JCAHPCの新スパコンシステム Miyabiの準備状況 (Part-I)

朴泰祐　　塙敏博

最先端共同HPC基盤施設 (JCAHPC)

筑波大学計算科学研究センター

東京大学 情報基盤センター

JCAHPCことはじめ: T2Kオープンスパコンアライアンス (2008~2013)

Open Supercomputer T2K
Open Supercomputer Alliance

- 筑波大・東大・京大による次世代スパコン技術の推進のための同時調達システム
- 計算科学・計算工学における研究・教育・システム共用に関する大学のリーダーシップの確立
- 2008年6月、3システムが国内1,2,4位の性能を達成

- オープン ハードウェアアーキテクチャ（コモディティ技術による）
- オープン ソフトウェアスタック（オープンソースミドルウェアとツール）
- オープン な利用とユーザ知識・アプリケーションの共有

Kyoto Univ.
416 nodes (61.2TF) / 13TB
Linpack Result:
Rpeak= 61.2TF (416 nodes)
Rmax = 50.5TF

Univ. Tokyo
952 nodes (140.1TF) / 31TB
Linpack Result:
Rpeak= 113.1TF (512+256 nodes)
Rmax = 83.0TF

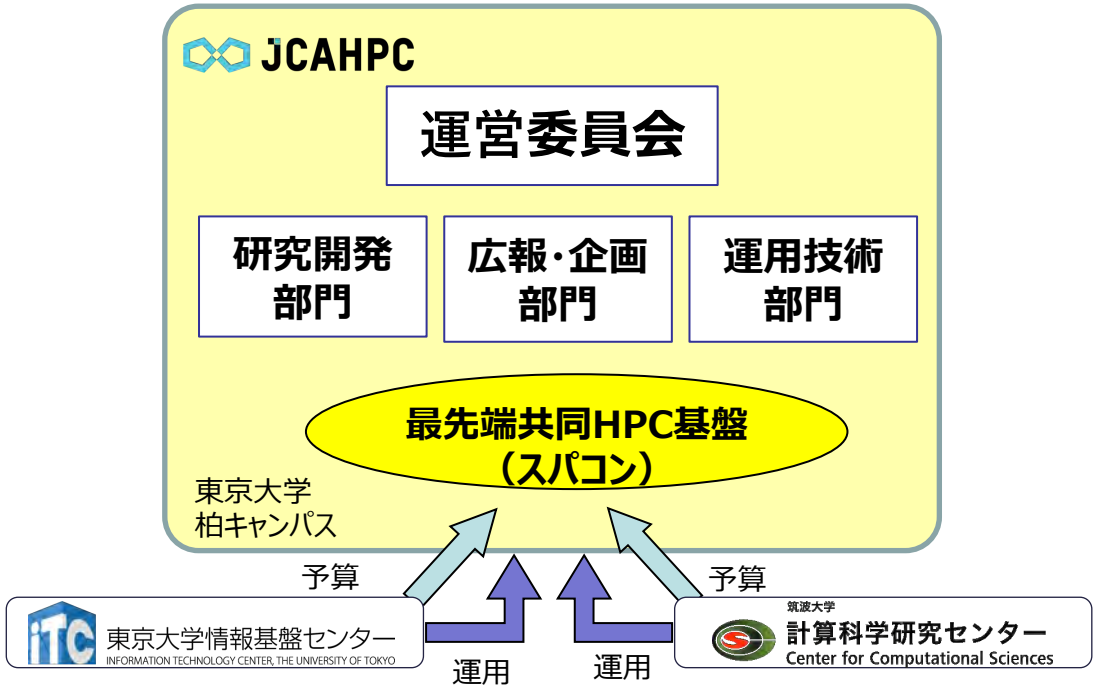
Univ. Tsukuba
648 nodes (95.4TF) / 20TB
Linpack Result:
Rpeak= 92.0TF (625 nodes)
Rmax = 76.5TF

T2KからJCAHPCへ

- T2Kオープンスパコンアライアンスの効果
  - 3台のスパコンは同時に調達・運用され，システム構築技術や性能チューニング技術等を共有し，大学間の強いHPC研究コミュニティの発展につながった。
- T2Kの後，異なる次期システム調達スケジュールやシステム開発ポリシーの違いにより，発展的に解消
  - 京大：4年間サイクルの調達
  - 筑波大：演算加速器系システムに傾注 (HA-PACS)
  - 東大：T2KだけでなくFACシステムとしてFX10を導入 (Oakleaf-FX, Oakbridge-FX)
- そしてOakforest-PACS@JCAHPC
  - 2013年，筑波大学と東京大学による新たなスパコン導入の枠組み→ **JCAHPC**
  - T2Kを越える，より強固な連携によるシステム調達→ **Oakforest-PACS (OFP)**

JCAHPCの体制

- 2大学が調達と運用に関して共に責任を持つ
  - 国内初の試み、世界的にも例を見ない
  - 日本で最大規模のシステムを実現
- 筑波大学と東京大学の間の密な連携・協力
  - 両センターの教員、技術職員が参加
  - 文科省下で前例のない試み
- 仕様に加え調達プロセスを一本化、単一のシステム



2024年度からのJCAHPC運営委員会の体制

歴代委員	施設長	副施設長	研究開発部門長	運用支援部門長/副部門長	広報・企画部門長	運営委員メンバー
第4期 (2024/4-)	朴泰祐（筑波大）	千葉滋（東京大）	中島研吾 (東京大)	塙敏博 (東京大)/ 建部修見 (筑波大)	高橋大介 (筑波大)	9名

