Cygnus-BD for data-driven and AI-driven Science

Osamu Tatebe

Center for Computational Sciences, University of Tsukuba

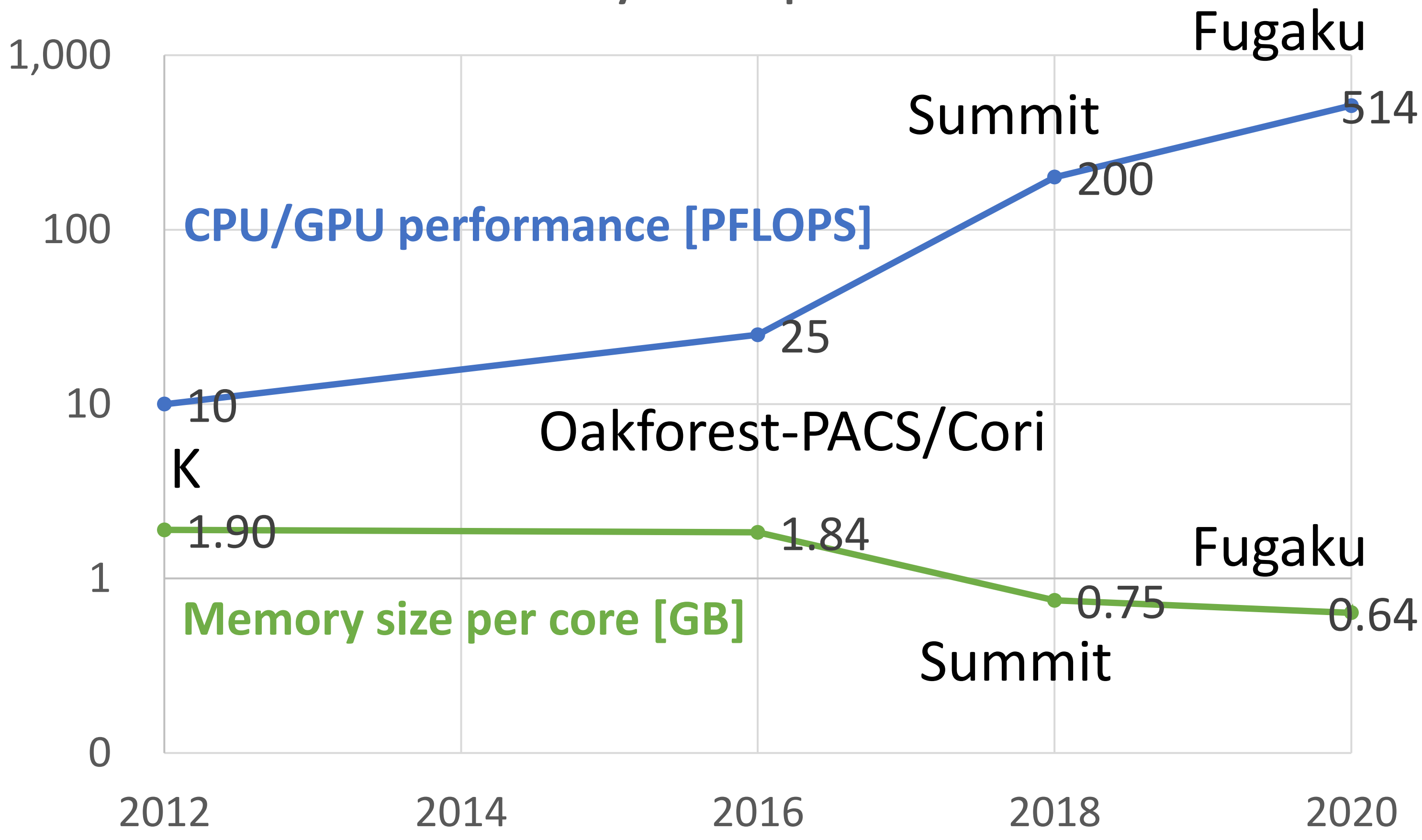
MCRP 2022 (Next Year)

- Oakforest-PACS will be shutdown in March, 2022
  - JCAHPC Seminar in May, 2022
- 15% of compute time of Wisteria-O (25.9 PFlops) operated by U Tokyo will be provided for MCRP 2022
  - Mini Fugaku (7,680 nodes)
  - 2.2GHz 48c A64FX, 32GB mem, 1 TiB/s mem BW
  - 6D Tofu-D interconnect
- Cygnus provided as well as this year

