FUJITSU

shaping tomorrow with you

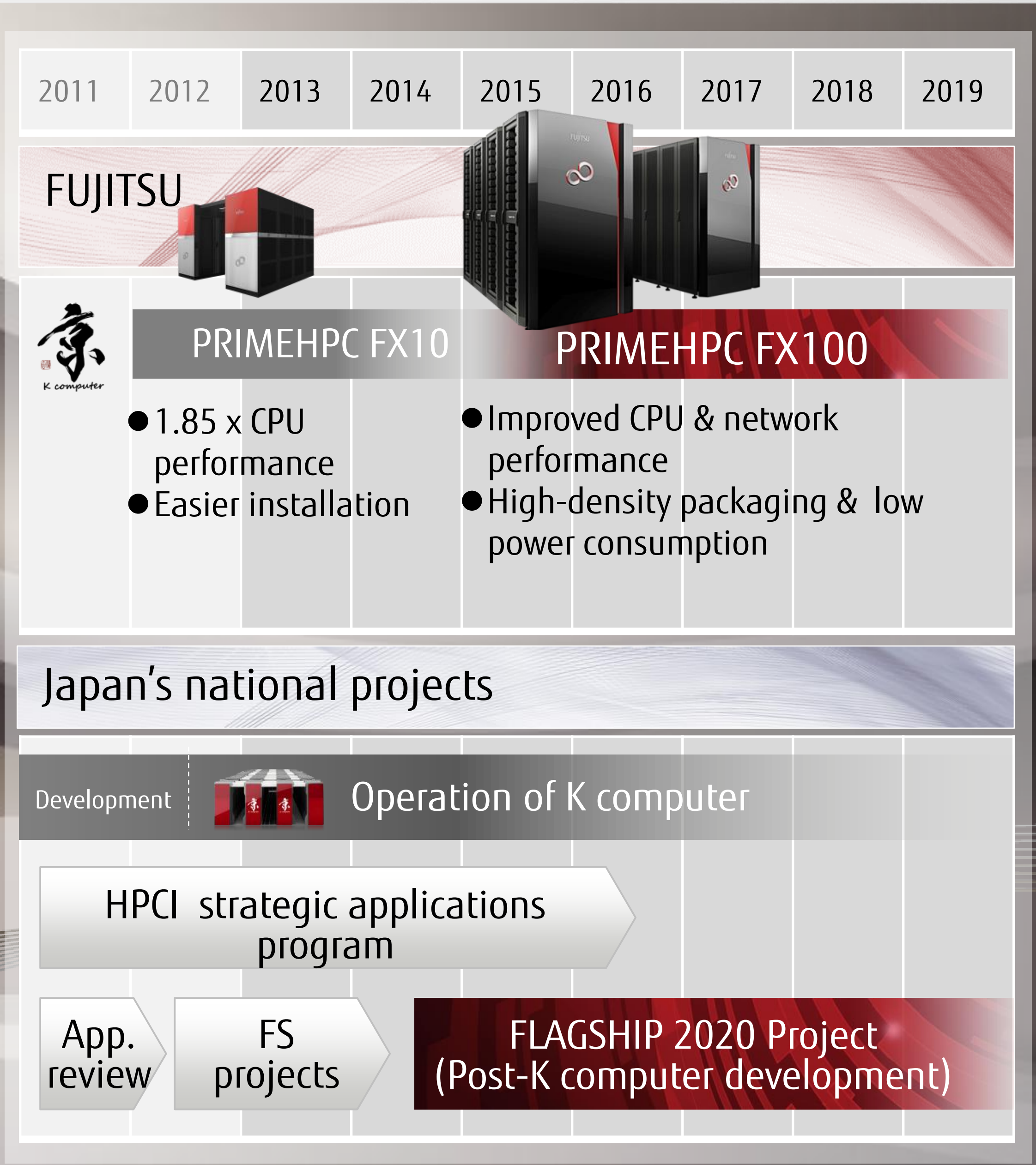
Fujitsu's new supercomputer, delivering the next step in Exascale capability

Toshiyuki Shimizu

November 19th, 2014

Copyright 2014 FUJITSU LIMITED

Past, PRIMEHPC FX100, and roadmap for Exascale



PRIMEHPC FX10

- ● 1.85 x CPU performance
- ● Easier installation

PRIMEHPC FX100

- ● Improved CPU & network performance
- ● High-density packaging & low power consumption

Japan’s national projects

Operation of K computer

HPCI strategic applications program

App. review

FS projects

FLAGSHIP 2020 Project
(Post-K computer development)

K computer and PRIMEHPC FX10 in operation

Many applications running and being developed for science and industries

PRIMEHPC FX100 is ready

CPU and interconnect inherits K computer architectural concept

Towards Exascale

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

Copyright 2014 FUJITSU LIMITED

PRIMEHPC FX100, design concept and approach

Provide steady progress for users

- Natural extent of performance profile of K computer and FX10
- Facilitate the evolution of applications

Challenge to state-of-art technologies for future generation

- 20nm CMOS technology
- HMC
- 25G optical connection

Copyright 2014 FUJITSU LIMITED

Natural extent of perf. profile of K and FX10

FUJITSU

Original high performance CPU for wide range of real applications

Highly scalable interconnect

	FX100	FX10	K computer
Double Flops / CPU	**Over 1 TF**	235 GF	128 GF
Single Flops / CPU	**Over 2 TF**	235 GF	128 GF
Max. # of threads	**32**	16	8
Memory / process	**32 GB**	32 GB	16 GB
SIMD width	**256 bit**	128 bit	128 bit
Byte per flop	**0.4 ~ 0.5**		
Interconnect	**Tofu 6D mesh/torus**		
Interconnect BW	**12.5 GB/s**	5 GB/s	5 GB/s

Compatibility with K computer and PRIMEHPC FX10

Binary compatible and make full use of performance by recompile
Compiler and libraries allow users to access new features

Copyright 2014 FUJITSU LIMITED

Features to facilitate evolution of applications

For further scalability

- Many core implementation and VISIMPACT
- Assistant cores for OS jitter reduction and offloading of house keeping tasks
- Tofu, 6D mesh/torus direct network for application optimization

Increase opportunities of SIMDization

- Compiler support of automatic/directive based detection and SIMDization
- Stride load/store, indirect load/store
- Permutation, Concatenate

Critical enhancements and optimizations

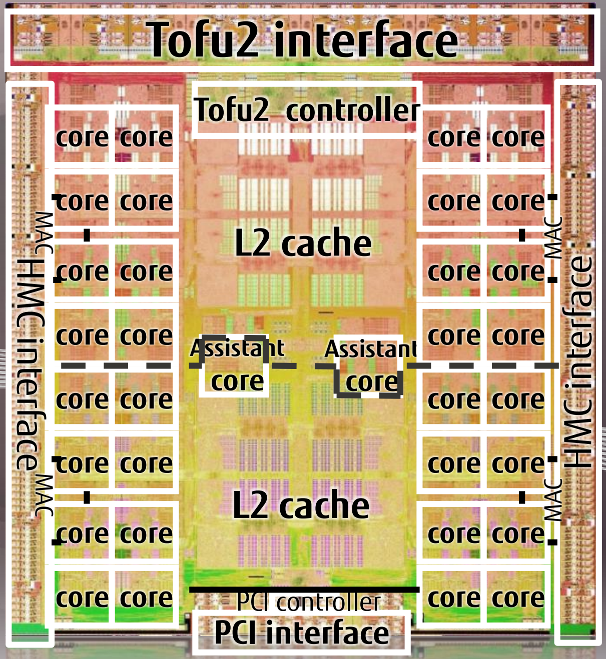
- Increase L1 ways and capacities, also allows flexible sector cache usage
- Increase out-of-order resources
- Implement better branch prediction
- Shorten the latency of message passing by cache injection

Copyright 2014 FUJITSU LIMITED

Fujitsu designed SPARC64 XIfx

- 256 bit wide SIMD
- 2x assistant cores
- HMC support
- Tofu interconnect 2 integrated

Architecture	SPARC V9 + HPC-ACE2
# of cores	32 compute + 2 assistant cores
Execution units	FMA x 2 (256 bit wide SIMD)
Cache	L1 inst. cache : 64 KB / core L1 data cache : 64 KB / core L2 cache : 24 MB / node
Main memory	32 GB / node
Memory bandwidth	240 GB/s (Read) + 240 GB/s (Write)
Technology	20nm CMOS, 3,750M Tr, 1,001 signal pins



