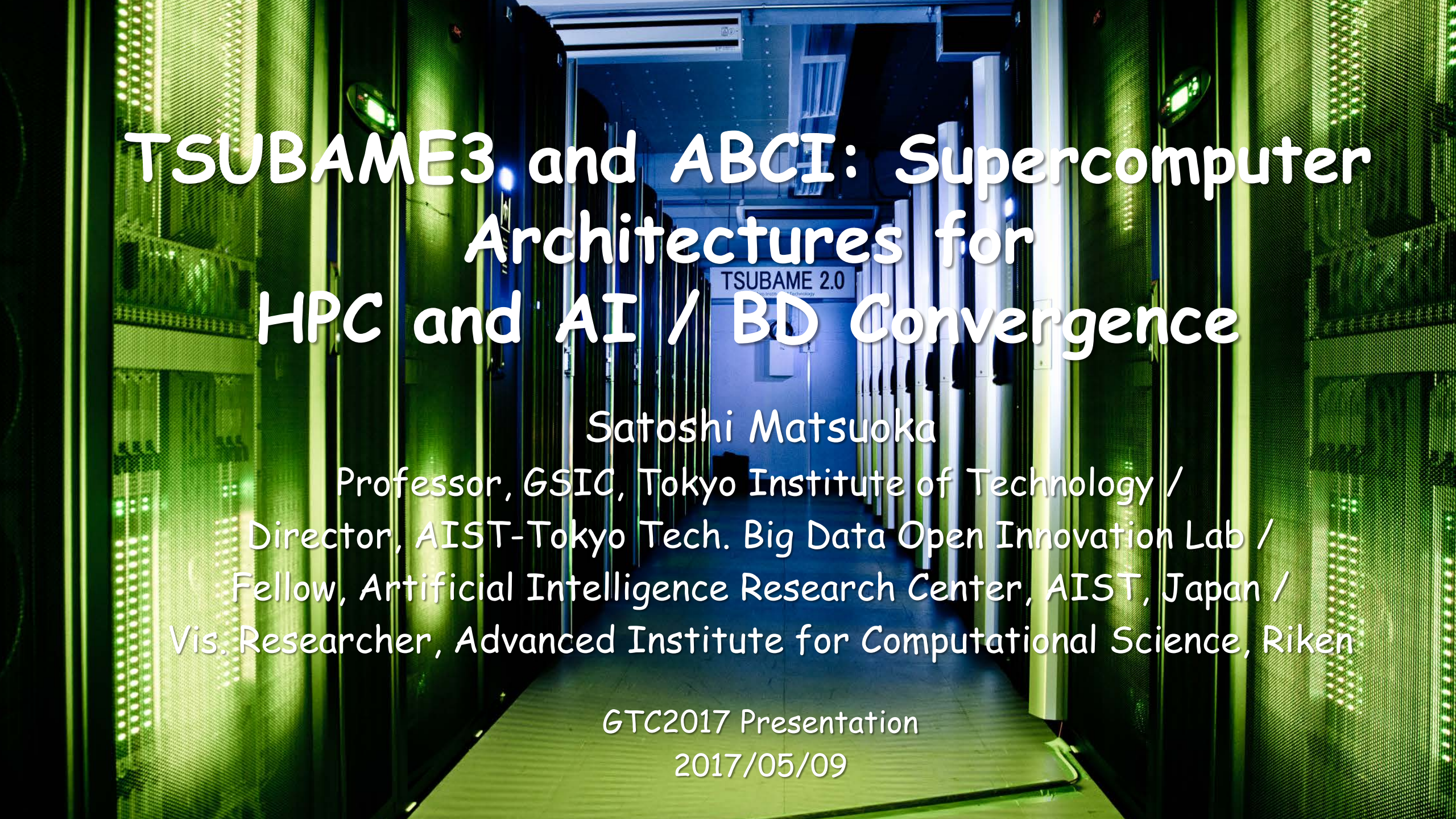


TSUBAME3 and ABCI: Supercomputer Architectures for HPC and AI / BD Convergence

Satoshi Matsuoka
Professor, GSIC, Tokyo Institute of Technology /
Director, AIST-Tokyo Tech. Big Data Open Innovation Lab /
Fellow, Artificial Intelligence Research Center, AIST, Japan /
Vis. Researcher, Advanced Institute for Computational Science, Riken

GTC2017 Presentation
2017/05/09

Tremendous Recent Rise in Interest by the Japanese Government on Big Data, DL, AI, and IoT

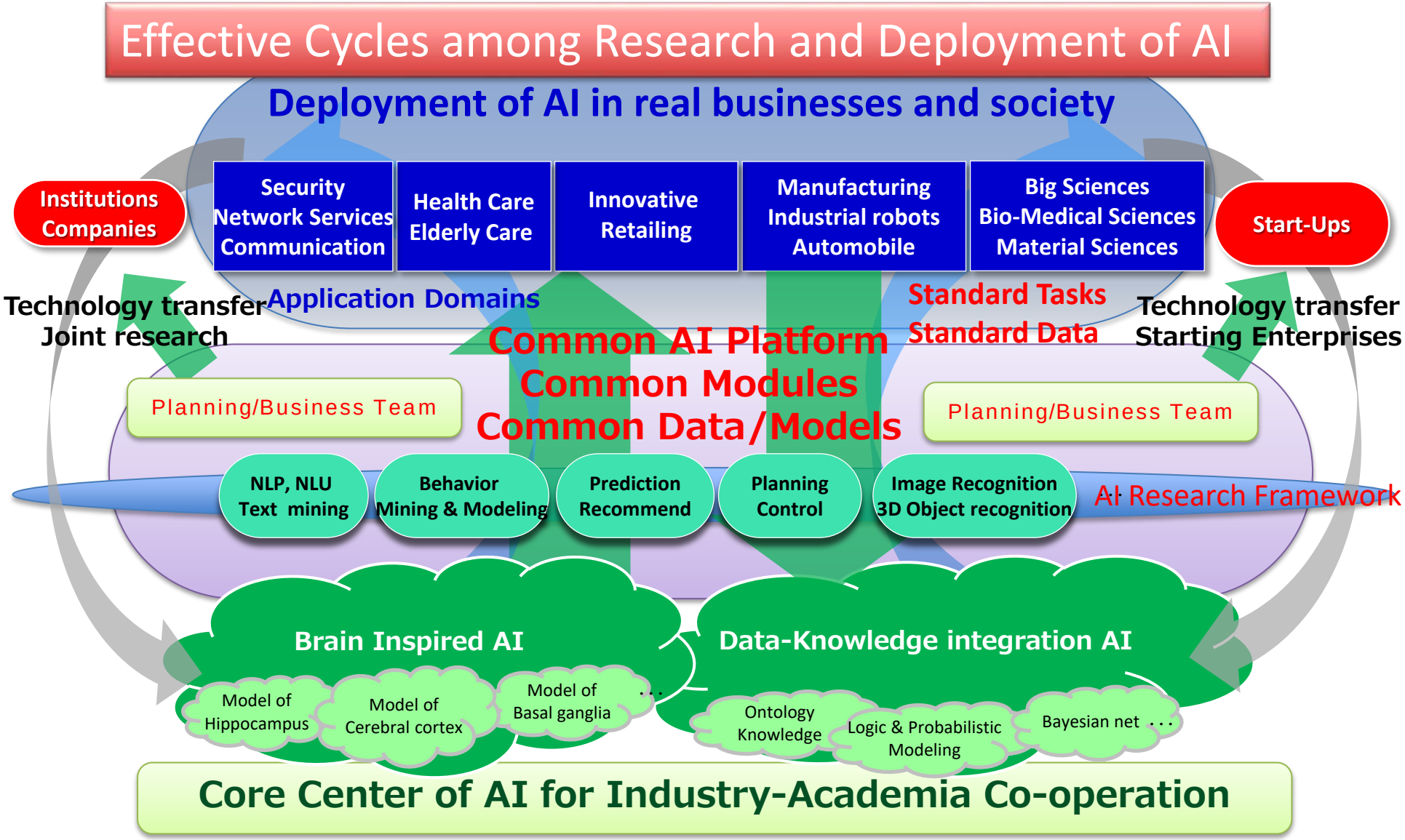
- Three national centers on Big Data and AI launched by three competing Ministries for FY 2016 (Apr 2015-)
  - METI – AIRC (Artificial Intelligence Research Center): AIST (AIST internal budget + > $200 million FY 2017), April 2015
    - Broad AI/BD/IoT, industry focus
  - MEXT – AIP (Artificial Intelligence Platform): Riken and other institutions ($~50 mil), April 2016
    - A separate Post-K related AI funding as well.
    - Narrowly focused on DNN
  - MOST – Universal Communication Lab: NICT ($50~55 mil)
    - Brain –related AI
  - $1 billion commitment on inter-ministry AI research over 10 years

Vice Minsiter Tsuchiya@MEXT Annoucing AIP estabishment

2015- AI Research Center (AIRC), AIST

Now > 400+ FTEs

Director:
Jun-ichi Tsujii



Matsuoka : Joint appointment as "Designated" Fellow since July 2017

National Institute for Advanced Industrial Science and Technology (AIST)

独立行政法人
産業技術総合研究所

経済産業省
Ministry of Economy, Trade and Industry

Ministry of Economics Trade and Industry (METI)

Joint Lab established Feb. 2017 to pursue BD/AI joint research using large-scale HPC BD/AI infrastructure

Tokyo Institute of Technology / GSIC

TSUBAME
Tokyo Institute of Technology

GSIC
Global Scientific Information and Computing Center

Resources and Acceleration of AI / Big Data, systems research

Tsubame 3.0/2.5 Big Data /AI resources

REAL WORLD BIGDATA
RWBC-OIL

AIST-Tokyo Tech Real World Big-Data Computation Open Innovation Laboratory (RWBC-OIL)

Director: Satoshi Matsuoka

AIST Artificial Intelligence Research Center (AIRC)

AiRC

Application Area
Natural Langauge Processing
Robotics
Security

Joint Research on AI / Big Data and applications

Industrial Collaboration in data, applications

Basic Research in Big Data / AI algorithms and methodologies

ITCS Departments

ABCI
AI Bridging Cloud Infrastructure

Industry

YAHOO! JAPAN
IT LAB
DENSO
DENSO IT LABORATORY, INC.

Other Big Data / AI research organizations and proposals
JST BigData CREST
JST AI CREST
Etc.

Characteristics of Big Data and AI Computing

As BD / AI

Graph Analytics e.g. Social Networks

Sort, Hash, e.g. DB, log analysis

Symbolic Processing: Traditional AI

As HPC Task

Integer Ops & Sparse Matrices

Data Movement, Large Memory

Sparse and Random Data, Low Locality

Acceleration, Scaling

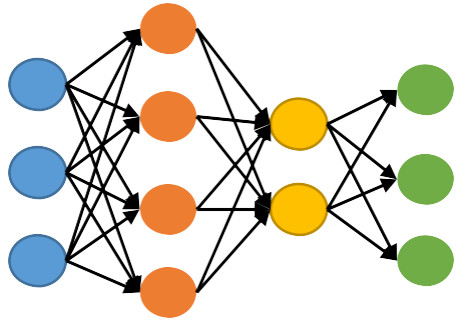
Opposite ends of HPC computing spectrum, but HPC simulation apps can also be categorized likewise

Acceleration via Supercomputers adapted to AI/BD

As BD / AI

Dense LA: DNN

Inference, Training, Generation



As HPC Task

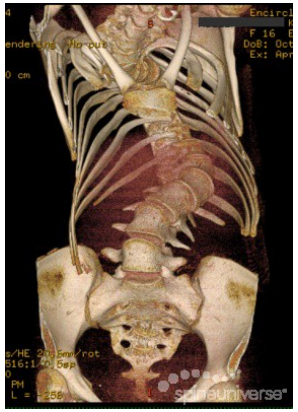
Dense Matrices, Reduced Precision

Dense and well organized neworks and Data

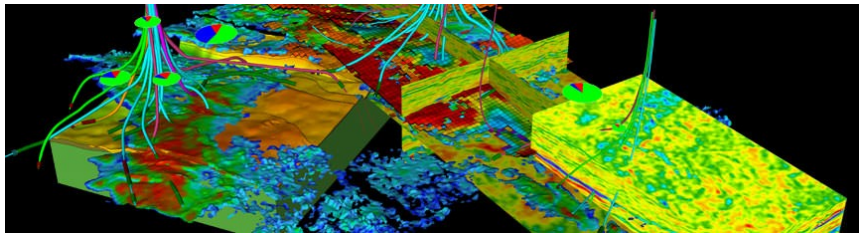
Acceleration, Scaling

(Big Data) BYTES capabilities, in bandwidth and capacity, unilaterally important but often missing from modern HPC machines in their pursuit of FLOPS…

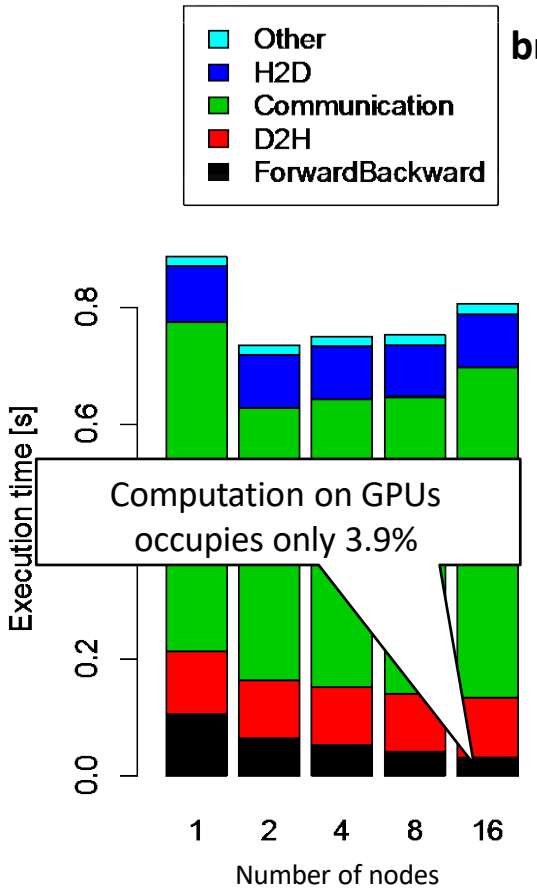
- **Need <u>BOTH bandwidth and capacity</u> (BYTES) in a HPC-BD/AI machine:**
  - Obvious for lefthand sparse ,bandwidth-dominated apps
  - But also for righthand DNN: Strong scaling, large networks and datasets, in particular for future 3D dataset analysis such as CT-scans, seismic simu. vs. analysis…)



(Source: https://www.spineuniverse.com/image-library/anterior-3d-ct-scan-progressive-kyphoscoliosis)



(Source: http://www.dgi.com/images/cvmain_overview/CV4DOverview_Model_001.jpg)



Our measurement on breakdown of one iteration of CaffeNet training on TSUBAME-KFC/DL (Mini-batch size of 256)

Proper arch. to support large memory cap. and BW, network latency and BW important

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

AI/BD Centers & Labs in National Labs & Universities

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

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

Web access and merchandice

Y! YAHOO! JAPAN

SoftBank

NTT

IoT Communication, location & other 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

"Big" Data

2017 Q2 TSUBAME3.0 Leading Machine Towards Exa & Big Data

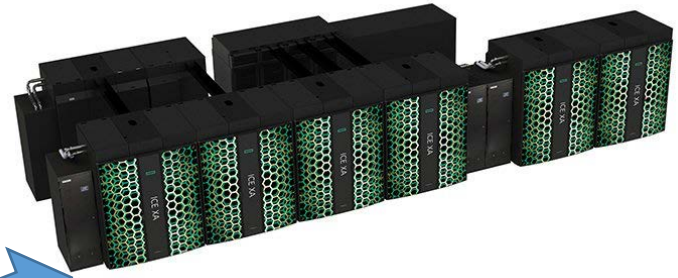
1. "Everybody's Supercomputer" - High Performance (12~24 DP Petaflops, 125~325TB/s Mem, 55~185Tbit/s NW), innovative high cost/performance packaging & design, in mere 180m^2…
2. "Extreme Green" – ~10GFlops/W power-efficient architecture, system-wide power control, advanced cooling, future energy reservoir load leveling & energy recovery
3. "Big Data Convergence" – BYTES-Centric Architecture, Extreme high BW & capacity, deep memory hierarchy, extreme I/O acceleration, Big Data SW Stack for machine learning, graph processing, …
4. "Cloud SC" – dynamic deployment, container-based node co-location & dynamic configuration, resource elasticity, assimilation of public clouds…
5. "Transparency" - full monitoring & user visibility of machine & job state, accountability via reproducibility

2006 TSUBAME1.0
80 Teraflops, #1 Asia #7 World
"Everybody's Supercomputer"

2010 TSUBAME2.0
2.4 Petaflops #4 World
"Greenest Production SC"

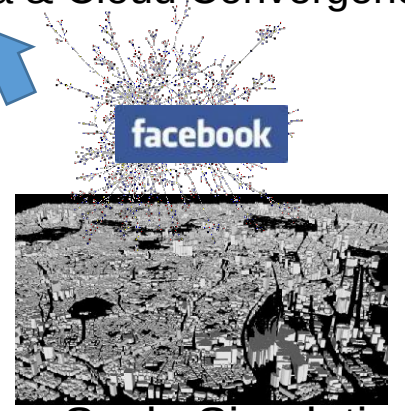
2011 ACM Gordon Bell Prize

2013
TSUBAME2.5
upgrade
5.7PF DFP
/17.1PF SFP
20% power
reduction



2013 TSUBAME-KFC
#1 Green 500

2017 TSUBAME3.0+2.5
~18PF(DFP) 4~5PB/s Mem BW
10GFlops/W power efficiency
Big Data & Cloud Convergence



Large Scale Simulation
Big Data Analytics
Industrial Apps

TSUBAME-KFC/DL: TSUBAME3 Prototype [ICPADS2014]

