

# Towards Next JCAHPC System

Toshihiro Hanawa

Information Technology Center

The University of Tokyo

# Agenda

- Introduction of JCAHPC
  - Oakforest-PACS
- Design for next “Oakforest-PACS II” system
- Porting to GPU
- Projects on JHPCN
  - GPU Direct Storage
  - Programming for GPU
- (Feasibility Study for “Fugaku-next”, Operation Technical Research Team)
- (Mixed Precision / Transprecision using FPGA)

# JCAHPC:

## Joint Center for Advanced HPC



筑波大学  
University of Tsukuba



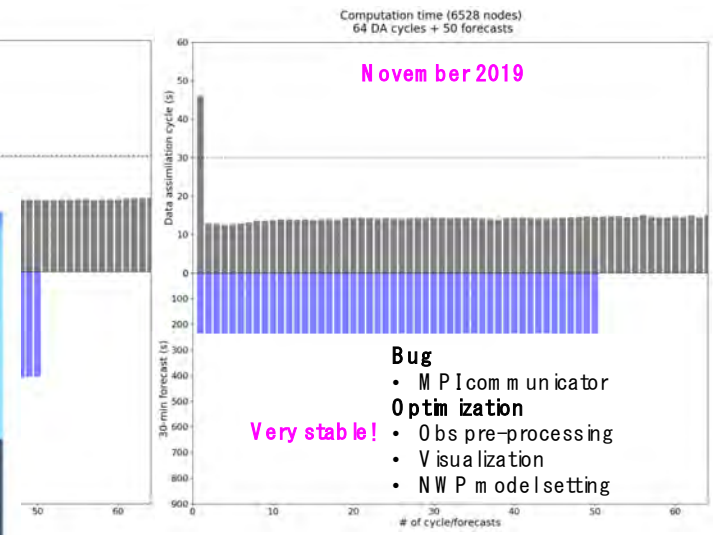
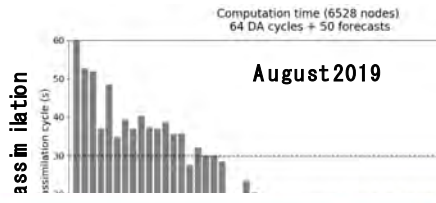
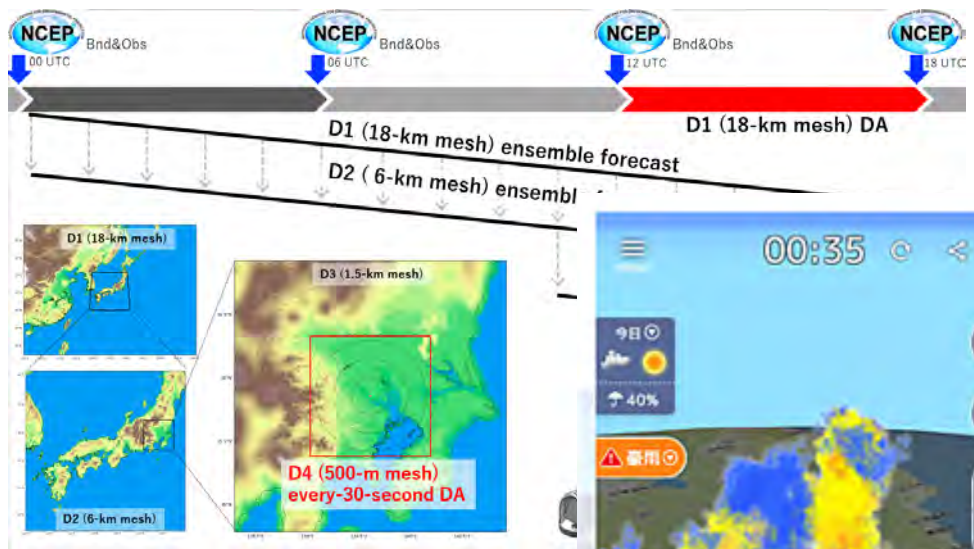
東京大学  
THE UNIVERSITY OF TOKYO



- Established in 2013
  - University of Tsukuba & University of Tokyo
    - Budgets of 2 Supercomputing Centers are combined
  - Promotion on Computational Science, Design/Procurement/Operation of Large-scale Systems
- Oakforest-PACS (OFP), 1<sup>st</sup> System of JCAHPC
  - 8,208 Intel Xeon Phi (KNL), 25PF, Fujitsu
  - Top500 (#6 (Nov.2016), #1 in Japan)
    - National Flagship System “in fact” (Oct.2019-Mar.2021) after shutdown of the K computer
  - Retired on March 31, 2022 (#39 (Nov.2021))
- We are starting procurement for OFP-II, successor of OFP, whose operation starts in April 2024



# Real-Time Prediction of Severe Rainstorm by OFP



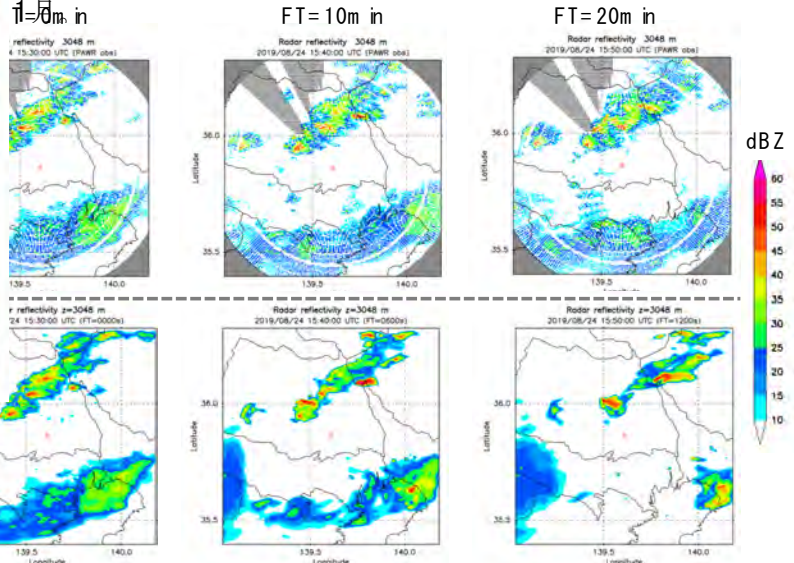
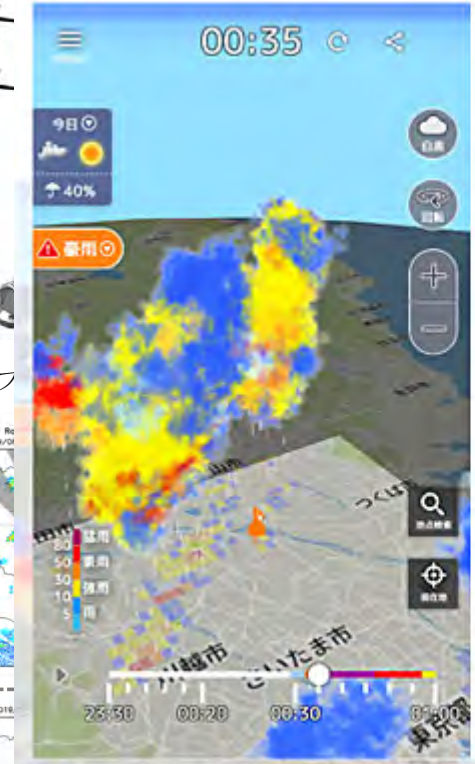
全体のワークフロー

タ同化、下段は30分予報にかかった時間(秒)。



PAWR Obs

SCALE-LETKF Analysis



SCALE Forecast

[c/o Dr. Takemasa Miyoshi (RIKEN R-CCS)]

2019年8月24日の事例についてのテスト結果。(上)レーダー観測と(下)SCALE-LETKFによる解析で得られたレーダー反射強度 (dBZ) を示す。

2019年8月24日の事例についてのテスト結果。(上)レーダー観測と(下)SCALE-LETKFによる予報で得られたレーダー反射強度 (dBZ) を示す。

# HPCI Urgent Call for Fighting against COVID-19 in Japan (FY.2020)

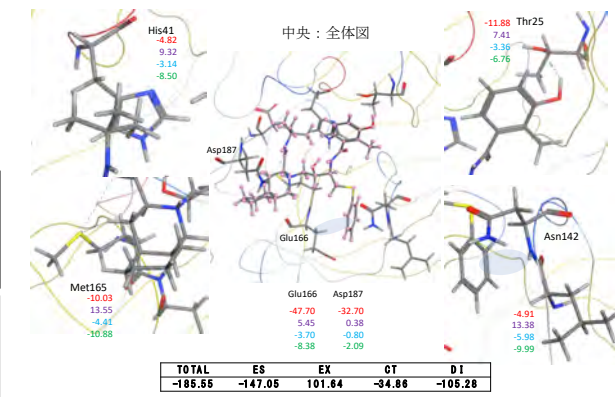
by 8 SC Centers of Natl. Univ., AIST etc.

6 of 14 accepted projects use U.Tokyo's Systems



Project Name	PI	System
Fragment molecular orbital calculations on the main protease of COVID-19	Yuji Mochizuki (Rikkyo U.)	OFP
Study on the evaluation of arrhythmogenic risk of COVID-19 candidate drugs	Toshiaki Hisada (UT Heart)	
Prediction of dynamical structure of Spike protein of SARS-COVID19	Yuji Sugita (RIKEN)	
Computer-assisted search for inhibitory agents for SARS-CoV-2	Tyuji Hoshino (Chiba U.)	OBCX
Prediction and Countermeasure for virus droplet Infection under Indoor Environment: Case studies for massively-parallel simulation on Fugaku	Makoto Tsubokura (Kobe U.)	
Spreading of polydisperse droplets in a turbulent puff of saturated exhaled air	Marco Edoardo Rosti (QIST)	

