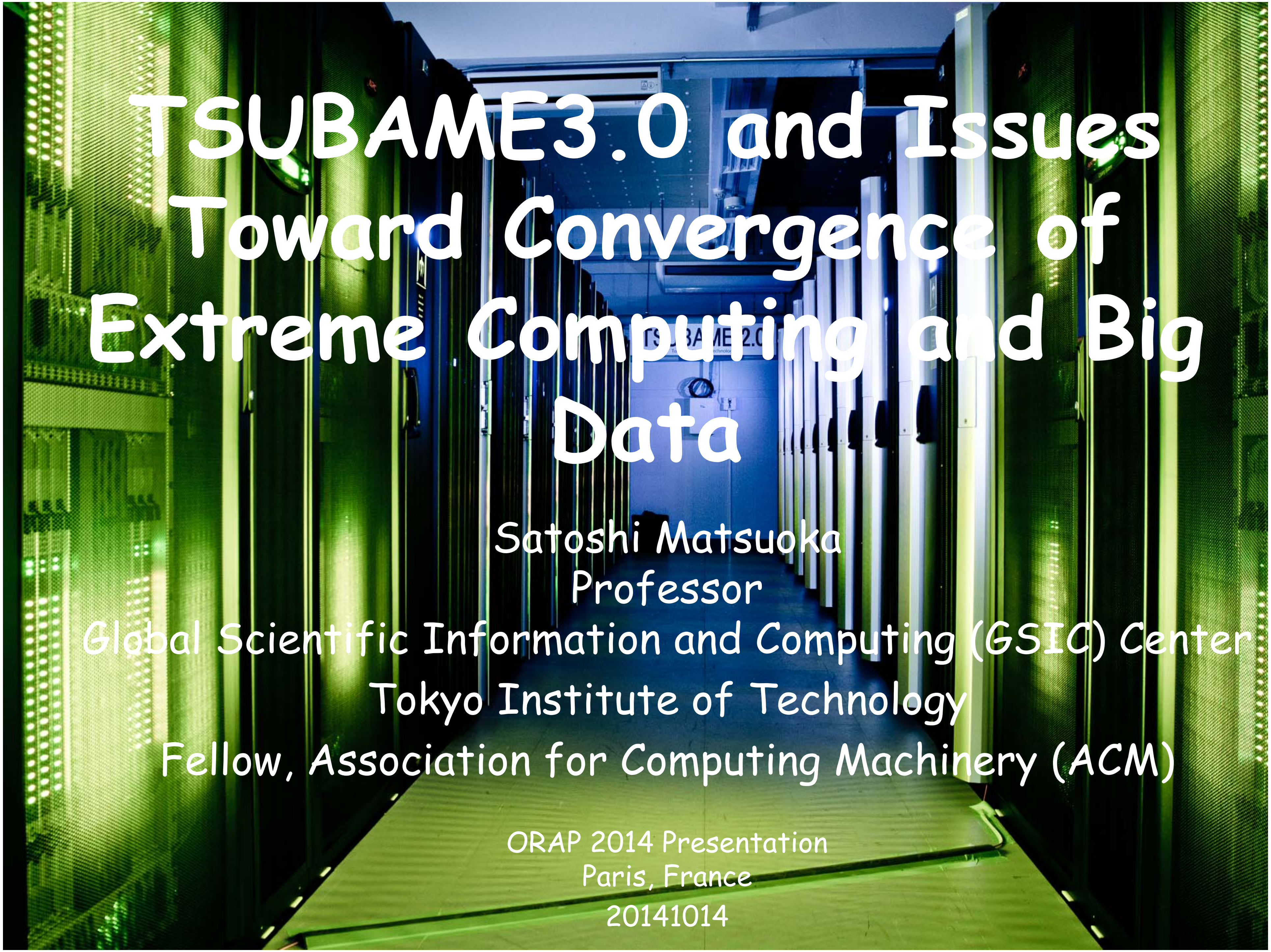


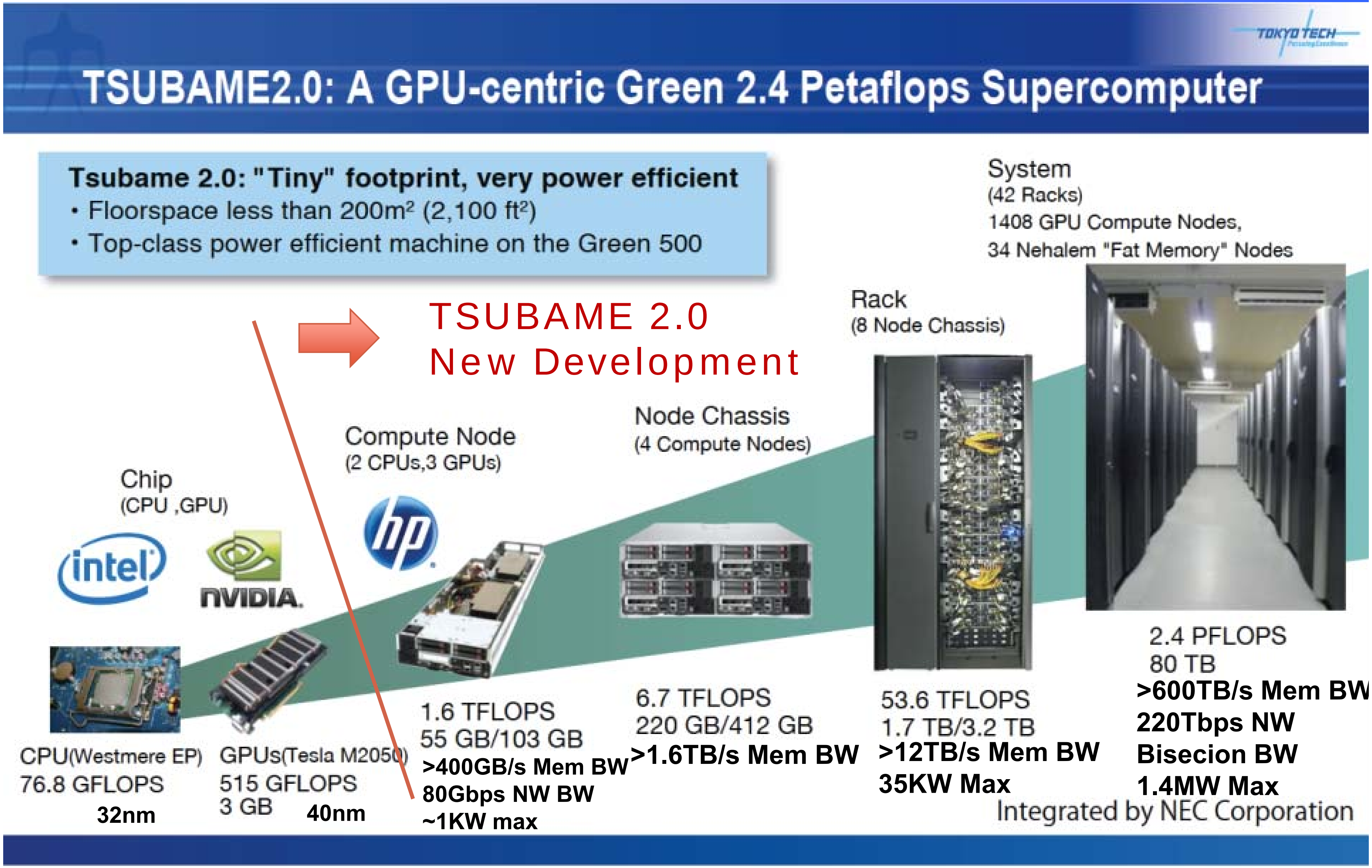
TSUBAME3.0 and Issues Toward Convergence of Extreme Computing and Big Data

Satoshi Matsuoka
Professor
Global Scientific Information and Computing (GSIC) Center
Tokyo Institute of Technology
Fellow, Association for Computing Machinery (ACM)

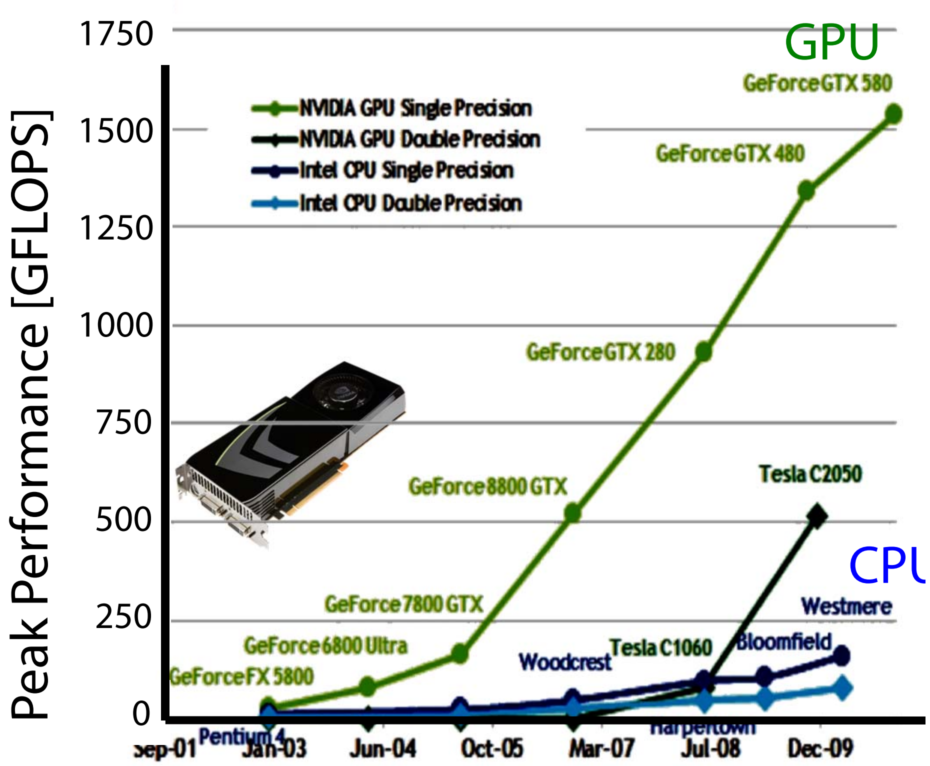
ORAP 2014 Presentation
Paris, France
20141014

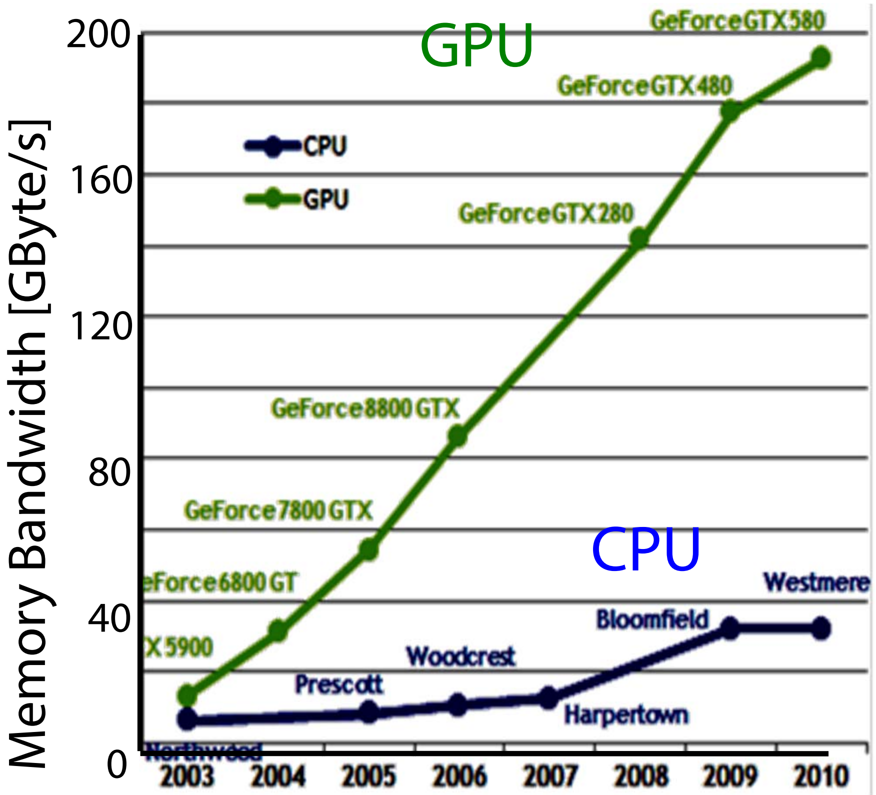
TSUBAME2.0 Nov. 1, 2010

“The Greenest Production Supercomputer in the World”



Performance Comparison of CPU vs. GPU

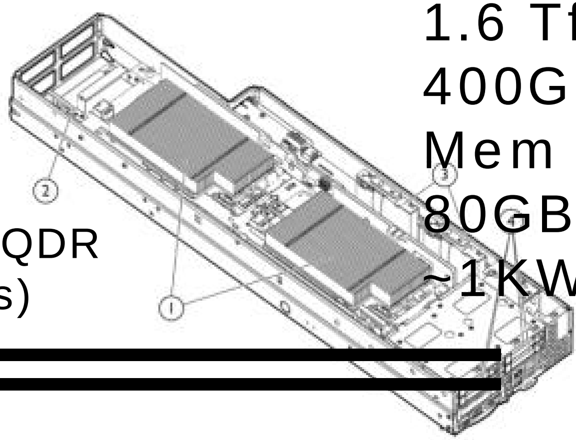




x5-6 socket-to-socket advantage in both compute and memory bandwidth,
Same power
(200W GPU vs. 200W CPU+memory+NW+...)

TSUBAME2.0 Compute Node

Thin Node



1.6 Tflops
400GB/s
Mem BW
80GBps NW
~1KW max

Infiniband QDR
x2 (80Gbps)

HP SL390G7 (Developed for TSUBAME 2.0)

GPU: NVIDIA Fermi M2050 x 3
515GFlops, 3GByte memory /GPU
CPU: Intel Westmere-EP 2.93GHz x2 (12cores/node)
Multi I/O chips, 72 PCI-e (16 x 4 + 4 x 2) lanes --- 3GPUs + 2 IB QDR
Memory: 54, 96 GB DDR3-1333
SSD:60GBx2, 120GBx2

Total Perf
2.4PFlops
Mem: ~100TB
SSD: ~200TB

TSUBAME2.0 Very Green…

"Greenest Production Supercomputer in the World"

the Green 500 (#3 overall)

Nov. 2010, June 2011

(#4 Top500 Nov. 2010)

<<

3 times more power efficient than a laptop!

TSUBAME Wins Awards…

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

Scott Lathrop
SC11 Conference Chair

Thom H. Dunning, Jr.
Gordon Bell Chair

ACM Gordon Bell Prize 2011
2.0 Petaflops Dendrite Simulation

Special Achievements in Scalability and Time-to-Solution

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

TSUBAME Three Key Application Areas

"Of High National Interest and Societal Benefit to the Japanese Taxpayers"

1. Safety/Disaster & Environment
2. Medical & Pharmaceutical
3. Manufacturing & Materials

Plus

Co-Design for general IT Industry and Ecosystem impact (IDC, Big Data, etc.)

Lattice-Boltzmann-LES with Coherent-structure SGS model [Onodera&Aoki2013]

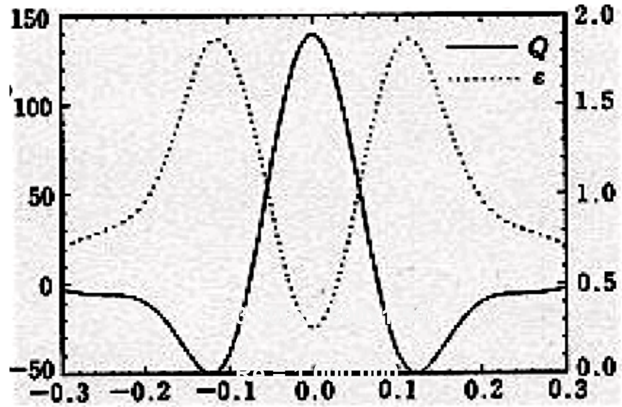
Coherent-structure Smagorinsky model

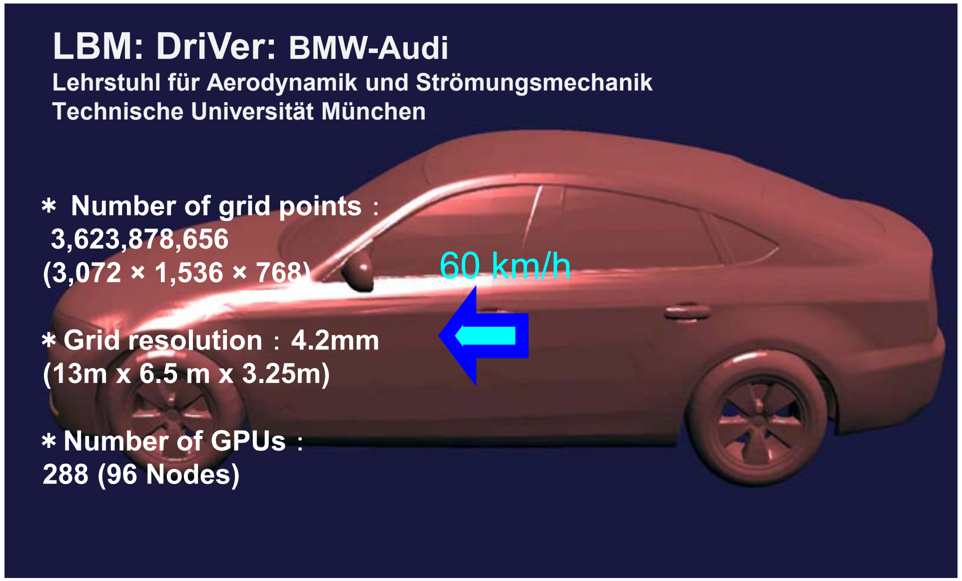
The model parameter is locally determined by the second invariant of the velocity gradient tensor.

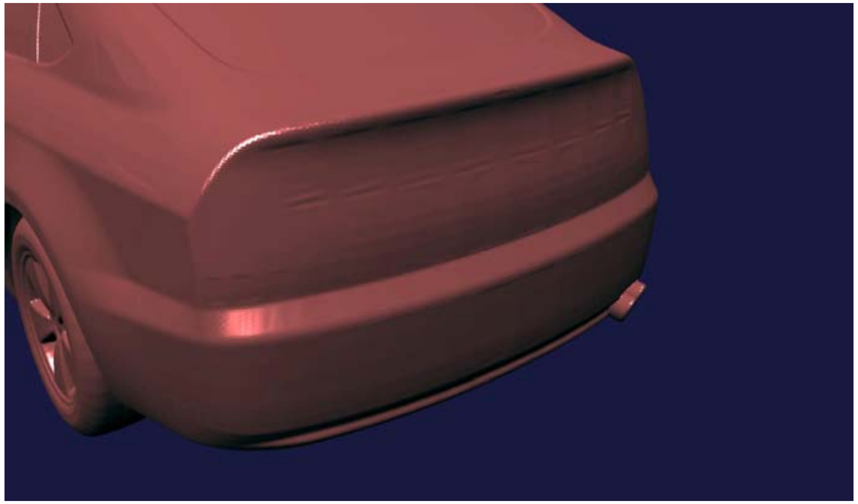
$$\nu_{SGS} = C\Delta^2|S| \qquad C = C_1|F_{CS}|^{3/2} \qquad F_{CS} = \frac{Q}{E}$$

◎ Turbulent flow around a complex object
◎ Large-scale parallel computation

Second invariant of the velocity gradient tensor(Q) and Energy dissipation(ε)







Computational Area – Entire Downtown Tokyo

Major part of Tokyo

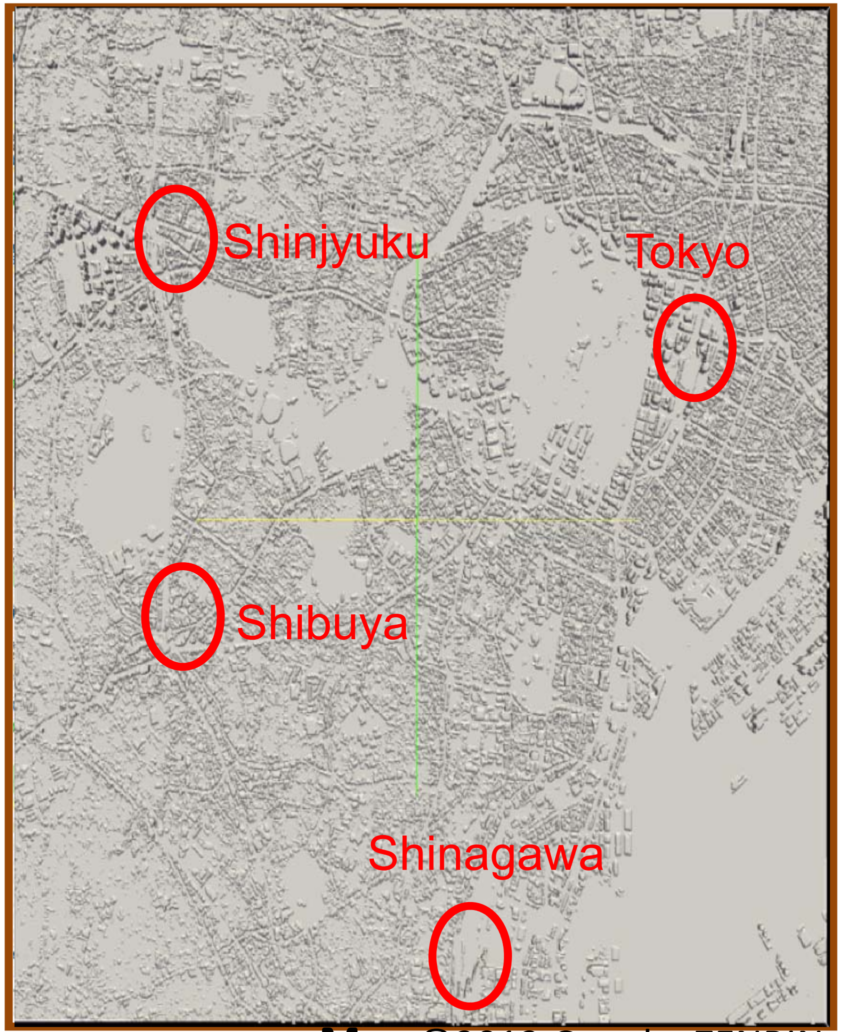
Including Shnjuku-ku, Chiyoda-ku, Minato-ku, Meguro-ku, Chuou-ku,

