Title: Converging HPC and Big Data / AI Infrastructures at Scale with BYTES-Oriented Architectures

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

IPDPS 2017 WS-HPBDC Keynote Presentation
2017/05/29
Orlando, Florida, USA

TSUBAME2.0 Nov. 1, 2010
"The Greenest Production Supercomputer in the World"

- GPU-centric (> 4000) high performance & low power
- Small footprint (~200m2 or 2000 sq.ft), low TCO
- High bandwidth memory, optical network, SSD storage…

TSUBAME 2.0
New Development

Chip
(CPU ,GPU)

CPU(Westmere EP)
76.8 GFLOPS
32nm

GPUs(Tesla M2050)
515 GFLOPS
3 GB 40nm

Compute Node
(2 CPUs,3 GPUs)

1.6 TFLOPS
55 GB/103 GB
>400GB/s Mem BW
80Gbps NW BW
~1KW max

Node Chassis
(4 Compute Nodes)

6.7 TFLOPS
220 GB/412 GB
>1.6TB/s Mem BW

Rack
(8 Node Chassis)

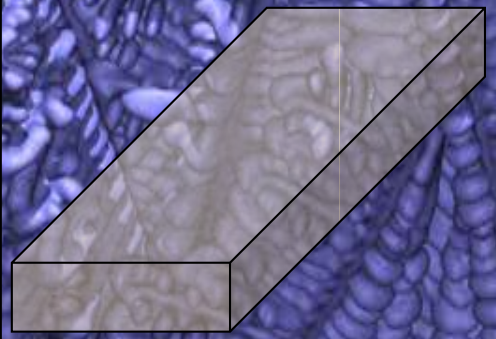
53.6 TFLOPS
1.7 TB/3.2 TB
>12TB/s Mem BW
35KW Max

System
(42 Racks)
1408 GPU Compute Nodes,
34 Nehalem "Fat Memory" Nodes

2.4 PFLOPS
80 TB
4224 GPUs
>600TB/s Mem BW
220Tbps NW
Bisecion BW
1.4MW Max

Integrated by NEC Corporation

ACM Gordon Bell Prize 2011
2.0 Petaflops Dendrite Simulation



Lightweight and durable metal alloy for future automobiles

Special Achievements in Scalability and Time-to-Solution
"Peta-Scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer"
2 Petaflops (3.4 Petaflops on TSUBAME 2.5)

Comparing K Computer to TSUBAME 2.5

東京工業大学
Tokyo Institute of Technology

GSIC
Global Scientific Information and Computing Center

Perf ≒
Cost ≪

独立行政法人理化学研究所
計算科学研究機構
RIKEN Advanced Institute for Computational Science

K Computer (2011)

TSUBAME2.0(2010)
→ TSUBAME2.5(2013)
17.1 Petaflops SFP
5.76 Petaflops DFP
$45mil / 6 years (incl. power)

BG/Q Sequoia (2011)
22 Petaflops SFP/DFP

11.4 Petaflops SFP/DFP
$1400mil 6 years
(incl. power)
x30 TSUBAME2

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

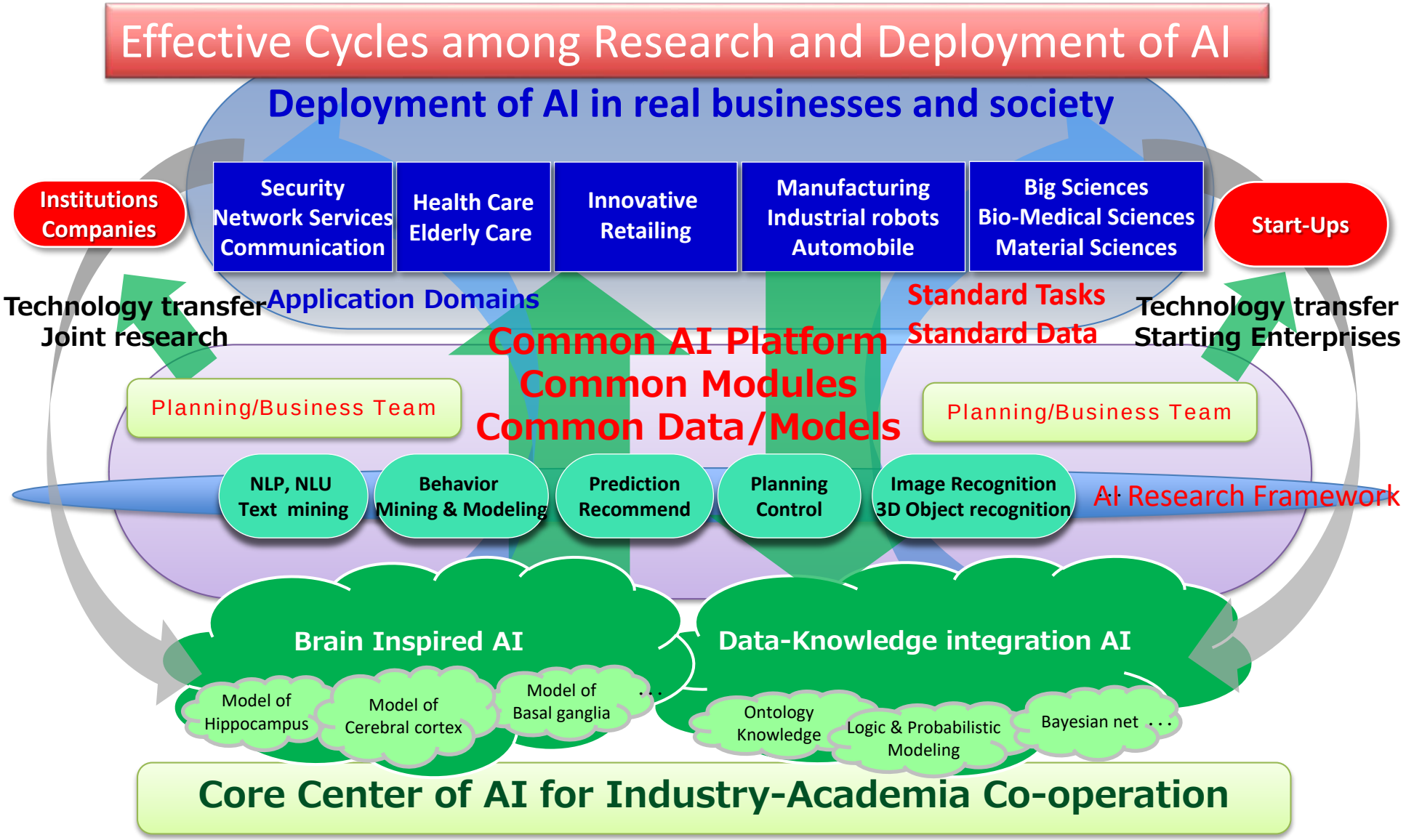
- Three national centers on Big Data and AI launched by three competing Ministries for FY 2016 (Apr 2015-)
  - METI – AIRC (Artificial Intelligence Research Center): AIST (AIST internal budget + > $200 million FY 2017), April 2015
    - Broad AI/BD/IoT, industry focus
  - MEXT – AIP (Artificial Intelligence Platform): Riken and other institutions ($~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)

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

Tsubame 3.0/2.5 Big Data /AI resources

Ministry of Economics Trade and Industry (METI)

Resources and Acceleration of AI / Big Data, systems research

AIST Artificial Intelligence Research Center (AIRC)

Application Area
Natural Langauge Processing
Robotics
Security

Joint Research on AI / Big Data and applications

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

Director: Satoshi Matsuoka

Basic Research in Big Data / AI algorithms and methodologies

ITCS Departments

Industrial Collaboration in data, applications

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

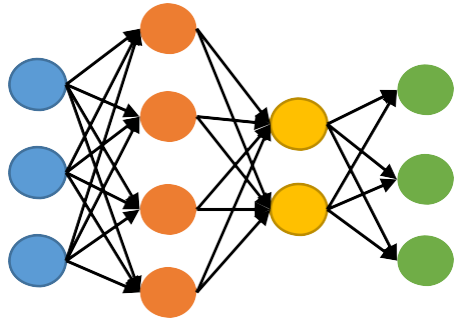
Acceleration, Scaling

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



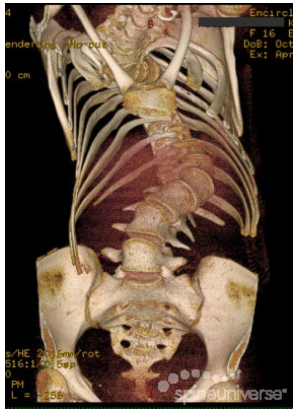
As HPC Task

Dense Matrices, Reduced Precision
Dense and well organized neworks and Data

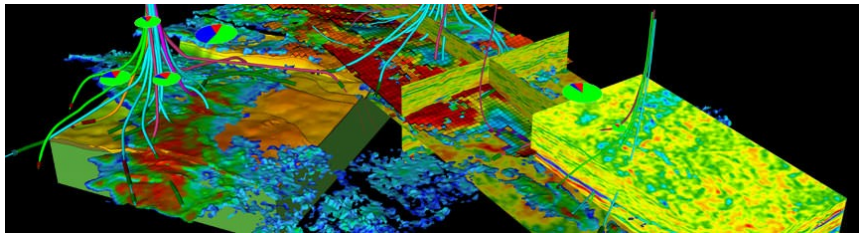
Acceleration, Scaling

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

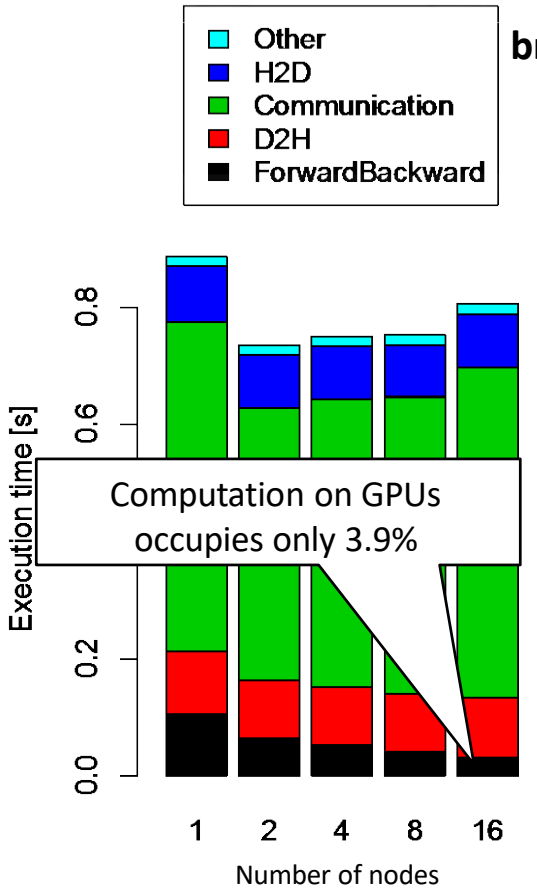
- **Need <u>BOTH bandwidth and capacity</u> (BYTES) in a HPC-BD/AI machine:**
  - Obvious for lefthand sparse ,bandwidth-dominated apps
  - But also for righthand DNN: Strong scaling, large networks and datasets, in particular for future 3D dataset analysis such as CT-scans, seismic simu. vs. analysis…)



(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

"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

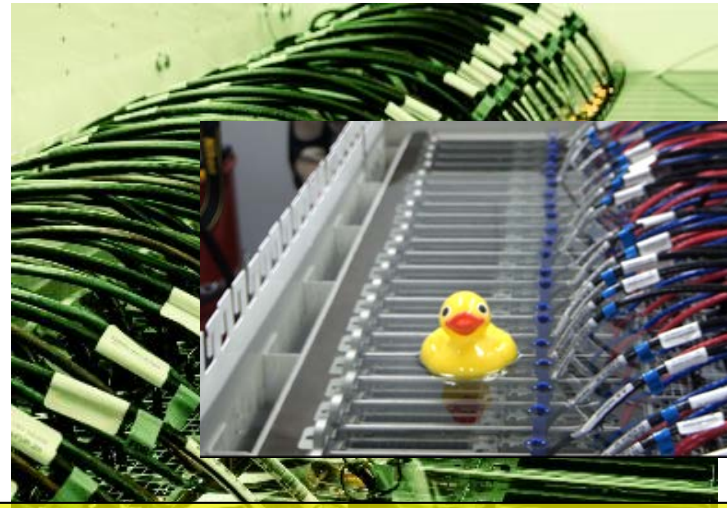
Over $1B Govt. AI investment over 10 years

AI&Data Infrastructures

Massive "Big" Data in Training

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

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

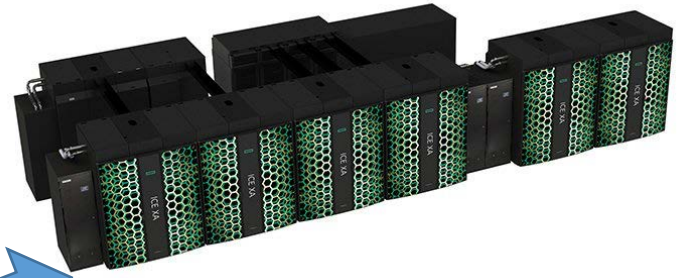
1. "Everybody's Supercomputer" - High Performance (12~24 DP Petaflops, 125~325TB/s Mem, 55~185Tbit/s NW), innovative high cost/performance packaging & design, in mere 180m^2…
2. "Extreme Green" – ~10GFlops/W power-efficient architecture, system-wide power control, advanced cooling, future energy reservoir 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"

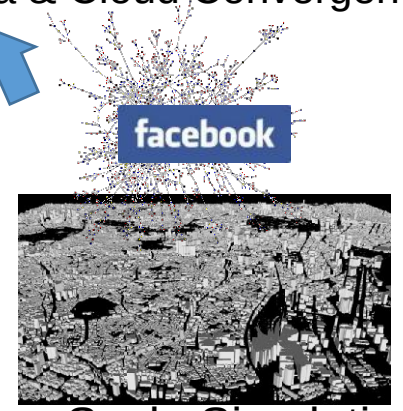
2011 ACM Gordon Bell Prize

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



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

2013 TSUBAME-KFC
#1 Green 500



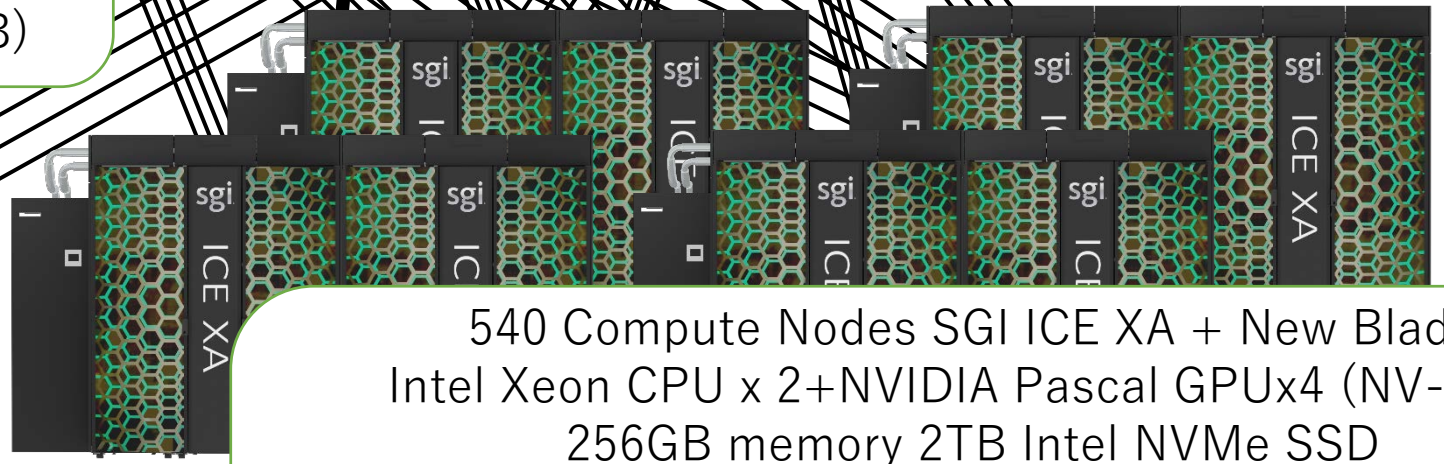
Large Scale Simulation
Big Data Analytics
Industrial Apps

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

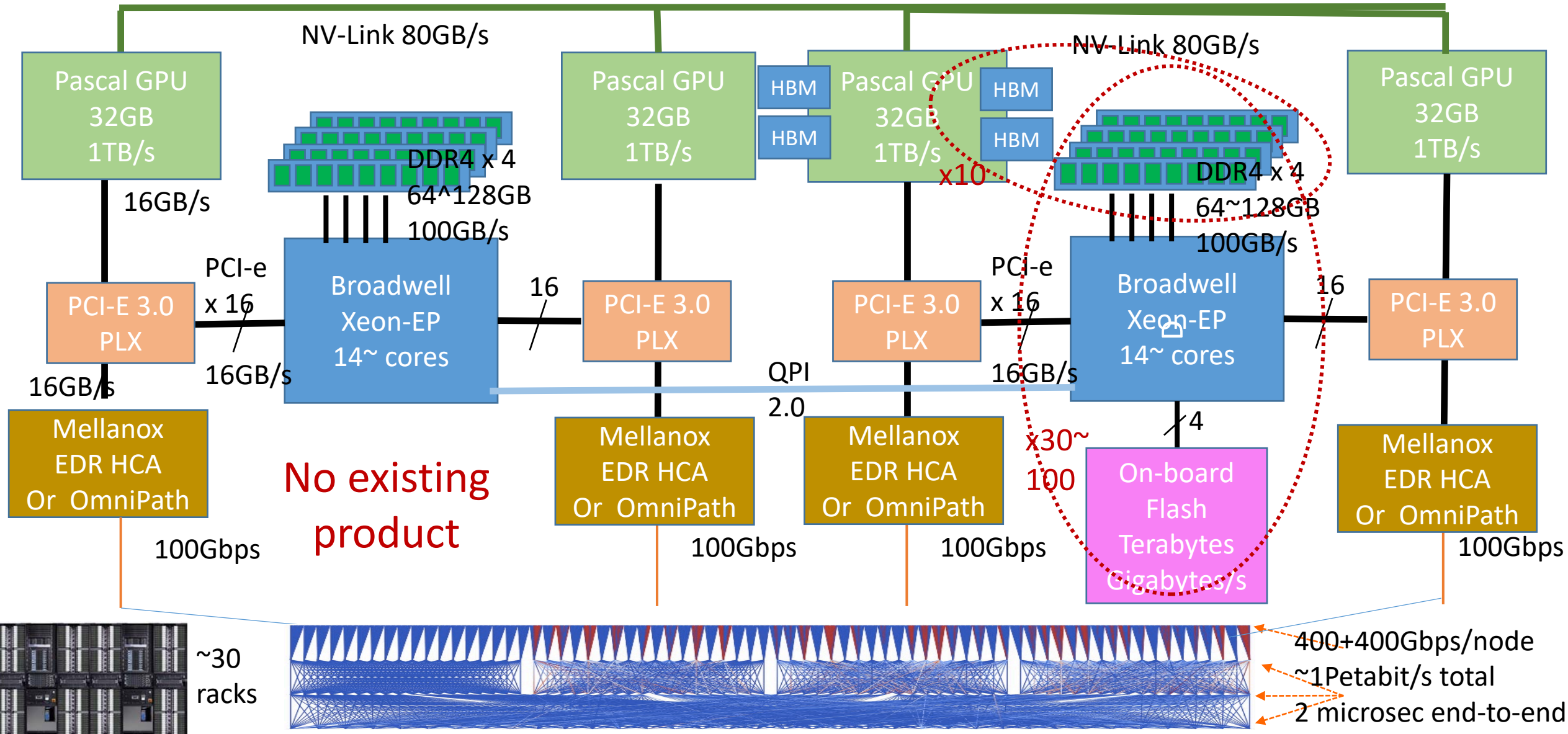
DDN Storage
(Lustre FS 15.9PB+Home 45TB)



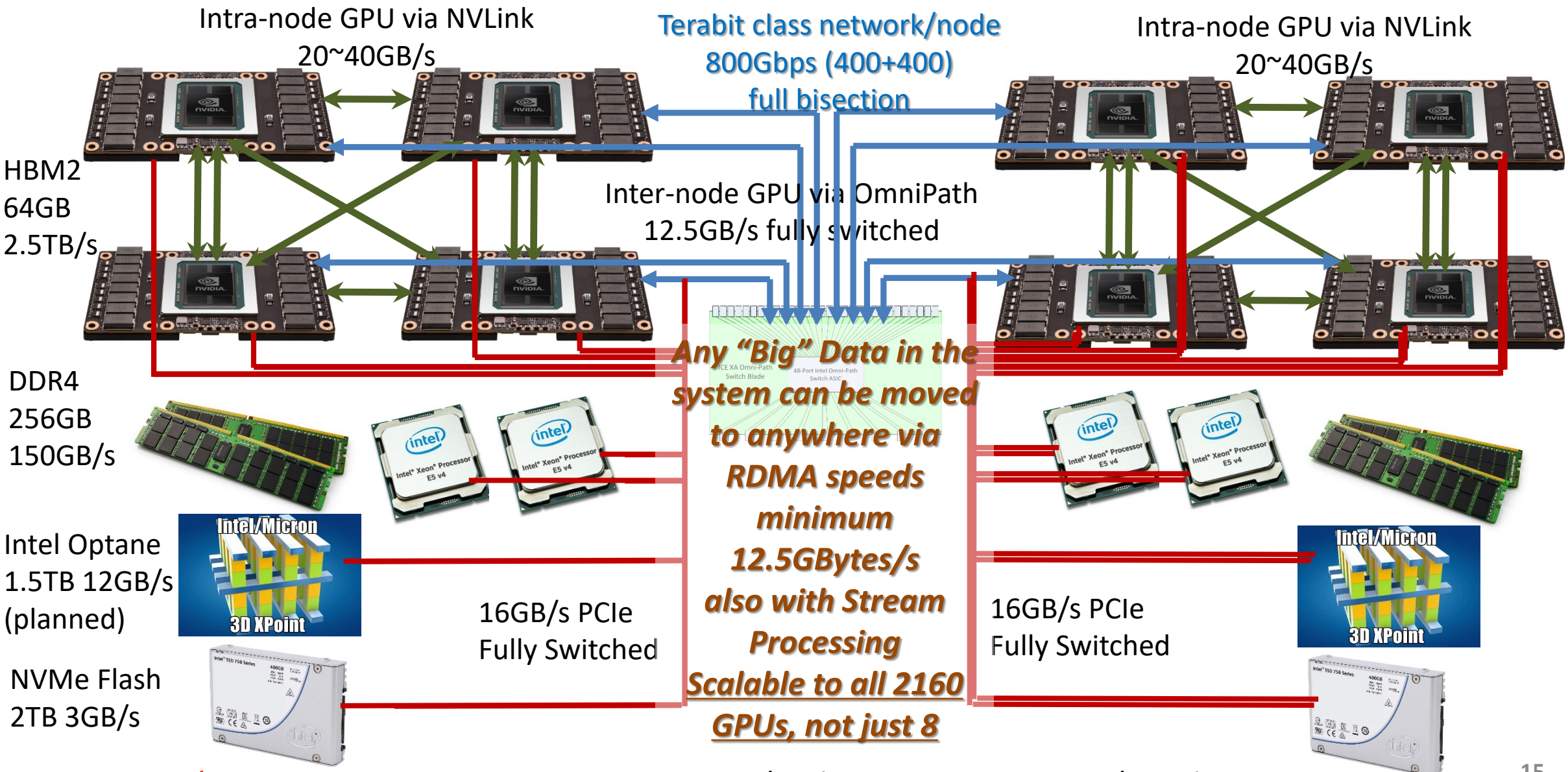
540 Compute Nodes SGI ICE XA + New Blade
Intel Xeon CPU x 2+NVIDIA Pascal GPUx4 (NV-Link)
256GB memory 2TB Intel NVMe SSD
47.2 AI-Petaflops, 12.1 Petaflops

