FUJITSU Supercomputer PRIMEHPC FX1000 and FX700

Toshiyuki Shimizu

Nov. 21st, 2019

FUJITSU LIMITED

Supercomputer "Fugaku", Formerly Known as Post-K

FUJITSU

- High app performance using co-design approach
- The best GF/W, Green500 #1 (prototype)
- The first Armv8 + SVE CPU, A64FX

Focus	Approach
Application performance	Co-design w/ application developers and Fujitsu-designed CPU core w/ high memory bandwidth utilizing HBM2
Power efficiency	Leading-edge Si-technology, Fujitsu's proven low power & high performance logic design, and power-controlling knobs
Usability	Arm®v8-A ISA with Scalable Vector Extension ("SVE"), and Arm standard Linux

FUJITSU Supercomputer PRIMEHPC FX1000 and FX700

- New PRIMEHPC utilizes "Fugaku" technologies

Fugaku「富岳」

Fujitsu PRIMEHPC FX1000

Fujitsu PRIMEHPC FX700

8-CPU 2U rack mountable air cool

384-CPU / rack, water cool, highest performance

384-CPU / rack, water cool, highest performance

Concept and Specifications

	FX700	FX1000	Fugaku「富岳」
Model	FX700	FX1000	Fugaku「富岳」
Concept	Cooperation with standard technologies	Application performance, energy efficiency, and scalability	
CPU	A64FX x8	A64FX x384 / rack	
Interconnect	InfiniBand EDR	TofuD	
Cooling	Air	Water	
Dimension	2U rack mountable	Custom: 800 mm x 1,400 mm x 2,000 mm	
Software stack	OpenHPC, BCM	Fujitsu Technical Computing Suite	

A64FX CPU Summary

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

12x compute cores
1x assistant core

PCIe Controller   Tofu Interface

CMG   CMG
C   C
HBM2   HBM2
N O C
CMG   CMG
C   C
HBM2   HBM2

CMG : Core Memory Group   NOC : Network on Chip
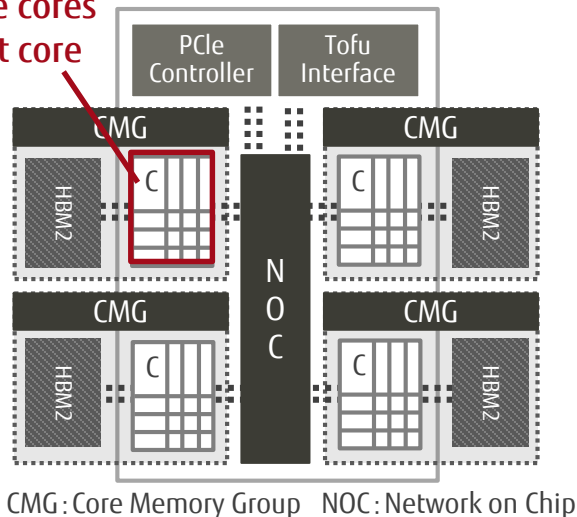
	A64FX
ISA (Base, extension)	Armv8.2-A, SVE
Process technology	7 nm
Peak DP performance	2.7+ TFLOPS
SIMD width	512-bit
# of cores	48 + 4
Memory capacity	32 GiB (HBM2 x4)
Memory peak bandwidth	1024 GB/s
PCIe	Gen3 16 lanes
High speed interconnect	TofuD integrated

A64FX Leading-edge Si-technology

- **TSMC 7nm FinFET & CoWoS**
  - Broadcom SerDes, HBM I/O, and SRAMs
  - 8.786 billion transistors
  - 594 signal pins
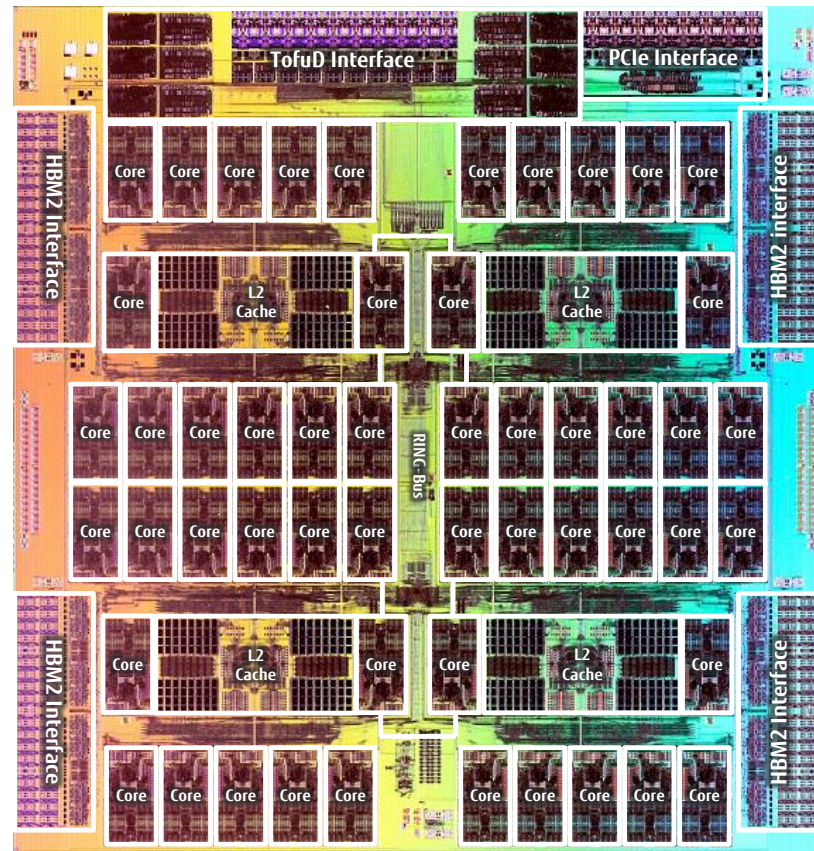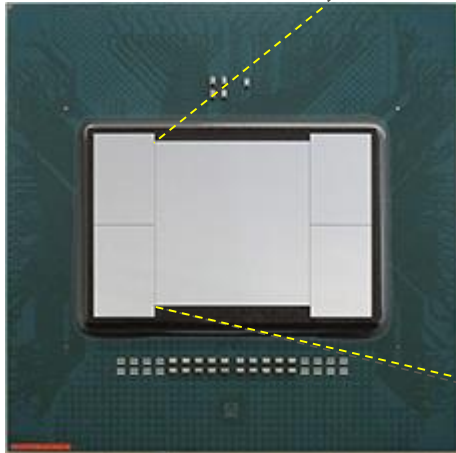
A64FX Tofu Interconnect D

- 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.
Port bandwidth	6.8 GB/s
Injection bandwidth	40.8 GB/s
	Measured
Put throughput	6.35 GB/s
Ping-pong latency	0.49~0.54 μs
One hop latency	Approx. 80 ns

Fugaku System Software
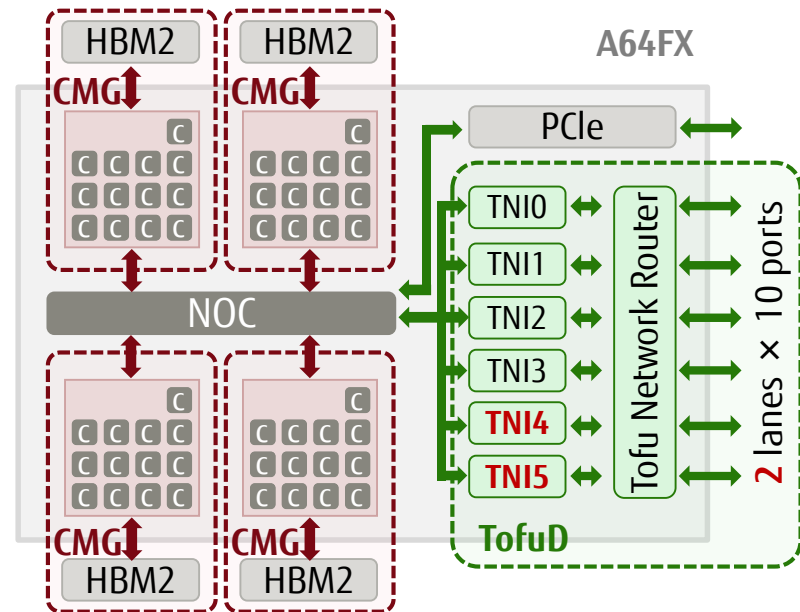
FUJITSU

■ RIKEN and Fujitsu are developing a software stack for Fugaku

Fugaku applications

Fujitsu Technical Computing Suite / RIKEN-developed 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 for Fugaku	OpenMP, COARRAY, Math.libs.
		Compilers (C, C++, Fortran)
		Debugging and tuning tools

arm

Linux OS / McKernel (Lightweight kernel)

