Tsubame 2.0: 2-Petaflops Performance of a GPU Stencil Application

Takayuki Aoki

Global Scientific Information and Computing Center (GSIC)
Tokyo Institute of Technology

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

Supercomputer in the world

TOP500 SUPERCOMPUTER SITES

2011 November

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu	705024	10510.00	11280.38	12659.9
2	National Supercomputing Center in Tianjin China	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010 NUDT	186368	2566.00	4701.00	4040.0
3	DOE/SC/Oak Ridge National Laboratory United States	Cray XT5-HE Opteron 6-core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.0
4	National Supercomputing Centre in Shenzhen (NSCS) China	Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 / 2010 Dawning				
5	GSIC Center, Tokyo Institute of Technology Japan	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows / 2010 NEC/HP	73278	1192.00	2287.63	1398.6

32 Million US$

TSUBAME 2.0

System (58 racks)

1442 nodes: 2952 CPU sockets,
4264 GPUs

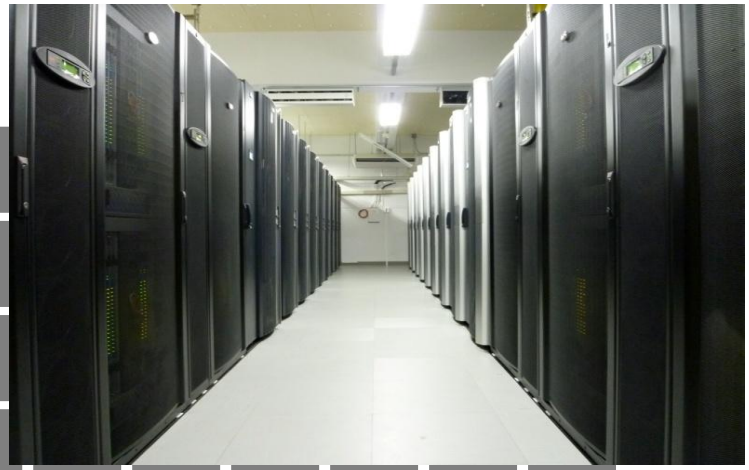
Performance: 224.7 TFLOPS (CPU) ※ Turbo boost
2196 TFLOPS (GPU)
Total: **2420** TFLOPS

Rack (30 nodes)

Performance: 51.0 TFLOPS
Memory: 2.03 TB

Compute Node
(2 CPUs, 3 GPUs)

Performance: 1.7 TFLOPS
Memory: 58.0GB(CPU)
+9.7GB(GPU)



