BDEC Japan update for Open High Performance Computing and Big Data / Artificial Intelligence Infrastructure

Satoshi Matsuoka
Professor, GSIC, Tokyo Institute of Technology / Director, AIST-Tokyo Tech. Big Data Open Innovation Lab / Fellow, Artificial Intelligence Research Center, AIST, Japan / Vis. Researcher, Advanced Institute for Computational Science, Riken

BDEC 2017
UPDATE:
Post K development

Yutaka Ishikawa
RIKEN AICS
An Overview of Post K

- CPU architecture
  - ARMv8-A + SVE + Fujitsu’s extension
  - FP64/FP32/FP16

- Completion of Functional design of system software and start of implementation

**Schedule**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Q2</td>
<td>Q3</td>
<td>Q4</td>
<td>Q1</td>
<td>Q2</td>
<td>Q3</td>
<td>Q4</td>
<td>Q1</td>
<td>Q2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FP64/FP32/FP16**

- McKernel is a light-weight kernel with Linux API.
  - New features, such as for manycore and deep memory hierarchy, can be implemented without modification of Linux
  - It runs on Intel Xeon and Xeon phi, and Fujitsu FX100 (SPARC)

- McKernel is running on the Oakforest-PACS supercomputer, 25 PF in peak, at JCAHPC organized by U. of Tsukuba and U. of Tokyo
Collaborations

- **DOE-MEXT**
  - Optimized Memory Management, Efficient MPI for exascale, Dynamic Execution Runtime, Storage Architectures, Metadata and active storage, Storage as a Service, Parallel I/O Libraries, MiniApps for Exascale CoDesign, Performance Models for Proxy Apps, OpenMP/XMP Runtime, Programming Models for Heterogeneity, LLVM for vectorization, Power Monitoring and Control, Power Steering, Resilience API, Shared Fault Data, etc.

- **CEA, France**
  - Programming Language
  - Runtime Environment
  - Energy-aware batch job scheduler
  - Large DFT calculations and QM/MM
  - Application of High Performance Computing to Earthquake Related Issues of Nuclear Power Plant Facilities
  - KPIs (Key Performance Indicators)

- **RIKEN AIP (Center for Advanced Intelligence Project)**
  - Massively parallel and distributed search, Machine Learning, etc.
<table>
<thead>
<tr>
<th>Peak Rank</th>
<th>Institution</th>
<th>System</th>
<th>Rpeak</th>
<th>Nov. 2016 Top500</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U-Tokyo/Tsukuba U JCAHP</td>
<td>Oakforest-PACS - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path</td>
<td>24.9</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>Tokyo Institute of Technology GSIC</td>
<td>TSUBAME 3.0 HPE/SGI ICE-XA custom NVIDIA Pascal P100 + Intel Xeon, Intel OmniPath</td>
<td>12.1</td>
<td>NA</td>
</tr>
<tr>
<td>3</td>
<td>Riken AICS</td>
<td>K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu</td>
<td>11.3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>Tokyo Institute of Technology GSIC</td>
<td>TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.93GHz, Infiniband QDR, NVIDIA K20x NEC/HPE</td>
<td>5.71</td>
<td>40</td>
</tr>
<tr>
<td>5</td>
<td>Kyoto University</td>
<td>Camphor 2 – Cray XC40 Intel Xeon Phi 68C 1.4Ghz</td>
<td>5.48</td>
<td>33</td>
</tr>
<tr>
<td>6</td>
<td>Japan Aerospace eXploration Agency</td>
<td>SORA-MA - Fujitsu PRIMEHPC FX100, SPARC64 XIfx 32C 1.98GHz, Tofu interconnect 2</td>
<td>3.48</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>Information Tech. Center, Nagoya U</td>
<td>Fujitsu PRIMEHPC FX100, SPARC64 XIfx 32C 2.2GHz, Tofu interconnect 2</td>
<td>3.24</td>
<td>35</td>
</tr>
<tr>
<td>8</td>
<td>National Inst. for Fusion Science(NIFS)</td>
<td>Plasma Simulator - Fujitsu PRIMEHPC FX100, SPARC64 XIfx 32C 1.98GHz, Tofu interconnect 2</td>
<td>2.62</td>
<td>48</td>
</tr>
<tr>
<td>9</td>
<td>Japan Atomic Energy Agency (JAEA)</td>
<td>SGI ICE X, Xeon E5-2680v3 12C 2.5GHz, Infiniband FDR</td>
<td>2.41</td>
<td>54</td>
</tr>
<tr>
<td>10</td>
<td>U-Tokyo- Inst. for Solid State Physics</td>
<td>Sekirei - SGI ICE XA, Xeon E5-2680v3 12C 2.5GHz, Infiniband FDR HPE/SGI</td>
<td>1.52</td>
<td>86</td>
</tr>
</tbody>
</table>
Established as Japanese integrated high performance computing infrastructure in 2011