Fugaku system hardware

Fugaku — Under development w/ RIKEN

FX1000 System Software

■ **RIKEN developed system software** could be used as Open Source Software

Applications
Fujitsu Technical Computing Suite / RIKEN-developed system software

Management software	File system	Programming environment
System management for high availability & power saving operation	FEFS Lustre-based distributed file system	XcalableMP
Job management for higher system utilization & power efficiency		MPI (Open MPI, MPICH)
		OpenMP, COARRAY, Math.libs.
		Compilers (C, C++, Fortran)
		Debugging and tuning tools

Linux OS / McKernel (Lightweight kernel)

FX1000 hardware

PRIMEHPC FX700

FUJITSU

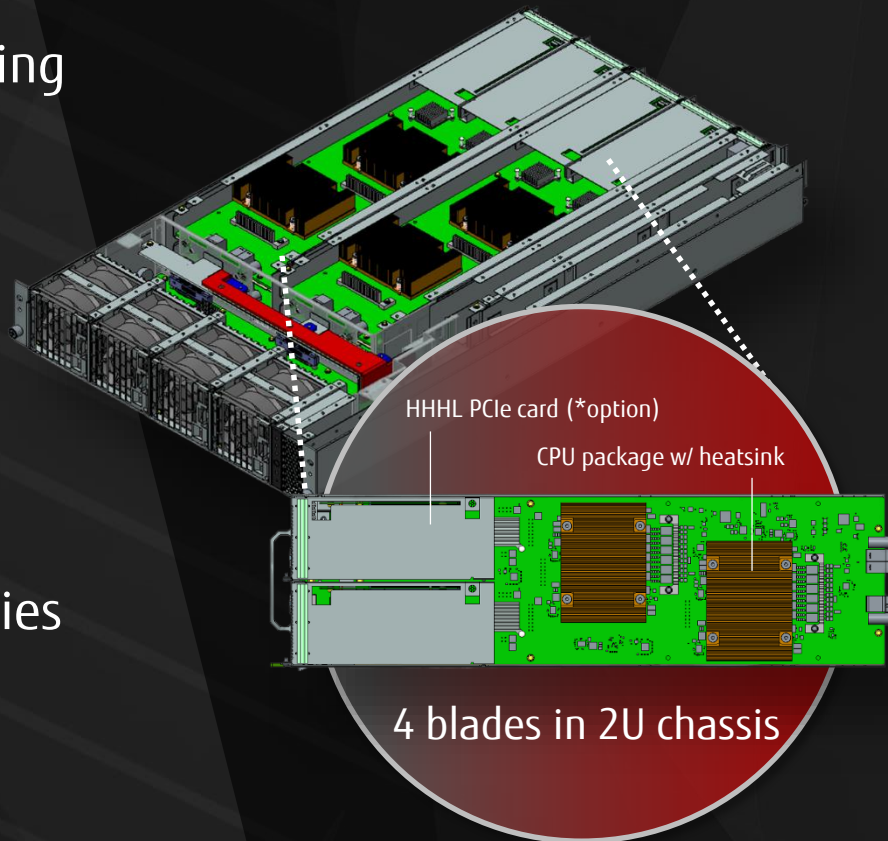- **High-performance Arm server featuring the A64FX CPU**
  - Same CPU as Fugaku and FX1000

- **Easy deployment and flexible configuration**
  - Air-cooled, 2U rack-mountable chassis
  - From 2 to 8 CPUs per chassis
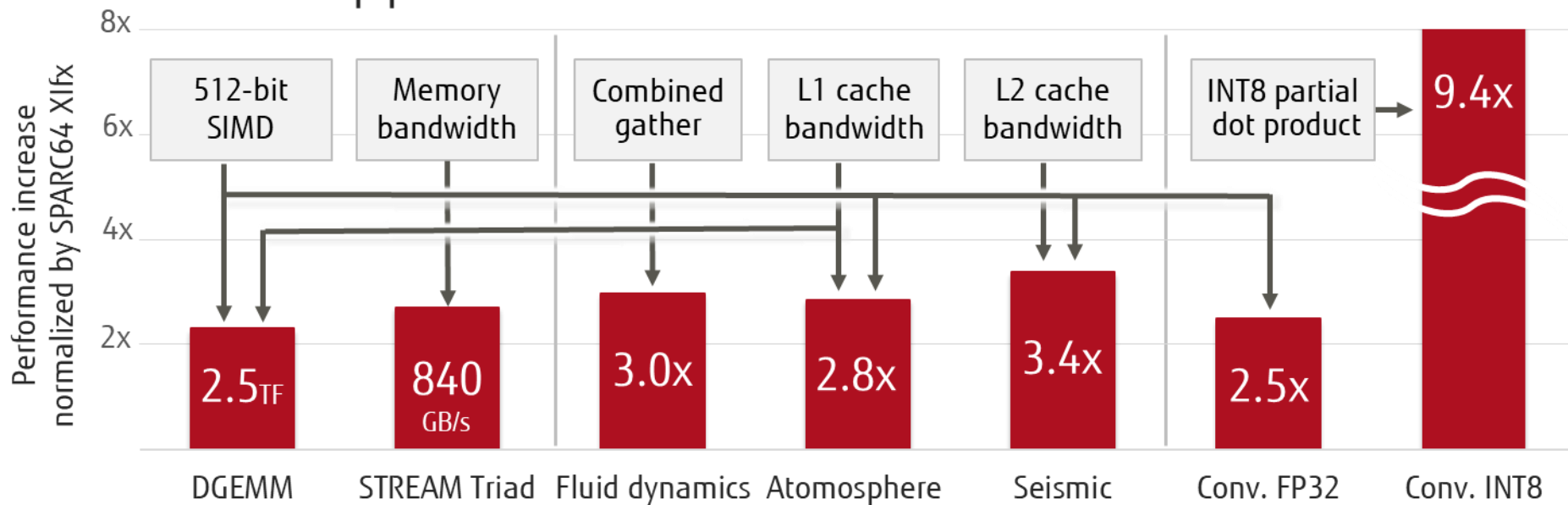
- **Utilize open and standard technologies**
  - InfiniBand EDR
  - RHEL 8, OpenHPC, Bright Cluster Manager, etc.

HHHL PCIe card (*option)

CPU package w/ heatsink

4 blades in 2U chassis

A64FX CPU Performance Evaluation (1/2)

■ Over 2.5x faster in HPC & AI benchmarks than SPARC64 XIfx

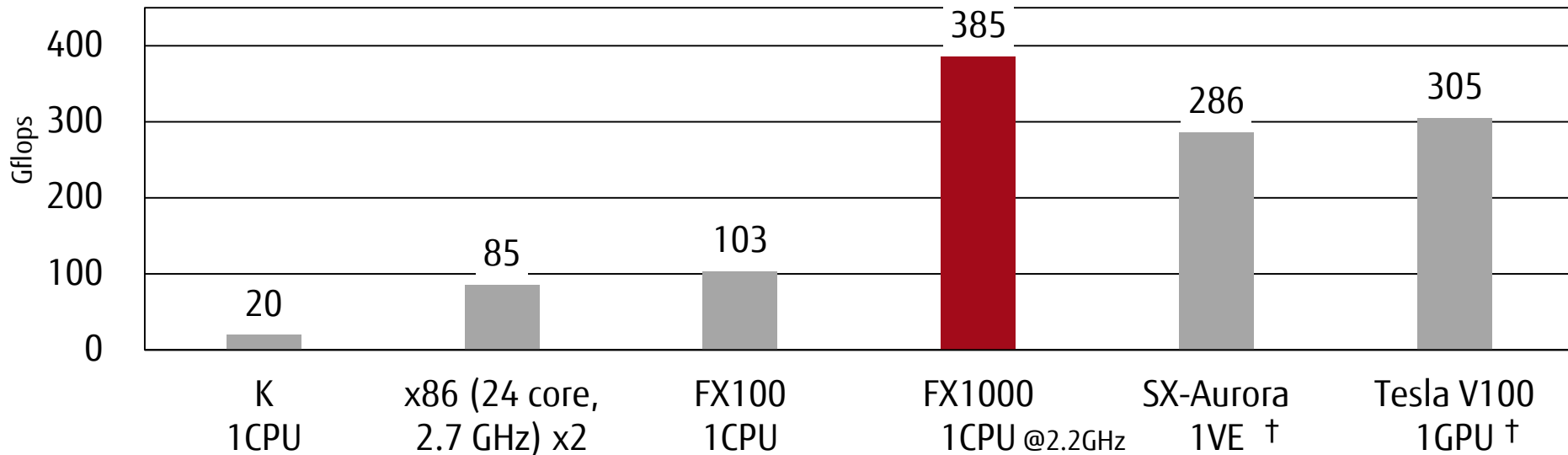A64FX chip performance measurements & architectural contributions

