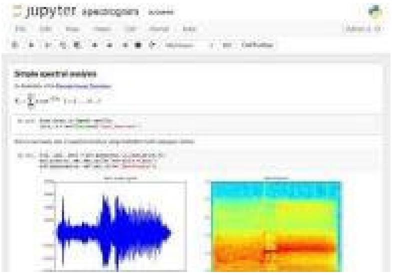
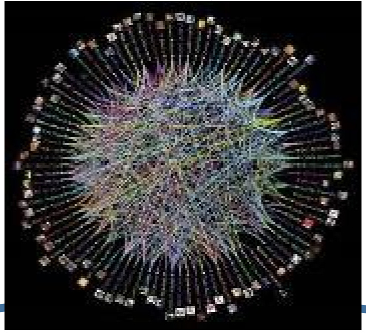
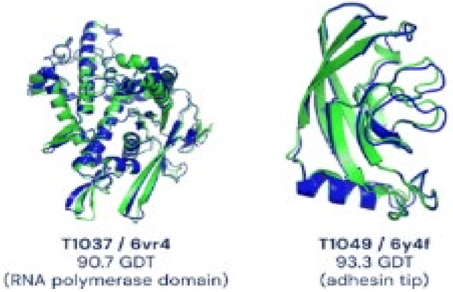
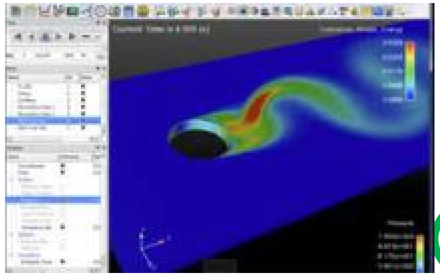
TSUBAME4.0:

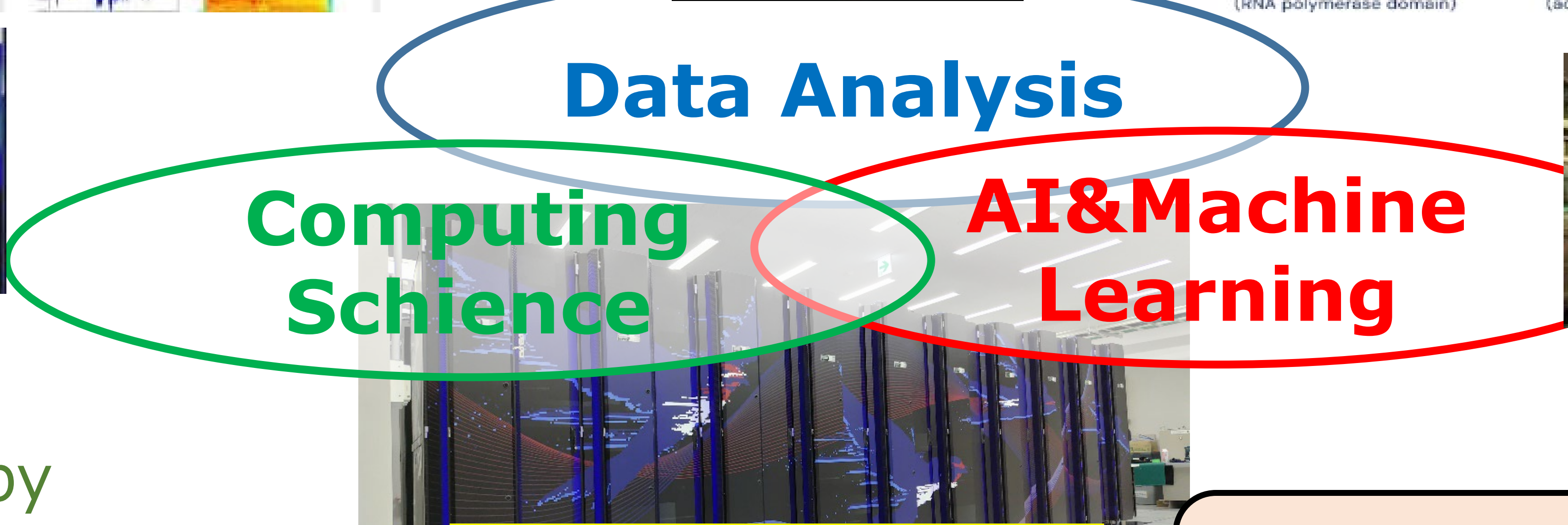
Supercomputer for Everybody for collaboration of Computing Science/Data Analysis/AI&Machine Learning