Variety of computer systems are connected via high speed academic backbone network and provided as HPCI resources to users in Japan and overseas

FY2017 Allocated computing resources

K computer

~4 Pflops x Yr

Other systems in total

~5.6 Pflops x Yr

K computer

~4 Pflops x Yr

Other systems in total

~5.6 Pflops x Yr

HPCI : High Performance Computing Infrastructure

The Flagship System

HPCI

The Second Layer System

Other Systems in Universities , National Labs, etc.

K computer

~4 Pflops x Yr

Other systems in total

~5.6 Pflops x Yr

SINET-5

: 100Gbps

: 10Gbps

International line

*2 : Joint Center for Advanced High Performance Computing

*3 : The Institute of Statistical Mathematics

*4 : Japan Agency for Marine-Earth Science and Technology
### HPCI projects call results for FY 2017

#### Number of submitted & awarded proposals for FY 2017 regular call projects*

<table>
<thead>
<tr>
<th></th>
<th>Submitted*³</th>
<th>Awarded*³</th>
<th>Ratio*³</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>K computer</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General Use</td>
<td>51(53)</td>
<td>31(31)</td>
<td>61(58)%</td>
</tr>
<tr>
<td>Junior Researcher Promotion</td>
<td>16(21)</td>
<td>11(13)</td>
<td>69(62)%</td>
</tr>
<tr>
<td>Industrial (non-proprietary)</td>
<td>29(30)</td>
<td>25(28)</td>
<td>86(93)%</td>
</tr>
<tr>
<td>Total</td>
<td>96(104)</td>
<td>67(72)</td>
<td>70(69)%</td>
</tr>
<tr>
<td><strong>Other HPCI system</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General Use</td>
<td>141(128)</td>
<td>64(59)</td>
<td>45(46)%</td>
</tr>
<tr>
<td>Industrial (non-proprietary)</td>
<td>14(11)</td>
<td>5(10)</td>
<td>36(91)%</td>
</tr>
<tr>
<td>Total</td>
<td>155(139)</td>
<td>69(69)</td>
<td>45(50)%</td>
</tr>
</tbody>
</table>

*1 : Trial call projects are not included.
*2 : Results of “Term A” projects. “Term B” projects call will start from April.
*3 : Numbers in parentheses indicate those for FY 2016
*4 : Includes “concurrent use with K computer”
Research application areas of awarded projects

**K computer**
(project number based)

- **2017**
  - Mathematical science: 13%
  - Physics and space physics: 11%
  - Material science and chemistry: 28%
  - Engineering and manufacturing: 30%
  - Bio and life science: 2%
  - Environment, disaster prevention and mitigation: 4%
  - Information and computer science: 13%
  - Nuclear and fusion engineering: 17%
  - Others: 1%

- **2016**
  - Mathematical science: 13%
  - Physics and space physics: 15%
  - Material science and chemistry: 29%
  - Engineering and manufacturing: 27%
  - Bio and life science: 1%
  - Environment, disaster prevention and mitigation: 4%
  - Information and computer science: 13%
  - Nuclear and fusion engineering: 19%
  - Others: 2%

**Other HPCI system**
(project number based)

- **2017**
  - Mathematical science: 4%
  - Physics and space physics: 3%
  - Material science and chemistry: 6%
  - Engineering and manufacturing: 1%
  - Bio and life science: 1%
  - Environment, disaster prevention and mitigation: 3%
  - Information and computer science: 26%
  - Nuclear and fusion engineering: 19%
  - Others: 2%

- **2016**
  - Mathematical science: 3%
  - Physics and space physics: 1%
  - Material science and chemistry: 14%
  - Engineering and manufacturing: 28%
  - Bio and life science: 14%
  - Environment, disaster prevention and mitigation: 22%
  - Information and computer science: 26%
  - Nuclear and fusion engineering: 19%
  - Others: 28%

**Total**

**General Use / Junior Researcher Promotion**

**Industrial (non-proprietary)**

Copyright 2017 RIST
U-Tokyo/Tsukuba-U JCAHPC “Oakforest-PACS”
24.9 Petaflops KNL/OmniPath