Early TSUBAME3 Architecture for Proposal

Ultra High BW, Deep Mem Hierarchy, Low Latency NW



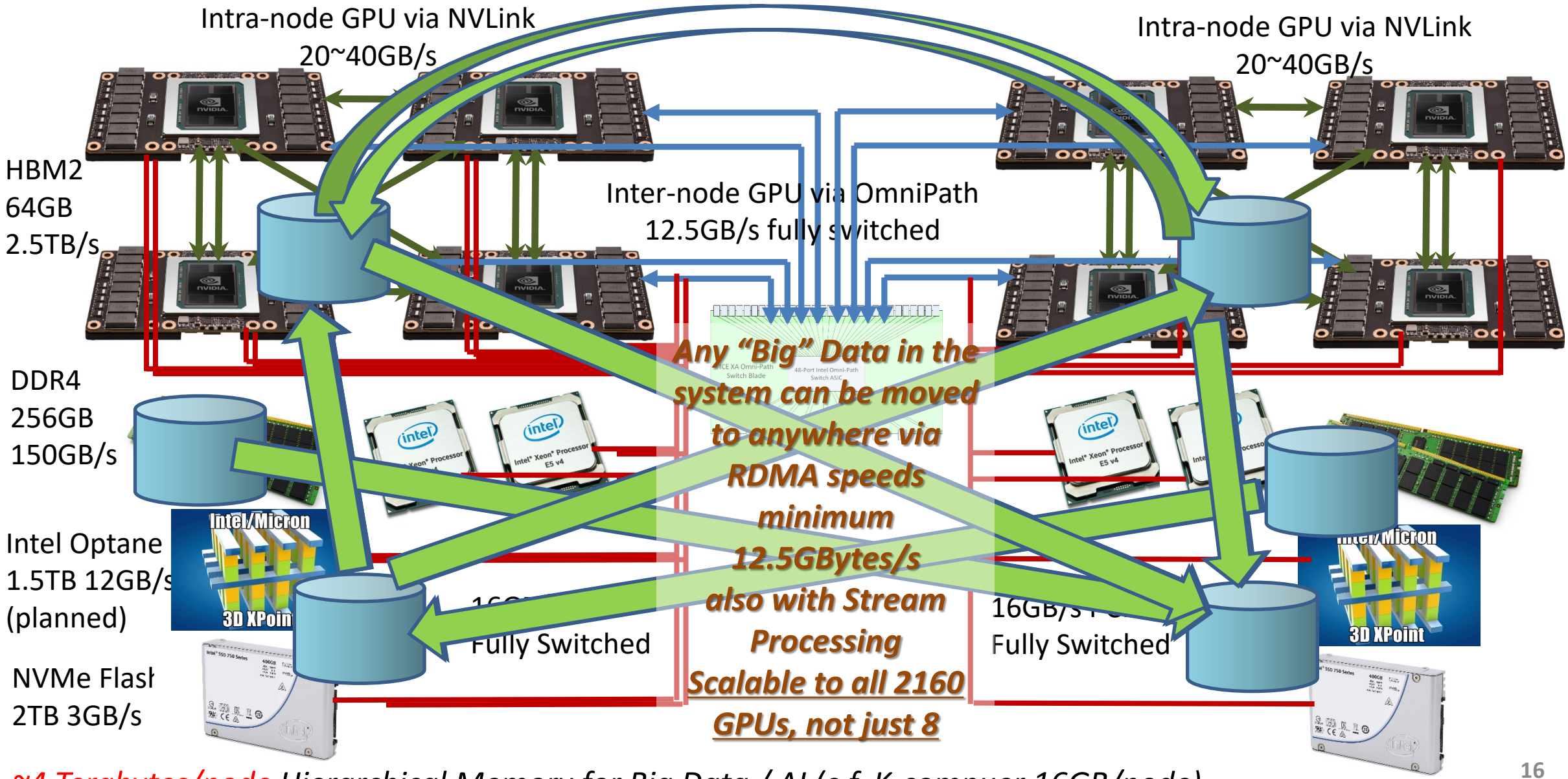
TSUBAME3: *A Massively BYTES Centric Architecture* for Converged BD/AI and HPC



~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

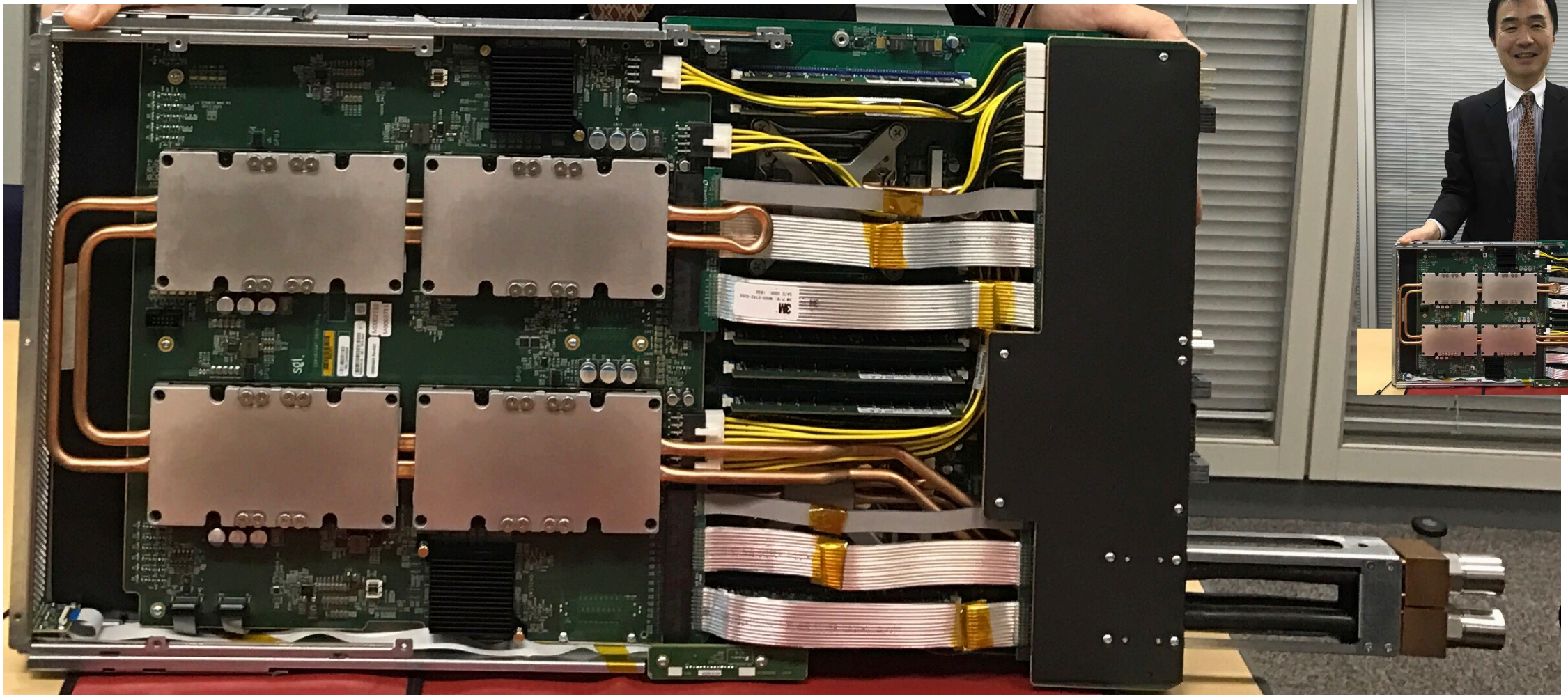


~4 Terabytes/node Hierarchical Memory for Big Data / AI (c.f. K-compuer 16GB/node)

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

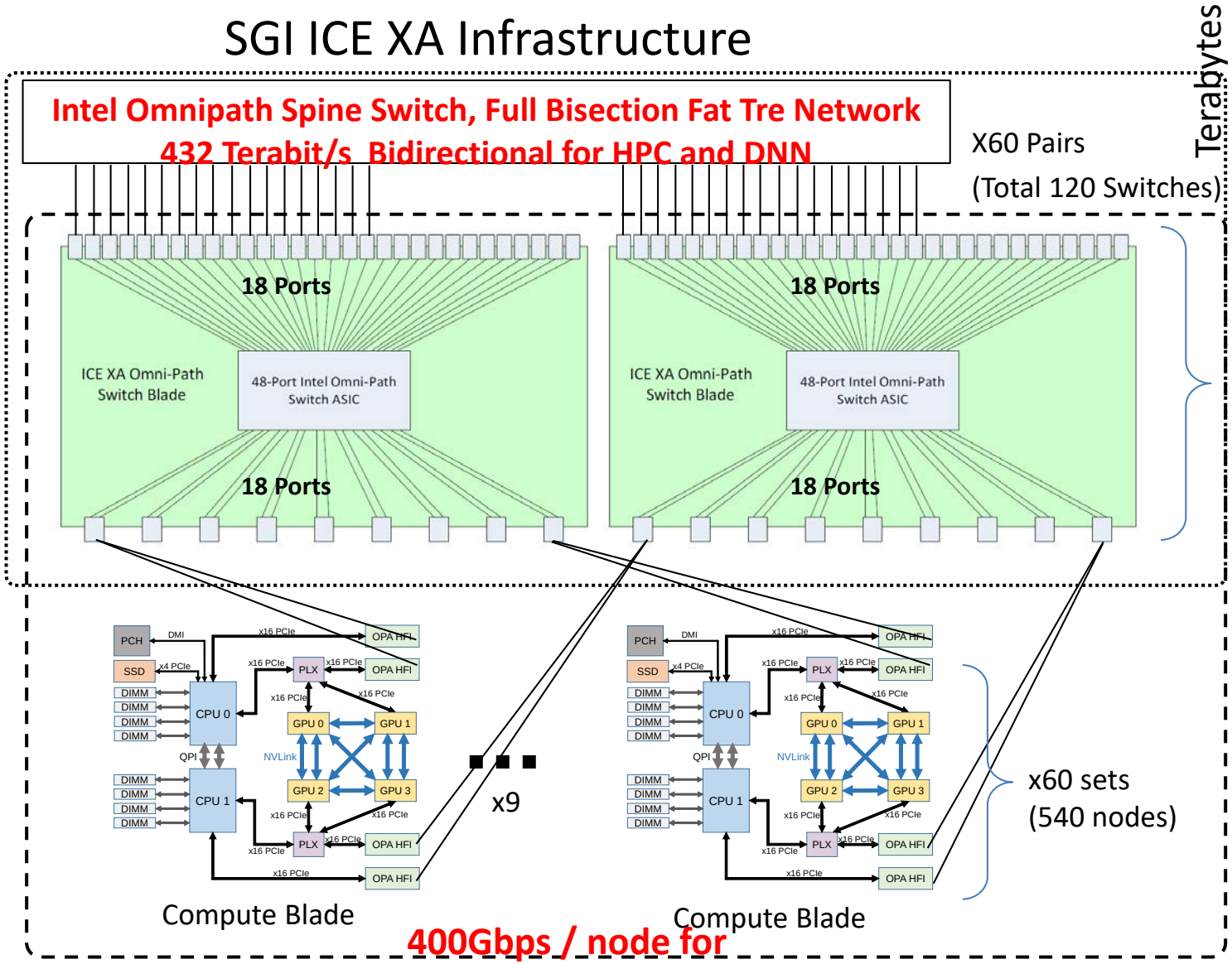
TSUBAME3.0 Co-Designed SGI ICE-XA Blade (new)

- 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



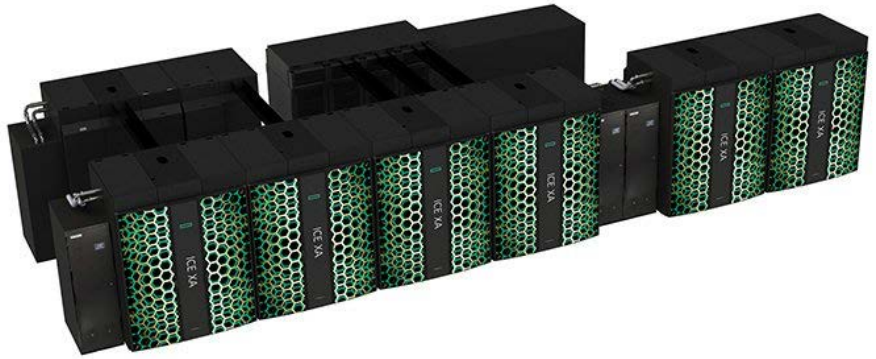
Ultra high performance & bandwidth "Fat Node"

- 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

Node Performance Comparison T2/2.5/3

Metric	TSUBAME2.0 (2010)	TSUBAME2.5 (2013)	TSUBAME3.0 (2017)	Factor
CPU Cores x Freq (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 Peak (TFLOPS)	1.58	3.93	21.2	13.4 & 5.39
GPU FP32 Peak (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
Interconnect Bandwidth (Gbps)	80	80	400	5.00

TSUBAME3.0 Datacenter

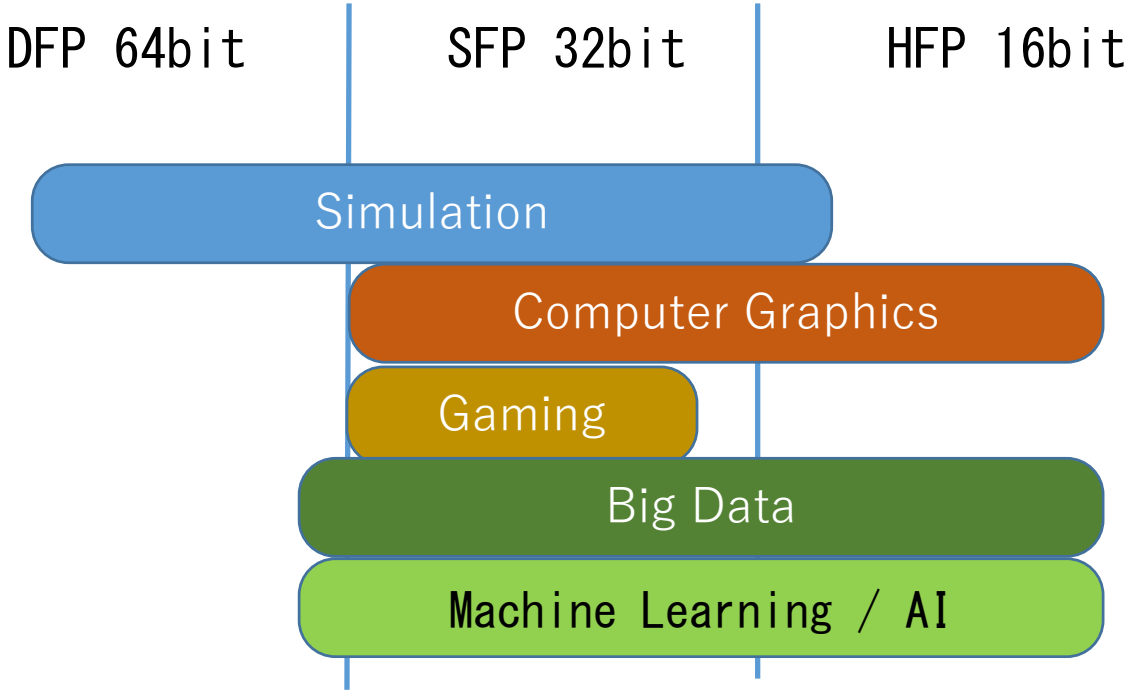


15 SGI ICE-XA Racks
2 Network Racks
3 DDN Storage Racks
20 Total Racks

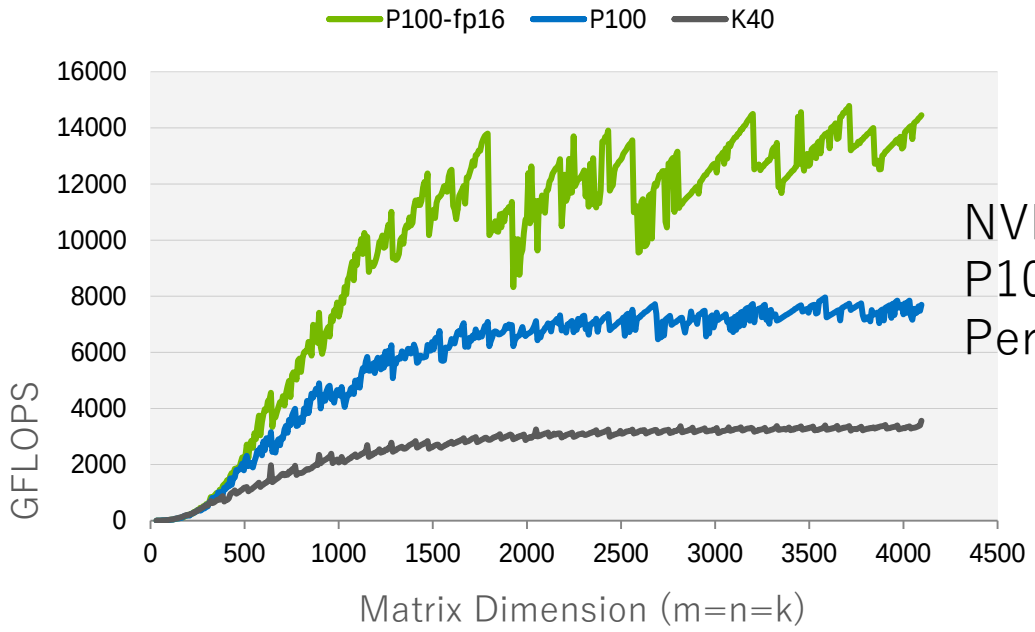
Compute racks cooled with
32 degrees warm water,
Yearound ambient cooling
Av. PUE = 1.033

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

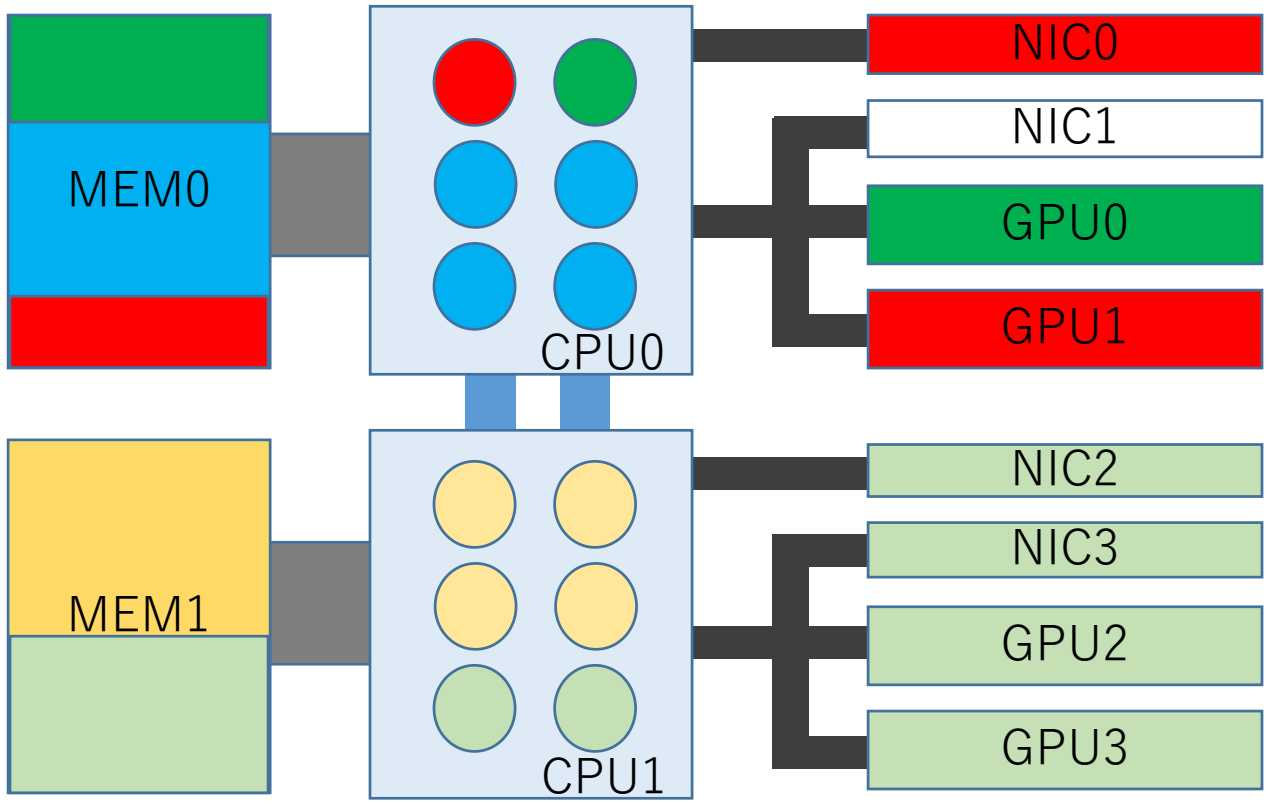


Tokyo Tech GSIC leads Japan in aggregated AI-capable FLOPS TSUBAME3+2.5+KFC, in all Supercomuters and CloudsNV



NVIDIA Pascal P100 DGEMM Performane

TSUBAME3.0 Container-Based Fine-grained Spatial Resource Allocations of Fat Nodes



Job	Allocated Resource
1	CPU 2Cores, NIC0, GPU1, 32GB Mem
2	CPU 8 Cores, 64GB Mem
3	CPU 4 Cores, GPU0、16GB Mem
4	CPU 8 Cores, 64GB Mem
5	CPU 4 Cores, NIC2&3, GPU2&3, 48G Mem

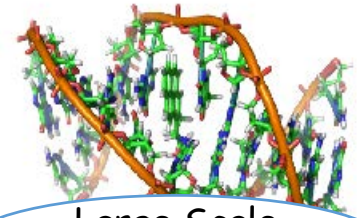
Resource Isolation via UGE Containers (future Docker etc.)

Container configuration and deployment tied to Univa Grid Engine

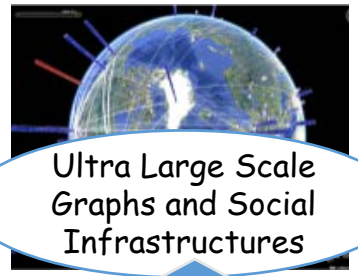
JST-CREST "Extreme Big Data" Project (2013-2018)

From FLOPS Centric to BYTES Centric HPC

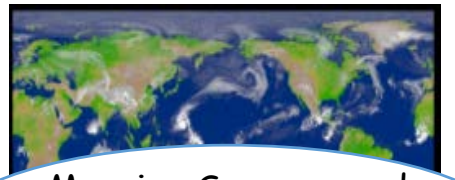
Given a top-class supercomputer, how fast can we accelerate next generation big data c.f. conventional Clouds?



