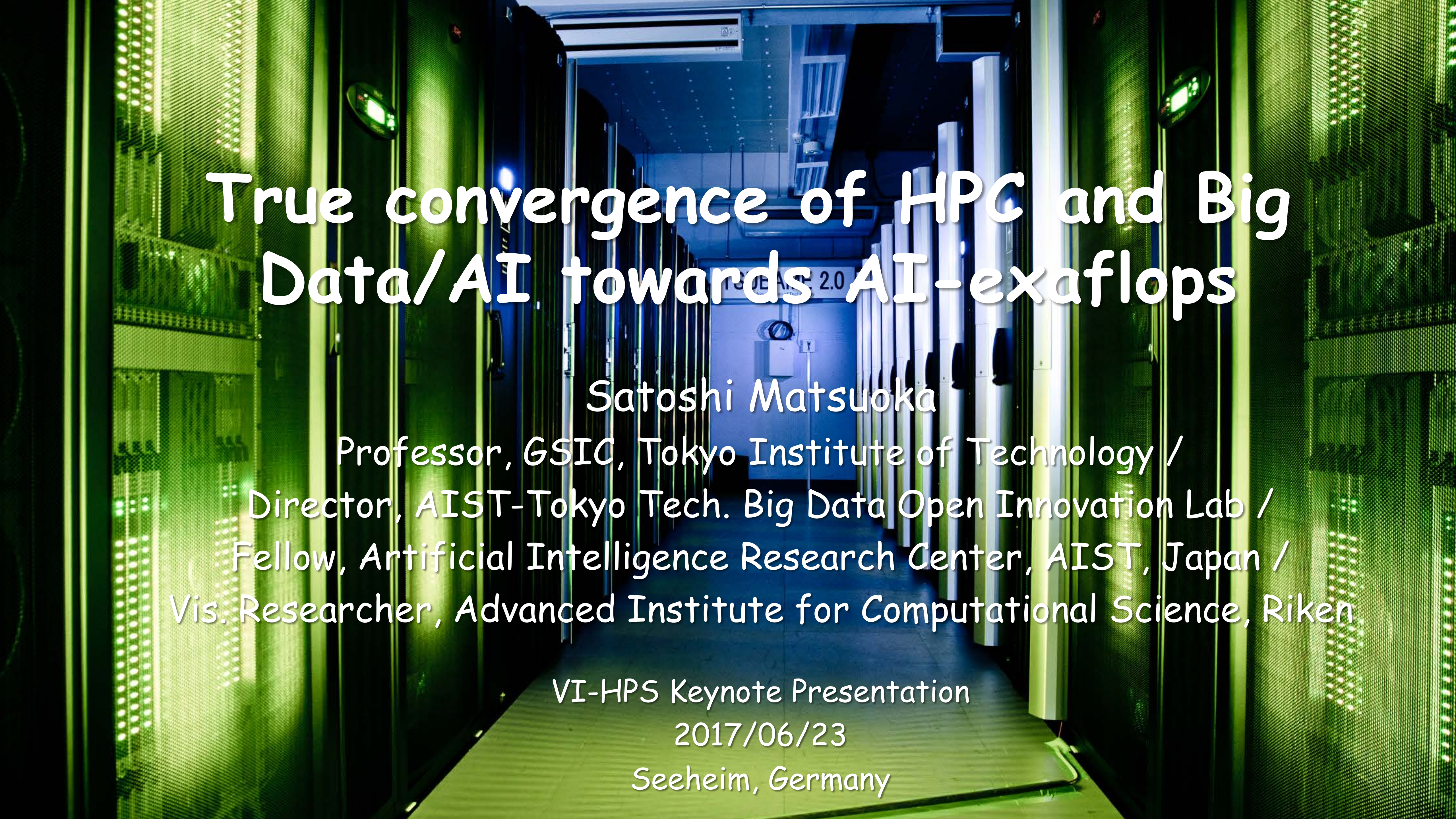
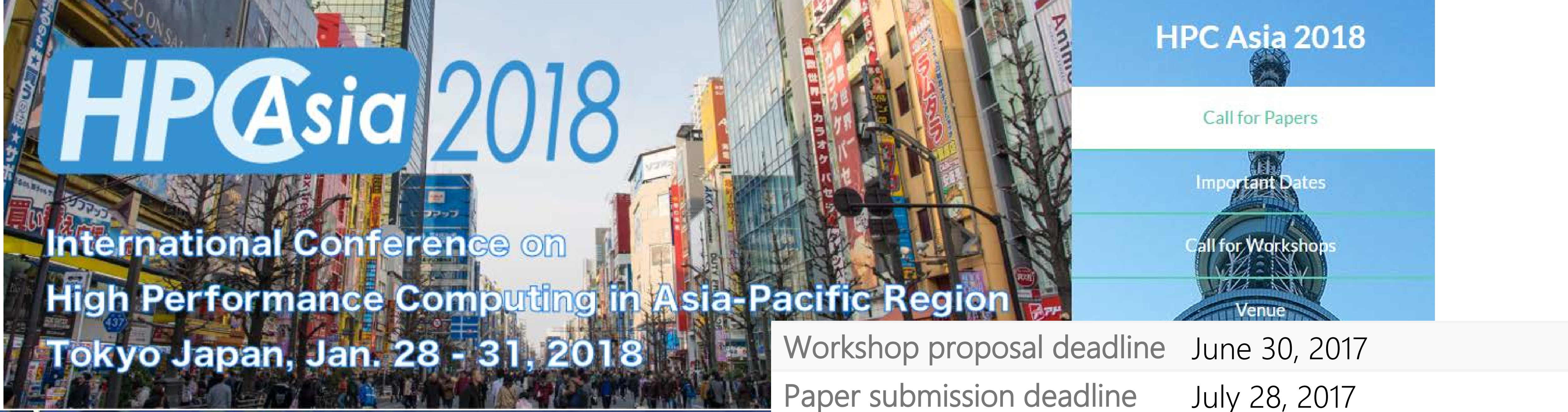


True convergence of HPC and Big Data/AI towards AI-exaflops

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

VI-HPS Keynote Presentation
2017/06/23
Seeheim, Germany

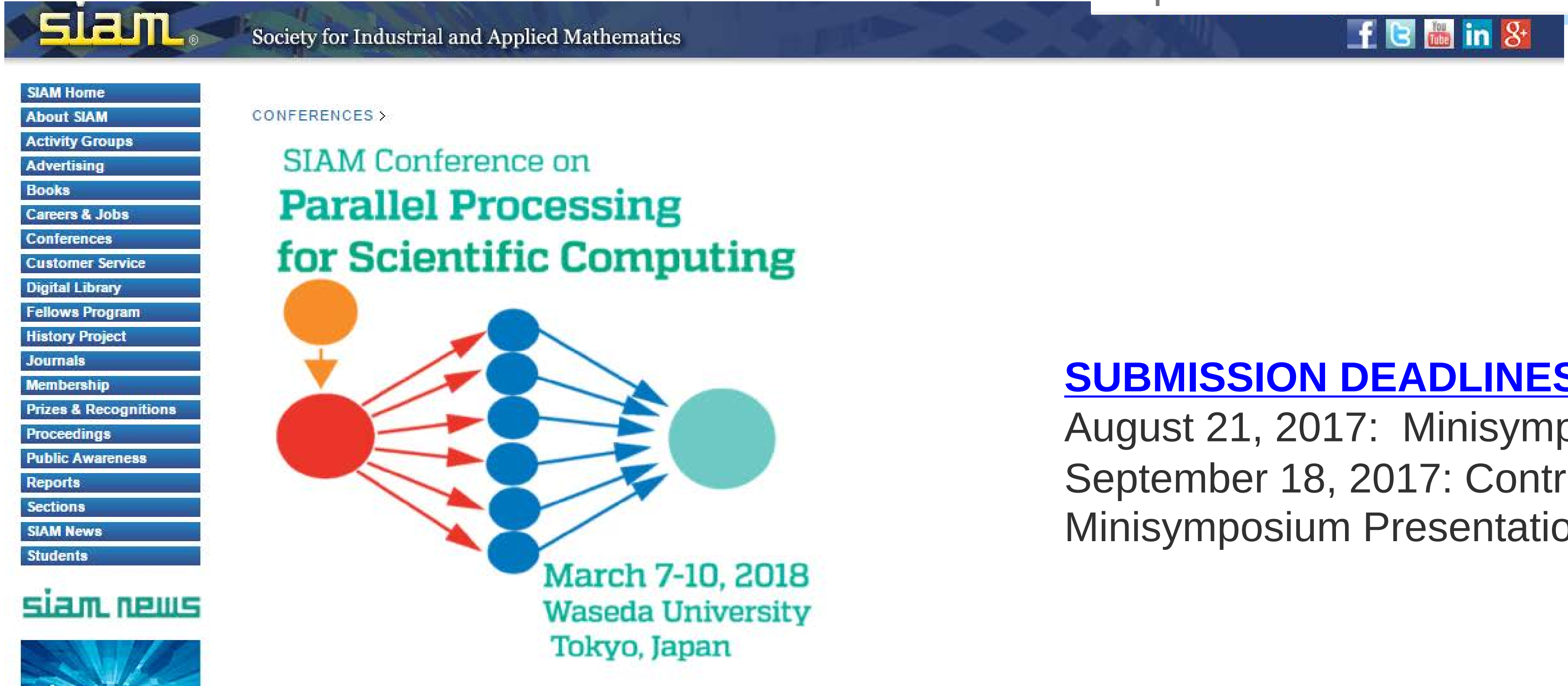


HPCAsia 2018

International Conference on High Performance Computing in Asia-Pacific Region

Tokyo Japan, Jan. 28 - 31, 2018

Workshop proposal deadline	June 30, 2017
Paper submission deadline	July 28, 2017



SIAM Conference on

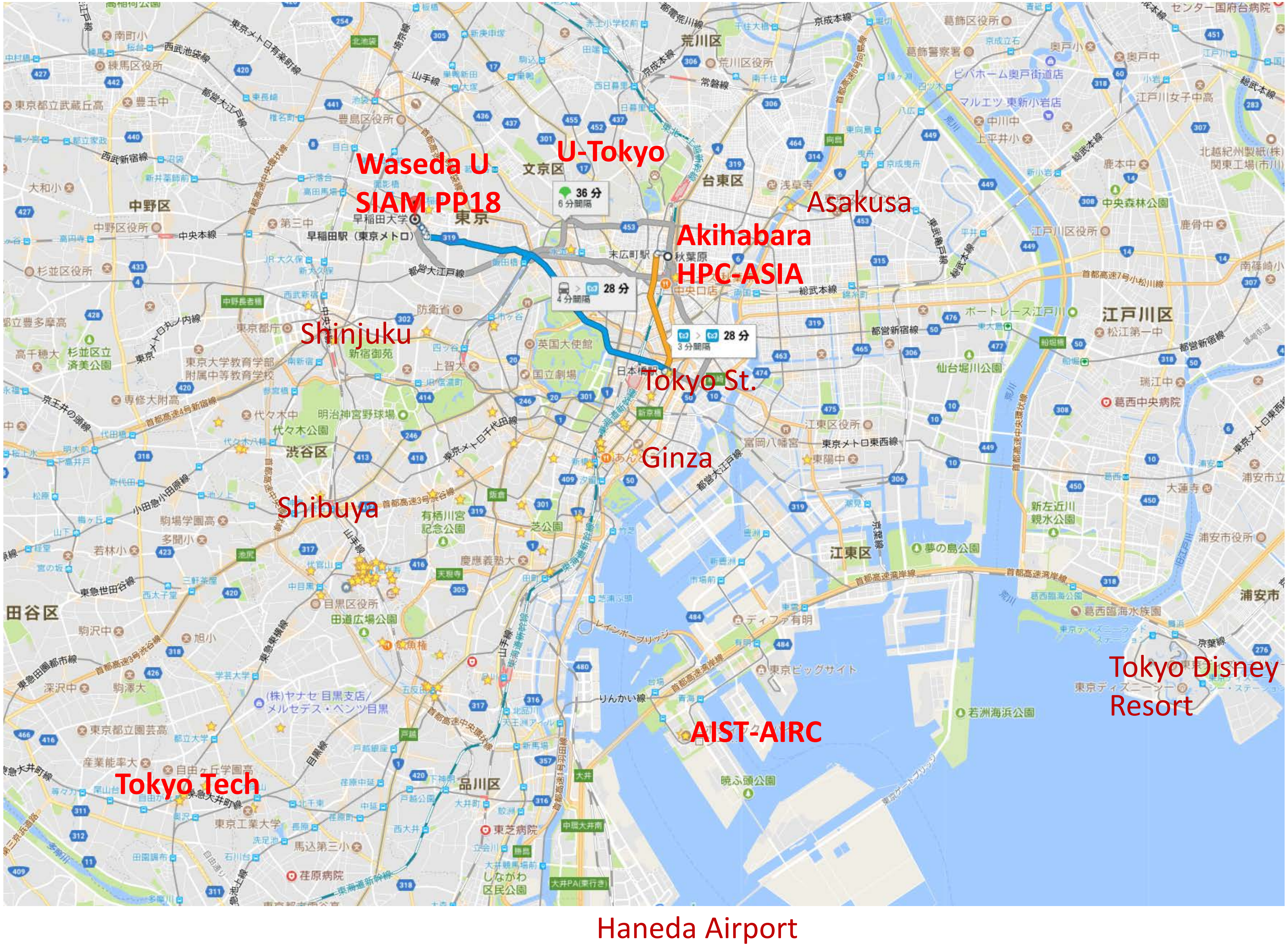
Parallel Processing for Scientific Computing

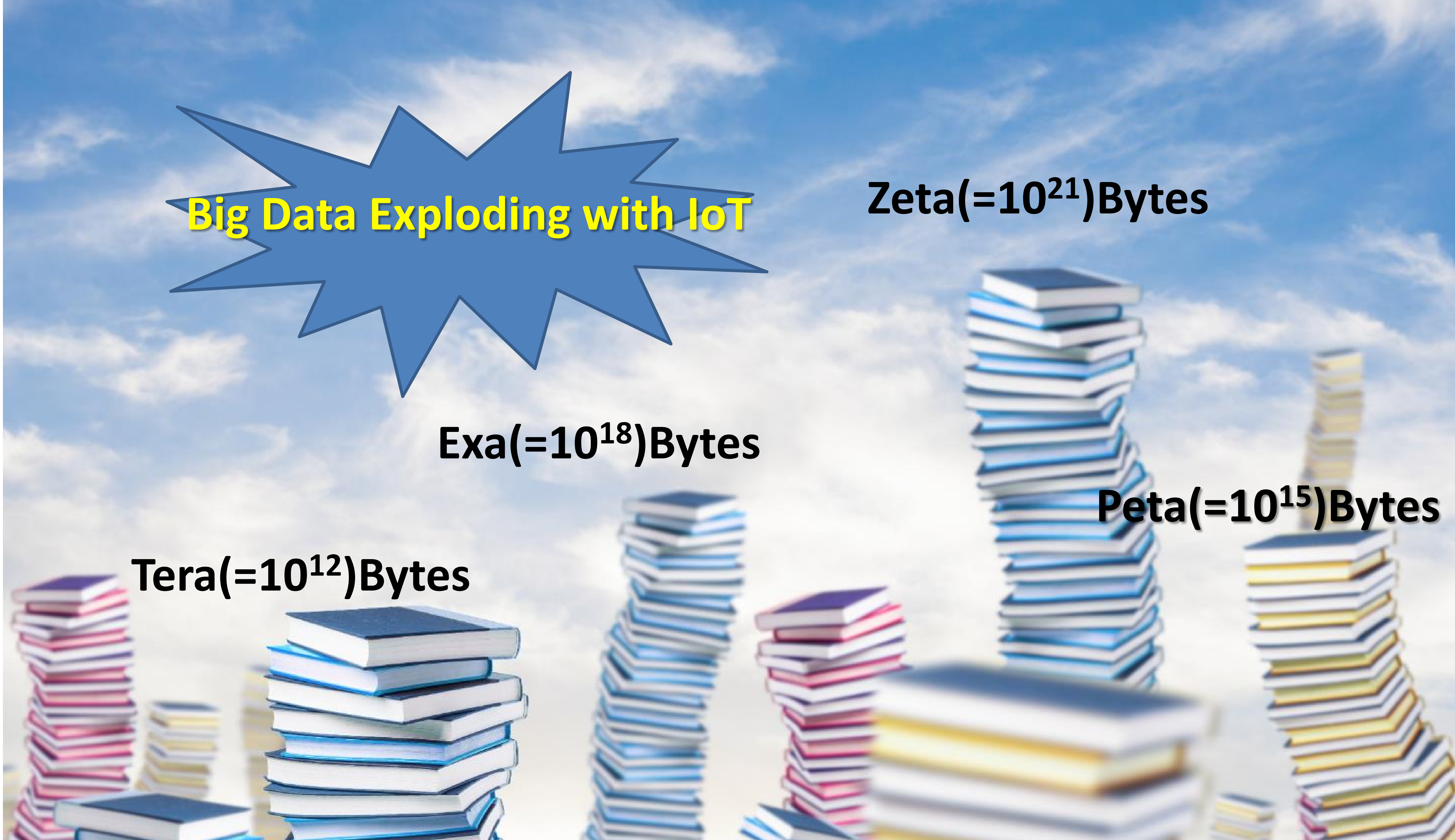
March 7-10, 2018
Waseda University
Tokyo, Japan

SUBMISSION DEADLINES

August 21, 2017: Minisymposium Proposal Submissions

September 18, 2017: Contributed Lecture, Poster and Minisymposium Presentation Abstracts


Waseda U
SIAM PP18
U-Tokyo
36 分
6 分間隔
Asakusa
Akihabara
HPC-ASIA
28 分
4 分間隔
28 分
3 分間隔
Shinjuku
Tokyo St.
Ginza
Shibuya
Tokyo Disney
Resort
AIST-AIRC
Tokyo Tech
Haneda Airport



Big Data Exploding with IoT

Zeta($=10^{21}$)Bytes

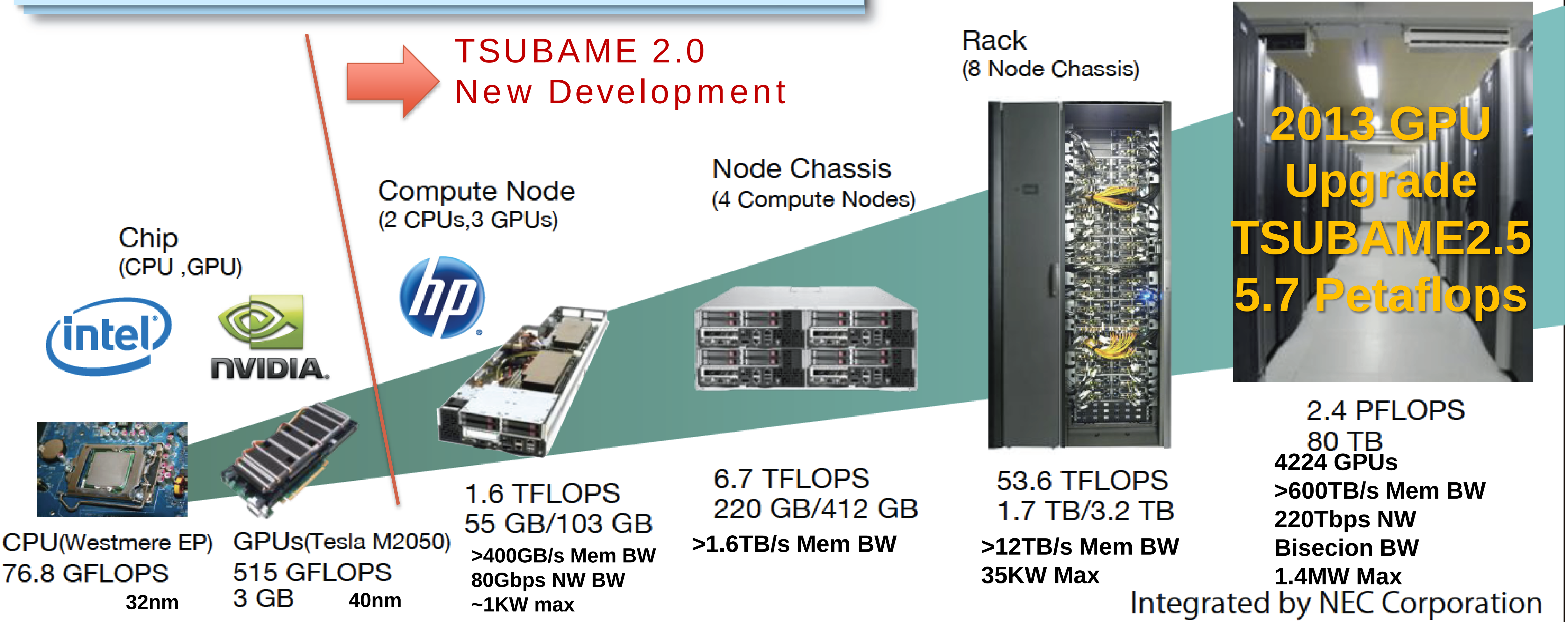
Exa($=10^{18}$)Bytes

Peta($=10^{15}$)Bytes

Tera($=10^{12}$)Bytes

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

2013 GPU Upgrade TSUBAME2.5 5.7 Petaflops

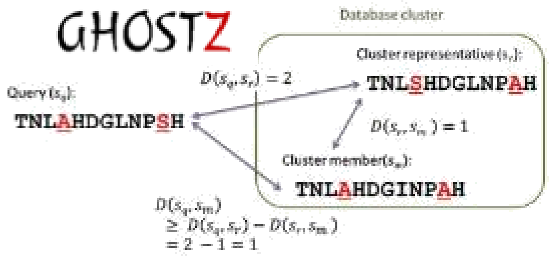
2.4 PFLOPS
80 TB
4224 GPUs
>600TB/s Mem BW
220Tbps NW
Bisecion BW
1.4MW Max

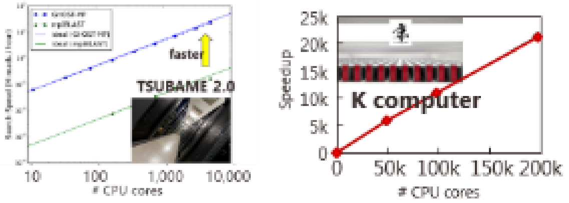
Integrated by NEC Corporation

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

Genomics

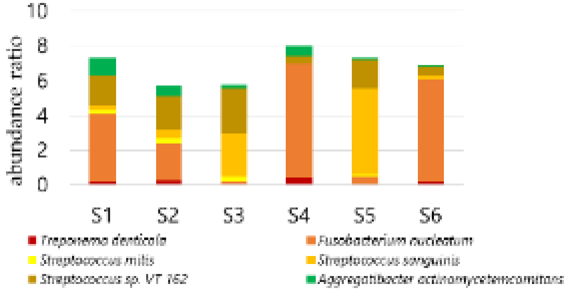
Ultra-fast Seq. Analysis





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

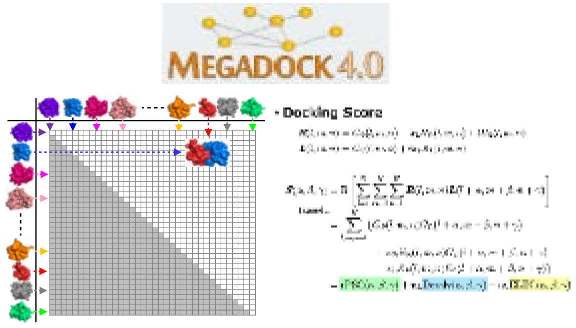
Oral/Gut Metagenomics



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

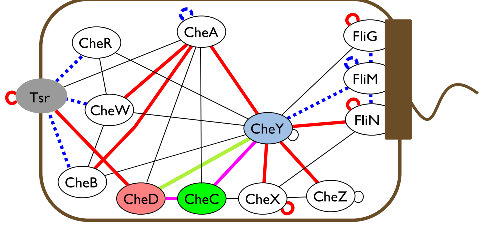
Protein-Protein Interactions

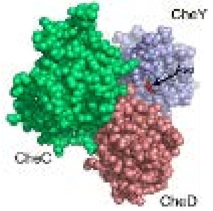
Exhaustive PPI Prediction System



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

Pathway Predictions

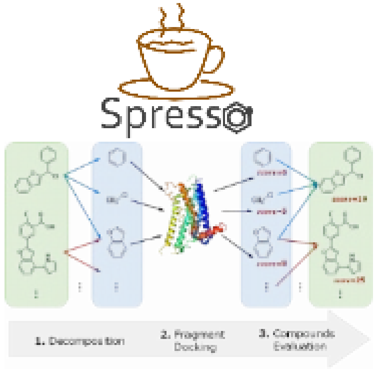




- Matsuzaki *et al.*, *Protein Pept Lett* (2014)

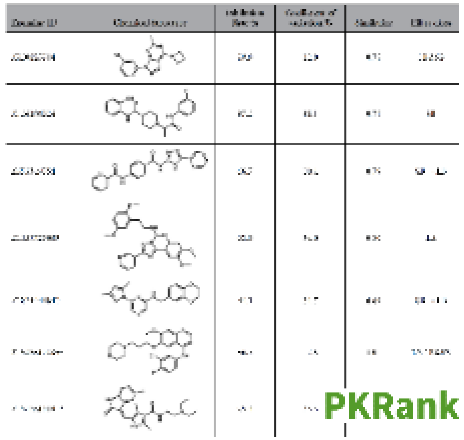
Drug Discovery

Fragment-based Virtual Screening



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

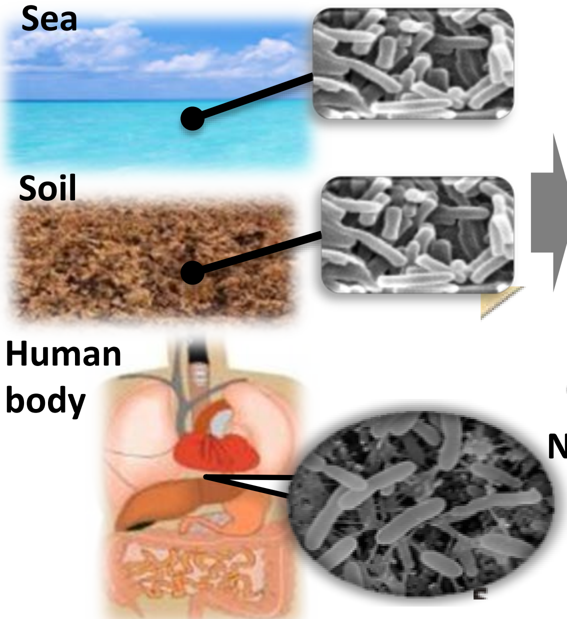
Learning-to-Rank VS



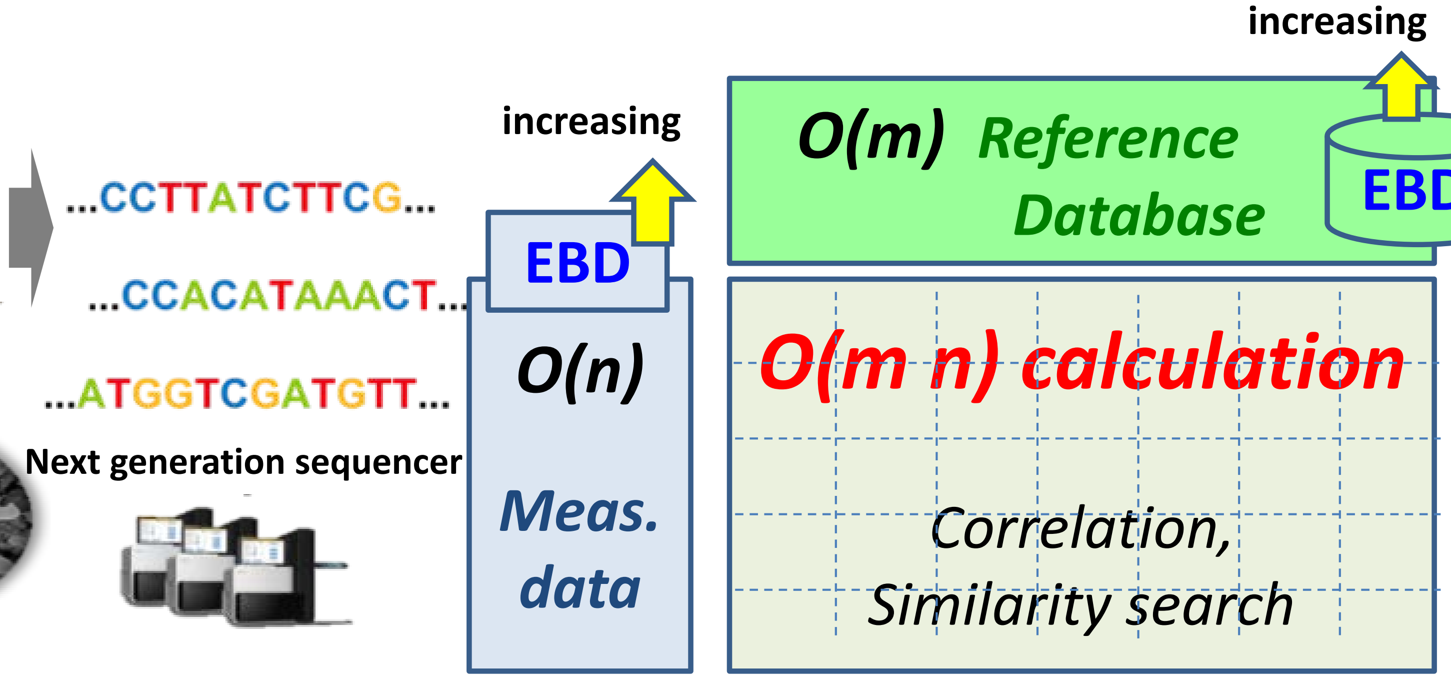
- Suzuki *et al.*, *AROB2017* (2017)

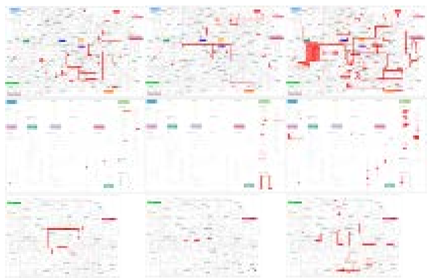
EBD vs. EBD : Large Scale Homology Search for Metagenomics

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

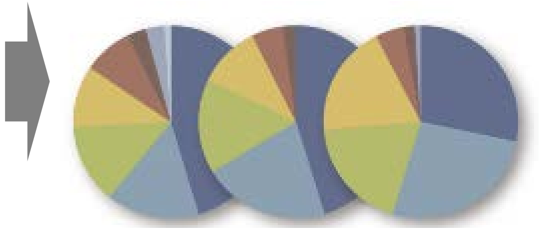


Various environments





Metabolic Pathway

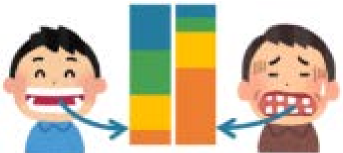


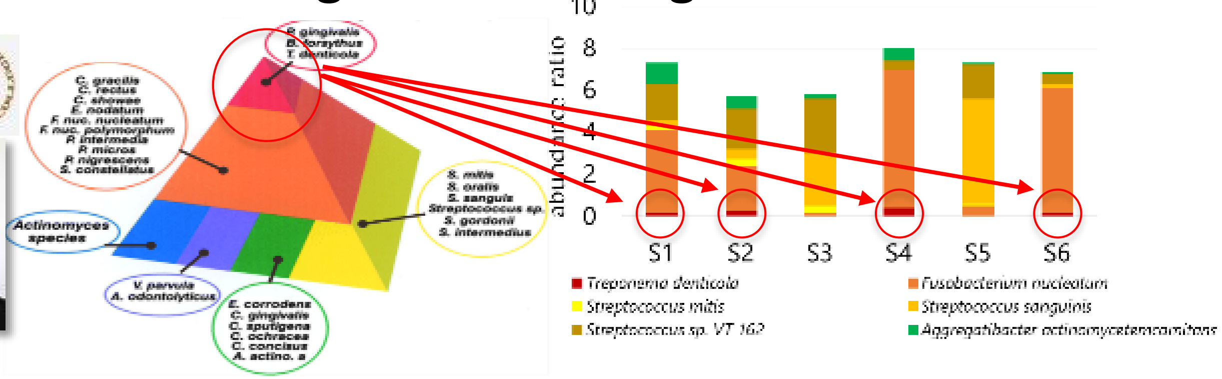
Taxonomic composition

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





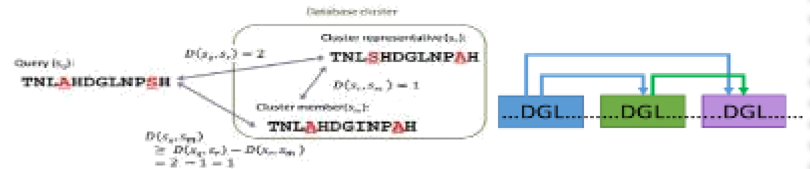
Development of Ultra-fast Homology Search Tools

x100,000 ~ x1,000,000 c.f. high-end BLAST WS (both FLOPS and BYTES)

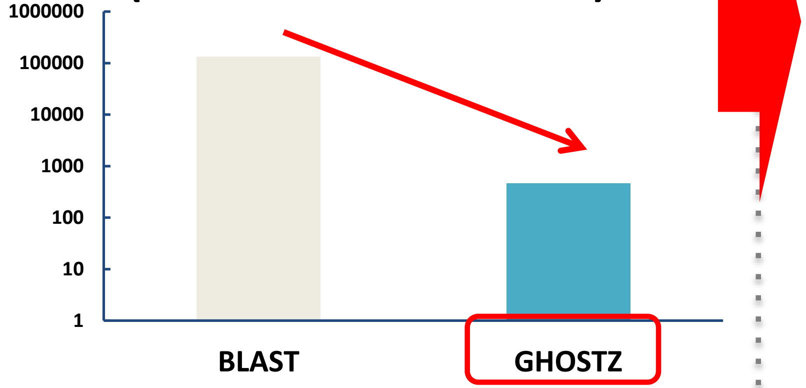
GHOSTZ

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

Subsequence sequence clustering



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

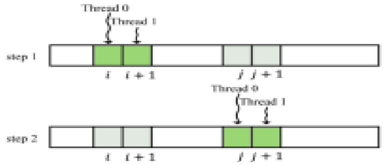


×240 faster than conventional algorithm

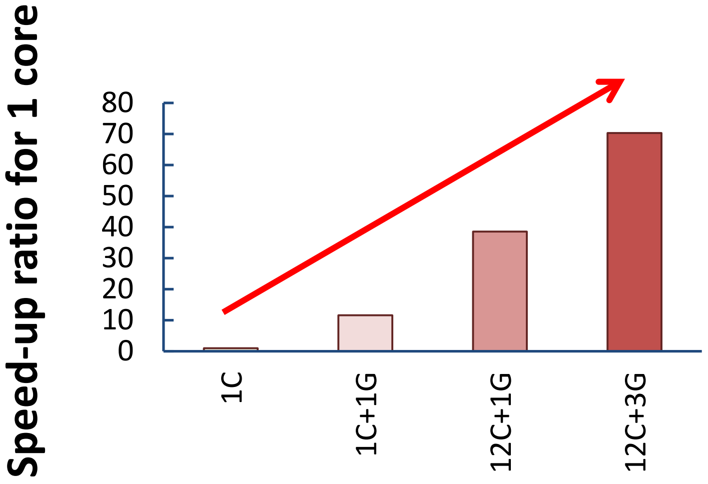
GHOSTZ-GPU

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

Multithread on GPU



TSUBAME 2.5 Thin node GPU



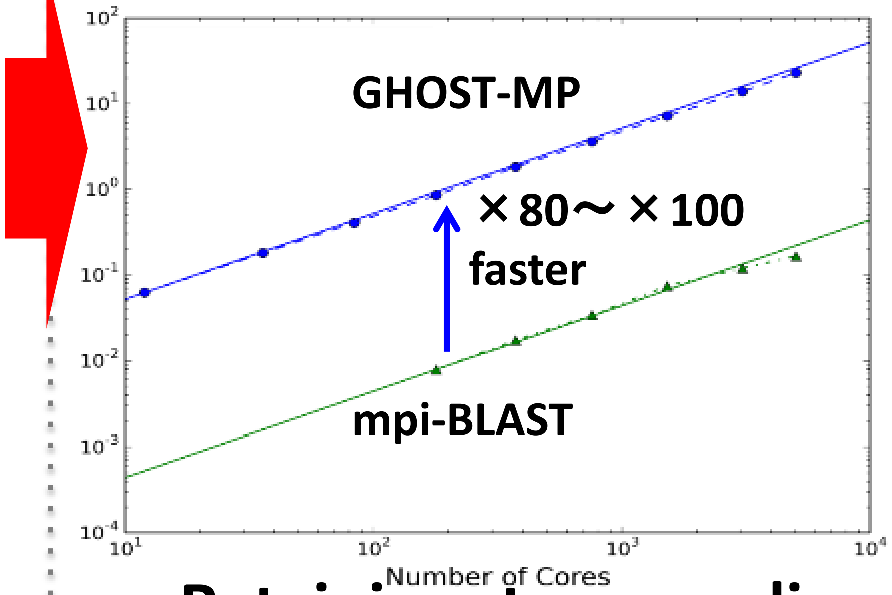
×70 faster than 1 core using 12 cores + 3 GPUs

