Arm's Race with Post-K and it's Game Changing Processor

Satoshi Matsuoka
Director, Riken Center for Computational Science /
Professor, Tokyo Institute of Technology

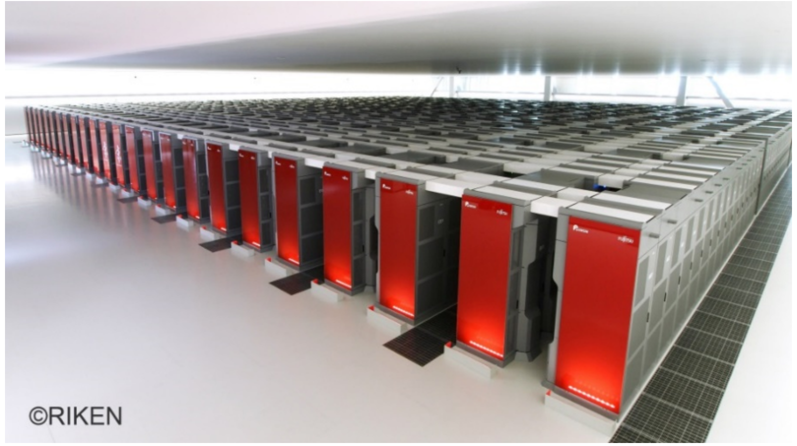
w/Mitsuhisa Sato, Yutaka Ishikawa, Riken-CCS

ISC 2018 GoingArm Workshop

K computer

Specifications

- Massively parallel, general purpose supercomputer
- No. of nodes : 88,128
- Peak speed: 11.28 Petaflops
- Memory: 1.27 PB
- Network: 6-dim mesh-torus (Tofu)



Top 500 ranking

LINPACK measures the speed and efficiency of linear equation calculations
Real applications require more complex computations.

- No.1 in Jun. & Nov. 2011
- No.10 in Nov. 2017

Graph 500 ranking

"Big Data" supercomputer ranking
Measures the ability of data-intensive loads

- No.1 in Nov. 2017

HPCG ranking

Measures the speed and efficiency of solving linear equation using HPCG
Better correlate to actual applications

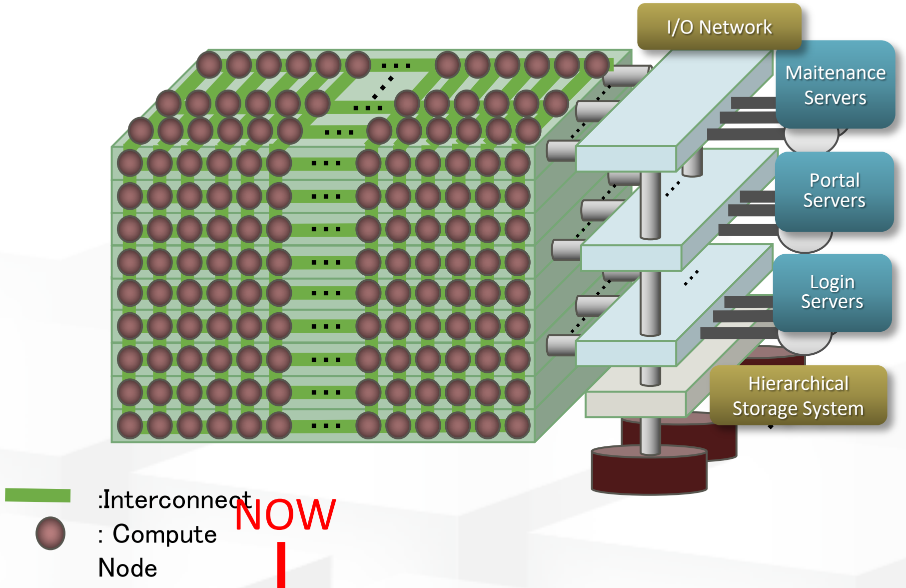
- No. 1 in Nov. 2017

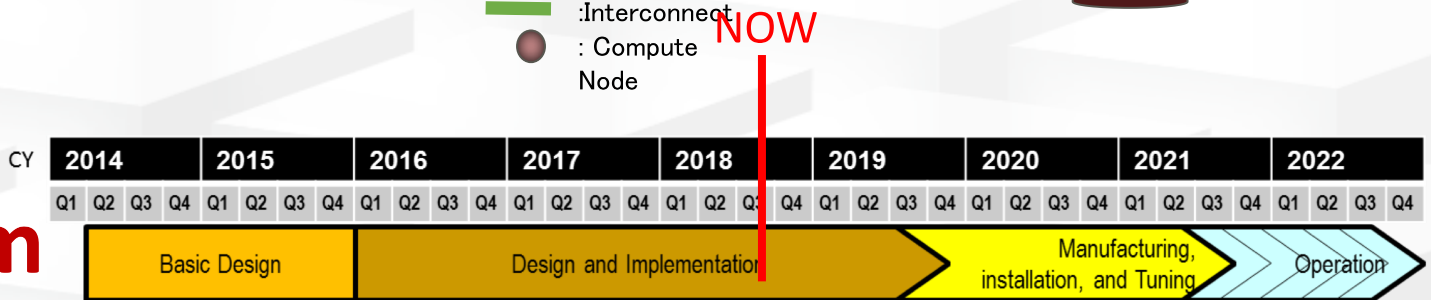
K computer achieved balance of processor speed, memory, and network.
high performance for wide areas of science.

Japan Flagship 2020 “Post K” Supercompute

- ✓CPU
  - Many core, Xeon-Class ARM v8 cores + 512 bit SVE (scalable vector extensions)
  - Multi-hundred petaflops peak total
  - Power Knob feature
- ✓Memory
  - ✓3-D stacked DRAM, Terabyte/s BW /chip
- ✓Interconnect
  - TOFU3 CPU-integrated 6-D torus network
- I/O acceleration with massive SDs
- 30MW+ Power, liquid cooled
- Riken co-design with **Fujitsu**
- **? Million cores in system**

Prime Minister Abe visiting K Computer 2013





Post-K: The Game Changer

1. **Heritage of the K-Computer, HP in simulation via extensive Co-Design**

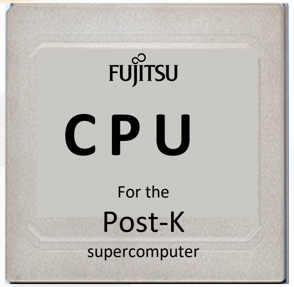
- High performance: up to x100 performance of K in real applications
- Multitudes of Scientific Breakthroughs via Post-K application programs
- Simultaneous high performance and ease-of-programming

2. New Technology Innovations of Post-K

- **High Performance, esp. via high memory BW**
  Performance boost by "factors" c.f. mainstream CPUs in many HPC & Society5.0 apps
- **Very Green e.g. extreme power efficiency**
  Ultra Power efficient design & various power control knobs
- **Arm Global Ecosystem & SVE contribution**
  ARM Ecosystem: 21 billion chips/year, SVE co-design and world's first implementation by Fujitsu, to become global std.
- **High Perf. on Society5.0 apps incl. AI**
  Architectural features for high perf on Society 5.0 apps based on Big Data, AI/ML, CAE/EDA, Blockchain security, etc.

Global leadership not just in the machine & apps, but as cutting edge IT

ARM: Massive ecosystem from embedded to HPC

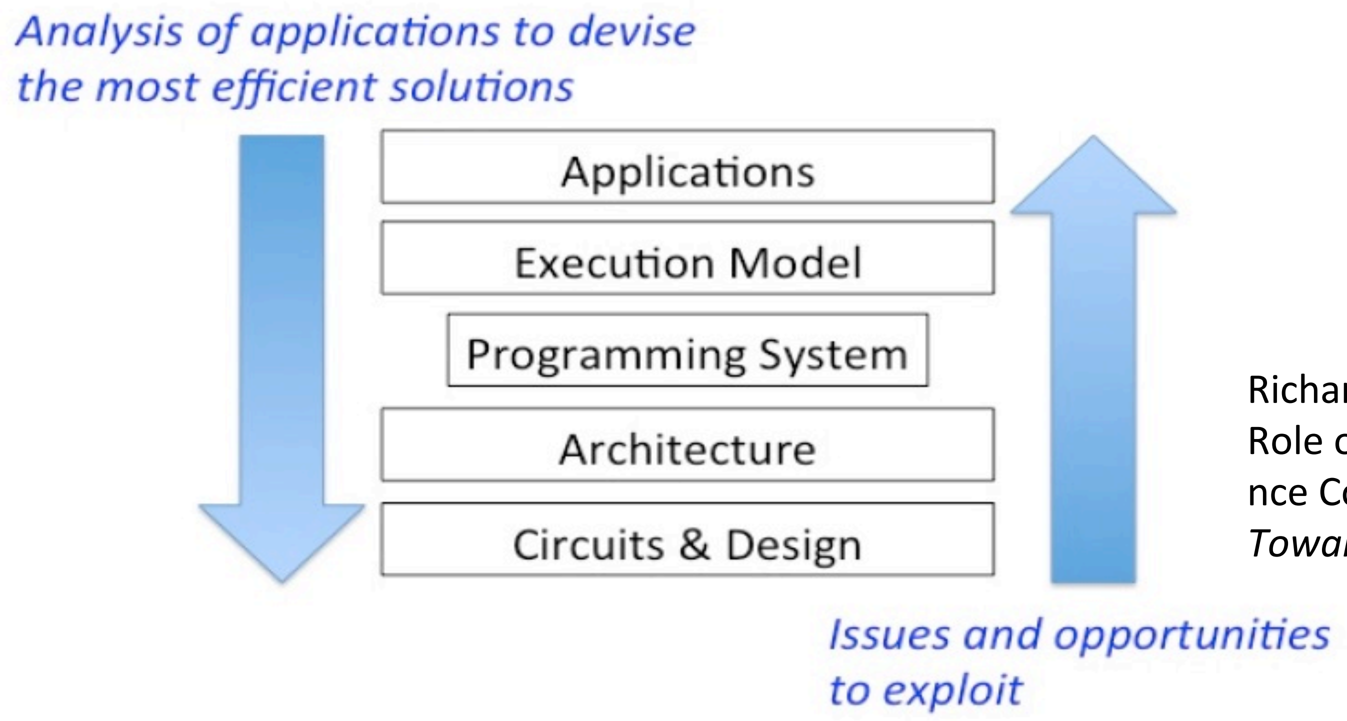


Technology not just limited to Post-K, but into societal IT infrastructures e.g. Clouds

ARM for HPC - Co-design Opportunities, ISC 2018 BOF
June 25, 2018

Co-design for Post-K

(slides by Mitsuhisa Sato **Team Leader of Architecture Development Team)**
Deputy project leader, FLAGSHIP 2020 project
Deputy Director, RIKEN Center for Computational Science (R-CCS)



Richard F. BARRETT, et.al. “On the Role of Co-design in High Performance Computing”, *Transition of HPC Towards Exascale Computing*

Co-design from Apps to Architecture

- **Architectural Parameters to be determined**
  - #SIMD, SIMD length, #core, #NUMA node, O3 resources, specialized hardware
  - cache (size and bandwidth), memory technologies
  - Chip die-size, power consumption
  - Interconnect
- **We have selected a set of target applications**
- **Performance estimation tool**
  - Performance projection using Fujitsu FX100 execution profile to a set of arch. parameters.
- **Co-design Methodology (at early design phase)**

1. **Setting set of system parameters**
2. **Tuning target applications under the system parameters**
3. **Evaluating execution time using prediction tools**
4. **Identifying hardware bottlenecks and changing the set of system parameters**

Target applications representatives of almost all our applications in terms of computational methods and communication patterns in order to design architectural features.

