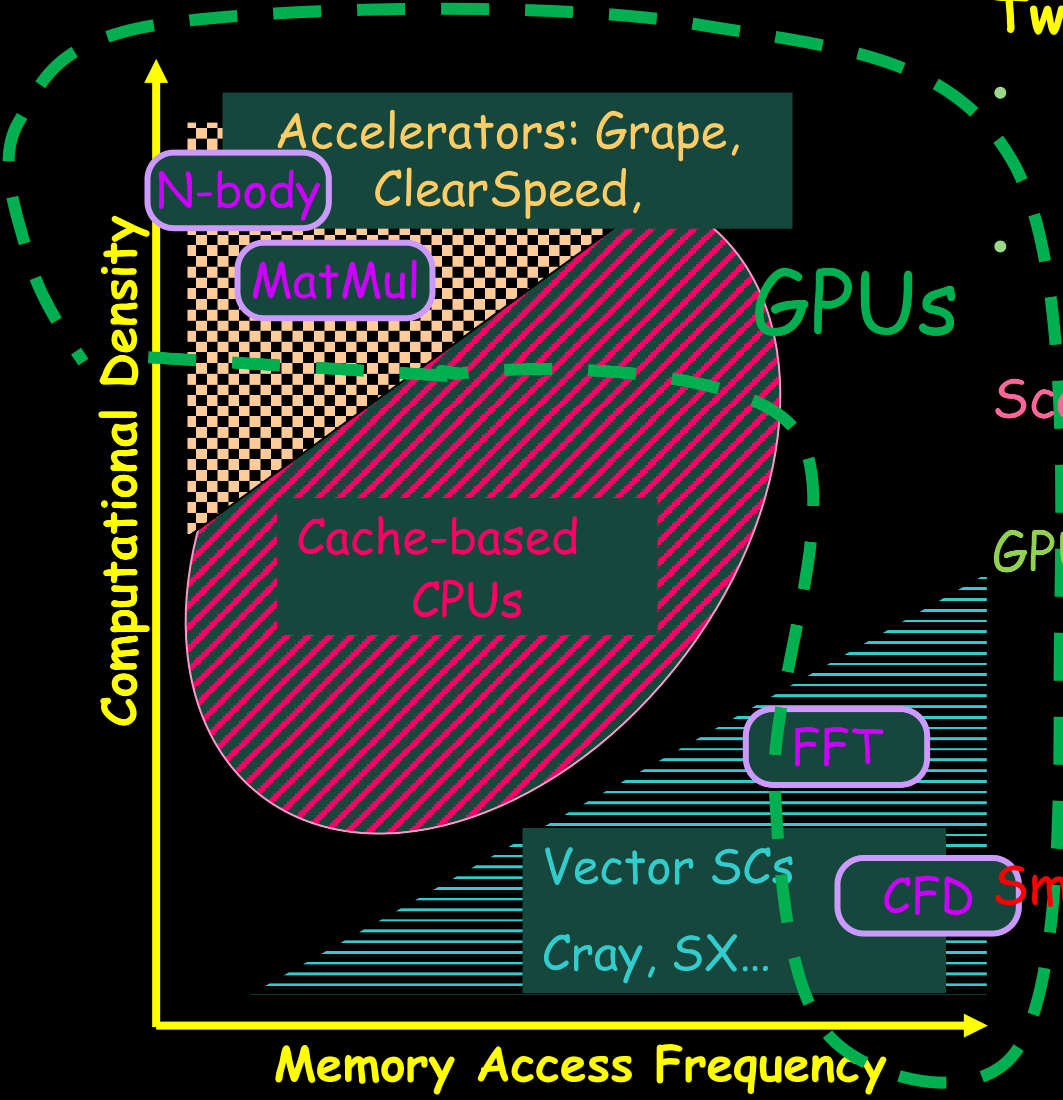
Tsubame 2.0 Experiences-Petascale Computing with GPUs Works

Satoshi Matsuoka
Tokyo Institute of Technology

GTC 2011, Beijing, China,
2011/12/14

GPUs as Modern-Day Vector Engines



Our Research@ Tokyo Tech

Two types of HPC Workloads

- High Computational Density
  => Traditional Accelerators
- High Memory Access Frequency
  => Traditional Vector SCs

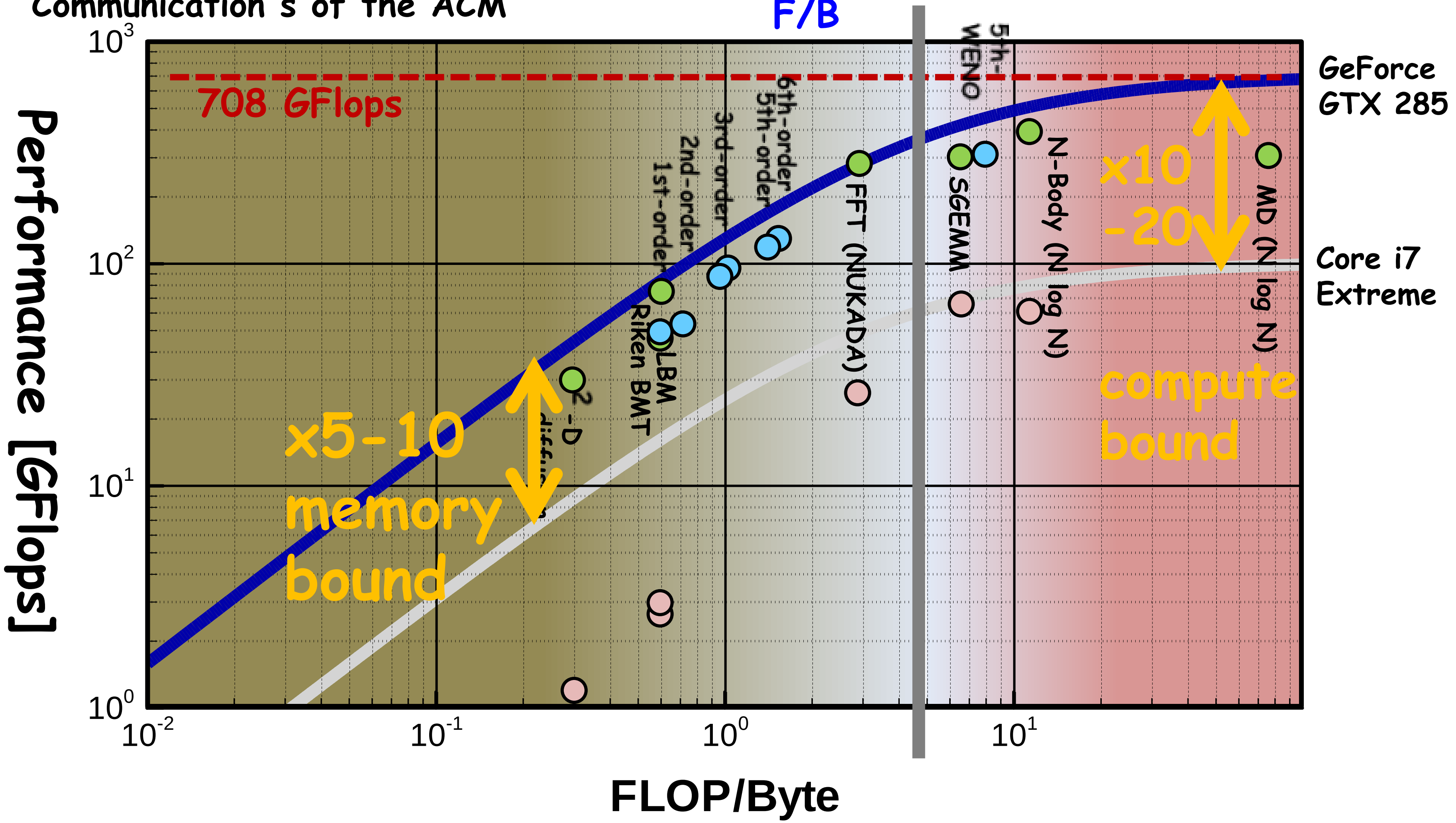
Scalar CPUs are so-so at both
=> Massive system for speedup

GPUs are both ***modern-day vector engines and high compute density accelerators***
=> Efficient Element of Next-Gen Supercomputers

Small memory, limited CPU-GPU BW, high-overhead communication with CPUs and other GPUs, lack of system-level SW support incl. fault tolerance, programming, …

GPU vs. CPU Performance

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



FLOP/Byte =
F/B

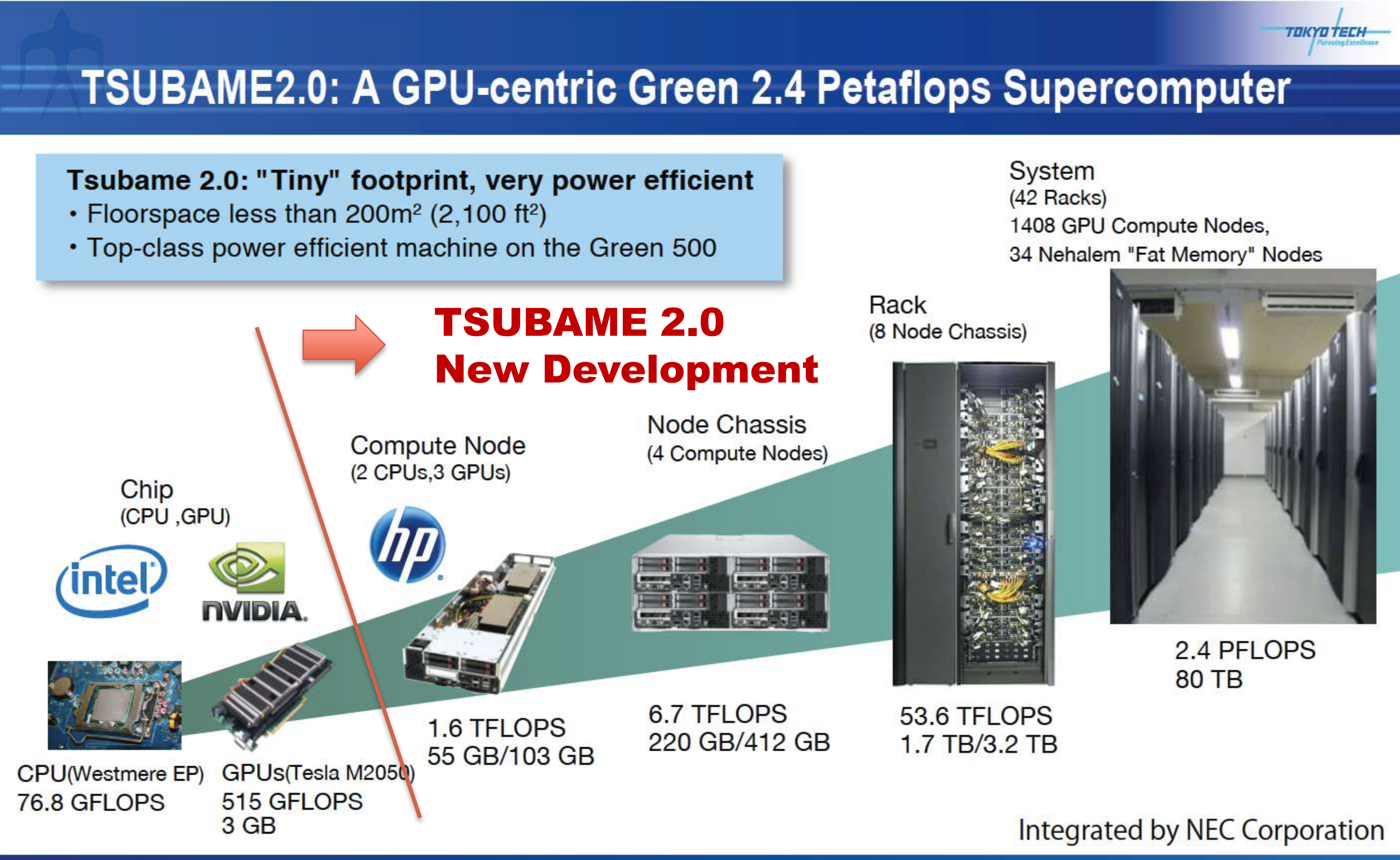
x5-10
memory
bound

x10
-20

compute
bound

TSUBAME2.0 Nov. 1, 2011

"The Greenest Production Supercomputer in the World"



TOKYO TECH
Pursuing Excellence

TSUBAME2.0: A GPU-centric Green 2.4 Petaflops Supercomputer

Tsubame 2.0: "Tiny" footprint, very power efficient

- Floorspace less than 200m² (2,100 ft²)
- Top-class power efficient machine on the Green 500

TSUBAME 2.0
New Development

Chip
(CPU ,GPU)

CPU(Westmere EP)
76.8 GFLOPS

GPUs(Tesla M2050)
515 GFLOPS
3 GB

Compute Node
(2 CPUs,3 GPUs)

1.6 TFLOPS
55 GB/103 GB

Node Chassis
(4 Compute Nodes)

6.7 TFLOPS
220 GB/412 GB

Rack
(8 Node Chassis)

53.6 TFLOPS
1.7 TB/3.2 TB

System
(42 Racks)
1408 GPU Compute Nodes,
34 Nehalem "Fat Memory" Nodes

2.4 PFLOPS
80 TB

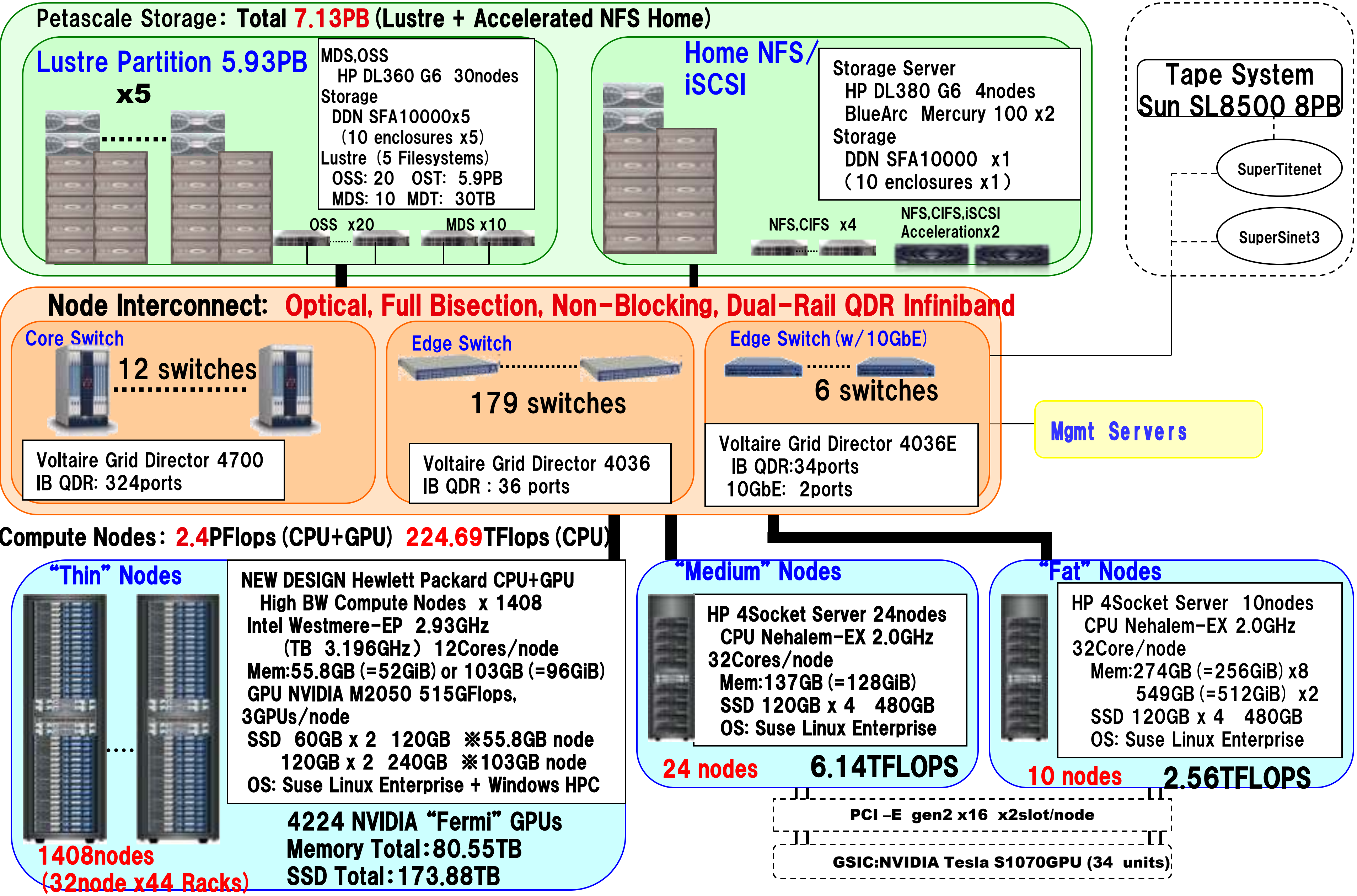
Integrated by NEC Corporation

Highlights of TSUBAME 2.0 Design (Oct. 2010) w/NEC-HP

- 2.4 PF Next gen multi-core x86 + next gen GPGPU
  - 1432 nodes, Intel Westmere/Nehalem EX
  - 4224 NVIDIA Tesla (Fermi) M2050 GPUs
  - ~100,000 total CPU and GPU "cores", High Bandwidth
  - **1.9 million "CUDA cores", 32K x 4K = 130 million CUDA threads(!)**

- 0.72 Petabyte/s aggregate mem BW,
  - Effective 0.3-0.5 Bytes/Flop, restrained memory capacity (100TB)
- Optical Dual-Rail IB-QDR BW, full bisection BW(Fat Tree)
  - 200Tbits/s, Likely fastest in the world, still scalable
- Flash/node, ~200TB (1PB in future), 660GB/s I/O BW
  - >7 PB IB attached HDDs, 15PB Total HFS incl. LTO tape
- Low power & efficient cooling, comparable to TSUBAME 1.0 (~1MW); PUE = 1.28 (60% better c.f. TSUBAME1)
- Virtualization and Dynamic Provisioning of Windows HPC + Linux, job migration, etc.

TSUBAME2.0 System Overview（2.4 Pflops/15PB）



Petascale Storage: Total 7.13PB (Lustre + Accelerated NFS Home)

Lustre Partition 5.93PB
x5

MDS,OSS
HP DL360 G6 30nodes
Storage
DDN SFA10000x5
（10 enclosures x5）
Lustre（5 Filesystems）
OSS: 20 OST: 5.9PB
MDS: 10 MDT: 30TB

OSS x20
MDS x10

Home NFS/iSCSI

Storage Server
HP DL380 G6 4nodes
BlueArc Mercury 100 x2
Storage
DDN SFA10000 x1
（10 enclosures x1）

NFS,CIFS x4
NFS,CIFS,iSCSI Accelerationx2

Tape System
Sun SL8500 8PB

SuperTitenet
SuperSinet3

Node Interconnect: Optical, Full Bisection, Non-Blocking, Dual-Rail QDR Infiniband

Core Switch
12 switches
Voltaire Grid Director 4700
IB QDR: 324ports

Edge Switch
179 switches
Voltaire Grid Director 4036
IB QDR : 36 ports

Edge Switch (w/10GbE)
6 switches
Voltaire Grid Director 4036E
IB QDR:34ports
10GbE: 2ports

Mgmt Servers

Compute Nodes: 2.4PFlops (CPU+GPU) 224.69TFlops (CPU)

"Thin" Nodes

NEW DESIGN Hewlett Packard CPU+GPU
High BW Compute Nodes x 1408
Intel Westmere-EP 2.93GHz
(TB 3.196GHz) 12Cores/node
Mem:55.8GB (=52GiB) or 103GB (=96GiB)
GPU NVIDIA M2050 515GFlops,
3GPUs/node
SSD 60GB x 2 120GB ※55.8GB node
120GB x 2 240GB ※103GB node
OS: Suse Linux Enterprise + Windows HPC

1408nodes
(32node x44 Racks)

4224 NVIDIA "Fermi" GPUs
Memory Total:80.55TB
SSD Total:173.88TB

"Medium" Nodes

HP 4Socket Server 24nodes
CPU Nehalem-EX 2.0GHz
32Cores/node
Mem:137GB (=128GiB)
SSD 120GB x 4 480GB
OS: Suse Linux Enterprise

24 nodes 6.14TFLOPS

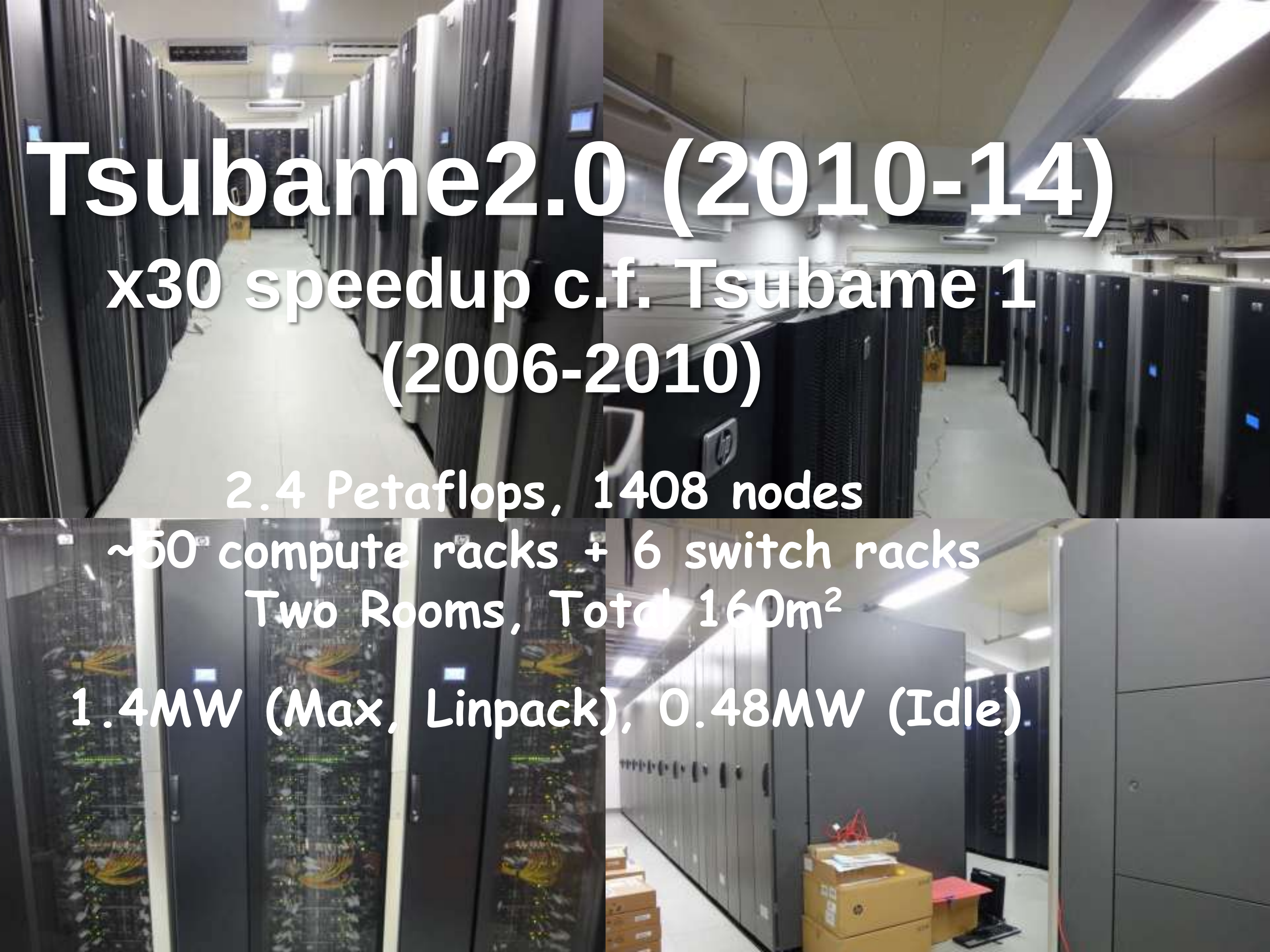
"Fat" Nodes

HP 4Socket Server 10nodes
CPU Nehalem-EX 2.0GHz
32Core/node
Mem:274GB (=256GiB) x8
549GB (=512GiB) x2
SSD 120GB x 4 480GB
OS: Suse Linux Enterprise

