Overview of Supercomputer Systems

Supercomputing Division
Information Technology Center
The University of Tokyo
### Supercomputers at ITC, U. of Tokyo

#### (retired, March 2014)

<table>
<thead>
<tr>
<th>System</th>
<th>Model</th>
<th>Peak Performance</th>
<th>Nodes</th>
<th>Memory</th>
<th>Performance/Node</th>
<th>Memory/Node</th>
<th>Disk Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oakleaf-fx</td>
<td>(Fujitsu PRIMEHPC FX10)</td>
<td>1.13 PFLOPS</td>
<td>4800</td>
<td>32 GB, 1.1 PB + 2.1 PB</td>
<td>236.5 GFLOPS</td>
<td>32 GB</td>
<td>1 PB</td>
</tr>
<tr>
<td>T2K-Todai</td>
<td>(Hitachi HA8000-tc/RS425)</td>
<td>140 TFLOPS</td>
<td>952</td>
<td>200 GB, 128 GB</td>
<td>147.2 GFLOPS</td>
<td>32 GB</td>
<td>1 PB</td>
</tr>
<tr>
<td>Yayoi</td>
<td>(Hitachi SR16000/M1)</td>
<td>54.9 TFLOPS</td>
<td>56</td>
<td>11200 GB</td>
<td>980.48 GFLOPS</td>
<td>200 GB</td>
<td>556 TB</td>
</tr>
</tbody>
</table>

“Oakbridge-fx” with 576 nodes installed in April 2014 (separated) (136TF)

Total Users > 2,000
• HPCI
• Supercomputer Systems in SCD/ITC/UT
• Overview of Fujitsu FX10 (Oakleaf-FX)
• Post T2K System
Innovative High Performance Computing Infrastructure (HPCI)

• HPCI
  – Seamless access to K computer, supercomputers, and user's machines
  – Distributed shared storage system

• HPCI Consortium
  – Providing proposals/suggestions to the government and related organizations
    • Plan and operation of HPCI system
    • Promotion of computational sciences
    • Future supercomputing
  – 38 organizations
  – Operations started in Fall 2012
    • https://www.hpci-office.jp/
SPIRE/HPCI
Strategic Programs for Innovative Research

• Objectives
  – Scientific results as soon as K computer starts its operation
  – Establishment of several core institutes for comp. science

• Overview
  – Selection of the five strategic research fields which will contribute to finding solutions to scientific and social Issues
    • Field 1: Life science/Drug manufacture
    • Field 2: New material/energy creation
    • Field 3: Global change prediction for disaster prevention/mitigation
    • Field 4: Mono-zukuri (Manufacturing technology)
    • Field 5: The origin of matters and the universe
  – A nation wide research group is formed by centering the core organization of each research area designated by MEXT.
  – The groups are to promote R&D using K computer and to construct research structures for their own area
HPCI戦略プログラム
Strategic Programs for Innovative Research

予測医療と革新的創薬
臓器レベルでの疾患を再現する階層統合シミュレーションを実現し、予測医療に貢献。また、標的タンパク質に強く結合する薬の候補化合物の設計を行い、創薬プロセスを加速。

世界に先駆けて次世代デバイスを提案
ナノスケールデバイスをもとにしてシミュレーションし、機能・材料特性予測を実現することで、次世代デバイスの設計手法を提案。超高性能・超低消費電力端末等の実現に貢献する。

次世代ものづくり
設計プロセスの革新
流動の物理メカニズム理解に基づいた高度な設計制御技術を確立することで、環境（CO2, NOx）と製品性能のバランスを目指した将来の製品競争力強化に資する革新のものづくりを実現

大質量星の超新星爆発の解明
非平衡座標系に対する動的解析により超新星爆発を高精度で模擬シミュレーションを実行し、大質量星が重力崩壊から超新星爆発に至る過程を解明する。

集中豪雨や地震の予測
雲解像モデル、地震動モデル等を駆使して、集中豪雨の位置や強度の推定を高精度に予測し、防災対策に資する。

物質・エネルギー創成
原子レベルでの精密計算による量子・核動力
次世代ナノスケールデバイス
集中豪雨や地震の予測
地震波伝播計算と破壊力学
大規模シミュレーション

防災・減災に資する
地球変動予測
November 2013

AICS, RIKEN:
K computer (11.28 PF, 1.27 PiB)

Univ. of Tokyo:
FX10 (300.8 TF, 59 TB)
GreenBlade8000 (242.5 TF, 38 TB)
2548x (10.6 TF, 24 TB)

