PCCC24 2024年12月5日(木)

最先端共同HPC基盤施設の新スパコンシステムMiyabiの運用開始に向けて / 東大情報基盤センターの目指す『計算・データ・学習』とデータ利活用を融合した革新的なスーパーコンピューティング

塙　敏博

最先端共同HPC基盤施設 (JCAHPC)

東京大学 情報基盤センター

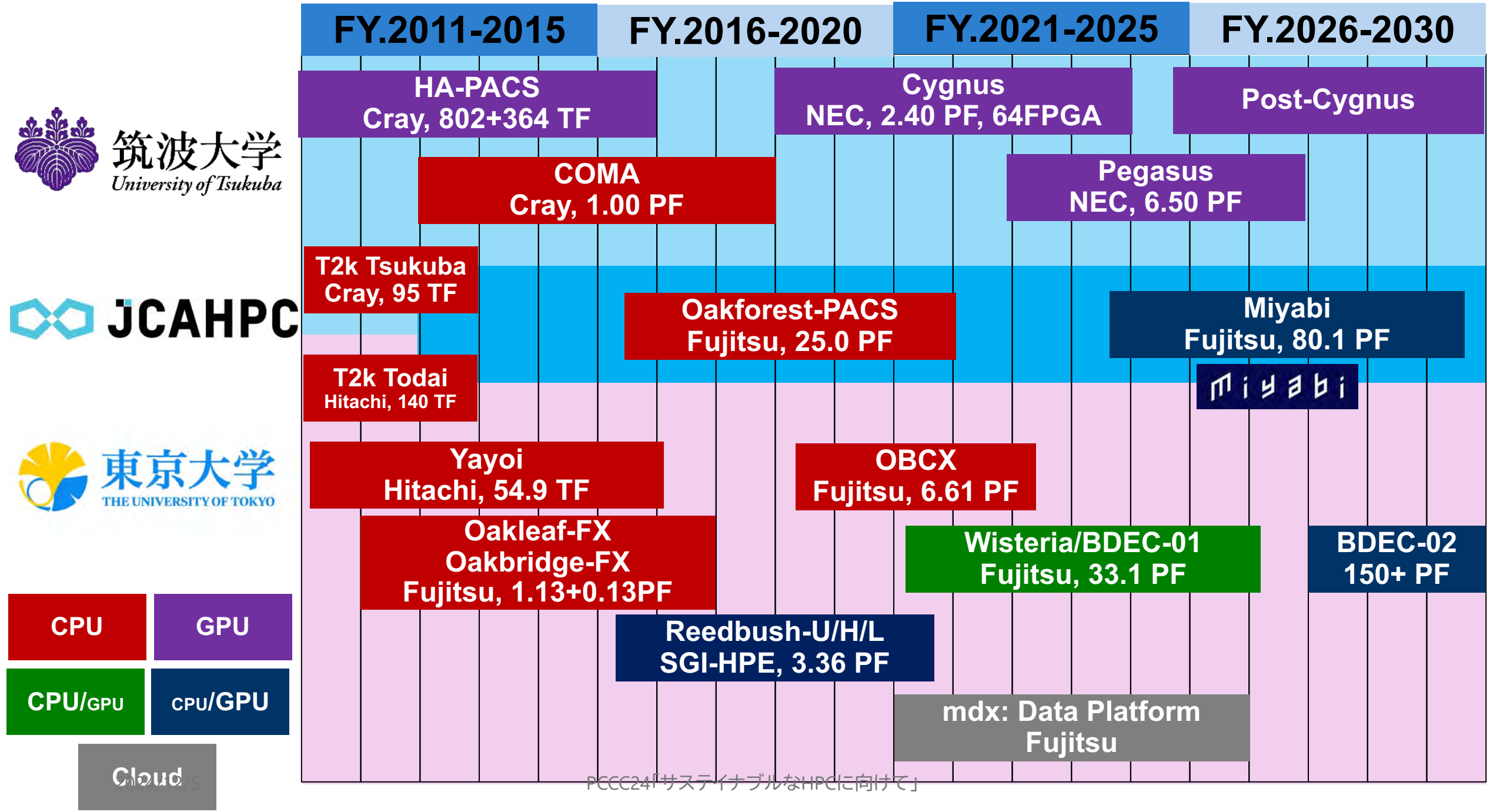
筑波大学 計算科学研究センター(客員)

最先端共同HPC基盤施設

http://jcahpc.jp/eng/index.html

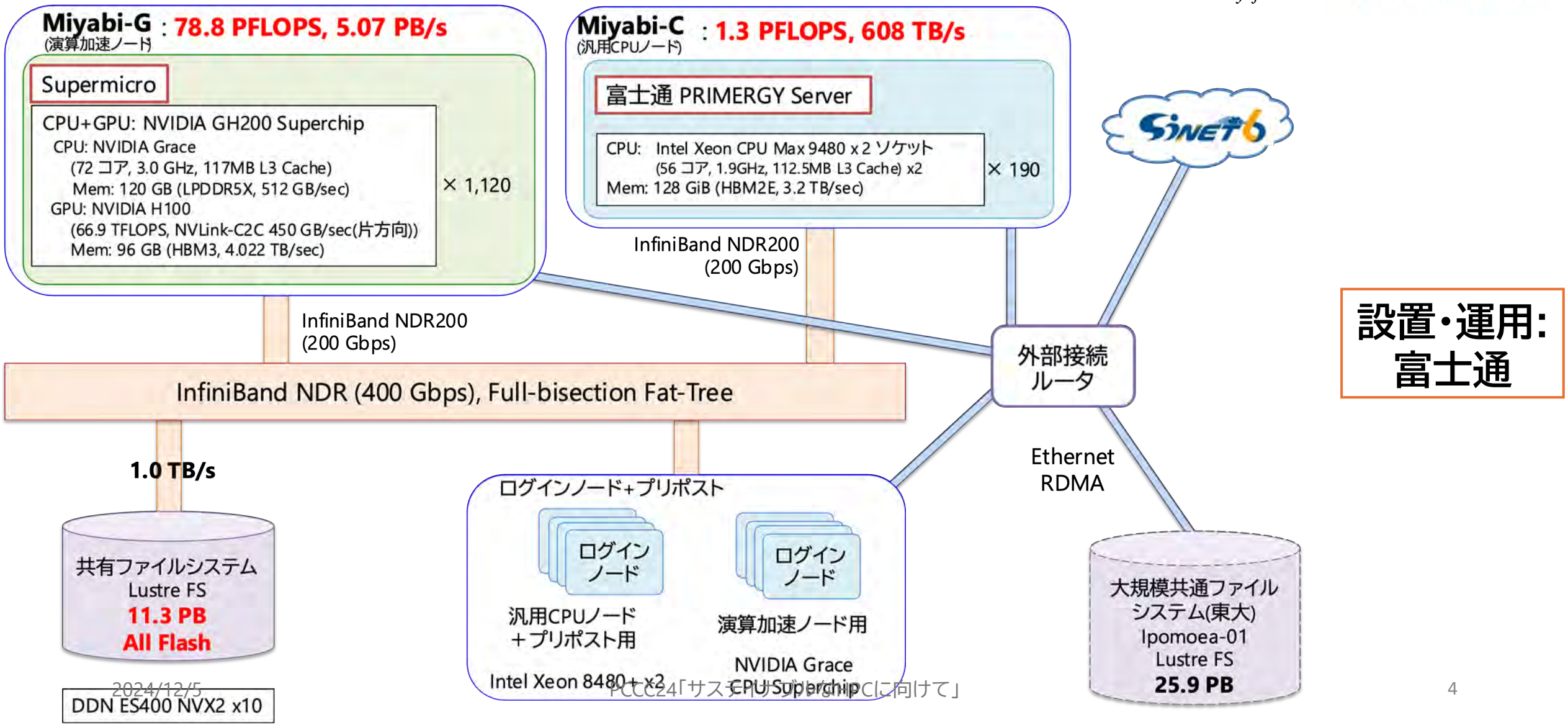
JCAHPC
筑波大学 University of Tsukuba
東京大学 THE UNIVERSITY OF TOKYO

- 2013年創立
  - JCAHPC(Joint Center for Advanced High Performance Computing)
    - 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
  - 最先端計算科学の推進
  - 大規模システムの設計・導入・運用を共同で実施する我が国でも初の試み:より大規模なシステムを効率的に導入可能
- Oakforest-PACS(OFP):JCAHPC第一号機
  - Intel Xeon Phi 8,208ノード, 25PF(富士通)
  - 2016年11月時点でTop500の6位(国内1位)
    - 2022年3月末で退役
  - 「京」の退役後,「富岳」登場までの2019・2020年度は事実上の「National Flagship System」としての役割を担う
- OFP-II(OFP後継機)へ向けた試み


