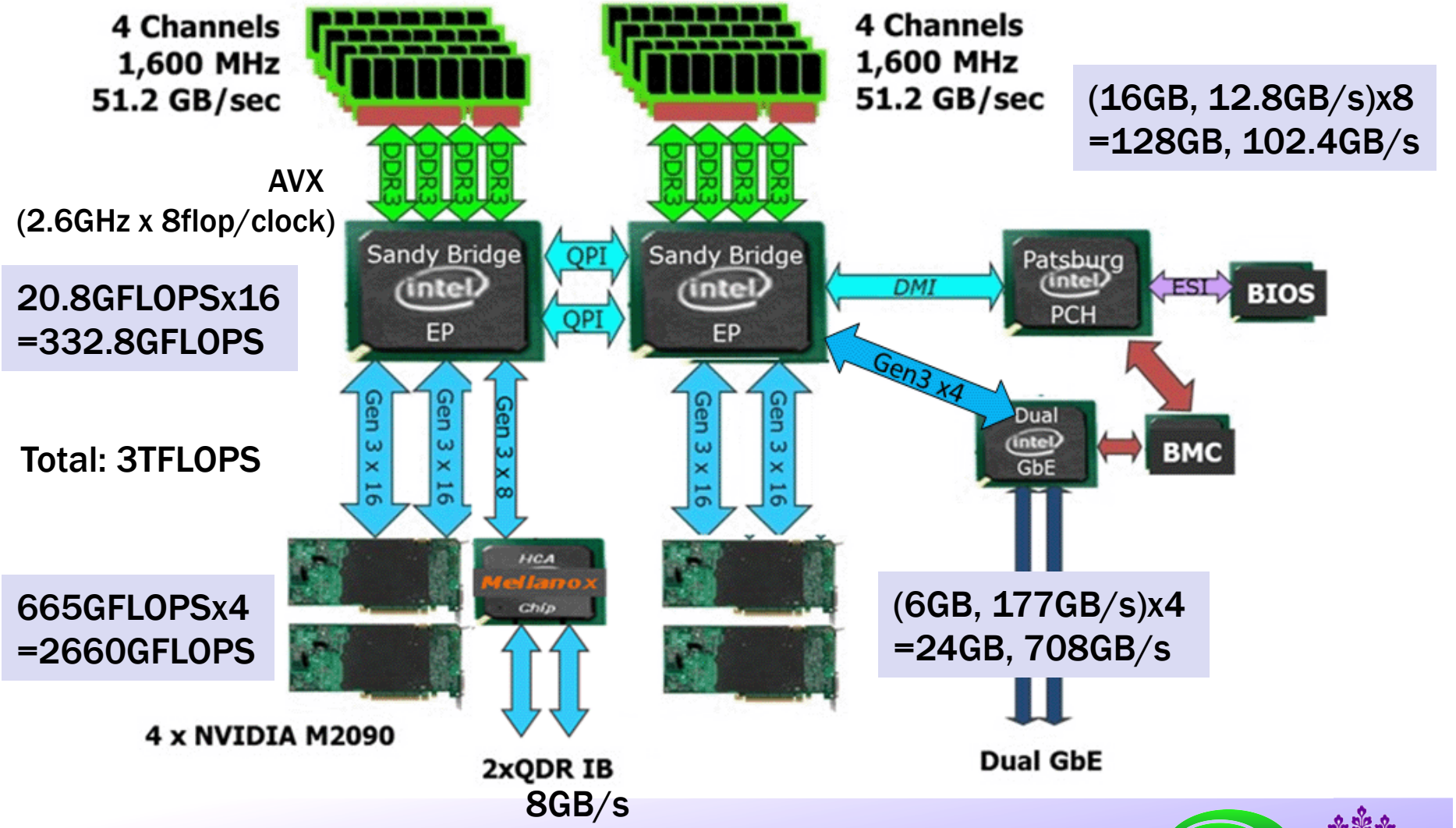
800 TB (physical 1PB) RAID-6
Luster cluster file system
Infiniband x 2
Dual MDS and OSS config.
⇒ high reliability

File server (disk array only)
(DDN S2A9550)

