

# The System Design of the Next Generation Supercomputer: Post-K Supercomputer

Shinji Sumimoto, Ph.D.

Next Generation Technical Computing Unit

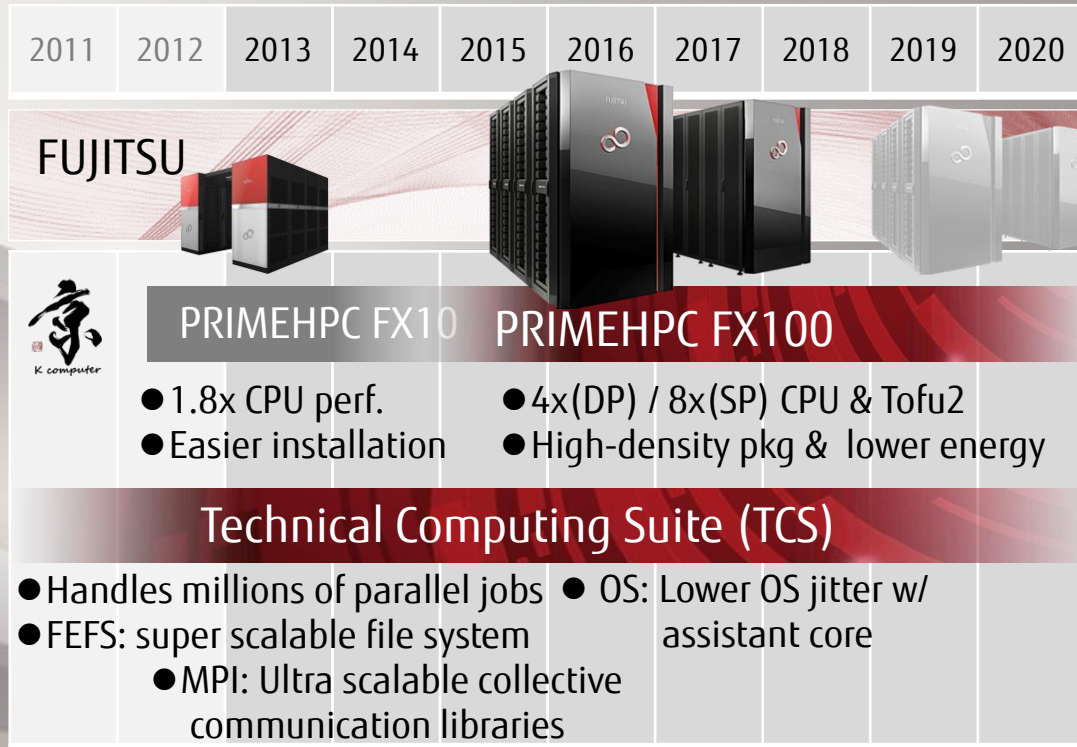
FUJITSU LIMITED

Oct. 25<sup>th</sup>, 2016

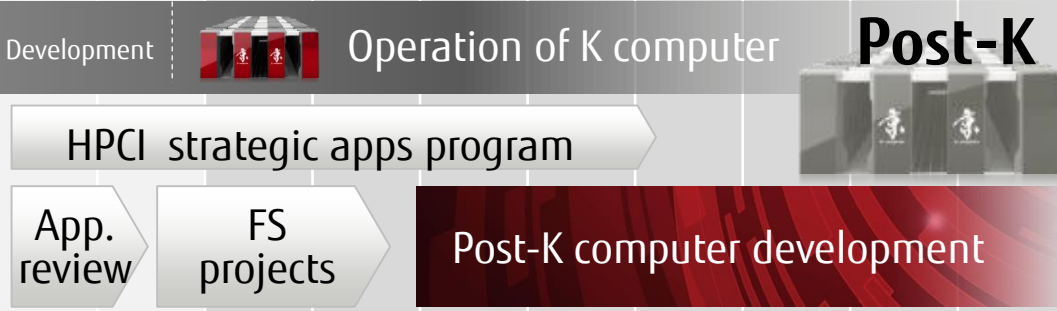
- Fujitsu in Technical Computing
  - PRIMEHPC FX100 overview
- The System Design of the Next Generation Supercomputer: Post-K
  - Background: FLAGSHIP2020 Project Overview
  - Requirements
  - System Design
  - Effectiveness for Meteorology

# Fujitsu in Technical Computing

# Past, PRIMEHPC FX100, and "Roadmap for Exascale"



## Japan's National Projects



## K computer and PRIMEHPC FX10 in operation

Many applications are currently running and being developed for science and various industries

## PRIMEHPC FX100 in operation

The CPU and interconnect inherit the K computer architectural concept, featuring state-of-the-art technologies