GHOST-MP

Kakuta, *et al*. (submitted)

MPI + OpenMP hybrid pallelization

TSUBAME 2.5



Retaining strong scaling **up to 100,000 cores**

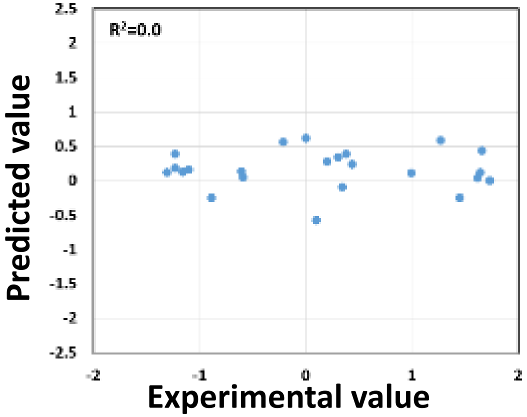
Plasma Protein Binding (PPB) Prediction by Machine Learning

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	○	△	◎

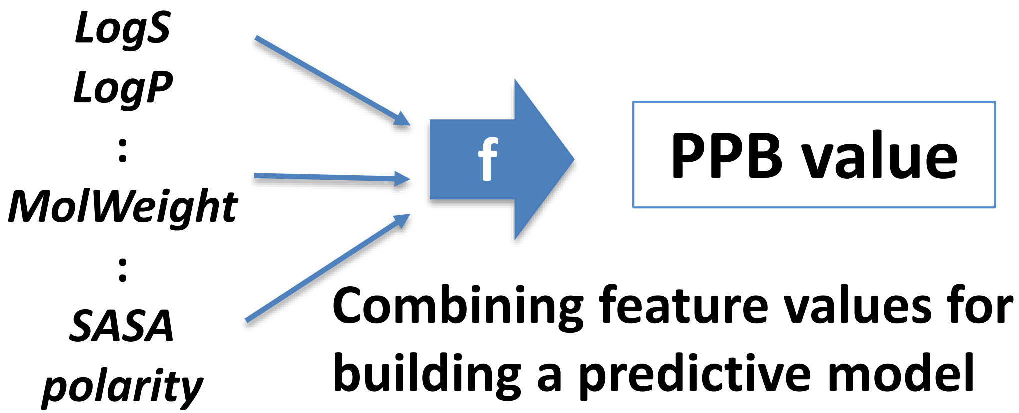
- Candidate peptides are tend to be degraded and excreted faster than small molecule drugs
- Strong needs to design bio-stable peptides for drug candidates



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

Solutions

Compute Feature Values
（more than 500 features）


LogS

LogP

:

MolWeight

:

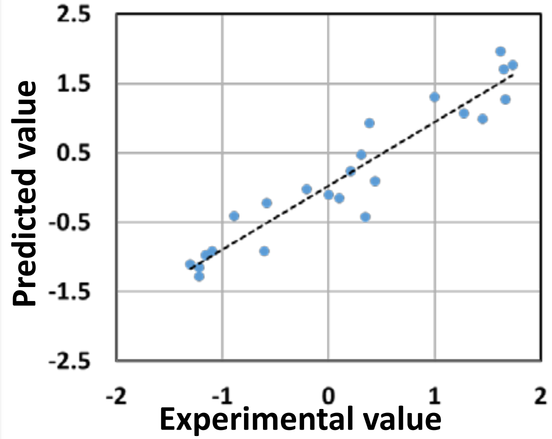
SASA

polarity

f

PPB value

Combining feature values for building a predictive model



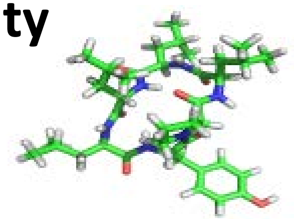
$R^2 = 0.905$

A constructed model can explain peptide PPB well

Molecular Dynamics Simulation for Membrane Permeability
Application for peptide drug discovery

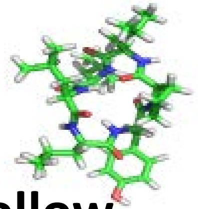
Problems

1) Single residue mutation can drastically change membrane permeability



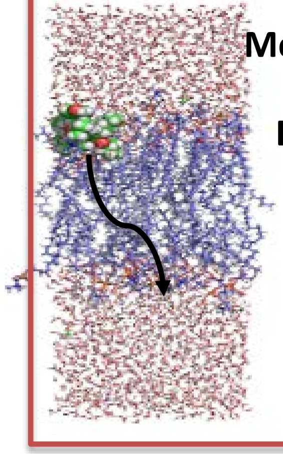
Sequence : D-Pro, D-Leu, D-Leu, L-Leu, D-Leu,
Membrane permeability : 7.9 × 10^{-6}cm/s

×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.

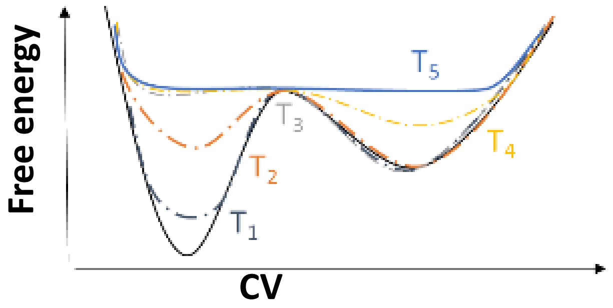
Ex) Membrane thickness : 40 Å
Peptide membrane permeability : 7.9×10^{-6} cm/s

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)

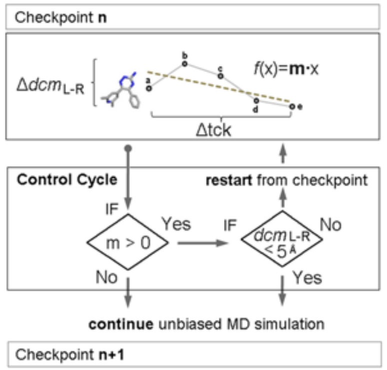
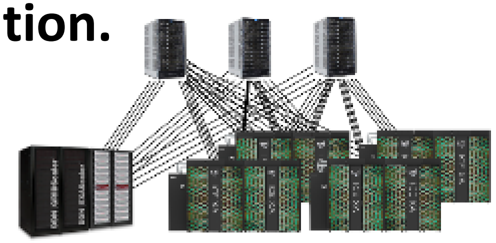


Figure 1. Scheme of the ligand–receptor distance vector ($dcm_{L\text{-}R}$) supervision algorithm implemented in the supervised molecular dynamics (SuMD) technique.

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

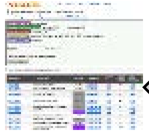
(Tonomachi, Kawasaki-city: planned 2017, PI Yutaka Akiyama)

Tokyo Tech

AIRC

Tokyo Tech's research seeds

①Drug Target selection system

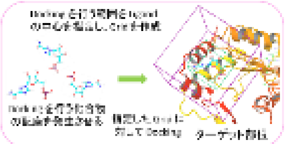


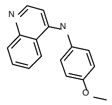
Minister of Health, Labour and Welfare Award of the 11th annual Merit Awards for Industry-Academia-Government Collaboration

Drugs for NTDs

②Glide-based Virtual Screening

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

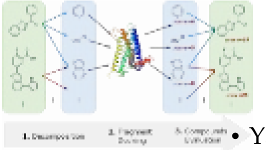




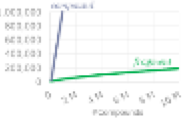
- Yoshino *et al., PLOS ONE* (2015)
- Chiba *et al., Sci Rep* (2015)

③Novel Algorithms for fast virtual screening against huge databases

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



Spresso

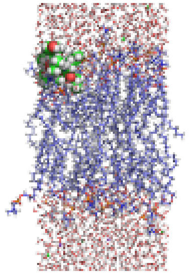


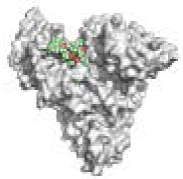
- Yanagisawa *et al., GIW* (2016)

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)

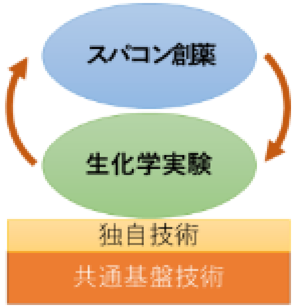
PeptiDream

Catalyst Inc.
(株)カタリスト

DiscoveResource
ディスカヴァリソース(株)

SCHRÖDINGER.
シュレーディンガー(株)

IMSBIO
(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)
Munincipie Govt (Kawasaki), JP & US Pharma

EBD App2: Miyoshi Group (Weather Forecast Application)

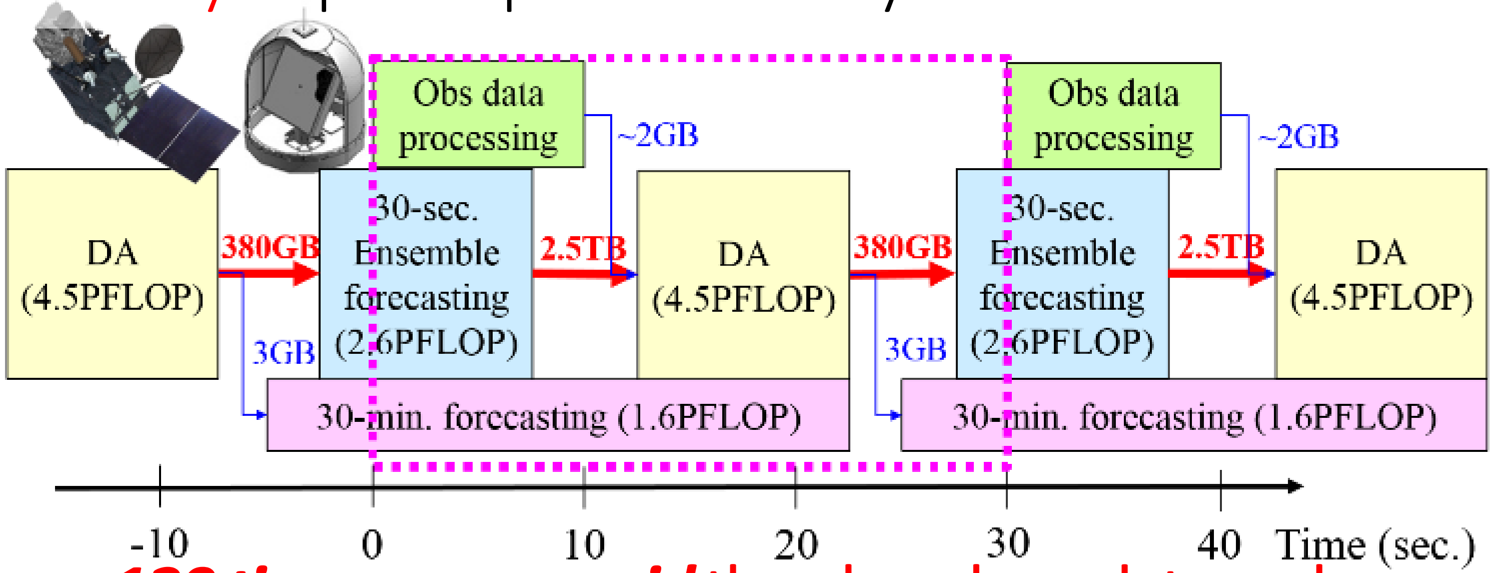
Only in 10 minutes!

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


Obs data processing
~2GB
DA (4.5PFLOP)
380GB
3GB
30-sec. Ensemble forecasting (2.6PFLOP)
2.5TB
DA (4.5PFLOP)
380GB
3GB
Obs data processing
~2GB
30-sec. Ensemble forecasting (2.6PFLOP)
2.5TB
DA (4.5PFLOP)
30-min. forecasting (1.6PFLOP)
30-min. forecasting (1.6PFLOP)
-10
0
10
20
30
40
Time (sec.)

120 times more rapid than hourly update cycles

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

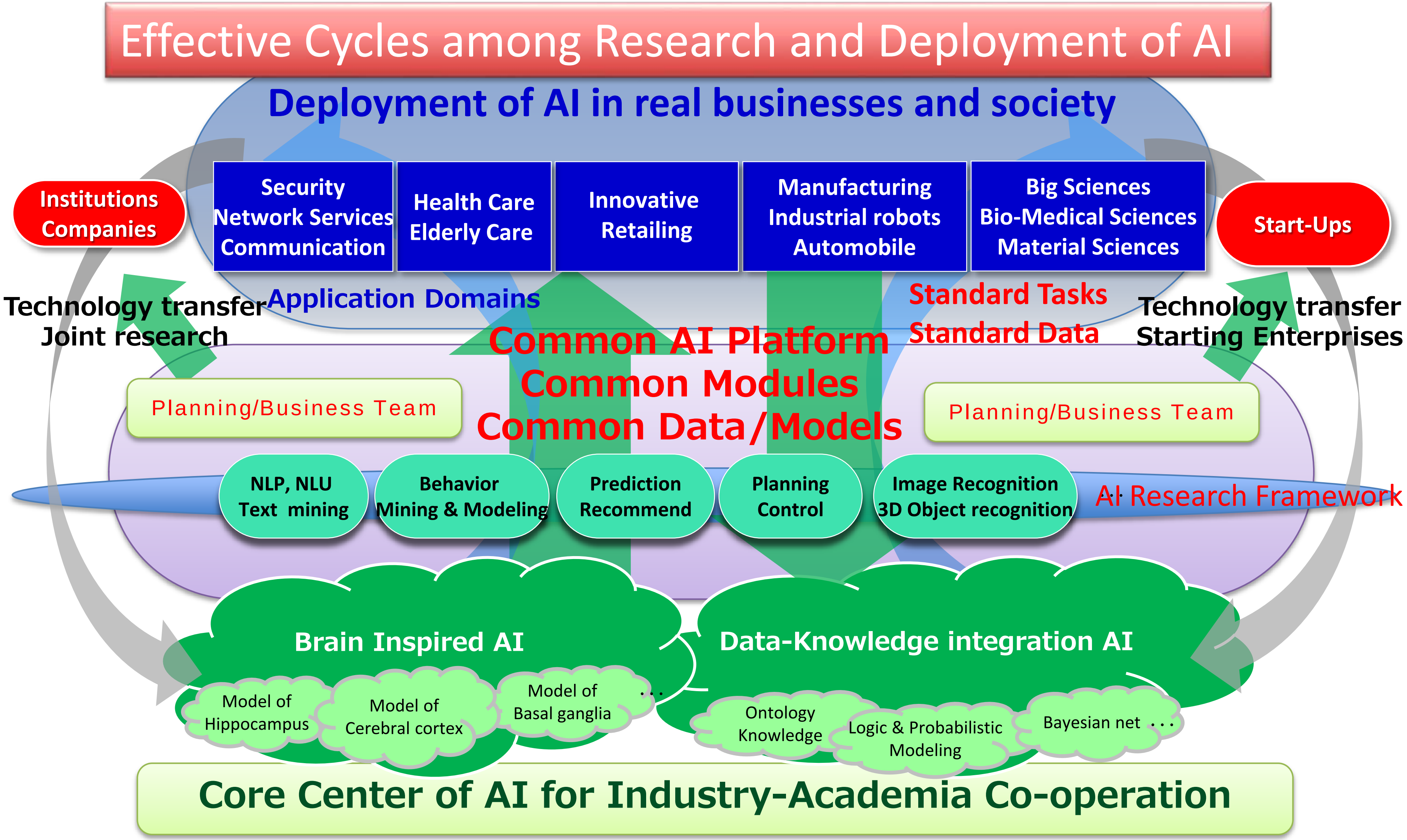
- Three national centers on Big Data and AI launched by three competing Ministries for FY 2016 (Apr 2015-)
  - METI – AIRC (Artificial Intelligence Research Center): AIST (AIST internal budget + > $200 million FY 2017), April 2015
    - Broad AI/BD/IoT, industry focus
  - MEXT – AIP (Artificial Intelligence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

Director:
Jun-ichi Tsujii

Now > 400+ FTEs



Effective Cycles among Research and Deployment of AI

Deployment of AI in real businesses and society

Institutions
Companies

Security
Network Services
Communication

Health Care
Elderly Care

Innovative
Retailing

Manufacturing
Industrial robots
Automobile

Big Sciences
Bio-Medical Sciences
Material Sciences

Start-Ups

Technology transfer
Joint research

Application Domains

Standard Tasks
Standard Data

Technology transfer
Starting Enterprises

Common AI Platform
Common Modules
Common Data/Models

Planning/Business Team

Planning/Business Team

NLP, NLU
Text mining

Behavior
Mining & Modeling

Prediction
Recommend

Planning
Control

Image Recognition
3D Object recognition

AI Research Framework

Brain Inspired AI

Model of
Hippocampus

Model of
Cerebral cortex

Model of
Basal ganglia

・・・

Data-Knowledge integration AI

Ontology
Knowledge

Logic & Probabilistic
Modeling

Bayesian net ・・・

Core Center of AI for Industry-Academia Co-operation

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

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

National Institute of Advanced Industrial Science and Technology
AIST
独立行政法人
産業技術総合研究所

National Institute for Advanced Industrial Science and Technology (AIST)

経済産業省
Ministry of Economy, Trade and Industry

Ministry of Economics Trade and Industry (METI)

AIST Artificial Intelligence Research Center (AIRC)

AIRC

Application Area
Natural Langauge Processing
Robotics
Security

ABCI
AI Bridging Cloud Infrastructure

Joint Research on AI / Big Data and applications

REAL WORLD BIGDATA
RWBC-OIL

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

Director: Satoshi Matsuoka

Industrial Collaboration in data, applications

Industry

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

Tokyo Institute of Technology / GSIC

TSUBAME
Tokyo Institute of Technology

GSIC
Global Scientific Information and Computing Center

Resources and Acceleration of AI / Big Data, systems research

Tsubame 3.0/2.5 Big Data /AI resources

ITCS Departments

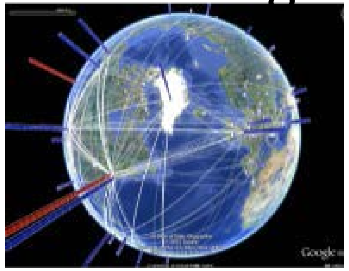
Basic Research in Big Data / AI algorithms and methodologies

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

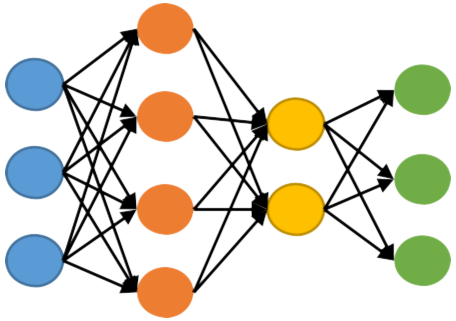
Integer Ops & Sparse Matrices
Data Movement, Large Memory
Sparse and Random Data, Low Locality

Acceleration, Scaling

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

As BD / AI

Dense LA: DNN
Inference, Training, Generation



As HPC Task

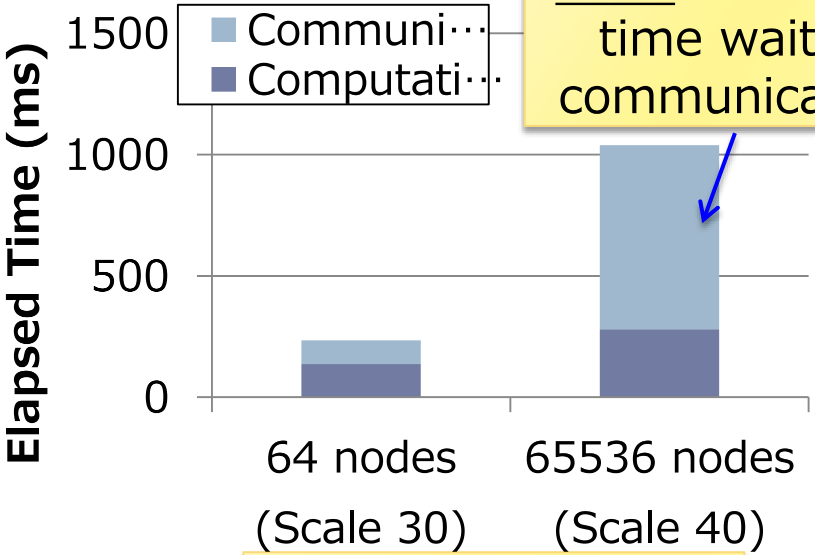
Dense Matrices, Reduced Precision
Dense and well organized neworks and Data

Acceleration, Scaling

Acceleration via Supercomputers adapted to AI/BD

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

#1 38621.4 GTEPS
(#7 10.51PF Top500)

Effective x13 performance c.f. Linpack

≫

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

LLNL-IBM Sequoia
1.6 million CPUs
1.6 Petabyte mem

#3 23751 GTEPS
(#4 17.17PF Top500)

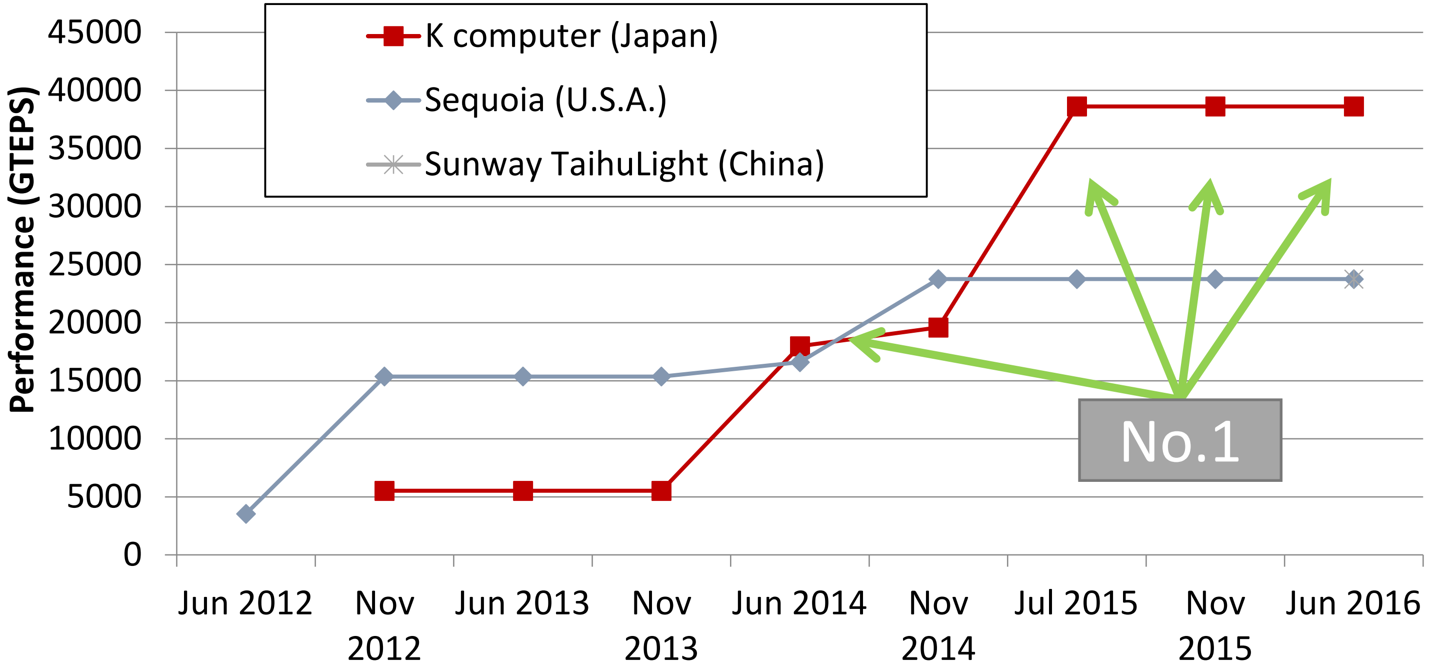
TaihuLight
10 million CPUs
1.3 Petabyte mem

#2 23755.7 GTEPS
(#1 93.01PF Top500)

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)

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

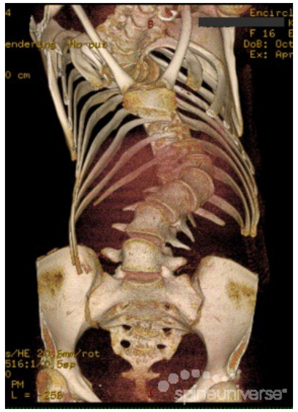
Single Rack High Density Oil Immersion
168 NVIDIA K80 GPUs + Xeon
413+TFlops (DFP)
1.5PFlops (SFP)
~60KW/rack

2013年11月/2014年6月
Word #1 Green500

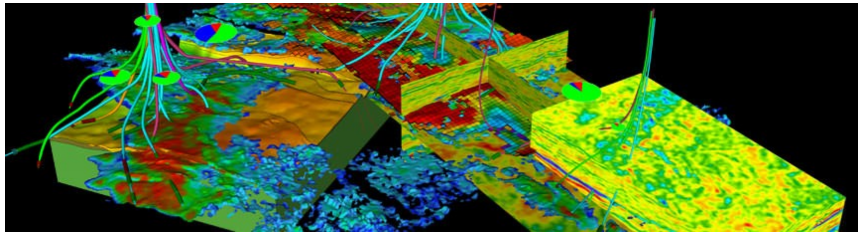
Container Facility
20 feet container (16m^2)
Fully Unmanned Operation

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

- **Need BOTH bandwidth and capacity (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…)

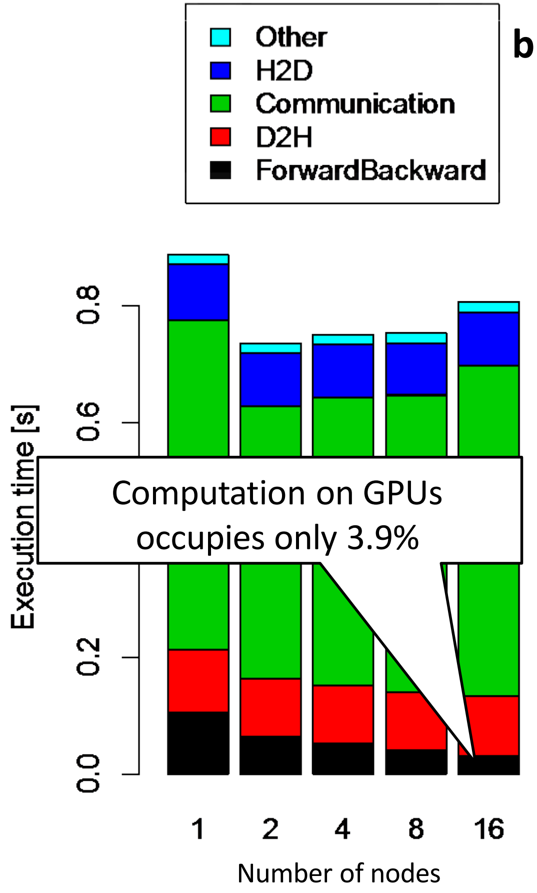


(Source: https://www.spineuniverse.com/image-library/anterior-3d-ct-scan-progressive-kyphoscoliosis)



(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)



Computation on GPUs occupies only 3.9%

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

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

TOP500 CERTIFICATE

“NEC/Sun/ClearSpeed/Voltaire TSUBAME, a Sun x4600 node cluster, at the GSIC Center, Tokyo Institute of Technology Japan

is ranked

No. 1 in Asia

[illegible]

Congratulations from the TOP500 Editors

2006 TSUBAME1.0
80 Teraflops, #1 Asia #7 World
“Everybody’s Supercomputer”

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

SC11

ACM Gordon Bell Prize
Special Achievements in Scalability and Time-to-Solution

Takashi Shimokawabe, Takayuki Aoki,
Tomohiro Takaki, Akinori Yamanaka,
Akira Nukada, Toshio Endo,
Naoya Maruyama, Satoshi Matsuoka

Peta-Scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer

2011 ACM Gordon Bell Prize

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

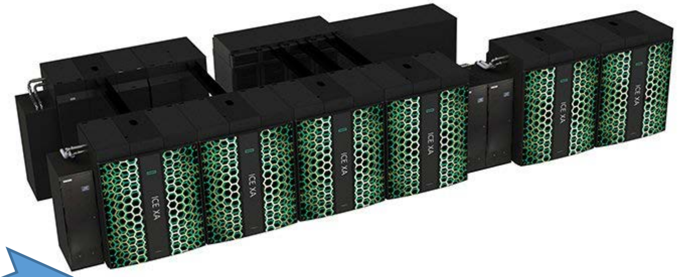
THE GREEN 500
sponsored by
SUPERMICRO

GSIC Center, Tokyo Institute of Technology

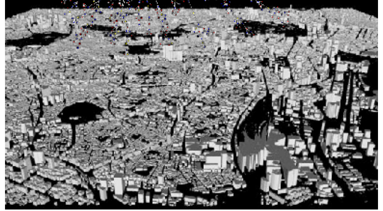
1st

November 2013

2013 TSUBAME-KFC
#1 Green 500

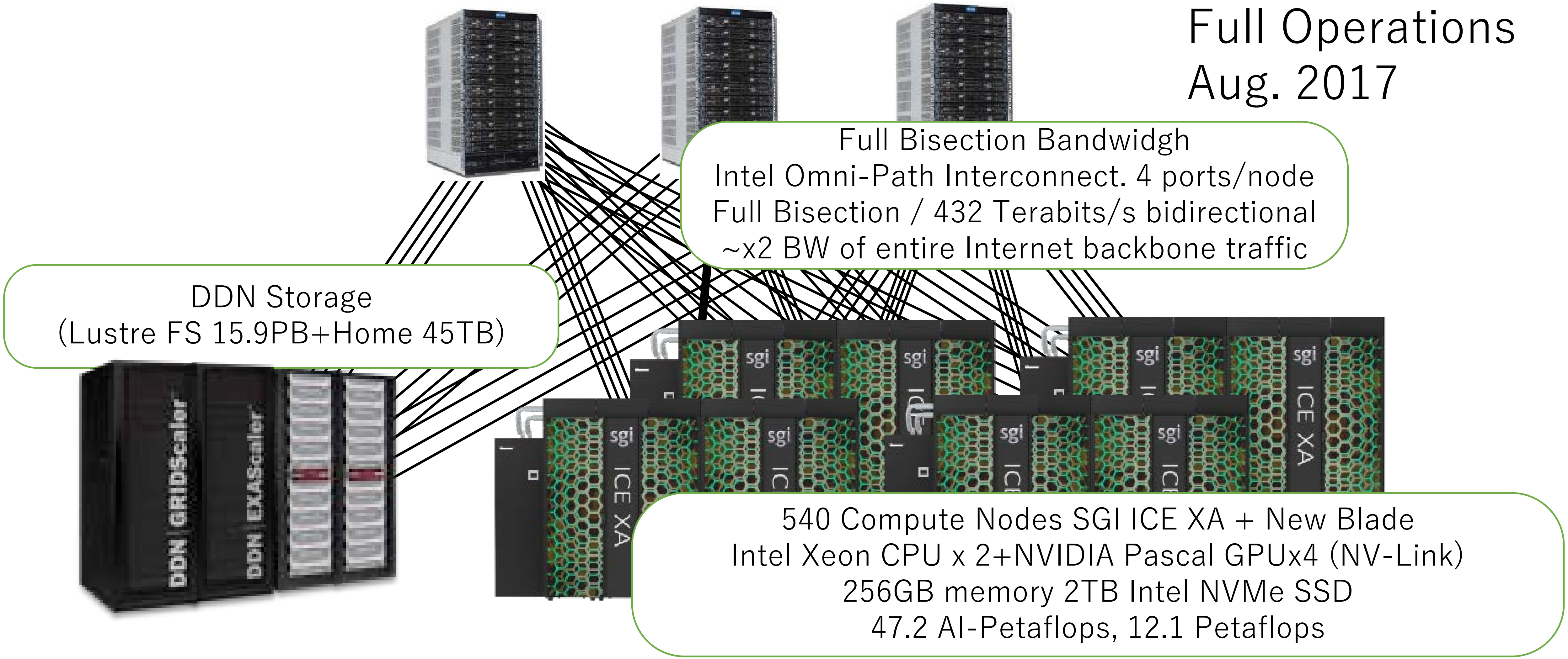


2017 TSUBAME3.0+2.5
~18PF(DFP) 4~5PB/s Mem BW
10GFlops/W power efficiency
Big Data & Cloud Convergence



Large Scale Simulation
Big Data Analytics
Industrial Apps

Overview of TSUBAME3.0 (#1 June 2017 Green 500)
BYTES-centric Architecture, Scalability to all 2160 GPUs, all nodes, the entire memory hierarchy



Full Operations
Aug. 2017

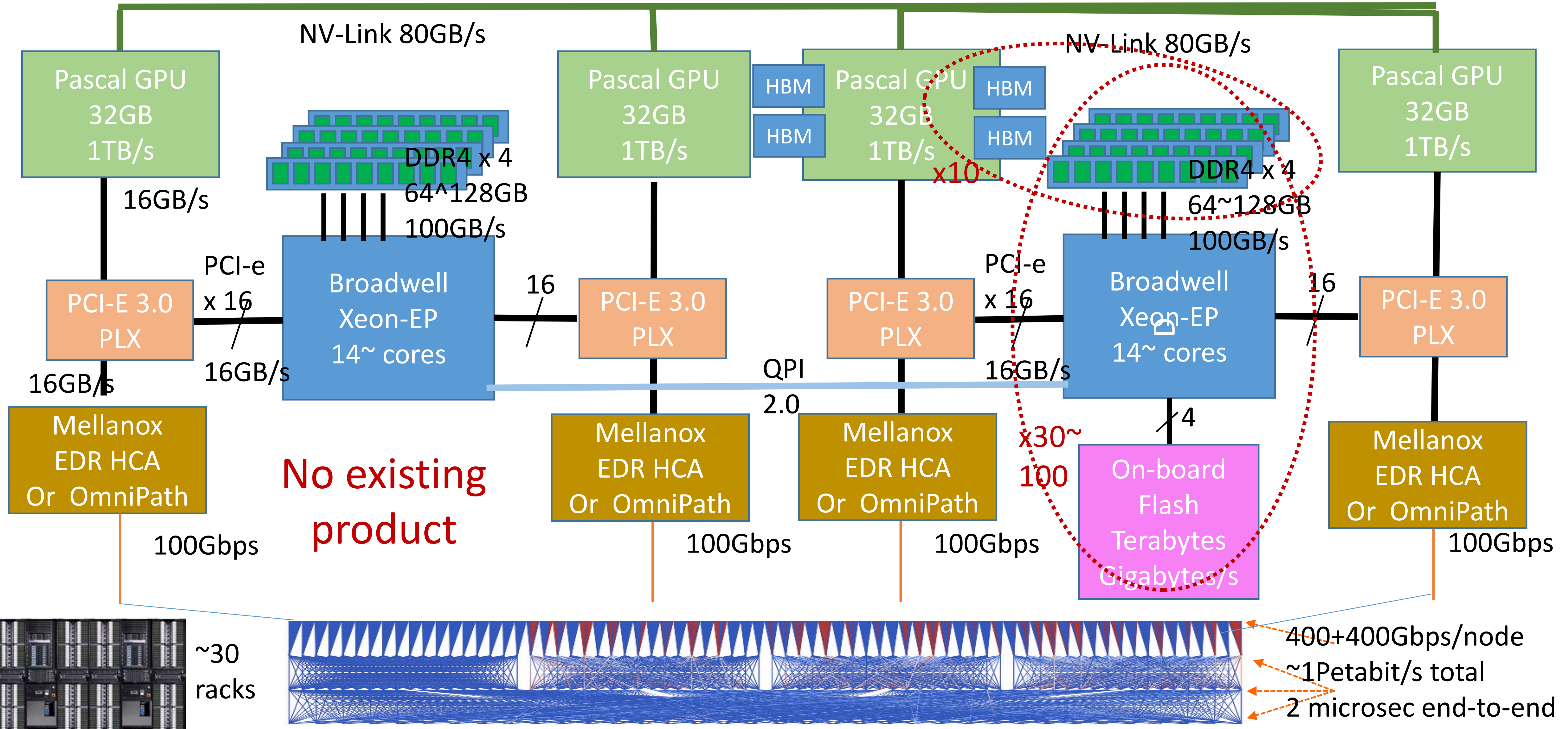
Full Bisection Bandwidgh
Intel Omni-Path Interconnect. 4 ports/node
Full Bisection / 432 Terabits/s bidirectional
~x2 BW of entire Internet backbone traffic

