Miyabiのストレージシステム

建部 修見
筑波大学計算科学研究センター

University of Tsukuba
Center for Computational Sciences

Supercomputers in CCS and JCAHPC

- CCS
  - Cygnus – 2.3PFlops GPU/FPGA multi-hybrid accelerated supercomputer (2019.4 – 2025.3)
  - Pegasus – 8.1PFlops big memory supercomputer (2023.4 – 2028.1)
  - Post Cygnus (tentative) under procurement – unified memory supercomputer (2025.10 –)
- JCAHPC
  - Oakforest-PACS – 25PFlops manycore supercomputer (2017.4 – 2022.3)
  - Miyabi – 80PFlops GH and Xeon CPU Max supercomputer (2025.1 –)

JCAHPC

- Joint Center of U Tsukuba and U Tokyo
- Established in 2013
- The first system is Oakforest-PACS (OFP) in 2016, the fastest supercomputer in Japan
  - OFP was shutdown in March 2022

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

May 10th, 2016

Joint Center for Advanced High Performance Computing (JCAHPC)

JCAHPC introduces the Post T2K (Oakforest-PACS) system with 25 PFLOPS based on Intel®'s next-generation manycore processors

Summary

Joint Center for Advanced High Performance Computing (JCAHPC) decided to introduce its new supercomputer system (Post T2K system) with 25 PFLOPS peak performance. The new system is named "Oakforest-PACS" with 8,208 of next-generation manycore processors developed by Intel Corporation, and is planned to start its full operation on December 1st, 2016. The Oakforest-PACS will be the fastest supercomputer system in Japan at that time. The Oakforest-PACS will be installed at the Kashiwa Research Complex II building in the Kashiwa-no-Ha (Oakleaf) campus, the University of Tokyo.

JCAHPC (http://jcahpc.jp/) was established in 2013 under agreement between the Center for Computational Sciences, University of Tsukuba (CCS) and the Information Technology Center, the University of Tokyo (ITC). JCAHPC consists of more than 20 faculty and staff members from both CCS and ITC. The primary mission of JCAHPC is designing, installing and operating the Post T2K System (Oakforest-PACS). CCS and ITC will cooperate for the procurement, installation and operation of the Oakforest-PACS under JCAHPC agreement. In addition, CCS and ITC will develop system software, numerical libraries, and large-scale applications for the Oakforest-PACS in collaboration made possible by the establishment of JCAHPC. JCAHPC is a new model for collaboration in research and development between supercomputer centers in Japan. JCAHPC is an advanced organization based on T2K Open Supercomputer Alliance by University of Tsukuba, University of Tokyo and Kyoto University (http://www.open-supercomputer.org/) established in 2006, where each of three universities introduced its own system, respectively.

Design of Miyabi

- Target performance: 100 to 150 PFlops
- GPU required for that performance
  - Most application codes of U Tokyo and Fugaku do not use GPU
  - GPU architecture selected at least 18 months before the operation to ensure a transition period
- GPU selection
  - Evaluate performance, ease of porting, and technical support using seven benchmarks from U Tsukuba and U Tokyo
  - Winner is NVIDIA
  - Major applications have been ported and optimized for NVIDIA GPU

Design of Storage System for Miyabi

- Estimated memory size: 500 TB
- Requested Storage Size and Performance
  - 25 PB capacity # 50x memory capacity (~= capacity for OFP)
  - 1.2 TB/s BW # 5 min to dump 70% of memory
  - 100 KIOPS for file creation # for 20K processes
- All flash not possible within budget
- Investigating a hybrid storage of flash and HDD
- It turned out the overhead of moving data between tiers has a big impact on storage performance
- QLC can solve the problem, although some capacity is sacrificed

Miyabi (OFP-II)

- Announced in November 2023
- Starting from January 2025
- **80.1 PFlops (3.2x OFP)**
  - Miyabi-G: 1,120 nodes of Grace Hopper
  - Miyabi-C: 190 nodes of Xeon CPU Max
- 11.3 PB, 1.0 TB/s, 100 KIOPS
  - **All flash Lustre**

JCAHPC decided to introduce "OFP-II"

Overview

- JCAHPC decided to introduce a new supercomputer system "OFP-II" with a total peak performance of 79.5 PFLOPS.
- OFP-II will be jointly procured and operated by the University of Tokyo and the University of Tsukuba.
- OFP-II will be the first large-scale system with NVIDIA GH200 Grace Hopper Superchip
- Operation will be started in January 2025.

