Next Supercomputer at CCS: Pegasus – Big memory supercomputer

Background

- Data-driven and AI-driven Science requires large memory size and storage performance, but memory capacity per CPU core decreases
- Introduces Persistent Memory in compute nodes to accelerate large-scale data analysis and big data for better cost performance, power consumption, and application performance
- Fosters new fields of large-scale data analysis, new applications of big data AI, and system software research

Pegasus Highlights

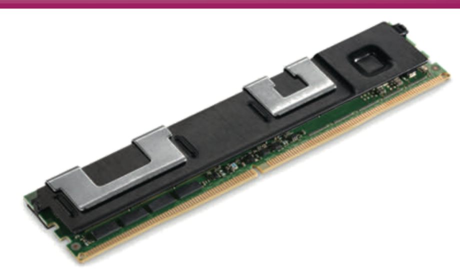
- **Intel Xeon Sapphire Rapids, NVIDIA H100** Tensor Core GPU with PCIe and **51TFlops** of extreme performance, and **2 TiB of persistent memory** strongly drive Big Data and AI
- The **world's first system** with **NVIDIA H100 PCIe GPUs connected via PCIe Gen5**
- First system announced in Japan that will utilize **NVIDIA Quantum-2** InfiniBand networking

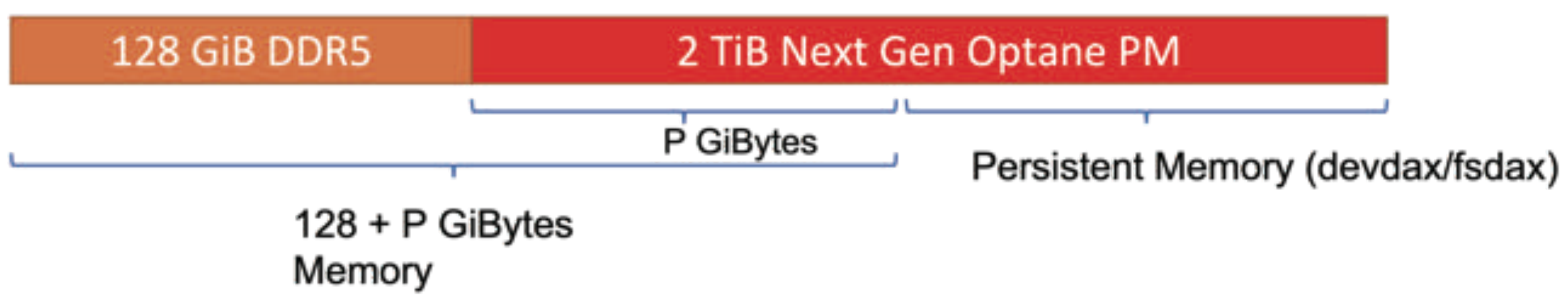
System name	Pegasus
Total performance	6.1 PFlops
Total memory size	255 TiByte (15 TiByte DDR5 + 240 TiByte Persistent Memory)
Number of nodes	120
Interconnects	Full bisection fat-tree network interconnected by the NVIDIA Quantum-2 InfiniBand platform
Parallel file system	7.1PB DDN EXAScaler (40 GB/s)

PEGASUS

Persistent Memory

- Persistent memory is configured as **Filesystem DAX** (fsdax) by default
  - DAX (Direct Access) mounted XFS is available at **/pmem**
- Pmem mode can be configured by the environment variables
  - MV_SIZE – use Pmem for **extended memory** and specify the **memory size** in GiB or ALL
  - USE_DEVDAX – use **Device DAX** (devdax) instead of fsdax





Specification

- Compute nodes (NEC LX 102Bk-6) x 120

CPU	Intel Xeon
GPU	NVIDIA H100 Tensor Core GPU with PCIe (51 TFlops in FP64 Tensor Core)
Memory	128GiB DDR5 (282 GB/s)
Persistent memory	2TiB Intel Optane persistent memory 300 series
SSD	2 x 3.2TB NVMe SSD (7 GB/s)
Networking	NVIDIA Quantum-2 InfiniBand platform (200 Gbps)

