JCAHPCスーパーコンピュータ Miyabi-Gのメモリ特性及び性能評価

筑波大学・計算科学研究センター/JCAHPC
朴　泰祐

背景

- これまでのCPUとGPUは独立したプロセッサだった
  - CPUとGPUのメモリが異なる、Cache-Coherentではないという点が課題
  - 分断されたメモリを扱うため、プログラミングが煩雑
    - CPU/GPUを分けたメモリ管理、明示的なデータ転送が必要
- **NVIDIA GH200 Grace Hopper Superchip**
  - CPUとGPUが1つのモジュールの上で密に結合
  - NVLink-C2Cによる高速で低レイテンシなデータ転送
- **Miyabi-G: 78.8P FLOPS, GH200 x 1120 nodes**
  - 筑波大学CCSと東京大学ITCが共同運営する最先端共同HPC基盤施設（JCAHPC）が導入・運用
  - 2025年1月運用開始

筑波大学CCSが運用する2台のH100ベースのスーパーコンピュータ

- Miyabi-G (GH200) vs Pegasus (Xeon+H100)

	Miyabi-G (1120 nodes)	Pegasus (150 nodes)
Operation started	Jan. 2025	Dec. 2022
CPU	NVIDIA Grace CPU (Arm Neoverse V2 CPU, 72core)	Intel Xeon Platinum 8468 2.1GHz (48core)
GPU	NVIDIA Hopper H100	NVIDIA H100 Tensor Core GPU with PCIe
Network	InfiniBand NDR200 x 1	InfiniBand NDR200 ×1
CPU･GPU connection	NVLinkC2C	PCIe Gen5 × 16Lane

Arm vs x86

同じHopper
(仕様は違う

同じ相互結合網

CPU-GPU接続が違う

Miyabi vs Pegasus: 性能仕様の違い

	Pegasus (theoretical peak)	Miyabi-G (theoretical peak)
CPU (FP64) [TFLOPS]	3.2256	3.456
GPU (FP64) [TFLOPS]	26	34
Memory Bandwidth (CPU) [GB/s]	282	512
Memory Capacity (CPU) [GiB]	111.7	128
L1d cache (CPU) [KB/core]	-	64
L2 cache (CPU) [MB/core]	-	1
Memory Bandwidth (GPU) [TB/s]	2	4.022
Memory Capacity (GPU) [GiB/s]	89.4	80
L1d cache (GPU) [MB / GPU]	29	34
L2 cache (GPU) [MiB / GPU]	50,00	60,00
GPU-CPU network (unidirection) [GB/s]	64	450
Inter-node Connection [Gbps]	200	200

