# New Challenge for HPC and AI by Big Memory (Data) Supercomputer *Pegasus*


Taisuke Boku

Director, Center for Computational Sciences, University of Tsukuba

(slides courtesy by O. Tatebe, R. Kobayashi and A. Nukada)

# CCS at University of Tsukuba

- Center for Computational Sciences
- Established in 1992
  - 12 years as Center for Computational Physics
  - Reorganized as Center for Computational Sciences in 2004
- Daily collaborative researches with two kinds of faculty members (45 in total)
  - Computational Scientists
    who have NEEDS (applications)
  - Computer Scientists
    who have SEEDS (system & solution)
- One of national supercomputer centers under MEXT, but we are Research Center (others are service centers)

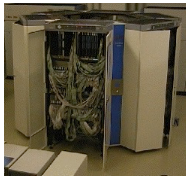
# History of PACS (PAX) series development at CCS

- 1977: research started by T. Hoshino and T. Kawai
- 1978: PACS-9 (with 9 nodes) completed
- 1996: CP-PACS, the first vendor-made supercomputer at CCS, ranked as #1 in TOP500

1978
1st gen: PACS-9

1980
2nd gen. PACS-32

1989
5th gen, QCDPAX



1996
6th gen: CP-PACS
Ranked #1 in TOP500

2006
7th gen: PACS-CS

2012~2013
8th gen: GPU cluster HA-PACS

2014
9th gen: COMA

2019
10th gen: Cygnus

| Year | Name | Performance |
|---|---|---|
| 1978 | PACS-9 | 7 KFLOPS |
| 1980 | PACS-32 | 500 KFLOPS |
| 1983 | PAX-128 | 4 MFLOPS |
| 1984 | PAX-32J | 3 MFLOPS |
| 1989 | QCDPAX | 14 GFLOPS |
| 1996 | CP-PACS | 614 GFLOPS |
| 2006 | PACS-CS | 14.3 TFLOPS |
| 2012~13 | HA-PACS | 1.166 PFLOPS |
| 2014 | COMA (PACS-IX) | 1.001 PFLOPS |
| 2019 | Cygnus (PACS-X) | 2.5 PFLOPS |

- ***co-design*** by computer scientists and computational scientists toward "practically high speed computer"
- Application-driven development
- Sustainable development experience
- Two streams of supercomputer operation
  - Our own unique strategy for advanced research → Cygnus, Pegasus
  - JCAHPC: widely spreading supercomputer resource service → OFP, OFP-II (planned)

# HPC technology contributing to AI: Computation

- Neural Network Processing
  - early '80s, Neural Network started to be studied for machine learning
  - supported by only poor processing power (CPU), no special hardware
  - just one middle layer, and results are not sufficient (low computation power)
    ⇨ not a Deep Learning (deep layered CNN-base machine learning)
- Accelerators, especially GPU
  - after GPU became attractive for numerical computing, GPU started to be introduced for NN
  - very regular and large capacity of computing ⇨ good for SIMD implementation
  - GPU is now the main player of AI (DL) and GPU vendors continue to build AI-oriented GPU as well as HPC use: NVIDIA and AMD (followed by Intel too)
- CPU instruction set
  - to support ML, FP16 (16-bit half precision) and BFloat16 (long mantissa) are introduced
    ⇨ recently with FP8
  - SIMD vector instruction is good to support them in up to 512bit

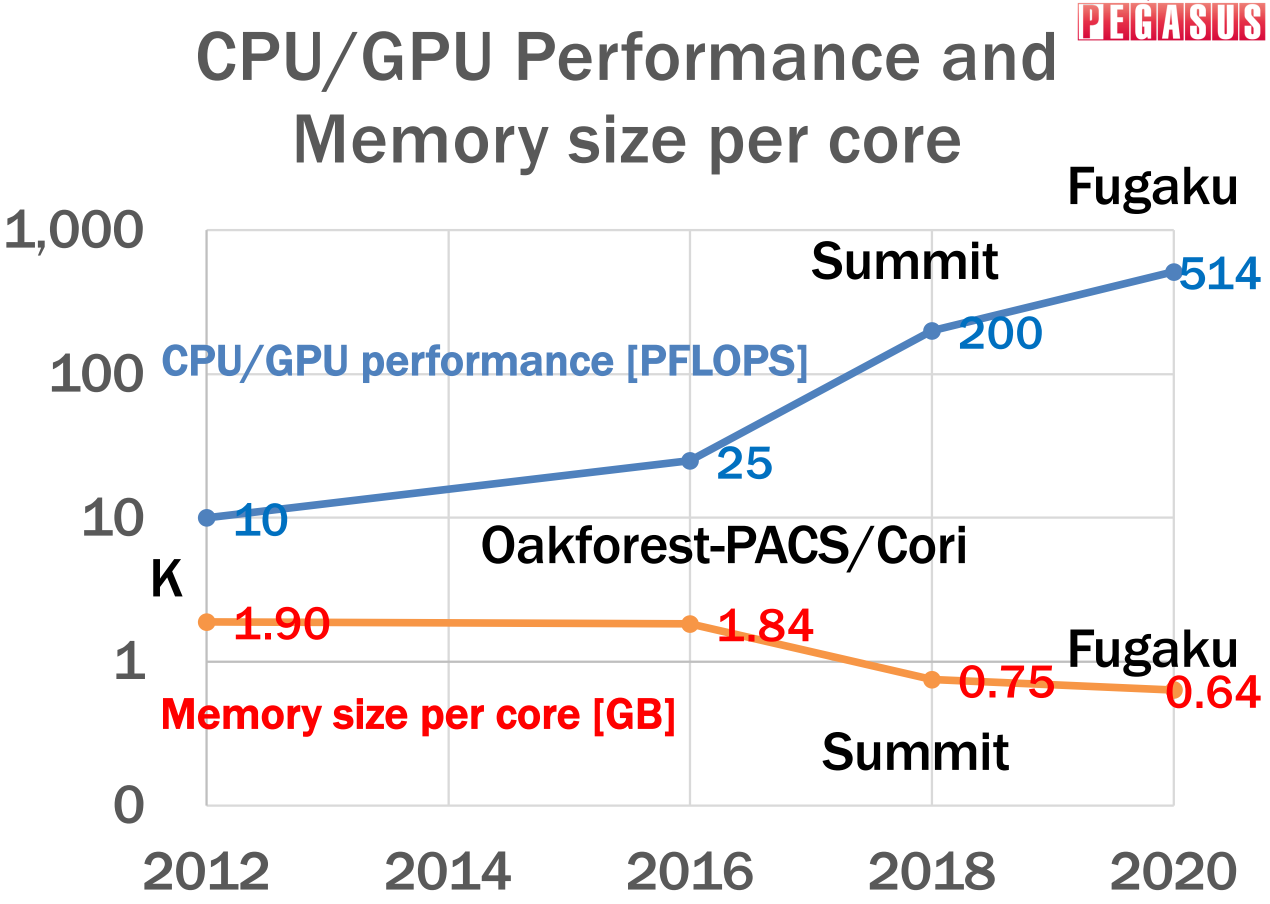
# AI contribution to HPC: efficient data analysis and reduction