Chassis with 8 nodes,
2U size

Computation node (Fujitsu next generation PRIMERGY)
with single chip Intel Xeon Phi (Knights Landing, 3+TFLOPS)
and Intel Omni-Path Architecture card (100Gbps)
## Specification of Oakforest-PACS system

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value/Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total peak performance</td>
<td>25 PFLOPS</td>
</tr>
<tr>
<td>Total number of compute nodes</td>
<td>8,208</td>
</tr>
<tr>
<td>Compute node</td>
<td></td>
</tr>
<tr>
<td>Product</td>
<td>Fujitsu Next-generation PRIMERGY server for HPC (under development)</td>
</tr>
<tr>
<td>Processor</td>
<td>Next-generation of Intel® Xeon Phi™ (Code name: Knights Landing), &gt;60 cores</td>
</tr>
<tr>
<td>Memory</td>
<td></td>
</tr>
<tr>
<td>High BW</td>
<td>16 GB, &gt; 400 GB/sec (MCDRAM, effective rate)</td>
</tr>
<tr>
<td>Low BW</td>
<td>96 GB, 115.2 GB/sec (DDR4-2400 x 6ch, peak rate)</td>
</tr>
<tr>
<td>Interconnect</td>
<td></td>
</tr>
<tr>
<td>Product</td>
<td>Intel® Omni-Path Architecture</td>
</tr>
<tr>
<td>Link speed</td>
<td>100 Gbps</td>
</tr>
<tr>
<td>Topology</td>
<td>Fat-tree with (completely) full-bisection bandwidth</td>
</tr>
<tr>
<td>Login node</td>
<td></td>
</tr>
<tr>
<td>Product</td>
<td>Fujitsu PRIMERGY RX2530 M2 server</td>
</tr>
<tr>
<td># of servers</td>
<td>20</td>
</tr>
<tr>
<td>Processor</td>
<td>Intel Xeon E5-2690v4 (2.6 GHz 14 core x 2 socket)</td>
</tr>
<tr>
<td>Memory</td>
<td>256 GB, 153 GB/sec (DDR4-2400 x 4ch x 2 socket)</td>
</tr>
</tbody>
</table>
## Specification of Oakforest-PACS system (I/O)

<table>
<thead>
<tr>
<th>Parallel File System</th>
<th>Type</th>
<th>Lustre File System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Capacity</td>
<td></td>
<td>26.2 PB</td>
</tr>
<tr>
<td>Metadata</td>
<td>Product</td>
<td>DataDirect Networks MDS server + SFA7700X</td>
</tr>
<tr>
<td># of MDS</td>
<td>4 servers x 3 set</td>
<td></td>
</tr>
<tr>
<td>MDT</td>
<td>7.7 TB (SAS SSD) x 3 set</td>
<td></td>
</tr>
<tr>
<td>Object storage</td>
<td>Product</td>
<td>DataDirect Networks SFA14KE</td>
</tr>
<tr>
<td># of OSS (Nodes)</td>
<td>10 (20)</td>
<td></td>
</tr>
<tr>
<td>Aggregate BW</td>
<td>500 GB/sec</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fast File Cache System</th>
<th>Type</th>
<th>Burst Buffer, Infinite Memory Engine (by DDN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total capacity</td>
<td>940 TB (NVMe SSD, including parity data by erasure coding)</td>
<td></td>
</tr>
<tr>
<td>Product</td>
<td>DataDirect Networks IME14K</td>
<td></td>
</tr>
<tr>
<td># of servers (Nodes)</td>
<td>25 (50)</td>
<td></td>
</tr>
<tr>
<td>Aggregate BW</td>
<td>1,560 GB/sec</td>
<td></td>
</tr>
</tbody>
</table>
K computer “Still the best” for Bandwidth (Data-centric) workloads (It’s the Bandwidth!)

