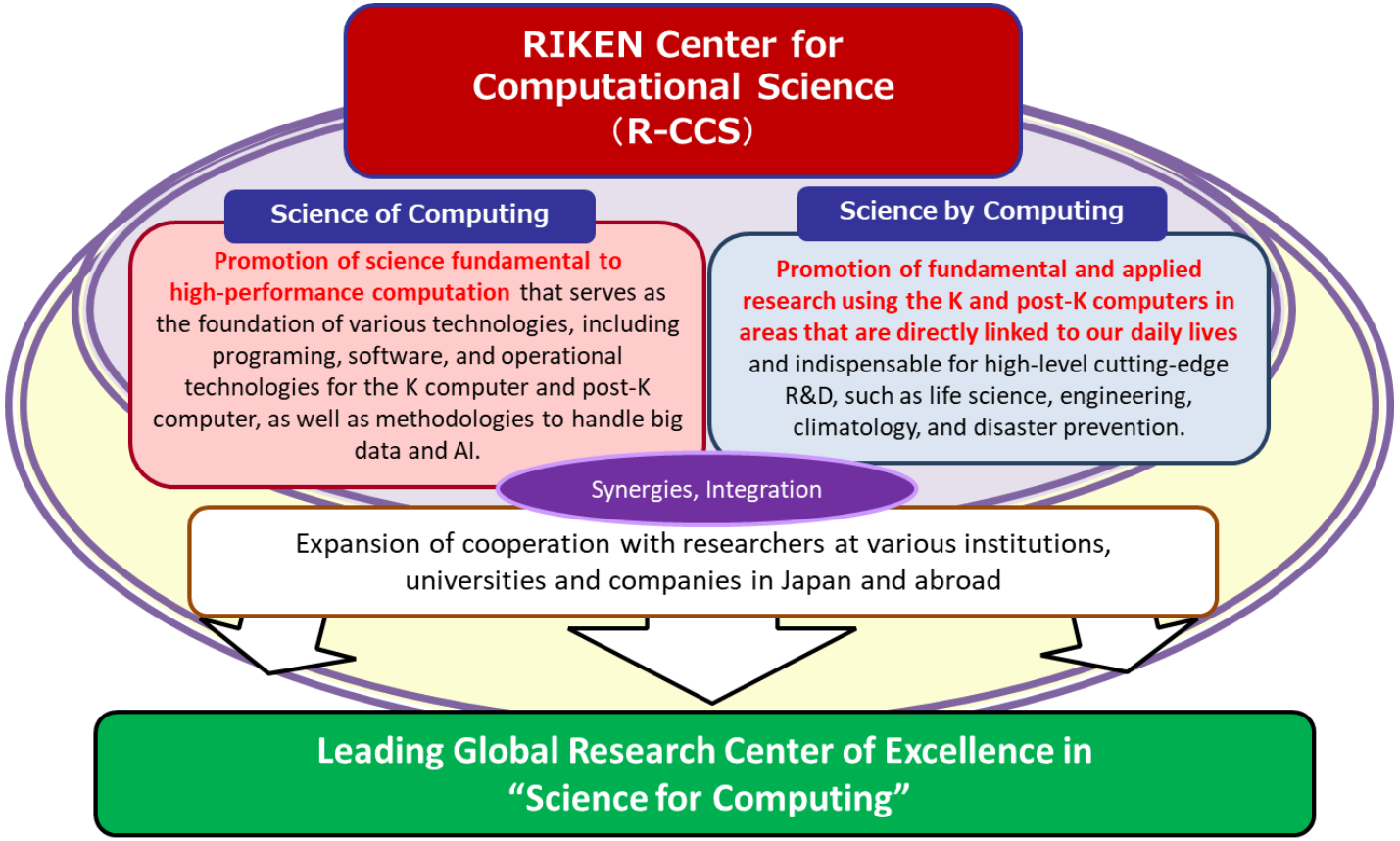
Fugaku: The first 'Exascale' Supercomputer with Real Application Performance as the Primary Target, and Towards the Future

- *Satoshi Matsuoka*
  - *Director, RIKEN Center for Computational Science & Professor, Tokyo Institute of Technology*

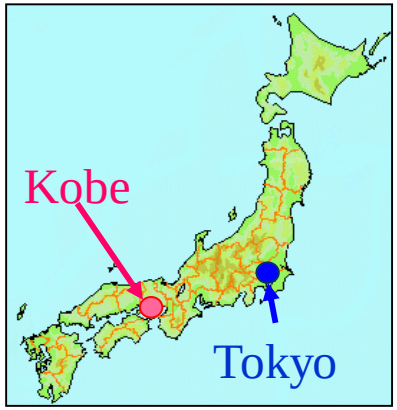
- **20190908 PPAM 2019 Keynote, Bialystok, Poland**

RIKEN R-CCS

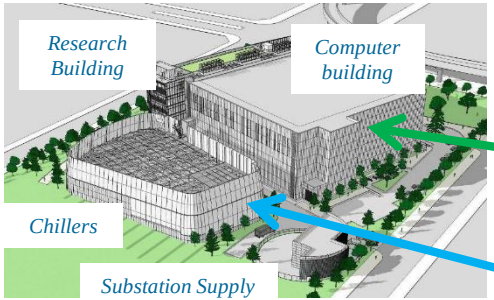
Riken R-CCS: Leadership HPC Research Center
“Science of Computing by Computing for Computing”

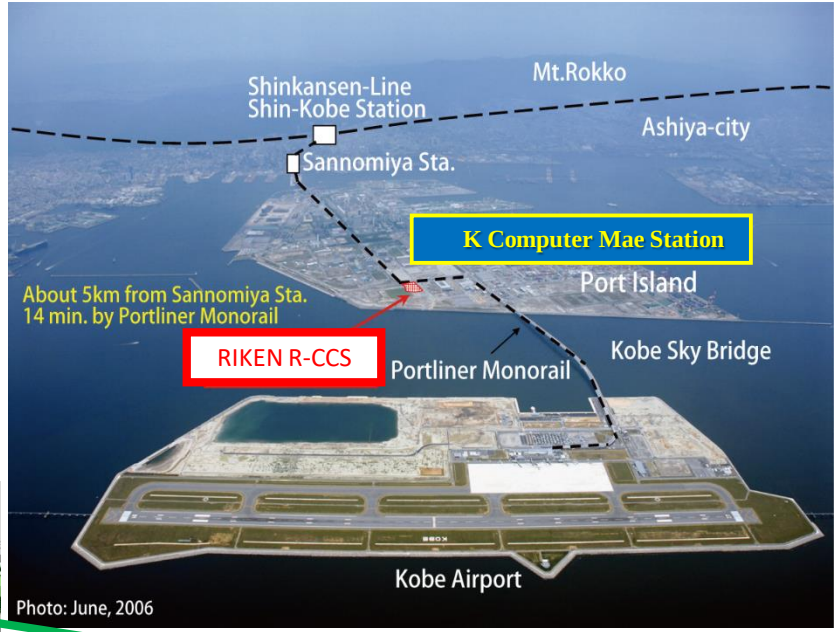


R-CCS with K Computer



423 km (263 miles) west of Tokyo





Computer room 50 m x 60 m = 3,000 m^2
Electric power up to 15 MW
Water cooling system

Gas-turbine co-generation 5 MW x 2

K computer (decommissioned Aug. 16, 2019)

Specifications

- Massively parallel, general purpose supercomputer
- No. of nodes: 88,128
- Peak speed: 11.28 Petaflops
- Memory: 1.27 PB
- Network: 6-dim mesh-torus (Tofu)

Top 500 ranking

LINPACK measures the speed and efficiency of linear equation calculations.
Real applications require more complex computations.

- No.1 in Jun. & Nov. 2011
- No.20 in Jun 2019

First supercomputer in the world to retire as #1 in major rankings (Graph 500)

Graph 500 ranking

"Big Data" supercomputer ranking
Measures the ability of data-intensive

- No. 1 for 9 consecutive editions since 2015

HPCG ranking

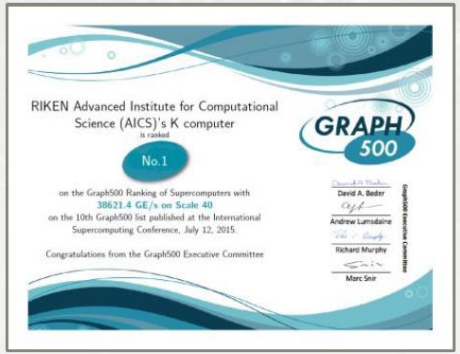
Measures the speed and efficiency of solving linear equation using HPCG
Better correlate to actual applications

- No. 1 in Nov. 2017, No. 3 since Jun 2018

ACM Gordon Bell Prize

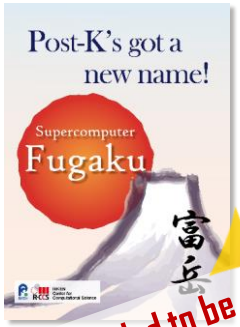
"Best HPC Application of the Year"

- Winner 2011 & 2012. several finalists



Japanese HPCI (High Peformance Computing Infrastructure)

- ***HPCI, High Performance Computing Infrastructure*** **established in 2012 as national HPC infrastructure, is a system connecting the Tier 0 flagship system and part of the Tier 1 major Universities and National Lab systems by high speed academic network (SINET-5).**
- World's top class ***HPCI*** computing resources with variety of system type are provided to ***open call***.



Scheduled to be operational around 2021

RIKEN R-CCS
K computer

Computing resources allocated at the Public Call in FY 2019

K computer

1.5 PFlops x Year

Other HPCI machines in total

14.2 PFlops x Year

Shared Storage

45 PB

SINET-5 (100 Gbps)

- : node
- : domestic line
- : international line

As of June 2019

K-Computer Shutdown Ceremony 30 Aug 2019

019年（令和元年）8月31日（土曜日）　12版　企業1　12

スパコン「京」歴史に幕
7年で計算速度一時トップ

理化学研究所は30日、神戸市の拠点でスーパーコンピューター「京（けい）」の電源を停止してシャットダウンした。2012年に本格稼働し、一時は計算速度の性能ランキングで世界首位に輝いた京だが、各国でスパコンの開発競争と性能の向上が急速に進むなかで、およそ7年の歴史に幕を下ろした。

30日午後に京が設置されている理研の計算科学研究センターでセレモニーがあり、理研や官公庁、企業などの関係者ら約200人が出席した。理研の松本紘理事長は「たいへん使いやすい良いコンピューターだった。一抹のさみしさはあるが（後継のスパコン）富岳（ふがく）に精神を引き継ぎたい」と述べた。

京の開発元である富士通の時田隆仁社長も「様々な分野で多くの成果を創出できたことを誇りに思う。京は今も世界最高クラスの技術を誇っており、技術的アプローチが正しかった」と話した。

その後、多数の装置が並ぶ京の計算機室で、代表4人の手で電源のスイッチが順に落とされ、出席者が見守るなか、午後3時25分ごろに計算機の発する機械音が鳴りやみ、機能を停止した。

京は近く解体が始まり、同じ場所に富岳が設置される。京の最大100倍以上という計算速度を目指し、21年ごろに運用が始まる予定だ。

「京」の前でメッセージを掲げる高校生（30日、神戸市）

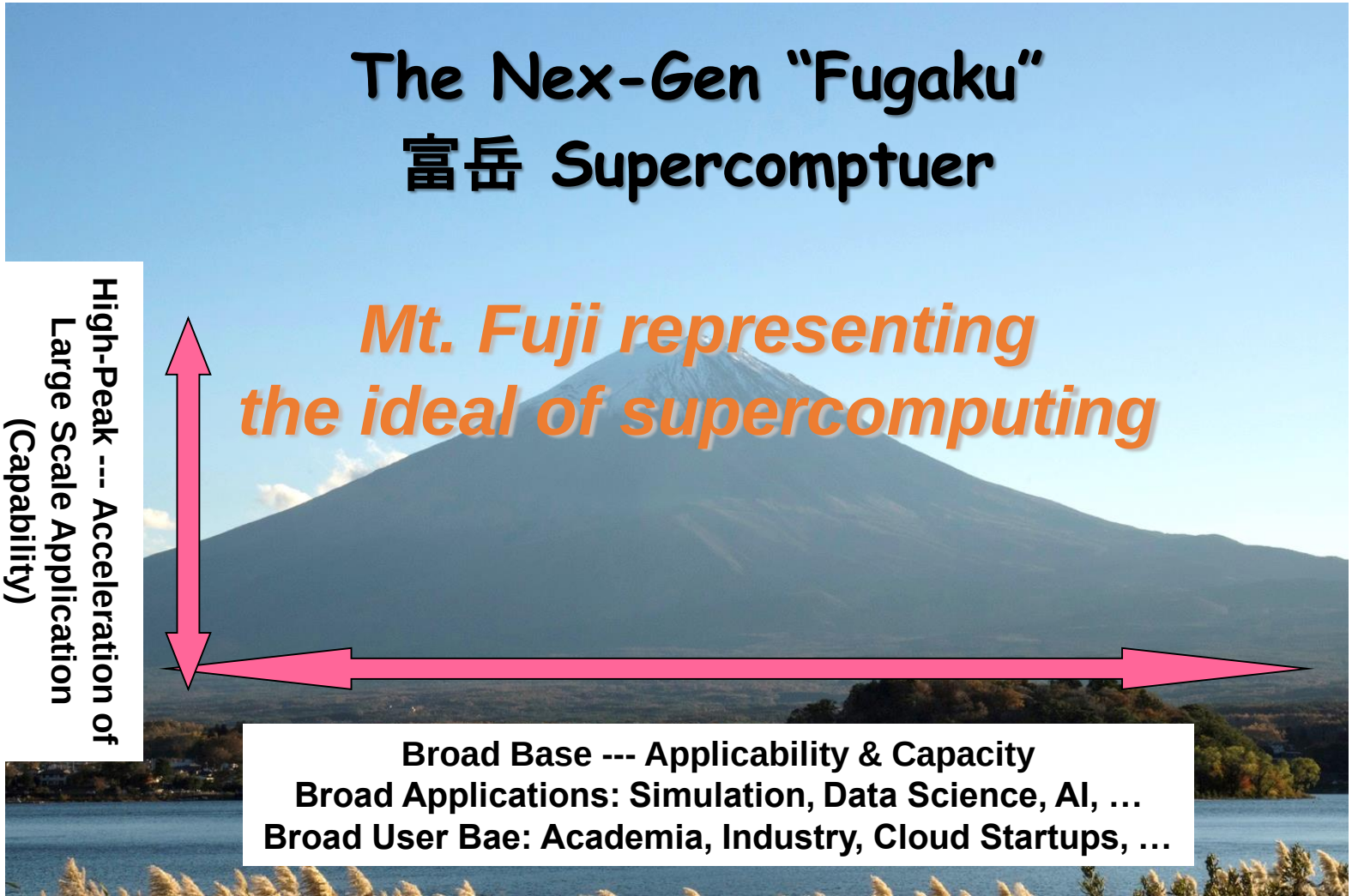
ソニーと資
オリンパスにも

ソニーは30日、保有するオリンパス株すべてを売却したと発表した。ソニーが投資ファンドの米サード・ポイントにオリンパス株の売却を求められていた一方、オリンパスにも米ファンドの影がみえる。両社は共同出資会社を通じた医療機器事業は継続する。既に発売した2製品に続く道筋を両社で示せるかが問われている。