- efficient parameter space search
  - reducing the parameter search space by machine learning
    - climate simulation
    - astrophysics
    - life science
  - data matching on large data space
    - text base docking for creation of medicine
- efficient data analysis
  - finding the characteristics of phenomena
  - machine learning base data sorting and collection
- surrogate modeling

**HPC for AI ⇔ AI for HPC** ⇨ GPU is important (of course!) but not enough

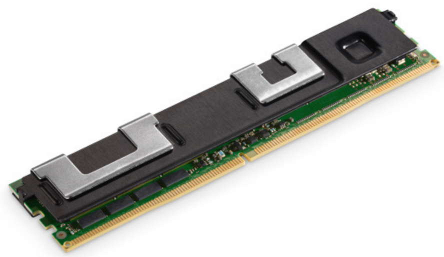
# Why we need Big Memory

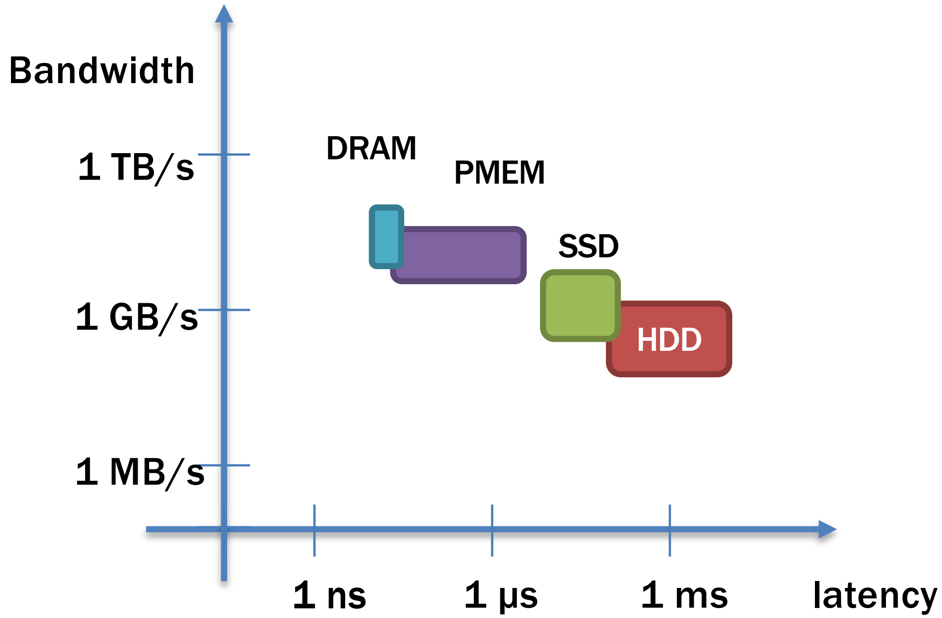
- CPU performance 50x, but memory size 3.8x in 8 years
- It matters for Data-driven and AI-driven Science
  - Memory size and Storage performance are really important
- Introduce Persistent Memory (PMEM)
  - Memory mode for memory size and direct mode for storage performance



