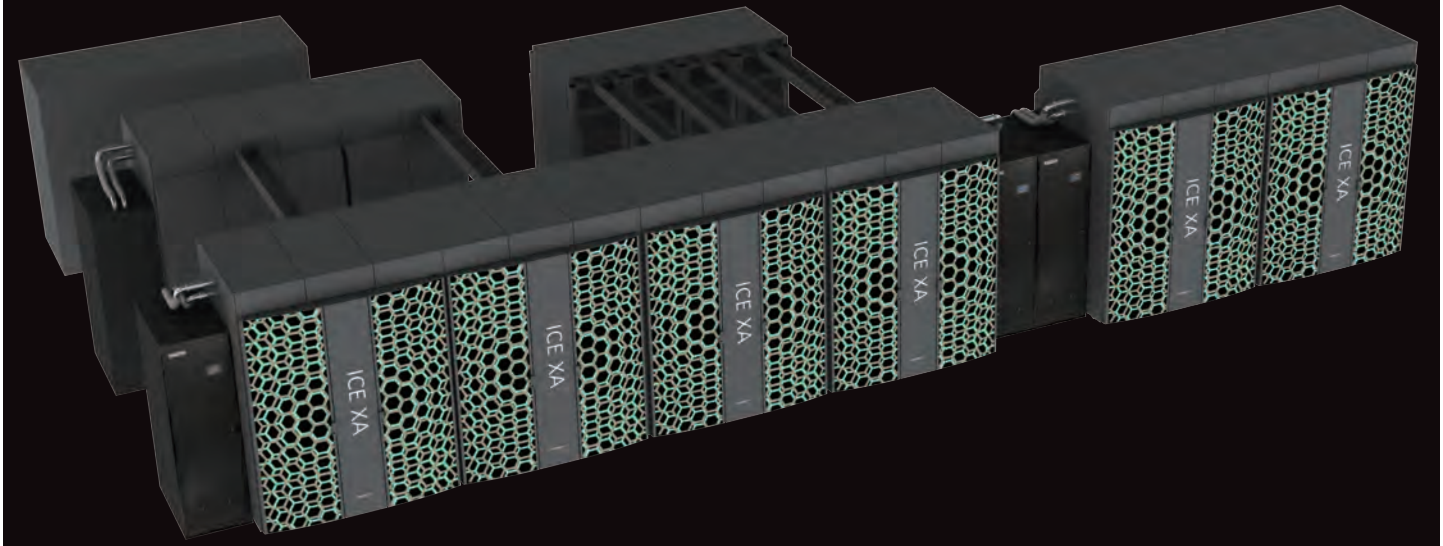
Global Scientific Information and Computing Center, Tokyo Institute of Technology

HARDWARE SOFTWARE SPECIFICATIONS

TSUBAME3.0

- 🟦 GPU Accelerated Compute Node
- 🟪 High-Speed Network Interconnect
- 🟥 High-Speed and Highly Reliable Storage Systems
- 🟩 Efficient Cooling System for Green Operation
- ⬛ System and Application Software



Peak Perf.: 12.15 PFlops (DP), 24.3 PFlops (SP), 47.2PFlops (HP)
Full bisection bandwidth Fat-Tree, qual-rail Omni-Path network
15.9PB Lustre filesystem at 150GB/s
Warm water cooling using cooling towers

