Fujitsu High Performance CPU for the Post-K Computer

August 21st, 2018
Toshio Yoshida
FUJITSU LIMITED

FUJITSU
shaping tomorrow with you

Key Message

- A64FX is the new Fujitsu-designed Arm processor

  - It is used in the post-K computer

- A64FX is the first processor of the Armv8-A SVE architecture

  - Fujitsu, as a lead partner, collaborated closely with Arm on the development of SVE

- A64FX achieves high performance in HPC and AI areas

  - Our own microarchitecture maximizes the capability of SVE

Outline

- Fujitsu Processor Development

- A64FX
  - Overview
  - Microarchitecture
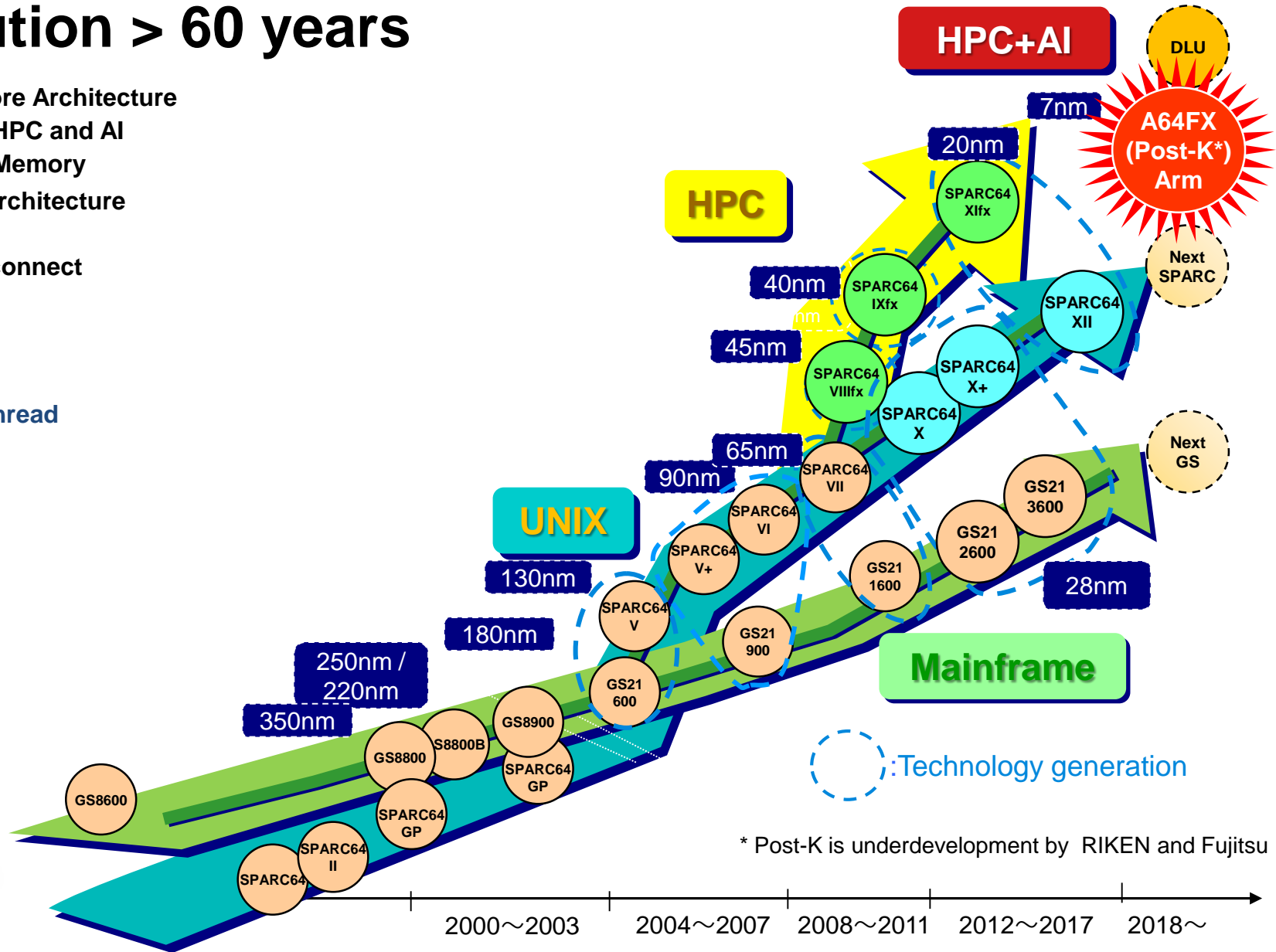 - Performance
  - Power Management
  - RAS

- Software Development

- Summary

Fujitsu Processor Development

Persistent Evolution > 60 years

HPC+AI

DLU

7nm

A64FX (Post-K*) Arm

Performance

Scalable Many-core Architecture
512-bit SIMD for HPC and AI
High Bandwidth Memory
Virtual Machine Architecture
Software on Chip
High-speed Interconnect

HPC-ACE
System on Chip
Hardware Barrier

Multi-core Multi-thread

L2$ on Die

Non-Blocking $
O-O-O Execution
Super-Scalar
Single-chip CPU

Store Ahead
Branch History
Prefetch

Reliability

$ ECC
Register/ALU Parity
Instruction Retry
$ Dynamic Degradation
Error Checkers/History

HPC

20nm

SPARC64 XIfx

Next SPARC

40nm

SPARC64 IXfx

SPARC64 XII

45nm

SPARC64 VIIIfx

SPARC64 X+

SPARC64 X

65nm

SPARC64 VII

Next GS

90nm

SPARC64 VI

GS21 3600

UNIX

SPARC64 V+

GS21 2600

130nm

SPARC64 V

GS21 1600

Mainframe

GS21 900

180nm

28nm

250nm / 220nm

GS21 600

350nm

GS8900

GS8800 S8800B

SPARC64 GP

GS8600

SPARC64 GP

◯ :Technology generation

SPARC64 II

* Post-K is underdevelopment by RIKEN and Fujitsu

SPARC64

2000～2003   2004～2007   2008～2011   2012～2017   2018～

DNA of Fujitsu Processors

A64FX inherits DNA from Fujitsu technologies used in the mainframes, UNIX and HPC servers

MF

High reliability

Stability
Integrity
Continuity

UNIX

High speed & flexibility

Thread performance
Software on Chip
Large SMP

HPC

High performance-per-watt

Execution and memory throughput
Low power
Massively parallel

(C) RIKEN

A64FX™

CPU w/ extremely high throughput

High performance
Massively parallel
Low power
Stability and integrity

A64FX Designed for HPC/AI

A64FX = CPU with extremely high throughput

1. High Performance ♔

HPC/AI apps.   >> General purpose CPU
Various data types
(FP64/32/16,   INT64/32/16/8)

2. High Throughput ♔

Vector   : 512-bit wide SIMD x 2 pipes /core
Memory   : HBM2 (extremely high B/W)
Scalable : 48 cores, Tofu interconnect

