CCS International Symposium 2020, U Tsukuba, 6th Oct 2020

The Supercomputer “Fugaku” and A64FX Manycore Processor

Mitsuhisa Sato Team Leader of Architecture Development Team
Deputy project leader, FLAGSHIP 2020 project
Deputy Director, RIKEN Center for Computational Science (R-CCS)
Professor (Cooperative Graduate School Program), University of Tsukuba

Tetsuya Odajima and Yuetsu Kodama,
FLAGSHIP 2020 project, R-CCS

FLAGSHIP2020 Project “Fugaku”

Missions

- Building the Japanese national flagship supercomputer “Fugaku “(a.k.a post K), and
- Developing wide range of HPC applications, running on Fugaku, in order to solve social and science issues in Japan (application development projects was over at the end of march, 2020)

Overview of Fugaku architecture

Node: Manycore architecture

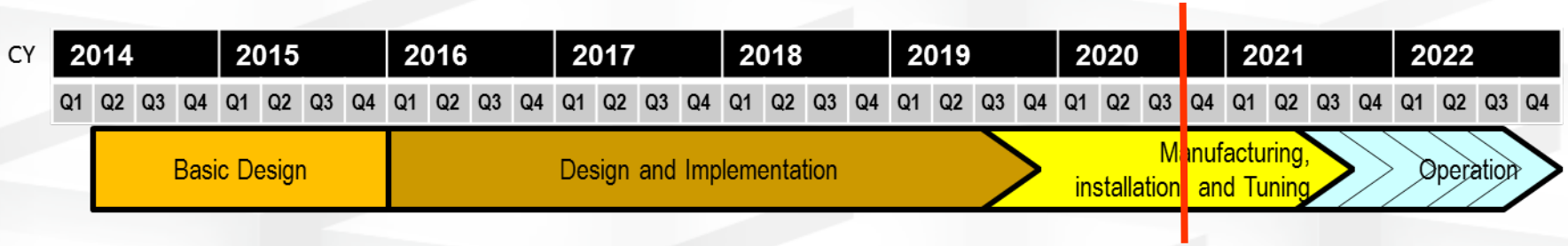
- Armv8-A + SVE (Scalable Vector Extension)
- SIMD Length: 512 bits
- # of Cores: 48 + (2/4 for OS) (> 3.0 TF / 48 core)
- Co-design with application developers and high memory bandwidth utilizing on-package stacked memory (HBM2) 1 TB/s B/W
- Low power : 15GF/W (dgemm)

Network: TofuD

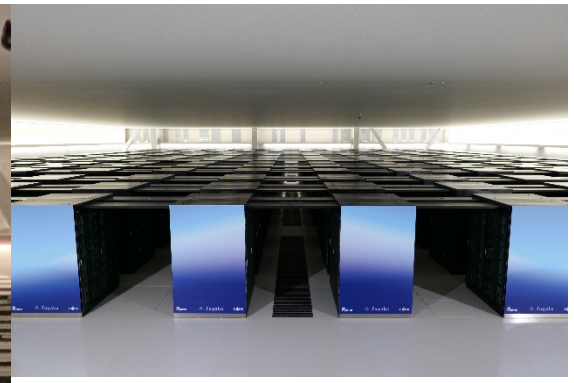
- Chip-Integrated NIC, 6D mesh/torus Interconnect

Status and Update

- March 2019: The Name of the system was decided as “Fugaku”
- Aug. 2019: The K computer decommissioned, stopped the services and shutdown (removed from the computer room)
- Oct 2019: access to the test chips was started.
- Nov. 2019: Fujitsu announce FX1000 and FX700, and business with Cray.
- Nov 2019: Fugaku clock frequency will be 2.0GHz and boost to 2.2 GHz.
- Nov 2019: Green 500 1st position!
- Oct-Nov 2019: MEXT announced the Fugaku “early access program” to begin around Q2/CY2020
- Dec 2019: Delivery and Installation of “Fugaku” was started.
- May 2020: Delivery completed
- June 2020: 1st in Top500, HPCG, Graph 500, HPL-AI at ISC2020



Supercomputer “Fugaku”



432 racks
158,976 nodes

Half exaflops in DP
1 Exaflops in SP!

Fugaku won 1st position in 4 benchmarks!

Bencmark	1st	Score	Unit	2nd	Score	1st/ 2nd
TOP500 (LINPACK)	Fugaku	415.5	PFLOPS	Summit（US）	148.6	2.80
HPCG	Fugaku	13.4	PFLOPS	Summit（US）	2.93	4.57
HPL-AI	Fugaku	1.42	EFLOPS	Summit（US）	0.55	2.58
Graph500	Fugaku	70,980	GTEPS	太湖之光 TaihuLight (China)	23,756	2.99

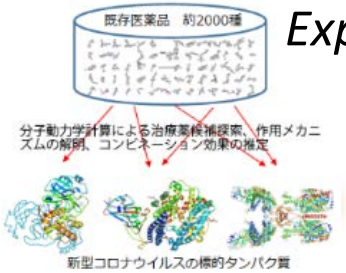
2 to 4 times faster in every benchmark!

MEXT Fugaku Program: Fight Against COVID19

Fugaku resources made available a year ahead of general production (more research topics under international solicitation)

A partner of international COVID-19 HPC Consortium

Medical-Pharma



Exploring new drug candidates for COVID-19

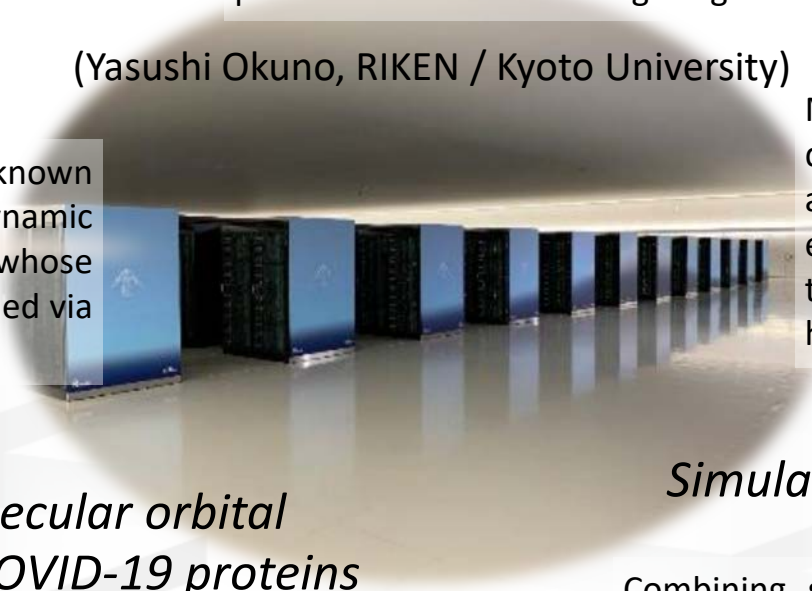
Large-scale MD to search & identify therapeutic drug candidates showing high affinity for COVID-19 target proteins from 2000 existing drugs

(Yasushi Okuno, RIKEN / Kyoto University)

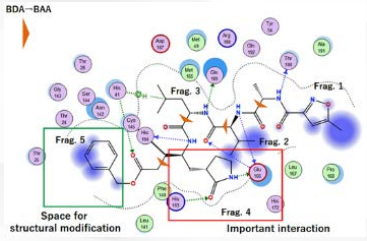
Prediction of conformation… dynamics of proteins on the surface of SARS-Cov-2

GENESIS MD to interpolate unknown experimentally undetectable dynamic behavior of spike proteins, whose static behavior has been identified via Cryo-EM

(Yuji Sugita, RIKEN)



Fragment molecular orbital calculations for COVID-19 proteins



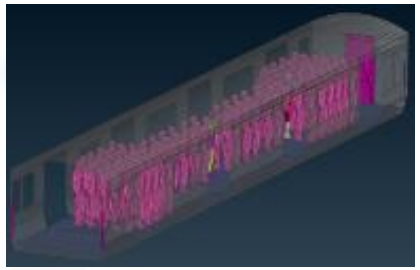
Large-scale, detailed interaction analysis of COVID-19 using Fragment Molecular Orbital (FMO) calculations using ABINIT-MP

(Yuji Mochizuki, Rikkyo University)

Societal-Epidemiology

Prediction and Countermeasure for Virus Droplet Infection under the Indoor Environment

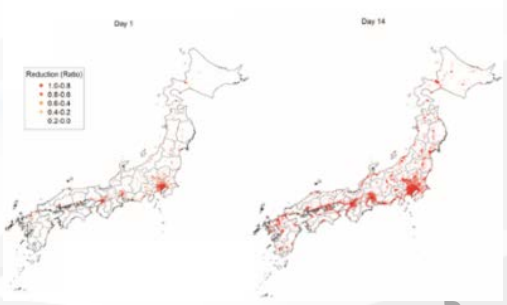
Massive parallel simulation of droplet scattering with airflow and hat transfer under indoor environment such as commuter trains, offices, classrooms, and hospital rooms



