演算加速クラスタによる計算科学の推進

朴　泰祐

筑波大学計算科学研究センター

計算機システム運用委員長／副センター長

http://www.ccs.tsukuba.ac.jp

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

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
バンド幅重視クラスタ
PACS-CS

2012~2013
GPU演算加速クラスタ
HA-PACS

完成年	名称	性能
1978年	PACS-9	7 KFLOPS
1980年	PACS-32	500 KFLOPS
1983年	PAX-128	4 MFLOPS
1984年	PAX-32J	3 MFLOPS
1989年	QCDPAX	14 GFLOPS
1996年	CP-PACS	614 GFLOPS
2006年	PACS-CS	14.3 TFLOPS
2012~13年	HA-PACS	1.166 PFLOPS
2014年	COMA (PACS-IX)	1.001 PFLOPS

- 計算科学者＋計算機工学者の共同開発による「実行性能重視スパコン」
- Application-drivenな開発
- 持続的な開発による経験の蓄積

この他：科研費特別推進研究によるハイブリッドクラスタ FIRST

CCSにおいて現在運用中の2系列のスパコン

- HA-PACS (PACS-VIII)
  - GPU cluster、一部に独自開発のGPU間直接通信ネットワーク(TCA)を実装
  - PFLOPSクラスのGPUクラスタにより accelerated computing によるコード開発とプロダクトランを実施
  - novice user よりも professional user をターゲットとした先進的マシン

- COMA (PACS-IX)
  - Many Core cluster
  - PFLOPSクラスの many core architecture クラスタにより many core processor の性能特性を理解しつつ次世代の many core 向けコード開発とプロダクトランを実施
  - 筑波大・東大で共同設置した「最先端共同HPC基盤施設 (JCAHPC)」において2015年度に導入予定の many core システムの先行研究と各種テスト実施
  - 汎用並列処理向け、比較的広いユーザ層を対象

筑波大最初の演算加速クラスタFIRST

- 専用アクセラレータ (Blade-GRAPE)搭載クラスタ
- Hewlett Packard社
- 2005年完成
- 各ノードをdual socket Xeon + GRAPE-6 board (Blade-GRAPE) で構成したヘテロクラスタ
- 計算宇宙物理学用
- 256 nodes/512 cores + 1024 GRAPE-6 chip
- Host: 3.1 TFLOPS
  Blade-GRAPE: 33TFLOPS

HA-PACS計画

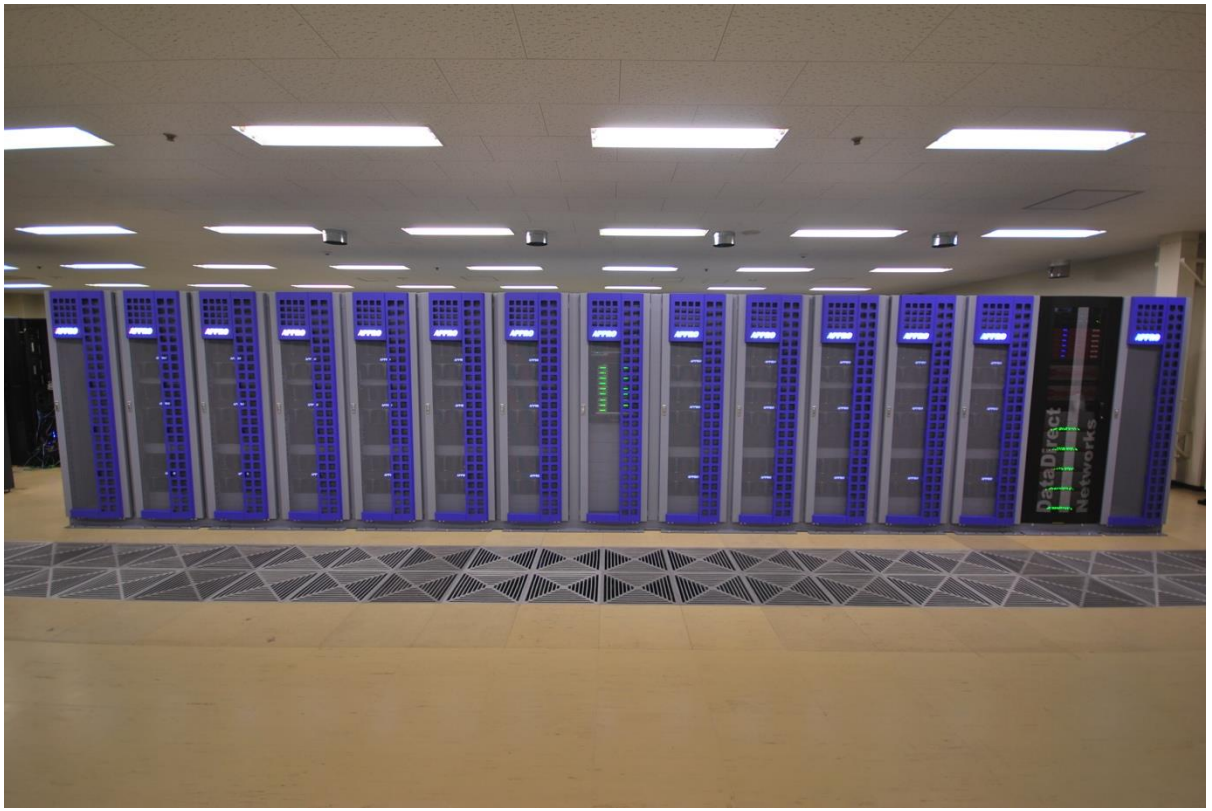
- 今後の(プロ向け)PC Cluster型スパコンに関する課題
  - performance / space (footprint)
  - performance / power
  - programming
- アクセラレータに基づくスパコン導入とプログラム開発を推進
  - HA-PACS (Highly Accelerated Parallel Advanced system for Computational Sciences)
  - HA-PACS Base Cluster: コモディティによる高密度・高性能GPUクラスタ
  - HA-PACS/TCA: TCA技術(後述)に基づく次世代 accelerated PC clusterの研究

HA-PACS Base Cluster

- Appro International 社
- CCSとして初めてGPUを本格採用したクラスタ
- 268ノード（2 Xeon E5-2670 +
  4 Fermi M2090)
- Mellanox IniniBand QDR x 2rail, Fat Tree
- File Server: DDN
  500 TB (RAID6+Lustre)
- 802 TFLOPS
  (HPL: 421 TFLOPS)
- TOP500 #41
  （国内 #5）

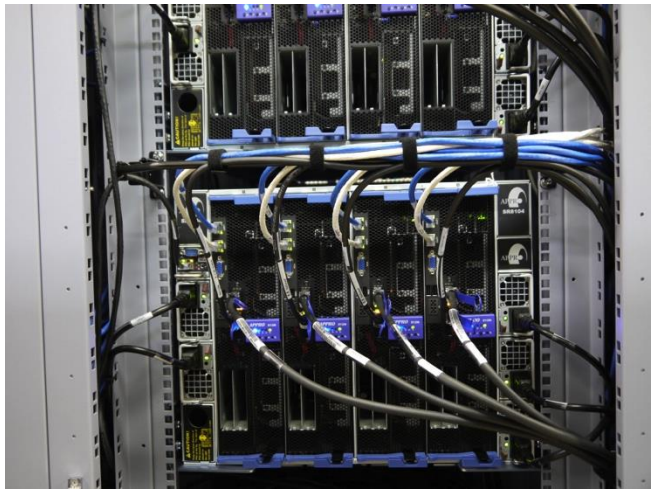
HA-PACS Base Cluster

Front view



Side view

HA-PACS Base Cluster

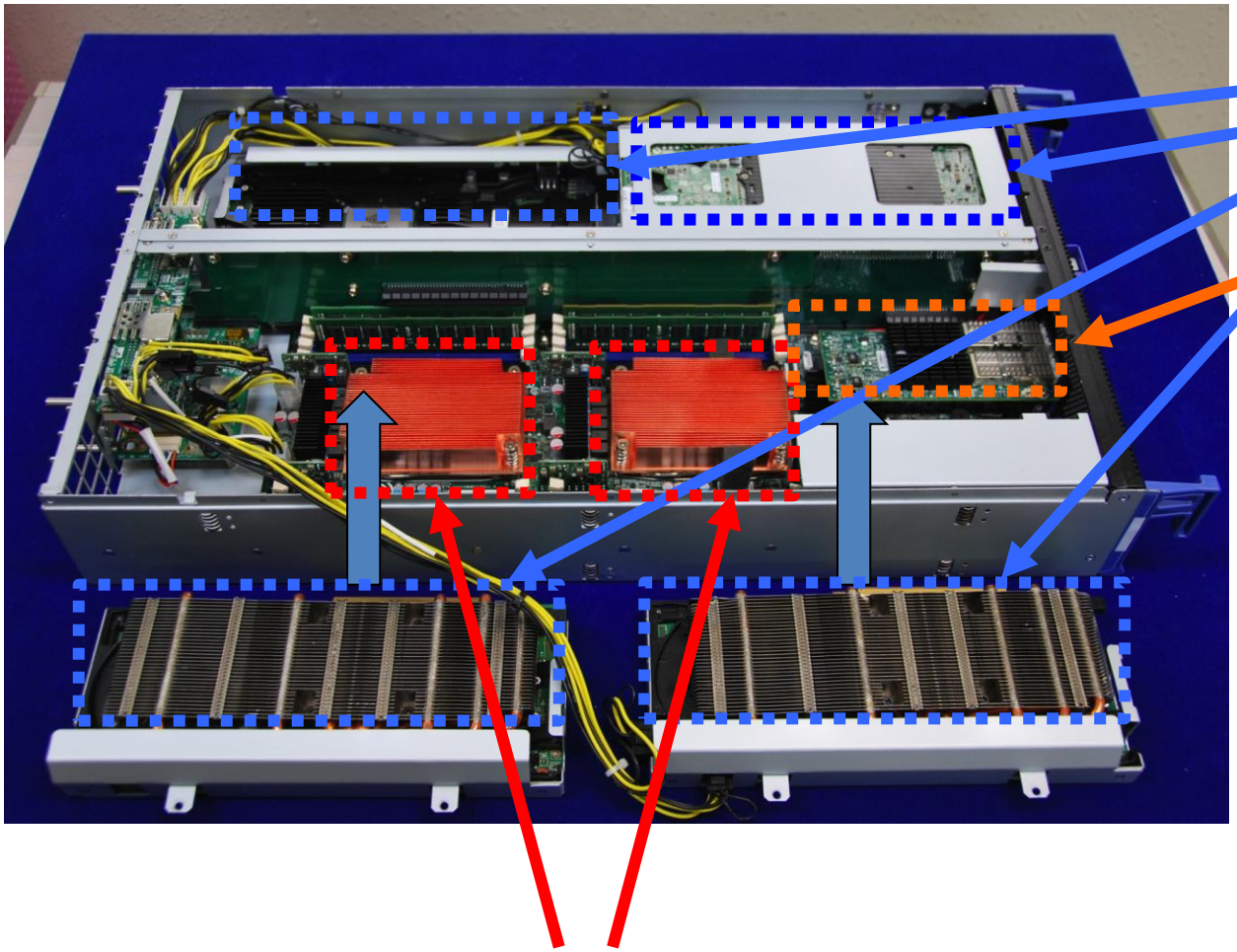


Rear view of one blade chassis with 4 blades

Front view of 3 blade chassis

Rear view of Infiniband switch and cables
(yellow=fibre, black=copper)

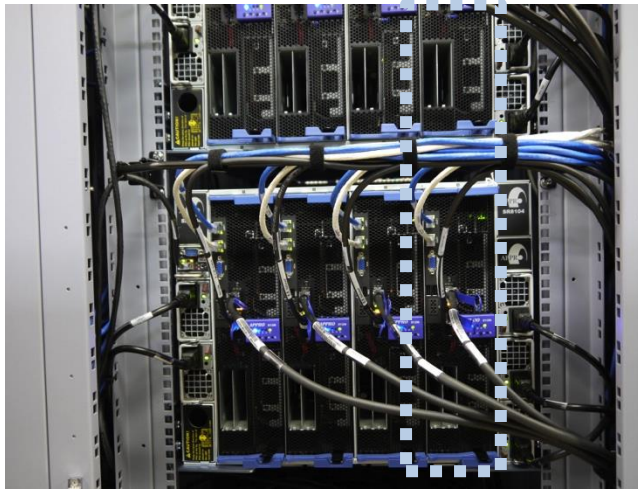
HA-PACS: base cluster (computation node)



