TSUBAME3.0: A Green, Accelerated, Big-Data Supercomputer

Satoshi Matsuoka
Professor, GSIC, Tokyo Institute of Technology / Fellow, Artificial Intelligence Research Center, AIST, Japan / Director, AIST-Tokyo Tech. Big Data Open Innovation Lab

Presenter:
Toshio Endo
Associate Professor, GSIC, Tokyo Institute of Technology
Denver, Colorado, USA
TSUBAME2.0 Nov. 1, 2010
“The Greenest Production Supercomputer in the World”

- GPU-centric (> 4000) high performance & low power
- Small footprint (~200m² or 2000 sq.ft), low TCO
- High bandwidth memory, optical network, SSD storage...

**TSUBAME 2.0**
New Development

➤ ➤ 2000 Users
➤ ➤ 20% Industry Workloads
➤ A Variety of HPC, emerging BD/AI workloads

2013 GPU Upgrade
TSUBAME2.5
5.7 Petaflops

- > 2000 Users
- > 20% Industry Workloads
- A Variety of HPC, emerging BD/AI workloads

CPU (Westmere EP)
76.8 GFLOPS
32nm

GPU (Tesla M2050)
515 GFLOPS
3 GB

1.6 TFLOPS
55 GB/103 GB
>400GB/s Mem BW
80Gbps NW BW
~1KW max

6.7 TFLOPS
220 GB/412 GB
>1.6TB/s Mem BW

53.6 TFLOPS
1.7 TB/3.2 TB
>12TB/s Mem BW
35KW Max

4224 GPUs
80 TB
>600TB/s Mem BW
220Tbps NW
Bisection BW
1.4MW Max

Integrated by NEC Corporation
2017 Q2 TSUBAME3.0 Leading Machine Towards Exa & Big Data

1. “Everybody’s Supercomputer” - High Performance (12~24 DP Petaflops, 125~325TB/s Mem, 55~185Tbit/s NW), innovative high cost/performance packaging & design, in mere 180m²...

2. “Extreme Green” – ~10GFlops/W power-efficient architecture, system-wide power control, advanced cooling, future energy reservoir load leveling & energy recovery

3. “Big Data Convergence” – BYTES-Centric Architecture, Extreme high BW & capacity, deep memory hierarchy, extreme I/O acceleration, Big Data SW Stack for machine learning, graph processing, ...

4. “Cloud SC” – dynamic deployment, container-based node co-location & dynamic configuration, resource elasticity, assimilation of public clouds...

5. “Transparency” - full monitoring & user visibility of machine & job state, accountability via reproducibility

2006 TSUBAME1.0
80 Teraflops, #1 Asia #7 World
“Everybody’s Supercomputer”

2010 TSUBAME2.0
2.4 Petaflops #4 World
“Greenest Production SC”

2013 TSUBAME2.5
upgrade
5.7PF DFP /17.1PF SFP
20% power reduction

2017 TSUBAME3.0+2.5
~18PF(DFP) 4~5PB/s Mem BW
10GFlops/W power efficiency
Big Data & Cloud Convergence

2011 ACM Gordon Bell Prize

Large Scale Simulation
Big Data Analytics
Industrial Apps
Overview of TSUBAME3.0
BYTES-centric Architecture, Scalability to all 2160 GPUs, all nodes, the entire memory hierarchy

Full Operations
Aug. 2017

Full Bisection Bandwidth
Intel Omni-Path Interconnect. 4 ports/node
Full Bisection / 432 Terabits/s bidirectional
~x2 BW of entire Internet backbone traffic

DDN Storage
(Lustre FS 15.9PB+Home 45TB)

540 Compute Nodes SGI ICE XA + New Blade
Intel Xeon CPU x 2+NVIDIA Pascal GPUx4 (NV-Link)
256GB memory 2TB Intel NVMe SSD
12.1 Petaflops (DP), 47.2 AI-Petaflops
TSUBAME3.0 Co-Designed SGI ICE-XA Blade (new)
- No exterior cable mess (power, NW, water)
- Plan to become a future HPE product
Given a top-class supercomputer, how fast can we accelerate next generation big data c.f. Clouds?

Bring HPC rigor in architectural, algorithmic, and system software performance and modeling into big data.
Characteristics of Big Data and AI Computing

As BD / AI
Graph Analytics e.g. Social Networks
Sort, Hash, e.g. DB, log analysis
Symbolic Processing: Traditional AI

As HPC Task
Integer Ops & Sparse Matrices
Data Movement, Large Memory
Sparse and Random Data, Low Locality

Opposite ends of HPC computing spectrum, but HPC simulation apps can also be categorized likewise

Acceleration, Scaling

As BD / AI
Dense LA: DNN
Inference, Training, Generation

As HPC Task
Dense Matrices, Reduced Precision
Dense and well organized networks and Data

Acceleration, Scaling

Acceleration via Supercomputers adapted to AI/BD
TSUBAME3: A Massively BYTES Centric Architecture for Converged BD/AI and HPC

- **Intra-node GPU via NVLink**: 20~40GB/s
- **Inter-node GPU via OmniPath**: 12.5GB/s fully switched

**Memory**:
- HBM2 64GB 2.5TB/s
- DDR4 256GB 150GB/s
- Intel Optane 1.5TB 12GB/s (planned)
- NVMe Flash 2TB 3GB/s

**Storage (planned)**:
- 2TB 3GB/s

**Network**:
- Terabit class network/node 800Gbps (400+400) full bisection

**Scalability**:
- Any "Big" Data in the system can be moved to anywhere via RDMA speeds minimum 12.5GBytes/s also with Stream Processing
- Scalable to all 2160 GPUs, not just 8

