FUJITSU

shaping tomorrow with you

The First SVE Enabled Arm Processor: A64FX and Building up Arm HPC Ecosystem

Shinji Sumimoto, Ph.D.
Next Generation Technical Computing Unit
FUJITSU LIMITED

Jan. 14th, 2019

Outline of This Talk

FUJITSU

- The First SVE Enabled Arm Processor: A64FX
  - A64FX: High Performance Arm CPU
- Arm HPC Ecosystem Development
  - Arm HPC Software Topics
    - Activities with Arm, Linaro and OSS Community
    - OSS Application Porting Updates

FUJITSU

A64FX: High Performance Arm CPU

- From presentation slides of Hotchips 30th and Cluster 2018
- Inheriting Fujitsu HPC CPU technologies with commodity standard ISA

A64FX Chip Overview

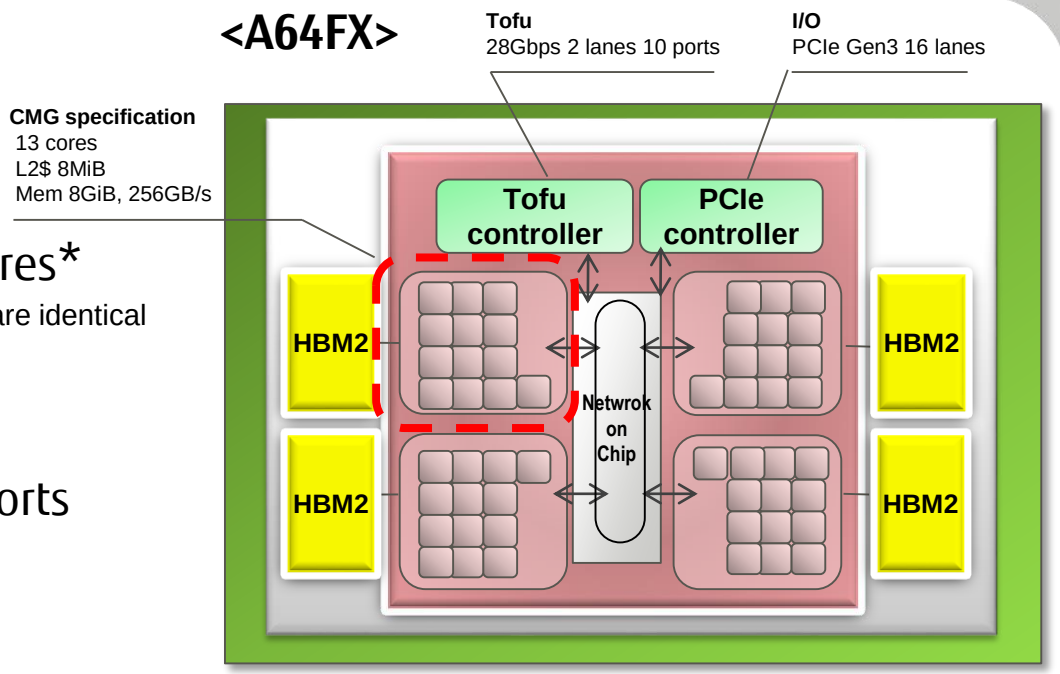
FUJITSU

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
- TofuD 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)

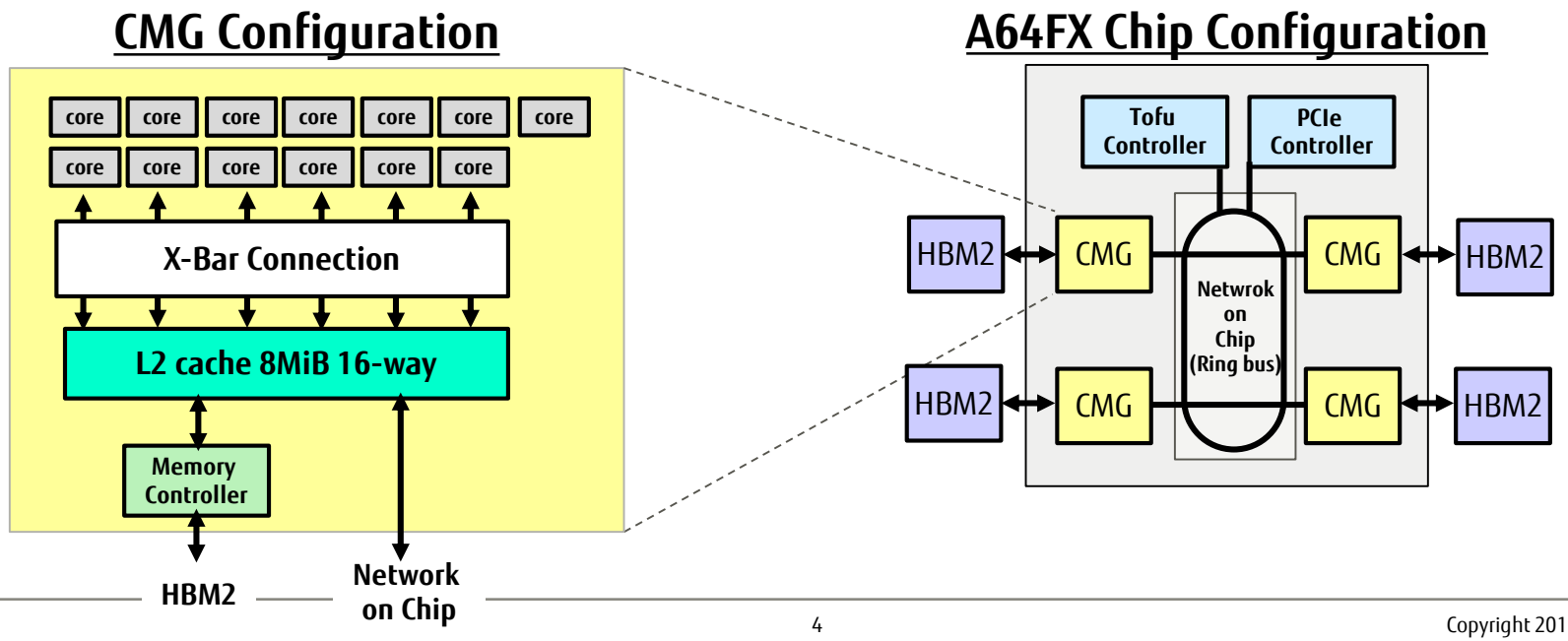
- >2.7TFLOPS (>90%@DGEMM)
- Memory B/W 1024GB/s (>80%@Stream Triad)

	A64FX (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 Many-Core Architecture

FUJITSU

- Consisting of 4 CMGs (Core Memory Group), ToFu and PCIe Controller
  - 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.
  - CMGs keep cache coherency by ccNUMA with on-chip directory
  - The X-bar connection realizes high efficiency for the L2 cache throughput
  - NUMA-aware software techniques enable linear scalability up to 48 cores
- Providing High I/O Performance by Wide Ring On-chip-network