1. TOP500 List
   - 2011: 2
   - 2012: 4
   - 2013: 4
   - 2014: 4
   - 2015: 7

2. Gordon Bell Prize
   - Finalist

3. HPC Challenge Awards
   - (HPC, Random Access, STREAM, FFT)
   - 2011: 4
   - 2012: 2

4. Graph500
   - 2011: 4
   - 2012: 2
The Graph500 – 2015~2016 – 4 Consecutive world #1
K Computer #1 Tokyo Tech[EBD CREST] Univ. Kyushu [Fujisawa Graph CREST], Riken AICS, Fujitsu

<table>
<thead>
<tr>
<th>List</th>
<th>Rank</th>
<th>GTEPS</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>November 2013</td>
<td>4</td>
<td>5524.12</td>
<td>Top-down only</td>
</tr>
<tr>
<td>June 2014</td>
<td>1</td>
<td>17977.05</td>
<td>Efficient hybrid</td>
</tr>
<tr>
<td>November 2014</td>
<td>2</td>
<td></td>
<td>Efficient hybrid</td>
</tr>
<tr>
<td>June, Nov 2015</td>
<td>1</td>
<td>38621.4</td>
<td>Hybrid + Node Compression</td>
</tr>
</tbody>
</table>

88,000 nodes, 660,000 CPU Cores, 1.3 Petabyte mem, 20GB/s Tofu NW

88,000 nodes, 660,000 CPU Cores, 1.3 Petabyte mem

LLNL-IBM Sequoia 1.6 million CPUs 1.6 Petabyte mem

TaihuLight 10 million CPUs 1.3 Petabyte mem

*Problem size is weak scaling “Brain-class” graph

73% total exec time wait in communication

Effective x13 performance c.f. Linpack

64 nodes (Scale 30)
65536 nodes (Scale 40)

Elapsed Time (ms)

Communi…

TaihuLight
88,000 nodes, 660,000 CPU Cores, 1.3 Petabyte mem, 20GB/s Tofu NW

Effective x13 performance c.f. Linpack

LLNL-IBM Sequoia 1.6 million CPUs 1.6 Petabyte mem

TaihuLight 10 million CPUs 1.3 Petabyte mem

*Problem size is weak scaling “Brain-class” graph

73% total exec time wait in communication
Two Big Data CREST Programs (2013-2020) ~$60 mil

Advanced Core Technologies for Big Data Integration

Research Supervisor: Masaru Kitsuregawa
Director General, National Institute of Informatics

Advanced Application Technologies to Boost Big Data Utilization for Multiple-Field Scientific Discovery and Social Problem Solving

Research Supervisor: Yuzuru Tanaka
Professor, Graduate School of Information Science and Technology, Hokkaido University
Tremendous Recent Rise in Interest by the Japanese Government on Big Data, DL, AI, and IoT

• Three national centers on Big Data and AI launched by three competing Ministries for FY 2016 (Apr 2015-)
  – METI – AIRC (Artificial Intelligence Research Center): AIST (AIST internal budget + > $200 million FY 2017), April 2015
    • Broad AI/BD/IoT, industry focus
  – MEXT – AIP (Artificial Intelligence Platform): Riken and other institutions ($~50 mil), April 2016
    • A separate Post-K related AI funding as well.
    • Narrowly focused on DNN
  – MOST – Universal Communication Lab: NICT ($50~55 mil)
    • Brain –related AI
  – $1 billion commitment on inter-ministry AI research over 10 years
Matsuoka: Joint appointment as “Designated” Fellow since July 2017
Two AI CREST Programs (under AIP, MEXT) (2016-2023) ~$40 mil x 2

Intelligent Information Processing Systems Creating Co-Experience Knowledge and Wisdom with Human-Machine Harmonious Collaboration

Research Supervisor: Norihiro Hagita (Board Director, Director, Intelligent Robotics and Communication Laboratories, Advanced Telecommunications Research Institute International)

Development and Integration of Artificial Intelligence Technologies for Innovation Acceleration

Research Supervisor: Minoru Etoh (Senior Vice President, General Manager of Innovation Management Department, NTT DOCOMO, INC.)
Estimated Compute Resource Requirements for Deep Learning
[Source: Preferred Network Japan Inc.]

To complete the learning phase in one day

<table>
<thead>
<tr>
<th>Image/Video Recognition</th>
<th>Auto Driving</th>
<th>Robots / Drones</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>10P (Image) ~ 10E (Video)</strong> Flops</td>
<td><strong>100P ~ 1E Flops</strong></td>
<td><strong>1E ~ 100E Flops</strong></td>
</tr>
<tr>
<td>学習データ: 1億枚の画像 10000クラス分類 数千ノードで6ヶ月 [Google 2015]</td>
<td>一人あたりゲノム解析で約10M個のSNPs 100万人で100PFlops、1億人で1EFlops</td>
<td>1台あたり年間1TB 100万台〜1億台から得られたデータで学習する場合</td>
</tr>
<tr>
<td><strong>10P ~ Flops</strong></td>
<td><strong>1E ~ 100E Flops</strong></td>
<td><strong>1E ~ 100E Flops</strong></td>
</tr>
<tr>
<td>1万人の5000時間分の音声データ 人工的に生成された10万時間の音声データを基に学習 [Baidu 2015]</td>
<td>自動運転車1台あたり1日 1TB 10台〜1000台、100日分の走行データの学習</td>
<td>1台あたり年間1TB 100万台〜1億台から得られたデータで学習する場合</td>
</tr>
</tbody>
</table>