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

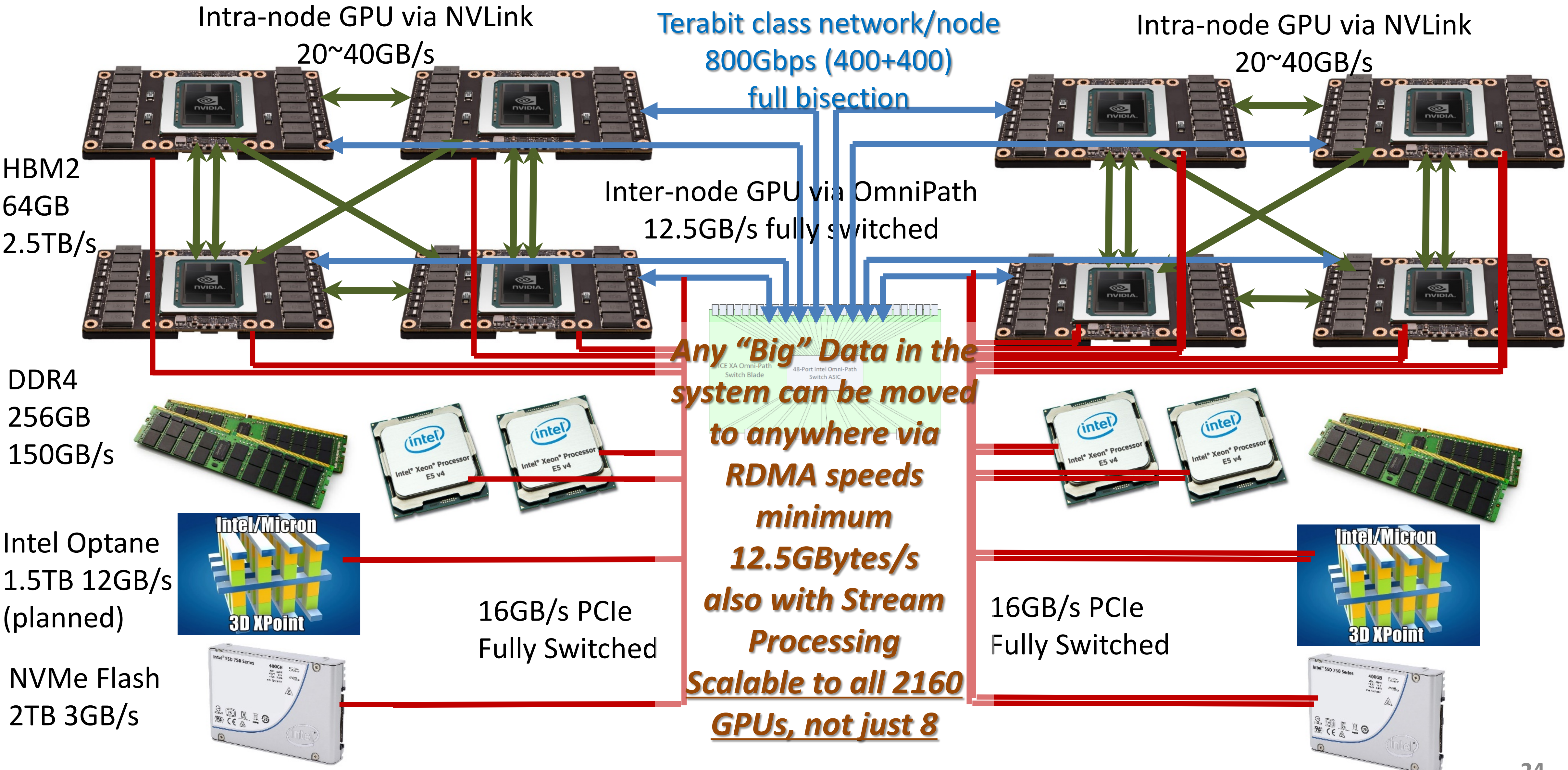
540 Compute Nodes SGI ICE XA + New Blade
Intel Xeon CPU x 2+NVIDIA Pascal GPUx4 (NV-Link)
256GB memory 2TB Intel NVMe SSD
47.2 AI-Petaflops, 12.1 Petaflops

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



Intra-node GPU via NVLink
20~40GB/s

Terabit class network/node
800Gbps (400+400)
full bisection

Intra-node GPU via NVLink
20~40GB/s

HBM2
64GB
2.5TB/s

Inter-node GPU via OmniPath
12.5GB/s fully switched

DDR4
256GB
150GB/s

Intel Optane
1.5TB 12GB/s
(planned)

NVMe Flash
2TB 3GB/s

16GB/s PCIe
Fully Switched

16GB/s PCIe
Fully Switched

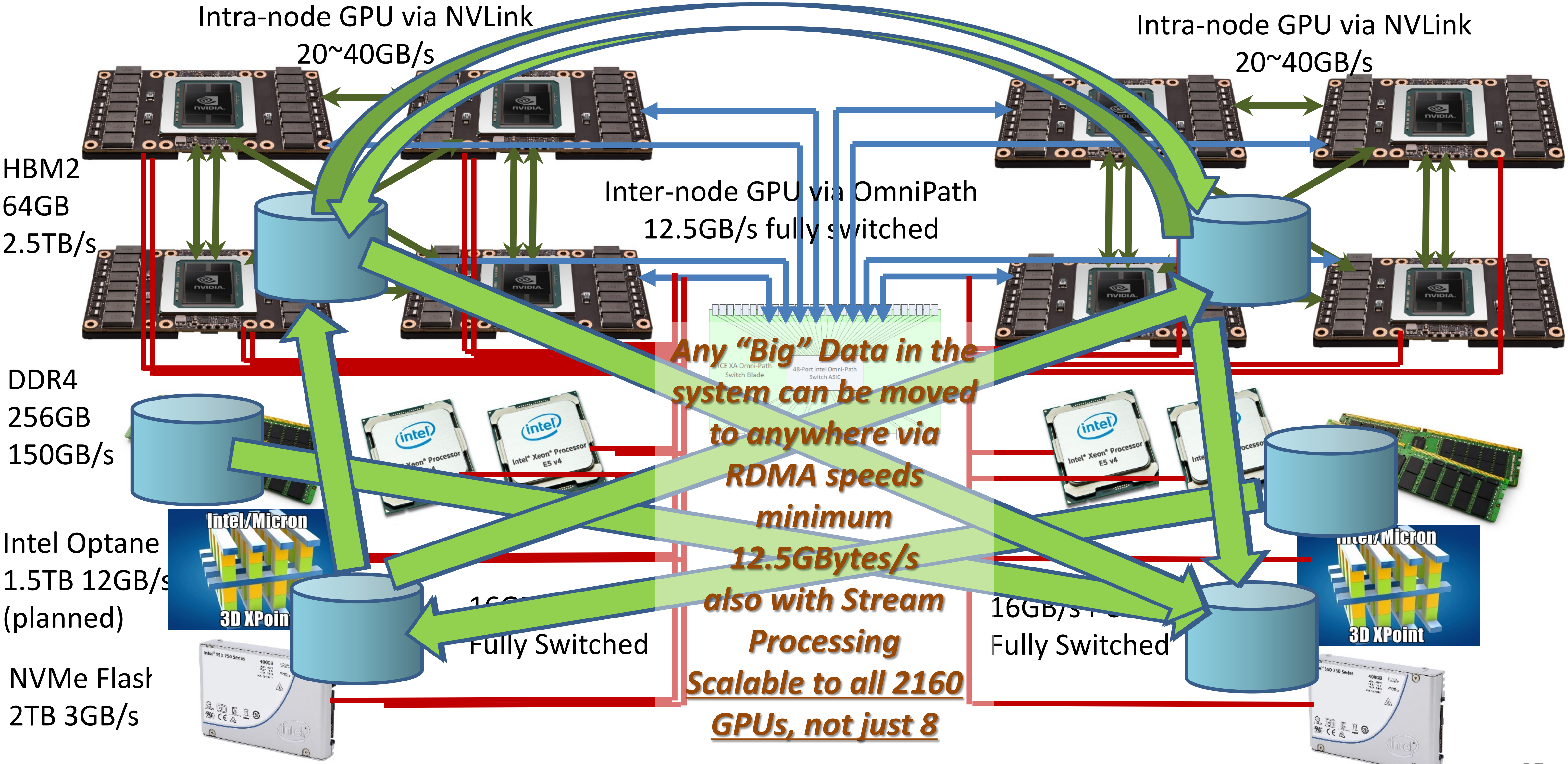
Any "Big" Data in the system can be moved to anywhere via RDMA speeds minimum 12.5GBytes/s also with Stream Processing
<u>Scalable to all 2160 GPUs, not just 8</u>

NVIDIA
intel
Intel® Xeon® Processor E5 v4
Intel/Micron
3D XPoint
ICE XA Omni-Path Switch Blade
48-Port Intel Omni-Path Switch ASIC
Intel® SSD 750 Series
400GB

<span>~4 Terabytes/node</span> 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



Intra-node GPU via NVLink
20~40GB/s

Intra-node GPU via NVLink
20~40GB/s

HBM2
64GB
2.5TB/s

Inter-node GPU via OmniPath
12.5GB/s fully switched

DDR4
256GB
150GB/s

Intel Optane
1.5TB 12GB/s
(planned)

NVMe Flash
2TB 3GB/s

Fully Switched

16GB/s
Fully Switched

Any "Big" Data in the system can be moved to anywhere via RDMA speeds minimum 12.5GBytes/s also with Stream Processing
<u>Scalable to all 2160 GPUs, not just 8</u>

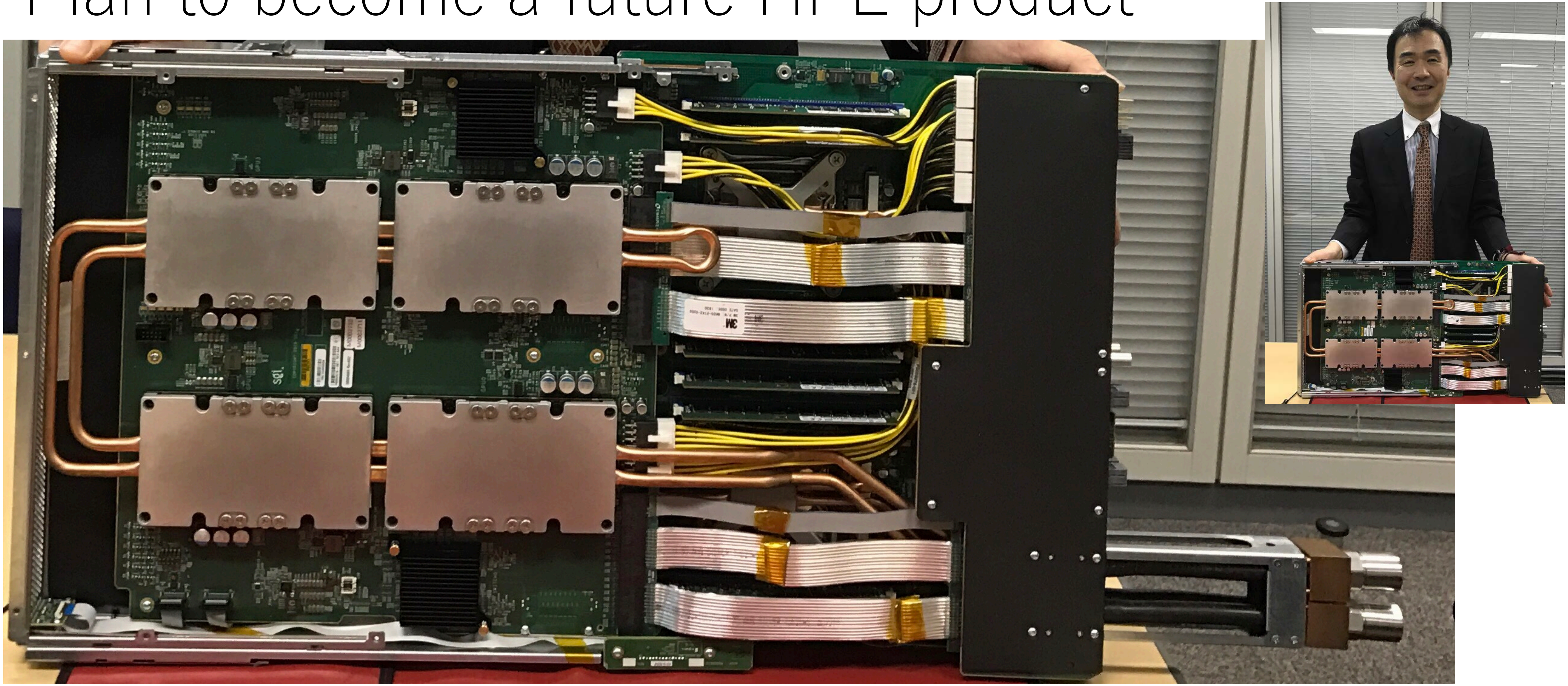
NVIDIA
intel
Intel® Xeon® Processor E5 v4
Intel/Micron
3D XPoint
Intel® SSD 750 Series
400GB
ICE XA Omni-Path Switch Blade
48-Port Intel Omni-Path Switch ASIC

~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*

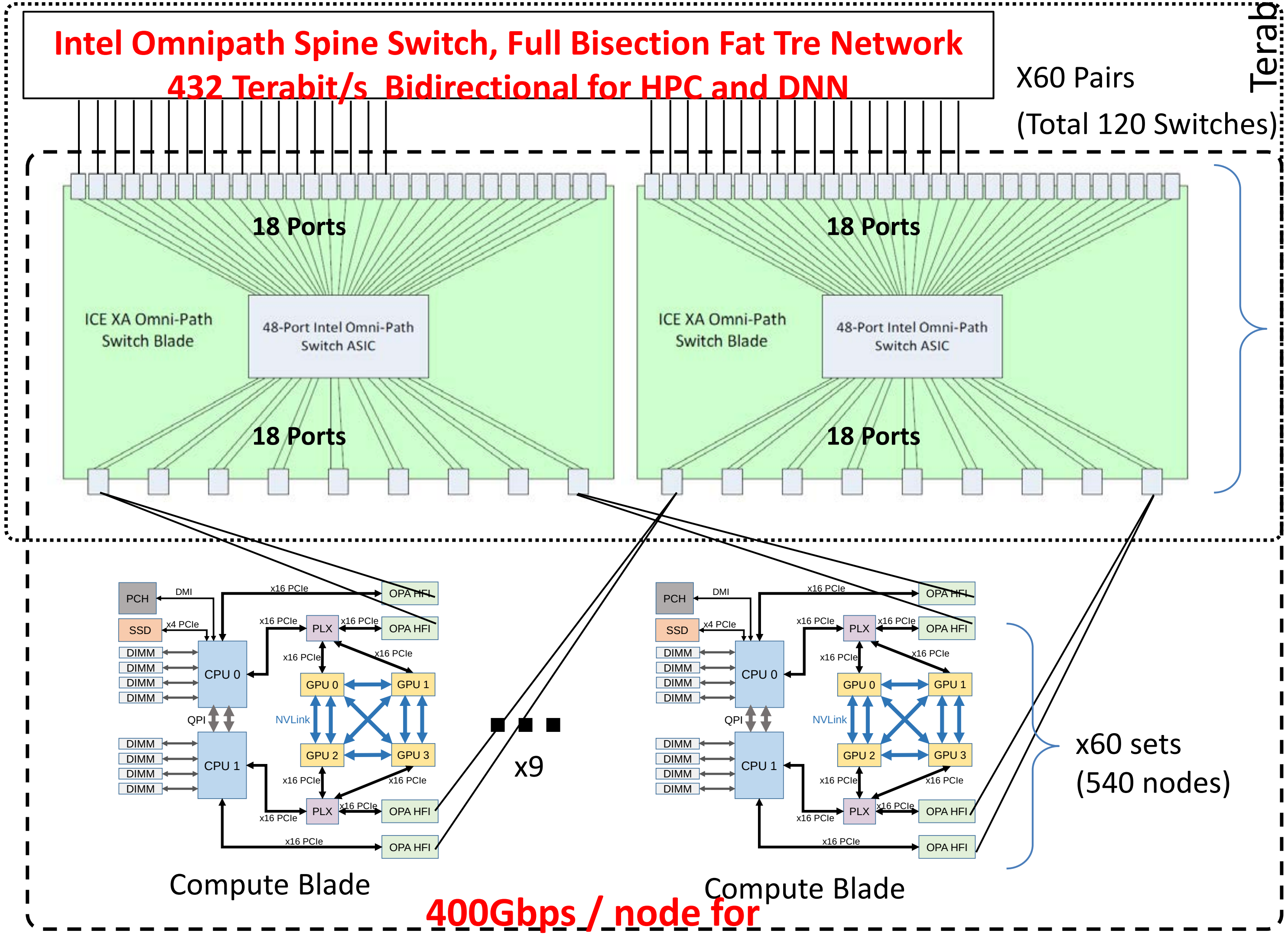
TSUBAME3.0 Co-Designed SGI ICE-XA Blade (new)

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

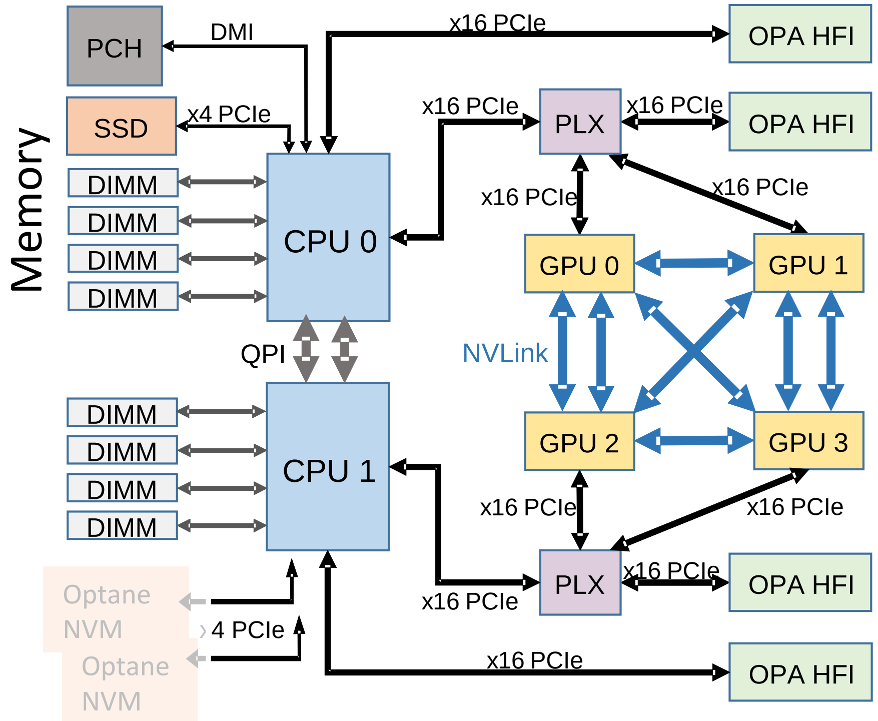


TSUBAME3.0 Compute Node SGI ICE-XA, a New GPU Compute Blade Co-Designed by SGI and Tokyo Tech GSIC

SGI ICE XA Infrastructure



400Gbps / node for HPC and DNN



Ultra high performance & bandwidth "Fat Node"

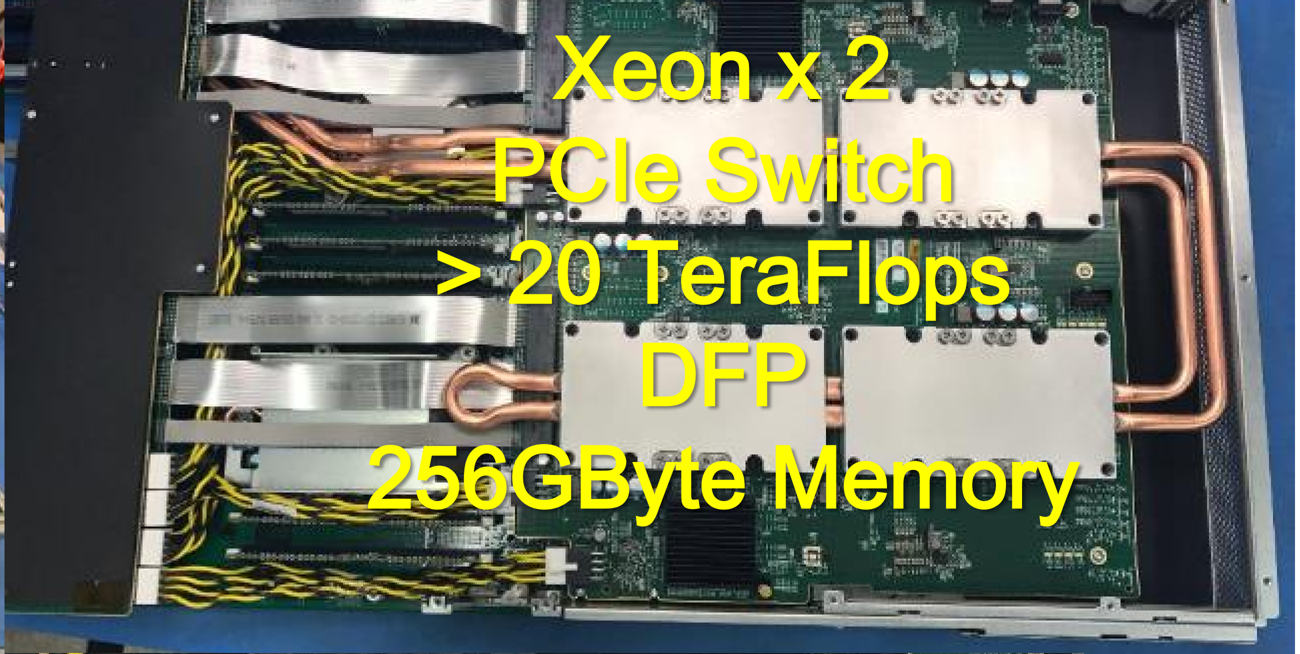
- High Performance: 4 SXM2(NVLink) NVIDIA Pascal P100 GPU + 2 Intel Xeon 84 AI-TFLops
- High Network Bandwidth – Intel Omnipath 100GBps x 4 = 400Gbps (100Gbps per GPU)
- High I/O Bandwidth - Intel 2 TeraByte NVMe
  - > 1PB & 1.5~2TB/s system total
  - Future Octane 3D-Xpoint memory Petabyte or more directly 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