Background

In March 2013, Center for Computational Sciences, University of Tsukuba (CCS), and Information Technology Center, the University of Tokyo (ITC) agreed to establish the Joint Center for Advanced High-Performance Computing (JCAHPC). JCAHPC is an organization to build and operate a stat-of-the-art, large-scale, high-preformance computing infrastructure by introducing supercomputer systems to be jointly procured, and has aimed to provide large-scale, ultra-high-speed computing capabilities to promote cutting-edge computational science and to contribute to the promotion of academics and science and technology in Japan. As a result, Oakforest-PACS started operation in October 2016 and ended in March 2022 after 5 years and 6 months of operation.

JCAHPC has been considering the design, construction, and operation of the successor to the Oakforest-PACS since November 2019. JCAHPC decided to introduce **a new supercomputer system "OFP-II" with a total peak performance of 79.5 PFLOPS**. The operation is scheduled to start in January 2025. Inheriting the philosophy of the introduction of Oakforest-PACS, OFP-II aims to promote novel computational science methods leveraging AI, such as AI for HPC/Science, to offer a platform supporting Society 5.0 by integration of simulation, data analysis, and machine learning, in addition to support users of large-scale applications.

Specification of OFP-II system

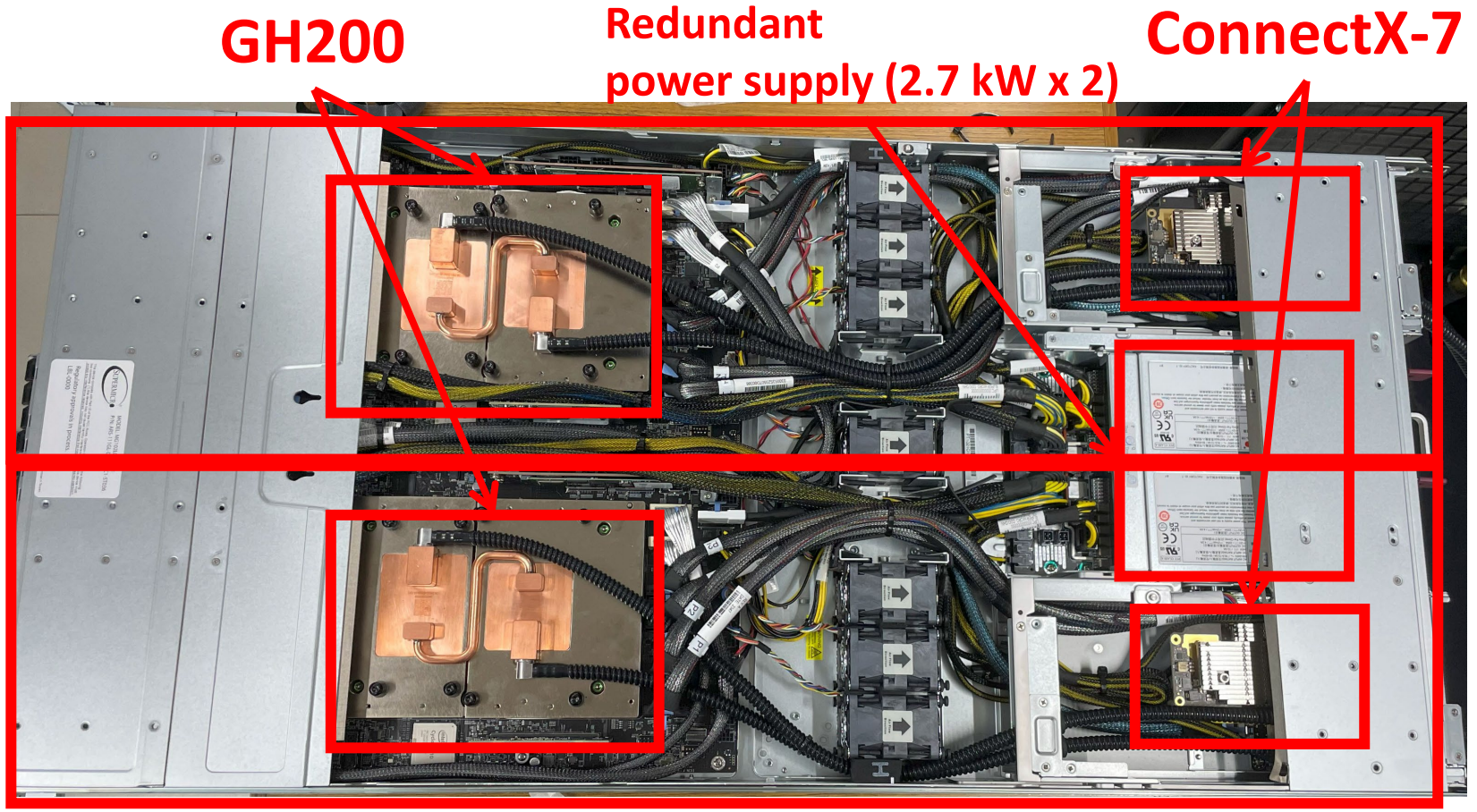
- Theoretical Peak Performance: 79.5 PFLOPS
- OFP-II has General-purpose CPU node adnd Accelerator node
  - General-purpose CPU node: 190 nodes with two Intel Xeon CPU Max 9480 per node
  - Accelerator node: 1,120 nodes with NVIDIA GH200 Grace Hopper Superchip per node
- Shared File System: NVMe SSD 10.3 PB