(Makoto Tsubokura, RIKEN / Kobe University)

Simulation analysis of pandemic phenomena

Combining simulations & analytics of disease propagation w/contact tracing apps, economic effects of lockdown, and reflections social media, for effective mitigation policies



(Nobuyasu Ito, RIKEN)

KPIs on Fugaku development in FLAGSHIP 2020 project

3 KPIs (key performance indicator) were defined for Fugaku development

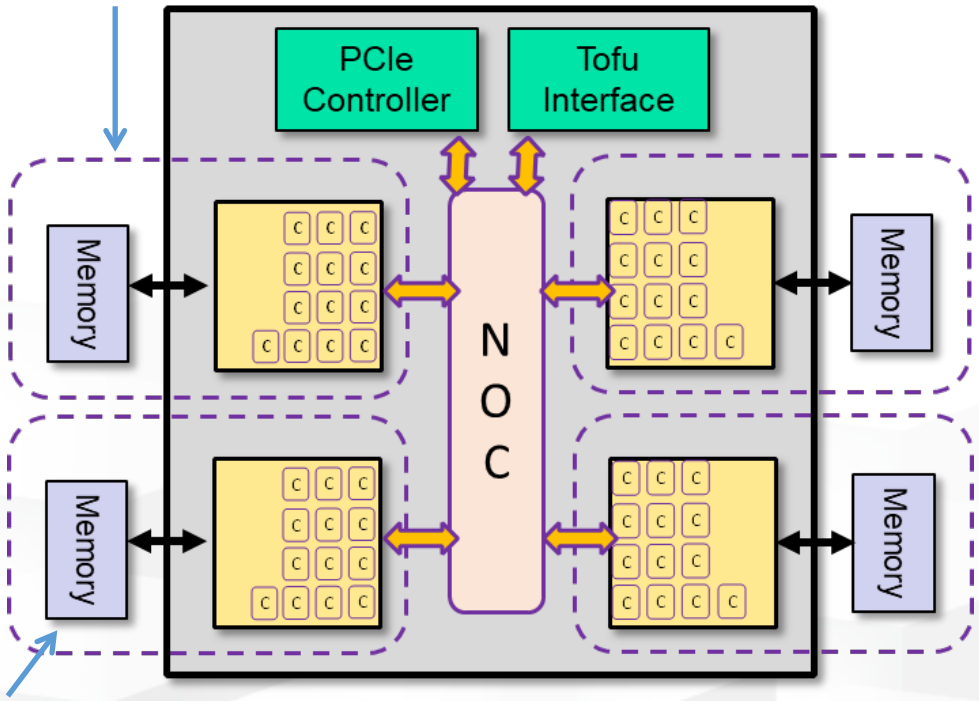
- **1. Extreme Power-Efficient System**
  - Maximum performance under Power consumption of 30 - 40MW (for system)
  - Approx. 15 GF/W (dgemm) confirmed by the prototype CPU => 1st in Green 500 !!!
- **2. Effective performance of target applications**
  - It is expected to exceed 100 times higher than the K computer's performance in some applications
  - 125 times faster in GENESIS (MD application), 120 times faster in NICAM+LETKF (climate simulation and data assimilation) were estimated
- **3. Ease-of-use system for wide-range of users**
  - Co-design with application developers
  - Shared memory system with high-bandwidth on-package memory must make existing OpenMP-MPI program ported easily.
  - No programming effort for accelerators such as GPUs is required.

CPU Architecture: A64FX

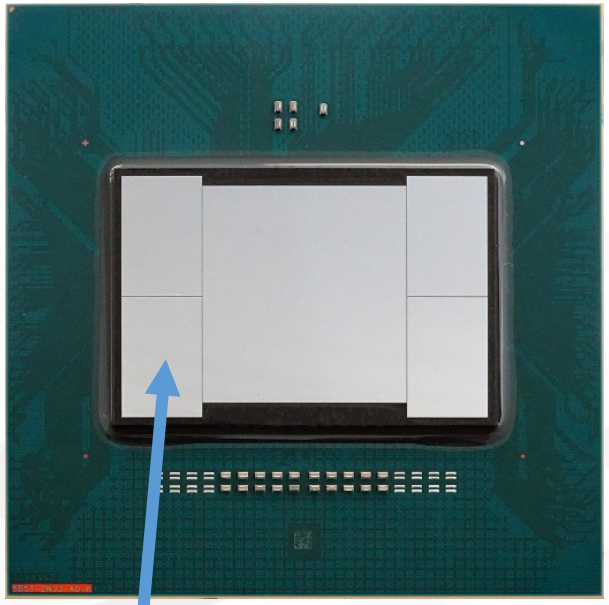
- **Armv8.2-A (AArch64 only) + SVE (Scalable Vector Extension)**
  - FP64/FP32/FP16 (https://developer.arm.com/products/architecture/a-profile/docs)
- **SVE 512-bit wide SIMD**
- **# of Cores: 48 + (2/4 for OS)**
- Co-design with application developers and high memory bandwidth utilizing on-package stacked memory: **HBM2(32GiB)**
- Leading-edge Si-technology (7nm FinFET), low power logic design **(approx. 15 GF/W (dgemm))**, and power-controlling knobs
- Clock frequency:
  - 2.0 GHz(normal), 2.2 GHz (boost)
- Peak performance
  - 3.0 TFLOPS@2GHz (>90% @ dgemm)
  - Memory B/W 1024GB/s (>80% stream)
  - Byte per Flops: 0.33

- "Common" programing model will be to run each MPI process on a NUMA node (CMG) with OpenMP-MPI hybrid programming.
- 48 threads OpenMP is also supported.

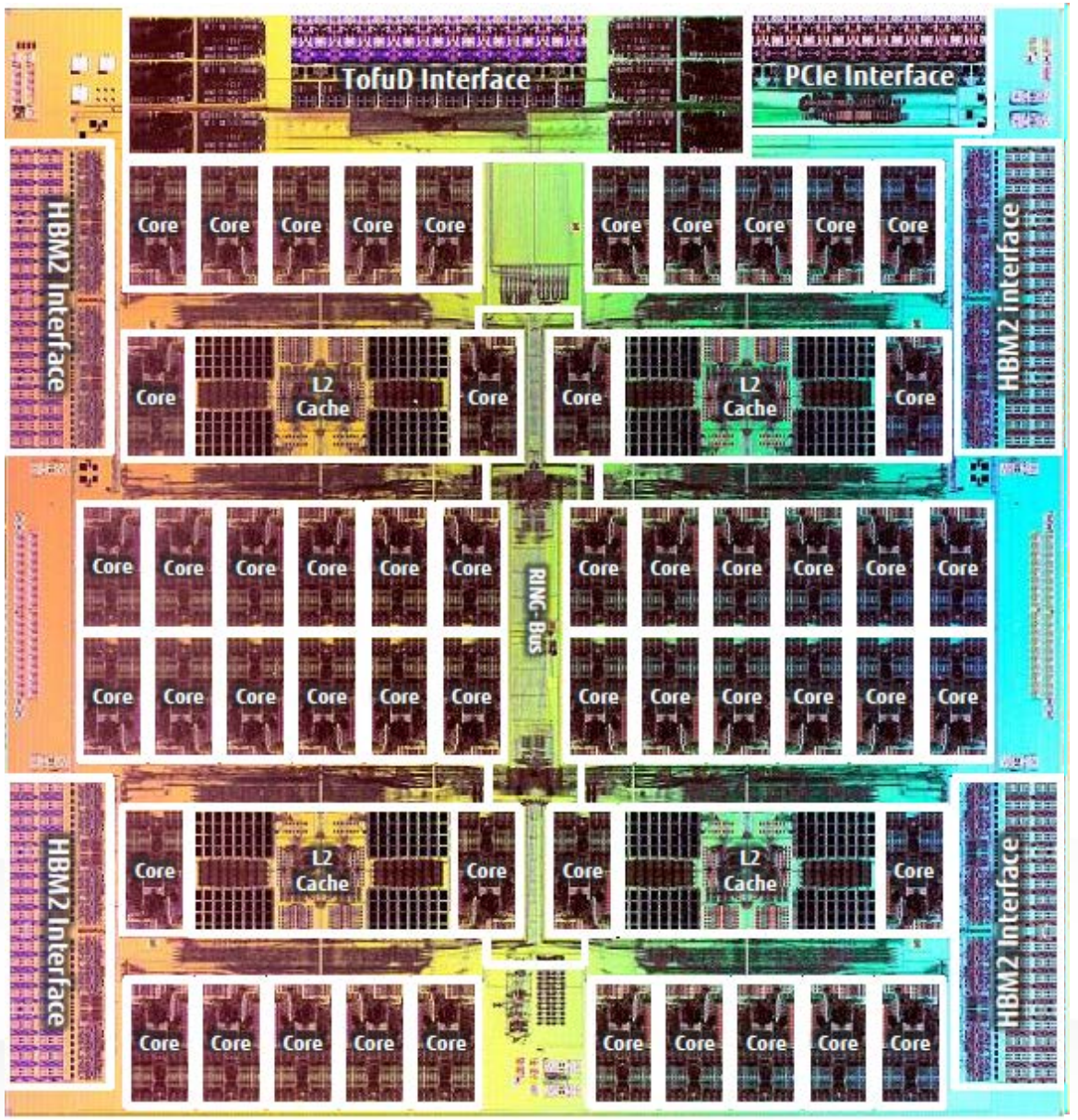
CMG(Core-Memory-Group): NUMA node
12+1 core