System software TCS supports the FX100 with newly developed technologies

## Towards Exascale

RIKEN and FUJITSU are working together for the Post-K computer

# Features of Fujitsu high-end supercomputer and Post-K



FUJITSU designed high performance CPU

Dedicated high performance interconnect Tofu

Application compatibility throughout generations

Post-K

## PRIMEHPC Series



© RIKEN

### K computer

VISIMPACT  
SIMD extension HPC-ACE  
Direct network Tofu  
CY2010~  
128GF, 8-core/CPU



### FX10

VISIMPACT  
HPC-ACE  
Direct network Tofu  
CY2012~  
236.5GF, 16-core/CPU



### FX100

**SMaC**  
Tofu interconnect 2  
HMC & Optical connections  
CY2015~  
1TF~, 32-core/CPU



## Provide steady progress for users

- Continue to keep performance portability among K computer, FX10 and FX100
- Facilitate the evolution of applications

## Challenge to state-of-art technologies for future generation

- 20nm CMOS technology
- Hybrid Memory Cube (HMC)
- 25Gbps optical connection

# PRIMEHPC FX100 Overview

## Tofu Interconnect 2

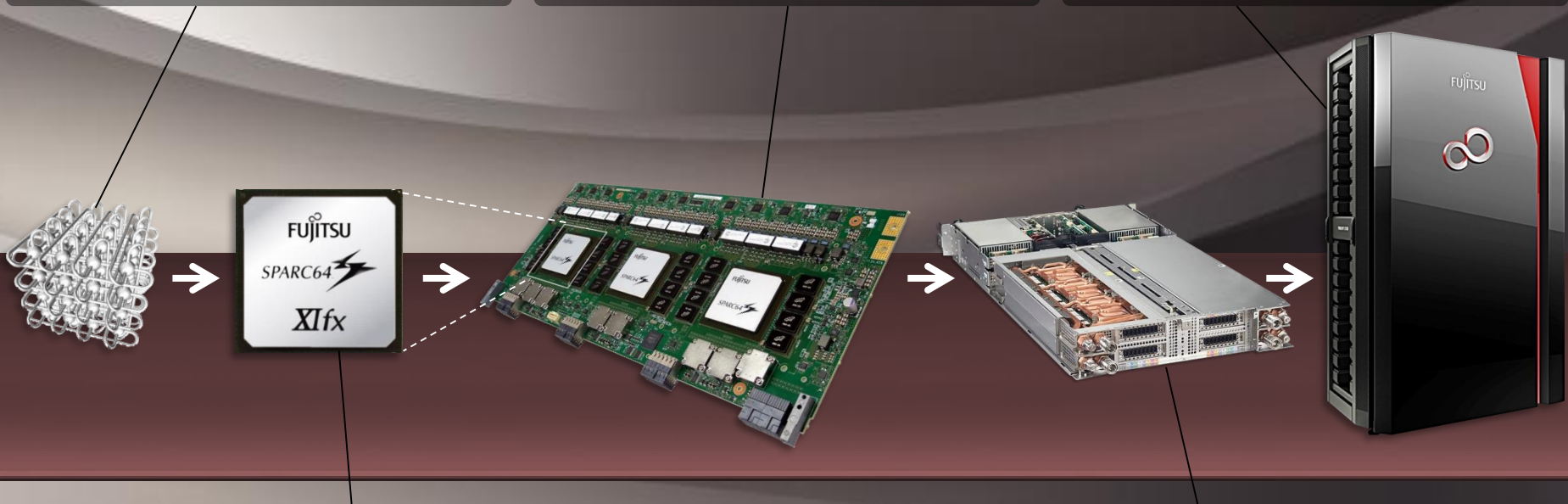
- 12.5 GB/s × 2 (in/out)/link
- 10 links/node
- Optical technology

## CPU Memory Board

- Three CPUs
- 3 x 8 Micron's HMCs
- 8 opt modules, for inter-chassis connections

## Cabinet

- Up to 216 nodes/cabinet
- High-density
- 100% water cooled with EXCU (option)



## Fujitsu designed SPARC64 XIfx

- 1TF~(DP)/2TF~(SP)
- 32 + 2 core CPU
- HPC-ACE2 support
- Tofu2 integrated

## Chassis

- 1 CPU/1 node
- 12 nodes/2U Chassis
- Water cooled

# The System Design of Post-K

- Background: FLAGSHIP2020 Project Overview
- Requirements of Post-K
- Design for Application Performance Portability
- Effectiveness for meteorology