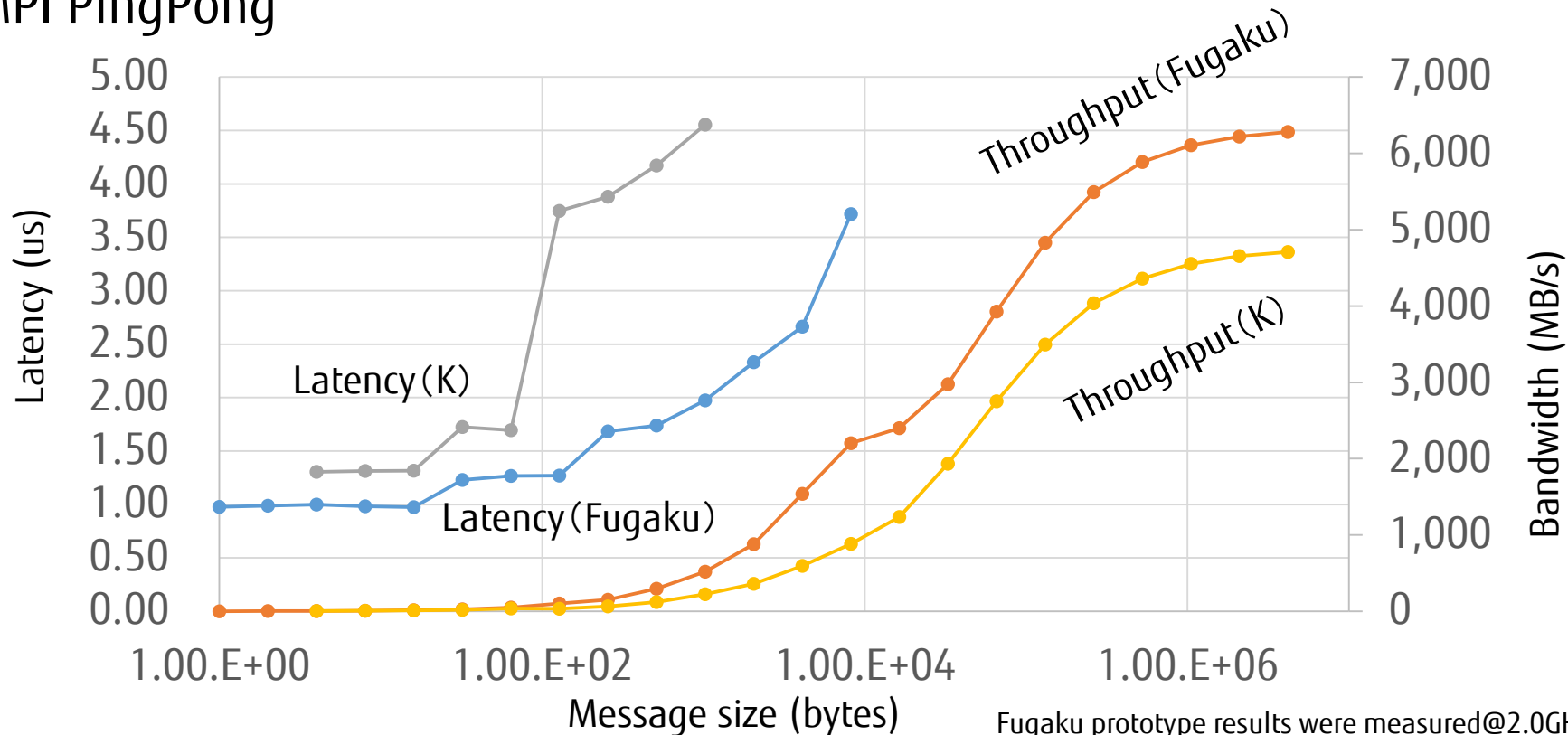# A64FX CPU Performance Evaluation (2/2)

■ Himeno Benchmark

■ Stencil calculation to solve Poisson's equation by Jacobi method

† "Performance evaluation of a vector supercomputer SX-aurora TSUBASA",
SC18,   https://dl.acm.org/citation.cfm?id=3291728

TofuD: MPI_Send/Receive Latency and BW

■ MPI PingPong

TofuD: MPI_Bcast Performance

- **New 3D bidirectional pipeline algorithm using 6 TNIs (Tofu Network IF)**
  - Sends data to six, $\pm X$, $\pm Y$, $\pm Z$, directions simultaneously
  - Reduces the constant overhead by up to "X+Y+Z-3" long pipelines

2D bidirectional pipeline w/ 4-dir
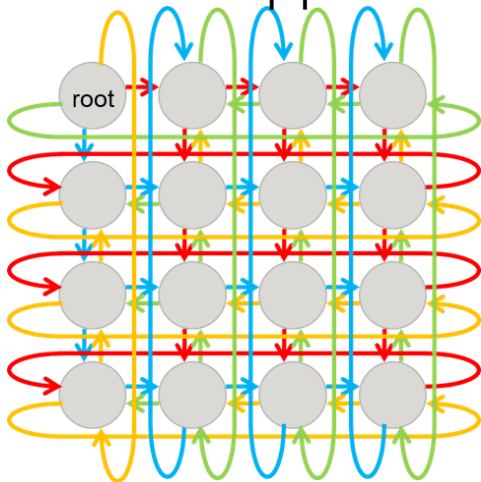
Allgather also utilizes the 6-dir simultaneous transfer

MPI_Bcast on 384 nodes

- New 3D on Fugaku, 6 TNIs
- 3D on Fugaku, 3 TNIs
- 3D on K, 3 TNIs
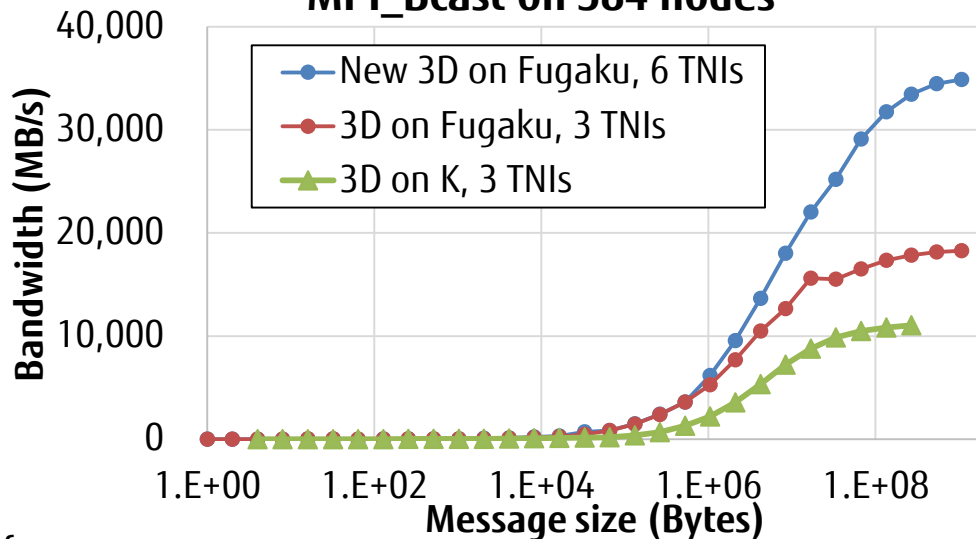
Fugaku prototype results were measured@2.0GHz

TofuD: MPI_Allreduce Performance

FUJITSU

■ # of items supported in hardware reduction, Tofu barrier interface, TBI

	K	Fugaku	Usage & func @ Fugaku
FP	1	3	Complex + Double
INT	1	6	3x w/ MAXLOC

■ MPI library implementation

■ Choose the best algorithm for VL

• TBI, Recursive doubling, Reduce & bcast

■ Multiple TBI calls (K calls TBI once)

MPI_Allreduce on 384 nodes

K w/o TBI

Fugaku w/o TBI

Fugaku w/ TBI

K w/ TBI

MPI_Allreduce (us)

