Global Scientific Information and Computing Center, Tokyo Institute of Technology

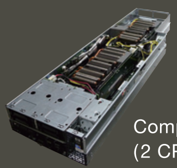
HARDWARE
SOFTWARE
SPECIFICATIONS

Global Scientific Information and Computing Center

TSUBAME 2.0

HARDWARE AND SOFTWARE SPECIFICATIONS

- Large-Scale GPU-Equipped High-Performance Compute Nodes
- High-Speed Network Interconnect
- High-Speed and Highly Reliable Storage Systems
- Low Power Consumption and Green Operation
- System and Application Software



Compute Node
(2 CPUs , 3GPUs)

1.7 TFLOPS
58.0 GB (CPU) + 9.7 GB (GPU)

Rack (30 nodes)

51.0 TFLOPS
2.03 TB

System (58 Racks)

1442nodes
2952 CPU sockets :
224.7 TFLOPS
※Turbo boost

4264GPUs :
2196 TFLOPS

Total :
2420 TFLOPS

Memory :
103.9 TB

Large-Scale GPU-Equipped High-Performance Compute Nodes

Compute nodes consist of three types of nodes: Thin, Medium and Fat nodes. Thin nodes, which provide most of the overall compute performance, are equipped with two CPUs and three Fermi core GPUs in a compact design 17/2 inches in width and 2U size in height. In addition, two QDR Infiniband HCAs are connected to dedicated PCI Express Buses to secure the communications bandwidth. Power supply units are organized with 3+1 redundancy, improving the node reliability significantly.

Thin Node 1408 nodes

HP ProLiant SL390s

GPU : NVIDIA Tesla M2050 (Fermi Core)×3 515GFLOPS VRAM 3GB/GPU
CPU : Intel Xeon X5670 2.93GHz ×2
6 core/socket 76.7 GFLOPS (12cores/node) ※Turbo boost : 3.196GHz
Memory : 58GB DDR3 1333MHz (partly 103GB)
SSD : 60GB ×2 (120GB/node) (partly 120GB ×2 (240GB/node))

Medium Node 24 nodes

HP ProLiant DL580 G7

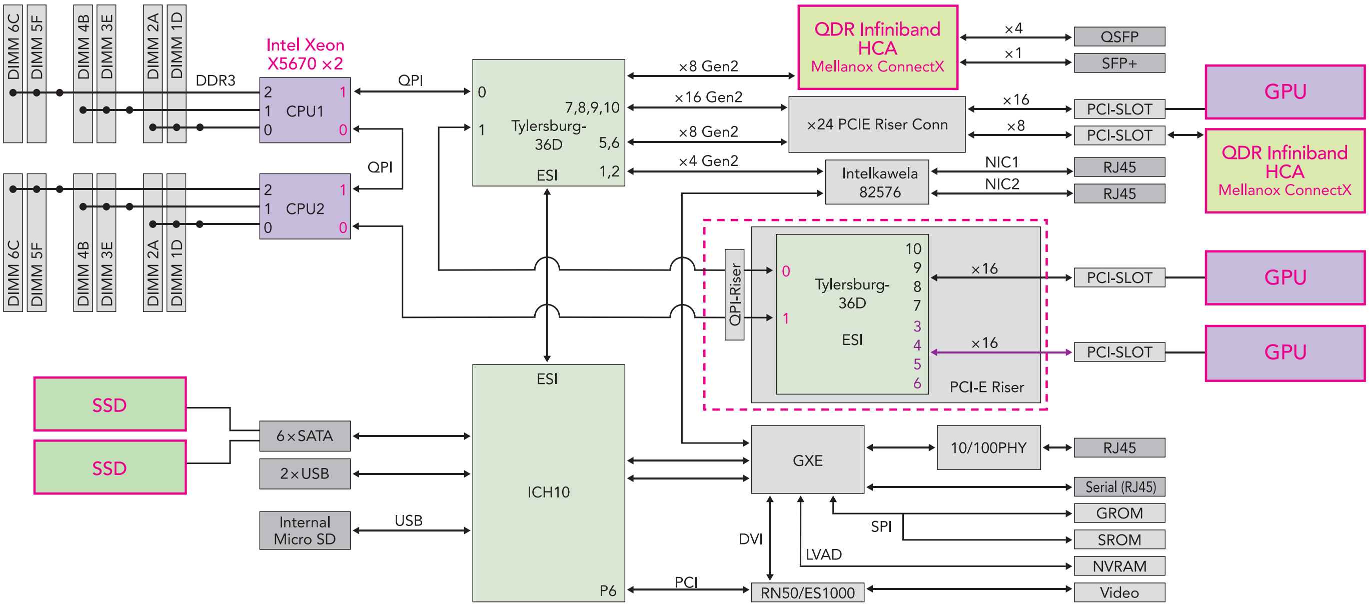
CPU: Intel Xeon X7550
(Nehalem-EX)
2.0 GHz ×4 sockets
(32cores/node)
GPU: NVIDIA Tesla S1070
Memory: 137 GB (DDR3 1066MHz)
SSD: 120GB ×4 (480GB/node)
Infiniband: QDR

