Convergence of Supercomputing and Extreme Big Data on the TSUBAME Supercomputer Exascale

松岡聡・東工大
Satoshi Matsuoka
Tokyo Institute of Technology

OpenSFS @ Tokyo
2013/10/17

Themes of the Day…

- How do you respond to the followings?
- "We don't need to invest in all that supercomputer R&D stuff; we invest into clouds, mobiles, etc., for big data and we will just leverage off those…"
- "Sure, supercomputers are pretty big, but giants *Google/Amazon/…* will have enough resource in the cloud for big data, so we will just use those…"

The current "Big Data" are not really that Big…

- Typical "real" definition: "Mining people's privacy data to make money"
- Corporate data are usually in data warehoused silo -> limited volume, in Gigabytes~Terabytes, seldom Petabytes.
- Processing involve simple O(n) algorithms, or those that can be accelerated with DB-inherited indexing algorithms
- Executed on re-purposed commodity "web" servers linked with 1Gbps networks running Hadoop/HDFS
- Vicious cycle of stagnation in innovations…
- **Breaking Down of Corporate Silos⇒ Convergence with Supercomputing with Extreme Big Data**

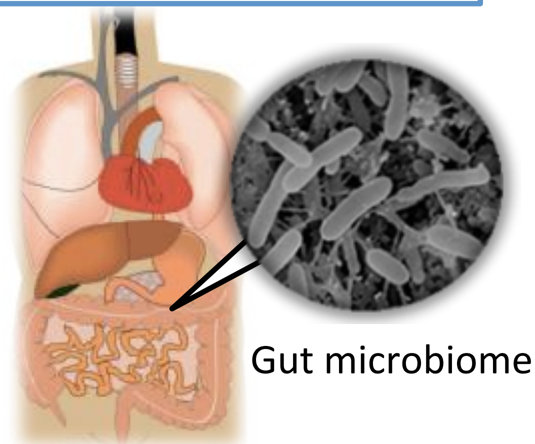
We will have tons of unknown genes

Metagenome analysis

[Slide Courtesy Yutaka Akiyama @ Tokyo Tech.]

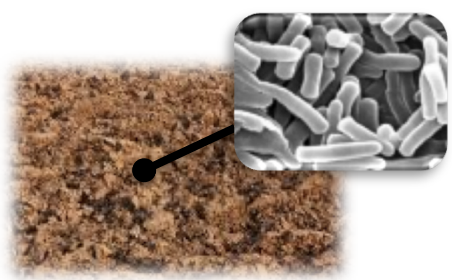
- Directly sequencing uncultured microbiomes obtained from target environment and analyzing the sequence data
  - Finding novel genes from unculturable microorganism
  - Elucidating composition of species/genes of environments

Examples of microbiome

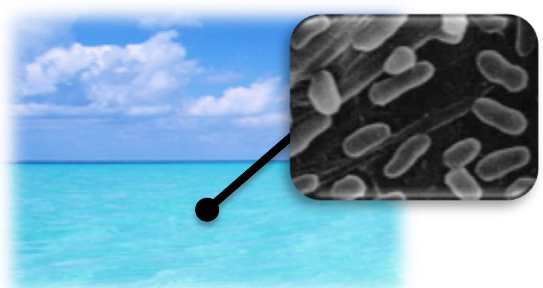


Gut microbiome

Human body



Soil

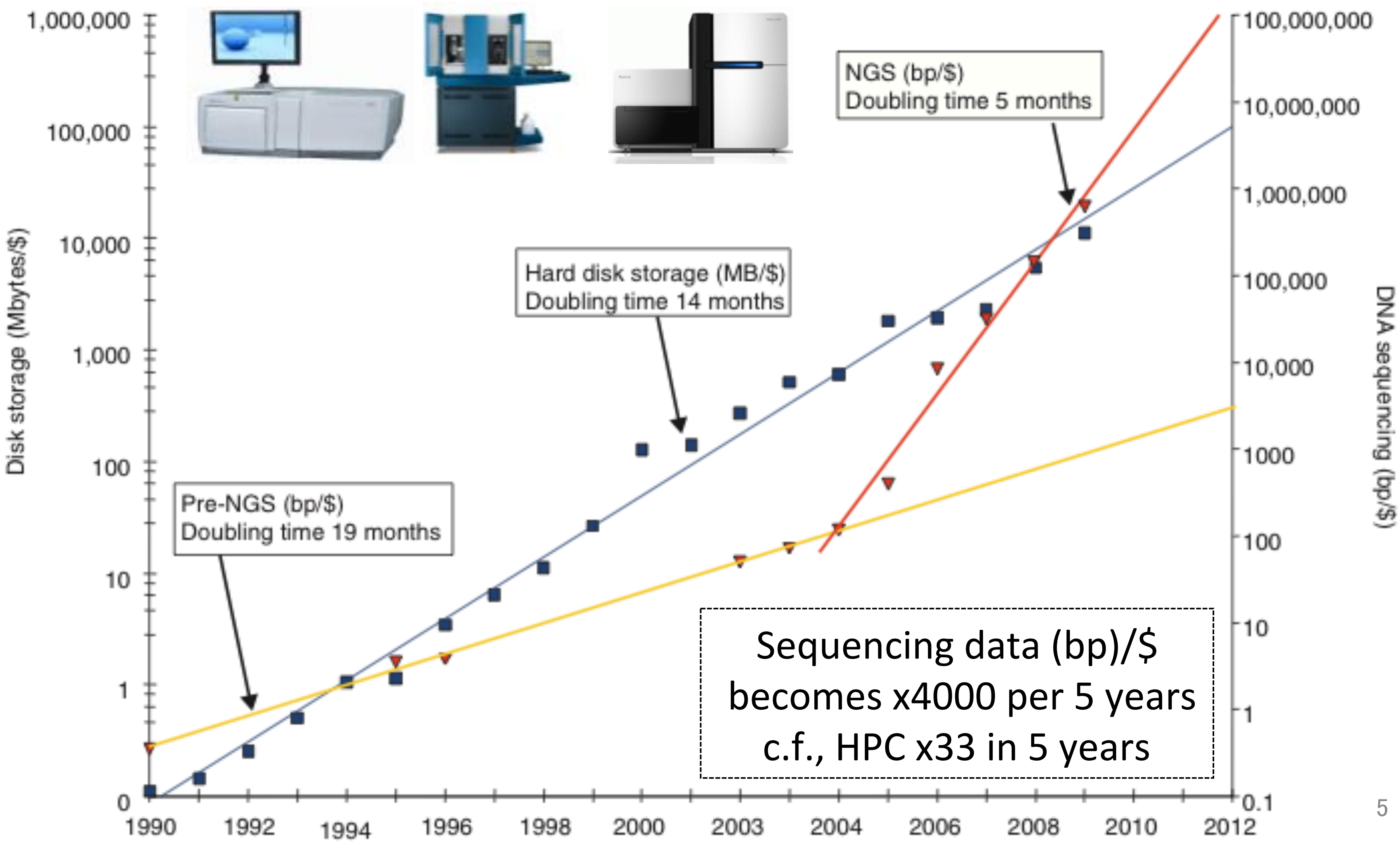


Sea

Extreme Big Data in Genomics

Impact of new generation sequencers

[Slide Courtesy Yutaka Akiyama @ Tokyo Tech.]



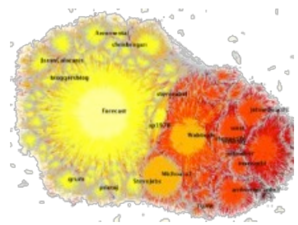
Sequencing data (bp)/$
becomes x4000 per 5 years
c.f., HPC x33 in 5 years

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

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

Slide courtecy David A. Bader @ Georgia Tech

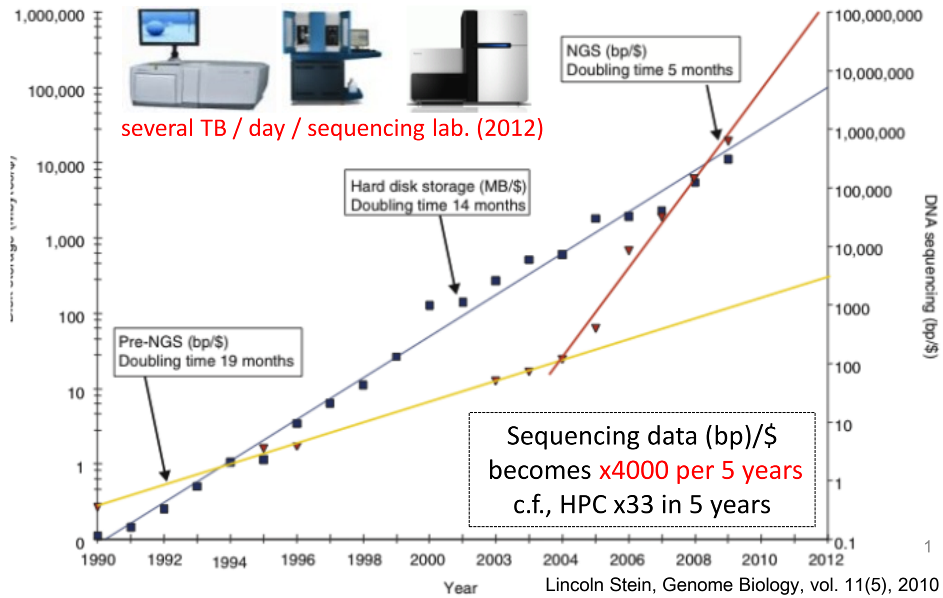
- Facebook:
  - ~1 billion users
  - average 130 friends
  - 30 billion pieces of content shared / month
- Twitter:
  - 500 million active users
  - 340 million tweets / day
- Internet – 100s of exabytes / year
  - 300 million new websites per year
  - 48 hours of video to You Tube per minute
  - 30,000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



Extreme Big Data in Genomics

ɔact of new generation sequencers

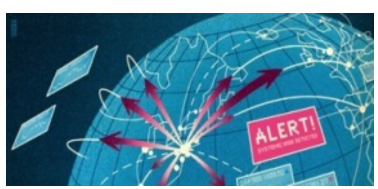
[Slide Courtesy Yutaka Akiyama @ Tokyo Tech.]



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

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**



Future "Extreme Big Data"

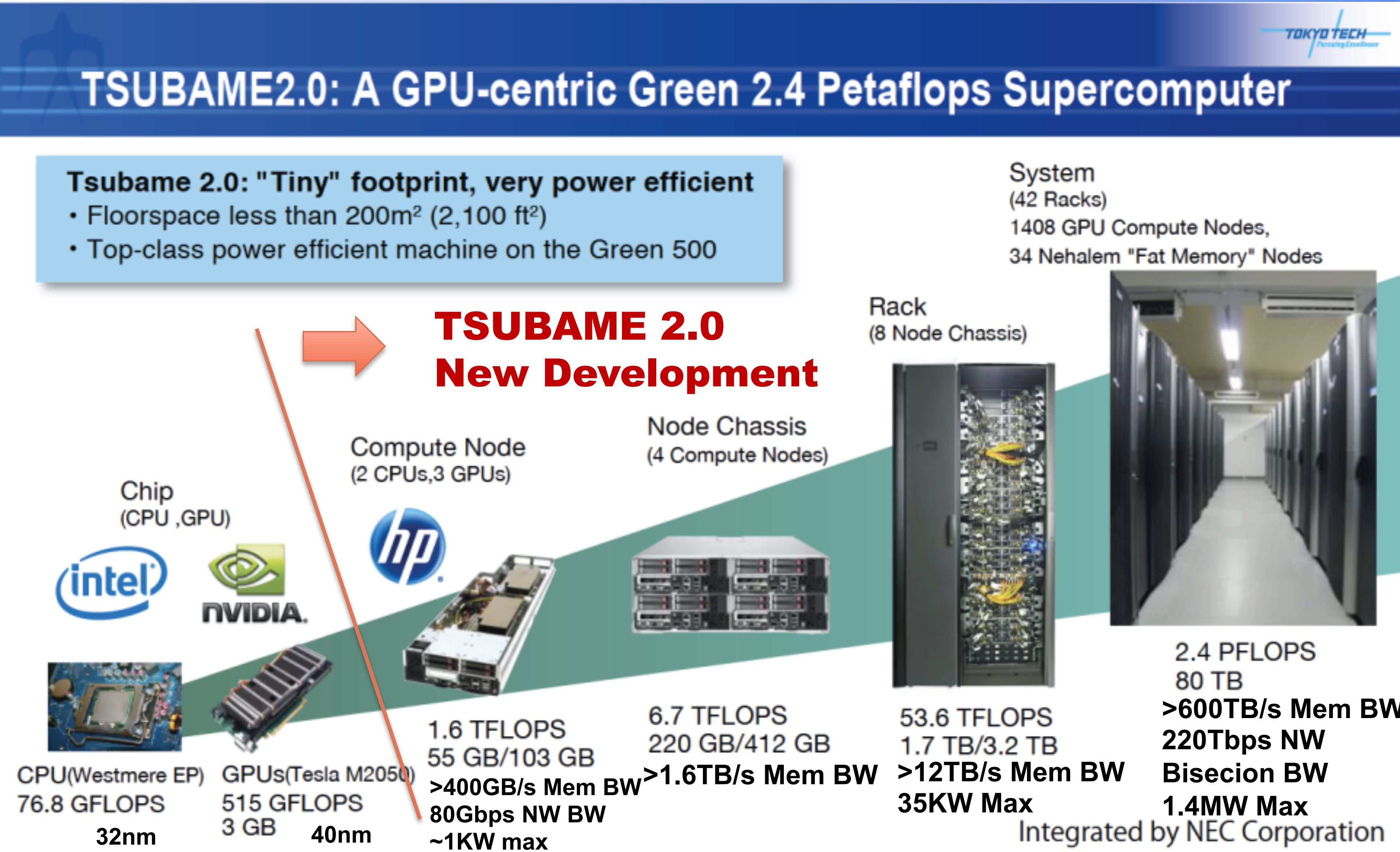
- **<u>NOT mining Tbytes Silo Data</u>**
- **Peta~Zetabytes of Data**
- **Ultra High-BW Data Stream**
- **Highly Unstructured, Irregular**
- **Complex correlations between data from multiple sources**
- **Extreme Capacity, Bandwidth, Compute All Required**

“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), …
- Will become the NORM for competitiveness reasons.

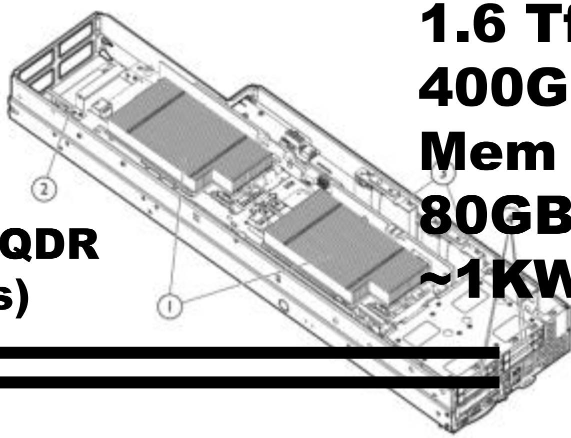
TSUBAME2.0 Nov. 1, 2010

"The Greenest Production Supercomputer in the World"



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)

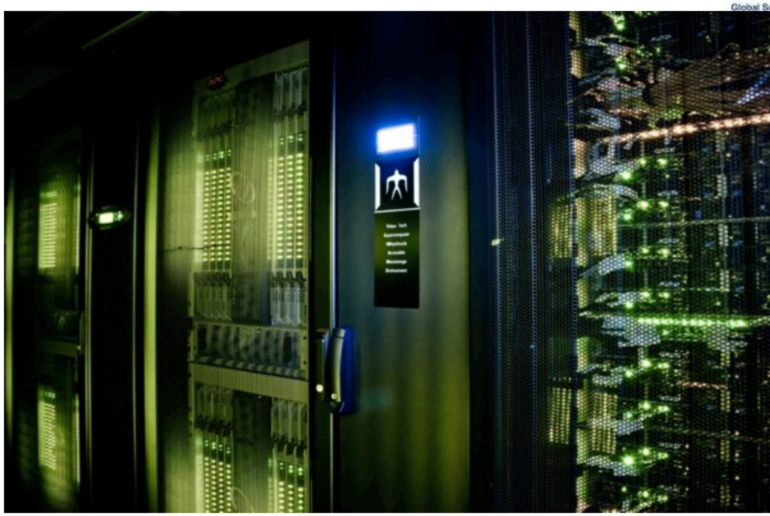
GPU: NVIDIA Fermi M2050 x 3
515GFlops, 3GByte memory /GPU
CPU: Intel Westmere-EP 2.93GHz x2 (12cores/node)
Multi I/O chips, 72 PCI-e (16 x 4 + 4 x 2) lanes --- 3GPUs + 2 IB QDR
Memory: 54, 96 GB DDR3-1333
SSD：60GBx2, 120GBx2

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

2010: TSUBAME2.0 as No.1 in Japan

東京工業大学
Tokyo Institute of Technology

GSIC



Total 2.4 Petaflops
#4 Top500, Nov. 2010

>

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

All Other Japanese Centers on the Top500 COMBINED 2.3 PetaFlops

TOP500® SUPERCOMPUTER SITES

TSUBAME Wins Awards…

"Greenest Production Supercomputer in the World"

the Green 500

Nov. 2010, June 2011

(#4 Top500 Nov. 2010)

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

Special Achievements in Scalability and Time-to-Solution

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

TSUBAME Wins Awards…

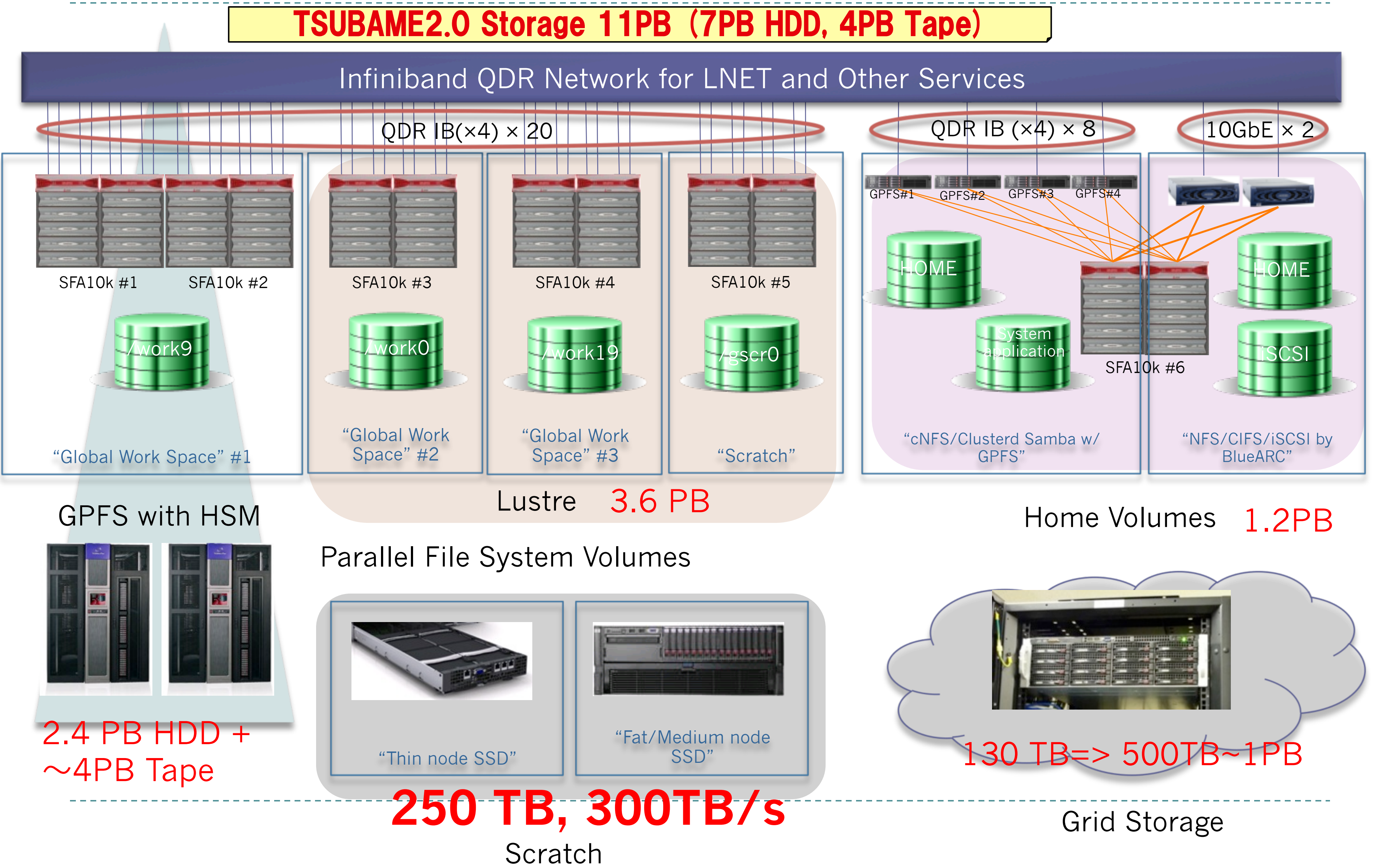
Commendation for Sci &Tech by Ministry of Education 2012 (文部科学大臣表彰)

Prize for Sci & Tech, Development Category

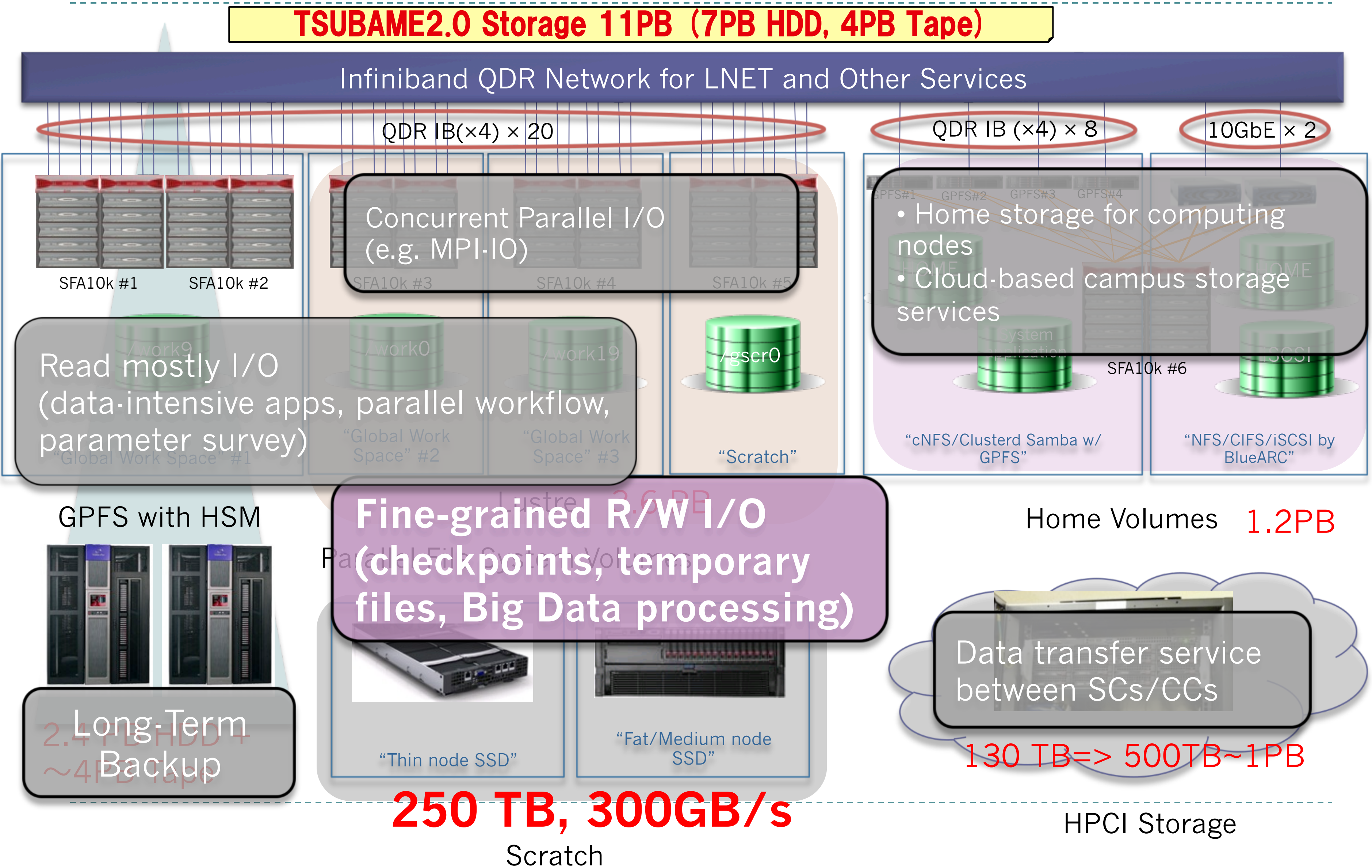
Development of Greenest Production Peta-scale Supercomputer

