Greenest Production Supercomputer in the World

Satoshi Matsuoka
Tokyo Institute of Technology

3.11 Japan Earthquake

- Top 10 Supercomputers in Japan (**96% perf. Share**) all near Tokyo!
  - 18 of 26 Japanese SCs on the Top500

- All had to be stopped due to rolling blackouts
- Despite all the research, we were not ready for this…

TOP500 Sublist Generator

R_{max} and R_{peak} values are in TFlops. For more details about other fields, check the TOP500 description.

26 entries found.

Rank	Site	System	Cores	R_{max}	R_{peak}
4	GSIC Center, Tokyo Institute of Technology Japan	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows NEC/HP	73278	1192	2287.63
33	Japan Atomic Energy Agency (JAEA) Japan	BX900 Xeon X5570 2.93GHz , Infiniband QDR Fujitsu	17072	191.4	200.08
42	University of Tokyo/Institute for Solid State Physics Japan	SGI Altix ICE 8400EX Xeon X5570 4-core 2.93 GHz, Infiniband SGI	15360	161.8	180.02
54	Japan Agency for Marine - Earth Science and Technology Japan	SX-9/E/1280M160 NEC	1280	122.4	131.07
59	JAXA Japan	Fujitsu FX1, Quadcore SPARC64 VII 2.52 GHz, Infiniband DDR Fujitsu	12032	110.6	121.28
70	Information Technology Center, The University of Tokyo Japan	Hitachi opteron QC 2.3 GHz Myrinet 10G Hitachi	15104	101.74	138.96
74	Institute of Physical and Chemical Res. (RIKEN) Japan	RIKEN Intergrated Cluster of Clusters, Xeon X5570 2.93GHz, Infiniband DDR Fujitsu	9048	97.94	106.04
88	GSIC Center, Tokyo Institute of Technology Japan	Sun Fire x4600/x6250, Opteron 2.4/2.6 GHz, Xeon E5440 2.833 GHz, ClearSpeed CSX600, nVidia GT200; Voltaire Infiniband NEC/Sun	31024	87.01	163.19
96	Center for Computational Sciences, University of Tsukuba Japan	Appro Xtreme-X3 Server - Quad Opteron Quad Core 2.3 GHz, Infiniband Appro International	10368	77.28	95.39
102	National Institute for Environmental Studies Japan	Asterism ID318, Intel Xeon E5530, NVIDIA C2050, Infiniband NSSOL / SGI Japan	144640	74.84	177.12
131	National Institute for Fusion Science (NIFS) Japan	Hitachi SR16000 Model L2/128, Power6 4.7Ghz, Infiniband Hitachi	4096	56.65	77

So What Can We Do, As Computer Scientists?

- Green Computing (Energy and Volatility)
  - JST CREST Megascale Project [2003-2007]
  - JST CREST Ultra Low-Power HPC [2007-2012]
  - Tsubame2.0 "Green Supercomputer" [2010-2014]
  - Green Supercomputing [2011-2015]
- Using IT to Analyze and Optimize Energy
  - Various Environmental and Device Simulations (on Tsubame2.0)
  - Supercomputing the Smart Grid – Post Petascale Graph Processing

JST-CREST "MegaScale" [2003-2007]

- *"Embedded tech. meets HPC"*
- Increased importance w/emphasis on mobile/ubiquitous computing
- Attain ~x2 with effective **DVS**

=> 2004 CTWatch J. (Matsuoka ed.)

- Can't attain x10 tho…
  - Low hanging fruit achieved with simple DVS
  - Memory, NW, I/O… starts to dominate
  - Simple slow cores limited to weak scaling
  - Conclusion: new devices, new architectures, cooling, apps need to be exploited and optimized as a whole

JST-CREST
MegaProto (SC2005)
17Sockets Effecion/1U
64GFlops/300W

MegaProto [SC05] Computing Element

- V1: TM5900
- 0.9GFlops@7W
- ⇒ 112.5MFlops/W
- L2C=512KB

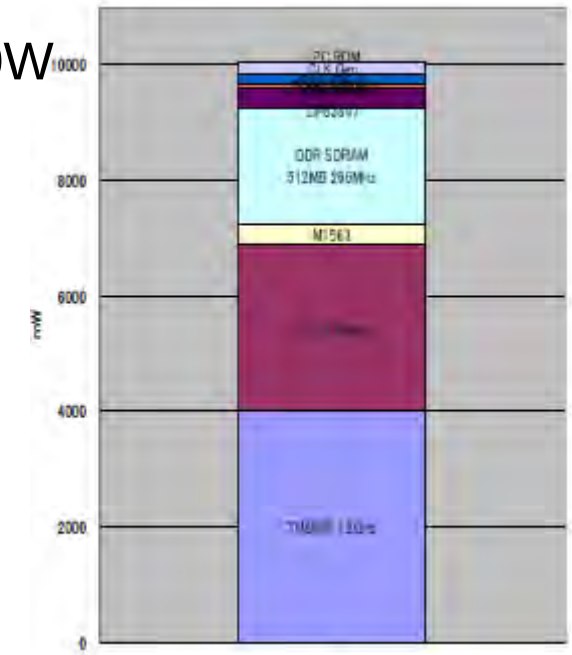
- V2: Effecion TM8800
- 2.4GFlops@5W
- ⇒ 480MFlops/W
- 512MB DDR-SDRAM

- 256MB SDRAM
  - (512M DDR in V2)
- 512KB flash

MegaProto Packaging Metrics

- Goal: 300W/1U system
  - CPU = 4-5W
  - Card = 10W⇒ x17=170W
  - MB (Networks, interface, HDD) = 70W
    - Low power commodity fabric still problematic
  - Fans = ~20W
  - Power Supply Loss = 85% ~ 40W
  - **Total: ~300W**
    - **(c.f. Xeon 3Ghz Dual 1U ~ 400W)**
  - Current: .933 GF x 16 / 300W = 50MFlops/W (15GigaFlops/1U)
  - **Next Version: 2.4 GF x 16 / 300W = 130MFlops/W (~40GigaFlops/1U)**
    - x3 improvement in perf/power

10W



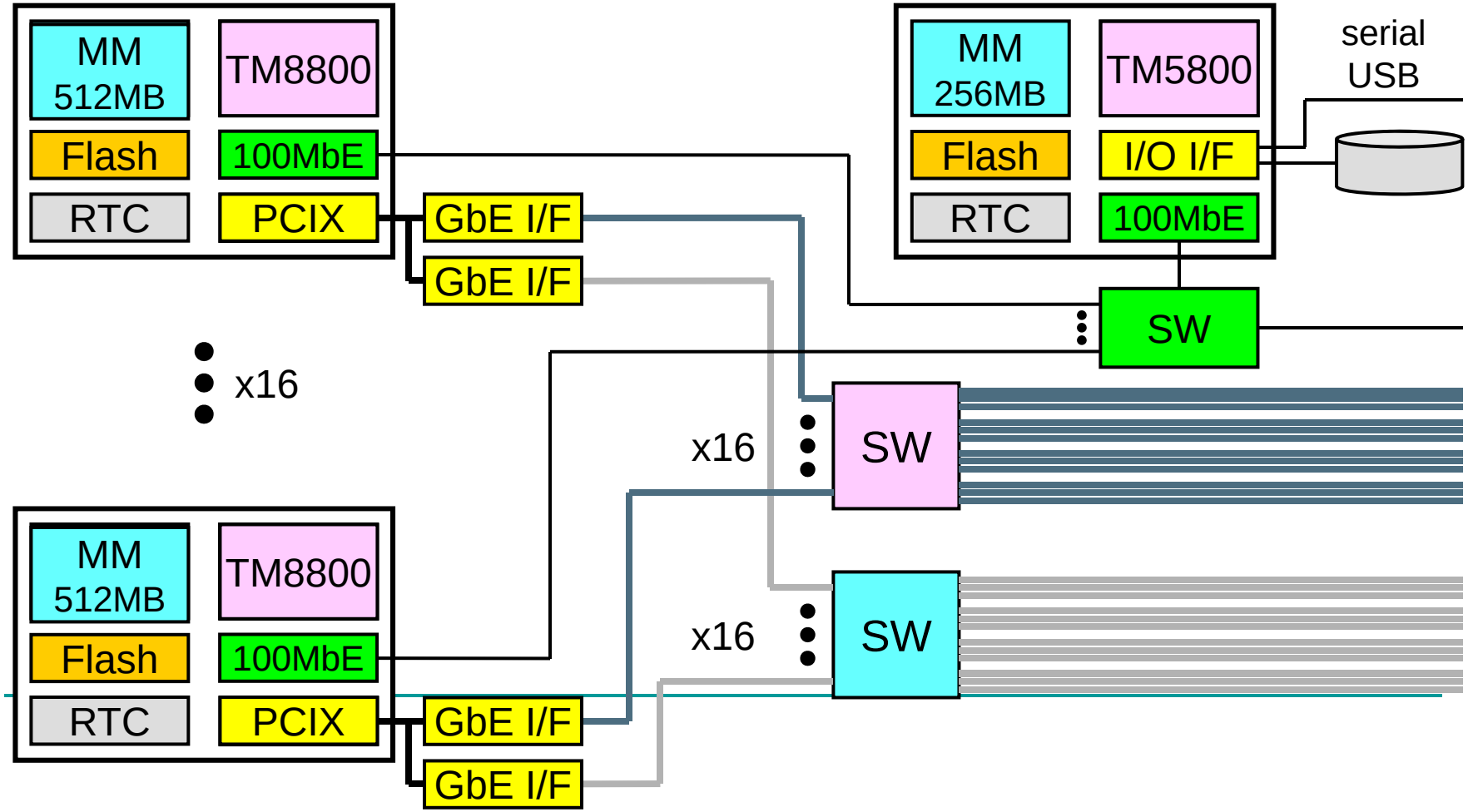
Power Budget for next gen TM8800 node

MegaProto System

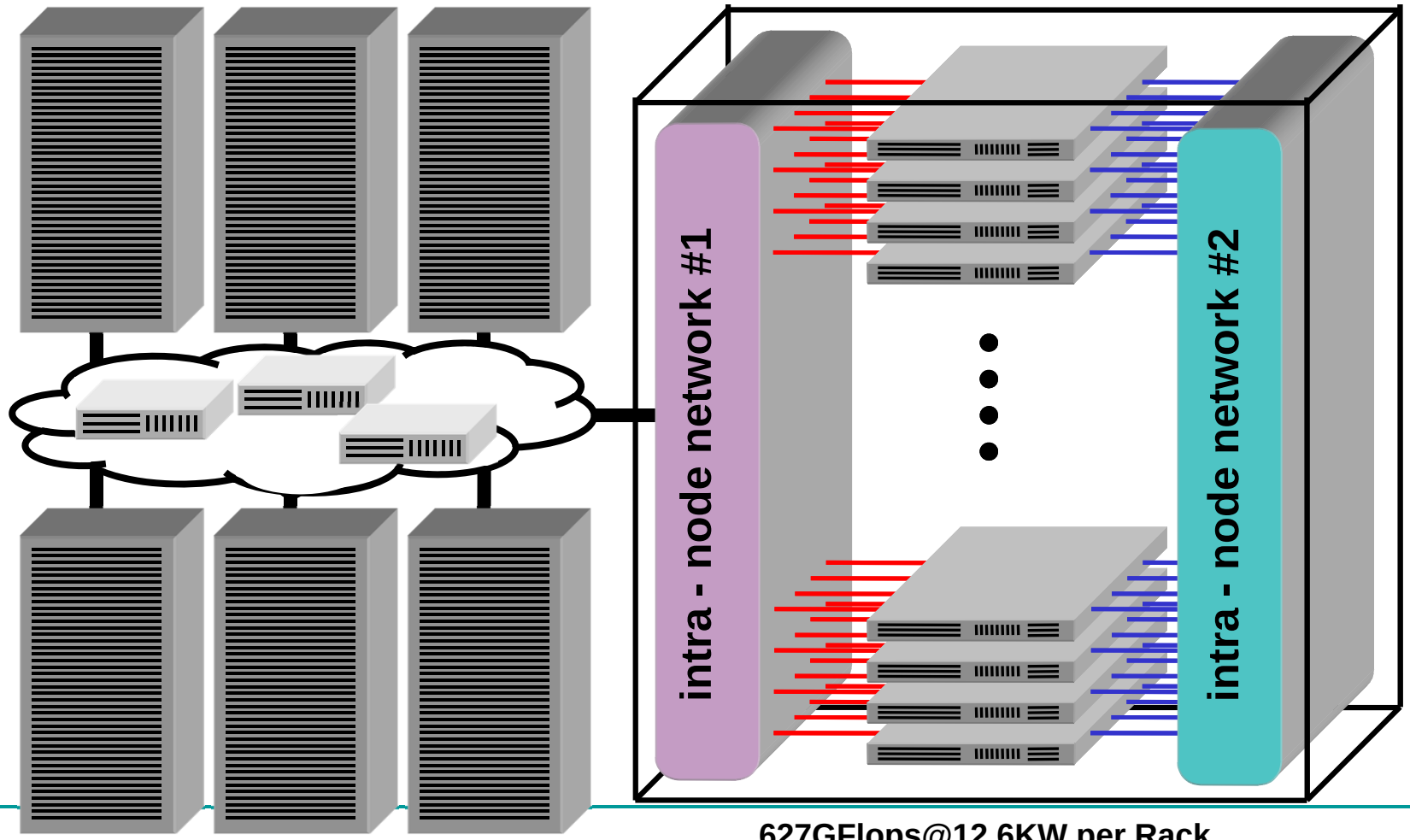
- Fabrication by IBM Japan (same fab team as BlueGene/L)
- 8 Units delivered, working
- 5 Units in our lab

MegaProto Integrated Network

- Low power, HP building block
- Reduce external switch requirements



MegaProto System Configuration



627GFlops@12.6KW per Rack
=> V2 1.61TFlops, 1.41TB/s MemBW,
1.3Gbps NetBW

The TSUBAME 1.0 “Supercomputing Grid Cluster” (Production) Spring 2006

Unified IB network

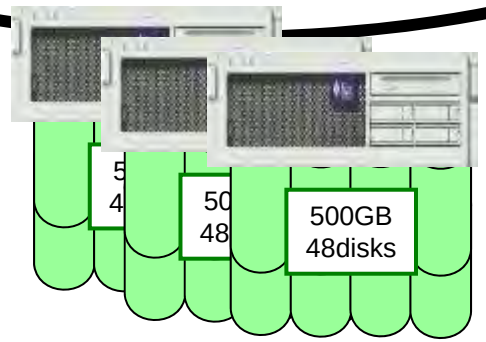
Voltaire ISR9288 Infiniband 10Gbps
x2 (DDR next ver.)
~1310+50 Ports
~13.5Terabits/s (3Tbits bisection)

10Gbps+External Network

“Fastest Supercomputer in Asia” 7th on the 27th Top500@38.18TF

Sun Galaxy 4 (Opteron Dual core 8-socket)
10480core/655Nodes
21.4Terabytes
50.4TeraFlops
OS Linux (SuSE 9, 10)
NAREGI Grid MW

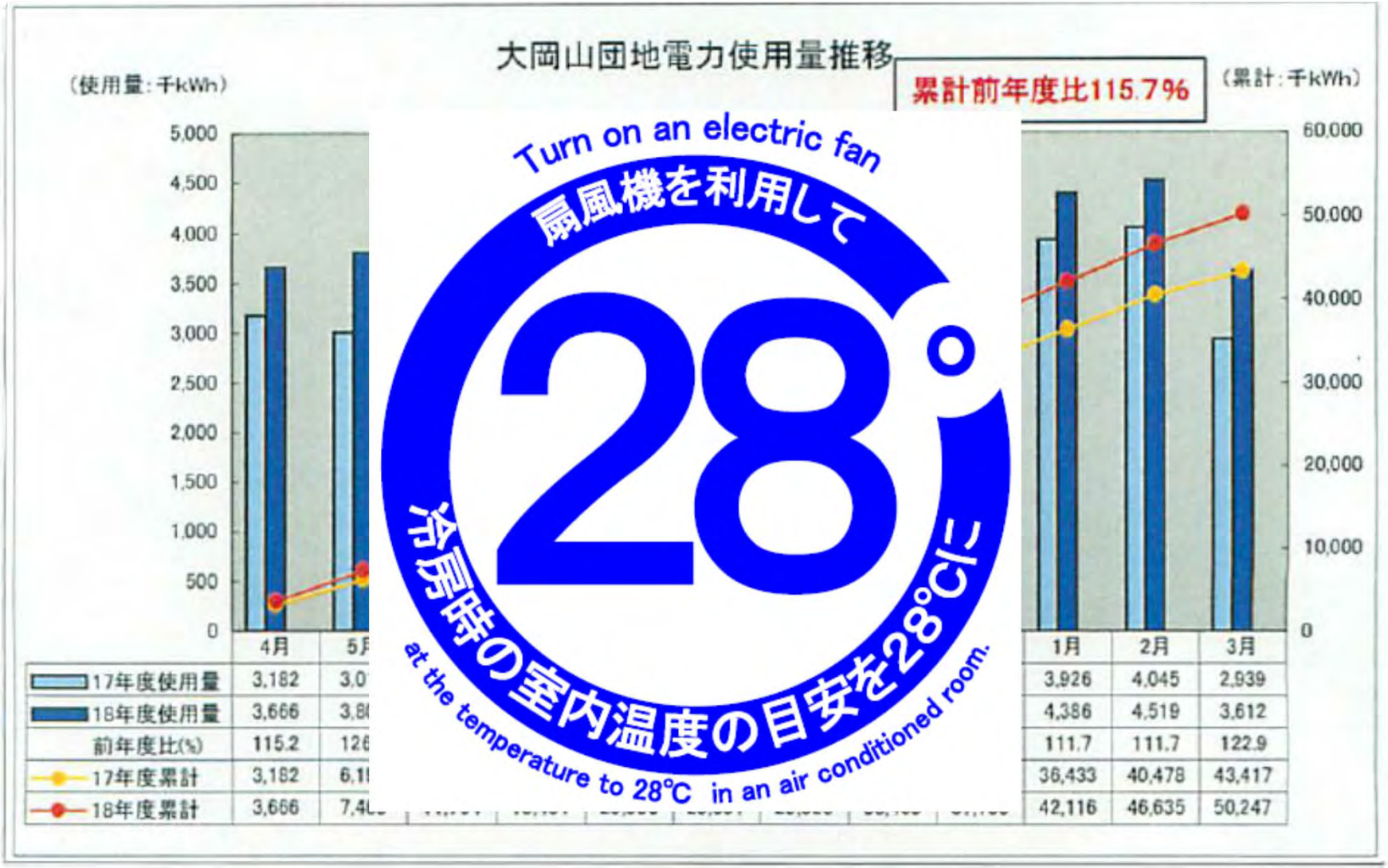
NEC SX-8i
(for porting)



Storage
1.0 Petabyte (Sun “Thumper”)
0.1Petabyte (NEC iStore)
<u>Lustre FS, NFS, WebDAV (over IP)</u>
50GB/s aggregate I/O BW

ClearSpeed CSX600
SIMD accelerator
360 boards,
35TeraFlops(Current))

You know you have a problem when, …



	4月	5月	1月	2月	3月
17年度使用量	3,182	3,0	3,926	4,045	2,939
18年度使用量	3,666	3,8	4,386	4,519	3,612
前年度比(%)	115.2	126	111.7	111.7	122.9
17年度累計	3,182	6,1	36,433	40,478	43,417
18年度累計	3,666	7,4	42,116	46,635	50,247

Biggest Problem is Power…

Machine	CPU Cores	Watts	Peak GFLOPS	Peak MFLOPS/ Watt	Watts/ CPU Core	Ratio c.f. TSUBAME
TSUBAME(Opteron)	10480	800,000	50,400	63.00	76.34	
TSUBAME2006 (w/360CSs)	11,200	810,000	79,430	98.06	72.32	
TSUBAME2007 (w/648CSs)	11,776	820,000	102,200	124.63	69.63	1.00
Earth Simulator	5120	6,000,000	40,000	6.67	1171.88	0.05
ASCI Purple (LLNL)	12240	6,000,000	77,824	12.97	490.20	0.10
AIST Supercluster (Opteron)	3188	522,240	14400	27.57	163.81	0.22
LLNL BG/L (rack)	2048	25,000	5734.4	229.38	12.21	1.84
Next Gen BG/P (rack)	4096	30,000	16384	546.13	7.32	4.38
TSUBAME 2.0 (2010Q3/4)	160,000	810,000	1,024,000	1264.20	5.06	10.14

TSUBAME 2.0 x24 improvement in 4.5 years…? ➔ ~ x1000 over 10 years

Arithmetic INTENSITY: FLOP/Byte

GP GPU

FLOP = number of FP operation for applications
Byte = Byte number of memory access for applications

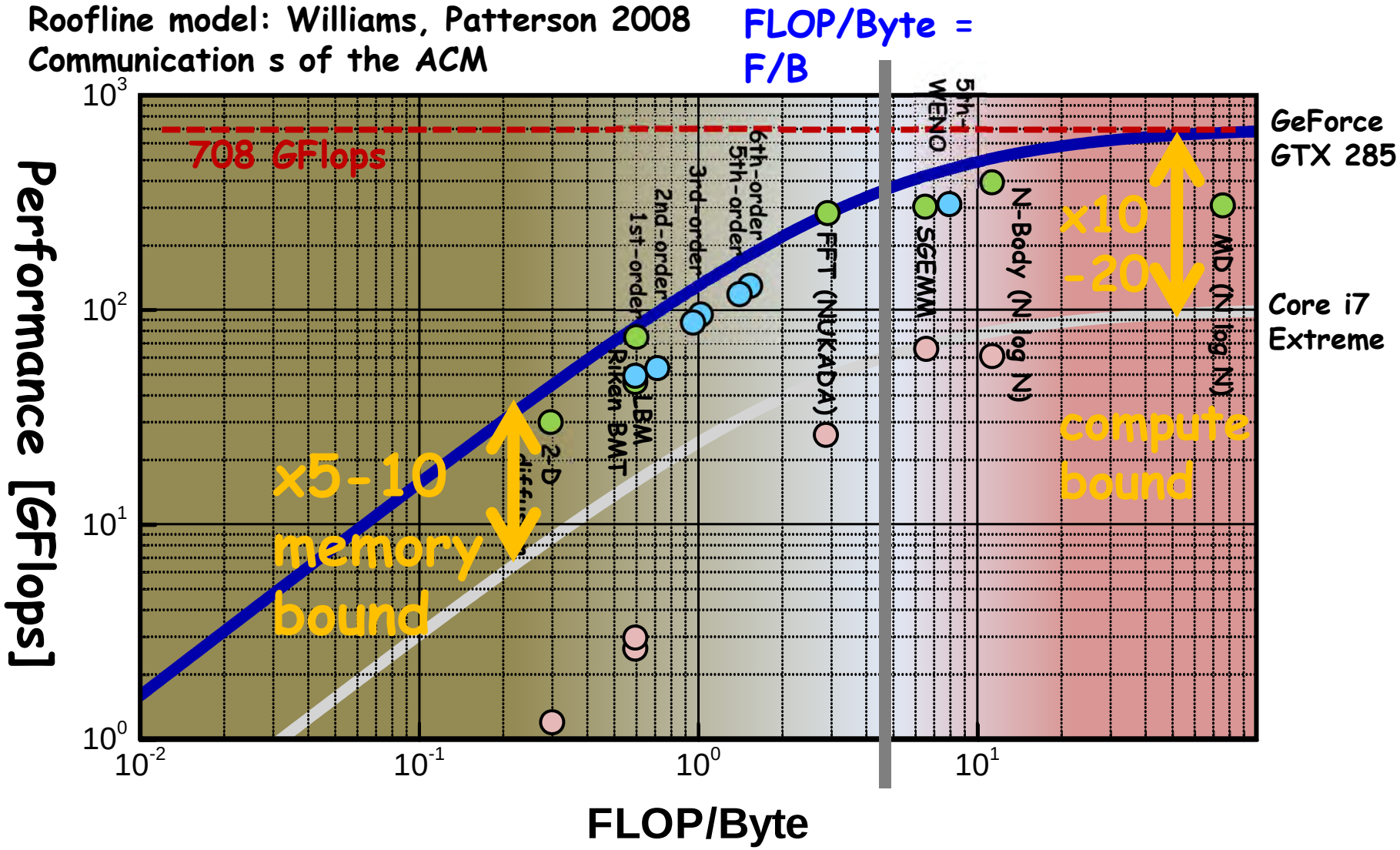
F = Peak Performance of floating point operation
B = Peak Memory Bandwidth

$$\text{Performance} = \frac{\text{FLOP}}{\text{FLOP/F} + \text{Byte/B} + \alpha}$$

$$= \frac{\text{FLOP/Byte}}{\text{FLOP/Byte} + \text{F/B} + \alpha}\text{F}$$

GPU vs. CPU Performance

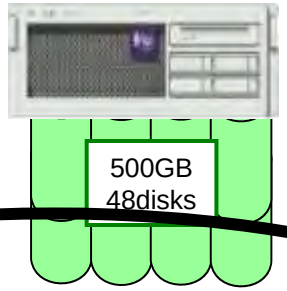
Roofline model: Williams, Patterson 2008
Communication s of the ACM



TSUBAME 1.2 Experimental Evolution (Oct. 2008)

Voltaire ISR9288 Infiniband x8
10Gbps x2 ~1310+50 Ports
~13.5Terabits/s
(3Tbits bisection)

NEC SX-8i



Storage
1.5 Petabyte (Sun x4500 x 60)
0.1Petabyte (NEC iStore)
Lustre FS, NFS, CIF, WebDAV (over IP)
60GB/s aggregate I/O BW

10Gbps+External NW

Unified Infiniband network

87.1TF Linpack (The First GPU Supercomputer on Top500) [IPDPS10]

Sun x4600 (16 Opteron Cores)
32~128 GBytes/Node
10480core/655Nodes
21.4TeraBytes
50.4TeraFlops
OS Linux (SuSE 9, 10)
NAREGI Grid MW

NEW Deploy:
GCOE TSUBASA
Harpertown-Xeon
90Node 720CPU
8.2TeraFlops

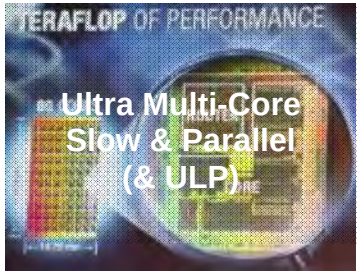
PCI-e

ClearSpeed CSX600
SIMD accelerator
648 boards,
52.2TeraFlops
SFP/DFP

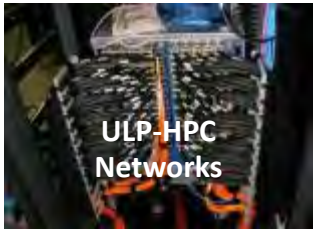
New Production Experiment

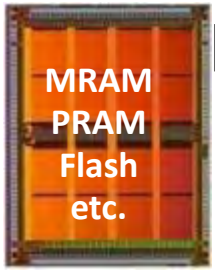
170 Nvidia Tesla 1070, ~680 Tesla cards
High Performance in Many BW-Intensive Apps
10% power increase over TSUBAME 1.0

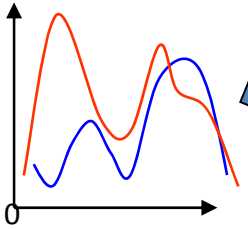
JST-CREST "Ultra Low Power (ULP)-HPC" Project 2007-2012



ULP-HPC SIMD-Vector (GPGPU, etc.)