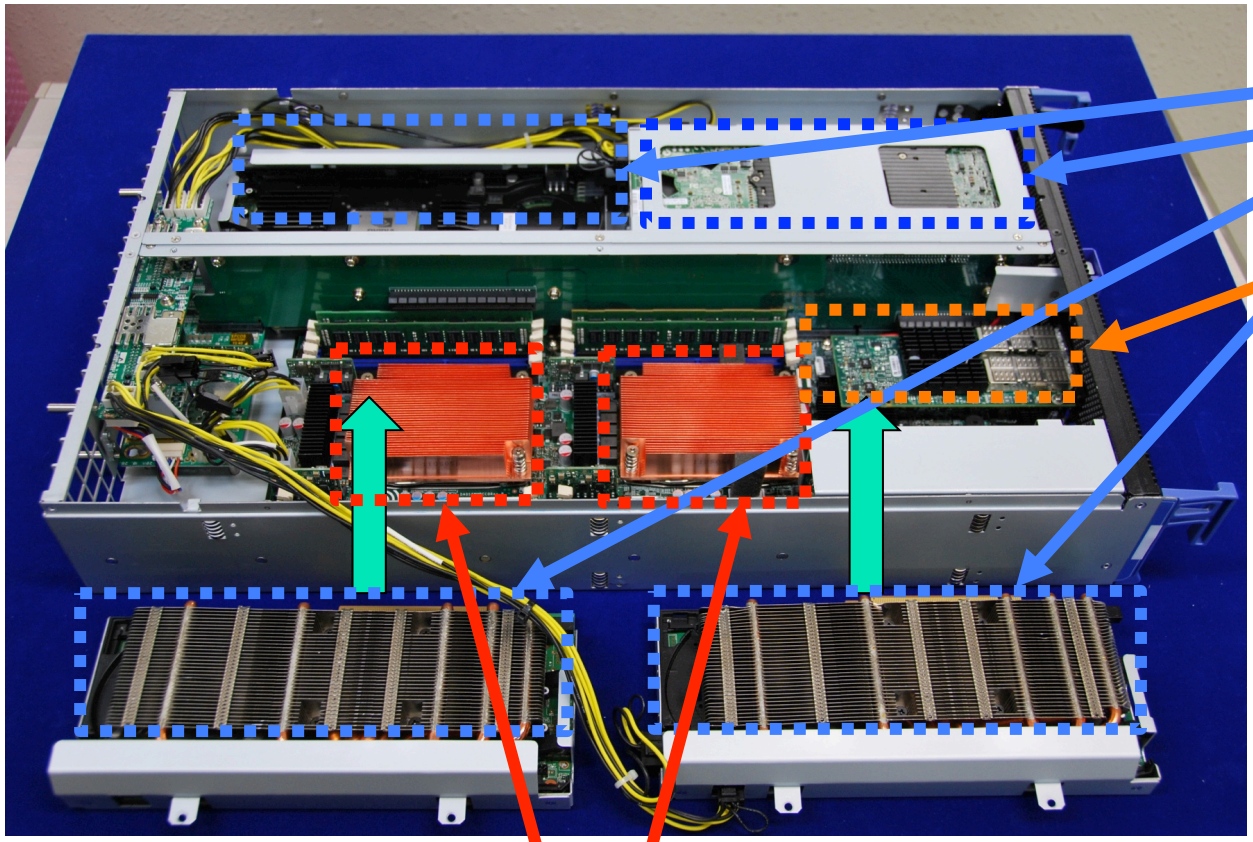
HA-PACS base cluster (2012年)

- 268 node, 4 GPU/node, 802 TFLOPS (peak), 421.6 TFLOPS (Linpack) #39 on 2012.6
- 計算ノード: Appro GreenBlade 8240
- ネットワーク: InfiniBand QDR x 2rail, Mellanox ConnectX3
- ファイルサーバ: DDN SFA10000 (500TB, RAID6, Lustre)

HA-PACS: base cluster (computation node)



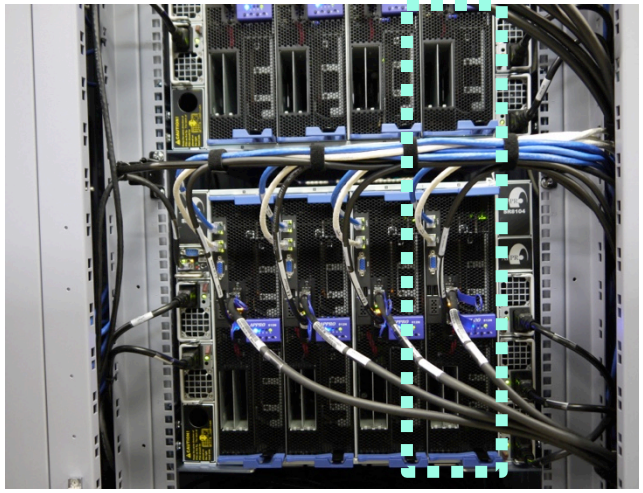
HA-PACS: base cluster (computation node)