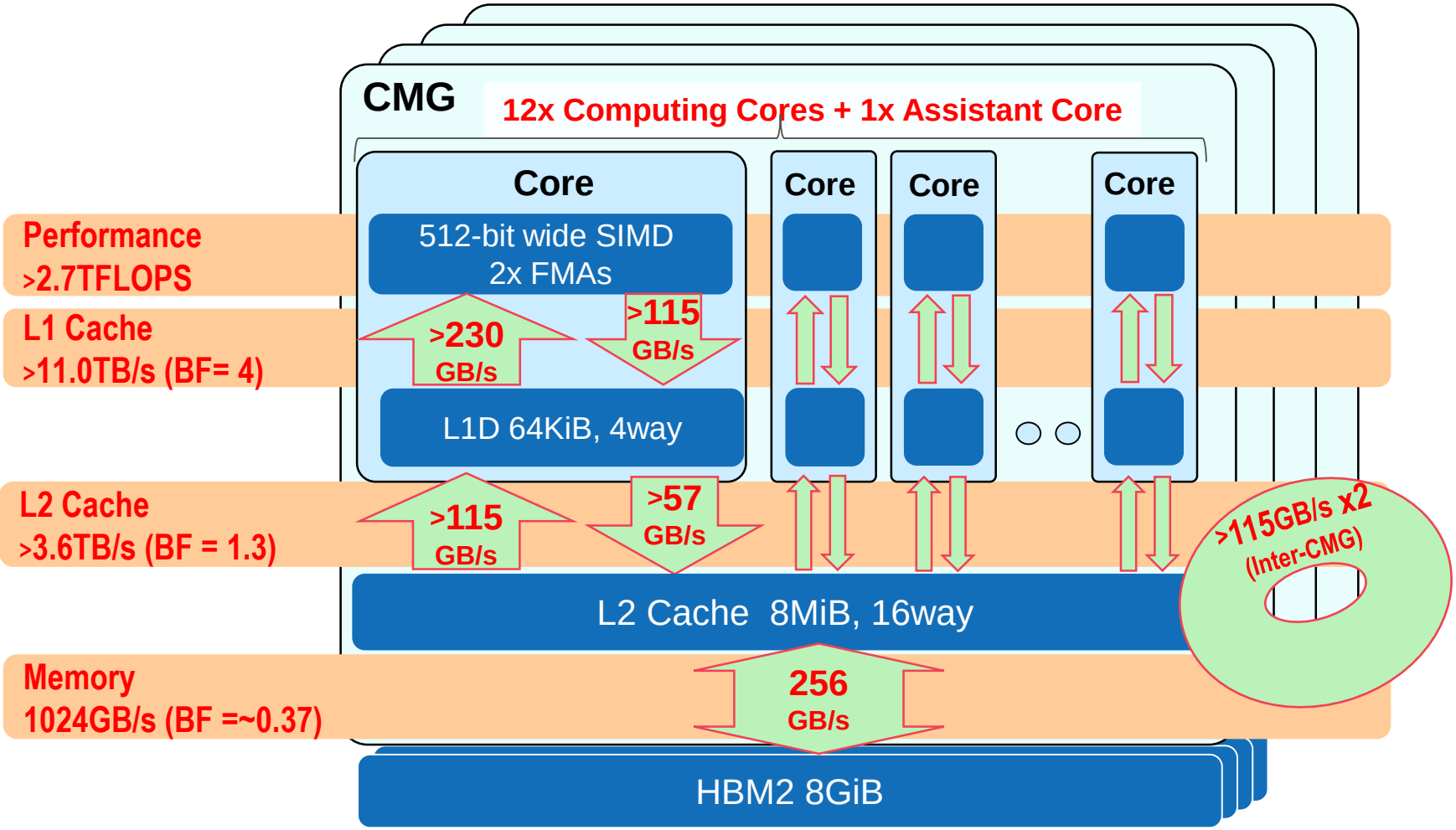


A64FX Memory System

FUJITSU

Extremely high bandwidth

- Asynchronous Processing in cores, caches and memory controllers
- Maximizing the capability of each layer's bandwidth



A64FX Core Features

FUJITSU

- Optimizing SVE architecture for wide range of applications with Arm including AI area by FP16 INT16/INT8 Dot Product
- Developing A64FX core micro-architecture to increase application performance

	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

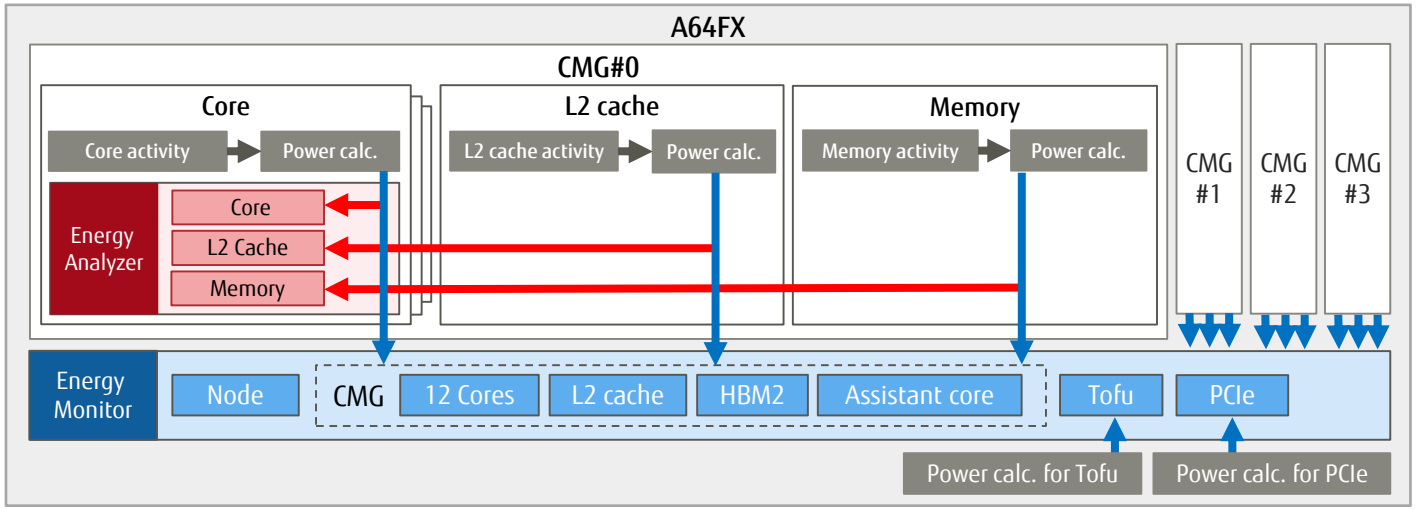
A64FX: Power monitor and analyzer

FUJITSU

- Energy monitor (per chip)
  - Node power via Power API*1 (~msec)
  - Averaged power of a node, CMG (cores, an L2 cache, a memory) etc.
- Energy analyzer (per core)
  - Power profiler via PAPI*2 (~nsec)
  - Fine grained power analysis of a core, an L2 cache and a memory

