Global Scientific Information and Computing Center, Tokyo Institute of Technology

HARDWARE SOFTWARE SPECIFICATIONS

Global Scientific Information and Computing Center

TSUBAME 2.5

HARDWARE AND SOFTWARE SPECIFICATIONS

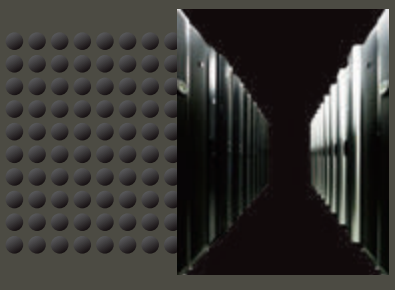
- Large-Scale GPU-Equipped High-Performance Compute Nodes
- High-Speed Network Interconnect
- High-Speed and Highly Reliable Storage Systems
- Low Power Consumption and Green Operation
- System and Application Software

Compute Node
(2 CPUs , 3GPUs)

4.08 TFLOPS.
58.0 GB (CPU) + 18 GB (GPU)

Rack (30 nodes)

122 TFLOPS
2.28 TB



System (58 Racks)

1442nodes
2952 CPU sockets :
224.7 TFLOPS
※Turbo boost

4360GPUs:
5.562 PFLOPS

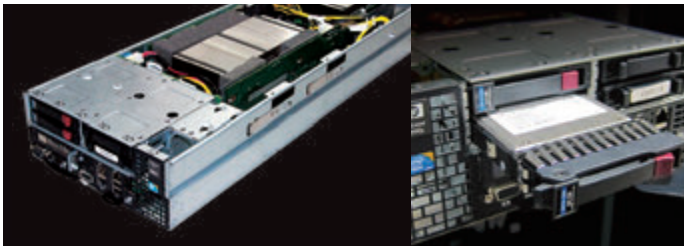
Total:
5.787 PFLOPS

Memory:
116 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 Kepler 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 communication bandwidth. Power supply units are organized with 3+1 redundancy, improving the node reliability significantly.

Thin Node 1408 nodes



HP ProLiant SL390s G7

CPU: Intel Xeon X5670 (Westmere-EP, 2.93GHz) ×2 sockets
six-core per socket, total 12 cores per node.
GPU: NVIDIA Tesla K20X (GK110) ×3. 1.31TFLOPS and VRAM 6GB per GPU
Memory: 58GB DDR3 1333MHz, partly 103GB
SSD: 120GB(60GB×2), partly 240GB(120GB×2)
Network: 4X QDR InfiniBand ×2

Medium Node 24 nodes

HP ProLiant DL580 G7

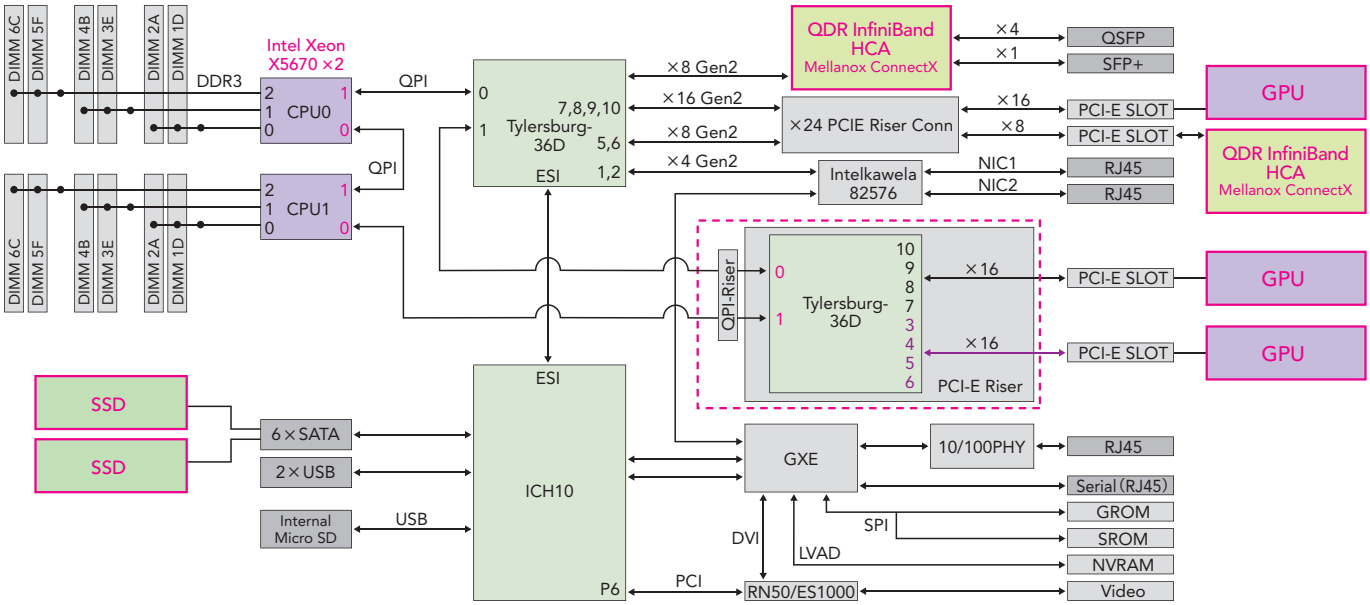
CPU: Intel Xeon X7550
(Nehalem-EX)
2.0 GHz ×4 sockets
(32cores/node)
GPU: NVIDIA Tesla S1070
(NVIDIA Tesla C1060×4) or
NextIO vCORE Express 2070
(NVIDIA Tesla M2070×4)
Memory: 137 GB (DDR3 1066MHz)
SSD: 120GB ×4 (480GB/node)
Network: 4X QDR InfiniBand