× 1.31

× 1.82

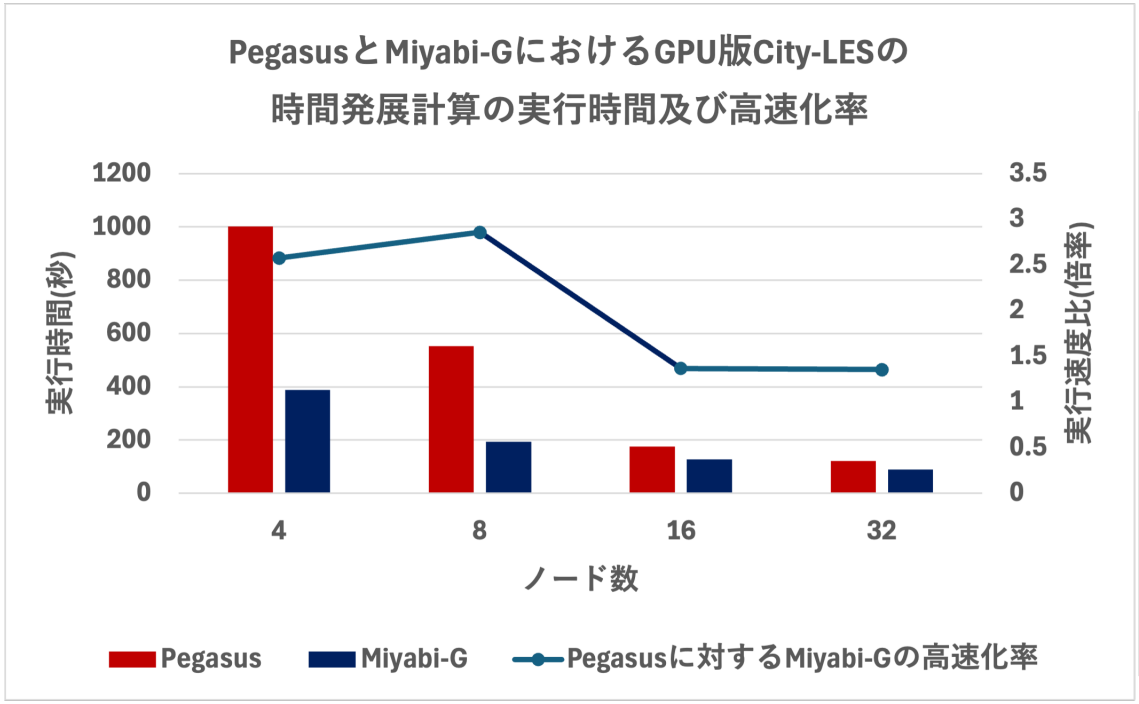
× 2.01

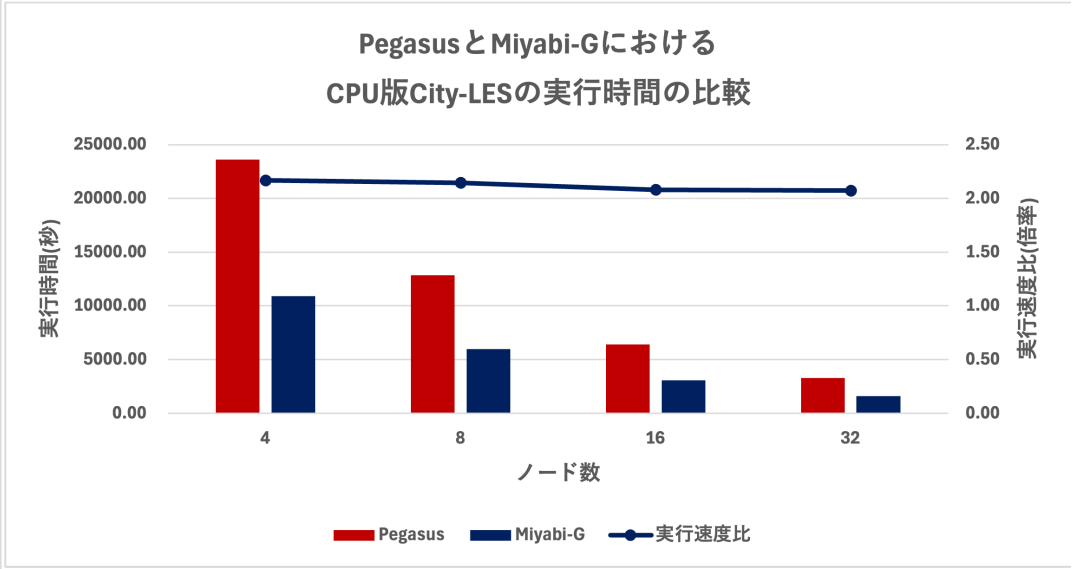
× 7.03

City-LES（都市気象シミュレーション）における性能差

- Miyabi-GはPegasusの1.5～3倍の性能
  - GPU memory b/w ⇨ 2x
- CPUのみのバージョンでもMiyabi-GはPegasusの2倍高速
  - CPU memory b/w ⇨ 1.8x, CPUコア数は48:72

※GPU版=CPU版の25倍高速 (4 nodes)

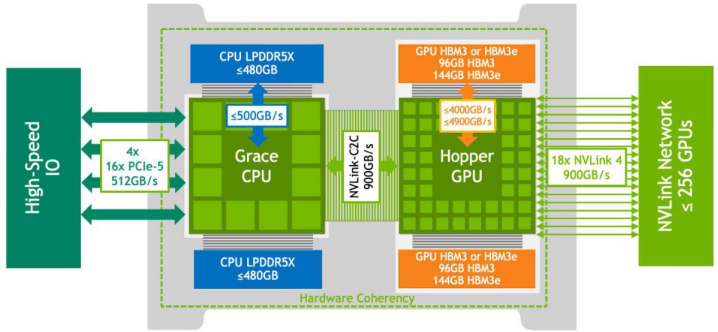




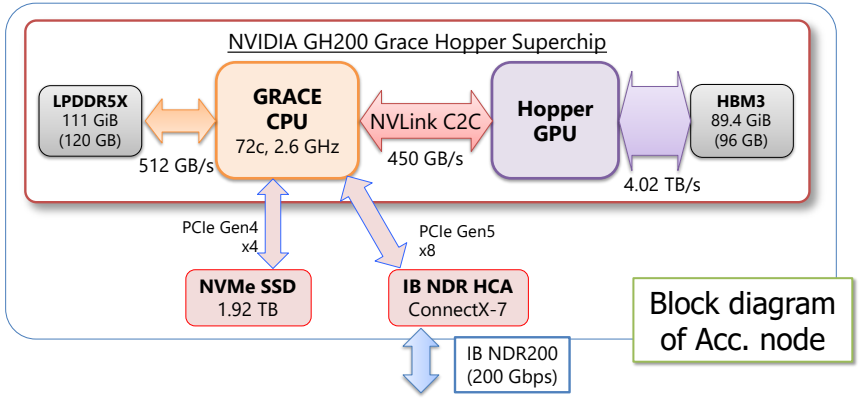
* 阿部, 佐藤, 朴, 藤田, 日下, "ドライミスト効果を持つ都市気象コードのGH200 vs Xeon+H100上の性能比較", 情報処理学会第199回HPC研究会, 柏, 2025年5月.