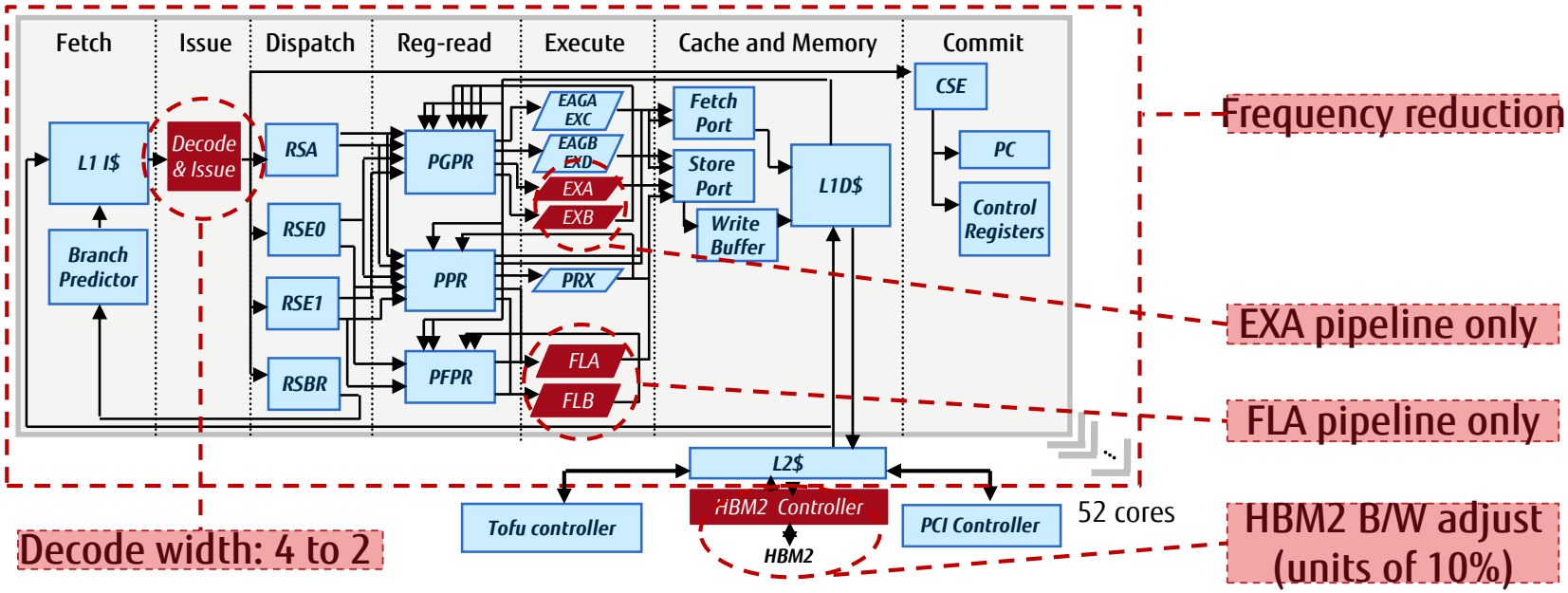
*1: Sandia National Laboratory

*2: Performance Application Programming Interface



A64FX: Power Knobs to reduce power consumption

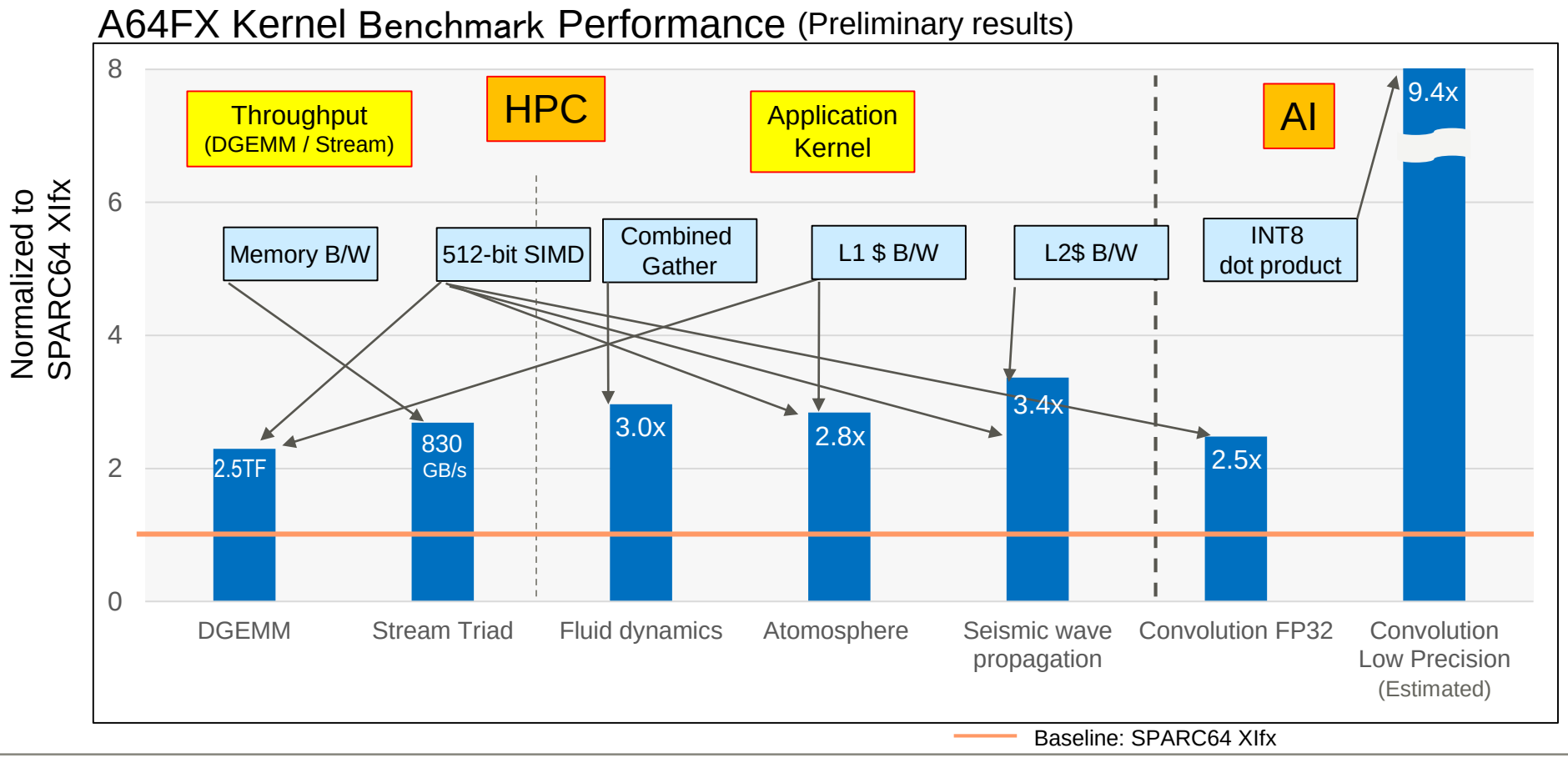
- "Power knob" limits units' activity via user APIs
- Performance/W would be optimized by utilizing Power knobs, Energy monitor & analyzer