# FLAGSHIP 2020 project overview:

IDC HPC User Forum @ Austin: from talk of Project Leader Prof. Ishikawa

## An Overview of Flagship 2020 project



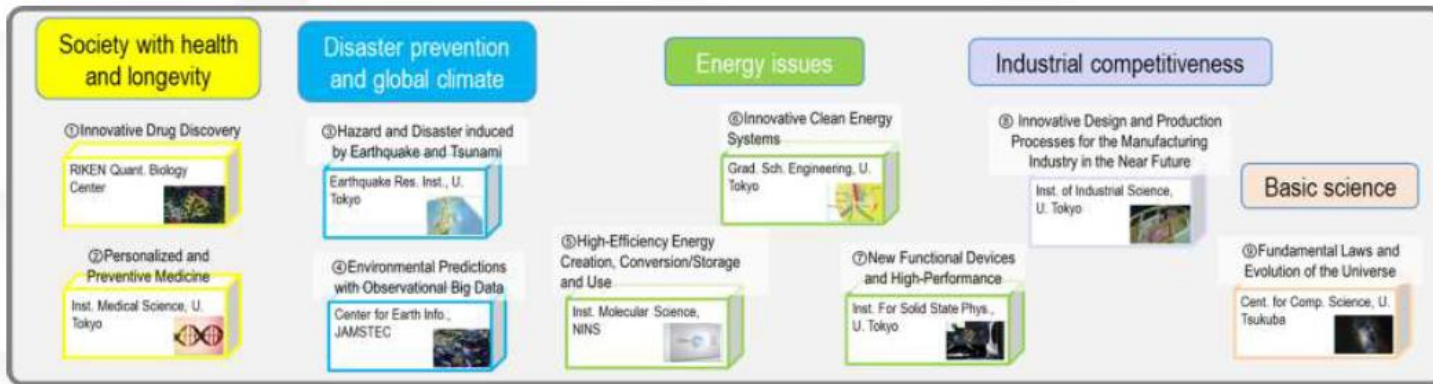
- Developing the next Japanese flagship computer, so-called "post K"
- Developing a wide range of application codes, to run on the "post K", to solve major social and science issues



Vendor partner



The Japanese government selected 9 social & scientific priority issues and their R&D organizations.



### Target Applications' Characteristics



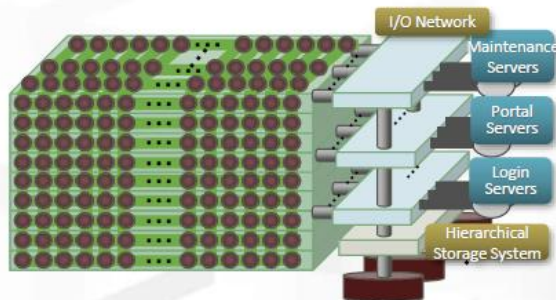
Target Application		
Program	Brief description	Co-design
① GENESIS	MD for proteins	Collective comm. (all-to-all), Floating point perf (FPP)
② Genomon	Genome processing (Genome alignment)	File I/O, Integer Perf.
③ GAMERA	Earthquake simulator (FEM in unstructured & structured grid)	Comm., Memory bandwidth
④ NICAM+LETK	Weather prediction system using Big data (structured grid stencil & ensemble Kalman filter)	Comm., Memory bandwidth, File I/O, SIMD
⑤ NTChem	molecular electronic (structure calculation)	Collective comm. (all-to-all, allreduce), FPP, SIMD,
⑥ FFB	Large Eddy Simulation (unstructured grid)	Comm., Memory bandwidth,
⑦ RSDFT	an ab-initio program (density functional theory)	Collective comm. (bcast), FPP
⑧ Adventure	Computational Mechanics System for Large Scale Analysis and Design (unstructured grid)	Comm., Memory bandwidth, SIMD
⑨ CCS-QCD	Lattice QCD simulation (structured grid Monte Carlo)	Comm., Memory bandwidth, Collective comm. (allreduce)

### An Overview of post K



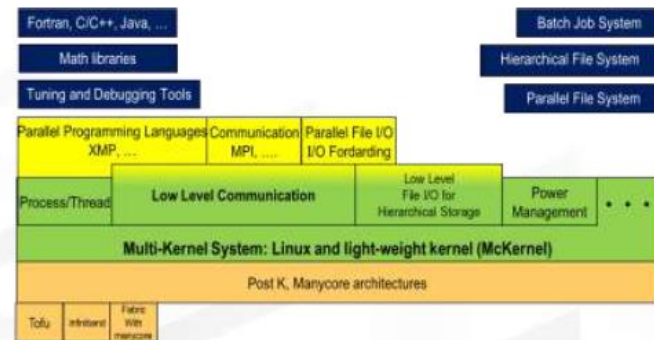
#### ● Hardware

- Manycore architecture
- 6D mesh/torus Interconnect
- 3-level hierarchical storage system
  - Silicon Disk
  - Magnetic Disk
  - Storage for archive



