密結合加速機構研究開発

塙　敏博
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
先端計算科学推進室

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 w/o accelerator,
    10位 Roadrunner Rpeak=1.4PFLOPS w PowerXCell)
- **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 Gen.2 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**

次世代PACSシステム：HA-PACS

- HA-PACS (Highly Accelerated Parallel Advanced system for Computational Sciences)
  - 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
  - TCA (Tightly Coupled Accelerator) 部
    - PCIeバスを介した acceleration device 間の直接通信要素技術の開発
    - ノード間に跨がる acceleration device の通信遅延の大幅削減

➡ 次世代accelerated computingに向けた基盤技術の開発と，それに対応する並列アプリケーションの開発

HA-PACS: TCA 部

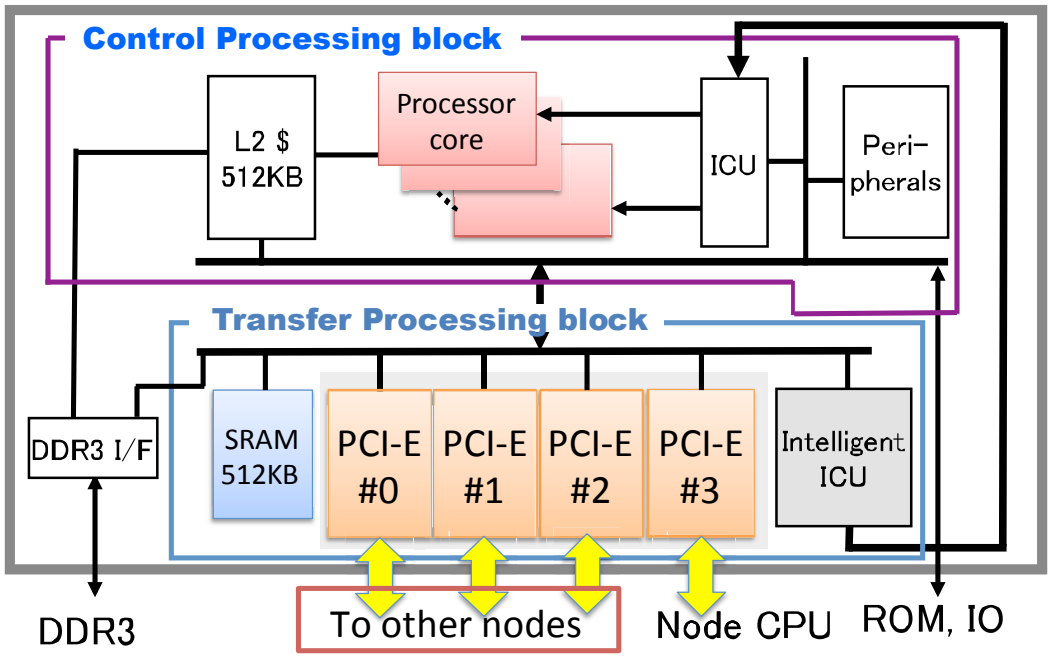
- **TCA: Tightly Coupled Accelerator 技術**
  - アクセラレータ (GPU) 間結合技術
  - PCIe を利用したデバイス間直接結合
    - 現在の全ての acceleration device （その他のI/O deviceも）は PCIe によって結合され、全ては PCIe end-point (device) として実装されている
    - Intelligent な PCIe 間結合機構により、論理的にdevice (end-point) 間の直接結合が可能
- **PEARL: PCI Express Adaptive and Reliable Link**
  - JST-CREST「実用化を目指した組込みシステム用ディペンダブル・オペレーティングシステム」課題「省電力でディペンダブルな組込み並列システム向け計算プラットフォーム」で開発
  - PCIe Gen.2 リンクによるノード間直接結合を実現

