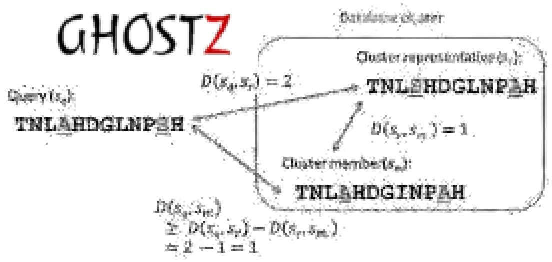
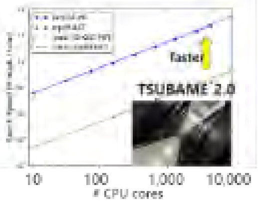
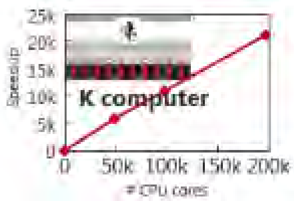
HPC and BD/AI Convergence Example [Yutaka Akiyama, Tokyo Tech]

Genomics

Ultra-fast Seq. Analysis

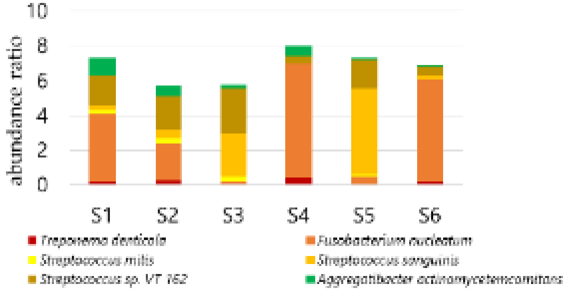






- Suzuki *et al.*, *Bioinformatics* (2015)
- Suzuki *et al.*, *PLOS ONE* (2016)

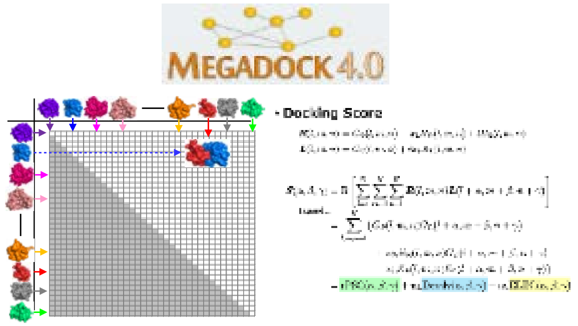
Oral/Gut Metagenomics



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

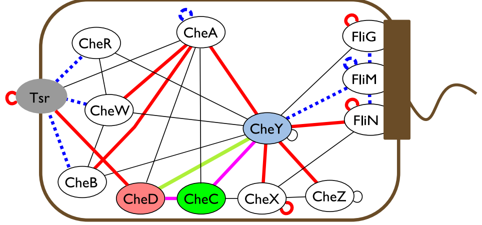
Protein-Protein Interactions

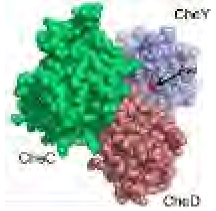
Exhaustive PPI Prediction System



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

Pathway Predictions

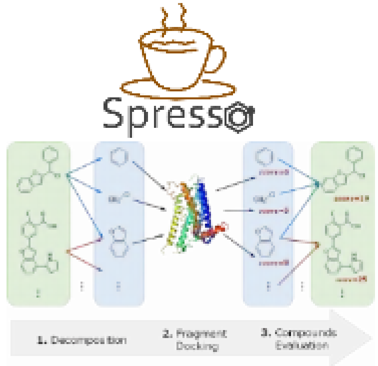




- Matsuzaki *et al.*, *Protein Pept Lett* (2014)