# Persistent Memory

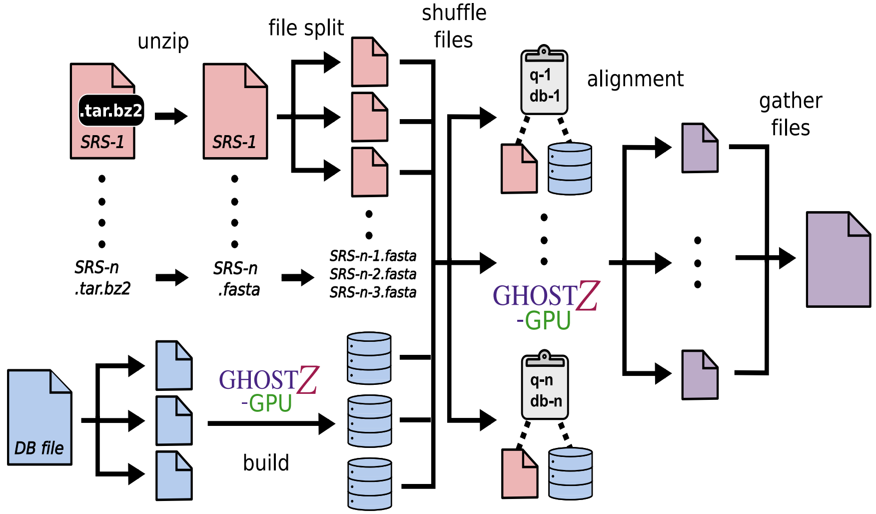
- One order better price/capacity
- Minimum latency is ~60 ns (similar to DRAM)
- ~Half of bandwidth
- Memory mode
  - Larger memory space without much performance penalty
  - Possible to use DRAM as last level cache
- App direct mode
  - Direct access to byte-addressable persistent memory and high-performance storage

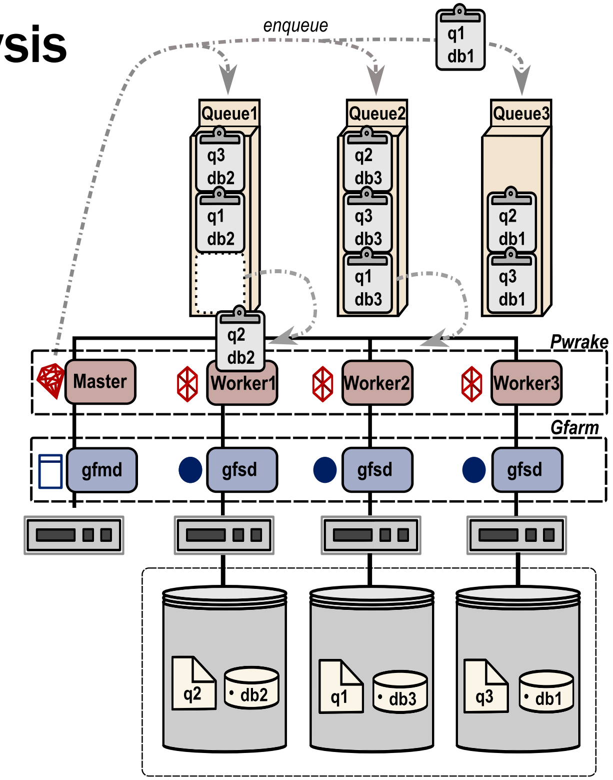




# Large dataset science (1): Genomic data accumulation and analysis

- environmental genome analysis ⇒ specific genome analysis
- large dataset ⇒ distributed processing with query
- large data capacity on computation node enhances performance



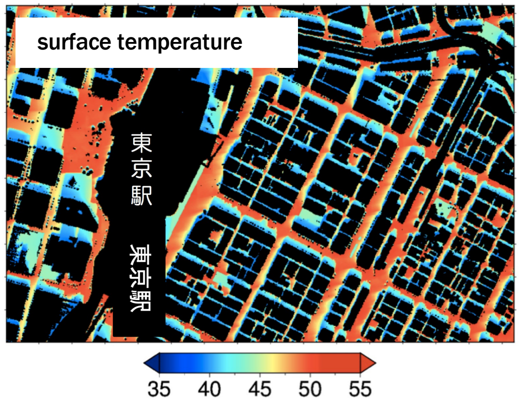



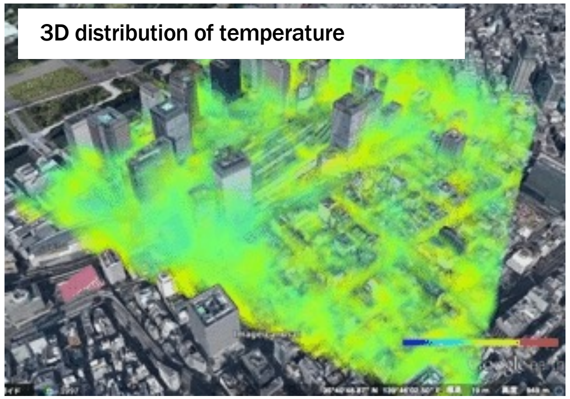
# Large dataset science (2): Local climate simulation

## Ultra high resolution of local climate simulation by multi-physics

- City-LES: urban climate simulation by LES, solar effect, building structure, surface material