- **TSMC 7nm FinFET**
- **CoWoS technologies for HBM2**



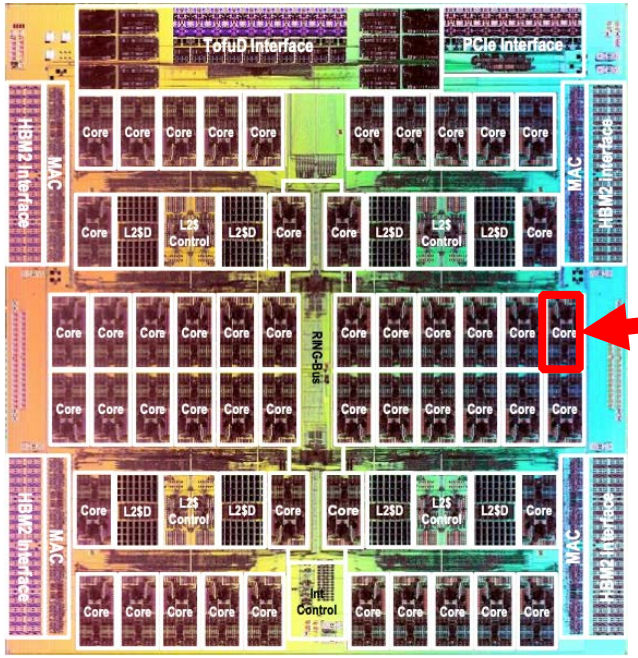
HBM2



Comparison of Die-size

- **A64FX: 52 cores (48 cores), 400 mm² die size (8.3 mm²/core), 7nm FinFET process (TSMC)**
- **Xeon Skylake: 20 tiles (5x4), 18 cores, ~485 mm² die size (estimated) (26.9 mm²/core), 14 nm process (Intel)**
- **More than 3 times larger per core.**

A64FX:
400 mm²
(20 x 20)

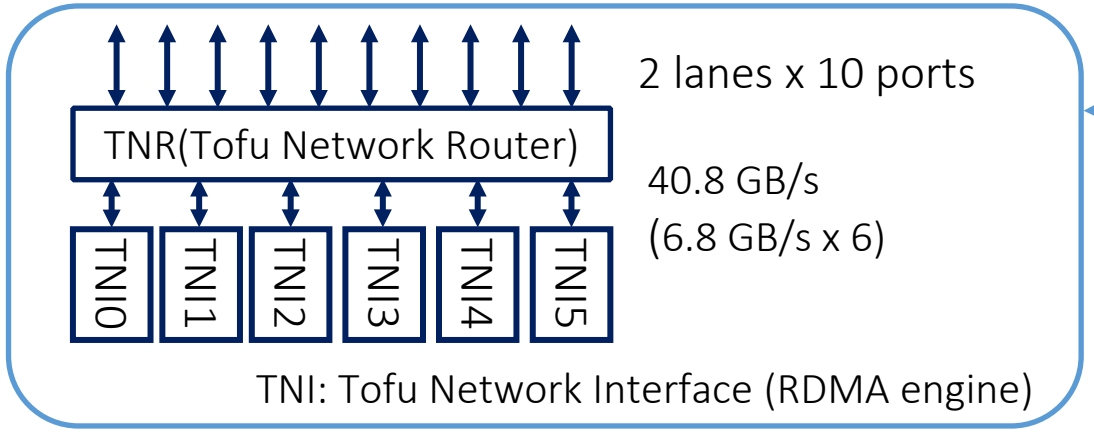


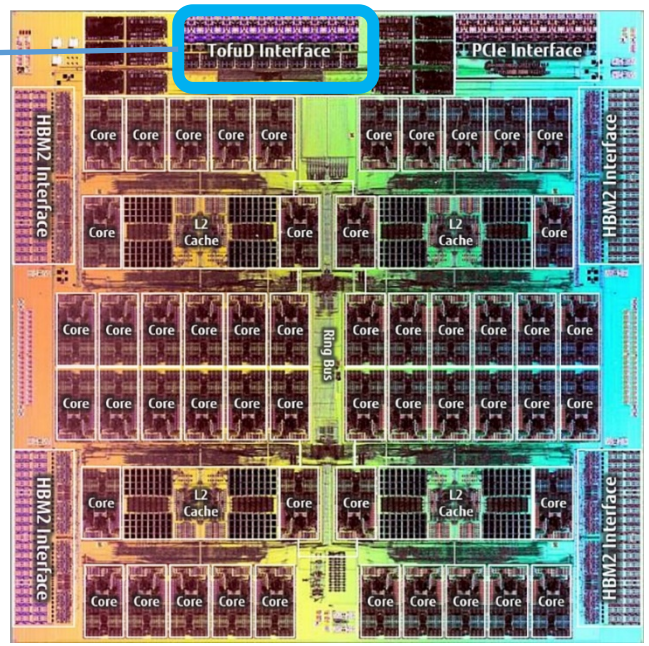
https://www.fujitsu.com/jp/solutions/business-technology/tc/catalog/ff2019-post-k-computer-development.pdf

https://en.wikichip.org/wiki/intel/microarchitectures/skylake_(server)

Xeon Skylake, High Core Count:
4 x 5 tiles, 18 cores, 2 tiles used for memory interface
485 mm² (22 x 22)

TofuD Interconnect





- 6 RDMA Engines
- Hardware barrier support
- Network operation offloading capability

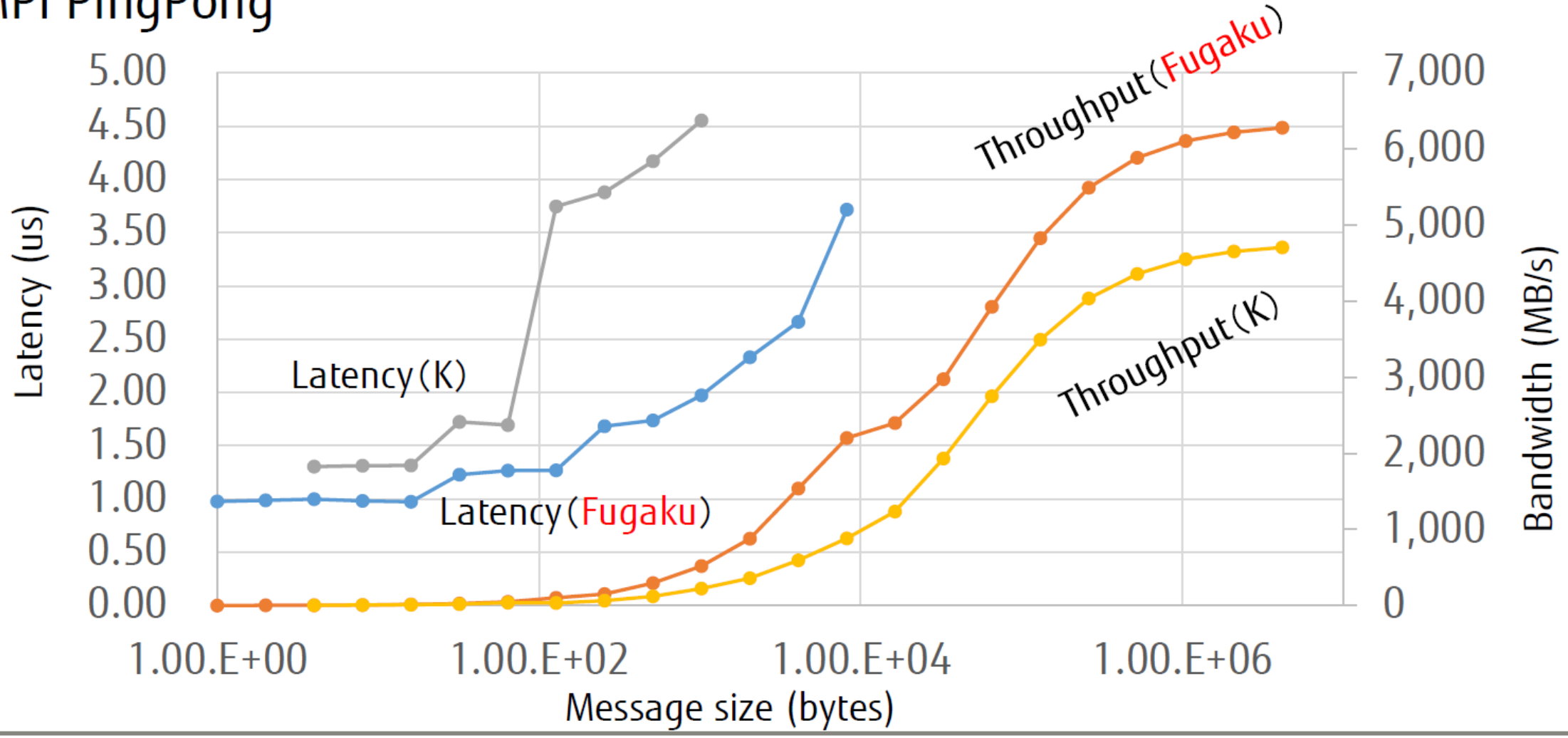
8B Put latency	0.49 – 0.54 usec
1MiB Put throughput	6.35 GB/s

rf. Yuichiro Ajima, et al. , “The Tofu Interconnect D,” IEEE Cluster 2018, 2018.

