AHUG, 20 June 2019 @ ISC 2019, Frankfurt

The post-K project and Fujitsu ARM-SVE enabled A64FX processor

for energy-efficiency and sustained application performance

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

FLAGSHIP2020 Project

☐ 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

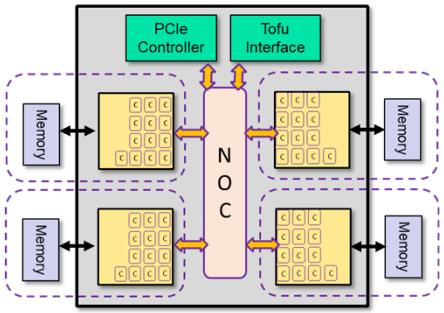
☐ Overview of Fugaku architecture

Node: Manycore architecture

- Armv8-A + SVE (Scalable Vector Extension)
- SIMD Length: 512 bits
- # of Cores: 48 + (2/4 for OS) (> 2.7 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

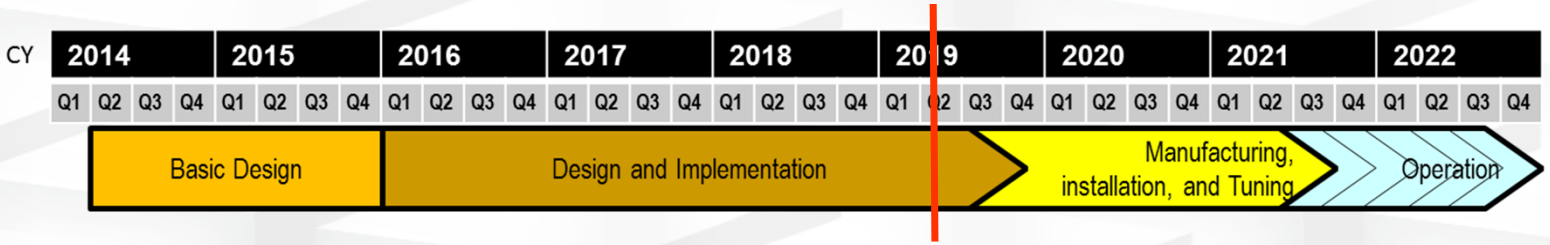


Fujitsu A64FX processor

Prototype board

☐ Status and Update

- "Design and Implementation" completed
- The official contract with Fujitsu to manufacture, ship, and install hardware for Fugaku is done
- RIKEN revealed #nodes > 150K
- The Name of the system was decided as "Fugaku"
- RIKEN announced the Fugaku early access program to begin around Q2/CY2020

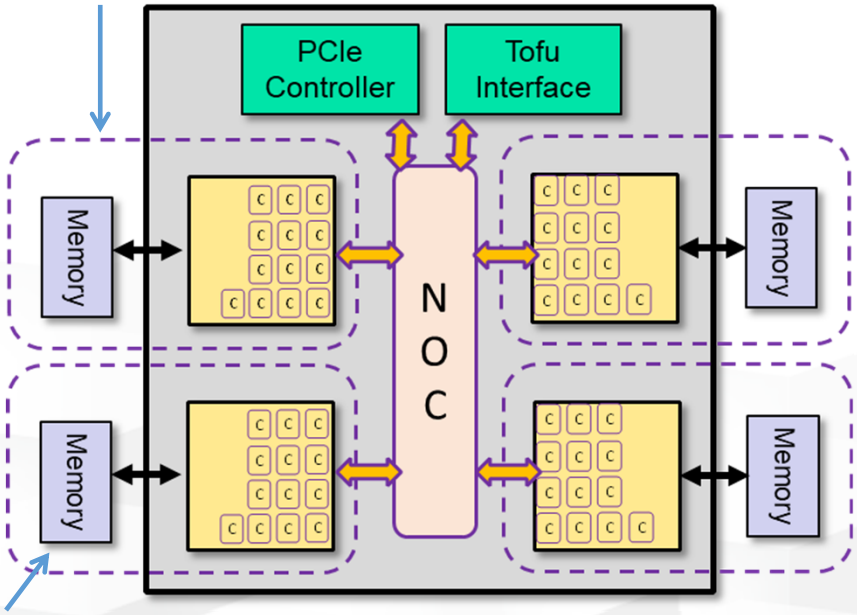


CPU Architecture: A64FX

- **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
- PCIe Gen3 16 lanes
- Peak performance
  - > 2.7 TFLOPS (>90% @ dgemm)
  - Memory B/W 1024GB/s (>80% stream)
  - Byte per Flops: approx. 0.4

- "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.