JCAHPCの成果

- 2大学の主要スパコン予算を合算することで
  - 2016年11月、OFPは「京」コンピュータを抜き国内最高性能システムとなった（世界第6位）
  - HPCI一般課題及び両大学の各種プログラムで活用され、日本を代表するスパコンとして広く世界に認知された
- メニーコア・アーキテクチャを採用したOFPにより
  - メニーコア特有のアプリケーションチューニングのノウハウ
  - 「富岳」の基本的なシステムソフトウェア開発に貢献
    - メニーコア向けOS McKernelの開発
    - メニーコア向けOmni並列コンパイラの開発
- 大規模高性能システムとして
  - 「京」から「富岳」への端境期の資源不足をカバー、数々の大規模アプリケーションの開発
  - 国際協力：LBNL/NERSC (Cori)、KISTI (Nurion)

⇒ Post-OFPシステムを引き続きJCAHPCとして共同調達・運営することを決定

スパコンの変遷@筑波大CCS＆東大ITC

JCAHPC

FY13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

HA-PACS Cray
802 TFLOPS+364 TFLOPS

COMA Cray
1 PFLOPS

T2K Tsukuba
Cray-Appro
95TFLOPS

T2K Tokyo Hitachi
40TF, 31.3TB

Oakleaf-FX: Fujitsu PRIMEHPC FX10, SPARC64 IXfx
1.13 PFLOPS, 150 TB

Oakbridge-FX
136.2 TFLOPS, 18.4 TB

Yayoi: Hitachi SR16000/M1 IBM Power-7
54.9 TFLOPS, 11.2 TB

Cygnus NEC
Intel Xeon SKX + NVIDIA V100 + Intel Stratix10
2.4 PFLOPS + 64 FPGA

Post-Cygnus ???
CPU+GPU unified memory system

Pegasus NEC
Intel Xeon SRP + DCPMM + NVIDIA H100
6 PFLOPS

Oakforest-PACS (OFP) Fujitsu, Intel Xeon Phi
25PFLOPS, 919.3TB

Miyabi (OFP-II) Fujitsu
NVIDIA GH200, Intel SRP-HBM
78.8PF + 1.3PF

Wisteria-Mercury (4yrs)
Intel Xeon SPR + NVIDIA H100

Wisteria/BDEC-01 Fujitsu,
FX1000+(Intel ICX+NVIDIA A100)
Integration of "Simulation, Data, and Learning"
33.1 PFLOPS

BDEC-02
150+ PF

Oakbridge-CX Fujitsu
Intel Xeon CLX
6.61 PFLOPS

Reedbush-U/H, HPE
Intel BDW + NVIDIA P100
1.93 PFLOPS

Reedbush-L HPE
1.43 PFLOPS

mdx Fujitsu
"Data platform"

JCAHPC第2期システム共同調達のポリシー

- システムの基本仕様
  - 演算加速装置(GPU)を含む超並列PCクラスタ
  - HPC・AI向け:最先端プロセッサ＋GPU
  - GPU搭載部とCPU部の2つのサブシステムを相互結合網と共有ファイルシステムで結合
    → HPCとAIの双方で絶対的な性能を持つGPU搭載部をメインに
    → GPU化しづらいアプリケーションの継続的吸収のためCPU部も
  - フラットに利用可能な柔軟性を持つ相互結合網
    → 共有ファイルシステムへのアクセス、スケジューリングの容易さ
- スケールメリットを活かす
  - 超大規模な単一ジョブ実行も可能とする
  - 大学基盤センター・スパコンセンターとして最大規模のシステムへ
- GPUへの移行をスムーズに行うため、まずGPU種をプレベンチマークにより決定
  - ユーザの準備期間の確保
  - GPUベンダーの協力の下、GPU移植作業を先行して開始

GPU移植・移行の計画

- NVIDIA Japanの協力
- 3,000人以上のOFP利用者：2つの形態
- 「自己移植（Self Porting）」：様々なオプション
  - 1週間のハッカソン（ミニキャンプ），3ヶ月に1回，オンライン・ハイブリッド，Slack併用
  - 毎月開催される「相談会」（Zoom，非ユーザーも自由に参加できる）
  - 素晴らしく充実した「移行ポータルサイト」，各種講習会
    - https://jcahpc.github.io/gpu_porting/
- **「サポート移植（Surpported Porting）」，2022年10月開始**
  - **多くのユーザーを有するコミュニティコード（17種類，次頁），OpenFOAM（NVIDIA）**
  - **外注のための予算も確保（落札ベンダーが担当する予定）**
  - **「サポート移植」グループメンバー（主に若手）はハッカソン・相談会にも積極的に参加**
- **基本的にOpenACC/StdPar（Standard Parallelism）推奨**

Miyabi システム・ハイライト

- Miyabi = Miyabi-G + Miyabi-C + Interconnect + File System + misc.
- Miyabi-G: GPU・CPU一体型ノード
  - NVIDIA GH200 (Grace-Hopper Superchip) x 1120 nodes
  - Theoretical peak (FP64): 78.8 PFLOPS (GPU+CPU)
- Miyabi-C: CPUノード
  - Intel Xeon CPU Max 9480 x 2sockets x 190 nodes
  - Theoretical peak (FP64): 1.3 PFLOPS
- システム総性能：80.1 PFLOPS
- 2023年11月：富士通により落札・契約
  ⇒ 2024年12月納入完了
  ⇒ 2025年1月より稼働開始

PCCC Workshop in すずかけ台
2024年6月28日(金)

JCAHPCの新スーパーコンピュータ Miyabiの準備状況

塙　敏博　　朴　泰祐

最先端共同HPC基盤施設 (JCAHPC)