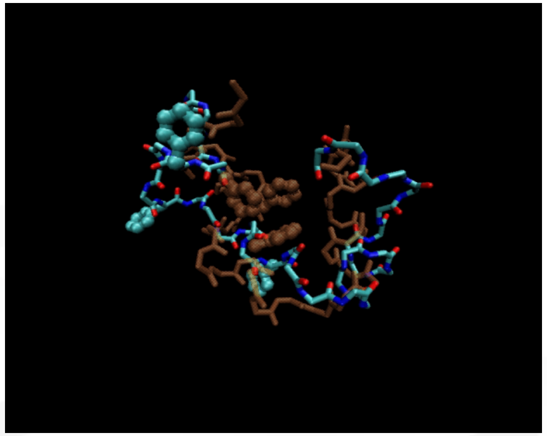
	Target Application	
	Program	**Brief description**
①	GENESIS	MD for proteins
②	Genomon	Genome processing (Genome alignment)
③	GAMERA	Earthquake simulator (FEM in unstructured & structured grid)
④	NICAM+LETK	Weather prediction system using Big data (structured grid stencil & ensemble Kalman filter)
⑤	NTChem	molecular electronic (structure calculation)
⑥	FFB	Large Eddy Simulation (unstructured grid)
⑦	RSDFT	an ab-initio program (density functional theory)
⑧	Adventure	Computational Mechanics System for Large Scale Analysis and Design (unstructured grid)
⑨	CCS-QCD	Lattice QCD simulation (structured grid Monte Carlo)

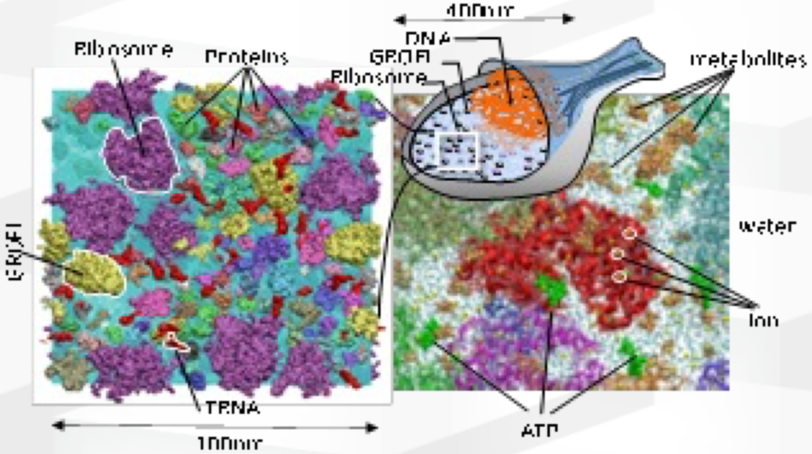
Genesis MD: proteins in a cell environment

Protein simulation before K

- Simulation of a protein in isolation

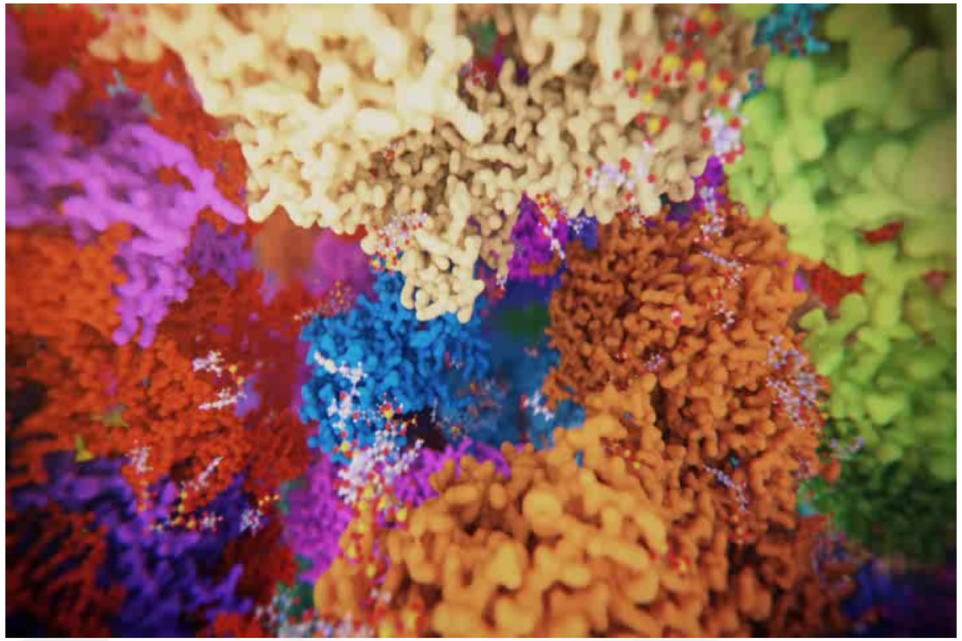
Folding simulation of Villin, a small protein with 36 amino acids





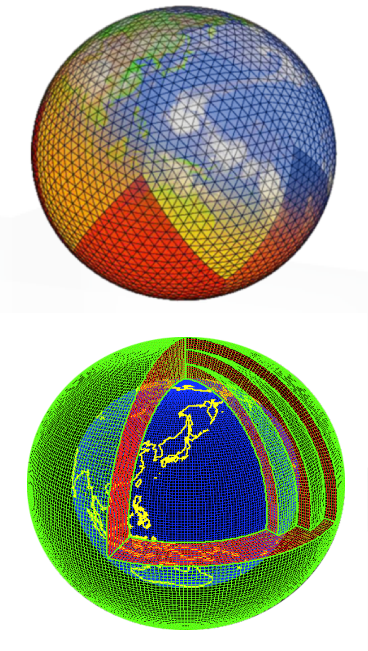
Protein simulation with K

- all atom simulation of a cell interior
- cytoplasm of Mycoplasma genitalium

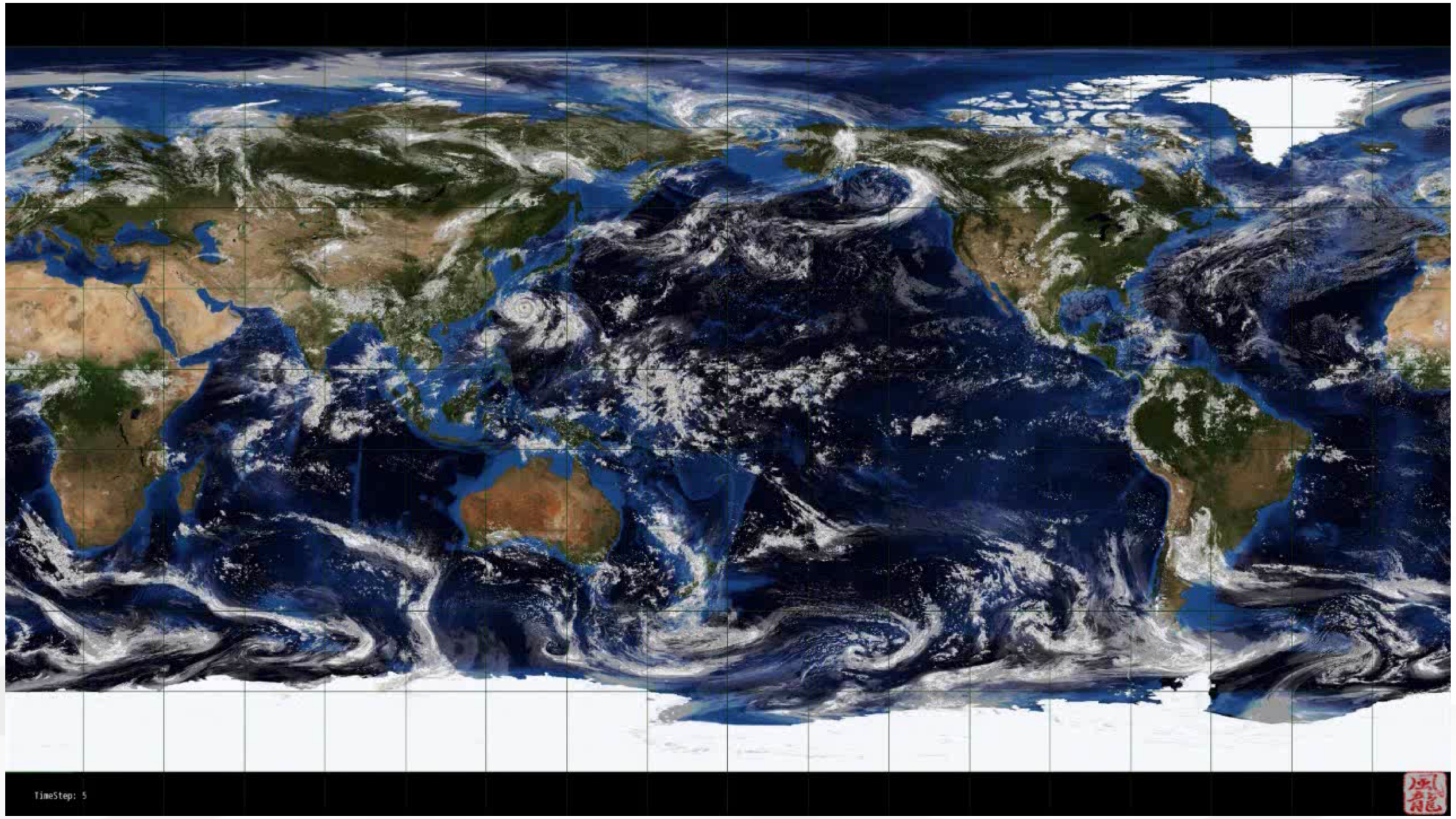


NICAM: Global Climate Simulation

- Global cloud resolving model **with 0.87 km-mesh** which allows resolution of cumulus clouds
- Month-long forecasts of Madden-Julian oscillations in the tropics is realized.



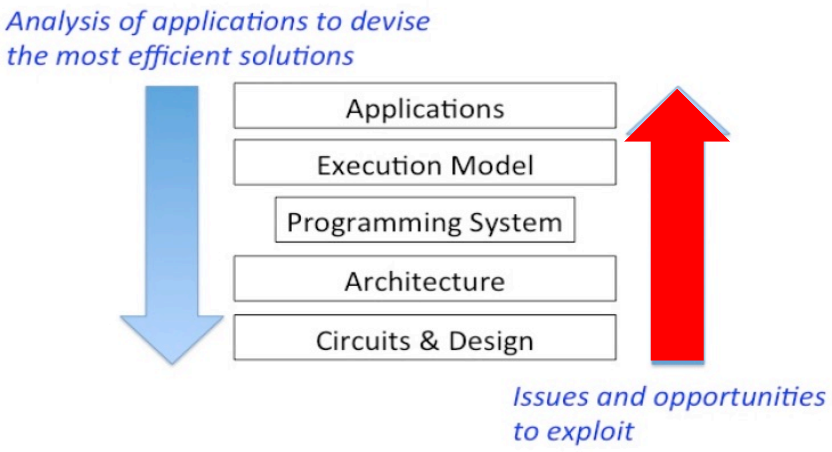
Global cloud resolving model



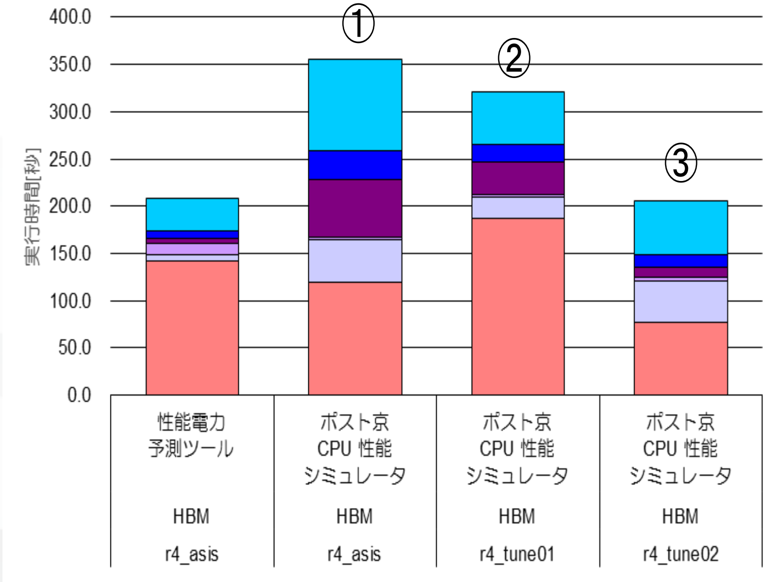
Miyamoto et al (2013) , Geophys. Res. Lett., 40, 4922–4926, doi:10.1002/grl.50944.