Low Power
High Perf Model

Power Optimize using Novel Components in HPC

Auto-Tuning for Perf. & Power

モデルと実測の Bayes 的融合

- Bayes モデルと事前分布

$$y_i \sim N(\mu_i, \sigma_i^2)$$

$$\mu_i \mid \beta, \sigma_i^2 \sim N(x_i^T \beta, \sigma_i^2 / \kappa_0)$$

$$\sigma_i^2 \sim \mathrm{Inv}\text{-}\chi^2(\nu_0, \sigma_0^2)$$

モデルによる所要時間の推定

- n 回実測後の事後予測分布

所要時間の実測データ

$$y_i \mid (y_{i1}, y_{i2}, \cdots, y_{in}) \sim t_{\nu_n}(\mu_{in}, \sigma_{in}^2 \kappa_{n+1} / \kappa_n)$$

$$\nu_n = \nu_0 + n, \quad \kappa_n = \kappa_0 + n, \quad \mu_n = (\kappa_0 x_i^T \beta + n\bar{y}_i) / \kappa_n$$

$$\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + \sum (y_m - \bar{y}_i)^2 + \kappa_0 n (\bar{y}_i - x_i^T \beta)^2 / \kappa_n$$

$$\bar{y}_i = \frac{1}{n} \sum_m y_m$$

ABCLibScript: アルゴリズム選択

```
!ABCLib$ static select region start
!ABCLib$ parameter (in CacheS, in NB, in NPrc)
!ABCLib$    select sub region start
!ABCLib$    according estimated
!ABCLib$    (2.0d0*CacheS*NB)/(3.0d0*NPrc)
            対象1（アルゴリズム1）
!ABCLib$    select sub region end
!ABCLib$    select sub region start
!ABCLib$    according estimated
!ABCLib$    (4.0d0*ChcheS*dlog(NB))/(2.0d0*NPrc)
            対象2（アルゴリズム2）
!ABCLib$    select sub region end
!ABCLib$ static select region end
```

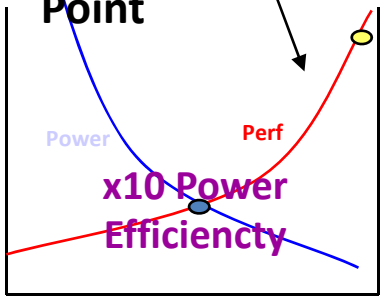
実行起動前自動チューニング指定、アルゴリズム選択処理の指定

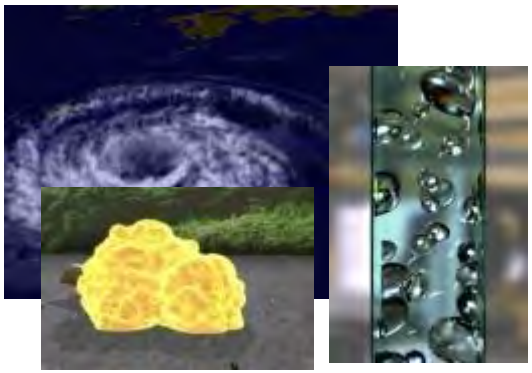
コスト定義関数で使われる入力変数

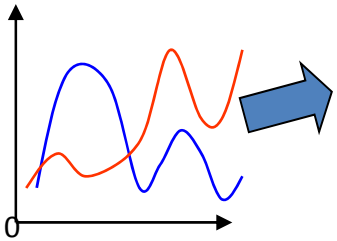
コスト定義関数

対象領域1、2

Optimization Point





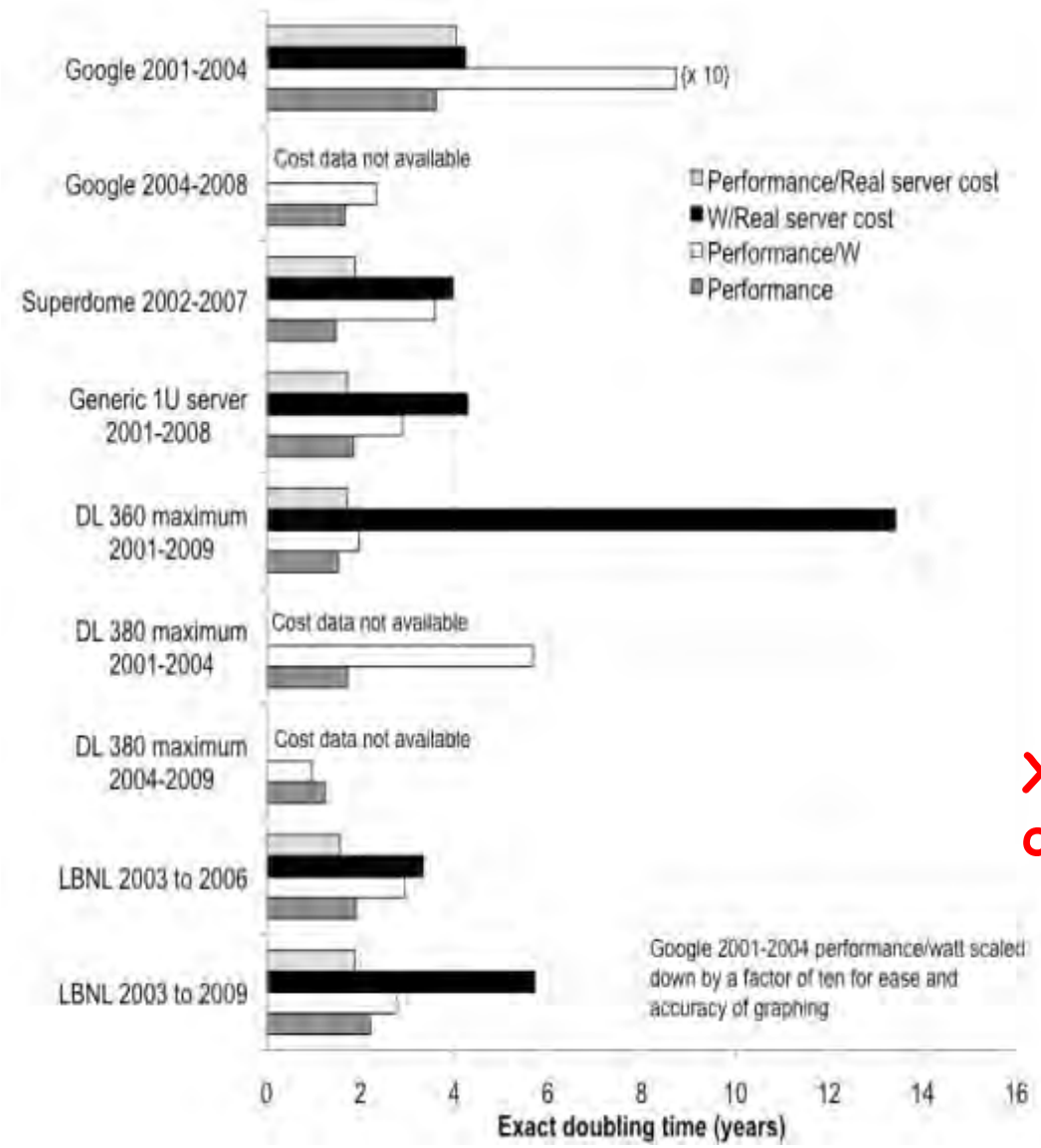


Perf. Model
Algorithms

Power-Aware and Optimizable Applications

x1000
Improvement in 10 years

Modern Server Speedups- x2 in 2 years



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

"Performance per server and performance per thousand dollars of purchase cost double every two years or so, which tracks the typical doubling time for transistors on a chip predicted by the most recent incarnation of Moore's law"

X32 in 20 years -> x30 discrepancy

Aggressive Power Saving in HPC

Methodologies	Enterprise/Business Clouds	HPC
Server Consolidation	Good	NG!
DVFS (Differential Voltage/Frequency Scaling)	Good	Poor
New Devices	Poor (Cost & Continuity)	Good
New Architecture	Poor (Cost & Continuity)	Good
Novel Cooling	Limited (Cost & Continuity)	Good (high thermal density)

How do we achive x1000?

Process Shrink x100

X

Many-Core GPU Usage x5

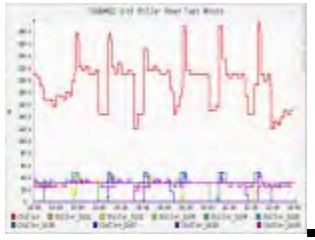
X

DVFS & Other SW x1.5

X

Efficient Cooling x1.4

x1000 !!!



ULP-HPC Project 2007-12

Ultra Green Supercomputing Project 2011-15

A precise measurement tool for power dissipation of CUDA kernels

motivation

To test the power dissipation of CUDA kernel precisely:

- Labview programming
- Data process and analysis
- Manual operations are required

Problem: It is complex to operate with the equipment and require manual operation

Goal

- Automatic power dissipation measurement
- Automatic data process and analysis
- User friendly APIs

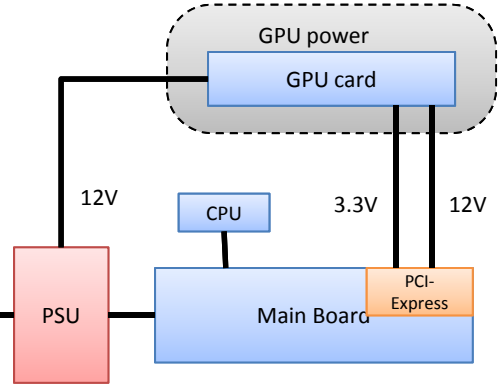
measurement

Power supply

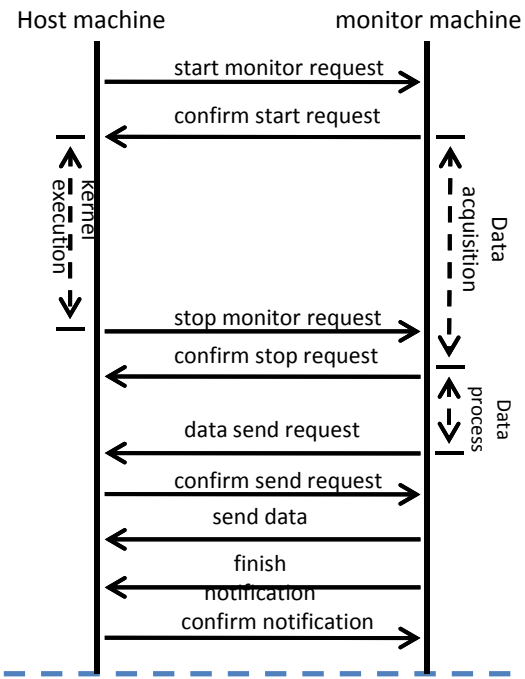
PSU:12V power line

PCI-Express bus:

- 12V power line
- 3.3V power line

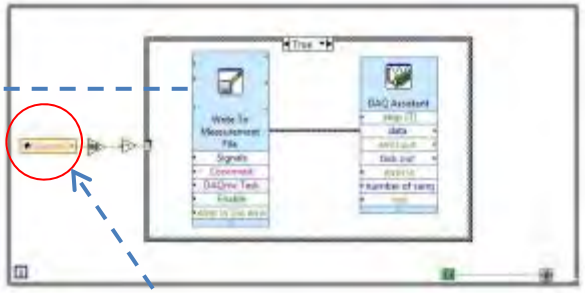


communication



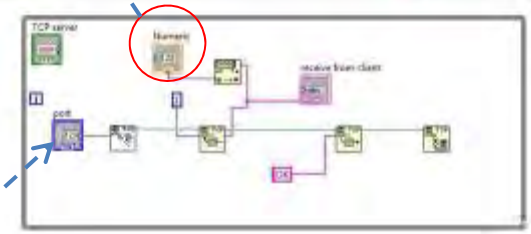
Data acquisition and process

Local files



Marker:

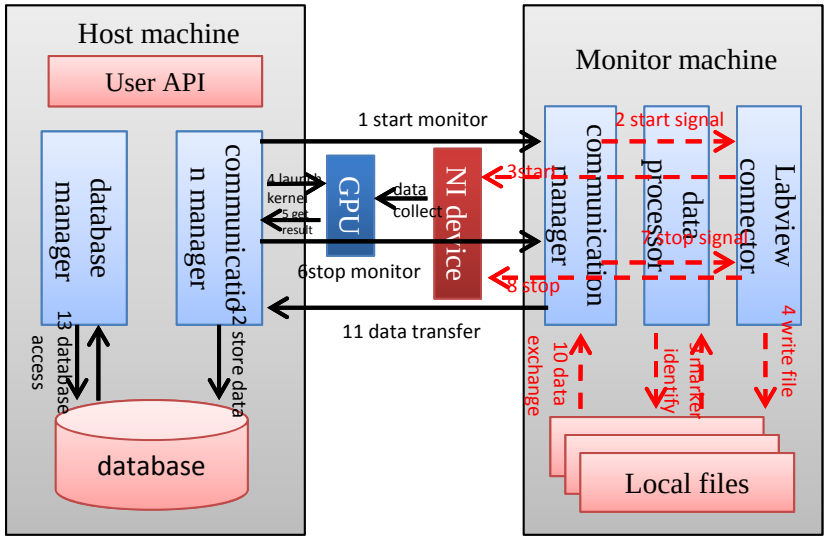
- heartbeat line benchmark
- Launch it before and after
- kernel Analyze and identify the markers, only keep the data between the two marker



TCP connection

A precise measurement tool for power dissipation of CUDA kernels

User process



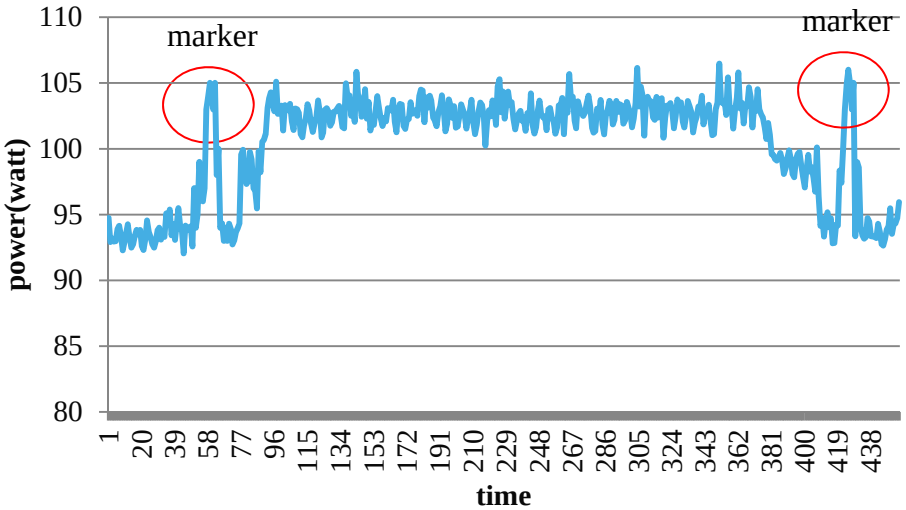
User APIs

```
//Example:
int main()
{
...
...
P_startMonitor();
P_marker();
Launch_kernel();
cudaThreadSynchronize();
P_marker();
P_stopMonitor();
int num ;
Struct record* rec;
num
=P_receiveData(&rec);
P_storeData(rec,num);
...
}
```

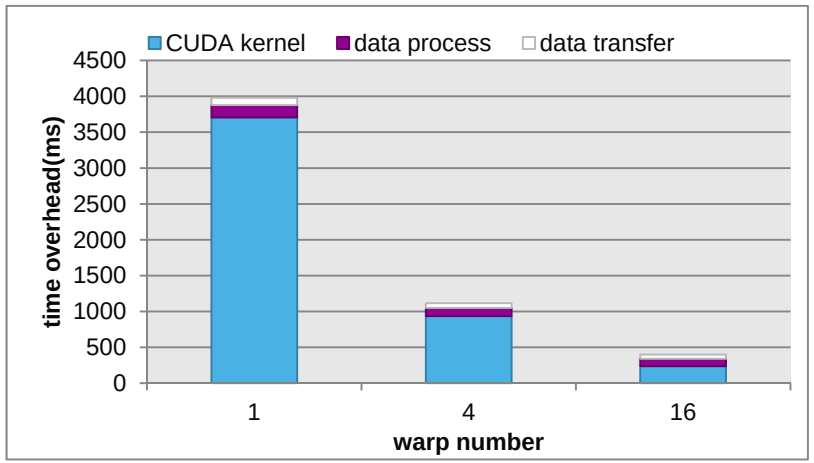
```
//marker kernel;
int P_marker();
//send start monitor request;
int P_startMonitor();
//send stop monitor request;
int P_stopMonitor();
//wait for data;
int P_receiveData(struct record** rec);
//store data into database;
int P_storeData(struct record* rec, int num);
//half-half benchmark;
int P_half_test(int time);
//computing intensive benchmark;
int P_compute_test(int time);
//memory access intensive benchmark;
int P_memory_test(int time);
//print error information;
int P_getError();
```

experiment

- Benchmark
  - Matrix multiplication with size 256*256
  - 10^5 samples per second



- Definition
  - Cuda kernel: the cuda kernel execution time
  - Data process: marker indentify process cost
  - Data transmission: data transfer from monitor machine to host



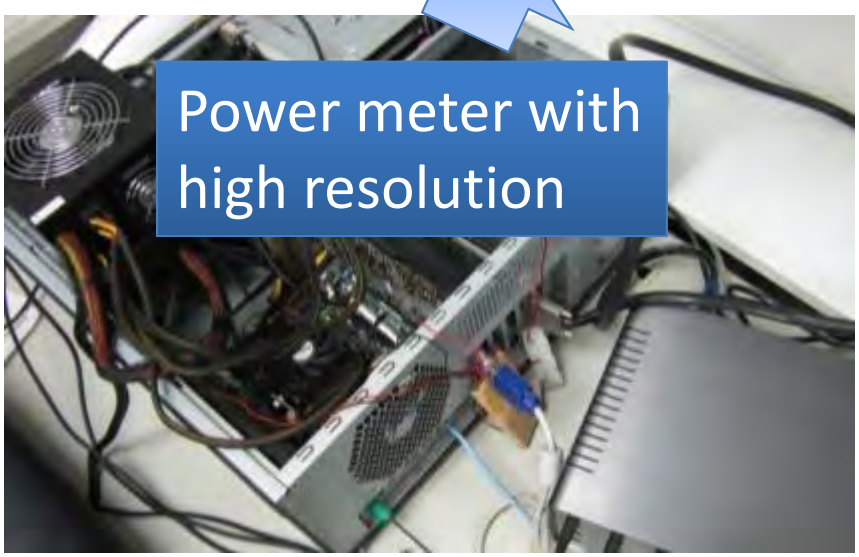
Statistical Power Modeling of GPUs [IEEE IGCC10]

GPU performance counters

$$p = \sum_{i=1}^{n} \alpha_i c_i + \beta$$

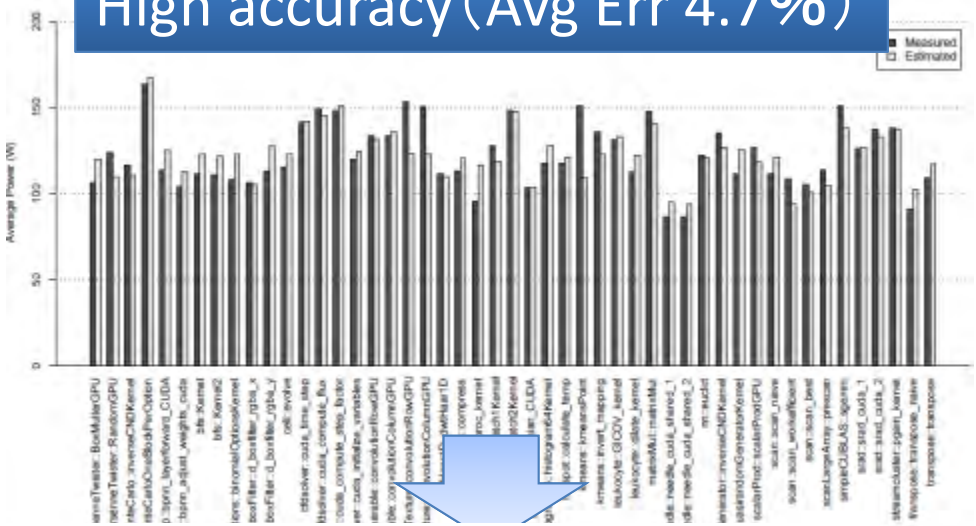
Average power consumption

- Prevents overtraining by ridge regression
- Determines optimal parameters by cross fitting



- Estimates GPU power consumption GPU statistically
- Linear regression model using performance counters as explanatory variables

High accuracy（Avg Err 4.7%）



Accurate even with DVFS

Future: Model-based power opt.

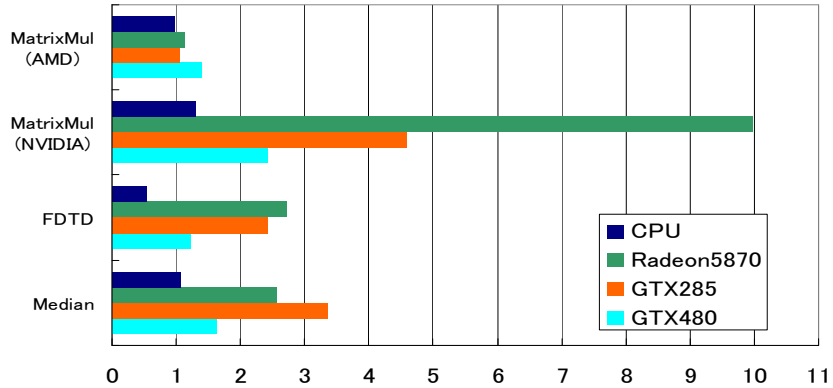
Linear model shows sufficient accur**acy**

Possibility of optimization of **Exascale** systems with O(10^8) processors

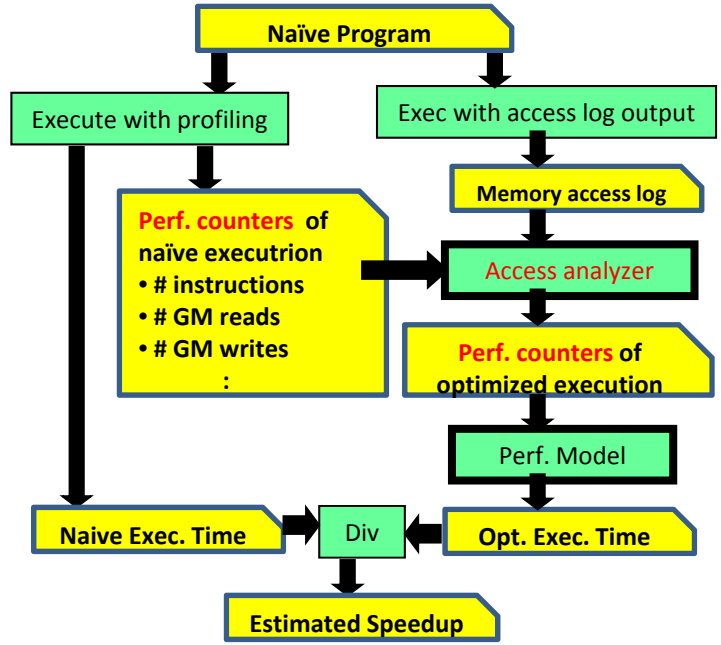
Optimization Effect Predictors for GPUs

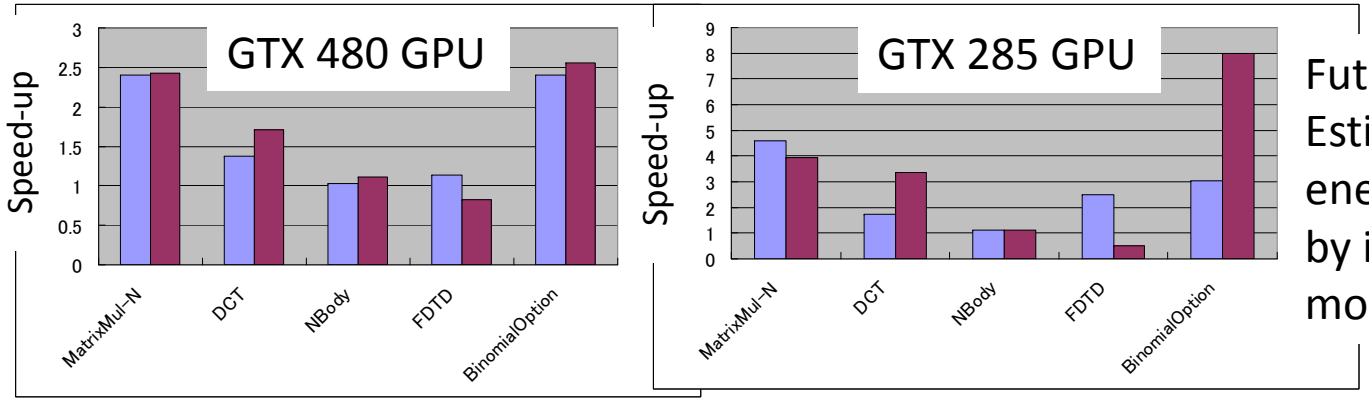
Optimization effects heavily depend on GPU architecture and generation

⇒ Estimate effects before making efforts



Speed-up by "shared memory" optimization





Comparing measured speed up and estimation

Future work:
Estimating improvement of energy consumption by integrating power model

Low Power Scheduling in GPU Clusters

[IEEE IPDPS-HPPAC 09]

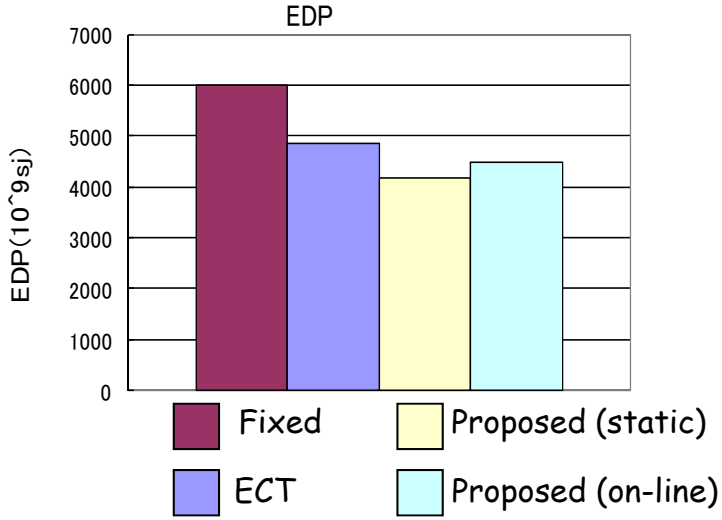
Objective: Optimize CPU/GPU Heterogeneous

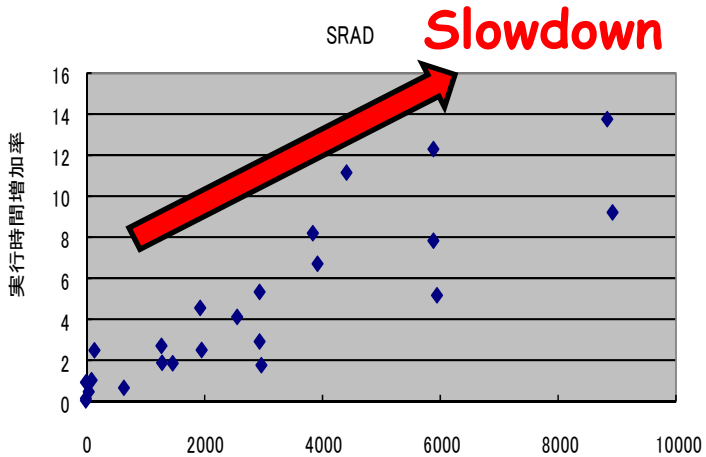
- Optimally schedule mixed sets of jobs executable on either CPU or GPU but w/different performance & power
- Assume GPU accel. factor (%) known

30% Improvement Energy-Delay Product

TODO: More realistic environment

- Different app. power profile
- PCI bus vs. memory conflict
  - GPU applications slow down by 10 % or more when co-scheduled with memory-intensive CPU app.