10km×10km

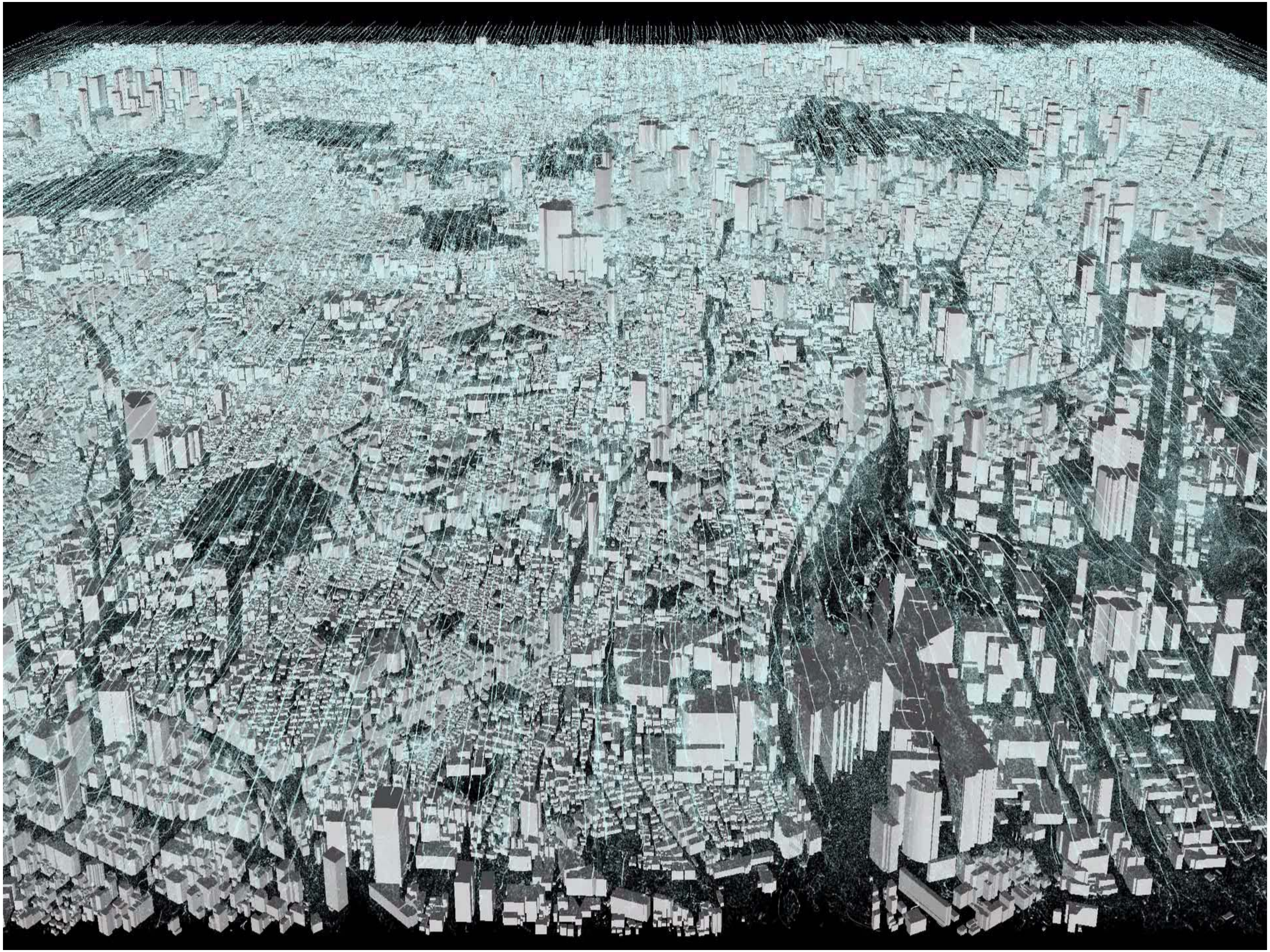
Building Data:

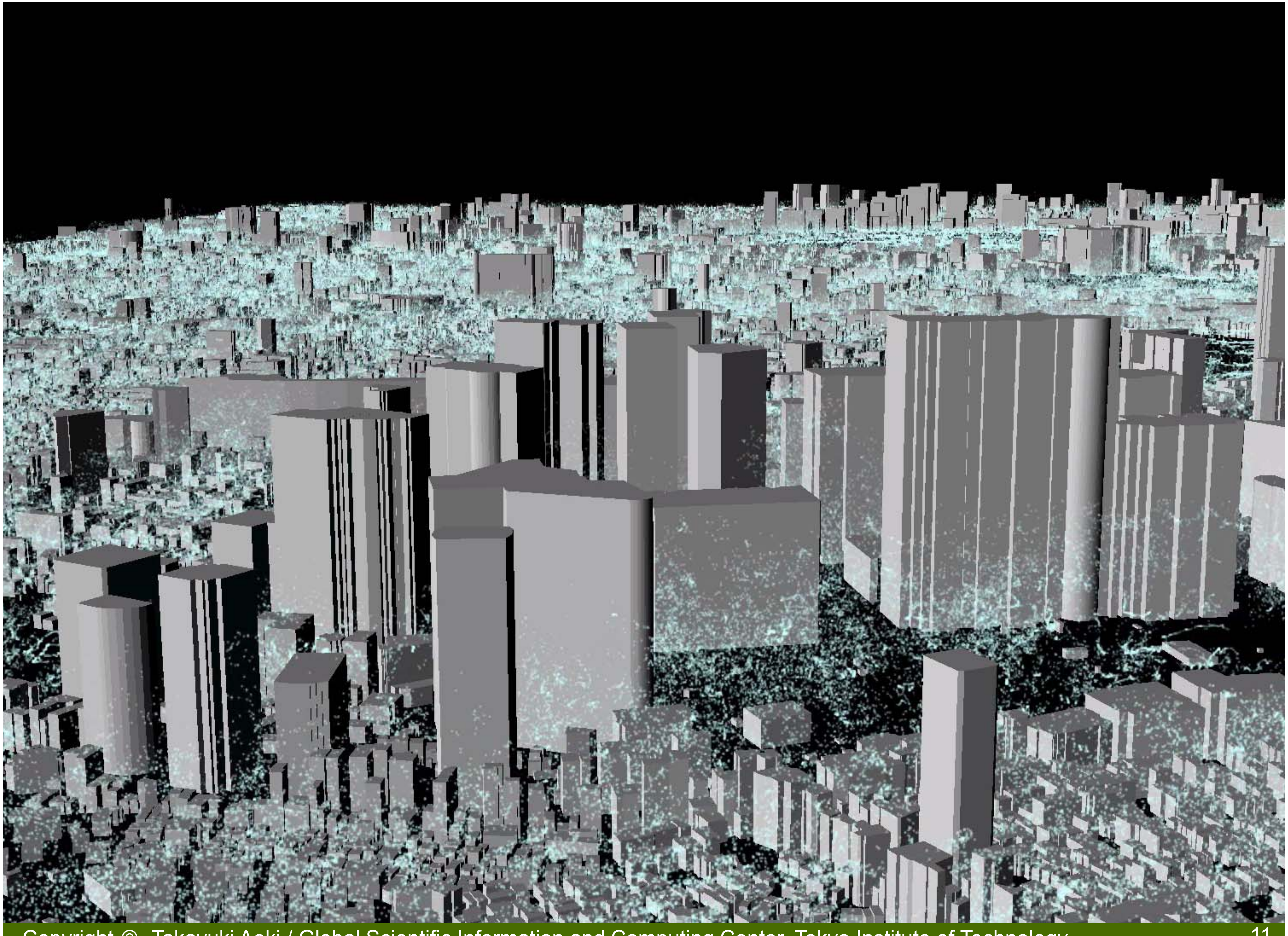
Pasco Co. Ltd.
TDM 3D

Achieved 0.592 Petaflops using over 4000 GPUs (15% efficiency)



Map ©2012 Google, ZENRIN





アステラス製薬とのデング熱等の熱帯病の特効薬の創薬

HOME SERVICES NEWS EDUCATION ABOUT US Search Log In Sign Up

Business Wire
A Berkshire Hathaway Company

astellas
Leading Light for Life

TOKYO TECH
Pursuing Excellence

いいね！ Tweet 3 Share 0 Pin it

:h 21, 2013 03:09 AM Eastern Daylight Time

Release Versions

- English
- Chinese
- EON: Enhanced Online News

Company Information Center

ASTELLAS PHARMA INC.

TOKYO:4503

Tokyo Institute of Technology and Astellas Launch Collaborative Research for New Anti-Dengue Virus Drugs for Neglected Tropical Diseases

- IT drug-discovery research through use of Tokyo Tech's Supercomputer TSUBAME2.0 -

TOKYO--(BUSINESS WIRE)--Tokyo Institute of Technology ("Tokyo Tech"; Tokyo, Japan; President: Yoshinao Mishima) and Astellas Pharma Inc. ("Astellas")(TOKYO:4503)(President and CEO: Yoshihiko Hatanaka) today announced that they have signed a joint research agreement for drug discovery research utilizing Tokyo Tech's TSUBAME2.0 supercomputer to efficiently discover candidates for the treatment of neglected tropical diseases ("NTDs") caused by dengue virus.

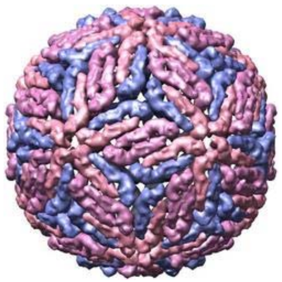
NTDs, prevalent mainly among the poor in tropical areas of developing countries, are infectious diseases spread by parasites or bacteria. As it is estimated that approximately one billion people are affected with NTDs worldwide, NTDs are a serious healthcare issue that is being addressed on a global scale. Among them, diseases caused by dengue virus, such as dengue fever/dengue hemorrhagic fever are with high unmet medical needs for treatment and development of new therapeutic drugs. There is no existing drug to treat dengue fever/dengue hemorrhagic fever in the market as well as under development, and the effectiveness of some vaccines to prevent dengue virus currently under development is unclear at this time.

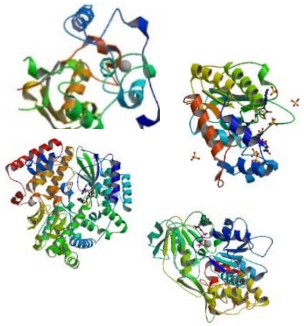
Under the collaborative agreement, Tokyo Tech which has cutting-edge computation technique, and Astellas will cooperate on an

Accelerate In-silico screeninig and data mining

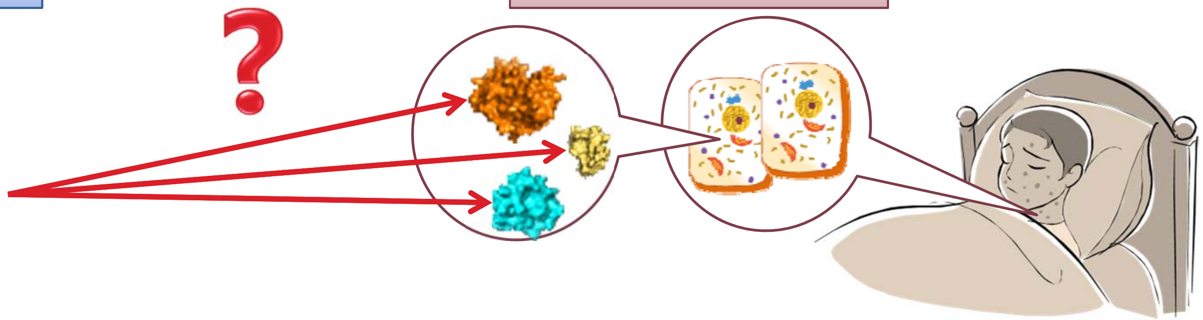
Discovery of Dengue-Human Interactome w/GPU Docking [Akiyama et. Al., Tokyo Tech]

Dengue virus enzymes

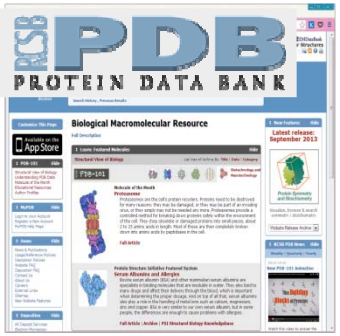




Human proteins



Protein name	Structure (PDB ID)
Protease	3U1I
Methyltransferase	1R6A
Polymerase	3VWS
Helicase	2JLR



Human protein structures were collected from the public database PDB using the following criteria:

- ✓ >25 residues
- ✓ X-ray resolution better than 3.25 Å
- ✓ No mutation

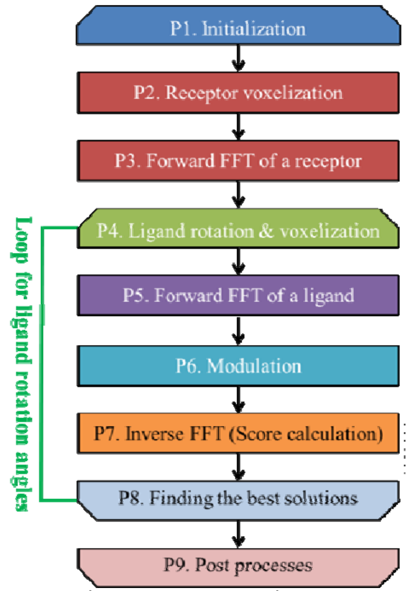
#Structures (PDB-chains)	30,544
#Proteins (UniProt IDs)	3,353

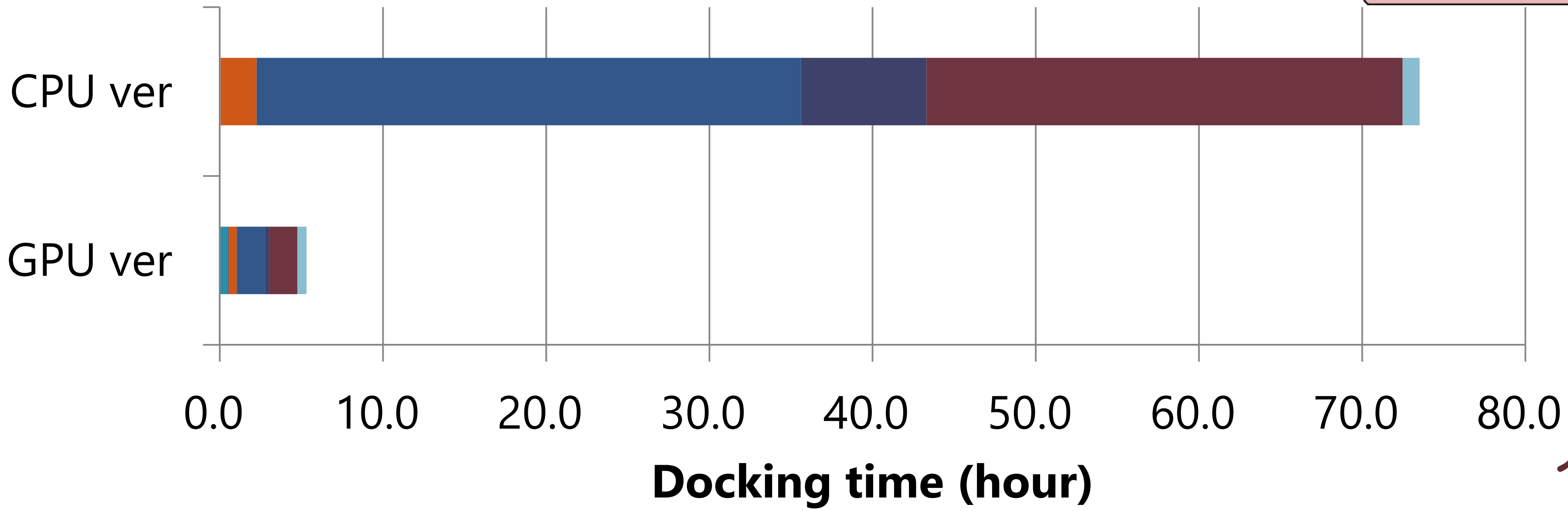
$4 \times 30{,}544 = 122{,}176$ dockings

June 15, 2013

Comparison of each process (1 CPU core vs. 1 CPU core and 1 GPU)

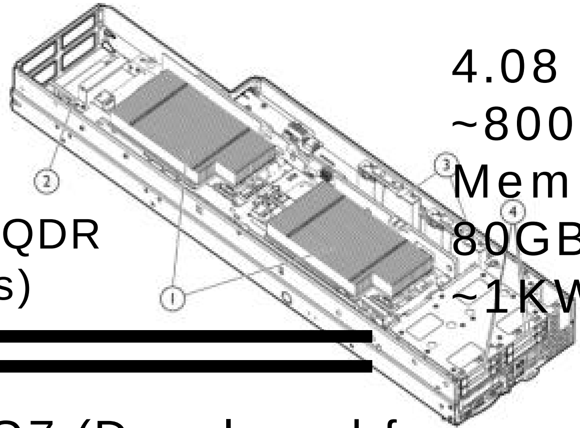
- **Comparison of CPU version and GPU version**
  1. FFT, Modulation: 20-30-fold faster
  2. Voxelization,
     Finding the best solutions: 2-6-fold faster
  3. Only initialization process slows down
     because of GPU initialization





TSUBAME2.0⇒2.5 Thin Node Upgrade

Thin Node

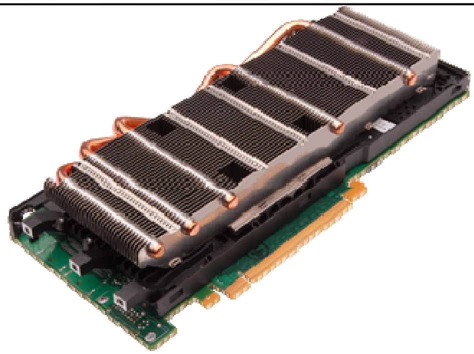


Peak Perf.
4.08 Tflops
~800GB/s Mem BW
80GBps NW
~1KW max

Infiniband QDR x2 (80Gbps)

HP SL390G7 (Developed for TSUBAME 2.0, Modified for 2.5)
GPU: NVIDIA Kepler K20X x 3
1310GFlops, 6GByte Mem(per GPU)
CPU: Intel Westmere-EP 2.93GHz x2
Multi I/O chips, 72 PCI-e (16 x 4 + 4 x 2) lanes --- 3GPUs + 2 IB QDR
Memory: 54, 96 GB DDR3-1333
SSD:60GBx2, 120GBx2

NVIDIA Fermi M2050 1039/515 GFlops



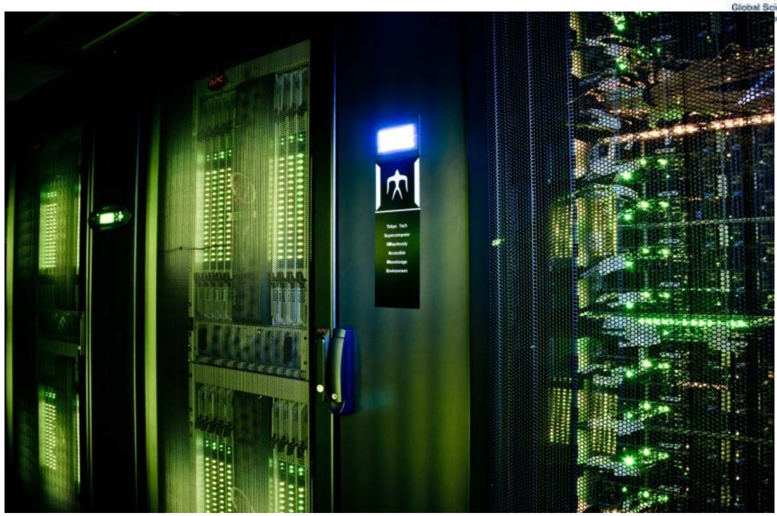
NVIDIA Kepler K20X 3950/1310 GFlops

Application	TSUBAME2.0 Performance	TSUBAME2.5 Performance	Boost Ratio
Top500/Linpack 4131 GPUs (PFlops)	1.192	2.843	2.39
Green500/Linpack 4131 GPUs (GFlops/W)	0.958	3.068	3.20
Semi-Definite Programming Nonlinear Optimization 4080 GPUs (PFlops)	1.019	1.713	1.68
Gordon Bell Dendrite Stencil 3968 GPUs (PFlops)	2.000	3.444	1.72
LBM LES Whole City Airflow 3968 GPUs (PFlops)	0.592	1.142	1.93
Amber 12 pmemd 4 nodes 8 GPUs (nsec/day)	3.44	11.39	3.31
GHOSTM Genome Homology Search 1 GPU (Sec)	19361	10785	1.80
MEGADOC Protein Docking 1 node 3GPUs (vs. 1CPU core)	37.11	83.49	2.25

2013: TSUBAME2.5 No.1 in Japan* in Single Precision FP, 17 Petaflops
(*but not in Linpack)

東京工業大学 Tokyo Institute of Technology

GSIC



17.1 Petaflops SFP
5.76 Petaflops DFP
$45mil / 6 years
(incl. power)

~=

北海道大学 / 東北大学 TOHOKU UNIVERSITY / 東京大学 THE UNIVERSITY OF TOKYO / 大阪大學 / 名古屋大学 / 九州大学 KYUSHU UNIVERSITY / 京都大学

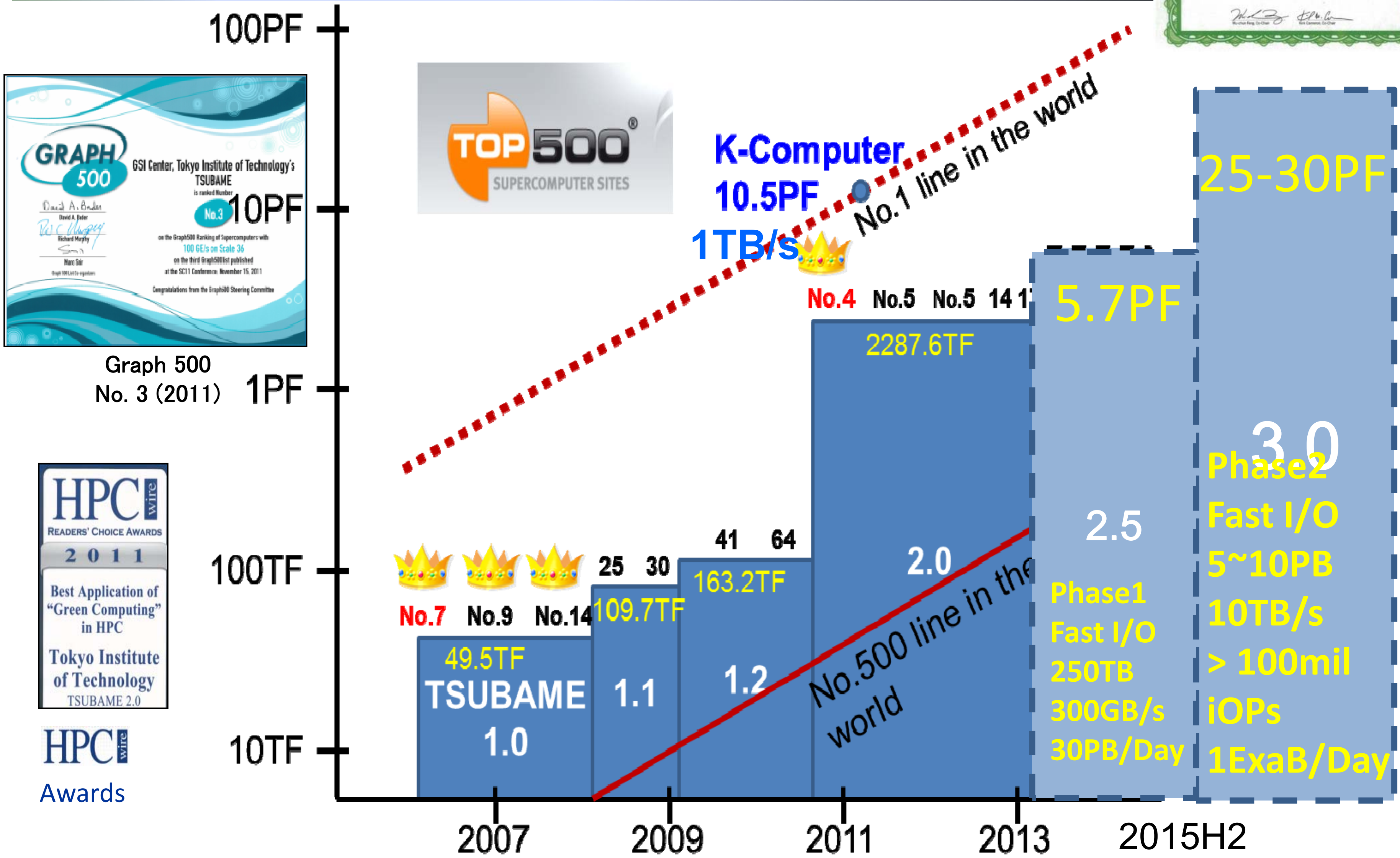
All University Centers COMBINED 9 Petaflops SFP

K Computer
11.4 Petaflops SFP/DFP
$1400 mil / 6years

Technological Comparisons (TSUBAME2 Deploying State-of-Art Tech.)