Co-design of Apps for Architecture

- **Tools for performance tuning**
  - Performance estimation tool
    - Performance projection using Fujitsu FX100 execution profile
    - Gives "target" performance
  - **Post-K processor simulator**
    - **Based on gem5, O3, cycle-level simulation**
    - **Very slow, so limited to kernel-level evaluation**



- **Co-design of apps**
  - 1. Estimate "target" performance using performance estimation tool
  - 2. Extract kernel code for simulator
  - 3. Measure exec time using simulator
  - 4. Feed-back to code optimization
  - 5. Feed-back to compiler

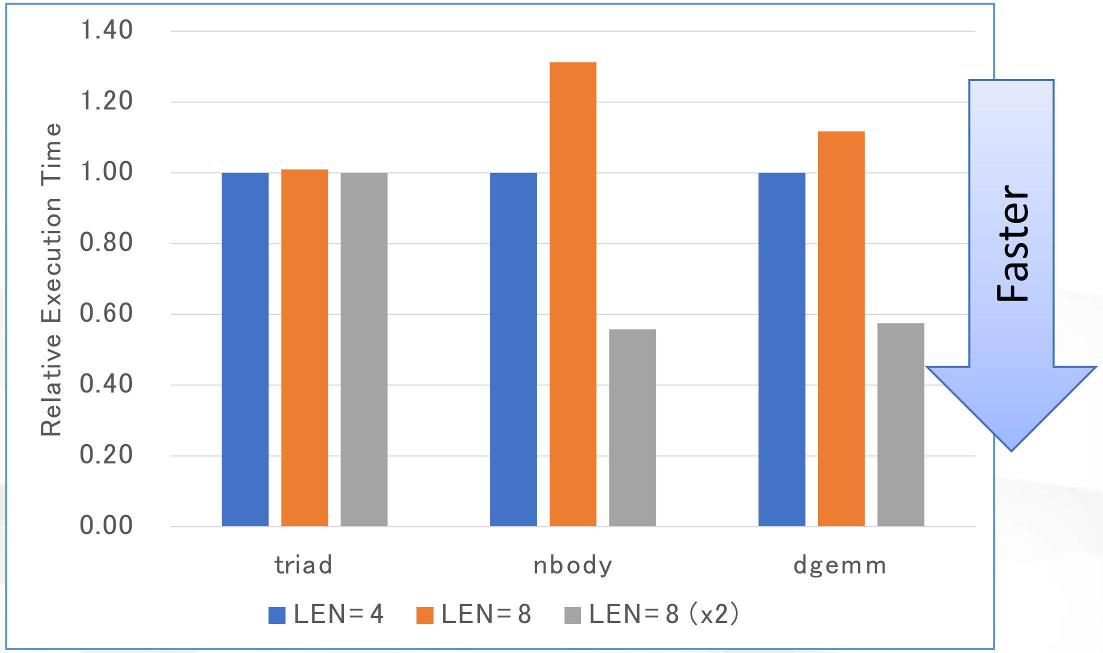


ARM for HPC - Co-design Opportunities

- **ARM SVE Vector Length Agnostic feature is very interesting, since we can examine vector performance using the same binary.**
- **We have investigated how to improve the performance of SVE keeping hardware-resource the same. (in "Rev-A" paper)**
  - ex. "512 bits SVE x 2 pipes" vs. "1024 bits SVE x 1 pipe"
  - Evaluation of **Performance and Power** (in "coolchips" paper) by using our gem-5 simulator (with "white" parameter) and ARM compiler.
  - Conclusion: Wide vector size over FPU element size will improve performance if there are enough rename registers and the utilization of FPU has room for improvement.

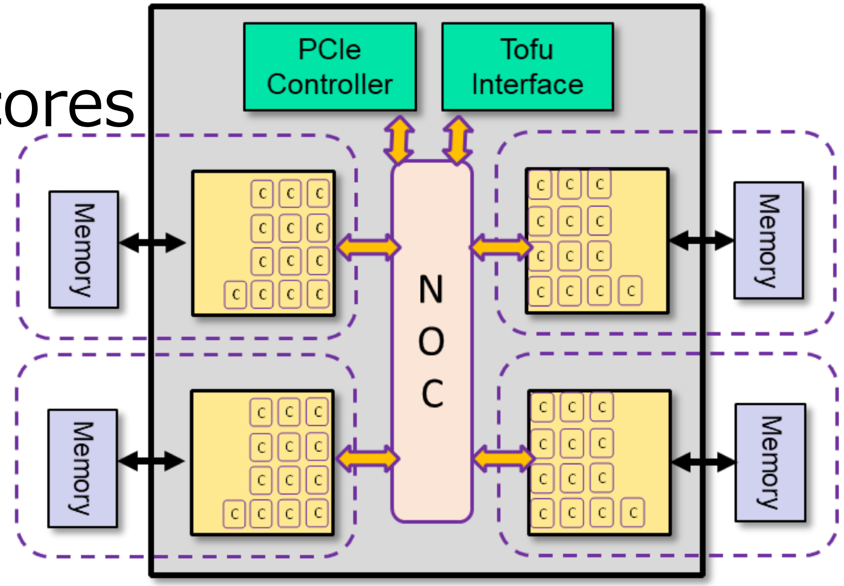
Note that these researches are not relevant to "post-K" architecture.

- Y. Kodama, T. Oajima and M. Sato. "Preliminary Performance Evaluation of Application Kernels Using ARM SVE with Multiple Vector Lengths", In Re-Emergence of Vector Architectures Workshop (Rev-A) in 2017 IEEE International Conference on Cluster Computing, pp. 677-684, Sep. 2017.
- T. Odajima, Y. Kodama and M. Sato, "Power Performance Analysis of ARM Scalable Vector Extension", In IEEE Symposium on Low-Power and High-Speed Chips and Systems (COOL Chips 21), Apr. 2018



Post K Processor is…

- **an Many-Core ARM CPU…**
  - 48 compute cores + 2 or 4 assistant (OS) cores
  - Brand new core design
  - Near Xeon-Class Integer performance core
  - ARM V8 --- 64bit ARM ecosystem
  - Tofu 3 + PCIe 3 external connection

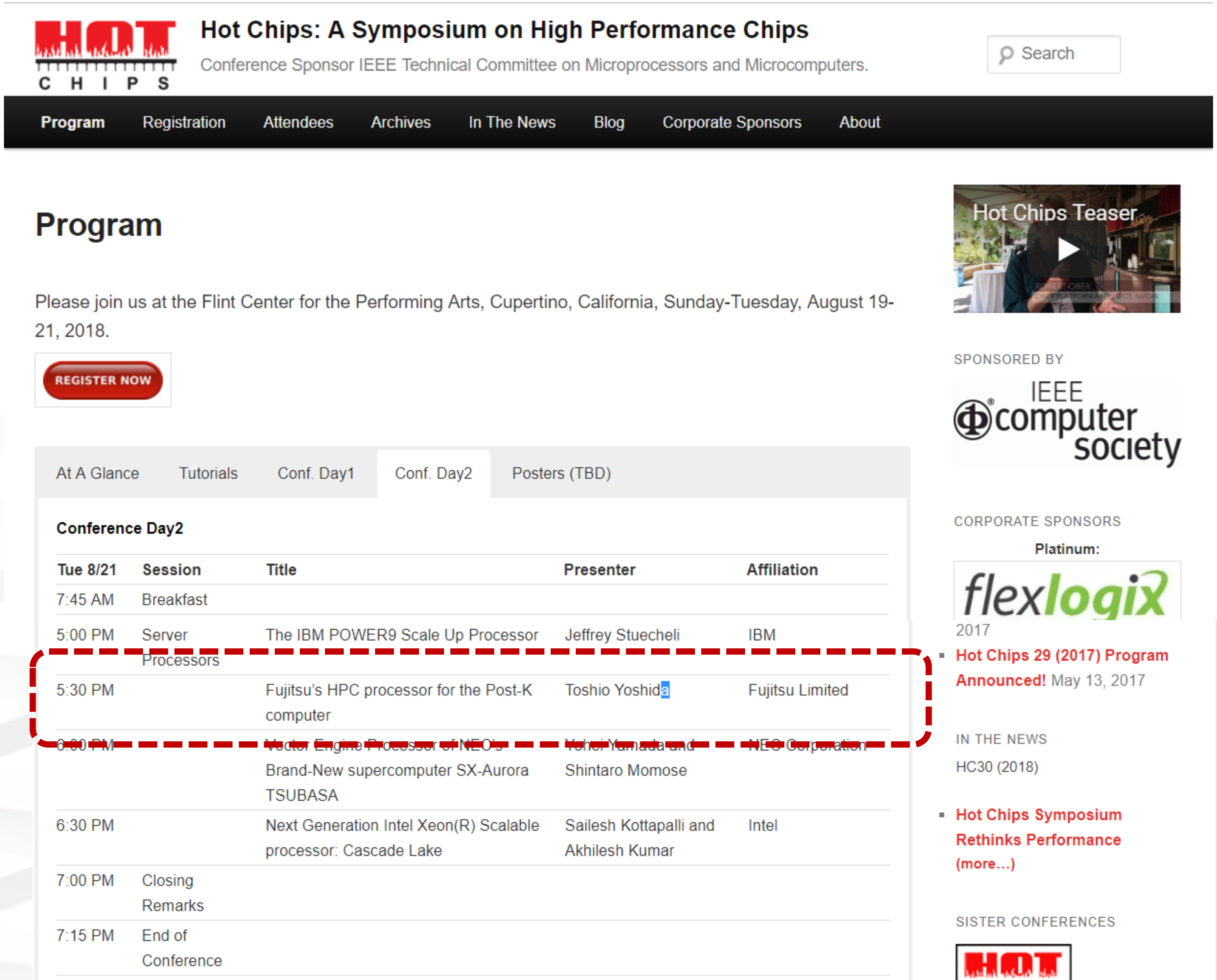


- **…but also a GPU-like processor**
  - SVE 512 bit vector extensions (ARM & Fujitsu)
    - Integer (1, 2, 4, 8 bytes) + Float (16, 32, 64 bytes)
  - Cache + scratchpad local memory (sector cache)
  - Multi-stack 3D mem – ~TB/s Mem BW (Bytes/DPF ~0.4)
    - Streaming memory access, strided access, scatter/gather etc.
  - Intra-chip barrier synch. and other memory enhancing features

- **GPU-like High performance in HPC, AI/Big Data, Blockchain…**

Post K Processor and other Details

- Aug. 21, Hotchips 2018 @ Stanford U
- Other details (new TOFU, detailed performance, Post-K machine config., etc.) forthcoming towards Fall, 2018.
- Gem5 Simulator availability under NDA from Riken
- Early chip availability up to Fujitsu



Hot Chips: A Symposium on High Performance Chips

Conference Sponsor IEEE Technical Committee on Microprocessors and Microcomputers.

Search

Program | Registration | Attendees | Archives | In The News | Blog | Corporate Sponsors | About

Program

Please join us at the Flint Center for the Performing Arts, Cupertino, California, Sunday-Tuesday, August 19-21, 2018.