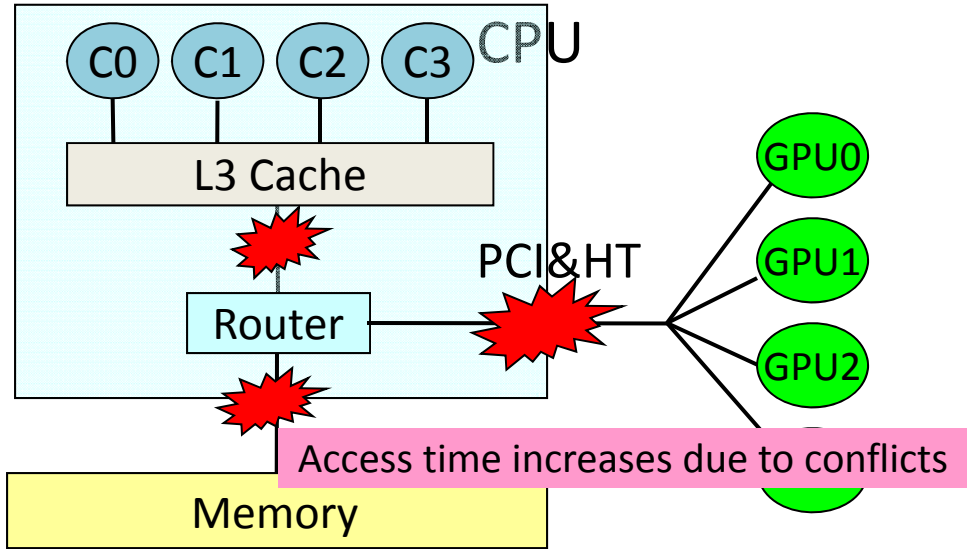
Background CPU process BW (MB/s)

Intelligent Task Scheduling for GPU clusters

- considering conflicts on memory bus and PCI-E bus -

T. Hamano, et al. IPSJ SIGHPC-124 (domestic)

4 processes share a node.



Status of each process is one of the following:

(1) Memory access

(2) PCI transfer (assume blocking cuMemcpy)

(3) No memory access

Speed-down ratio of a process = r (percentage of each status) × α (conflict coefficient)

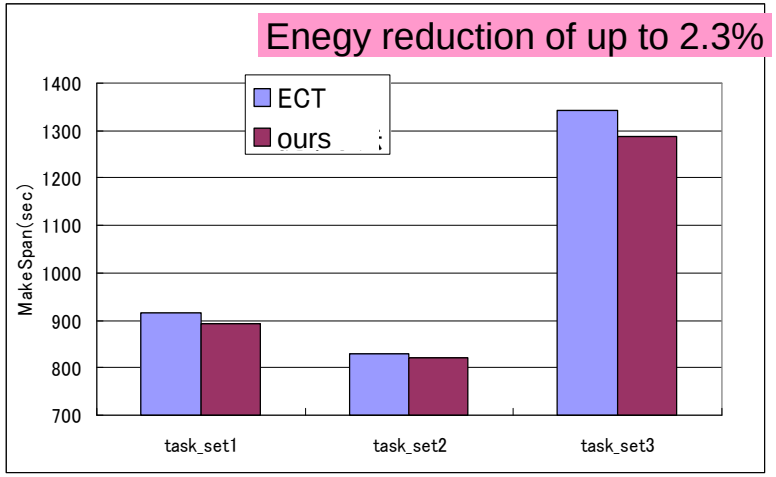
$$\sum_{s0}\sum_{s1}\sum_{s2}\sum_{s3}\alpha(s0,\{s1,s2,s3\})r_0(s0)r_1(s1)r_2(s2)r_3(s3)$$

We integrated this model into ECT(Earliest Complete Time) scheduling.

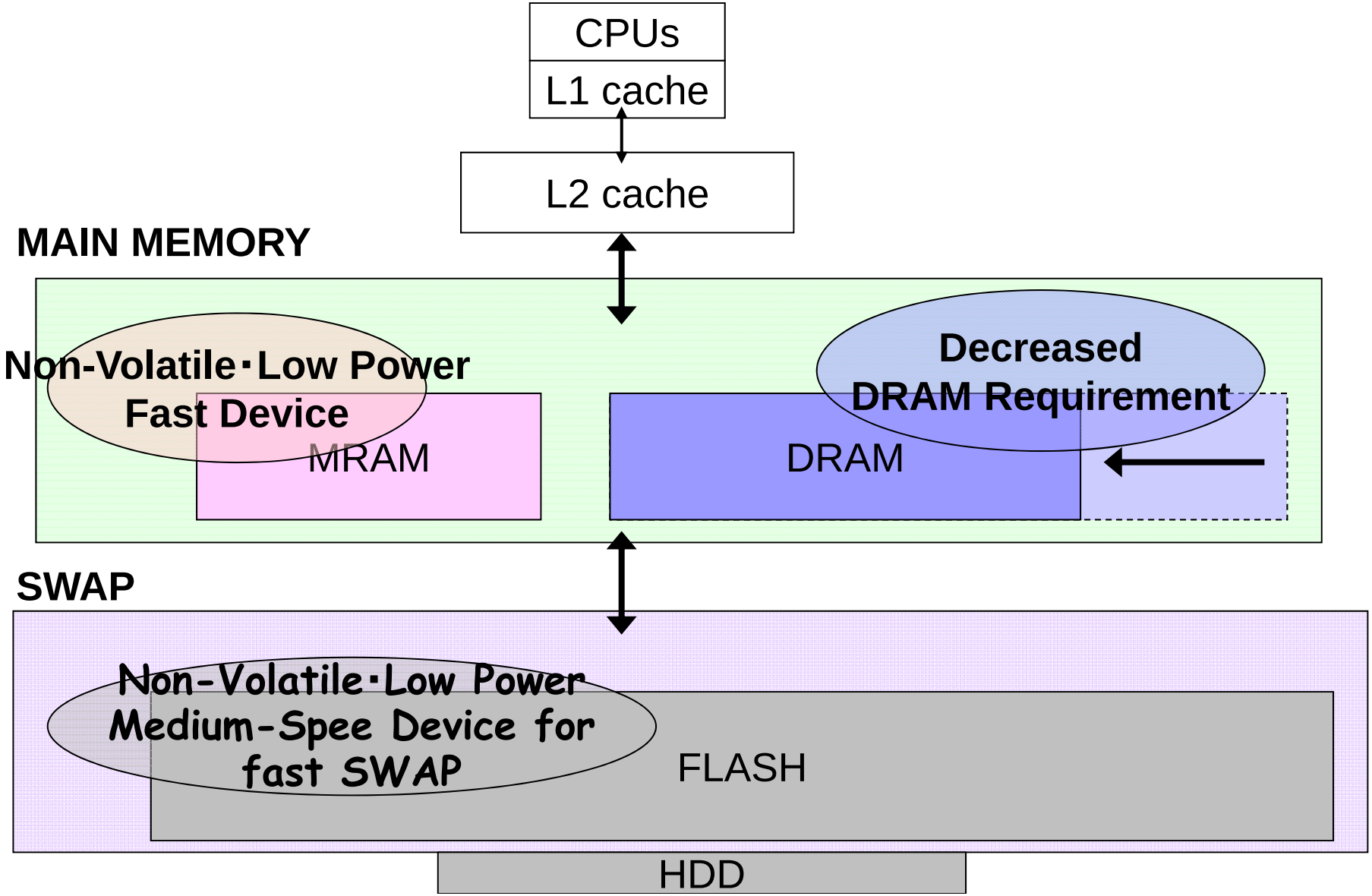
Assuming the following information of each job is known

- Execution time (at stand alone case)
- Total memory access (from performance counter)
- Total PCI-E transfer (from CUDA profiler)

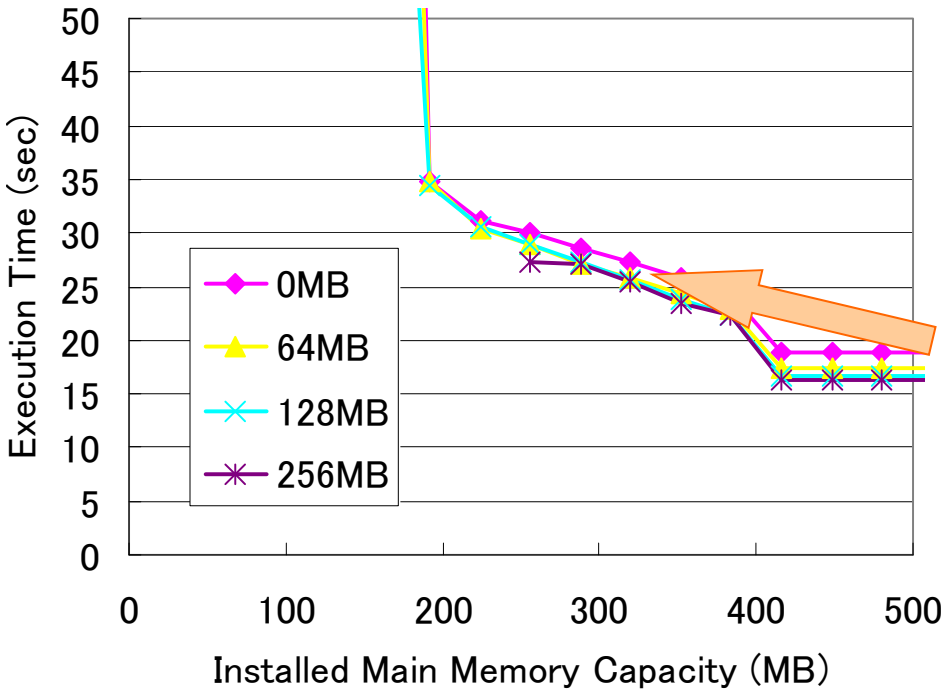
Actual execution time is estimated by our model.



ULP-HPC Research Example: Next-Gen ULP Memory System ([IEEE IPDPS-HPPAC 08])



Performance vs. Power Consumption

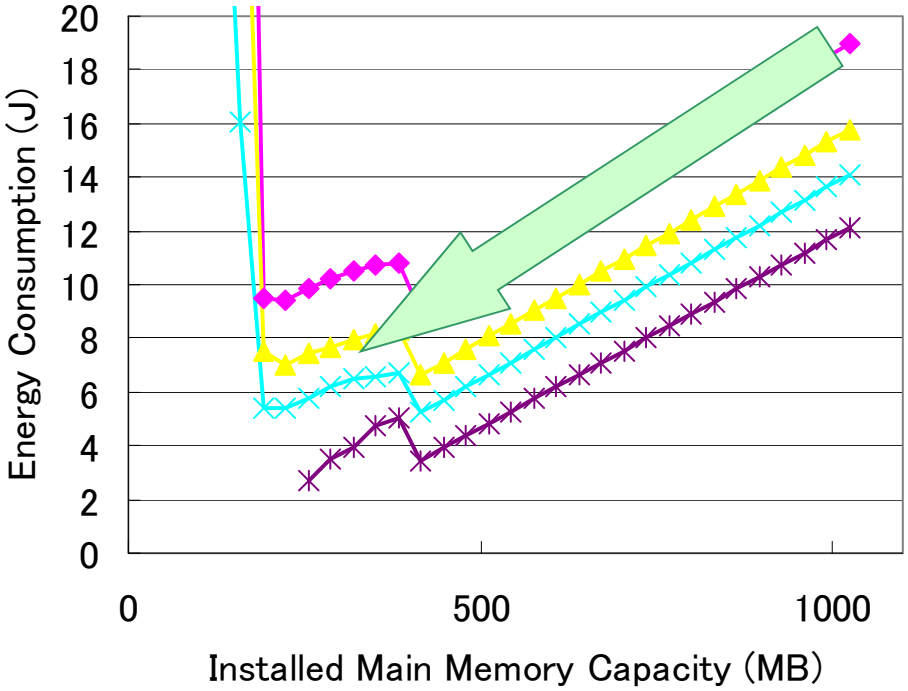


Application : NAS CG (class B)

Standard system: DRAM 1GB

Proposed system:

(MRAM 128MB + DRAM 192MB)



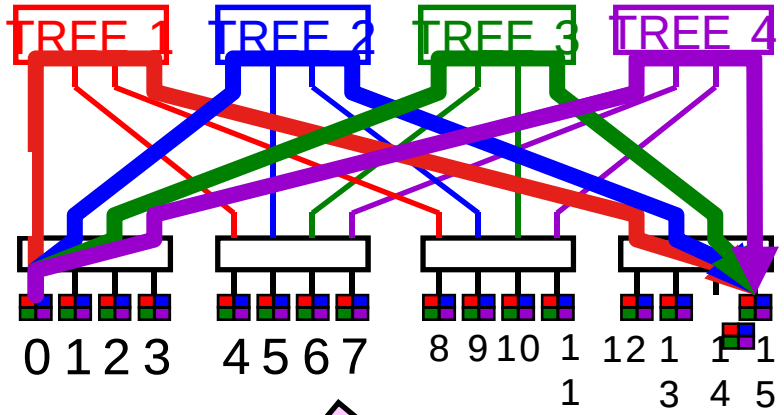
Execution overhead : x 1.6
Memory Energy Consumption : reduced to 28%

Ultra-Low-Power HPC Interconnects

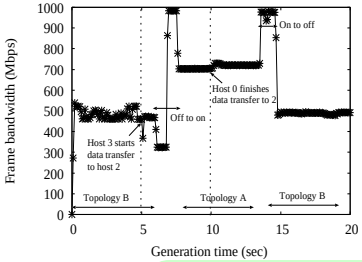
- **Objective: Power efficiency of interconnects is improved by 100 times within 10 years**
- Using a large number of small switches
  - No more expensive monster-switches (1,000-port switch)

Switch

Tree 1-4: all On



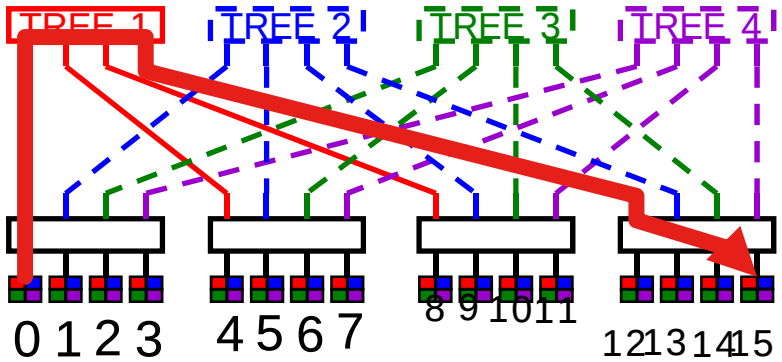
Proc



Traffic decreases
(Links becomes off)

Tree 1: On

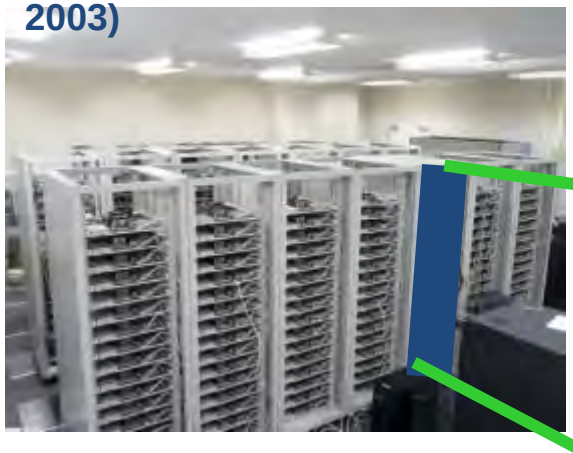
Tree2-4: Off



Reference
Michihiro Koibuchi, Tomohiro Otsuka, Tomohiro Kudoh, Hideharu Amano, ``A Switch-tagged Routing Methodology for PC Clusters with VLAN Ethernet", IEEE Transactions on Parallel and Distributed Systems, Vol.22, No.2, pp.217-230, Feb 2011

Research Outcome

SuperNova cluster （225 hosts, 93rd supercomputer on Nov, 2003)

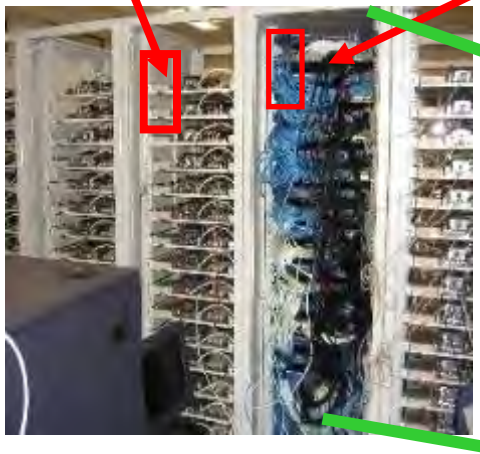


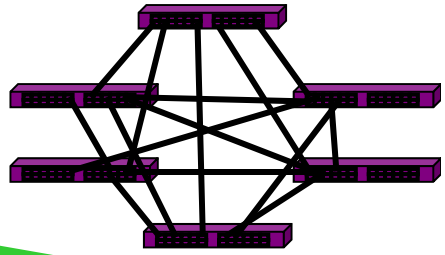
Before

Monster switch
（312 port@1）

PC

Switch





After (ULP-HPC)

Small switch
（48port@8)

Power：using small switches（-87%）×cyclic topology(+60%）
×On/Off link regulation（-37%）≒-87%(1/8)

Cost：using small switches（-98%) ×cyclic topology(+60%）
×On/Off link regulation(±0%) ≒ -97%(1/30)

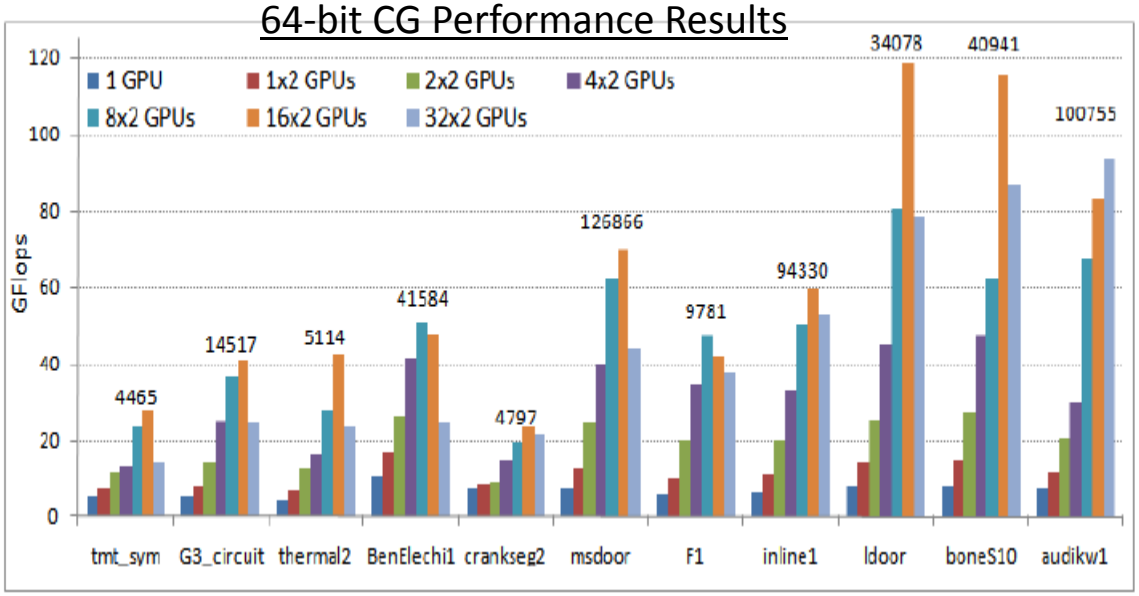
NII Group (Michihiro Koibuchi (koibuchi@nii.ac.jp), Masato Yoshimi （myosimi@mail.doshisha.ac.jp）)

"Auto-Tuning" Sparse Iterative Solvers on a Multi-GPU Clusters

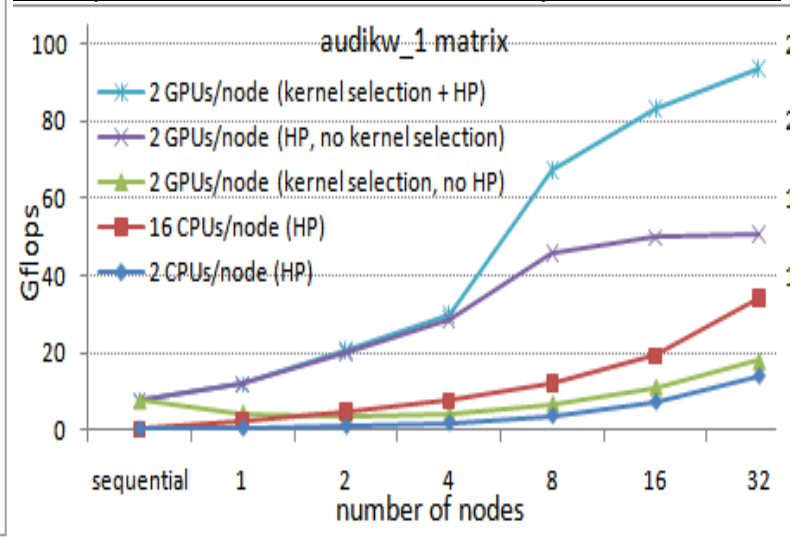
(A. Cevahir, A. Nukada, S. Matsuoka)

- High communication due to irregular sparsity
  - GPU computing is fast, communication bottleneck is more severe
- We propose techniques to achieve scalable parallelization
  - A new sparse matrix-vector multiplication (MxV) kernel [*ICCS'09*]
  - **Auto-selection of optimal running MxV kernel** [*ISC'10*]
  - Minimization of communication between GPUs [*ISC'10*]
- Two methods: **CG and PageRank (on 100CPUs)**

64-bit CG Performance Results



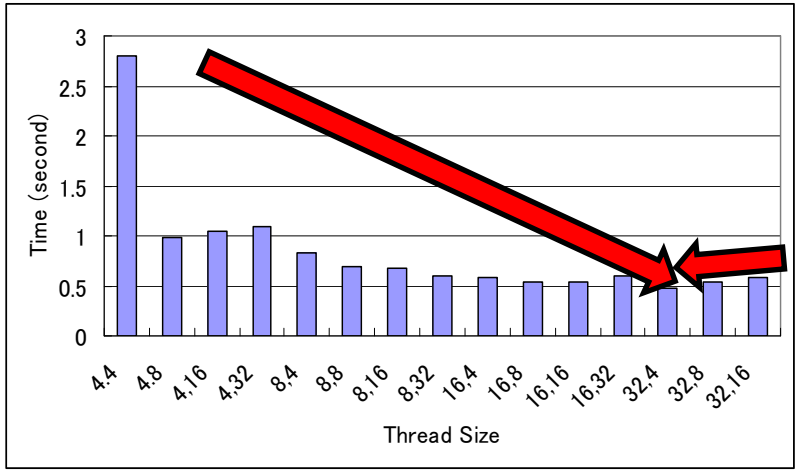
Comparisons of different CG implementations



Towards Auto-Tuning of Stencil Computation on GPU-accelerated Clusters

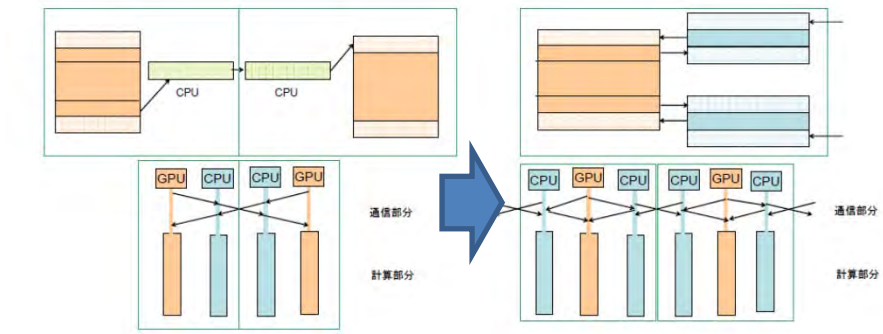
- Programming on GPU clusters gets harder because of
  - Hybrid programming with CUDA/MPI/threads
  - # of tuning parameters is increasing
  - Hiding latency

→ Auto-tuning is more important



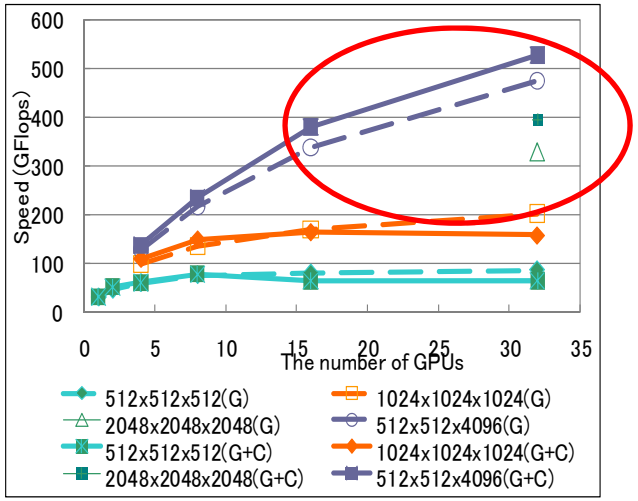
We find the optimal parameters, such as block size, by auto-tuning

→ x60 faster than 1 CPU core



- 3-HOP communication has been required, for GPU<->CPU<->CPU<->GPU

→ 1-HOP method is proposed, and auto-tune between communication methods

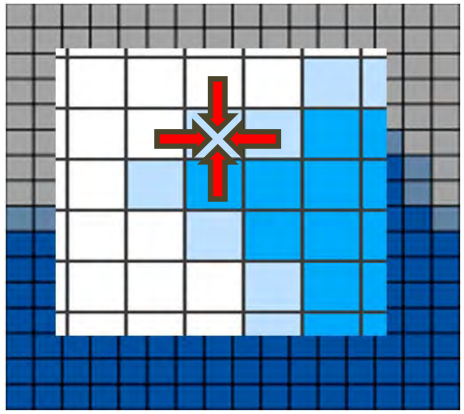


Up to 20% speed-up with 1-HOP method

526GFlops is achieved with 32GPUs on TSUBAME 1.2

→ x1170 faster than 1 CPU core

Scalability of Parallel Stencil Computations with Automatic GPU Code Generation and Auto-Tuning



Parallel stencil codes get complex because of

- Domain decomposition
- MPI+CUDA comm.
- Overlapping comm.