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

GPU M2050

Details of Compute Node

GP GPU

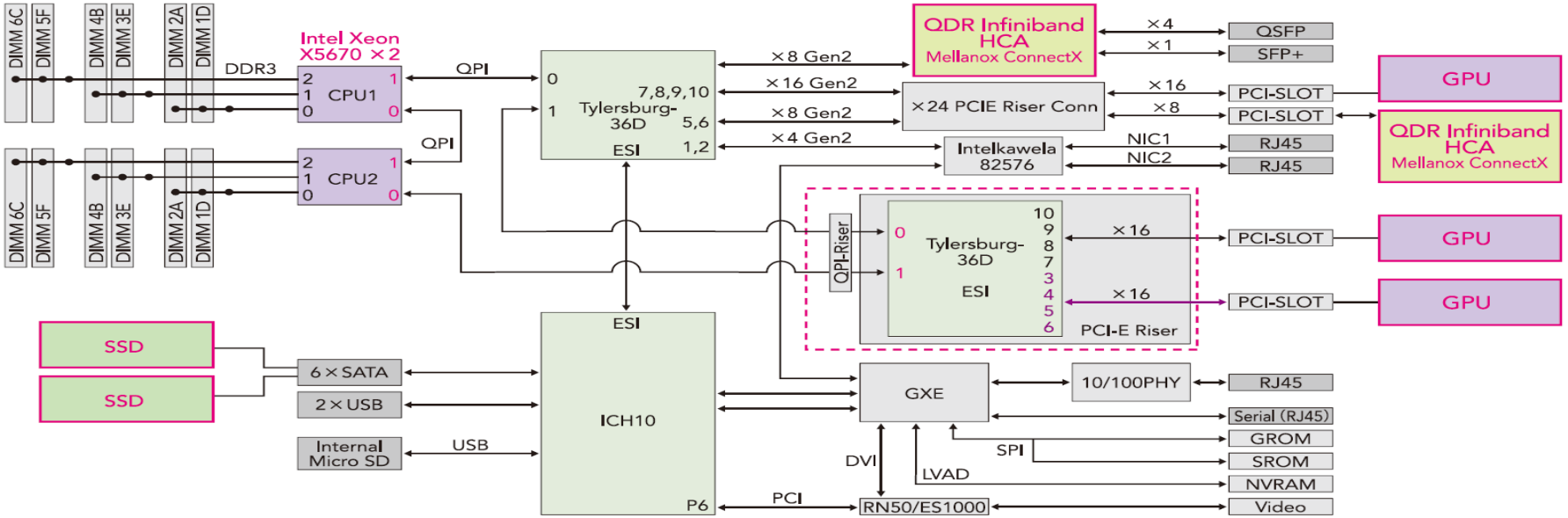
HP ProLiant SL390s