Large Scale Metagenomics



Ultra Large Scale Graphs and Social Infrastructures



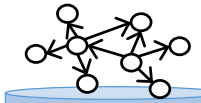
Massive Sensors and Data Assimilation in Weather Prediction

Co-Design

Co-Design

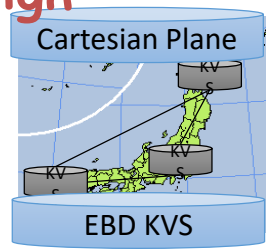
Co-Design

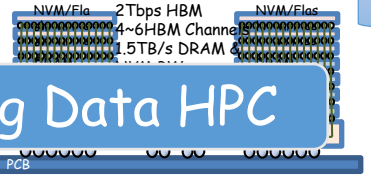
EBD Bag



Graph Store

EBD System Software incl. EBD Object System





Exascale Big Data HPC

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

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

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

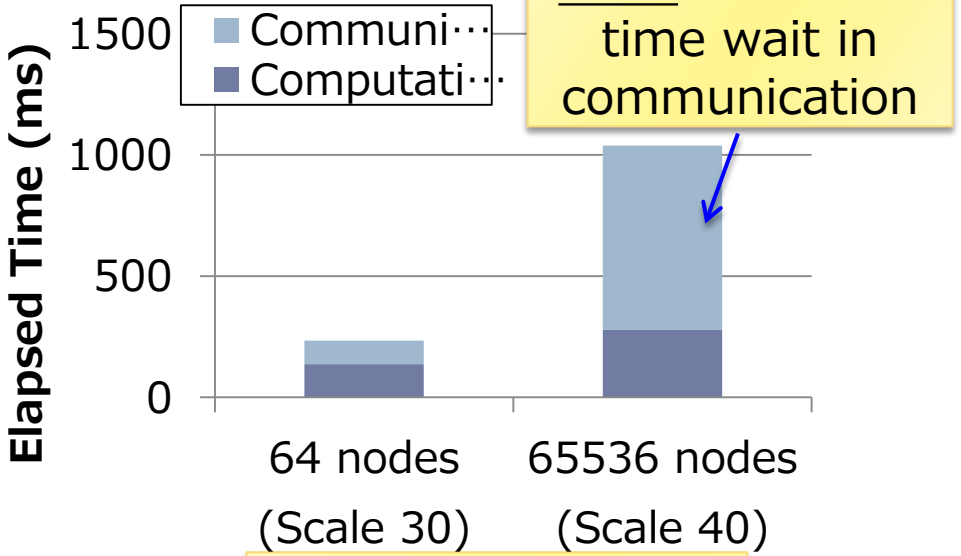
Cloud IDC
Very low BW & Efficiency
Highly available, resilient

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

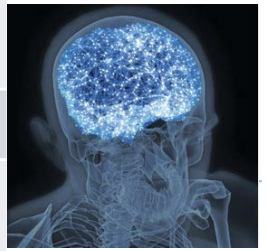
Effective x13 performance c.f. Linpack

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

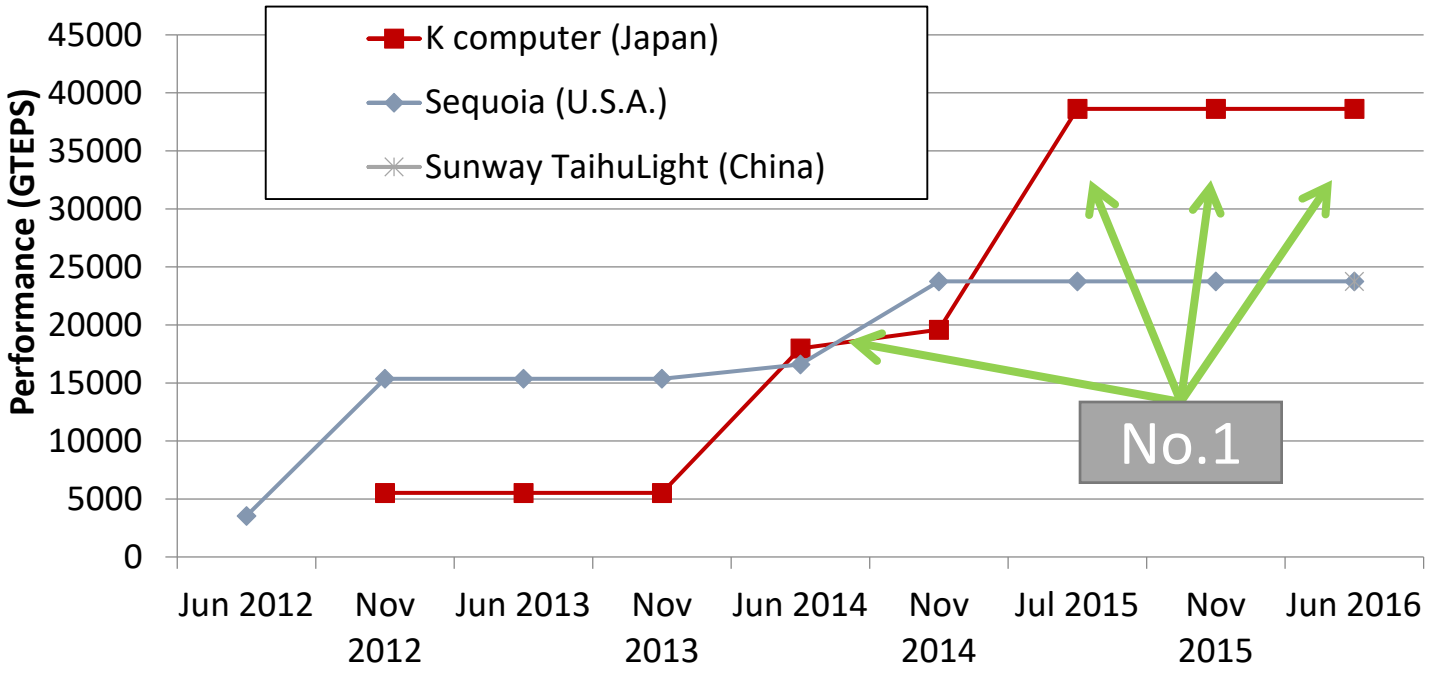
BYTES Rich Machine + Superior BYTES algoithm



BYTES, not FLOPS!

K-computer No.1 on Graph500: 4th Consecutive Time

- What is Graph500 Benchmark?
  - Supercomputer benchmark for data intensive applications.
  - Rank supercomputers by the performance of **Breadth-First Search** for very huge graph data.



This is achieved by a combination of high machine performance and **our software optimization**.

- Efficient Sparse Matrix Representation with Bitmap
- Vertex Reordering for Bitmap Optimization
- Optimizing Inter-Node Communications
- Load Balancing

  etc.

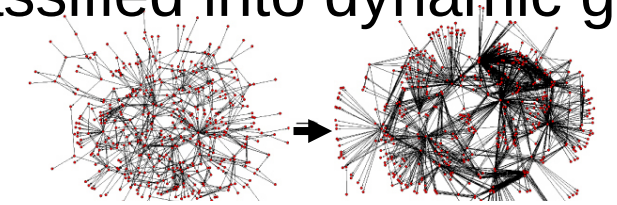
- Koji Ueno, Toyotaro Suzumura, Naoya Maruyama, Katsuki Fujisawa, and Satoshi Matsuoka, "**Efficient Breadth-First Search on Massively Parallel and Distributed Memory Machines**", in proceedings of 2016 IEEE International Conference on Big Data (IEEE BigData 2016), Washington D.C., Dec. 5-8, 2016 (to appear)

Distributed Large-Scale Dynamic Graph Data Store

Keita Iwabuchi[1, 2], Scott Sallinen[3], Roger Pearce[2],
Brian Van Essen[2], Maya Gokhale[2], Satoshi Matsuoka[1]
1. Tokyo Institute of Technology (Tokyo Tech)
2. Lawrence Livermore National Laboratory (LLNL)
3. University of British Columbia

Dynamic Graphs (temporal graph)

- the structure of a graph changes dynamically over time
- many real-world graphs are classified into dynamic graph



Source: Jakob Enemark and Kim Sneppen, “Gene duplication models for directed networks with limits on growth”, Journal of Statistical Mechanics: Theory and Experiment 2007

Sparse Large Scale-free

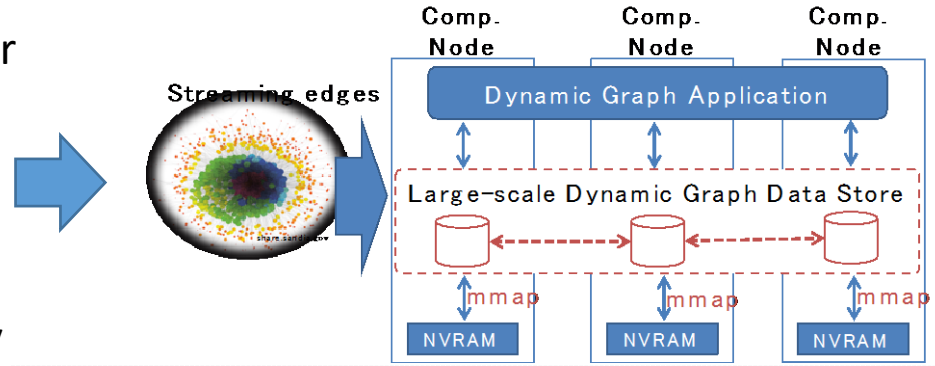
- social network, genome analysis, WWW, etc.
  - e.g., Facebook manages 1.39 billion active users as of 2014, with more than 400 billion edges

- Most studies for large graphs have not focused on a dynamic graph data structure, but rather a static one, such as Graph 500
- Even with the large memory capacities of HPC systems, many graph applications require additional out-of-core memory (this part is still at an early stage)

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

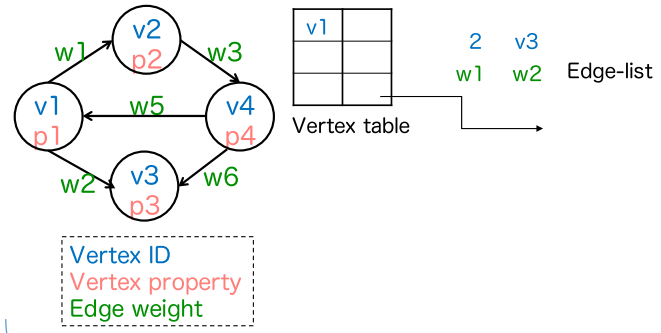
K Computer large memory but very expensive DRAM only



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)

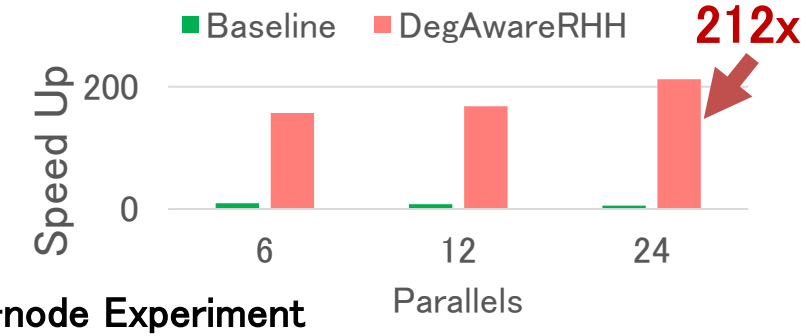
C.f. STINGER (single-node, on memory)

STINGER

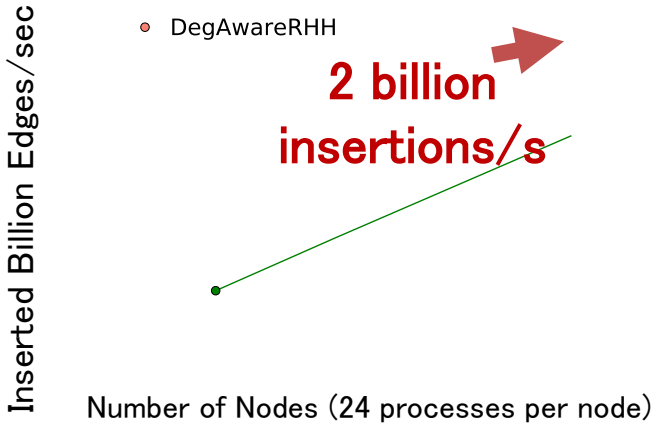
- A state-of-the-art dynamic graph processing framework developed at Georgia Tech

Baseline model

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



Multi-node Experiment

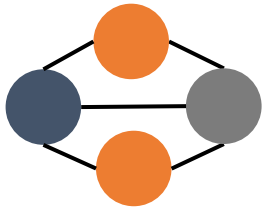


Dynamic graph store w/ world's top graph update performance and scalability

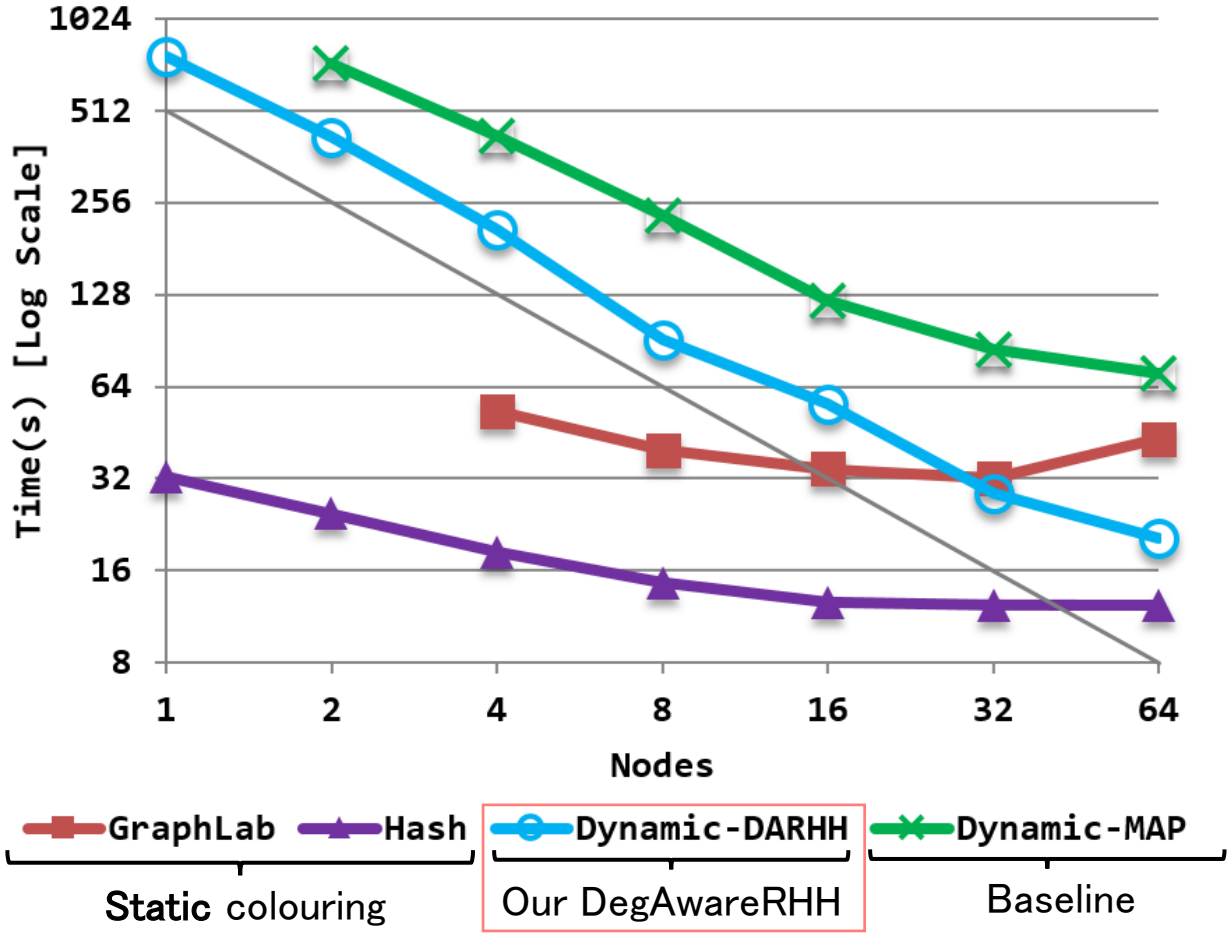
K. Iwabuchi, S. Sallinen, R. Pearce, B. V. Essen, M. Gokhale, and S. Matsuoka, Towards a distributed large-scale dynamic graph data store. In 2016 IEEE Interna- tional Parallel and Distributed Processing Symposium Workshops (IPDPSW)

Large-scale Graph Colouring (vertex coloring)

SC'16



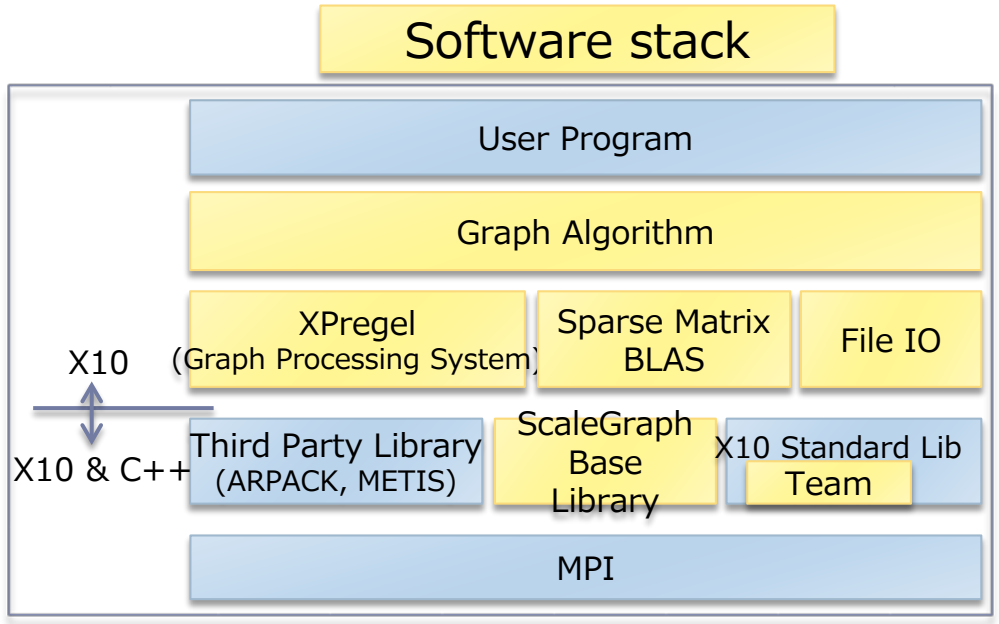
- Color each vertices with the minimal #colours so that **no** two adjacent vertices have the same colour
- Compare our dynamic graph colouring algorithm on DegAwareRHH against:
  1. two static algorithms including GraphLab
  2. an another graph store implementation with same dynamic algorithm (Dynamic-MAP)

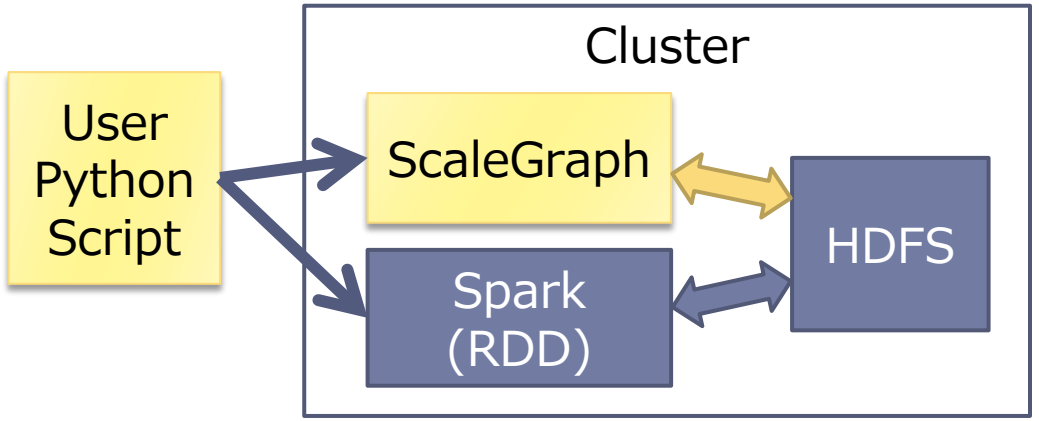


Scott Sallinen, Keita Iwabuchi, Roger Pearce, Maya Gokhale, Matei Ripeanu, "Graph Coloring as a Challenge Problem for Dynamic Graph Processing on Distributed Systems", SC' 16

ScaleGraph Large-scale Graph Processing Framework enhanced w/ User-Friendly Python / Spark Interface

- ScaleGraph [Suzumura]
  - X10-based open source **Highly Scalable Large Scale Graph Analytics Library** beyond the scale of billions of vertices and edges on Distributed Systems
    - **XPregel**: Pregel-based bulk synchronous parallel graph processing framework
    - Built-in graph algorithms (Centrality, Connected Component, Clustering, etc.)
- NEW Development: Python Interface
  - Allow users to use ScaleGraph with Spark* by easy python interface

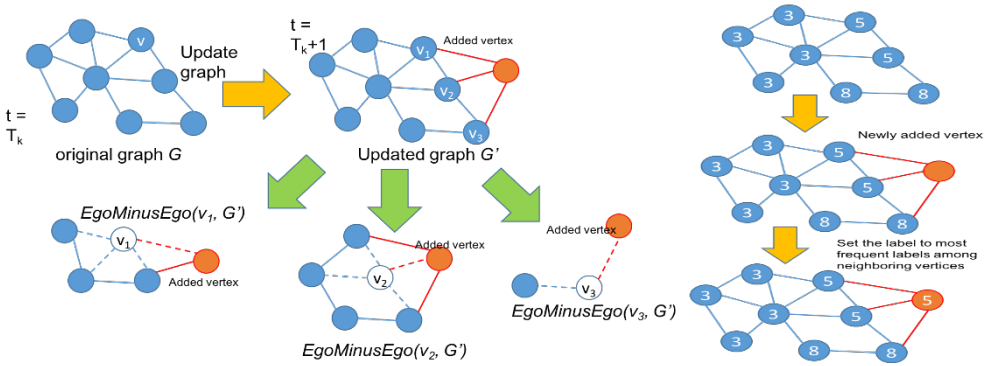


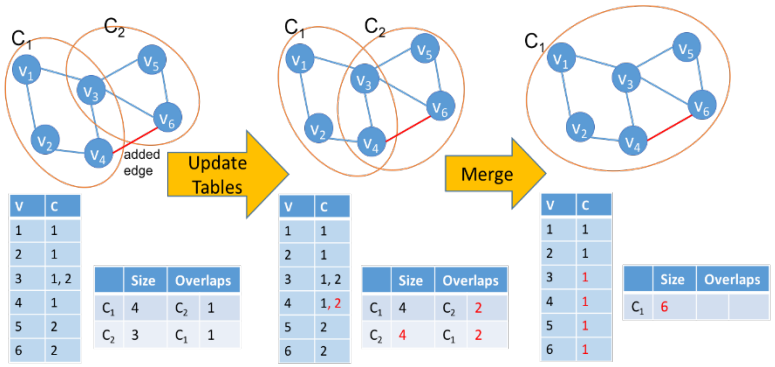


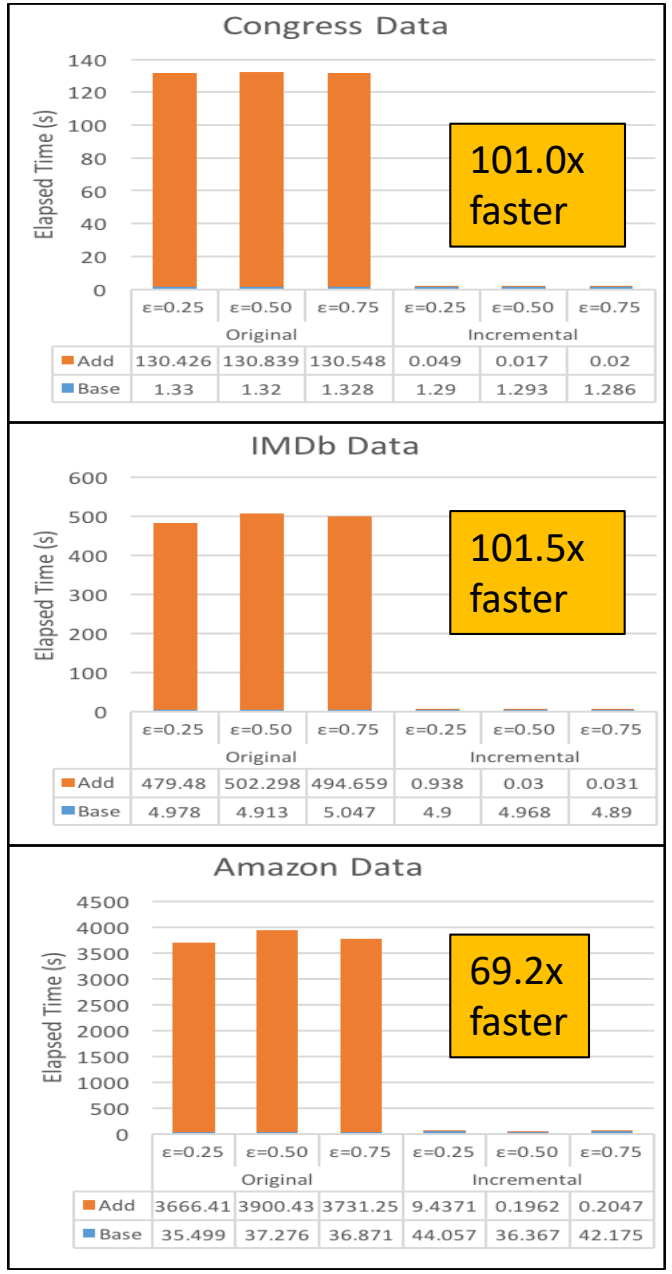
*Apache Spark: http://spark.apache.org/

Incremental Graph Community Detection

- Background
  - Community detection for large-scale **time-evolving and dynamic graphs** has been one of important research problems in graph computing.
  - It is time-wasting to compute communities entire graphs every time from scratch.
- Proposal
  - **An incremental community detection algorithm** based on core procedures in a state-of-the-art community detection algorithm named DEMON.
    - Ego Minus Ego, Label Propagation and Merge