東京大学 情報基盤センター

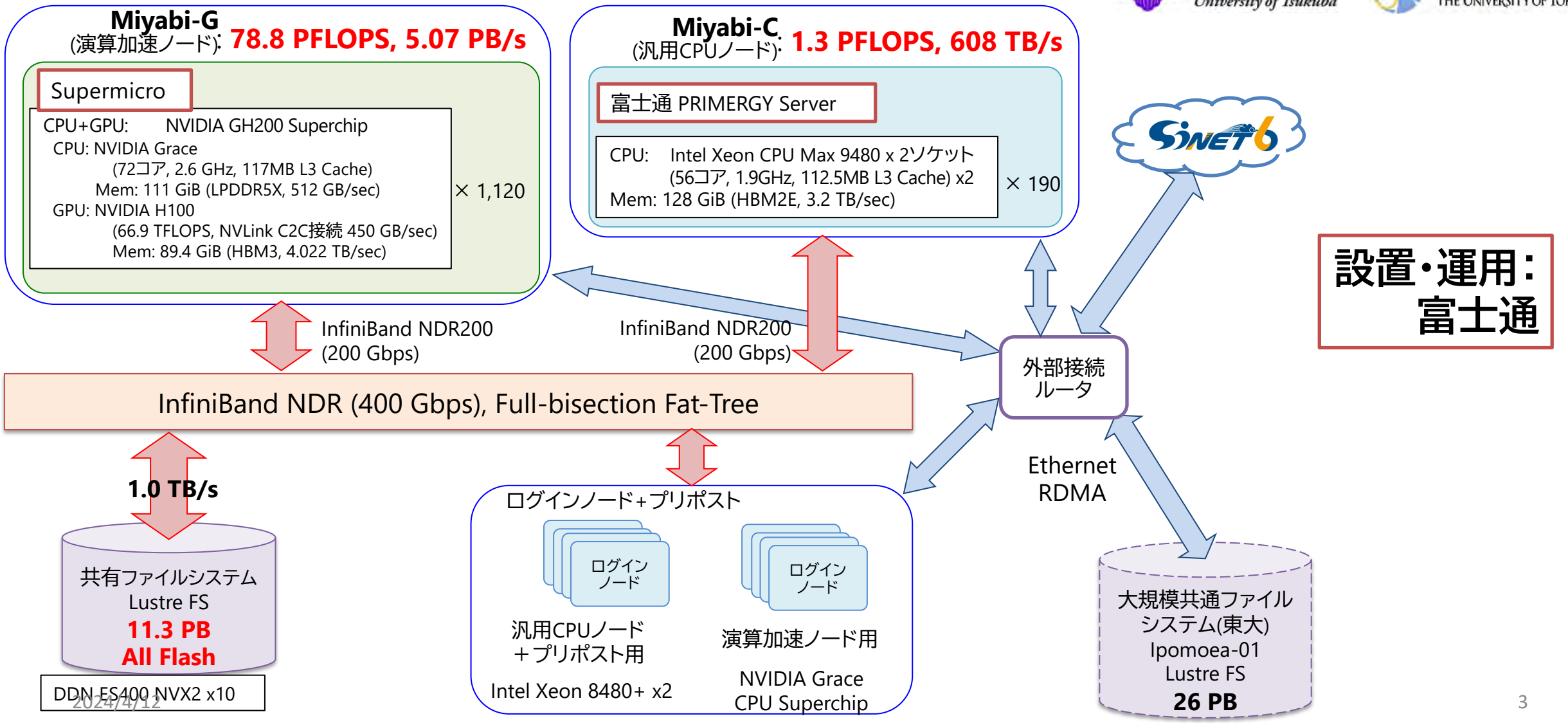
筑波大学 計算科学研究センター

Miyabiについて

- 2024年4月名称を公開
- “Miyabi”という名前には、単に理論性能が優れているというだけでなく、その持てる能力を難なく発揮できるように、という思いが込められています。
- 両センターは緊密な協力のもと新システム Miyabiの導入および運用を遂行し，「計算・データ・学習」融合による，高度に洗練された計算科学を推進することによって，安心・安全・信頼に基づく社会の実現に貢献していきます。

