資料４

TSUBAME3, ABCI, Post-K での超高速・超スケーラブル深層学習のHPCによる進化

理化学研究所　計算科学研究センター
センター長　　松岡　聡

DNN Circa 2018
Basically achieved w/HPC underneath

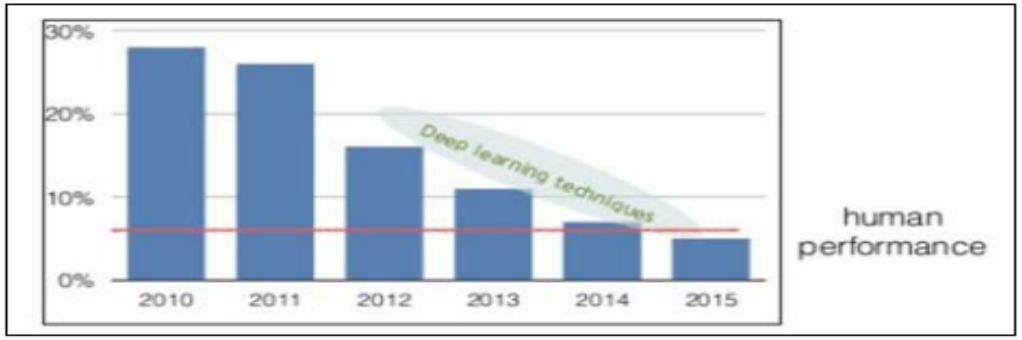
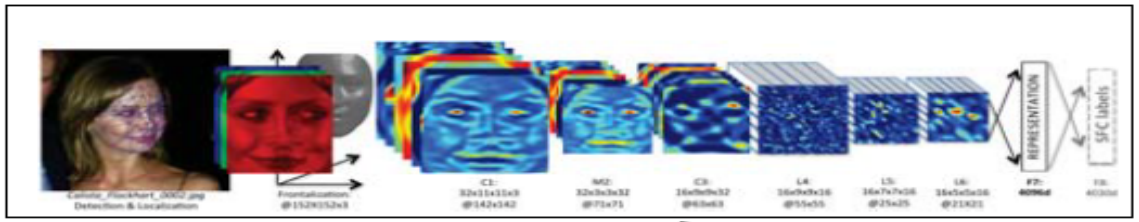
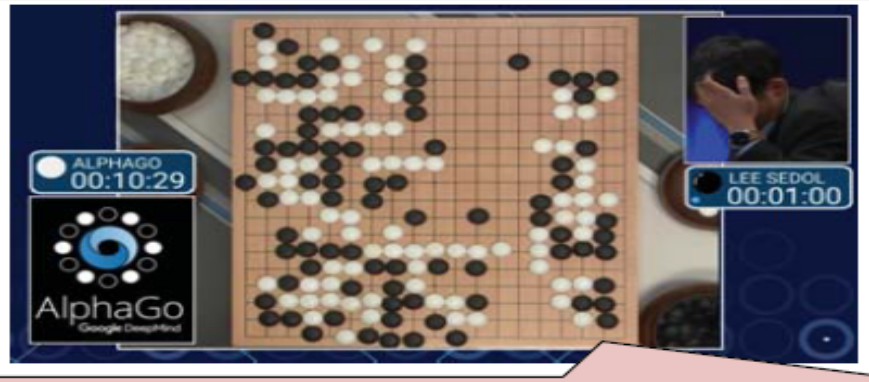


Image Classification Error Rate
Armatures: 5.1 97.53%
Professionals: 2%
ResNet (Microsoft):3.57%





Facial Recognition
Humans 97.53%
DeepFace (Facebook):97:35%
FaceNet (Google)99.63%

Game of Go: Human Pro vs AI
AlphaGo (Google DeepMind)が
4勝1敗(5戦中)

(Slide courtesy Naonori Ueda @ Riken AIP)

現代のAIはHPCにより「復興」

2012年6月「キャットペーパー」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.

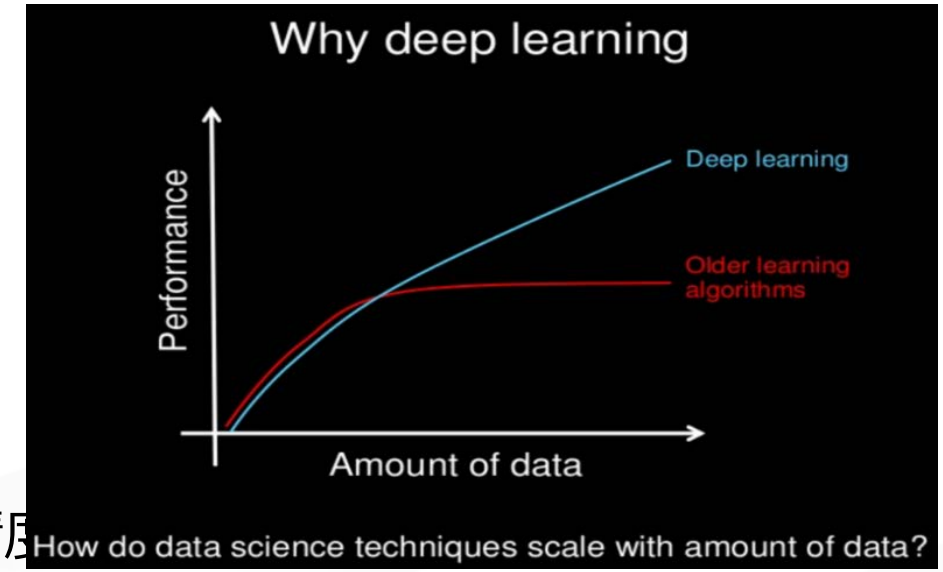


Fig. 2: Andrew Ng (Baidu) "What Data Scientists Should Know about Deep Learning"

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があがるが、(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

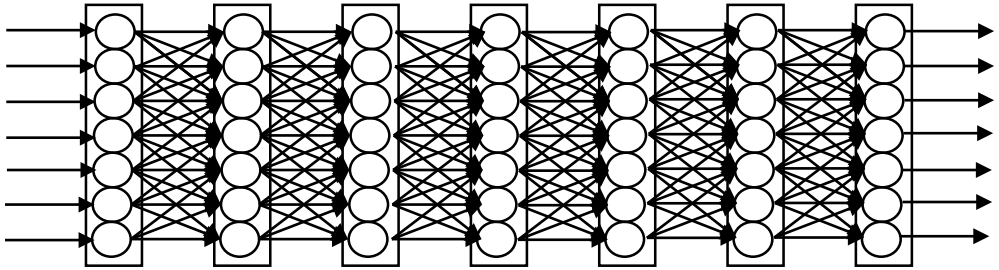
学習の計算パワーを上げるには、通常のシミュレーションと同じく
(1) 個々のCPUでの速度をハード・ソフトの工夫で高速化
(2) それを大規模並列化する

スパコンによる大規模な計算

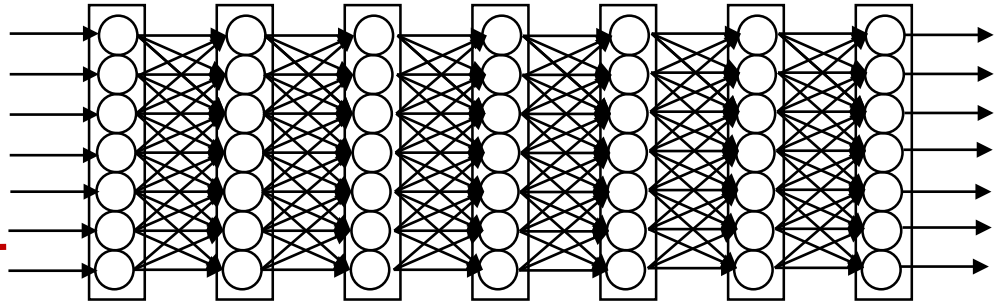
4 Layers of Parallelism in DNN Training

- 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 bandwidth bound
  - TOFU: Extremely strong reduction, x6 EDR Infiniband

Inter-Node



- Model Parallelism (domain decomposition)
  - Split and parallelize the layer calculations in propagation => Network latency bound
  - Low latency required for (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

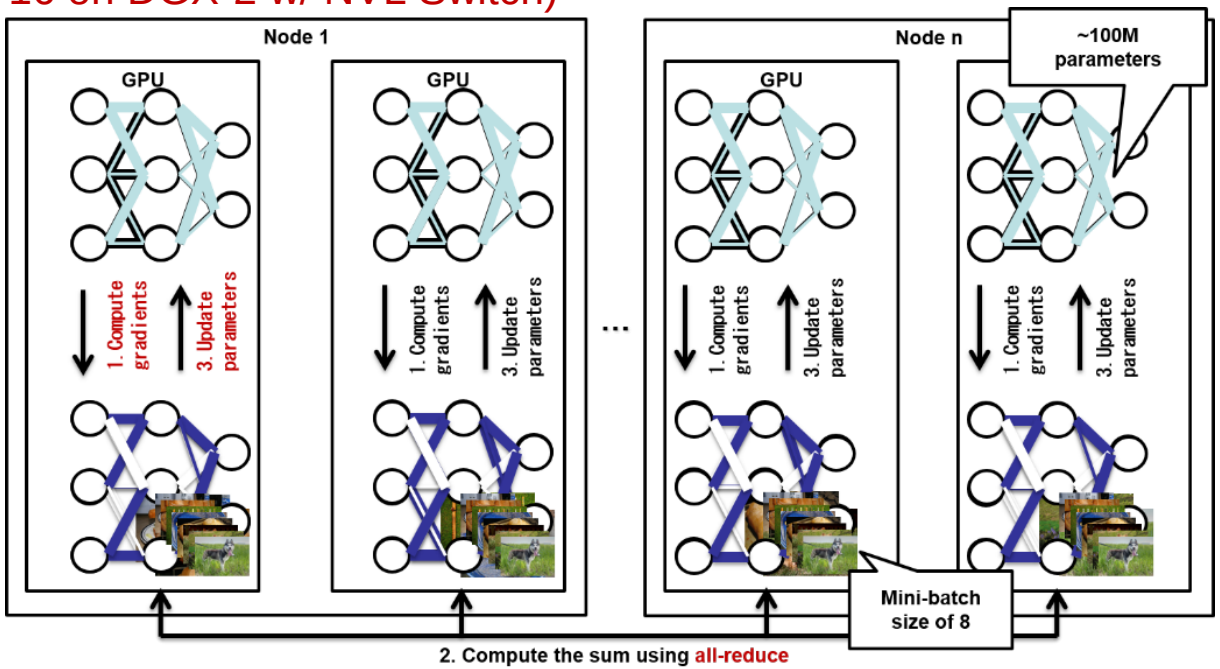
Intra-Node (Chip level)

Massive amount of total parallelism, only possible via supercomputing

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

Deep Learning is "All about Scale" Massive Parallelization is the key

- **Data-parallel training with (Asynchronous) Stochastic Gradient Descent**
  - Replicate network to all the nodes, feed different data, average the gradients periodically
  - Network All-Reduce Reduction in Megabytes~Gigabytes becomes the bottleneck at scale
  - NVIDIA: NVLink Hardware + NICL library (up to 8 GPUs on DGX-1, 16 on DGX-2 w/ NVL Switch)



Network becomes the bottleneck

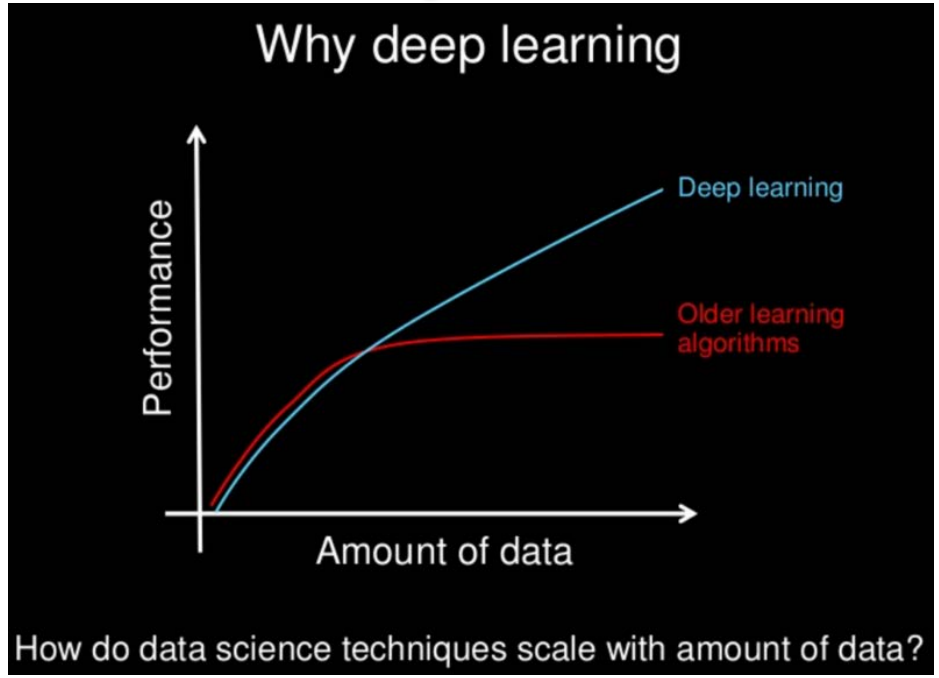


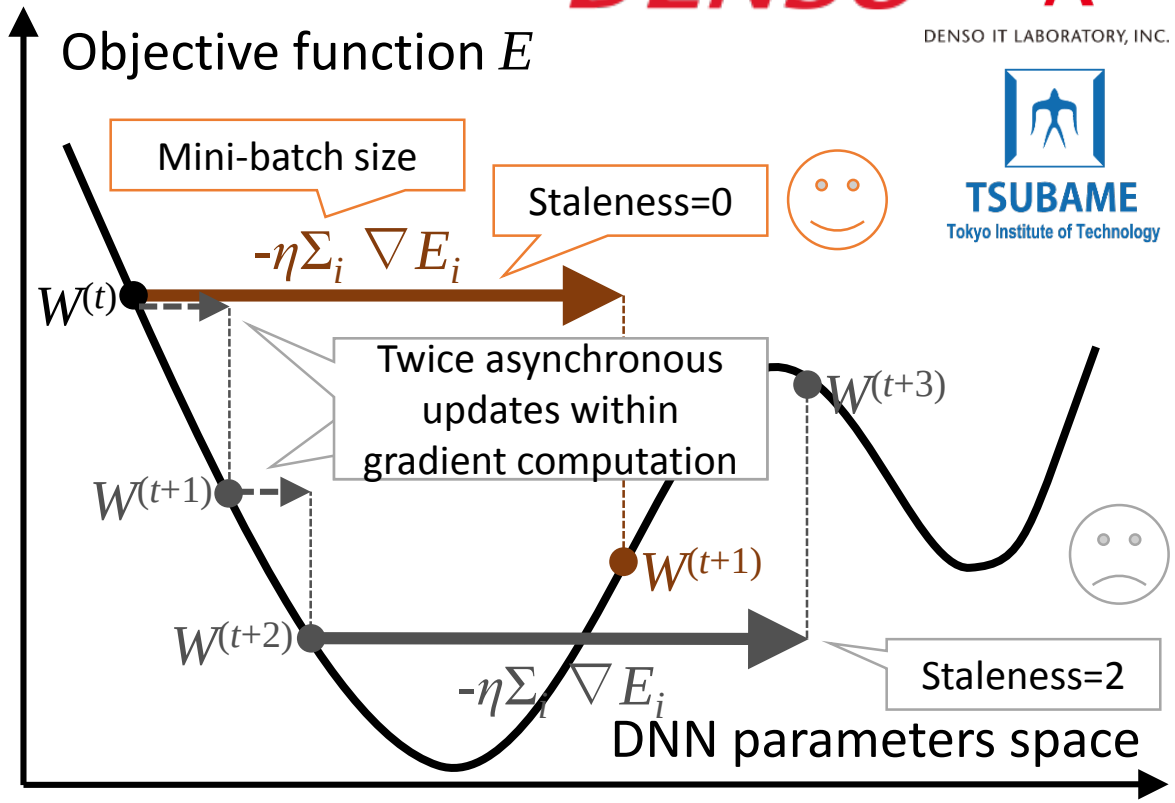
Fig. 2: Andrew Ng (Baidu) "What Data Scientists Should Know about Deep Learning"