REGISTER NOW

At A Glance | Tutorials | Conf. Day1 | Conf. Day2 | Posters (TBD)

Conference Day2

Tue 8/21	Session	Title	Presenter	Affiliation
7:45 AM	Breakfast			
5:00 PM	Server Processors	The IBM POWER9 Scale Up Processor	Jeffrey Stuecheli	IBM
5:30 PM		Fujitsu's HPC processor for the Post-K computer	Toshio Yoshida	Fujitsu Limited
6:00 PM		Vector Engine Processor of NEC's Brand-New supercomputer SX-Aurora TSUBASA	Yohei Yamada and Shintaro Momose	NEC Corporation
6:30 PM		Next Generation Intel Xeon(R) Scalable processor: Cascade Lake	Sailesh Kottapalli and Akhilesh Kumar	Intel
7:00 PM	Closing Remarks			
7:15 PM	End of Conference			

Hot Chips Teaser

SPONSORED BY

IEEE computer society

CORPORATE SPONSORS

Platinum:

flexlogix

2017

- Hot Chips 29 (2017) Program Announced! May 13, 2017

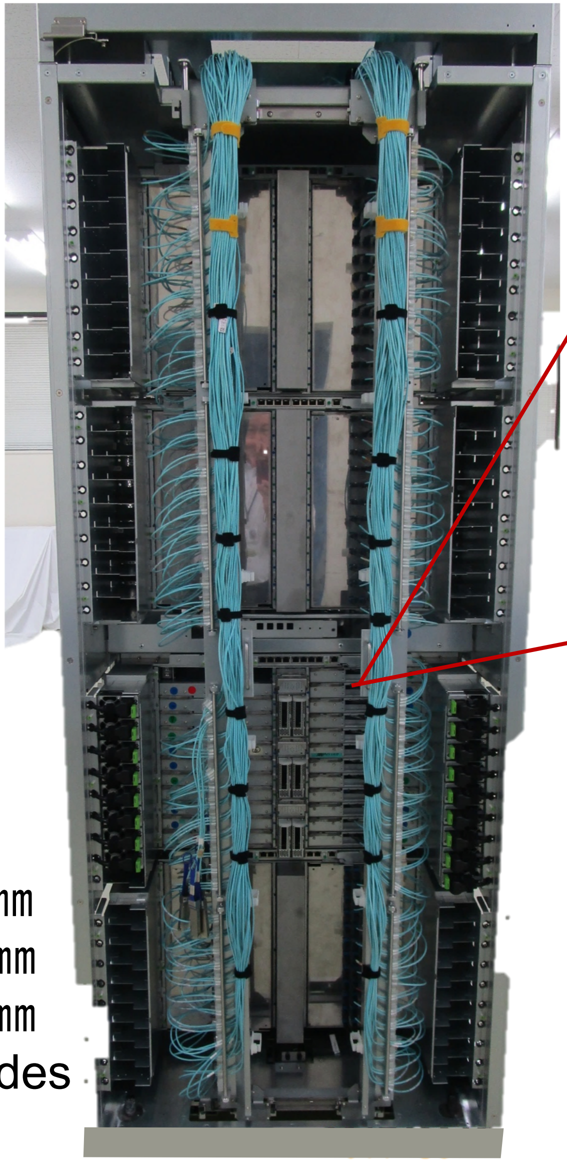
IN THE NEWS

HC30 (2018)

- Hot Chips Symposium Rethinks Performance (more...)

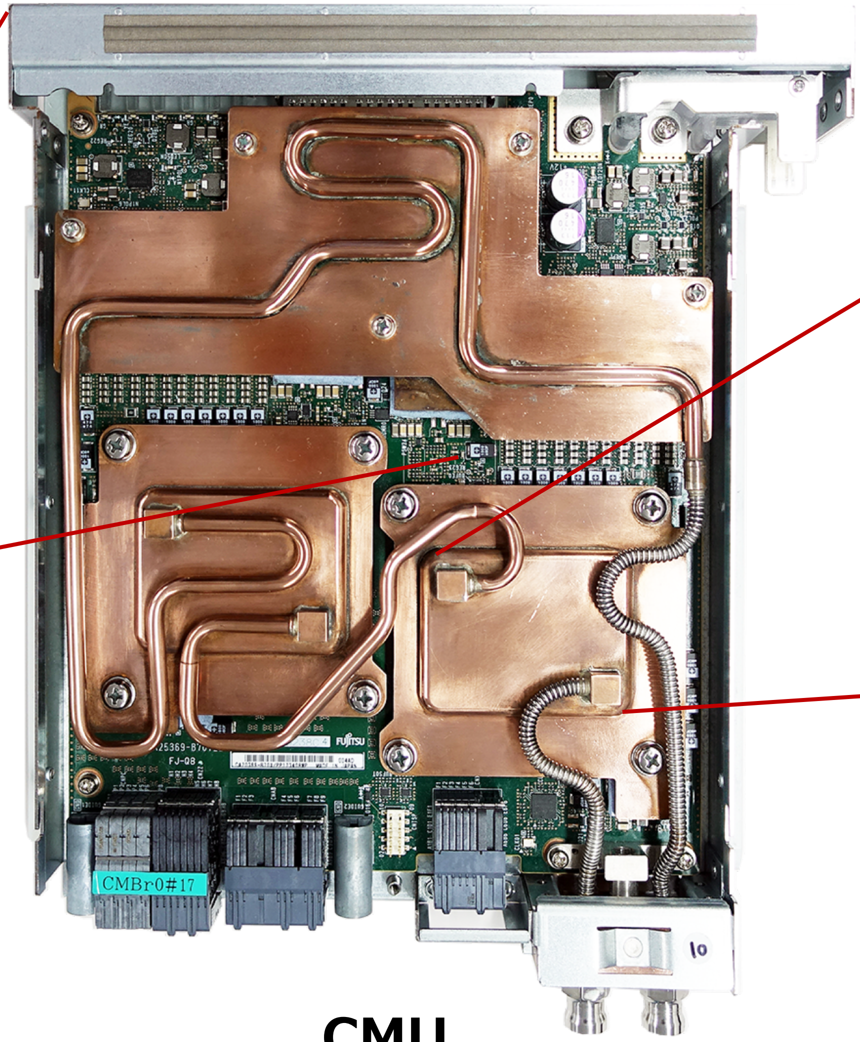
SISTER CONFERENCES

Post-K Chassis, PCB (w/DLC), and CPU Package



W 800mm
D1400mm
H2000mm
384 nodes

230 mm



280 mm

CMU

60 mm

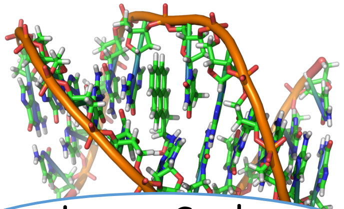
60 mm

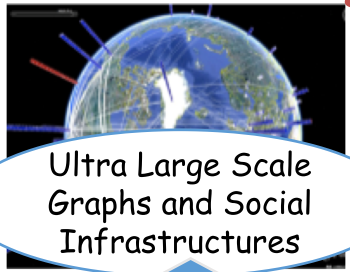
CPU Package

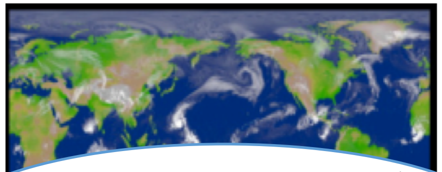
A0 Chip Booted in June
Undergoing Tests

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

Future Non-Silo Extreme Big Data Scientific Apps







Large Scale Metagenomics

Ultra Large Scale Graphs and Social Infrastructures

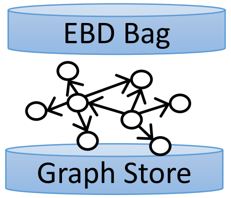
Massive Sensors and Data Assimilation in Weather Prediction

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

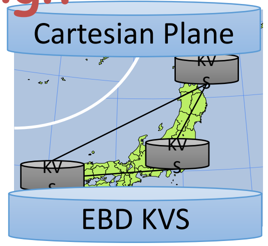
Co-Design

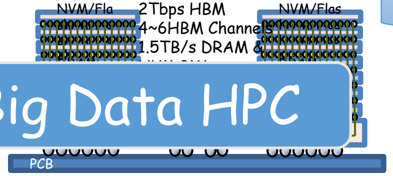
Co-Design

Co-Design



EBD System Software incl. EBD Object System





Exascale Big Data HPC

Bring HPC rigor in architectural, algorithmic, and system software performance and modeling into big data

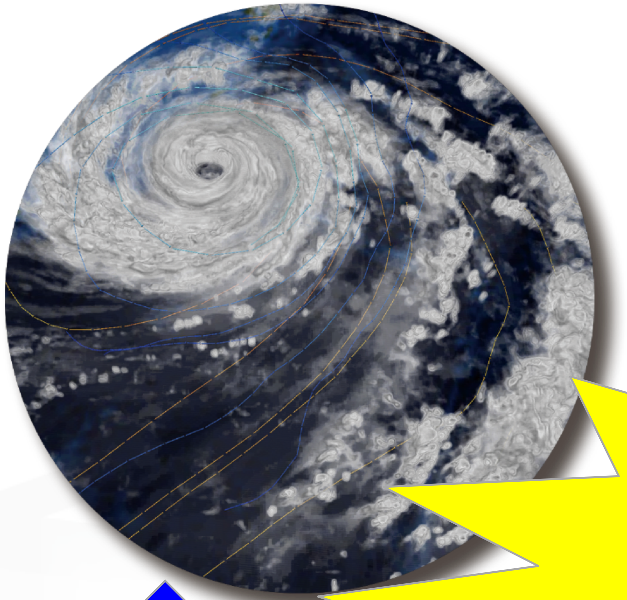
Convergent Architecture (Phases 1~4)
Large Capacity NVM, High-Bisection NW

Cloud IDC
Very low BW & Efficiency
Highly available, resilient

Supercomputers
Compute&Batch-Oriented
More fragile

Pioneering “Big Data Assimilation” Era

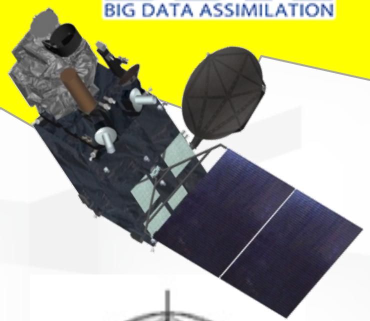
High-precision Simulations



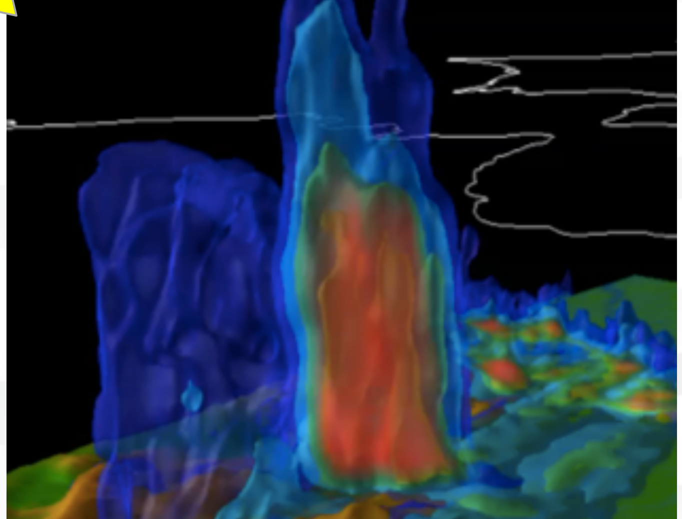
国立研究開発法人
科学技術振興機構
Japan Science and Technology Agency
CREST

Future-generation technologies available 10 years in advance

BDA
BIG DATA ASSIMILATION



High-precision observations



Mutual feedback

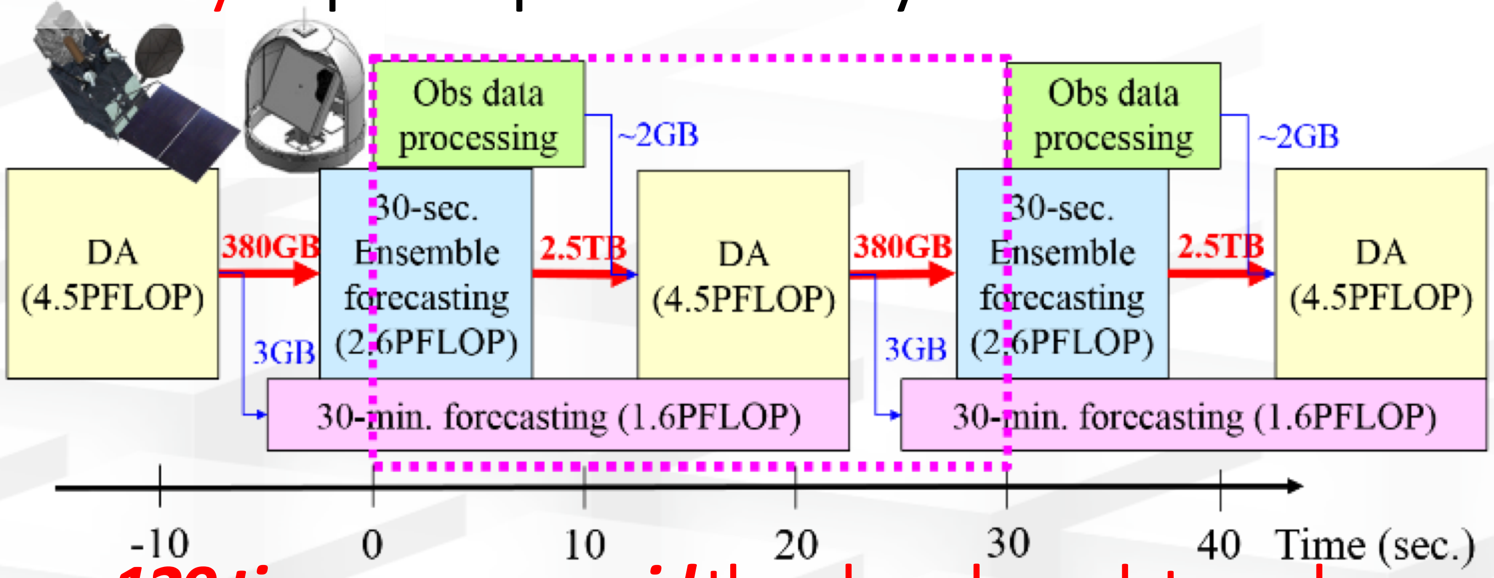
Takemasa Miyoshi, Riken-CCS (Weather Data Assimilation)