Satoshi Matsuoka, Toshio Endo, Takayuki Aoki

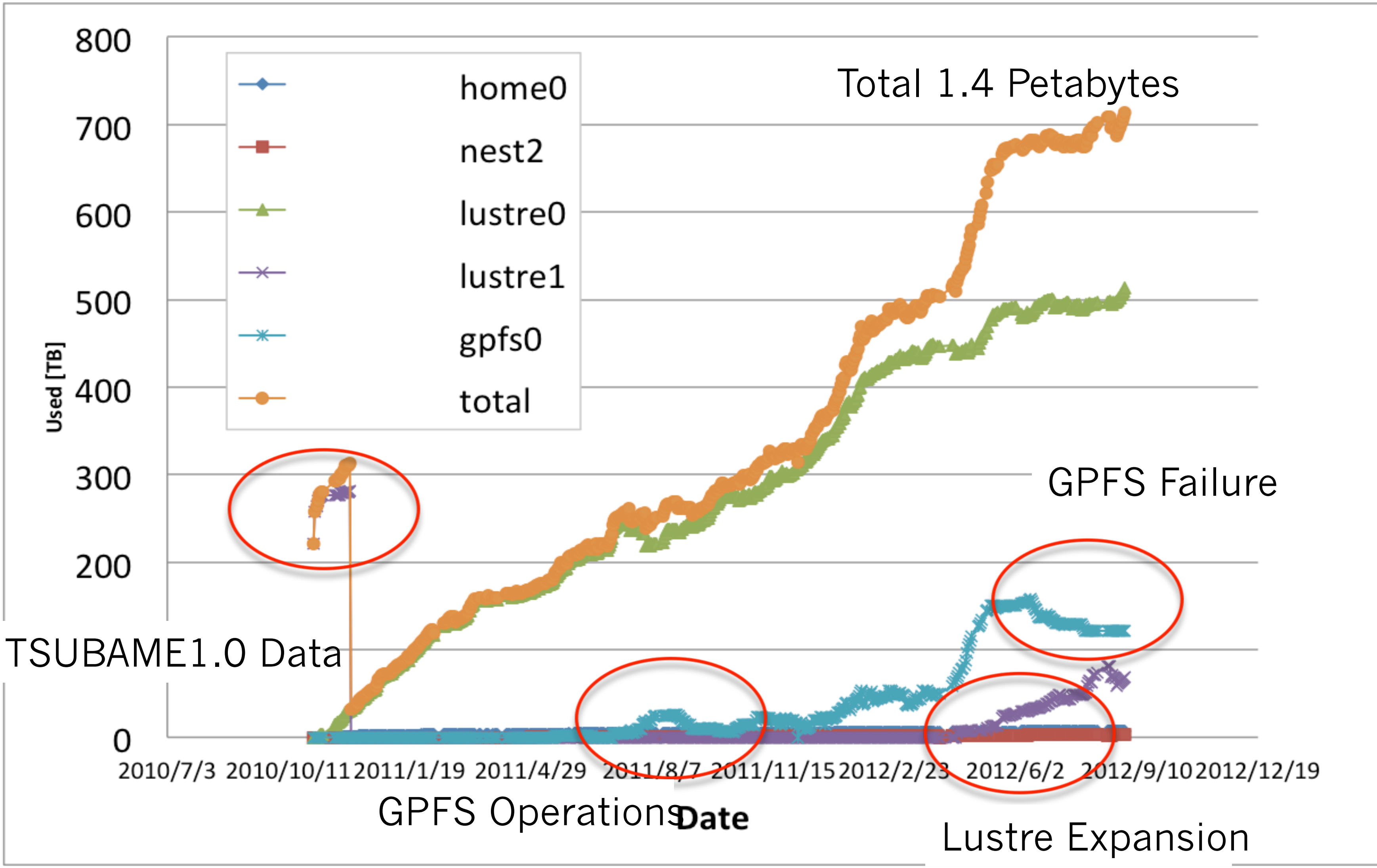
TSUBAME2.0 Storage Overview



TSUBAME2.0 Storage Overview

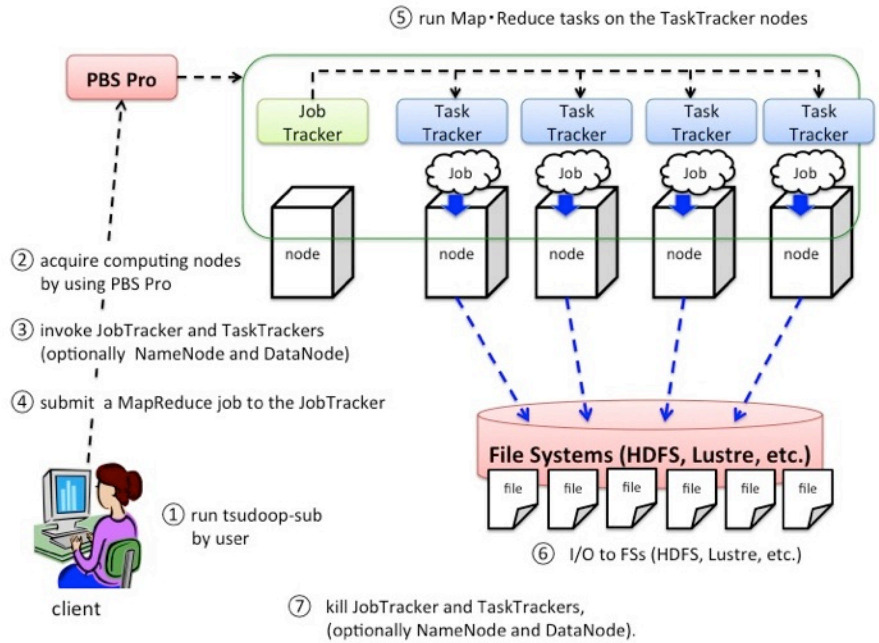


TSUBAME2.0 Storage Usage

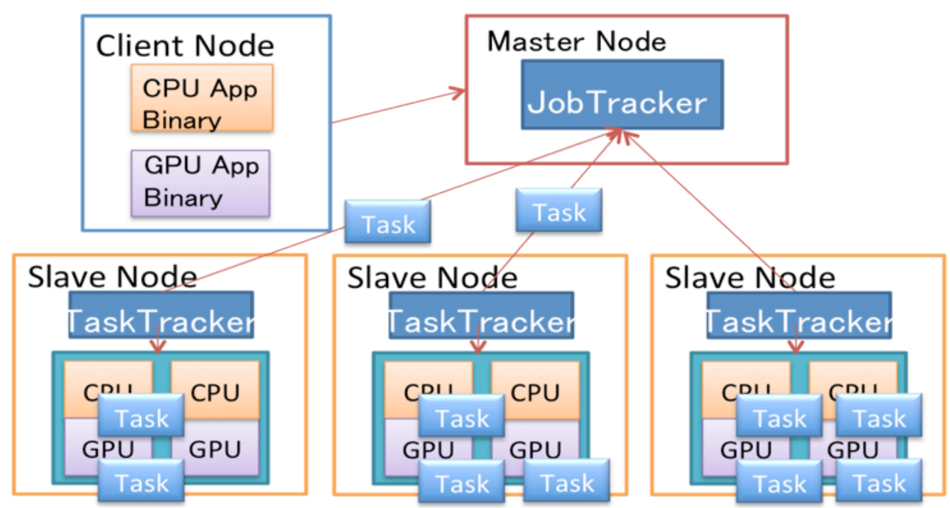


Hadoop on TSUBAME (Tsudoop)

- Script-based invocation
  - acquire computing nodes via PBS Pro
  - deploy a Hadoop environment on the fly (incl. HDFS)
  - execute a user MapReduce jobs
- Various FS support
  - HDFS by aggregating local SSDs
  - Lustre, GPFS (to appear)

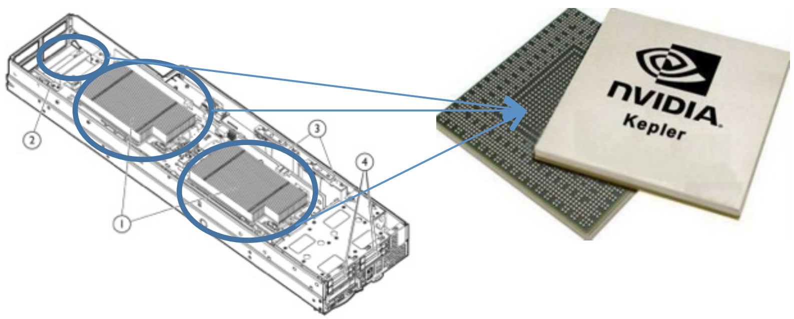


- Customized Hadoop for executing CUDA programs (experimental)
  - Hybrid Map Task Scheduling
    - Automatically detects map task characteristics by monitoring
    - Scheduling map tasks to minimize overall MapReduce job execution time
    - Extension of Hadoop Pipes features



Towards TSUBAME 3.0
Interim Upgrade TSUBAME2.0 to 2.5 (Early Fall 2013)

- Upgrade the TSUBAME2.0s GPUs NVIDIA Fermi M2050 to Kepler K20X



TSUBAME2.0 Compute Node
Fermi GPU 3 x 1408 = 4224 GPUs

SFP/DFP peak from 4.8PF/2.4PF => 17PF/5.7PF

c.f. The K Computer 11.2/11.2

Acceleration of Important Apps
Considerable Improvement
Summer 2013

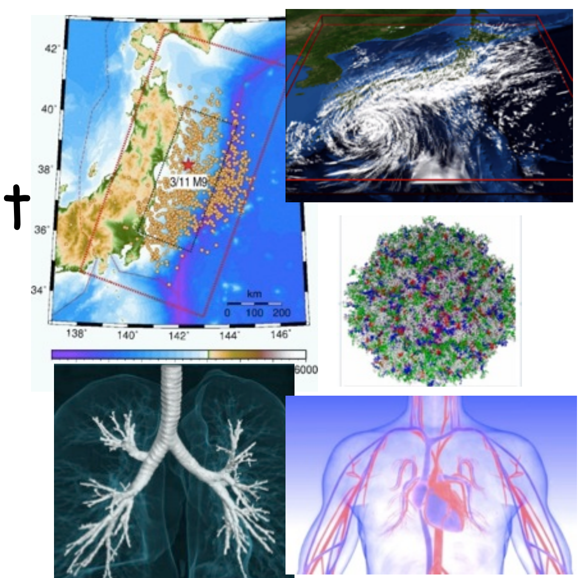
2011/12/20 22:03

Service List

service	assigned nodes		running jobs		users
S	100%	352 / 352 nodes	[illegible]	[illegible]	[illegible]
S96	100%	41 / 41 nodes	53%	41 / 76 jobs	4
G	99%	475 / 477 nodes	100%	62 / 62 jobs	12
V	83%	364 / 437 nodes	80%	1531 / 1904 jobs	[illegible]
L128	100%	10 / 10 nodes	66%	10 / 15 jobs	1
L128F	100%	10 / 10 nodes	71%	[illegible]	[illegible]
L256	37%	3 / 8 nodes	100%	[illegible]	[illegible]
L512	100%	2 / 2 nodes	100%	2 / 2 jobs	1
H/X	93%	301 (+ 95) / 420 nodes	100%	87 / 87 jobs	8
ALL	93%	1558 (+ 95) / 1757 nodes	73%	1921 / 2615 jobs	93

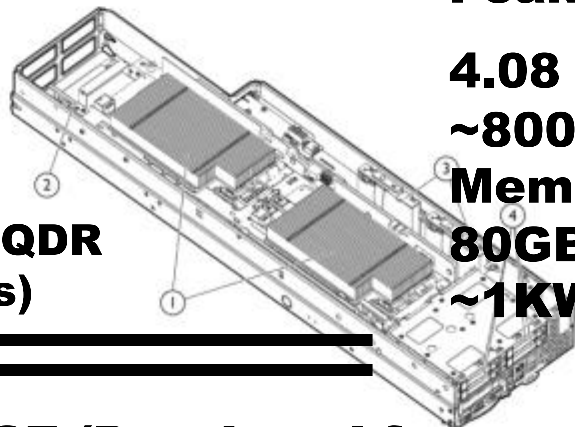
Significant Capacity Improvement at low cost & w/o Power Increase

TSUBAME3.0 2H2015



TSUBAME2.0⇒2.5 Thin Node Upgrade

Thin Node



Infiniband QDR x2 (80Gbps)

Peak Perf.

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

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

Productized as HP ProLiant
SL390s
Modified for TSUABME2.5

NVIDIA Fermi M2050 1039/515 GFlops

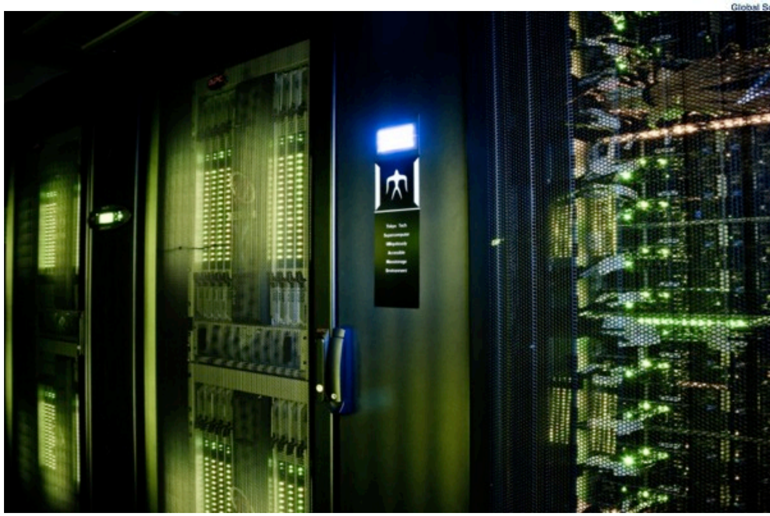
NVIDIA Kepler K20X 3950/1310 GFlops

	TSUBAME2.0	TSUBAME2.5
Thin Node x 1408 Units		
Node Machine	HP Proliant SL390s	← No Change
CPU	Intel Xeon X5670 (6core 2.93GHz, Westmere) x 2	← No Change
GPU	NVIDIA Tesla M2050 x 3 ● 448 CUDA cores (Fermi) ➢ SFP 1.03TFlops ➢ DFP 0.515TFlops ● 3GiB GDDR5 memory ● 150GB Peak, ~90GB/s STREAM Memory BW	NVIDIA Tesla K20X x 3 ● 2688 CUDA cores (Kepler) ➢ SFP 3.95TFlops ➢ DFP 1.31TFlops ● 6GiB GDDR5 memory ● 250GB Peak, ~180GB/s STREAM Memory BW
Node Performance (incl. CPU Turbo boost)	● SFP 3.40TFlops ● DFP 1.70TFlops ● ~500GB Peak, ~300GB/s STREAM Memory BW	● SFP 12.2TFlops ● DFP 4.08TFlops ● ~800GB Peak, ~570GB/s STREAM Memory BW
TOTAL System		
Total System Performance	● SFP 4.80PFlops ● DFP 2.40PFlops ● Peak ˜0.70PB/s, STREAM ˜0.440PB/s Memory BW	● **SFP 17.1PFlops (x3.6)** ● **DFP 5.76PFlops (x2.4)** ● Peak ~1.16PB/s, STREAM ~0.804PB/s Memory BW (x1.8)

2013: TSUBAME2.5 No.1 in Japan in Single Precision FP, 17 Petaflops

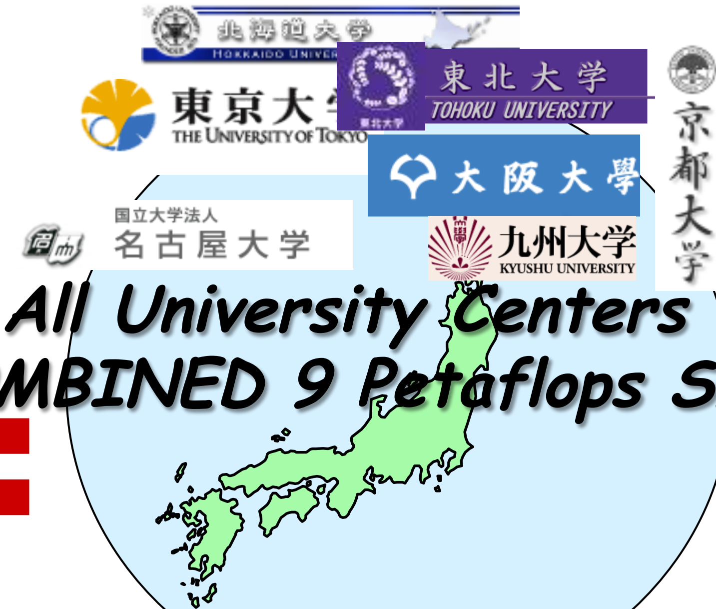
東京工業大学
Tokyo Institute of Technology

GSIC



Total
17.1 Petaflops SFP
5.76 Petaflops DFP

~=



All University Centers COMBINED 9 Petaflops SFP

K Computer
11.4 Petaflops SFP/DFP

Linpack Benchmark

- Linpack: Dense matrix solver by LU decomposition with pivotting
  - Used in Top500/Green500 supercomputer ranking!
- On TSUBAME2.5, we adopted “In-core” algorithm, where the whole matrix data are placed on GPU device memory
  - K20X on T2.5 has 2x larger memory than M2050 on T2.0
  - PCIe communication has relatively larger effects

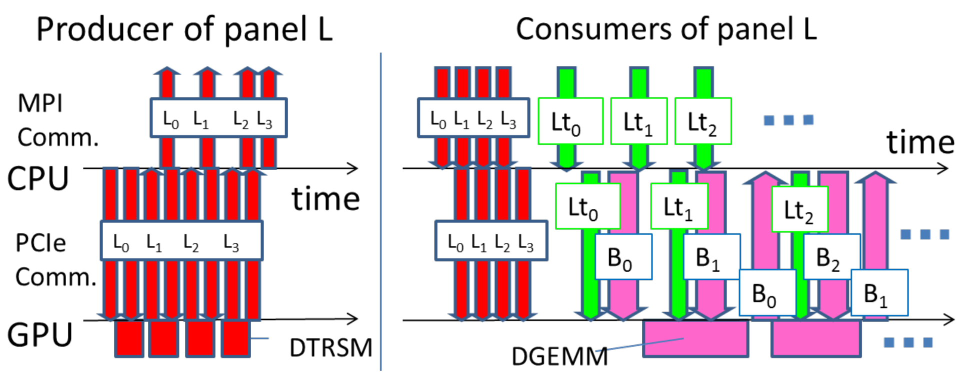
	TSUBAME2.0	TSUBAME2.5
Algorithm	Out-of-core	In-core
N (matrix size)	2,490,368	1,760,000
NB (block size)	1024	192
Speed (PFlops)	**1.192** **2.39x**	**2.843**
Rank in Top500	**No. 4** in 11/2010	TBA in 11/2013
Power (MWatt)	1.244	0.958
Speed/Power (GFlops/Watt)	**0.958** **>2.50x**	**>2.40**
Rank in Green500	**No. 2** in 11/2010	TBA in 11/2013

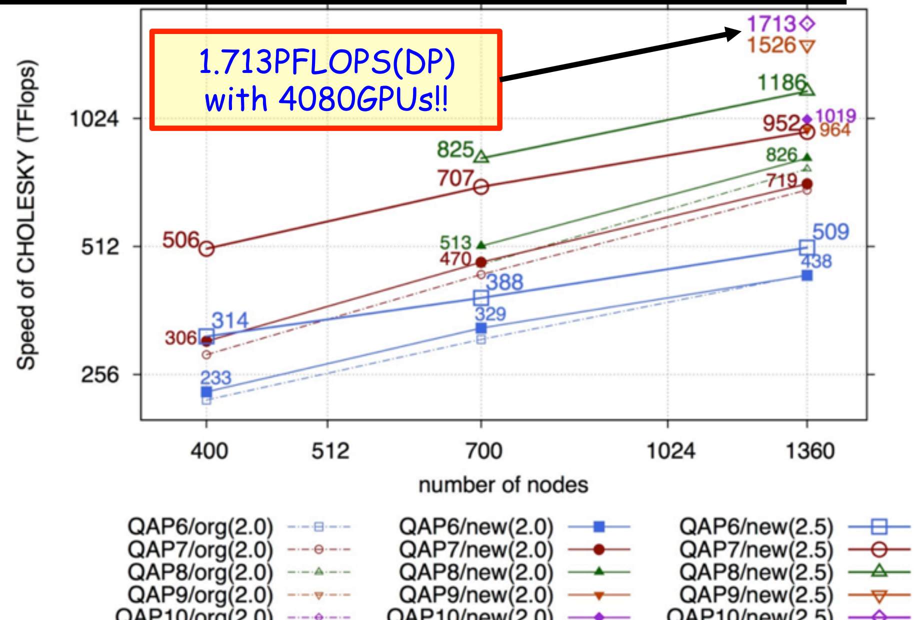
High-Performance General Solver for Extremely Large-scale Semidefinite Programming Problems [Fujisawa]

1. Mathematical Programming : one of the most important mathematical programming
2. Many Applications : combinatorial optimization, control theory, structural optimization, quantum chemistry, sensor network location, data mining, etc.

Parallel Algorithm of Cholesky Factorization

GPU computation, PCI-e communication, and MPI communication are overlapped

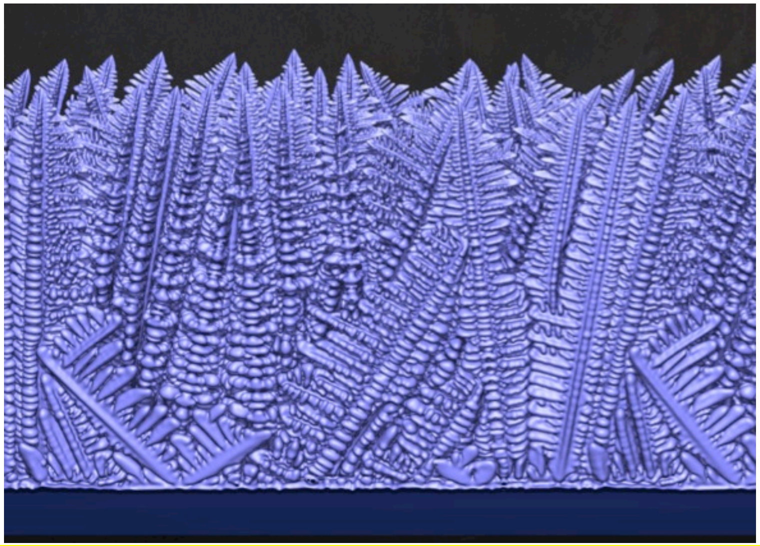




- SDPARA is a parallel implementation of the interior-point method for Semidefinite Programming
  Parallel computation for two major bottleneck parts
  - ELEMENTS ⇒ Computation of Schur complement matrix (SCM)
  - CHOLESKY ⇒ Cholesky factorization of Schur complement matrix (SCM)
- SDPARA could attain high scalability using 16,320 CPU cores on the TSUBAME 2.5 supercomputer and some techniques of processor affinity and memory interleaving when the computation of SCM (ELEMENTS) constituted a bottleneck.
- With 4,080 NVIDIA GPUs on the TSUBAME 2.0 & 2.5 supercomputer, our implementation achieved 1.019 PFlops(TSUBAME 2.0) & 1.713PFlops(TSUBAME 2.5) in double precision for a large-scale problem (CHOLESKY) with over two million constraints.

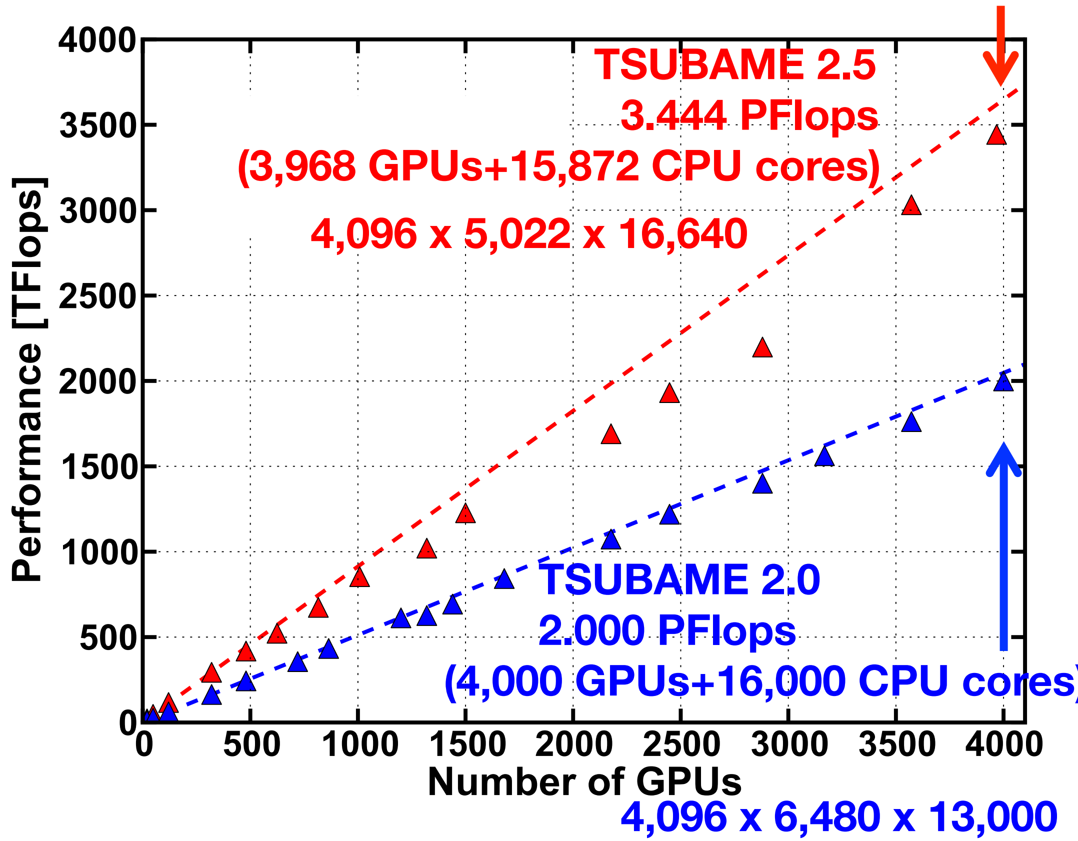
Phase-field simulation for Dendritic Solidification [Shimokawabe, Aoki et. al.] Gordon Bell 2011 Winner

Weak scaling on TSUBAME (Single precision)
Mesh size（1GPU+4 CPU cores）:4096 x 162 x 130



Developing lightweight strengthening material by controlling microstructure

→ **Low-carbon society**

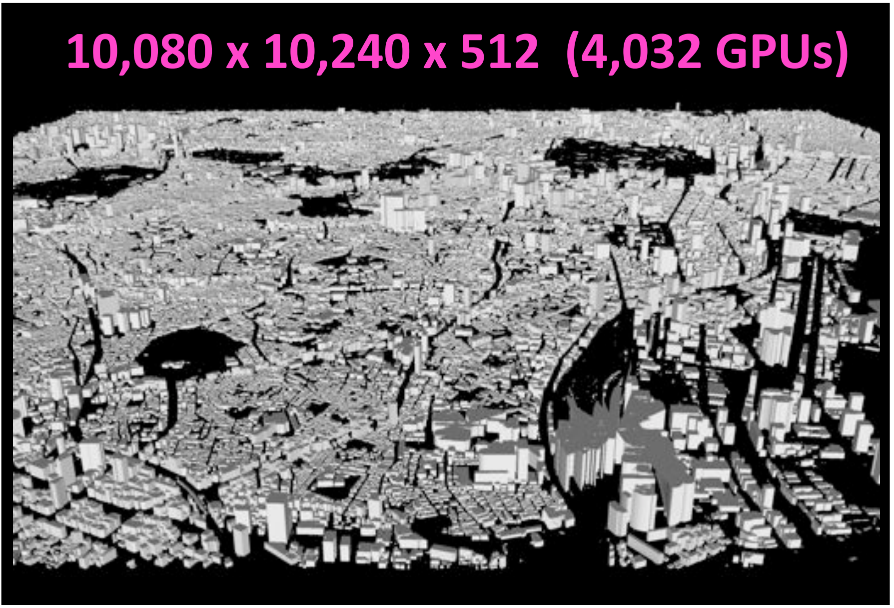


- Peta-Scale phase-field simulations can simulate the multiple dendritic growth during solidification required for the evaluation of new materials.
- 2011 ACM Gordon Bell Prize Special Achievements in Scalability and Time-to-Solution

Peta-scale stencil application :