Copyright 2014 FUJITSU LIMITED

Tofu interconnect 2, and other features

Tofu2

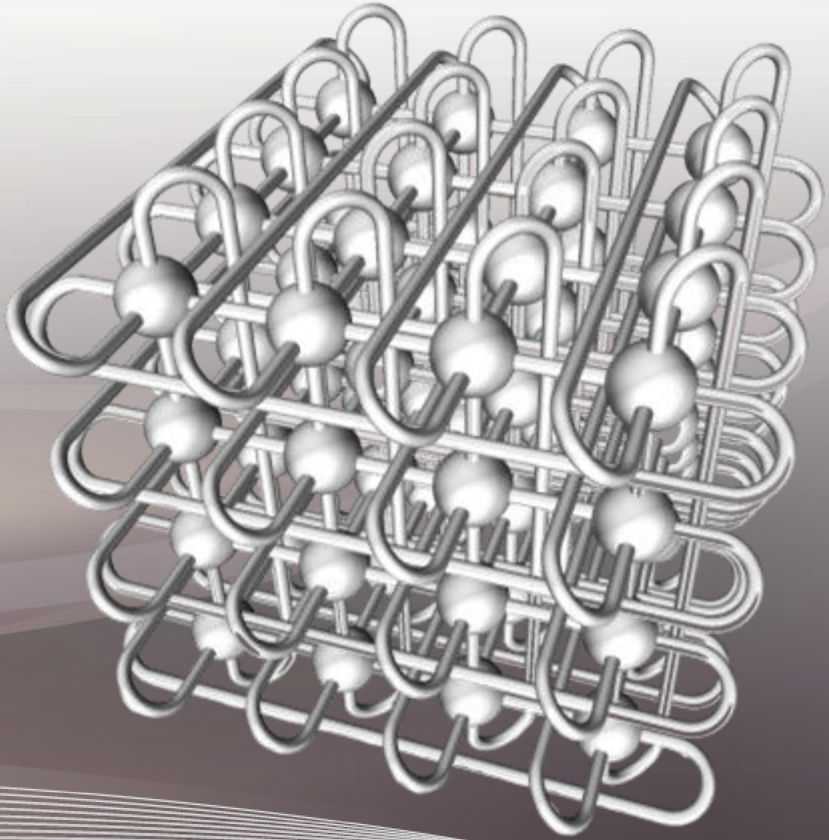
- Compatible with K computer
- Low latency collective communication utilizing multiple RDMA engines
- Hardware barrier
- Optical inter chassis connection

19 inch rack mountable chassis

- 12 nodes / 2U

Water cooling

- Including CPU, memory, optical modules
- 90% of parts are cooled by water



Copyright 2014 FUJITSU LIMITED

Feature and configuration of FX100

Tofu interconnect 2

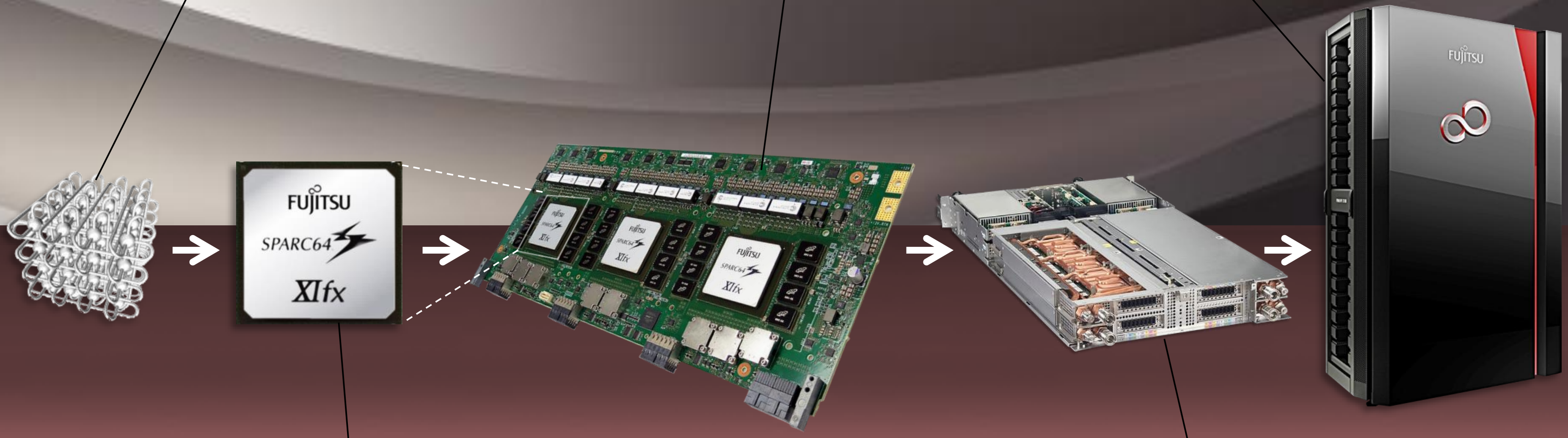
- 12.5 GB/s×2(in/out)/link
- 10 links/node
- Optical technology

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

Copyright 2014 FUJITSU LIMITED

System software for PRIMEHPC FX10 and FX100

FUJITSU

Tuned Linux OS for HPC applications

- Supports large pages and OS jitter minimization

Combination of self-developed software and customized OSS

- System management software and languages are self-developed
- File system and MPI are developed based on OSS and the results were fed back to the communities

Single system images with x86 and hybrid configurations

System management portal and HPC portal

Technical Computing Suite

Management

- System management
  - Single system image
  - Single action IPL
  - Fail safe capability
- Job management
  - Highly efficient scheduler

File system (FEFS)

- Lustre based
- Higher scalability (thousands of IO servers)
- Higher IO performance (1.4 TB/s)

Programing environment

- Compiler
  - Fortran, XPF, C, C++
  - Automatic parallelization
  - SIMD support
- MPI: OpenMPI based
- Tools and math libraries

Copyright 2014 FUJITSU LIMITED

Features and evaluations

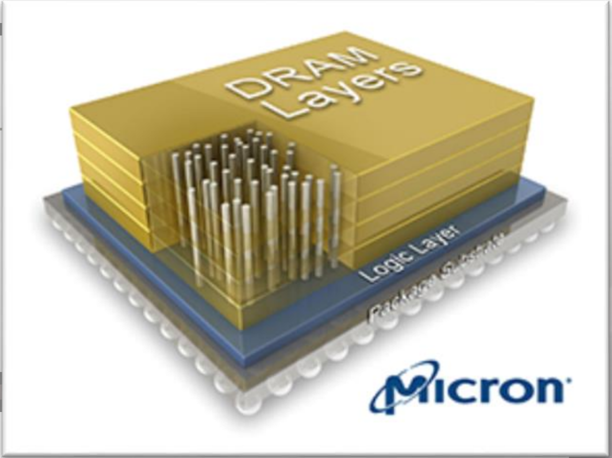
- HMC
- SIMD
- Tofu interconnect 2
- VISIMPACT
- Assistant core
- Real applications

Copyright 2014 FUJITSU LIMITED

Hybrid Memory Cube (HMC) support

HMC

- Higher density at BW
- Higher capacity and higher BW at package
- Lower power consumption at BW



