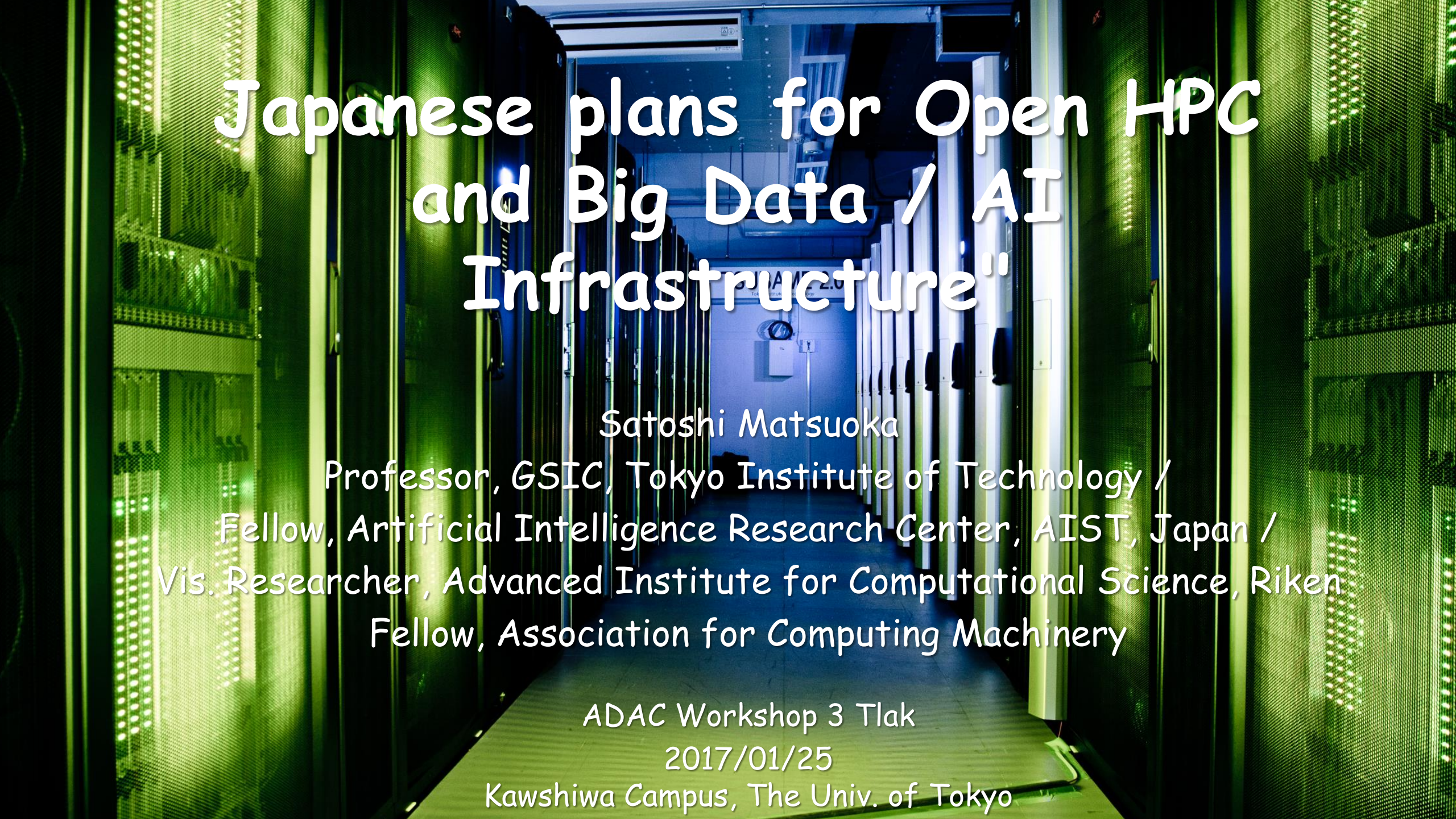


Japanese plans for Open HPC and Big Data / AI Infrastructure"

Satoshi Matsuoka

Professor, GSIC, Tokyo Institute of Technology /

Fellow, Artificial Intelligence Research Center, AIST, Japan /

Vis. Researcher, Advanced Institute for Computational Science, Riken

Fellow, Association for Computing Machinery

ADAC Workshop 3 Tlak

2017/01/25

Kawshiwa Campus, The Univ. of Tokyo

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

Comparing K Computer to TSUBAME 2.5

東京工業大学
Tokyo Institute of Technology

GSIC
Global Scientific Information and Computing Center

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

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)

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

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

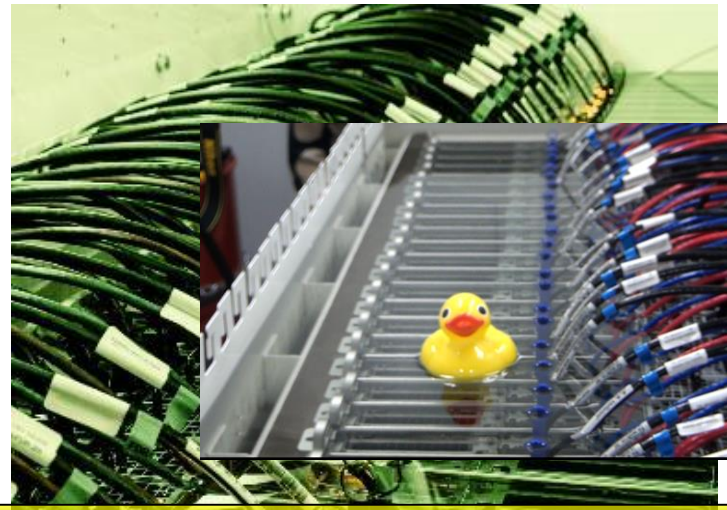
Tsubame current & future plans

- TSUBAME 2.5 (Production) Sep. 2013 – Mar 2019 (and beyond)
  - TSUBAME2.0 Nov. 2010-Sep. 2013, upgrade M2050 GPU -> K20X
  - 1424 nodes / 4224 GPUs, to be reduced to ~1300 nodes upon TSUBAME3 deployment
  - 5.7Petaflops (DFP), 17.1Petaflops (SFP)
- TSUBAME-KFC/DL (experimental, T3 Proto) – Oct 2013 – (Sep 2018 and beyond)
  - Upgrade to KFC/DL Oct. 2015 K20X GPU -> K80 GPU
  - 42 nodes / 336 GPU chips, 0.5/1.5 PF DFP/SFP
  - Oil immersion, ambient cooling, PUE < 1.09
- TSUBAME 3.0 (Production) Aug 2017 ~2021 (and beyond) – **Bid opens Jan 30th**
  - 12~15 Petaflops DFP depending on who wins
  - Parallel production to TSUBAME2.5
  - Focus on BD / AI workloads, not just traditional HPC => ~100PF max for AI combined with 2.5
- New IDC space construction for Tsubame3 and staggered operations beyond (T3+T4)
  - Power (4MW) + ambient cooling + storage (up to 100PB HDD) + high floor load (> 1 Ton / m^2)
  - To be completed March 2017
  - Power/Energy minimization for joint op in development

4

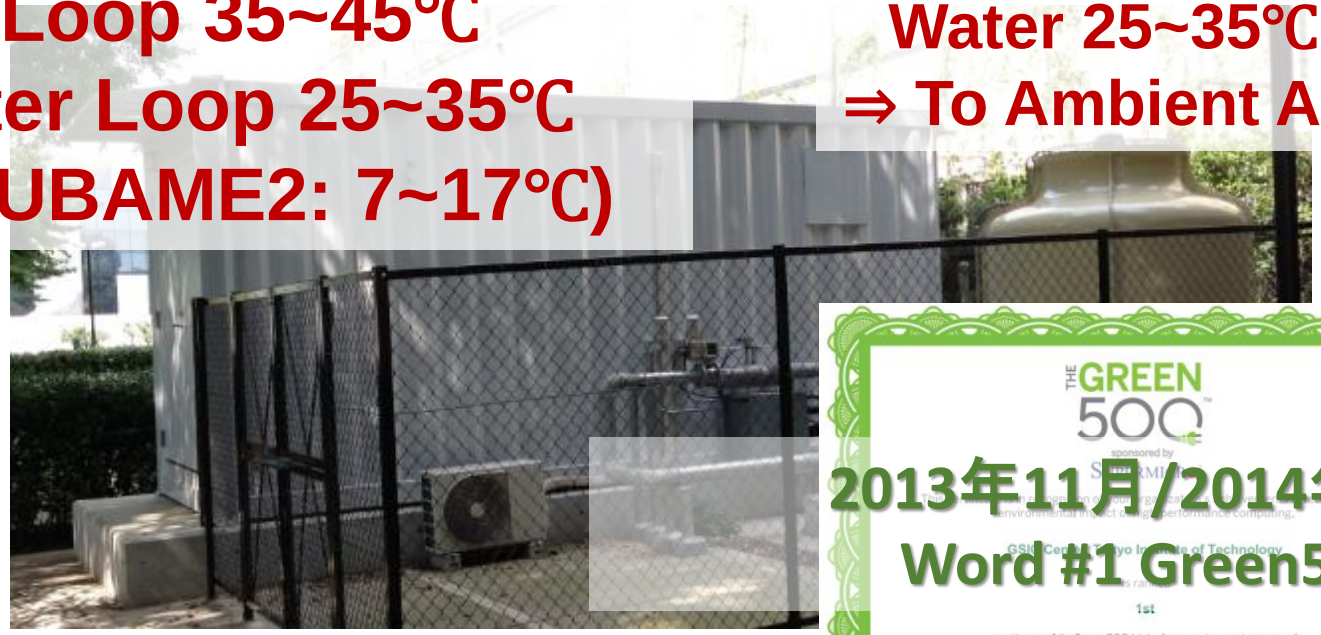
TSUBAME-KFC/DL: TSUBAME3 Prototype [ICPADS2014]

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 180m²…
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" – 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"

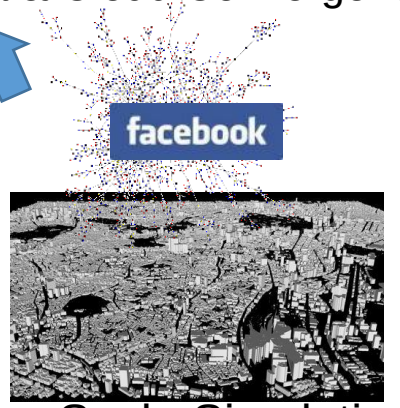
2010 TSUBAME2.0
2.4 Petaflops #4 World
"Greenest Production SC"

2011 ACM Gordon Bell Prize

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

2013 TSUBAME-KFC
#1 Green 500

2016 TSUBAME3.0+2.5
~20PF(DFP) 4~5PB/s Mem BW
10GFlops/W power efficiency
Big Data & Cloud Convergence



Large Scale Simulation
Big Data Analytics
Industrial Apps

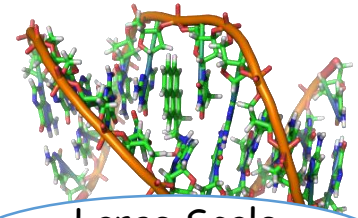
TSUBAME3 (some proposals) Bid open Jan 30th

- Extremely efficient power, FLOPS/W > 10 GFlops /W
- Extremely efficient cooling, PUE = 1.03 (annual avg), hot water cooling
- NVIDIA Pascal GPU accelerated architecture – TSUBAME2 Heritage
- Rich system interconnect BW – Intel Omipath, 1-to-1 GPU-to-HCA injection BW
- Rich accelerated I/O hierarchy – High Cap. / BW local NVMe on every node, aggregatable into single namespace via BeeGFS, staging from Lustre
  - Up to 1 Petabyte capacity
  - No Burst Buffer but exploits local storage to match performance much cheaper
- Full container based management for resource provisioning, isolation and execution environment control of accelerated systems
- Accelerated BD/AI/ML Software stack (partially) converged with HPC
- Future 100 Petabyte object store capacity extensions

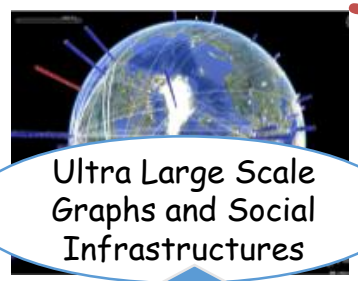
JST-CREST "Extreme Big Data" Project (2013-2018)

Future Non-Silo Extreme Big Data Scientific Apps

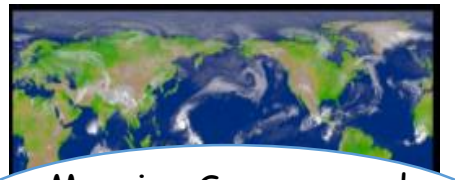
Given a top-class supercomputer, how fast can we accelerate next generation big data c.f. Clouds?



Large Scale Metagenomics



Ultra Large Scale Graphs and Social Infrastructures



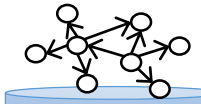
Massive Sensors and Data Assimilation in Weather Prediction

Co-Design

Co-Design

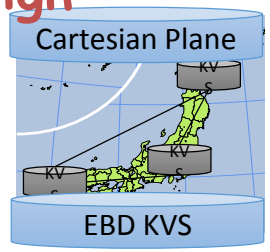
Co-Design

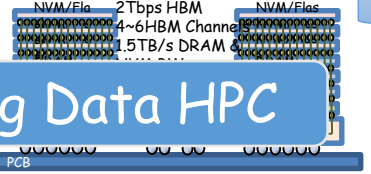
EBD Bag



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 regading Architectural, algorithmic, and system software evolution?

Use of GPUs?

Cloud IDC
Very low BW & Efficiency
Highly available, resilient

Supercomputers
Compute&Batch-Oriented
More fragile

Extreme Big Data (EBD) Team

Co-Design EHPC and EDB Apps

- **Satoshi Matsuoka (PI), Toshio Endo, Hitoshi Sato (Tokyo Tech.) (EBD Software System)**



- **Osamu Tatebe (Univ. Tsukuba) (EBD-I/O)**

- **Michihiro Koibuchi (NII) (EBD Network)**

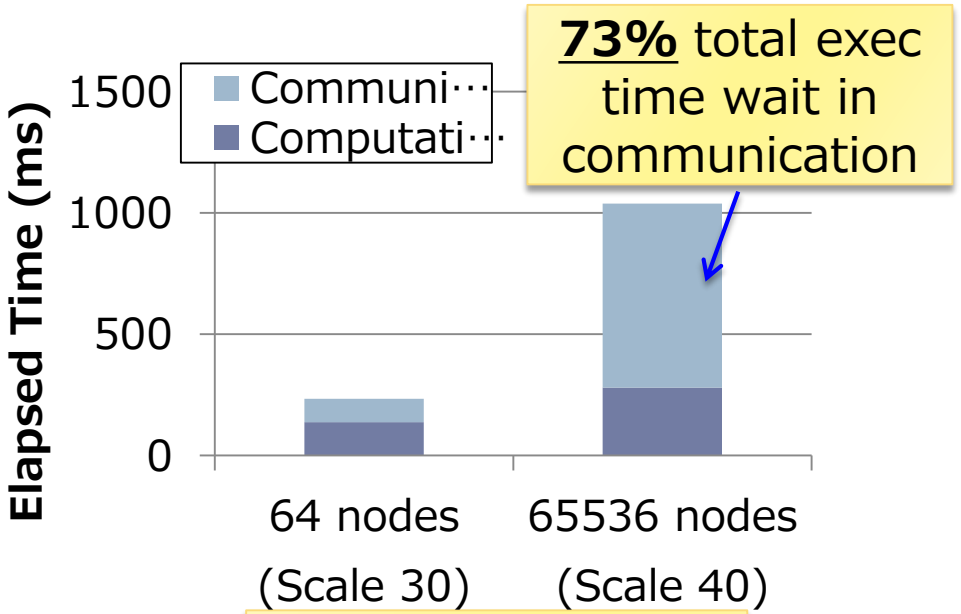
- **Yutaka Akiyama, Ken Kurokawa (Tokyo Tech) (EBD App1 Genome)**


- **Takemasa Miyoshi (Riken AICS) (EBD App2 Weather, data assim.)**

- **Toyotaro Suzumura (IBM Watson / Columbia U)(EBD App3 Social Simulation)**
- **(now merged into Matsuoka Team)**

