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>CPU Type</th>
<th>Total Peak performance</th>
<th>Total number of nodes</th>
<th>Total memory</th>
<th>Peak performance / node</th>
<th>Main memory per 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>150 TB</td>
<td>236.5 GFLOPS</td>
<td>32 GB</td>
<td>1.1 PB + 2.1 PB</td>
</tr>
<tr>
<td>T2K-Todai (Hitachi HA8000-tc/RS425)</td>
<td>AMD Quad Core Opteron 2.3GHz</td>
<td>140 TFLOPS</td>
<td>952</td>
<td>32000 GB</td>
<td>147.2 GFLOPS</td>
<td>32 GB, 128 GB</td>
<td>1 PB</td>
</tr>
<tr>
<td>Yayoi (Hitachi SR16000/M1)</td>
<td>IBM POWER 7 3.83GHz</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
  – Infrastructure (Supercomputers & Distributed Shared Storage System)
    • Seamless access to K, SC’s (9 Univ’s), & user's machines
  – Promotion of Computational Science
    • Strategic Programs for Innovative Research (SPIRE)
  – R&D for Future Systems (Post-peta/Exascale)

• HPCI Consortium
  – Providing proposals/suggestions to the government and related organizations, operations of infrastructure
  – 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業
予測する生命科学・医療
および創薬基盤

戦略プログラム
戦略的プログラム

次世代ものづくり

設計プロセスの革新
熱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削減）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的ものづくりを実現

物質と宇宙の起源と構造
大質量星の超新星爆発の解明

世界に先駆けた次世代デバイスを提案
ナノスケールデバイスをまるごと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「超高性能」、「超低消費電力」端末

防災・減災に資する
地球変動予測

集中豪雨や地震の予測
雲解像モデル、強震動モデル等を駆使して、集中豪雨の位置や被害の損壊規模を高精度に予測し、防災・減災対策に資する。
• 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
- 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, 65th in 45th TOP500)
  - Additional 576 Nodes with 136 TF (Oakbridge-FX, 2014.04-)
Work Ratio

FY.2014: 83.6% Average
Oakleaf-FX + Oakbridge-FX
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
- Economics
Research Area based on CPU Hours
FX10 in FY.2014 (2014.4~2015.3E)

Oakleaf-FX + Oakbridge-FX
Type of Users based on CPU Hours FX10 in FY.2014 (2014.4~2015.3E)

- General Group Users
- HPCI
- JHPCN
- Industry
- Education
- HPC-Challenge
- Personal Users
- Young Researcher

Oakleaf-FX + Oakbridge-FX
Service Fee

• Not FREE

• Service Fee = Cost for Electricity (System+A/C)
  – 2M USD for Oakleaf-FX (2 MW)
  – 1M USD for T2K (1 MW) (~March 2014)
Services for Industry

• Originally, only academic users have been allowed to access our supercomputer systems.

• Since FY.2008, we started services for industry
  – mainly for spread of large-scale parallel computing
  – 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 are needed

• Currently only Oakleaf-FX is open for industry
  – Normal usage (more expensive than academic users)
  – Trial usage with discount rate
  – Research collaboration with academic rate
  – 3-4 groups for each year
  – Open-Source/In-House Codes (NO ISV/Commercial App.)
Education

• Oakleaf-FX only

• 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 to faculty members to introduce hands-on tutorial of supercomputer system into graduate/undergraduate classes.
  – Up to 12 nodes of Oakleaf-FX
  – Proposal
  – Not limited to Classes of the University of Tokyo
• HPCI
• Supercomputer Systems in SCD/ITC/UT
• **Overview of Fujitsu FX10 (Oakleaf-FX)**

• Post T2K System
Features of FX10 (Oakleaf-FX)

• Well-Balanced System
  – 1.13 PFLOPS for Peak Performance
  – Max. Power Consumption < 1.40 MW
    • < 2.00MW including 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 !
• Open-Source Libraries/Applications
• Highly Scalable for both of Flat MPI and Hybrid
FX10 System (Oakleaf-FX)

- 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

- **CPU**: SPARC64™ IXfx 1.848 GHz
- **CPU**: SPARC64™ VIII fx 2.000 GHz
- **Number of Cores/Node**: 16 / 8
- **Size of L2 Cache/Node**: 12 MB / 6 MB
- **Peak Performance/Node**: 236.5 GFLOPS / 128.0 GFLOPS
- **Memory/Node**: 32 GB / 16 GB
- **Memory Bandwidth/Node**: 85 GB/sec (DDR3-1333) / 64 GB/sec (DDR3-1000)

*Copyright 2011 FUJITSU LIMITED*
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>Open Source</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BLAS, LAPACK, ScaLAPACK, FFTW, SuperLU, PETSc, METIS, SuperLU_DIST, Parallel NetCDF</td>
<td></td>
</tr>
<tr>
<td><strong>Applications</strong></td>
<td>OpenFOAM, ABINIT-MP, PHASE, FrontFlow/blue, 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, ANSYS etc.)**
• HPCI
• Supercomputer Systems in SCD/ITC/UT
• Overview of Fujitsu FX10 (Oakleaf-FX)

• Post T2K System
Post T2K System

- >25 PFLOPS, FY.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
- Programming is still difficult, although Intel compiler works.
  - (MPI + OpenMP)
  - Tuning for performance (e.g. prefetching) is essential
  - Some framework for helping users needed