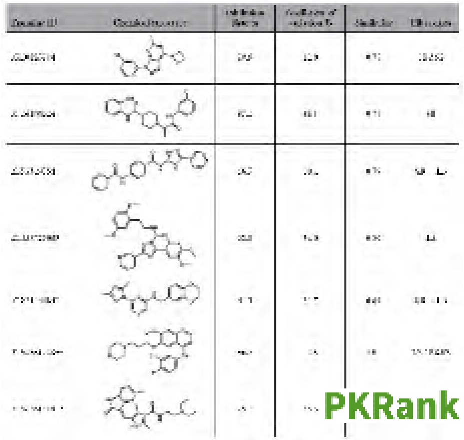
Drug Discovery

Fragment-based Virtual Screening



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

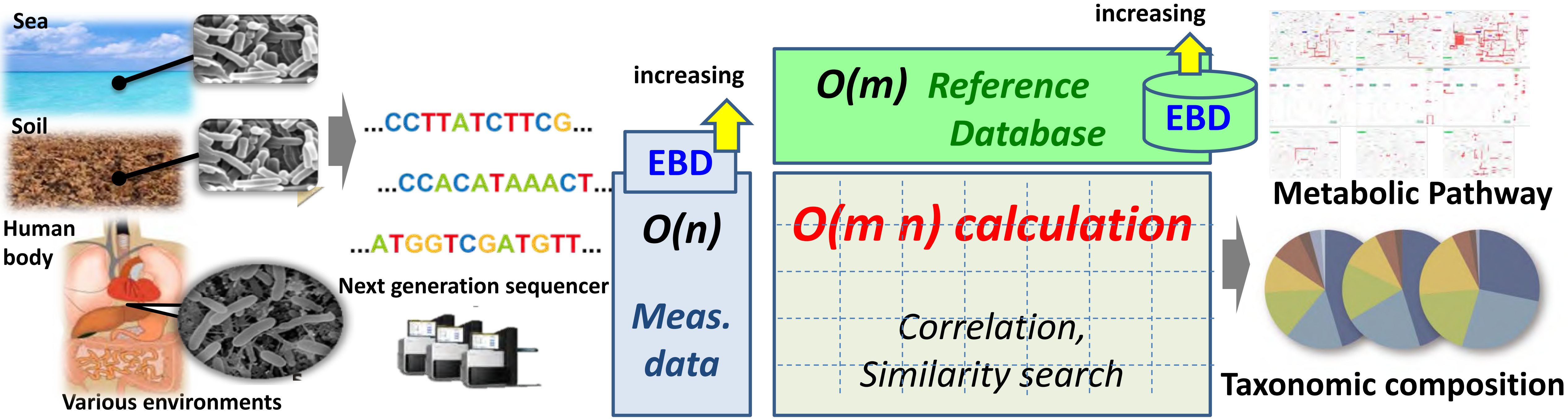
Learning-to-Rank VS



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

EBD vs. EBD : Large Scale Homology Search for Metagenomics

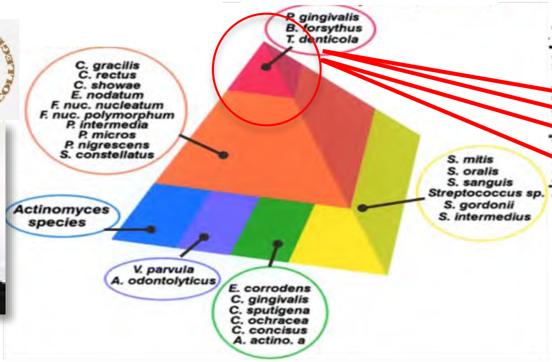
- Revealing uncultured microbiomes and finding novel genes in various environments
- Applied for human health in recent years

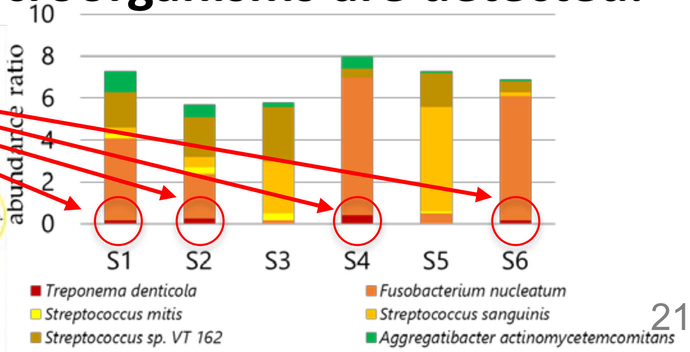


Metagenomic analysis of periodontitis patients

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

High risk microorganisms are detected.



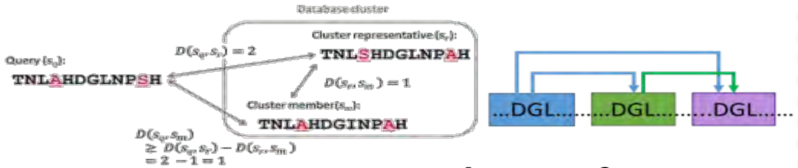


Development of Ultra-fast Homology Search Tools

GHOSTZ

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

Subsequence sequence clustering



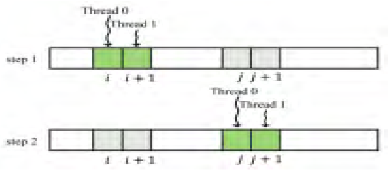
computational time for 10,000 sequences (sec.) (3.9 GB DB、1CPU core)

