HARDWARE AND SOFTWARE SPECIFICATIONS

Global Scientific Information and Computing Center, Tokyo Institute of Technology

TSUBAME 2.0

- 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

System (64 Ranks)
- 14824 cores
- 24072 GPU sockets
- 24GFLOPS
- 8 Turbo boost
- 48V6400x
- 2141 TFLOPS
- Total
- 24GFLOPS
- Memory: 128 x 1 T

Compute Node
- (2 CPUs x 3 GPUs)
- 2 TFLOPS
- 32GB (CPU) + 4GB (GPU)

Flash (10 nodes)
- 5.5 TFLOPS
- (34 TFLOPS)
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. All nodes in width and 2U size are height. In addition, two ODR Infiniband HCAAs are connected to dedicated PCI Express Buses to secure the communications bandwidth. Power supply units are organized with 3+1 redundancy, improving the nodes' reliability significantly.

Thin Node: 1408 nodes

- **HP ProLiant SL390s**
  - GPU: NVIDIA Tesla M2050 (Fermi Core)-3
  - CPU: Intel Xeon X5670 (6 cores/6 threads)
  - Memory: 16GB DDR3-1333Mhz (partly 16GB)
  - SSD: 600GB x 2 (192GB/node) (partly 192GB x 2 (160GB/node))

Medium Node: 24 nodes

- **HP ProLiant DL580 G7**
  - CPU: Intel Xeon X5750 (24 cores/24 threads)
  - Memory: 128GB (DDRII 1066Mhz)
  - SSD: 1.9TB x 2 (256GB/node)
  - Infiniband: ODR

Fat Node: 10 nodes

- **HP ProLiant DL580 G7**
  - CPU: Intel Xeon X5750 (24 cores/24 threads)
  - Memory: 128GB (8 nodes)
  - SSD: 1.9TB x 2 (256GB/node)
  - Infiniband: ODR

Block Diagram of Thin Node

Details of GPU:

- **NVIDIA GPU Tesla M2050**
  - PCIe Express (Gen 2 x16)
  - Peak performance: 1.15 TFLOPS (double precision)
  - Number of CUDA cores: 448 cores

Details of SM (Streaming Multiprocessor):

- **CUDA cores x32**
  - 14 SMs
  - Memory bandwidth: 156.1GB/s (DDR3)
  - Memory clock: 1.3GHz (VDD)
  - ECC memory: support
  - Uniform Cache: 512KB
**High-Speed Network Interconnect**

Compute nodes of TSUBAME2.0 interconnected with Dual-Rail QDR InfiniBand networks of Fat-Tree type full bisection bandwidth achieve 200Gbps, enabling communication between the compute nodes is extremely low in microseconds-order time, therefore resulting in high-speed performance and high-speed connection to highly reliable storages. This network is linked by more than 8000 optical fiber cables in a total length of 100km.

![Diagram of network interconnect](image)

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

- 1 VSC Rack (Thin Node, 816)
- 15 nodes

**Medium Node x 24**

**Fat Node x 10**

- Edge Switch #1
- Edge Switch #2
- Edge Switch #3
- Edge Switch #4

**Inter-Node Connection Network**

Core switch: Voltaire GridDirector 41064x
Edge switch: Voltaire GridDirector 4000FX4, 6000FX4

**InfiniBand QDR Network for LNET and Other Services**

**GPFS with HSM**

- Storage: 2.4 PB + 4PB (Tape)
- Servers: HP Proliant DL580 G7
- FC: 36Gb/s
- Network: Gigabit Ethernet
- Storage: 14TB

**Lustre 3.6PB**

- Storage: 14TB
- Network: Gigabit Ethernet
- Storage: 30TB

**Home**

- Storage: 1PB
- Network: Gigabit Ethernet

**High-Speed and Highly Reliable Storage Systems**

TSUBAME2.0 provides 11PB of massive storage volumes to serve various purposes, including about 180TB of SSDs embedded in compute nodes for scratch I/O, 5PB 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

Cooling: Modular Cooling System

- Water group controller
- Power grid network controller
- Water group monitoring module
- Heat exchanger module
- Water group
- Water inlet
- Heat exchanger
- Water outlet

The rack-contained water-cooling system with a built-in heat exchanger is employed, allowing high-density cooling up to 5MW per rack (the world's top class) as being 10 times larger than what is used in typical data centers. Homogenous cooling air is provided through the role of the sensor with automatic opening/closing doors where a fan is unnecessary. Power consumption is minimized with a completely automatic temperature control to avoid heat removal from 96% to 67% by water cooling. Moreover, the air-conditioning doors contribute to a great reduction.

- Peak power consumption of air-conditioning equipment: 575 (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 that the performance boost is more than 30 times compared to TSUBAME 1.2, 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.

<table>
<thead>
<tr>
<th>OS</th>
<th>SUSE Linux Enterprise Server 11 SP1 Windows 2008 R2 SP1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch System</td>
<td>PBS Professional</td>
</tr>
</tbody>
</table>

ISV Application Software

<table>
<thead>
<tr>
<th>ANSYS Fluent, Winbench</th>
<th>Discovery Studio</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABAQUS</td>
<td>Business</td>
</tr>
<tr>
<td>ABAQUS CAE</td>
<td>Mathematics</td>
</tr>
<tr>
<td>MD Nastran</td>
<td>MATLAB</td>
</tr>
<tr>
<td>Patran</td>
<td>Maple</td>
</tr>
<tr>
<td>LS-DYNA</td>
<td>AVS/Express</td>
</tr>
<tr>
<td>Geostatian</td>
<td>AVS/Express PCE</td>
</tr>
<tr>
<td>Gauss View</td>
<td>Delight</td>
</tr>
<tr>
<td>AMBER</td>
<td>FGI Compiler</td>
</tr>
<tr>
<td>Milpro</td>
<td>Intel Compiler</td>
</tr>
<tr>
<td>Materials Studio</td>
<td>TotalView Debugger</td>
</tr>
</tbody>
</table>

Published by Global Scientific Information and Computing Center, Tokyo Institute of Technology

E-mail: tsubame@gsic.titech.ac.jp

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