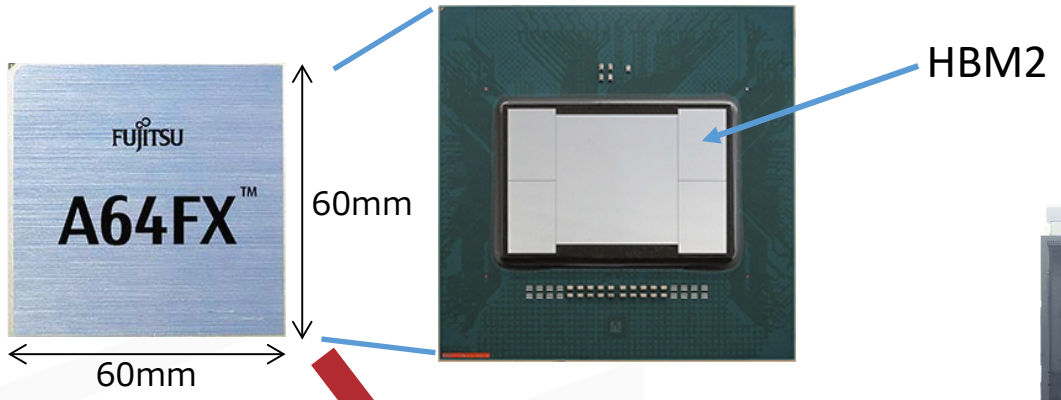
TofuD: MPI_Send/Receive Latency and BW

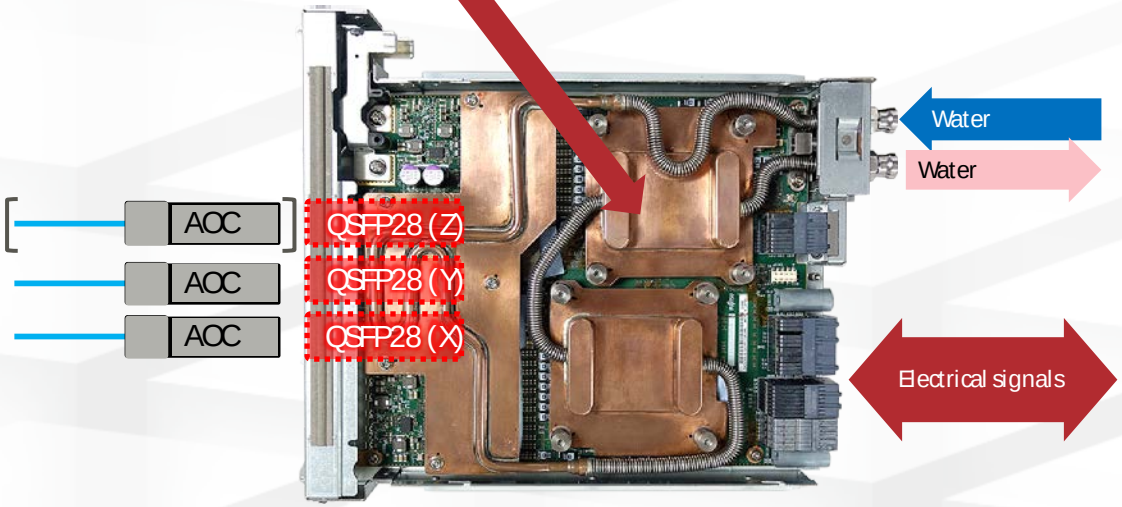
FUJITSU

MPI PingPong

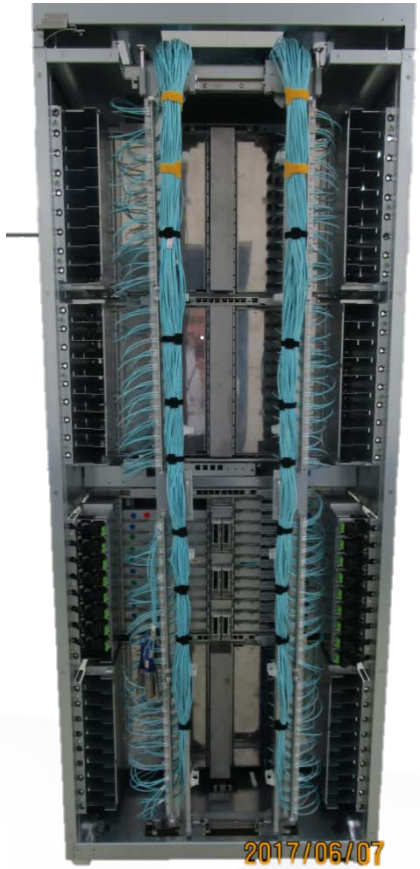


Fugaku prototype board and rack





2 CPU / CMU



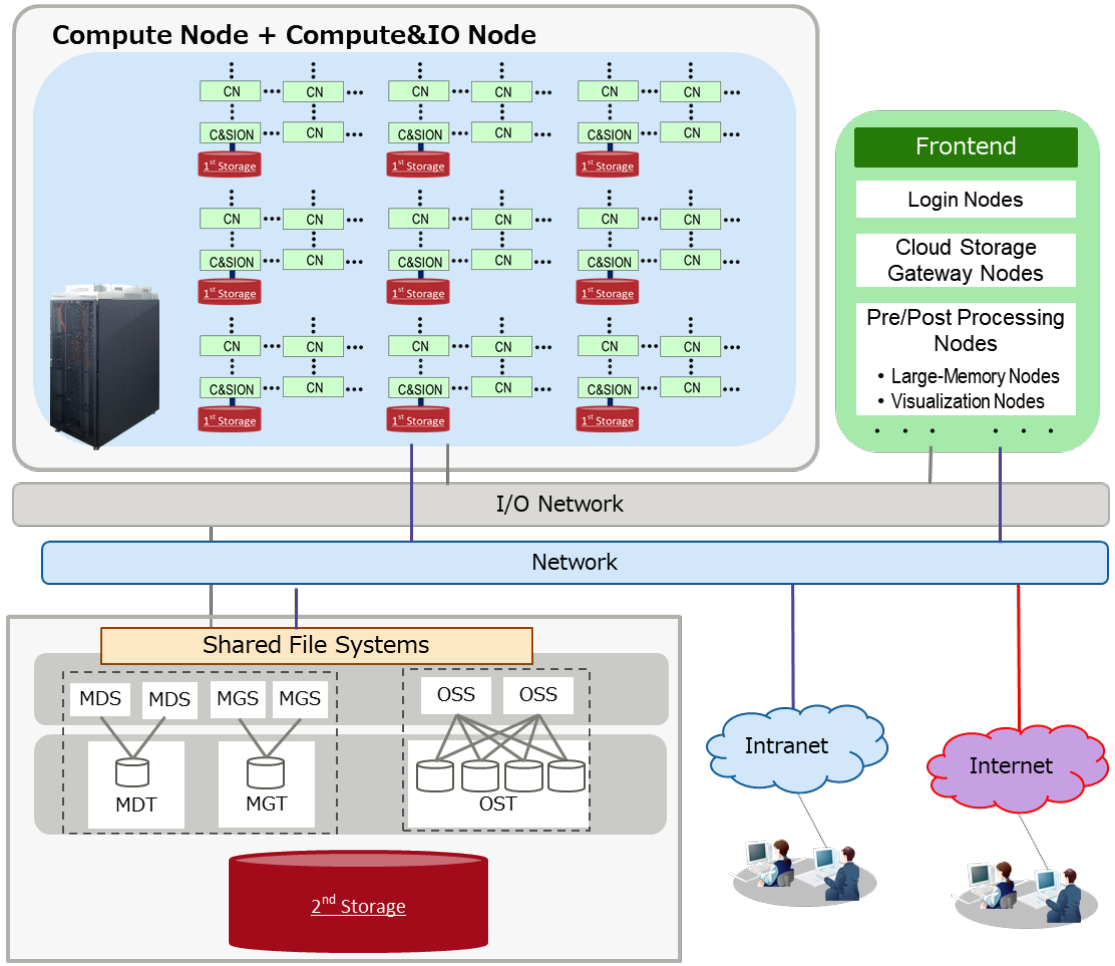
Shelf: 48 CPUs (24 CMU)
Rack: 8 shelves = 384 CPUs (8x48)

Fugaku System Configuration

- **158,976node**

Boost mode: 3.3792TF x 150k+ = 500+ PF

- **Two types of nodes**
  - Compute Node and Compute & I/O Node connected by Fujitsu TofuD, 6D mesh/torus Interconnect
- **3-level hierarchical storage system**
  - 1st Layer
    - One of 16 compute nodes, called Compute & Storage I/O Node, has SSD about 1.6 TB
    - Services
      - Cache for global file system
      - Temporary file systems
        - Local file system for compute node
        - Shared file system for a job
  - 2nd Layer
    - Fujitsu FEFS: Lustre-based global file system
  - 3rd Layer
    - Cloud storage services



Advances from the K computer

	K computer	Fugaku	ratio
# core	8	48	
Si tech. (nm)	45	7	
Core perf. (GFLOPS)	16	64(70)	4(4.4)
Chip(node) perf. (TFLOPS)	0.128	3.072 (3.379)	24 (26.4)
Memory BW (GB/s)	64	1024	
B/F (Bytes/FLOP)	0.5	0.33	
#node / rack	96	384	4
#node/system	82,944	158,976	
System perf.(DP PFLOPS)	10.6	488 (537) 977(1070)	42.3(52.2) 84.6(104.4)

Si Tech

SVE

CMG&Si Tech

HBM

More than **7.6 M General-purpose cores**!