A64FX Chip Level Application Performance

FUJITSU

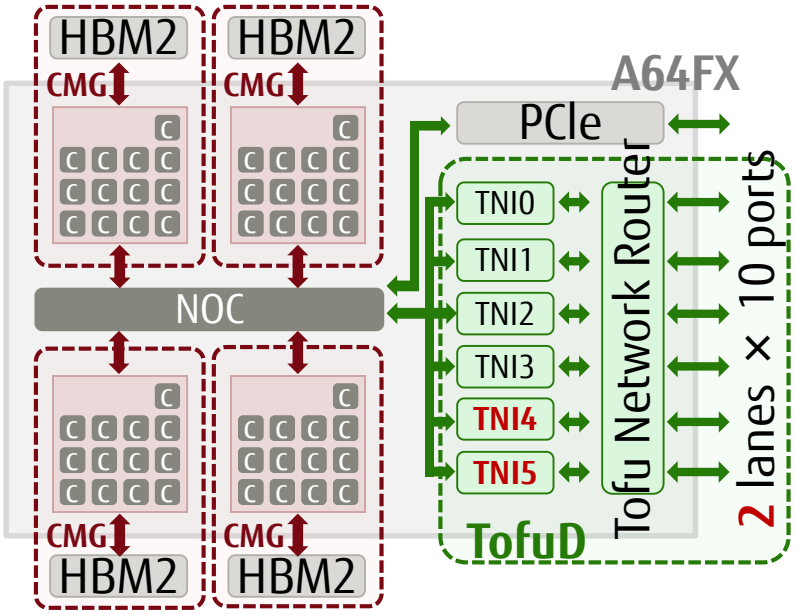
- Boosting application performance up by micro-architectural enhancements, 512-bit wide SIMD, HBM2 and semi-conductor process technologies
  - > 2.5x faster in HPC/AI benchmarks than that of SPARC64 XIfx tuned by Fujitsu compiler for A64FX micro-architecture and SVE



A64FX: Tofu interconnect D

- Integrated w/ rich resources
  - Increased TNIs achieves higher injection BW & flexible comm. patterns
  - Increased barrier resources allow flexible collective comm. algorithms
- Memory bypassing achieves low latency
  - Direct descriptor & cache injection

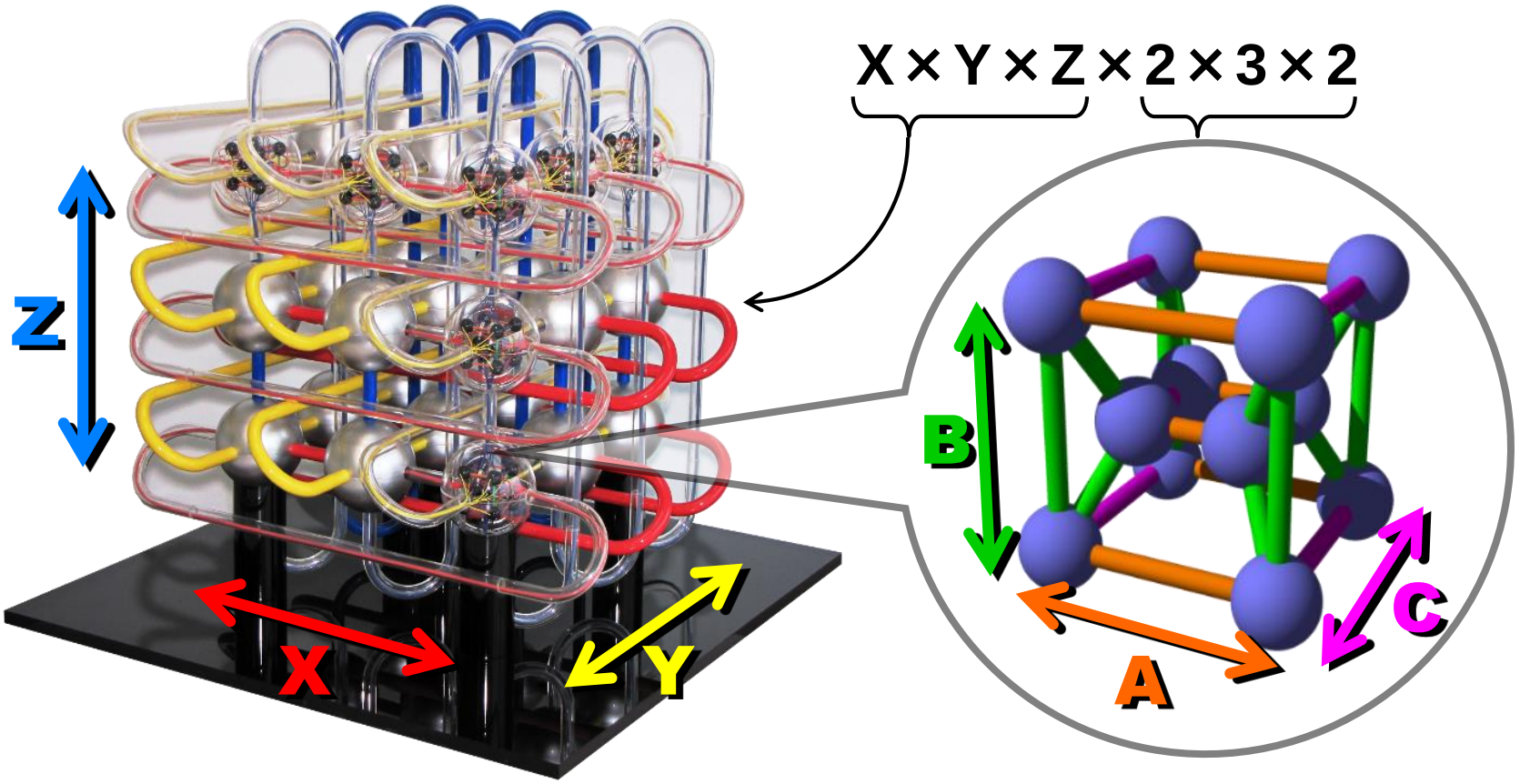
	TofuD spec
Data rate	28.05 Gbps
Link bandwidth	6.8 GB/s
Injection bandwidth	40.8 GB/s
	Measured
Put throughput	6.35 GB/s
PingPong latency	0.49~0.54 μs