GPU (M2090) x 4

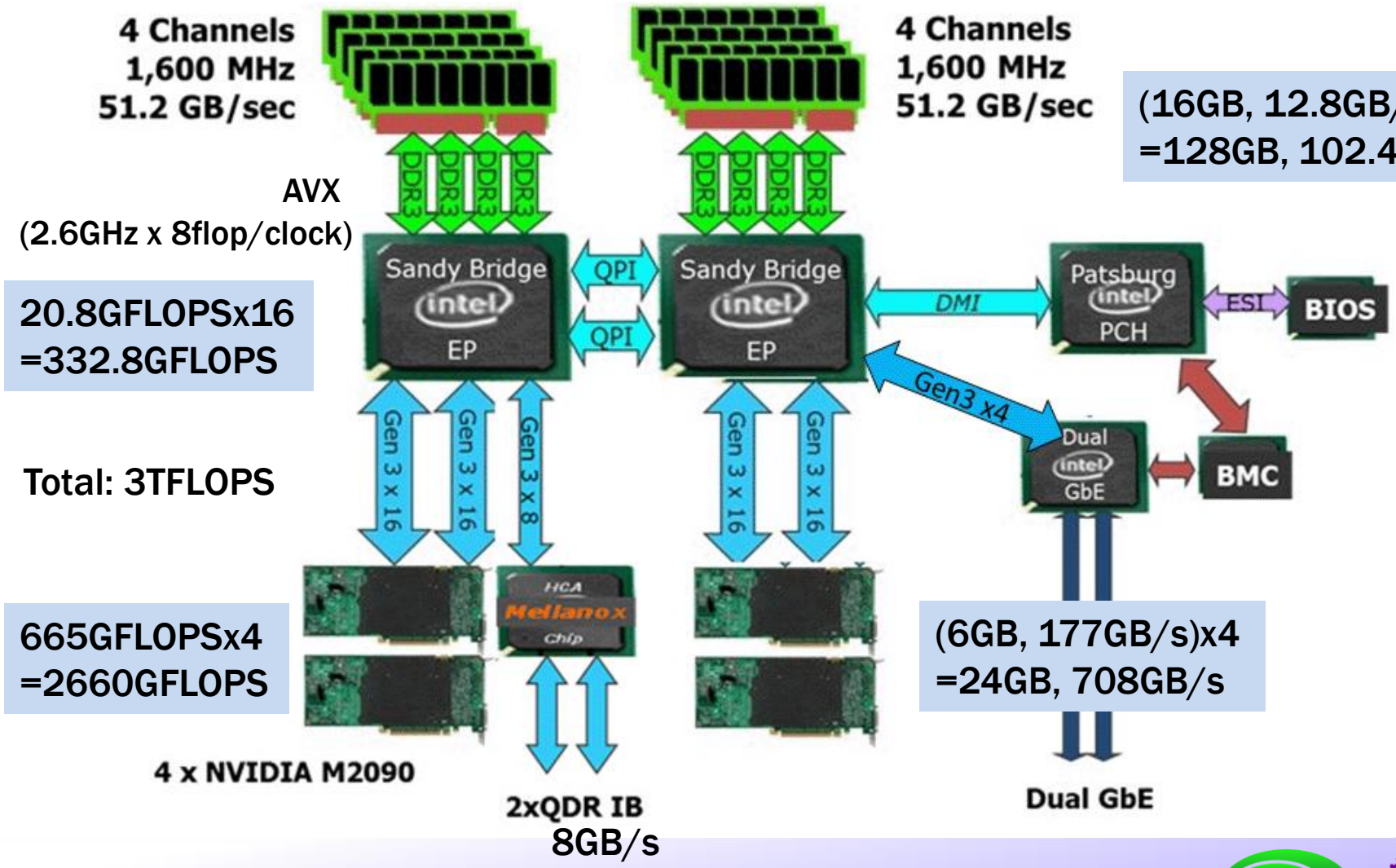
InfiniBand QDRx2

CPU (SandyBridge-EP) x 2



計算ノードシャーシ（4 node）

Computation node of Base Cluster



HA-PACS Base Cluster + TCA
(2013/11稼働開始)

- HA-PACS Base Cluster = 2.99 TFlops x 268 node = 802 TFlops
- HA-PACS/TCA = 5.69 TFlops x 64 node = 364 TFlops (Green500 #3)
- TOTAL: 1.166 PFlops

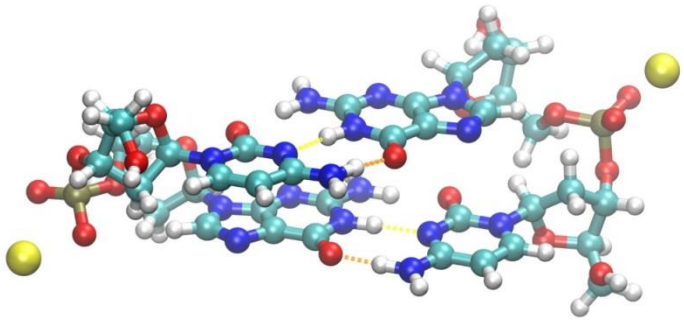
Linpack性能（TOP500, Green500）

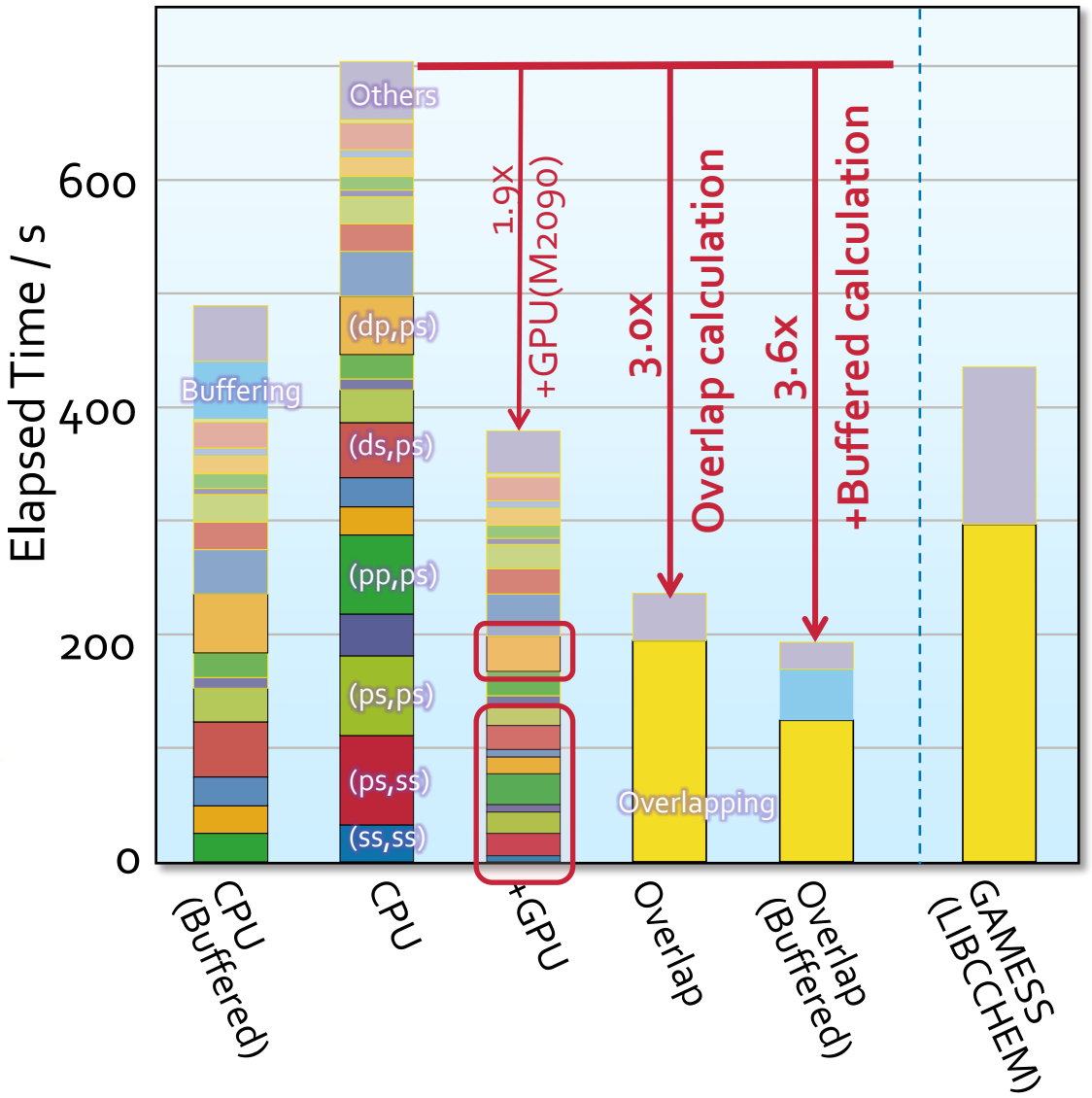
- TOP500 (HPL)
  - 421.6 TFLOPS （世界第41位、2012年6月）
  - GPUクラスタとしては世界第7位
  - 計算効率（対ピーク性能）: 54.2%

- Green500
  - 1151.91 MFLOPS/W （世界第24位、2012年6月）
  - GPUクラスタとしては世界第3位
  - 大規模GPUクラスタ（TOP50以内）としては世界第1位
    ⇒ SandyBridge-EPにより x80 lanes の PCIe gen2 が利用可能、
    4台の M2090 をバンド幅律速なしに接続

HA-PACSの成果例：HF計算の性能評価

- Model DNA (CG)2
  - HF/6-31G(d)
  - 126 atom, 1,208 AO
  - 14 SCF iterations
- HA-PACS 1node
  - 16 CPU cores
    - Intel SandyBridge-E5, 2.6GHz
  - 4 GPU(NVIDIA M2090)
- Software
  - OpenFMO
  - GAMESS
    - Version: 1 MAY 2013 (R1)
    - GPU support (LIBCCHEM)





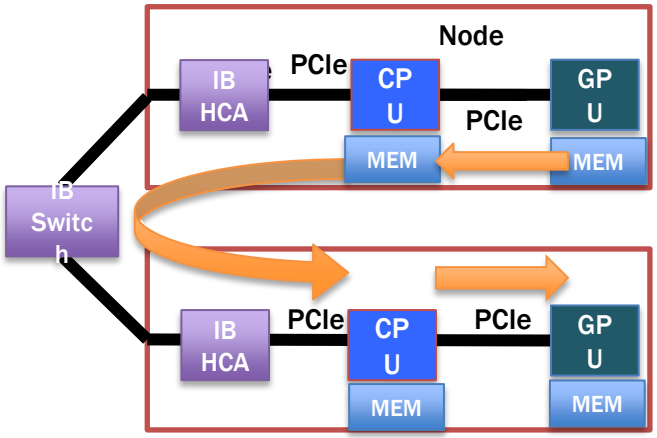
HA-PACSにおける新技術開拓

- ポストペタスケール～エクサスケールシステムの有力な候補として演算加速装置（Accelerating Devices: AD）を多数用いた超並列システムを想定
  - AD: GPU, GRAPE-DR, MIC, FPGA, etc.
- 超並列AD環境におけるHPCの問題点
  - AD間の結合：AD性能の向上に伴う通信ボトルネックの顕在化
  - 記憶装置：限られたメモリサイズ
  - ADとホストCPUの通信：バンド幅、協調計算モデル⇒複雑化
  - プログラミング：直交する多くのパラダイム⇒生産性の低下
- エクサ時代の Strong Scaling 問題
  - コア当たりメモリ容量の縮小により weak scaling に依存した性能向上が限界に
  - strong scaling では通信レイテンシが本質的問題となる