Fig. 3: Simplified DL workflow with ASGD per iteration:
1. Compute gradient
2. Exchange gradients via all-reduce; and
3. Update network parameters

Predicting Statistics of Asynchronous SGD Parameters for a Large-Scale Distributed Deep Learning System on GPU Supercomputers

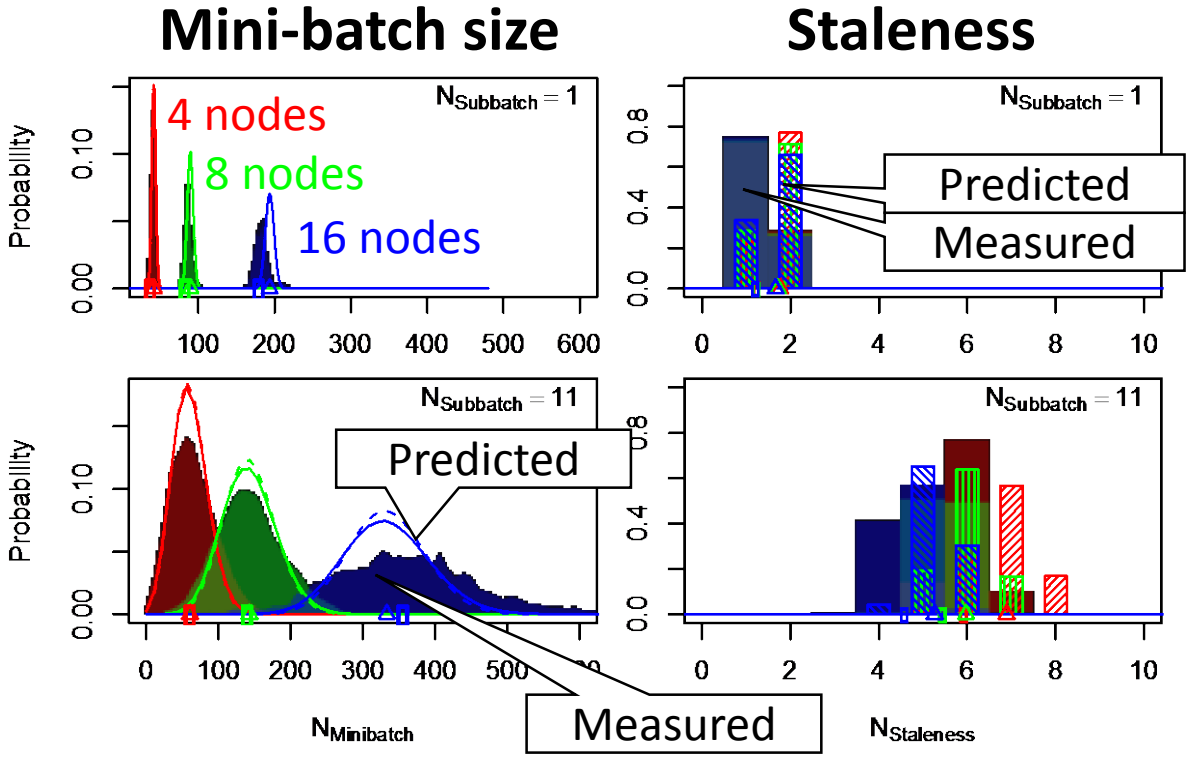
Background

- In large-scale Asynchronous Stochastic Gradient Descent (ASGD), mini-batch size and gradient staleness tend to be large and unpredictable, which increase the error of trained DNN



Proposal

- We propose a empirical performance model for an ASGD deep learning system SPRINT which considers probability distribution of mini-batch size and staleness



($N_{Subbatch}$: # of samples per one GPU iteration)

- Yosuke Oyama, Akihiro Nomura, Ikuro Sato, Hiroki Nishimura, Yukimasa Tamatsu, and Satoshi Matsuoka, "**Predicting Statistics of Asynchronous SGD Parameters for a Large-Scale Distributed Deep Learning System on GPU Supercomputers**", in proceedings of 2016 IEEE International Conference on Big Data (IEEE BigData 2016), Washington D.C., Dec. 5-8, 2016

Interconnect Performance as important as GPU Performance to accelerate DL

- **ASGD DL system SPRINT (by DENSO IT Lab) and DL speedup prediction with performance model**

$$T_{Epoch} = \frac{N_{File} \times T_{GPU}}{N_{Node} \times N_{GPU} \times N_{Subbatch}}$$

  - Data measured on T2 and KFC (both FDR) fitted to formulas
  - Allreduce time ($\in T_{GPU}$) dep. on #nodes and #DL_parameters

$$T_{Barrier} + (\alpha \log_2(N_{Node}) + \beta) \times N_{Param}$$

The Optimal Predicted Configurations of CNN-A on TSUBAME-KFC/DL

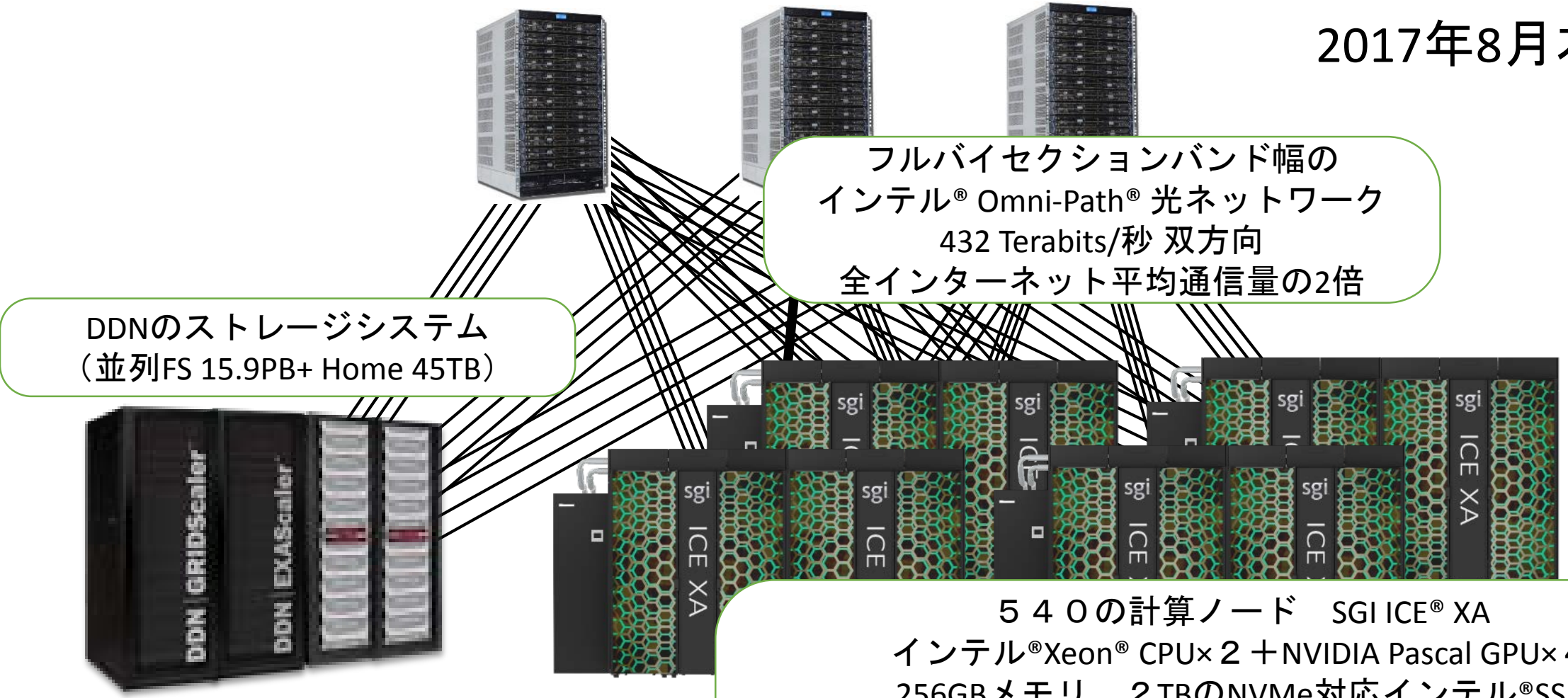
	N_{Node}	$N_{Subbatch}$	Average mini-batch size	Epoch time[s]	Speedup
Baseline K80+ FDR	8	8	165.1	1779	-
FP16	7	22	170.1	1462	**1.22**
EDR IB	12	11	166.6	1245	**1.43**
FP16 + EDR IB	8	15	171.5	1128	**1.58**

Fig. 4: Oyama et al. "Predicting Statistics of Asynchronous SGD Parameters for a Large-Scale Distributed Deep Learning System on GPU Supercomputers

- **Other approaches == similar improvements:**
  - Cuda-Aware CNTK optimizes communication pipeline ➔ 15%—23% speedup
    (Banerjee et al. "Re-designing CNTK Deep Learning Framework on Modern GPU Enabled Clusters")
  - Reduced precision (FP[16|8|1]) to minimize msg. size w/ no or minor accuracy loss

TSUBAME 3.0 のシステム概要
全GPU・全ノード・全メモリ階層での BYTES中心スケーラブルアーキテクチャ

2017年8月本稼働



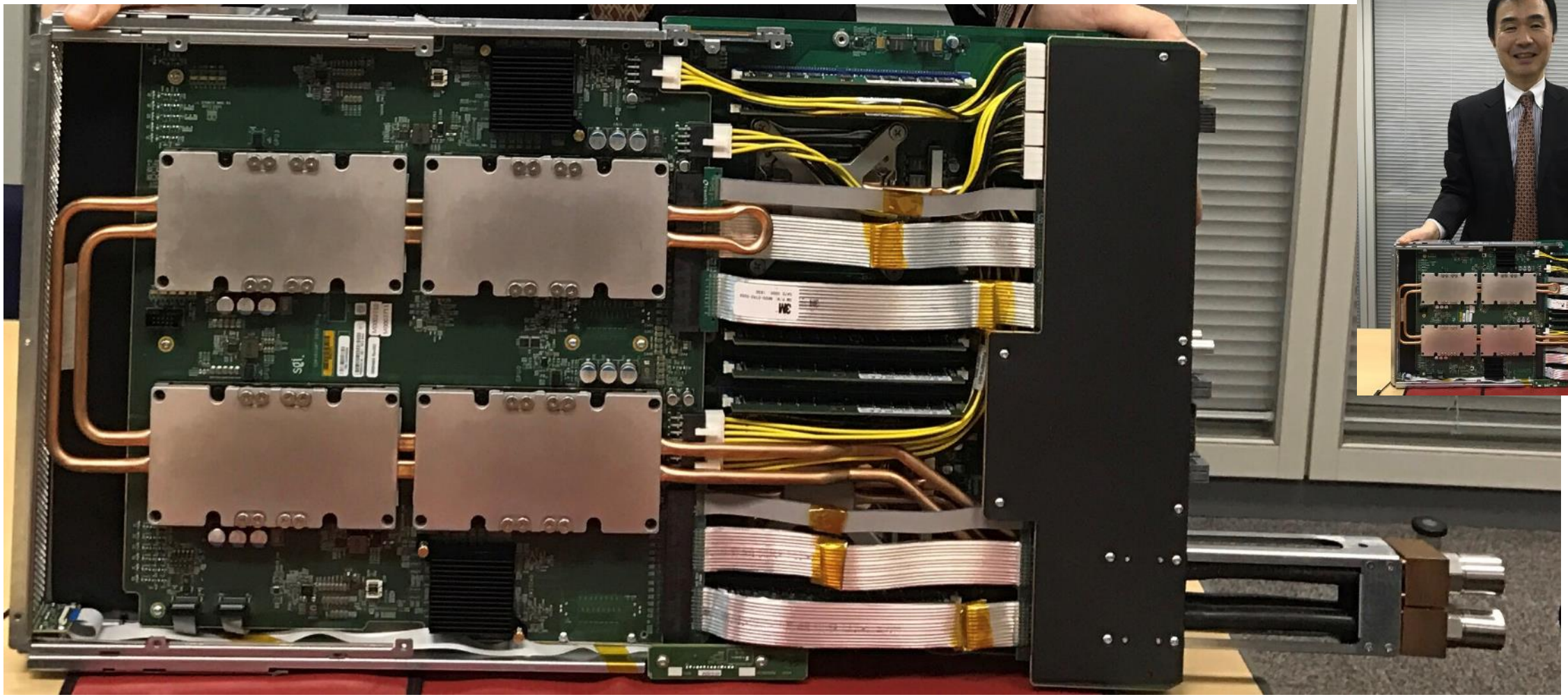
フルバイセクションバンド幅の
インテル® Omni-Path® 光ネットワーク
432 Terabits/秒 双方向
全インターネット平均通信量の2倍

DDNのストレージシステム
（並列FS 15.9PB+ Home 45TB）

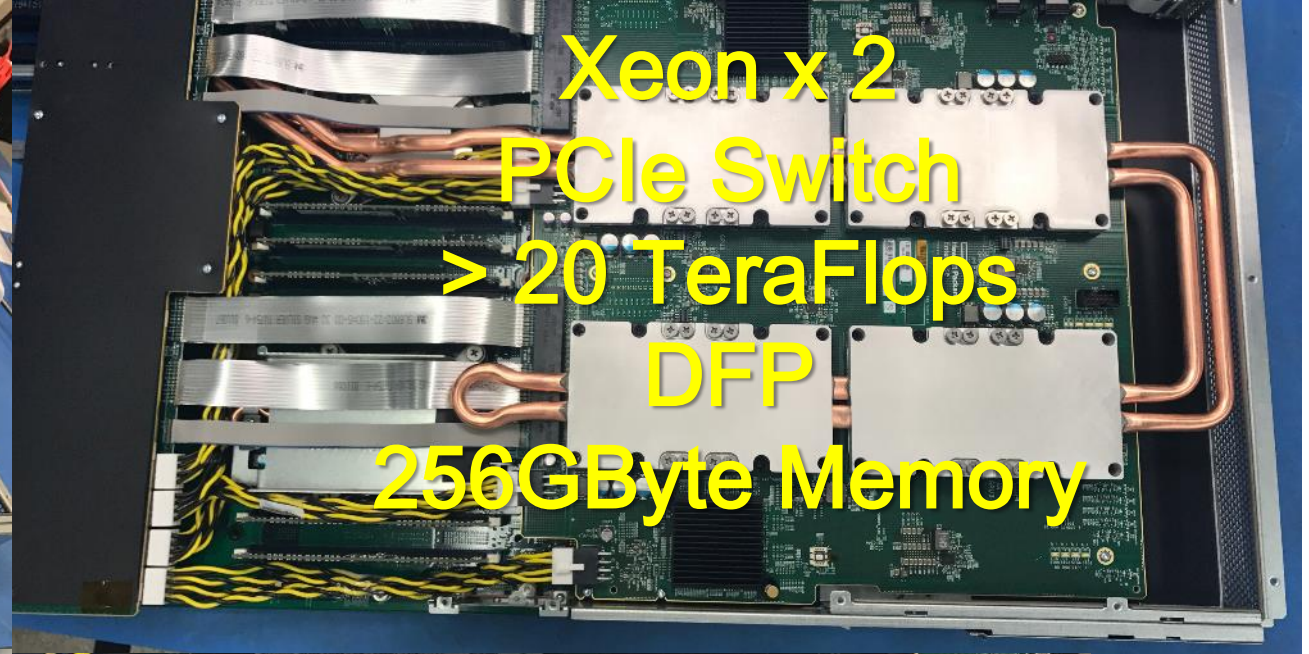
５４０の計算ノード　SGI ICE® XA
インテル®Xeon® CPU×２＋NVIDIA Pascal GPU×４
256GBメモリ、２TBのNVMe対応インテル®SSD
47.2 AI-Petaflops, 12.1 Petaflops