GH200のアーキテクチャ

- Grace CPU (ARM 72 cores)
  - アーキテクチャがARM、proprietary (x86依存)なソフトウェアは注意が必要
  - 128bit SVE SIMD, 3.4TFLOPS@3GHz
    - 富岳（A64FX）で動作実績のあるプログラムはそのまま実行できる
    - ただし、SVE幅が異なるため512bit幅を前提にしている場合は性能的に問題生じる可能性もあり
  - LPDDR5X Memory 120GB, 512GB/s
- Hopper GPU (H100)
  - 67 TFLOPS (DP, Tensor)
  - HBM3 Memory 4.02TB/s
  - CUDAのコアは、SXM版やPCIe版のH100同じ
    - CUDAやOpenACC等のプログラムは特に変更なく実行できる
  - CPU-GPU間がNVLink-C2Cで接続されているが、アプリ視点でPCIeバスと違いはない
    - NVLink-C2Cは高速なバス、 cudaMemcpy()が高速になると捉えて差し支えない
    - NVLink-C2C@450GB/s は PCIe Gen.5 x16@64GB/s より7倍高速



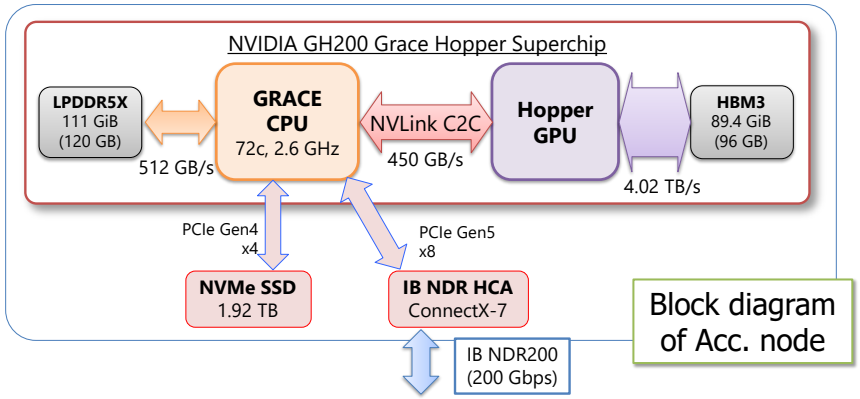
画像 : Data Sheetより引用
https://resources.nvidia.com/en-us-grace-cpu/grace-hopper-superchip



JCAHPC資料より引用
https://www.itc.u-tokyo.ac.jp/OFP-II/poster_OFP2-JCAHPC.pdf

GH200とSystem Allocated Memory

- CPUとGPUはアドレス空間を共有しCache-Coherent
- カーネルからはNUMAノードとして見える
  - CPUがNode 0, GPUがNode 1
  - CPU NUMA用のAPIを適用可能
  - 各NUMA domainのメモリ特性が全く異なる点が特徴
- System Allocated Memory（SAM）
  - 普通に確保したメモリ
  - CPUもGPUもアクセスでき、Cache-Coherent
  - First-Touchに基づくPage割当が行われる（CPU NUMAと同等）
  - メモリアクセスに基づくPage Migrationが行われる
- 従来と同様にCUDA APIを使う事も可能