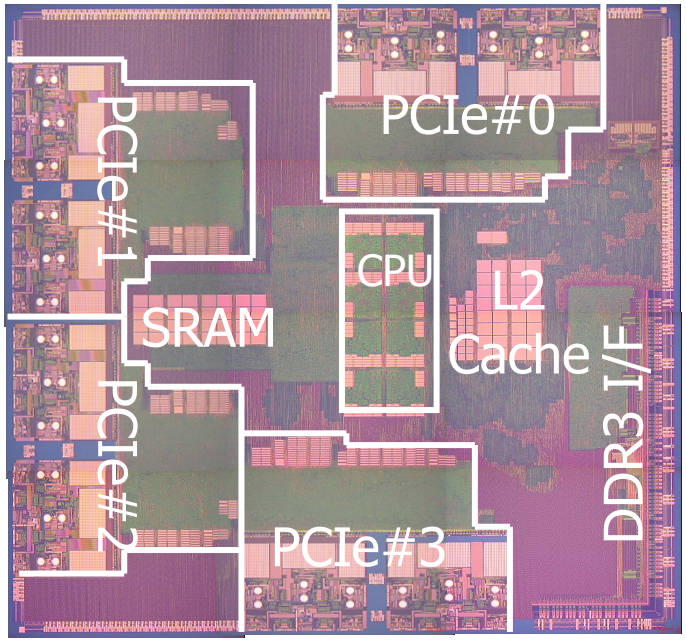
⇒ **PEARLのHPC向け改良版によりTCAを実現**

PEARL: PCI Express Adaptive & Reliable Link

- **PCI Express**を高速シリアル通信としてそのまま利用
- **PEACH (PCI Express Adaptive Communication Hub)**チップ経由で各ノードを接続
  => CPU同士の通信を実現
  - 本来 CPUにはデバイスしか接続できない
- 高性能に加えて，低消費電力・省電力，耐故障を可能に

PEACH チップの構成 [Otani et al., ISSCC2011]

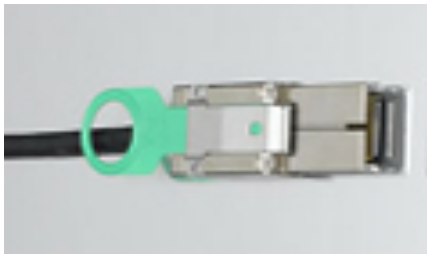


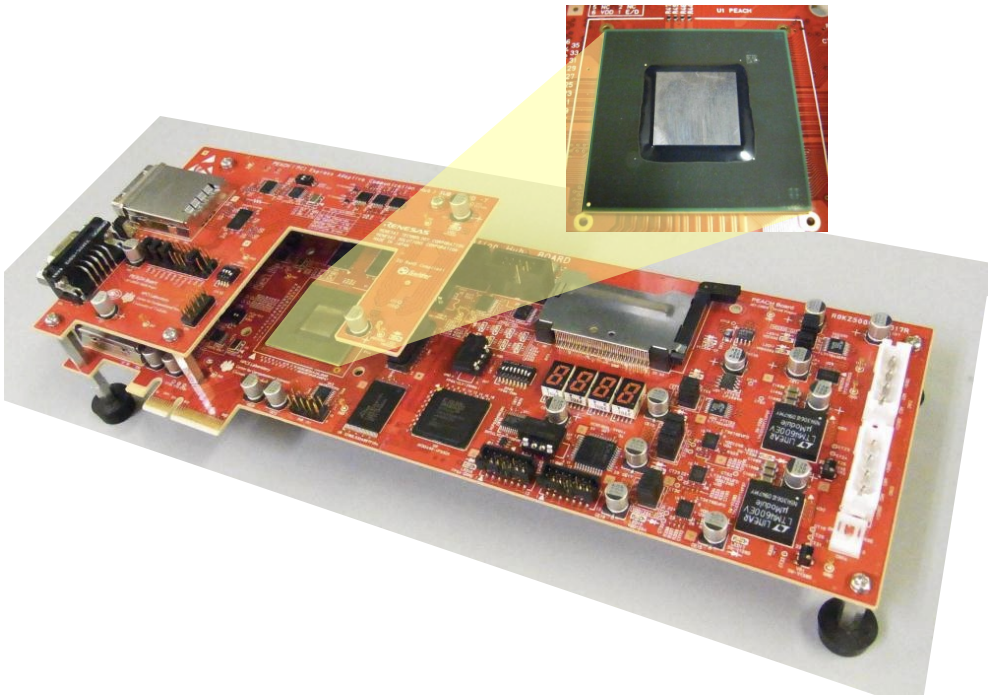


- CPU: ルネサス M32R 4コア SMP (max. 400MHz)
- 通信リンク： PCI Express Gen2 x4レーン (20Gbps) *4ポート

