

「ポスト京からさらに未来のHPCとAIの融合」



理化学研究所 計算科学研究センター
センター長 松岡 聡

文部科学省 ポスト「京」の利活用促進・成果創出加速に関するワーキンググループ



20181108

2018 4/1 より理研計算科学センター長に 計算の 計算による 計算のための科学

理化学研究所 計算科学研究センター
(R-CCS)

計算の科学

「京」やポスト「京」を支えるプログラミング手法・ソフトウェア・運用技術や、ビッグデータ・AIなどへの対応を実現する手法など、様々な技術のベースとなる高性能計算の本質に関する研究を推進

計算による科学

生命科学、工学、気象・気候、防災・減災など私たちの生活に直結し、国民の関心事の高い最先端の研究開発に欠くことのできない、基礎研究や応用研究を「京」やポスト「京」を用いて推進

シナジー、融合

国内外の各機関、大学、企業の研究者との協力を更に拡大

世界中からトップ研究センターと認知される
国際的な「計算のための科学」の知の拠点



1. コ・デザインによる京から受け継ぐシミュレーションでの優位性

- ◎ アプリにおける高演算性能：京と比較で最大100倍以上の性能向上
- ◎ 重点・萌芽課題における多くの科学的成果創出の準備・期待
- ◎ 高性能の達成と、容易なプログラミング・高実用性の両立

**マシン自身だけでなく
半導体やITのテクノロジー
でも世界をリード**

2. ポスト京/A64fx の新たな技術イノベーション

◎ チップ自身の高演算性能（高速メモリ→高メモリバンド幅）

CPUチップ単位で、多くのHPC & Society 5.0アプリで従来CPU数倍の性能

◎ 高い省電力・グリーン性能

「パワーノブ」含む省電力技術導入に、汎用CPUでは世界トップレベルの高効率

◎ Arm「エコシステム」の充実

年間30億個生産されるArmプロセッサの命令セットを採用

クラウドにも展開が容易、今回成果のSVEのグローバルスタンダード化

◎ Society 5.0アプリへの展開

HPC, ビッグデータ, AI, CAE/EDA, 次世代コンピュータセキュリティ・

ブロックチェーンなどのアプリにおいて、GPUに匹敵する性能を発揮すると期待。



ARM：スマホ等の機器、
車載チップ等で
巨大なエコシステム

Society 5.0への適用 → シミュレーション・ビッグデータ・AIの加速

更にクラウド等のインフラ、および技術としても世界に展開普及も

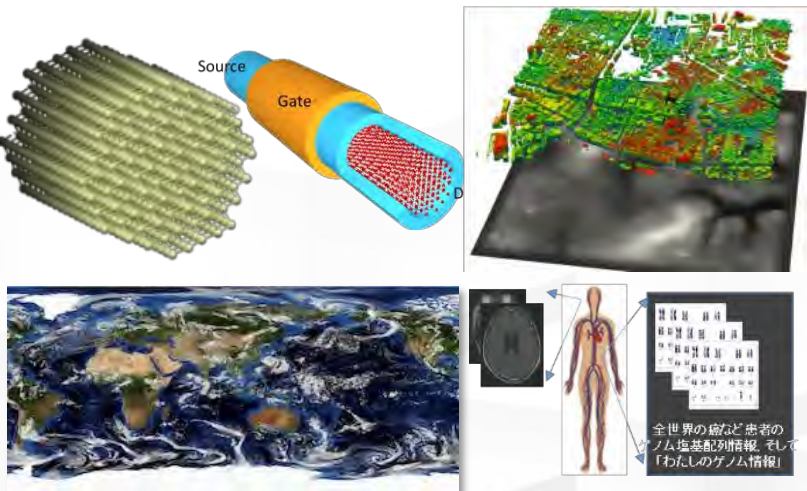
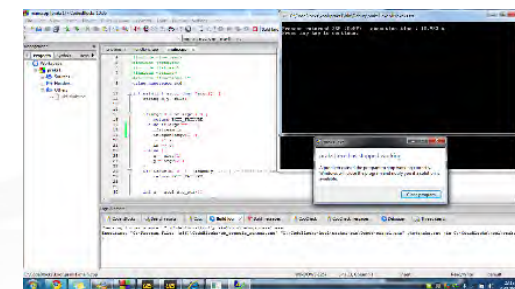
ポスト「京」におけるコ・デザイン

システムとアプリケーションの協調的な開発
手法 (Co-design)

「計算による科学」

● 重点9課題：アプリケーション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様々な社会的・科学的課題を選定

「計算の科学」



各種アプリケーションの様々な計算上の特性を一般化・特徴化

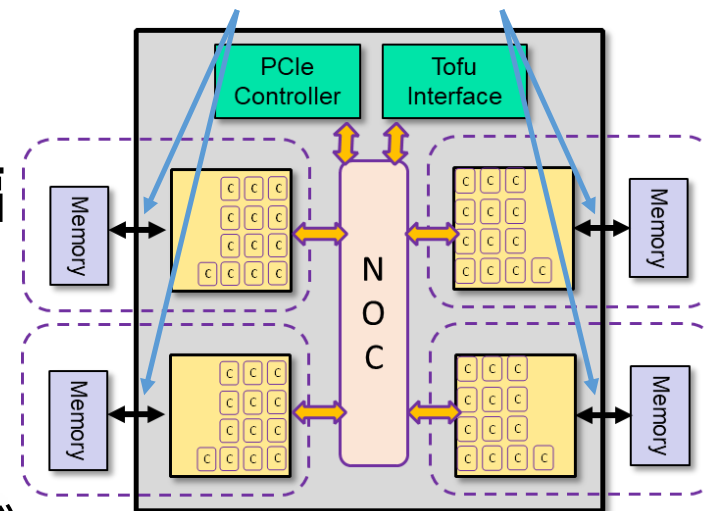
各アプリ特性を加味したシステム設計、及びそれに適したアプリの最適化

- 理研CCS(IBM AICS)と開発企業、及び重点課題機関の密な研究開発体制
- コ・デザインにより多くのアプリから9つの代表的な「ターゲットアプリケーション」の選定
- 複数のターゲットアプリで、コ・デザインの結果、京と比較して**100倍以上**の性能向上
- プログラミングを含む使い易さの向上・高い電力性能などの技術成果も

1. 高いメモリバンド幅によるHPCアプリの高演算性能

- 種々のベンチマークから、多くのHPCアプリケーションはメモリ律速
 - ポスト京：FIBREベンチマーク、DoE ECP ベンチマーク
 - メモリコントローラのビジー率の高さ
- ハイエンドXeon比較でポスト京プロセッサが圧倒的メモリバンド幅
 - Xeon: 6~8チャンネル DDR4、~100GB/s (STREAM)
 - ポスト京: 4 Stack HBM2 2Ghz, 1TB/s(スペック), ~840GB/s (STREAM)
 - CPUとして世界初の採用、バンド幅・レーテンシ・コヒーレンシをすべて満たす世界トップクラスのメモリコントローラ的设计
- 最新・ハイエンドのIntel Xeon Platinum 比で、計算で2倍、メモリでは8倍の性能、非常に省電力
 - Volta GPUクラスのパフォーマンスや電力性能比
 - メモリ周りの最適化：単純にHBMをボルトオンしたのではない
 - 大規模HPCだけでなく、ビッグデータ・AI・セキュリティなどSociety 5.0 アプリで高性能