Fat Node 10 nodes

HP ProLiant DL580 G7

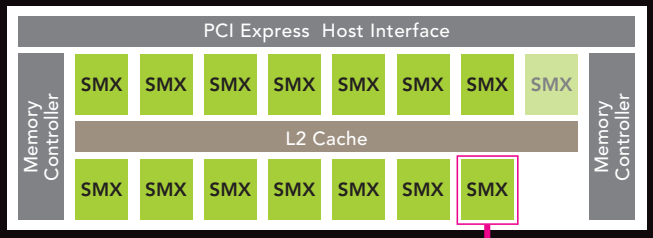
CPU: Intel Xeon X7550
(Nehalem-EX)
2.0 GHz ×4 sockets
(32cores/node)
GPU: NVIDIA Tesla S1070
(NVIDIA Tesla C1060×4)
Memory: 274 GB (8 nodes),
548 GB (2 nodes)
DDR3 1066MHz
SSD: 120GB ×5 (600GB/node)
Network: 4X QDR InfiniBand

Block Diagram of Thin Node

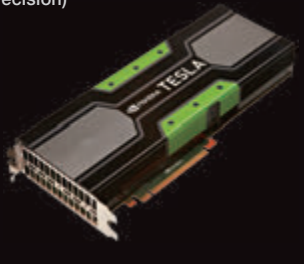


Details of GPU

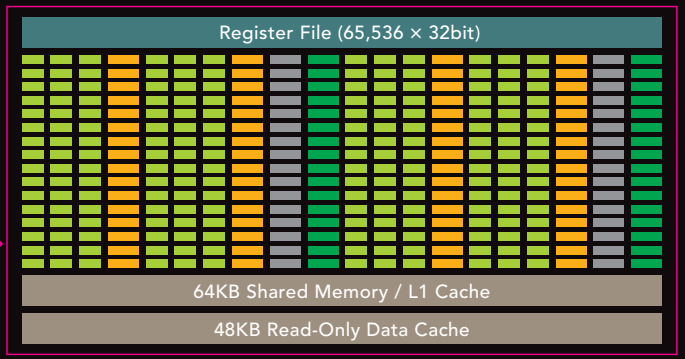
K20X Architecture (Kepler GK110 Core)



Peak performance : 1.31 TFLOPS (double precision)
3.95 TFLOPS (single precision)
Shader clock : 732 MHz
Number of CUDA cores (SP) : 2,688 cores
Streaming Multiprocessor (SMX) : 14 SMX
Writable L2 cache: 1.5MB
Memory bandwidth : 250GB / sec
Memory clock : 2.6GHz (GDDR5)
ECC support:
internal and external memory
On-board memory: 6GB



Details of SM (Streaming Multiprocessor)