TofuD: 6D Mesh/Torus Network

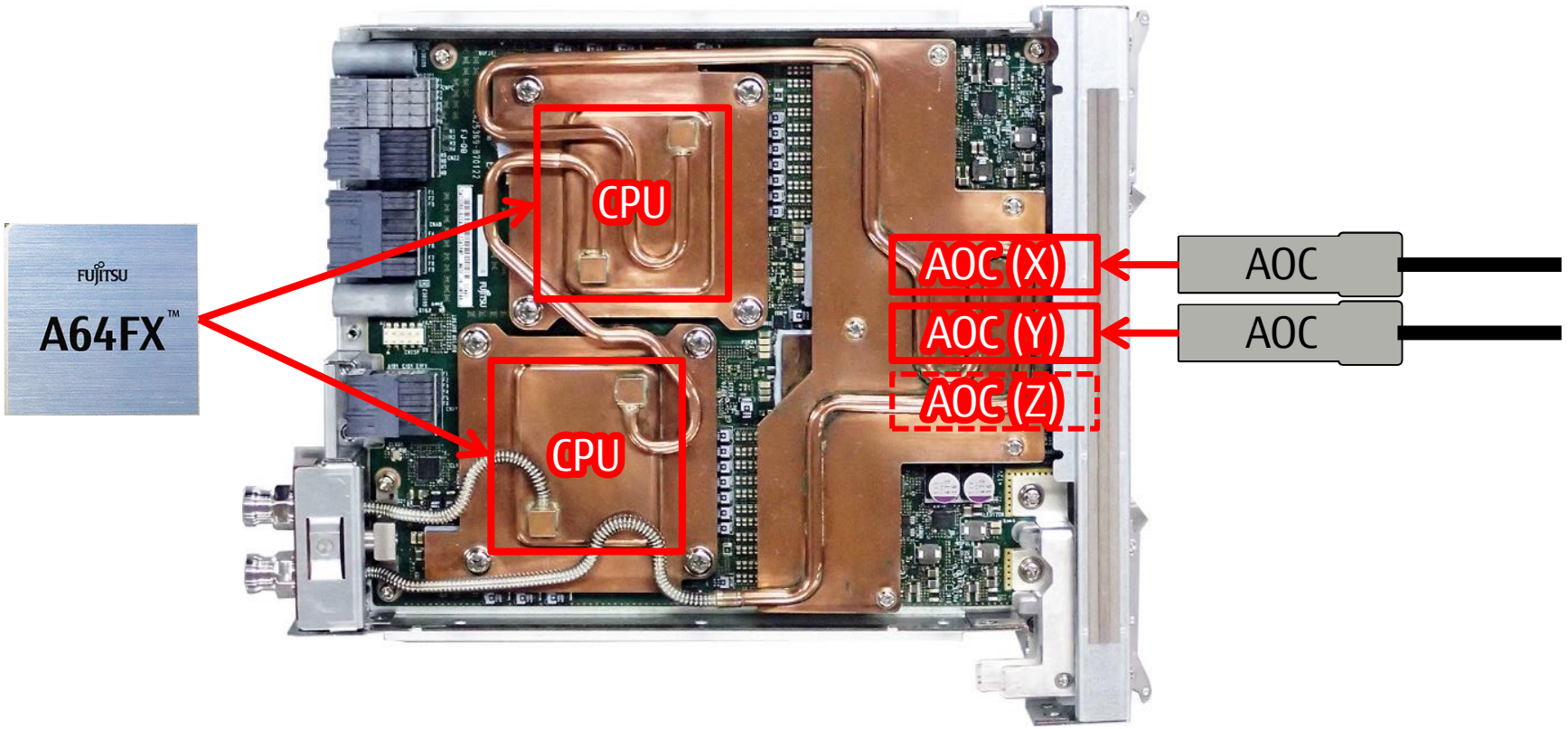
FUJITSU

- Six coordinates: (X, Y, Z) × (A, B, C)
  - X, Y and Z: sizes are depends on the system size
  - A, B and C: sizes are fixed to 2, 3, and 2 respectively
- Tofu stands for “torus fusion”



TofuD: Packaging – CPU Memory Unit

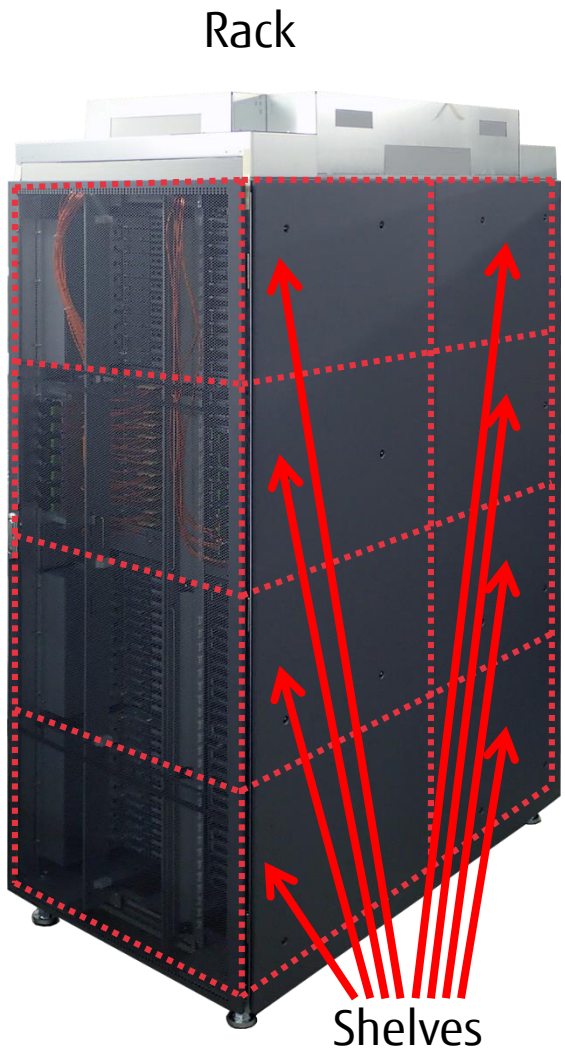
- Two CPUs connected with C-axis
  - X × Y × Z × A × B × C = 1 × 1 × 1 × 1 × 1 × 2
- Two or three active optical cable cages on board
  - Each cable is shared by two CPUs



TofuD: Packaging – Rack Structure

FUJITSU

- Rack
  - 8 shelves
  - 192 CMUs or 384 CPUs
- Shelf
  - 24 CMUs or 48 CPUs
  - X × Y × Z × A × B × C = 1 × 1 × 4 × 2 × 3 × 2
- Top or bottom half of rack
  - 4 shelves
  - X × Y × Z × A × B × C = 2 × 2 × 4 × 2 × 3 × 2



TofuD: Put Latencies & Throughput& Injection Rate from Clustrer 2018

FUJITSU

- TofuD: Evaluated by hardware emulators using the production RTL codes
  - Simulation model: System-level included multiple nodes

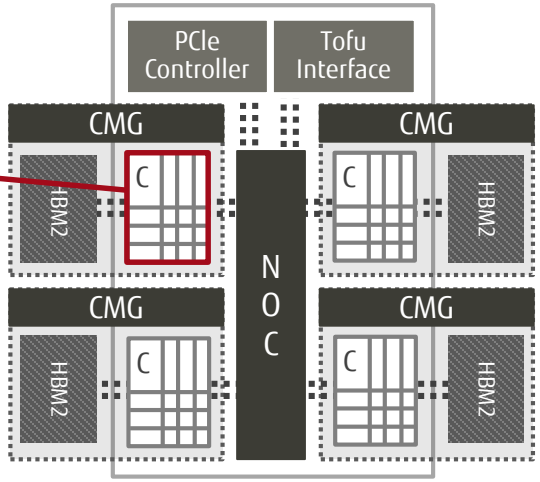
	Communication settings	Latency
Tofu	Descriptor on main memory	1.15 μs
	Direct Descriptor	0.91 μs
Tofu2	Cache injection OFF	0.87 μs
	Cache injection ON	0.71 μs
TofuD	To/From far CMGs	0.54 μs
	To/From near CMGs	0.49 μs

	Put throughput	Injection rate
Tofu	4.76 GB/s (95%)	15.0 GB/s (77%)
Tofu2	11.46 GB/s (92%)	45.8 GB/s (92%)
TofuD	6.35 GB/s (93%)	38.1 GB/s (93%)