オリンパスが実施した自社株買いに応募し、ソニーは約800億円で全株を売却した。500億円で取得したオリンパス株を、2回に分けて計1500億円で売却した格

ウィッシュの画面はあえて雑多な雰囲気にしている

The Nex-Gen "Fugaku"
富岳 Supercomptuer



Arm64fx & Fugaku 富岳 /Post-K are:

- **Fujitsu-Riken design A64fx ARM v8.2 (SVE), 48/52 core CPU**
  - ***HPC Optimized:*** Extremely high package high memory BW (1TByte/s), on-die Tofu-D network BW (~400Gbps), high SVE FLOPS (~3Teraflops), various AI support (FP16, INT8, etc.)
  - Gen purpose CPU – Linux, Windows (Word), other SCs/Clouds
  - Extremely power efficient – > **10x power/perf efficiency for CFD benchmark** over current mainstream x86 CPU
- **Largest and fastest supercomputer to be ever built circa 2020**
  - > 150,000 nodes, superseding LLNL Sequoia
  - > 150 PetaByte/s memory BW
  - Tofu-D 6D Torus NW, 60 Petabps injection BW (10x global IDC traffic)
  - 25~30PB NVMe L1 storage
  - many endpoint 100Gbps I/O network into Lustre
  - The first 'exascale' machine (not exa64bitflops but in apps perf.)

Brief History of R-CCS towards Fugaku

January 2006
Next Generation Supercomputer Project (K Computer) start

July 2010
RIKEN AICS established
August 2010
HPCI Project start
September 2010
K computer installation starts
First meeting of SDHPC (Post-K)

June 2011
#1 on Top 500
November 2011
#1 on Top 500 > 10 Petaflops
ACM Gordon Bell Award
End of FY 2011 (March 2012)
SDHPC Whitepaper

April 2012
Post-K Feasibility Study start
3 Arch Teams and 1 Apps Team
June 2012
K computer construction complete
September 2012
K computer production start
November 2012
ACM Gordon bell Award

End of FY2013 (Mar 2014)
Post-K Feasibility Study Reports

April 2014
Post-K project start
June 2014
#1 on Graph 500

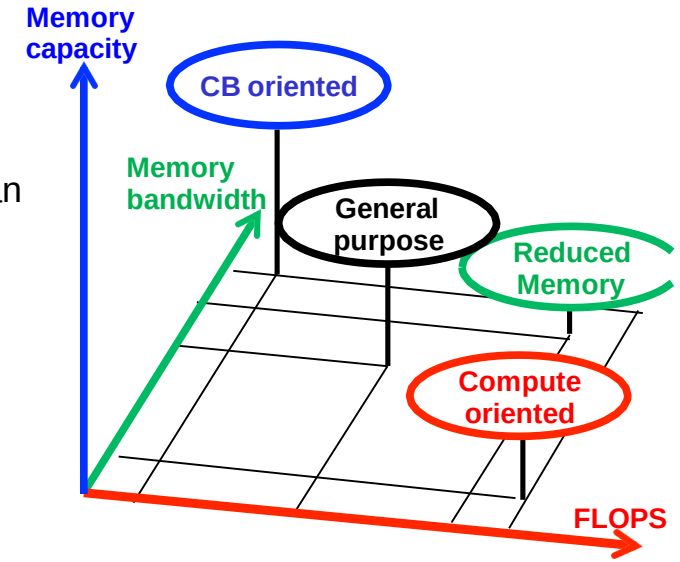
April 2018
AICS renamed to RIKEN R-CCS. Satoshi Matsuoka becomes new Director
Aug 2018
Arm A64fx announce at Hotchips
Oct 2018
NEDO 100x processor project start
Nov 2018
Post-K Manufacturing approval by Prime Minister's CSTI Committee

March 2019
Post-K Manufacturing start
May 2019
Post-K named "Supercomputer Fugaku"
July 2019
Post-Moore Whitepaper start
Aug 2019
K Computer shutdown
Dec 2019
Fugaku installation start (planned)

SDHPC (2011-2012) Candidate of ExaScale Architecture

https://www.exascale.org/mediawiki/images/a/aa/Talk-3-kondo.pdf

- Four types of architectures are considered
  - General Purpose (GP)
    - Ordinary CPU-based MPPs
    - e.g.) K-Computer, GPU, Blue Gene, x86-based PC-clusters
  - Capacity-Bandwidth oriented (CB)
    - With expensive memory-I/F rather than computing capability
    - e.g.) Vector machines
  - Reduced Memory (RM)
    - With embedded (main) memory
    - e.g.) SoC, MD-GRAPE4, Anton
  - Compute Oriented (CO)
    - Many processing units
    - e.g.) ClearSpeed, GRAPE-DR



SDHPC (2011-2012) Performance Projection

- Performance projection for an HPC system in 2018
  - Achieved through continuous technology development
  - Constraints: 20 – 30MW electricity & 2000sqm space

Node Performance

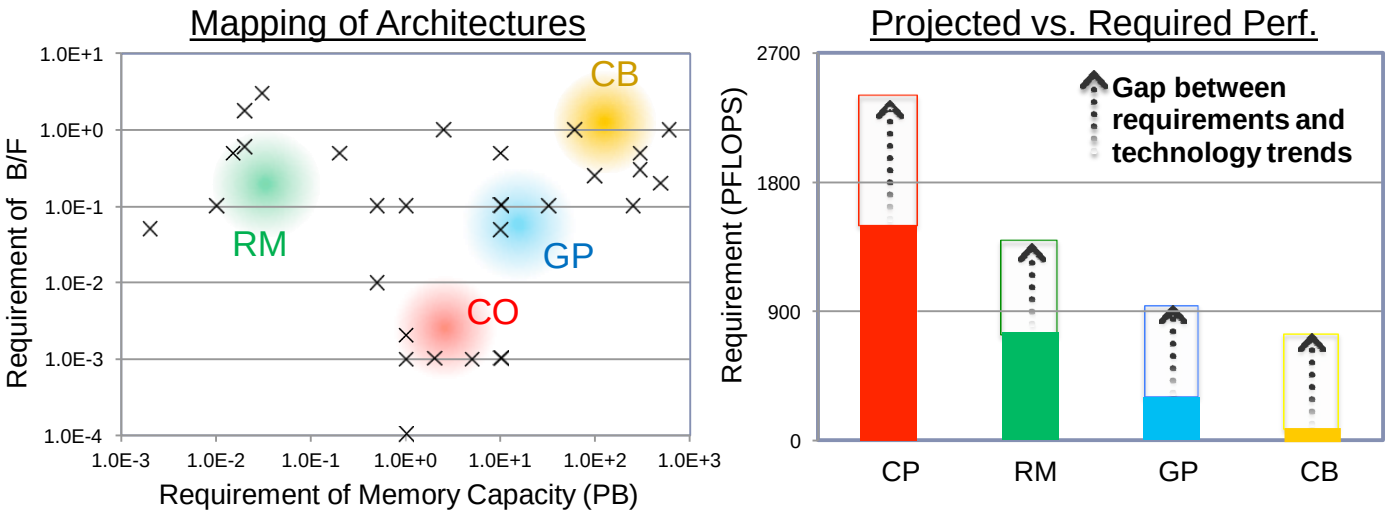
	Total CPU Performance (PetaFLOPS)	Total Memory Bandwidth (PetaByte/s)	Total Memory Capacity (PetaByte)	Byte / Flop
General Purpose	200~400	20~40	20~40	0.1
Capacity-BW Oriented	50~100	50~100	50~100	1.0
Reduced Memory	500~1000	250~500	0.1~0.2	0.5
Compute Oriented	1000~2000	5~10	5~10	0.005

Ne

	Injection	P-to-P	Bisection	Latency	Latency	idth
High-radix (Dragonfly)	32 GB/s	32 GB/s	2.0 PB/s	200 ns	1000 ns	
				1 EB	10TB/s	
				100 times larger than main memory	For saving all data in memory to disks within 1000-sec.	

SDHPC (2011-2012) Gap Between Requirement and Technology Trends

- Mapping four architectures onto science requirement
- Projected performance vs. science requirement
  - Big gap between projected and required performance



Needs national research project for science-driven HPC systems

Post-K Feasibility Study (2012-2013)

- **3 Architecture Teams, from identified architectural types in the SDHPC report**
  - General Purpose --- balanced
  - Compute Intensive --- high flops and/or low memory capacity & high memory BW
  - Large Memory Capacity --- also w/high memory BW
- **The A64fx processor satisfied multiple roles - basically balanced but also compute intensive**

- **Application Team (Tomita, Matsuoka)**
  - Put all the K-Computer applications stakeholders into one room
  - Templated reporting of science impact possible on exascale machines and their computational algorithms / requirements
  - 600 page report (English summary available)

Post-K Application Feasibility Study 2012-2013

https://hpci-aplfs.r-ccs.riken.jp/document/roadmap/roadmap_e_1405.pdf

Computational Science Roadmap -Overview-

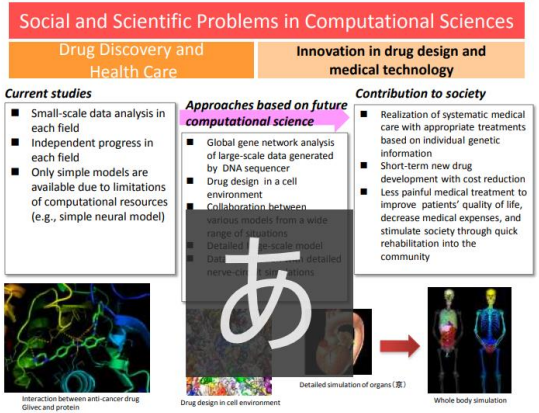
Social Contributions and Scientific Outcomes Aimed for by Innovations through Large-Scale Parallel Computing

May, 2014

Feasibility Study on Future HPC Infrastructures

(Application Working Group)

mechanisms, such as blood clot formation in the heart or brain infarctions, and will be effective in improving patients' Quality of Life (QOL) through the development of minimally invasive treatments, which only pose a slight burden to the patient, and of the medical devices required for these treatments. It will further be effective in revitalizing society through patients' early re-entry into the community and in reducing costs of medical treatment.



The supercomputer's vast computational power will undoubtedly greatly contribute to the development of various aspects in the field of life science, such as detailed neural and cellular simulations, simulations over extended periods of time and space, and almost real-time assimilation[4] of those data. Eventually it could form an important scientific basis for innovative drug design and medical technologies.

The table below lists the computational performance required in the future for the respective areas of drug discovery and healthcare.

[4] One of the methods to merge different observational and experimental data into a numerical model at a high degree.

