FUJITSU

shaping tomorrow with you

# FUJITSU HPC
# The Next Step Toward Exascale

November 17th, 2015

**Toshiyuki Shimizu**

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

FUJITSU


## 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

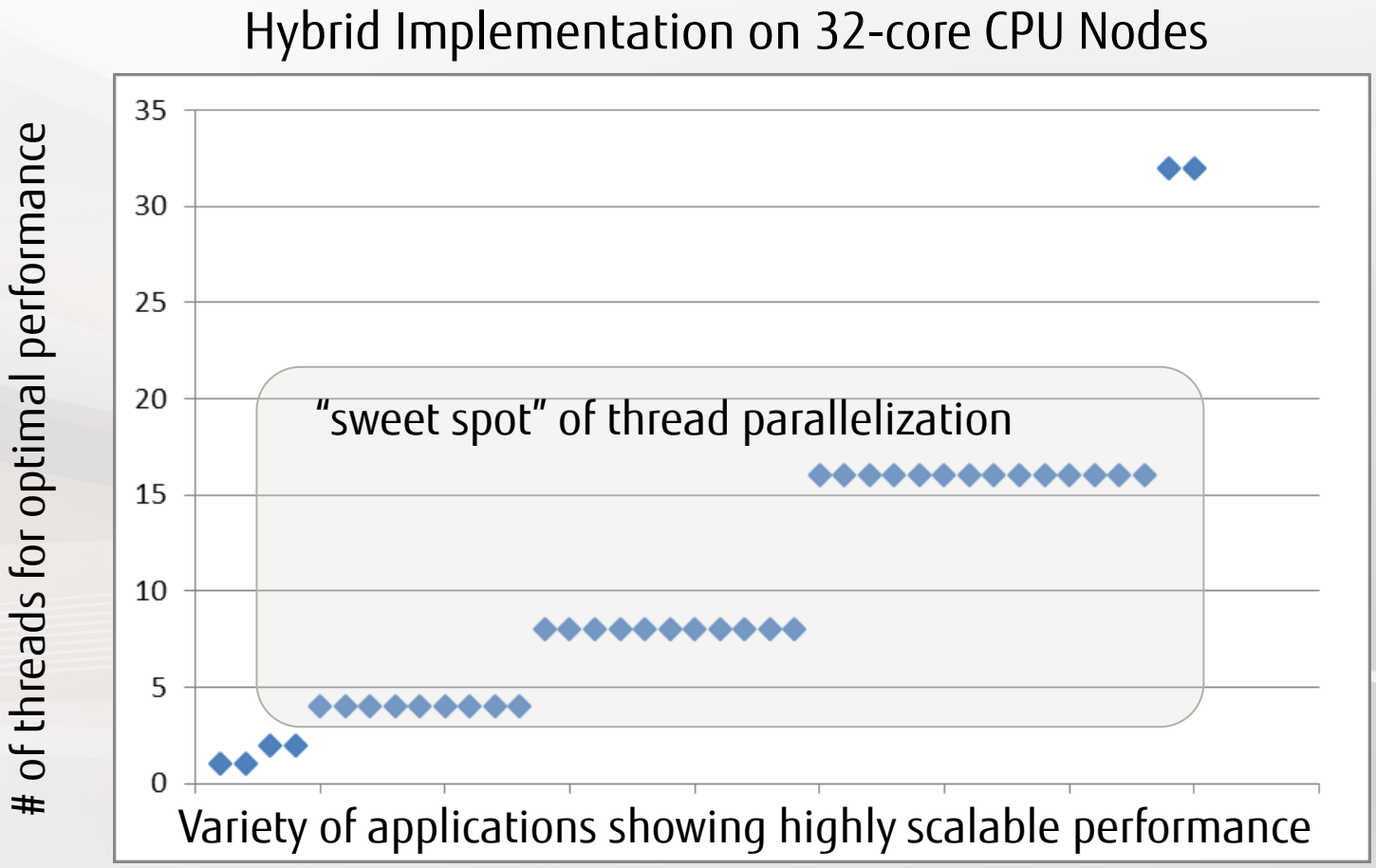
## Towards Exascale

RIKEN selected Fujitsu as a partner for the basic design of the Post K computer

# Toward Higher Performance beyond 100PF

FUJITSU

- Hybrid and hierarchical implementation must be chosen!
  - Four and sixteen thread hybrid parallelization shows the best performance for most of the applications