3. High Efficiency ♔

Performance
   (D|S|H)GEMM   >90%
   Stream Triad   >80%
Perf-per-watt   >> General purpose CPU

4. Standard ♔

Binary compatibility with
Armv8.2-A + SVE + SBSA* level3

*Arm's "Server Base System Architecture"

A64FX Chip Overview

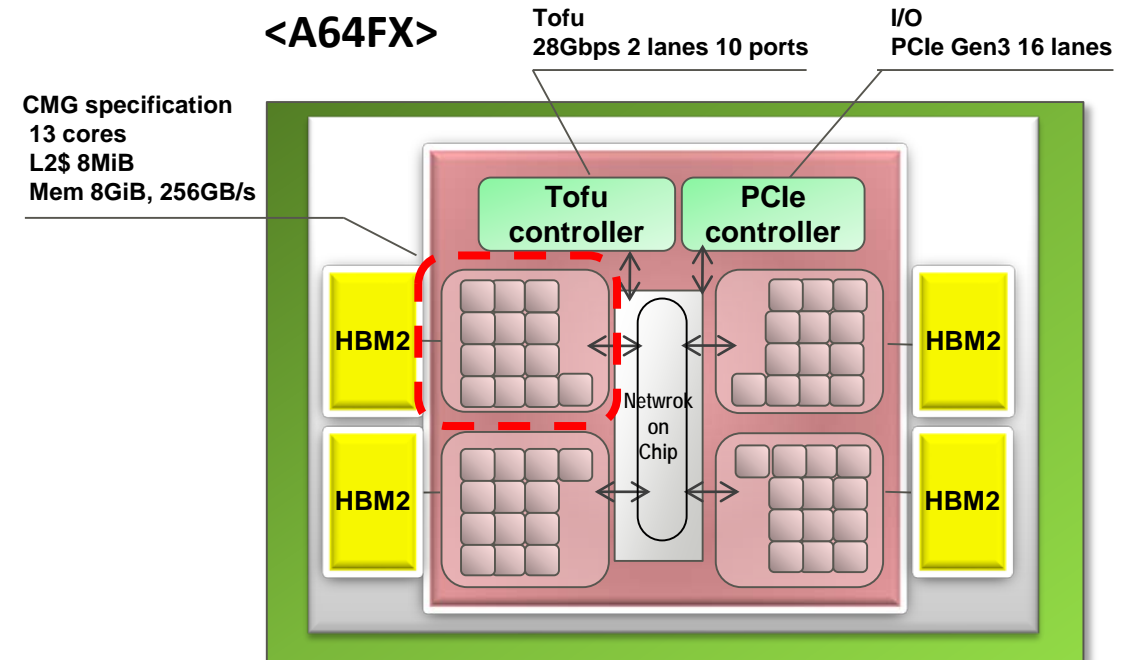## ■ Architecture Features

- Armv8.2-A (AArch64 only)

- SVE 512-bit wide SIMD

- 48 computing cores + 4 assistant cores*

  *All the cores are identical

- HBM2   32GiB
- Tofu   6D Mesh/Torus
               28Gbps x 2 lanes x 10 ports

- PCIe Gen3 16 lanes

■ 7nm FinFET

- 8,786M transistors
- 594 package signal pins

■ Peak Performance (Efficiency)

- >2.7TFLOPS (>90%@DGEMM)
- Memory B/W 1024GB/s (>80%@Stream Triad)

\<A64FX\>

Tofu
28Gbps 2 lanes 10 ports

I/O
PCIe Gen3 16 lanes

CMG specification
13 cores
L2$ 8MiB
Mem 8GiB, 256GB/s

Tofu controller

PCIe controller

HBM2

HBM2

Netwrok on Chip

HBM2

HBM2

	A64FX (Post-K)	SPARC64 XIfx (PRIMEHPC FX100)
ISA (Base)	Armv8.2-A	SPARC-V9
ISA (Extension)	SVE	HPC-ACE2
Process Node	7nm	20nm
Peak Performance	>2.7TFLOPS	1.1TFLOPS
SIMD	512-bit	256-bit
# of Cores	48+4	32+2
Memory	HBM2	HMC
Memory Peak B/W	1024GB/s	240GB/s x2 (in/out)

A64FX Features

- ■ Collaboration with Arm to develop and optimize SVE for a wide range of applications
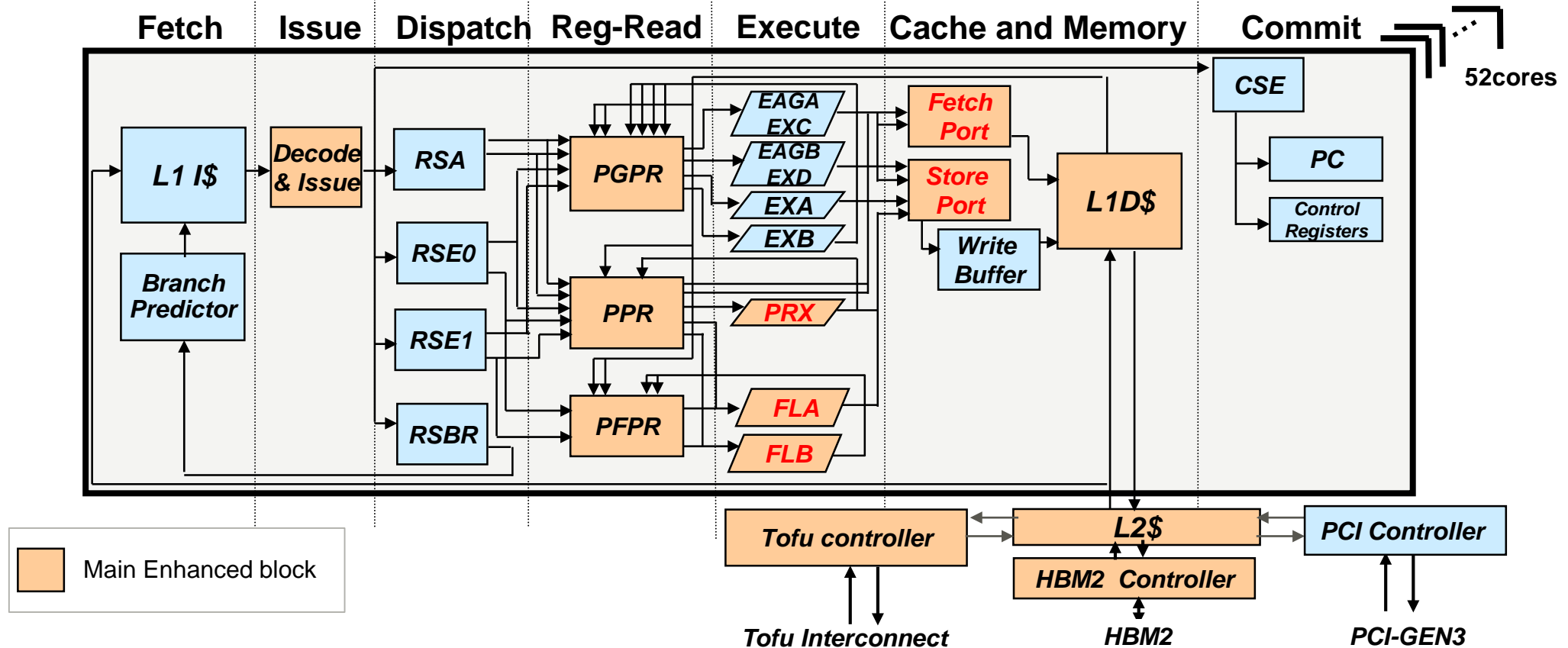 - FP16 and INT16/8 dot product are introduced for AI applications