Subject	Performance (PFLOPS)	Memory bandwidth (PB/s)	Memory size per case (PB)	Storage size per case (PB)	Elapse Time /Case (hour)	Number of Cases	Total operation count (EFLOP)	Summary and numerical method	Problem size	Notes
Personal Genome Analysis	0.0054	0.0001	1.6	0.1	0.7	200000	2700	Sequence matching	Cancer Genome Analysis: Short read mapping and mutation identification of 200,000 people's genome	1 case = 1 person Integer operations are dominant. "Total operation count" = total instruction count (Total FLOP = 46 EFLO
Gene Network Analysis	25	89	0.08	0.016	0.34	26000	780000	Baysian network estimation and L1-regularization	40,000 transcripts x 26,000 data sets consisting of 2,800,000 arrays	
MD and Free-energy calculation for drug design and so on	1000	400	0.0001		0.0012	1000000	4300000	Molecular dynamics simulation with all-atom model	Number of Cases: 100,000 ligands X 10 target proteins	B/F=0.4. Supposed to run 100-1000 cases simultaneously. Memory size per case is estimated for a 100 nod run.
MD simulations under cellular environments or MD simulations of Virus	490	49	0.2	1.2	48	10	50000	Molecular dynamics simulations with all-atom grain-grained model	100,000,000 particles	B/F=0.1
Simulations of cellular signaling pathways	42	100	10	10	240			opic lattice reaction-diffusion simulation	1,000 to 10,000 cells	integer operations
Precise Structure-Based Drug Design	0.83	0.14	1	0.001	1	100	200	quantum chemical calculations on the interactions between proteins and drugs	proteins (500 residues) + ligands in solution	1TB/s IO speed require to dump 1TB dataset pe second
Design of Biological Devices	1.1	0.19	1	0.001	1	100	400	Spectroscopic analyses of proteins (200-500 residues)	more than 100,000 orbitals	1TB/s IO speed require to dump 1TB dataset pe second
Multi-scale simulation of a blood clot	400	64	1	1	170	10	2500000	Semi-implicit FDM simulation of fluid-structure interaction with chemical factors	Length:100mm, D:100um, Calculation Time:10s, Grid size:0.1um, Velocity:10^-2m/s, Delta T:1us	
High Intensity Focused Ultrasound	380	460	54	64	240	10	3300000	Explicit FDM simulation of sound wave and heat transfer	Calculation Area:400mm^3, Grid: 225x10^12, Steps: 1459200, FLOP/grid/step: 1000	
Simulations of Brain and Neural Systems	* 6.9	* 7.6	* 56	* 3600	0.28	100	700	Single compartment model	100 billion neuons, 10000 synapses/neuron, 10^5steps	
Data assimulation of whole insect brain via communication between a phisological experiment and a simulation, Parameter estimator in insect brain simulation	* 71	* 60	* 0.2	* 20	28	20	140000	Multi-compartment HH model with local Crank-Nicolson method, evolutionary algorithm	1000 neurons, 10^6 genes, 100 generations	Supposing 100 MB/s communication to exter environment will be required

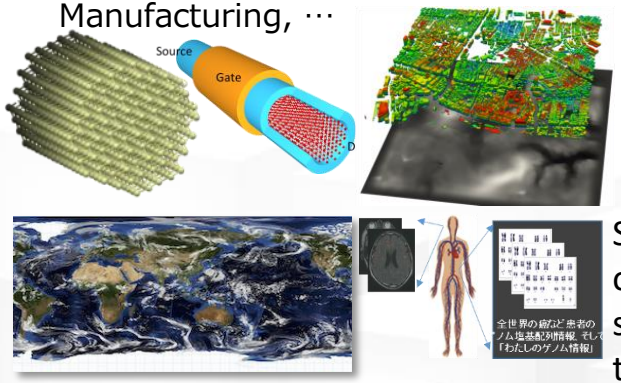
*Figures marked with a * are still under examination. The website will show more accurate figures as they become available.*

Co-Design Activities in Fugaku

Multiple Activities since 2011

Science by Computing

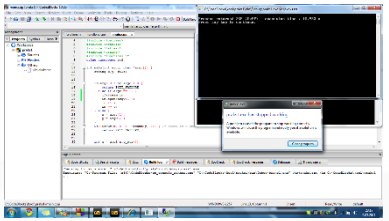
・9 Priority App Areas: High Concern to General Public: Medical/Pharma, Environment/Disaster, Energy, Manufacturing, …



Select representatives fr om 100s of applications signifying various compu tational characteristics

Science of Computing

Design systems with param eters that consider various application characteristics



- **Extremely tight collabrations between the Co-Design apps centers, Riken, and Fujitsu, etc.**
- **Chose 9 representative apps as “target application” scenario**
- **Achieve up to x100 speedup c.f. K-Computer**
- **Also ease-of-programming, broad SW ecosystem, very low power, …**

Research Subjects of the Post-K Computer

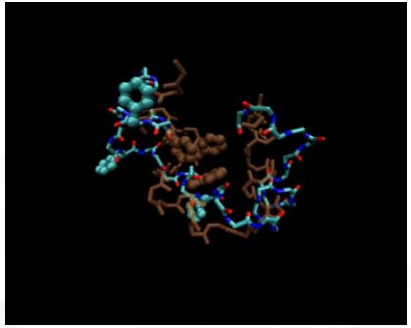
The post K computer will expand the fields pioneered by the K computer, and also challenge new areas.

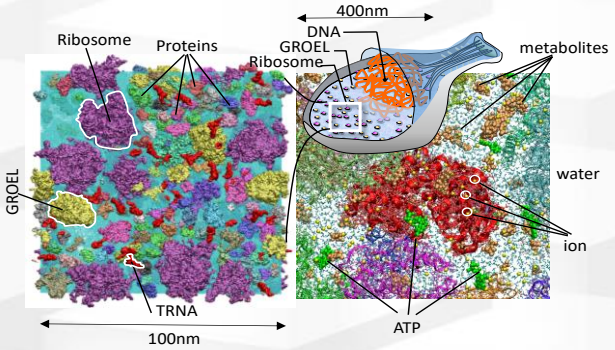
Genesis MD: proteins in a cell environment

Protein simulation before K

- Simulation of a protein in isolation

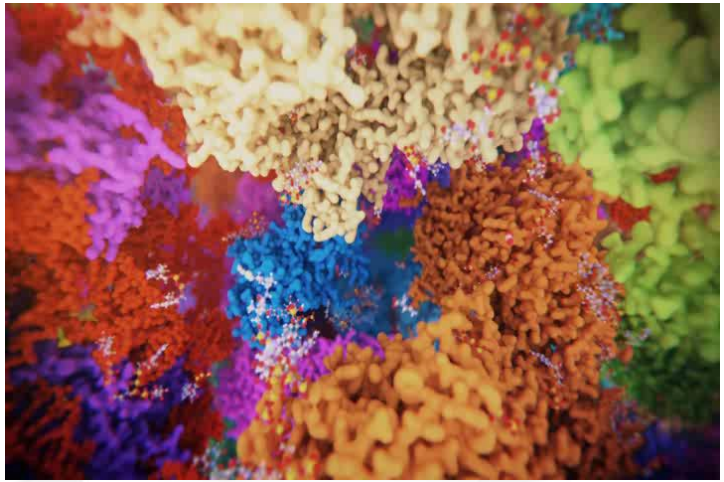
Folding simulation of Villin, a small protein with 36 amino acids





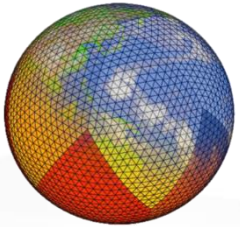
Protein simulation with K

- all atom simulation of a cell interior
- cytoplasm of Mycoplasma genitalium

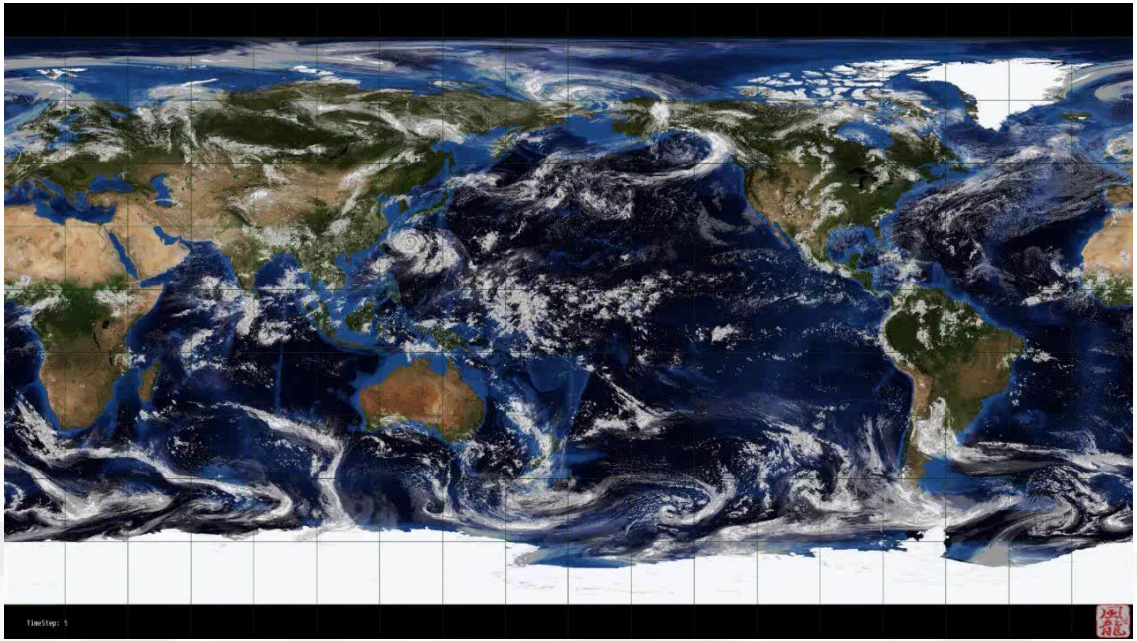


NICAM: Global Climate Simulation

- Global cloud resolving model **with 0.87 km-mesh** which allows resolution of cumulus clouds
- Month-long forecasts of Madden-Julian oscillations in the tropics is realized.



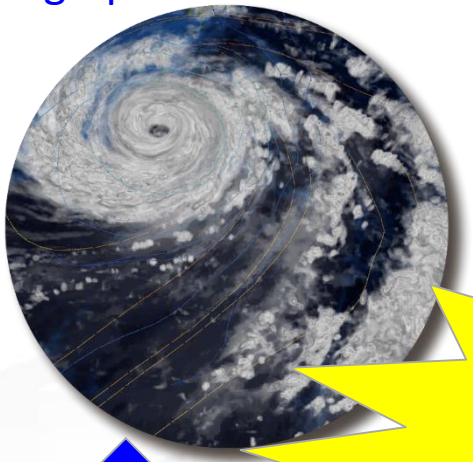
Global cloud resolving model



Miyamoto et al (2013) , Geophys. Res. Lett., 40, 4922–4926, doi:10.1002/grl.50944.

“Big Data Assimilation” NICAM+LETKF

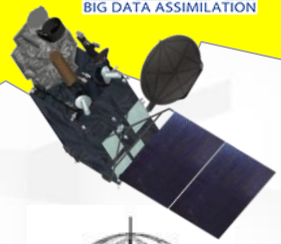
High-precision Simulations



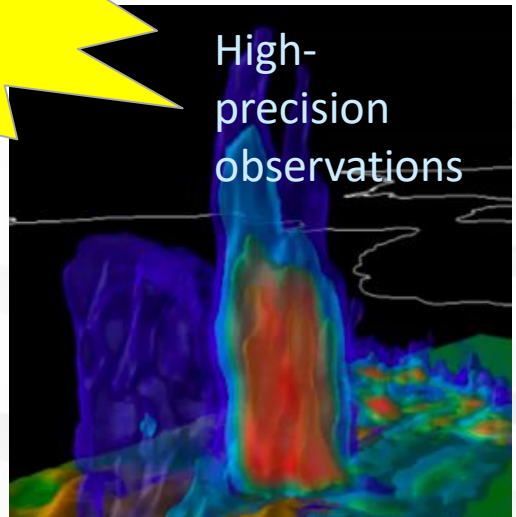
国立研究開発法人 科学技術振興機構 Japan Science and Technology Agency CREST

Future-generation technologies available 10 years in advance

BDA BIG DATA ASSIMILATION



Mutual feedback



High-precision observations

Co-design from Apps to Architecture

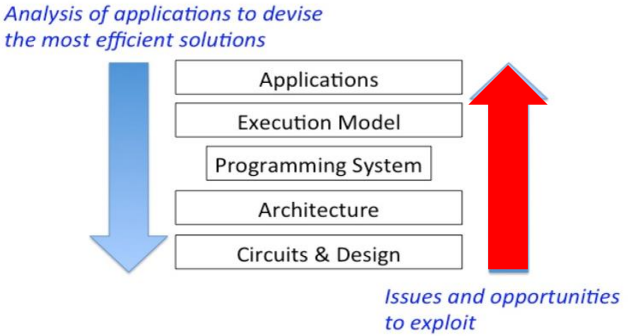
- **Architectural Parameters to be determined**
  - #SIMD, SIMD length, #core, #NUMA node, O3 resources, specialized hardware
  - cache (size and bandwidth), memory technologies
  - Chip die-size, power consumption
  - Interconnect
- **We have selected a set of target applications**
- **Performance estimation tool**
  - Performance projection using Fujitsu FX100 execution profile to a set of arch. parameters.
- **Co-design Methodology (at early design phase)**
  1. Setting set of system parameters
  2. Tuning target applications under the system parameters
  3. Evaluating execution time using prediction tools
  4. Identifying hardware bottlenecks and changing the set of system parameters

Target applications representatives of almost all our applications in terms of computational methods and communication patterns in order to design architectural features.