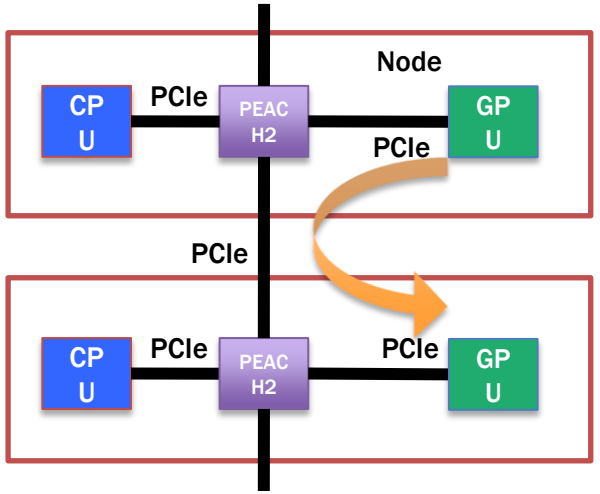
HA-PACS計画の礎：「エクサスケール計算技術開拓による先端学際計算科学教育研究拠点の充実」（代表：佐藤三久、H23～H25）
⇒ JST-CREST：「ポストペタスケール時代に向けた演算加速機構・通信機構統合環境の研究開発」（代表：朴泰祐、H24～H30）で継続

基本コンセプト：TCA

- TCA (Tightly Coupled Accelerators)
  - PCIe によるAD間直接通信（ノード内・ノード間）を実現
  - PEACH2チップ（FPGAによるプロトタイプ）によるインテリジェントなPCIeスイッチ＋コントローラ
  - ホストCPU・メモリ・結合網に依存しないAD間直接通信



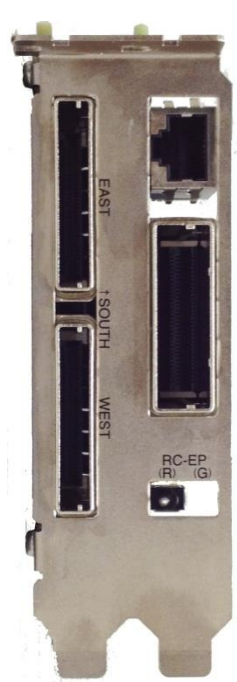
通常のノード間GPU間通信



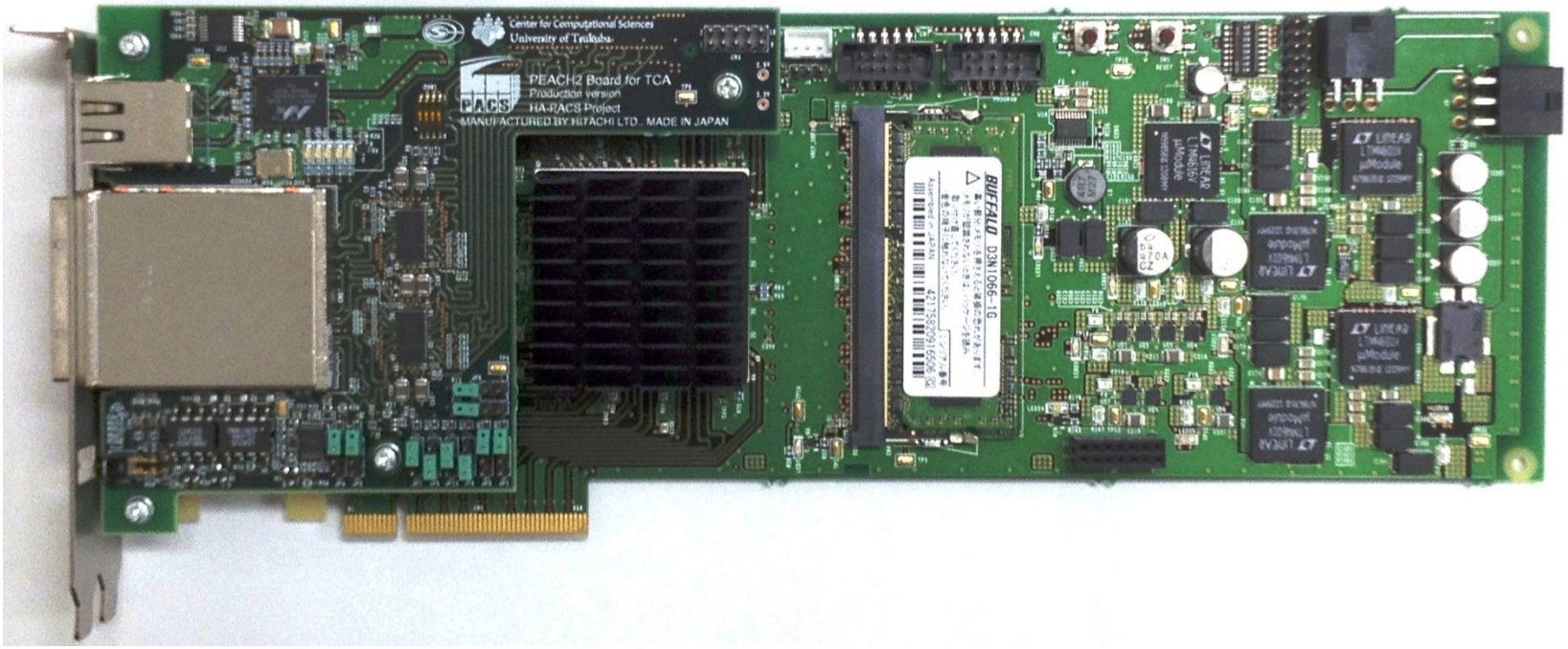
TCAによるノード間GPU間通信

PEACH2 board

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

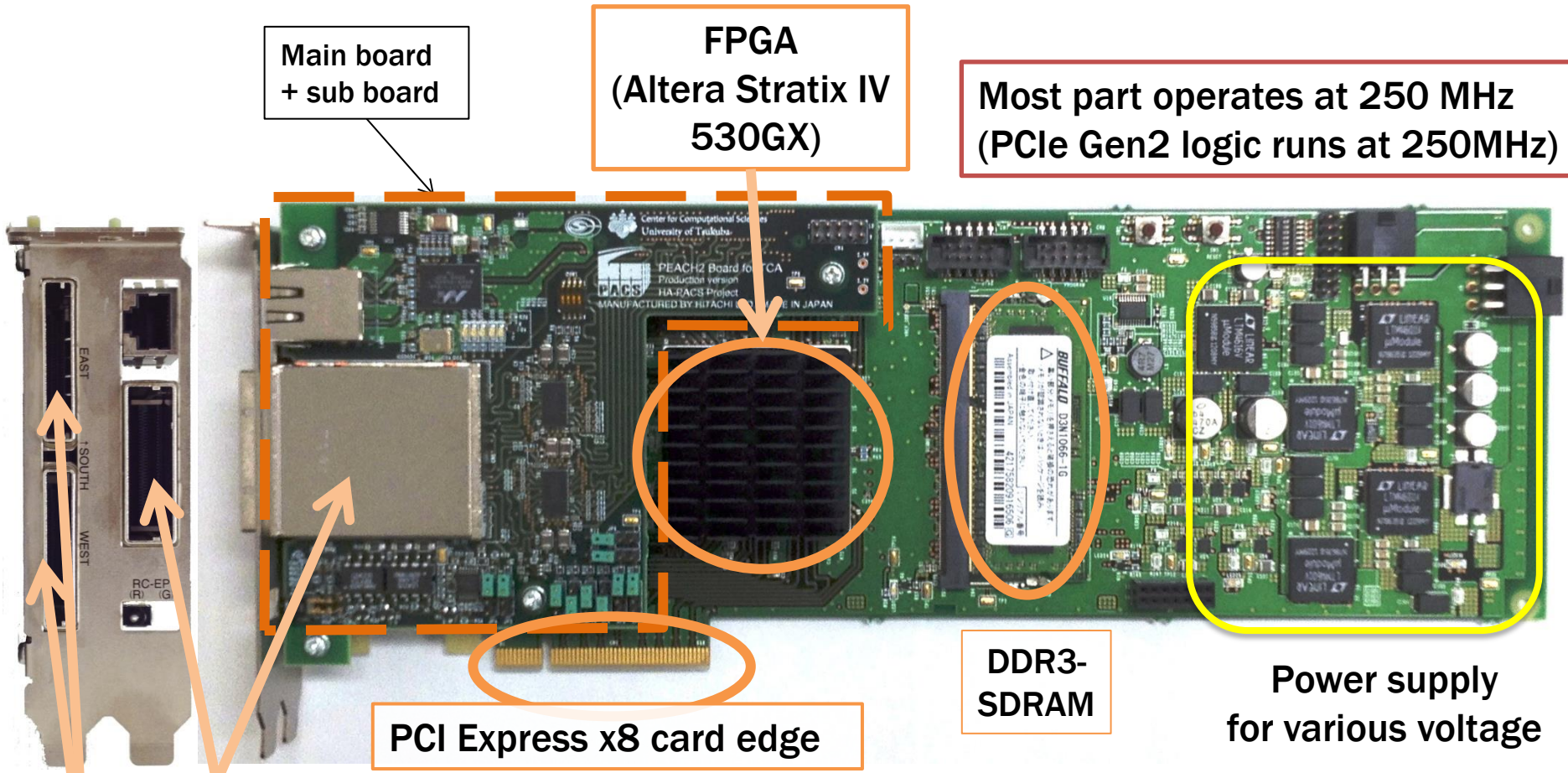


Side View



Top View

PEACH2 board



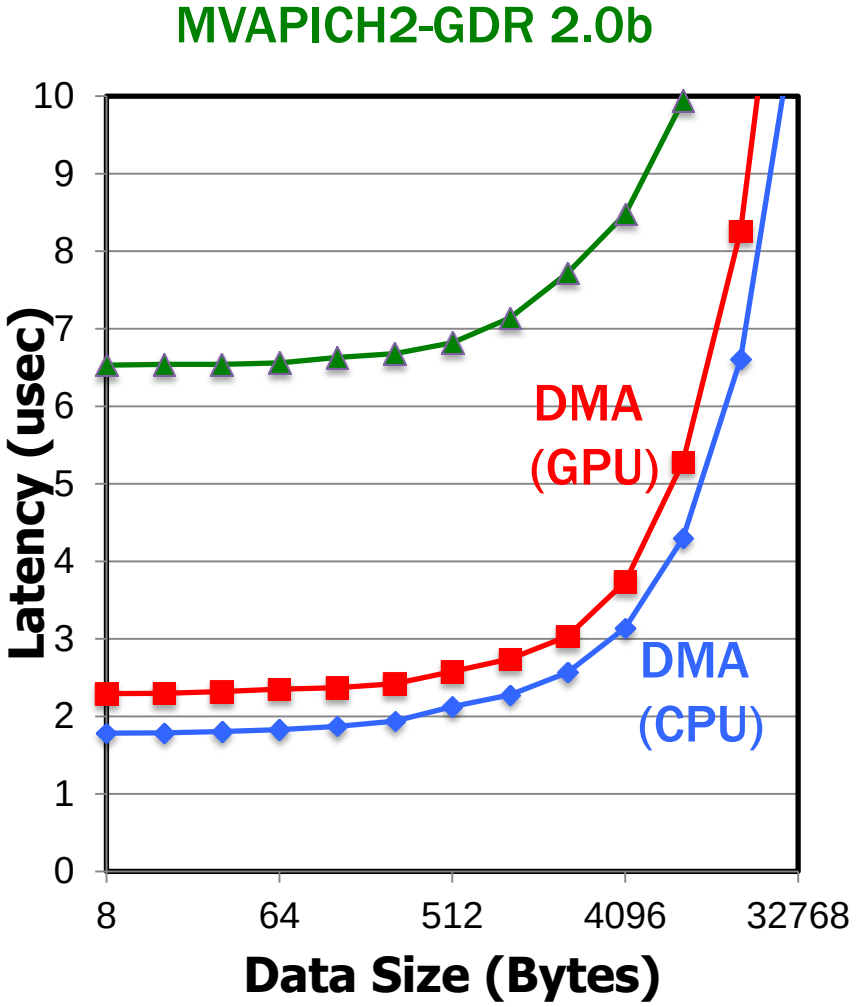
HA-PACS/TCA computation node inside

Ping-pong Latency

Minimum Latency
(nearest neighbor comm.)