Physis: Code generation framework for multi-GPUs

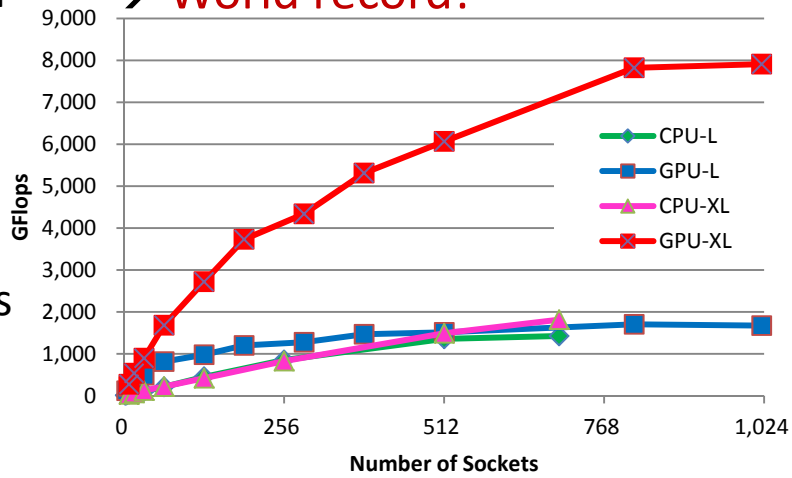
User writes computation function of a single point

```
__stencil__ void average(int x, int y, int z, grid3d_real g)
{
    if (x == 0 || y == 0 || z == 0) return;
    float ret = psGridGet(g,0,0,0)
       + psGridGet(g,-1,0,0) + psGridGet(g,1,0,0)
       + psGridGet(g,0,-1,0) + psGridGet(g,0,1,0)
       + psGridGet(g,0,0,-1) + psGridGet(g,0,0,1);
    psGridEmit(g, ret / 7.0);
}
```

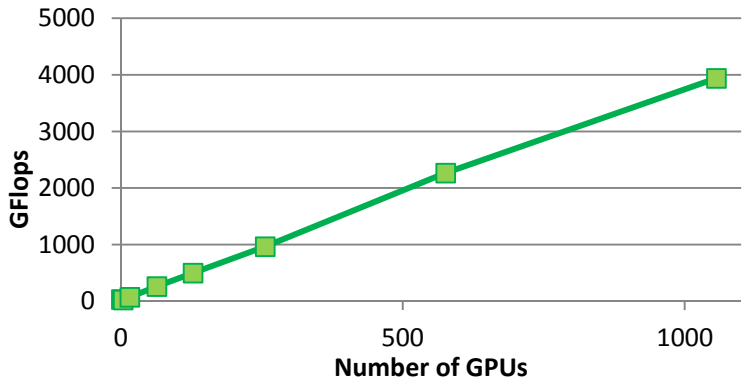
Parallel stencil code for GPU clusters is generated

8TF is achieved in Himeno bench with 1024GPUs on TSUBAME2.0

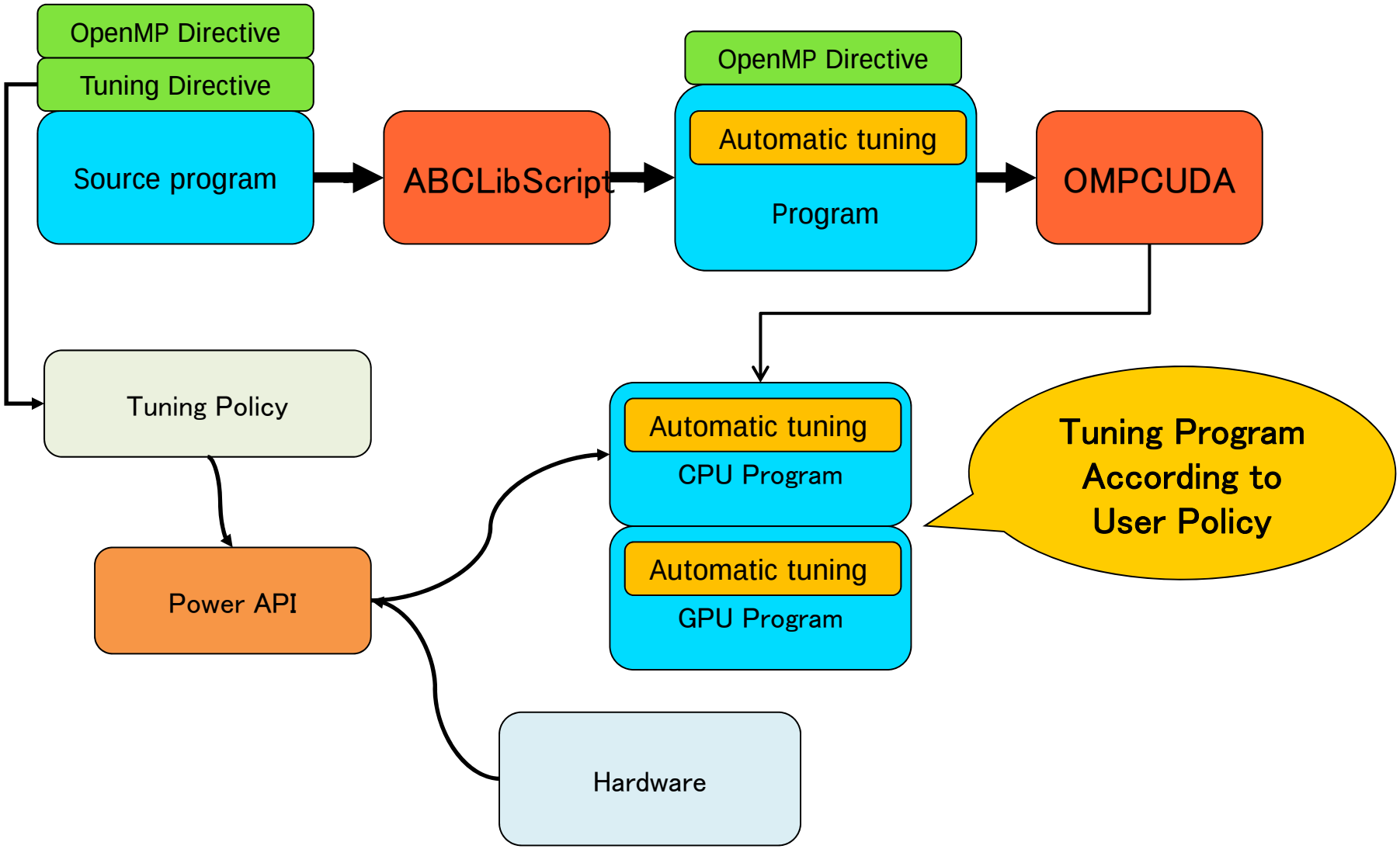
→ World record!



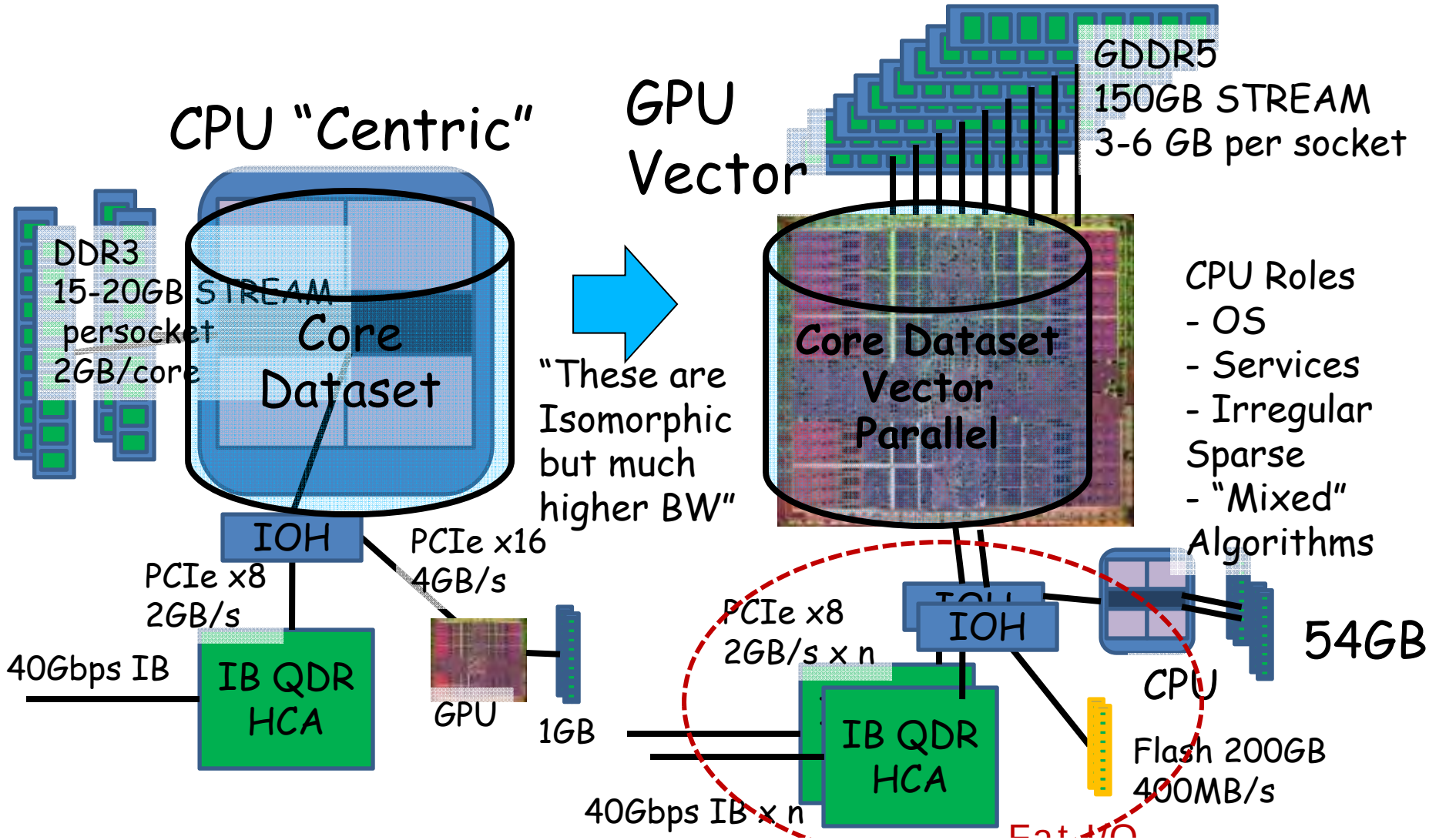
Weak scalability of seismic simulation



Energy and Performance Tuning Framework for GPU Accelerator Environment

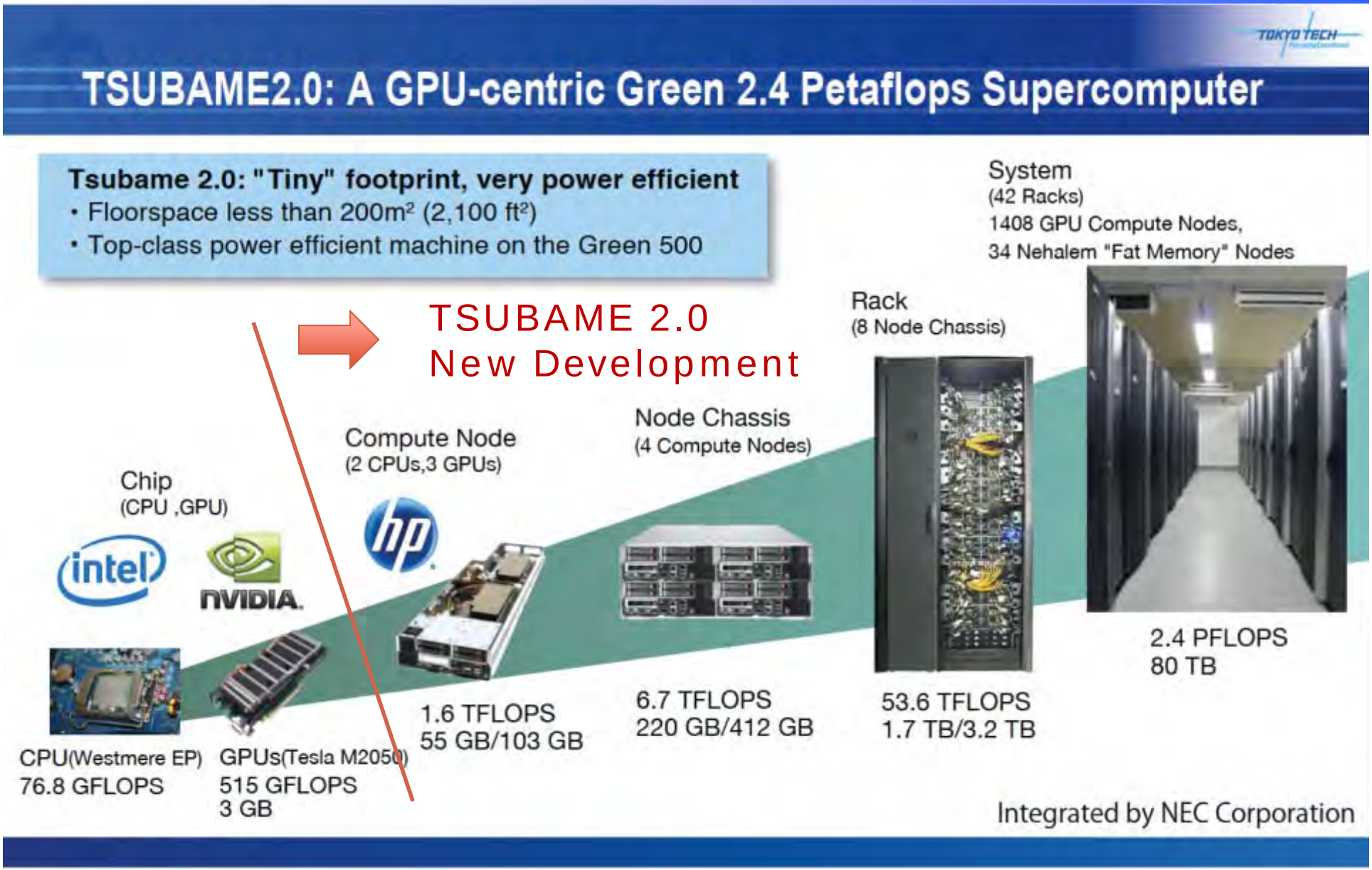


From TSUBAME 1.2 to 2.0: From CPU Centric to GPU Centric Nodes for Scaling High Bandwidth in Network & I/O!!



TSUBAME2.0 Nov. 1, 2011

"The Greenest Production Supercomputer in the World"

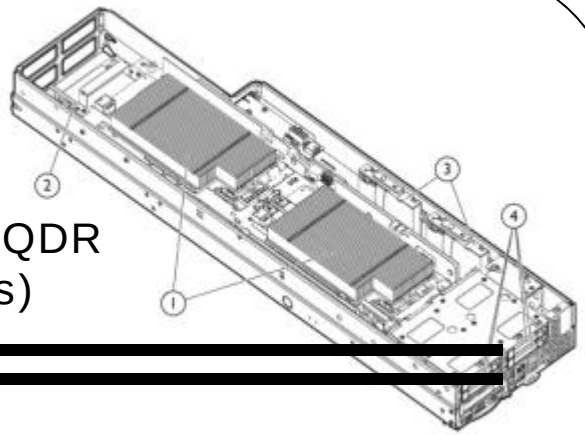


2.4 Petaflops, 1408 nodes
~50 compute racks + 6 switch racks
Two Rooms, Total 160m²
1.4MW (Max, Linpack), 0.48MW (Idle)

GPU M2050

TSUBAME2.0 Compute Nodes

Thin Node



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)
Memory: 54, 96 GB DDR3-1333
SSD:60GBx2, 120GBx2

1408nodes:

4224GPUs:59,136 SIMD Vector Cores, 2175.36TFlops (Double FP)

2816CPUs, 16,896 Scalar Cores: 215.99TFlops

Total: 2391.35TFLOPS

Memory: 80.6TB (CPU) + 12.7TB (GPU)

SSD: 173.9TB

IB QDR

PCI-e Gen2x16 x2
NVIDIA Tesla S1070 GPU

HP 4 Socket Server
CPU: Intel Nehalem-EX 2.0GHz x4 (32cores/node)
Memory: 128, 256, 512GB DDR3-1066
SSD:120GB x4 (480GB/node)

34 nodes:
8.7TFlops

Memory:
6.0TB+GPU

SSD: 16TB+

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

Details of Compute Node

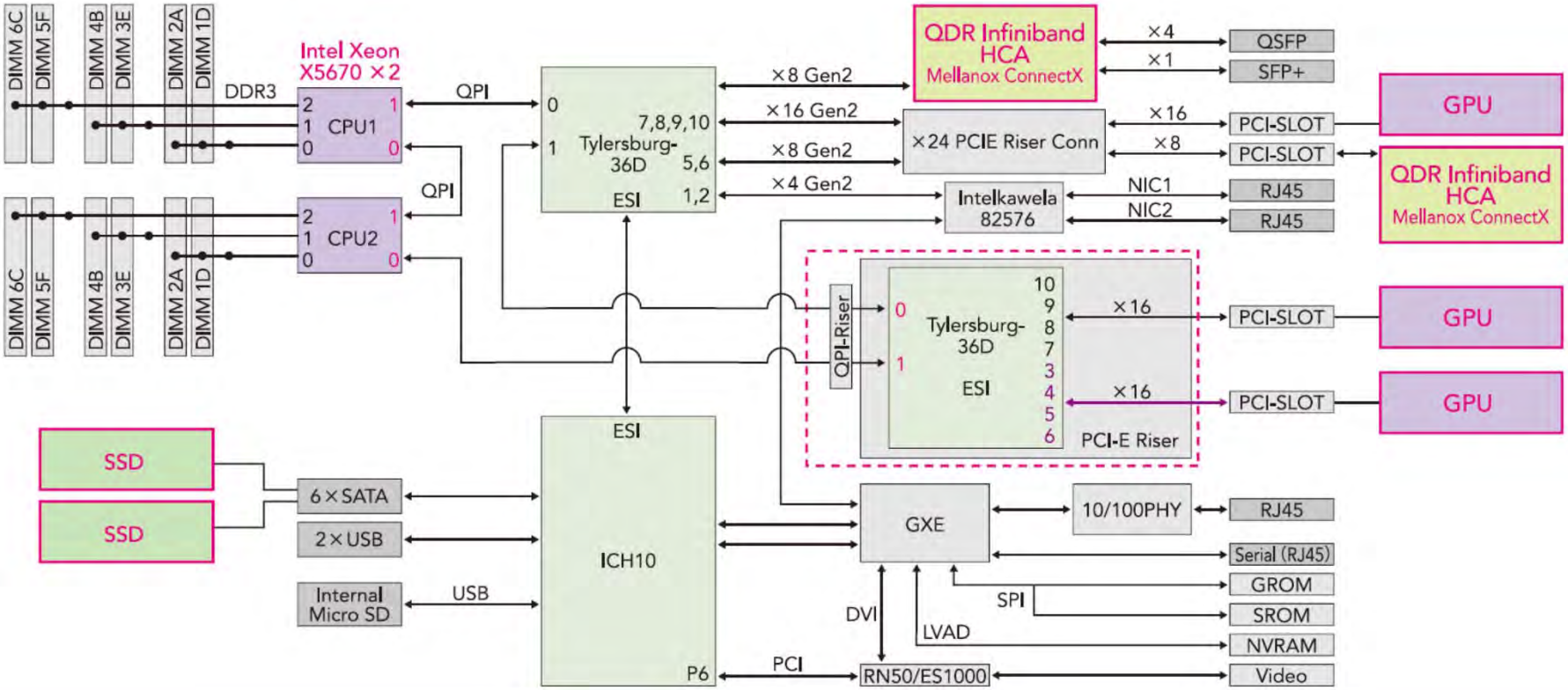
HP ProLiant SL390s

GPU : NVIDIA Tesla M2050 (Fermi Core) ×3 515GFLOPS VRAM 3GB/GPU

CPU : Intel Xeon X5670 2.93Ghz ×2
6 core/socket 76.7 GFLOPS (12cores/node) ※ Turbo boost: 3.196GHz

Memory : 58GB DDR3 1333MHz 一部 103GB

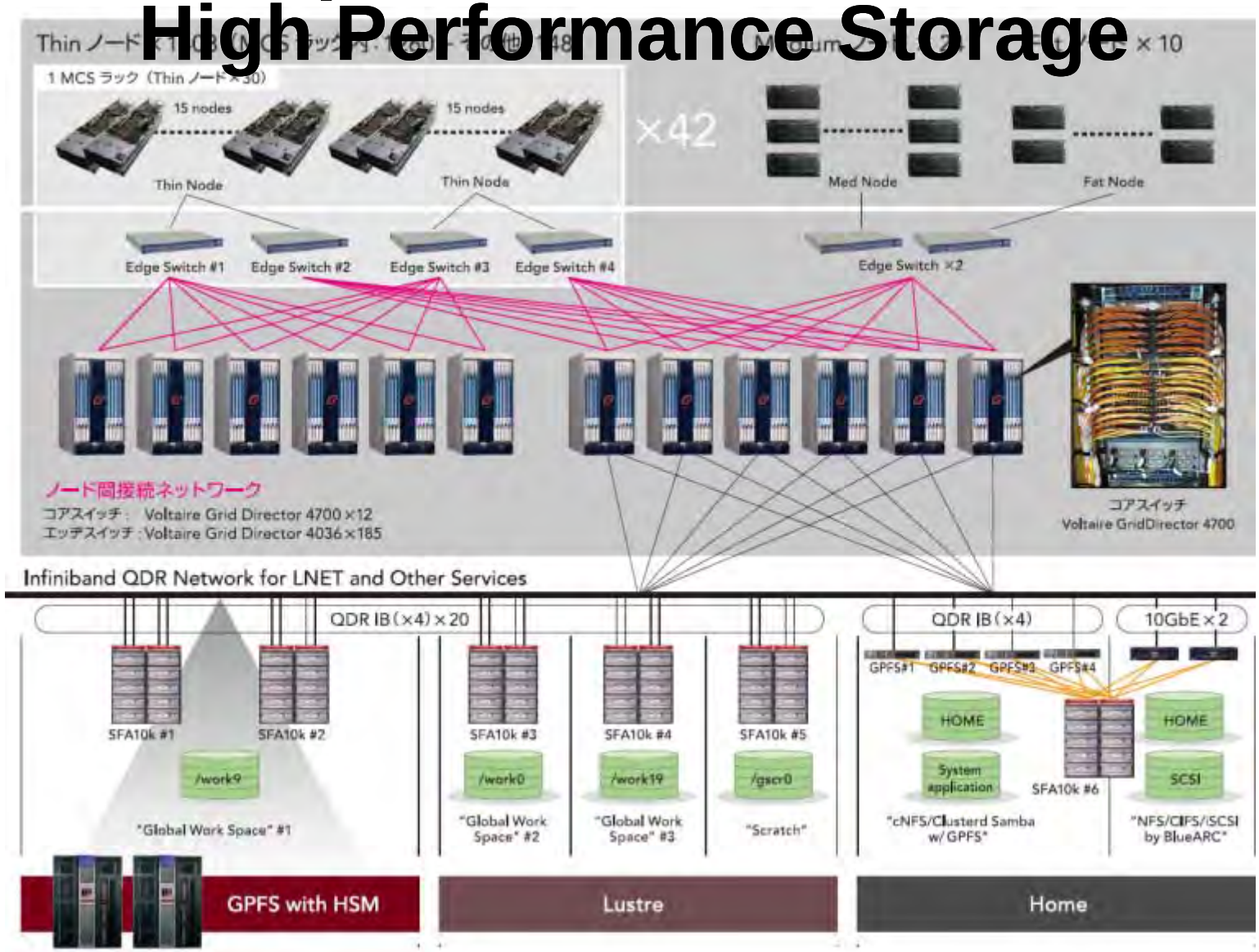
SSD : 60GB ×2 (120GB/node) 一部 120GB ×2 (240GB/node)



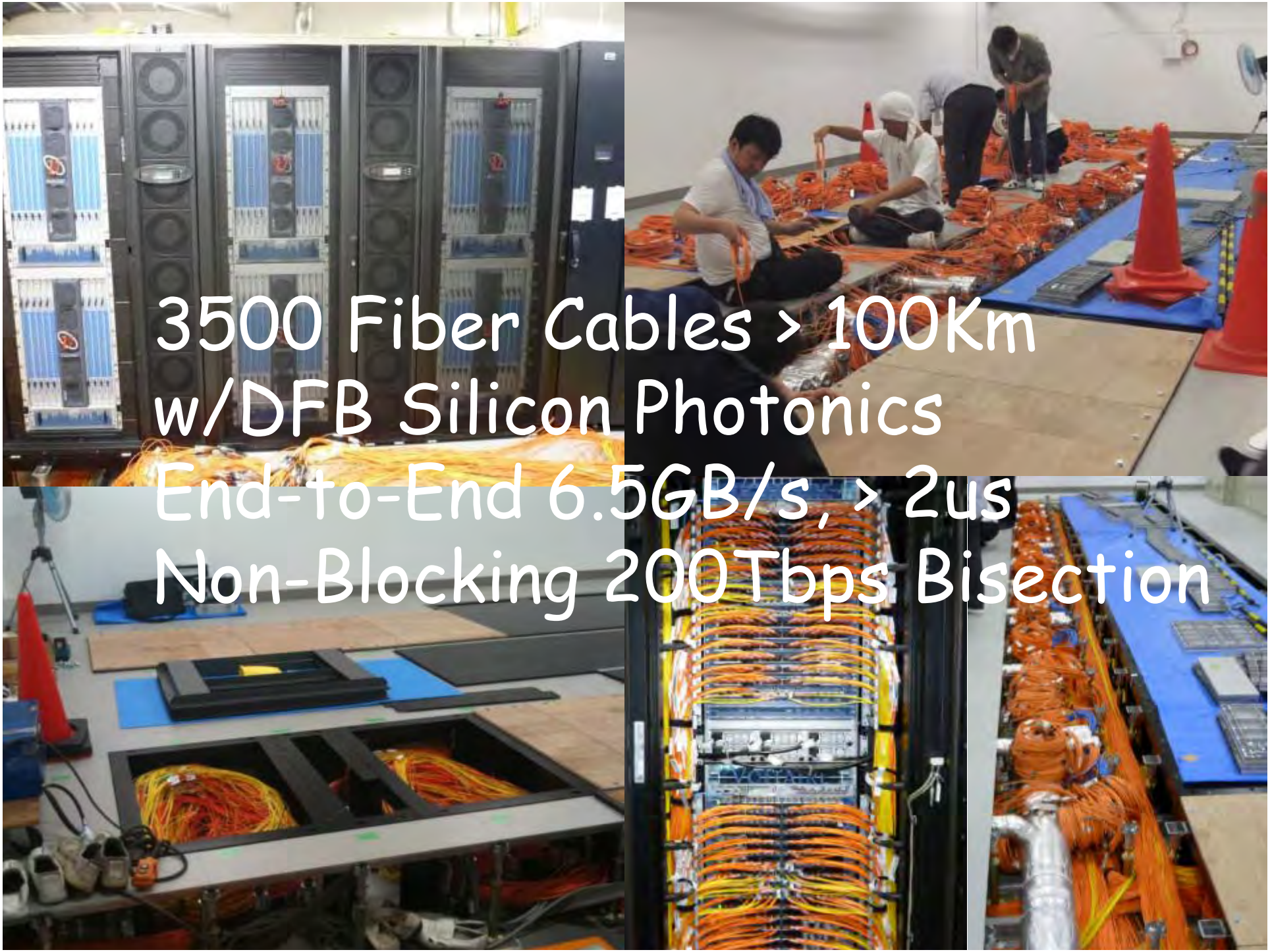
Copyright © Takayuki Aoki / Global Scientific Information and Computing Center, Tokyo Institute of Technology

Full Bisection Multi-Rail Optical Network and High Performance Storage

GP GPU

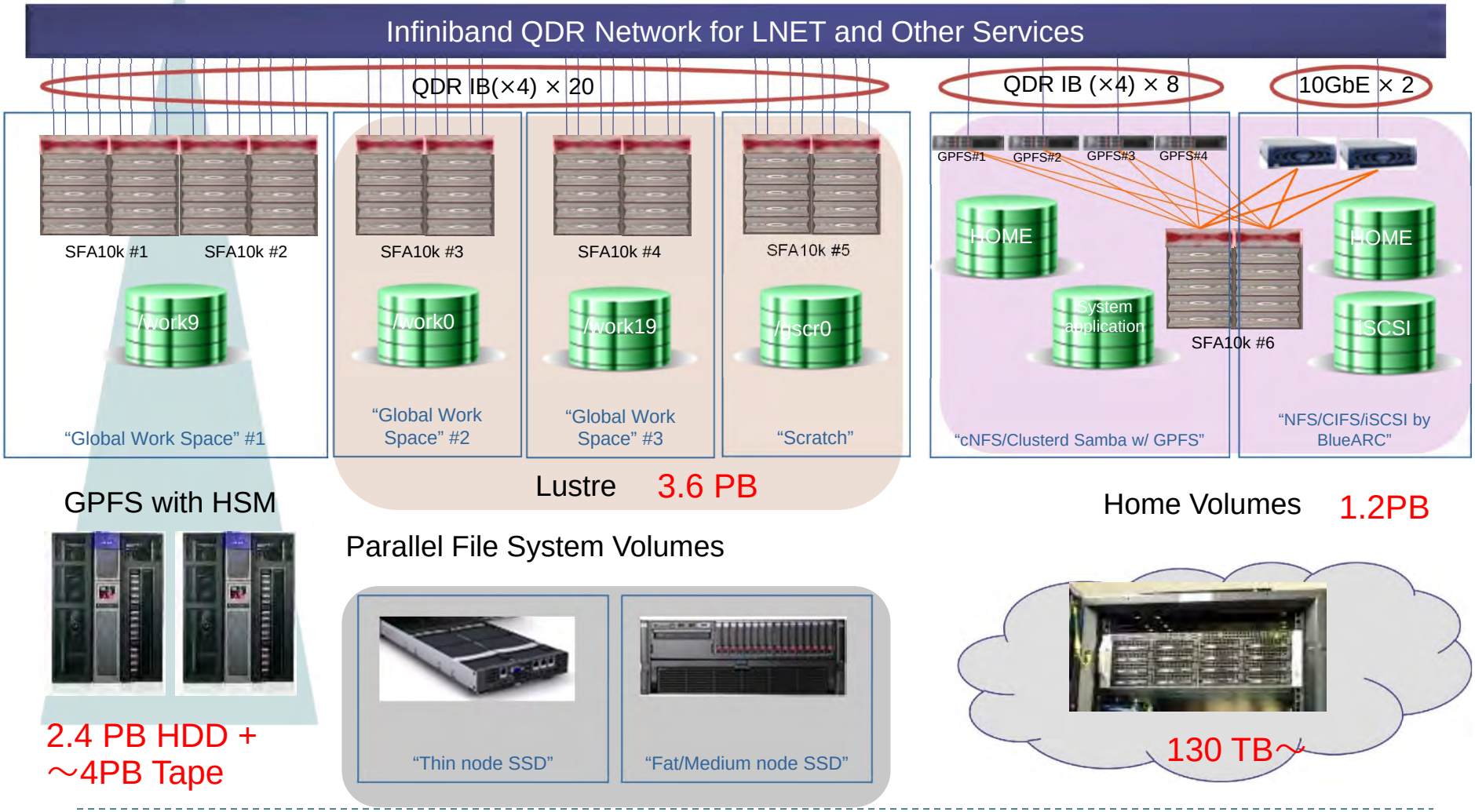


Copyright © Takayuki Aoki / Global Scientific Information and Computing Center, Tokyo Institute of Technology


3500 Fiber Cables > 100Km
w/DFB Silicon Photonics
End-to-End 6.5GB/s, > 2us
Non-Blocking 200Tbps Bisection

TSUBAME2.0 Storage Overview

TSUBAME2.0 Storage 11PB (7PB HDD, 4PB Tape)



Tsubame2.0 Efficient Cooling Infrastructure

HP's water-cooled rack

Completely closed racks with their own heat exchanger.