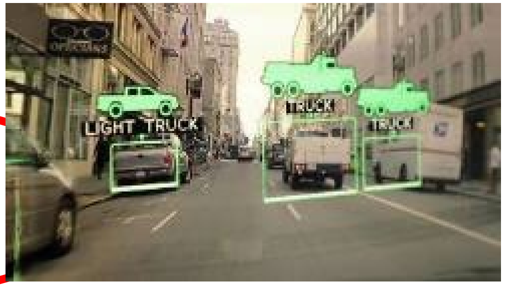




Data Analysis

Computing Schience

AI&Machine Learning



Integrated by
Hewlett Packard Enterprise

TSUBAME4.0
Operation is started in Apr 2024 !

It will remain an important infrastructure on the unified university (Autumn 2024~)

TSUBAME4.0 Overview

Computing Nodes:
240 HPE Cray XD665
(4x H100 + 2x 96-core EPYC)

Total computation speed:

- **66.8 PFlops** (FP64)
- **952 PFlops** (FP16 for AI)

Storage:
HPE Cray ClusterStor E1000

Total capacity:

- 44 PByte (Hard disk part)
- 327 TByte (SSD part)

System in 30 racks

- Compute: 23 racks
- Storage&mgmt.: 7 racks

Installed in Suzukakedai campus, Tokyo Tech

Integrated by HPE

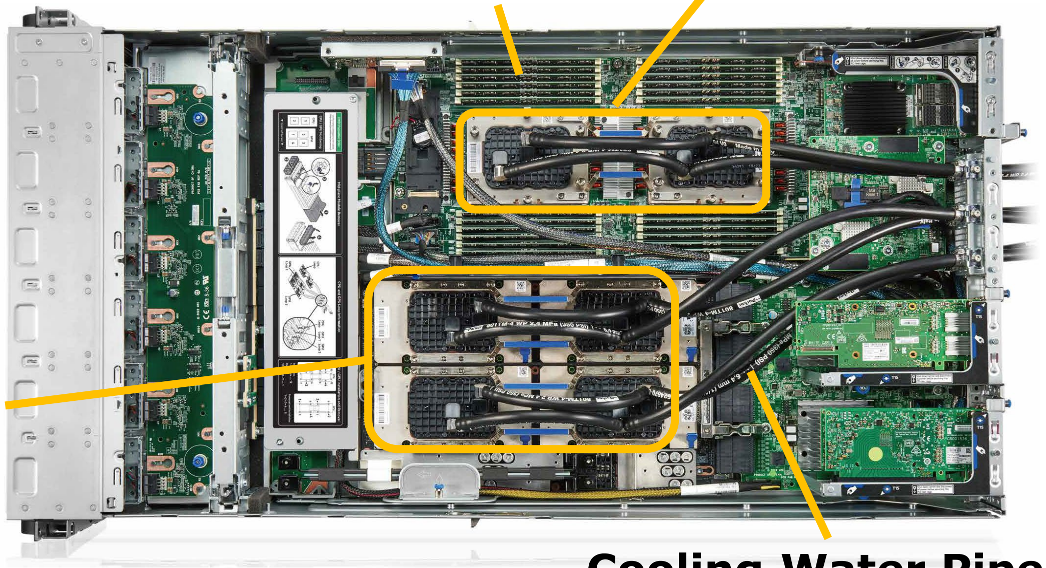
TSUBAME4.0 Specifications

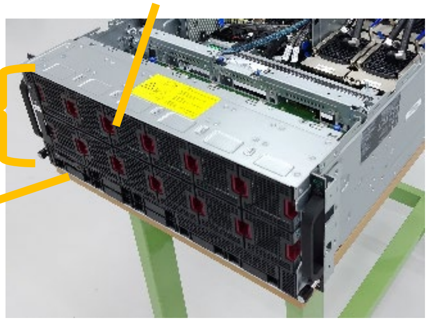
	TSUBAME3.0(2017-)	TSUBAME4.0(Apr 2024-)
Computational Performance		
• FP64 Matrix	12PFlops	**66.8PFlops (5.5x)**
• FP64 Vector		**34.7PFlops (2.8x)**
• Deep Learning (FP16 Matrix)	47PFlops	**952PFlops (20x)**
GPU Memory Bandwidth	1.56 PB/s	**3.07 PB/s (1.97x)**
Number of Nodes	540 Nodes (homogeneous config)	**240 nodes (homogeneous config)**
GPUs	2160 NVIDIA P100	**960 NVIDIA H100**
Cooling / Inlet Water Temperature	Free Cooling with Cooling Tower 32℃	**Chiller 20℃**
Power Consumption (incl. cooling)	1080kW (Spec. value) 400~600kW(Operation)	**1820kW (Spec. value) 450~800kW(Expected. Op.)**