Workshop on Challenges for New Extreme Scale Applications

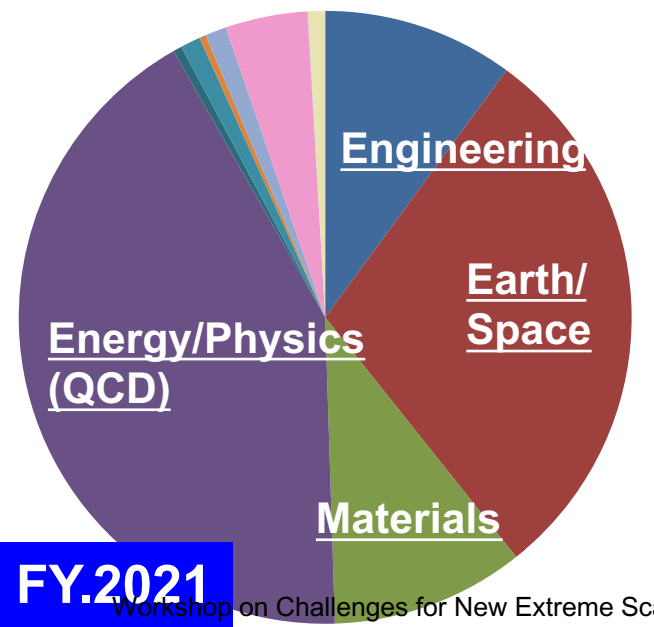
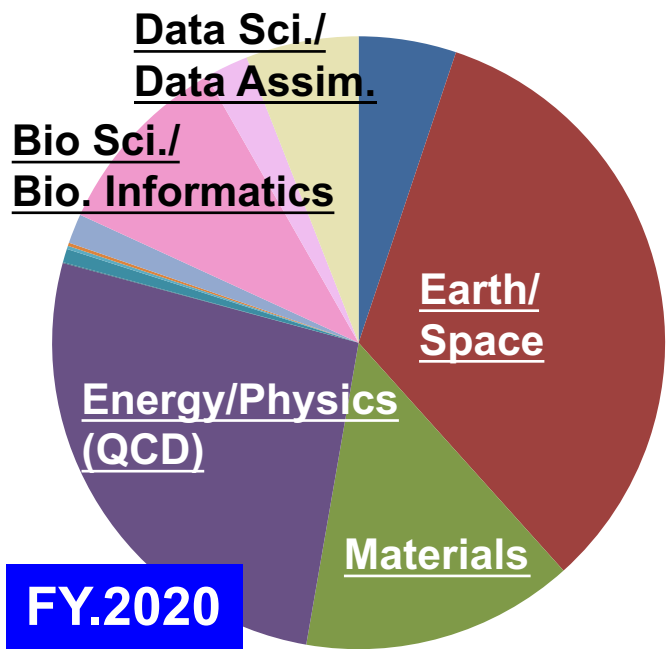
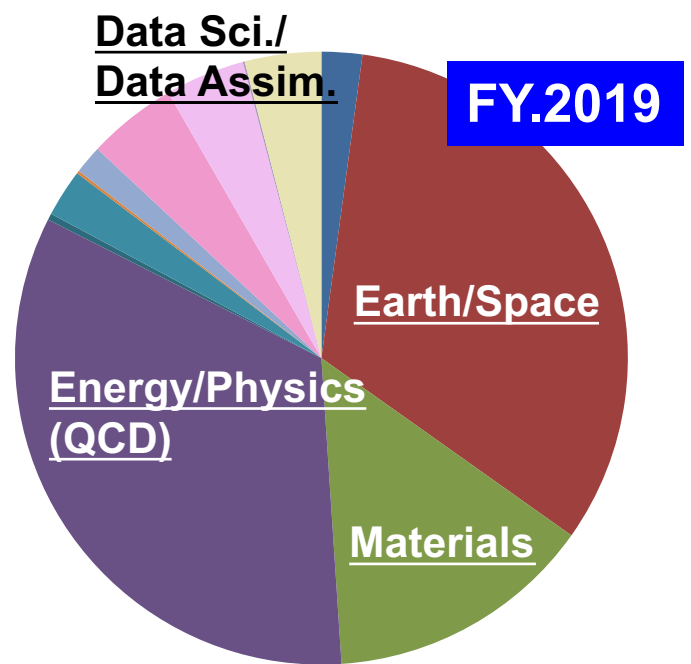
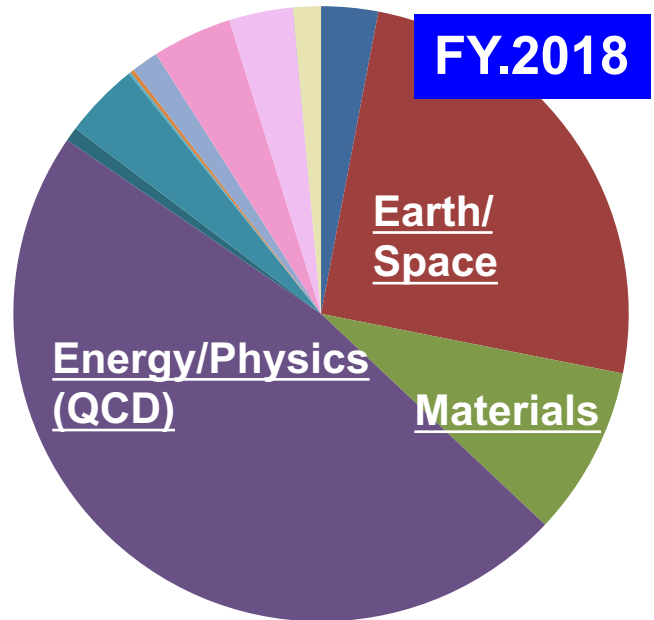
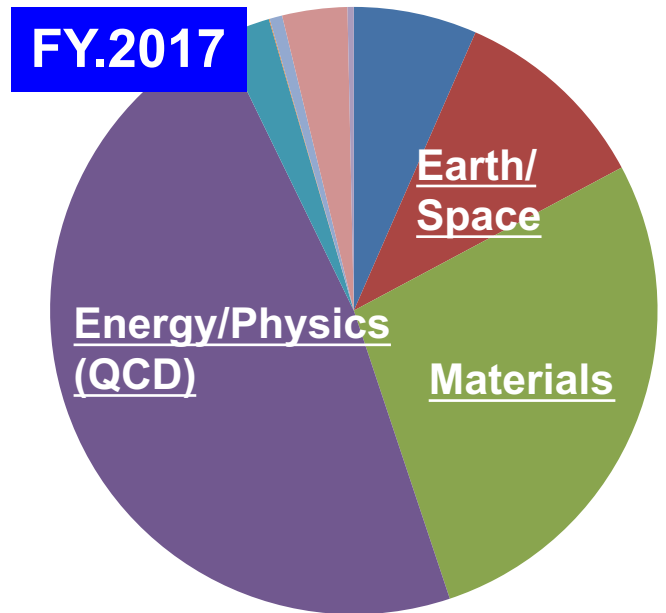


[c/o Prof. Y. Mochizuki (Rikkyo U.)]



[c/o Prof. M.Tsubokura (Kobe U.)]

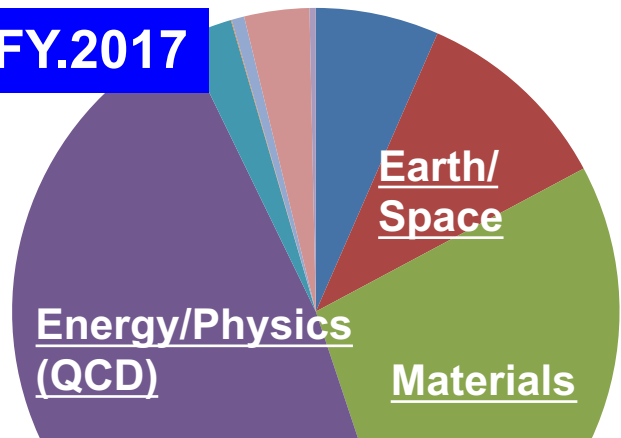
# How was OFP used ...



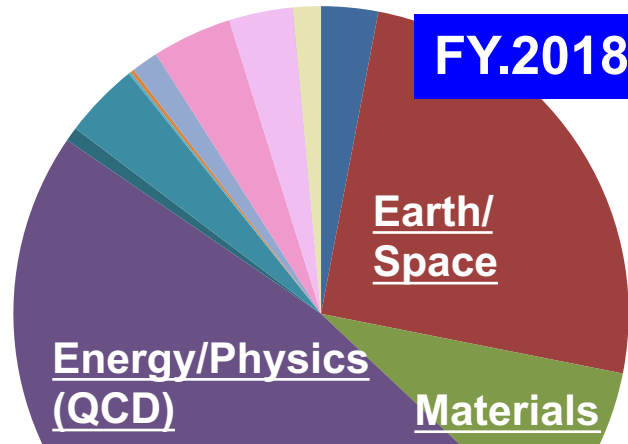
- Engineering
- Earth & Space
- Materials
- Energy & Physics
- Info: System
- Info: Algorithms
- Info: AI
- Education
- Industry
- Bio Science
- Bio Informatics
- Social Science & Economy
- Data Science & Data Assimilation

# How was OFP used ...

**FY.2017**

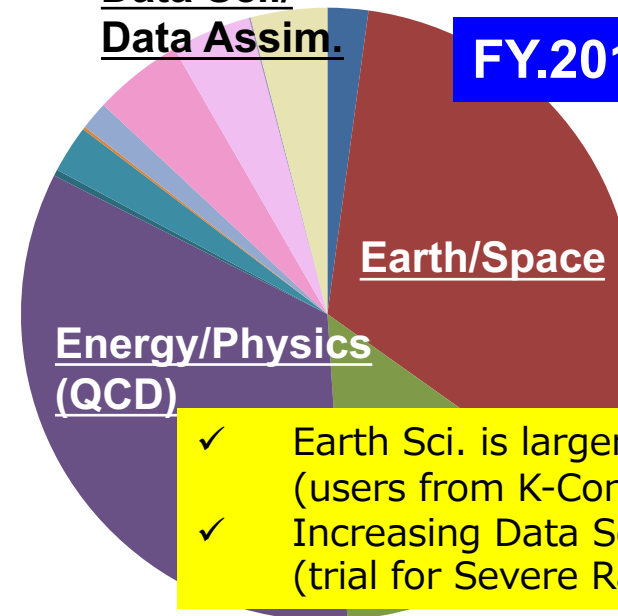


**FY.2018**



Data Sci./  
Data Assim.

**FY.2019**



- ✓ FY2017, 2018: Half occupied by QCD users
- ✓ Increasing Earth/Space (Atmosphere&Ocean, Solid Earth, Astrophysics)
  - ◆ From Oakleaf-FX (FX10)

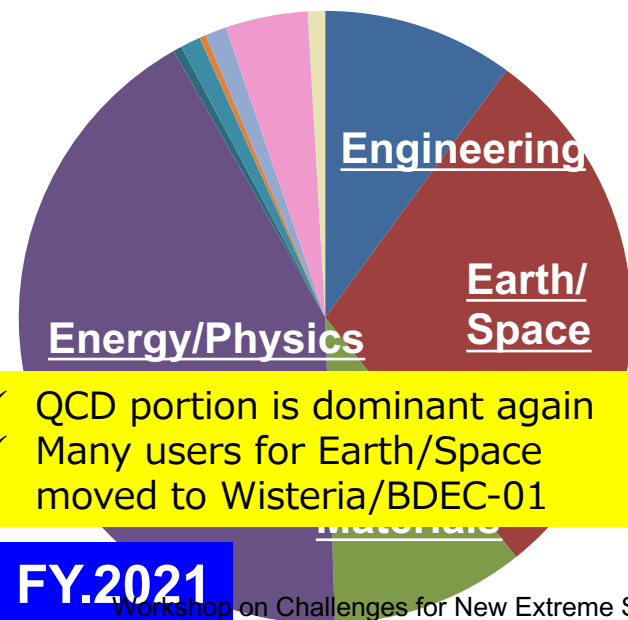
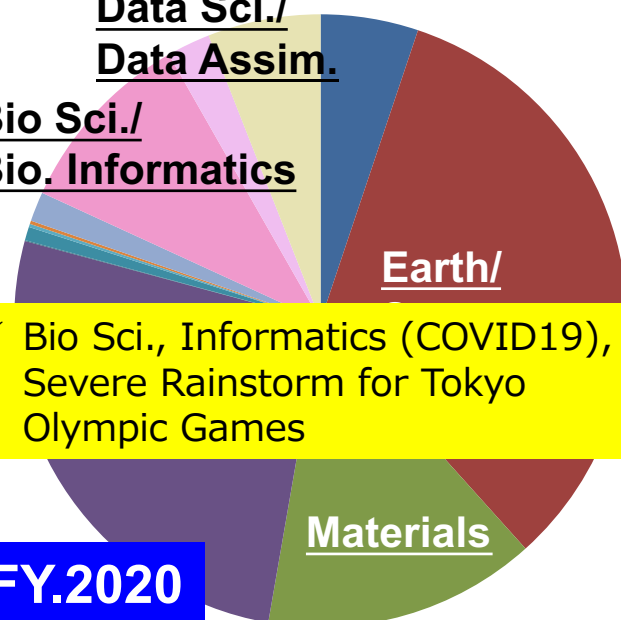
- ✓ Earth Sci. is larger than QCD (users from K-Computer?)
- ✓ Increasing Data Sci. / Assim (trial for Severe Rainstorm)

Data Sci./  
Data Assim.