	A64FX (Post-K)	SPARC64 XIfx (PRIMEHPC FX100)	SPAR64 VIIIfx (K computer)
ISA	Armv8.2-A + SVE	SPARC-V9 + HPC-ACE2	SPARC-V9 + HPC-ACE
SIMD Width	512-bit	256-bit	128-bit
Four-operand FMA	✓ Enhanced	✓	✓
Gather/Scatter	✓ Enhanced	✓	
Predicated Operations	✓ 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	✓

* Utilizing AArch64 implementation-defined system registers

A64FX Core Pipeline

- A64FX enhances and inherits superior features of SPARC64
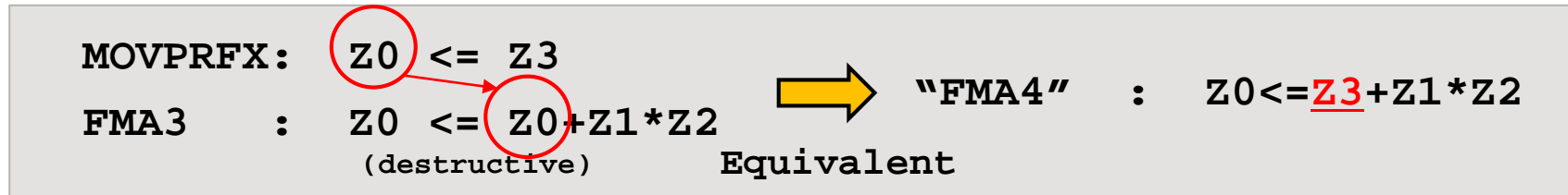 - Inherits superscalar, out-of-order, branch prediction, etc.
  - Enhances SIMD and predicate operations
    - 2x 512-bit wide SIMD FMA + Predicate Operation + 4x ALU (shared w/ 2x AGEN)
    - 2x 512-bit wide SIMD load or 512-bit wide SIMD store

Fetch | **Issue** | **Dispatch** | **Reg-Read** | **Execute** | **Cache and Memory** | **Commit**

52cores

L1 I$
Branch Predictor

Decode & Issue

RSA
RSE0
RSE1
RSBR

PGPR
PPR
PFPR

EAGA EXC
EAGB EXD
EXA
EXB
PRX
FLA
FLB

Fetch Port
Store Port
Write Buffer
L1D$

CSE
PC
Control Registers

Tofu controller
L2$
HBM2 Controller
PCI Controller

Tofu Interconnect
HBM2
PCI-GEN3

Main Enhanced block
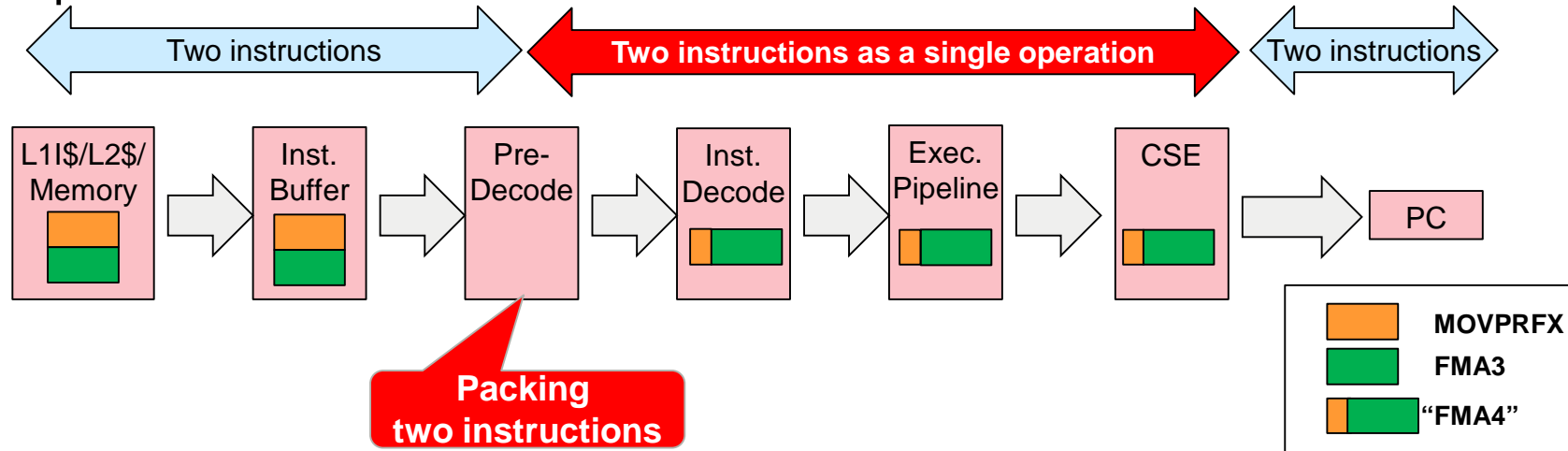
Four-operand FMA with Prefix Instruction