#### ● System Software

- Multi-Kernel: Linux with Light-weight Kernel
- File I/O middleware for 3-level hierarchical storage system and application
- Application-oriented file I/O middleware
- MPI+OpenMP programming environment
- Highly productive programming language and libraries



## ■ Goals:

- World's top class application performance with solving social and scientific priority issues
- Reliable system operation with limited power consumption

## ■ Requirements:

- World's top class application performance with limited power consumption
- Keeping system reliability as much as possible as well as K computer.
- Easy migration of existing application from existing systems including K computer to expand system use

■ Especially, performance portability of existing application on K computer is important

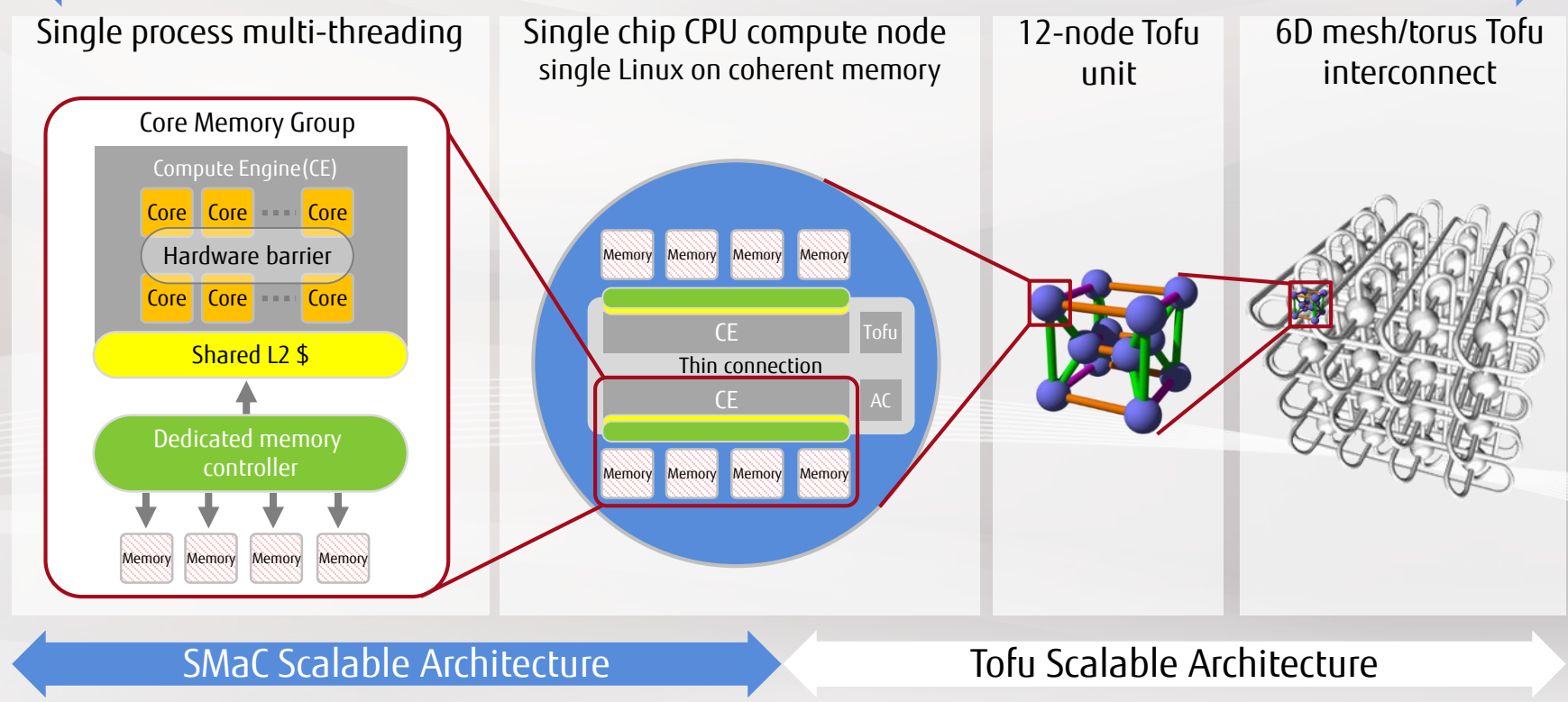
- Application Execution Model Backward Compatibility
  - Processor Architecture
  - System Architecture
- System Balance among Processor Core, Memory, Interconnect and Storage I/O Performance
  - Trying to keep system balance with limited power consumption and cost.
- Binary Compatibility Mechanism for Future Generations
  - Preferring to keep portability without re-compiling applications
- Execution Environment: Compiler, Runtime System, MPI, Batch Script etc.
  - Backward Compatibility of System Operation and Application Execution Environment