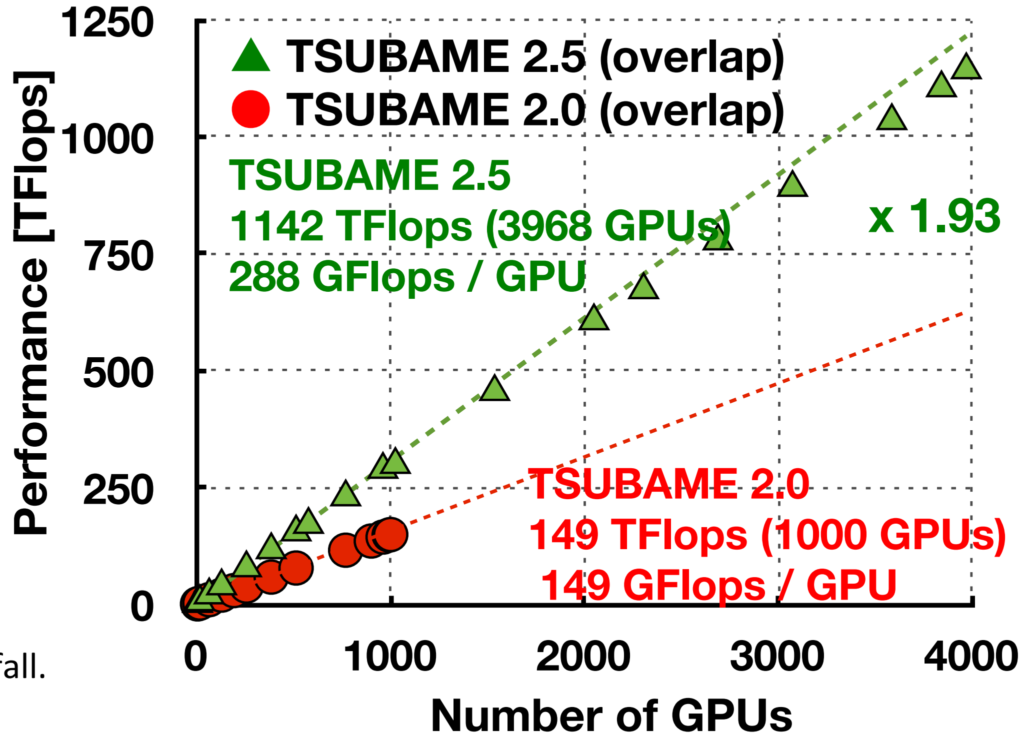
A Large-scale LES Wind Simulation using Lattice Boltzmann Method [Onodera, Aoki]

Large-scale Wind Simulation for a 10km x 10km Area in Metropolitan Tokyo



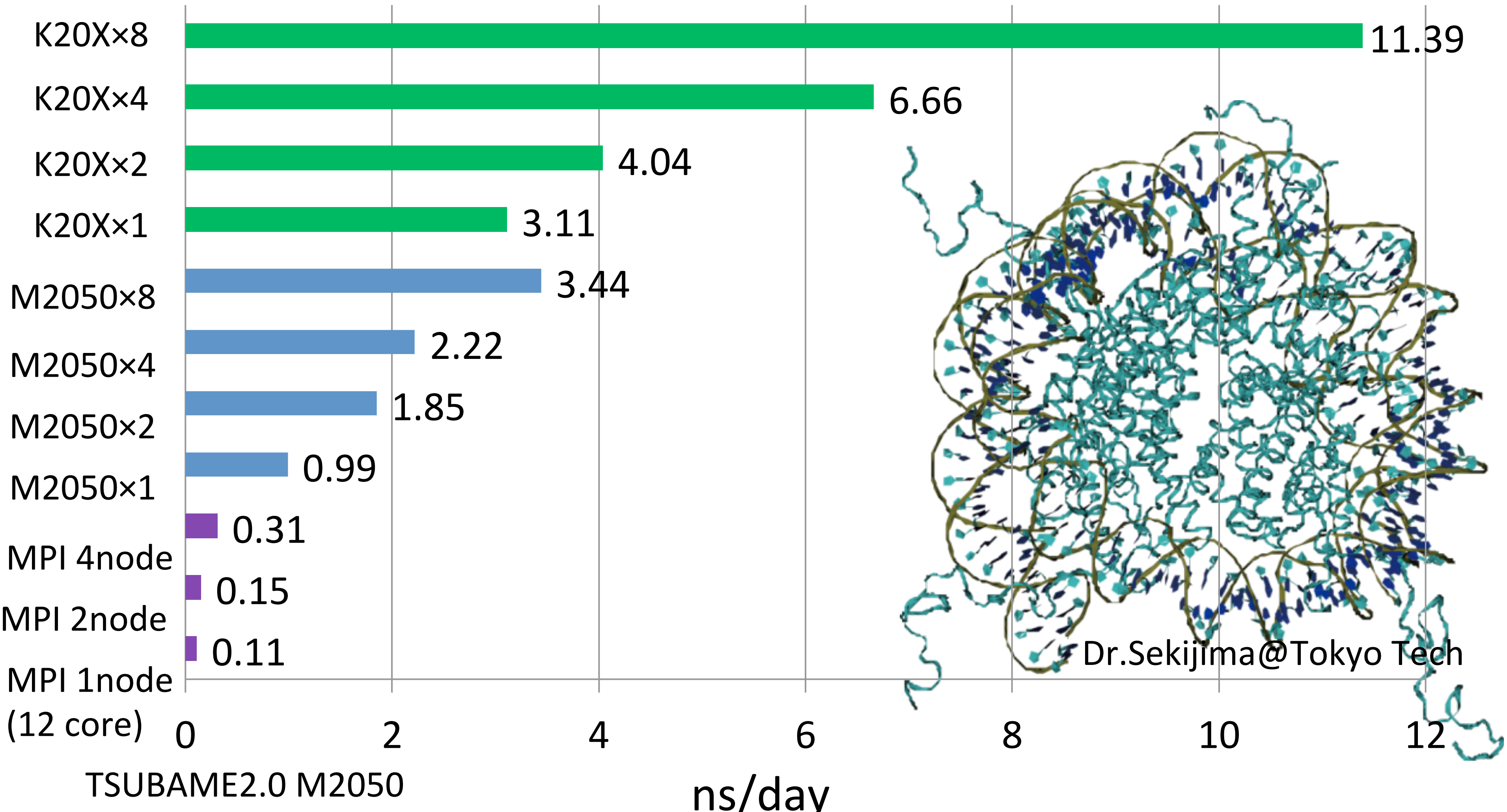
The above peta-scale simulations were executed as the TSUBAME Grand Challenge Program, Category A in 2012 fall.

Weak scalability in single precision (N = 192 x 256 x 256)



- The LES wind simulation for the area 10km × 10km with 1-m resolution has never been done before in the world.
- We achieved 1.14 PFLOPS using 3968 GPUs on the TSUBAME 2.5 supercomputer.

AMBER pmemd benchmark
Nucleosome = 25,095 atoms

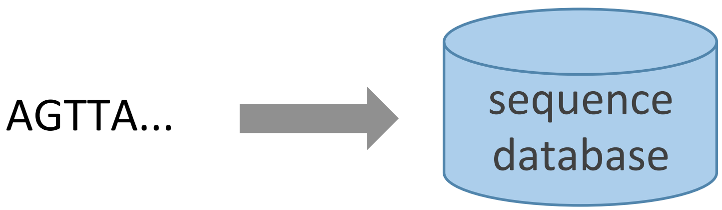


TSUBAME2.0 M2050
TSUBAME2.5 K20X

GHOSTM: A GPU-Accelerated HOmology Search Tool for Metagenomics

Suzuki *et al.*, *PLoS ONE*, 7(5): e36060. (2012)

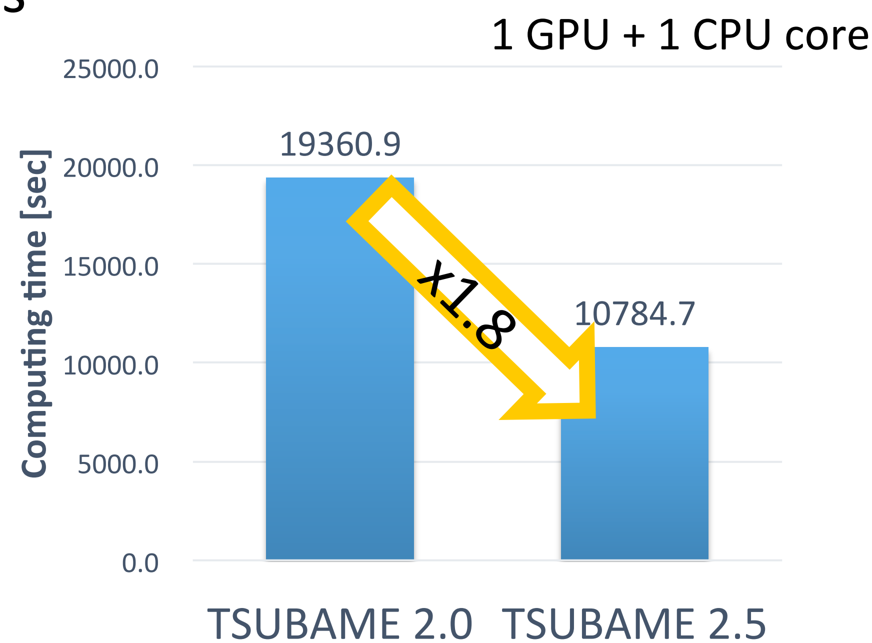
Homology search is one of important methods to annotate DNA sequences



Search similar sequences

Data

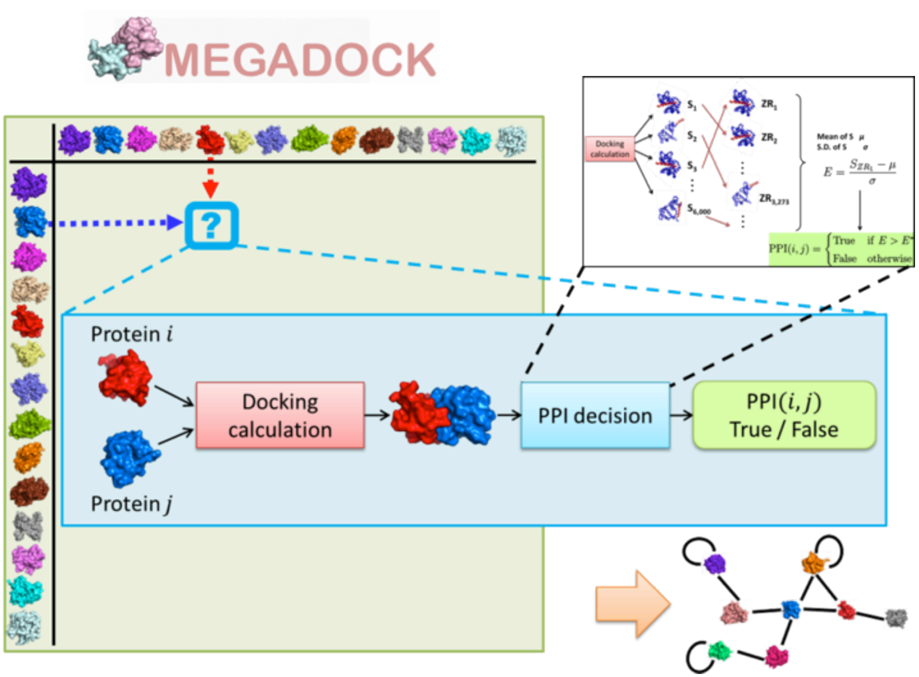
- soil metagenomic data (SRR407548)
  150 bp, 100,000 entries
- KEGG GENES amino acid sequences
  4 GB, (May, 2013)



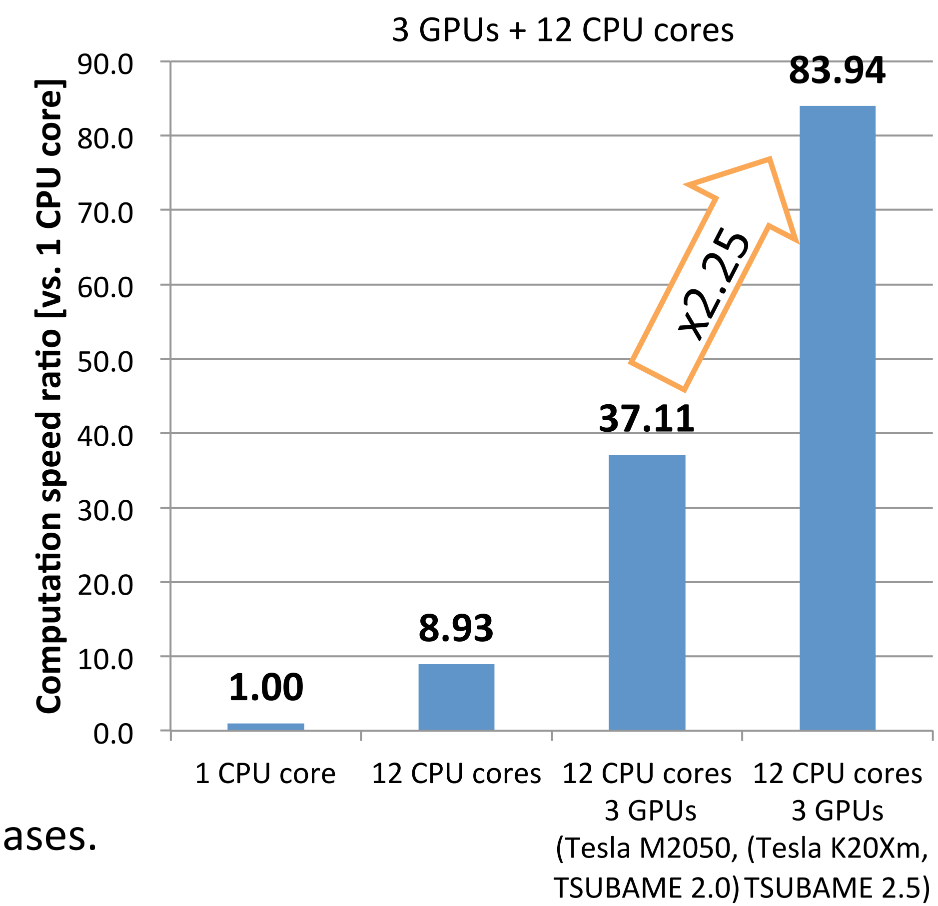
MEGADOCK-GPU

M. Ohue, *et al.*, *Protein Pept Lett*, (2013).
T. Shimoda, *et al.*, *ParBio2013* (2013).

Predicting protein-protein interaction network via protein-protein docking calculations



Docking calculations for 352 pairs

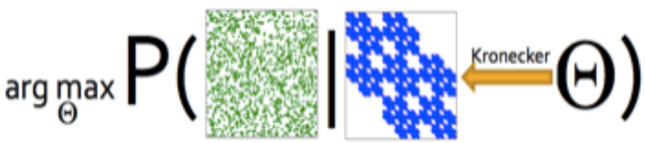


Protein-protein interaction network is very important to understand cell behavior and diseases.

Application	TSUBAME2.0 Performance	TSUBAME2.5 Performance	Boost Ratio
Top500/Linpack (PFlops)	1.192	2.843	2.39
Green500/Linpack (GFlops/W)	0.958	> 2.400	> 2.50
Semi-Definite Programming Nonlinear Optimization (PFlops)	1.019	1.713	1.68
Gordon Bell Dandrite Stencil (PFlops)	2.000	3.444	1.72
LBM LES Whole City Airflow (PFlops)	0.600	1.142	1.90
Amber 12 pmemd 4 nodes 8 GPUs (nsec/day)	3.44	11.39	3.31
GHOSTM Genome Homology Search (Sec)	19361	10785	1.80
MEGADOC Protein Docking (vs. 1CPU core)	37.11	83.49	2.25

Graph500 "Big Data" Benchmark

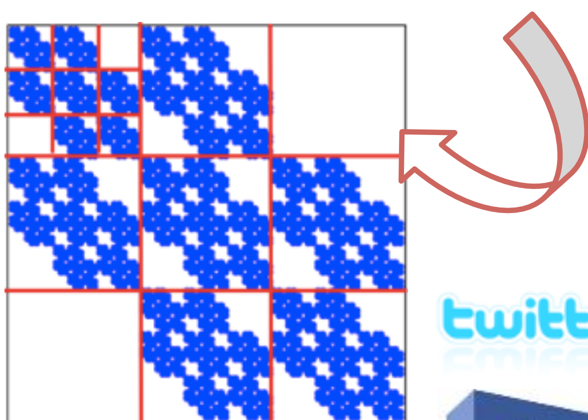
Kronecker graph BSP Problem



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

November 15, 2010

Graph 500 Takes Aim at a New Kind of HPC

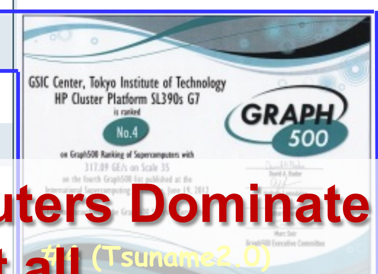
Richard Murphy (Sandia NL => Micron)

" **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

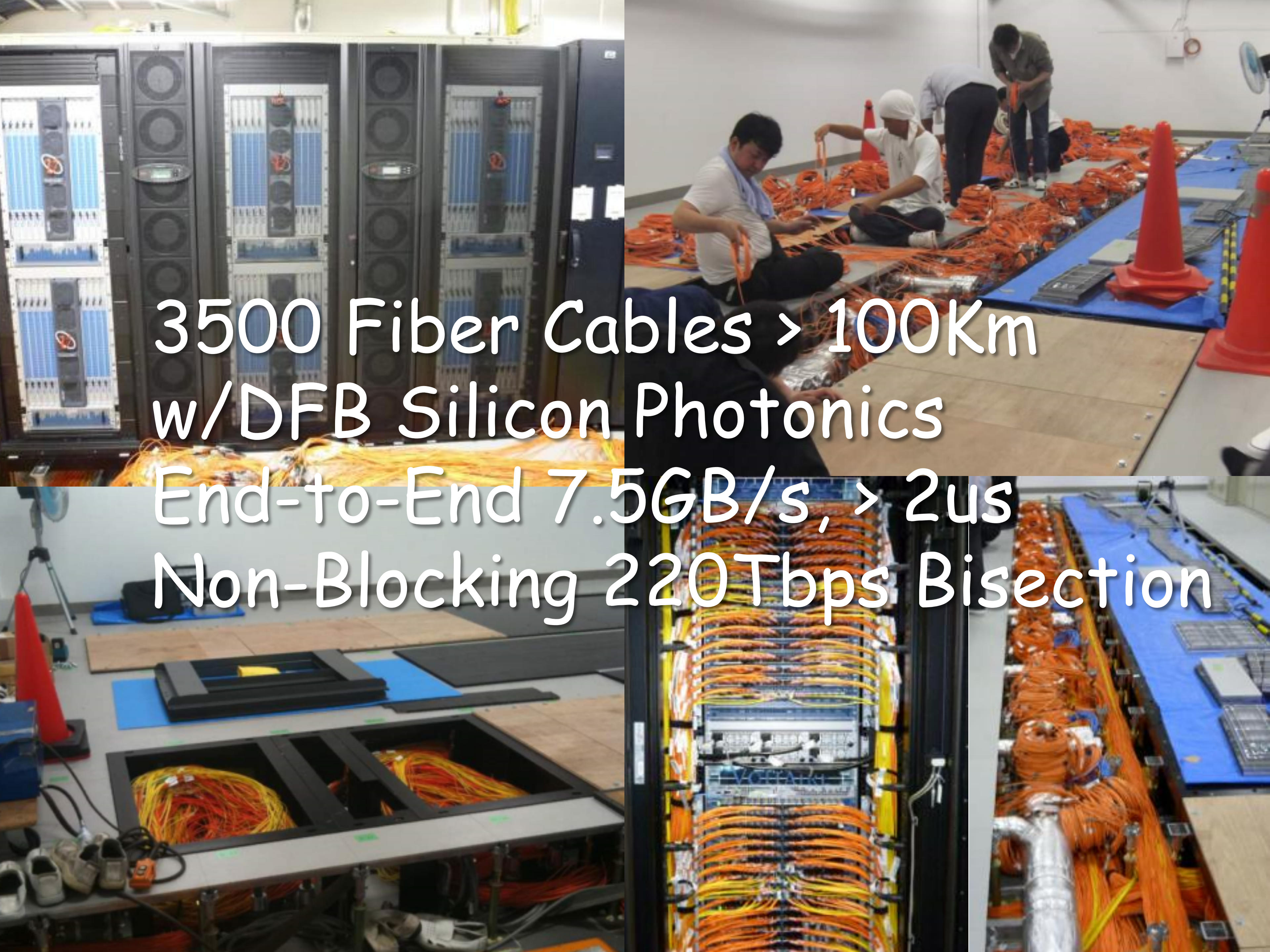
Rank	Installation Site	Machine	Number of nodes	Number of cores	Problem scale	GTEPS
1	DOE/SC/Argonne National Laboratory	Mira/BlueGene/Q	32768	524288	38	3541.00
1	LLNL	Sequoia/Blue Gene/Q	32768	524288	38	3541.00
2	DARPA Trial Subset, IBM Development Engineering	Power 775, POWER7 8C 3.836 GHz	1024	32768	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 Research [illegible]	NNSA/SC Blue Gene/Q Prototype II	[illegible]	65536	[illegible]	[illegible]



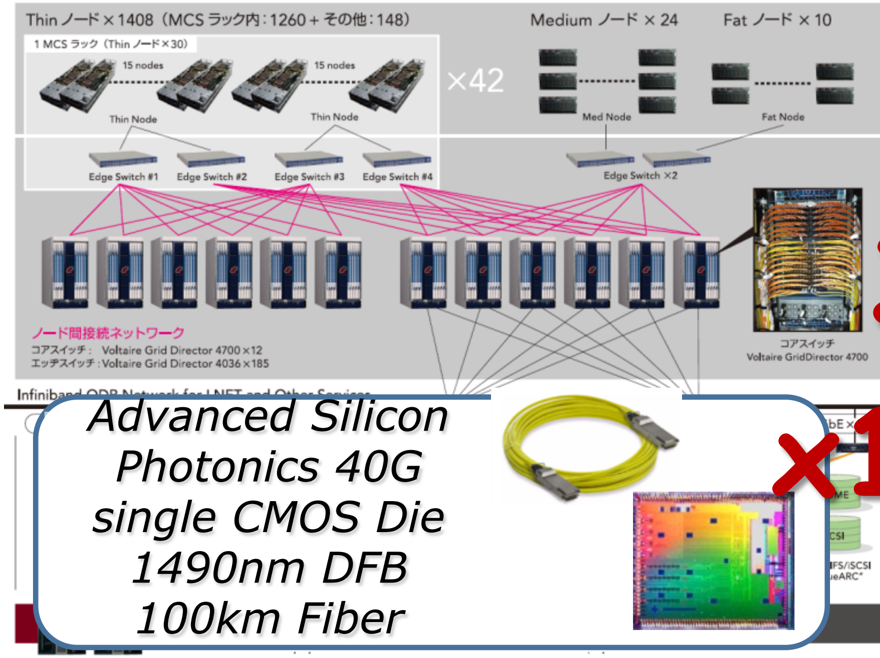
Reality: Top500 Supercomputers Dominate
No Cloud IDCs at all

#4 (Tsuname2.0)

TSUBAME2.0 #3(Nov.2011) #4(Jun.2012)

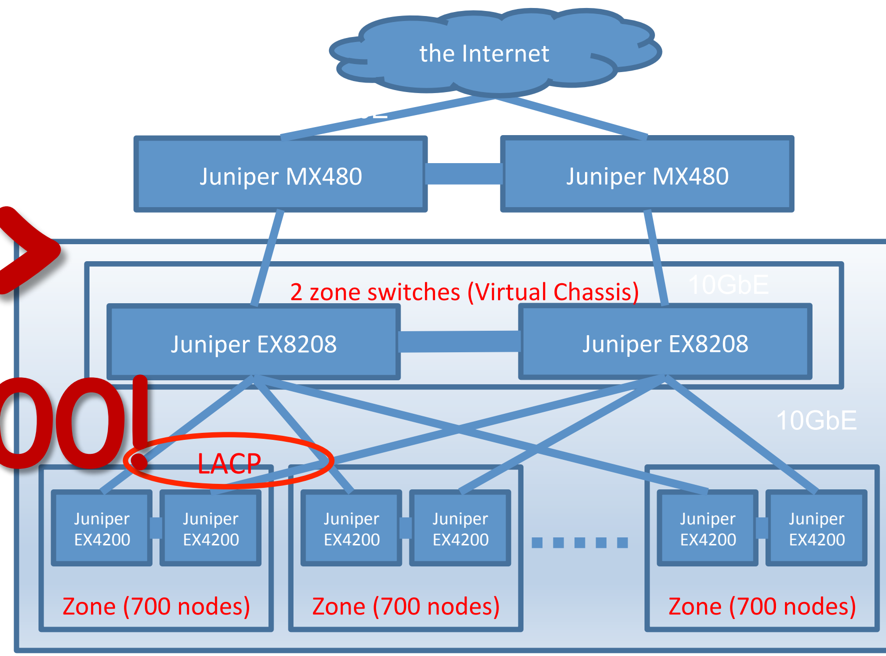

3500 Fiber Cables > 100Km
w/DFB Silicon Photonics
End-to-End 7.5GB/s, > 2us
Non-Blocking 220Tbps Bisection

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



×1000!