Big Data Assimilation

for severe weather forecast

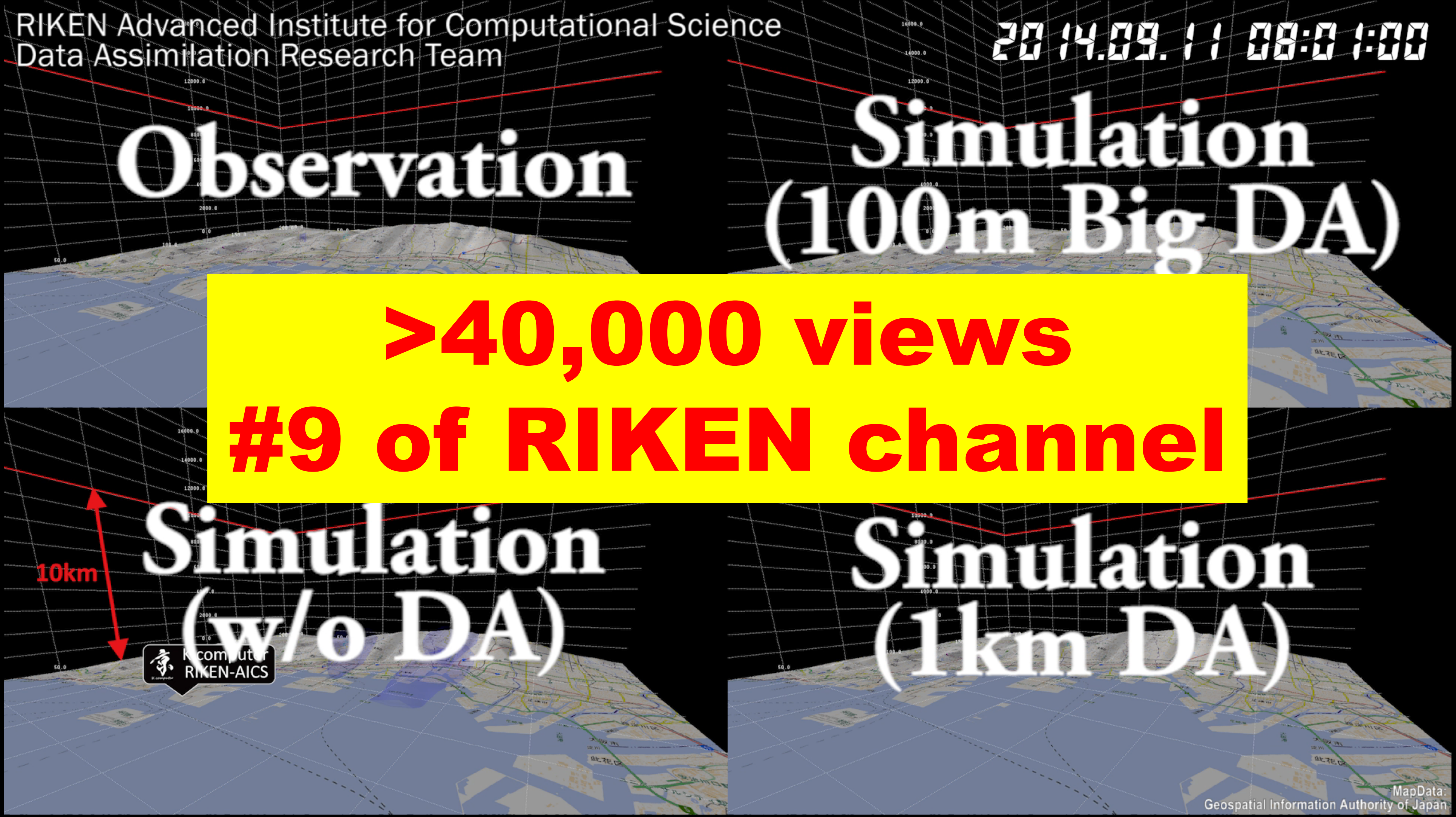
Goal ： Pinpoint (100-m resol.) forecast of severe local weather by updating 30-min forecast every 30 sec!

Revolutionary super-rapid 30-sec. cycle



120 times more rapid than hourly update cycles

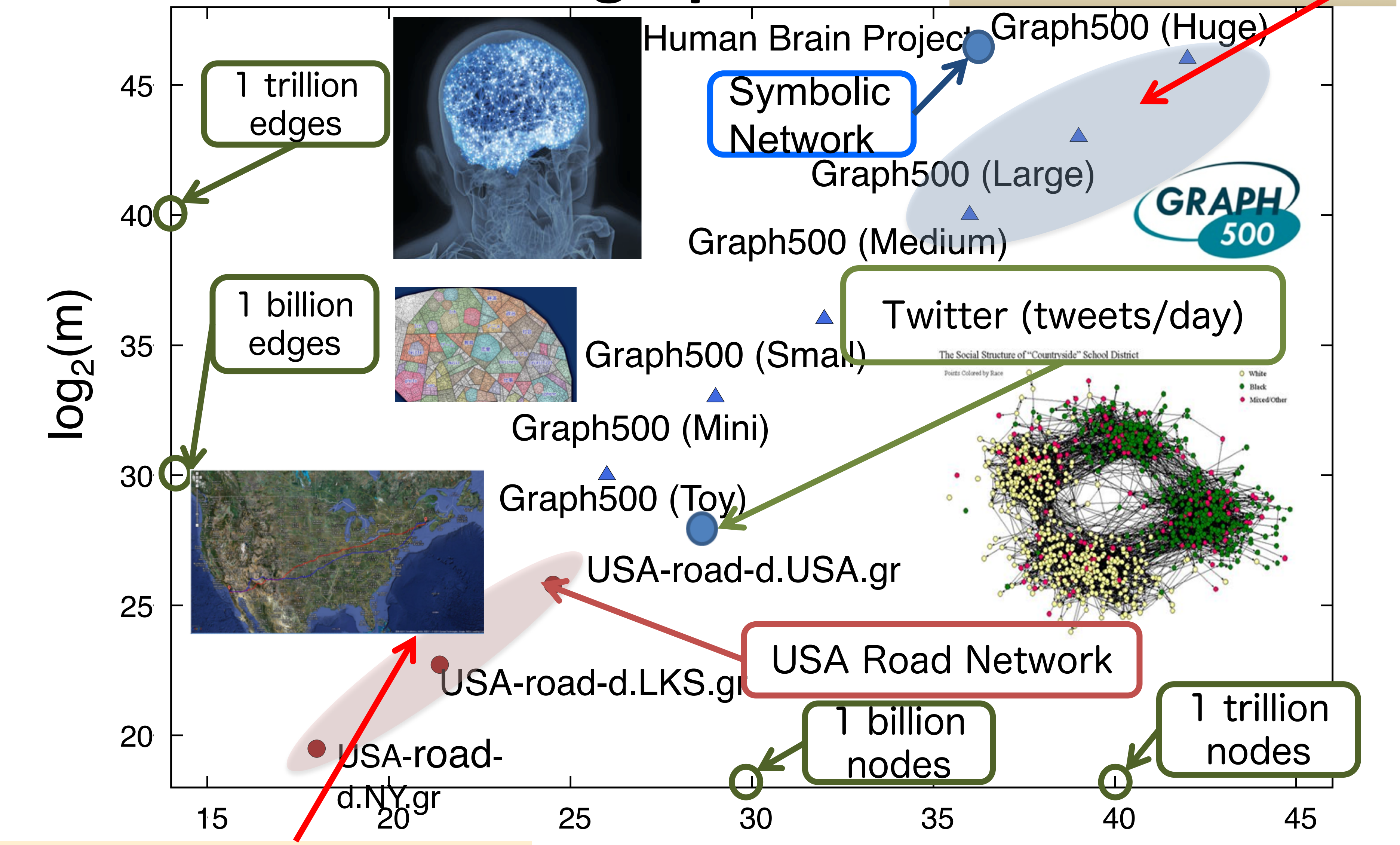
9/11/2014, sudden local rain



>40,000 views

#9 of RIKEN channel

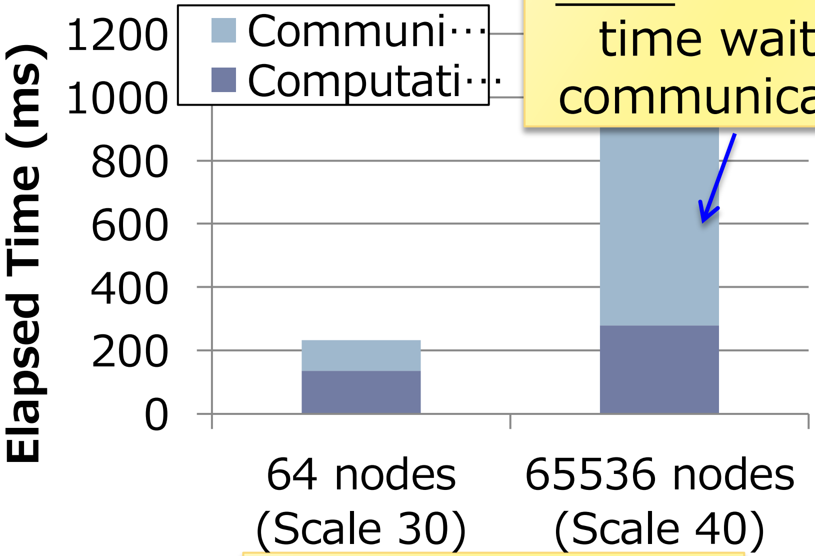
The size of graphs



Mobile Phone : SONY SO-01F
Snapdragon S4 1.7GHz 4core: 2GB RAM
1.03GTEPS: 235.06MTEPS/W

Sparse BYTES: The Graph500 – 2015~2016 – world #1 x 4

K Computer #1 Tokyo Tech[Matsuoka EBD CREST] Univ. Kyushu [Fujisawa Graph CREST], Riken AICS, Fujitsu



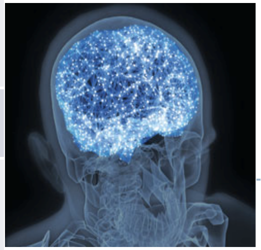
73% total exec time wait in communication