Oil Immersive Cooling＋ Hot Water Cooling + High Density Packaging + Fine-Grained Power Monitoring and Control, *upgrade to /DL Oct. 2015*

High Temperature Cooling
Oil Loop 35~45℃
⇒ Water Loop 25~35℃
(c.f. TSUBAME2: 7~17℃)

Cooling Tower:
Water 25~35℃
⇒ To Ambient Air

Single Rack High Density Oil Immersion
168 NVIDIA K80 GPUs + Xeon
413+TFlops (DFP)
1.5PFlops (SFP)
~60KW/rack

Container Facility
20 feet container (16m²)
Fully Unmanned Operation

THE GREEN 500
sponsored by SUPERMICRO

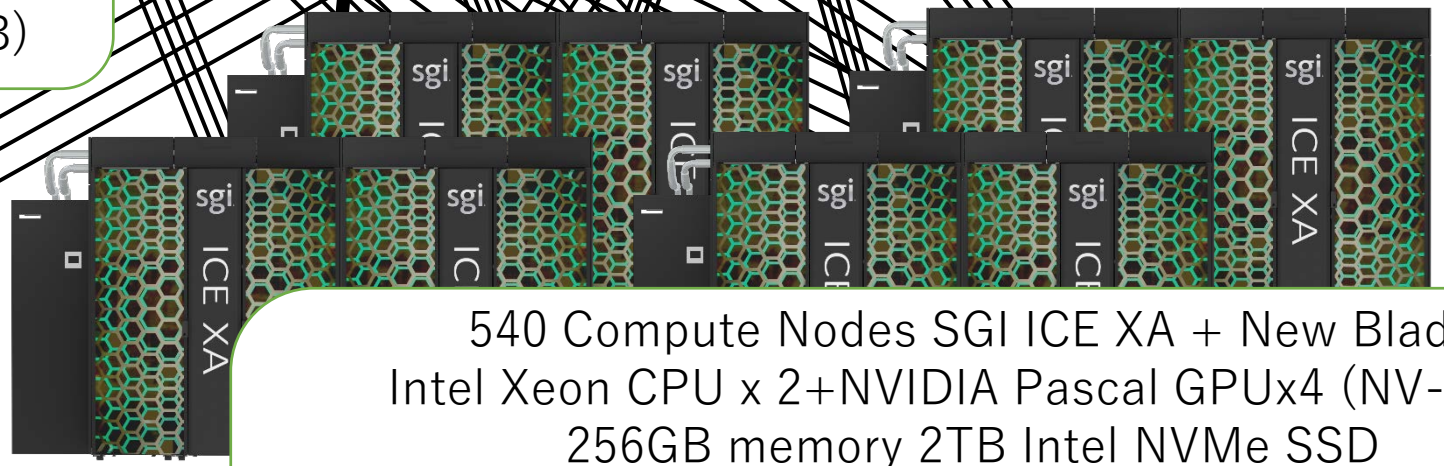
2013年11月/2014年6月
Word #1 Green500

Overview of TSUBAME3.0
BYTES-centric Architecture, Scalaibility to all 2160 GPUs, all nodes, the entire memory hiearchy

Full Operations
Aug. 2017

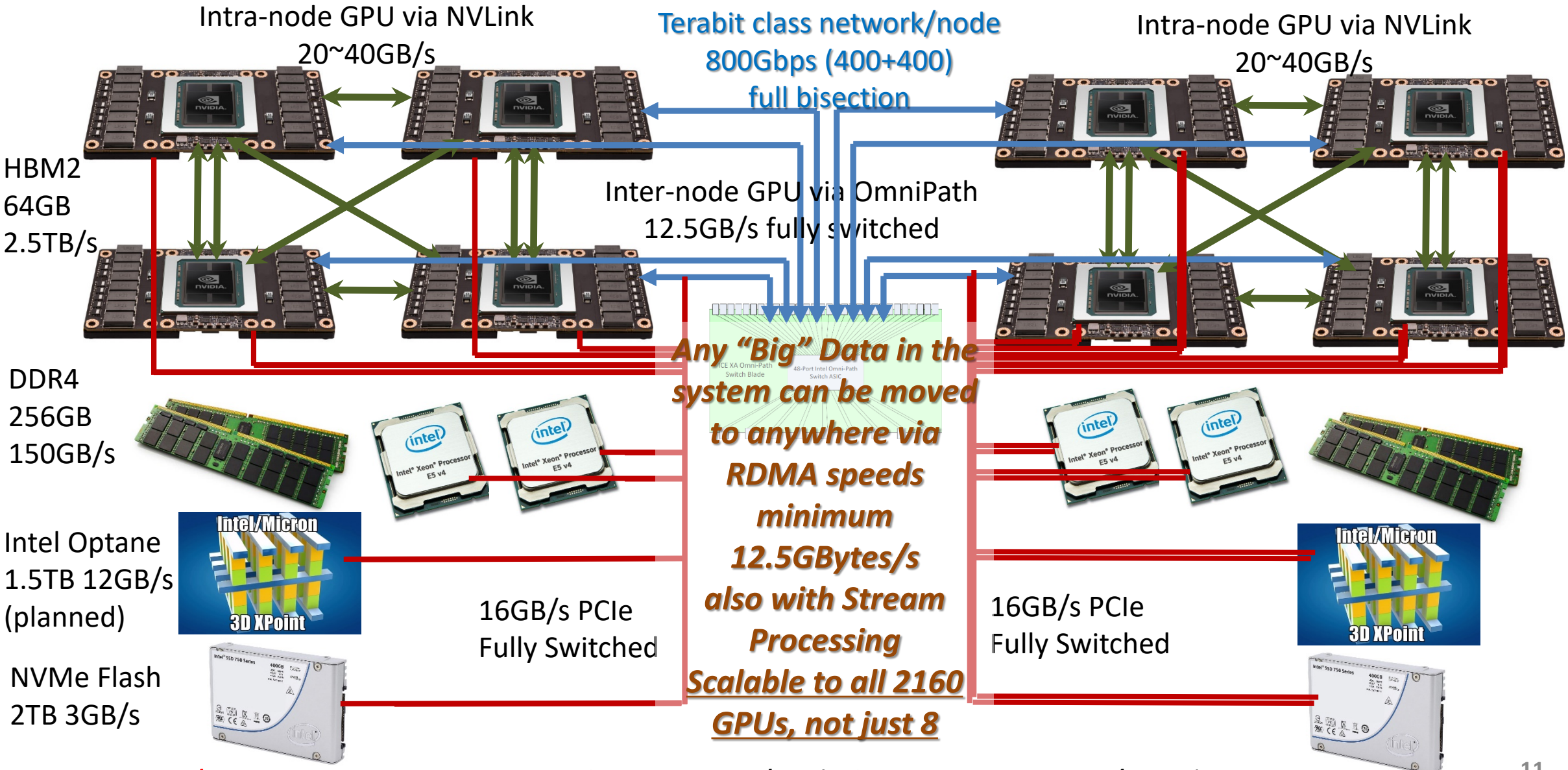
Full Bisection Bandwidgh
Intel Omni-Path Interconnect. 4 ports/node
Full Bisection / 432 Terabits/s bidirectional
~x2 BW of entire Internet backbone traffic

DDN Storage
(Lustre FS 15.9PB+Home 45TB)



540 Compute Nodes SGI ICE XA + New Blade
Intel Xeon CPU x 2+NVIDIA Pascal GPUx4 (NV-Link)
256GB memory 2TB Intel NVMe SSD
47.2 AI-Petaflops, 12.1 Petaflops

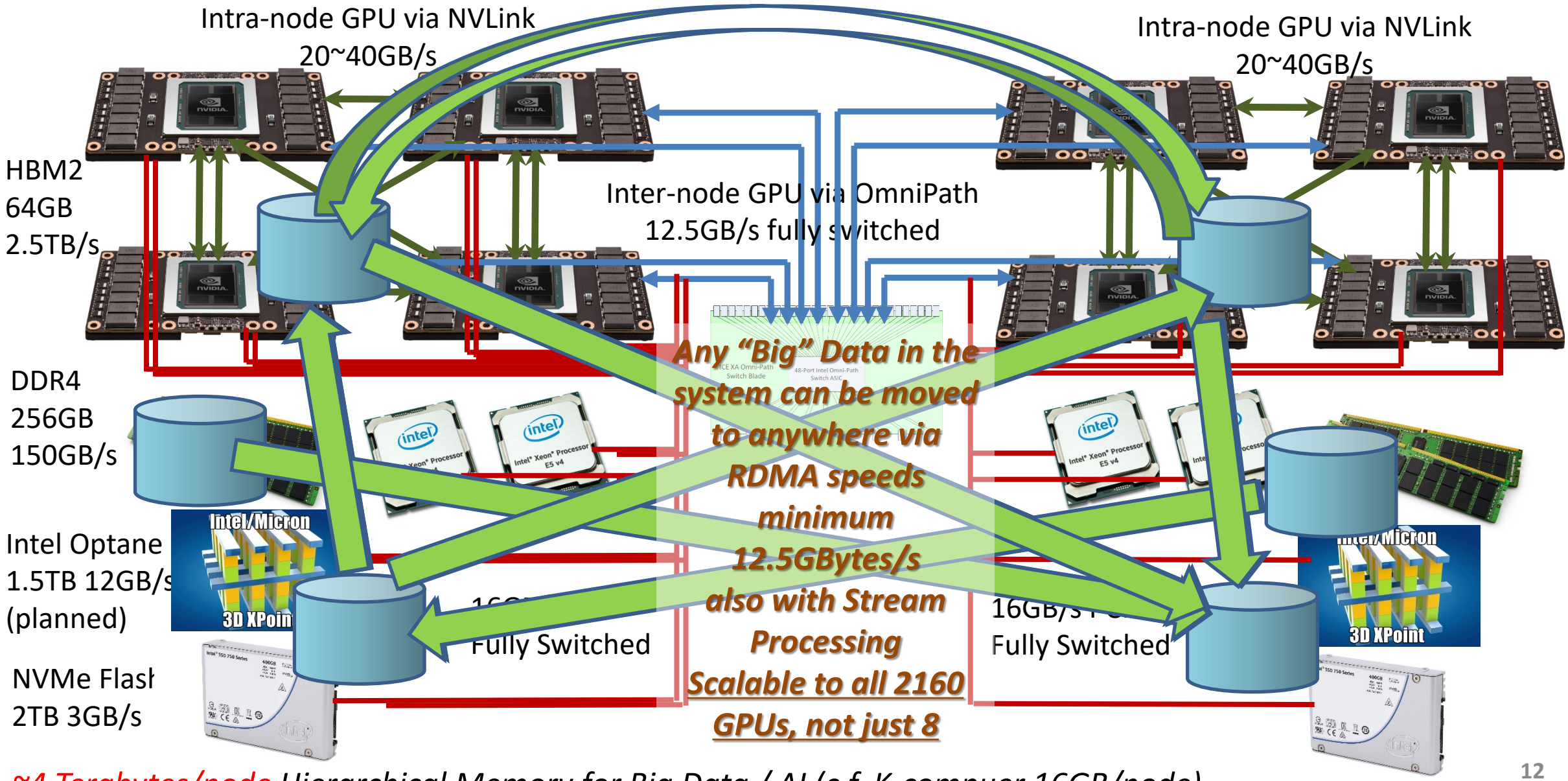
TSUBAME3: *A Massively BYTES Centric Architecture* for Converged BD/AI and HPC



~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

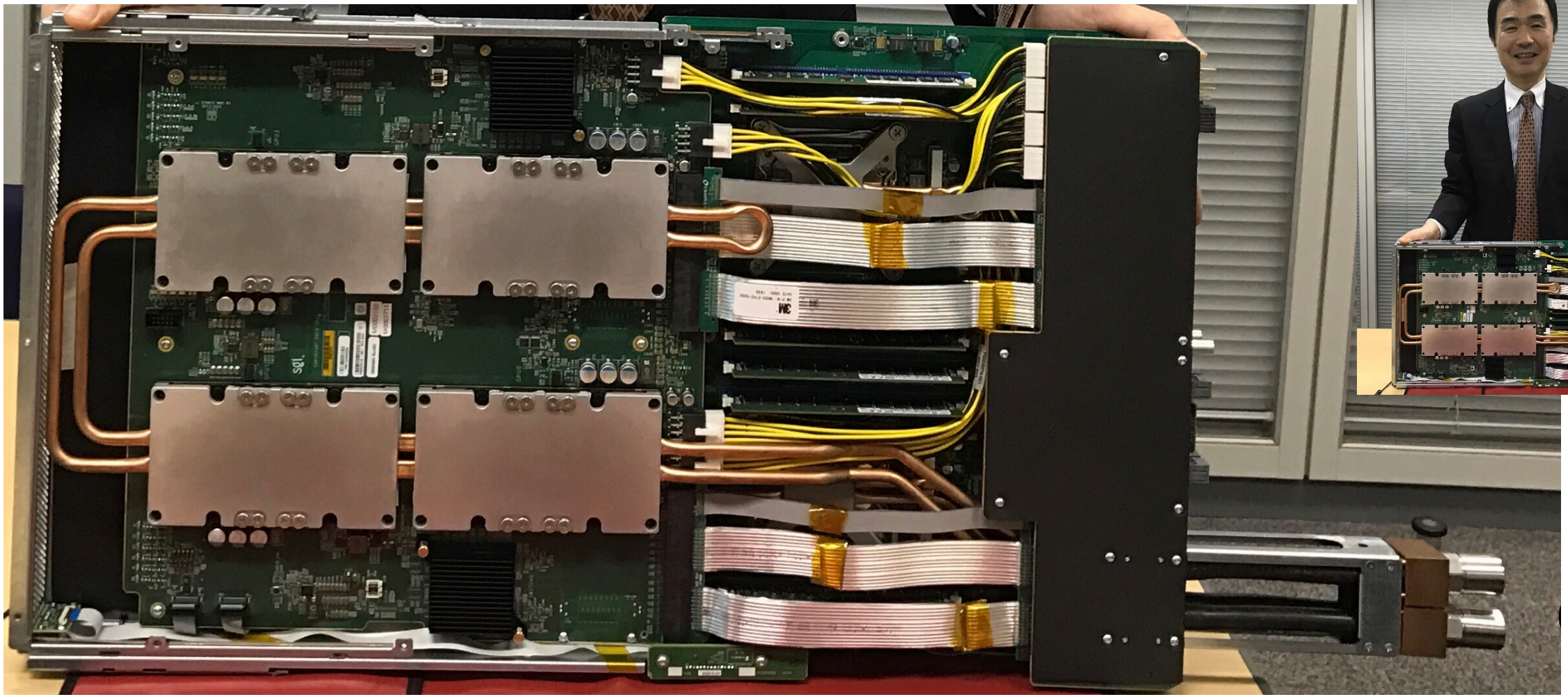


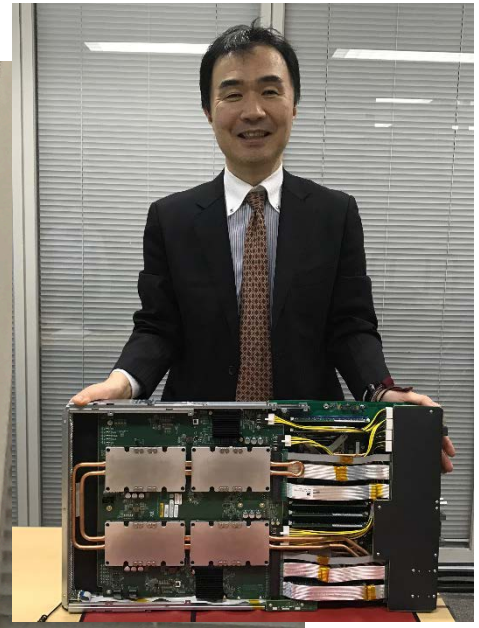
~4 Terabytes/node Hierarchical Memory for Big Data / AI (c.f. K-compuer 16GB/node)