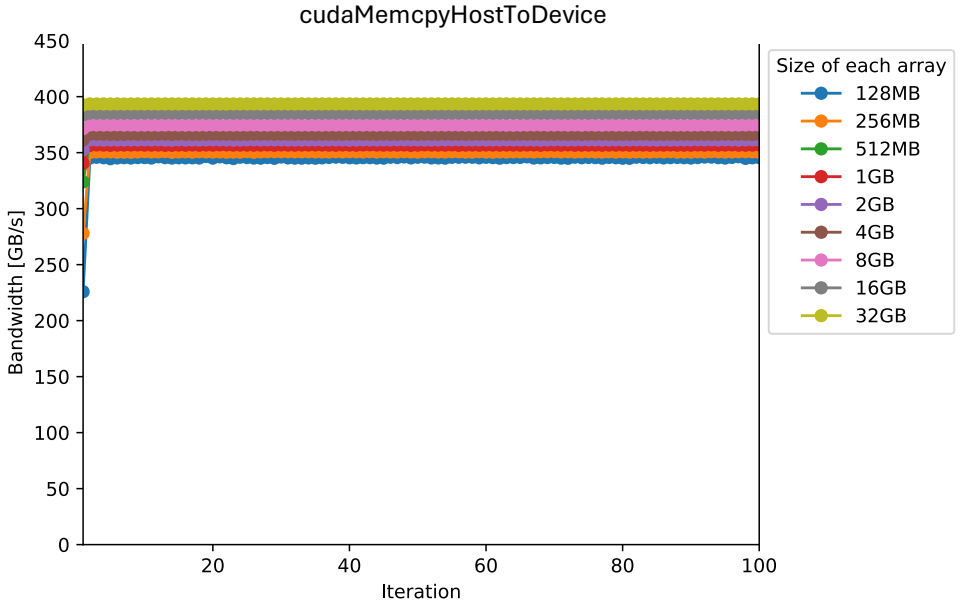


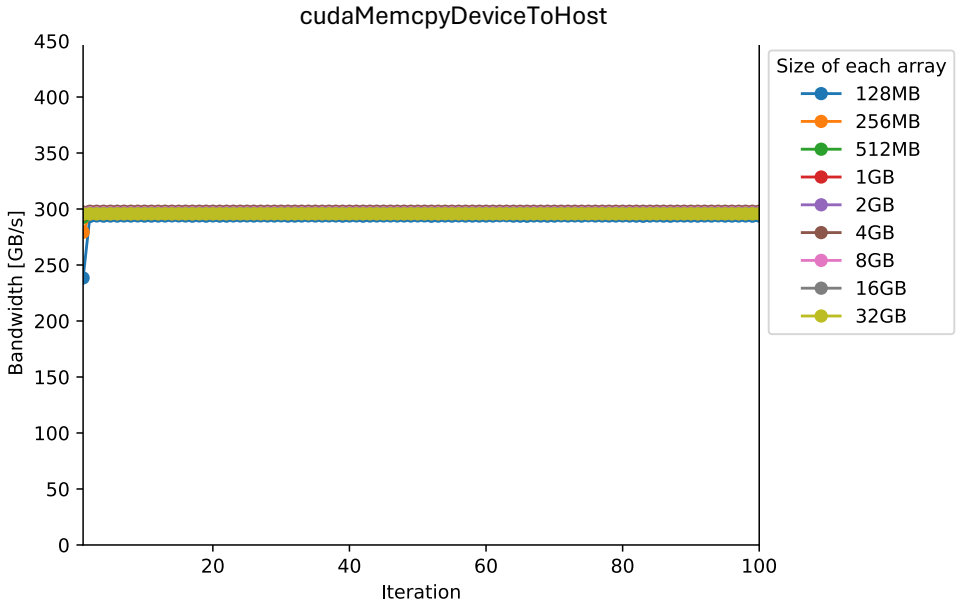
SAM Page Migration

- CPUメモリにあるMemory Pageに対して、GPUがアクセスするとGPUメモリに移動する場合がある（Page Migration）
  - ページ毎にカウンタがあり、条件を満たすとMigrationされる
  - 1回のアクセスでは移動せず、複数回のアクセスが必要
  - アプリケーションは何もすることなく、システムが自動的にデータを移動する
  - GPU⇔LPDDR5X@450GB/s が GPU⇔HBM3@4TB/s になり、次回以降のアクセスが高速に
- 逆方向（GPU→CPU）は発生していないと思われる
  - メモリアクセスが契機のGPU→CPUのMigrationは観測できていない
  - 公式ドキュメント等に、Page Migrationの詳細な情報がなく、詳細な調査は今後の課題
  - CUDAとOpenACCで振る舞いが違う？

NVLink-C2Cの性能評価

- CUDA APIを用いてNVLink-C2Cの性能評価を行う
  - SAMは使わず、従来手法を用いてメモリ管理をする際の性能(cudaMalloc+cudaMemcpy)
  - HostToDeviceで最大400GB/s、DeviceToHostで最大300GB/sの性能
  - 性能揺らぎがなく、サイズ毎に性能が安定
  - 1CUDA Streamで逐次実行
    - 多重化すればさらに性能が向上する可能性
      ⇨ ただしNVLINK-C2CとCPU memory性能で律速される