×240 faster than conventional algorithm

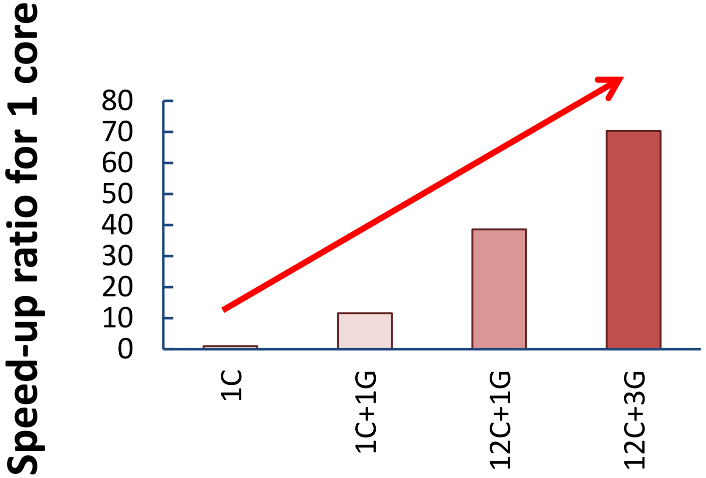
GHOSTZ-GPU

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

Multithread on GPU



TSUBAME 2.5 Thin node GPU



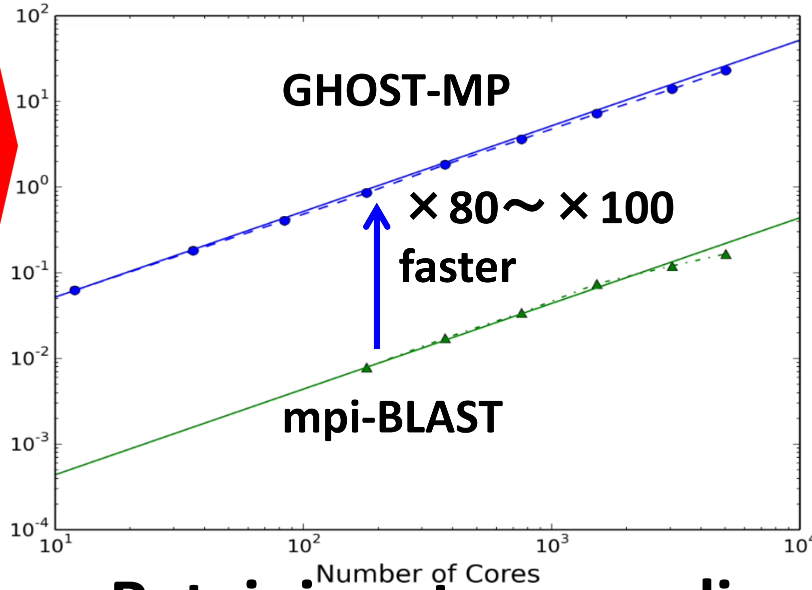
×70 faster than 1 core using 12 cores + 3 GPUs

GHOST-MP

Kakuta, *et al*. (submitted)

MPI + OpenMP hybrid pallelization

TSUBAME 2.5



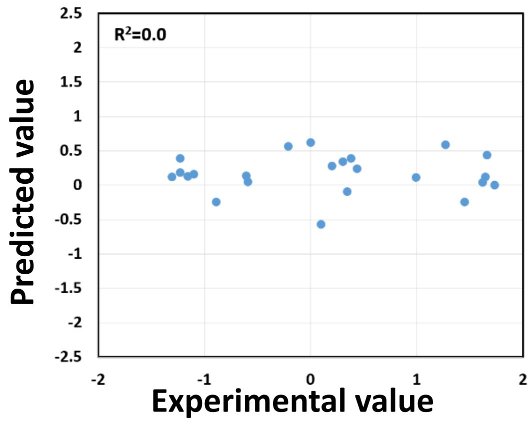
Retaining strong scaling up to 100,000 cores

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

Problems

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

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

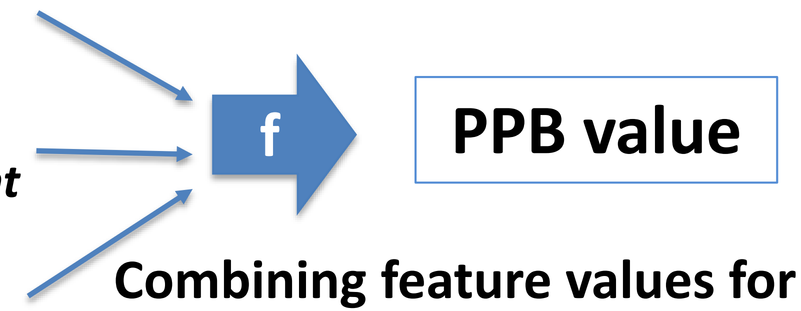


Previous PPB prediction software for small molecule can not predict peptide PPB

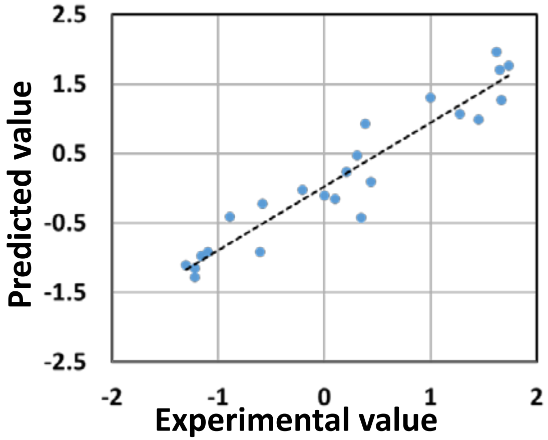
Solutions

Compute Feature Values (more than 500 features)

LogS
LogP
:
MolWeight
:
SASA
polarity



Combining feature values for building a predictive model



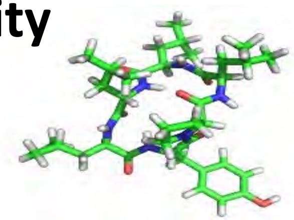
R^2 = 0.905

A constructed model can explain peptide PPB well

Molecular Dynamics Simulation for Membrane Permeability
Application for peptide drug discovery

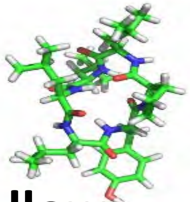
Problems

1) Single residue mutation can drastically change membrane permeability



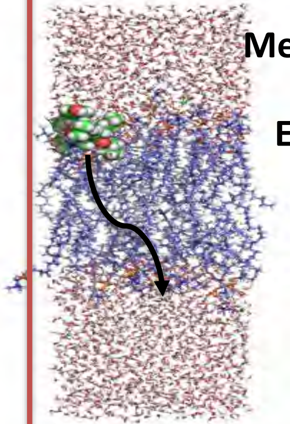
Sequence : D-Pro, D-Leu, D-Leu, L-Leu, D-Leu,
Membrane permeability : 7.9 × 10 $^{-6}$cm/s

× 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.