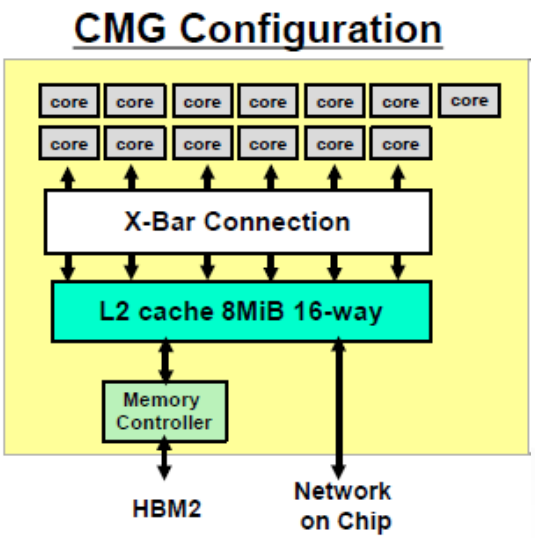
CMG(Core-Memory-Group): NUMA node
12+1 core

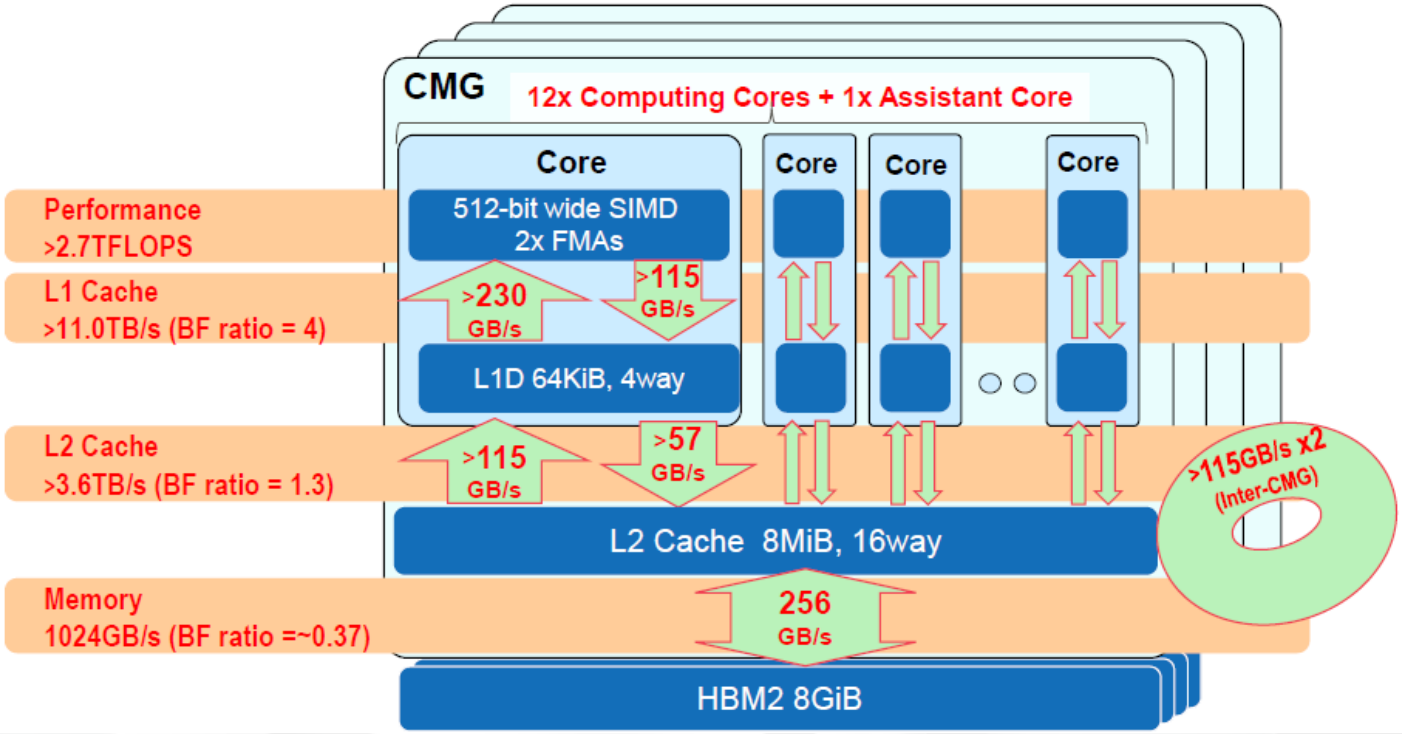


HBM2: 8GiB

CMG (Core Memory Group)

- CMG: 13 cores (12+1) and L2 cache (8MiB 16way) and memory controller for HBM2 (8GiB)
- X-bar connection in a CMG maximize efficiency for throughput of L2 (>115 GB/s for R, >57 GB/s for W)
- Assistant core is dedicated to run OS demon, I/O, etc
- 4 CMGs support cache coherency by ccNUMA with on-chip directory (> 115GB/s x 2 for inter-CMGs)



