Current Status on the Development of the K Computer

Mitsuo YOKOKAWA
Next-Generation Supercomputer R&D Center
RIKEN
What is “K computer”?

- “京 (kei)” is a nickname of the next-generation supercomputer system.
  - The name was chosen from public applications this July.
  - It comes from a target performance of the system, that is 10petaFLOPS.
  - “京” is a Japanese prefix number which stands for $10^{16}$, or 10 peta.

- “K computer” is assigned for English documents.
- A logo 写 is written in Japanese calligraphy is also determined.

- Another meaning of the “京” is “a big gate.”
  - A new era of computational science is coming though the gate “京” by hoping promised future success.
Contents

- Outline of the project
- System configuration of the K computer
- Facilities
Outline of the project
Outline of the next-generation supercomputer project

- Objectives are
  - to develop the world's most advanced and high-performance supercomputer,
  - to develop and deploy its usage technologies including application software
as one of Japan's Key Technologies of National Importance.

- Period of the project: FY2006-FY2012 (7-year project)

- RIKEN (The Institute of Physical and Chemical Research) plays the central role of the project in developing the.
Goals of the project

- Development and installation of the most advanced high performance supercomputer system with LINPACK performance of 10 petaflops.

- Development and deployment of application software, which should be made to attain the system maximum capability, in various science and engineering fields.

- Establishment of an “Advanced Institute for Computational Science” as one of the Center of Excellences around supercomputing facilities.
## Schedule of the project

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual design</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detailed design</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prototype, evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Production, installation, and adjustment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tuning and improvement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Applications</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Next-Generation Integrated Nanoscience Simulation</td>
<td>Development, production, and evaluation</td>
<td>Verification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Next-Generation Integrated Life Simulation</td>
<td>Development, production, and evaluation</td>
<td>Verification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Buildings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computer building</td>
<td>Design</td>
<td>Construction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research building</td>
<td>Design</td>
<td>Construction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We are here.
System configuration of K computer
System configuration

**Compute Nodes**
- Number of nodes > 80K
- Number of cores > 640K
- Total memory capacity > 1PB

**Interconnect Network**
Six-dimensional mesh/torus

- Local File System
- Global I/O Networks
- Global File System
- Frontend Servers
- Users
- Internet

Networks for Control and Management

- Control Servers
- System Configuration
- Management Servers
- Job & User Management

SNA+MC2010, Oct. 17-20, Tokyo, Japan
Compute nodes and network

- **Compute nodes (CPUs):** > 80,000
  - Number of cores: > 640,000
- **Peak performance:** > 10PFLOPS
- **Memory:** > 1PB (16GB/node)

- **6-dimensional mesh/torus network: Tofu**
  - 10 connections to each adjacent node
- **Peak bandwidth:** 5GB/s x 2 for each connection
- **Logically 3-dimensional torus network**

**Compute node**
- SPARC64™ VIIIfx
- CPU: 128GFLOPS (8 cores)
- Core SIMD(4FMA) 16GFLOPS
- L2$: 6MB
- MEM: 16GB
- 5GB/s(peak) x 2 per connection
- 5GB/s(peak) x 2 in total

**Diagram:**
- Courtesy of FUJITSU Ltd.
CPU features (Fujitsu SPARC64™ VIIIfx)

- 8 cores
- 2 SIMD operation circuit
  - 2 Multiply & add floating-point operations (SP or DP) are executed in one SIMD instruction
- 256 FP registers (double precision)
- Hardware barrier among cores
- Pre-fetch instruction
- Shared 6MB L2 Cache (12-way)
  - Software controllable cache (sectored cache)

Performance
- 16GFLOPS/core, 128GFLOPS/CPU
- 45nm CMOS process, 2GHz
- 22.7mm x 22.6mm
- 760 M transistors
- 58W (at 30°C by water cooling)

Reference: SPARC64™ VIIIfx Extensions
System environments

- OS: Linux based OS on compute nodes

- Two-level large-scale distributed file system with local file system and global file system

  - Users’ permanent files reside in the global file system
  - Staging functions
    - Files on the global file system which are used in a job are staged into the local file system.
    - Data generated during a job execution are moved back to the global file system after the job finished.

- Batch job-oriented system
  - Interactive environments are also available in debugging.
**Flow of batch job processing**

- **Job submission**: Users submit their jobs with directions to a scheduler.
- **File staging**: Input files used in the job are copied from a global file system to a local file system.
- **Job execution**: The job is executed on the assigned compute nodes.
- **File de-staging**: Output files are moved back to the global file system.

**Scheduler** assigns the job to appropriate compute nodes according to the job description in the script.

**Check the job status**

**Input data** flows from the global file system to the local file system before being executed on the compute nodes. After execution, the output data is moved back to the global file system.
Facilities for the system
AICS: location of the K computer in Kobe

AICS (Advanced Institute for Computational Science) was established at RIKEN this July.

450km (280miles) west from Tokyo
Photo (as of May 31, 2010)
Layout of the buildings

- Research Building
- Computer building
- Chillers
- Substation Supply
Cooling system
Features of the buildings (1/3)

Basement
- Soil improvement work was done at the depth of 20m below ground for prohibiting liquefaction damage.
- Basement of 2m was made so that the research and computer buildings are constructed on it.

Seismic isolated structure
- Three kinds of seismic isolated equipment are adapted to ensure S-grade (6 Japanese quake scale) quake-resistance.
Features of the buildings (2/3)

- Prevention from damage by salt in the seawater
  - Aluminum panels and glass wall are used at outer skins of buildings to prevent damages by seawater.

- Large computer room without pillars
  - Flexibility of a layout of computer cabinets and cabling of interconnect network and workability at installation are considered.
    - Average load capacity is 1,000kg/m² to support all cabinets of K computer.
    - Depth of the raised floor is 1.5m (5 ft.).
First installation of the K computer

- Eight racks are housed on the 3rd floor of the computer building on October 1, 2010.

- 98.3 TFLOPS in peak performance.

- The system will be in operational in 2012.
Thank you for your attention!

A photo in the early evening