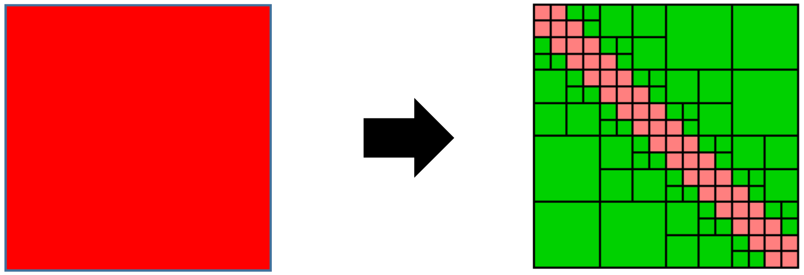
Hierarchical matrix(H-matrix) for CNN acceleration

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



dense matrix　　　　Hierarchical matrix

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

- Back ground

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

  - Significant savings in memory when compressed

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

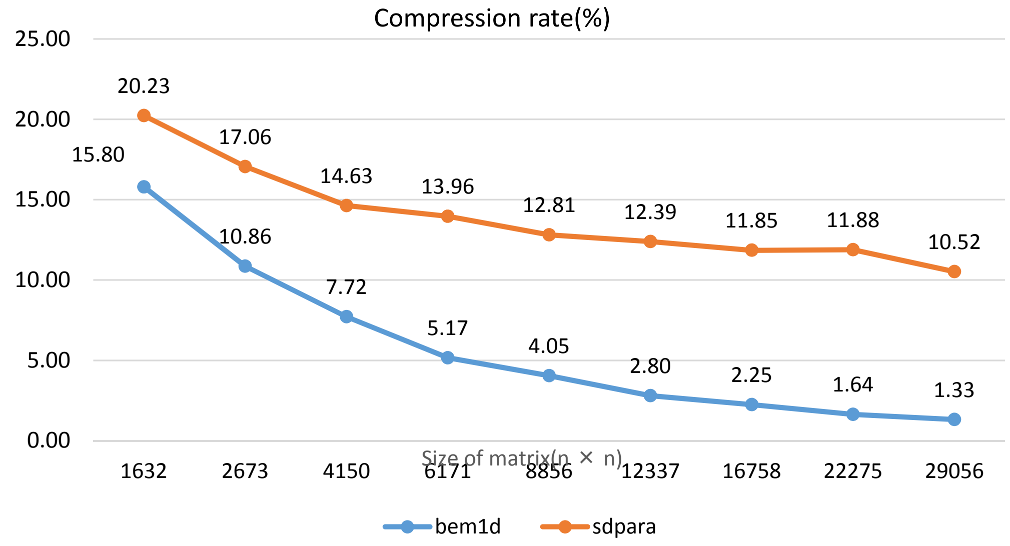
  - Computational complexity

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

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

Preliminary Results – Compression rate of matrices

SDPARA

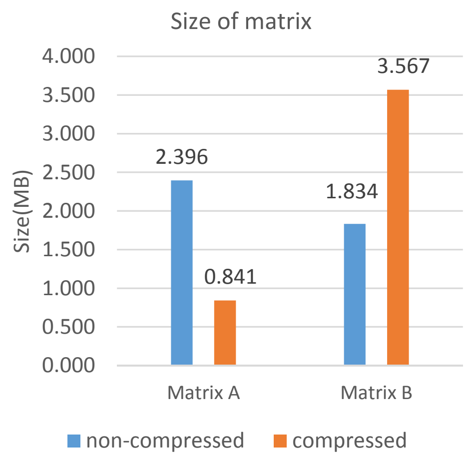


Compressive rate = (uncompressed size) / (compressed size)

We can compress the matrix in some applications.

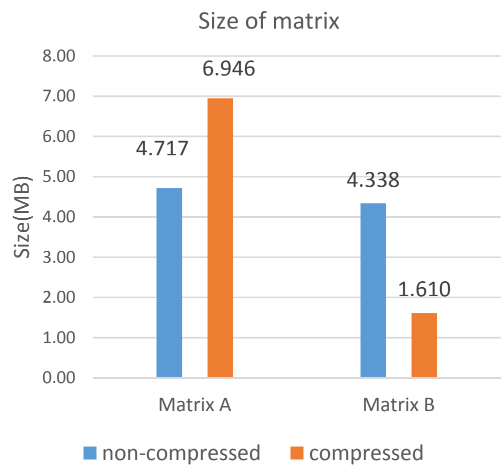
- bem1d: 1-Dimention Boundary element method
- sdpara: A parallel implementation of the inter-point method for Semi-Define Programming(SDP)

Deep Learning (CNN)



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

→ Matrix A successfully compressed!



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

→ Matrix B successfully compressed!

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

Power optimization using Deep Q-Network

Kento Teranishi

・Background

Power optimization by frequency control in existing research

Performance counter
Temperature
Frequency,…

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

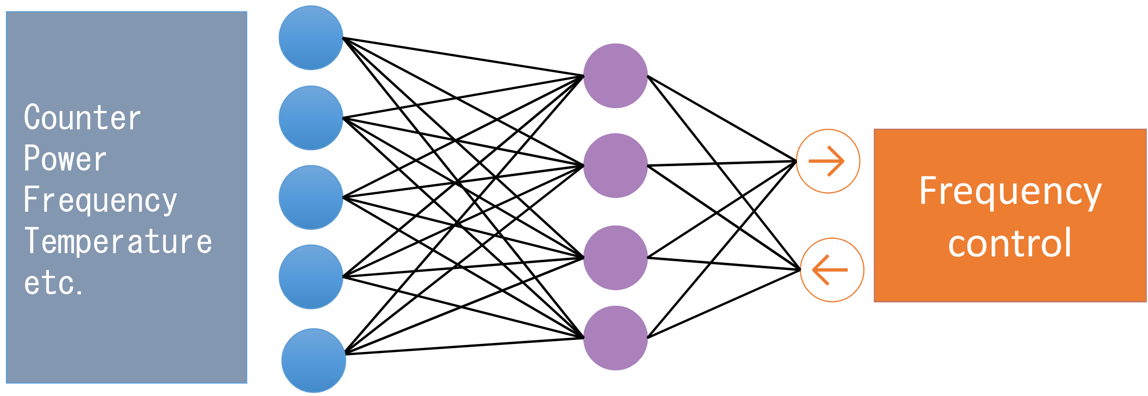
Frequency

- Detailed analysis is necessary
- Low versatility

Use Deep Learning for analysis.