GPU (M2090) x 4

InfiniBand QDRx2

CPU (SandyBridge-EP) x 2



計算ノードシャーシ（4 node）

現在のリソース利用

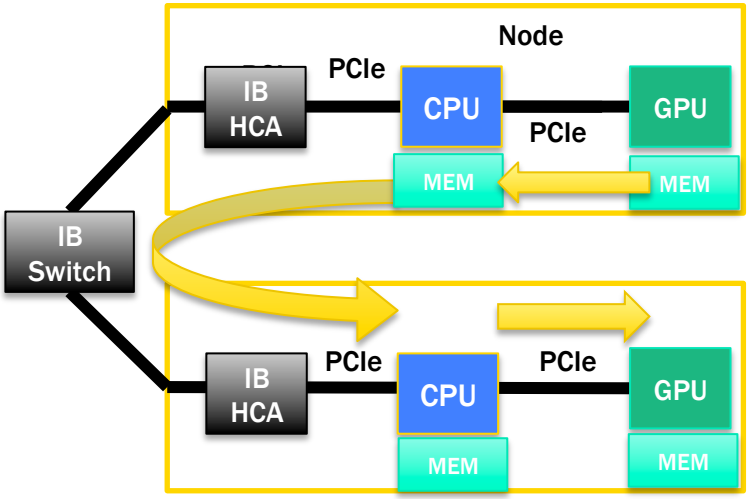
- T2K-Tsukuba
  - 学際共同利用：約55%
  - HPCI：約20%
  - 大規模一般利用（有償）：約20%
  - 計算基礎科学連携拠点：約5%
- HA-PACS
  - 学際共同利用：100%
- 利用状況（H25.4~6月平均ジョブ稼働率）
  - T2K-Tsukuba: 69.9%　(システム稼働率：96.6%)
  - HA-PACS: 87.4%　　(システム稼働率：96.6%)

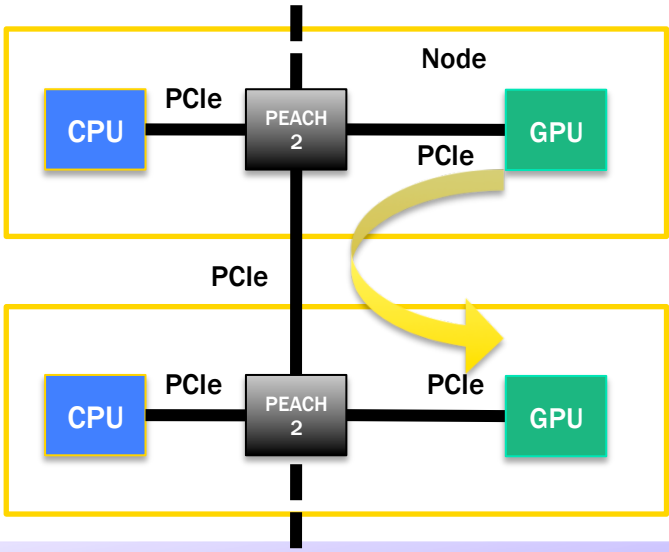
HA-PACS/TCA

- TCA: Tightly Coupled Accelerator
  - アクセラレータ間直接通信ネットワーク機構 (GPUs)
  - PCIeをアクセラレータ間直接通信リンクとして利用
    - 現在のアクセラレータやその他のI/OデバイスはPCIeによりCPUに結合されている
    - インテリジェントなPCIeスイッチと外部リンクにより理論的にはあらゆるデバイスを結合可能
  - このコンセプトの下でノードを跨がるアクセラレータ間並列結合システムを構築し強スケーリング可能なアプリケーションを実行可能な環境を提供
- HA-PACS/TCA
  - TCA機構の有効性を実証するための実験クラスタとしてHA-PACS base cluster を拡張し、TCA機構を組み込む

HA-PACS/TCA (Tightly Coupled Accelerator)