Miyabi

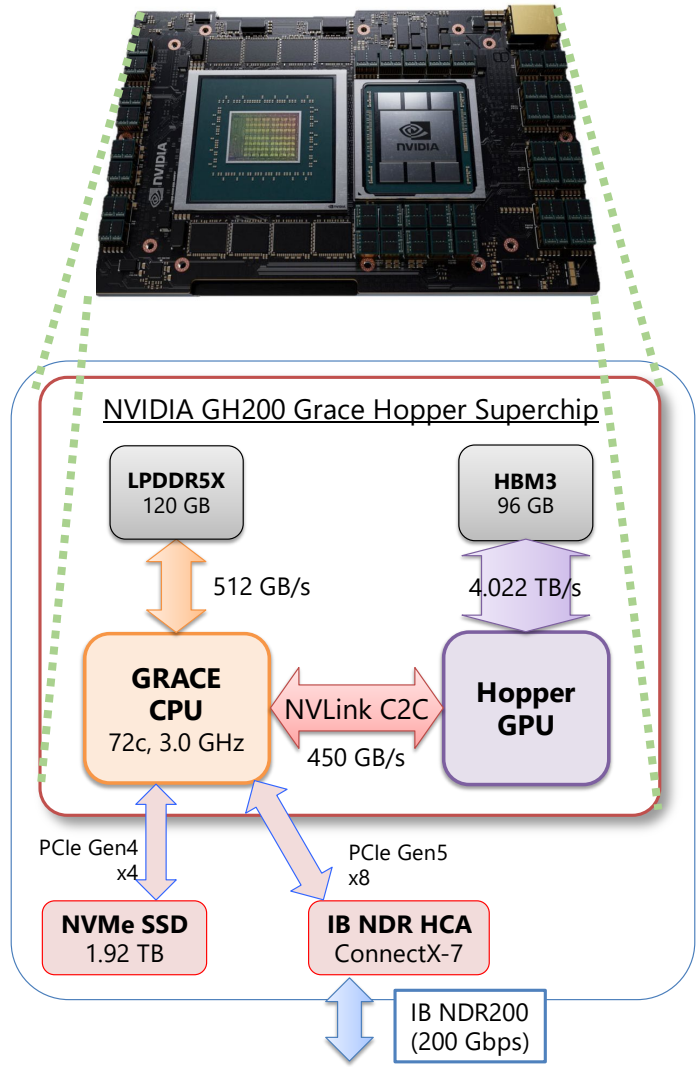
Half of Miyabi-G (the other half is on the back side)