10 nodes 2.56TFLOPS

PCI –E gen2 x16 x2slot/node

GSIC:NVIDIA Tesla S1070GPU (34 units)



Tsubame2.0 (2010-14)

x30 speedup c.f. Tsubame 1 (2006-2010)

2.4 Petaflops, 1408 nodes

~50 compute racks + 6 switch racks

Two Rooms, Total 160m^2

1.4MW (Max, Linpack), 0.48MW (Idle)

TSUBAME2.0 Storage

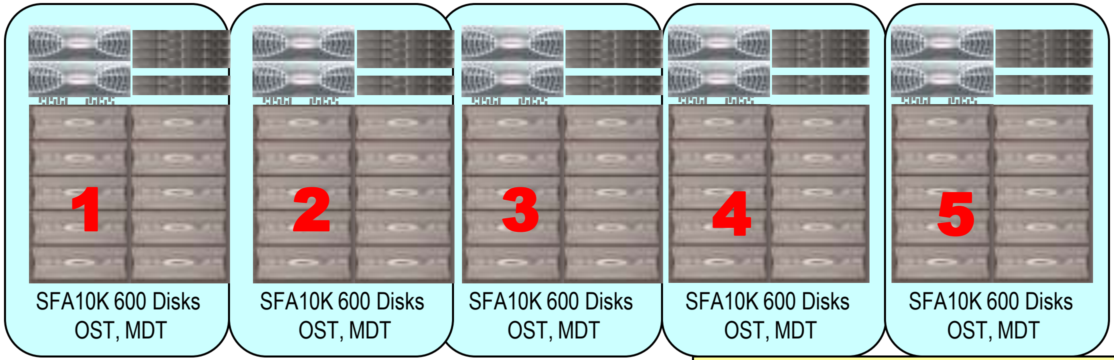
Multi-Petabyte storage consisting of Luster Parallel Filesystem Partition and NFS/CIFS/iSCSI Home Partition + Node SSD Acceleration

Lustre Parallel Filesystem Partition, 5.93PB

MDS:HP DL360 G6 x10
- CPU:Intel Westmere-EP x2 socket (12 Cores)
- Memory:51GB (=48GiB)
- IB HCA:IB 4X QDR PCI-e G2 x1port

OSS:HP DL360 G6 x20
- CPU:Intel Westmere-EP x2 socket (12 Cores)
- Memory:25GB (=24GiB)
- IB HCA：IB 4X QDR PCI-e G2 x2port

Storage:DDN SFA10000 x5
- Total Capacity:5.93PB
  2TB SATA x 2950 Disks + 600GB SAS x 50 Disks

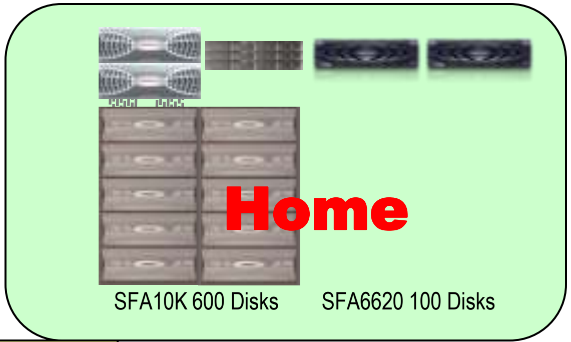


Home Partition 1.2PB

NFS/CIFS:HP DL380 G6 x4
- CPU:Intel Westmere-EP x2 socket (12 Cores)
- Memory:51GB (=48GiB)
- IB HCA:IB 4X QDR PCI-e G2 x2port

NFS/CIFS/iSCSI:BlueArc Mercury100 x2
- 10GbE x2

ストレージ:DDN SFA10000 x1
- Total Capacity:1.2PB
  2TB SATA x 600 Disks



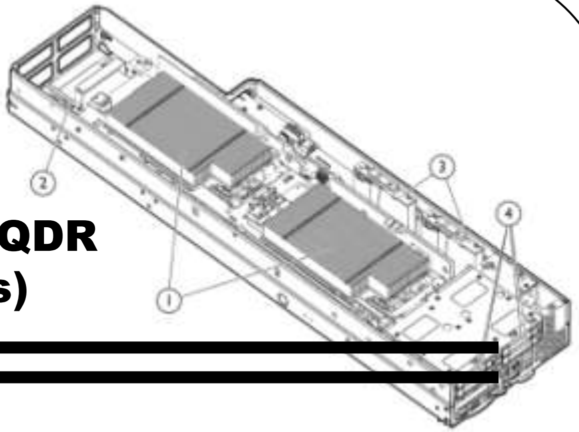
7.13PB HDD +200TB SSD + 8PB Tape

200 TB SSD (SLC, 1PB MLC future)
7.1 PB HDD (Highly redundunt)
4-8 PB Tape (HFS+Backup)
All Inifiniband+10GbE Connected

Lustre + GPFS
NFS, CIFS, WebDAV,...
Tivoli HFS
GridScaler + BlueArc

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

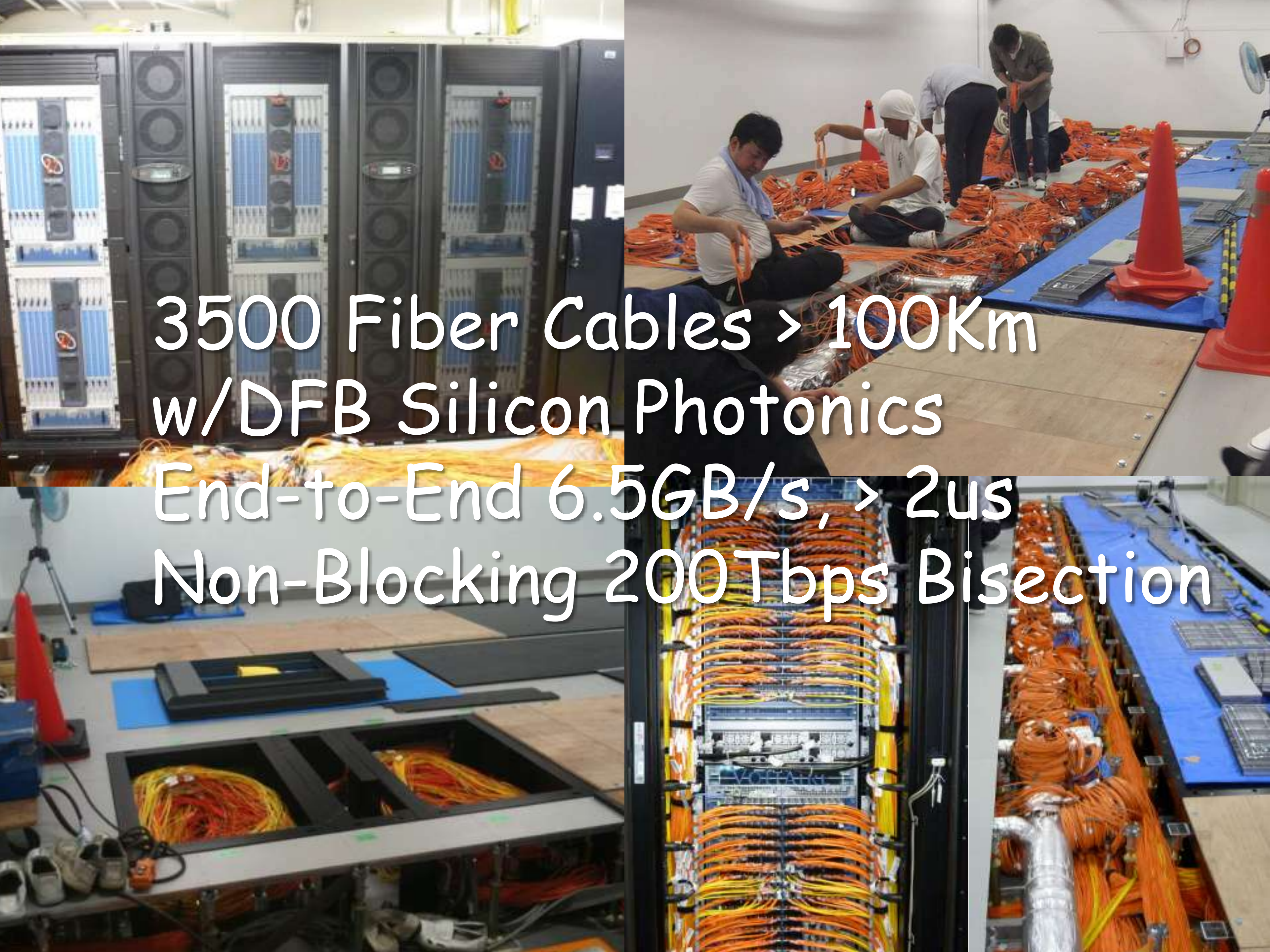
SSD: 16TB+

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

SL390 Compute Node

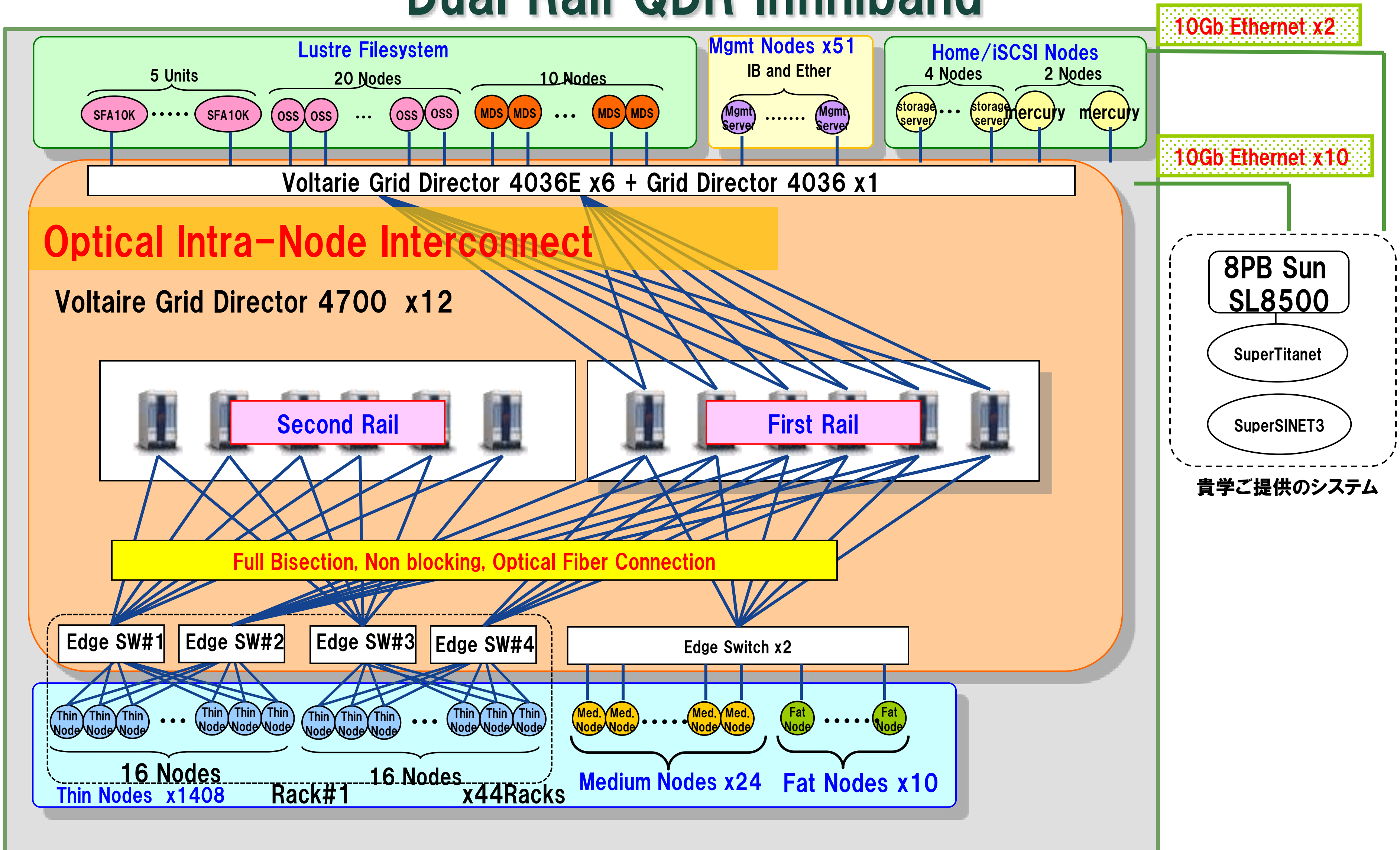
Collaborative Development w/HP

3 GPUs, 2CPUs, 50-100GB Mem
120-240GB SSD, QDR-IB x 2



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

TSUBAME 2.0 Full Bisection Fat Tree, Optical, Dual Rail QDR Infiniband



TSUBAME2.0ネットワーク全体図

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

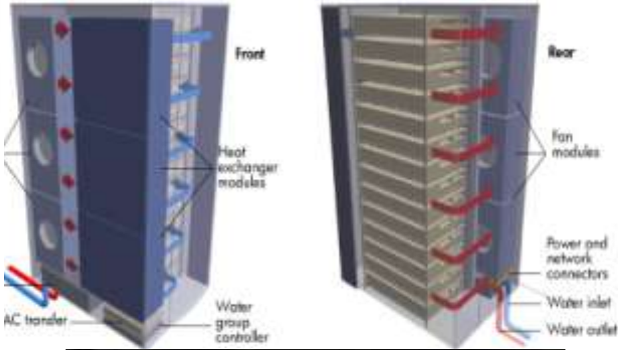
Automatic door opening mechanism controlling both racks

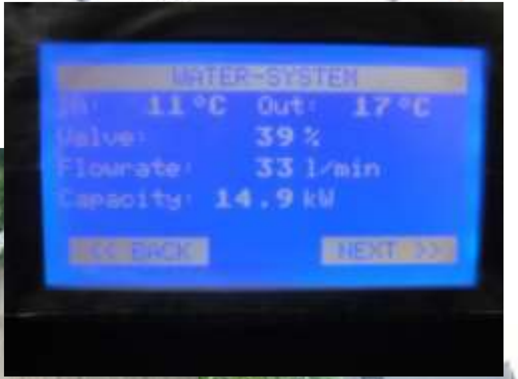
Adjustable temperature set point

Removes 95% to 97% of heat inside racks

Polycarbonate front door reduces ambient noise considerably

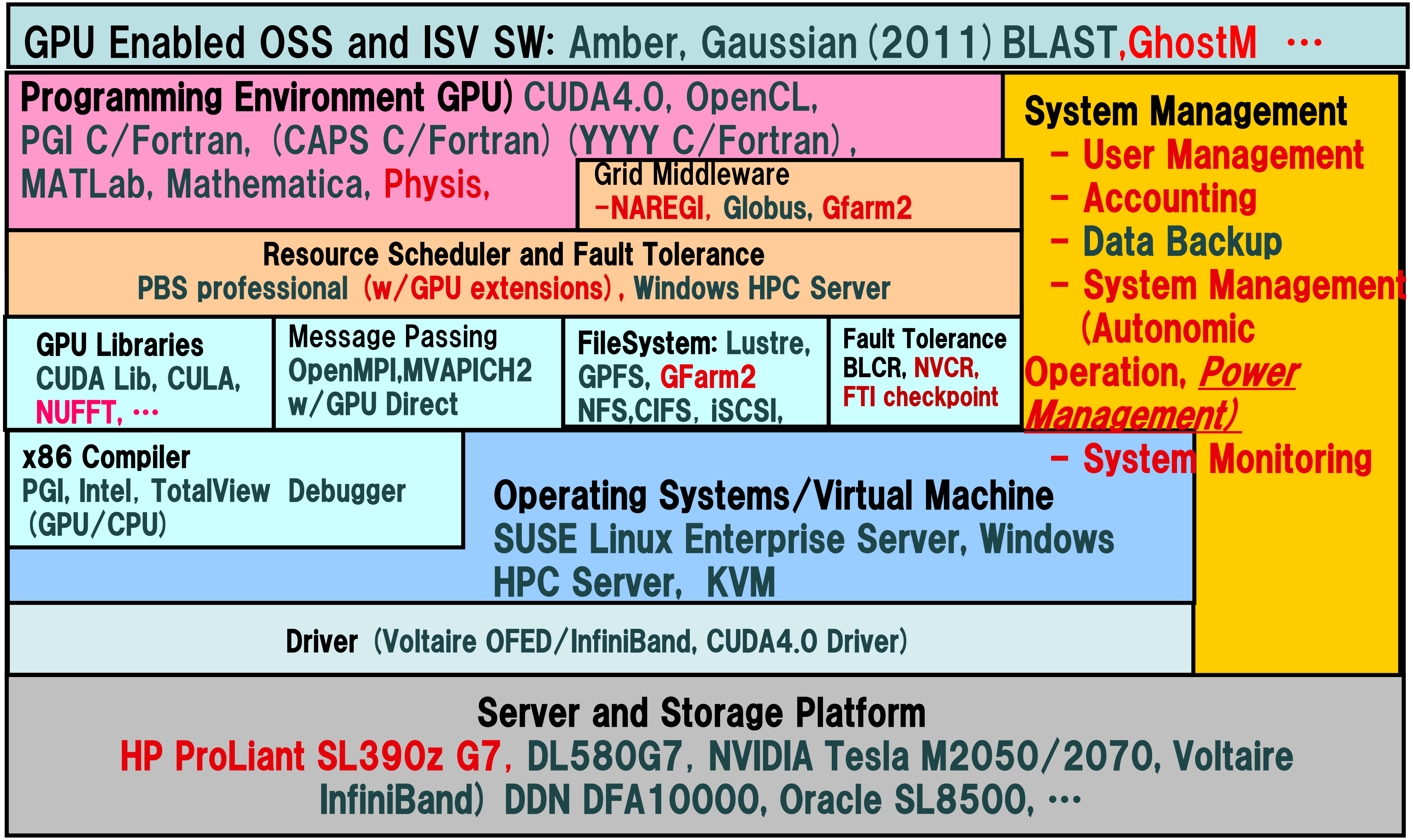
~= Entire Earth Simulator (rack = 50TF)





```
WATER-SYSTEM
In:      11 °C   Out:   17 °C
Valve:           39 %
Flowrate:        33 l/min
Capacity:  14.9 kW
```

TSUBAME2.0 Software Stack（Red: R&D at Tokyo Tech)



GPU Enabled OSS and ISV SW: Amber, Gaussian (2011) BLAST,GhostM …

Programming Environment GPU) CUDA4.0, OpenCL, PGI C/Fortran,（CAPS C/Fortran）(YYYY C/Fortran) , MATLab, Mathematica, Physis,

Grid Middleware
-NAREGI, Globus, Gfarm2

Resource Scheduler and Fault Tolerance
PBS professional（w/GPU extensions) , Windows HPC Server

GPU Libraries
CUDA Lib, CULA, NUFFT, …

Message Passing
OpenMPI,MVAPICH2 w/GPU Direct

FileSystem: Lustre, GPFS, GFarm2 NFS,CIFS, iSCSI,

Fault Tolerance
BLCR, NVCR, FTI checkpoint

x86 Compiler
PGI, Intel, TotalView Debugger (GPU/CPU)