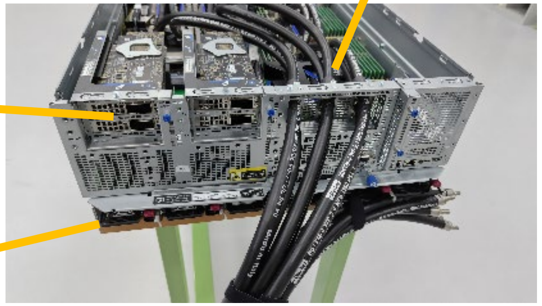
730kW Today

TSUBAME4.0 Node – HPE Cray XD665 4U Server

CPU: 2x AMD EPYC 9654
96 cores, 2.4~3.55GHz
Memory: 24 x 32GiB DDR5-4800
768 GiB in total
GPU: 4x NVIDIA H100
SXM5 94GB HBM2e
Network: 4x InfiniBand NDR200
SSD: 1.92TB NVMe







TSUBAME4.0 Node Specifications

	TSUBAME3.0	**TSUBAME4.0**
CPU	Intel Xeon 2680v4 ×2	**AMD EPYC 9654 ×2**
• Clock, #cores	2.4GHz, 28 cores(=14×2)	**2.4GHz, 192 cores (=96×2)**
Main Memory	DDR3-2400 4ch×2	**DDR5-4800 12ch×2**
• Size	256GiB	**768GiB**
Network	OmniPath 100Gbps×4	**InfiniBand NDR 200Gbps×4**
OS	SUSE Linux Enterprise 12	**RedHat Enterprise Linux 8**
GPU	NVIDIA P100 SXM×4	**NVIDIA H100 SXM5 94GB HBM2e ×4 ***
Specs per GPU:		
• Speed (FP64)	5.3TFlops	**66.9TFlops (Matrix), 33.4TFlops(Vector)**
• Mem Size	16GB	**94GB**
• Mem Speed	0.73TB/s	**2.39TB/s**

*: H100 customized variant aka HPC SKU (memory size and speed differ from normal H100)

GPU Model Used in TSUBAME4

NVIDIA H100 SXM5 94GB HBM2e

- 4GPUs × 240nodes = 960GPUs in total

	H100 PCIe model	H100 94GB Model (TSUBAME4)	H100 SXM5 (Normal Model)
Speed (FP64)	51TFlops (Mat) 26TFlops (Vec)	< **67TFlops (Mat)** **34TFlops (Vec)**	
Speed (FP16)	756TFlops (Mat)	< **990TFlops (Mat)**	
Mem Size	80GB	< **94GB**	> 80GB
Mem Speed	2.0TB/s	< **2.39TB/s**	< **3.35TB/s**

HBM2e (0.4TB/s?)× 5　HBM2e (0.4TB/s?)× 6　HBM3 (0.67TB/s?) × 5

Larger GPU memory size is important for recent AI tasks (LLM models, AlphaFold/OmegaFold…)

⇔ Trade off with lower memory speed ➔ Longer computing time

CPU Model Used in TSUBAME4

AMD EPYC 9654 x 2CPUs per node

- Genoa 96cores with chiplet technology
- 12 DDR5-4800 ch × 2CPUs = 24 ch
- (96 cores x 2CPUs) × 240nodes = 46,080 in total