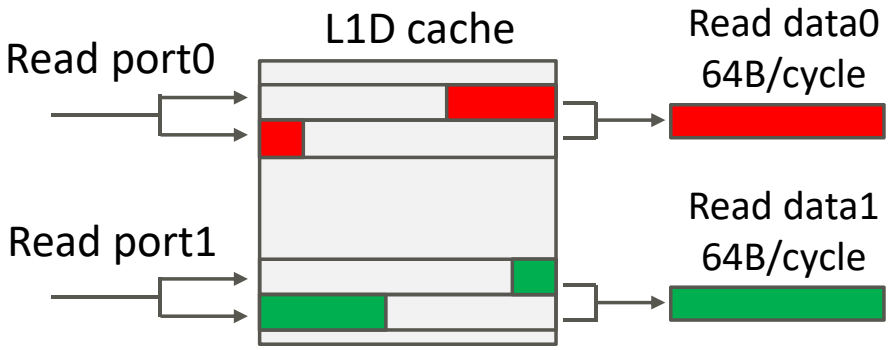


Figures from the slide presented in Hotchips 30 by Fujitsu

A64FX Optimized Load Efficiency for Apps Performance

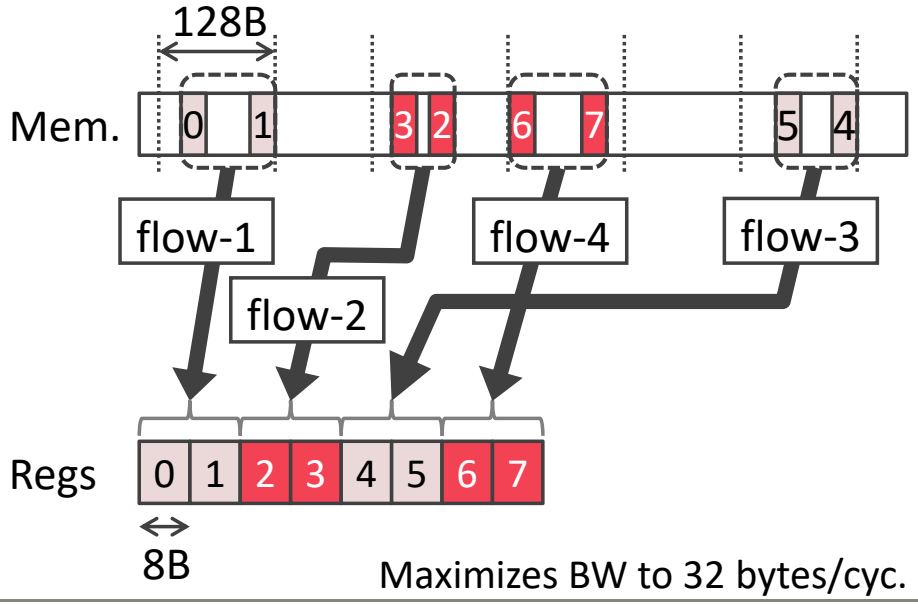
FUJITSU

- 128 bytes/cycle sustained bandwidth even for unaligned SIMD load



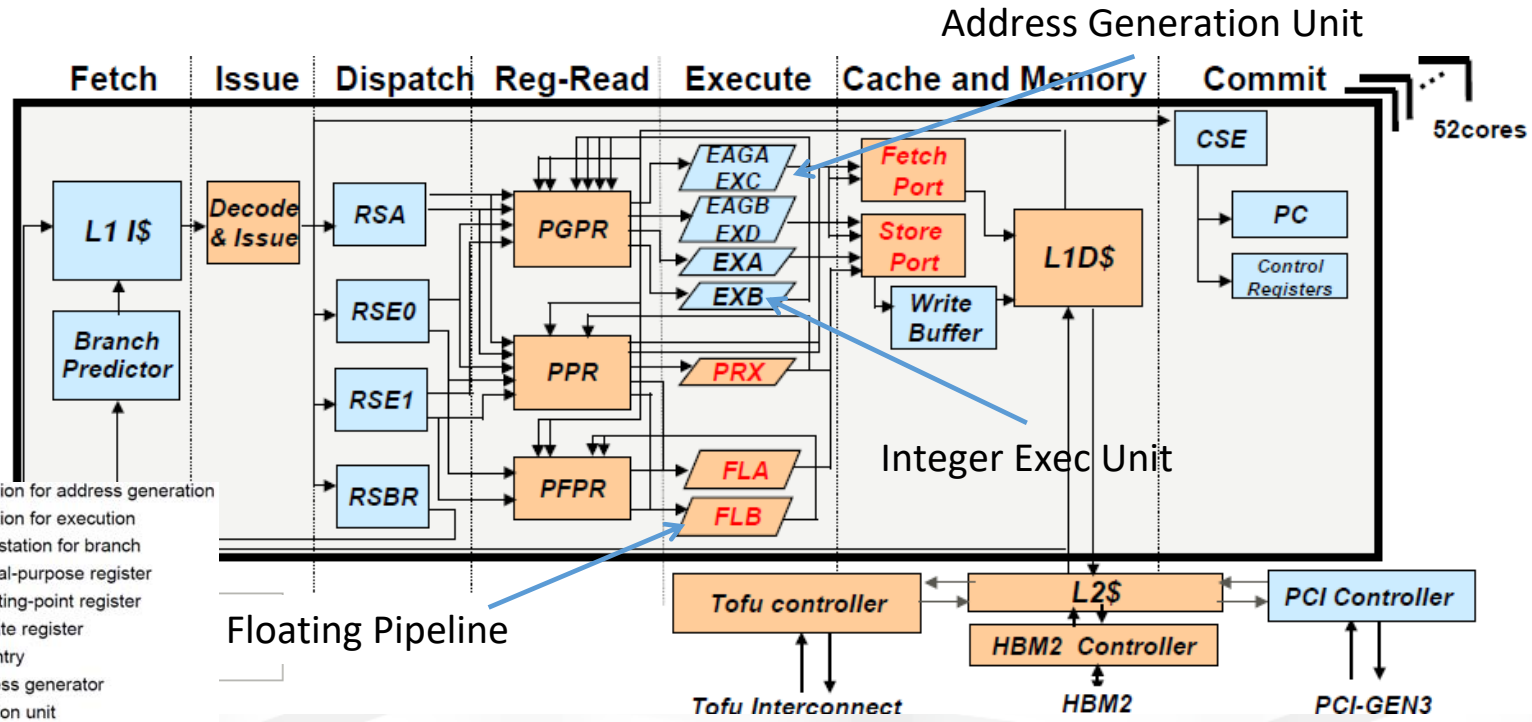
- "Combined Gather" doubles gather (indirect) load's data throughput, when target elements are within a "128-byte aligned block" for a pair of two regs, even & odd

Suggested through Co-design work w/ app teams



FX64A Core Pipeline

- **Superscalar Arch with out-of-order, branch prediction, inherited from Fujitsu SPARC**
- **L1D cache: 64 KiB, 4 ways, "Combined Gather" mechanism on L1**
- **SIMD and predicate operations**
  - 2x 512-bit wide SIMD FMA + Predicate Operation + 4x ALU (shared w/ 2x AGEN)
  - 2x 512-bit wide SIMD load or 512-bit wide SIMD store



Figures from the slide presented in Hotchips 30 by Fujitsu

- RSA: Reservation station for address generation
- RSE: Reservation station for execution
- RSBR: Reservation station for branch
- PGPR: Physical general-purpose register
- PFPR: Physical floating-point register
- PPR: Physical predicate register
- CSE: Commit stack entry
- EAG: Effective address generator
- EX : Integer execution unit
- FL : Floating-point execution unit
- PRX : Predicate execution unit

Low-power Design & Power Management

- **Leading-edge Si-technology (7nm FinFET)**
- **Low power logic design (15 GF/W @ dgemm)**