SR16000/M1 (54.9 TF, 10.94 TB)
T2K (75.36 TF, 16 TB / 140 TF, 31.25 TB)
EastHubPCCluster (10 TF, 5.71 TB / 13 TF, 8.1 TB)

GPU Cluster (CPU 4.5 TF, GPU 16.48 TF, 1.5 TB)
WestHubPCCluster (12.37 TF, 8.25 TB)
RENKEI-VPE: VM Hosting

Kyushu Univ.:
FX10 (181.6 TF, 24 TB)
CX400 (811.9 TF, 184.5 TB)
HA8000-tc (712.5 TF, 24.7 TB)
SR16000 VM1 (8.19 TF, 16 TB)

Hokkaido Univ.:
SR16000/M1 (172 TF, 22 TB)
BS2000 (44 TF, 14 TB)

Tohoku Univ.:
SX9 (29.4 TF, 18 TB)
Express5800 (1.74 TF, 3 TB)

Univ. of Tsukuba:
T2K (95.4 Tiflops, 20 TB)
HA-PACS (802 Tiflops, 34.3 TB)
FIRST (36.1 Tiflops, 1.6 TB)

Osaka Univ.:
SX9 (16 TF, 10 TB)
SX8R (5.3 TF, 3.3 TB)
PCCluster (22.7 TF, 4.6 TB)

Nagoya Univ.:
FX10 (90.8 TF, 12 TB)
CX400 (471 TF, 43 TB)

Tohoku Univ.:
SX9 (29.4 TF, 18 TB)
Express5800 (1.74 TF, 3 TB)

Univ. of Tokyo:
FX10 (1.13 PF, 150 TB)
SR16000/M1 (54.9 TF, 10.94 TB)
T2K (75.36 TF, 16 TB / 140 TF, 31.25 TB)
EastHubPCCluster (10 TF, 5.71 TB / 13 TF, 8.1 TB)
GPU Cluster (CPU 4.5 TF, GPU 16.48 TF, 1.5 TB)
WestHubPCCluster (12.37 TF, 8.25 TB)
RENKEI-VPE: VM Hosting

Kyoto Univ.:
XE6 (300.8 TF, 59 TB)
GreenBlade8000 (242.5 TF, 38 TB)

Tokyo Institute of Technology:
TSUBAME2.5 (5.79 PF, 150 TB)
• HPCI
• **Supercomputer Systems in SCD/ITC/UT**
• Overview of Fujitsu FX10 (Oakleaf-FX)

• Post T2K System
Current Supercomputer Systems
University of Tokyo

• Total number of users ~ 2,000 (50% from outside of UT)
• Hitachi HA8000 Cluster System (T2K/Tokyo) (2008.6-2014.3)
  – Cluster based on AMD Quad-Core Opteron (Barcelona)
  – 140.1 TFLOPS
• Hitachi SR16000/M1 (Yayoi) (2011.10-)
  – Power 7 based SMP with 200 GB/node
  – 54.9 TFLOPS
• Fujitsu PRIMEHPC FX10 (Oakleaf-FX) (2012.04-)
  – SPARC64 IXfx
  – Commercial version of K computer
  – 1.13 PFLOPS (1.043 PFLOPS for LINPACK, 36th in 43rd TOP500)
  – Additional 576 Nodes with 136 TF (Oakbridge-FX, 2014.04-)
<table>
<thead>
<tr>
<th></th>
<th>HA8000 (T2K)</th>
<th>SMP (Yayoi) SR16000/M1</th>
<th>PRIMEHPC FX10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>AMD Quad Core Opteron 2.3GHz</td>
<td>IBM Power7 3.83GHz</td>
<td>FUJITSU SPARC64iXfx 1.8GHz</td>
</tr>
<tr>
<td>Total # of core</td>
<td>15232</td>
<td>1792</td>
<td>76800</td>
</tr>
<tr>
<td>Total Peak FLOPS</td>
<td>140 TFLOPS</td>
<td>54.9 TFLOPS</td>
<td>1.13 PFLOPS</td>
</tr>
<tr>
<td>Total # of nodes</td>
<td>952</td>
<td>56</td>
<td>4800</td>
</tr>
<tr>
<td>Total Memory</td>
<td>32 TB</td>
<td>11200 GB</td>
<td>150 TB</td>
</tr>
<tr>
<td># of core / node</td>
<td>16</td>
<td>32</td>
<td>16</td>
</tr>
<tr>
<td>Perk FLOPS / node</td>
<td>147.2 GFLOPS</td>
<td>980.5 GFLOPS</td>
<td>236.5 GFLOPS</td>
</tr>
<tr>
<td>Memory / node</td>
<td>32 GB, 128 GB</td>
<td>200 GB</td>
<td>32 GB</td>
</tr>
<tr>
<td>Network</td>
<td>Myrinet 10G Full-bisection</td>
<td>Hierarchical Full-bisection</td>
<td>Tofu 6D Mesh/Torus</td>
</tr>
<tr>
<td>Storage</td>
<td>1 PB</td>
<td>556 TB</td>
<td>1.1PB + 2.1 PB</td>
</tr>
</tbody>
</table>
Supercomputers in U.Tokyo