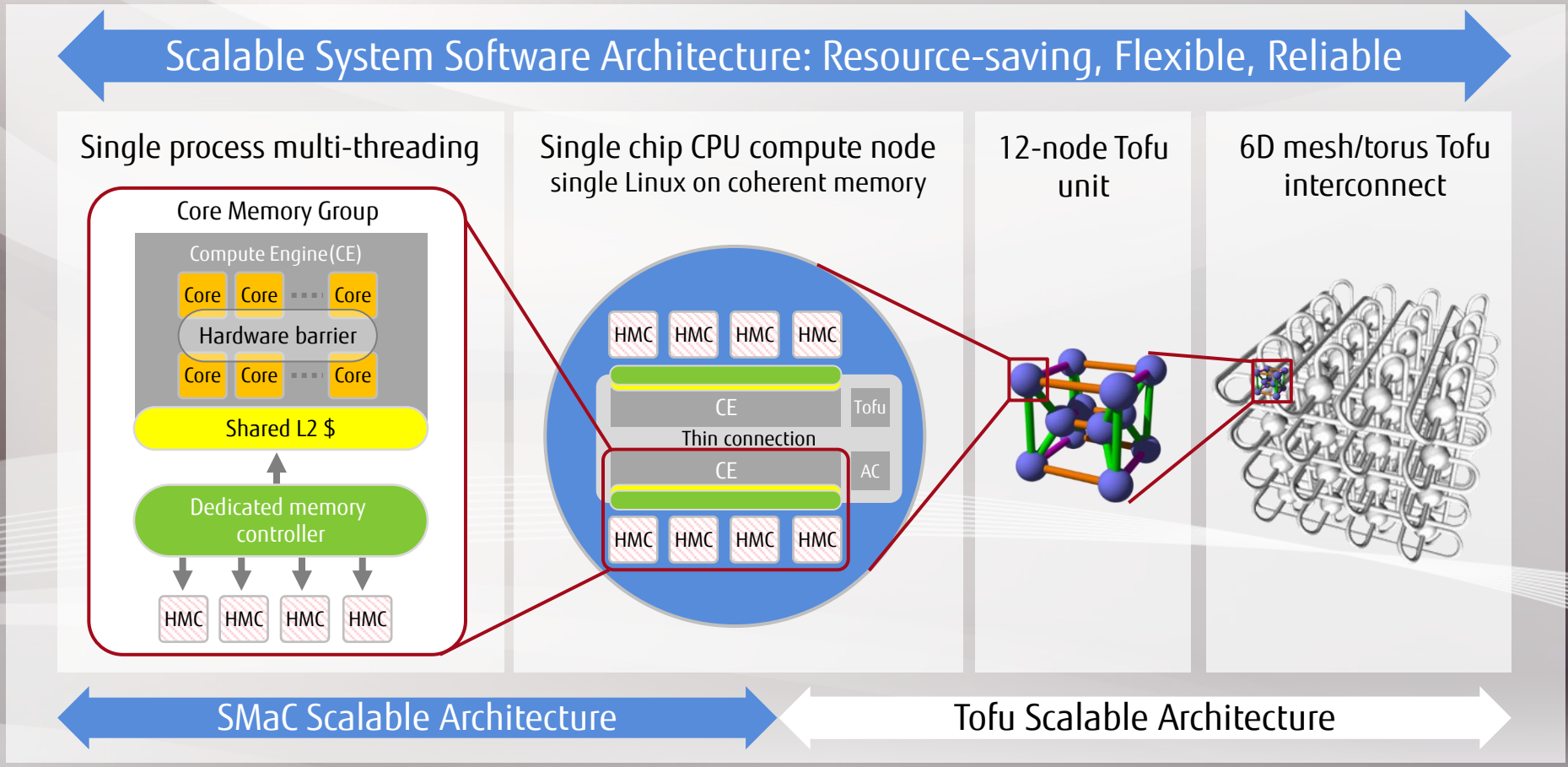


# Fujitsu's Approach for FX100 and Beyond

FUJITSU

## Using State-of-the-Art SW/HW Technologies for Application Performance

- System software is also enhanced for long term evolution
- A scalable, many-core micro architecture concept, "SMaC," has been developed
- Scalable interconnect "Tofu"



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

FUJITSU

- For Higher Application Performance
  - Thin and low overhead system design
- Resource-saving
  - Small system memory footprint
  - Power-saving control
- Usability and Compatibility
  - OSS, ISV support, asset protection
- Flexibility
  - Supports computer science research and data analytic science research in addition to computational science
- Reliability
  - Stable operation, immediate fault detection, minimization of downtime through fast recovery
- Maintainability
  - System updates occur during operation, minimizing maintenance time by enabling fault log sampling

# SMaC (Scalable Many Core) Concept & Approach

FUJITSU

- Many core-oriented, long-lasting architecture
- Scalable performance improvement by increasing the number of cores
  - Increasing the number of cores would be safe, even in the post-Moore's Law era using 3D stacking and alternative newer technologies

**Middle-sized, general purpose, out-of-order, superscalar processor core**

- Good performance for variety of apps
- Low power

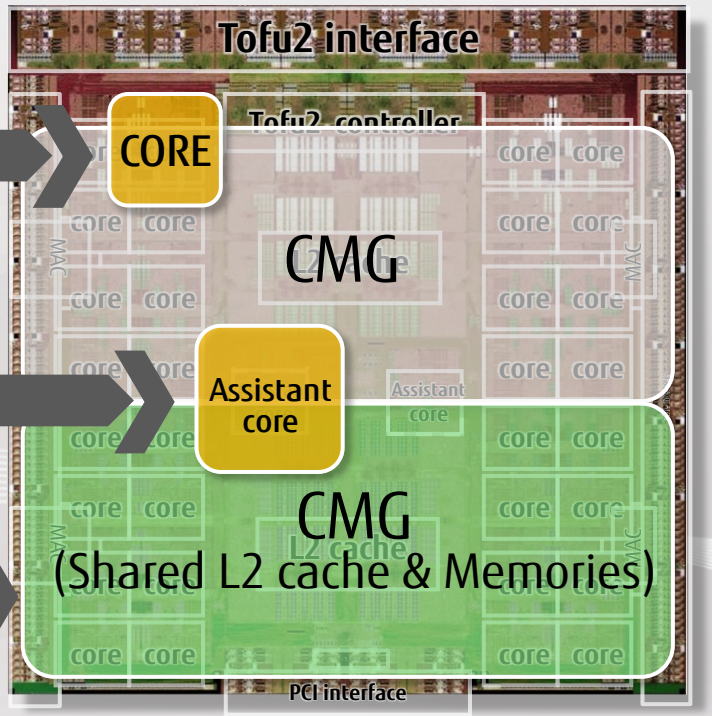
**Assistant core**

- OS jitter reduction and assistance for IO, async MPI
- Highly scalable nodes

**Core Memory Group (CMG), many core building block, ccNUMA integration**

- Hierarchal structure for hybrid parallel model
- Optimized area and performance