Hiroki Kanezashi and Toyotaro Suzumura, An Incremental Local-First Community Detection Method for Dynamic Graphs, Third International Workshop on High Performance Big Graph Data Management, Analysis, and Mining (BigGraphs 2016), to appear

GPU-based Distributed Sorting [Shamoto, IEEE BigData 2014, IEEE Trans. Big Data 2015]

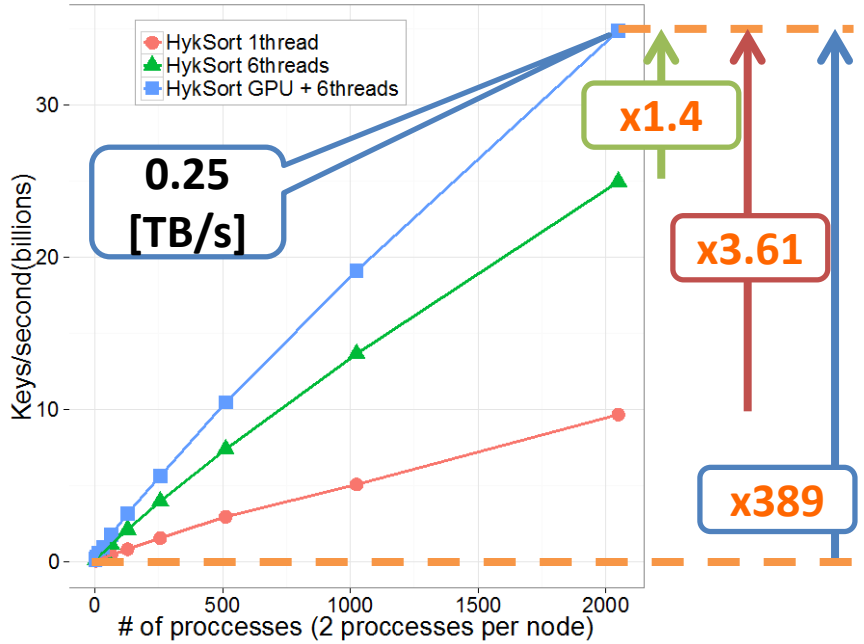
EBD Algorithm Kernels

- Sorting: Kernel algorithm for various EBD processing
- Fast sorting methods
  - Distributed Sorting: Sorting for distributed system
    - Splitter-based parallel sort
    - Radix sort
    - Merge sort
  - Sorting on heterogeneous architectures
    - Many sorting algorithms are accelerated by many cores and high memory bandwidth.
- Sorting for large-scale heterogeneous systems remains unclear

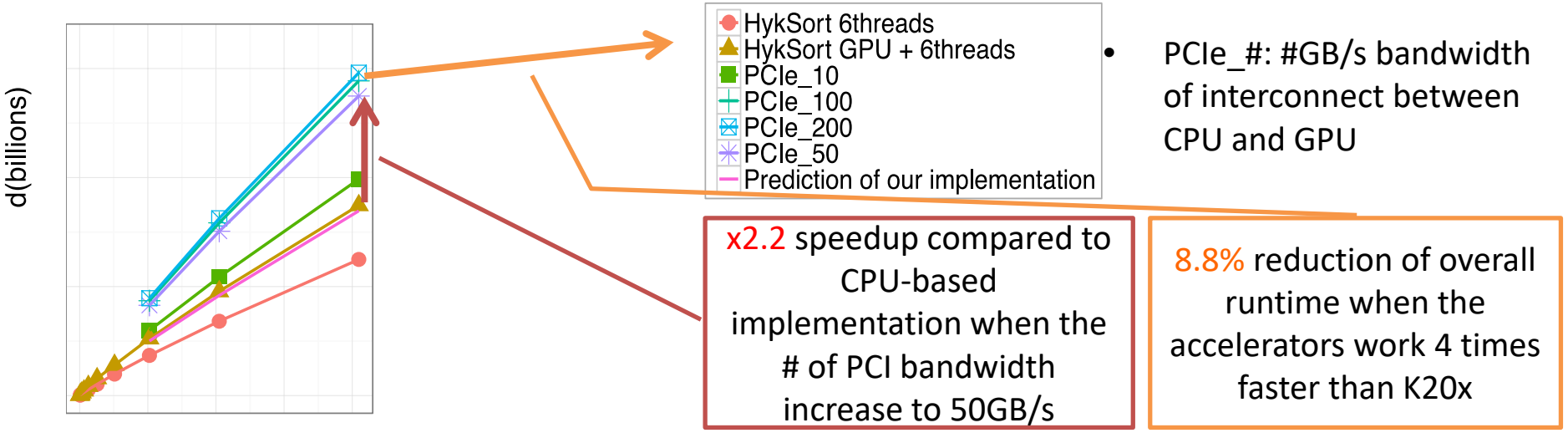
- We develop and evaluate bandwidth and latency reducing GPU-based HykSort on TSUBAME2.5 via latency hiding
  - Now preparing to release the sorting library

GPU implementation of splitter-based sorting (HykSort)

- Weak scaling performance (Grand Challenge on TSUBAME2.5)
  - 1 ~ 1024 nodes (2 ~ 2048 GPUs)
  - 2 processes per node
  - Each node has 2GB 64bit integer
- C.f. Yahoo/Hadoop Terasort: 0.02[TB/s]
  - Including I/O



Performance prediction



- PCIe_#: #GB/s bandwidth of interconnect between CPU and GPU

x2.2 speedup compared to CPU-based implementation when the # of PCI bandwidth increase to 50GB/s

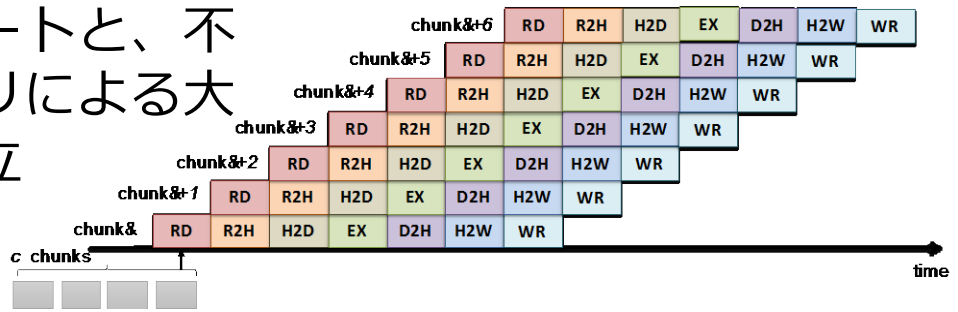
8.8% reduction of overall runtime when the accelerators work 4 times faster than K20x

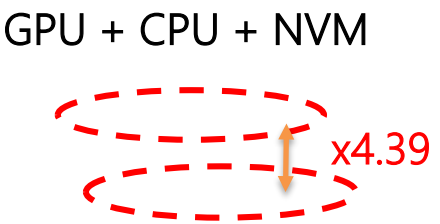
Xtr2sort: Out-of-core Sorting Acceleration using GPU and Flash NVM [IEEE BigData2016]

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





Out-of-core GPU-MapReduce for Large-scale Graph Processing [IEEE Cluster 2014]

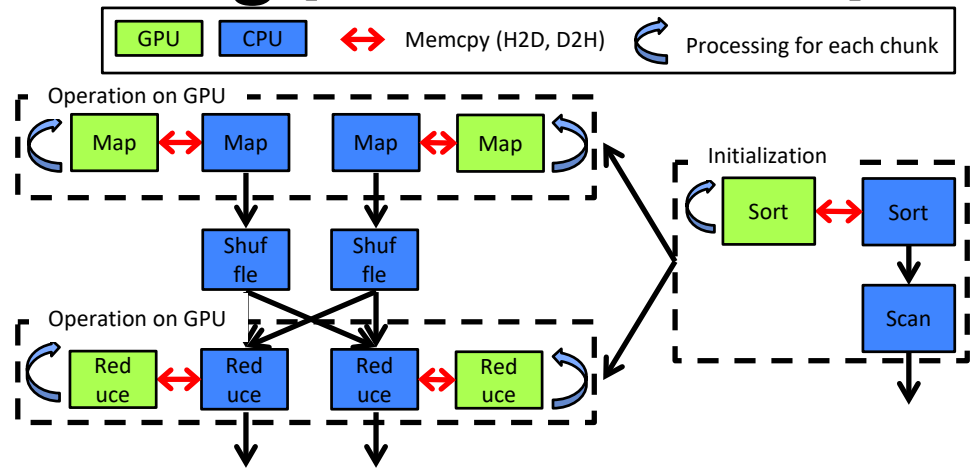
Emergence of large-scale graphs
- SNS, road network, smart grid, etc.
- Millions to trillions of vertices/edges

→ Need for fast graph processing on supercomputers

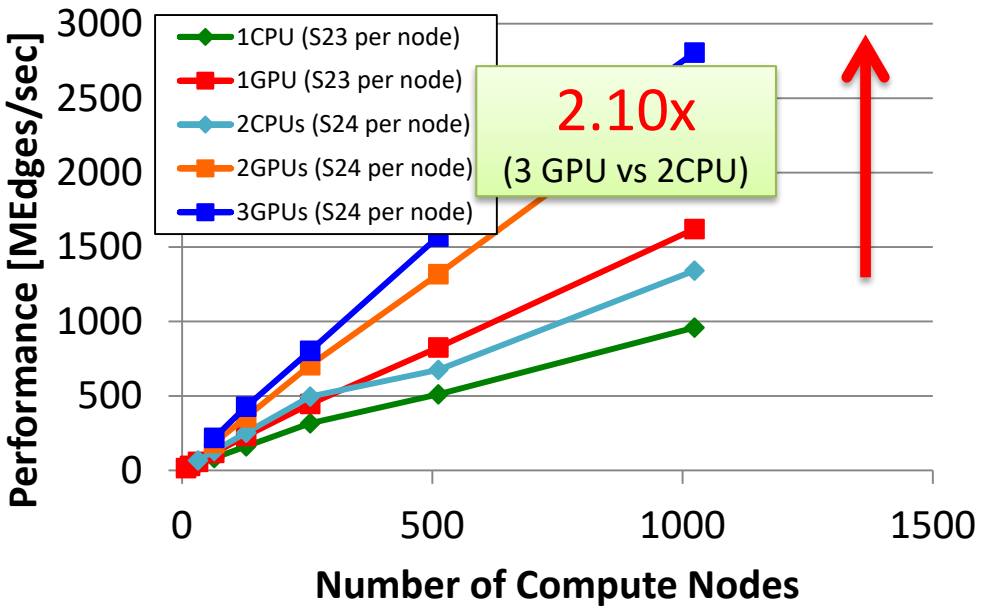
Problem: GPU memory capacity limits scalable large-scale graph processing

Proposal: Out-of-core GPU memory management on MapReduce
- Stream-based GPU MapReduce
- Out-of-core GPU sorting

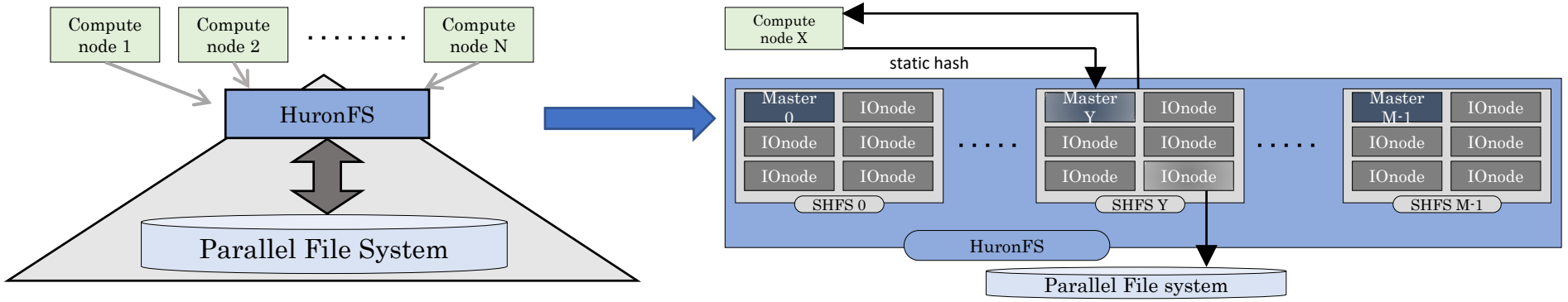
Experimental Results:
performance improvement over CPUs
- Map: 1.41x, Reduce: 1.49x, Sort: 4.95x speedup
- Overlapping communication effectively



Weak scaling on TSUBAME2.5

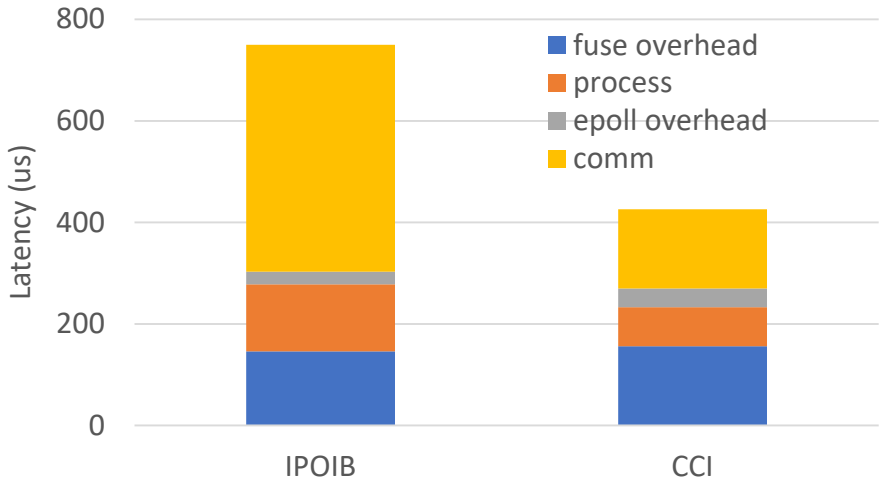


Hierarchical, UseR-level and ON-demand File system(HuronFS) (IEEE ICPADS 2016) w/LLNL

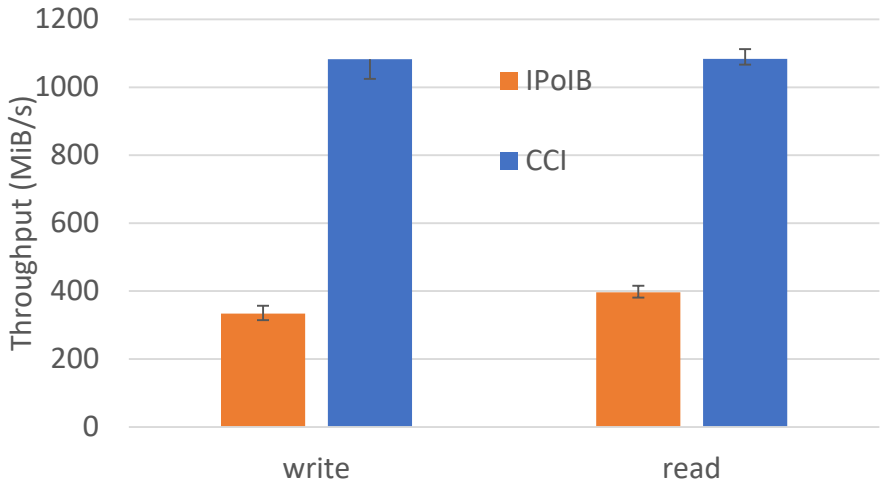


- HuronFS*:* dedicated dynamic instances to provide “burst buffer” for caching data
- I/O requests from *Compute Nodes* are forwarded to HuronFS
- The whole system consists of several SHFS (Sub HuronFS)
  - Workload are distributed among all the SHFS using hash of file path
- Each SHFS consists of a Master and several IOnodes
  - Masters: controlling all IOnodes in the same SHFS and handling all I/O requests
  - IOnodes: storing actual data and transferring data with Compute Nodes
- Supporting TCP/IP, Infiniband (CCI framework)
- Supporting Fuse, LD_PRELOAD

HuronFS Basic IO Performance

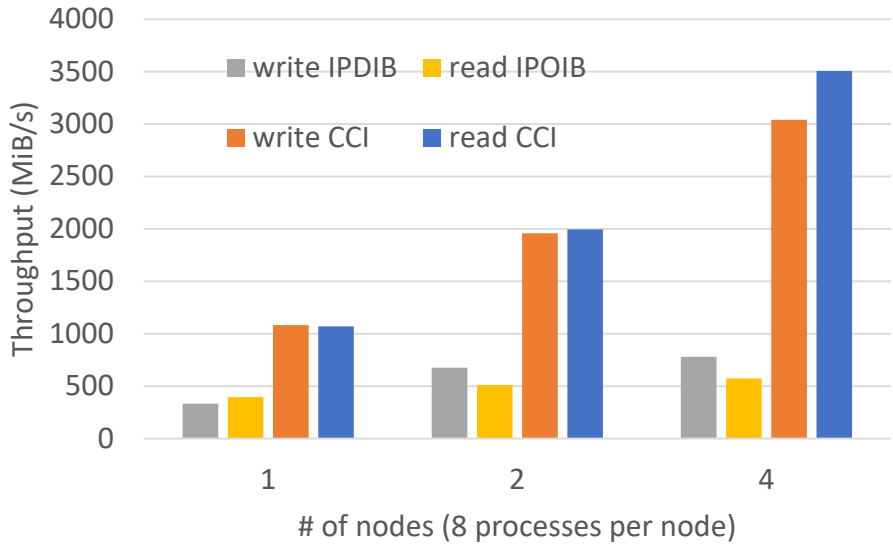


Latency



Throughput from single client

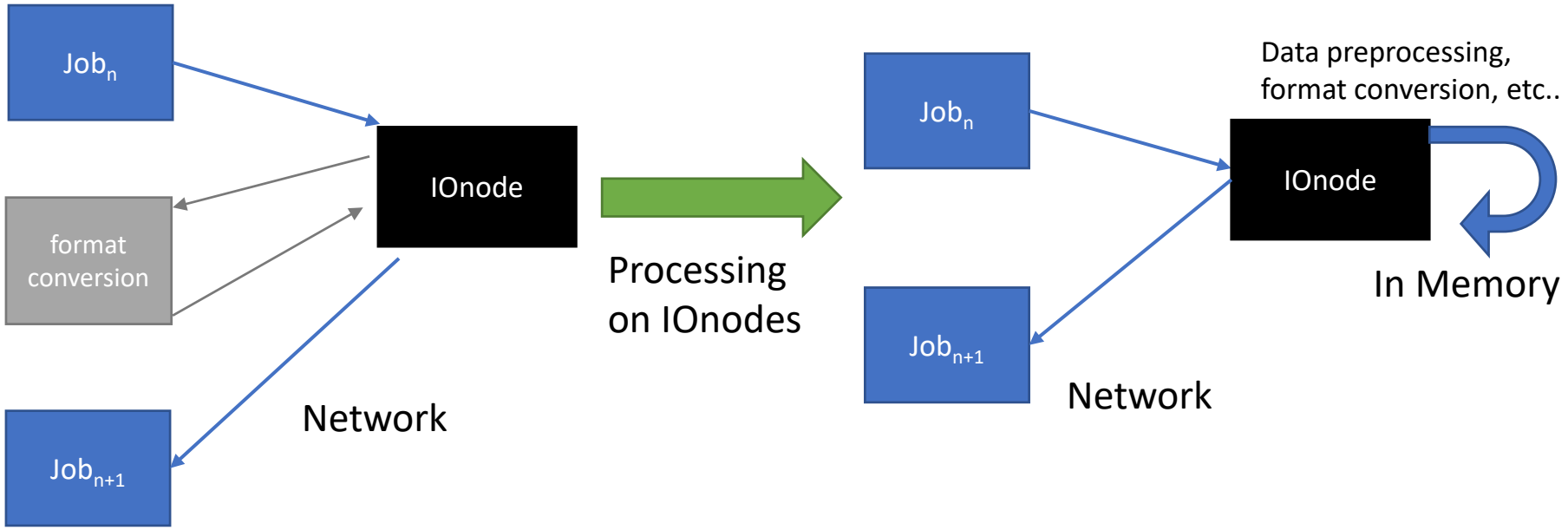
Inifinband	4X FDR 56 Gb/sec mellanox
CPU	Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz
Mem	251G



Throughput from single IOnode

Plans

- Continuing researching on auto buffer allocation
- Utilizing computation power on IOnodes
  - Data preprocessing
  - Format conversion



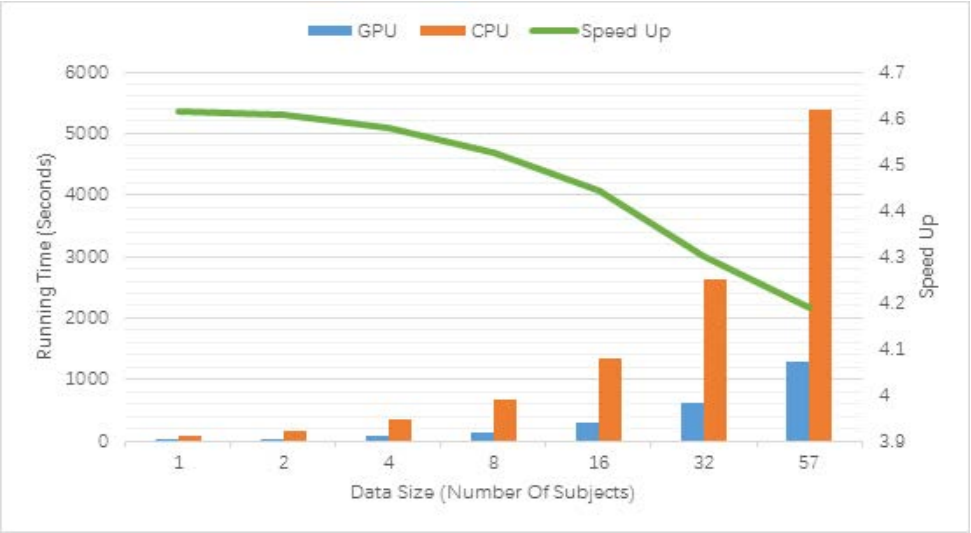
GPU-Based Fast Signal Processing for Large Amounts of Snore Sound Data

- Background

Snore sound (SnS) data carry very important information for diagnosis and evaluation of Primary Snoring and Obstructive Sleep Apnea (OSA). With the increasing number of collected SnS data from subjects, how to handle such large amount of data is a big challenge. In this study, we utilize the Graphics Processing Unit (GPU) to process a large amount of SnS data collected from two hospitals in China and Germany to accelerate the features extraction of biomedical signal.