A Major Northern Japanese Cloud Datacenter (2013)



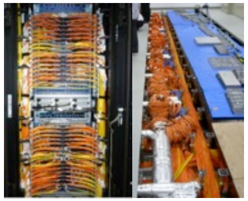
~1500 nodes compute & storage
Full Bisection Multi-Rail
Optical Network
Injection 80GBps/Node
Bisection 220Terabps

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

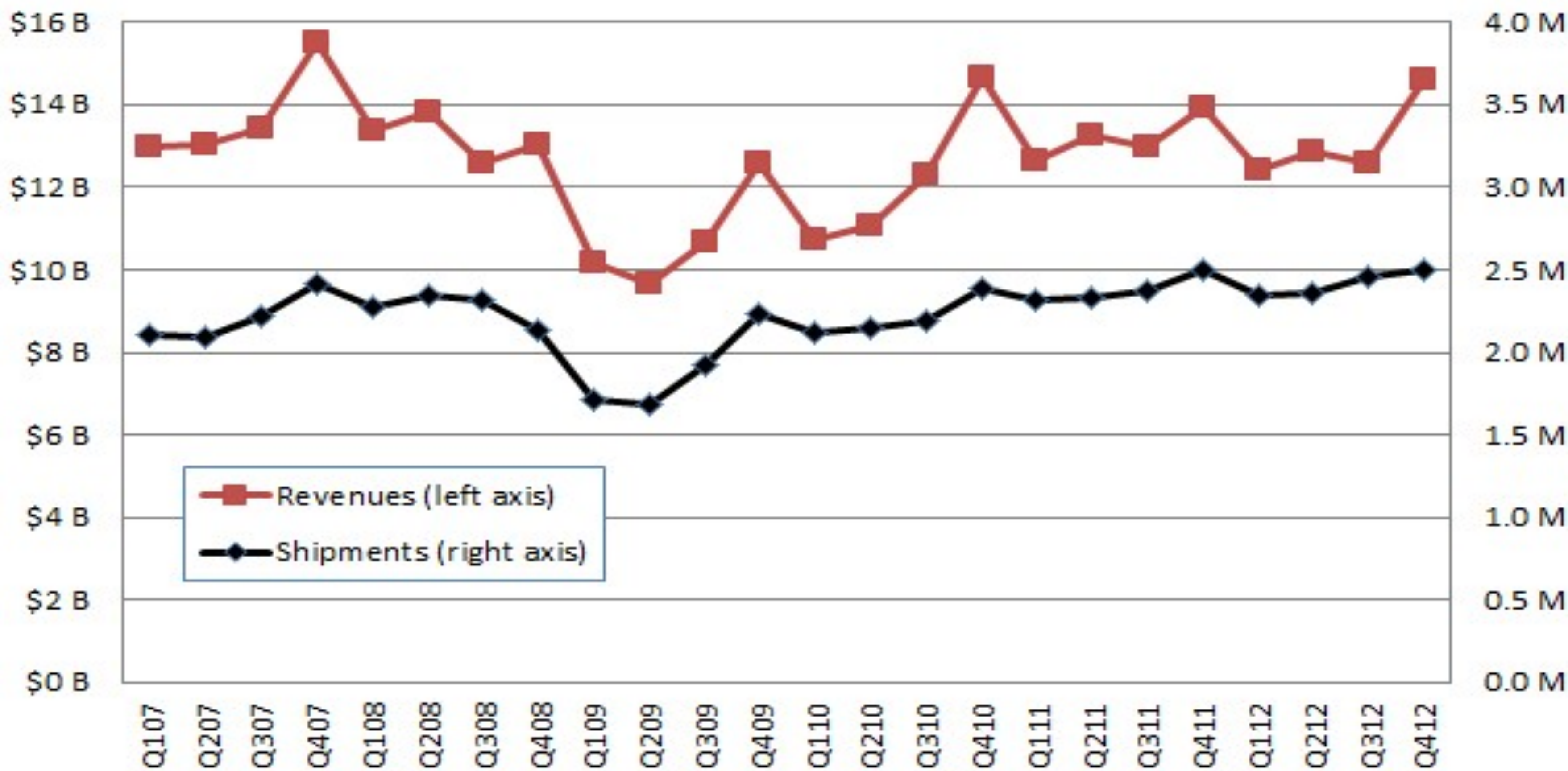
But what does "220Tbps" mean?

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

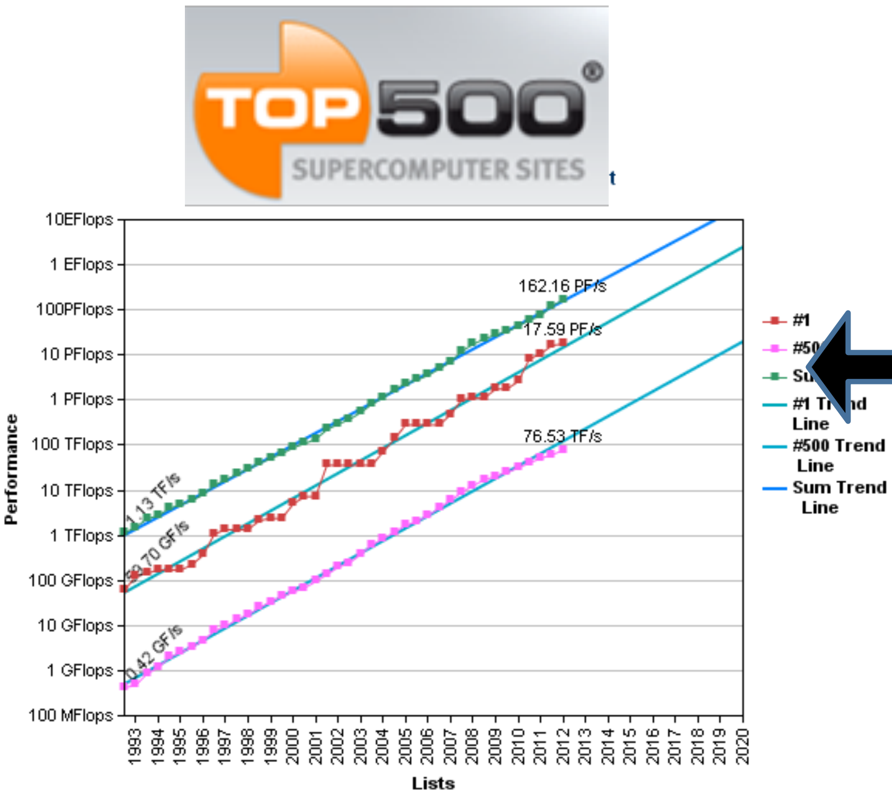


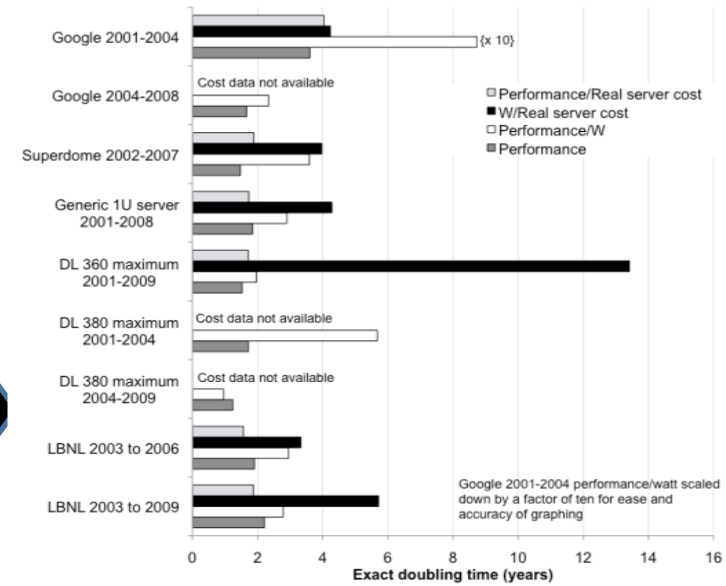
Global Server Shipments are Flat – ~40% Capacity Growth Rate (~30% for non-HPC)



Service rate proportional to North-South bandwidth
IDC Capacity Growth CAGR thus ~30%

"Convergence" of Supercomputing and Big Data with supercomputing leadership





Source: Assessing trends over time in performance, costs, and energy use for servers, Intel, 2009.

HPC: x1000 in 10 years

CAGR ~= 100%

IDC: x30 in 10 years

Server unit sales flat (replacement demand)

CAGR ~= 30-40%

Attendees:
US: 25
Europe: 11
Japan 9

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

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

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

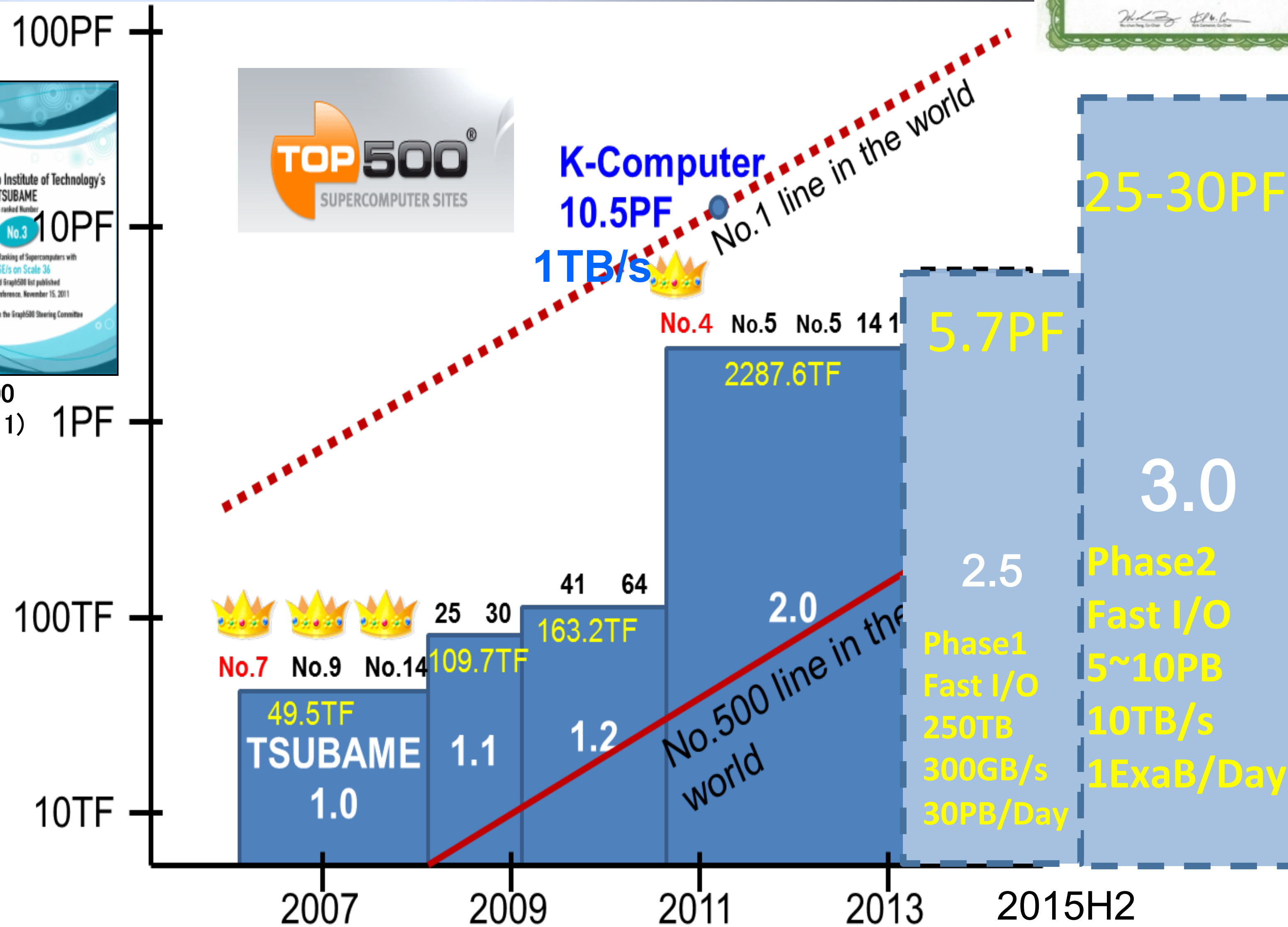
BIG DATA AND EXTREME-SCALE COMPUTING

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

TSUBAME Evolution

Graph 500
No. 3 (2011)

HPC wire
Awards



Focused Research Towards Tsubame 3.0 and Beyond towards Exa

- **Green Computing**: Ultra Power Efficient HPC
- **High Radix Bisection Networks** – HW, Topology, Routing Algorithms, Placement…
- **Fault Tolerance** – Group-based Hierarchical Checkpointing, Fault Prediction, Hybrid Algorithms
- **Scientific "Extreme" Big Data** – Ultra Fast I/O, Hadoop Acceleration, Large Graphs => **Convergence**
- **New memory systems** – Pushing the envelops of low Power vs. Capacity vs. BW, exploit the deep hierarchy with new algorithms to decrease Bytes/Flops
- **Post Petascale Programming** – OpenACC and other many-core programming substrates, Task Parallel
- **Scalable Algorithms for Many Core** – Apps/System/HW Co-Design

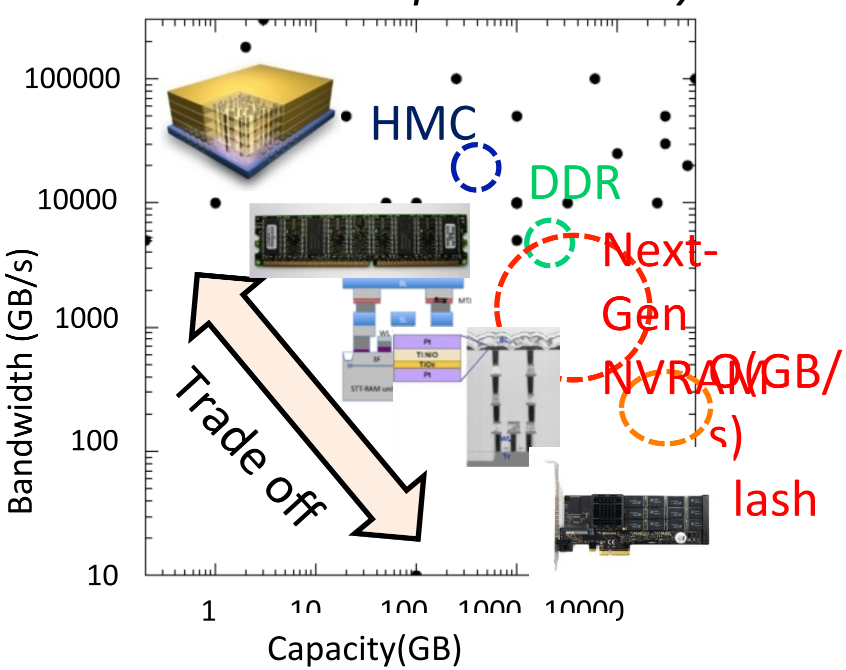
"Software Technology that Deals with Deeper Memory Hierarchy in Post-petascale Era"

2012-2017, PI: Toshio Endo, Tokyo Tech

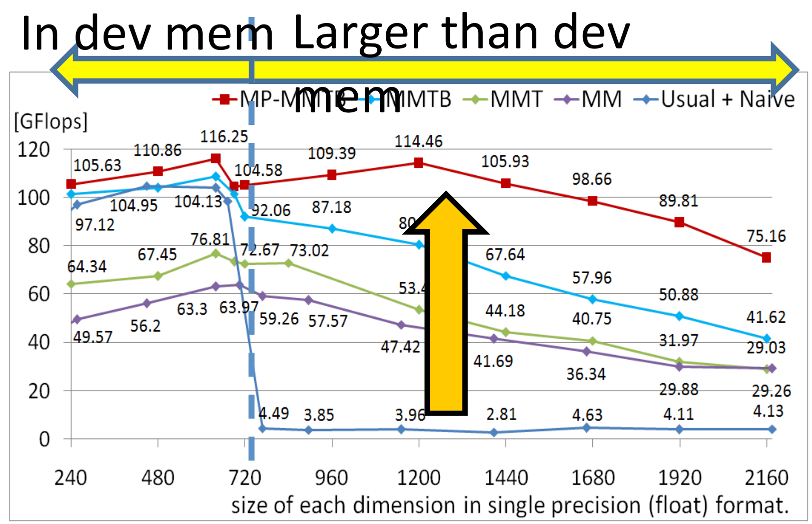
Growing "Memory wall" will be an obstacle to larger and fast simulations in post-petascale era

- Deeper memory hierarchy and locality improvement are keys
- Goals: ~100PB/s and ~100PB simulations on Exaflops system

Towards deeper hierarchy



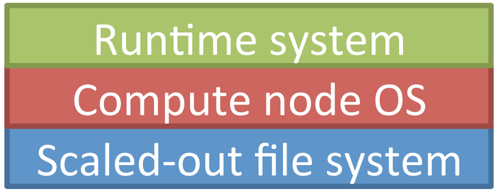
Locality Improvement of Stencil Computations

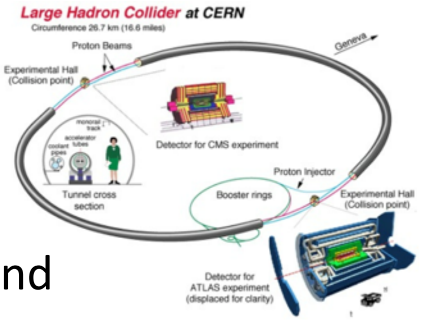


JST CREST "System Software for Post Petascale Data Intensive Science" (FY2011-15)

Co-PI	Institute	
Osamu Tatebe	University of Tsukuba	Project Leader
Yoshihiro Oyama	University of Electro-Communications	

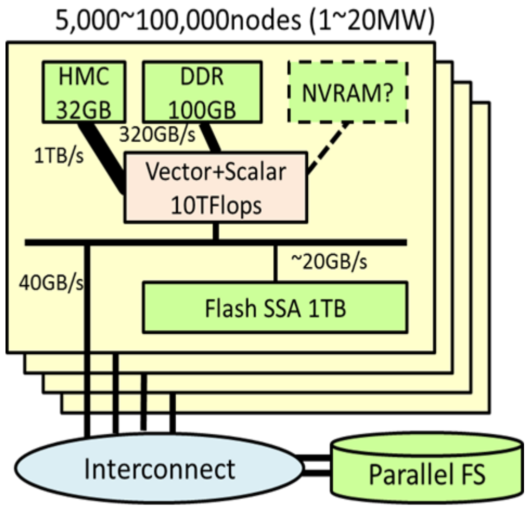
- Objective
  - R&D based on scale-out file system architecture
  - Target snapshot, 100 TB/s
- Research topics
  - Scale-out distributed file system
    - Scale out to O(10K) I/O servers by utilizing access locality
    - Metadata server clustering to scale the performance out
  - Compute node OS
    - Kernel driver, process scheduling, client caching, operation offload
  - Runtime for Data-Intensive Computing
    - Efficient runtime of workflow execution, MapReduce, and MPI-IO for the scale-out distributed file system

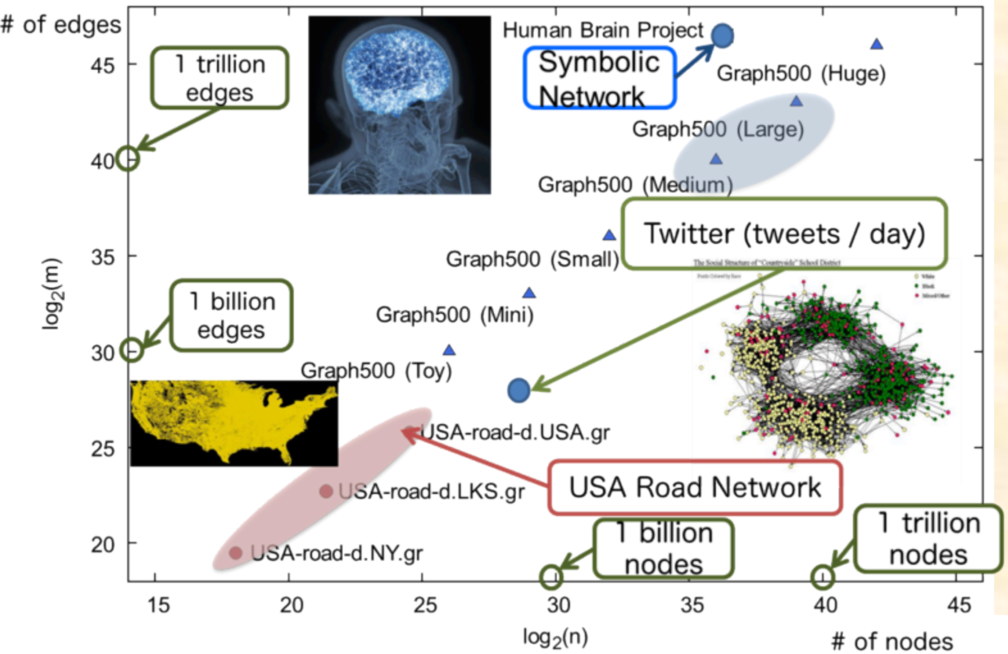




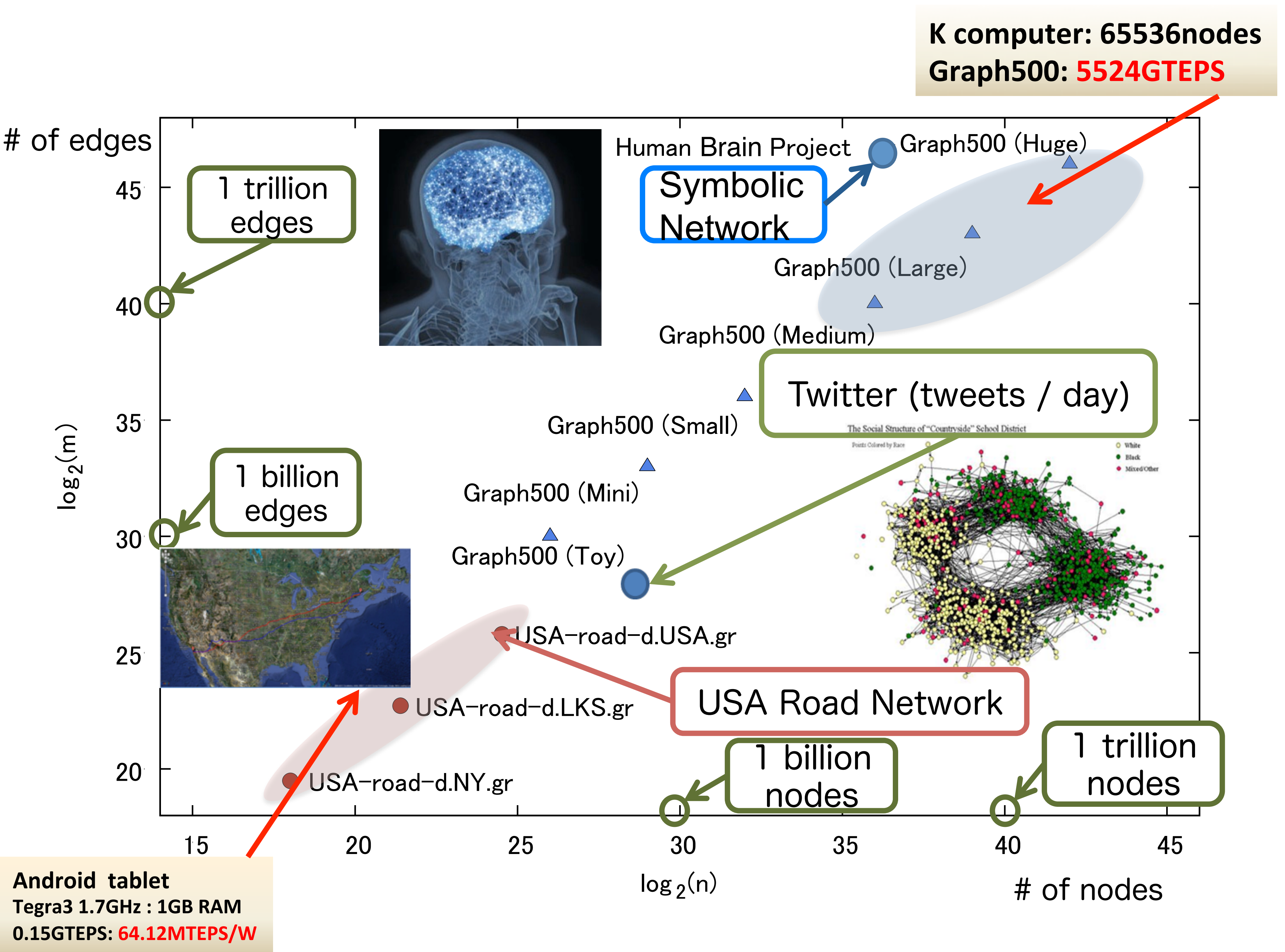
JST CREST: Advanced Computing and Optimization Infrastructure for Extremely Large-Scale Graphs on Post Peta-Scale Supercomputers

- 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



K computer: 65536nodes
Graph500: 5524GTEPS
# of edges
1 trillion edges
Human Brain Project
Symbolic Network
Graph500 (Huge)
Graph500 (Large)
Graph500 (Medium)
Twitter (tweets / day)
The Social Structure of "Countryside" School District
Points Colored by Race
White
Black
Mixed/Other
Graph500 (Small)
Graph500 (Mini)
1 billion edges
Graph500 (Toy)
USA-road-d.USA.gr
USA-road-d.LKS.gr
USA-road-d.NY.gr
USA Road Network
1 billion nodes
1 trillion nodes
45
40
35
30
25
20
$\log_2(m)$
15
20
25
30
35
40
45
$\log_2(n)$
# of nodes
Android tablet
Tegra3 1.7GHz : 1GB RAM
0.15GTEPS: 64.12MTEPS/W

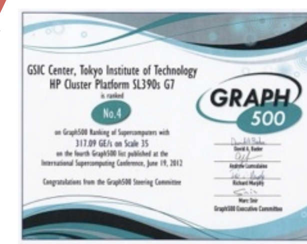
Our achievements (Super computer) : **Graph500**

Rank	1st 2010/11	2nd 2011/06
1	SCALE36 / 7.0G / 8192 node	SCALE38 / 18.47 G / 32768 node
2	SCALE32 / 5.6G / 9544 node	SCALE38 / 18.36 G / 32768 node
3	SCALE29 / 1.3G / 128 node	SCALE37 / 43.38 G / 4096 node

4th List

University of Tokyo FX10

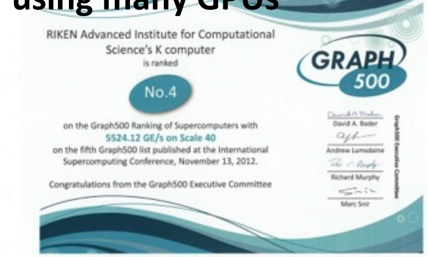
Rank	3rd 2011/11	4th 2012/06
1	SCALE32 / 253G / 4096 node	SCALE38 / 3541G / 32768 node
2	SCALE37 / 113G / 1800 node	SCALE35 / 508G / 1024 node
3	SCALE37 / 103G / 4096 node	SCALE38 / 358G / 4800 node
4	SCALE36 / 100G / 1366 node	SCALE35 / 317G / 1366 node



TITECH TSUBAME 2.0

TITECH TSUBAME 2.0
The first implementation using many GPUs

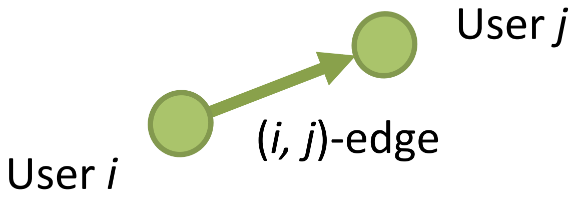
Rank	5th 2012/11
1	SCALE40 / 15363G / 65536 node
2	SCALE39 / 10461G / 32768 node
3	SCALE38 / 5848G / 15384 node
4	SCALE40 / 5524G / 65536 node



K Computer in AICS, Japan

Twitter network (Application of Graph500 Benchmark)

Follow-ship network 2009



41 million vertices and 2.47 billion edges

Our NUMA-optimized BFS on 4-way Xeon system

69 ms / BFS

⇒ **21.28 GTEPS**

Six-degrees of separation

Frontier size in BFS

with source as User 21,804,357

Lv	Frontier size	Freq. (%)	Cum. Freq. (%)
0	1	0.00	0.00
1	7	0.00	0.00
2	6,188	0.01	0.01
3	510,515	1.23	1.24
4	29,526,508	70.89	72.13
5	11,314,238	27.16	99.29
6	282,456	0.68	99.97
7	11536	0.03	100.00
8	673	0.00	100.00
9	68	0.00	100.00
10	19	0.00	100.00
11	10	0.00	100.00
12	5	0.00	100.00
13	2	0.00	100.00
14	2	0.00	100.00
15	2	0.00	100.00
Total	41,652,230	100.00	-

