Fujitsu has a leading role in HPC for >30 years

- **AP3000**
- **PRIMEPOWER**
- **HPC2500**
  - Most Efficient Performance in Top500 (Nov. 2008)

**VPP5000**
- World's Fastest Vector Processor (1999)
- No.1 in Top500 (June and November 2011)

**VPP300/700**
- Developed with NAL

**FX1**
- K computer
- Most Efficient Performance in Top500 (Nov. 2008)

**FX10**
- Trans-Exascale

**PRIMEQUEST**
- Japan's First Vector (Array) Supercomputer (1977)

**VP Series**
- Japan's First Vector (Array) Supercomputer (1977)

**F230-75APU**

**NWT**
- Developed with NAL

**World's Fastest Vector Processor (1999)**

**No.1 in Top500 (Nov. 1993)**
- Gordon Bell Prize (1994, 95, 96)

**World's Most Scalable Supercomputer (2003)**

**AP1000**

**AP3000**

**VPP500**

**SPARC Enterprise**

**PRIMERGY**
- PRIMERGY CX400
- PRIMERGY BX900 / BX400 Cluster node
- PRIMERGY RX200 Cluster node

**PRIMERGY**
- Cluster node

**Number 1 in Top500 (June and November 2011)**

**Japan's Largest Cluster in Top500 (July 2004)**

**World's Most Scalable Supercomputer (2003)**

**Vector Supercomputer Series**

**Scalar MPP Series**

**Scalar Supercomputer Series**

**x86 Cluster Series**

**1980**
- Japan's First Vector (Array) Supercomputer (1977)

**1990**
- No.1 in Top500 (Nov. 1993)

**1995**
- Gordon Bell Prize (1994, 95, 96)

**2000**

**2005**
- Japan's Largest Cluster in Top500 (July 2004)

**2010**
- Next x86 generation

**2015**
- PRIMERGY CX400

**2020**
- PRIMERGY BX900 / BX400 Cluster node
- PRIMERGY RX200 Cluster node

*NWT: Numerical Wind Tunnel*
World-wide Network of Partners and Resources

**FUJITSU Japan**
- Global lead in High Performance Computing
- Strategy, Development, Services and Support

**FUJITSU Technology Solutions**
- PRIMERGY based HPC Ecosystem
- Services and Support

**FUJITSU Systems Europe, ict GmbH - a FUJITSU company**
- HPC application champions
- Benchmarking

**FUJITSU Laboratories**
- Research & Development
  - e.g. Open PetaScale Libraries Network

Partners and Resources:
- Exa
- ANSYS
- MSC Software
- MAGMA
- Platform Computing
- Intel
- Microsoft
- redhat.
- NVIDIA
- Mellanox Technologies
- oerc
- ANU
- Imperial College London
- JAIST
PRIMERGY x86 HPC Servers
Fujitsu provides the largest reach with x86 Open Standards

- The only vendor to provide Mainframe on x86 technology (BS2000 SQ series)
- Micro-Server to x86 MC & Mainframe
- SME to Fortune 500 customers
- Complete PRIMERGY HW & Ecosystem
- Solutions from SBS to Cloud
HPC Market and our offering

Customer Needs
- Density
- Reliability
- Network Throughput
- Low Power consumption
- Simultaneous data access
- Storage Size
- Storage Throughput

Increasing with market segmentation and size

Project type
- Highly customized projects
- Customized projects
- Bespoke / Limited Customization
- Packaged Turn-Key Solutions (Ready To Go)

Market segmentation
- Workgroup < 100K$
- Departmental > 100K$
- Divisional > 250K$
- Super Computer > 500K$

Fujitsu HPC ECO System
- PRIMEHPC FX10
- PRIMERGY CX400
- PRIMERGY BX900
- PRIMERGY Rack Servers
- Power Workstation

Increasing with market segmentation and size
Meet diverse HPC needs

- **Scalability, Compute density**
  - BX400
  - BX900
  - RX2530 M1
  - RX2540 M1
  - CX400 M1

