FUJITSU

shaping tomorrow with you

Fujitsu A64FX Processor Technical presentation

Feb 4th, 2025

Presenter: John Wagner - Senior HPC Solutions Architect

FUJITSU TECHNOLOGY SOLUTIONS

This document includes descriptions of assistant cores and TofuD which are only for Fugaku and FX1000.

Contents

What is the A64FX Processor

- Fujitsu Processor Development
- DNA of Fujitsu Processors
- Technical Info of A64FX
- A64FX Specifications and Efficiency
- A64FX Features
- Execution Unit

FUJITSU History of HPC System

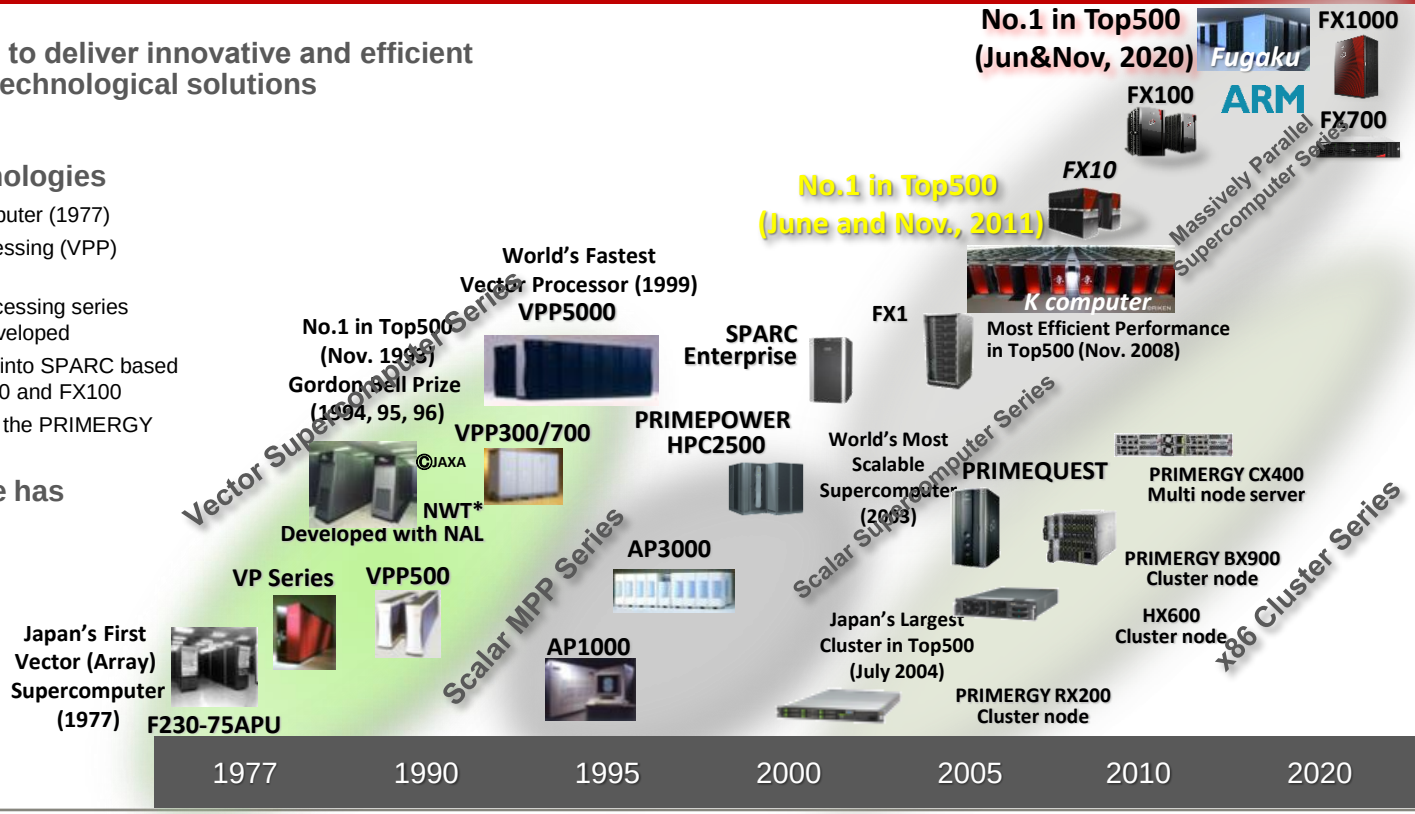
Delivering high-end computing solutions for over 40 years

FUJITSU's goal has always been to deliver innovative and efficient systems through our advanced technological solutions

Diverse product range and technologies

- Starting with Vector array supercomputer (1977)
- Developed into Vector Parallel Processing (VPP) system
- In addition, a Massively Parallel Processing series based on SPARC processor was developed
- The MPP range technology evolved into SPARC based supercomputers of K-computer, FX10 and FX100
- Today this range is accompanied by the PRIMERGY Intel based product series

And finally, the ARM architecture has been chosen to take us into the 'Exascale' processing era



*NWT: Numerical Wind Tunnel

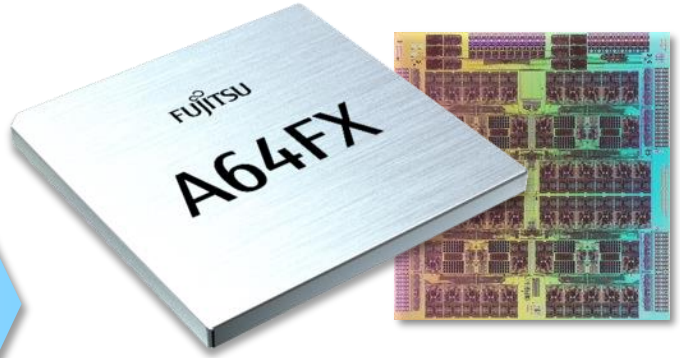
DNA of Fujitsu Processors

A64FX inherits DNA from technologies in the HPC, UNIX servers, & Mainframes

High performance-per-watt
Execution & memory throughput
Low power
Massively parallel

High speed & flexibility
Thread performance
Software on Chip
Large SMP

High reliability
Stability
Integrity
Continuity



CPU with extremely high throughput
High performance
Low power consumption
High reliability

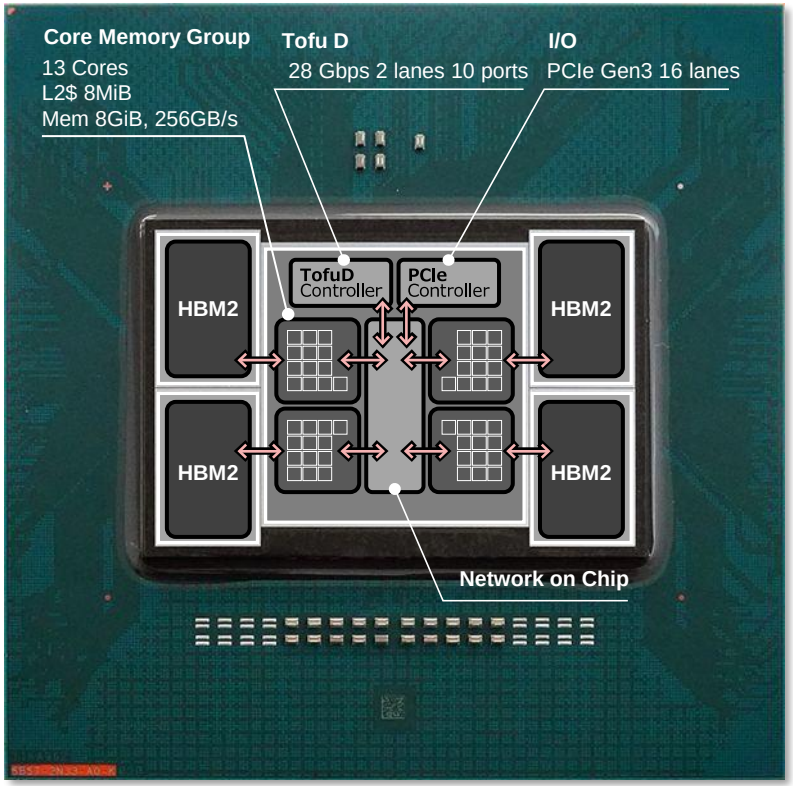
Technical Info of A64FX

Architecture Features

- Arm V8.2-A (aarch64 only)
- SVE 512-bit wide SIMD x 2 pipeline per core
- 48 computing cores + 4 assistant cores[*1]
  All 52 cores are identical
- Frequency 1.8GHz, 2.0GHz, 2.2GHz[*1]
- L1I$ size: 3MiB (64KiB x 48 computing core)
- L1D$ size: 3MiB (64KiB x 48 computing core)
- L2 cache size: 32MiB (8MiB x 4 CMG[*2])
- HBM2 32GiB
- PCIe Gen3 16 lanes
- TofuD[*1] 6D Mesh/Torus,
  28Gbps x 2 lanes x 10 ports

7nm Fin FET

- 8,786M transistors
- 594 package signal pins



*1: 2.2GHz, Assistant cores and TofuD are only available for Fugaku and FX1000
*2: CMG: Core Memory Group

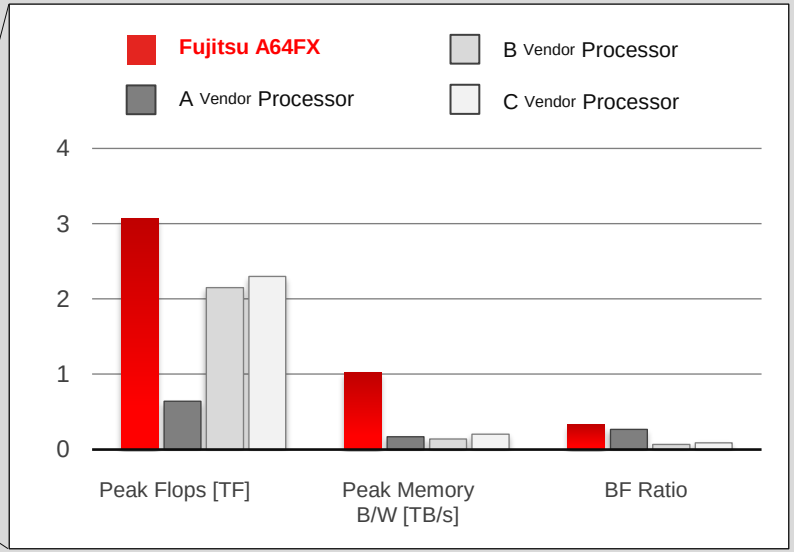
A64FX Specifications and Efficiency

A64FX peak performance and efficiency

- 3.072TFLOPS@2.0GHz, >90%@DGEMM
- Memory B/W 1024GB/s, >80%@Stream Triad

Specifications

	Fujitsu A64FX	A Vendor Processor	B Vendor Processor	C Vendor Processor
ISA	**Arm v8.2-A**	**Arm v8-A**	**x86**	**x86**
Vector instructions	SVE	Neon	AVX512	AVX256
Process Node	7nm	16nm	14nm	7nm
# of Cores	48	32	28	64
Memory	HBM2	DDR4	DDR4	DDR4
Peak FLOPS	3.072TF	0.64TF	2.15TF	2.30TF
Peak Memory B/W	1024GB/s	171GB/s	141GB/s	205GB/s
BF Ratio	0.333	0.267	0.066	0.089



A64FX Features

Collaboration with Arm to develop and optimize SVE for a wide range of applications

- FP16 and INT16/8 dot product are introduced for AI applications

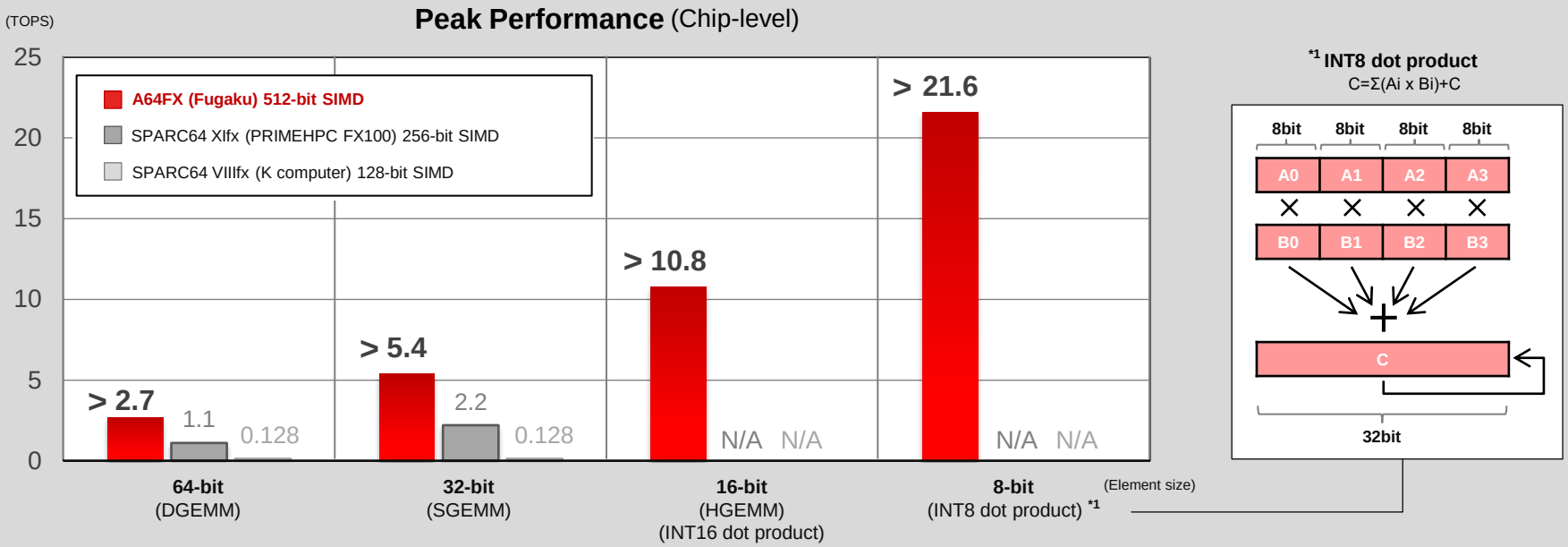
	A64FX	SPARC64 Xifx [PRIMEHPC FX100]	SPARC64 VIIIfx [K computer]
ISA	**Armv8.2-A + SVE**	**SPARC-V9 + HPC-ACE2**	**SPARC-V9 + HPC-ACE**
SIMD Width	✓ 512-bit	256-bit	128-bit
Predicated Operations	✓ Enhanced	✓	✓
Four-operand FMA	✓ Enhanced	✓	✓
Gather/Scatter	✓ Enhanced	✓	–
Math. Acceleration	✓ Further enhanced	✓ Enhanced	✓
Compress	✓ Enhanced	✓	–
First Fault Load	✓ New	–	–
FP16	✓ New	–	–
INT16/ INT8 Dot Product	✓ New	–	–
HW Barrier* / Sector Cache*	✓ Further enhanced	Enhanced	✓
Many core architecture	✓ Further enhanced (48 cores)	Enhanced (32 cores)	✓ (8 cores)
High Band Width	✓ Further enhanced (1024GB/sec)	Enhanced (240GB/sec x2 in/out)	✓ (64GB/sec)

