

スーパーコンピュータ「富岳」の開発

理化学研究所 計算科学研究センター
副センター長
フラッグシップ2020 副プロジェクトリーダー
アーキテクチャ開発チーム・チームリーダー
筑波大学 連携大学院教授

佐藤 三久

フラッグシップ2020 プロジェクトリーダー
石川裕

富岳

スーパーコンピュータ「富岳」(2021 ~)



ふがくん

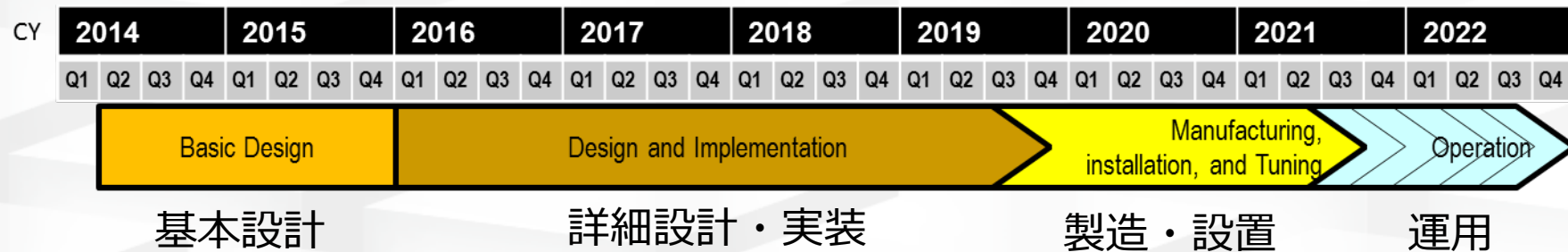


- **富岳開発プロジェクト・フラグシップ2020の概要**
- **富岳システムの概要**
 - プロセッサ A64FX
 - 性能
 - ソフトウェア
- **成果・その他**
- **開発プロジェクトを振り返って**
 - 情報科学の研究とスパコン・プロジェクト

フラッグシップ°2020プロジェクト

- 目的
 - 「京」の次のスーパーコンピュータ（ポスト「京」=「富岳」）の開発
 - ポスト京で我が国の科学および社会的・諸課題を解決するアプリケーションの開発
- プロジェクトは、2014年度から開始
 - 理研が開発主体
 - ベンダーパートナーは、富士通
- 2020年度末で開発プロジェクトは終了、2021年3月から共用開始

プロジェクトが開始されるまえに、2年(2012-2013)のフェージビリティプロジェクトがあった。



Post-K(富岳)の開発については、次の3つの開発目標を設定し、設計を進めた。

- **1. 高い電力効率**
 - 施設の最大電力量は、30 - 40MW (for system)に設定し、電力設備を準備した
- **2. 「ターゲット・アプリケーション」での高効率・高性能**
 - 目安として、いくつかのアプリケーションで、「京」の100倍以上の性能を達成できること。
 - ベンチマーク (Top500等) は目標ではない。
- **3. 広いユーザーに対して、使いやすいシステム**
 - GPUなどのアクセラレータは不採用
 - 京でのアプリケーションが継続して使える

我が国の科学技術の発展、産業競争力の強化に資するため、イノベーションの創出や国民の安全・安心の確保につながる最先端の研究基盤として、2021~22年の運用開始を目標に、世界最高水準の汎用性のあるスーパーコンピュータの実現を目指す。

開発目標

- 最大で「京」の100倍のアプリケーション実効性能
- 消費電力 30~40MW（「京」は12.7MW）

Co-design

- システムとアプリケーションを協調的に開発（Co-design）
- アプリケーションの対象として、健康長寿、防災・減災、エネルギー、ものづくり分野等の社会的・科学的課題を選定

システムの特徴

- 世界最高水準の

- ★消費電力性能
- ★計算能力
- ★ユーザーの利便・使い勝手の良さ
- ★画期的な成果の創出

⇒ 総合力のあるスーパーコンピュータ



【文部科学省 HPCI計画推進委員会（平成28年8月）】

- 基本設計評価後に、最先端の半導体の設計・製造について、加工技術開発の困難さ等から世界的に遅延。ポスト「京」においては、開発スケジュールに12か月から24か月の遅延。
- 目標性能及び経費等の観点から確認を行い、新たな技術を採用して国費総額を変更せずに当初の開発目標を達成する見込みであると評価。
- さらに、ユーザの利便や使い勝手の良さを向上するため、新たな付加価値の創出に向けた取組を実施。

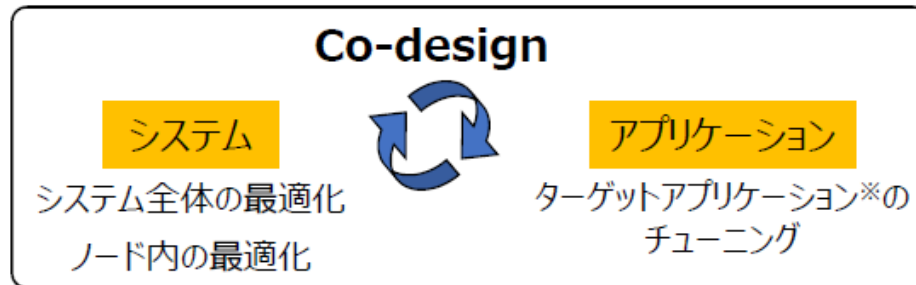
世界最高水準のHPC向け汎用スパコンを開発

最先端半導体技術によるスパコン

- ・自主開発により、いち早く最先端技術を取り入れて高性能かつ低電力なスパコンを実現。

高効率かつ省電力なシステム

- ・ソフトウェアとハードウェアのCo-designにおいて、アーキテクチャの最適化設計による演算回路及びメモリ帯域の高効率利用、アプリケーションの演算特性に適応した電力制御による省電力化を実現。



※ 計算科学的手法を網羅するように、ターゲットアプリケーションを設定

使い勝手の良いシステム

- ・Armアーキテクチャの採用、Linuxディストリビューションを含むオープンソースの活用及びOpen HPCやArm HPCユーザコミュニティとの連携によりArmエコシステムの構築を目指し、多様なアプリケーションユーザの利用を促進。
- ・ポスト「京」の仕様等に関する説明会を開催（平成30年1月）し、さらに2回程度の開催を予定。
- ・Co-designの成果に基づくチューニングマニュアル等の整備・公開、チューニング環境の提供を予定。

Codesign of "Fugaku"

3 Design Targets:

- **1. Extreme Power-Efficient System**
 - Maximum performance under Power consumption of 30 - 40MW (for system)
- **2. Effective performance of target applications**
 - It is expected to exceed 100 times higher than the K computer's performance in some applications
- **3. Ease-of-use system for wide-range of users**

Cool (Low-power) technology is important!!



Codesign

Codesign to meet these
3 design targets

Technologies and Architectural Parameters to be determined

- **Basic Architecture Design (by Feasibility Studies)**
 - Manycore approach, O3 cores, some parameters on chip configuration and SIMD
- **Instruction Set Architecture and SIMD Instructions**
 - Fujitsu collaborated with Arm, contributing to the design of the SVE as a lead partner
- **Chip configuration**
- **Memory technology**
 - DDR, HBM, HMC ...
- **Cache structure**
- **Out of order (O3) resources**
- **Enhancement for Target Applications**
- **Interconnect between Nodes**
 - SerDes, topologies "Tofu" or other network?