- **HPC specific components**
  - Intel & Mellanox IB Interconnect
  - Intel Xeon Phi / NVIDIA GPGPU

- **Capability**
  - Scalability, Infrastructure density

- **Flexibility to address specific customer requirements**
New dual-socket PRIMERGY M1 servers

The datacenter standard without compromise

Server blade with maximum scalability for most demanding applications

Platform for cloud, HPC, and large scale-out computing

---

**Factor 1.2x - 1.7x application related performance vs. predecessor.**

---

**Head / File Server / Login Node**

**Head / Compute / FileServer / Login / NAS Node**

**Compute / Accelerator Node**

---

**Copyright 2014 FUJITSU**
CX400 M1: with high focus on HPC

Platform for HPC and large scale-out computing

- 2U / 4 CPU nodes or 2 nodes for CPU + Accelerator
- Latest Technology: Haswell CPU + DDR4 memory for max. performance
- Double the density for accelerators: CX2570 M1
- Increased I/O: 2x PCIe slot and up to 6 x 2.5” HDD per node, including SATA DOM option
- Support of Fujitsu Cool-safe® Liquid Cooling Technology coming soon

Compact server node density with high power efficiency to realize large scale-out solutions for HPC and cloud computing at lower overall costs

Dual socket server node in a highly condensed half-wide, 1U form factor

Dual socket server node for ambitious high performance computing, analytics and visualization solutions

CX400 M1
Chassis

CX2550 M1
Server Nodes

CX2570 M1
Standard server node for PRIMERGY CX2550 M1

- New condensed half-width-1U server form factor enables to plug up to 4 CX2550 M1 into a 2U chassis
- Intel® Xeon® processor E5-2600 v3 product family, advanced Turbo Boost 2.0 technology
- 16 DIMMs per server node with up to 1,024 GB DDR4 memory and up to 2,133 MHz DRAM bandwidth
- Variable local storage: 6x 2.5” or 1x SATA-DOM per Node
- Fan less server node with shared power and cooling
- Hot-plug for server nodes, power supplies and disk drives enable enhanced availability and easy serviceability

HPC Usage Scenarios

- Head Node
- Compute Node
- File Server Node
- Login Node
- NAS
- Accelerator Card Node
FUJITSU Server PRIMERGY CX2570 M1
Server Node

HPC optimized server node for PRIMERGY CX400 M1

- New condensed half-width-2U server form factor enables to plug up to 2 CX2570 M1 into a 2U chassis
- Intel® Xeon® processor E5-2600 v3 product family, advanced Turbo Boost 2.0 technology
- 16 DIMMs per server node with up to 1,024 GB DDR4 memory and up to 2,133 MHz DRAM bandwidth
- 2 optional GPGPU or co-processor compute cards per node, selected from the market leading NVIDIA® Tesla™ or the Intel® Xeon Phi™ product family
- Variable local storage: 6x 2.5” or 1x SATA-DOM per Node
- Fanless server node with shared power and cooling
- Hot-plug for server nodes, power supplies and disk drives enable enhanced availability and easy serviceability.

HPC Usage Scenarios

- Head Node
- Compute Node
- File Server Node
- Login Node
- NAS
- Accelerator Card Node
FUJITSU Server PRIMERGY RX2540 M1
2 socket 2U Rack Server

The versatile powerhouse

- Intel® Xeon® E5-2600 v3 product family with up to 18 core per processor and Turbo Boost 2.0
- Intel® Advanced Vector Extensions 2 (AVX2) to improve compute-intensive performance
- High scalability of up to 1536GB RAM (24 DIMMs), up to 24*) storage drives and 8 PCIe slots Gen3
- 12G SAS support to accelerate database solutions
- Cool-safe® Advanced Thermal Design: Support higher temperature range of 5 – 40°C
- Support of LTO drives*)
- Modular concept enables individual & cost saving configuration of drives, LAN (DynamicLoM) and power supply units
- “Wire-Once” flexible virtualization by ServerView Virtual I/O Manager (VIOM)
- Simplified power management and two hot-plug PSU with 96% efficiency result in leading energy efficiency in its class