88,000 nodes,
660,000 CPU Cores
1.3 Petabyte mem
20GB/s Tofu NW

K computer

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



≫

Effective x13 performance c.f. Linpack

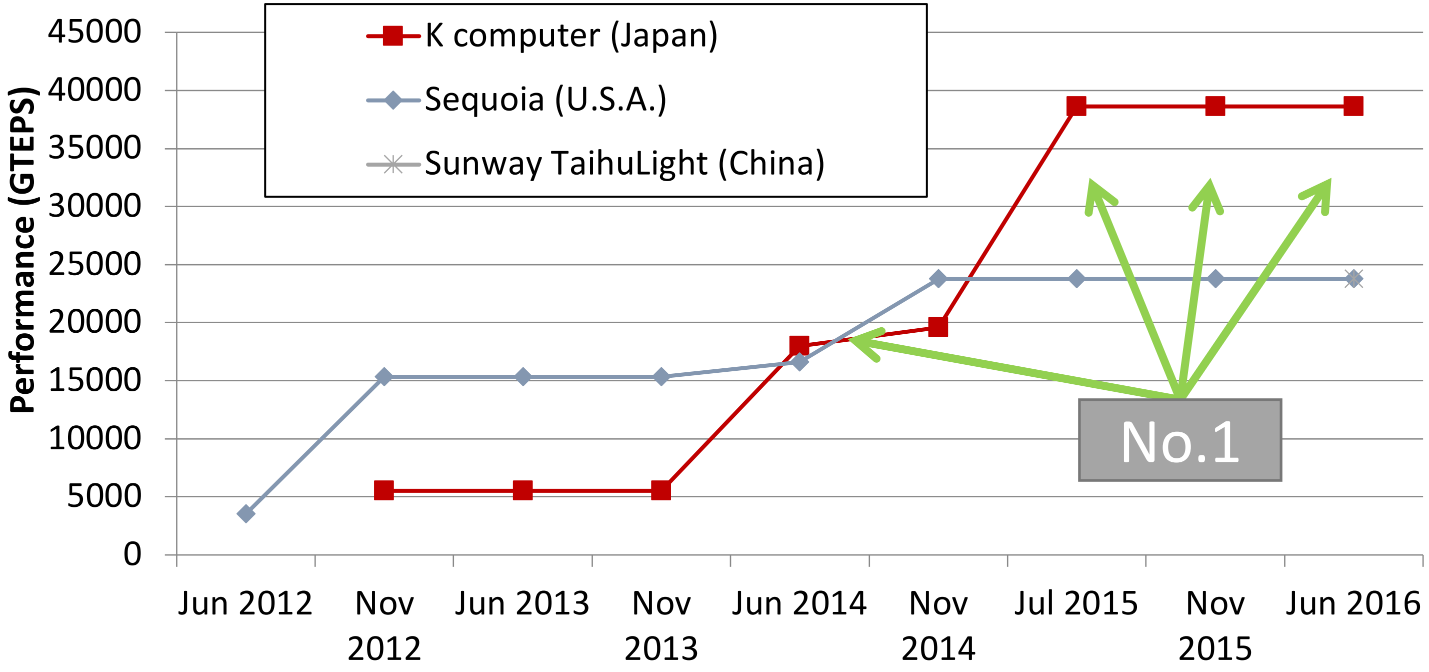
LLNL-IBM Sequoia
1.6 million CPUs
1.6 Petabyte mem

TaihuLight
10 million CPUs
1.3 Petabyte mem

BYTES, not FLOPS!

K-computer No.1 on Graph500: 5 Consecutive Times

- 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)

Modern AI is enabled by Supercomputing

- 25 years of AI winter after failure of symbolic logic based methods (e.g., Prolog, ICOT) -> resurrection by DNN, basic algorithms in the 1980s but too expensive -> HPC made machines 10 million times faster in 30 years -> expensive training now possible
- Recent trends require more supercomputing power
  - Deeper, more complex networks (Capsule Networks)
  - Complex, multidimensional data (e.g., 3-D Hi-Res images)
  - Increasing training sets (incl. GANs)
  - Coupling with high-fidelity simulations
  - Etc.

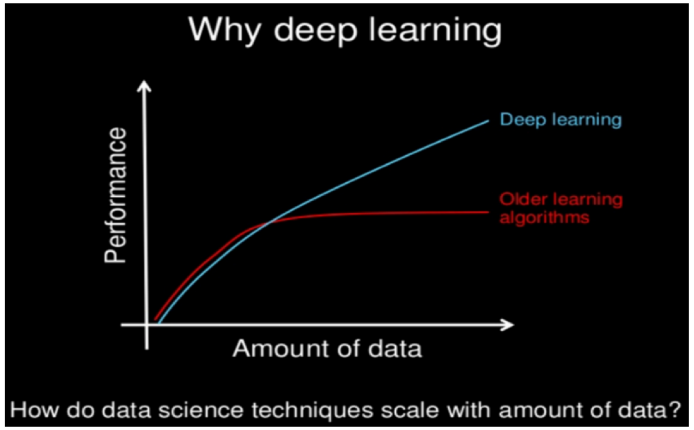
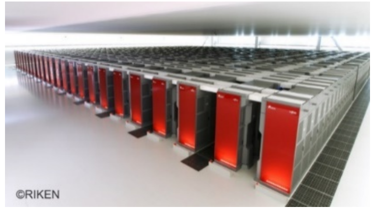
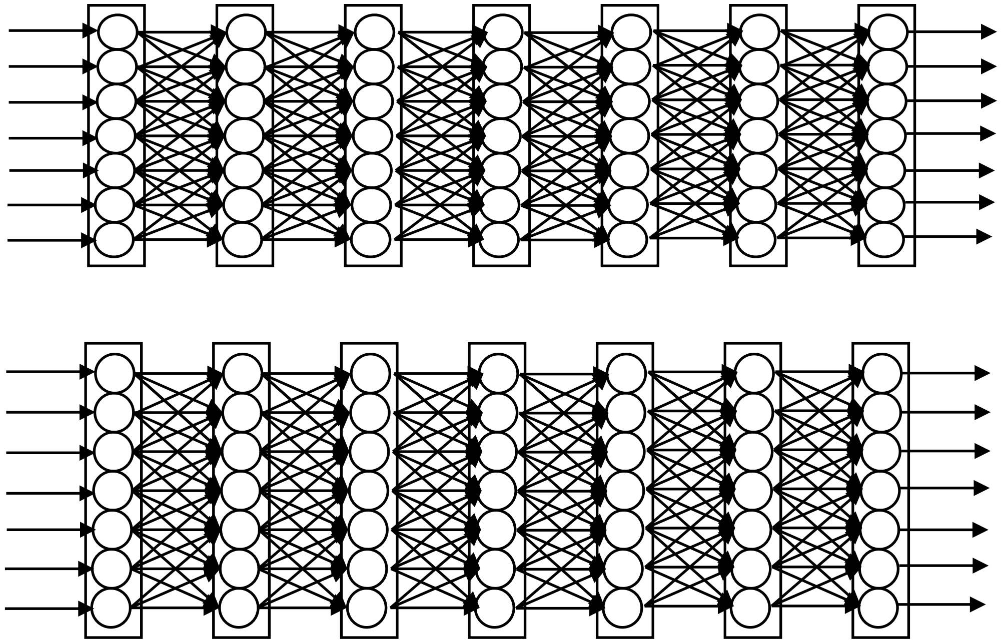


Fig. 2: Andrew Ng (Baidu) "What Data Scientists Should Know about Deep Learning"



4 Layers of Parallelism in DNN Training well supported in Post-K

- Hyper Parameter Search
  - Searching optimal network configs & parameters
  - Parallel search, massive parallelism required
- Data Parallelism
  - Copy the network to compute nodes, feed different batch data, average => network reduction bound
  - TOFU: Extremely strong reduction, x6 EDR Infiniband
- Model Parallelism (domain decomposition)
  - Split and parallelize the layer calculations in propagation
  - Low latency required (bad for GPU) -> strong latency tolerant cores + low latency TOFU network
- Intra-Chip ILP, Vector and other low level Parallelism
  - Parallelize the convolution operations etc.
  - SVE FP16+INT8 vectorization support + extremely high memory bandwidth w/HBM2
- Post-K could become world's biggest & fastest platform for DNN training!



Deep Learning is "All about Scale" Massive Parallelization is the key

- **Data-parallel training with (Asynchronous) Stochastic Gradient Descent**
  - Replicate network to all the nodes, feed different data, average the gradients periodically
  - Network All-Reduce Reduction in Megabytes~Gigabytes becomes the bottleneck at scale
  - NVIDIA: NVLink Hardware + NICL library (up to 8 GPUs on DGX-1, 16 on DGX-2 w/ NVL Switch)

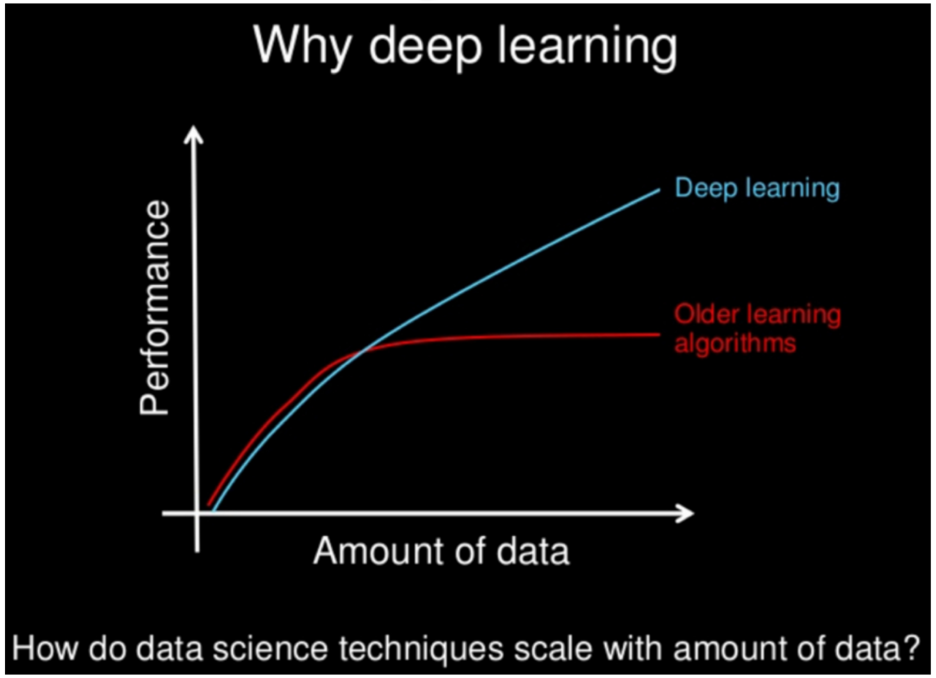


Fig. 2: Andrew Ng (Baidu) "What Data Scientists Should Know about Deep Learning"

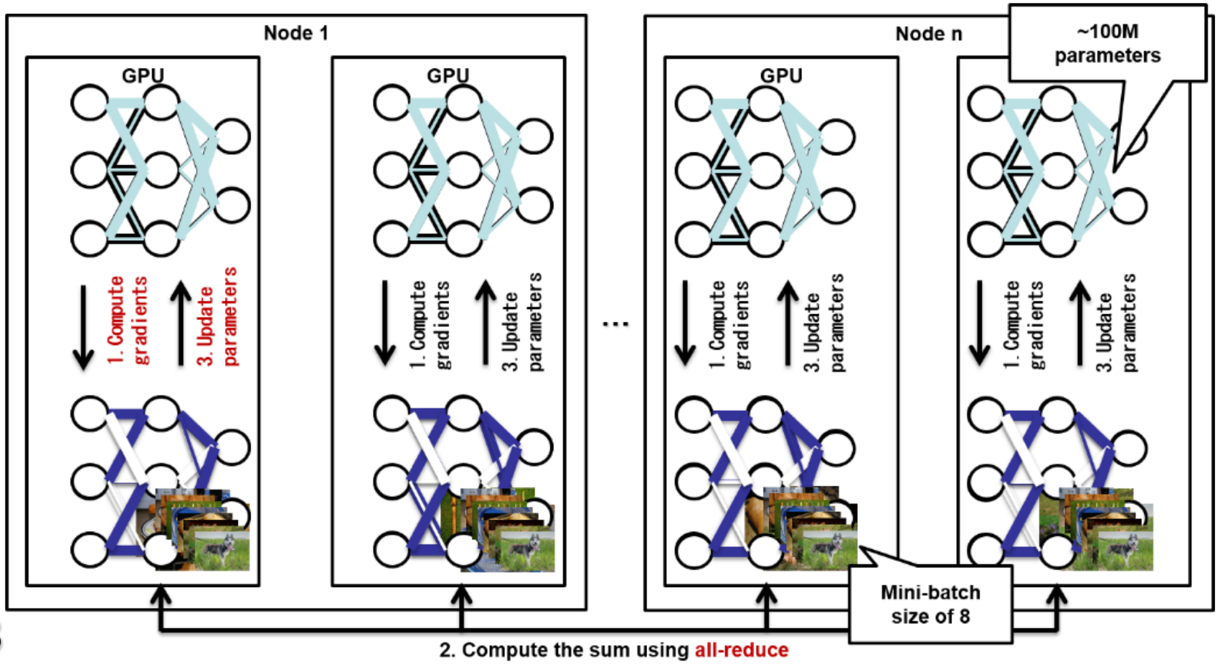


Fig. 3: Simplified DL workflow with ASGD per iteration:
1. Compute gradient
2. Exchange gradients via all-reduce; and
3. Update network parameters