Bio Sci./  
Bio. Informatics

- ✓ Bio Sci., Informatics (COVID19), Severe Rainstorm for Tokyo Olympic Games

**FY.2020**

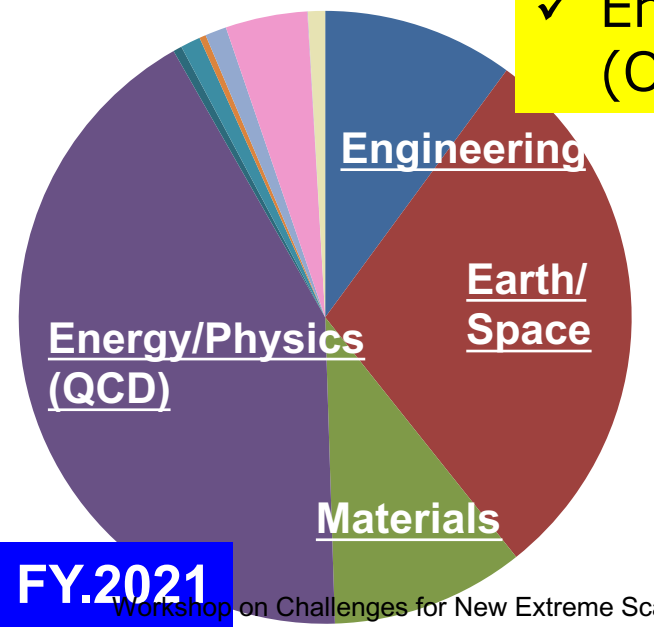
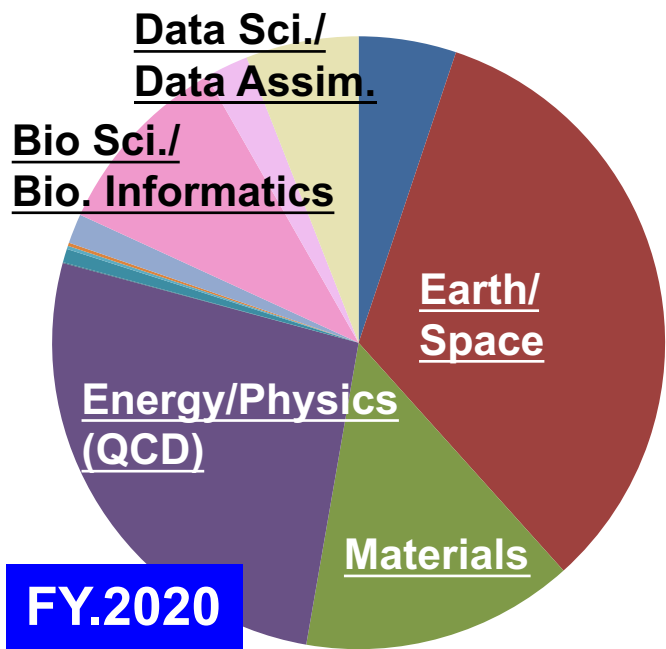
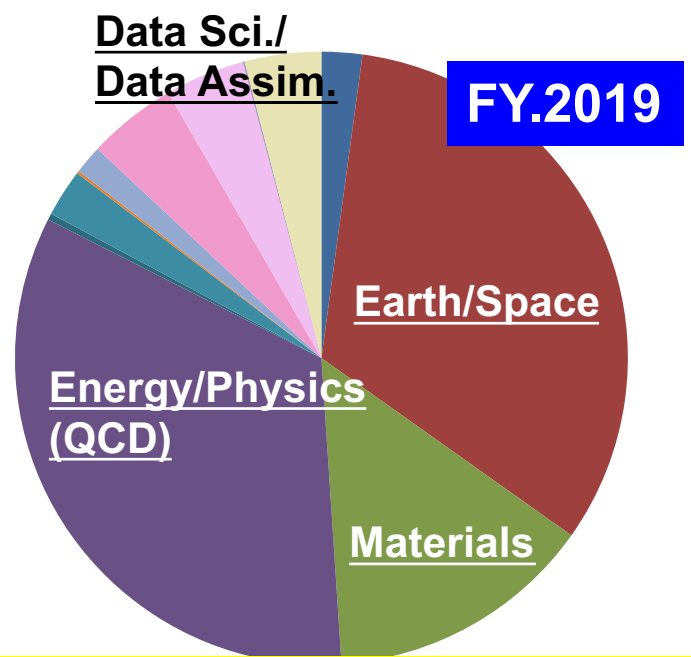
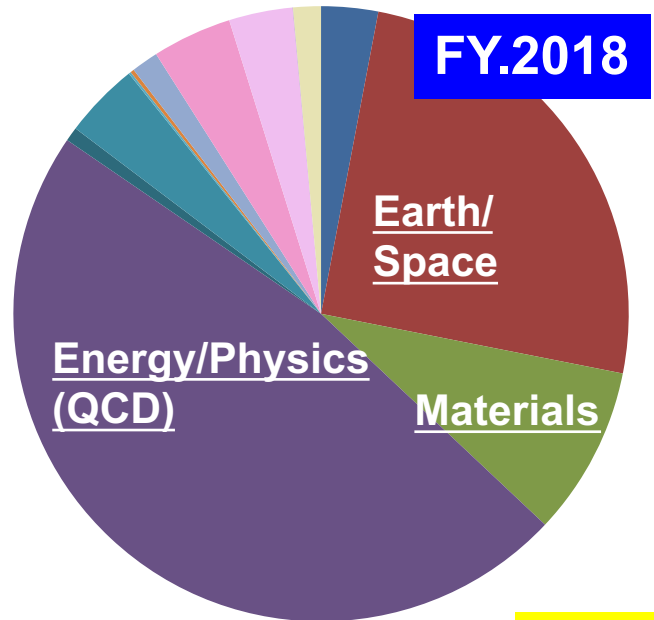
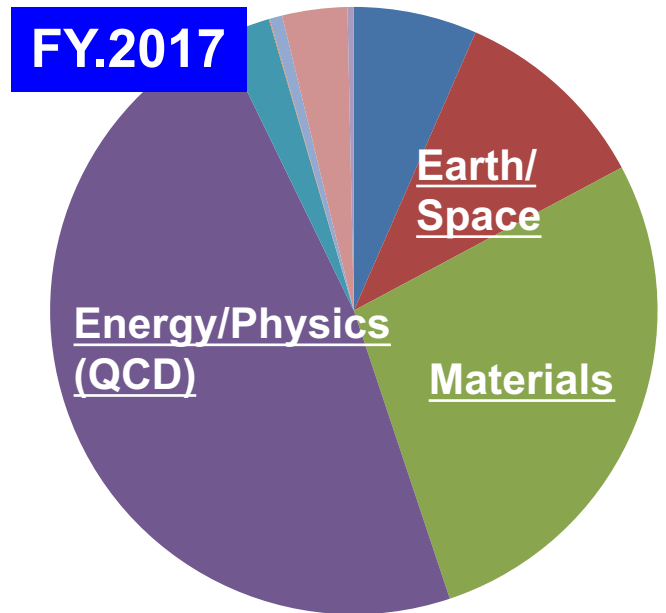


- ✓ QCD portion is dominant again
- ✓ Many users for Earth/Space moved to Wisteria/BDEC-01

**FY.2021**

- Engineering
- Earth & Space
- Materials
- Energy & Physics
- Info: System
- Info: Algorithms
- Info: AI
- Education
- Industry
- Bio Science
- Bio Informatics
- Social Science & Economy
- Data Science & Data Assimilation

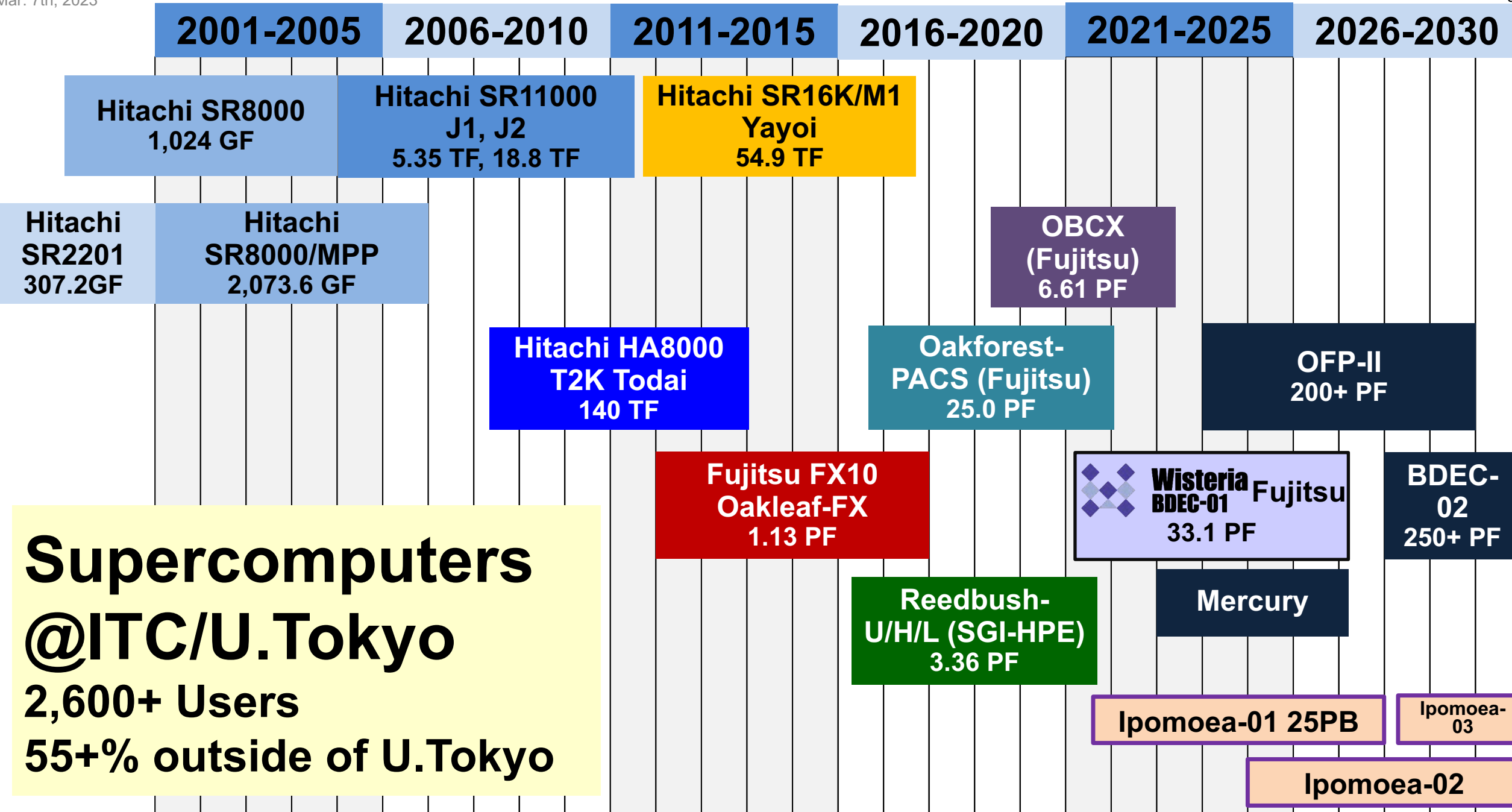
# How was OFP used ...