# APPLICATION EXECUTION MODEL BACKWARD COMPATIBILITY

# System Architecture for Performance Portability

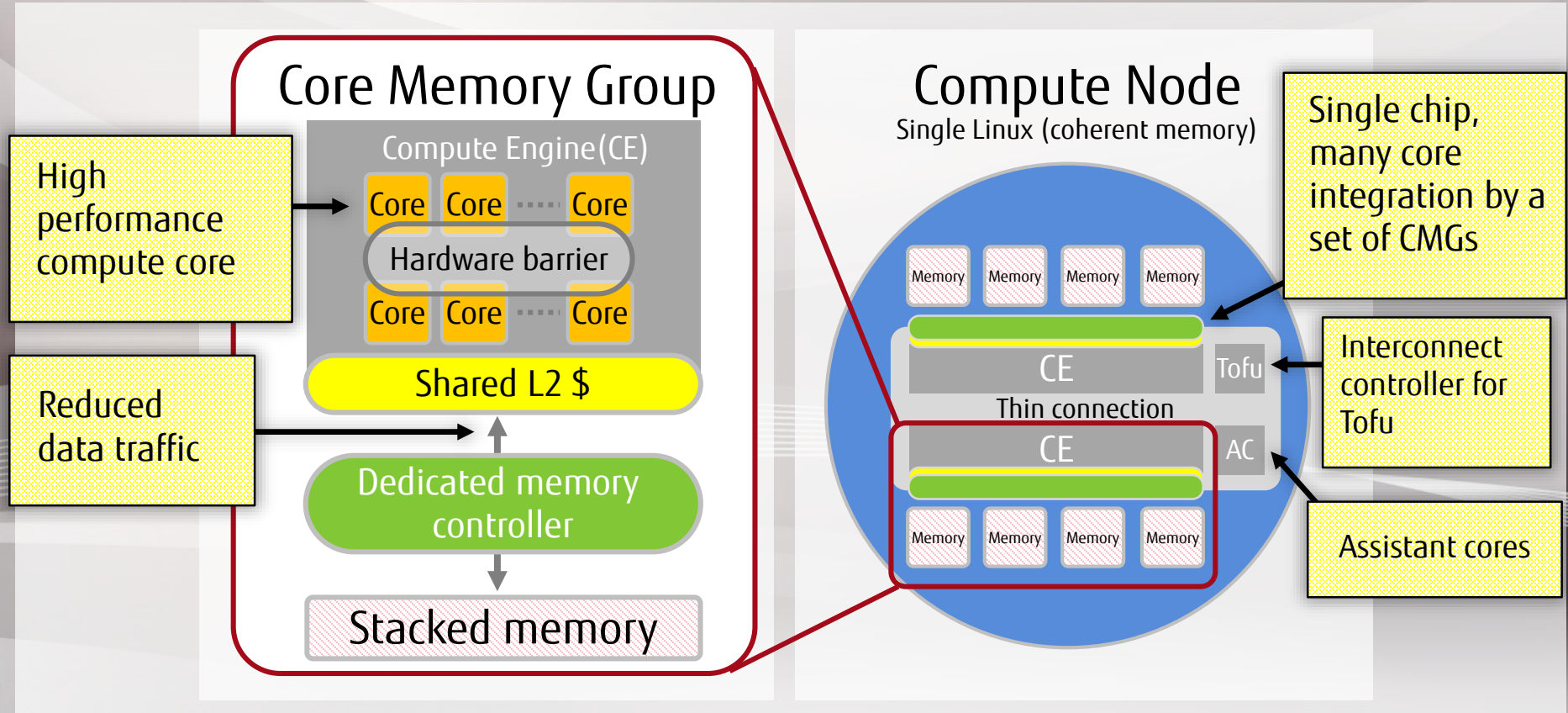
- A scalable, many-core micro architecture concept: "SMaC,"
  - Single Process Multi-Threaded Model in a Socket: CMG
  - Scalable interconnect: "Tofu"
- High Performance Lustre Based Cluster File System: FEFS

Scalable System Software Architecture: Resource-saving, Flexible, Reliable



# Core Memory Group (CMG) Structure

- Cores in the group share the same L2 cache
- Dedicated memory and memory controller for the CMG provide high BW and low latency data access
- Loosely coupled CMGs using tagged coherent protocol share data with small silicon overhead
- Hierarchical configuration promises good core/performance scalability