1.5 x width of normal rack+rear ext.

Cooling for high density deployments

35kW of cooling capacity single rack

- **Highest Rack Heat Density ever**
- **3000CFM Intake airflow with 7C chiller water**

up to 2000 lbs of IT equipment

Uniform air flow across the front of the servers

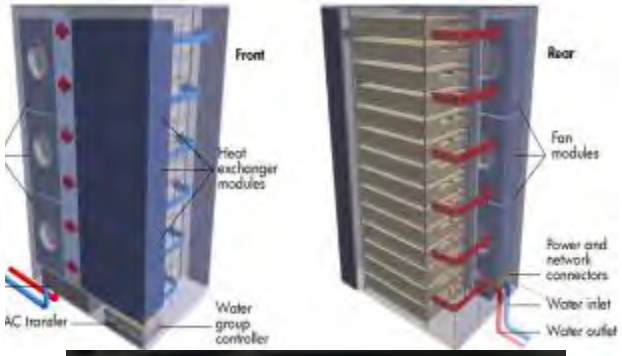
Automatic door opening mechanism controlling both racks

Adjustable temperature set point

Removes 95% to 97% of heat inside racks

Polycarbonate front door reduces ambient noise considerably

~= Entire Earth Simulator (rack = 50TF)

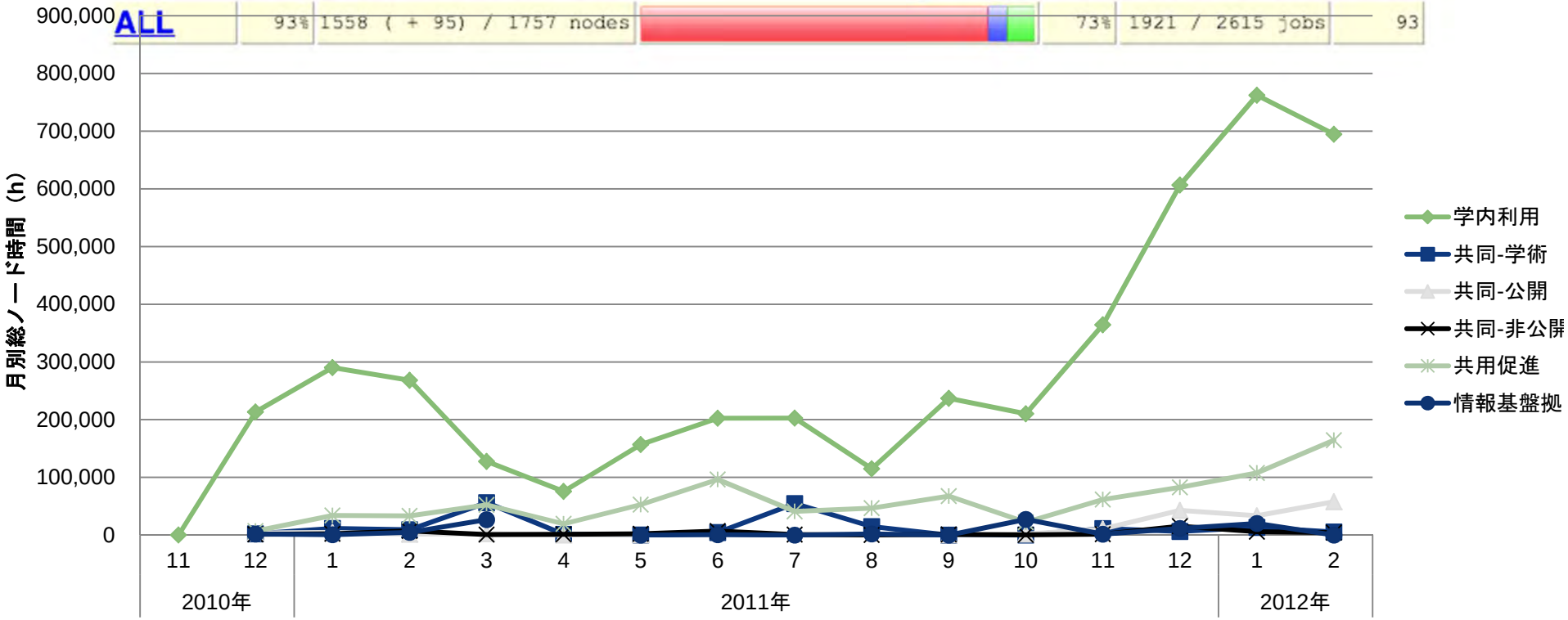


~7000 campus users
~2000 SC Users
~50 Industry User Groups
>90% System Utilization
50~60% GPU Utilization across ALL jobs

TSUBAME2.0における利用者種別による使用時間の集計（2012/02末時点）

Service List

service	assigned nodes		running jobs		users
S	100%	352 / 352 nodes	38%	175 / 452 jobs	46
S96	100%	41 / 41 nodes	53%	41 / 76 jobs	4
G	99%	475 / 477 nodes	100%	62 / 62 jobs	12
V	83%	364 / 432 nodes	80%	1531 / 1904 jobs	34
L128	100%	10 / 10 nodes	66%	10 / 15 jobs	1
L128F	100%	10 / 10 nodes	71%	10 / 14 jobs	1
L256	37%	3 / 8 nodes	100%	3 / 3 jobs	1
L512	100%	2 / 2 nodes	100%	2 / 2 jobs	1
H/X	95%	301 (+ 95) / 420 nodes	100%	87 / 87 jobs	8
ALL	93%	1558 (+ 95) / 1757 nodes	73%	1921 / 2615 jobs	93



TSUBAME2.0 World Rankings (Nov. 2010 Announcement Green500!!!)

The Top 500 (Absolute Performance)

#1 : ~2.5 PetaFlops: China Defense Univ. Dawning Tianhe 1-A
#2 : 1.76 Petaflops: US ORNL Cray XT5 Jaguar
#3 : 1.27 PetaFlops: China Shenzen SC Nebulae
#4 : 1.19 PetaFlops: Japan Tokyo Tech. HP/NEC TSUBAME2.0
#5 : 1.054 PetaFlops: US LLBL Cray XE6 Hopper
#~33 (#2 Japan) : 0.191 Petaflops : JAEA Fujitsu

The Green 500 (Performance/Power Efficiency)

#1: 1684.20 : US IBM Research BG/Q Prototype (116)
#2: 958.35: Japan Tokyo Tech/HP/NEC Tsubame 2.0 (4)
#3: 933.06 : US NCSA Hybrid Cluster Prototype (403)
#4: 828.67: Japan Riken "K" Supercomputer Prototype (170)
#5-7: 773.38: Germany Julich etc.IBM QPACE SFB TR (207-209)
(#2+ 1448.03: Japan NAO Grape-DR Prototype) (383) (Added in Dec.)

TSUBAME2.0 "Greenest Production Supercomputer in the World"
Nov., 2010, June 2011 (two in a row!)

THE GREEN 500™

sponsored by

SUPERMICRO

This certificate is in recognition of your organization's achievements in reducing the environmental impact of high-performance computing.

GSIC Center, Tokyo Institute of Technology

Is recognized as the

Greenest Production Supercomputer in the World

on the world's Green500 List of computer systems as of

November 2010

Wu-chun Feng, Co-Chair

Kirk Cameron, Co-Chair

Petaflops? Gigaflops/W?

x66,000 faster

x3 power efficient

<<

x44,000 Data

Laptop: SONY Vaio type Z (VPCZ1)
CPU: Intel Core i7 620M (2.66GHz)
MEMORY: DDR3-1066 4GBx2
OS: Microsoft Windows 7 Ultimate 64bit
HPL: Intel(R) Optimized LINPACK Benchmark for Windows (10.2.6.015)
256GB HDD

18.1 GigaFlops Linpack

369 MegaFlops/W

Supercomputer: TSUBAME 2.0
CPU: 2714 Intel Westmere 2.93 Ghz
GPU: 4071 nVidia Fermi M2050
MEMORY: DDR3-1333 80TB + GDDR5 12TB
OS: SuSE Linux 11 + Windows HPC Server R2
HPL: Tokyo Tech Heterogeneous HPL
11PB Hierarchical Storage

1.192 PetaFlops Linpack

1043 MegaFlops/W

27th List: The TOP10

Rank						ax ps]	Power [MW]
1						62	9.90
2						66	4.04
3						59	6.95
4						71	2.58
5						92	1.40
6	D					10	3.98
7	N					88	4.10
8						54	2.91
9	l'Energie Atomique (CEA)	Bull	Bull bullx super-node S6010/S6030	France	8	1.050	4.59

HPC Wire Article "TSUBAME 2.0 is the Clear Overall Winner"

HPC wire

Since 1986 - Covering the Fastest Computers in the World and the People Who Run Them

言語を選択 Translation Disclaimer Subscribe | Sign In

Home | News | Features | Blogs | HPC Markets | Whitepapers | Multimedia

Visit additional Tabor Communication Publications

datanami

Digital Manufacturing report

HPC In the Cloud

Today's Top Feature

Feb 07, 2012

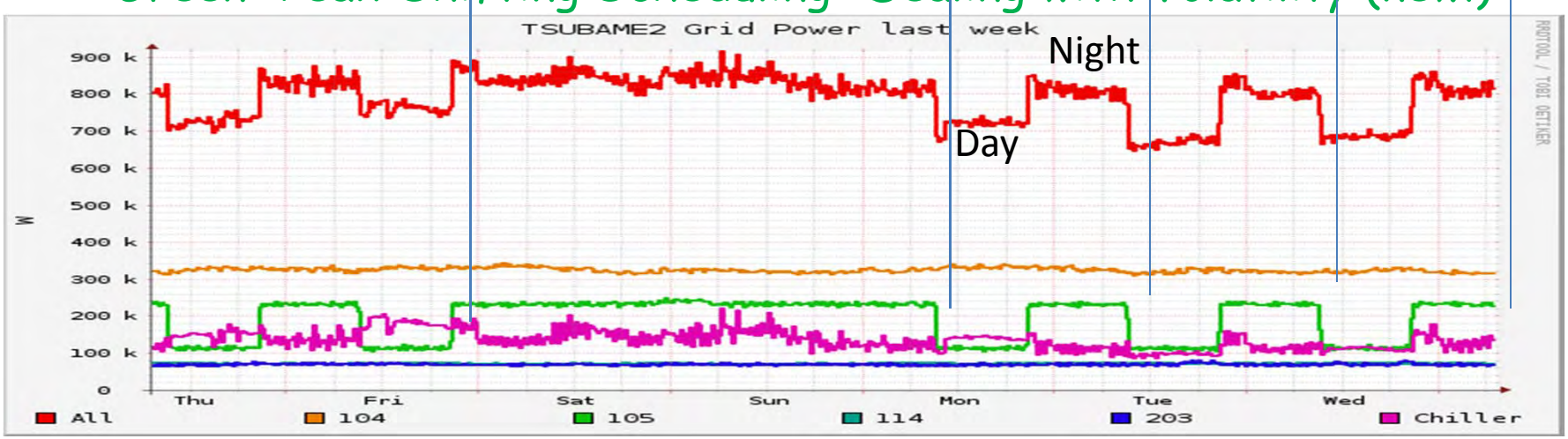
TOP500 SUPERCOMPUTER SITES | GRAPH 500 | THE GREEN 500

Number Crunching, Data Crunching and Energy Efficiency: the HPC Hat Trick

In the world of high performance computing, there are three distinct metrics in play: number crunching speed, data crunching speed, and energy efficiency. Can a computer excel at all three, or is our best recourse to try for something less than a hat trick? Read more...

Green Technologies in Production in Tsubame2.0---ULPHPC Feedback

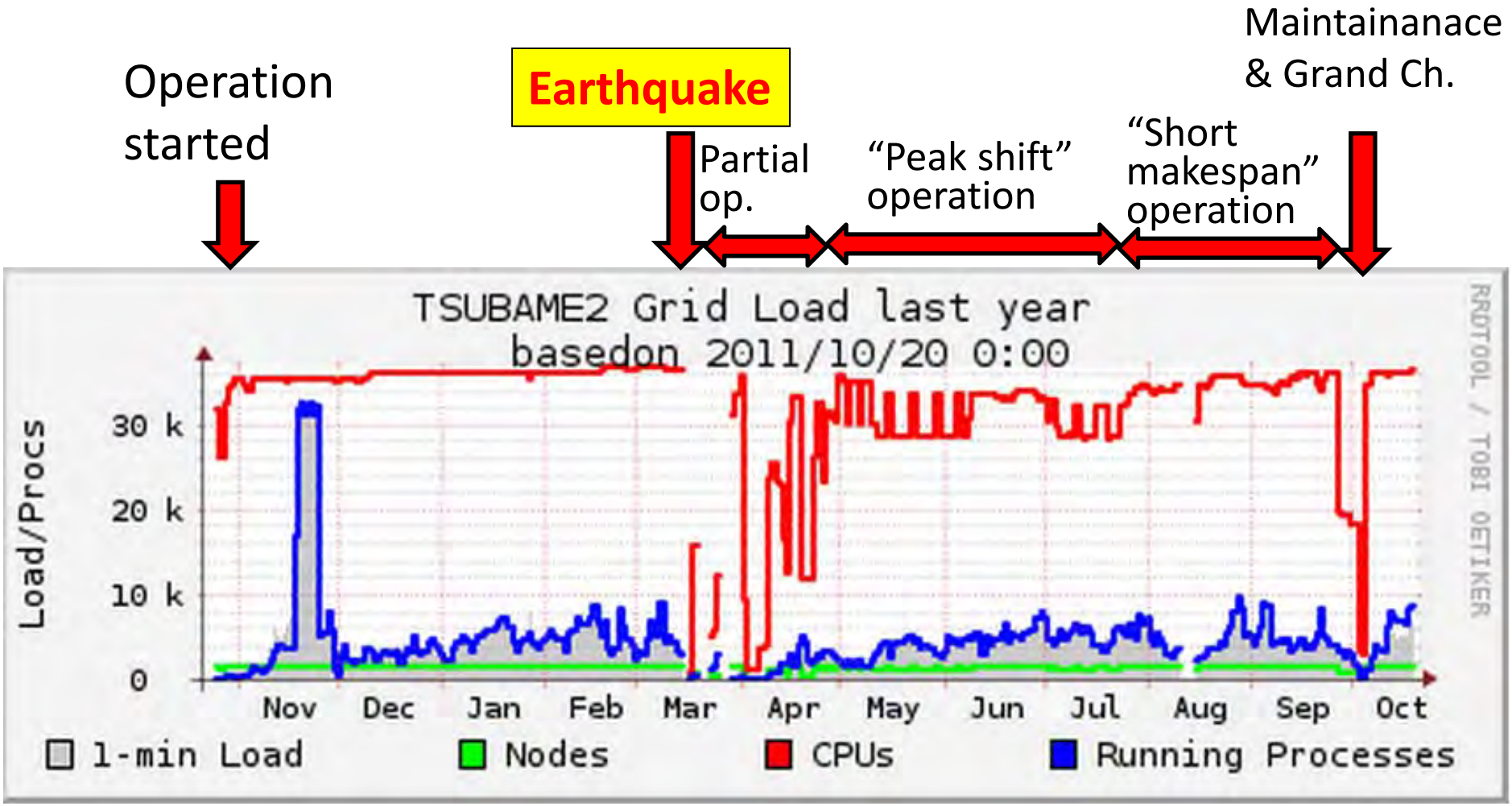
- Extensive Many-Core(GPU)–Multi-Core(CPU) Architecture
- Power efficient devices: 3000 SSDs (I/O), Integrated Fiber Optics, 92% efficient PSU, Fans and other LP devices
- Efficient hybrid Air-Liquid Cooling (HP-MCS Racks, PUE < 1.3)
- Extensive Thermal and Power Sensing (30,000 thermal sensors, power sensors / node & rack, etc.)
- Proactive Node/Rack Power Capping (used constantly)
- Power efficient software stack (DVFS etc.) and GPU-apps
- "Green" Peak-Shifting Scheduling– Dealing with Volatility (new!)



TSUBAME2.0 Power&PUE

	Power consumption by IT facility	Power consumption by Cooling facility	Total power consumption	PUE
Typical operation	750kW	230kW	980kW	1.31
Under high utilization (GPU DGEMM)	1610kW	410kW	2020kW	1.25

The First Year of TSUBAME2.0



Operation History towards Shortage of Power

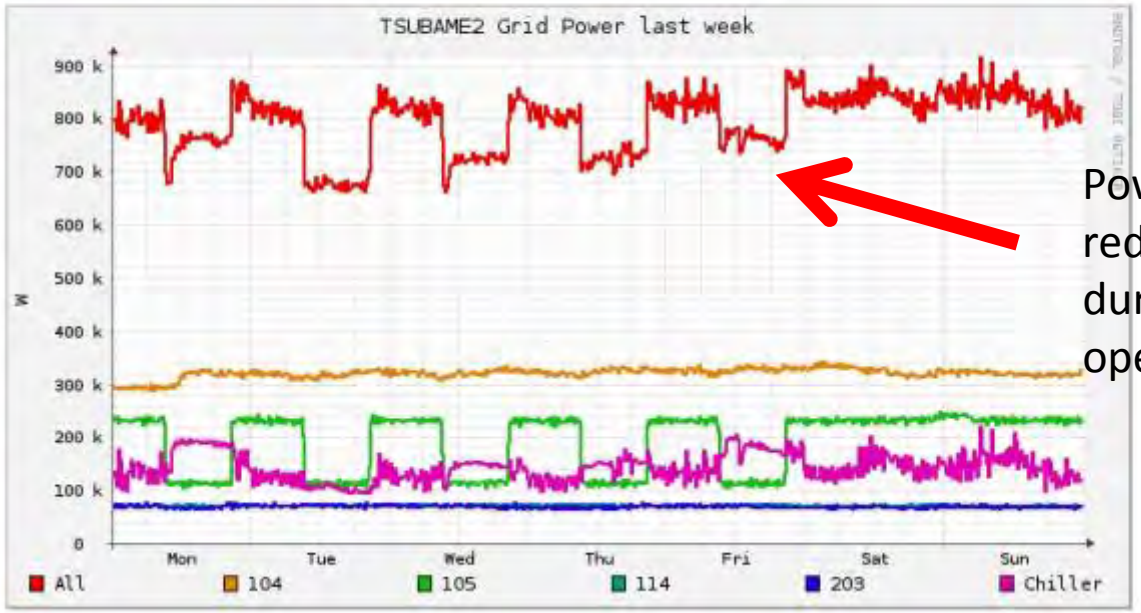
Date	Operation	# of nodes in operation
2010/11/1	Operation started	S 300, G/V 480, H/X 420
2011/3/11	Earthquake	
3/14	System shut down for rolling outage	0
3/17	Partially started, but stopped soon	S 160, G/V 100, H/X 150 -> 0
3/2--4/6	Maintainance	
4/6--4/7	Grand Challenge applications run	
4/8--4/24	Partial operation	S 280, G/V 100
4/25—6/8, 7/1—7/24	“Peak shift” operation	S 300, G/V 480, + X 420 (only nighttime)
6/9—6/30	Full operation temporarily	S 300, G/V 480, H/X 420
7/25—9/25	“Short makespan” operation	S 300, G/V 280, H/X 420 Y 200 (makespan <1h)
9/26—10/6	Grand Challenge applications run	

Realtime Measurement & Visualization of Power Consumption

- Now realtime power consumption information is available at http://mon.g.gsic.titech.ac.jp/powermon/

TSUBAME 2.0 Power Monitoring System

TSUBAME 2.0 All Power Summary



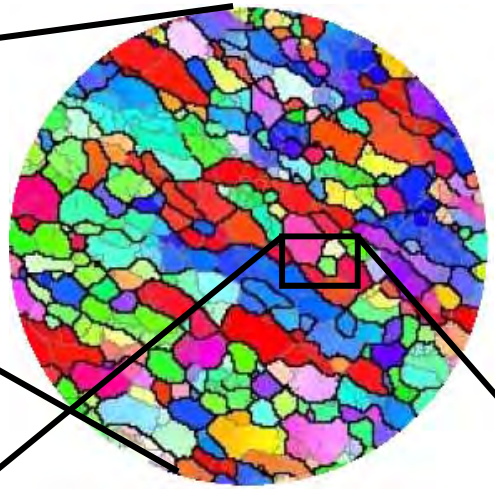
Power is succefully reduced in daytime during peak shift operation

Metal Dendritic Solidification

GP GPU

Mechanical Structure

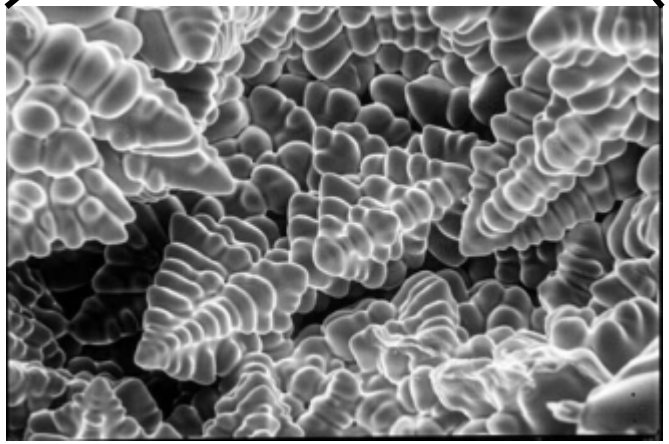
Material Microstructure



Low-carbon society

Improvement of fuel efficiency by reducing the weight of transportation

Developing lightweight strengthening material by controlling microstructure



Dendritic Growth

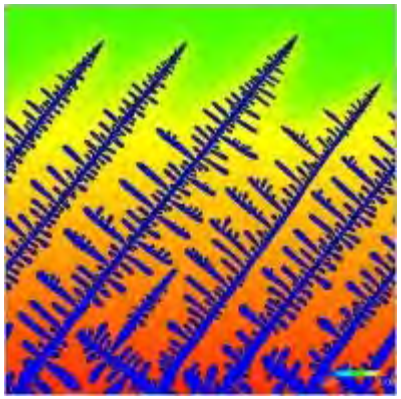
Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