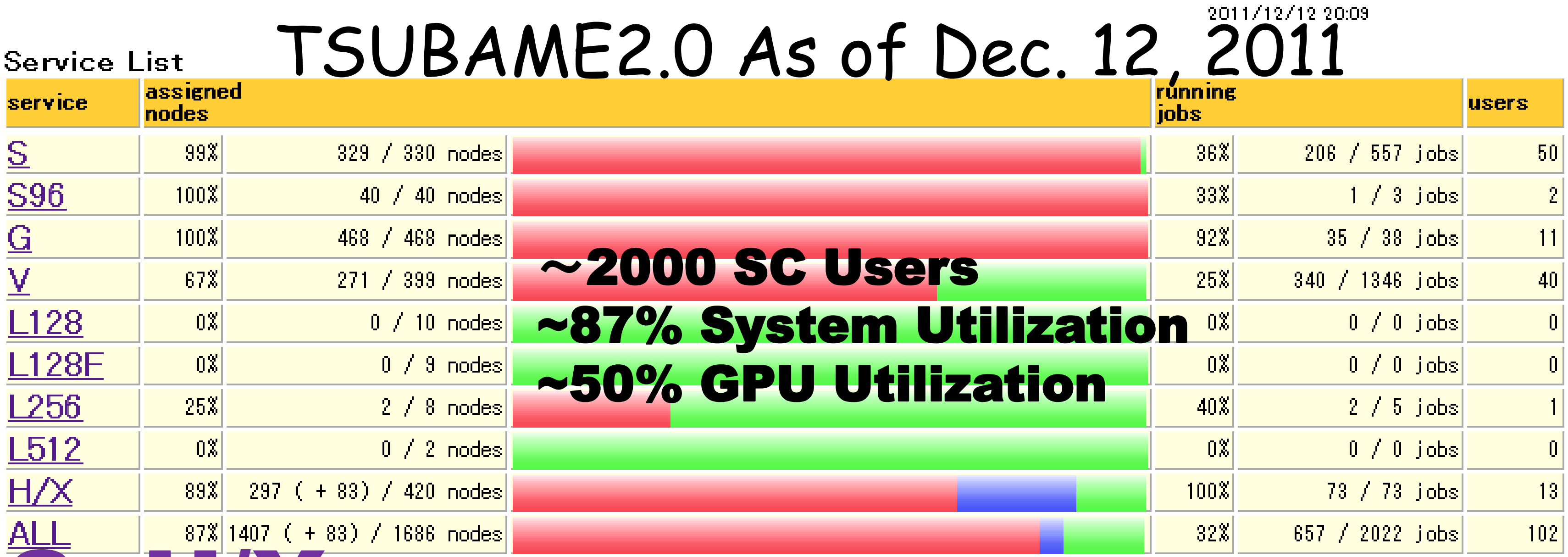
Operating Systems/Virtual Machine
SUSE Linux Enterprise Server, Windows HPC Server, KVM

System Management
- User Management
- Accounting
- Data Backup
- System Management (Autonomic Operation, *Power Management)*
- System Monitoring

Driver（Voltaire OFED/InfiniBand, CUDA4.0 Driver)

Server and Storage Platform
HP ProLiant SL390z G7, DL580G7, NVIDIA Tesla M2050/2070, Voltaire InfiniBand）DDN DFA10000, Oracle SL8500, …

TSUBAME 2.0 Cloud Service Utilization



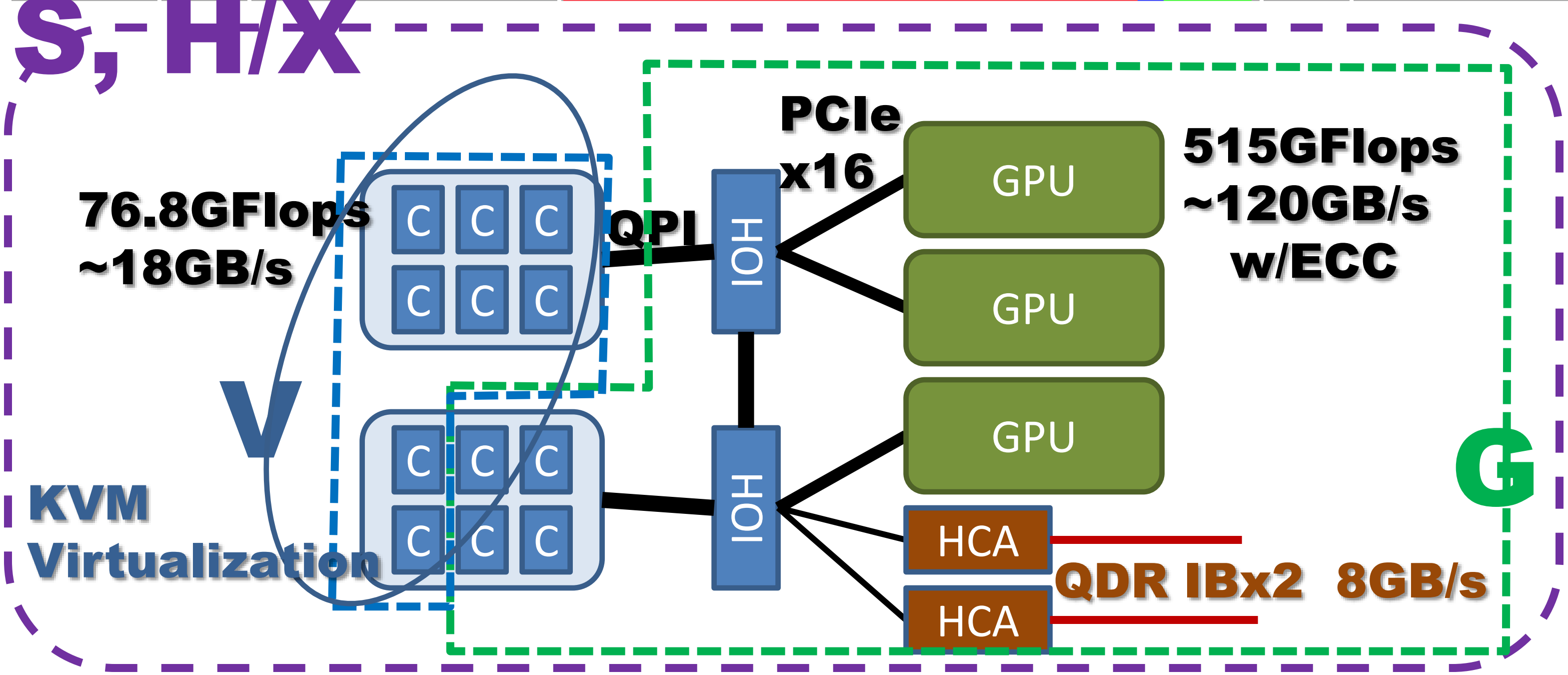
TSUBAME2.0 As of Dec. 12, 2011

2011/12/12 20:09

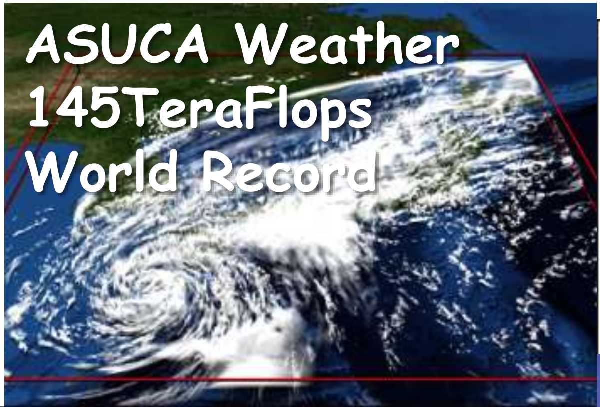
Service List

service	assigned nodes			running jobs		users
S	99%	329 / 330 nodes		36%	206 / 557 jobs	50
S96	100%	40 / 40 nodes		33%	1 / 3 jobs	2
G	100%	468 / 468 nodes		92%	35 / 38 jobs	11
V	67%	271 / 399 nodes		25%	340 / 1346 jobs	40
L128	0%	0 / 10 nodes		0%	0 / 0 jobs	0
L128F	0%	0 / 9 nodes		0%	0 / 0 jobs	0
L256	25%	2 / 8 nodes		40%	2 / 5 jobs	1
L512	0%	0 / 2 nodes		0%	0 / 0 jobs	0
H/X	89%	297 (+ 83) / 420 nodes		100%	73 / 73 jobs	13
ALL	87%	1407 (+ 83) / 1686 nodes		32%	657 / 2022 jobs	102

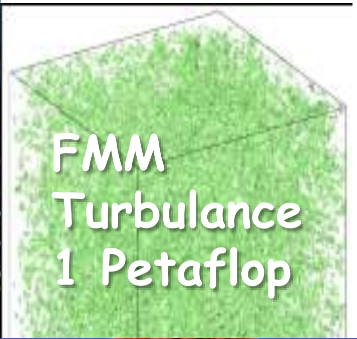
~2000 SC Users
~87% System Utilization
~50% GPU Utilization



Tsubame 2.0's Achievements



ASUCA Weather
145TeraFlops
World Record



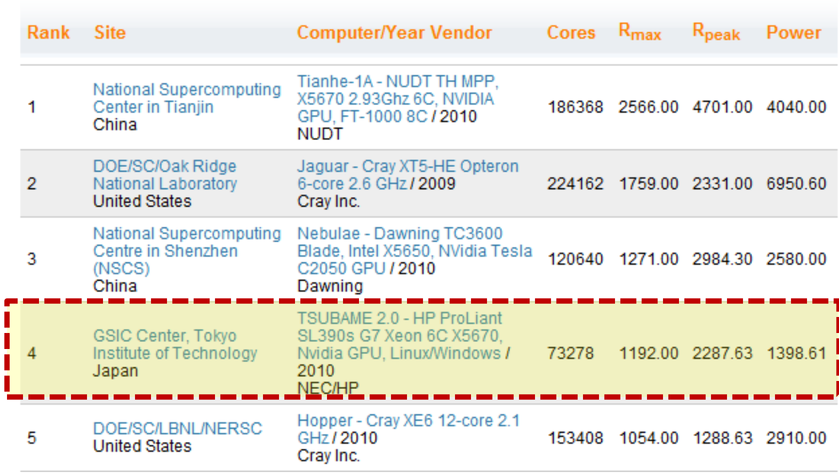
FMM
Turbulance
1 Petaflop

Dendrite
Crystallization
2.0 PetaFlops
2011 Gordon
Bell Award!!

Blood Flow 600
TeraFlops
2011 Gordon Bell
Award Honorable
Mention

Over 10 Petascale Applications

4th Fastest Supercomputer in the World (Nov. 2010 Top500)

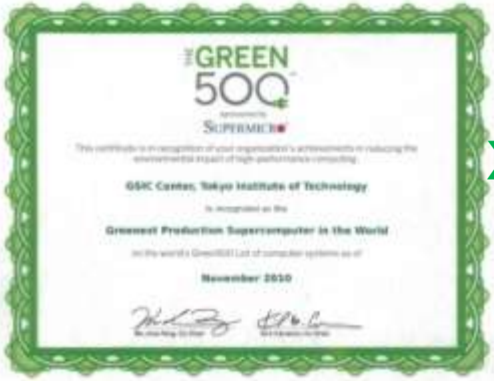


Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	National Supercomputing Center in Tianjin China	Tianhe-1A - NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C / 2010 NUDT	186368	2566.00	4701.00	4040.00
2	DOE/SC/Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.60
3	National Supercomputing Centre in Shenzhen (NSCS) China	Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU / 2010 Dawning	120640	1271.00	2984.30	2580.00
4	GSIC Center, Tokyo Institute of Technology Japan	TSUBAME 2.0 - HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows / 2010 NEC/HP	73278	1192.00	2287.63	1398.61
5	DOE/SC/LBNL/NERSC United States	Hopper - Cray XE6 12-core 2.1 GHz / 2010 Cray Inc.	153408	1054.00	1288.63	2910.00

Fruit of Years of Collaborative Research – Info-Plosion, JST CREST Ultra Low Power HPC…

World's Greenest Production Supercomputer

Nov. 2010, June 2011



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

x66,000 faster

x3 power efficient

>>

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)

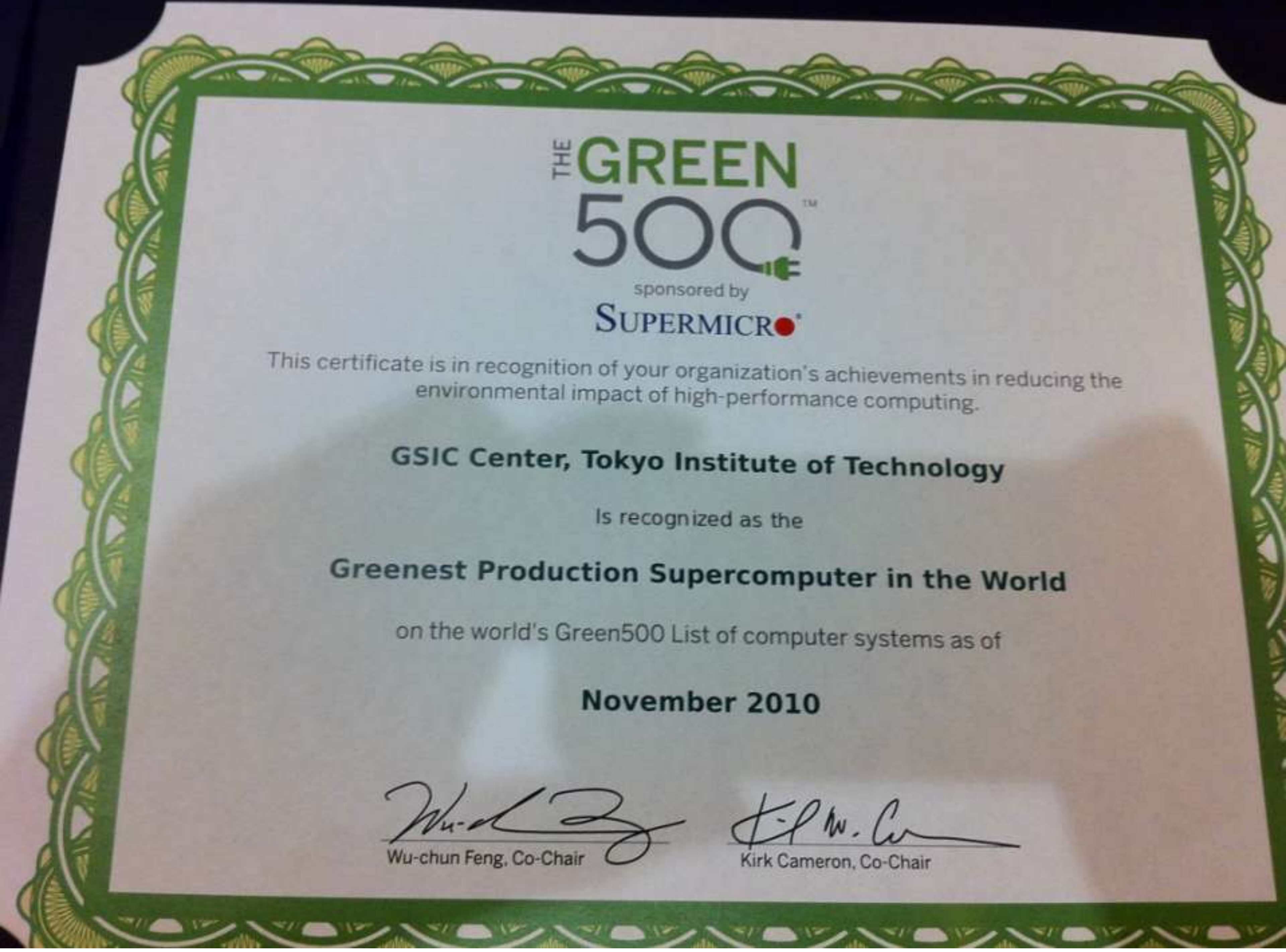
#3: 933.06 : US NCSA Hybrid Cluster Prototype **(403)**

#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

Example Grand Challenge, Petascale Applications on TSUBAME2.0 in 2011 (~10 apps)

PetaFLOPS Phase-Field Simulation (Aoki)

Metal dendritic solidification is simulated

Phase-field Model:

Time integration of Phase-field

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

M_ϕ Mobility
a Interface anisotropy
ΔS Entropy of fusion
ΔT Undercooling

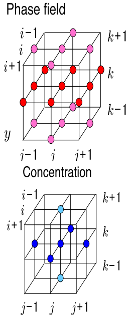
Time integration of solute concentration

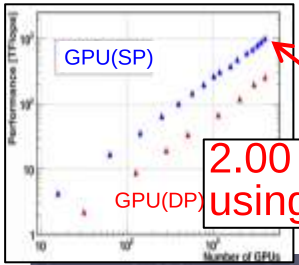
$$\frac{\partial c}{\partial t} = \nabla \cdot [D_S \phi \nabla c_S + D_L (1-\phi) \nabla c_L]$$

$$c_S = \frac{kc}{1-\phi+k\phi},\ c_L = \frac{c}{1-\phi+k\phi},\ k = c_S/c_L$$

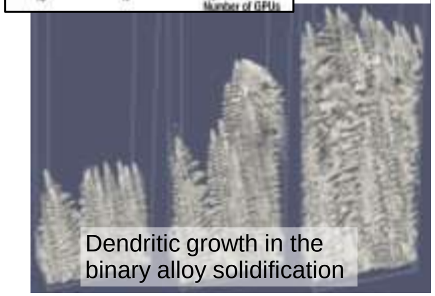
D_S D_L Diffusion coeff. in solid and liquid

Access pattern





2.00 PFlops in SP using 4,000 GPUs



Dendritic growth in the binary alloy solidification

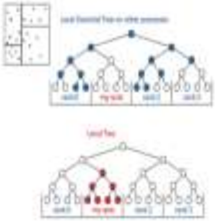
(Mesh:768 × 1632 × 3264)

SC11 Gordon Bell Winner + Tech Paper, fastest Stencil ever

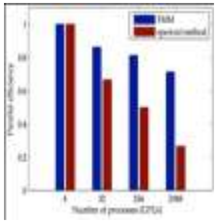
Turbulance Simulation using FMM (Yasuoka)

Vortex method with fast multipole method (FMM) is used

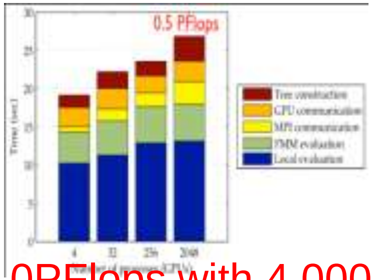
Q criteria in isotropic turbulence



Efficiency in weak scaling (4×10^6 particles per proc)



Breakdown of exec time



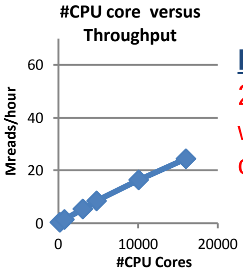
1.0PFlops with 4,000GPUs

Fastest FMM ever? Can hit 1PF w/ 4000 GPUs

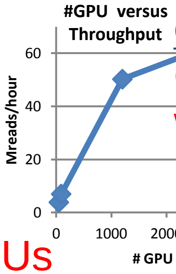
BLASTX for Millions of DNAseq (Akiyama)

Metagenome Analysis for Bacteria in Soil

Data: 224million DNA reads(75b) /set
Pre-filtering: reduces to 71million reads
BLASTX: 71million DNA vs. nr-aa DB (4.2G)



BLASTX
24.4 Million/hour with 16008 CPU cores



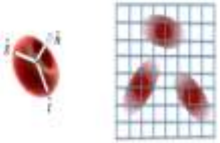
GHOSTM
60.6 Million/hour with 2,520 GPUs

GHOSTM is our original CUDA app almost compatible to BLASTX.

Can cope next gen Giga Sequencers

Multiphysics Biofluidics Simulation (Bernaschi)

Simulation of blood flows that accounts for from red blood cells to endothelial stress

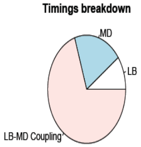


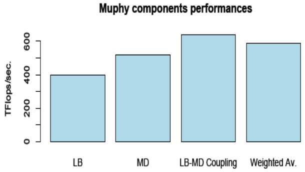
Red blood cells as ellipsoidal particles

Multiphyics simulation with *MUPHY* software

Fluid: blood plasma → Lattice Boltzmann(LB)

Body: RBC → Extended MD

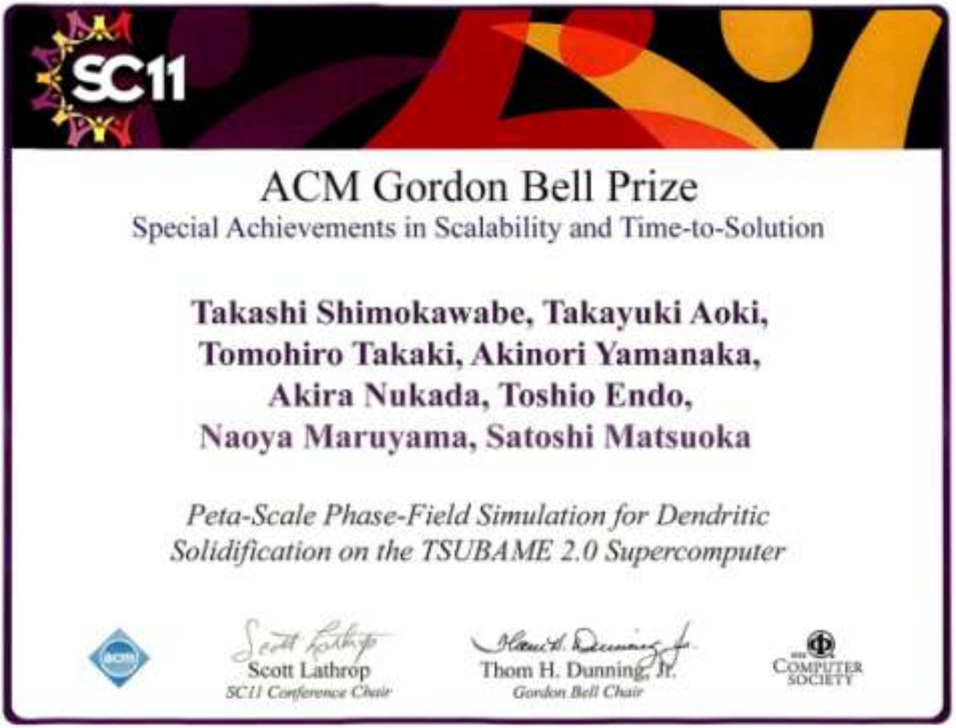




0.6PFlops with 4,000GPUs

450M RBCs are simulated

SC11 Gordon Bell Honorable Mention Fastest LBM ever?