* Utilizing aarch64 implementation-defined system registers and available with FUJITSU Software Technical Computing Suite or FUJITSU Software Compiler package

Execution Unit

Extremely high throughput

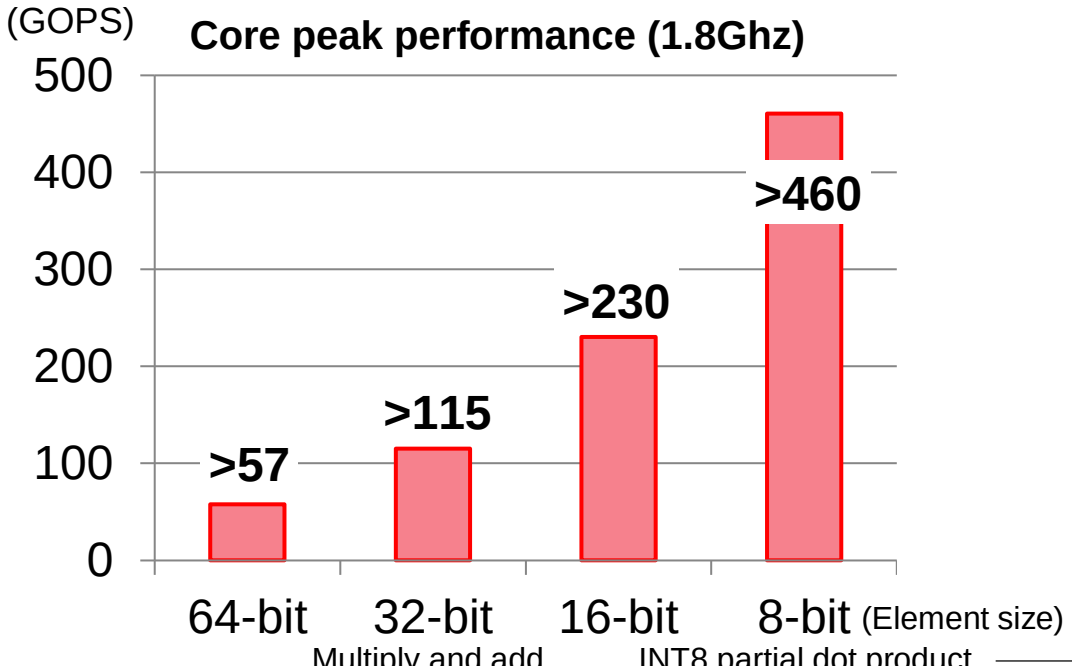
- 512-bit wide SIMD x 2 Pipelines x 48 Cores
- >90% execution efficiency in (D|S|H)GEMM and INT16/8 dot product

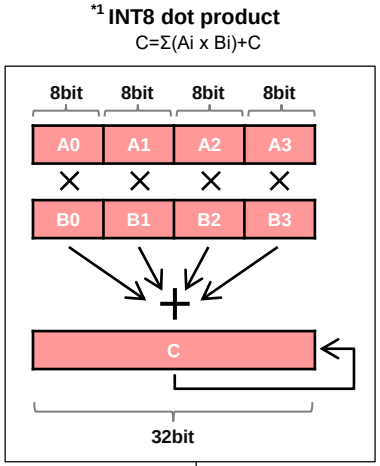


A64FX technologies: Core performance

High calc. throughput of Fujitsu original CPU core w/ SVE

- 512-bit wide SIMD x 2 pipelines and new integer functions





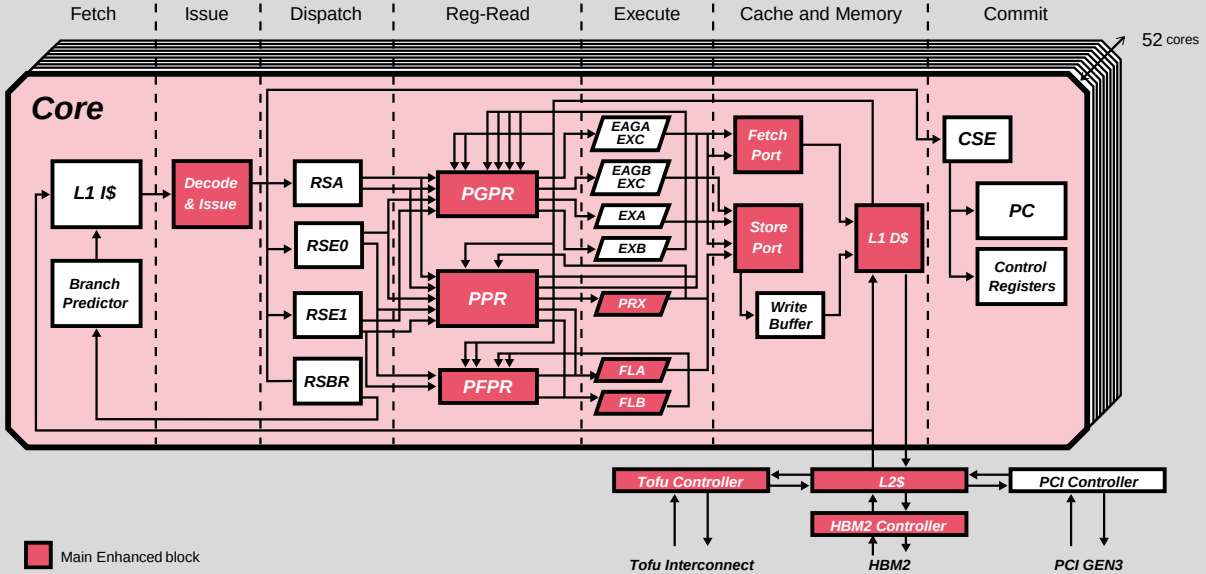
Why HPC Performance is High

- A64FX Core Pipeline
- SIMD Width
- Four-operand FMA with Prefix Instruction
- Gather/ Scatter (Level 1 Cache)
- Sector Cache
- Many-Core Architecture
- High Bandwidth
- High Performance in Benchmark
- High Performance on Real Applications

A64FX Core Pipeline

A64FX inherits and enhances superior features of the SPARC64 VIIIfx (K computer's CPU)

- Inherits superscalar, out-of-order, branch prediction, etc.
- Enhances SIMD and predicate operations (PR)
  - 2x 512-bit wide SIMD FMA + Predicate Oper. + 4x ALU (shared with 2x AGEN)
  - 2x 512-bit wide SIMD load or 512-bit wide SIMD store



SIMD Width

SIMD Width Contribution to Computing Performance

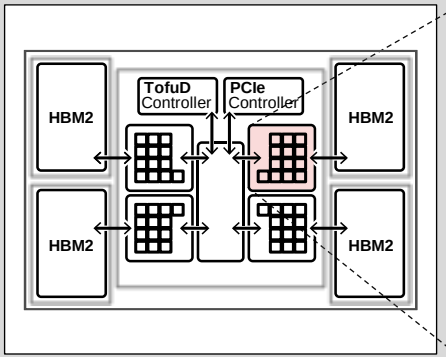
- Peak performance [double-precision] is 3.072 TFlops

4 [CMG*s] × 12 [Compute cores] (A) × 2 [Floating point pipelines] × 8 [SIMD Register Length(512) / dfloat variable(64)] × 2 [Number of operations for FMA] (B)

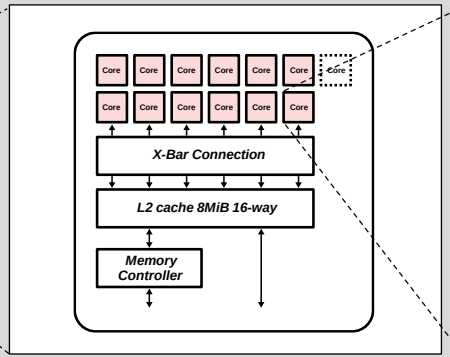
× 2.0GHz [Operating frequency] = **3.072** [Tflops]

A

A64FX

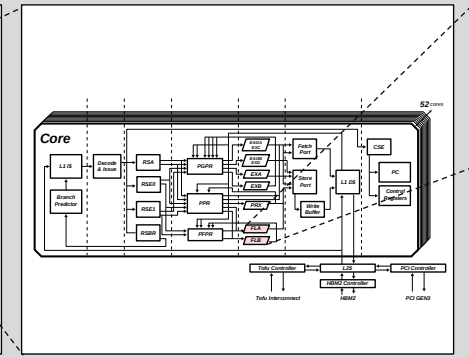


CMG* Configuration

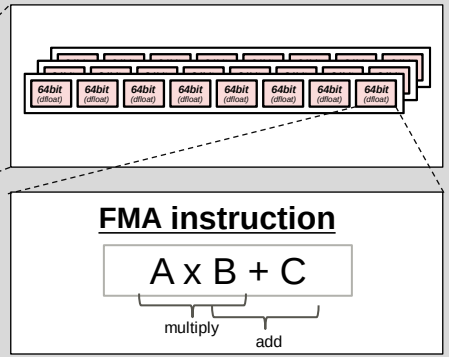


B

CPU core pipeline structure



SIMD Register Length = 512bit



* CMG: Core memory Group

Four-operand FMA with Prefix Instruction

MOVPRFX as a prefix instruction

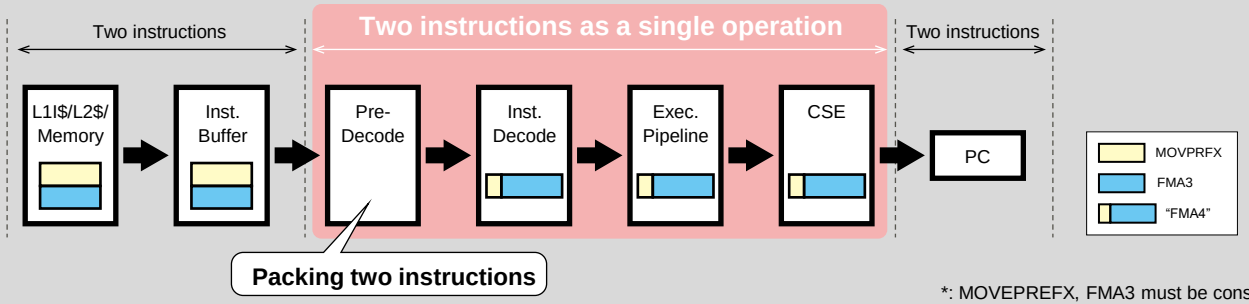
- For SVE, four-operand “FMA4” requires a prefix instruction (MOVPRFX) followed by destructive 3-operand FMA3

A64FX Implementation for MOVPRFX

- A64FX hides the overhead of its main pipeline by packing MOVPRFX and the following instruction into a single operation

MOVPRFX as a prefix instruction

A64FX Implementation for MOVPRFX



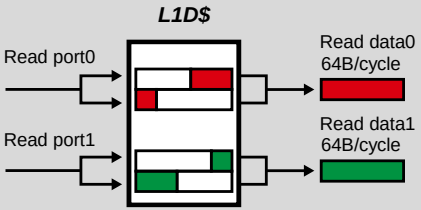
*: MOVEPREFX, FMA3 must be consecutive

Gather/Scatter (Level 1 Cache)

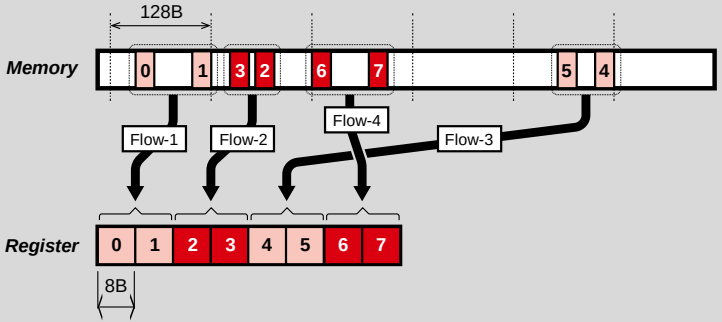
L1 cache throughput maximizes core performance

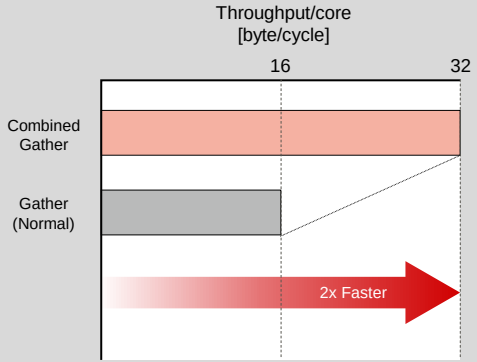
- **Sustained throughput** for 512-bit wide SIMD load
  - An unaligned SIMD load crossing cache line keeps the same throughput
- **Combined Gather** mechanism increasing gather throughput
  - Gather processing is important for real HPC applications
  - A64FX introduces “Combined Gather” mechanism enabling to return up to two consecutive elements in a “128-byte aligned block” simultaneously

Sustained throughput



Combined Gather





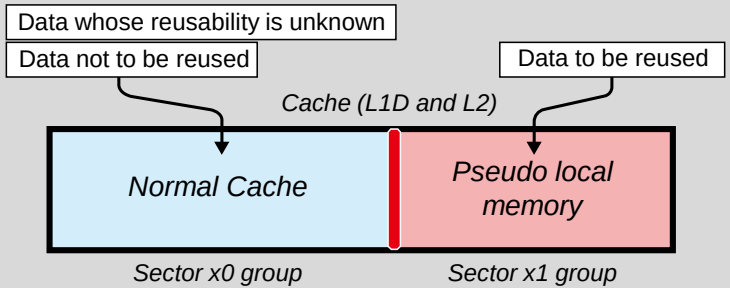
Sector Cache

Concept

- Software-controlled cache mechanism
  - **Conventional caches' issue:** Hindering performance improvements because of evicting the data including reusable data from the cache when registering other data
  - **The sector cache mechanism:** To achieve higher performance, splitting the cache into two sectors frequently reused data in a sector separately and allows software control

Sector Cache Technology

- Newly implemented the sector cache mechanism into the L1D as well as L2
- Using sector x1group for data to be reused to improve reusability of cache and reduce cache miss