GPU-Equipped High-Performance Compute Nodes

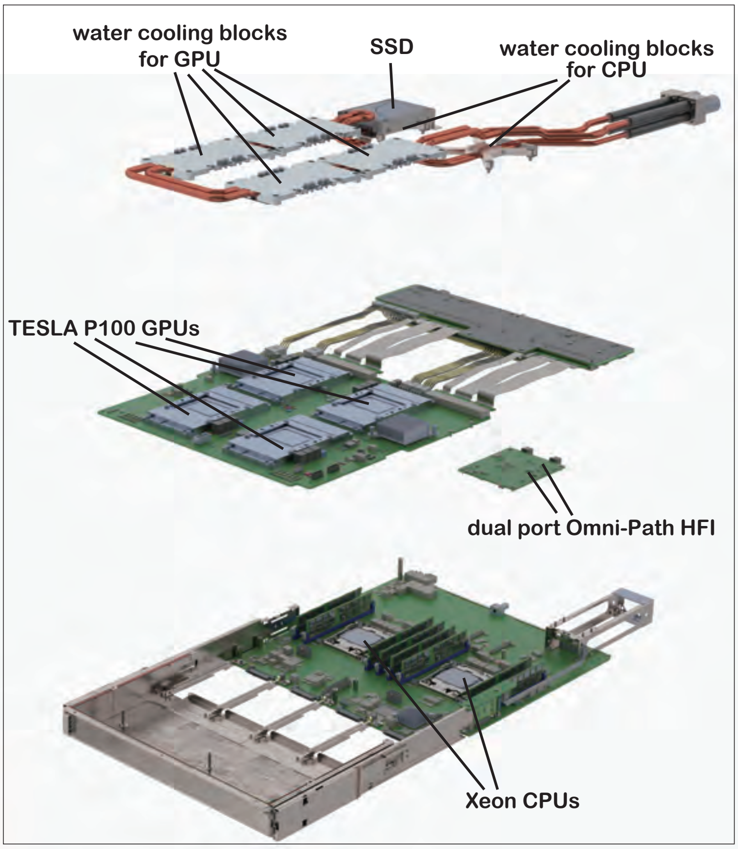
TSUBAME3.0 system includes 540 compute nodes, which provides 12.15 PFlops performance in total. Each compute node is equipped with two CPUs and four GPUs in a compact design blade. In addition, high-speed network with four Omni-Path HFIs and large capacity SSD accelerate applications' performance in Big Data and AI areas.

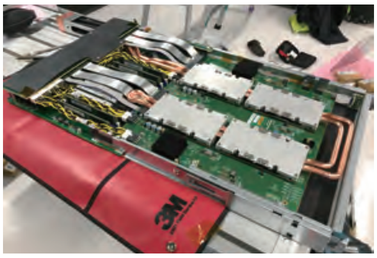
HPE SGI ICE-XA (SGI 8600) IP139-SXM2 540 nodes

CPU:	Intel Xeon E5-2680 V4 (Broadwell-EP, 2.4GHz) ×2 sockets 14 cores per socket, total 28 cores per node.
GPU:	NVIDIA TESLA P100 for NVlink-Optimized servers ×4.
Memory:	256GB (DDR4-2400 32GB module ×8)
SSD:	Intel DC P3500 2TB (NVMe, PCI-E 3.0 x4, R2700/W1800)
Network:	Intel Omni-Path Architecture HFI (100Gbps) ×4

SGI ICE XA cabinet
(front door opened)

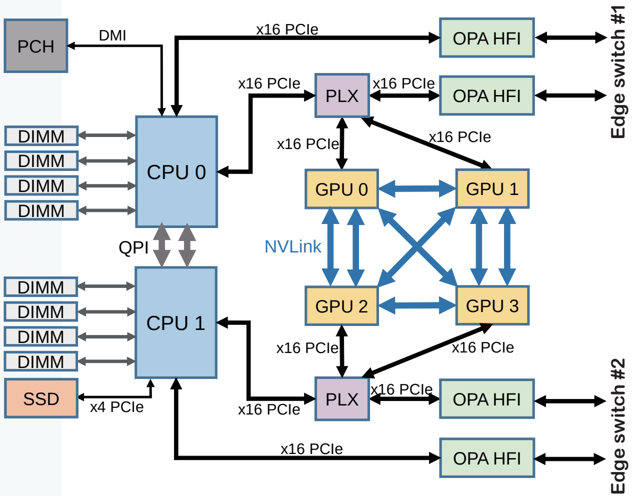
Each E-Rack contains up to four chassis, and each chassis contains nine compute nodes.