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**

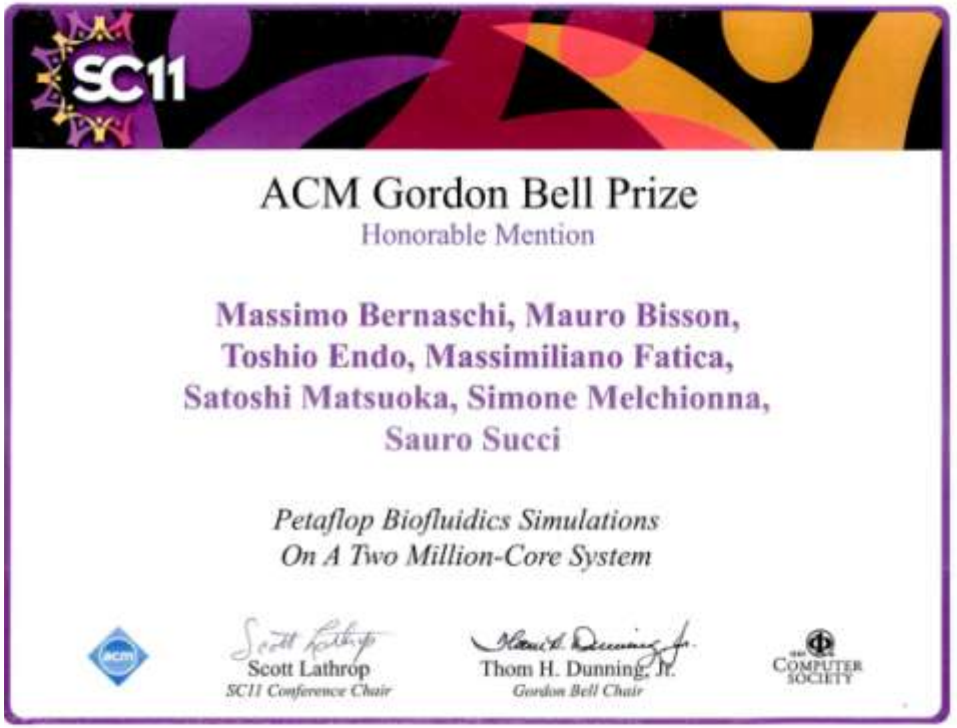
*Peta-Scale Phase-Field Simulation for Dendritic
Solidification on the TSUBAME 2.0 Supercomputer*

Scott Lathrop
SC11 Conference Chair

Thom H. Dunning, Jr.
Gordon Bell Chair

acm

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*

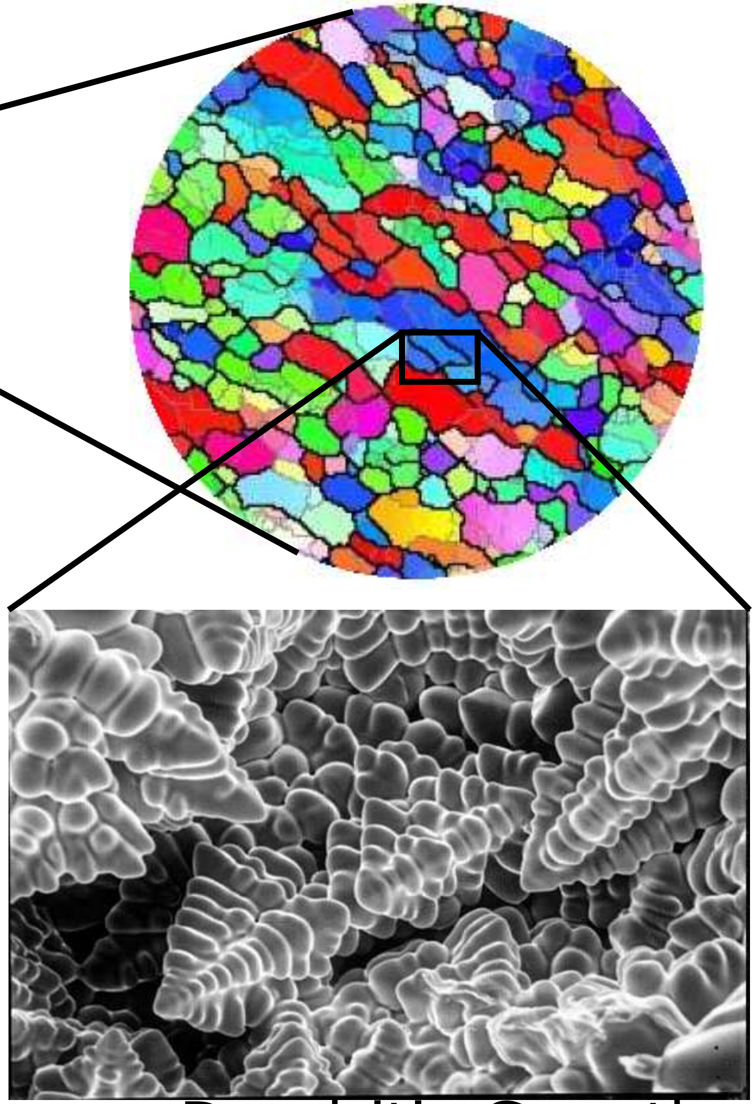
Scott Lathrop
SC11 Conference Chair

Thom H. Dunning, Jr.
Gordon Bell Chair

acm

COMPUTER SOCIETY

Background



Low-carbon society

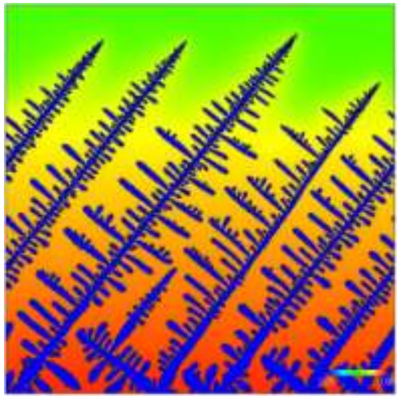
Improvement of fuel efficiency by reducing the weight of transportation and mechanical structures

Developing lightweight strengthening material by controlling microstructure

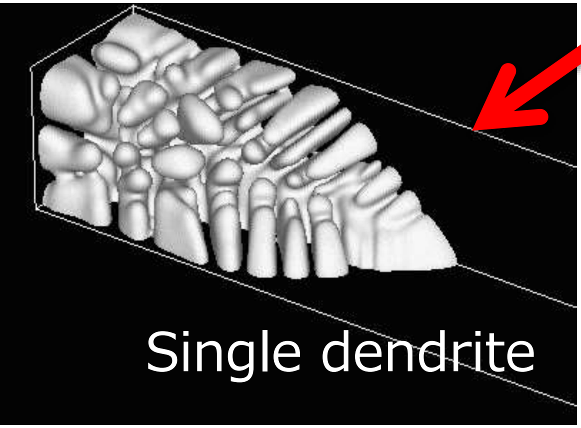
Impact of Peta-scale Simulation on Material Science

Previous Research

2D



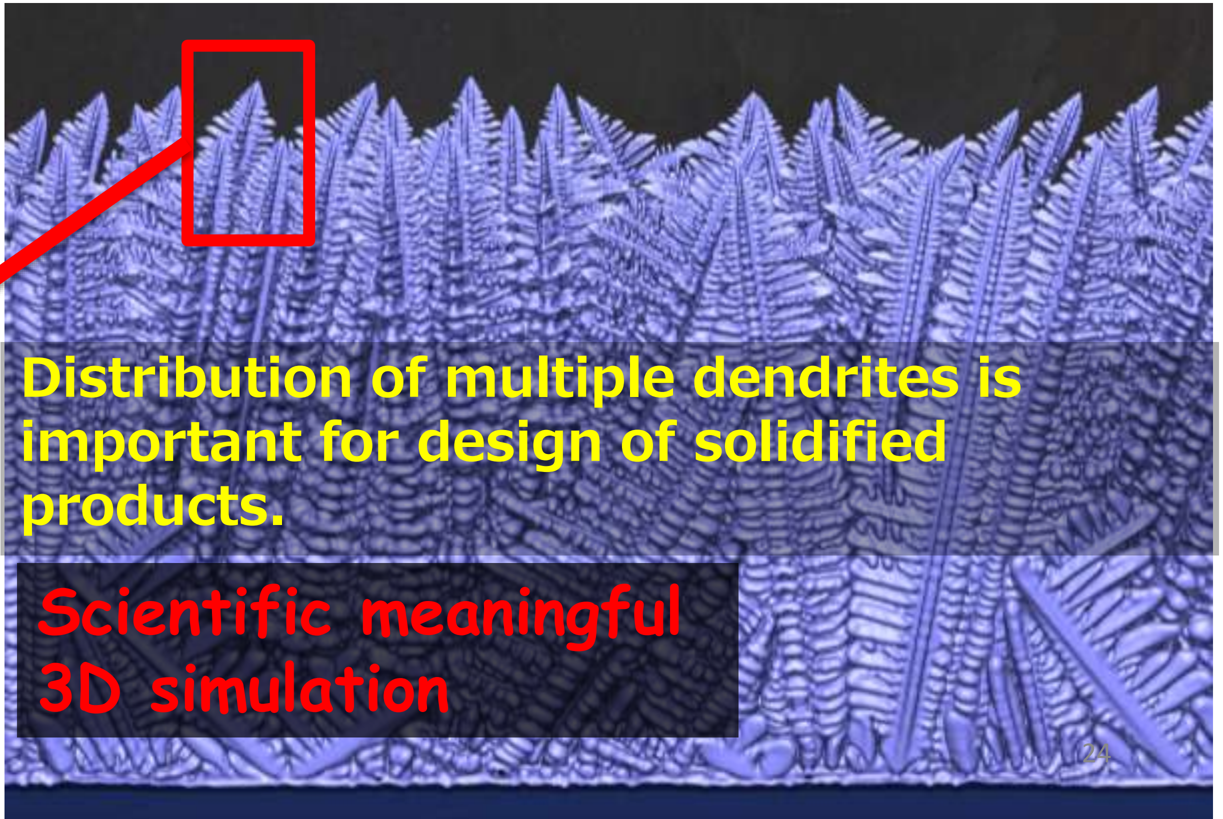
3D simple shape



Single dendrite

Peta-scale Simulation

- ✓ GPU-rich Supercomputer
- ✓ Optimization for Peta-scale computing



Distribution of multiple dendrites is important for design of solidified products.

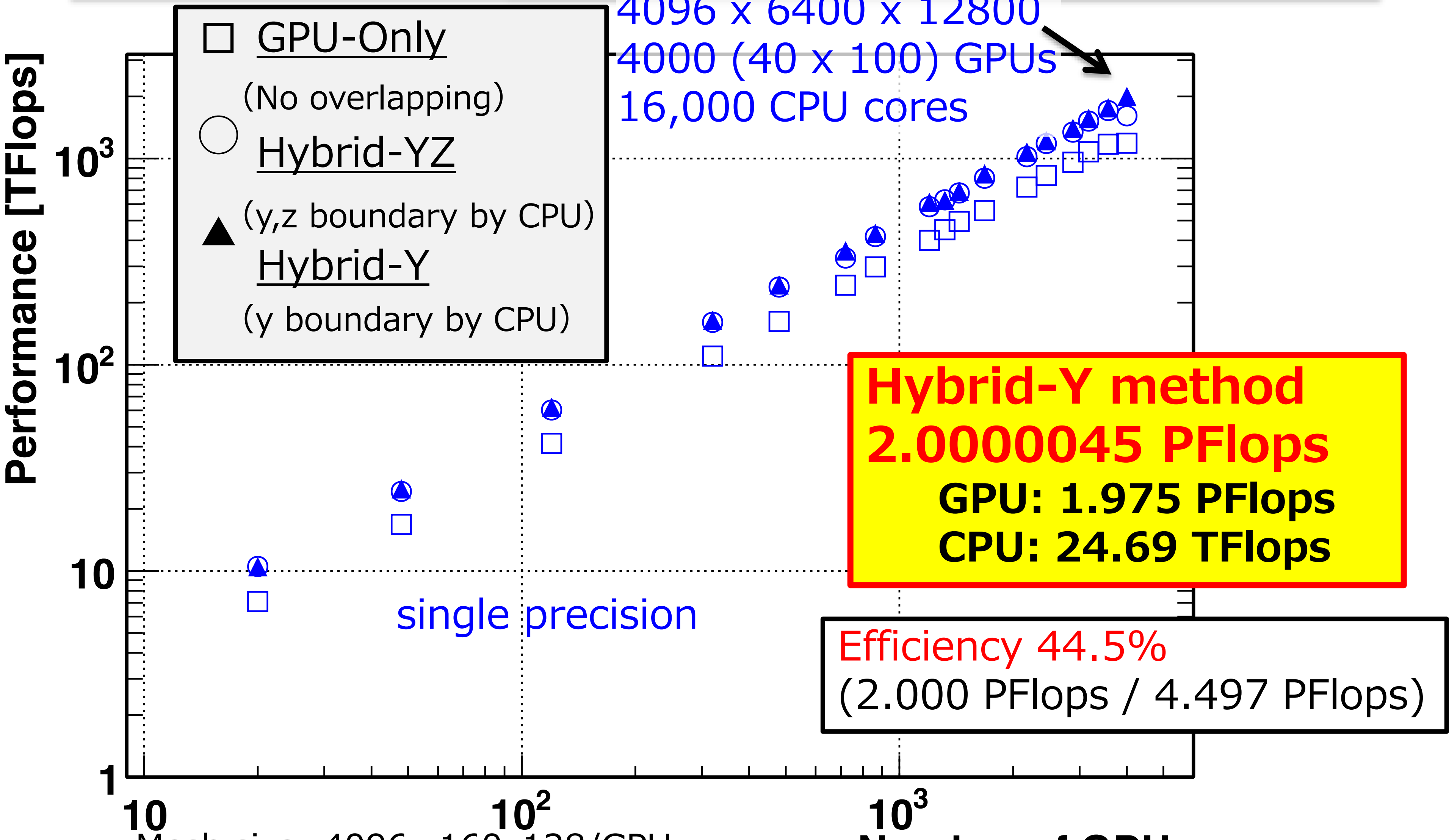
Scientific meaningful 3D simulation

Large-scale Phase-field Simulation

4096 x 1024 x 4096 (periodic boundary)

(Special thanks to Mr. Kuroki for 3D rendering.)

Weak scaling results on TSUBAME 2.0



4096 x 6400 x 12800
4000 (40 x 100) GPUs
16,000 CPU cores

single precision

Hybrid-Y method
2.0000045 PFlops
GPU: 1.975 PFlops
CPU: 24.69 TFlops

Efficiency 44.5%
(2.000 PFlops / 4.497 PFlops)

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

Power consumption and Efficiency

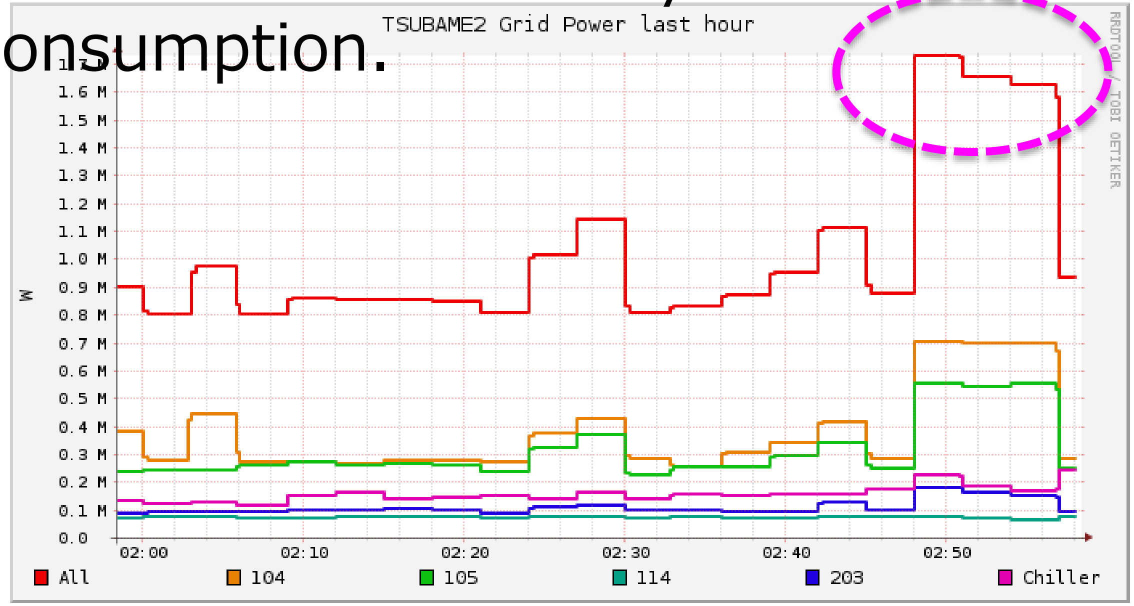
- The power consumption by application executions on TSUBAME 2.0 is measured in detail.
- Our phase-field simulation (real application)
  - ✓ 2.000 PFlops (single precision)
  - ✓ Performance to the peak: **44.5%**
  - ✓ **Green computing: 1468 MFlops/W**

2PFlops-Simulation

~1.36 MW

(Total:***1729kW***)

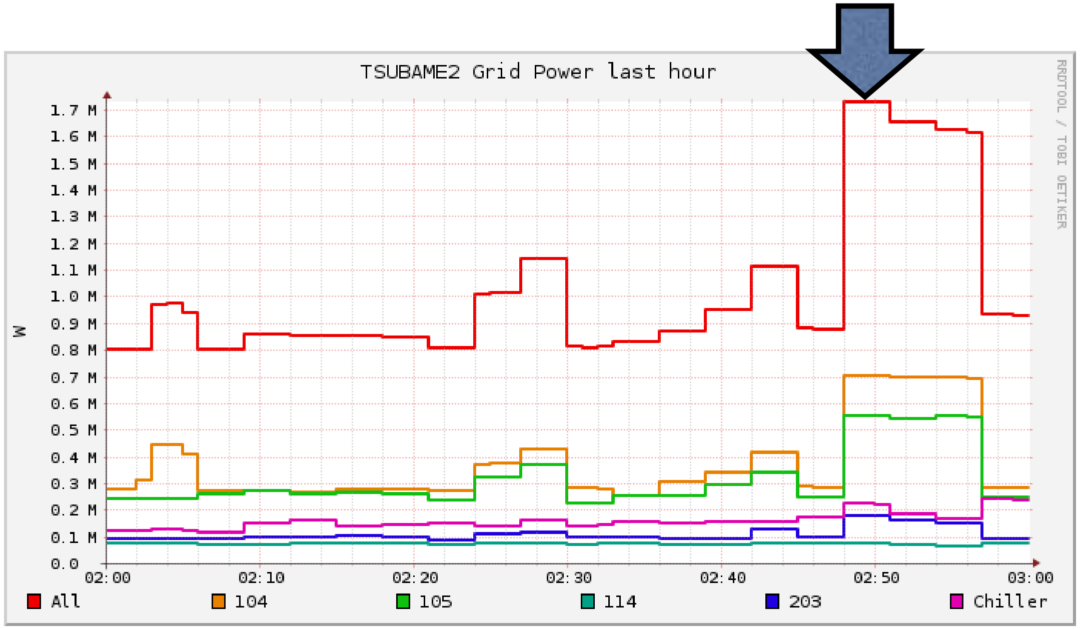
→ We obtained the simulation results by small electric power consumption.



Ref.
Linpack benchmark

- ✓ 1.192 PFlops (DP)
- ✓ Efficiency 52.1%
- ✓ 827.8 MFlops/W

Power Consumption during 2.0PFlops Phase-Field Run



2011/10/5 2:00—3:00

Compute node:

1362kW

Storage:

73kW

Cooling:

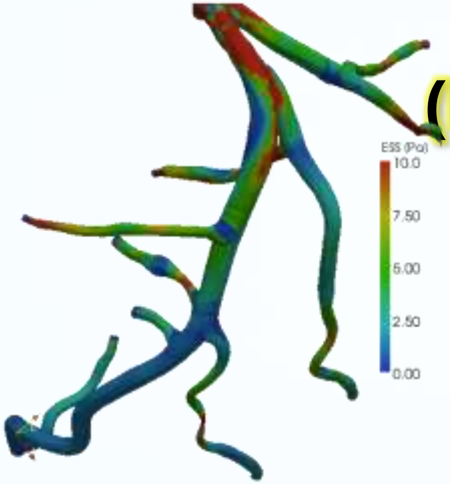
294kW at max

Total:

1729kW at max

MUPHY: Multiphysics simulation of blood flow

(Melchionna, Bernaschi et al.)



Multiphyics simulation with *MUPHY* software

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

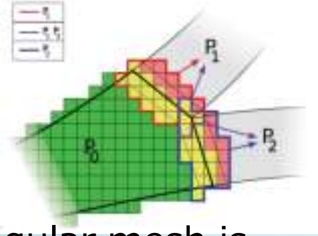
Realistic geometry (from CAT scan)

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

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

Fluid: Blood plasma

Lattice Boltzmann

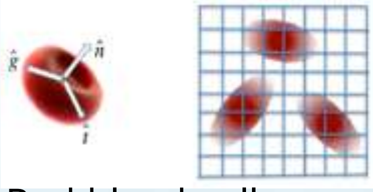


Irregular mesh is divided by using PT-SCOTCH tool, considering cutoff distance

coupled

Body: Red blood cell

Extended MD



Red blood cells (RBCs) are represented as ellipsoidal particles

CARDIOVASCULAR HEMODYNAMICS

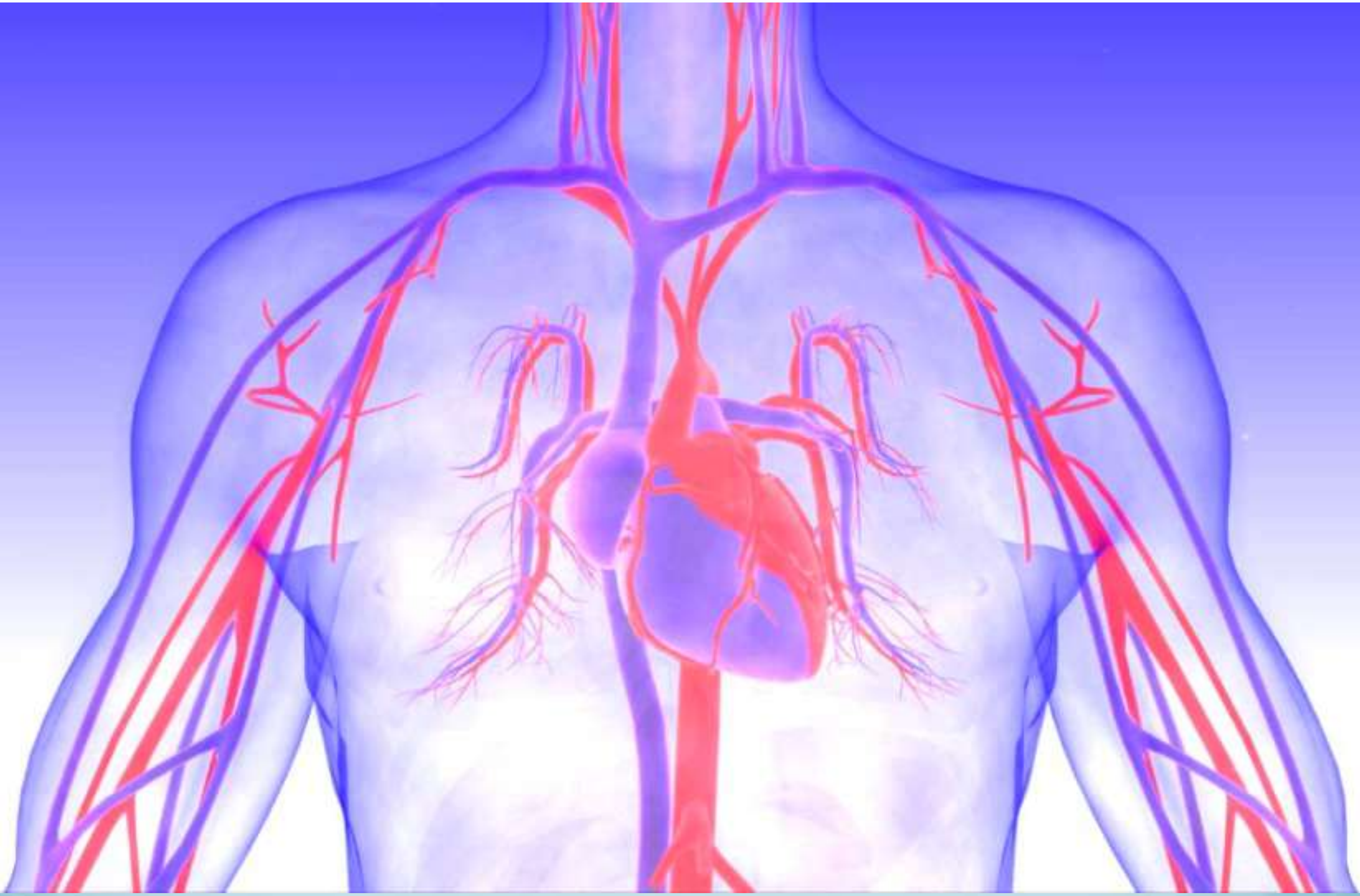
A topic with enormous impact on society

Plaque rupture is followed by flow interruption and leads to heart attack.