A64FX: Summary

- Arm SVE, high performance and efficiency
  - DP performance >2.7 TFLOPS, >90%@DGEMM
  - Memory BW 1024 GB/s, >80%@STREAM Triad

CMG
12x compute cores
1x assistant core



CMG : Core Memory Group NOC : Network on Chip

	A64FX
ISA (Base, extension)	Armv8.2-A, SVE
Process technology	7 nm
Peak DP performance	>2.7 TFLOPS
SIMD width	512-bit
# of compute cores	48
Memory capacity	32 GiB (HBM2 x4)
Memory peak bandwidth	1024 GB/s
PCIe	Gen3 16 lanes
High speed interconnect	TofuD integrated

FUJITSU

Arm HPC Software Topics: Activities with Linaro and OSS Community

- With Arm and Linaro
- With OSS Community: Open MPI and Lustre

Activities with Arm and Linaro

FUJITSU

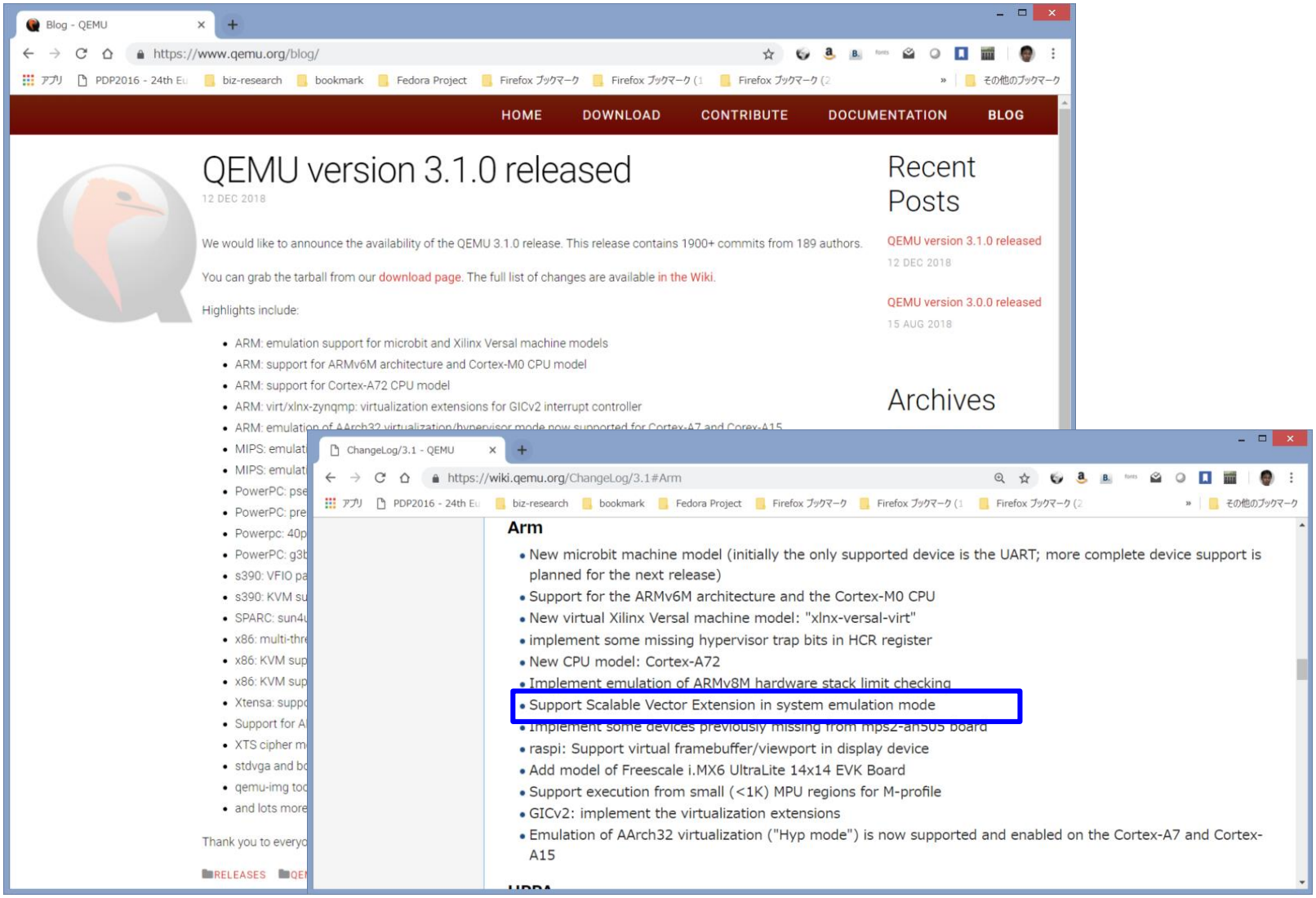
- LLVM SVE upstreaming and OSS porting with Arm
  - Variable Vector Length Support for LLVM Community in cooperation with Arm
- OpenHPC with Linaro:
  - Mr. Okamoto(Fujitsu) has been selected a 2018-2019 TSC(Technical Steering Committee) member
- Development Status with Linaro
  - LLVM/Clang for aarch64 Improvement: now ongoing
    - Register allocation, Software pipelining support, Vectorization/SIMDization
    - Pushing SVE support to the LLVM community in cooperation with Arm, Variable Vector Length Support is critical issue to introduce to LLVM tree.
  - QEMU/SVE Development: for building SVE software development
    - V3.1.0 released: https://www.qemu.org/2018/12/12/qemu-3-1-0/

QEMU/SVE Development with Linaro

https://www.qemu.org/2018/12/12/qemu-3-1-0/

FUJITSU

- Finally, Version 3.1.0 supports SVE in system emulation mode!



Post-K Software Stack

FUJITSU

- Post-K system supports SBSA/SBBR
  - Keeping binary compatibility with the other Aarch64 based systems.

Post-K Applications

FUJITSU Technical Computing Suite / RIKEN Advanced System Software

Management Software	Hierarchical File I/O Software	Programming Environment
System management for highly available & power saving operation	Application-oriented file I/O middleware	XcalableMP
Job management for higher system utilization & power efficiency	Lustre-based distributed file system FEFS	MPI (Open MPI, MPICH)
		OpenMP, COARRAY, Math Libs.
		Compilers (C, C++, Fortran)
		Debugging and tuning tools

arm Linux OS / McKernel (Lightweight Kernel)

Post-K System Hardware

Post-K
Under Development
w/ RIKEN

Open MPI: from SC18 BoF Slides

https://www.open-mpi.org/papers/sc-2018

FUJITSU

Open MPI: from SC18 BoF Slides

https://www.open-mpi.org/papers/sc-2018

FUJITSU

MPI for the Post-K Computer

- Post-K MPI based on Open MPI
  - Support A64FX (Armv8.2-A+SVE) and TofuD
  - Plan to use Open MPI 4.0 and PMIx 2.1
- Contribution to Open MPI from post-K MPI
  - Persistent collectives [*see next page*]
  - Datatype for half-precision floating point [*early 2019*]
  - Thread parallelization of pack/unpack [*early 2019*]

Half-precision(FP16)datatype development started in cooperation with ANL and Mellanox

FUJITSU

MPI-4.0 or MPI-3.2 ?