A compute node (cover opened)

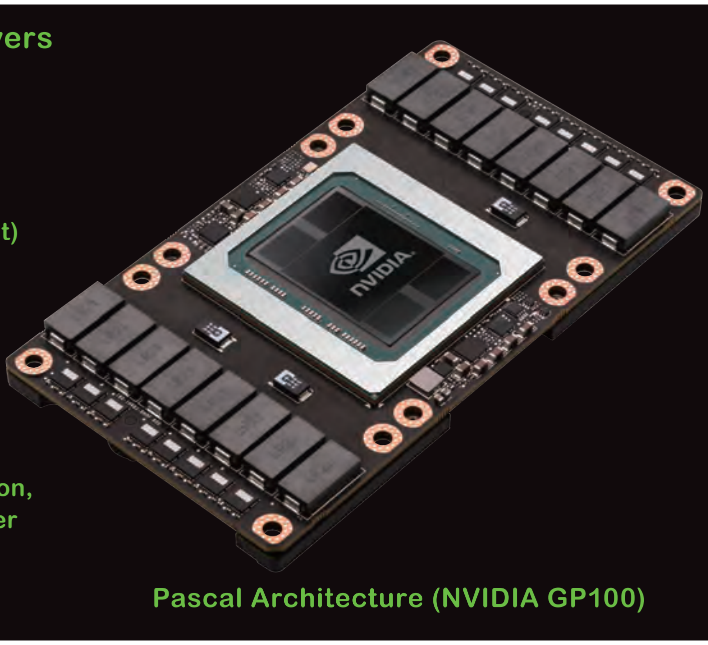
Block Diagram



Tesla P100 for NVlink-Optimized servers

Peak performance :
5.3 TFLOPS (double precision)
10.6 TFLOPS (single precision)
21.2 TFLOPS (half precision)
Shader clock : 1328 MHz (1480MHz with boost)
Number of CUDA cores (SP) : 3,584
Streaming Multiprocessors : 56
On-board memory: 16GB HBM2
Memory bandwidth : 720GB / sec
TDP: 300W

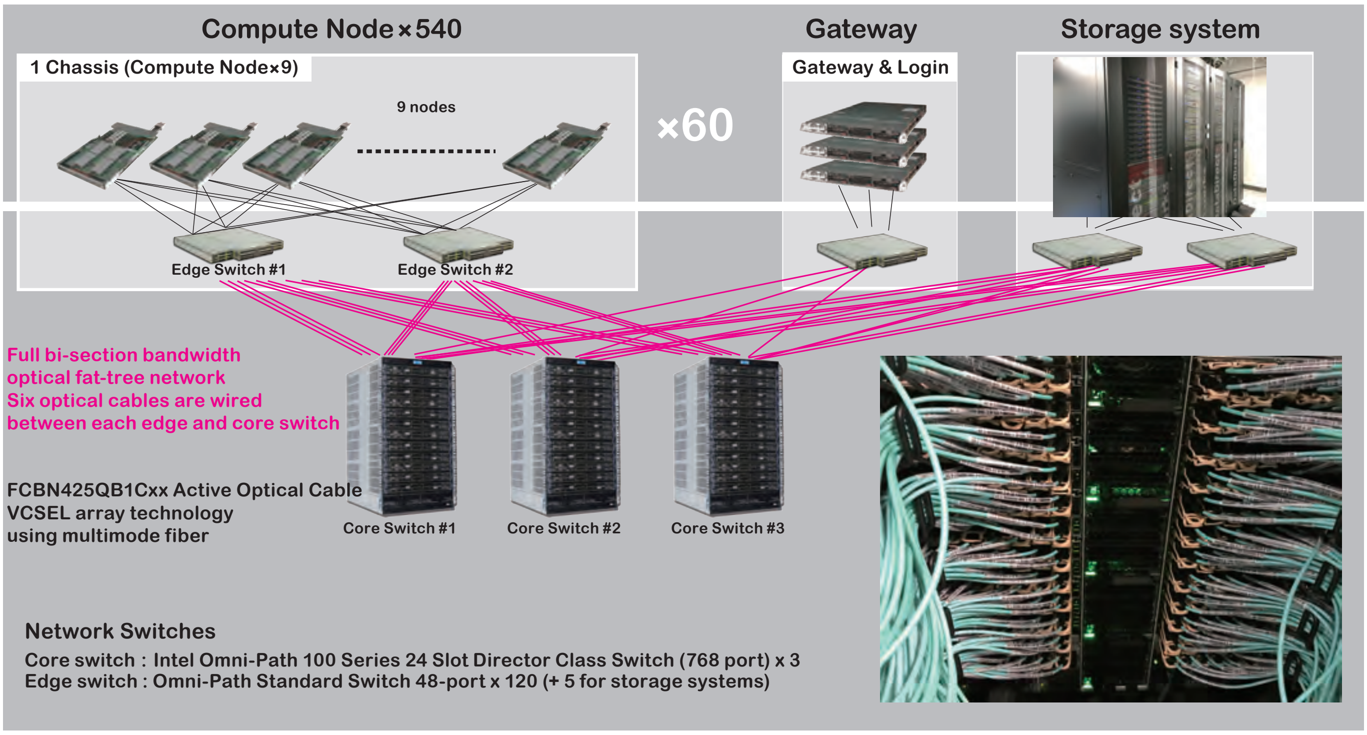
NVLink is direct interconnect between GPUs. Bandwidth of a link is 20GB/s for each direction, and four NVLinks enable 160GB/s data transfer in addition to the PCI-E bandwidth.