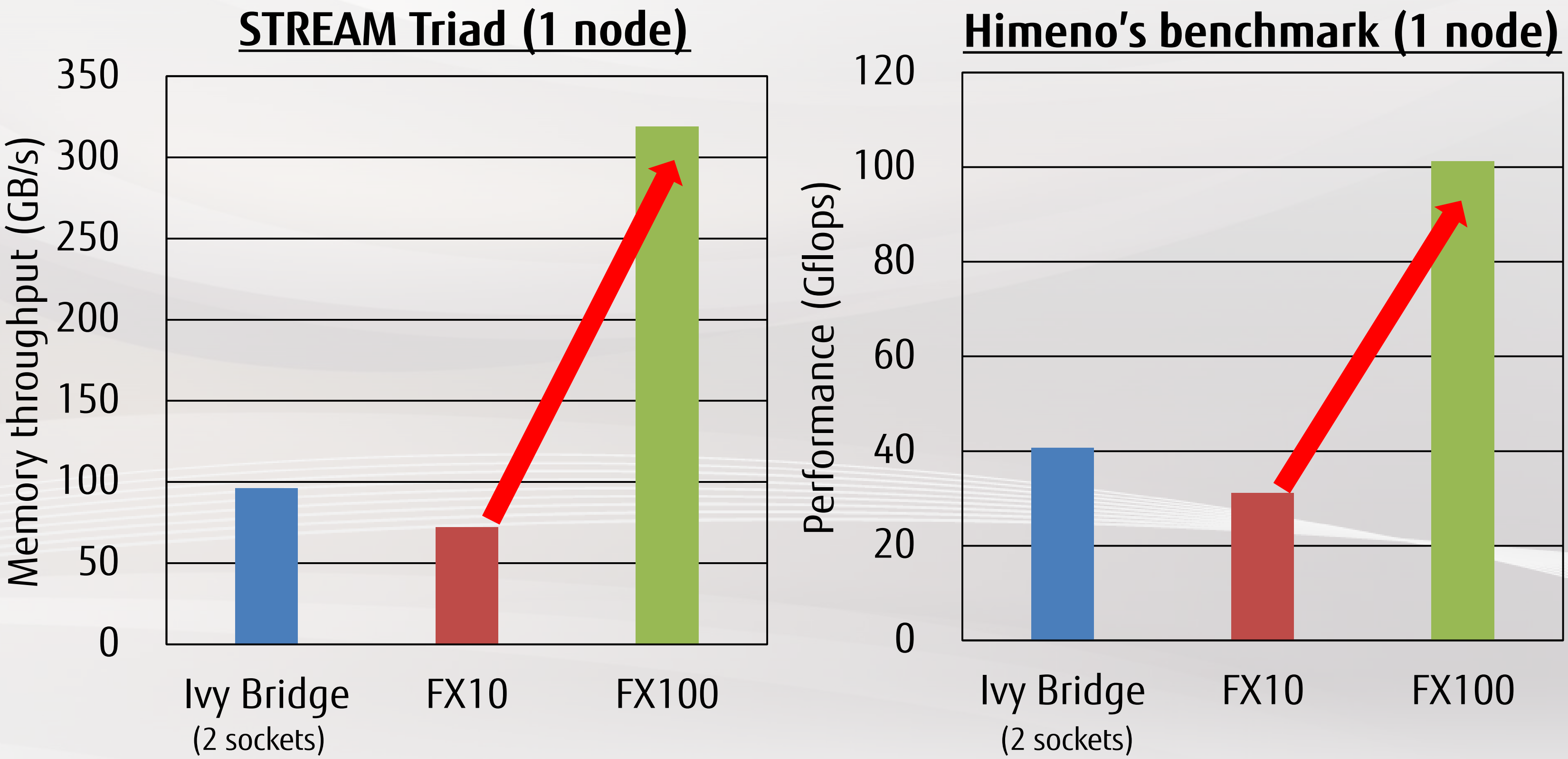
Comparable capacity and bandwidth to those of K computer and FX10

Per CPU count	Capacity	Bandwidth
HMC x 8	**32GB**	**480GB/s**
DDR4-DIMM x 8	32~128GB	154GB/s
GDDR5 x 16	8GB	320GB/s

Copyright 2014 FUJITSU LIMITED

Improving memory throughput

By using HMC, node memory throughput increase 3x, 4x

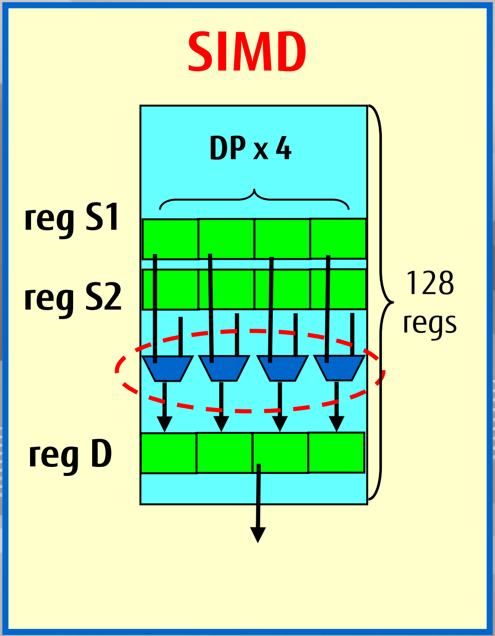


Copyright 2014 FUJITSU LIMITED

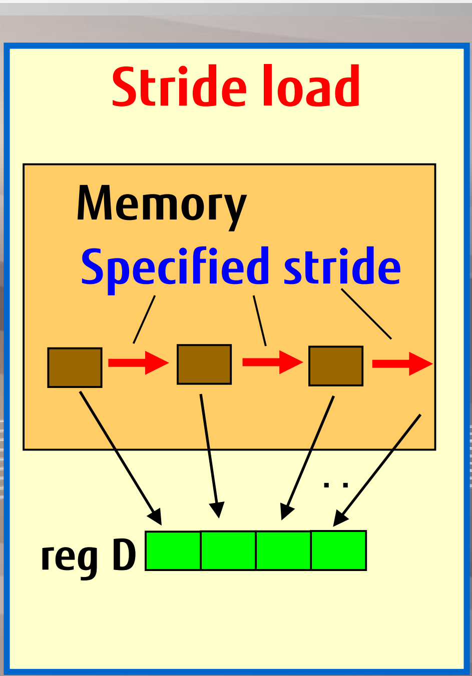
SIMD extension of HPC-ACE2

256-bit wide SIMD with 128 FPRs

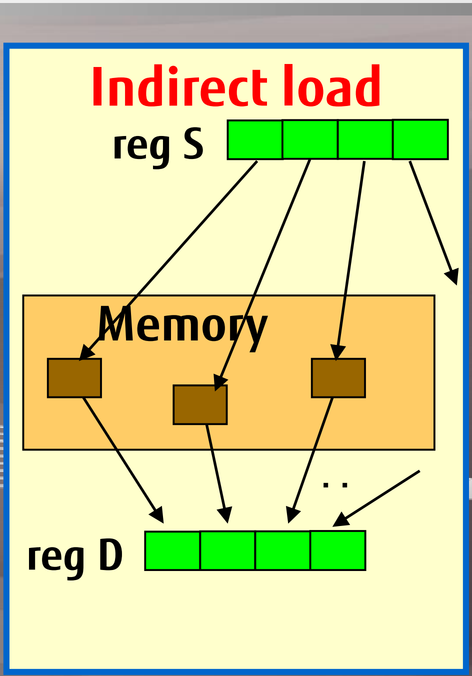
- Double precision x 4, single precision x 8, 8-byte integer x 4
- Stride Load/Store
- Indirect (list) Load/Store
- Permutation, Concatenate