- PIO: CPU to CPU: 0.8 us
- DMA:CPU to CPU: 1.8 us
  GPU to GPU: 2.3 us

cf. MV2-GDR 2.0b: 6.5 us (w/ GDR),
17 us (w/o GDR)

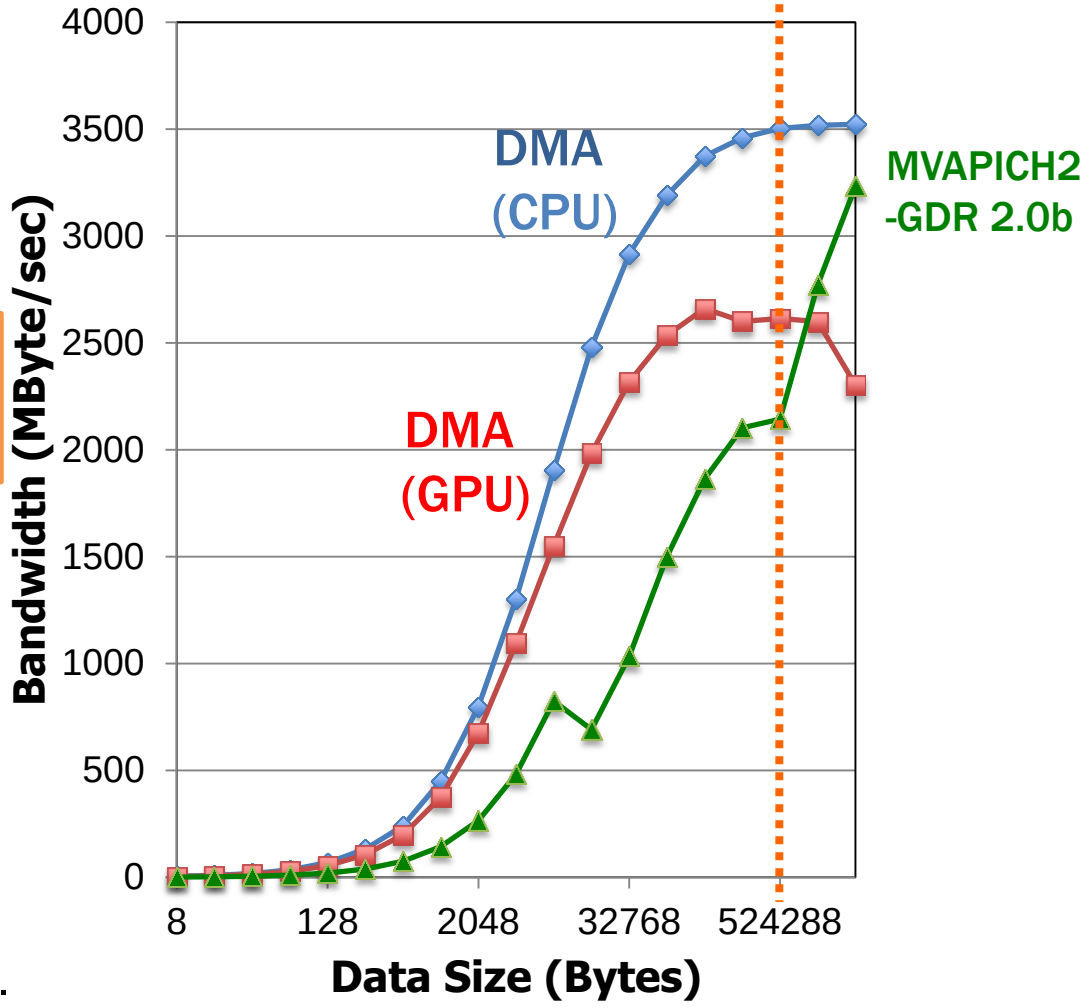


Ping-pong Bandwidth

- Max. 3.5 GByte/sec
  - 95% of theoretical peak
  - Converge to the same peak if hop count increases

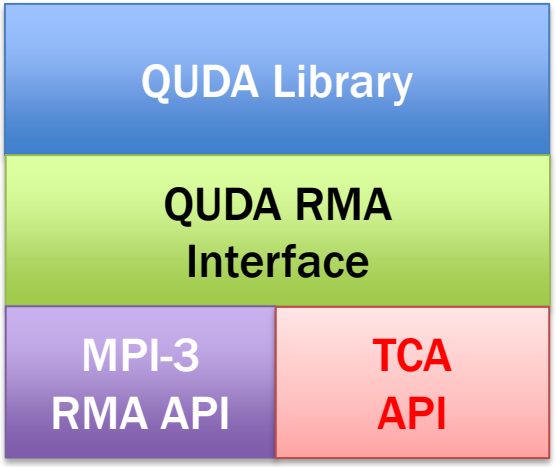
Max Payload Size = 256byte
Theoretical peak (detailed):
4GB/sec × 256 / (256 + 24) = 3.66 GB/s

- GPU – GPU DMA performance is up to 2.6 GByte/sec.
  - better than MV2GDR under < 1MB
  - Over QPI: limited to 360MB/s
  - SB(SandyBridge): limited to 880MB/s due to PCIe sw perf.



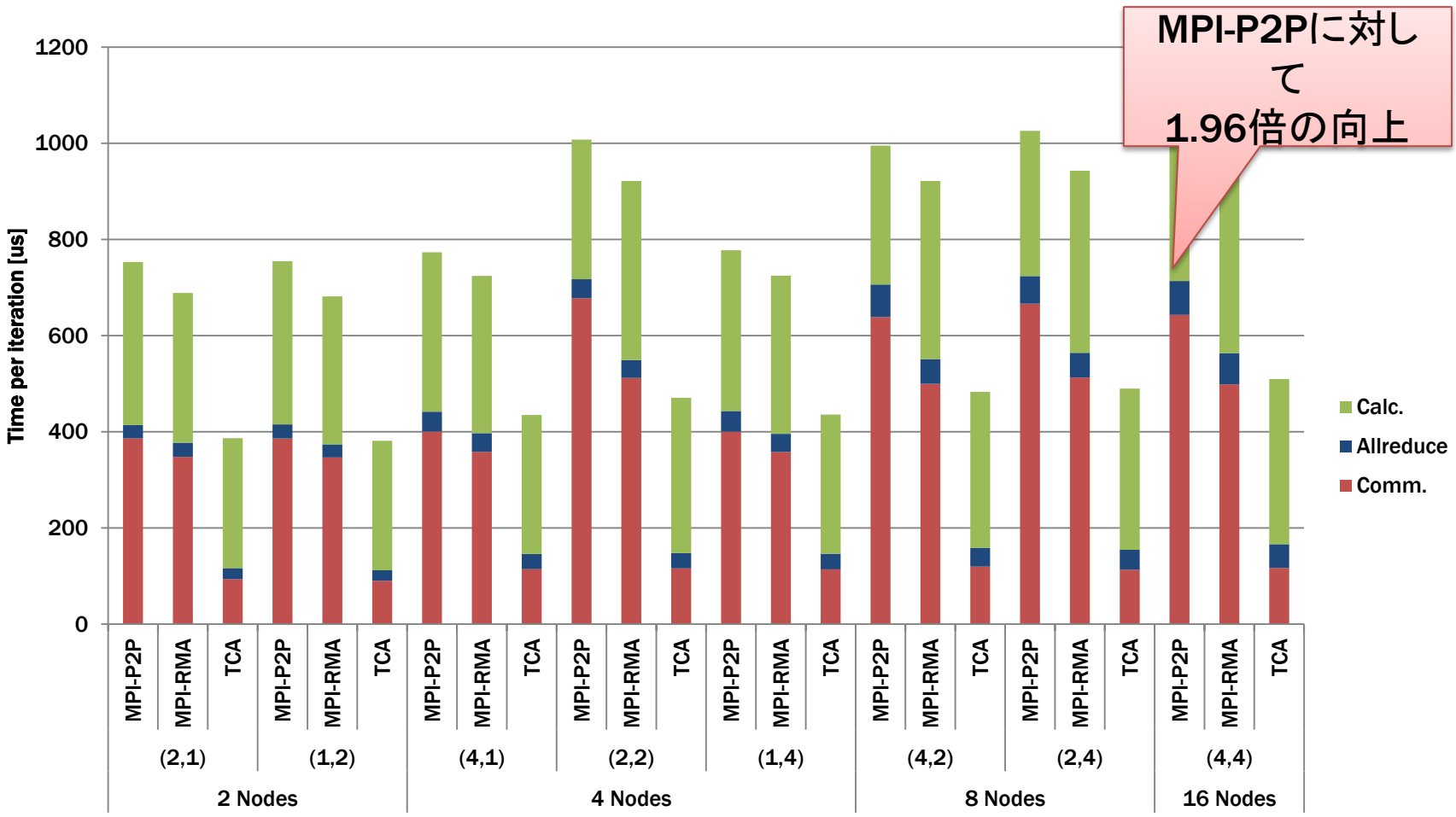
QUDA QCD Library のTCA向け実装

- QUDA: The open source Lattice QCD library
  - NVIDIA GPU向けのLQCDライブラリ
    - 全ての計算をNVIDIA GPU上で行う
- MPI実装のみで、そのままではTCAを適用できない
  - TCAはRDMA Writeで通信を行う
  - RMAをサポートするようにQUDAを拡張する
  - QUDAの通信ライブラリを再構築し、RMA通信ベースで設計
    - MPI-3 RMAでの実装
    - TCAでの実装
- NVIDIA社、Mike Clarkとの共同研究



通信抽象化レイヤーによって複数の通信APIをサポートしている

QUDA Small Model (8^4)