Extreme Big Data (EBD)

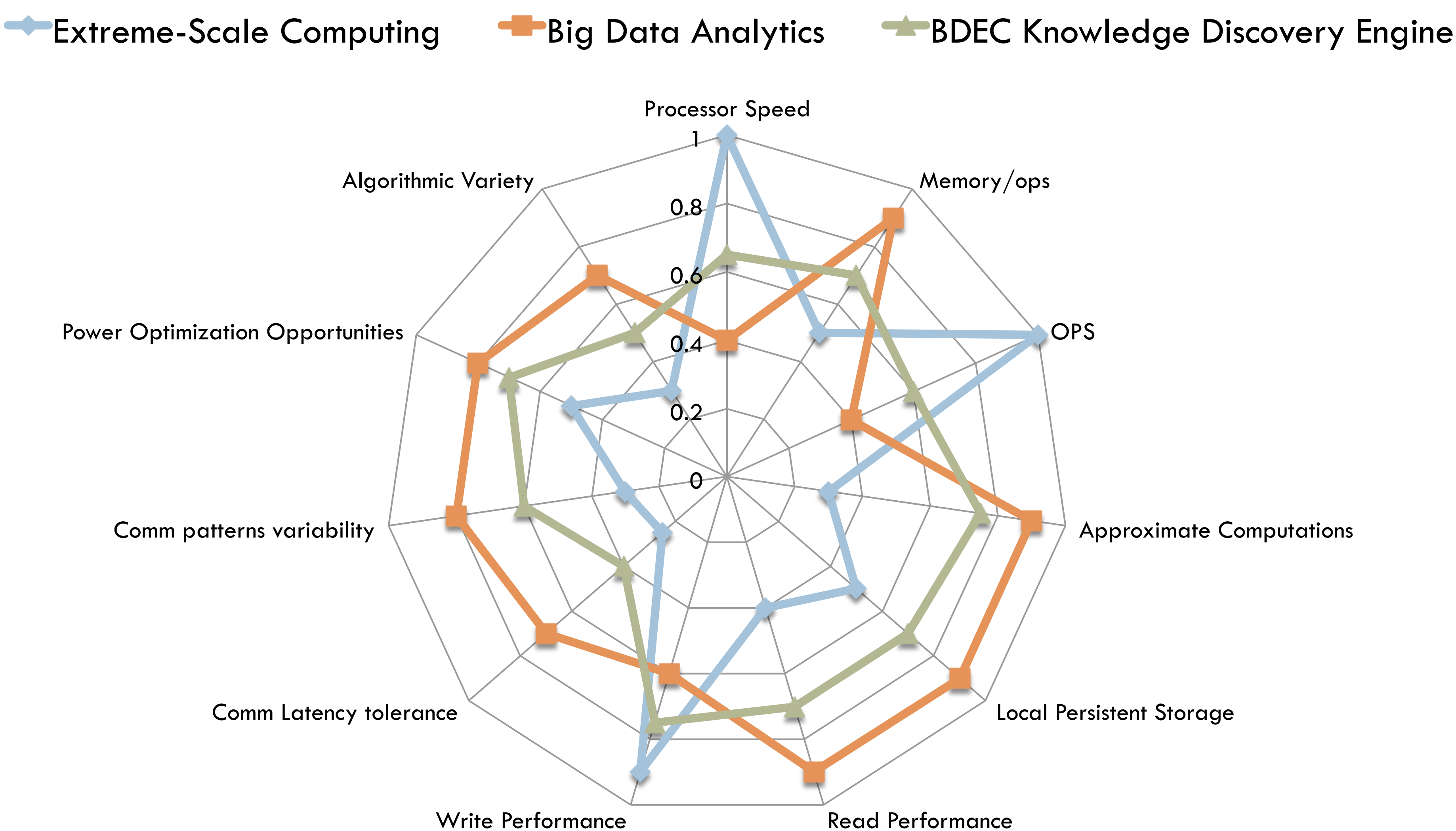
Next Generation Big Data Infrastructure Technologies Towards Yottabyte/Year

Principal Invesigator
Satoshi Matsuoka

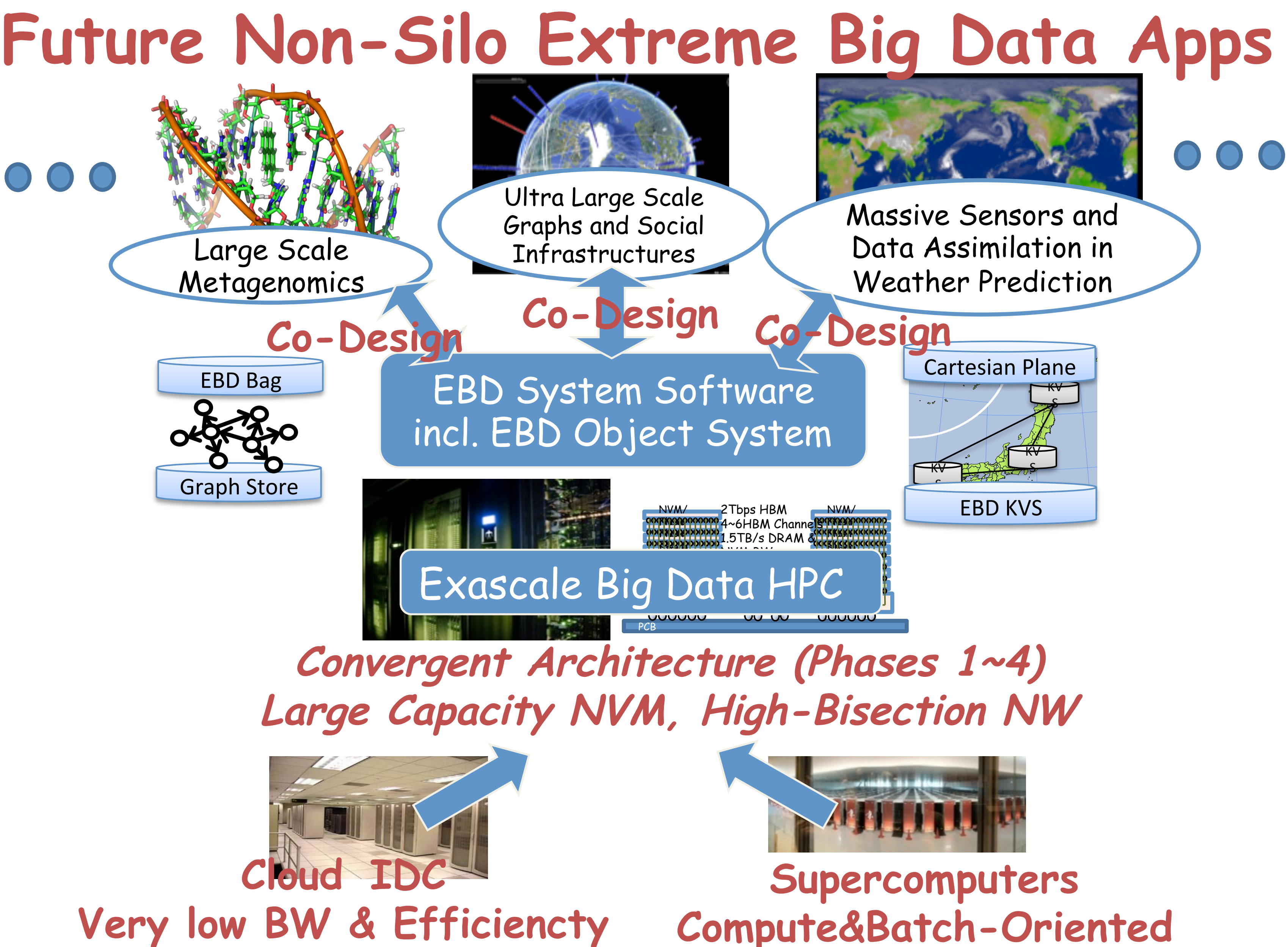
Global Scientific Information and Computing Center
Tokyo Institute of Technolgoy

Extreme Big Data not just traditional HPC!!! --- Analysis of required system properties ---

[Slide courtesy Alok Choudhary, Northeastern U]



EBE Research Scheme



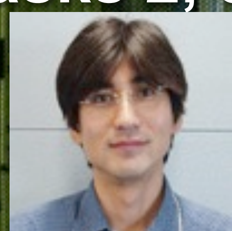
Extreme Big Data (EBD) Team

Co-Design EHPC and EDB Apps

- **Satoshi Matsuoka (PI), Toshio Endo, Hitoshi Sato (Tokyo Tech.) (Tasks 1, 3, 4, 6)**



- **Osamu Tatebe (Univ. Tsukuba) (Tasks 2, 3)**



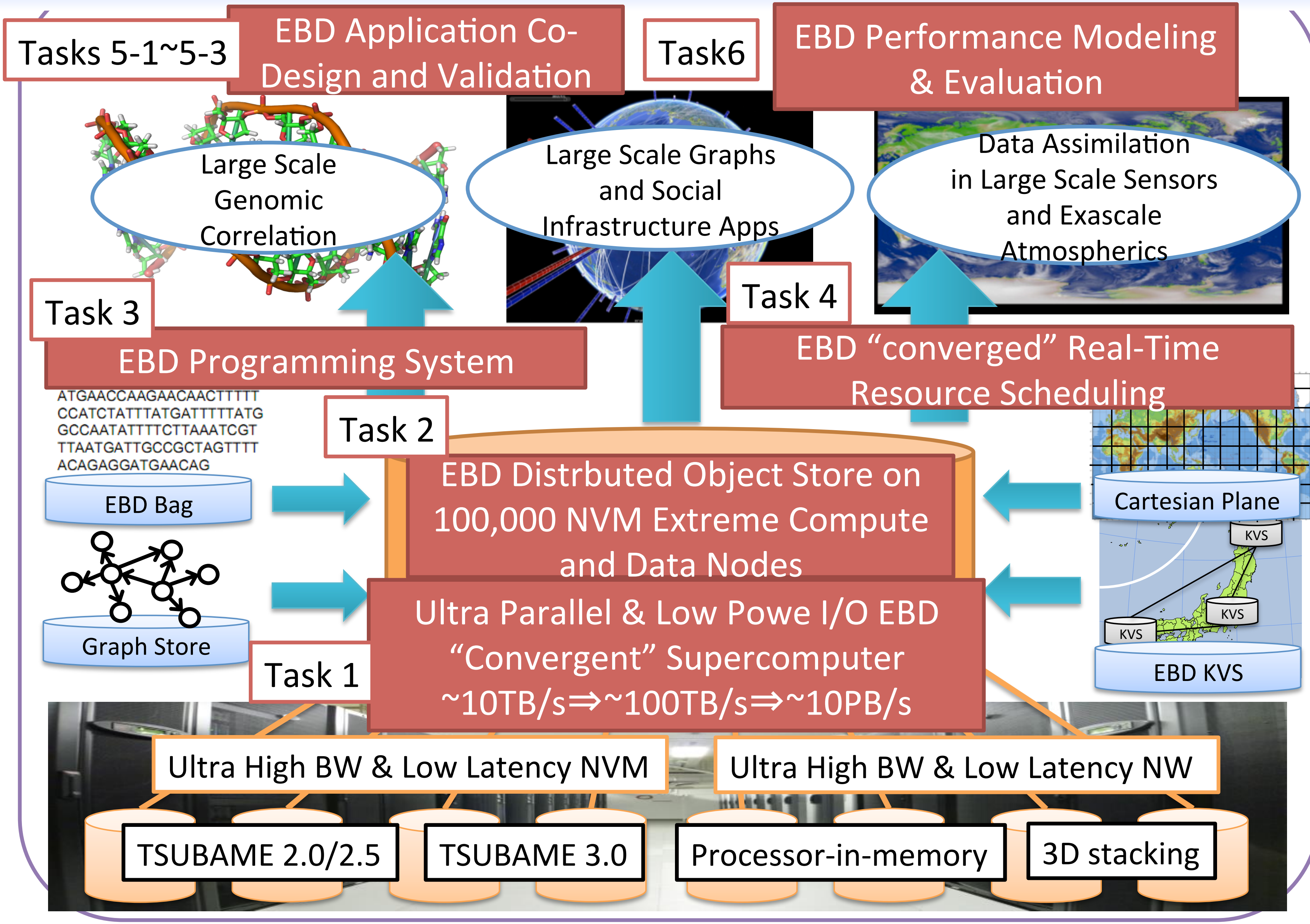
- **Michihiro Koibuchi (NII) (Tasks 1, 2)**

- **Yutaka Akiyama, Ken Kurokawa (Tokyo Tech, 5-1)**

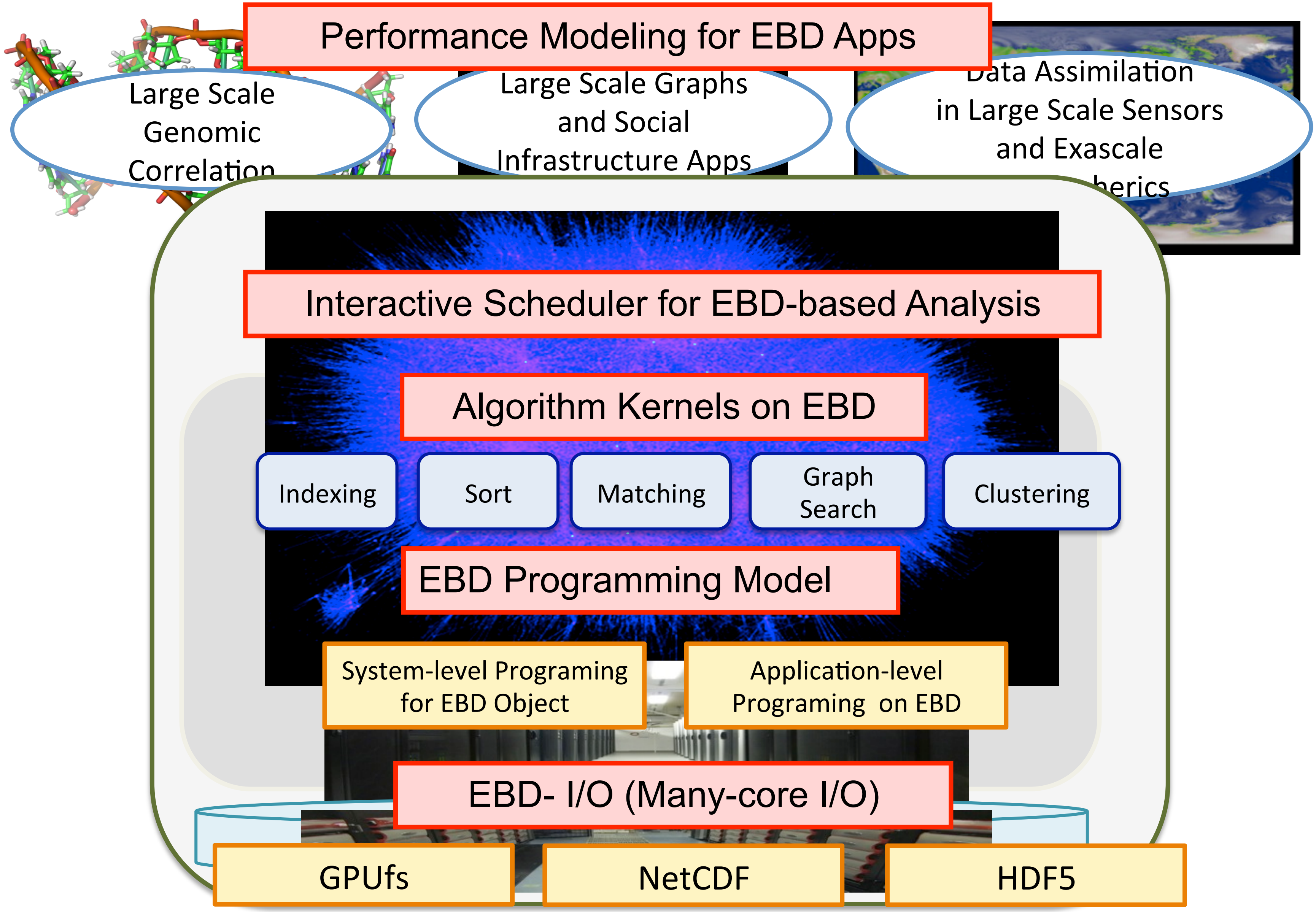

- **Toyotaro Suzumura (Tokyo Tech. and IBM Lab, 5-2)**

- **Takemasa Miyoshi (Riken AICS, 5-3)**

100,000 Times Fold EBD "Convergent" System Overview



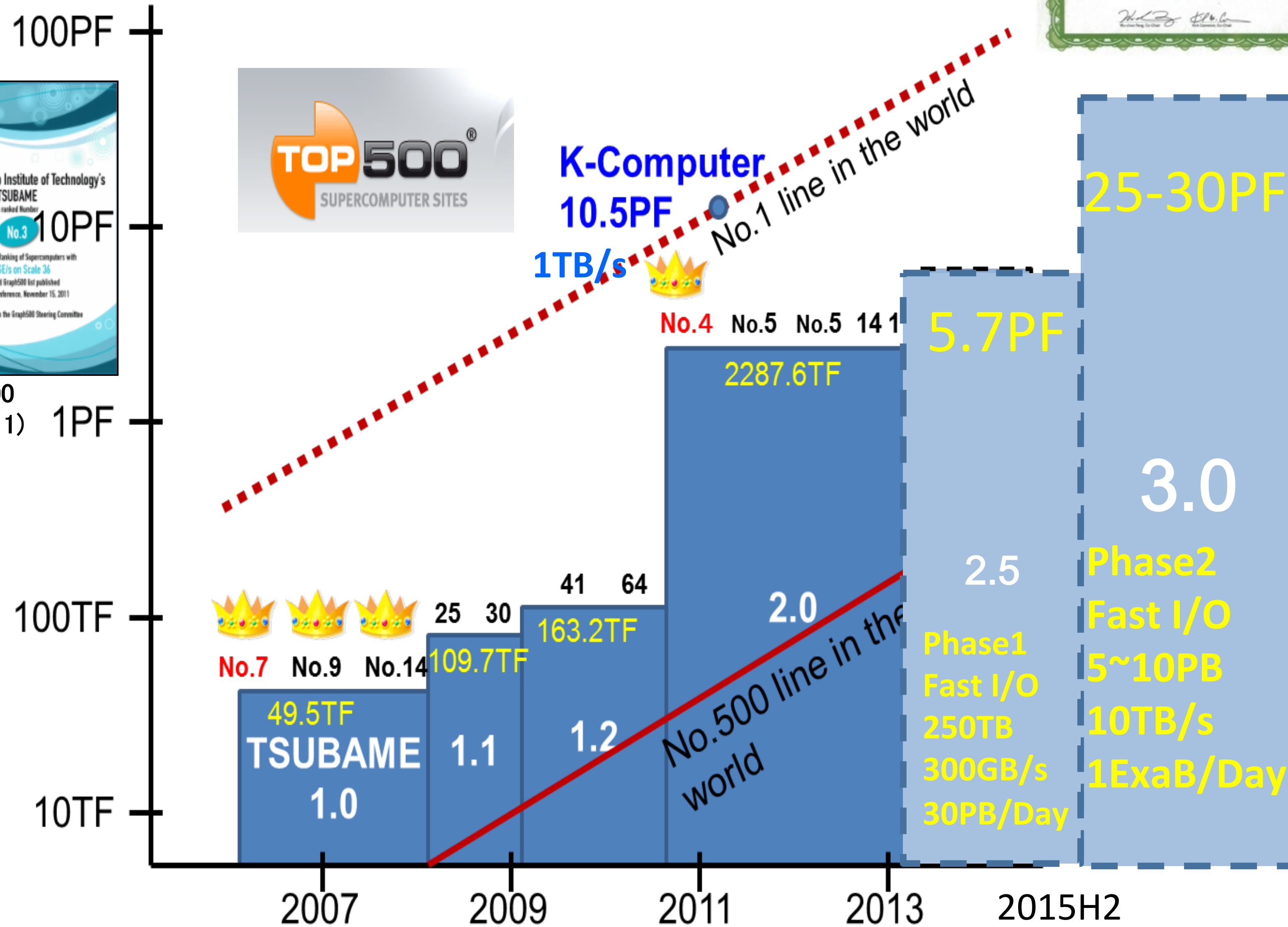
EBD System Software (Matsuoka Group)



TSUBAME Evolution Towards EBD (Matsuoka Group)

Graph 500
No. 3 (2011)

HPC wire
Awards

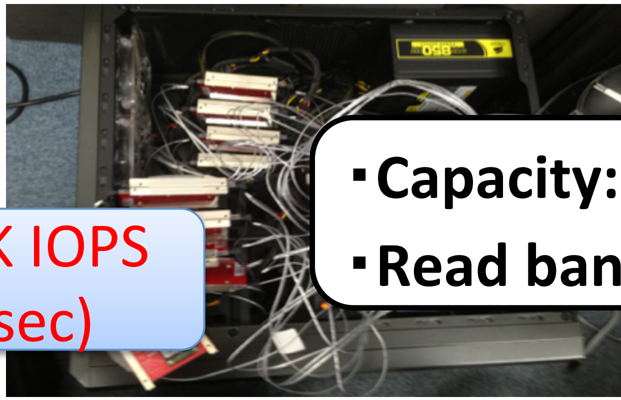


EBD- I/O
(Many-core I/O)

Preliminary I/O Evaluation on GPU and NVRAM

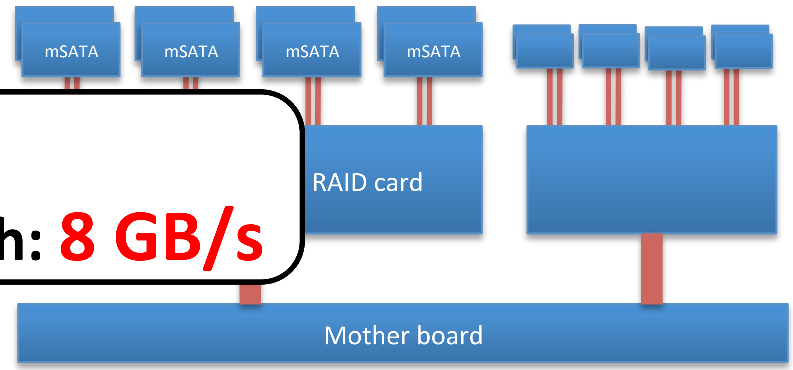
How to design local storage for next-gen supercomputers ?

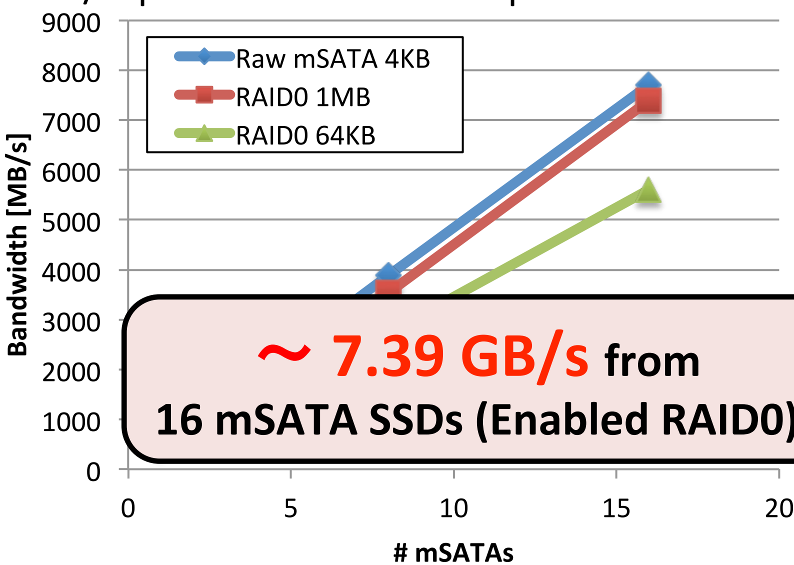
- Designed a local I/O prototype using 16 mSATA SSDs



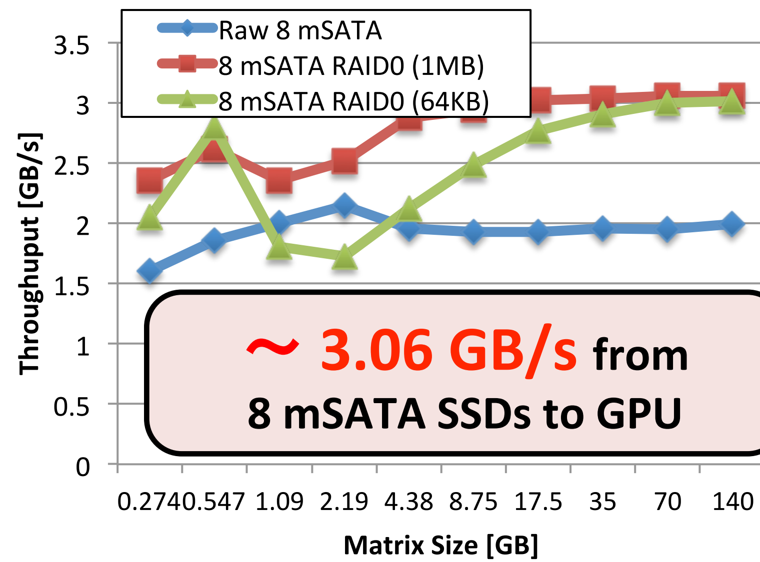
～320K IOPS (3 μ sec)