Fat Node 10 nodes

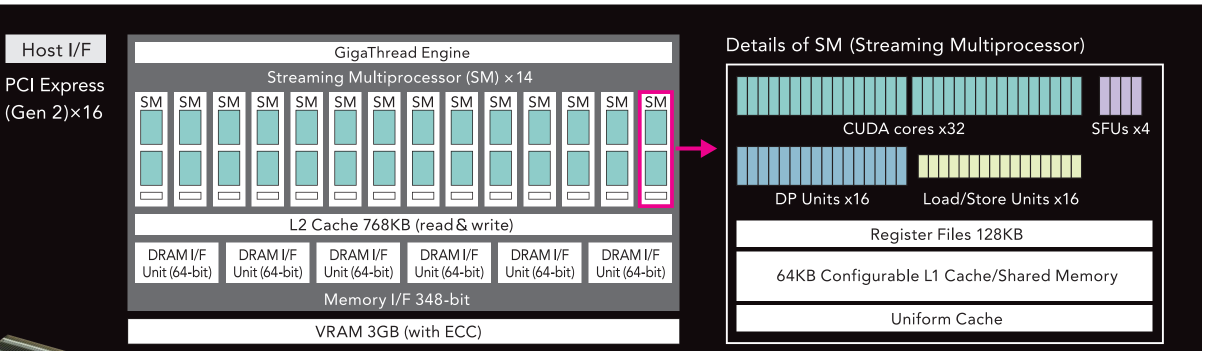
HP ProLiant DL580 G7

CPU: Intel Xeon X7550
(Nehalem-EX)
2.0 GHz ×4 sockets
(32cores/node)
GPU: NVIDIA Tesla S1070
Memory: 274 GB (8 nodes),
548 GB (2 nodes)
DDR3 1066MHz
SSD: 120GB x5 (600GB/node)
Infiniband: QDR