Requirement for Peta-scale Computing

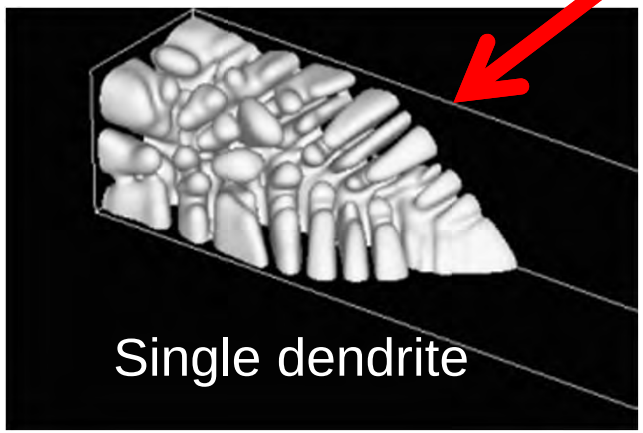
GP GPU

Previous works

2D computation

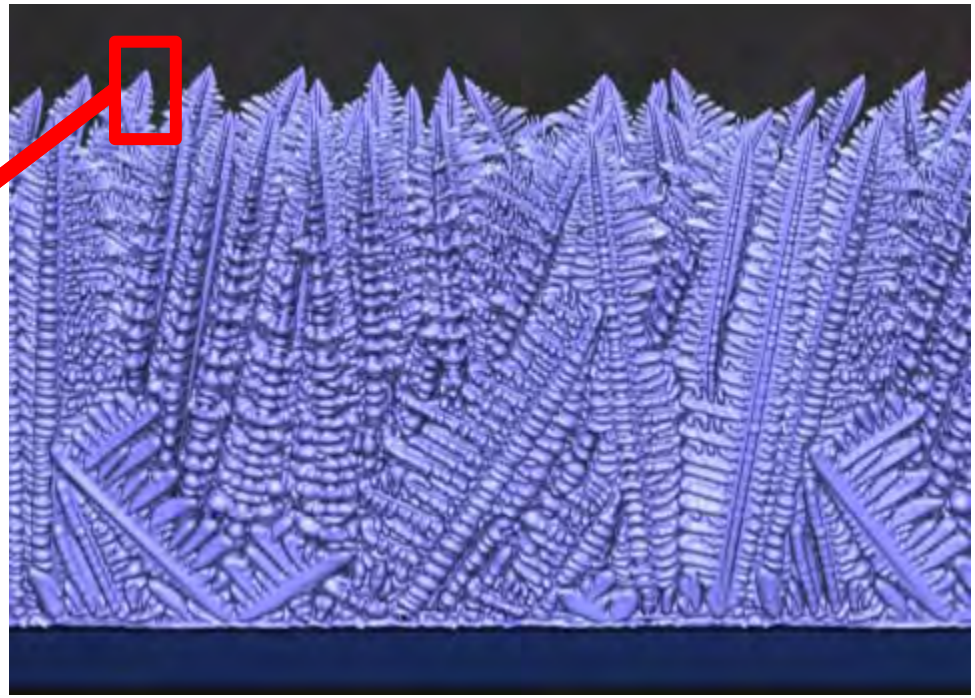


3D computation



×1000 large-scale computation

on TSUBAME 2.0

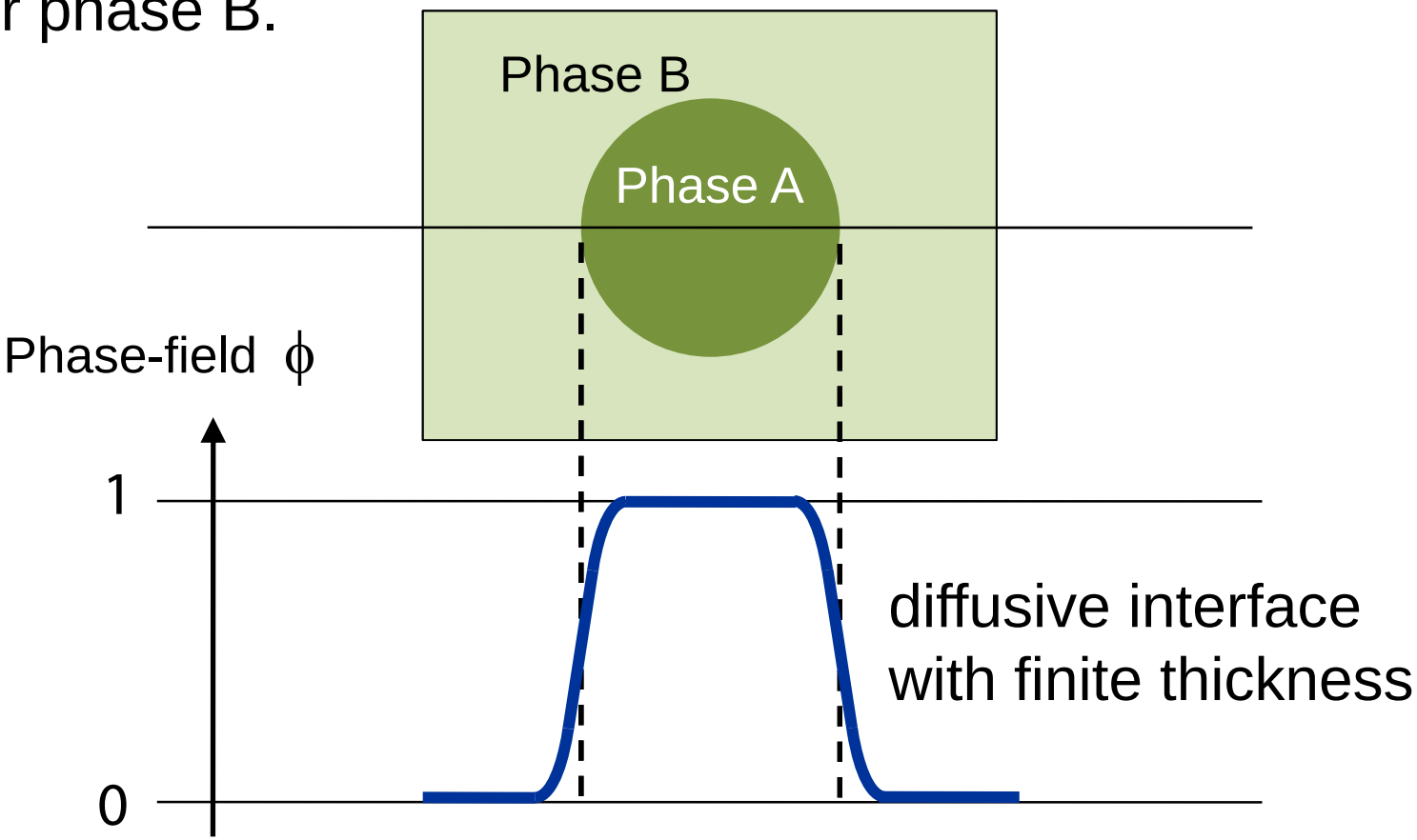


← ~ mm scale →

Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

Phase-Field Model

The phase-field model is derived from non-equilibrium statistical physics and f = 0 represents the phase A and f = 1 for phase B.



Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

Al-Si: Binary Alloy

GP GPU

Time evolution of the phase-field ϕ

（Allen-Cahn equation)

$$\frac{\partial \phi}{\partial t} = M_\phi \left[\nabla \cdot (a^2 \boldsymbol{\nabla} \phi) + \frac{\partial}{\partial x} \left(a \frac{\partial a}{\partial \phi_x} |\boldsymbol{\nabla} \phi|^2 \right) + \frac{\partial}{\partial y} \left(a \frac{\partial a}{\partial \phi_y} |\boldsymbol{\nabla} \phi|^2 \right) + \frac{\partial}{\partial z} \left(a \frac{\partial a}{\partial \phi_z} |\boldsymbol{\nabla} \phi|^2 \right) - \Delta S \Delta T \frac{dp(\phi))}{d\phi} - W \frac{dq(\phi)}{d\phi} \right]$$

Time evolution of the condensation: c

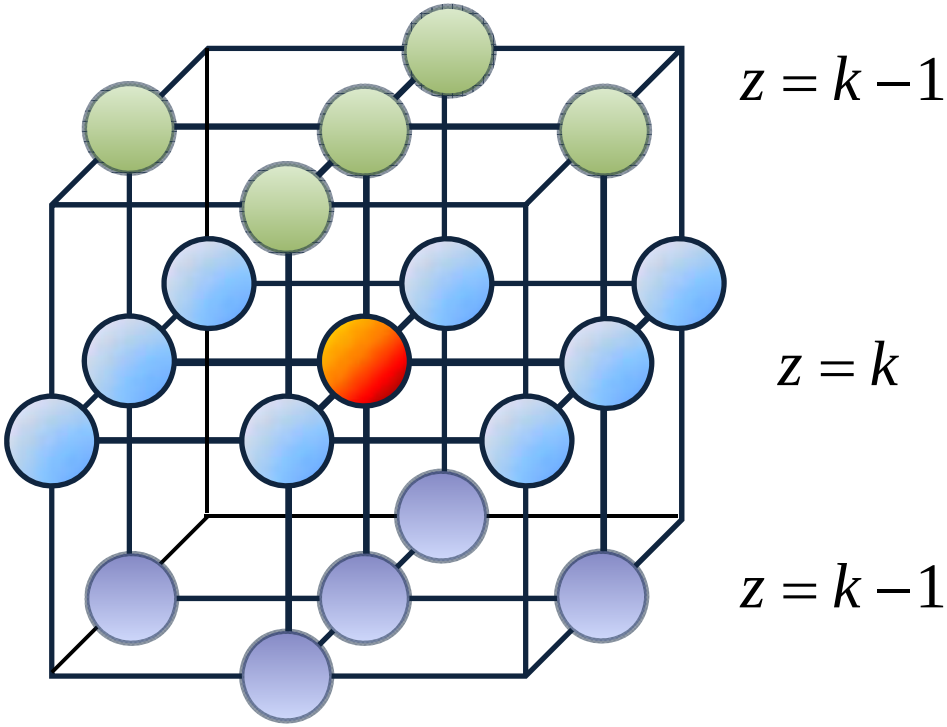
$$\frac{\partial c}{\partial t} = \boldsymbol{\nabla} \cdot [D_S \phi \boldsymbol{\nabla} c_S + D_L (1 - \phi) \boldsymbol{\nabla} c_L]$$

Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

Finite Difference Method

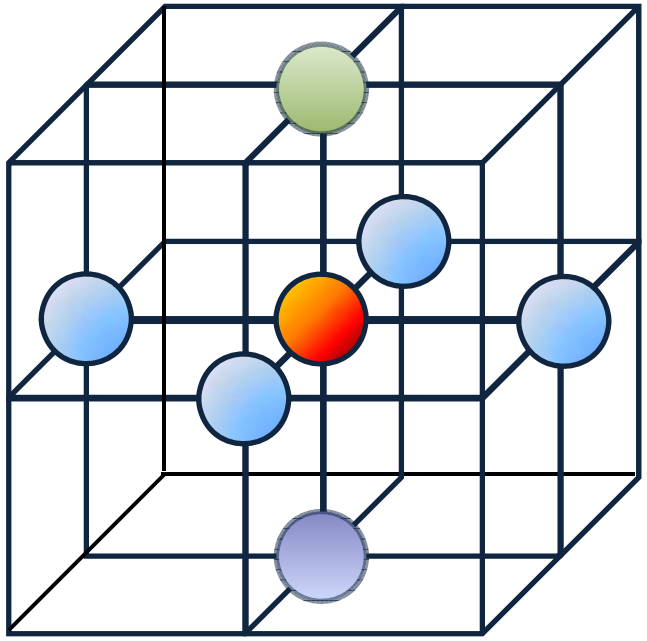
<u>Phase Field</u> : ϕ

19 points to solve $\phi_{i,j,k}$



<u>Condensation</u> : c

7 points to solve $c_{i,j,k}$



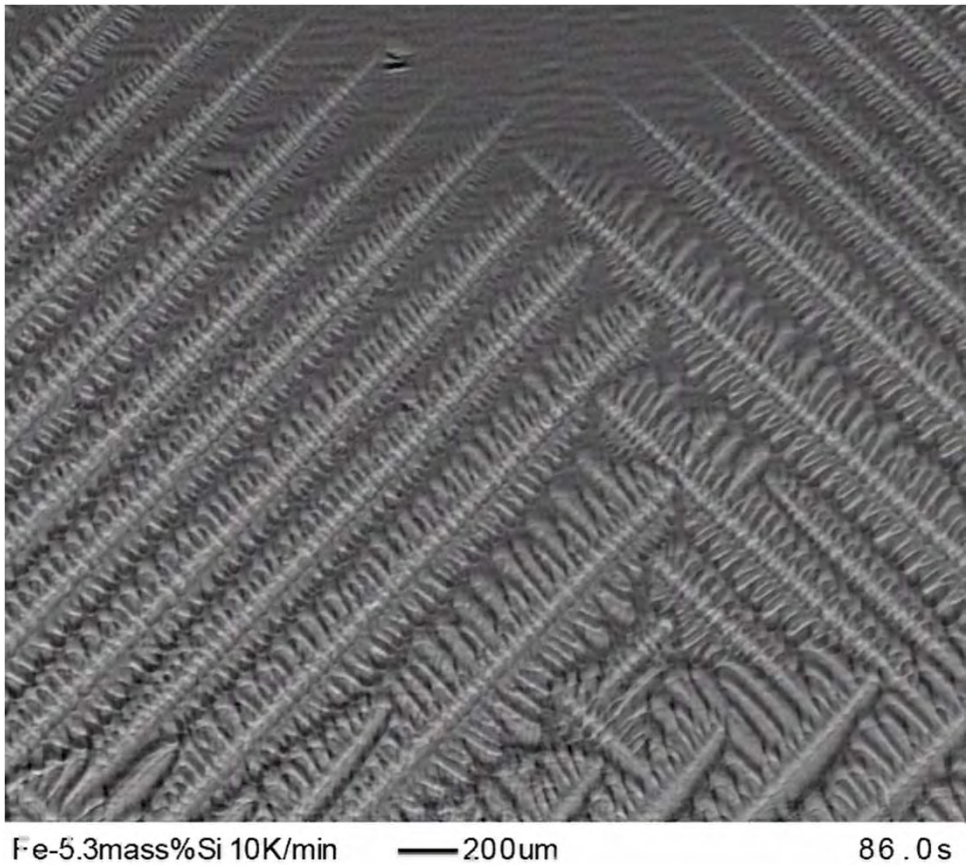
Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

Comparison with Experiment

GP GPU

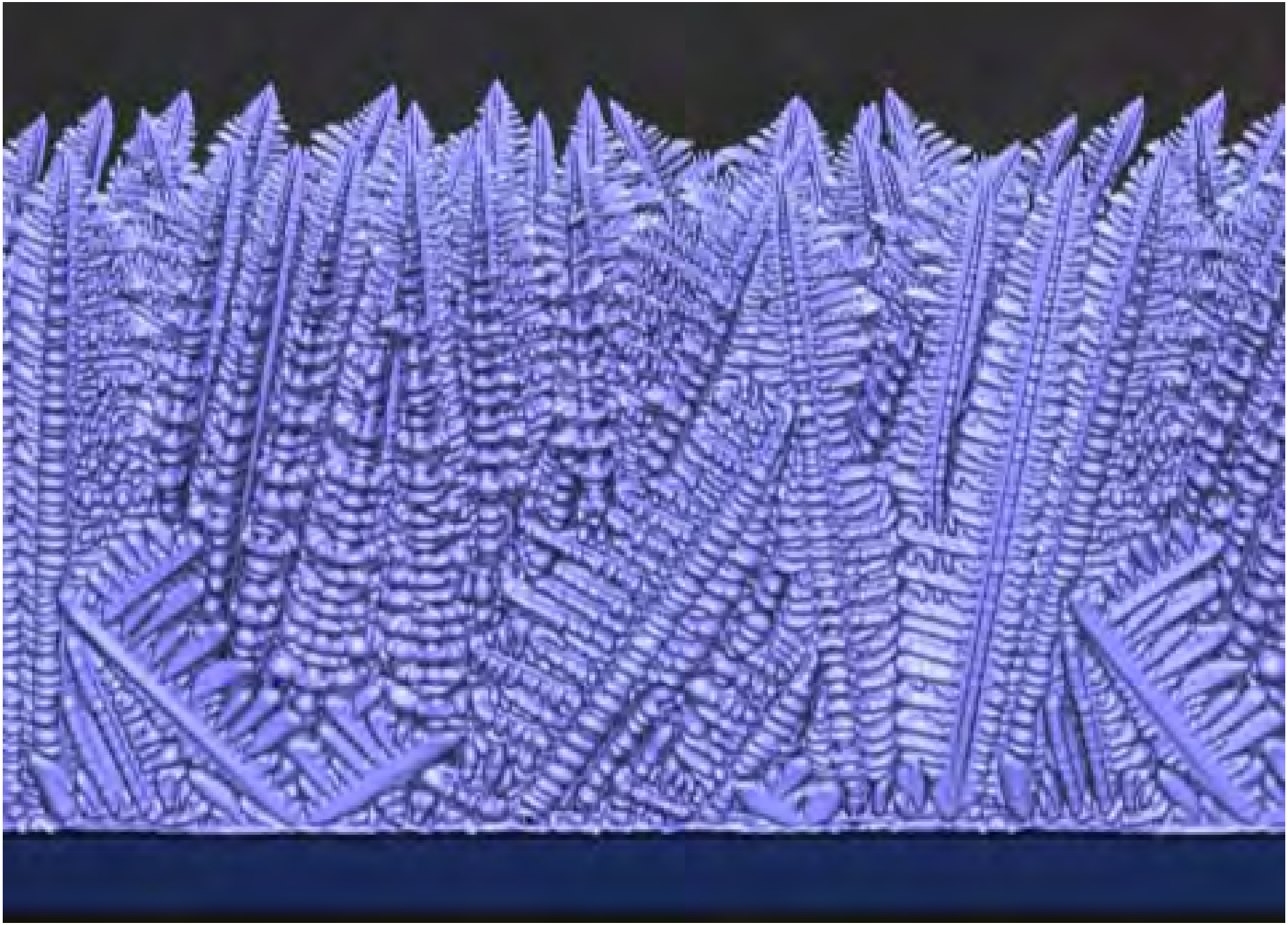
Observation:

X-ray imaging of Solidification of a binary alloy at Spring-8 in Japan by Prof. Yasuda (Osaka University in Japan)

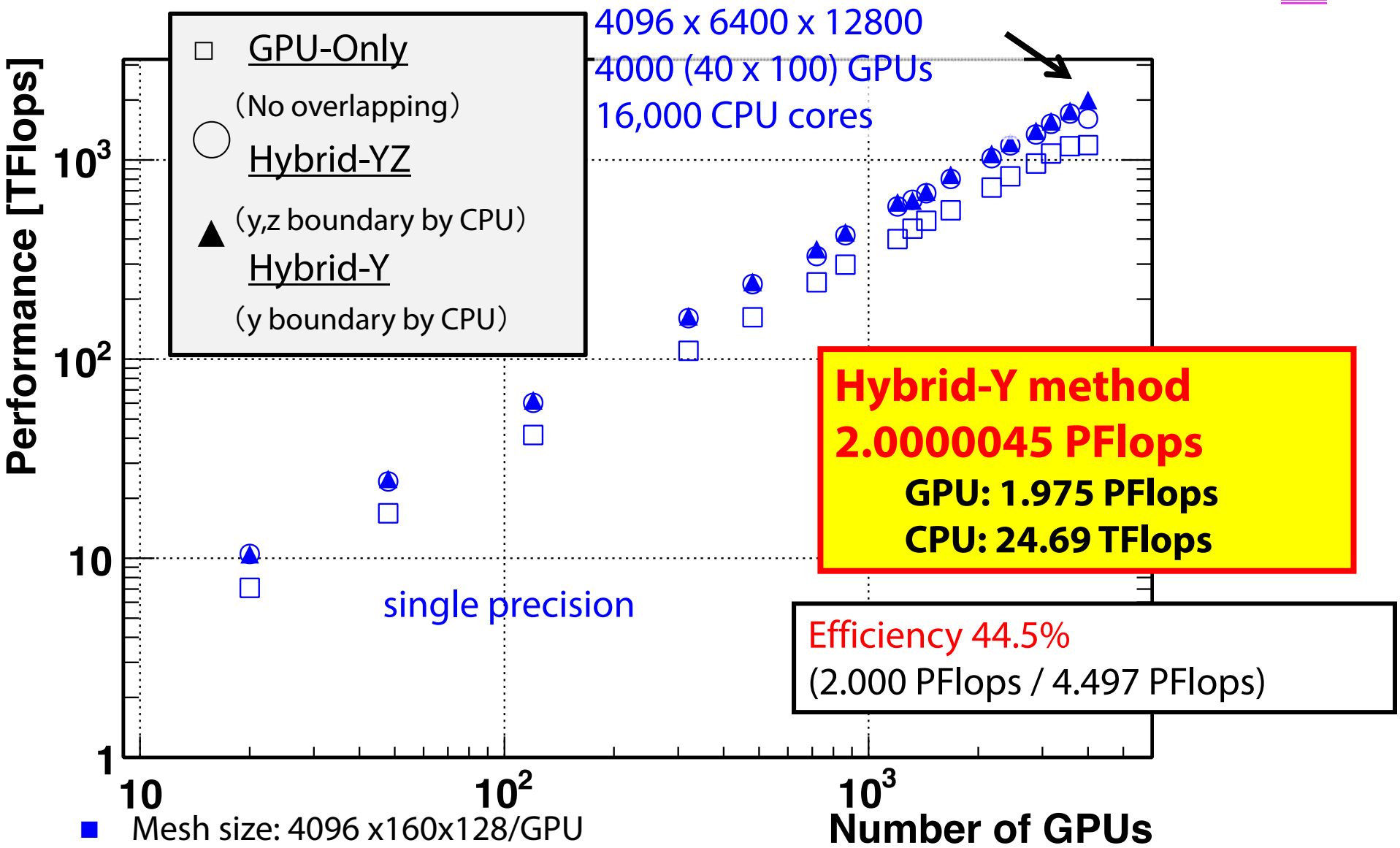


Phase-field simulation

Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology



Weak scaling results on TSUBAME 2.0



- Mesh size: 4096 x160x128/GPU
- NVIDIA Tesla M2050 card / Intel Xeon X5670 2.93 GHz on TSUBAME 2.0

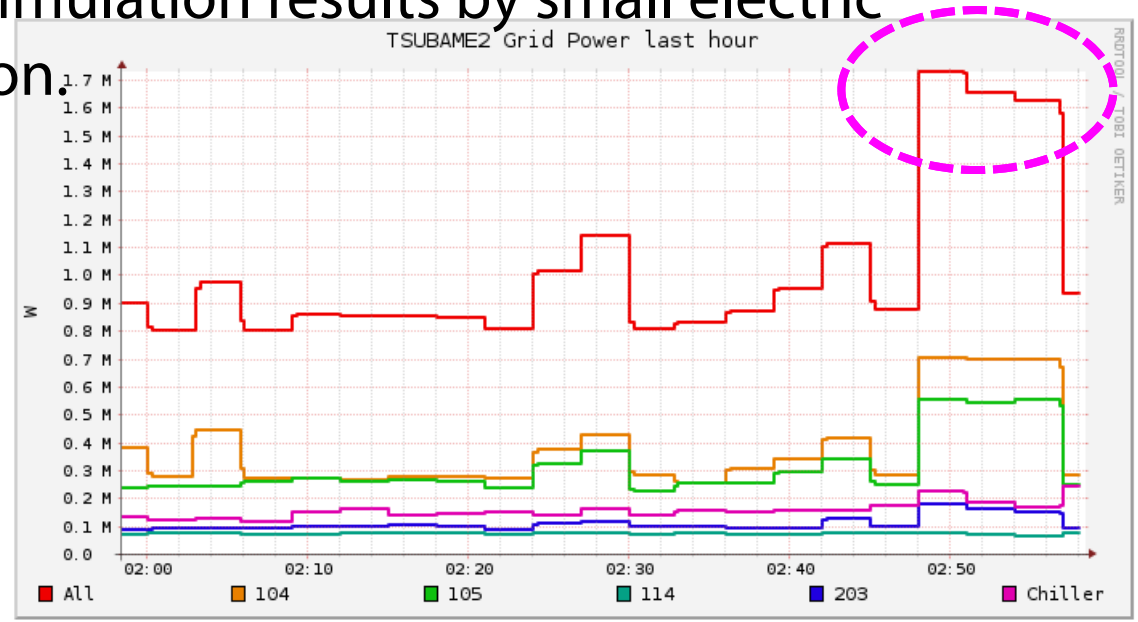
Power consumption and Efficiency

- The power consumption by application executions on TSUBAME 2.0 is measured in detail.
- Our phase-field simulation (real application)
  - ✓ 2.000 PFlops (single precision)
  - ✓ Performance to the peak: **44.5%**
  - ✓ **Green computing: 1468 MFlops/W**

2PFlops-Simulation

~1.36 MW
(Total:***1729kW***)

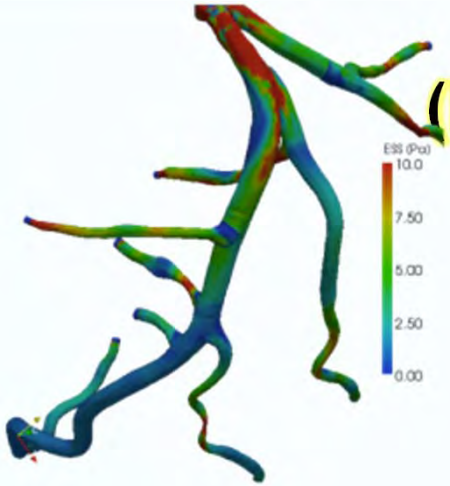
➡ We obtained the simulation results by small electric power consumption.



Ref.
Linpack benchmark
- ✓ 1.192 PFlops (DP)
- ✓ Efficiency 52.1%
- ✓ 827.8 MFlops/W

MUPHY: Multiphysics simulation of blood flow

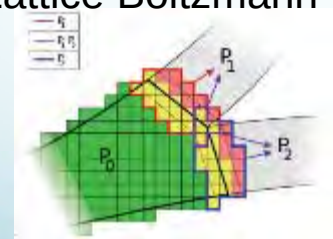
(Melchionna, Bernaschi et al.)



Multiphyics simulation with *MUPHY* software

Fluid: Blood plasma

Lattice Boltzmann

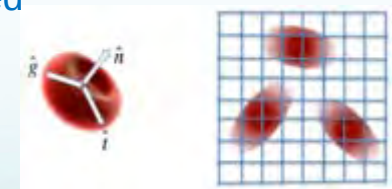


Irregular mesh is divided by using PT-SCOTCH tool, considering cutoff distance

coupled

Body: Red blood cell

Extended MD



Red blood cells (RBCs) are represented as ellipsoidal particles

Combined Lattice-Boltzmann (LB) simulation for plasma and Molecular Dynamics (MD) for Red Blood Cells

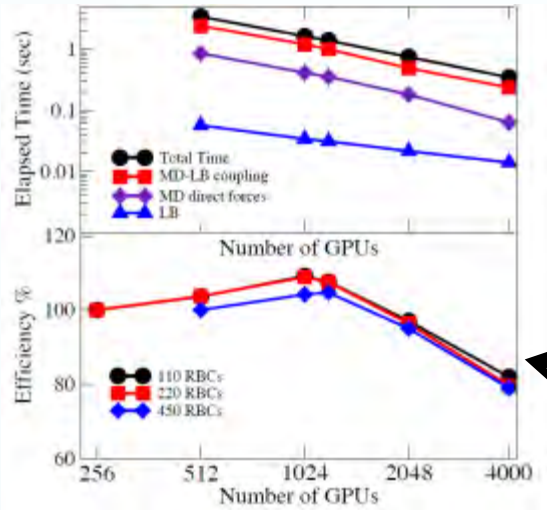
Realistic geometry (from CAT scan)

Two-levels of parallelism: CUDA (on GPU) + MPI

- 1 Billion mesh node for LB component
- 100 Million RBCs

Results on Tsubame2 Supercomputer (2) : Using 4,000 GPUs

Strong Scaling Results

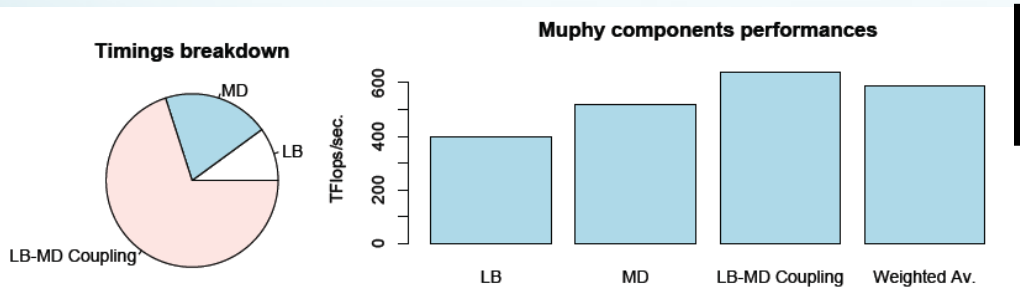


Elapsed time per timestep for 1G mesh nodes and 450M RBCs (log scale)

Parallel efficiency for 110, 220, 450M RBCs

~80% with 4K GPUs

Speeds per Component



0.6PFlops with 4,000GPUs
for 1G mesh nodes, 450M RBCs

A complete heartbeat at microsecond resolution can be simulated in 48hours

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

THE COMPUTER SOCIETY

SC11

ACM Gordon Bell Prize

Honorable Mention

Massimo Bernaschi, Mauro Bisson, Toshio Endo, Massimiliano Fatica, Satoshi Matsuoka, Simone Melchionna, Sauro Succi