FY
05 06 07 08 09 10 11 12 13 14 15 16 17 18 19

Hitachi SR11000/J2
18.8TFLOPS, 16.4TB
Fat nodes with large memory

Hitachi SR16000/M1 based on IBM Power-7
54.9 TFLOPS, 11.2 TB
Our last SMP, to be switched to MPP

Hitachi HA8000 (T2K)
140TFLOPS, 31.3TB
(Flat) MPI, good comm. performance

Fujitsu PRIMEHPC FX10 based on SPARC64 IXfx
1.13 PFLOPS, 150 TB
Turning point to Hybrid Parallel Prog. Model

Our last SMP, to be switched to MPP

Post T2K
O(10¹-10²)PFLOPS

Peta

Exa
Research Area based on CPU Hours FX10 in FY.2013 (2013.4~2014.3E)

- Engineering
- Earth/Space
- Material
- Energy/Physics
- Information Sci.
- Education
- Industry
- Bio
- Bio-Mechanics
- Economics
Services for Industry (FX10)

- Originally, only academic users have been allowed to access our supercomputer systems.
- Since FY.2008, we started services for industry
  - supports to start large-scale computing for future business
  - not compete with private data centers, cloud services …
  - basically, results must be open to public
  - max 10% total comp. resource is open for usage by industry
  - special qualification processes/special (higher) fee for usage
- Currently Oakleaf-FX is open for industry
  - Normal usage (more expensive than academic users)
    - 4 groups (FY.2014) (1 IT, 3 manufacturing), fundamental research
  - Trial usage with discount rate
  - Research collaboration with academic rate (e.g. Taisei)
  - Open-Source/In-House Codes (NO ISV/Commercial App.)
Training & Education (FX10)

• 2-Day “Hands-on” Tutorials for Parallel Programming by Faculty Members of SCD/ITC (Free)
  – Fundamental MPI (3 times per year)
  – Advanced MPI (2 times per year)
  – OpenMP for Multicore Architectures (2 times per year)
  – Participants from industry are accepted.

• Graduate/Undergraduate Classes with Supercomputer System (Free)
  – We encourage faculty members to introduce hands-on tutorial of supercomputer system into graduate/undergraduate classes.
  – Up to 12 nodes (192 cores) of Oakleaf-FX
  – Proposal-based
  – Not limited to Classes of the University of Tokyo, 2-3 of 10

• RIKEN AICS Summer/Spring School (2011~)
• HPCI
• Supercomputer Systems in SCD/ITC/UT
• **Overview of Fujitsu FX10 (Oakleaf-FX)**
• Post T2K System
Features of FX10 (Oakleaf-FX)

- **Well-Balanced System**
  - Peak Performance: 1.13 PFLOPS, 398 TB/sec
  - Max. Power Consumption < 1.40 MW (<2.00MW with A/C)
    - Strict Requirement after March 11, 2011
    - 1.043 PFLOPS for Linpack with 1.177 MW (excluding A/C)

- **6-Dim. Mesh/Torus Interconnect**
  - Highly Scalable Tofu Interconnect
  - 5.0x2 GB/sec/link, 6 TB/sec for Bi-Section Bandwidth

- **High-Performance File System**
  - FEFS (Fujitsu Exabyte File System) based on Lustre

- **Flexible Switching between Full/Partial Operation**

- **K compatible (16 cores/node, K: 8 cores/node)**

- **Open-Source Libraries/Applications**

- **Highly Scalable for both of Flat MPI and Hybrid (OpenMP + MPI)**
• Aggregate memory bandwidth: 398 TB/sec.
• Local file system for staging with 1.1 PB of capacity and 131 GB/sec of aggregate I/O performance (for staging)
• Shared file system for storing data with 2.1 PB and 136 GB/sec.
• External file system: 3.6 PB
SPARC64™ IXfx