- ✓ The number of cores in a CMG
- ✓ The number of CMGs in a chip
- ✓ How to connect cores to shared L2 in a CMG
- ✓ The number of ways, the size, and throughput of the L1 and L2 caches
- ✓ The topology of network-on-chip to connect CMGs
- ✓ The die size of the chip
- ✓ The number of chips in a node

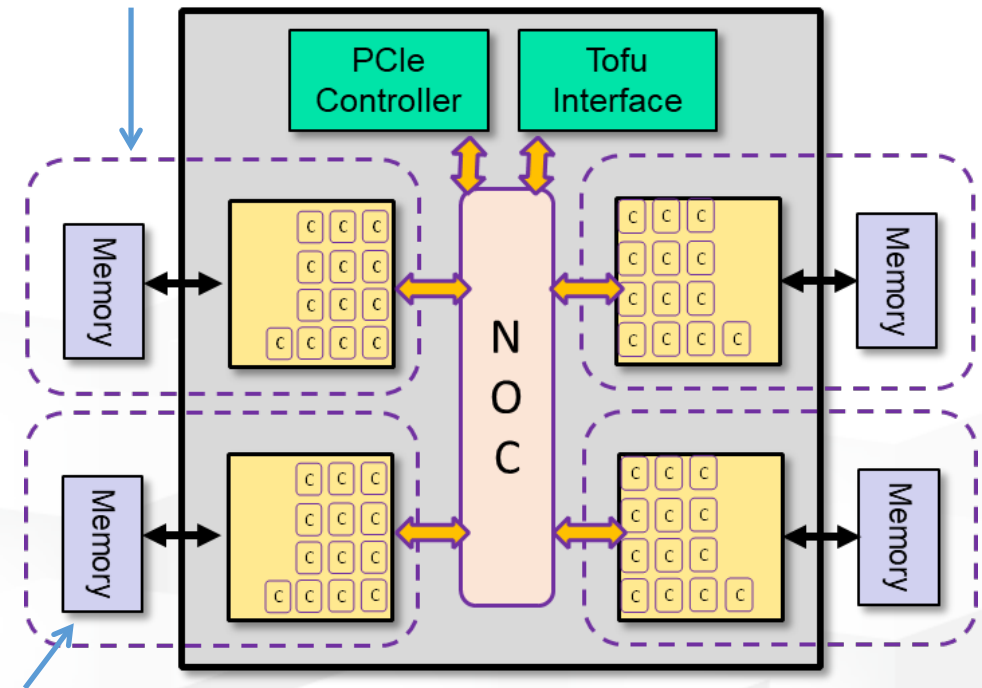
Supercomputer "Fugaku" and A64FX processor

- **Ultra-scale "general-purpose" manycore system: 158,976 nodes (1 processor/node, total 7.6 M cores, theoretical peak 537PFLOPS (DP))**
- **Arm-based manycore processor: Fujitsu A64FX (Armv8.2-A SVE 512bit SIMD, #core 48 + 2/4, 3TF@2.0GHz, boost to 2.2GHz)**
 - 12 cores in a cluster of cores called CMG, connected to L2 and HBM memory chips
- **Advanced Memory technology: HBM2 32 GiB, 1024 GB/s bandwidth, packaged in CPU chip**
- **Scalable Interconnect: ToFu-D interconnect**



- ◆ Standard programming model is OpenMP-MPI hybrid programming. running each MPI process on a NUMA node (CMG).
- ◆ 48 threads OpenMP is also supported.

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

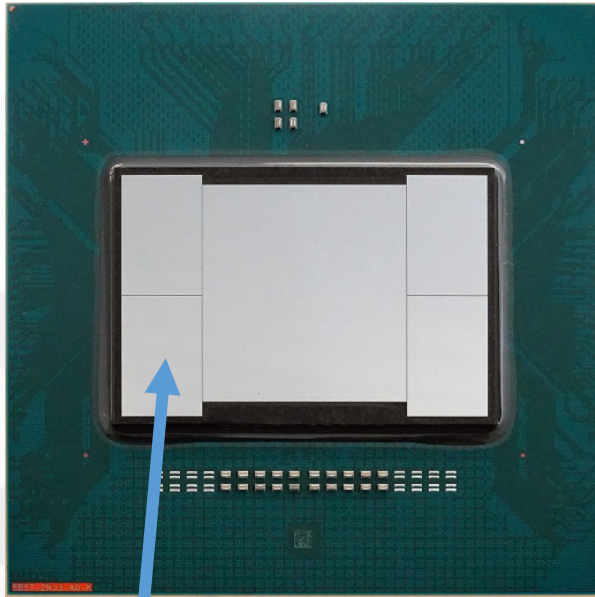


HBM2: 8GiB

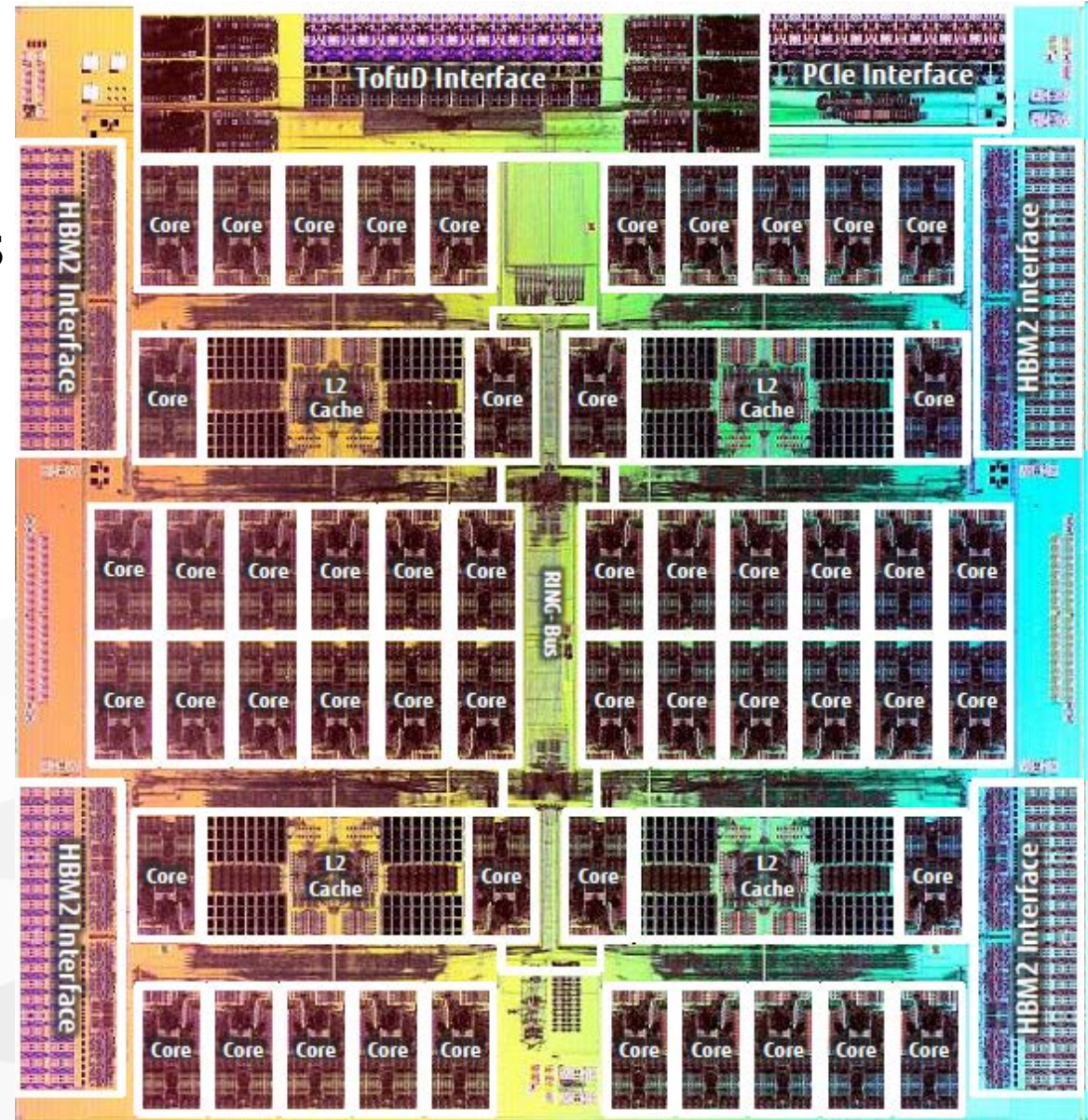
Diagram of A64FX processor

Die Photograph of A64FX processor

- TSMC 7nm FinFET
- 400 mm²
- HBM2 chips are mounted on Si-interposer connected by TSMC CoWoS technology