- **True GPU-direct**
  - 現在のGPUクラスタではGPU間通信に多数のホップが必要 (3-5回のメモリコピー)
  - 強スケーリングを実現する大規模GPUクラスタのためには超低レイテンシで高バンド幅を持つGPU間直接リンクが必要
- **PEACH2**
  - TCAをGPU間結合で実現するFPGAベースのPCIeスイッチ
  - CPUを経由することなくノードを跨ぐGPU間直接通信が可能

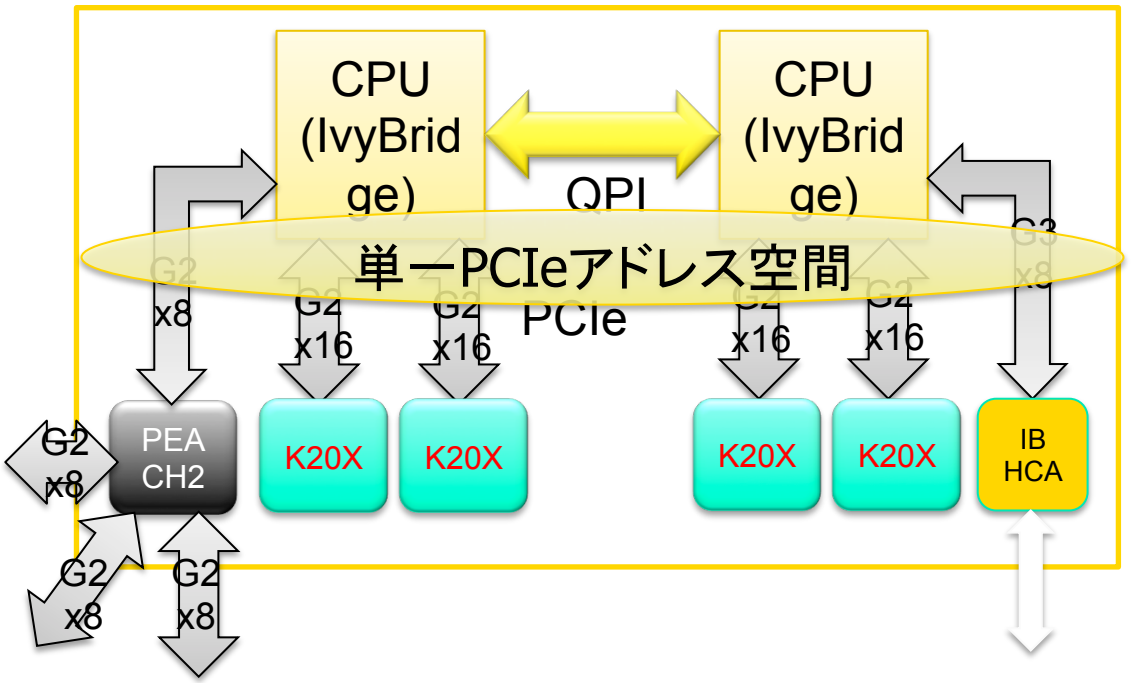




HA-PACS/TCAノード構成

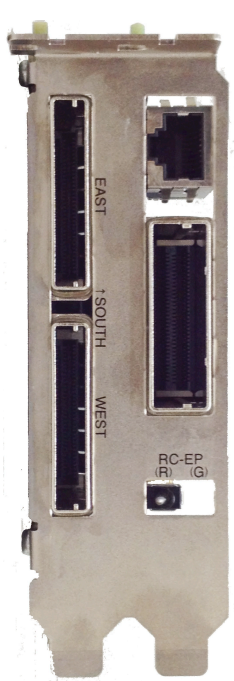
- CPUからのGPUの見え方は均一
- PEACH2から全てのGPUにアクセス可能
  - Kepler + CUDA 5.0 “GPUDirect Support for RDMA” によりGPUメモリをアクセス
- 他の3ノードと接続