- ## MOVPRFX as a prefix instruction

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

  ```
  MOVPRFX:   Z0 <= Z3                              "FMA4"  :  Z0<=Z3+Z1*Z2

  FMA3    :  Z0 <= Z0+Z1*Z2
             (destructive)    Equivalent
  ```
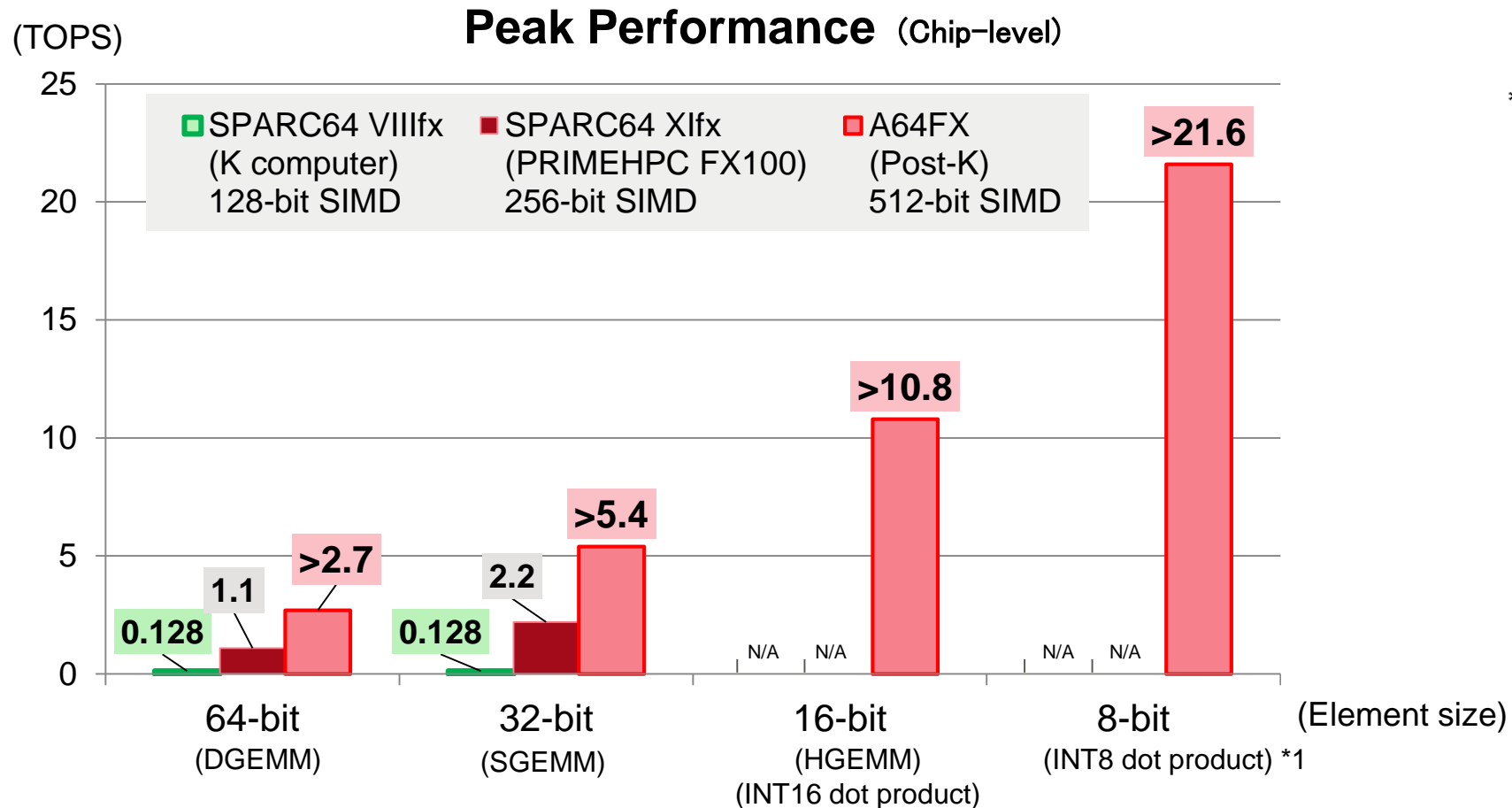
- ## A64FX implementation for MOVPRFX

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

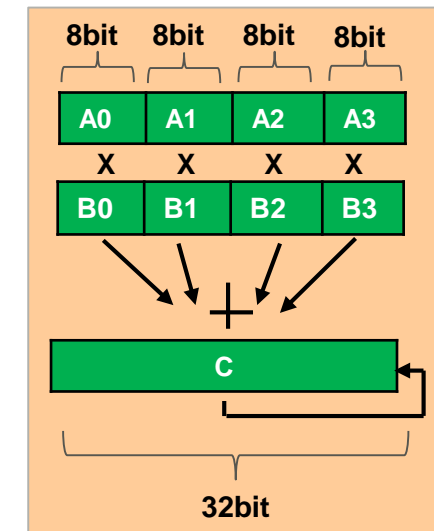 Two instructions | Two instructions as a single operation | Two instructions

  L1I$/L2$/Memory → Inst. Buffer → Pre-Decode → Inst. Decode → Exec. Pipeline → CSE → PC

  Packing two instructions

	MOVPRFX
	FMA3
	"FMA4"

All Rights Reserved. Copyright © FUJITSU LIMITED 2018

9

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

Peak Performance （Chip-level）

(TOPS)

Legend:
- ☐ SPARC64 VIIIfx (K computer) 128-bit SIMD
- ■ SPARC64 XIfx (PRIMEHPC FX100) 256-bit SIMD
- ☐ A64FX (Post-K) 512-bit SIMD

Element size	SPARC64 VIIIfx	SPARC64 XIfx	A64FX
64-bit (DGEMM)	0.128	1.1	>2.7
32-bit (SGEMM)	0.128	2.2	>5.4
16-bit (HGEMM) (INT16 dot product)	N/A	N/A	>10.8
8-bit (INT8 dot product) *1	N/A	N/A	>21.6

(Element size)

*1 **INT8 dot product**

$$C = \Sigma \ (A_i \times B_i) + C$$

8bit 8bit 8bit 8bit

| A0 | A1 | A2 | A3 |

x x x x
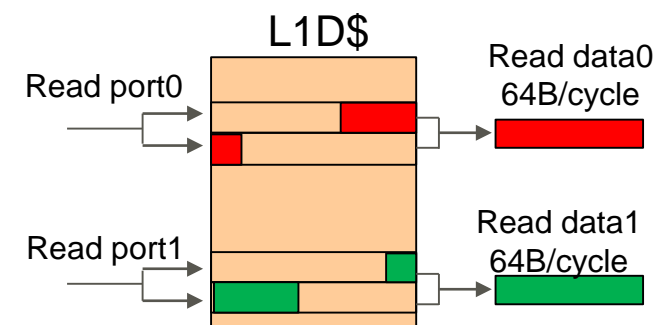
| B0 | B1 | B2 | B3 |

+

| C |

32bit

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

L1D\$

Read port0 → Read data0 64B/cycle

Read port1 → Read data1 64B/cycle

  - "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

< Combined Gather >

128B

Memory | 0 | 1 | 3 | 2 | 6 | 7 | 5 | 4 |

flow-4   flow-3

flow-1   flow-2

Register | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

8B

Throughput/Core [byte/cycle]