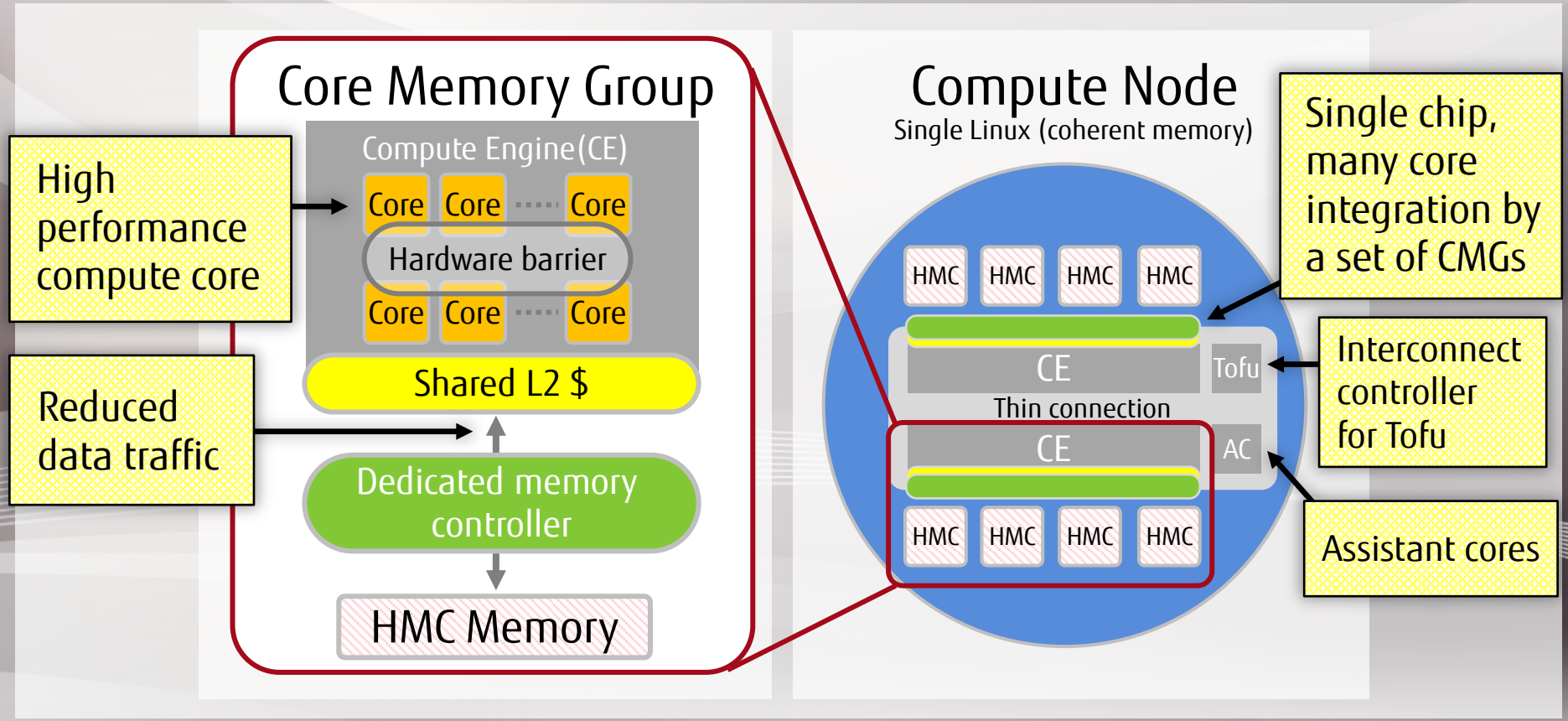
## FX100 CPU implementation



# Core Memory Group (CMG) Structure

FUJITSU

- 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

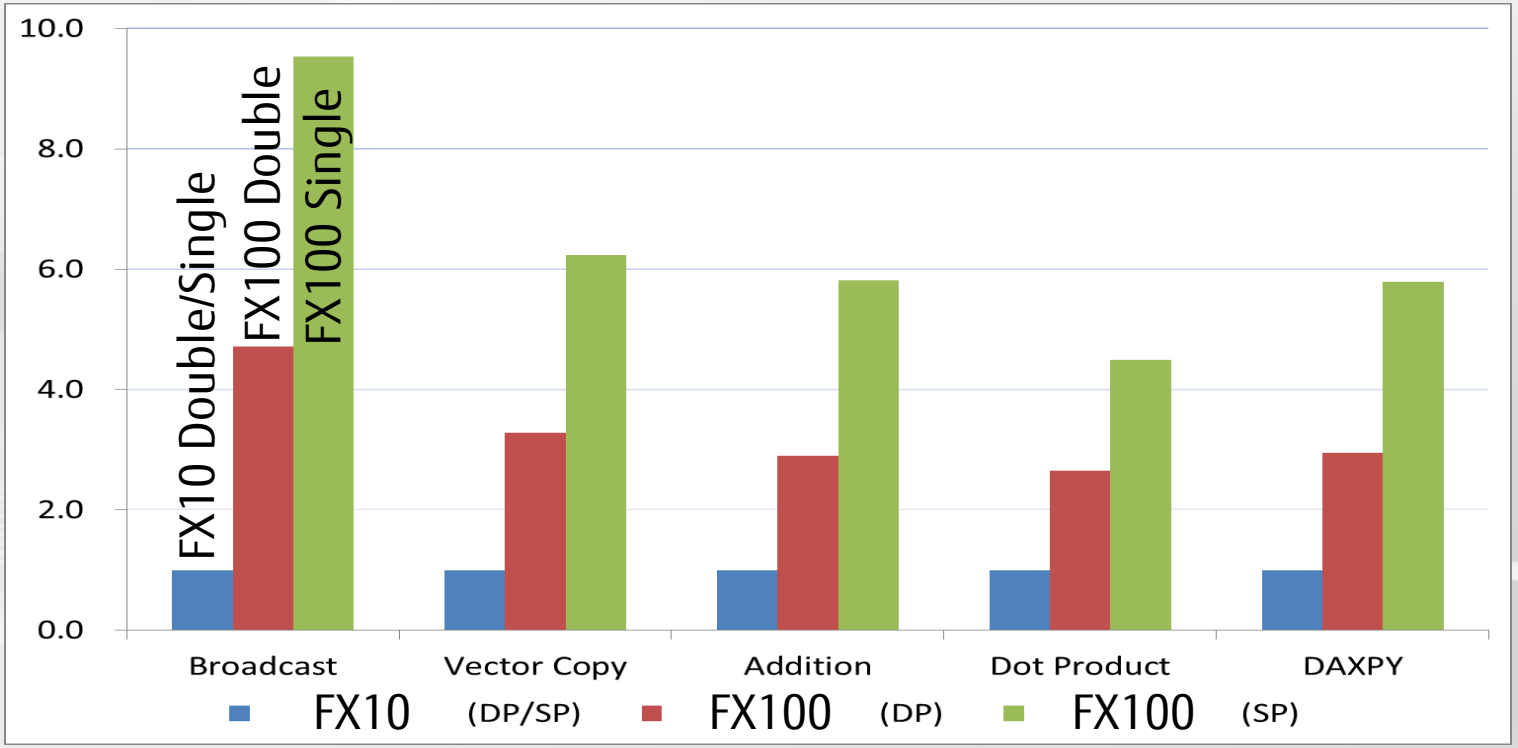


# High Performance Compute Core (SIMD extensions)

FUJITSU

- DP 3x, SP 6x faster than FX10 in basic kernels
  - High BW memory & Improved L1 cache pipelines contribute to exceed a peak performance increase of 2.3x

## Basic Kernels Performance per Core



# Reducing Data Traffic (XFILL & Sector Cache)

FUJITSU

- XFILL inst. marks the cache line filled with new data
- Sector cache holds specific data loaded by attributed load inst.