**TOKYO2020 model around Tokyo Station, 1m grid**



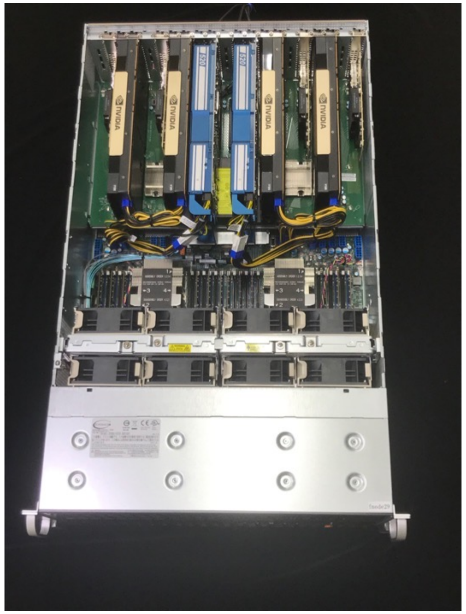




- large scale in-transit analysis on large capacity date
- completely GPUized (up to 15x performance of CPU)

# How to use GPU + Big Memory (PMEM) for HPC/BD/AI ?

- GPU
  - totally important for both HPC and AI
  - powerful GPU contributes HPC ↔ AI coupling
- Ultra large capacity of medium~high speed memory
  - astrophysics, climate, life science: requiring large capacity of working set with relatively low FLOPS requirement
- In-Situ processing
  - bypassing the slow in/out of large capacity of file data
    ⇨ PMEM technology support
- High-speed distributed storage
  - even as a distributed storage, much faster solution is possible (shown later)

# Cygnus (PACS-X): Extreme Computing with multi-hybrid accelerators

System integration by NEC

**The world first practical supercomputer with Multi-Hybrid (GPU + FPGA) Accelerating Architecture: 320 GPUs (V100) + 64 FPGAs (Stratix10) in 80 nodes**

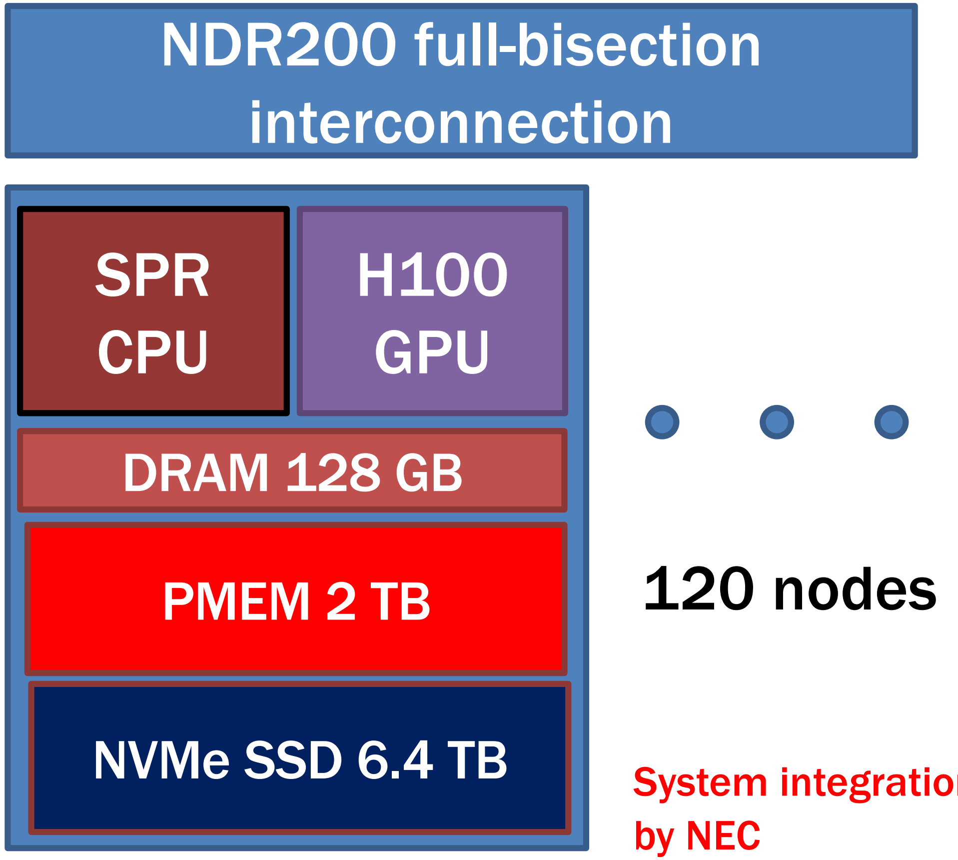
# Pegasus (PACS-XI) : Big Memory Supercomputer

- Strategy of current Cygnus (GPU+FPGA)
  - accelerating traditional HPC, especially for multi-physical simulation with multiple phenomena, by coupling of GPU + FPGA
  - GPU: SIMD-type of spatial parallelism
    FPGA: pipelining x spatial parallelism
  - GPU and FPGA compensate with each other to fill the gap for various parallelism in various applications or even in an application
  - AI (deep learning) is mainly done on GPU, and FPGA partially support (ex. sorting on database search)
- New concept of Big Memory Supercomputer
  - much larger simulation on HPC applications: astrophysics, climate, bioscience
  - much faster distributed file system with large scale cluster computing