機械学習、深層学習は学習データが大きいほど高精度になる 現在は人が生み出したデータが対象だが、今後は機械が生み出すデータが対象となる

各推定値は1GBの学習データに対して1日で学習するためには 1TFlops必要だと計算

<table>
<thead>
<tr>
<th>2015</th>
<th>2020</th>
<th>2025</th>
<th>2030</th>
</tr>
</thead>
<tbody>
<tr>
<td>10PF</td>
<td>100PF</td>
<td>1EF</td>
<td>10EF</td>
</tr>
</tbody>
</table>

It’s the FLOPS too! (in reduced precision)
So both are important in the infrastructure
The current status of AI & Big Data in Japan
We need the triage of **algorithms/infrastructure/data** but we lack the **infrastructure** dedicated to AI & Big Data (c.f. Google)
The current status of AI & Big Data in Japan

We need the triage of algorithms/infrastructure/data but we lack the infrastructure dedicated to AI & Big Data (c.f. Google)

Application-based Solution providers of ML (e.g. Pharma, Semiconductors)
Custom ML/DL Software

Analysis of automotive cameras
Performance analysis & improvement of DL

Use of Massive Scale Data now Wasted

Petabytes of Drive
Recording Video

Web access and merchandice

AI&Data Processing Infrastructures

Investigating the Application of

Use of Massive Scale Data

FA&Robots

Web access and merchandice

IoT Communication, location & other data

Chainer OSS DL Framework
Many applications in manufacturing, web, pharma, etc.

Machine Learning Algorithms

Preferred Networks

"Chainer" OSS DL Framework
Many applications in manufacturing, web, pharma, etc.

Acceleration & Scaling of DL & other ML Algorithms & SW

Fujitsu

"Chainer" OSS DL Framework
Many applications in manufacturing, web, pharma, etc.
The current status of AI & Big Data in Japan

We need the triage of **algorithms/infrastructure/data** but we lack the **infrastructure** dedicated to AI & Big Data (c.f. Google)

**Insufficient to Counter the Giants**

- (Google, Microsoft, Baidu etc.)

in their own game

**AI&Data Infrastructures**

**Massive “Big” Data in Training**

**Use of Massive Scale Data now Wasted**

- Denso
- Toyota
- JARI
- FANUC
- Denso ITLAB
- MIZUHO
- DeNA
- ABEJA
- FA&Robots
- SoftBank
- NTT

**Application-based Solution providers of ML (e.g. Pharma, Semiconductors)**

- Custom ML/DL Software

**Massive Rise in Computing Requirements**

- Windows Azure
- Intel Cloud
- SAKURA Internet
- Amazon Web Services

**Preferred Networks**

- "Chainer" OSS DL Framework
- Many applications in manufacturing, web, pharma, etc.

**Investigating the Application of**

- Petabytes of Drive
- Recording Video
- Web access and merchandice

**Machine Learning Algorithms**

- Analy 車載カメラ映像解析
- Perf AI&Data 高性能化高速化に関する基礎研究

**Use of Massive Scale Data location & other data**
The “Chicken or Egg Problem” of AI-HPC Infrastructures

• “On Premise” machines in clients => “Can’t invest in big in AI machines unless we forecast good ROI. We don’t have the experience in running on big machines.”

• Public Clouds other than the giants => “Can’t invest big in AI machines unless we forecast good ROI. We are cutthroat.”