・Objective

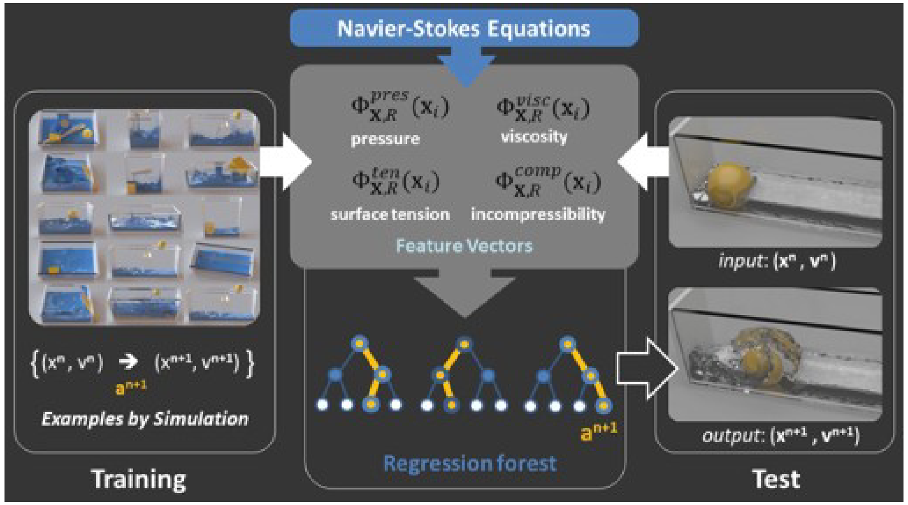
Implement the computer control system using Deep Q-Network.

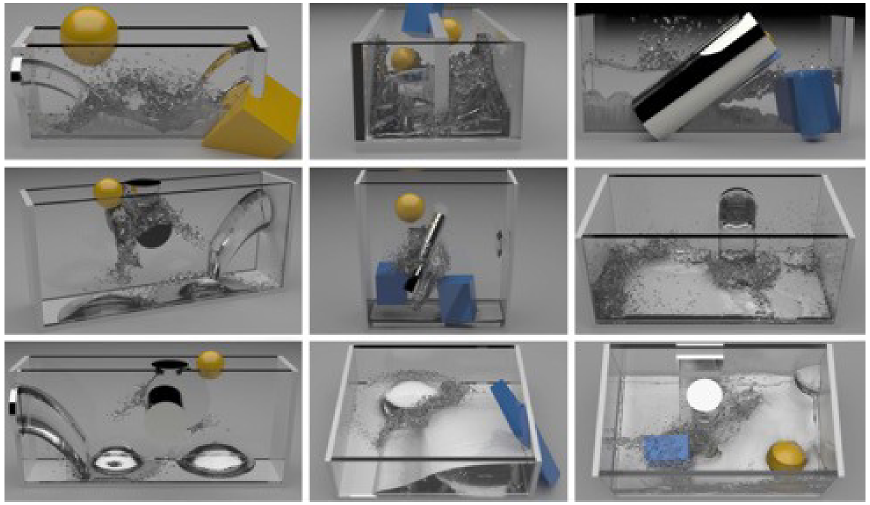


Deep Q-Network (DQN)

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

Using ML to Approximate Sciences - Fluid Dynamics Example (slide courtesy of Bill Dally @ NVIDIA)





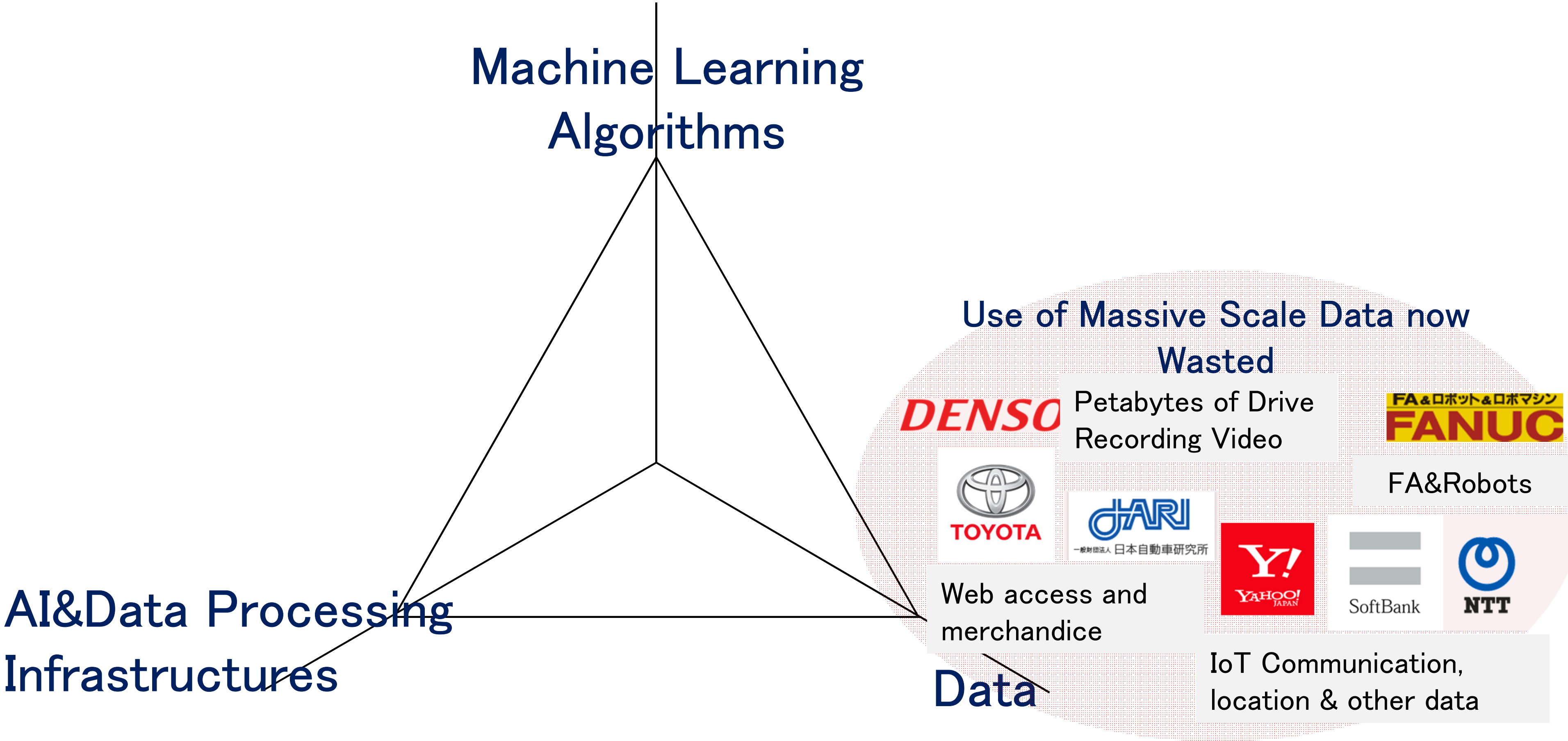
"… Implementation led to a speed-up of one to three orders of magnitude compared to the state-of-the-art position-based fluid solver and runs in real-time for systems with up to 2 million particles"

"Data-driven Fluid Simulations using Regression Forests" http://people.inf.ethz.ch/ladickyl/fluid_sigasia15.pdf