PEACHボード

- **PCIe Card ElectroMechanical (CEM)規格に準拠, PCIe x4スロットに挿入して使用**
  - 3ポート目はドータボードを使って接続





PEACH ⇒ PEACH2 への進化

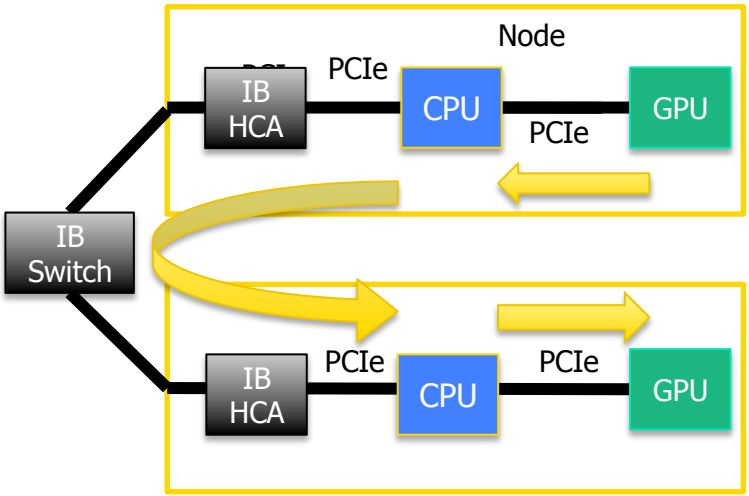
- バンド幅
  - GPUのGen2 x16に対して Gen2 x4 しかない
    => Gen2 x8 （本来はGen3 x8にしたいが…）
  - DMAコントローラの性能・機能不足
    => Chaining DMA, Scatter/Gather
- レイテンシ
  - 内蔵プロセッサによるハンドリング
    => ハードワイヤードロジック
- **PCIe Gen.2 のIPを持つFPGAで実装**