Liquid Cooled "Hot Pluggable" ICE-XA Blade

Smaller than 1U server, no cables or pipes



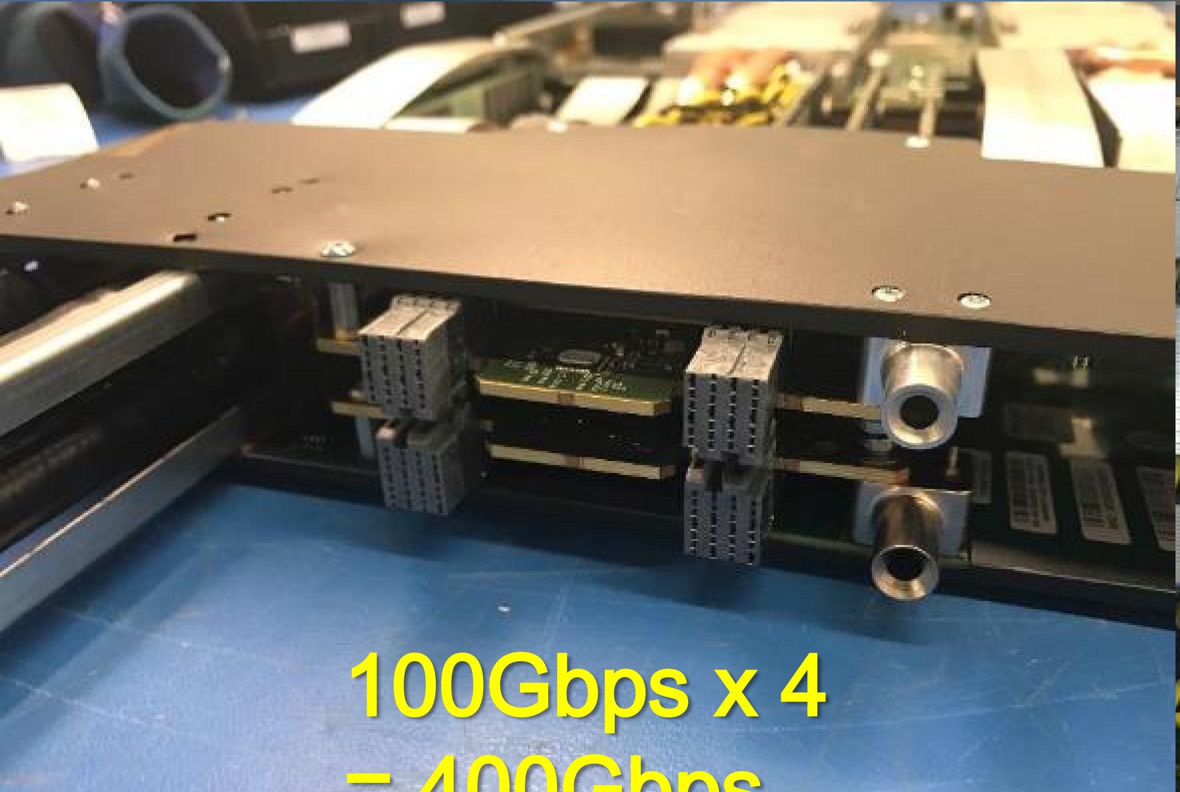
Xeon x 2

PCIe Switch

> 20 TeraFlops

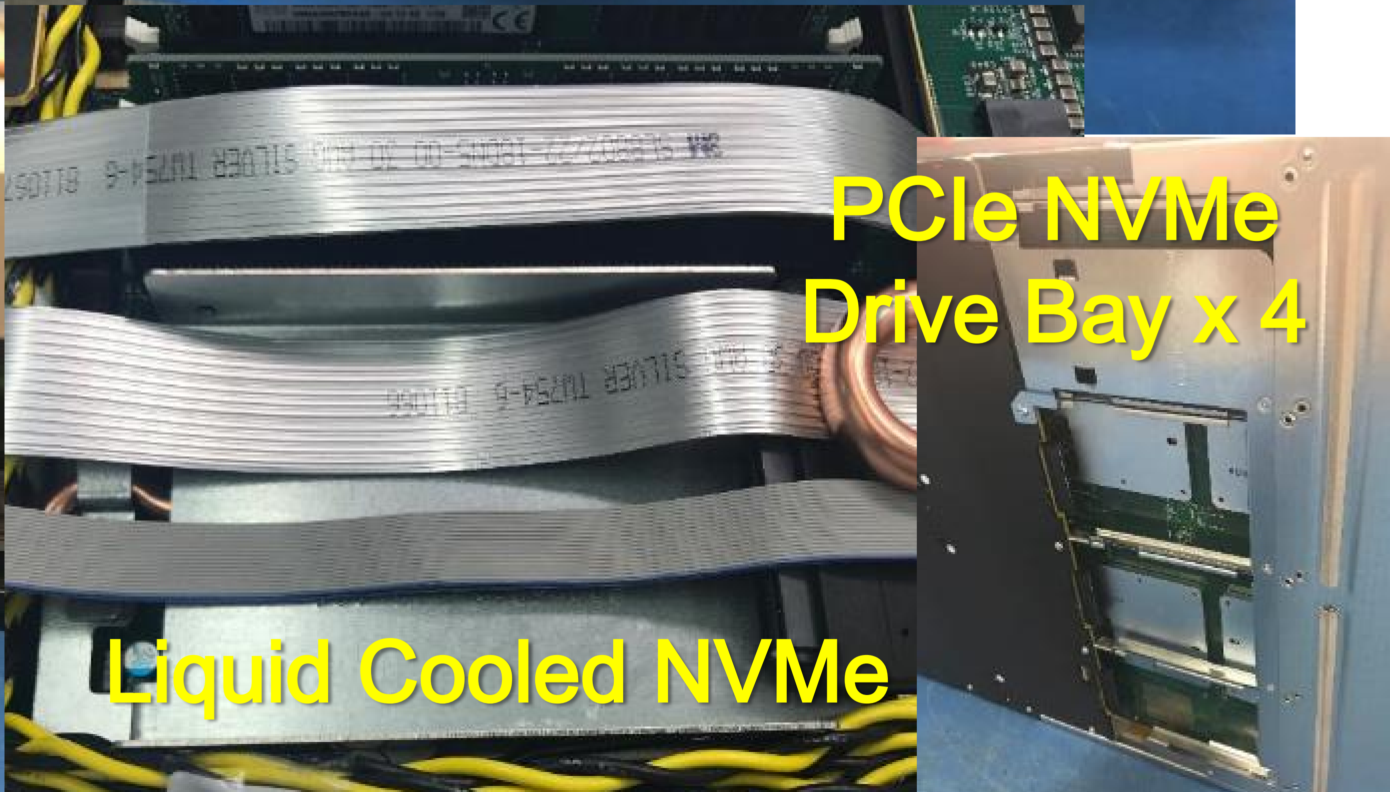
DFP

256GByte Memory



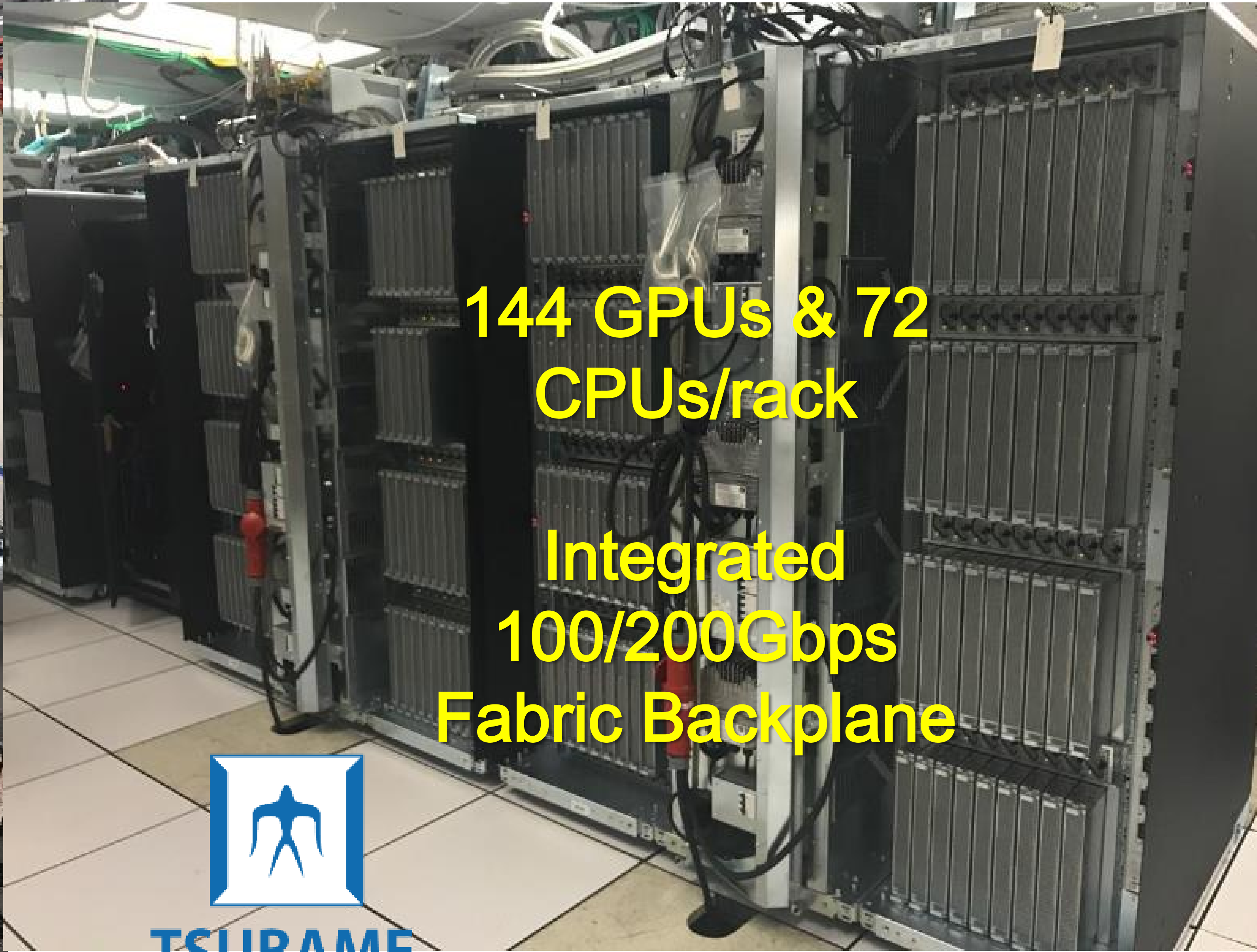
100Gbps x 4

= 400Gbps



PCIe NVMe Drive Bay x 4

Liquid Cooled NVMe

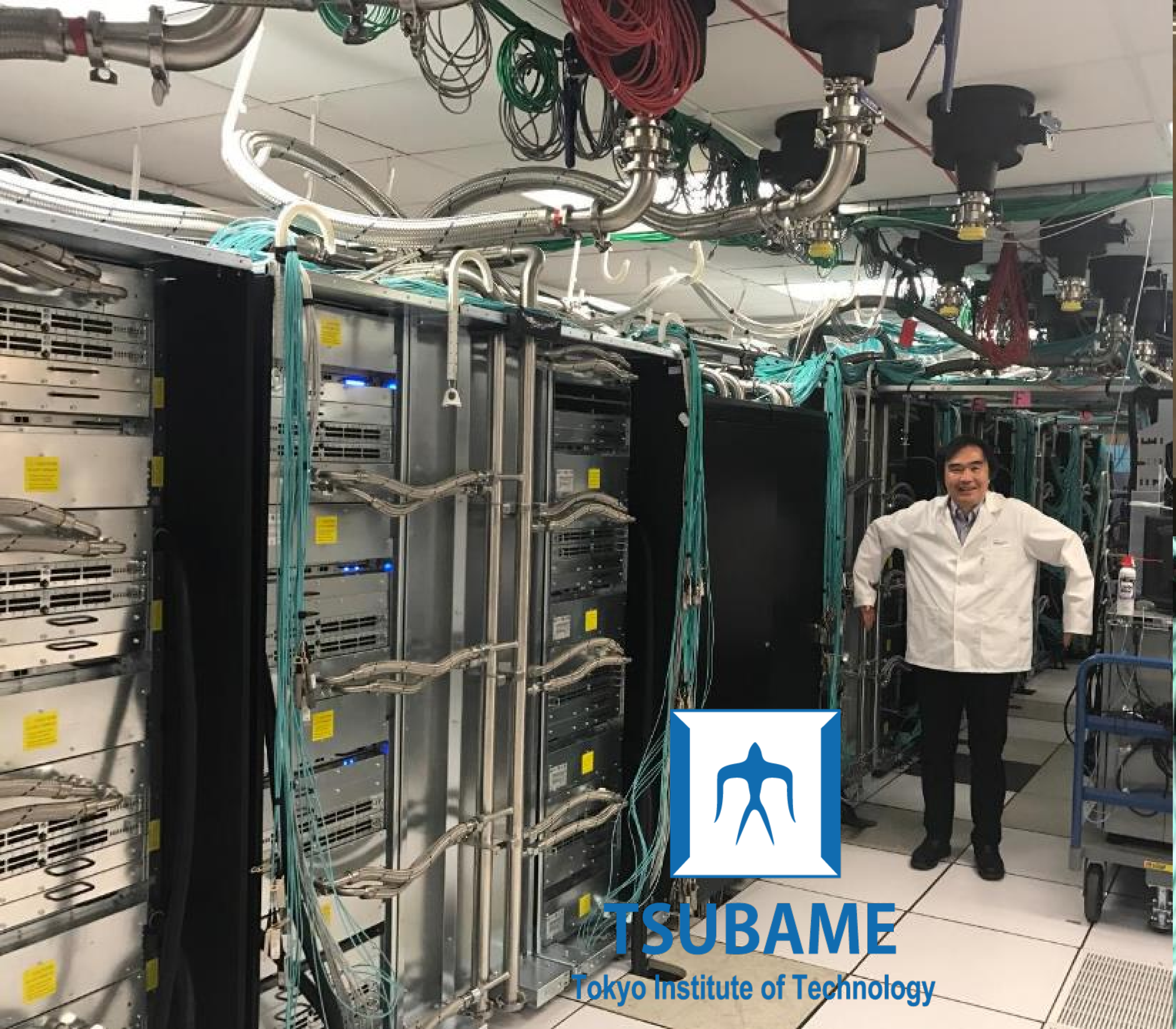


144 GPUs & 72 CPUs/rack

Integrated 100/200Gbps Fabric Backplane

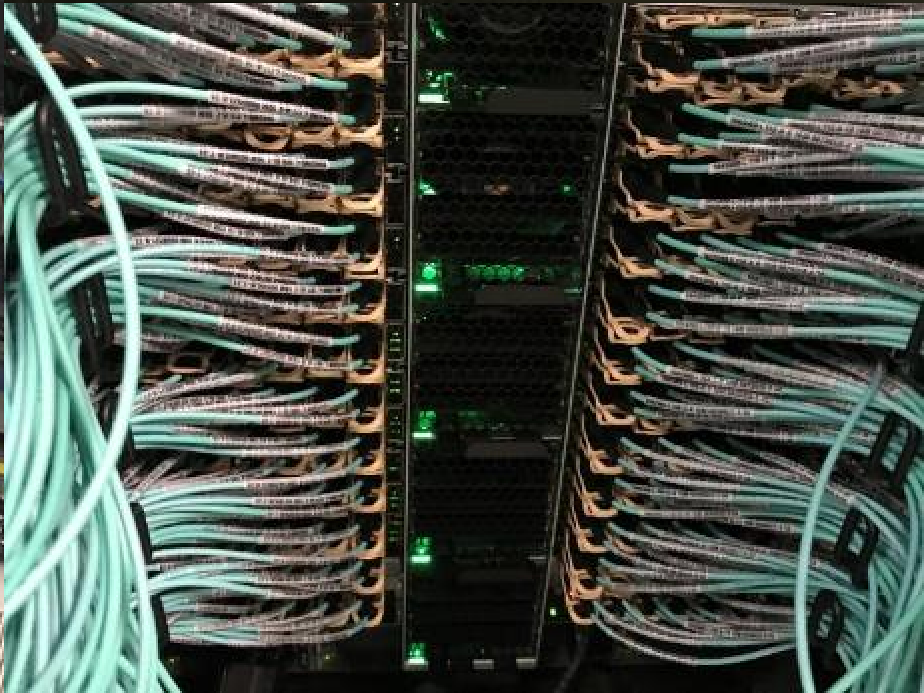
TSUBAME

Tokyo Institute of Technology

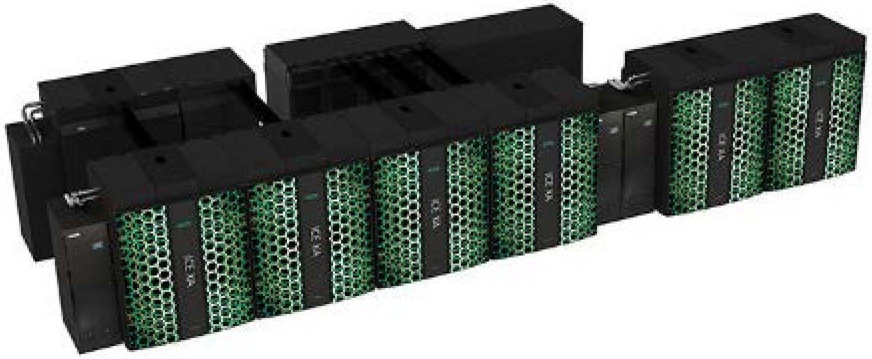


TSUBAME

Tokyo Institute of Technology



TSUBAME3.0 Datacenter

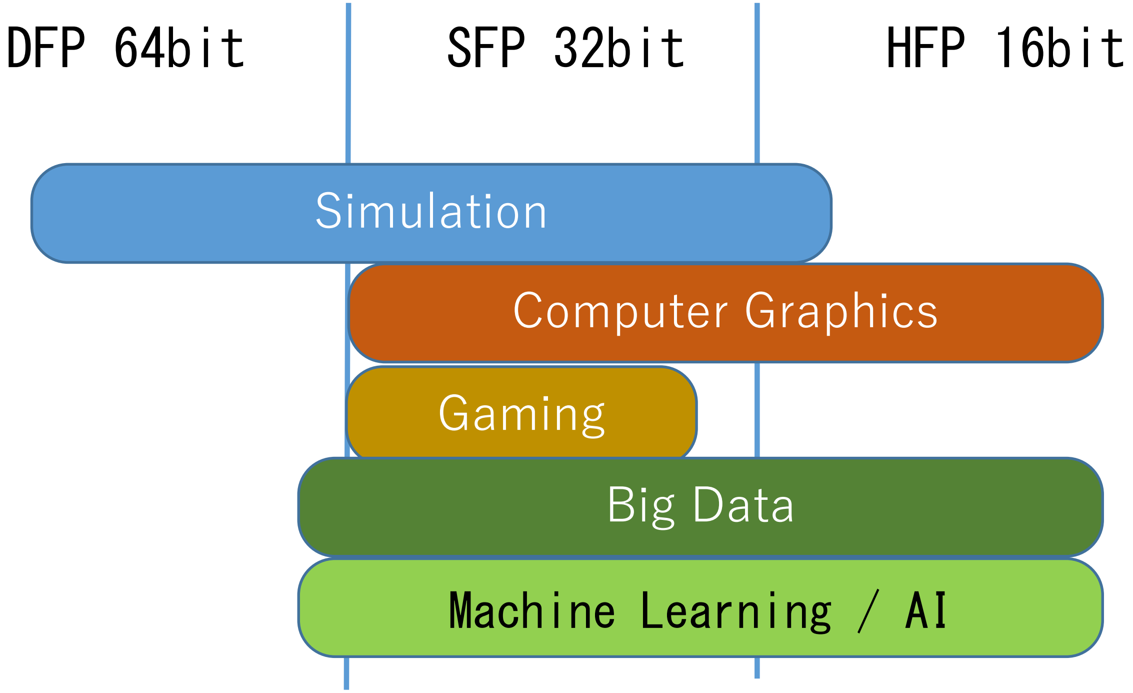


15 SGI ICE-XA Racks
2 Network Racks
3 DDN Storage Racks
20 Total Racks

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

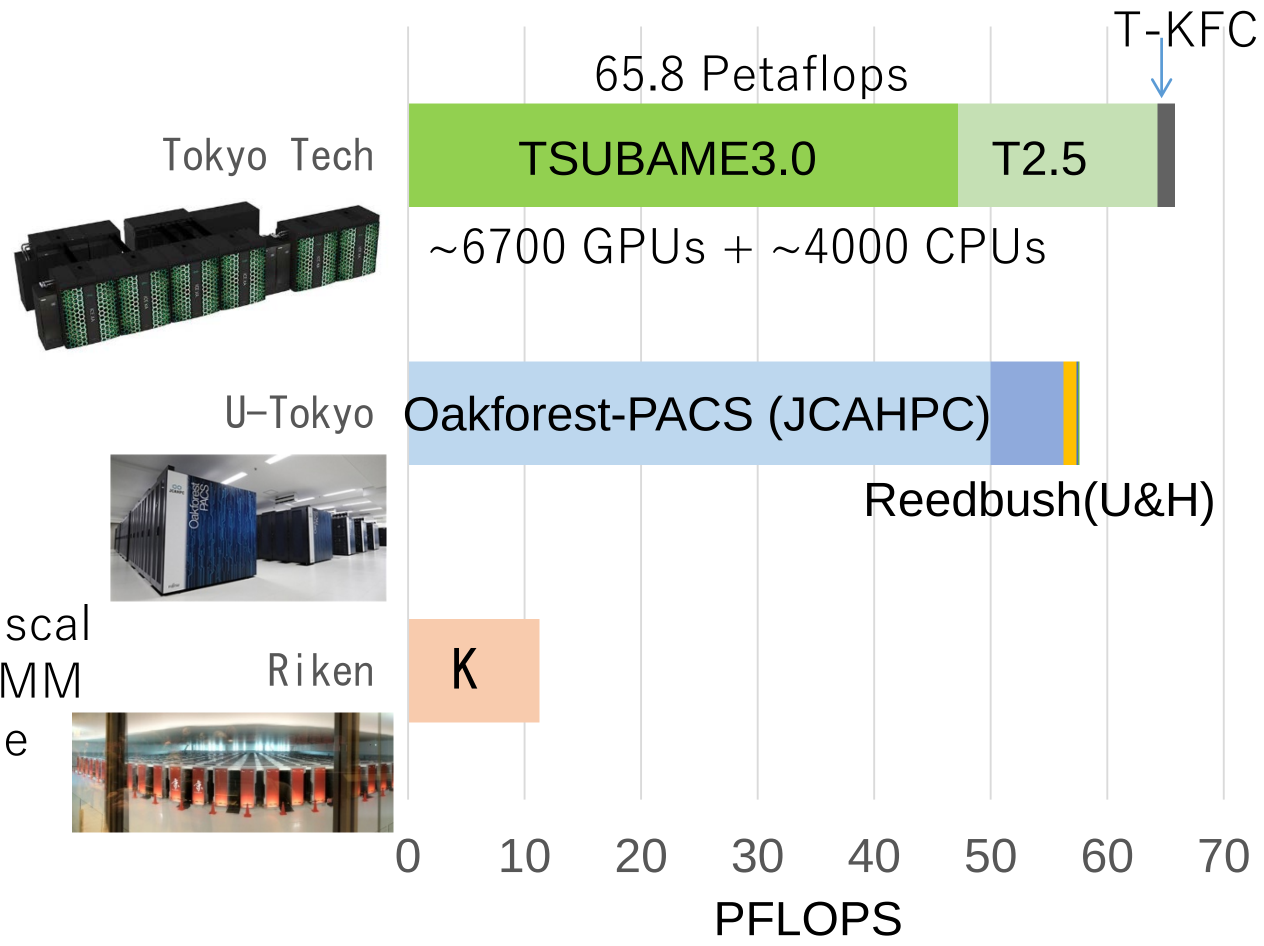
Japanese Open Supercomputing Sites Aug. 2017 (pink=HPCI Sites)

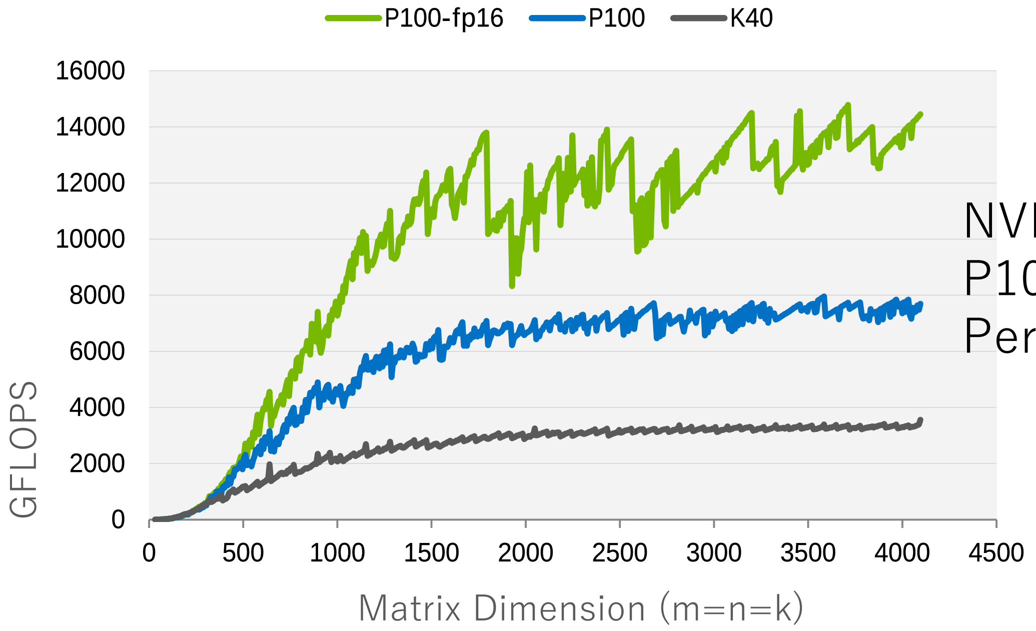
Peak Rank	Institution	System	Double FP Rpeak	Nov. 2016 Top500
1	U-Tokyo/Tsukuba U JCAHP	Oakforest-PACS - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path	24.9	6
2	Tokyo Institute of Technology GSIC	TSUBAME 3.0 - HPE/SGI ICE-XA custom NVIDIA Pascal P100 + Intel Xeon, Intel OmniPath	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

Site Comparisons of AI-FP Perfs



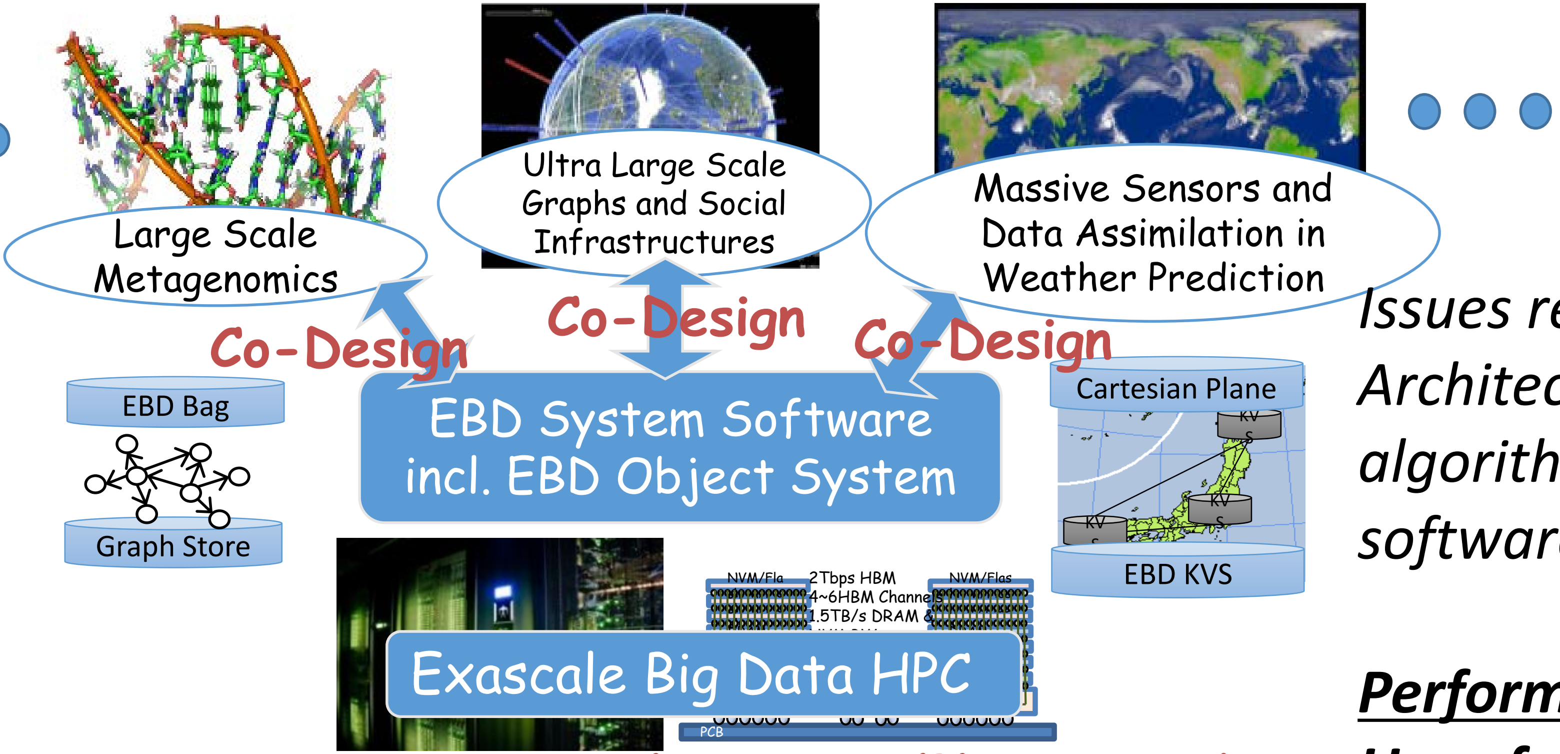


NVIDIA Pascal P100 DGEMM Performane

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

From FLOPS Centric to BYTES Centric HPC

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



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

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

<u>Performance Model?</u>
<u>Use of accelerators e.g. GPUs?</u>

Cloud IDC
Very low BW & Efficiency
Highly available, resilient

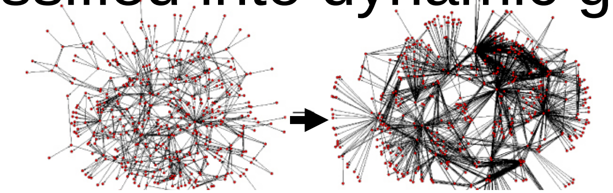
Supercomputers
Compute&Batch-Oriented
More fragile

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

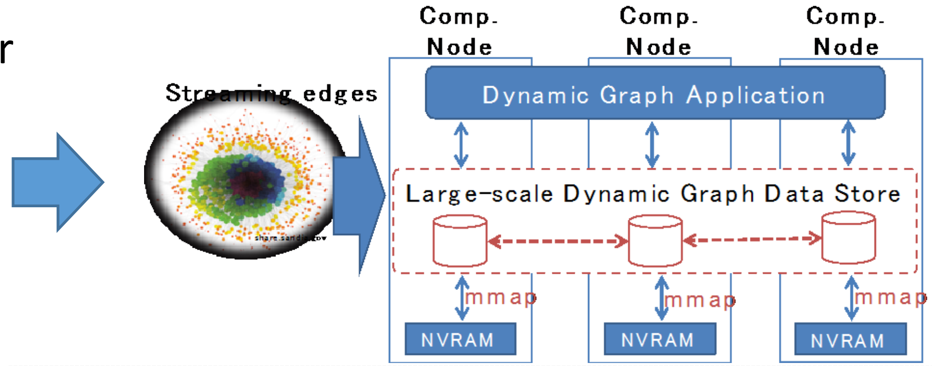
- social network, genome analysis, WWW, etc.
  - e.g., Facebook manages 1.39 billion active users as of 2014, with more than 400 billion edges

- 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

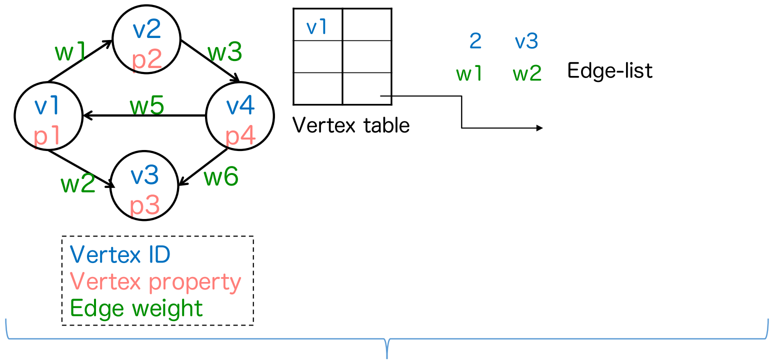
K Computer large memory but very expensive DRAM only



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)

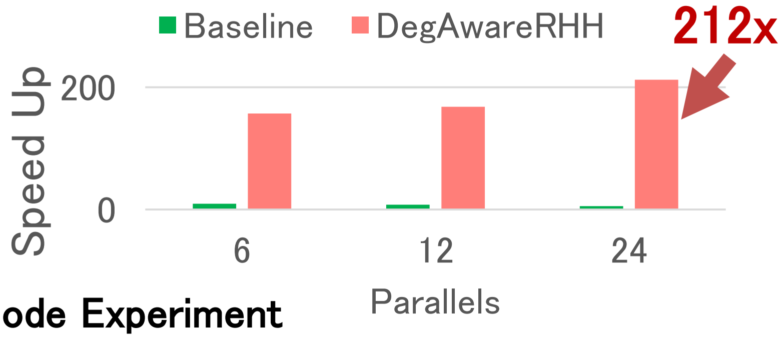
C.f. STINGER (single-node, on memory)

STINGER

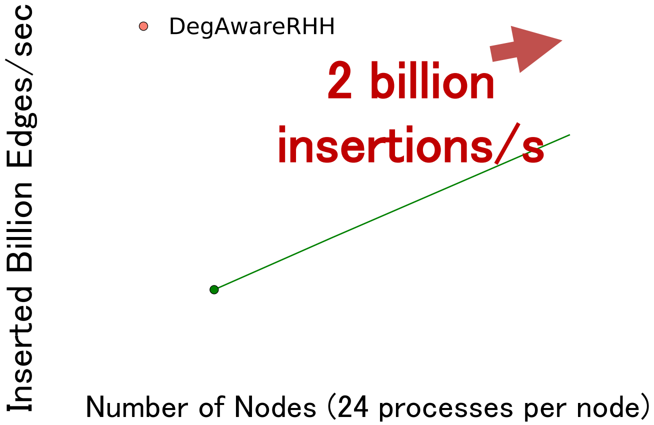
- A state-of-the-art dynamic graph processing framework developed at Georgia Tech

Baseline model

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



Multi-node Experiment



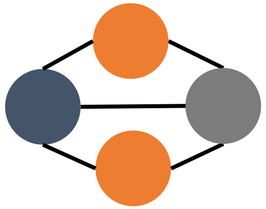
Dynamic graph store w/ world's top graph update performance and scalability

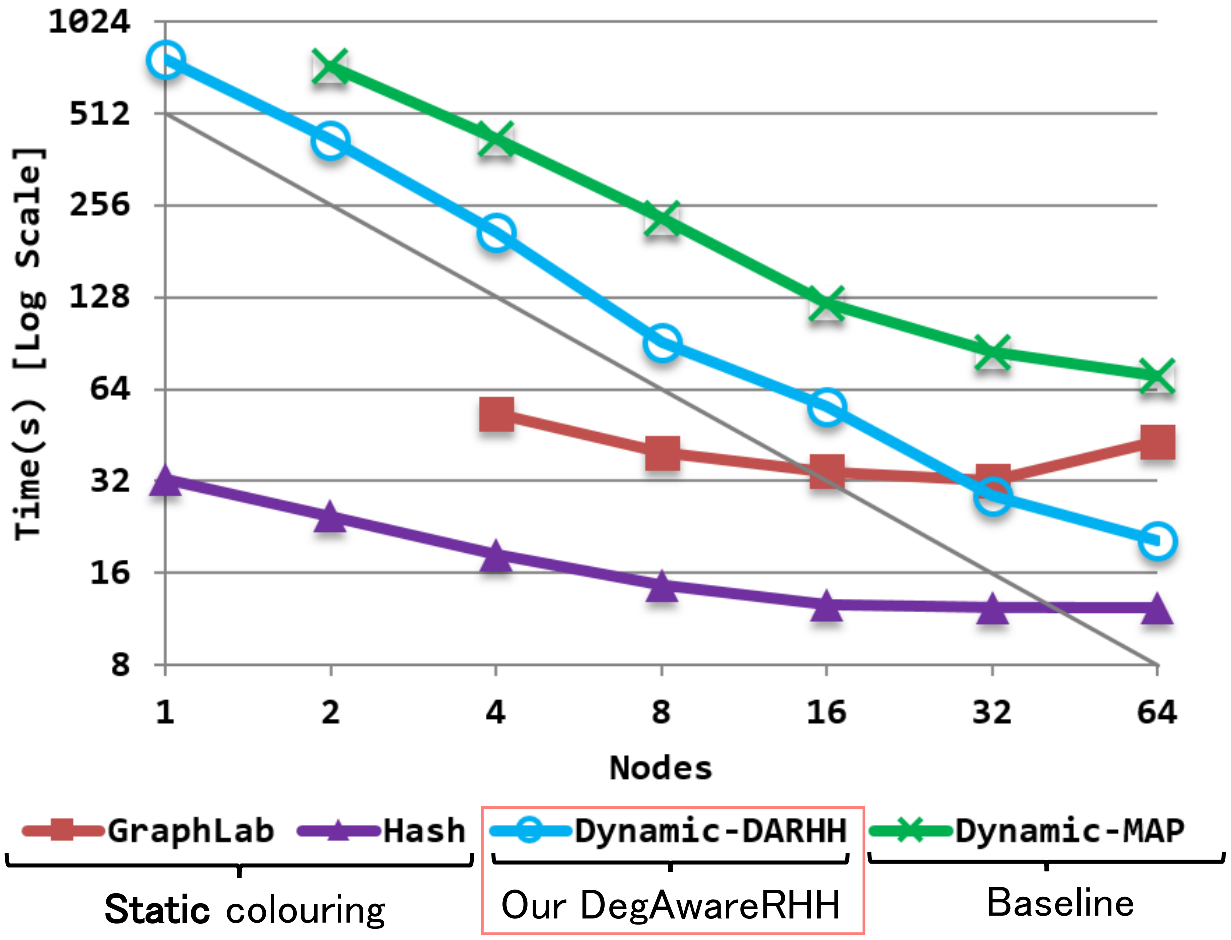
K. Iwabuchi, S. Sallinen, R. Pearce, B. V. Essen, M. Gokhale, and S. Matsuoka, Towards a distributed large-scale dynamic graph data store. In 2016 IEEE Interna- tional Parallel and Distributed Processing Symposium Workshops (IPDPSW)

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)

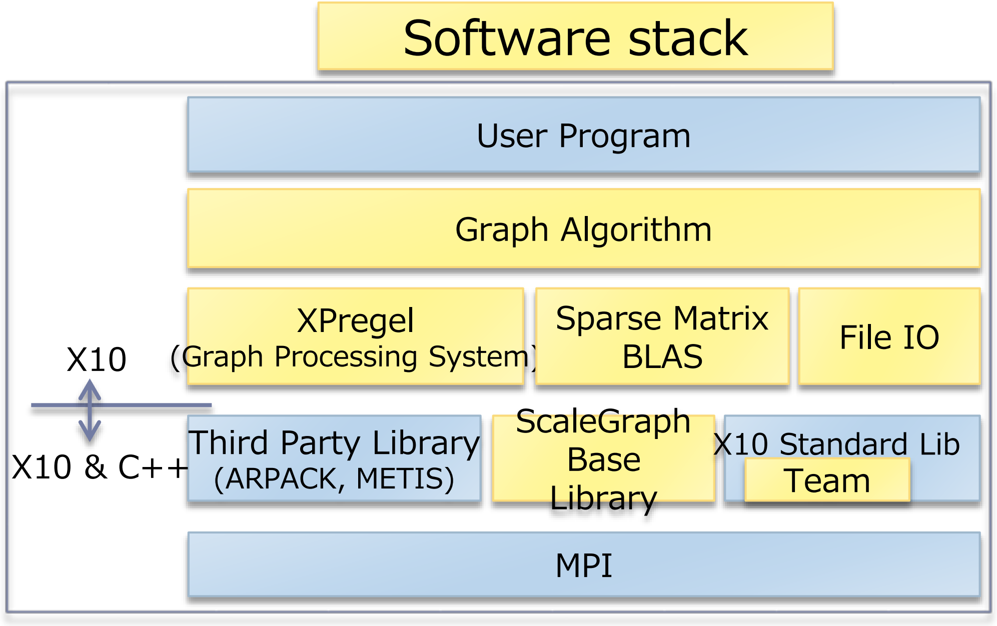


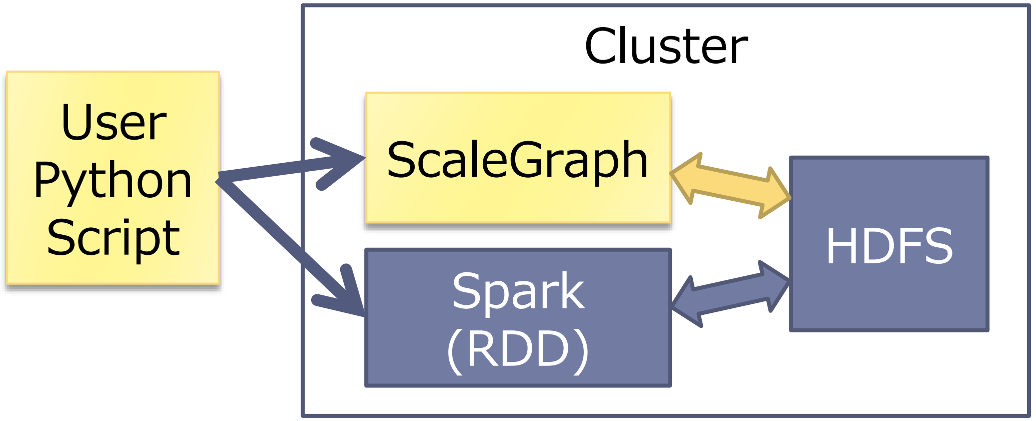


Scott Sallinen, Keita Iwabuchi, Roger Pearce, Maya Gokhale, Matei Ripeanu, "Graph Coloring as a Challenge Problem for Dynamic Graph Processing on Distributed Systems", SC' 16

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

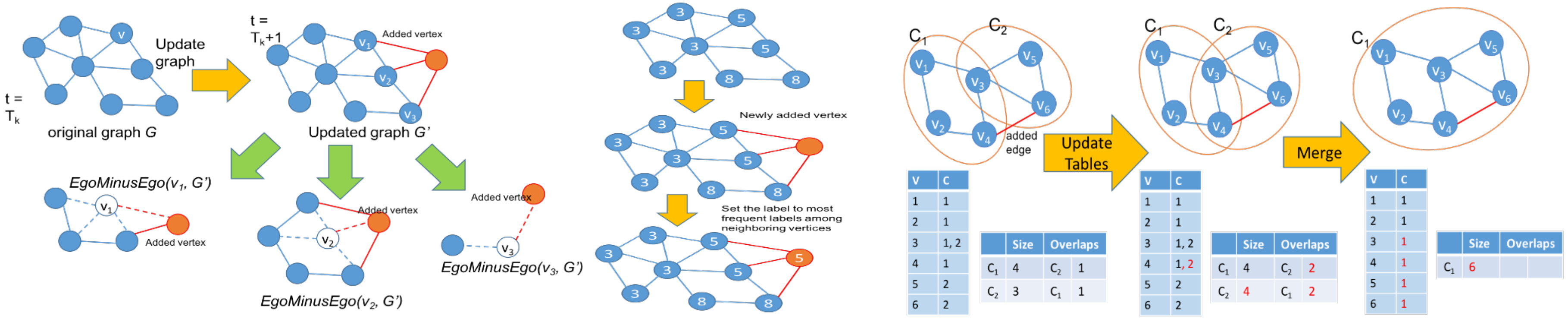




*Apache Spark: http://spark.apache.org/

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



V	C
1	1
2	1
3	1, 2
4	1
5	2
6	2

	Size	Overlaps	
C1	4	C2	1
C2	3	C1	1

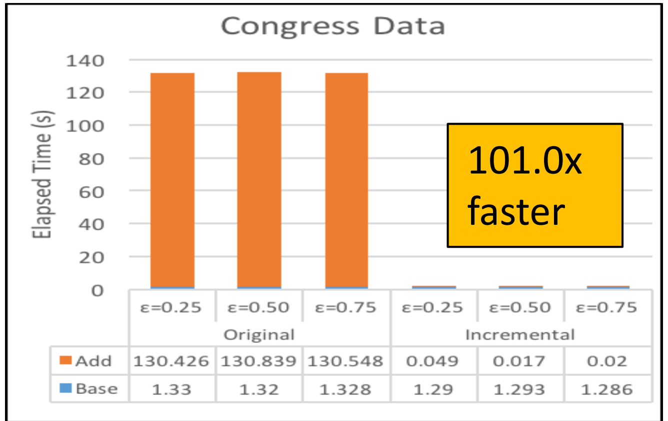
V	C
1	1
2	1
3	1, 2
4	1, 2
5	2
6	2

	Size	Overlaps	
C1	4	C2	2
C2	4	C1	2

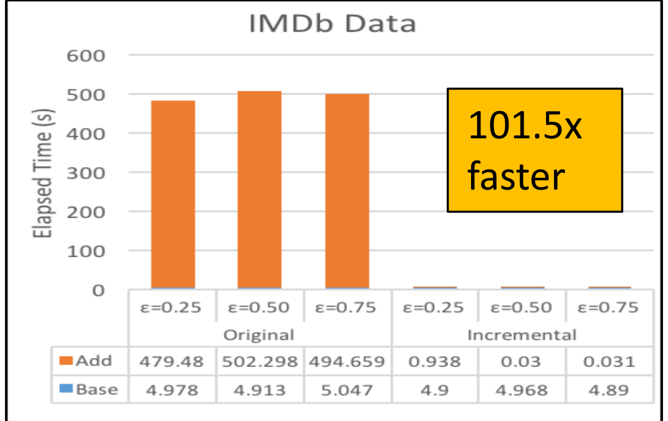
V	C
1	1
2	1
3	1
4	1
5	1
6	1

	Size	Overlaps
C1	6	

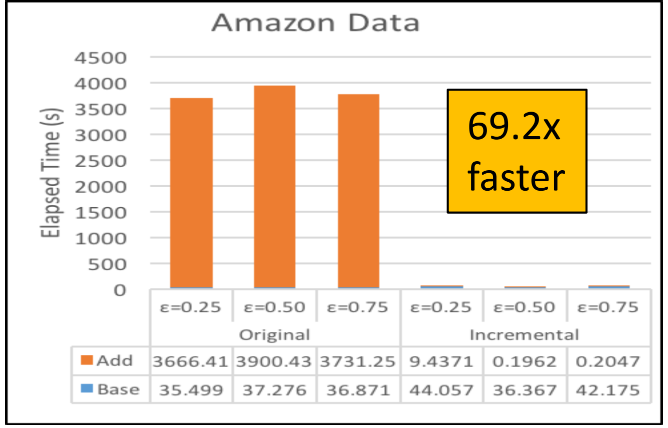
Hiroki Kanezashi and Toyotaro Suzumura, An Incremental Local-First Community Detection Method for Dynamic Graphs, Third International Workshop on High Performance Big Graph Data Management, Analysis, and Mining (BigGraphs 2016), to appear