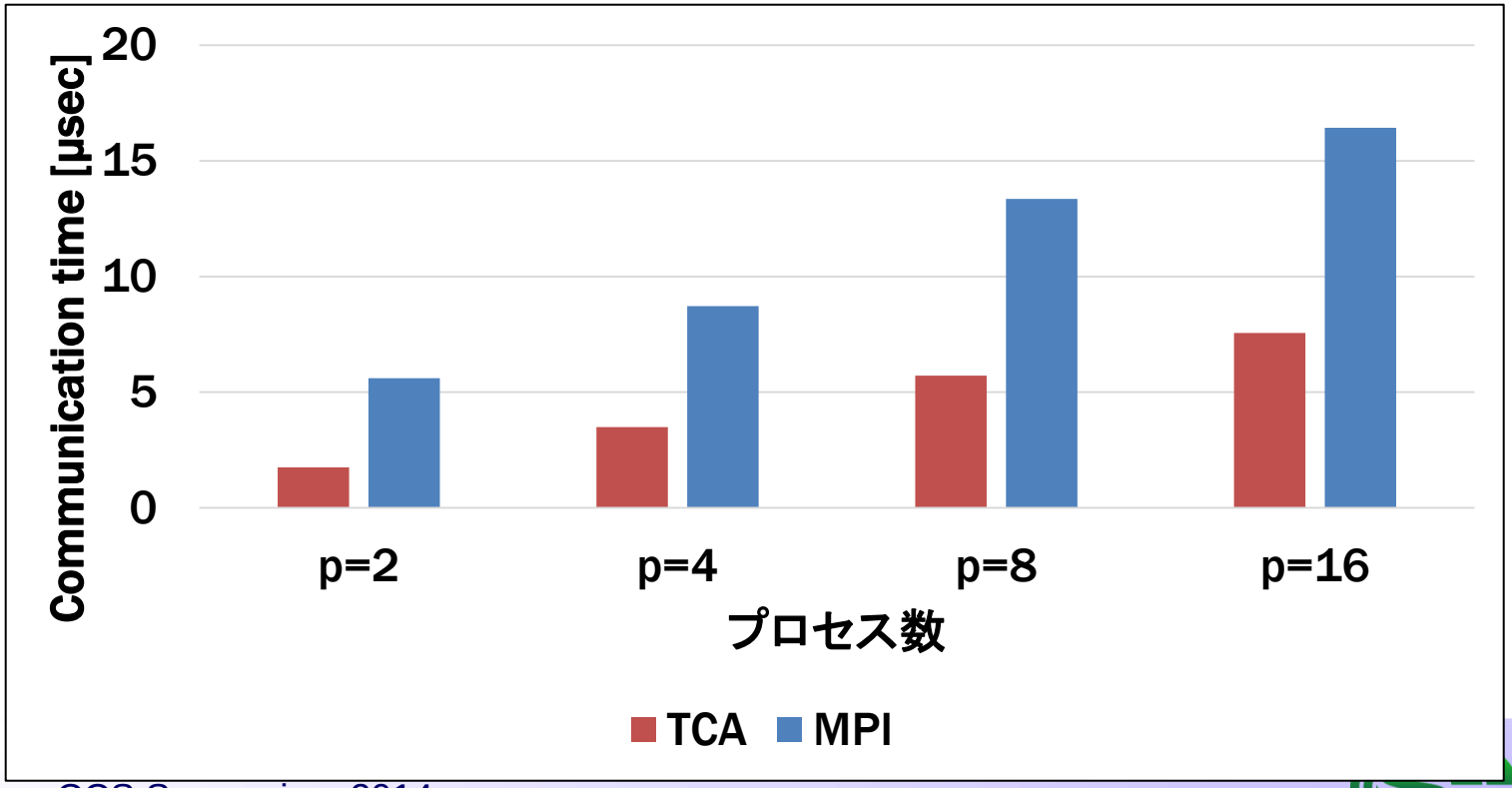


各次元のメッセージ長 =
2 × (24KB /各次元のノード数)

Allreduceの性能

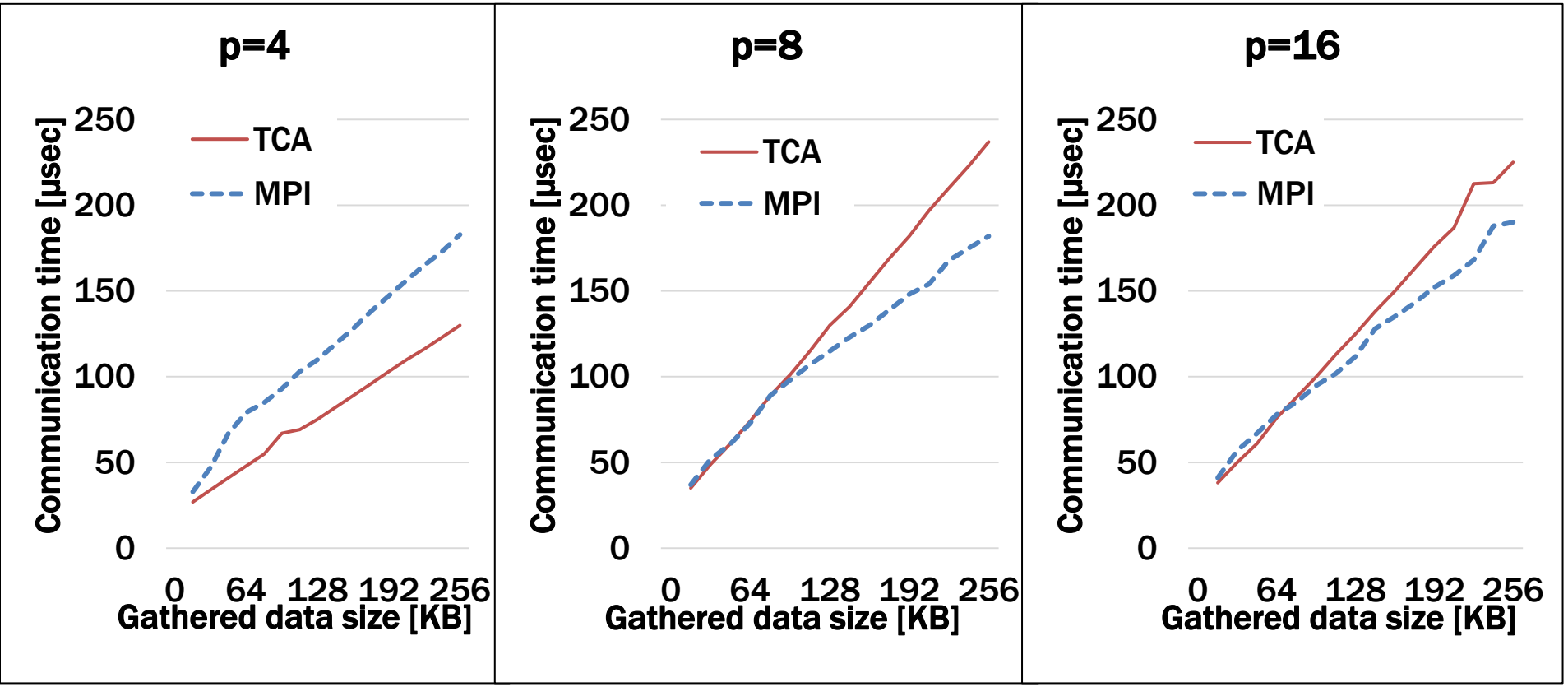
8 Bytesのスカラー値をAllreduceする通信時間（50回の平均）

- MPIを用いた実装の**半分ほどの通信時間**で済む
- TCAの低レイテンシという特徴が有効に働いている

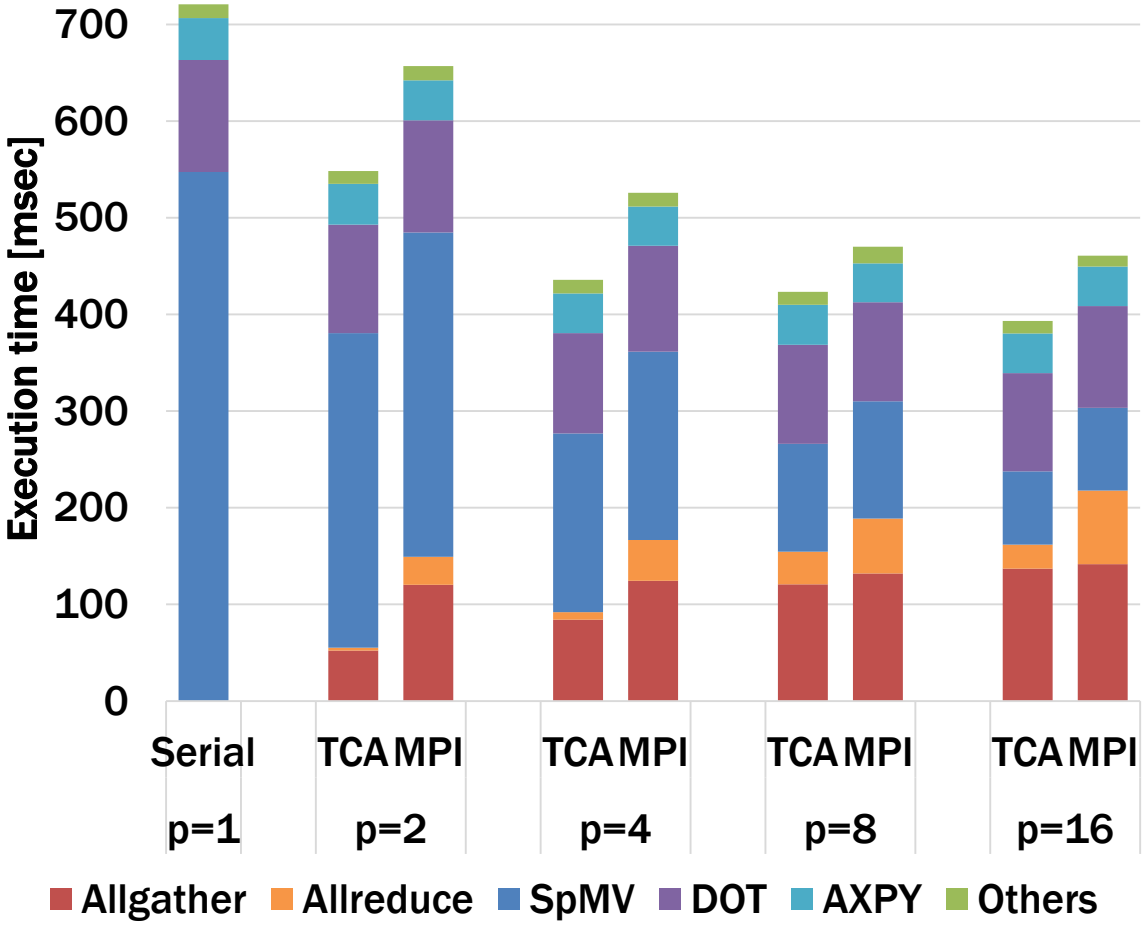


Allgatherの性能:

- HA-PACS/TCAの1サブクラスタ(最大16ノード)における測定結果
- プロセス数pが増えるとMPI (MVAPICH 2 GDR 2.0b) との性能差が小さくなる(TCAでは通信経路の衝突が起こるため)



CG法の性能

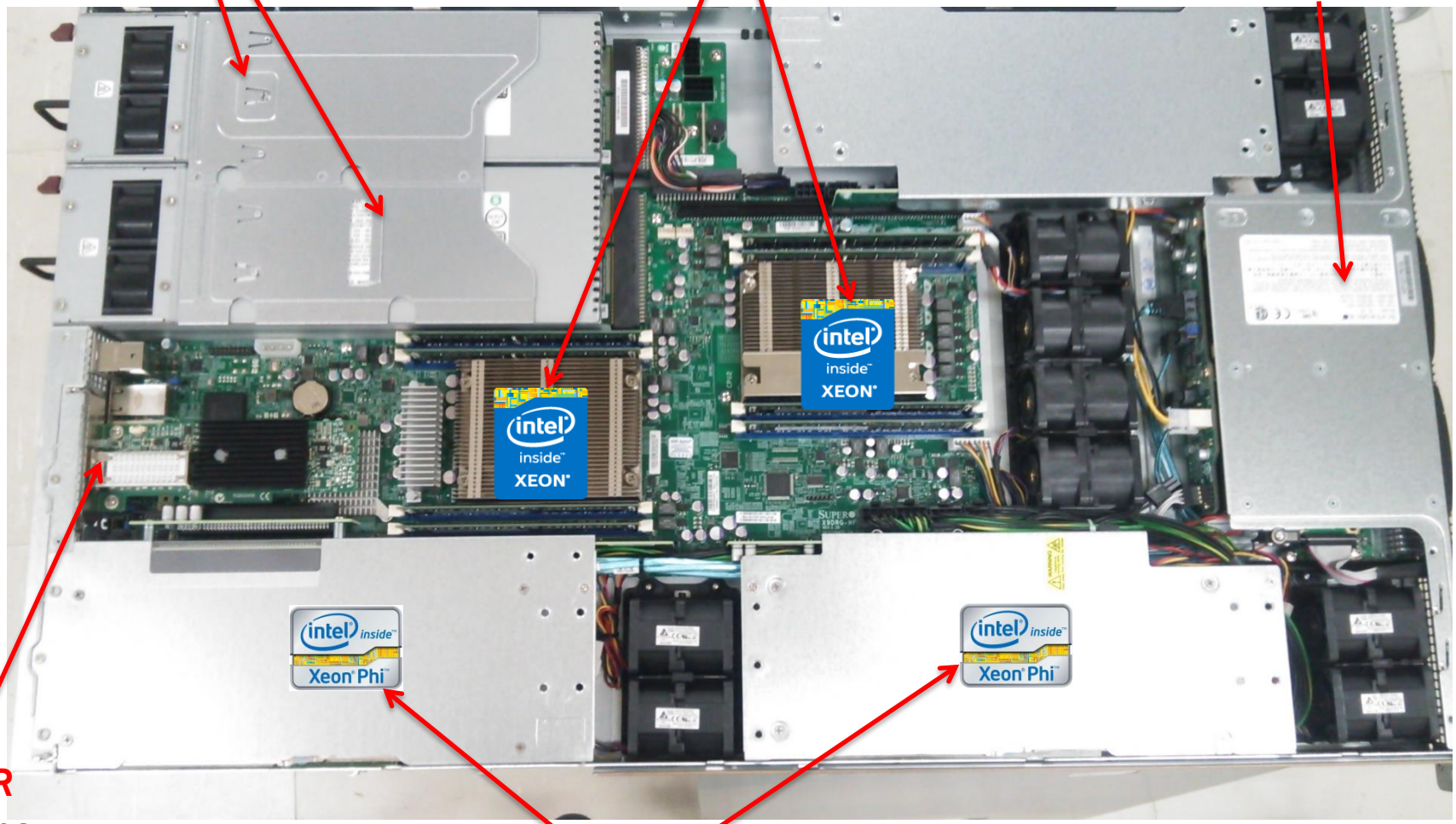


- 1,000回反復を行った時の実行時間
  - GPUは1ノード辺り1GPUを使用
  - プロセスランク0の内訳
- 疎行列名: nd6k
  - 行数: 18,000
  - 非零要素数: 6,897,316
  - 先程の倍の大きさの行列
  - Univ. Florida Sparse Matrix Collectionより取得