9

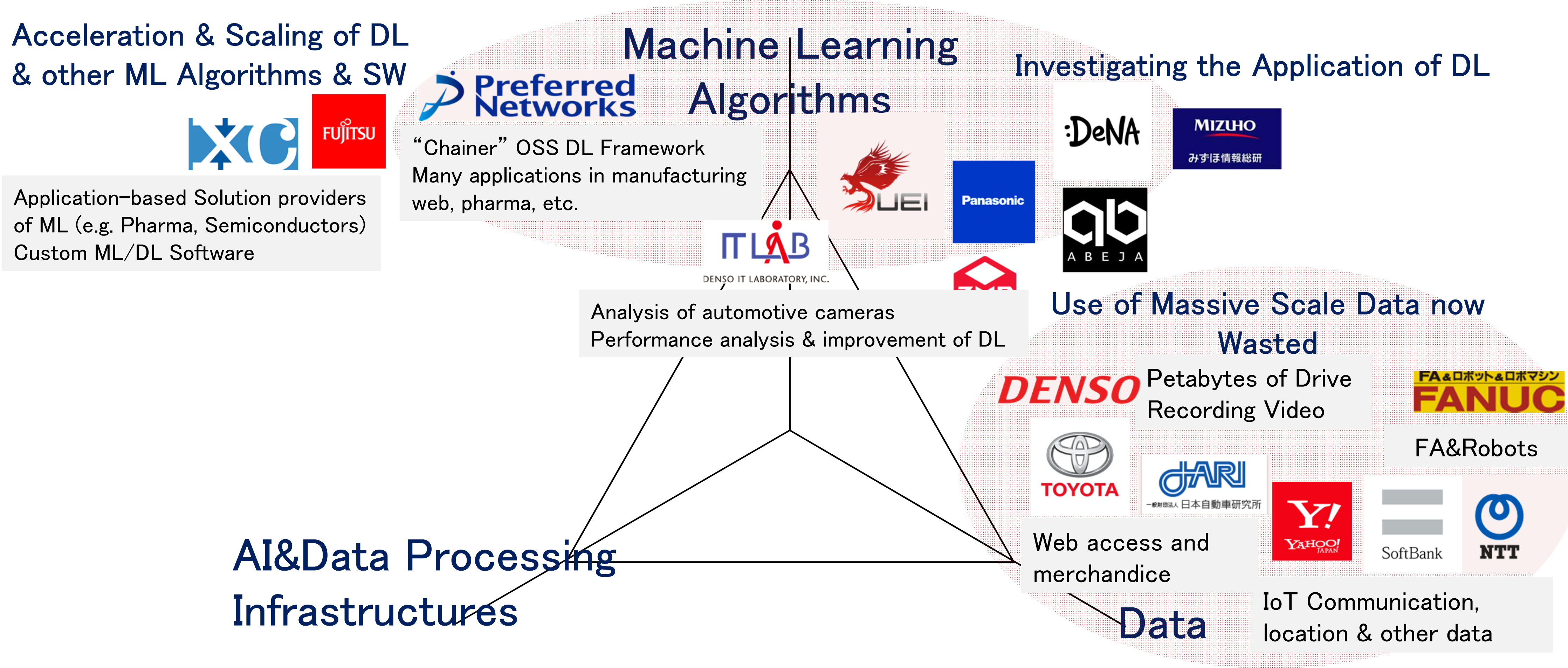
The current status of AI & Big Data in Japan

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



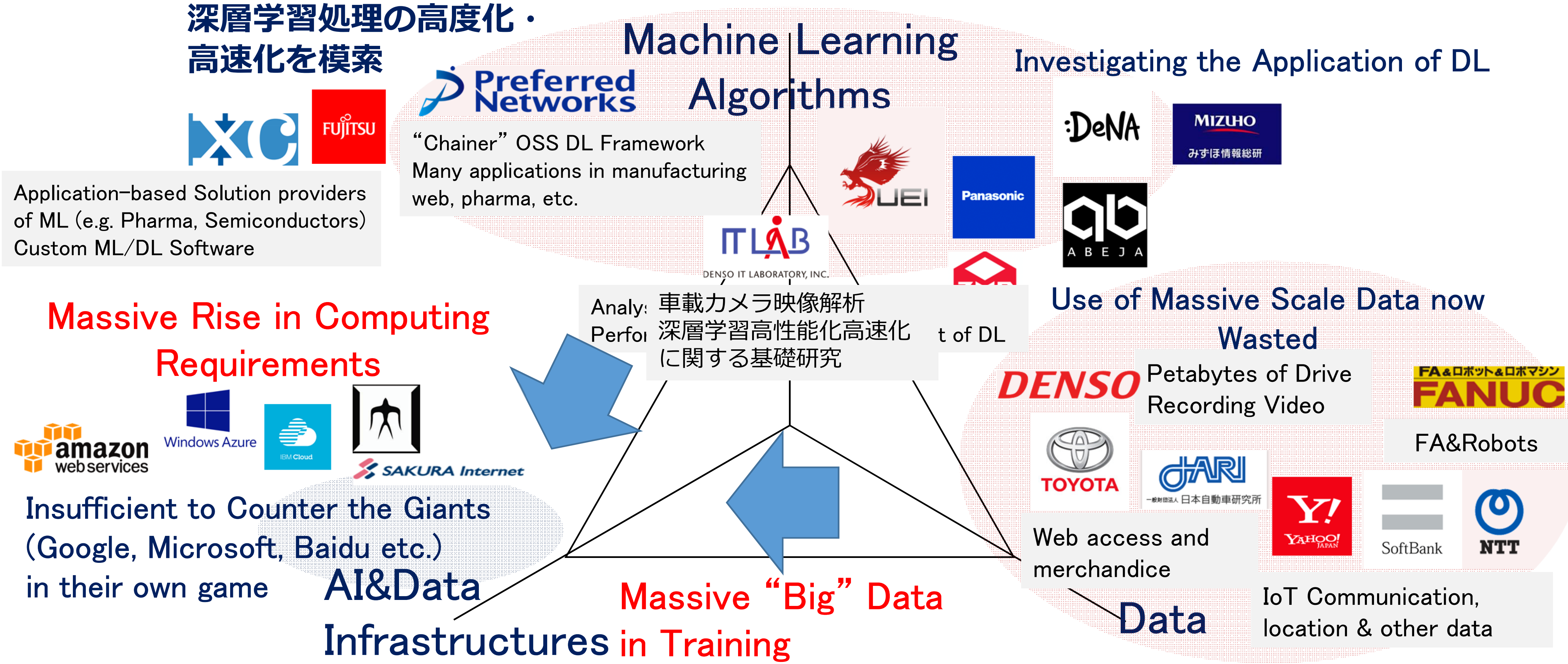
The current status of AI & Big Data in Japan