	Original ε=0.25	Original ε=0.50	Original ε=0.75	Incremental ε=0.25	Incremental ε=0.50	Incremental ε=0.75
Add	130.426	130.839	130.548	0.049	0.017	0.02
Base	1.33	1.32	1.328	1.29	1.293	1.286



	Original ε=0.25	Original ε=0.50	Original ε=0.75	Incremental ε=0.25	Incremental ε=0.50	Incremental ε=0.75
Add	479.48	502.298	494.659	0.938	0.03	0.031
Base	4.978	4.913	5.047	4.9	4.968	4.89



	Original ε=0.25	Original ε=0.50	Original ε=0.75	Incremental ε=0.25	Incremental ε=0.50	Incremental ε=0.75
Add	3666.41	3900.43	3731.25	9.4371	0.1962	0.2047
Base	35.499	37.276	36.871	44.057	36.367	42.175

EBD Algorithm Kernels

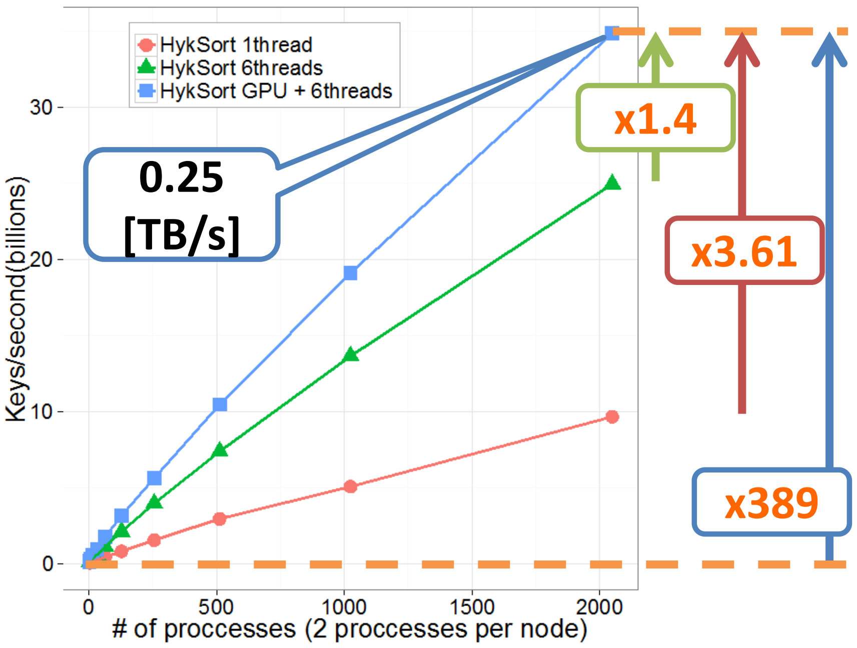
GPU-based Distributed Sorting [Shamoto, IEEE BigData 2014, IEEE Trans. Big Data 2015]

- 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



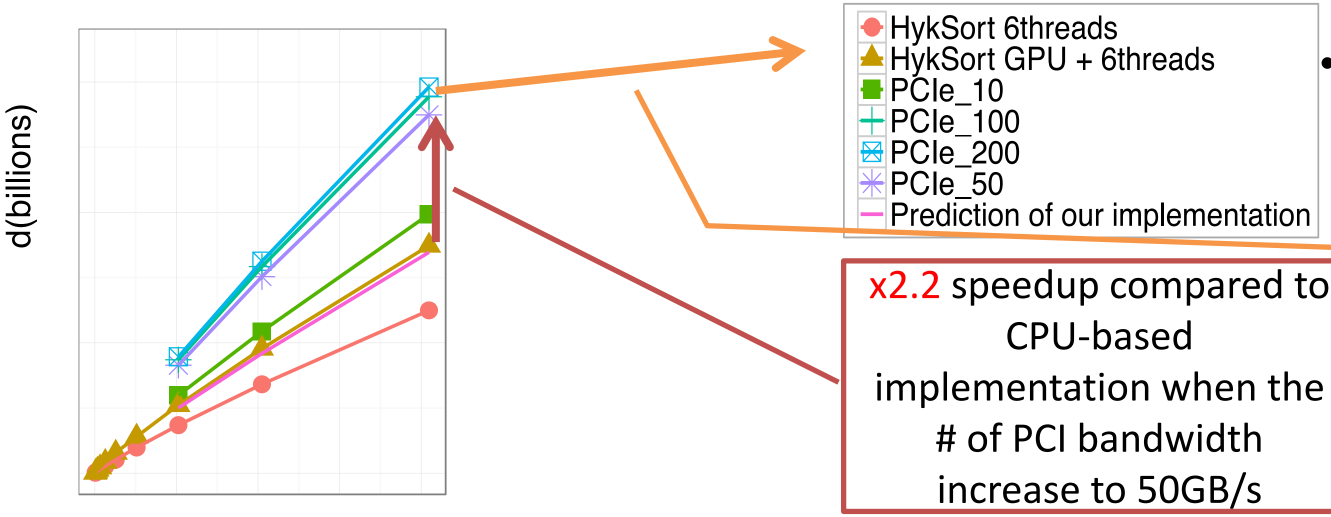
0.25 [TB/s]

x1.4

x3.61

x389

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

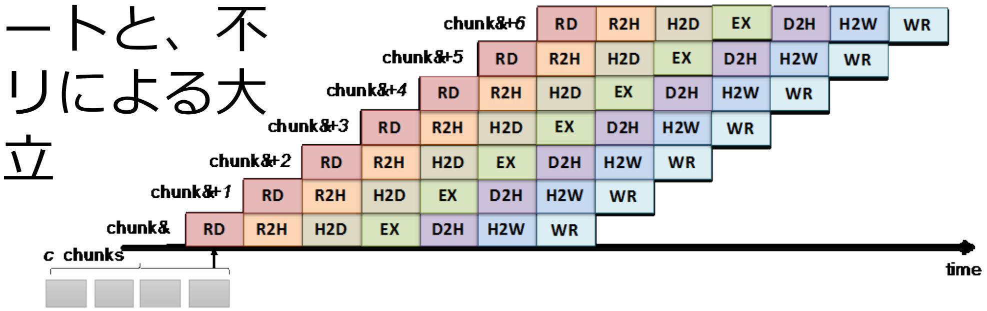
8.8% reduction of overall runtime when the accelerators work 4 times faster than K20x

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

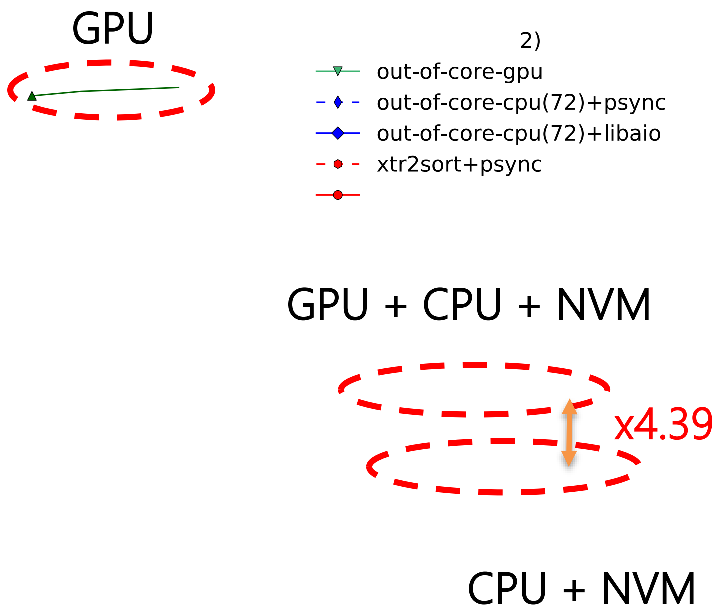
How to combine deepening memory layers for future HPC/Big Data workloads, targeting Post Moore Era?

- 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



chunk&+6 RD R2H H2D EX D2H H2W WR
chunk&+5 RD R2H H2D EX D2H H2W WR
chunk&+4 RD R2H H2D EX D2H H2W WR
chunk&+3 RD R2H H2D EX D2H H2W WR
chunk&+2 RD R2H H2D EX D2H H2W WR
chunk&+1 RD R2H H2D EX D2H H2W WR
chunk& RD R2H H2D EX D2H H2W WR
c chunks
time



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

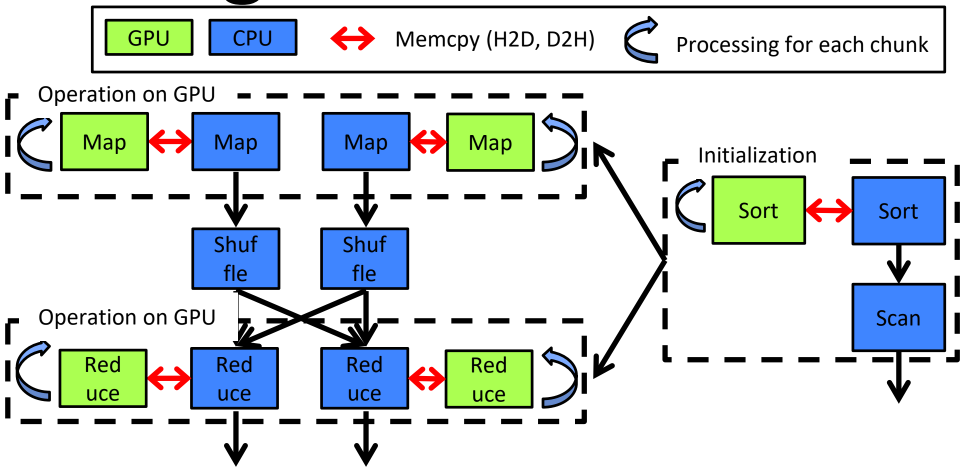
Problem: GPU memory capacity limits scalable large-scale graph processing

Proposal: Out-of-core GPU memory management on MapReduce

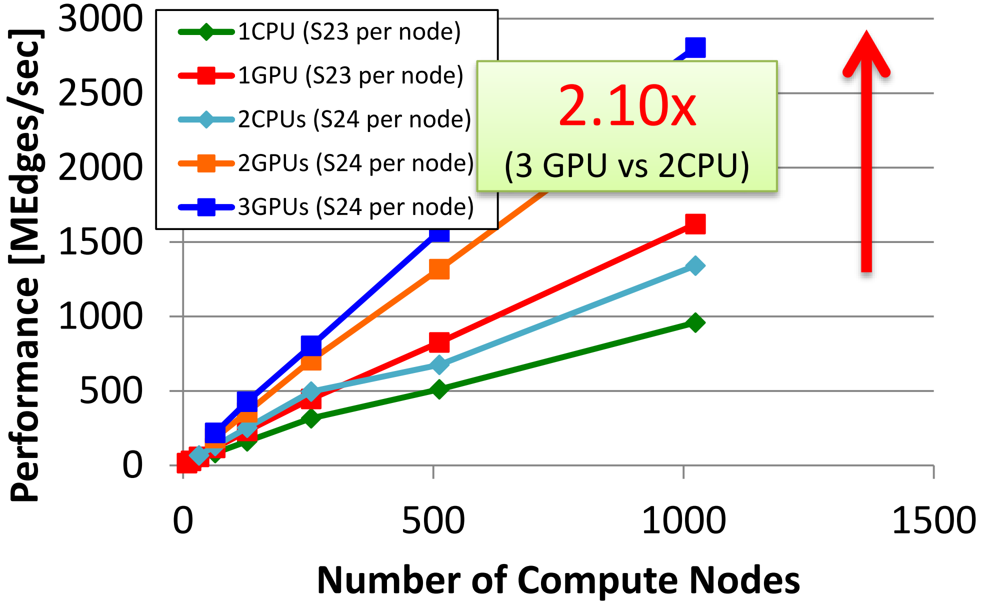
- Stream-based GPU MapReduce
- Out-of-core GPU sorting

Experimental Results:
performance improvement over CPUs

- Map: 1.41x, Reduce: 1.49x, Sort: 4.95x speedup
- Overlapping communication effectively

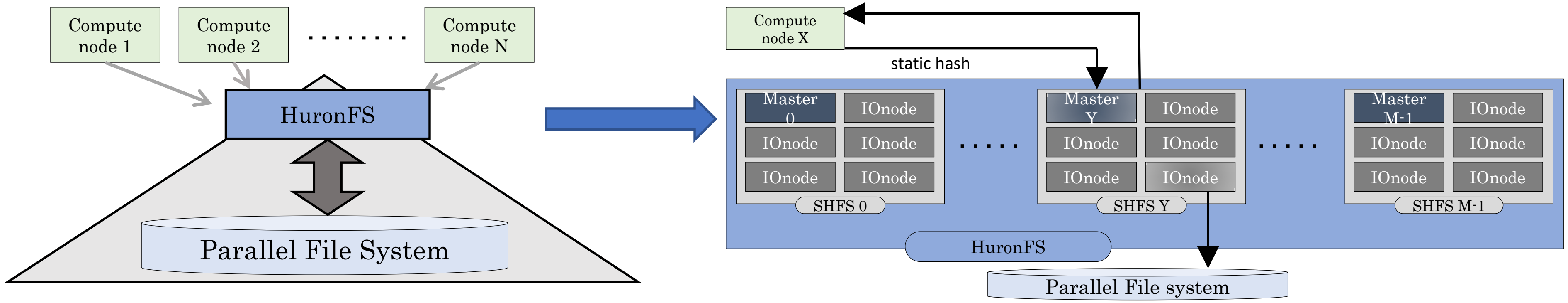


Weak scaling on TSUBAME2.5



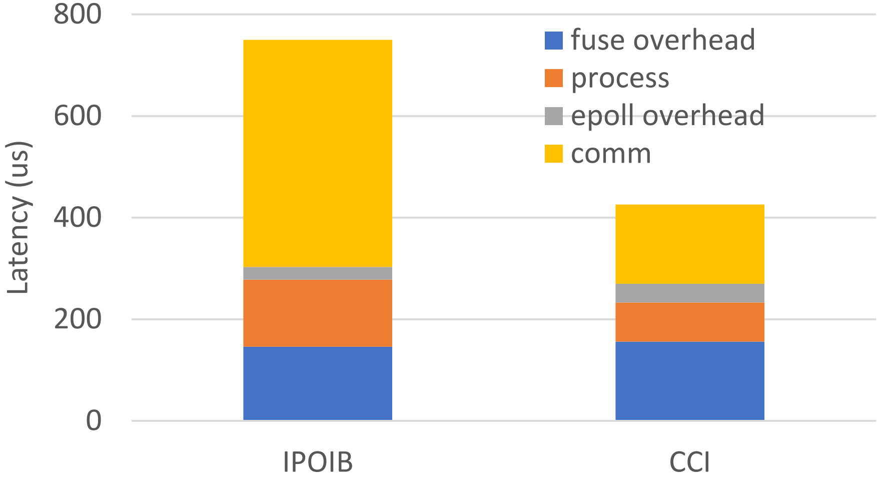
2.10x
(3 GPU vs 2CPU)

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

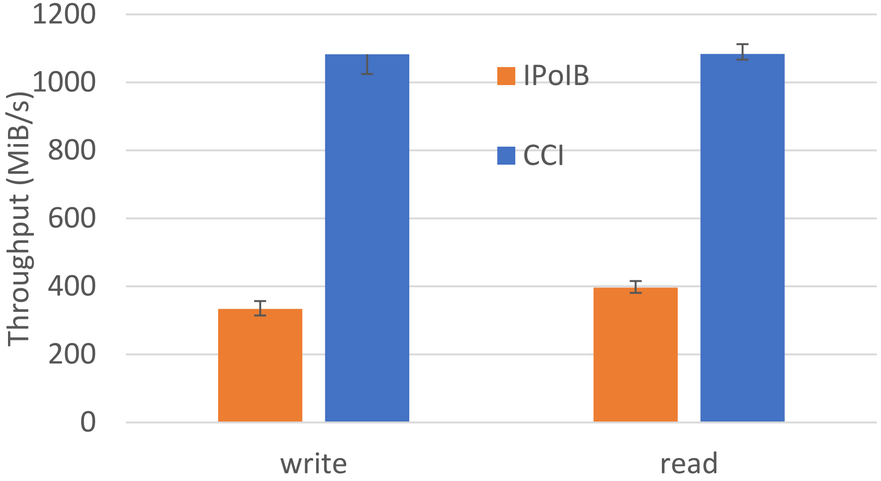


- HuronFS*:* dedicated dynamic instances to provide “burst buffer” for caching data
- I/O requests from *Compute Nodes* are forwarded to HuronFS
- The whole system consists of several SHFS (Sub HuronFS)
  - Workload are distributed among all the SHFS using hash of file path
- Each SHFS consists of a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

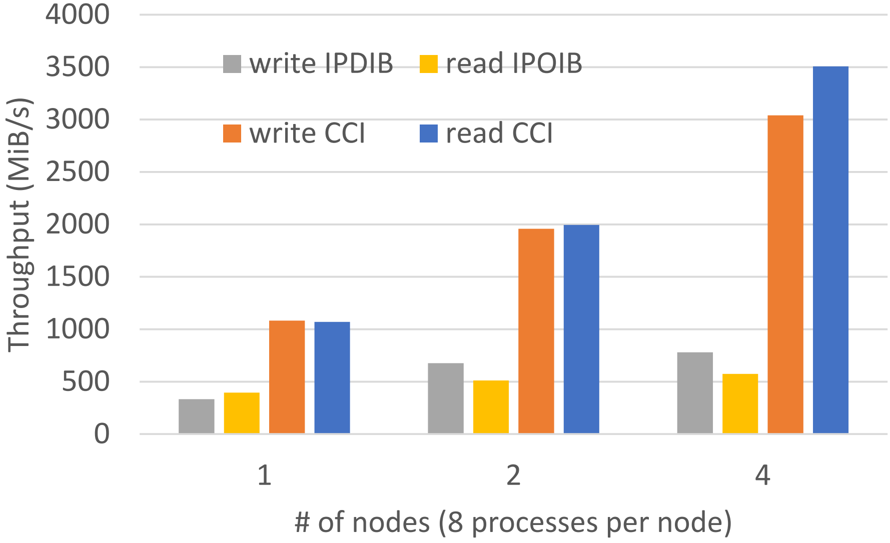


Latency



Throughput from single client

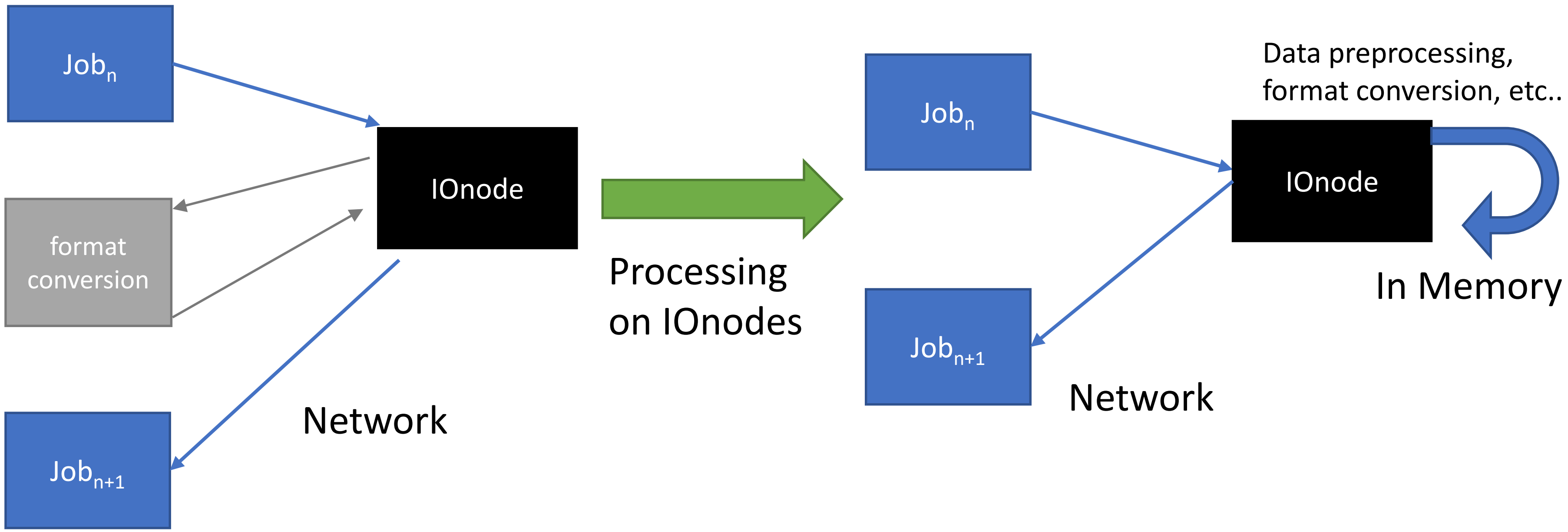
Inifinband	4X FDR 56 Gb/sec mellanox
CPU	Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz
Mem	251G



Throughput from single IOnode

Plans

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



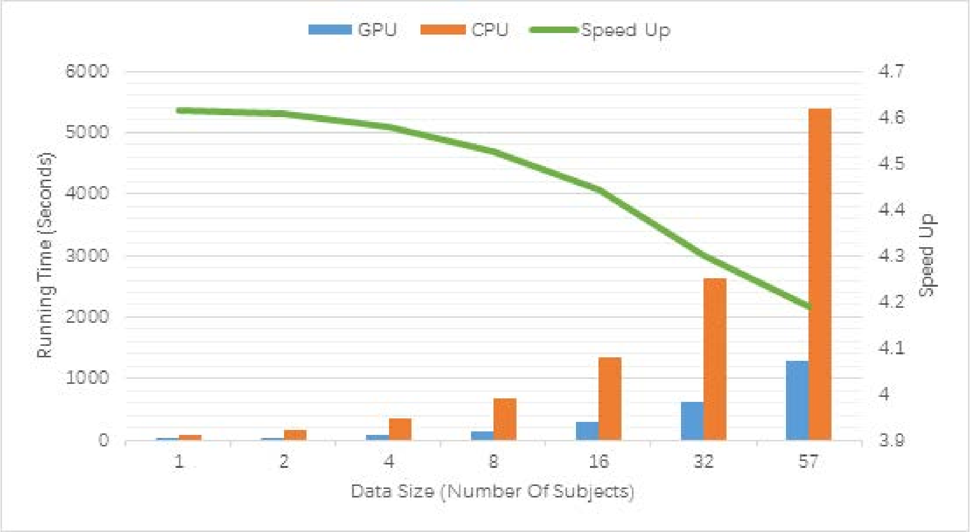
GPU-Based Fast Signal Processing for Large Amounts of Snore Sound Data

- 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

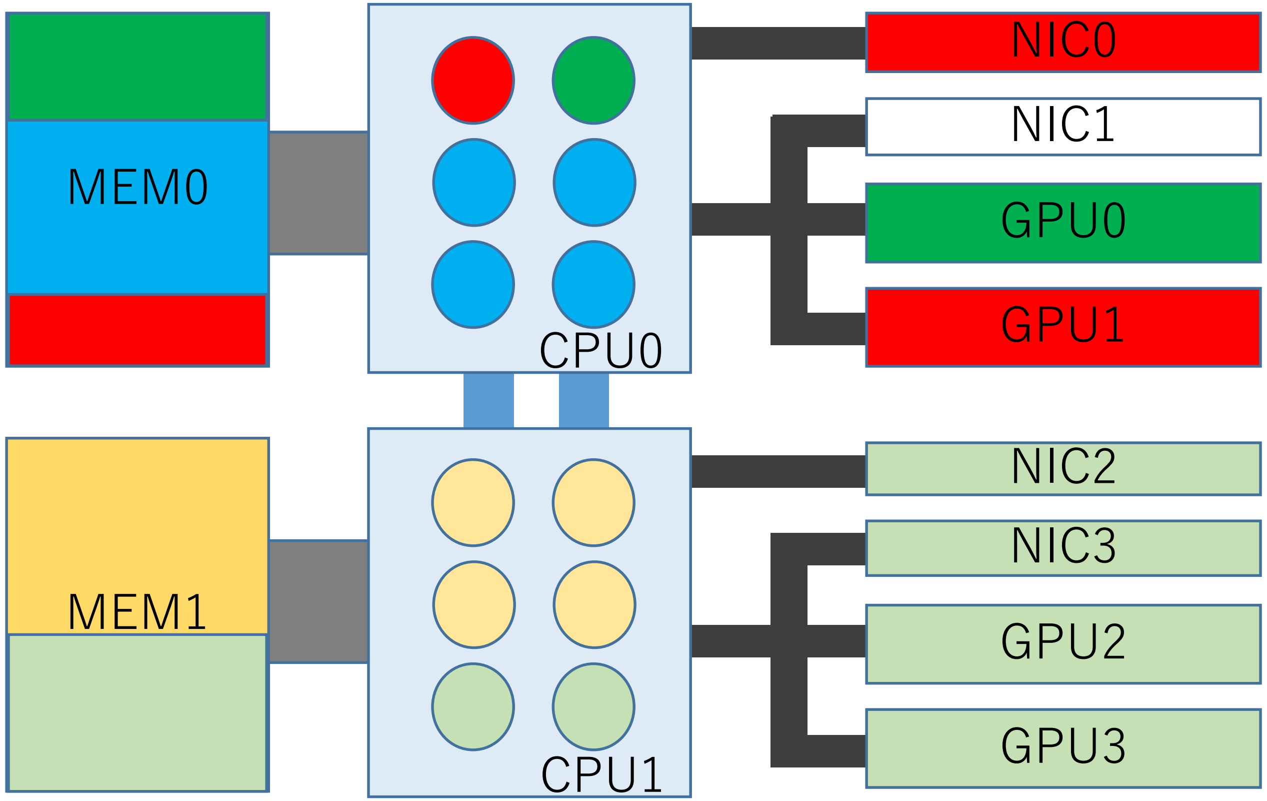
Subjects	Total Time (hours)	Data Size (GB)	Data format	Sampling Rate
57 (China + Germany)	187.75	31.10	WAV	16 kHz, Mono

- 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.

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

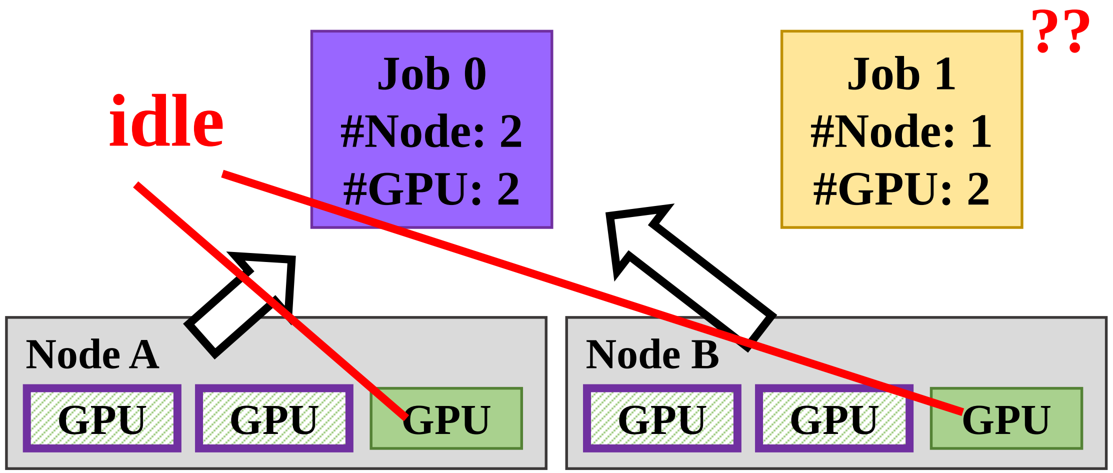
Resource Isolation via UGE Containers (future Docker etc.)

Container configuration and deployment tied to Univa Grid Engine

Background

A kind of resource assignment fragmentation

Multi-GPU batch-queue systems have many idle GPUs despite having jobs waiting, due to the *scattered idle-GPU problem* [1].



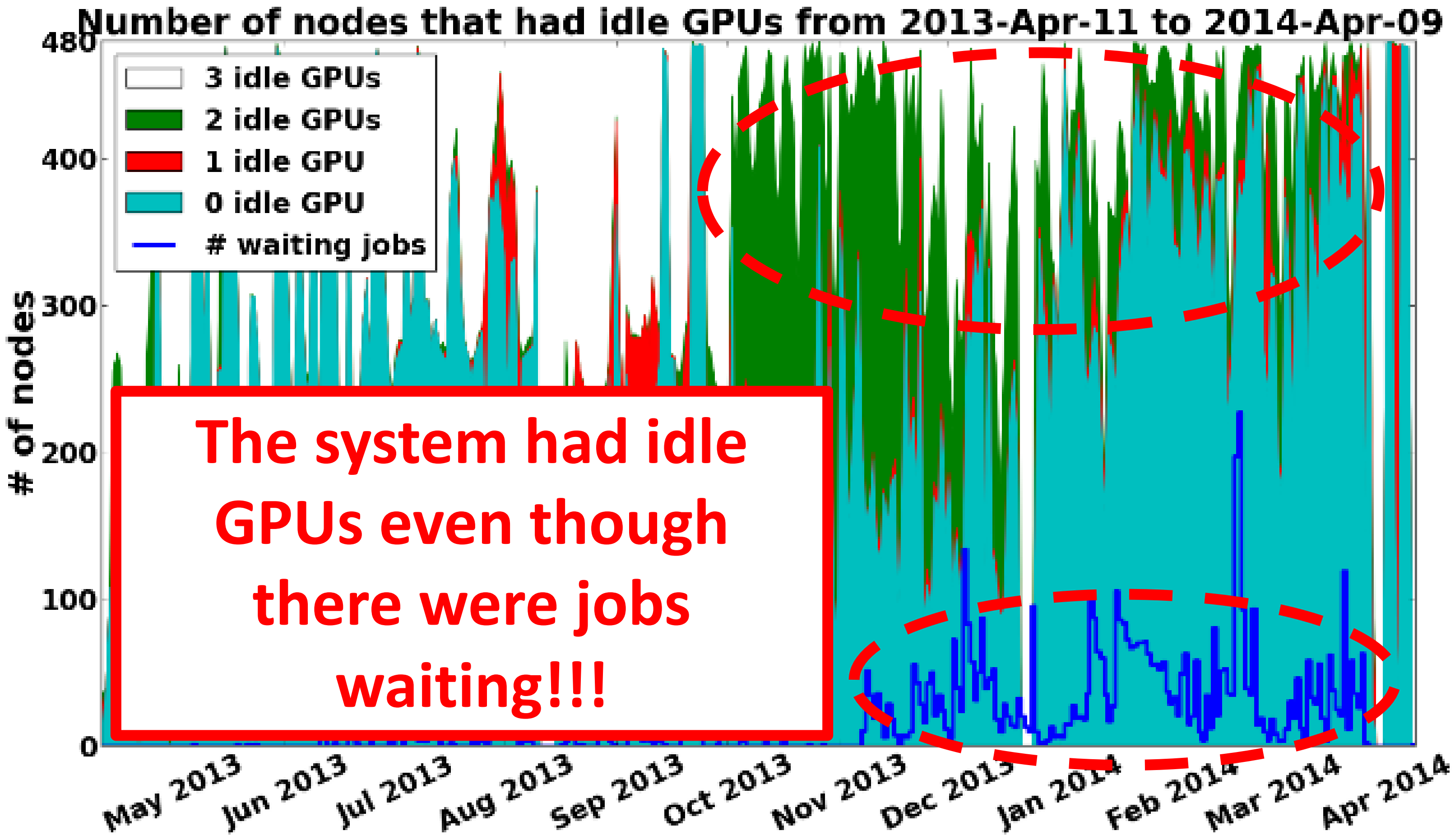
Scenario: Job 1 requests two GPUs on one node but each node has only one unoccupied GPU left.

Result: Job 1 cannot run and two GPUs are left idle.

[1] P. Markthub et al. “Using rCUDA to Reduce GPU Resource-Assignment Fragmentation Caused by Job Scheduler,” PDCAT2014

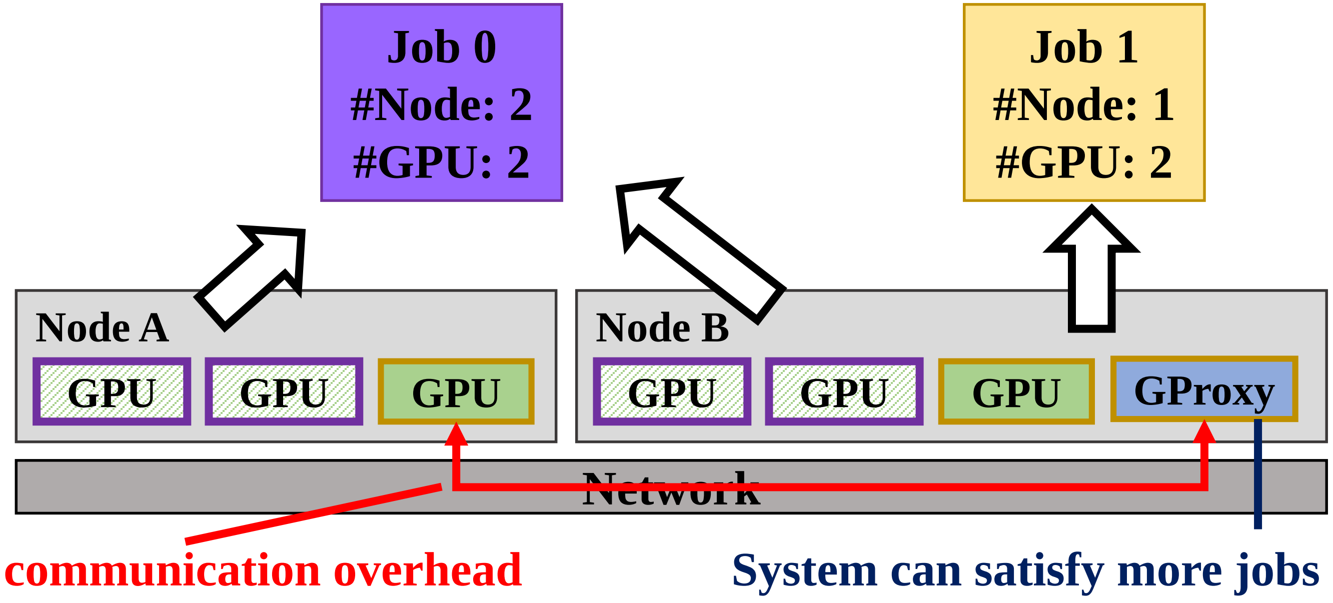
Idle-GPU Problem in Multi-GPU Batch-Queue Systems

TSUBAME2.5’s G Queue (GPU Queue)



The system had idle GPUs even though there were jobs waiting!!!

Previous Solution & Problems



Previous Solution [1]:

- Enable the system to serve more jobs by creating a GPU proxy that links with a remote GPU.
- Proven to reduce job waiting times as much as 25%.