**Hierarchical Memory**:
- ~2.3 Terabytes/node Hierarchical Memory for Big Data / AI (c.f. K-computer 16GB/node)

⇒ Over 1 Petabytes in TSUBAME3, Can be moved at 54 Terabyte/s or 1.7 Zetabytes / year
TSUBAME3: A Massively BYTES Centric Architecture for Converged BD/AI and HPC

- Intra-node GPU via NVLink: 20~40GB/s
- Inter-node GPU via OmniPath: 12.5GB/s fully switched

HBM2
64GB
2.5TB/s

DDR4
256GB
150GB/s

Intel Optane
1.5TB 12GB/s (planned)

NVMe Flash
2TB 3GB/s

Any “Big” Data in the system can be moved to anywhere via RDMA speeds minimum 12.5GBytes/s also with Stream Processing Scalable to all 2160 GPUs, not just 8

~2.3 Terabytes/node Hierarchical Memory for Big Data / AI (c.f. K-computer 16GB/node)

⇒ Over 2 Petabytes in TSUBAME3, Can be moved at 54 Terabyte/s or 1.7 Zetabytes / year
TSUBAME3.0 Compute Node SGI ICE-XA, a New GPU Compute Blade Co-Designed by SGI and Tokyo Tech GSIC

SGI ICE XA Infrastructure

Intel Omnipath Spine Switch, Full Bisection Fat Tre Network
432 Terabit/s Bidirectional for HPC and DNN

Ultra high performance & bandwidth “Fat Node”
• High Performance: 4 SX M2(NVLink) NVIDIA Pascal P100 GPU + 2 Intel Xeon 84 AI-TFLops
• High Network Bandwidth – Intel Omnipath 100GBps x 4 = 400Gbps (100Gbps per GPU)
• High I/O Bandwidth - Intel 2 TeraByte NVMe
  • > 1PB & 1.5~2TB/s system total
  • Future Octane 3D-Xpoint memory Petabyte or more directly accessible
• Ultra High Density, Hot Water Cooled Blades
  • 36 blades / rack = 144 GPU + 72 CPU, 50-60KW, x10 thermals c.f. IDC

400Gbps / node for HPC and DNN
Basic Requirements for AI Cloud System

**Application**
- Easy use of various ML/DL/Graph frameworks from Python, Jupyter Notebook, R, etc.
- Web-based applications and services provision

**System Software**
- HPC-oriented techniques for numerical libraries, BD Algorithm kernels, etc.
- Supporting long running jobs / workflow for DL
- Accelerated I/O and secure data access to large data sets
- User-customized environment based on Linux containers for easy deployment and reproducibility

**OS**

**Hardware**
- Modern supercomputing facilities based on commodity components
Oil Immersive Cooling + Hot Water Cooling + High Density Packaging + Fine-Grained Power Monitoring and Control, upgrade to /DL Oct. 2015

High Temperature Cooling
Oil Loop 35~45°C
⇒ Water Loop 25~35°C
(c.f. TSUBAME2: 7~17°C)

Cooling Tower:
Water 25~35°C
⇒ To Ambient Air

High Density Oil Immersion Upgrade 2015 for DNN Training
1.25 TB SSD/Node
168 NVIDIA K80 GPUs + Xeon
413+TFlops (DFP)
1.5PFlops (SFP)
~60KW/rack

Container Facility
20 feet container (16m²)
Fully Unmanned Operation
Warm Water Cooling Distribution in T3

Rooftop free cooling tower

- **1MB Cooling Capacity**
  - Outgoing 32 degrees C
  - Return 40 degrees C

On the ground chillers (shared with Tsubame2)

- **2MW Cooling Capacity**
  - Outgoing 17 degrees C
  - Return 24 degrees C

Compute Node
HPE SGI ICE XA

Backup Heat Exchanger

Storage Interconnect SW

In-Room Air-Con for Humans

- **100KW Max**
**Tsubame3 Highly Efficient Datacenter**

Machine PUE $\approx 1.03$ (~$1.1 \text{ w/storage}$)

- **Power, water, and cabling** are all above with ceiling support, for space efficiency and freedom of layout. IDC space $\approx 130 \text{ m}^2$

- Reinforced “Slab-Like” flat floor surface. Over $1 \text{t/m}^2$ floor load

- Year-round free “warm-water” cooling with cooling tower, PUE = 1.03, machine power $\approx$ facility power

- Over $100 \text{t}$ total floor load

- **420V Tri-Phase AC Power**

- **420V High Voltage** to minimize electrical distribution loss and cheaper cabling

- **Efficient Warm-Water Cooling**

- **Ultra High Density**

- **Low Electrical Distribution Loss**

- Piping and cabling hang from the ceiling
Liquid Cooled

“Hot Pluggable” ICE-
XA Blade

Smaller than 1U server,
no cables or pipes

100Gbps x 4
= 400Gbps

Xeon x 2
PCIe Switch
> 20 TeraFlops
DFP

256GByte Memory

PCle NVMe
Drive Bay x 4

Liquid Cooled NVMe
144 GPUs & 72 CPUs/rack
Integrated 100/200Gbps Fabric Backplane

15 Compute Racks
4 DDN Storage Racks
3 Peripheral & SW racks
Total 22 Racks
Running Benchmarks (HPL, HPCG)
Tokyo Tech + HPE/SGI Japan Team

Top 500 Submission = 1.998 Petaflops @ 144Nodes

Green 500 Submission = 14.11 Gflops/W
(Nov. 2016 #1 = 9.46 Gflops/W)

...Announcement at ISC17 Opening, on 19th June 2017