- SVE increases core performance
- Silicon tech. and scalable architecture (CMG) to increase node performance
- HBM enables high bandwidth

Value in blankets
Indicate the number
At boost mode (2.2GHz)

Benchmark Results on test chip A64FX

- **CloverLeaf (UK Mini-App Consortium), Fortran/C**
  - A hydrodynamics mini-app to solve the compressible Euler equations in 2D, using an explicit, second-order method
  - Stencil calculation

- **TeaLeaf (UK Mini-App Consortium), Fortran**
  - A mini-application to enable design-space explorations for iterative sparse linear solvers
  - https://github.com/UK-MAC/TeaLeaf_ref.git
  - Problem size: Benchmarks/tea_bm_5.in, end_step=10 -> 3

- **LULESH (LLNL), C**
  - Mini-app representative of simplified 3D Lagrangian hydrodynamics on an unstructured mesh, indirect memory access

Processors for comparison

	A64FX	TX2 (ThunderX2)	SKL (Skylake)
# cores	48 (1 CPU)	56 (28 x 2 sockets)	24 (12 x 2 sockets)
Clock	2.0 GHz (Normal)	2.0 GHz	2.6 GHz (※)
SIMD	SVE 512-bit	NEON 128-bit	AVX512 512-bit
Memory Peak bandwidth	HBM2 1,024 GB/s	DDR4-8ch 341 GB/s	DDR4-6ch 256 GB/s
Network	TofuD	InfiniBand FDR x 1	InfiniBand HDR x 1
Compiler Options	Fujitsu compiler 4.1.0 -Kfast,openmp	Arm HPC compiler 19.1 -Ofast -fopenmp -march=armv8.1-a	Intel compiler 19.1 -O3 -qopenmp -march=native

(※) AVX512 instruction is executed at 90% peak Feq.

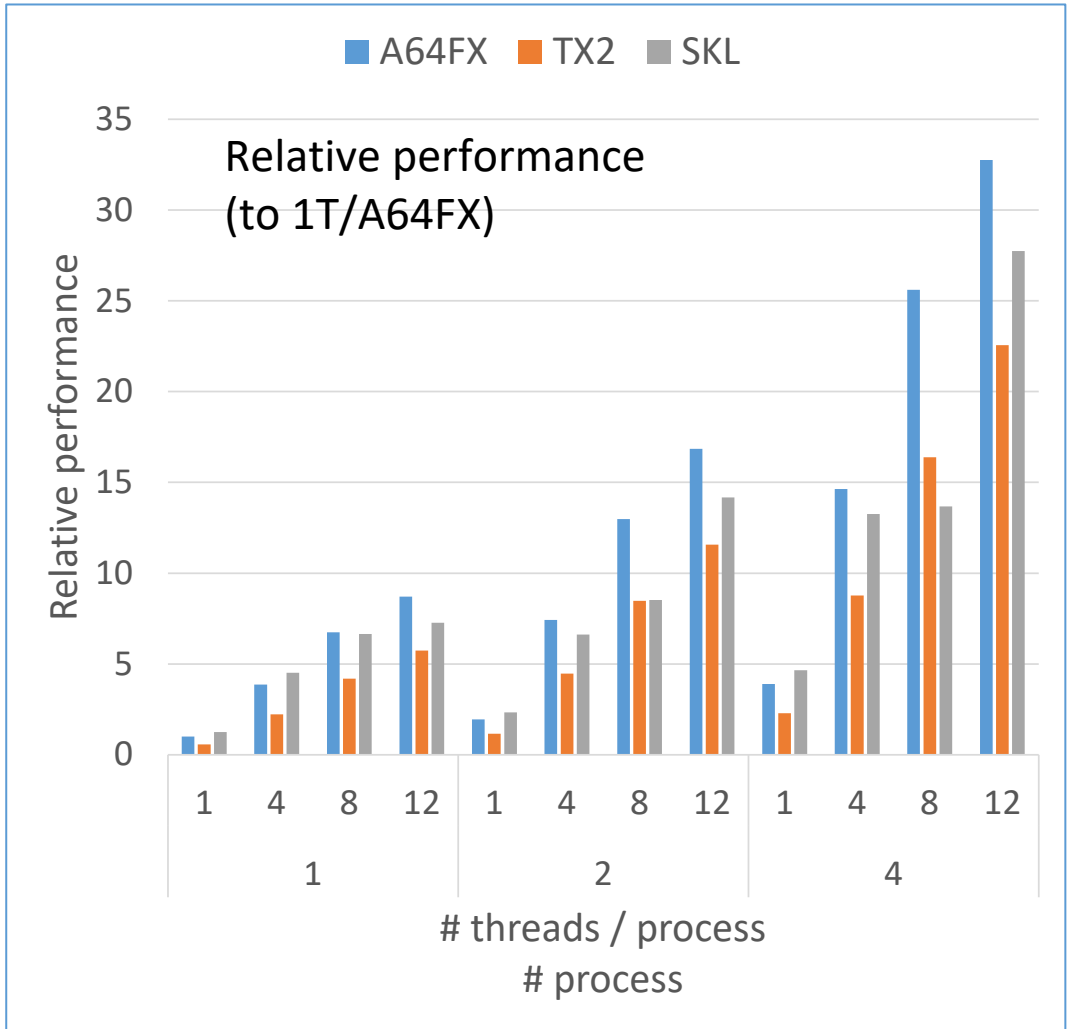
Threads and sockets and nodes

- **#threads ≦ 12**
  - A64FX: execute on only CMG0
  - TX2, SKL: execute on only Socket0
- **12 < #threads ≦ 24**
  - A64FX: execute on CMG0 and CMG1
  - TX2, SKL: execute on one node (max #threads: 12 on a socket)
- **24 < #threads ≦ 48**
  - A64FX: execute no one node
  - TX2, SKL: execute on two node (max #threads: 12 on a socket)

Disclaimer:
The software used for the evaluation, such as the compiler, is still under development and its performance may be different when the supercomputer Fugaku starts its operation.