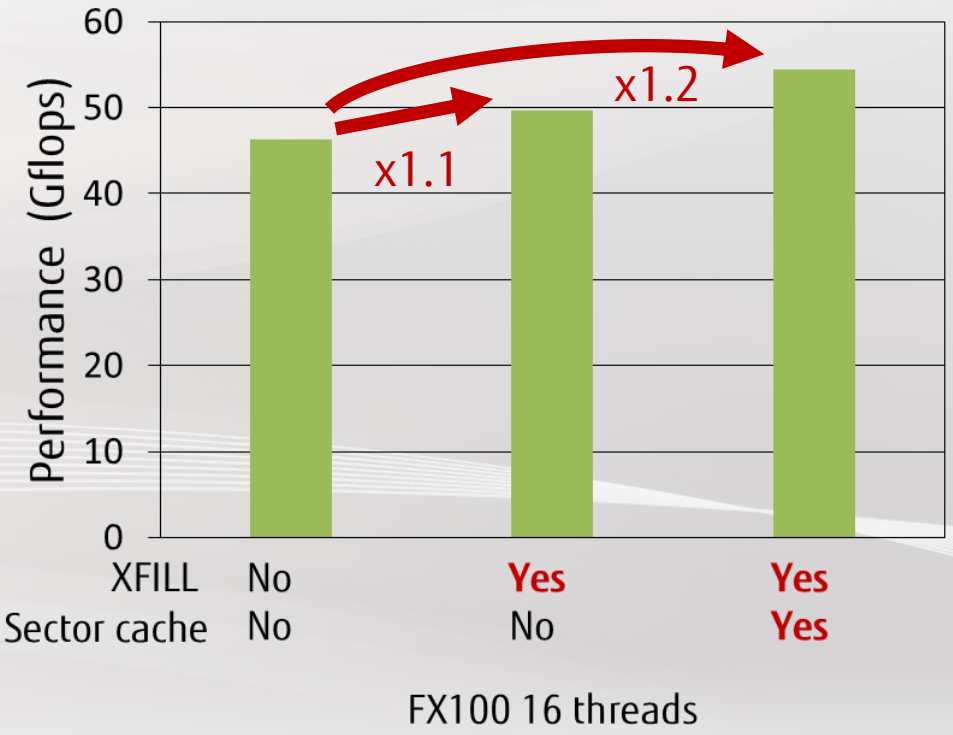
## Himeno's Benchmark

```
!OCL CACHE_SECTOR_SIZE(18,6)
!OCL CACHE_SUBSECTOR_ASSIGN(...)
do k=2,kmax-1
 do j=2,jmax-1
  do i=2,imax-1
   s0=...p(i, j, k)... &
        ...p(i, j+1, k)... &
        ...p(i, j-1, k)...
   ...
   wrk2(i, j, k)=...
  enddo
 enddo
enddo
!OCL END_CACHE_SECTOR_SIZE
!OCL END_CACHE_SUBSECTOR
```

Specify vars except array *p*

## Speedup by XFILL & Sector Cache

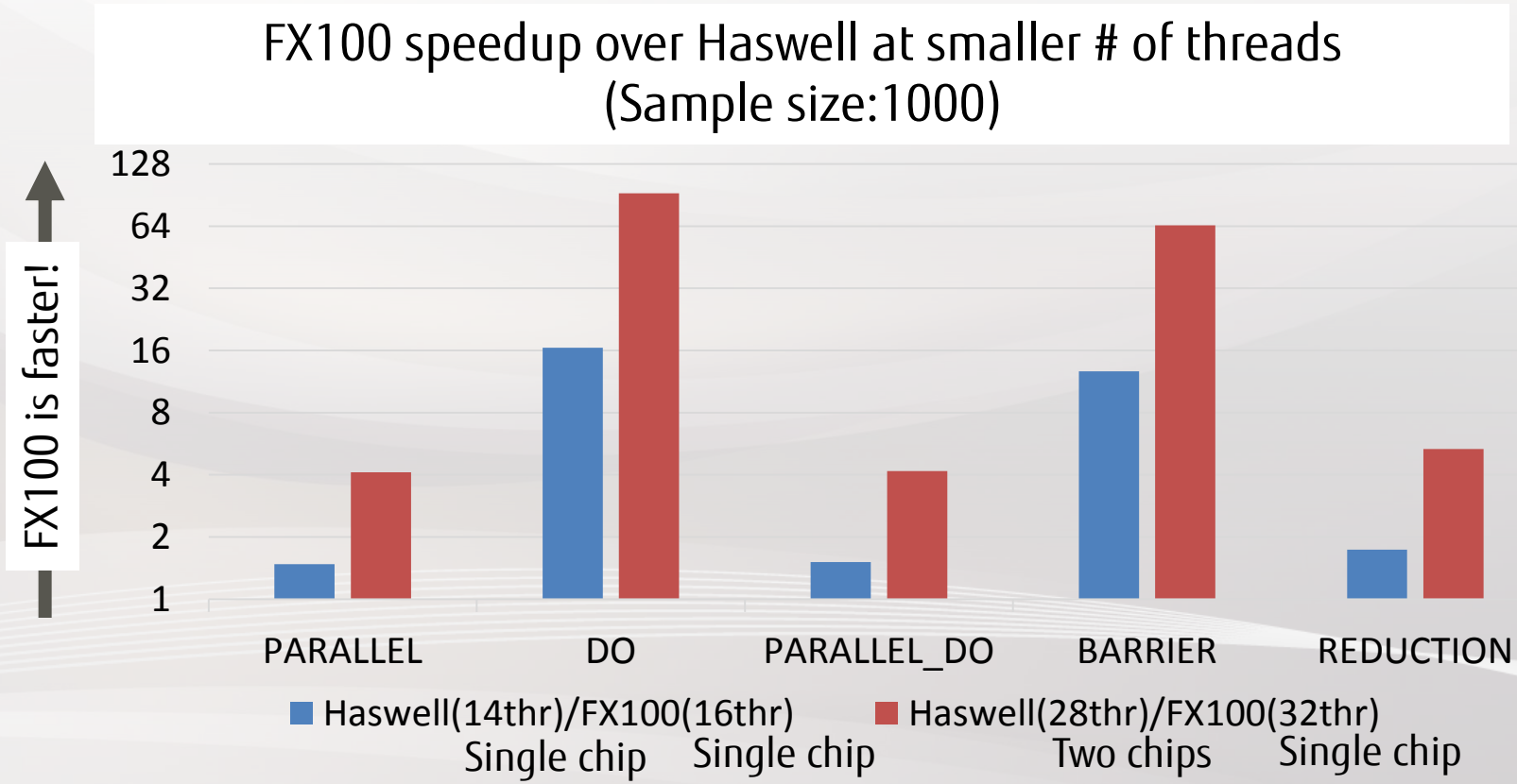
**(Problem size: SP 4098x130x130, parallelized by *j*)**