Cygnus-BD background

- 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
  - Memory mode for memory size and direct mode for storage performance

CPU/GPU Performance and Memory size per core

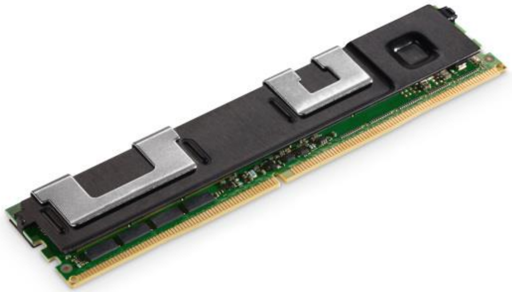


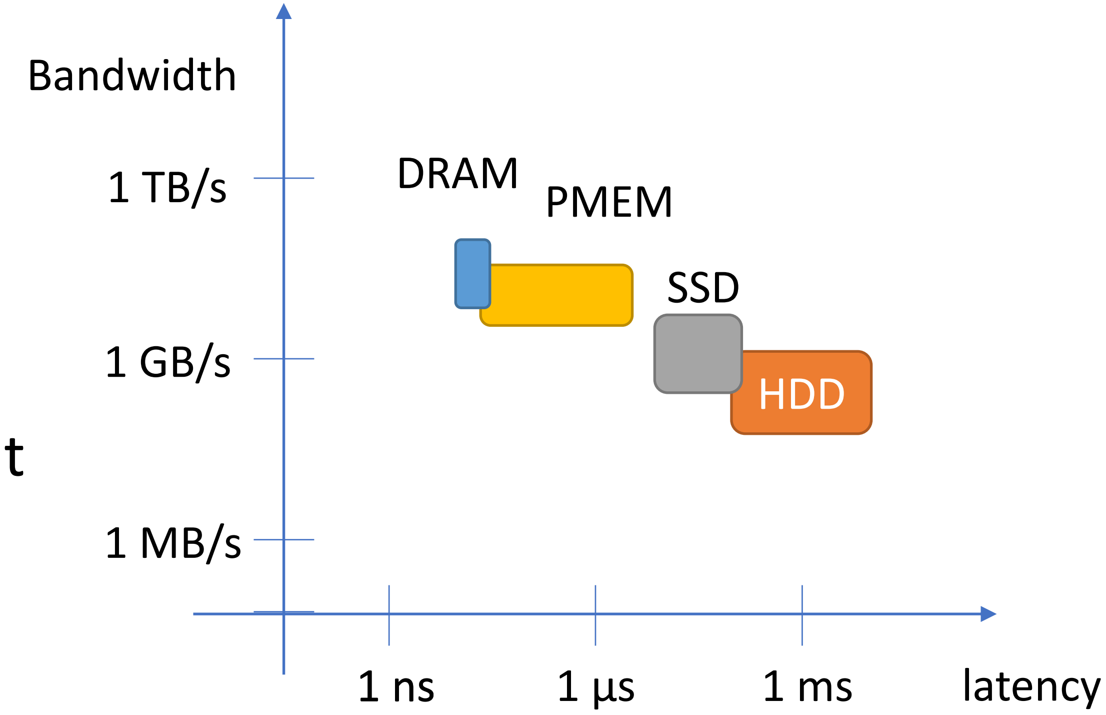
Design Goal of Cygnus-BD

- Accelerates large-scale data analysis and big data AI by utilizing persistent memory for large memory space and high performance storage
- Fosters new fields of large-scale data analysis, new applications of big data AI, and system software research

Persistent Memory

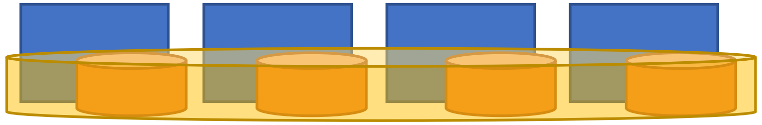
- One order better cost performance
- Minimum latency is ~60 ns (similar to DRAM)
- Half of bandwidth
- Memory mode
  - Larger memory space without much performance penalty
- App direct mode
  - Direct access to byte-addressable persistent memory and high-performance storage