16   32

Combined Gather

Gather (Normal)   2x faster
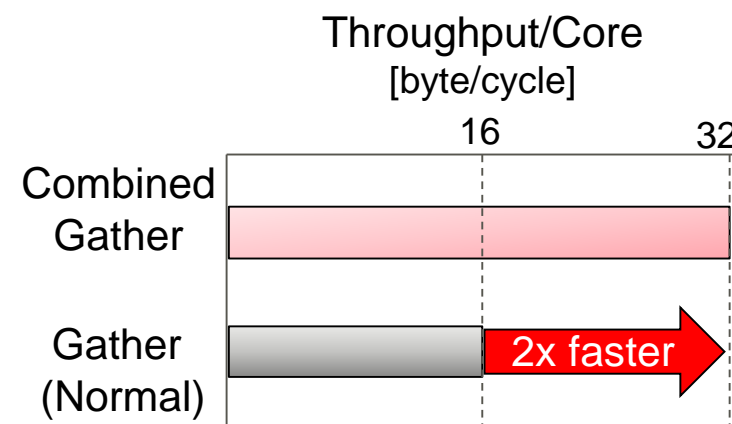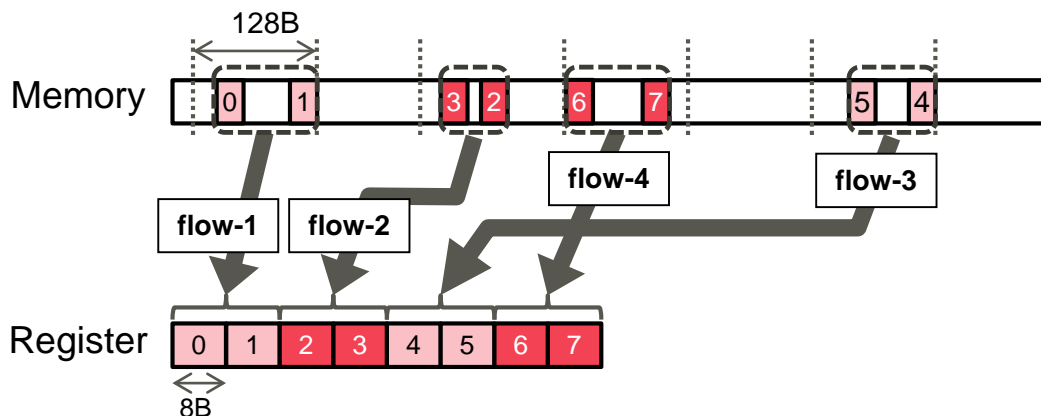
Many-Core Architecture

- ■ A64FX consists of four CMGs (Core Memory Group)
  - A CMG consists of 13 cores, an 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
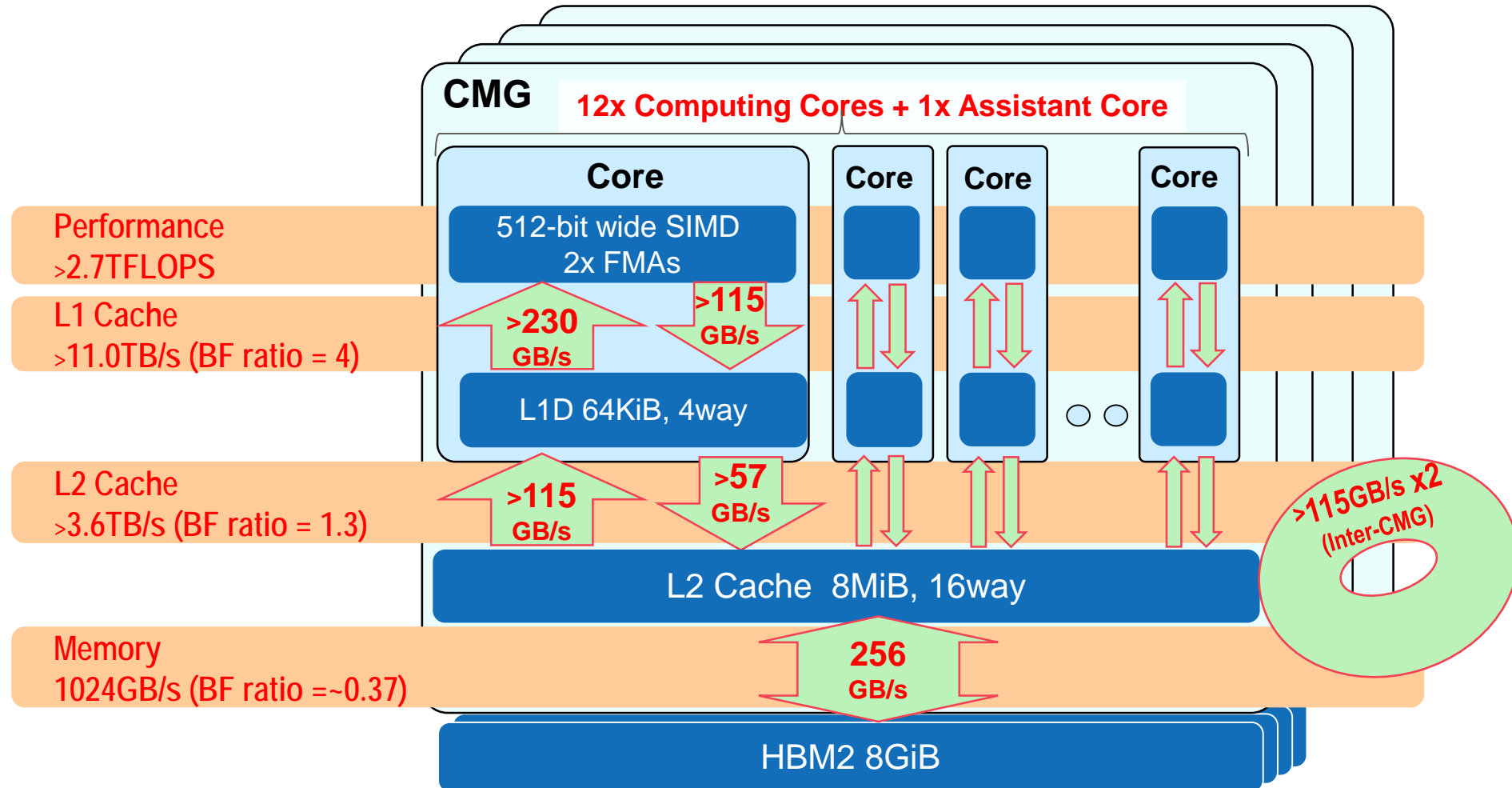
CMG Configuration
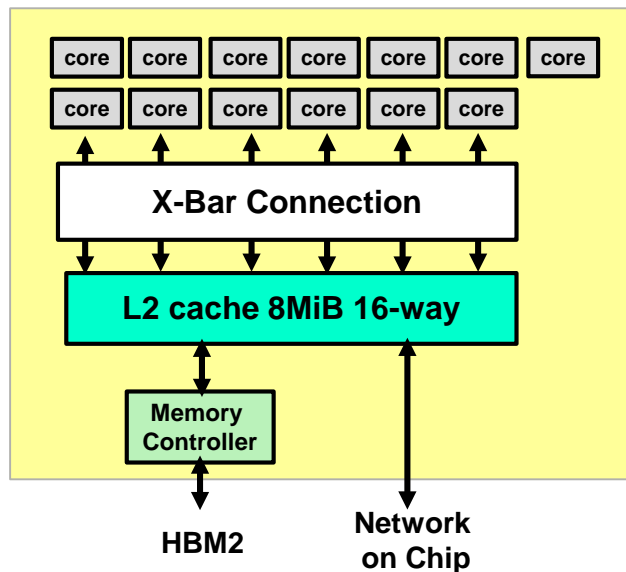
A64FX Chip Configuration
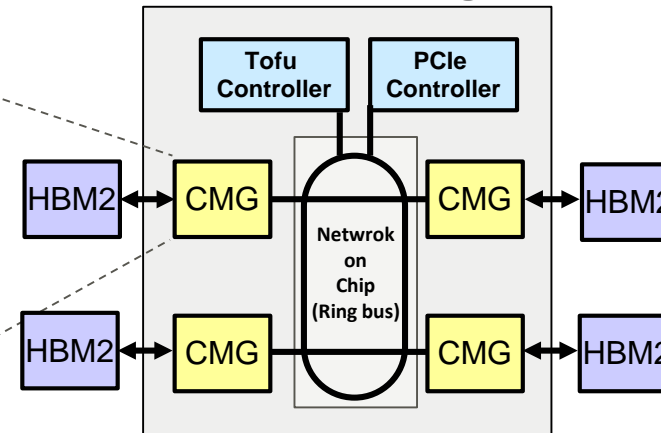
High Bandwidth