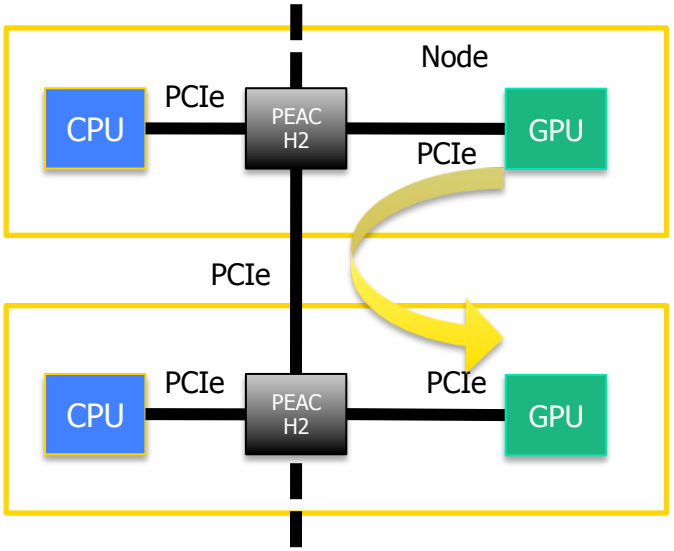
Altera社 FPGA (Stratix IV GX)を使用予定

PEACH2を用いた HA-PACS/TCA 部

- HPC向けに展開
  - 高バンド幅，低レイテンシ化
  - 低消費電力には目をつぶる
- GPU Computing向け
  - HA-PACS/TCA

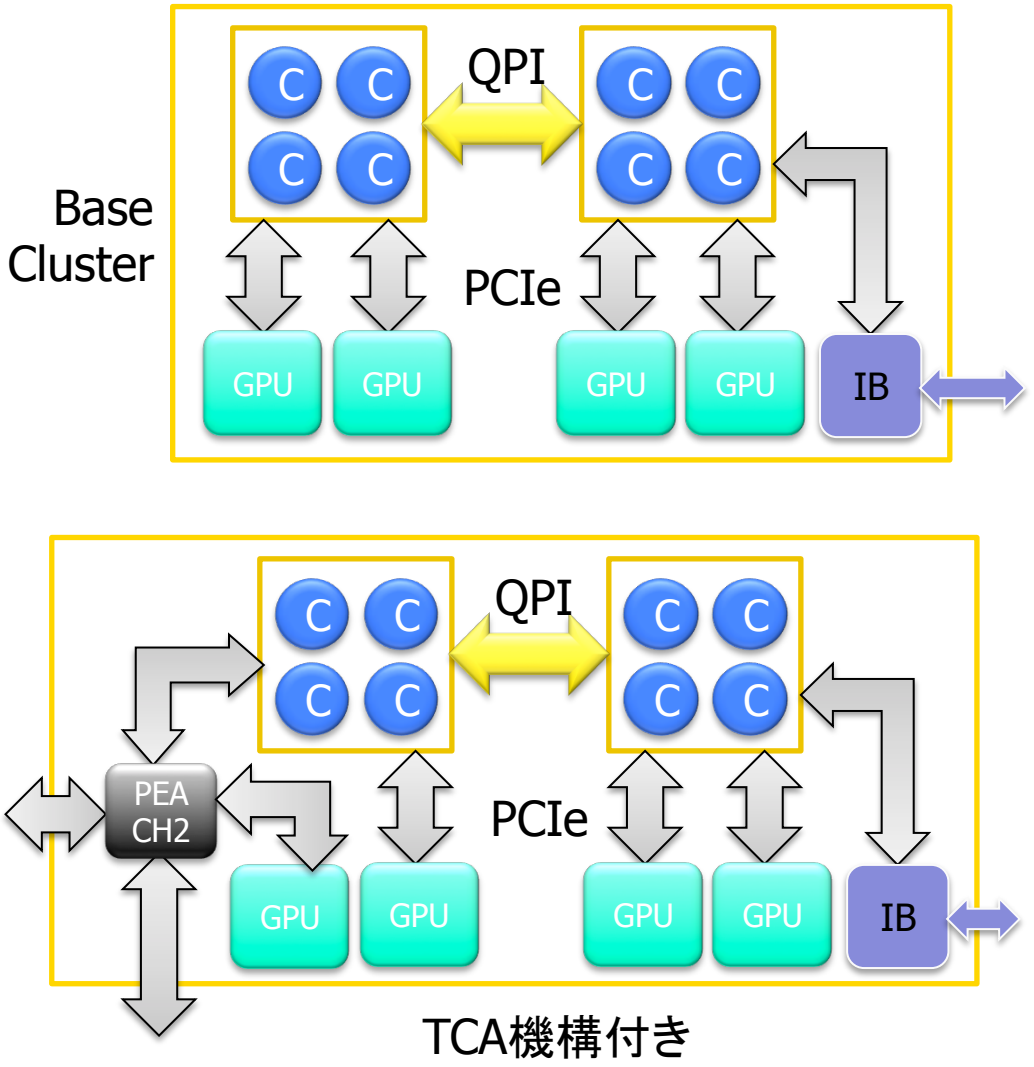
- True GPU-direct
  - GPU間直接通信プロトコルが必要
    ⇒ NVIDIAからの技術協力の下開発中





HA-PACS/TCAのノード実装イメージ

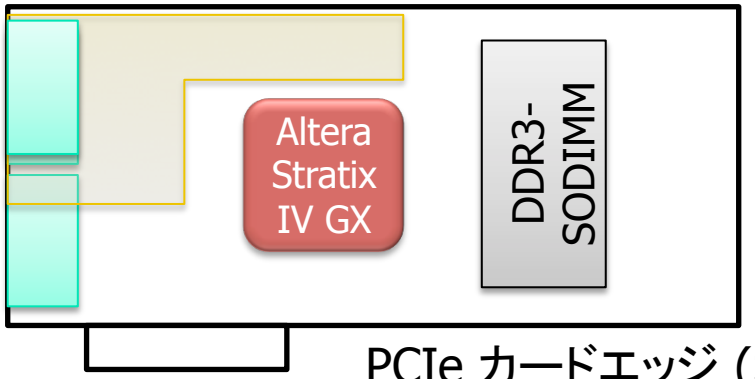
- Base Cluster に加えて、PEACH2によるTCA機構を持たせたノードにより拡張
- HA-PACS/TCAで全ノードをPEARL結合するわけではない
  （現在のPEACH2のノード数制約、channel数制約）
- TCA機構により十数ノード程度を結合したサブクラスタを構築、それらをIBで結合