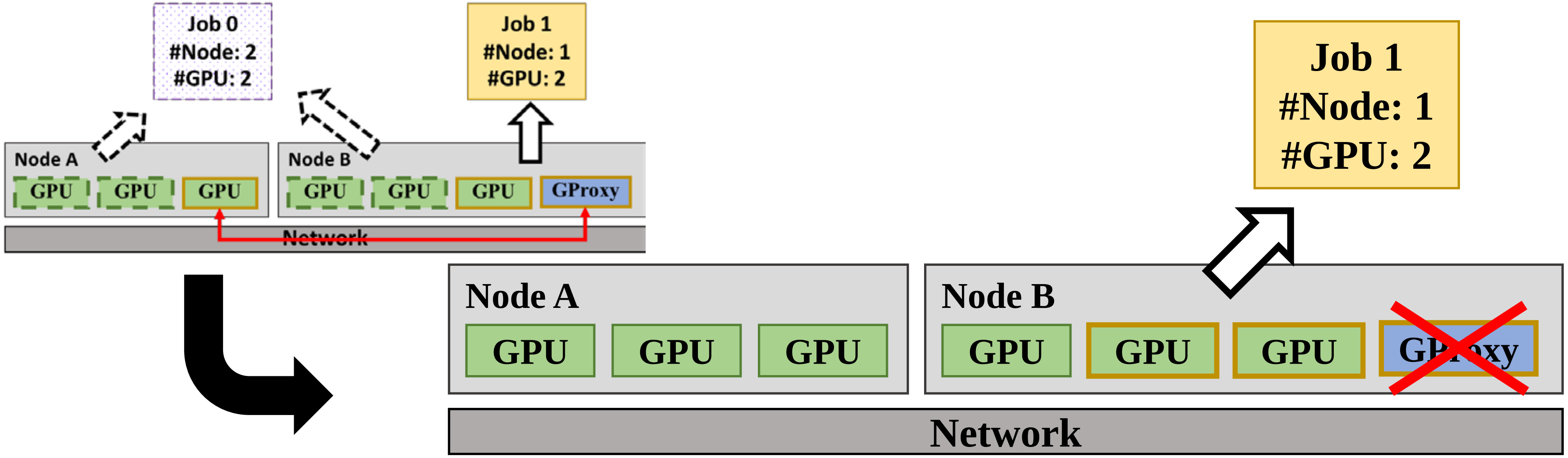
Problems:

- Remote GPU execution overhead
- Network congestion

The execution times of GPU communication intensive applications (e.g. LAMMPS, SRAD) may increase more than 5 times!!!

[1] P. Markthub et al. “Using rCUDA to Reduce GPU Resource-Assignment Fragmentation Caused by Job Scheduler,” PDCAT2014

New Solution Overview



Migrate execution on a remote GPU to a local GPU when it becomes available can solve the performance problems

Propose:

Low-overhead remote GPU execution middleware

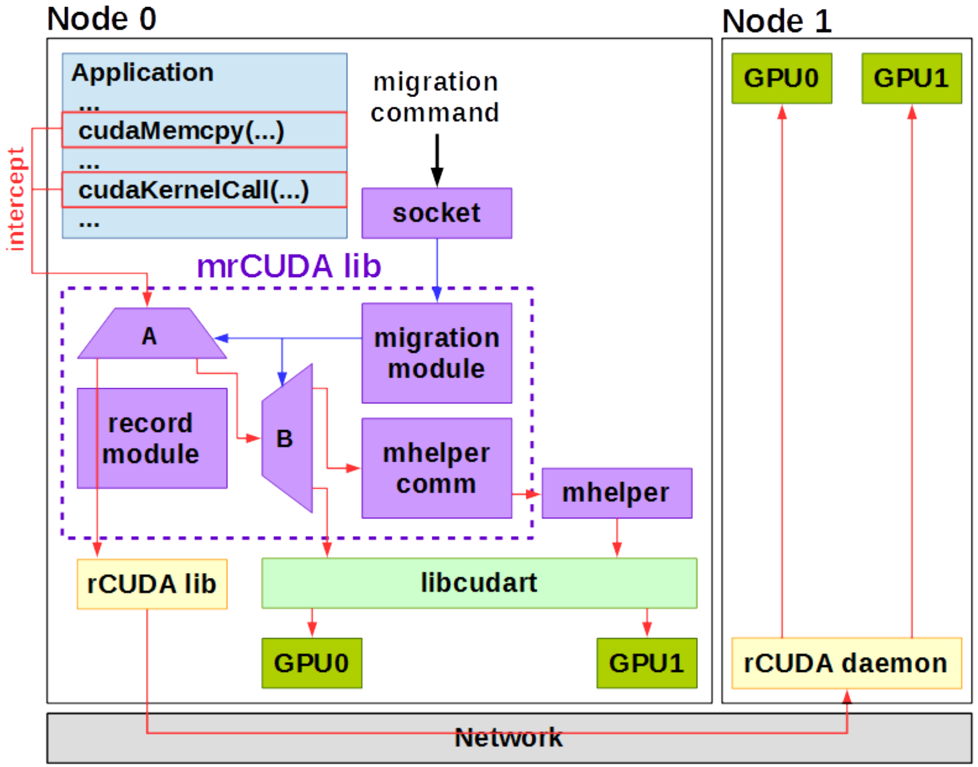
1. **mrCUDA:** an extension of *rCUDA* [1] to enable remote-to-local GPU migration.
2. **MRQ:** a heuristic extension of job scheduling algorithms to make the best out of mrCUDA.

[1] F. Silla, "Is remote GPU virtualization useful?" http://rcuda.net/pub/rCUDA barna 15.pdf, September 2015.

mrCUDA

Objective: Enable seamless and on-demand remote-to-local GPU migration on rCUDA

- rCUDA handles remote GPU execution, while mrCUDA handles GPU migration.
- GPU migration starts after mrCUDA receives a migration command via its special UNIX socket.



*A,B: Selectors

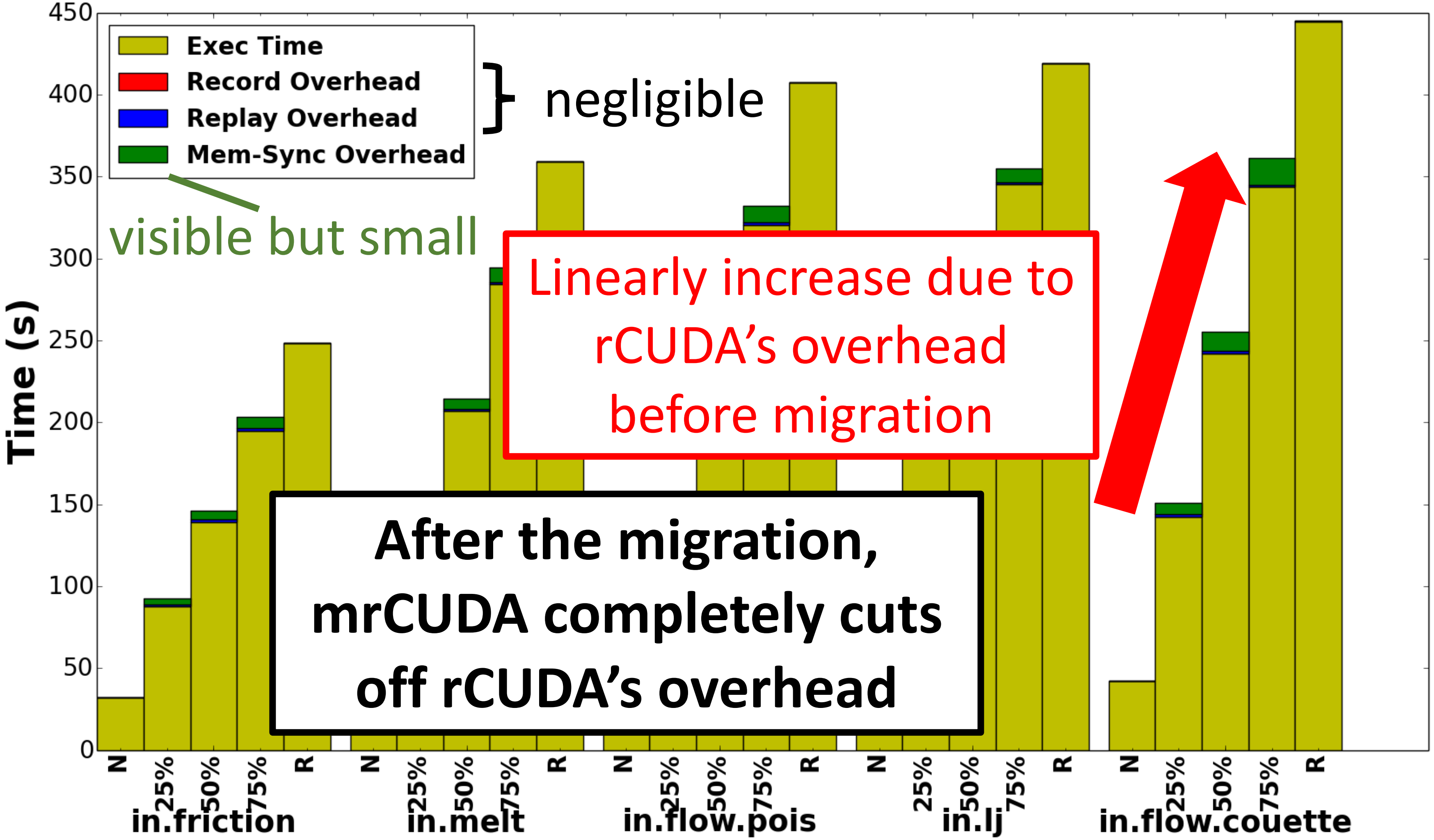
Migration Algorithm – a modified version of Replay Method [1]:

- Intercept all CUDA invocations.
- **Before migration:** Pass all intercepted calls to *rCUDA* while recording some CUDA calls (e.g. cudaMalloc).
- **During migration:** Replay the recorded calls in order and memsync GPU data.
- **After migration:** Pass all intercepted calls to *libcudart* without recording.

To recreate remote GPU's states on local GPU

[1] A. Nukada et al. "NVCR: A transparent checkpoint-restart library for NVIDIA CUDA," IPDPWS2011

Case Study: Migrating remote CUDA Execution of LAMMPS



} negligible

visible but small

Linearly increase due to rCUDA's overhead before migration

After the migration, mrCUDA completely cuts off rCUDA's overhead

*2 nodes, Tesla K20c, InfiniBand 4xFDR

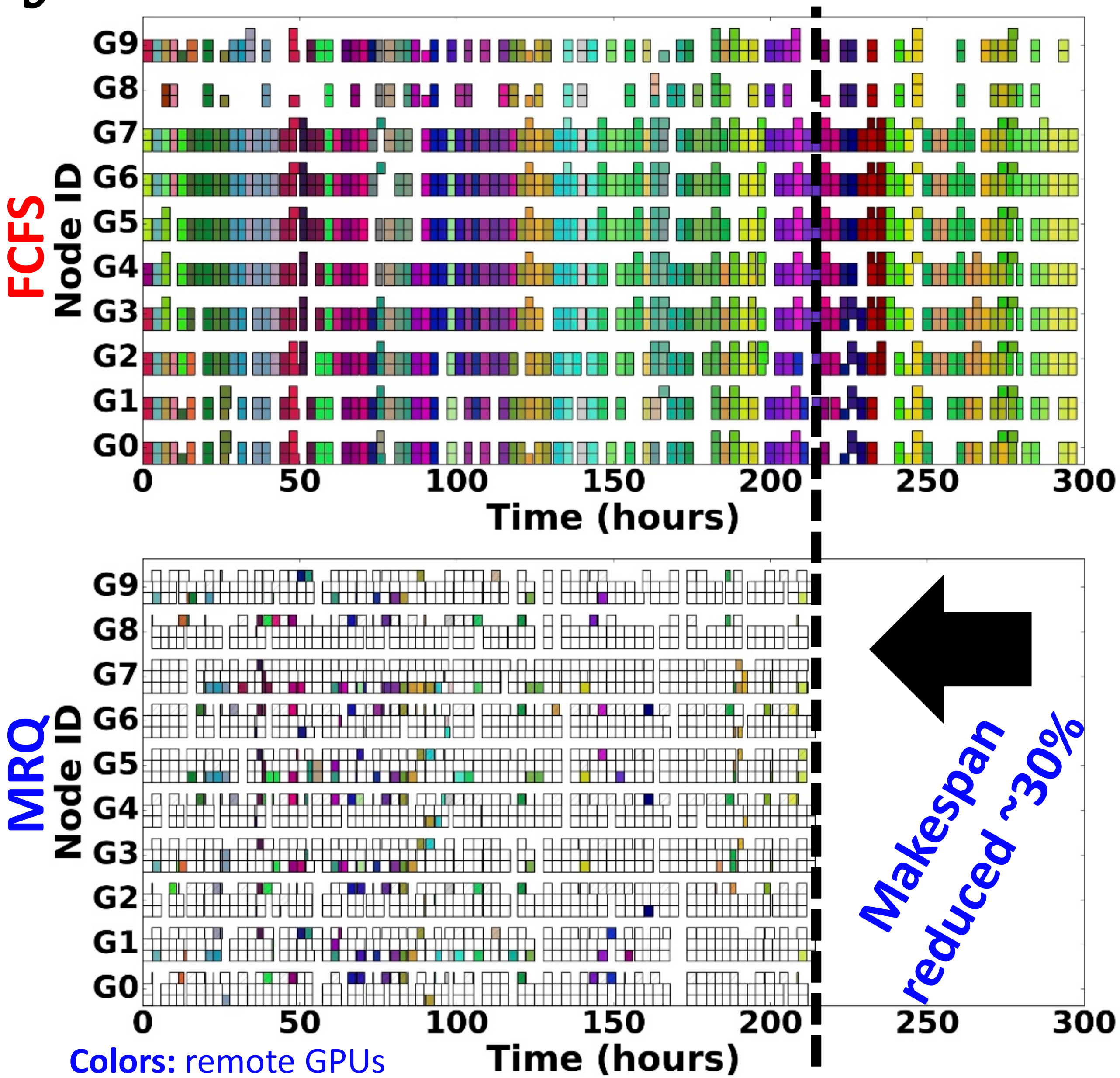
***x%:** migrate after finish x% of total iterations

GPU Occupancy Patterns

Systems can server more jobs concurrently with MRQ.

MRQ uses the same scheduling policy as FCFS.

Jobs do not experience significant execution time expansion, mainly thanks to GPU migration.



Makespan reduced ~30%

Colors: remote GPUs

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

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

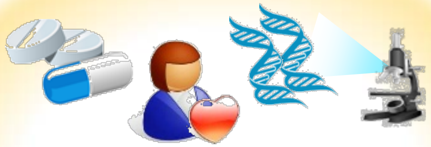
To complete the learning phase in one day

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]

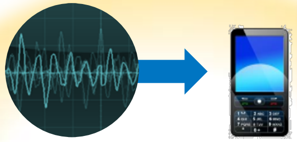
Bio / Healthcare



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

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

Image Recognition

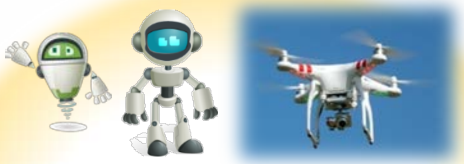


10P～ Flops
1万人の5000時間分の音声データ
人工的に生成された10万時間の
音声データを基に学習 [Baidu 2015]

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

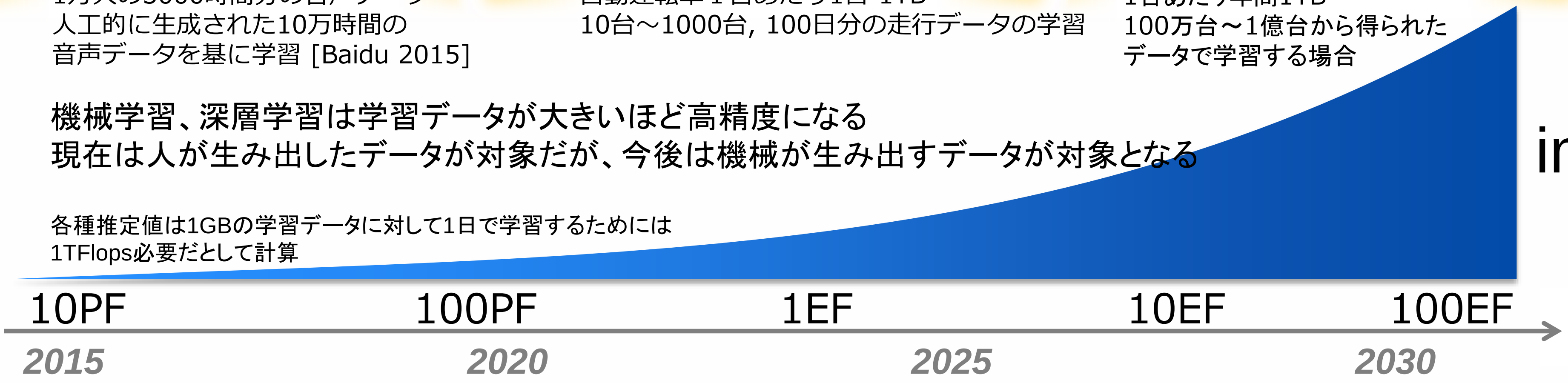
Robots / Drones



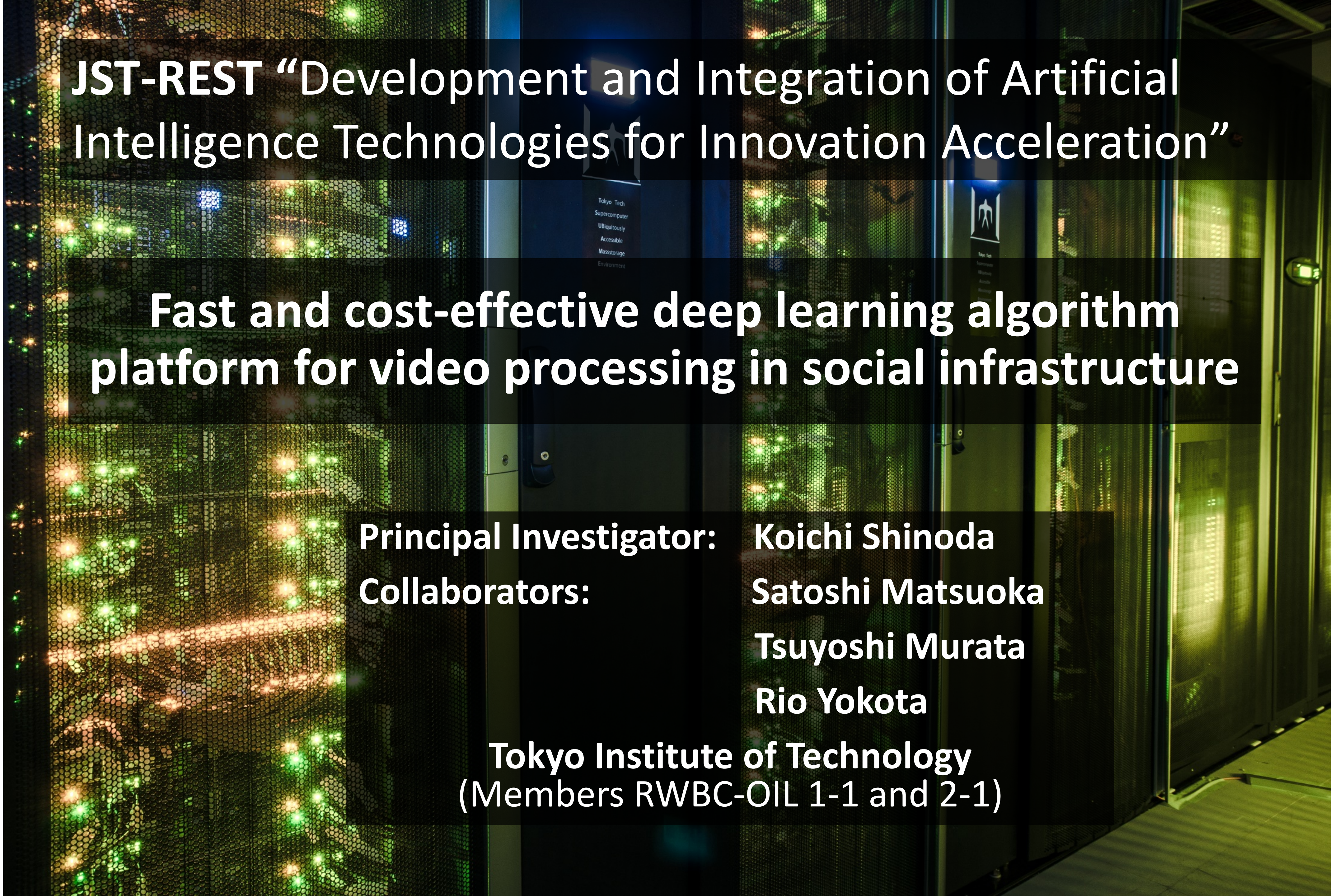
1E～100E Flops
1台あたり年間1TB
100万台～1億台から得られた
データで学習する場合

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



So both are important in the infrastructure



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

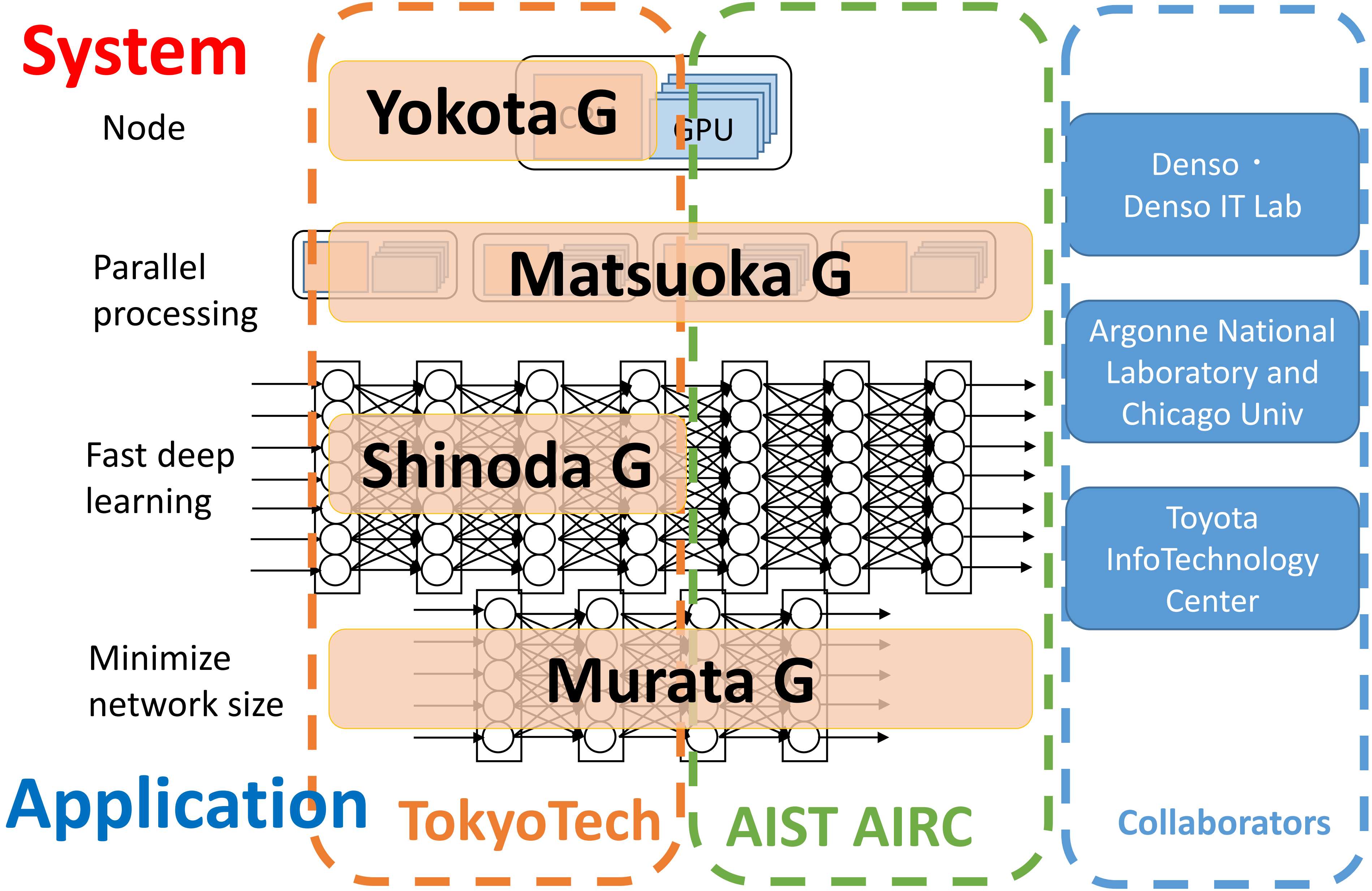
- Video processing in smart society for safety and security
  - Intelligent transport systems
    Drive recorder video
  - Security systems
    Surveillance camera video
- Deep learning
  - Much higher performance than before
  - IT giants with large computational resources has formed a monopoly

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



System

Node

Parallel processing

Fast deep learning

Minimize network size

Application

Yokota G

Matsuoka G

Shinoda G

Murata G

TokyoTech

AIST AIRC

Denso・
Denso IT Lab

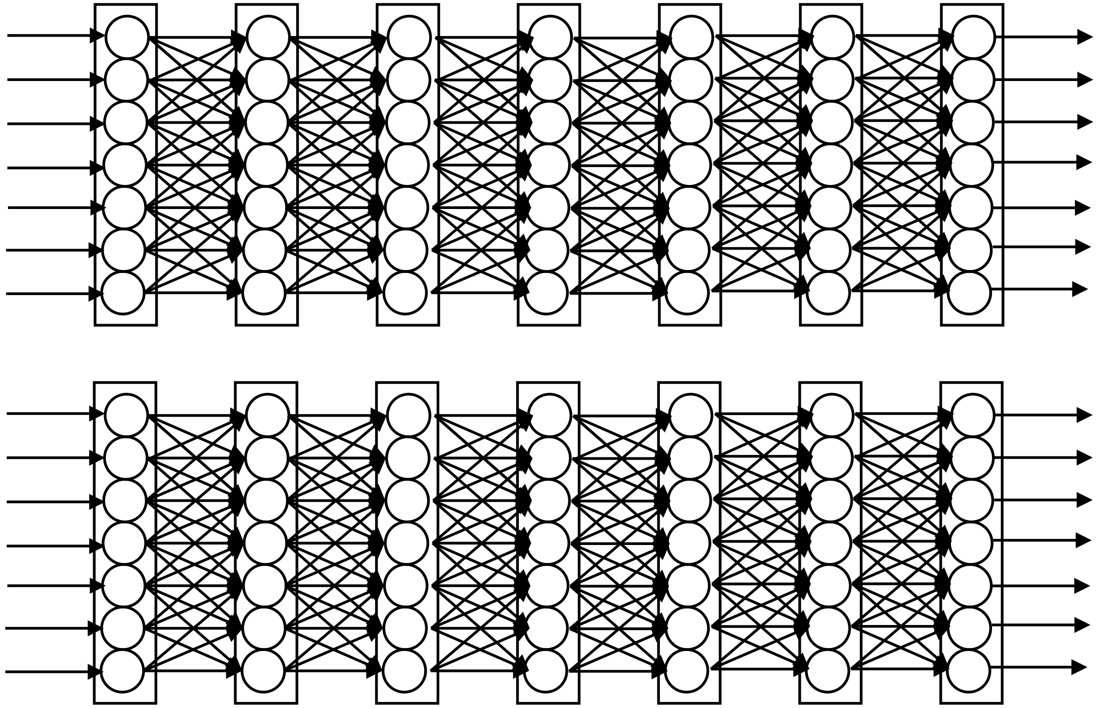
Argonne National Laboratory and Chicago Univ

Toyota InfoTechnology Center

Collaborators

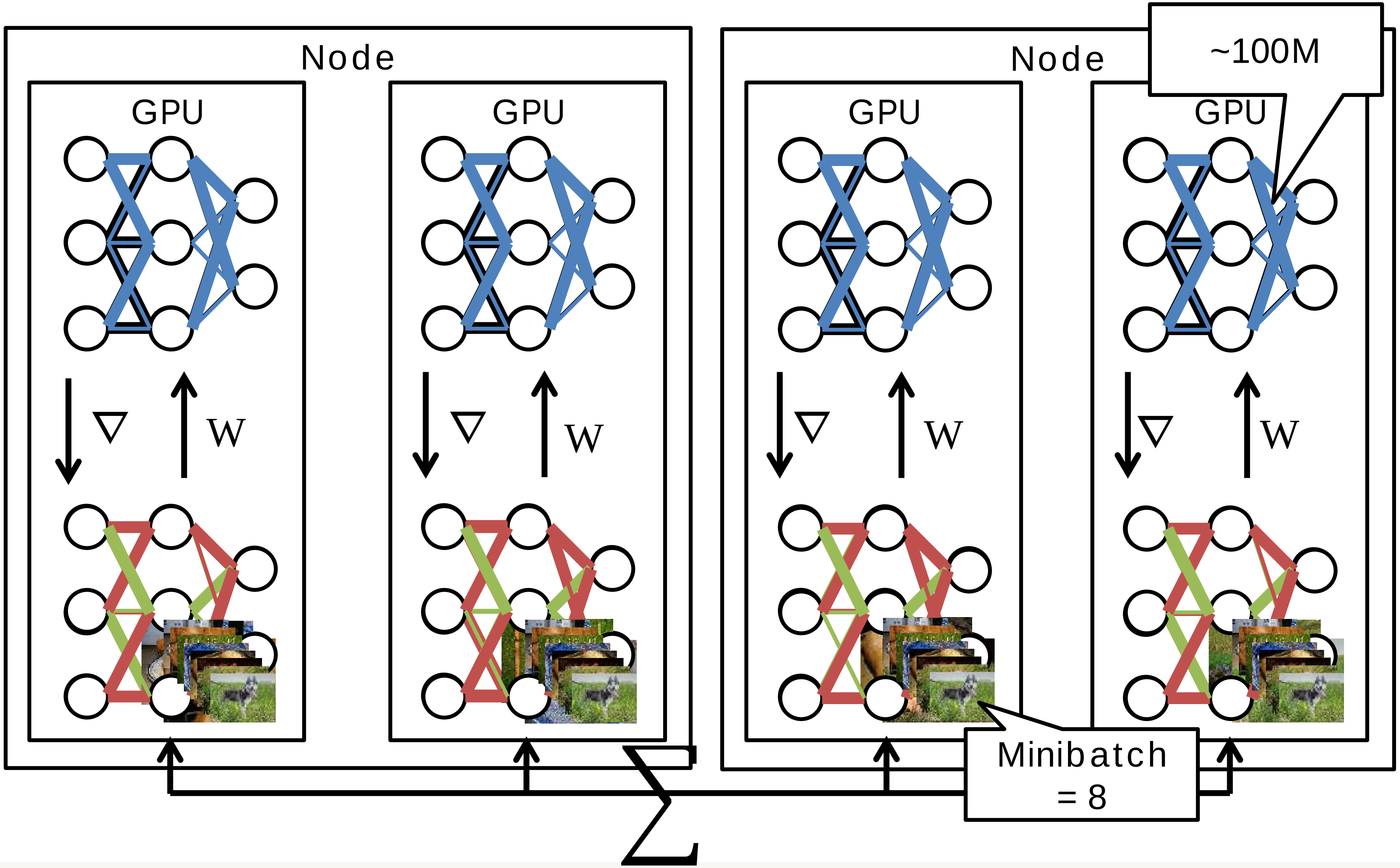
4 Layers of Parallelism in DNN Training

- Hyper Parameter Search
  - Searching optimal network configurations and parameters
  - Often use evolutionary algorithms
- Data Parallelism
  - Split and parallelize the batch data
  - Synchronous, asynchronous, hybrid, …
- Model Parallelism
  - Split and parallelize the layer calculations in forward/backward propagation
- ILP and other low level Parallelism
  - Parallelize the convolution operations etc. (in reality matrix multiply)



Parallelizing Deep Neural Network Training

Data Parallel SGD(Stochastic Gradient Descent)



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

Background

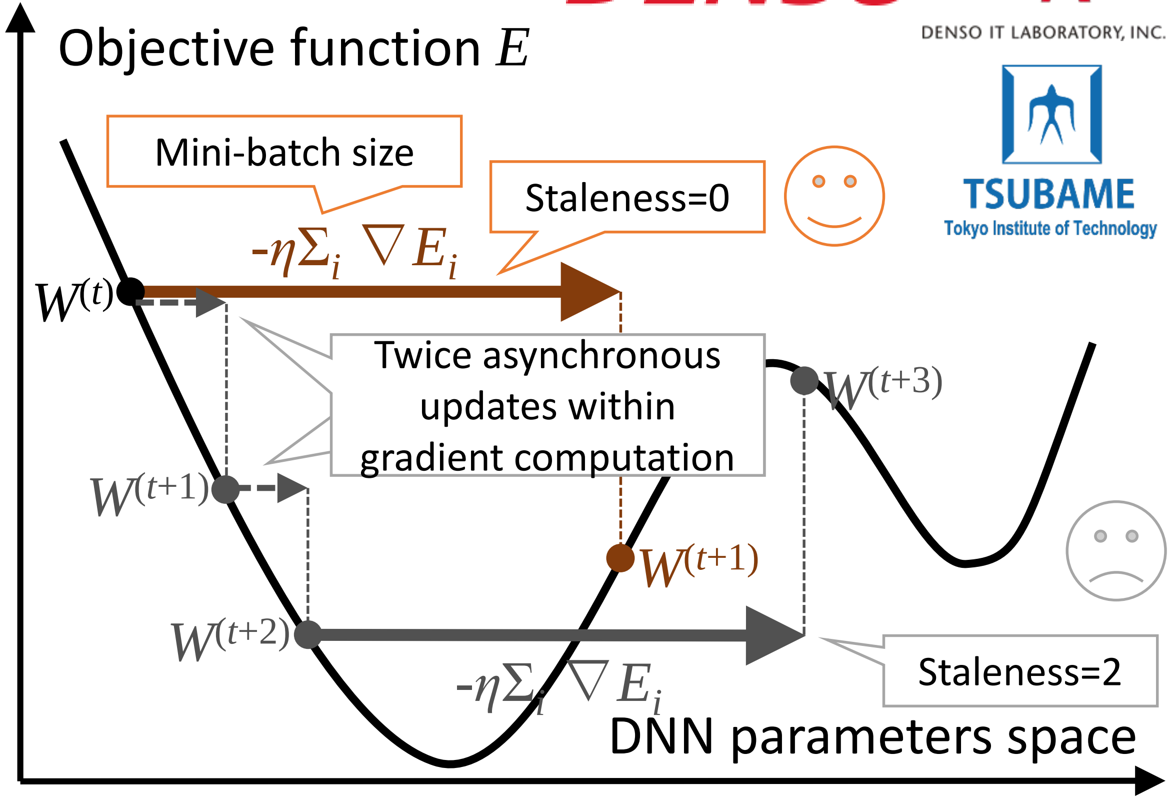
- In large-scale Asynchronous Stochastic Gradient Descent (ASGD), mini-batch size and gradient staleness tend to be large and unpredictable, which increase the error of trained DNN

DENSO ITLAB

DENSO IT LABORATORY, INC.