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

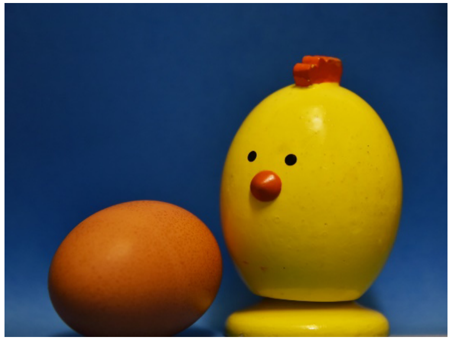


The current status of AI & Big Data in Japan

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



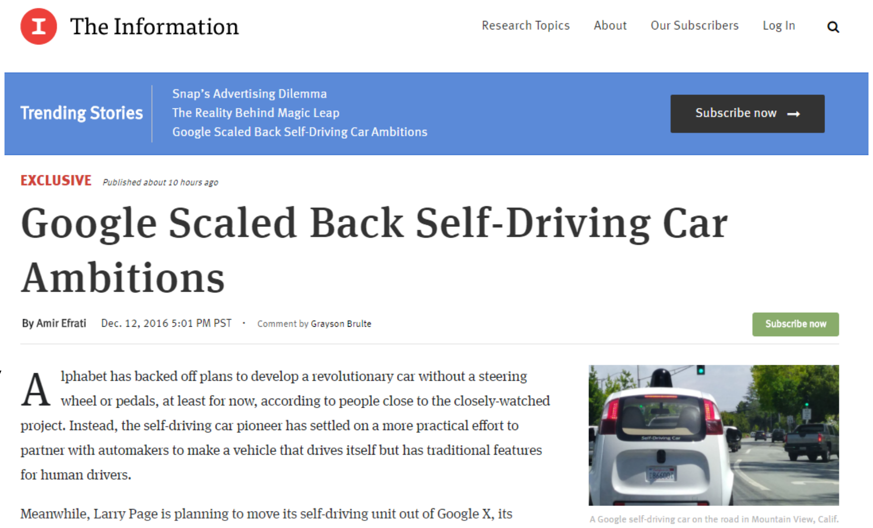
The "Chicken or Egg Problem" of AI-HPC Infrastructures



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

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

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


The Information
Research Topics About Our Subscribers Log In

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

EXCLUSIVE *Published about 10 hours ago*

Google Scaled Back Self-Driving Car Ambitions

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

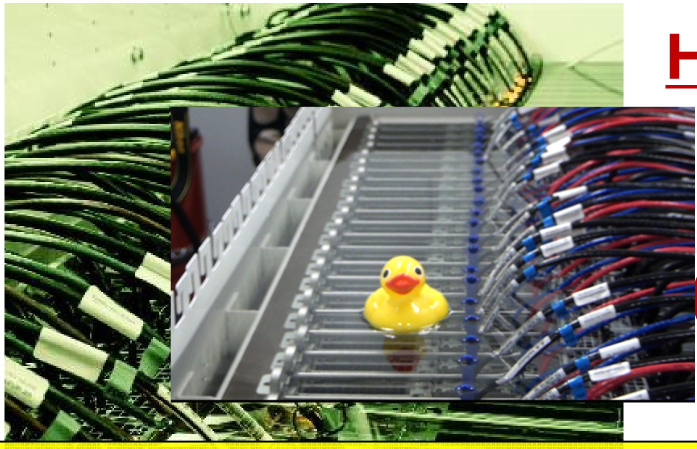
Alphabet has backed off plans to develop a revolutionary car without a steering wheel or pedals, at least for now, according to people close to the closely-watched project. Instead, the self-driving car pioneer has settled on a more practical effort to partner with automakers to make a vehicle that drives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

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

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

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

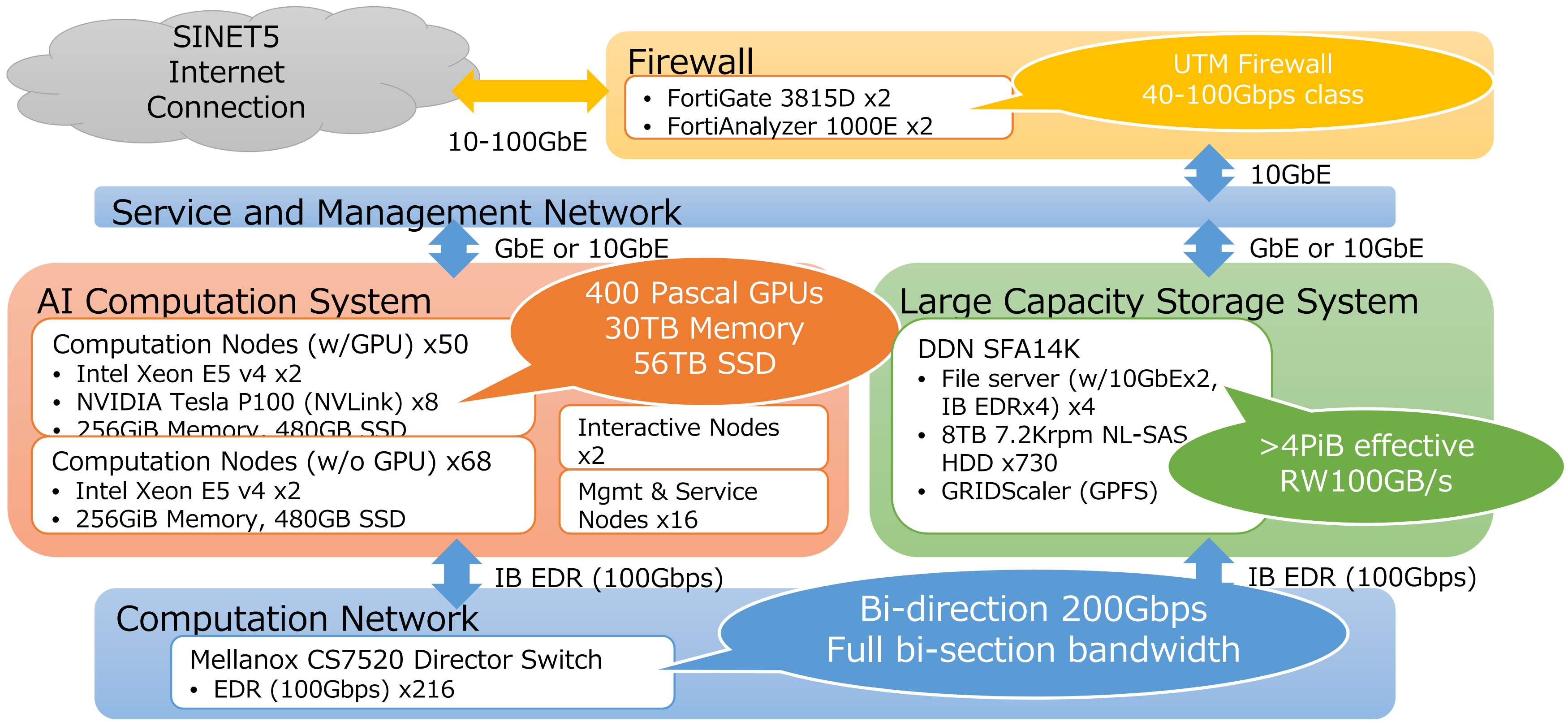
Single Rack High Density Oil Immersion
168 NVIDIA K80 GPUs + Xeon
413+TFlops (DFP)
1.5PFlops (SFP)
~60KW/rack