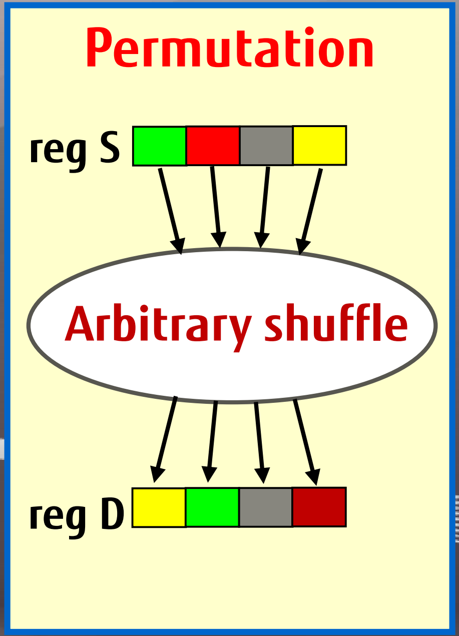
SIMD



Stride load



Indirect load



Permutation

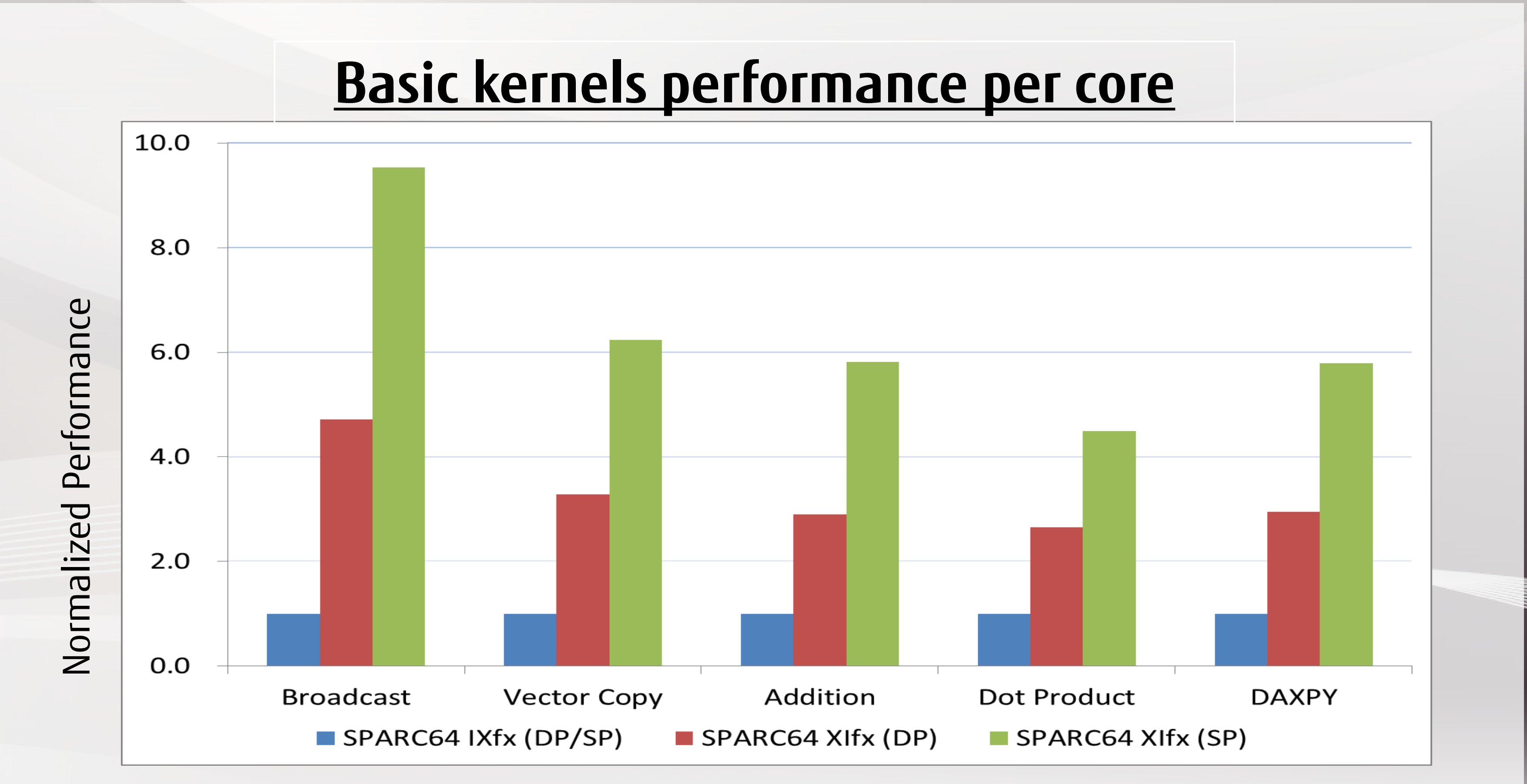
Copyright 2014 FUJITSU LIMITED

Wider SIMD extensions

FUJITSU

DP 3x, SP 6x faster than FX10 in basic kernels

- Improved L1 cache pipelines



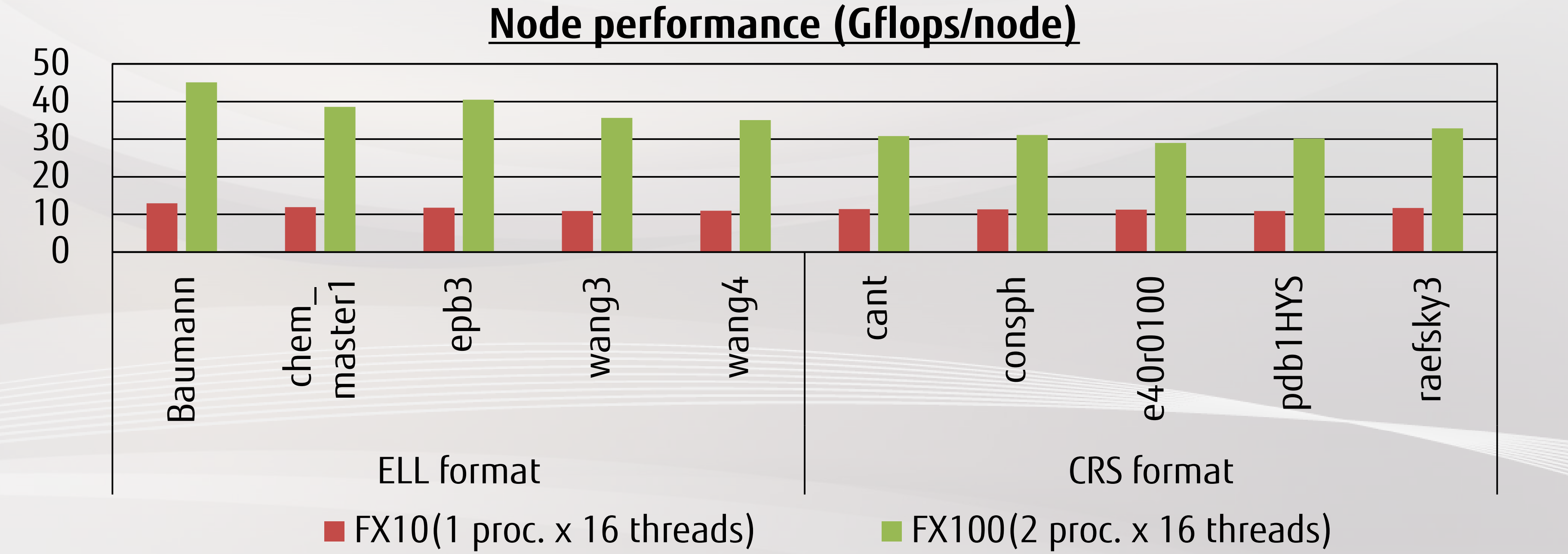
Copyright 2014 FUJITSU LIMITED

Effect of indirect LOAD

Sparse Matrix-Vector Multiplication

- 10 matrices from Florida Sparse Matrix Collection†
- ELL (ELLPACK) format and CRS (Compressed Row Storage) format

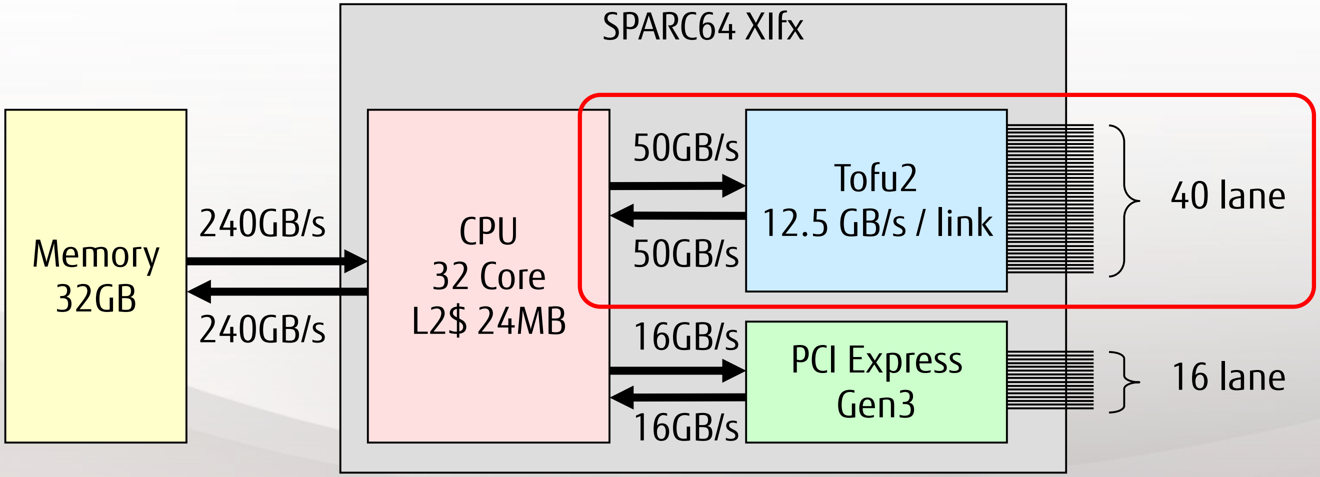
- **3.4x faster in ELL format, 2.7x faster in CRS format**
- **Indirect LOAD instruction helps SIMD acceleration**