# Pegasus: world first combination of H100+SPR+PMEM+NDR200

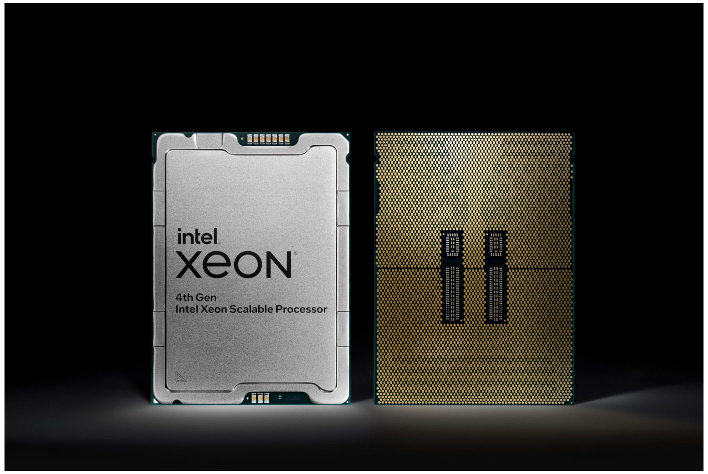
- total Performance
  - 120 nodes, 6.5 PFlops, 240 TiB (PMEM)
- node components
  - NVIDIA H100 PCIe (gen5)
  - Intel Sapphire Rapids (SPR)
  - Intel Optane ver3
- Interconnection Network
  - NDR200: NVIDIA Quantum-2 IB with 200Gbops full bisection b/w
- parallel file system (DDN)
  - 7.1 PByte, 40 GB/s

**HPL: 3.47 PF (54%) (as on Feb. 2023)**
**MPI pingpoing: 23.9GB/s (95.7%)**



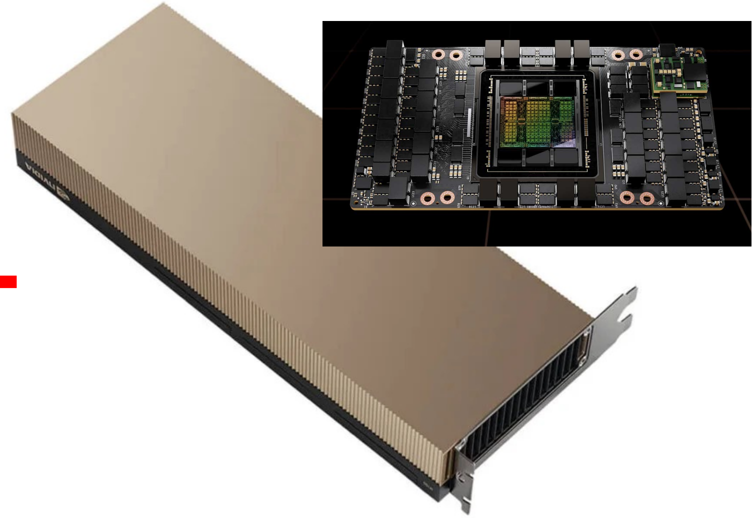

System integration by NEC

# Combination of SPR + Optane300 + H100




48 core, 2.1GHz

256GB PMEM + 16GB DRAM
x 8 modules



26 TFLOPS (FP64)
51 TFLOPS (FP64-Tensor)
80 GB HBM3 (2TB/s)

- all the brand-new parts combination (first in the world)
- large capacity memory + high speed CPU and GPU for HPC
- large capacity memory + ultra high speed tensor calculation for AI
- MemVerge software supports straight-forward extension of current app. to Optane