- Acoustic features of SnS data

we extract **11** acoustic features from a large amount of SnS data, which can be visualized to help doctors and specialists to diagnose, research, and remedy the diseases efficiently.



Results of GPU and CPU based systems for processing SnS data

Snore sound data information

Subjects	Total Time (hours)	Data Size (GB)	Data format	Sampling Rate
57 (China + Germany)	187.75	31.10	WAV	16 kHz, Mono

- Result

We set 1 CPU (with Python2.7, numpy 1.10.4 and scipy 0.17 packages) for processing 1 subject's data as our baseline. Result show that the GPU based system is almost 4.6×faster than the CPU implementation. However, the speed-up decreases when increasing the data size. We think that this result should be caused by the fact that, the transmission of data is not hidden by other computations, as will be a real-world application.

* Jian Guo, Kun Qian, Huijie Xu, Christoph Janott, Bjorn Schuller, Satoshi Matsuoka, "GPU-Based Fast Signal Processing for Large Amounts of Snore Sound Data", *In proceedings of 5th IEEE Global Conference on Consumer Electronics (GCCE 2016)*, October 11-14, 2016.

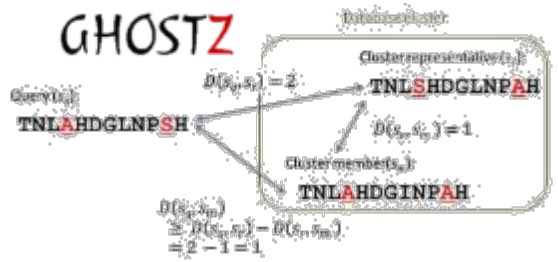
Open Source Release of EBD System Software (install on T3/Amazon/ABCI)

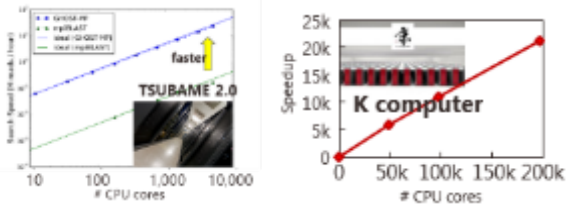
- mrCUDA
  - rCUDA extension enabling remote-to-local GPU migration
  - https://github.com/EBD-CREST/mrCUDA
  - GPU 3.0
  - Co-Funded by NVIDIA
- Huron FS (w/LLNL)
  - I/O Burst Buffer for Inter Cloud Environment
  - https://github.com/EBD-CREST/cbb
  - Apache License 2.0
  - Co-funded by Amazon
- ScaleGraph 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

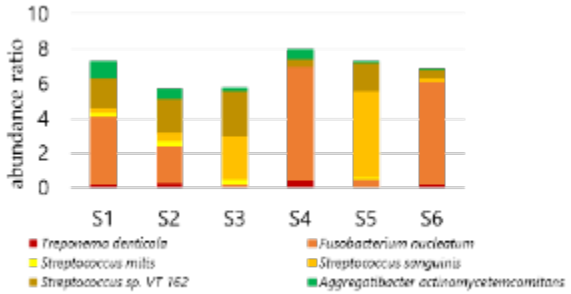
Ultra-fast Seq. Analysis





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

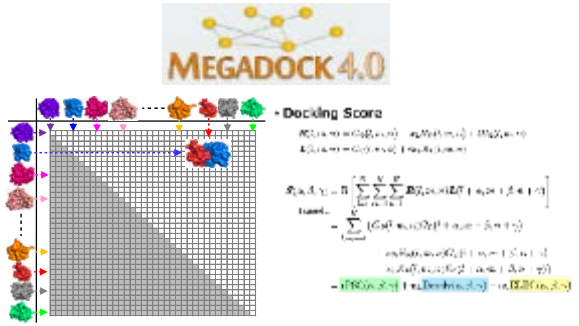
Oral/Gut Metagenomics



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

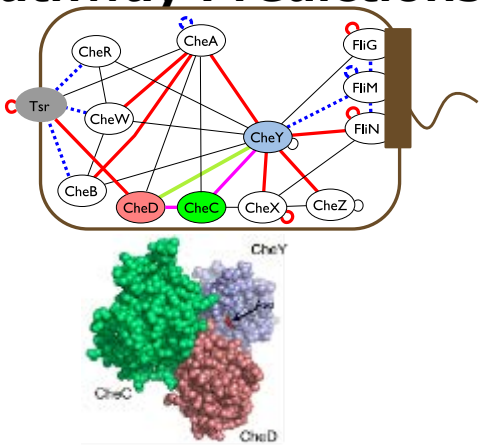
Protein-Protein Interactions

Exhaustive PPI Prediction System



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

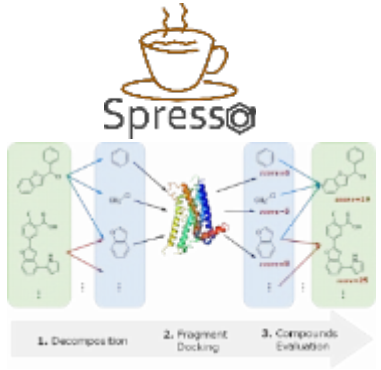
Pathway Predictions



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

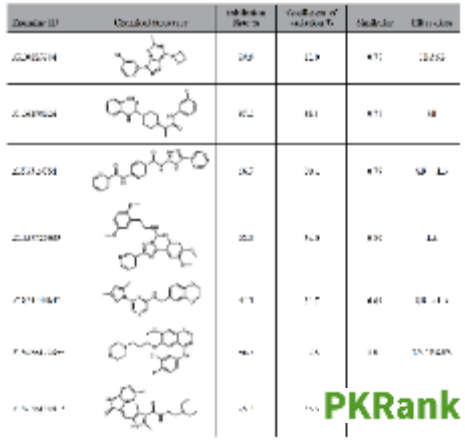
Drug Discovery

Fragment-based Virtual Screening



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

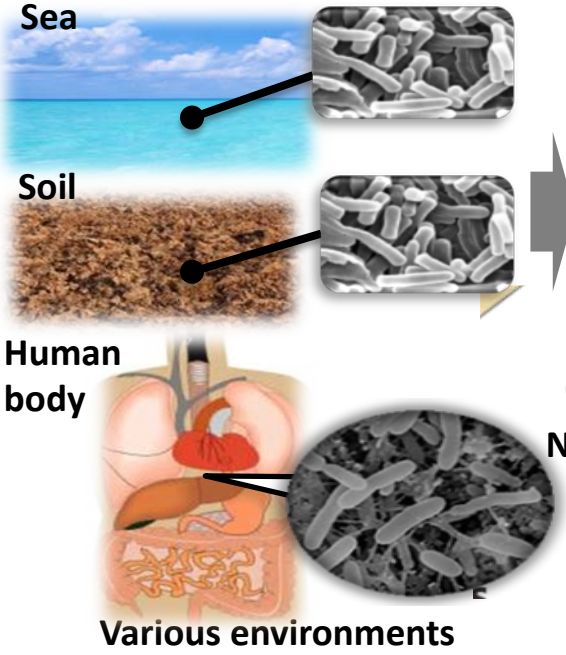
Learning-to-Rank VS

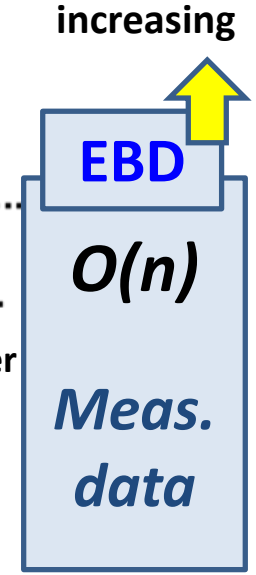


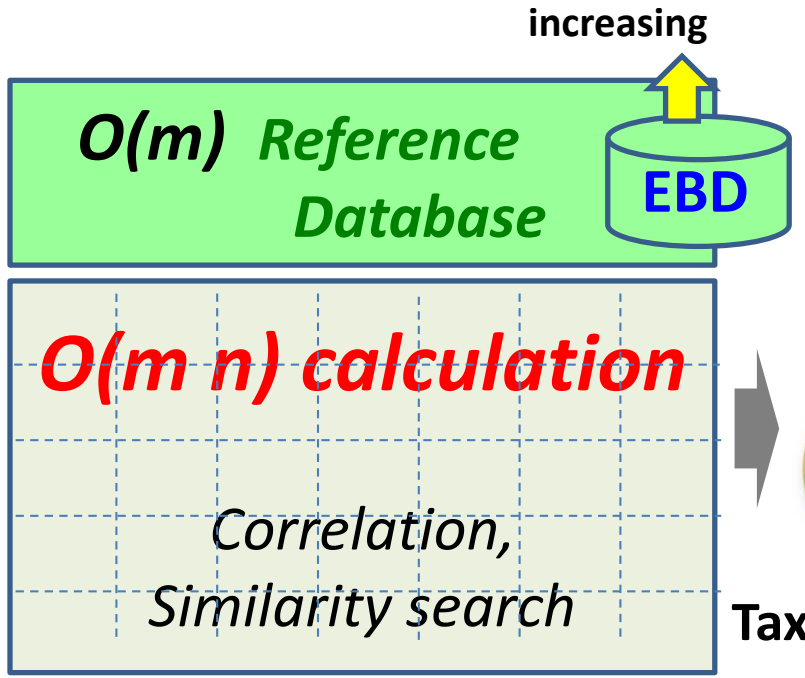
- Suzuki *et al.*, *AROB2017* (2017)

EBD vs. EBD : Large Scale Homology Search for Metagenomics

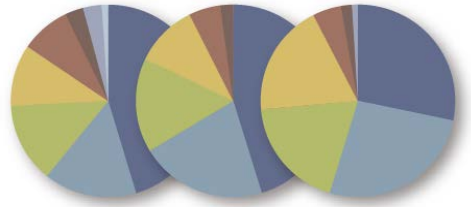
- Revealing uncultured microbiomes and finding novel genes in various environments
- Applied for human health in recent years







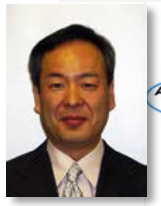

Metabolic Pathway

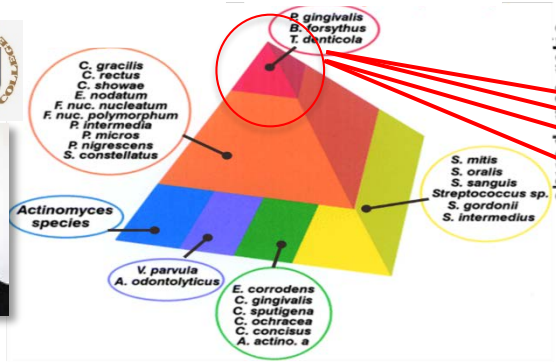

Taxonomic composition

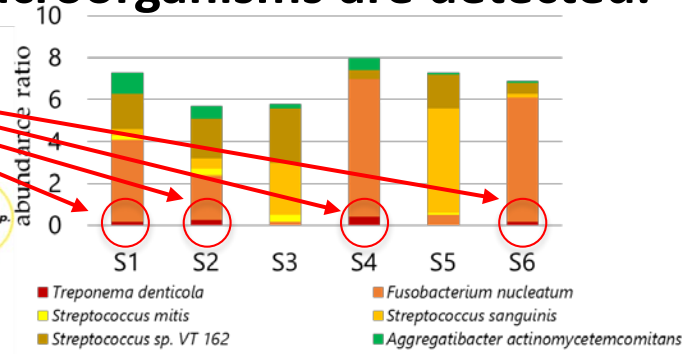
High risk microorganisms are detected.

Metagenomic analysis of periodontitis patients

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





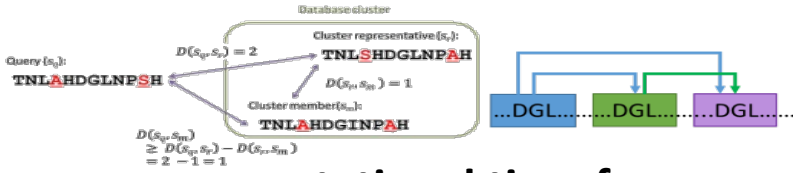


Development of Ultra-fast Homology Search Tools

GHOSTZ

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

Subsequence sequence clustering



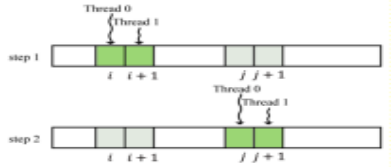
computational time for 10,000 sequences (sec.) (3.9 GB DB、1CPU core)

× 240 faster than conventional algorithm

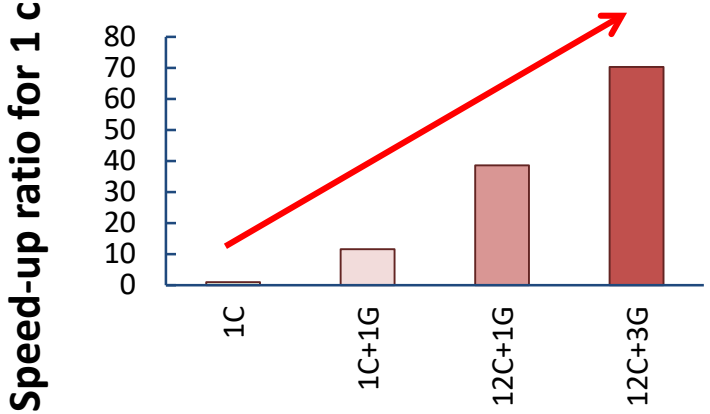
GHOSTZ-GPU

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

Multithread on GPU



TSUBAME 2.5 Thin node GPU



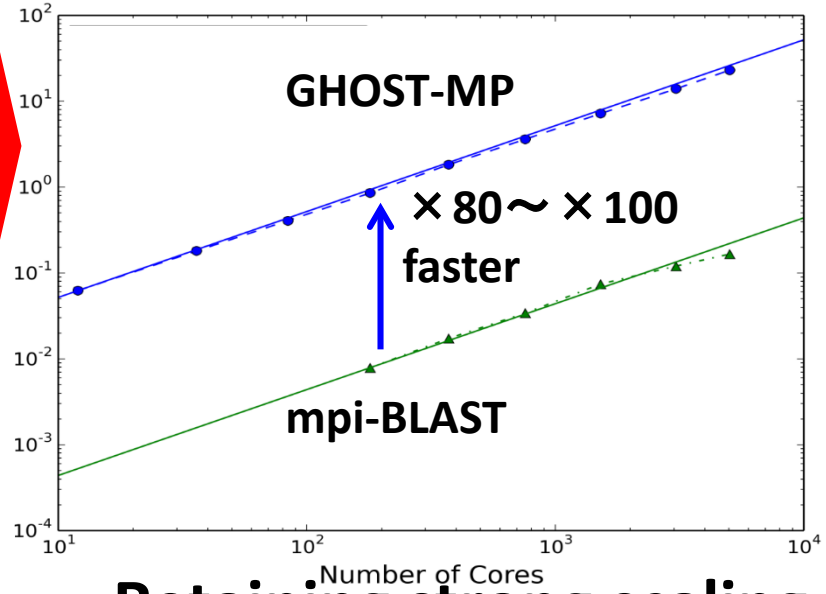
× 70 faster than 1 core using 12 cores + 3 GPUs

GHOST-MP

Kakuta, *et al*. (submitted)

MPI + OpenMP hybrid pallelization

TSUBAME 2.5



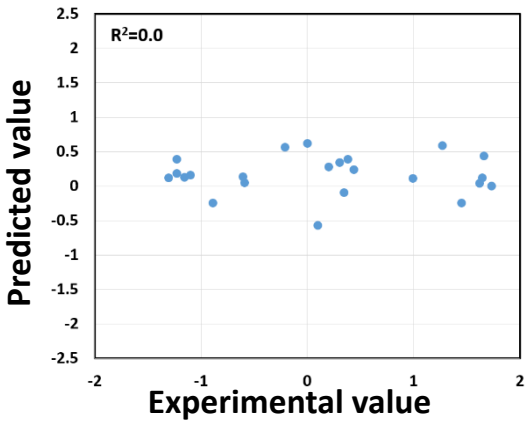
Retaining strong scaling up to 100,000 cores

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

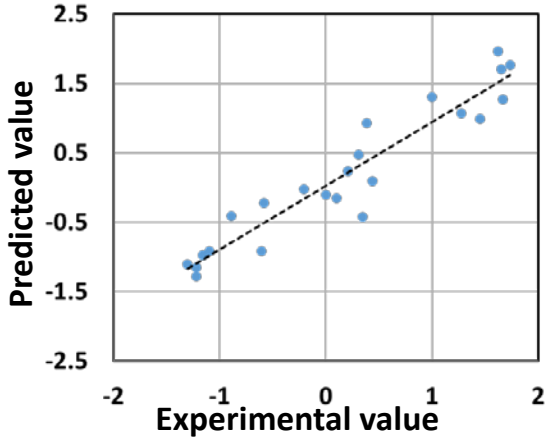
Compute Feature Values
(more than 500 features)

LogS
LogP
:
MolWeight
:
SASA
polarity

f

PPB value

Combining feature values for building a predictive model



$R^2 = 0.905$

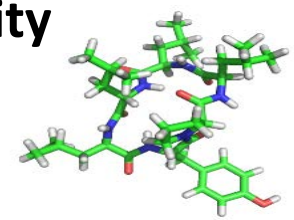
A constructed model can explain peptide PPB well

Molecular Dynamics Simulation for Membrane Permeability
Application for peptide drug discovery

Problems

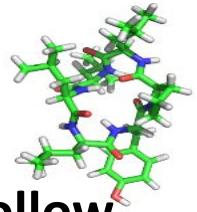
1) Single residue mutation can drastically change membrane permeability

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

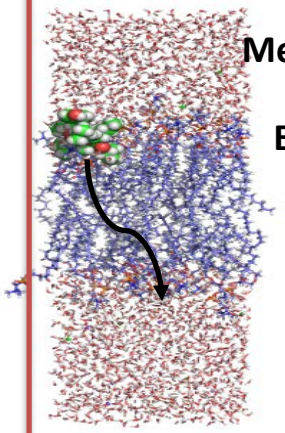


× 0.006

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.

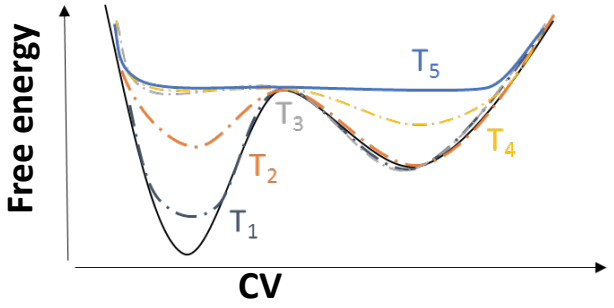
Ex) Membrane thickness : 40 Å
Peptide membrane permeability : 7.9 × 10^{-6} cm/s

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)

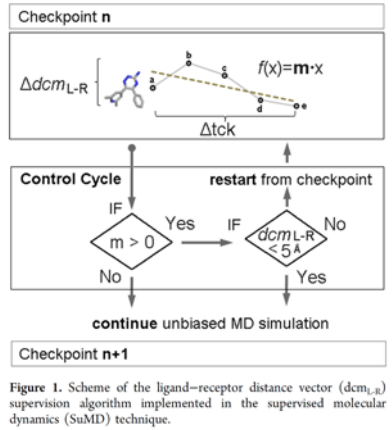
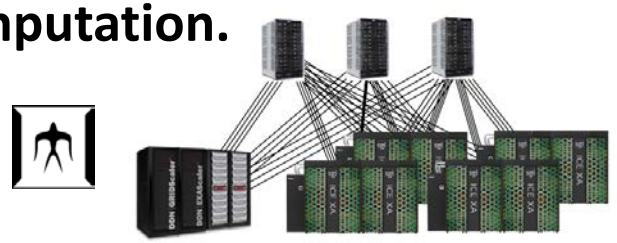


Figure 1. Scheme of the ligand–receptor distance vector (dcm_{L-R}) supervision algorithm implemented in the supervised molecular dynamics (SuMD) technique.