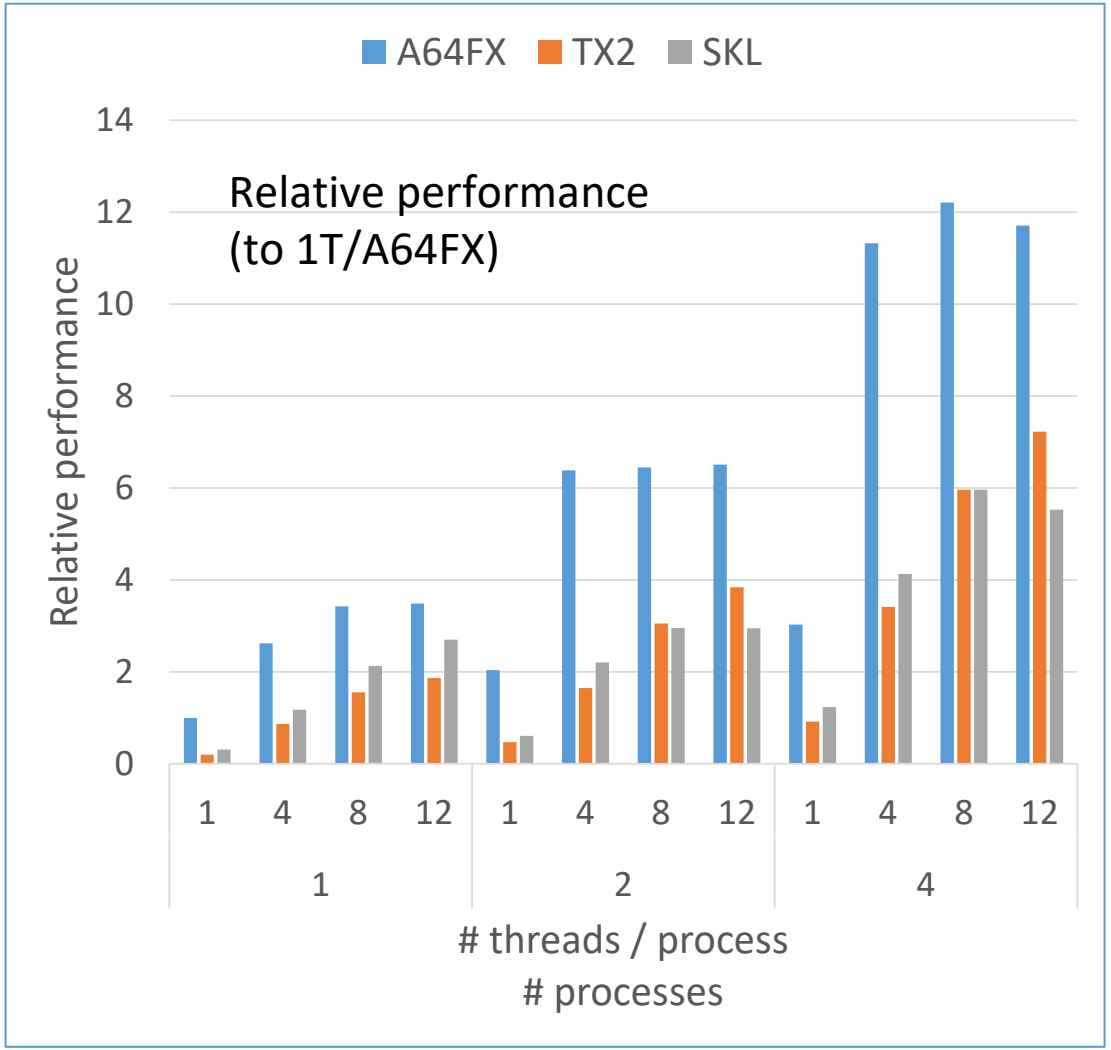
CloverLeaf

- Good scalability by increasing the number of threads within CMG.
- The performance of one A64FX is comparable (better) to that of two nodes (4 chips) of Skylake



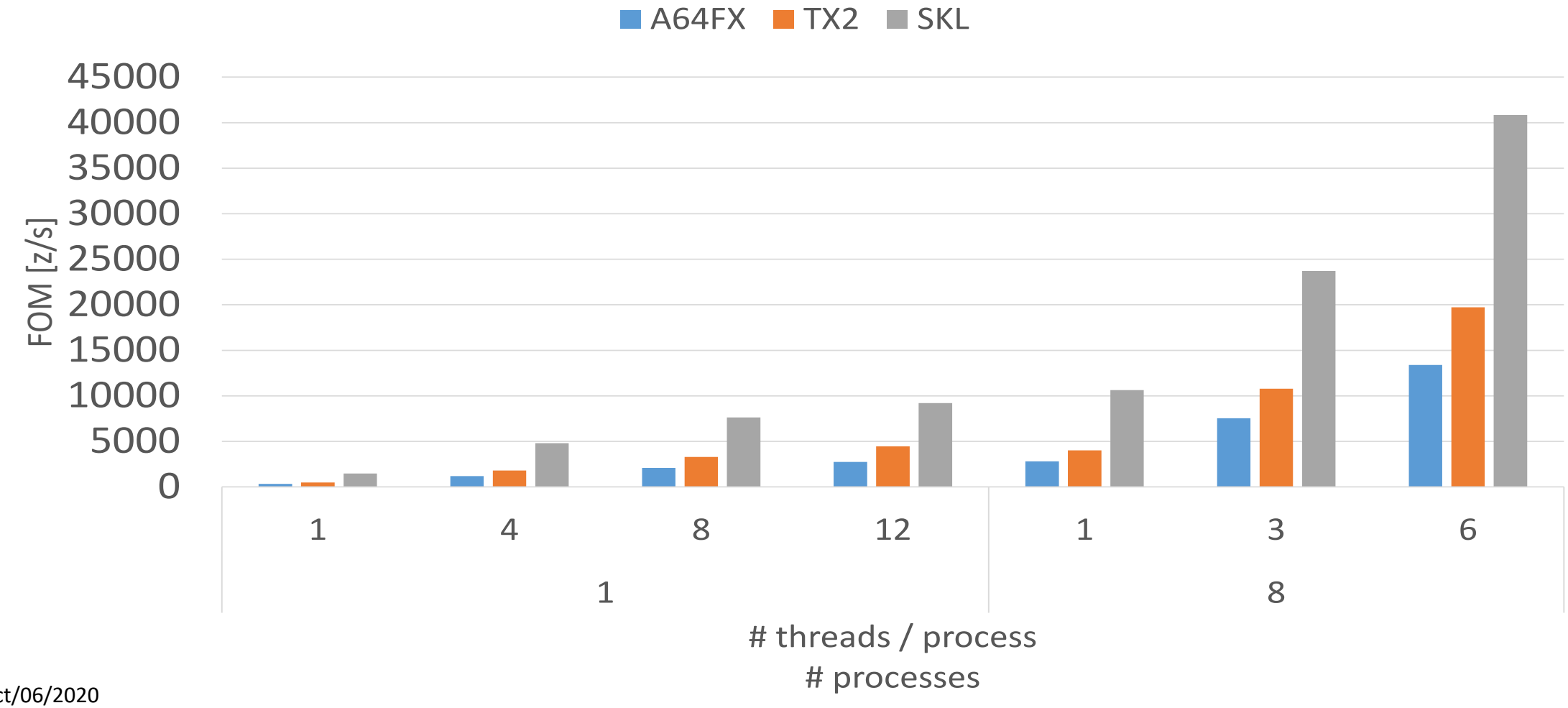
TeaLeaf

- Memory bandwidth intensive application. The speedup is limited for more than 4 threads due to the memory bandwidth.
- The performance of one A64FX is twice better than that of two nodes (4 chips) of Skylake. It reflects the difference of total memory bandwidth.



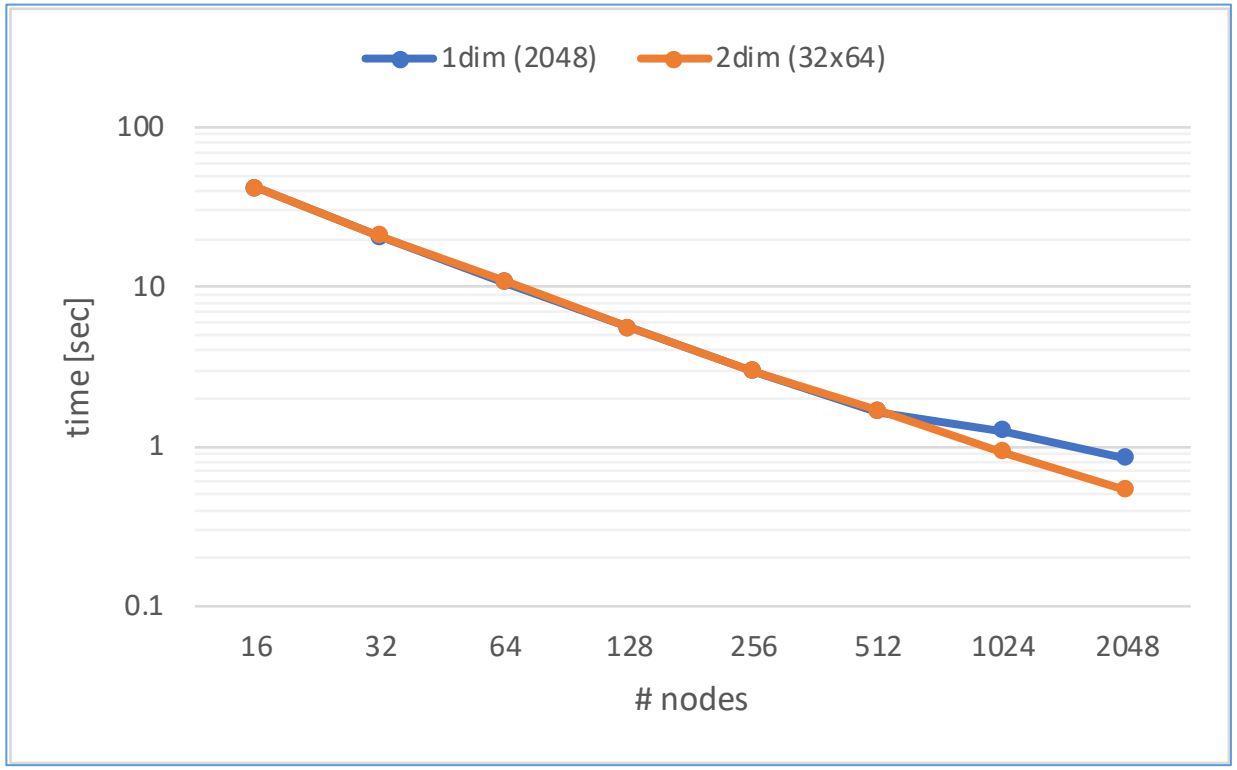
LULESH

- A64FX performance is less than Thx2 and Intel one
- We found low vectorization (SIMD (SVE) instructions ratio is a few %)
- We need more code tuning for more vectorization using SIMD