- **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

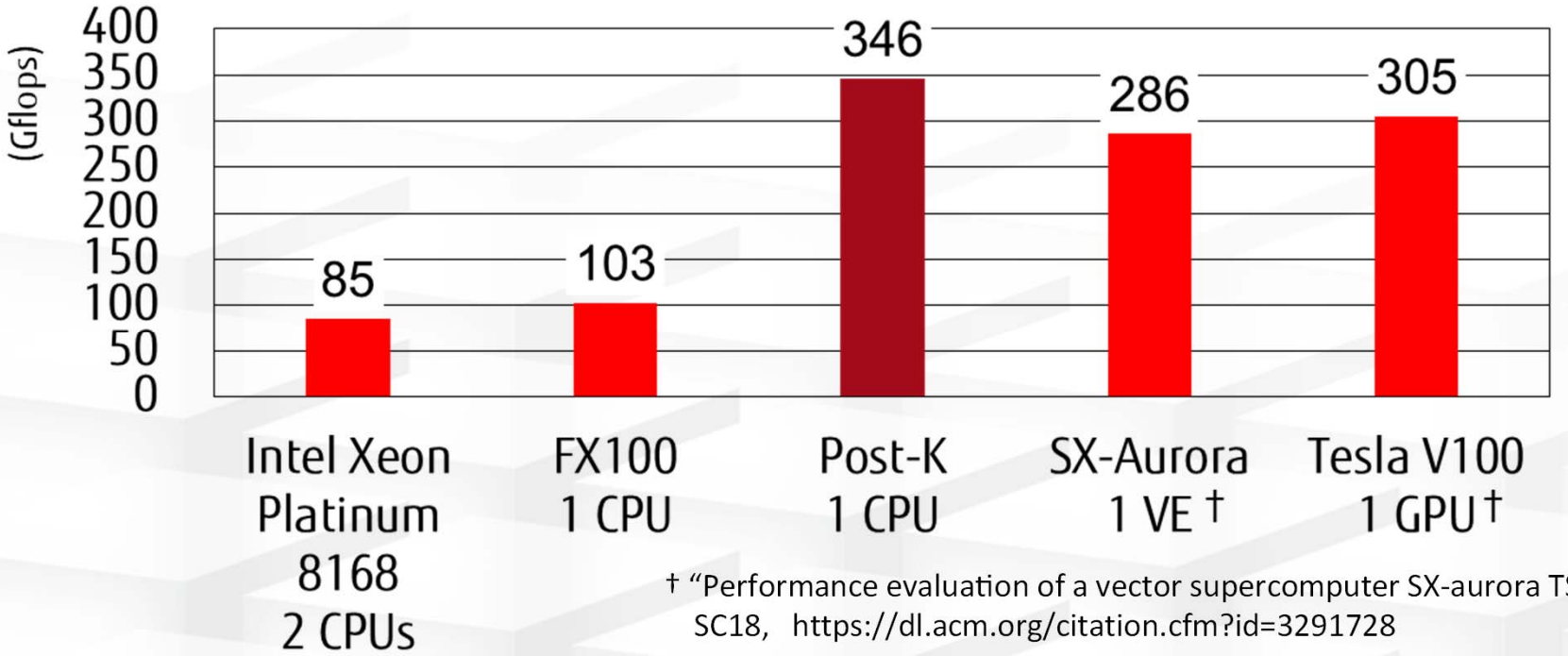
- **"Eco-mode" : FLA only with lower "stand-by" power for ALUs**
  - Reduce the power-consumption for memory intensive apps.
  - 4 apps out of 9 target applications select "eco-mode" for the max performance under the limitation of our power capacity (Even using HBM2!)
- **Retention mode: power state for de-activation of CPU with keeping network alive**
  - Large reduction of system power-consumption at idle time

- **HPL & Stream**
  - > 2.5TF / node for dgemm
  - > 830GB/s /node for stream triad

- **Himeno Benchmark (Fortran90)**

Peak Performance

	Fugaku	K
Peak DP (double precision)	400+ Pflops (x34+)	11.3 Pflops
Peak SP (single precision)	800+ Pflops (x70+)	11.3 Pflops
Peak HP (half precision)	1600+ Pflops (x141+)	--
Total memory bandwidth	150+ PB/sec (x29+)	5.2PB/sec



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

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

- **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**

Target Application's Performance

- **Performance Targets**
  - 100 times faster than K for some applications (tuning included)
  - 30 to 40 MW power consumption

https://postk-web.r-ccs.riken.jp/perf.html

Predicted Performance of 9 Target Applications

As of 2019/05/14

Area	Priority Issue	Performance Speedup over K	Application	Brief description
Health and longevity	1. Innovative computing infrastructure for drug discovery	**x125+**	**GENESIS**	MD for proteins
	2. Personalized and preventive medicine using big data	**x8+**	**Genomon**	Genome processing (Genome alignment)
Disaster prevention and Environment	3. Integrated simulation systems induced by earthquake and tsunami	**x45+**	**GAMERA**	Earthquake simulator (FEM in unstructured & structured grid)
	4. Meteorological and global environmental prediction using big data	**x120+**	**NICAM+ LETKF**	Weather prediction system using Big data (structured grid stencil & ensemble Kalman filter)
Energy issue	5. New technologies for energy creation, conversion / storage, and use	**x40+**	**NTChem**	Molecular electronic (structure calculation)
	6. Accelerated development of innovative clean energy systems	**x35+**	**Adventure**	Computational Mechanics System for Large Scale Analysis and Design (unstructured grid)
Industrial competitivenes s enhancement	7. Creation of new functional devices and high-performance materials	**x30+**	**RSDFT**	Ab-initio program (density functional theory)
	8. Development of innovative design and production processes	**x25+**	**FFB**	Large Eddy Simulation (unstructured grid)
Basic science	9. Elucidation of the fundamental laws and evolution of the universe	**x25+**	**LQCD**	Lattice QCD simulation (structured grid Monte Carlo)