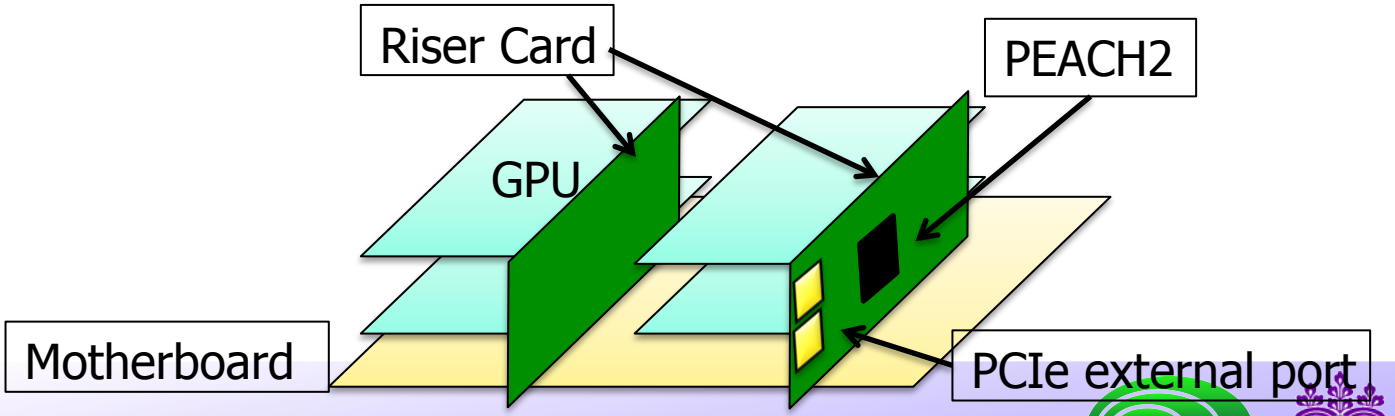
PEACH2ボード実装イメージ（案）

- **PCIe規格準拠 （2スロット占有）**

PCIe ケーブルコネクタ
- GPU ＊ 1
- ノード間 ＊ 2



- **または，ライザーボードに内蔵**



HA-PACS TCA部 スペックとスケジュール

- ノード詳細は未定、PEACH2によるPEARL技術により
  ノード内の数個のGPUをTCA結合
- ノード数は未定、数十ノードで200TFLOPS以上
  ⇒ Base Cluster 部と合わせ 1PFLOPS
- 2012/03までにPEACH2を完成、2013/03までにTCA完成予定
  - 現在，市販FPGA評価ボードを用いて設計・評価中
- 2013/04以降，TCAを活用したアプリの開発