Largest core counts as the current-gen x86 CPUs

➔ Higher throughput in data analysis

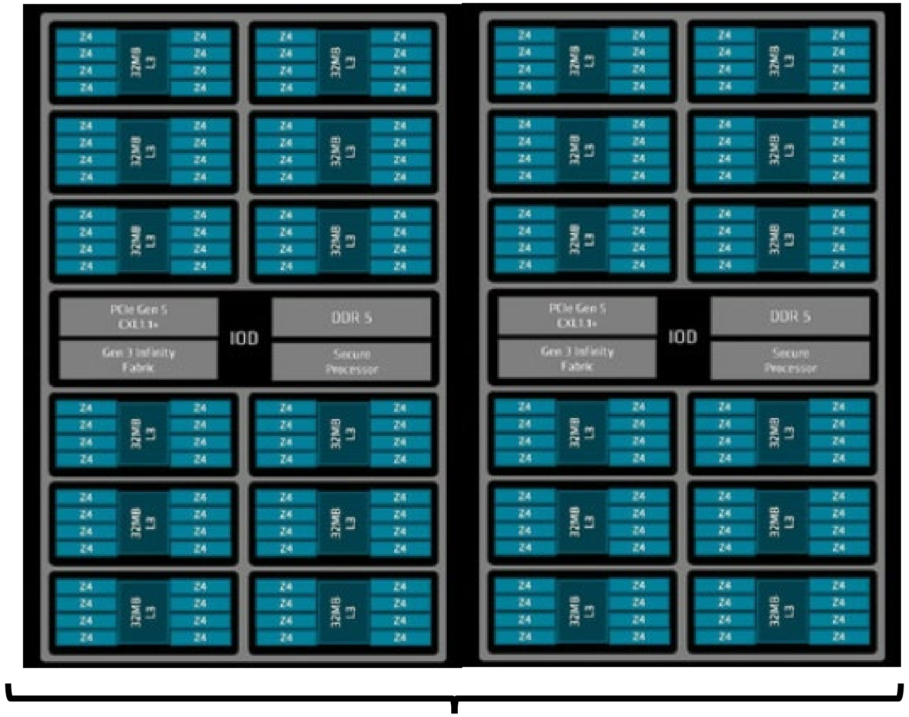


Photo by AMD

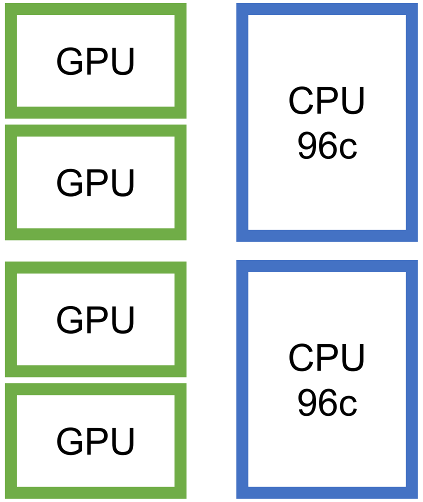
96 cores x 2CPUs = 192 cores

Dynamic Node Partitioning in TSUBAME4

TSUBAME3.0 vs TSUBAME4.0

- #CPU cores per node: 28 cores vs 192 cores
- #Nodes: 540 nodes vs 240 nodes