KPIs on Fugaku development in FLAGSHIP 2020 project

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

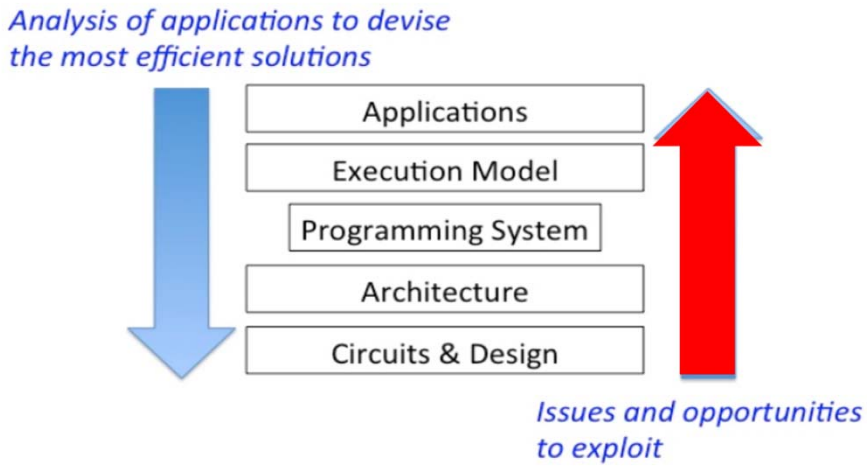
- **1. Extreme Power-Efficient System**
  - Maximum performance under Power consumption of 30 - 40MW (for system)
  - Approx. 15 GF/W (dgemm) confirmed by the prototype CPU

- **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

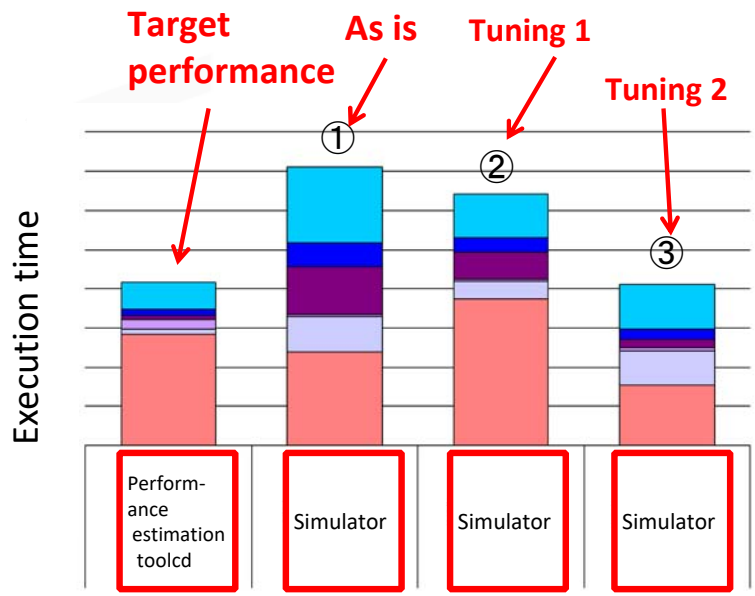
- **3. Ease-of-use system for wide-range of users**
  - Shared memory system with high-bandwidth on-package memory must make existing OpenMP-MPI program ported easily.
  - No programming effort for accelerators such as GPUs is required.
  - Co-design with application developers

Co-design of Apps for Architecture

- **Tools for performance tuning**
  - Performance estimation tool
    - Performance projection using Fujitsu FX100 execution profile
    - Gives "target" performance
  - Post-K processor simulator
    - Based on gem5, O3, cycle-level simulation
    - Very slow, so limited to kernel-level evaluation



- **Co-design of apps**
  - 1. Estimate "target" performance using performance estimation tool
  - 2. Extract kernel code for simulator
  - 3. Measure exec time using simulator
  - 4. Feed-back to code optimization
  - 5. Feed-back to compiler



RIKEN Arm-SVE gem5 simulator

- **Fugaku (PostK) Fujitsu A64FX processor simulator based on gem-5**
  - The processor simulator will give a detail performance results including estimated executing time, cache-miss, the number of instruction executed in O3.
  - The user can understand how the compiled code for SVE is executed on A64FX processor for optimization.
  - NDA with RIKEN/Fujitsu is required.

- **Open version of Arm-SVE gem5 simulator in docker file (x86)**
  - Arm-SVE gem5 with “open parameters” (free) and gcc for Arm-SVE included
  - Can be used for architecture exploration
  - Available on Linaro docker hub:

    https://hub.docker.com/r/linaro/gem5-riken-open

- **Compilers for Fujitsu A64FX processor**
  - Fujitsu Compilers : Fortran, C, C++. Fully-tuning for “postK” architecture.
  - Arm Compiler : LLVM-based compiler to generate code forArmv8-A + SV. C,C++ by Clang, Fortran by Flang

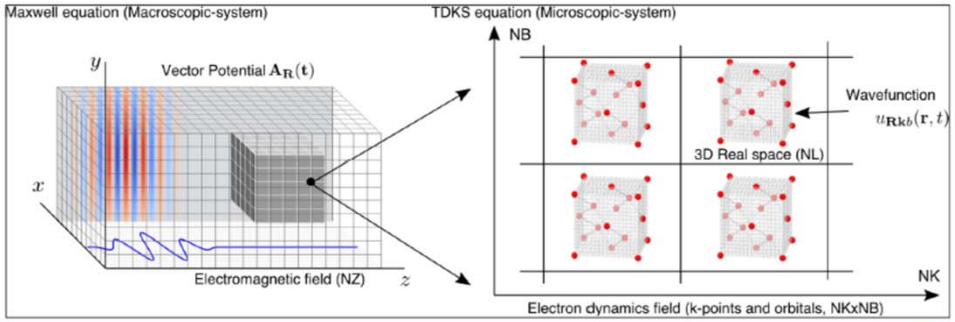
Performance study using Post-K simulator