Persistent Collectives in MPI-4.0

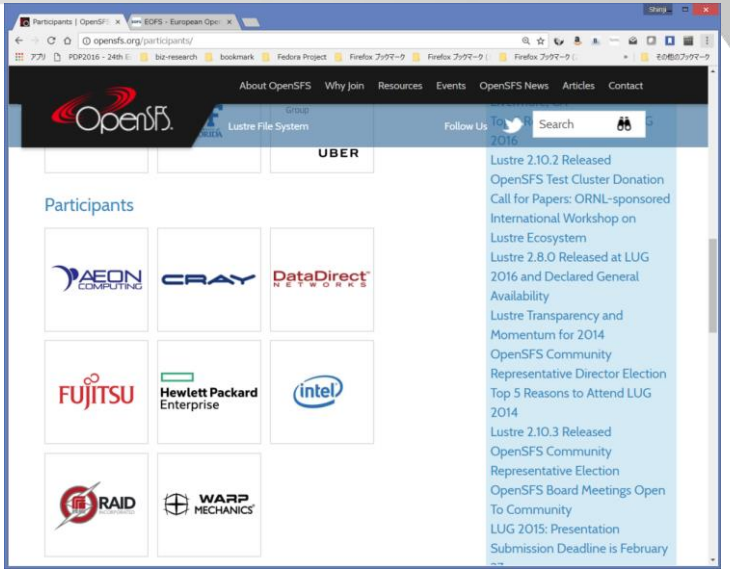
- Persistent collectives are in Open MPI 4.0.x
- Overlap computation & communication and reduce communication initialization cost
- Use MPIX_ prefix because standardization is not complete

```
MPIX_Bcast_init(
  buf, count, ..., &req);
for (...) {
  MPI_Start(&req);
  // ... your computation
  MPI_Wait(&req, &stat);
}
MPI_Request_free(&req);
```

Lustre Community: OpenSFS and EOFS

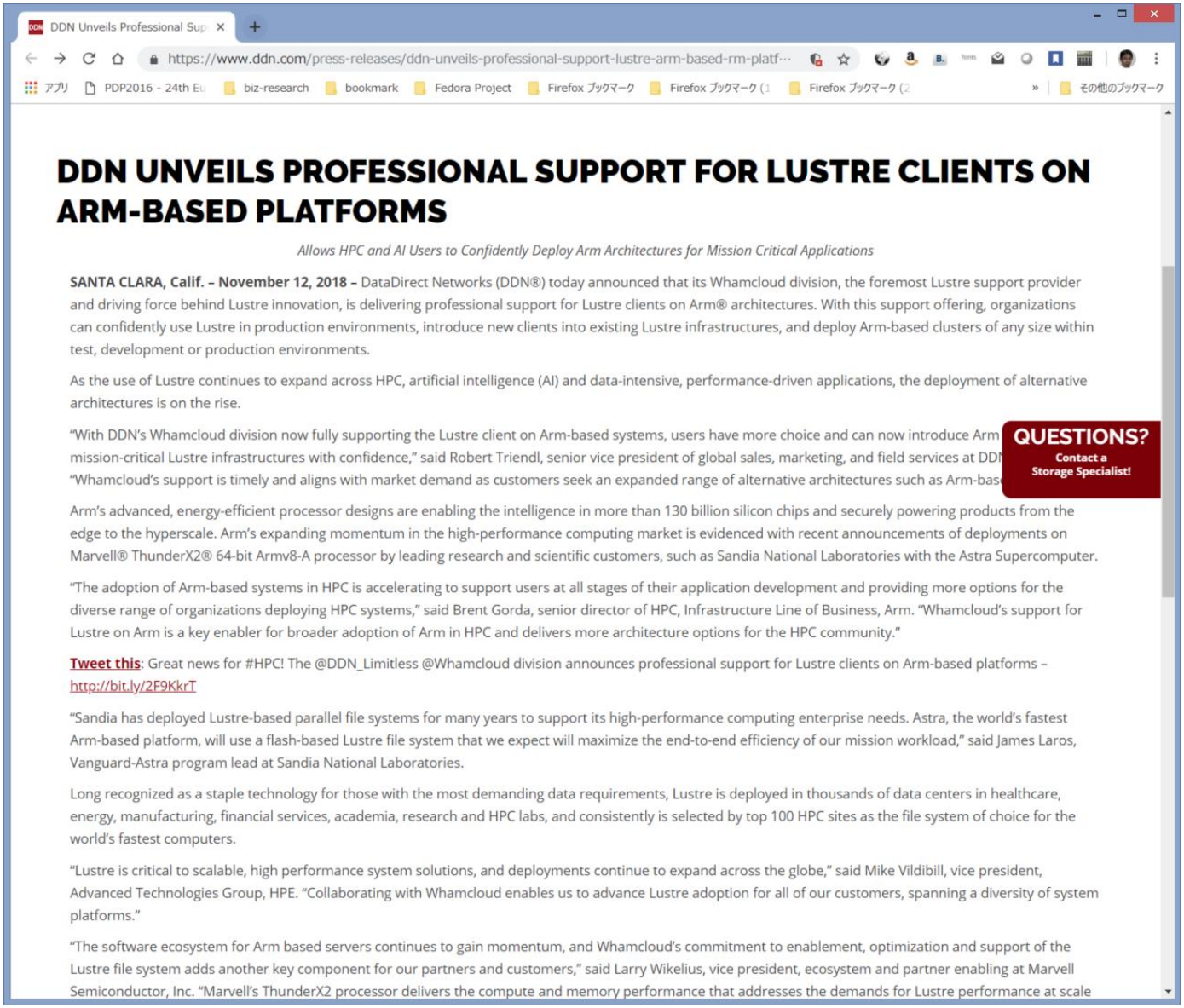
- OpenSFS: US Based Non-profit Organization
  - President: Sarp Oral (ORNL)
- EOFS: EU Based Non-profit Organization
  - President: Frank Baetke (HPE)
- Lustre for Arm
  - Fujitsu is member of OpenSFS and will support Lustre based products.
- Two Major Events
  - Lustre User Group(LUG)
    - LUG 19@Houston, 2019/5/15-17 http://opensfs.org/events/
  - Lustre Admins and Devs workshop(LAD)
    - LAD 18@Paris, 2018/9/24-25 https://www.eofs.eu/events/lad18

Slides Archives are on each site



2018/11: Whamcloud has started Lustre client support on Arm based platforms

https://www.ddn.com/press-releases/ddn-unveils-professional-support-lustre-arm-based-rm-platforms/



DDN Unveils Professional Sup

https://www.ddn.com/press-releases/ddn-unveils-professional-support-lustre-arm-based-rm-platf···

アプリ | PDP2016 - 24th E | biz-research | bookmark | Fedora Project | Firefox ブックマーク | Firefox ブックマーク (1 | Firefox ブックマーク (2 | その他のブックマーク

DDN UNVEILS PROFESSIONAL SUPPORT FOR LUSTRE CLIENTS ON ARM-BASED PLATFORMS

Allows HPC and AI Users to Confidently Deploy Arm Architectures for Mission Critical Applications

SANTA CLARA, Calif. – November 12, 2018 – DataDirect Networks (DDN®) today announced that its Whamcloud division, the foremost Lustre support provider and driving force behind Lustre innovation, is delivering professional support for Lustre clients on Arm® architectures. With this support offering, organizations can confidently use Lustre in production environments, introduce new clients into existing Lustre infrastructures, and deploy Arm-based clusters of any size within test, development or production environments.