	TSUBAME2.5	BG/Q Sequoia	K Computer
Single Precision FP	17.1 Petaflops	20.1 Petaflops	11.3 Petaflops
Green500 (MFLOPS/W)	**3,068.71 (6th)**	**2,176.58 (26th)**	**830.18 (123rd)**
Standard Operational Power (inc. Cooling)	~0.8MW	5~6MW?	10~11MW
Hardware Architecture	Many-Core (GPU) + Multi-Core Hetero	Multi-Core Homo	Multi-Core Homo
Maximum HW Threads	> 1 Billion	~6 million	~700,000
Memory Technology	GDDR5+DDR3	DDR3	DDR3
Network Technology	Luxtera Silicon Photonics	Standard Optics	Copper
Non Volatile Memory / SSD	SSD Flash all nodes ~250TBytes	None	None
Power Management	Node/System Active Power Cap	Rack-level measurement only	Rack-level measurement only
Virtualization	KVM (G & V queues, Resource segragation)	None	None

TSUBAME Evolution
Towards Exascale and Extreme Big Data



Focused Research Towards Tsubame 3.0 and Beyond towards Exa

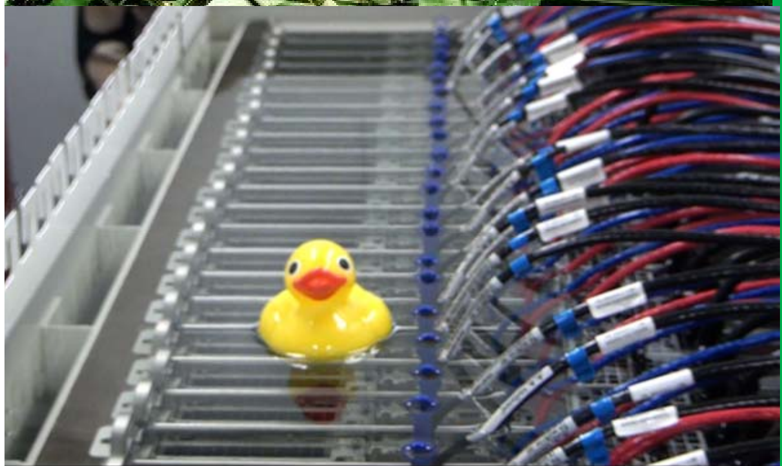
- **New memory systems** – Pushing the envelops of low Power vs. Capacity, Communication and Synchroniation Reducing Algorithms (CSRA)
- **Post Petascale Networks** – HW, Topology, Routing Algorithms, Placement…
- **Green Computing**: Ultra Power Efficient HPC
- **Scientific "Extreme" Big Data** – Ultra Fast I/O, Hadoop Acceleration, Large Graphs
- **Fault Tolerance** – Group-based Hierarchical Checkpointing, Fault Prediction, Hybrid Algorithms
- **Post Petascale Programming** – OpenACC and other many-core programming substrates, Task Parallel
- **Scalable Algorithms for Many Core** – Communication and Synchronization Reducing Algorithm (CSRA)

TSUBAME-KFC

by

GSIC, Tokyo Institute of Technology
NEC, NVIDIA, Green Revolution Cooling,
SUPERMICRO

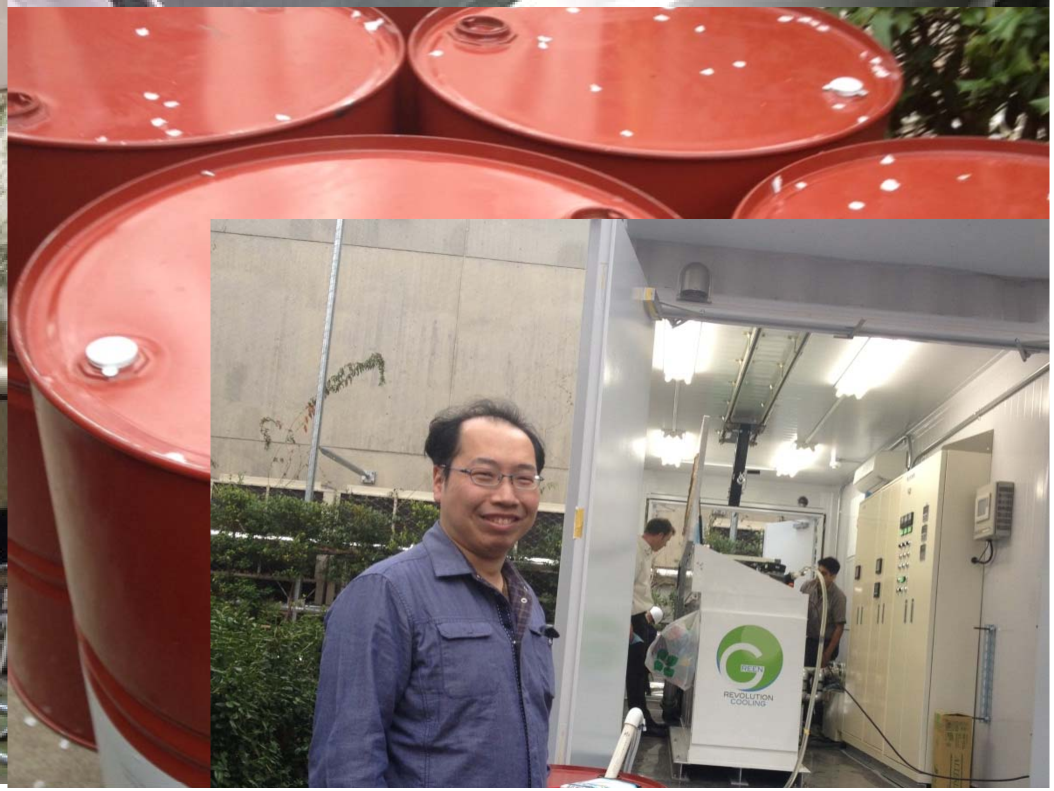
TSUBAME-KFC
(Kepler Fluid Cooling)



A TSUBAME3.0 prototype system
with advanced next gen cooling
40 compute nodes are oil-submerged
1200 liters of oil (Exxon PAO ~1 ton)
#1 Nov. 2013 Green 500!!

Single Node	5.26 TFLOPS DFP
System (40 nodes)	210.61 TFLOPS DFP 630TFlops SFP
Storage (3SSDs/node)	1.2TBytes SSDs/Node Total 50TBytes ~50GB/s BW





REVOLUTION
COOLING

Oil

ExxonMobil SpectraSyn Polyalphaolefins (PAO)

	4	6	8
Kinematic Viscosity@40C	19 cSt	31 cSt	48 cSt
Specific Gravity@15.6C	0.820	0.827	0.833
Flash point (Open Cup)	220 C	246 C	260 C
Pour point	-66 C	-57 C	-48 C

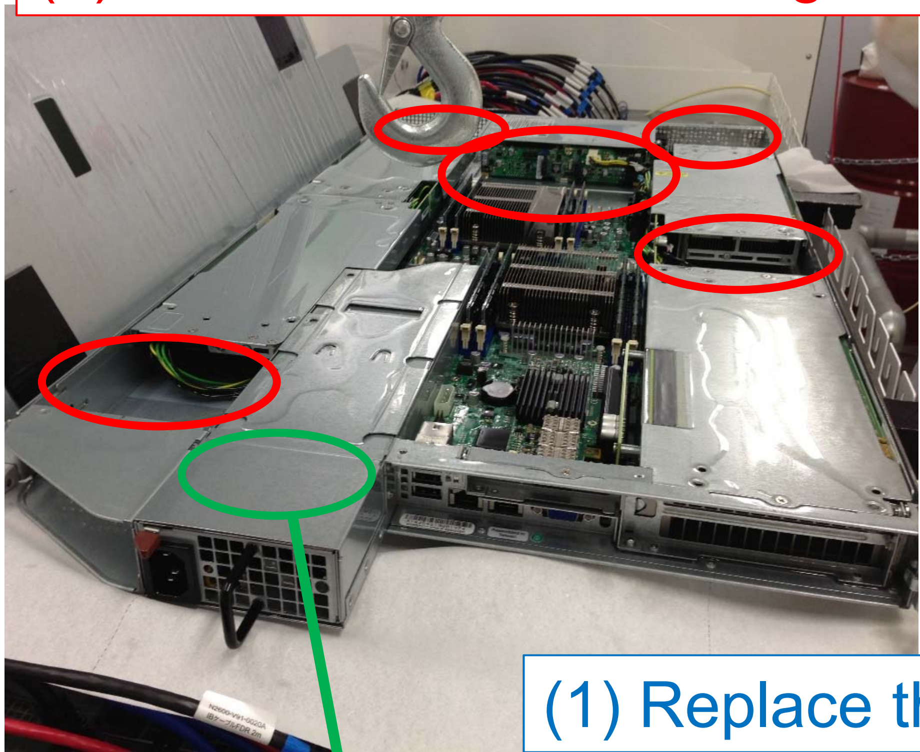
Fire Station at Den-en Chofu

Flash point of oil must be higher than 250 degrees C,
Otherwise it is a hazardous material under the Fire Defense Law in Japan.

Still the officer at the fire station requested us to follow the safety regulations of hazardous material: sufficient clearance around the oil, etc.

Compute Node

(2) Removed twelve cooling fans



NEC LX 1U-4GPU Server, 104Re-1G

- 2X Intel Xeon E5-2620 v2 Processor (Ivy Bridge EP, 2.1GHz, 6 core)
- 4X NVIDIA Tesla K20X GPU
- 1X Mellanox FDR InfiniBand HCA

CentOS 6.4 64bit
Intel Compiler, GCC,
CUDA 5.5
OpenMPI 1.7.2

(1) Replace thermal grease with thermal sheet

(3) Update firmware of power unit
to operate with cooling fan stopped.

Optimizations for efficiency

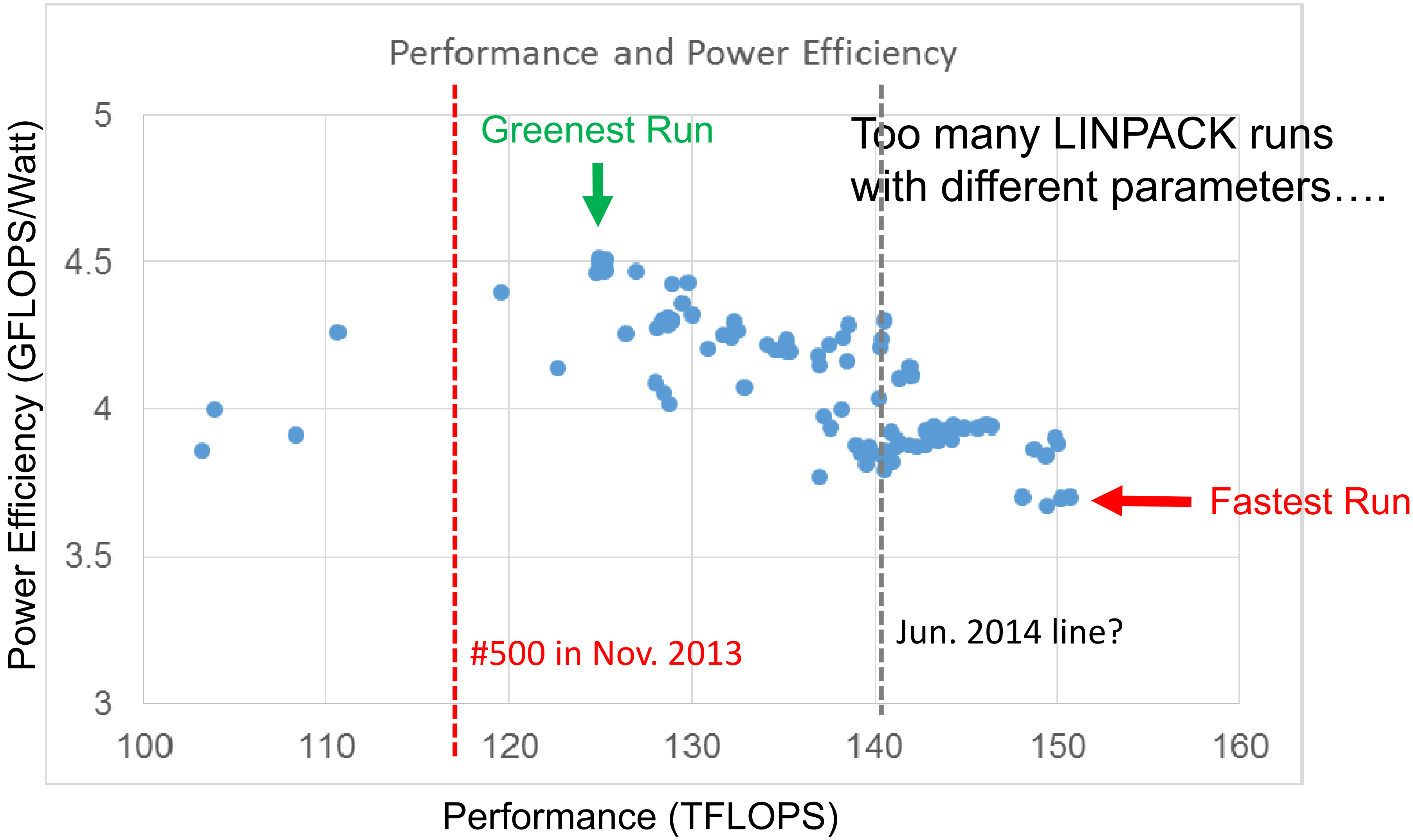
Lower performance leads higher efficiency

- Tuning for HPL parameters
  - Especially, block size (NB), and process grid (P&Q)
- Adjusting GPU clock and voltage
  - Available GPU clocks (MHz):
    614 (best), 640, 666, 705, 732 (default), 758, 784

and advantages of hardware configuration

- GPU:CPU ratio = 2:1
- Low power Ivy Bridge CPU (this also lower the perf.)
- Cooling system. No cooling fans. Low temperature.

Green500 submission



#1 in Green 500 List (Nov. 2013)

- 1st achievement as Japanese supercomputer
- #1 again in June 2014
- TSUBAME 2.5 is also ranked #6

The Green500 List

Listed below are the November 2013 The Green500's energy-efficient supercomputers ranked from 1 to 10.

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	4,503.17	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	27.78
2	3,631.86	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
5	3,130.95	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
6	3,068.71	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.930GHz, Infiniband QDR, NVIDIA K20x	922.54
7	2,702.16	University of Arizona	iDataPlex DX360M4, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR14, NVIDIA K20x	53.62
8	2,629.10	Max-Planck-Gesellschaft MPI/IPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	269.94
9	2,629.10	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	55.62
10	2,358.69	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C 2.000GHz, Infiniband FDR, Nvidia K20m	71.01

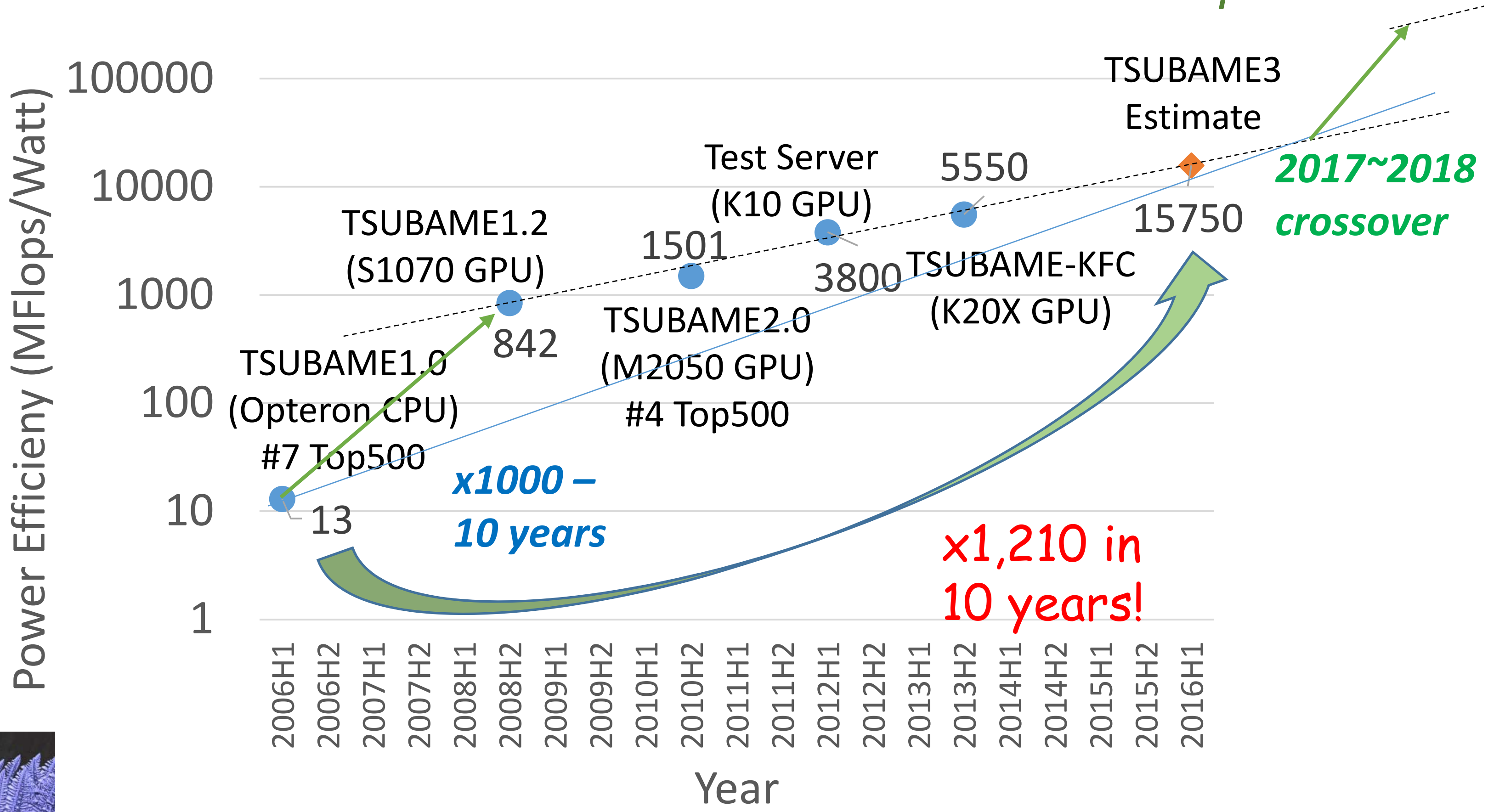
TSUBAME-KFC

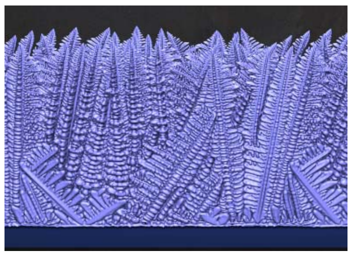
TSUBAME 2.5

* Performance data obtained from publicly available sources including TOP500

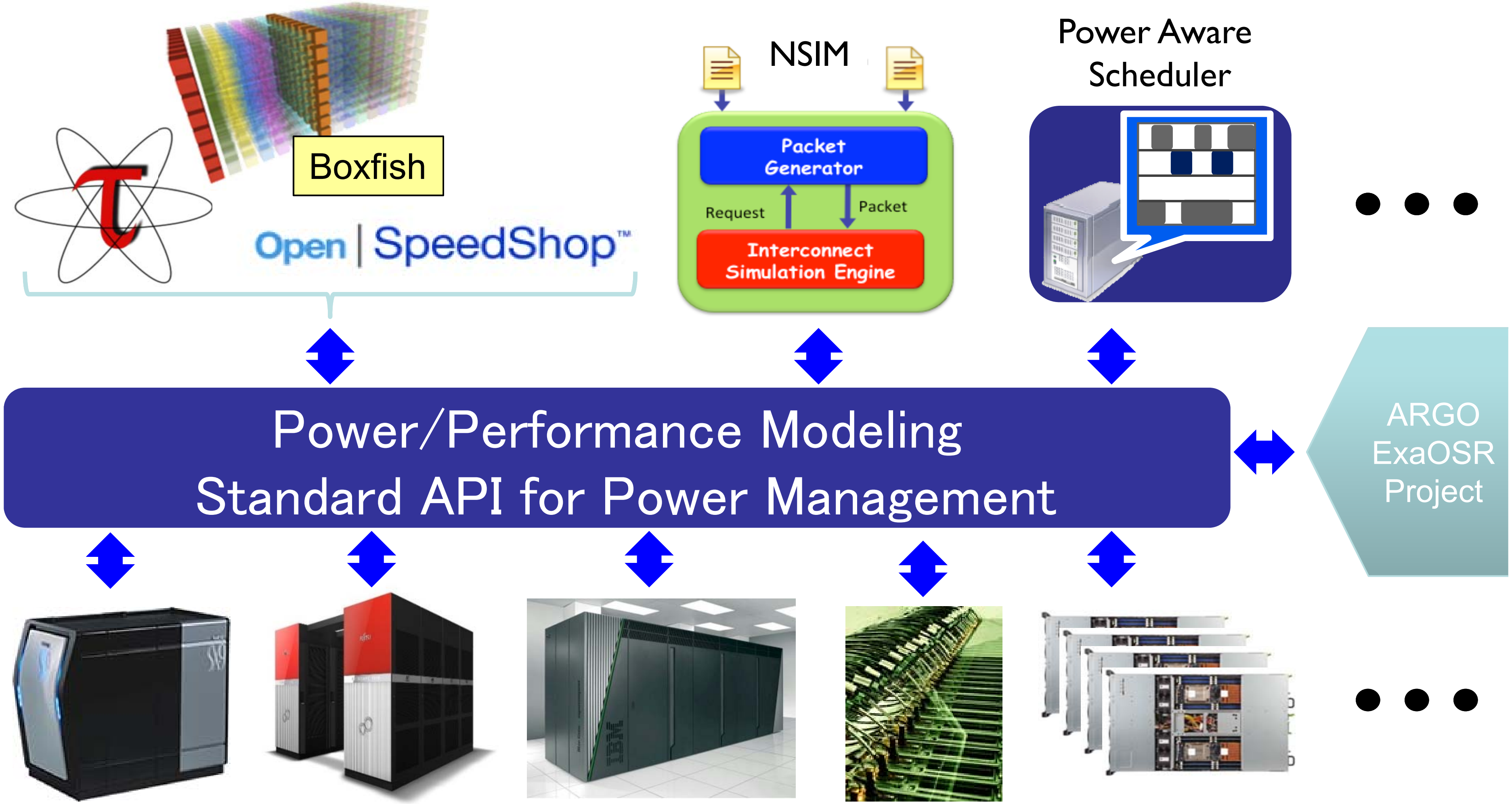
Power Efficiency of GB Dendrite Simulation since 2006

Need another leap!