NEC LX B1000E Blade Enclosure

NEC LX 102Bk-6

NEC LX 124Rk-2

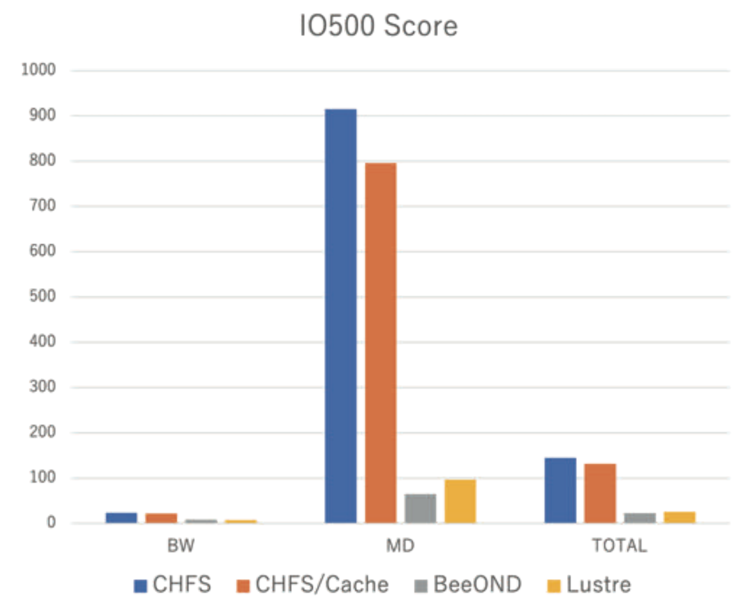
- Login nodes (NEC LX 124Rk-2) x 3
  - 2 x Intel Xeon (Sapphire Rapids)
  - 256 GiB DDR5 Memory, NVMe SSD, InfiniBand x 2, 100GbE
- Parallel File System (DDN ES200NV/ES7990X/SS9012)
  - DDN EXAScaler (Lustre)
  - MDS/MDT (4.2 billion inodes)
    - Active/Standby MDS
    - 1.92 TB NVMe SSD x 11 (8D + 2P + 1HS)
    - InfiniBand HDR100 x 4
  - OSS/OST (7.1 PF available)
    - 4 x Active/Active OSSs
    - 18 TB 7,200rpm NL-SAS x 534 ((33drives x 8pools + 3HS) x 2)
    - 8D + 2P Declustered RAID
    - InfiniBand HDR100 x 8

Software Component

- Ubuntu
- Intel oneAPI (C++/C/Fortran, oneMKL, MPI, VTune, Trace Analyzer&Collector)
- NVIDIA HPC SDK (C++/C/Fortran/Cuda, cuBLAS, cuTENSOR, cuFFT, …, Open MPI, NVSHMEM, NCCL, Profilers, Debugger)
- Open Source SDK (GNU Compilers, Python, PMDK, Open MPI)
- Tensorflow, Keras, PyTorch, …
- JypyterHub, TensorBoard, Nextcloud, Gfarm

CHFS/Cache – Caching File System for Node-Local Persistent Memory / Storage

- Design with CHFS without degradation of metadata and bandwidth performance
  - Relaxed reasonably consistency with PFS
  - User's assumption to FS does not change
- Demonstrates high bandwidth, metadata performance, latency, and scalability that are nearly identical to CHFS without caching



Osamu Tatebe, Hiroki Ohtsuji, "Caching Support for CHFS Node-local Persistent Memory File System", Proceedings of 3rd Workshop on Extreme-Scale Storage and Analysis (ESSA 2022), pp.1103-1110, 10.1109/IPDPSW55747.2022.00182, 2022