Vector length

K: Double w/ TBI
INT32 w/o TBI
Double w/o TBI
INT32 w/ TBI
Double w/ TBI

Fugaku prototype results were measured@2.0GHz
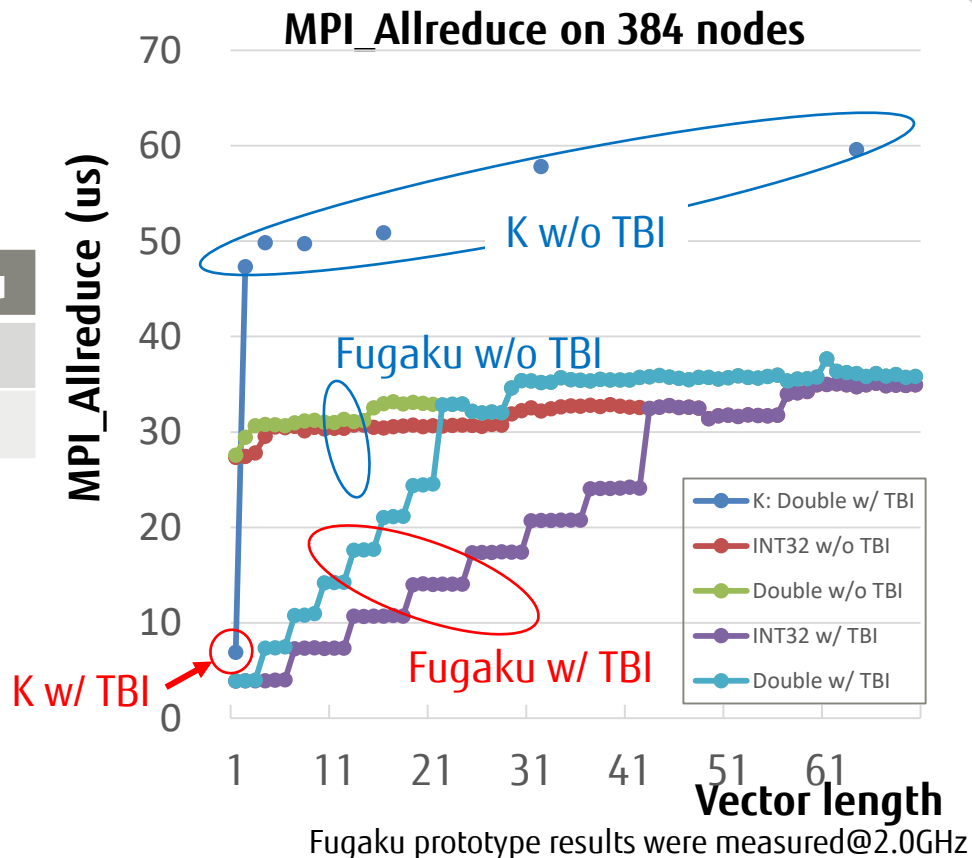
Optimization for Extreme Scalability & Efficiency

- **OS jitter reduction**
  - Utilizes Tickless mode of Linux
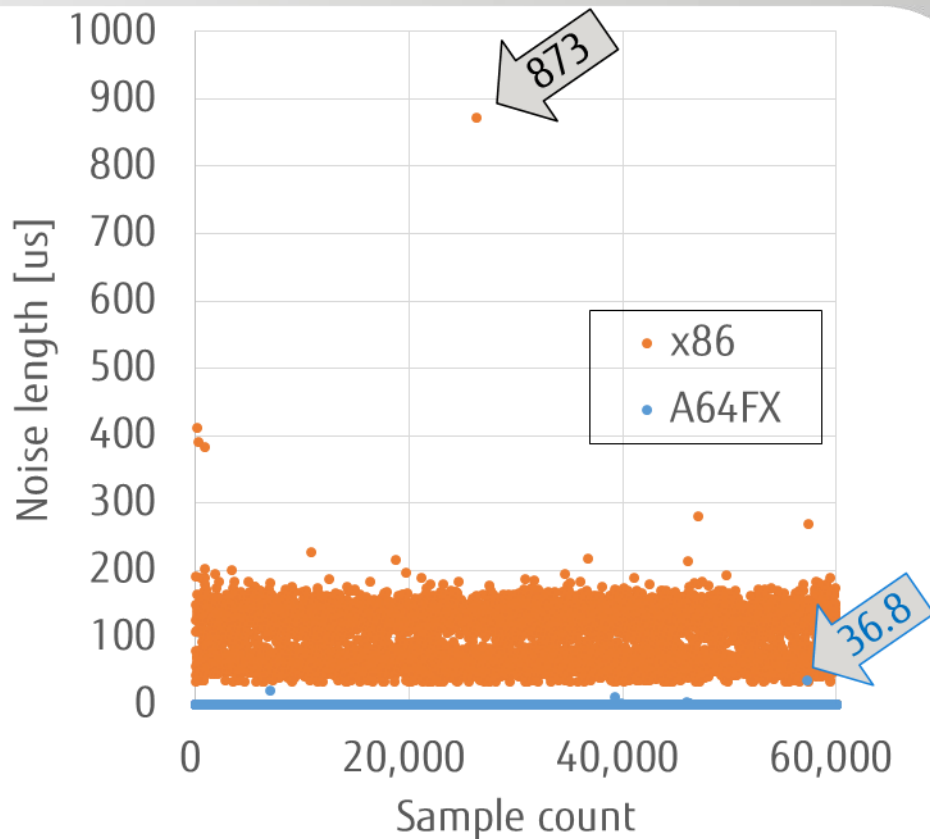 - Stops unneeded daemons & services
  - Offloads OS interrupts & daemons to the A64FX assistant cores
- **OS jitter measurement by FWQ**

	Mean noise ratio	Longest noise
Fugaku	7.1E-07	37 µs
K	6.6E-5	85 µs
x86	3.7E-03	873 µs

Fugaku prototype results were measured@2.0GHz

Green500, Nov. 2019

FUJITSU

A64FX prototype –

Fujitsu A64FX 48C 2GHz

ranked **#1** on the list

768x general purpose A64FX CPU w/o accelerators

- 1.9995 PFLOPS @ HPL, 84.75%
- 16.876 GF/W
- Power quality level 2

The GREEN 500

HOME   GREEN500   LISTS ▾   RESOURCES ▾   ABOUT ▾   MEDIA KIT

Home / Lists / November 2019

NOVEMBER 2019

- The most energy-efficient system and No. 1 on the Green500 is a new Fujitsu A64FX prototype installed at Fujitsu, Japan. It achieved 16.9 GFlops/Watt power-efficiency during its 2.0 Pflop/s Linpack performance run. It is listed on position 160 in the TOP500.
- In second position is the NA-1 system, a PEZY Computing / Exascaler Inc. system which is currently being readied at PEZY Computing, Japan for a future installation at NA Simulation in Japan. It achieve 16.3 GFlops/Watt power efficiency. It is on position 421 in the TOP500.
- The No 3 on the Green500 is AiMOS, a new IBM Power systems at the Rensselaer Polytechnic Institute Center for Computational Innovations (CCI), New York, USA. It achieved 15.8 GFlops/Watt and is listed at position 25 in the TOP500.

Green500 List for November 2019

Listed below are the November 2019 The Green500's energy-efficient supercomputers ranked from 1 to 10.

Note: Shaded entries in the table below mean the power data is derived and not meassured.