# Pegasus outlook

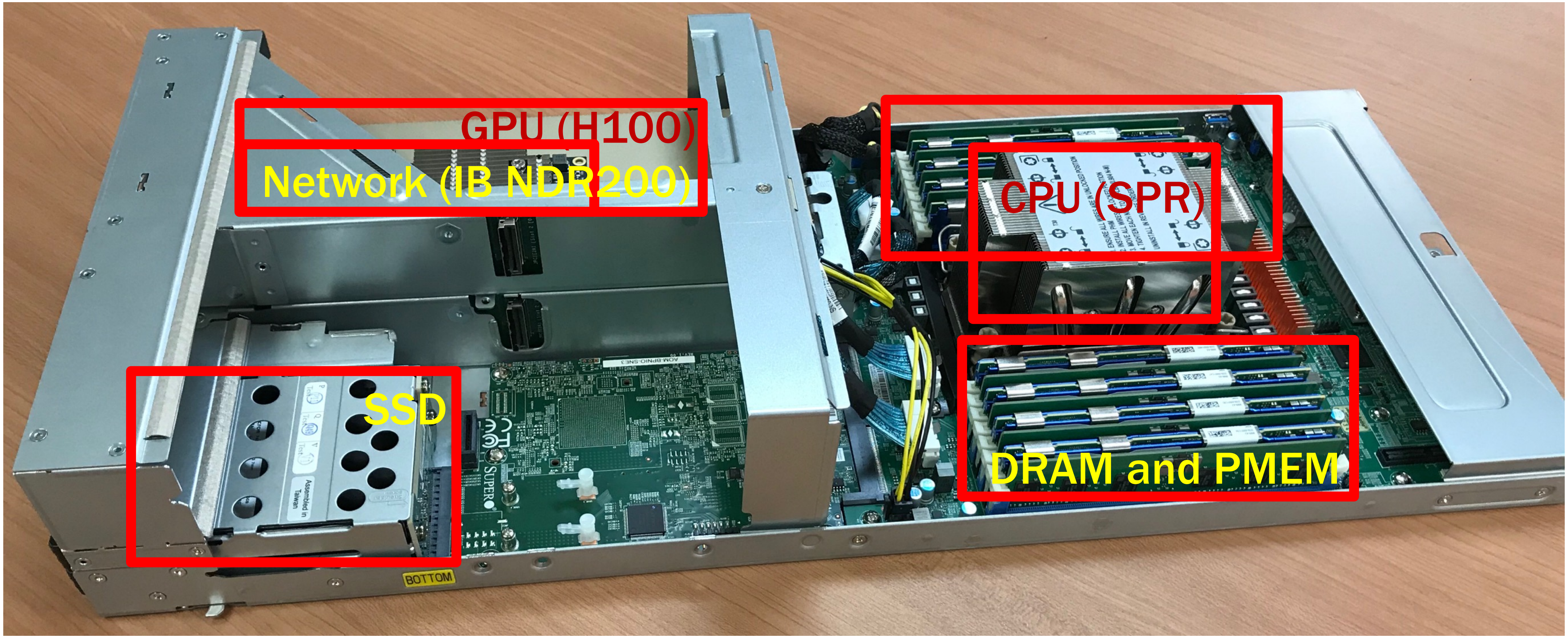

Pegasus at CCS, Univ. of Tsukuba (2022/12~)

file server

computation nodes (120)

IB switch
+login server

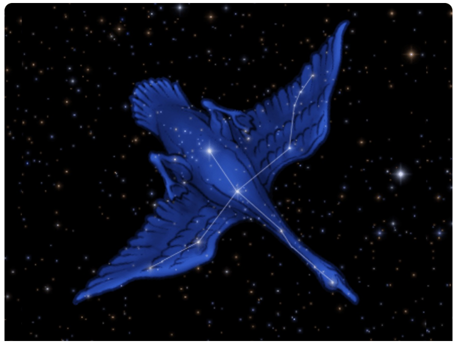
# Computation node



# Comparison of Cygnus and Pegasus

| | Cygnus (2019) | Pegasus (2022) |
|---|---|---|
| #nodes | 81 (C: 162, G: 324) | 120 (C: 120, G: 120) |
| PFLOPS (DP) | 2.3 | 6.5 (2.8x) |
| CPU | 0.16 | 0.5 (3.1x) |
| GPU | 2.18 | 6.0 (2.7x) |
| FPGA (SP) | 0.64 | 0 |
| DRAM (TiB) | 10.2 | 30.7 (3.0x) |
| PMEM (TiB) | 0 | 240 |
| Storage (PB) | 2.4 | 7.1 (3x) |

# Cygnus and Pegasus



Cygnus

Pegasus


blue Cygnus and red Pegasus
are the sister systems

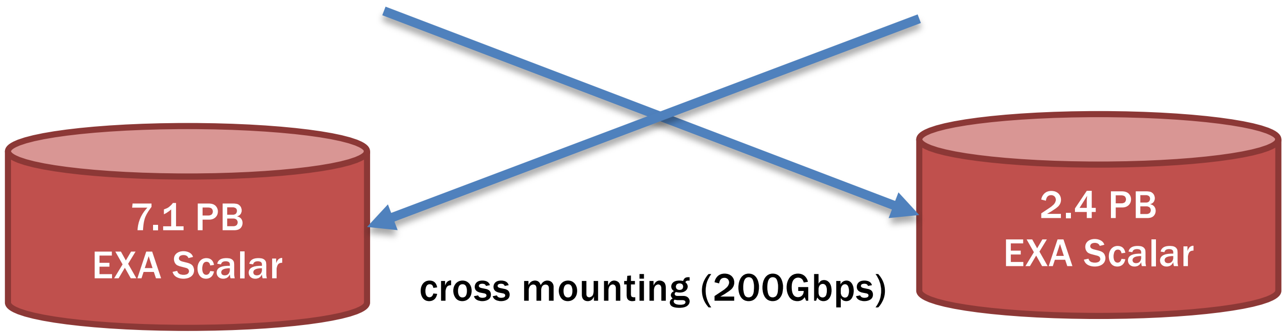
# Pegasus & Cygnus : twin system for Extreme Computing and Big Data