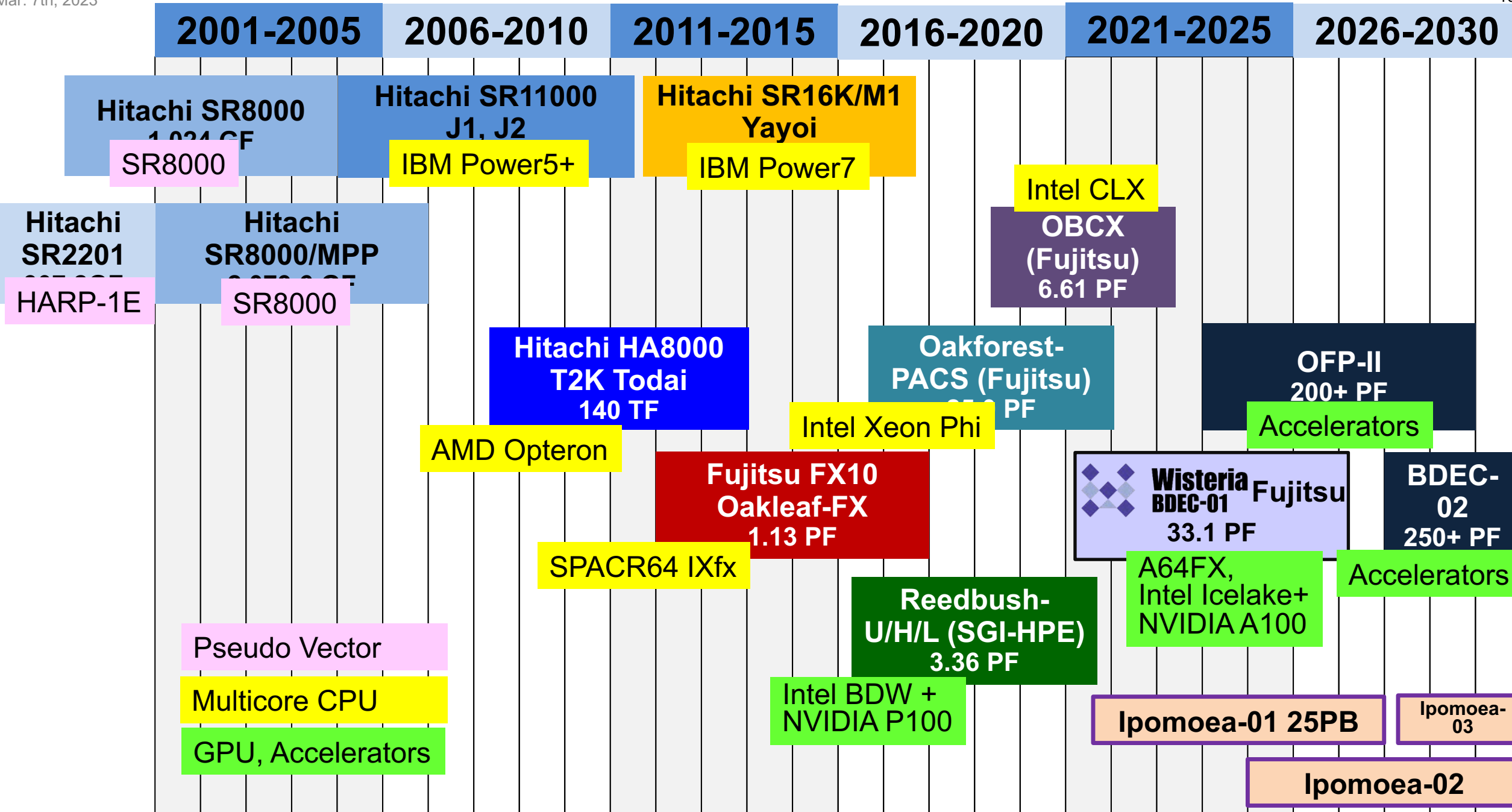
✓ Engineering is not so big ratio (Oakleaf-FX, Reedbush-U, OBCX)

- Earth & Space
- Materials
- Energy & Physics
- Info: System
- Info: Algorithms
- Info: AI
- Industry
- Bio Science
- Bio Informatics
- Social Science & Economy
- Data Science & Data Assimilation



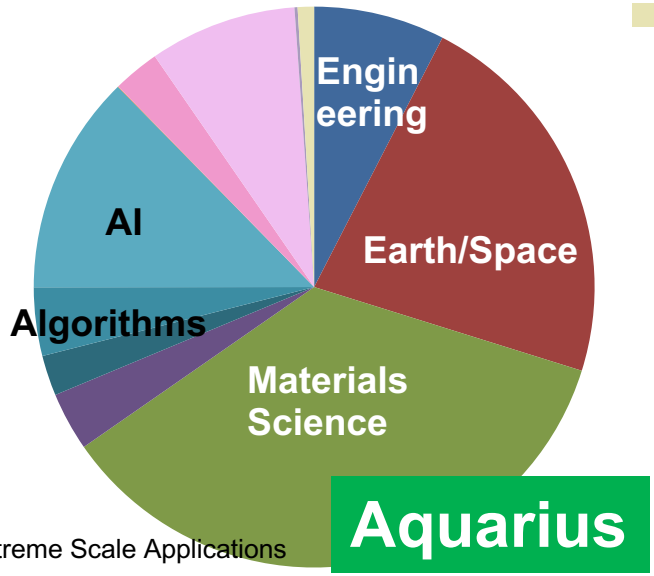
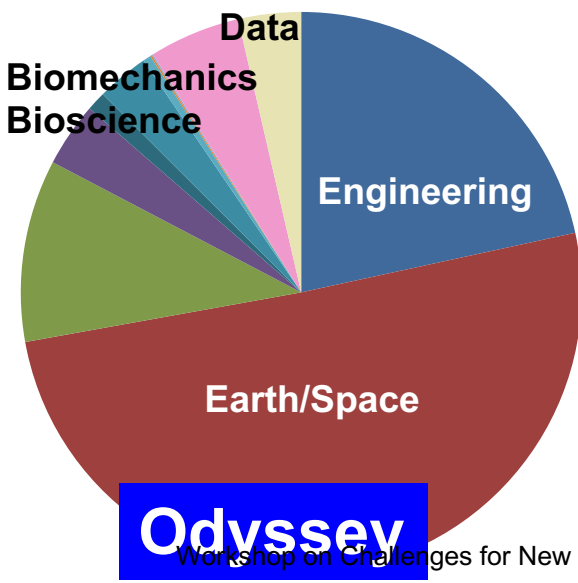
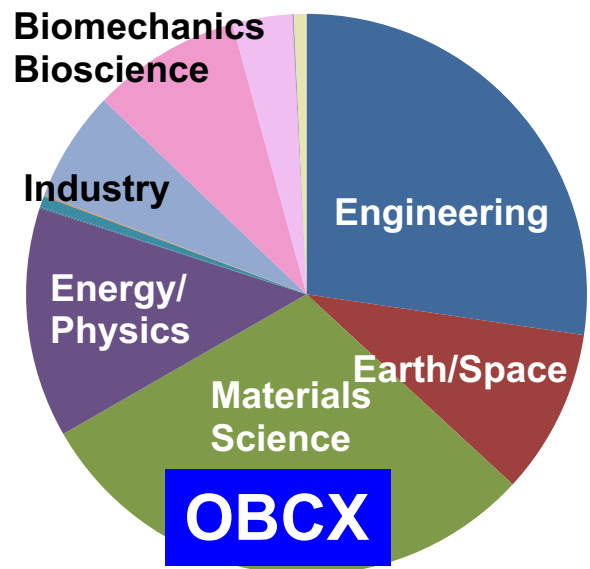
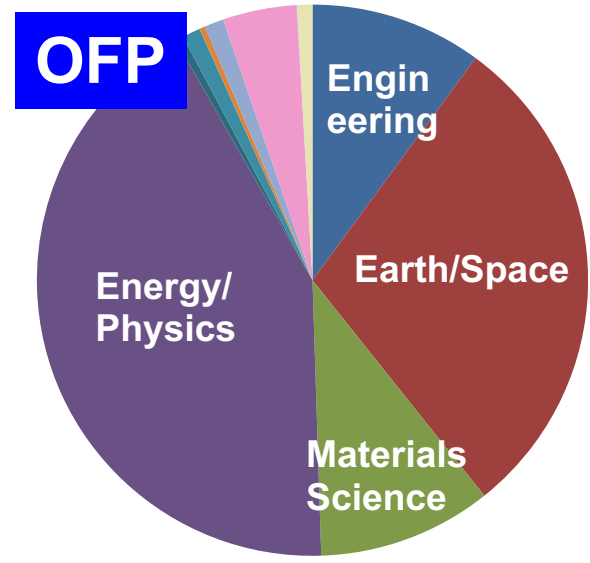
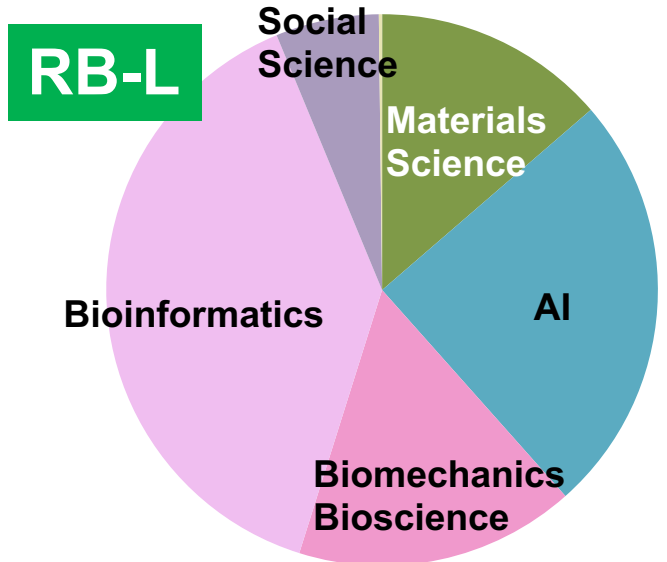
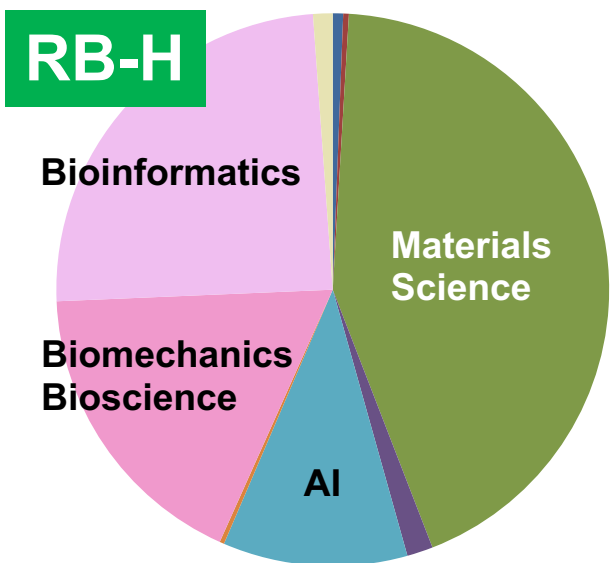