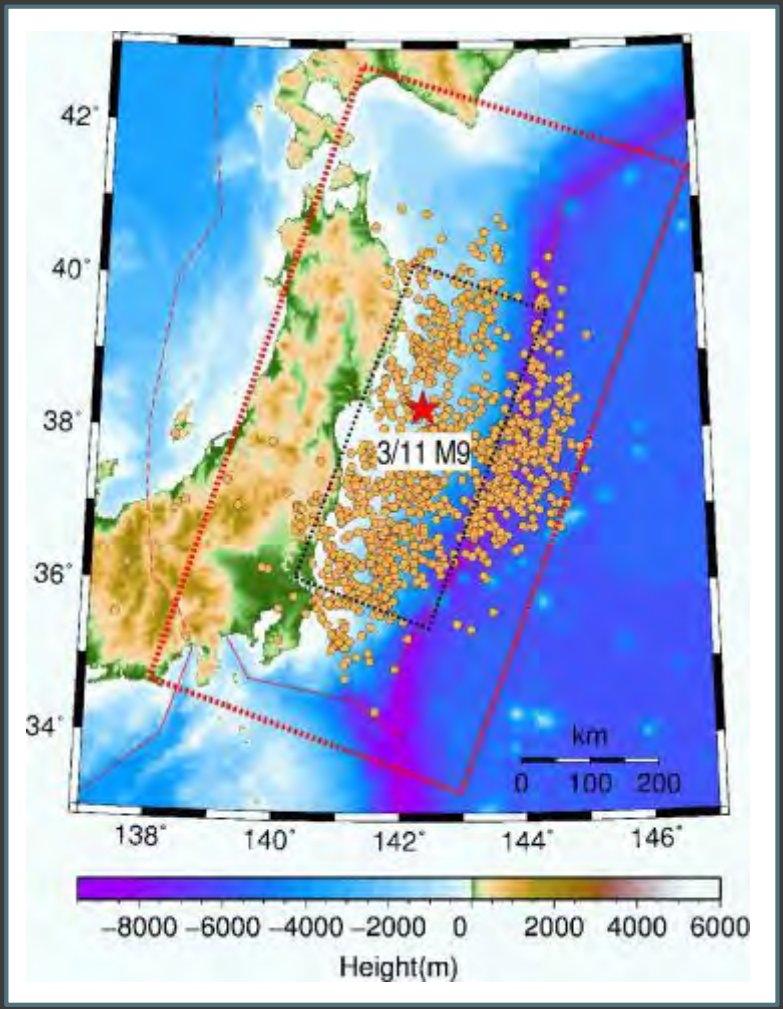
Petaflop Biofluidics Simulations On A Two Million-Core System

Scott Lathrop
SC11 Conference Chair

Thom H. Dunning, Jr.
Gordon Bell Chair

THE COMPUTER SOCIETY

2011 MAGNITUDE 9 TOHOKU-OKI EARTHQUAKE



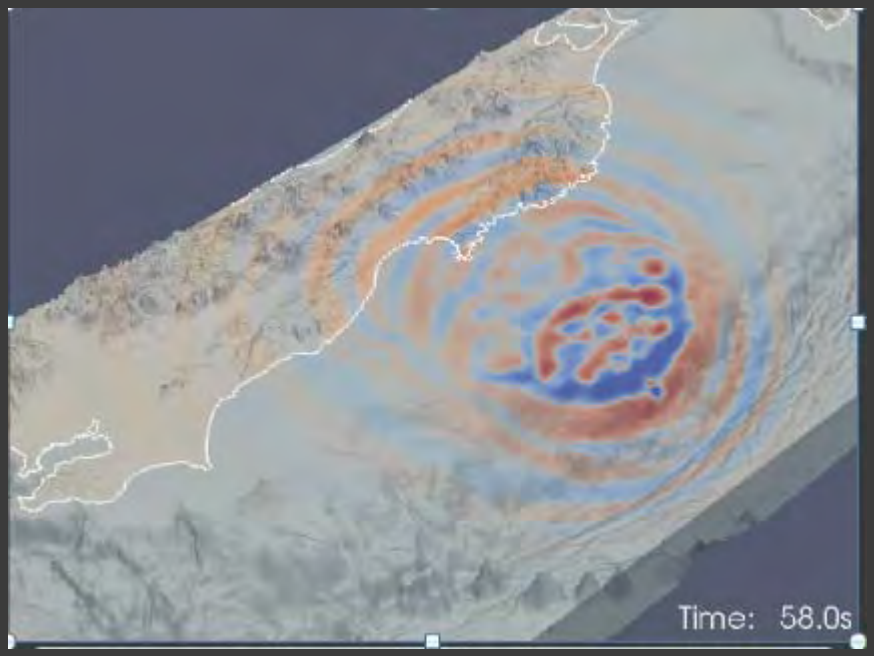
Aftershock Distribution

- Fatalities: 19,508
  - Strong shakings and devastating tsunamis
- Large source area
  - 500km x 200 km
  - Inner black rectangle
- Large FDM region required
  - 960km x 480km in horizontal
  - 240km in depth
  - Outer red rectangle

TARO OKAMOTO
TOKYO INSTITUTE OF TECHNOLOGY

2011 MAGNITUDE 9 TOHOKU-OKI EARTHQUAKE

FDTD Simulation of Wave Propagation



Main part of the FDM region

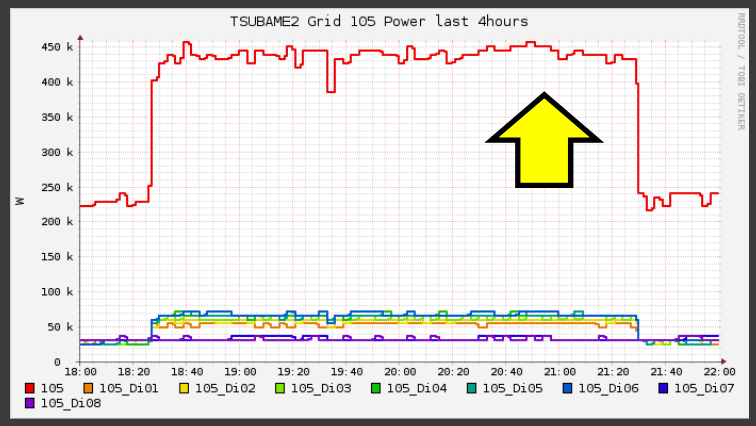
- Finite-Difference Time Domain (Okamoto et al. 2010)
  - Topography, ocean layer, and heterogeneity
  - Grid size: 6400 x 3200 x 1600
  - Grid spacing: 150 m
  - Time interval: 0.005 s
  - **1000 GPUs of TSUBAME-2.0**
  - Preliminary source model
- Visualization
  - Vertical ground motion on land ocean bottom

TARO OKAMOTO
TOKYO INSTITUTE OF TECHNOLOGY

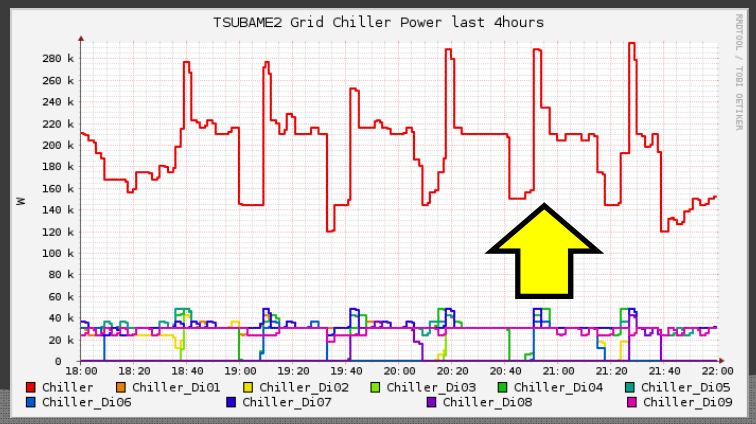
2011 MAGNITUDE 9 TOHOKU-OKI EARTHQUAKE

Power Consumption during 700-node Run

Compute nodes (partial)



Chiller (shared by all jobs)



Compute node:

903**kW in total**

550kW for This app

(estimate from 540nodes)

Storage:

72kW

Cooling:

345kW at max

(shared by all jobs)

Total:

1320kW at max

TARO OKAMOTO
TOKYO INSTITUTE OF TECHNOLOGY

Next Generation Numerical Weather Prediction[SC10]

Collaboration: Japan Meteorological Agency

Meso-scale Atmosphere Model:

Cloud Resolving Non-hydrostatic model
[Shimokawabe et. al. SC10 BSP Finalist]

ex. WRF(Weather Research and Forecast)

Typhoon ~ 1000km

1~ 10km
Tornado,
Down burst,
Heavy Rain

WSM5 (WRF Single Moment 5-tracer) Microphysics*

Represents condensation, precipitation and thermodynamic effects of latent heat release

1 % of lines of code, 25 % of elapsed time

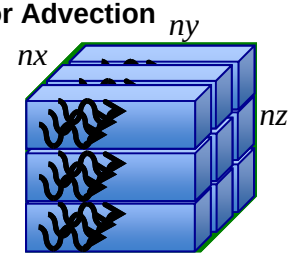
⇒ 20 x boost in microphysics (1.2 - 1.3 x overall improvement)

ASUCA : full GPU Implementation

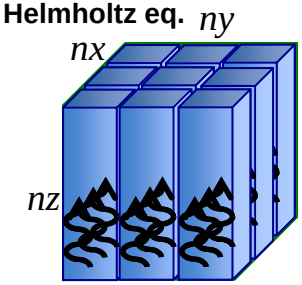
developed by Japan Meteorological Agency

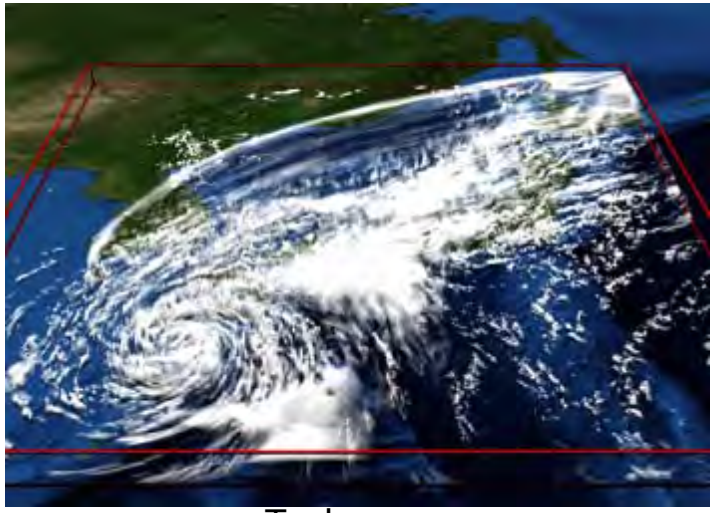
TSUBAME 2.0 : 145 Tflops
World Record !!!

Block Division for Advection

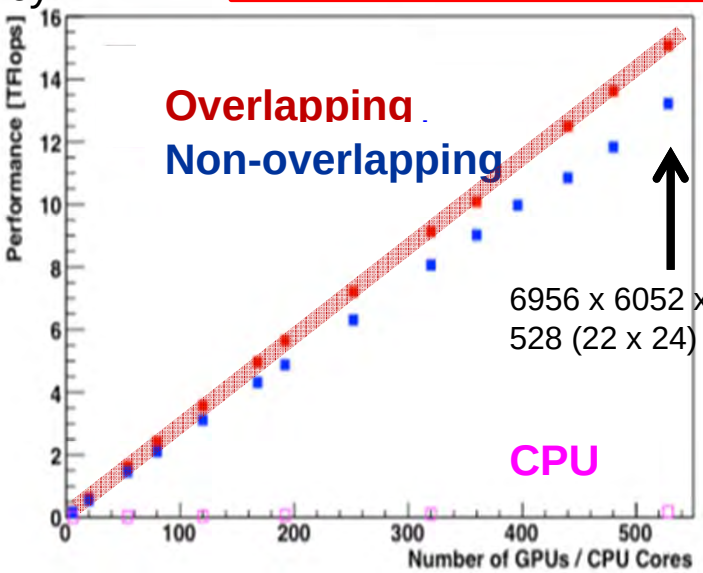


for 1-D Helmholtz eq.



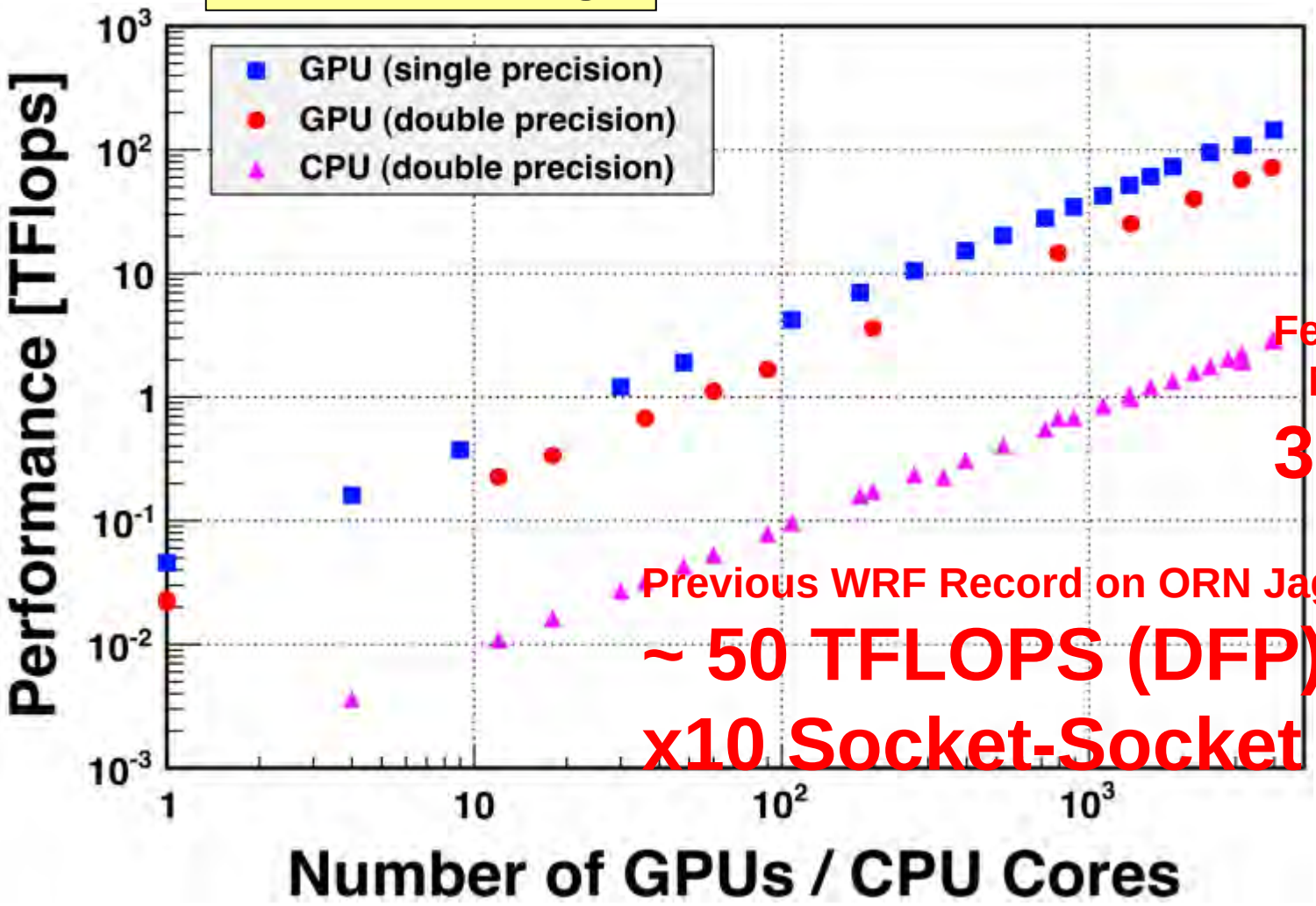


Typhoon



TSUBAME 2.0 Performance

Weak Scaling



145.0 Tflops
Single precision

76.1 Tflops
Doublele precision

Fermi core Tesla M2050

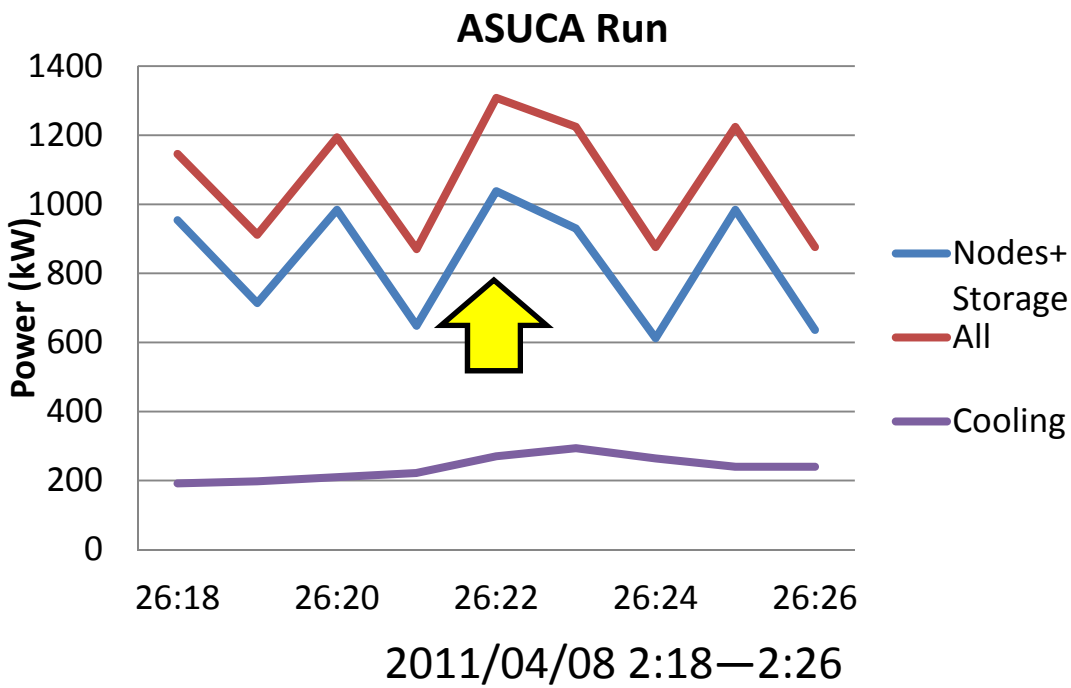
3990 GPUs

Previous WRF Record on ORN Jaguar

~ 50 TFLOPS (DFP)

x10 Socket-Socket

Power Consumption during Full TSUBAME2 Test with ASUCA



Compute node:

*960**kW***

Storage:

78kW

Cooling:

270kW max

Total:

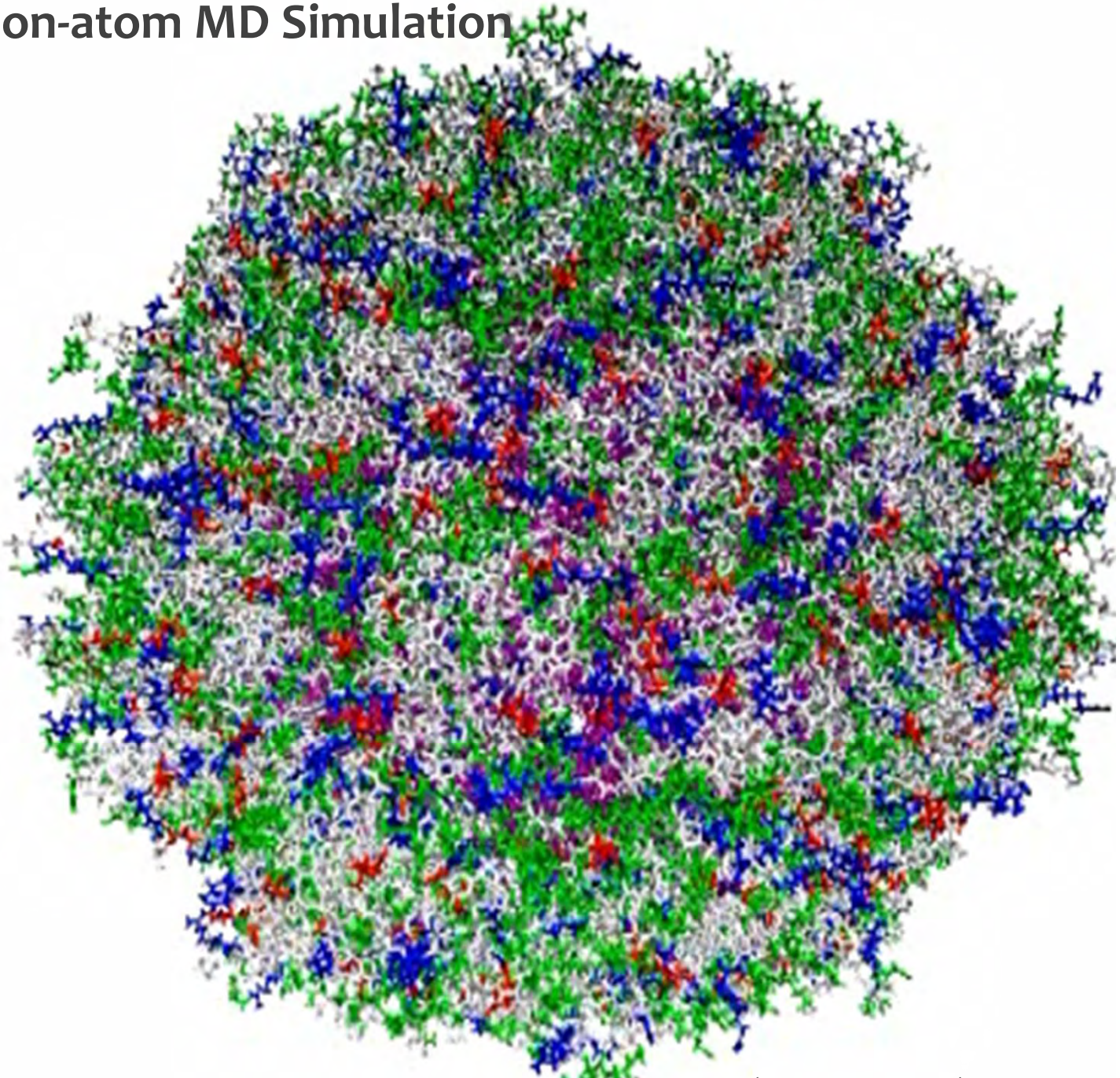
1308kW max

100-million-atom MD Simulation

M. Sekijima (Tokyo Tech), Jim Phillips (UIUC)

- NAMD is a parallel molecular dynamics code developed at University of Illinois.
- This evaluation is result of an interdisciplinary collaboration between UIUC and Tokyo Tech.
- The 100-million-atom benchmark in this work was assembled by replicating a million-atom satellite tobacco mosaic virus (STMV) simulation on a 5x5x4 grid.
- One STMV (Satellite Tobacco Mosaic Virus) includes 1,066,628 atoms.

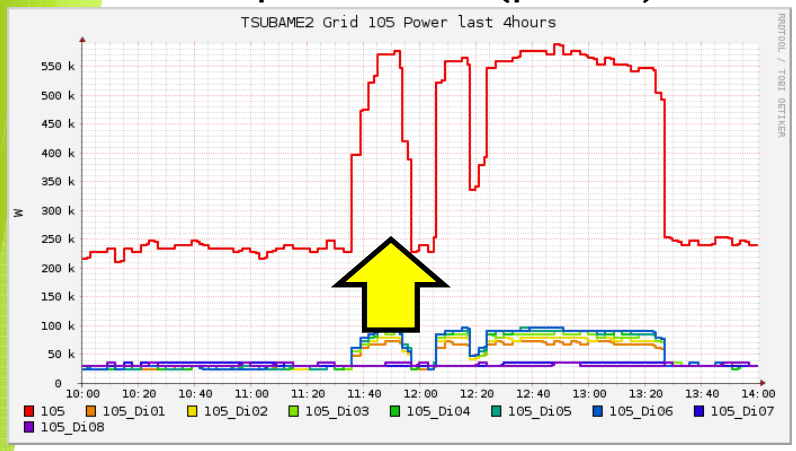
100-million-atom MD Simulation



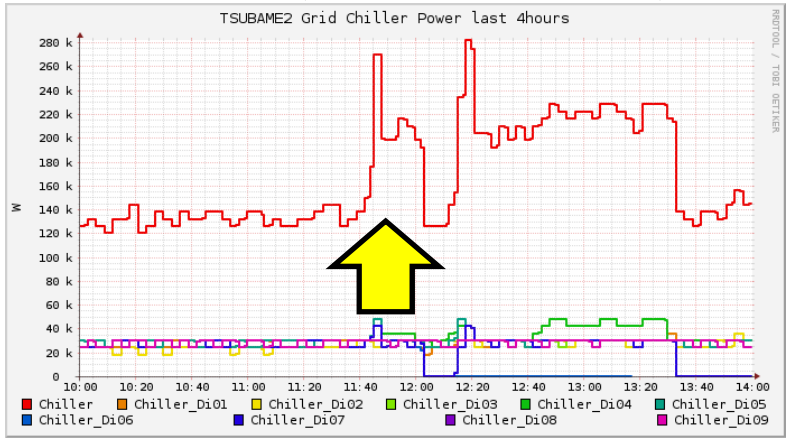
M. Sekijima (Tokyo Tech), Jim Phillips (UIUC)

Power Consumption during 700-node Run

Compute nodes (partial)



Chiller (shared by all jobs)



Compute node:

1115kW in total

706kW for This app

(estimate from 540nodes)

Storage:

72kW

Cooling:

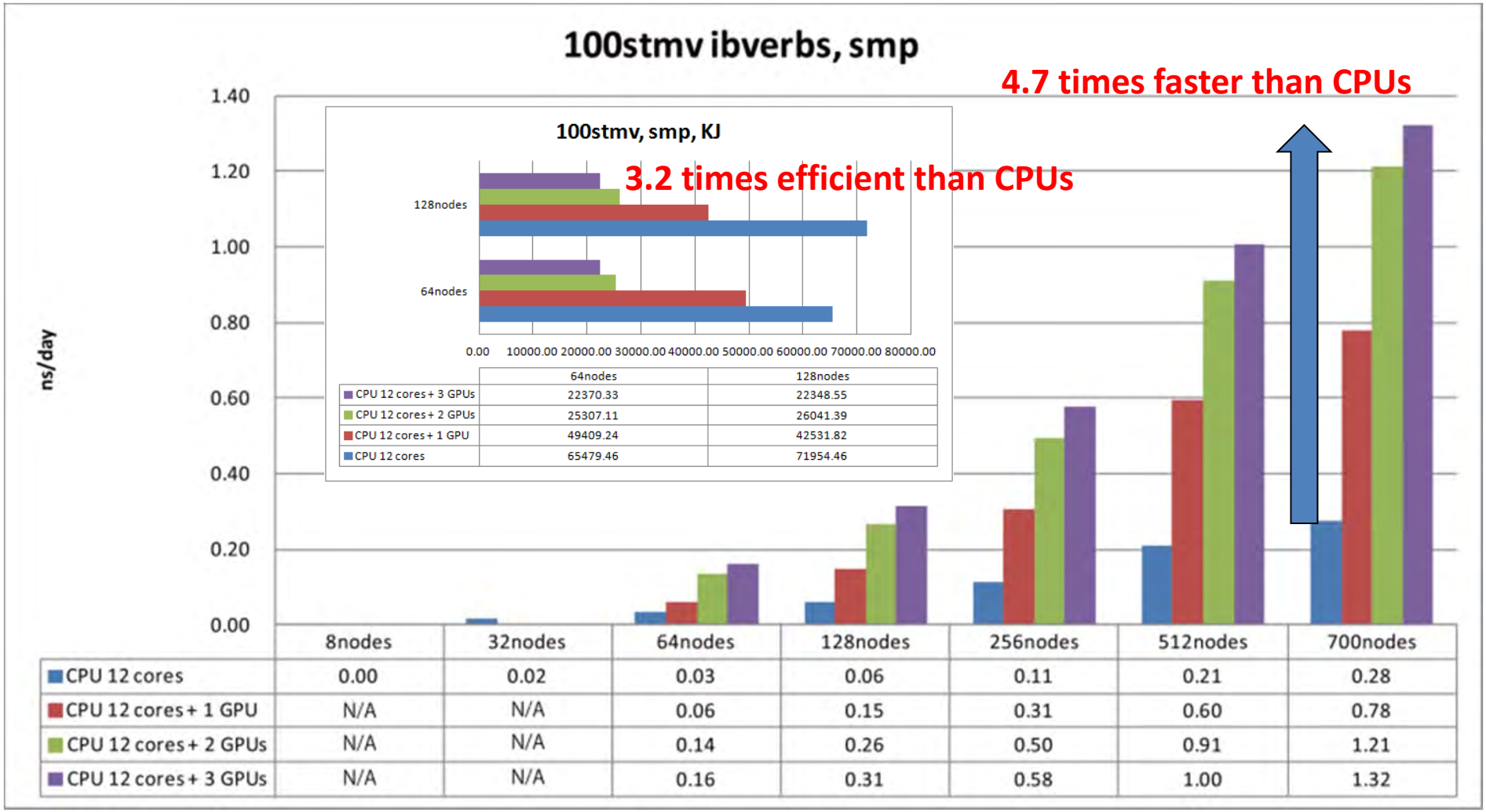
340kW max

(shared by all jobs)