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

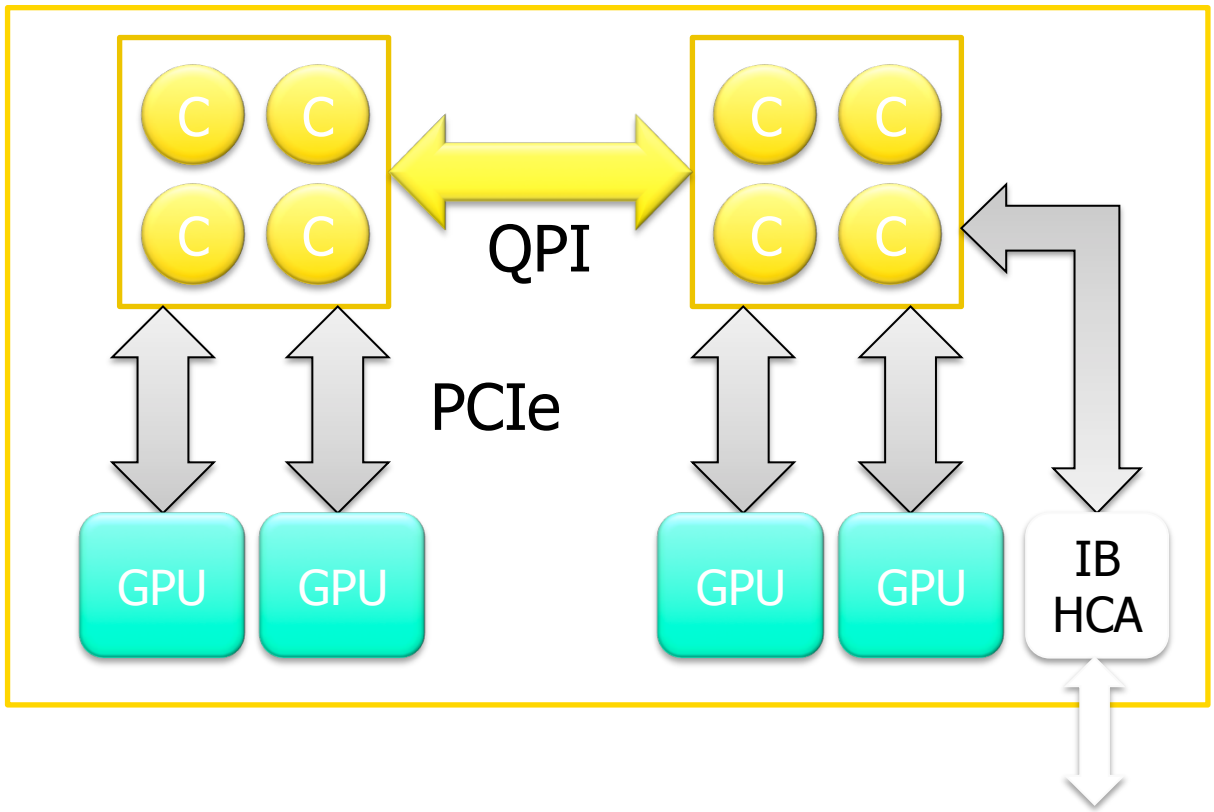
- **Base Cluster** 部は **2012/01**に完成
- **TCA** 部は**2013/03**完成予定（ただしテストシステムは**2012**前半）