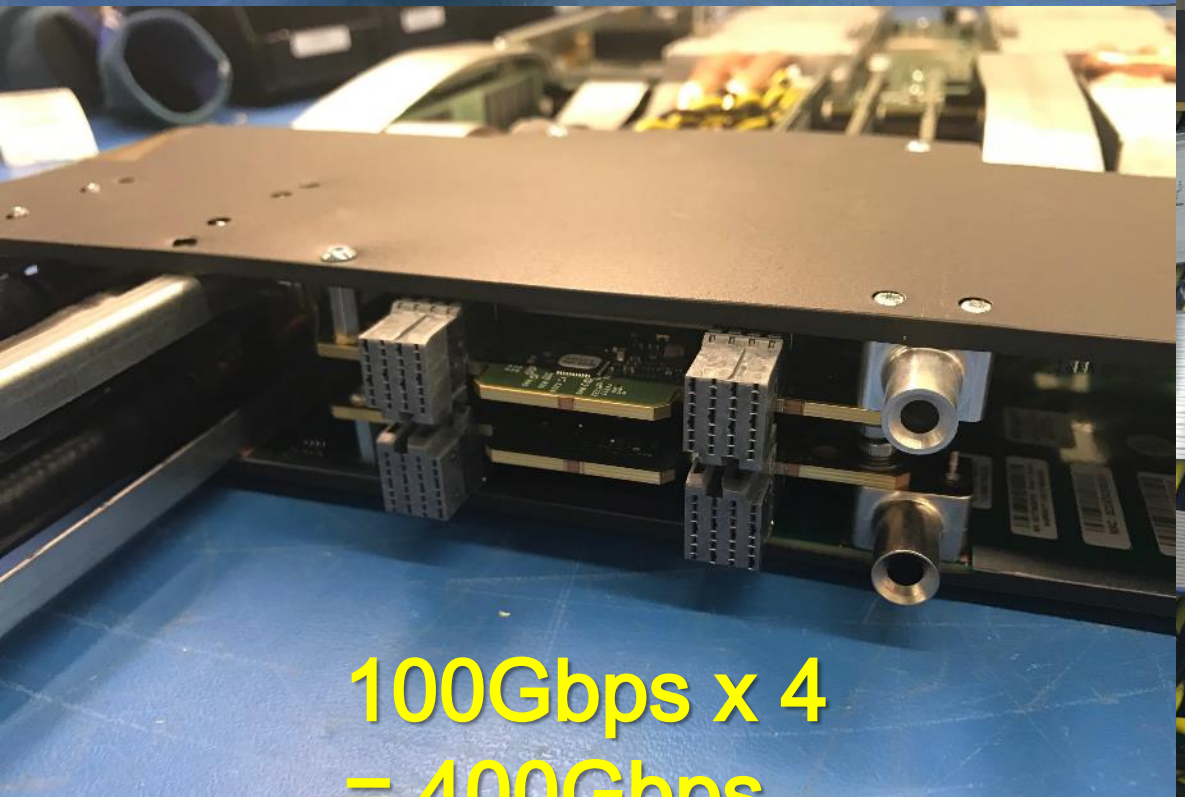
TSUBAME3.0 Co-Designed SGI ICE-XA Blade (new)

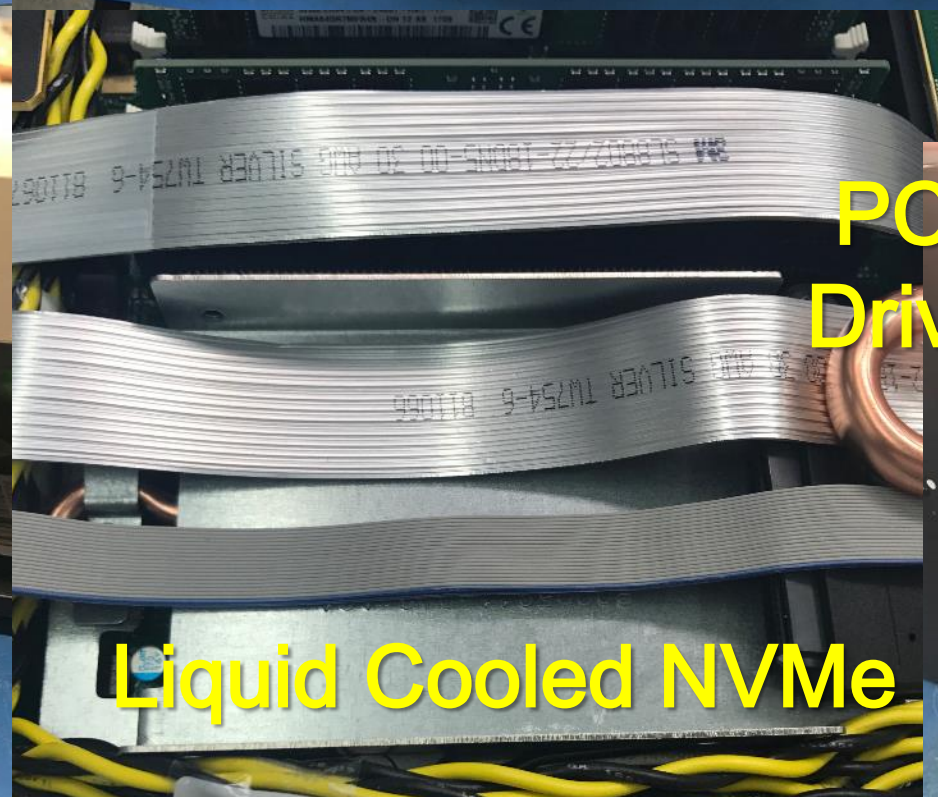
- No exterior cable mess (power, NW, water)
- Plan to become a future HPE product



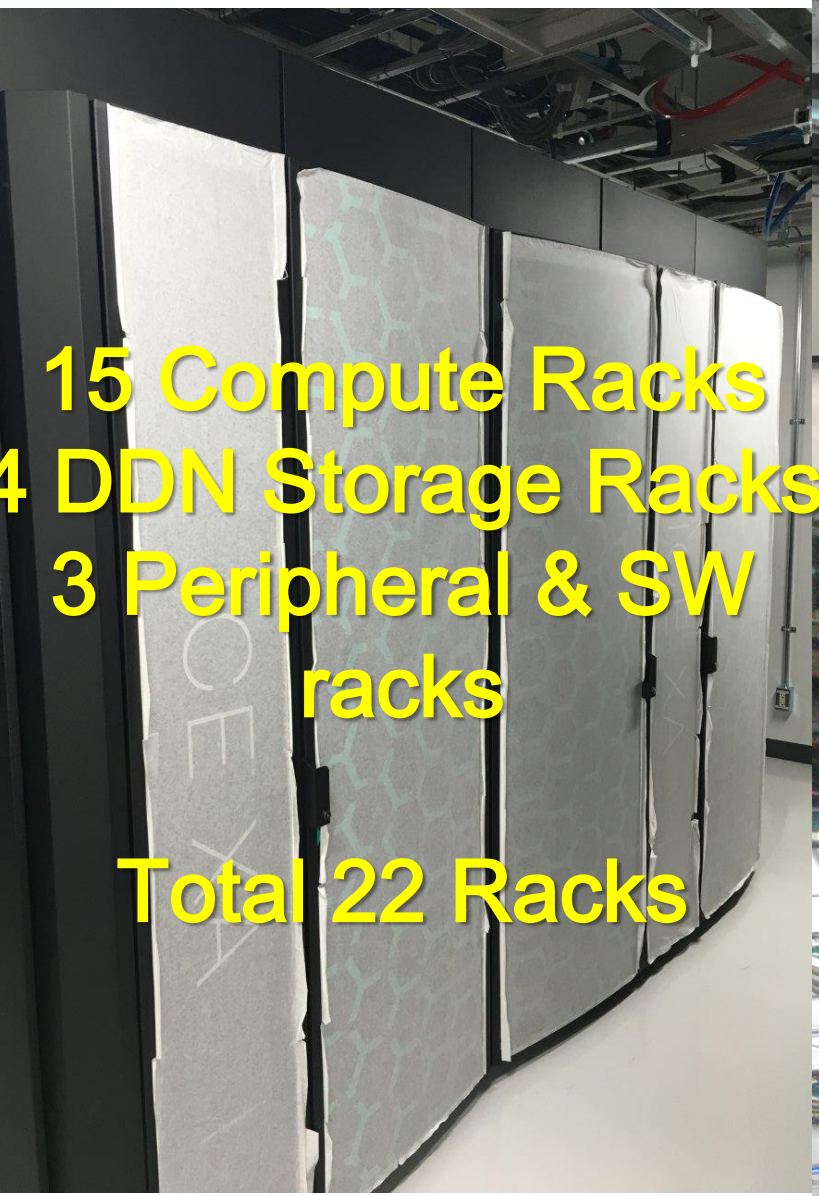

Liquid Cooled
“Hot Pluggable” ICE-XA Blade
Smaller than 1U server, no cables or pipes