Pascal Architecture (NVIDIA GP100)

High-Speed Interconnect

Compute nodes of TSUBAME3.0 interconnected with Omni-Path Architecture of full bi-section bandwidth fat-tree network achieving 432Tbps. End-to-End latency between the compute nodes is extremely low in microsecond-order time, resulting in high-speed performance and high-speed connection to highly reliable storages.



Full bi-section bandwidth
optical fat-tree network
Six optical cables are wired
between each edge and core switch

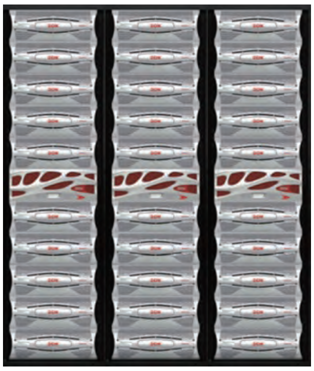
FCBN425QB1Cxx Active Optical Cable
VCSEL array technology
using multimode fiber

Network Switches

Core switch : Intel Omni-Path 100 Series 24 Slot Director Class Switch (768 port) x 3
Edge switch : Omni-Path Standard Switch 48-port x 120 (+ 5 for storage systems)

Lustre

/gs/hs0 "Global Work Space" #1
/gs/hs1 "Global Work Space" #2
/gs/hs2 "Global Work Space" #3

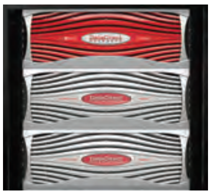


DDN EXAScaler x 3sets : Total 15.9PB, 150GB/s
CIFS gateway server x 2

Each EXAScaler consists of
SFA14KXE + SS8462 x 10
EF4024(MDS) x 2 + ED4024(MDT) x 2
10TB 7.2Krpm NL-SAS HDD x 700 (incl. 20 spare)
5.3PB effective capacity (7.0PB physical)

Home

HOME "NFS for compute nodes"

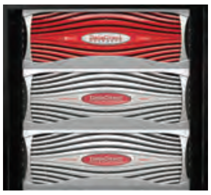


DDN GridDirector : 45TB

SFA7700X + SS8460 x 2
300GB 2.5" 10Krpm SAS HDD x 226
(Data: 200, Meta: 20, Spare: 6)
NAS Gateway(NFS) x 4