This is the first cause of mortality in western society.

It is essential to forecast where and when plaques form

The only possibility to access the patient-specific risk map (shear stress patterns) is through computing the complete arterial geometry!



Results on Tsubame2 Supercomputer (1)

Cluster of Nvidia M2050 GPUs connected by QDR Infiniband.
Scaling study up to 512 nodes (each node has 3 GPUs).
Very fast parallel I/O (read 100 GB in ~10 sec)

1 billion mesh nodes

GPUs	Time (s)	Efficiency
256	0.07616	N.A.
512	0.03852	98.86 %
1,024	0.01995	95.37 %
1,536	0.01343	94.43 %

Lattice Boltzmann Scaling (time per step)

LB kernel:1 GPU ~200 BG/P cores
1536 GPUs equivalent to full BlueGene/P

1 billion mesh nodes + 100 million RBC

GPUs	Time (s)	Efficiency
256	0.44453	N.A.
512	0.25601	86.82%
1,024	0.14062	79.03%

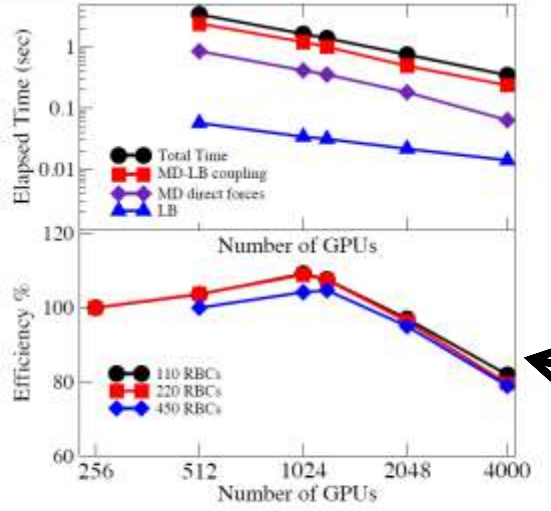
Lattice Boltzmann + Cell Dynamics Scaling (time per step)

Time to completion on stationary flow: 23 minutes

New run on FULL TSUBAME2.0 (4000 GPUs) just completed with an improved algorithm, exhibiting petascale performance(!)

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

Strong Scaling Results

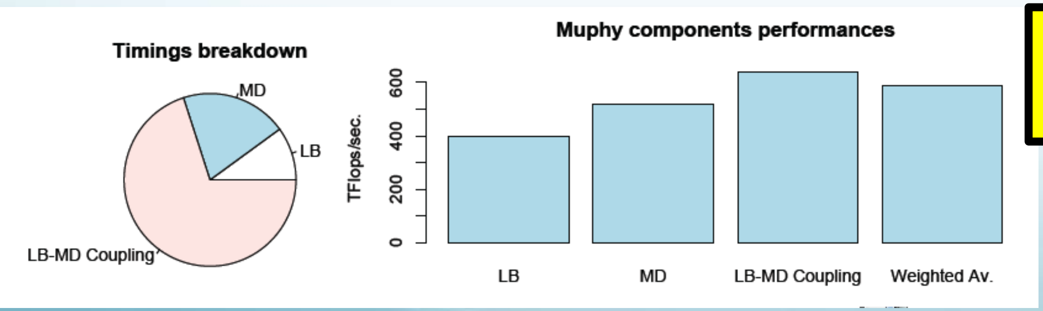


Elapsed time per timestep for 1G mesh nodes and 450M RBCs (log scale)

Parallel efficiency for 110, 220, 450M RBCs

~80% with 4K GPUs

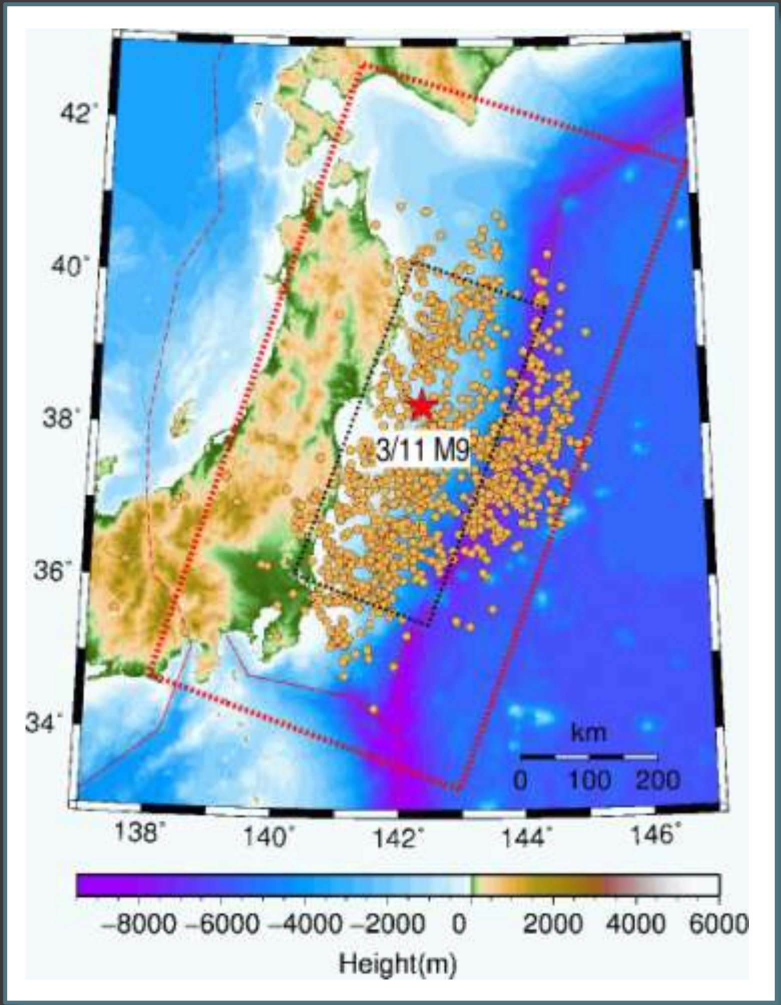
Speeds per Component



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

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

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

FDTD Simulation of Wave Propagation

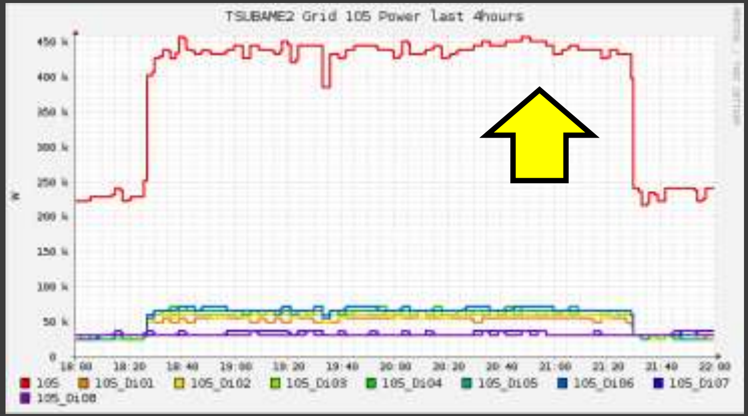
Main part of the FDM region

TARO OKAMOTO
TOKYO INSTITUTE OF TECHNOLOGY

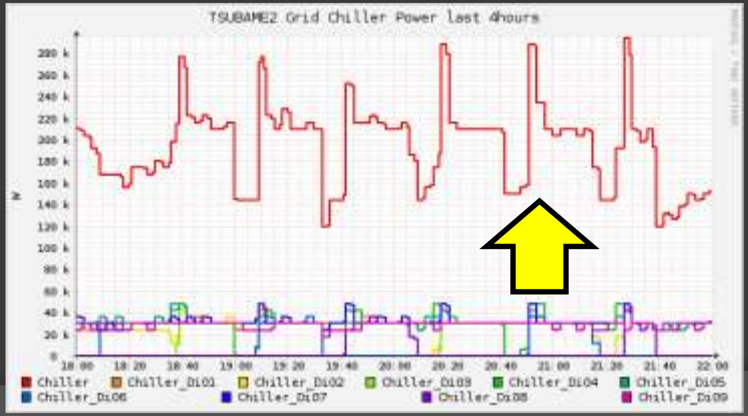
2011 MAGNITUDE 9 TOHOKU-OKI EARTHQUAKE

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

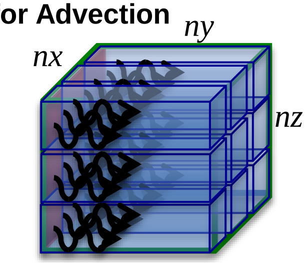
⇒ 20 x boost in microphysics (1.2 - 1.3 x overall improvement)

ASUCA : full GPU Implementation

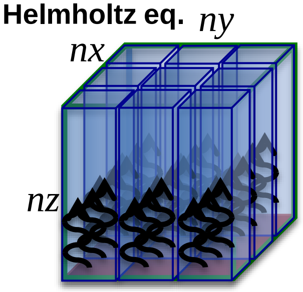
developed by Japan Meteorological Agency

TSUBAME 2.0 : 145 Tflops World Record !!!

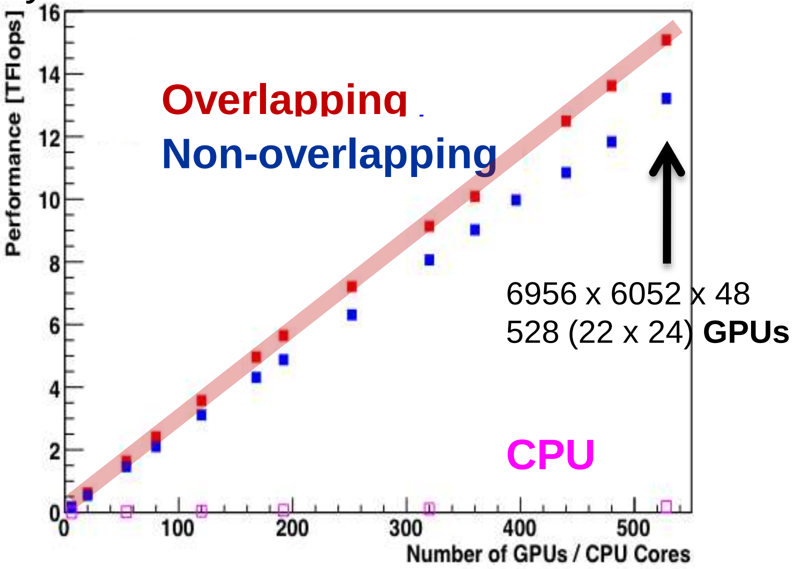
Block Division for Advection

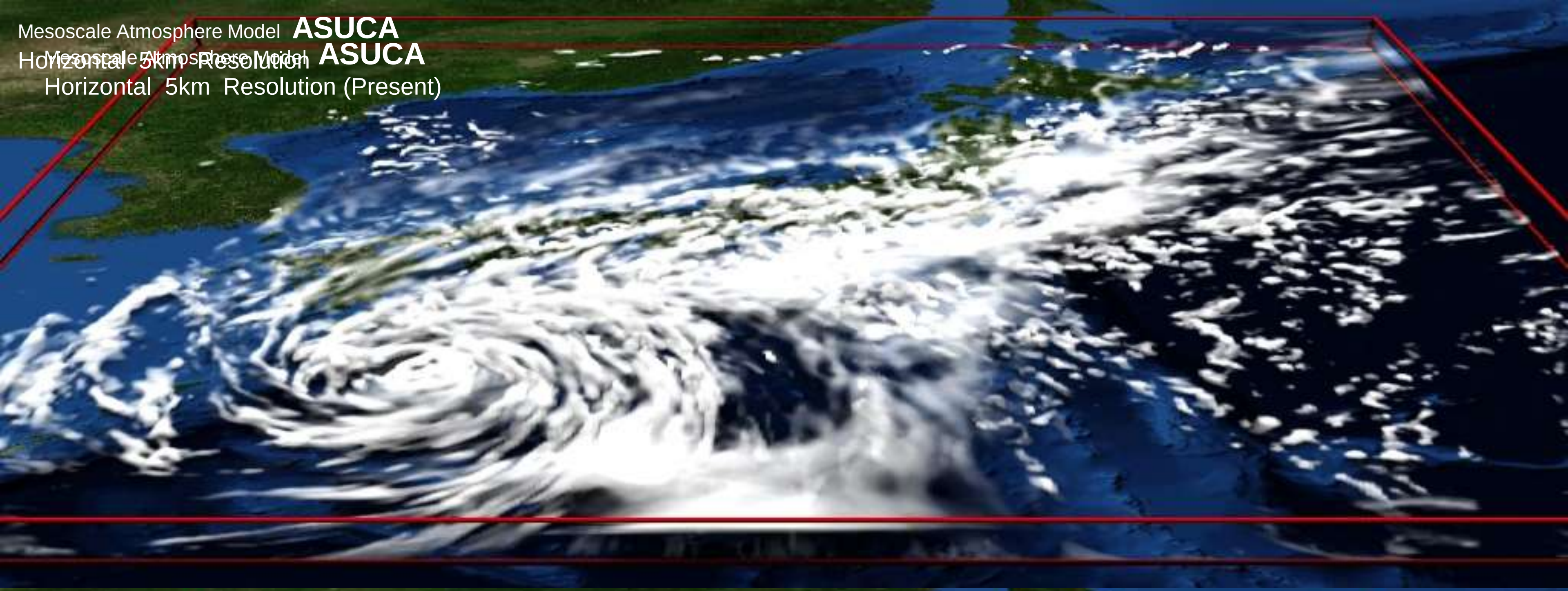


for 1-D Helmholtz eq.



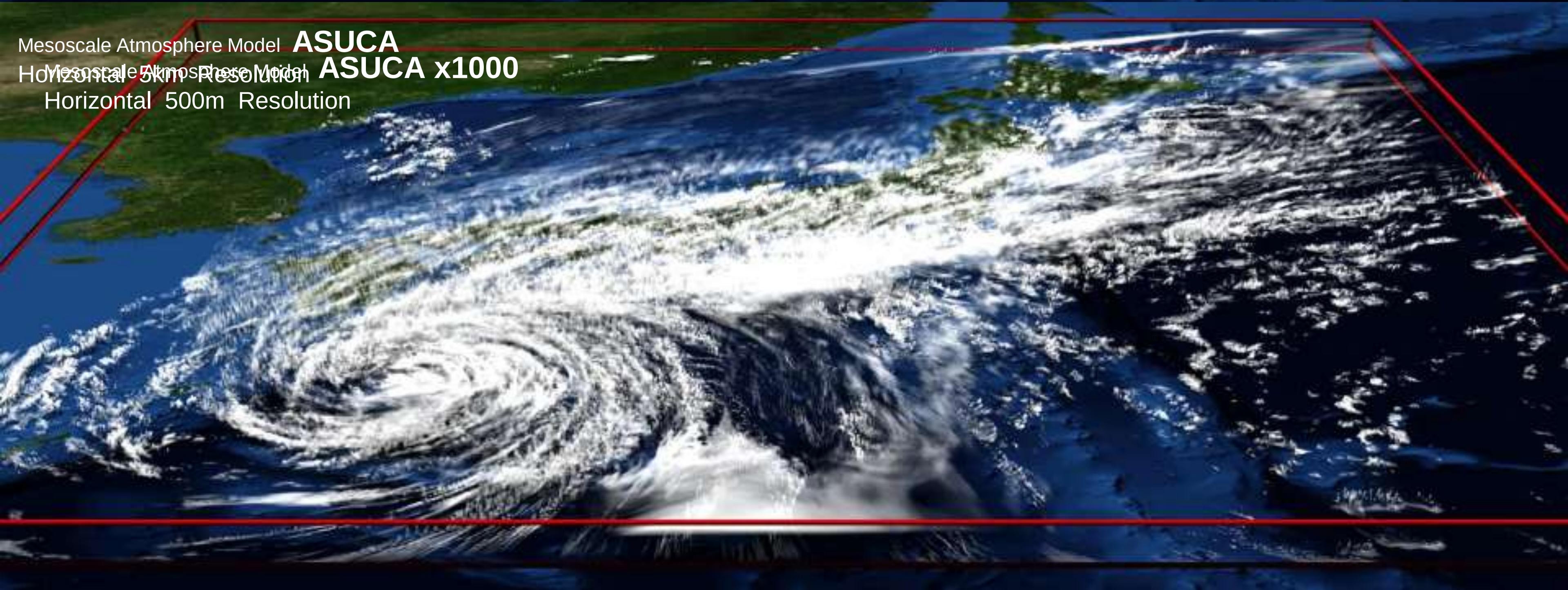
Typhoon





Mesoscale Atmosphere Model **ASUCA**
Horizontal 5km Resolution

Mesoscale Atmosphere Model **ASUCA**
Horizontal 5km Resolution (Present)

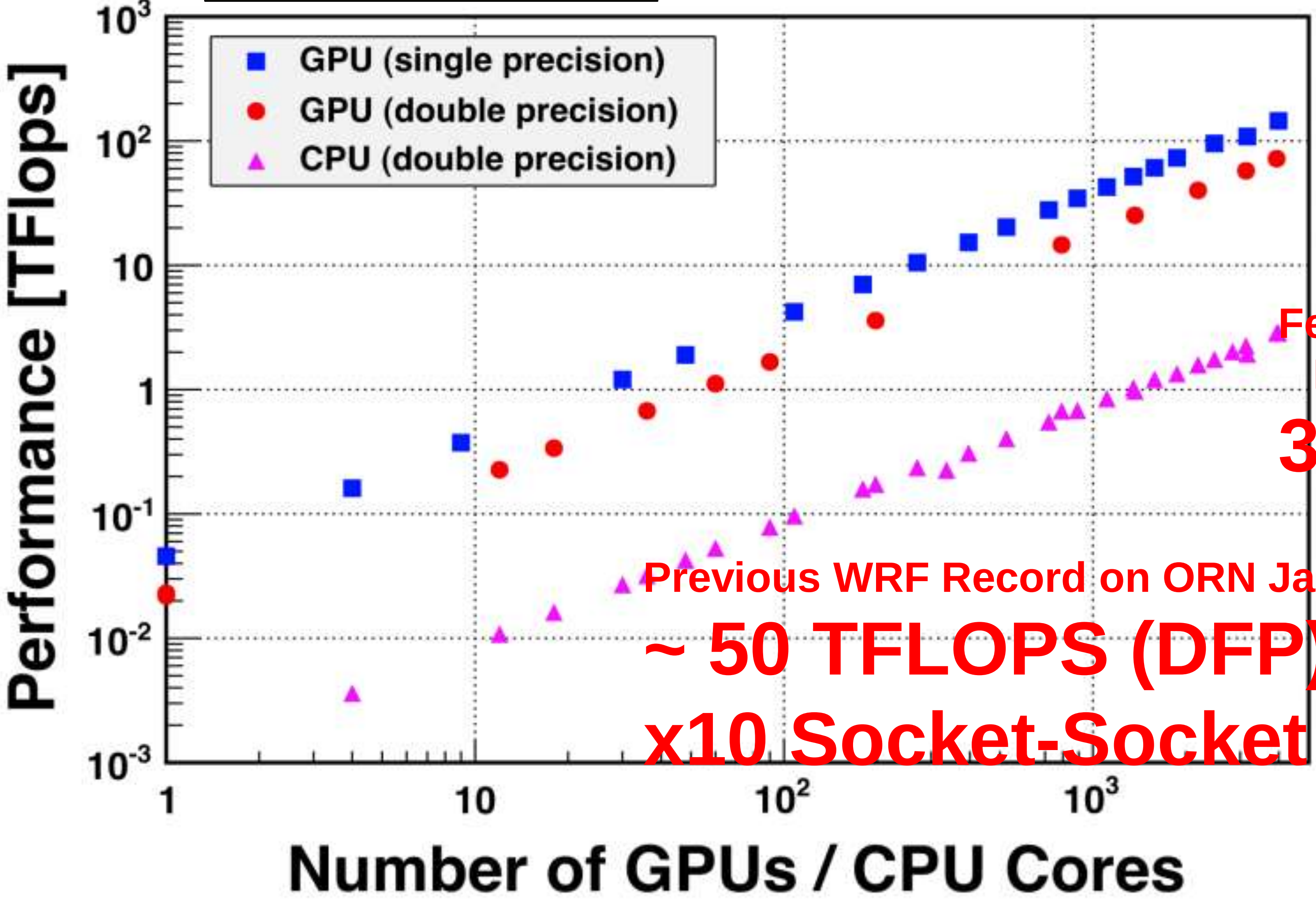


Mesoscale Atmosphere Model **ASUCA**
Horizontal 5km Resolution

Mesoscale Atmosphere Model **ASUCA x1000**
Horizontal 500m Resolution

TSUBAME 2.0 Performance

Weak Scaling



145.0 Tflops
Single precision

76.1 Tflops
Doublele precision

Fermi core Tesla
M2050

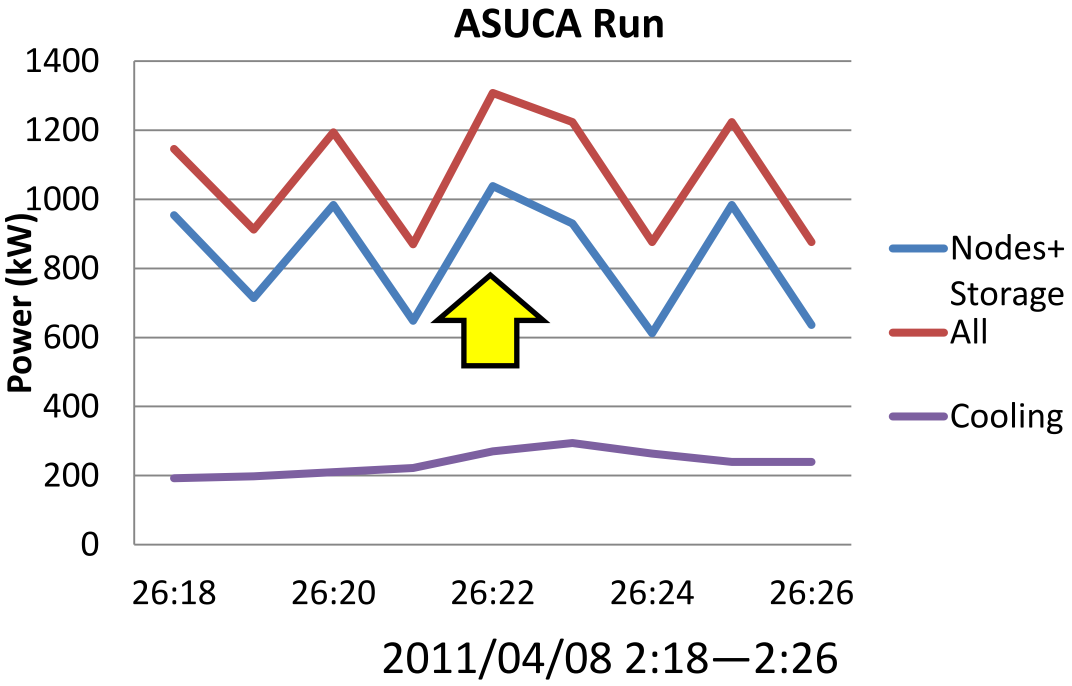
3990 GPUs

Previous WRF Record on ORN Jaguar

~ 50 TFLOPS (DFP)

x10 Socket-Socket

Power Consumption during Full TSUBAME2 Test with ASUCA



2011/04/08 2:18—2:26

Compute node:

