HA-PACSプロジェクト紹介

児玉祐悦
筑波大学計算科学研究センター
次世代計算システム開発室

筑波大：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
- 2006年のPACS-CSは第7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	7千回/秒
1980年	PAXS-32	50万回/秒
1983年	PAX-128	4百万回/秒
1984年	PAX-32J	3百万回/秒
1989年	QCDPAX	140億回/秒
1996年	CP-PACS	6140億回/秒
2006年	PACS-CS	14.3兆回/秒

- 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

Exa-scale時代を睨んだ次世代PACSシステム

- post-Peta Scale⇒Exa Scale時代のHPCシステム
  - 何らかのアクセラレータ技術の導入
  - アクセラレータによって「本当に」加速されるアプリケーションの開発（アルゴリズムレベルを含む）
  - 「汎用」アクセラレータ技術
    - GPGPU
    - Intel MIC (Many Integrated Core)
    - Cell Broadband Engine
    - FPGA
    - GRAPE-DR
  - 現時点で最も performance/cost ratio の高いアクセラレータ⇒GPGPU

GPGPUを積極的に利用した大規模並列クラスタをベースにアクセラレータによる大規模科学技術計算アプリケーションの加速を目指す

GPUコンピューティング：現在のHPCの潮流

- GPU clusters in TOP500 2011/6
  - 2位 天河Tienha-1A （Rpeak=4.7PFLOPS）
  - 4位 星雲Nebulae (Rpeak=3PFLOPS)
  - 5位 TSUBAME2.0 (Rpeak=2.3PFLOPS)
  - (1位 K Computer Rpeak=8.8PFLOPS)
- GPU搭載MPP
  - Cray XK series
- 特徴
  - 圧倒的な peak performance / cost 比
  - 圧倒的な peak performance / power 比
  - 超並列型はTOP500に連なっているが定常的に大規模（PFLOPSクラス）アプリケーションが走っている状態ではない
    ⇒ 超並列GPUアプリケーションは発展途上

GPUクラスタの問題点

- GPGPU(& 一般的なアクセラレータハード)による高性能計算の問題点
  - データ入出力:I/O busによる制約
  - ex) GPGPU: PCIe Gen2 x16 が標準的
    - 理論ピーク性能: 8GB/s (I/O)
      ⇔ 665 GFLOPS (NVIDIA M2090)
  - アクセラレータ間のノード間直接通信は不可能
    ⇒ CPUを介した間接的通信による通信レイテンシの増大
    - ex) GPGPU:
      GPU mem ⇒ CPU mem ⇒ (MPI) ⇒ CPU mem ⇒ GPU mem
- GPU(アクセラレータ)のノード間直接結合に関する要素技術研究が必要

開発体制

- 先端計算科学推進室(室長:梅村)
  - 超並列GPUアプリケーション開発
    - 素粒子物理学
    - 宇宙物理学
    - 生命科学
    - 原子核・量子物性
    - 地球環境
    - データ基盤
    - (計算機システム)
- 次世代計算システム開発室(室長:朴)
  - システム開発

「HA-PACSにおける次世代計算科学」
梅村雅之

次世代PACSシステム：HA-PACS

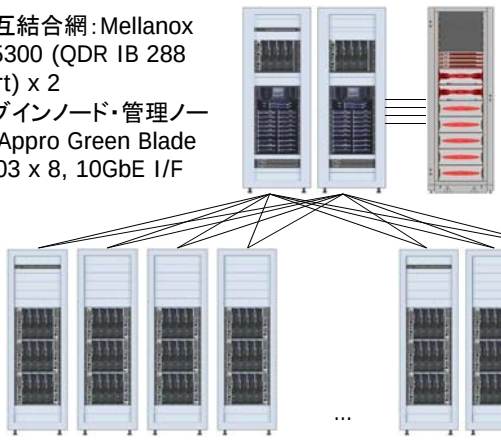
- HA-PACS (Highly Accelerated Parallel Advanced system for Computational Sciences)
- 2つのシステムから構成
  - Base Cluster 部
    - 最先端CPUと最先端GPUの組み合わせによる標準的な大規模クラスタを構築
    - 先進的 I/O bus 技術の導入により高効率で GPGPU 技術を利用
    - GPU技術を生かした次世代アクセラレータ対応の大規模並列アプリケーションを開発 & プロダクトラン
  - TCA (Tightly Coupled Accelerator) 部
    - PCIeバスを介した acceleration d
    - ノード間に跨がる acceleration de