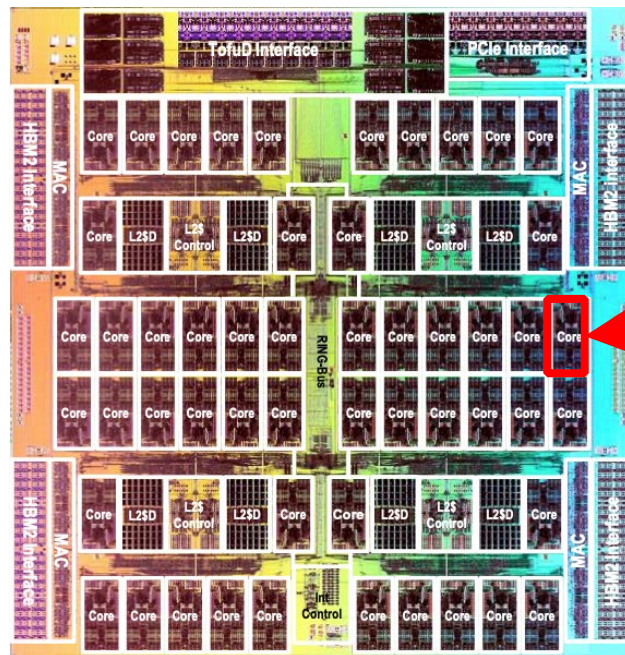
HBM2



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)
- A64FX core is more than 3 times smaller 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_\(server\)](https://en.wikichip.org/wiki/intel/microarchitectures/skylake_(server))

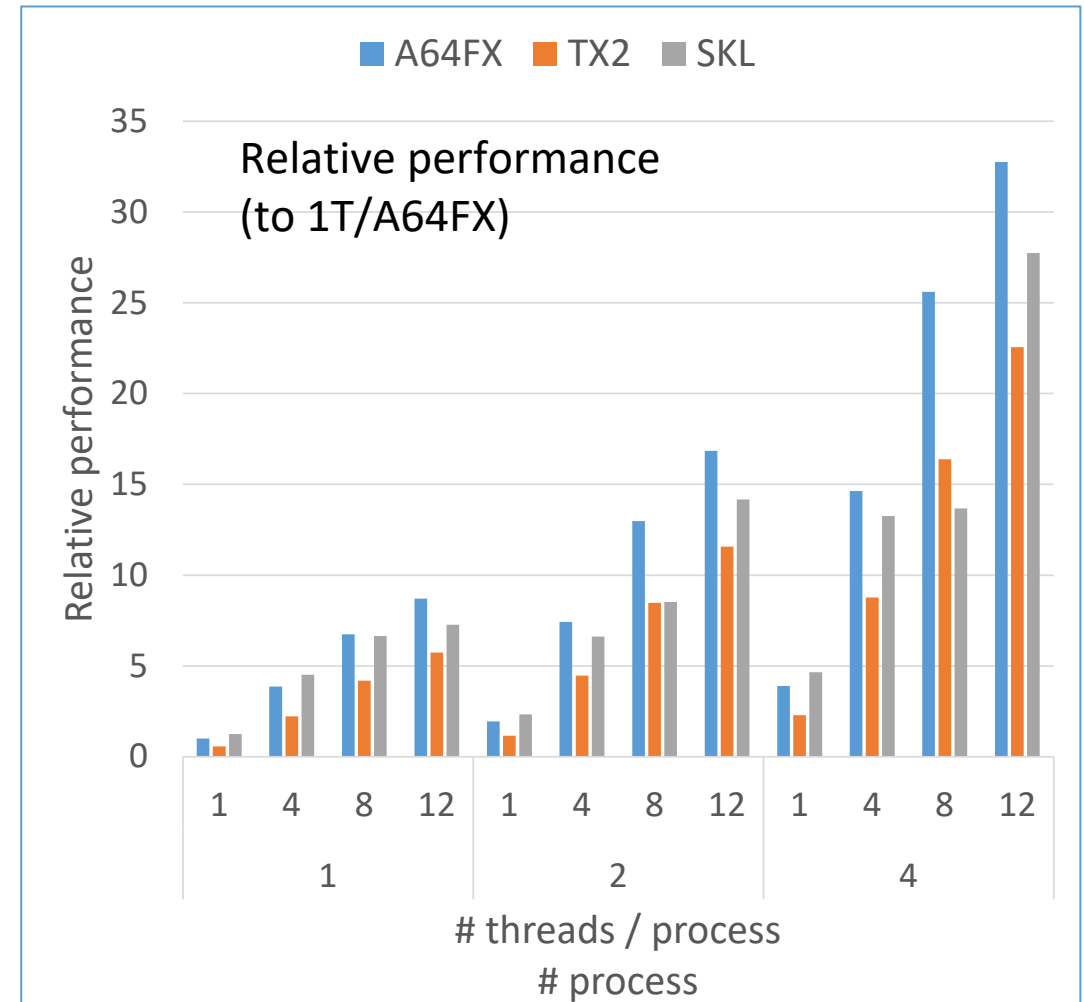
Xeon Skylake, High
Core Count:
4 x 5 tiles, 18 cores, 2
tiles used for memory
interface
485 mm² (22 x 22)

Benchmark result of CloverLeaf

Taken from UK benchmarks:
A hydrodynamics mini-app to solve the compressible Euler equations in 2D, using an explicit, second-order method



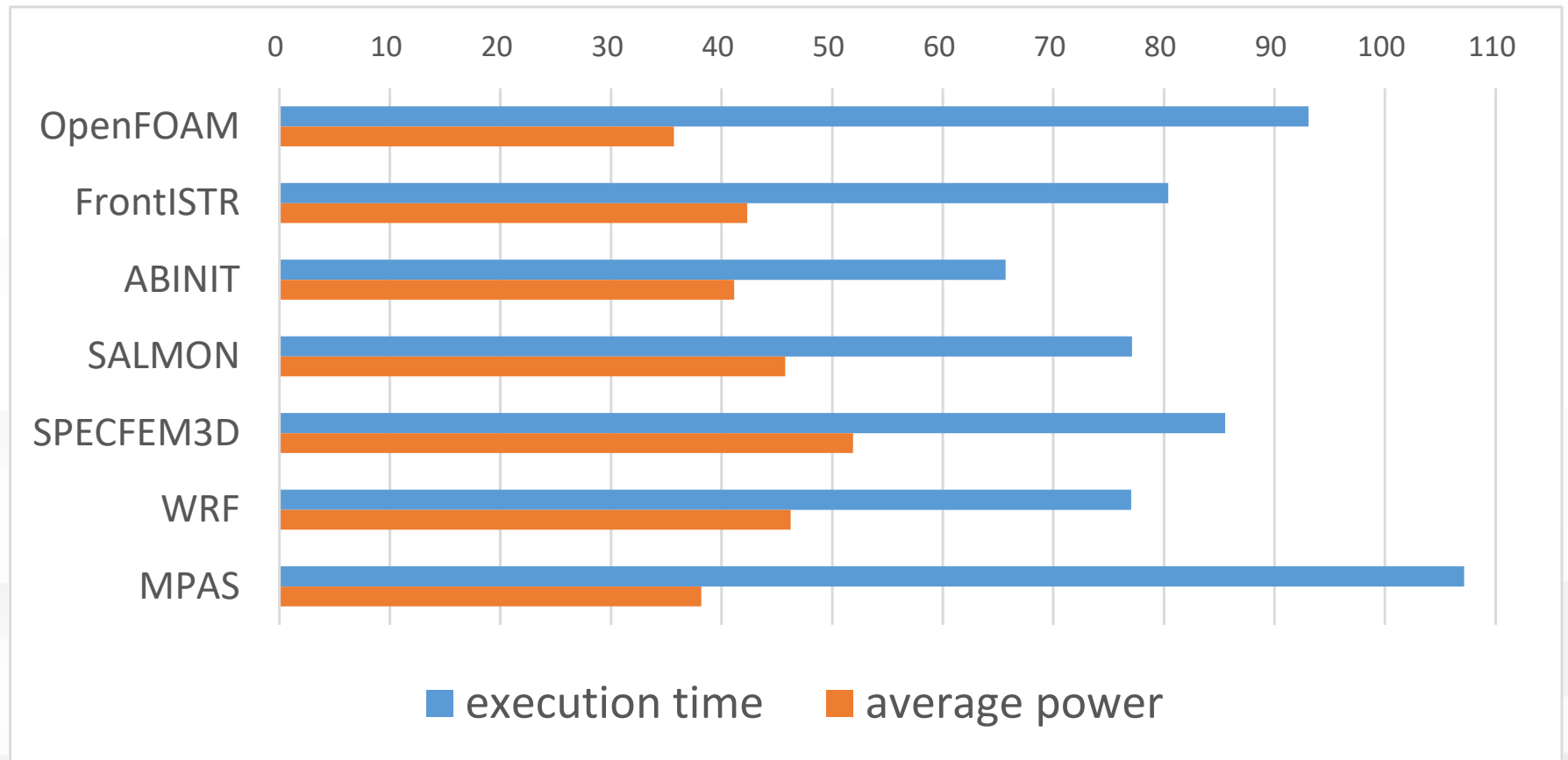
- Comparison with two nodes of TX2 (dual) and Skylake (dual)
- Good scalability by increasing the number of threads within CMG.
- The performance of one A64FX is comparable (better) to that of two nodes (4 sockets) of Skylake