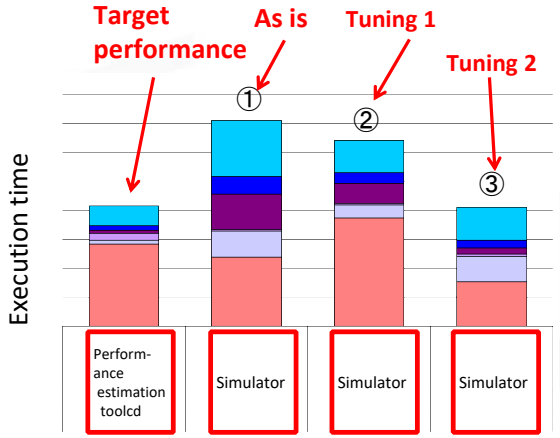
	Target Application	
	Program	**Brief description**
①	GENESIS	MD for proteins
②	Genomon	Genome processing (Genome alignment)
③	GAMERA	Earthquake simulator (FEM in unstructured & structured grid)
④	NICAM+LETK	Weather prediction system using Big data (structured grid stencil & ensemble Kalman filter)
⑤	NTChem	molecular electronic (structure calculation)
⑥	FFB	Large Eddy Simulation (unstructured grid)
⑦	RSDFT	an ab-initio program (density functional theory)
⑧	Adventure	Computational Mechanics System for Large Scale Analysis and Design (unstructured grid)
⑨	CCS-QCD	Lattice QCD simulation (structured grid Monte Carlo)

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

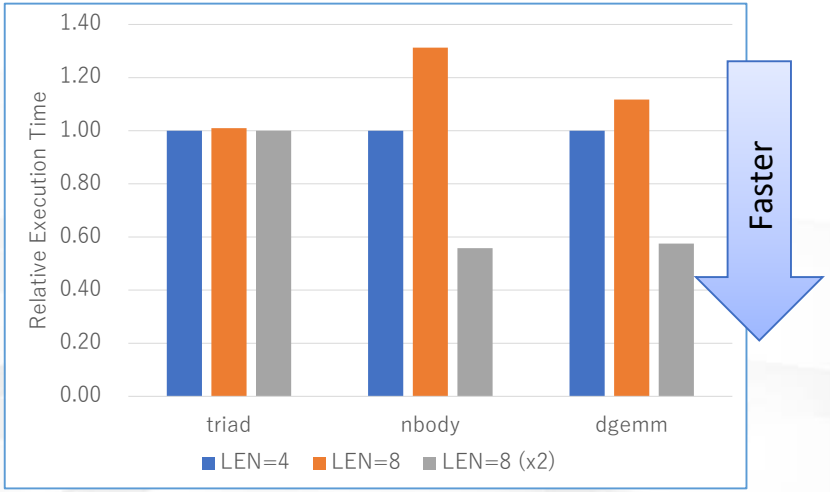


ARM for HPC - Co-design Opportunities

- **ARM SVE Vector Length Agnostic feature is very interesting, since we can examine vector performance using the same binary.**
- **We have investigated how to improve the performance of SVE keeping hardware-resource the same. (in "Rev-A" paper)**
  - ex. "512 bits SVE x 2 pipes" vs. "1024 bits SVE x 1 pipe"
  - Evaluation of **Performance and Power** (in "coolchips" paper) by using our gem-5 simulator (with "white" parameter) and ARM compiler.
  - Conclusion: Wide vector size over FPU element size will improve performance if there are enough rename registers and the utilization of FPU has room for improvement.

Note that these researches are not relevant to "post-K" architecture.

- Y. Kodama, T. Oajima and M. Sato. "Preliminary Performance Evaluation of Application Kernels Using ARM SVE with Multiple Vector Lengths", In Re-Emergence of Vector Architectures Workshop (Rev-A) in 2017 IEEE International Conference on Cluster Computing, pp. 677-684, Sep. 2017.
- T. Odajima, Y. Kodama and M. Sato, "Power Performance Analysis of ARM Scalable Vector Extension", In IEEE Symposium on Low-Power and High-Speed Chips and Systems (COOL Chips 21), Apr. 2018



Arm64fx & Fugaku (富岳) are:

- **Fujitsu-Riken design A64fx ARM v8.2 (SVE), 48/52 core CPU**
  - ***HPC Optimized:*** Extremely high package high memory BW (1TByte/s), on-die Tofu-D network BW (~400Gbps), high SVE FLOPS (~3Teraflops), various AI support (FP16, INT8, etc.)
  - Gen purpose CPU – Linux, Windows (Word), other SCs/Clouds
  - Extremely power efficient – > **10x power/perf efficiency for CFD benchmark** over current mainstream x86 CPU
- **Largest and fastest supercomputer to be ever built circa 2020**
  - > 150,000 nodes, superseding LLNL Sequoia
  - > 150 PetaByte/s memory BW
  - Tofu-D 6D Torus NW, 60 Petabps injection BW (10x global IDC traffic)
  - 25~30PB NVMe L1 storage
  - ~10,000 endpoint 100Gbps I/O network into Lustre
  - The first 'exascale' machine (not exa64bitflops but in apps perf.)

Fugaku: The Game Changer

1. **Heritage of the K-Computer, HP in simulation via extensive Co-Design**
   - High performance: up to x100 performance of K in real applications
   - Retain BYTES/FLOP of K (0.4~0.5) for real application performance
   - Simultaneous high performance and ease-of-programming

2. New Technology Innovations of Fugaku

- **High Performance, esp. via high memory BW**
  Performance boost by “factors” c.f. mainstream CPUs in many HPC & Society5.0 apps via BW & Vector acceleration
- **Very Green e.g. extreme power efficiency**
  Ultra Power efficient design & various power control knobs
- **Arm Global Ecosystem & SVE contribution**
  Top CPU in ARM Ecosystem of 21 billion chips/year, SVE co-design and world’s first implementation by Fujitsu
- **High Perf. on Society5.0 apps incl. AI**
  Architectural features for high perf on Society 5.0 apps based on Big Data, AI/ML, CAE/EDA, Blockchain security, etc.

Global leadership not just in the machine & apps, but as cutting edge IT

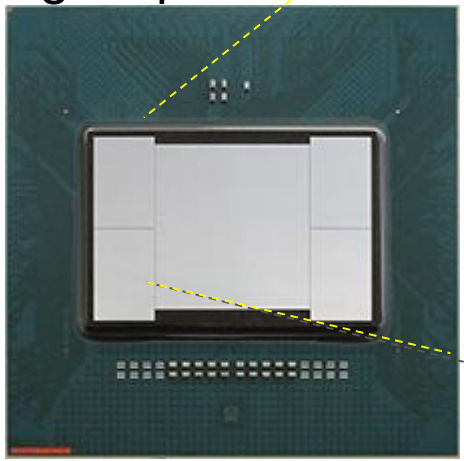
ARM: Massive ecosystem from embedded to HPC

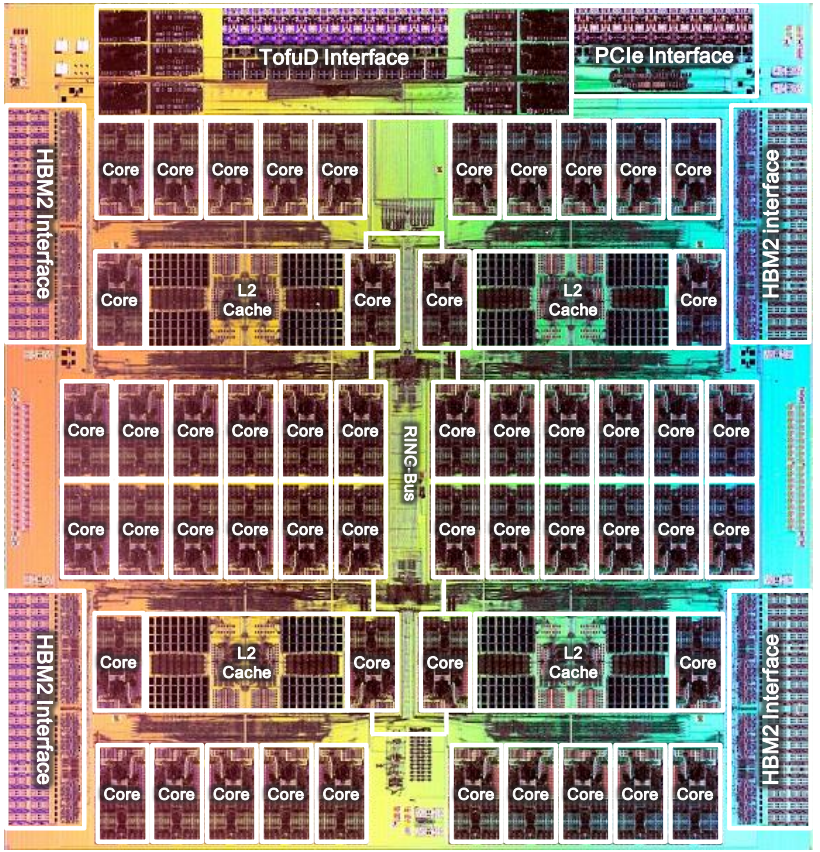
Technology not just limited to Fugaku, but into societal IT infrastructures e.g. Clouds

A64FX Leading-edge Si-technology

FUJITSU

- TSMC 7nm FinFET & CoWoS
  - Broadcom SerDes, HBM I/O, and SRAMs
  - 87.86 billion transistors
  - 594 signal pins





“Fugaku” Chronology

(Disclaimer: below includes speculative schedules and subject to change)

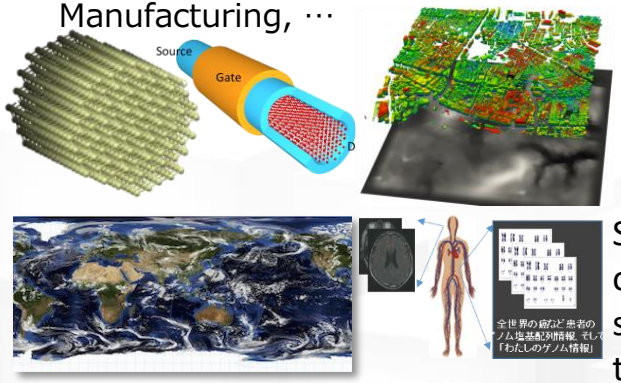
- May 2018 A0 Chip came out, almost bug free
- 1Q2019 B0 Chip on hand, bug free, exceeded perf. target
- Mar 2019 “Fugaku” manufacturing budget approval by the Diet, actual manufacturing contract signed **(now w/Society 5.0 AI mission also)**
- Aug 2019 End of K-Computer operations
- 4Q2019 “Fugaku” installation starts
- 1H2020 “Fugaku” preproduction operation starts
- 1~2Q2021 “Fugaku” production operation starts (hopefully)
- And of course we move on…

Co-Design Activities in Fugaku

Multiple Activities since 2011

Science by Computing

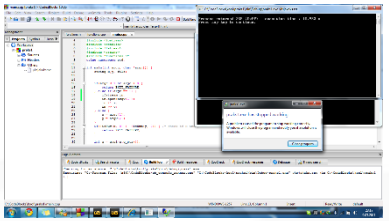
・9 Priority App Areas: High Concern to General Public: Medical/Pharma, Environment/Disaster, Energy, Manufacturing, …



Select representatives fr
om 100s of applications
signifying various compu
tational characteristics

Science of Computing

Design systems with param
eters that consider various
application characteristics



- **Extremely tight collabrations between the Co-Design apps centers, Riken, and Fujitsu, etc.**
- **Chose 9 representative apps as “target application” scenario**
- **Achieve up to x100 speedup c.f. K-Computer**
- **Also ease-of-programming, broad SW ecosystem, very low power, …**

Research Subjects of the Post-K Computer

The post K computer will expand the fields pioneered by the K computer, and also challenge new areas.

Co-design from Apps to Architecture

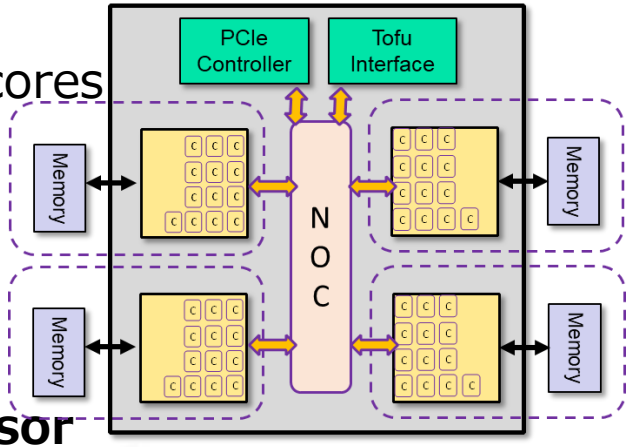
- **Architectural Parameters to be determined**
  - #SIMD, SIMD length, #core, #NUMA node, O3 resources, specialized hardware
  - cache (size and bandwidth), memory technologies
  - Chip die-size, power consumption
  - Interconnect
- **We have selected a set of target applications**
- **Performance estimation tool**
  - Performance projection using Fujitsu FX100 execution profile to a set of arch. parameters.
- **Co-design Methodology (at early design phase)**
  1. Setting set of system parameters
  2. Tuning target applications under the system parameters
  3. Evaluating execution time using prediction tools
  4. Identifying hardware bottlenecks and changing the set of system parameters

Target applications representatives of almost all our applications in terms of computational methods and communication patterns in order to design architectural features.