「密結合加速機構研究開発」
塙 敏博

HA-PACS: Base Cluster 部

8/29 開札により（株）アルゴグラフィックス/クレイ・ジャパン・インクの共同提案に決定

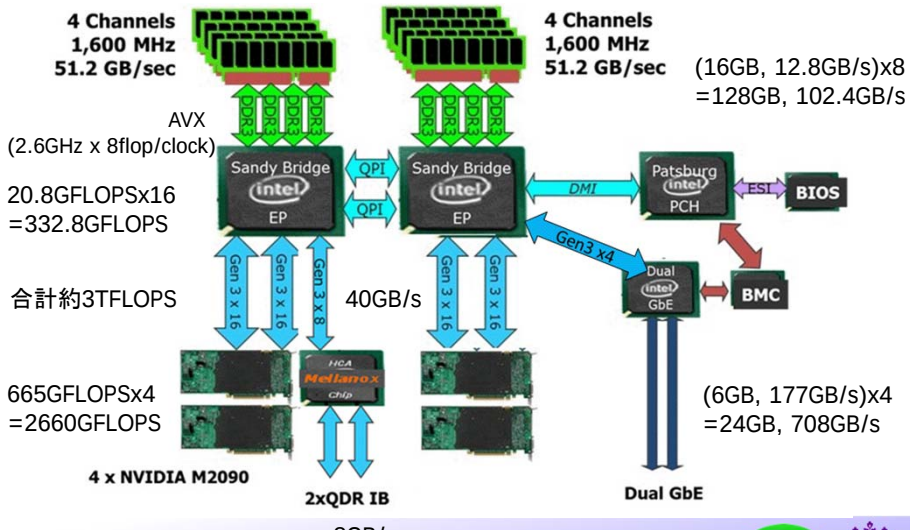
相互結合網：Mellanox IS5300 (QDR IB 288 port) x 2
ログインノード・管理ノード：Appro Green Blade 8203 x 8, 10GbE I/F



ストレージ：DDN SFA10000, QDR IB 接続, Lusterファイルシステム, ユーザ領域 504TB

計算ノード：Appro Green Blade 8204 (8U enc. 4 node) 268 node (67 enc./23 rack)

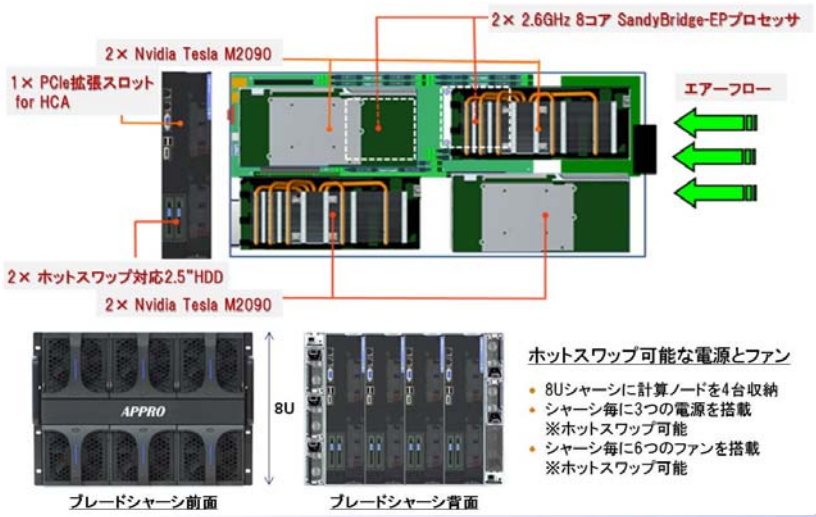
HA-PACS: Base Cluster 部（計算ノード）



HA-PACS: Base Cluster 部（GPU）

- Nvidia M2090
  - 512基のCUDAコアを搭載
  - 構成可変な合計64KBの共有/L1キャッシュを搭載。768KBのL2キャッシュ搭載。
  - マルチカーネルサポート
  - 統合アドレス空間をサポート
  - 単精度･倍精度の両演算にてFMA命令サポート(倍精度665GFLOPS、単精度1331GFLOPS)
  - ECC対応6GB GDDR5、177GB/秒
  - 消費電力225W

HA-PACS: Base Cluster 部（ブレードシステム）



HA-PACS: Base Cluster 部（全体）

- 268ノードを2台のIBスイッチにより接続
- CPU性能 89TFLOPS + GPU性能 713TFLOPS = 合計 802TFLOPS
- メモリ容量CPU:34TByte、メモリバンド幅27TByte/秒, GPU:6.4TByte, 190TByte/秒
- バイセクションバンド幅2.1TByte/秒
- ストレージ504TByte
- 消費電力　408kW（分電盤での測定が可能）
- 26ラック（5.5m x 10m 含む作業空間）
- 2012年1月末導入

まとめ

- **HA-PACS: 筑波大学の次世代GPUクラスタ**
  - Base Cluster 部による大規模並列GPUアプリケーション開発を進め、次世代アクセラレータ技術によるアルゴリズムレベルからのアプリケーションを育成
  - TCA 部によるアクセラレータ間直接通信要素技術を開発し、次世代 accelerated computing への基盤技術につなげる
- **Base Cluster 部は 2012/01に完成、TCA 部はその1年後（ただしテストシステムは2012前半に完成）**
- **総合的に 1PFLOPS peak 性能のシステムを構築し、先進的大規模科学技術計算を集中的に実行**