Performance and Power-efficiency of HPC OSS

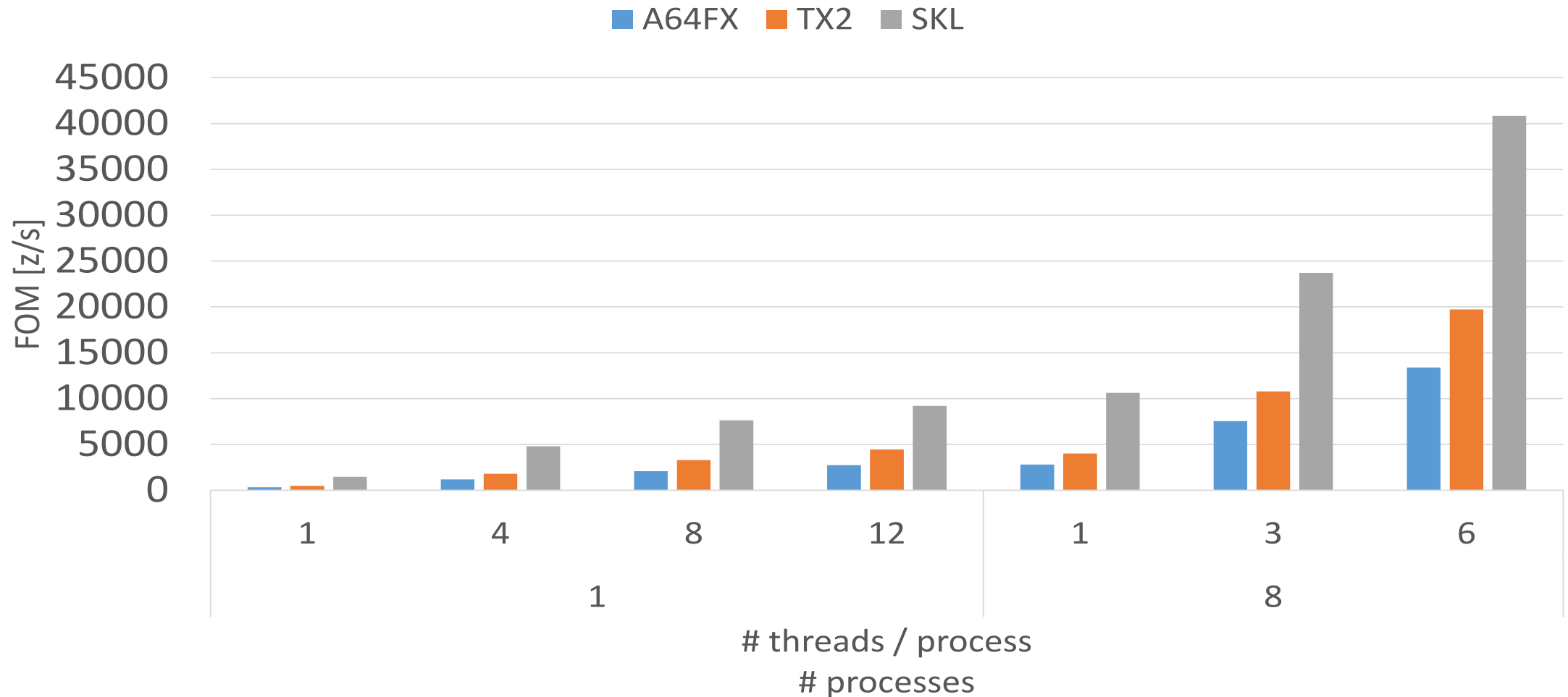
- Several Open-source software were already ported and evaluated.
- Evaluation using one chip A64FX and dual chips of Xeon.
- The almost same performance to dual sockets of Xeon with half of power consumption.

Performance and power efficiency of open-source applications (results are shown in %, relative to Intel Xeon Platinum 8268 (Cascadelake, 2.90 GHz, 24 cores/socket) (dual sockets))



LULESH

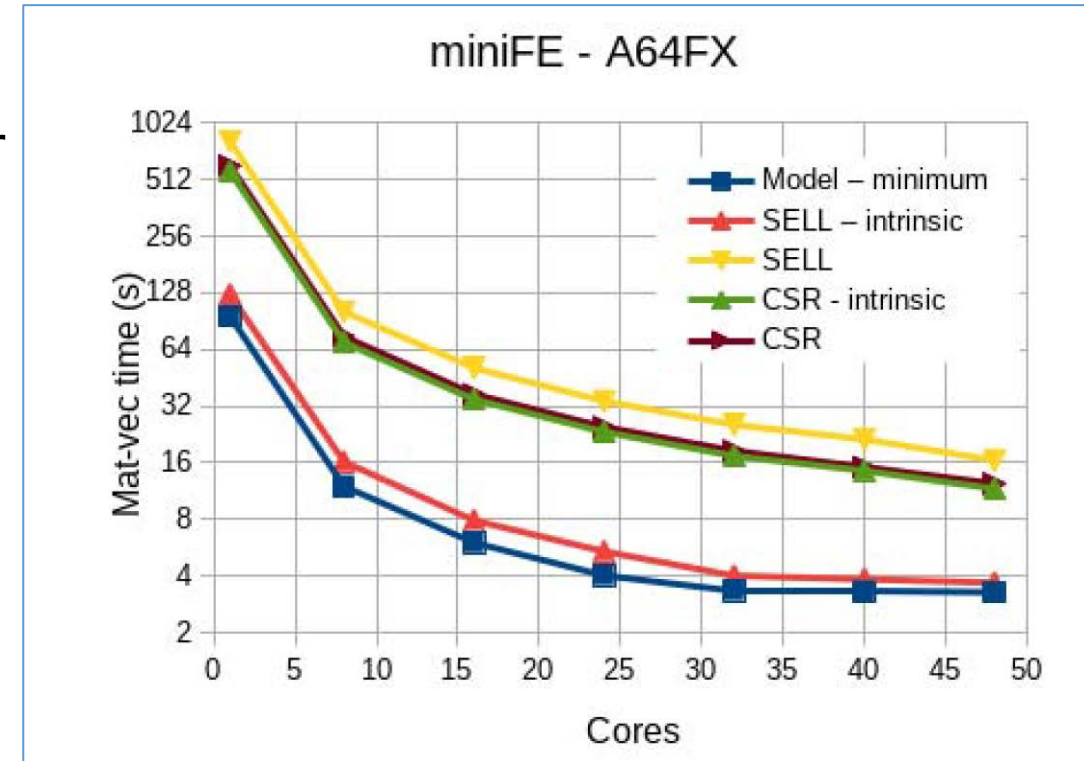
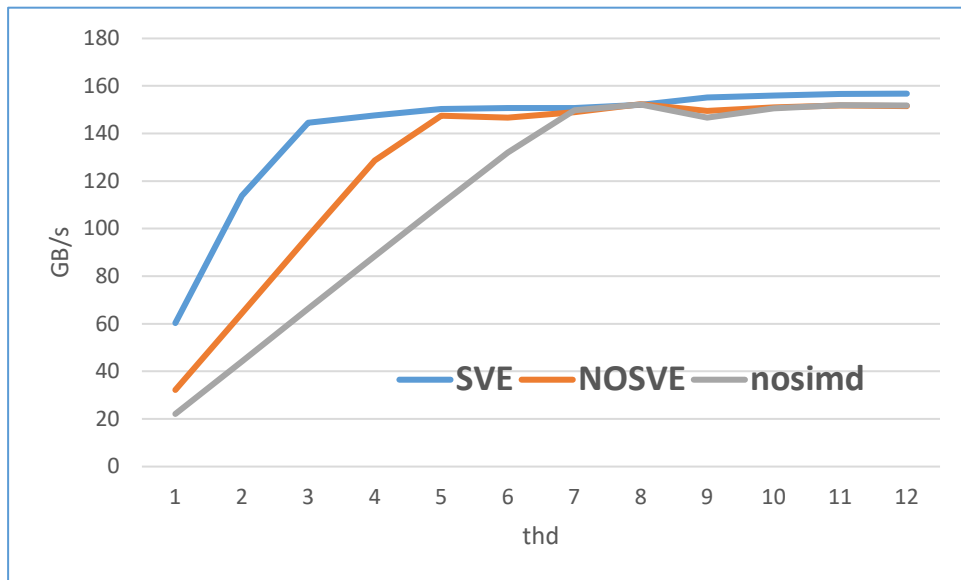
- A64FX performance is less than Thx2 and Intel one
- We found low vectorization (SIMD (SVE) instructions ratio is a few %)
- We need more code tuning for more vectorization using SIMD