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

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

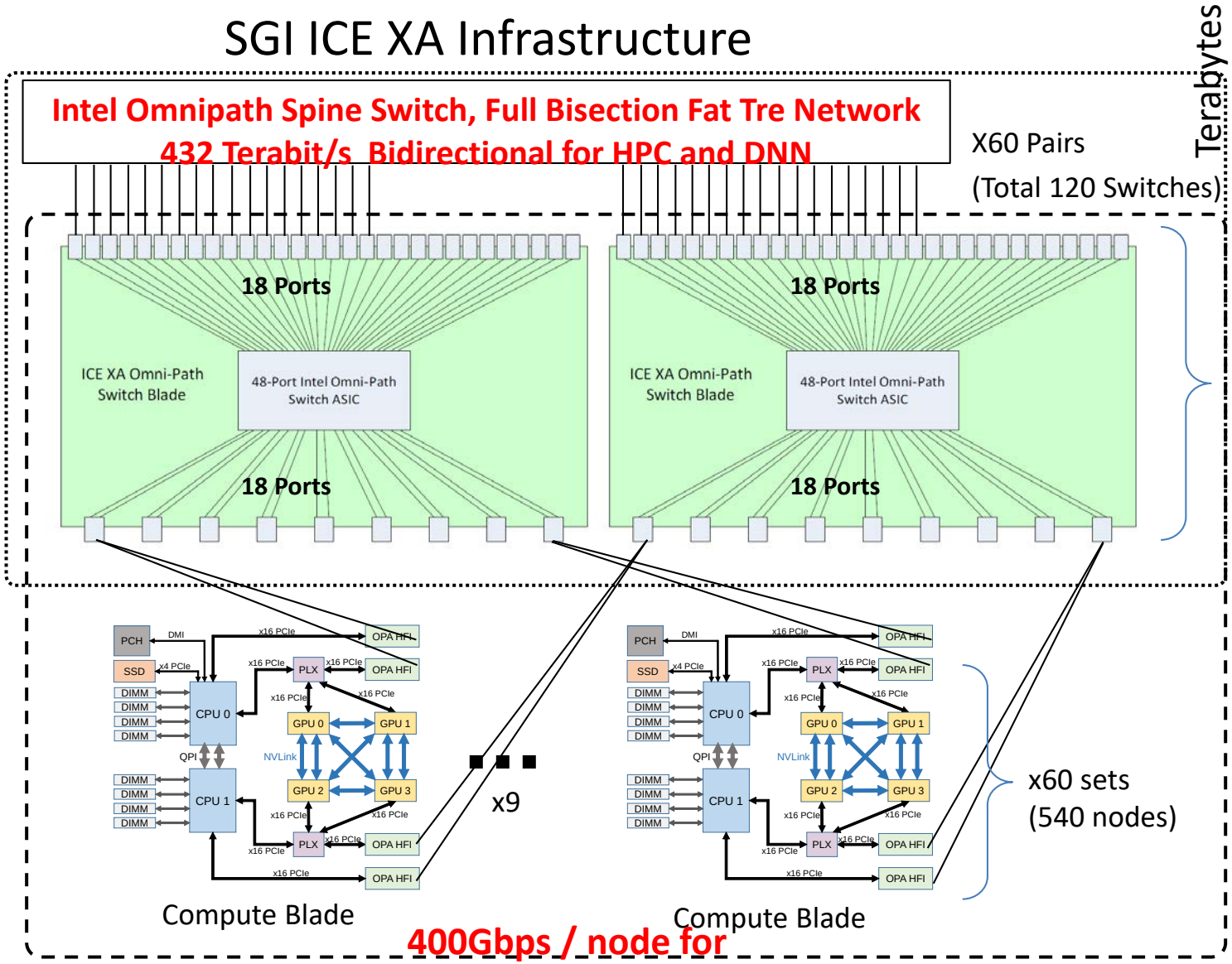
- No exterior cable mess (power, NW, water)
- Plan to become a future HPE product

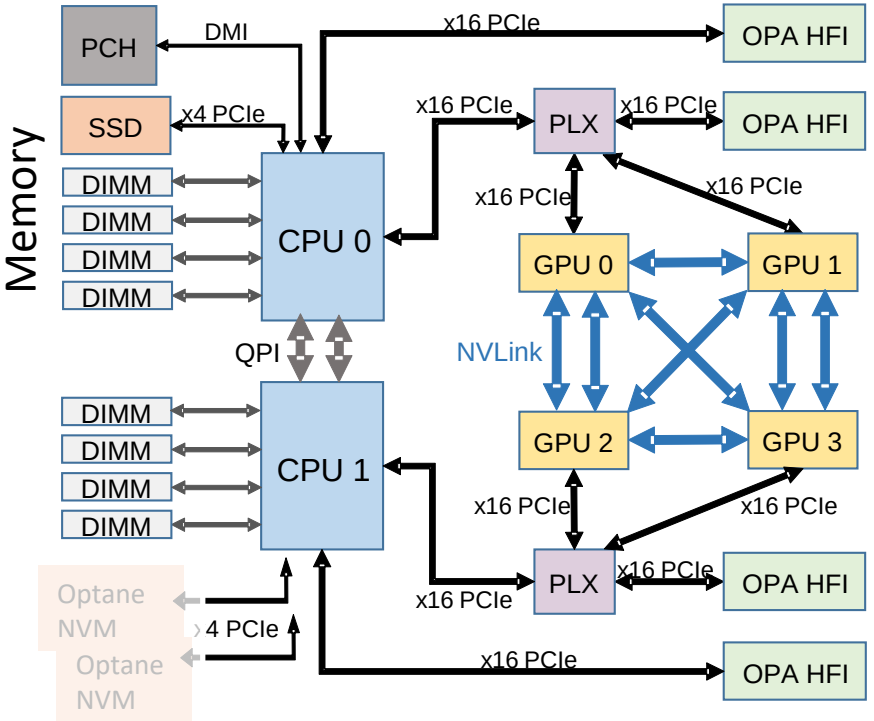




TSUBAME3.0 Compute Node SGI ICE-XA, a New GPU Compute Blade Co-Designed by SGI and Tokyo Tech GSIC

SGI ICE XA Infrastructure





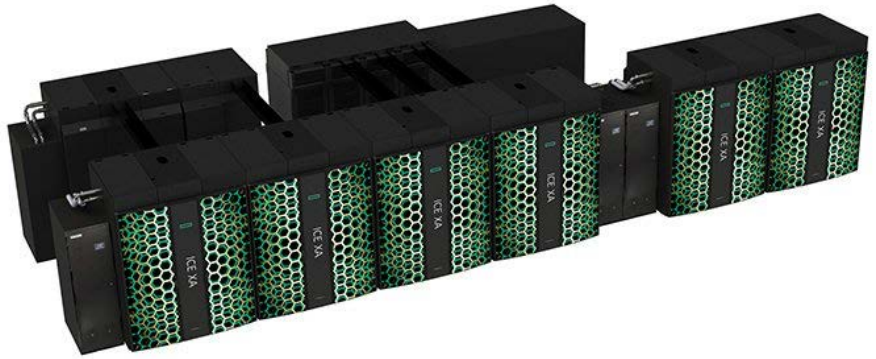
Ultra high performance & bandwidth "Fat Node"

- High Performance: 4 SXM2(NVLink) NVIDIA Pascal P100 GPU + 2 Intel Xeon 84 AI-TFLops
- High Network Bandwidth – Intel Omnipath 100GBps x 4 = 400Gbps (100Gbps per GPU)
- High I/O Bandwidth - Intel 2 TeraByte NVMe
  - > 1PB & 1.5~2TB/s system total
  - Future Octane 3D-Xpoint memory Petabyte or more directly accessible
- Ultra High Density, Hot Water Cooled Blades
  - 36 blades / rack = 144 GPU + 72 CPU, 50-60KW, x10 thermals c.f. IDC

TSUBAME 2.0/2.5/3.0 Node Performances

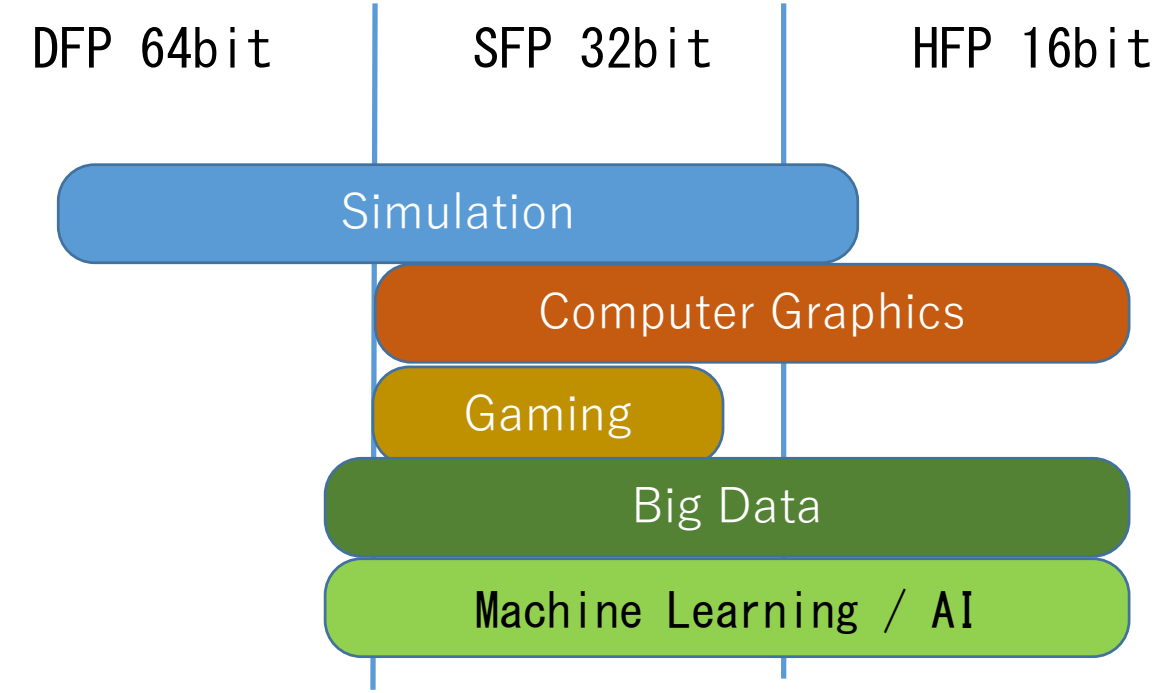
指標	TSUBAME2.0 (2010)	TSUBAME2.5 (2013)	TSUBAME3.0 (2017)	Factor
CPU Cores x Frequency (GHz)	35.16	35.16	72.8	2.07
CPU Memory Capacity (GB)	54	54	256	4.74
CPU Memory Bandwidth (GB/s)	64	64	153.6	2.40
GPU CUDA Cores	1,344	8,064	14,336	1.78
GPU FP64 (TFLOPS)	1.58	3.93	21.2	13.4 & 5.39
GPU FP32 (TFLOPS)	3.09	11.85	42.4	13.7 & 3.58
GPU FP16 (TFLOPS)	3.09	11.85	84.8	27.4 & 7.16
GPU Memory Capacity (GB)	9	18	64	7.1 & 3.56
GPU Memory Bandwidth (GB/s)	450	750	2928	6.5 & 3.90
SSD Capacity (GB)	120	120	2000	16.67
SSD READ (MB/s)	550	550	2700	4.91
SSD WRITE (MB/s)	500	500	1800	3.60
Network Injection BW (Gbps)	80	80	400	5.00

TSUBAME3.0 Datacenter

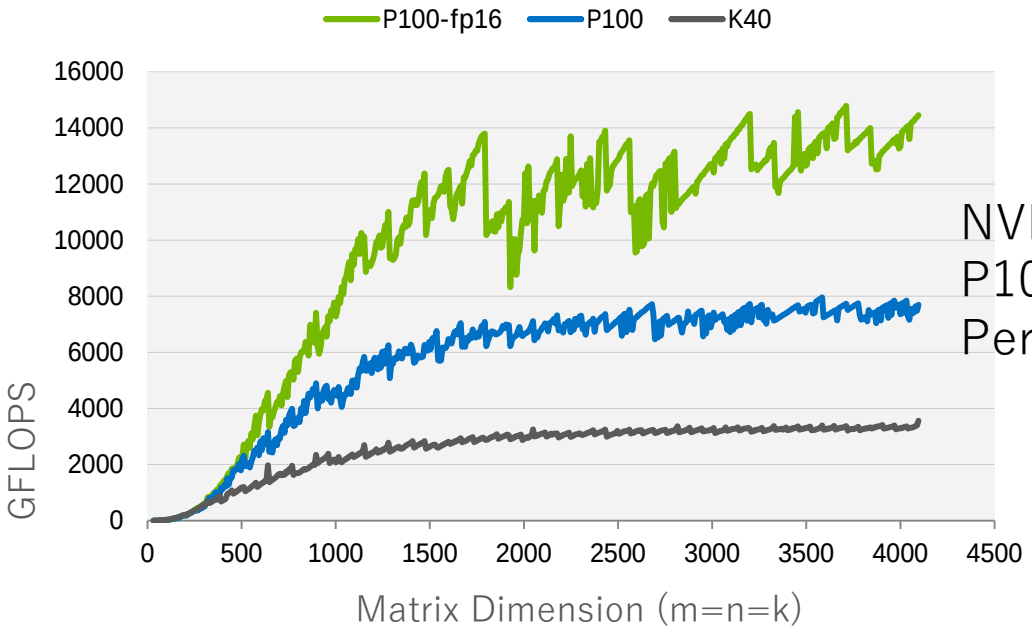


15 SGI ICE-XA Racks
2 Network Racks
3 DDN Storage Racks
20 Total Racks

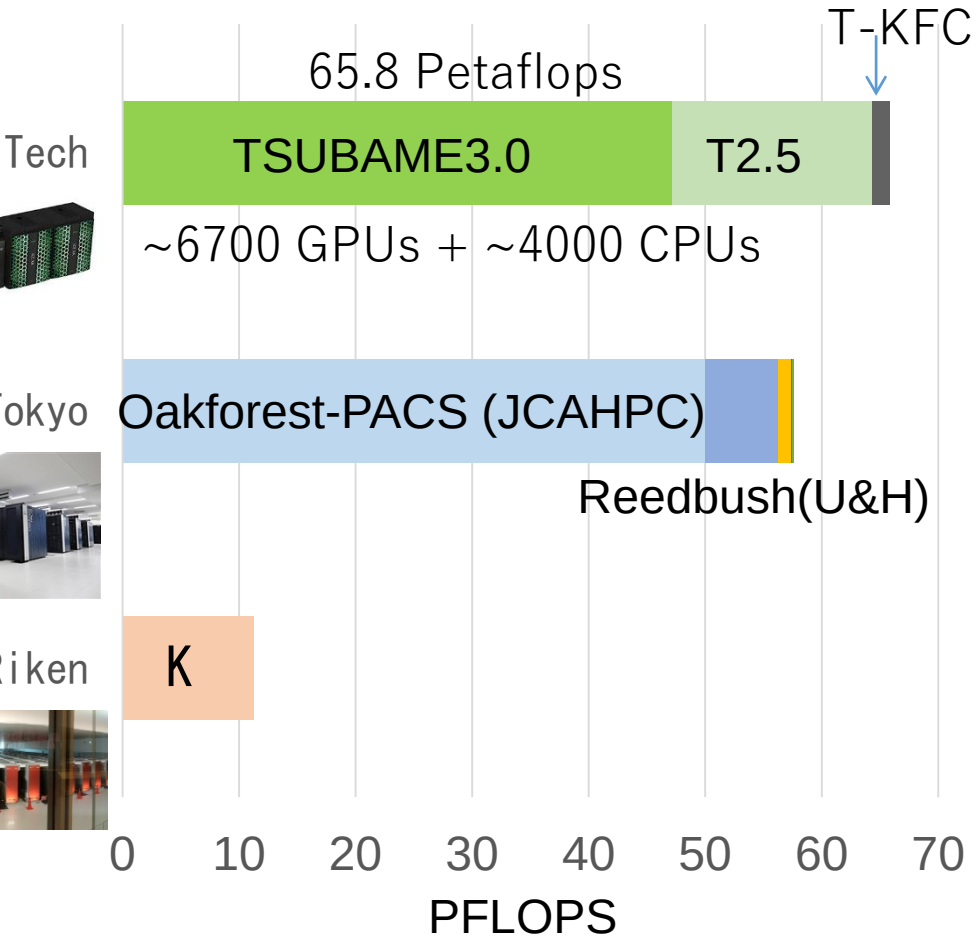
Compute racks cooled with
32 degrees warm water,
Yearound ambient cooling
Av. PUE = 1.033



Tokyo Tech GSIC leads Japan in aggregated AI-capable FLOPS TSUBAME3+2.5+KFC, in all Supercomuters and CloudsNV



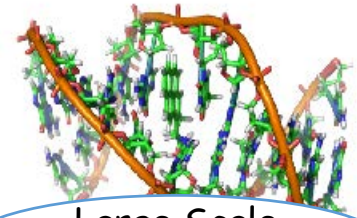
NVIDIA Pascal P100 DGEMM Performane



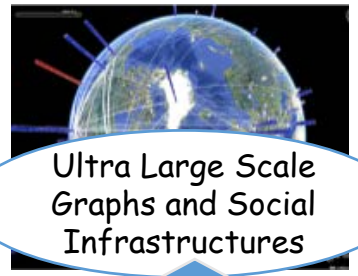
JST-CREST "Extreme Big Data" Project (2013-2018)

From FLOPS Centric to BYTES Centric HPC

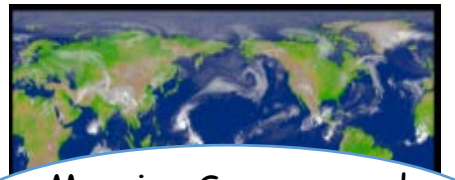
Given a top-class supercomputer, how fast can we accelerate next generation big data c.f. conventional Clouds?