*) from 2/2015
Fujitsu PRIMERGY RX2530 M1*
2 socket 1U Rack Server Node

Maximum productivity in a 1U housing

- Intel® Xeon® E5-2600 v3 product family with up to 18 core per processor and Turbo Boost 2.0
- Intel® Advanced Vector Extensions 2 (AVX2) to improve compute-intensive performance
- High scalability of up to 1536GB RAM (24 DIMMs), up to 10 storage drives and 4 PCIe slots Gen3
- 12G SAS support
- Cool-safe® Advanced Thermal Design: Support higher temperature range of 5 – 40°C
- Modular concept enables individual & cost saving configuration of drives, LAN (DynamicLoM) and power supply units
- Simplified power management and two hot-plug PSU with 96% efficiency result in leading energy efficiency in its class

*) from 1/2015
FUJITSU Server PRIMERGY BX900
Server Blade PRIMERGY BX2560 M1

Maximum scalability for most demanding applications

- Blade chassis holds up to 18 compute blades
- Versatile solution includes optional storage blades and mid-plane connection units
  - Integrated dual-channel 10 Gbit/s Converged Network Adapter offering 8 physical functions and additional slots for mezzanine cards
  - BX900 S3 supports InfiniBand FDR 56Gb connection blade
  - Storage blades for integrated storage solution
- Compute nodes include:
  - Housing two Intel® Xeon® E5-2600 v3 processors up to 18 cores and 45 MB smart cache
  - Leading 22nm manufacturing process leads to up to ~40% more performance in same power envelope
  - Up to 16 slots for high-speed (up to 2,133 MHz) DDR4 memory, which are accessed via 4 channels per CPU (up to 512 GB capacity)
  - 2 x 2.5” SSD or 1 x 2.5” SATA HDD
- Save time and conserve valuable IT resources by simplifying remote management with the new, CIM compliant, integrated Remote Management Controller (iRMC S4)

HPC Usage Scenarios

- Head Node
- Compute Node
- File Server Node
- Login Node
- NAS
- Accelerator Card Node
HPC compute node – Many Core trend

- Customer model size and needs to shorten simulation time are increasing rapidly
- Further significant performance development for Xeon DP CPUs will not continue to improve with factors 2-3x
- Accelerator / GP GPU technology
  - Helps to
    - Significantly increase performance for HPC workloads
      - E.g. Xeon Phi provides x1.5 performance boost over Xeon DP CPU (Haswell)
    - Reduce cost overhead
    - Improve density
  - Challenges at
    - Limited scalability
    - Application efficiency (requires adoption / modification)
    - Cooling capabilities in Data Centre
A path towards Exascale …

… enforces a deployment of parallelism at each level to the ultimate extent:

- Node level (distributed memory)
- Multi socket (shared memory on nodes)
- CPU level (number of cores)
- Instruction level (SIMD)

Challenges

- Node parallelism ➞ ultra-high-speed interconnect
- CPU parallelism ➞ higher memory bandwidth ➞ greater complexity of memory hierarchy
- Core parallelism ➞ increase of system errors
- Amdahl’s Law ➞ every portion of serial code lowers the overall performance

**Floating-point operations per second**

<table>
<thead>
<tr>
<th>Name</th>
<th>FLOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>exaFLOPS</td>
<td>$10^{18}$</td>
</tr>
<tr>
<td>petaFLOPS</td>
<td>$10^{15}$</td>
</tr>
</tbody>
</table>

Towards Many Core Architectures
Cluster Interconnect

- Interconnect is a key performance factor in parallel execution environment
  - Compute performance increased by Factor 12 in past 5 Years
  - Interconnect performance need to stay aligned

- What’s next?
  - Mellanox: EDR 100Gb/s (Q2/2015)
  - Intel Omniscale (Q4/2015)
Interconnect comparison