- PEACH2を除けばHA-PACS base clusterと同等の構成(CPUは異なる)
  - CPUソケット内蔵のPCIe（80レーン）を全て使用

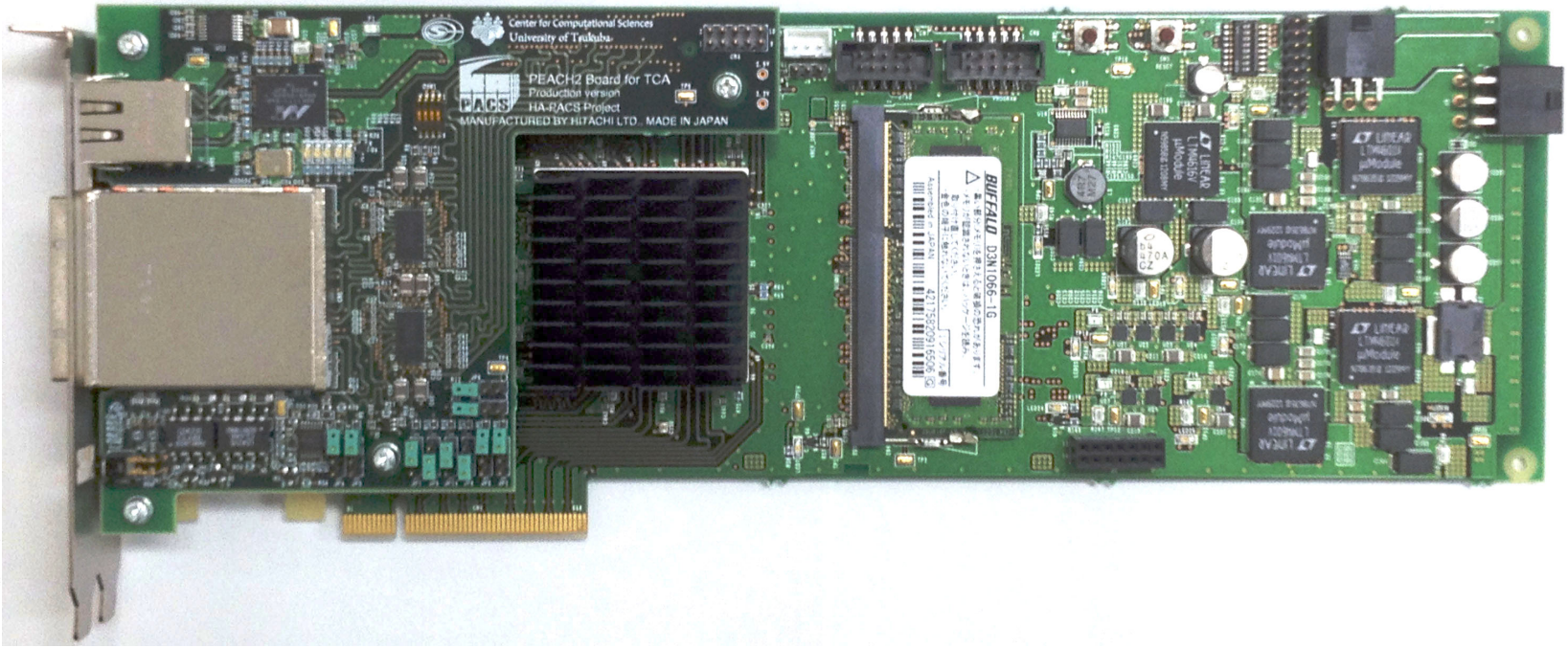


PEACH2 board

- PCI Express Gen2 x8 peripheral board
  - Compatible with PCIe Spec.