Miyabi (OFP-II)の概要 (1/3)
2025年1月運用開始

JCAHPC
筑波大学 University of Tsukuba
東京大学 THE UNIVERSITY OF TOKYO

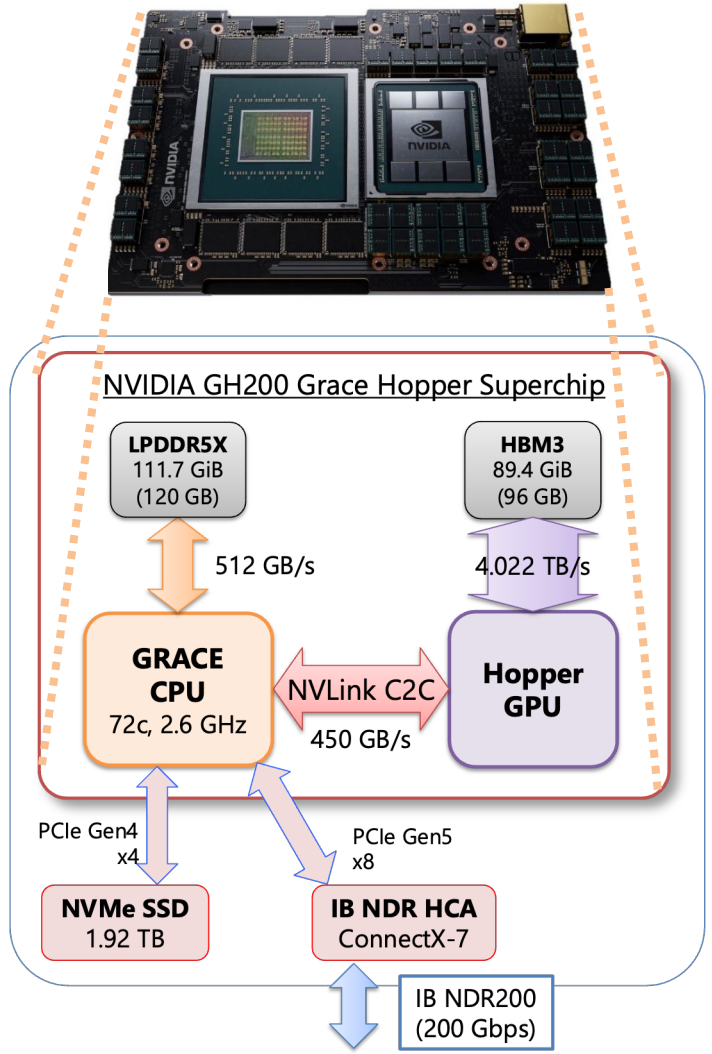


設置・運用：富士通

Miyabi(OFP-II)の概要 (2/3)

JCAHPC
筑波大学 University of Tsukuba
東京大学 THE UNIVERSITY OF TOKYO

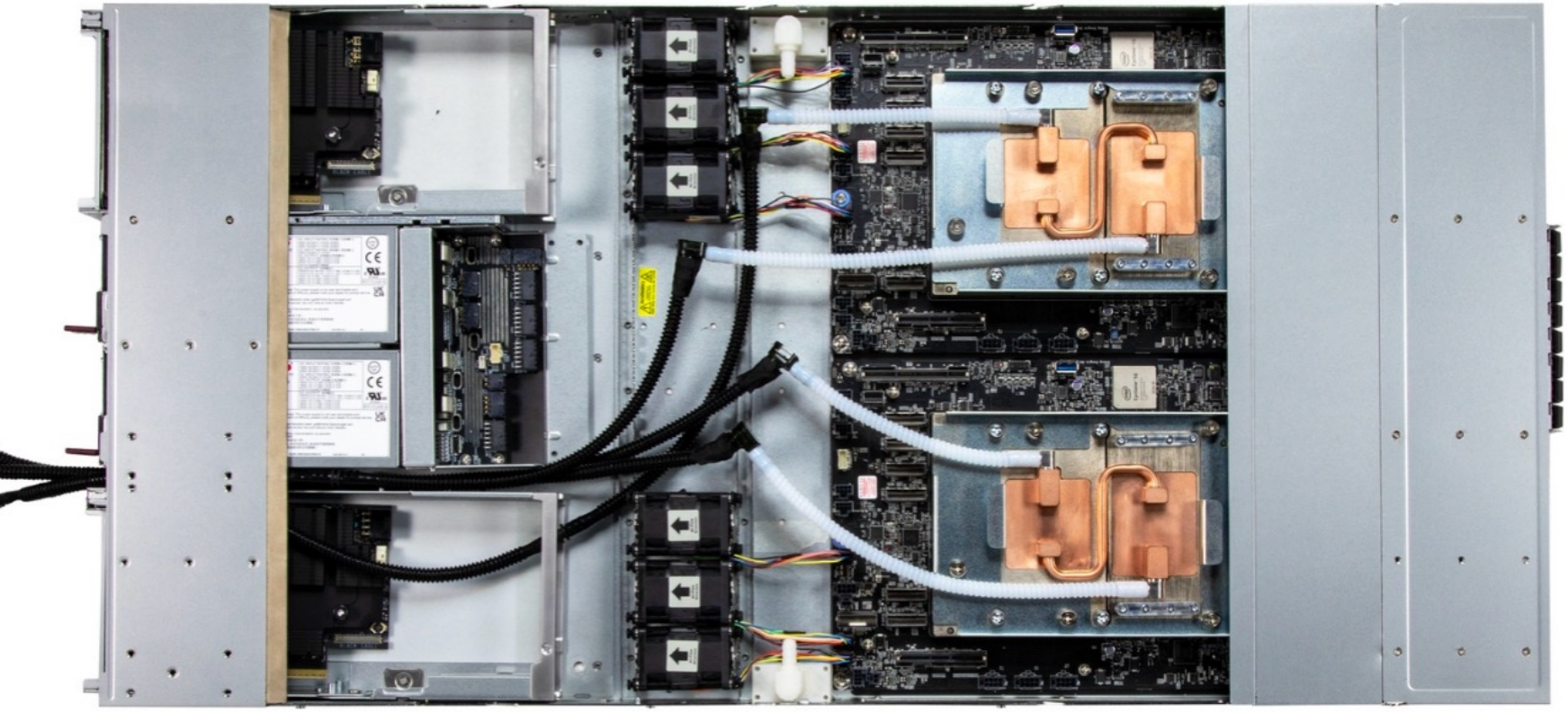
- **Miyabi-G: 演算加速ノード: NVIDIA GH200**
  - 計算ノード: NVIDIA GH200 Grace-Hopper Superchip
    - Grace: 72c, 3.45 TF, 120 GB, 512 GB/sec (LPDDR5X)
    - H100: 66.9 TF DP-Tensor Core, 96 GB, 4,022 GB/sec (HBM3)
      - CPU-GPU間はキャッシュコヒーレント
    - NVMe SSD for each GPU: 1.9TB, 8.0GB/sec, GPUDirect Storage
  - **合計 (CPU+GPUの合計値)**
    - **1,120 ノード, 78.8 PF, 5.07 PB/sec, IB-NDR 200**
- **Miyabi-C: 汎用CPUノード: Intel Xeon Max 9480 (SPR)**
  - 計算ノード: Intel Xeon Max 9480 (1.9 GHz, 56c) x 2
    - 6.8 TF, 128 GiB, 3,200 GB/sec (HBM2e only)
  - 合計
    - **190 ノード, 1.3 PF, IB-NDR 200**
    - **372 TB/sec for STREAM Triad (Peak: 608 TB/sec)**