Large Scale Metagenomics



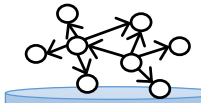
Ultra Large Scale Graphs and Social Infrastructures



Massive Sensors and Data Assimilation in Weather Prediction

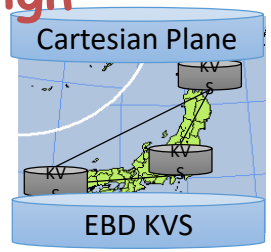
Co-Design Co-Design Co-Design

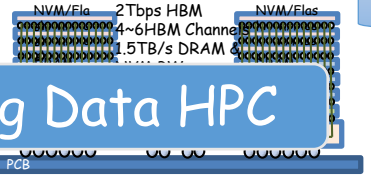
EBD Bag



Graph Store

EBD System Software incl. EBD Object System





Exascale Big Data HPC

Convergent Architecture (Phases 1~4)
Large Capacity NVM, High-Bisection NW

Issues regarding Architecture, algorithms, system software in co-design

Performance Model?
Use of accelerators e.g. GPUs?

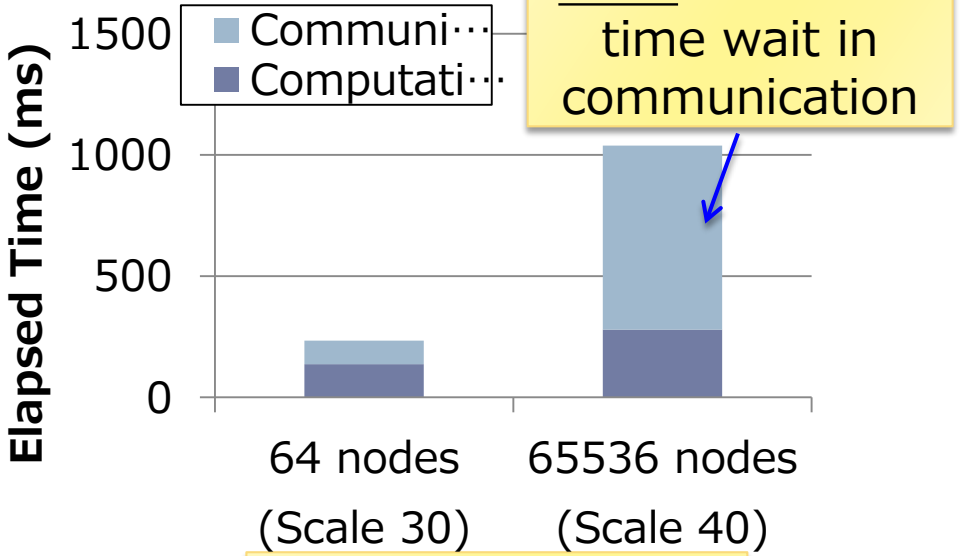
Cloud IDC
Very low BW & Efficiency
Highly available, resilient

Supercomputers
Compute&Batch-Oriented
More fragile

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

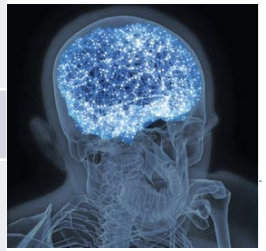
Effective x13 performance c.f. Linpack

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, Nov 2015 June Nov 2016	**1**	**38621.4**	Hybrid + Node Compression

BYTES Rich Machine + Superior BYTES algoithm

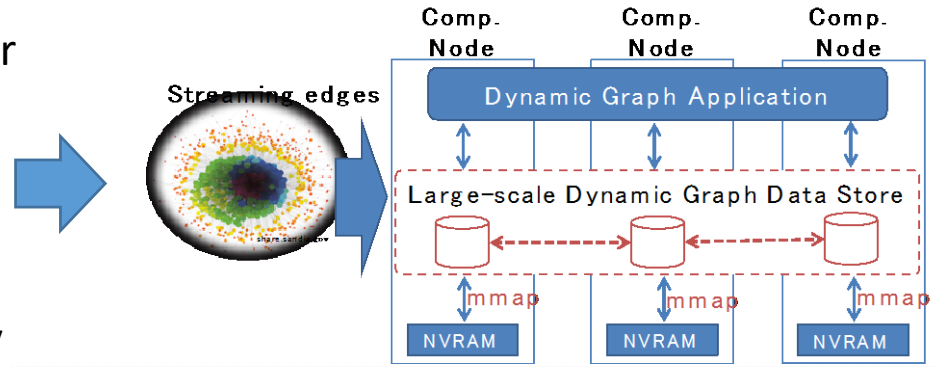


BYTES, not FLOPS!

Distributed Large-Scale Dynamic Graph Data Store (work with LLNL, [SC16 etc.])

Based on K-Computer results, adaping to (1) deep memory hierarchy, (2) rapid dynamic graph changes

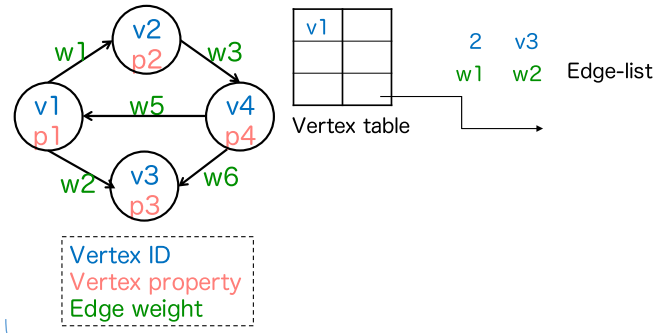
K Computer large memory but very expensive DRAM only



Node Level Dynamic Graph Data Store

Follows an adjacency-list format and leverages an open address hashing to construct its tables

Develop algorithms and SW exploiting large hierarchical memory



Extend for multi-processes using an async MPI communication framework

Dynamic Graph Construction (on-memory & NVM)

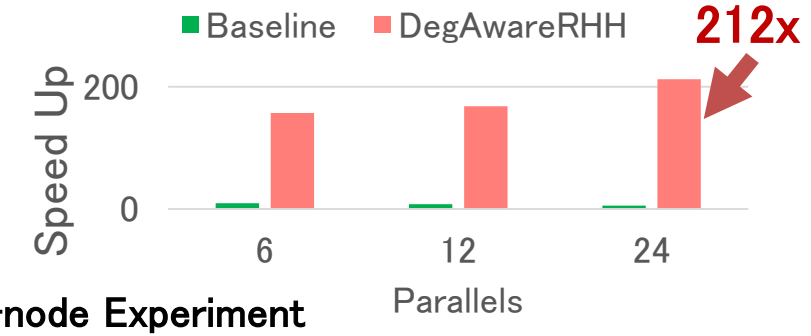
C.f. STINGER (single-node, on memory)

STINGER

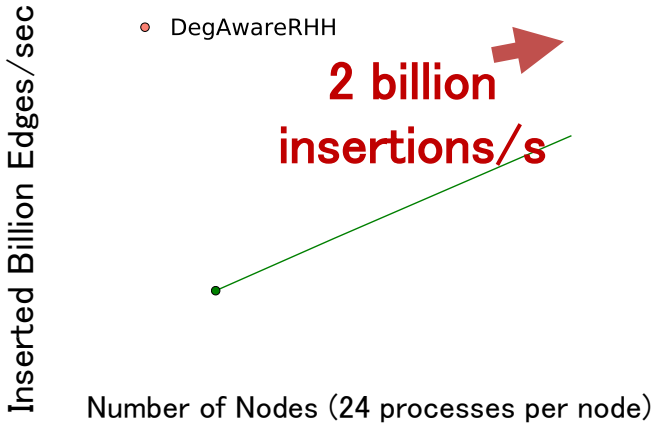
- A state-of-the-art dynamic graph processing framework developed at Georgia Tech

Baseline model

- A naïve implementation using *Boost* library (C++) and the MPI communication framework



Multi-node Experiment



Dynamic graph store w/ world's top graph update performance and scalability

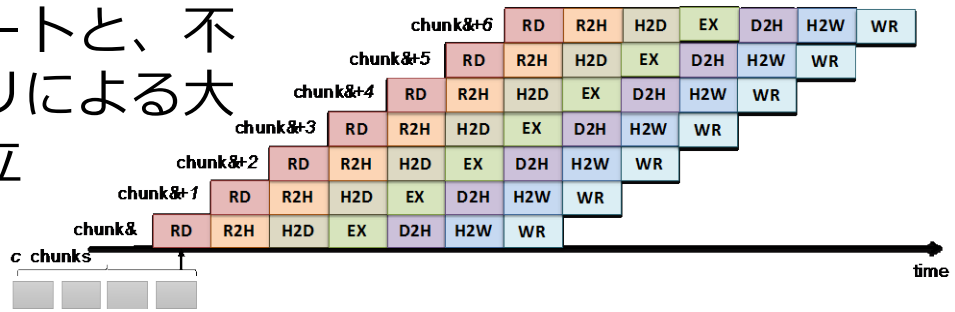
K. Iwabuchi, S. Sallinen, R. Pearce, B. V. Essen, M. Gokhale, and S. Matsuoka, Towards a distributed large-scale dynamic graph data store. In 2016 IEEE Interna- tional Parallel and Distributed Processing Symposium Workshops (IPDPSW)

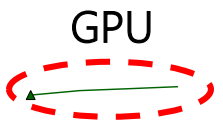
Xtr2sort: Out-of-core Sorting Acceleration using GPU and Flash NVM [IEEE BigData2016]

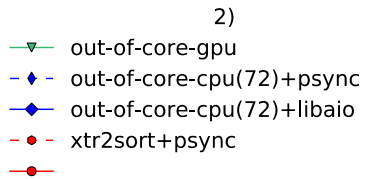
How to combine deepening memory layers for future HPC/Big Data workloads, targeting Post Moore Era?

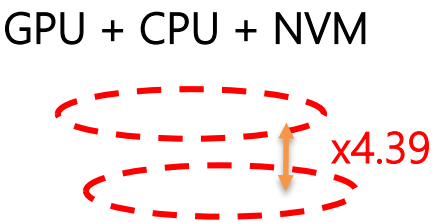
- Sample-sort-based Out-of-core Sorting Approach for Deep Memory Hierarchy Systems w/ GPU and Flash NVM
  - I/O chunking to fit device memory capacity of GPU
  - Pipeline-based Latency hiding to overlap data transfers between NVM, CPU, and GPU using asynchronous data transfers, e.g., cudaMemCpyAsync(), libaio

BYTES中心のHPCアルゴリズム：GPUのバンド幅高速ソートと、不揮発性メモリによる大容量化の両立









Estimated Compute Resource Requirements for Deep Learning [Source: Preferred Network Japan Inc.]

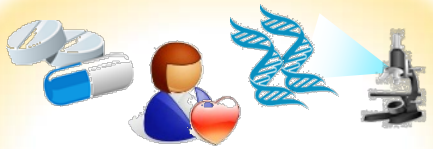
To complete the learning phase in one day

P:Peta
E:Exa
F:Flops

Image/Video Recognition

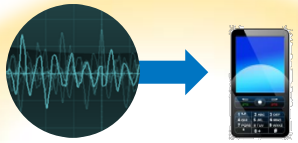
10P（Image）～ 10E（Video） Flops
学習データ：1億枚の画像 10000クラス分類
数千ノードで6ヶ月 [Google 2015]

Bio / Healthcare



100P ～ 1E Flops
一人あたりゲノム解析で約10M個のSNPs
100万人で100PFlops、1億人で1EFlops

Image Recognition

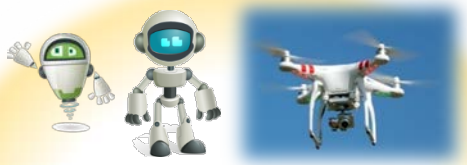


10P～ Flops
1万人の5000時間分の音声データ
人工的に生成された10万時間の
音声データを基に学習 [Baidu 2015]

Auto Driving

1E～100E Flops
自動運転車１台あたり1日 1TB
10台～1000台, 100日分の走行データの学習

Robots / Drones



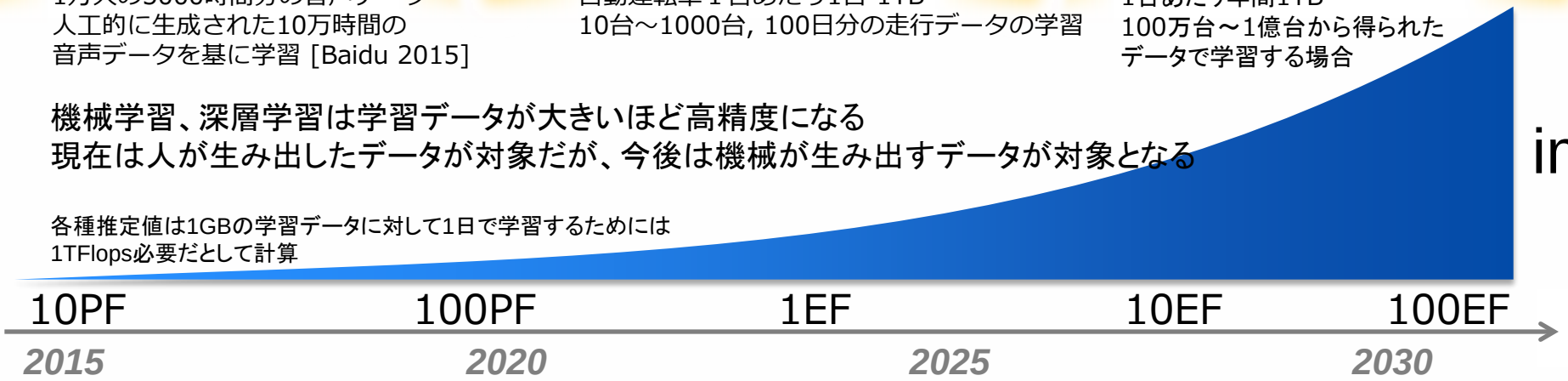
1E～100E Flops
1台あたり年間1TB
100万台～1億台から得られた
データで学習する場合

It’s the FLOPS (in reduced precision) and BW!

So both are important in the infrastructure

機械学習、深層学習は学習データが大きいほど高精度になる
現在は人が生み出したデータが対象だが、今後は機械が生み出すデータが対象となる

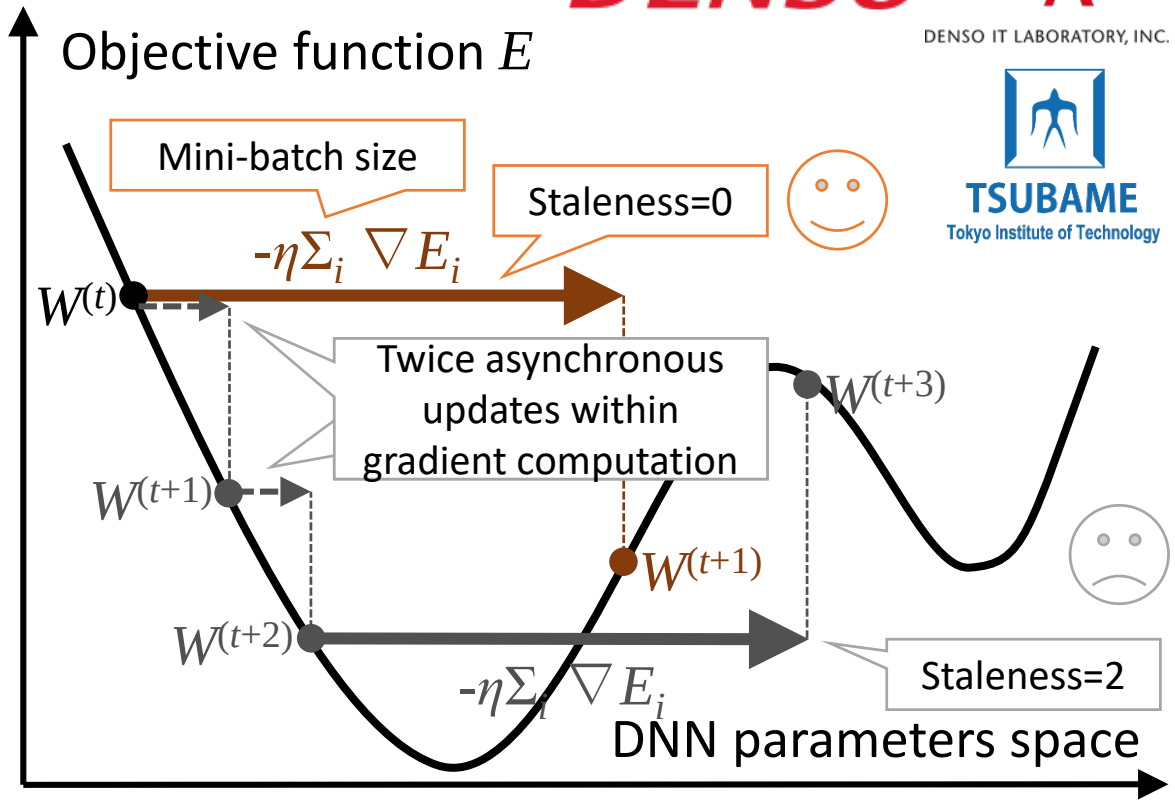
各種推定値は1GBの学習データに対して1日で学習するためには
1TFlops必要だとして計算