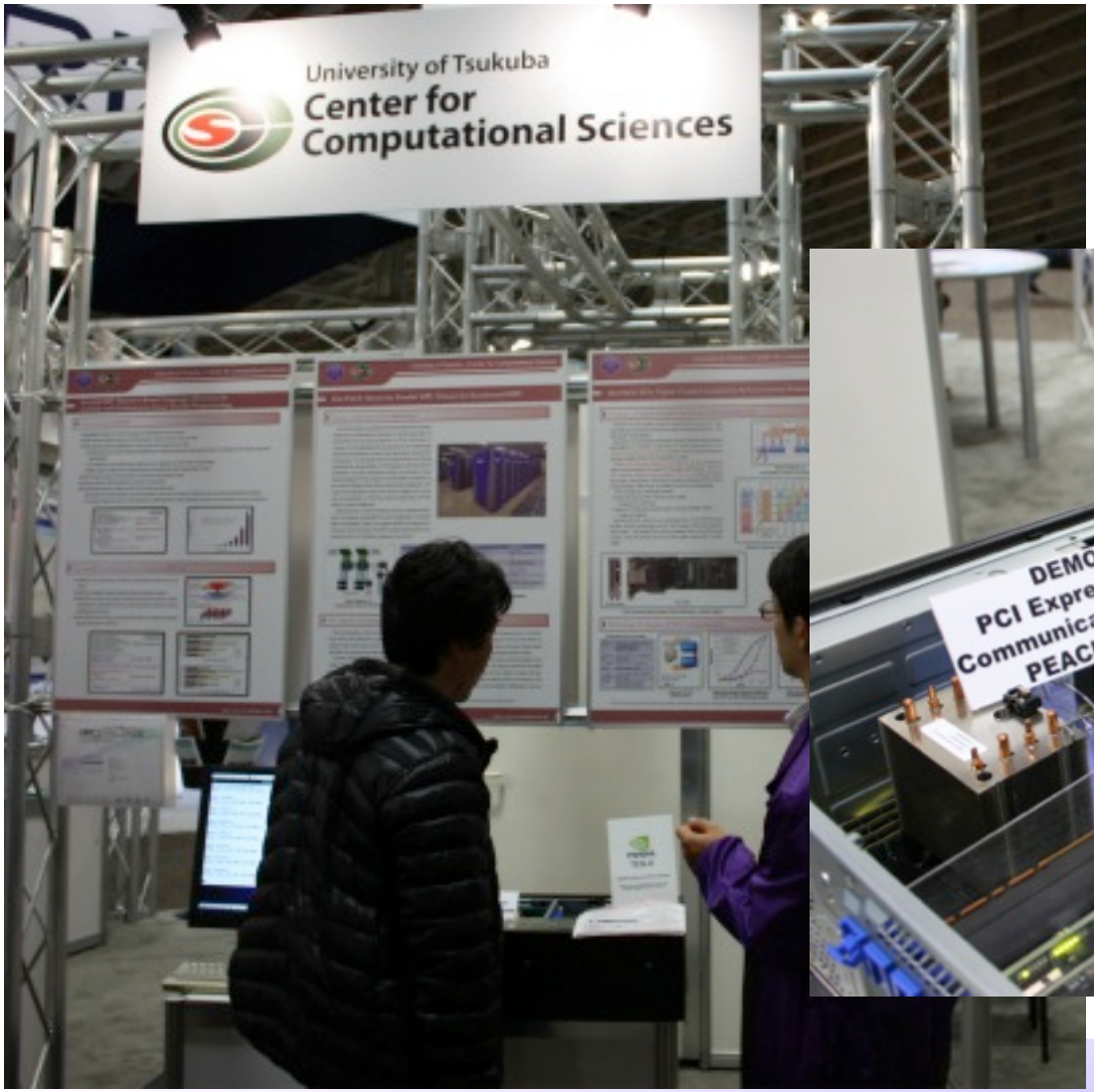


Side View



Top View

Demonstration at SC12

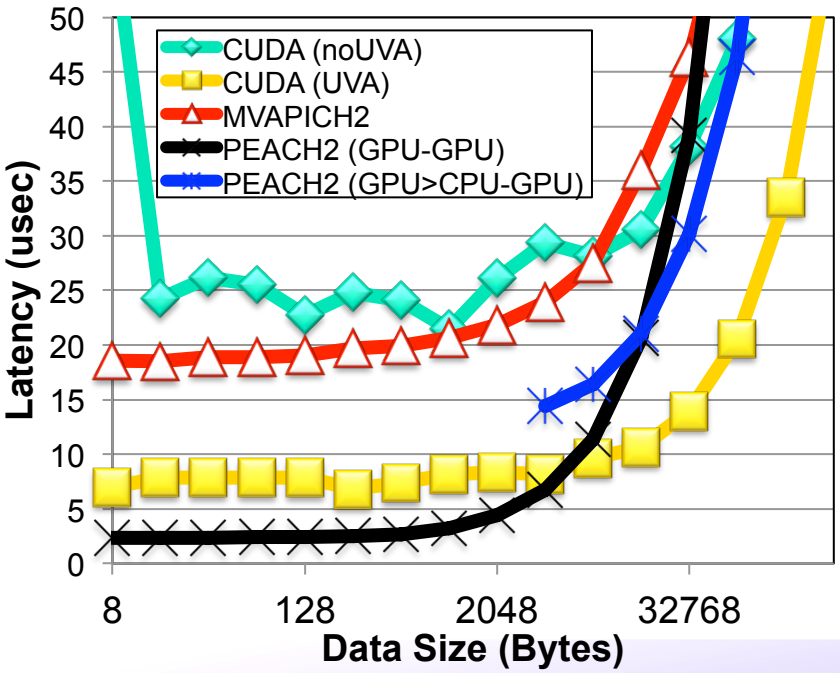


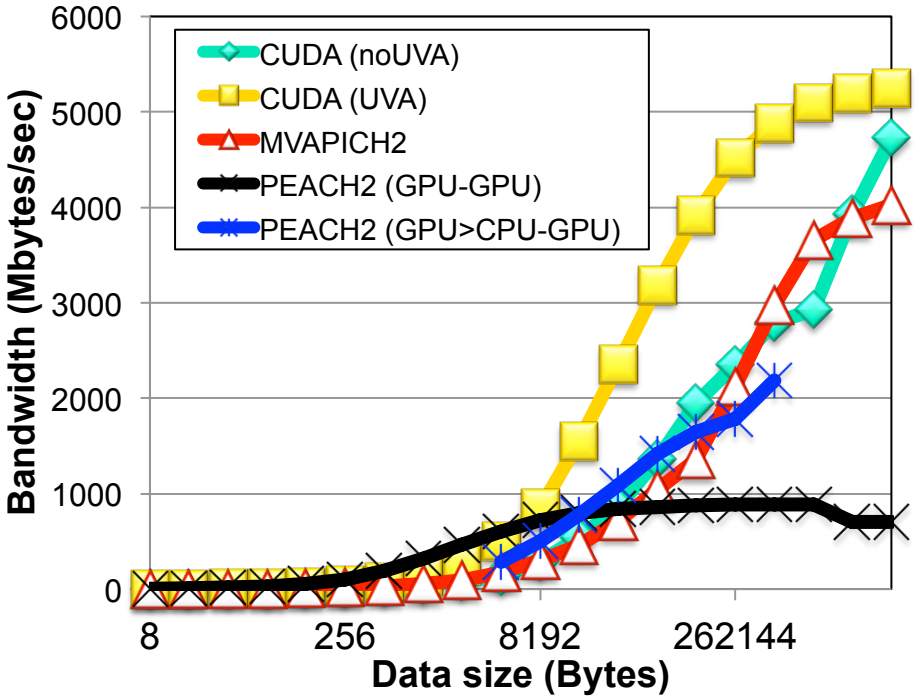
TCA通信とMPI通信の比較

別ノードのGPUデバイスメモリ間通信

- PEACH2: TCAによる通信
- CUDA: ノード内GPU間のcudaMemcpy()
  - No-UVA: without Unified Virtual Address
  - UVA: with Unified Virtual Address
- MVAPICH2: CUDA device間直接通信ON
  - InfiniBand FDR10
    (Mellanox Connect-X3 directly connected by cable)

- PEACH2の性能はMVAPICH2より高く、さらにノード内GPU間CUDA通信よりも高い





HA-PACS/TCA実験システム導入

- システム
  - ノード：Cray-3623G4-SM (Supermicro社M/B)
  - CPU: Intel Xeon E5-2680 v2 (IvyBridge)
    2.8GHz 10core * 2
  - GPU: NVIDIA K20X * 4
  - ノード数：64
  - ネットワーク：InfiniBand QDR * 2 rail
  - 総性能：CPU部= 224GFLOPS * 64 = 28.6TFLOPS
    GPU部= 1.31TFLOPS * 4 * 64 = 335.4TFLOPS
    合計 364TFLOPS
- HA-PACS base clusterとの結合
  - TCA部をbase clusterとInfiniBand QDR * 2 を40ポートで結合
  - **システム総ピーク性能は1.17PFLOPSに**
  - 2013年10月末インストール

Accelerated Computing に向けての研究

- 今後のエクサスケールコンピューティング、EFLOPSコンピュータの実現に向け、演算加速機構は重要な役割を持つ
  - 現実的な範囲でEFLOPSを実現するためのEFLOPS/W達成
  - 大規模並列環境において演算加速装置と通信機構は一体化される必要がある
  - 大規模演算加速装置・通信機構環境における新たなアプリケーション／アルゴリズムの開発が必要
  - 文部科学省 Feasibility Study において筑波大を中心とするグループは上記研究を推進・提唱
- CCS独自資源としては引き続き accelerated computing を軸としたマシン導入・アプリ開発を継続する予定
  ⇔ JCAHPCにおける汎用システム路線