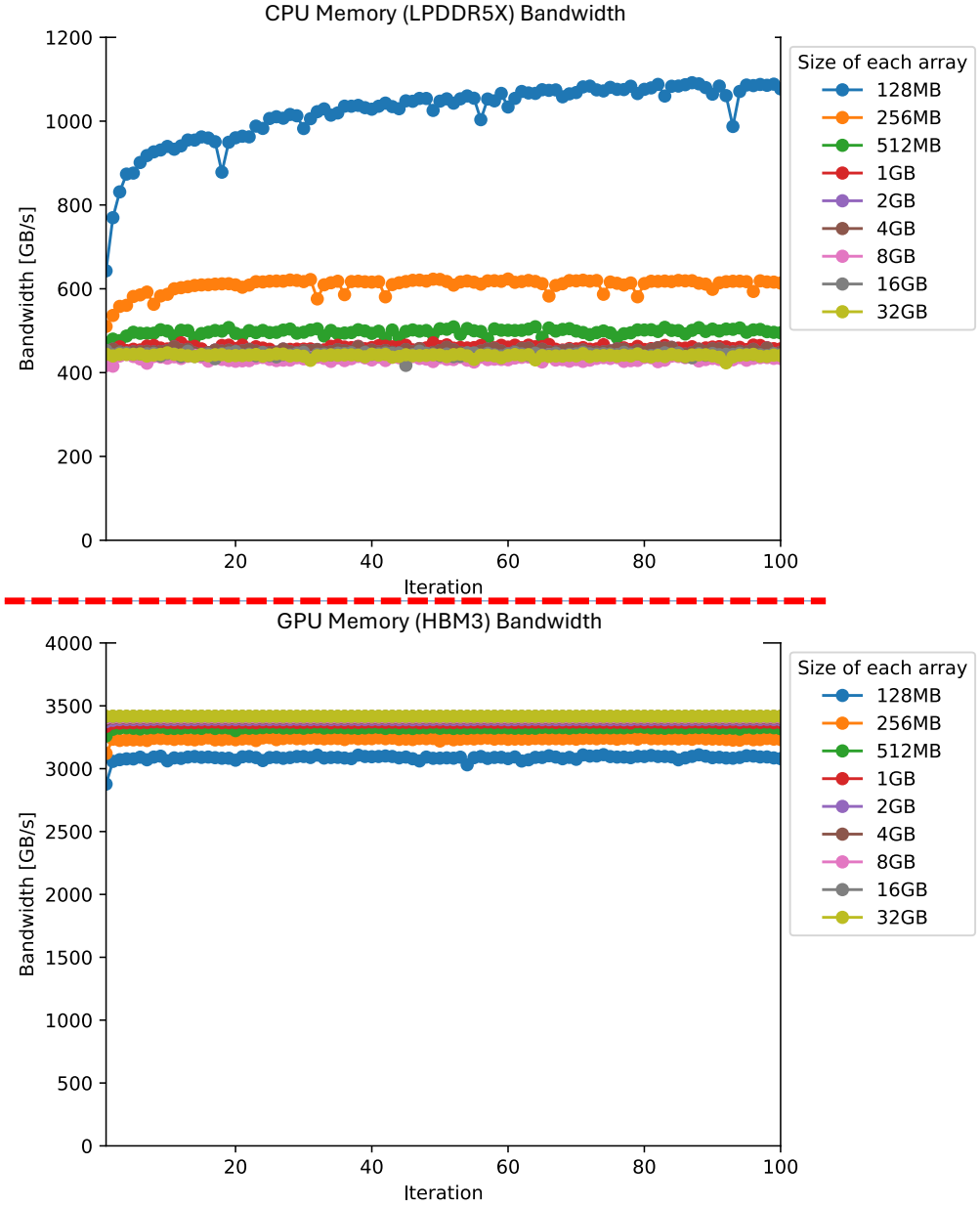




*吉田, 藤田, 白井, 朴, 辻, "NVIDIA GH200におけるSystem-Allocated Memoryの性能評価", 情報処理学会第199回HPC研究会, 柏, 2025年5月.

GH200メモリの性能評価

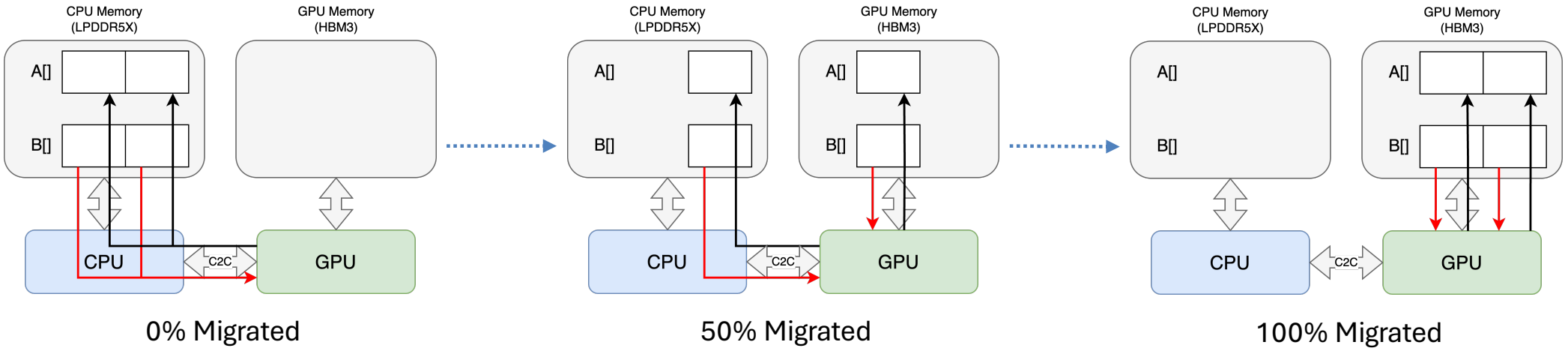
- CPUメモリの帯域
  - 約400GB/s~420GB/s
  - 小さいサイズで性能が高いのは、Cacheの影響と思われる
- GPUメモリの帯域
  - 配列が大きいほど性能が高く最大で約3.4TB/s
  - 性能揺らぎがなく、サイズ毎に性能が安定
- NVLink-C2Cの性能は450GB/s
  - CPUから見たら、LPDDR5XとHBM3の帯域はsame order



*吉田, 藤田, 白井, 朴, 辻, "NVIDIA GH200におけるSystem-Allocated Memoryの性能評価", 情報処理学会第199回HPC研究会, 柏, 2025年5月.