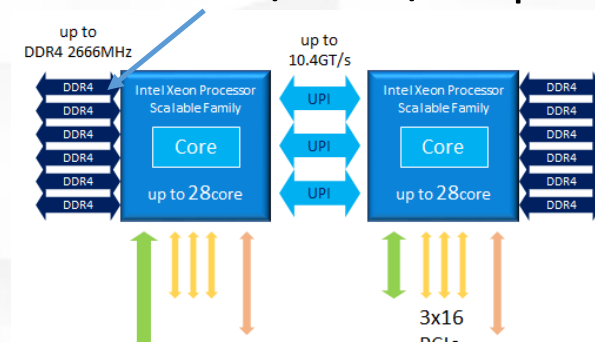
PostK Proc: HBM2
256GB/s x 4



Skylake Xeon

6ch DDR4 2666

21.3GB/s x 6 / chip



③ポスト「京」の電力性能等の向上に向けた取組は適切か。

【前回指摘事項】

- アプリケーションの実行速度や消費電力の最適化を可能とするエコモード等の複数のモードの導入や、アプリケーションの特性に応じた消費電力のコントロールを可能とするパワーノブAPI※の導入等、運用における電力削減に向けた取組を実施している。
※API : Application Programming Interface
- 性能を重視しつつ最先端の省電力設計により、CPUとして世界最高水準の電力性能効率を達成し、全体では開発目標である30～40MW以下の消費電力を達成する見込み。

【消費電力の低減の取組】

- ・エコモード：利用する演算回路を半分にすることによって、底上げ電力を半分に低減するモード。
⇒ メモリバンド幅が律速するアプリケーション開発において、大きな電力削減効果。
- ・リテンションモード：ジョブを割り当てられていないノードをアイドル状態ではなく、節電状態にすることにより、消費電力を約半分に削減するモード。
⇒ ノードが使われない時の電力を大幅に削減。
- ・パワーノブAPI：プログラム及びジョブスケーラから、プログラムの特性に応じた電力制御を可能とし、運用時の電力効率を向上。

3. 中間評価以降の成果予定と目標の達成見込み

①ポスト「京」が世界の他のシステムに対し総合力で卓抜するものであることが示されているか。

【前回指摘事項】

- 世界の他の汎用的なシステムと比較して、消費電力当たりの性能が世界最高水準であることが見込まれる。
- 計算能力を左右する要素であるCPUの演算性能やメモリバンド幅、ノード間通信性能について、汎用的なシステムとしてはそれぞれ世界最高水準であることが見込まれる。

【最新のチップ等との比較】

	ピーク性能 (倍精度浮動小数点演算)	メモリバンド幅 (STREAM triadベンチマーク)	Linpack効率	電力性能	ネットワーク性能 (Interconnect)
ポスト「京」/A64fx	2.7 TFlops以上	840 GB/sec	85 %以上	約15 GFlops/Watt	40.8GB/sec
Oakforest-PACS/Xeon Phi KNL	3.0464 TFlops	490 GB/sec	54.4 %	4.986 GFlops/Watt	12.5 GB/sec ^{※3}
Niagara/Xeon Skylake ^{※1}	1.536 Tflops	104.5 GB/sec	66.7 %	4.546 GFlops/Watt	6.3 GB/sec ^{※3}
Summit/GPU Volta GV100 ^{※2}	7.8 Tflops	855 GB/sec	65.2 %	13.889 GFlops/Watt	4.2 GB/sec ^{※3}
DGX-1 SaturnV Volta/GPU Tesla V100 ^{※2}	7.8 Tflops	855 GB/sec	58.8 %	15.113 GFlops/Watt	6.3 GB/sec ^{※3}

※1 Skylake (Xeon Gold 6148 20C 2.4GHz) の1 CPUあたりの性能は2ソケット版性能公開情報に基づき、1ソケット性能を計算。

※2 NVLINK接続におけるGPU 1ソケットのピーク性能。メモリバンド幅はGPU 1ソケットあたりの性能。

※3 ネットワークコントローラをチップに内蔵しておらず、100Gbps (12.5GB/sec) の外付けInfiniband等を有する。

Niagaraの場合CPUが2ソケットに100Gbps Infinibandを1つ搭載、Summitの場合GPUが6ソケットに100Gbps Infinibandを2つ搭載、DGX-1 SaturnV Voltaの場合GPUが8ソケットに100Gbps Infinibandを4つ搭載、それぞれのネットワーク性能はソケット当たりの値を計算 (小数点第2位四捨五入)。

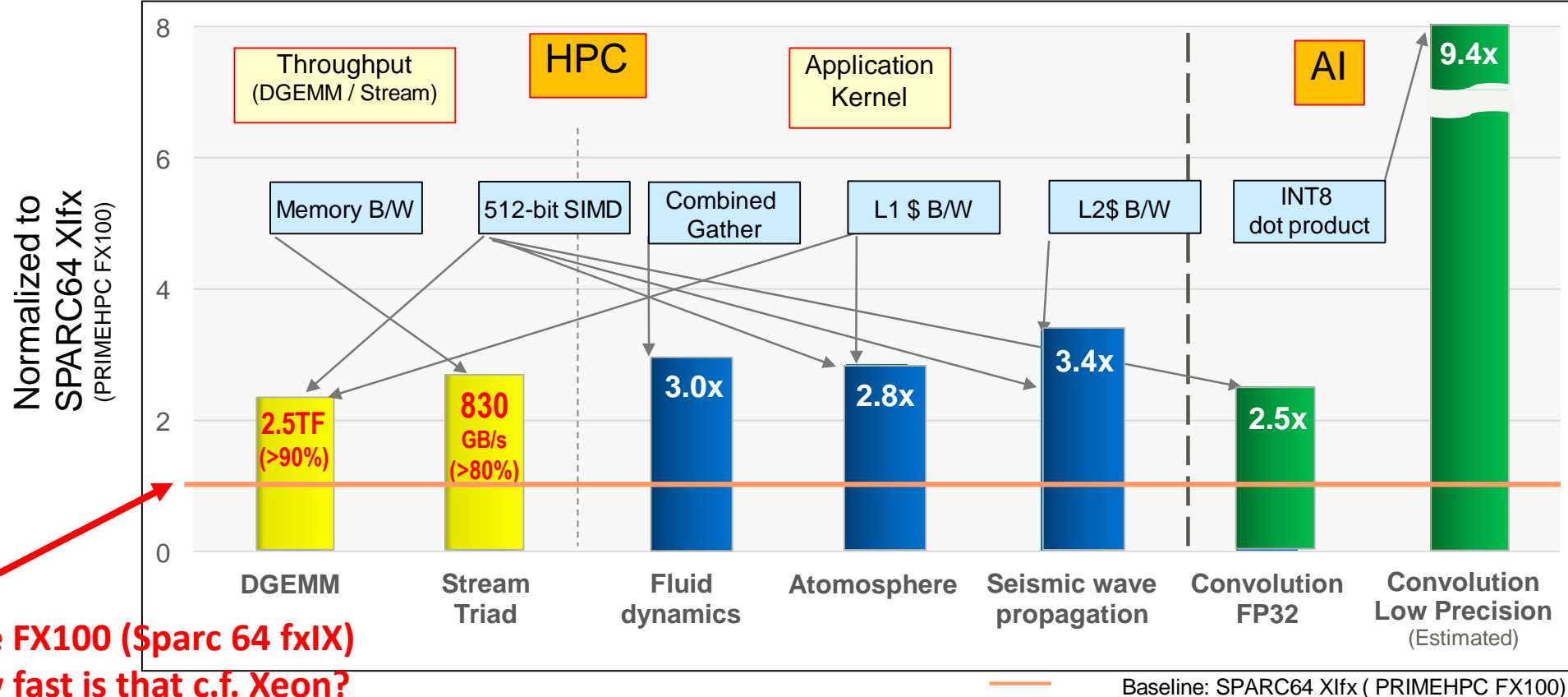
メモリバウンドな多くのアプリにおいては、それらの電力性能はGPUを上回る期待もある

Performance

■ A64FX boosts performance up by microarchitectural enhancements, 512-bit wide SIMD, HBM2 and process technology

- > 2.5x faster in HPC/AI benchmarks than SPARC64 Xlfx (Fujitsu's previous HPC CPU)
- The results are based on the Fujitsu compiler optimized for our microarchitecture and SVE

A64FX Benchmark Kernel Performance (Preliminary results)