Future vision – International Collaboration for Power



Existing Collaboration with Livermore and Sandia NL

Extreme Big Data (EBD)

Next Generation Big Data Infrastructure Technologies Towards Yottabyte/Year (2013H2-2018H1)

Principal Invesigator
Satoshi Matsuoka

Global Scientific Information and Computing Center
Tokyo Institute of Technolgy / JST CREST

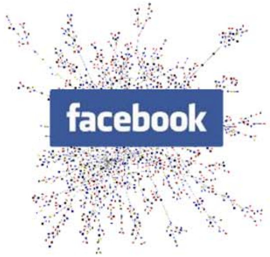
Convergence of HPC and Big Data

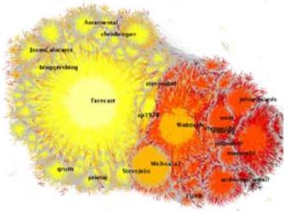
- The current "Big Data" are not really that Big…
  - Typical definition: "Mining people's privacy data to make money"
- But "Extreme Big Data" will change everything
  - "Breaking down of Silos" (Rajeeb Harza, Intel VP of Technical Computing)
- Already happening in Science & Engineering due to Open Data movement
- More complex analysis algorithms: O(n log n), O(m x n), …
- Fundamental to next gen IT Infrastructure – Clouds hosting convergent machines

Extreme Big Data Example in Social NW rates and volumes are immense

Slide courtesy David A. Bader @ Georgia Tech

- Facebook:
  - ~1 billion users
  - average 130 friends
  - 30 billion pieces of content shared / month
- Twitter:
  - 500 million active users
  - 340 million tweets / day
- Internet – 100s of exabytes / year
  - 300 million new websites per year
  - 48 hours of video to You Tube per minute
  - 30,000 YouTube videos played per second



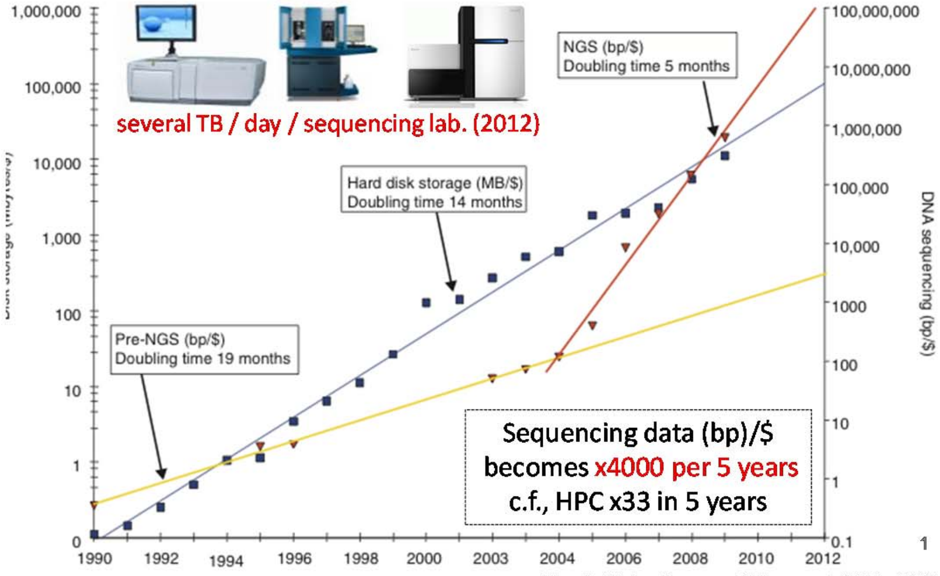


Continuous Billion-Scale Social Simulation with Real-Time Streaming Data (Toyotaro Suzumura/IBM-Tokyo Tech)

- **Applications**
  - Target Area: Planet (Open Street Map)
  - **7 billion people**
- **Input Data**
  - Road Network (Open Street Map) for Planet: **300 GB** (XML)
  - Trip data for 7 billion people
    - **10 KB (1 trip) x 7 billion = 70 TB**
  - Real-Time Streaming Data (e.g. Social sensor, physical data)
- **Simulated Output for 1 Iteration**
  - **700 TB**

Extreme Big Data in Genomics

[Slide Courtesy Yutaka Akiyama @ Tokyo Tech.]



Lincoln Stein, Genome Biology, vol. 11(5), 2010

Future "Extreme Big Data"

- <u>NOT mining Tbytes Silo Data</u>
- Peta~Zetabytes of Data
- Ultra High-BW Data Stream
- Highly Unstructured, Irregular
- Complex correlations between data from multiple sources
- Extreme Capacity, Bandwidth, Compute All Required

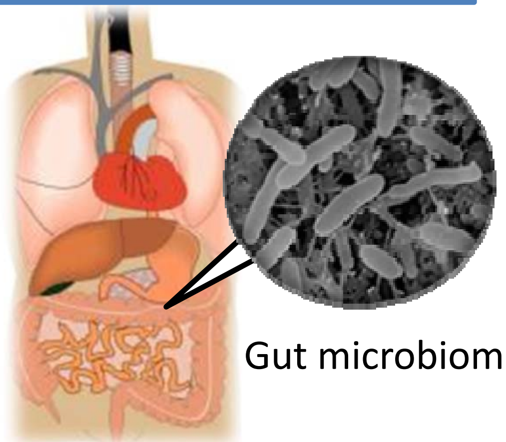
We will have tons of unknown genes

Metagenome analysis

[Slide Courtesy Yutaka Akiyama @ Tokyo Tech.]

- Directly sequencing uncultured microbiomes obtained from target environment and analyzing the sequence data
  - Finding novel genes from unculturable microorganism
  - Elucidating composition of species/genes of environments

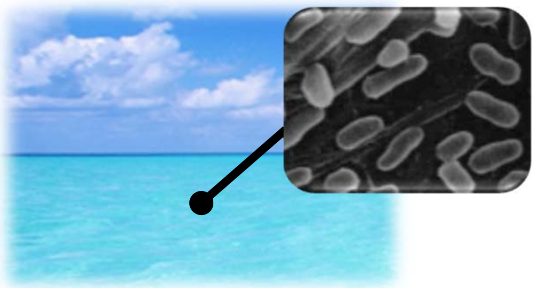
Examples of microbiome



Gut microbiome

Human body

Soil



Sea

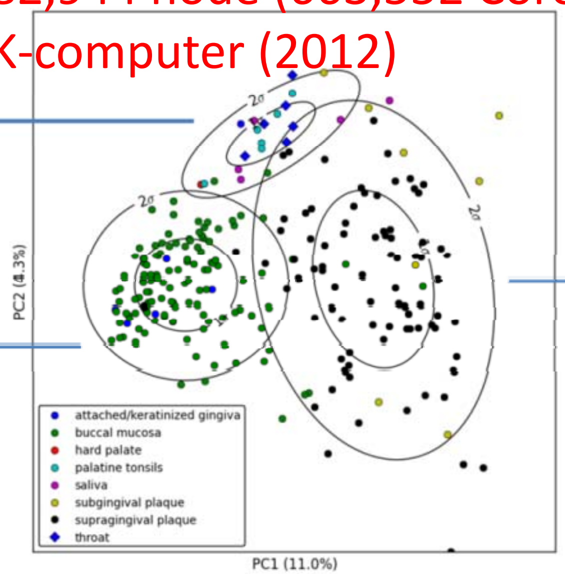
Results from Akiyama group@Tokyo Tech

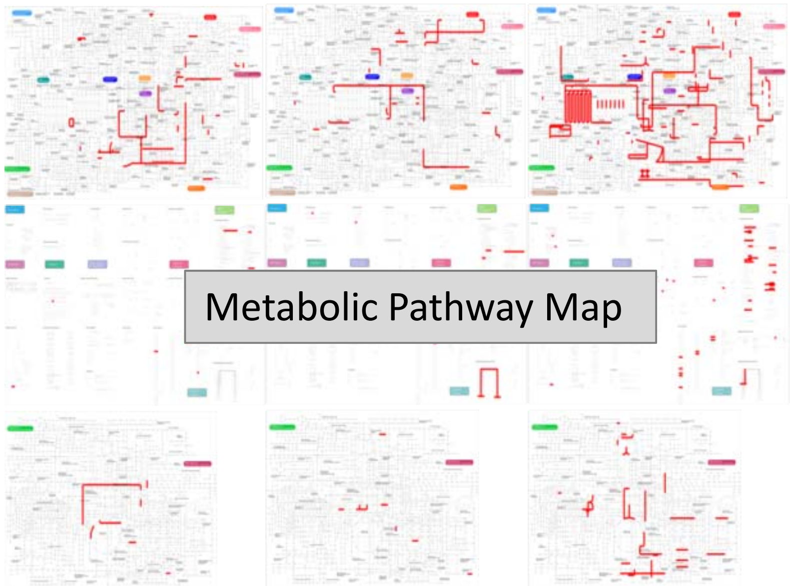
Ultra high-sensitive "big data" metagenome sequence analysis of human oral microbiome

- Required > **1 million node*hour product** on K-computer
- World's most sensitive sequence analysis (based on amino acid similar

- Discovered at least three microbiome clusters with <u>functional</u> differences.
  (Integrated 422 experiment samples taken from 9 different oral parts)

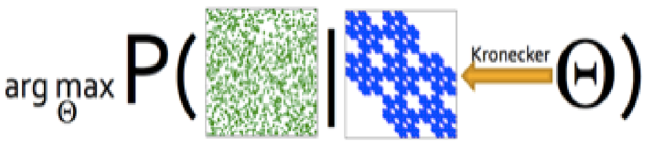
572.8 M Reads / hour
82,944 node (663,552 Cores)
K-computer (2012)





Graph500 “Big Data” Benchmark

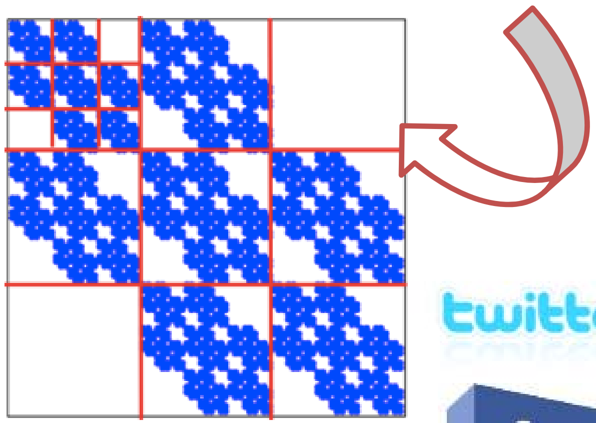
Kronecker graph BSP Problem



A: 0.57, B: 0.19
C: 0.19, D: 0.05

1	1	0
1	1	1
0	1	1

G_1



G_4 adjacency matrix

twitter facebook amazon.com

November 15, 2010 HPCwire

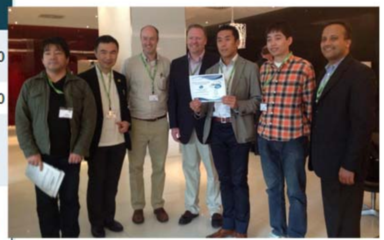
Graph 500 Takes Aim at a New Kind of HPC
Richard Murphy (Sandia NL => Micron)

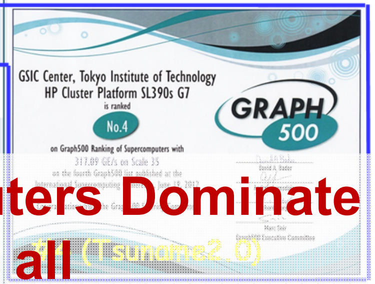
“ I expect that this ranking may at times look very different from the TOP500 list. Cloud architectures will almost certainly dominate a major chunk of part of the list.”

The 4th Graph500 List (Jun2012) TSUBAME #4 w/GPUs

Toyotaro Suzumura, Koji Ueno, Tokyo Institute of Technology

Rank	Installation Site	Machine	Number of nodes	Number of cores	Problem scale	GTEPS
1	DOE/SC/Argonne National Laboratory	Mira/BlueGene/Q	32768	524288	38	3541.00
1	LLNL	Sequoia/Blue Gene/Q	32768	524288	38	3541.00
2	DARPA Trial Subset, IBM Development Engineering	Power 775, POWER7 8C 3.836 GHz	1024	32768	35	508.05
3	Information Technology Center, The University of Tokyo	Oakleaf-FX (Fujitsu PRIMEHPC FX 10)	4800	76800	38	358.10
4	GSIC Center, Tokyo Institute of Technology	TSUBAME	1366	16392	35	317.09
5	Brookhaven National Laboratory	BLUE GENE/Q	1024	16384	34	294.29
6	DOE/SC/Argonne National Laboratory	Vesta/BlueGene/Q	1024	16384	34	292.36
[illegible]	NASA [illegible] / Parallel [illegible], Intel Labs	Pleiades - SGI ICE X, dual [illegible] Infiniband, E5-2670 "sandybridge"	[illegible]	[illegible]	[illegible]	[illegible]
8	NERSC/LBNL	XE6	[illegible]	[illegible]	[illegible]	[illegible]
9	NNSA and IBM [illegible]	NNSA/SC Blue Gene/Q [illegible]	[illegible]	65536	[illegible]	[illegible]

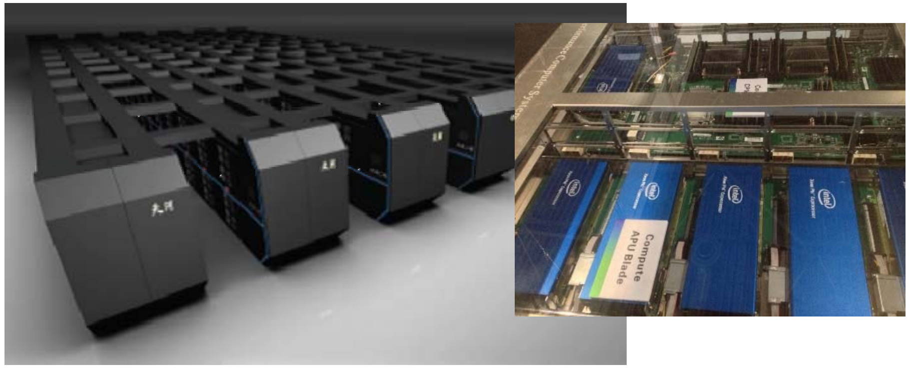




Reality: Top500 Supercomputers Dominate
No Cloud IDCs at all
TSUBAME2.0 #3(Nov.2011) #4(Jun.2012)

Top Supercomputers vs. Global IDC

K Computer (#1 2011-12) Riken-AICS
Fujitsu Sparc VIII-fx Venus CPU
88,000 nodes, 800,000CPU cores
~11 Petaflops (10^{16})
1.4 Petabyte memory, 13 MW Power
864 racks、3000m^2



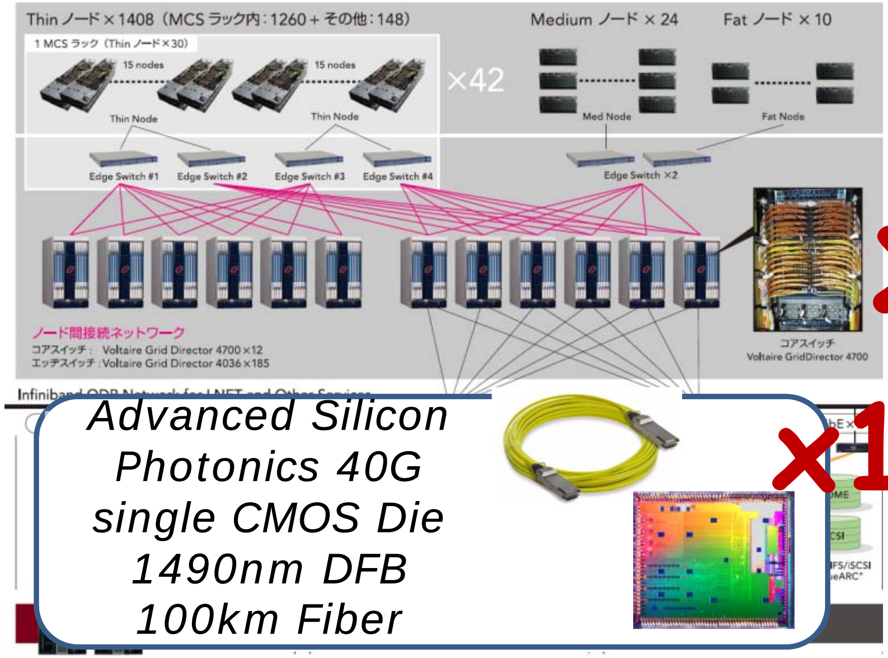
Tianhe2 (#1 2013) China Gwanjou
48,000 KNC Xeon Phi + 36,000 Ivy Bridge Xeon
18,000 nodes, >3 Million CPU cores
54 Petaflops (10^{16})
0.8 Petabyte memory, 20 MW Power
??? racks、???m^2

C.f. Amazon ~= 500,000 Nodes, ~5 million Cores

#1 2012 IBM BlueGene/Q "Sequoia"
Lawrence Livermore National Lab
IBM PowerPC System-On-Chip
98,000 nodes, 1.57million Cores
~20 Petaflops
1.6 Petabytes, 8MW, 96 racks

DARPA study
2020 Exaflop (10^{18})
100 million~
1 Billion Cores

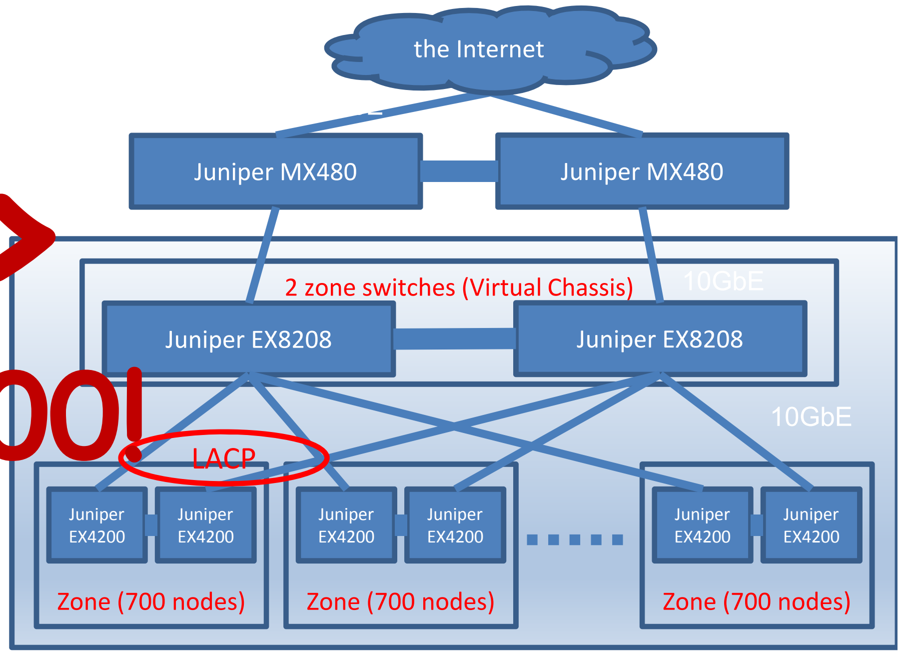
Supercomputer Tokyo Tech. Tsubame 2.0 #4 Top500 (2010)



Advanced Silicon Photonics 40G single CMOS Die 1490nm DFB 100km Fiber

> > ×1000!