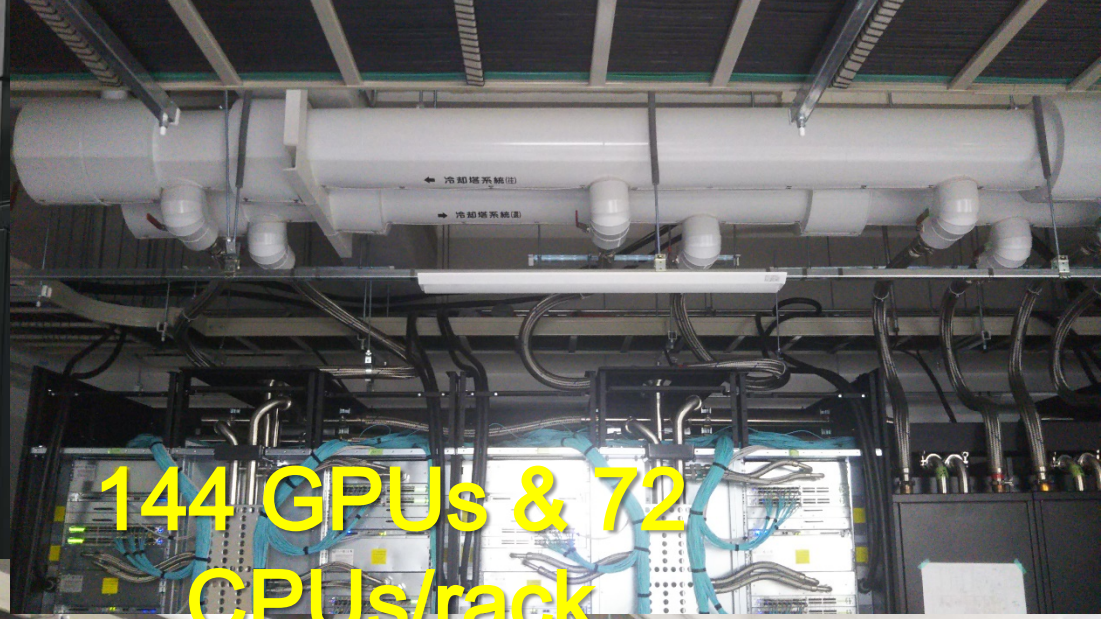

Xeon x 2
PCIe Switch
> 20 TeraFlops
DFP
256GByte Memory

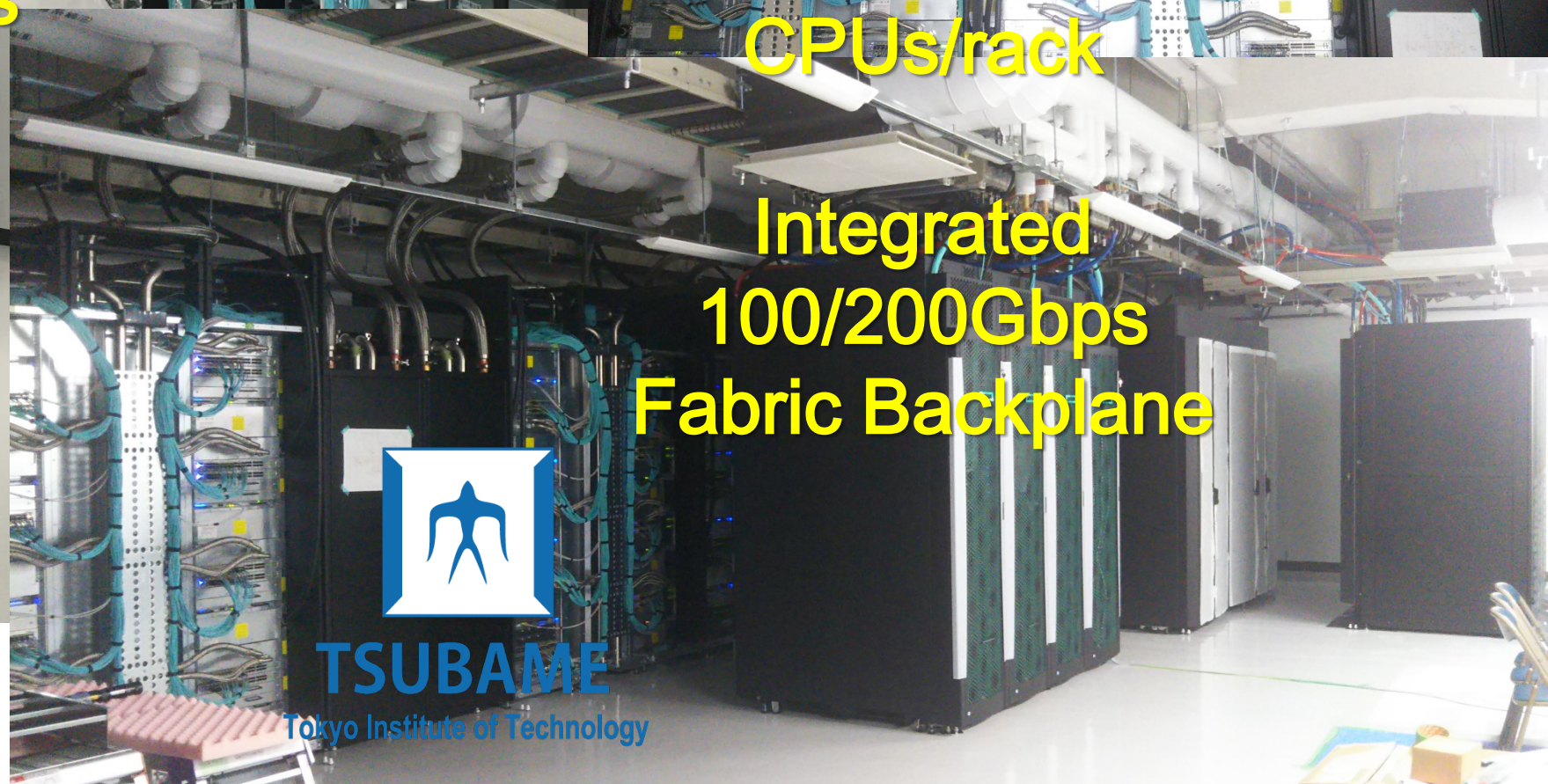

100Gbps x 4
= 400Gbps


Liquid Cooled NVMe


PCIe NVMe
Drive Bay x 4

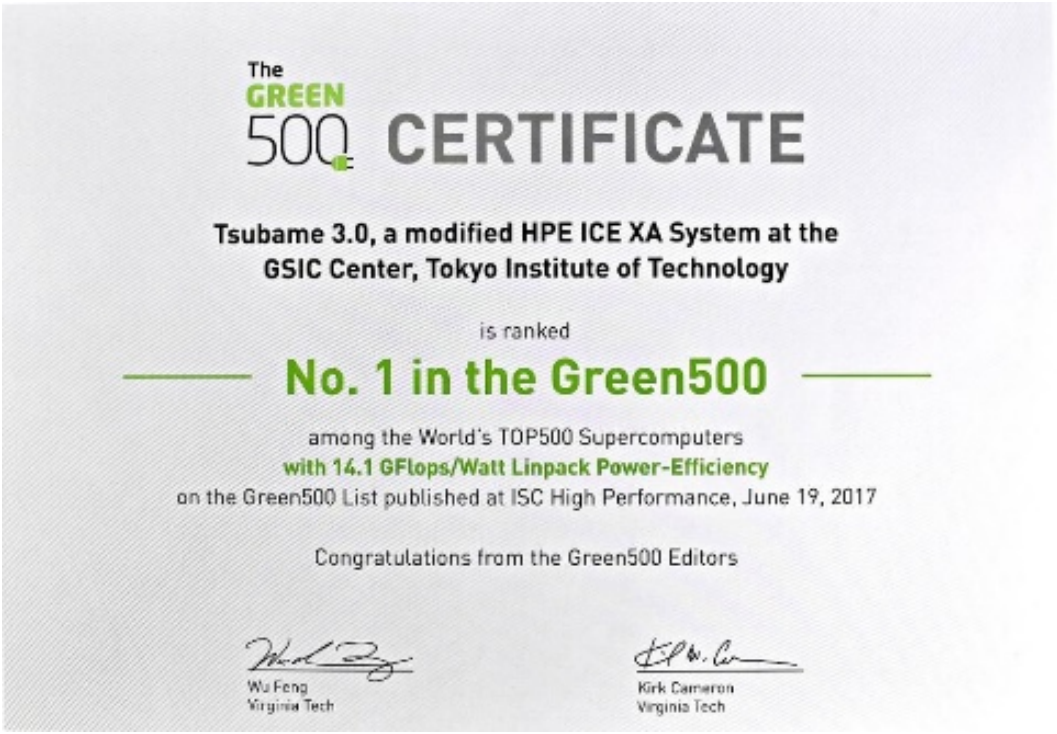

15 Compute Racks
4 DDN Storage Racks
3 Peripheral & SW racks
Total 22 Racks


144 GPUs & 72 CPUs/rack


Integrated 100/200Gbps Fabric Backplane
TSUBAME
Tokyo Institute of Technology

TSUBAME3.0 became the first large production petaflops-scale supercomputer in the world to be #1 on the “Green500” power efficiency W world ranking of supercomputers

14.1 Gigaflops/W is more than x10 more efficient than PCs and Smartphones!


The GREEN 500 CERTIFICATE

Tsubame 3.0, a modified HPE ICE XA System at the GSIC Center, Tokyo Institute of Technology

is ranked

No. 1 in the Green500

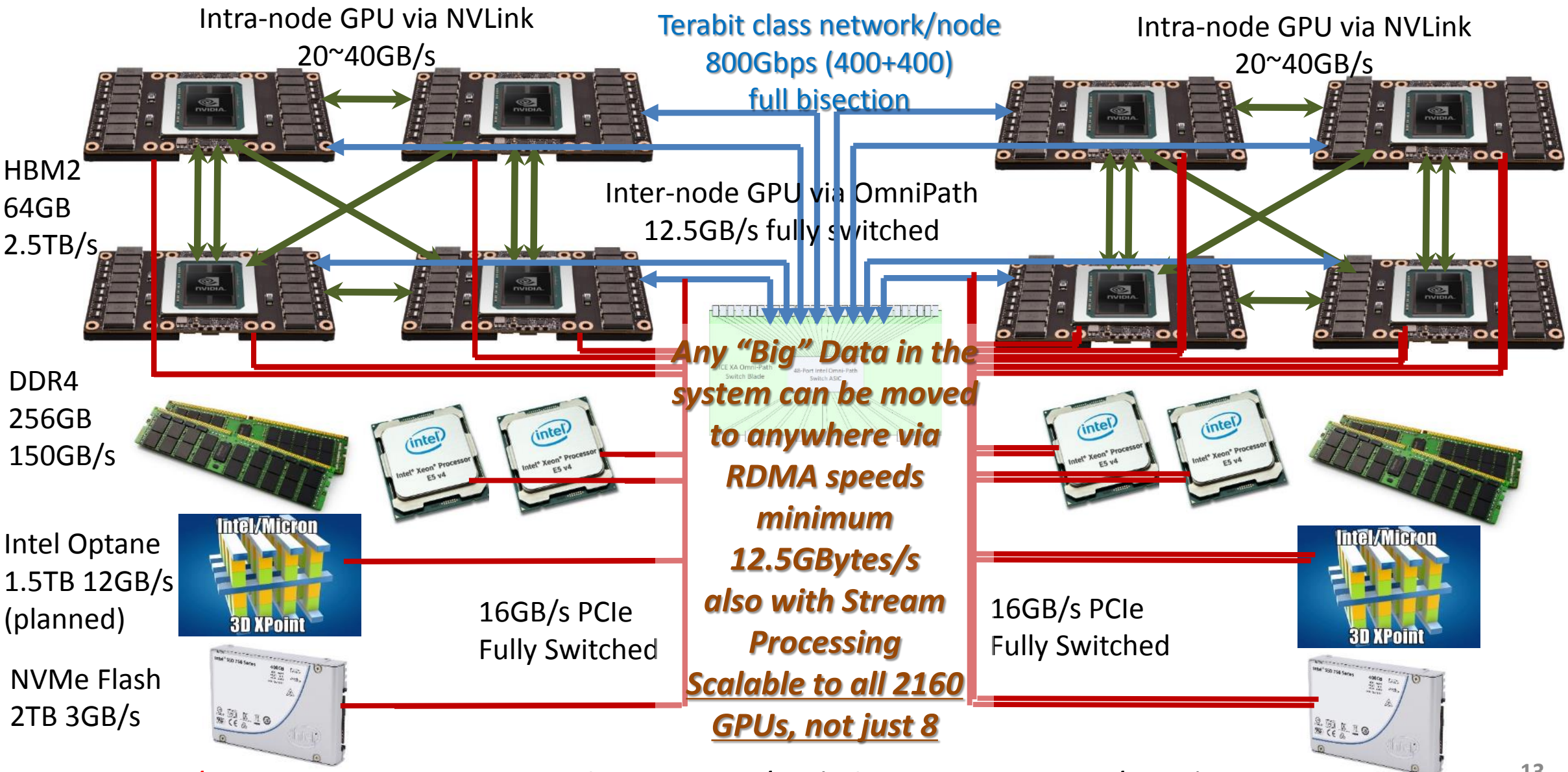
among the World's TOP500 Supercomputers
with 14.1 GFlops/Watt Linpack Power-Efficiency
on the Green500 List published at ISC High Performance, June 19, 2017

Congratulations from the Green500 Editors

Wu Feng
Virginia Tech

Kirk Cameron
Virginia Tech

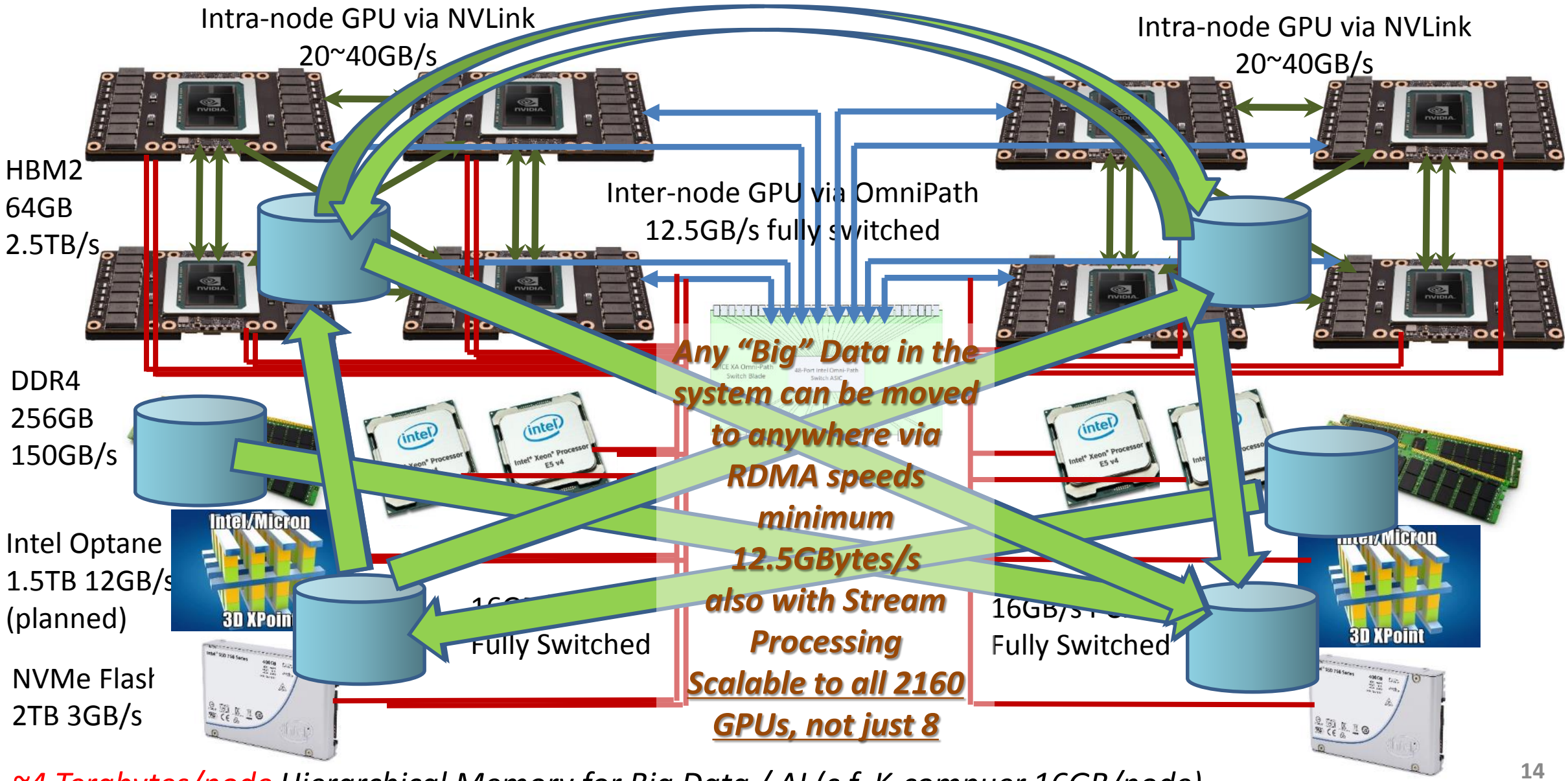
TSUBAME3: *A Massively BYTES Centric Architecture* for Converged BD/AI and HPC



~4 Terabytes/node Hierarchical Memory for Big Data / AI (c.f. K-compuer 16GB/node)

➔ *Over 2 Petabytes in TSUBAME3, Can be moved at 54 Terabyte/s or 1.7 Zetabytes / year*

TSUBAME3: *A Massively BYTES Centric Architecture* for Converged BD/AI and HPC



~4 Terabytes/node Hierarchical Memory for Big Data / AI (c.f. K-compuer 16GB/node)

➔ *Over 2 Petabytes in TSUBAME3, Can be moved at 54 Terabyte/s or 1.7 Zetabytes / year*

The current status of AI & Big Data in Japan

We need the triage of advanced algorithms/infrastructure/data but we lack the cutting edge infrastructure dedicated to AI & Big Data (c.f. HPC)

Tokyo Tech

AIRC

Joint RWBC Open Innov. Lab (OIL) (Director: Matsuoka)

AIST NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST) AIST-AIRC

東京工業大学 Tokyo Institute of Technology

Riken -AIP

NICT-UCRI

AI Venture Startups

Preferred Networks

UEI

R&D ML Algorithms & SW

Big Companies AI/BD R&D (also Science)

IT LAB DENSO IT LABORATORY, INC.

Panasonic

ABEJA

MIZUHO みずほ情報総研

DeNA

ZMP

Seeking Innovative Application of AI & Data

AI/BD Centers & Labs in National Labs & Universities

Use of Massive Scale Data now Wasted

DENSO Petabytes of Drive Recording Video

FANUC FA&ロボット&ロボマシン FA&Robots

TOYOTA

JARI 一般財団法人 日本自動車研究所

Y! YAHOO! JAPAN

SoftBank

NTT

Web access and merchandice

IoT Communication, location & other data

"Big" Data

Massive Rise in Computing Requirements (1 AI-PF/person?)

SAKURA Internet

Google Cloud Platform

Windows Azure

amazon web services

In HPC, Cloud continues to be insufficient for cutting edge research => dedicated SCs dominate & racing to Exascale

Over $1B Govt. AI investment over 10 years

AI&Data Infrastructures

Massive "Big" Data in Training

世界最大級・超省電力・オープン
ＡＩインフラストラクチャ

東京大学 THE UNIVERSITY OF TOKYO　AIST

東京大学 柏IIキャンパス

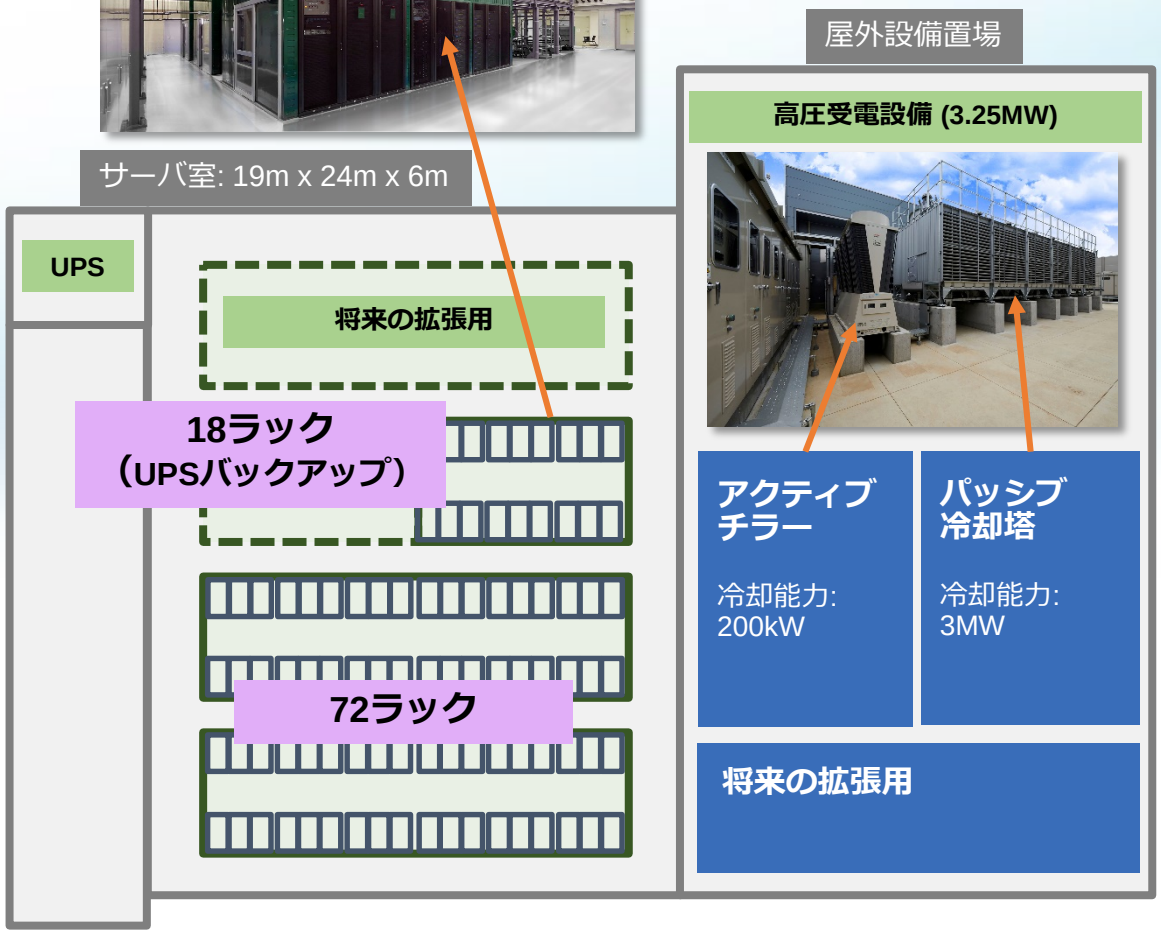
ABCI　AI Bridging Cloud Infrastructure

- 世界トップレベルの計算処理能力とデータ処理能力
- ＡＩとビッグデータのアルゴリズム、ソフトウェア、応用開発のためのオープンかつ専用の計算インフラストラクチャ
- 我が国における産学官共同のＡＩ研究開発を加速するオープンイノベーションプラットフォーム