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

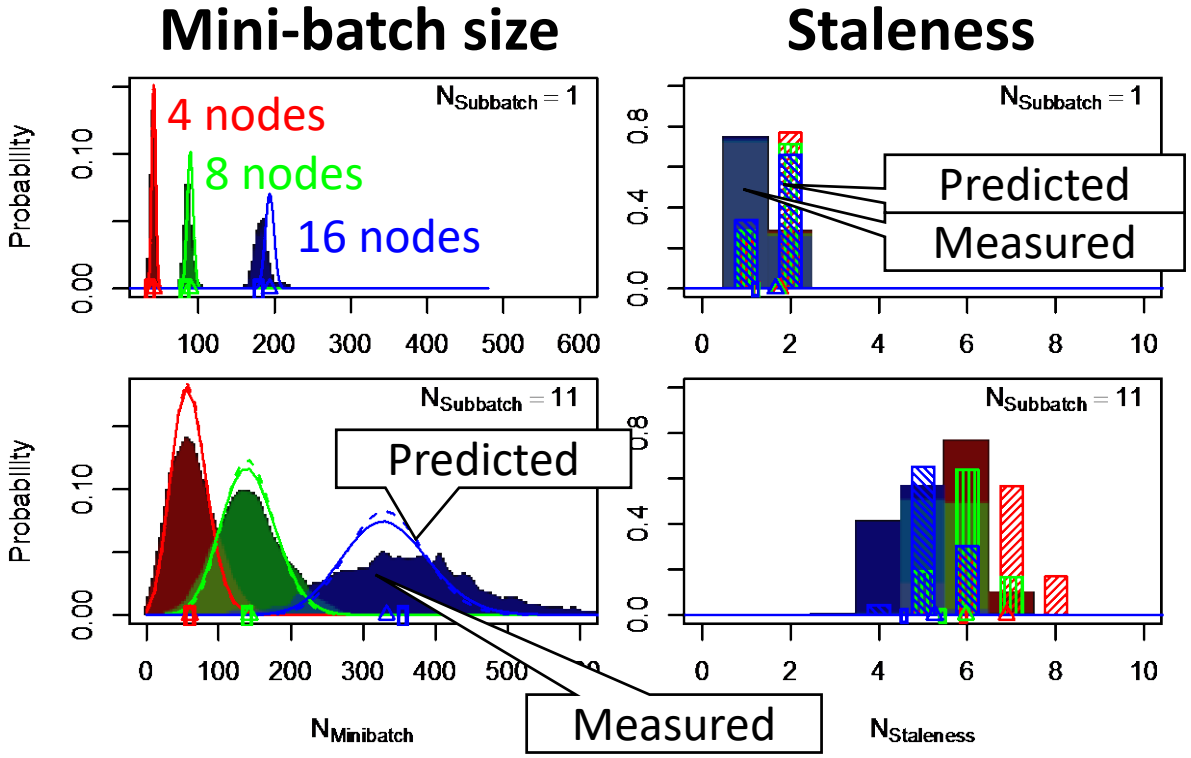
Background

- 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

Performance Prediction of Future HW for CNN

- Predicts the best performance with two future architectural extensions
  - FP16: precision reduction to double the peak floating point performance
  - EDR IB: 4xEDR InfiniBand (100Gbps) upgrade from FDR (56Gbps)

→ Not only # of nodes, but also fast interconnect is important for scalability

TSUBAME-KFC/DL ILSVRC2012 dataset deep learning
Prediction of best parameters (average minibatch size 138±25%)

	N_Node	N_Subbatch	Epoch Time	Average Minibatch Size
(Current HW)	8	8	1779	165.1
FP16	7	22	1462	170.1
EDR IB	12	11	1245	166.6
FP16 + EDR IB	8	15	1128	171.5

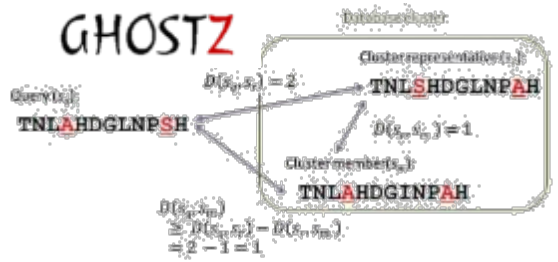
Open Source Release of EBD System Software (install on T3/Amazon/ABCI)

- mrCUDA
  - rCUDA extension enabling remote-to-local GPU migration
  - https://github.com/EBD-CREST/mrCUDA
  - GPU 3.0
  - Co-Funded by NVIDIA
- Huron FS (w/LLNL)
  - I/O Burst Buffer for Inter Cloud Environment
  - https://github.com/EBD-CREST/cbb
  - Apache License 2.0
  - Co-funded by Amazon
- ScaleGraph Python
  - Python Extension for ScaleGraph X10-based Distributed Graph Library
  - https://github.com/EBD-CREST/scalegraphpython
  - Eclipse Public License v1.0
- GPUSort
  - GPU-based Large-scale Sort
  - https://github.com/EBD-CREST/gpusort
  - MIT License
- Others, including dynamic graph store

HPC and BD/AI Convergence Example [Yutaka Akiyama, Tokyo Tech]

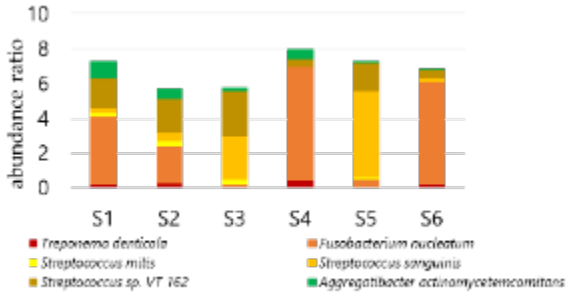
Genomics

Ultra-fast Seq. Analysis



- Suzuki *et al.*, *Bioinformatics* (2015)
- Suzuki *et al.*, *PLOS ONE* (2016)

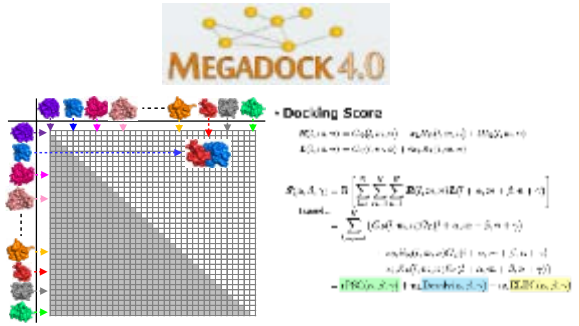
Oral/Gut Metagenomics



- Yamasawa *et al.*, *IIBMP* (2016)

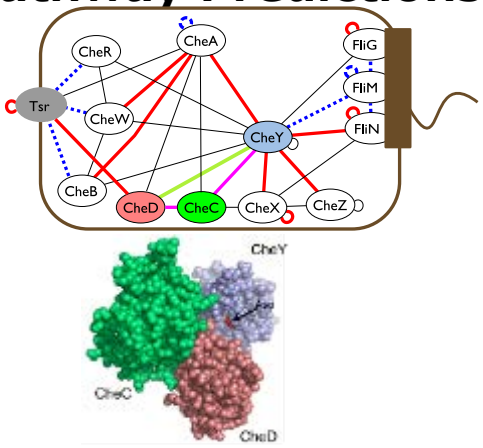
Protein-Protein Interactions

Exhaustive PPI Prediction System



- Ohue *et al.*, *Bioinformatics* (2014)

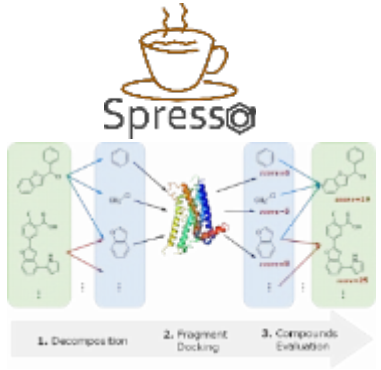
Pathway Predictions



- Matsuzaki *et al.*, *Protein Pept Lett* (2014)

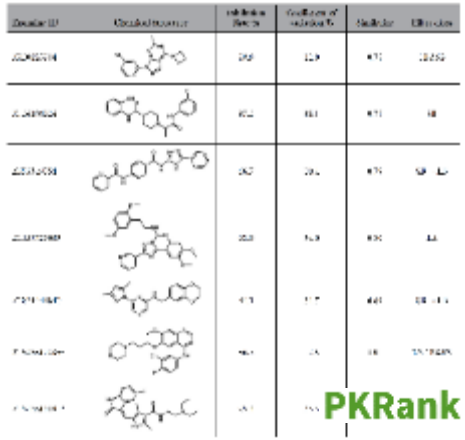
Drug Discovery

Fragment-based Virtual Screening



- Yanagisawa *et al.*, *GIW* (2016)

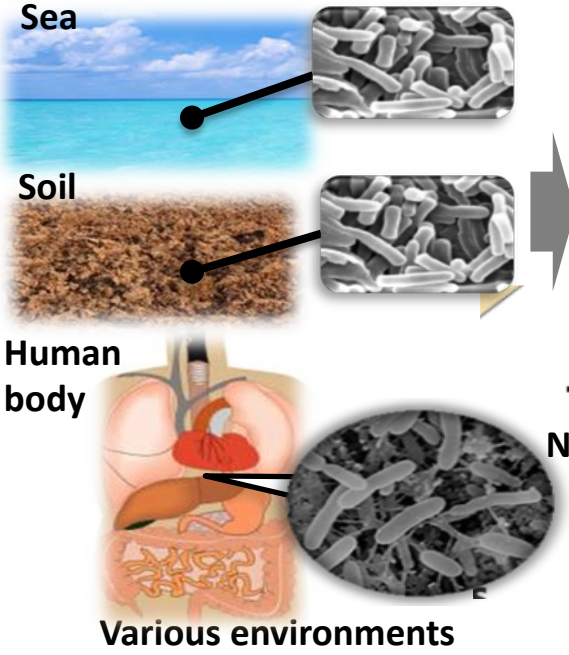
Learning-to-Rank VS



- Suzuki *et al.*, *AROB2017* (2017)

EBD vs. EBD : Large Scale Homology Search for Metagenomics

- Revealing uncultured microbiomes and finding novel genes in various environments
- Applied for human health in recent years



...CCTTATCTTCG...
...CCACATAAACT...
...ATGGTCGATGTT...

Next generation sequencer

increasing

EBD

O(n)

Meas. data

increasing

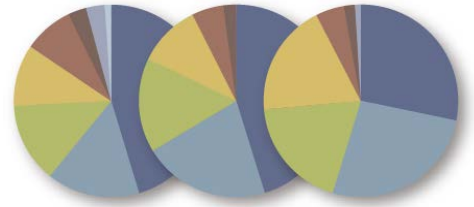
O(m) Reference Database

EBD

O(m n) calculation

Correlation, Similarity search

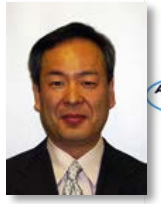
Metabolic Pathway



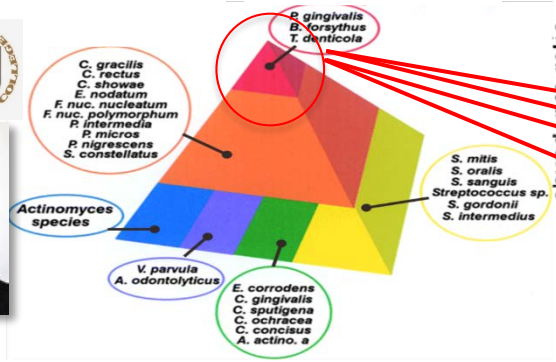
Taxonomic composition

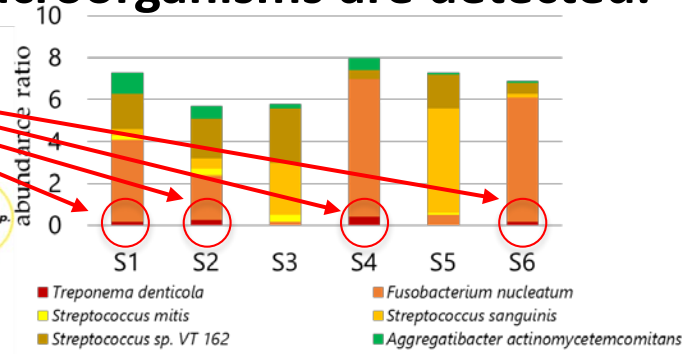
Metagenomic analysis of periodontitis patients

- with Tokyo Dental College, Prof. Kazuyuki Ishihara
- Comparative metagenomic analysis bewtween healthy persons and patients



High risk microorganisms are detected.



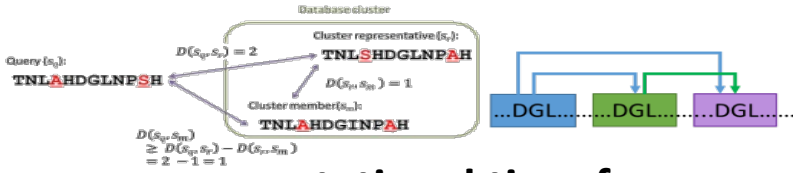


Development of Ultra-fast Homology Search Tools

GHOSTZ

Suzuki, *et al*. *Bioinformatics*, 2015.

Subsequence sequence clustering



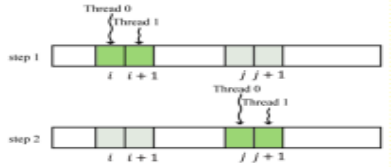
computational time for 10,000 sequences (sec.) (3.9 GB DB、1CPU core)

×240 faster than conventional algorithm

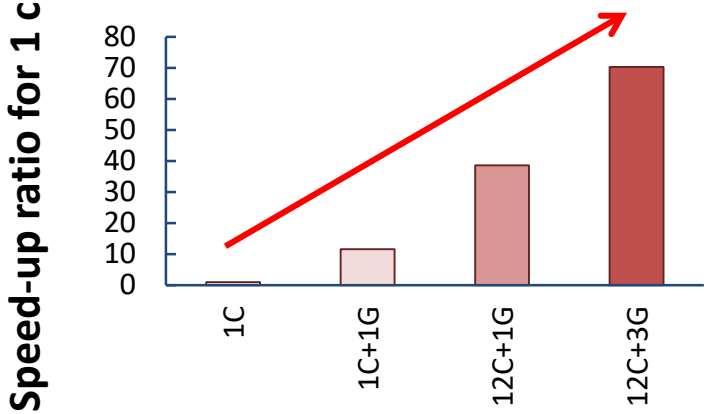
GHOSTZ-GPU

Suzuki, *et al*. *PLOS ONE*, 2016.

Multithread on GPU



TSUBAME 2.5 Thin node GPU



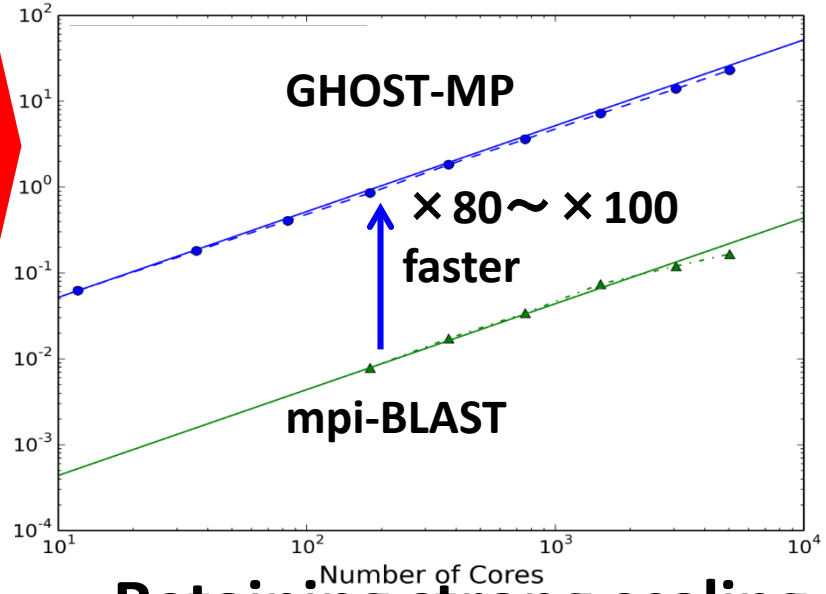
×70 faster than 1 core using 12 cores + 3 GPUs

GHOST-MP

Kakuta, *et al*. (submitted)

MPI + OpenMP hybrid pallelization

TSUBAME 2.5



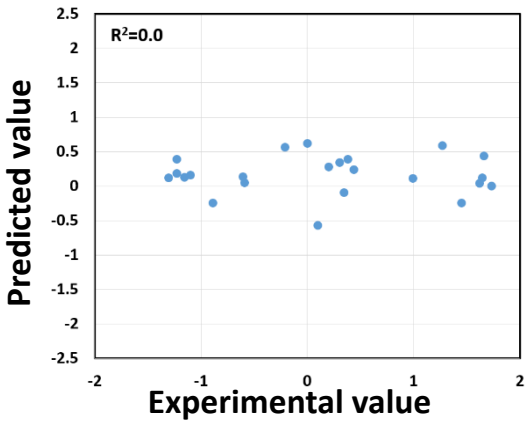
Retaining strong scaling up to 100,000 cores

Plasma Protein Binding (PPB) Prediction by Machine Learning
Application for peptide drug discovery

Problems

	Small molecule drug	Peptide drug	Antibody drug
Molecular weight	~1,000	600~2,500	150,000~
Number of targets	◎	○	△
Target specificity	△	○	◎
PPI inhibition	×	○	○
Bio-stability	○	△	◎