The Graph500 – 2015~2016 – 4 Consecutive world #1

K Computer #1 Tokyo Tech[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

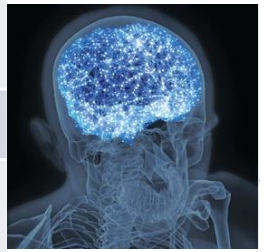
Effective x13 performance c.f. Linpack

TaihuLight
10 million CPUs
1.3 Petabyte mem

LLNL-IBM Sequoia
1.6 million CPUs
1.6 Petabyte mem

IBM

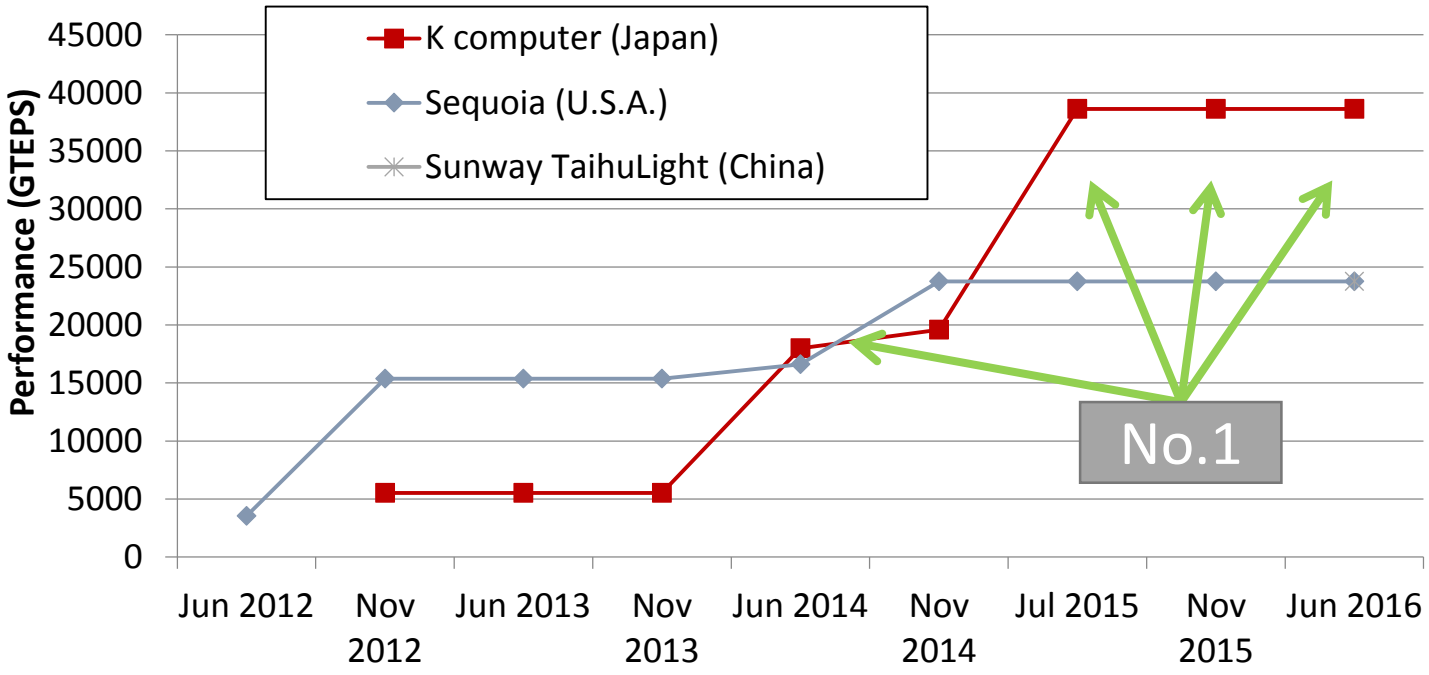
*Problem size is weak scaling
"Brain-class" graph



List	Rank	GTEPS	Implementat
November 2013	4	5524.12	Top-down o
June 2014	1	17977.05	Efficient hybrid
November 2014	2		Efficient hybrid
June, Nov 2015 June Nov 2016	**1**	**38621.4**	Hybrid + Node Compression

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**.

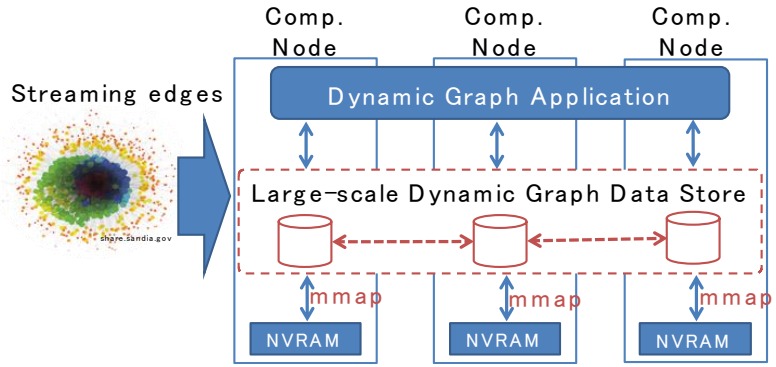
- Efficient Sparse Matrix Representation with Bitmap
- Vertex Reordering for Bitmap Optimization
- Optimizing Inter-Node Communications
- Load Balancing

  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)

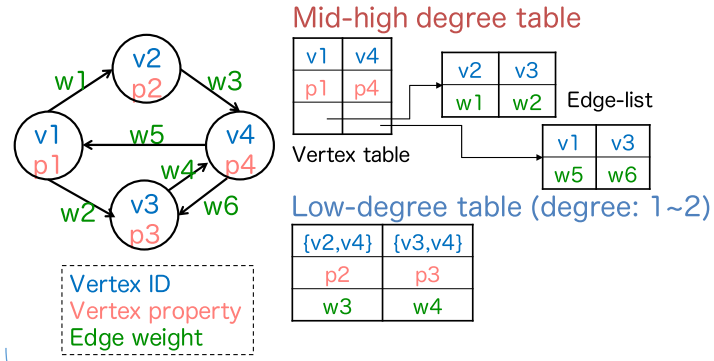
Towards a Distributed Large-Scale Dynamic Graph Data Store

Goal: to develop the data store for large-scale dynamic graph analysis on supercomputers



Node Level Dynamic Graph Data Store

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



Extend for multi-processes using an async MPI communication framework

Dynamic Graph Construction (on-memory)

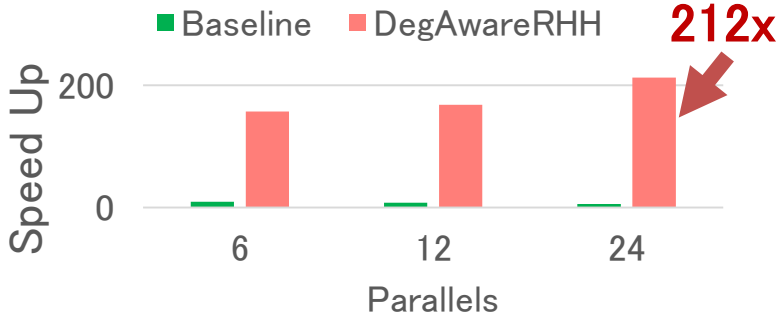
Against STINGER (single-node)

STINGER

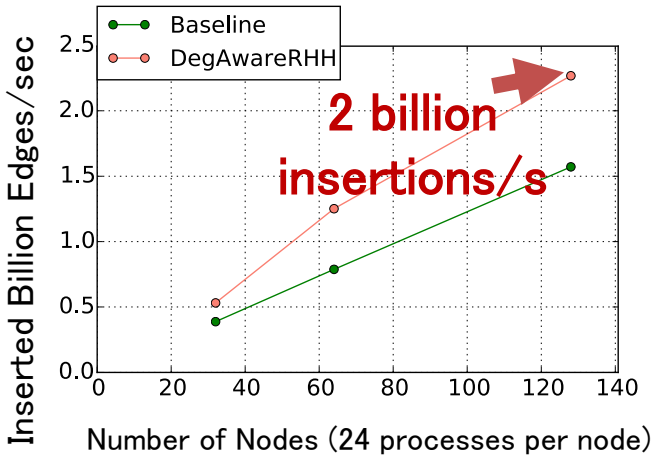
- A state-of-the-art dynamic graph processing framework developed at Georgia Tech

Baseline model

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



Multi-node Experiment

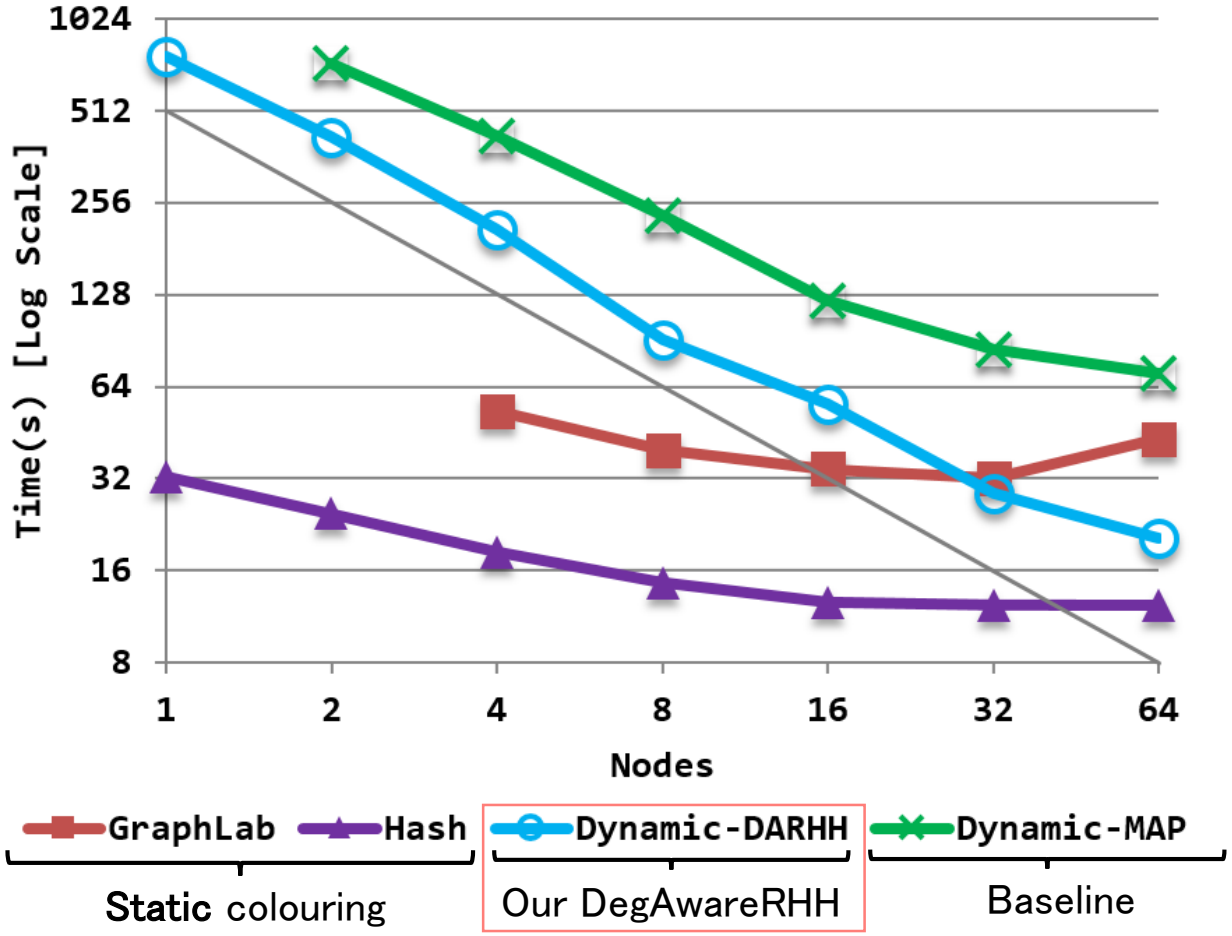


K. Iwabuchi, S. Sallinen, R. Pearce, B. V. Essen, M. Gokhale, and S. Matsuoka, Towards a distributed large-scale dynamic graph data store. In 2016 IEEE Interna- tional Parallel and Distributed Processing Symposium Workshops (IPDPSW)

Large-scale Graph Colouring (vertex coloring)

SC'16

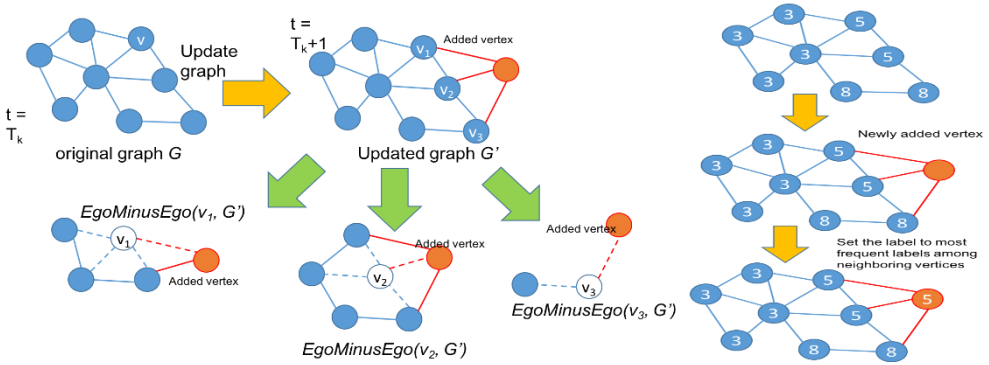
- Color each vertices with the minimal #colours so that **no** two adjacent vertices have the same colour
- Compare our dynamic graph colouring algorithm on DegAwareRHH against:
  1. two static algorithms including GraphLab
  2. an another graph store implementation with same dynamic algorithm (Dynamic-MAP)

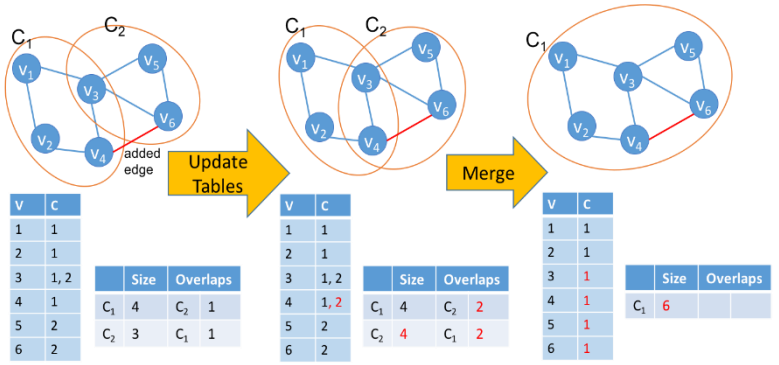


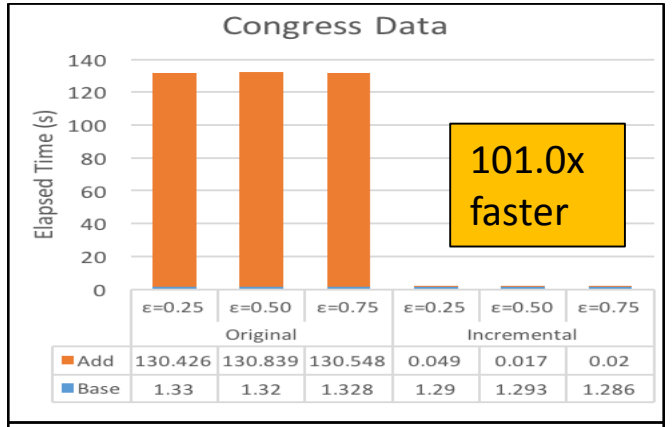
Scott Sallinen, Keita Iwabuchi, Roger Pearce, Maya Gokhale, Matei Ripeanu, "Graph Coloring as a Challenge Problem for Dynamic Graph Processing on Distributed Systems", SC' 16