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

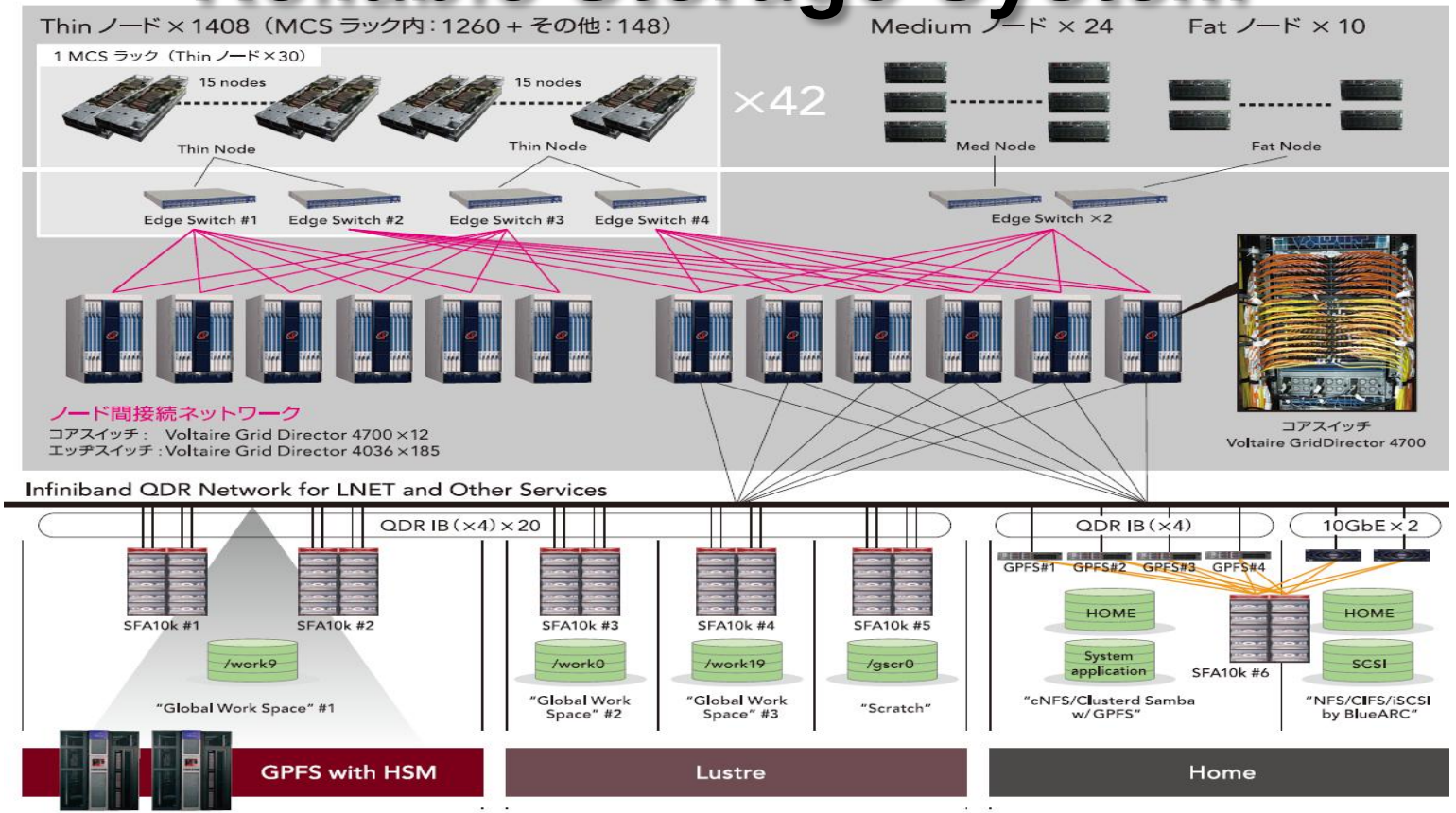
Memory : 58GB DDR3 1333MHz —部 103GB

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