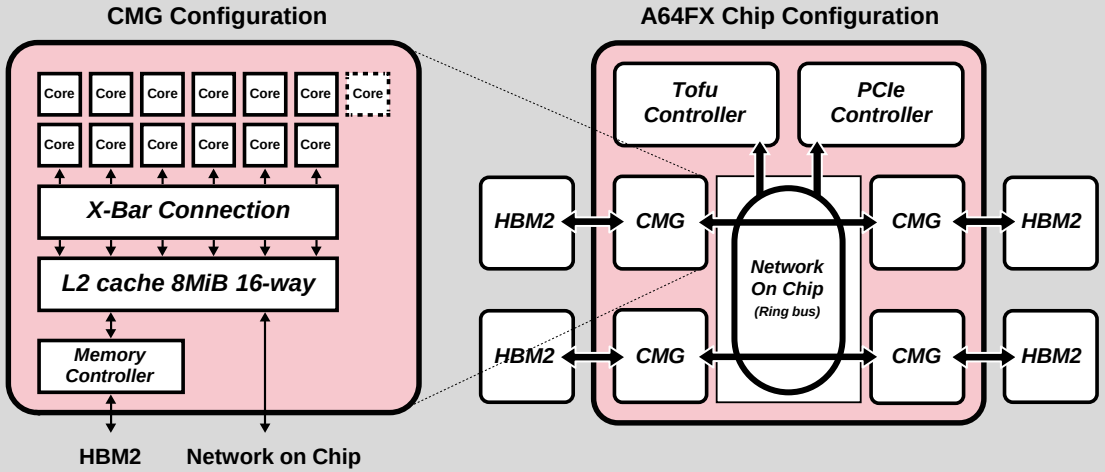
* “Sector Cache” is available with FUJITSU Software Technical Computing Suite

Many Core Architecture

A64FX consists of four CMGs (Core Memory Group)

- A CMG consists of 13 cores, L2 cache and a memory controller
  - One out of 13 cores is an assistant core which handles daemon, I/O, etc.
- Four CMGs keep cache coherency by ccNUMA with on-chip directory
- X-bar connection in a CMG maximizes high efficiency for throughput of the L2 cache
- Process binding in a CMG allows linear scalability up to 48 cores

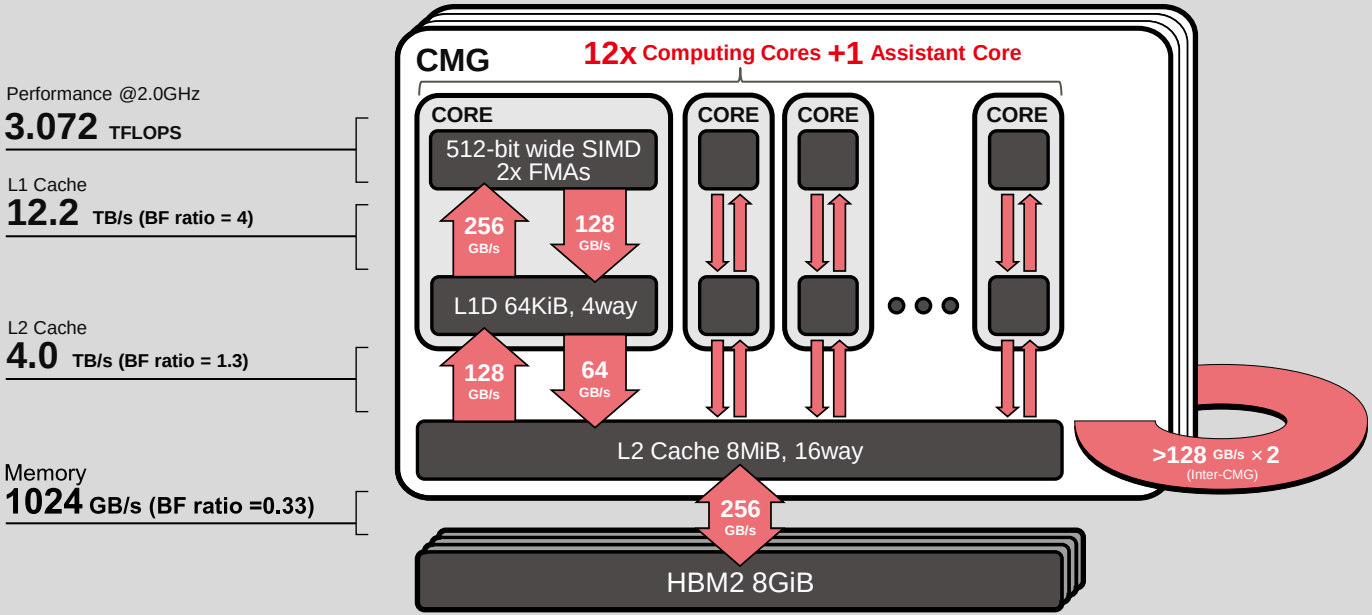
On-chip-network with a wide ring bus secures I/O performance



High Bandwidth

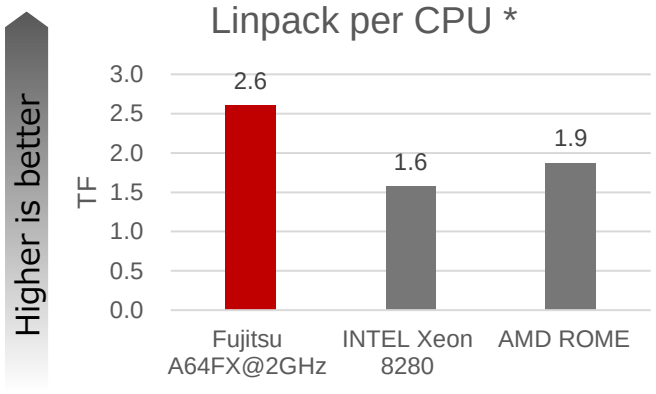
Extremely high bandwidth in caches and memory

- A64FX has out-of-order mechanisms in cores, caches and memory controllers
  It maximizes the capability of each layer's bandwidth



High Performance in Benchmark

Compute intensive benchmark

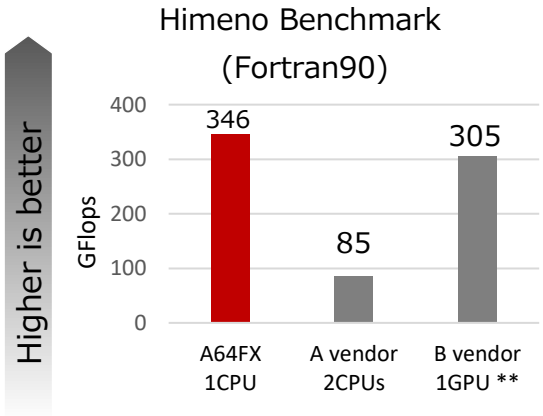


Linpack Benchmark:

The benchmark to solve a dense system of linear equations

* Calculated from top500.org 2019/11

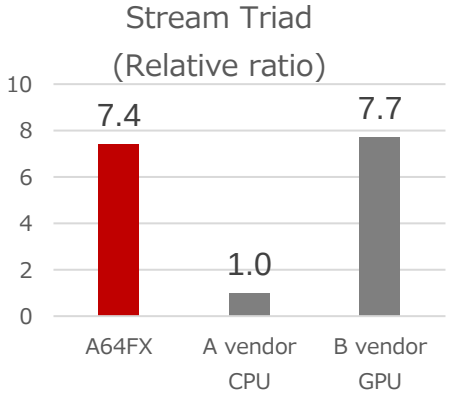
Memory intensive benchmark



Himeno Benchmark:

Stencil calculation to solve Poisson's equation by Jacobi method

** Performance evaluation of a B vendor's vector supercomputer system



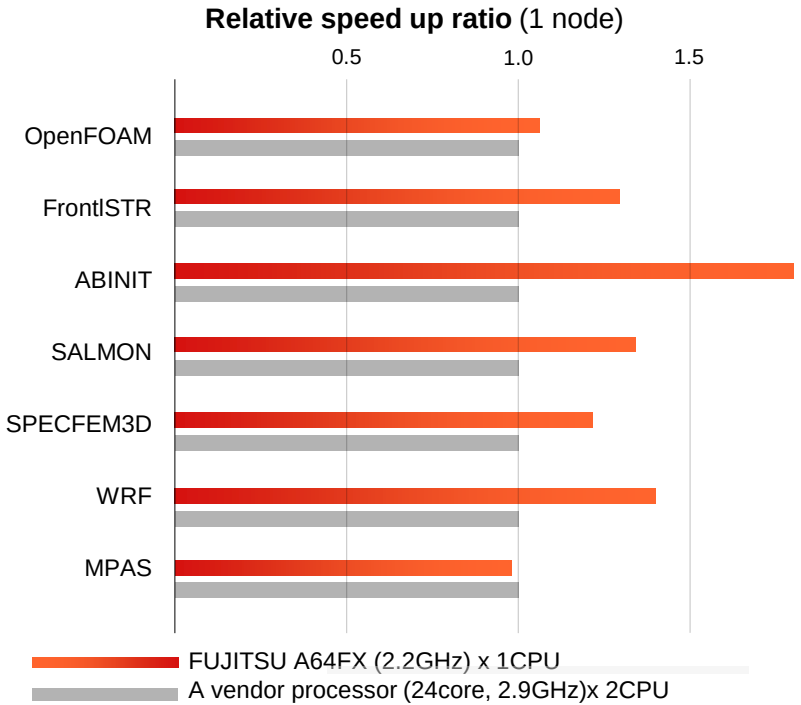
Stream Triad:

The benchmark to measure memory B/W for simple vector kernels

High Performance on Real Apps

The performance on 1node is evaluated for seven OSS applications

- Measured on PRIMEHPC FX1000, A64FX 2.2GHz
- Up to 1.8x faster over A vendor Processor x2
- High memory B/W and 512-bit wide SIMD work effectively with these applications



Why Power Consumption is Low

- Low Power Consumption Technology
- Custom Design
- 2.5D Packaging Technology
- Power Management
- Green500, Nov. 2019 Result
- High Performance in Power Efficiency

Low Power Consumption Technology

7nm Process

Leakage Power Reduction

- Water cooling system (Fugaku)
- Multi-Vth standard cell
- Transistor channel width optimization
  - Fujitsu in-house design tool reduces channel width of cells, while still meeting operating frequency

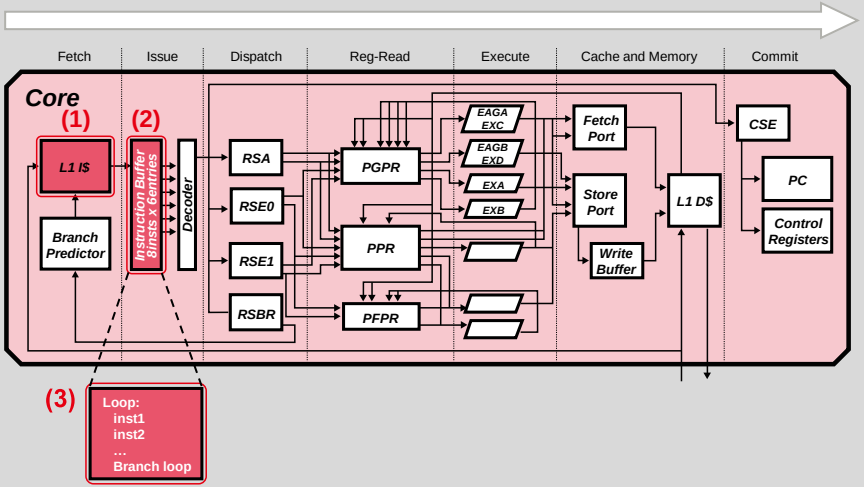
Dynamic Power Reduction

- Clock gating to in-active latches
- Micro-architecture optimization

e.g. Short Loop Detector

The basic operation is fetched from (1)L1I $ and decoded via (2) Instruction Buffer.

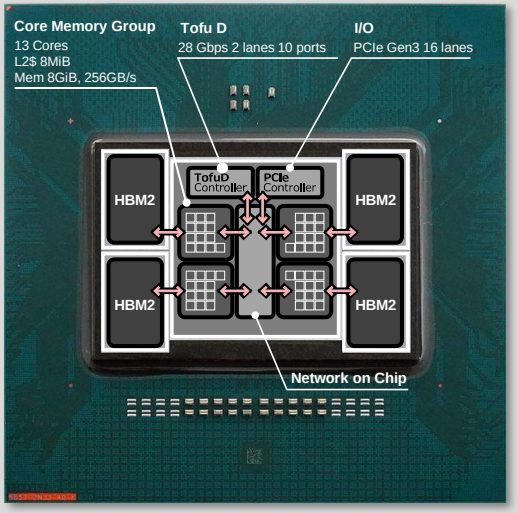
However, if (3) "loop contained in the Instruction Buffer" is detected, the instruction fetch from (1)L1I$ is stopped **to save power** and fetch the instruction from the (2)Instruction Buffer.