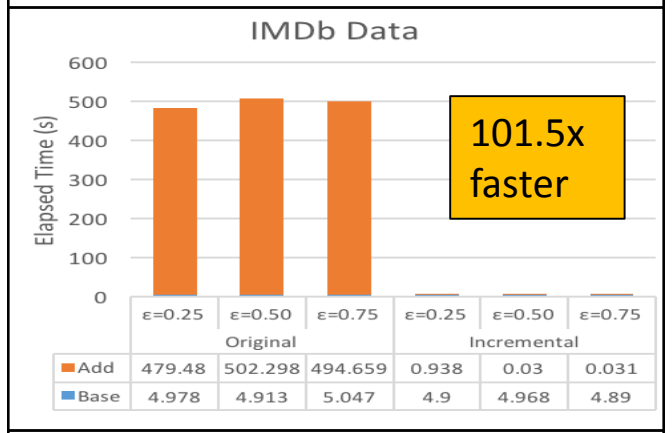
Incremental Graph Community Detection

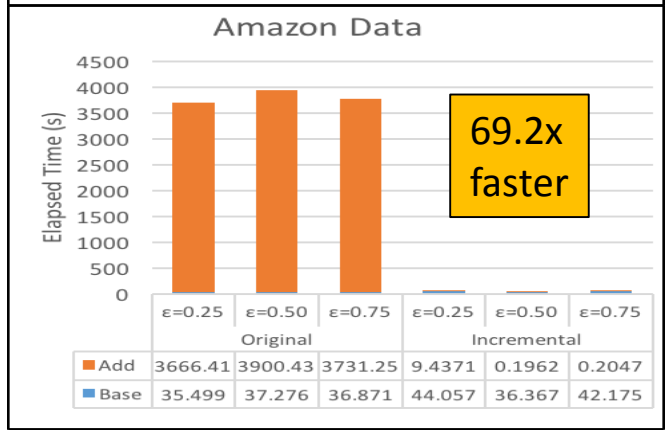
- Background
  - Community detection for large-scale **time-evolving and dynamic graphs** has been one of important research problems in graph computing.
  - It is time-wasting to compute communities entire graphs every time from scratch.
- Proposal
  - **An incremental community detection algorithm** 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]

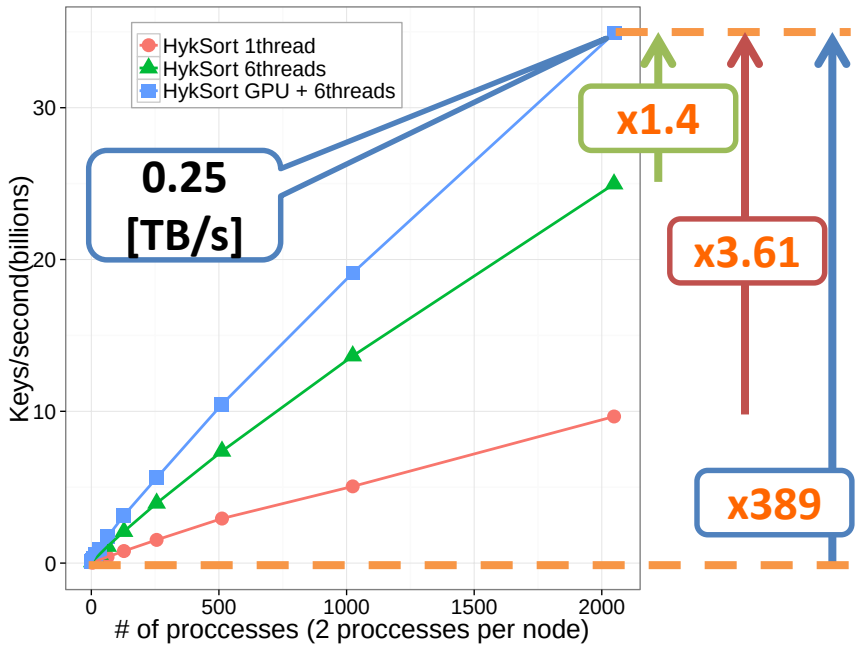
EBD Algorithm Kernels

- 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

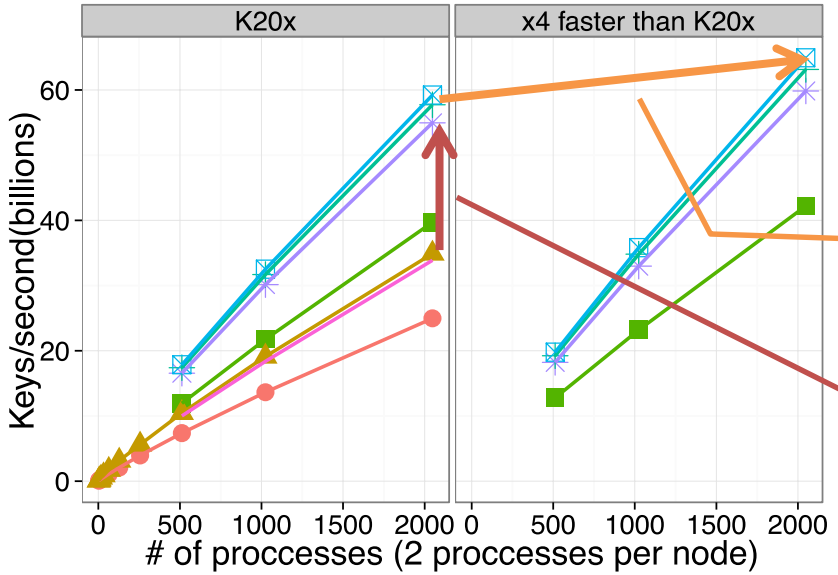
- We develop and evaluate bandwidth and latency reducing GPU-based HykSort on TSUBAME2.5 via latency hiding
  - Now preparing to release the sorting library

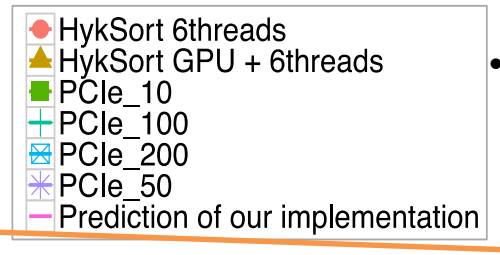
GPU implementation of splitter-based sorting (HykSort)

- 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

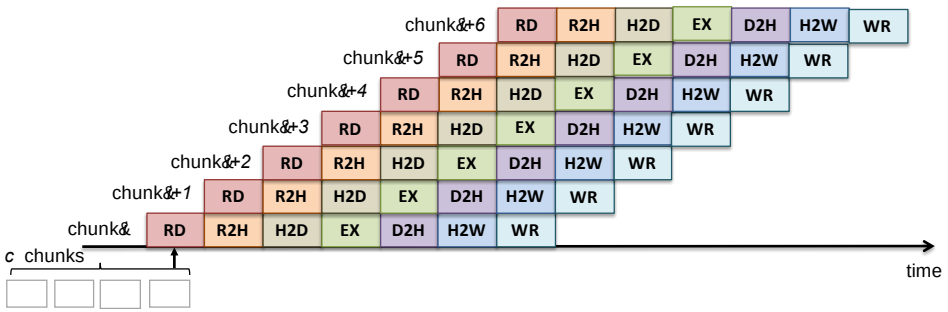
x2.2 speedup compared to CPU-based implementation when the # of PCI bandwidth increase to 50GB/s

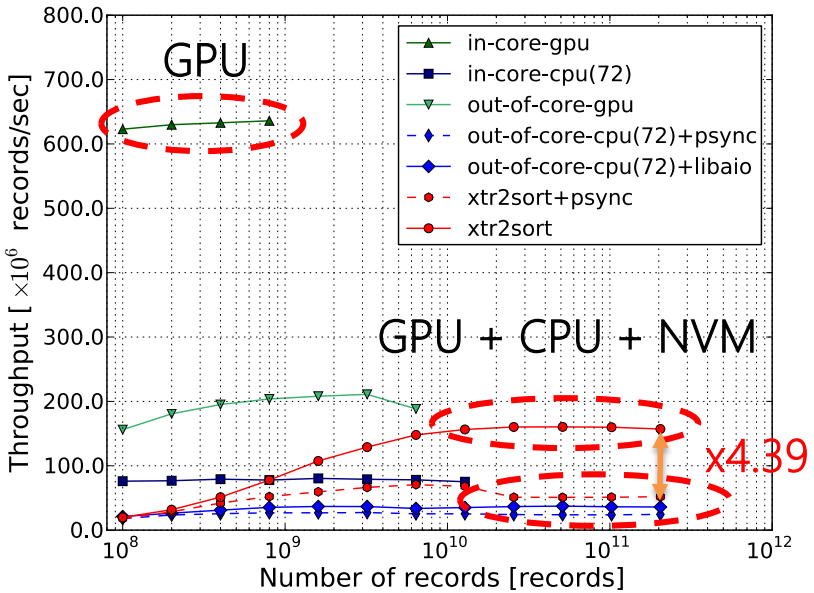
8.8% reduction of overall runtime when the accelerators work 4 times faster than K20x

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

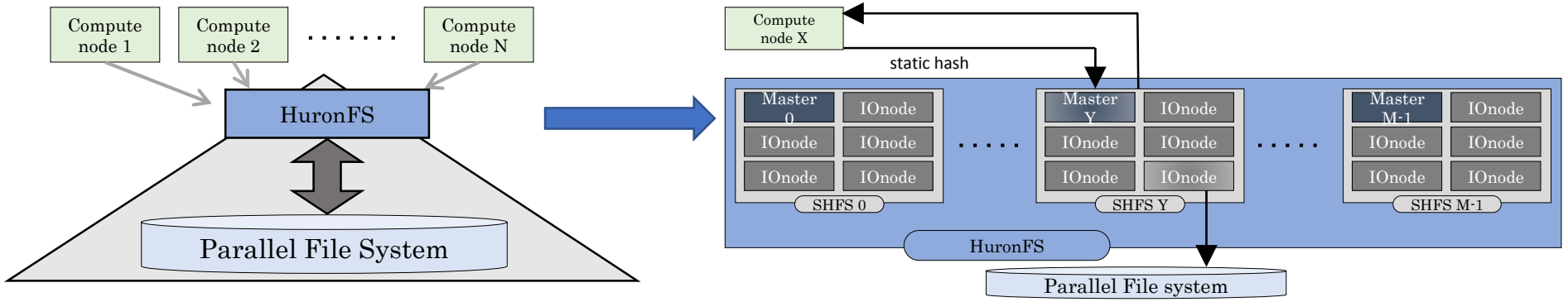
How to combine deepening memory layers for future HPC/Big Data workloads, targeting Post Moore Era?

- 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



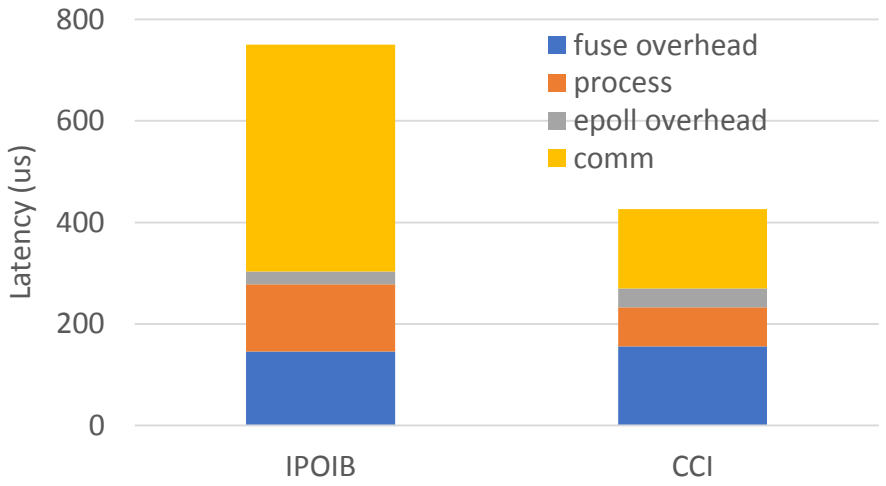


Hierarchical, UseR-level and ON-demand File system(HuronFS) (IEEE ICPADS 2016) w/LLNL



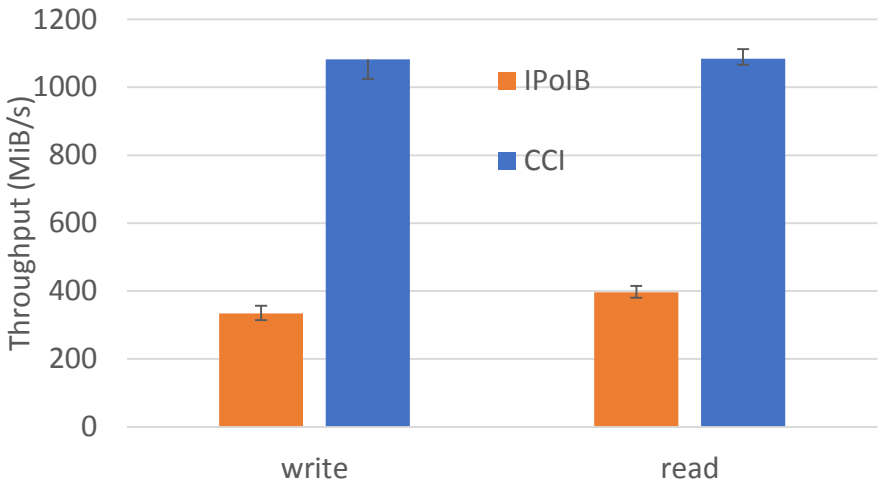
- HuronFS*:* dedicated dynamic instances to provide “burst buffer” for caching data
- I/O requests from *Compute Nodes* are forwarded to HuronFS
- The whole system consists of several SHFS (Sub HuronFS)
  - Workload are distributed among all the SHFS using hash of file path
- Each SHFS consists of a 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

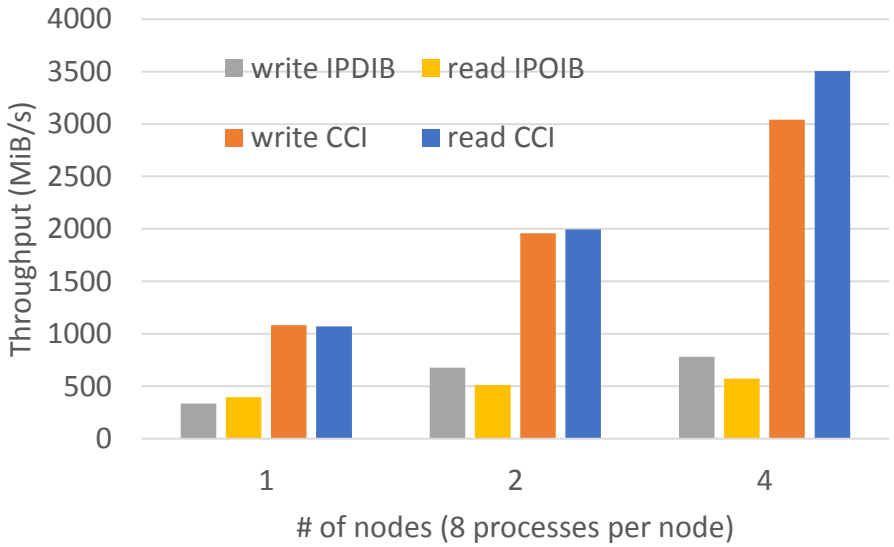


Latency

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



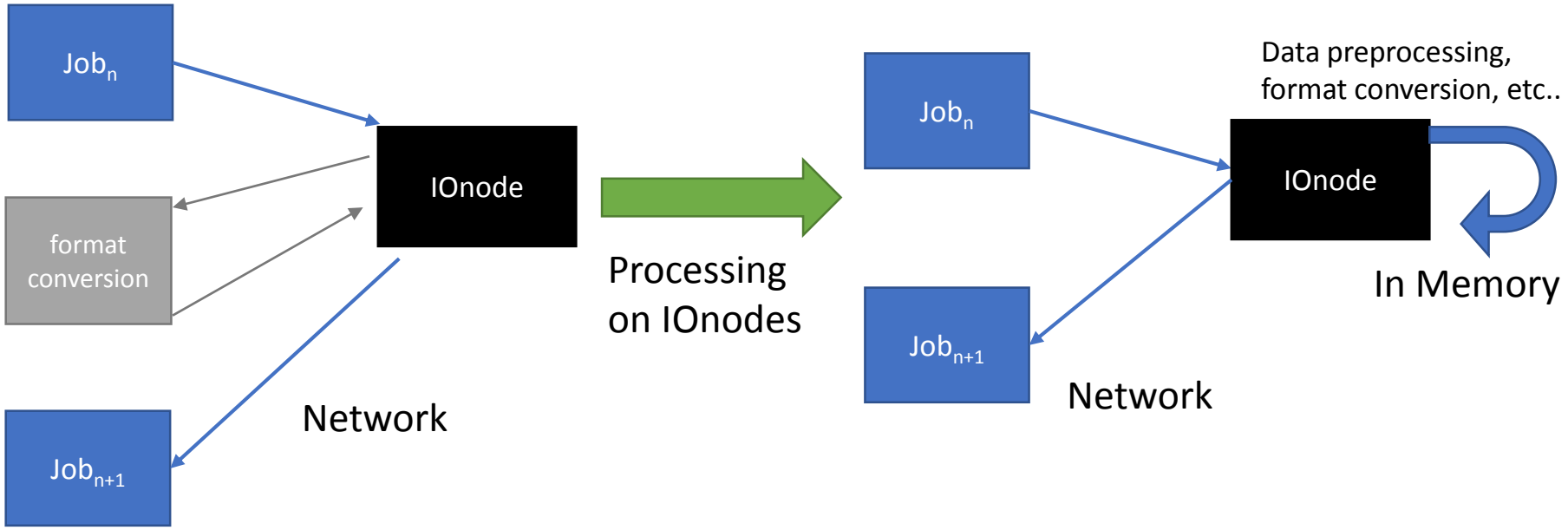
Throughput from single client



Throughput from single IOnode

Plans

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



Solving the Python Performance Problem

Fortran

+ top performance
+ HPC legacy
- hard to maintain
- Not used in BD/AI

Dichotomy Problem

Python

+ ease of programming
often used in BD/AI
+ general-purpose tools
- big runtime overhead

Dillema

- Performance or ease-of-programming?

