Key Software Technologies for Future High Performance Computing
Key software issues for future HPC

- Scalability
- Manageability
- Power Efficiency
- Productivity
Key software issues for future HPC

- Scalability
- Manageability
- Power Efficiency
- Productivity
Technologies for Scalability

Background:
The performance are always top-priority on HPC application. Scalability is the most important for performance on massively parallel computing.

Issues in this area:
- Improve application parallelization
- Reduce OS Jitter affection
- Keep I/O performance scalability
Technologies for Scalability

- Improve application parallelization

**Technology:** VISIMPACT + Tofu Optimized MPI

- Easy and efficient inter-core Parallelization automated by VISIMPACT

- Threads-Processes hybrid parallelization with Tofu optimized MPI library.
Technologies for Scalability

**Reduce OS Jitter affection**

**Technology: Tuned Linux OS for PRIMEHPC**

- Minimized negative effect of OS jitter ultimately by:
  - Core-binding technology
  - Deliberately selected and tuned system service

- **x86 cluster w/o TCS**
- **x86 cluster w/ TCS**
- **FX10 OS w/ TCS**

*TCS: Technical Computing Suite (Fujitsu’s System Software Product)*

---

**PRIMEHPC Series**

- Compiler
- MPI
- Tools
- PRIMEHPC OS
- Job Manager
- System Manager
- Mgmt. Nodes
- Main Storage

---

Copyright 2013 FUJITSU LIMITED
Technologies for Scalability

Keep I/O performance scalability

**Technology:** FEFS (Fujitsu Exabyte File System)

- Lustre based scalable file system
- Supports up to 8 Exa bytes capacity
- Achieves superb performance on K computer
  - Write: 965 GB/s
  - Read: 1,486 GB/s

Collaborative work with RIKEN
Key software issues for future HPC

- Scalability
- Manageability
- Power Efficiency
- Productivity
Technologies for Manageability

**Background:**
For system administrator, managing system is a key issue. We have already achieved nearly 100,000 nodes system on K computer.

**Issues in this area:**
- Availability
- Operability
Technologies for Manageability

- **Availability**

**Technology:** Automatic Management Node failover
- Immediate failover by Hot Stand-by

**Technology:** Automatic MDS/OSS failover
- Monitored by TCS system manager

**MDS: Active/Standby**
- IB SW
- IB multirail
- Node failover
- FC multipath
- RAID1+0

**OSS Active/Active**
- IB SW
- RAID6
- RAID6
Technologies for Manageability

**Operability**

**Technology: Centric Management** provides single system image for:
- System Installation / Update
- Node status Monitoring (Hardware / Software)
- Power Control / Monitoring
- Support PRIMEHPC / x86 Hybrid Cluster system

**Technology: Flexible cluster management** provides various physical/logical partitioning.

**Technology: QoS/Directory Quota on FEFS** facilitates sharing global storage across multi cluster system.
Key software issues for future HPC

- Scalability
- Manageability
- Power Efficiency
- Productivity
Technologies for Power Efficiency

Background:

$$\text{Power Efficiency} = \frac{\text{“Actual” Throughput}}{\text{Total system power}}$$

Customer Requirement:
✓ Increasing “actual” system throughput
✓ Keeping total power at target value

Issues in this area:
- Maximize application efficiency
- Maximize Resource utilization
- System-wide Power Management

Copyright 2013 FUJITSU LIMITED
 Technologies for Power Efficiency

Maximize application efficiency

**Technology:** Optimized OS and languages achieves good efficiency for many applications.

<table>
<thead>
<tr>
<th>Application</th>
<th>Nodes</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINPACK</td>
<td>88,128</td>
<td>93%</td>
</tr>
<tr>
<td>NICAM</td>
<td>81,920</td>
<td>8%</td>
</tr>
<tr>
<td>Seism3D</td>
<td>82,944</td>
<td>18%</td>
</tr>
<tr>
<td>PHASE</td>
<td>82,944</td>
<td>20%</td>
</tr>
<tr>
<td>RSDFT</td>
<td>82,944</td>
<td>52%</td>
</tr>
<tr>
<td>FrontFlow/blue</td>
<td>80,000</td>
<td>3%</td>
</tr>
<tr>
<td>Lattice QCD</td>
<td>82,944</td>
<td>16%</td>
</tr>
<tr>
<td>ZZ-EFSI</td>
<td>82,944</td>
<td>46%</td>
</tr>
</tbody>
</table>

Table is provided by Dr. Minami of RIKEN
Technologies for Power Efficiency

Maximize Resource Utilization

**Technology:** Various job allocation method to increase node/core utilization even on Torus system
- Torus mode/Mesh mode allocation
- Node simplex/share allocation
- Heterogeneous hybrid parallel job allocation

System-wide Power Management

**Technology:** Centric Power Control helps to integrate center-wide power capping with:
- Interface to control power of nodes or storages
- Power consumption monitoring
- Control power saving mode w/ job manager
Key software issues for future HPC

- Scalability
- Manageability
- Power Efficiency
- Productivity
Background:
Because of increasing complexity of node and network architecture, developing applications become more difficult.

Issues in this area:
- Tuning and Debugging
- Portability
Technologies for Productivity

- **Tuning and Debugging**

  **Technology:**
  - Supports world’s standard debugger (DDT)
  - Profiler
    - GUI Detailed PA information
    - Optimize communication on Tofu
  - Rank Mapping Optimization (RMATT)
Technologies for Productivity

Portability

Technology: Compatibility with K computer
- Binary compatibility
- Same Architecture applied
  1cpu/node, VISIMPACT, Tofu interconnect

Technology: Continues to supports:
- The latest international standards
  Fortran 2008, C 11, C++ 11
- De facto standards
  GNU C/C++ extensions, OpenMP 4.0, MPI 3.0

Technology: Supports Generic Linux OS
- POSIX compliant Linux and generic libraries
Now, we are ready for 100 PFlops!

What’s next?
Activities for Exascale Computing

Roadmap of Japan’s National Project

<table>
<thead>
<tr>
<th>National projects</th>
<th>Development</th>
<th>Operation of K computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>App. review</td>
<td>HPCI</td>
<td>Exa-system development project</td>
</tr>
<tr>
<td>FS projects</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Participate co-design for Exascale System software

- Light-weight Micro kernel next to Linux
- File I/O performance improvement