SAMとMigrationの性能評価

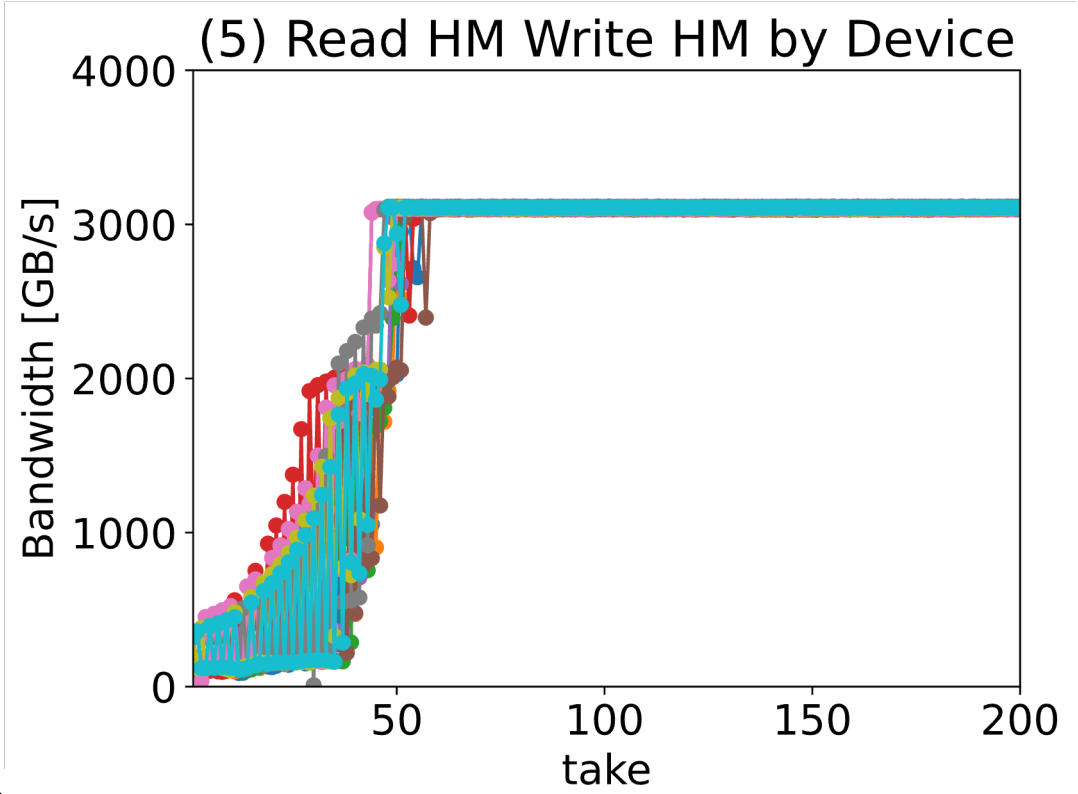
- CPU上に確保した配列で、GPUがコピーを行う
  - 最初はCPU上にある配列同士のコピーをGPUが行うため性能が悪い
  - Migrationが進むと、GPUメモリが利用され、性能が徐々に上がる
  - 全てMigrationされると、GPUメモリ同士のコピーになり、最大性能になる



Migration性能評価結果 (8GB data size)

- 配列コピーを合計100回実行し、それぞれの実行で性能を測定
  （グラフはこれを10回行ったものを重ねている）
  - 配列はmalloc() + first touchで最初はCPUメモリにある
  - GPUからread & write する
  - 回を重ねると性能が向上していき cudaMalloc() でHBMに確保したデータをGPUでアクセスした場合に近づく
    ⇨ 徐々にmigrateされていく
- migrate後の上限性能が通常のHBMアクセスより悪い (< 3.4TB/s)

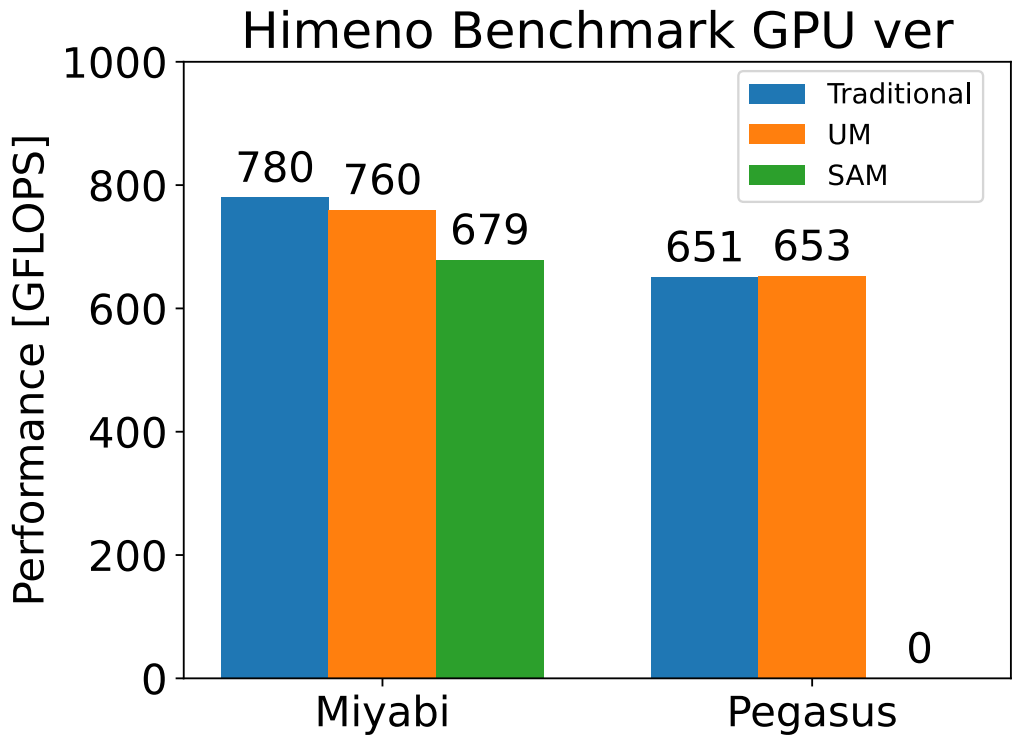
```
/* malloc(sizeof(double)*N) */
for(i=0;i < N; i++)
    A[i] = B[i];
```



*吉田, 藤田, 白井, 朴, 辻, "NVIDIA GH200におけるSystem-Allocated Memoryの性能評価", 情報処理学会第199回HPC研究会, 柏, 2025年5月.