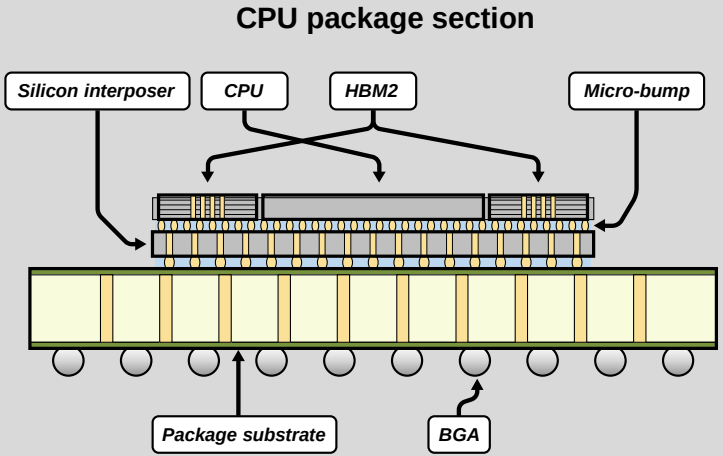


2.5D Packaging Technology

High density packaging technology

- Heterogeneously integrate CPU chip and 3D stacked memory into single package using 2.5D packaging technology
- The CPU chip includes four HBM2 memory controllers and interfaces for PCI-Express and ToFu D networks





Power Management

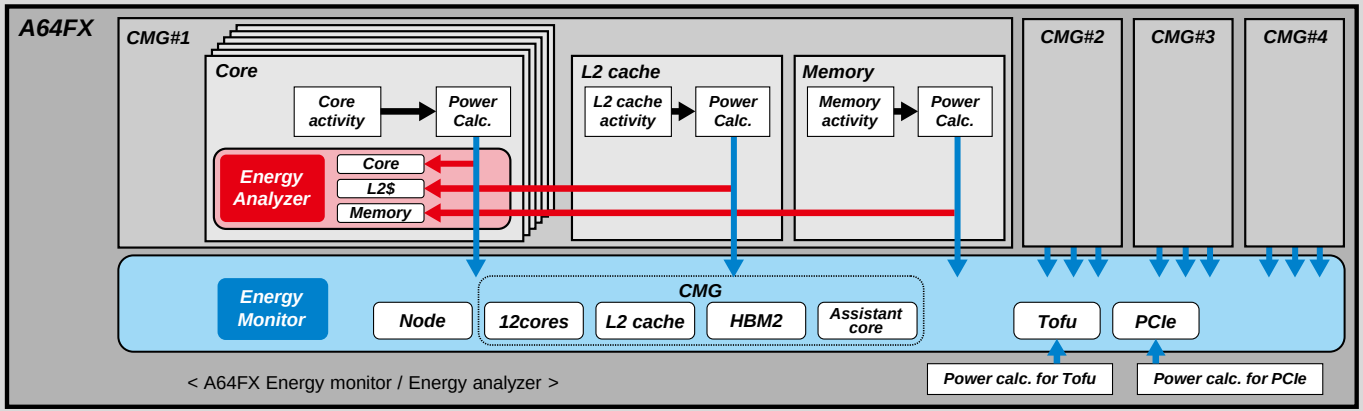
“Energy monitor” / “Energy analyzer” for activity-based power estimation

- Energy monitor (per chip) : Node power via Power API* (~msec)
  - Average power estimation of a node, CMG (cores, an L2 cache, a memory) etc.
- Energy analyzer (per core) : Power profiler via PAPI** (~nsec)
  - Fine grained power analysis of a core, an L2 cache and a memory
- **Enabling chip-level power monitoring and detailed power analysis of applications**

*Suggested by Sandia National Laboratories and is available with FUJITSU Software Technical Computing Suite

** Performance Application Programming Interface

Enabling chip-level power monitoring and detailed power analysis of applications



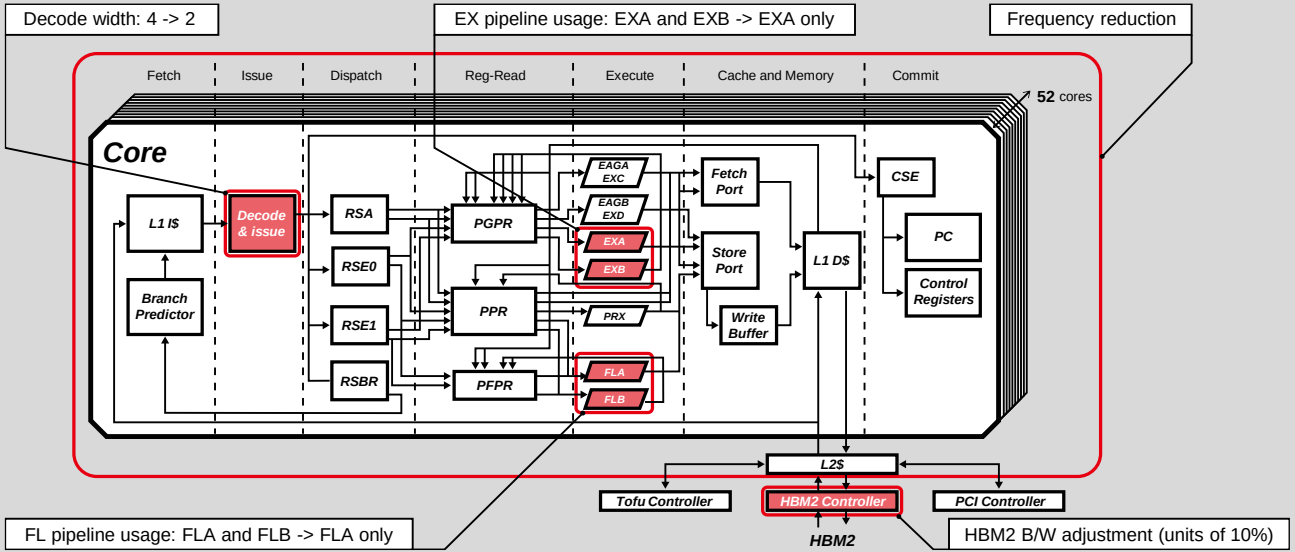
< A64FX Energy monitor / Energy analyzer >

Power Management (Cont.)

“Power Knob” for power optimization

- A64FX provides power management function called Power Knob*
  - Applications can change hardware configurations for power optimization
- Power knob and Energy monitor/analyzer help users optimize power consumption of their applications

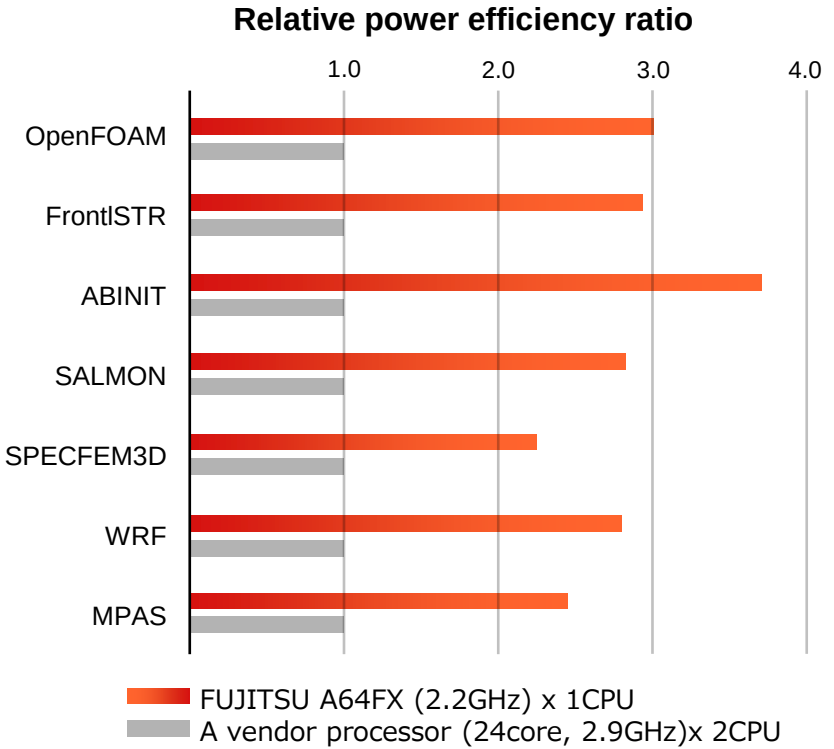
*__Power knob__ is available in FUJITSU Software Technical Computing Suite



High Performance in Power Efficiency

The power efficiency on 1node is evaluated for seven OSS applications

- Measured on PRIMEHPC FX1000, A64FX 2.2GHz
- Up to 3.7x more efficient over A vendor Processor x2
- High power efficiency is achieved by energy-conscious design and implementation



Why is Reliability High

- **Fujitsu Mission Critical Technologies**
- **A64FX inherits K computer's CPU Reliability**

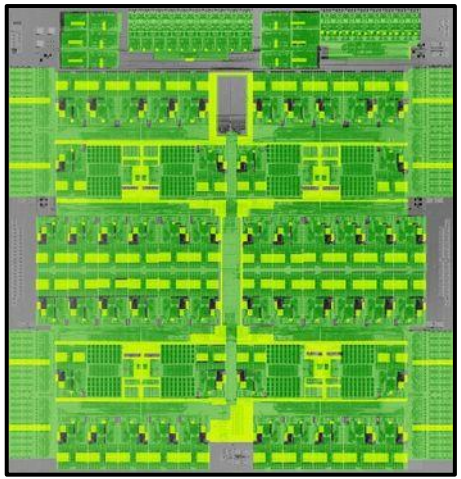
Fujitsu Mission Critical Technologies

Large systems require extensive RAS capability of CPU and interconnect

- A64FX has a **mainframe class RAS** for integrity and stability
  It contributes to very low CPU failure rate and high system stability
  - ECC or duplication for all caches
  - Parity check for execution units
  - Hardware instruction retry
  - Hardware lane recovery for Tofu links
  - **~128,400 error checkers in total**

High reliability from Fujitsu MF technologies

Units	Error Detection and Correction
Cache (Tag)	ECC, Duplicate & Parity
Cache (Data)	ECC, Parity
Register	ECC (INT), Parity(Others)
Execution Unit	Parity, Residue
Core	Hardware Instruction Retry
PCI Express	Multi layer CRC Error Recovery
TofuD	Hardware Lane Recovery



A64FX RAS Diagram

- 1bit error Correctable
- 1bit error Detectable
- 1bit error Harmless

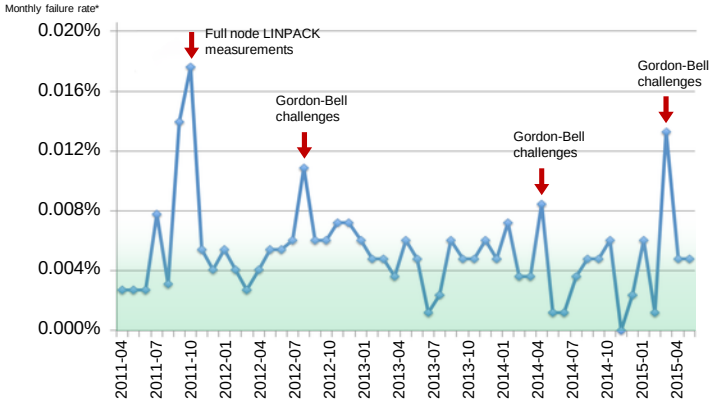
A64FX inherits K computer's CPU Reliability

FUJITSU

A64FX inherits K computer's RAS technologies

- Failure trend of K computer's 80,000+ CPUs [SPARC64 VIIIfx] was low

Monthly Failure Rate of CPUs



Failure trend of CPUs is almost stable except high load terms

*Monthly failure rate = $\frac{\text{Failure counts in the month}}{\text{Number of installed CPUs}}$

Number of CPUs = **82,944**
(Since July 2012)

Comparison with Blue Waters

AFR: Annual Failure Rate (Average failure rate per year)
FIT: Failure In Time (1FIT = 1 failure per 10^9hours)

Units	Parameter	K computer (April 2011 – June 2015)	Blue Wates*
CPU	# of Parts	82,944	49,238
	AFR	0.06%	0.23%
	FIT	72.00	265.15
DIMM	# of Parts	663,552	197,032
	AFR	0.016%	0.112%
	FIT	18.02	127.84
	FIT / GB	9.01	15.98

*C.Di Martio et al., Lessons learned from the analysis of system failures at petascale: the case of blue waters.
44th international conference on Dependable Systems and Networks (DSN 2014),2014.

CPU failure rates of the K computer are about one quarter compared to that of Blue Waters.

(Source: ISC 2015 Long term failure analysis of 10 petascale supercomputer, RIKEN)

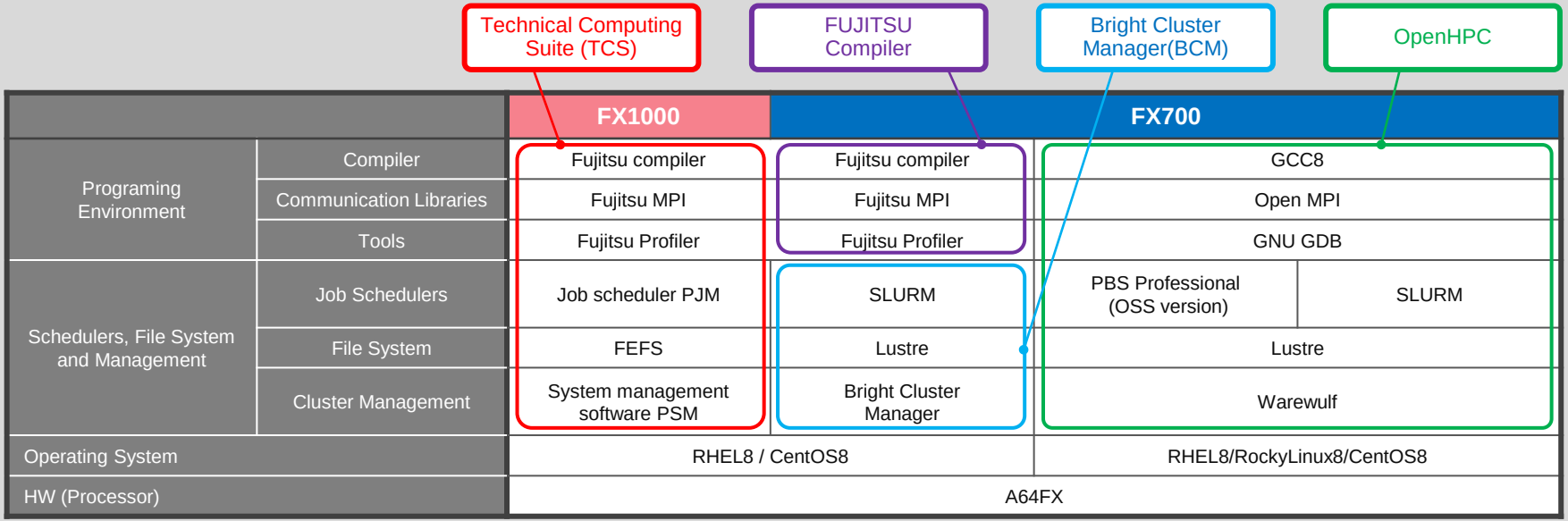
Fujitsu Software Environment

- **Software stack**
- **Fujitsu Language System**
- **Advantages of Fujitsu Compilers**