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

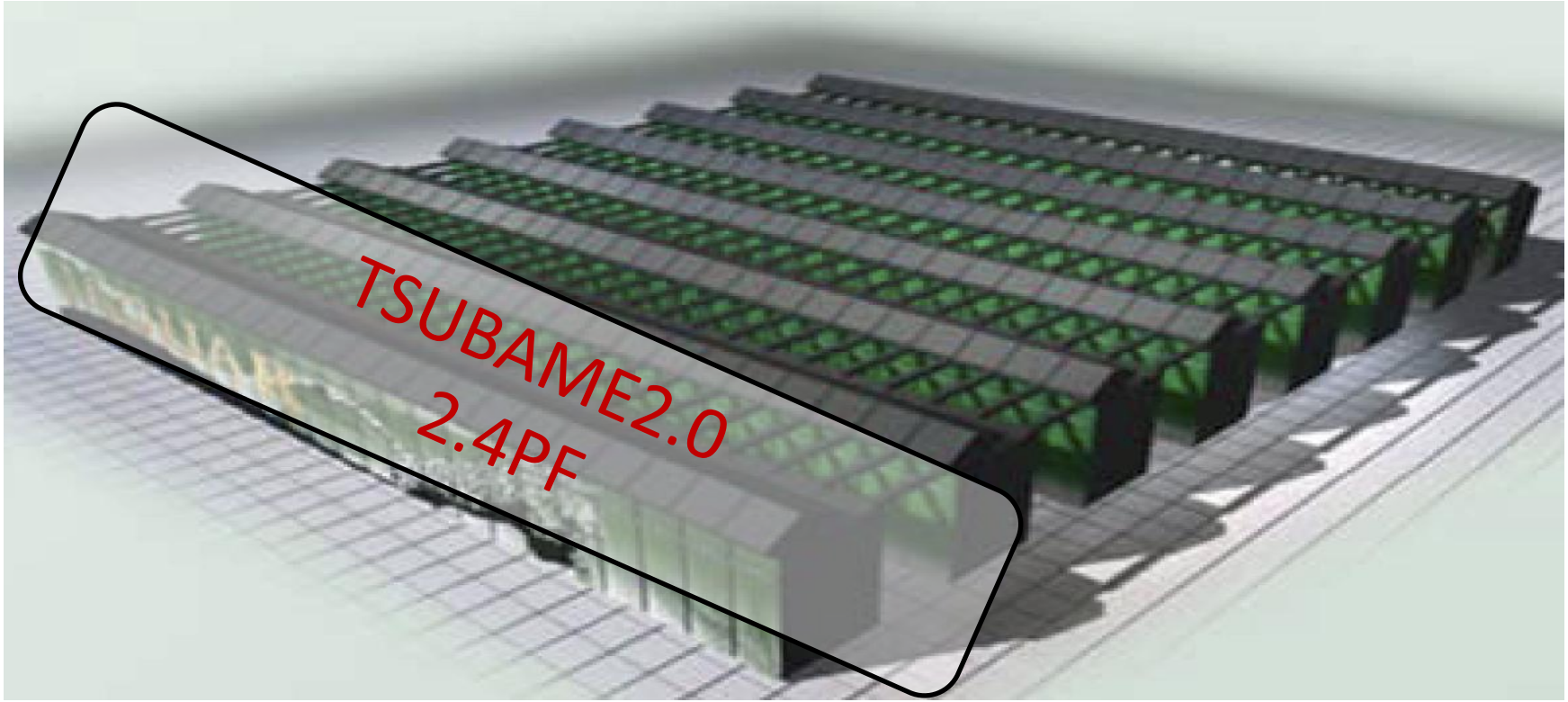
High-Speed Network and Reliable Storage System