**Pegasus**

6.5 PFlops
GPU
Pmem

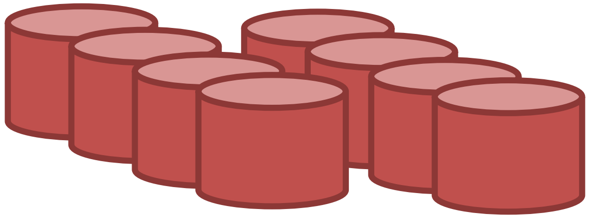
**Cygnus**

2.3 PFlops
GPU
FPGA



# Research of ad hoc parallel file system

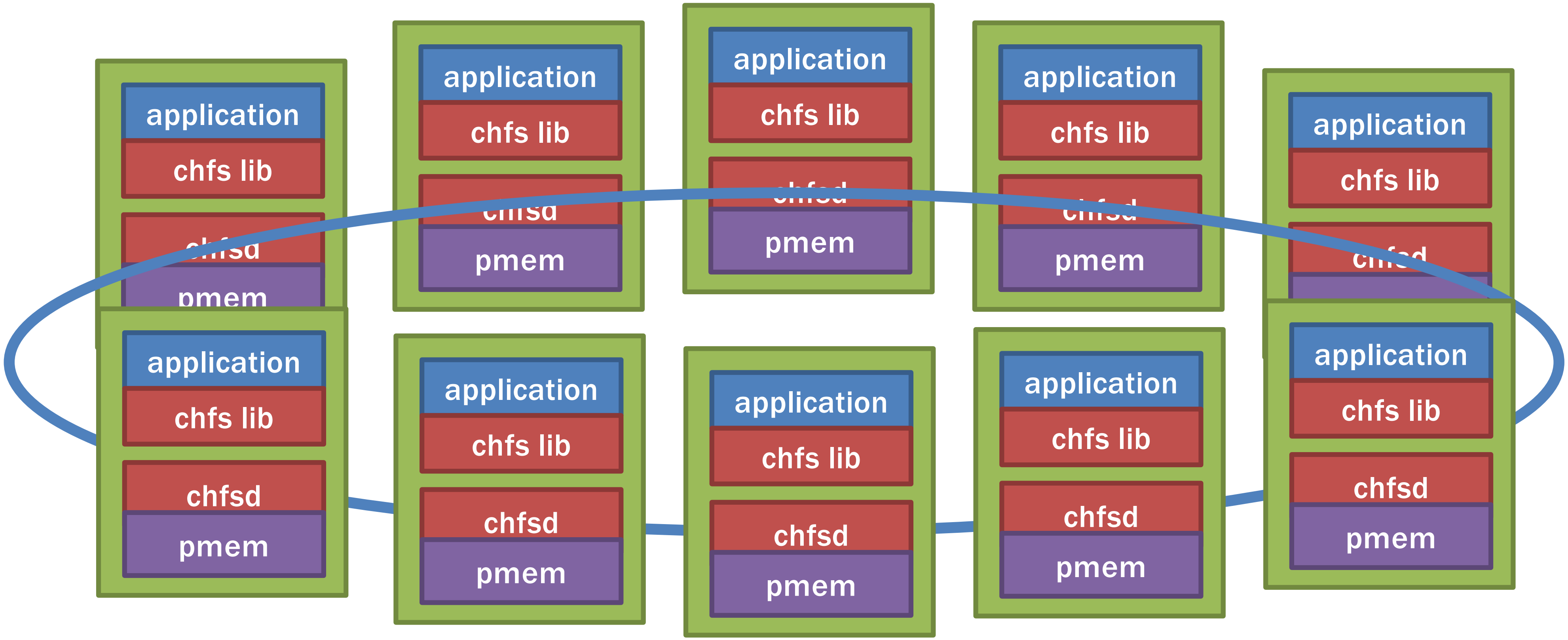
- Temporal parallel file system using node-local storage
- Fill the performance gap between CPU/GPU and storage



- We are developing CHFS (Consistent Hash File System) ad hoc file system to utilize persistent memory
  - No metadata server, no sequential processing for performance and scalability

* O. Tatebe, et. al., **"CHFS: Parallel Consistent Hashing File System for Node-local Persistent Memory", HPC Asia 2022**

# System Architecture of CHFS




Compute nodes

# GPU performance (preliminary): V100/Cygnus vs H100/Pegasus

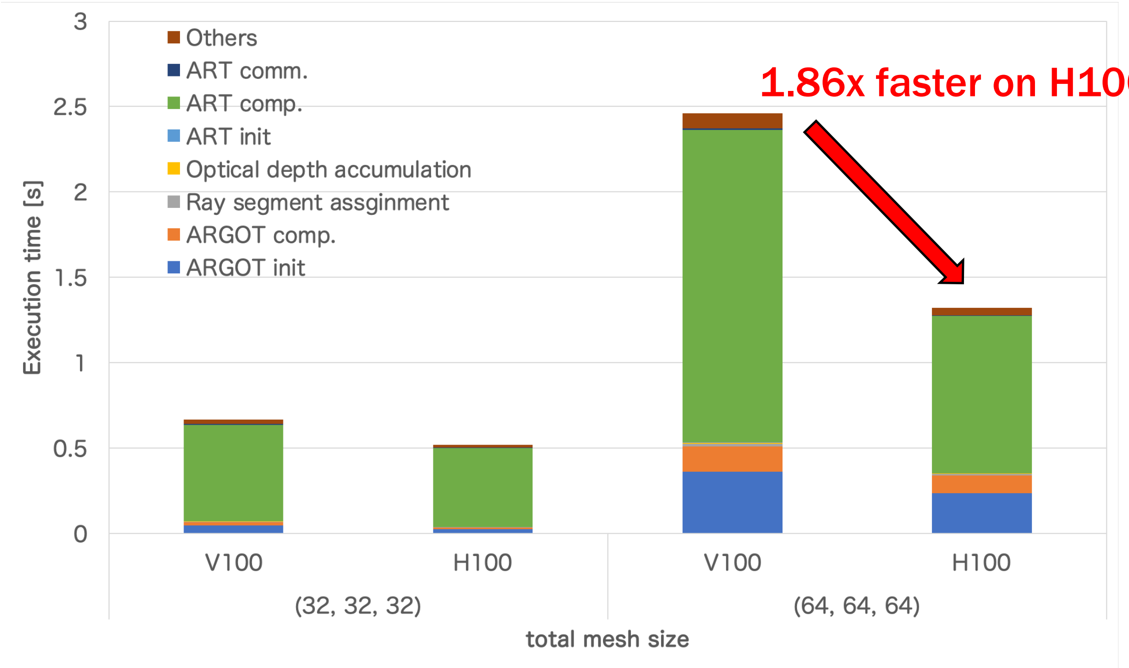
(by R. Kobayashi@CCS)