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



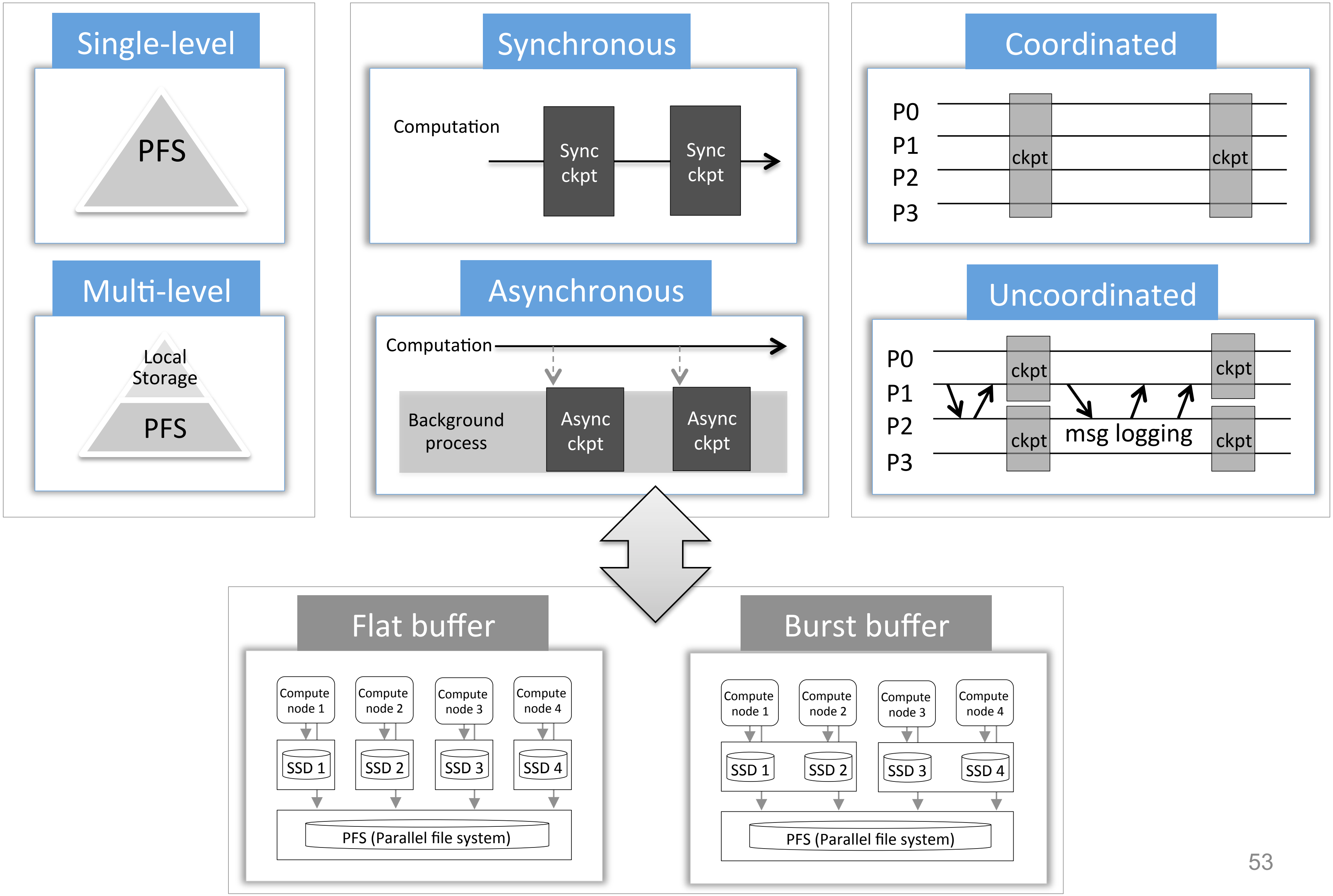


～ 7.39 GB/s from 16 mSATA SSDs (Enabled RAID0)



～ 3.06 GB/s from 8 mSATA SSDs to GPU

Target C/R strategies & Storage designs



Multi-level Asynchronous C/R Model

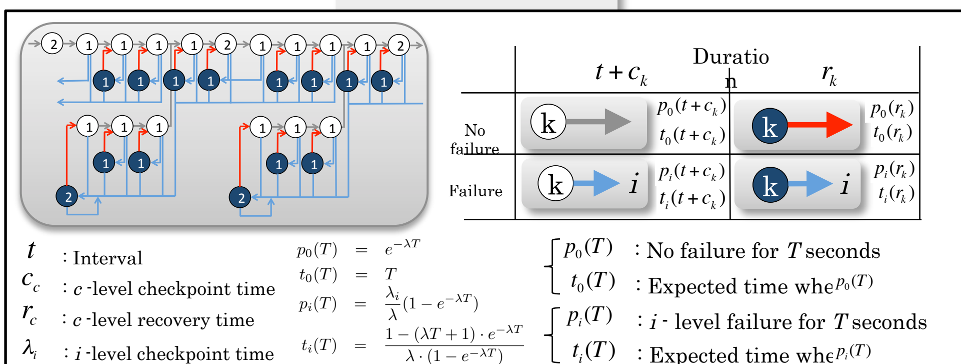
- Compute checkpoint/restart "*Efficiency*" *for C/R strategy comparison*
  - *Efficiency* : Fraction of time an application spends only in computation in optimal checkpoint interval

$$Efficiency = \frac{ideal\ \ runtime}{expected\ \ runtime}$$

ideal runtime : No failure and No checkpoint

expected runtime : Computed by the models

$$f : (L_{i=1\ldots N}, O_{i=1\ldots N}, R_{i=1\ldots N})$$



- Input: Each level of
  - L_i : Checkpoint Latency
  - O_i : Checkpoint overhead
  - R_i : Restart time
- Output: "*Efficiency*"

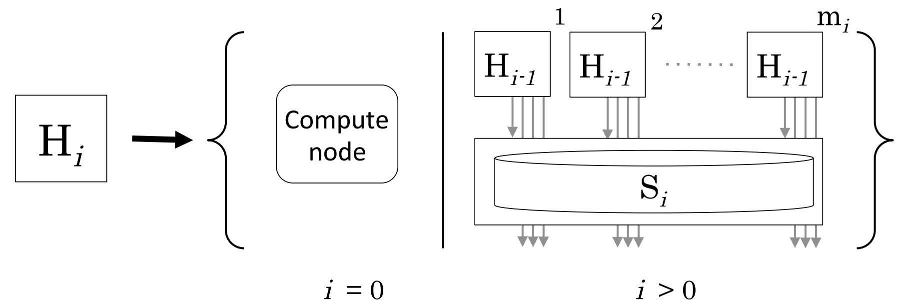
Efficiency

Source: Sato, K., Maruyama, N., Mohror, K., Moody, A., Gamblin, T., de Supinski, B. R. and Matsuoka, S.: Design and Modeling of a Non-Blocking Checkpointing System (SC12)

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

Recursive Structured Storage Model (Collaboration with DoE LLNL)

- Generalization of storage architectures with "*context-free grammar*"
  - A tier i hierarchical entity (H_i), has a storage (S_i)shared by (m_i) upper hierarchical entities (H_{i-1})
  - $H_{i=0}$ is a compute node
  - $H_N \{m_1, m_2, \ldots, m_N\}$



Storage Model: $H_N \{m_1, m_2, \ldots, m_N\}$

r_i	Sequential read throughput from compute nodes ($H_{i=0}$)
w_i	Sequential write throughput from compute nodes ($H_{i=0}$)
m_i	The number of a upper hierarchical entities (H_{i-1}) sharing S_i

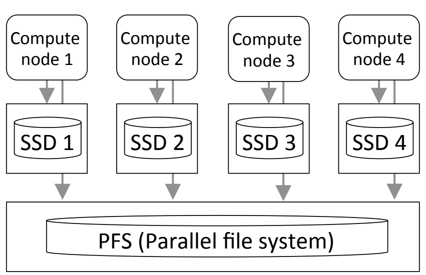
<# of C/R nodes per S_i >

=

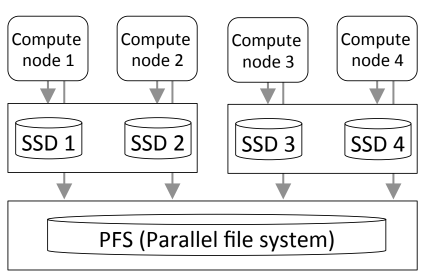
$$\frac{K^*}{\text{<\# of } S_i\text{>} \; (= \Pi^N_{k=i+1} m_k)}$$

*K: C/R cluster size

Example

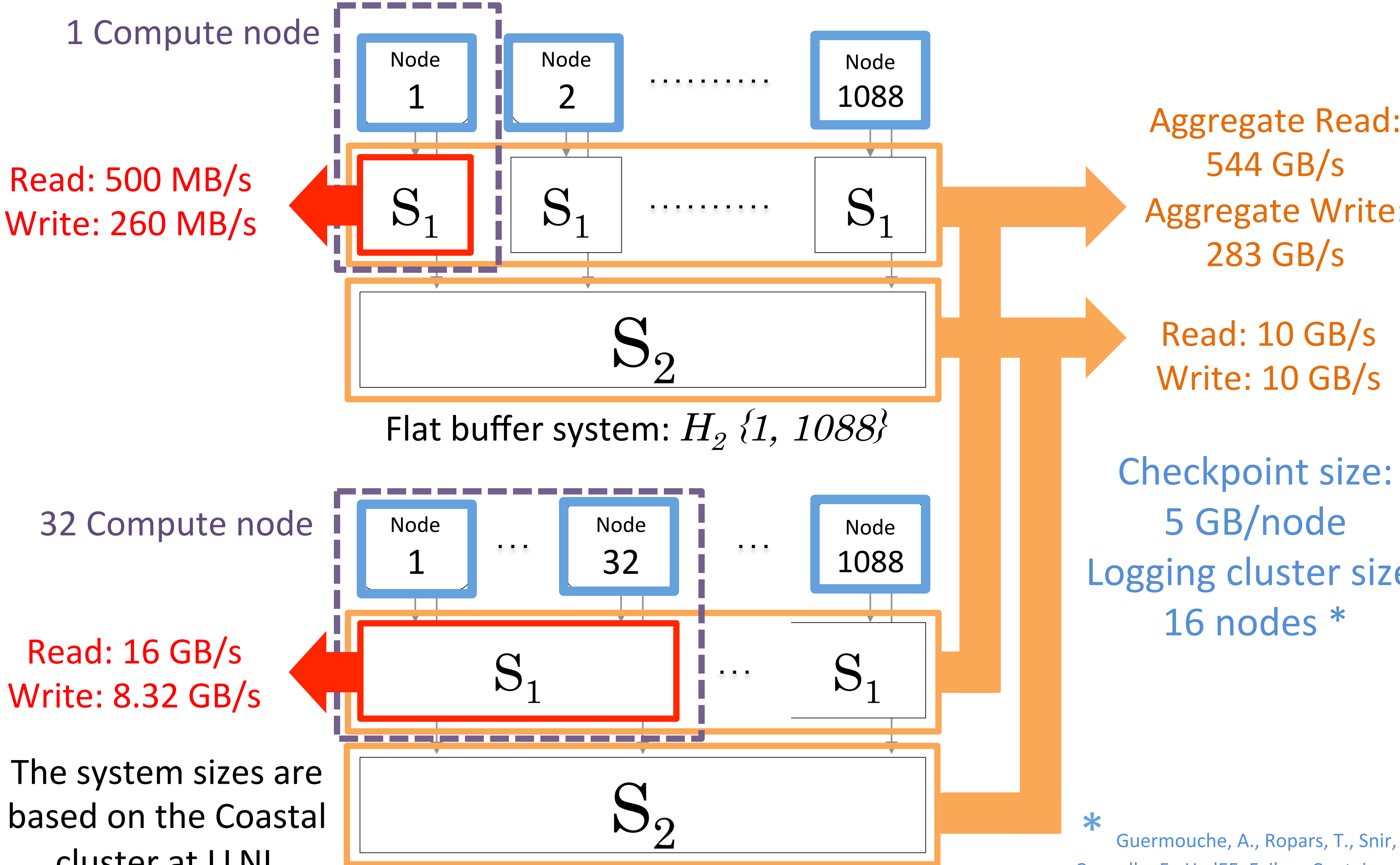


Flat buffer system: $H_2 \{1, 4\}$



Burst buffer system: $H_2 \{2, 2\}$

Experimental Setup



Burst buffer system: H_2 {32, 34}

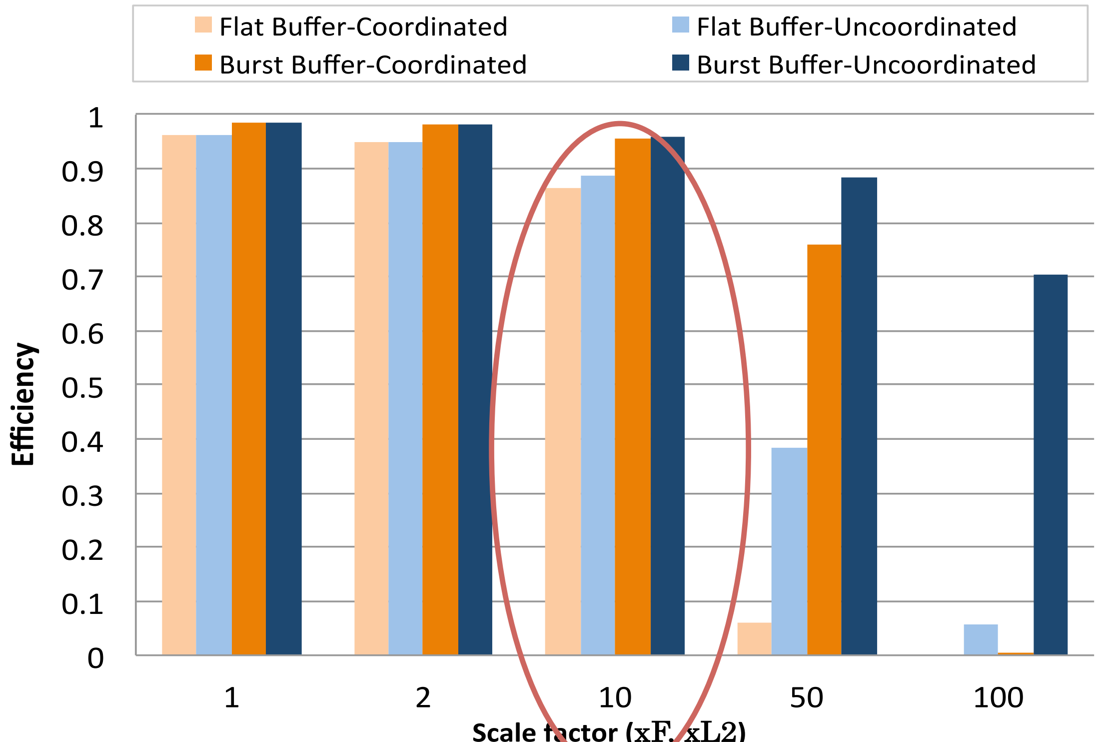
Checkpoint size:
5 GB/node
Logging cluster size:
16 nodes *

The system sizes are based on the Coastal cluster at LLNL (88.5TFLOPS)

* Guermouche, A., Ropars, T., Snir, M. and Cappello, F.: HydEE: Failure Containment without Event Logging for Large Scale Send-Deterministic MPI Applications

Efficiency with Increasing Failure Rates and Checkpoint Costs

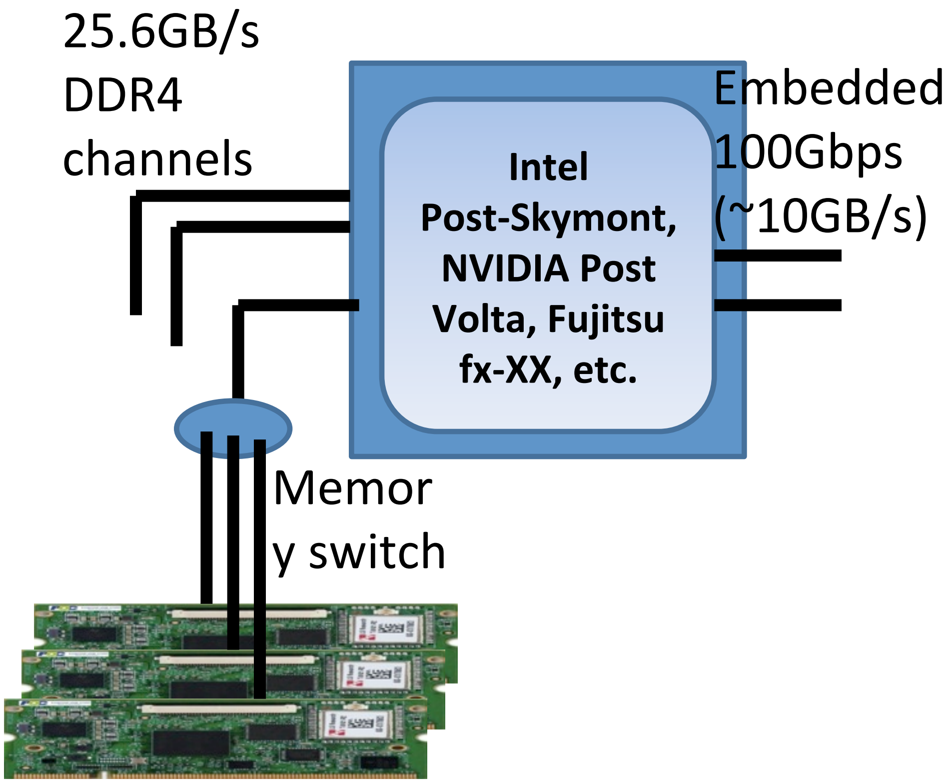
- Assuming message logging overhead is 0
- The burst buffer system always achieves a higher efficiency
  ⇒ Stores checkpoints on fewer nodes



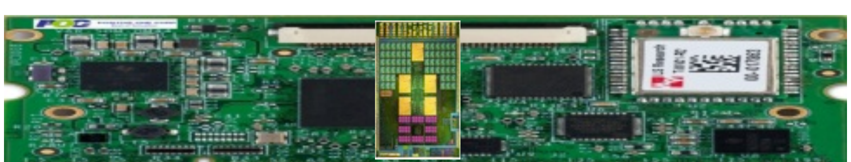
- All systems works equally well up to x10 => TSUBAME4.0 (2020) can go exascale
- Uncoordinated checkpointing: 70% efficiency on systems two orders of magnitude larger (if logging overhead is 0)
  ⇒ Partial restart exploit the bandwidth of both burst buffers and the PFS

Phase3 Scaling up to Petabyte/s I/O EBD 2017-18

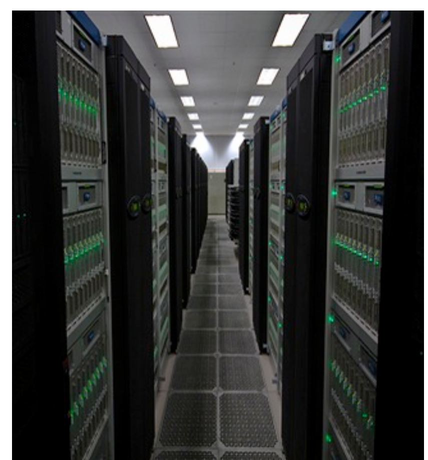
DRAM+Flash(+Processor) 100 ExaB/Day, 30 ZetaB/Year



25.6GB/s x 3~4 =80~100GB/s (DDR4-3200)
4~6 channels=>320~600GB/s
(12~24 DIMMS per socket)
4.8 Teraflops 10W, $500?



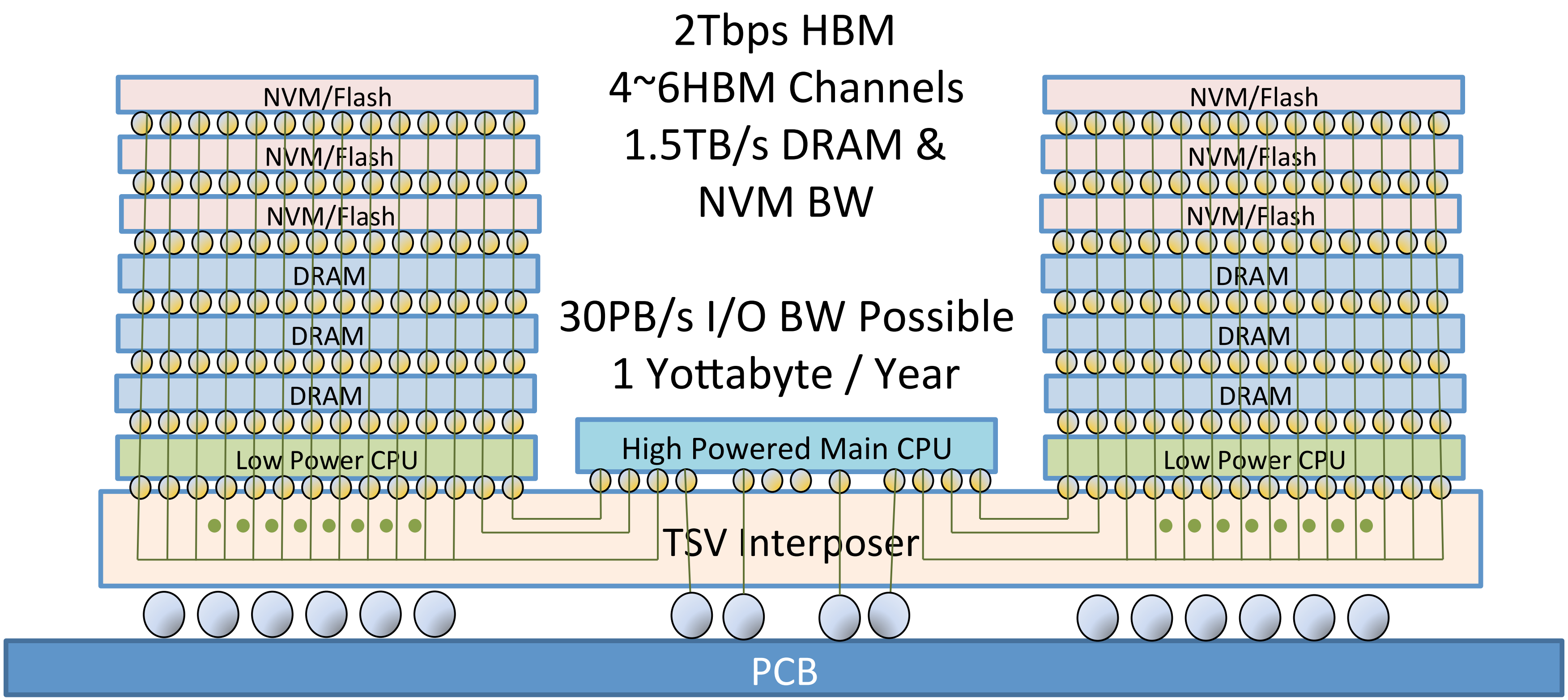
Rack
4 cabinets/64 nodes
25TB DRAM
786TB Flash
50 TB/s DRAM BW
1.54TB/s Flash BW
1.28TB/s NW BW
384TFLops
30.7KW, $1 mil



IDC/SC
650 Racks (~ES)
41,600nodes
16PB DRAM
511PB Flash
25.6PB/s DRAM BW
1PB/s Flash BW
(x1000 K-comp HDD)
250PFlops DFP
500PFlops SFP
830TB/s NW BW
20MW, $700 million

Phase4: 2019-20 DRAM+NVM+CPU with 3D/2.5D Die Stacking

-The Ultimate Convergence of BD and EC-



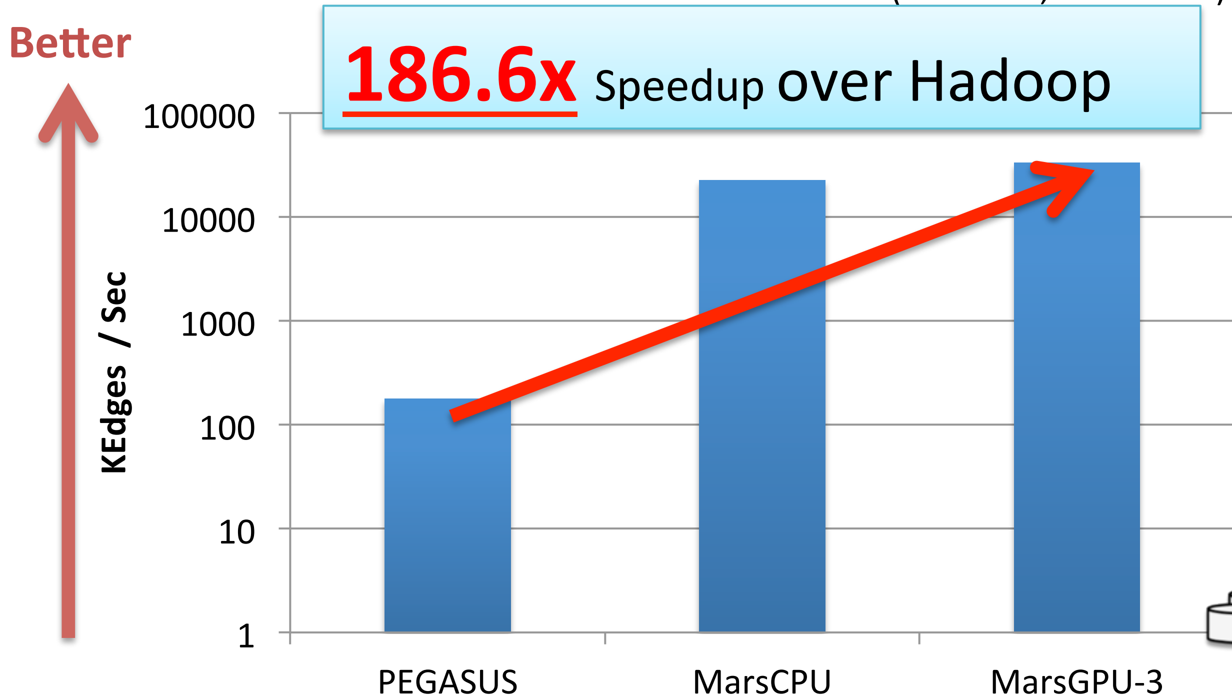
EBD Programming Model

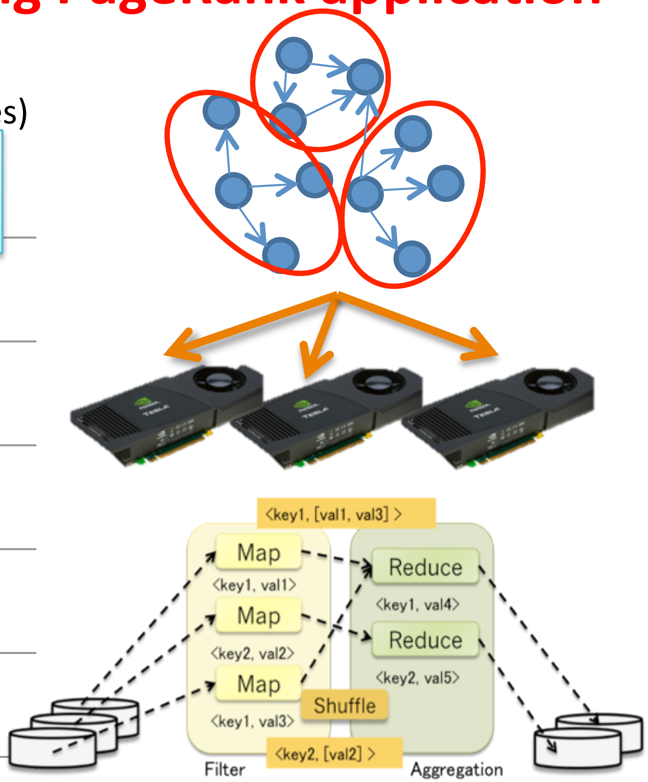
A scalable MapReduce-based large scale graph processing algorithm using multi-GPU [CCGrid 2013]

How to utilize multi-GPU for large-scale graph processing ?