FY.2011-2015
FY.2016-2020
FY.2021-2025
FY.2026-2030
筑波大学
University of Tsukuba
JCAHPC
東京大学
THE UNIVERSITY OF TOKYO
HA-PACS
Cray, 802+364 TF
Cygnus
NEC, 2.40 PF, 64FPGA
Post-Cygnus
COMA
Cray, 1.00 PF
Pegasus
NEC, 6.50 PF
T2k Tsukuba
Cray, 95 TF
Oakforest-PACS
Fujitsu, 25.0 PF
Miyabi
Fujitsu, 80.1 PF
Miyabi
T2k Todai
Hitachi, 140 TF
Yayoi
Hitachi, 54.9 TF
OBCX
Fujitsu, 6.61 PF
Oakleaf-FX
Oakbridge-FX
Fujitsu, 1.13+0.13PF
Wisteria/BDEC-01
Fujitsu, 33.1 PF
BDEC-02
150+ PF
Reedbush-U/H/L
SGI-HPE, 3.36 PF
mdx: Data Platform
Fujitsu
CPU
GPU
CPU/GPU
CPU/GPU
Cloud

Miyabiの概要 (1/3)
2025年1月運用開始 80.1 PFLOPS

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

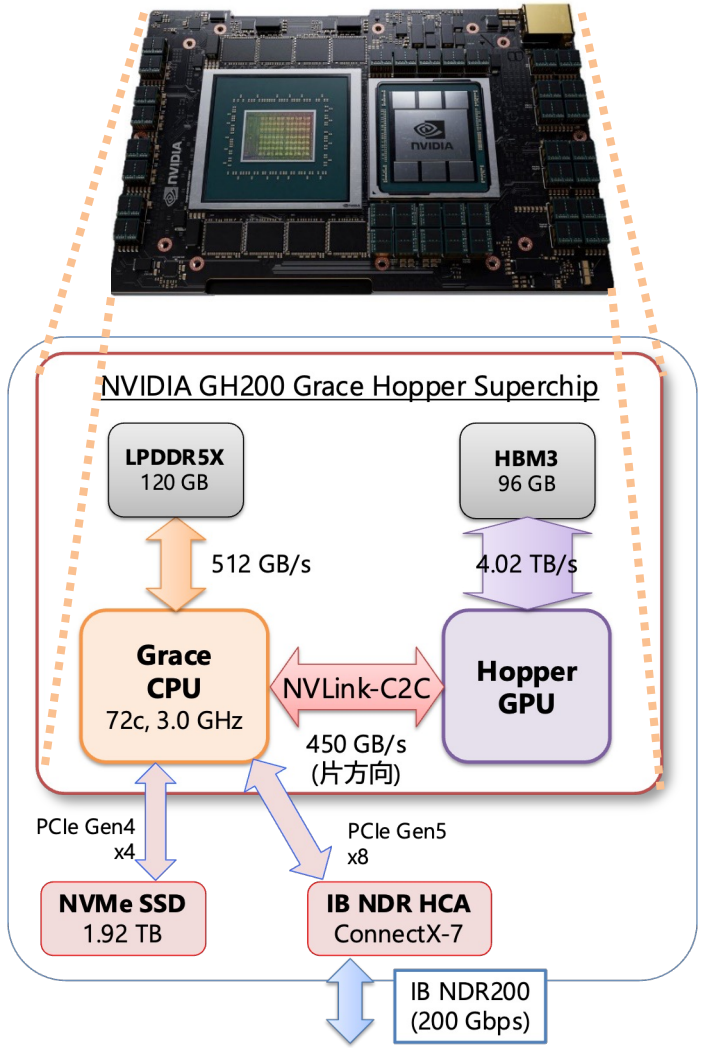


設置・運用: 富士通

Miyabiの概要 (2/3)

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

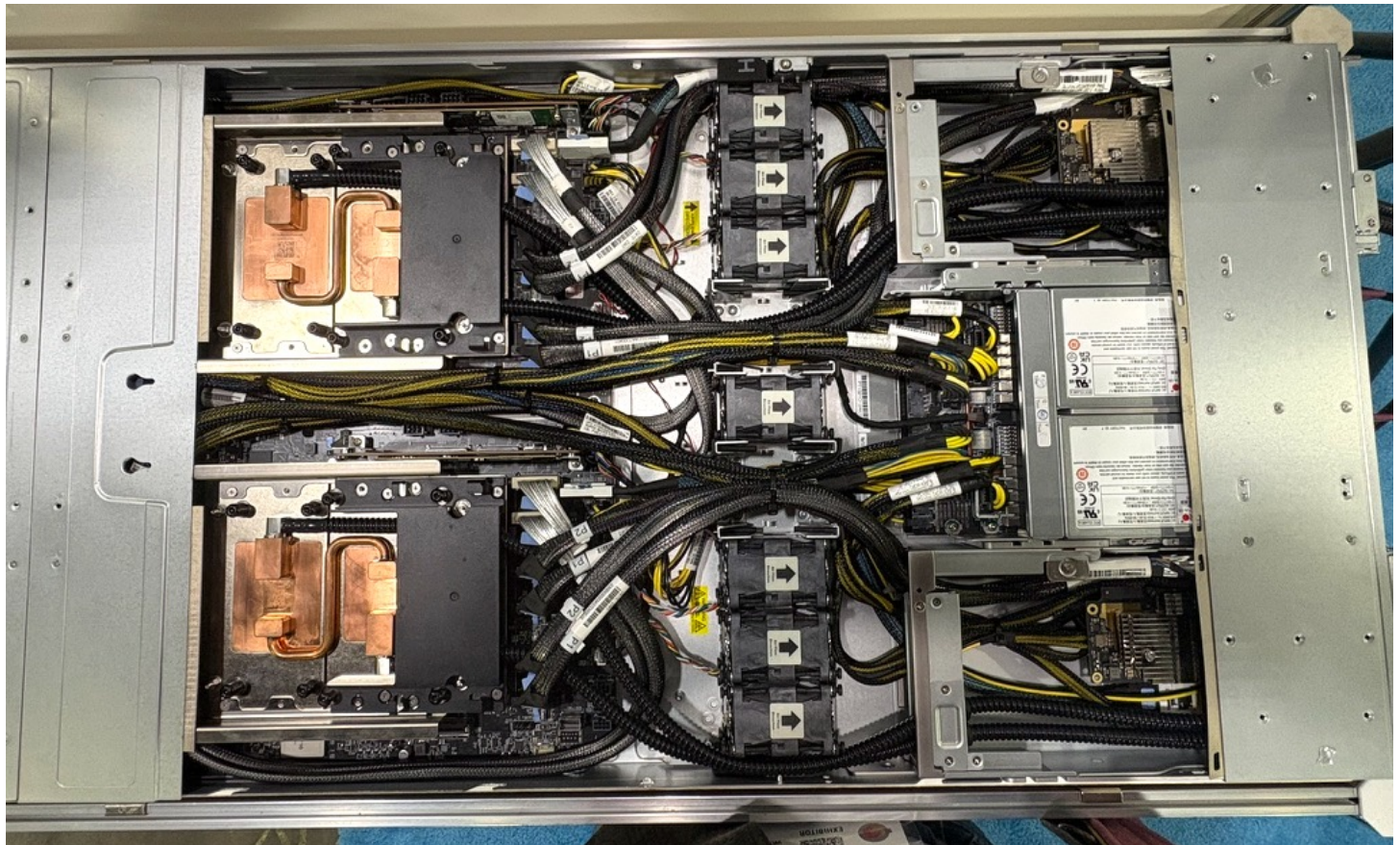
- **Miyabi-G: 演算加速ノード: NVIDIA GH200**
  - 計算ノード: NVIDIA GH200 Grace-Hopper Superchip
    - Grace: 72c, 3.45 TF, 120 GB, 512 GB/sec (LPDDR5X)
    - H100: 66.9 TF DP-Tensor Core, 96 GB, 4,022 GB/sec (HBM3)
      - CPU-GPU間はキャッシュコヒーレント
    - NVMe SSD for each GPU: 1.9TB, 8.0GB/sec, GPUDirect Storage
  - **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)**

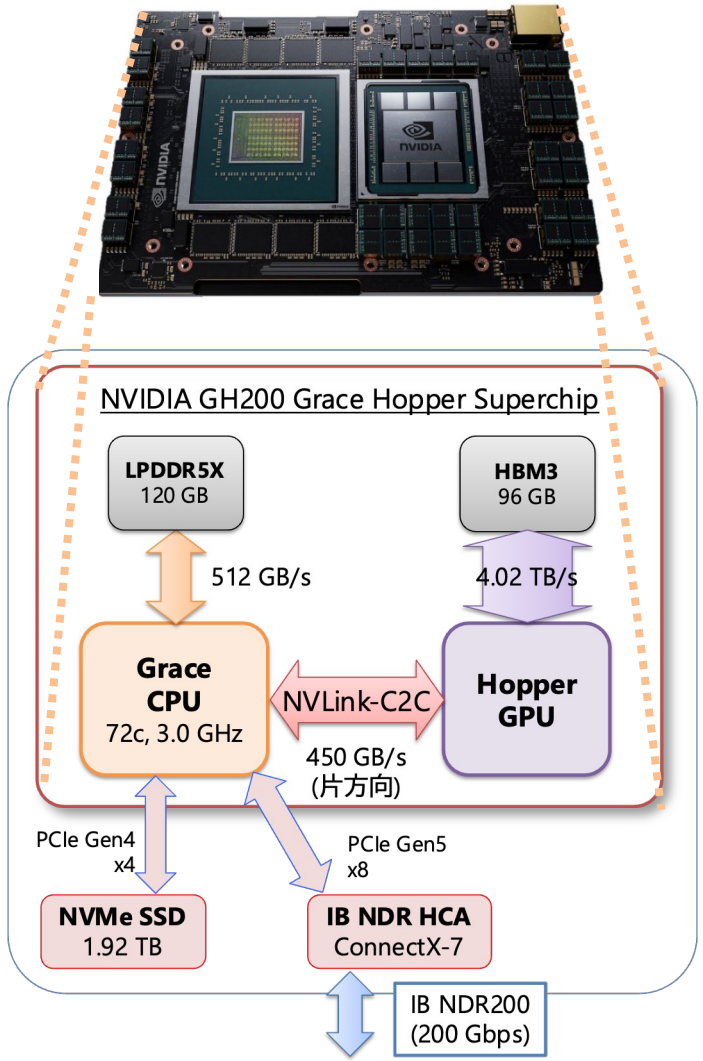


Miyabiの概要 (2/3)