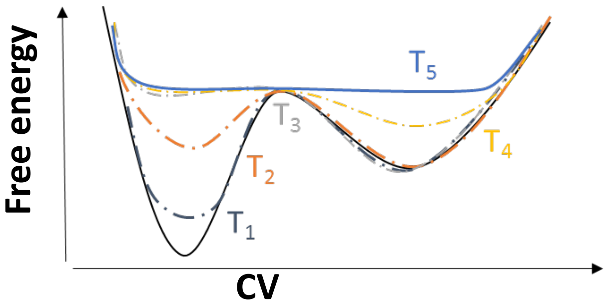
Ex) Membrane thickness : 40 Å
Peptide membrane permeability : 7.9 × 10^{-6} cm/s

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)

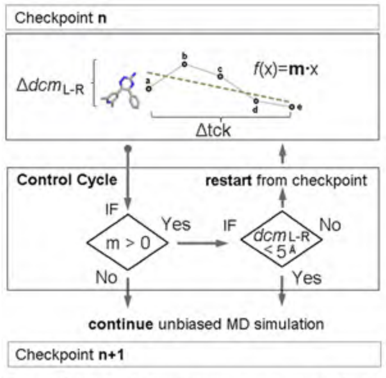
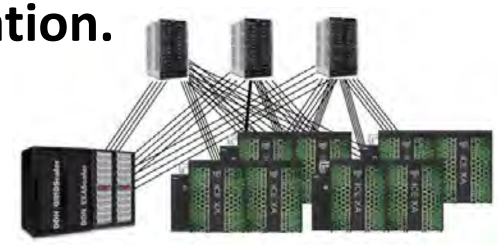


Figure 1. Scheme of the ligand–receptor distance vector (dcm_{L-R}) supervision algorithm implemented in the supervised molecular dynamics (SuMD) technique.

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

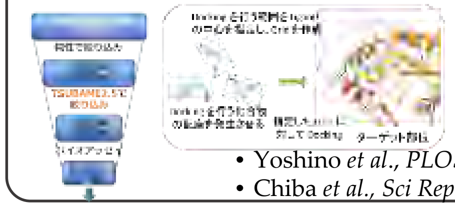
①Drug Target selection system

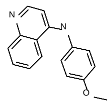
Minister of Health, Labour and Welfare Award of the 11th annual Merit Awards for Industry-Academia-Government Collaboration

Drugs for NTDs

②Glide-based Virtual Screening

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

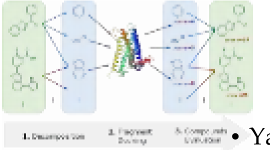




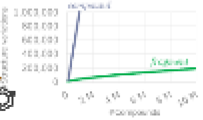
- Yoshino *et al.*, *PLOS ONE* (2015)
- Chiba *et al.*, *Sci Rep* (2015)

③Novel Algorithms for fast virtual screening against huge databases

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



Spresso

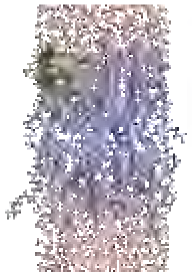


- Yanagisawa *et al.*, *GIW* (2016)

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)

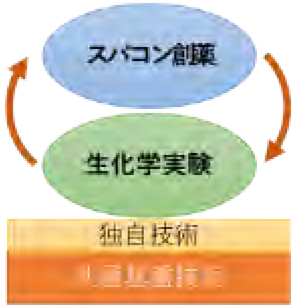
PeptiDream

Catalyst Inc.
(株)カタリスト

DiscoveryResource
ディスカヴァリソース(株)

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) Municiple Govt (Kawasaki), JP & US Pharma