University of Tsukuba
Center for Computational Sciences

SALMON: Electron Dynamics Simulator

- Main developers: Center for Computational Sciences, U. Tsukuba
- Coupled Maxwell-TDDFT multi-scale simulation
- Open-source application (99% Fortran, 1% C), Apache 2.0 license
  - https://salmon-tddft.jp/

SALMON



- We have been developing a cycle-level simulator for the post-K processor using gem5.
- Collaboration with U. Tsukuba
- Kernel evaluation using single core

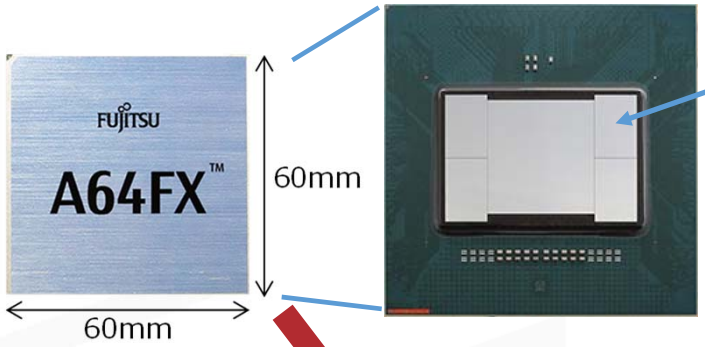
	Post-K Simulator	KNL
Execution time [msec]	4.2	5.5
Number of L1D misses	29569	–
L1D miss rate	1.19%	–
Number of L2 misses	20	
L2 miss rate	0.01%	

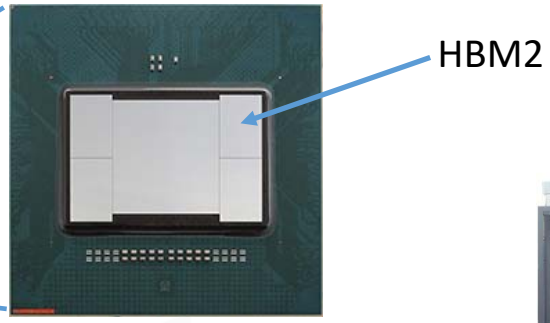
3.4 msec by further optimization

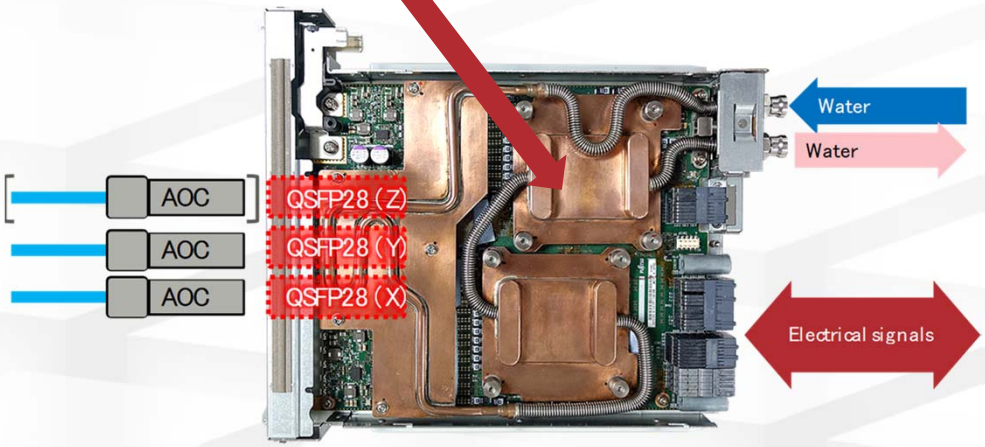
- 1.3 times faster than KNL per core
- With further optimization (inst. scheduling) exec time reduced to 3.4 msec (1.6 times faster)
- This is the evaluation on L1. OpenMP Multicore execution will be much faster due to HBM memory

Fugaku prototype board and rack

- **"Fujitsu Completes Post-K Supercomputer CPU Prototype, Begins Functionality Trials", HPCwire June 21, 2018**