Software Stack for A64FX System

Both Fujitsu & OSS vendor SW stack are available for application development

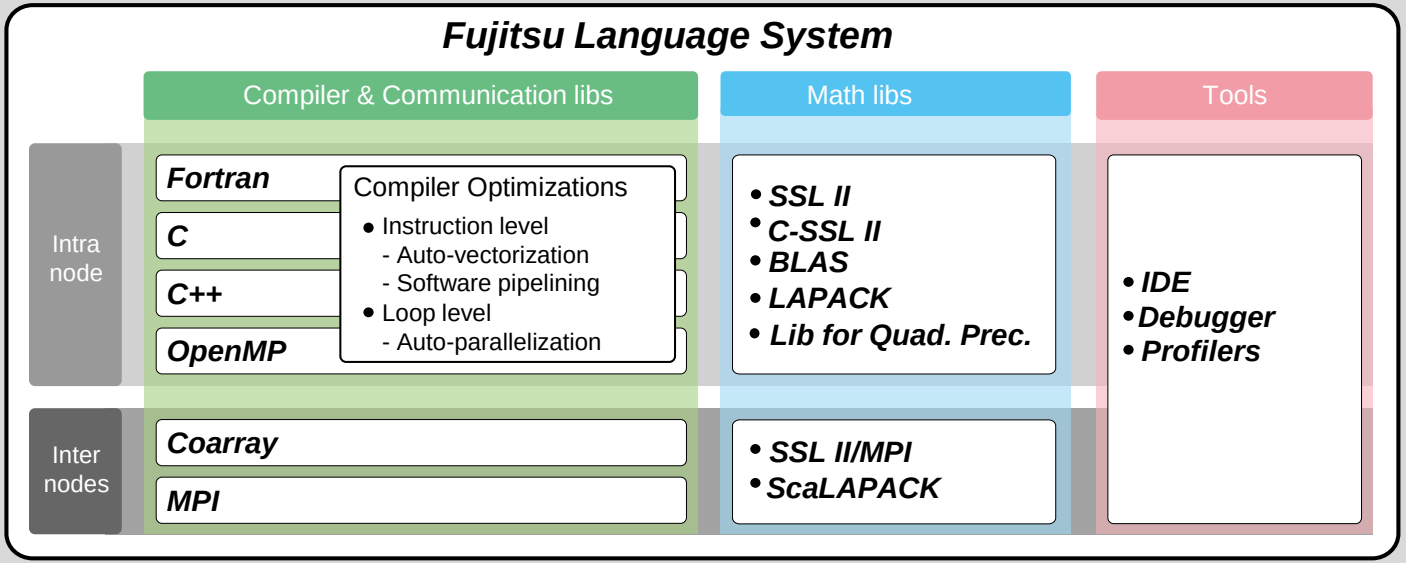
- FJ compilers are optimized for the u-architecture, maximizing SVE and HBM2 performance
- We collaboratively work with RIKEN / Linaro / OSS / ISVs and contribute to Arm HPC ecosystem



Fujitsu Language System

Fujitsu Language system

- Develops a variety of programming tools for various programming models
- Designs and develops Software that exploits Hardware performance



Advantages of Fujitsu Compilers

FUJITSU

Advanced optimizations to accelerate applications

- Proven vectorization technologies to utilize Armv8-A with SVE
- Software pipelining improves instruction-level parallelism to get objects suitable for micro-architecture

Language standard support

- Continuing to support new standards of Fortran, C/C++ and gnu c extension

Support multilevel parallelization

- Auto-parallelization technologies and OpenMP support for thread-level parallelism
- MPI and Fortran Coarray support for process-level parallelism

Fujitsu compiler for C/C++ has two modes (Trad mode and Clang mode) to achieve high performance in a wide range of applications

- Trad mode: Fujitsu's original compiler
- Clang mode: Clang/LLVM compiler with Fujitsu's enhancement

Fujitsu Compiler Commands

FUJITSU

Compiler can be used in two modes:

- has a cross compiler on x86 front-end servers
- native on the A64FX compute nodes

Cross-compiler on an x86 server

Language	Command
Fortran	`frtpx [option list] [file list]`
C	`fccpx [options list] [file list]`
C++	`FCCpx [options list] [file list]`

Native-compiler on A64FX

Language	Command
Fortran	`frt [option list] [file list]`
C	`fcc [options list] [file list]`
C++	`FCC [options list] [file list]`

A64FX platforms

FUJITSU

FX1000

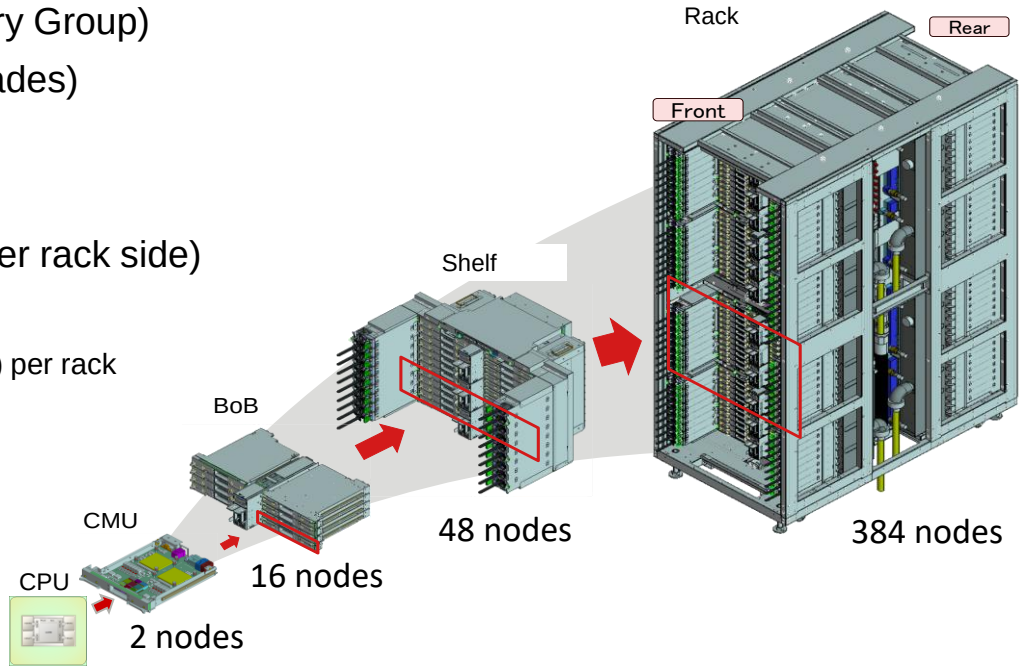
FX700

FX1000 integrated system

FX1000 rack integration

Rack characteristics

- 2 nodes per CMU (Core Memory Group)
- 8 CMU's per BoB (Bunch of Blades)
- 16 nodes per BoB
- 3 BoB's per shelf
- 8 shelves per rack (4 shelves per rack side)
- 384 nodes per rack
  - 1.18PF (2 GHz) or 1.298PF (2.2GHz) per rack