Miyabi-G Compute Node

- 2 nodes in 1U, Direct Liquid Cooling

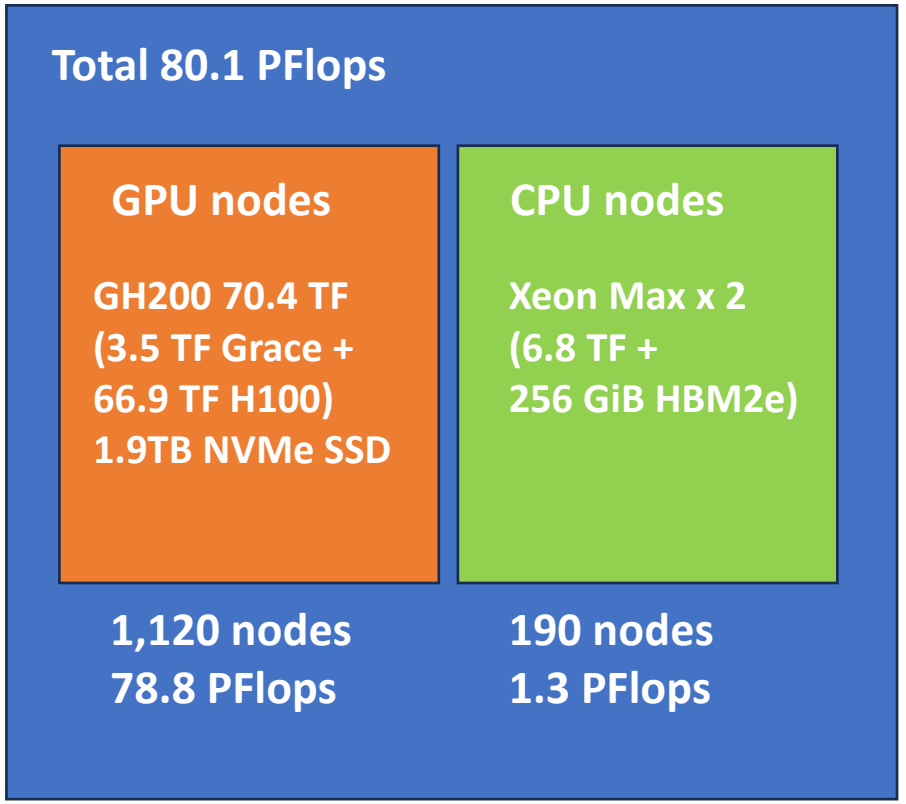


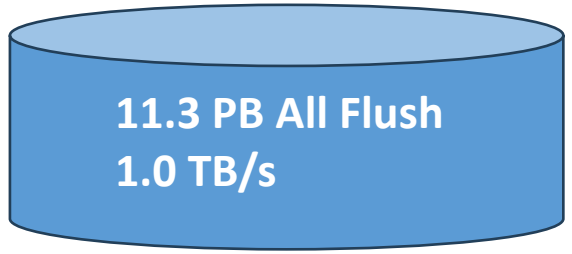
GH200



Specification of Miyabi

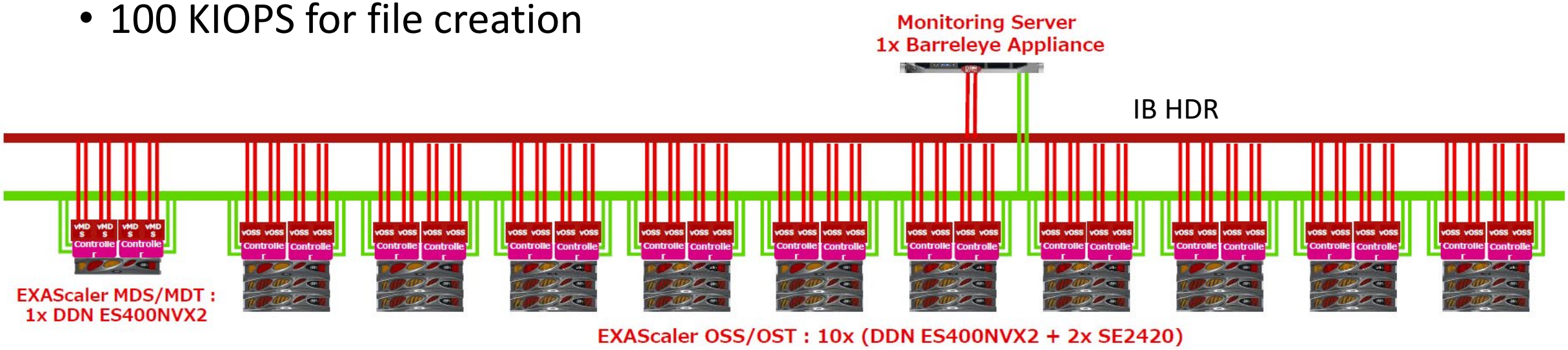
Total Performance	**80.1** PFlops	
	GPU Nodes	CPU Nodes
Performance	78.8 PFlops	1.3 PFlops
# nodes (# racks)	1,120 (19 racks)	190 (3 racks)
Processor	**GH200 70.35 TF** (72c, 3.45TF Grace + 66.9TF H100)	**Xeon Max** 9450 **6.8 TF** (56c, 3.4TF) x 2
Memory	201.2 GiB (111.8GiB LPDDR5X (512 GB/s) + 89.4 GiB HBM3 (4 TB/s))	128 GiB HBM2e (3.2 TB/s)
NVMe SSD	1.9 TB / 8.0 GB/s	-
Network	InfiniBand NDR 200	
Storage	All Flush DDN EXAScaler 11.3 PB, 1.0 TB/s, 23.5G inodes, 100 KIOPS	