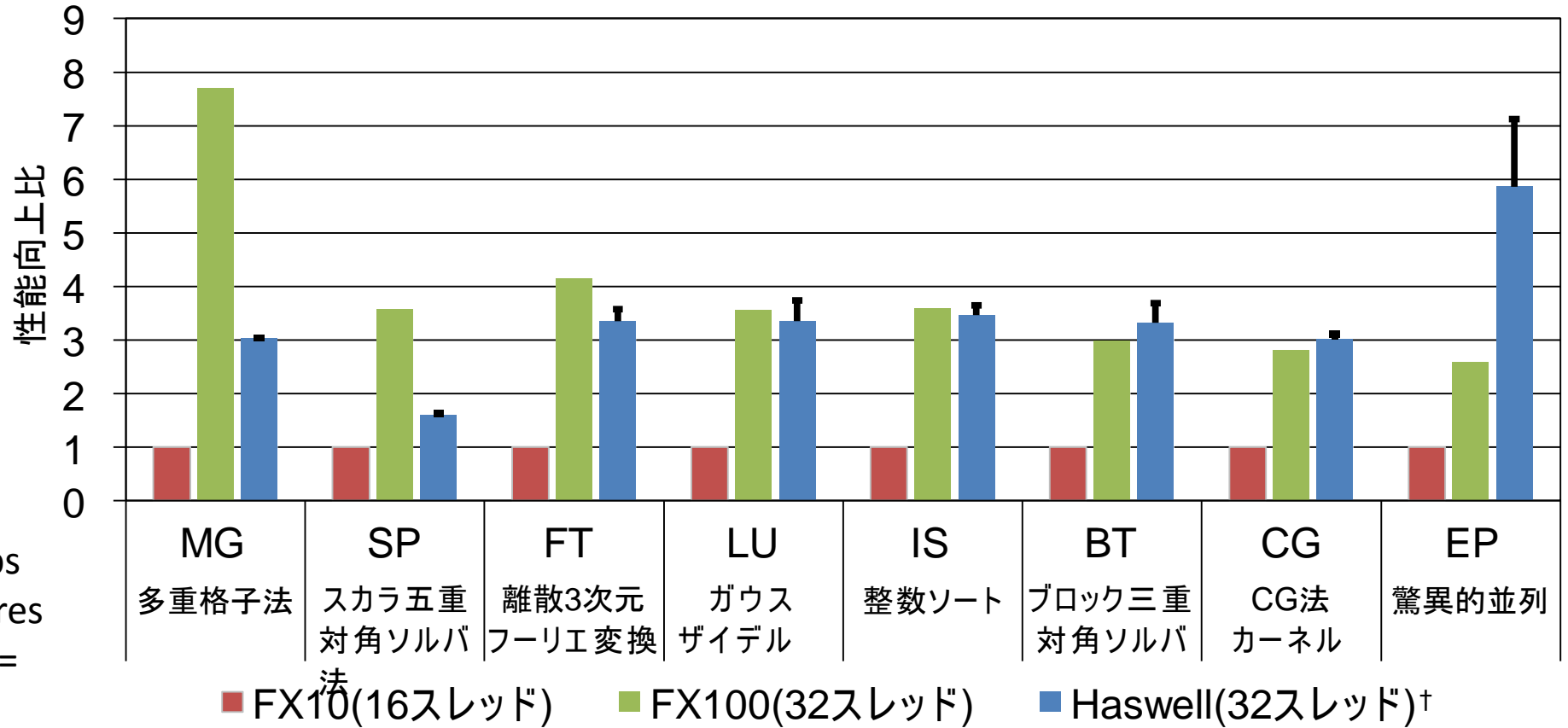
Baseline FX100 (Sparc 64 fxIX)
But how fast is that c.f. Xeon?

NAS Parallel Benchmark of FX100

[Slide by Ikuo Miyoshi, Fujitsu, SSKen2015]

■ OpenMP版を用いてノードあたり演算性能を評価

FX10に対するノードあたり性能向上比



[†] エラーバーはCPU周波数を1.9GHzに固定しない場合の性能を表す

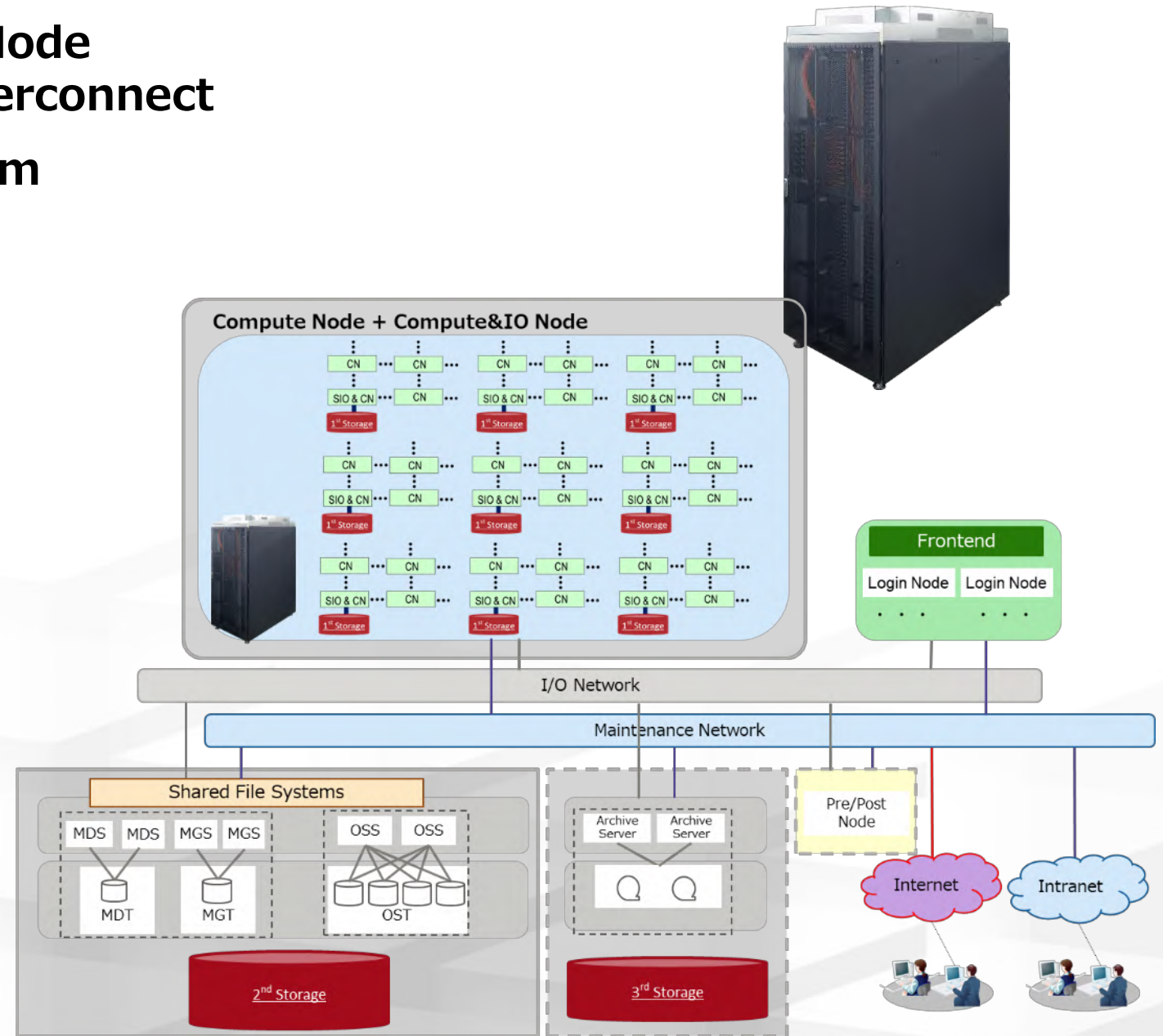
FX100は、FX10比平均3.9倍、Haswell比平均1.3倍のノードあたり演算性能