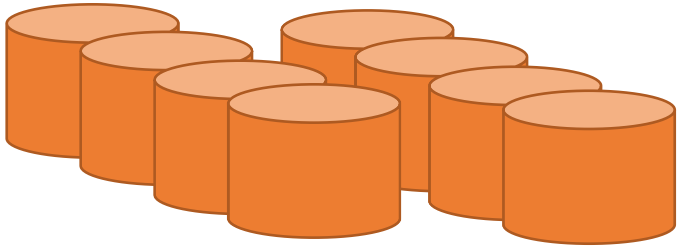




Research of Ad hoc parallel file system

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





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

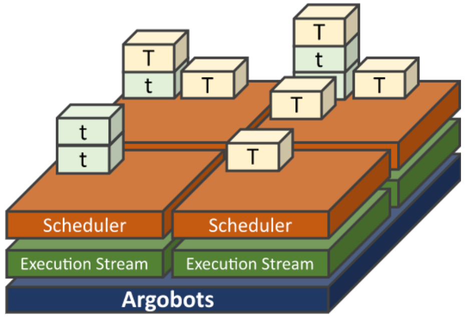
Design goal of CHFS

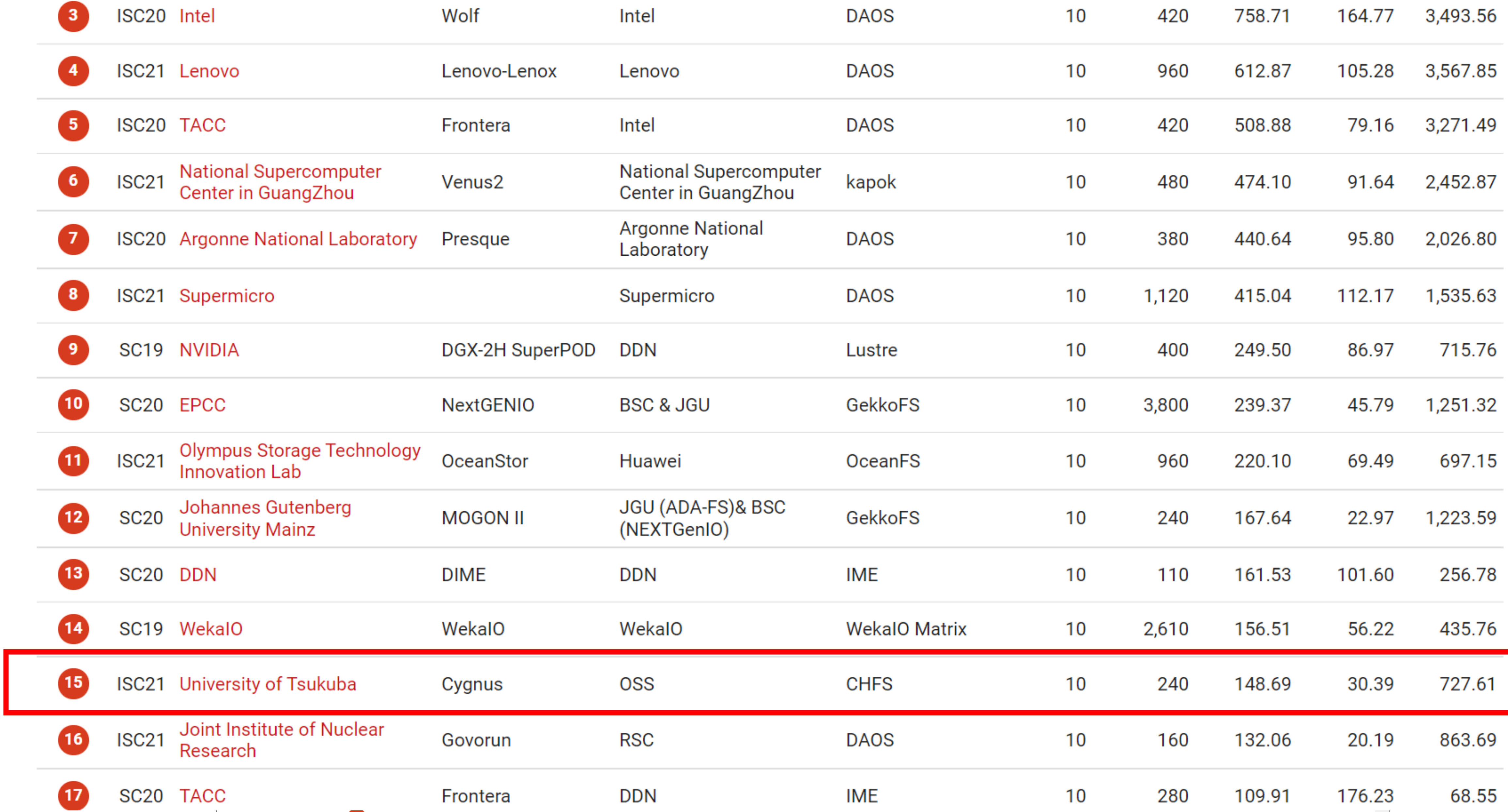
- Utilize persistent memory performance
  - In-memory persistent key-value store (not block-based file system)
- Reduce metadata overhead and achieve scalable performance improvement
  - No dedicated metadata server
  - No sequential execution
  - Based on highly parallel distributed key-value store without any central data structure
- Improve single-shared-file performance
  - File is divided into fixed-size chunks to distribute a single file among servers