CMU: CPU Memory Unit

FUJITSU

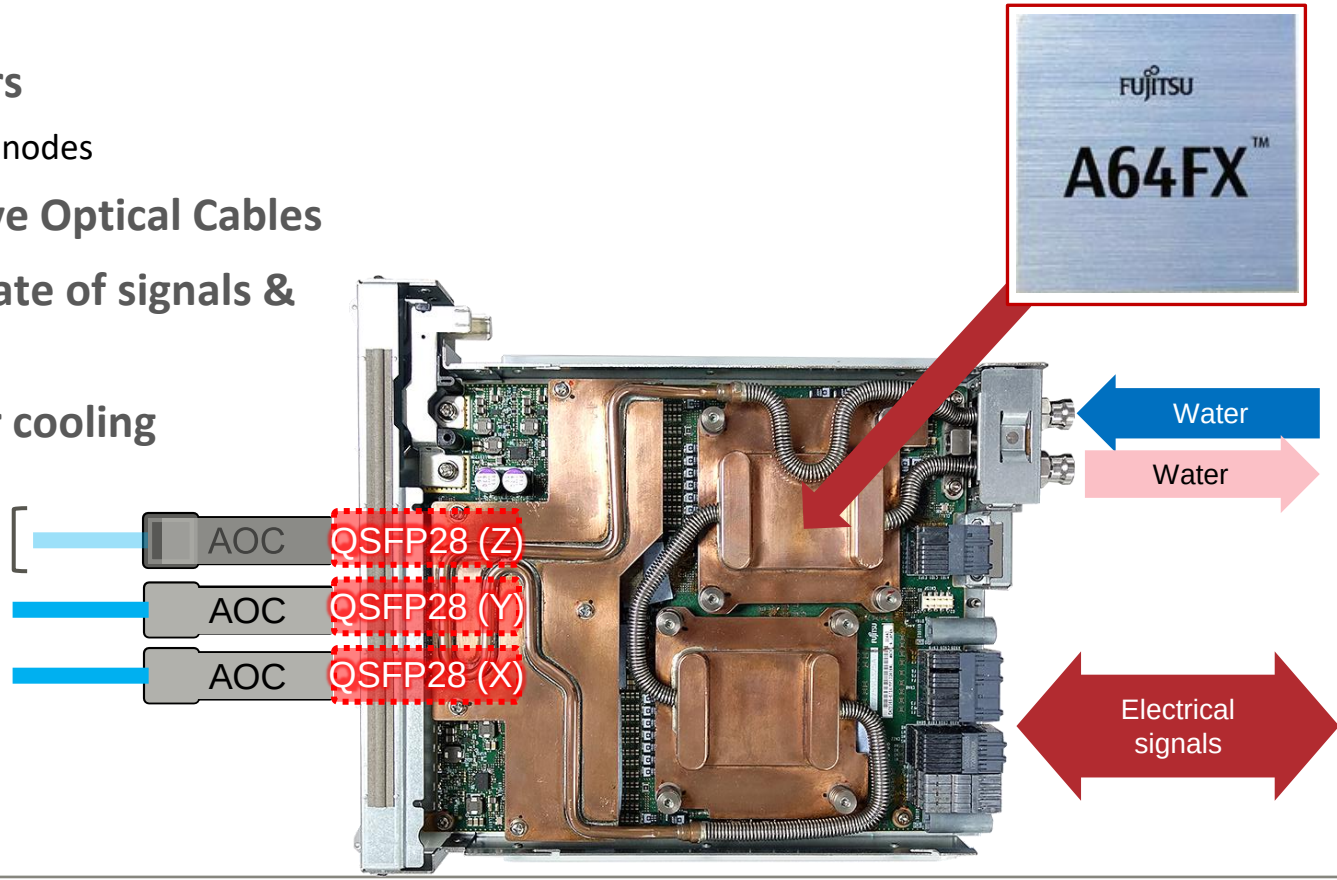
2x A64FX Processors

- Run as 2 separate nodes

QSFP28 x3 for Active Optical Cables

Single-side blind mate of signals & water

~100% direct water cooling

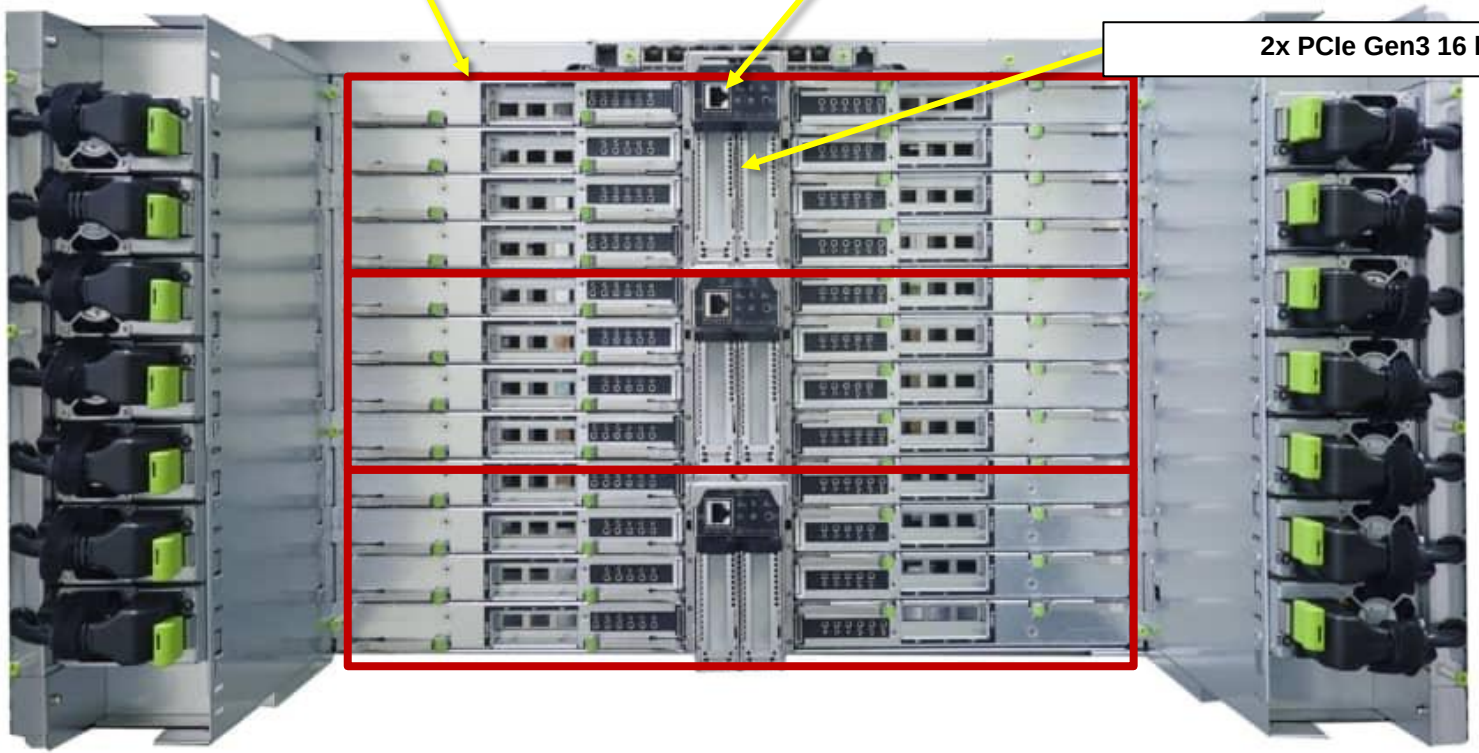


FX1000 shelf configuration

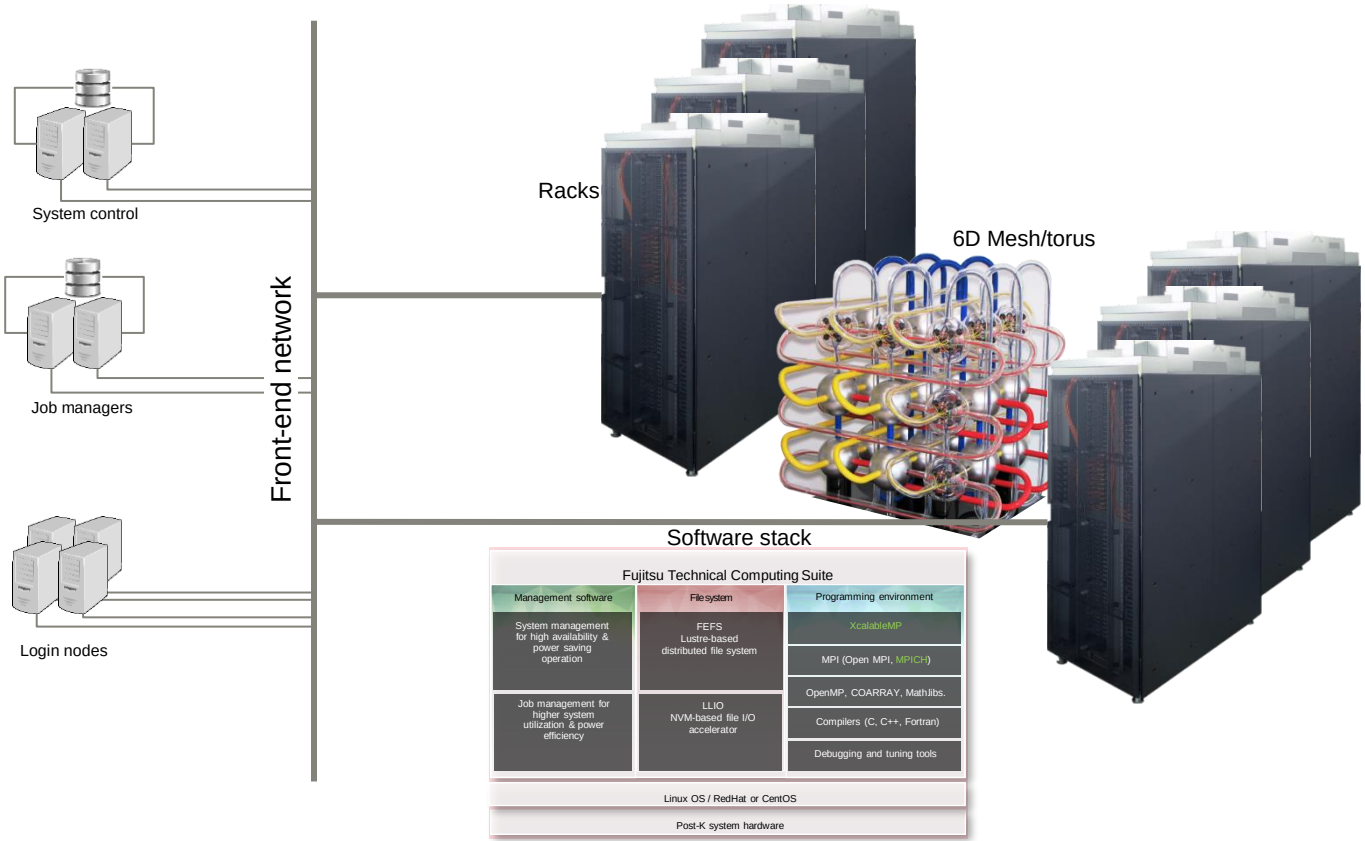
BoB – Bunch of blades
Supports 8 blades each with 2 nodes

1 Ethernet port per Bob for management/IPMI

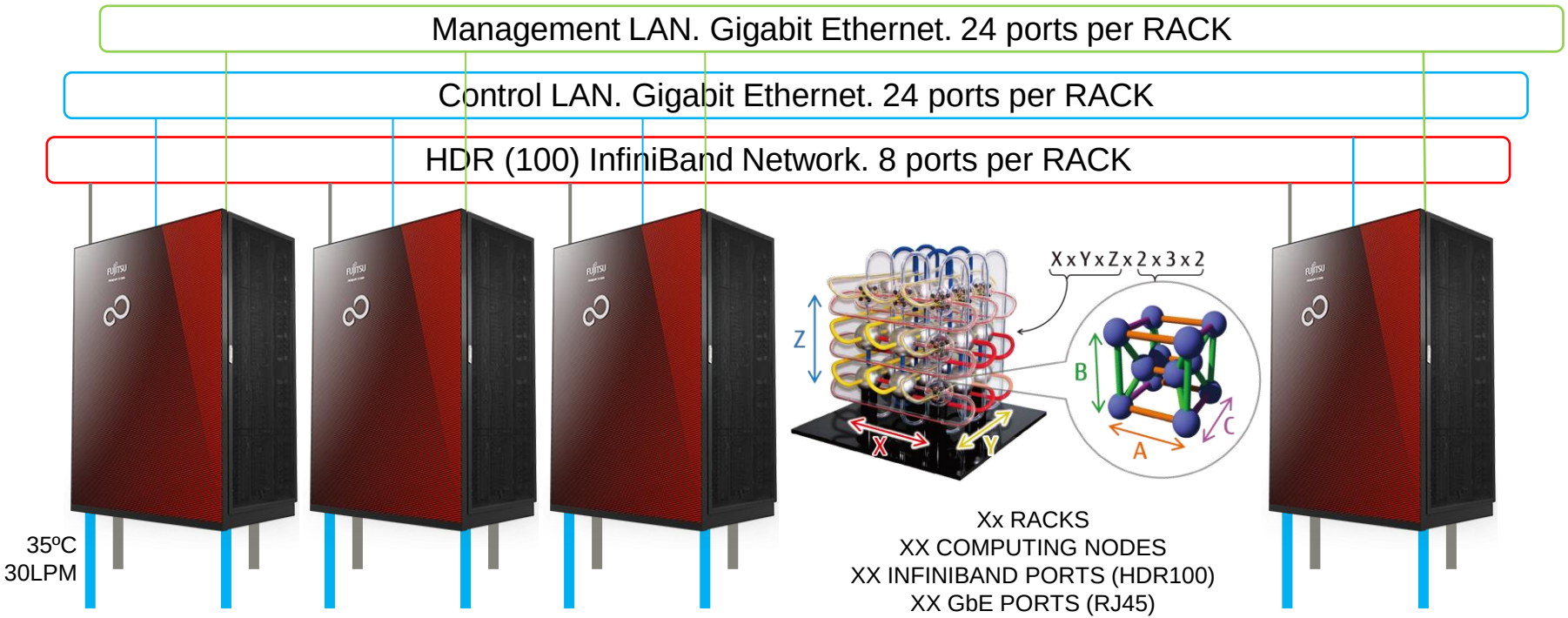
2x PCIe Gen3 16 lane slots



FX-1000 system



Hardware Configuration Overview



Tofu Interconnect (1/2)

FUJITSU

Architecture

- 6D Mesh/Torus topology (Node axis (x, y, z, a, b, c))
- Arbitrary combination of (x, y, z) axis and (a, b, c) axis enables the application to view as simple 3D torus

Bandwidth per link: 6.8GB/s x bi-direction

Concurrent data transfer: 40GB/s x bi-direction

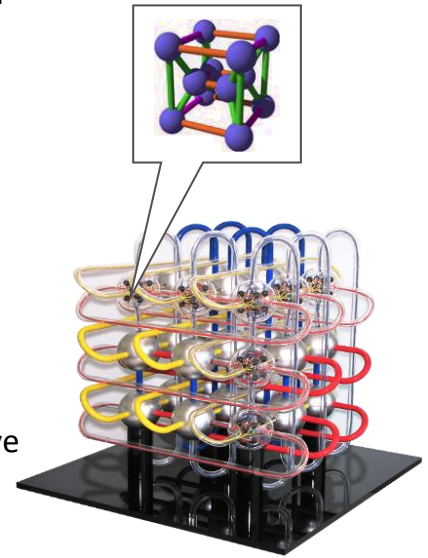
Hardware support of collective communication (Barrier, Reduction)

- Hardware Reduction can support up to 3 elements of double precision data

Features

- Approx. 390K node scalability with direct interconnection network for massive parallel jobs
- Overcome issues with normal 3D torus topology to realize fault-tolerant high operability interconnect
- Non-blocking DMA engine to enable overlap of compute and communication

Tofu unit
(12 Node)

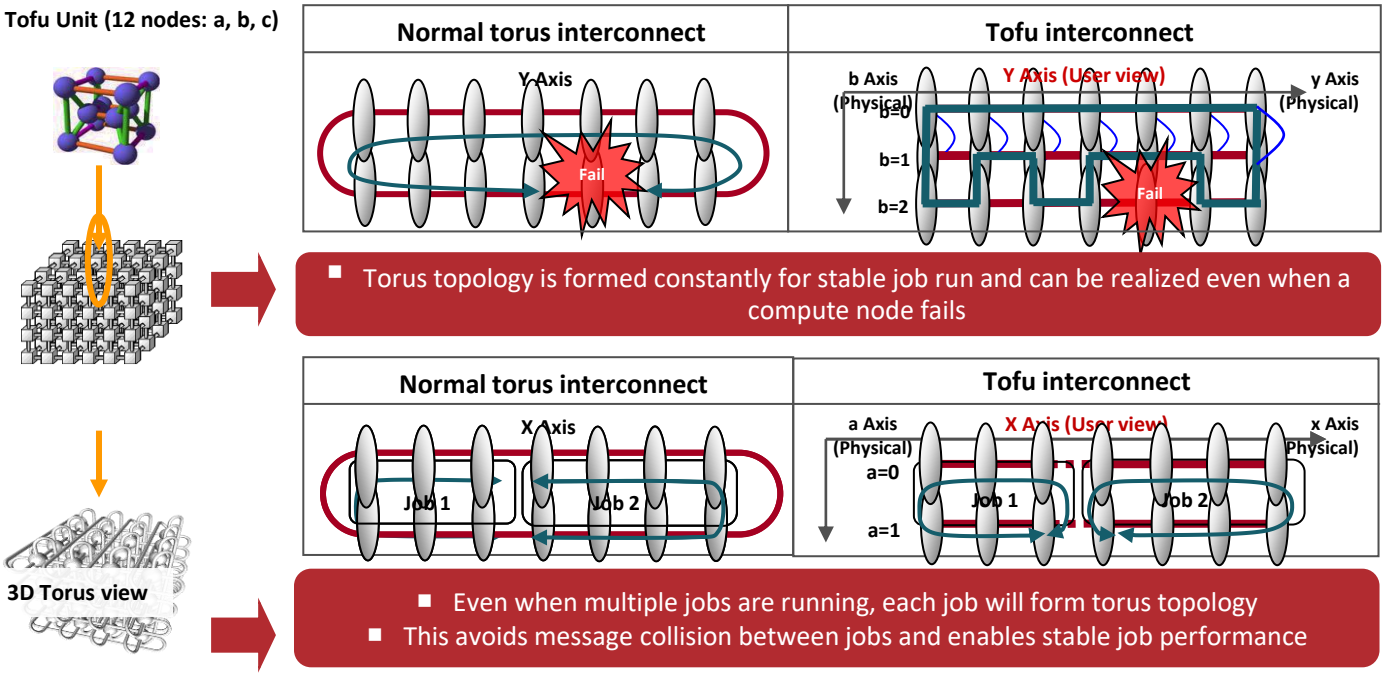


Tofu Interconnect Concept

Tofu Interconnect (2/2)

Fault-tolerant high operability interconnect

- Formation of torus topology avoiding a failed node

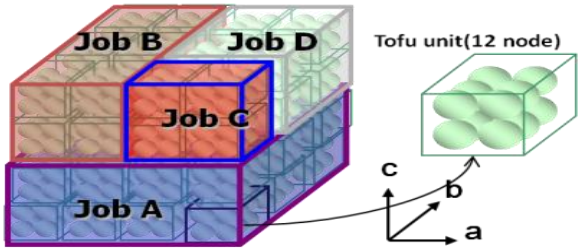


System/Job optimized node allocation

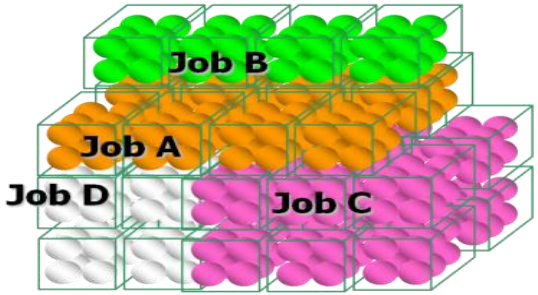
FUJITSU

Allocation of nodes for an MPI parallel job

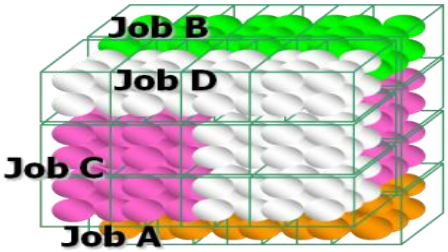
- Rectangular job node allocation
  - Guarantees communication performance between the adjacent nodes of the job
  - Prevents communication conflict with other jobs
  - Single node failure only aborts one job
- Non-Rectangular job node allocation
  - Can use non-adjacent nodes
  - Increases system node occupancy
  - Can be affected by node failures outside the job allocated nodes
  - Shares communication bandwidth with other jobs