A Major Northern Japanese Cloud Datacenter (2013)



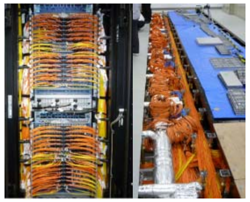
<u>~1500 nodes</u> compute & storage
Full Bisection Multi-Rail Optical Network
<u>Injection 80GBps/Node</u>
<u>Bisection 220Terabps</u>

8 zones, Total <u>5600 nodes</u>,
<u>Injection 1GBps/Node</u>
<u>Bisection 160Gigabps</u>

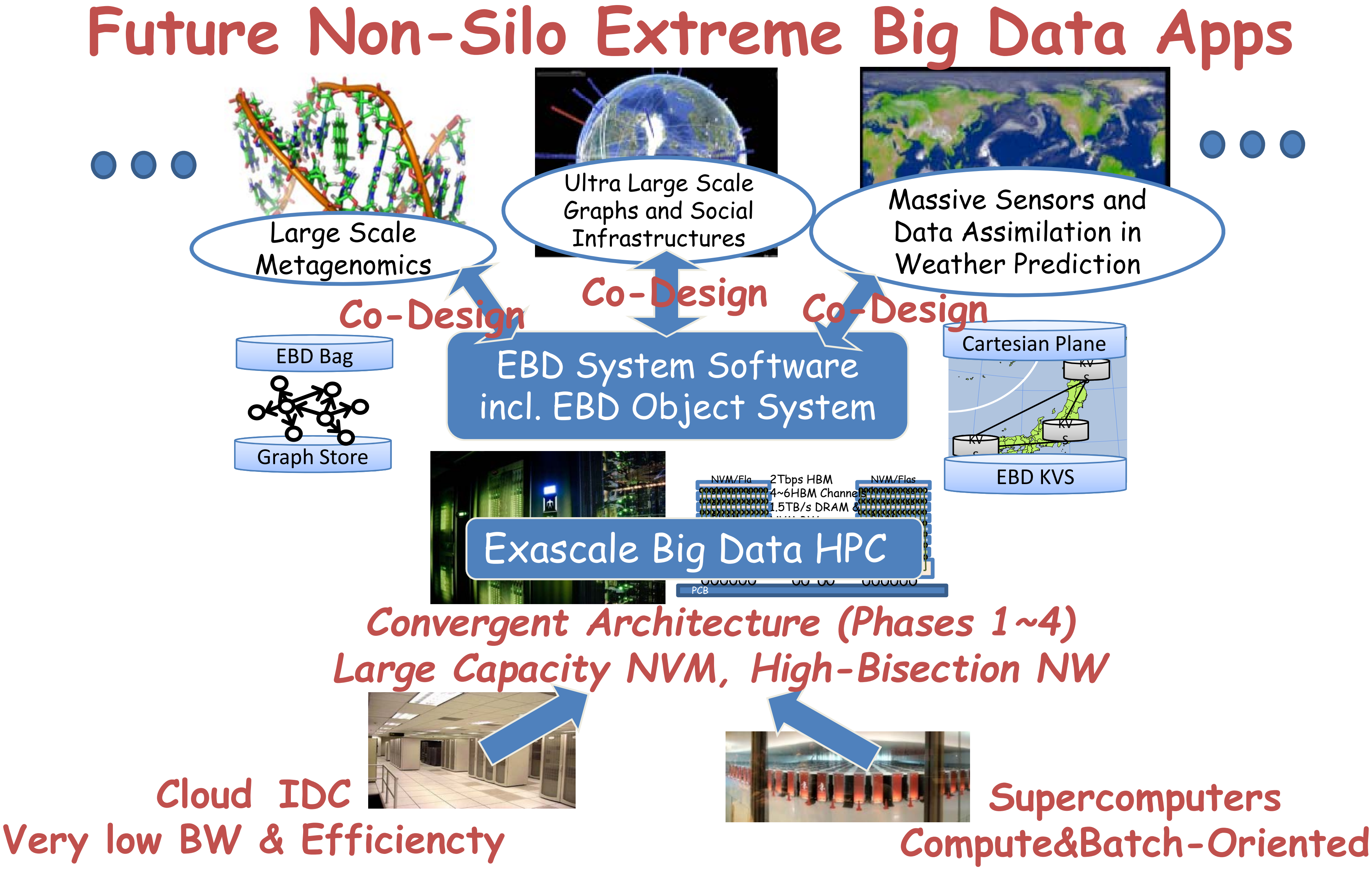
But what does "220Tbps" mean?

Global IP Traffic, 2011-2016 (Source Cicso)							
	2011	2012	2013	2014	2015	2016	CAGR 2011-2016
By Type (PB per Month / Average Bitrate in Tbps)							
Fixed Internet	23,288	32,990	40,587	50,888	64,349	81,347	28%
	71.9	101.8	125.3	157.1	198.6	251.1	
Managed IP	6,849	9,199	11,846	13,925	16,085	18,131	21%
	21.1	28.4	36.6	43.0	49.6	56.0	
Mobile data	597	1,252	2,379	4,215	6,896	10,804	78%
	1.8	3.9	7.3	13.0	21.3	33.3	
Total IP traffic	30,734	43,441	54,812	69,028	87,331	110,282	29%
	94.9	134.1	169.2	213.0	269.5	340.4	

TSUBAME2.0 Network has TWICE the capacity of the Global Internet, being used by 2.1 Billion users



JST-CREST Extreme Big Data Research Scheme (2013-2018)

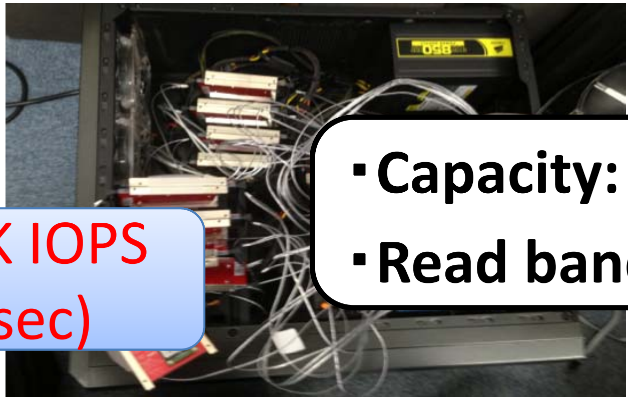


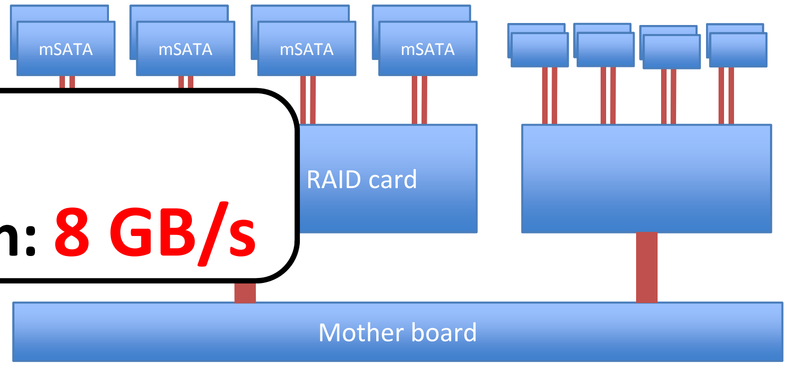
EBD- I/O
(Many-core I/O)

Preliminary I/O Evaluation on GPU and NVRAM for TSUBAME3.0(2016)

How to design local storage for next-gen supercomputers ?

- Designed a local I/O prototype using 16 mSATA SSDs

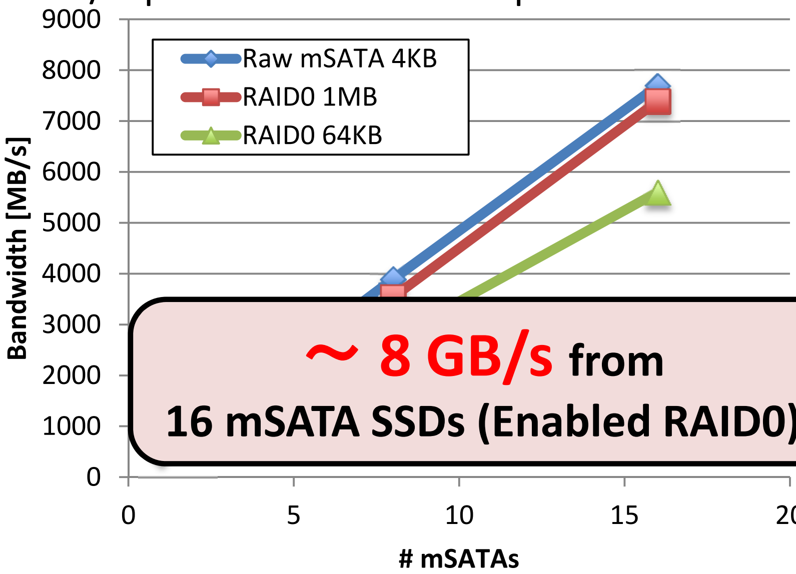




～320K IOPS
(3 μ sec)

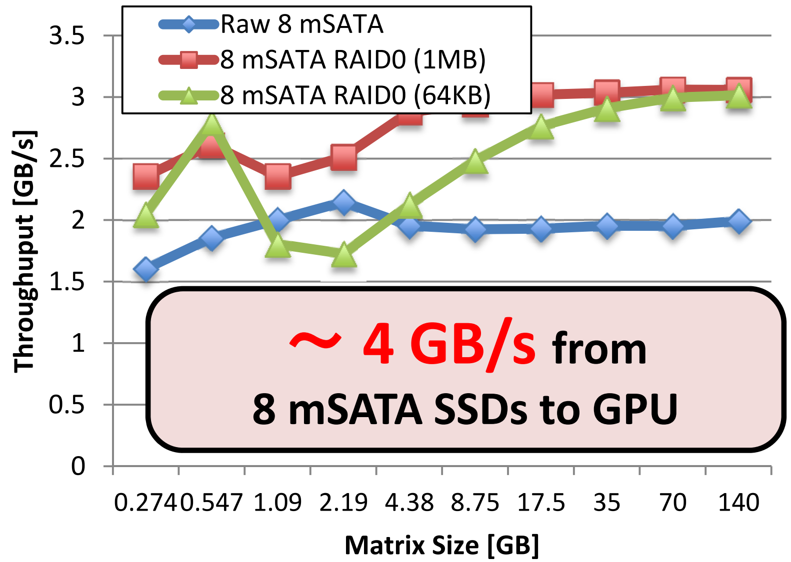
- **Capacity: 4TB**
- **Read bandwidth: 8 GB/s**

I/O performance of multiple mSATA SSD



～ 8 GB/s from 16 mSATA SSDs (Enabled RAID0)

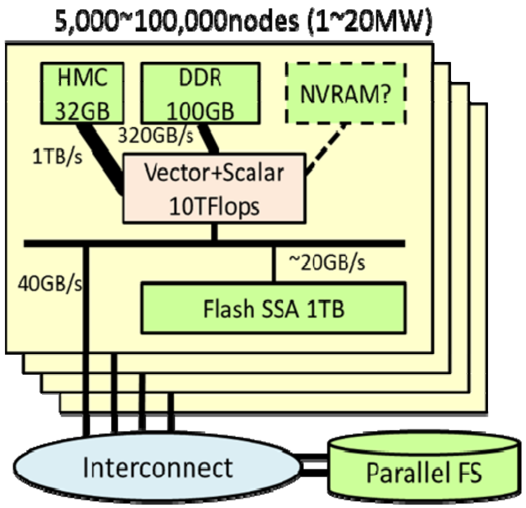
I/O performance from GPU to multiple mSATA SSDs

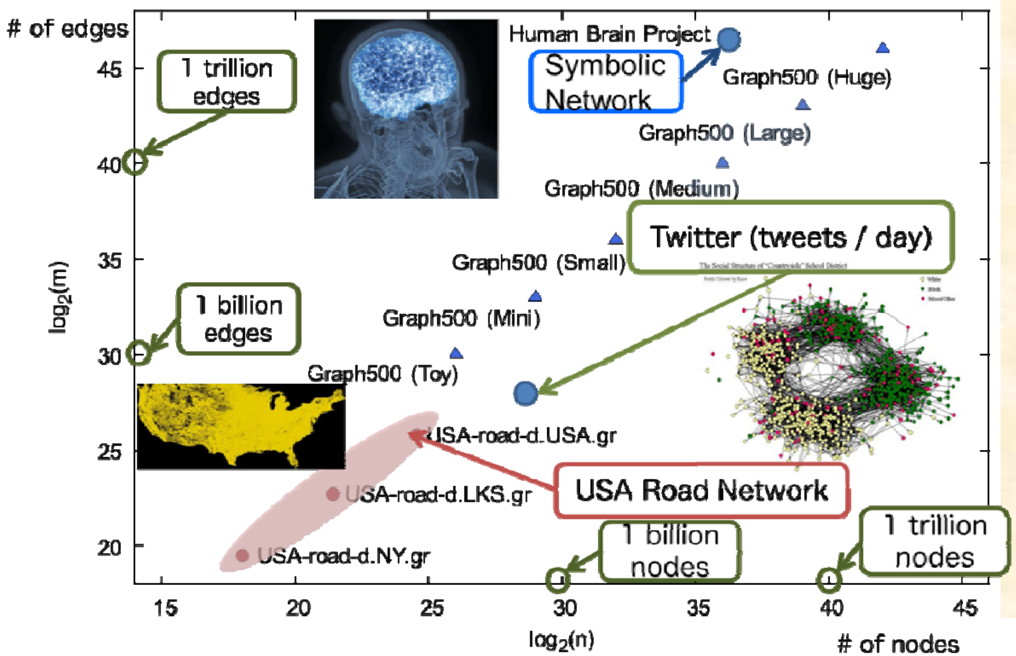


～ 4 GB/s from 8 mSATA SSDs to GPU

JST CREST: Advanced Computing and Optimization Infrastructure for Extremely Large-Scale Graphs on Post Peta-Scale Supercomputers

- Innovative Algorithms and implementations
  - Optimization, Searching, Clustering, Network flow, etc.
- Extreme Big Graph Data for emerging applications
  - **2^{30} ~ 2^{42} nodes** and **2^{40} ~ 2^{46} edges**
  - **Over 1M threads** are required for real-time analysis
- Many applications on post peta-scale supercomputers
  - Analyzing massive cyber security and social networks
  - Optimizing smart grid networks
  - Health care and medical science
  - Understanding complex life system





- Example: Symbolic Network
  - **Human Brain Project** **http://www.humanbrainproject.eu/**
  - Understanding the human brain is one of the greatest challenges facing 21st century science
  - **89 billion neurons**(nodes)
  - **1 trillion connections**(edges)
  - Over 10^{17} bytes memory(storage) and 10^{18} Flops for brain simulator

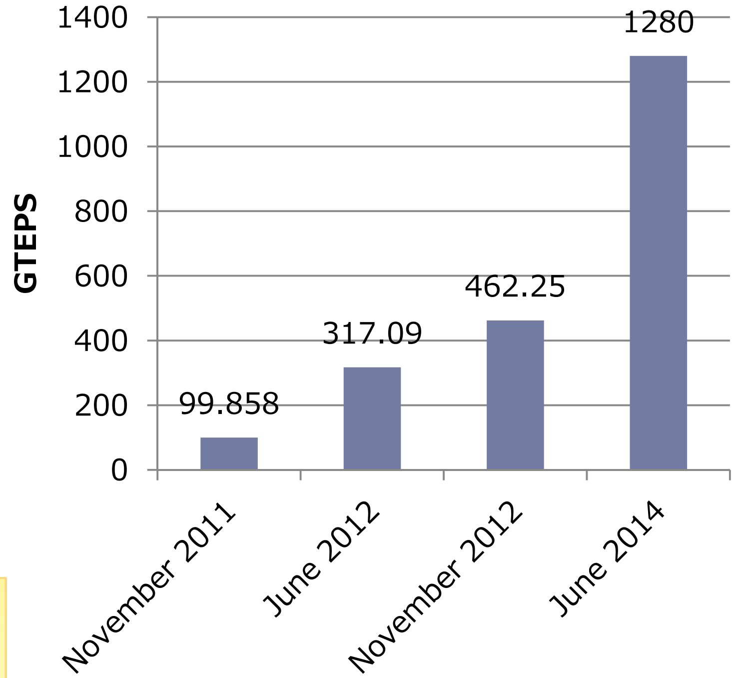
The Graph500 – June 2014
K Computer and TSUBAME 2.0 & 2.5

Graph500 ranking history for TSUBAME2.0 and 2.5

List	Rank	GTEPS	Implementation
November 2011	3	99.858	Top-down only
June 2012	4	317.09	GPU
November 2012	20	462.25	GPU
June 2014	12	1280	**Efficient hybrid**

*Every score is obtained using TSUBAME2.0 1366 nodes or TSUBAME 2.5 1024 nodes

BFS performance on TSUBAME2.0 and 2.5



Graph500 ranking history for K Computer

List	Rank	GTEPS	Implementation
November 2013	4	5524.12	Top-down only
June 2014	1	17977.05	**Efficient hybrid**

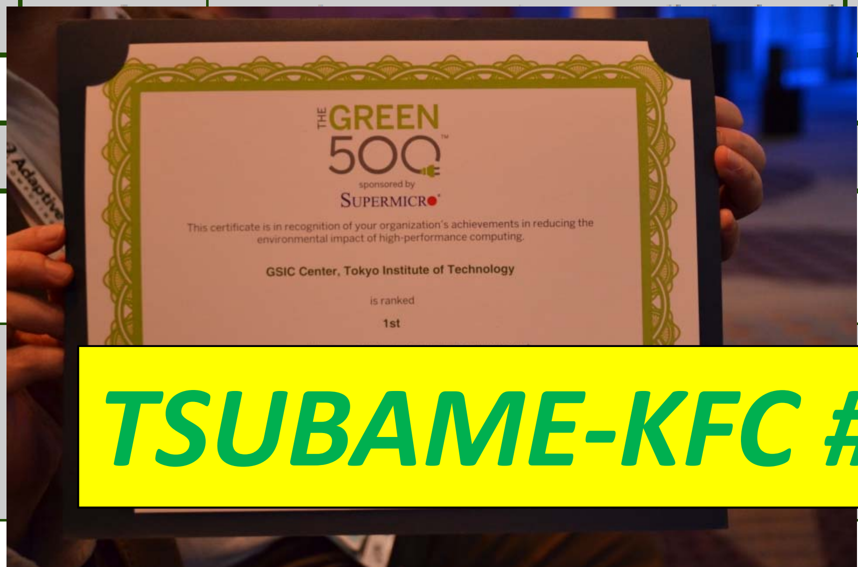

RIKEN Advanced Institute for Computational Science (AICS)'s K computer
is ranked
No.1
on the Graph500 Ranking of Supercomputers with
17977.1 GE/s on Scale 40
on the 8th Graph500 list published at the International Supercomputing Conference, June 22, 2014.
Congratulations from the Graph500 Executive Committee
GRAPH 500
David A. Bader
Andrew Lumsdaine
Richard Murphy
Marc Snir
Graph500 Executive Committee

2013/11 Green Graph500 Ranking

- **TEPS (Traversal Edges Per Watt)**
- **http://green.graph500.org**

In the **Big Data** category:

Rank	MTEPS/W	Site	Machine	G500 rank	Scale	GTEPS	Nodes
1	**6.72**	Tokyo Institute of Technology	TSUBAME KFC	47	32	44.01	32
2							6384
3			DO				2768
4			E				1
5			DC				5536
6							1
7							64



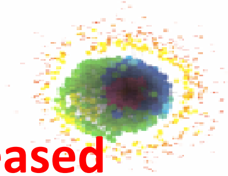
TSUBAME-KFC #1 For Big Data!