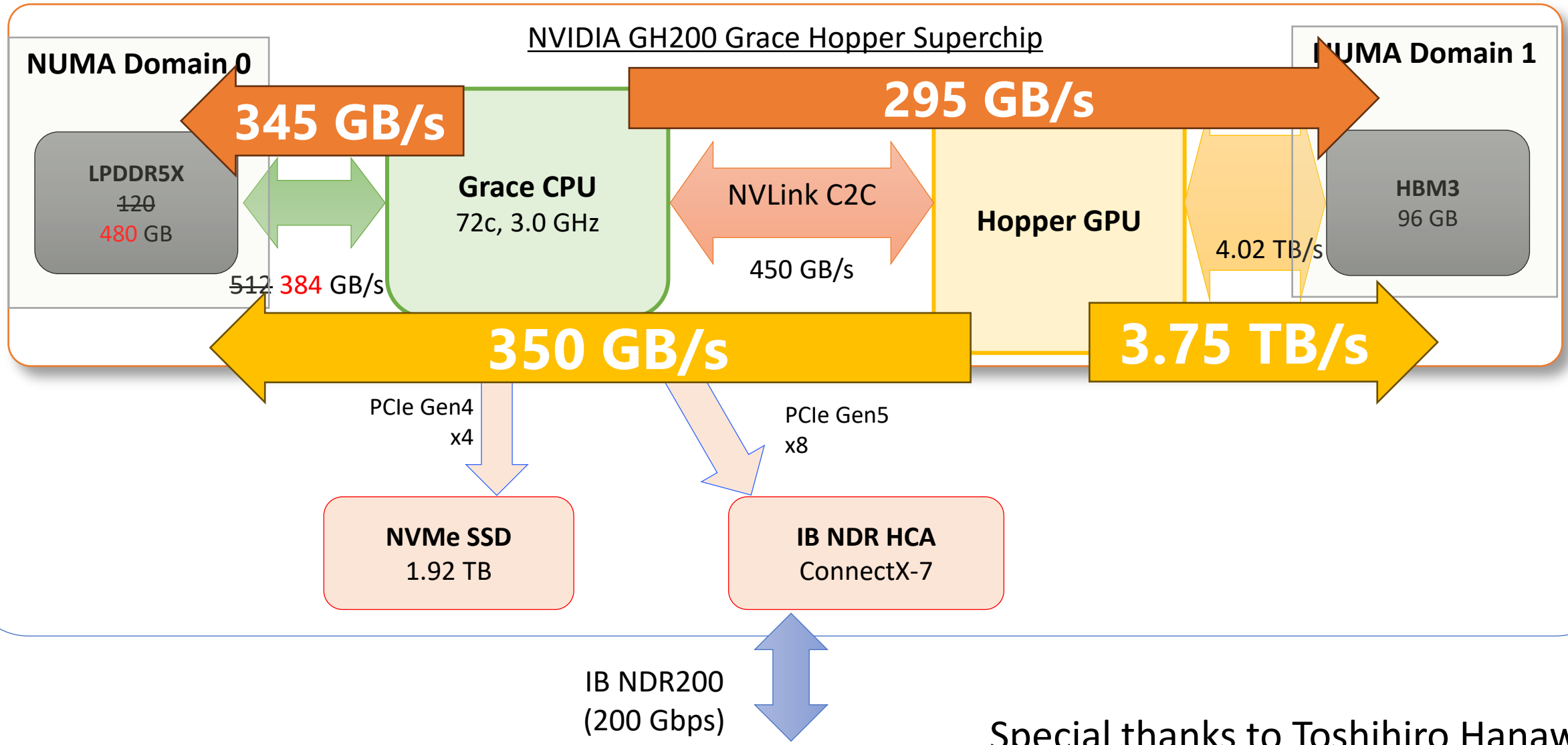


All flash Lustre of Miyabi

- Effective capacity: 11.3 PB (DCR RAID6, 30TB QLC 460 + 20)
  - 8 MDT (4 MDS), 80 OST (40 OSS)
- Theoretical Peak BW: 1.0 TB/s
- Effective Write BW: 0.65 TB/s
- Effective Read BW: 0.85 TB/s
- # inodes: 23.5 Billion
- 100 KIOPS for file creation