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.**

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)

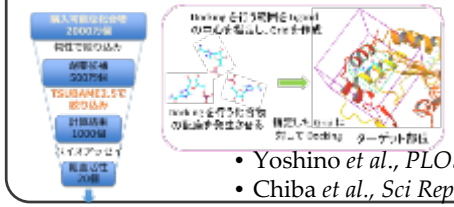
Tokyo Tech's research seeds

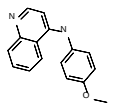
①Drug Target selection system

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**

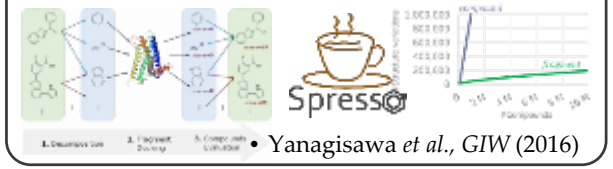




- Yoshino *et al.*, *PLOS ONE* (2015)
- Chiba *et al.*, *Sci Rep* (2015)

③Novel Algorithms for fast virtual screening against huge databases

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

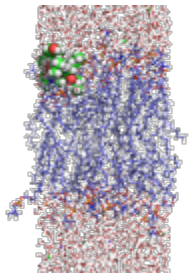


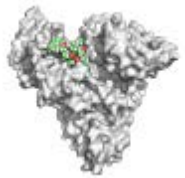
- Yanagisawa *et al.*, *GIW* (2016)

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

DiscoveReSource 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

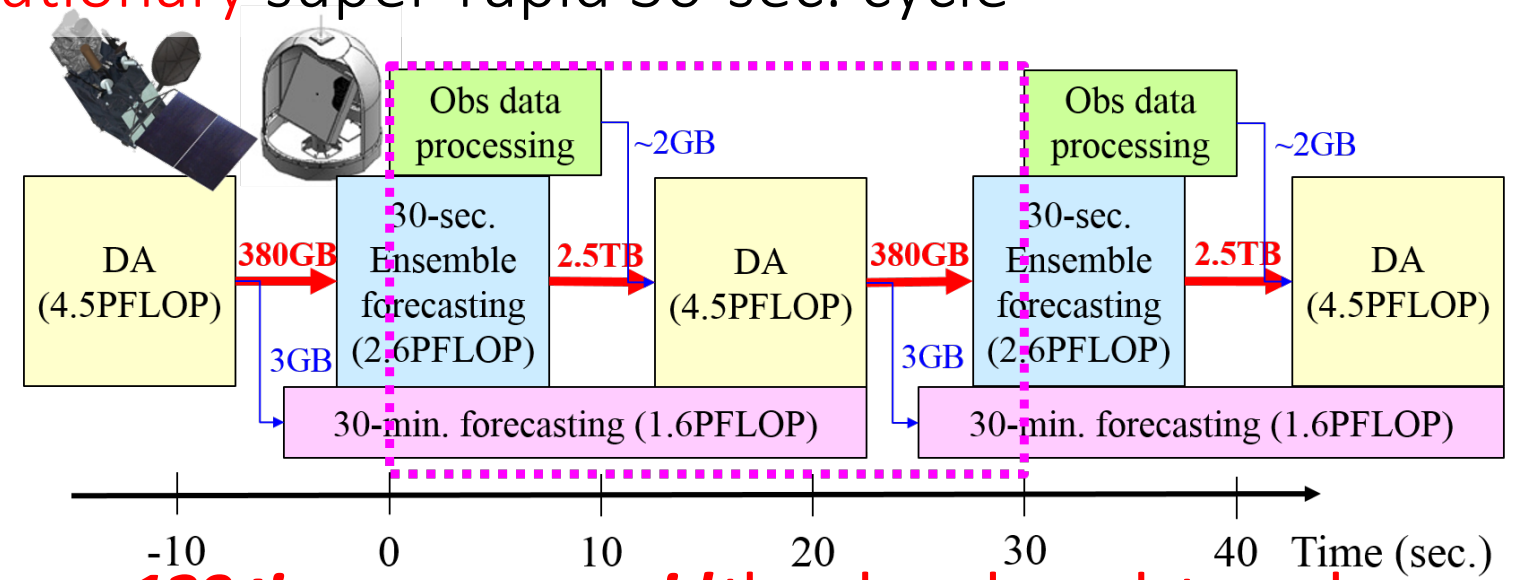
Investments from JP Govt., Tokyo Tech. (TSUBAME SC) Muninciple Govt (Kawasaki), JP & US Pharma

EBD App2: Miyoshi Group (Weather Forecast Application)

Big Data Assimilation
for severe weather forecast

Goal：Pinpoint (100-m resol.) forecast of severe local weather by updating 30-min forecast every 30 sec!

Revolutionary super-rapid 30-sec. cycle



120 times more rapid than hourly update cycles

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

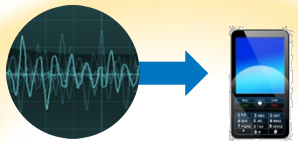
Image/Video Recognition

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

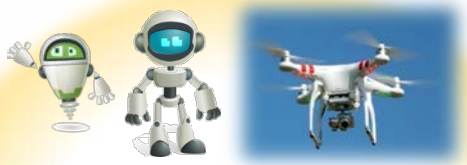


10P～ Flops
1万人の5000時間分の音声データ
人工的に生成された10万時間の
音声データを基に学習 [Baidu 2015]

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



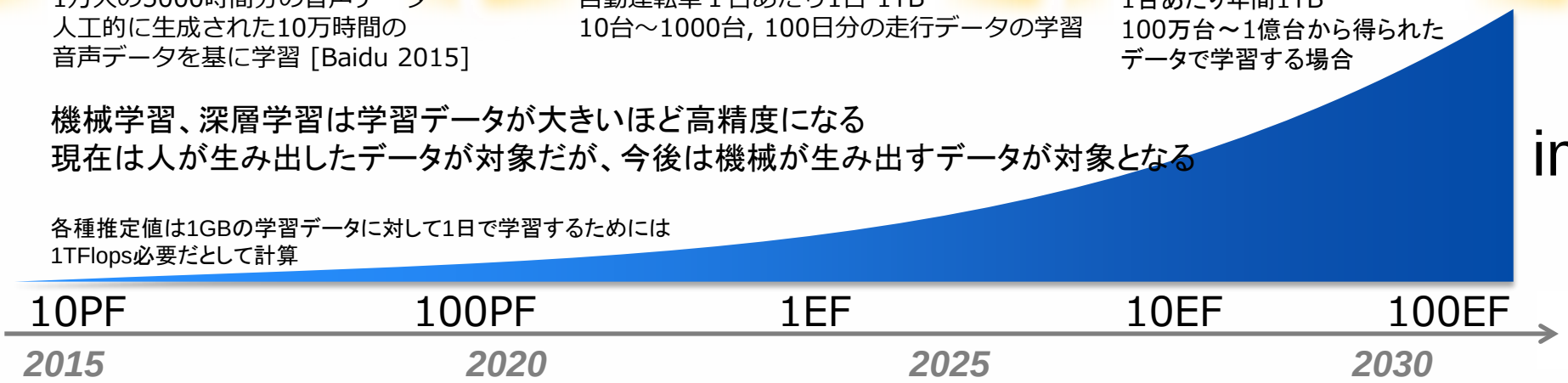
1E～100E Flops
1台あたり年間1TB
100万台～1億台から得られた
データで学習する場合

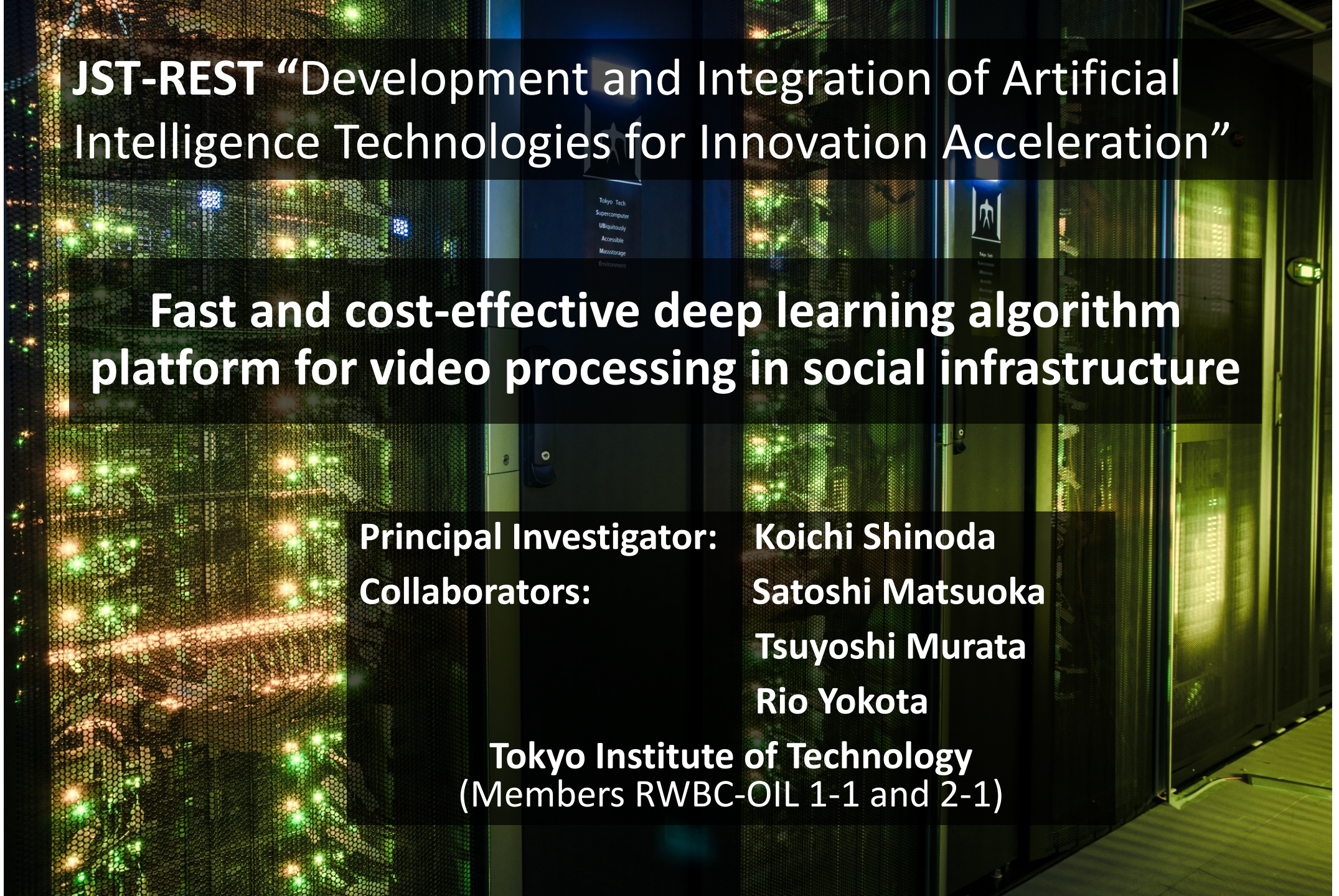
It’s the FLOPS (in reduced precision) and BW!

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





JST-REST “Development and Integration of Artificial Intelligence Technologies for Innovation Acceleration”

Fast and cost-effective deep learning algorithm platform for video processing in social infrastructure

Principal Investigator: **Koichi Shinoda**

Collaborators: **Satoshi Matsuoka**

Tsuyoshi Murata

Rio Yokota

Tokyo Institute of Technology

(Members RWBC-OIL 1-1 and 2-1)

Background

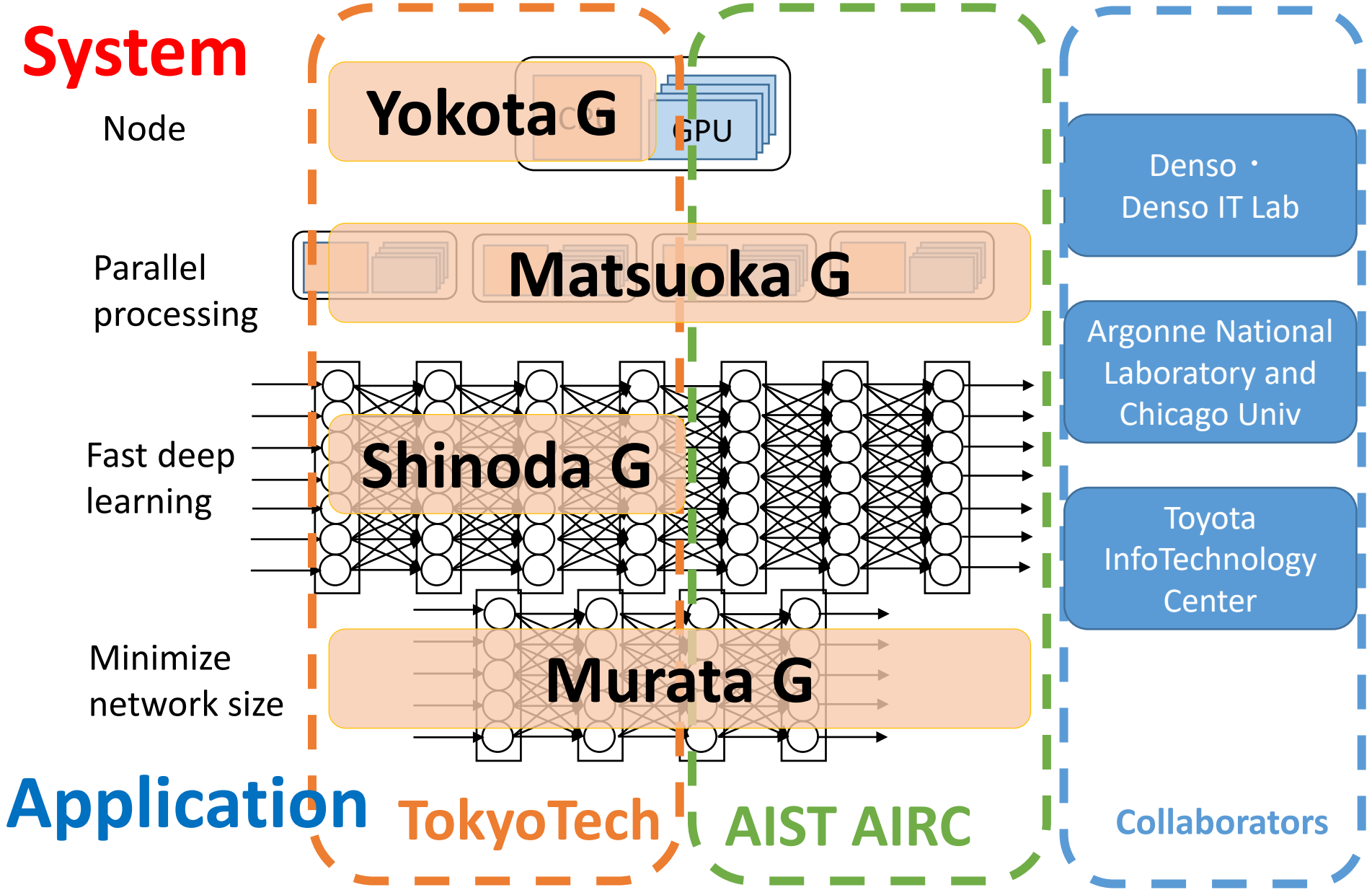
- Video processing in smart society for safety and security
  - Intelligent transport systems
    Drive recorder video
  - Security systems
    Surveillance camera video
- Deep learning
  - Much higher performance than before
  - IT giants with large computational resources has formed a monopoly

Problems：

- Real-time accurate recognition of small objects and their movement
- Edge-computing without heavy traffic on Internet
- Flexible framework for training which can adapt rapidly to the environmental changes

Research team

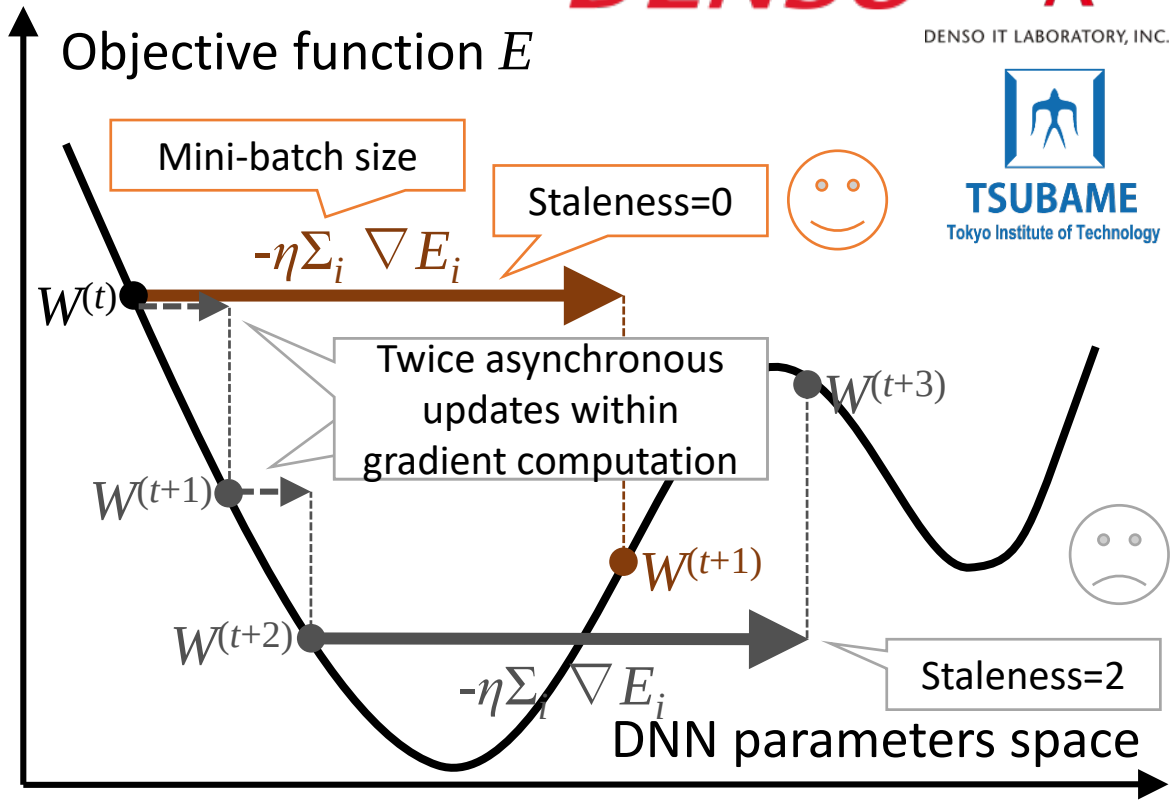
Reference



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

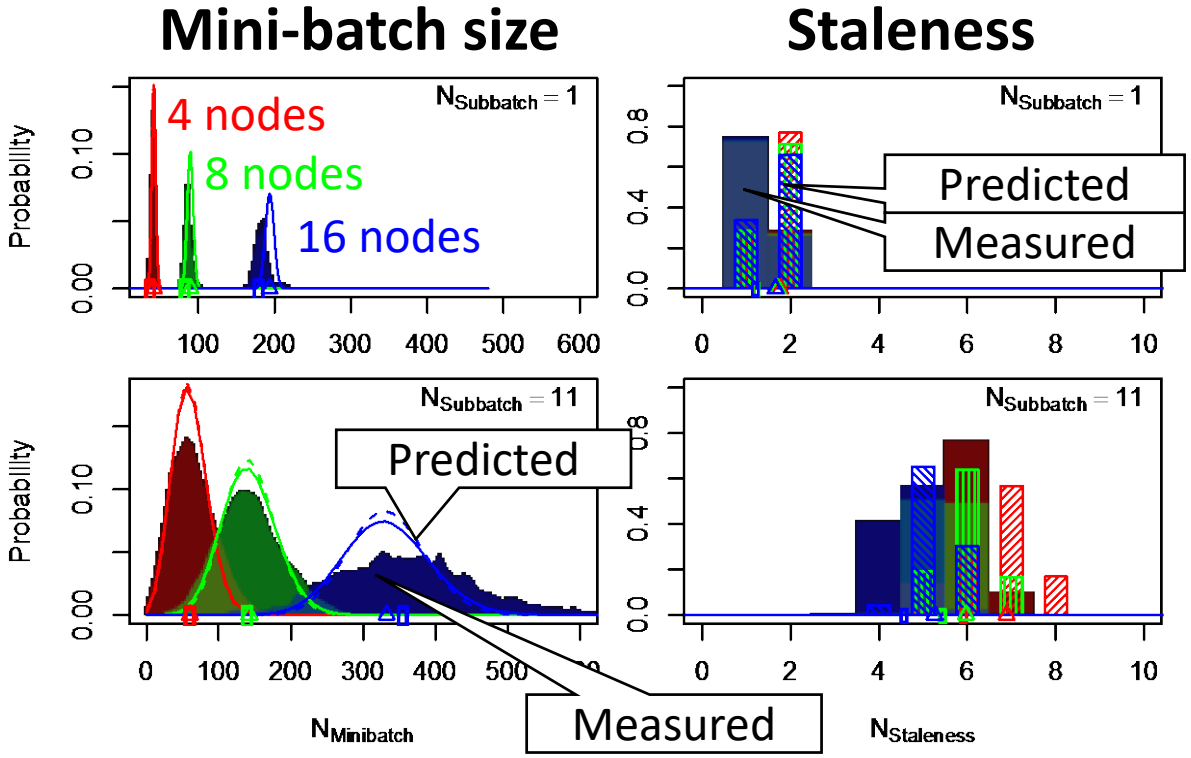
Background

- 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)

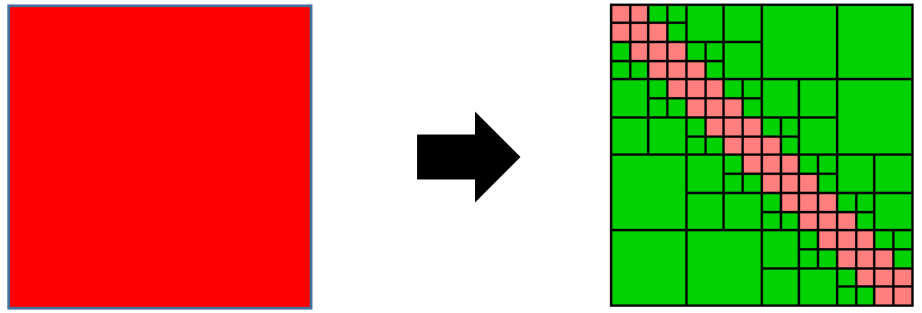
→ Not only # of nodes, but also fast interconnect is important for scalability

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

Hierarchical matrix(H-matrix) for CNN acceleration

- Hierarchical matrix is an efficient data-sparse representations of certain densely populated matrices.
- CNN(Convolutional Neural Network)



dense matrix　　Hierarchical matrix

The H-matrix approximation of dense matrix.
The red blocks are dense matrices. The green block are low-rank matrices with rank k.

- Back ground

  - Hierarchical matrix(H-matrix) is a an approximated form represent $n \times n$ correlations of n objects, which usually requires a $n \times n$ huge dense matrix.

  - Significant savings in memory when compressed

$$O(n^2) \Longrightarrow O(kn \log n)$$

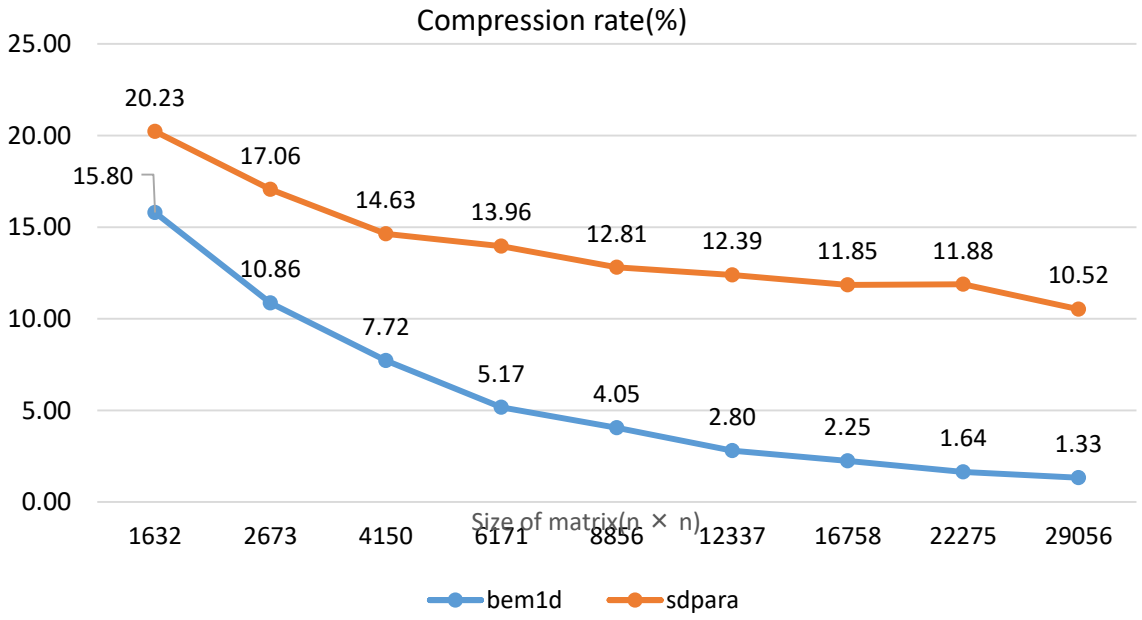
  - Computational complexity

$$O(n^3) \Longrightarrow O(k^2 n \log n^2)$$

  such as matrix-matrix multiplication, LU factorization, Inversion…

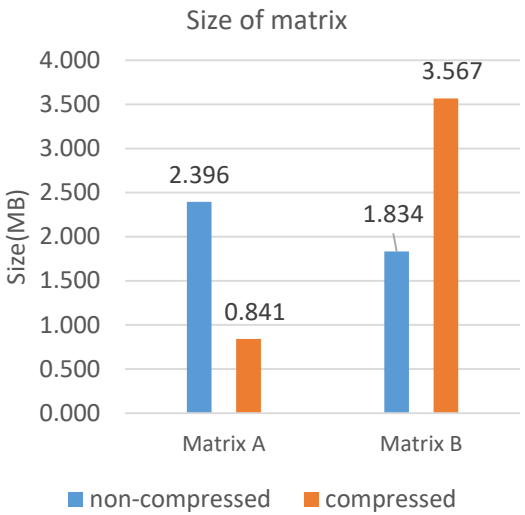
Preliminary Results – Compression rate of matrices

SDPARA



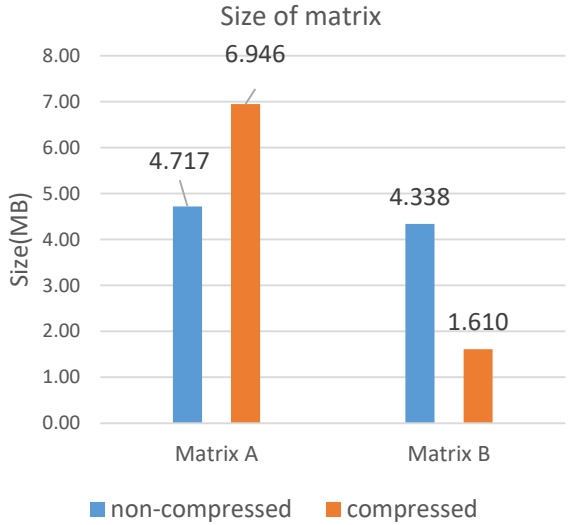
Compressive rate = (uncompressed size) / (compressed size)

Deep Learning (CNN)



(m, n, k) = (1764, 1350, 178)

→ Matrix A successfully compressed!