Block Diagram of Thin Node



Details of GPU



NVIDIA GPU
Tesla M2050

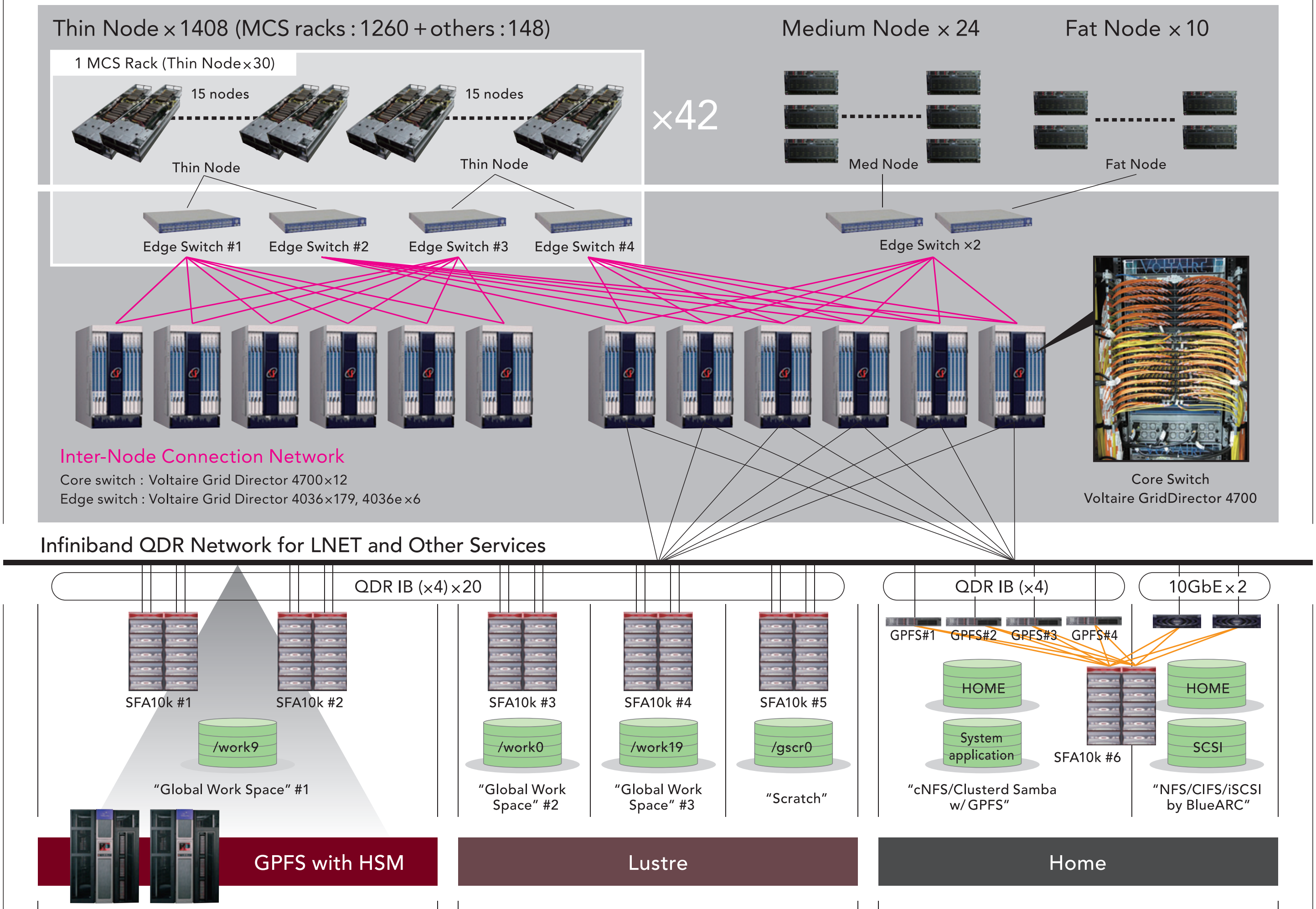
Peak performance : 515 GFLOPS
(double precision)
1030 GFLOPS
(single precision)
Shader clock : 1.15GHz
Number of CUDA cores (SP) : 448 cores

Streaming Multiprocessor (SM) : 14 SMs
CUDA core (SP) / SM : 32 cores
DP unit / SM : 16
SFU / SM : 4 units
Warp scheduler / SM : 2 units
Shared memory / SM : 16KB or 48KB
Writable L1 cache / SM : 48KB or 16KB
Writable L2 cache : 768 KB

Memory bandwidth : 150.2GB / sec
Memory clock : 1.565GHz (GDDR5)
ECC memory : support
On-board memory : 3GB

High-Speed Network Interconnect

Compute nodes of TSUBAME2.0 interconnected with Dual-Rail QDR InfiniBand networks of Fat-Tree type full bi-section bandwidth achieve 200Tbps. End-to-End latency between the compute nodes is extremely low in microsecond-order time, therefore resulting in high-speed performance and high-speed connection to highly reliable storages.
This network is linked by more than 3000 optical fiber cables in a total length of 100km.