Rectangular node allocation for jobs



Non-Rectangular node allocation for jobs

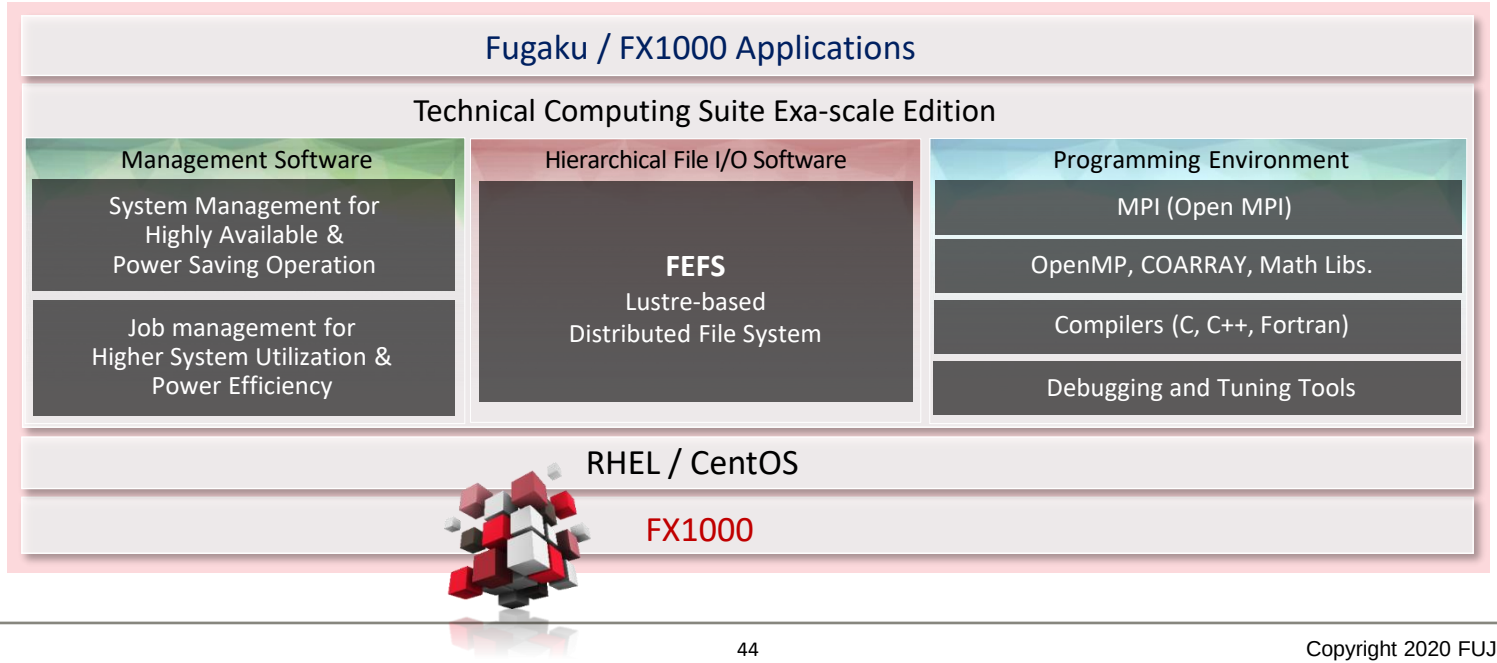


FX100 Exa-scale Edition Software Stack

FUJITSU

Software Stack based on FUJITSU Technical Computing Suite

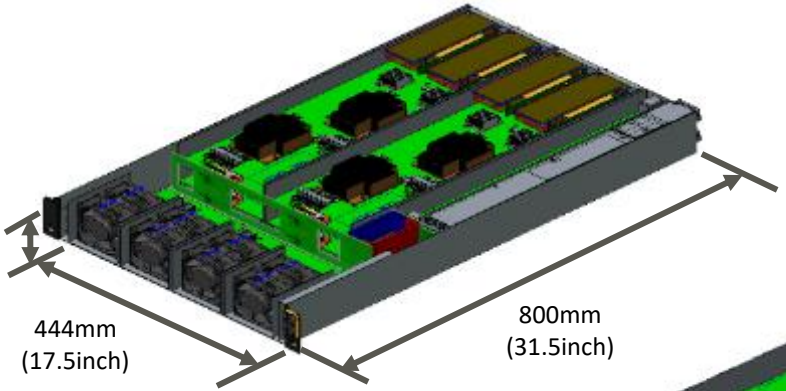
- Exa-scale Edition Software Stack will utilize a Linux-based operation system
- System management provide provisioning, deployment function and full supervision



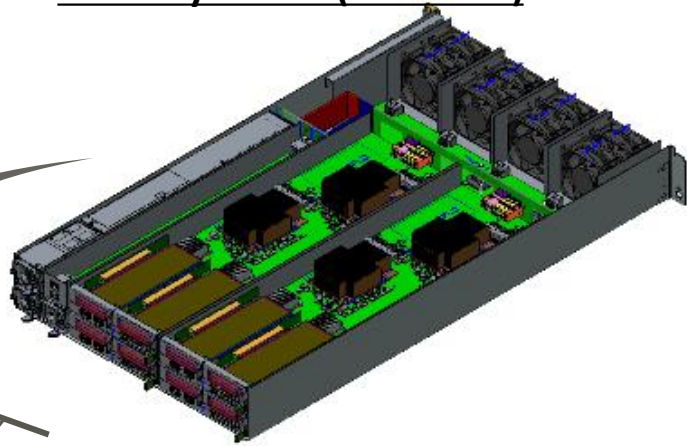
FX700 - 2U system

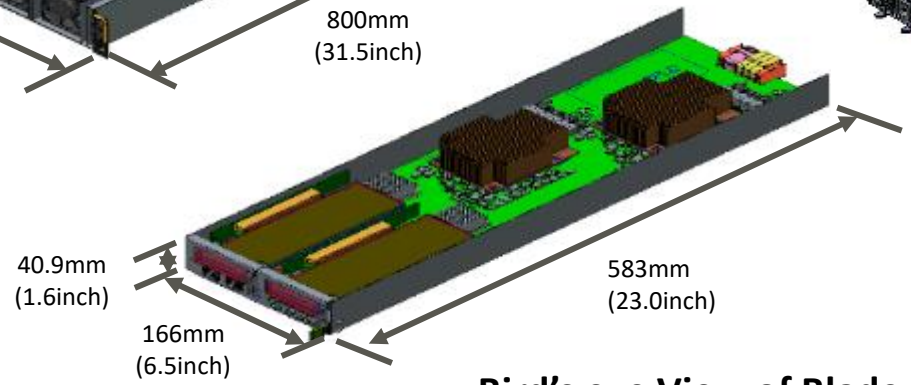
FX700 Chassis

Bird's eye View (Front side)



Bird's eye View (Rear side)

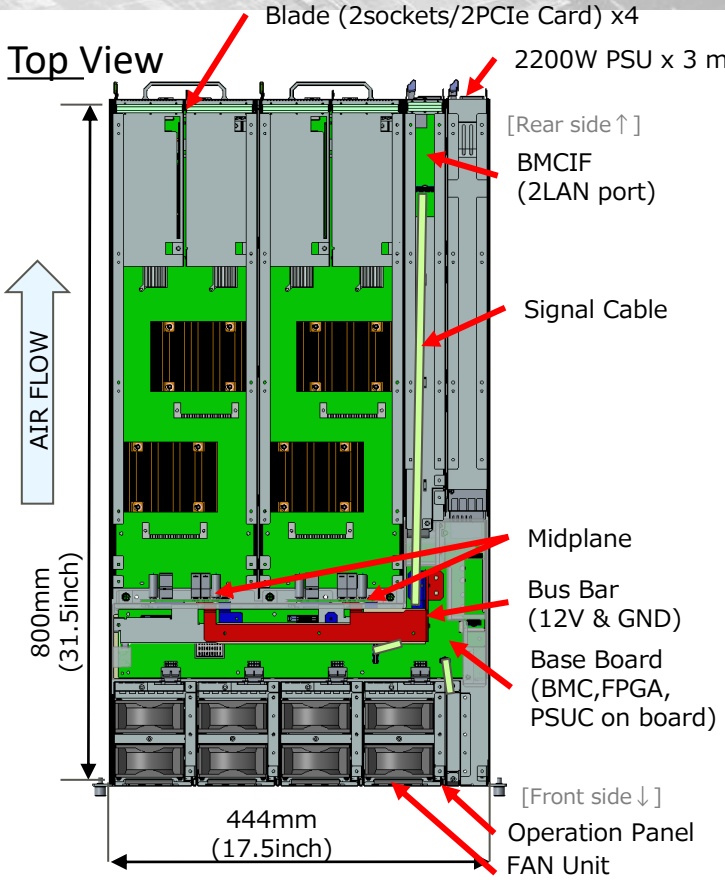




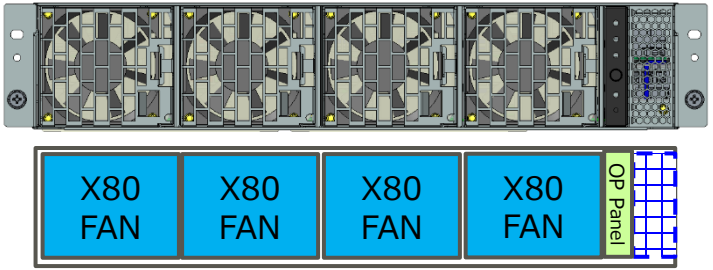
Bird's eye View of Blade

FX700 Chassis

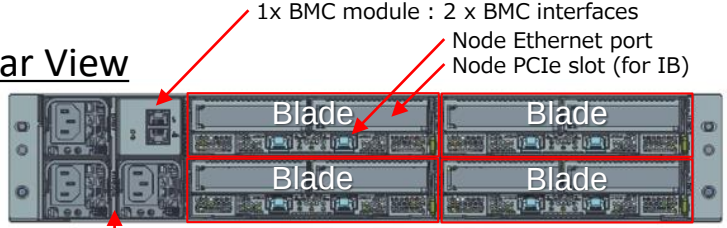
Top View



Front View

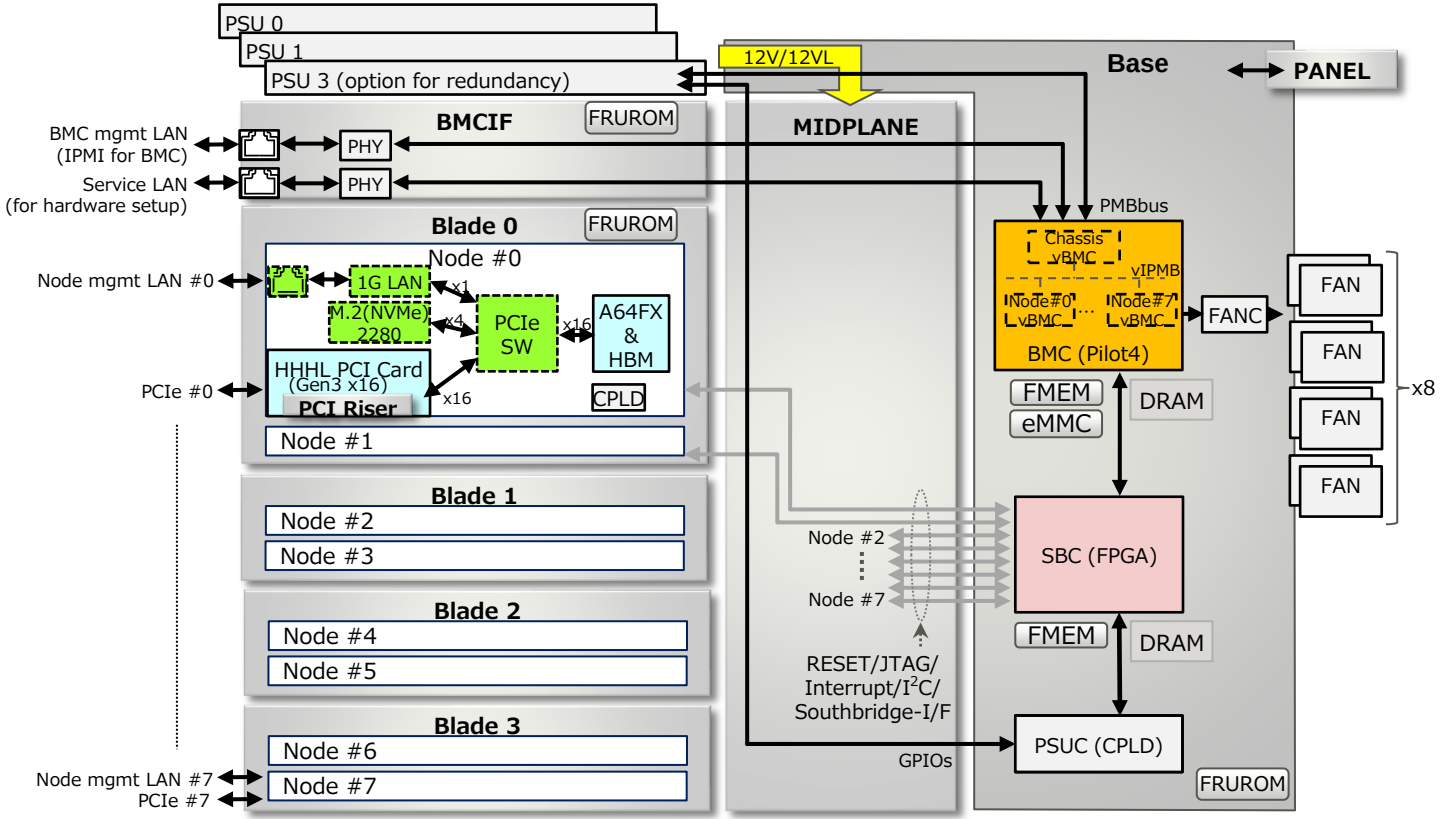


Rear View



3x PSU slots: 2200W PSU (200-240VAC)
(2+1 redundant configuration)
PCIe slot: HHHL (max. 167.65mm in length)

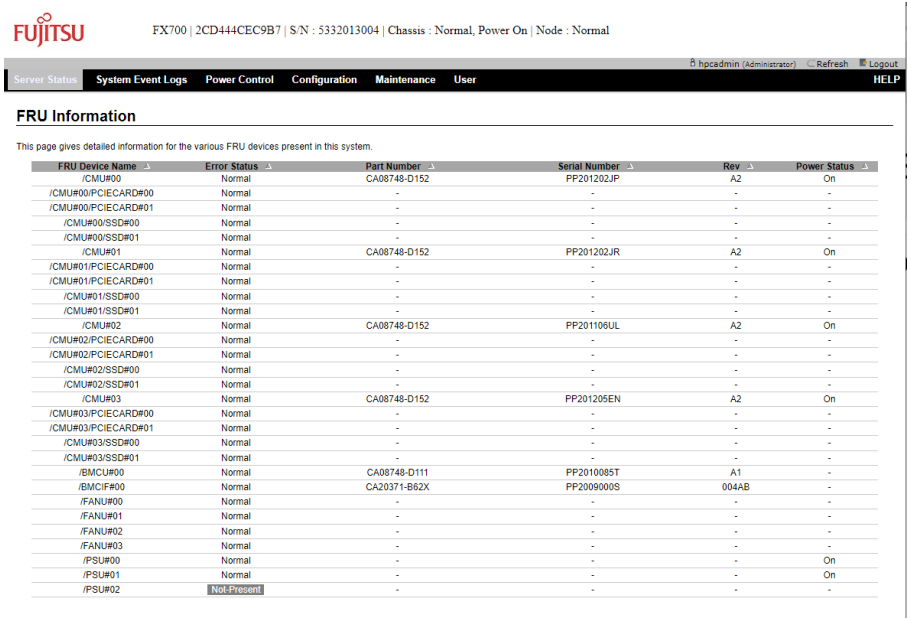
FX700 Block Diagram



FX700 BMC interface

Web portal

- Server Status
- System Event Logs (can be downloaded)
- Power Control – control power and boot option (disk, network, UEFI shell) for each node
- Configuration (Network, SNMP, NTP …etc)
- Maintenance – update firmware, change mode of components to enable maintenance
- User management – control who can connect to the BMC