理論ピーク性能	**: 550 PFlops (半精度)**
	37 PFlops (倍精度)
実効性能(Linpack)	**: 19.88 PFlops (世界5位, 国内1位)**
	12.054 GFlops/W (世界8位)
使用電力(最大)	**: 2.3 MW**
年間平均PUE	**: 1.1以下 (推定値)**

2018年8月1日運用開始

ＡＩデータセンター棟
“1ラック70kW級スパコンのコモディティ化を可能に”



- **低コストで軽量な建物**
- **耐荷重2トン/m²のコンクリートスラブ上にクーリングポッドとラックを直接設置**
- **ラック数**
  - 初期: 90 (ABCIは41台使用)
  - 最大: 144
- **電力容量**
  - 3.25 MW (ABCIは最大2.3MW使用)
- **冷却能力**
  - 3.2 MW (夏季の最小値)
  - ラックあたり水冷60kW＋空冷10kW

ABCI Procurement Benchmarks

- **Big Data Benchmarks**
  - (SPEC CPU Rate)
  - Graph 500
  - MinuteSort
  - Node Local Storage I/O
  - Parallel FS I/O

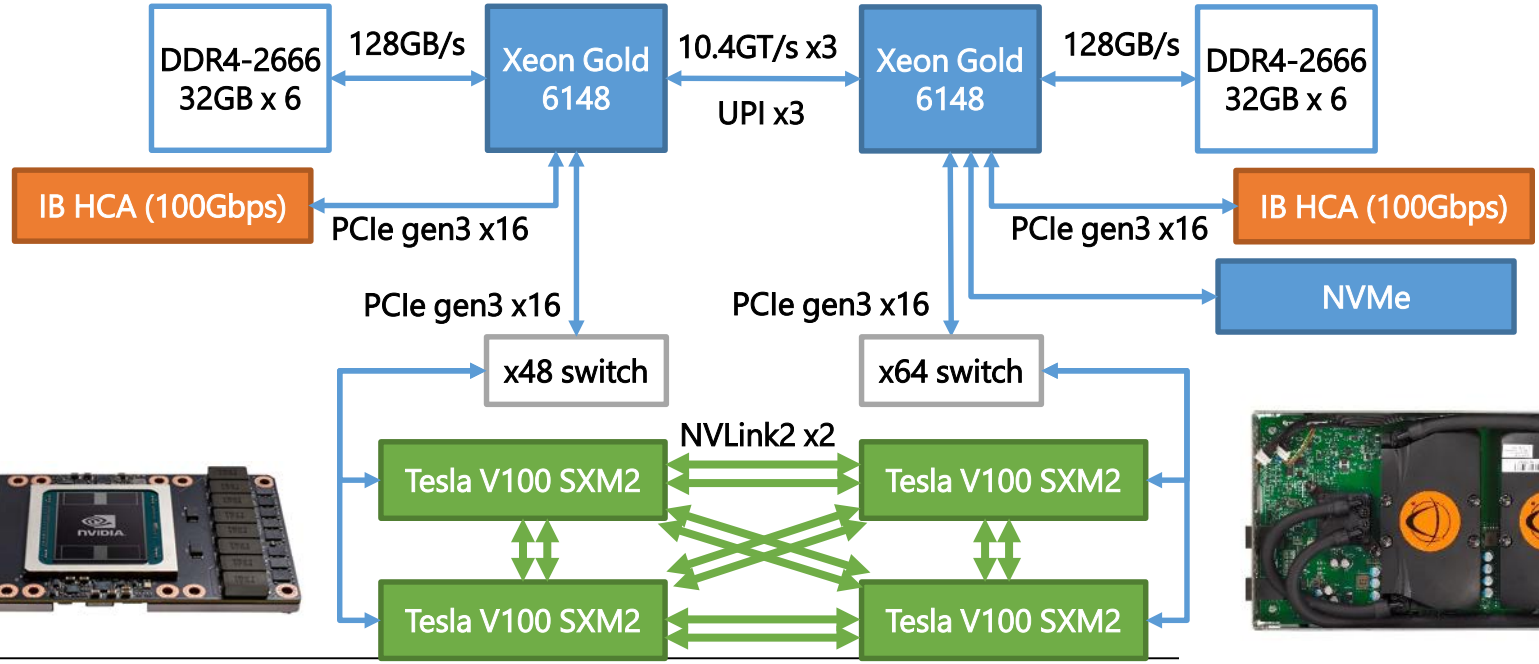
No traditional HPC Simulation Benchmarks except SPEC CPU.
Plan on “open-sourcing”

- **AI/ML Benchmarks**
  - Low precision GEMM
    - CNN Kernel, defines “AI-Flops”
  - Single Node CNN
    - AlexNet and GoogLeNet
    - ILSVRC2012 Dataset
  - Multi-Node Scalable CNN
    - Caffe+MPI
  - Large Memory CNN
    - Convnet on Chainer
  - RNN / LSTM
    - Neural Machine Translation on Torch

ABCI Computing Node

"TSUBAME3 Commoditized for IDC" (Fujitsu)

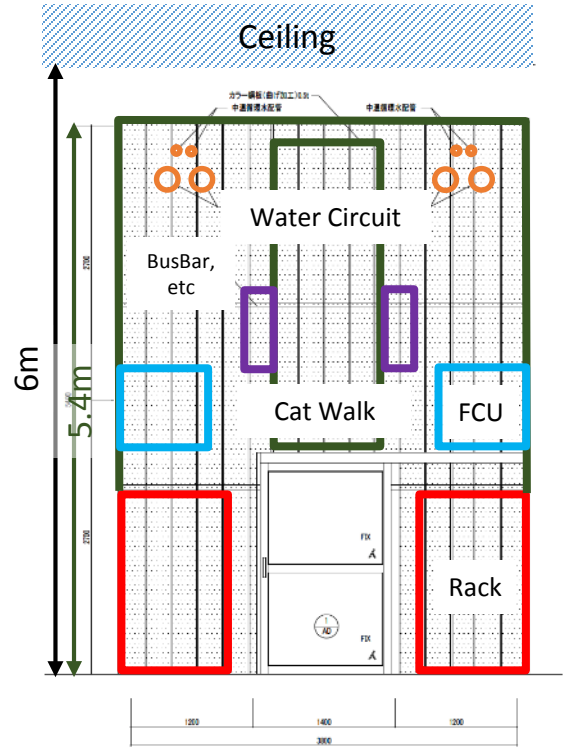
FUJITSU PRIMERGY Server (2 servers in 2U)	
CPU	Intel Xeon Gold 6148 (27.5M Cache, 2.40 GHz, 20 Core) x2
GPU	NVIDIA Tesla V100 (SXM2) x4
Memory	384GiB
Local Storage	1.6TB NVMe SSD (Intel SSD DC P4600 u.2) x1
Interconnect	InfiniBand EDR x2



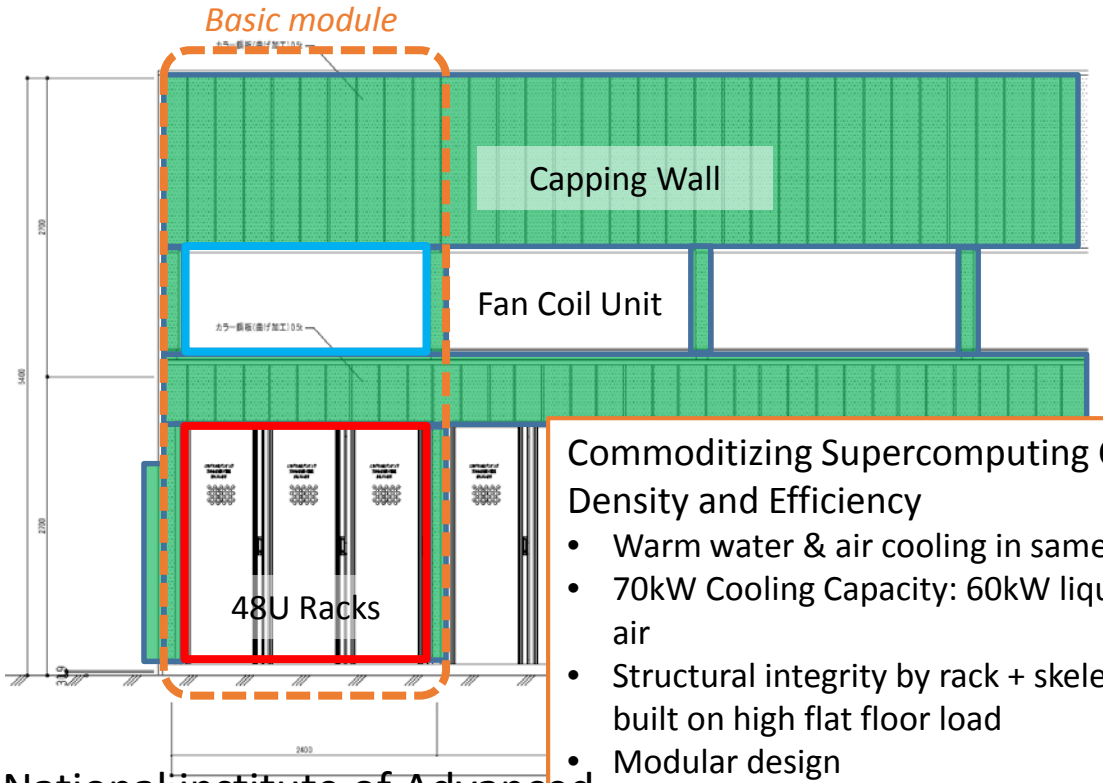
Cooling Pod

Implementing 70kW/Rack in AI Cloud Datacenter

front view



side view



Commoditizing Supercomputing Cooling Density and Efficiency

- Warm water & air cooling in same rack
- 70kW Cooling Capacity: 60kW liquid + 10KW air
- Structural integrity by rack + skeleton frame built on high flat floor load
- Modular design

National institute of Advanced Industrial Science and Technology


January 16, 2017


October 30, 2017


Jan 2018


Jan 2018


Feb 2018


Apr 2018
FCU-2R-01
FCU-2R-02

	AAIC（AIST AI Cloud）	TSUBAME3.0	ABCI（AI橋渡しクラウド）	Summit	TPU 3.0 Pod
研究機関	産総研	東工大	**産総研**	ORNL	Google
運用開始年	2017	2017	**2018**	2018	2018
ノード数	50	540	**1088**	4608	unknown
スループットプロセッサ	NVIDIA Tesla P100	NVIDIA Tesla P100	**NVIDIA Tesla V100**	NVIDIA Tesla V100	TPU 3.0
GPU数	400	2160	**4352**	27648	unknown
理論性能(FP64)	2.2 PF	12.2 PF	**37.2 PF**	200 PF	unknown
理論性能(DL)	8.6 PF	47.2 PF	**550 PF**	3.3 EF	100 PF / Pod
TOP500*	287	19	**5 (日本一位)**	1	unknown
GREEN500*	7	6	**8**	5	unknown
ノード数 / ラック	6	36	**34**	16	unknown
GPU数 / ラック	48	144	**136**	96	unknown
kW / ラック	22 kW	64.8 kW	**67.33 kW**	45-55 kW (est.)	unknown
DL性能 / ラック	0.9 PF	3.1 PF	**17 PF**	12 PF	12.5 PF (100 PF / 8 rack)

(*2018年6月時点)

Comparing ABCI to Classical IDC
AI IDC CAPX/OPEX accelerartion by > x100

Perf >
x400~600
Power Eff >
x200~300




Traditional Xeon IDC

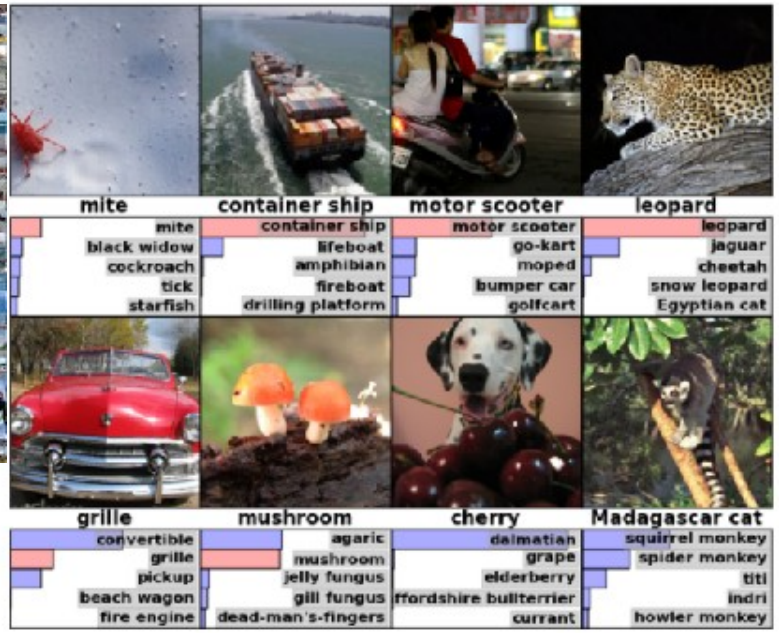
~10KW/rack PUE 1.5~2
15~20 1U Xeon Servers
2 Tera AI-FLOPS(SFP) / server
30~40 Tera AI-FLOP / rack
Low cooling efficiency

ABCI “Open Source” IDC

~70KW/rack PUE 1.0x
~500 Tera AI-FLOPS(HFP) / server
~17 Peta AI-FLOPs / rack
Inexpensive, very high cooling efficiency (PUE~1.1)

Training ImageNet in Minutes

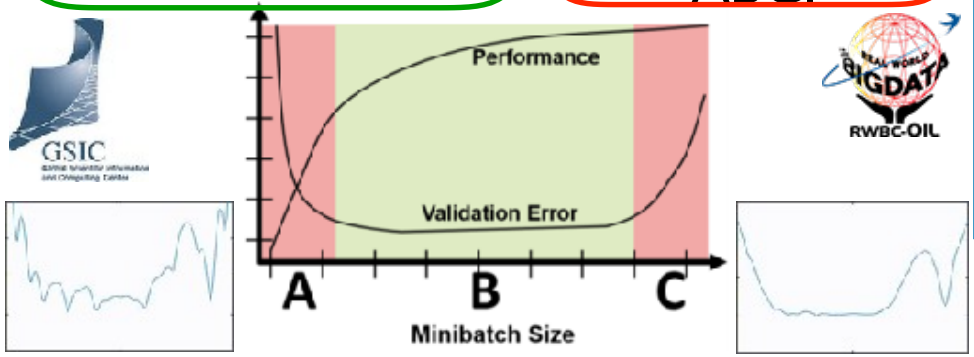
Rio Yokota, Kazuki Osawa, Yohei Tsuji, Yuichiro Ueno, Hiroki Naganuma, Shun Iwase, Kaku Linsho
Tokyo Institute of Technology
+Akira Naruse (NVIDIA)