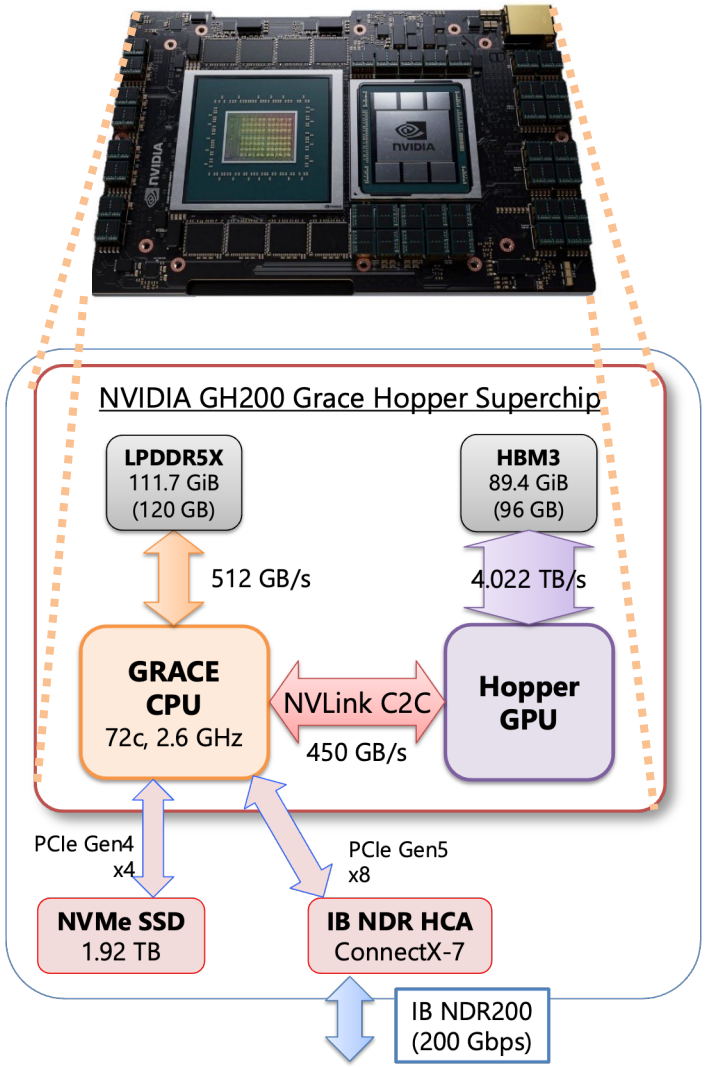


Miyabi(OFP-II)の概要 (2/3)

JCAHPC
筑波大学 University of Tsukuba
東京大学 THE UNIVERSITY OF TOKYO

- Supermicro ARS-111GL-DNHR-LCC
  - 1U 2ノード、直接水冷

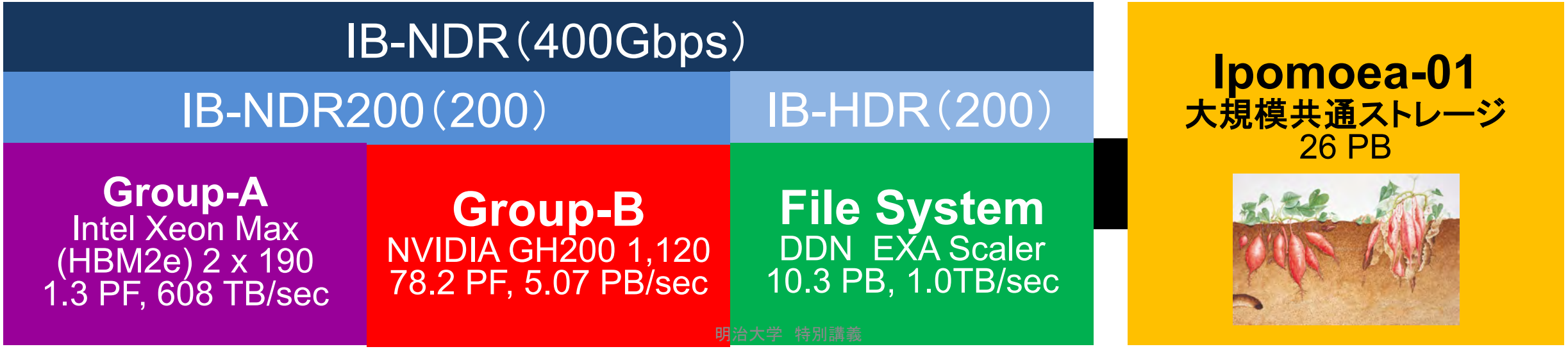




Miyabi(OFP-II)の概要 (3/3)

JCAHPC
筑波大学 University of Tsukuba
東京大学 THE UNIVERSITY OF TOKYO

- ファイルシステム: **DDN EXAScalar, Lustre FS**
  - 10.3 PB (NVMe SSD) 1.0TB/sec, “Ipomoea-01” (26 PB) も利用可能
- **Group-A/B の全ノードはフルバイセクションバンド幅Fat Treeで接続**
  - **(400Gbps/8)×(32×20＋16×1) = 32.8 TB/sec**
- **2025年1月運用開始、Group-A/B間の通信は h3-Open-SYS/WaitIO により実現**