使用コード: NAS Parallel Benchmarks Ver. 3.3.1 OpenMP版 クラスC

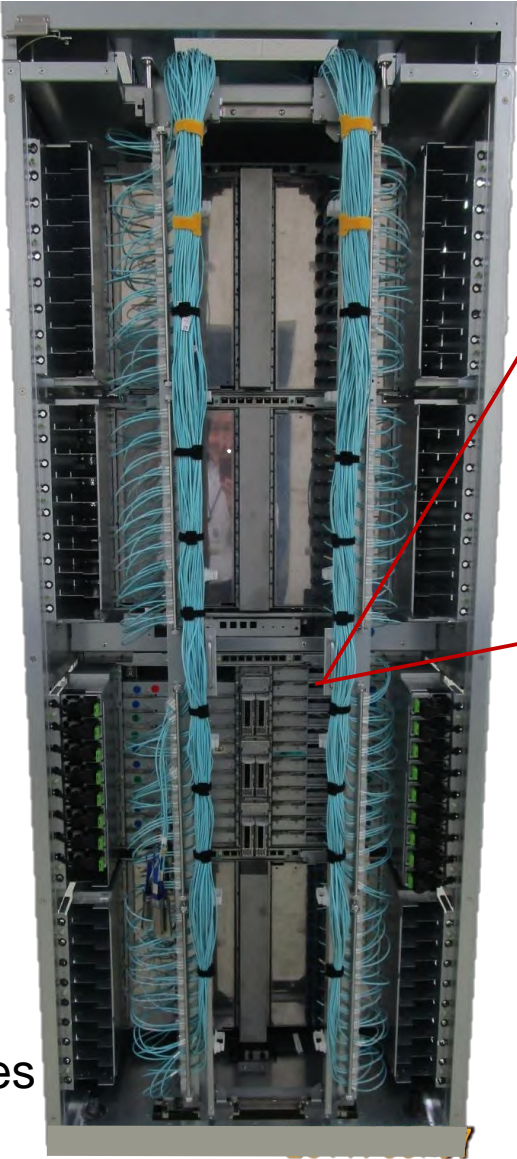
Note: Haswell:
 1 node = 2 chips
 32 threads&cores
 FX100: 1 node =
 1 chip
 32 threads&cores

An Overview of Post-K Hardware

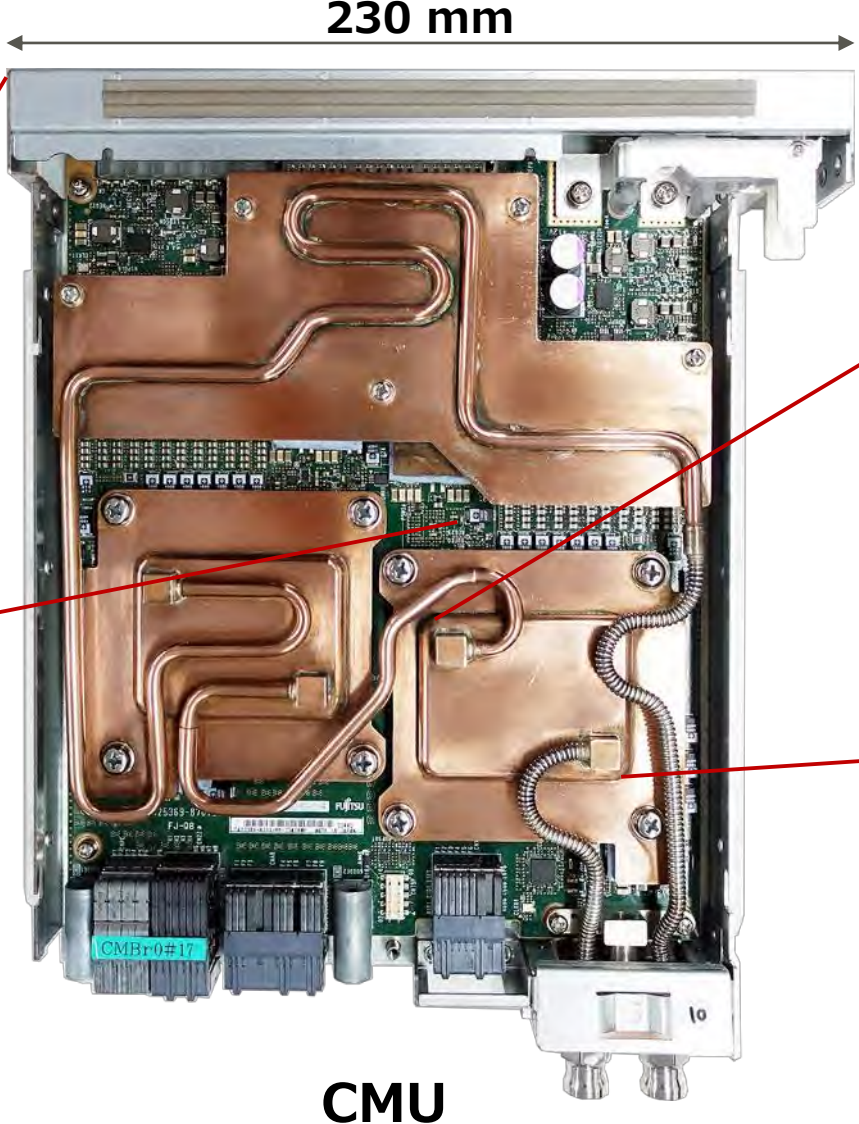
- Compute Node, Compute + I/O Node connected by 6D mesh/torus Interconnect
- 3-level hierarchical storage system
 - 1st Layer
 - Cache for global file system
 - Temporary file systems
 - Local file system for compute node
 - Shared file system for a job
 - 2nd Layer
 - Lustre-based global file system
 - 3rd Layer
 - Storage for archive
- **>100,000 nodes**
- **Approaching 10 million CPU cores**



Post-K Chassis, PCB (w/DLC), and A64fx CPU Package



W 800mm
D1400mm
H2000mm
384 nodes



CMU



60 mm

280 mm

60 mm

CPU Package

**A0 Chip Booted in June
Undergoing Tests
B0 underway**

2. ARM サーバ&HPC エコシステムの充実

- 年間のプロセッサの生産量：x86 3億個 vs. ARM 30億個

- サーバチップのハードウェアエコシステムの確立

今回新たにSVE (Scalable Vector Extension)を世界初めて提案・実装。

これはARM v8の正式プロファイルであり、v9では組み込みなどへも普及

本プロジェクトの成果がスパコン・クラウドからIoTまでグローバルデファクトスタンダードへ

- Cavium: 2018年5月に製品版の量産を発表

- Competition: Intelに対するSecondary Choice
- Sustainability: ARM ベンダーのSecondary choice

- サーバベンダーのコミット：HPE, Cray, Fujitsu

- メジャークラウドのコミット：Microsoft, Google, etc.

- HPCソフトウェアエコシステムの確立

- 米国：DoE Sandia/Los Alamos NL, NERSC

- 欧州：European Exascale, CEA(仏), BSC(西), EPCC&Bristol (英)

- 中国：NUDT-Tianhe 3、上海交通大学

ARM HPC ECOSYSTEM事例: 英国

(slides courtesy Prof. Simon McIntosh-Smith)

- 英国はARMの本拠地、採用に積極的
- 多くのアプリやソフトが移植中
- だが、現状彼らが用いているARMチップ(Cavium)では、性能向上は少ない
- その他フランス(CEA), スペイン(BSC), Sandia NL(US), など



'Isambard', a new Tier 2 HPC service from GW4.
Named in honour of Isambard Kingdom Brunel



I.K. Brunel 1804-1859



- The Isambard project's focus will be on the top 10 most heavily used codes on Archer in 2017:
 - VASP, CASTEP, GROMACS, CP2K, UM, HYDRA, NAMD, Oasis, SBLI, NEMO
 - Note: 8 of these 10 codes are written in **FORTRAN**
- Additional important codes for project partners:
 - OpenFOAM, OpenIFS, WRF, CASINO, LAMMPS, ...
- We want to collaborate wherever possible!
 - Accelerate the adoption of Arm in HPC