Implementation of CHFS

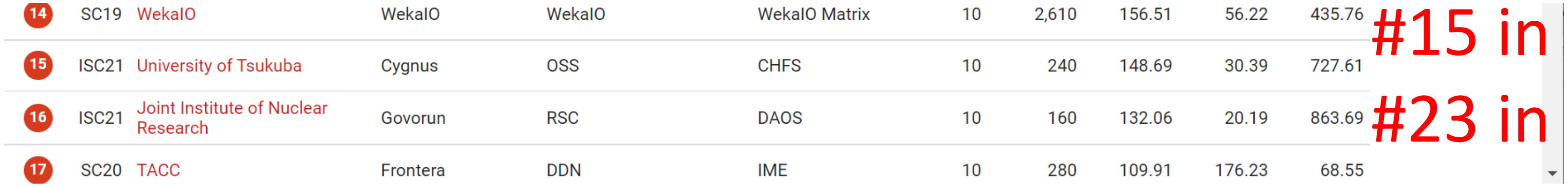
- Mochi-Margo [JCST 2020]
  - https://mochi.readthedocs.io/en/latest/
  - Communication library using Mercury and Argobots
- Mercury [Cluster 2013]
  - Async RPC, RDMA communication library
  - libfabric, CCI, shared memory plug-ins
- Argobots [IEEE TPDS 2018]
  - Light-weight thread library
- pmemkv
  - cmap – concurrent hash map

Mochi





3	ISC20	Intel	Wolf	Intel	DAOS	10	420	758.71	164.77	3,493.56
4	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	10	960	612.87	105.28	3,567.85
5	ISC20	TACC	Frontera	Intel	DAOS	10	420	508.88	79.16	3,271.49
6	ISC21	National Supercomputer Center in GuangZhou	Venus2	National Supercomputer Center in GuangZhou	kapok	10	480	474.10	91.64	2,452.87
7	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	10	380	440.64	95.80	2,026.80
8	ISC21	Supermicro		Supermicro	DAOS	10	1,120	415.04	112.17	1,535.63
9	SC19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	249.50	86.97	715.76
10	SC20	EPCC	NextGENIO	BSC & JGU	GekkoFS	10	3,800	239.37	45.79	1,251.32
11	ISC21	Olympus Storage Technology Innovation Lab	OceanStor	Huawei	OceanFS	10	960	220.10	69.49	697.15
12	SC20	Johannes Gutenberg University Mainz	MOGON II	JGU (ADA-FS)& BSC (NEXTGenIO)	GekkoFS	10	240	167.64	22.97	1,223.59
13	SC20	DDN	DIME	DDN	IME	10	110	161.53	101.60	256.78
14	SC19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	10	2,610	156.51	56.22	435.76
15	ISC21	University of Tsukuba	Cygnus	OSS	CHFS	10	240	148.69	30.39	727.61
16	ISC21	Joint Institute of Nuclear Research	Govorun	RSC	DAOS	10	160	132.06	20.19	863.69
17	SC20	TACC	Frontera	DDN	IME	10	280	109.91	176.23	68.55



14	SC19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	10	2,610	156.51	56.22	435.76
15	ISC21	University of Tsukuba	Cygnus	OSS	CHFS	10	240	148.69	30.39	727.61
16	ISC21	Joint Institute of Nuclear Research	Govorun	RSC	DAOS	10	160	132.06	20.19	863.69
17	SC20	TACC	Frontera	DDN	IME	10	280	109.91	176.23	68.55

#15 in 10 node list

#23 in full list

Summary

- 15% of compute time of Wisteria-O operated by U Tokyo will be provided for MCRP 2022
- Cygnus-BD will be introduced in 2022
  - Big memory and high-performance storage for data-driven and AI-driven science
- Research of ad hoc parallel file system
  - Better and scalable performance utilizing persistent memory
  - #15 in 2021 June IO500 10 node list, #23 in full list