2013年11月/2014年6
Word #1 Green500

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

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

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



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

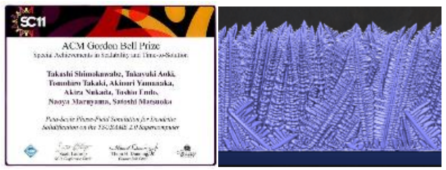
1. **“Everybody’s Supercomputer” - High Performance (12~24 DP Petaflops, 125~325TB/s Mem, 55~185Tbit/s NW), innovative high cost/performance packaging & design, in mere 180m²…**
2. **“Extreme Green” – ~10GFlops/W power-efficient architecture, system-wide power control, advanced cooling, future energy reservoir load leveling & energy recovery**
3. **“Big Data Convergence” – Extreme high BW &capacity, deep memory hierarchy, extreme I/O acceleration, Big Data SW Stack for machine learning, graph processing, …**
4. **“Cloud SC” – dynamic deployment, container-based node co-location & dynamic configuration, resource elasticity, assimilation of public clouds…**
5. **“Transparency” - full monitoring & user visibility of machine & job state, accountability via reproducibility**

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

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

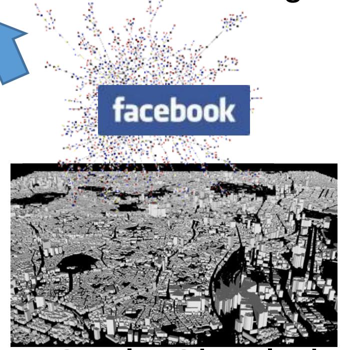
2010 TSUBAME2.0
2.4 Petaflops #4 World
“Greenest Production SC”

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



2011 ACM Gordon Bell Prize

2013 TSUBAME-KFC
#1 Green 500



Large Scale Simulation
Big Data Analytics
Industrial Apps

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

東京工業大学
Tokyo Institute of Technology
GSIC
Global Scientific Information and Computing Center

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

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

X7~10
>>
(effectively more due to optimized DL SW Stack on GPUs)

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

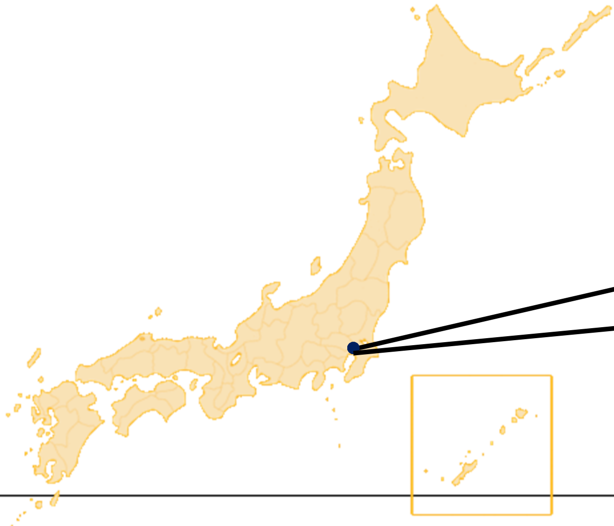
K Computer (2011)

Deep Learning
FP32 11.4 Petaflops

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

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

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

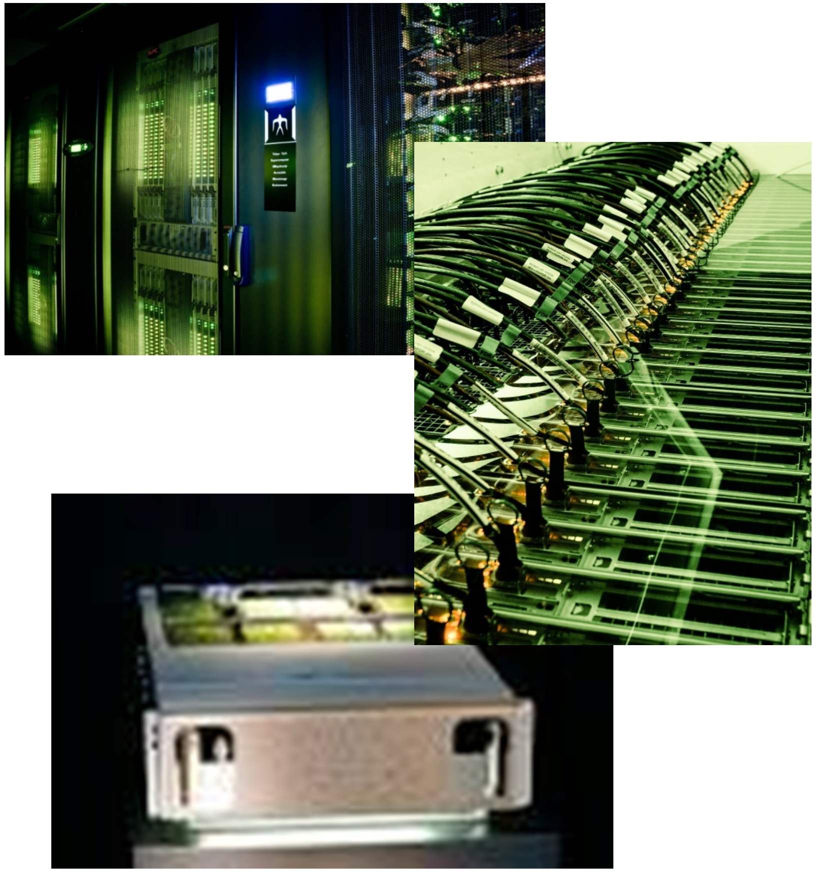


- 130~200 AI-Petaflops
- < 3MW Power
- < 1.1 Avg. PUE
- Operational 2017Q3~Q4

Univ. Tokyo Kashiwa Campus

ABCI – 2017Q4~ 2018Q1

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



Cloud Infrastructure

- **Ultra-dense IDC design from ground-up**
  - Custom inexpensive lightweight "warehouse" building w/ substantial earthquake tolerance
- **Cloud ecosystem**
  - Wide-ranging Big Data and HPC standard software stacks
- **Extreme green**
  - Ambient warm liquid cooling, large Li-ion battery storage, and high-efficiency power supplies, etc.
- **Advanced cloud-based operation**
  - Incl. dynamic deployment, container-based virtualized provisioning, multitenant partitioning, and automatic failure recovery, etc.
  - Joining HPC and Cloud Software stack for real

National Institute for Advanced Industrial Science and Technology (AIST)

Ministry of Economics Trade and Industry (METI)

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

ラボ長（産総研研究職 or 東工大教員/クロスアポ） 松岡

- 副ラボ長（産総研研究職）
- 副ラボ長（産総研事務職）
- ラボ研究主幹（産総研研究職）
- ラボ構成員

GSIC

Tsubame 3.0/2.5 Big Data /AI resources

Resources and Acceleration of AI / Big Data, systems research

Joing Organization＠Odaiba

AIST-TokyoTech AI/Big Data Open Innovation Laboratory (OIL)

AIST Artificial Intelligence Research Center (AIRC)

Joint Research on AI / Big Data and applications

Application Area
Natural Langauge Processing
Robotics
Security

Industrial Collaboration in data, applications

Basic Research in Big Data / AI algorithms and methodologies

“Smart AI Technology Research Organization”

Other Big Data / AI research organizations and proposals

Industry

DENSO IT LABORATORY, INC.