COMA (PACS-IX)

- Cray社 CS300 ベース
- Intel Xeon Phi (KNC: Knights Corner)を全面採用
- 393ノード（2 Xeon E5-2670v2 +
  2 Xeon Phi 7110P)
- Mellanox IniniBand FDR, Fat Tree
- 2014/04時点でXeon Phi搭載クラスタとして日本最大
- File Server: DDN 1.5PB (RAID6+Lustre)
- 1.001 PFLOPS (HPL: 746 TFLOPS)
  June '14 TOP500 #51
- HPL効率 74.7%

COMA (PACS-IX) 計算ノード (Cray 1U 1027GR)



COMA (PACS-IX) overview

- T2K-Tsukubaの運転終了後に導入
  - H26年3月末～4月初旬
- システム構成
  - 計算ノード：汎用CPU＋メニーコアプロセッサ
  - ノード構成
    - CPU x 2: Intel Xeon E5-2670v2
    - MIC x 2: Intel Xeon Phi 7110P
    - Memory: CPU=64GB MIC=16GB
    - Network: IB FDR Full-bisection b/w Fat Tree
  - ノード数： 393
  - ピーク性能： CPU=157.2 TFlops　MIC=843.8 TFlops

    TOTAL: 1001 TFlops = **1.001 PFLOPS**
- システムベンダー: Cray Inc.

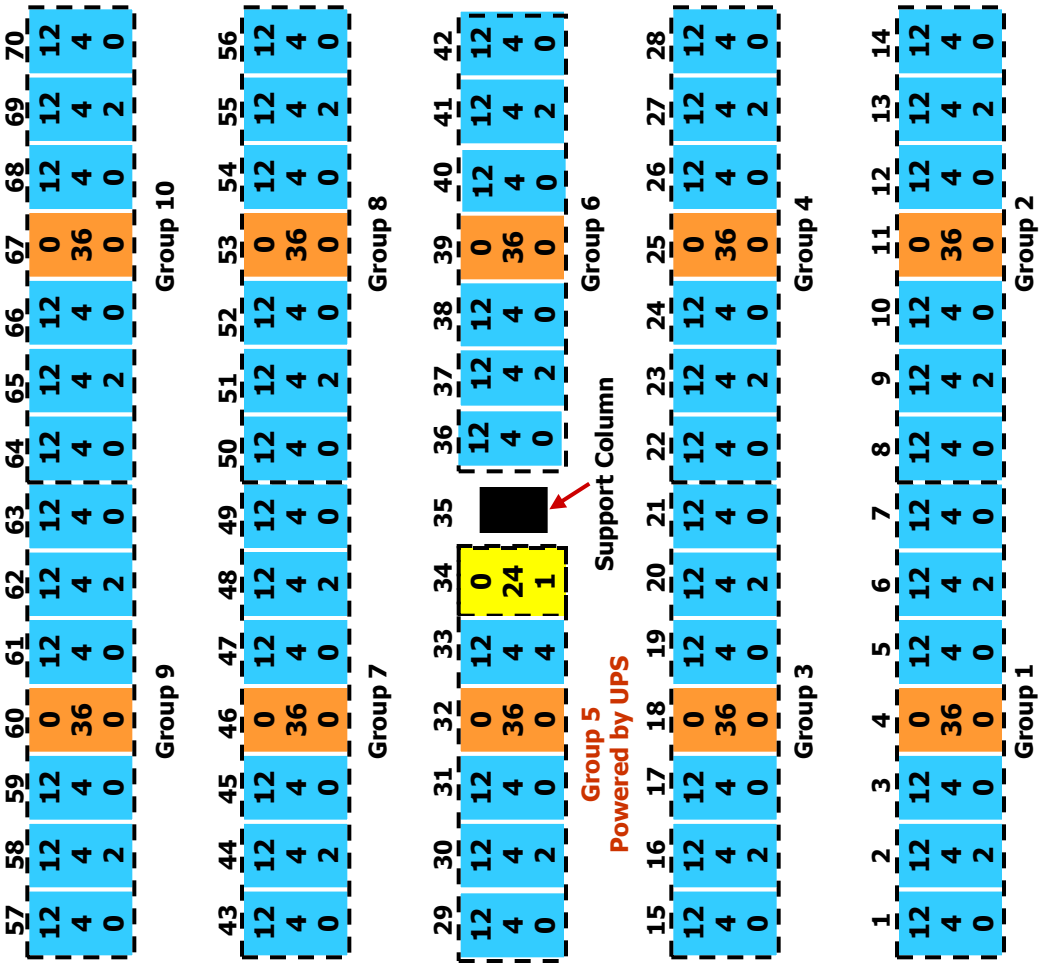
What is COMA ?

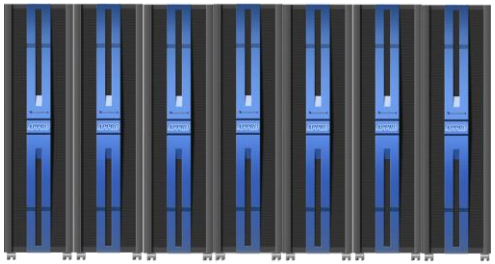
- Cluster of Many-core Architecture processor
- COMA＝「かみのけ座」
  - 代表的な銀河団の一つ
  - 銀河＝星の集まり（＝Many Core）
  - 銀河団＝銀河の集まり（＝Cluster）
- 同時にPACSシリーズ第9世代のマシンとなるため、”PACS-IX”のコード名を併用

T2K-Tsukuba (COMA導入前のシステム)

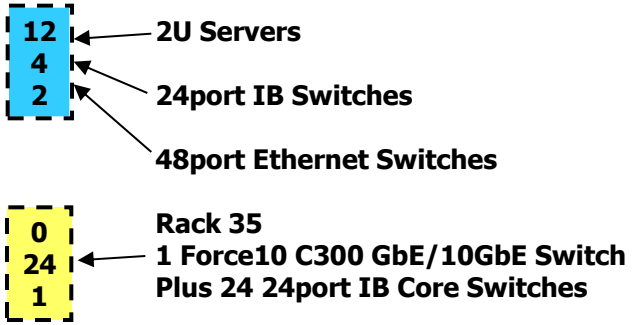
#20 at TOP500 on June 2008 (Linpack: 76.46 TFLOPS)

T2K-Tsukuba floor plan (76.5TF Linpack, 2008)

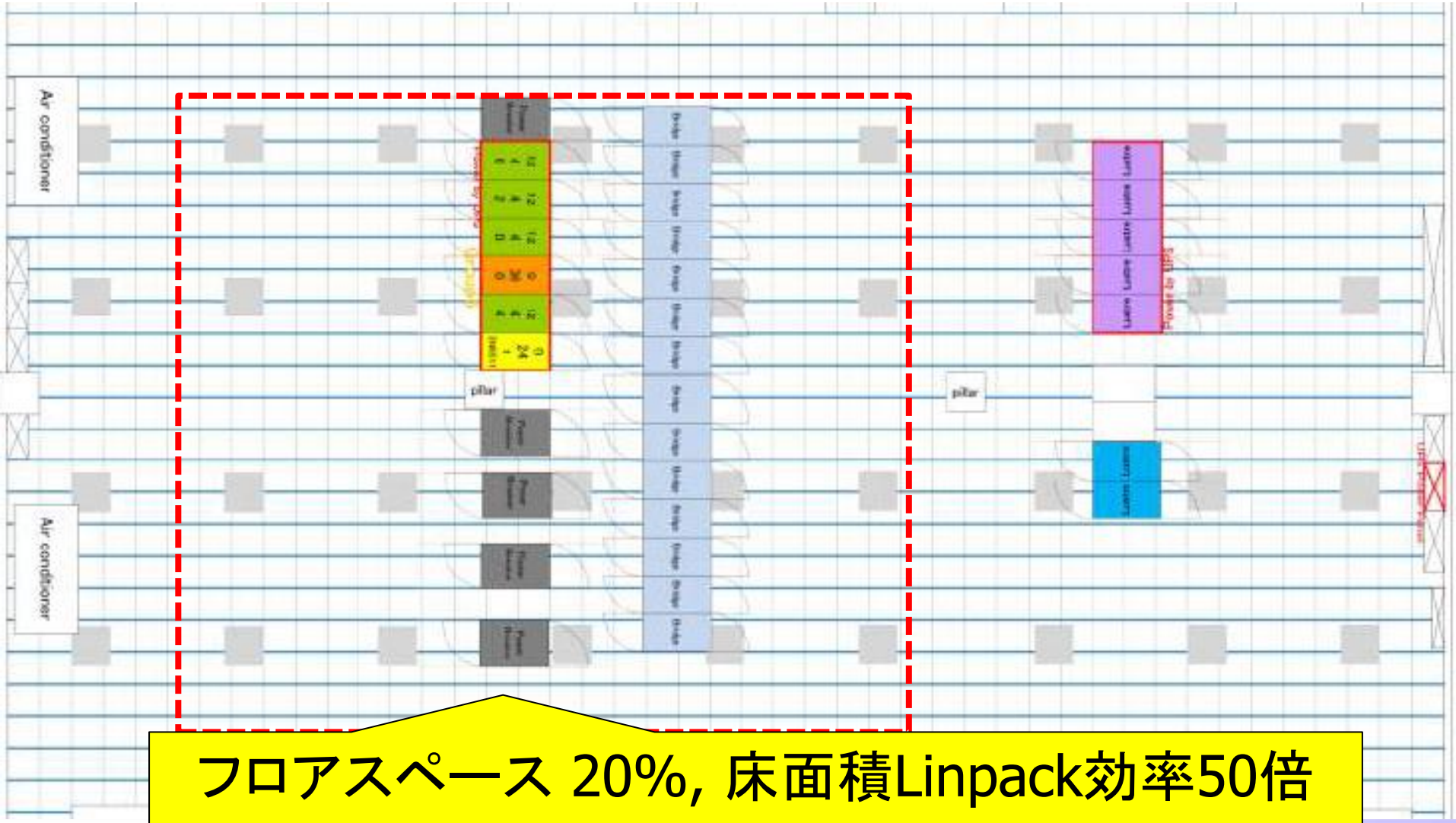




638 Compute Server Nodes
10 Spare Nodes
20 Management Nodes
22 File Server(I/O) Nodes
2 Management Nodes



COMA floor plan (746TF Linpack, 2014)