	Target Application	
	Program	**Brief description**
①	GENESIS	MD for proteins
②	Genomon	Genome processing (Genome alignment)
③	GAMERA	Earthquake simulator (FEM in unstructured & structured grid)
④	NICAM+LETK	Weather prediction system using Big data (structured grid stencil & ensemble Kalman filter)
⑤	NTChem	molecular electronic (structure calculation)
⑥	FFB	Large Eddy Simulation (unstructured grid)
⑦	RSDFT	an ab-initio program (density functional theory)
⑧	Adventure	Computational Mechanics System for Large Scale Analysis and Design (unstructured grid)
⑨	CCS-QCD	Lattice QCD simulation (structured grid Monte Carlo)

Fugaku's FUjitsu A64fx Processor is…

- **an Many-Core ARM CPU…**
  - 48 compute cores + 2 or 4 assistant (OS) cores
  - Brand new core design
  - Near Xeon-Class Integer performance core
  - ARM V8 --- 64bit ARM ecosystem
  - Tofu-D + PCIe 3 external connection



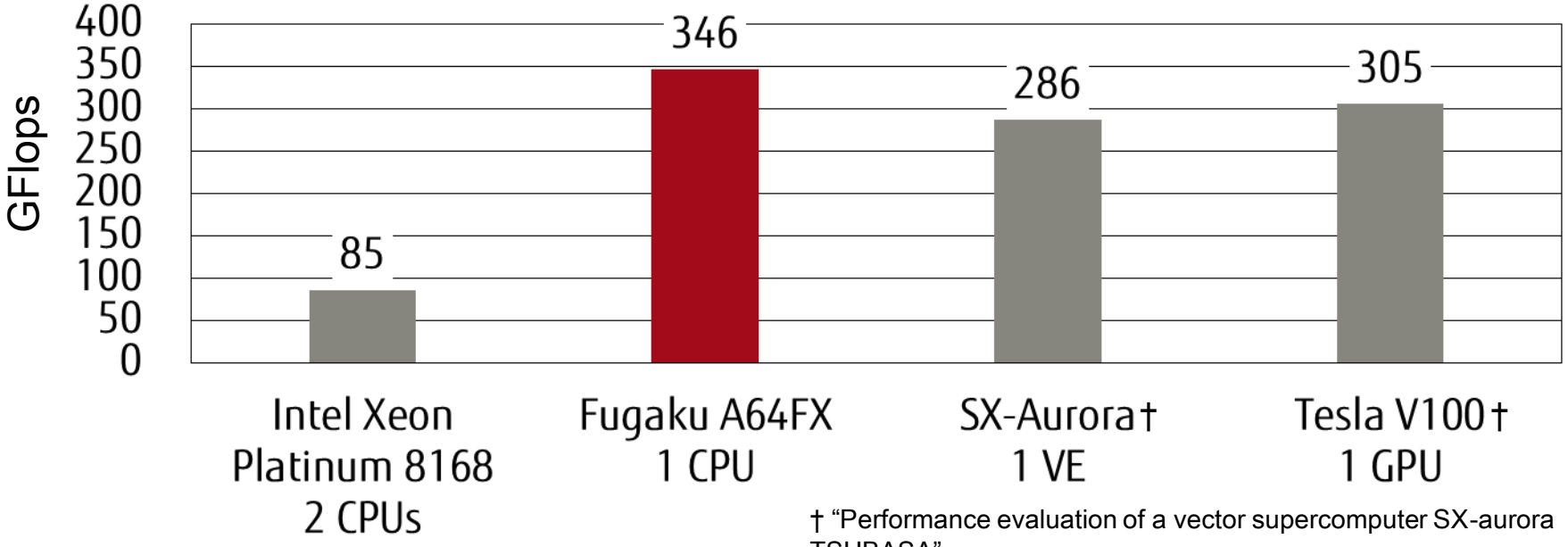
- **…but also an accelerated GPU-like processor**
  - SVE 512 bit x 2 vector extensions (ARM & Fujitsu)
    - Integer (1, 2, 4, 8 bytes) + Float (16, 32, 64 bytes)
  - Cache + scratchpad-like local memory (sector cache)
  - HBM2 on package memory – Massive Mem BW (Bytes/DPF ~0.4)
    - Streaming memory access, strided access, scatter/gather etc.
  - Intra-chip barrier synch. and other memory enhancing features

- **A64fx/SVE a great target for OpenACC (w/o data movement)**

“Fugaku” CPU Performance Evaluation (2/3)

FUJITSU

- Himeno Benchmark (Fortran90)
  - Stencil calculation to solve Poisson’s equation by Jacobi method

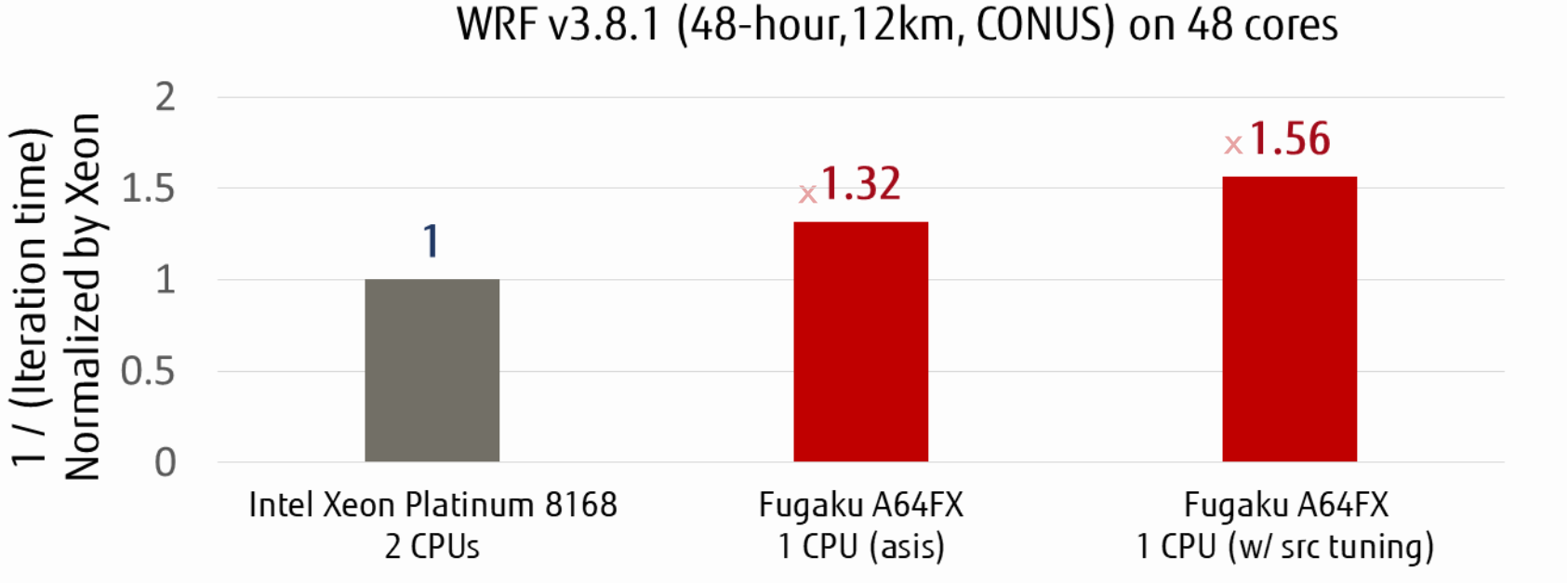


† “Performance evaluation of a vector supercomputer SX-aurora TSUBASA”, SC18, https://dl.acm.org/citation.cfm?id=3291728

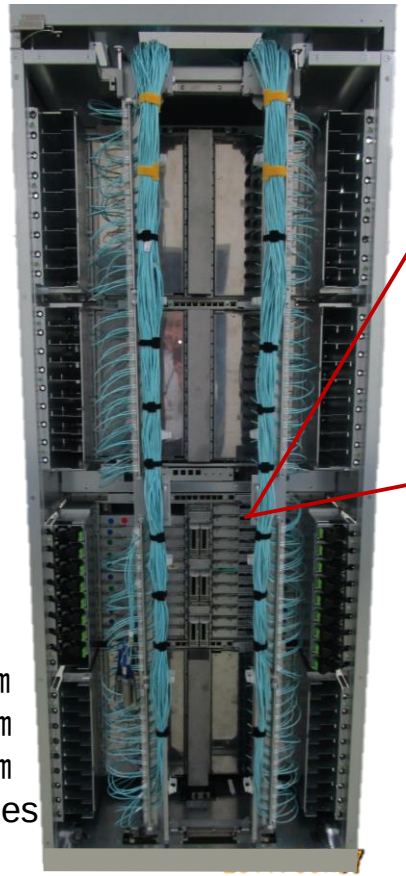
“Fugaku” CPU Performance Evaluation (3/3)

FUJITSU

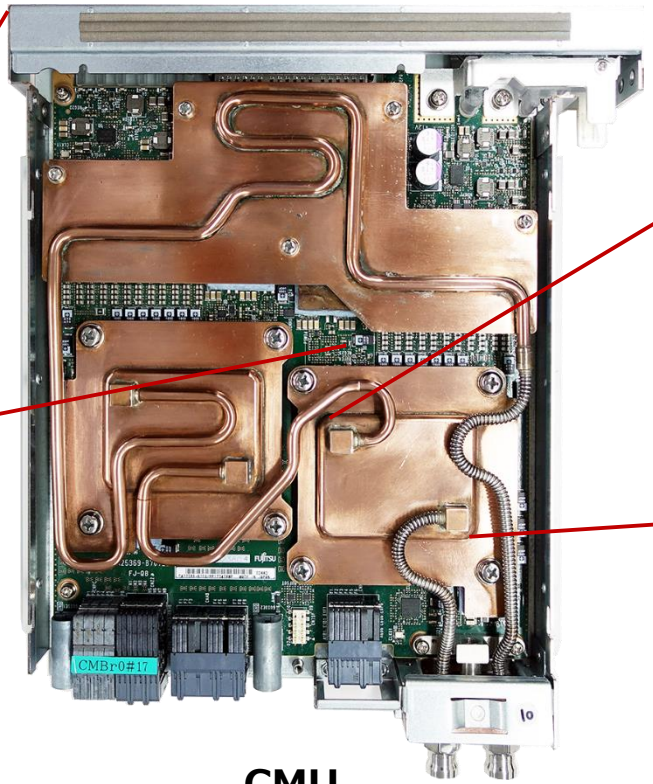
- WRF: Weather Research and Forecasting model
  - Vectorizing loops including IF-constructs is key optimization
  - Source code tuning using directives promotes compiler optimizations



Fugaku Chassis, PCB (w/DLC), and CPU Package



W 800mm
D1400mm
H2000mm
384 nodes



CMU

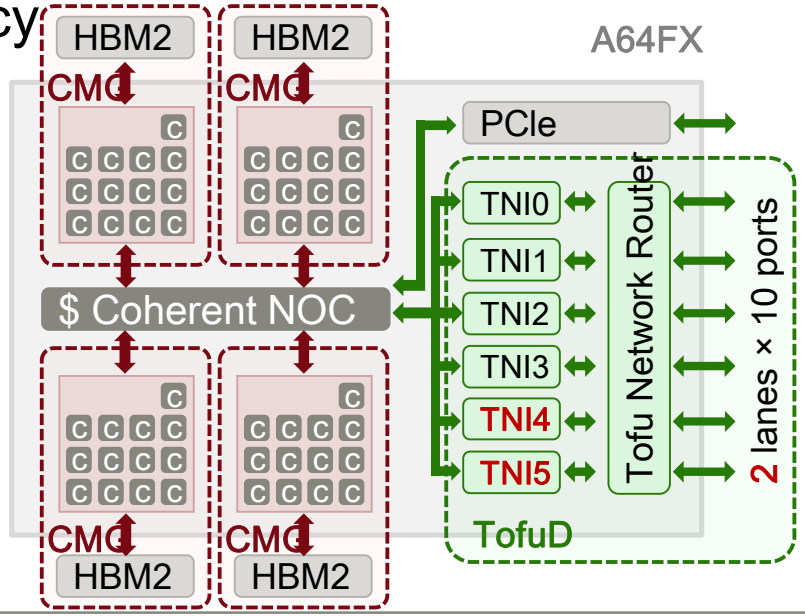
CPU Package

A0 Chip Booted in June
Undergoing Tests

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

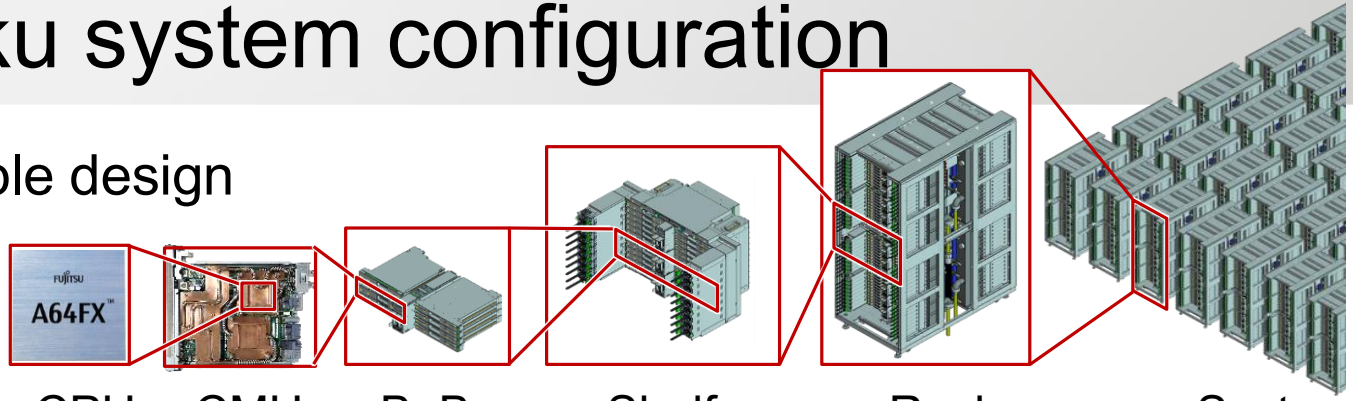
	TofuD spec
Port bandwidth	6.8 GB/s
Injection bandwidth	40.8 GB/s
	Measured
Put throughput	6.35 GB/s
Ping-pong latency	0.49~0.54 μs