**Supercomputers @ITC/U.Tokyo**  
 2,600+ Users  
 55+% outside of U.Tokyo



# Research Area (FY.2021: U.Tokyo)

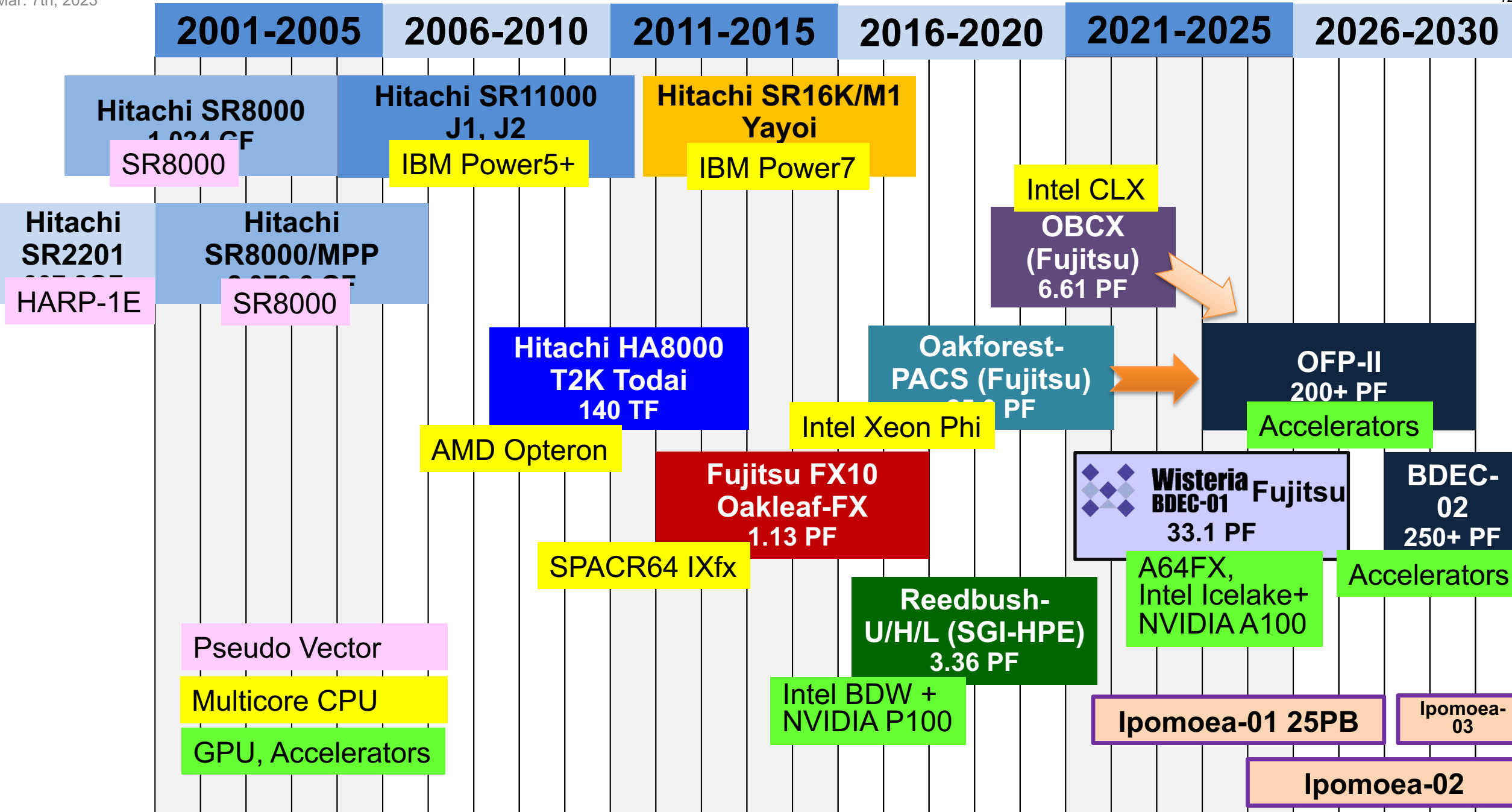
## Odyssey, Aquarius: After Aug., RB-H, RB-L: Nov.E



- Engineering
- Earth/Space
- Material
- Energy/Physics
- Info. Sci. : System
- Info. Sci. : Algorithms
- Info. Sci. : AI
- Education
- Industry
- Bio
- Bioinformatics
- Social Sci. & Economics
- Data

■ CPU

■ GPU



# Towards 2<sup>nd</sup> Gen. System



- Nov. 2019: Agreed to continue design and operation of the 2nd generation system as JCAHPC
- Feb. 2021: Codename as “Oakforest-PACS II”
- Nov. 2022: Start procurement procedure: **Request for Information (RFI)**
- Apr. 2024: Target for operation start
- OFP was the top machine in the HPCI 2<sup>nd</sup> Tier system
  - OFP could run 2,000 node jobs at any time and kept supporting Japanese compute resources between K-computer decommission and Fugaku operation start



- Inherit from **the philosophy of OFP: Support for users of large-scale applications continuously**
- **New Usage: AI for HPC/Science**
- **Co-design with applications**

# Development for New Types of Applications

- Highly-developed simulation by integration of “Simulation+Data+Learning”:  
enabling more tightly-coupled cooperation with data
  - Data-assimilation + Simulation
  - Simulation parameter estimation by machine learning + Ensemble calculation  
→ AI for HPC / Science
  - Advancing efforts on Wisteria/BDEC-01 (U. Tokyo) or Cygnus+Pegasus (U. Tsukuba)
    - Wisteria/BDEC-01 consists of both “Odyssey” (simulation) node and “Aquarius” (Data+Learning) node
    - Cygnus: Extreme Computing with multi-hetero accelerator, Pegasus: Bigdata and AI
- Requirement for Carbon-Neutrality and Power efficiency

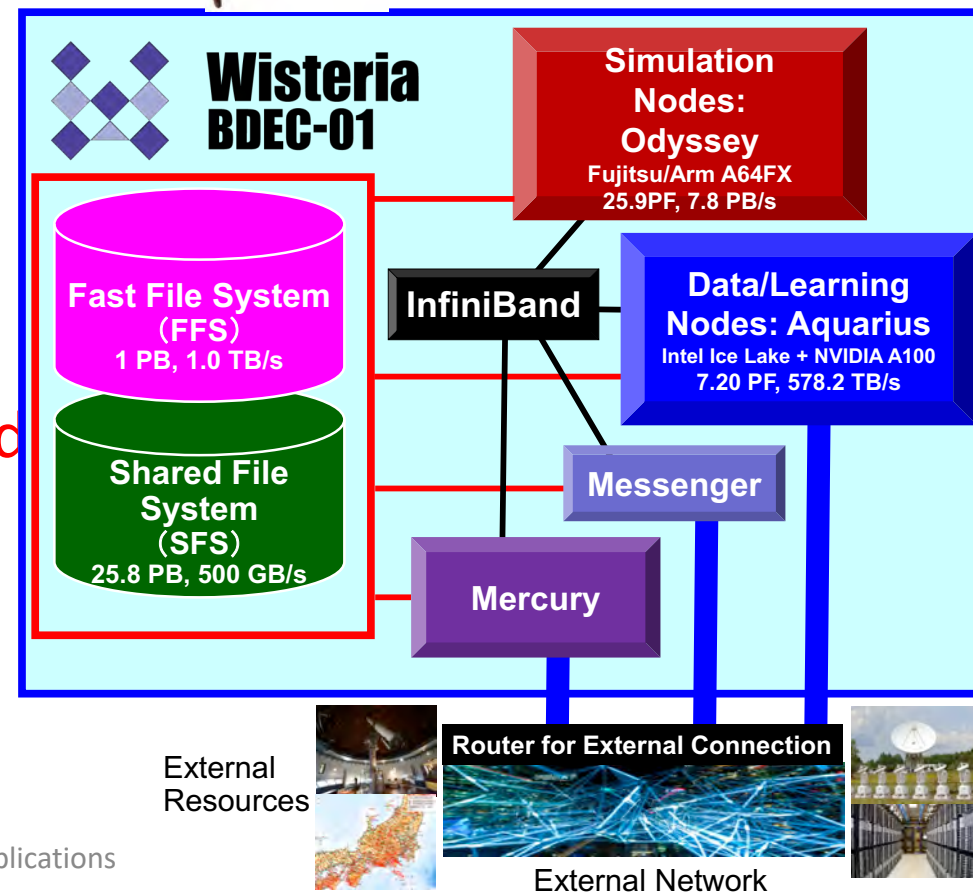


Introduction of GPU for most of compute resources  
(and still needs small ratio of CPU-only nodes)

# Pre-Benchmark for OFP-II Design



- >3,000 users for OFP and our current systems
- It is difficult to perfectly migrate users to GPU cluster
  - U.Tokyo considered this for Wisteria/BDEC-01, but finally gave up (Fall 2019)
- **Varieties of GPU's**
- **Varieties of Programming Environment for GPU**
  - OpenACC
  - OpenMP 5.0 Target offloading
  - Standard Language Support
    - **Better for (OpenMP+MPI) users of OFP than Fall 2019**
  - CUDA / HIP / SYCL
- **Wisteria-Mercury for the prototype of OFP-II: Testbed for Porting Applications to GPU**
  - Original plan was a node-Group extension of Wisteria/BDEC-01
- Pre-benchmarks for a decision of GPU on OFP-II and Mercury
  - Performance Estimation on H/W for OFP-II timeframe



# Seven Benchmarks



筑波大学  
University of Tsukuba



東京大学  
THE UNIVERSITY OF TOKYO


Benchmarking by GPU vendors: Feb. – May. 2022

**A:** Benchmark for General CPU  
**B:** Already GPU-enabled

Code	Description	Lang.	Parallelization	GPU	Category
P3D	3-D Poisson's Equation by Finite Volume Method	C	OpenMP	N/A	A
GeoFEM/ICCG	Finite Element Method	Fortran	OpenMP, MPI	N/A	
H-Matrix	Hierarchical-Matrix calculation	Fortran	OpenMP, MPI	N/A	
QCD	Quantum-Chromo Dynamics simulation	Fortran	OpenMP, MPI	CUDA	B
N-Body	N-Body simulation using FDPS	C++	OpenMP, MPI	CUDA	
GROMACS	Molecular Dynamics simulation	C++	OpenMP, MPI	CUDA, HIP, SYCL	
SALMON	Ab-initio quantum-mechanical simulator for optics and nanoscience	Fortran	OpenMP, MPI	(OpenACC)	A