Rank	TOP500 Rank	System	Cores	Rmax [TFlop/s]	Power [kW]	Power Efficiency (GFlops/watts)
1	159	**A64FX prototype** - Fujitsu A64FX, Fujitsu A64FX 48C 2GHz, Tofu interconnect D , Fujitsu / Fujitsu Numazu Plant / Japan	36,864	1,999.5	118	16.876
2	420	**NA-1** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz , PEZY Computing / Exascaler Inc. / PEZY Computing K.K. / Japan	1,271,040	1,303.2	80	16.256
3	24	**AiMOS** - IBM Power System AC922, IBM POWER9 20C 3.45GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Volta GV100 , IBM / Rensselaer Polytechnic Institute Center for Computational Innovations (CCI) / United States	130,000	8,045.0	510	15.771
4	373	**Satori** - IBM Power System AC922, IBM POWER9 20C 2.4GHz, Infiniband EDR, NVIDIA Tesla V100 SXM2 , IBM / MIT/MGHPCC Holyoke, MA / United States	23,040	1,464.0	94	15.574
5	1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM / DOE/SC/Oak Ridge National Laboratory / United States	2,414,592	148,600.0	10,096	14.719

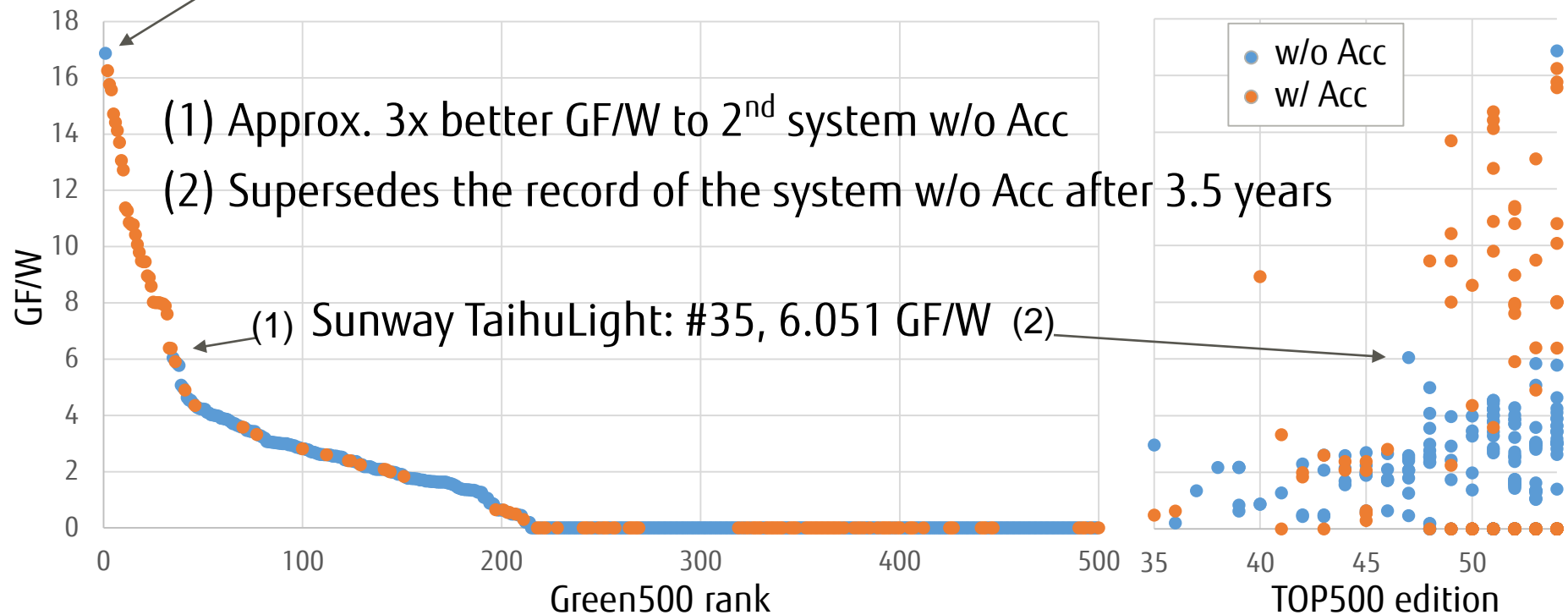https://www.top500.org/green500/lists/2019/11/
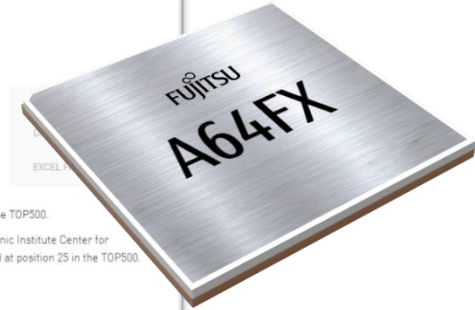
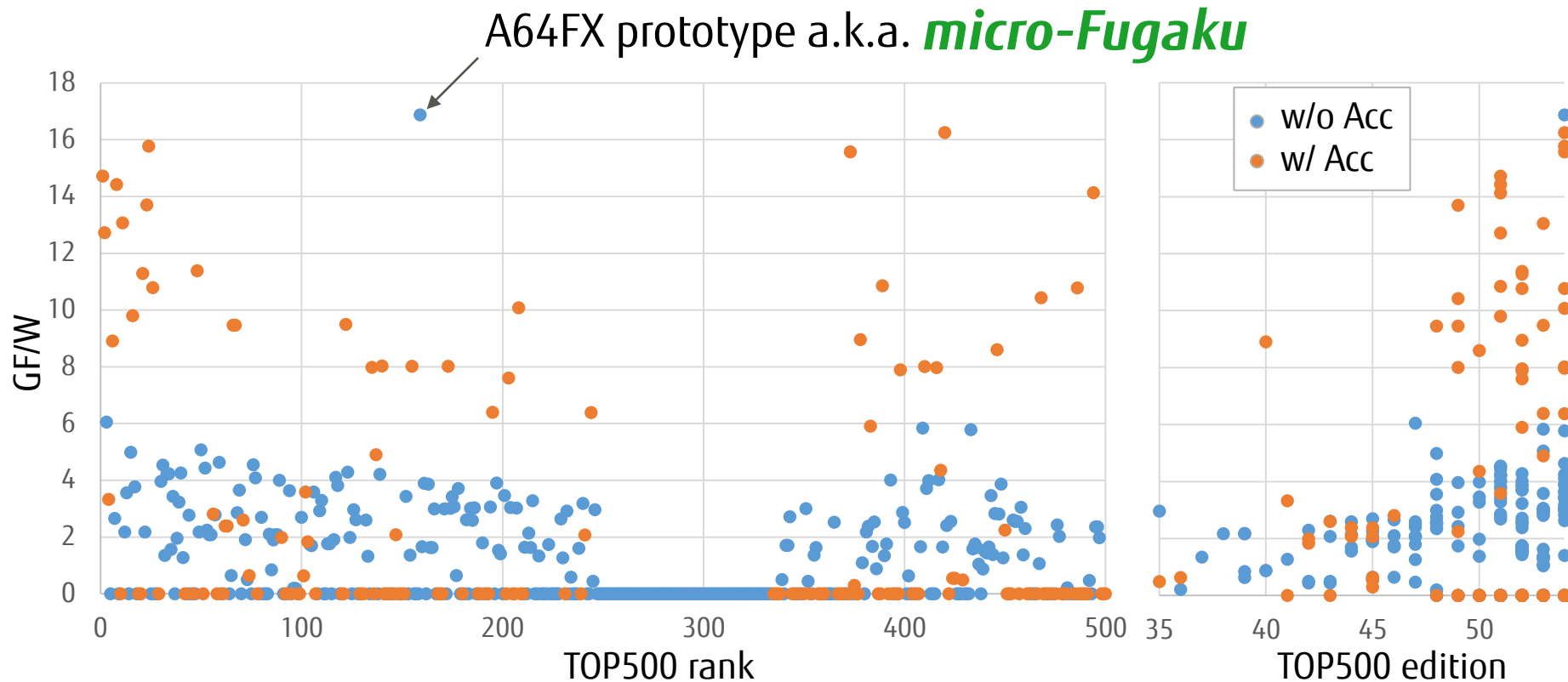# SC19 Green500 ranking and 1st appeared TOP500 edition

FUJITSU

A64FX prototype: #1, 16.876 GF/W

(1) Approx. 3x better GF/W to 2nd system w/o Acc

(2) Supersedes the record of the system w/o Acc after 3.5 years

(1) Sunway TaihuLight: #35, 6.051 GF/W (2)