# CMG & SMaC Effect on OpenMP Microbenchmark

FUJITSU

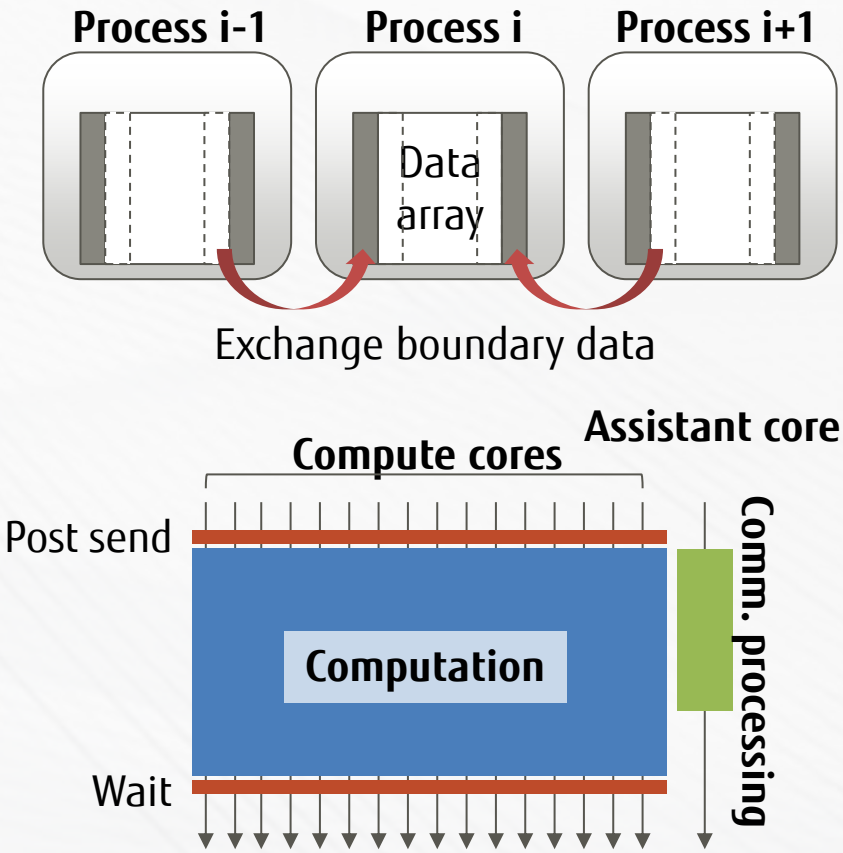
- FX100 outperforms Haswell due to the optimal implementation of the OpenMP library and the hardware inter-core barrier in the CPU
  - A single chip CPU of FX100 is effective (larger gap at 32 vs 28 threads)

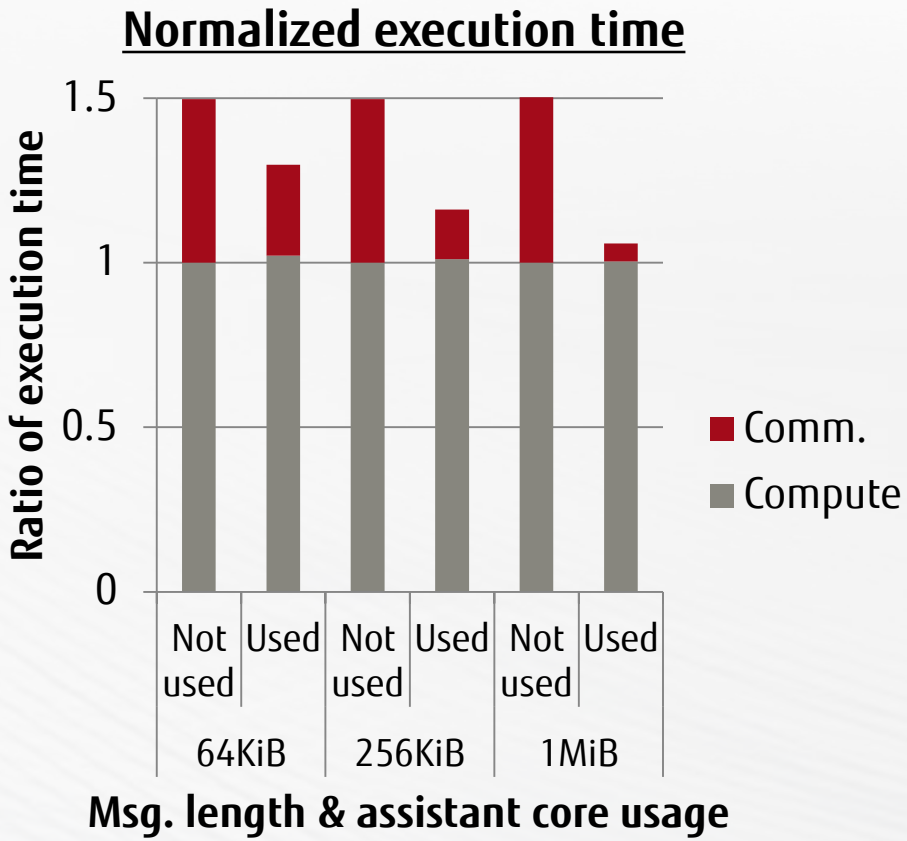



Configuration of Haswell: 2chips/node, E5-2697 v3 @ 2.60GHz

# Overlapping Execution of Non-blocking Comm.

- An assistant core is used in the MPI library
  - Boundary data transfer of stencil code