Since a single node is more precious, node partitioning is even more important



More instance types are defined

current plan, to be changed

Balanced types

- 192cores + 4GPUs
- 96cores + 2GPUs
- 48cores + 1GPU
- 24cores + 0.5GPU

GPU types

- 8cores + 1GPU
- 4cores + 0.5GPU

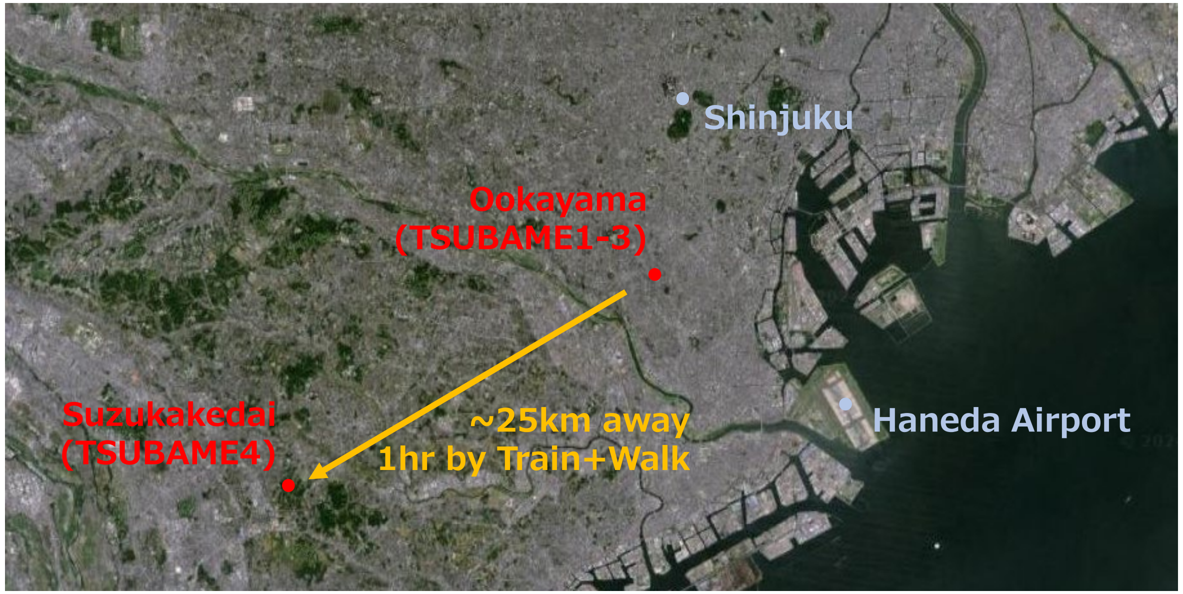
CPU types

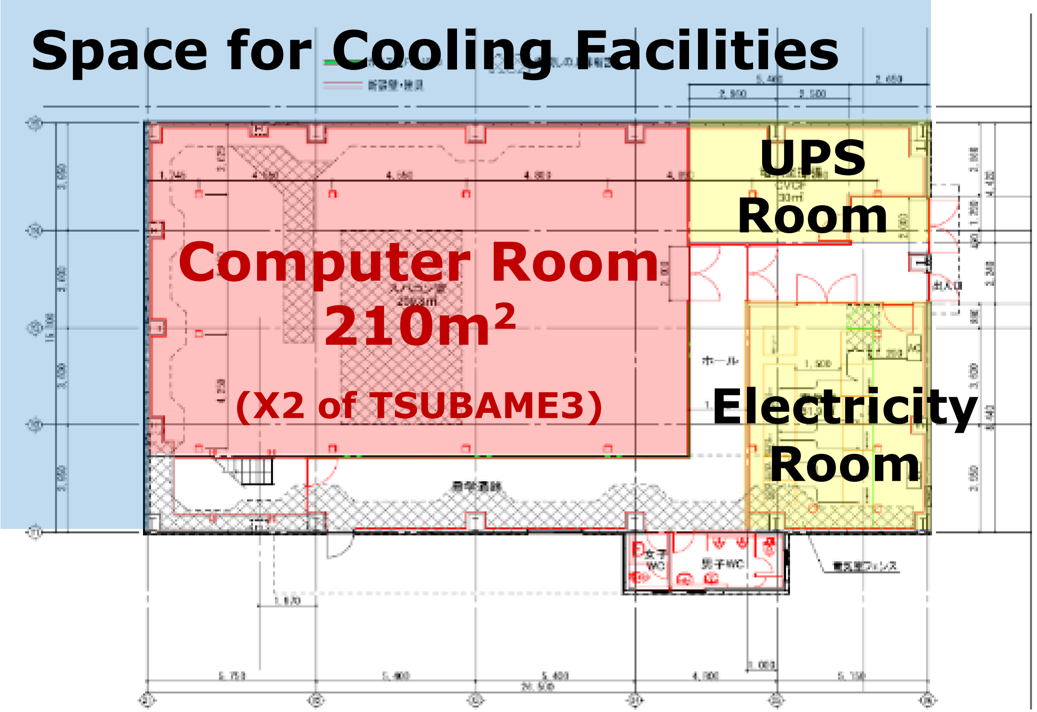
- 160cores
- 80cores
- 40cores
- 16cores
- 8cores
- 4cores