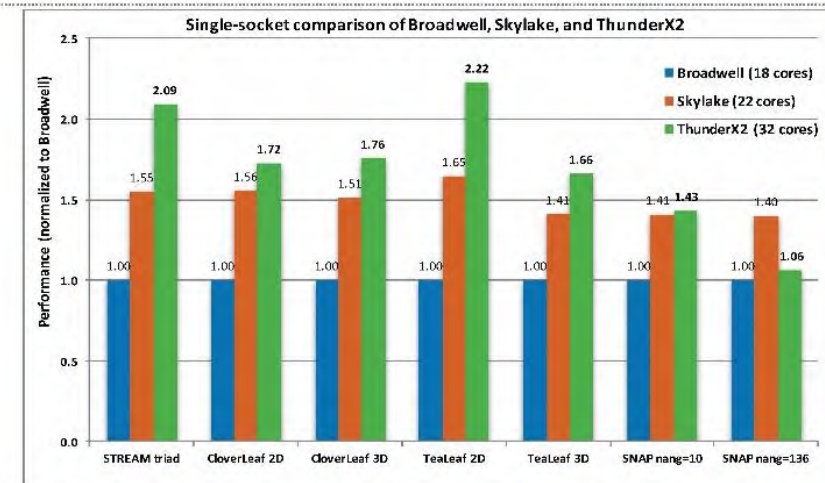


Isambard system specification (red = new info):

- Cray "Scout" system – XC50 series
 - Aries interconnect
- **10,000+** Armv8 cores
 - Cavium ThunderX2 processors
 - 2x 32core @ >2GHz per node
- Cray software tools
- Technology comparison:
 - x86, Xeon Phi, Pascal GPUs
- Phase 1 installed March 2017
- The Arm part arrives early 2018

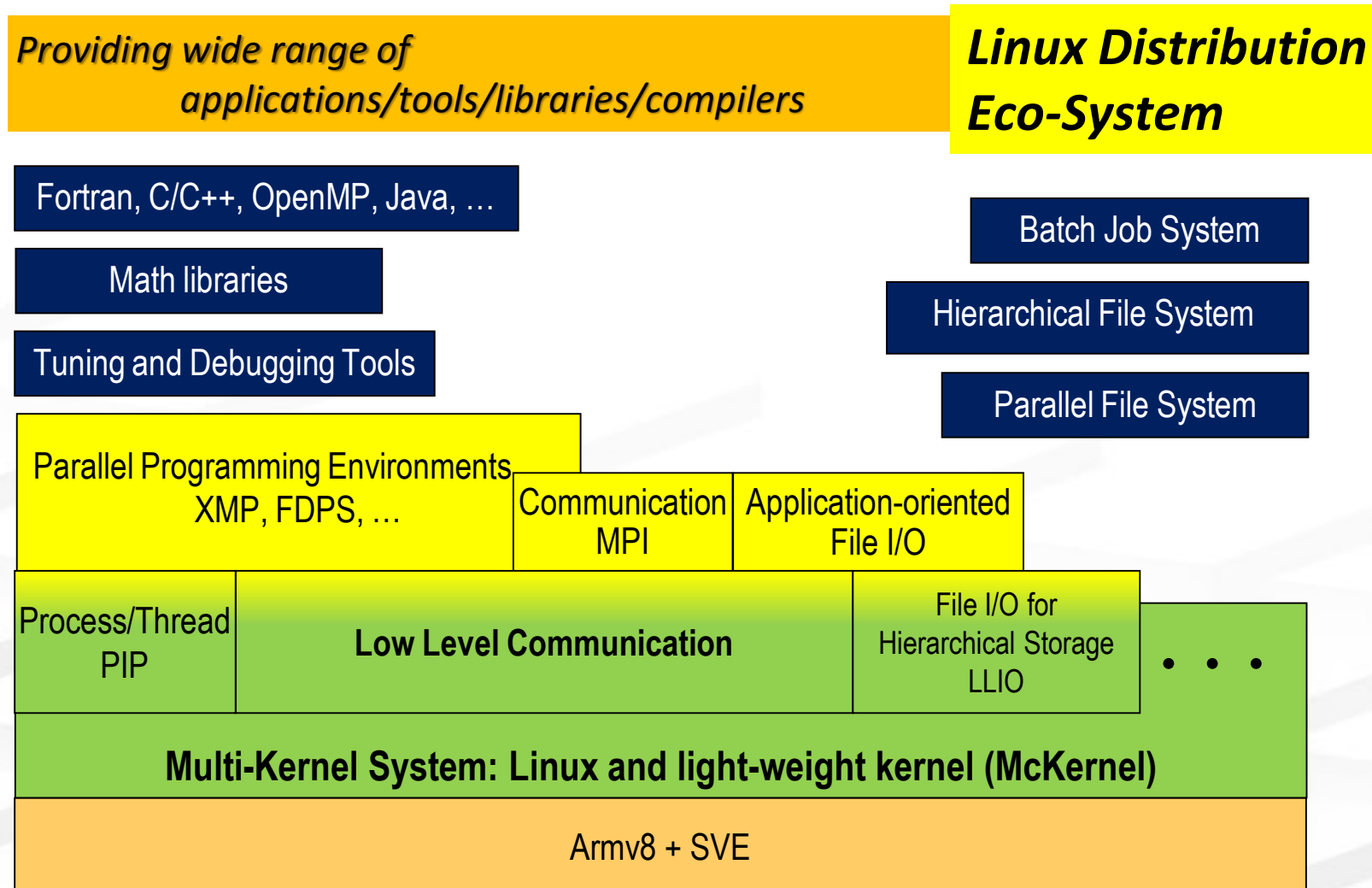


I.K. Brunel 1804-1859



An Overview of Post-K HPC System Software Stack

Easy of use is one of our KPIs (Key Performance Indicators)



3. 中間評価以降の成果予定と目標の達成見込み

- ユーザの利便性・使い勝手の良さについては、企業などでも使われている一般のソフトウェアを容易に利用でき、また、Armエコシステム構築により豊富なソフトウェアが利用できるようになる。

【ユーザの利便性・使い勝手についての調査】

- ・ ユーザの利便性・使い勝手について、ハイペリオンリサーチ社※に2020～2021年頃に設置が見込まれる世界の他のシステムとの比較調査を委託。

※ ハイペリオンリサーチ社：米国のIT専門調査会社。顧客にはIBM、Intel、Arm、富士通などのベンダや米国エネルギー省など。

<https://hyperionresearch.com/>

【結果概要】

- ・ ポスト「京」は、システムとアプリケーションの協調開発（Co-design）により、
 - ① 既存ソフトウェアの移植や新しいソフトウェアの開発を支援するエコシステム
 - ② 幅広い分野のアプリケーションに対し高い性能と信頼性を発揮するアーキテクチャであることから、ユーザの利便性・使い勝手について高い評価。
- ・ アメリカは、既存ソフトウェアの移植や新しいソフトウェアの開発に対して投資をしており、ユーザの利便性・使い勝手について高い評価。
しかしながら、2021年に導入予定のA21は、チャレンジングな最新技術によるチップを採用する予定であり、ユーザの利便性・使い勝手についての評価は低い。
- ・ ヨーロッパは、市販システムの導入を進めるとともに、プロトタイプシステムの研究開発を進めている。市販システムについてはユーザの利便性・使い勝手についての評価は高いが、プロトタイプシステムについては22か国が参加する研究開発について幅広いユーザやアプリケーションに対する要求を調整するのは困難であり、ユーザの利便性・使い勝手についての評価は低い。
- ・ 中国は、世代ごと、システムごとに独立して開発されており、共通のエコシステムが形成されていないため、ユーザの利便性・使い勝手についての評価は低い。