† http://www.cise.ufl.edu/research/sparse/matrices/

Copyright 2014 FUJITSU LIMITED

Tofu2

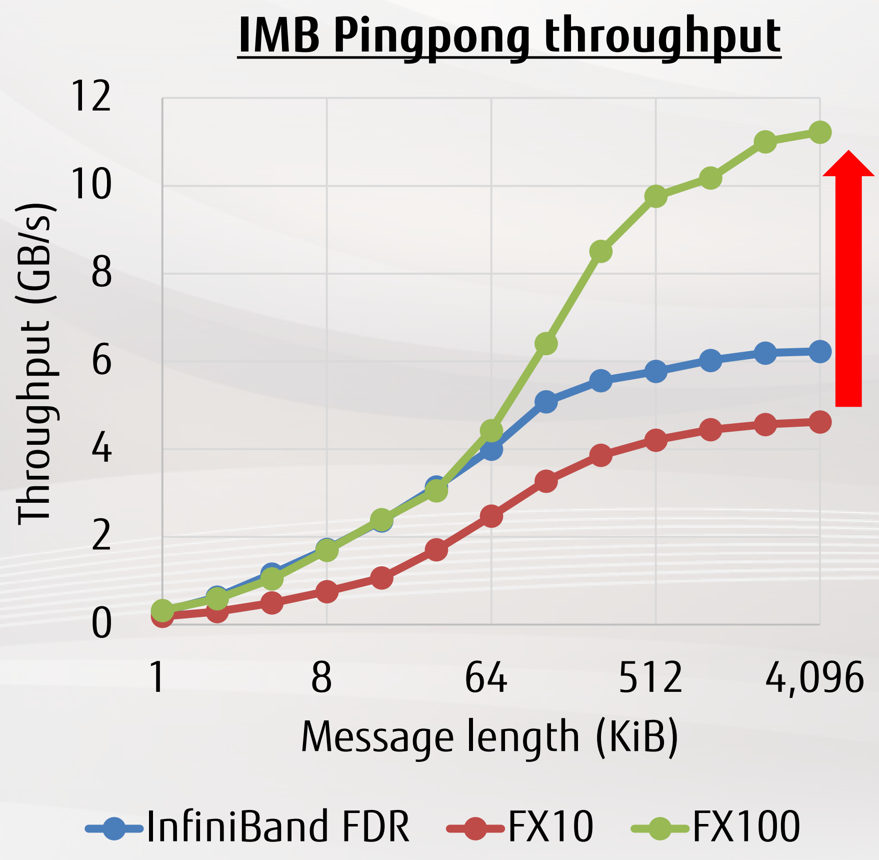


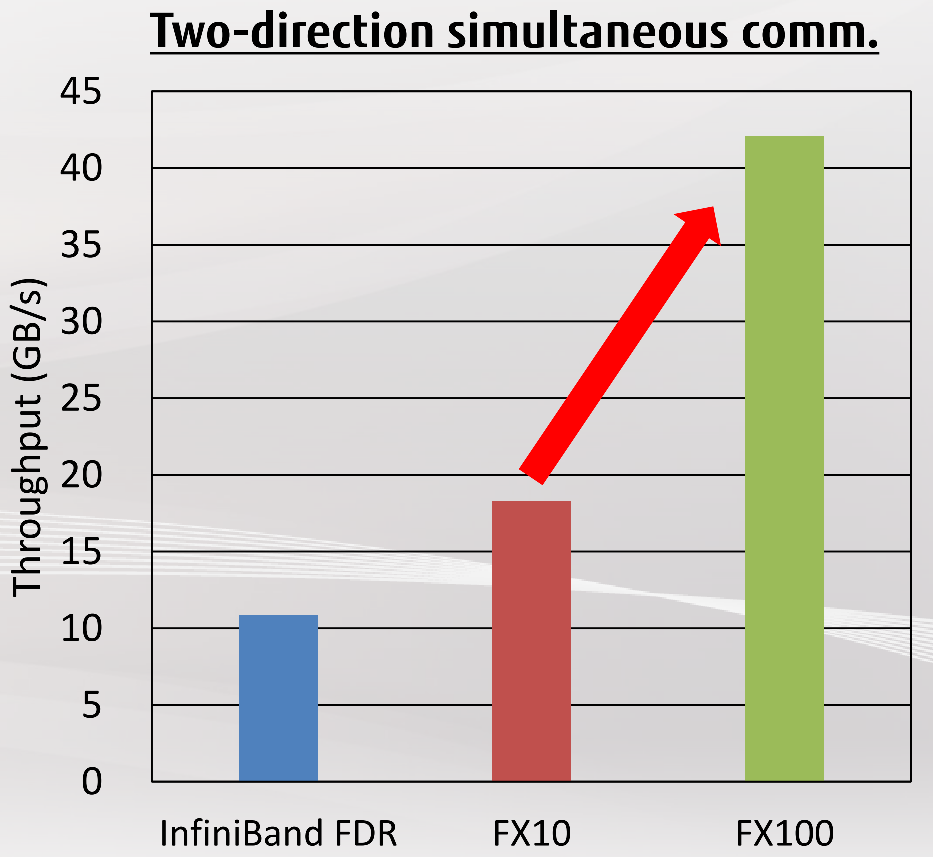
	Tofu	Tofu2
System	K computer and FX10	FX100
CPU	SPARC64 VIIIfx/IXfx	SPARC64 XIfx
Integration	No, dedicated LSI ICC is required	Yes, integrated into the CPU chip
Topology	6D mesh/torus topology	←
Link bandwidth	5 GB/s (6.25 Gbps x 8 lanes x 10 dirs)	12.5 GB/s (25 Gbps x 4 lanes x 10 dirs)
Node bandwidth	20 GB/s x in/out	50 GB/s x in/out
Other features	-	Cache injection, atomic operation Optical connection(2/3 of links are optical)

Copyright 2014 FUJITSU LIMITED

Communication performance

- Throughput improves 2.4x higher than FX10
- Good simultaneous multiple direction transfer