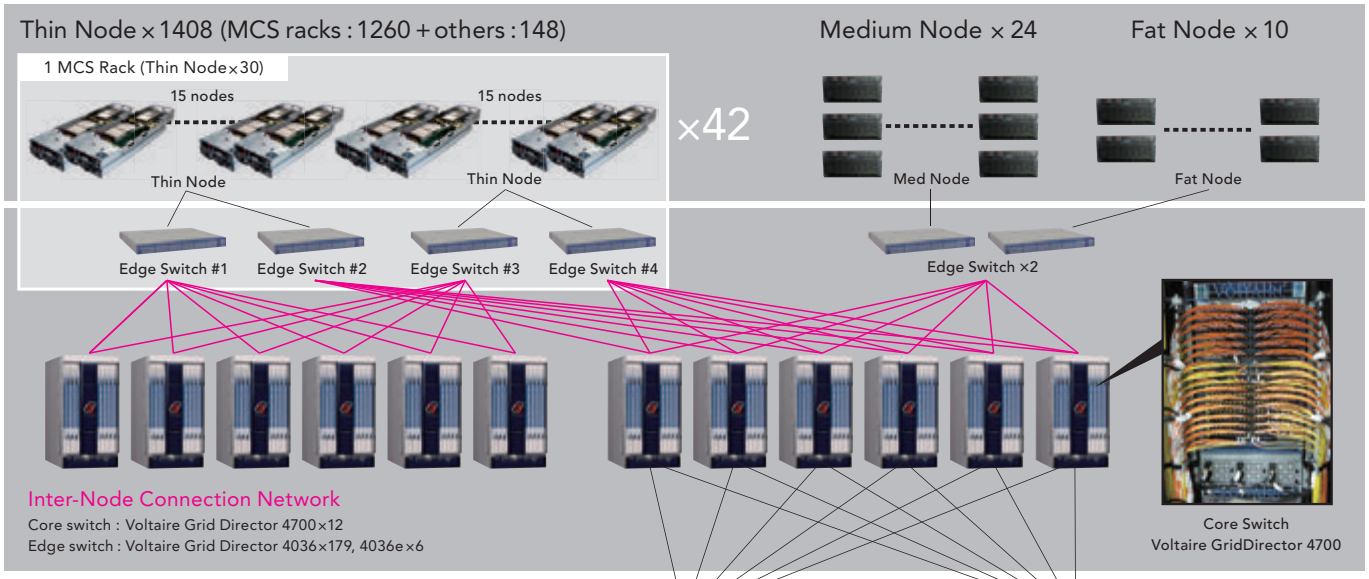


CUDA core (SP) / SMX : 192 cores
DP unit / SMX : 64 SFU / SM : 32 units
Warp scheduler / SMX : 4 units
Shared memory / SMX : 16KB or 32KB or 48KB
Writable L1 cache / SMX : 48KB or 32KB or 16KB
Read-only data cache / SMX: 48KB

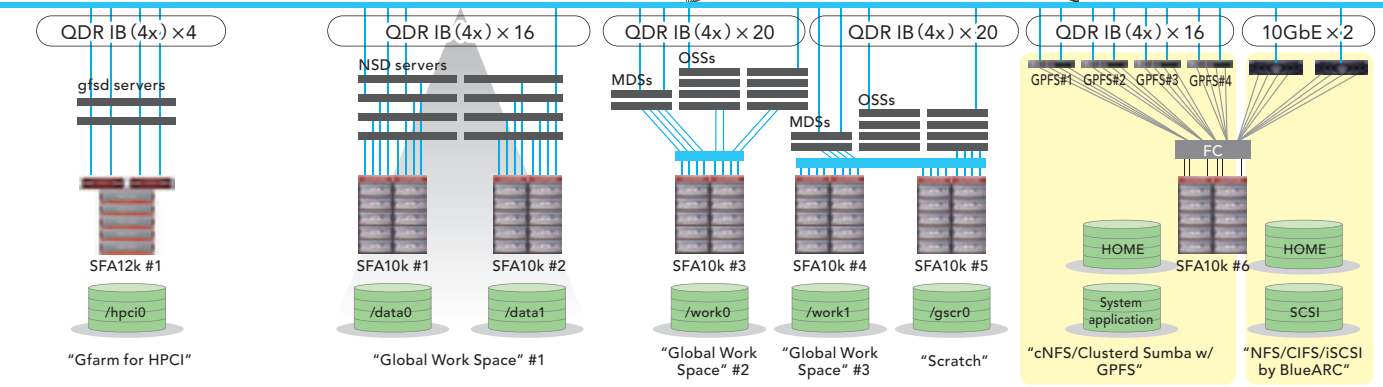
Core
DP Unit
LDST
SFU

High-Speed Network Interconnect

Compute nodes of TSUBAME2.5 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.