# **SYSTEM PARAMETER BALANCE AMONG PROCESSOR CORE, MEMORY, INTERCONNECT AND STORAGE I/O**

# Post-K's System Balance

FUJITSU original CPUs steadily increase their fundamental performance

Support a programming model, hybrid parallel execution

Uncompromised **system balance** for the best use of applications

From FX100, an **assistant core** and **CMG** is introduced

	Post-K	FX100	FX10	K computer
Double Flops / CPU	<b>TBD</b>	1 TF	235 GF	128 GF
Single Flops / CPU		2 TF	235 GF	128 GF
SIMD width		256 bit	128 bit	128 bit
# of CMG(# of cores/CMG)		2(17)	1(16)	1(8)
# of cores / CPU		32 + 2xAC	16	8
Memory / CPU		32 GB	32 GB	16 GB
Memory BW		480 GB/s	83.5 GB/s	64 GB/s

## Compatible highly scalable FUJITSU original interconnect

- Optimal implementation keeping high application scalability
- Flexible and efficient communication patterns for application performance
- Non blocking CPU off loadable DMA engines for calculation and communication overlapping

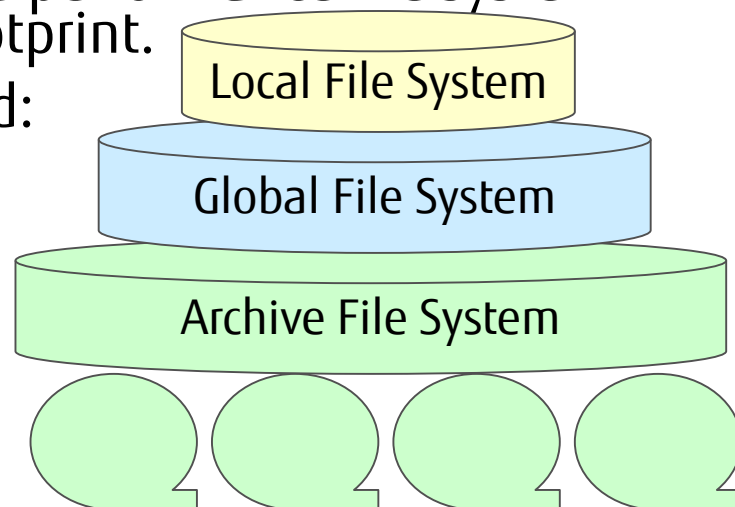
	Post-K	FX100	FX10	K computer
Interconnect	Tofu 6D mesh/torus			
Interconnect BW	<b>TBD</b>	12.5 GB/s	5 GB/s	5 GB/s
# of DMA engines		4	4	4
Node injection BW		50 GB/s	20 GB/s	20 GB/s
Collective operations		Yes	Yes	Yes

## ■ K computer File System Design

- How should we realize High Speed and Redundancy together?
- How do we avoid I/O conflicts between Jobs?
- These are not realized in single file system.
  - Therefore, we have introduced Integrated Layered File System.
- K computer achieved 1 TB/s sustained file I/O performance

## ■ Next Generation File System/Storage Design

- Another trade off targets: Power, Capacity, Footprint
  - Difficult to realize enough capacity and performance file system in limited power consumption and footprint.
- Third Storage layer for Capacity is needed:  
Three Layered File System
  - Local File System for Performance
  - Global File System for Shared Use
  - Archive File System for Capacity



# **BINARY COMPATIBILITY MECHANISM FOR FUTURE GENERATIONS**

Post-K fully utilizes FUJITSU proven supercomputer microarchitecture

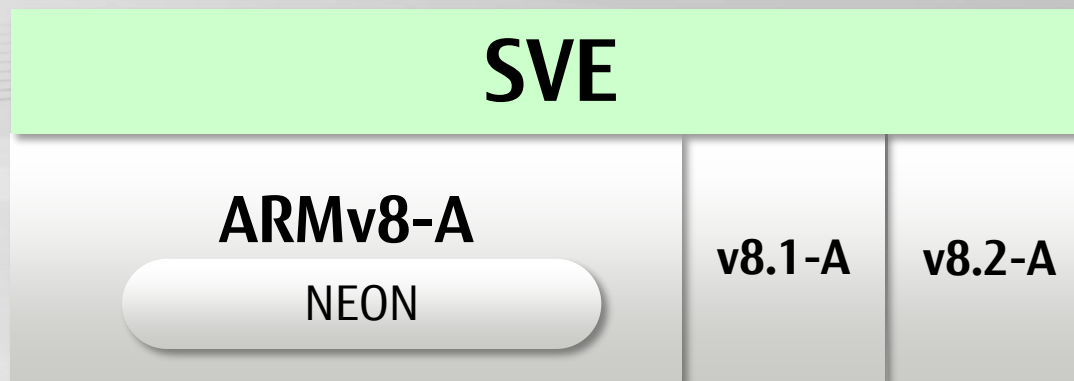
FUJITSU, as a lead partner of ARM SVE development, is contributing to complement ARM SVE (Scalable Vector Extension), for application performance efficiency