How to improve the performance of sparse-matrix code

- **Storage format is important:**
 - Sliced ELLPACK format shows significantly better performance than CSR, but only when it is vectorized manually using intrinsics."
 - CSR is not good even with manual vectorizing.
- **Vectorizing with SVE is important to get memory bandwidth.**

Memory bandwidth with hardware prefetch

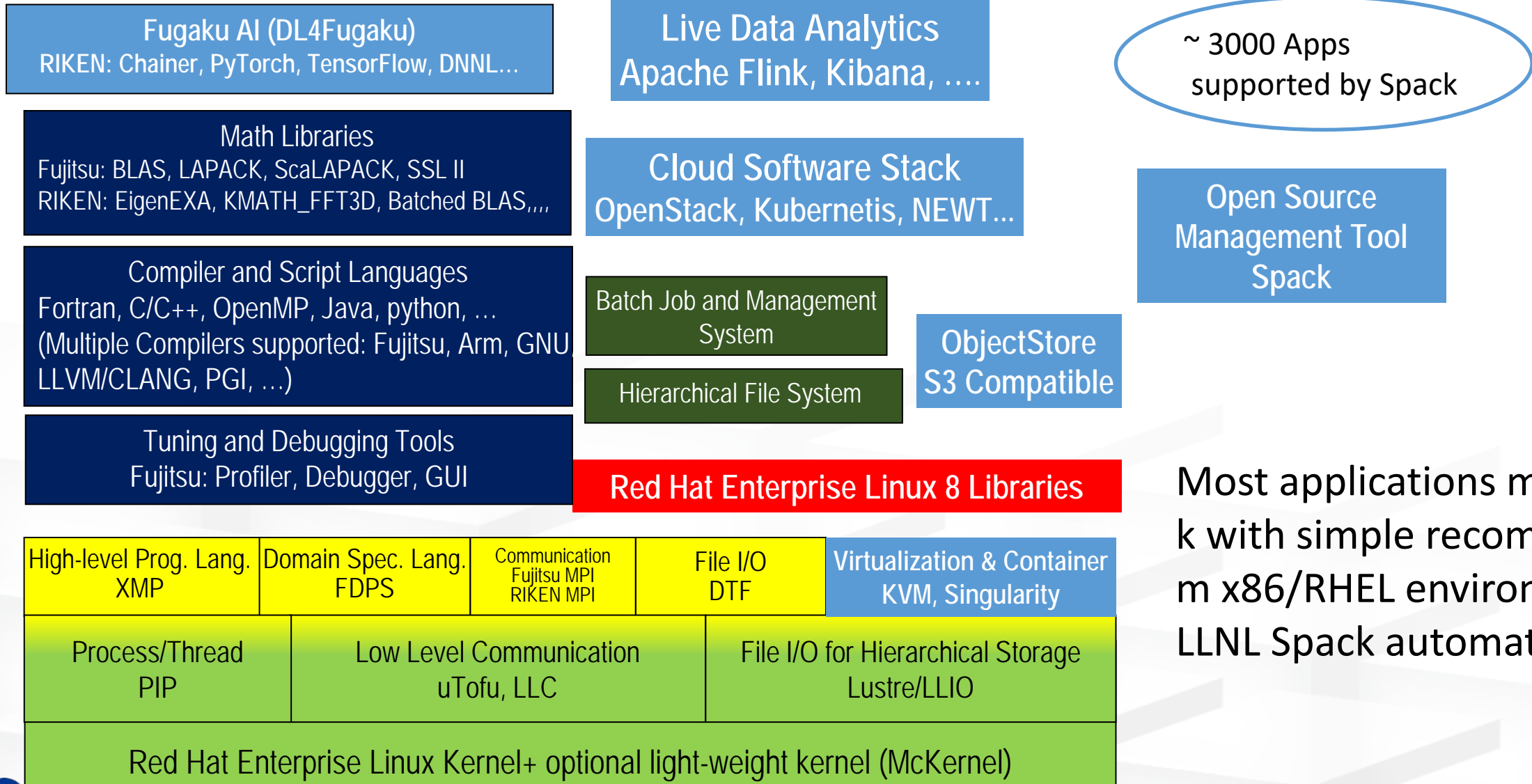


B. Brank, S. Nassyr, F. Pouyan and D. Pleiter, "Porting Applications to Arm-based Processors," EAHPC Workshop, *IEEE CLUSTER 2020*, Kobe, Japan, 2020, pp. 559-566, doi: 10.1109/CLUSTER49012.2020.00079.

Summary of A64FX performance characteristics

- For core-to-core comparison in intspeed, integer performance is $\frac{1}{4}$ of Xeon
- For chip-to-chip comparison in SPEC OMP, 48 threads performance of one chip is 65% to one chip of recent high-end Xeon (Cascade Lake)
 - **NOTE:** Performance of memory-intensive benchmarks is extremely good in A64FX thanks to HBM.
- For some scientific workload, the almost same performance to dual sockets of Xeon with half of power consumption (UK benchmark and HPC OSS)
- High SIMD rate is important to get performance
 - Need to tune memory access pattern
 - We found many benchmark programs are not well-vecterized.
- Power efficiency of A64FX is very good (double efficiency than Xeon?)

Fugaku System Software Stack



Most applications may 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.
 - 4 compilers (Fujitsu, gcc, LLVM/Arm, Cray), OpenMP 4.x is supported.
 - uTofu low-level comm. APIs for Tofu-D interconnect.
- Container and Virtual machine (KVM, Singularity, ...)
- DL4Fugaku: AI framework for A64FX and Fugaku, used in Chainer, PyTorch, TensorFlow
- Many Open-source software are already 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.
 - XscalableMP directive-based PGAS Language
 - FDPS: DLS for Framework for Developing Particle Simulators.
 - EigenExa: Eigen-value math library for large-scale parallel systems.