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

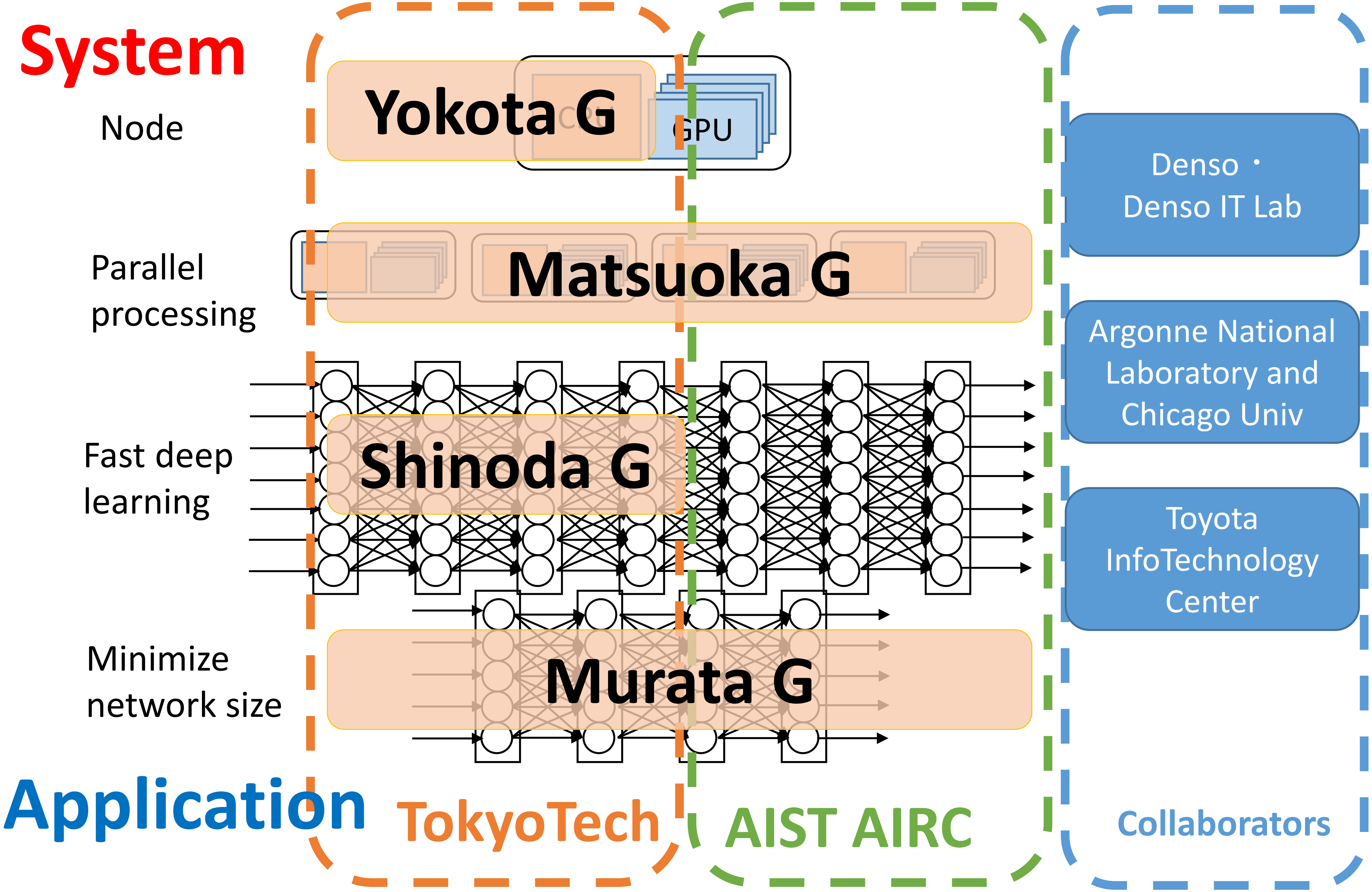
- Video processing in smart society for safety and security
  - Intelligent transport systems
    Drive recorder video
  - Security systems
    Surveillance camera video
- Deep learning
  - Much higher performance than before
  - IT giants with large computational resources has formed a monopoly