Miyabi(OFP-II) 仕様まとめ

JCAHPC

	演算加速ノード	汎用CPUノード
理論ピーク性能	78.8 PFLOPS	1.29 PFLOPS
ノード数	1,120	190
合計メモリ容量	241.9 TB	23.75 TiB
合計メモリバンド幅	5.07 PB/sec	608 TB/sec
インタコネクト トポロジ	InfiniBand NDR200 (200 Gbps) Full-bisection Fat Tree	

共有ファイルシステム		Lustre FS
MDS	サーバ	DDN ES400NVX2
	サーバ数(VM)	1 (4)
	inode数	appx. 23.5 B
OSS	サーバ	DDN ES400NVX2
	サーバ数	10 set
	容量	11.3 PB (All Flash)
	理論バンド幅	1.0 TB/sec

項目		演算加速ノード	汎用CPUノード
サーバ		Supermicro ARS-111GL-DNHR-LCC	FUJITSU Server PRIMERGY CX2550 M7
CPU	プロセッサ名	NVIDIA GH200 Grace Hopper Superchip, NVIDIA Grace	Intel Xeon CPU Max 9480（Sapphire Rapids）
	プロセッサ数 (コア数)	1 (72)	2 (56+56)
	周波数	3.0 GHz	1.9 GHz
	理論演算性能	3.456 TFLOPS	6.8096 TFLOPS
	メモリ	LPDDR5X	HBM2E
	メモリ容量	120 GB	128 GiB
	メモリ帯域幅	512 GB/s	3.2 TB/s
GPU	プロセッサ名	NVIDIA Hopper	-
	プロセッサ数	1	
	SM数	132	
	理論演算性能	66.9 TFLOPS	
	メモリ	HBM3	
	メモリ容量	96 GB	
	メモリ帯域幅	4.02 TB/s	
	CPU-GPU間接続	NVLink C2C 450 GB/sec キャッシュコヒーレント	
NVMe SSD		1.92 TB, PCIe Gen4 x4	-

GPU移植・移行の計画

- NVIDIA Japanの協力
- 3,000人以上のOFP利用者：2つの形態
- 「自己移植（Self Porting）」：様々なオプション
  - 1週間のハッカソン（ミニキャンプ），3ヶ月に1回，オンライン・ハイブリッド，Slack併用
  - 毎月開催される「相談会」（Zoom，非ユーザーも自由に参加できる）
  - 素晴らしく充実した「移行ポータルサイト」，各種講習会
    - https://jcahpc.github.io/gpu_porting/
- **「サポート移植（Surpported Porting）」，2022年10月開始**
  - **多くのユーザーを有するコミュニティコード（19種類，次頁），OpenFOAM（NVIDIA）**
  - **外注のための予算も確保（富士通が担当する予定）**
  - **「サポート移植」グループメンバー（主に若手）はハッカソン・相談会にも積極的に参加**
- **基本的にOpenACC/StdPar（Standard Parallelism）推奨**