■ 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
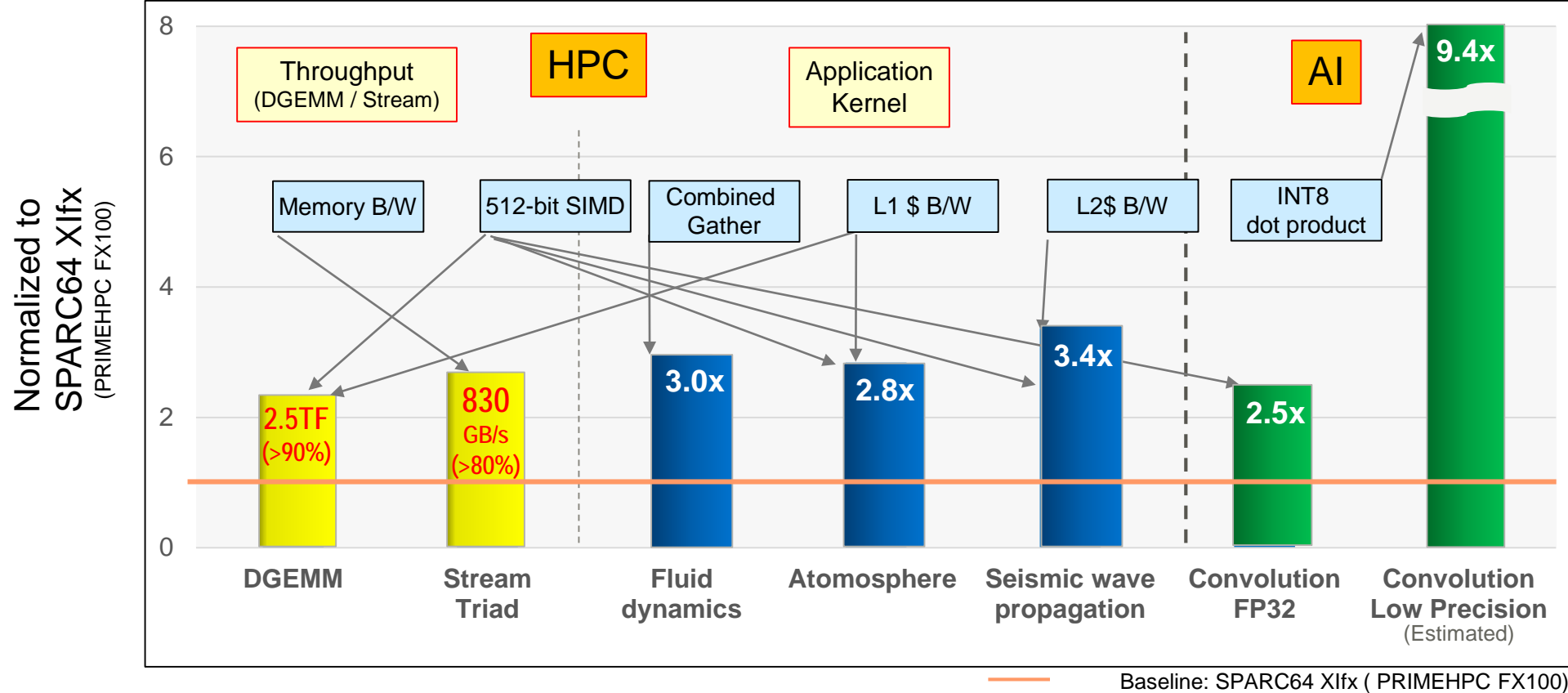
CMG

12x Computing Cores + 1x Assistant Core

Performance
>2.7TFLOPS

Core
512-bit wide SIMD
2x FMAs

Core **Core** **Core**

>230 GB/s **>115 GB/s**

L1 Cache
>11.0TB/s (BF ratio = 4)

L1D 64KiB, 4way

L2 Cache
>3.6TB/s (BF ratio = 1.3)

>115 GB/s **>57 GB/s**

>115GB/s x2 (Inter-CMG)

L2 Cache 8MiB, 16way

Memory
1024GB/s (BF ratio =~0.37)

256 GB/s

HBM2 8GiB

Performance

- A64FX boosts performance up by microarchitectural enhancements, 512-bit wide SIMD, HBM2 and process technology
  - > 2.5x faster in HPC/AI benchmarks than SPARC64 XIfx (Fujitsu's previous HPC CPU)
    - The results are based on the Fujitsu compiler optimized for our microarchitecture and SVE

A64FX Benchmark Kernel Performance (Preliminary results)

Normalized to SPARC64 XIfx (PRIMEHPC FX100)

Throughput (DGEMM / Stream)	
HPC	Application Kernel
AI	

Memory B/W — 512-bit SIMD — Combined Gather — L1 $ B/W — L2$ B/W — INT8 dot product

- DGEMM: 2.5TF (>90%)
- Stream Triad: 830 GB/s (>80%)
- Fluid dynamics: 3.0x
- Atomosphere: 2.8x
- Seismic wave propagation: 3.4x
- Convolution FP32: 2.5x
- Convolution Low Precision (Estimated): 9.4x

Baseline: SPARC64 XIfx (PRIMEHPC FX100)

Power Management

- "Energy monitor" / "Energy analyzer" for activity-based power estimation

  ✓ Energy monitor (per chip) : Node power via Power API* (~msec)   *Sandia National Laboratory
  - Average power estimation of a node, CMG (cores, an L2 cache, a memory) etc.