960kW

Storage:

78kW

Cooling:

270kW max

Total:

1308kW max

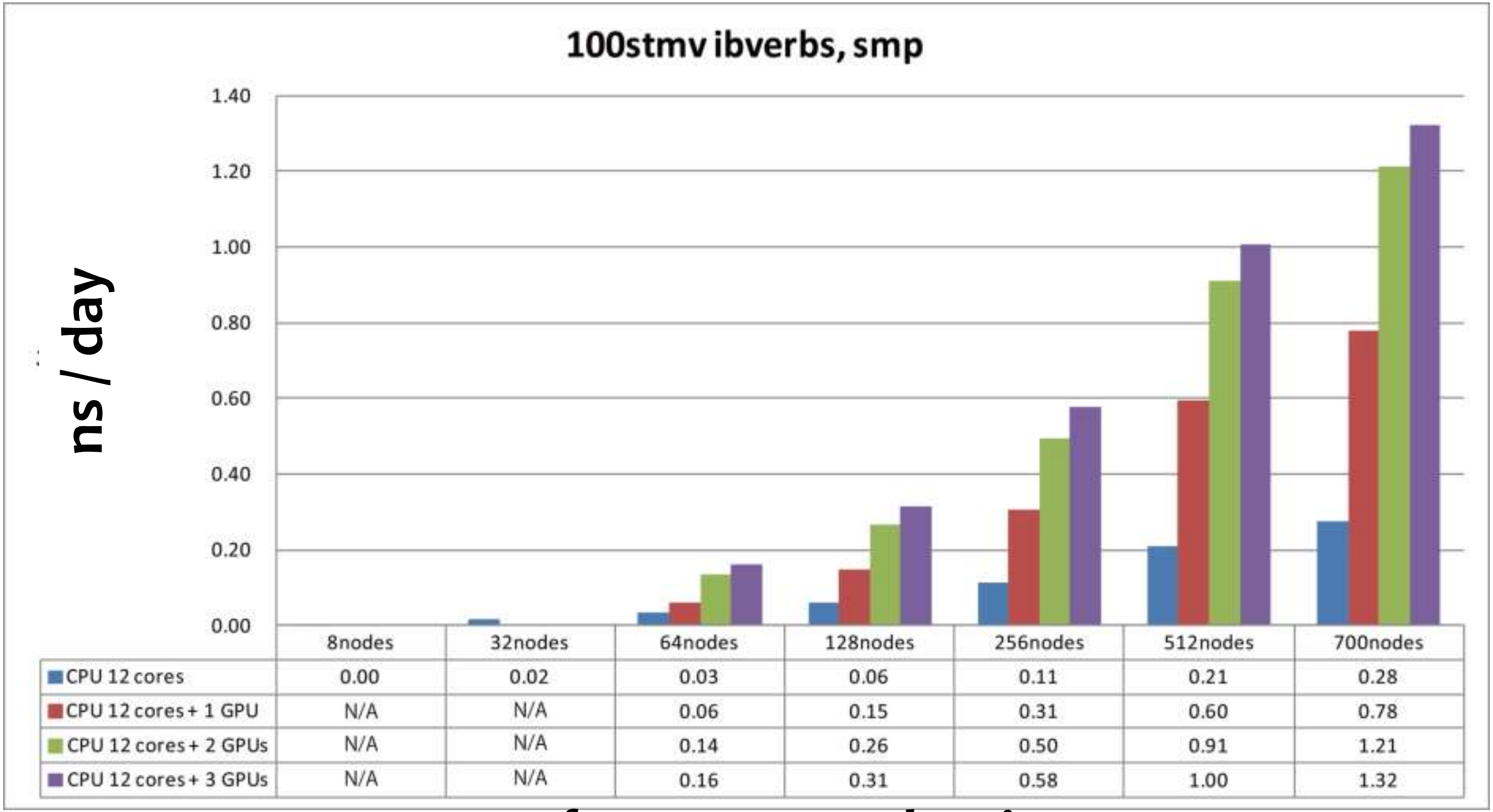
100-million-atom MD Simulation

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

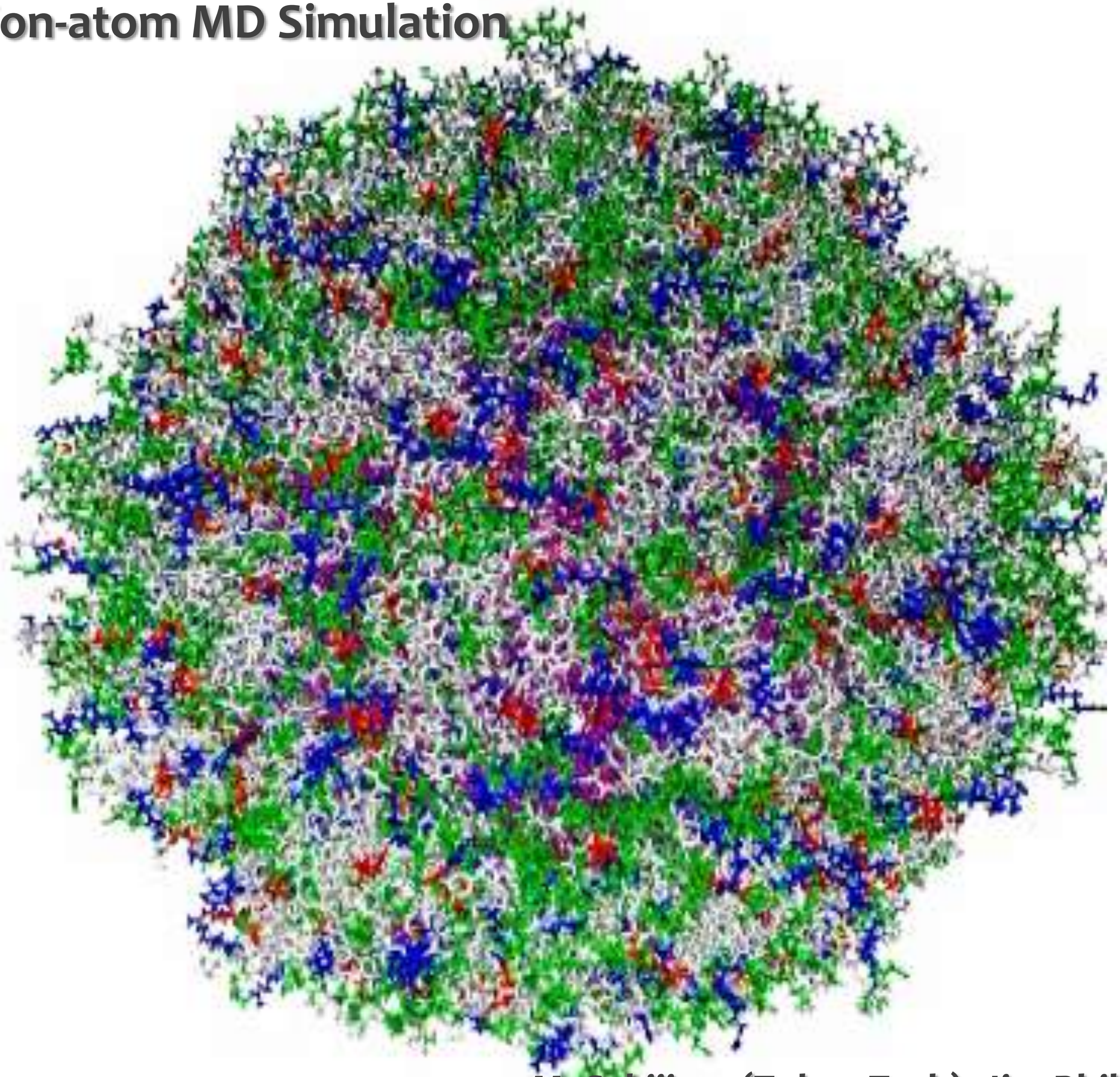
M. Sekijima (Tokyo Tech), Jim Phillips (UIUC)



	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

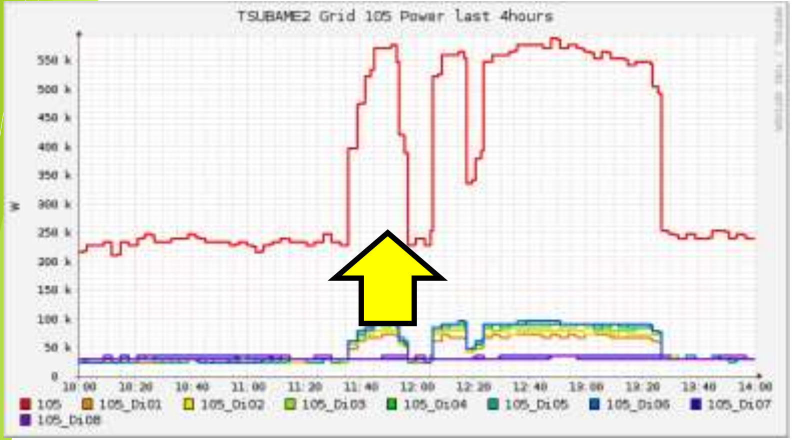
100-million-atom MD Simulation



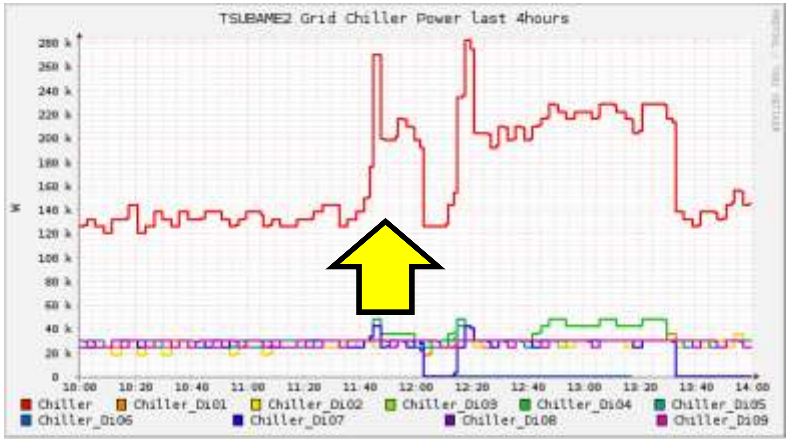
M. Sekijima (Tokyo Tech), Jim Phillips (UIUC)

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)

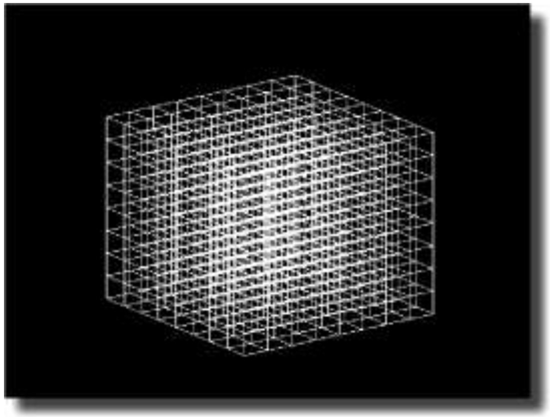
Total:

1527kW max

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

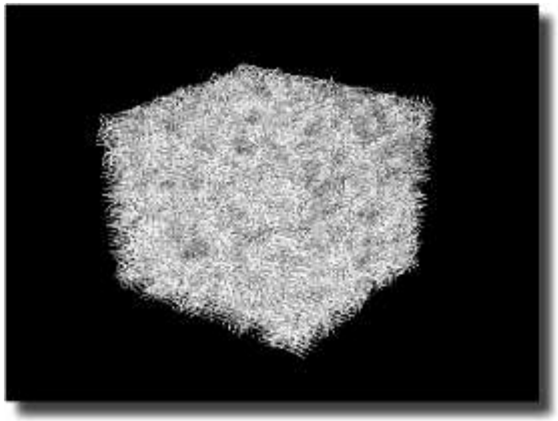
Petaflops scale turbulence simulation on TSUBAM 2.0

Isotropic turbulence



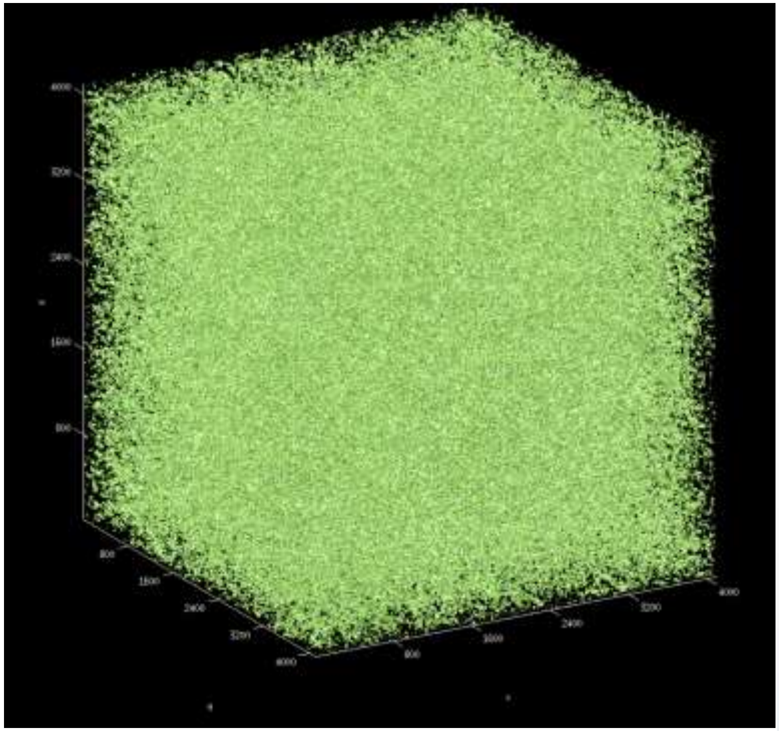
Pseudo Spectral Method (2/3 dealiasing)

Re : 500
N : 2048^3



Vortex Particle Method (Reinitialized CSM)

Re : 500
N : 2048^3



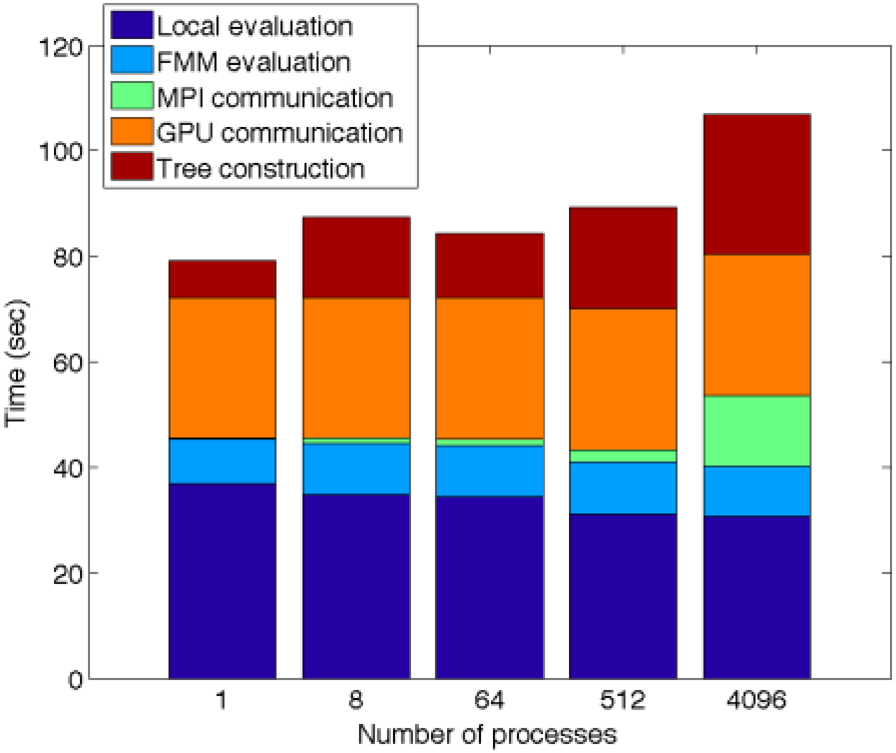
8 billion particles

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

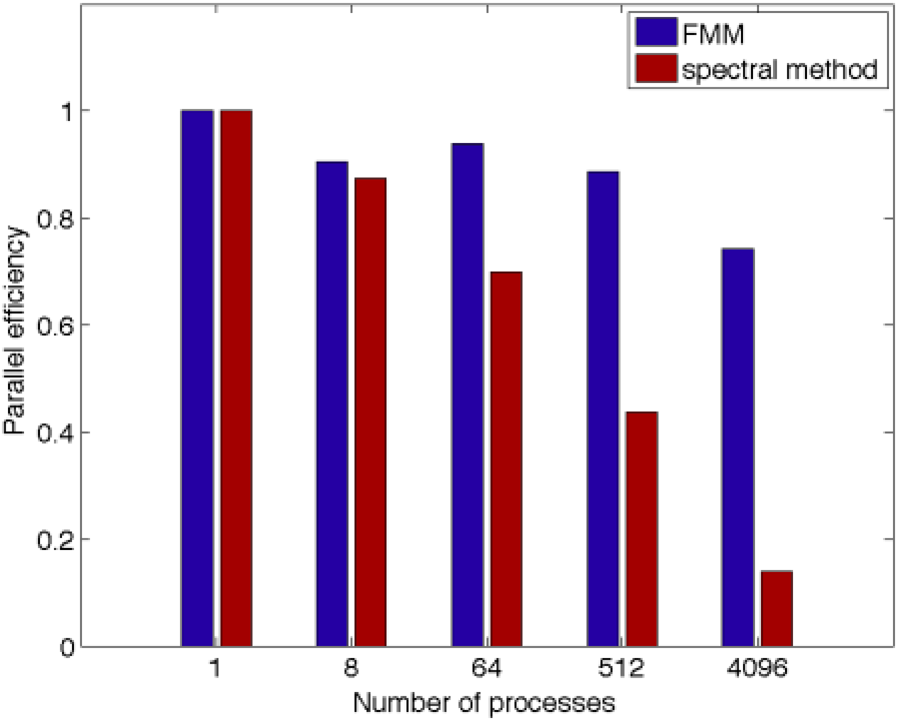
Petaflops scale turbulence simulation on TSUBAME 2.0