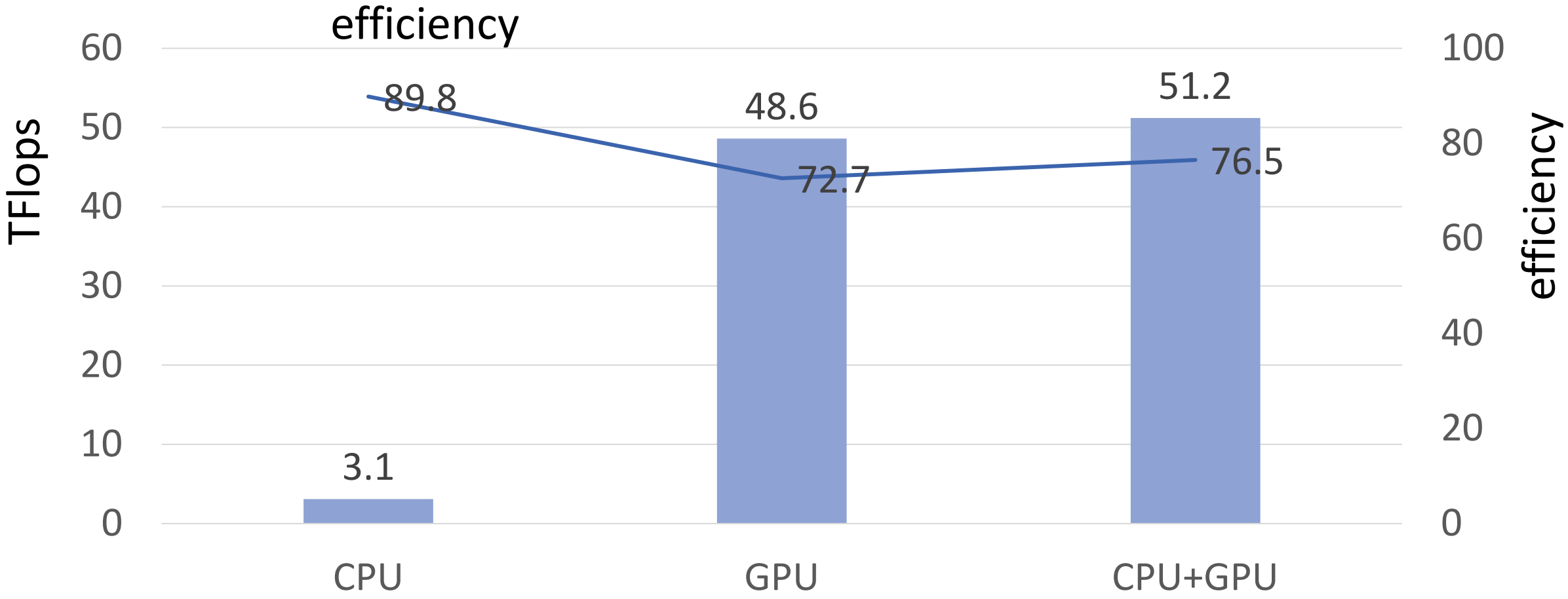


Memory bandwidth of air-cooled eval machine



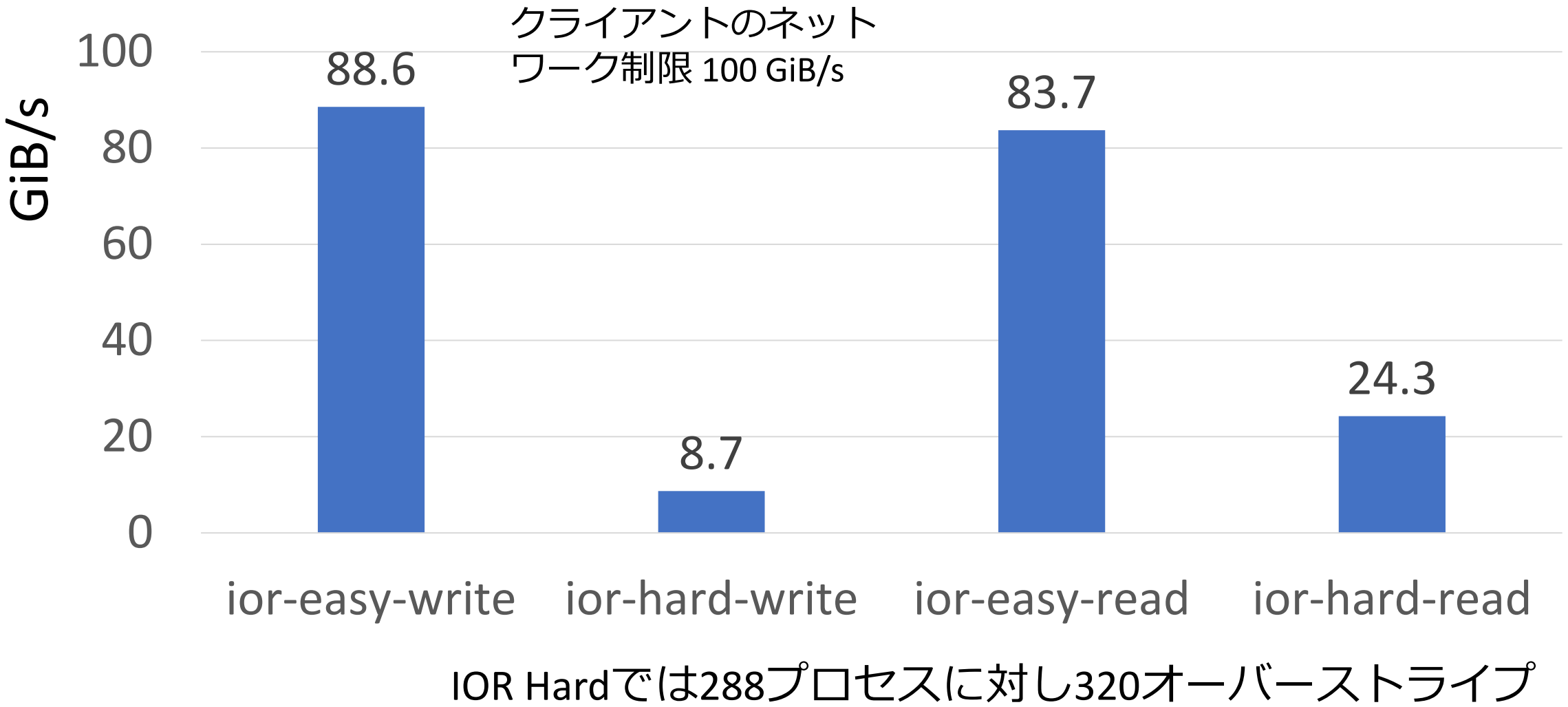
Special thanks to Toshihiro Hanawa