(m, n, k) = (324, 298, 1908)

→ Matrix B successfully compressed!

We can compress the matrix in some applications.

- bem1d: 1-Dimention Boundary element method
- sdpara: A parallel implementation of the inter-point method for Semi-Define Programming(SDP)

In CNN system application, Sgemm(Single precision floating General Matrix Multiplication) $C = \alpha AB + \beta C$ accounts for large part of calculation (around 70%).

Power optimization using Deep Q-Network

Kento Teranishi

- Background

Power optimization by frequency control in existing research

Performance counter
Temperature
Frequency,…

$$P = f(x_1, x_2, \ldots)$$
$$T_{exe} = g(x_1, x_2, \ldots)$$

→ Frequency

- Detailed analysis is necessary
- Low versatility

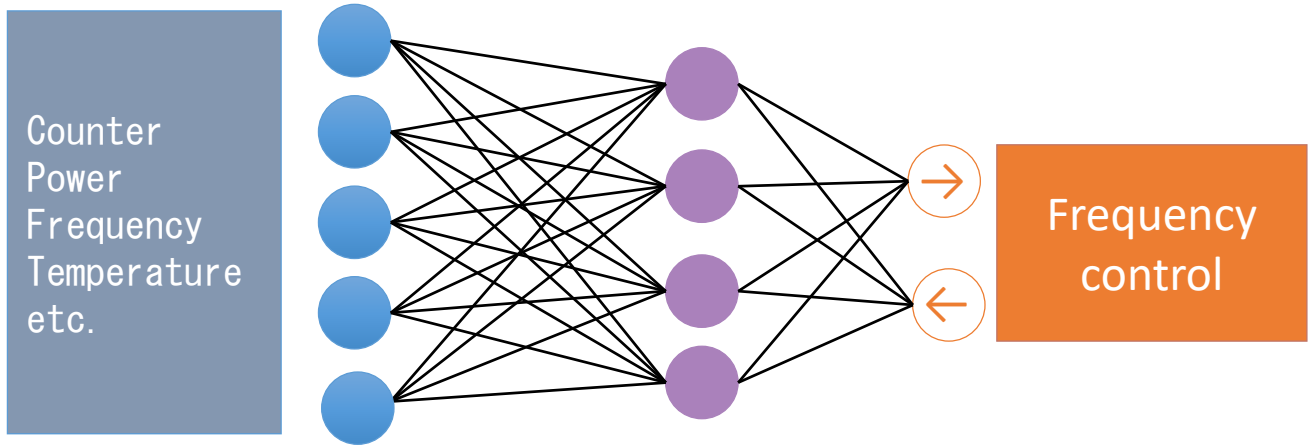
→ Use Deep Learning for analysis.

- Objective

Implement the computer control system using Deep Q-Network.

Deep Q-Network (DQN)

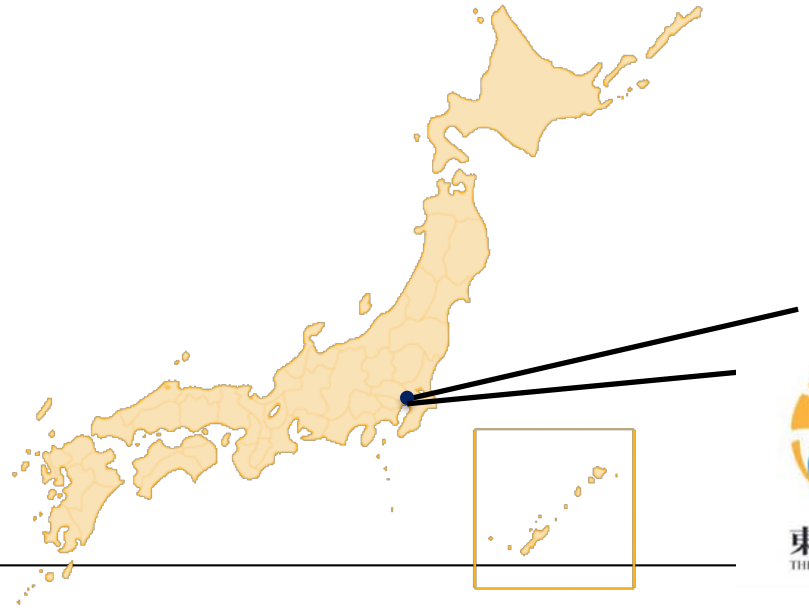
Deep reinforcement learning
Calculate action value function Q from neural network
Used for game playing AI, robot car, AlphaGO.



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



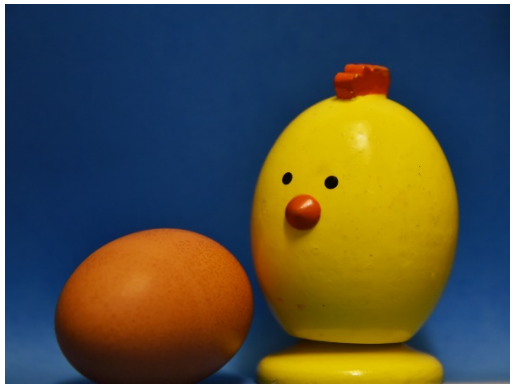
- 130~200 AI-Petaflops
- < 3MW Power
- < 1.1 Avg. PUE
- Operational 2017Q4 ~2018Q1

東京大学 THE UNIVERSITY OF TOKYO

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

Univ. Tokyo Kashiwa Campus

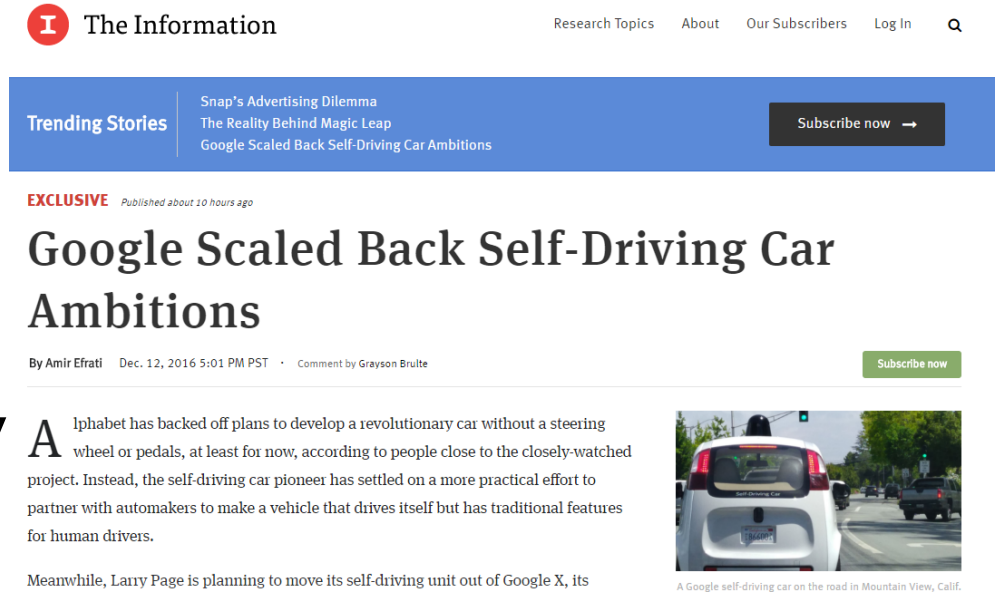
The “Chicken or Egg Problem” of AI-HPC Infrastructures



- “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

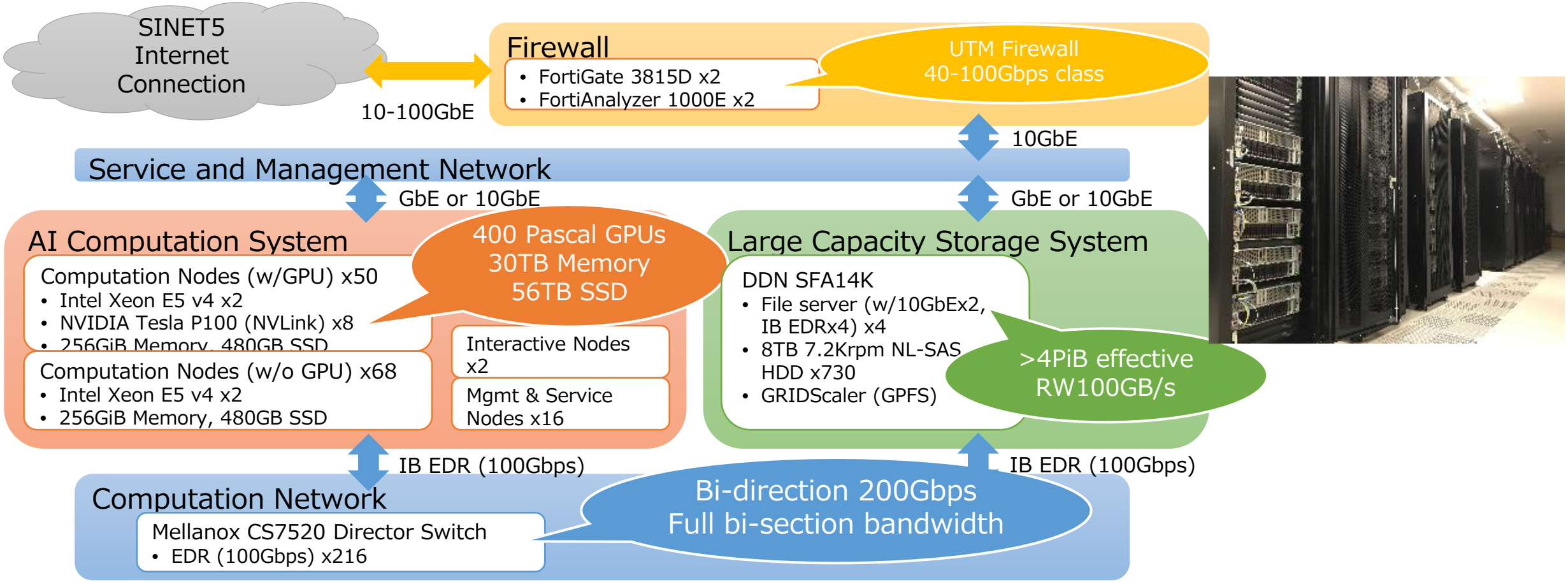
Alphabet has backed off plans to develop a revolutionary car without a steering wheel or pedals, at least for now, according to people close to the closely-watched project. Instead, the self-driving car pioneer has settled on a more practical effort to partner with automakers to make a vehicle that drives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

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 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 BW & Low latency 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.**
- **Ultra-Green (PUE<1.1), High Thermal (60KW) Rack**
  - **Custom, warehouse-like IDC building and internal pods**
  - **Final "commoditization" of HPC technologies into Clouds**

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

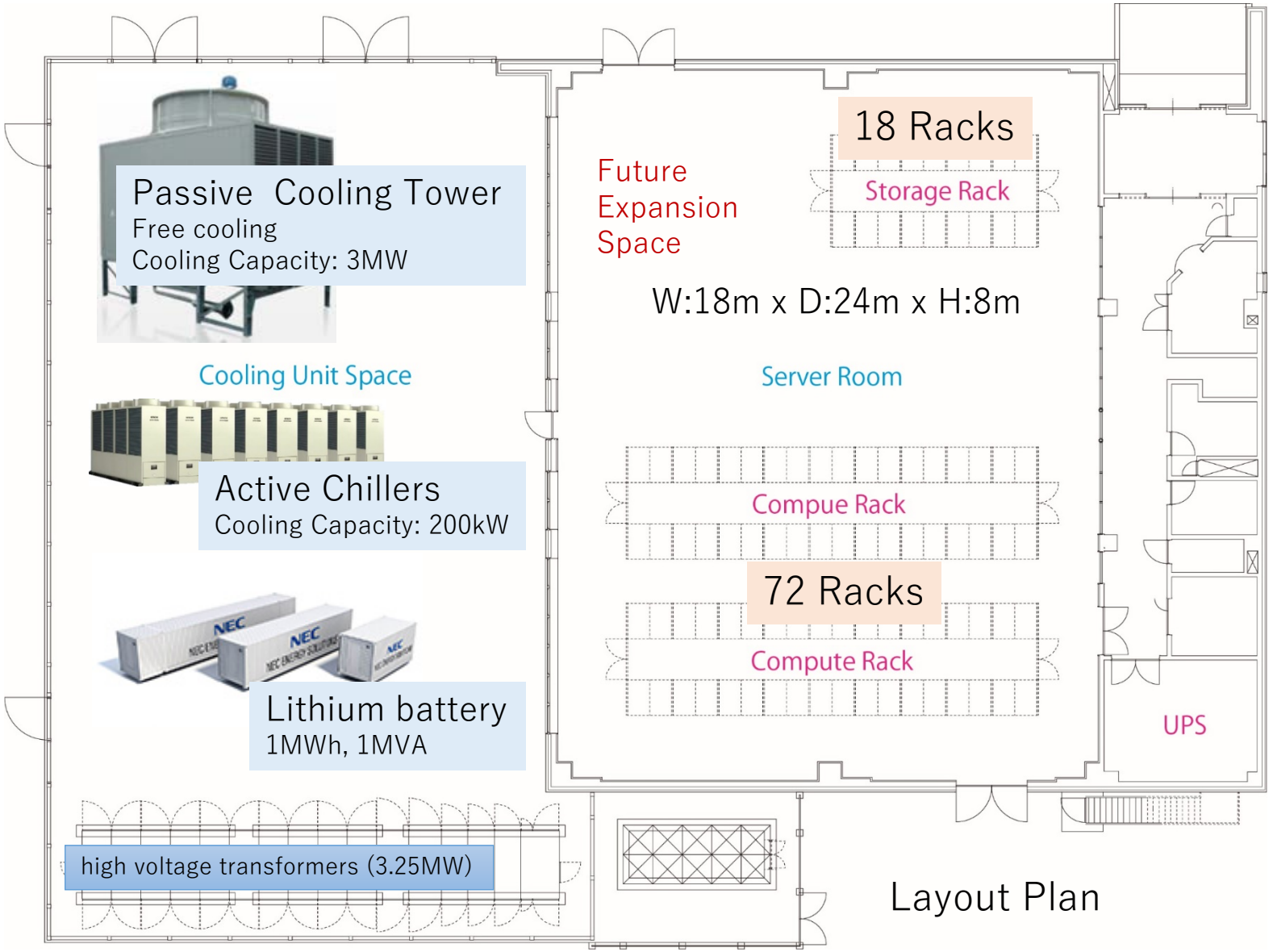
イメージスケッ

Reference Image

引用元: NEC導入事例

ABCI Cloud Data Center

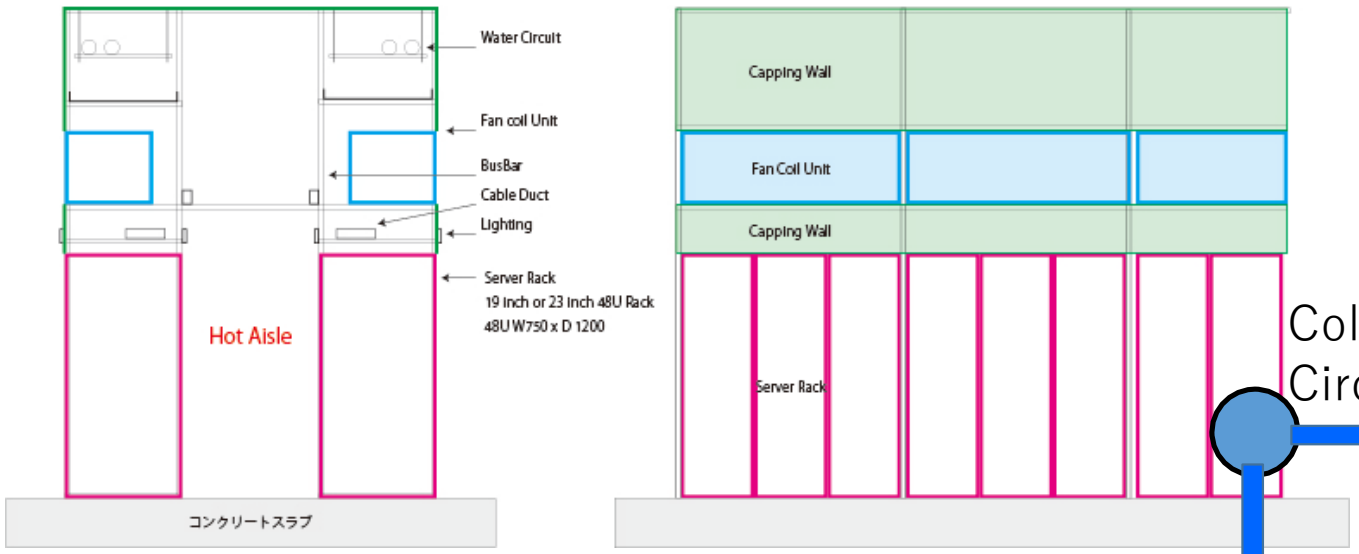
“Commoditizing 60KW/rack Supercomputer”



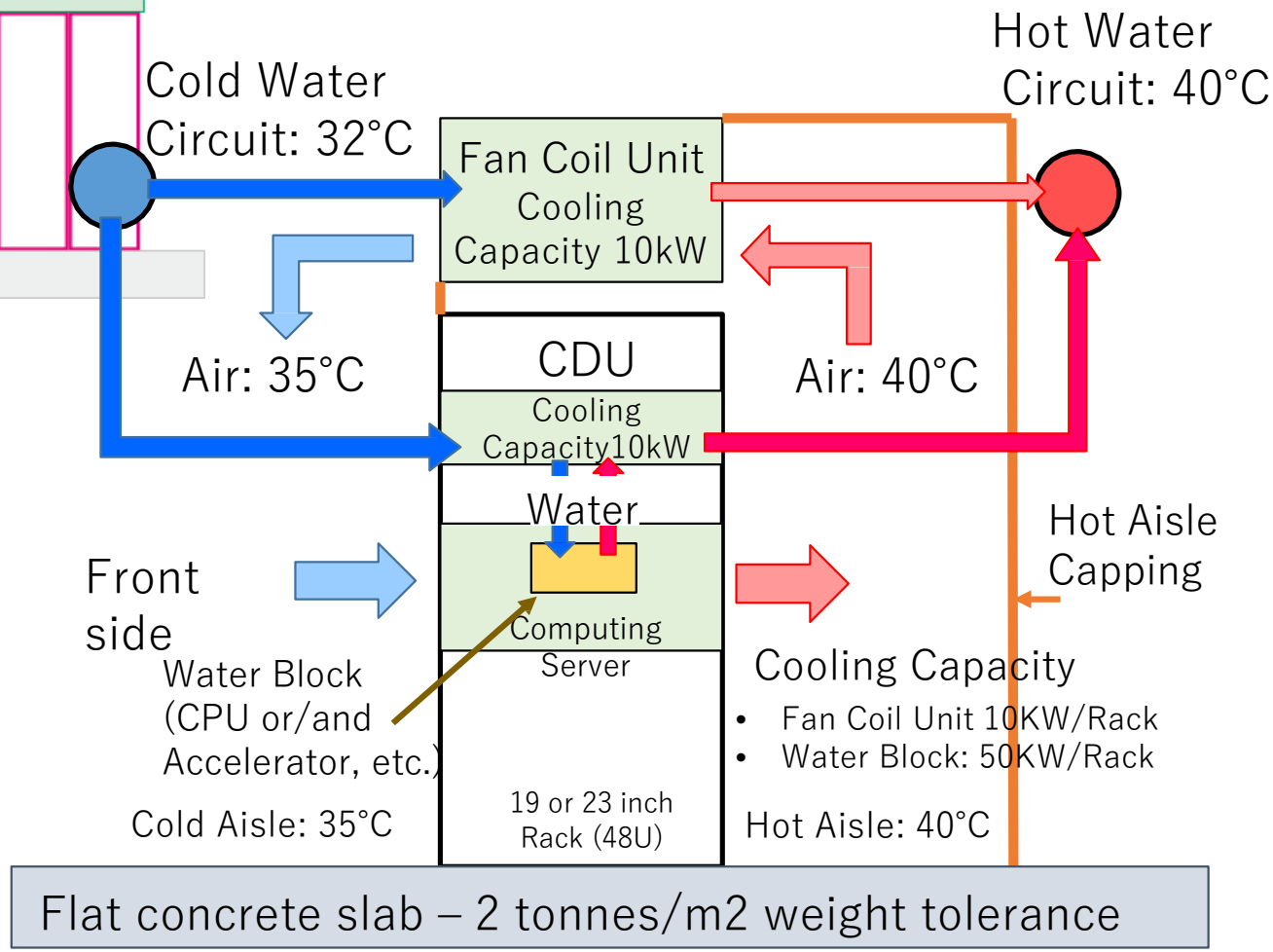
Layout Plan

- 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

ABCI Procurement Benchmarks

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

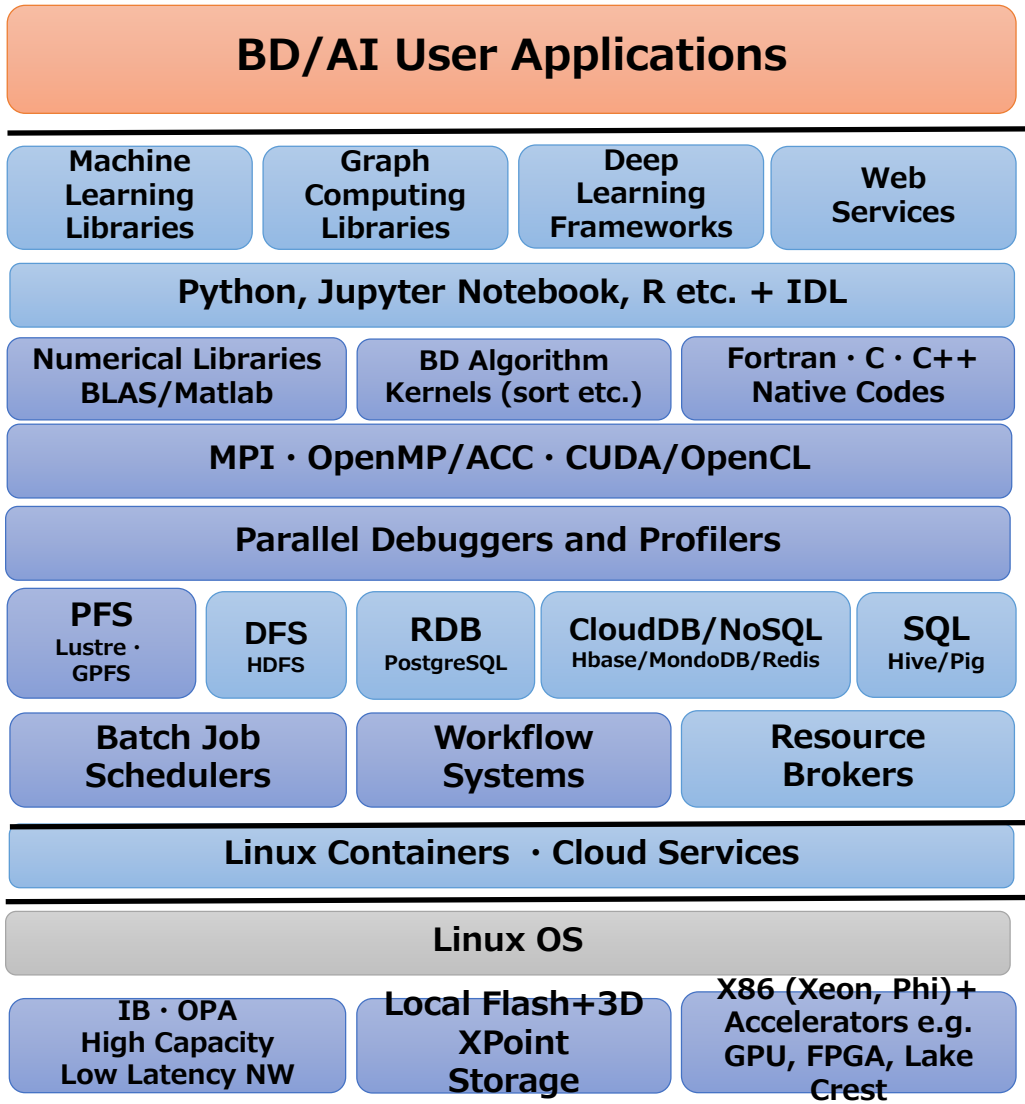
No traditional HPC Simulation Benchmarks Except SPECCPU