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

Reference

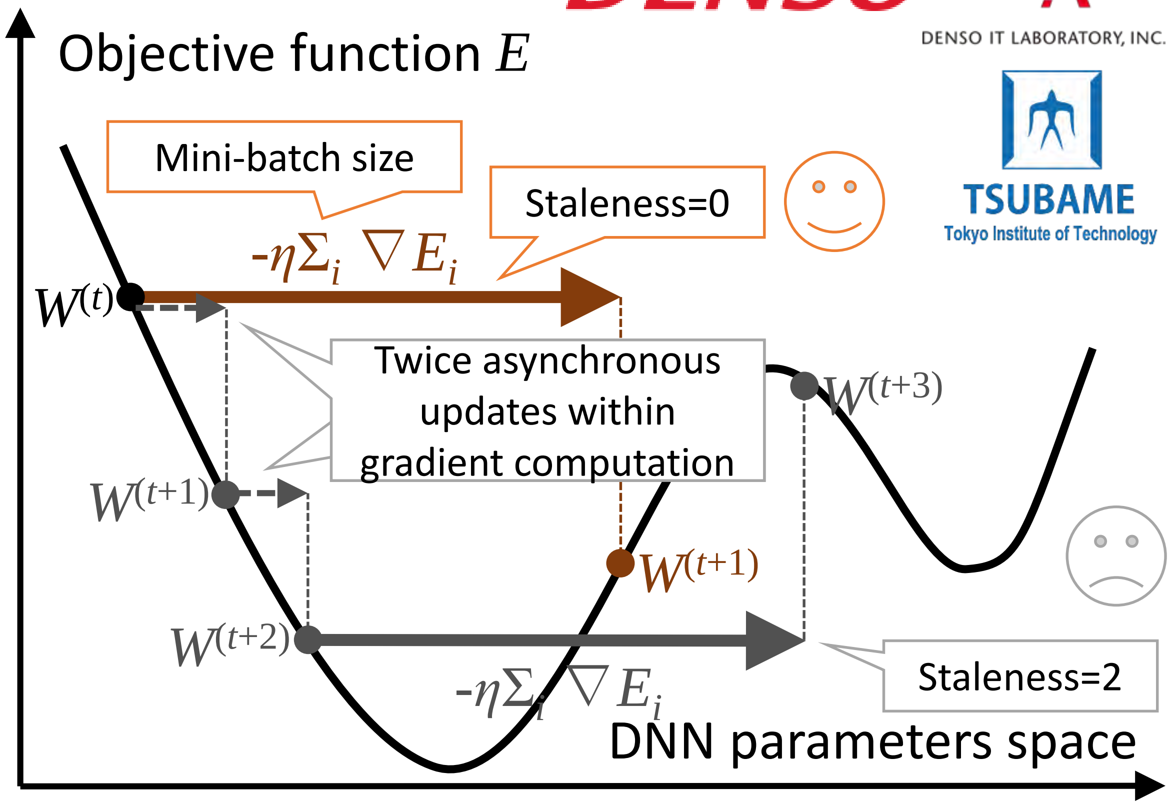
Research team



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

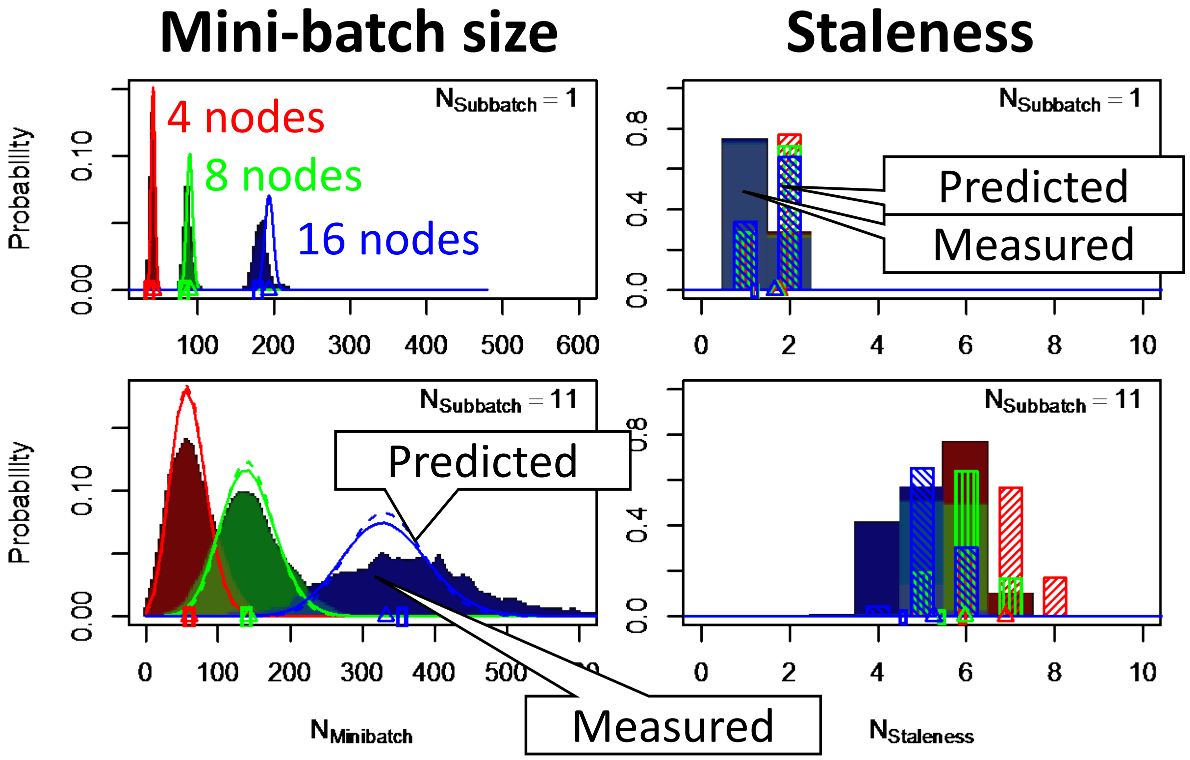
Background

- 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

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

National Institute for Advanced Industrial Science and Technology (AIST)

独立行政法人 産業技術総合研究所

経済産業省 Ministry of Economy, Trade and Industry

Ministry of Economics Trade and Industry (METI)

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

Resources and Acceleration of AI / Big Data, systems research

REAL WORLD BIGDATA RWBC-OIL

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

Director: Satoshi Matsuoka

AIST Artificial Intelligence Research Center (AIRC)

Joint Research on AI / Big Data and applications

ITCS Departments

Basic Research in Big Data / AI algorithms and methodologies

Application Area
Natural Langauge Processing
Robotics
Security

ABCI
AI Bridging Cloud Infrastructure

Industrial Collaboration in data, applications

Industry

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

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

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)