HPL performance (1 node)



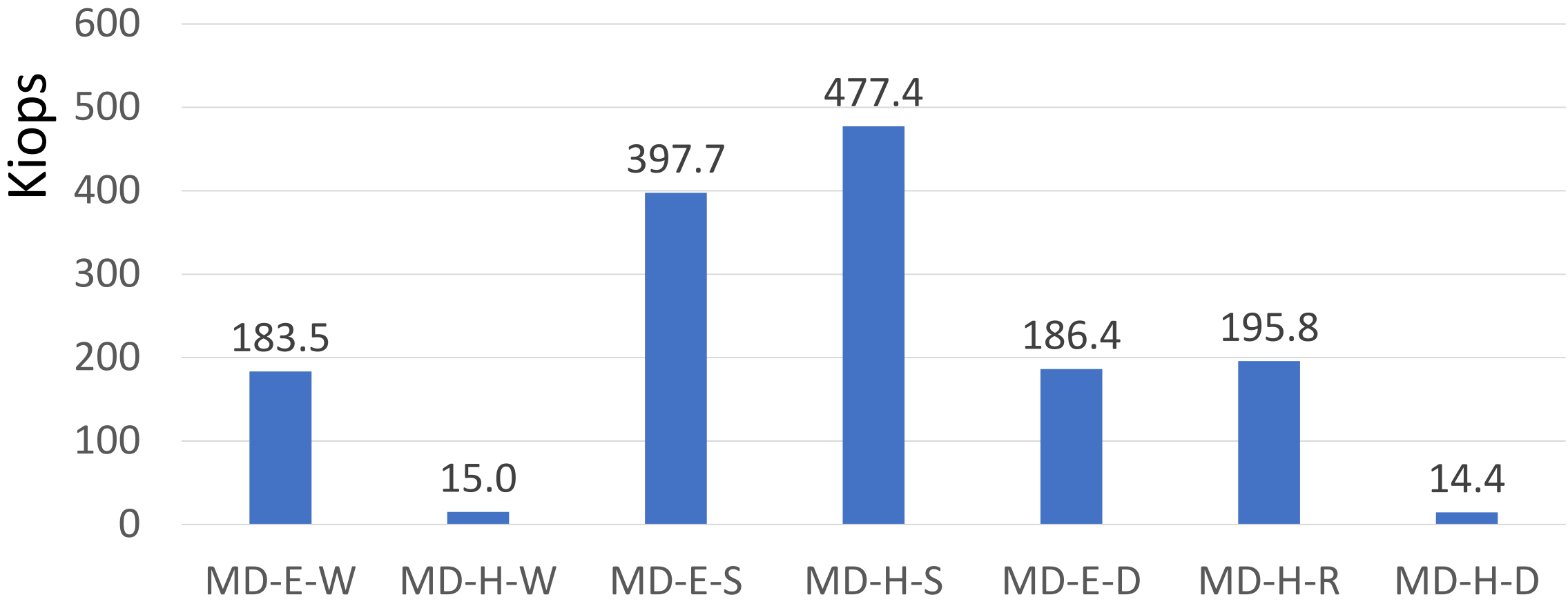
Special thanks to Toshihiro Hanawa

IO500 Bandwidth (4 nodes, 72 ppn, 60 sec)