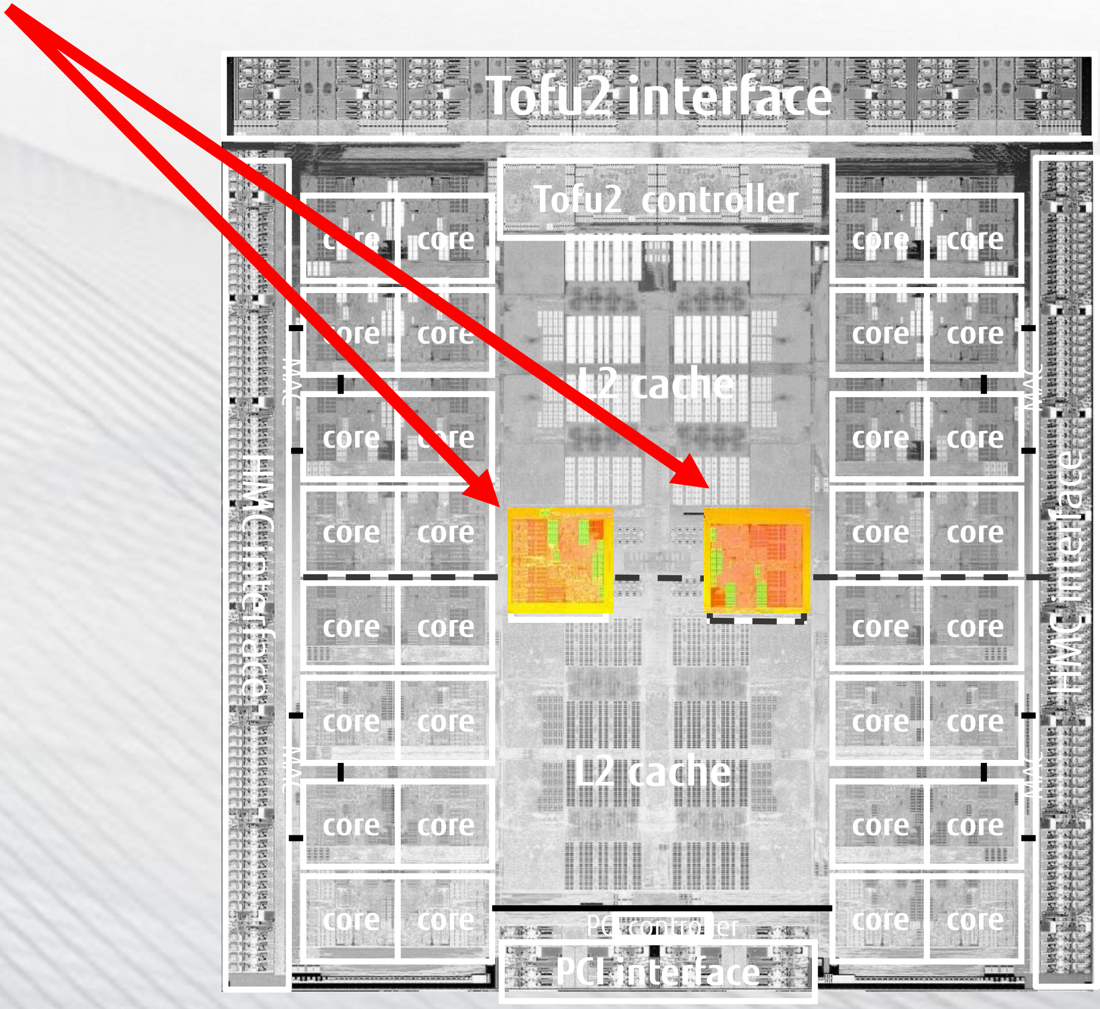




Copyright 2014 FUJITSU LIMITED

Assistant core

- Two assistant cores available

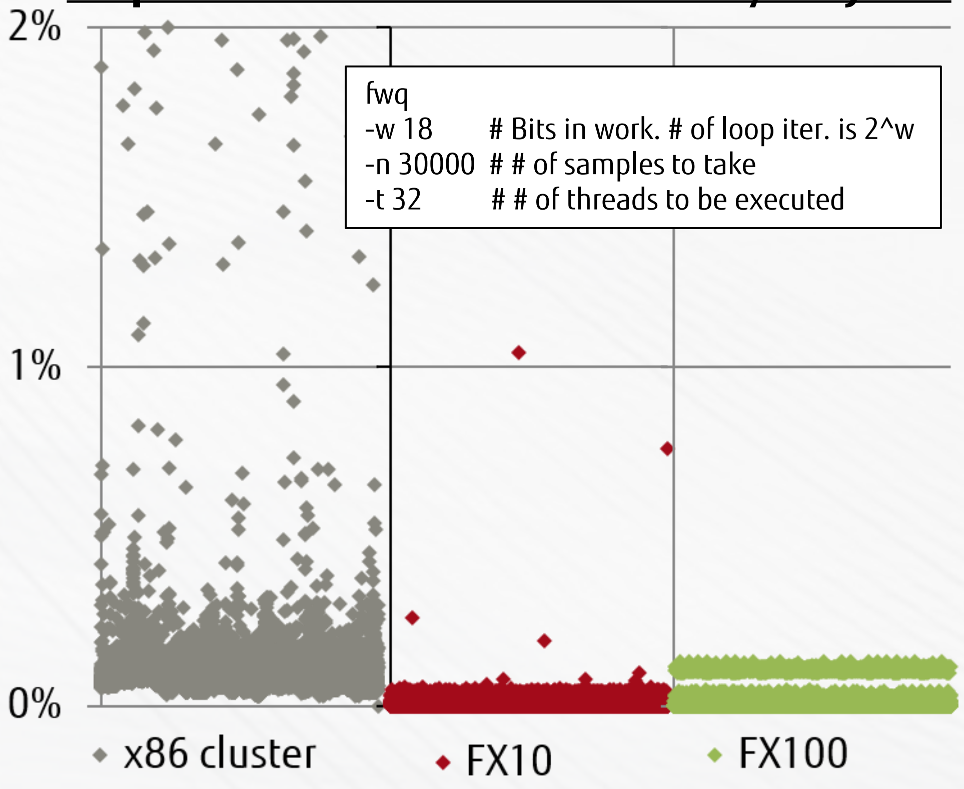


Copyright 2014 FUJITSU LIMITED

OS jitter reduction utilizing assistant core

Offloading of daemons, IO processing, MPI asynchronous communication to the assistant core reduces OS jitter

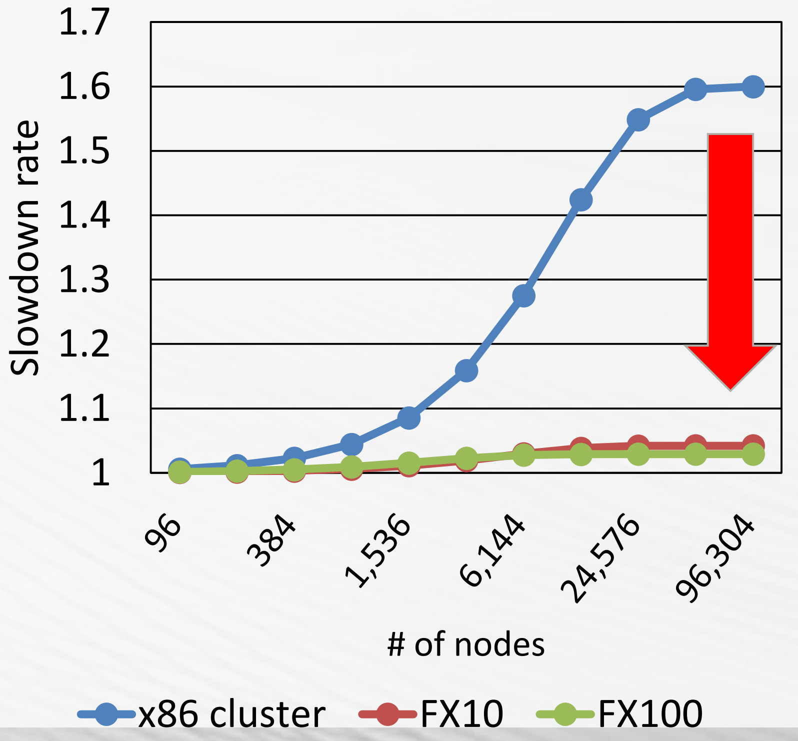
Dispersion of calc. time caused by OS jitter



```
fwq
-w 18       # Bits in work. # of loop iter. is 2^w
-n 30000  # # of samples to take
-t 32       # # of threads to be executed
```

(OS optimization undergo)

Estimated slowdown caused by OS jitter (comm. interval=1ms)