3つのクラスタの電力

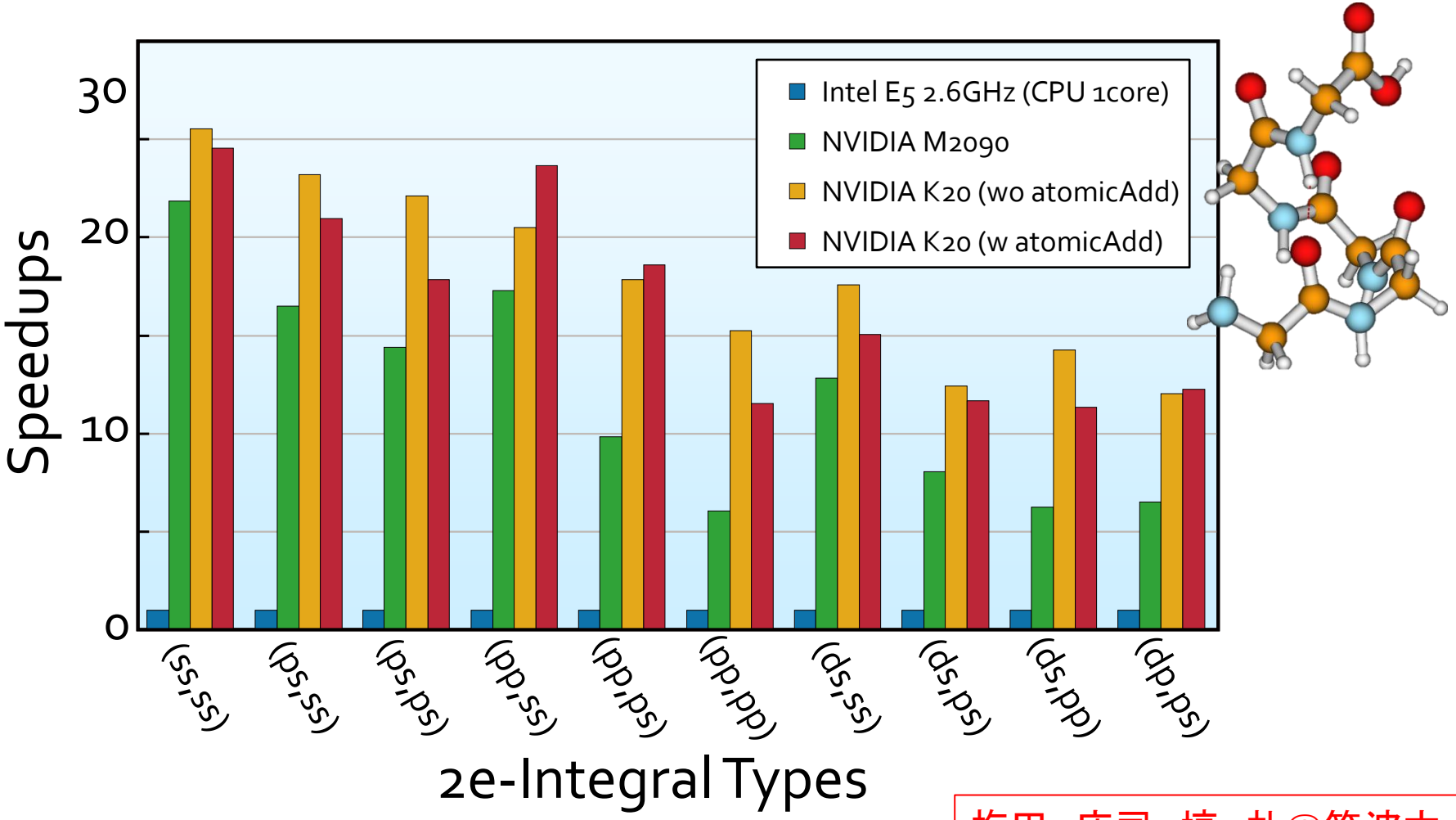
Syste	Linpack性能 (TF)	平均消費電力 (kW)
T2K-Tsukuba	76.5	420
HA-PACS (GPU)	421	250
COMA (MIC)	746	215 (利用率60%)

- Linpack (HPL) 単純換算ではCOMAはT2K-Tsukubaの10~15倍程度の電力効率
- 実アプリケーションの実行効率は？

アプリケーション開発・プロジェクト

- 筑波大学CCSにおける重点開発アプリケーション
  - 素粒子物理学：格子色力学(QCD)
  - 宇宙物理学：重力輻射流体、宇宙生命科学
  - 物性物理学：TDDFT
  - 生命科学：QM/MM、HF行列、FMO
  - 地球環境：LES
  - 計算機科学：GPU/MIC programming language (XMP)
- これら全てについてGPU化、MIC化を推進

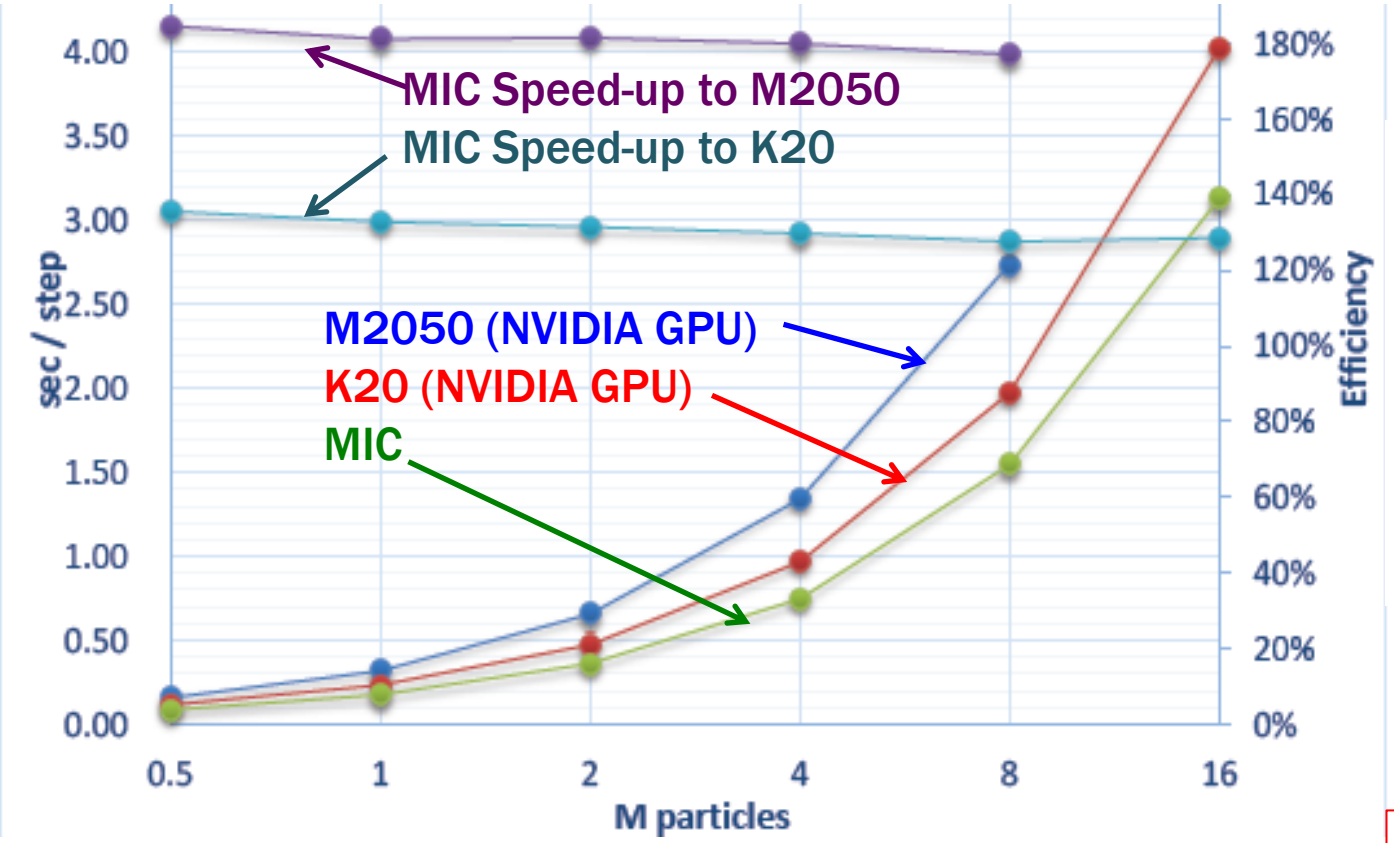
FMOにおける二電子積分のGPU加速



梅田、庄司、塙、朴@筑波大

GPU vs ManyCore (1)

- 重力tree code, 非常に computation bound, well tuned for MIC with intrinsic
  4 threads/core (240 threads) で最高性能

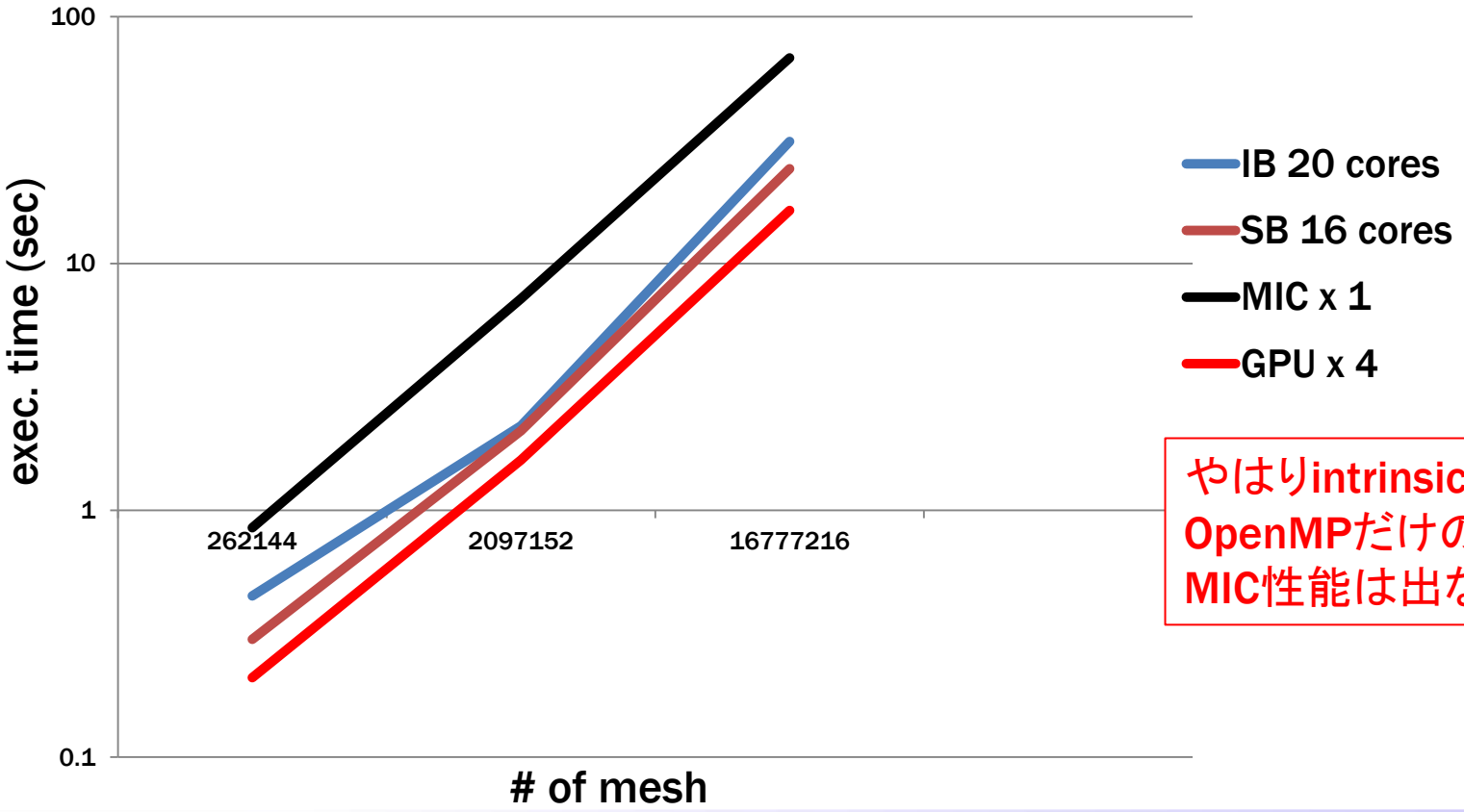


高橋、扇谷、朴@筑波大

データ提供：吉川耕司博士@筑波大

GPU vs ManyCore (2)