As the use of Lustre continues to expand across HPC, artificial intelligence (AI) and data-intensive, performance-driven applications, the deployment of alternative architectures is on the rise.

"With DDN's Whamcloud division now fully supporting the Lustre client on Arm-based systems, users have more choice and can now introduce Arm mission-critical Lustre infrastructures with confidence," said Robert Triendl, senior vice president of global sales, marketing, and field services at DDN "Whamcloud's support is timely and aligns with market demand as customers seek an expanded range of alternative architectures such as Arm-bas

QUESTIONS?
Contact a Storage Specialist!

Arm's advanced, energy-efficient processor designs are enabling the intelligence in more than 130 billion silicon chips and securely powering products from the edge to the hyperscale. Arm's expanding momentum in the high-performance computing market is evidenced with recent announcements of deployments on Marvell® ThunderX2® 64-bit Armv8-A processor by leading research and scientific customers, such as Sandia National Laboratories with the Astra Supercomputer.

"The adoption of Arm-based systems in HPC is accelerating to support users at all stages of their application development and providing more options for the diverse range of organizations deploying HPC systems," said Brent Gorda, senior director of HPC, Infrastructure Line of Business, Arm. "Whamcloud's support for Lustre on Arm is a key enabler for broader adoption of Arm in HPC and delivers more architecture options for the HPC community."

Tweet this: Great news for #HPC! The @DDN_Limitless @Whamcloud division announces professional support for Lustre clients on Arm-based platforms – http://bit.ly/2F9KkrT

"Sandia has deployed Lustre-based parallel file systems for many years to support its high-performance computing enterprise needs. Astra, the world's fastest Arm-based platform, will use a flash-based Lustre file system that we expect will maximize the end-to-end efficiency of our mission workload," said James Laros, Vanguard-Astra program lead at Sandia National Laboratories.

Long recognized as a staple technology for those with the most demanding data requirements, Lustre is deployed in thousands of data centers in healthcare, energy, manufacturing, financial services, academia, research and HPC labs, and consistently is selected by top 100 HPC sites as the file system of choice for the world's fastest computers.

"Lustre is critical to scalable, high performance system solutions, and deployments continue to expand across the globe," said Mike Vildibill, vice president, Advanced Technologies Group, HPE. "Collaborating with Whamcloud enables us to advance Lustre adoption for all of our customers, spanning a diversity of system platforms."

"The software ecosystem for Arm based servers continues to gain momentum, and Whamcloud's commitment to enablement, optimization and support of the Lustre file system adds another key component for our partners and customers," said Larry Wikelius, vice president, ecosystem and partner enabling at Marvell Semiconductor, Inc. "Marvell's ThunderX2 processor delivers the compute and memory performance that addresses the demands for Lustre performance at scale

FUJITSU

Arm HPC Software Topics: OSS Application Porting Updates

OSS apps porting at Arm HPC Users Group

FUJITSU

(http://arm-hpc.gitlab.io/)

- Twelve primary OSS applications are listed and being tested in the Users Group for each compilers, collaboratively w/ Arm

Application	Lang.	GCC	LLVM	Arm	Fujitsu
LAMMPS	C++	Modified	Modified	Modified	Modified
GROMACS	C	Modified	Modified	Modified	Modified
GAMESS*	Fortran	Modified	Modified	Modified	Modified
OpenFOAM	C++	Modified	Modified	Modified	Modified
NAMD	C++	Modified	Modified	Modified	Modified
WRF	Fortran	Modified	Modified	Modified	Modified
Quantum ESPRESSO	Fortran	Ok in as is	Ok in as is	Ok in as is	Modified
NWChem	Fortran	Ok in as is	Modified	Modified	ongoing
ABINIT	Fortran	Modified	Modified	Modified	Modified
CP2K	Fortran	Ok in as is	Issues found	Issues found	ongoing
NEST*	C++	Ok in as is	Modified	Modified	Modified
BLAST*	C++	Ok in as is	Modified	Modified	Modified

* Registered by Fujitsu

Issue of CP2K (known issue)

FUJITSU

- flang rejects valid empty constructor
  https://github.com/flang-compiler/flang/issues/239 (Closed)
  https://github.com/flang-compiler/flang/issues/615 (New Ticket, Segfault)

dbcsr_data_types.F90

```
module dbcsr_data_types
  TYPE dbcsr_mempool_type
  END TYPE dbcsr_mempool_type

  TYPE dbcsr_memtype_type
    LOGICAL                         :: mpi = .FALSE.
    TYPE(dbcsr_mempool_type), POINTER :: pool => Null()
  END TYPE dbcsr_memtype_type
end module
```

dbcsr_data_types_user.F90

```
module a
  use dbcsr_data_types, only: dbcsr_memtype_type

  type foo
    type(dbcsr_memtype_type) :: val = dbcsr_memtype_type()
  end type
end module
```

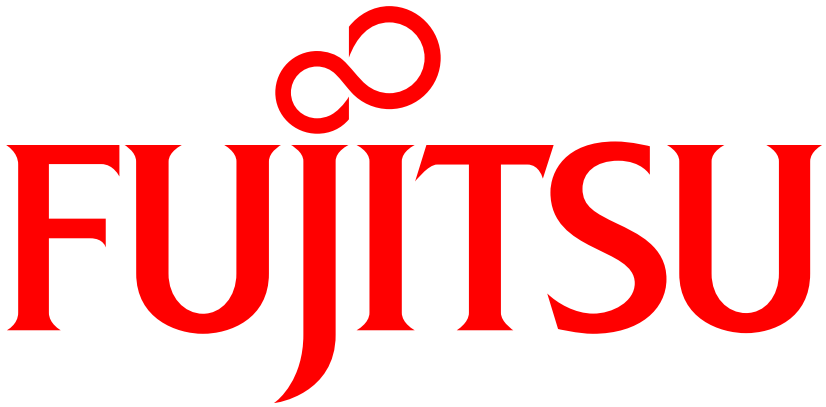
```
[eco@cn-r05-01 work]$ gfortran -c dbcsr_data_types.F90 && gfortran -c dbcsr_data_types_user.F90
[eco@cn-r05-01 work]$
[eco@cn-r05-01 work]$ flang -c dbcsr_data_types.F90 && flang -c dbcsr_data_types_user.F90
F90-F-0155-Empty structure constructor() - type dbcsr_memtype_type (dbcsr_data_types_user.F90: 5)
F90/x86-64 Linux Flang - 1.5 2017-05-01: compilation aborted
[eco@cn-r05-01 work]$
```

Summary

FUJITSU

- A64FX: High Performance Arm CPU
  - > 2.5 TFLOPS singe-chip degemm performance
  - Arm is already not only mobile CPU but also high-end HPCs

- Arm HPC Ecosystem Development
  - Arm HPC Software Topics
    - Activities with Arm, Linaro and OSS Community
    - Porting and Evaluation of HPC Application
  - Will need continuous efforts



shaping tomorrow with you