Fugaku system configuration

FUJITSU

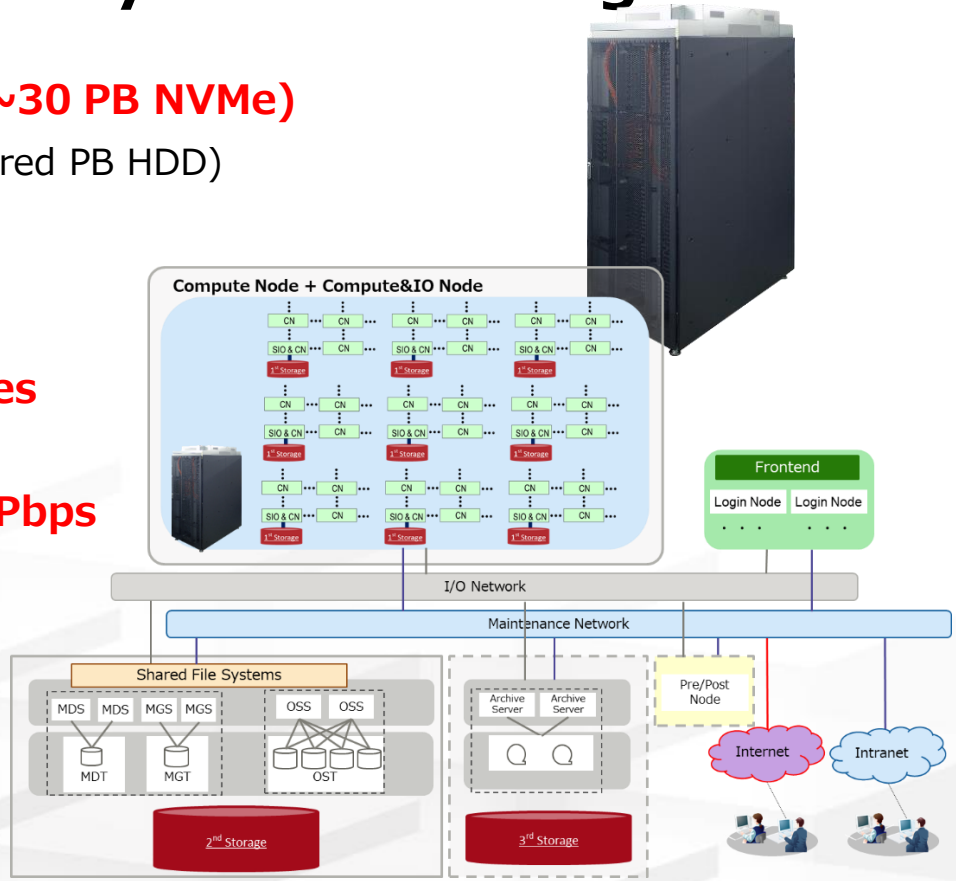
- Scalable design



Unit	# of nodes	Description
CPU	1	Single socket node with HBM2 & Tofu interconnect D
CMU	2	CPU Memory Unit: 2x CPU
BoB	16	Bunch of Blades: 8x CMU
Shelf	48	3x BoB
Rack	384	8x Shelf
System	150k+	As a Fugaku system

Overview of Fugaku System & Storage

- **3-level hierarchical storage**
  - 1st Layer: GFS Cache + Temp FS **(25~30 PB NVMe)**
  - 2nd Layer: Lustre-based GFS (a few hundred PB HDD)
  - 3rd Layer: Off-site Cloud Storage
- **Full Machine Spec**
  - **>150,000 nodes**
    ~8 million High Perf. Arm v8.2 Cores
  - **> 150PB/s memory BW**
  - **Tofu-D 10x Global IDC traffic @ 60Pbps**
  - **~10,000 I/O fabric endpoints**
  - **> 400 racks**
  - **~40 MegaWatts Machine+IDC**
    PUE ~ 1.1 High Pressure DLC
  - **NRE pays off: ~= 15~30 million state-of-the art competing CPU Cores for HPC workloads (both dense and sparse problems)**



Fugaku Performance Estimate on 9 Co-Design Target Apps

Performance target goal

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

Peak performance to be achieved

	PostK	K
Peak DP (double precision)	>400+ Pflops (34x +)	11.3 Pflops
Peak SP (single precision)	>800+ Pflops (70x +)	11.3 Pflops
Peak HP (half precision)	>1600+ Pflops (141x +)	--
Total memory bandwidth	>150+ PB/sec (29x +)	5,184TB/sec

Geometric Mean of Performance Speedup of the 9 Target Applications over the K-Computer

> 37x+

As of 2019/05/14

Category	Priority Issue Area	Performance Speedup over K	Application	Brief description
Health and longevity	1. Innovative computing infrastructure for drug discovery	125x +	GENESIS	MD for proteins
	2. Personalized and preventive medicine using big data	8x +	Genomon	Genome processing (Genome alignment)
Disaster prevention and Environment	3. Integrated simulation systems induced by earthquake and tsunami	45x +	GAMERA	Earthquake simulator (FEM in unstructured & structured grid)
	4. Meteorological and global environmental prediction using big data	120x +	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	40x +	NTChem	Molecular electronic simulation (structure calculation)
	6. Accelerated development of innovative clean energy systems	35x +	Adventure	Computational Mechanics System for Large Scale Analysis and Design (unstructured grid)
Industrial competitiveness enhancement	7. Creation of new functional devices and high-performance materials	30x +	RSDFT	Ab-initio simulation (density functional theory)
	8. Development of innovative design and production processes	25x +	FFB	Large Eddy Simulation (unstructured grid)
Basic science	9. Elucidation of the fundamental laws and evolution of the universe	25x +	LQCD	Lattice QCD simulation (structured grid Monte Carlo)

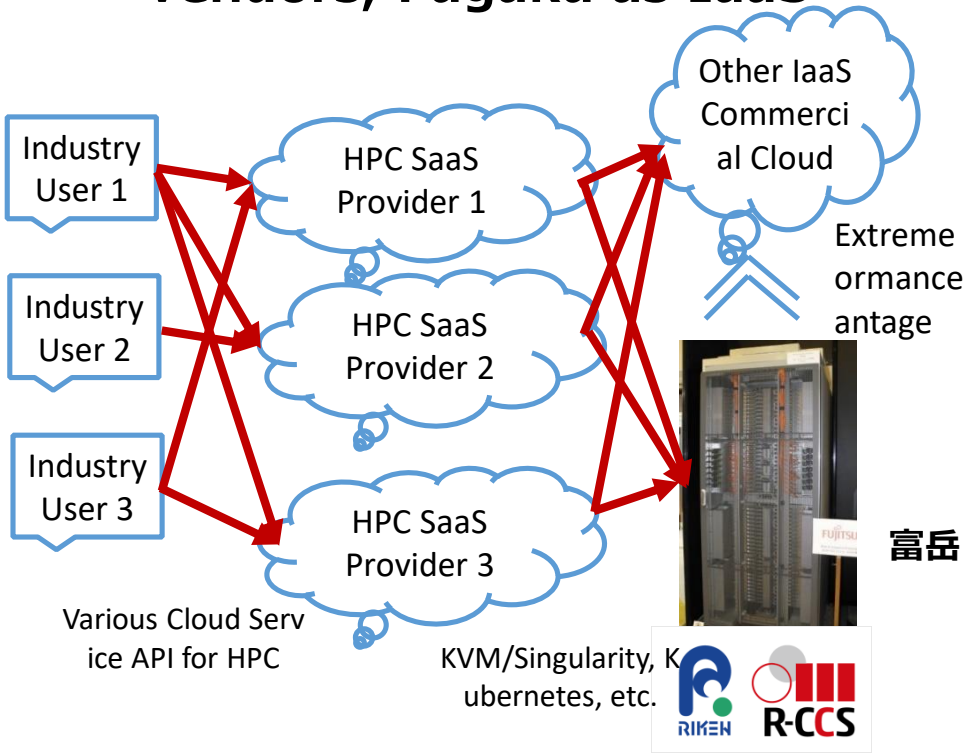
Fugaku Programming Environment

- **Programing Languages and Compilers provided by Fujitsu**
  - Fortran2008 & Fortran2018 subset
  - C11 & GNU and Clang extensions
  - C++14 & C++17 subset and GNU and Clang extensions
  - OpenMP 4.5 & OpenMP 5.0 subset
  - Java

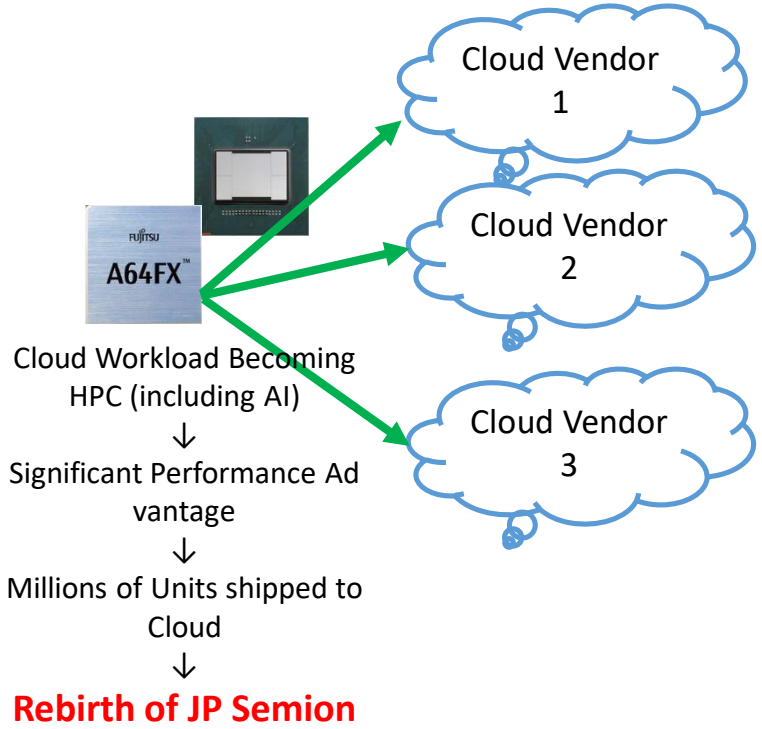
    GCC and LLVM will be also available
- **Parallel Programming Language & Domain Specific Library provided by RIKEN**
  - XcalableMP
  - FDPS (Framework for Developing Particle Simulator)
- **Process/Thread Library provided by RIKEN**
  - PiP (Process in Process)
- **Script Languages provided by Linux distributor**
  - E.g., Python+NumPy, SciPy
- **Communication Libraries**
  - MPI 3.1 & MPI4.0 subset
    - Open MPI base (Fujitsu), MPICH (RIKEN)
  - Low-level Communication Libraries
    - uTofu (Fujitsu), LLC(RIKEN)
- **File I/O Libraries provided by RIKEN**
  - Lustre
  - pnetCDF, DTF, FTAR
- **Math Libraries**
  - BLAS, LAPACK, ScaLAPACK, SSL II （Fujitsu）
  - EigenEXA, Batched BLAS （RIKEN）
- **Programming Tools provided by Fujitsu**
  - Profiler, Debugger, GUI
- **NEW: Containers (Singularity) and other Cloud APIs**
- **NEW: AI software stacks (w/ARM)**
- **NEW: DoE Spack Package Manager**

Fugaku Cloud Strategy

- **Industry use of Fugaku via intermediary cloud SaaS vendors, Fugaku as IaaS**



- **A64fx and other Fugaku Technology being incorporated into the Cloud**



A64fx in upcoming Stony Brook Cray System

National Science Foundation
WHERE DISCOVERIES BEGIN

SEARCH

HOME | RESEARCH AREAS | FUNDING | AWARDS | DOCUMENT LIBRARY | NEWS | ABOUT NSF

Awards

Search Awards
Recent Awards
Presidential and Honorary Awards
About Awards

How to Manage Your Award

Grant Policy Manual
Grant General Conditions
Cooperative Agreement Conditions
Special Conditions
Federal Demonstration Partnership
Policy Office Website

Award Abstract #1927880

Category II : Ookami: A high-productivity path to frontiers of scientific discovery enabled by exascale system technologies

NSF Org:	OAC Office of Advanced Cyberinfrastructure (OAC)
Initial Amendment Date:	July 11, 2019
Latest Amendment Date:	August 29, 2019
Award Number:	1927880
Award Instrument:	Cooperative Agreement
Program Manager:	Robert Chadduck OAC Office of Advanced Cyberinfrastructure (OAC) CSE Direct For Computer & Info Scie & Enginr
Start Date:	October 1, 2019
End Date:	September 30, 2024 (Estimated)
Awarded Amount to Date:	$2,780,373.00
Investigator(s):	Robert Harrison robert.harrison@stonybrook.edu (Principal Investigator) Barbara Chapman (Co-Principal Investigator) Matthew Jones (Co-Principal Investigator) Alan Calder (Co-Principal Investigator)
Sponsor:	SUNY at Stony Brook WEST 5510 FRK MEL LIB Stony Brook, NY 11794-0001 (631)632-9949
NSF Program(s):	Innovative HPC
Program Reference Code(s):	
Program Element Code(s):	7619

ABSTRACT

The State University of New York proposes to procure and operate for at least four years the first computer outside of Japan with the A64fx processor developed by Fujitsu for the Japanese path to exascale computing (i.e., computers capable of 10^18 operations per second). The ARM-based, multi-core, 512-bit SIMD-vector processor with ultrahigh-bandwidth memory promises to retain familiar and successful programming models while achieving very high performance for a wide range of applications including simulation and big data. The testbed significantly extends current NSF-sponsored HPC technologies and will enable the community to evaluate and demonstrate the potential of this technology for deployment in multiple settings. Through integration with NSF's Extreme Science and Engineering Discovery Environment (XSEDE), the system will be widely accessible and fully leverages existing cyber infrastructure including the XDMoD monitoring system.