Software Ecosystem for HPC in AI

Different SW Ecosystem between HPC and AI/BD/Cloud

How to achieve convergence—for real, for rapid tech transfer

Existing Clouds

- BD/AI User Applications
- Machine Learnig MLlib/ Mahout/Chainer | Graph Processing GraphX/ Giraph /ScaleGraph | SQL/Non-SQL Hive/Pig
- Java · Scala · Python + IDL
- MapReduce Framework Spark/Hadoop
- RDB PostgresQL | CloudDB/NoSQL Hbase/Cassandra/MondoDB
- Distributed Filesysem HDFS & Object Store | Coordination Service ZooKeeper
- VM(KVM), Container(Docker), Cloud Services (OpenStack)
- Linux OS
- Ethernet TOR Swtiches High Latency/Low Capacity NW | Local Node Storage | x86 CPU

Application Layer

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

System Software Layer

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

OS Layer

Hardware Layer

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

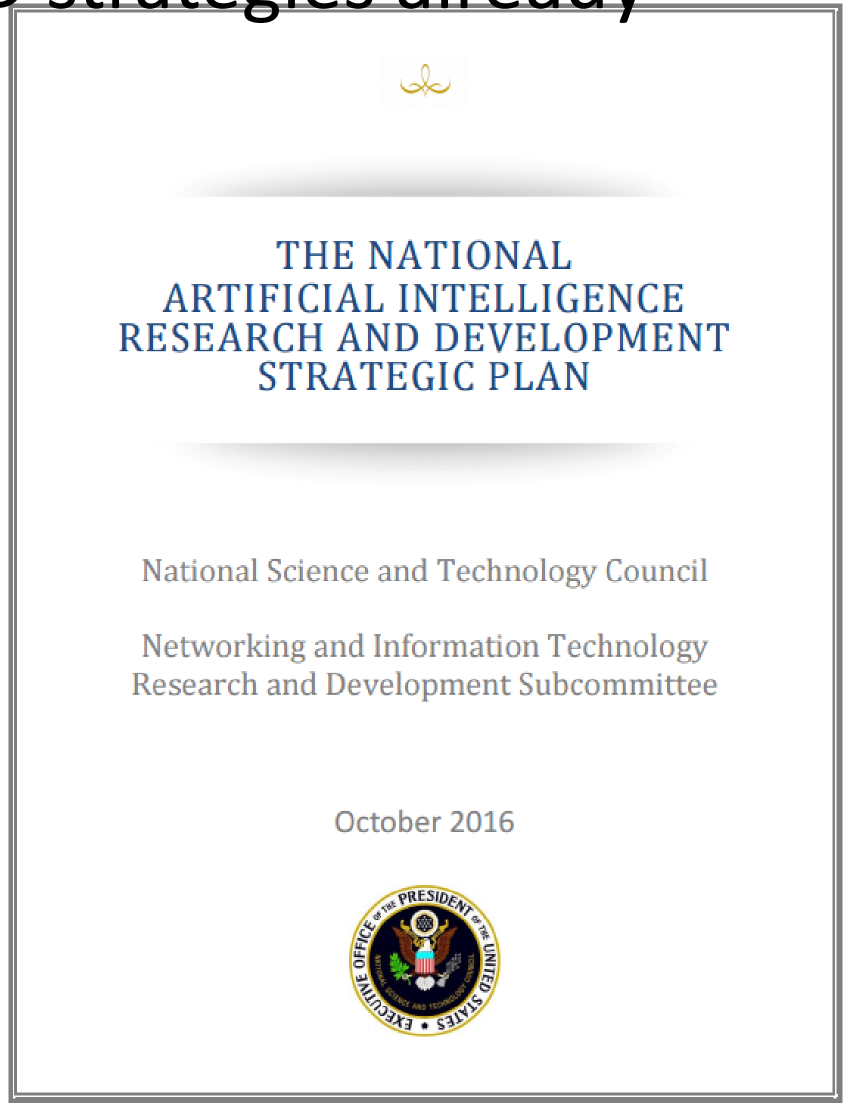
Existing Supercomputers

- HPC User Code
- Numerical Libraries LAPACK, FFTW | Various DSLs | Workflow Systems
- Fortran · C · C++ + IDL
- MPI · OpenMP/ACC · CUDA/OpenCL
- Parallel Debuggers and Profilers
- Parallel Filesystem Lustre, GPFS, | Batch Job Schedulers PBS Pro, Slurm, UGE
- Linux OS
- InfiniBand/OPA High Capacity Low Latency NW | High Performance SAN+Burst Buffers | X86 + Accelerators e.g. GPUs, FPGAs