SC19 TOP500 ranking and GF/W

A64FX prototype a.k.a. *micro-Fugaku*
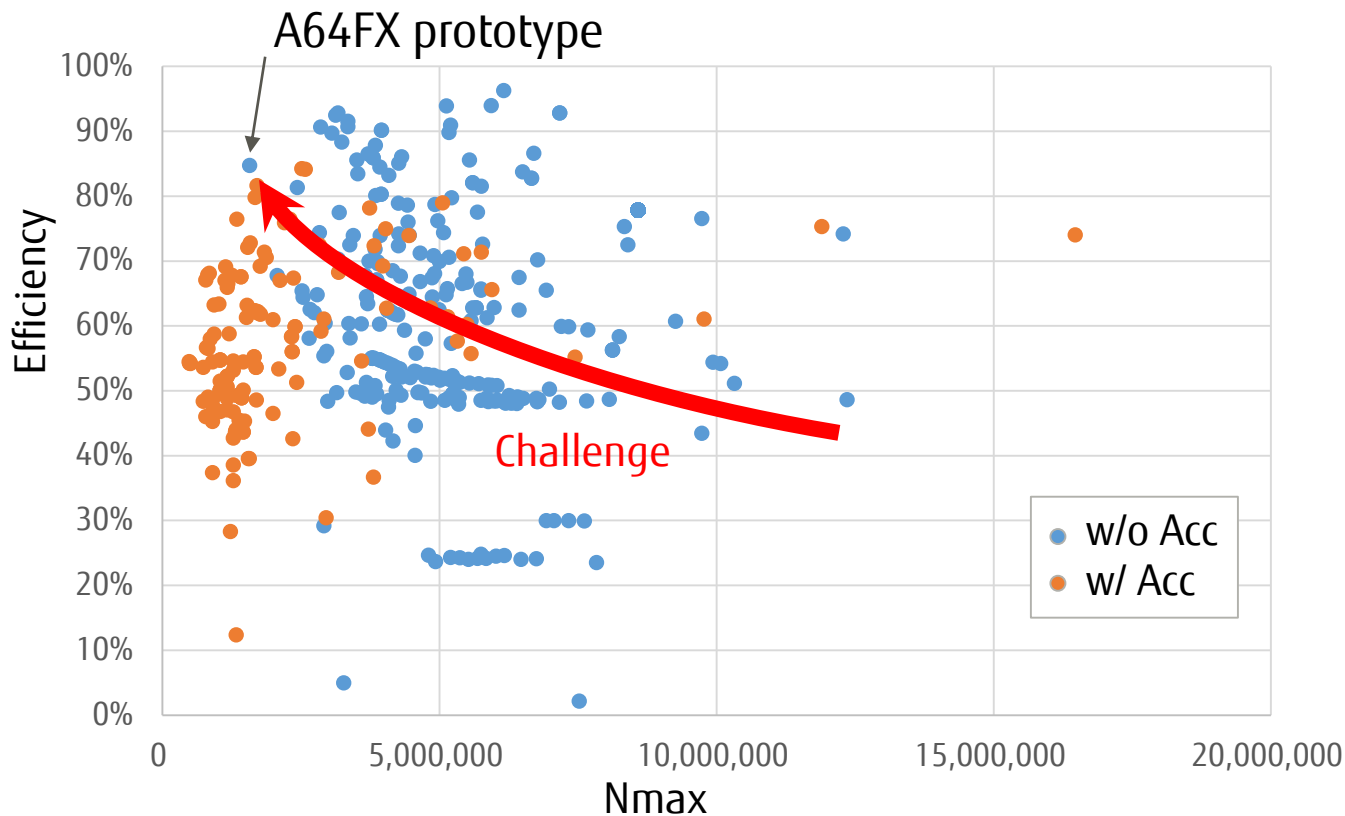
How Fujitsu worked

- **Key for GF/W is {energy efficient HW} x {parallel/exec efficiency}**
- **A64FX is designed for energy efficient**
  - Fujitsu's proven CPU microarchitecture & 7nm FinFET
  - SoC design: Tofu interconnect D integrated
  - CoWoS: 4x HBM2 for main memory integrated
- **Superior parallel/exec efficiency**
  - Math. libraries are tuned for application efficiency
  - Comm. libs are also tuned utilizing long experience of Tofu @ K computer
  - Performance tuning is efficiently done utilizing rich performance analyzer/monitor

SC19 TOP500 calculation efficiency

- ■ Superior efficiency 84.75% with small Nmax

- ■ Results of:
  - ■ Optimized communication and math. libs
  - ■ Optimization of overlapped communication

Fujitsu Numazu Plant: A64FX prototype system

Package air conditioner

CDU

Power distribution board

384-CPU rack   384-CPU rack

Shelf 2,3,6,7

A64FX prototype

Shelf 0,1,4,5

Power distribution board

Shelf 2,3,6,7

Shelf 0,1,4,5

Cooling water

Power meter for backup

Ethernet SW

Ethernet SW

Ethernet SW

Control LAN

Management LAN

High precision power meter

ntp server

Head nodes

Subsystem: 1.5kW
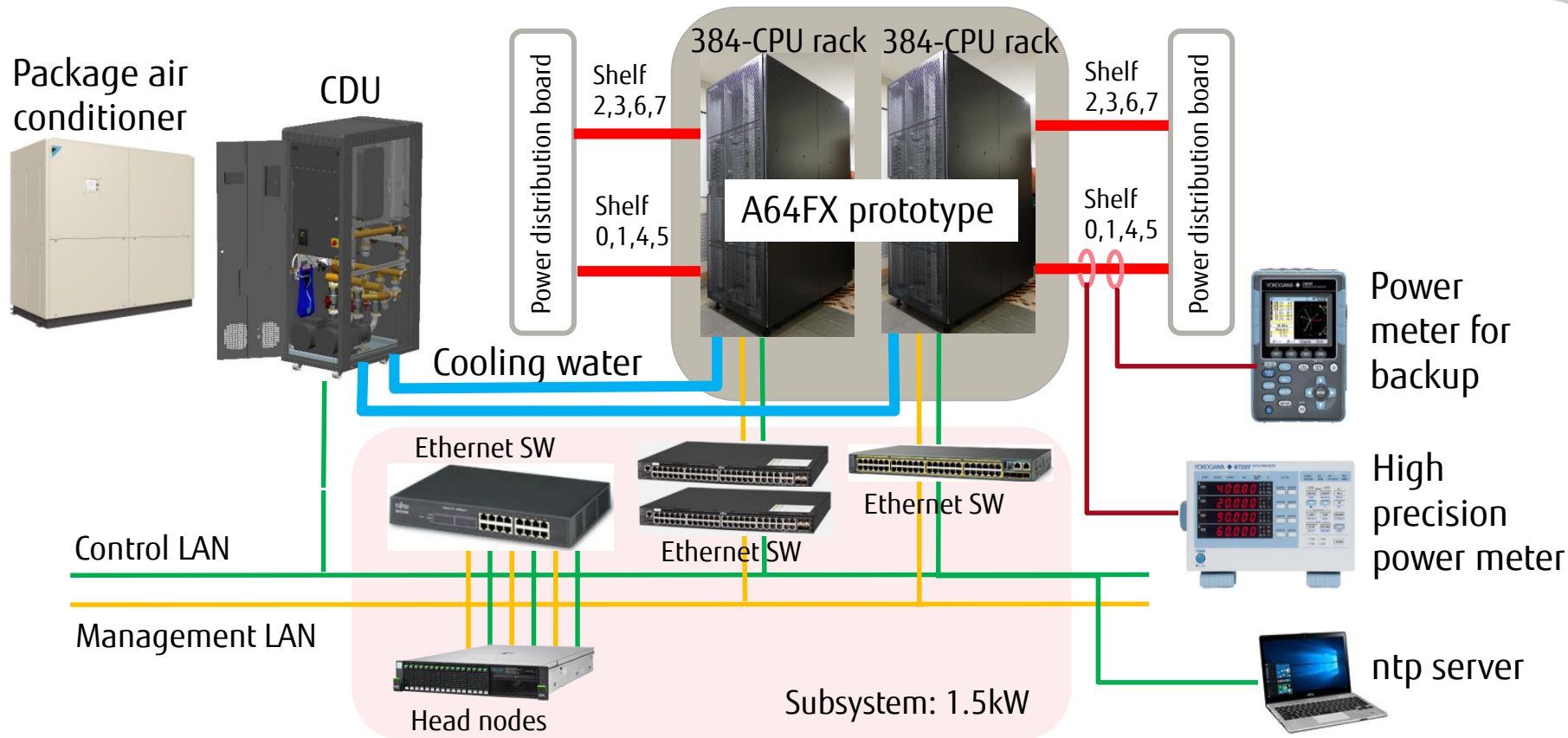
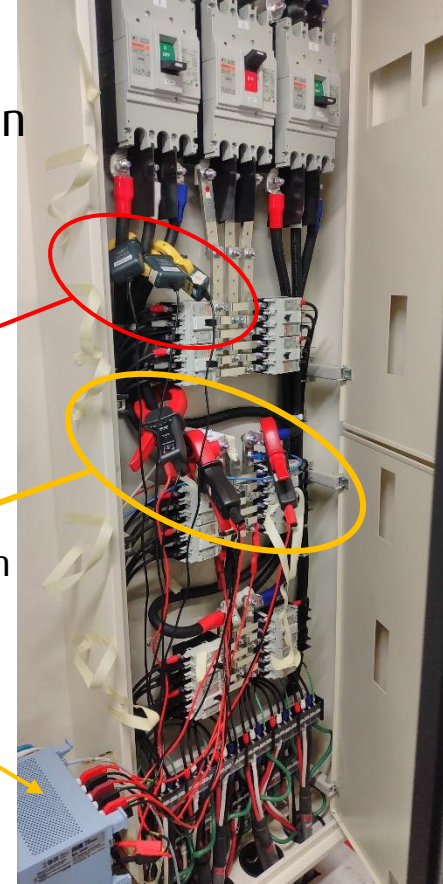# A64FX prototype @ Fujitsu Numazu Plant