Total:

1527kW max

M. Sekijima (Tokyo Tech), Jim Phillips (UIUC)

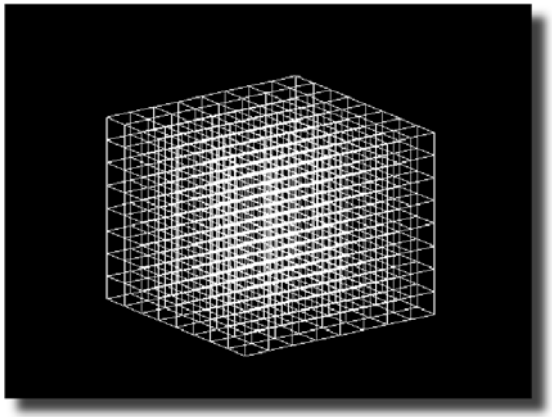


	8nodes	32nodes	64nodes	128nodes	256nodes	512nodes	700nodes
CPU 12 cores	0.00	0.02	0.03	0.06	0.11	0.21	0.28
CPU 12 cores + 1 GPU	N/A	N/A	0.06	0.15	0.31	0.60	0.78
CPU 12 cores + 2 GPUs	N/A	N/A	0.14	0.26	0.50	0.91	1.21
CPU 12 cores + 3 GPUs	N/A	N/A	0.16	0.31	0.58	1.00	1.32

Performance Evaluation

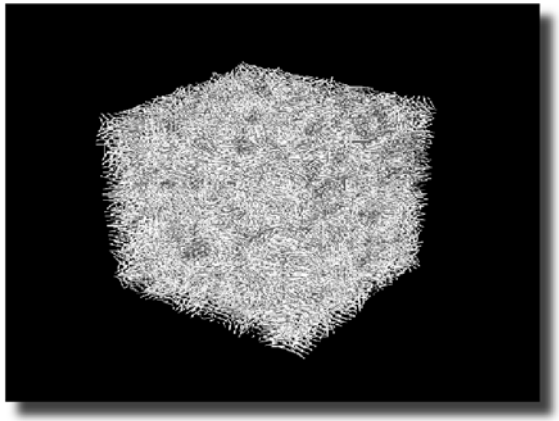
Petaflops scale turbulence simulation on TSUBAME 2.0

Isotropic turbulence



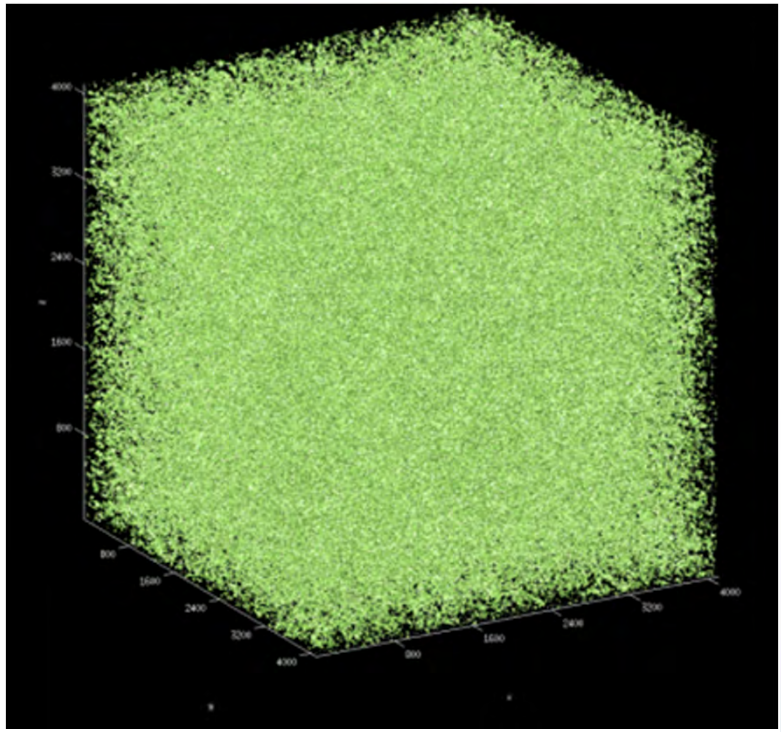
Pseudo Spectral Method
(2/3 dealiasing)

Re_L : 500
N : 2048^3



Vortex Particle Method
(Reinitialized CSM)

Re_L : 500
N : 2048^3

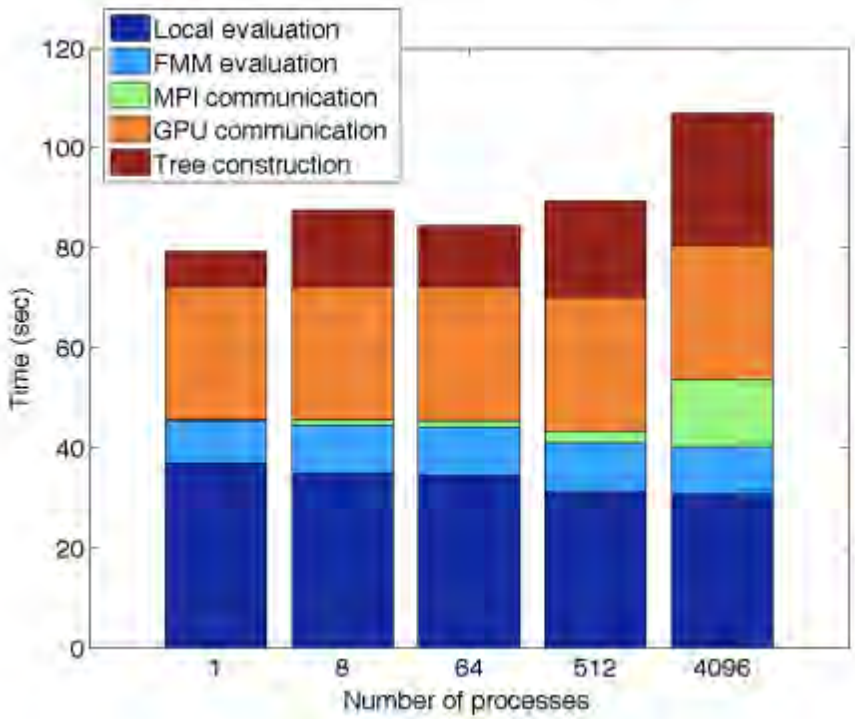


8 billion particles

R. Yokota (KAUST) , L. A. Barba (Boston Univ), T. Narumi (Univ of Electro Communications), K. Yasuoka (Keio Univ)

Petaflops scale turbulence simulation on TSUBAME 2.0

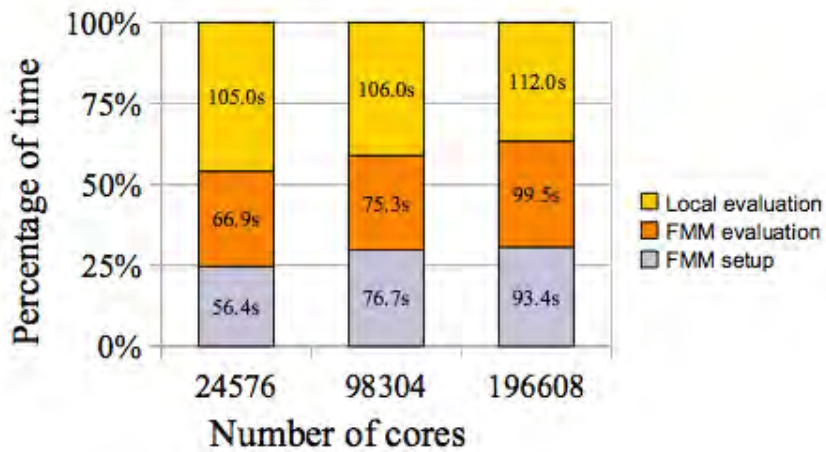
Present work



64 billion in 100 seconds
1.0 PFlops

Rahimian et al. (2010 Gordon Bell)

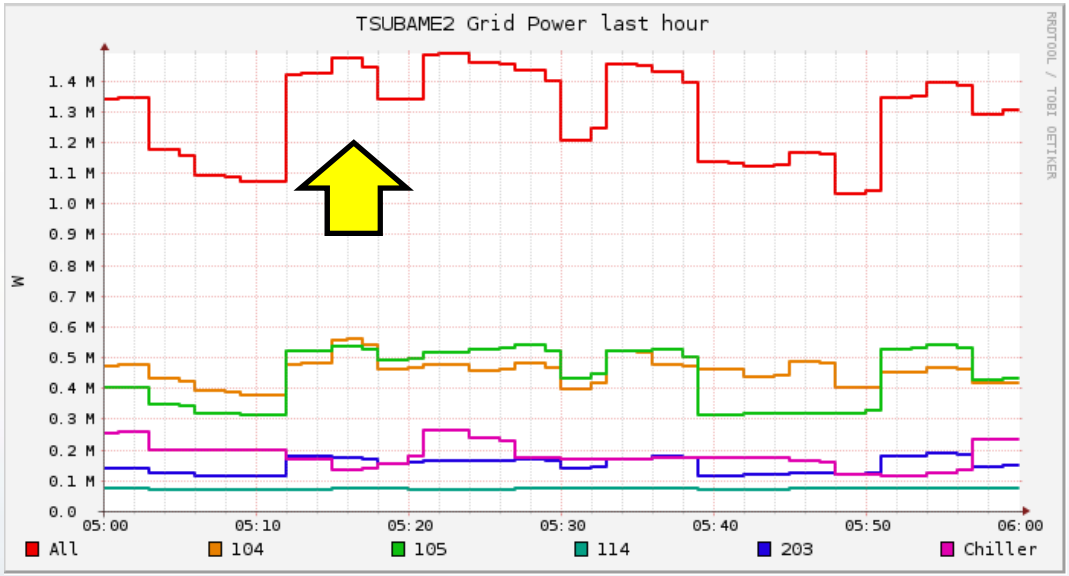
p	24576	98304	196608
Time (sec)	228.3	258	304.9
Efficiency	1.00	0.88	0.75



90 billion in 300 seconds
0.7 PFlops

R. Yokota (KAUST) , L. A. Barba (Boston Univ), T. Narumi (Univ of Electro Communications), K. Yasuoka (Keio Univ)

Power Usage during Full System Test



2011/10/4 5:00—6:00

Compute node:
1190kW

Storage:
72kW

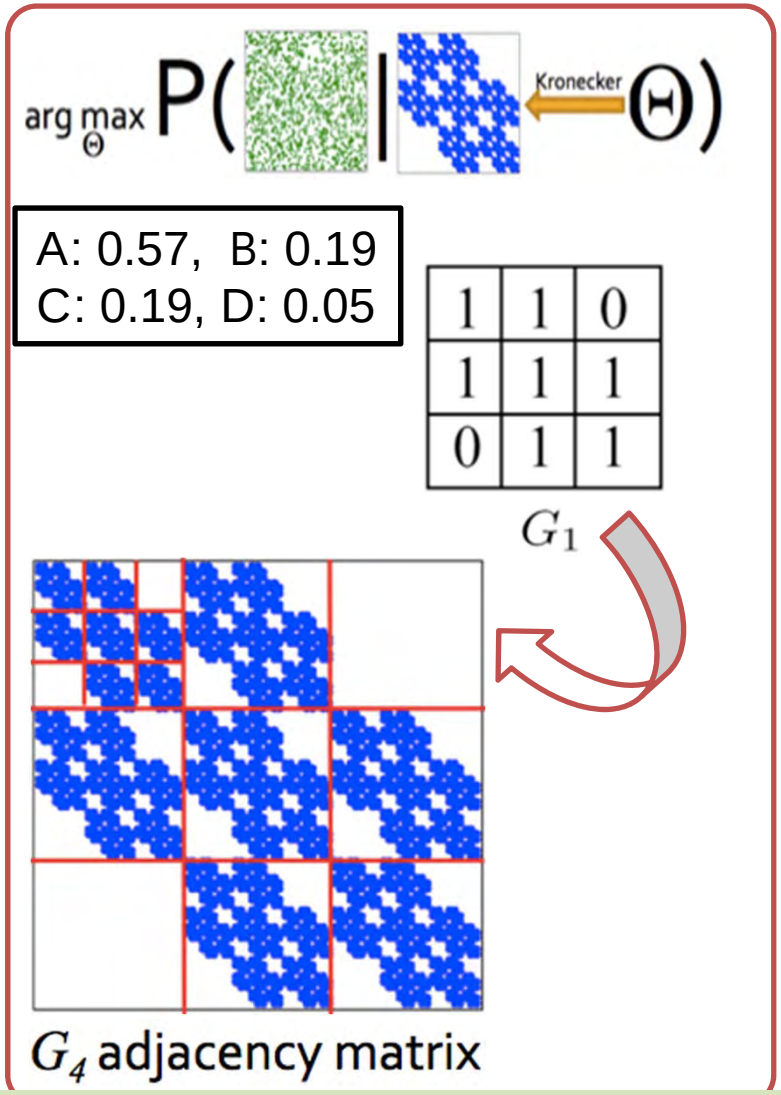
Cooling:
240kW

Total:
1502kW

R. Yokota (KAUST) , L. A. Barba (Boston Univ), T. Narumi (Univ of Electro Communications), K. Yasuoka (Keio Univ)

Graph500 on TSUBAME 2.0

Kronecker graph



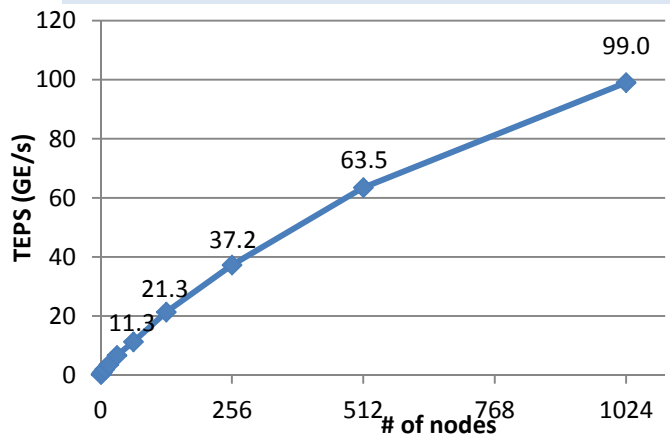
- Graph500 is a new benchmark that ranks supercomputers by executing a large-scale graph search problem.
- The benchmark is ranked by so-called **TEPS (Traversed Edges Per Second)** that measures the number of edges to be traversed per second by searching all the reachable vertices from one arbitrary vertex with each team's optimized BFS (Breadth-First Search) algorithm.

Toyotaro Suzumura, Koji Ueno, Tokyo Institute of Technology

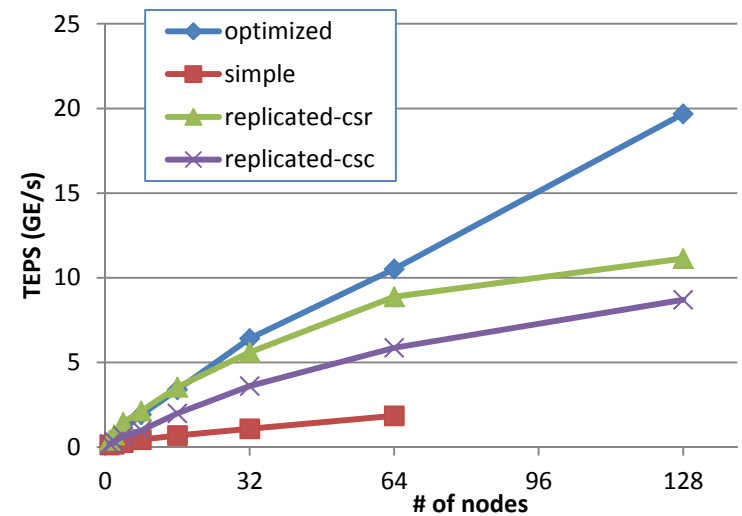
Highly Scalable Graph Search Method for the Graph500 Benchmark

- Our early study reveals that the provided reference implementations are not scalable in a large-scale distributed environment.
- We propose an optimized method based on 2D based partitioning and other various optimization methods such as communication compression and vertex sorting.
- Our optimized implementation can solve BFS (Breadth First Search) of large-scale graph with 2^{36} (68.7 billion) vertices and 2^{40} (1.1 trillion) edges for 10.58 seconds with 1366 nodes and 16392 CPU cores on TSUBAME 2.0
- This record corresponds to 103.9 GE/s (TEPS)

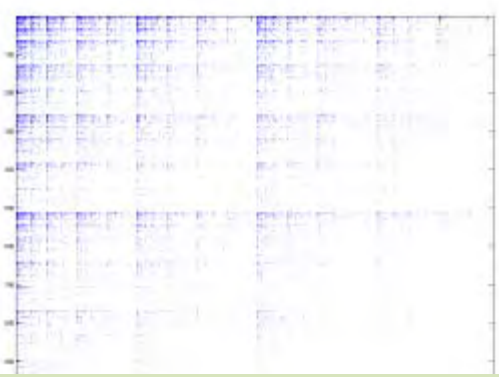
Performance of Our Optimized Implementation with Scale 26 per 1 node



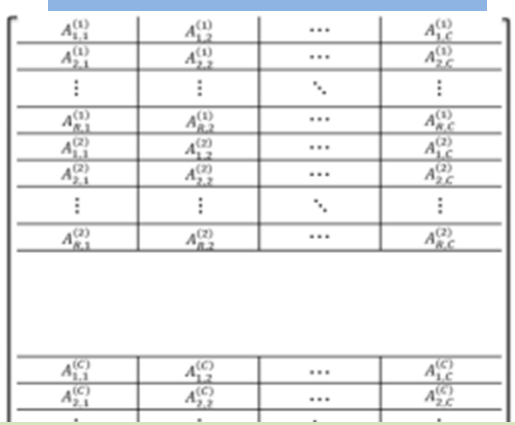
Performance Comparison with Reference Implementations (simple, replicated-csr and replicated-csc) and Scale 24 per 1 node



Vertex Sorting by utilizing the scale-free nature of the Kronecker Graph

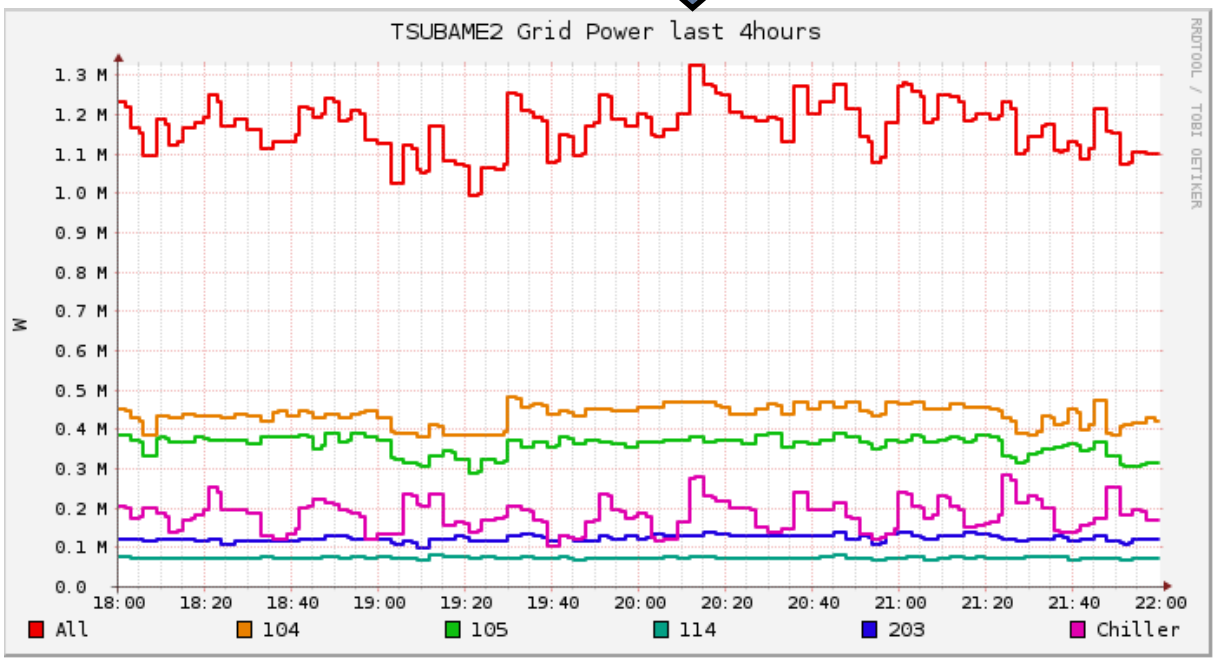


2D Partitioning Optimization



Toyotaro Suzumura, Koji Ueno, Tokyo Institute of Technology

Power Consumption during Graph500 Run on TSUBAME 2.0



2011/10/4 18:00—22:00

Compute node:
902kW

Storage:
75kW

Cooling:
346kW max

Total:
1323kW max

Toyotaro Suzumura, Koji Ueno, Tokyo Institute of Technology

TSUBAME2.0 Power Consumption with Petascale Applications

	Compute nodes & Network(kW)	Storage (kW)	Cooling (kW)	Total (kW)	Cooling/Total
Typical Production	750	72	230	980	23.5%
Earthquake (2000 GPUs)	550/903	72	345	1320	26.1%
NAMD MD (2000 GPUs)	706/1115	72	340	1527	22.3%
ASUCA Weather	960	78	270	1308	20.6%
Turbulence FMM	1190	72	240	1502	16.0%
Graph500	902	75	346	1323	26.2%
Phase-field	1362	73	294	1729	17.0%
GPU DGEMM	1538	72	410	2020	20.3%
Linpack (Top500)	1417	72	-	-	-

Awards and Accolades

- Green500 Greenest Production Supercomoputer (Nov.2010/Jun2011)
- ACM Gordon Bell Award 2011 x 2
- SC11 Tech Paper "Perfect Score" Award
- ACM Fellow (Matsuoka), CUDA Fellow (Aoki)
- HPCWire Editor & Reader's Choice Awards(2011) – Industry Collab. & Green
- "Clear Overall Winner in Number Crunching, Data Crunching and Energy Efficiency" HPCWire Feb.
- Commendation for Science and Technology by the Minister of MEXT
- Etc. etc. … (~20 related awards)

DoE Exascale Parameters

x1000 power efficiency in 10 years

System attributes	"2010"		"2015"		"2018-20"	
System peak	2 PetaFlops		100-200 PetaFlops		1 ExaFlop	
Power	Jaguar 6 MW	TSUBAME 1.3 MW	15 MW		20 MW	
System Memory	0.3PB	0.1PB	5 PB		32-64PB	
Node Perf	125GF	1.6TF	0.5TF	7TF	1TF	10TF
Node Mem BW	25GB/s	0.5TB/s	0.1TB/s	1TB/s	0.4TB/s	4TB/s
Node Concurrency	12	O(1000)	O(100)	O(1000)	O(1000)	O(10000)
#Nodes	18,700	1442	50,000	5,000	1 million	100,000
Total Node Interconnect	1.5GB/s	8GB/s	20GB/s		200GB/s	

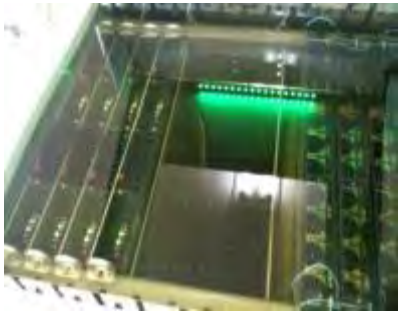
Ultra Low Power HPC Towards TSUBAME3.0

2006
TSUBAME1

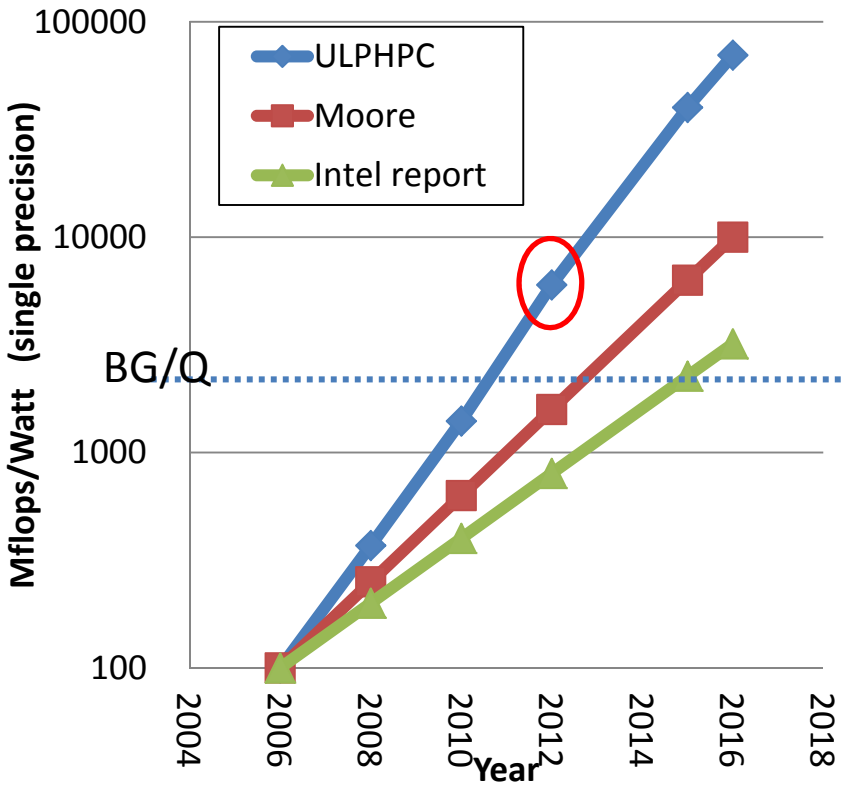
2008
TSUBAME1.2
1st GPU SC

2010
TSUBAME2.0
Greenest Production

2012
3.0 Proto1, 2, …



Year	Machine	Peak GFlops	Estim. App GFlops	KWatt	Estim. Mflops /W
2006 T1.0	Sun X4600	160	128	1.3	100
2008 T1.2	Sun X4600 + Tesla S1070 2GPUs	1300	630	1.7	370
2010 T2.0	HP SL390s with Tesla M2050 3GPUs	3200	1400	1.0	1400
2012 Proto	Sandy Bridge EP w/ 8 Kepler GPUs	20000	7000	1.1	>6000



Power Efficiency Compared

Machine	Power (incl. cooling)	Linpack Perf (PF)	Linpack Mflops/W	Factor
Tsubame1.0 (2006Q1)	1.8MW	0.038	21	2368
ORNL Jaguar (2009Q4)	~9MW	1.76	196	256
Tsubame2.0 (2010Q4)	1.8MW	1.2	667	75 *x31.6*
K Computer (2011Q2)	~16MW	10	625	80
BlueGene/Q (2012Q1)	~12MW?	17	1417	35.3
Tsubame3.0 (2015Q1)	1.8MW	20	11000	4.6 *~x16*
EXA (2018Q4)?	20MW	1000	50000	1