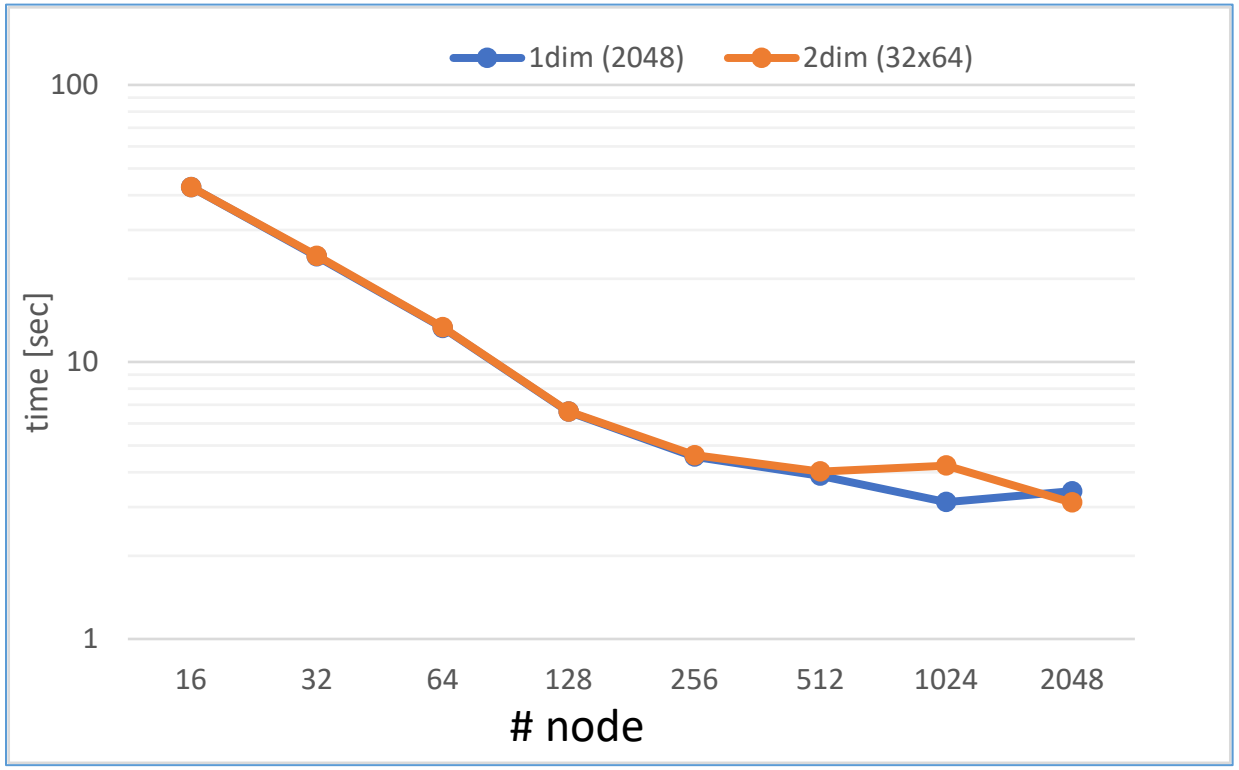


Scalability for Multi-nodes

- **Strong scaling in CloverLeaf and TeaLeaf (FlatMPI) up to 2048 nodes**
- **CloverLeaf : Good scalability for 2D**
- **TeaLeaf: Limited by communication (helo and dot)**



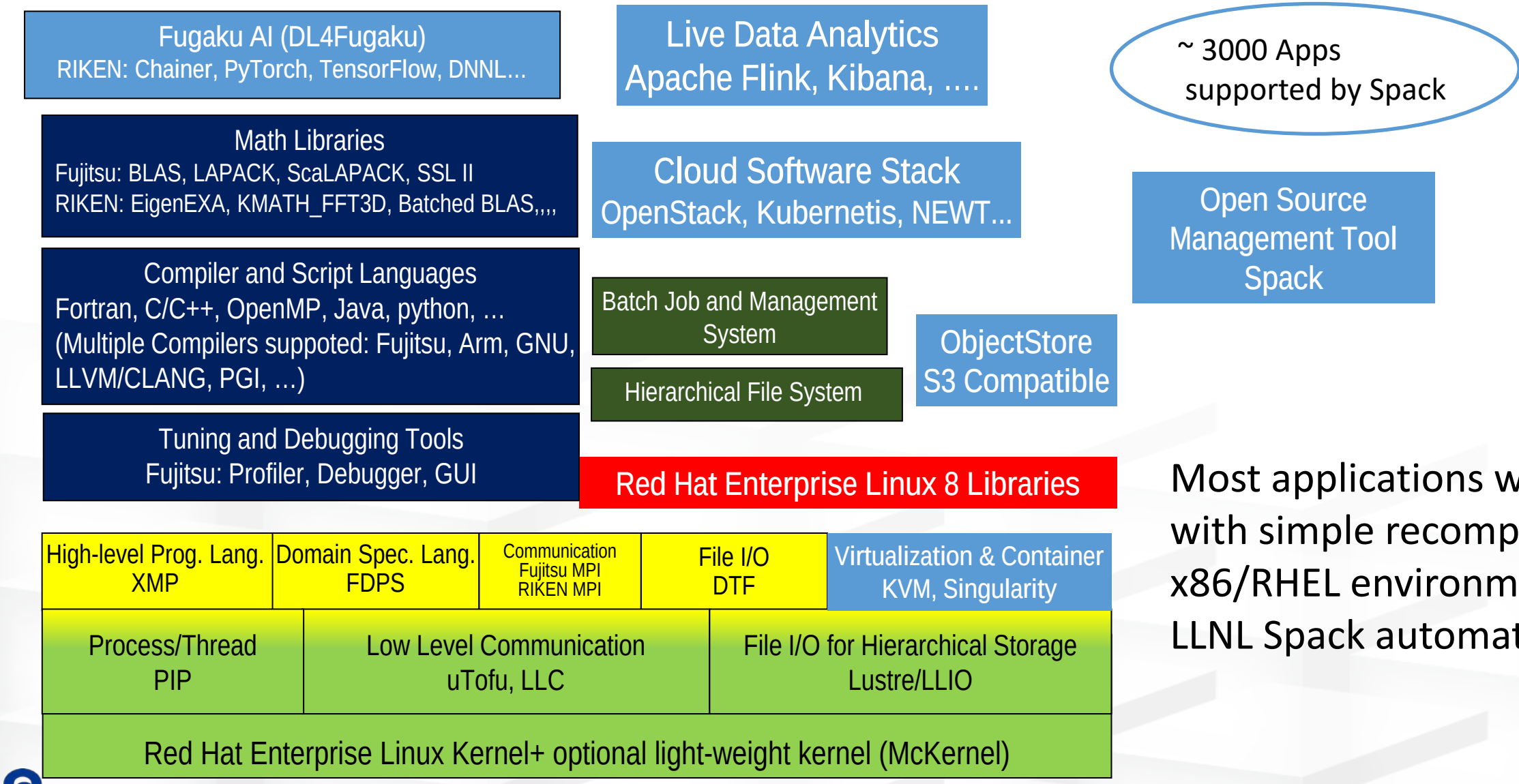
CloverLeaf



TeaLeaf

Problem Size: InputDecks/clover_bm2048_short.in

Fugaku / Fujitsu FX1000 System Software Stack



Most applications will work with simple recompile from x86/RHEL environment.
LLNL Spack automates this.

System software and Programming models & languages for “Fugaku”

- Standard programming model is OpenMP (for NUMA node(CMG)) + MPI
  - Both OpenMPI (by Fujitsu) and MPICH (by Riken) are supported.
  - OpenMP 4.x is supported by Fujitsu compiler. LLVM-based compiler and gcc available.
  - uTofu low-level comm. Layer for Tofu-D interconnect.
- Container and Virtual machine (KVM, Singularity, …)
- DL4Fugaku: AI framework for Fugaku, used in Chainer, PyTorch, TensorFlow
- Many Open-source software will be ported using Spack