Himeno Benchmark

- Himeno BenchmarkにSAMを適用
  - 問題サイズXL (512x512x1024)
  - 3000 Iterationで固定
    - SAMのMigrationは十分完了したと考えられる
- SAMの性能はTraditionalに及ばない
  - Copy Benchmarkで現れたデータサイズが大きいとSAMの性能が出にくい特性が影響していると考えられる
- SAMでは、一切のメモリ管理が不要という利点
- 今回は利用していないがOpenACCを用いれば、CPU OpenMPと同じ複雑度でGPUプログラミングが可能
- Himeno Benchmarkは構造が単純なのでより複雑なアクセスパターンのアプリケーションやdeep copyを必要とするケースなどの評価が必要

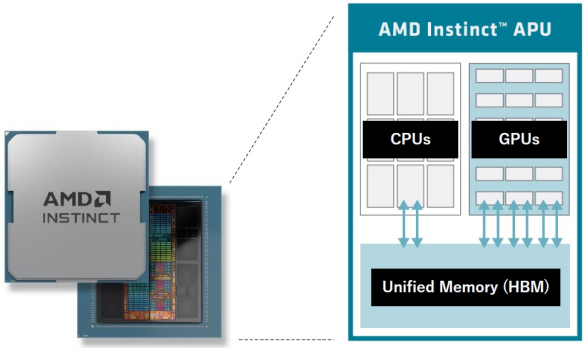


*吉田, 藤田, 白井, 朴, 辻, "NVIDIA GH200におけるSystem-Allocated Memoryの性能評価", 情報処理学会第199回HPC研究会, 柏, 2025年5月.