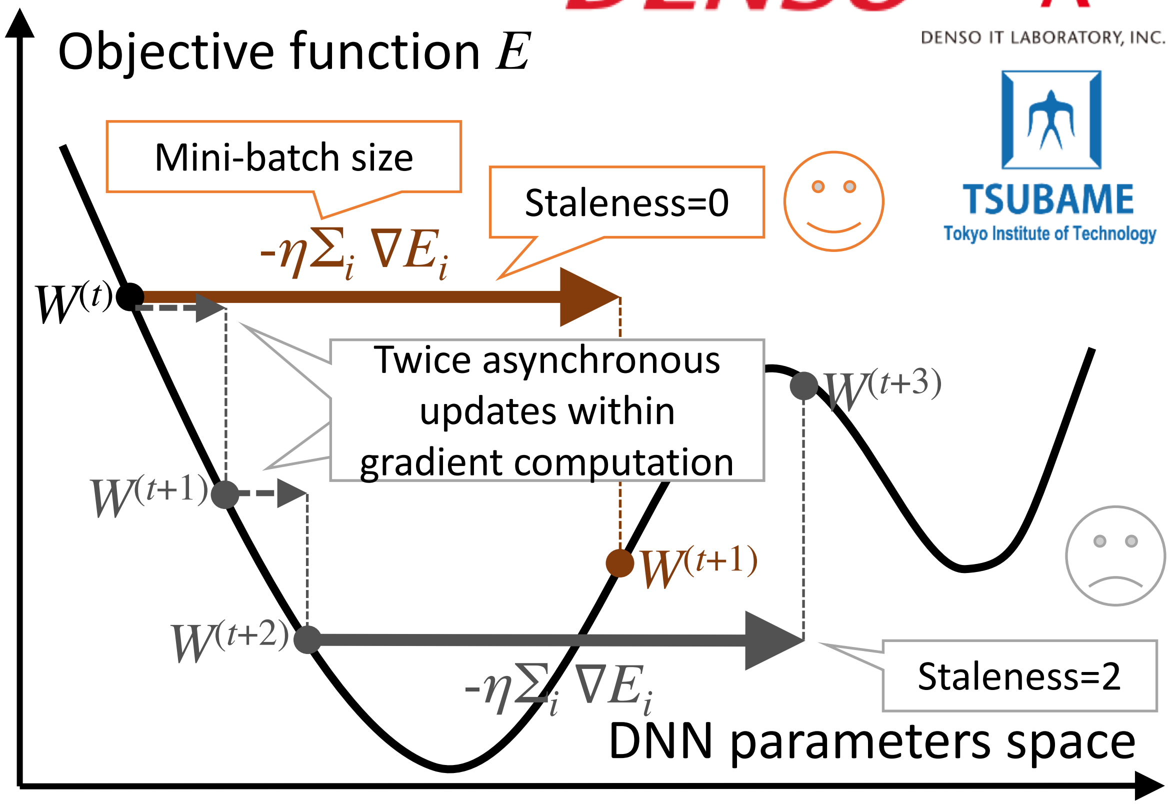
Example AI Research: Predicting Statistics of Asynchronous SGD Parameters for a Large-Scale Distributed Deep Learning System on GPU Supercomputers

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

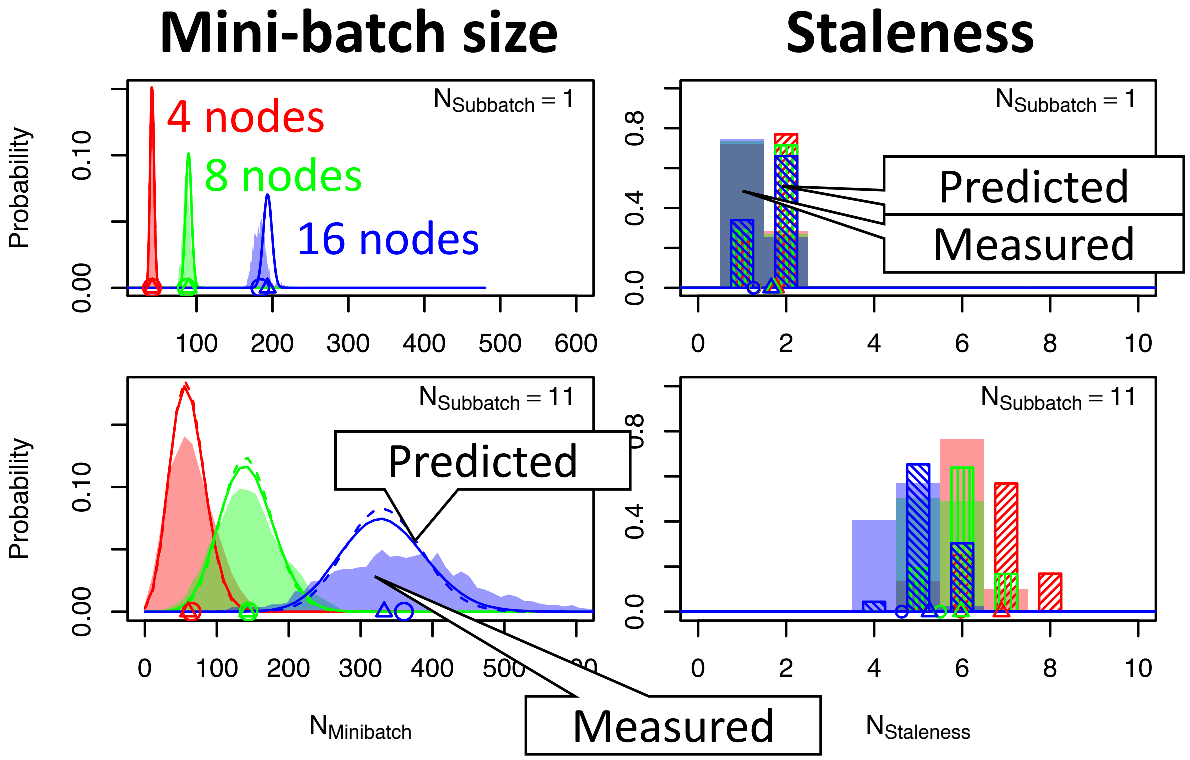
DENSO IT LABORATORY, INC.

TSUBAME
Tokyo Institute of Technology



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

Interconnect Performance as important as GPU Performance to accelerate DL

- **ASGD DL system SPRINT (by DENSO IT Lab) and DL speedup prediction with performance model**

$$T_{Epoch} = \frac{N_{File} \times T_{GPU}}{N_{Node} \times N_{GPU} \times N_{Subbatch}}$$

  - Data measured on T2 and KFC (both FDR) fitted to formulas
  - Allreduce time ($\in T_{GPU}$) dep. on #nodes and #DL_parameters

$$T_{Barrier} + (\alpha \log_2(N_{Node}) + \beta) \times N_{Param}$$

The Optimal Predicted Configurations of CNN-A on TSUBAME-KFC/DL

	N_{Node}	$N_{Subbatch}$	Average mini-batch size	Epoch time[s]	Speedup
Baseline	8	8	165.1	1779	-
FP16	7	22	170.1	1462	**1.22**
EDR IB	12	11	166.6	1245	**1.43**
FP16 + EDR IB	8	15	171.5	1128	**1.58**

Fig. 4: Oyama et al. "Predicting Statistics of Asynchronous SGD Parameters for a Large-Scale Distributed Deep Learning System on GPU Supercomputers

- **Other approaches == similar improvements:**
  - Cuda-Aware CNTK optimizes communication pipeline ➔ 15%—23% speedup **(Banerjee et al. "Re-designing CNTK Deep Learning Framework on Modern GPU Enabled Clusters")**
  - Reduced precision (FP[16|8|1]) to minimize msg. size w/ no or minor accuracy loss

Massive Scale Deep Learning on Post-K

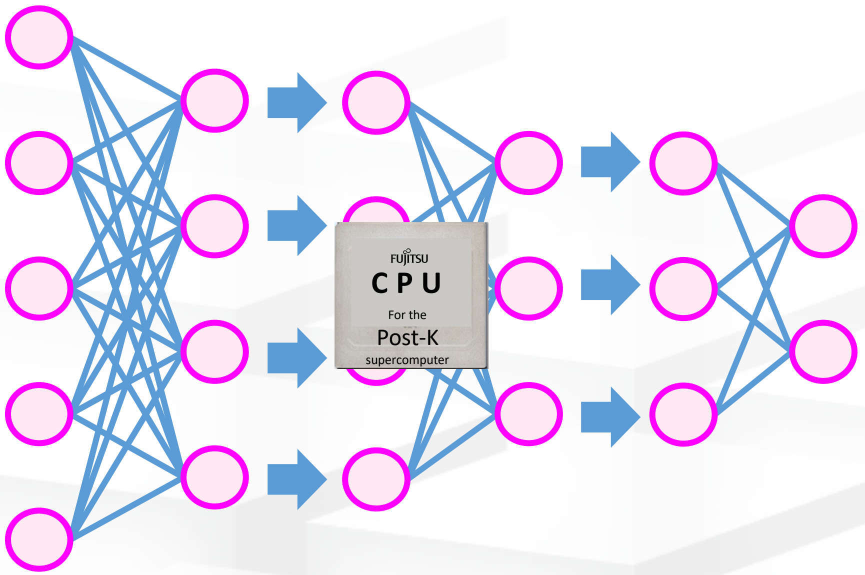
Post-K Processor

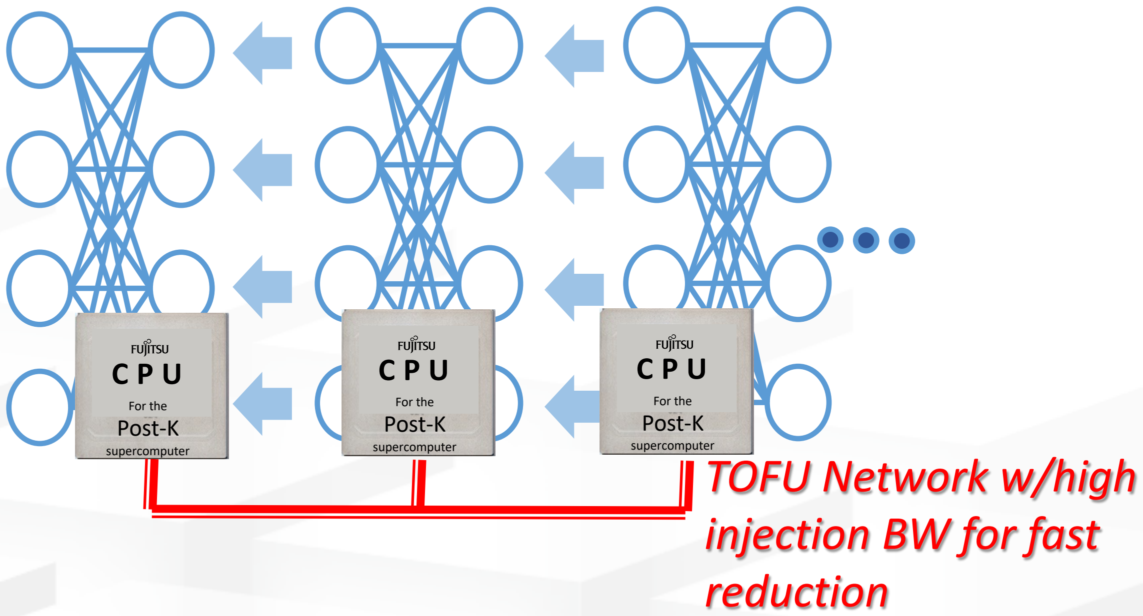
- ◆High perf FP16&Int8
- ◆**High mem BW for convolution**
- ◆**Built-in scalable Tofu network**