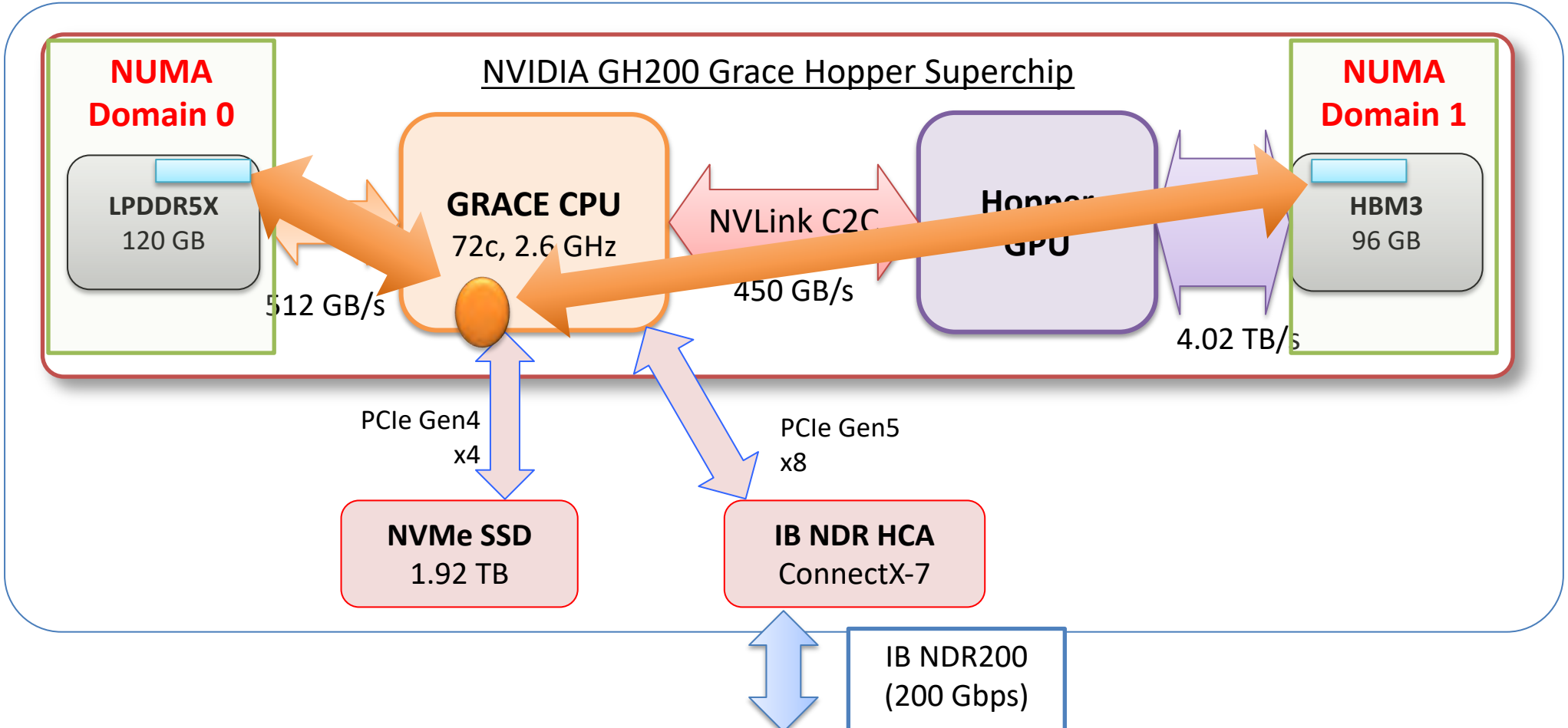
Category	Name (Organizations)	Target, Method etc.	Language
Engineering (5)	FrontISTR (U.Tokyo)	Solid Mechanics, FEM	Fortran
	FrontFlow/blue (FFB) (U.Tokyo)	CFD, FEM	Fortran
	FrontFlow/red (AFFr) (Advanced Soft)	CFD, FVM	Fortran
	FFX (U.Tokyo)	CFD, Lattice Boltzmann Method (LBM)	Fortran
	CUBE (Kobe U./RIKEN)	CFD, Hierarchical Cartesian Grid	Fortran
Biophysics (3)	ABINIT-MP (Rikkyo U.)	Drug Discovery etc., FMO	Fortran
	UT-Heart (UT Heart, U.Tokyo)	Heart Simulation, FEM etc.	Fortran, C
	Lynx (Simula, U.Tokyo)	Cardiac Electrophysiology, FVM	C
Physics (3)	MUTSU/iHallMHD3D (NIFS)	Turbulent MHD, FFT	Fortran
	Nucl_TDDFT (Tokyo Tech)	Nuclear Physics, Time Dependent DFT	Fortran
	Athena++ (Tohoku U. etc.)	Astrophysics/MHD, FVM/AMR	C++
Climate/ Weather/ Ocean (4)	SCALE (RIKEN)	Climate/Weather, FVM	Fortran
	NICAM (U.Tokyo, RIKEN, NIES)	Global Climate, FVM	Fortran
	MIROC-GCM (AORI/U.Tokyo)	Atmospheric Science, FFT etc.	Fortran77
	Kinaco (AORI/U.Tokyo)	Ocean Science, FDM	Fortran
Earthquake (4)	OpenSWPC (ERI/U.Tokyo)	Earthquake Wave Propagation, FDM	Fortran
	SPECFEM3D (Kyoto U.)	Earthquake Simulations, Spectral FEM	Fortran
	hbi_hacapk (JAMSTEC, U.Tokyo)	Earthquake Simulations, H-Matrix	Fortran
	sse_3d (NIED)	Earthquake Science, BEM (CUDA Fortran)	Fortran

Miyabiの特徴

- CG1 (GH200, Grace-Hopper: 1 socket/node)
  - 他の多くのGH200システムは 4ソケット/ノード
    - Cray/HPE: CSCS Alps, LANL Venado, Bristol Isambard AI, Cyfronet Helios,…
    - Eviden/Bull/Atos: Julich JUPITER, CEA Jules Verne
  - TACC “Vista” by Dell, CG1 (Gigabyte製？)
- InfiniBand NDR200, Fat-tree フルバイセクションBW
  - Cray/HPE: SlingShot
  - Eviden: BXIv2 (CEA-JV), IB-NDR200 (JUPITER)
  - InfiniBand: Vista、トポロジ不明
- GPUDirect Storageが利用可能 ➔ノード内NVMe-SSD, Lustre Filesystem (All flash)
  - SlingShot は未サポート (BXIもおそらく未サポート)
- Intel Xeon Sapphire Rapids HBMの計算ノードと密結合
- 富士通による設計と運用、Supermicro製計算ノード