- Interconnect is breaking through the 100Gb/s

**Mellanox EDR**
- EDR is continues development of existing technology (downward compatible)
- Based on PCIe Controller
- CPU offloading and RDMA
- 36 Ports Switch radix
  - Managed, Unmanged (both in FTS portfolio)
- 648 ports director switch
- Max Cluster size 11.664 ports non blocking

**Intel Omniscale**
- Based on Truescale, but with major enhancements
- Direct out of CPU (KNL first, Xeon DP later)
- New protocol: no downward compatibility
- NO CPU offloading, NO RDMA
- 48 Ports Switch radix
- 1152 ports director switch with silicon photonics interface
- Max Cluster size 27.648 ports non blocking
Fujitsu Software HPC Cluster Suite (HCS)

- A complete HPC software stack solution
The need for a HPC software stack

- **HPC cluster general characteristics**
  - HPC clusters consist primarily of compute nodes with exactly the same hardware
  - Clusters can have a few to thousands of compute nodes
  - The software used on each compute node is exactly the same
  - Compute nodes do not have keyboards or displays attached to them

- **Fundamental operational requirements**
  - Bare metal deployment with no intervention
  - Central management and deployment of all needed software components
  - A way to control resource usage across the compute nodes
  - Ability to run many applications (Serial & Parallel) simultaneously on the cluster
  - High-speed inter node communication and access to large data storage areas
  - Some sort of shared storage is needed
  - Monitoring and management of nodes
The Fujitsu HPC Cluster Suite (HCS)

- Comprehensive software stack for managing Fujitsu PRIMERGY HPC clusters
  - Easy-to-use cluster management
  - Popular workload managers
  - General HPC Open Source Software
  - Highly scalable parallel file system
  - Graphical end-user interface for simplified usage

- Alliance with leading ISVs
- Fully validated HPC solution

Fujitsu HCS coverage
Front-end interface with 2 main roles

- **Process Manager – the End-user interface**
  - Simplified usage of HPC applications for end-users
  - Enables sharing of expertise and results between team members
  - Can be used from remote locations
  - Job control through workload manager
  - File explorer for user and system disk areas
  - Service view of applications

- **Cluster Manager – the Administrator interface**
  - Manage nodes and node groups
  - Define network, software, mount-points
  - Monitor cluster usage
  - Define alert conditions
Simplicity + Expertise = Gateway Application Catalogue

Extendable web desktop – Integrated and intuitive HPC environment

Download pre-built interface packages from Fujitsu Application Catalogue web site.

Import extended application processes into local Gateway.

Incorporate business-relevant HPC methods into project workflow.

User Gateway site

Fujitsu web site
HPC Software Stack integrated with PRIMERGY

- HPC cluster accessible through web browser with single sign-on.
  - User works inside an adaptable and intuitive desktop within the browser.

- Direct access to run directory and application-specific data
  - Application Catalogue packages increasingly provided with data monitors.
  - Allows users to track application progress through model-related metrics.
  - User-selectable tabular and plotting formats.

- Create and retain parameter profiles for different experiments
  - Entry form for main parameters for given application.
  - Lists of pre-defined options for selected parameters.
  - File navigator to select inputs and set output location.
  - Accounting project code for system usage tracking.

- Workflow Editing
  - Projects space management
  - Drag-drop workflow graph designer
  - Wide range of variables types and attributes
  - Flexibility and fine-grain control
  - Preset preferred/fixed parameters
Gateway Summary

- Intuitive and adaptable HPC user interface
- Natural desktop working practice, but within your browser
- Supports multiple applications
- Tools to run, monitor and organise HPC workloads
- Download and import extended application methods
- Run own scripts and onboard standardised interface
Fujitsu HPC Cluster Suite: Summary

Simplify HPC to lower cost and risk, and increase access

Fujitsu Ease of Use for more HPC population

Build in Expertise to realise more value from HPC
FEFS – High speed Parallel file system
Why Parallel File system?