Unprecedened DL scalability

High Performance and Ultra-Scalable Network for massive scaling model & data parallelism

High Performance DNN Convolution





Low Precision ALU + High Memory Bandwidth + Advanced Combining of Convolution Algorithms (FFT+Winograd+GEMM)

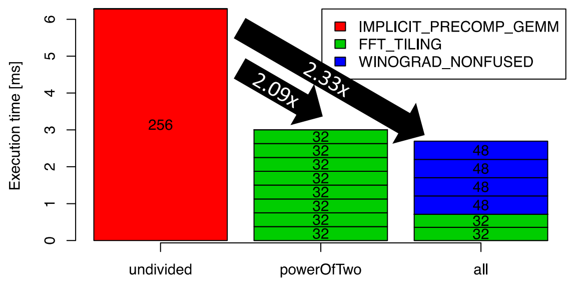
Unprecedented Scalability of Data/

Selecting the Optimal Convolution Kernel

- **NEW! Micro Batching: Tokyo Tech. and ETH [Oyama, Tan, Hoefler & Matsuoka]**
  - Use the “micro-batch” technique to select the best convolution kernel
    - Direct, GEMM, FFT, Winograd
    - Optimize both speed and memory size
  - On high-end GPUs, in many cases Winograd or FFT chosen over GEMM
    - They are faster but use more memory
  - Currently implemented as cuDNN wrapper, applicable to all frameworks
  - For Post-K, (1) Winograd/FFT are selected more often, and (2) performance will be similar to GPUs in such cases

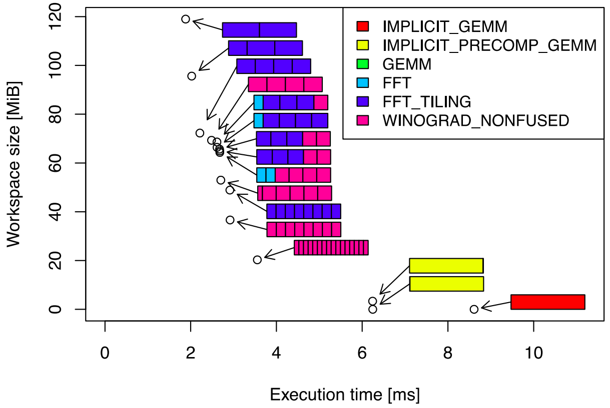
Evaluation: WR using Dynamic Programming

- μ-cuDNN achieved **2.33x** speedup on forward convolution of AlexNet conv2



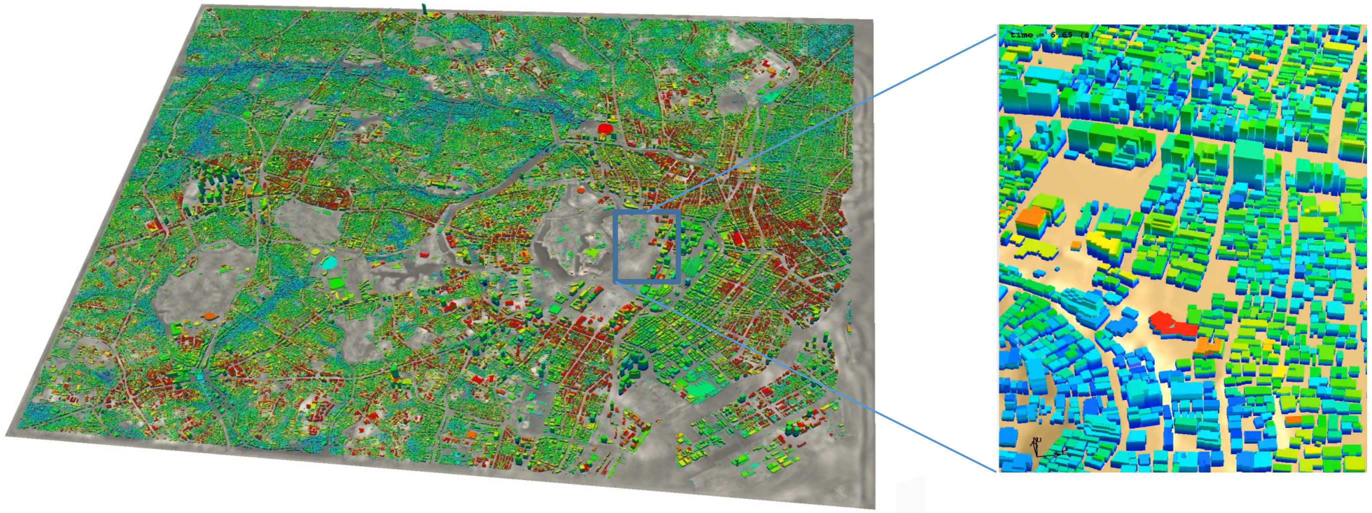
`cudnnConvolutionForward` of AlexNet conv2 on NVIDIA Tesla P100-SXM2
Workspace size of 64 MiB, mini-batch size of 256
Numbers on each rectangles represent micro-batch sizes

Evaluation: WD using Integer LP

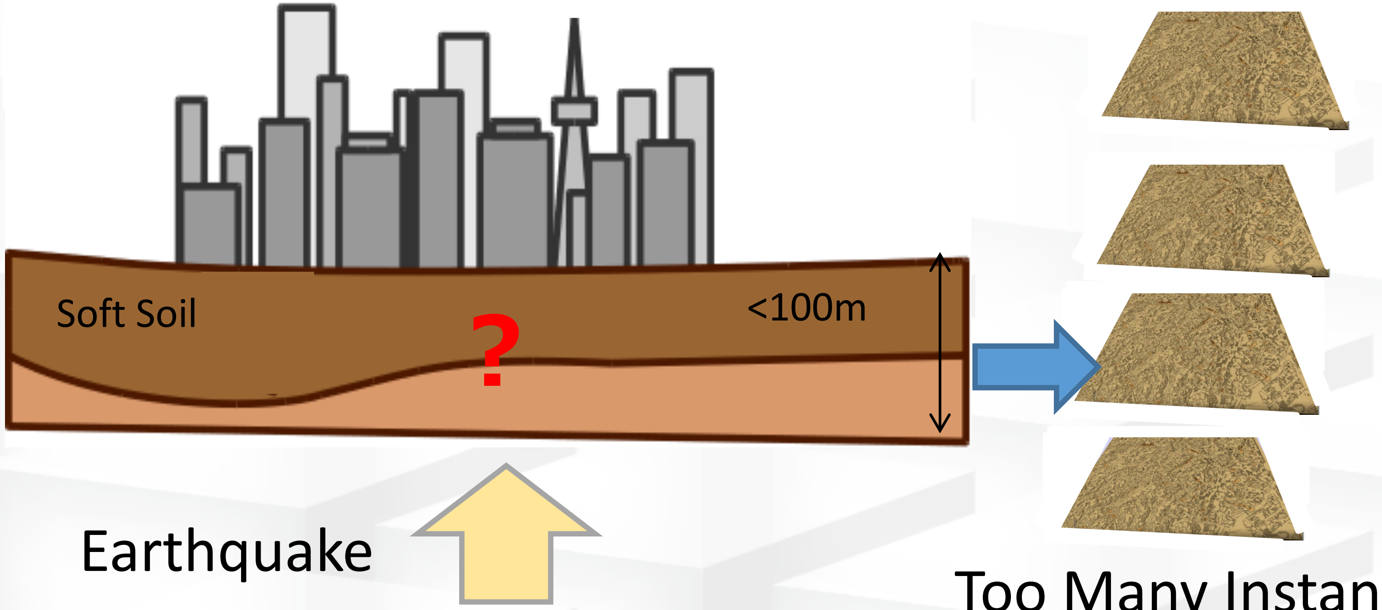


A desirable configuration set of AlexNet conv2 (Forward)
Mini-batch size of 256, P100-SXM2
Each bar represents proportion of micro-batch sizes and algorithms

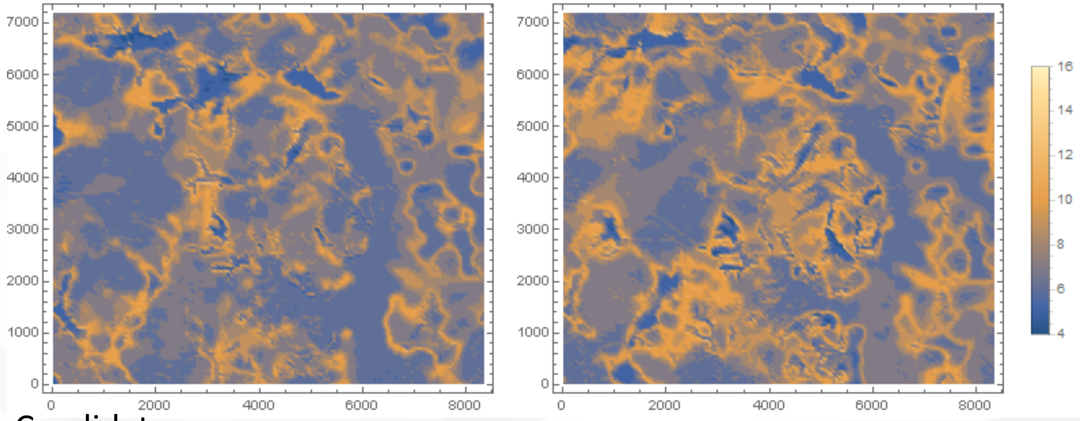
Large Scale simulation and AI coming together
[Ichimura et. al. Univ. of Tokyo, IEEE/ACM SC17 Best Poster]



130 billion freedom earthquake of entire Tokyo on K-Computer (ACM Gordon Bell Prize Finalist, SC16,17 Best Poster)



Too Many Instances

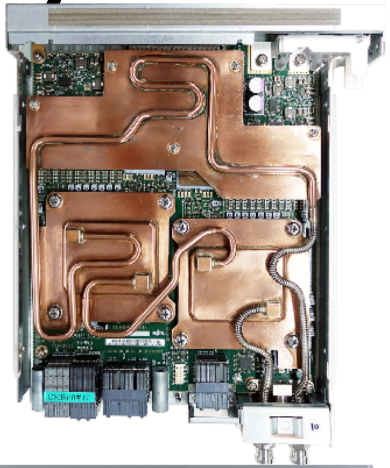


Candidate Underground Structure 1

Candidate Underground Structure 2

AI Trained by Simulation to generate candidate soft soil structure

Post-K CPU New Innnovations: Summary



1. **Ultra high bandwidth using on-package memory & matching CPU core**
   - Recent studies show that majority of apps are memory bound, some compute bound but can use lower precision e.g. FP16
   - Comparison w/mainstream CPU: much faster FPU, almost order magnitude faster memory BW, and ultra high performance accordingly
   - Memory controller to sustain massive on package memory (OPM) BW: difficult for coherent memory CPU, first CPU in the world to support OPM
2. **Very Green e.g. extreme power efficiency**
   - Power optimized design, clock gating & power knob, efficient cooling
   - Power efficiency much better than CPUs, comparable to GPU systems
3. **Arm Global Ecosystem & SVE contribution**
   - Annual processor production: x86 3-400mil, ARM 21bil, (2~3 bil high end)
   - Rapid upbringing HPC&IDC Ecosystem (e.g. Cavium, HPE, Sandia, Bristol,…)
   - SVE(Scalable Vector Extension) -> Arm-Fujitsu co-design, future global std.
3. **High Performance on Society5.0 apps including AI**
   - Next gen AI/ML requires massive speedup => high perf chips + HPC massive scalability across chips
   - Post-K processor: support for AI/ML acceleration e.g. Int8/FP16+fast memory for GPU-class convolution, fast interconnect for massive scaling
   - Top performance in AI as well as other Society 5.0 apps