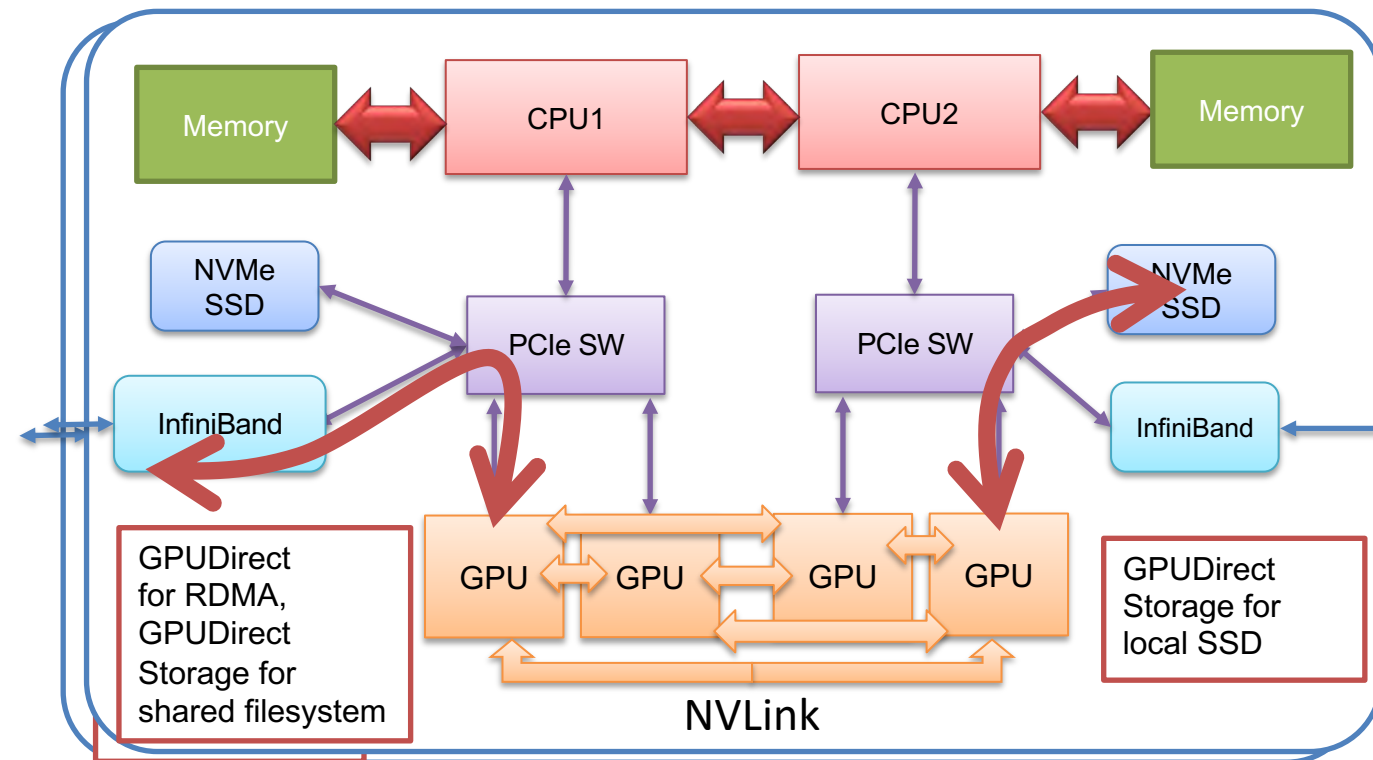
# GPU for OFP-II

- Officially selected NVIDIA's GPU for OFP-II (Jun. 2022)
    - Of course, GPU for Wisteria-Mercury must be same architecture
  - Reason for selection
    1. Performance of benchmarks
    2. Ease of porting efforts
    3. Support environment
    4. Fortran Support
- 
- Porting Activities
    - Basically, porting by users themselves
    - Special Support for big user groups and community-code providers: 16 groups
      - Kernel optimization, etc. by JCAHPC & NVIDIA
  - Support for Users
    - Mini-camp, Tutorial, Consultation
    - Portal, Materials, Videos
      - [https://jcahpc.github.io/gpu\\_porting/](https://jcahpc.github.io/gpu_porting/) (in Japanese)

Many thanks for support by JCAHPC and NVIDIA members !!

# GPU Configuration for OFP-II

- CPU-GPU combination options
  - CPU: Intel/AMD x86-64 based + GPU: H100 or successor connected by PCIe, available NVLink among GPUs
  - CPU: NVIDIA Grace (Arm-based) + GPU: H100 or successor connected by NVLink
    - Cache Coherent between CPU-GPU
- GPUDirect \* features
  - GPUDirect for RDMA: Direct communication between GPUs over nodes for GPU memory contents thru InfiniBand
  - GPUDirect Storage: Direct file IO between GPU memory contents and storage devices
    - Local SSD drives
    - Shared filesystem thru InfiniBand



Assumed GPU Node Configuration by x86-64 based

# Co-Design for Target Application



- **Study on Next Generation Accelerator and Its File IO**

- PI: Prof. Hanawa (U. Tokyo), vice PI: Prof. Tatebe (U. Tsukuba)
- Establish to easily handle data on the GPU by direct file IO or overlap between compute and file IO
- Realize efficient GPU-to-file IO processing in real applications: Astrophysics, City LES, Machine Learning

Create benchmark  
based on these projects

- **Practical Acceleration Methods to Achieve High Performance for Large-scale Applications**

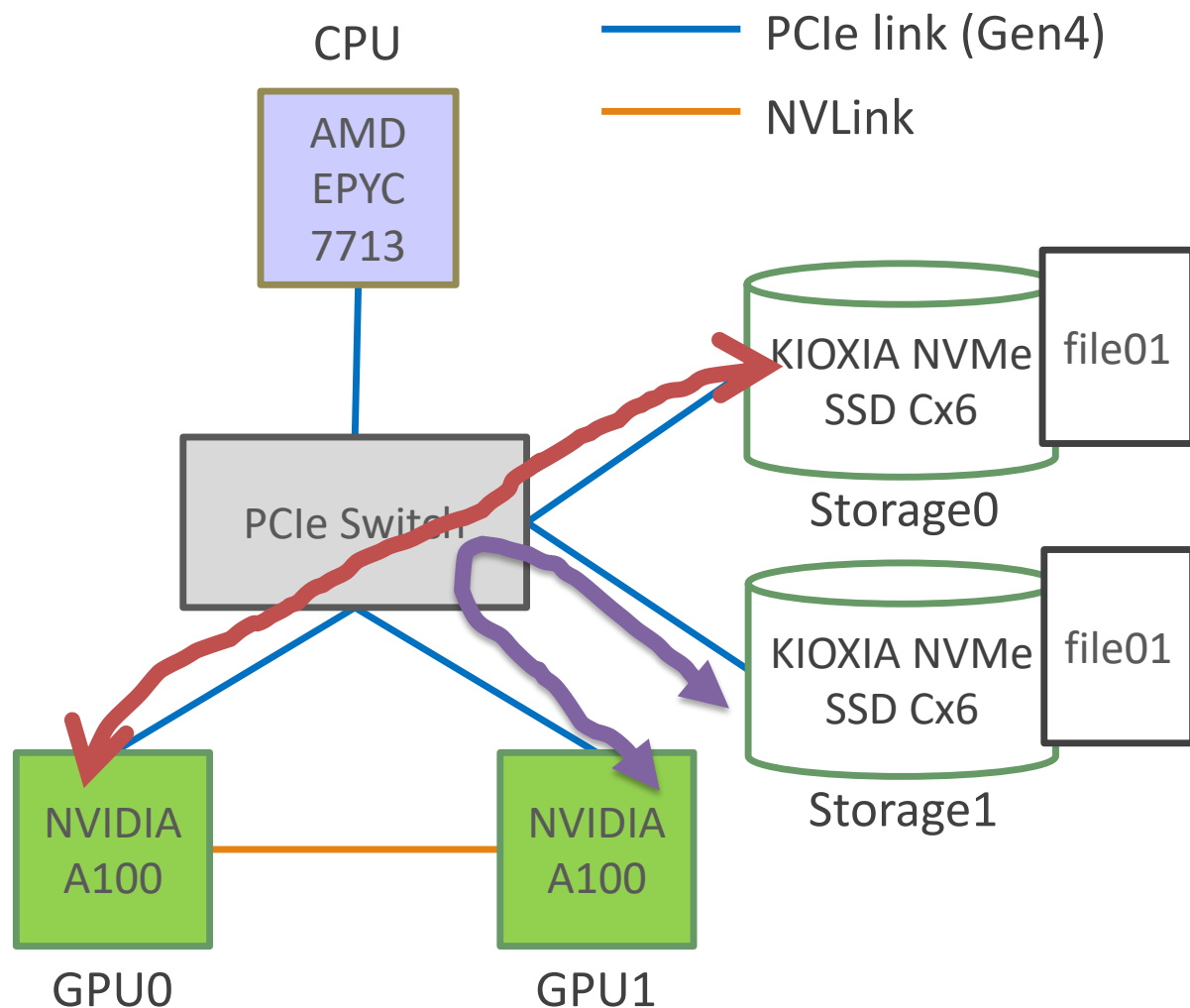
- PI: Prof. Shimokawabe (U. Tokyo), vice PI: Prof. Nukada (U. Tsukuba)
- Port CPU applications running on supercomputers to GPU-equipped supercomputers
- Establish methods for such porting strategy based on directive or standard parallelism supported by programming languages

Accepted as a JHPCN project [JHPCN: <https://jhpcn-kyoten.itc.u-tokyo.ac.jp/en/>]

Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures)