# Application Evaluation

FUJITSU

- CCS QCD Miniapp
  - Quantum chromodynamics, a linear equation solver with a large sparse coefficient matrix appearing in a lattice QCD problem
- NAS parallel benchmarks FT class C by OpenMP parallel
  - Time integration of a 3D partial differential equation using FFT (512^3)
- MHD
  - Simulation of Jovian and Kronian magnetosphere and space weather
- GT5D
  - Gyrokinetic toroidal 5D Eulerian code

# CCS QCD "Miniapp" (OpenMP single node)

FUJITSU

- Compiler improves cache hit rate using hints of directive & sector cache, better than FX10

**Typical implementation to enable sector cache**

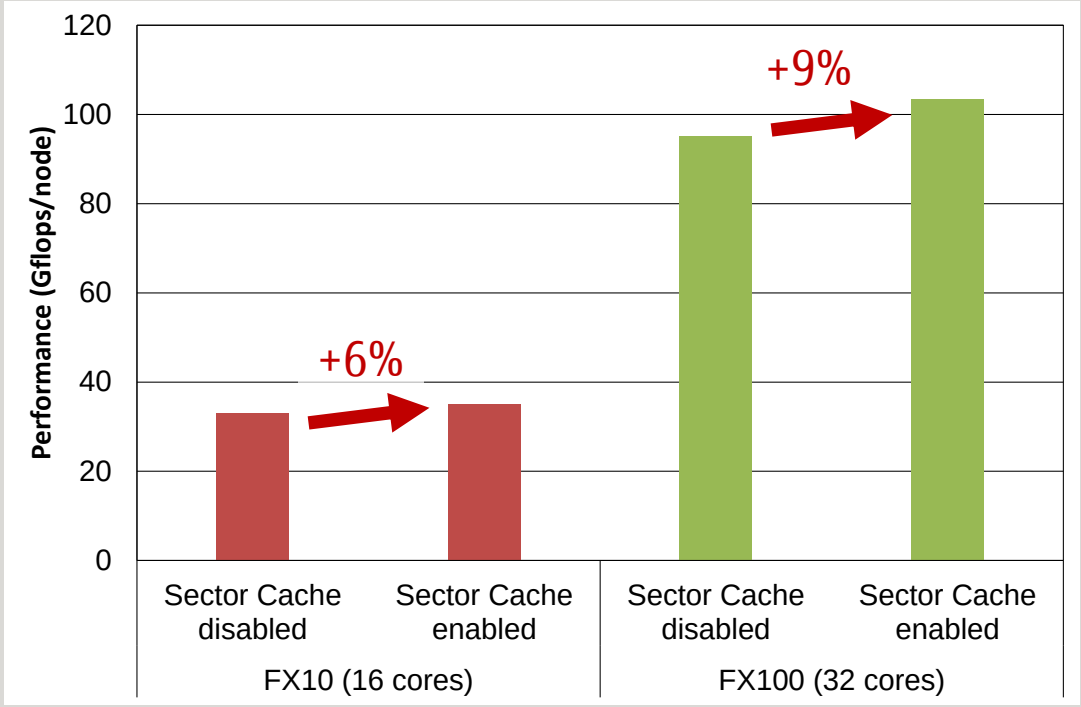
```
!OCL CACHE_SECTOR_SIZE(19,5)
!OCL CACHE_SUBSECTOR_ASSIGN(ue,uo,yde,fclinve)
!$OMP PARALLEL DO SCHEDULE(STATIC,1)
 do ix=1,NX
  do iy=1,NY
   do iz=1,NZ
     ...
     gy11=yo(...,iy+1,...)+...
     ...
     gy11=yo(...,iy-1,...)+...
     ...
   enddo
  enddo
 enddo
!$OMP END PARALLEL DO
!OCL END_CACHE_SUBSECTOR
!OCL END_CACHE_SECTOR_SIZE
```