- **Implemented a graph processing algorithm on multi-GPU**
- **Confirmed speedup on multi-GPU using PageRank application**

Performance evaluation on TSUBAME 2.0 (SCALE 27, 128 nodes)



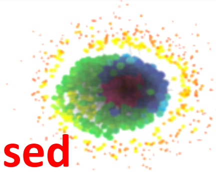


Large Scale BFS Using NVRAM

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

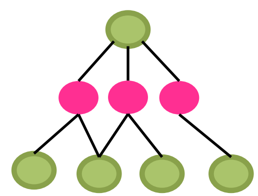
Using NVRAMs for large scale graph processing has possibilities of minimum performance degradation

2.Hybrid-BFS

Switch two approaches

Top-down

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



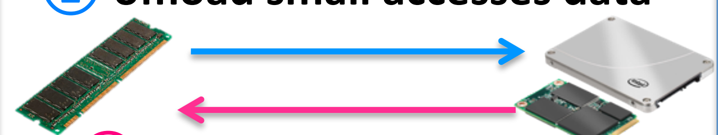
Bottom-up

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

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

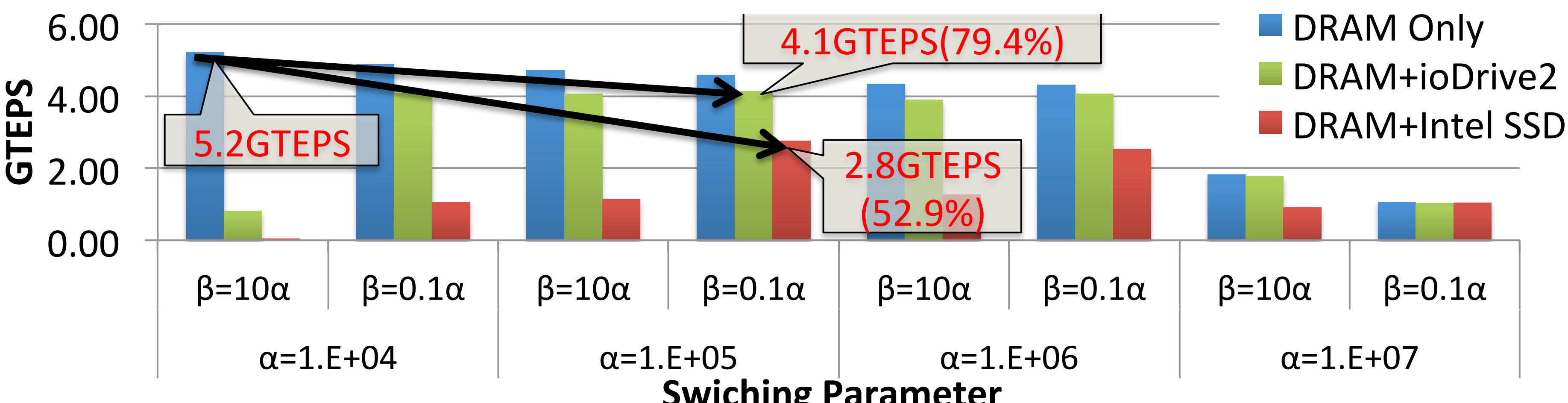
3.Proporsal

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)



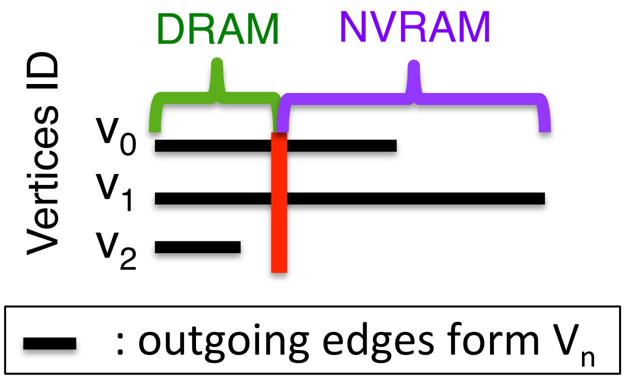
5. Current Work

EBD Algorithm Kernels

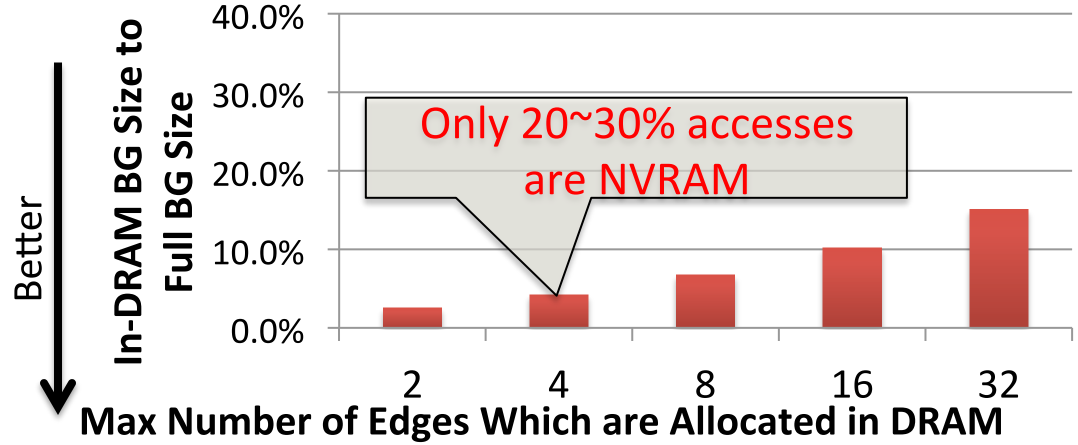
In Bottom-up approach,
all un-visited vertex have to do is find a edge which is connected to frontier's vertex.

A lot of edges are not accessed

Each vertex allocate only a few edges to DRAM



Simulation : Reduce Bottom-up Graph(BG), Scale 27



6. Related Work and Summary

- Pearce, et al. : 1 TB DRAM and 12 TB NVRAM(Fusion-io ioDrive)
  52 MTEPS [Scale 36 : 69G vertices, 1100G edges]

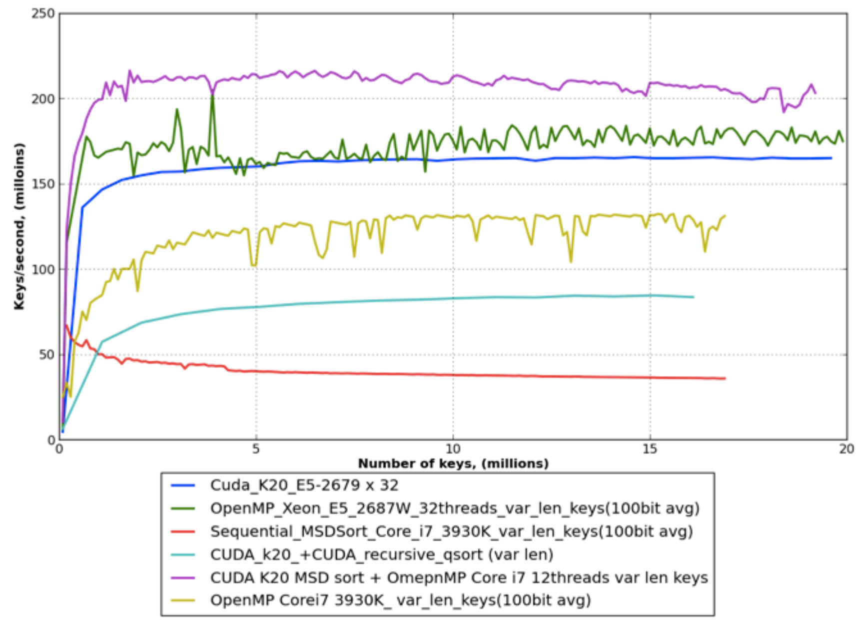
- We could reduce half the size of DRAM with 20.6% performance degradation
  (**4.1** GTEPS) **[Scale 27 : 130M** vertices, **2.1G edges]**

Roger Pearce, Maya Gokhale, Nancy M. Amato, "Scaling Techniques for Massive Scale-Free Graphs in Distributed (External) Memory" Parallel and Distributed Processing Symposium, International, 2013 IEEE 27th International Symposium on Parallel and Distributed Processing

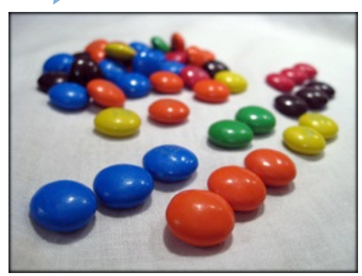
Algorithm Kernels on EBD

High Performance Sorting

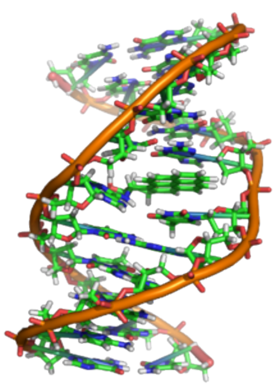
Fast algorithms:



le
icot
anana
iwi



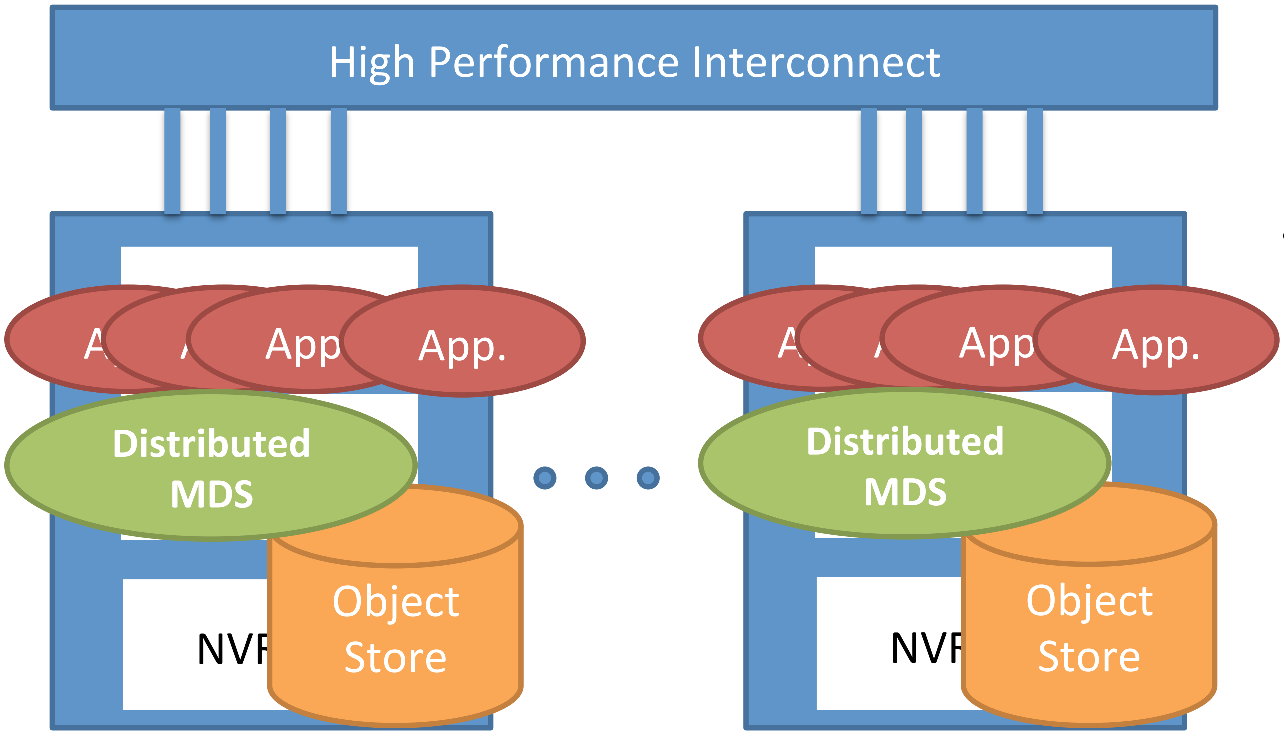
Scalability



Efficient implementation

R&D of EDB Distributed Object Store (co-PI: Osamu Tatebe, U-Tsukuba)

- Key design issues for Scaled-out IOPS and I/O bandwidth
  - Scalable distributed MDS (1M IOPS)
  - High Performance local object store
  - Efficient parallel access (100 TB/s) and parallel query



Parallel access and parallel query

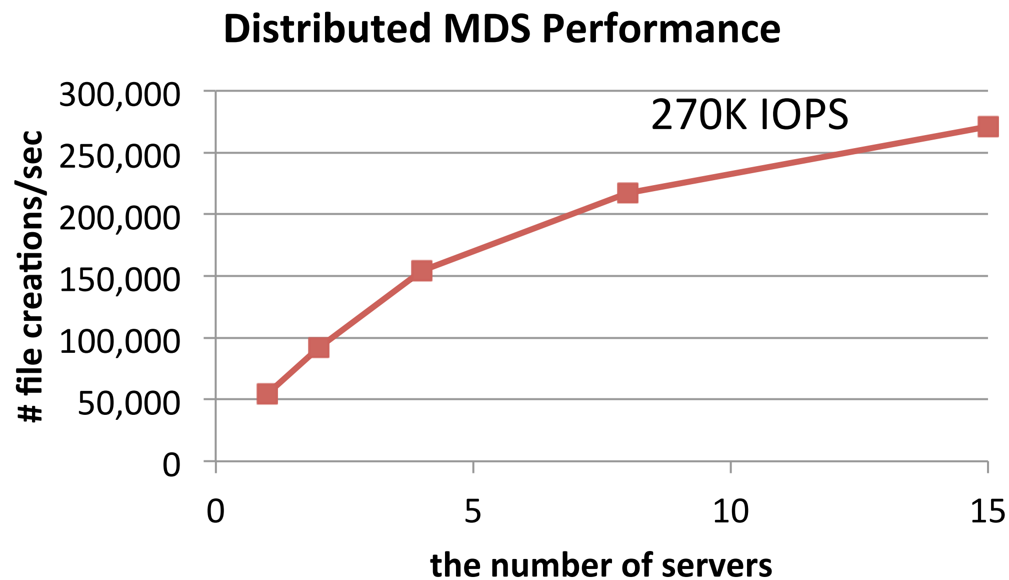
Scalable distributed MDS

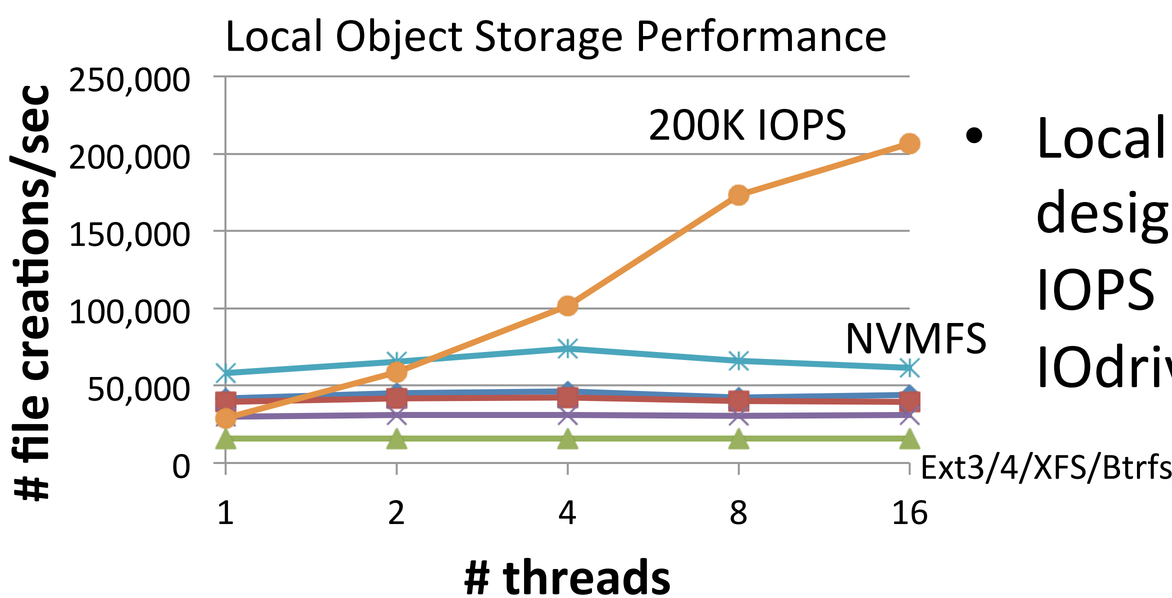
High Performance Object Store for NVRAM/flash

R&D of EDB Distributed Object Store

- Early distributed MDS design achieves 270K IOPS using 15 MDSs. [not published yet]

	Ours	GIGA+	skyFS	Lustre
IOPS	**270K**	98K	100K	80K
#servers (#cores)	15 (240)	32 (256)	32 (512)	1 (16)

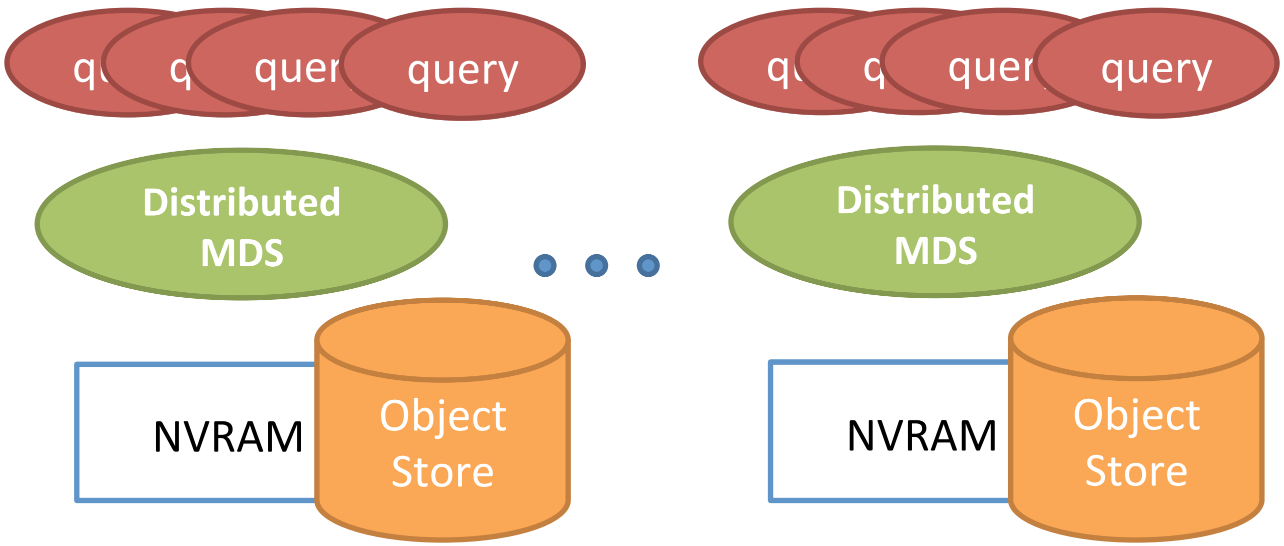




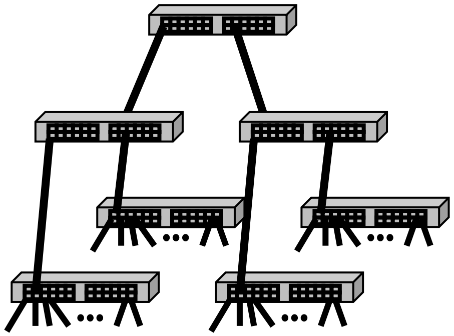
- Local Object Storage design achieves 200K IOPS using FusionIO IOdrive. [SWoPP 2013]

This year's goal

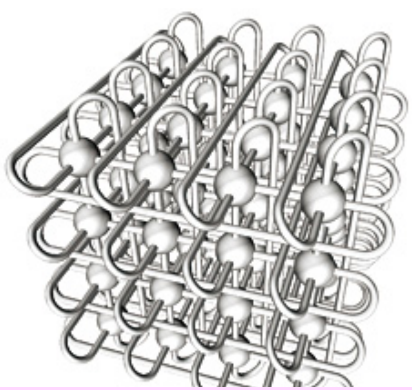
- Conceptual Design of object store
  - Distributed metadata server for O(100K) clients
  - Local object store for NVRAM/Flash
  - Parallel query to maximize data locality



EBD Interconnect (Koibuchi Group)

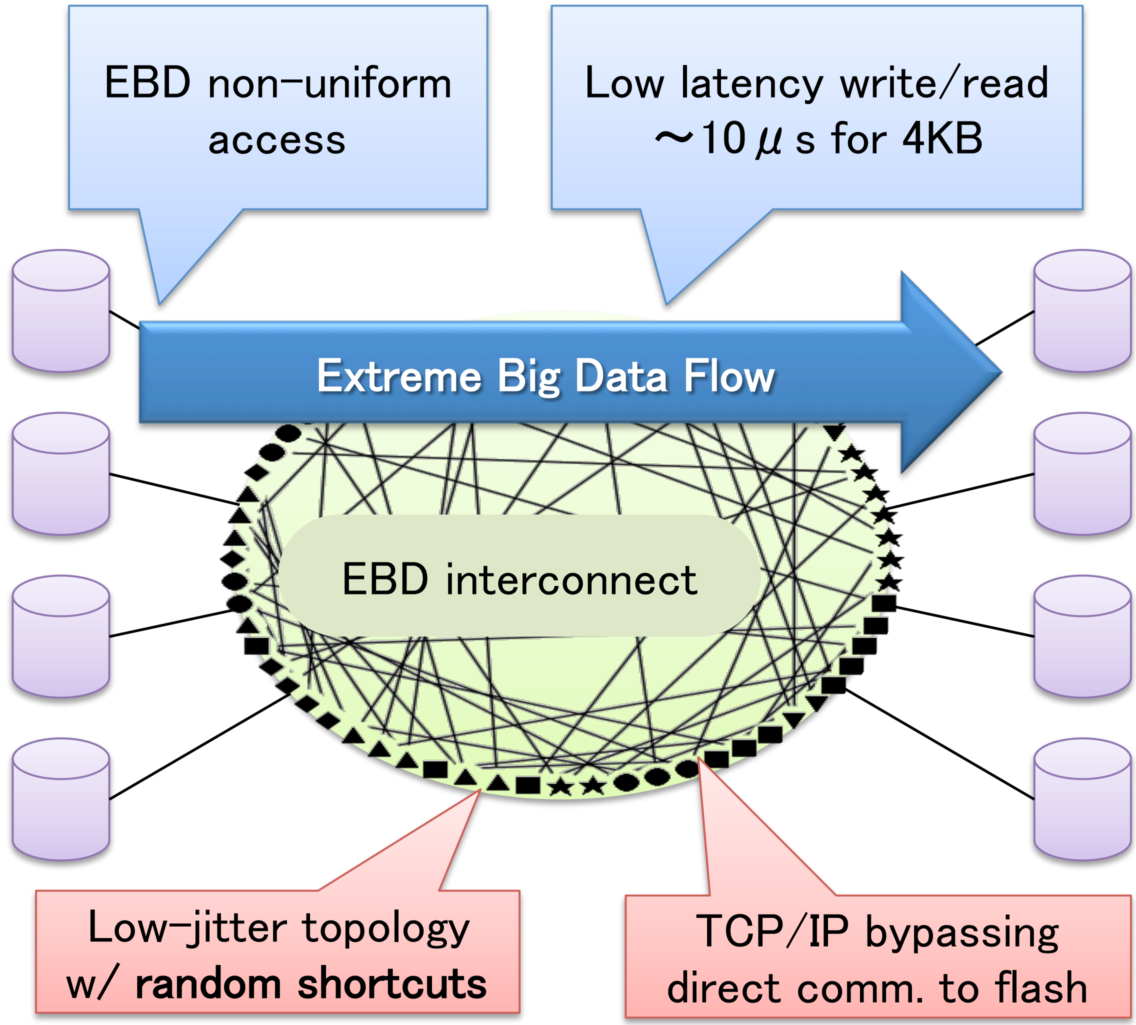


Typical Data Centers
-Poor scalability
- 1GbE + 10GbE
- TCP/IP basis



K Computer

Supercomputers
- Dedicated to neighboring and uniform access

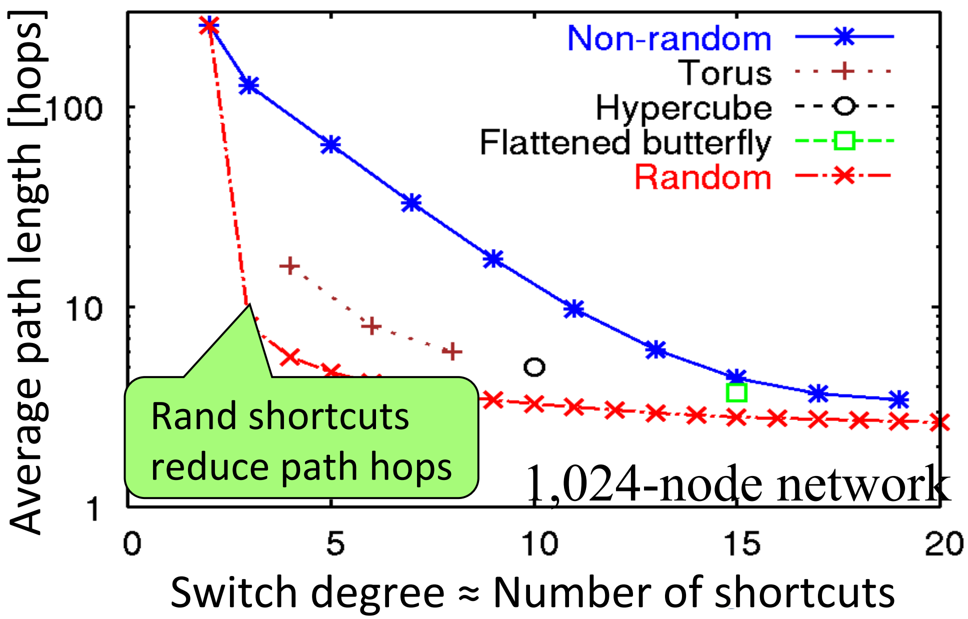


Our current technology:
Rand Topology[ISCA12] Deadlock-free routing[IEEE Trans.12]
Cabling Layout[HPCA13] Virtual routing method[IPDPS09]

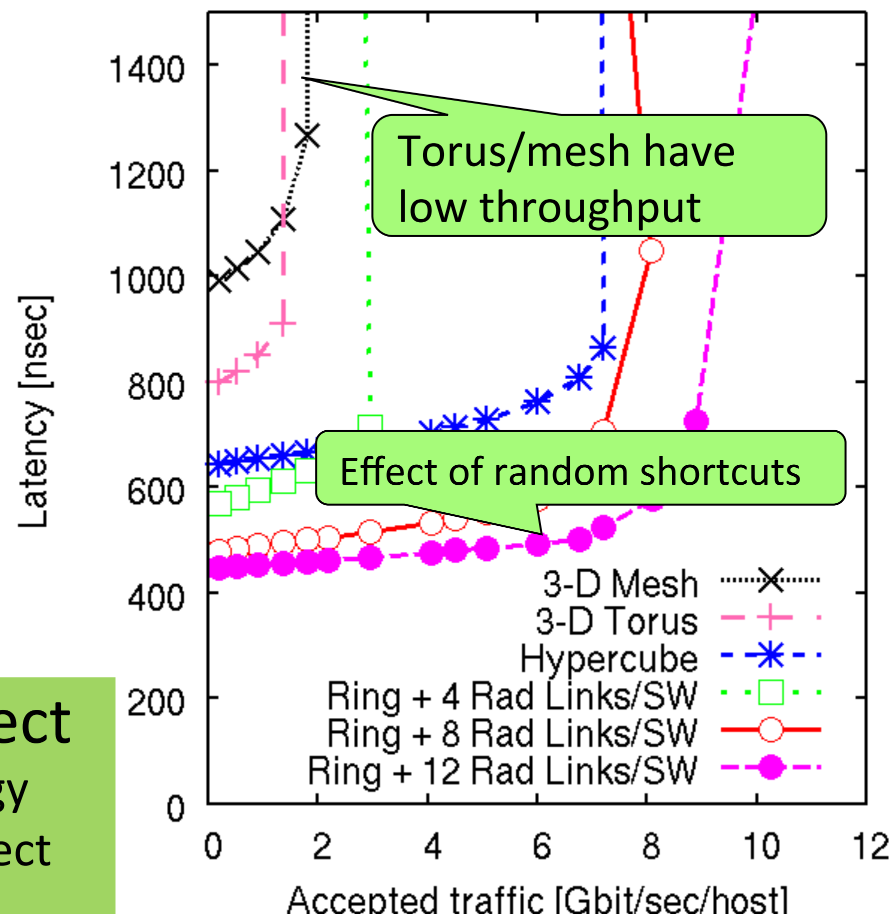
EBD Interconnects (Cont'd)

Layout-conscious Random Topology and routing

Topology Graph Analysis



Simulation results under non-uniform shuffle access



(before) ⇨ (after) EBD interconnect

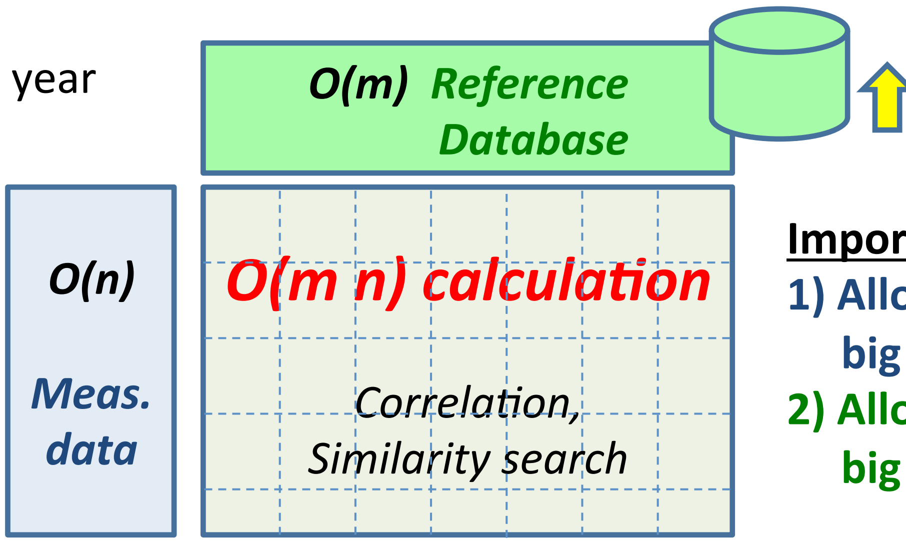
- Torus/Fat tree → Random shortcut topology
- TCP/IP comm. to HDD → TCP/IP bypass direct comm. to flash
- 1G-10/20Gbps tech. → Modern InfiniBand tech.