Solution

- Python for development, Fortran at runtime.

take legacy Fortran

- reuse legacy code
- keep battle-tested implementations

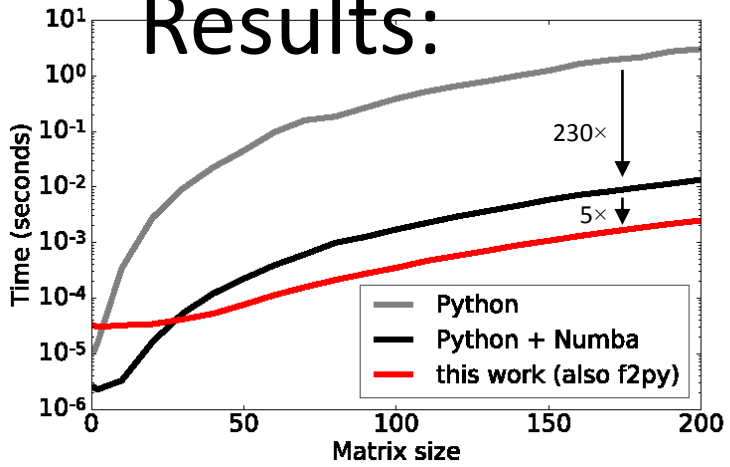
migrate to Python

- happens once
- semi-automatic or automatic
- performance-critical data is retained (via Type Hints)
- user can easily extend functionality

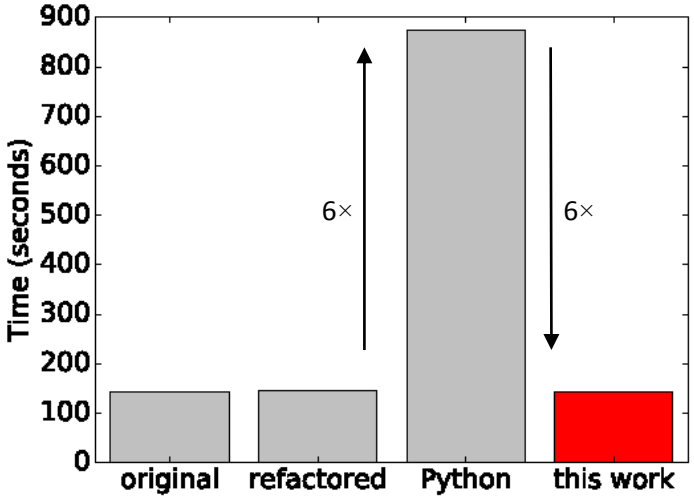
translate performance-critical kernels to Fortran

- JIT: at runtime
- fully automatic
- original performance is retained
- user doesn't interact with Fortran

Results:



DGEMM performance the same as Fortran. 5× better than Numba. [1]



Migrated Miranda IO benchmark retains original performance. [1]

[1] Mateusz Bysiek, Aleksandr Drozd, Satoshi Matsuoka. "Migrating Legacy Fortran to Python While Retaining Fortran-Level Performance Through Transpilation and Type Hints". In: Proceedings of the 6th Workshop on Python for High-Performance and Scientific Computing. PyHPC 2016. Salt Lake City, Utah, USA. ACM, 2016, URL: http://conferences.computer.org/pyhpc/2016/papers/5220a009.pdf

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
- CBB
  - 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 in development…

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

- Three projects and centers on Big Data and AI launched by three competing Ministries for FY 2016 (Apr 2016-)
  - MEXT – AIP (Artificial Intelligence Platform): Riken and other institutions ($~50 mil)
    - A separate Post-K related AI funding as well.
  - METI – AIRC (Artificial Intelligence Research Center): AIST (AIST internal budget + $~8 mil)
  - MOST – Universal Communication Lab: NICT ($50~55 mil)
  - $1 billion commitment on inter-ministry AI research over 10 years
- However, lack of massive platform and expertise in parallel computing c.f. Google, FB, Baidu…
  - MEXT attempts to suggest use of K computer
    -> community revolt “we want to use lots of GPUs like Google!”
  - MEXT Vice Minister Sadayuki Tsuchiya himself visits Matsuoka at Tokyo Tech Feb 1st, 2016.
    - “What is GPU and why is it so good for DL/AI?”
    - “Can you and TSUBAME can contribute to the MEXT projects directly over multiple years, with appropriate funding?”
  - Similar talks with METI & AIRC
    - “Can TSUBAME be utilized to cover the necessary research workload at AIRC?” --- Satoshi Sekiguchi, Director of Informatics, AIST

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

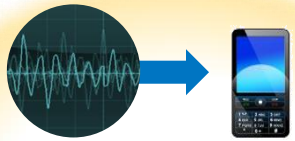
Image/Video Recognition

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

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

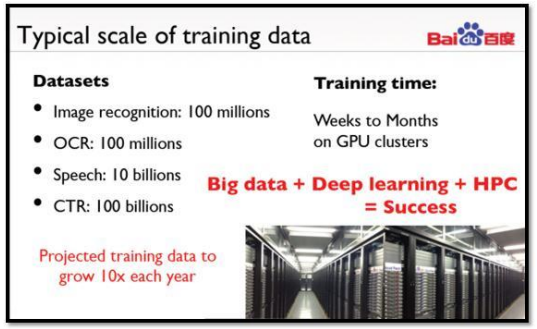
各種推定値は1GBの学習データに対して1日で学習するためには
1TFlops必要だとして計算

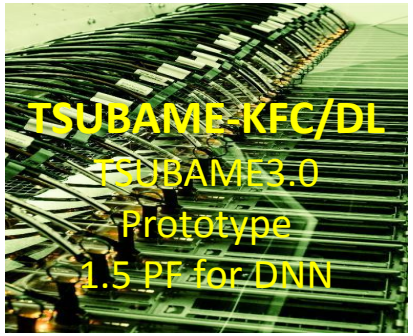
10PF 100PF 1EF 10EF 100EF

2015 2020 2025 2030

Research on Advanced Deep Learning Applications (Part of JST Extreme Big Data Project 2013-2018)

- Deep Learning IS HPC!
  - Training models – mostly dense MatVec
  - Data Access for training target data sets
  - Sharing updated training parameters in neural networks
- Goals
  - Accelerate DL applications in EBD architectures ?
    - Extreme-scale Parallelization, Fast Interconnects, Storage I/O, etc.
  - Performance bottlenecks of multi-node parallel DL algorithms on current HPC systems ?
- Current Status
  - Official Collaboration w/DENSO IT Lab signed November
  - Profiling based bottleneck identification and performance modeling & optimization of a real DL application on TSUBAME
    - Great result, joint paper being prepared for submission
  - > 100 million images, 1500 GPUs (6 Pflops) 1 week grand challenge run
  - Compete w/Google, MS, Baidu etc. in ILSVRC in ImageNet with shallow network
    - To fit within smaller platforms e.g. Jetson
    - Got reasonable results, about 10% accuracy with 15-layer CNN
  - Denso Lab continues to run workloads on TSUBAME2.5 and TSUBAME-KFC/DL
  - In talks with other companies, e.g. Yahoo! Japan

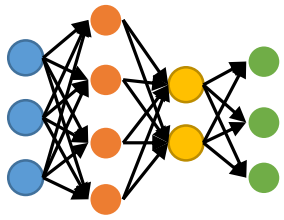


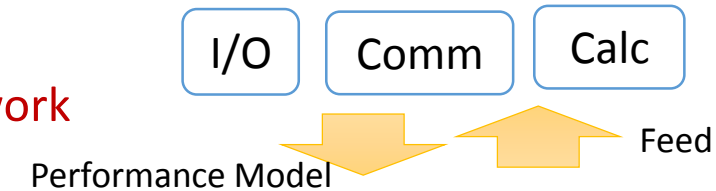


Many companies (ex. Baidu, etc.) employ GPU-based Cluster Architectures, similar to TSUBAME2 & KFC

DENSO

Real DL Applications

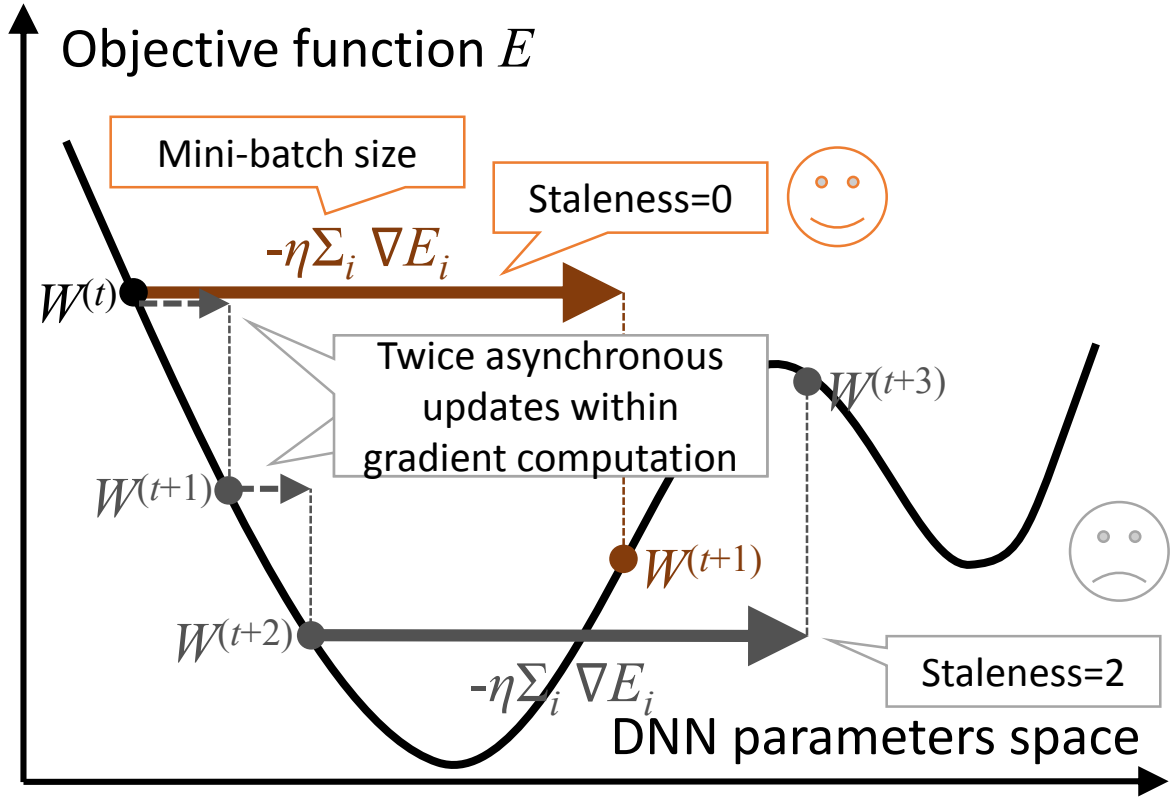




Predicting Statistics of Asynchronous SGD Parameters for a Large-Scale Distributed Deep Learning System on GPU Supercomputers

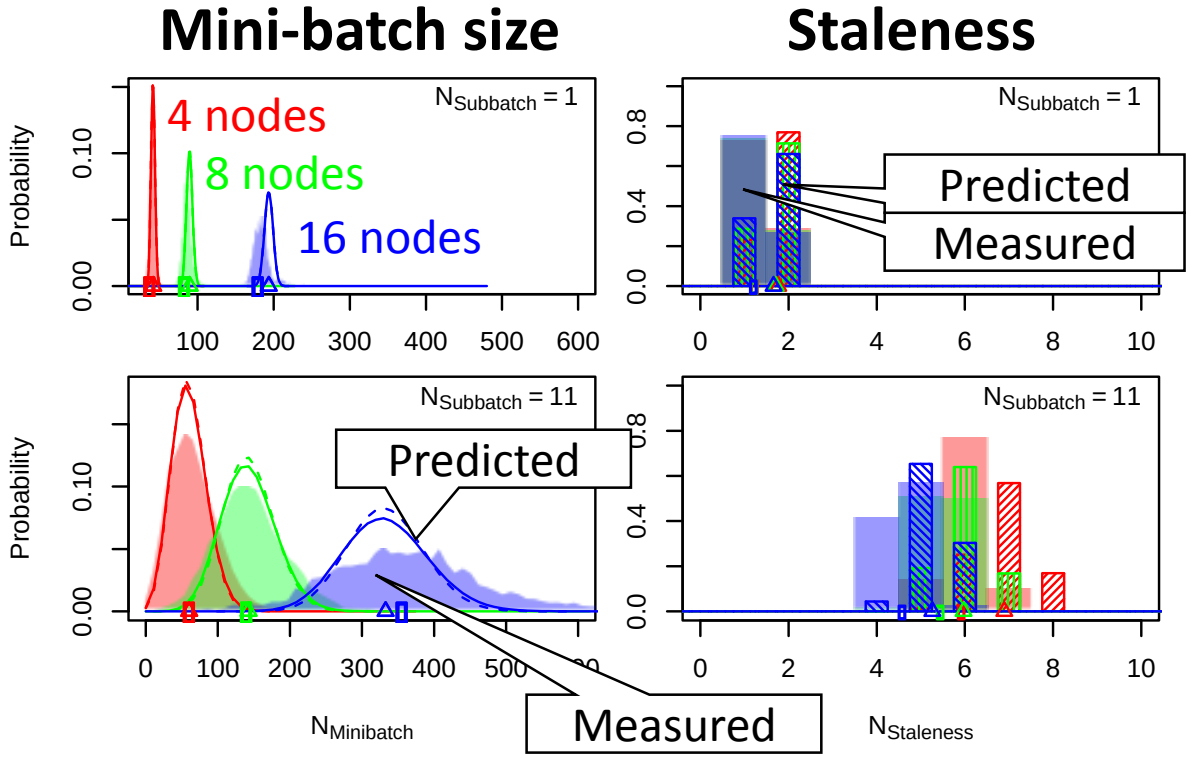
Background

- 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 (to appear)

Approach and Contribution

- **Approach**: Proposing a performance model for an ASGD deep learning system SPRINT which considers probability distribution of mini-batch size and staleness
  - Takes CNN structure and machine specifications as input
  - Predicts time to sweep entire dataset (epoch time) and the distribution of the statistics
- **Contribution**
  - Our model predicts epoch time, average mini-batch size and staleness with 5%, 9%, 19% error in average respectively on several supercomputers
  - Our model steadily choose the fastest machine configuration that nearly meets a target mini-batch size

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

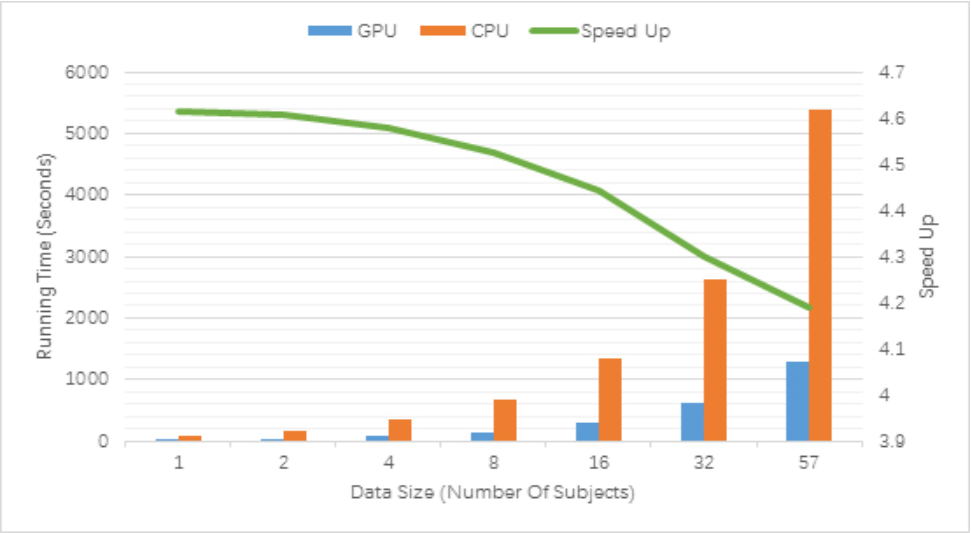
GPU-Based Fast Signal Processing for Large Amounts of Snore Sound Data

- 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

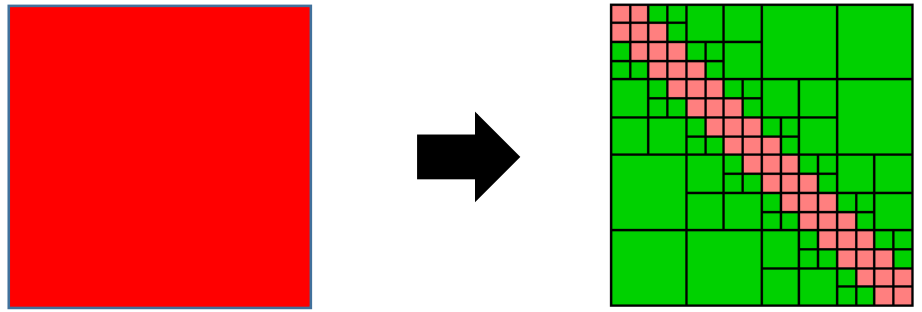
- Result

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.