ARM V8+SVE brings out the real strength of FUJITSU's microarchitecture

ISA	Functions for Perf.	Post-K	FX100	FX10	K computer
SVE w/ FJ contribution	FMA	✓	✓	✓	✓
	Math. acc. prim.*	✓ Enhanced	✓	✓	✓
Fujitsu Extension	Inter core barrier	✓	✓	✓	✓
	Sector cache	✓ Enhanced	✓	✓	✓
	Prefetch	✓ Enhanced	✓	✓	✓

\*Mathematical acceleration primitives include trigonometric functions, sine & cosines, and exponential...

- ❑ HPC Extension Instruction Set of AArch64
  - ❑ SVE is not an extension of NEON, independent of NEON for HPC processing
  - ❑ SVE instruction and NEON instruction are able to execute independently
  - ❑ SVE includes various amount instruction set to support higher SIMD execution
  
- ❑ Post-K processor ISA will be ARM with SVE



# SIMD ISA Extension Comparison

		SVE	HPC-ACE2	AVX-512
Base ISA		ARMv8-A	SPARC V9	Intel 64
SIMD	Bit width	<b>128~2048</b>	256	512
	SP Elements	<b>4~64</b>	8	16
	DP Elements	<b>2~32</b>	4	8
GP Registers (#)		31+SP (Same as v8-A)	32+32	16
Vector Registers (#)		32	128	32
Predicate Registers (#)		<b>16</b>	(Included in Vector Regs.)	8



- ISA does not fix Vector Length

- SVE supports VL from 128 to 2048 bit with multiples of 128 bit
- VL is set by processor before executing a binary dynamically

- Single execution binary can be executed on processors with multiple VLs

- Vector-Length Agnostic (VLA) programming enables ABI Compatibility

## Execution Binary Portability

Execution Binary does not depend on processor's VL



512bit SIMD

Increasing dynamic instruction steps to double

Execution Binary/a.out

Reducing dynamic instruction steps to half



256bit SIMD

# EFFECTIVENESS FOR METEOROLOGY ON POST-K

## ■ Achievements with K computer

- NICAM performance on K computer is good scalability up to 81920node x 8 threads with 0.9 PFLOPS

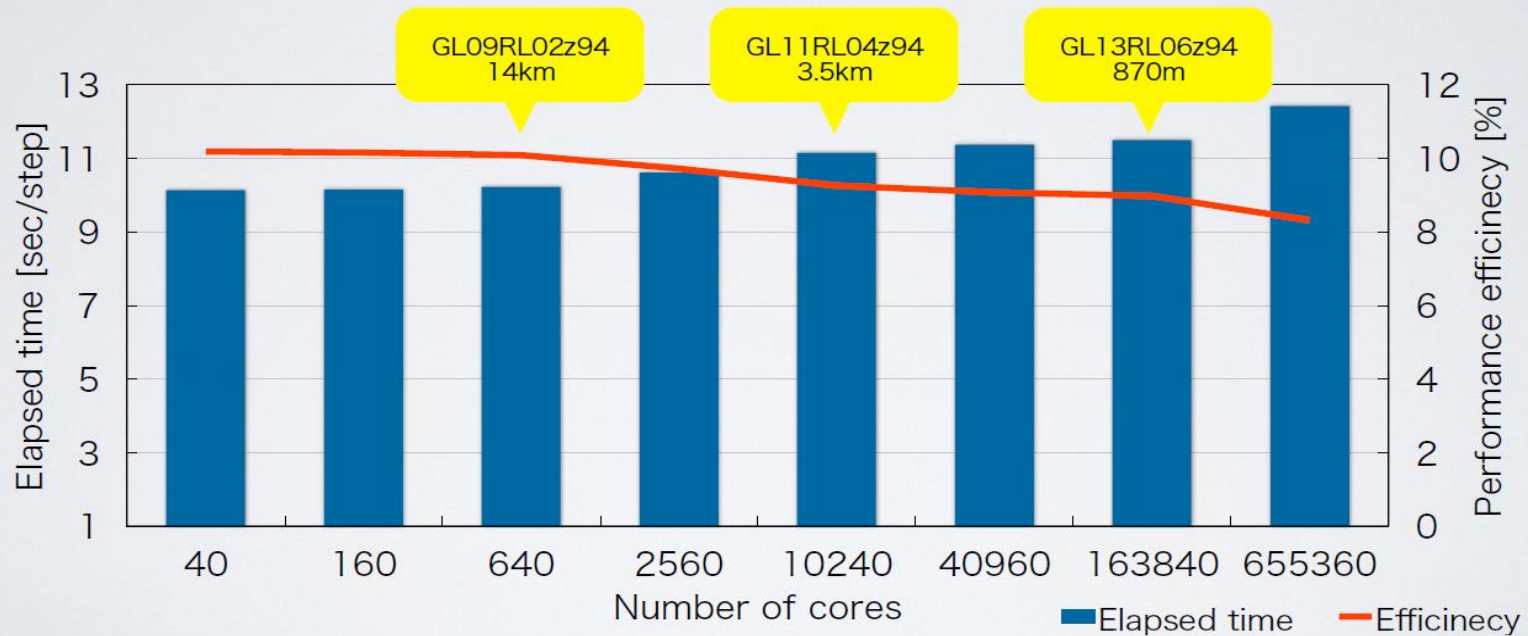
- “Recent performance of NICAM on the K-computer and activities towards post-petascale computing”, Hisashi Yashiro (Riken/AICS), Workshop on Scalability (ECMWF, 14-15 April, 2014)
- <http://www.ecmwf.int/sites/default/files/elibrary/2014/13821-recent-performance-nicam-k-computer-and-activities-towards-post-petascale-computing.pdf>