Reserve L2$ 2.5MB for sector1

Assign single use data to sector1
Reusable data would be stored in sector0

Size: $32^4$
https://github.com/fiber-miniapp/ccs-qcd

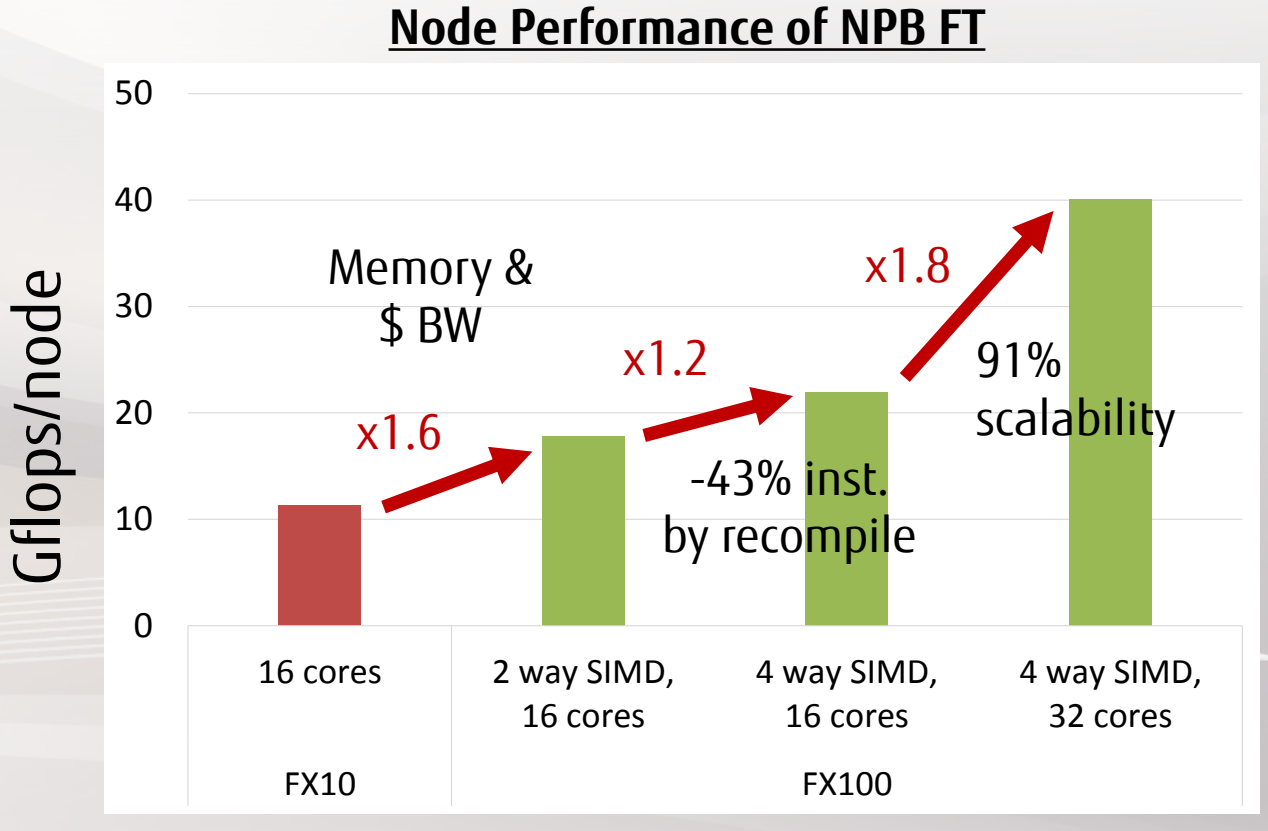
**Node Performance of CCS QCD**



# NAS Parallel Benchmark FT (OpenMP Single Node)

FUJITSU

- 3.5x Improvement per node over the FX10
  - 256bit SIMD x SMaC high performance node
  - Scalable thread performance

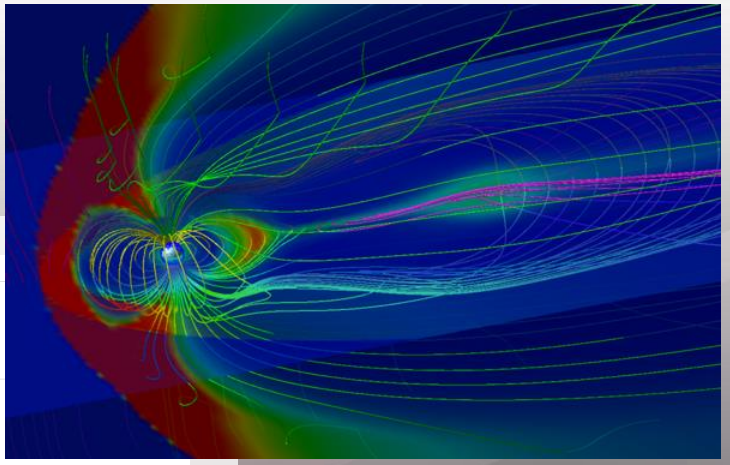


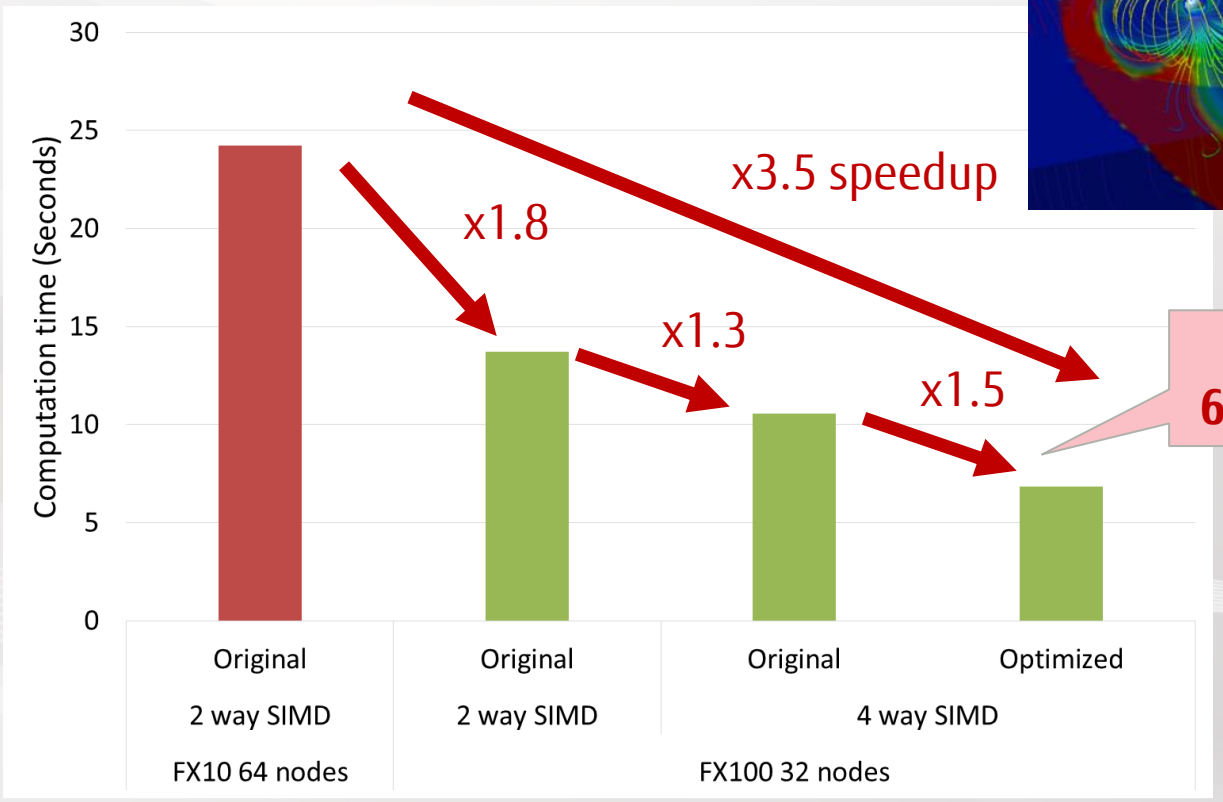

NAS Parallel Benchmarks Ver. 3.3.1 OpenMP Class C

# MHD Sim. of Jovian and Kronian Magnetosphere & Space Weather

FUJITSU

- Binary of FX10 runs on FX100 (1.8x)
- Recompile for FX100 utilizes 4 way SIMD (1.3x)
- Source optimization for FX100 attains 3.5x of FX10 at the same # of cores (1,024 cores)