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

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

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


THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN

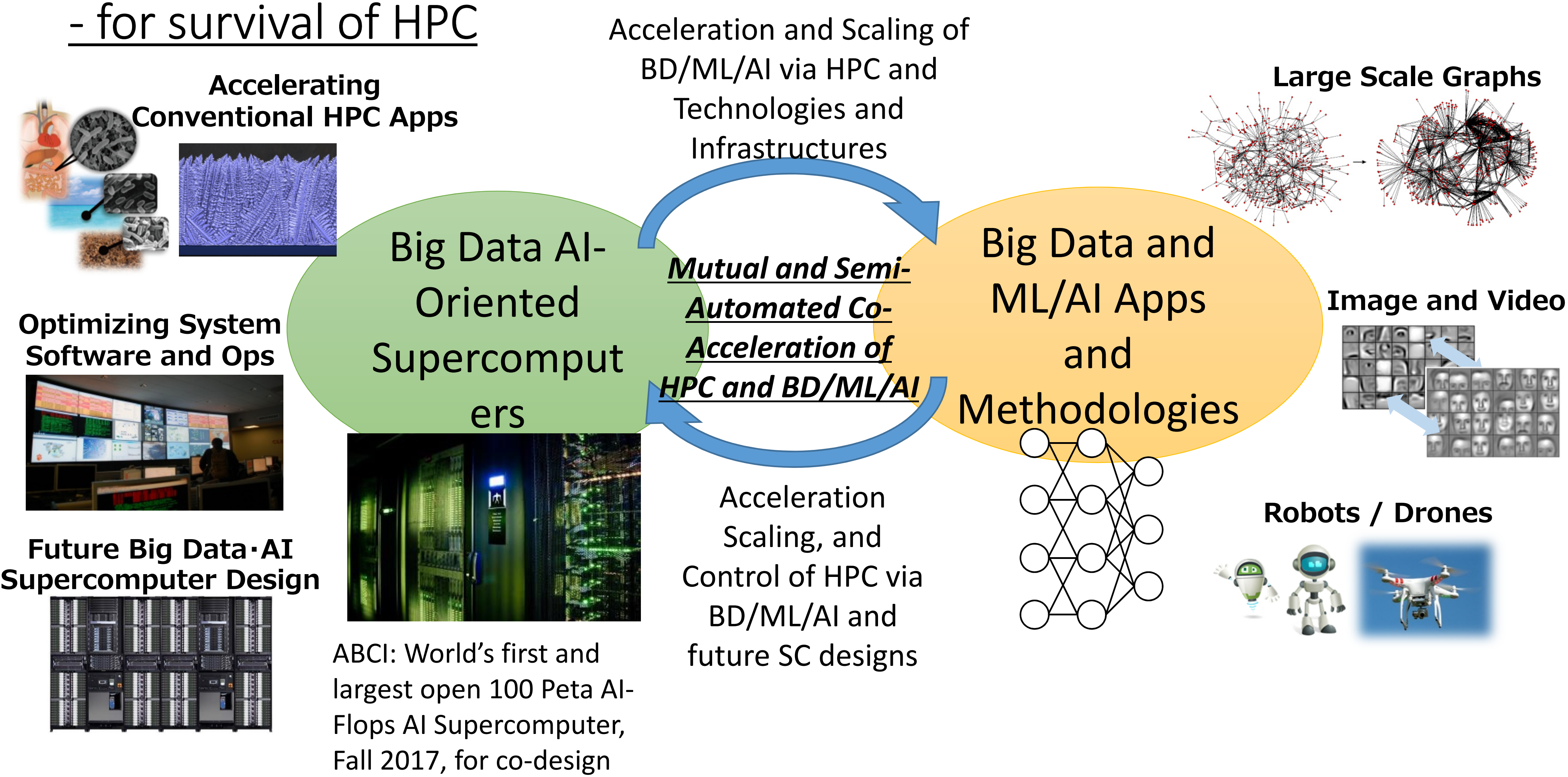
National Science and Technology Council

Networking and Information Technology Research and Development Subcommittee

October 2016

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

- for survival of HPC



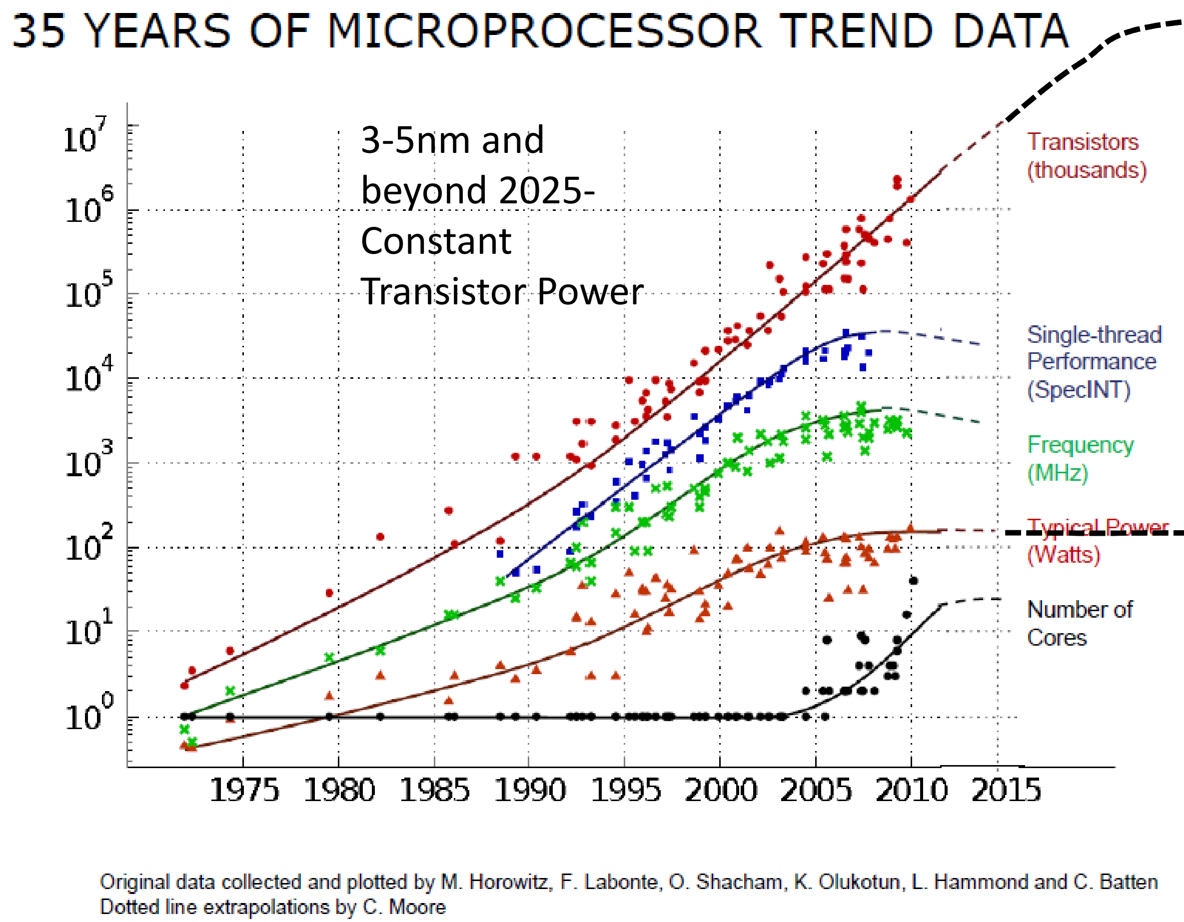
What is worse: Moore's Law will end in the 2020's