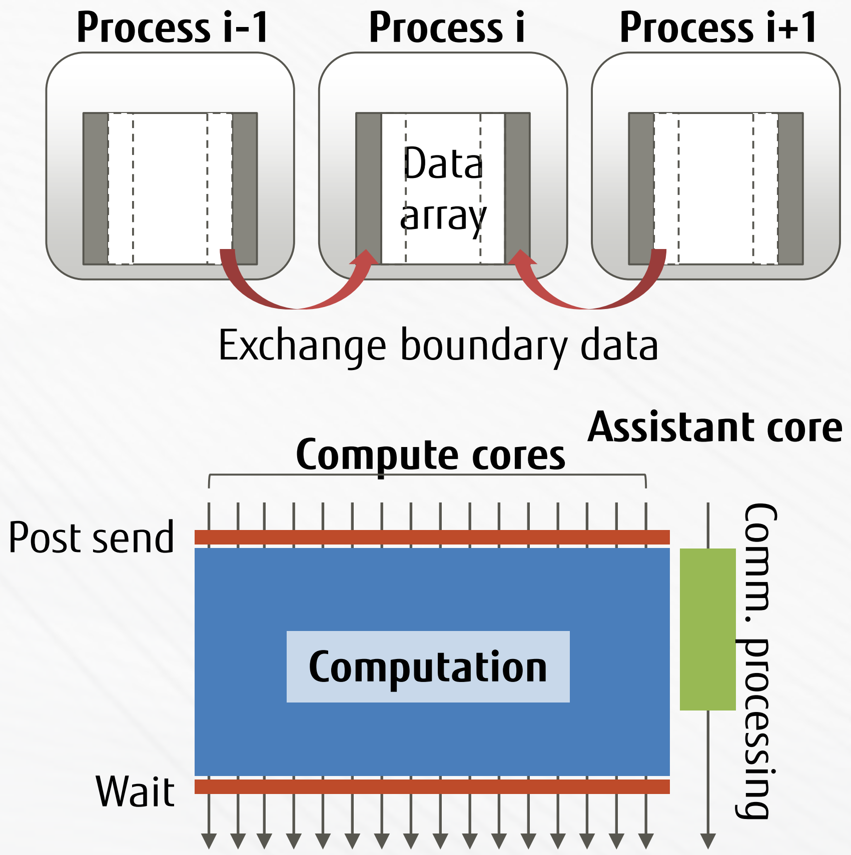


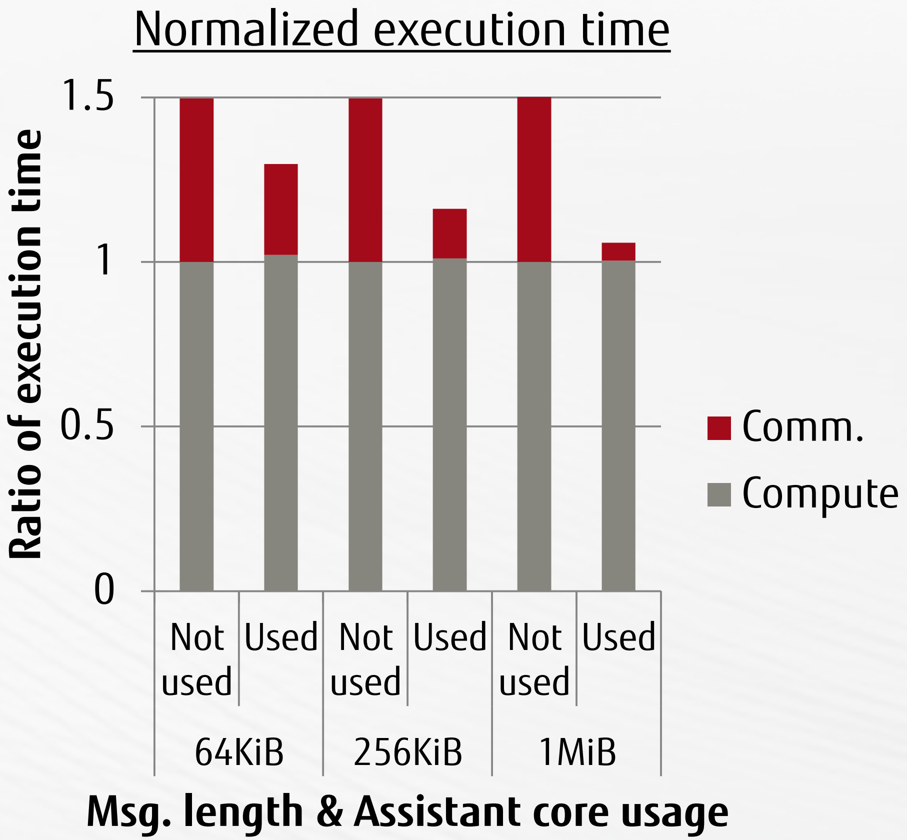
Copyright 2014 FUJITSU LIMITED

Overlapping execution of non-blocking comm.

Assistant core is used in MPI library

Boundary data transfer of stencil code





Copyright 2014 FUJITSU LIMITED

Application evaluation

- NAS Parallel Benchmarks FT Class C by OpenMP parallel
- CCS QCD Miniapp †
- NICAM-DC-MINI ††

† https://github.com/fiber-miniapp/ccs-qcd

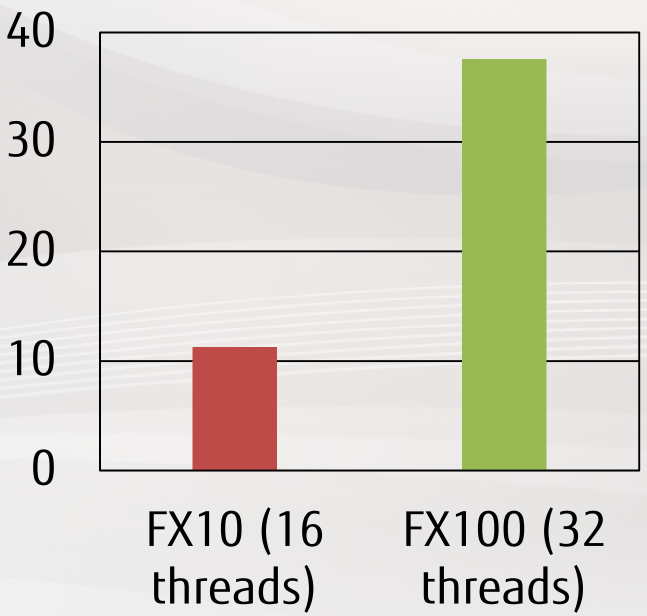
†† https://github.com/fiber-miniapp/nicam-dc-mini

NAS Parallel Benchmarks FT Class C (OpenMP)

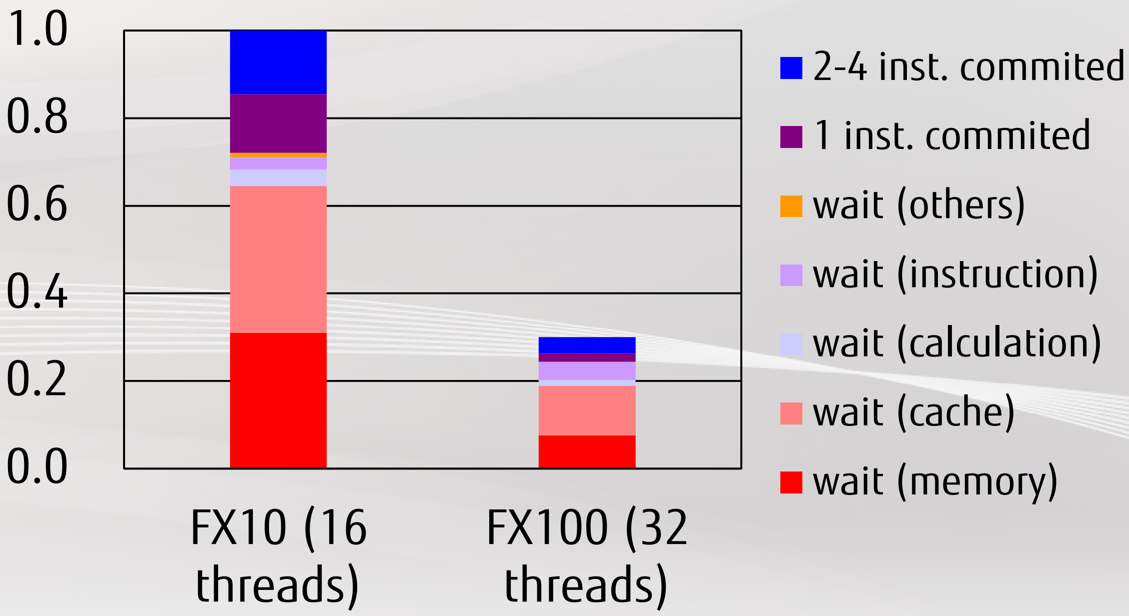
Time integration of a 3D partial differential equation using FFT (512^3)

- **3.3x faster on FX100 with 32 threads**
- **Node performance is enhanced by higher cache/memory throughput, as well as increased CPU cores and SIMD width**

Node performance (Gflops/node)