EBD Algorithm Kernels

Large Scale BFS Using NVRAM

1. Introduction

- Large scale graph processing in various domains



DRAM resources has increased

- Spread of Flash Devices

Prof : Price per bit, Energy consumption

Cons: Latency, Throughput

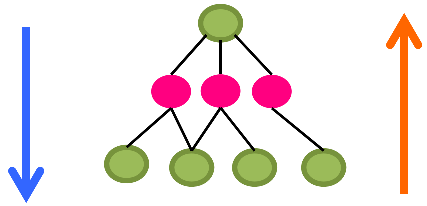
Using NVRAMs for large scale graph processing has possibilities of minimum performance degradation

2.Hybrid-BFS

Switch two approaches

Top-down

$n_{frontier} < \frac{n_{all}}{\beta}$



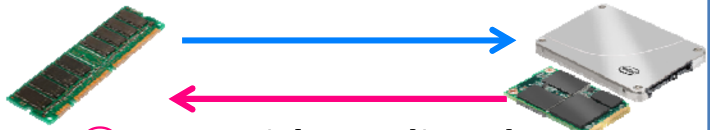
Bottom-up

$n_{frontier} > \frac{n_{all}}{\alpha}$

# of frontiers: $n_{frontier}$, # of all vertices: n_{all}, parameter : α, β

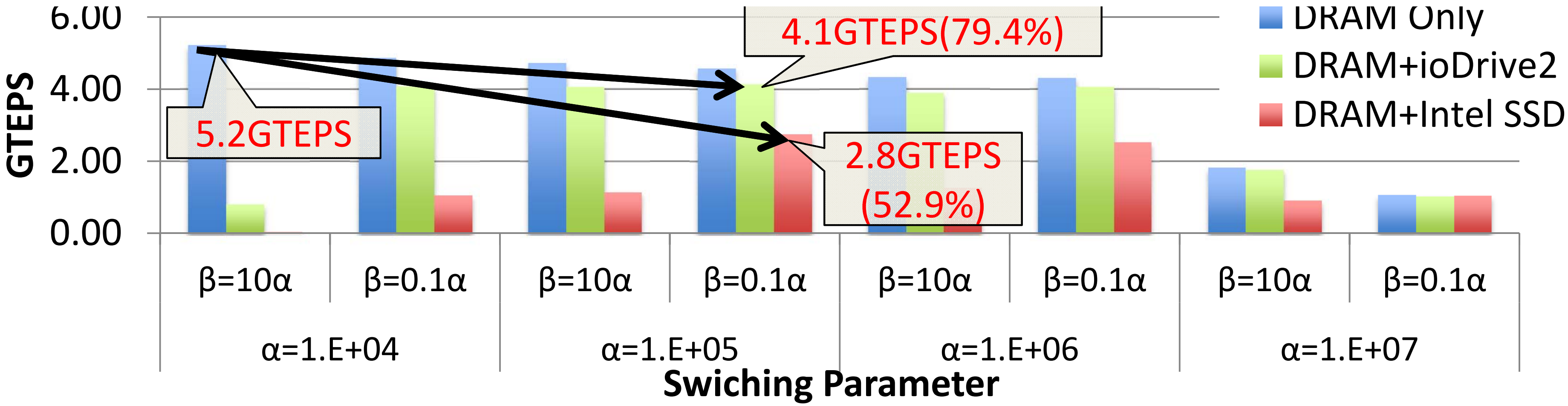
3.Proporsal

① **offload small accesses data**



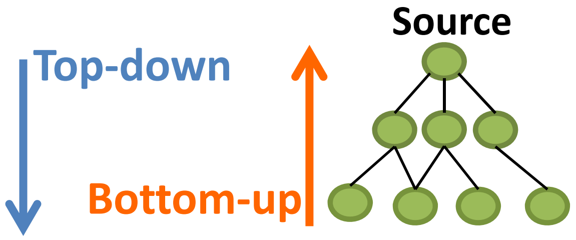
② **BFS with reading data from NVRAM**

4.Evaluation (Offload Top-down Graph : we could reduce half the size of DRAM [128GB -> 64 GB] at Scale 27)

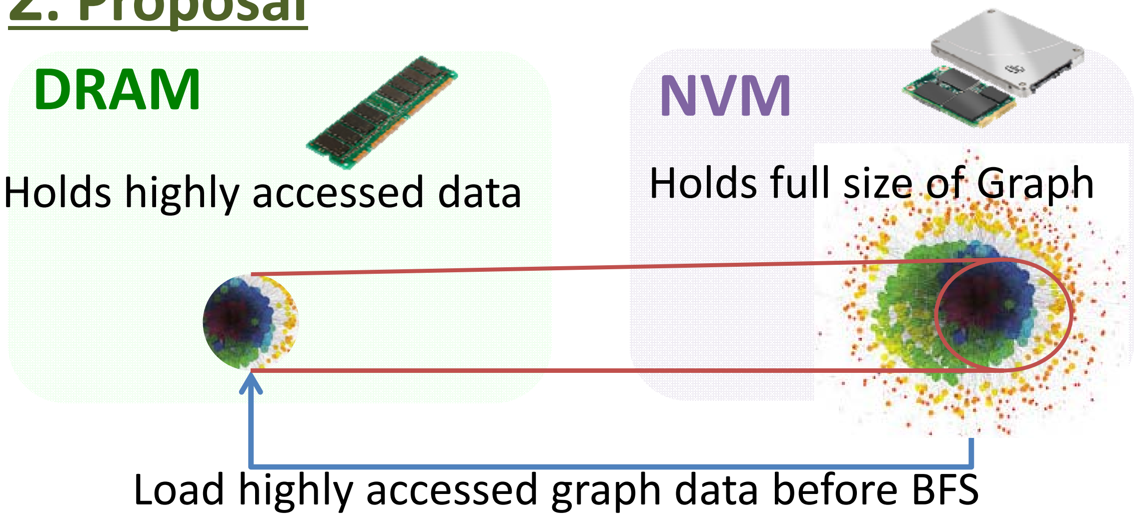


Large Scale Graph Processing Using NVM

1. Hybrid-BFS (Beamer'11)

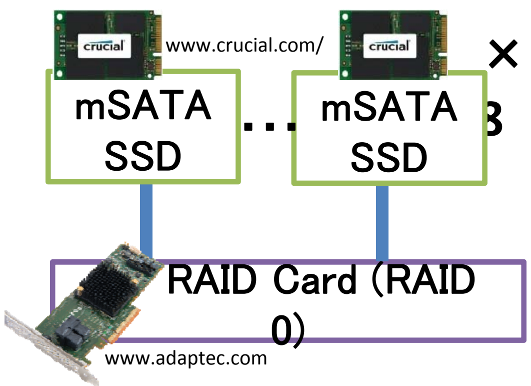


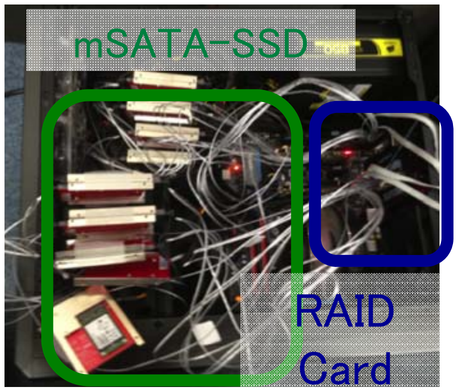
2. Proposal

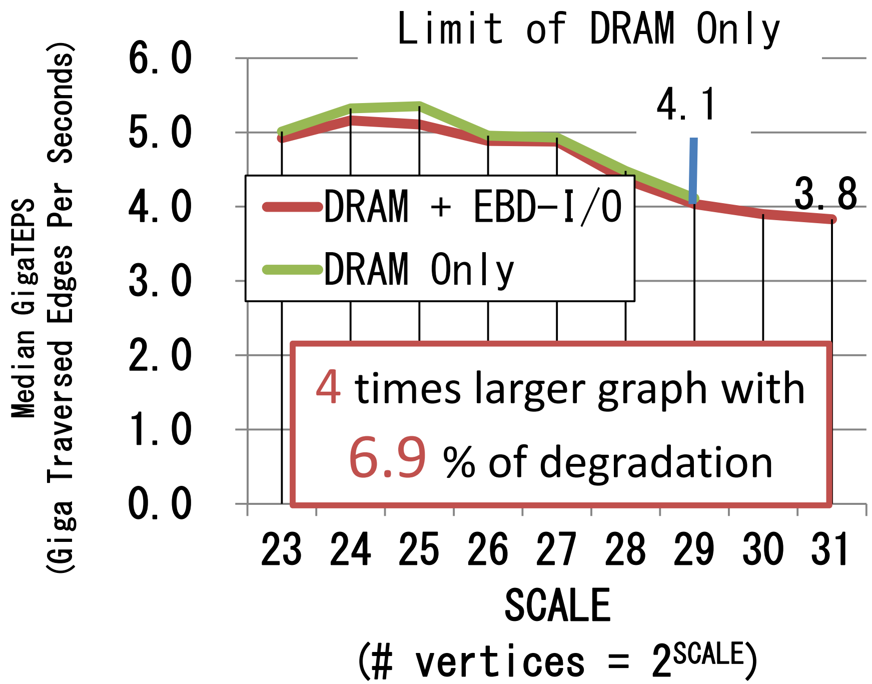


3. Experiment

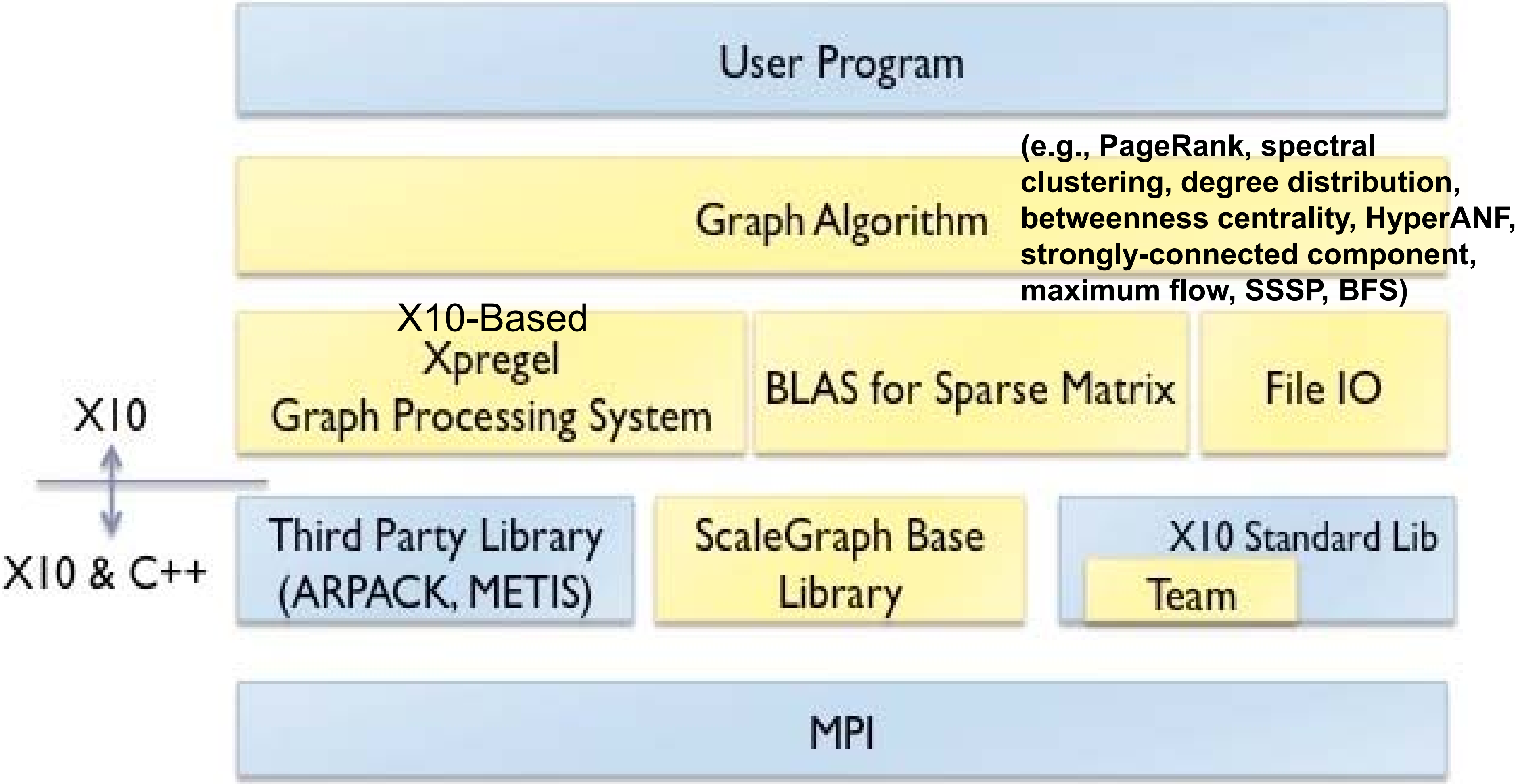
CPU	Intel Xeon E5-2690 × 2
DRAM	256 GB
NVM	EBD-I/O 2TB × 2







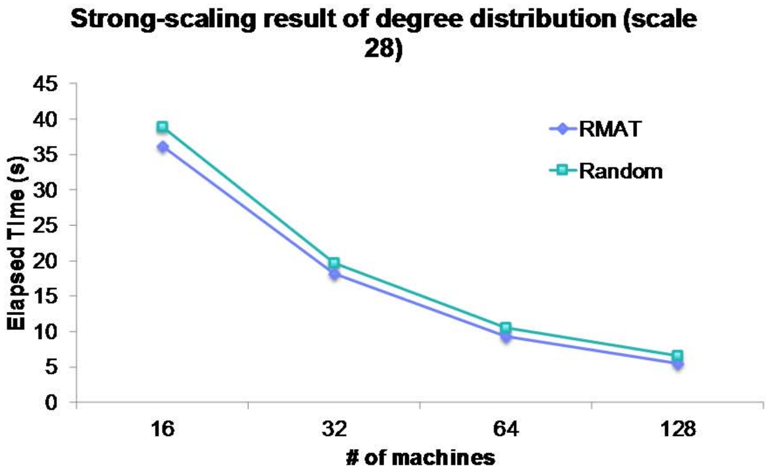
ScaleGraph : Large-Scale Graph Analytics Library for HPC-Big Data Convergent Architecture [Suzumura, Ueno et. al.]



XPregel – X10-based Pregel-like Graph Programming System for convergent architectures

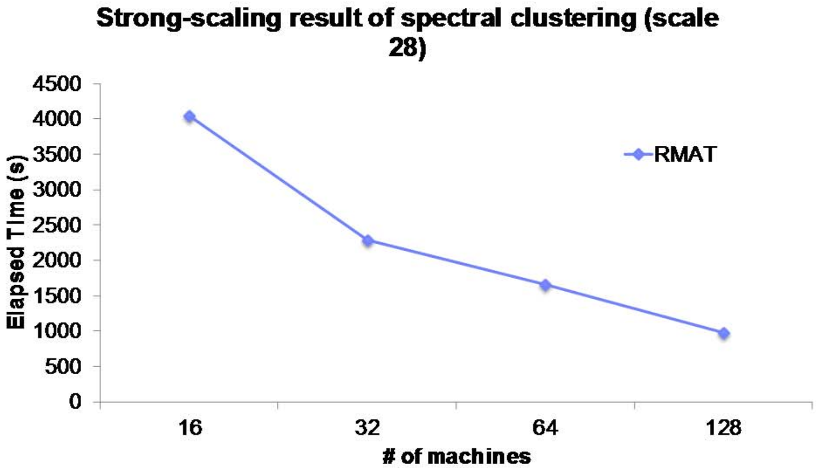
- XPregel optimizations on supercomputers
  1. Utilize MPI collective communication.
  2. Avoid serialization, which enables utilizing fast supercomputer interconnects
  3. Destination of messages computed by a simple bit manipulation thanks to vertex id renumbering.
  4. Optimized message communication when all vertices send the same message to all the neighbor vertices.
  5. Simple API in X10 language.

Degree Distribution



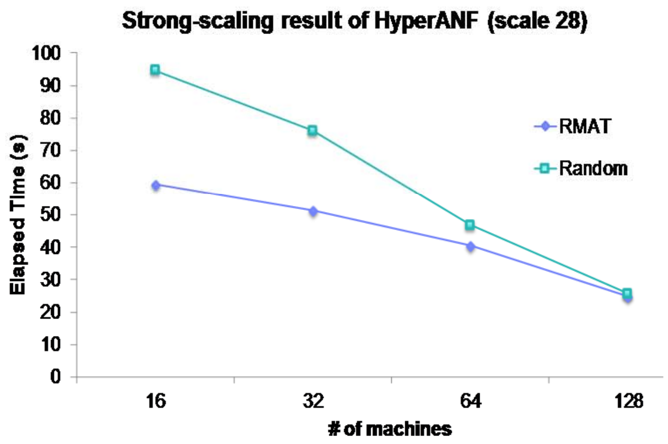
The scale-28 graphs we used have 2^{28} (≈268 million) of vertices and 16×2^{28} (≈4.29 billion) of edges

Spectral Clustering



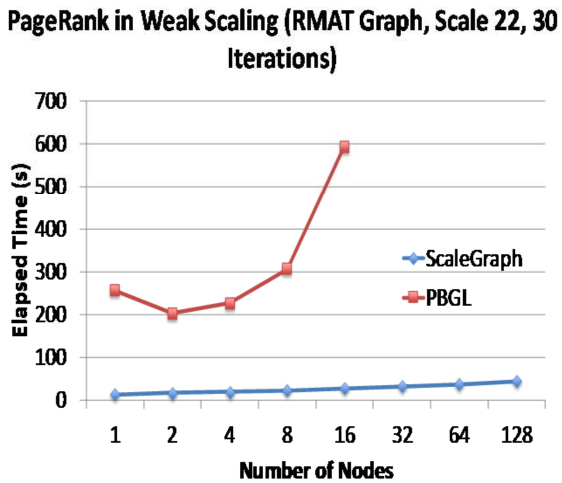
The scale-28 graphs we used have 2^{28} (≈268 million) of vertices and 16×2^{28} (≈4.29 billion) of edges

Degree of Separation

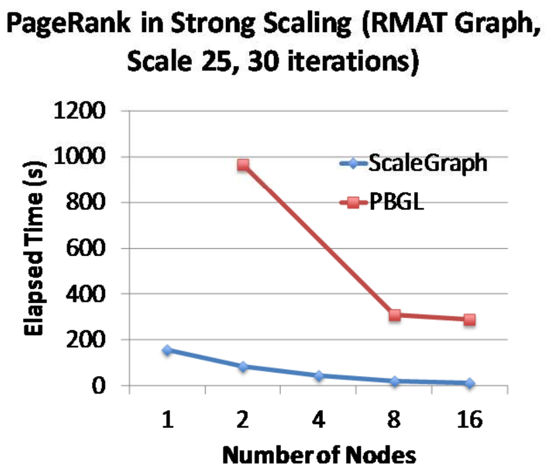


The scale-28 graphs we used have 2^{28} (≈268 million) of vertices and 16×2^{28} (≈4.29 billion) of edges

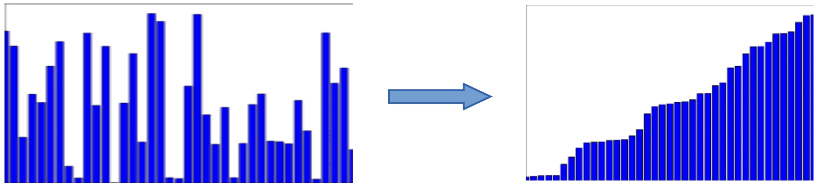
ScaleGraph vs. PBGL



ScaleGraph vs. PBGL

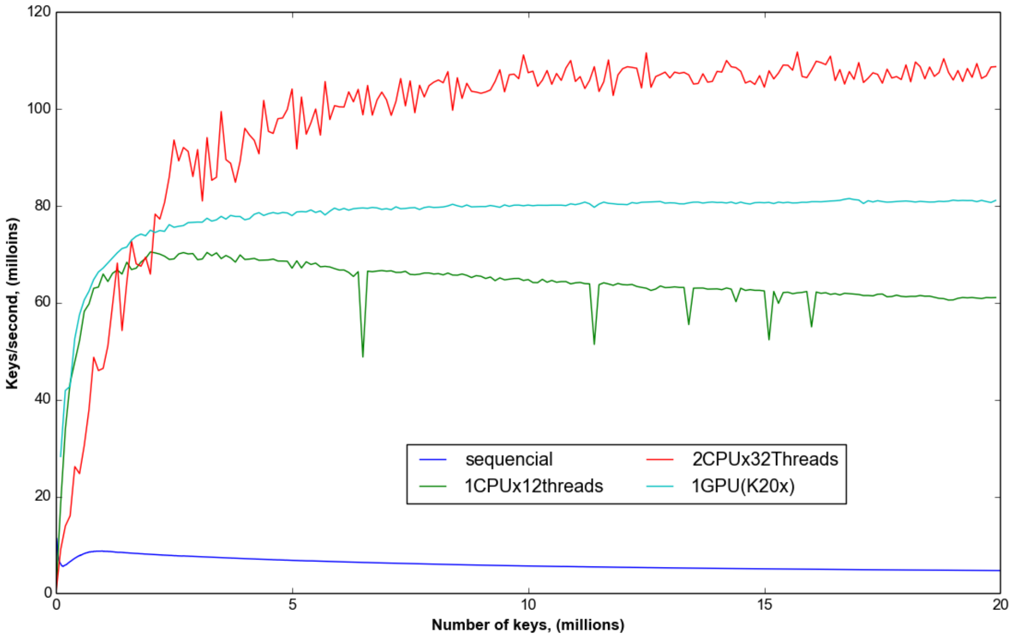


Sorting for EBD [BigData2013]
using single node to the utmost capacity



Sorting
One of the fundamental primitives
Extremely well studied
Variety of data types, sizes, hardware architectures and characteristics leave lots of space for improvement.

- Sorting long/variable length keys (strings)
- Implementations for GPUs and multi /many-core CPUs
- Hybrid parallelization scheme combining data-parallel and task-parallel stages
- Trimming keys to reduce host-to-device communication overheads
- Up to **100 million string keys per second**



MSD radix sort

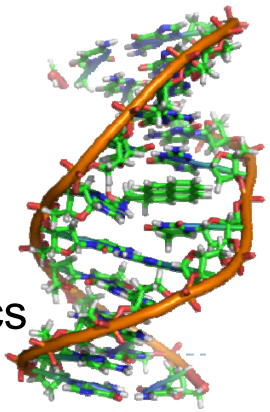
apple
apricot
banana
kiwi

Don't have to examine all characters

Processing textual data (e.g. corpus linguistics)

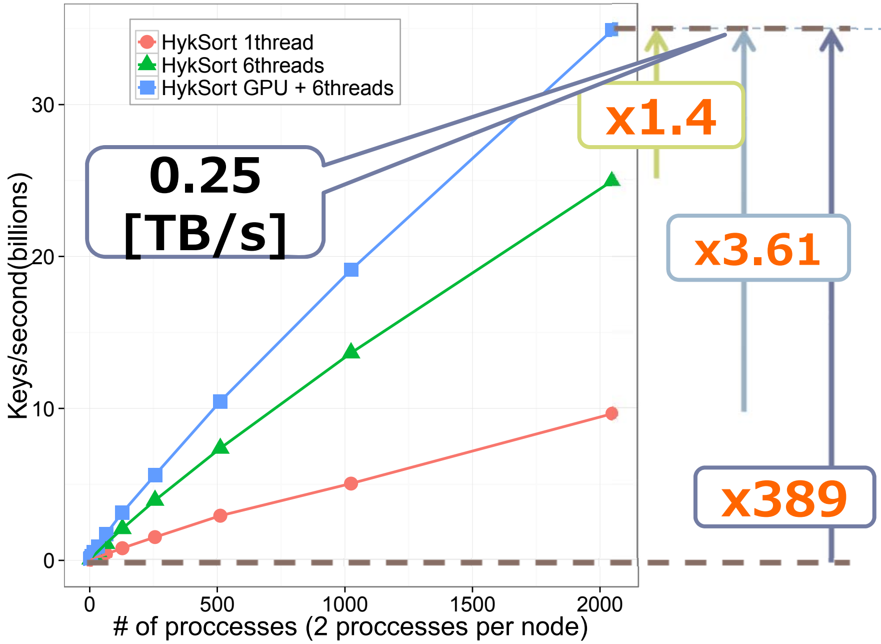
High efficiency on small alphabets

Computational genomics (A,C,G,T)