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&Robots

TOYOTA

JARI 日本自動車研究所

YAHOO! JAPAN

SoftBank

NTT

Web access and merchandise

IoT Communication, location & other data

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

SAKURA Internet

Google Cloud Platform

Windows Azure

amazon web services

Over $1B Govt. AI investment over 10 years

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

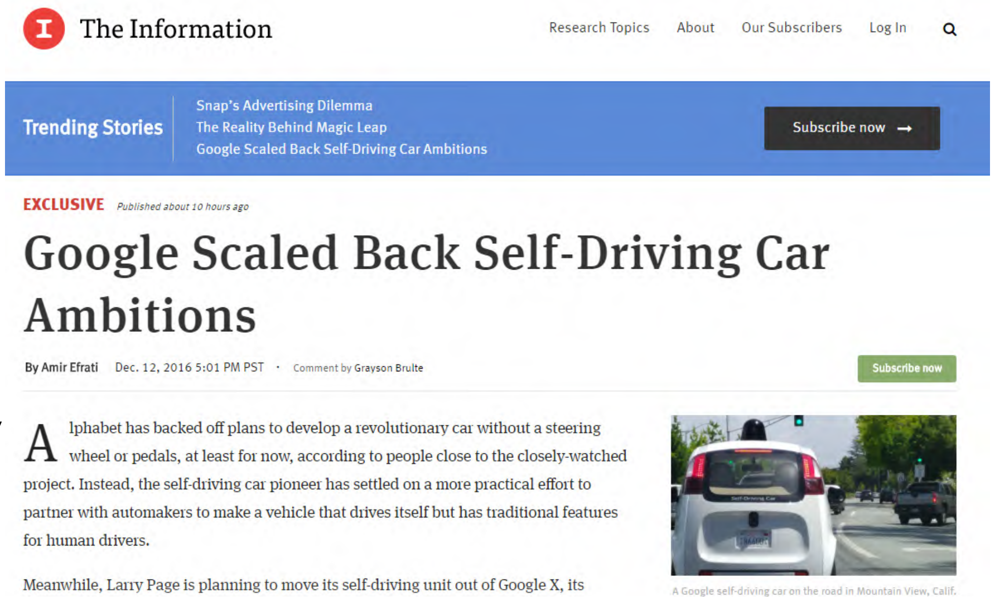
AI&Data Infrastructures

Massive "Big" Data in Training

"Big" Data

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.

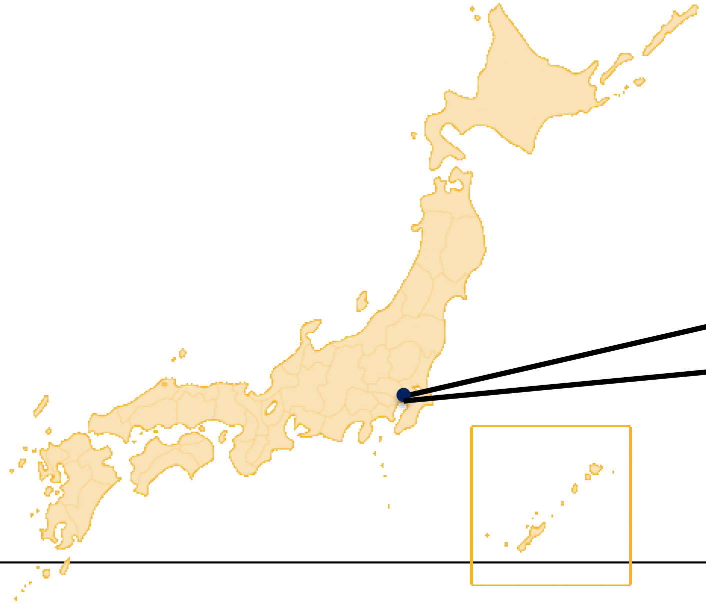
Meanwhile, Larry Page is planning to move its self-driving unit out of Google X, its