What does this mean for science? Compared with the best CPUs anticipated during the deployment period, A64fx offers 2-4x better performance on memory-intensive applications such as sparse-matrix solvers found in many engineering and physics codes.

HPCwire

Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them

Home
Technologies
Sectors
AI/ML/DL
Exascale
Specials
Resource Library
Podcast
Events
Job Bank

Cray ARM-based 'Ookami' to Serve as Testbed for Computational Studies at Stony Brook

August 16, 2019

STONY BROOK, N.Y., August 16, 2019 – A $5 million grant from the National Science Foundation (NSF) to the Institute of Advanced Computational Science (IACS) at Stony Brook University will enable researchers nationwide to test future supercomputing technologies and advance computational and data-driven research on the world's most pressing challenges.

Serving as a testbed for advanced computer technologies, the Ookami system is expected to signal a new generation of high-speed U.S. supercomputers. Using a Cray ARM-based system, Ookami will deliver remarkably high performance for scientific applications, in part due to its blazing-fast memory. Robert J. Harrison, PhD, professor of applied mathematics and statistics and director of IACS, expects that these advanced technologies will enable researchers to more quickly and effectively conduct computational investigations. The project is led by IACS faculty in partnership with co-PI Matt Jones, PhD at the State University of New York at Buffalo, whose team will lead the capture of detailed operational metrics and provision of extensive

Ookami

- Test bed for NSF researchers
  - First planned deployment of the Post-K processor outside of Japan
- Collaboration with Riken CCS
  - http://www.riken.jp/en/research/labs/r-ccs/
- Installation 3Q 2020
- $5M award NSF OAC 1942140 for purchase and operations

Node	
Processor	A64FX
#Cores	48+4
Peak DP	2.76 TOP/s
Peak INT8	22.08 TOP/s
Memory	32GB@1TB/s
System	
#Nodes	176
Peak DP	486 TOP/s
Peak INT8	3886 TOP/s
Memory	5.6 TB
Disk	0.5 PB
Comms	IB HDR-100

Pursuing Convergence of HPC & AI (1)

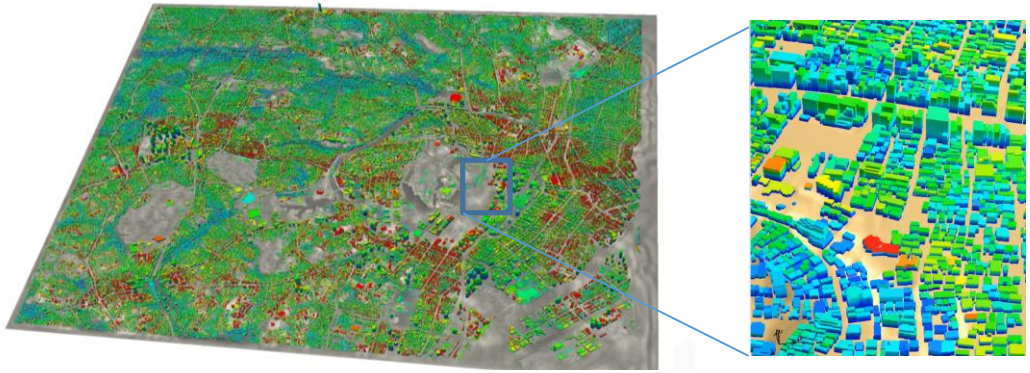
- **Acceleration of Simulation (first principles methods) with AI (empirical method) : *AI for HPC***
  - Interpolation & Extrapolation of long trajectory MD
  - Reducing parameter space on Paretho optimization of results
  - Adjusting convergence parameters for iterative methods etc.
  - AI *replacing* simulation when exact physical models are unclear, or excessively costly to compute
- **Acceleration of AI with HPC: *HPC for AI***
  - HPC Processing of training data -data cleansing
  - Acceleration of (Parallel) Training: Deeper networks, bigger training sets, complicated networks, high dimensional data…
  - Acceleration of Inference: above + real time streaming data
  - Various modern training algorithms: Reinforcement learning, GAN, Dilated Convolution, etc.

R-CCS Persuit of Convergence of HPC & AI (2)

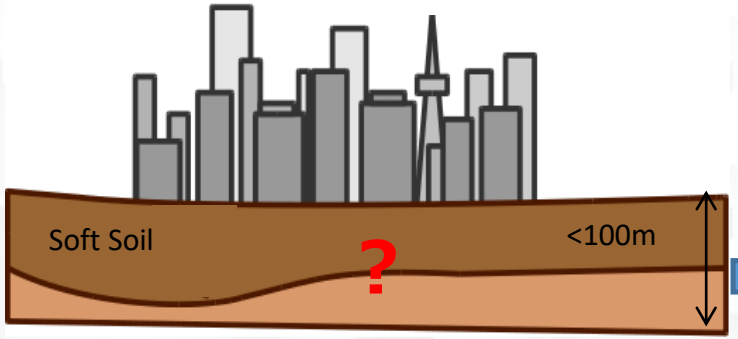
- **Acceleration of Simulation (first principles methods) with AI (empirical method) : *AI for HPC***
  - ***Most*** R-CCS research & operations teams investigating use of AI for HPC
  - 9 priority co-design issues area teams also extensive plans
  - Essential to deploy AI/DL frameworks efficiently & at scale on A64fx/Fugaku

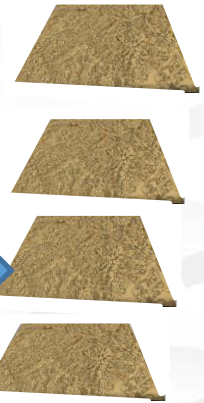
- **Acceleration of AI with HPC: *HPC for AI***
  - New teams instituted in Science of Computing to accelerate AI
    - Kento Sato (High Performance Big Data Systems)
    - Satoshi Matsuoka (High Performance AI Systems)
    - Masaaki Kondo Next Gen (High Performance Architecture)
  - Collaborations with various projects

Large Scale simulation and AI coming together
[Ichimura et. al. Univ. of Tokyo, IEEE/ACM SC17 Best Poster 2018 Gordon Bell Finalist]

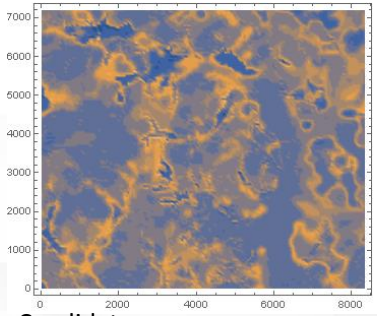


130 billion freedom earthquake of entire Tokyo on K-Computer (2018 ACM Gordon Bell Prize Finalist, SC16,17 Best Poster)



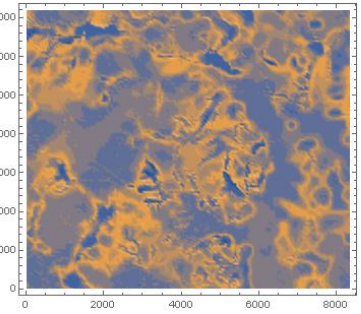


Too Many Instances



Candidate Underground Structure 1

AI Trained by Simulation to generate candidate soft soil structure

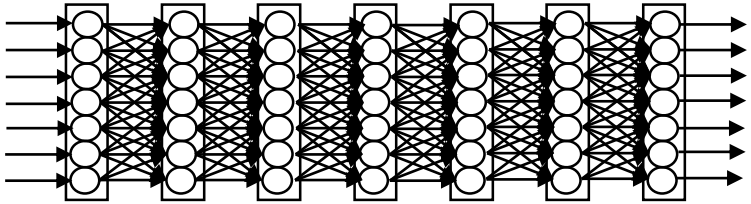


Candidate Underground Structure 2

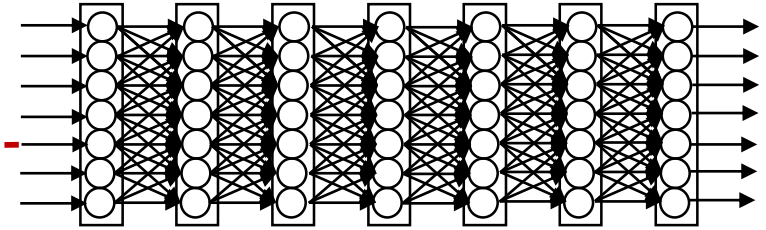
4 Layers of Parallelism in DNN Training

- Hyper Parameter Search
  - Searching optimal network configs & parameters
  - Parallel search, massive parallelism required
- Data Parallelism
  - Copy the network to compute nodes, feed different batch data, average => network reduction bound
  - TOFU: Extremely strong reduction, x6 EDR Infiniband

Inter-Node



- Model Parallelism (domain decomposition)
  - Split and parallelize the layer calculations in propagation
  - Low latency required (bad for GPU) -> strong latency tolerant cores + low latency TOFU network



- Intra-Chip ILP, Vector and other low level Parallelism
  - Parallelize the convolution operations etc.
  - SVE FP16+INT8 vectorization support + extremely high memory bandwidth w/HBM2

Intra-Node

- Post-K could become world's biggest & fastest platform for DNN training!

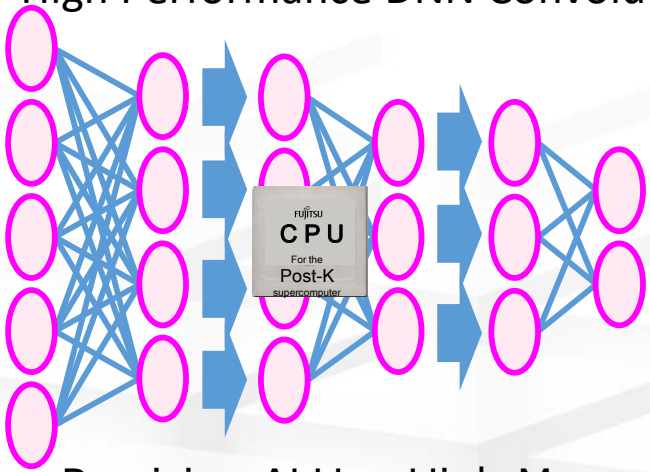
Massive amount of total parallelism, only possible via supercomputing

Massive Scale Deep Learning on Post-K

Post-K Processor

- ◆High perf FP16&Int8
- ◆**High mem BW for convolution**
- ◆**Built-in scalable Tofu network**

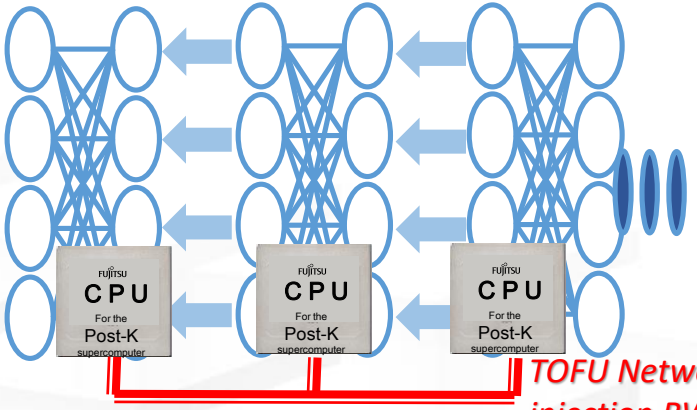
High Performance DNN Convolution



Low Precision ALU + High Memory Bandwidth + Advanced Combining of Convolution Algorithms (FFT+Winograd+GEMM)

Unprecedened DL scalability

High Performance and Ultra-Scalable Network for massive scaling model & data parallelism

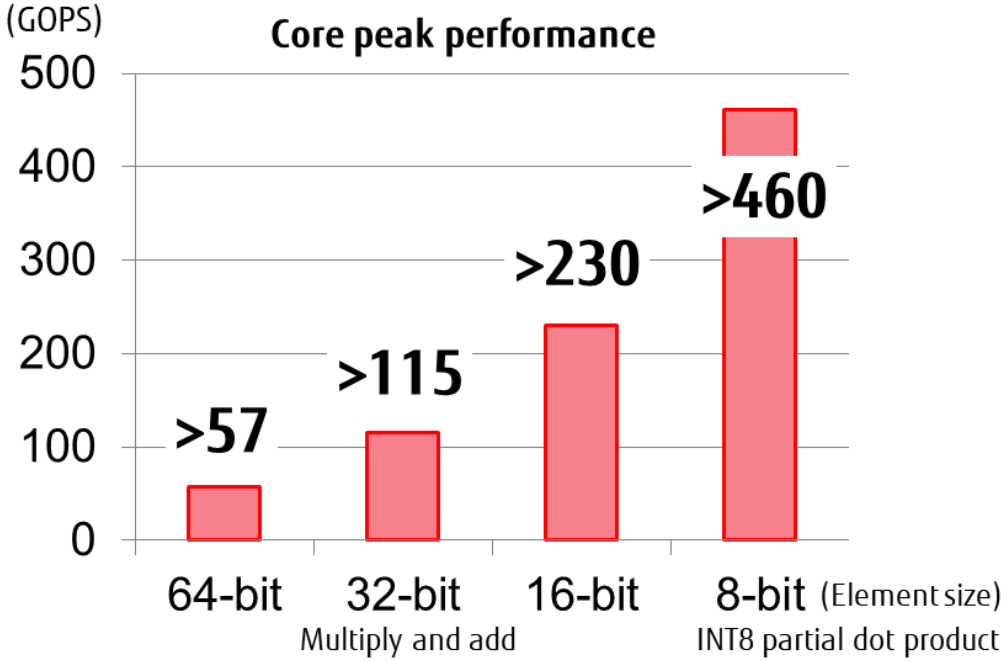


TOFU Network w/high injection BW for fast reduction

Unprecedented Scalability of Data/

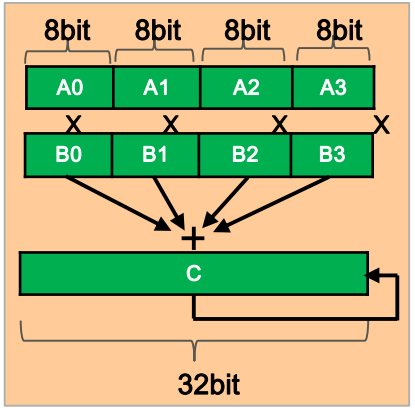
A64FX technologies: Core performance

- High calc. throughput of Fujitsu's original CPU core w/ SVE
  - 512-bit wide SIMD x 2 pipelines and new integer functions
  - 8 bit inference => training possible?