  ✓ Energy analyzer (per core) : Power profiler via PAPI** (~nsec)   ** Performance Application Programming Interface
  - Fine grained power analysis of a core, an L2 cache and a memory

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

<A64FX Energy monitor/ Energy analyzer>
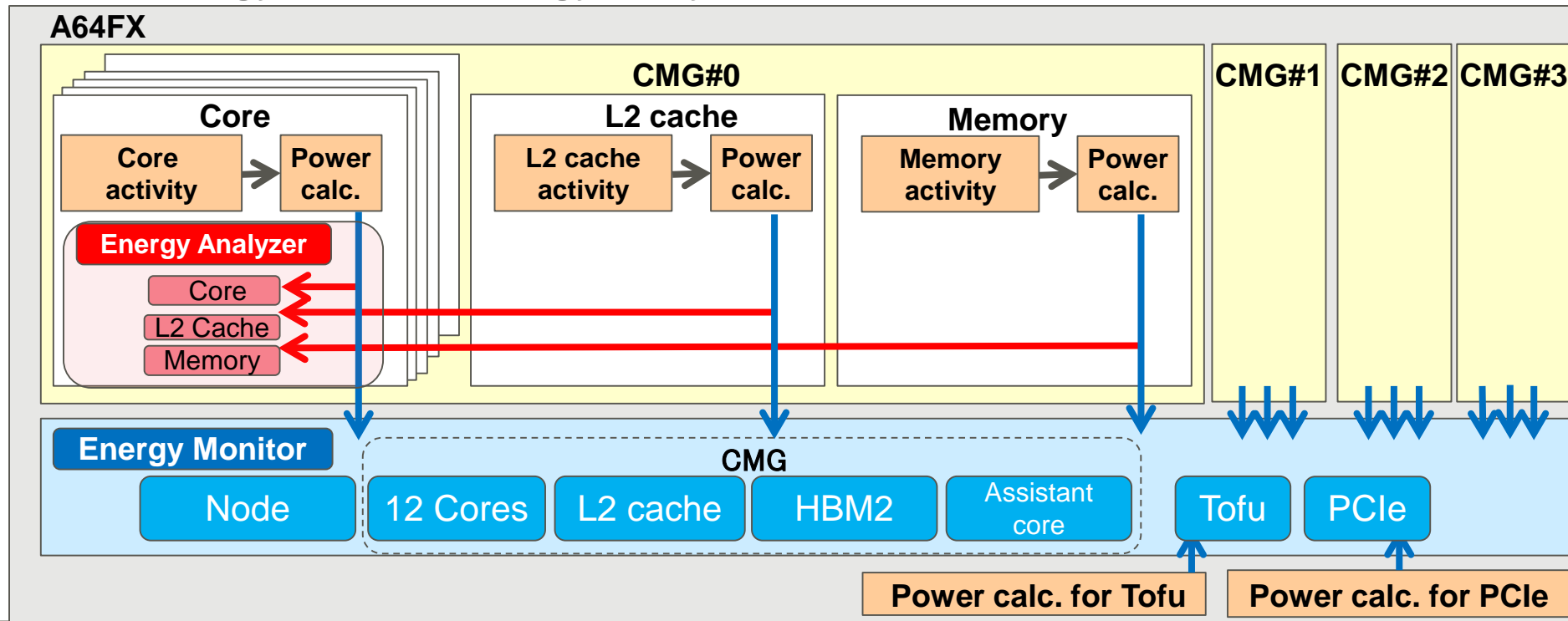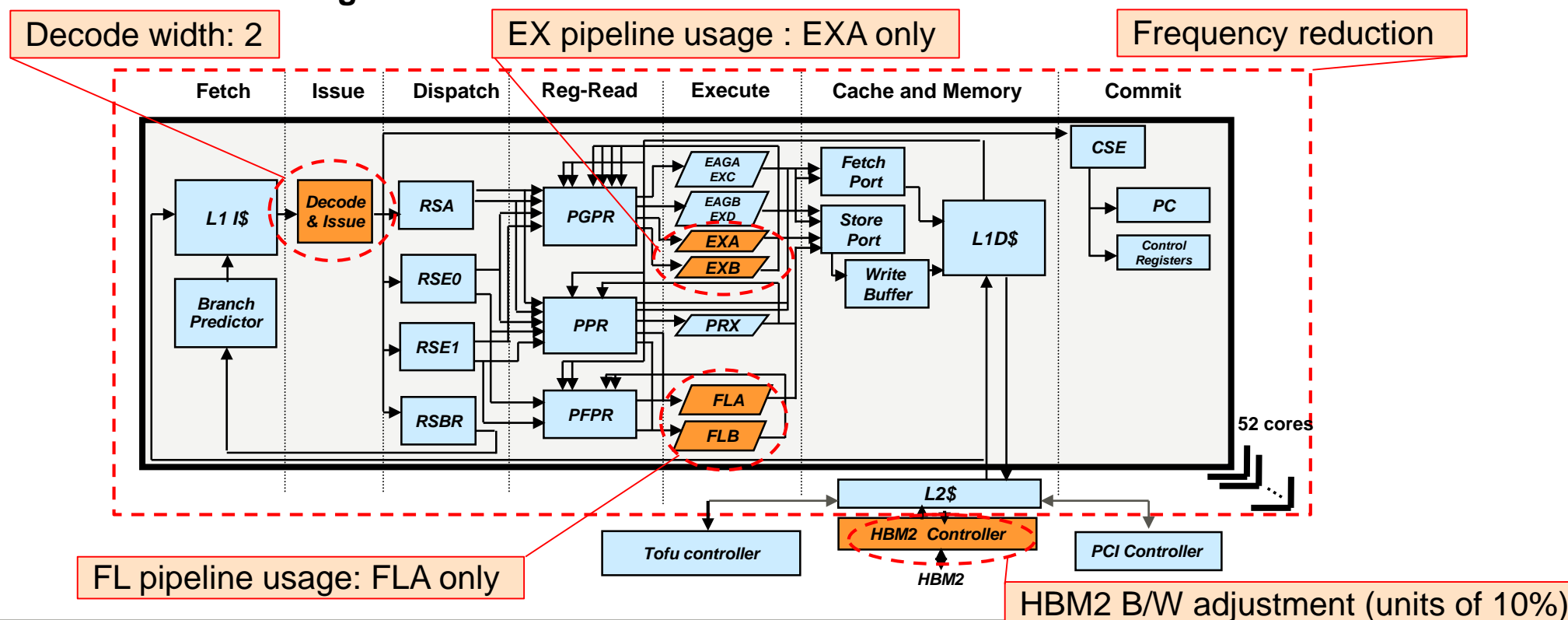
Power Management (Cont.)