TSUBAME3.0

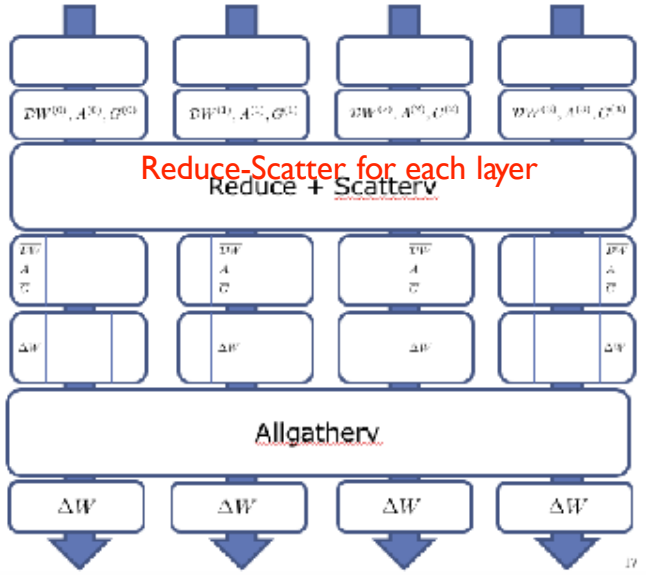
ABCI

	GPU/KNL	Time
Facebook	512	30 min
Preferred Networks	1024	15 min
UC Berkeley	2048	14 min
Tencent	2048	6.6 min
Tokyo Tech + AIST	4096	? min



Source Ben-nun & Hoefler https://arxiv.org/pdf/1802.09941.pdf

Parallel Scalability

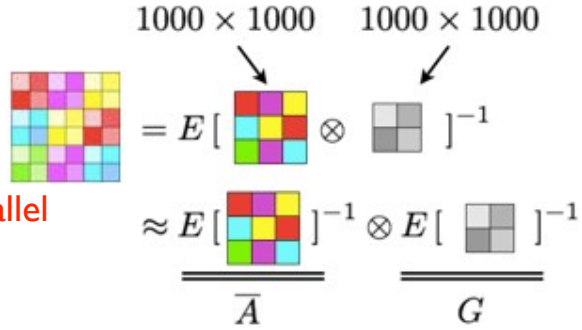


Reduce-Scatter for each layer

Inter-node RabenseifnerAllReduce

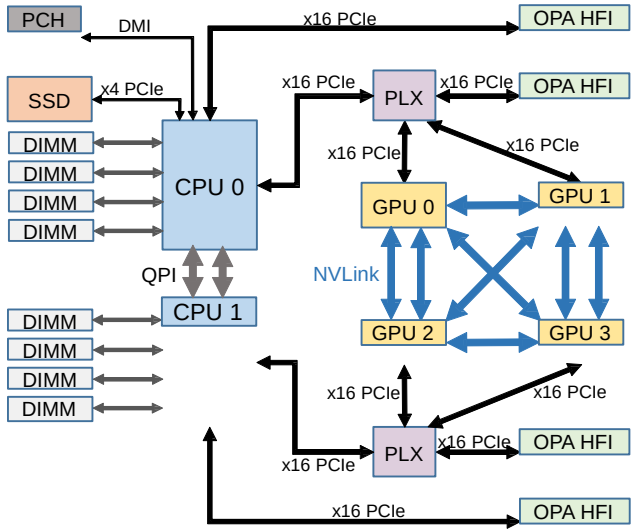
Cross-over from data-parallel to model-parallel

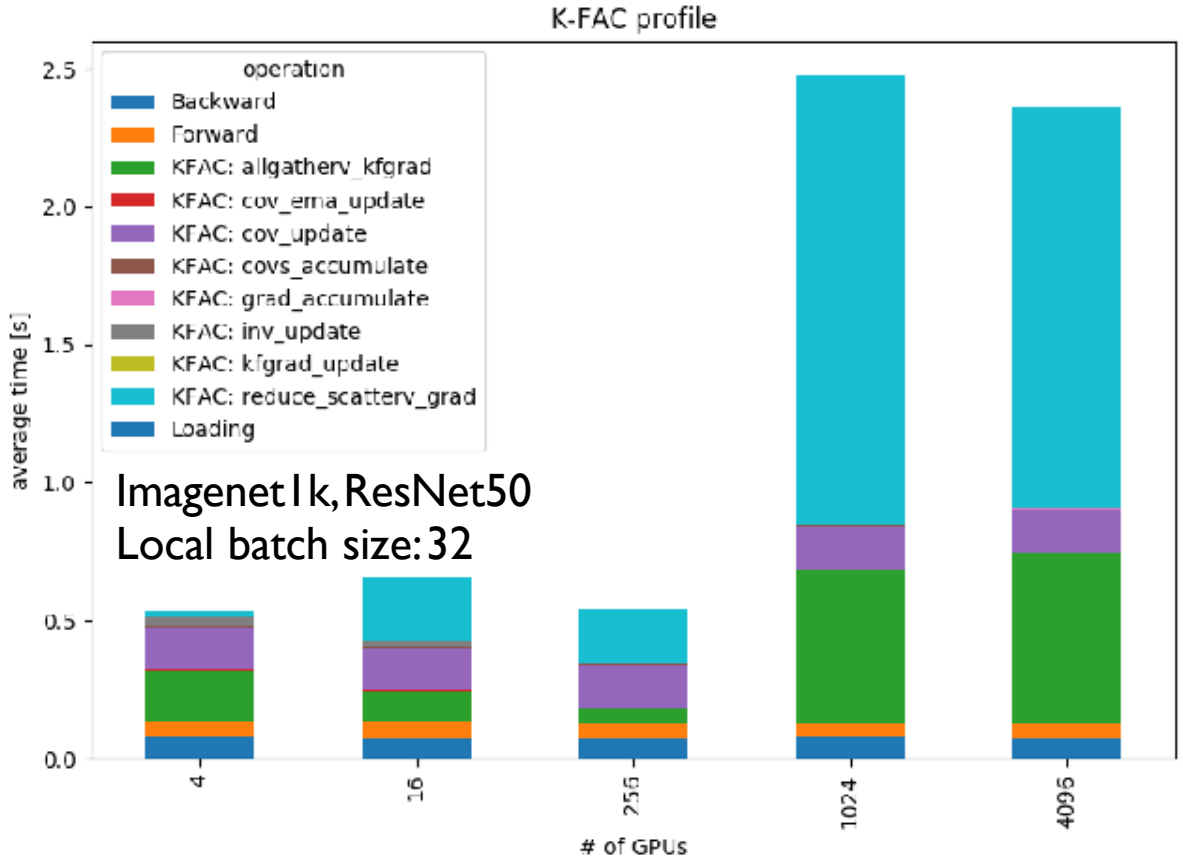
Cross-over from model-parallel to data-parallel



Kronecker Factorization

Intra-node RingAllReduce





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

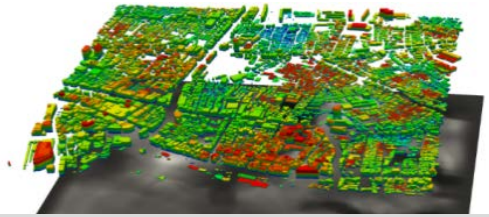
国内外の各機関、大学、企業の研究者との協力を更に拡大

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

R-CCSの特筆すべき業績等

次世代地震被害予測システムのコア技術で受賞

「京」全体で開発ソフトウェアGAMERAを実行し、**従来の1,000倍以上の規模**の地盤振動問題を解くことで、高性能計算技術の**世界最高峰の国際会議SC14,SC15でゴードン・ベル賞ファイナリストに選出**され、**SC16,SC17で最優秀ポスター賞**を受賞。



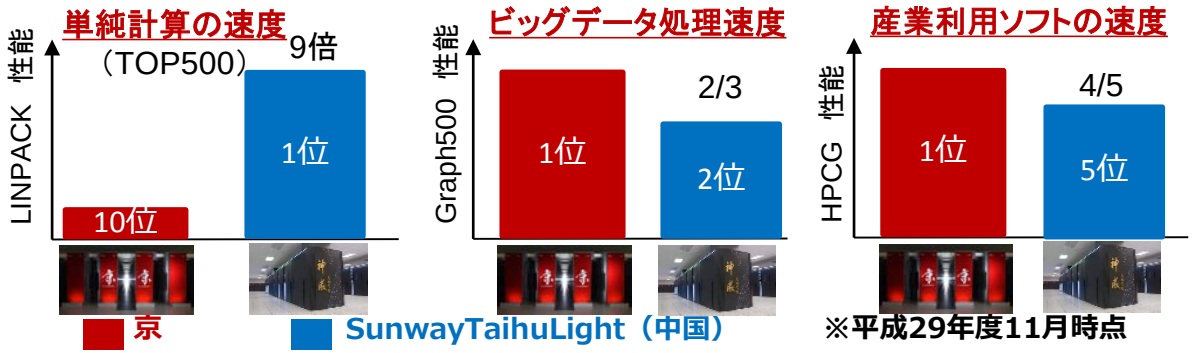
都市圏を丸ごとモデル化

災害被害シミュレーション

（社会科学）復旧シミュレーション

Graph500,HPCGで世界1位を獲得

「京」の性能を引き出す独自のアルゴリズムを研究開発した成果として、大規模グラフ解析に関するスパコン性能指標**Graph500**で、平成27年6月から平成29年11月まで**6期連続（通算7期）**で**世界1位を獲得**。産業利用ソフトで用いる計算手法の処理速度に関するスパコン性能指標**HPCG**で、平成28年11月から平成29年11月まで**3期連続**で**世界1位を獲得**。



※平成29年度11月時点

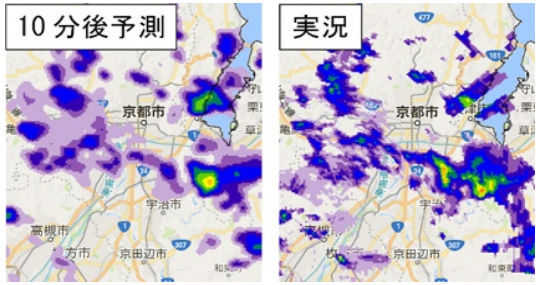
スパコン向けソフト開発を大幅に容易に

並列化や最適化のプログラミングを自動化することで大規模スパコン向けシミュレーションソフトの開発コストを大幅削減。
高性能計算技術の**世界最高峰の国際会議SC16で最優秀論文賞**を受賞。

ソフト開発コストの大幅削減等が期待される。

「3D降水ナウキャスト手法」を用いた降水予報

解像度100mで30秒ごとに新しい観測データを取り込んでリアルタイムに予測を実行するシステム（3D降水ナウキャスト手法）を構築し、**世界初となる30秒更新10分後**までの降水予報を開始した。

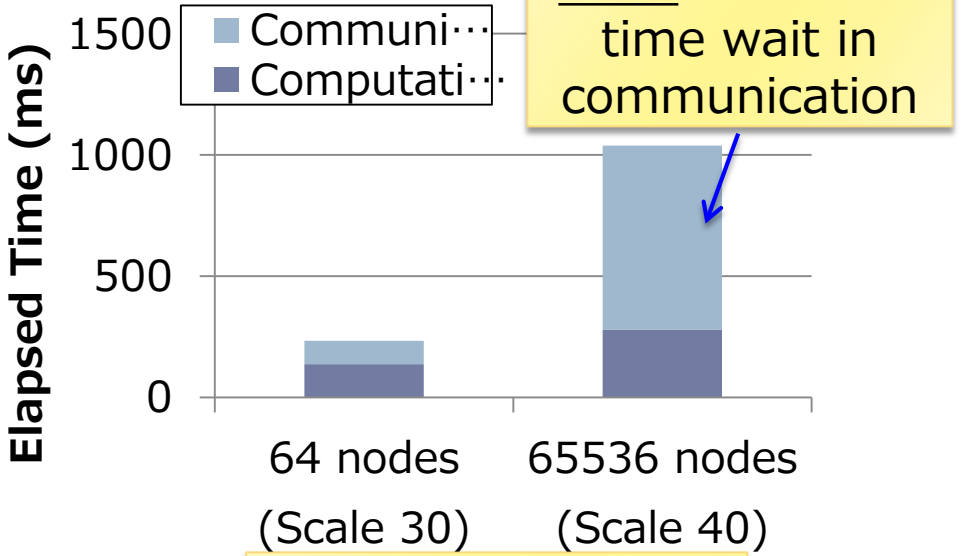


将来、超高速かつ超高精細な天気予報が可能になると期待される。

Sparse BYTES: The Graph500 – 2015~2016 – world #1 x 4

K Computer #1 Tokyo Tech[Matsuoka EBD CREST] Univ.

Kyushu [Fujisawa Graph CREST], Riken AICS, Fujitsu



73% total exec time wait in communication

88,000 nodes,
660,000 CPU Cores
1.3 Petabyte mem
20GB/s Tofu NW

K computer

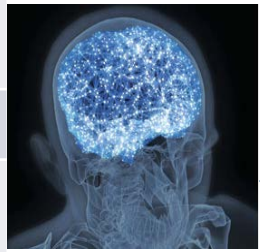
Effective x13 performance c.f. Linpack

LLNL-IBM Sequoia
1.6 million CPUs
1.6 Petabyte mem

TaihuLight
10 million CPUs
1.3 Petabyte mem

List	Rank	GTEPS	Implementat
November 2013	4	5524.12	Top-down o
June 2014	1	17977.05	Efficient hybrid
November 2014	2	19585.2	Efficient hybrid
June 2015 June 2018	**1**	**38621.4**	Hybrid + Node Compression

BYTES Rich Machine + Superior BYTES algoithm



BYTES, not FLOPS!