Campus storage

HOME "CIFS for other machines"



DDN GridDirector : 36TB

SFA7700X + SS8460 x 2
300GB 2.5" 10Krpm SAS HDD x 186
(Data: 160, Meta: 20, Spare: 6)
NAS Gateway(CIFS) x 4

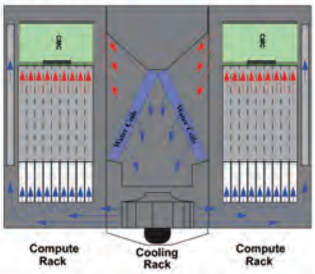
High-Speed and Highly Reliable Storage Systems

TSUBAME3.0 provides massive storage volumes to serve various purposes, including 1.08 PB of SSDs embedded in compute nodes for scratch I/O, 15.9PB of parallel file systems such as Lustre for high speed parallel I/O, and 45 + 36 TB of home storage volumes for providing campus cloud storage services.

Efficient Cooling System for Green Operation

Warm Water Cooling

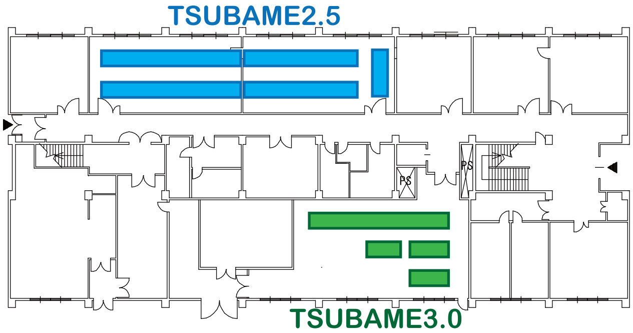
Cooling system requires additional electric power. In order to minimize it, we employ low-power evaporative cooling tower located on the rooffloor of the building. It is estimated that the cooling tower provides less than 32 degrees water even in summer, and the compute nodes of TSUBAME3.0 are designed to work with the warm water.



CPUs and GPUs are direct water-cooled. The other components are air-cooled using rack-contained built-in heat exchanger (Water Coils in the figure) between water and air, allowing low-cost, high-maintainability, and high energy efficiency. The racks for storage and I/O systems employ rear door cooling technology.

Cooling towers on the roof of the building

Small space installation



Highly efficient cooling of TSUBAME3.0 enables high density compute nodes. The space required for installation is less than half of TSUBAME2.5.

System Software

The job management system and the cluster management system are working together to manage user environment as well as distributing computational resources to the insufficient part by taking from the node pool. Some jobs do not use all of computing resources (CPU cores, GPUs, memory, etc.) on allocated compute nodes. UNIVA Grid Engine supports cgroup and Docker to enable efficient node sharing between multiple jobs.

OS	SUSE Linux Enterprise Server 12 SP2
Batch System	UNIVA Grid Engine

ISV (commercial) Software

(* GPU full or partial support)

Compilers and Debuggers

Intel Compiler (C/C++/Fortran)
PGI Compiler*
(C/C++/Fortran, OpenACC, CUDA Fortran)
Arm Forge*

Applications

ANSYS Workbench*, Mechanical*
ANSYS CFD, Fluent*, HFSS*
COMSOL Multiphysics
LS-DYNA
Gaussian*, Gauss View
MATLAB*
AVS/Express, AVS/Express PCE
Schrödinger Small-Molecule Drug Discovery Suite*

ABAQUS*, ABAQUS CAE
MSC Nastran*, Patran, Marc*
CST STUDIO SUITE* (MW-Studio*)
AMBER*
Materials Studio, Discovery Studio
Mathematica*
Maple*

Yellow: The license for all users White: The license for Tokyo Tech users

Published by Global Scientific Information and Computing Center, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, JAPAN
TEL : +81-3-5734-2087 FAX : +81-3-5734-3198 E-mail : tsubame@gsic.titech.ac.jp

https://www.gsic.titech.ac.jp/

201906