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

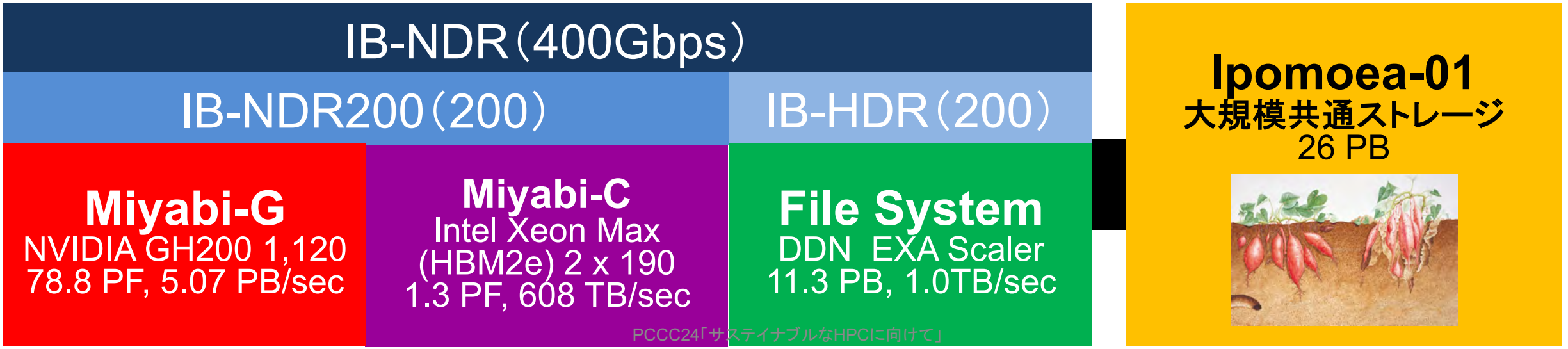
- Supermicro ARS-111GL-DNHR-LCC
  - 1U 2ノード、直接水冷





Miyabiの概要 (3/3)

- ファイルシステム: **DDN EXAScalar, Lustre FS**
  - 11.3 PB (NVMe SSD) 1.0TB/sec, “Ipomoea-01” (26 PB) も利用可能
- **Miyabi-G/C 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、Miyabi-G/C間の通信は h3-Open-SYS/WaitIO により実現**



Miyabi(OFP-II) 仕様まとめ

JCAHPC

	Miyabi-G (演算加速ノード)	Miyabi-C (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		Miyabi-G (演算加速ノード)	Miyabi-C (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 (片方向) 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**

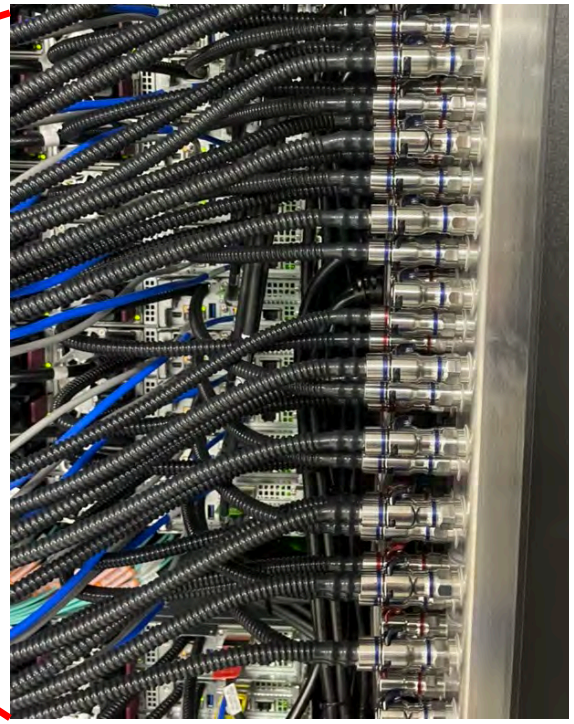
Category	Name (Organizations)	Target, Method etc.	Language
Engineering (5)	FrontISTR (U.Tokyo)	Solid Mechanics, FEM	Fortran
	FrontFlow/blue (FFB) (U.Tokyo)	CFD, FEM	Fortran
	FrontFlow/red (AFFr) (Advanced Soft)	CFD, FVM	Fortran
	FFX (U.Tokyo)	CFD, Lattice Boltzmann Method (LBM)	Fortran
	CUBE (Kobe U./RIKEN)	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

Miyabi-Gのラック内冷却機器

リアドアパネル内

マニフォールド・
水冷パイプ



In Rack CDU(ラック毎)

64th TOP500 List (Nov, 2024)

R_{max}: Performance of Linpack (TFLOPS)
R_{peak}: Peak Performance (TFLOPS), Power: kW

http://www.top500.org/

	Site	Computer/Year Vendor	Cores	R_{max} (PFLOPS)	R_{peak} (PFLOPS)	GFLOPS/W	Power (kW)
1	**El Capitan, 2024, USA** DOE/NNSA/LLNL	HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS	11,039,616	1,742.00 (=1.742 EF)	2,746.38 63.4 %	58.99	29,581
2	**Frontier, 2021, USA** DOE/SC/Oak Ridge National Laboratory	HPE Cray EX235a, AMD Optimized 3rd Gen. EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11	9,066,176	1.353.00	2,055.72 65.8 %	54.98	24,607
3	**Aurora, 2023, USA** DOE/SC/Argonne National Laboratory	HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel	9,264,128	1,012.00	1,980.01 51.1 %	26.15	38,698
4	**Eagle, 2023, USA** Microsoft	Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR	2,073,600	561.20	846.84 66.3 %		
5	**HPC 6, 2024, Italy** Eni S.p.A.	HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9	3,143,520	477.90	606.97 66.3 %	56.48	8,461
6	**Fugaku, 2020, Japan** R-CCS, RIKEN	Fujitsu PRIMEHPC FX1000, Fujitsu A64FX 48C 2.2GHz, Tofu-D	7,630,848	442.01	537.21 82.3 %	14.78	29,899
7	**Alps, 2024, Switzerland** Swiss Natl. SC Centre (CSCS)	HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11	2,121,600	434.90	574.84 75.7 %	61.05	7,124
8	**LUMI, 2023, Finland** EuroHPC/CSC	HPE Cray EX235a, AMD Optimized 3rd Gen. EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11	2,752,704	379.70	531.51 71.4 %	53.43	7,107
9	**Leonard, 2023, Italy** EuroHPC/Cineca	BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64GB, Quad-rail NVIDIA HDR100	1,824,768	241.20	306.31 78.7 %	32.19	7,494
10	**Tuolumne, 2024, USA** DOE/NNSA/LLNL	HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS	1,161,216	208.10	288.88 72.0 %	61.45	3,387
13	**Venado, 2024, USA** DOE/NNSA/LANL	HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11	481,440	98.51	130.44 75.5 %	59.29	1,662
16	**CHIE-3, 2024, Japan** SoftBank, Corp.	NVIDIA DGX H100, Xeon Platinum 8480C 56C 2GHz, NVIDIA H100, Infiniband NDR400, Ubuntu 22.04.4 LTS	163.200	91.94	138.32 66.5 %		
17	**CHIE-2, 2024, Japan** SoftBank, Corp.	NVIDIA DGX H100, Xeon Platinum 8480C 56C 2GHz, NVIDIA H100, Infiniband NDR400, Ubuntu 22.04.4 LTS	163.200	89.78	138.32 64.9 %		
18	**JETI, 2024, Germany** EuroHPC/FZJ	BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Linux, Modular OS	391,680	83.14	94.00 88.4 %	63.43	1,311
22	**CEA-HE, 2024, France** CEA	BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail BXI v2, EVIDEN	389,232	64.32	103.48 62.2 %	52.17	1,233
28	**Miyabi-G, 2024, Japan** JCAHPC	Fujitsu, Supermicro ARS 111GL DNHR LCC, Grace Hopper Superchip 72C 3GHz, Infiniband NDR200, Rocky Linux	80,640	46.80	72.80 64.3 %	47.59	983
36	**TSUBAME 4.0, 2024, Japan** Institute of Science Tokyo	HPE Cray XD665, AMD EPYC 9654 96C 2.4GHz, NVIDIA H100 SXM5 94 GB, Infiniband NDR200	172,800	39.62	61.60 64.3 %	48.55	816
58	**Wisteria/BDEC-01 (Odyssey), 2021, Japan** U.Tokyo	Fujitsu PRIMEHPC FX1000, A64FX 48C 2.2GHz, Tofu D	368,640	22.12	25.95 85.2 %	15.07	1,468