3. 中間評価以降の成果予定と目標の達成見込み

【調査結果】

	国	導入時期	プロセッサ	性能/開発目標※2	使い勝手 向上のための 新機能※5	新規コード 開発の 容易性※5	既存コードの 移植性※5
ポスト「京」	日本	2020年	富士通 A64FX (Armv8-A+SVE)	最大で「京」の100倍の アプリケーション実効性能	5/5	5/5	5/5
Summit	米国	2018年 第2四半期	IBM POWER9、 NVIDIA Volta/Tesla	アプリケーション※3でTitan (27.1PF) の 4倍の実効性能と6倍のスループット	5/5	4/5	4/5
Sierra	米国	2018年 第2四半期	IBM POWER9、 NVIDIA Tesla V100	アプリケーション※3でSequoia (20.1PF) の 4倍の実効性能と6倍のスループット	5/5	4/5	5/5
NERSC-9	米国	2020年 第4四半期	Intel Xeon Phi系列の 可能性	ミニアプリケーション群※4の実効性能結果に 基づく指標※5において Edison (2.6PF) の20倍、 NERSC-8 (27.9PF) の4倍	5/4	4/4	4/4
Crossroads	米国	2020年 第4四半期	Intel Xeon Phi系列の 可能性		5/5	4/5	5/5
A21	米国	2021年	Intel Cascade Lake-AP 後継の可能性	1エクサ以上の能力、50,000ノード以上、 5 PB主記憶、高いメモリバンド幅	2/4	2/4	3/4
EU Pre-Exascale #1	独	2021~2022年	EPI Arm CPU Gen1、 EPI Risc-V Accelerator Chip	EPIによるArm CPU設計をベースにした プロトタイプシステムの実現	2/2	2/3	2/3
EU Pre-Exascale #2	仏	2021~2022年	製品ベース	既存および計画されたアプリケーション実行に 合致する調達ベースの仕様になる可能性	4/4	4/4	5/4
Sunway 2020	中国	2020年 第4四半期※1	国産	LINPACK効率60%以上、1PBメモリ、 30GF/W電力性能	2/1	2/2	3/2
Sugon Exascale	中国	2020年 第4四半期※1	プロトタイプはHygon製x86 (AMDからのライセンス)		2/1	2/1	2/1
NUDT 2020	中国	2020年 第4四半期※1	国産		2/2	2/3	2/3

※1 1~1.5年の遅延する可能性

※2 演算性能はPEAK性能 (小数点以下2位を四捨五入)

※3 実効性能評価プログラム: LSMS、QBO、HACC、Nekbone等、スループット評価プログラム: miniFE、SANP、QMCPack 等

※4 SNAP、PENNANT、HPCG、MiniPIC、UMT、MILC、MiniDFT、GTC、Meraculous

※5 SSP: sustained system performance

※6 X/Y: Xの値はシステム開発者が想定するユーザおよびアプリケーションに対する評価であり、Yの値は設置されるスパコンセンターのユーザおよびアプリケーションに対する評価。評価基準については次頁に記載

現代のAIはHPCにより「復興」

2012年6月「キヤットペーパー」の衝撃

Googleの「ネコ認識」→機械学習によって、自らネコの概念を獲得し、識別。



実際に獲得された
ネコの「概念」
Le et al.(2012)

ディープニューラルネットワーク

AI研究のブレイクスルー。しかし原理は1970年代からあった

**Society5.0の実現に向けた
飛躍的な発展**

画像認識、ロボット・自動運転、自動翻訳...etc.

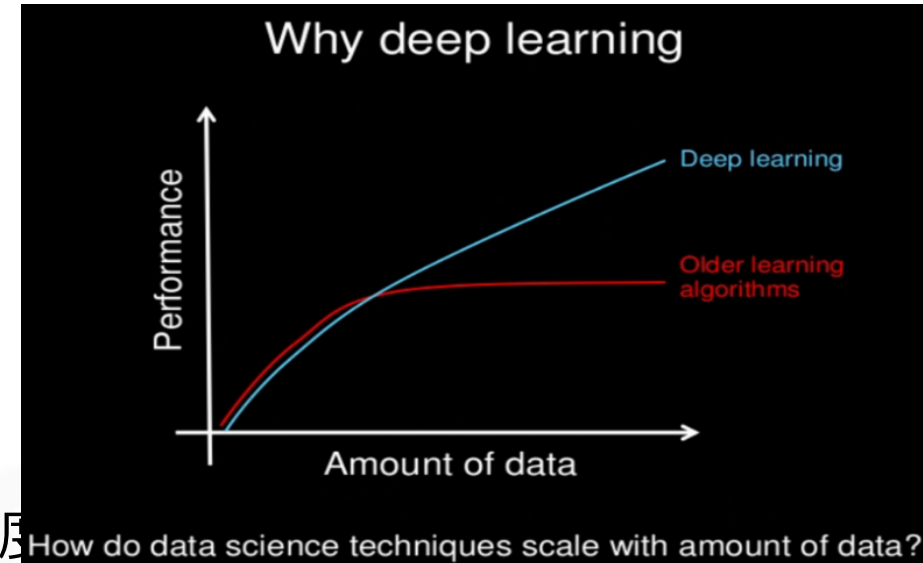


Fig. 2: Andrew Ng (Baidu) "What Data Scientists Should Know about Deep Learning"

従来の学習法と比べ、(1)学習データ量を増やすとどんどん精度上がるが、(2) その分大量の計算が必要になる
なので、**計算パワーが100万倍になって初めて可能になった**