A GPU is partitioned by MIG

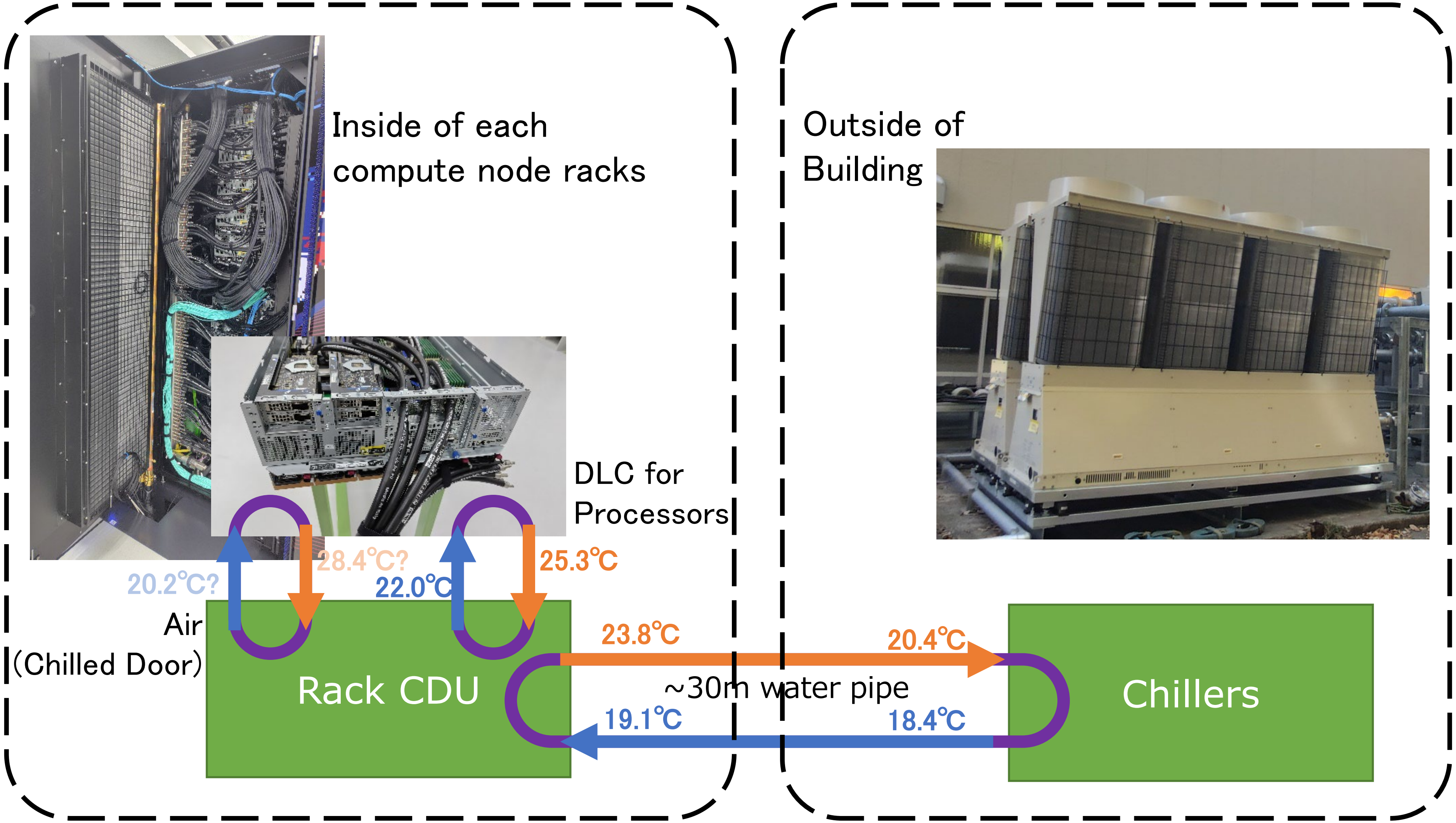
New Data Center for TSUBAME4.0

- Renovated old experimental factory for power plant research
- TSUBAME3 room in Ookayama was too small (~100m²)
- Suzukakedai is located in Kanagawa Pref (Yokohama City)
  - Applicable local laws are different





TSUBAME4 Cooling loops



Inside of each compute node racks

DLC for Processors

Air (Chilled Door)

20.2℃?

28.4℃?

22.0℃

25.3℃

Rack CDU

23.8℃

19.1℃

~30m water pipe

Outside of Building

20.4℃

18.4℃

Chillers

- System is cooled by chillers
  - TSUBAME3 compute nodes was cooled by cooling towers (storage, mgmt.: old chillers)
- Processors are cooled by water, other parts are cooled by air
  - 80~90% by water
  - Same as TSUBAME3
- Air temperature at CDU is not accurate
  - Measured via insulation materials
- Other sensors' accuracy is not verified