- **ARGOT: Astrophysical fundamental simulation code for early stage universe to analyze the born of first stars and galaxies**


**ARGOT method: point source radiation**

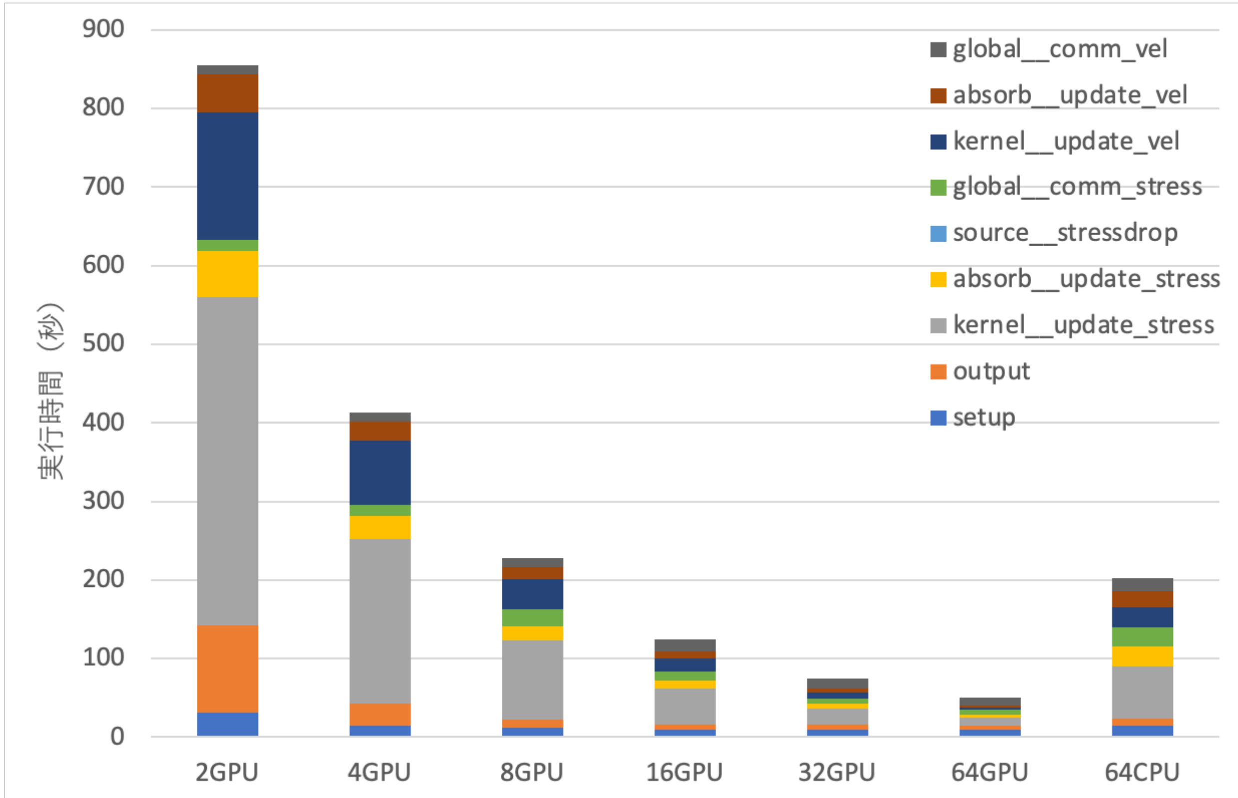
**ART method: photon diffusion radiation**

**heavy computation on GPU**



# DO CONCURRENT preliminary evaluation (by A. Nukada@CCS)

- OpenSWPC: seismic simulation of earth quake wave propagation




(original code by T. Furumura et.al.
= Fortran + OpenMP + MPI)
GPU/parallelization by
CUDA Fortran + OpenACC + MPI

Pegasus
CPU: Intel SPR 48core 2.1GHz
GPU: H100 PCIe
IB: NDR200 (Quantum-2 IB)

# Utilization of PMEM

- Large memory
  - MemVerge – using PMEM (2TB) in behind of DDR (128GB) to use DDR memory as "4th cache" to make balance between capacity and speed
  - just declaration on job script, no need for reprogramming
- High speed I/O
  - mounted as in the same manner with local SSD
  - will be used under CHFS ad hoc distributed file system
- Mixed
  - it is possible to provide some fraction for memory and other for storage
  - storage solution is "ad hoc" – life time is limited to the same job

# Summary

- **AI** is the latest important application and gets growing rapidly as a **practical application in human life**
- **HPC technologies** have been **contributing to AI (HPC for AI)** so far, and now it's time to use **AI technologies for efficient HPC solutions (AI for HPC)**
- Gap between **computation performance and memory capacity** is so serious
- Utilizing **Persisten Memory (PMEM)** both for large capacity memory and high performance shared file system (ad hoc) simultaneously, including efficient in-situ processing
- **GPU** continues to play an important role both for **HPC and AI**
  ⇨ **high performance GPU system with large capacity memory is ideal**
- Variation of coverage is so wide to support large scale data science and simulation, based on **HPC/BD/AI** and **GPU/PMEM** combination

⇨ ***Pegasus*** ***(started official operation from April 1st 2023)***