学習の計算パワーを上げるには、通常のシミュレーションと同じく

- (1) 個々のCPUでの速度をハード・ソフトの工夫で高速化
- (2) それを大規模並列化する



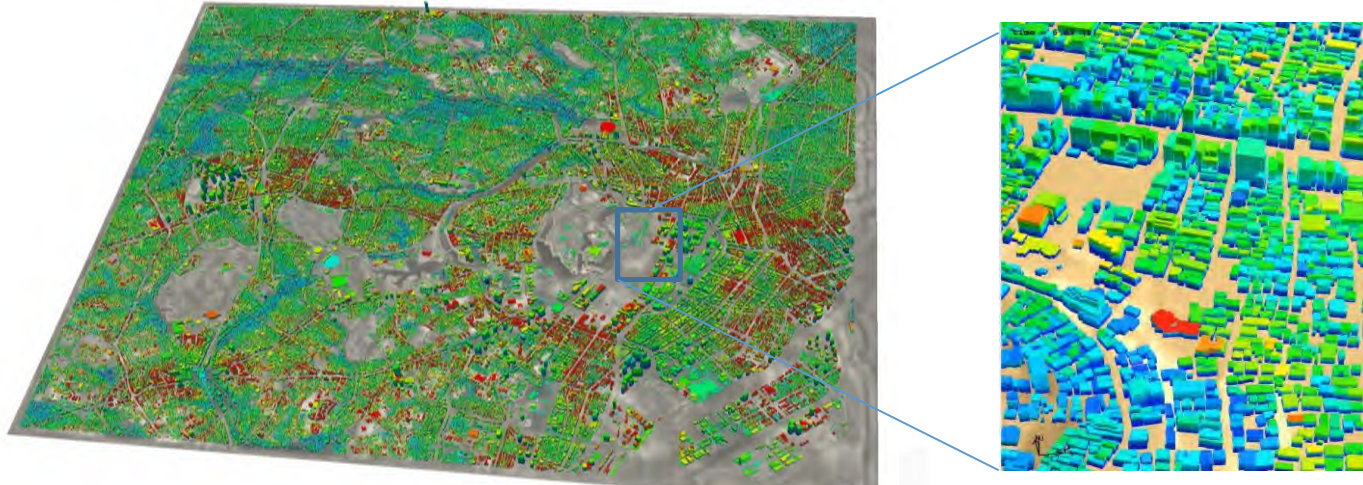
**スパコンによる
大規模な計算**

HPCとAIの「融合」とは

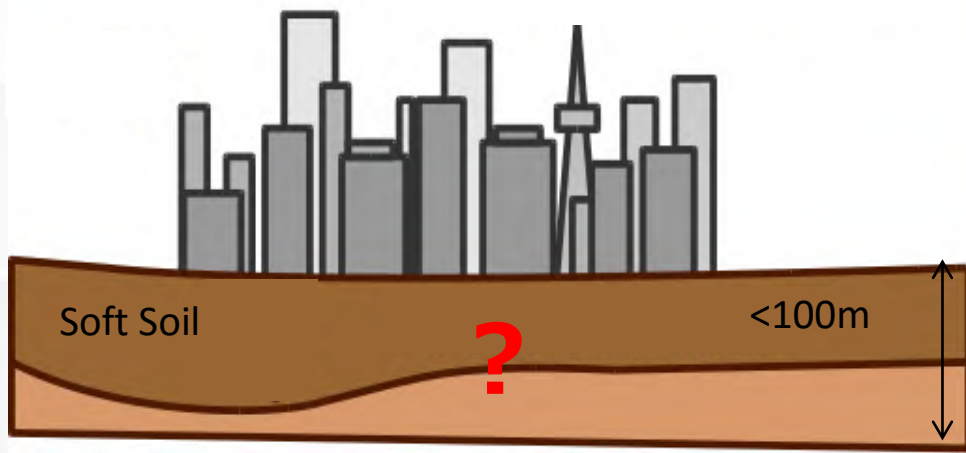
- **AI(経験則的手法)によるシミュレーション(第一原理的手法)の加速**
 - MDなどLong trajectory計算でのtrajectory外挿
 - 産業応用時の最適化時のパラメタ探索空間の削減
 - 収束計算加速の為のパラメタ調整、などなど
- **AI(経験則的手法)によるシミュレーション(第一原理的手法)の置き換え**
 - 精緻なモデルが不明な場合、計算コストが高すぎる場合
- **HPCによるAIの加速**
 - 学習データのビッグデータ処理-data cleansing
 - 学習の加速：大規模並列トレーニングの加速- より深いネットワーク(ResNET X00)、高次元・高解像度・画像以外の非定型データ
 - 推論の加速：ストリーミングデータに対するリアルタイムレスポンス
 - 様々な高度な学習法：強化学習、GAN、Dilated Convolution, etc.

Large Scale simulation and AI coming together

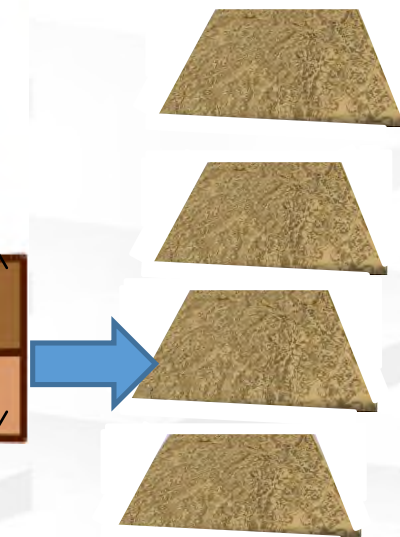
[Ichimura et. al. Univ. of Tokyo, IEEE/ACM SC17 Best Poster
2018 ACM Gordon Bell Prize Finalist]



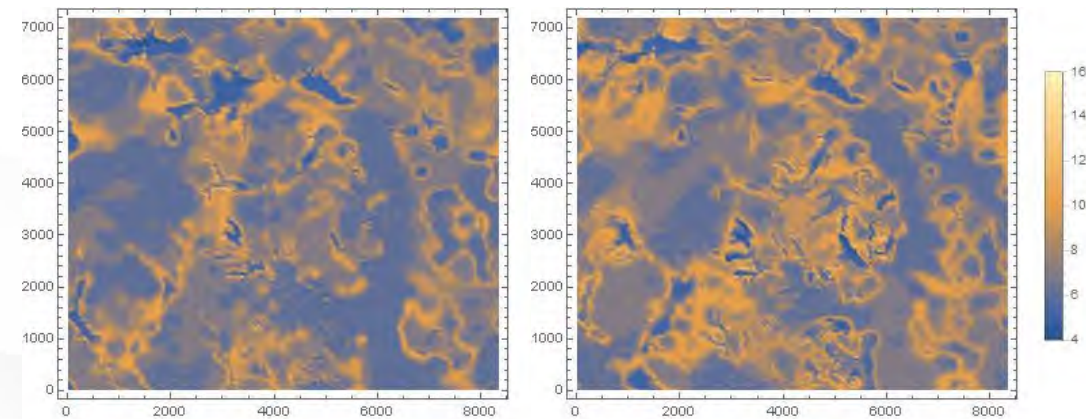
130 billion freedom
earthquake of entire Tokyo
on K-Computer (ACM
Gordon Bell Prize Finalist,
SC16,17 Best Poster)



Earthquake



Too Many Instances



Candidate
Underground
Structure 1

Candidate
Underground
Structure 2

AI Trained by Simulation
to generate candidate
soft soil structure

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)

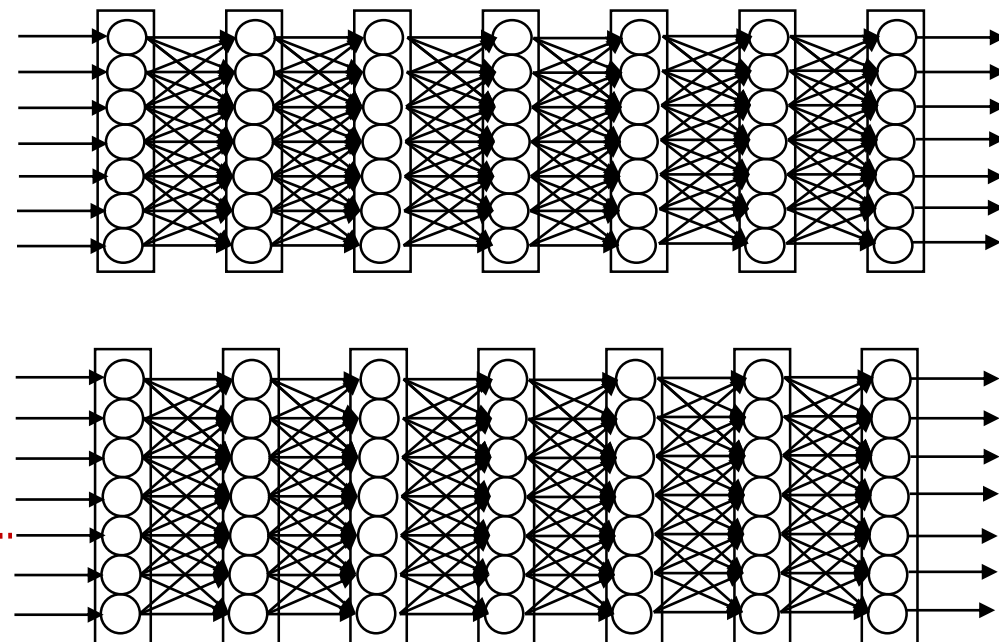
- Split and parallelize the layer calculations in propagation
- Low latency required (bad for GPU) -> strong latency tolerant cores + low latency TOFU network

- 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!

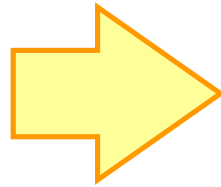


Massive amount of total parallelism, only possible via supercomputing

AI・機械学習の事例

高性能な深層学習の畳み込み演算

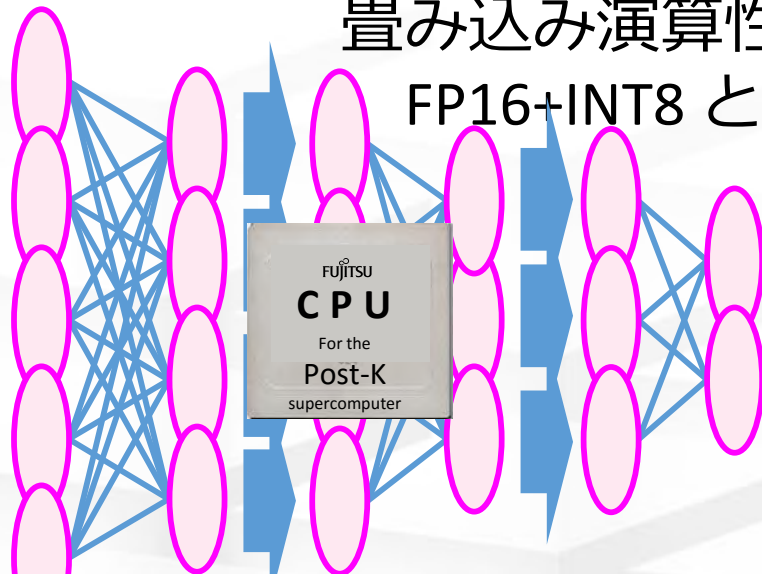
- ◆ 高性能低精度演算(FP16, INT8)
- ◆ 高メモリバンド幅
- ◆ 高性能・スケーラブルネットワーク