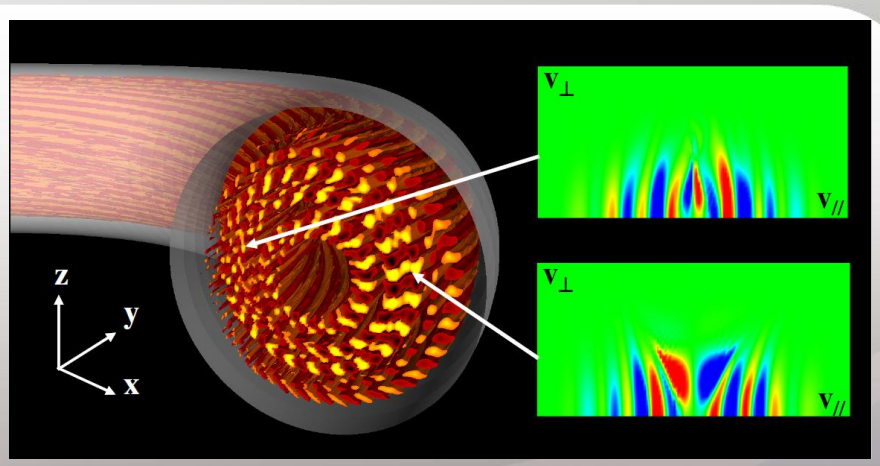



(Evaluated under the collaboration with Dr. Fukazawa @ Kyoto Univ.)

# GT5D: Gyrokinetic Toroidal 5D Eulerian Code[1]

FUJITSU

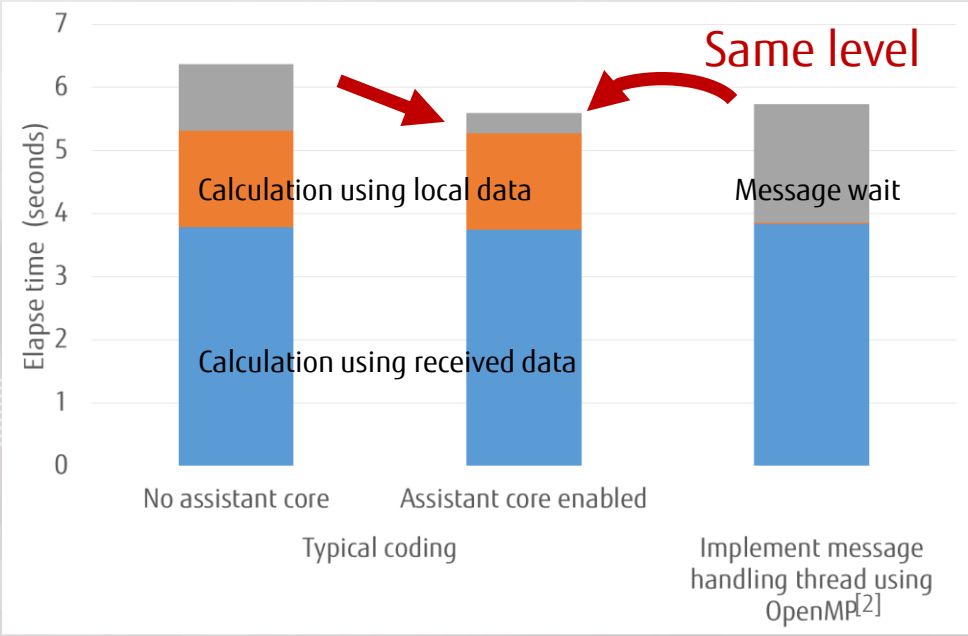
- Assistant core enables overlapping of communication and calculations
  - While computation progresses on compute cores using local data, the assistant cores handle communication tasks, independently and in parallel, until 'MPI_Waitall'
  - Minimal code modification retains portability and maintainability, free from the headache of manual code optimization for the same level of performance improvement




By courtesy of JAEA

**Typical implementation of the overlapping of communication and computation**

```
   MPI_Isend
   MPI_Irecv
!$OMP PARALLEL DO
   do i=...
      Computation using local data
      (independent from the communication)
   enddo
!$OMP END PARALLEL DO
   MPI_Waitall
   do i=...
      Computation using received data
   enddo
```




Size :256x256x64x128, 16 threads x 64 process, Range 14dx

[1] Y. Idomura et al., Nuclear Fusion 49, 065029 (2009)

[2] Y. Idomura et al., Int. J. HPC Appl. 28, 73 (2014)

# Summary, FX100, and Exascale...

FUJITSU

## FX100 Detailed Evaluation Unveiled

- Refined architectural concept **"SMaC"** is presented and evaluated
- Great application compatibility and scalable performance

## FLAGSHIP 2020 Basic Design has been Completed

- High application performance efficiency by consistent and conclusive approach

**PRIMEHPC Series**

**Exascale**


### K computer

VISIMPACT
SIMD extension HPC-ACE
Direct network Tofu

CY2010~
**128GF, 8-core/CPU**

### FX10

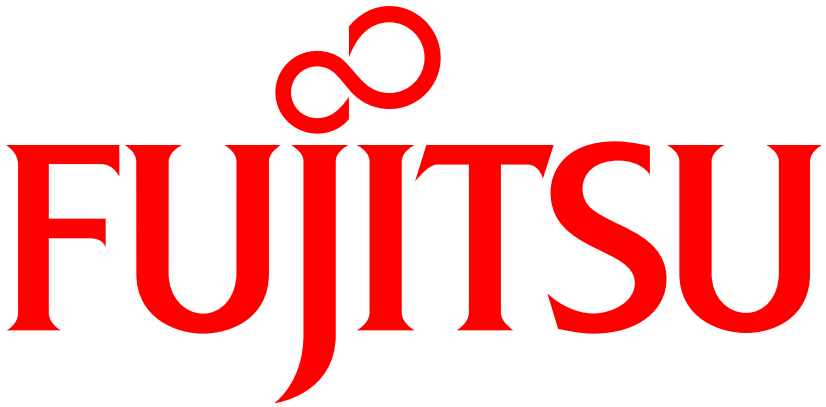
VISIMPACT
HPC-ACE
Direct network Tofu

CY2012~
**236.5GF, 16-core/CPU**

### FX100

SMaC
Tofu interconnect 2
HMC & Optical connections

CY2015~
**1TF~, 32-core/CPU**




shaping tomorrow with you