# Basic Performance Result(1) : Local SSD

## PCIe Topology & Measurement Methods



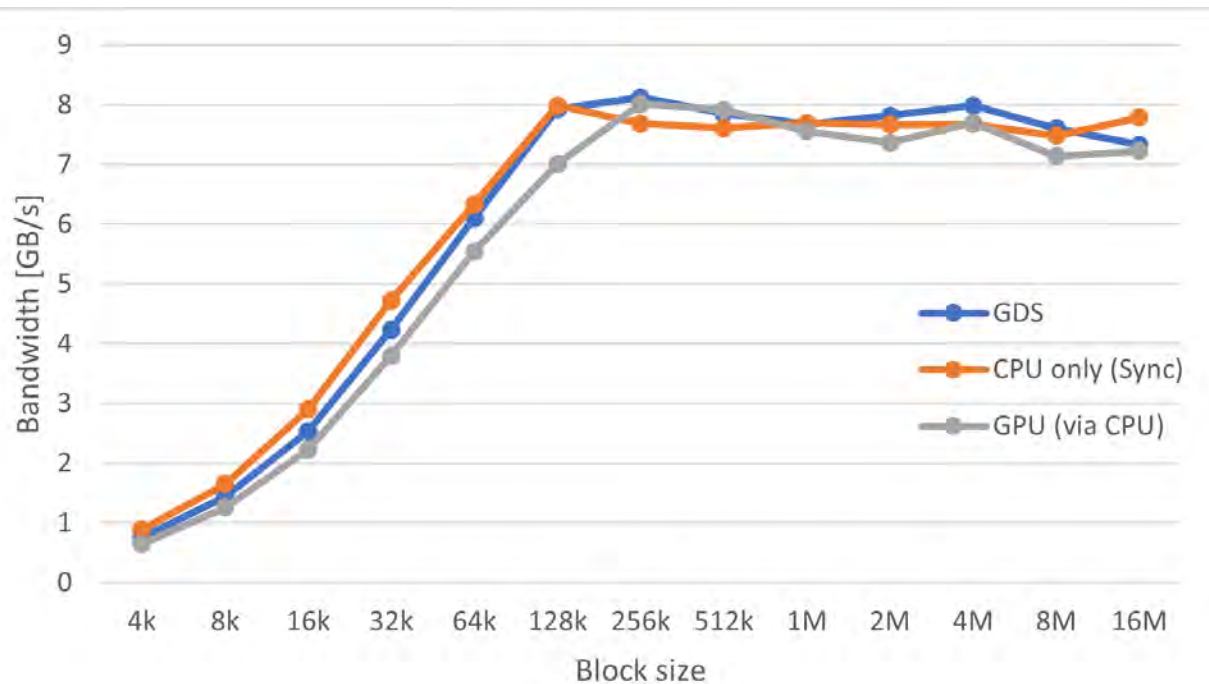
- **Data-transfer methods**

1. GPU ↔ Storage (GPUDirect Storage)
2. CPU ↔ Storage
3. GPU ↔ CPU ↔ Storage

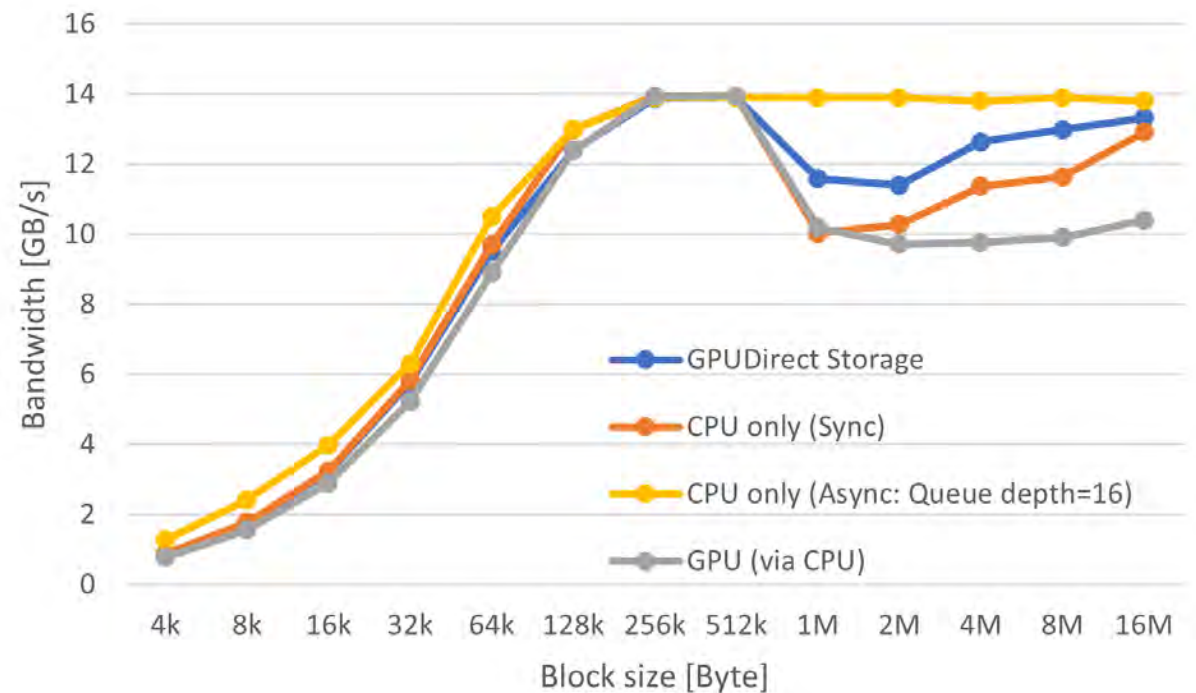
- **Settings**

- Filesize: 1GB
- Blocksize: Sequential 1MB / Random 4kB
- Runtime: 120s
- No. of threads: 2, 4, 8, ..., 256
- GPUs assigned to each NVMe
- Measured three times each

# Sequential Access



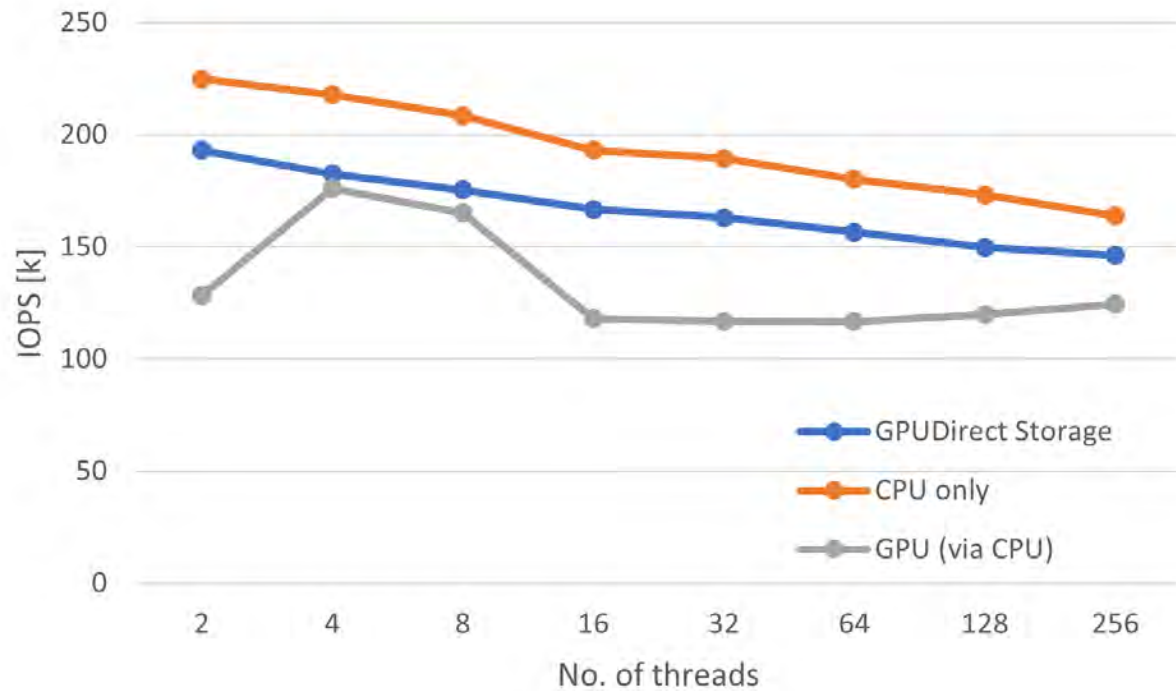
Sequential write (threads=8)



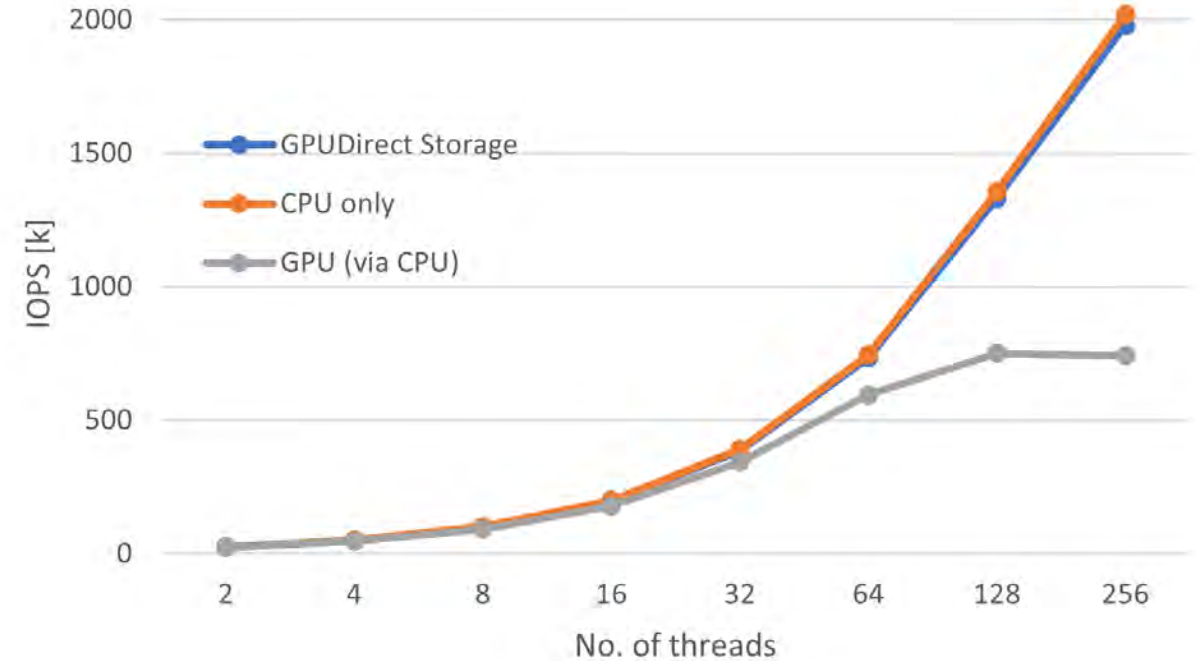
Sequential read (threads=16)

- Bandwidth is larger when:
  - block size is 256kB; the advantages of all transfer methods are not very different (sequential write)
  - block size is more than 256kB; async CPU has the highest bandwidth (sequential read)

# Random Access



Random write (block size=4kB)



Random read (block size=4kB)

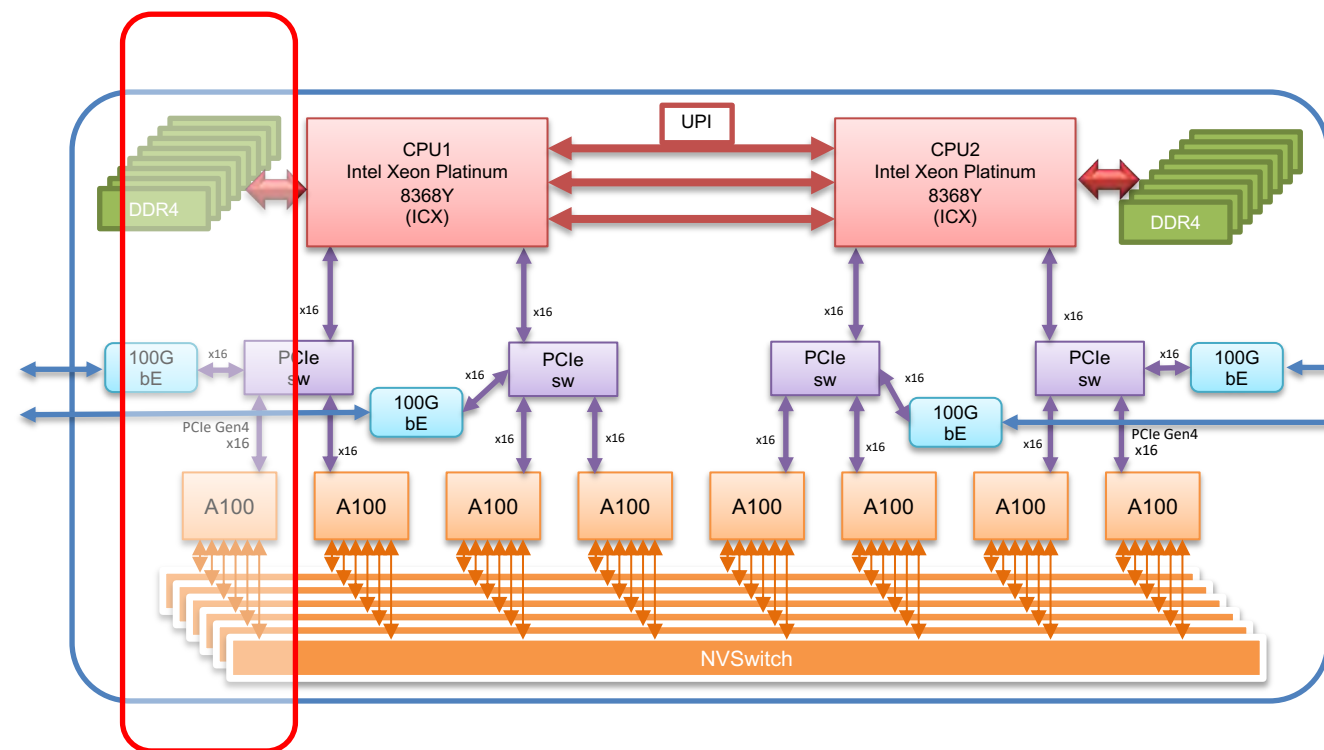
- IOPS is larger when:
  - no. of thread is smaller; CPU has the highest IOPS of the three transfer methods (random write)
  - no. of thread is bigger; CPU has the highest IOPS of the three transfer methods (random read)