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

ORNL Jaguar vs Tsubame 2.0

Similar Peak Performance, 1/5 the Size and Power



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

Supercomputer in the world

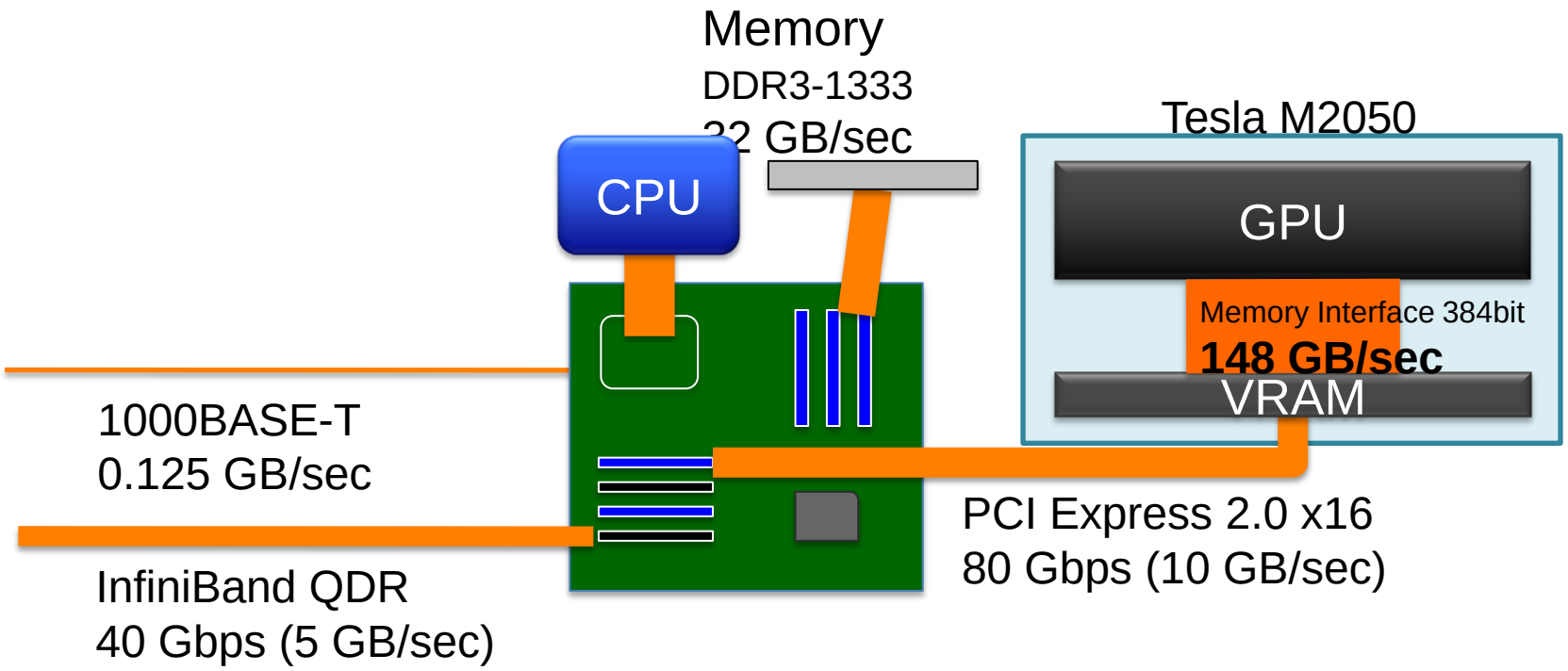
The Green500 list, June 2011

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	2097.19	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype 2	40.95
2	1684.20	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype 1	38.80
3	1375.88	Nagasaki University	DEGIMA Cluster, Intel i5, ATI Radeon GPU, Infiniband QDR	34.24
4	958.35	GSIC Center, Tokyo Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows	1243.80
5	891.88	CINECA / SCS - SuperComputing Solution	iDataPlex DX360M3, Xeon 2.4, nVidia GPU, Infiniband	160.00
6	824.56	RIKEN Advanced Institute for Computational Science (AICS)	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	9898.56
7	773.38	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54

TSUBAME2.0 PUE = 1.2 (Power Usage Effectiveness)

Heterogeneous Computer

- **Several Bandwidth Bottle Necks**



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

Next Generation

Weather Prediction

GP GPU

Collaboration: Japan Meteorological Agency

Meso-scale Atmosphere Model:

Cloud Resolving Non-hydrostatic model

Compressible equation taking consideration of sound waves.