Green 500 Ranking (Nov. 2024)

http://www.top500.org/

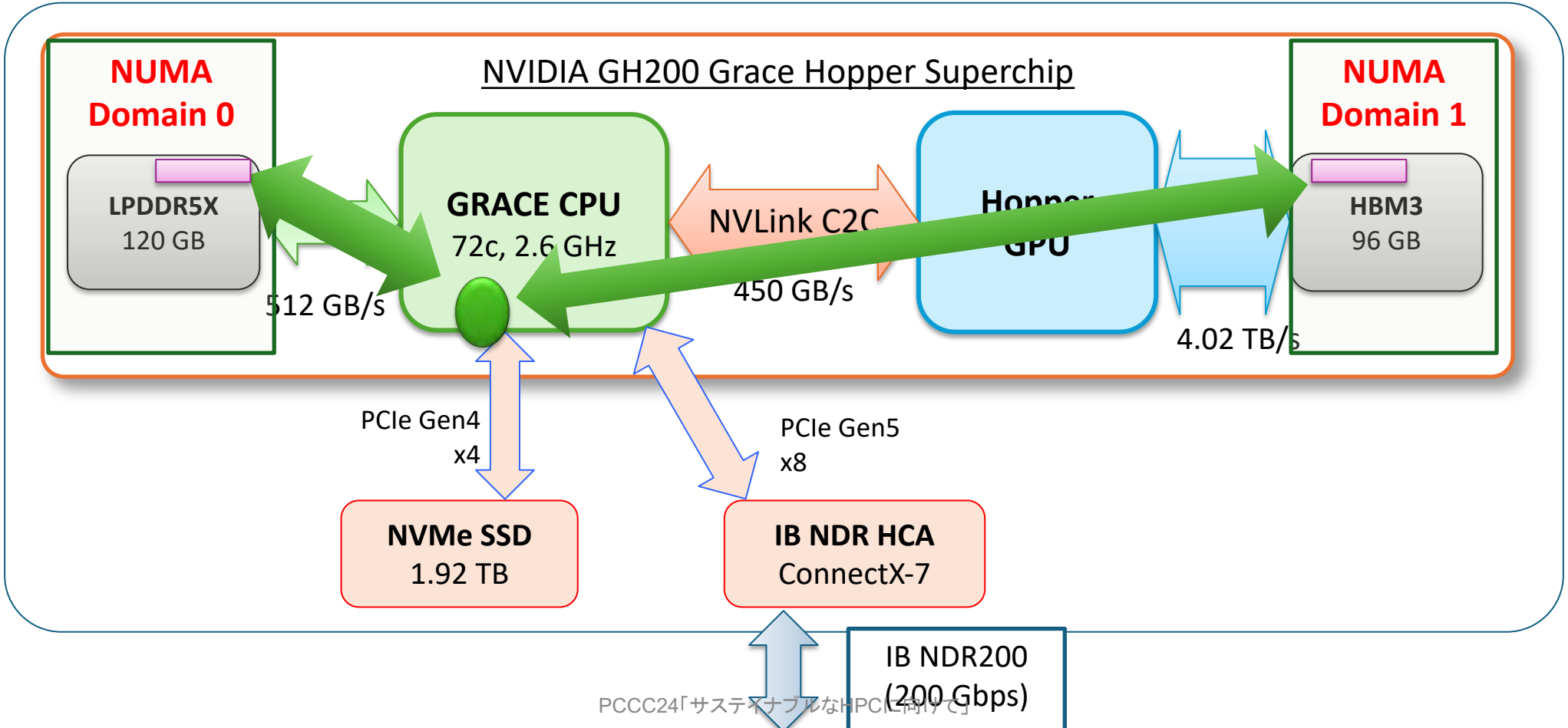
	TOP 500 Rank	System	Accelerator	Cores	HPL Rmax (Pflop/s)	Power (kW)	GFLOPS/W	Level
1	224	JEDI, EuroHPC/Julich, Germany	NVIDIA GH200	19,584	4.50	67	72.733	1
2	122	ROMEO-2025, ROMEO HPC Center - Champagne-Ardenne, France	NVIDIA GH200	47,328	9.86	160	70.912	1
3	442	Adastra2, GENCI-CINES, France	AMD Instrinct MI300A	16,128	2.53	37	69.098	1
4	155	Isambard-AI phase1, U. Bristol, UK	NVIDIA GH200	34,272	7.42	117	68.835	1
5	51	Capella, TU Dresden ZIH, Germany	NVIDIA H100 94GB	85,248	24.06	445	68.053	3
6	18	JETI, EuroHPC/Julich, Germany	NVIDIA GH200	391,680	83.14	1,311	67.963	1
7	69	Helios GPU, Cyfronet, Poland	NVIDIA GH200	89,760	19.14	317	66.948	2
8	371	Henri, Flatiron Institute, USA	NVIDIA H100 80GB	8,288	2.88	44	65.396	?
9	340	HoreKa-Teal, KIT, Germany	NVIDIA H100 94GB SXM5	13,616	3.12	50	62.964	1
10	49	rzAdams, DoE LLNL, US	AMD Instrinct MI300A	129,024	24.38	388	62.803	2
16	13	Venado, DoE LANL, US	NVIDIA GH200	481,440	98.51	1,662	59.287	1
30	36	TSUBAME 4.0, Science Tokyo	NVIDIA H100 94GB SXM5	172,800	39.62	816	48.565	3
33	28	**Miyabi-G, JCAHPC**	**NVIDIA GH200**	**221,952**	**46.80**	**983**	**47.588**	**3**
48	230	Pegasus, University of Tsukuba, Japan	NVIDIA H100 80GB	27,000	4.34	130	41.123	2
49	191	Wisteria/BDEC-01 (Aquarius), The University of Tokyo, Japan	NVIDIA A100 40GB	42,120	4.425	183	24.06	2

HPCG Ranking (Nov, 2024)

	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)
1	Fugaku	7,630,848	442.01	4	16.00
2	Frontier	8,699,904	1,206.00	1	14.05
3	Aurora	9,264,128	1,012.00	2	5.61
4	LUMI	2,752,704	379.70	5	4.59
5	Alps	2,121,600	434.90	7	3.67
6	Leonardo	1,824,768	241.20	9	3.11
7	Perlmutter	888,832	79.23	19	1.91
8	Sierra	1,572,480	94.64	14	1.80
9	Selene	555,520	63.46	23	1.62
10	JUWELS Booster	449,228	44.12	33	1.28
12	AOBA-S	64,512	17.22	76	1.09
17	Wisteria/Odyssey	348,640	22.12	58	0.818
19	ES4-SX-Aurora Tsubasa	43,776	9.99	119	0.748
21	Miyabi-G	221,952	46.80	28	0.645