Thin Node × 1408 (MCS racks : 1260 + others : 148)

1 MCS Rack (Thin Node × 30)

15 nodes

15 nodes

Thin Node

Thin Node

×42

Medium Node × 24

Fat Node × 10

Med Node

Fat Node

Edge Switch #1

Edge Switch #2

Edge Switch #3

Edge Switch #4

Edge Switch ×2

Inter-Node Connection Network

Core switch : Voltaire Grid Director 4700×12
Edge switch : Voltaire Grid Director 4036×179, 4036e ×6

Core Switch
Voltaire GridDirector 4700

Infiniband QDR Network for LNET and Other Services

QDR IB (×4) × 20

QDR IB (×4)

10GbE × 2

SFA10k #1

SFA10k #2

/work9

“Global Work Space” #1

SFA10k #3

/work0

“Global Work Space” #2

SFA10k #4

/work19

“Global Work Space” #3

SFA10k #5

/gscr0

“Scratch”

GPFS#1 GPFS#2 GPFS#3 GPFS#4

HOME

System application

SFA10k #6

HOME

SCSI

“cNFS/Clusterd Samba w/ GPFS”

“NFS/CIFS/iSCSI by BlueARC”

GPFS with HSM

Lustre

Home

GPFS with HSM 2.4 PB+4PB (Tape)

Servers: HP ProLaint DL380 G6 × 4
Intel Westmere EP × 2, 48GB Mem,
QDR (4×)IB × 2

HP ProLaint DL 360 × 4
Intel Westmere-EP × 2, 24GB Mem,
QDR (4×)IB × 2

Storage: DDN SFA 10k × 2
2TB SATA × 1200 disks

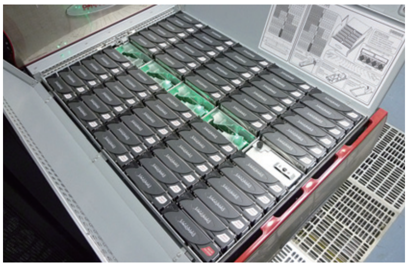
Tape: StorageTek SL8500 × 2
LTO4 × 5000 roles

Lustre 3.6 PB

MDS : HP ProLaint DL360 G6 × 6
Intel Westmere-EP × 2, 48GB Mem,
QDR (4×)IB × 2

OSS : HP ProLaint DL360 G6 × 12
Intel Westmere-EP × 2, 24GB Mem,
QDR (4×)IB × 2

Storage: DDN SFA 10k × 3,
2TB SATA × 1750 disks,
600GB SAS × 50 disks



Home 1.2 PB

cNFS (GridScaler)/Clustred Samba w/GPFS:
HP ProLaint DL380 G6 × 4
Intel Westmere EP × 2, 48GB Mem,
QDR (4×)IB × 2

NFS/CIFS/iSCSI:
BlueArc Mercury 100 × 2
10Gbps × 2

Storage: DDN SFA 10k × 1
2TB SATA × 600 disks

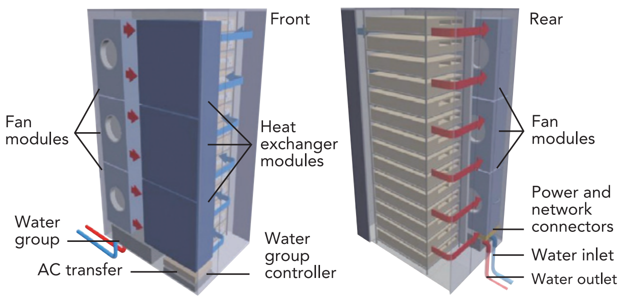
High-Speed and Highly Reliable Storage Systems