- 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



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

20 year Post-Dennard Many-Core Era

20-year Next-Gen Post-Moore era

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

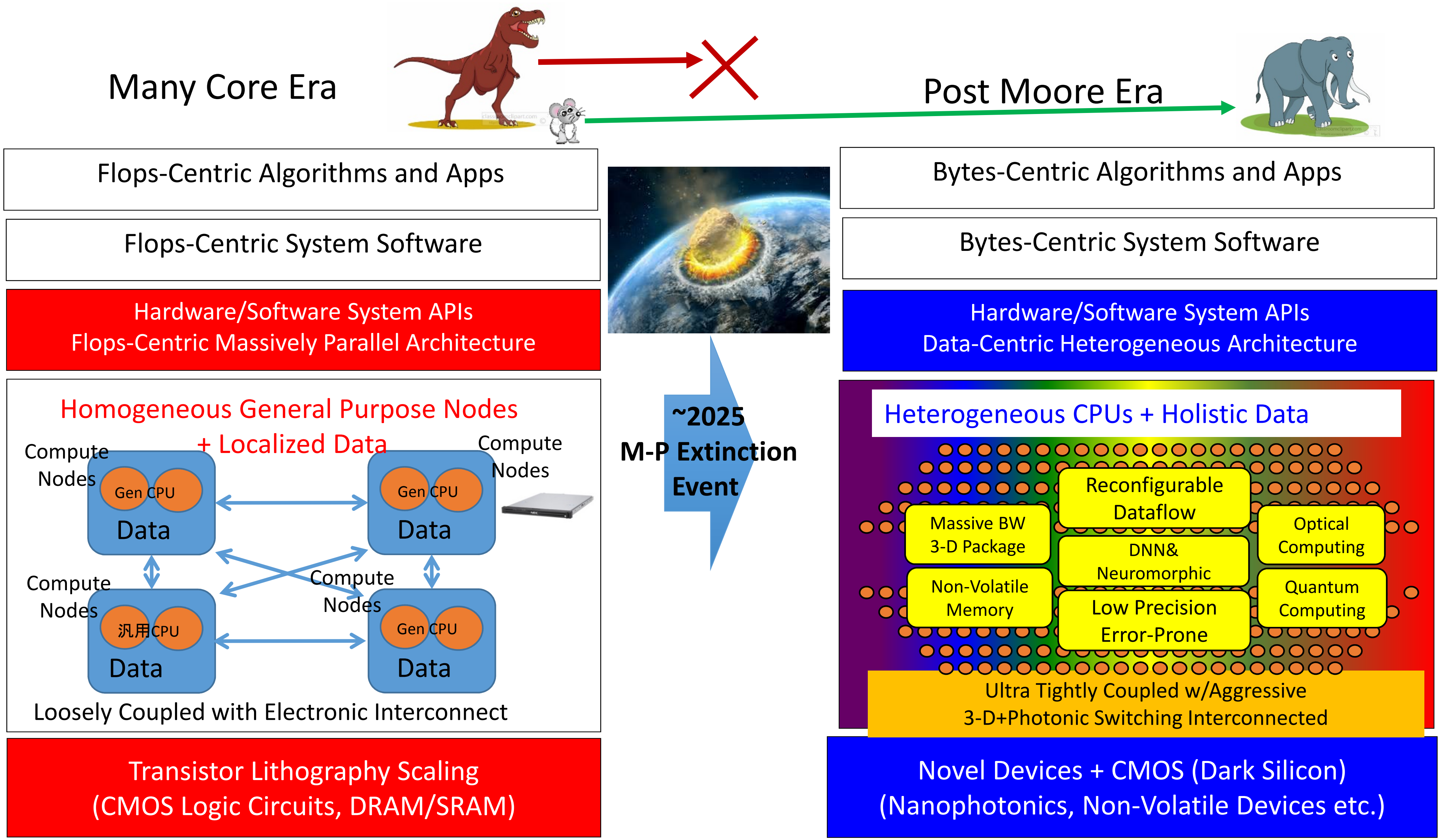
Need to realize the next 20-year era of supercomputing

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

- Systems people have been telling the algorithm people that “FLOPS will be free, bandwidth is important, so devise algorithms under that assumption”
- This will certainly be true until exascale in 2020…
- But when Moore’s Law ends in 2025-2030, constant transistor power (esp. for logic) = FLOPS will no longer be free!
- <u>So algorithms that simply increase arithmetic intensity will no longer scale beyond that point</u>
- 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*

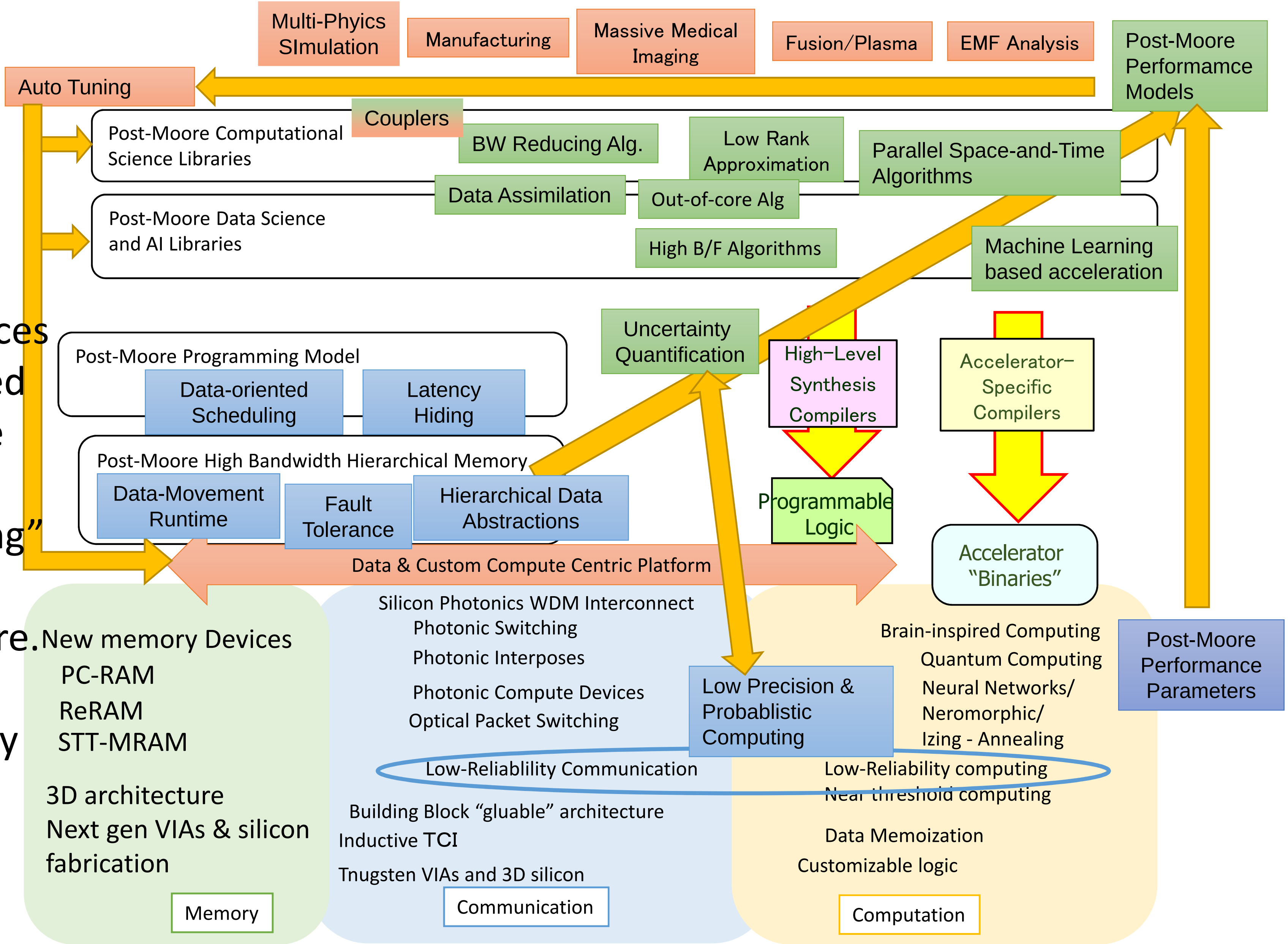


Many Core Era
Post Moore Era
Flops-Centric Algorithms and Apps
Flops-Centric System Software
Hardware/Software System APIs
Flops-Centric Massively Parallel Architecture
Homogeneous General Purpose Nodes
+ Localized Data
Compute Nodes
Gen CPU
Data
汎用CPU
Loosely Coupled with Electronic Interconnect
Transistor 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



Post Moore Era Supercomputing Workshop @ SC16

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