Breakdown of execution time

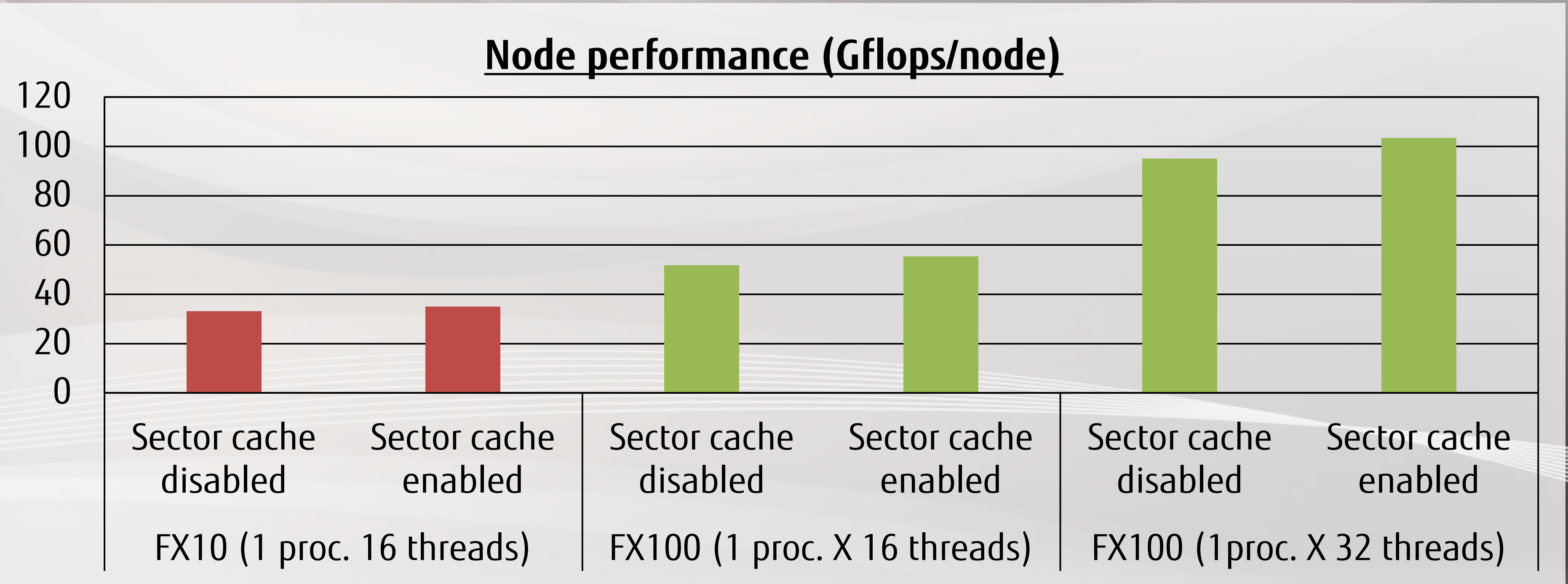


Copyright 2014 FUJITSU LIMITED

CCS QCD Miniapp†

A linear equation solver with a large sparse coefficient matrix appearing in a lattice QCD problem (32x32x32x32)

- **1.6x faster with 16 threads, 3.0x faster with 32 threads**
- **Enhanced memory bandwidth boosts the performance and Sector cache mechanism promotes data reuse on L2$**



† https://github.com/fiber-miniapp/ccs-qcd

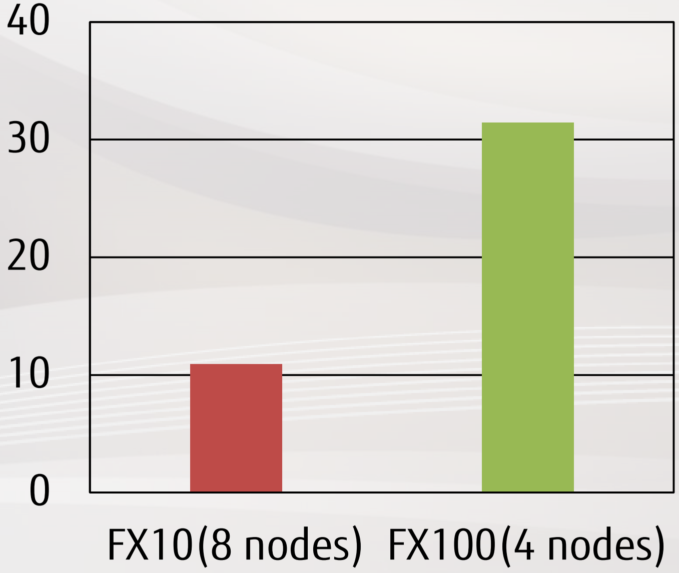
Copyright 2014 FUJITSU LIMITED

NICAM-DC-MINI†

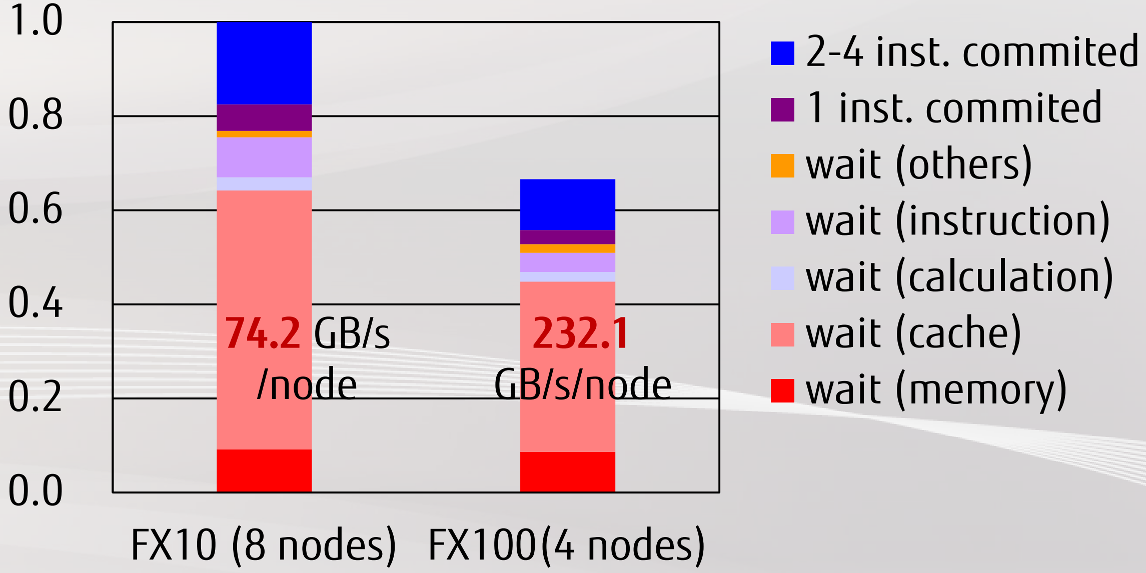
A miniapp based on NICAM-DC derived from NICAM
(Nonhydrostatic ICosahedral Atmospheric Model)

- **2.9x faster on FX100, including communications**
- **3.1x higher memory throughput in "Vertical Implicit" calculation, speeding up 1.5x with half the number of FX10 nodes**

Node performance (Gflops/node)



Breakdown of execution time ("vi_path2" routine)



† https://github.com/fiber-miniapp/nicam-dc-mini

Copyright 2014 FUJITSU LIMITED

Summary, FX100

FX100 provides steady progress for users, natural extent of perf. profile

- Single CPU/node architecture for multicore
- Good Byte/flop and scalability

Leads apps toward highly scalable and introduces new technologies

- Original CPU and interconnect
- Support for tens of millions of cores
  (VISIMPACT, Collective comm. HW)

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
VISIMPACT
HPC-ACE2
Tofu interconnect 2
HMC & Optical connections

CY2015~
1TF~, 32-core/CPU

Copyright 2014 FUJITSU LIMITED

Summary, FX100, and Exascale...

FLAGSHIP 2020 basic design has started

- High application perf. efficiency is our target
- Keep similar approach for application compatibility

Post-K computer

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

VISIMPACT
HPC-ACE2
Tofu interconnect 2
HMC & Optical connections

CY2015~
1TF~, 32-core/CPU

Copyright 2014 FUJITSU LIMITED

FUJITSU

shaping tomorrow with you