- ■ "Power knob" for power optimization

  - ■ A64FX provides power management function called "Power Knob"

    - Applications can change hardware configurations for power optimization

  - → Power knobs and Energy monitor/analyzer will help users to optimize power consumption of their applications

\<A64FX Power Knob Diagram\>

Decode width: 2

EX pipeline usage : EXA only

Frequency reduction

FL pipeline usage: FLA only

HBM2 B/W adjustment (units of 10%)

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

<A64FX RAS Diagram>

<A64FX RAS Mechanism>

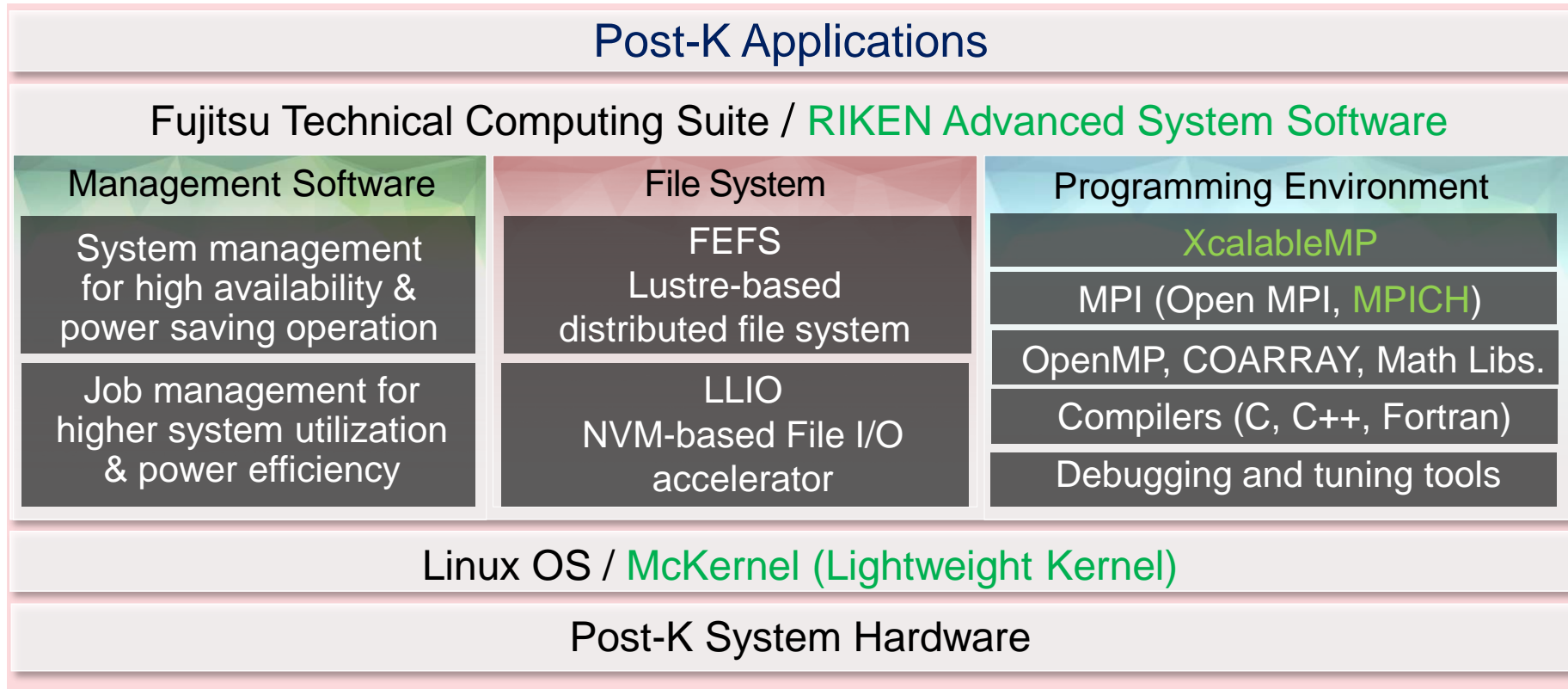| Units	Error Detection and Correction
Cache (Tag)	ECC, Duplicate & Parity
Cache (Data)	ECC, Parity
Register	ECC (INT), Parity(Others)
Execution Unit	Parity, Residue
Core	Hardware Instruction Retry
Tofu	Hardware Lane Recovery
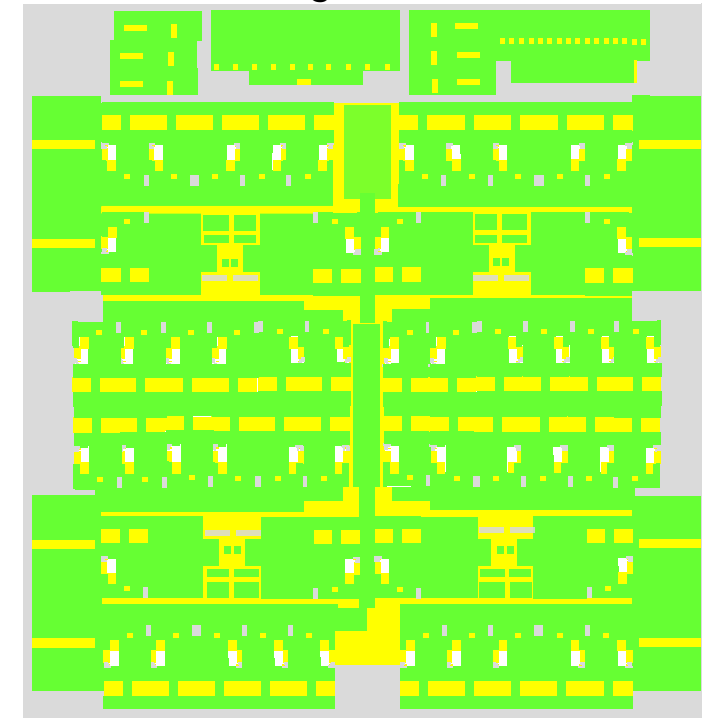
Green: 1 bit error Correctable
Yellow: 1 bit error Detectable
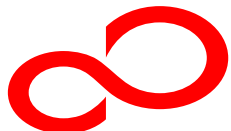Gray   : 1 bit error harmless

Software Development

- ■ RIKEN and Fujitsu are developing software stacks for the post-K computer
  - Fujitsu compilers are optimized for the microarchitecture, maximizing SVE and HBM2 performance
- ■ We collaboratively work with RIKEN / Linaro / OSS communities / ISVs and contribute to Arm HPC ecosystem

Post-K Applications		
Fujitsu Technical Computing Suite / RIKEN Advanced System Software		
Management Software	File System	Programming Environment
System management for high availability & power saving operation	FEFS Lustre-based distributed file system	XcalableMP
		MPI (Open MPI, MPICH)
Job management for higher system utilization & power efficiency	LLIO NVM-based File I/O accelerator	OpenMP, COARRAY, Math Libs.
		Compilers (C, C++, Fortran)
		Debugging and tuning tools
Linux OS / McKernel (Lightweight Kernel)		
Post-K System Hardware		

Summary

■A64FX is the first processor of the Armv8-A SVE architecture.
It is used for the post-K computer

■Fujitsu's proven microarchitecture achieves high performance in
HPC and AI areas

■Fujitsu collaboratively works with partners and continuously contributes to
Arm ecosystem

■We will continue to develop Arm processors

Abbreviations

- A64FX

  - RSA: Reservation station for address generation

  - RSE: Reservation station for execution

  - RSBR: Reservation station for branch

  - PGPR: Physical general-purpose register

  - PFPR: Physical floating-point register

  - PPR: Physical predicate register

  - CSE: Commit stack entry

  - EAG: Effective address generator

  - EX   :   Integer execution unit

  - FL   :   Floating-point execution unit

  - PRX : Predicate execution unit

  - Tofu: Torus-Fusion