• Large scale supercomputer centers => “Can’t invest big in AI machines unless we forecast good ROI. Can’t sacrifice our existing clients and our machines are full”

• Thus the giants dominate, AI technologies, big data, and people stay behind the corporate firewalls…
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 AI – HPC Convergence” – Extreme high BW & FLOPS, deep memory hierarchy, extreme I/O acceleration, 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

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

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

2013 TSUBAME-KFC
#1 Green 500

2016 TSUBAME3.0+2.5
~18PF(DFP) 3~4PB/s Mem BW
10GFlops/W power efficiency
Big Data & Cloud Convergence

Large Scale Simulation
Big Data Analytics
Industrial Apps
Overview of TSUBAME3.0

- 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
- 47.2 AI-Petaflops, 12.1 Petaflops

Full Operations
Aug. 2017
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 Tree Network 432 Terabit/s Bidirectional

X60 Pairs (Total 120 Switches)

High performance “Fat Node”
- High Performance 4 SXM2 (NVLink) NVIDIA Pascal P100 GPU + Xeon
- High Network Bandwidth – Intel Omnipath 100GBps x 4 = 400GBps
- High I/O Bandwidth - Intel 2 TeraByte NVMe
  - > 1PB & 1.5~2TB/s system total
- Ultra High Density, Hot Water Cooled Blades
  - 36 blades / rack = 144 GPU + 72 CPU, 50-60KW, x10 thermals c.f. IDC


Tokyo Tech GSIC leads Japan in aggregated AI-capable FLOPS TSUBAME3+2.5+KFC, in all Supercomputers and Clouds NV.

Site Comparisons of AI-FP Perfs

Tokyo Tech
- TSUBAME3.0
- T2.5
- ~6700 GPUs + ~4000 CPUs

U-Tokyo
- Oakforest-PACS (JCAHPC)
- Reedbush (U&H)

Riken
- K

NVIDIA Pascal
- P100 DGEMM Performance

Matrix Dimension (m=n=k)

GFLOPS

0 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000 11000 12000 13000 14000 15000 16000

0 500 1000 1500 2000 2500 3000 3500 4000 4500

DFP 64bit | SFP 32bit | HFP 16bit

Simulation

Computer Graphics

Gaming

Big Data

Machine Learning / AI
TSUBAME3.0 SGI ICE-XA Blade (new) - Plan to become a future HPE product
TSUBAME3.0 Datacenter

15 SGI ICE-XA Racks
2 Network Racks
3 DDN Storage Racks
20 Total Racks

Compute racks cooled with 32 degrees warm water, yearround ambient cooling
PUE = 1.033
AI R&D Investments in METI

<table>
<thead>
<tr>
<th>FY2015</th>
<th>FY2016</th>
<th>FY2017</th>
<th>FY2018</th>
<th>FY2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foundation of AIRC @AIST</td>
<td>9M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceleration of AI R&amp;D (FY 15 Supplementary Budget)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAIC, AIST AI Cloud →400x Tesla P100, Spark-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global Open Innovation Arena for AI R&amp;D (FY16 Supplementary Budget)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• ABCI, AI-Bridging Cloud Infrastructure →130PFLOPS(AI), PUE &lt; 1.1, &lt; 3MW</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Demonstration env. for Robotics/Industry 4.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• R&amp;D “base” for AI-accelerated Nanofabrication and Medical technologies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ABCI Prototype: AIST AI Cloud (AAIC)  
March 2017 (System Vendor: NEC)

- 400x NVIDIA Tesla P100s and Infiniband EDR accelerate various AI workloads including ML (Machine Learning) and DL (Deep Learning).
- Advanced data analytics leveraged by 4PiB shared Big Data Storage and Apache Spark w/ its ecosystem.
METI AIST-AIRC ABCI as the \textit{worlds first large-scale OPEN AI Infrastructure}

- **ABCI**: AI Bridging Cloud Infrastructure
  - Top-Level SC compute & data capability (130~200 AI-Petaflops)
  - Open Public & Dedicated infrastructure for AI & Big Data Algorithms, Software and Applications
  - Platform to accelerate joint academic-industry R&D for AI in Japan

- 130~200 AI-Petaflops
  - < 3MW Power
  - < 1.1 Avg. PUE
  - Operational 2017Q3~Q4

Univ. Tokyo Kashiwa Campus
ABCI – 2017Q4~ 2018Q1

- Extreme computing power
  - w/ 130~200 AI-PFlops for AI, ML, DL
  - x1 million speedup over high-end PC: 1 Day training for 3000-Year DNN training job
  - TSUBAME-KFC (1.4 AI-Pflops) x 90 users (T2 avg)

- Big Data and HPC converged modern design
  - For advanced data analytics (Big Data) and scientific simulation (HPC), etc.
  - Leverage Tokyo Tech’s “TSUBAME3” design, but differences/enhancements being AI/BD centric

- Ultra high bandwidth and low latency in memory, network, and storage
  - For accelerating various AI/BD workloads
  - Data-centric architecture, optimizes data movement

- Big Data/AI and HPC SW Stack Convergence
  - Incl. results from JST-CREST EBD
  - Wide contributions from the PC Cluster community desirable.