最先端共同HPC基盤施設に向けての取り組み

大規模メニーコア実験システム（調達中）

- 最先端共同HPC基盤施設のスパコンとして想定されているメニーコア・アーキテクチャに基づくプロセッサ上での大規模計算科学のための実験システム
- メニーコア・プロセッサを多数用いた並列クラスタを導入し、アプリケーション開発を行う
- 東京大学で開発中のメニーコア・プロセッサ向けOSの実験と実アプリケーションでの評価を行う
- HA-PACSで開発中の演算加速器（GPU）向けアプリケーションの移植・チューニングの研究基盤
- 筑波大学で開発中のPGAS言語（XcalableMP）のメニーコア・アーキテクチャ向け実装の研究基盤
- さらにTCAをメニーコア・プロセッサ・ベースのアクセラレータ技術にも応用

大規模メニーコア実験システムの仕様（抜粋）

- ノード構成・性能
  - メニーコア・アーキテクチャに基づく演算加速装置を持つPCクラスタ
    ⇒メニーコア部の性能 > 2TFLOPS
  - 一定のマルチコアを持つCPUをホストとする
    ⇒マルチコア部の性能 > 400GFLOPS
  - メモリバンド幅 > 119 GByte/s
    メモリ容量 > 64 GiByte
- 並列構成
  - InfiniBand FDR以上のネットワークリンク
  - バイセクションバンド幅を保証
- システム全体性能
  - メニーコア部総性能 > 565TFLOPS
  - マルチコア部総性能 > 112TFLOPS

大規模メニーコア実験システムの利用

- 研究
  - JCAHPCシステムに向けたアプリケーション開発
  - 筑波大学各種システムソフトプロジェクトの研究基盤
  - 東京大学との連携研究の研究基盤

- サービス
  - 現在のT2K-Tsukuba（H26.2月末でshut down）の後継として、JCAHPCシステム運用開始までの間の各種計算リソースサービス
    - 学際共同利用（HA-PACSと平行して）
    - HPCI（汎用プロセッサ部のみを提供の予定）
    - 一般利用（有償）

メニーコア・プロセッサについて

- 現状のメニーコア・プロセッサの技術的課題
  - 数十コアが搭載されているが、個々のコアの性能は最先端汎用CPUコアに比べて弱い
  - 多数のコアによる並列処理がベースで、throughput computing的な利用には向いているが、逐次部分が多いと性能ボトルネックが顕在化する
  - 演算加速装置として使った場合、ホストとの通信・ノード間通信にどう対応するか

- 演算加速装置として
  - GPUとの比較、今後のアクセラレータ系計算に対するプログラミングをどうするか
  - HA-PACSと大規模メニーコア実験システムの両方を運用することにより、様々な実験に対応

導入計画

- 現在調達中（最終仕様確定、入札準備中）
- H25.9中に入札、10月末に開札・契約
- H26.3末にインストール（3月と4月に分けて導入）
- 各種サービスをH26.4月より開始する予定
- 設置場所：T2K-Tsukubaを撤去した跡地
- ファイルサービスの継続性：T2K-Tsukubaに新ファイルサーバを設置し、データ移行をスムーズに

JCAHPCにおける筑波大の役割

- T2K-Tsukuba ⇒ 大規模メニーコア実験システムに引き継がれるスパコンリソースサービスの継続
  - 全システムの約1/3が筑波大持ち分（予定）
  - 学際共同利用
  - HPCI
  - 計算基礎科学連携拠点
  - その他（有償サービス）
- 超大規模システムにおける各種研究の展開
  - （全システム占有時を含めた）先進的計算科学研究・アプリケーション開発・実行
  - メニーコア・アーキテクチャを中心とする言語・ライブラリの開発

まとめ

- 筑波大学計算科学研究センターは前身の計算物理学研究センター発足から20年余り、計算機工学と計算科学の融合研究を継続
- CP-PACSを始めとする数々の成功体験の上で、システム開発と応用プログラム開発を平行して進め、多くの世界的研究成果を達成
- 全国共同利用施設として学際共同利用・HPCI等にけるサービスを提供、主として大規模計算に貢献
- 最先端共同HPC基盤施設における超大規模システム上で研究を展開することにより、さらなる発展を