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

AIST

METI AIST-AIRC ABCI
as the *worlds first large-scale OPEN AI Infrastructure*

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



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

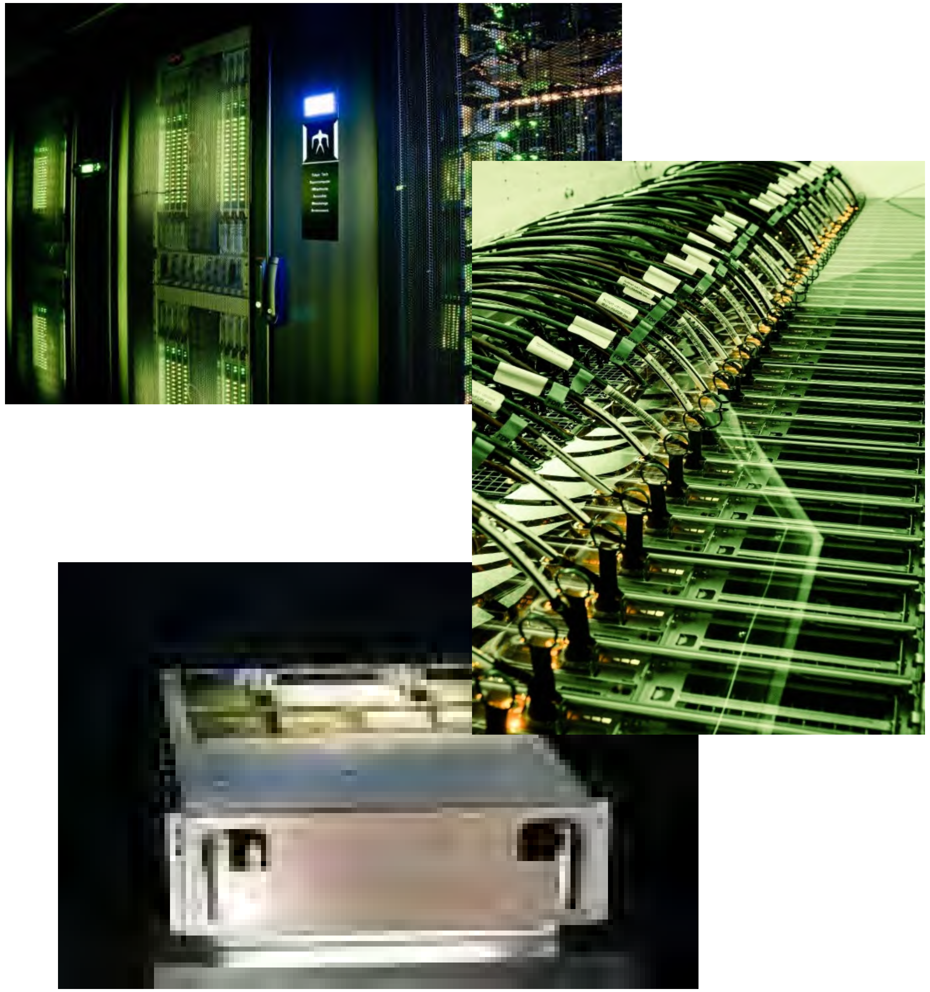
東京大学 THE UNIVERSITY OF TOKYO

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

Univ. Tokyo Kashiwa Campus

ABCI – 2017Q4~ 2018Q1

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



ABCI 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)**

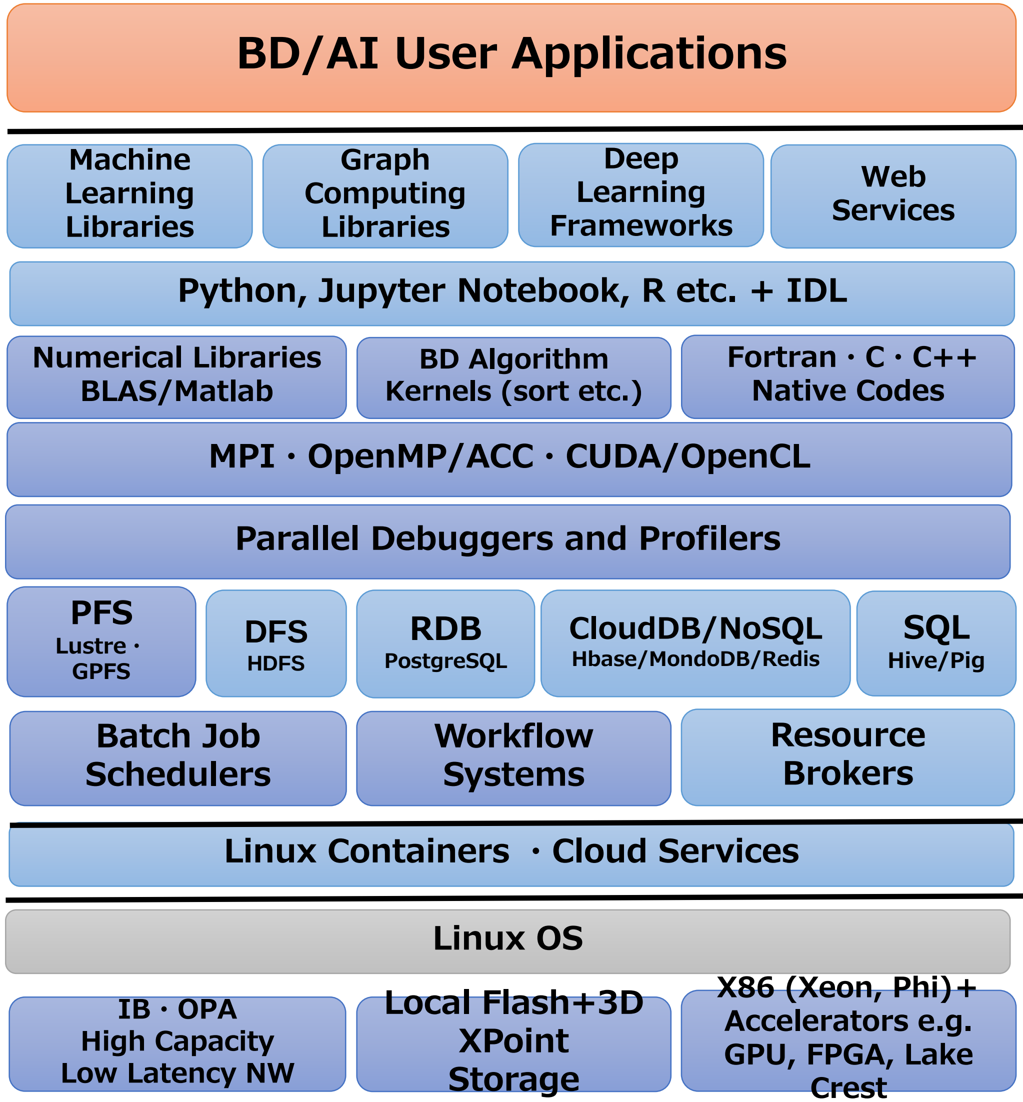
ABCI Procurement Benchmarks

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

No traditional HPC Simulation Benchmarks Except SPECCPU

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

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

Cutting Edge Research AI Infrastructures in Japan
Accelerating BD/AI with HPC
(and my effort to design & build them)

In Production

x5.8

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

Under Acceptance

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

x5.8

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

Being Manufactured

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

x2.8~4.2

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

x5.0~7.7

Draft RFC out
IDC under construction

1H 2019?
“**ExaAI**”
~1 AI-ExaFlop

Undergoing Engineering Study

~x1000 in 3.5 years

Built/funded

still under plans

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)