INT8 partial dot product

$C = \Sigma (Ai \times Bi) + C$



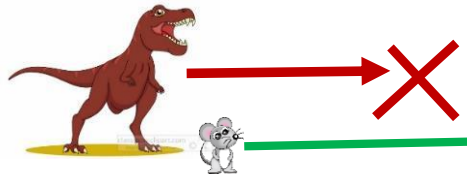
Large Scale Public AI Infrastructures in Japan

Inference 838.5PF
Training 86.9 PF

vs. Summit
Inf. 1/4
Train. 1/5

	Deployed	Purpose	AI Processor	Inference Peak Perf.	Training Peak Perf.	Top500 Perf/Rank	Green500 Perf/Rank
Tokyo Tech. TSUBAME3	July 2017	HPC + AI Public	NVIDIA P100 x 2160	45.8 PF (FP16)	22.9 PF / 45.8PF (FP32/FP16)	8.125 PF #22	13.704 GF/W #5
U-Tokyo Reedbush-H/L	Apr. 2018 (update)	HPC + AI Public	NVIDIA P100 x 496	10.71 PF (FP16)	5.36 PF / 10.71PF (FP32/FP16)	(Unranked)	(Unranked)
U-Kyushu ITO-B	Oct. 2017	HPC + AI Public	NVIDIA P100 x 512	11.1 PF (FP16)	5.53 PF/11.1 PF (FP32/FP16)	(Unranked)	(Unranked)
AIST-AIRC AICC	Oct. 2017	AI Lab Only	NVIDIA P100 x 400	8.64 PF (FP16)	4.32 PF / 8.64PF (FP32/FP16)	0.961 PF #446	12.681 GF/W #7
Riken-AIP Raiden	Apr. 2018 (update)	AI Lab Only	NVIDIA V100 x 432	54.0 PF (FP16)	6.40 PF/54.0 PF (FP32/FP16)	1.213 PF #280	11.363 GF/W #10
AIST-AIRC ABCI	Aug. 2018	AI Public	NVIDIA V100 x 4352	544.0 PF (FP16)	65.3 PF/544.0 PF (FP32/FP16)	19.88 PF #7	14.423 GF/W #4
NICT (unnamed)	Summer 2019	AI Lab Only	NVIDIA V100 x 1700程度	~210 PF (FP16)	~26 PF/~210 PF (FP32/FP16)	????	????
C.f. US ORNL Summit	Summer 2018	HPC + AI Public	NVIDIA V100 x 27,000	3,375 PF (FP16)	405 PF/3,375 PF (FP32/FP16)	143.5 PF #1	14.668 GF/W #3
Riken R-CCS Fugaku	2020 ~2021	HPC + AI Public	Fujitsu A64fx > x 150,000	> 4000 PO (Int8)	>1000PF/>2000PF (FP32/FP16)	> 400PF #1 (2020?)	> 15 GF/W ???
ABCI 2 (speculative)	2022 ~2023	AI Public	Future GPU ~ 5000	Similar	similar	~100PF	25~30GF/W ???

Many Core Era



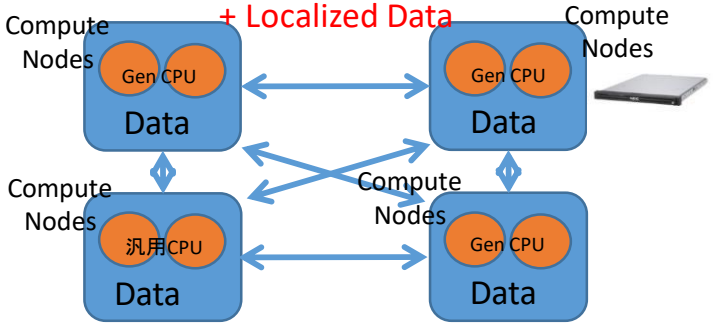
Post Moore
Cambrian Era

Flops-Centric Monolithic Algorithms and Apps

Flops-Centric Monolithic System Software

Hardware/Software System APIs
Flops-Centric Massively Parallel Architecture

Homogeneous General Purpose Nodes
+ Localized Data



Loosely Coupled with Electronic Interconnect

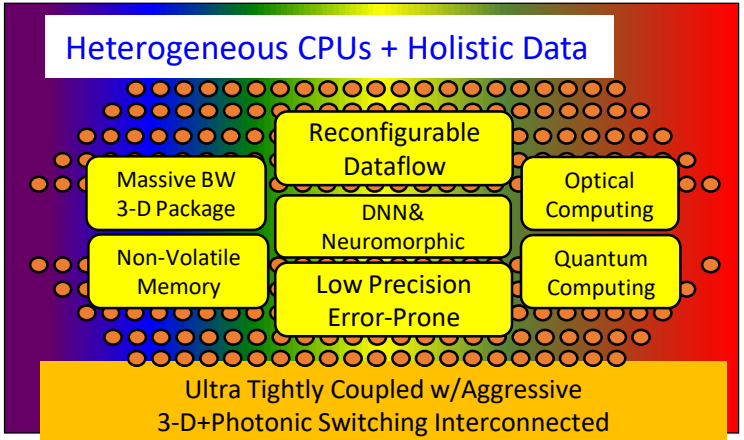
Transistor Lithography Scaling
(CMOS Logic Circuits, DRAM/SRAM)

~2025
M-P Extinction
Event

Cambrian Heterogeneous Algorithms and Apps

Cambrian Heterogeneous System Software

Hardware/Software System APIs
"Cambrian" Heterogeneous Architecture



Novel Devices + CMOS (Dark Silicon)
(Nanophotonics, Non-Volatile Devices etc.)

Our Strategies Towards Post-Moore Era

- **Basic Research on Post-Moore**
  - Funded 2017: DEEP-AI CREST (Matsuoka)
  - Funded 2018: NEDO 100x 2028 Processor Architecture (Matsuoka, Sano, Kondo, SatoK)
  - Funded 2019: Kiban-S Post-Moore Algorithms (NakajimaK etc.)
  - Submitted: Neuromorphic Architecture (Sano etc. w/Riken AIP, Riken CBS (Center for Brain Science))
  - In preparation: Cambrian Computing (w/HPCI Centers)
- **Author a Post-Moore Whitepaper towards Fugaku-next**
  - All-hands BoF last week at annual SWoPP workshop
  - Towards official “Feasibility Study” towards Fugaku-next
  - Similar efforts as K => Fugaku started in 2012

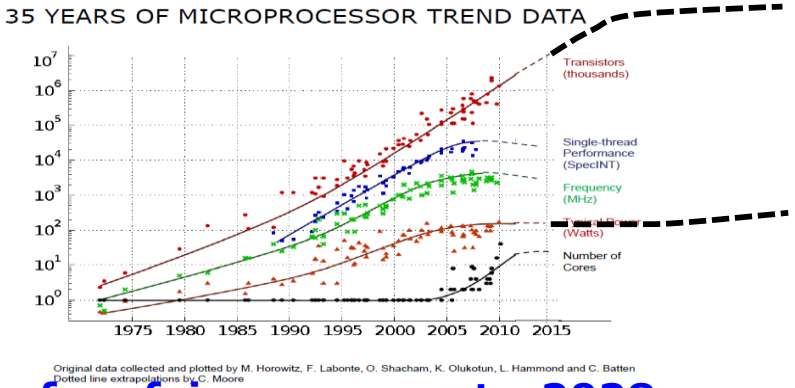
Basic Research #1: NEDO 100x Processor

2028: Post-Moore Era

~2015 ~25 Years Post-Dennard, Many-core Scaling era

2016~Moore's Law Slowing Down

2025~Post-Moore Era, end of transistor lithography (FLOPS) improvement



Research: Architectural investigation of perf. improvement ~2028

- 100x in 2028 c.f. mainstream high-end CPUs circa 2018 across applications

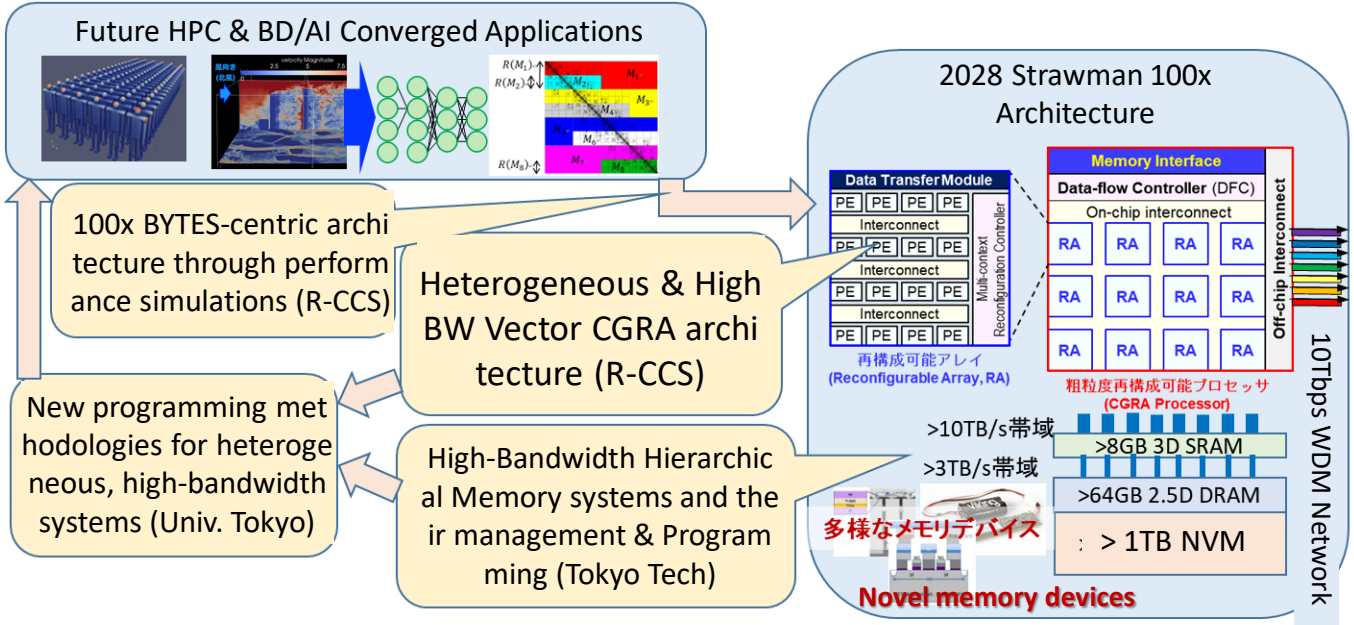
Key to performance improvement: from FLOPS to Bytes – data movement architectural optimization

- CGRA – Coarse-Grained Reconfigurable Vector Dataflow
- Deep & Wide memory architecture w/advanced 3D packaging & novel memory devices
- All-Photonic DWM interconnect w/high BW, low energy injection
- Kernel-specific HW optimization w/low # of transistors & associated system software, programming, and algorithms

NEDO 100x Processor

Towards 100x processor in 2028

- Various combinations of CPU architectures, new memory devices and 3-D technologies
- Perf. measurement/characterization/models for high-BW intra-chip data movement
- Cost models and algorithms for horizontal & hierarchical data movement
- Programming models and heterogeneous resource management



12 Apr, 2019

Collaborations w/External Partners

New generation HPC&AI converged appications

Survey and whitepaper of next-gen HPC & BD/AI coverged applications

BDEC2 (Big Data and Extreme Computing) main contributor, second meeting Feb 2019 @ Riken-CCS (85 particpants)

- The University of Tennessee
- Argonne National Laboratory
- Barcelona Supercomputing Center
- NCSA/University of Illinois
- Sandia National Laboratories
- Georgia Tech
- Shanghai Jiao Tong University
- Los Alamos National Laboratory
- JP HPCI Centers

Survey of next-gen hardware technologies

Various deep conversations with vendors home & abroad

Various future technology outlook towards Post-Moore computing

- HPE
- Intel
- Fujitsu
- NVIDIA
- ARM
- Cavium

Collaborations on basic post-moore research

LBNL Collaboration on post-moore machine archtctures

Surveying and simulating future BYTES-oriented post-moore architectures

Topics

PARADISE (PostMoore system simulator)
Hardware generators, Open Hardware
Advanced materials and devices
CRXO and EUREJKA
Quantum computing

Application Evaluation

Main Collaborators

John Shalf
Ramamorthy Ramesh
Dabn Armbrust
Dilip Vasudevan
David Donofrio
Patrick Naulleau
Chia Chang
Anastassia Butko