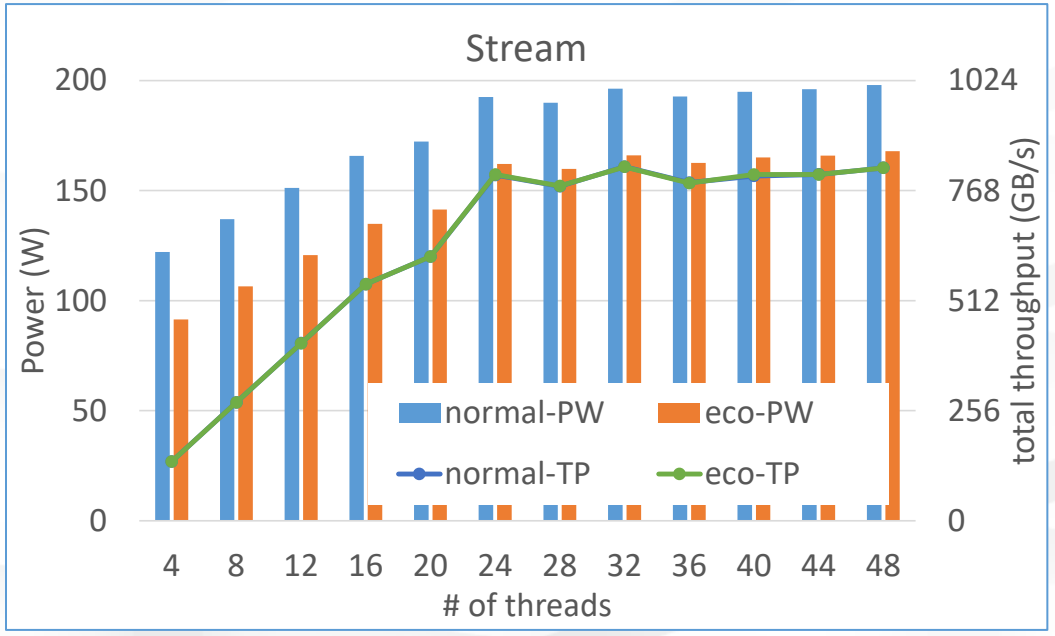
- System software and Programming tools, Math-Libs developed by RIKEN
  - McKernel: Light-weight Kernel enabling jitter-less environment for large-scale parallel program execution.
  - XcalableMP directive-based PGAS Language
  - FDPS: DLS for Framework for Developing Particle Simulators.
  - EigenExa: Eigen-value math library for large-scale parallel systems.

Low-power Design & Power Management

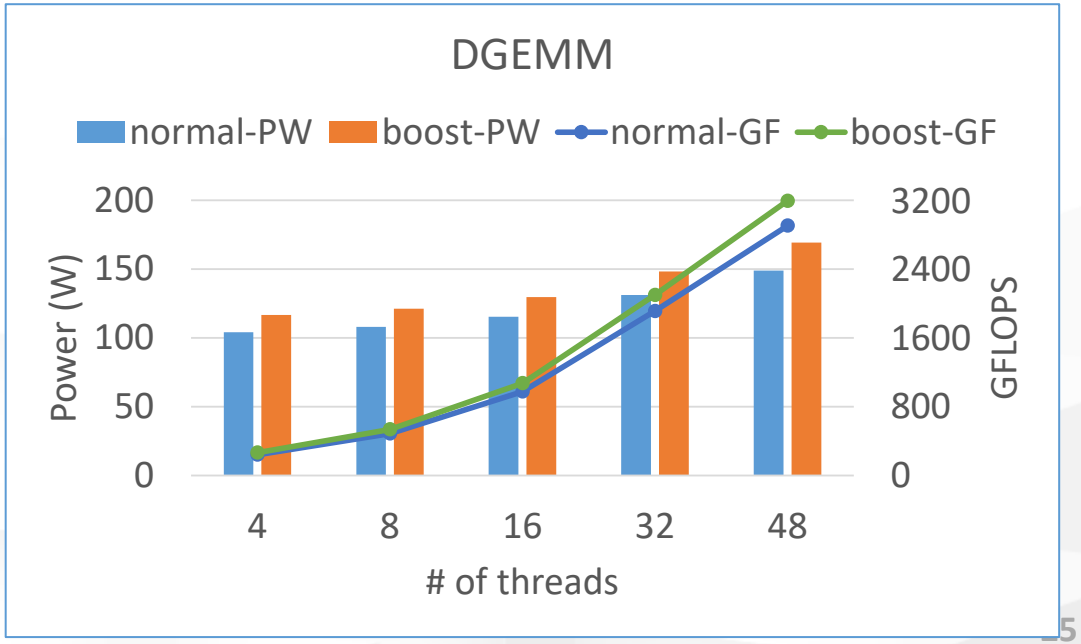
- **7nm FinFET (TSMC) with low-power logic design**
- **A64FX provides power management function called "Power Knob"**
  - FL pipeline usage: FLA only, EX pipeline usage : EXA only, Frequency reduction …
  - User program can change "Power Knob" for power optimization
  - "Energy monitor" facility enables chip-level power monitoring and detailed power analysis of applications
- **"Eco-mode" : FLA only with lower "stand-by" power for ALUs**
  - Reduce the power-consumption for memory intensive apps.
  - 4 apps out of 9 target applications select "eco-mode" for the max performance under the limitation of our power capacity (Even using HBM2!)
- **Retention mode: power state for de-activation of CPU with keeping network alive**
  - Large reduction of system power-consumption at idle time
- **"Power Knobs" can be controlled by Sandia PowerAPIs and setting running modes.**
  - We are now designing the accounting system to give incentive to make use of power-knobs
  - "Power budget" as well as node-hour budget.

Boost mode & Eco mode

- **Power & Performance of STREAM using Eco mode**
  - The performance is almost the same as that in normal mode (24 threads hits 80% of peak memory bandwidth
  - The power increases upto 24 threads.
  - 15%-25% reduction comparing to that in normal mode.



- **Power & Performance of DGEMM (in Fujitsu Lib) using Boost mode**
  - Reach to 95% out of peak performance
  - The performance is 10% better than that in normal mode.
  - The power increases by 13.7%
  - The power-efficiency decreases by 3.3 %

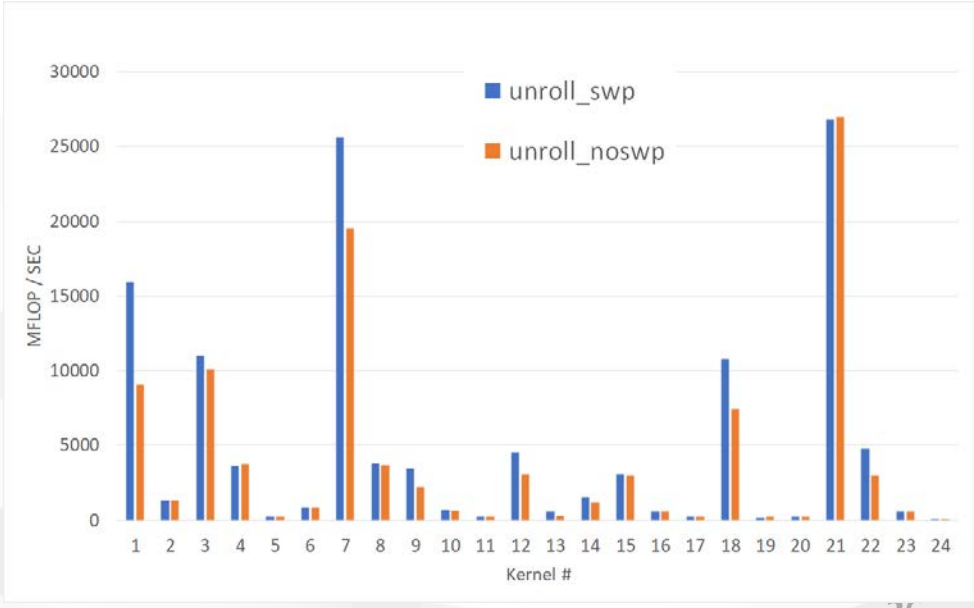


Tips on Performance tuning for A64FX

- **HPC-oriented design**
  - Small core ⇒ Less O3 resources
  - (Relatively) Long pipeline
    - 9 cycles for floating point operations
    - Core has only L1 cache
  - High-throughput, but long-latency
  - Pipeline often stalls for loops having complex body.
  - Compiler optimization (Fujitsu compiler)
    - SWP: software pipelining, loop fission, …
- **How to exploit SIMD**
  - SIMD is a key for performance on A64FX
  - OpenMP SIMD directives

	A64FX	Skylake
ReOrder Buffer	128 entries	224 entries
Reservation Station	60 (=10x2+20x2) entries	97 entries
Physical Vector Register	128 (=32 + 96) entries	168 entries
Load Buffer	40 entries	72 entries
Store Buffer	24 entries	56 entries

A64FX : https://github.com/fujitsu/A64FX
Skylake : https://en.wikichip.org/wiki/intel/microarchitectures/skylake_(server)



Performance improvement by SWP in Livermore Kernels by Fujitsu compiler

Concluding remarks

- **We are now sure to achieve 3 KPIs**
  - Power-efficiency
  - Effective Performance of applications.
  - Ease-of-use
- **Well-balanced system for several apps**
- **In 2020, Fugaku is partially used by early users, incl. COVID-19 apps**
- **"Startup Preparation Project" allocation is open for the usage upto March, 2021.**
- **Open to international users through HPCI, general allocation April 2021 (application starting Sept. 2020)**

- **For the next of Fugaku, …**
  - “Dark-side” of (our) co-design of HPC, … No so “disruptive” architecture., but, … ease-of-use
  - Will need application-specific accelerators for more power-efficiency in near future?
  - Or is there any room to improve on the existing processor architecture?