2m

Power distribution board
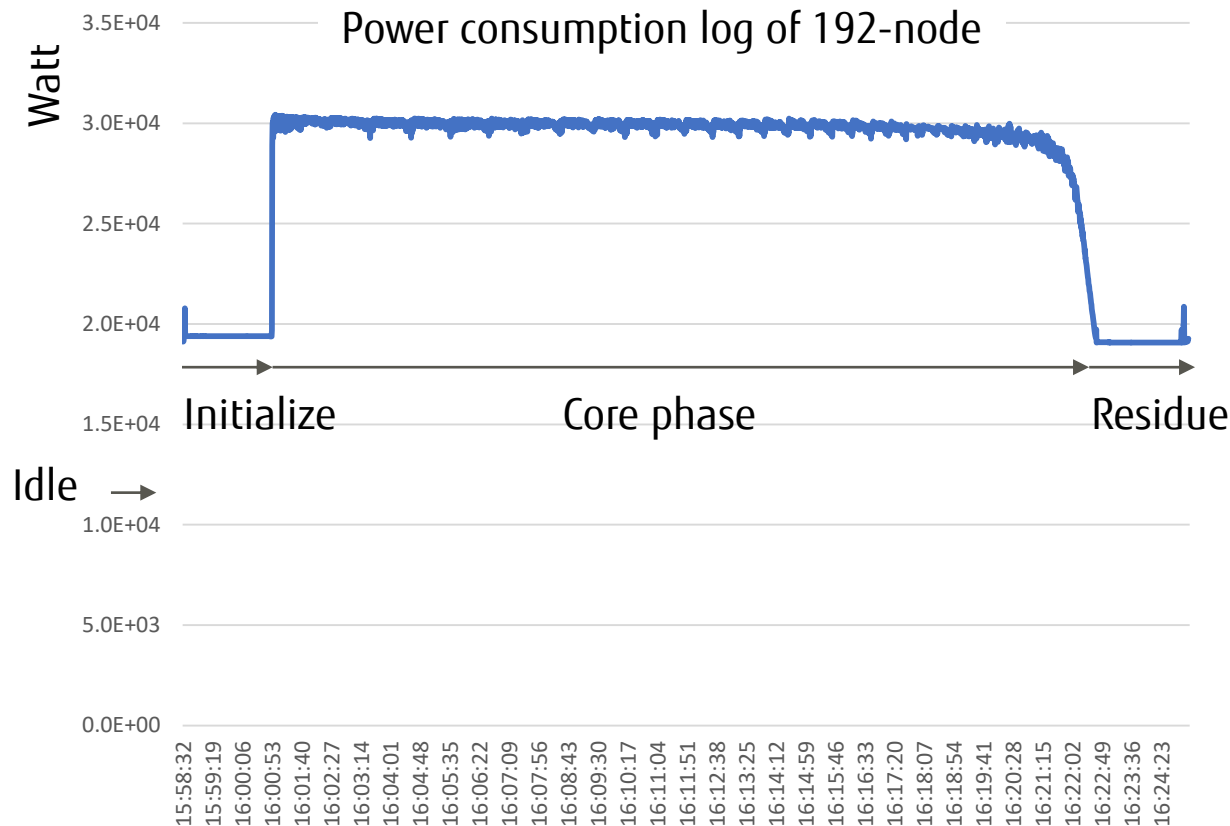
Clamps for backup

Clamps for high precision power meter

High precision power meter

Power consumption of ¼ A64FX prototype system

- Steady and high efficient computation are observed even from the point of power consumption

- Ave. of core phase:
  118.48 kW / system

- Idle power:
  46.92 kW / system

Power consumption log of 192-node

(Chart: Watt on the y-axis ranging from 0.0E+00 to 3.5E+04, with time along the x-axis from 15:58:32 to 16:24:23. The power stays near 2.0E+04 during Initialize, rises to about 3.0E+04 during Core phase, then drops back near 2.0E+04 at Residue. Phases labeled: Initialize, Core phase, Residue. Idle marker near 1.0E+04.)

Porting status @ Arm HPC Users Group

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

Application	Program	GCC	LLVM	Arm	Fujitsu
LAMMPS	C++	Modified(C)	Modified(C)	Modified(C)	Modified(C)
GROMACS	C	Modified(C)	Modified(C)	Modified(C)	Modified(C)
GAMESS	Fortran	Modified(C)	Modified(C)	Modified(C)	Modified(C)
OpenFOAM	C++	Modified(C)	Modified(C)	Modified(C)	Modified(C)
NAMD	C++	Modified(C)	Modified(C)	Modified(C)	Modified(C)
Siesta	Fortran	Yes	No	No	Modified(C)
WRF	Fortran	Modified(C)	Modified(C)	Modified(C)	Modified(C)
Quantum ESPRESSO	Fortran	Yes	Yes	Yes	Yes
NWChem	Fortran	Yes	Modified(C)	Modified(C)	Modified(S)
ABINIT	Fortran	Yes	Modified(S)	Modified(S)	Modified(S)
CP2K	Fortran	Yes	No	No	Modified(S)
NEST	C++	YES	Modified(C)	Modified(C)	Modified(C)
USQCD (MILC)	C	Yes	Modified(C)	Modified(C)	Modified(C)
BLAST	C++	Yes	Modified(C)	Modified(C)	Modified(C)
OpenMX	C	Yes	ongoing	ongoing	Yes

Registered by Fujitsu

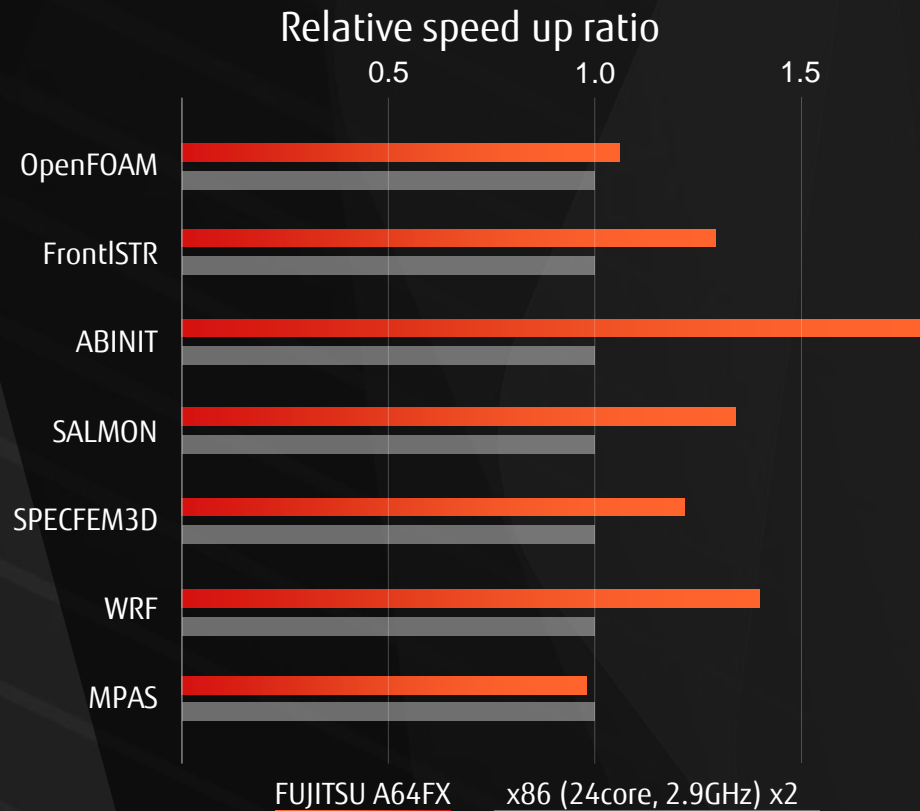Already registered

Not registered

Yes   : Ok in as is
Modified(C): Build env. modification
Modified(S): Src code modification
No   : Issues found
Ongoing   : Ongoing