Meso-scale

2000 km

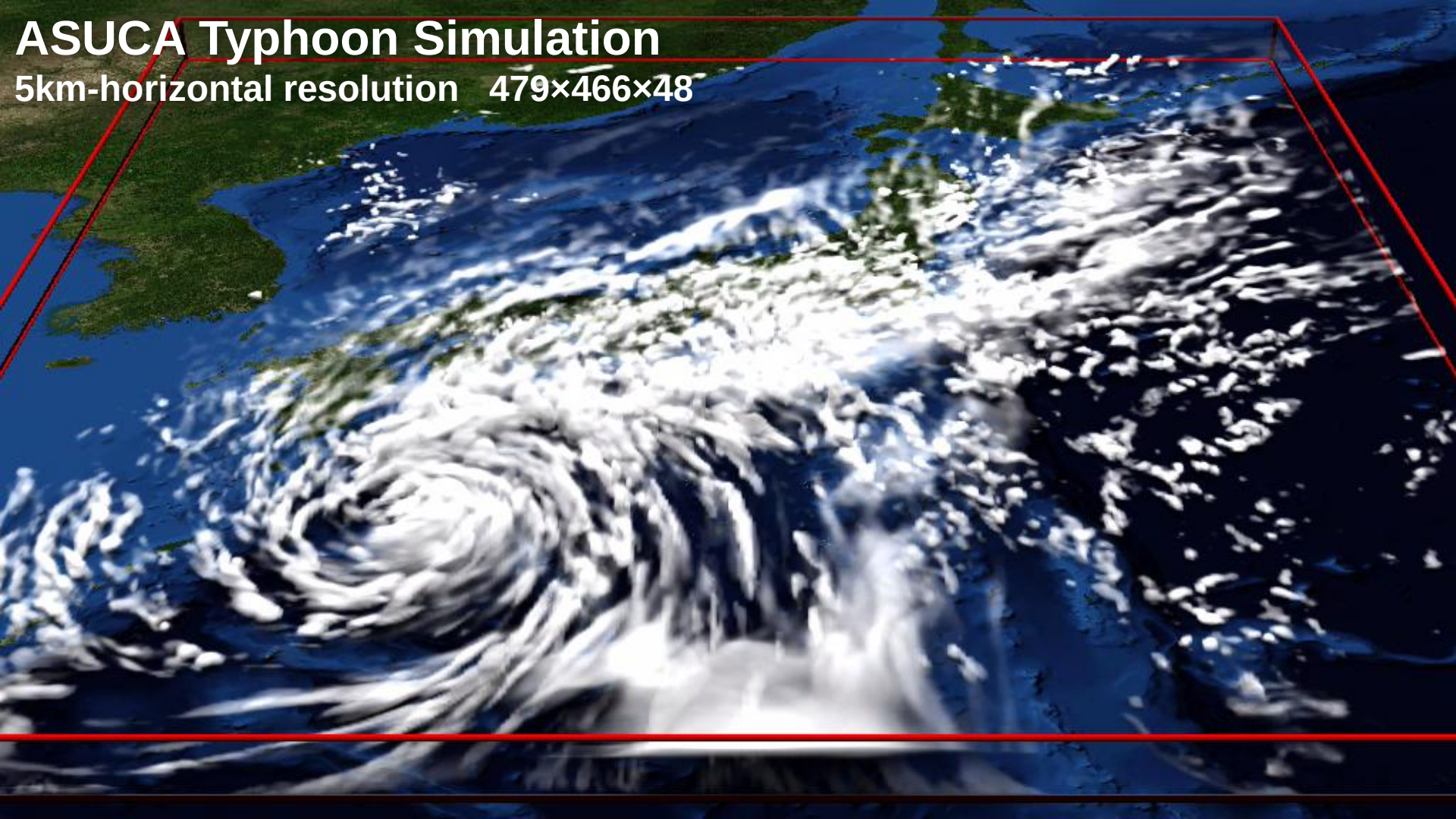
Typhoon

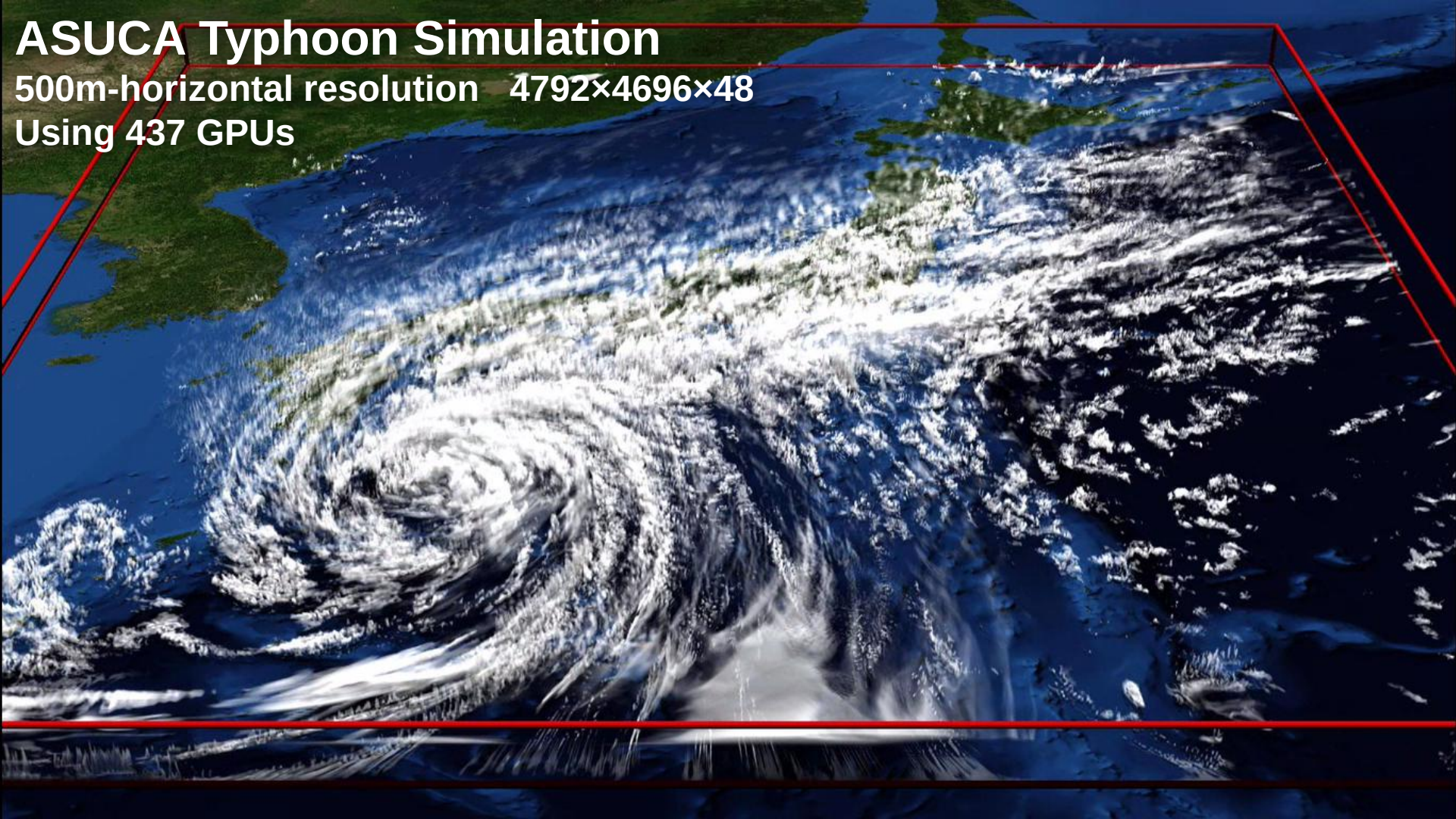
a few km

Tornado, Down burst

Heavy Rain

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


ASUCA Typhoon Simulation
5km-horizontal resolution 479×466×48


ASUCA Typhoon Simulation
500m-horizontal resolution 4792×4696×48
Using 437 GPUs

Two-Phase Flow Simulation