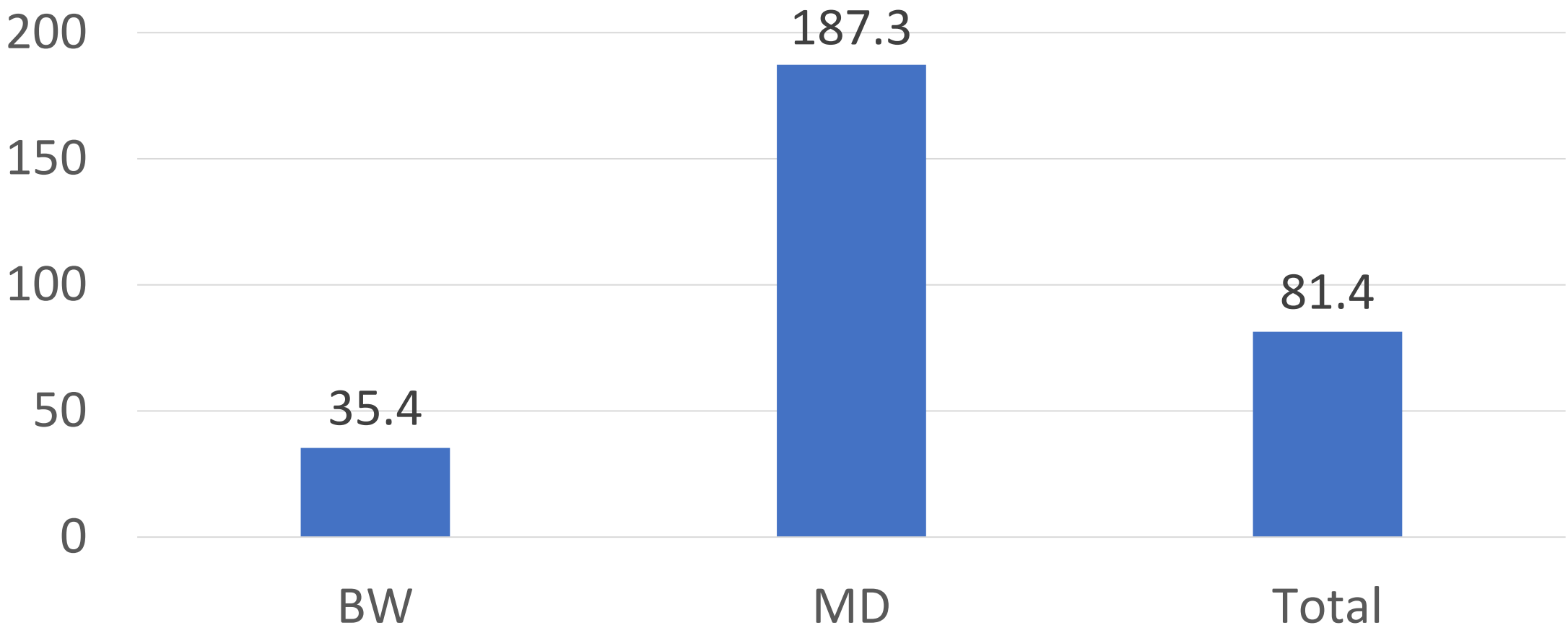
IOR Hardでは288プロセスに対し320オーバーストライプ

IO500 Metadata (4 nodes, 72 ppn, 60 sec)



8 MDTにディレクトリストライプ、64KiB DoM利用

IO500 (4 nodes, 72 ppn, 60 seconds)



学際共同利用プログラム

- 目的
  - 全国の学際的計算科学の発展に資する
- 研究分野
  - 素粒子、宇宙、原子核、物質科学、生命、地球環境、生物、化学、HPC、計算情報学、数値解析
- 募集期間
  - 12月中旬～1月後半
- 申込
  - https://www.ccs.tsukuba.ac.jp/kyodoriyou/gakusai/

一般利用

- 利用目的
  - 学術研究
- 申請資格
  - 学術研究を目的とする公的な機関に所属し、計算科学関連分野の研究を行う者
- 利用資格
  - 申請者と共同研究を行う民間企業を含む研究者
- 申請期間
  - 随時

学際ハブ拠点スパコンお試し利用

- 事業の目的
  - 新たな需要の開拓、企業での計算科学的手法の導入の裾野拡大
- 対象者
  - CCSと共同研究を継続的に実施する企業
- 制度
  - センター教員との共同研究を前提としたスパコンお試し企業利用
- 申込
  - https://www.ccs.tsukuba.ac.jp/kyodoriyou/hub-trial/

Summary

- Miyabi will be introduced in Jan 2025
  - 80PFlops GH and Xeon Max CPU supercomputer
- 11.3 PB 1.0 TB/s All flash storage system
  - 4-node IO500 score 81.4 (63rd 10-node research)