A64FX CPU Performance Evaluation for Real Apps

Open source software, Real apps on an A64FX, 2.2GHz of PRIMEHPC FX1000

- Up to 1.8x faster over the latest x86 processor (24core, 2.9GHz) x2

- High memory B/W and long SIMD length of A64FX work effectively with these applications

Relative speed up ratio

| | 0.5 | 1.0 | 1.5 |

- OpenFOAM
- FrontISTR
- ABINIT
- SALMON
- SPECFEM3D
- WRF
- MPAS

FUJITSU A64FX   x86 (24core, 2.9GHz) x2

A64FX Superior Power Efficiency for Real Apps

Performance / Energy consumption on an A64FX, 2.2GHz of PRIMEHPC FX1000

- Up to 3.7x more efficient over the latest x86 processor (24core, 2.9GHz) x2

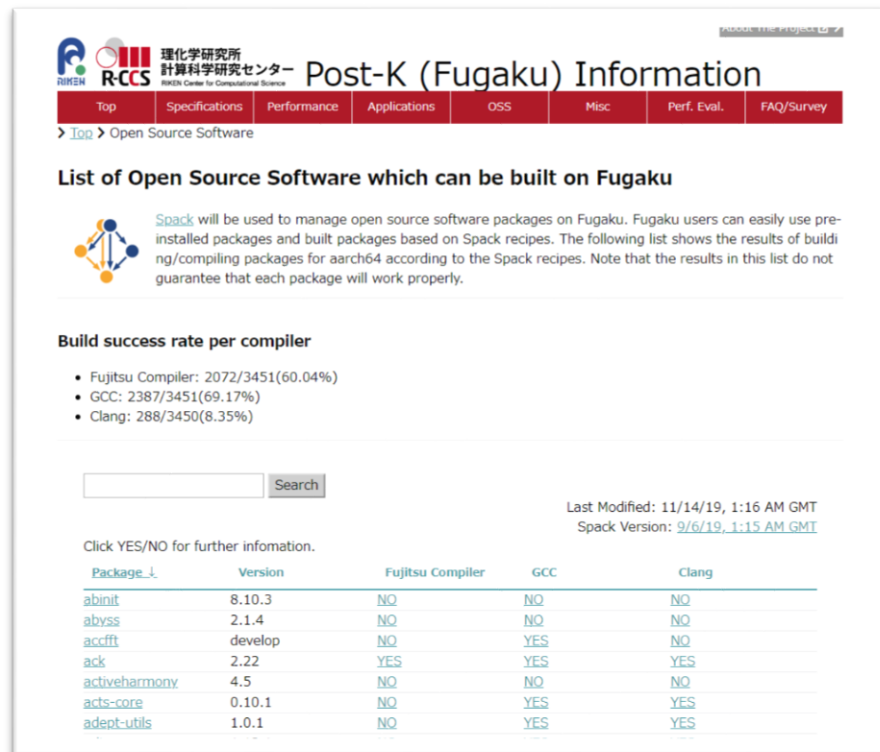- High efficiency is achieved by energy-conscious design and implementation

Relative power efficiency ratio

FUJITSU A64FX   x86 (24core, 2.9GHz) x2

26
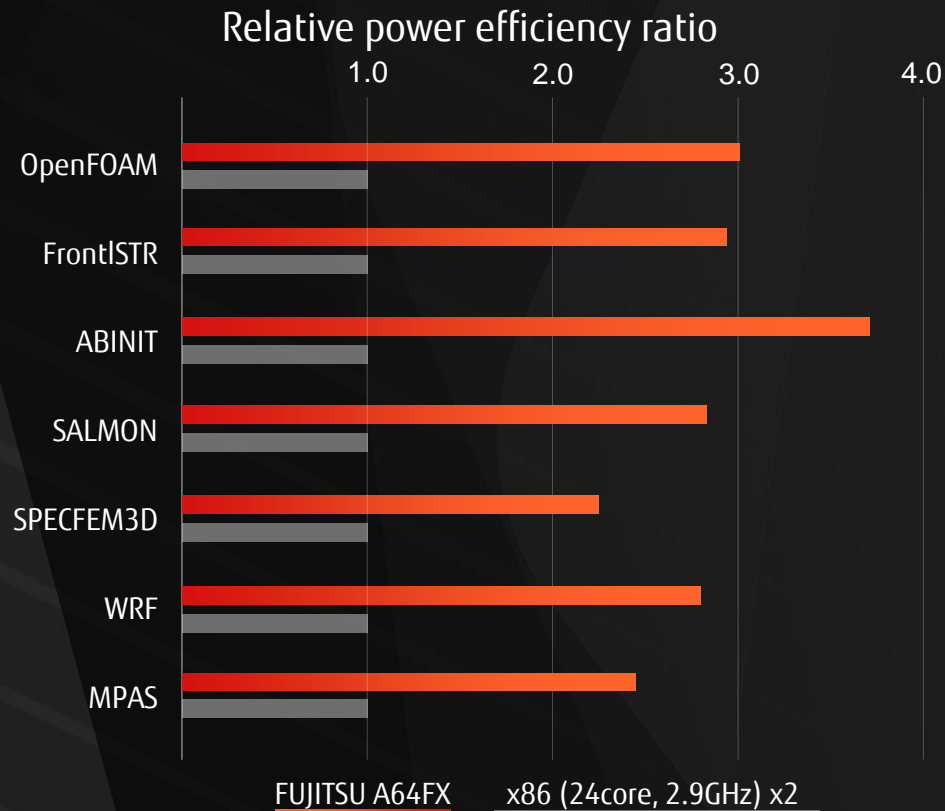
Line-up of OSS on AArch64 According to Spack Recipes

FUJITSU

- RIKEN and Fujitsu are maintaining and working to increase the number of OSS packages to be built / compiled for *aarch64*

of OSS to be built w/ compilers

Target	Fujitsu	GCC	Clang
aarch64	2,072/3,451 (60%)	2,387/3,451 (69%)	288/3,451 (8.4%)
x86	-	2,479/3,451 (72%)	768/3,574* (21%)

Spack Version: 9/6/19, 1:15 AM GMT, *10/31/19, 11:35 PM GMT

https://postk-web.r-ccs.riken.jp/oss/public/

理化学研究所 計算科学研究センター Post-K (Fugaku) Information
RIKEN Center for Computational Science

| Top | Specifications | Performance | Applications | OSS | Misc | Perf. Eval. | FAQ/Survey |

> Top > Open Source Software

List of Open Source Software which can be built on Fugaku

Spack will be used to manage open source software packages on Fugaku. Fugaku users can easily use pre-installed packages and built packages based Spack recipes. The following list shows the results of building/compiling packages for aarch64 according to the Spack recipes. Note that the results in this list do not guarantee that each package will work properly.

Build success rate per compiler

- Fujitsu Compiler: 2072/3451(60.04%)
- GCC: 2387/3451(69.17%)
- Clang: 288/3450(8.35%)

Last Modified: 11/14/19, 1:16 AM GMT
Spack Version: 9/6/19, 1:15 AM GMT

Click YES/NO for further infomation.

Package ↓	Version	Fujitsu Compiler	GCC	Clang
abinit	8.10.3	NO	NO	NO
abyss	2.1.4	NO	NO	NO
accfft	develop	NO	YES	NO
ack	2.22	YES	YES	YES
activeharmony	4.5	NO	NO	NO
acts-core	0.10.1	NO	YES	YES
adept-utils	1.0.1	NO	YES	YES

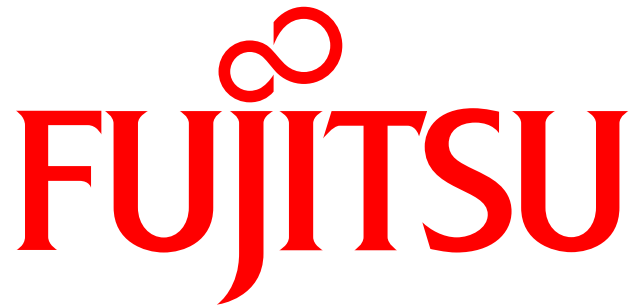# Summary: "Fugaku" and Fujitsu PRIMEHPC FX1000/FX700

- "Fugaku" is designed and runs applications at the highest level performance to be worthy of the name

- Arm HPC ecosystem and expanding apps portfolio are likened to the broad gradual slopes of Mt. Fuji

  - Fujitsu began production of "Fugaku", and announced PRIMEHPC FX1000 and FX700

  - Cray also announced introducing CS500 – Fujitsu A64FX Arm server

Fujitsu PRIMEHPC FX1000

FX700

https://www.riken.jp/pr/news/2019/20190827_1/