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

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

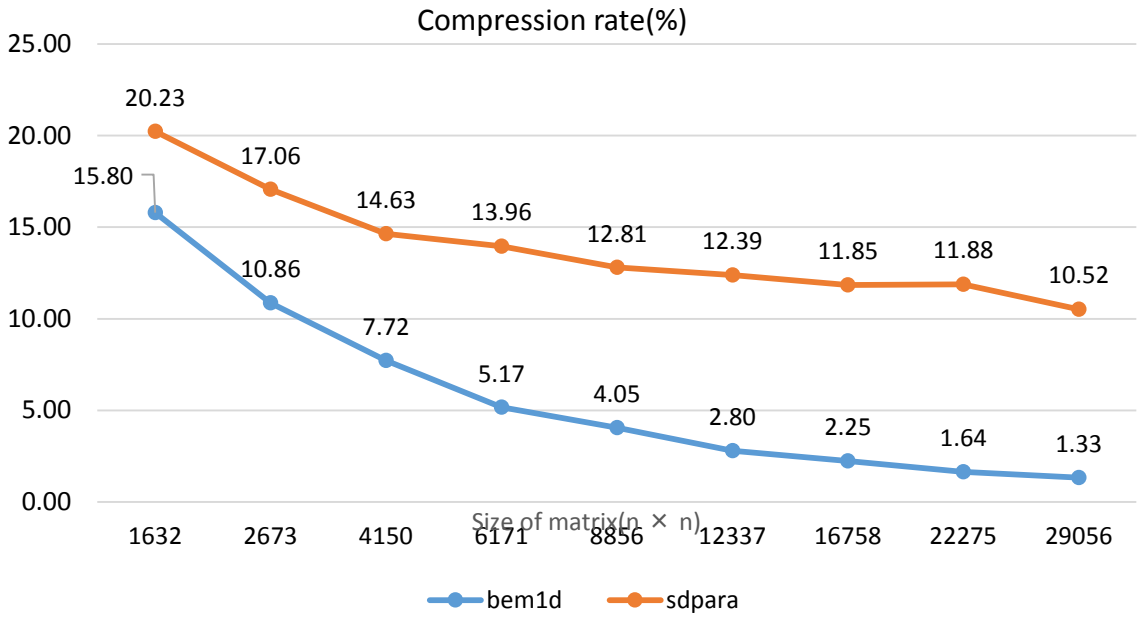
- Computational complexity

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

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

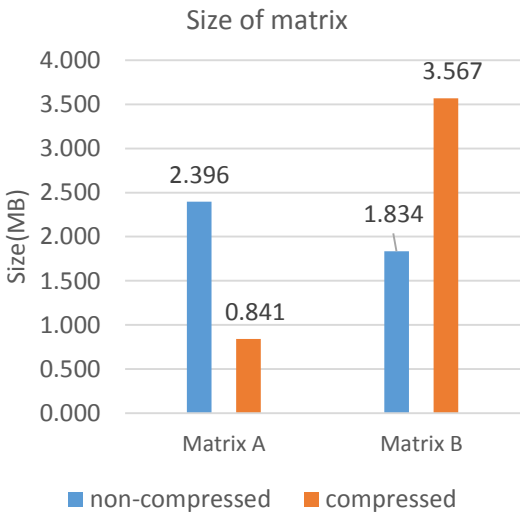
Preliminary Results – Compression rate of matrices

SDPARA



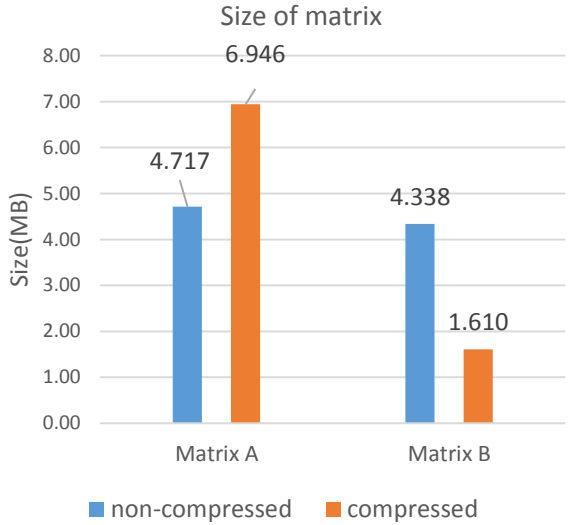
Compressive rate = (uncompressed size) / (compressed size)

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

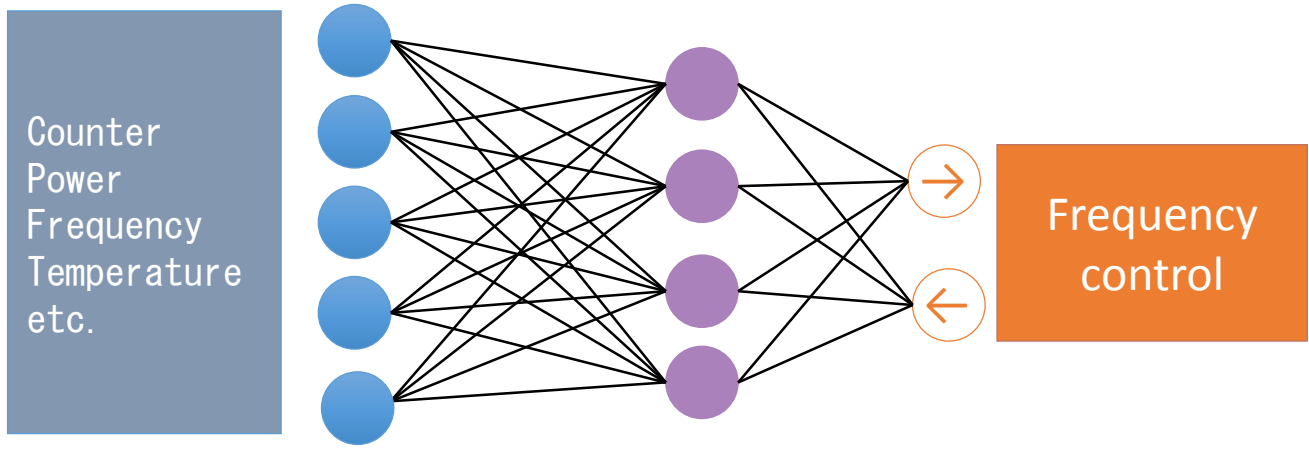
→ Use Deep Learning for analysis.

- 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.



Two AI CREST Programs (2016-2023) ~$40 mil x 2

Intelligent Information Processing Systems Creating Co-Experience Knowledge and Wisdom with Human-Machine Harmonious Collaboration

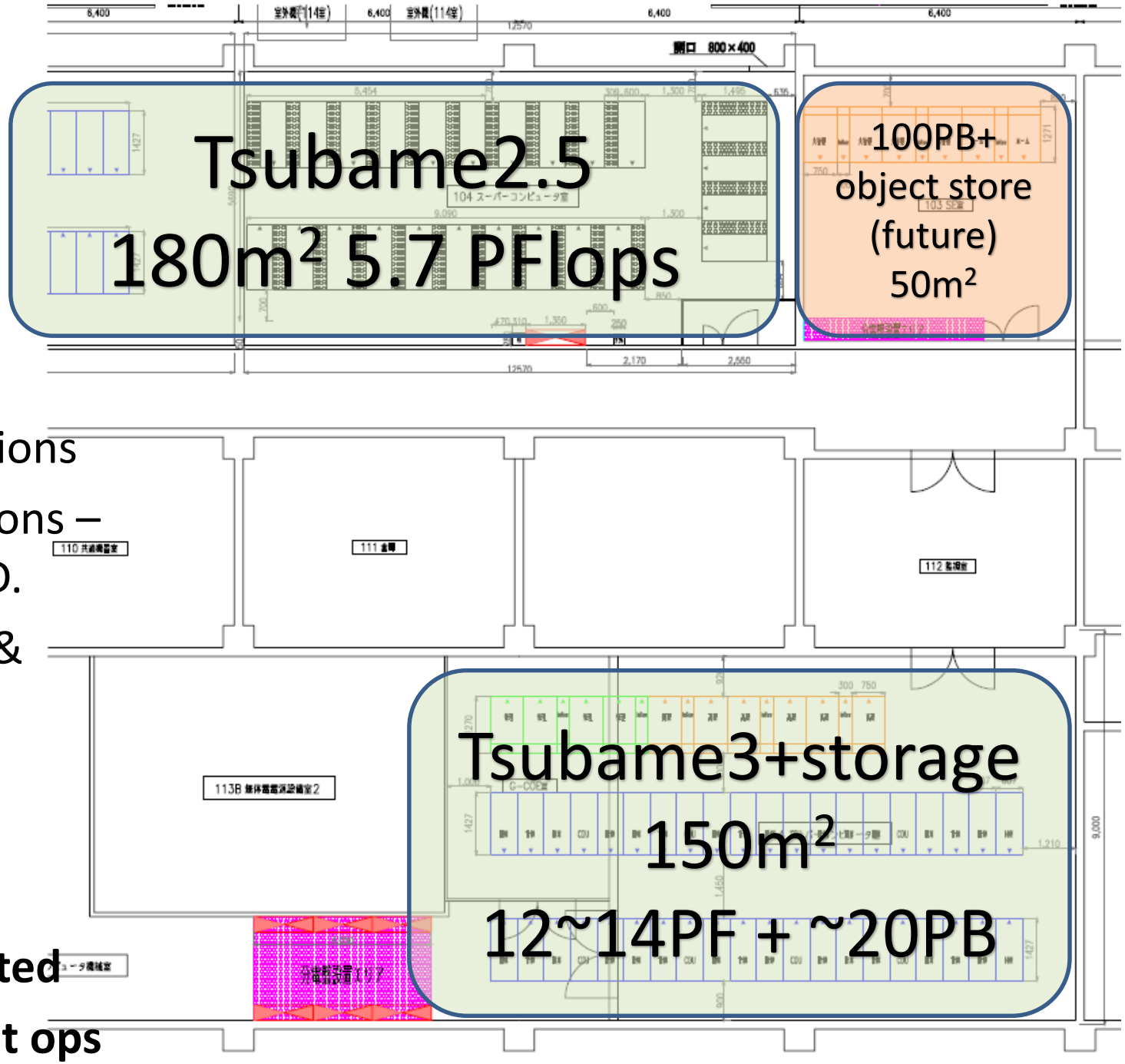
Research Supervisor: Norihiro Hagita (Board Director, Director, Intelligent Robotics and Communication Laboratories, Advanced Telecommunications Research Institute International)

Development and Integration of Artificial Intelligence Technologies for Innovation Acceleration

Research Supervisor: Minoru Etoh (Senior Vice President, General Manager of Innovation Management Department, NTT DOCOMO, INC.)

TSUBAME2&3 Joint Operation Plan

- New dedicated datacenter space for Tsubame3 => retain TSUBAME2
- Joint operation 2017~2019
  - TSUBAME3: mainline HPC operations
  - TSUBAME2.5: specialized operations – industry jobs, long running, AI/BD.
- Power capped not to exceed power & cooling limits (4MW)
- **Total 6~7000 GPUs, ~70Pflops for AI**
  - **Storage enhanced to cope w/capacity**
  - **Pending budgetary allocation**
- **Construction on new IDC space started**
- **Future: TSUBAME3+TSUBAME4 joint ops**



Comparison of Machine Learning / AI Capabilities of TSUBAME3+2.5 and K-Computer

東京工業大学
Tokyo Institute of Technology

GSIC
Global Scientific Information and Computing Center

TSUBAME2.5(2013)
+TSUBAME3.0(2017)

Deep Learning / AI Capabilities
FP16+FP32 up to ~70 Petaflops
+ up to 100PB online storage

x~7
>>
(effectively more due to optimized DL SW Stack on GPUs)

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

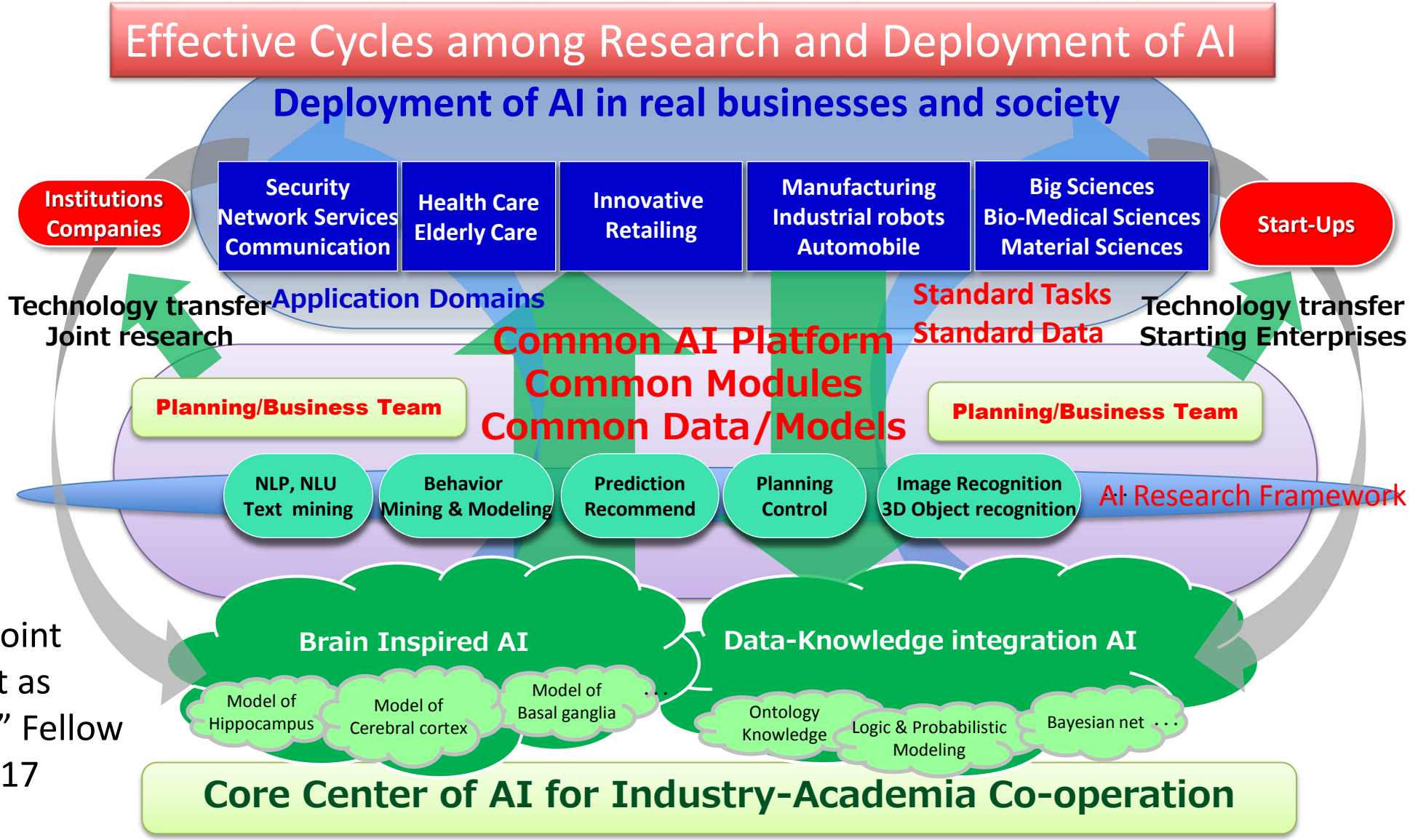
K Computer (2011)