API co-design for complicated I/O requirements

（co-PI：Yutaka Akiyama, Tokyo Tech)

300% increase per year

Metagenome sciences



Important issues:

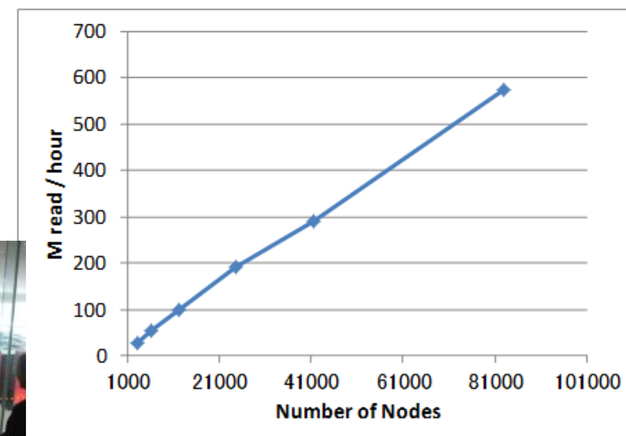
1) **Allocation of big Measured data**
2) **Allocation of big References (DB)**

Simple batch of ***BLASTX*** software

0.18 M Reads / hour
144core Xeon Cluster (2010)

3000-fold speed-up

GHOST-MP
OpenMP / MPI
load-balancing
data dispatcher

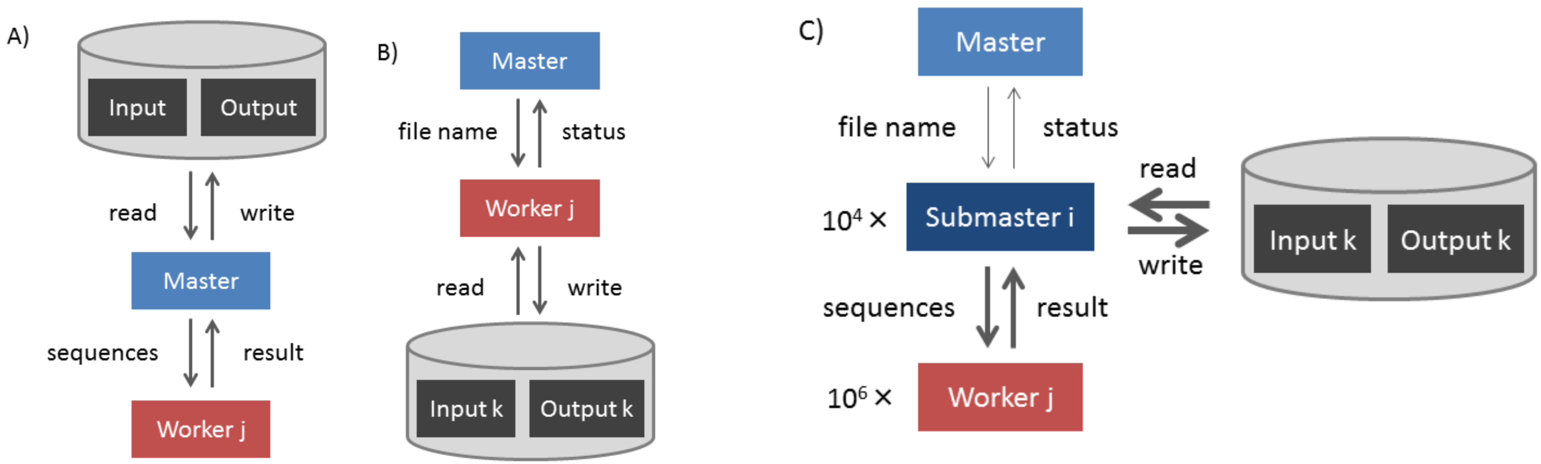


572.8 M Reads / hour
82944node K-computer (2012)

API co-design for complicated I/O requirements

1) Novel APIs for supporting abstraction of I/O

completely-centralized | completely-distributed | current complicated implementation on K-computer



simple. but need to “stage-out” millions of files

most efficient solution by now. however too much complicated.

New Idea: ***“EBD bag”*** (a kind of large-scale **Key-Value Store**)

Because most of results are **write-only**, and **independent** in time order

It **virtually enables completely distributed I/O programming** (B) efficiently..

2) System Evaluation through real big applications

Ultra-scale metagenome analysis, cancer genome, compound screening, etc.

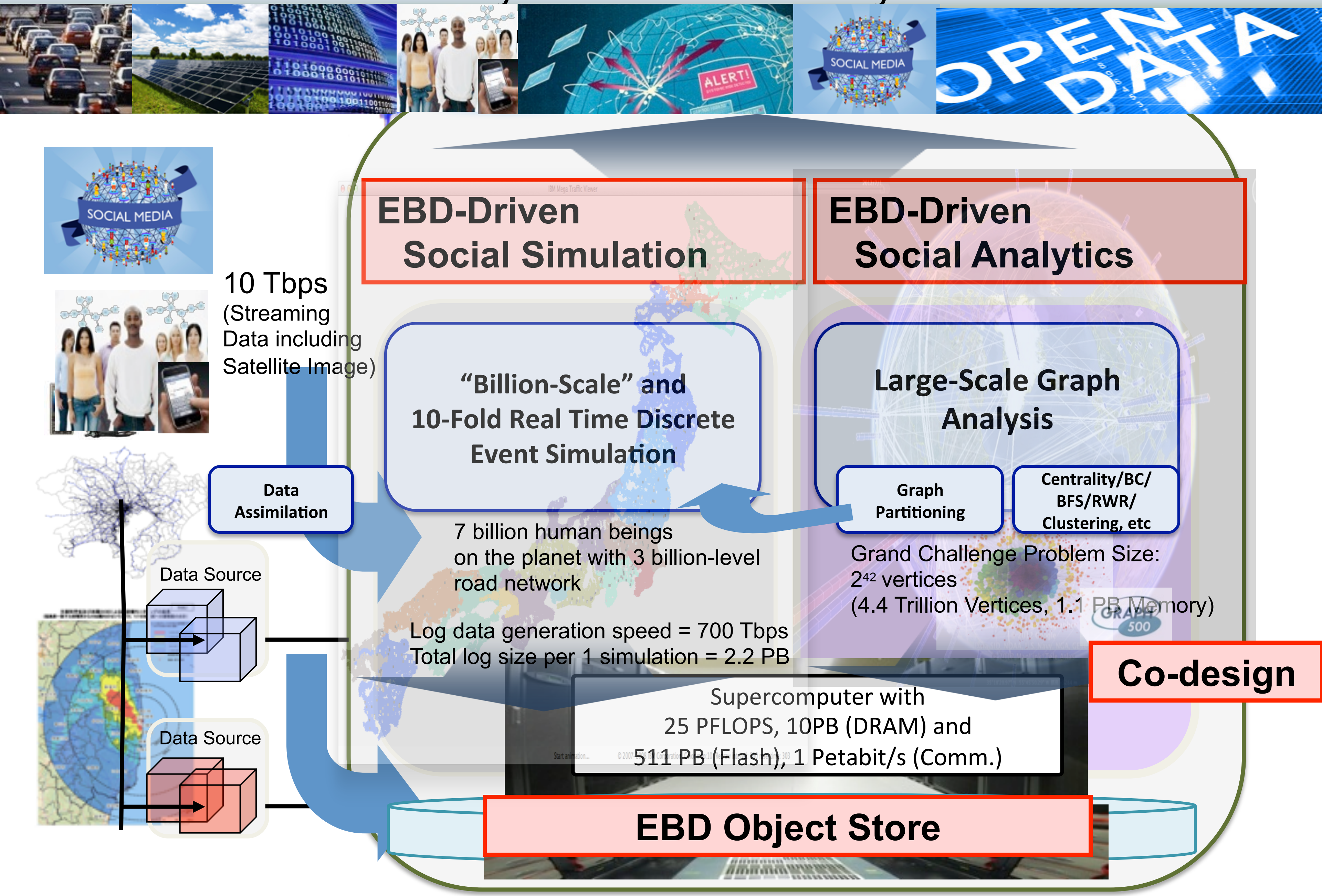
API co-design for complicated I/O requirements

Plan for H25 (FY2013)

1. Requirement Analysis and Schematic Design for New APIs
   - ✓ EBD vs. EBD collective analysis procedures
   - ✓ Proposal of the "EBD bag" function

2. Preparation for evaluation through real big applications
   - ✓ Ultra-scale Metagenome analysis: data collection and system prototyping (ex. human oral microbiome)
   - ✓ Cancer genome analysis
   - ✓ Estimation of near-future I/O requirements in related fields (genomics, proteomics, drug design, etc.)

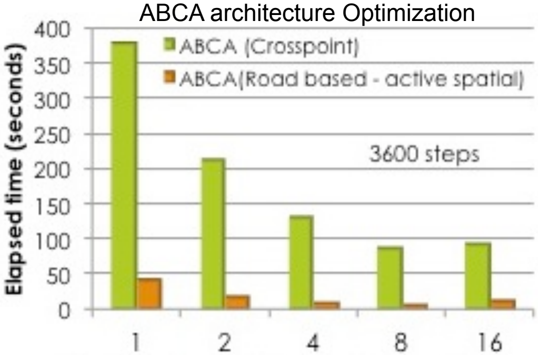
Suzumura Group
EBD Driven Planetary-Scale Social Analytics Infrastructure



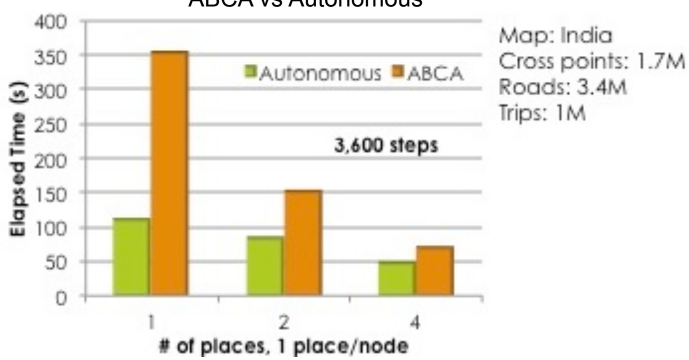
EBD Driven Planetary-Scale Social Analytics Infrastructure

A Study on Scalable Architecture and Optimization Methods for Billion-scale Social Simulation

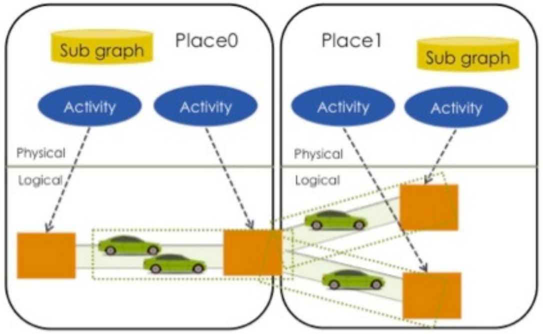
- Motivation & Goal for 2013 and 2014: Our previous design (ABCA) cannot cope with billion-scale simulation in real-time due to tremendous amount of data and I/O, so this study is to propose the best architecture that can deal with real-time billion-scale social simulation on the future hardware designed for extremely big data processing
- Study the performance characteristics of the agent-based social simulation implementations of each candidate architecture and optimization methods
  - We started investigating from billion-scale traffic simulation
- Current status: we have completed the implementations for the first two architectures and evaluated them in million-scale simulation.
- Plan by the end of this year: Complete implementations of three candidate architectures and evaluate them in billion-scale simulation
- Future: we plan to make the framework more flexible to support more complicated social simulation



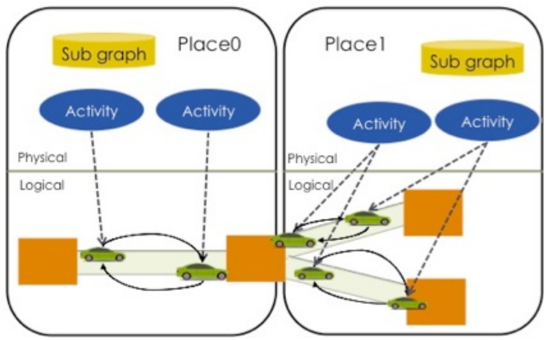
- Tokyo Map (~160K cross points, ~230K roads, 46K agents)
- Tsubame S queue machine



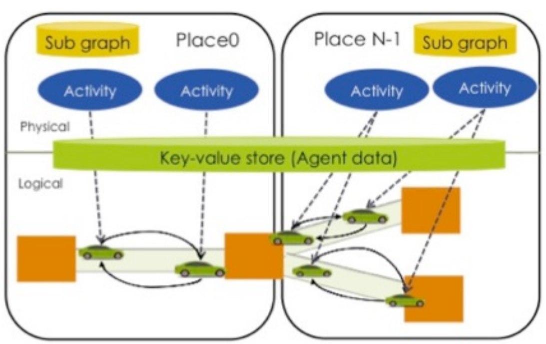
Agent-based Cellular Automata



Autonomous Architecture

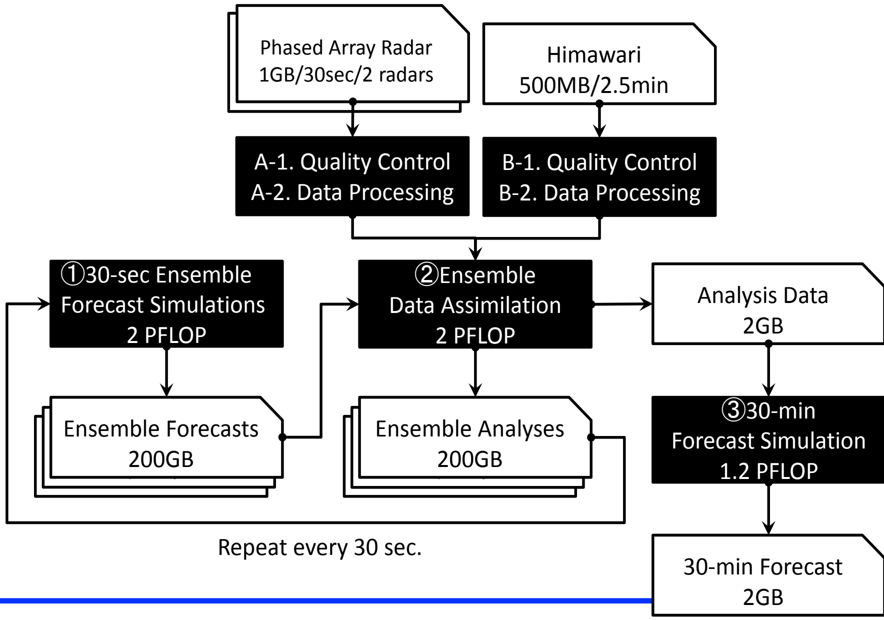


Data Storage Architecture



Fail-Safe EBD Workflow and Geometrical Search in Big Data Assimilation（co-PI : Takemasa Miyoshi, Riken AICS)

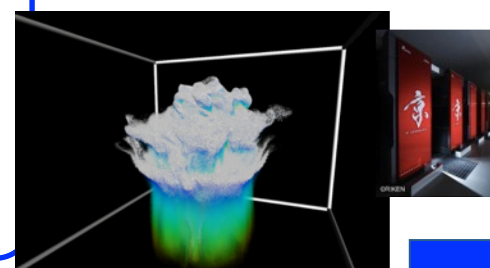
EBD Data Assimilation System in Weather Forecast (Proposed simultaneously to Prof. Tanaka's CREST)



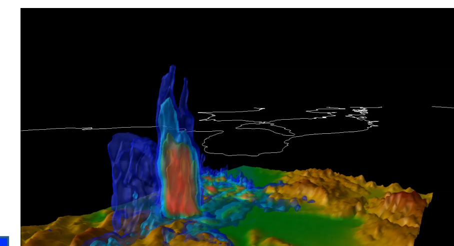
Weather Observation data keep flowing-in every 30s.

In case of hardware failure

Difficult to catch up once delayed

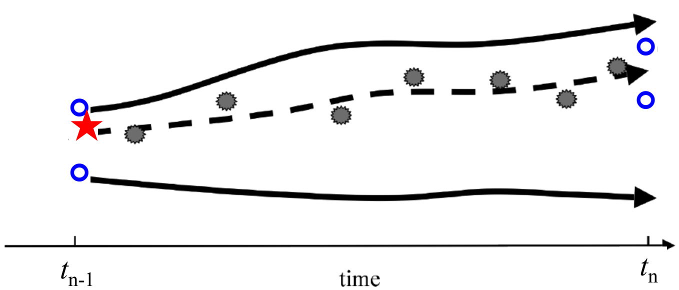


Failed Simulation



Phased Array Radar

4-dimensional Ensemble Kalman Filter 4D-LETKF



$$\tilde{\bar{\mathbf{x}}}_a(t_{n-1}) = \bar{\mathbf{x}}_a(t_{n-1}) + \mathbf{X}_a(t_{n-1})\bar{\mathbf{w}}_a(t_n)$$

$$\tilde{\mathbf{X}}_a(t_{n-1}) = \mathbf{X}_a(t_{n-1})\mathbf{W}_a(t_n)$$

$$\bar{\mathbf{w}}_a = \tilde{\mathbf{P}}_a \mathbf{Y}_b^T \mathbf{R}^{-1}(\mathbf{y} - H(\bar{\mathbf{x}}));$$

$$\mathbf{W}_a = [(K-1)\tilde{\mathbf{P}}_a]^{\frac{1}{2}}$$

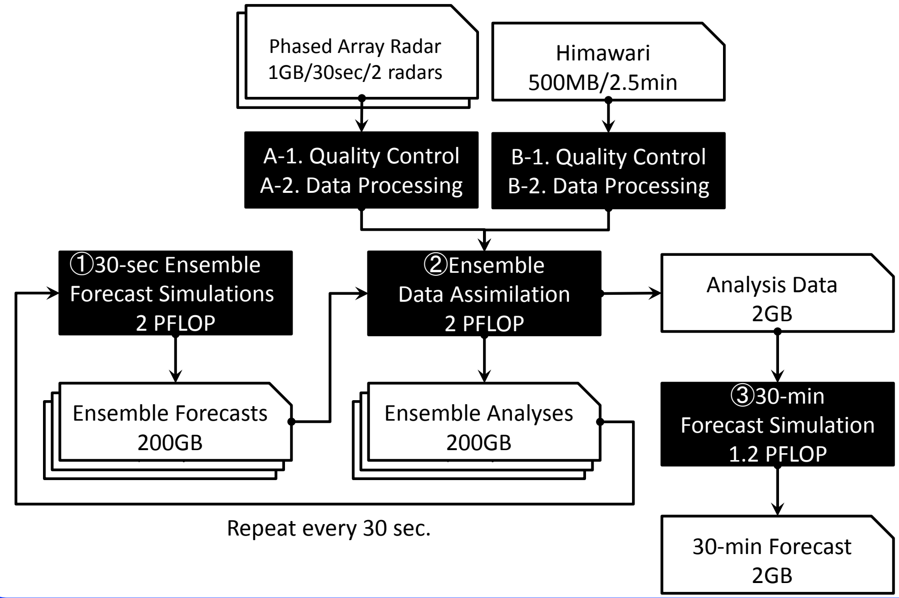
4D-LETKF enables processing multiple steps at one time

Observations at multiple times are treated simultaneously.

Highly reliable system enabling to catch up in case of delay

Fail-safe workflow

Next-generation Data Assimilation System
(Proposed simultaneously to Prof. Tanaka's CREST)



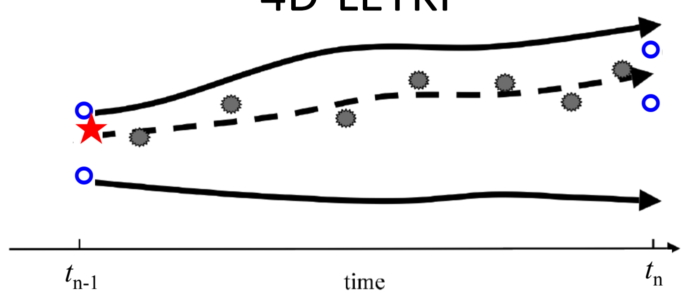
Observation data keep flowing-in every 30 sec.

In case of hardware failure

It is hard to catch up once it gets delayed

4D-LETKF enables processing multiple steps at one time

4-dimensional Ensemble Kalman Filter
4D-LETKF



$$\tilde{\bar{\mathbf{x}}}_a(t_{n-1}) = \bar{\mathbf{x}}_a(t_{n-1}) + \mathbf{X}_a(t_{n-1})\bar{\mathbf{w}}_a(t_n)$$

$$\tilde{\mathbf{X}}_a(t_{n-1}) = \mathbf{X}_a(t_{n-1})\mathbf{W}_a(t_n)$$

$$\bar{\mathbf{w}}_a = \tilde{\mathbf{P}}_a \mathbf{Y}_b^T \mathbf{R}^{-1}(\mathbf{y} - H(\bar{\mathbf{x}}));$$

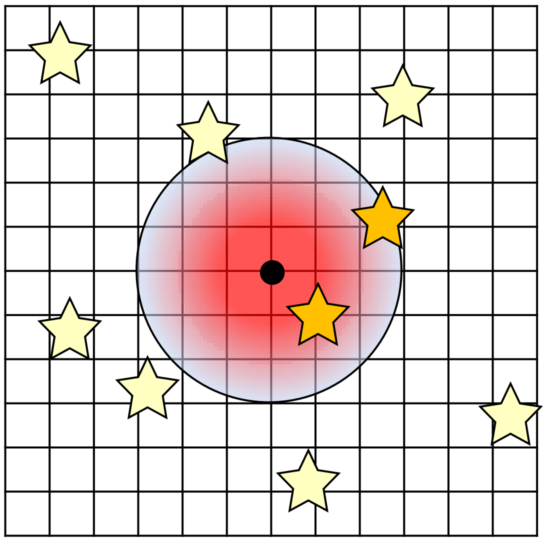
$$\mathbf{W}_a = [(K-1)\tilde{\mathbf{P}}_a]^{\frac{1}{2}}$$

Observations at multiple times are treated simultaneously.

Highly reliable system enabling to catch up in case of delay

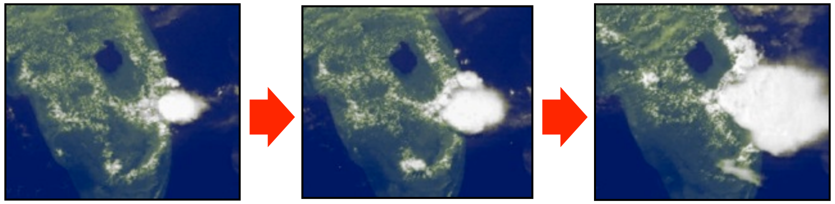
Optimizing LETKF local search

LETKF (Local Ensemble Transform Kalman Filter) includes geographical search of nearby observations around each grid point on Earth



Search $O(10^3)$ nearby observations out of total $O(10^6)$ at each of $O(10^7)$ grid points.

About half of the total LETKF computer time



Co-design of Hardware and Software

Optimizing the spatial search algorithm suitable for the converged EBD architecture

International Collaborators and Potential Industries

Alok Chaudhary
Professor, Northwestern U
Big data performance and benchmarking

Rick Stevens
Associate Laboratory Director,
Argonne National Laboratory
Convergence Architecture

Robert Ross
Math. and Computing Sciences,
Argonne National Laboratory
Distributed Big Data Objects

spec
Graph and Big Data Benchmarking

Gabriel Antoniu
Scientific leader of the KerData research team at INRIA Rennes
Distributed Filesystems

NEC hp Rambus Texas Instruments DataDirect Networks Micron nVIDIA …