ポスト「京」～The Game Changer～

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

2. ポスト京の新たな技術イノベーション

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

◎**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＆Society5.0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

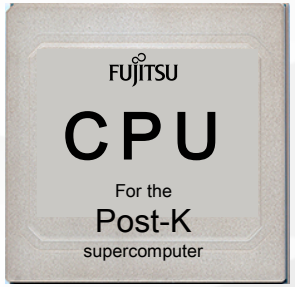
クラウドにも展開が容易、今回成果のSVEのグローバルスタンダード化

ARM：スマホ等の機器、車載チップ等で大きな実績

◎**Society5.0アプリへの展開**

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並みからそれ以上の性能を発揮すると期待。



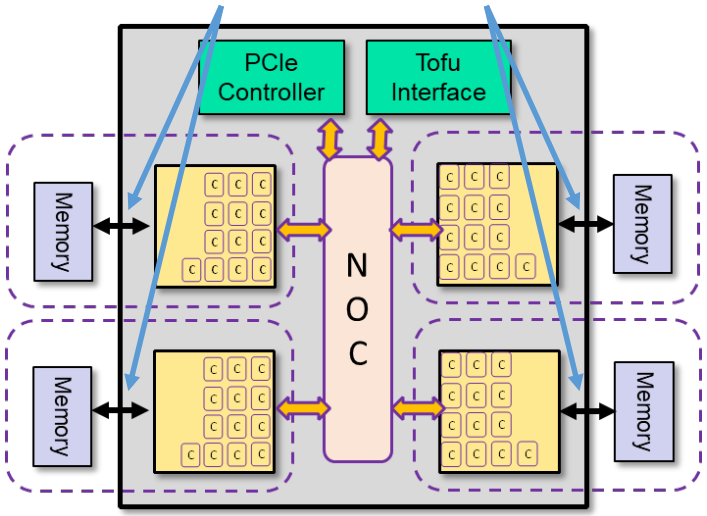
ポスト「京」は“二刀流”➡ビッグデータ・AIへの展開

更にインフラ、および技術としても世界に展開普及も

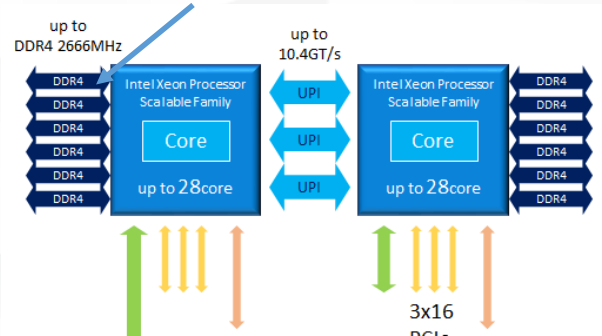
1. 高いメモリバンド幅によるHPCアプリの高演算性能

- **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の設計
- **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クラスの性能や電力性能比
  - 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



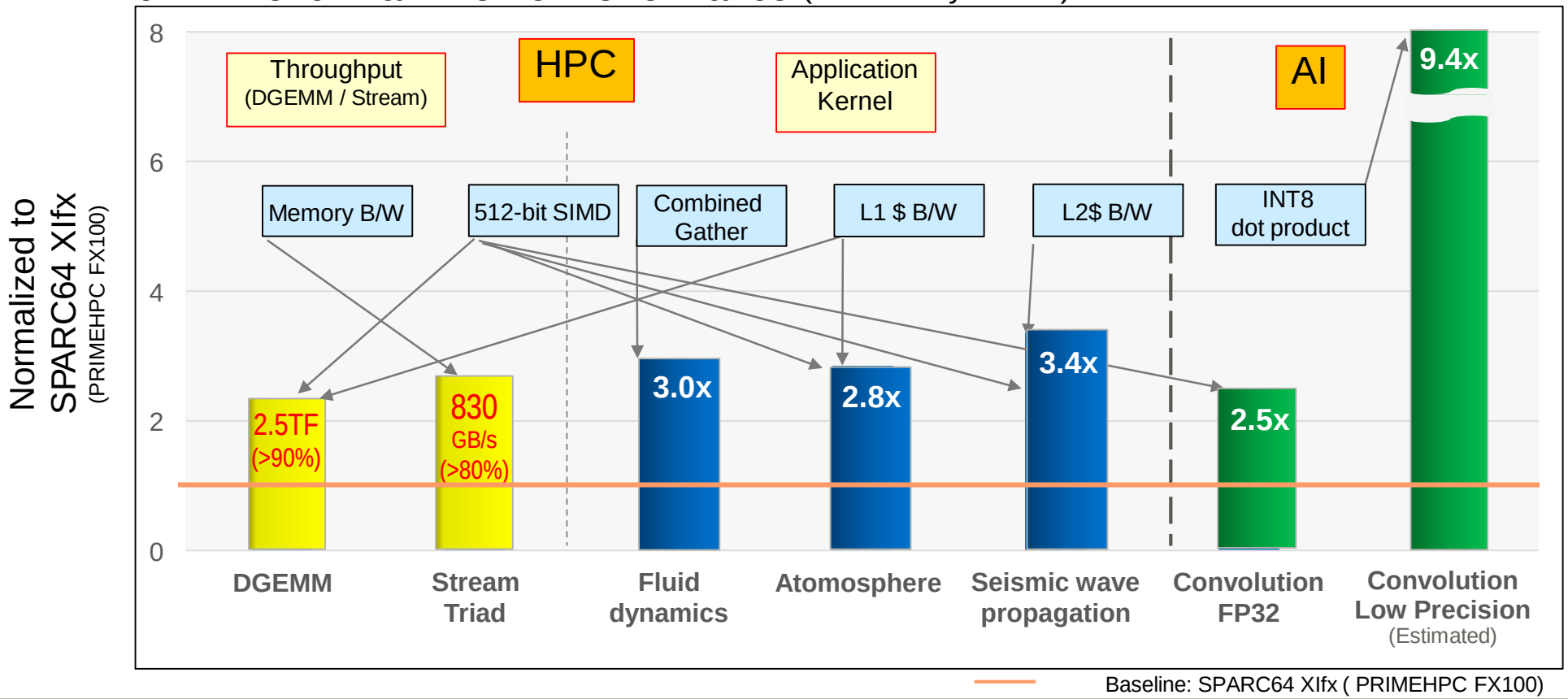
Skylake Xeon
6ch DDR4 2666
21.3GB/s x 6 / chip



A64fx Performance - A0 Silicon

- A64FX boosts performance up by microarchitectural enhancements, 512-bit wide SIMD, HBM2 and process technology
  - > 2.5x faster in HPC/AI benchmarks than SPARC64 XIfx (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)



Post-K A64fx A0 (ES) performance

	Performance / CPU					Machine Performance (HPC)		
	Peak TF (DFP)	Peak Mem. BW	Stream Triad	Theoretical B/F	DGEMM Efficiency	Linpack Efficiency	GF/W	Network BW Per Chip
Post-K A64fx (A0 Eng. Sample)	2.764/ 3.072	1024GB/s	840GB/s	0.37/ 0.33	94 %	87.7 %	GPU level	40.8GB/s(6.8x 6)
Intel KNL	3.0464	600GB/s	490GB/s	0.20	66%	54.4 %	4.9	12.5 GB/s
Intel Skylake	1.6128	127.8GB/s	97 GB/s	0.08	80 %	66.7 %	4.5	6.2GB/s
NVIDIA V100 DGX-1	7.8	900 GB/s	855GB/s	0.12		58 %	15.113	150GB/s 6.2GB/s

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**
  - 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

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

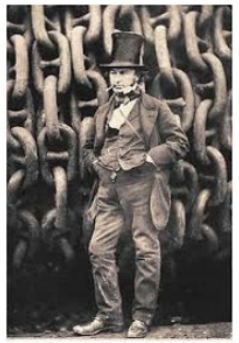
ARM HPC ECOSYSTEM事例: 英国
(slides courtesy Prof. Simon Mcintosh-Smith)

- 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

University of BRISTOL | Isambard | GW4

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

University of BRISTOL | UNIVERSITY OF BATH | UNIVERSITY OF EXETER | CARDIFF UNIVERSITY / PRIFYSGOL CAERDYDD | Met Office | EPSRC | CRAY THE SUPERCOMPUTER COMPANY | ARM


I.K.Brunel 1804-1859

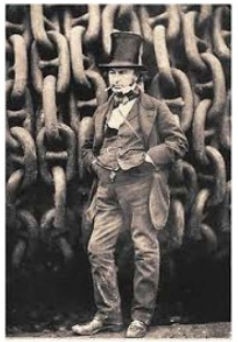
University of BRISTOL | Isambard | GW4

- 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

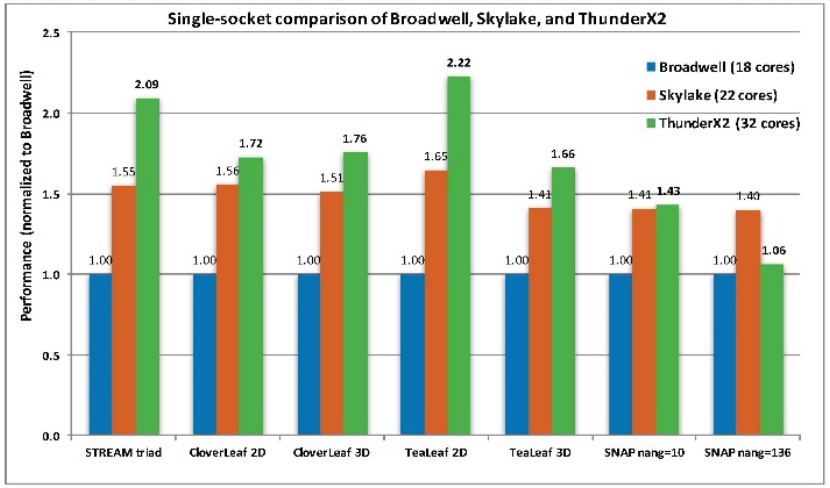
University of BRISTOL | Isambard | GW4

Isambard system specification (red = new info):

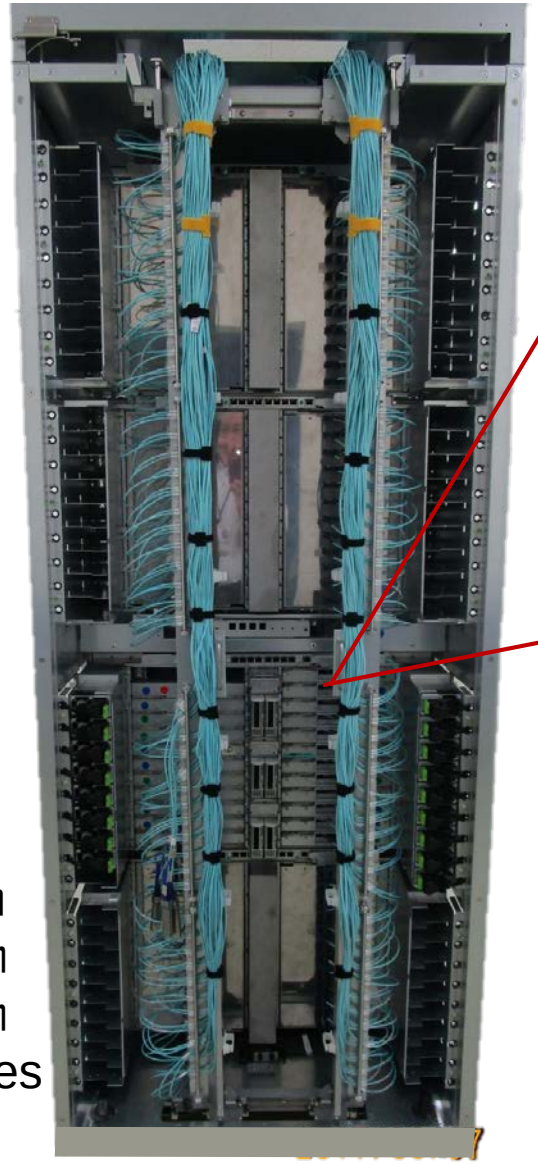
- Cray "Scout" system – XC50 series
  - Aries interconnect
- 10,000+ Armv8 cores
  - Cavium ThunderX2 processors
  - 2x 32core @ >2GHz per node
- Cray software tools
- Technology comparison:
  - x86, Xeon Phi, Pascal GPUs
- Phase 1 installed March 2017
- The Arm part arrives early 2018