TSUBAME2.0 provides 11PB of massive storage volumes to serve various purposes, including about 190TB of SSDs embedded in compute nodes for scratch I/O, 5.9PB of parallel file systems such as Lustre and GPFS for high speed parallel I/O, 1.2 PB of home storage volumes for providing campus cloud storage services, and over 4PB of tape libraries for hierarchical storage management handled with GPFS.

Low Power Consumption and Green Operation

Power performance in Linpack benchmark : 958.35 (MFLOPS/W)
Peak power consumption of system equipment : 1749 (KW)
Average power consumption of system equipment : 1130 (KW)
Idle power consumption for system equipment : 532 (KW)
Yearly average PUE : 1.277

Cooling : Modular Cooling System

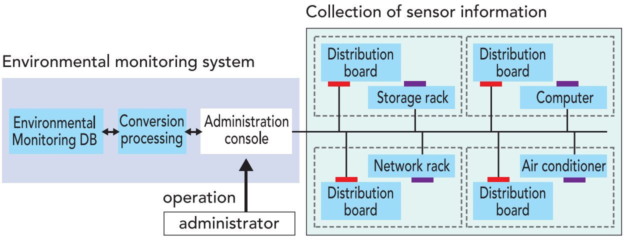


The rack-contained water-cooling system with a built-in heat exchanger is employed, allowing high-density cooling up to 35kW per rack (it is the world' s top class as being 10 times larger than what is used in typical data centers). Homogeneous cooling air is provided through the inlet of the server with automatic open/close doors where a humidifier is unnecessary. Power consumption is minimized with a completely automated temperature control to enable heat removal from 95% to 97% by water cooling. Moreover, polycarbonate doors contribute to a great noise reduction.

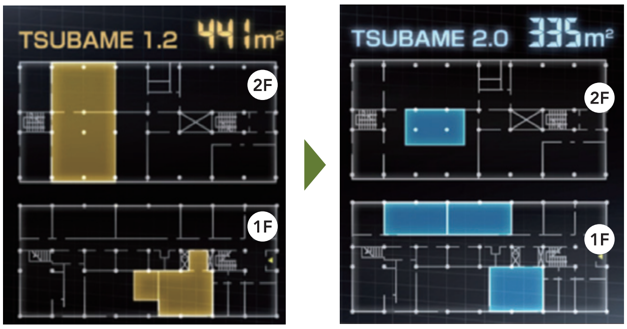
- Peak power consumption of air-conditioning equipment : 875 (KW)
- Average power consumption of air-conditioning equipment : 313 (KW)

Green Operation : Monitoring of Environment

Temperature, power consumption, etc., are observed in real-time not only in the computer room but also to compute nodes and to each rack.



Small space installation



Despite the fact the performance boost is more than 30 times compared to TSUBAME1.2, the space required for installation has narrowed down.

System and Application Software

"Dynamic provisioning" dynamically switched between Windows and Linux

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. Both batch schedulers for Linux and Windows manage to dynamically increase or reduce the compute nodes. The job scheduling also manages to support the execution of a virtual machine.

OS	SUSE Linux Enterprise Server 11 SP1 Windows HPC Server 2008 R2
Batch System	PBS Professional

ISV Application Software

ANSYS Fluent, Workbench
ABAQUS
ABAQUS CAE
MD Nastran
Patran
LS-DYNA
Gaussian
Gauss View
AMBER
Molpro
Materials Studio
Discovery Studio
Scigress
Mathematica
MATLAB
Maple
AVS/Express
AVS/Express PCE
EnSight
PGI Compiler
Intel Compiler
Total View Debugger

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

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