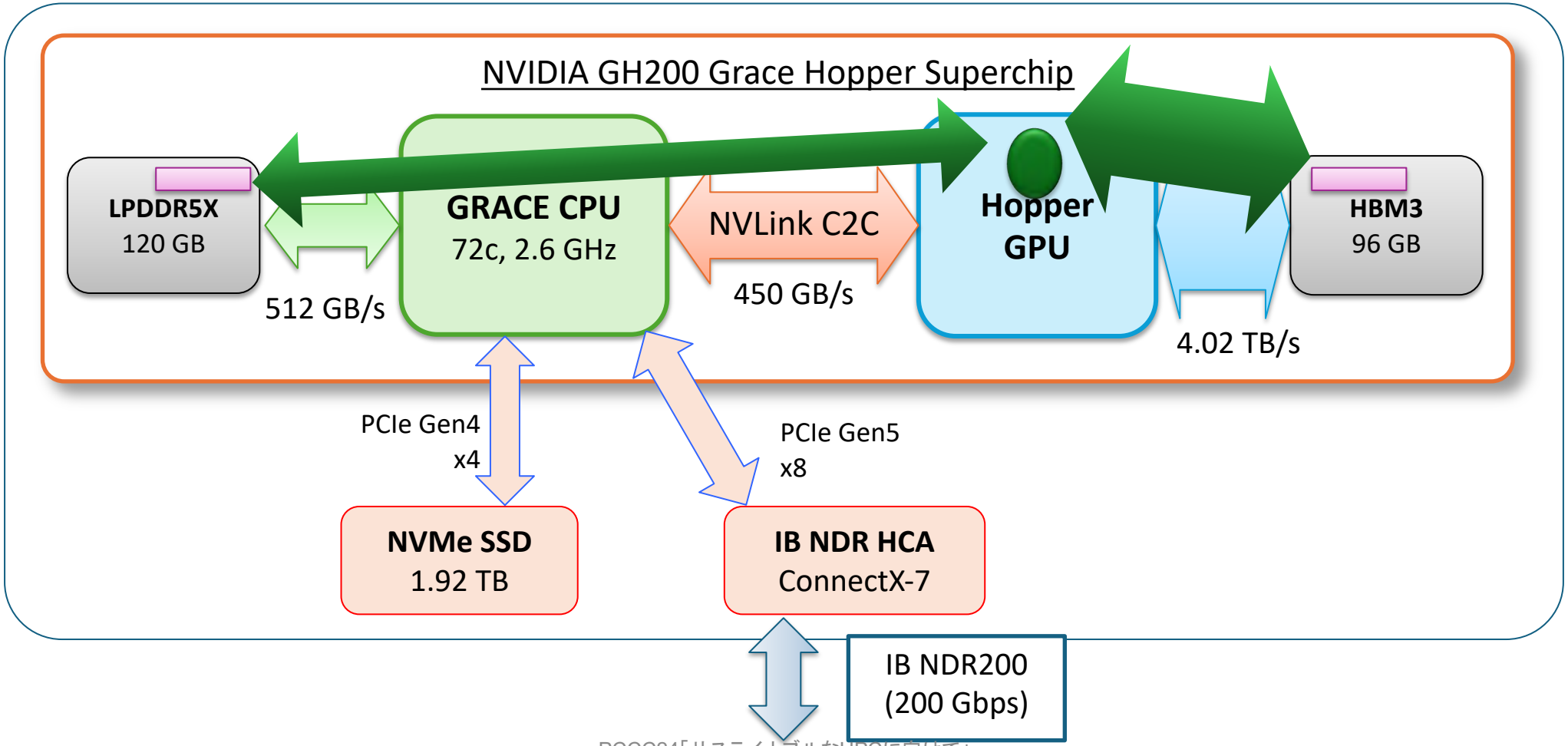
Graceからのメモリビュー

- 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
- 従来のCUDAのコードはH100とほぼ挙動は変わらない（メモリBW比相当）

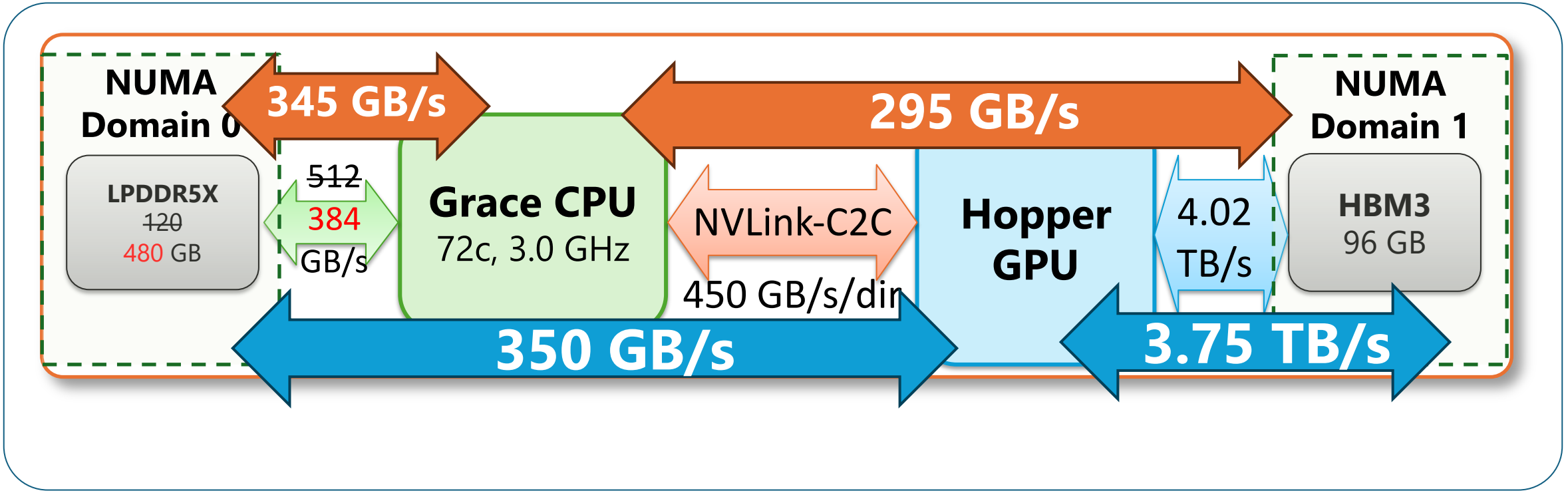


nvbandwidth

- https://github.com/NVIDIA/nvbandwidth
  - 8 GB (-b 8192)
- Host To Device BW (Copy kernel)
  - host_to_device_memcpy_sm
- Device To Host BW (Copy kernel)
  - device_to_host_memcpy_sm
- Device To Host Latency (Copy kernel)
  - host_device_latency_sm

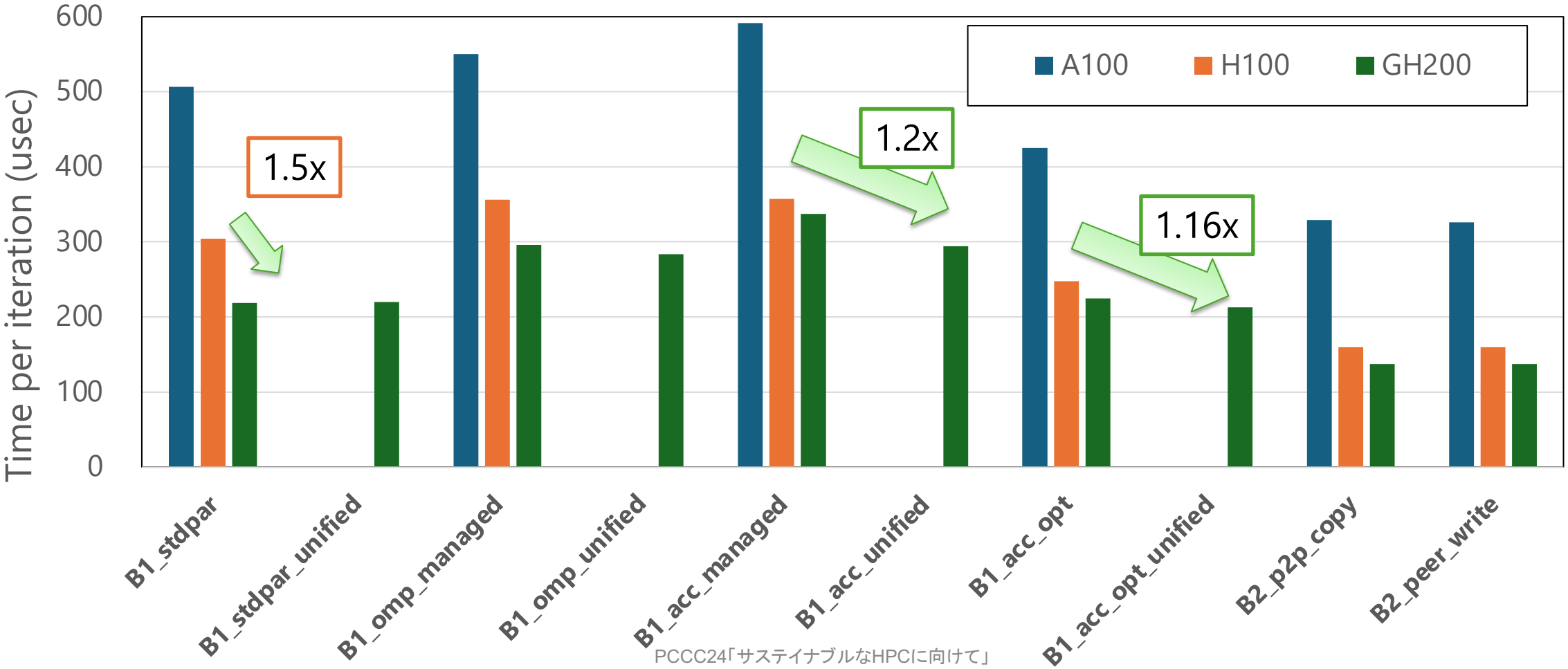
	Memcpy (H→D) (GB/sec)	Memcpy (D→H) (GB/sec)	Latency (D⇔H, single dir) (nsec)
GH200	368.40	351.34	769.08
SPR+H100 SXM5	36.82	39.93	1089.60

メモリバンド幅まとめ


NUMA
Domain 0
345 GB/s
295 GB/s
NUMA
Domain 1
LPDDR5X
~~120~~
480 GB
~~512~~
384
GB/s
Grace CPU
72c, 3.0 GHz
NVLink-C2C
450 GB/s/dir
Hopper
GPU
4.02
TB/s
HBM3
96 GB
350 GB/s
3.75 TB/s

