FUJITSU Supercomputer
PRIMEHPC FX100
Hardware and Software Overview
FUJITSU Supercomputers

- Fujitsu has been developing supercomputers nearly 40 years, and will continue its development to deliver the best application performance.

**K computer**
Peak performance: 11.28 petaflops

**PRIMEHPC FX10**
Peak performance: up to 23.2 petaflops

**PRIMEHPC FX100**
Peak performance: over 100 petaflops

Exascale
**PRIMEHPC FX100 Design Concept**

- Designed to be a massively parallel supercomputer system
  - High performance for a wide range of real applications

- Inherited the K computer features
  - General purpose CPU architecture for application productivity
  - 6D mesh/torus topology, hardware barrier synchronization, sector cache, etc.

- Introducing new technologies for Exascale computing
  - HPC-ACE2: Wide SIMD enhancements
  - Assistant cores: Dedicated cores for non-calculation operation
  - HMC: Leading-edge memory technology
**Over 1 TF high performance processor**

- 32 compute cores
- 2 assistant cores: Offloading non-calculation operations (Daemons, IOs, non-blocking MPI functions, etc.)

**HPC-ACE2: ISA enhancements**

- Two 256-bit wide SIMD units per core
- Various SIMD instructions (stride load/store, indirect load/store, permutation, etc.)

**HMC support**

- 480GB/s/node of theoretical memory throughput
Enenhanced Tofu interconnect

- Highly scalable, 6-dimensional mesh/torus topology
- Increased link bandwidth by 2.5 times to 12.5GB/s
- Added atomic memory operations

CPU-integrated interconnect controller

- Reduced communication latency
- Improved packaging density and energy efficiency

Optical cable connection between chassis

- Enable flexible installation
Enhanced software stack developed by Fujitsu

<table>
<thead>
<tr>
<th>Management software</th>
<th>High Performance File System</th>
<th>Programming Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>System management</td>
<td>lustre-based distributed file system (enhanced for FX100)</td>
<td>MPI, OpenMP, COARRAY</td>
</tr>
<tr>
<td>Job management</td>
<td></td>
<td>Compilers (C, C++, Fortran) Mathematical libraries</td>
</tr>
</tbody>
</table>

Linux-based OS enhanced for FX100

PRIMEHPC FX100
## The Evolution of FUJITSU Software

<table>
<thead>
<tr>
<th>2011 (K computer)</th>
<th>2015 (PRIMEHPC FX100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fortran, C, C++ with sophisticated optimization</td>
<td>COARRAY in Fortran 2008, C++11 with advanced vectorization for wide SIMDs</td>
</tr>
<tr>
<td>Scalable MPI over 100k procs</td>
<td>Asynchronous MPI comm. for low-latency and scalability</td>
</tr>
<tr>
<td>Large-scale job scheduler (over 80k nodes)</td>
<td>Flexible job allocation for high throughput computing</td>
</tr>
</tbody>
</table>

### Future

- Optimization strategy based on application characteristics
- Scalable MPI over 100k procs
- Power saving functions

---

**Copyright 2015 FUJITSU LIMITED**
FX100 Performance and the Effect of the New Technologies
FX100 greatly improves the performance of various types of programs

### Node Performance of Benchmarks and Applications

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>FX10 (16 cores, 128-bit SIMD)</th>
<th>FX100 (32 cores, 256-bit SIMD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPL</td>
<td>1.0x</td>
<td>4.2x</td>
</tr>
<tr>
<td>HPCG</td>
<td>2.0x</td>
<td>3.0x</td>
</tr>
<tr>
<td>NPB FT</td>
<td>3.0x</td>
<td>3.3x</td>
</tr>
<tr>
<td>NTChem Mini</td>
<td>4.0x</td>
<td>4.0x</td>
</tr>
<tr>
<td>NICAM-DC Mini</td>
<td>2.9x</td>
<td>3.1x</td>
</tr>
<tr>
<td>CCS QCD Mini</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UPACS-Lite</td>
<td></td>
<td>4.7x</td>
</tr>
</tbody>
</table>

* Fiber miniapp suite developed by RIKEN

† Calculation of compressible fluid dynamics. Developed by JAXA
Balanced Enhancement of FLOPS and Memory

- Over 1 TFLOPS and 480 GB/s memory bandwidth per chip
- PRIMEHPC series show high performance for both HPL and HPCG

Chip Performance of HPL and HPCG
Loop Vectorization by New SIMD Instructions

- Vectorizing complex loops is a key to get higher performance.
- FX100 introduces new SIMD instructions, such as non-continuous memory accesses, integer calculations, permutation, compression, etc.

The Effect of SIMD Compression (NPB EP)

<table>
<thead>
<tr>
<th>Relative perf.</th>
<th>Scalar FX10</th>
<th>Scalar FX100</th>
<th>SIMD*</th>
</tr>
</thead>
</table>

* With a loop fission to promote vectorization

256-bit register

src1

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

src2 (mask bits)

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
</table>

compress

dest

<table>
<thead>
<tr>
<th>B</th>
<th>C</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
</table>
By offloading MPI processing to assistant cores, non-blocking communications are performed simultaneously with computation.

Scalability Improvement by Overlapping (The Himeno Benchmark*)

* A stencil code solving the Poisson’s equation solution
† Halo exchanges are overlapped
Fujitsu MPI now supports MPI-3.0, including RMA functions!

Almost all FX100’s RMA functions start transfer asynchronously (no remote response required)

**MPI_Put Microbenchmark**

- **Latency**
  - Message Size vs. Latency for FX10 and FX100
  - FX100 shows lower latency for all message sizes.

- **Throughput**
  - Message Size vs. Throughput for FX10 and FX100
  - FX100 shows higher throughput for all message sizes.

**Legend**
- FX10
- FX100

1. Request msg.
2. Waiting for a response
3. No need to wait for a response
4. Atomic one-sided operation
5. Data

**Note:**
- >80% lower latency
- 5x faster throughput
Fujitsu MPI provides high-bandwidth collective functions optimized for Tofu

- High bandwidth of the Tofu interconnect 2 (peak 12.5GB/s per network engine)
- Driving 4 network engines in parallel
- Low latency communication protocol thanks to RDMA

Why are Fujitsu MPI's collectives so fast?

**Algorithm**
- Phase 1
- Phase 2
- Phase 3

**MPI_Allgather Throughput (36 nodes)**

- Throughput (GB/s)
- Message Size (bytes)

<table>
<thead>
<tr>
<th>Message Size</th>
<th>FX10</th>
<th>FX100</th>
</tr>
</thead>
<tbody>
<tr>
<td>1k</td>
<td>14</td>
<td>14.9x</td>
</tr>
<tr>
<td>32k</td>
<td>25</td>
<td>28.9x</td>
</tr>
<tr>
<td>1M</td>
<td>35</td>
<td>39.9x</td>
</tr>
<tr>
<td>32M</td>
<td>40</td>
<td>43.9x</td>
</tr>
</tbody>
</table>

2.9x faster
Summary

- FX100 achieves high performance of various applications by the new technologies and inherited features.
- This evolution is continuing to the next generations.

K computer

PRIMEHPC FX10

PRIMEHPC FX100