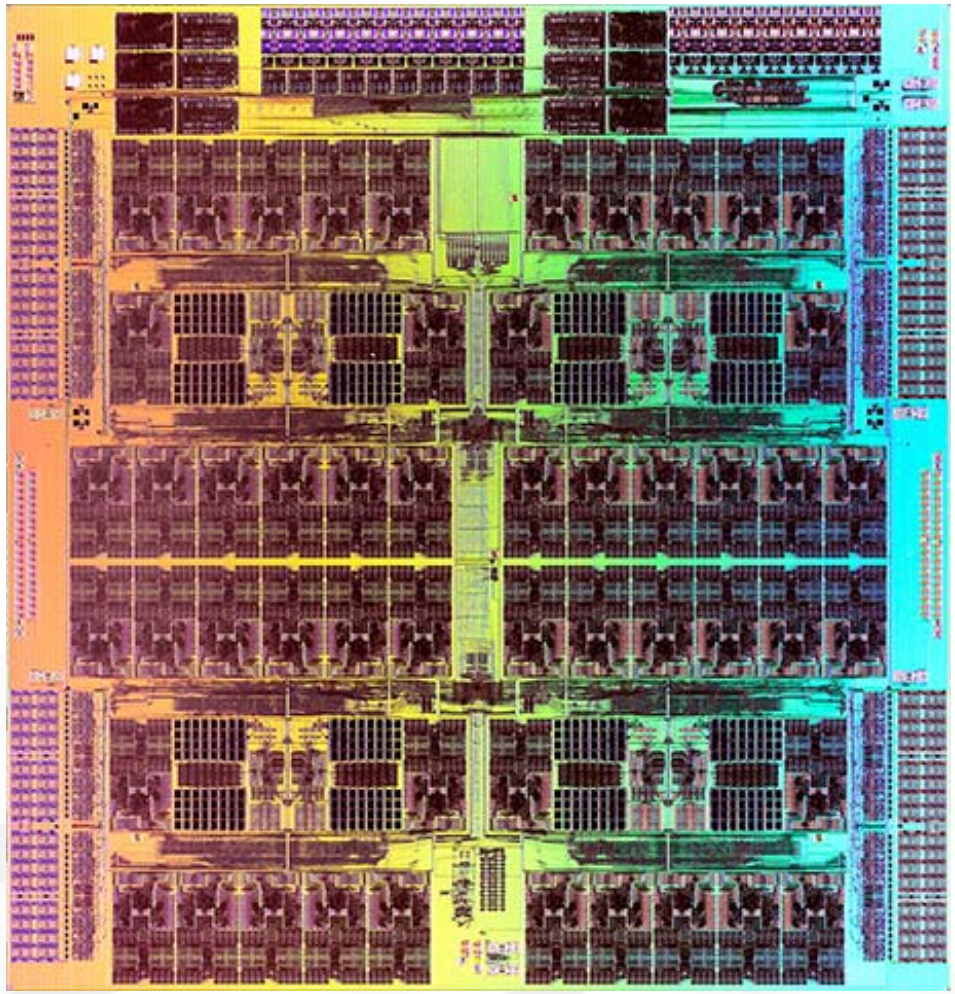




2 CPU / CMU

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

CPU-Die



画像：富士通提供

Advances from the K computer

	K computer	Fugaku	ratio
# core	8	48	
Si tech. (nm)	45	7	
Core perf. (GFLOPS)	16	> 56	3.5
Chip(node) perf. (TFLOPS)	0.128	>2.7	21
Memory BW (GB/s)	64	1024	
B/F (Bytes/FLOP)	0.5	0.4	
#node / rack	96	384	4
Rack perf. (TFLOPS)	12.3	1036.8	84
#node/system	82,944	> 150,000	
System perf.(DP PFLOPS)	10.6	> 405	38

Si Tech

SVE

CMG&Si Tech

HBM

More than **7.5 M General-purpose cores**!

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

Fugaku CPU New Innovations: Summary

1. **Ultra high bandwidth using on-package memory & matching CPU core**
   - Recent studies show that majority of apps are memory bound, some compute bound but can use lower precision e.g. FP16
   - Comparison w/mainstream CPU: much faster FPU, almost order magnitude faster memory BW, and ultra high performance accordingly
   - Memory controller to sustain massive on package memory (OPM) BW: difficult for coherent memory CPU, first CPU in the world to support OPM
2. **Very Green e.g. extreme power efficiency**
   - Power optimized design, clock gating & power knob, efficient cooling
   - Power efficiency much better than CPUs, comparable to GPU systems
3. **Arm Global Ecosystem & SVE contribution**
   - Annual processor production: x86 3-400mil, ARM 21bil, (2~3 bil high end)
   - Rapid upbringing HPC&IDC Ecosystem (e.g. Cavium, HPE, Sandia, Bristol,…)
   - SVE(Scalable Vector Extension) -> Arm-Fujitsu co-design, future global std.
4. **High Performance on Society5.0 apps including AI**
   - Next gen AI/ML requires massive speedup => high perf chips + HPC massive scalability across chips
   - Fujitsu A64FX processor: support for AI/ML acceleration e.g. Int8/FP16+fast memory for GPU-class convolution, fast interconnect for massive scaling
   - Top performance in AI as well as other Society 5.0 apps