- Candidate peptides are tend to be degraded and excreted faster than small molecule drugs
- Strong needs to design bio-stable peptides for drug candidates



Previous PPB prediction software for small molecule can not predict peptide PPB

Solutions

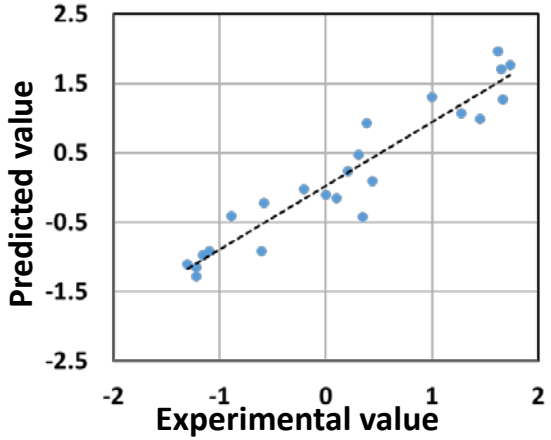
Compute Feature Values
(more than 500 features)

LogS
LogP
:
MolWeight
:
SASA
polarity

f

PPB value

Combining feature values for building a predictive model



$R^2 = 0.905$

A constructed model can explain peptide PPB well

RWBC-OIL 2-3: Tokyo Tech IT-Drug Discovery Factory Simulation & Big Data & AI at Top HPC Scale

(Tonomachi, Kawasaki-city: planned 2017, PI Yutaka Akiyama)

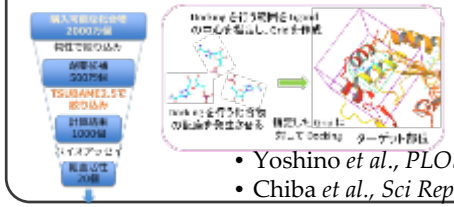
Tokyo Tech's research seeds

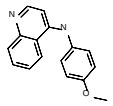
①Drug Target selection system

Minister of Health, Labour and Welfare Award of the 11th annual Merit Awards for Industry-Academia-Government Collaboration

②Glide-based Virtual Screening

TSUBAME's GPU-environment allows **World's top-tier Virtual Screening**

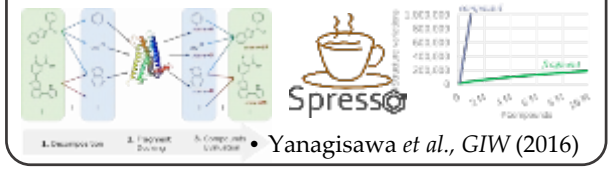




- Yoshino *et al.*, *PLOS ONE* (2015)
- Chiba *et al.*, *Sci Rep* (2015)

③Novel Algorithms for fast virtual screening against huge databases

Fragment-based efficient algorithm designed for **100-millions cmpds data**

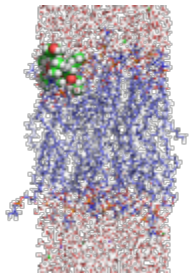


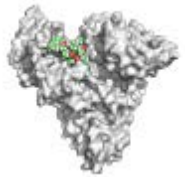
- Yanagisawa *et al.*, *GIW* (2016)

Drug Discovery platform powered by Supercomputing and Machine Learning

Application projects

New Drug Discovery platform especially for specialty peptide and nucl. acids.





Plasma binding (ML-based)

Membrane penetration (Mol. Dynamics simulation)

PeptiDream

Catalyst Inc. (株)カタリスト

DiscoveRessource ディスカヴァリソース(株)

SCHRÖDINGER シュレーディンガー(株)

(株)情報数理バイオ

KING SKYFRONT
Kawasaki Innovation Gateway at SKYFRONT

Multi-Petaflops Compute
Peta~Exabytes Data Processing Continuously

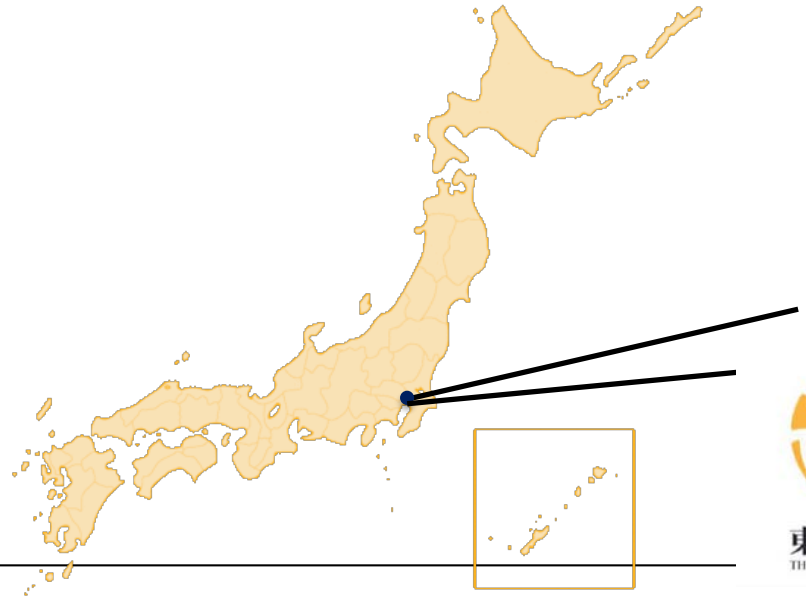
Cutting Edge, Large-Scale HPC & BD/AI Infrastructure Absolutely Necessary

Investments from JP Govt., Tokyo Tech. (TSUBAME SC) Muninciple Govt (Kawasaki), JP & US Pharma

METI AIST-AIRC ABCI

as the *worlds first large-scale OPEN AI Infrastructure*

- **ABCI**: AI Bridging Cloud Infrastructure
  - Top-Level SC compute & data capability for DNN (130~200 AI-Petaflops)
  - Open Public & Dedicated infrastructure for AI & Big Data Algorithms, Software and Applications
  - Platform to accelerate joint academic-industry R&D for AI in Japan



- 130~200 AI-Petaflops
- < 3MW Power
- < 1.1 Avg. PUE
- Operational 2017Q4 ~2018Q1

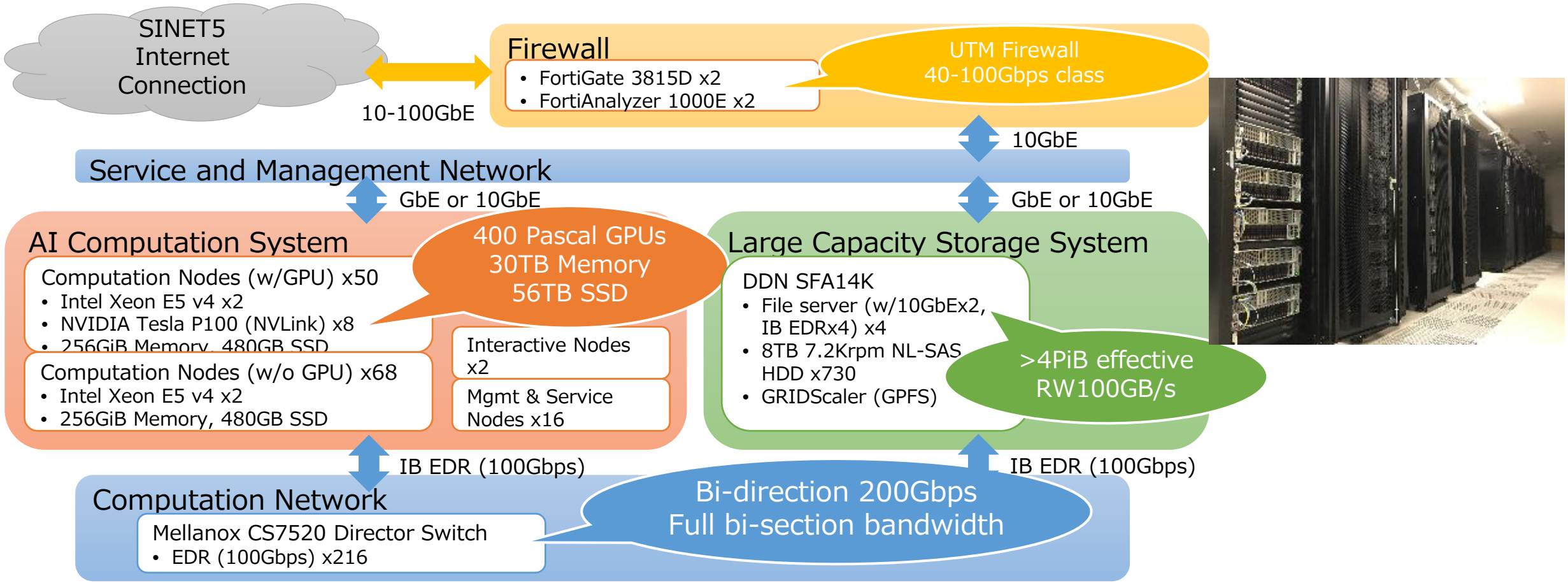
東京大学 THE UNIVERSITY OF TOKYO

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

Univ. Tokyo Kashiwa Campus

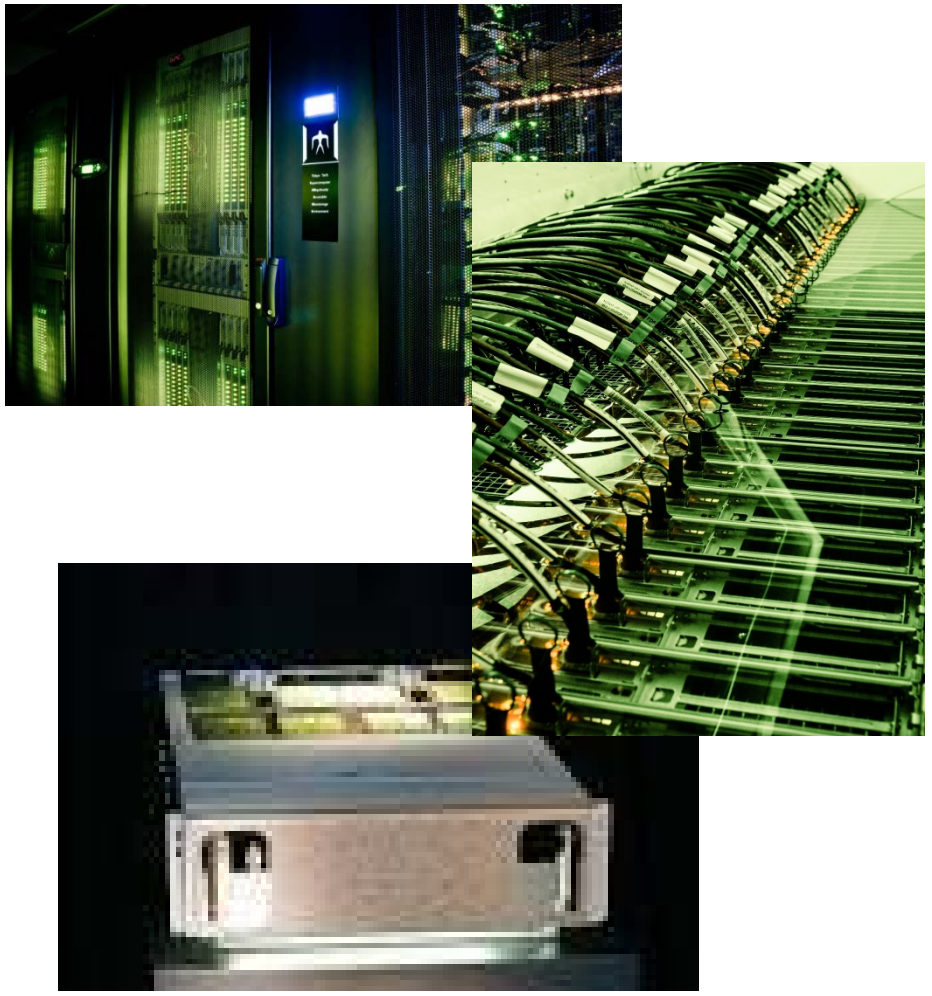
ABCI Prototype: AIST AI Cloud (AAIC) March 2017 (System Vendor: NEC)

- **400x NVIDIA Tesla P100s and Infiniband EDR** accelerate various AI workloads including ML (Machine Learning) and DL (Deep Learning).
- Advanced data analytics leveraged by **4PiB shared Big Data Storage and Apache Spark** w/ its ecosystem.



The "Real" ABCI – 2018Q1

- **Extreme computing power**
  - w/ **130~200 AI-PFlops** for AI/ML especially DNN
  - **x1 million speedup** over high-end PC: 1 Day training for 3000-Year DNN training job
  - TSUBAME-KFC (1.4 AI-Pflops) x 90 users (T2 avg)
- **Big Data and HPC converged modern design**
  - For advanced data analytics (Big Data) and scientific simulation (HPC), etc.
  - Leverage Tokyo Tech's "TSUBAME3" design, **but differences/enhancements being AI/BD centric**
- **Ultra high bandwidth and low latency in memory, network, and storage**
  - For accelerating various AI/BD workloads
  - Data-centric architecture, optimizes data movement
- **Big Data/AI and HPC SW Stack Convergence**
  - **Incl. results from JST-CREST EBD**
  - **Wide contributions from the PC Cluster community desirable.**



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 SPECCPU

- AI/ML Benchmarks
  - Low precision GEMM
    - CNN Kernel, defines “AI-Flops”
  - Single Node CNN
    - AlexNet and GoogLeNet
    - ILSVRC2012 Dataset
  - Multi-Node CNN
    - Caffe+MPI
  - Large Memory CNN
    - Convnet on Chainer
  - RNN / LSTM
    - To be determined

Cutting Edge Research AI Infrastructures in Japan Accelerating BD/AI with HPC

(and my effort to design & build them)

Tokyo Tech

1H 2019?
"**ExaAI**"
~1 AI-ExaFlop
Undergoing Engineering Study

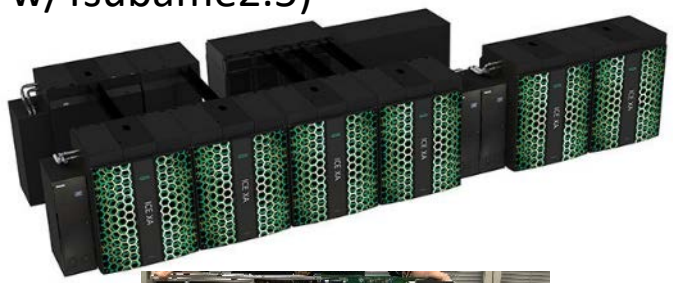
x5.0~7.7

Mar. 2018
ABCI (AIST-AIRC)
130-200 AI-PF

Draft RFC out
IDC under construction

x2.8~4.2

Aug. 2017
TSUBAME3.0 (Tokyo Tech./HPE)
47.2 AI-PF (65.8 AI-PF w/Tsubame2.5)



Being Manufactured

x5.8

Mar. 2017
AIST AI Cloud
(AIST-AIRC/NEC)
8.2 AI-PF

Under Acceptance

x5.8

In Production

Oct. 2015
TSUBAME-KFC/DL
(Tokyo Tech./NEC)
1.4 AI-PF(Petaflops)

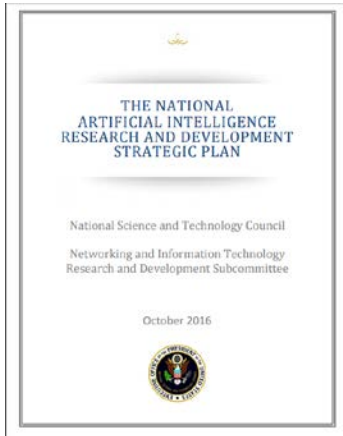
~x1000 in 3.5 years

Built/funded

still under plans

Mar. 2017
AI Supercomputer
Riken AIP/Fujitsu
4.1 AI-PF

R&D Investments into world leading AI/BD HW & SW & Algorithms and their co-design for cutting edge Infrastructure absolutely necessary (just as is with Japan Post-K and US ECP in HPC)

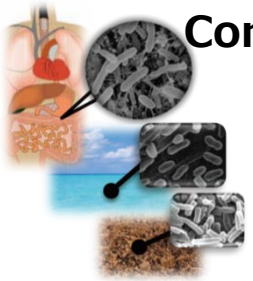


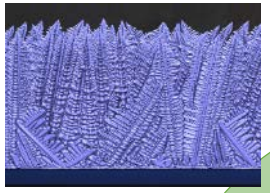
Backups

Co-Design of BD/ML/AI with HPC using BD/ML/AI - for survival of HPC

Acceleration and Scaling of BD/ML/AI via HPC and Technologies and Infrastructures

Accelerating Conventional HPC Apps





Optimizing System Software and Ops

Future Big Data・AI Supercomputer Design

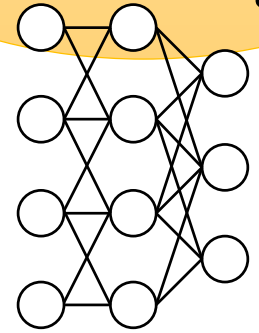
Big Data AI-Oriented Supercomputers

ABCI: World's first and largest open 100 Peta AI-Flops AI Supercomputer, Fall 2017, for co-design