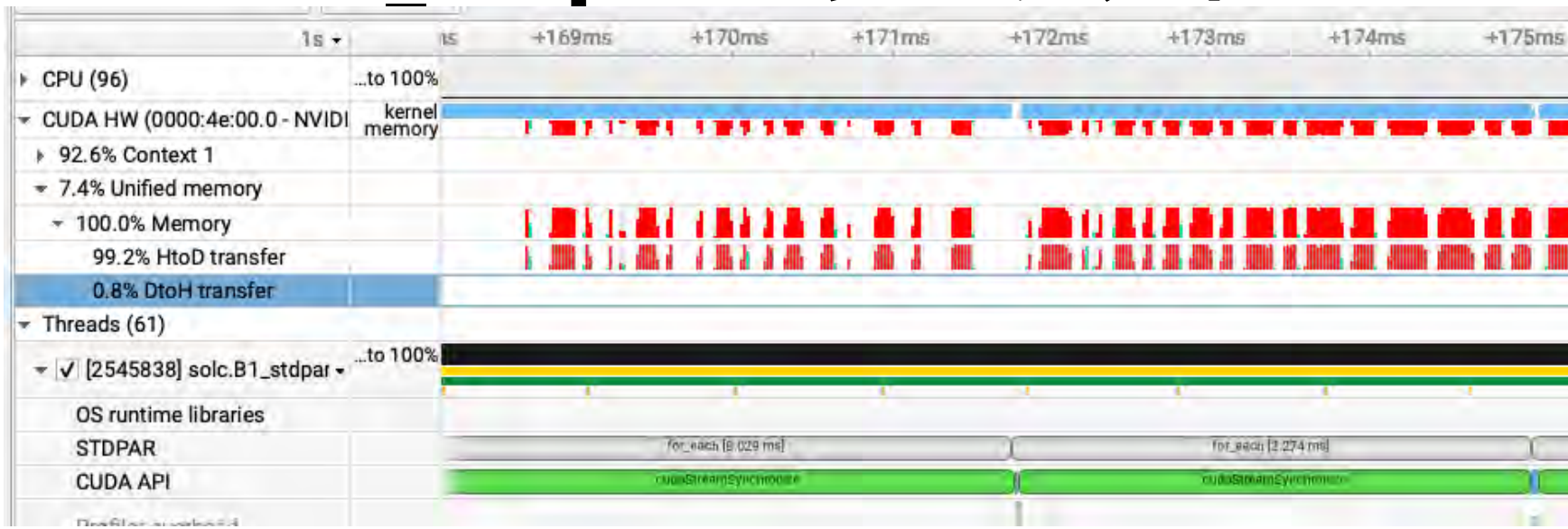
Poisson 3D: small (128x128x128)

- C言語、オリジナルは OpenMP並列化
  - 小サイズでは、managed ➔unifiedの効果が高い
  - stdparがGH200で性能が高い



Poisson 3D small B1_stdparのプロファイル

- 上: H100 SXM5
  下: GH200
  - 赤: データ転送
- データ転送の総量はさほど変わらない
- NVLink-C2C vs PCIeは約7~10倍の性能
- 64Kページ@GH200の効果
  - ページフォルトは1/4
    (ページフォルトをトリガにしてデータ転送が起こる)



NVIDIA HPC SDK コンパイラ

- 既存のコードをNVIDIA GPUに向けて移植する上では、ほぼ一択
  - 2ヶ月ごとにリリース、最新は 24.11

- GH200向けにはこのような感じでコンパイラを呼べば良い
- OpenACC
  - data指示文を省略可能

```
nvfortran -Ofast -acc=gpu -gpu=cc90,mem:unified:nomanagedalloc file.f90
```

- Stdpar (Fortran: Do-concurrent)
  - CPUとGPU協調動作での加速も将来的に期待、ただしコンパイラ任せ

```
nvfortran -Ofast -stdpar=gpu -gpu=cc90,mem:unified:nomanagedalloc file.f90
```

MIG (Multi-Instance GPU)

- GPUリソース(SM(演算コア), メモリ)を複数のインスタンスに分割する機能
  - https://www.nvidia.com/ja-jp/technologies/multi-instance-gpu/
- 想定される活用シーン
  - 1基のNVIDIA GH200の演算リソースを使い切れない場合
    ➔ 分割されたGPUを使えば, トークン消費量を抑えられる
  - 実装・開発中のコードのデバッグ・動作テスト・機能テスト
    ➔ (見かけの)GPU数が増えるので, ジョブ投入後の待ち時間が短縮される
  - 教育利用(主にGPUプログラミングの初心者・初級者向け)
    ➔ (見かけの)GPU数が増えるので, 多数の受講者のジョブが同時に実行される
- Miyabi-GでのMIG利用形態(GH200を4分割)
  - トークン消費量は通常のノード占有キューの 1/4
  - MIG#3 だけはGPUリソース(SM数)が少ないが気にしない

	MIG#0	MIG#1	MIG#2	MIG#3
GPUリソース	32 SMs, 24 GB	32 SMs, 24 GB	32 SMs, 24 GB	26 SMs, 24 GB
CPUリソース	18コア, 25 GiB	18コア, 25 GiB	18コア, 25 GiB	18コア, 25 GiB

東大情報基盤センターの目指す
『計算・データ・学習』とデータ利活用を融合した
革新的なスーパーコンピューティング

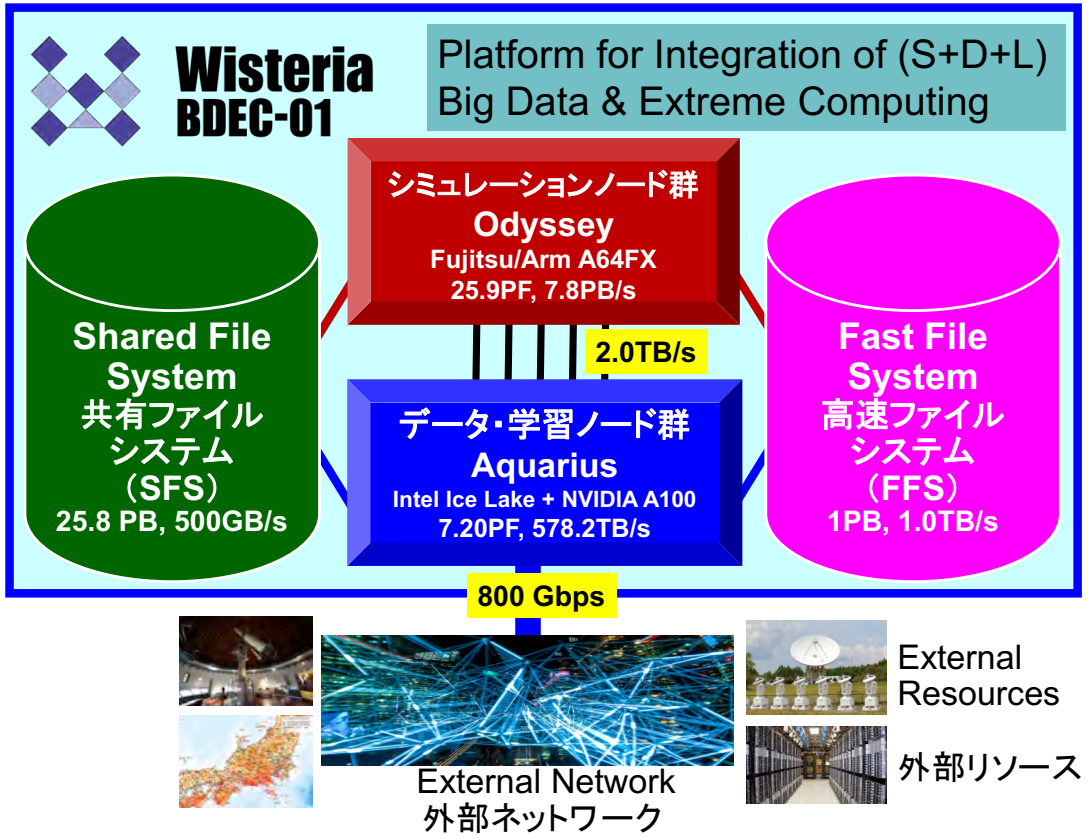
(シミュレーション(計算)+データ+学習)融合(S+D+L)

- 東大情報基盤センターでは, 2015年頃から「(S+D+L)融合」の重要性に注目し, それを実現するためのハードウェア, ソフトウェア, アプリケーション, アルゴリズムに関する研究開発を開始
  - BDEC計画(Big Data & Extreme Computing)
  - 「データ+学習」による, より高度な「シミュレーション」
    - AI for HPC, AI for Science
  - 地球科学関連では自然な発想(すでに実施されている)
- 2021年5月に運用を開始した「Wisteria/BDEC-01」は「BDEC計画」の1号機
  - Reedbush, Oakbridge-CXは「BDEC」のプロトタイプと位置づけられる
  - 「計算・データ・学習(S+D+L)」融合を実現する, 世界でも初めてのプラットフォーム

Wisteria/BDEC-01

BDEC:「計算・データ・学習（S+D+L）」融合のためのプラットフォーム（Big Data & Extreme Computing）

- 2021年5月14日運用開始
  - 東京大学柏Ⅱキャンパス
- 33.1 PF, 8.38 PB/sec.，富士通製
  - ~4.5 MVA（空調込み）, ~360m²
- Hierarchical, Hybrid, Heterogeneous (h3)
- **2種類のノード群**
  - **シミュレーションノード群（S, SIM）：Odyssey**
    - **従来のスパコン**
    - **Fujitsu PRIMEHPC FX1000 (A64FX), 25.9 PF**
      - **7,680ノード（368,640 コア）, 20ラック, Tofu-D**
  - **データ・学習ノード群（D/L, DL）：Aquarius**
    - **データ解析，機械学習**
    - **Intel Xeon Ice Lake + NVIDIA A100, 7.2 PF**
      - **45ノード（Ice Lake：90基，A100：360基）, IB-HDR**
    - **外部リソース（ストレージ，サーバー，センサーネットワーク他）に直接接続**
  - ファイルシステム：共有（大容量）＋高速