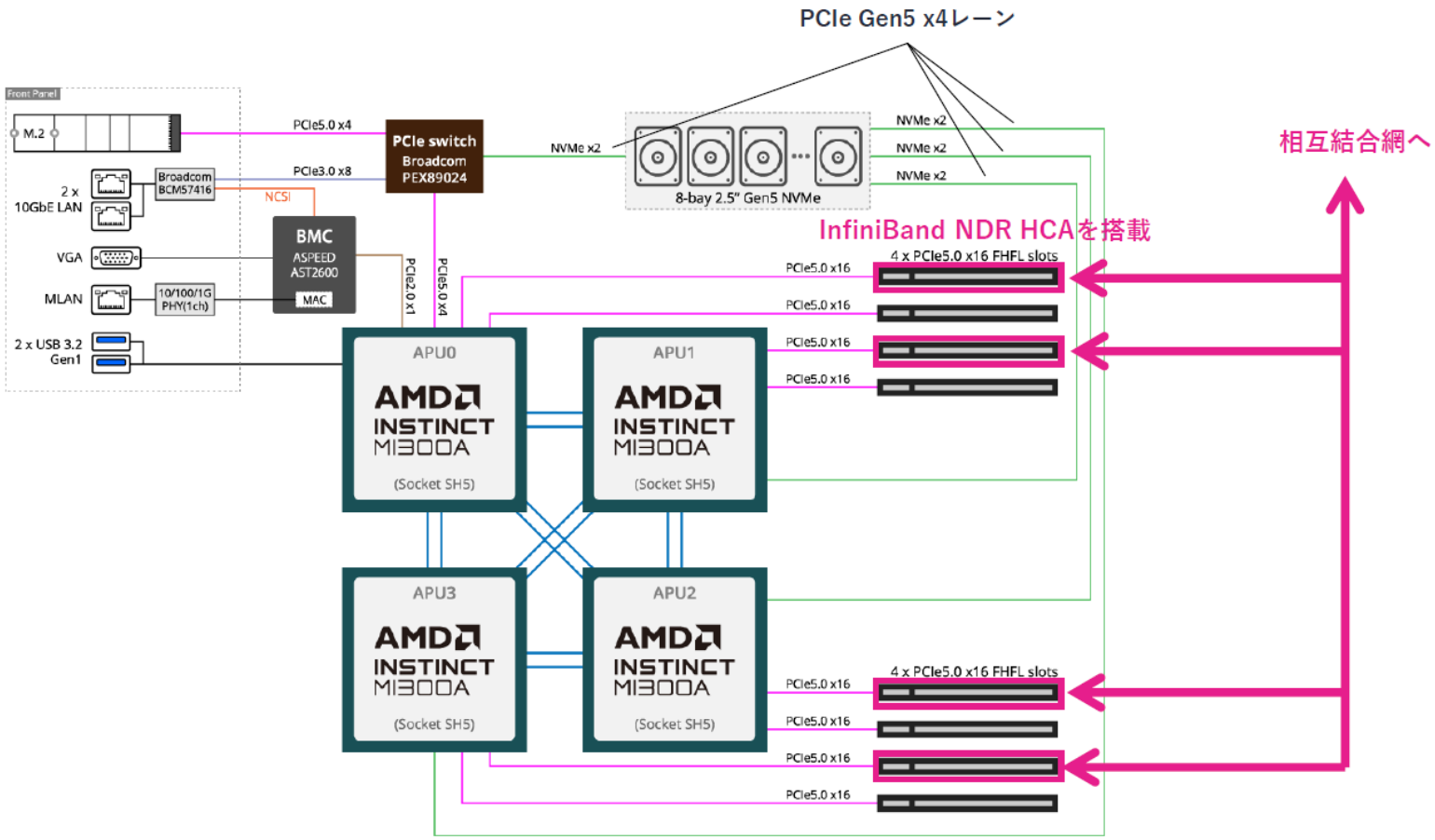
筑波大：Post Cygnusユニファイドメモリスーパーコンピュータ

- 2026年3月運用開始
- 総合性能
  - 24 nodes, 11.905 PFlops, 12 TiB HBM3
- ノード仕様
  - AMD Instinct MI300A APU x 4　（～EI Capitanの構成に近い）
    (24c EPYC Zen 4 CPU, 122.6 TFlops CDNA 3 GPU, 128GB HBM3)
  - 3.84 TB PCIe Gen5 NVMe SSD x 4
  - InfiniBand NDR (400 Gbps) x 4
- 並列ファイルシステム
  - 5.2 PByte DDN EXAScaler (100 GB/s)
    (DDN ES400NVX2 + SS9024 x 4)
- Prime vendor: NEC

NEC LX 401Bax-3GA

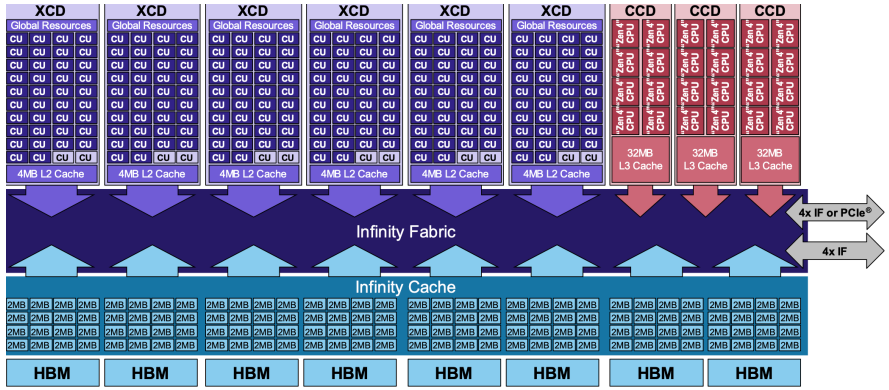


計算ノードブロックダイアグラム



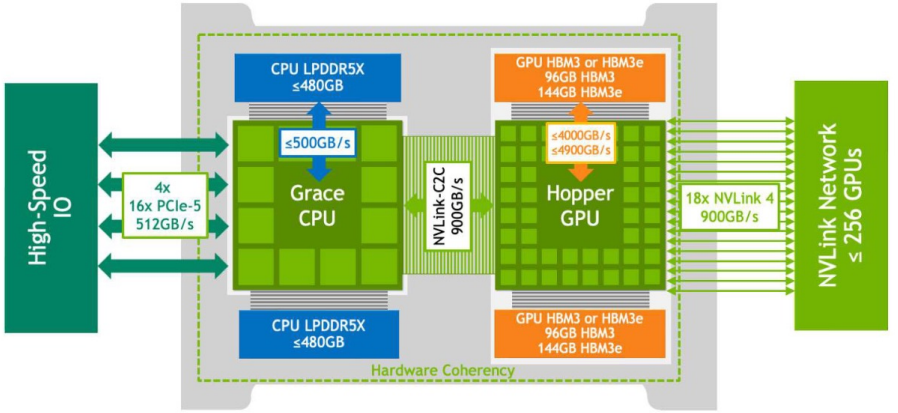
DDR-HBM NUMA (NVIDIA) vs flat-HBM (AMD) ?

- MI300A APU は flat HBM3 メモリをCPUとGPUが共有し、GH200のDDRLPX5 (CPU) + HBM3 (GPU) のNUMA構成とは全く違う
- 来年2月、異なるノード構成の2種類のGPUスーパーコンピュータを同時に運用する唯一の国立大学センターとなる
- それぞれの性能特性、アプリケーションとの相性などを詳細に調査する予定



AMD MI300A (post-Cygnus)

vs



NVIDIA GH200 (Miyabi-G)

おわりに

- 筑波大学CCSではスーパーコンピュータの一般利用での企業利用プログラムを開始します！（開始日は未定）
  - 成果公開型：これまでの一般利用とほぼ同じ料金
  - 成果非公開型：一般利用の３～４倍程度
  - Miyabi に関しては東大情報基盤センターと同じ料金
- 学際ハブ拠点事業における「スパコンお試し利用」も引き続き行っています！
- 詳しくはwebなどでのお知らせをチェック！