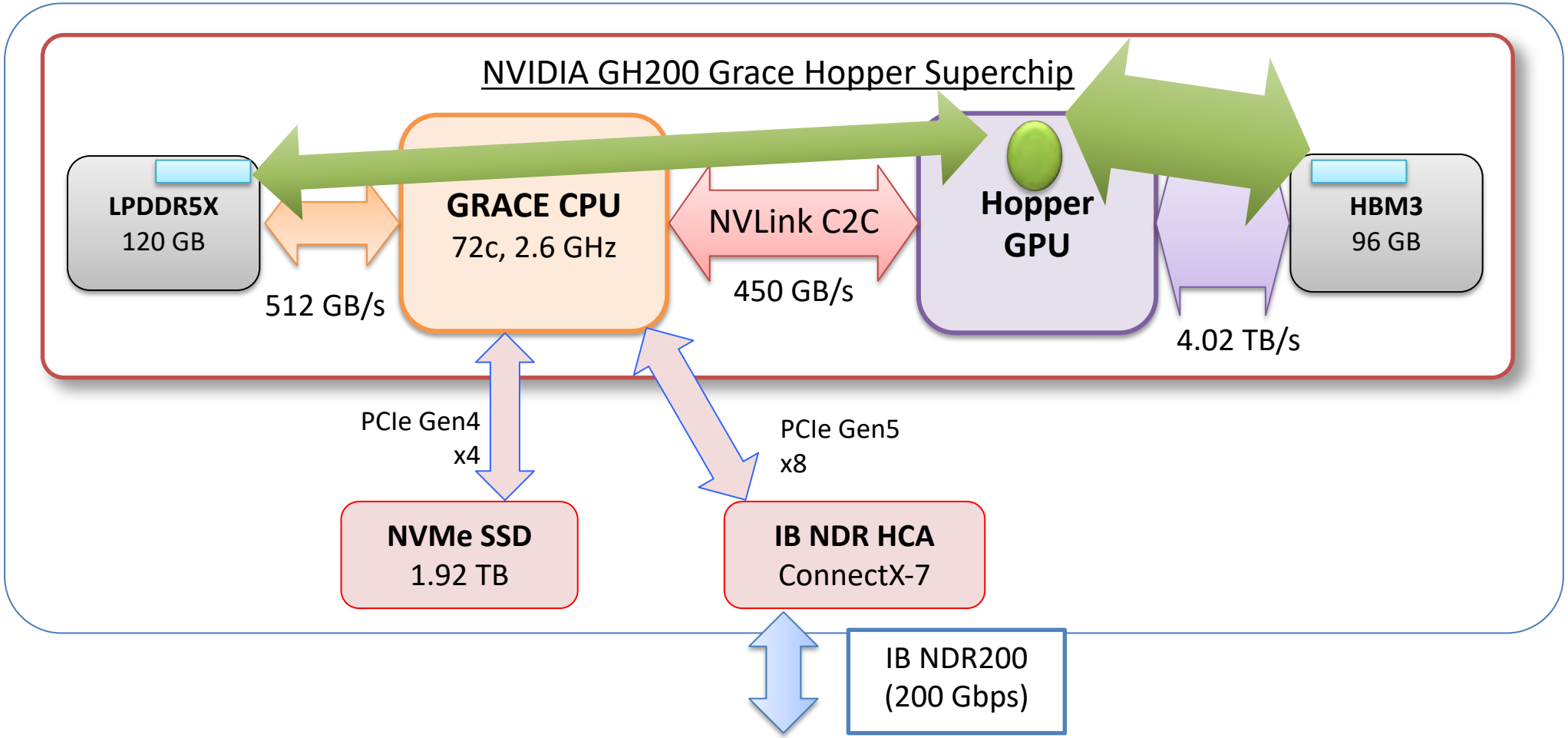
Graceからのメモリビュー

- NUMAとして見える
  - First Touchが大事
- 従来のCUDAのメモリモデルも使えて普通に動く＋転送が速い
  - cudaMalloc() + cudaMemcpy()



Hopperからのメモリビュー

- Graceでのアドレスを使って直接アクセス可能



GH200の予備評価

- 空冷版のGrace-Hopper (Supermicro製)で評価中
  - Miyabi-Gとの違い
    - Graceメモリ容量 120 GB ➔ 480 GB
    - Graceメモリバンド幅 512 GB/sec ➔ 384 GB/sec
- OS: Rocky Linux 9.4 aarch64
  - kernel: 5.14.0-427.{20,22}.1.el9_4.aarch64+64k
  - 64Kページ、irqbalance無効, transparent huge page 無効、とかもろもろ、、
  - 参考
    - https://docs.nvidia.com/grace-patch-config-guide.pdf
    - https://docs.nvidia.com/grace-performance-tuning-guide.pdf
- CUDA 12.4, CUDA Driver: 550.90.7
- OFED 24.04-0.6.6
- NVHPC 24.5-2
- コンテナ
  - Apptainer 1.3.2, Enroot (+caps) 3.5.0

- 評価結果は後日公開予定

導入に向けたスケジュール

- GPUアーキテクチャ決定: 2022年6月末
- 資料招請(RFI): 2022年11月上旬
- 意見招請(RFC): 2023年5月上旬
- 入札広告(RFP): 2023年8月下旬
- 開札: 2023年11月上旬
- 運用開始: 2025年1月15日

➔導入期間: 14ヶ月、
調達作業を始めて運用開始まで足掛け3年、、

OFP (2016年5月契約): 106円/ドル

2022年2月 ウクライナ危機

(135円/ドル)

(148円/ドル) ~4.5 MW, **150+**(~180) **PF**

(137円/ドル) ~3.0 MW, **100+**(~120) **PF**

2023年10月 ガザ・イスラエル問題

(146~151円/ドル) ~3.0 MW, **70+ PFLOPS**

2.0 MW, **80 PFLOPS**

高圧ケーブル不足➔解消

まとめ

- Miyabi (OFP-II)の導入中
  - 運用開始: 2025年1月、導入、運用：富士通
    - 設置場所：東京大学 柏キャンパス（OFPと同じ部屋、ずっと小さい）
  - Miyabiの合計性能 **80.1 PFLOPS**：OFP (25 PFLOPS)の**約3.2倍**の性能
    - Miyabi-G: 演算加速ノード, NVIDIA GH200 Superchip
      - CPU: Grace, 3.456 TFLOPS, 120 GB, 512 GB/s (LPDDR5X)+
        GPU: Hopper, 66.9 TFLOPS, 96 GB, 4.02 TB/s (HBM3), キャッシュコヒーレント, 1.9TB NVMe SSD
      - 合計: 1,120 node, 78.8 PFLOPS, 5.07 PB/s
    - Miyabi-C: 汎用CPUノード, Intel Xeon Max 9480
      - 計算ノード: 6.8 TFLOPS, 128GiB, 3.2 TB/s (HBM2e only)
      - 合計: 190 ノード, 1.3 PFLOPS, 608 TB/s
    - ストレージ: DDN Lustre 11.3 PB, All Flash
  - GH200を採用する国内初のオープンシステム、世界的に見ても特徴あるシステム