Deep Learning
FP32 11.4 Petaflops

Slightly faster than U-Tokyo under this metric

AI Research Center (AIRC), AIST

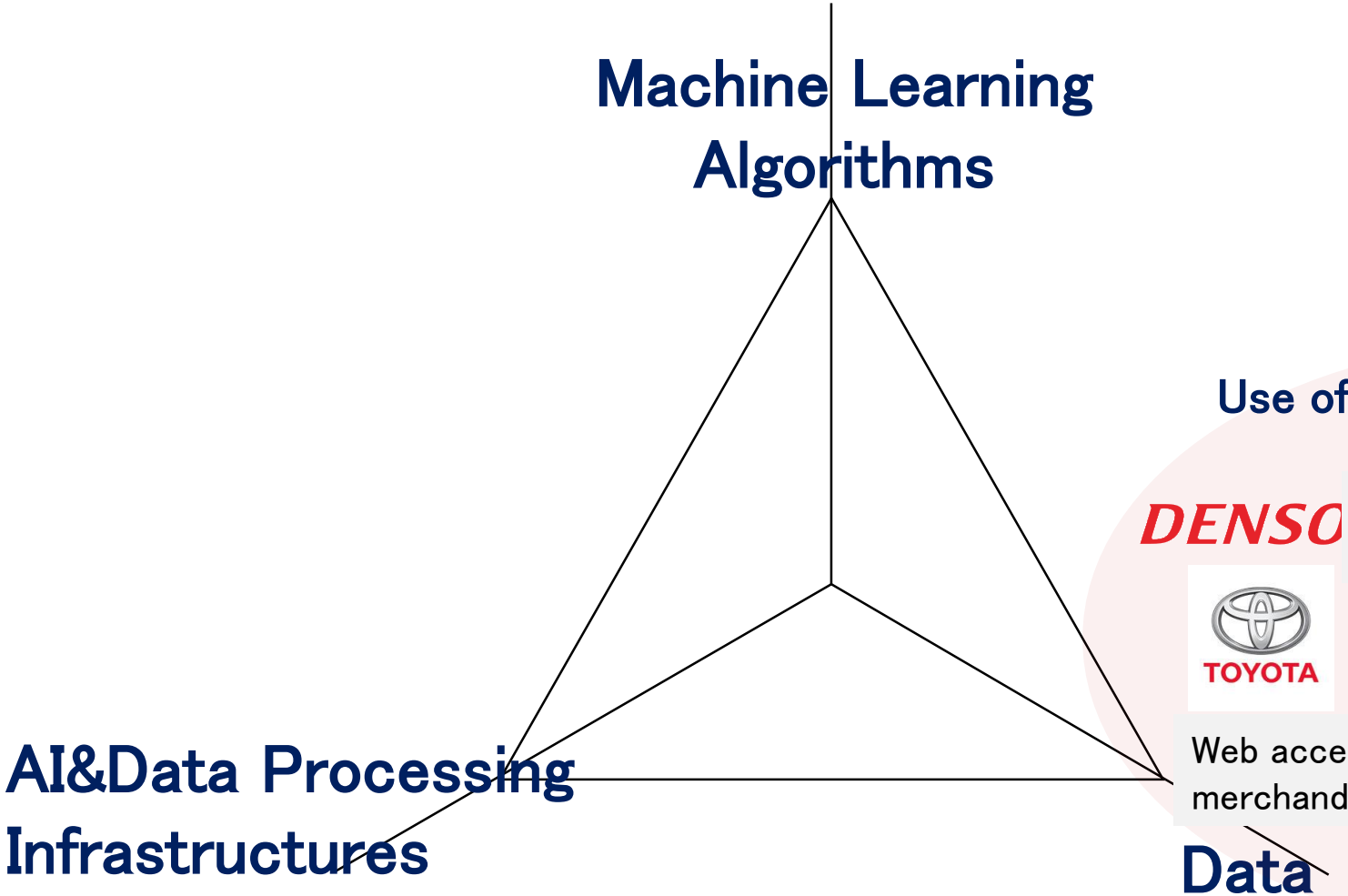
Now > 300+ FTEs



Matsuoka : Joint appointment as “Designated” Fellow since July 2017

The current status of AI & Big Data in Japan

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



Use of Massive Scale Data now Wasted

DENSO Petabytes of Drive Recording Video

FANUC FA&ロボット&ロボマシン FA&Robots

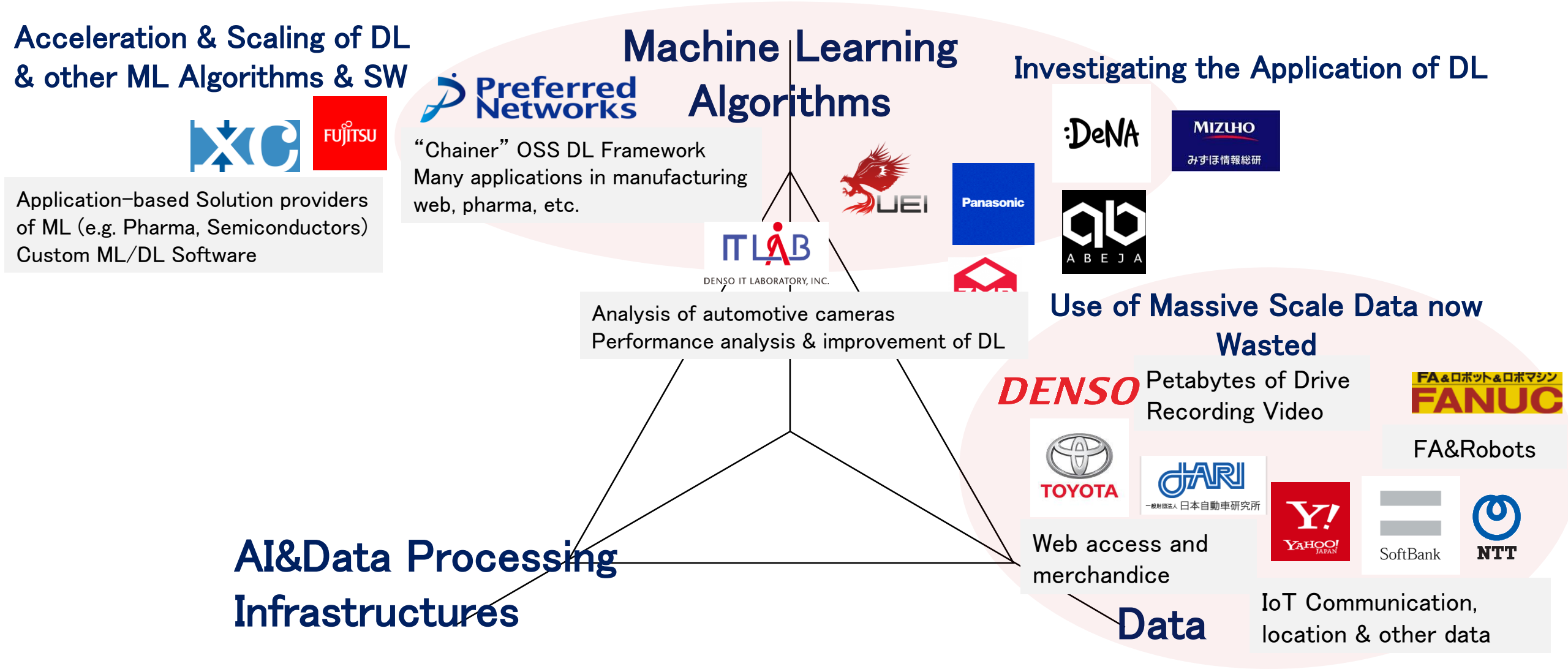
TOYOTA JARI 一般財団法人 日本自動車研究所

Y! YAHOO! JAPAN SoftBank NTT Web access and merchandice

IoT Communication, location & other data

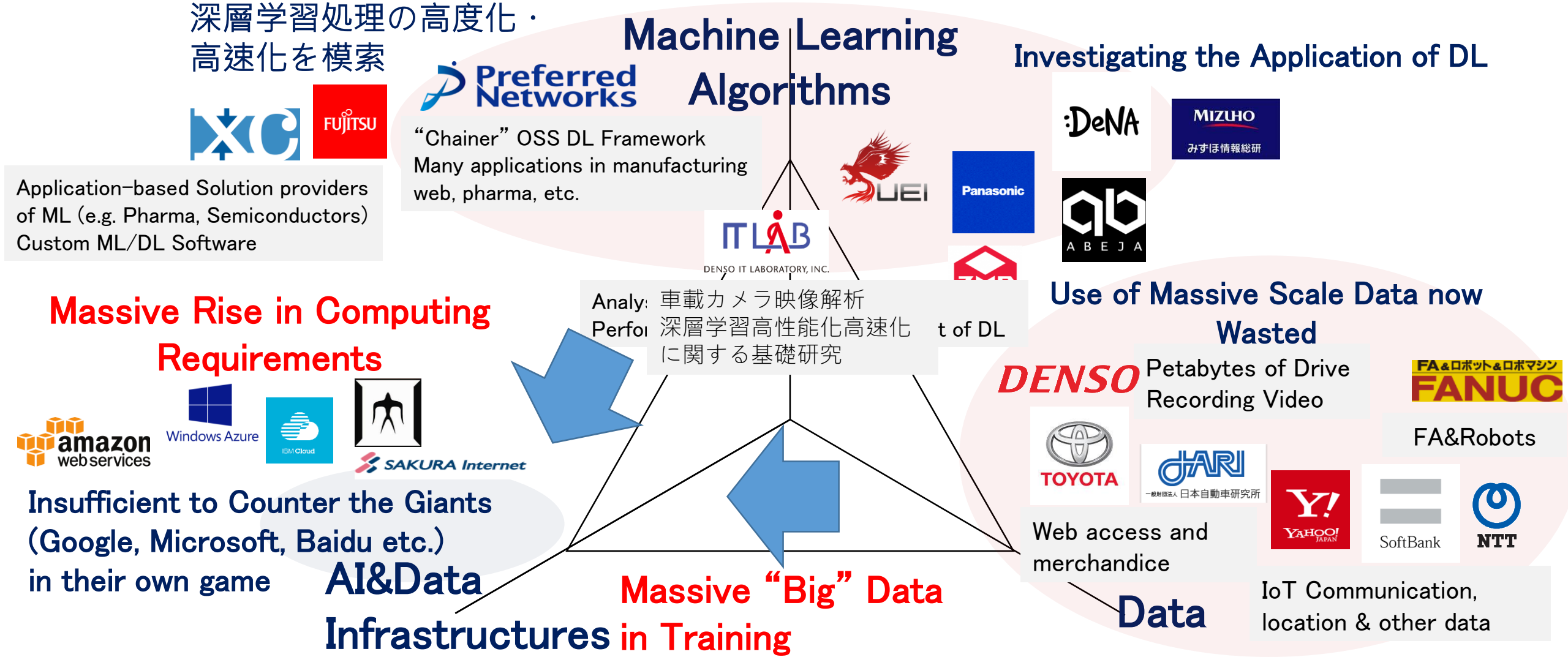
The current status of AI & Big Data in Japan

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

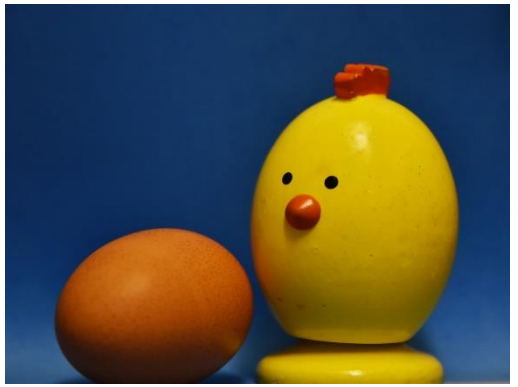


The current status of AI & Big Data in Japan

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



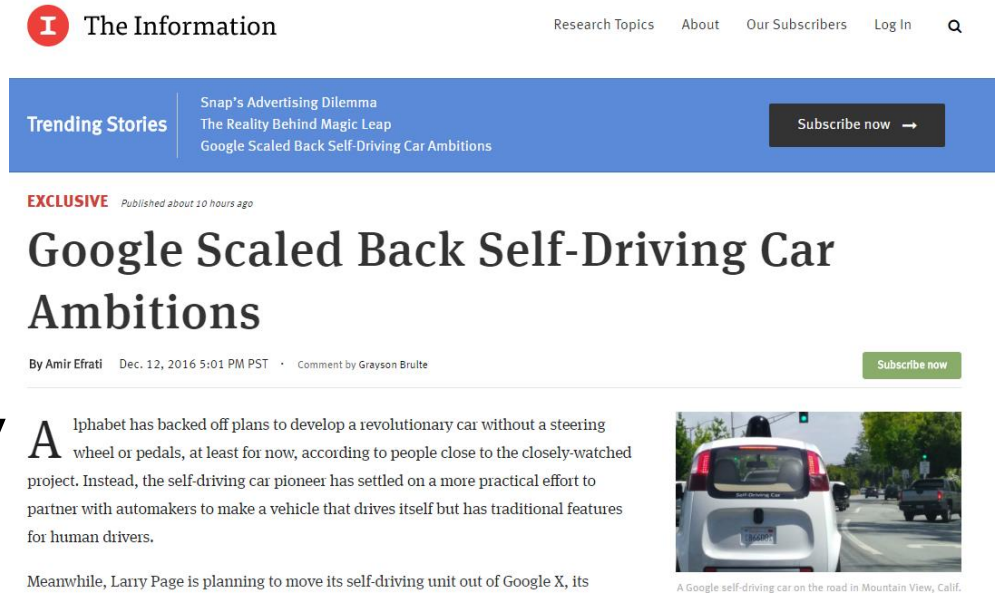
The “Chicken or Egg Problem” of AI-HPC Infrastructures