<table>
<thead>
<tr>
<th>CPU</th>
<th>SPARC64™ IXfx 1.848 GHz</th>
<th>SPARC64™ VIIIfx 2.000 GHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Cores/Node</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Size of L2 Cache/Node</td>
<td>12 MB</td>
<td>6 MB</td>
</tr>
<tr>
<td>Peak Performance/Node</td>
<td>236.5 GFLOPS</td>
<td>128.0 GFLOPS</td>
</tr>
<tr>
<td>Memory/Node</td>
<td>32 GB</td>
<td>16 GB</td>
</tr>
<tr>
<td>Memory Bandwidth/Node</td>
<td>85 GB/sec (DDR3-1333)</td>
<td>64 GB/sec (DDR3-1000)</td>
</tr>
</tbody>
</table>
Racks

• A “System Board” with 4 nodes
• A “Rack” with 24 system boards (= 96 nodes)
• Full System with 50 Racks, 4,800 nodes
Tofu Interconnect

• Node Group
  – 12 nodes
  – A/C-axis: on system board, B-axis: 3 system boards

• 6D: (X,Y,Z,A,B,C)
  – ABC 3D Mesh: connects 12 nodes of each node group
  – XYZ 3D Mesh: connects “ABC 3D Mesh” group
## Software of FX10

<table>
<thead>
<tr>
<th></th>
<th>Computing/Interactive Nodes</th>
<th>Login Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OS</strong></td>
<td>Special OS (XTCOS)</td>
<td>Red Hat Enterprise Linux</td>
</tr>
<tr>
<td><strong>Compiler</strong></td>
<td>Fujitsu</td>
<td>Fujitsu (Cross Compiler)</td>
</tr>
<tr>
<td></td>
<td>Fortran 77/90</td>
<td>Fortran 77/90</td>
</tr>
<tr>
<td></td>
<td>C/C++</td>
<td>C/C++</td>
</tr>
<tr>
<td></td>
<td>GNU</td>
<td>GNU (Cross Compiler)</td>
</tr>
<tr>
<td></td>
<td>GCC, g95</td>
<td>GCC, g95</td>
</tr>
<tr>
<td><strong>Library</strong></td>
<td>Fujitsu</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SSL II (Scientific Subroutine Library II), C-SSL II, SSL II/MPI</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Open Source</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BLAS, LAPACK, ScaLAPACK, FFTW, SuperLU, PETSc, METIS, SuperLU_DIST</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Parallel NetCDF</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Applications</strong></td>
<td>OpenFOAM, ABINIT-MP, PHASE, FrontFlow/blue</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FrontSTR, REVOCAP</td>
<td></td>
</tr>
<tr>
<td><strong>File System</strong></td>
<td>FEFS (based on Lustre)</td>
<td></td>
</tr>
<tr>
<td><strong>Free Software</strong></td>
<td>bash, tcsh, zsh, emacs, autoconf, automake, bzip2, cvs, gawk, gmake, gzip, make, less, sed, tar, vim etc.</td>
<td></td>
</tr>
</tbody>
</table>

**NO ISV/Commercial Applications (e.g. NASTRAN, ABAQUS, STAR-CD etc.)**
• HPCI
• Supercomputer Systems in SCD/ITC/UT
• Overview of Fujitsu FX10 (Oakleaf-FX)

• Post T2K System
Post T2K System

• 20-30 PFLOPS, FY.2015-2016
• Many-core based (e.g. (only) Intel MIC/Xeon Phi)
• Joint Center for Advanced High Performance Computing (JCAHPC, http://jcahpc.jp/)
  – University of Tsukuba
  – University of Tokyo

• New system will installed in Kashiwa-no-Ha (Leaf of Oak) Campus/U.Tokyo, which is between Tokyo and Tsukuba
Post T2K System

- 20-30 PFLOPS, FY.2015-2016
- Many-core based (e.g. (only) Intel MIC/Xeon Phi)
- Joint Center for Advanced High Performance Computing (JCAHPC, http://jcahpc.jp/)
  - University of Tsukuba
  - University of Tokyo
  - New system will installed in Kashiwa-no-Ha (Leaf of Oak) Campus/U.Tokyo, which is between Tokyo and Tsukuba
- Programming is still difficult, although Intel compiler works.
  - (MPI + OpenMP) + Y
  - Tuning for performance (e.g. prefetching) is essential
  - Some framework for helping users needed