Wisteria
BDEC-01

Simulation Nodes
Odyssey
25.9 PF, 7.8 PB/s

Fast File System（FFS）
1.0 PB, 1.0 TB/s

Shared File System（SFS）
25.8 PB, 0.50 TB/s

Data/Learning Nodes
Aquarius
7.20 PF, 578.2 TB/s

計算科学コード

シミュレーション
ノード群，Odyssey

計算結果

最適化されたモデル，パラメータ

Wisteria/BDEC-01

機械学習，DDA

観測データ

データ･学習ノード群
Aquarius

データ同化
データ解析

Wisteria
BDEC-01

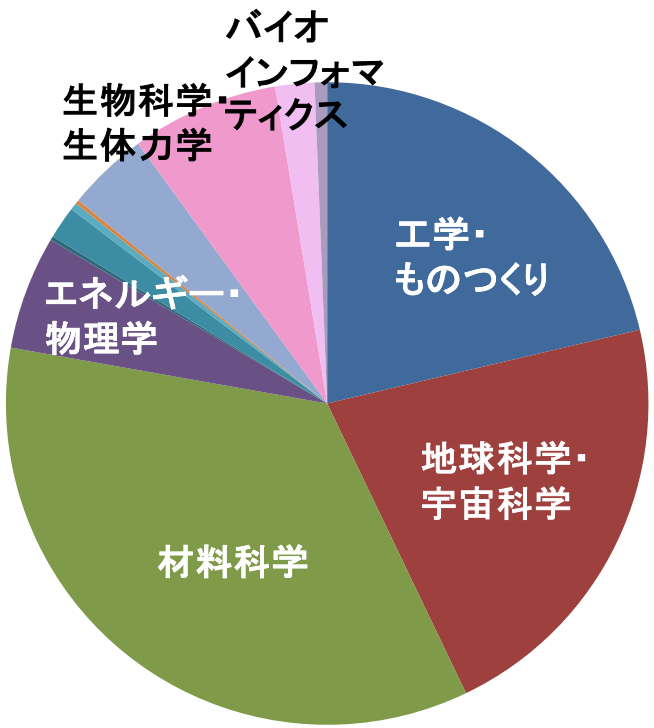
サーバー
ストレージ
DB
センサー群
他

外部ネットワーク

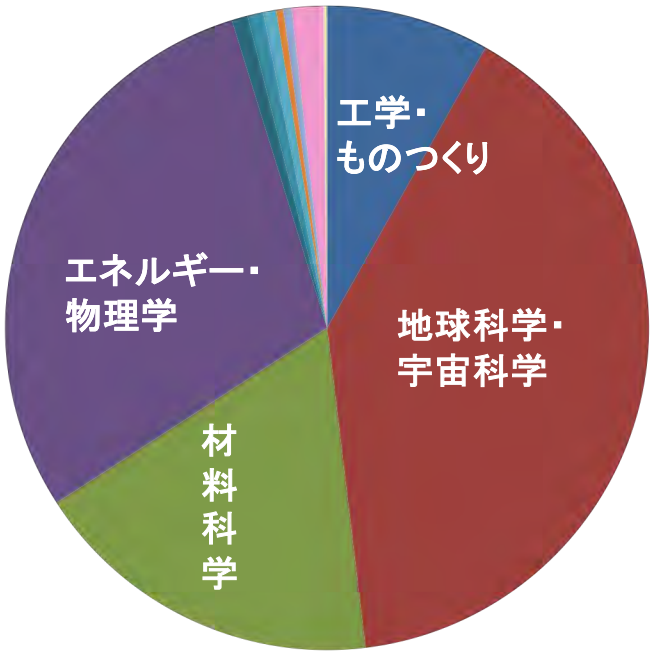
外部
リソース

2023年度分野別計算資源利用割合

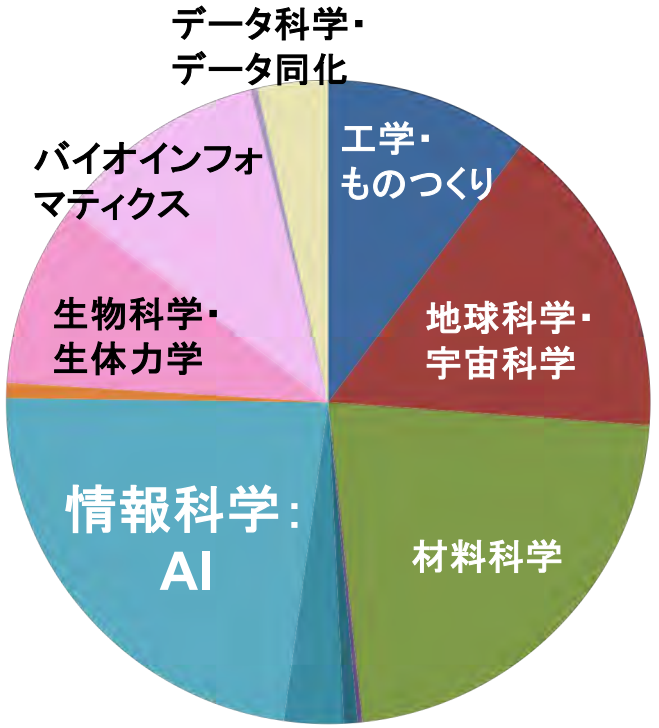
■汎用CPU, ■GPU



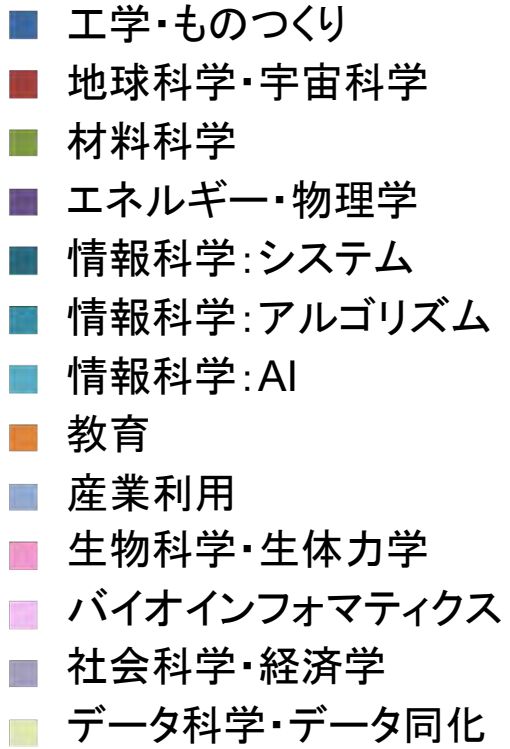
OBCX
CascadeLake
2023年9月末退役



Odyssey
A64FX

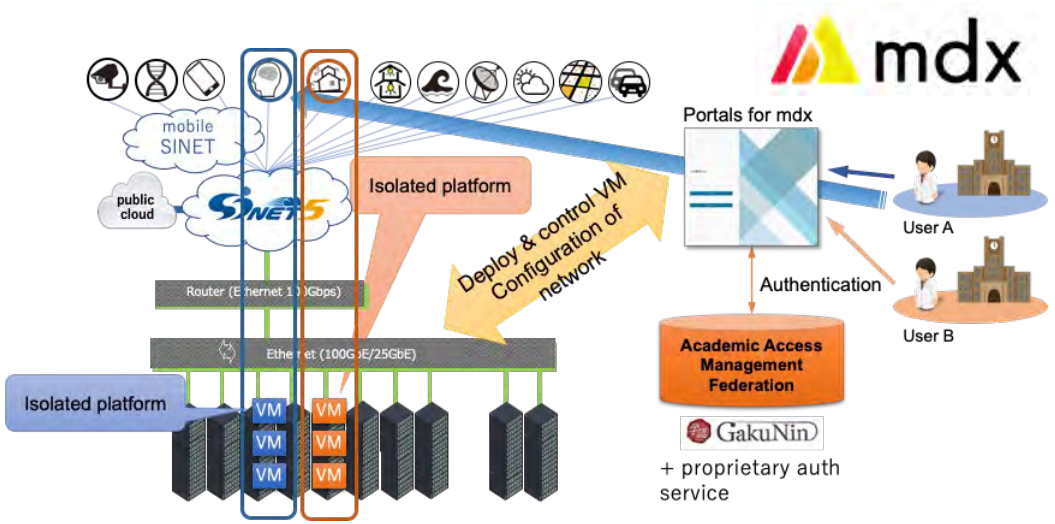


Aquarius
A100



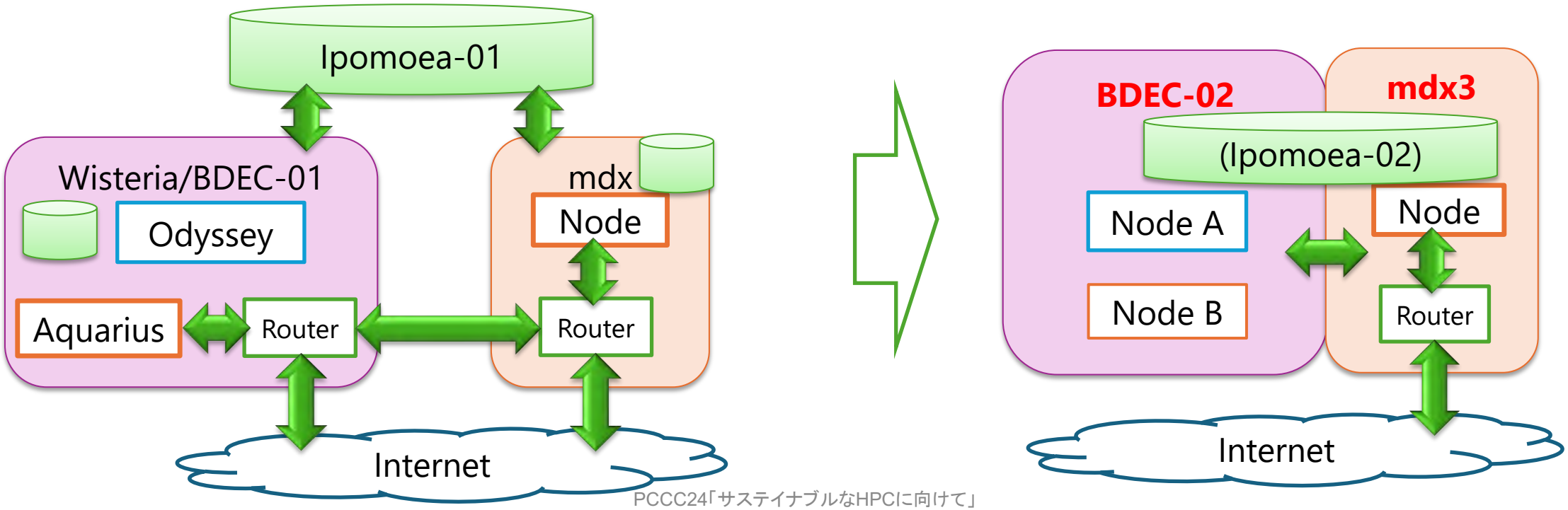
BDEC-02: BDEC計画 第2世代構想

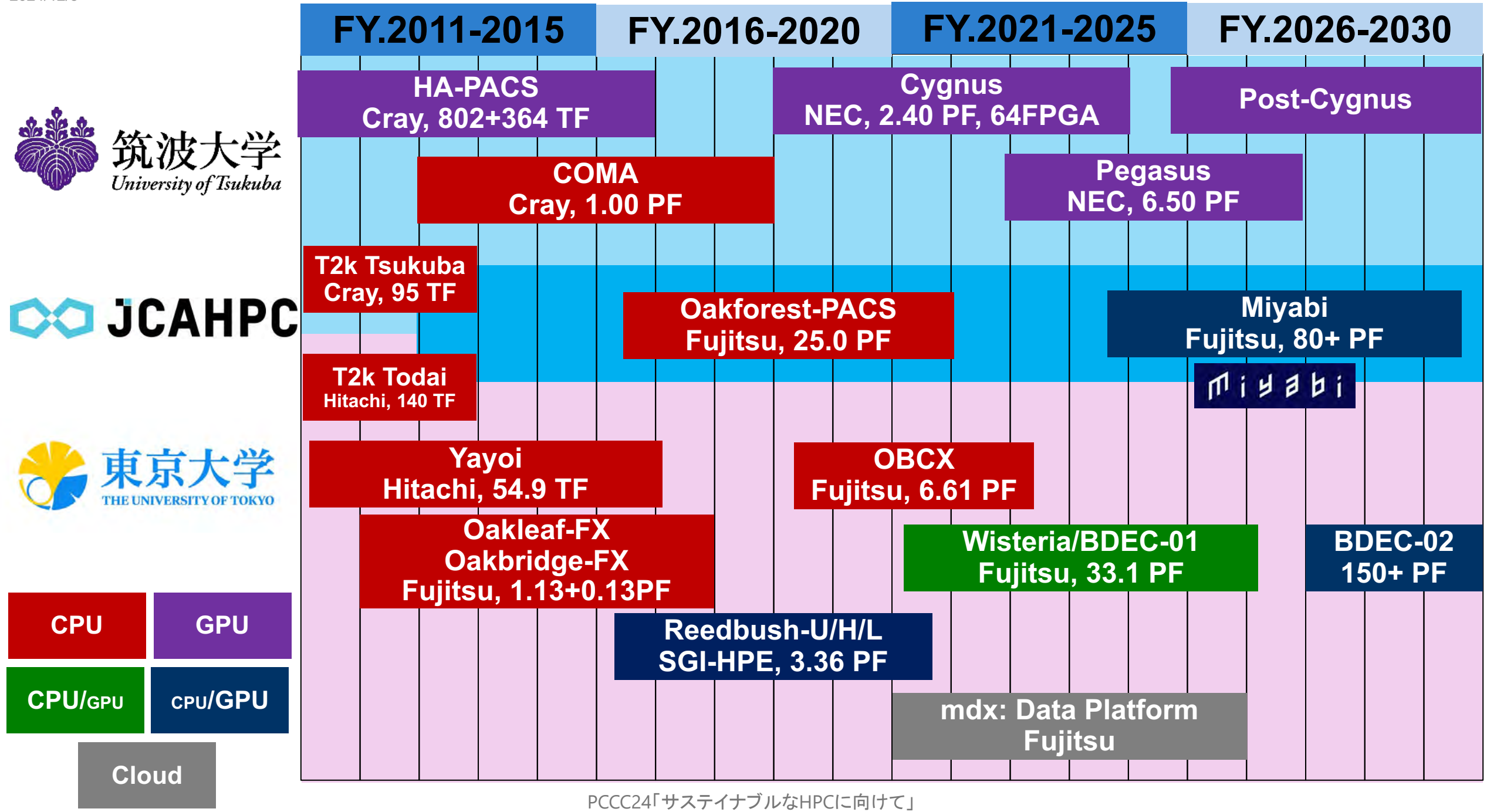
- 運用開始：2027年秋～2028年3月を想定
  - Wisteria/BDEC-01は2027年4月末で運用終了
- 当初は 200-250 PFLOPSを期待していたが、150 PFLOPS辺りが現実的か
  - とはいえHPL性能ありき、ではない
- AI for Scienceに向けたリアルタイム「計算・データ・学習」融合
- Miyabiに向けたGPU移行で培った知見 ➔ NVIDIA GPU、またはOpenACC / StdParと互換性のあるプログラミング環境
  - ユーザは2年余りのNVIDIA GPUへの移行の労力を払ってきた
  - Fortranは未だに重要
- 専用のデバイス/システムとの接続も想定
  - 量子コンピュータ等
  - h3-Open-BDEC, h3-Open-BDEC/QH
- mdx1のリプレースも必要な時期
  - クラウド指向のデータ利活用基盤
  - mdxの機能も包含できるよう設計
  - ストレージはシステム全体で共有



BDEC-02+mdx3 (仮)のコンセプト

- 資源をできるだけ共有し、冗長構成を避けることでコストを削減