Weak Scaling

Wall clock time



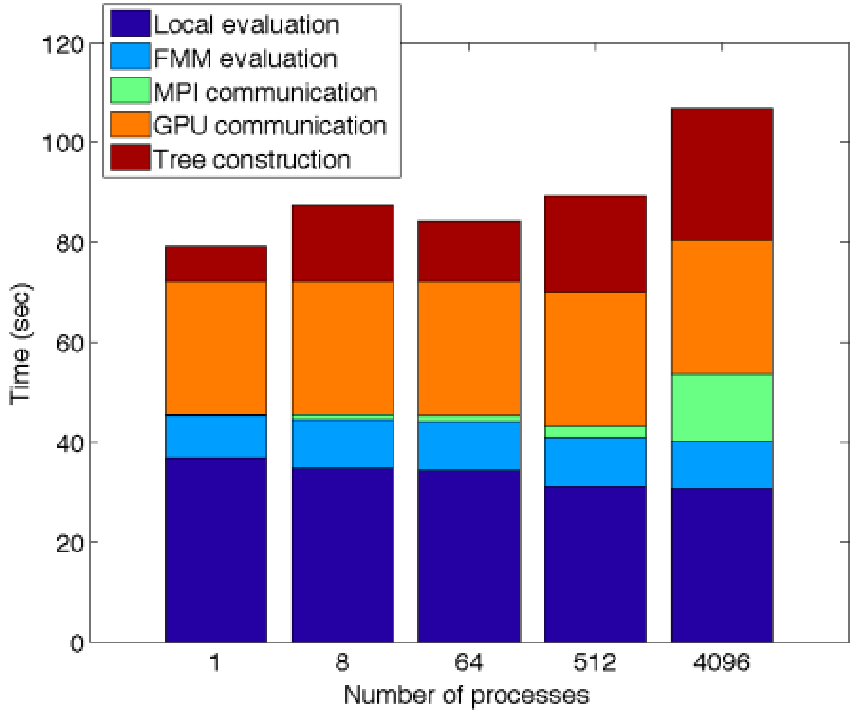
Parallel efficiency



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

Petaflops scale turbulence simulation on TSUBAM 2.0

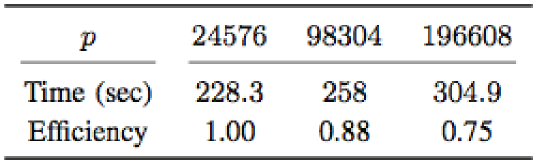
Present work



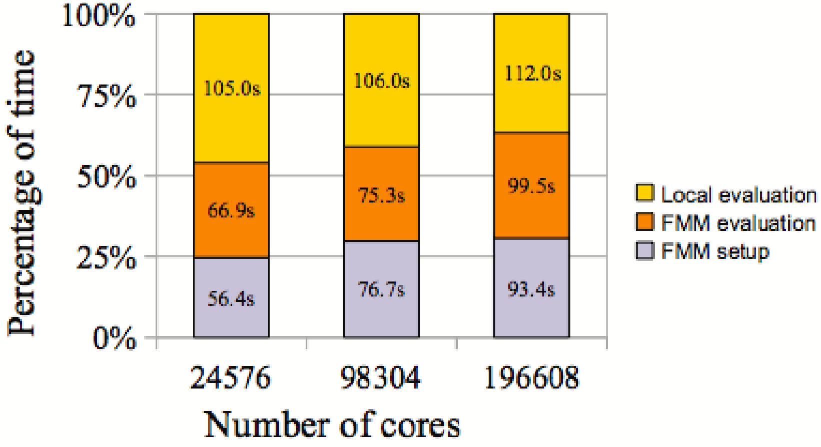
64 billion in 100 seconds

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

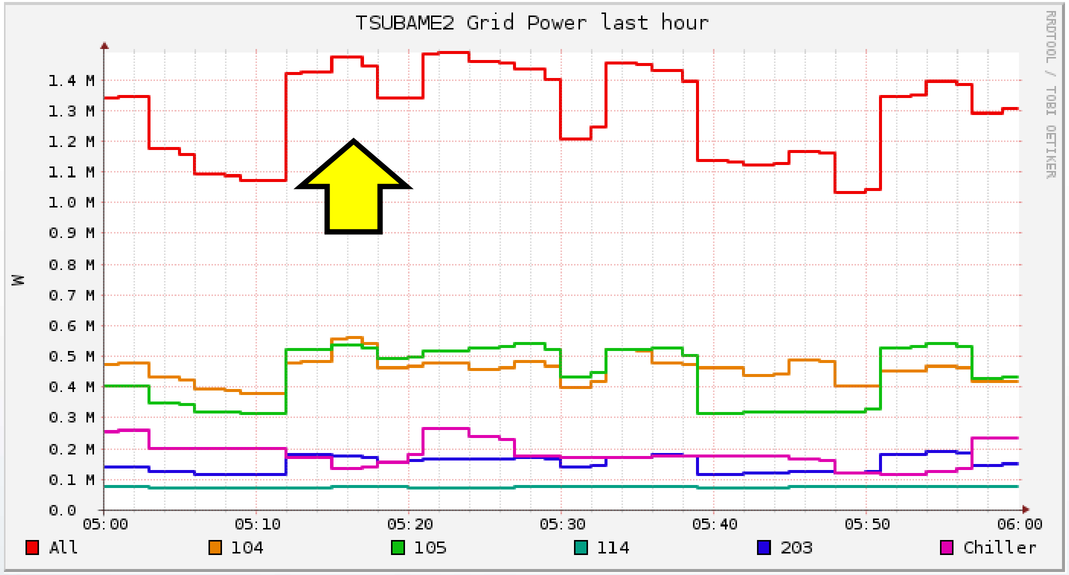


90 billion in 300 seconds

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

Large-Scale Metagenomics [Akiyama et. al. Tokyo Tech.]

Combined effective use of GPUs and SSDs on TSUBAME2.0.

Metagenome analysis: study of the genomes of uncultured microbes obtained from microbial communities in their natural habitats

Collecting bacteria in soil

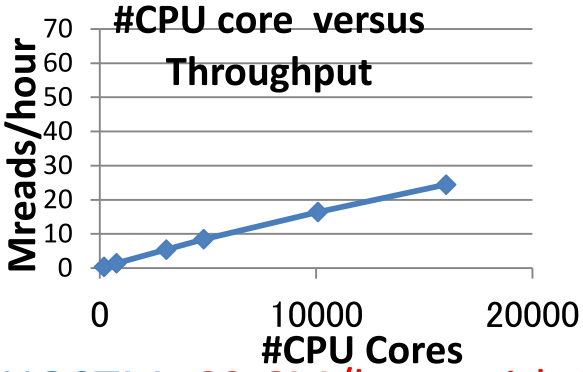
Two homology search tools are available:

1) BLASTX, standard software on CPUs
2) GHOSTM, our GPU-based fast software compatible with BLASTX

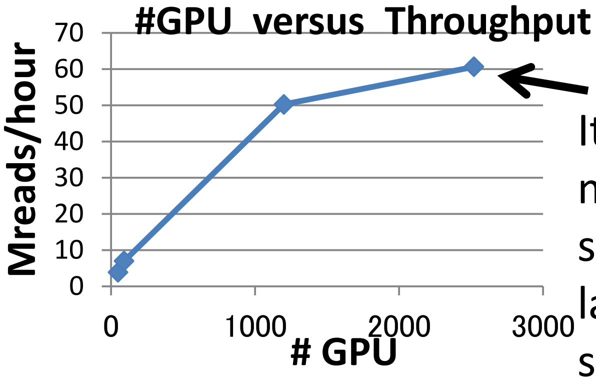
Data: 224million DNA reads(75b) /set
Pre-filtering: reduces to 71M reads
Search: 71M DNA vs. NCBI nr-aa DB (4.2GB)

Results on TSUBAME2.0

BLASTX: 24.4M/hour with 16K cores



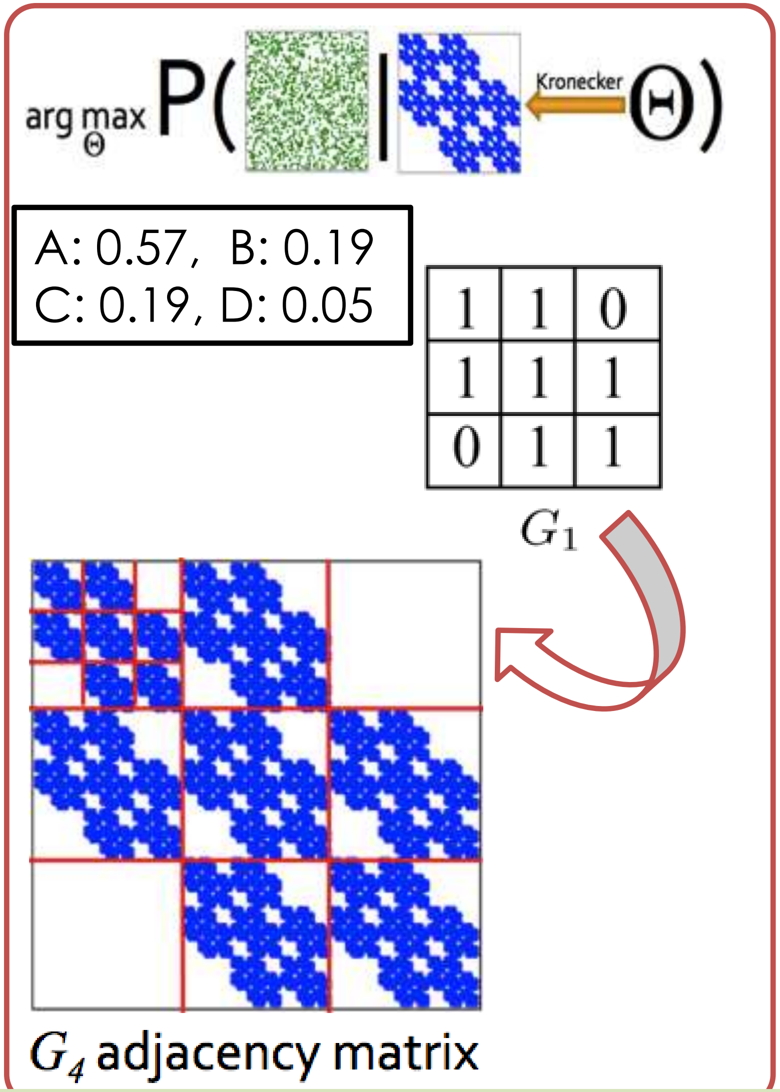
GHOSTM: 60.6M/hour with **2520 GPUs**



It would be more scalable with larger data sets

Graph500 on TSUBAME 2.0

Kronecker graph



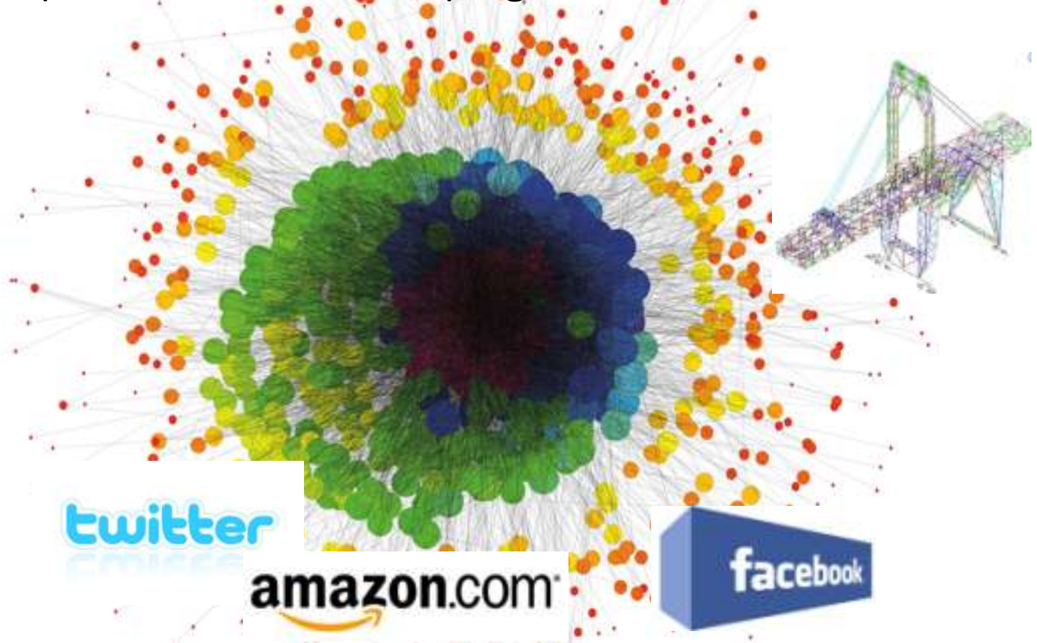
A: 0.57, B: 0.19
C: 0.19, D: 0.05

1	1	0
1	1	1
0	1	1

G_1

G_4 adjacency matrix

- 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.

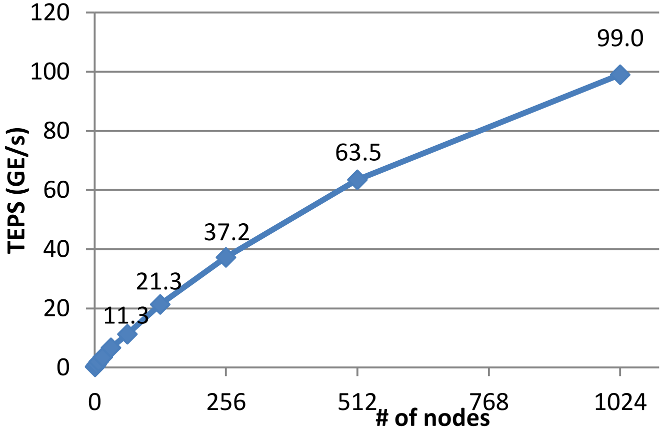


Toyotaro Suzumura, Koji Ueno, Tokyo Institute of Technology

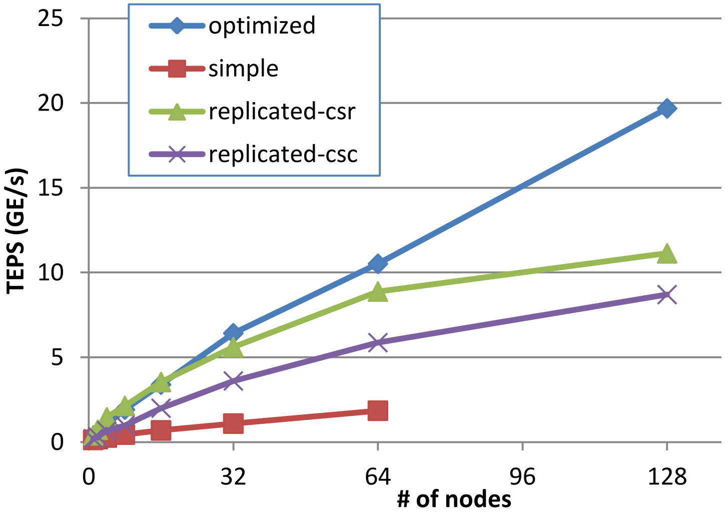
Highly Scalable Graph Search Method for the Graph500 Benchmark

- 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 2.0 103.9 GE/s (TEPS)
- ***#3 Graph 500 Nov. 2011***

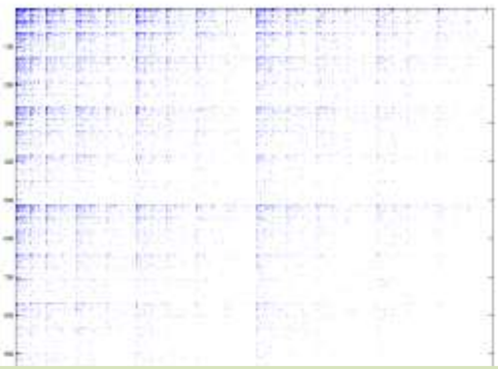
Performance of Our Optimized Implementation with Scale 26 per 1 node



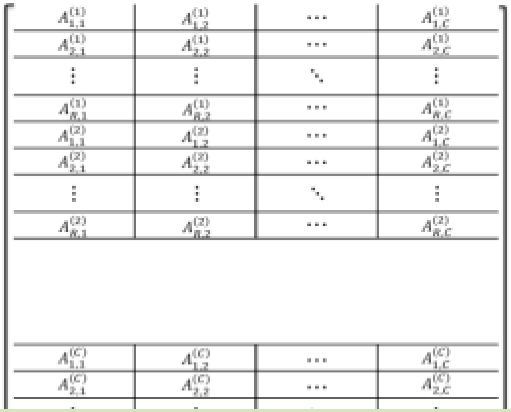
Performance Comparison with Reference Implementations (simple, replicated-csr and replicated-csc) and Scale 24 per 1 node



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

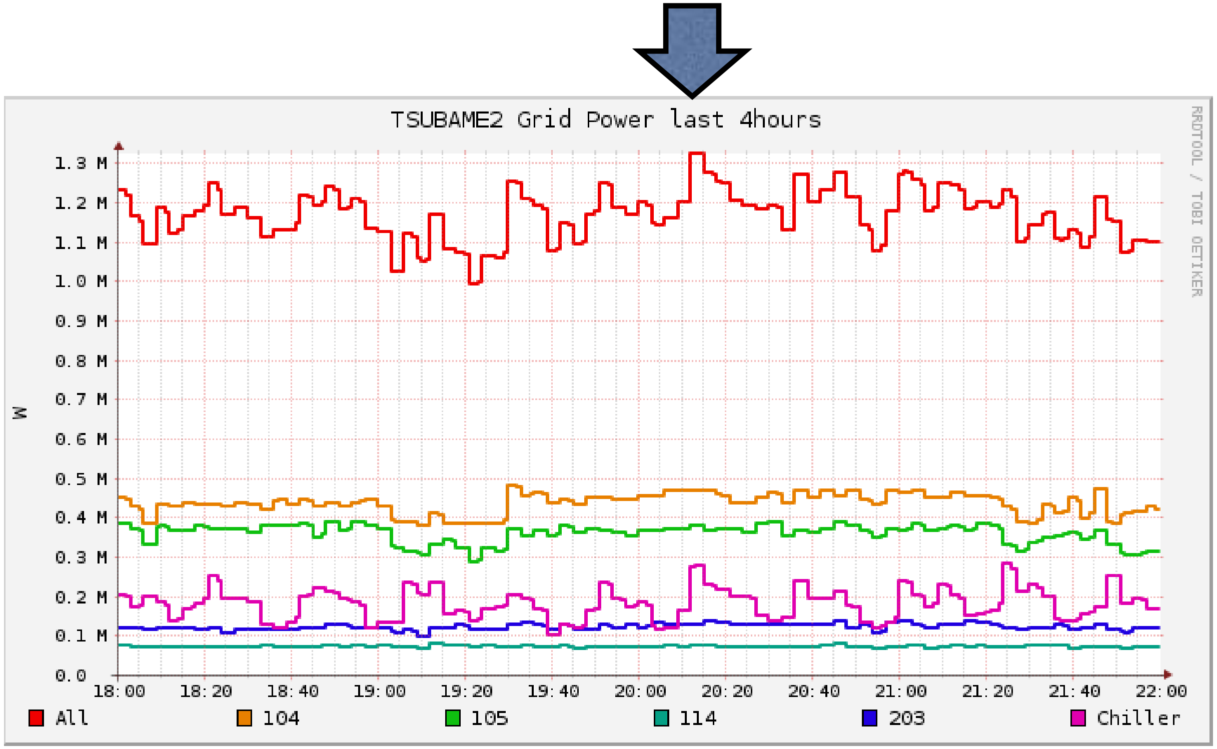


2D Partitioning Optimization



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

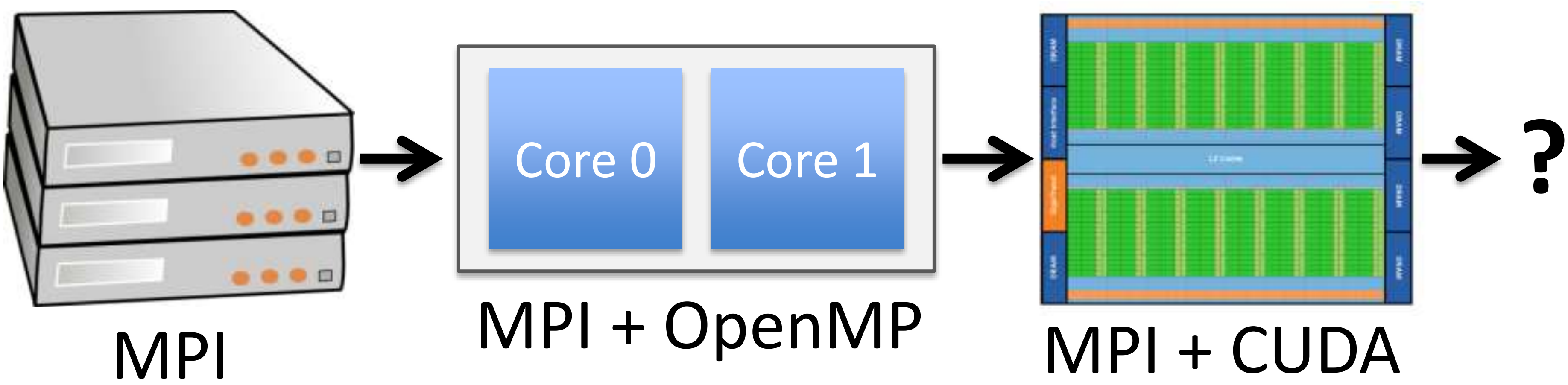
Total:

1323kW max

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	-	-	-

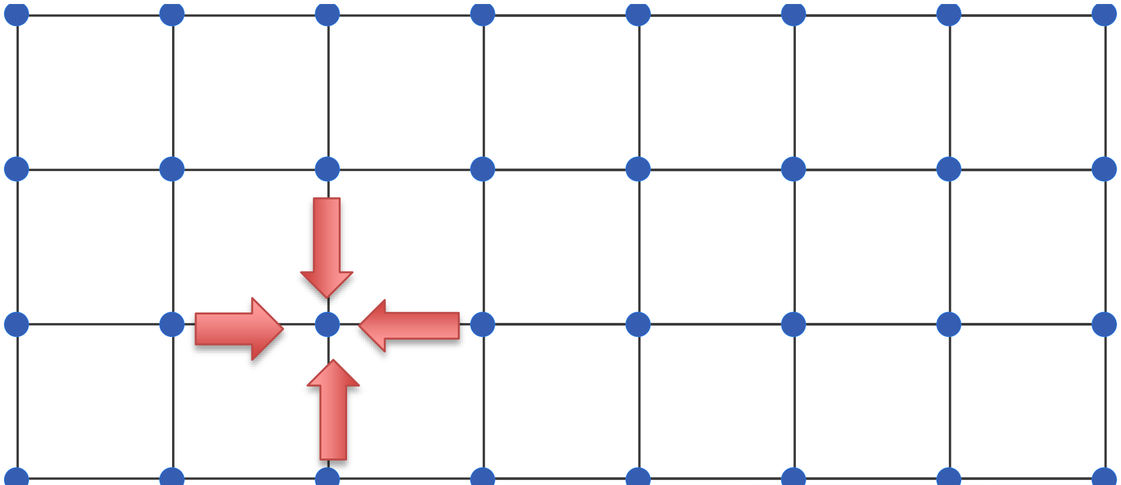
HPC Programming Model Trend



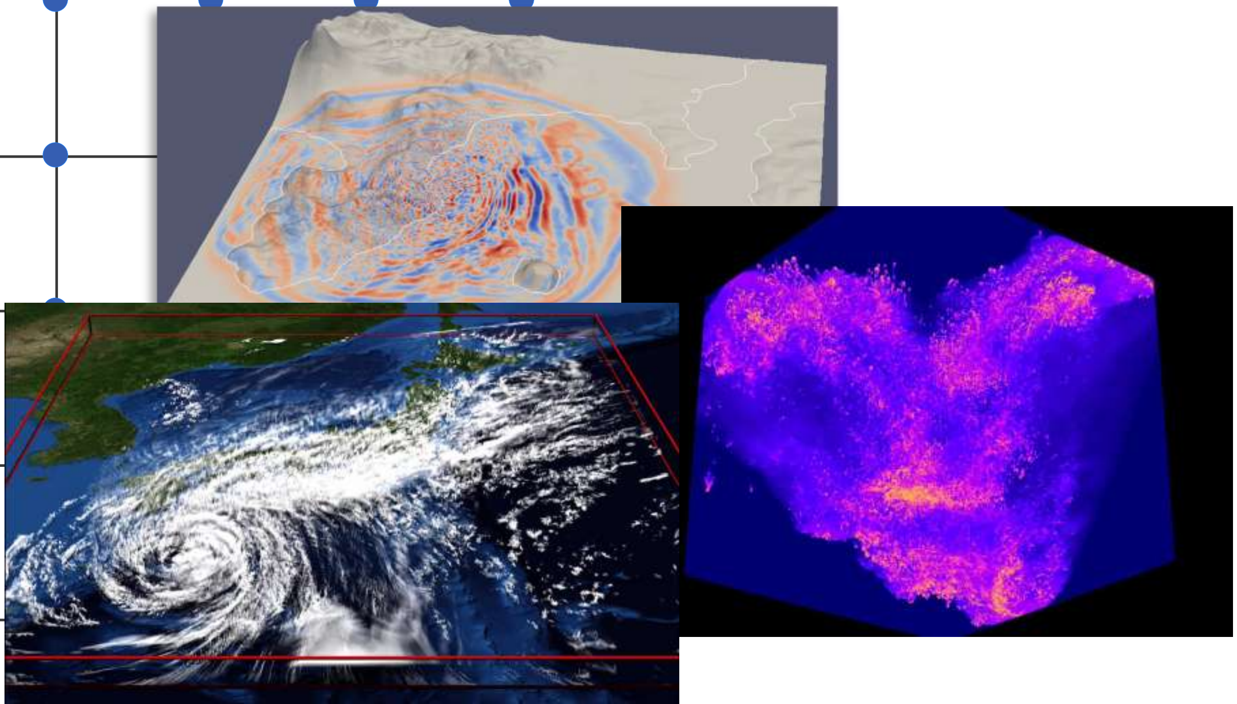
MPI → MPI + OpenMP → MPI + CUDA → ?