## ■ Effectiveness for meteorology on Post-K

- Wider SIMD with good system balance using meteorology application: IFS

## Weak scaling test

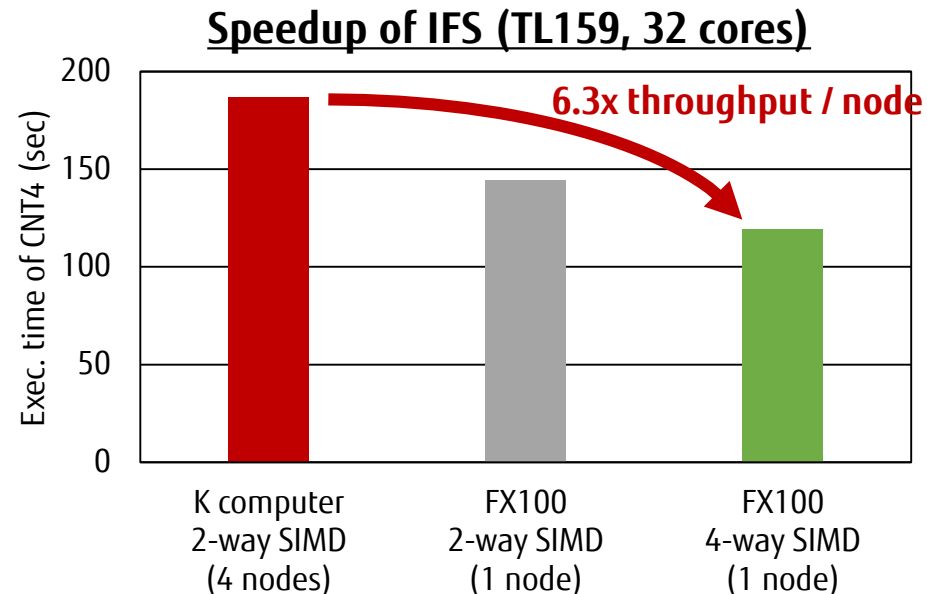
- Same problem size per node, same steps
  - Full configuration / full components
  - Realistic boundary / initial data set
- Good scalability up to 81920node x 8threads with 0.9PFLOPS



# Performance Portability from K to FX100 and Post-K

- SIMD width and memory bandwidth are enhanced from K to FX100
  - 4-way SIMD is supported on FX100 and Hybrid Memory Cube (HMC) provides higher memory bandwidth
- Speedup of IFS on FX100 is realized by wider SIMD and good system balance
  - 2-way SIMD can benefit from high memory bandwidth
  - 4-way SIMD accelerates calculations and drives memory bandwidth more
- Trying to keep system balance will be expected to provide higher performance on the next-generation machine: Post-K with more wider SIMD width

	K computer	FX100
Flops / CPU	128 Gflops	1 Tflops
SIMD width	128 bit	256 bit
Memory BW	64 GB/s	480 GB/s
Byte per flop	0.4 ~ 0.5	




## ■ Fujitsu in Technical Computing

### ■ PRIMEHPC FX100 Overview

## ■ The System Design of the Next Generation Supercomputer: Post-K

- Performance Portability: Trying to keep system balance with limited power consumption and cost will be expected to provide higher performance on the next-generation machine: Post-K with more wider SIMD width.
- Application Binary Compatibility: Scalable Vector Length will help to keep binary compatibility for the future systems without re-compilation of programs.



**FUJITSU**

shaping tomorrow with you