Mutual and Semi-Automated Co-Acceleration of HPC and BD/ML/AI

Acceleration Scaling, and Control of HPC via BD/ML/AI and future SC designs

Big Data and ML/AI Apps and Methodologies



Large Scale Graphs

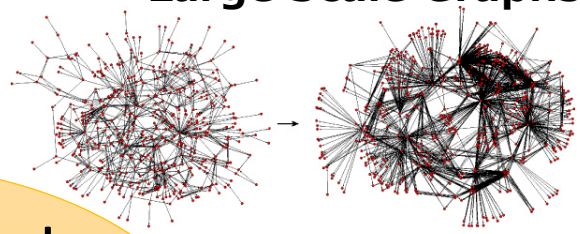
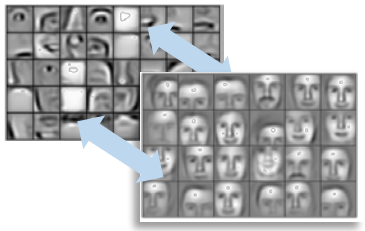
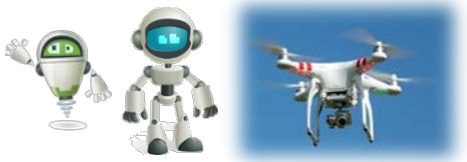


Image and Video

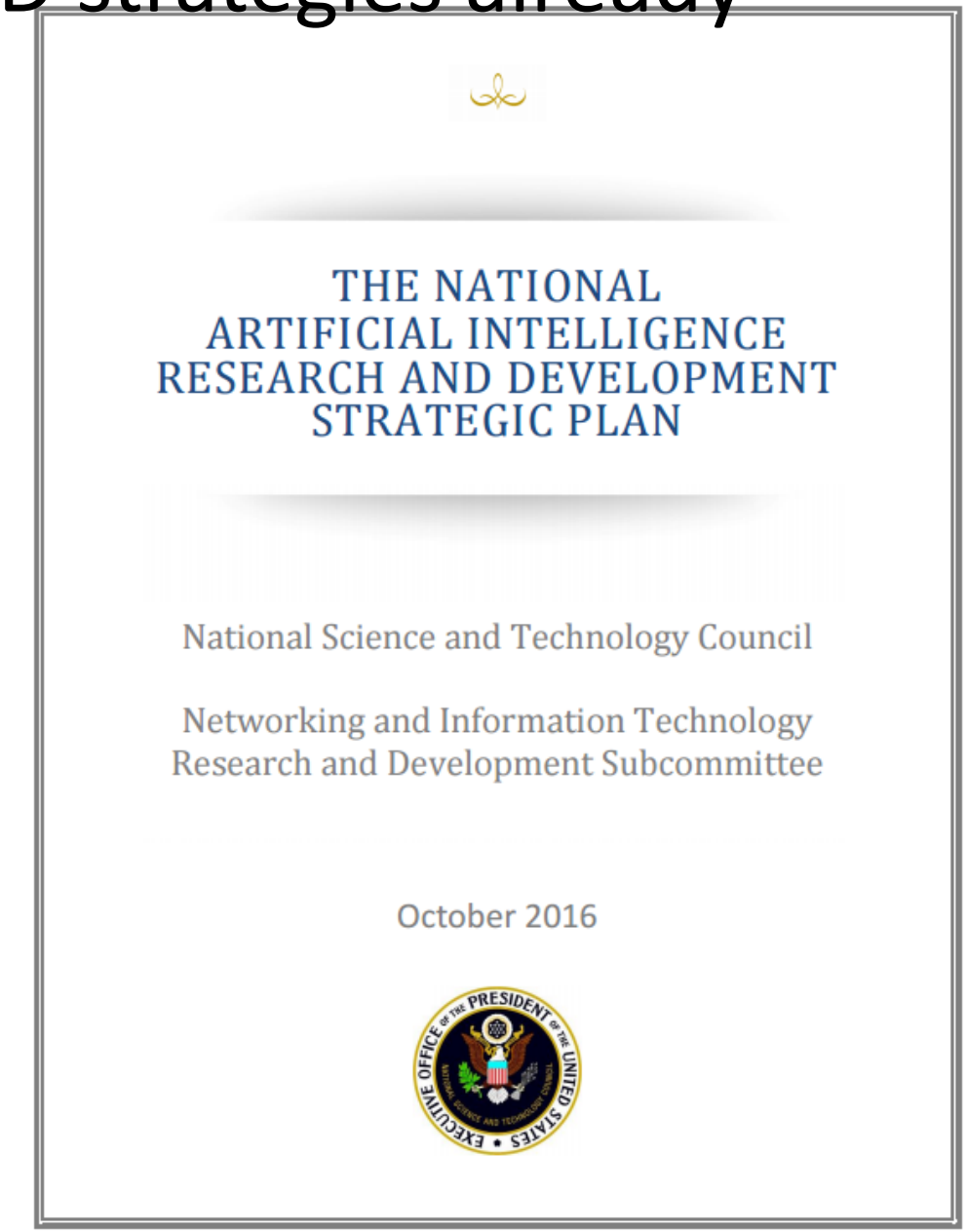


Robots / Drones

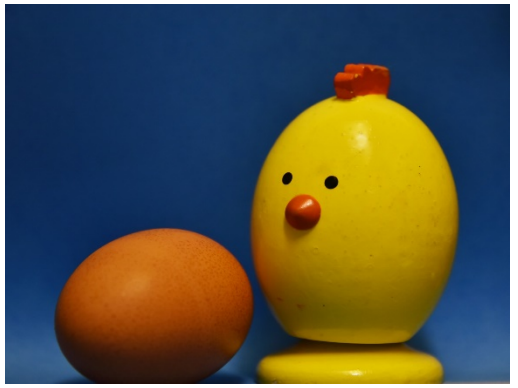


We are implementing the US AI&BD strategies already …in Japan, at AIRC w/ABCI

- Strategy 5: Develop shared public datasets and environments for AI training and testing. The depth, quality, and accuracy of training datasets and resources significantly affect AI performance. Researchers need to develop high quality datasets and environments and enable responsible access to high-quality datasets as well as to testing and training resources.
- Strategy 6: Measure and evaluate AI technologies through standards and benchmarks. Essential to advancements in AI are standards, benchmarks, testbeds, and community engagement that guide and evaluate progress in AI. Additional research is needed to develop a broad spectrum of evaluative techniques.



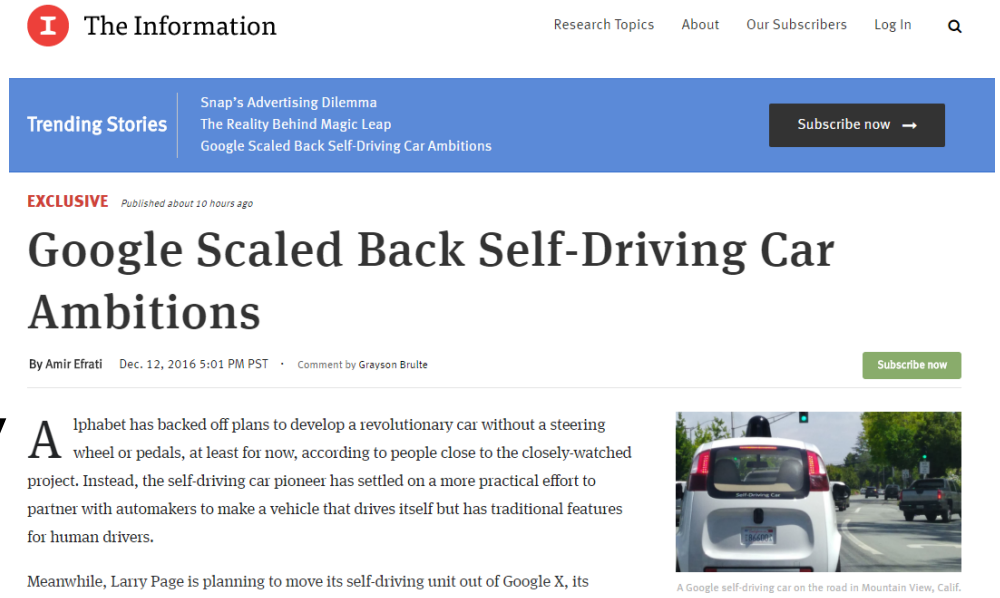
The “Chicken or Egg Problem” of AI-HPC Infrastructures



- “On Premise” machines in clients => “Can’t invest in big in AI machines unless we forecast good ROI. We don’t have the experience in running on big machines.”
- Public Clouds other than the giants => “Can’t invest big in AI machines unless we forecast good ROI. We are cutthroat.”
- Large scale supercomputer centers => “Can’t invest big in AI machines unless we forecast good ROI. Can’t sacrifice our existing clients and our machines are full”
- Thus the giants dominate, AI technologies, big data, and people stay behind the corporate firewalls…

But Commercial Companies esp. the "AI Giants"are Leading AI R&D, are they not?

- Yes, but that is because their shot–term goals could harvest the low hanging fruits in DNN rejuvenated AI
- But AI/BD research is just beginning––– if we leave it to the interests of commercial companies, we cannot tackle difficult problems with no proven ROI
  - Very unhealthy for research
- This is different from more mature fields, such as pharmaceuticals or aerospace, where there is balanced investments and innovations in both academia/government and the industry



The Information

Research Topics About Our Subscribers Log In

Trending Stories

Snap's Advertising Dilemma
The Reality Behind Magic Leap
Google Scaled Back Self-Driving Car Ambitions

Subscribe now →

EXCLUSIVE *Published about 10 hours ago*

Google Scaled Back Self-Driving Car Ambitions

By Amir Efrati Dec. 12, 2016 5:01 PM PST · Comment by Grayson Brulte

Subscribe now

Alphabet has backed off plans to develop a revolutionary car without a steering wheel or pedals, at least for now, according to people close to the closely-watched project. Instead, the self-driving car pioneer has settled on a more practical effort to partner with automakers to make a vehicle that drives itself but has traditional features for human drivers.

Meanwhile, Larry Page is planning to move its self-driving unit out of Google X, its

A Google self-driving car on the road in Mountain View, Calif.

Japanese Open Supercomputing Sites Aug. 2017 (pink=HPCI Sites)

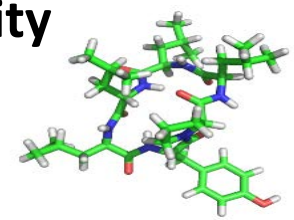
Peak Rank	Institution	System	Double FP Rpeak	Nov. 2016 Top500
1	U-Tokyo/Tsukuba U JCAHP	Oakforest-PACS - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path	24.9	6
2	Tokyo Institute of Technology GSIC	TSUBAME 3.0 - HPE/SGI ICE-XA custom NVIDIA Pascal P100 + Intel Xeon, Intel OmniPath	12.1	NA
3	Riken AICS	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	11.3	7
4	Tokyo Institute of Technology GSIC	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.93GHz, Infiniband QDR, NVIDIA K20x NEC/HPE	5.71	40
5	Kyoto University	Camphor 2 – Cray XC40 Intel Xeon Phi 68C 1.4Ghz	5.48	33
6	Japan Aerospace eXploration Agency	SORA-MA - Fujitsu PRIMEHPC FX100, SPARC64 XIfx 32C 1.98GHz, Tofu interconnect 2	3.48	30
7	Information Tech. Center, Nagoya U	Fujitsu PRIMEHPC FX100, SPARC64 XIfx 32C 2.2GHz, Tofu interconnect 2	3.24	35
8	National Inst. for Fusion Science(NIFS)	Plasma Simulator - Fujitsu PRIMEHPC FX100, SPARC64 XIfx 32C 1.98GHz, Tofu interconnect 2	2.62	48
9	Japan Atomic Energy Agency (JAEA)	SGI ICE X, Xeon E5-2680v3 12C 2.5GHz, Infiniband FDR	2.41	54
10	AIST AI Research Center (AIRC)	AAIC (AIST AI Cloud) – NEC/SMC Cluster, NVIDIA Pascal P100 + Intel Xeon, Infiniband EDR	2.2	NA

Molecular Dynamics Simulation for Membrane Permeability
Application for peptide drug discovery

Problems

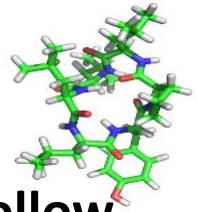
1) Single residue mutation can drastically change membrane permeability

Sequence : D-Pro, D-Leu, D-Leu, **L-Leu**, D-Leu,
Membrane permeability : **7.9** × 10^{-6} cm/s

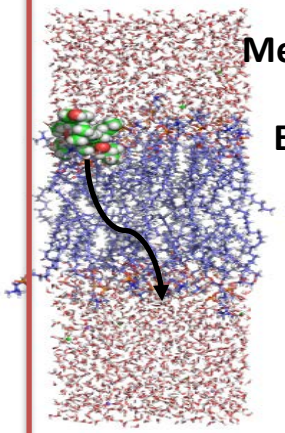


× 0.006

Sequence : D-Pro, D-Leu, D-Leu, **D-Leu**, D-Leu, L-Tyr
Membrane permeability : **0.045** × 10^{-6} cm/s



2) Standard MD simulation can not follow membrane permeation.



Membrane permeation is **millisecond** order phenomenon.

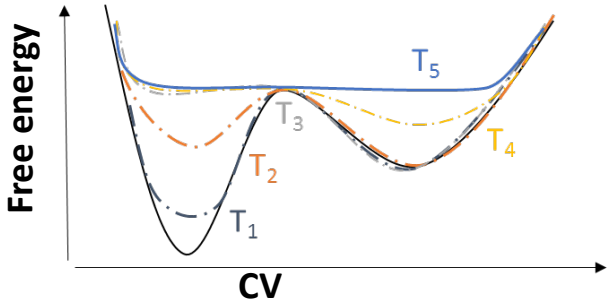
Ex) Membrane thickness : 40 Å
Peptide membrane permeability : 7.9 × 10^{-6} cm/s

Typical peptide membrane permeation takes
40 Å / 7.9 × 10^{-6} cm/s = 0.5 **millisecond**

Solutions

1) Apply enhanced sampling

Metadynamics (MTD)



Supervised MD (SuMD)

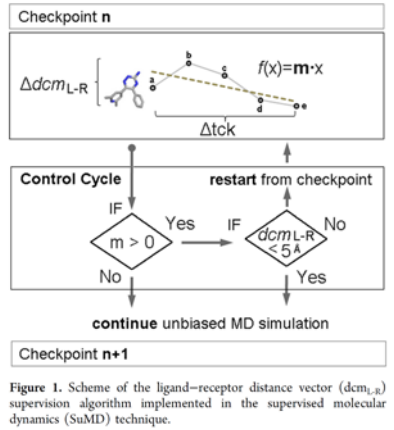
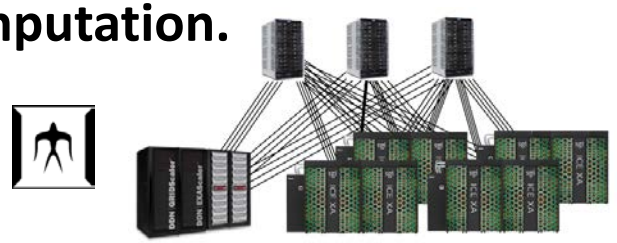


Figure 1. Scheme of the ligand–receptor distance vector (dcm_{L-R}) supervision algorithm implemented in the supervised molecular dynamics (SuMD) technique.

2) GPU acceleration and massively parallel computation.