今後の大規模なデータ学習を必要とするAI研究において、世界トップのマシンとなる！

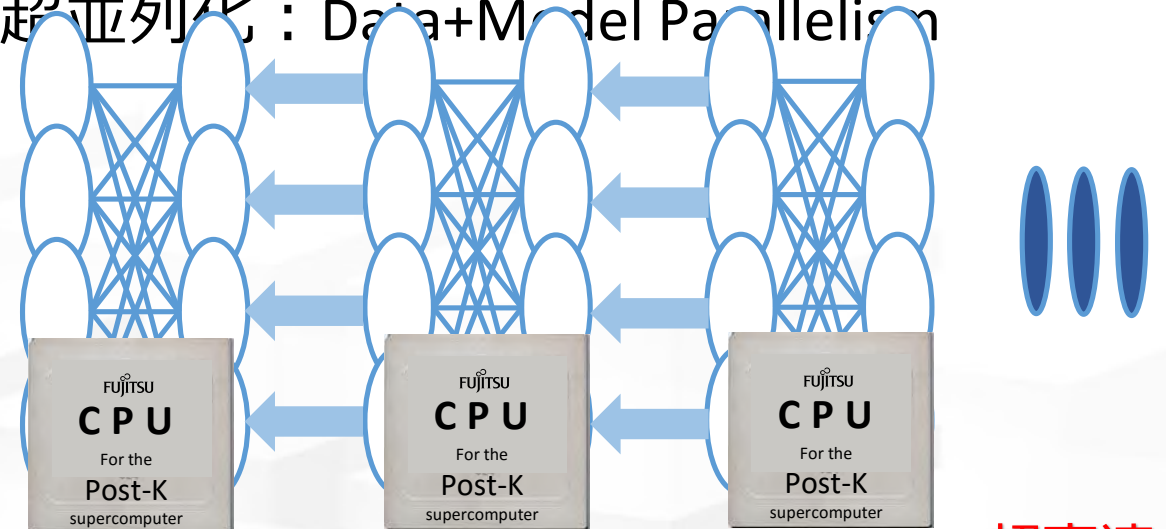
ワーク

- ① チップ単位の高い畳み込み演算性能
FP16+INT8 と高速メモリ



畳み込み演算の適切な選択 (FFT, Winograd) + 高メモリバンド幅 + 低精度演算機能により、GPUに匹敵or超える性能が見込める。

- ② 高いネットワーク通信性能による超並列化 : Data+Model Parallelism



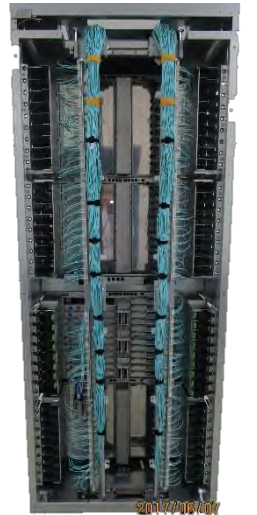
TOFU超高速ネットワーク

スケーラブルかつ低レイテンシ通信性能によって、大規模な並列化(model & data)が見込める

我が国の AI インフラの拡充に向けて ビッグデータ・AI の HPC加速 革新・実用・継続が重要

将来計画 2020年代

ポスト京も世界
トップクラス
AI-エクサ性能へ



運用中

2018年8月 **x5.0~7.7**
ABCI (産総研AIRC)
550 AI-ペタフロップス
以上(実際は数百?)

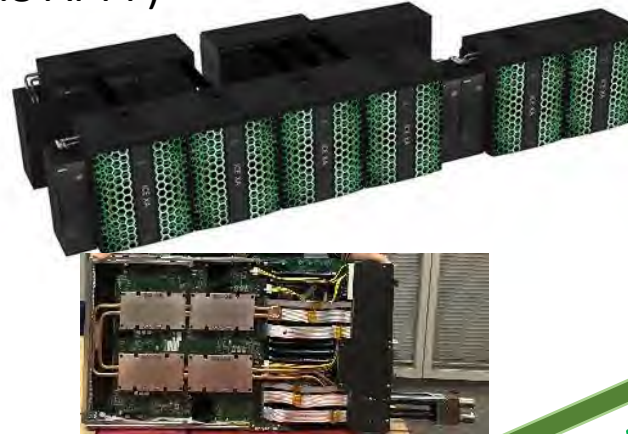
IDCや設備も
AI用に革新



x2.8~4.2

運用中

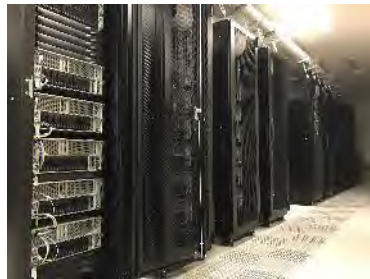
2017年8月
TSUBAME3.0
(東工大・HPE)
47.2 AI-PF (Tsubame2.5合算だと
65.8 AI-PF)



運用中

2017年4月 **x5.8**
AIST AI クラウド
(産総研AIRC・NEC)
8.2 AI-ペタフロップス

x5.8



3.5年で1000倍以上

運用中

2017年4月
RAIDEN
(理研AIP・富士通)
4.1 AI-ペタフロップス

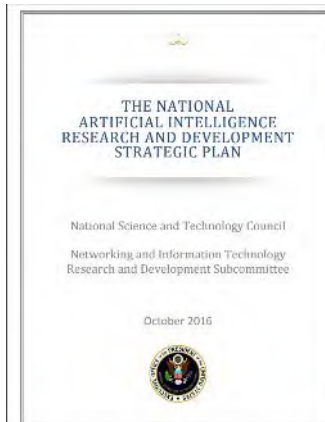


運用中



2015年10月
TSUBAME-KFC/DL
(東工大・NEC・NVIDIA)
1.4 AI-ペタフロップス

2016年10月 ホワイトハウス AI戦略白書
「AIにおける最先端のアルゴリズム・ソフト・ハードさらにそれらのコデザインのためにはオープンなインフラを中心としたコミュニティ形成が必須」



1. 高速メモリテクノロジーによる高メモリバンド幅、それに伴うHPCアプリの高性能演算性能

- コ・デザインアプリに加え、世界のベンチマーク調査から、多くのHPCアプリはメモリ性能律速、一部は演算性能律速→高いメモリ性能と、それとバランスした演算性能の向上が重要
- ポスト京CPUと現存トップサーバCPUと比較：何倍ものメモリ性能と、比較優位な演算性能
- 超高速メモリに対応し(CPUでは世界初)、演算性能も上げたCPUの設計が成功のカギ

2. 高い省電力・グリーン性能

- 電力最適化設計、クロックゲーティング、パワーストック等の高性能省電力設計
- CPUでは世界トップレベルの電力性能比、及び高効率冷却→クラウドなど一般ITにも適用可

3. ARM サーバ&HPC エコシステムの充実→使いやすさ・広範囲な成果普及

- 年間のプロセッサの生産量：x86 3億個 vs. ARM 30億個
- 高性能エコシステムの急速な確立 (e.g. Cavium, HPE, Sandia, Bristol, CEA, ...)
- ポスト京 SVE(Scalable Vector Extension) ->スパコン・クラウドからIoTまでグローバル標準

3. Society5.0：ポスト「京」の圧倒的性能優位性のHPC、ビッグデータ、AIなどへ適用

- 次世代の深層学習によるAIは莫大な演算量→大規模なHPCが必要
- Post-Kプロセッサ：FP16などのAI専用命令と、高バンド幅メモリ => 深層学習の中心である「畳み込み演算」の優れた性能と、大規模な並列化のスケールビリティ
- 世界トップのAIマシンの可能性、Blockchain Securityなど他のSociety5.0アプリでも同様