TSUBAME

Tokyo Institute of Technology



Objective function E

Mini-batch size

Staleness=0

$-\eta\Sigma_i \nabla E_i$

$W^{(t)}$

Twice asynchronous updates within gradient computation

$W^{(t+3)}$

$W^{(t+1)}$

$W^{(t+1)}$

$W^{(t+2)}$

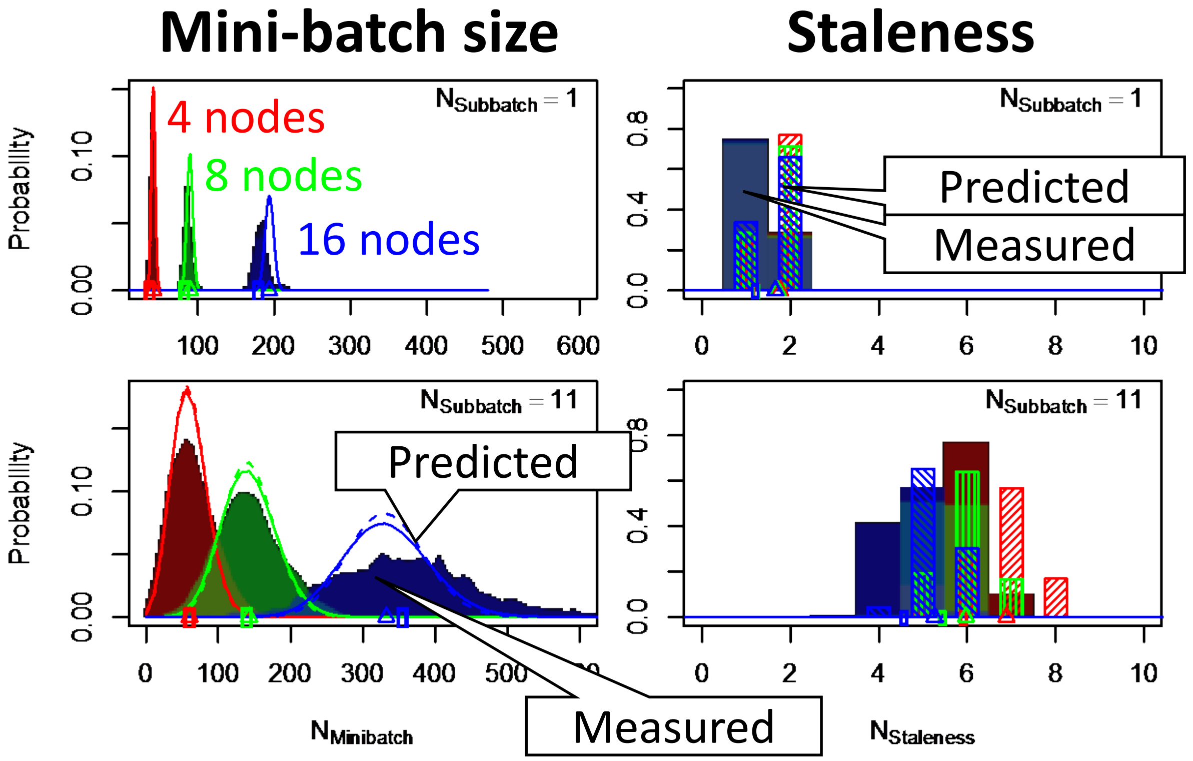
$-\eta\Sigma_i \nabla E_i$

Staleness=2

DNN parameters space

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 flops, but also **NW injection bandwidth 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

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)

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

AIRC

Riken -AIP

NICT-UCRI

AI Venture Startups

Preferred Networks

UEI

R&D ML Algorithms & SW

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

IT LAB DENSO IT LABORATORY, INC.

Panasonic

ABEJA

MIZUHO みずほ情報総研

DeNA

ZMP

Seeking Innovative Application of AI & Data

AI/BD Centers & Labs in National Labs & Universities

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

Over $1B Govt. AI investment over 10 years

Use of Massive Scale Data now Wasted

SAKURA Internet

Google Cloud Platform

Windows Azure

amazon web services

DENSO

Petabytes of Drive Recording Video

FA&ロボット&ロボマシン FANUC

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

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

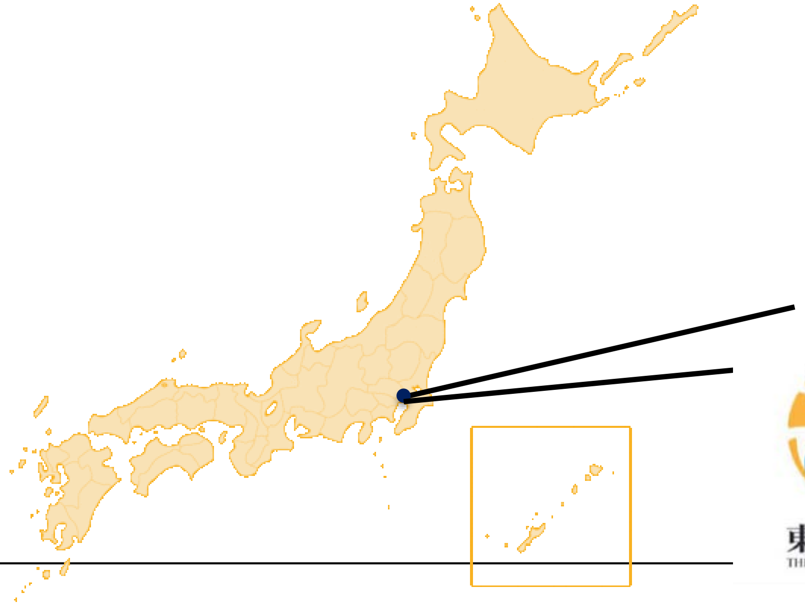
AI&Data Infrastructures

Massive "Big" Data in Training

"Big" Data

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

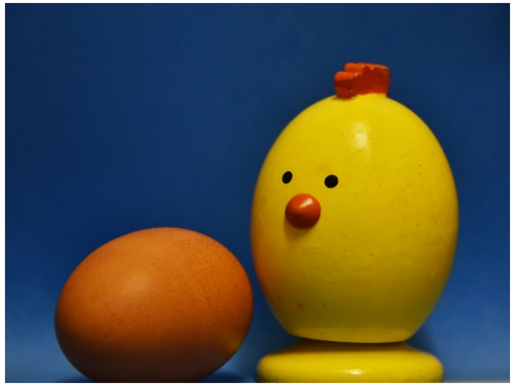
- **ABCI:** AI Bridging Cloud Infrastructure
  - Top-Level SC compute & data capability for DNN (130˜200 AI-Petaflops)
  - Open Public & Dedicated infrastructure for AI & Big Data Algorithms, Software and Applications
  - Platform to accelerate joint academic-industry R&D for AI in Japan



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

Univ. Tokyo Kashiwa Campus

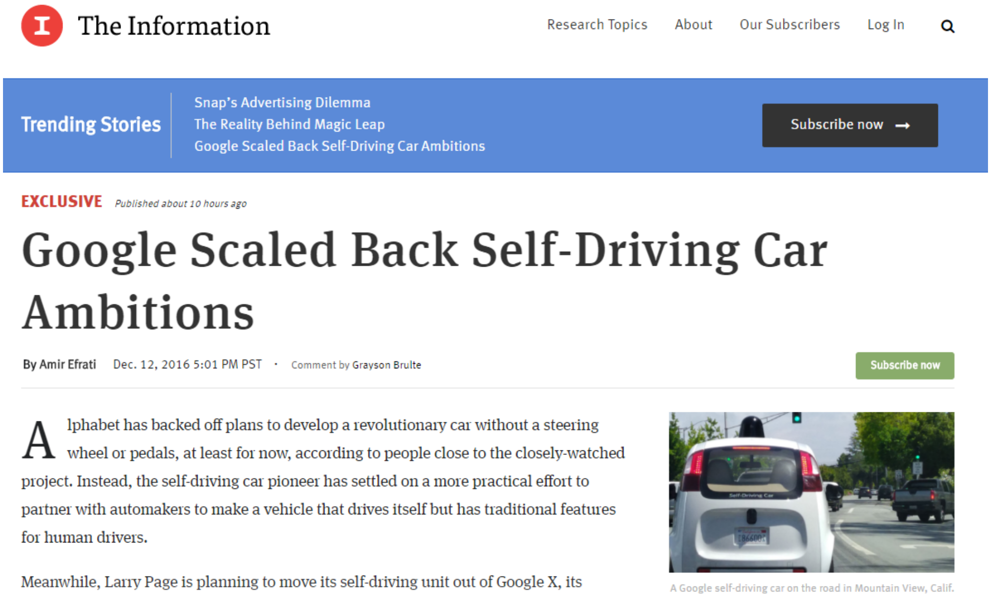
The “Chicken or Egg Problem” of AI-HPC Infrastructures



- “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
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

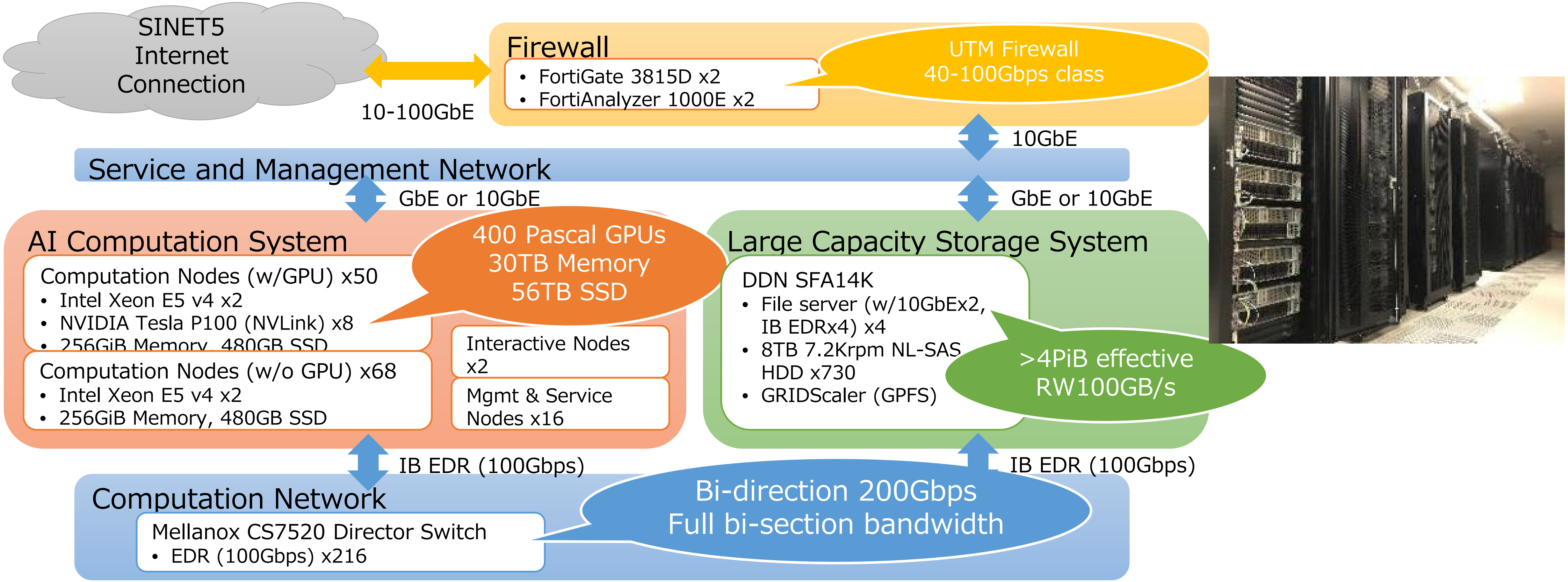
Alphabet has backed off plans to develop a revolutionary car without a steering wheel or pedals, at least for now, according to people close to the closely-watched project. Instead, the self-driving car pioneer has settled on a more practical effort to partner with automakers to make a vehicle that drives 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 (#3 June 2017 Green 500)

- **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.



SINET5 Internet Connection

10-100GbE

Firewall
- FortiGate 3815D x2
- FortiAnalyzer 1000E x2

UTM Firewall
40-100Gbps class

10GbE

Service and Management Network

GbE or 10GbE

GbE or 10GbE

AI Computation System

400 Pascal GPUs
30TB Memory
56TB SSD

Computation Nodes (w/GPU) x50
- Intel Xeon E5 v4 x2
- NVIDIA Tesla P100 (NVLink) x8
- 256GiB Memory, 480GB SSD

Computation Nodes (w/o GPU) x68
- Intel Xeon E5 v4 x2
- 256GiB Memory, 480GB SSD

Interactive Nodes x2

Mgmt & Service Nodes x16

Large Capacity Storage System

DDN SFA14K
- File server (w/10GbEx2, IB EDRx4) x4
- 8TB 7.2Krpm NL-SAS HDD x730
- GRIDScaler (GPFS)

>4PiB effective
RW100GB/s

IB EDR (100Gbps)

IB EDR (100Gbps)

Computation Network

Mellanox CS7520 Director Switch
- EDR (100Gbps) x216

Bi-direction 200Gbps
Full bi-section bandwidth

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

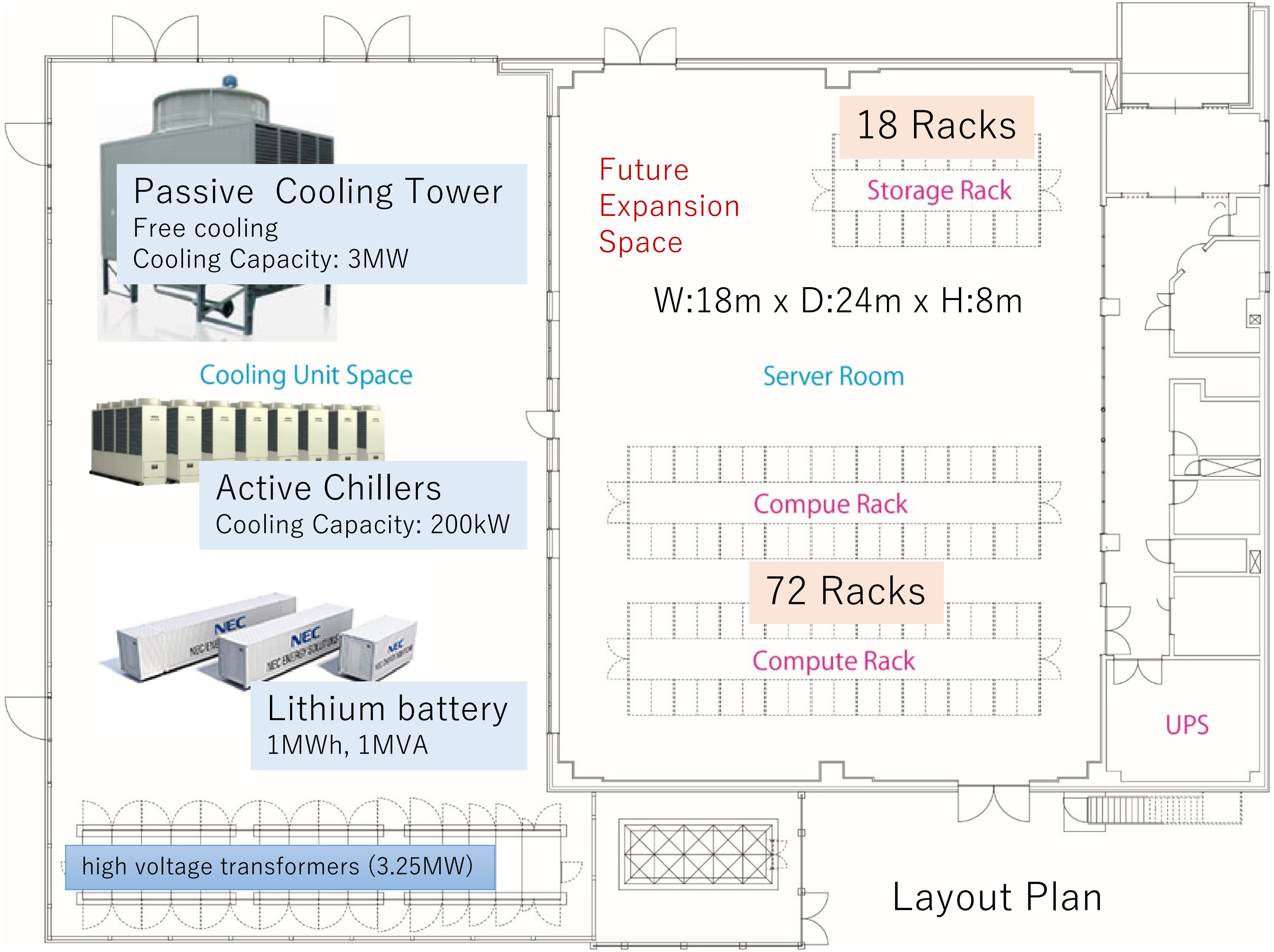
イメージスケッ

Reference Image

引用元: NEC導入事例

ABCI Cloud Data Center

“Commoditizing 60KW/rack Supercomputer”

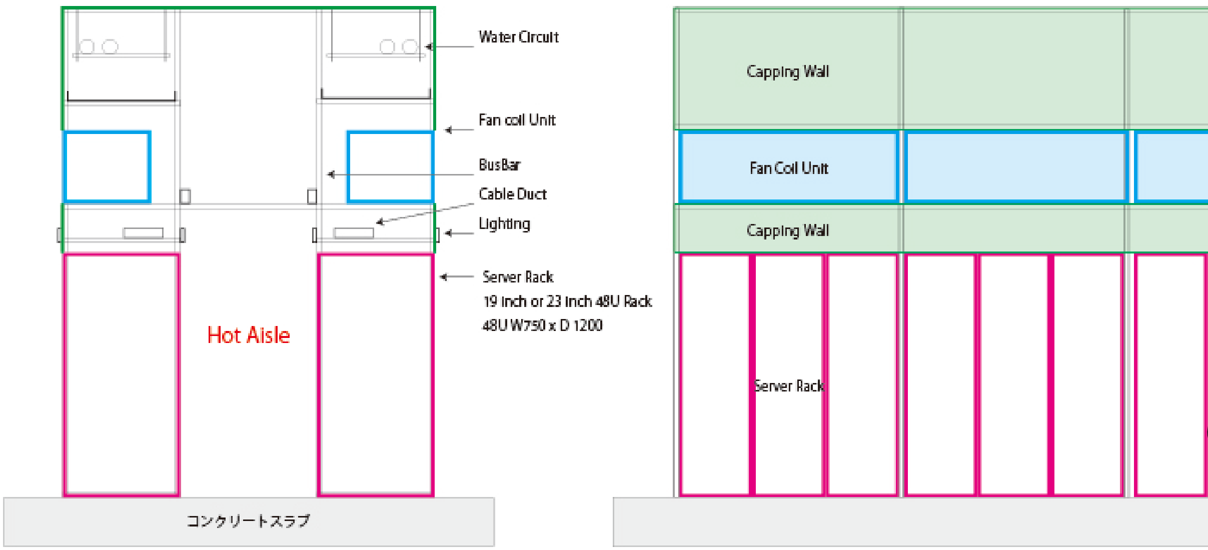


Layout Plan

- 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)

Data Center Image

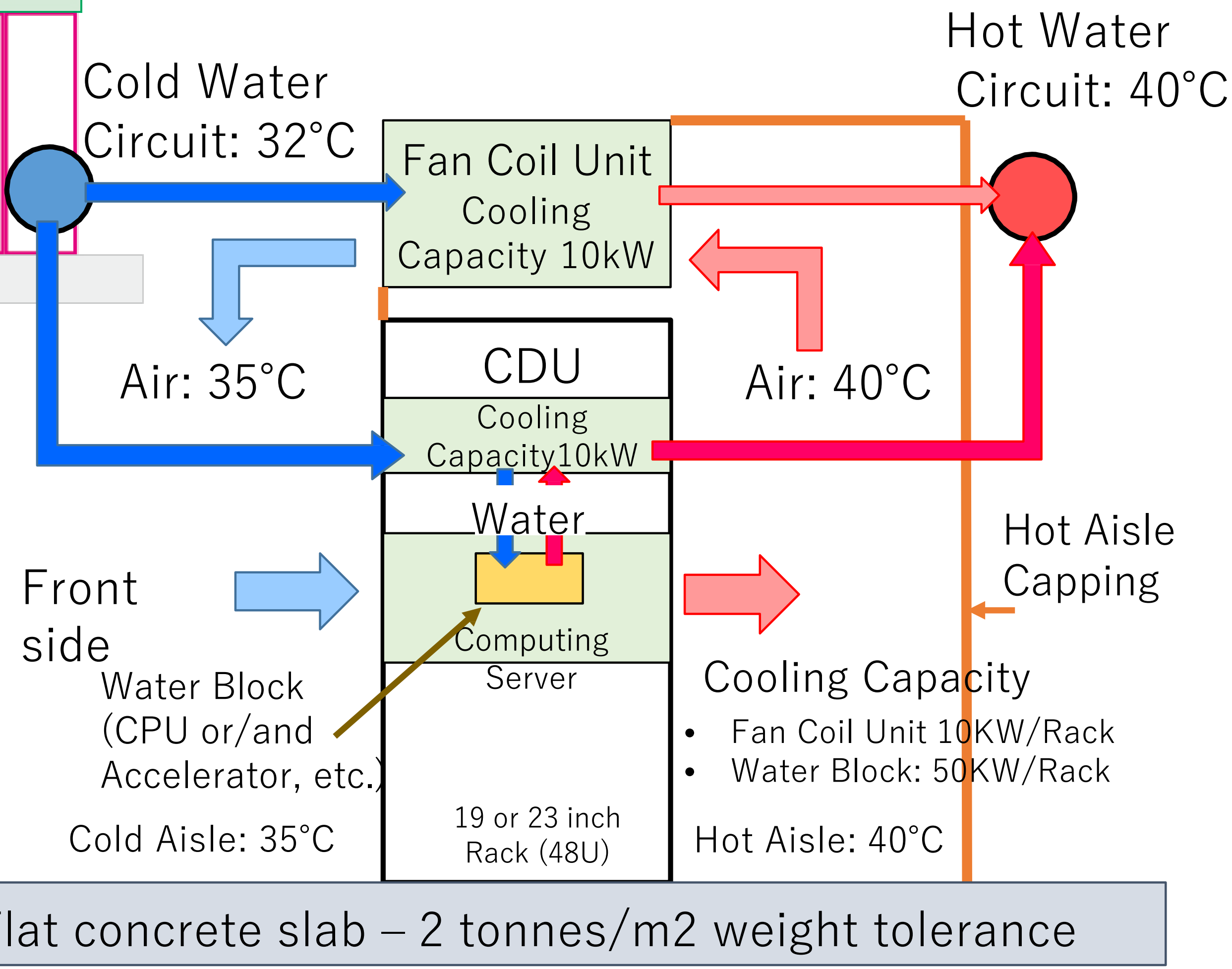
Implementing 60KW cooling in Cloud IDC – Cooling Pods



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

Cooling Block Diagram (Hot Rack)



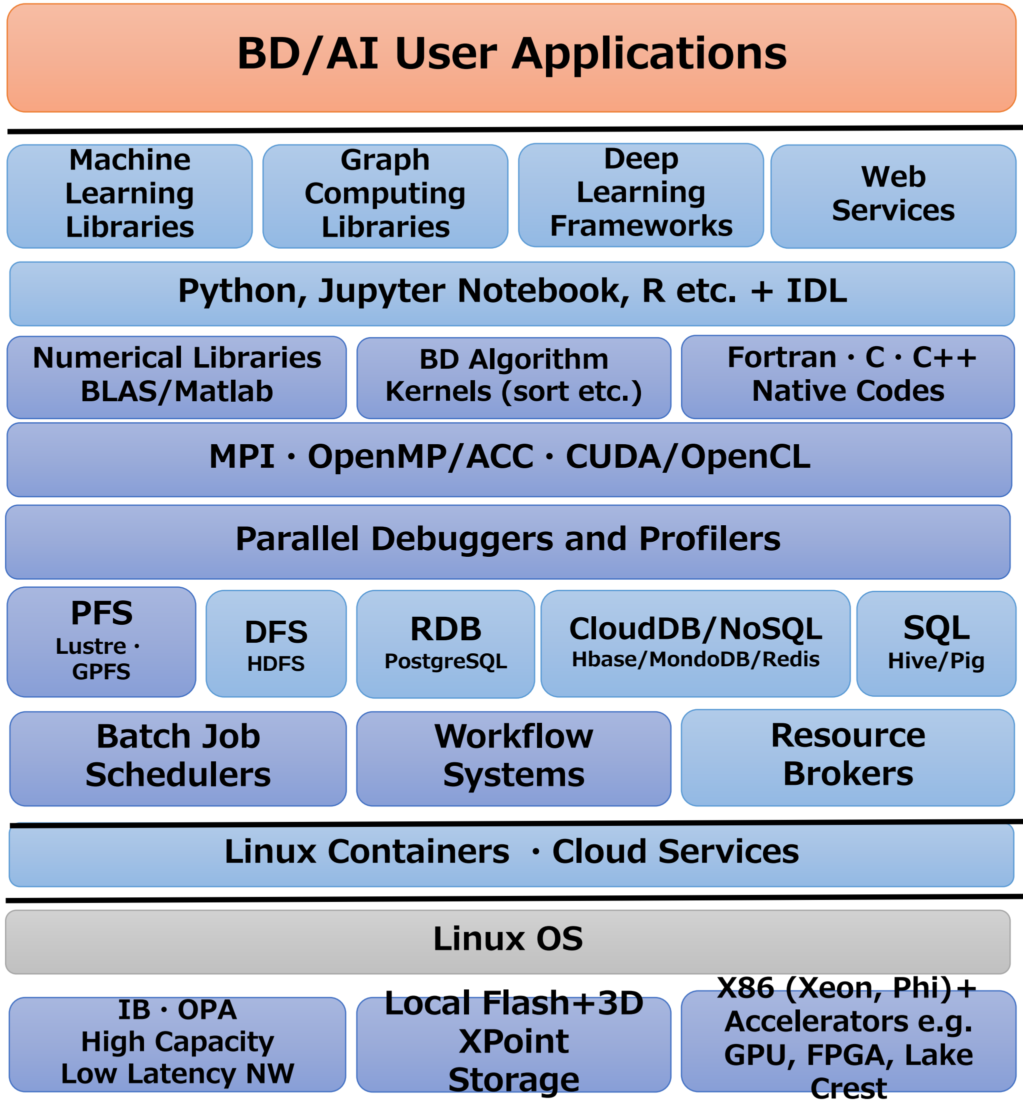
ABCI Procurement Benchmarks

- 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)

Basic Requirements for AI Cloud System



BD/AI User Applications

Machine Learning Libraries

Graph Computing Libraries

Deep Learning Frameworks

Web Services

Python, Jupyter Notebook, R etc. + IDL

Numerical Libraries BLAS/Matlab

BD Algorithm Kernels (sort etc.)

Fortran · C · C++ Native Codes

MPI · OpenMP/ACC · CUDA/OpenCL

Parallel Debuggers and Profilers

PFS Lustre · GPFS

DFS HDFS

RDB PostgreSQL

CloudDB/NoSQL Hbase/MondoDB/Redis

SQL Hive/Pig

Batch Job Schedulers

Workflow Systems

Resource Brokers

Linux Containers · Cloud Services

Linux OS

IB · OPA High Capacity Low Latency NW

Local Flash+3D XPoint Storage

X86 (Xeon, Phi)+ Accelerators e.g. GPU, FPGA, Lake Crest

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™)

K computer

Supercomputer K technologies

FY2018～

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

"Exascale" AI possible in 1H2019

Designed for Scalable Learning, technically superior to Google TPU2

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

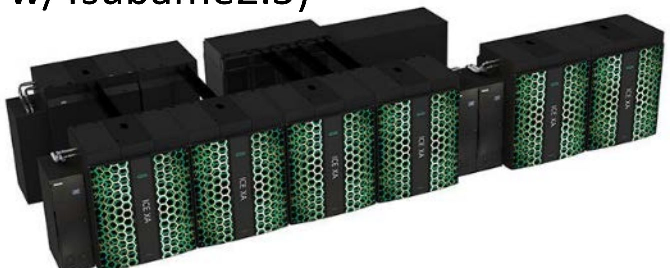
x5.8

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 AI-PF

x5.0~7.7

Draft RFC out
IDC under construction

1H 2019?
“**ExaAI**”
~1 AI-ExaFlop

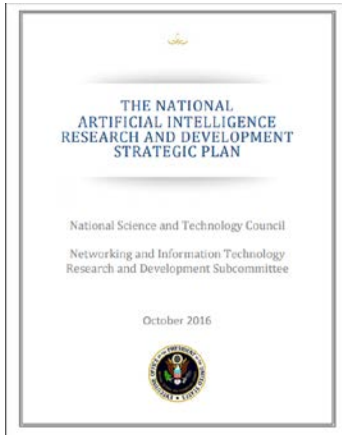
Undergoing Engineering Study

~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)

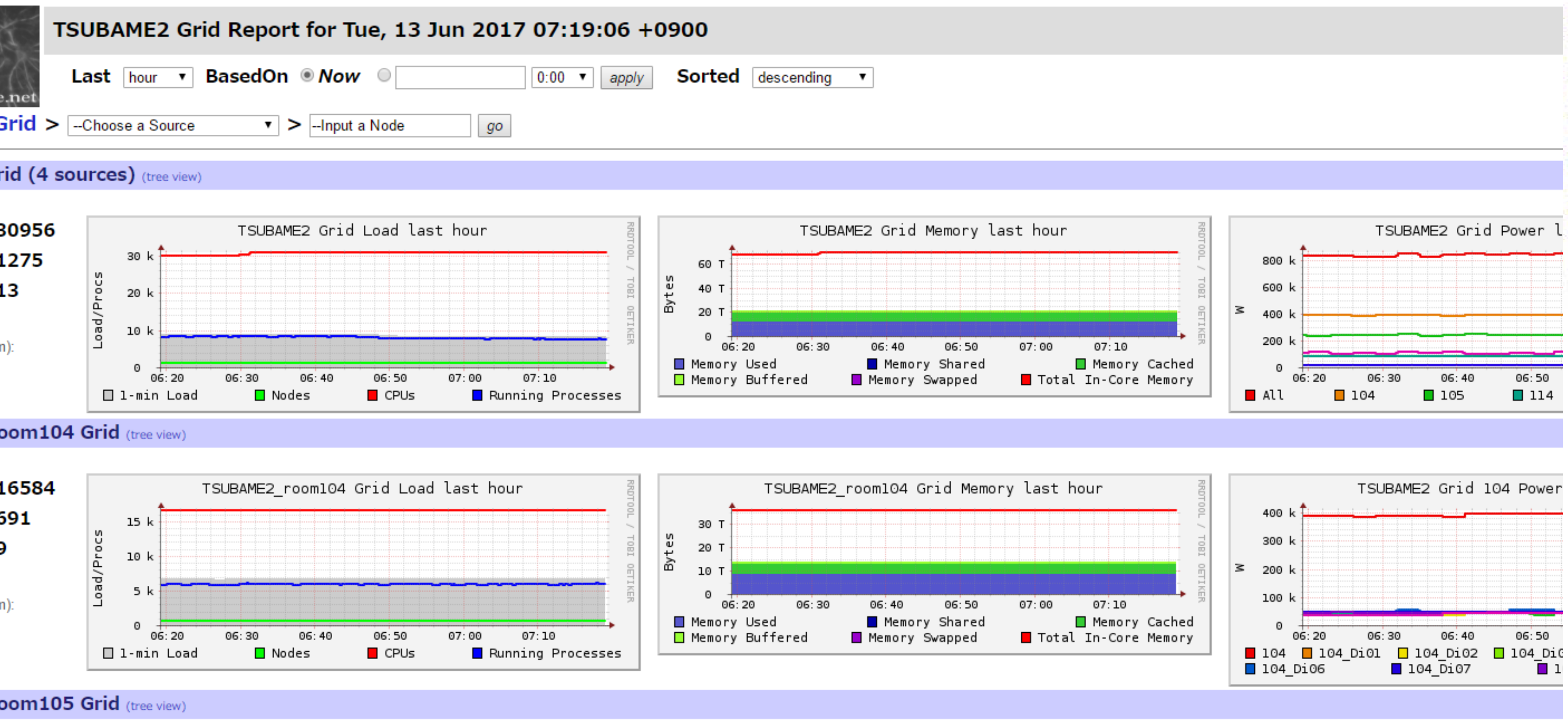


THE NATIONAL
ARTIFICIAL INTELLIGENCE
RESEARCH AND DEVELOPMENT
STRATEGIC PLAN

National Science and Technology Council

Networking and Information Technology
Research and Development Subcommittee

October 2016



TSUBAME2 Grid Report for Tue, 13 Jun 2017 07:19:06 +0900

Last hour ▼ BasedOn ● Now ○ 0:00 ▼ apply Sorted descending ▼

Grid > --Choose a Source ▼ > --Input a Node go

rid (4 sources) (tree view)

30956
1275
13

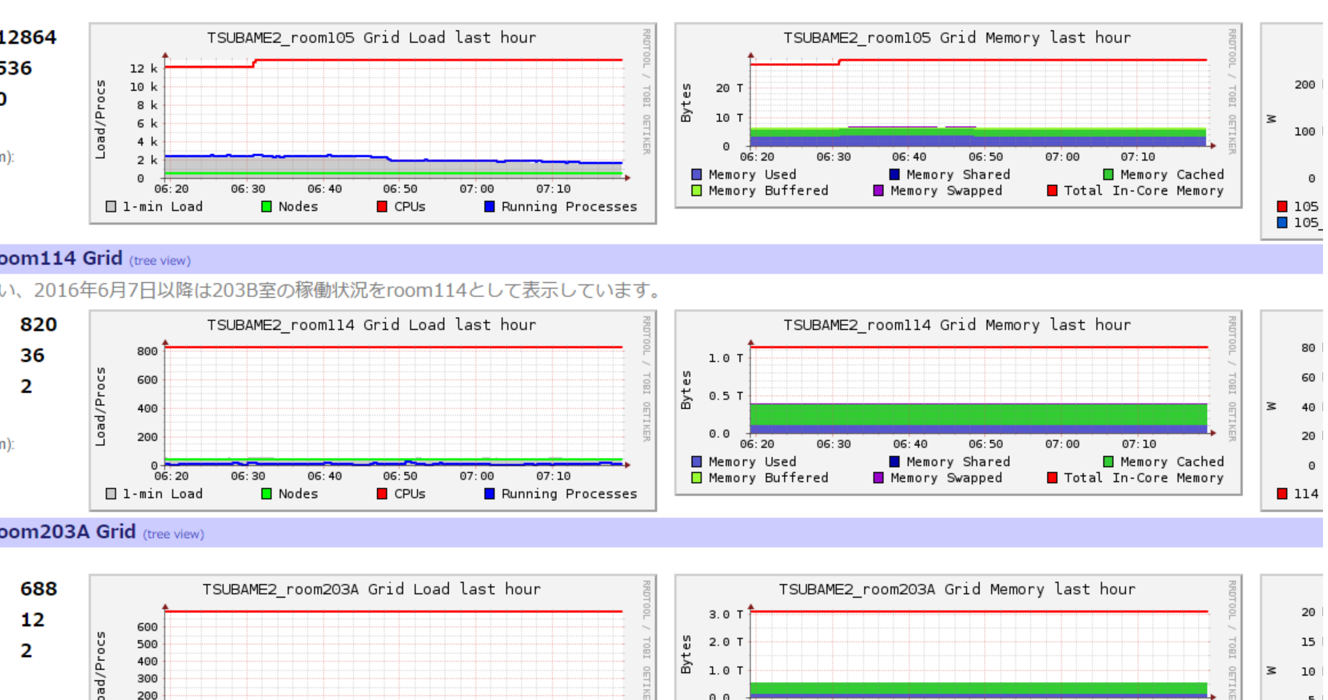
oom104 Grid (tree view)

16584
591
9

oom105 Grid (tree view)

12864
536
0

TSUBAME2 Grid Load last hour
TSUBAME2 Grid Memory last hour
TSUBAME2 Grid Power l
TSUBAME2_room104 Grid Load last hour
TSUBAME2_room104 Grid Memory last hour
TSUBAME2 Grid 104 Power
Load/Procs
Bytes
W
□ 1-min Load 🟩 Nodes 🟥 CPUs 🟦 Running Processes
🟦 Memory Used 🟦 Memory Shared 🟩 Memory Cached
🟩 Memory Buffered 🟪 Memory Swapped 🟥 Total In-Core Memory
🟥 All 🟧 104 🟩 105 🟩 114
🟥 104 🟧 104_Di01 🟨 104_Di02 🟩 104_Di
🟦 104_Di06 🟦 104_Di07
06:20 06:30 06:40 06:50 07:00 07:10
RRDTOOL / TOBI OETIKER



oom114 Grid (tree view)