- “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

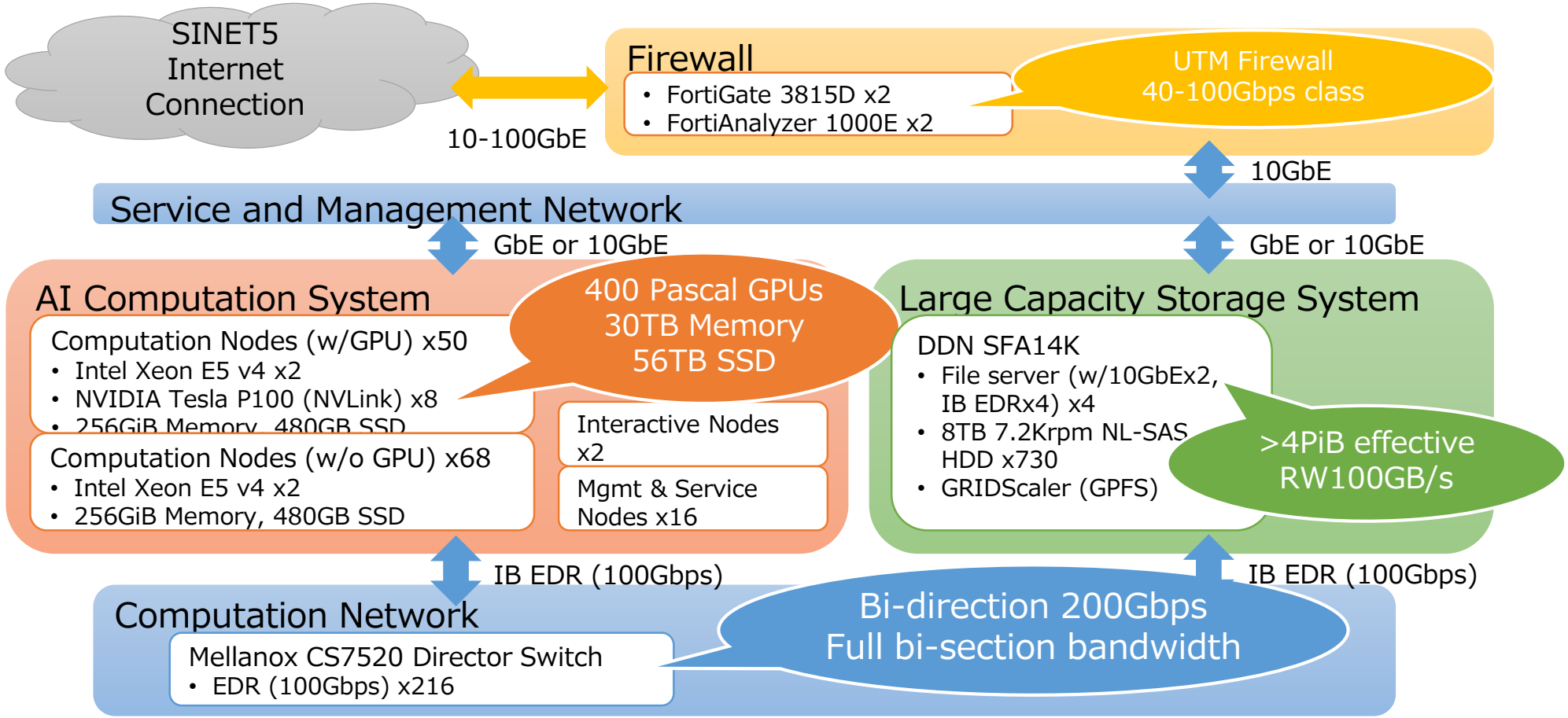
Alphabet has backed off plans to develop a revolutionary car without a steering wheel or pedals, at least for now, according to people close to the closely-watched project. Instead, the self-driving car pioneer has settled on a more practical effort to partner with automakers to make a vehicle that drives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

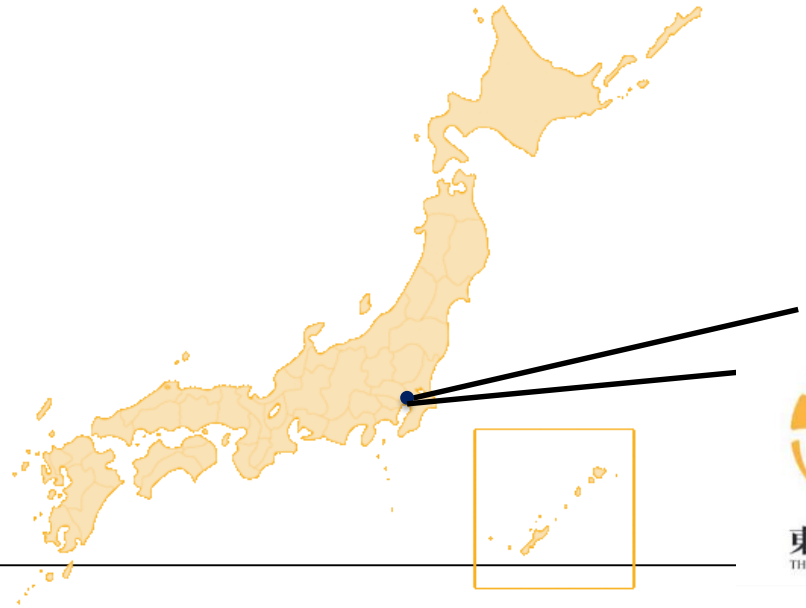
- **400x NVIDIA Tesla P100s and Infiniband EDR** accelerate various AI workloads including ML (Machine Learning) and DL (Deep Learning).
- Advanced data analytics leveraged by **4PiB shared Big Data Storage and Apache Spark** w/ its ecosystem.



METI AIST-AIRC ABCI

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

- **ABCI**: AI Bridging Cloud Infrastructure
  - Top-Level SC compute & data capability (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 2017Q3~Q4

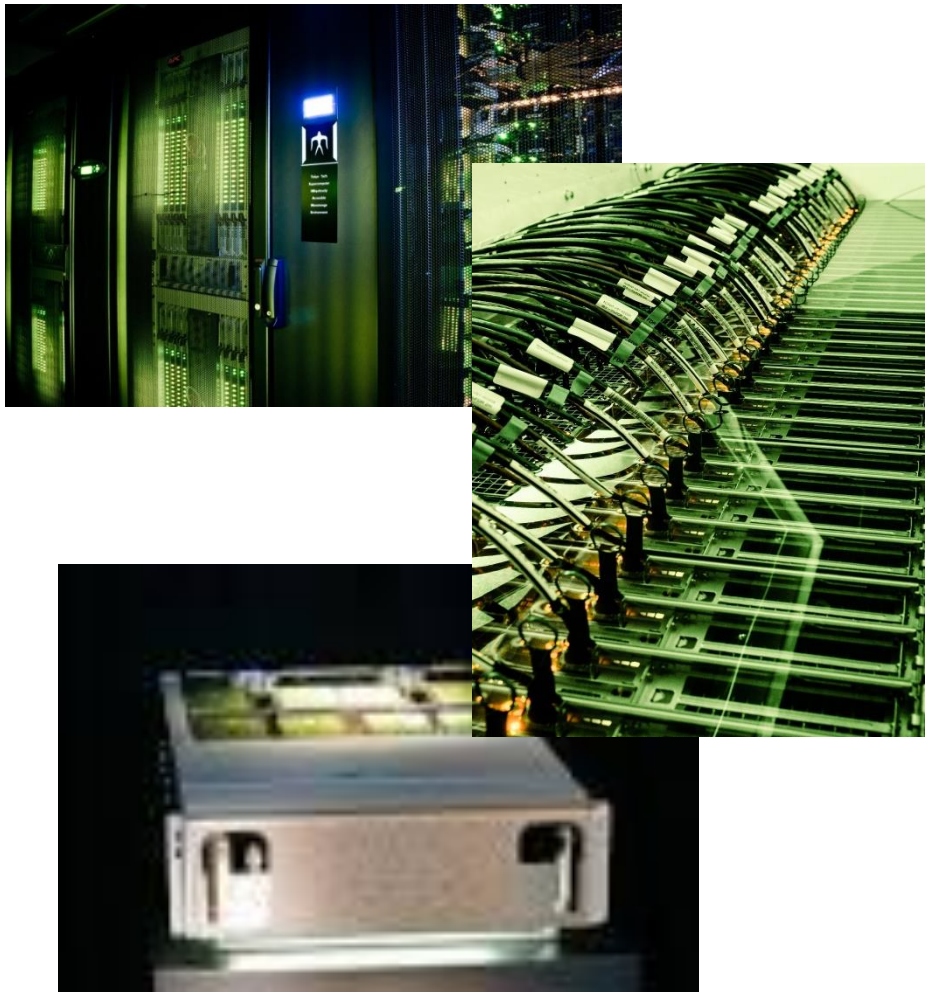
東京大学 THE UNIVERSITY OF TOKYO

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

Univ. Tokyo Kashiwa Campus

ABCI – 2017Q4~ 2018Q1

- **Extreme computing power**
  - w/ **130~200 AI-PFlops** for AI, ML, DL
  - **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.**



“SC Accelerated” Cloud IDC for AI

- **Ultra-dense IDC design from *ground-up***
  - Custom inexpensive lightweight “warehouse” building w/ substantial earthquake tolerance
  - ***Revolutionize traditional IDCs to accommodate commoditized SCs for AI, x10~x20 density***
- **Cloud ecosystem**
  - Big Data and HPC standard software stacks
- **Extreme green – >60KW/rack, PUE<1.05**
  - Intra-room Pod-based scalable design, liquid and air-cooled nodes can be mixed
  - Ambient warm liquid cooling, large Li-ion battery storage, and high-efficiency power supplies, etc.
- **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

National Institute for Advanced Industrial Science and Technology (AIST)

Matsuoka will be appointed 15% to AIST AI-OIL starting summer

ラボ長（産総研研究職 or 東工大教員/クロスアポ）
- 副ラボ長（産総研研究職）
- 副ラボ長（産総研事務職）
- ラボ研究主幹（産総研研究職）
- ラボ構成員

Director
Satoshi Matsuoka

GSIC

Ministry of Economics Trade and Industry (METI)

Resources and Acceleration of AI / Big Data, systems research

Tsubame 3.0/2.5 Big Data /AI resources

Joing Organization@Odaiba

AIST Artificial Intelligence Research Center (AIRC)

Application Area
Natural Langauge Processing
Robotics
Security

Joint Research on AI / Big Data and applications

AIST-TokyoTech AI/Big Data Open Innovation Laboratory (OIL) Will start Feb 20, 2017

School of Information Sciene and Engineering

Industrial Collaboration in data, applications

Basic Research in Big Data / AI algorithms and methodologies

Other Big Data / AI research organizations and proposals

Industry

DENSO IT LABORATORY, INC.

Software Ecosystem for HPC in AI

Different SW Ecosystem between HPC and AI/BD/Cloud
How to achieve convergence—for real, for rapid tech transfer

Existing Clouds

BD/AI User Applications

Machine Learnig MLlib/ Mahout/Chainer | **Graph Processing GraphX/ Giraph /ScaleGraph** | **SQL/Non-SQL Hive/Pig**

Java · Scala · Python + IDL

MapReduce Framework Spark/Hadoop

RDB PostgresQL | **CloudDB/NoSQL Hbase/Cassandra/MondoDB**

Distributed Filesysem HDFS & Object Store | **Coordination Service ZooKeeper**

VM(KVM), Container(Docker), Cloud Services (OpenStack)

Linux OS

Ethernet TOR Swtiches High Latency/Low Capacity NW | **Local Node Storage** | **x86 CPU**

Application Layer

- Cloud Jobs often Interactive w/resource control REST APIs
- HPC Jobs are Batch-Oriented, resource control by MPI

System Software Layer

- Cloud employs High Productivity Languages but performance neglected, focus on data analytics and dynamic frequent changes
- HPC employs High Performance Languages but requires Ninja Programmers, low productivity. Kernels & compilers well tuned & result shared by many programs, less rewrite
- Cloud focused on databases and data manipulation workflow
- HPC focused on compute kernels, even for data processing. Jobs scales to thousands of jobs, thus debugging and performance tuning
- Cloud requires purpose-specific computing/data environment as well as their mutual isolation & security
- HPC requires environment for fast & lean use of resources, but on modern machines require considerable system software support

OS Layer

Hardware Layer

- Cloud HW based on Web Server "commodity" x86 servers, distributed storage on nodes assuming REST API access
- HPC HW aggressively adopts new technologies such a s GPUs, focused on ultimate performance at higher cost, shared storage to support legacy apps

Existing Supercomputers

HPC User Code

Numerical Libraries LAPACK, FFTW | **Various DSLs** | **Workflow Systems**

Fortran · C · C++ + IDL

MPI · OpenMP/ACC · CUDA/OpenCL

Parallel Debuggers and Profilers

Parallel Filesystem Lustre, GPFS, | **Batch Job Schedulers PBS Pro, Slurm, UGE**

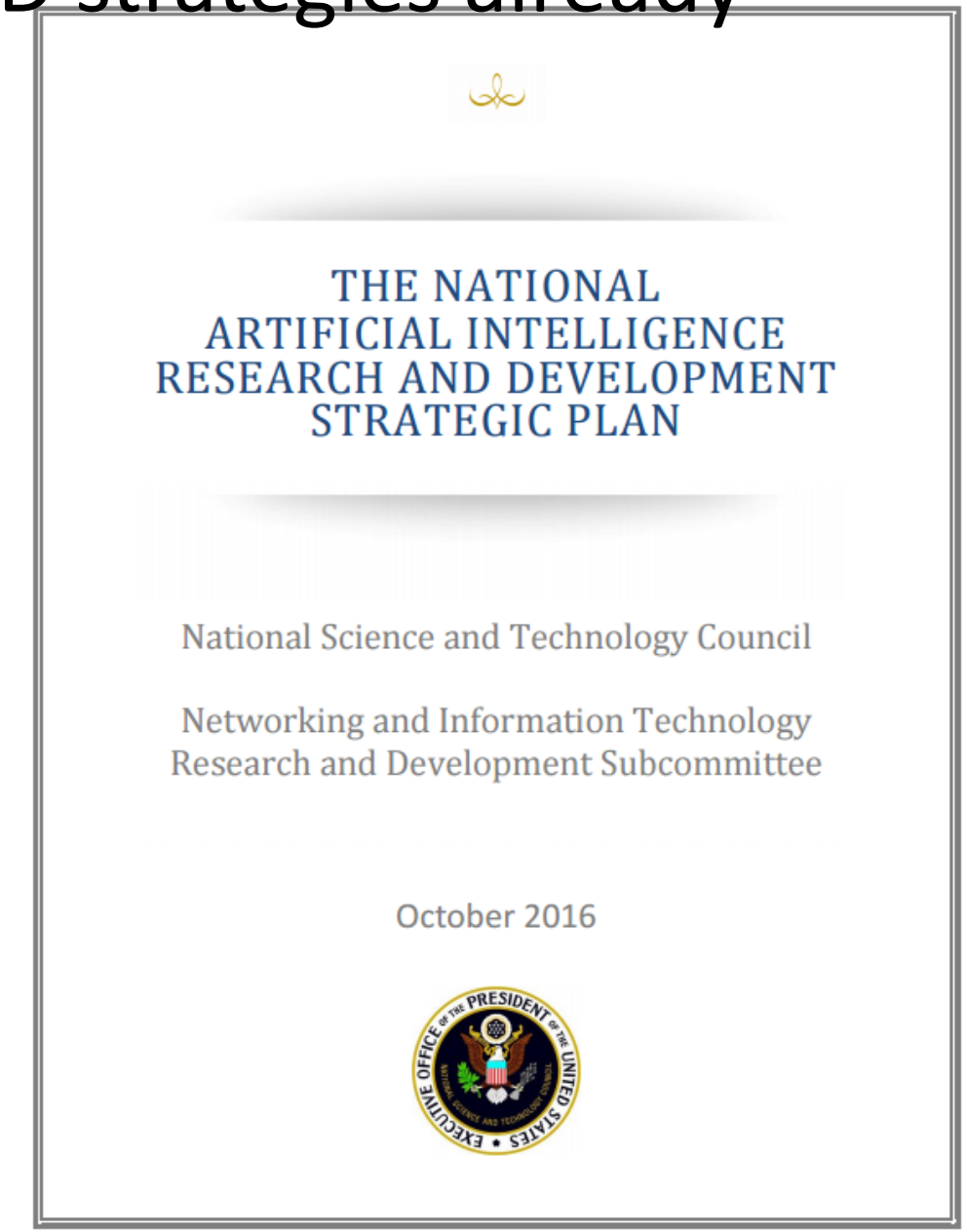
Linux OS

InfiniBand/OPA High Capacity Low Latency NW | **High Performance SAN+Burst Buffers** | **X86 + Accelerators e.g. GPUs, FPGAs**

Various convergence research efforts underway but no realistic converged SW Stack yet => achieving HPC – AI/BD/Cloud convergence key ABCI goal

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

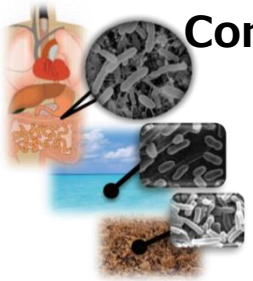
- Strategy 5: Develop shared public datasets and environments for AI training and testing. The depth, quality, and accuracy of training datasets and resources significantly affect AI performance. Researchers need to develop high quality datasets and environments and enable responsible access to high-quality 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.