Performance Tuning for A64FX processor

- **HPC-oriented design**

- Small core \Rightarrow 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)**

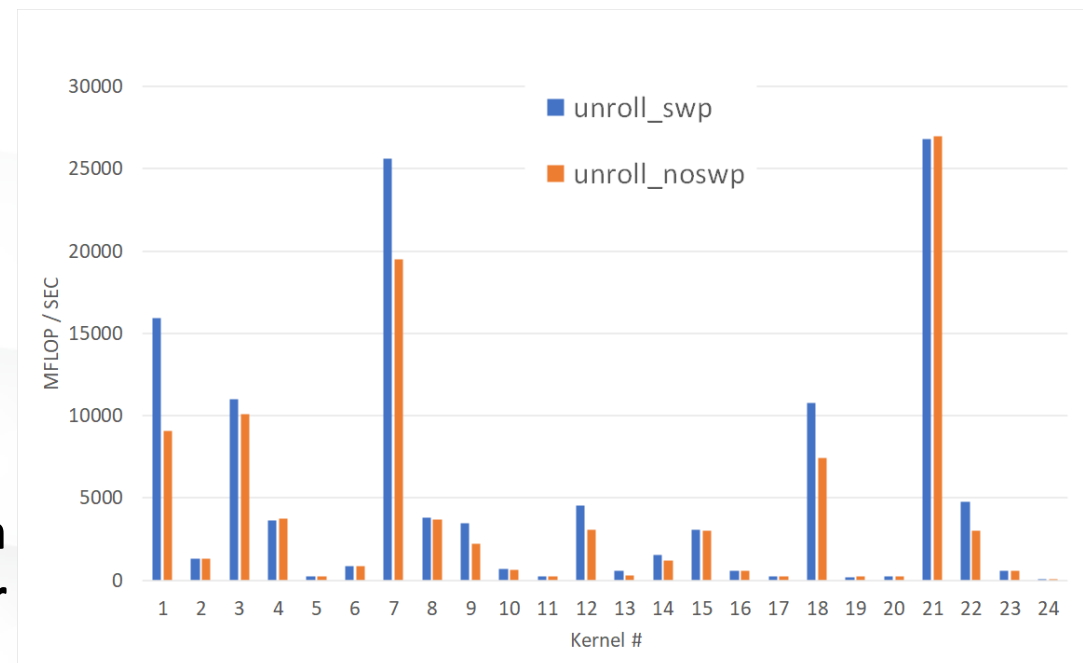
- SWP: software pipelining
 - \sim 20% speedup in Livermore Kernels
- Automatic and Manual loop fissions

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

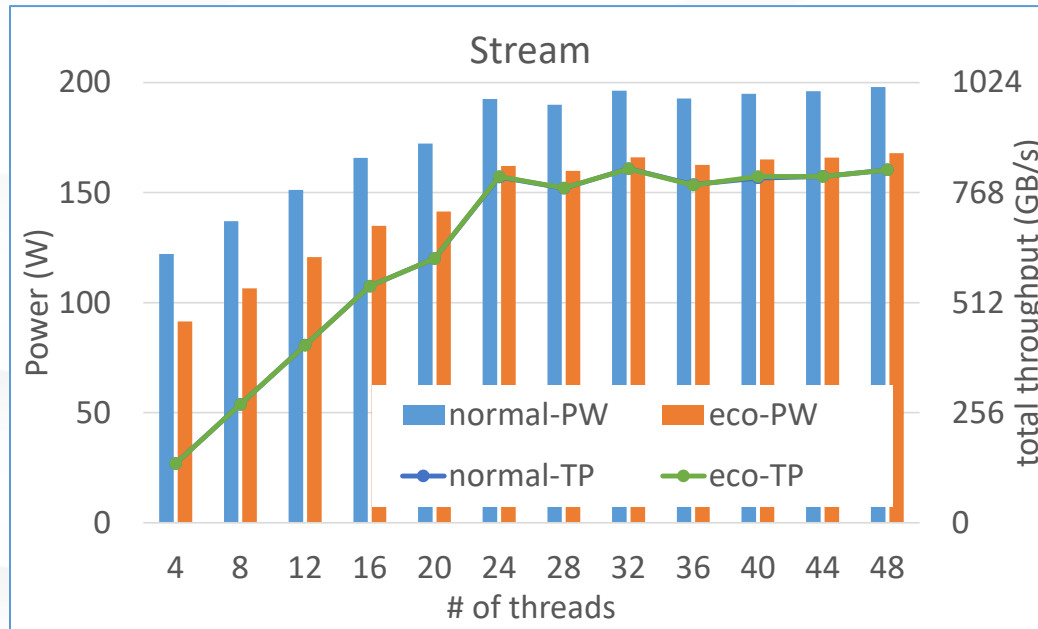
Performance improvement by SWP in Livermore Kernels by Fujitsu compiler



Evaluation power mode: Boost mode (2.2GHz) & Eco mode (1 SIMD pipeline)

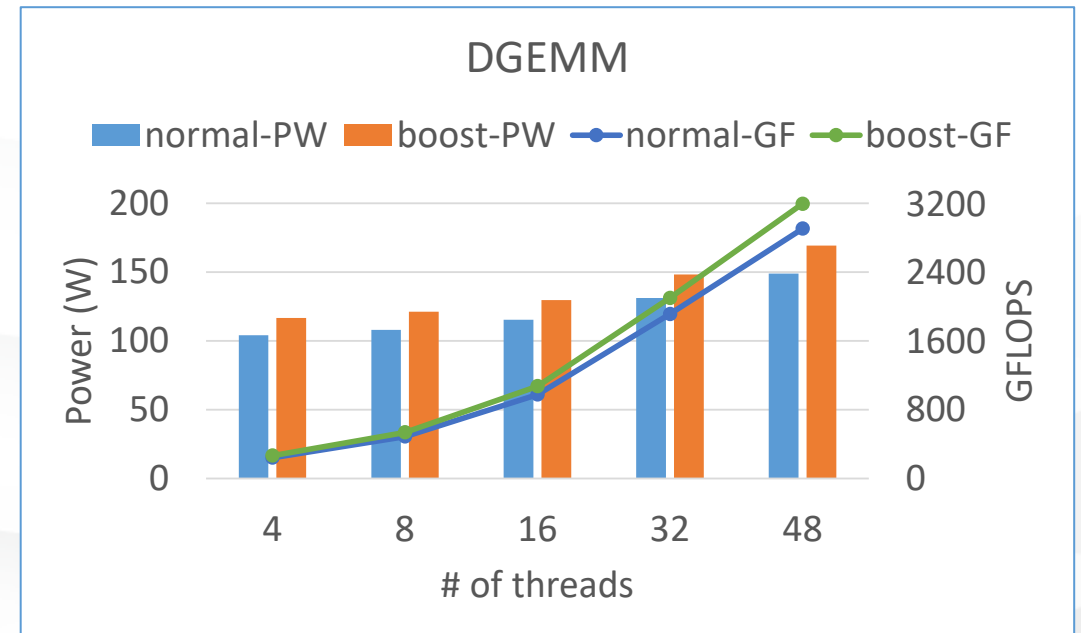
Power & Performance of STREAM using Eco mode

- 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 %



「京」から「富岳」への進歩

	京	富岳	比
ノードのコア数	8	48	
半導体技術 (nm)	45	7	
コア当たりの性能(GFLOPS)	16	64 (70)	4(4.4)
ノード当たりの性能. (TFLOPS)	0.128	3.07 (3.37)	24 (26.4)
ノードのメモリバンド幅(GB/s)	64	1024	
ラック当たりのノード数	96	384	4
ラックの性能(TFLOPS)	12.3	1180 (1297)	95(105)
システム全体のノード数	82,944	158,976	
システム性能 (PFLOPS)倍精度	10.6	488 (537)	42.3 (52.2)
単精度	10.6	977 (1070)	84.6 (104.4)

「ノード」とは、ネットワークにつながるコンピュータの単位
 富岳の場合は、ノードは1チップからなる。1チップは、複数の
 コア（コンピュータ）を内蔵する。

No.1 in Green500 at SC19!