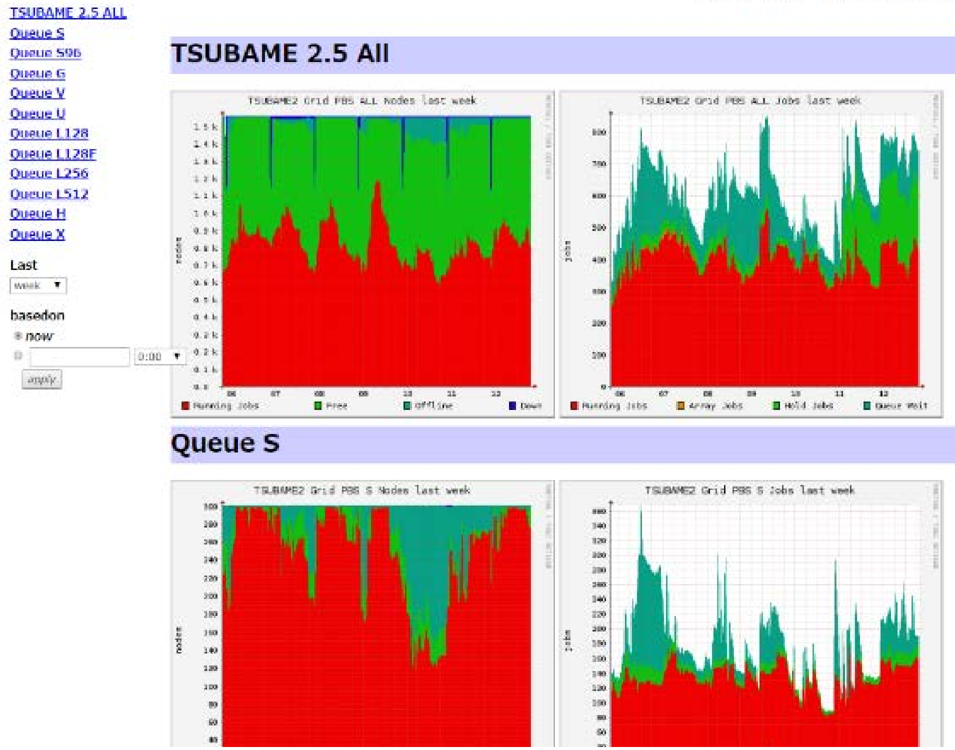
い、2016年6月7日以降は203B室の稼働状況をroom114として表示しています。

820
36
2

oom203A Grid (tree view)

688
12
2

TSUBAME2_room105 Grid Load last hour
TSUBAME2_room105 Grid Memory last hour
TSUBAME2_room114 Grid Load last hour
TSUBAME2_room114 Grid Memory last hour
TSUBAME2_room203A Grid Load last hour
TSUBAME2_room203A Grid Memory last hour
Load/Procs
Bytes
W
□ 1-min Load 🟩 Nodes 🟥 CPUs 🟦 Running Processes
🟦 Memory Used 🟦 Memory Shared 🟩 Memory Cached
🟩 Memory Buffered 🟪 Memory Swapped 🟥 Total In-Core Memory
🟥 105 🟦 105_
🟥 114
06:20 06:30 06:40 06:50 07:00 07:10



TSUBAME 2.5 ALL
Queue S
Queue S96
Queue G
Queue V
Queue U
Queue L128
Queue L128F
Queue L256
Queue L512
Queue H
Queue X

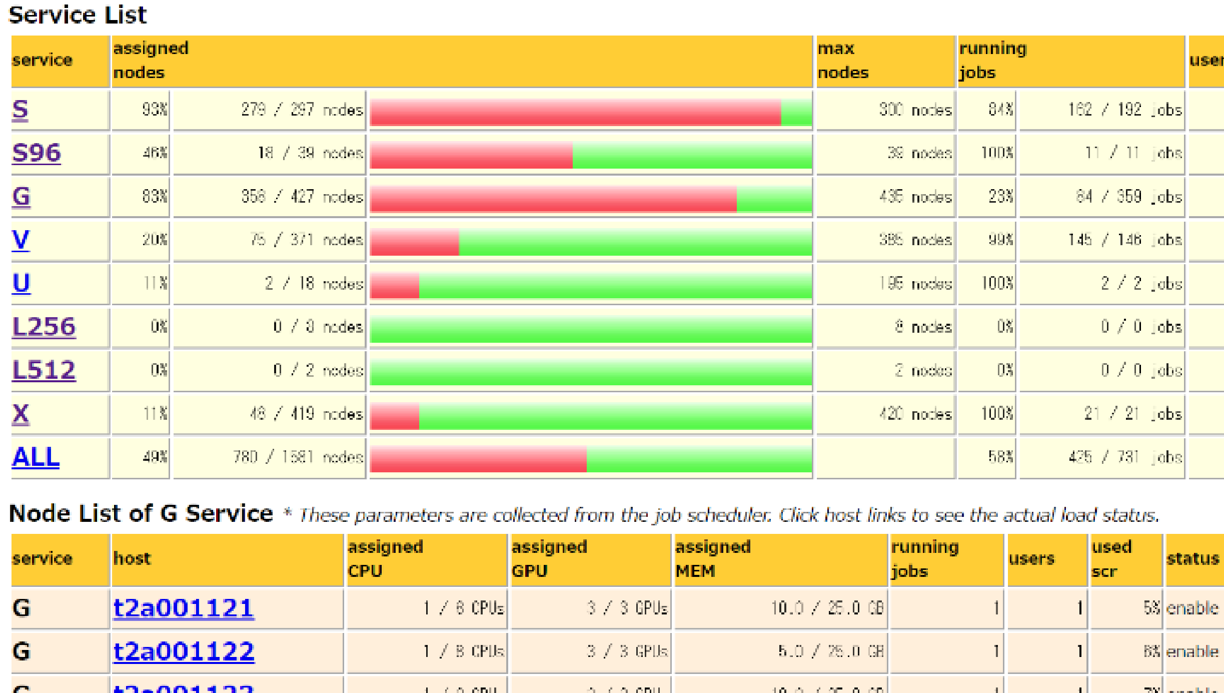
Last week ▼

basedon
● now
○ 0:00 ▼
apply

TSUBAME 2.5 All

Queue S

TSUBAME2 Grid PBS ALL Nodes last week
TSUBAME2 Grid PBS ALL Jobs last week
TSUBAME2 Grid PBS S Nodes last week
TSUBAME2 Grid PBS S Jobs last week
nodes
jobs
🟥 Running Jobs 🟩 Free 🟩 Offline 🟦 Down
🟥 Running Jobs 🟧 Array Jobs 🟩 Hold Jobs 🟩 Queue Wait



Service List

service	assigned nodes			max nodes	running jobs		user
S	93%	279 / 297 nodes		300 nodes	84%	162 / 192 jobs	
S96	46%	18 / 39 nodes		39 nodes	100%	11 / 11 jobs	
G	83%	356 / 427 nodes		435 nodes	23%	84 / 359 jobs	
V	20%	75 / 371 nodes		385 nodes	99%	145 / 146 jobs	
U	11%	2 / 18 nodes		195 nodes	100%	2 / 2 jobs	
L256	0%	0 / 8 nodes		8 nodes	0%	0 / 0 jobs	
L512	0%	0 / 2 nodes		2 nodes	0%	0 / 0 jobs	
X	11%	46 / 419 nodes		420 nodes	100%	21 / 21 jobs	
ALL	49%	780 / 1581 nodes			58%	425 / 731 jobs	

Node List of G Service

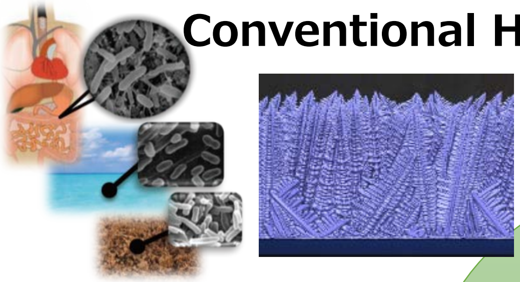
* *These parameters are collected from the job scheduler. Click host links to see the actual load status.*

service	host	assigned CPU	assigned GPU	assigned MEM	running jobs	users	used scr	status
G	t2a001121	1 / 8 CPUs	3 / 3 GPUs	10.0 / 25.0 GB	1	1	5%	enable
G	t2a001122	1 / 8 CPUs	3 / 3 GPUs	5.0 / 25.0 GB	1	1	6%	enable
G	t2a001123	1 / 8 CPUs	3 / 3 GPUs	10.0 / 25.0 GB	1	1	7%	enable

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

- for survival of HPC

Accelerating Conventional HPC Apps



Optimizing System Software and Ops

Future Big Data･AI Supercomputer Design

Big Data AI-Oriented Supercomputers

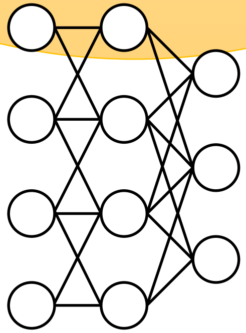
Acceleration and Scaling of BD/ML/AI via HPC and Technologies and Infrastructures

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

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

Big Data and ML/AI Apps and Methodologies

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



Large Scale Graphs

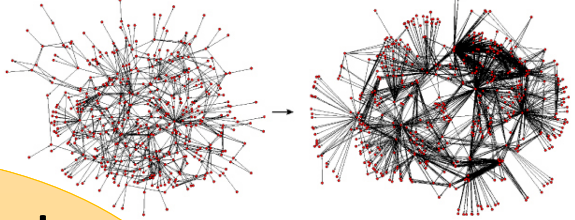
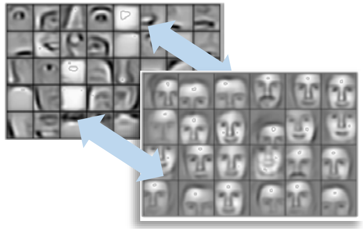
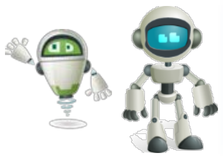


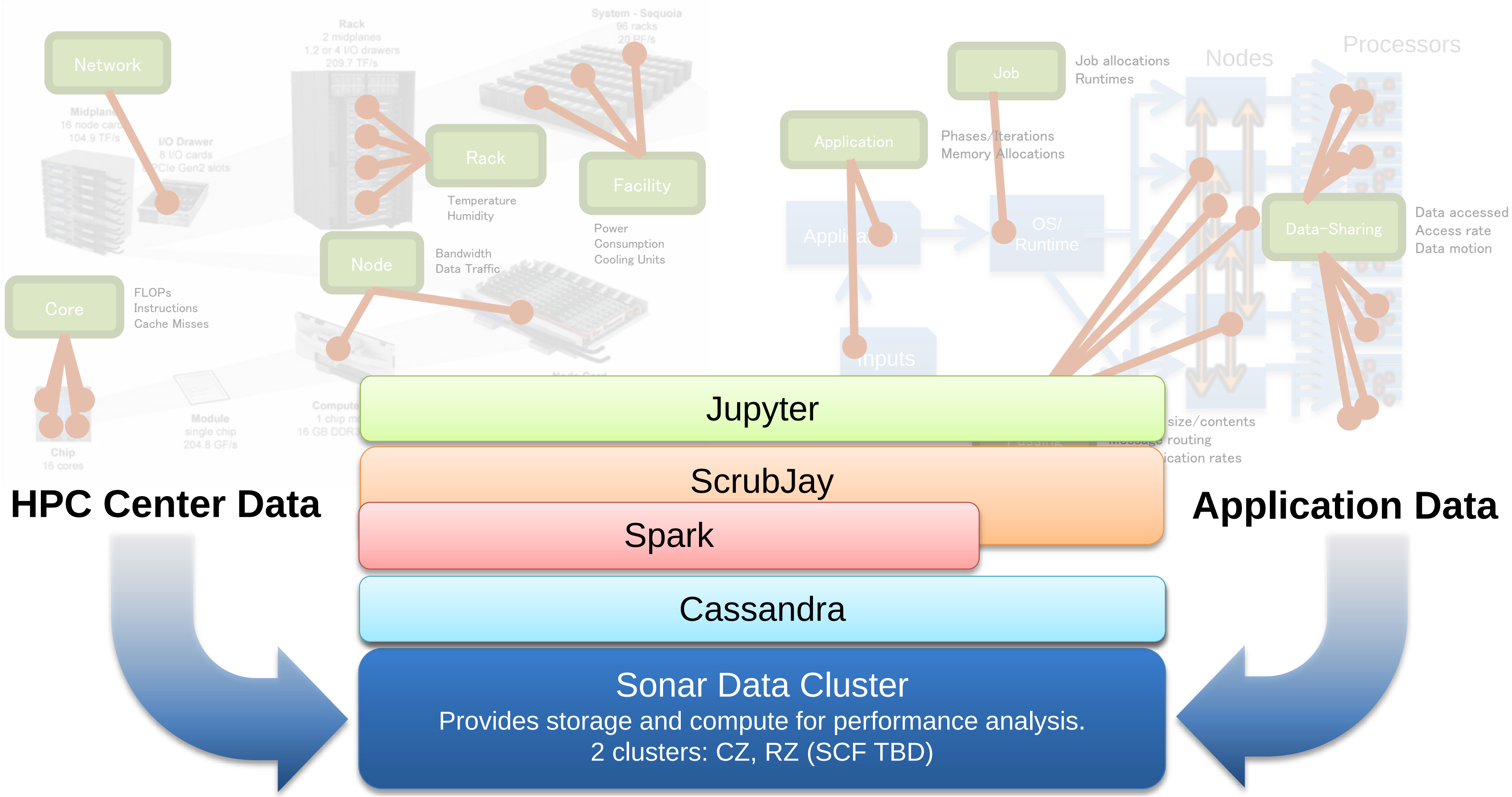
Image and Video



Robots / Drones



Sonar collects data from the HPC Center and applications, allowing users to access it with secure permissions
(Slide courtesy Todd Gamblin @ LLNL)



Jupyter

ScrubJay

Spark

Cassandra

Sonar Data Cluster

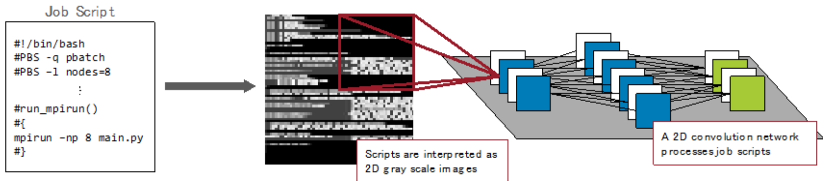
Provides storage and compute for performance analysis.
2 clusters: CZ, RZ (SCF TBD)

Sonar enables all LC users to research into the root causes of performance variation

Sonar | Prediction | Apollo | ScrubJay

We combine neural networks with queue simulation to predict resource utilization
(Slide courtesy Todd Gamblin @ LLNL)

Convolutional Neural Networks



```
#!/bin/bash
#PBS -q pbatch
#PBS -l nodes=8
        :
#run_mpirun()
#{
mpirun -np 8 main.py
#}
```

Scripts are interpreted as 2D gray scale images

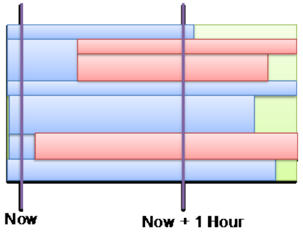
A 2D convolution network processesjob scripts

Job Runtime

User Estimate

CNN prediction

Queue Simulation



I/O Bandwidth

- CNN prediction:
  - Takes *only* job scripts and queue data as input
  - Leverage unstructured information (coding style, job scripts)
  - No preprocessing of job scripts required; fully automated
- Queue simulation:
  - Use predicted runtimes to simulate future job schedule
- We can predict many I/O bursts.
  - Some bursts can't be predicted b/c jobs enter the queue and run immediately
  - There may be periodicity or other patterns to these bursts
- We are investigating additional modeling techniques to predict bursts that cannot be simulated

Observed

Predicted
(30 min horizon)

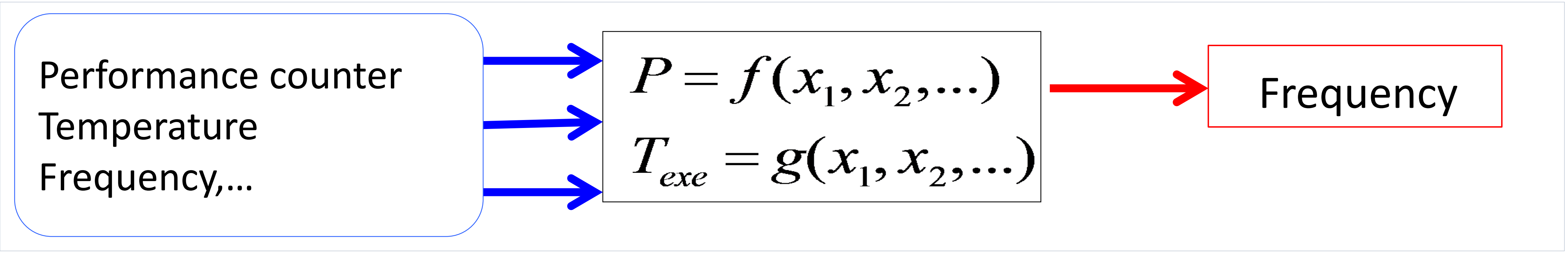
These results will provide input for resource-aware scheduling on the Flux project

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

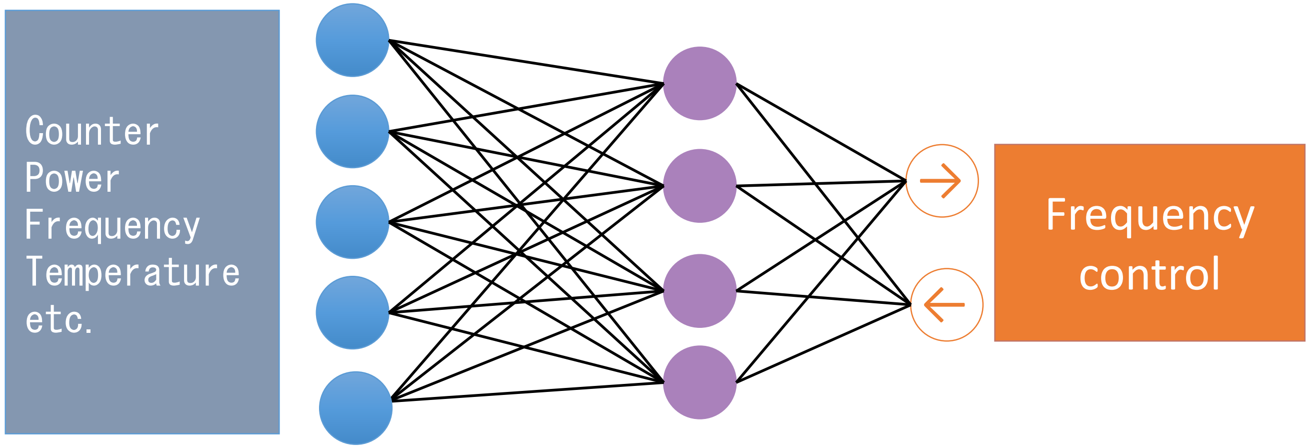
Use Deep Learning for analysis.

・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.

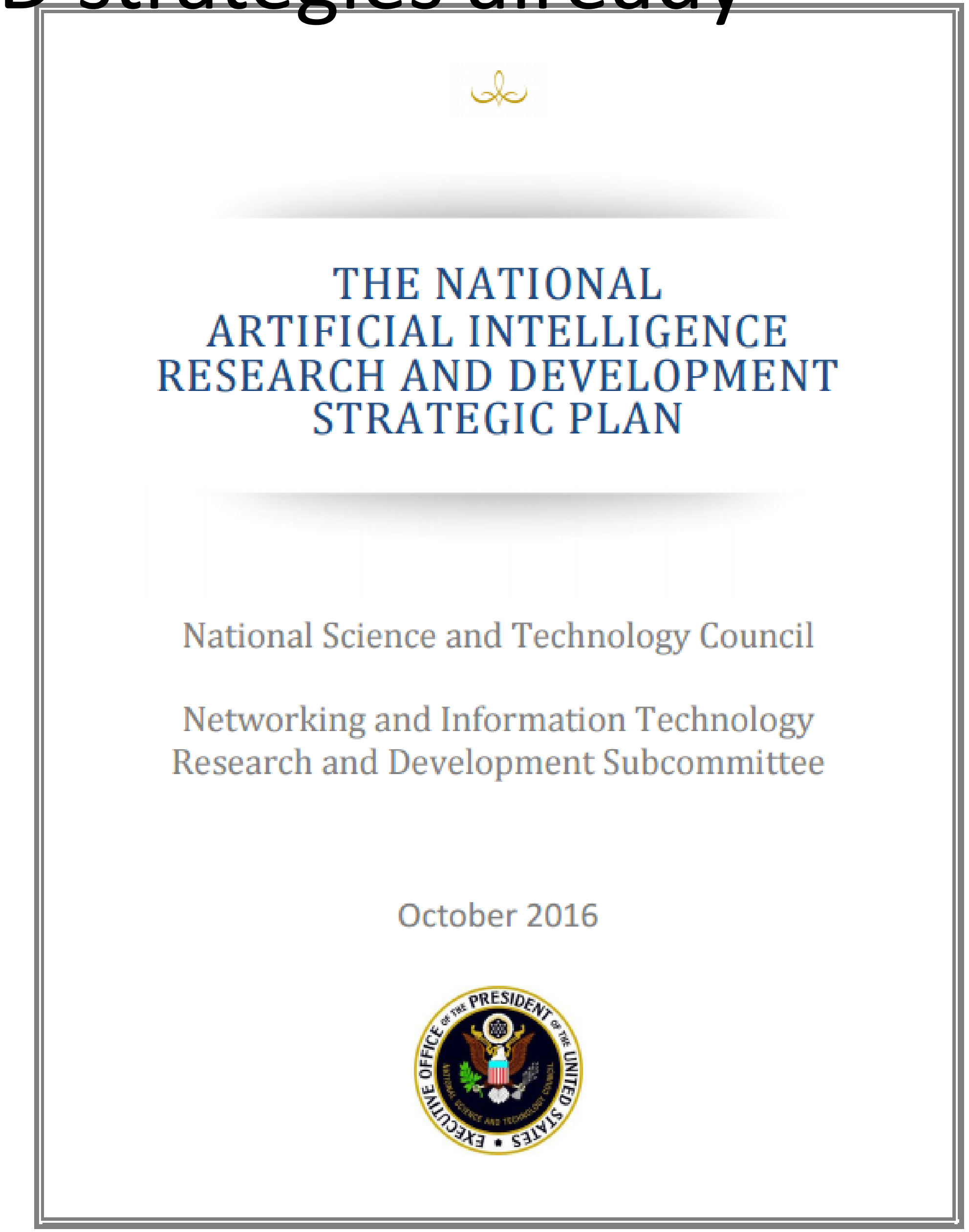


Counter
Power
Frequency
Temperature
etc.

Frequency control

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

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



THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN

National Science and Technology Council

Networking and Information Technology Research and Development Subcommittee

October 2016

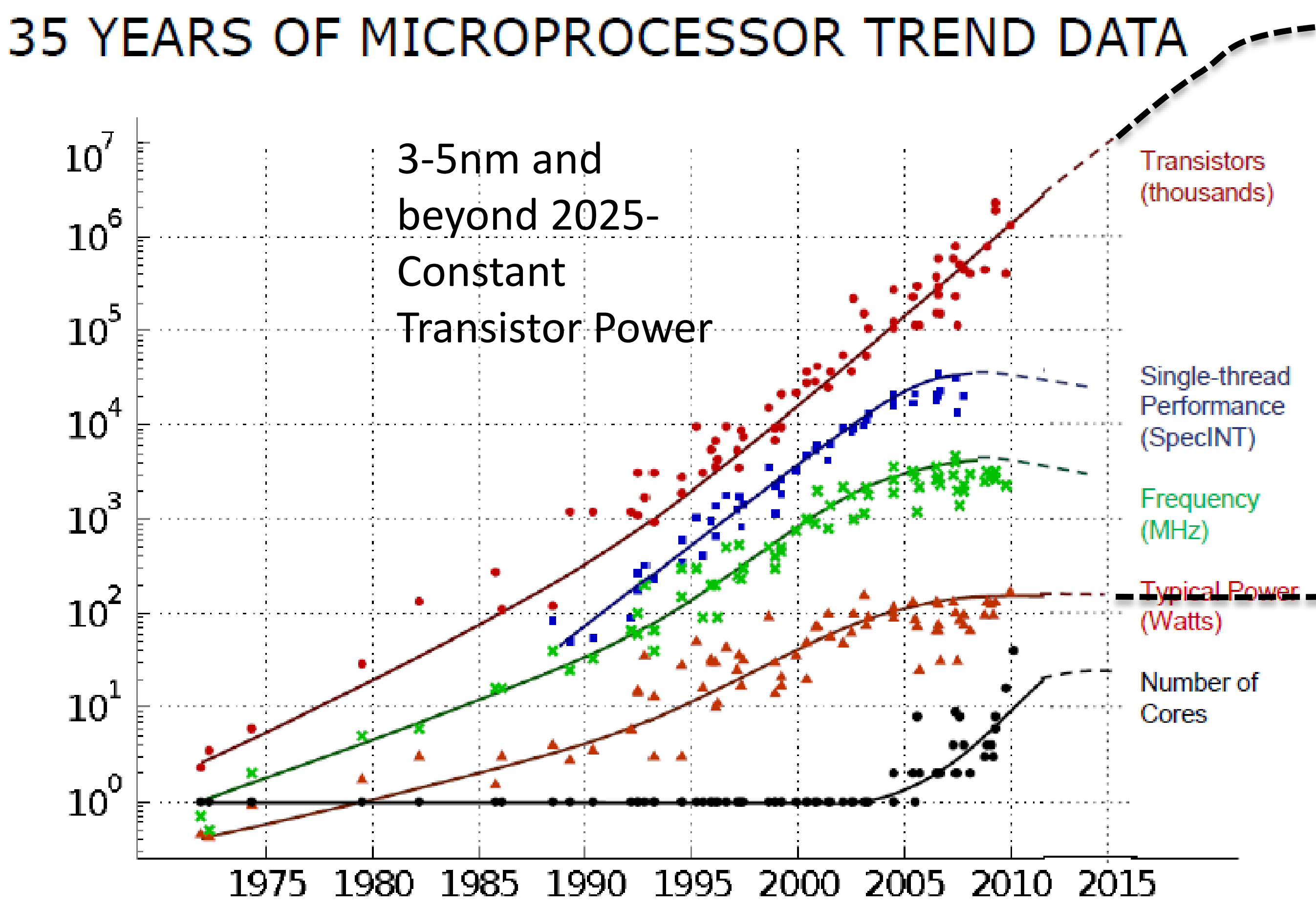
What is worse: Moore’s Law will end in the 2020’s

- 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
- 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***

The curse of constant transistor power shall soon be upon us

Gordon Moore

20 year Eras towards of End of Moore’s Law



3-5nm and beyond 2025- Constant Transistor Power

Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

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

- 1980s~2004 Dennard scaling, perf+ = single thread+ = transistor & freq+ = power+

20 year Post-Dennard Many-Core Era

- 2004~2015 feature scaling, perf+ = transistor+ = core#+, constant power
- 2015~2025 all above gets harder

20-year Next-Gen Post-Moore era

- 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*

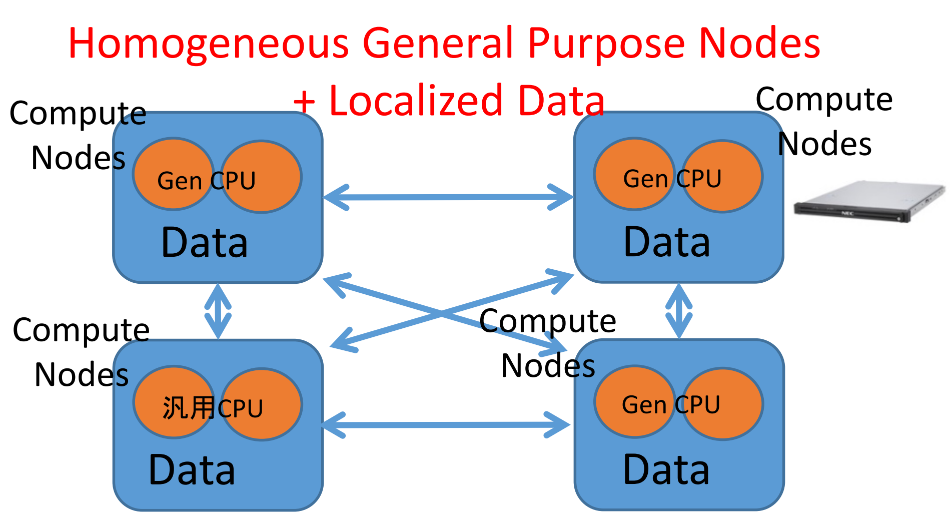
Many Core Era

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

Compute Nodes

Gen CPU

Data

Compute Nodes

Gen CPU

Data

Compute Nodes

汎用CPU

Data

Compute Nodes

Gen CPU

Data

Loosely Coupled with Electronic Interconnect

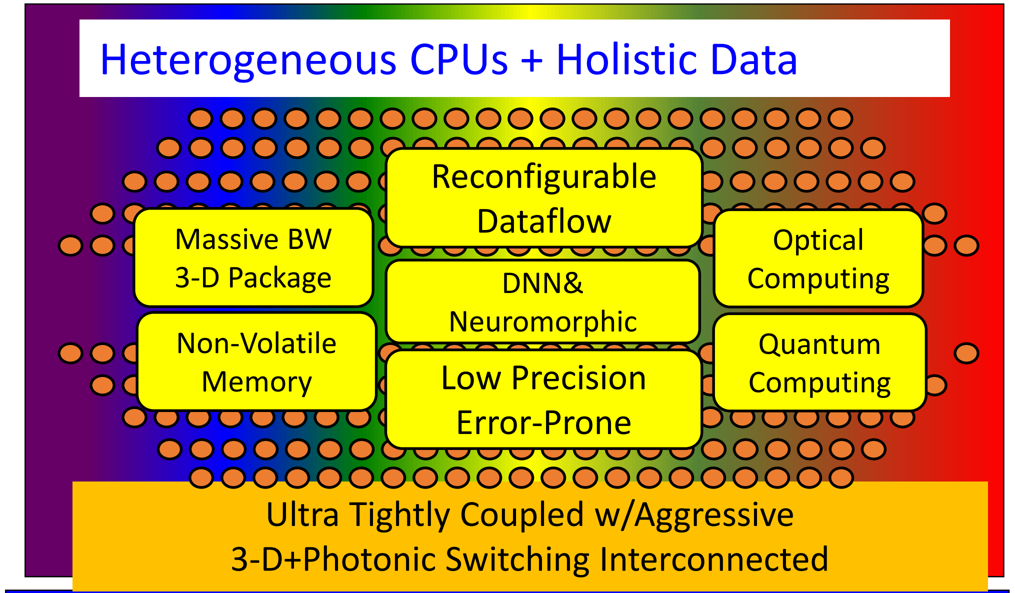
Transistor Lithography Scaling
(CMOS Logic Circuits, DRAM/SRAM)

~2025
M-P Extinction
Event

Bytes-Centric Algorithms and Apps

Bytes-Centric System Software

Hardware/Software System APIs
Data-Centric Heterogeneous Architecture



Heterogeneous CPUs + Holistic Data

Reconfigurable
Dataflow

Massive BW
3-D Package

Optical
Computing

DNN&
Neuromorphic

Non-Volatile
Memory

Quantum
Computing

Low Precision
Error-Prone

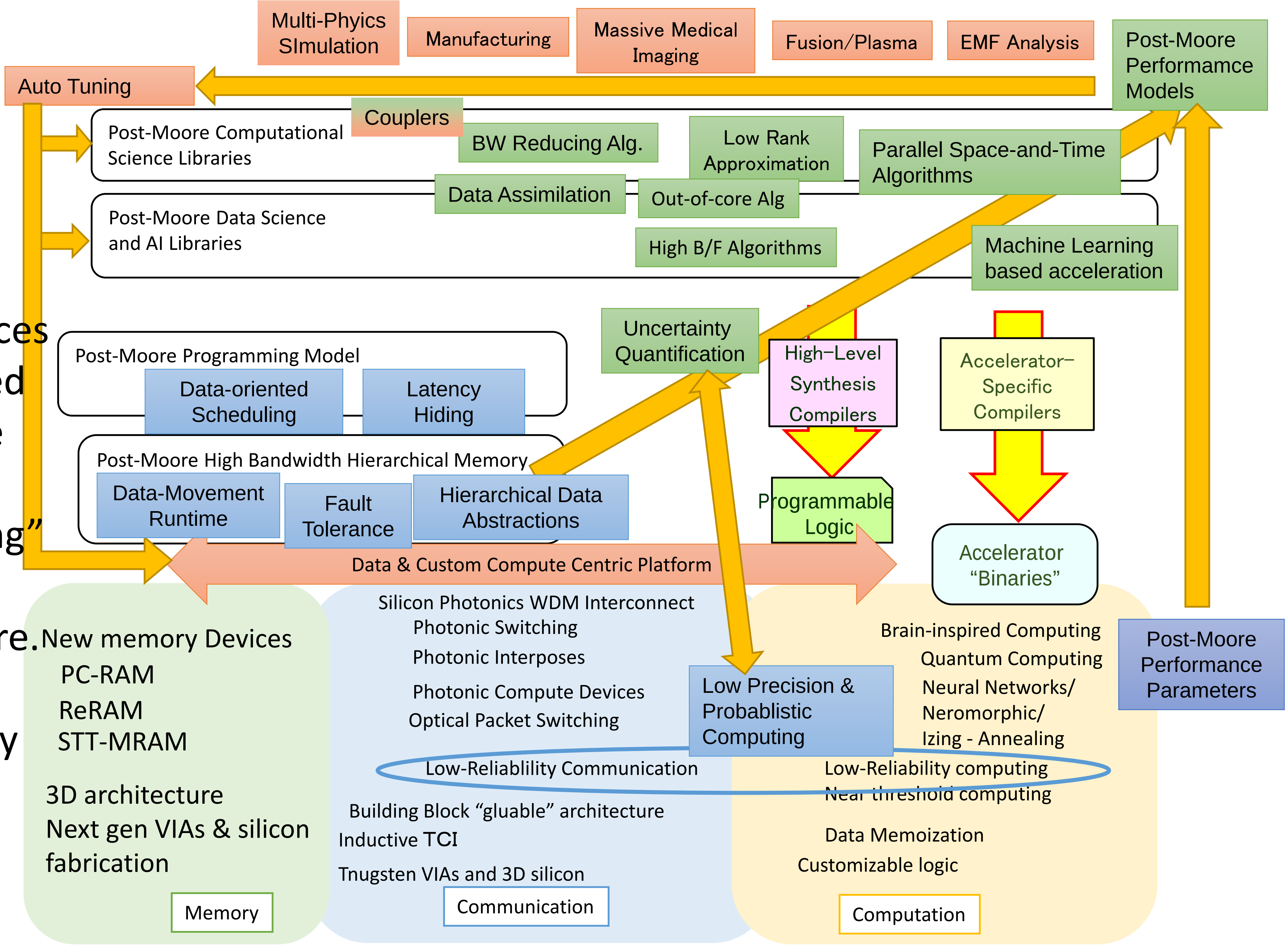
Ultra Tightly Coupled w/Aggressive
3-D+Photonic Switching Interconnected

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



Multi-Phyics SImulation

Manufacturing

Massive Medical Imaging

Fusion/Plasma

EMF Analysis

Post-Moore Performamce Models

Auto Tuning

Post-Moore Computational Science Libraries

Couplers

BW Reducing Alg.

Low Rank Approximation

Parallel Space-and-Time Algorithms

Data Assimilation

Out-of-core Alg

Post-Moore Data Science and AI Libraries

High B/F Algorithms

Machine Learning based acceleration

Uncertainty Quantification

High−Level Synthesis Compilers

Accelerator−Specific Compilers

Post-Moore Programming Model

Data-oriented Scheduling

Latency Hiding

Post-Moore High Bandwidth Hierarchical Memory

Data-Movement Runtime

Fault Tolerance

Hierarchical Data Abstractions

Programmable Logic

Accelerator “Binaries”

Data & Custom Compute Centric Platform

New memory Devices
PC-RAM
ReRAM
STT-MRAM

3D architecture
Next gen VIAs & silicon fabrication

Memory

Silicon Photonics WDM Interconnect
Photonic Switching
Photonic Interposes
Photonic Compute Devices
Optical Packet Switching

Low-Reliablility Communication

Building Block “gluable” architecture
Inductive TCI
Tnugsten VIAs and 3D silicon

Communication

Low Precision & Probablistic Computing

Brain-inspired Computing
Quantum Computing
Neural Networks/
Neromorphic/
Izing - Annealing

Low-Reliability computing
Near threshold computing

Data Memoization
Customizable logic

Computation

Post-Moore Performance Parameters