- Conceptually getting more complicated
- Needs deep understanding of architecture
- MPI + X
  - New architecture → New **“X”**

Example: Stencil Computation

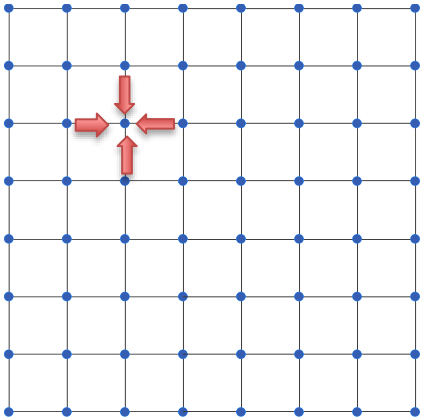


Pc’ = (Pc + Pn + Ps + Pw + Pe) * 1/5.0



Physis (Φύσις) Framework [SC11]

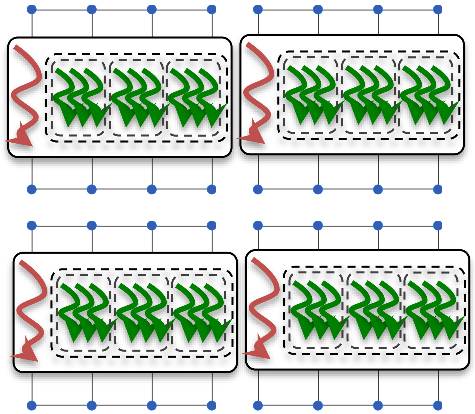
Physis (φύσις) is a Greek theological, philosophical, and scientific term usually translated into English as "nature.“ (Wikipedia:Physis)



Stencil DSL

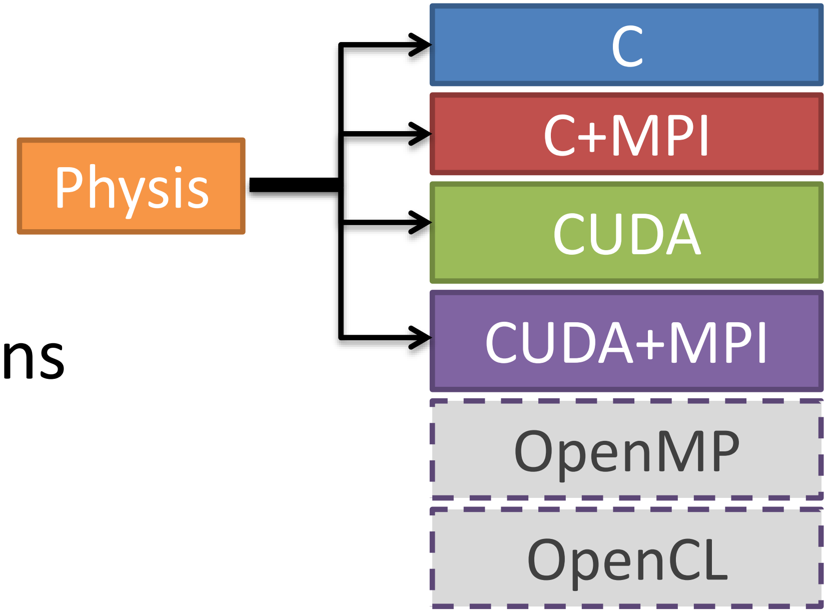
- Declarative
- Portable
- Global-view
- C-based

```
void diffusion(int x, int y, int z,
     PSGrid3DFloat g1, PSGrid3DFloat g2) {
float v = PSGridGet(g1,x,y,z)
 +PSGridGet(g1,x-1,y,z)+PSGridGet(g1,x+1,y,z)
 +PSGridGet(g1,x,y-1,z)+PSGridGet(g1,x,y+1,z)
 +PSGridGet(g1,x,y,z-1)+PSGridGet(g1,x,y,z+1);
PSGridEmit(g2,v/7.0);
}
```

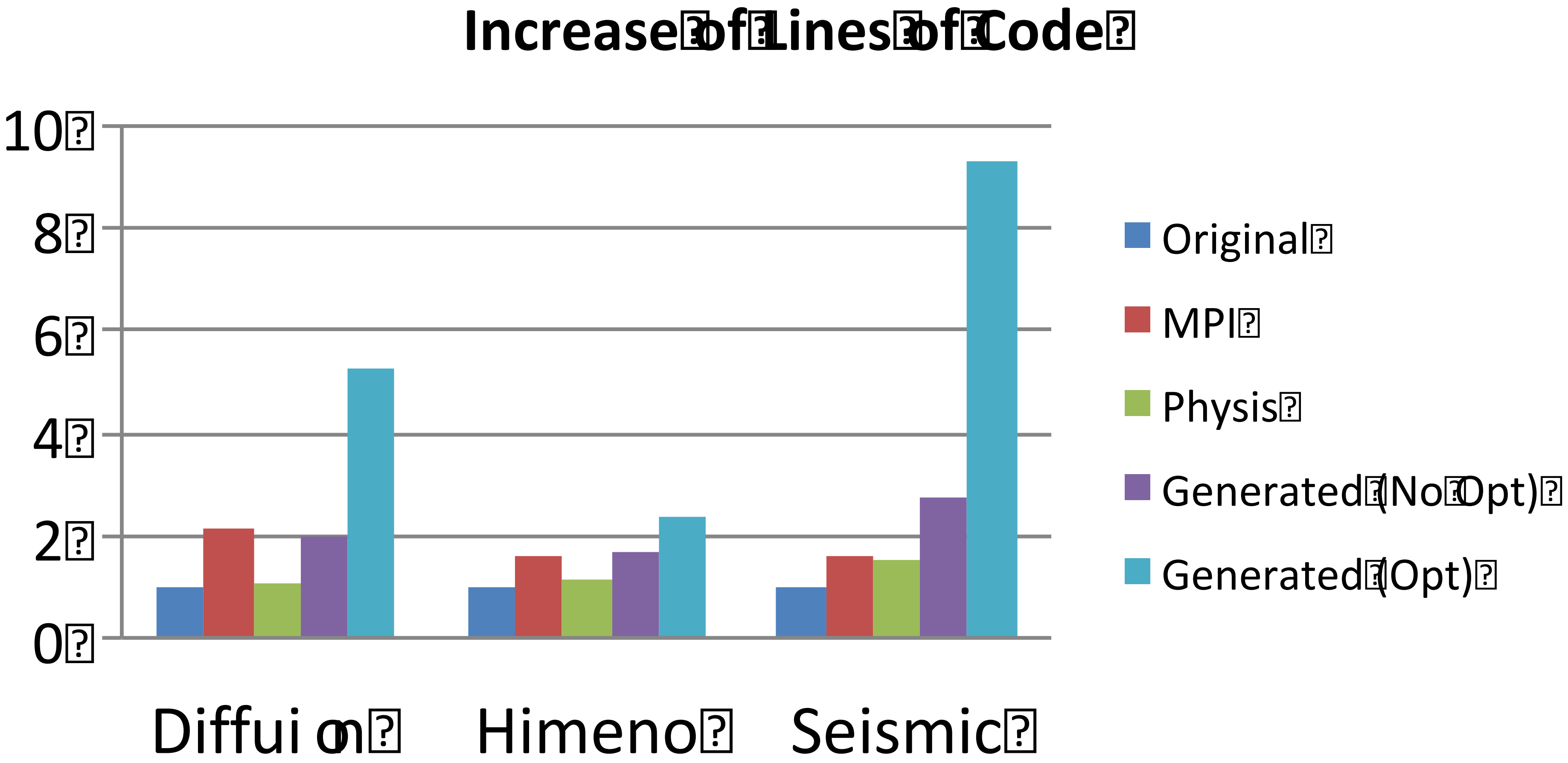


DSL Compiler

- Target-specific code generation and optimizations
- Automatic parallelization

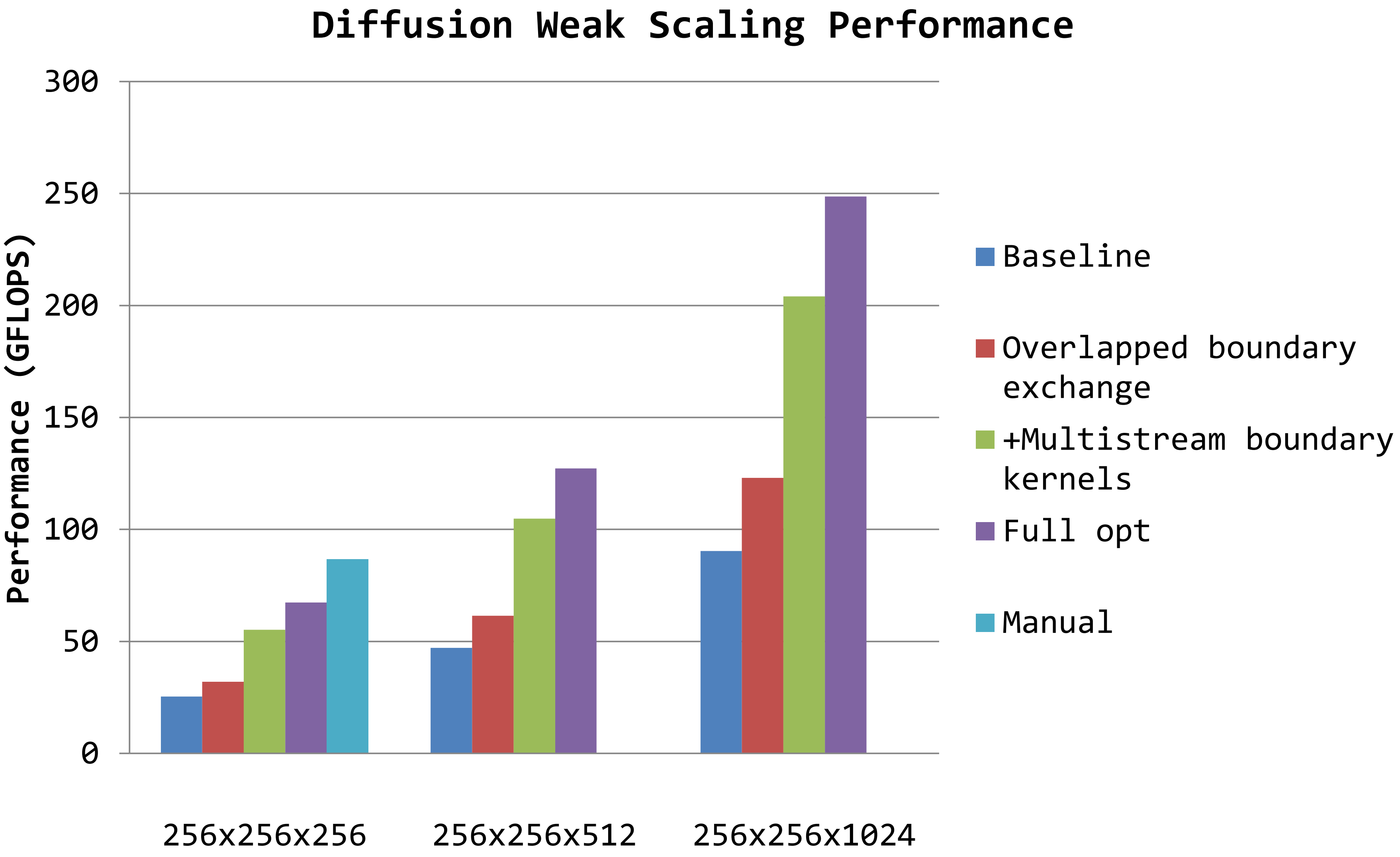


Productivity

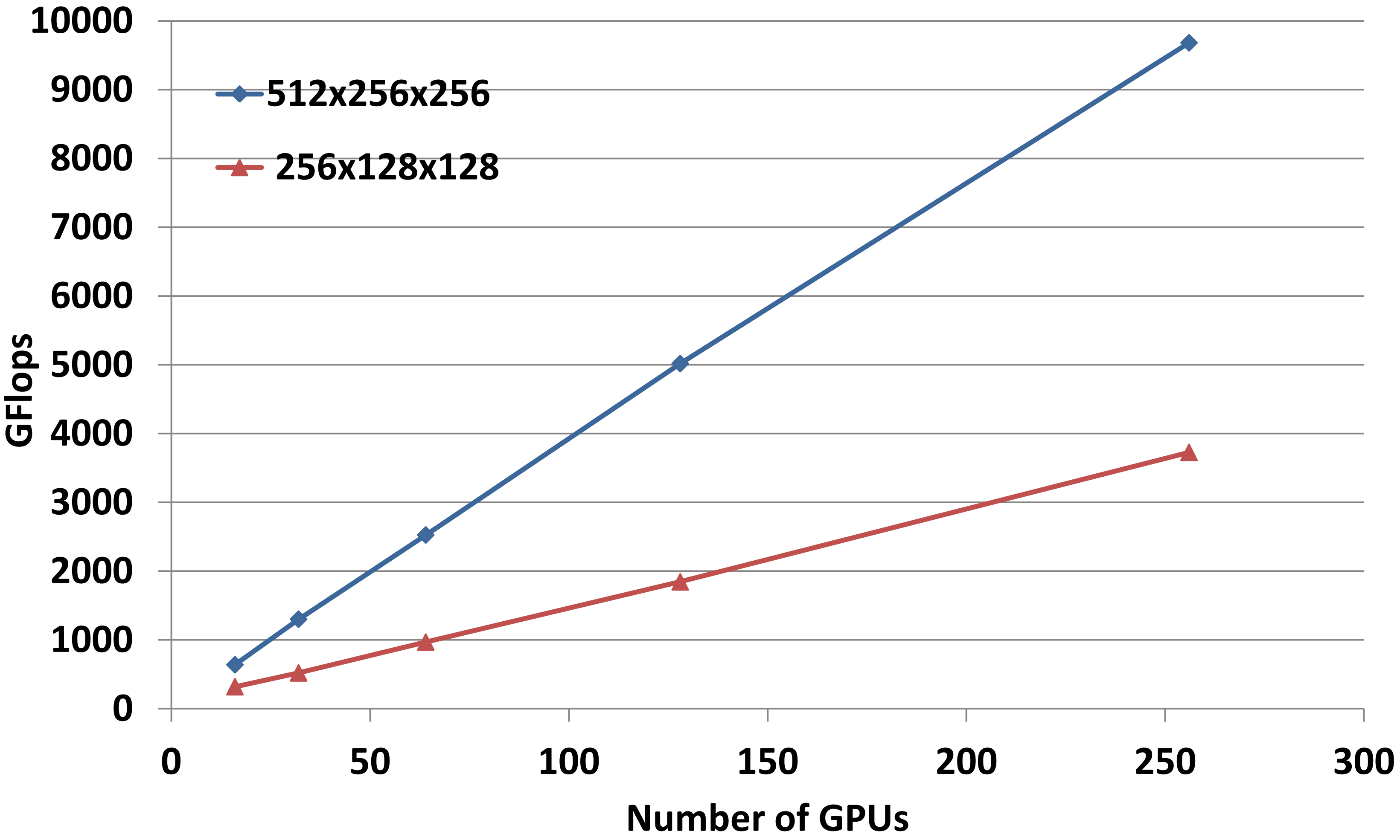


Similar size as sequential code in C

Optimization Effects

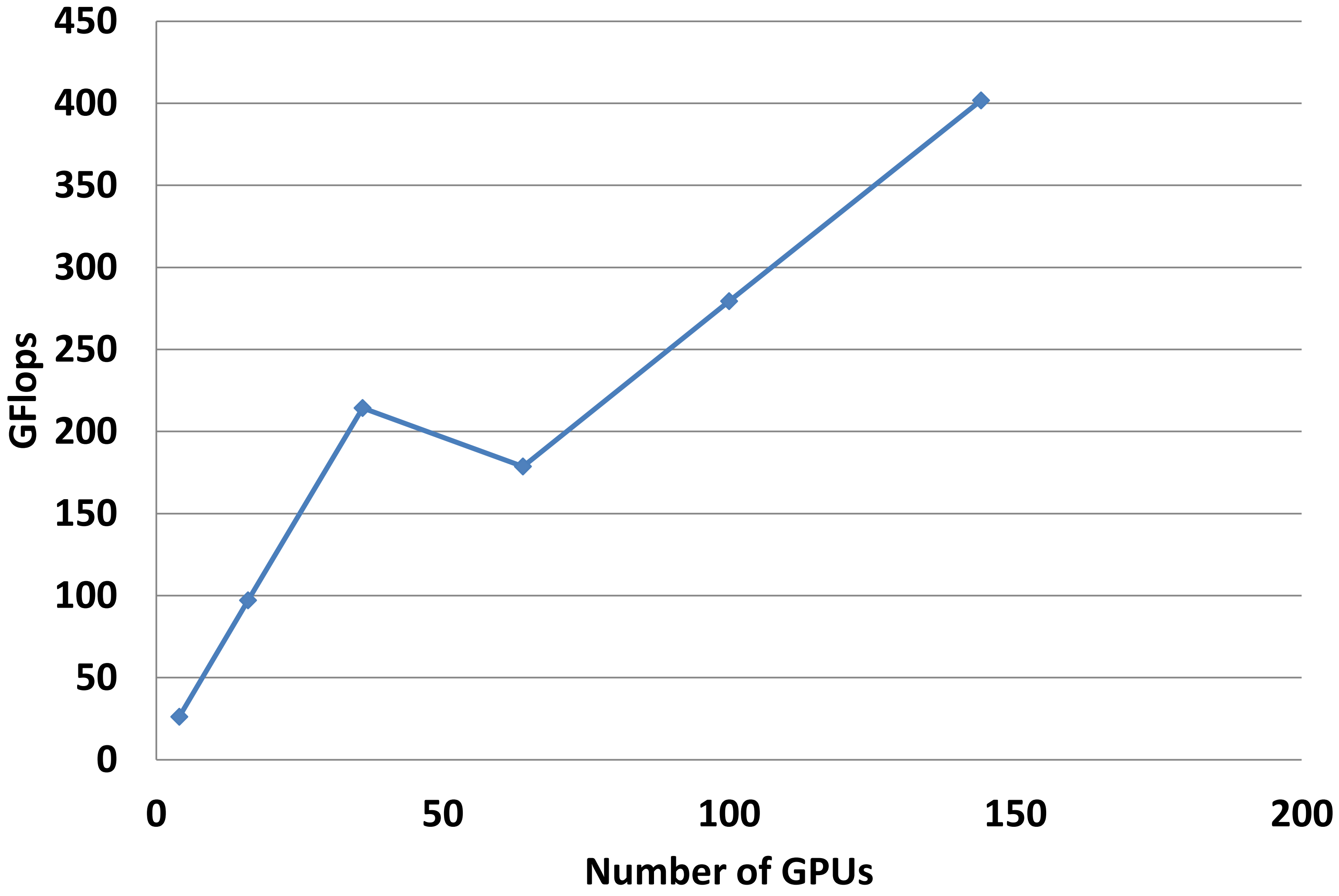


Diffusion Weak Scaling



Seismic Weak Scaling

Problem size: 256x256x256 per GPU



Summary

- Tsubame2.0 a year later since Nov. 2010
  - Over 2000 users, ~100 users online
  - ~90% system util, ~50% GPU util
  - System up 24/7, tolerated 3/11 disaster
  - Very power efficient, ~3/4 Tsubame 1.0
  - Collaborative R&D really paid off
- Accolades
  - 2011 Gordon Bell x 2
  - 2010-2011 Greenest Production SC Green500
  - 2010-2011 3 consecutive Top5 in Top500
  - 2011 #3 Green 500
  - Lots and lots of publications incl. 4 SC2011 papers
  - Many many more awards and press
- But most importantly, it works!