Announce from
Fujitsu at SC19



Green500, Nov. 2019

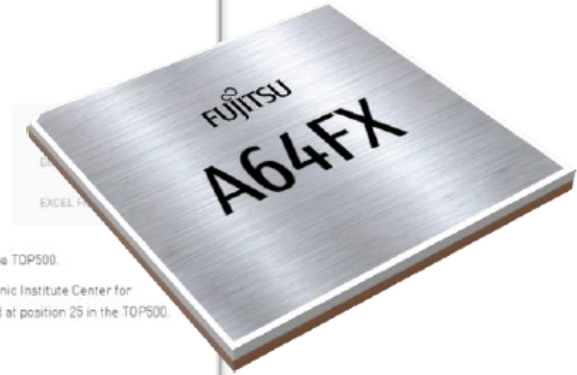
A64FX prototype –
Fujitsu A64FX 48C 2GHz
ranked **#1** on the list

768x general purpose A64FX
CPU w/o accelerators

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

The Green500 website screenshot shows the November 2019 list. The top entry is the Fujitsu A64FX prototype, ranked #1. The table below shows the top 5 systems.

Rank	Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	159	A64FX prototype - Fujitsu A64FX, Fujitsu A64FX 48C 2GHz, Tofu interconnect D, Fujitsu Fujitsu Numazu Plant Japan	36,864	1,999.5	118	16.876
2	420	NA-1 - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz, PEZY Computing / Exascale Inc, PEZY Computing K.K. Japan	1,271,040	1,303.2	80	16.256
3	24	AIMOS - IBM Power System AC922, IBM POWER9 20C 3.43GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Volta GV100, IBM Rensselaer Polytechnic Institute Center for Computational Innovations (CCI) United States	130,000	8,045.0	510	15.771
4	373	Satori - IBM Power System AC922, IBM POWER9 20C 2.40GHz, Infiniband EDR, NVIDIA Tesla V100 SXM2, IBM MIT/MGHPC Holyoke, MA United States	23,040	1,464.0	94	15.374
5	1	Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	10,096	14.719



TOP500, HPCG, HPL-AI, Graph500 の4冠

2020年6月時点

	ノード数	周波数 (GHz)	測定値	ピーク性能	効率	使用ノード数の割合	第2位との比較	
							性能	性能差 (倍率)
TOP500	152,064	2.2	415.53 PF	513.85 PF	80.9%	96%	148.60 PF	2.8
HPCG	138,240	2.2	13.36 PF	467.14 PF	2.9%	87%	2.92 PF	4.6
HPL-AI	126,720	2.0	1.42 EP	1.55 EP	91.3%	80%	0.55 EF	2.5
Graph500	92,160	2.2	70.98 Teps			58%	23.75 Teps	3.0

備考：性能値は小数点第3位以下切り捨て

● TOP500

- 密行列係数連立1次方程式をLU分解法で求解するプログラム(Linpack)を使用

● HPCG

- 疎行列係数連立一次方程式を共役勾配法 (conjugate gradient) で求解するプログラムを使用

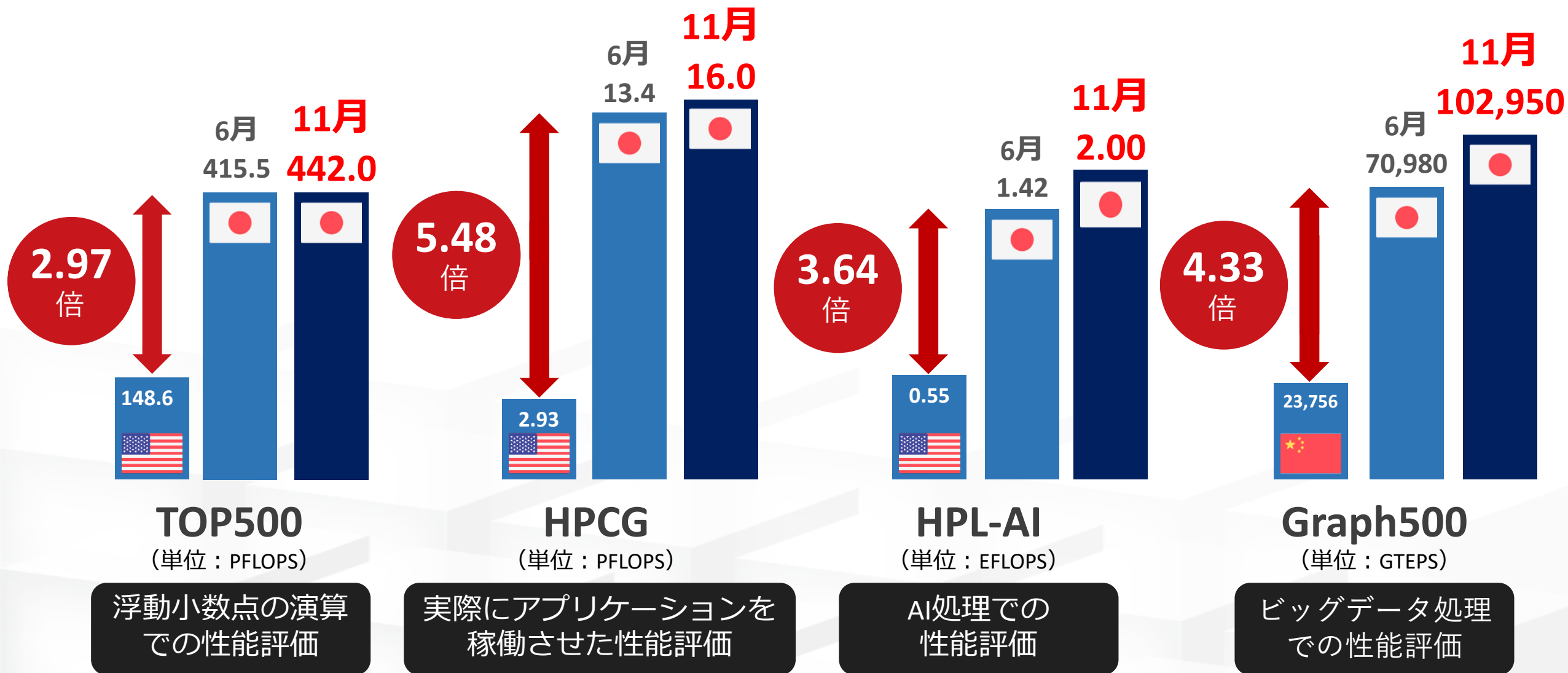
● HPL-AI

- 密行列係数連立1次方程式をLU分解かつ混合演算 (倍精度浮動小数点(FP64)、半精度浮動小数点(FP16))によって求解するプログラムを使用

● Graph500

- グラフ探索問題

「富岳」全系によるベンチマークテスト結果（6月の結果との比較）



2位に対して、3倍から5.5倍近い性能差を実現

目標は達成できたか？



Target applications

Performance relative to the K computer

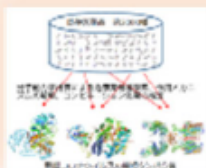
Power consumption

アプリケーション	利用形態	問題規模	ノード数/ジョブ	性能倍率	消費電力
GENESIS	多重	92,224原子	1	131倍	22 MW
GENOMON	多重	リード長150、14億リード(ペアードエンド)	96	23倍	20 MW
GAMERA	大規模単一	1兆自由度	147,456	63倍	21 MW
NICAM+ LETKF	大規模単一	全球3.5kmメッシュ、1024メンバENS同化	131,072	127倍	22 MW
NTChem	多重	720原子、19,680原子軌道	17,820	70倍	26 MW
ADVENTURE	多重	16.5億自由度	4,096	63倍	28 MW
RSDFT	多重	110,592原子、221,184バンド	10,368	38倍	30 MW
FFB	大規模単一	6,748億要素	158,976	51倍	29 MW
LQCD	大規模単一	192^4格子	147,456	38倍	20 MW