# Basic Performance Result (2): VM to Lustre (NVMe SSD)

- mdx: a Platform for Building Data-Empowered Society
  - Cloud-like Supercomputer system virtualized by VMware ESXi
  - <https://mdx.jp/en/>
- VM (1GPU, 18 vCPUs) ⇔ Lustre FS with NVMe SSD
  - GPU: PCIe pass through, 100GbE: SR-IOV
  - 10GB, iosize 1MB, Striping in maximum, best from 3 trials

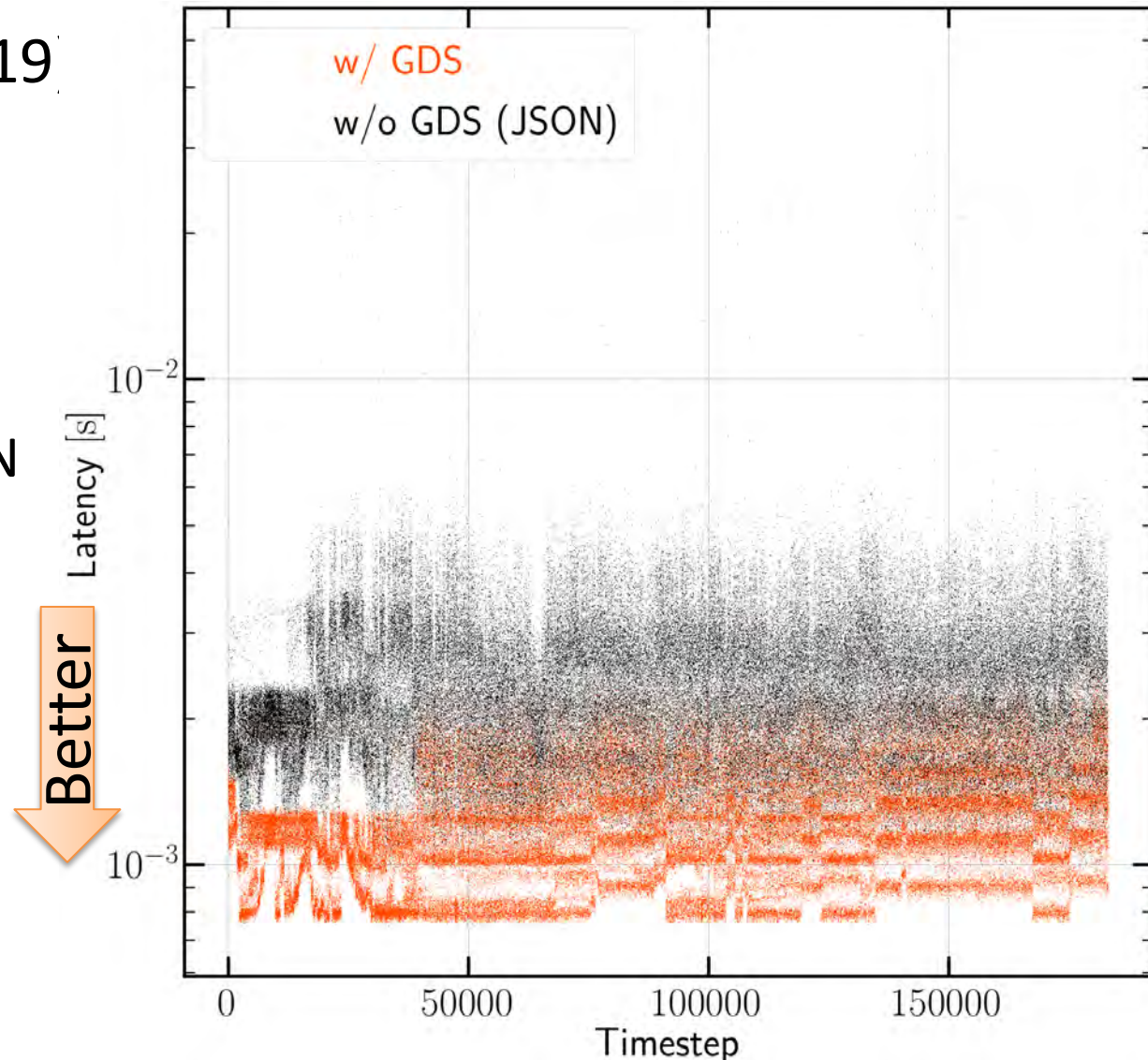
Need to consider optimal transfer method based on transfer size and number of threads

# of th	mode	Read (GiB/s)	Write (GiB/s)
4	nonGDS	2.57	2.93
	GDS	3.46	3.66
16	nonGDS	7.11	7.19
	GDS	8.88	8.83
32	nonGDS	2.05	8.75
	GDS	9.34	8.47
128	nonGDS	0.22	10.03
	GDS	0.24	8.80



# GDS Performance in Astrophysics

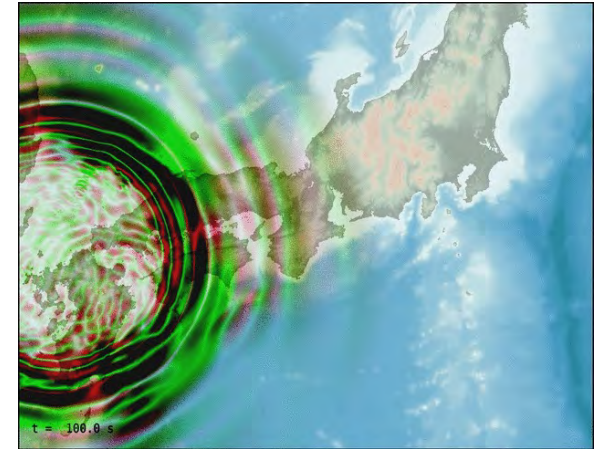
- GOTHIC (Miki & Umemura 2017, Miki 2019)
  - GPU-optimized tree code
- **GDS implementation:**  
**HDF5 via Virtual File Driver** (instead of directly using file API)
  - “cuFile” behavior can be controlled by JSON
    - Disable GDS: "gds\_rdma\_write\_support": false
- Galaxy collision between M31 and a past satellite galaxy
- **Very frequent output for fast-moving particles (~1k out of ~10M particles)**
- **GDS doubles the write performance**





# GPU Porting project

- **The research project [PI: Prof. Shimokawabe]**
  - The goal of this research project is to port CPU applications running on supercomputers to GPU-equipped supercomputers and to establish methods for such porting.
- **One of the target applications: OpenSWPC**
  - OpenSWPC: a simulation code for seismic wave propagation
  - Porting OpenSWPC, which is parallelized by OpenMP for CPU, to NVIDIA GPU using Fortran's standard parallelization syntax DO CONCURRENT.



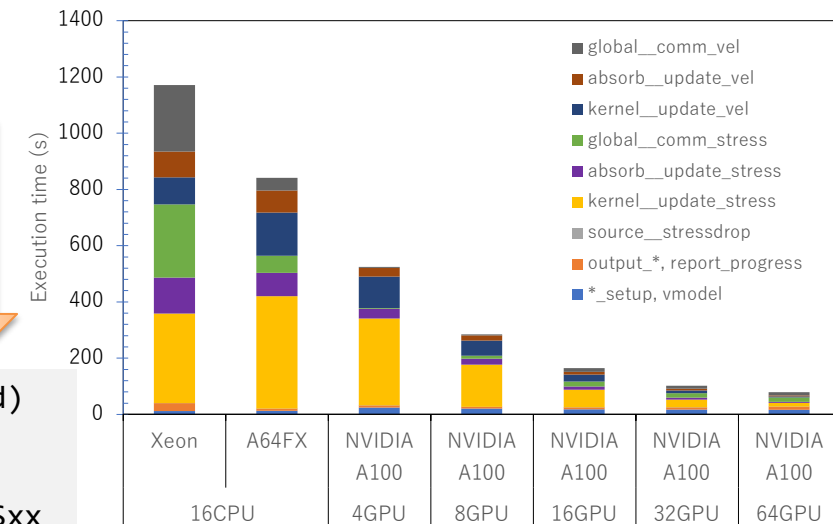
<https://github.com/tktmyd/OpenSWPC>

```
!$omp parallel private(j,k,dxSxx)
do j=jbeg, jend
  do k=kbeg, kend
    dxSxx = Sxx(k ,1,j) - Sxx(k ,0 ,j)
    Vx(k,0,j) = Vx(k,0,j) + gxe0(1) * dxSxx
  end do
end do
!$omp end parallel
```

2.2x faster 4x A100 vs 16x Xeon  
1.6x faster 4x A100 vs 16x A64FX  
3.2x faster in 16GPU vs 4GPU

Better

```
do concurrent (j=jbeg: jend, k=kbeg: kend)
  local(dxSxx)
  dxSxx = Sxx(k ,1,j) - Sxx(k ,0 ,j)
  Vx(k,0,j) = Vx(k,0,j) + gxe0(1) * dxSxx
end do
```



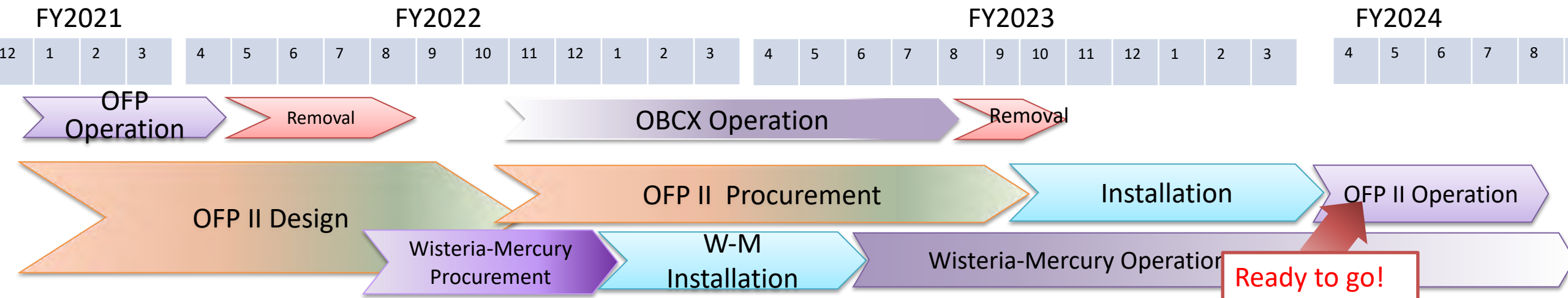
Performance of OpenSWPC on CPU/GPU

[ Seimi, Nukada (Univ. of Tsukuba), 2023]

# Summary (Oakforest-PACS II)

## Oakforest-PACS II (OFP-II)

- Design and Procurement for starting operation in Apr. 2024 (or later)
- **Target: 200+ PFLOPS, GPU nodes and CPU only nodes** (relatively small number)
- Designed and developed as a leading machine in the second class of HPCI, next to Fugaku
  - Advance co-design with applications
- **Promote GPU porting of existing applications using Wisteria-Mercury**



(Note) Arrows don't mean the exact period