- ARGOT (AstroPhysics)：輻射輸送＋化学反応
  - GPU - HA-PACS single node, SandyBridge, NVIDIA Fermi M2090
  - MIC - COMA single node, IvyBridge, Xeon Phi 7110P (without intrinsic)



やはりintrinsicなしのOpenMPだけのC記述ではMIC性能は出ない

データ提供：石川健一博士@広島大

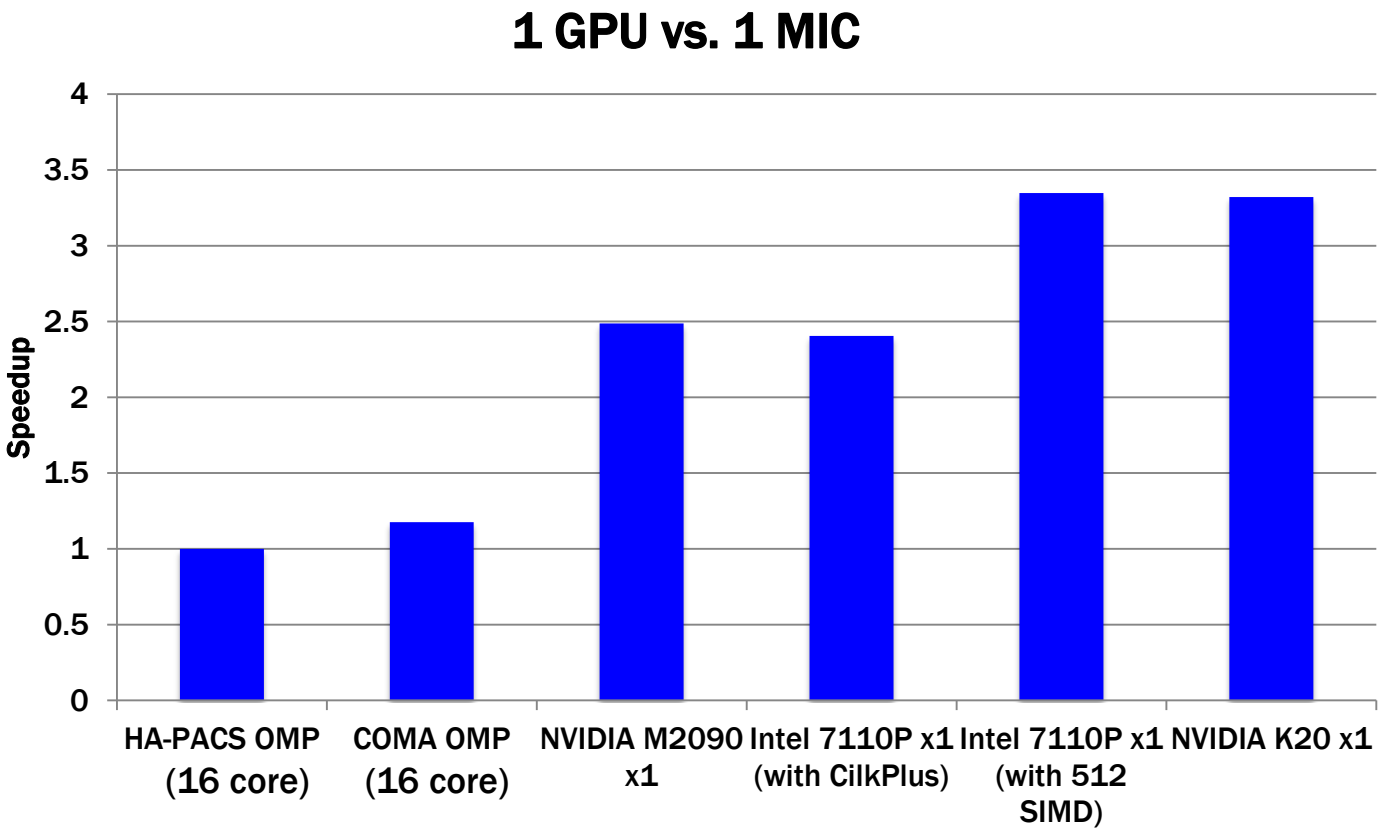
GPU vs ManyCore (3)

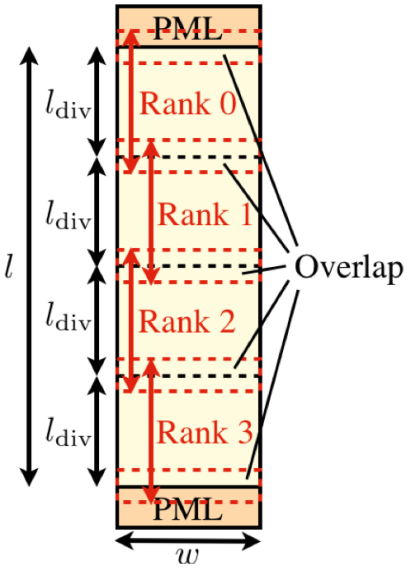
- QCD
  - GPU (Fermi M2090) – CUDA
  - MIC (Xeon Phi 7110P) –OpenMP, or with intrinsic tuned
  - CPU – OpenMP
- single node performance
  - GPU – HA-PACS single node, single GPU
  - MIC – COMA single node, single MIC, 240 threads (60 cores)
  - CPU – SandyBridge (HA-PACS) or IvyBridge (COMA), 16 threads (16 cores)

Resource	Time (sec)	Perf. (Gflops)
HA-PACS: CPU (DP) only	12.15	33.9
HA-PACS: CPU(DP)+GPU(SP)	6.44	74.2 (SP GPU part)
COMA: MIC(DP) only	105.64	3.90
COMA: MIC(DP)+MIC(SP, intrinsic)	6.19	NA

GPU vs ManyCore (4)

- MTDM: 導波管を流れる電磁波の2次元シミュレーション
  （非常に computation bound な計算）



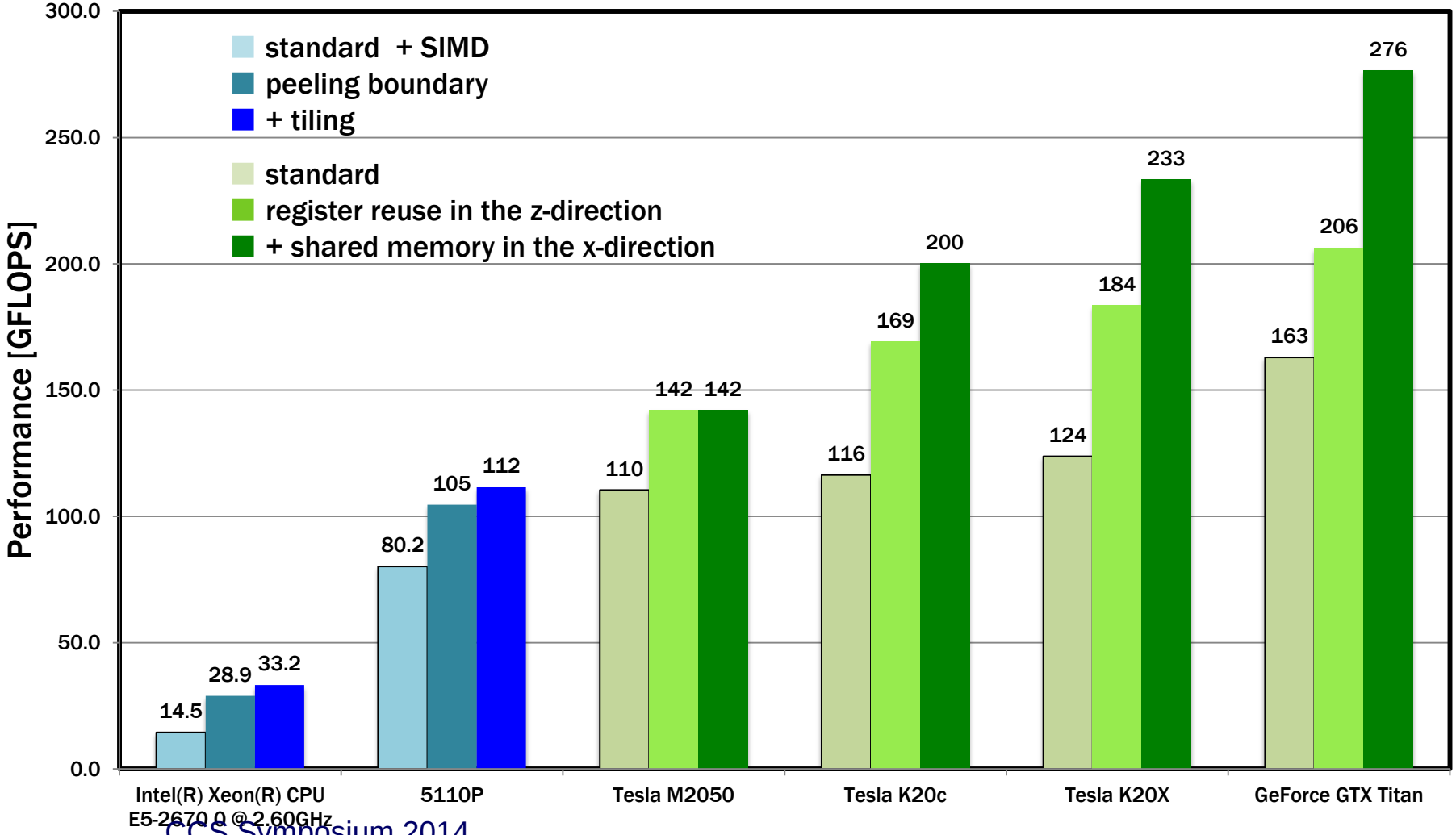


廣川@筑波大

GPU vs ManyCore (5)

資料提供:青木尊之博士@東工大

3次元拡散方程式のシミュレーション (7点 stencil)



MIC (KNC) の実効性能は？

- 「そのまま（OpenMPのみ）」では性能アップは厳しい
- KNCを知った上での様々なプログラミング／チューニング
  - SIMDを意識、特にintrinsicで明示的プログラミングを行うと劇的に性能向上する場合が多い
  - data localityへの配慮
  - スレッド数調整
- アプリケーションがMPI化されていて、かつ負荷分散調整の余地がある場合
  - Symetric mode (CPUとMICを同列にx86コアの集合として利用）が適用可能
    ⇒　ただし、上記特性があるので、プログラムには
    “#ifdef MIC” のような工夫が必要

今後の many-core への期待

- 最先端共同HPC基盤施設（JCAHPC: Joint Center for Advanced HPC）
  - 筑波大学と東京大学がT2K終了後のスパコン共同調達・共同運用を行う施設
  - 東京大学柏キャンパス内に設置
  - 2015年度を目処に、ピーク性能30PFLOPSを目指したスパコンを調達、運用
  - many core architecture に基づくプロセッサを想定
- COMAはこの先鞭として many-core base application の開発ベースとしての役割を果たす

まとめ

- 筑波大CCSではスペース性能・電力性能の観点から、accelerated computing を今後の重要な研究プラットフォームと捉え、HA-PACS, COMAの導入を行い、大規模並列演算加速計算を推進中
- 従来からの主な計算科学課題をGPU化、many-core化
- さらに学際共同利用プログラム等を通して国内の計算科学研究者に計算資源を提供
- HA-PACSの約2年間の運用を通じ、QCD、宇宙物理、生命科学等の重要課題でGPU化を推進
- COMAではmany-core applicationの開発を推進中だが、intrinsicを積極的に利用しないと性能的にはGPU（というかCPU）に及ばない
  ⇒ KNC -> KNL の性能向上に期待 ＋ コンパイラの性能向上
- CCS独自リソースとしては今後、アクセラレータを積極的に用いたシステムを中心に開発・導入を続ける（柏JCAHPCとの平行運用）