- Fujitsu’s offering for parallel file system is **FEFS** (Fujitsu Exabyte File System)
What is FEFS?

- Developed to be the world's fastest parallel file system
  - Originally developed for the demands of the K-Computer
    - Achieved 1TB/s single I/O performance
  - Based on Lustre, most popular distributed file system for HPC

- Brings value add/differences compared with Lustre
  - High scalability and performance
    - up to 1 TB/s performance and 8 ExaByte capacity
  - Additional Quality of Service functions
    - Fair Share QoS & Directory Quota
  - Highly reliability
    - fully validated with PRIMERGY and ETERNUS, additional bug fixed by Fujitsu, Built-in High Availability
  - Easy installation
    - FEFS design sheet creates configuration files automatically
    - HPC Cluster Suite provides automated installation scheme for I/O nodes
K computer* achieved the world’s highest file system throughput
IO throughput performance > 1 TB/s

Compute nodes: 82944
Global FEFS File system: 30PB

Read bandwidth: > 1 TB/s

*: K computer has been co-developed by RIKEN and Fujitsu limited.
FEFS characteristics

- Extremely Large capacity
  - Extra-large volume (100PB~1EB)
  - Massive number of clients (100k~1M) & I/O servers (1k~10k)

- High I/O Performance
  - Throughput of Single-stream (~GB/s) & Parallel IO (~TB/s)
  - Reducing file open latency (~10k ops)

- High Reliability and High Availability
  - Continuation of file service even if a component failure occurs

- I/O Usage Management
  - Fair-share QoS
  - Best-effort QoS

- FEFS is optimized for maximizing hardware performance while minimizing file I/O overhead

Copyright 2014 FUJITSU
Components of the FEFS file system

**Management Server (MGS)**
- Stores configuration information for all FEFS file systems in a cluster
- Plays no active part in I/O operations
- Is contacted by each component to understand the file system configuration
- Can co-exist on an MDS

**Metadata Server (MDS)**
- Stores file system metadata in one or more MDT’s
- Each MDS manages the names and directories of one FEFS file system
- Multiple MDS’s can be used as Active/Standby pairs to enable fail-over capability

**FEFS Clients**
- Usually computational or visualization nodes that mount one or more FEFS file systems
- FEFS clients see a single coherent and synchronised namespace regardless of how many OSS’s/OST’s are used to create the file system

**File Management Server**
- Monitors server and service health of the storage cluster
- Automates fail-over process when a failure occurs

**Object Storage Server (OSSs)**
- Provides file I/O services for one or more OST’s (typically between 2 and 8 OST’s)
- User file data is stored in one or more objects. Each object being on a separate OST
- Files are striped based on a per file, per directory or per file system specification

**Object Storage Target (OSTs)**
- Stores file data in one or more objects. Each object being on a separate OST
- Files are striped based on a per file, per directory or per file system specification

**Management Target (MGT)**
- Stores I/O storage cluster data on participating servers and general configuration control data
- Can co-exist on an MDS

**Metadata Target (MDT)**
- Stores metadata on storage attached to an MDS.
- Each file system has an MDT
- When placed on a shared storage system can be used by multiple MDS’s for failover
FEFS – extending the Lustre base

FEFS features

Large scale
- Max file size
- Max number of files
- Max client number
- Max stripe count
- 512KB block

High performance
- File striping
- Parallel I/O
- Client cache
- Server cache
- MDS response
- I/O zoning
- OS zoning
- OS jitter reduction

Lustre features

Network
- Tofu Interconnect
- IB Multi-rail
- IB/Ether
- LNET Router

Connectivity
- Lustre mount
- NFS export

Reliability
- Failover
- Journal / fsck
- RAS

Operations Management
- ACL
- Disk Quota
- Directory Quota
- Dynamic configuration change

New
Extended
Reused

Copyright 2014 FUJITSU
FUJITSU

shaping tomorrow with you