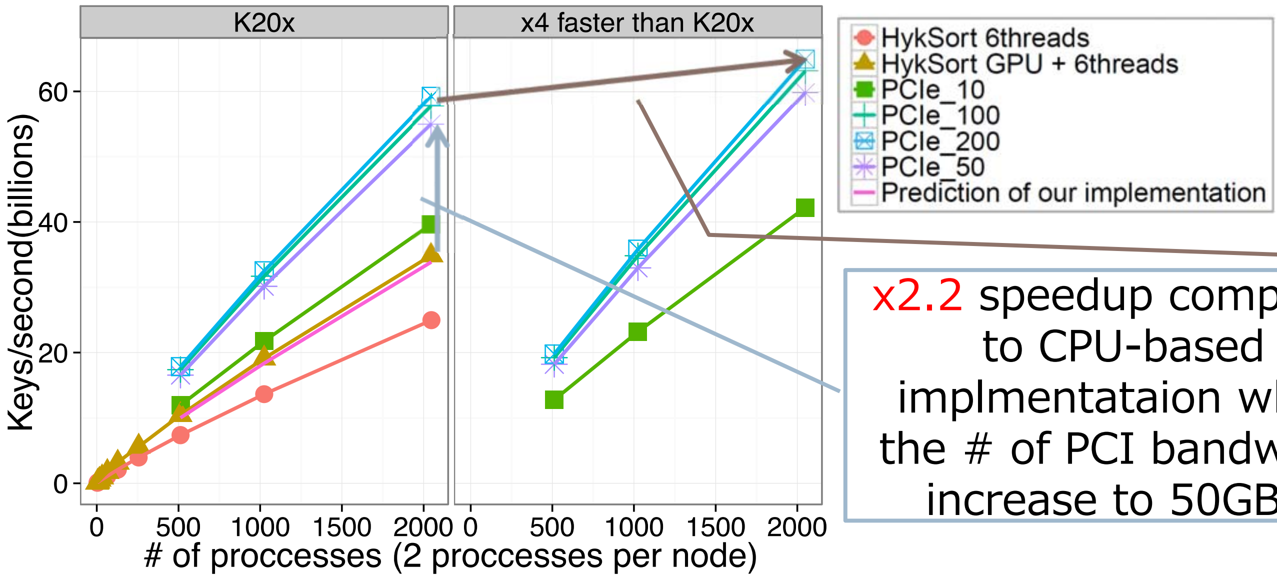
Sorting for EBD [IEEE BigData 2013]

Plugging in GPUs for large-scale sorting



- **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 and each node has 2GB 64bit integer
- Yahoo/Hadoop Terasort: 0.02[TB/s]
  - Including I/O

- **Performance prediction**



- PCIe_#: #GB/s bandwidth of interconnect between CPU and GPU

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

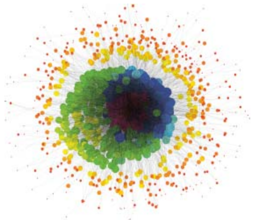
8.8% reduction of overall runtime when the accelerators work 4 times faster than K20x

Hamar (Highly Accelerated Map Reduce) [IEEE CCGrid 2013, IEEE Cluster 2014]

- A software framework for large-scale supercomputers w/ many-core accelerators and local NVM devices
  - Abstraction for deepening memory hierarchy
    - Device memory on GPUs, DRAM, Flash devices, etc.
- Features
  - Object-oriented
    - C++-based implementation
    - Easy adaptation to modern commodity many-core accelerator/Flash devices w/ SDKs
      - CUDA, OpenNVM, etc.
  - Weak-scaling over 1000 GPUs
    - TSUBAME2
  - Out-of-core GPU data management
    - Optimized data streaming between device/host memory
    - GPU-based external sorting
  - Optimized data formats for many-core accelerators
    - Similar to JDS format

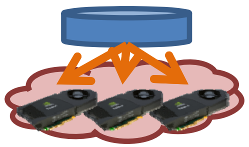
EBD Programming Model

Multi-GPU Map-Reduce GIM-V with Load Balance Optimization[CCGrid2013]



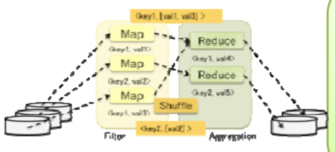
Graph Application

PageRank



Graph Algorithm

Multi-GPU GIM-V



MapReduce Framework

Multi-GPU Mars

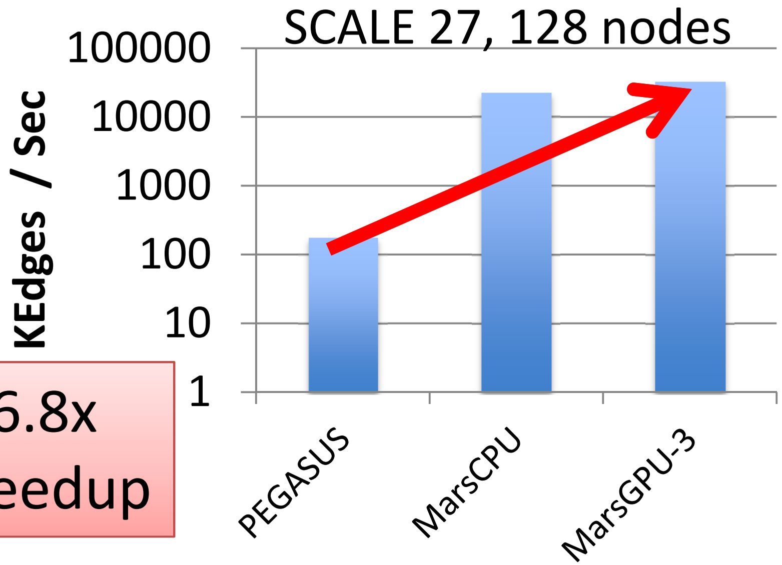
Platform

CUDA, MPI

Implement GIM-V on multi-GPUs MapReduce

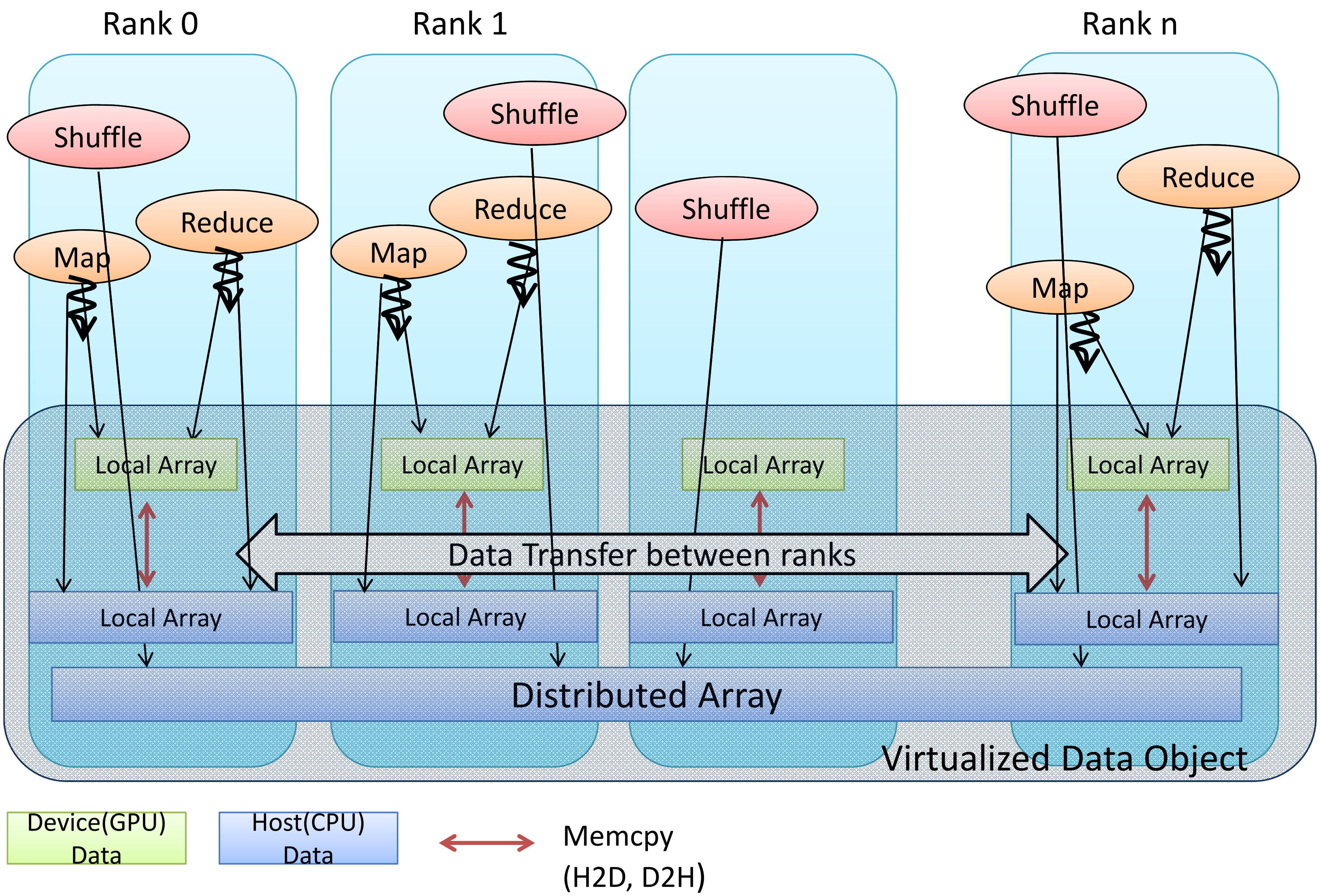
- Optimization for GIM-V
- Load balance optimization

Extend existing GPU MapReduce framework (Mars) for multi-GPU



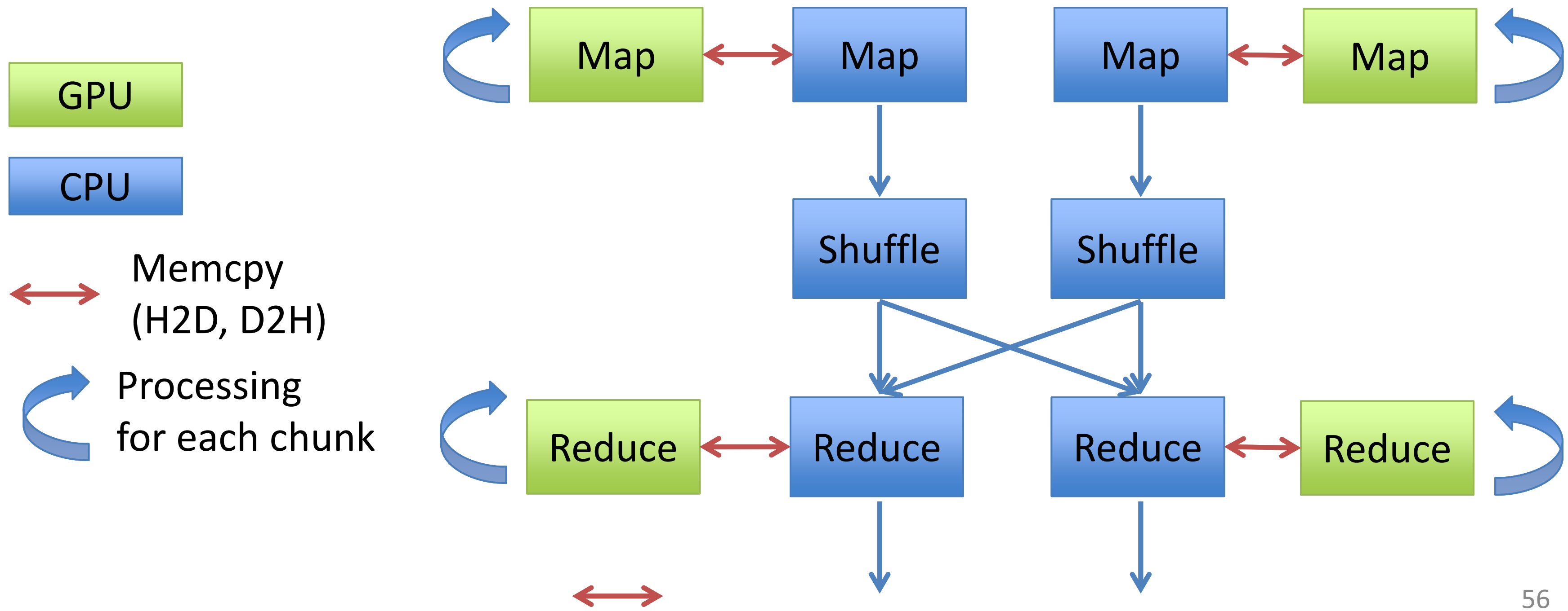
186.8x Speedup

Hamar Overview



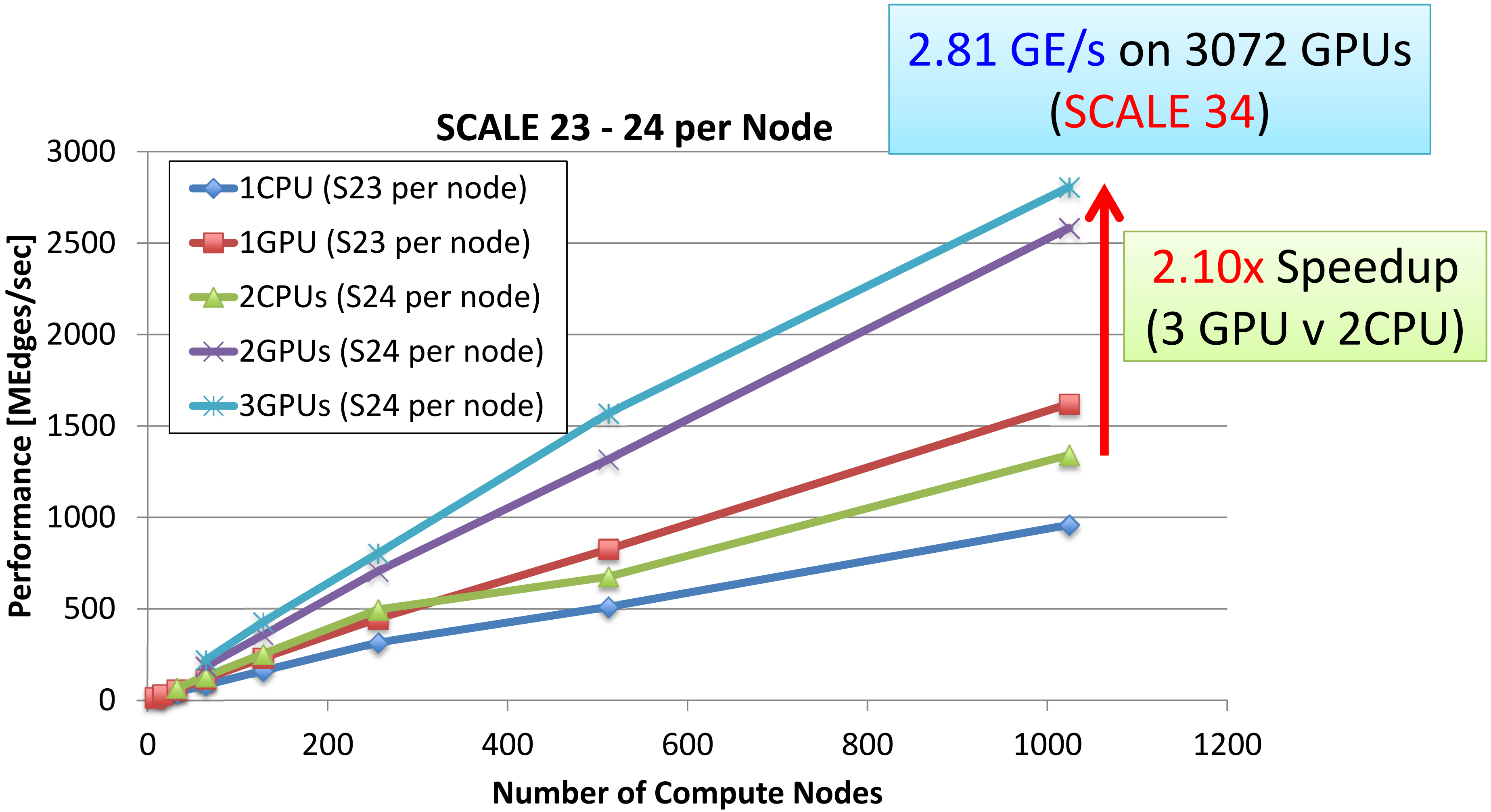
Highly Accelerated MapReduce with Out-of-core support on GPUs

- Hierarchical memory management for large-scale data parallel processing using multi-GPUs
  - Support out-of-core processing on GPU devices
  - Overlapping computation and communication

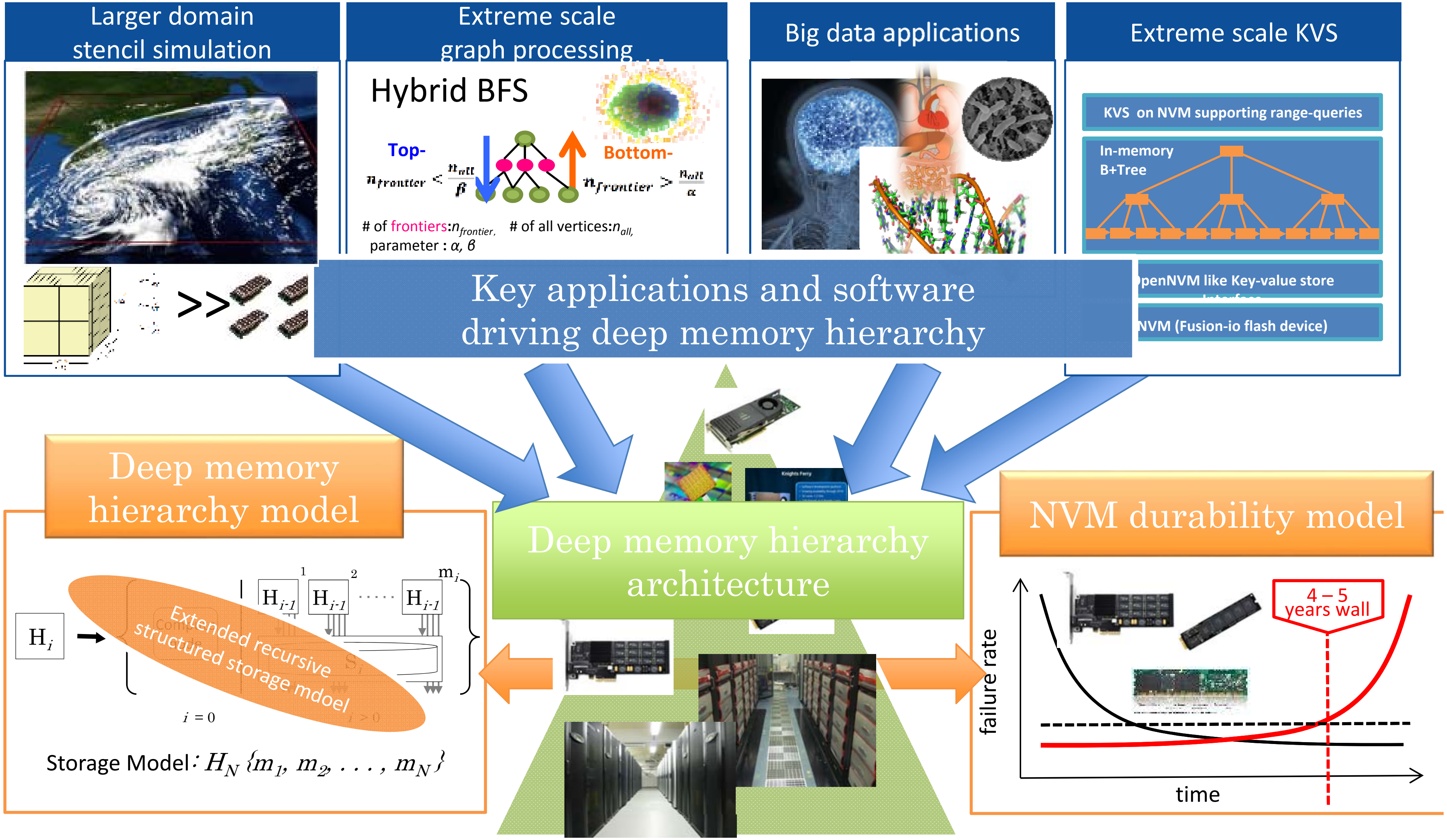


Weak Scaling Performance

- PageRank application on TSUBAME 2.5
- Data size is larger than GPU memory capacity



Future: Big Data & Deep memory hierarchy and modeling



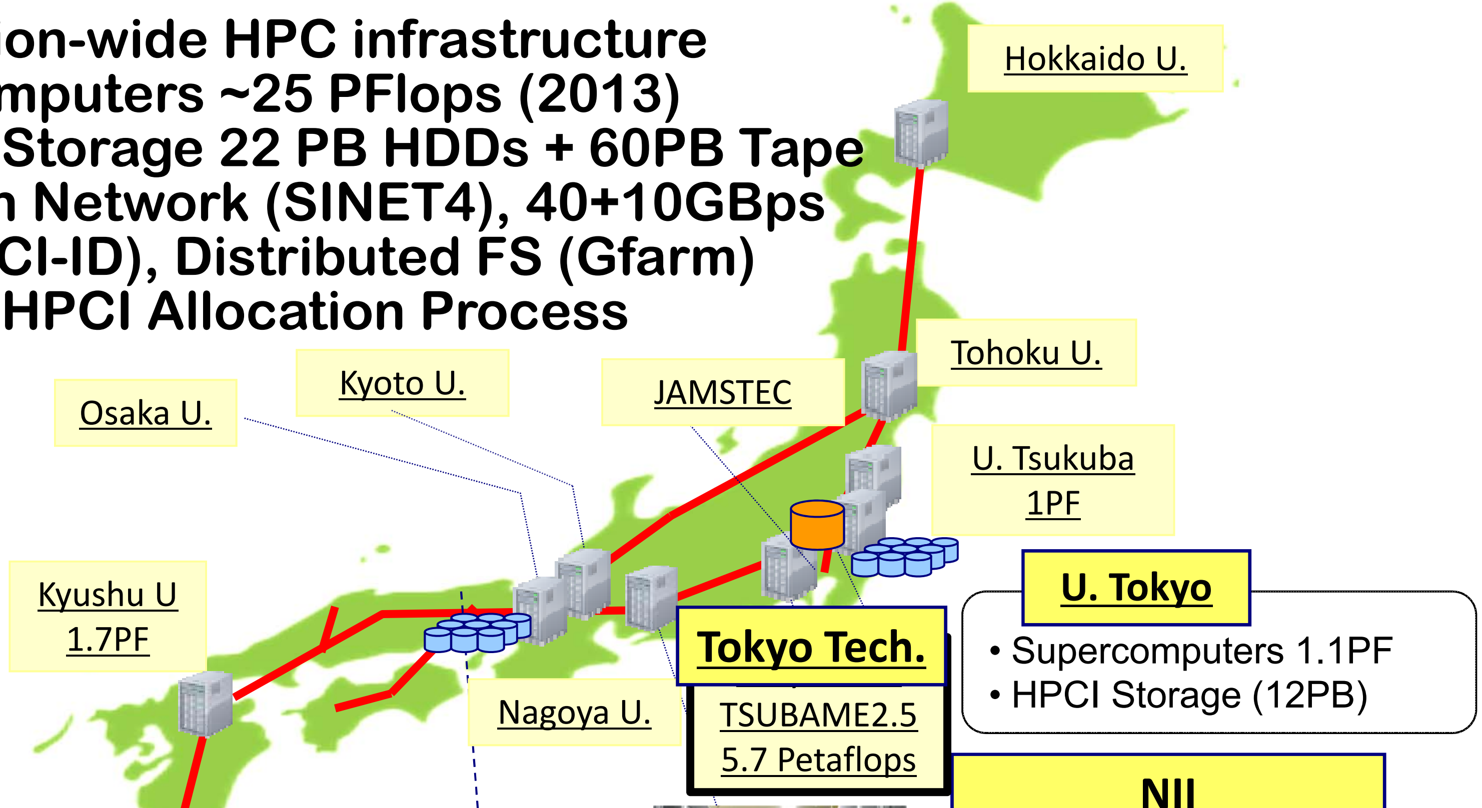
The Japanese Flagship 2020 “Post-K” Project