FRU Device Name	Error Status	Part Number	Serial Number	Rev	Power Status
/CMU#00	Normal	CA08748-D152	PP201202JP	A2	On
/CMU#00/PCIECARD#00	Normal	-	-	-	-
/CMU#00/PCIECARD#01	Normal	-	-	-	-
/CMU#00/SSD#00	Normal	-	-	-	-
/CMU#00/SSD#01	Normal	-	-	-	-
/CMU#01	Normal	CA08748-D152	PP201202JR	A2	On
/CMU#01/PCIECARD#00	Normal	-	-	-	-
/CMU#01/PCIECARD#01	Normal	-	-	-	-
/CMU#01/SSD#00	Normal	-	-	-	-
/CMU#01/SSD#01	Normal	-	-	-	-
/CMU#02	Normal	CA08748-D152	PP201106UL	A2	On
/CMU#02/PCIECARD#00	Normal	-	-	-	-
/CMU#02/PCIECARD#01	Normal	-	-	-	-
/CMU#02/SSD#00	Normal	-	-	-	-
/CMU#02/SSD#01	Normal	-	-	-	-
/CMU#03	Normal	CA08748-D152	PP201205EN	A2	On
/CMU#03/PCIECARD#00	Normal	-	-	-	-
/CMU#03/PCIECARD#01	Normal	-	-	-	-
/CMU#03/SSD#00	Normal	-	-	-	-
/CMU#03/SSD#01	Normal	-	-	-	-
/BMCU#00	Normal	CA08748-D111	PP2010085T	A1	-
/BMCIF#00	Normal	CA20371-B62X	PP2009000S	004AB	-
/FANU#00	Normal	-	-	-	-
/FANU#01	Normal	-	-	-	-
/FANU#02	Normal	-	-	-	-
/FANU#03	Normal	-	-	-	-
/PSU#00	Normal	-	-	-	On
/PSU#01	Normal	-	-	-	On
/PSU#02	Not-Present	-	-	-	-

FX700 chassis specification

FUJITSU

A64FX 2U Scalable Computing Server

- 8 nodes per 2U chassis
- Node characteristics
  - 48-core Armv8-A SVE (512bit SIMD)
  - 32GB HBM2
  - IB HDR100

FCS : 2020.3.31

High-performance

- 2.7 or 3.072 TF(DP) x 8 CPU

Low power consumption & Air cooled: ~2.6kW

Specifications	
Chassis	2U / 8 nodes maximum
CPU (A64FX)	2~8
Main memory (HBM2)	32GB (8GB x 4, on package)/CPU
DIMM	None
Internal boot disk	M.2 2280 NVMe SSD x1 /CPU (Option)
Node mgmt. LAN	1GbE x 1 port / CPU
PCIe slots	1 / CPU (Gen3 x16 lanes, HHHL, max 25W) Supports InfiniBand HDR100 (ConnectX6)
BMC Control/Service network	1GbE x 2 port / chassis
AC input (Freq.)	Base : 2,200W, AC 200-240V(50-60Hz) x 2 (non redundant) Option : 2,200W, AC 200-240V(50-60Hz) x 3 (2+1 redundant)
Dimension (W x D x H)	444mm x 800mm x 86.5mm
Weight	40kg
Operating conditions	Ambient: 5-35℃, Humidity: 20-80% (not condensed)
Safety	UL, CSA, CE
RoHS	RoHS2

FX700 Software Stack options

Commercial and OSS SW stack offerings

- FJ compilers are optimized for the μ-architecture, maximizing SVE and HBM2 performance
- We collaboratively work with RIKEN / Linaro / OSS / ISVs and contribute to Arm HPC ecosystem

		Bright Cluster Manager	OpenHPC
Programing Environment	Compiler	Fujitsu compiler GCC11	Fujitsu compiler GCC11
	Communication Libraries	Fujitsu MPI OpenMPI	Fujitsu MPI Open MPI
	Debuggers	GNU GDB	GNU GDB
Schedulers, File System and Management	Job Schedulers	SLURM/OpenPBS	SLURM/OpenPBS
	File System	Lustre/BeeGFS	Lustre/BeeGFS
	Cluster Management	**Bright Cluster Manager**	**Warewulf**
Operating System		**RHEL8.x** / Rocky Linux 8.x	RHEL / Rocky Linux 8.x
HW (Processor)		**A64FX**	

The future – Fujitsu Monaka processor

- **ARMV9-A SVE2 and 3D stacking**

3D microarchitecture – High Performance

FUJITSU

Fujitsu microarchitecture | **3D many-core architecture** | **Confidential Computing**

High-performance

- Cloud native 3D many-core design by Fujitsu's proven microarchitecture
- High memory bandwidths

Energy Efficient

- Leading-edge process technology
- Ultra-low voltage operation

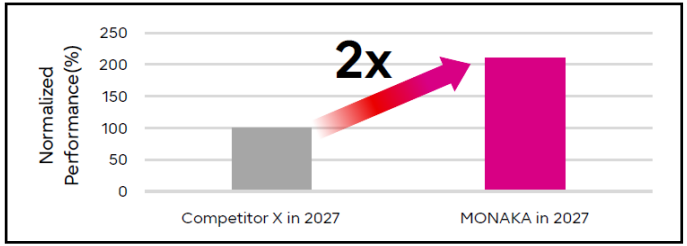
High Reliability

- Multiple VM Confidential Computing
- Mainframe class RAS for stable operation

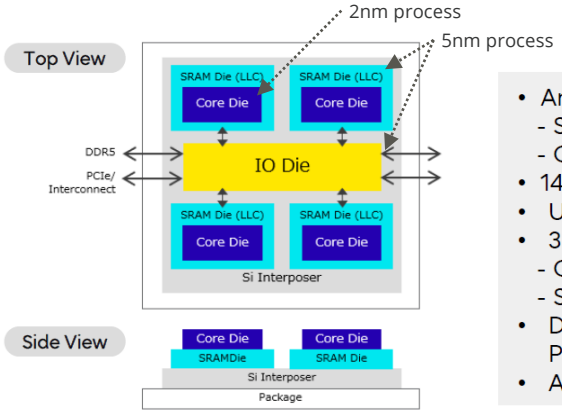
Easy to Use

- Open & de facto standard software stacks
- Fujitsu compiler technology
- Air cooling for easy deployment

Performance per Watt



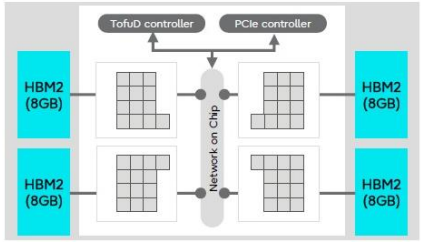
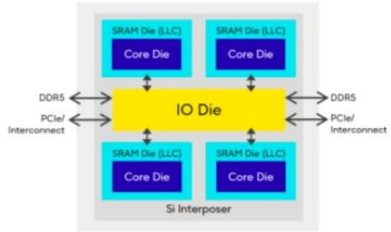
This presentation is based on results obtained from a project subsidized by the New Energy and Industrial Technology Development Organization (NEDO).



subject to change

- Armv9-A architecture
  - SVE2 for AI and HPC
  - Confidential Computing for security
- 144 cores x 2 sockets
- Ultra low voltage
- 3D chiplet
  - Core die 2nm
  - SRAM die/IO die 5nm
- DDR 12 channels
  PCI Express6.0(CXL3.0)
- Air cooling

Comparison – A64FX and Fujitsu-Monaka

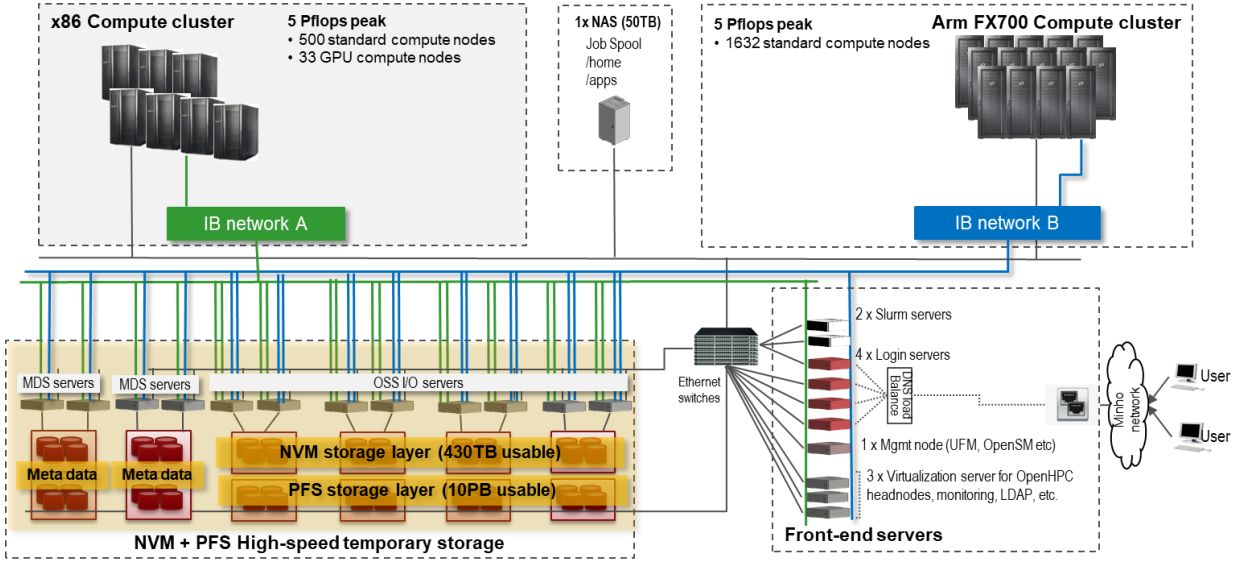
A64FX	FUJITSU-MONAKA
	
Armv8-A Architecture - SVE for HPC and AI	Armv9-A Architecture - SVE2 enhanced for HPC and AI - Confidential Computing
48 cores x 1 socket (48 cores per node)	144 cores x 2 sockets (288 cores per node)
Low voltage	Ultra low voltage
2.5D - CPU 7nm - HBM2	3D chiplet - Core die 2nm - SRAM die/IO die 5nm
HBM2 4 channels	DDR5 12 channels
PCI Express 3.0 Tofu Interconnect	PCI Express 6.0 (CXL3.0)
Air cooling and water cooling	Air cooling

Deucalion system

- **Hybrid cluster with a unified storage and job subsystem**

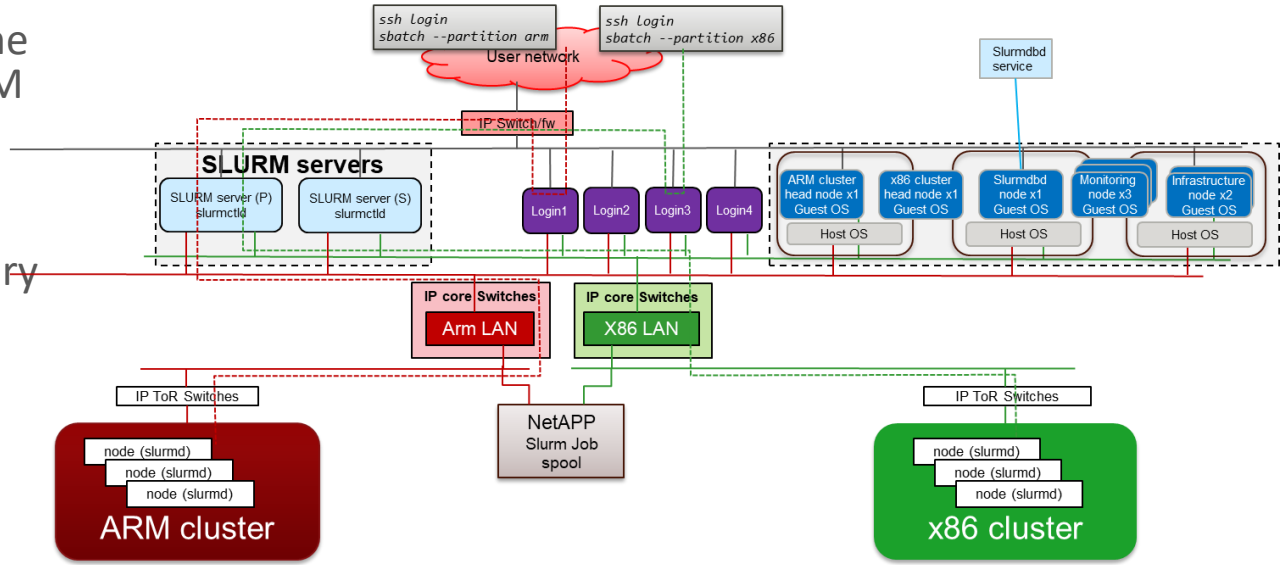
Deucalion Architecture

- Hybrid cluster with ARM, x86 , x86+GPU
- ARM compute cluster
  - 5 Pflops peak – A64FX
  - 1632 x Compute Nodes single processor
- x86 compute cluster
  - 5 Pflops peak - AMD Rome 7742
  - 530 x standard Compute Nodes (dual proc)
  - 33 x compute with GPU (NVIDIA A100)
- High speed temporary storage (430TB+10PB)
  - High speed storage with both an NVM tier and a traditional PFS disk based tier
- NAS shared storage (50TB)
  - Highly reliable NAS for common user files (homes, apps, job spool etc).
- Front-end servers
  - Common set of front-end servers (login nodes, Slurm nodes, management nodes).
  - Use of virtualized servers or containers for cluster management, monitoring and LDAP.



Job Scheduler - SLURM

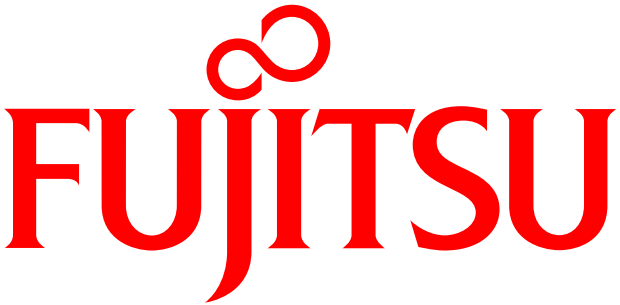
- SLURM manages both ARM and x86 clusters
- Users specify a partition name in their job scripts and SLURM allocates the job to nodes of the appropriate cluster accordingly
- SLURM uses HA with a primary and secondary server
- Shared job spool resides on the NetApp NFS server
- Slurmdbd runs on the slurmdbd VM
- slurmd runs on all the compute nodes



Abbreviations

A64FX

- B/W: Band Width
- BF: Band width-to-Flop
- BGA: Ball Grid Array
- CSE: Commit Stack Entry
- EAG: Effective Address Generator
- EX: Integer EXecution unit
- FL: FLoating-point execution unit
- PFPR: Physical Floating-Point Register
- PGPR: Physical General-Purpose Register
- PPR: Physical Predicate Register
- PRX: PRedicate eXecution unit
- RSA: Reservation Station for Address generation
- RSBR: Reservation Station for Branch
- RSE: Reservation Station for Execution
- Tofu: Torus-Fusion



shaping tomorrow with you