  - 今でもWisteria/BDEC-01とmdxは同一の部屋に設置されている、、
- 適材適所の使い分けを1システムに統合
  - mdx3: インタラクティブ、リアルタイム処理、コンテナベース
  - BDEC-02: バッチ処理




FY.2011-2015
FY.2016-2020
FY.2021-2025
FY.2026-2030
筑波大学
University of Tsukuba
JCAHPC
東京大学
THE UNIVERSITY OF TOKYO
HA-PACS
Cray, 802+364 TF
Cygnus
NEC, 2.40 PF, 64FPGA
Post-Cygnus
COMA
Cray, 1.00 PF
Pegasus
NEC, 6.50 PF
T2k Tsukuba
Cray, 95 TF
Oakforest-PACS
Fujitsu, 25.0 PF
Miyabi
Fujitsu, 80+ PF
Miyabi
T2k Todai
Hitachi, 140 TF
Yayoi
Hitachi, 54.9 TF
OBCX
Fujitsu, 6.61 PF
Oakleaf-FX
Oakbridge-FX
Fujitsu, 1.13+0.13PF
Wisteria/BDEC-01
Fujitsu, 33.1 PF
BDEC-02
150+ PF
Reedbush-U/H/L
SGI-HPE, 3.36 PF
mdx: Data Platform
Fujitsu
CPU
GPU
CPU/GPU
CPU/GPU
Cloud

まとめ

- Miyabi (OFP-II)の運用開始間近 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
    HPL性能 46.8 PFLOPS, OFP (13.55 PFLOPS)の**3.46倍**
    - Miyabi-G: 演算加速ノード, NVIDIA GH200 Superchip
      - CPU-GPU間はNVLink-C2Cにより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
      - HBM2eのみ使用 (DIMMは無し)
      - 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
- BDEC計画 第2世代: BDEC-02
  - mdxの機能を包含できるように ➔ mdx3 …?
  - インタラクティブ、リアルタイム処理 + バッチ処理を1システムに統合

利用説明会＋見学

- 日時：**2025 年 1 月 16 日(木)14:00～17:00**
- 場所：東京大学 柏キャンパス 第2総合研究棟3階 315会議室 or オンライン
- 詳細: https://www.cc.u-tokyo.ac.jp/events/seminar/20250116.php

申し込みはこちらから