- ターゲット・アプリケーションの性能は達成できた。
- 電力については、設計時の想定よりも低くなった。
- 20-30MWでの運用を想定

昨年度、共用前に前倒して実施されたコロナ対策のための計算科学

「富岳」による 新型コロナウイルスの治療薬候補同定

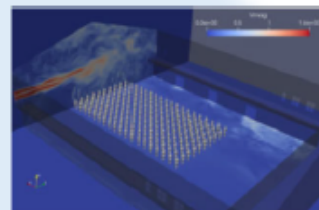


分子動力学計算により、約2000種の既存医薬品の中から、新型コロナウイルスの標的タンパク質に高い親和性を示す治療薬候補を探索・同定する。

(課題代表者；理化学研究所/京都大学 奥野 恭史)

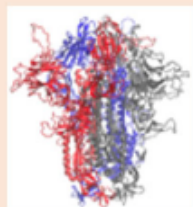
室内環境におけるウイルス飛沫感染の 予測とその対策

通勤列車内、オフィス、教室、病室といった室内環境において、新型コロナウイルスの特性を考慮した飛沫の飛散シミュレーションを行い、感染リスク評価を行った上で、感染リスク低減対策の提案を行う。



(課題代表者；理化学研究所/神戸大学 坪倉 誠)

「富岳」を用いた新型コロナウイルス 表面のタンパク質動的構造予測



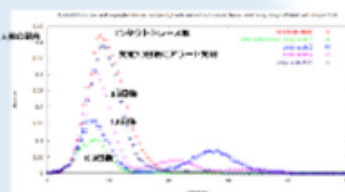
クライオ電子顕微鏡によって解かれたウイルス表面タンパク質の立体構造を初期モデルとして、その立体構造の動きを「富岳」を用いた分子動力学計算で予測する。

(課題代表者；理化学研究所 杉田 有治)



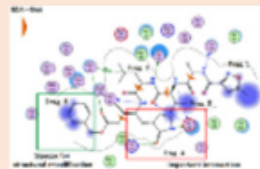
パンデミック現象および対策の シミュレーション解析

今後生じうる社会経済活動への影響を評価し、収束シナリオとその実現方法を探る。あわせてウイルスの変異などにより感染・発病の経過が変化した場合に起こりうる事象への対応を立案する。



(課題代表者；理化学研究所 伊藤 伸泰)

新型コロナウイルス関連タンパク質に対する フラグメント分子軌道計算

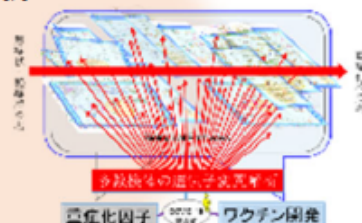


新型コロナウイルス関連タンパク質に対するフラグメント分子軌道計算を系統的に実施し、詳細な相互作用解析を行う。

(課題代表者；立教大学 望月 祐志)

新型コロナウイルス感染症重症化 に関するヒト遺伝子解析

新型コロナウイルスの重症化例および軽症ないし無症状感染例について、全ゲノムシーケンスを用いた解析を実施し、スパコンシミュレーションによる重症化リスク関連遺伝子変異を同定する。



(課題代表者；東京医科歯科大学 宮野 悟)

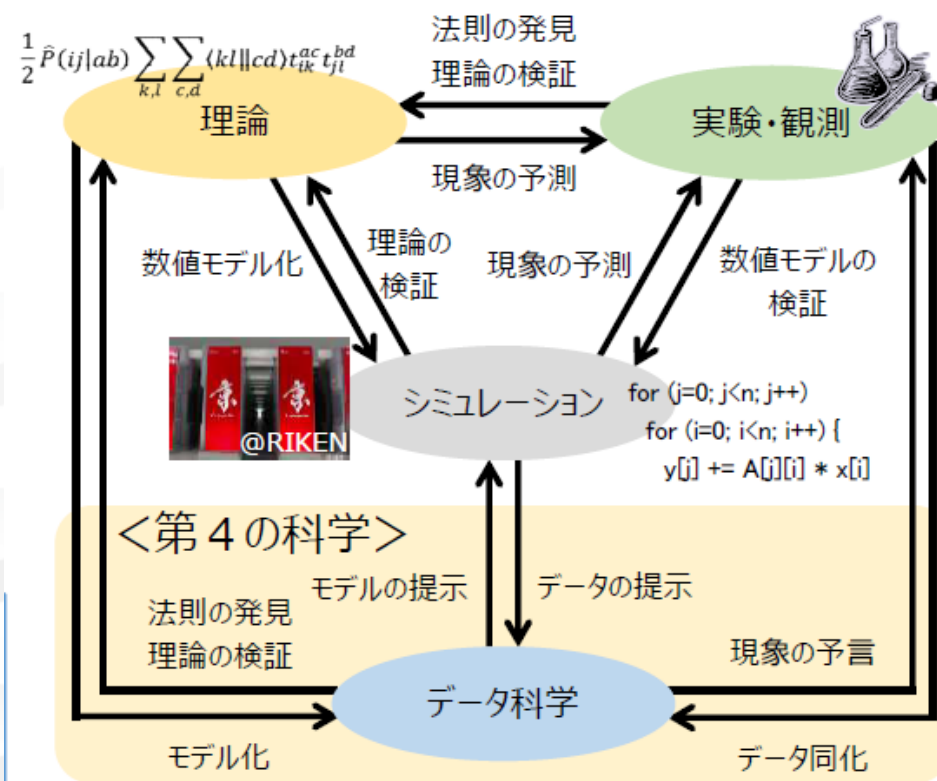
富岳でのアプリケーション

- 京では、システム全体でなにができるか（capability）を目標としていた
- 富岳では、京で数ジョブしか同時に動かすことができなかった大規模アプリを数十から数百ジョブを同時に動かすことができるようになる。（capacity）
- 実際の科学の研究では、いろいろなケースを試してみることが大切

- シミュレーションと
データ科学の融合

- AIを加速する仕組みも！

文科省のホームページより http://www.mext.go.jp/b_menu/shingi/chousa/shinkou/020/shiryo/_icsFiles/afieldfile/2018/09/18/1409147_4_1.pdf



- ディープラーニングでは、大量のパーティンを読み込ませて、「学習」する必要があるために、大量の計算が必要。
- アプリごとに別の学習が必要



スパコンやGPUが必要。

- 富岳は、ディープラーニングの学習のための演算を高速化する命令がある。
- また、ディープラーニングのニューラルネットワークが大きくなると並列処理が必要

● 科学技術計算とAI (ディープ・ラーニング)

- 時間のかかる計算の代わりに、あらかじめ学習をしておいたディープ・ラーニングで答えを求める
- 学習のデータが足りない場合に、シミュレーションした結果を学習に使うことが期待されている。

「富岳」の開発を振り返って

- プロジェクトリーダーの石川さんに感謝
- RWCP新情報処理開発機構からの経験が生きた、と思う
- このプロジェクトは、「開発」プロジェクト
 - 「研究」プロジェクトとは、まったく違う。最後にはシステムを稼働する必要がある。
- 目標設定は重要，プロジェクトの存在理由となる
 - 目標が達成できて、よかった。
- 7年は長い
 - 基本設計(2014-2015)時に、2019-2020年で使えるテクノロジー（例えば、7nmの半導体）を予想して、設計しなくてはならない。
 - 実は、10nmを使う予定で、7nmに変更して、
 - が、チップを作るのは設計が完了して(2017年) シリコンがでてくるまで、おおよそ、2年かかる。
- 「京」の反省点（海外にうれなかった）のは、解消された（？）

- **そもそも、CSの研究をしている人はスパコンには興味がない（なかった…）**
- **システム（CS）の研究と、それを使うアプリ側の研究のずれ**
 - 一緒に、研究しましょう、というのは理想だが、…
 - 結局、研究中のものは、アプリは使わない。
 - つかってもらうCSの研究はむずかしい。
 - post-Peta CRESTの例：アプリに近いものはいいが、基礎的なところはむずかしい。
- **そもそも、富岳の開発プロジェクトは、研究ではなかった。**
- **A64FXはまだまだ、ソフトの開発の余地があるので興味のある方はよろしく。**
- **また、クラウドとかデータ処理とか、計算科学のニーズも多様化している。**

- **我が国のプロセッサ開発の将来は？**
- **将来の若手の方に期待！**