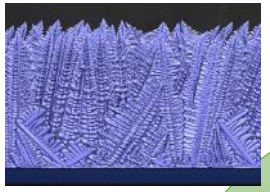


Co-Design of BD/ML/AI with HPC using BD/ML/AI - for survival of HPC

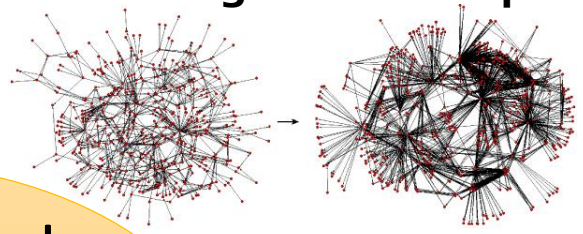
Acceleration and Scaling of BD/ML/AI via HPC and Technologies and Infrastructures

Accelerating Conventional HPC Apps





Large Scale Graphs



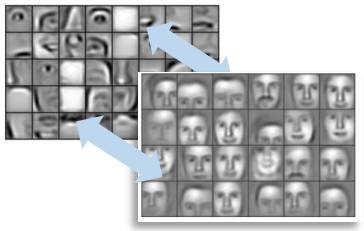
Big Data AI-Oriented Supercomputers

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

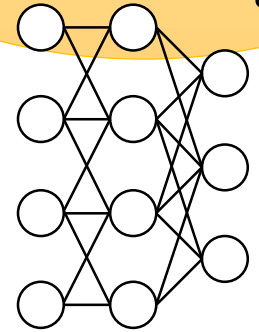
Image and Video



Future Big Data・AI Supercomputer Design

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



Robots / Drones

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

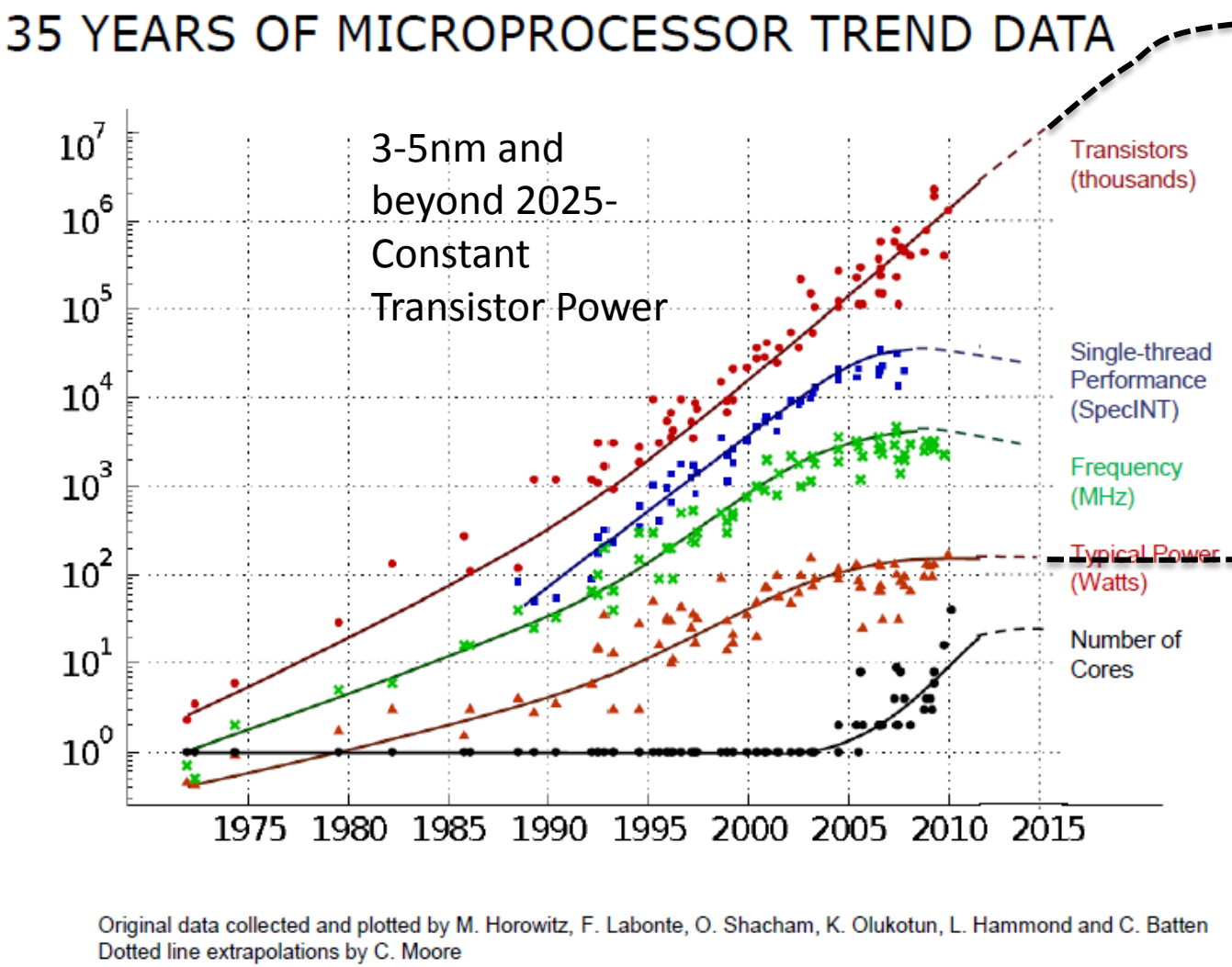
- Much of underlying IT performance growth due to Moore's law
  - "LSI: x2 transistors in 1~1.5 years"
  - Causing qualitative "leaps" in IT and societal innovations
  - The main reason we have supercomputers and Google…
- But this is slowing down & ending, by mid 2020s…!!!
  - End of Lithography shrinks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

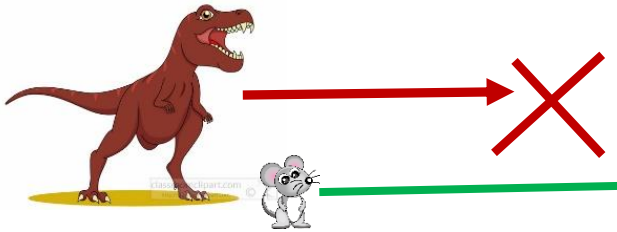
The “curse of constant transistor power”
- Ignorance of this is like ignoring global warming -

- 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

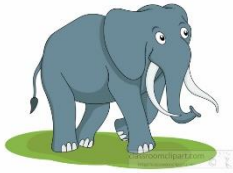
Performance growth via ***data-centric computing:***
"From FLOPS to BYTES"

- *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

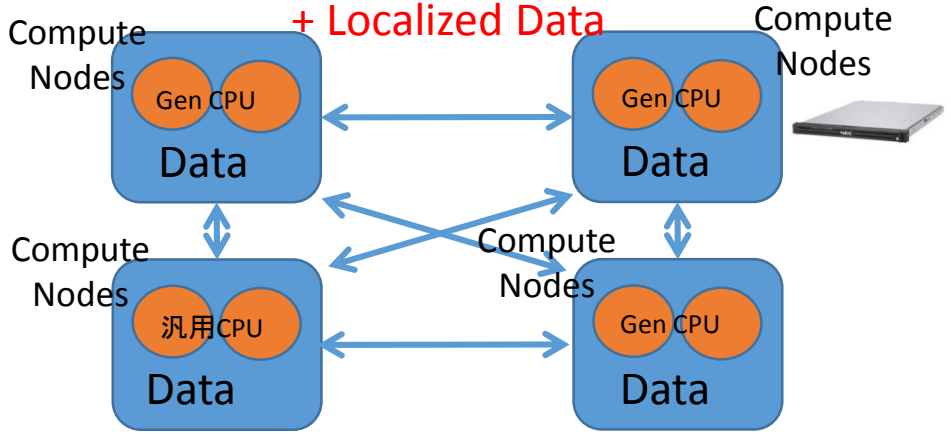


Flops-Centric Algorithms and Apps

Flops-Centric System Software

Hardware/Software System APIs
Flops-Centric Massively Parallel Architecture

Homogeneous General Purpose Nodes
+ Localized Data



Loosely Coupled with Electronic Interconnect

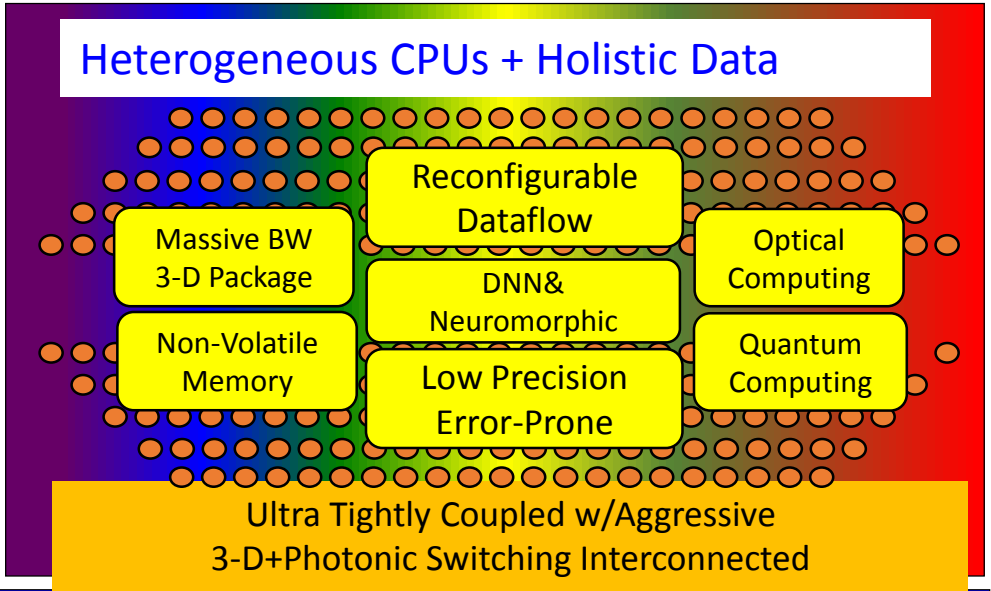
Transistor Lithography Scaling
(CMOS Logic Circuits, DRAM/SRAM)

~2025
M-P Extinction Event

Bytes-Centric Algorithms and Apps

Bytes-Centric System Software

Hardware/Software System APIs
Data-Centric Heterogeneous Architecture

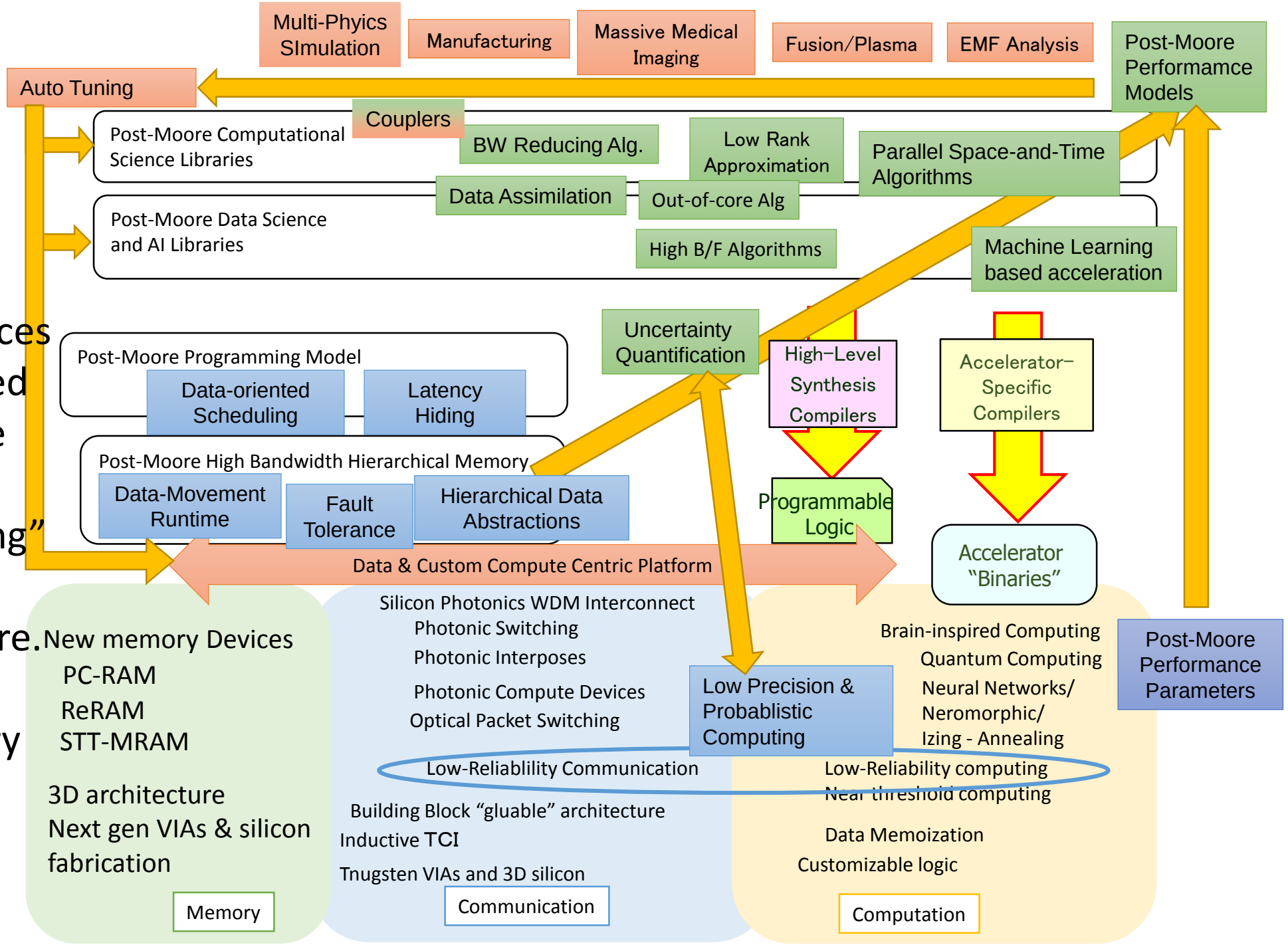


Novel Devices + CMOS (Dark Silicon)
(Nanophotonics, Non-Volatile Devices etc.)

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

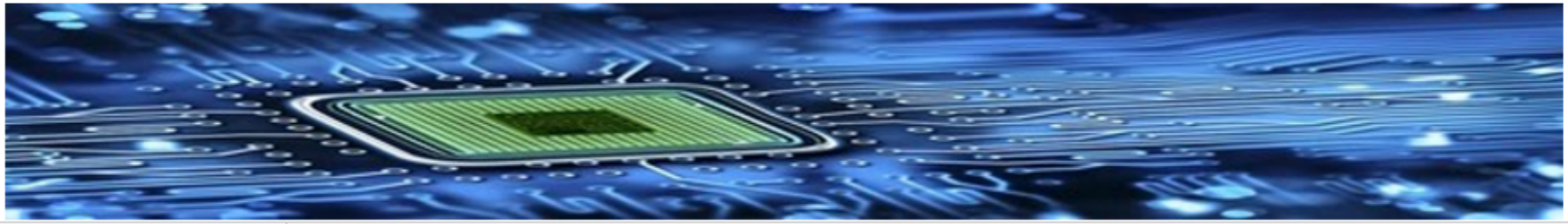
Device & arch. advances improving data-related parameters over time

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



Post Moore Era Supercomputing Workshop @ SC16

- https://sites.google.com/site/2016pmes/
- Jeff Vetter (ORNL), Satoshi Matsuoka (Tokyo Tech) et. al.



Search this sit

2016 Post-Moore's Era Supercomputing (PMES) Workshop Home

News

Call For Position Papers - Submission Deadline - June 17

Invited Speakers

Photos

Program

Resources

Workshop Venue

Sitemap

188

2016 Post-Moore's Era Supercomputing (PMES) Workshop Home

Co-located with SC16 in Salt Lake City
Monday, 14 November 2016

Workshop URL: http://j.mp/pmes2016
CFP URL: http://j.mp/pmes2016cfp
Submission URL (EasyChair): http://j.mp/pmes2016submissions
Submission questions: pmes16@easychair.org

This interdisciplinary workshop is organized to explore the scientific issues, challenges, and opportunities for supercomputing beyond the scaling limits of

News

PMES Submission Site Now Open!
PMES Workshop Confirmed for SC16!
Submissions open for PMES Position Papers on April 17

Important Dates

- Submission Site Opens: 17 April 2016