- AI/ML Benchmarks
  - Low precision GEMM
    - CNN Kernel, defines “AI-Flops”
  - Single Node CNN
    - AlexNet => RESNET?
    - ILSVRC2012 Dataset
  - Multi-Node CNN
    - Caffe+MPI (could allow other MPI-enabled frameworks)
  - Large Memory CNN
    - Convnet on Chainer
  - RNN / LSTM
    - OpenNMT RNN (collaboration w/NICT UCL)

Current ABCI Benchmarking Status

- Extensive discussions w/processor and system vendors, experienced DL researchers and users
- Collaboration with NICT (National Institute of Communication Technologies of Japan)
- Dedicated benchmark team formed within AIRC
- Early benchmarks and tweaking on AAIC
- Also looking at other HPC-DL benchmarks such as ANL CANDLE DNN benchmark
- Will be able to directly contribute to IRDS (along with other HPC benchmarks)
- We also have a generation of GPUs for DNN, from G200 to Latest Pascal, with Volta forthcoming (DGX-1V)
  - G200-S1070 (2008), Fermi M2050 (2010), Kepler K20c/K20x (2013), /K40/K80 (2014), Maxwell Geforce cards (2015), Pascal P100 (2016), Volta V100 (2017)

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

Fujitsu Deep Learning Processor (DLU™)

FUJITSU

京 K computer

Supercomputer K technologies

FY2018～

DLU™

(Deep Learning Unit)

FUJITSU DLU™ Deep Learning Unit

DLU™ features

- Architecture designed for Deep Learning
- High performance HBM2 memory
- Low power design

→ Goal: 10x Performance/Watt compared to others

- Massively parallel： Apply supercomputer interconnect technology

→ Ability to handle large scale neural networks

→ TOFU Network derivative for massive scaling

Designed for Scalable Learning, technically superior to Google TPU2

"Exascale" AI possible in 1H2019

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

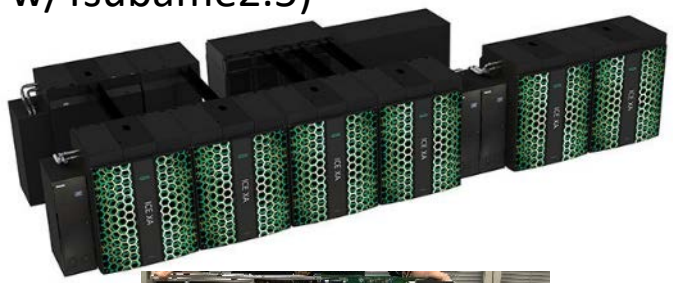
x5.0~7.7

Mar. 2018
ABCI (AIST-AIRC)
>130 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)

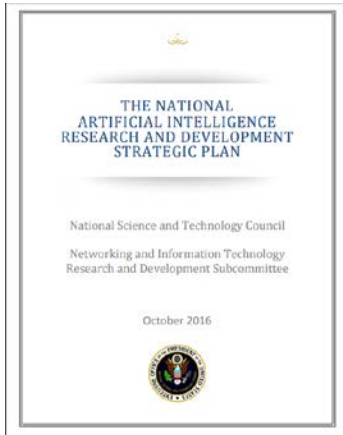
~x1000 in 3.5 years

Built/funded

still under plans

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

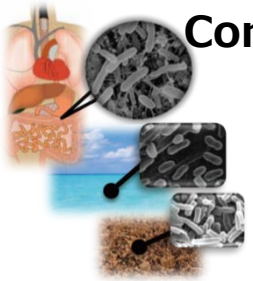
R&D Investments into world leading AI/BD HW & SW & Algorithms and their co-design for cutting edge Infrastructure absolutely necessary (just as is with Japan Post-K and US ECP in HPC)

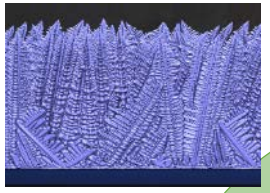


Co-Design of BD/ML/AI with HPC using BD/ML/AI - for survival of HPC

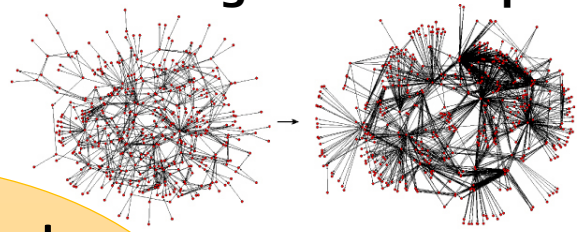
Acceleration and Scaling of BD/ML/AI via HPC and Technologies and Infrastructures

Accelerating Conventional HPC Apps





Large Scale Graphs



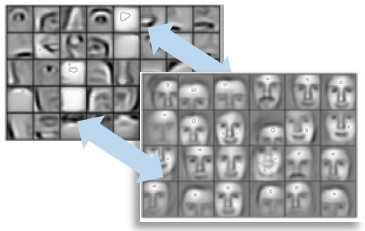
Big Data AI-Oriented Supercomputers

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

Big Data and ML/AI Apps and Methodologies

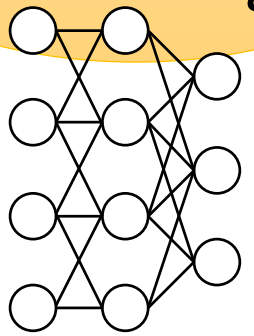
Optimizing System Software and Ops

Image and Video



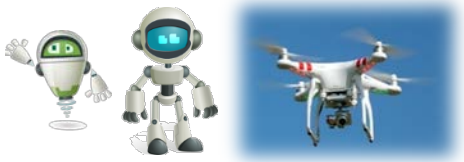
ABCI: World's first and largest open 100 Peta AI-Flops AI Supercomputer, Fall 2017, for co-design

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



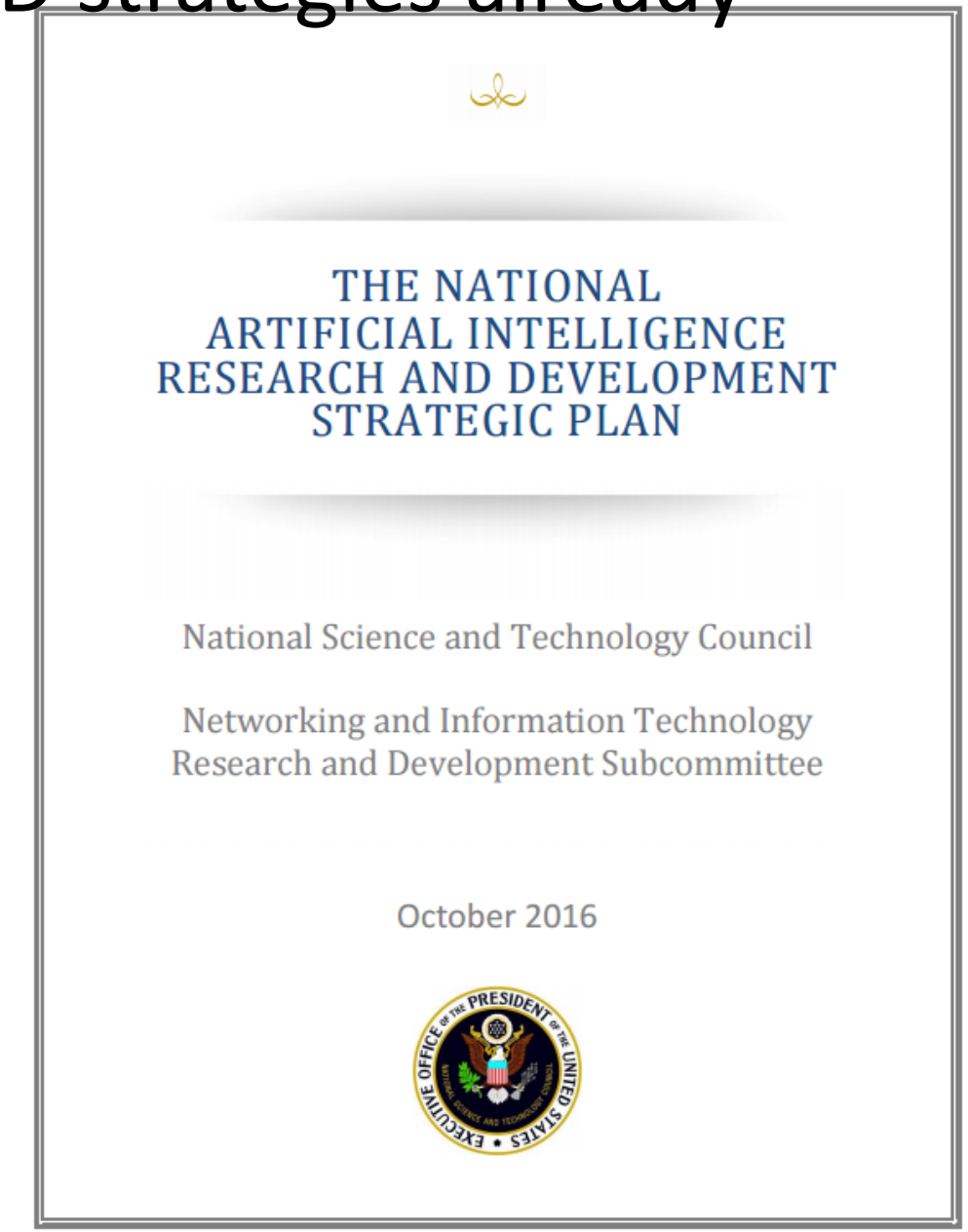
Future Big Data・AI Supercomputer Design

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.



What is worse: Moore's Law will end in the 2020's

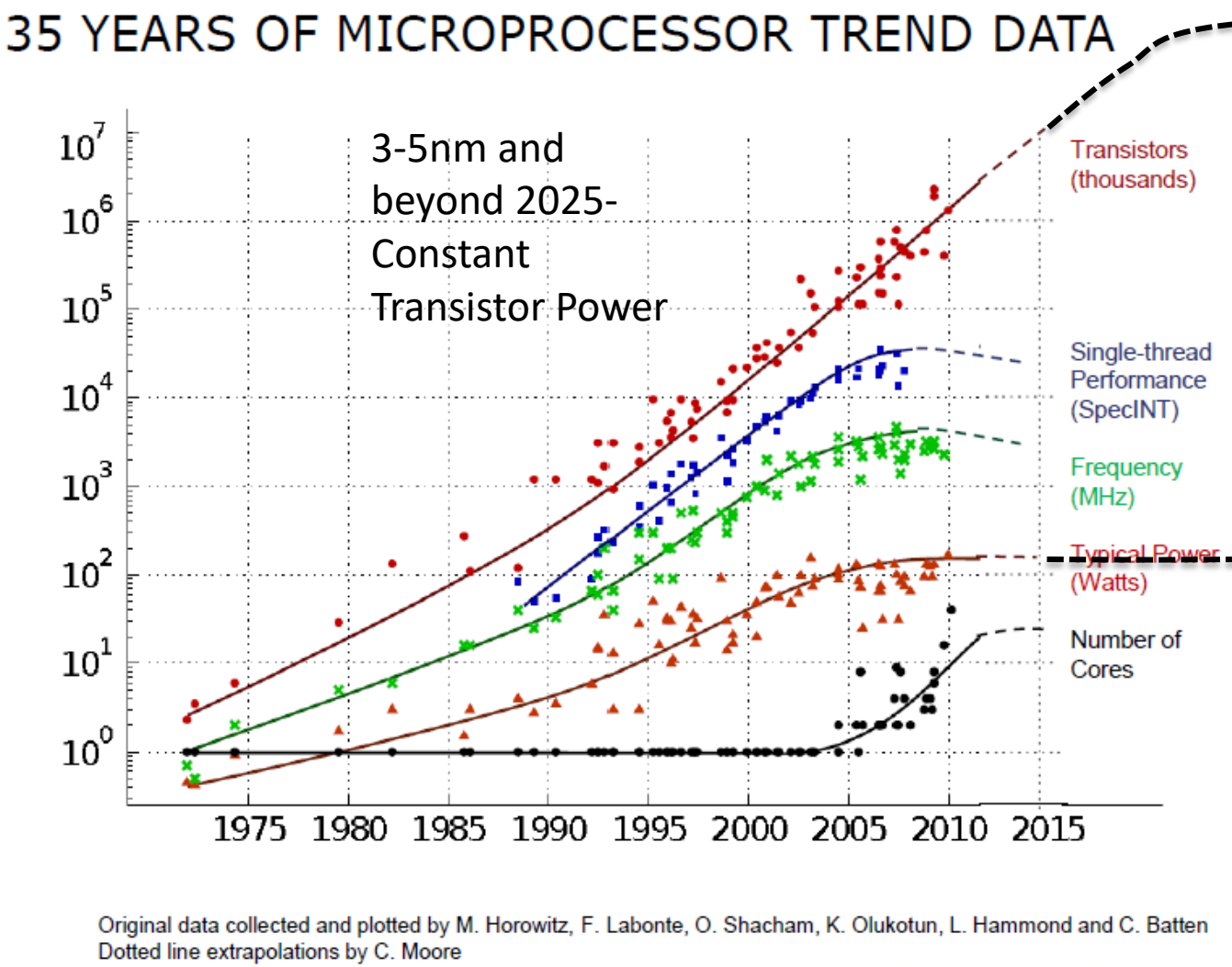
- Much of underlying IT performance growth due to Moore's law
  - "LSI: x2 transistors in 1~1.5 years"
  - Causing qualitative "leaps" in IT and societal innovations
  - The main reason we have supercomputers and Google…
- But this is slowing down & ending, by mid 2020s…!!!
  - End of Lithography shrinks
  - End of Dennard scaling
  - End of Fab Economics

Gordon Moore

The curse of constant transistor power shall soon be upon us

- How do we *sustain* "performance growth" beyond the "end of Moore"?
  - Not just one-time speed bumps
  - ***Will affect all aspects of IT, including BD/AI/ML/IoT, not just HPC***
  - ***End of IT as we know it***

20 year Eras towards of End of Moore's Law



20-year Moore-Dennard Single Core ILP-Vector Killer-Micro Era

20 year Post-Dennard Many-Core Era

20-year Next-Gen Post-Moore era

- 1980s~2004 Dennard scaling, perf+ = single thread+ = transistor & freq+ = power+
- 2004~2015 feature scaling, perf+ = transistor+ = core#+, constant power
- 2015~2025 all above gets harder
- 2025~ post-Moore, **constant feature&power = flat performance**

Need to realize the next 20-year era of supercomputing

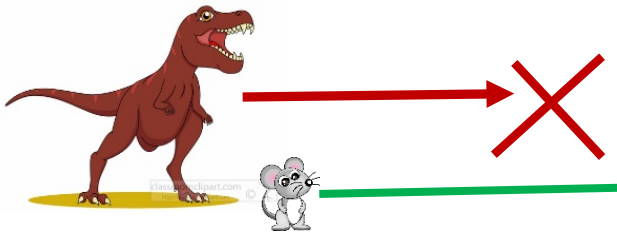
The “curse of constant transistor power”
- Ignorance of this is like ignoring global warming -

- Systems people have been telling the algorithm people that “FLOPS will be free, bandwidth is important, so devise algorithms under that assumption”
- This will certainly be true until exascale in 2020…
- But when Moore’s Law ends in 2025-2030, constant transistor power (esp. for logic) = FLOPS will no longer be free!
- So algorithms that simply increase arithmetic intensity will no longer scale beyond that point
- Like countering global warming – need disruptive change in computing – in HW-SW-Alg-Apps etc. for the next 20 year era

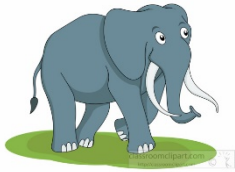
Performance growth via ***data-centric computing:***
"From FLOPS to BYTES"

- *Identify the new parameter(s) for scaling over time*
- Because data-related parameters (e.g. capacity and bandwidth) *will still likely continue to grow towards 2040s*
- Can grow transistor# for compute, but CANNOT use them AT THE SAME TIME(Dark Silicon) => **multiple computing units specialized to type of data**
- **Continued capacity growth**: 3D stacking (esp. direct silicon layering) and low power NVM (e.g. ReRAM)
- **Continued BW growth:** Data movement energy will be **capped constant** by dense 3D design and advanced optics from silicon photonics technologies
- Almost back to the old "vector" days(?), but no free lunch – latency still problem, locality still important, *need* ***general algorithmic acceleration thru data capacity and bandwidth,*** *not FLOPS*

Many Core Era



Post Moore Era

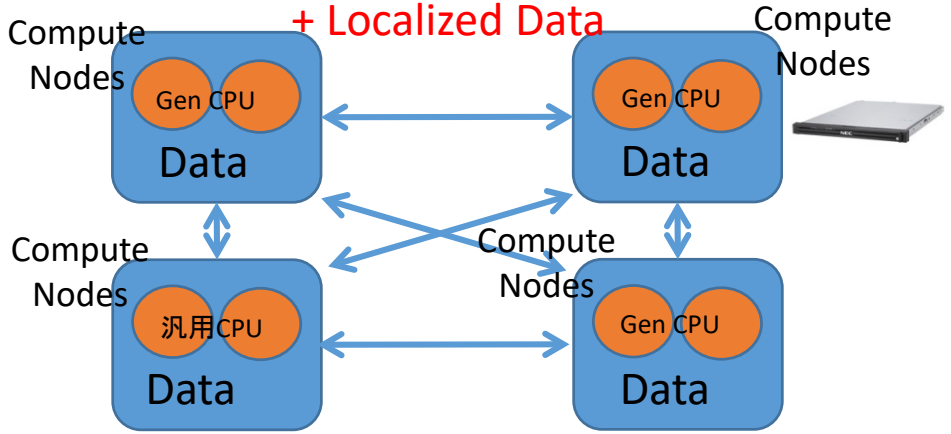


Flops-Centric Algorithms and Apps

Flops-Centric System Software

Hardware/Software System APIs
Flops-Centric Massively Parallel Architecture

Homogeneous General Purpose Nodes
+ Localized Data



Loosely Coupled with Electronic Interconnect

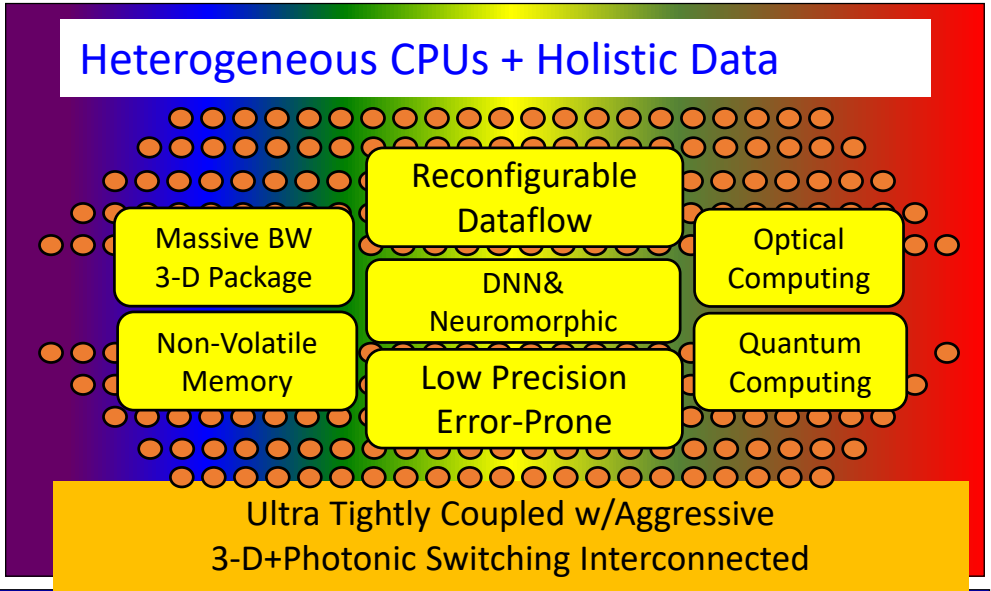
Transistor Lithography Scaling
(CMOS Logic Circuits, DRAM/SRAM)

~2025
M-P Extinction Event

Bytes-Centric Algorithms and Apps

Bytes-Centric System Software

Hardware/Software System APIs
Data-Centric Heterogeneous Architecture



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

Post-Moore is NOT a More-Moore device as a panacea

Device & arch. advances improving data-related parameters over time

"Rebooting Computing" in terms of devices, architectures, software. Algorithms, and applications necessary
=> Co-Design even more important
c.f. Exascale