- RFC just out, includes 10 BD/ML benchmarks
  - No HPC benchmarks
ABCI Cloud Infrastructure

- **Ultra-dense IDC design from ground-up**
  - Custom inexpensive lightweight “warehouse” building w/ substantial earthquake tolerance
  - **x20 thermal density of standard IDC**

- **Extreme green**
  - Ambient warm liquid cooling, large Li-ion battery storage, and high-efficiency power supplies, etc.
  - **Commoditizing supercomputer cooling technologies to Clouds (60KW/rack)**

- **Cloud ecosystem**
  - Wide-ranging Big Data and HPC standard software stacks

- **Advanced cloud-based operation**
  - Incl. dynamic deployment, container-based virtualized provisioning, multitenant partitioning, and automatic failure recovery, etc.
  - Joining HPC and Cloud Software stack for real
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AI-FLOPS Peak AI Performance</strong></td>
<td>47.2 Pflops (DFP 12.1 Pflops) 3.1 PetaFlops/rack</td>
<td>130<del>200 Pflops, (DFP NA) 3</del>4 PetaFlops/rack</td>
<td>11.3 Petaflops 12.3 Tflops/rack</td>
</tr>
<tr>
<td><strong>System Packaging</strong></td>
<td>Custom SC (ICE-XA), Liquid Cool</td>
<td>19 inch rack (LC), ABCI-IDC</td>
<td>Custom SC (LC)</td>
</tr>
<tr>
<td><strong>Operational Power incl. Cooling</strong></td>
<td>Below 1MW</td>
<td>Approx. 2MW</td>
<td>Over 15MW</td>
</tr>
<tr>
<td><strong>Max Rack Thermals &amp; PUE</strong></td>
<td>61KW, 1.033</td>
<td>50-60KW, below 1.1</td>
<td>~20KW, ~1.3</td>
</tr>
<tr>
<td><strong>Node Hardware Architecture</strong></td>
<td>Many-Core (NVIDIA Pascal P100) + Multi-Core (Intel Xeon)</td>
<td>Many-Core AI/DL oriented processor (incl. GPUs)</td>
<td>Heavyweight Multi-Core</td>
</tr>
<tr>
<td><strong>Memory Technology</strong></td>
<td>HBM2+DDR4</td>
<td>On Die Memory + DDR4</td>
<td>DDR3</td>
</tr>
<tr>
<td><strong>Network Technology</strong></td>
<td>Intel OmniPath, 4 x 100Gbps / node, full bisection, optical NW</td>
<td>Injection/bisection scaled down c.f. to save cost &amp; IDC friendly</td>
<td>Copper Tofu 6-D torus custom NW</td>
</tr>
<tr>
<td><strong>Per-node non volatile memory</strong></td>
<td>2TeraByte NVMe/node</td>
<td>&gt; 400GB NVMe/node</td>
<td>None</td>
</tr>
<tr>
<td><strong>Power monitoring and control</strong></td>
<td>Detailed node / whole system power monitoring &amp; control</td>
<td>Detailed node / whole system power monitoring &amp; control</td>
<td>Whole system monitoring only</td>
</tr>
<tr>
<td><strong>Cloud and Virtualization, AI</strong></td>
<td>All nodes container virtualization, horizontal node splits, Cloud API dynamic provisioning, ML Stack</td>
<td>All nodes container virtualization, horizontal node splits, Cloud API dynamic provisioning, ML Stack</td>
<td>None</td>
</tr>
<tr>
<td><strong>Procurement Benchmarks</strong></td>
<td>HPC-Oriented Benchmarks</td>
<td>BD &amp; DNN Benchmarks</td>
<td>HPC Benchmarks</td>
</tr>
</tbody>
</table>
Fujitsu Deep Learning Processor (DLU™)

**DLU™ features**

- Architecture designed for Deep Learning
- High performance HBM2 memory
- Low power design
  - Goal: 10x Performance/Watt compared to others
- Massively parallel: Apply supercomputer interconnect technology
  - Ability to handle large scale neural networks
  - TOFU Network derivative for massive scaling

“Exascale” AI possible in 1H2019
Software Ecosystem for HPC in AI
Different SW Ecosystem between HPC and AI/BD/Cloud
How to achieve convergence—for real, for rapid tech transfer