- 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

補助スライド

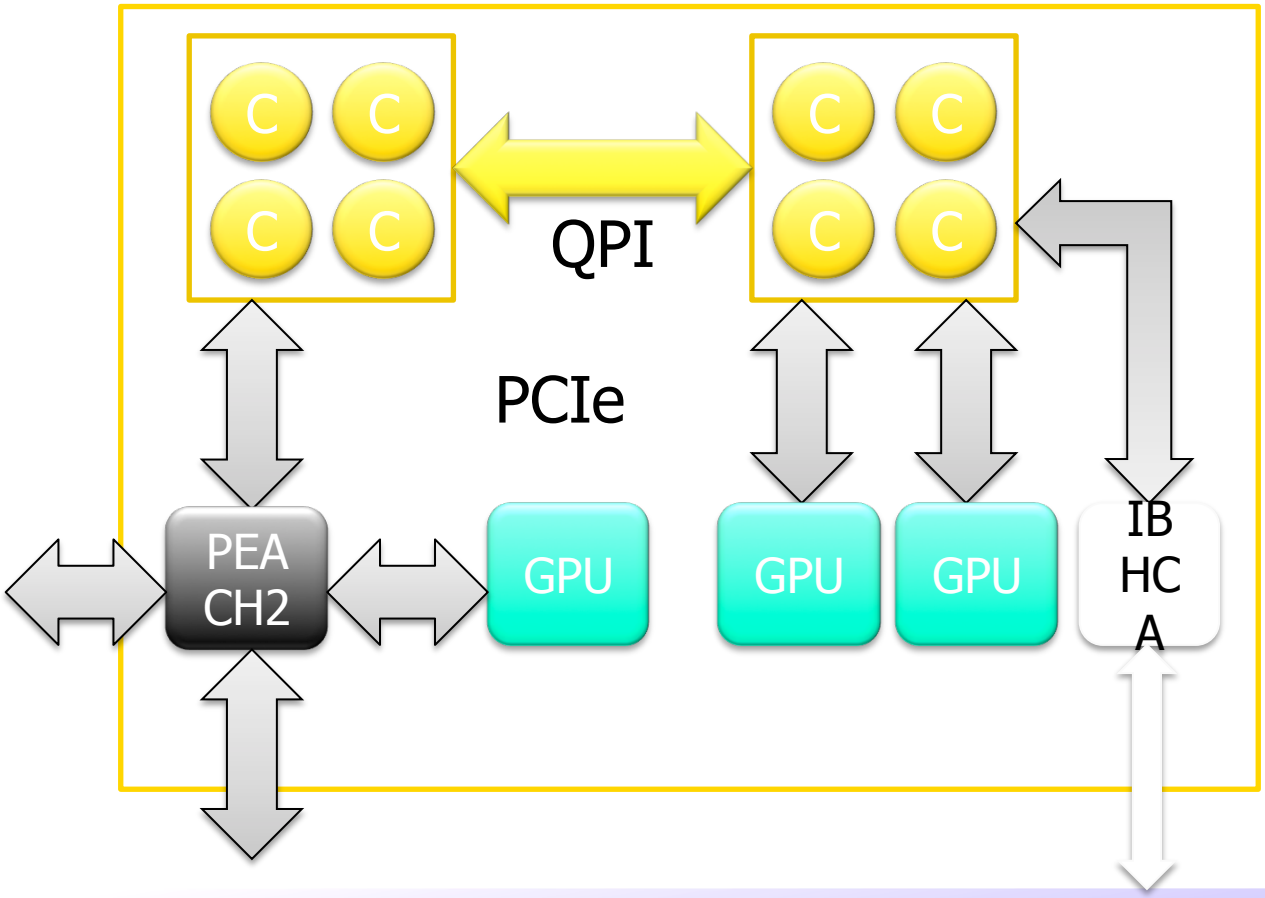
PEARLによるアクセラレータ直接結合

- PEARLなし



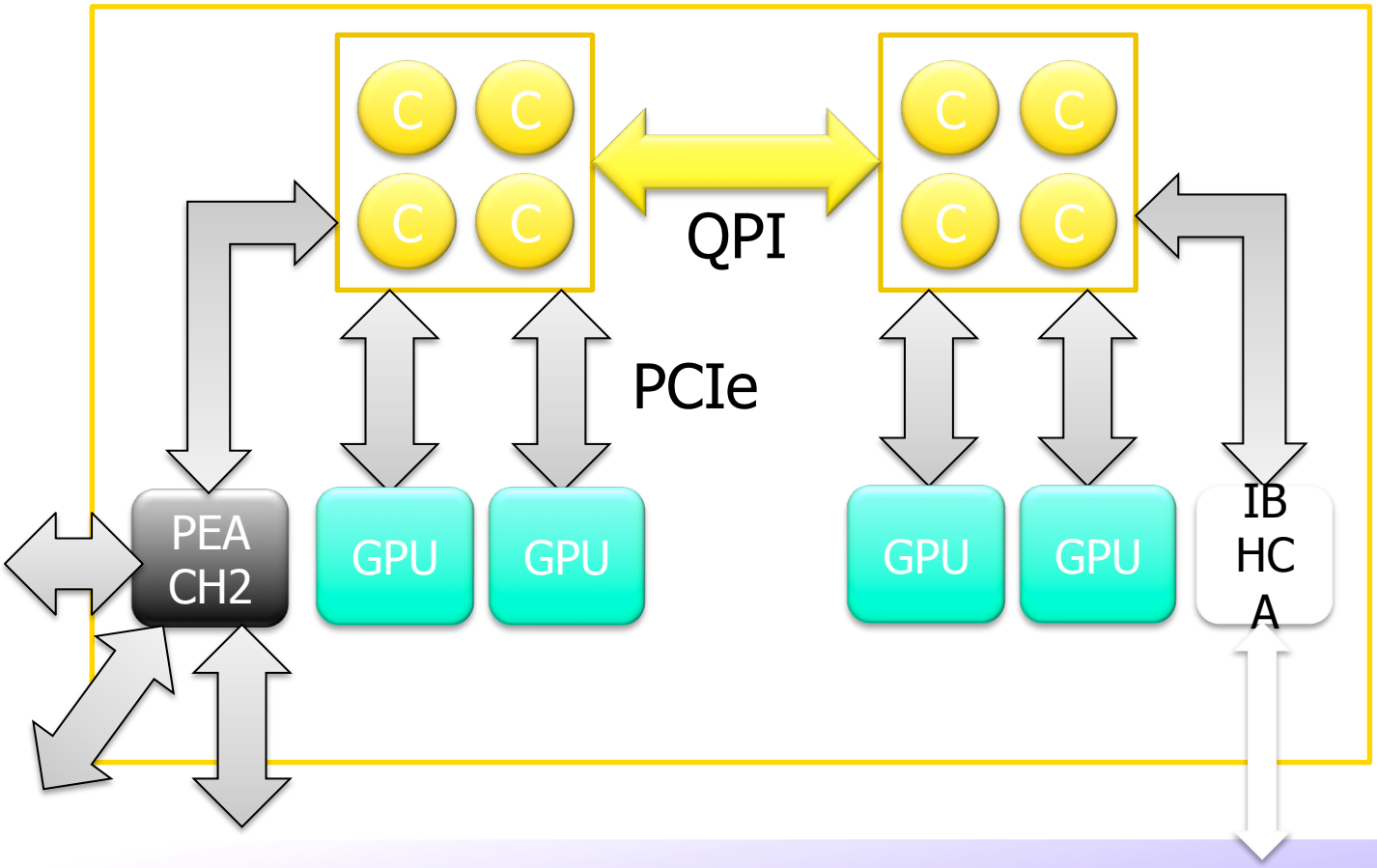
PEARLによるアクセラレータ直接結合

- GPU接続例 1

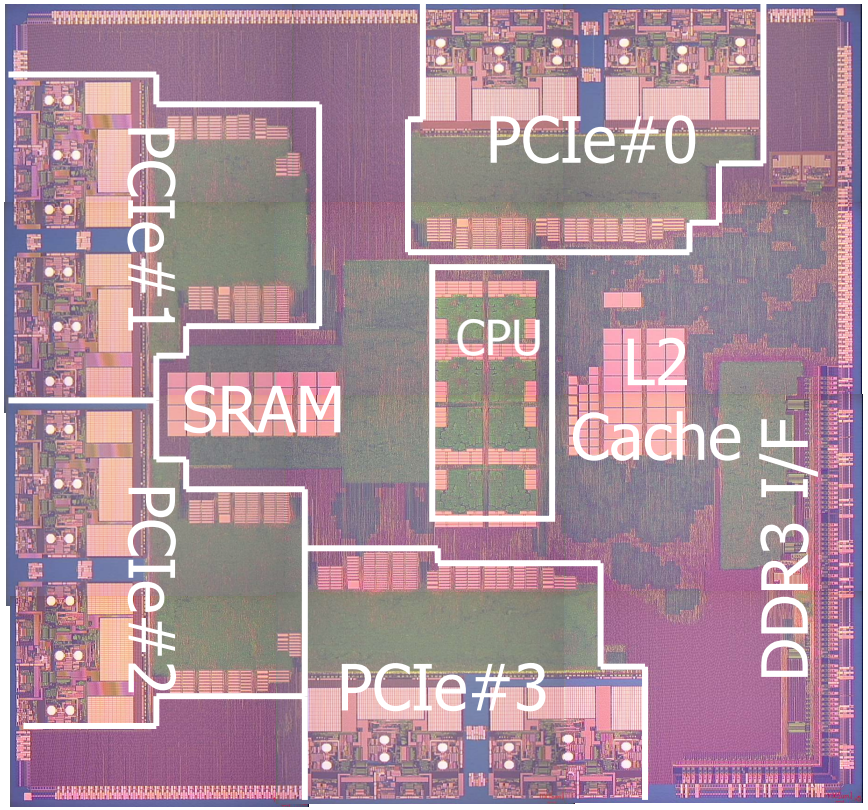


PEARLによるアクセラレータ直接結合

- GPU接続例２



PEACHチップ [Otani et al., ISSCC2011]



チップ諸元

- CPU: ルネサス M32R 4コア SMP (max. 400MHz)
- 通信リンク： PCI Express Gen2 x4 レーン (20Gbps) *4ポート
- プロセス： 45nm Low Power, triple-Vth, 8-layer metal
- チップサイズ: 11mm×11mm
- 消費電力: 3.2W
  - InfiniBandの2分の1以下
  - レーン速度選択による省電力機能

Speed		4 lanes	2 lanes	1 lane
Gen2	5 Gbps	1.00	0.50	0.28
Gen1	2.5 Gbps	0.84	0.42	0.24