MD engine

GROMACS
DESMOND
} on GPU

- **Millisecond order phenomenon can be simulated.**
- **Hundreds of peptides can be calculated simultaneously on TSUBAME.**

ABCI Cloud Infrastructure

- **Ultra-dense IDC design from ground-up**
  - Custom inexpensive lightweight “warehouse” building w/ substantial earthquake tolerance
  - **x20 thermal density of standard IDC**
- **Extreme green**
  - Ambient warm liquid cooling, large Li-ion battery storage, and high-efficiency power supplies, etc.
  - **Commoditizing supercomputer cooling technologies to Clouds (60KW/rack)**
- **Cloud ecosystem**
  - Wide-ranging Big Data and HPC standard software stacks
- **Advanced cloud-based operation**
  - Incl. dynamic deployment, container-based virtualized provisioning, multitenant partitioning, and automatic failure recovery, etc.
  - Joining HPC and Cloud Software stack for real
- **Final piece in the commoditization of HPC (into IDC)**

ABCI AI-IDC CG Image

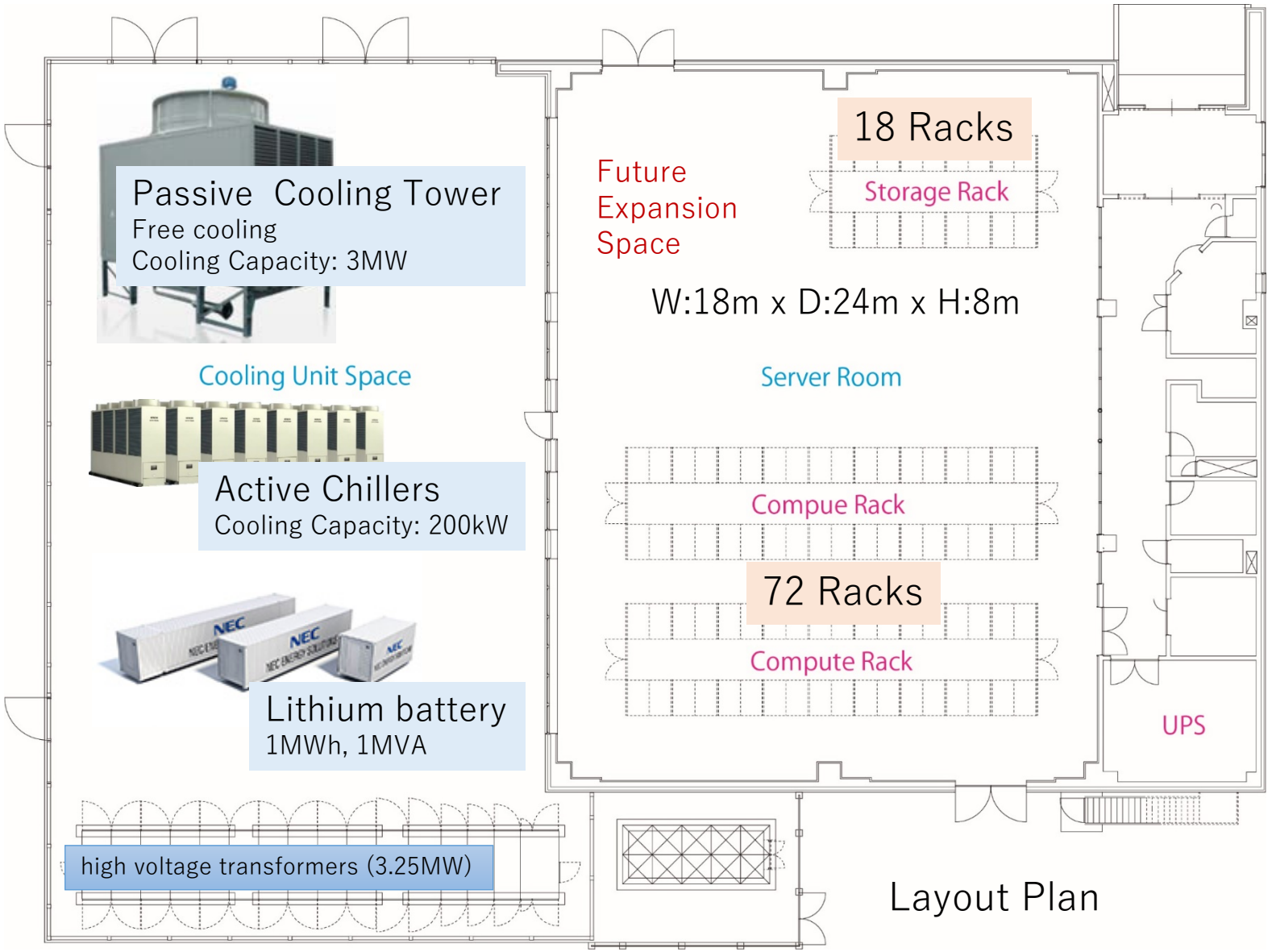
イメージスケッ

Reference Image

引用元: NEC導入事例

ABCI Cloud Data Center

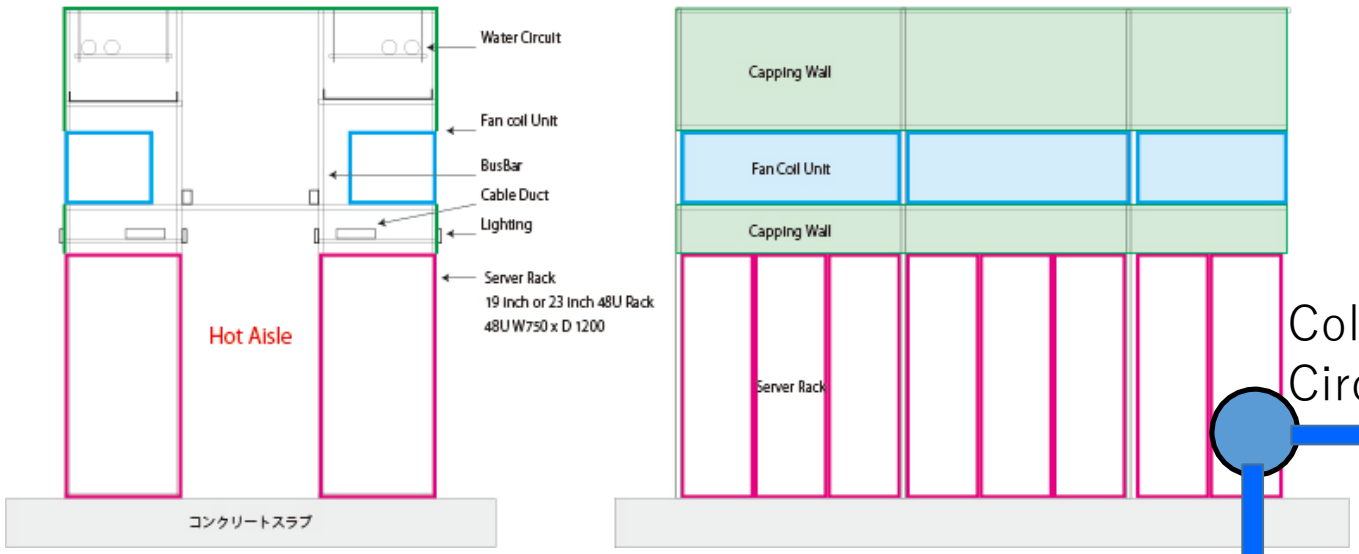
“Commoditizing 60KW/rack Supercomputer”



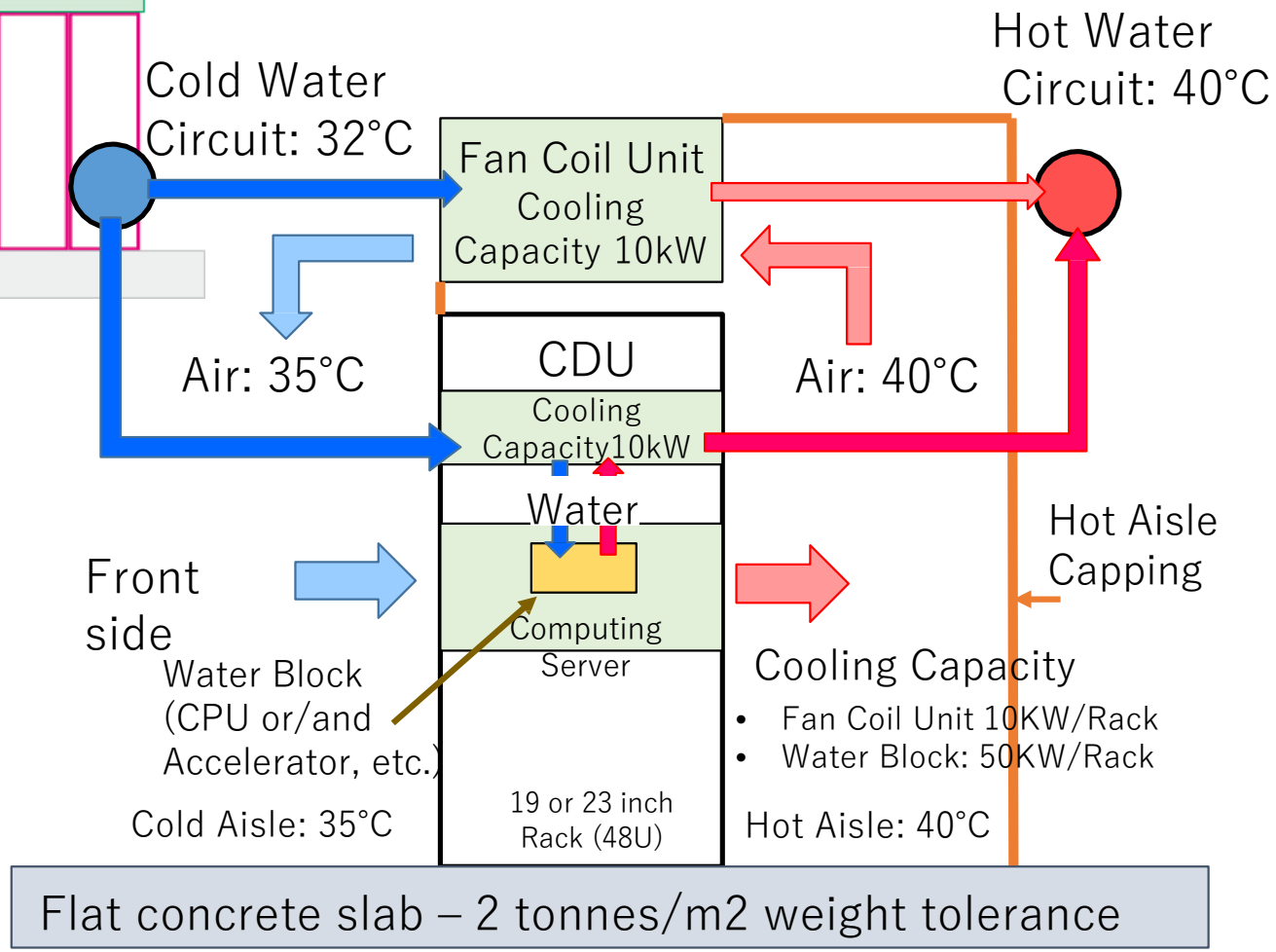
Layout Plan

- Single Floor, inexpensive build
- Hard concrete floor 2 tonnes/m2 weight tolerance for racks and cooling pods
- Number of Racks
  - Initial: 90
  - Max: 144
- Power Capacity
  - 3.25 MW (MAX)
- Cooling Capacity
  - 3.2 MW (Minimum in Summer)

Implementing 60KW cooling in Cloud IDC – Cooling Pods



Cooling Block Diagram (Hot Rack)



Commoditizing Supercomputing Cooling Density and Efficiency

- Warm water cooling – 32C
- Liquid cooling & air cooling in same rack
- 60KW Cooling Capacity, 50KW Liquid+10KW Air
- Very low PUE
- Structural integrity by rack + skeleton frame built on high flat floor load

TSUBAME3.0&ABCI Comparison

- TSUBAME3: “Big Data and AI-oriented Supercomputer”
  ABCI: “Supercomputer-oriented next-gen IDC template for AI & Big Data”
- The two machines are sisters but above dictates their differences
  - Hardware: TSUBAME3 still emphasizes DFP performance as well as extreme injection and bisection interconnect bandwidth. ABCI does not require high DFP performance, and reduces interconnect requirement for cost reduction and IDC friendliness
  - TSUBAME3 node & machine packaging is custom co-designed as a supercomputer based on SGI/HPE ICE-XA, with extreme performance density (3.1 PetaFlops/rack) thermal density (61KW/rack), extremely low PUE=1.033. ABCI aims for similar density and efficiency in a 19inch IDC ecosystem
  - Both will converge HPC and BD/AI/ML software stacks, but ABCI adoption of the latter will be quicker and comprehensive given the nature of the machine
- The major theme of ABCI is “How to disseminate TSUBAME3-class AI-Oriented supercomputing in the Cloud” ==> other performance parameters are similar to TSUBAME3
  - Compute and data parameters are similar except for the interconnect
  - Thermal density (50~60KW/rack c.f. 3~6KW/rack for standard IDC), PUE<1.1 (standard IDC 1.5~3)
- We are also building ABCI-IDC, as the proof-of-concept datacenter building infrastructure that will be a template for future high-density high-performance “convergent” datacenters

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

AIRC

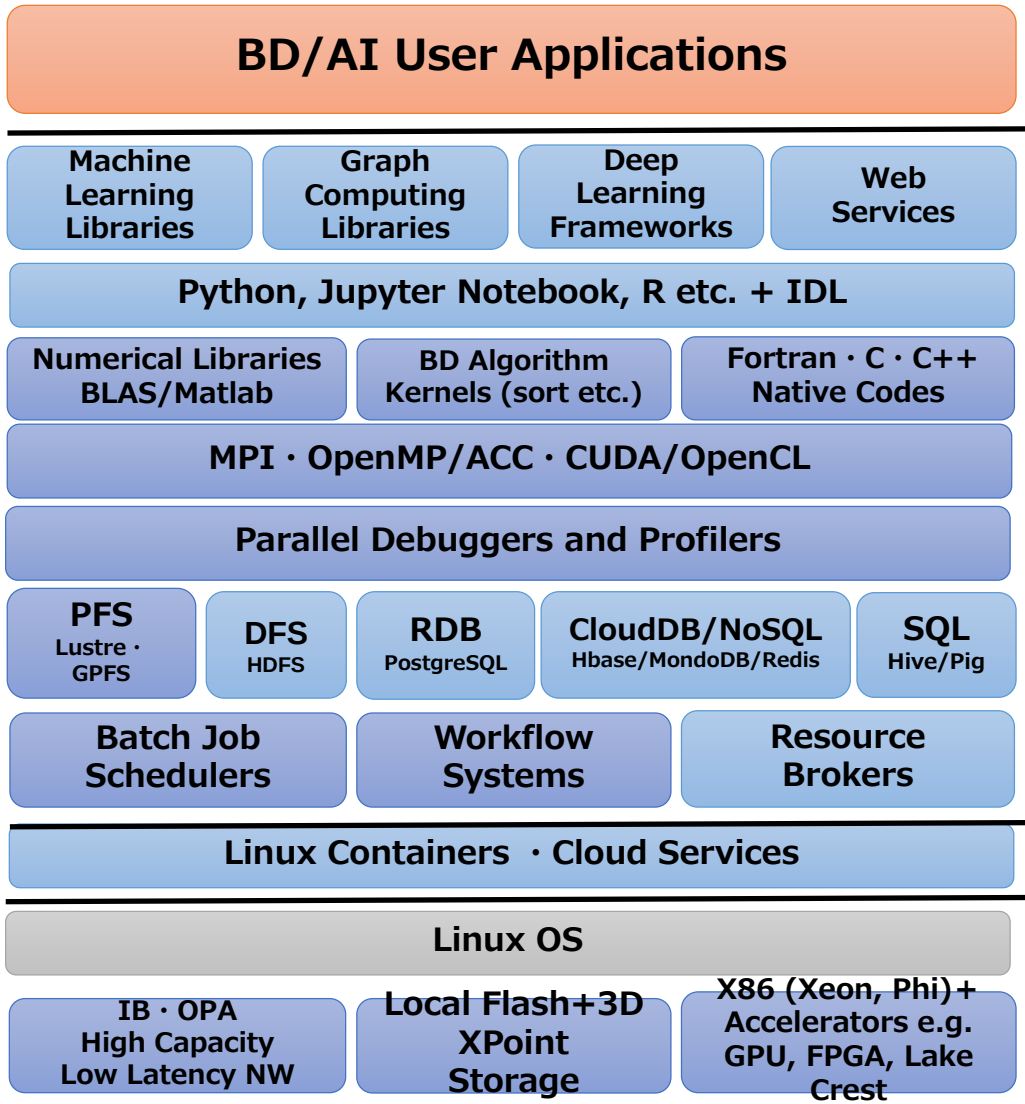
TSUBAME3.0&ABCI Comparison Chart

TSUBAME
Tokyo Institute of Technology

GSIC
Global Scientific Information and Computing Center

	TSUBAME3 (2017/7)	ABCI (2018/3)	C.f.: K (2012)
AI-FLOPS Peak AI Performance	47.2 Pflops (DFP 12.1 PFlops) 3.1 PetaFlops/rack	130~200 Pflops, (DFP NA) 3~4 PetaFlops/rack	11.3 Petaflops 12.3 Tflops/rack
System Packaging	Custom SC (ICE-XA), Liquid Cool	19 inch rack (LC), ABCI-IDC	Custom SC (LC)
Operational Power incl. Cooling	Below 1MW	Approx. 2MW	Over 15MW
Max Rack Thermals & PUE	61KW, 1.033	50-60KW, below 1.1	~20KW, ~1.3
Node Hardware Architecture	Many-Core (NVIDIA Pascal P100) + Multi-Core (Intel Xeon)	Many-Core AI/DL oriented processor (incl. GPUs)	Heavyweight Multi-Core
Memory Technology	HBM2+DDR4	On Die Memory + DDR4	DDR3
Network Technology	Intel OmniPath, 4 x 100Gbps / node, full bisection, inter-switch optical network	Both injection & bisection BW will be scaled down c.f. T3 to save cost & IDC friendly	Copper Tofu 6-D torus custom interconnect
Per-node non volatile memory	2TeraByte NVMe/node	> 400GB NVMe/node	None
Power monitoring and control	Detailed node / whole system power monitoring & control	Detailed node / whole system power monitoring & control	Whole system monitoring only
Cloud and Virtualization, AI	**All nodes container virtualization, horizontal node splits, Cloud API dynamic provisioning, ML Stack**	**All nodes container virtualization, horizontal node splits, Cloud API dynamic provisioning, ML Stack**	None

Basic Requirements for AI Cloud System



Application

- ✓ Easy use of various ML/DL/Graph frameworks from Python, Jupyter Notebook, R, etc.
- ✓ Web-based applications and services provision

System Software

- ✓ HPC-oriented techniques for numerical libraries, BD Algorithm kernels, etc.
- ✓ Supporting long running jobs / workflow for DL
- ✓ Accelerated I/O and secure data access to large data sets
- ✓ User-customized environment based on Linux containers for easy deployment and reproducibility

OS

Hardware

- ✓ Modern supercomputing facilities based on commodity components