Existing Clouds
- BD/AI User Applications
  - Machine Learning (ML), Mahout/Chainer
  - MapReduce Framework (Spark/Hadoop)
  - Distributed Filesystem (HDFS & Object Store)
  - VM(KVM), Container(Docker), Cloud Services (OpenStack)
  - Linux OS

Application Layer
- Cloud Jobs often Interactive w/resource control REST APIs
- HPC Jobs are Batch-Oriented, resource control by MPI

Existing Supercomputers
- HPC User Code
  - Numerical Libraries (LAPACK, FFTW)
  - Various DSLs
  - Workflow Systems
  - Fortran · C · C++ + IDL
  - MPI · OpenMP/ACC · CUDA/OpenCL
  - Parallel Debuggers and Profilers
  - Parallel Filesystem (Lustre, GPFS,)
  - Batch Job Schedulers (PBS Pro, Slurm, UGE)

System Software Layer
- Cloud employs High Productivity Languages but performance neglected, focus on data analytics and dynamic frequent changes
- HPC employs High Performance Languages but requires Ninja Programmers, low productivity. Kernels & compilers well tuned & result shared by many programs, less rewrite
- Cloud focused on databases and data manipulation workflow
- HPC focused on compute kernels, even for data processing. Jobs scales to thousands of jobs, thus debugging and performance tuning
- Cloud requires purpose-specific computing/data environment as well as their mutual isolation & security
- HPC requires environment for fast & lean use of resources, but on modern machines require considerable system software support

Hardware Layer
- Cloud HW based on Web Server “commodity” x86 servers, distributed storage on nodes assuming REST API access
- HPC HW aggressively adopts new technologies such as GPUs, focused on ultimate performance at higher cost, shared storage to support legacy apps

InfiniBand/OPA
High Capacity
Low Latency NW

Various convergence research efforts underway but no realistic converged SW Stack yet => achieving HPC – AI/BD/Cloud convergence key ABCI goal
National Institute for Advanced Industrial Science and Technology (AIST)

Director: Satoshi Matsuoka

Ministry of Economics Trade and Industry (METI)

AIST Artificial Intelligence Research Center (AIRC)

Application Area
Natural Language Processing
Robotics
Security

Matsuoka appointed 15% to AIST AI-OIL

Joint Research on AI / Big Data and applications

AIST-TokyoTech AI/Big Data Open Innovation Laboratory (OIL)
“Real World Big Data”

Industrial Collaboration in data, applications

Resources and Acceleration of AI / Big Data, systems research

Tsubame 3.0/2.5
Big Data / AI resources

School of Informatics

Other Big Data / AI research organizations and proposals

Joint Organization@Odaiba

Industry

Basic Research in Big Data / AI algorithms and methodologies

Director: Satoshi Matsuoka
We are implementing the US AI&BD strategies already...

in Japan, at AIRC w/ABCI

• Strategy 5: Develop shared public datasets and environments for AI training and testing. The depth, quality, and accuracy of training datasets and resources significantly affect AI performance. Researchers need to develop high quality datasets and environments and enable responsible access to high-quality datasets as well as to testing and training resources.

• Strategy 6: Measure and evaluate AI technologies through standards and benchmarks. Essential to advancements in AI are standards, benchmarks, testbeds, and community engagement that guide and evaluate progress in AI. Additional research is needed to develop a broad spectrum of evaluative techniques.
Co-Design of BD/ML/AI with HPC using BD/ML/AI
- for survival of HPC

Big Data AI-Oriented Supercomputers

Accelerating Conventional HPC Apps

Optimizing System Software and Ops

Future Big Data•AI Supercomputer Design

Accelerating BD/ML/AI via HPC and Technologies and Infrastructures

Mutual and Semi-Automated Co-Acceleration of HPC and BD/ML/AI

Big Data and ML/AI Apps and Methodologies

Acceleration Scaling, and Control of HPC via BD/ML/AI and future SC designs

ABCI: World’s first and largest open 100 Peta AI-Flops AI Supercomputer, Fall 2017, for co-design

Robots / Drones

Image and Video

Large Scale Graphs
But Commercial Companies esp. the “AI Giants” are Leading AI R&D, are they not?

- Yes, but that is because their shot-term goals could harvest the low hanging fruits in DNN rejuvenated AI
- But AI/BD research is just beginning --- if we leave it to the interests of commercial companies, we cannot tackle difficult problems with no proven ROI
  - Very unhealthy for research
- This is different from more mature fields, such as pharmaceuticals or aerospace, where there is balanced investments and innovations in both academia/government and the indu