Bridging the gap to Exascale

Japan's High Performance Computing Infrastructure (HPCI)

HPCI: a nation-wide HPC infrastructure

- Supercomputers ~25 PFlops (2013)
- National Storage 22 PB HDDs + 60PB Tape
- Research Network (SINET4), 40+10GBps
- SSO (HPCI-ID), Distributed FS (Gfarm)
- National HPCI Allocation Process

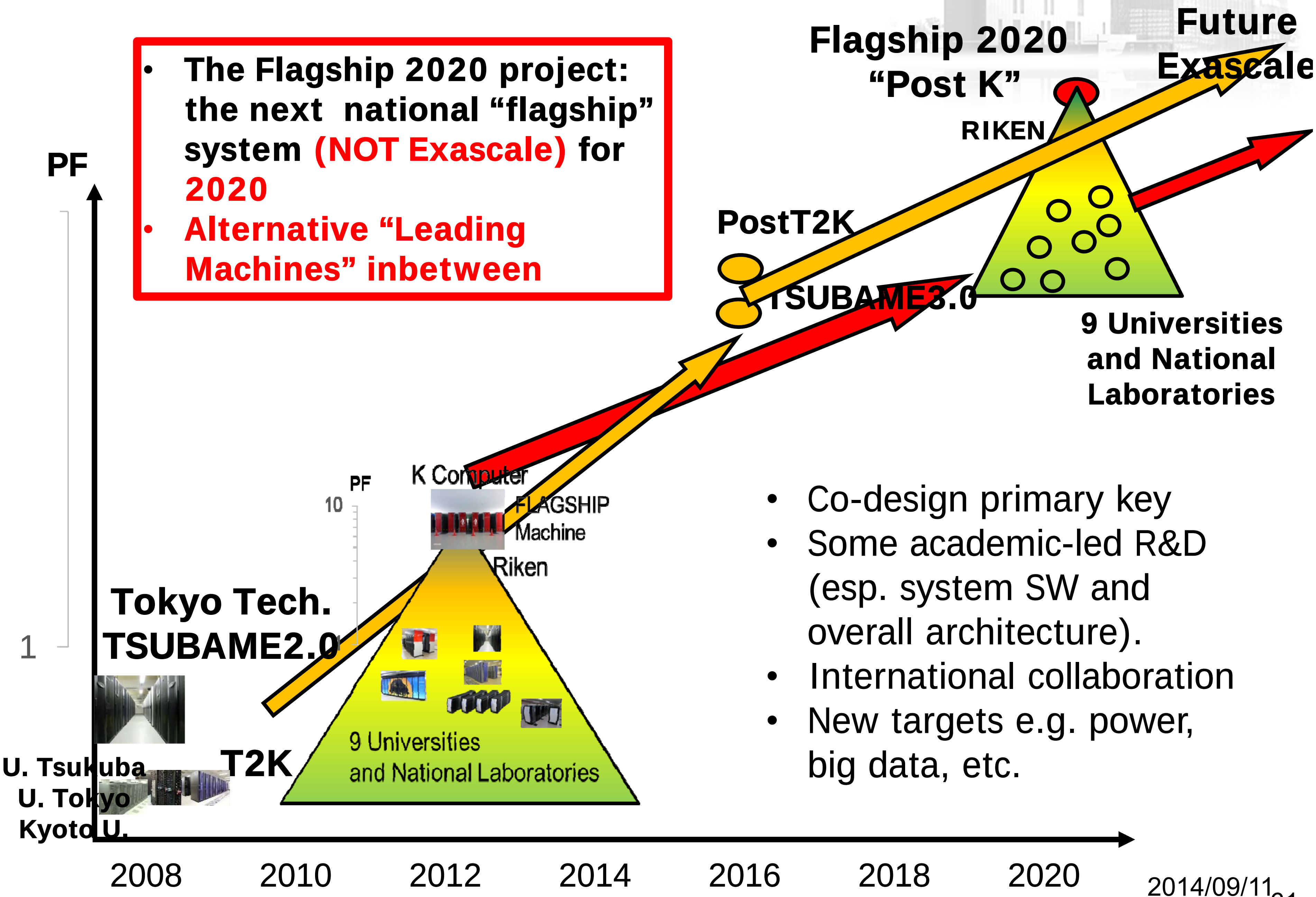


Riken AICS

- "K" computer 11Petaflops
- HPCI Storage (10PB)

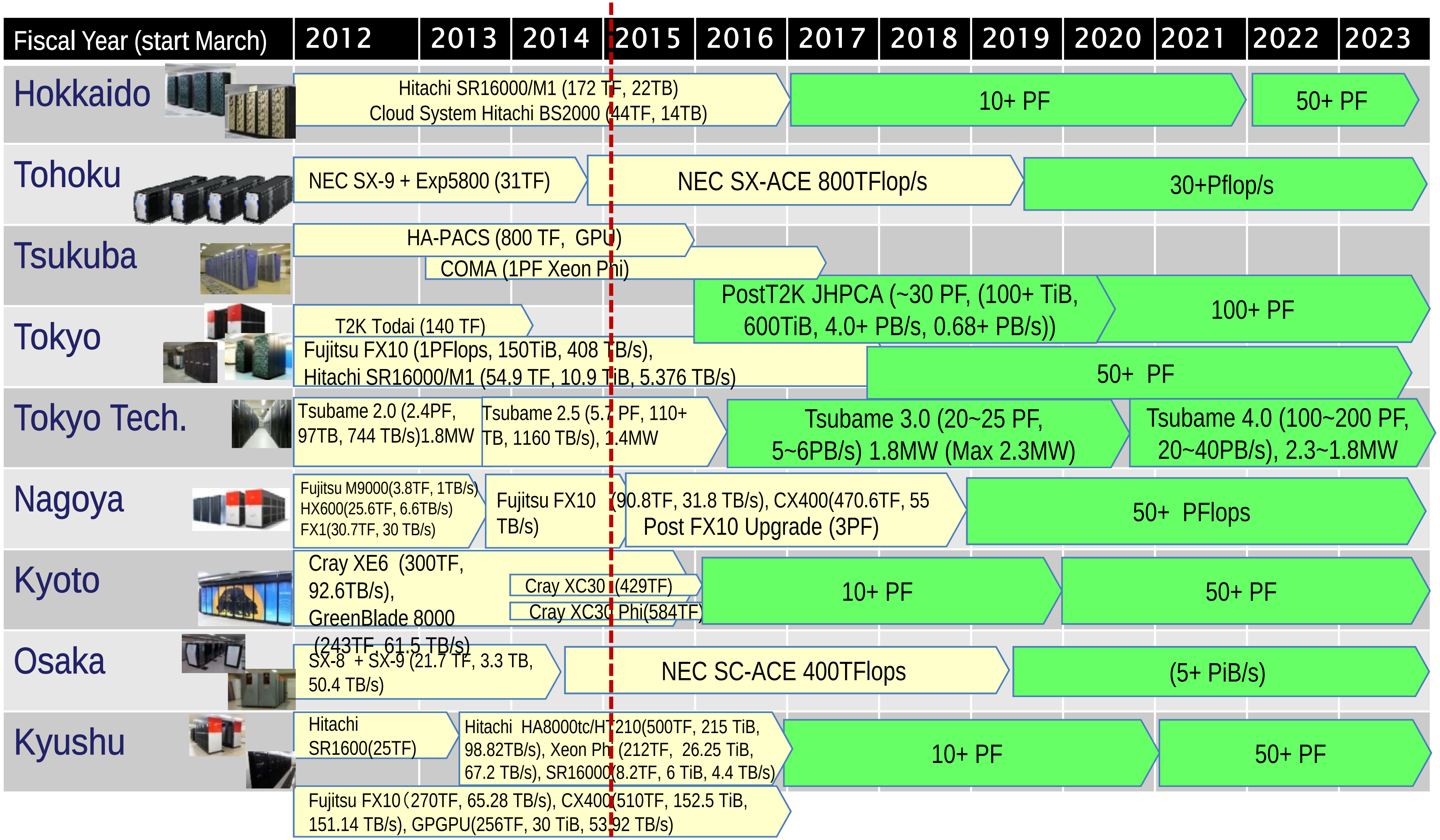
Towards the Next Flagship Machine & Beyond

- **The Flagship 2020 project: the next national "flagship" system (NOT Exascale) for 2020**
- **Alternative "Leading Machines" inbetween**



- Co-design primary key
- Some academic-led R&D (esp. system SW and overall architecture).
- International collaboration
- New targets e.g. power, big data, etc.

Japanese "Leading Machine" Candidates
Roadmap of the 9 HPCI University Centers

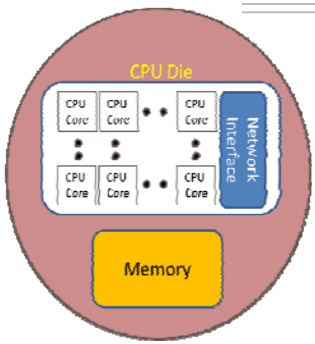


~17PF April 2015, Japan-wide ~40PF(incl. K)

Flagship 2020 "Post K" Supercomputer

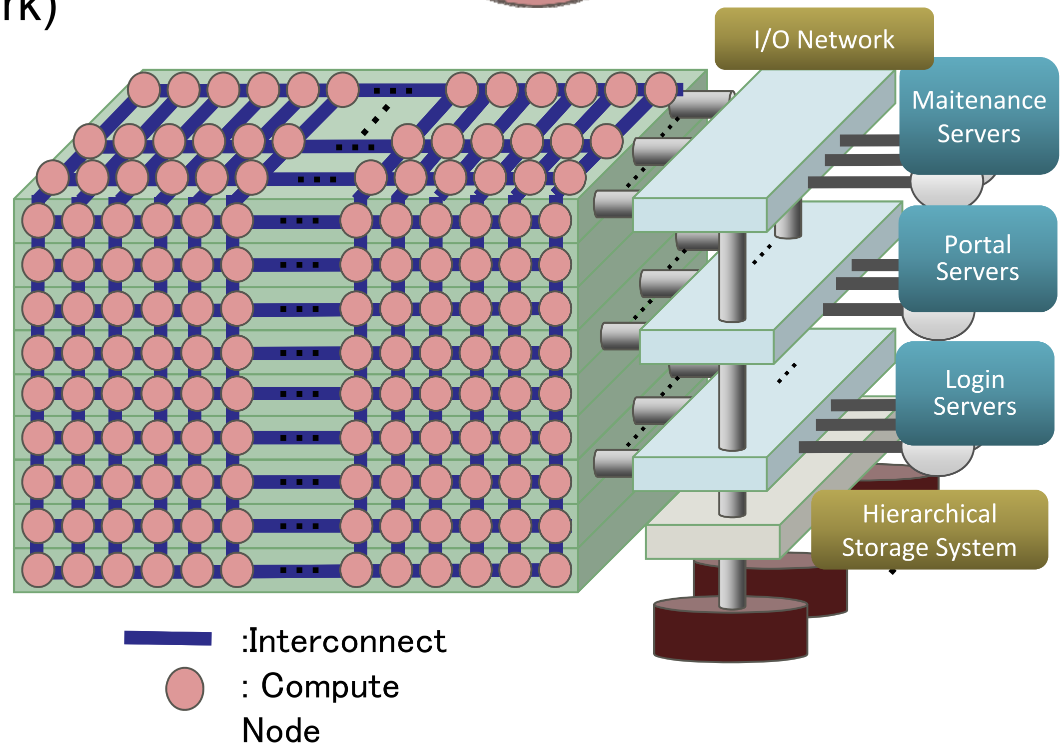
✓ CPU
  - Many-core with Interconnect interface integrated on chip
  - Power Knob feature for saving power

✓ Interconnect
  - TOFU (mesh/torus network)



Co-design may include:

- Compute Node Features
  - FP performance
  - Memory hierarchy, control, capacity, and bandwidth
- Network Performance
- I/O Performance



Current status of the Flagship 2020 project

- The project currently procured development of the basic design of the Flagship 2020 supercomputer

- In the specification RFP:
  - Constraints are:
    - Power capacity (about 30MW)
    - Space for system installation (in Kobe AICS building)
    - Budget (money) for development (NRE) and productio
    - … some degree of compatibility to the current K comp

  - The system should be designed to maximize the performance of applications in each computational science field.
    - "Co-design" is a keyword!
- FLASH! Fujitsu announced as the winner Oct 1st

Co-design elements in HPC systems

- Hardware/architecture
  - Node architecture (#core, #SIMD, etc...)
  - cache (size and bandwidth)
  - network (topologies, latency and bandwidth)
  - memory technologies (HBM and HMC, ...)
  - specialized hardware
  - #nodes
  - Storage, file systems
  - ... system configurations
- System software
  - Operating system for many core architecture
  - communication library (low level layer, MPI, PGAS)
  - Programming model and languages, DSLs,
  - Power, Reslience, ...
- Algorithm and math lib
  - Dense and Sparse solver
  - Eigen solver
  - ... Domain-specific lib and framework
- And, Applications!

Previous meeting
Fukuoka, Japan
Feb. 27-28
Adjacent Big Data Workshop
Feb. 26

Next meeting
Barcelona Spain, Jan 28-30

Charleston, South Carolina, USA, April 30- May 1

Exec Committee
Pete Beckman
Jean-Yves Berthou
Jack Dongarra
Yutaka Ishikawa
Satoshi Matsuoka
Philippe Ricoux

BIG DATA AND EXTREME-SCALE COMPUTING

http://www.exascale.org/bdec/

TSUBAME4 2021~ K-in-a-Box (Golden Box) BD/EC Convergent Architecture

1/500 Size, 1/150 Power, 1/500 Cost, x5 DRAM+ NVM Memory

10 Petaflops, 10 Petabyte Hiearchical Memory (K: 1.5PB), 10K nodes

50GB/s Interconnect (200-300Tbps Bisection BW)

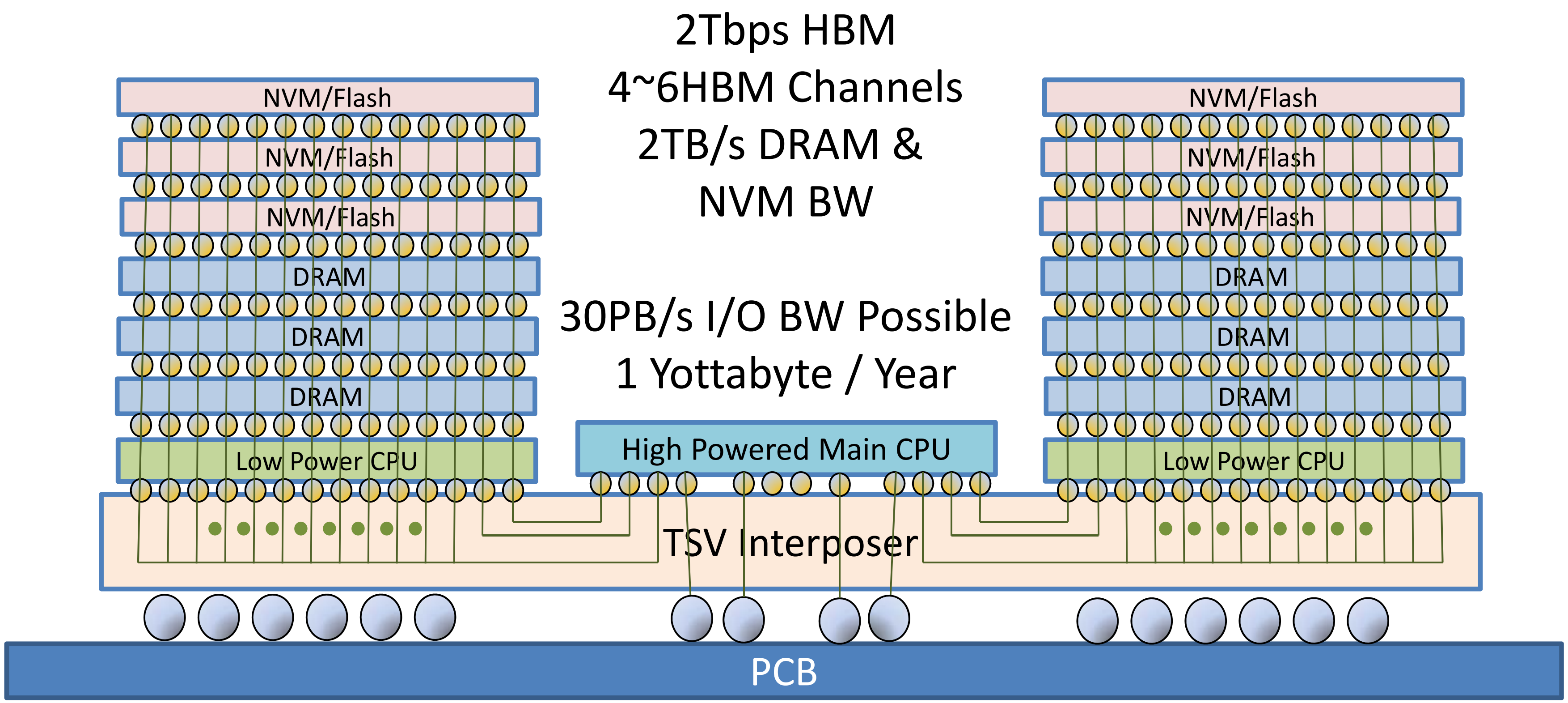
(Conceptually similar to HP “The Machine”)

Datacenter in a Box

Large Datacenter will become “Jurassic”

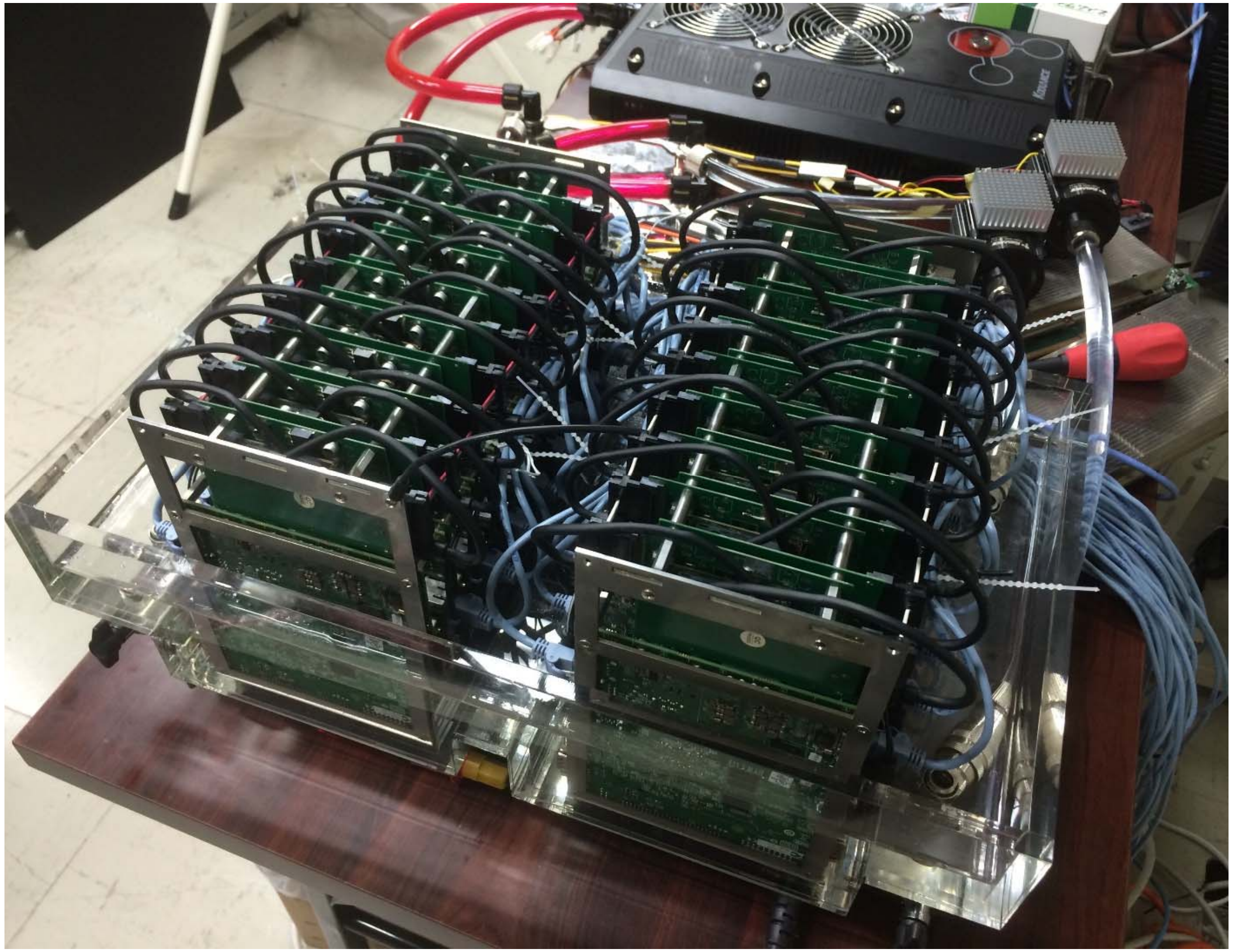
Tsubame 4: 2020- DRAM+NVM+CPU with 3D/2.5D Die Stacking

-The Ultimate Convergence of BD and EC-



Direct Chip-Chip Interconnect with planar VCSEL optics

GoldenBox Proto1 (NVIDIA K1-based) To be shown at SC14 Tokyo Tech. Booth…



- 36 Node Tegra K1, 11TFlops SFP
- ~700GB/s BW
- ~350Watts
- Integrated mSata SSD, ~7GB/s I/O
- Ultra dense, Oil immersive cooling
- Same SW stack as TSUBAME

2022: x10 Flops, x10 Mem Bandwidth, silicon photonics, x10 NVM, x10 node density