InfiniBand QDR Network for LNET and Other Services



HPCI

Gfarm: ~ 600TB

gsfd server:
HP ProLiant DL360p Gen8 × 2
Intel XeonE5 2640 × 2,
64 GB Mem,
QDR IB (4x) × 2

Storage:
DDN SFA12k × 1
4TB SAS HDD × 155 disks

GPFS with HSM

GPFS 2.4 PB

NSD server: HP ProLiant DL380 G6 × 4
Intel Westmere EP × 2,
48GB Mem, QDR IB (4x) × 2

HP ProLiant DL360 G6 × 4
Intel Westmere EP × 2,
24GB Mem,
QDR IB (4x) × 2

Storage: DDN SFA10k × 2
2TB SATA HDD × 1180 disks
600GB SAS HDD × 20 disks

Lustre

Lustre 3.6 PB

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

OSS : HP ProLiant DL360 G6 × 16
Intel Westmere-EP × 2,
24GB Mem,
QDR IB (4x) × 2

Storage: DDN SFA 10k × 3,
2TB SATA HDD × 1770 disks,
600GB SAS HDD × 30 disks

Home

Home 1.2 PB

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

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

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

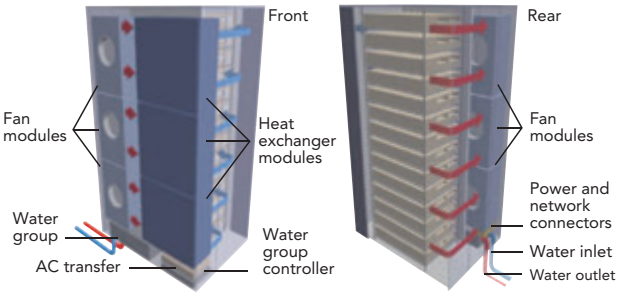
High-Speed and Highly Reliable Storage Systems

TSUBAME2.5 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 : 3068.71 (MFLOPS/W)
Peak power consumption of system equipment : 1620 (KW)
Average power consumption of system equipment* : 698 (KW)
Idle power consumption for system equipment : 470 (KW)
Yearly average PUE : 1.285
[*] Average power consumption is an annual average of TSUBAME2.0's.

Cooling : Modular Cooling System

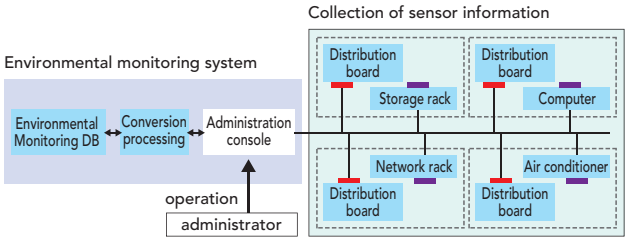


The rack-contained water-cooling system with a built-in heat exchanger is employed, allowing high-density cooling up to 35kW per rack, which is the top class in the world. Homogeneous cooling air is provided through the inlet of the server with automatic open/close doors where a humidifier is unnecessary. 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.

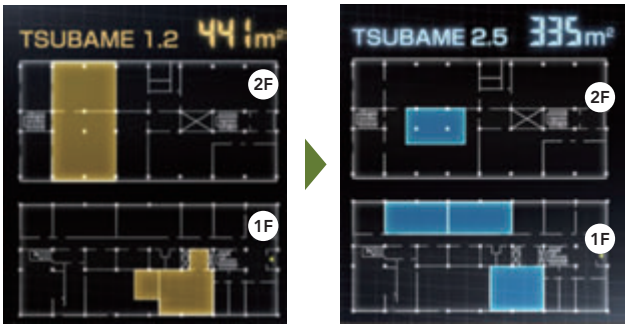
- Peak power consumption of air-conditioning equipment : 460 (KW)
- Average power consumption of air-conditioning equipment* : 204 (KW)

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 70 times compared to TSUBAME1.2, the space required for installation has narrowed down.

System 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 (commercial) Software

(* GPU full or partial support)(As of November, 2013)

Compilers, Debuggers and Libraries

Intel Compiler (C/C++/Fortran)
PGI Compiler*
(C/C++/Fortran, OpenACC, CUDA Fortran)
Total View Debugger*
CAPS Compiler* (HMPP, OpenACC)
CULA* (Numerical Libraries for CUDA)

Applications

ANSYS Fluent*, Workbench*
MSC Nastran*
LS-DYNA
Gaussian, Gauss View
Molpro
Scigress
MATLAB*
AVS/Express, AVS/Express PCE
ABAQUS*, ABAQUS CAE
Patran
CST STUDIO SUITE* (MW-Studio*)
AMBER*
Materials Studio, Discovery Studio
Mathematica*
Maple*
EnSight

: The license for all users : The license for Tokyo Tech users : The license for industrial 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

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

201311