I.K.Brunel 1804-1859

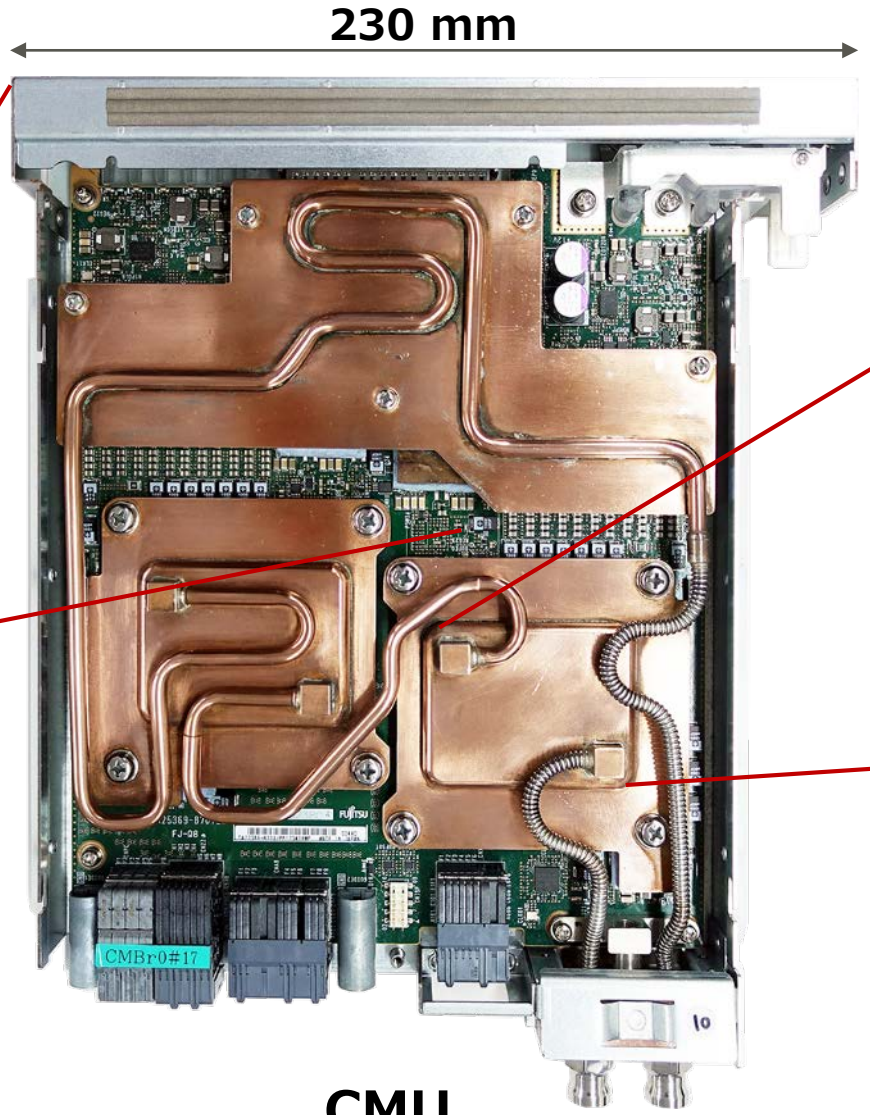
University of BRISTOL | Isambard | GW4



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



W 800mm
D1400mm
H2000mm
384 nodes



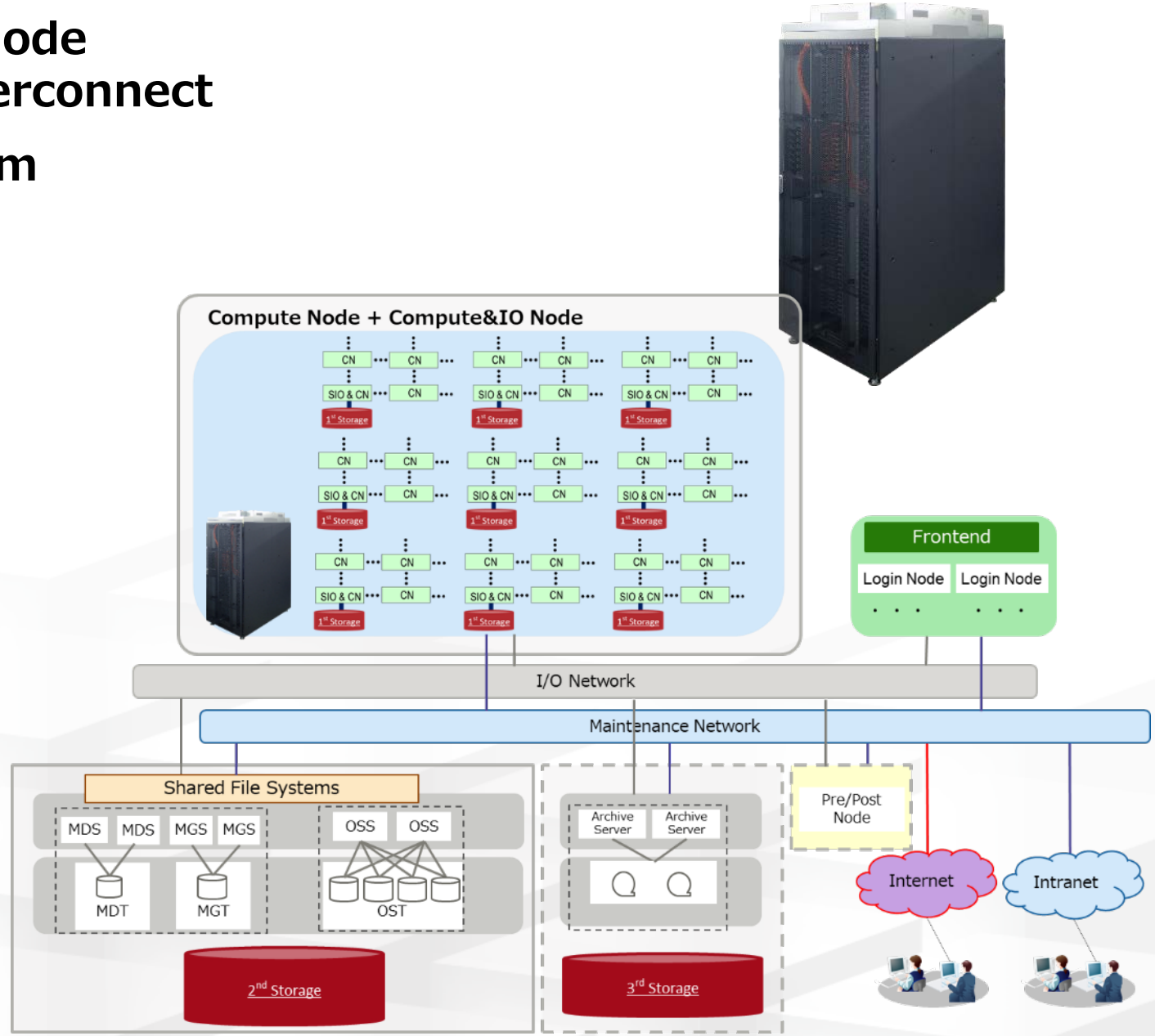
CMU

CPU Package

A0 Chip Booted in June
Undergoing Tests
B0 underway

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**



Society5.0：ポスト「京」の性能優位性

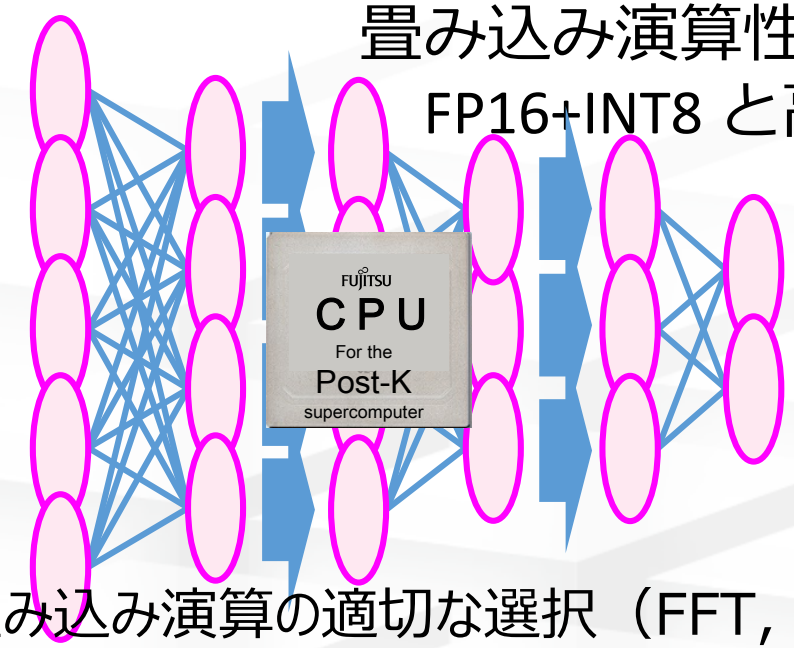
AI・機械学習の事例

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**

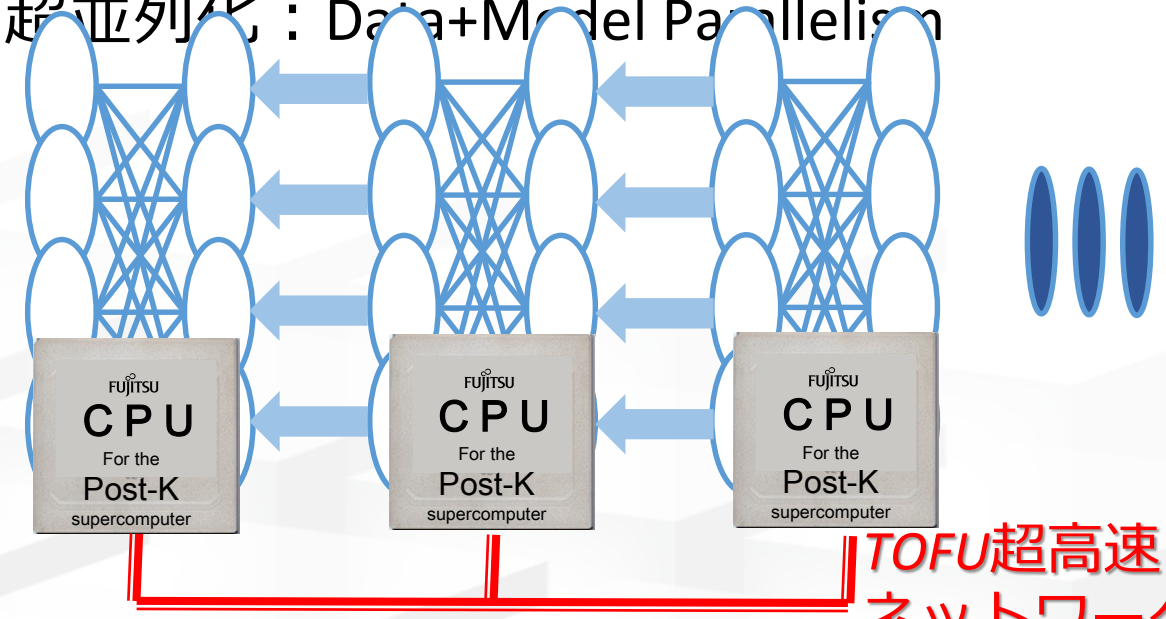
今後の大規模なデータ学習を必要とするAI研究において、世界トップのマシンとなる！

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



スケーラブルかつ低レー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

A64fx vs GPU vs TPU

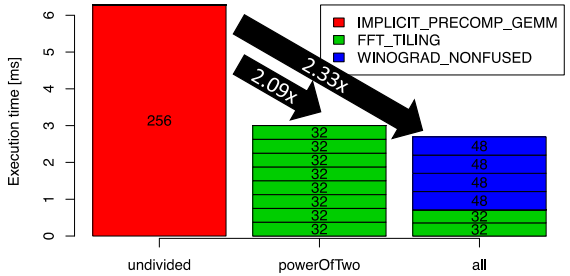
	A64fx CPU	GPU (Volta, Turing)	TPU ※Google オリジナルのプロセッサ
チップの性能	FP16+INT8、および高メモリバンド幅 ⇒FLOPS性能 〇 ⇒高メモリバンド幅による畳み込み◎	FLOPS演算性能　◎ 高メモリバンド幅による畳み込み◎	演算性能　〇～◎ ただしproprietaryで入手不可能
通信スケーラビリティ	通信のスケーラビリティ　◎(数十万) 6-D Torus, 高インジェクションBW モデル、データ並列とも大規模にスケール	Up to 16 on DGX2◎ 大規模△ 16を超えると大幅にダウン、かつModel 並列には適さない	？？ (2D Proprietary Torus)、京ほどはスケールしない
ソフトウェアエコシステム	エコシステム：ARM　AI, HPCすべて◎ ほとんどのXeon向けのフレームワークは容易に移植可能 (ただし性能を得るには研究開発が必要)	AIエコシステム　◎ (AI以外では〇)	×　Google独自開発 TensorFlowのみ対象 他のアプリ領域の柔軟性が低く、開発したプログラムの他サービスへの移行困難

どのようにして適切な畳み込みカーネルを選択するか

- **ETHとの共同研究：Micro Batching [Oyama, Tan, Hoefler & Matsuoka, IEEE Cluster2018]**
  - DNNにおいて、micro-batch技法を用い適切な畳み込みカーネルを自動選択
  - 速度とメモリのバランスを最適化
  - GPU上の実験では、多くの場合GEMMでなくWinograd or FFTが選ばれる
  - cuDNNのラッパーなので、全てのDNNフレームワークに透過的に適用可能
  - ポスト京チップでは、(1)更にWinograd/FFTの選択のケースが増える、(2)その場合GPUと速度は同じ、GEMMを盲目的に用いている場合よりは高速化

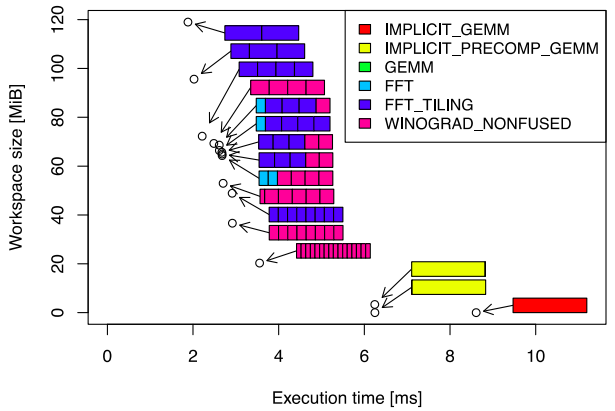
Evaluation: WR using Dynamic Programming

- μ-cuDNN achieved **2.33x** speedup on forward convolution of AlexNet conv2



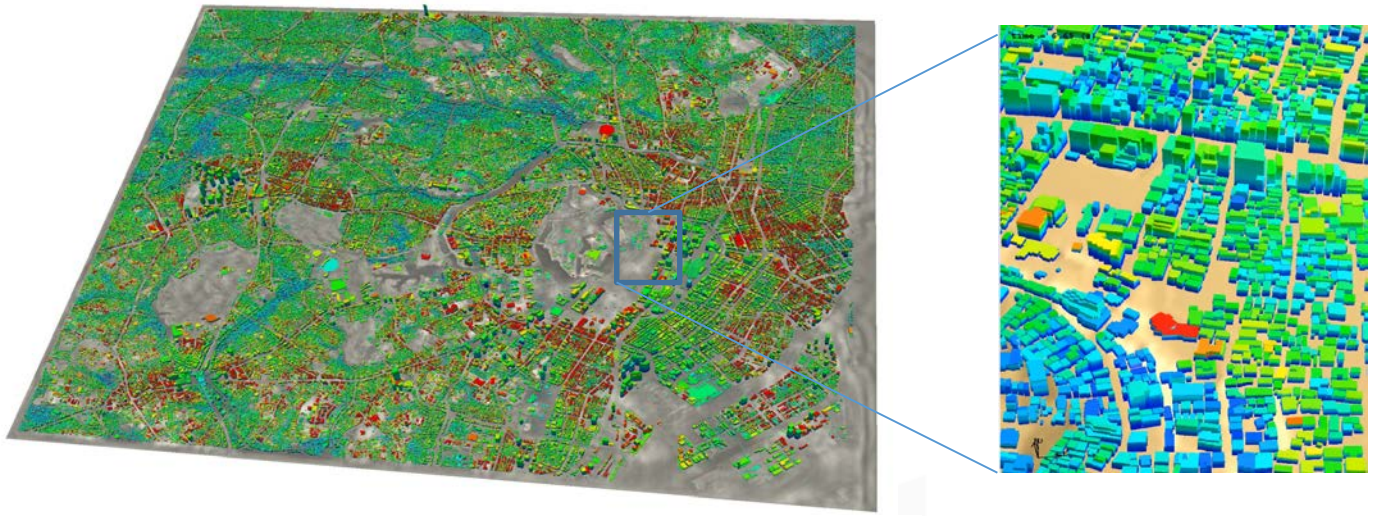
`cudnnConvolutionForward` of AlexNet conv2 on NVIDIA Tesla P100-SXM2
Workspace size of 64 MiB, mini-batch size of 256
Numbers on each rectangles represent micro-batch sizes

Evaluation: WD using Integer LP

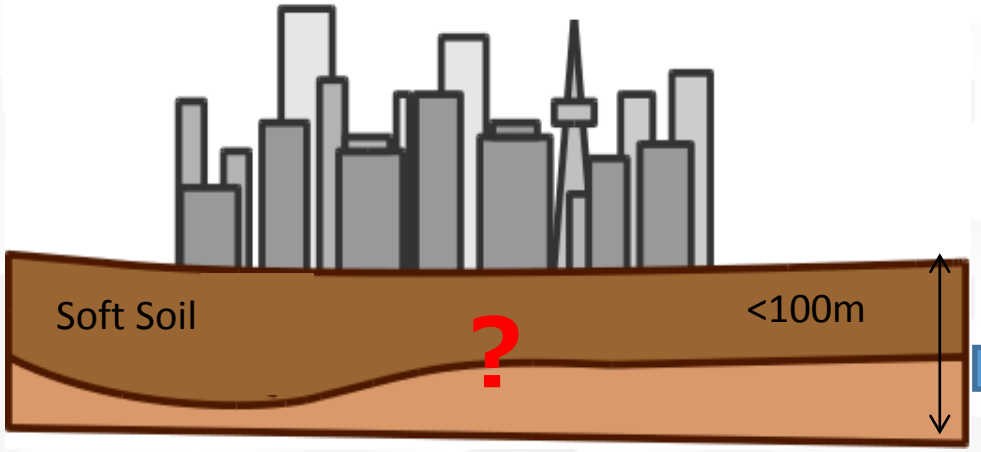


A desirable configuration set of AlexNet conv2 (Forward)
Mini-batch size of 256, P100-SXM2
Each bar represents proportion of micro-batch sizes and algorithms

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

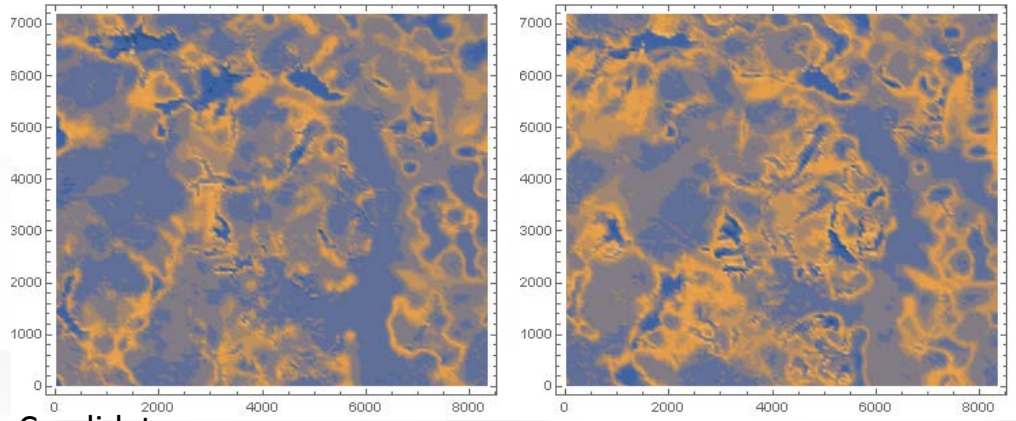


130 billion freedom earthquake of entire Tokyo on K-Computer (ACM Gordon Bell Prize Finalist, SC16,17 Best Poster)



Earthquake

Too Many Instances



Candidate Underground Structure 1

Candidate Underground Structure 2

AI Trained by Simulation to generate candidate soft soil structure

我が国の AI インフラの拡充に向けて
東工大・産総研・理研によるビッグデータ・AI の HPC加速
ここでも革新・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

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

x5.8

運用中

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

x5.8

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)

x2.8~4.2

運用中

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

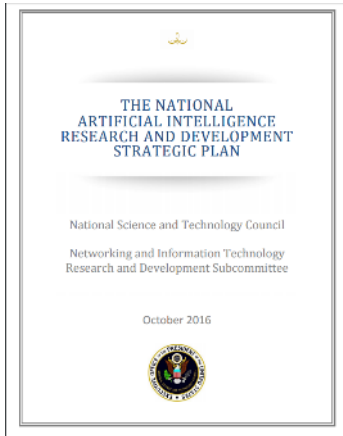
x5.0~7.7

2018年8月
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

3.5 年で1000倍以上

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」


THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN

National Science and Technology Council

Networking and Information Technology Research and Development Subcommittee

October 2016