GP GPU

Particle Method

ex. SPH

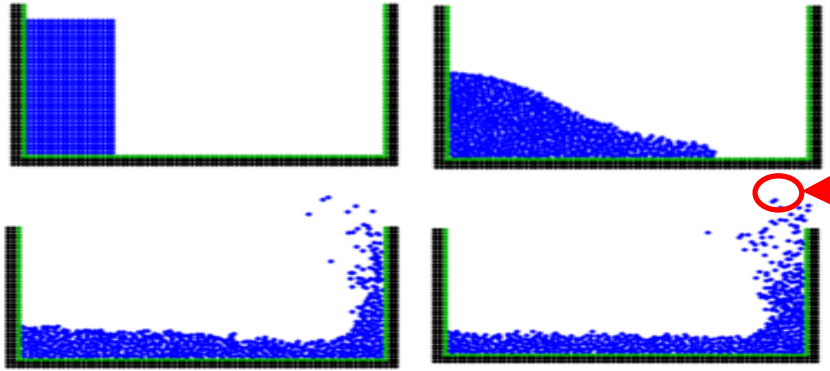
Low accuracy

< $10^{6\text{-}7}$ particles

Mesh Method (Surface Capture)

- Navier-Stokes solver : Fractional Step
- Time integration : 3rd TVD Runge-Kutta
- Advection term : 5th WENO
- Diffusion term : 4th FD
- Poisson : MG-BiCGstab
- Surface tension : CSF model
- Surface capture : CLSVOF(THINC + Level-Set)

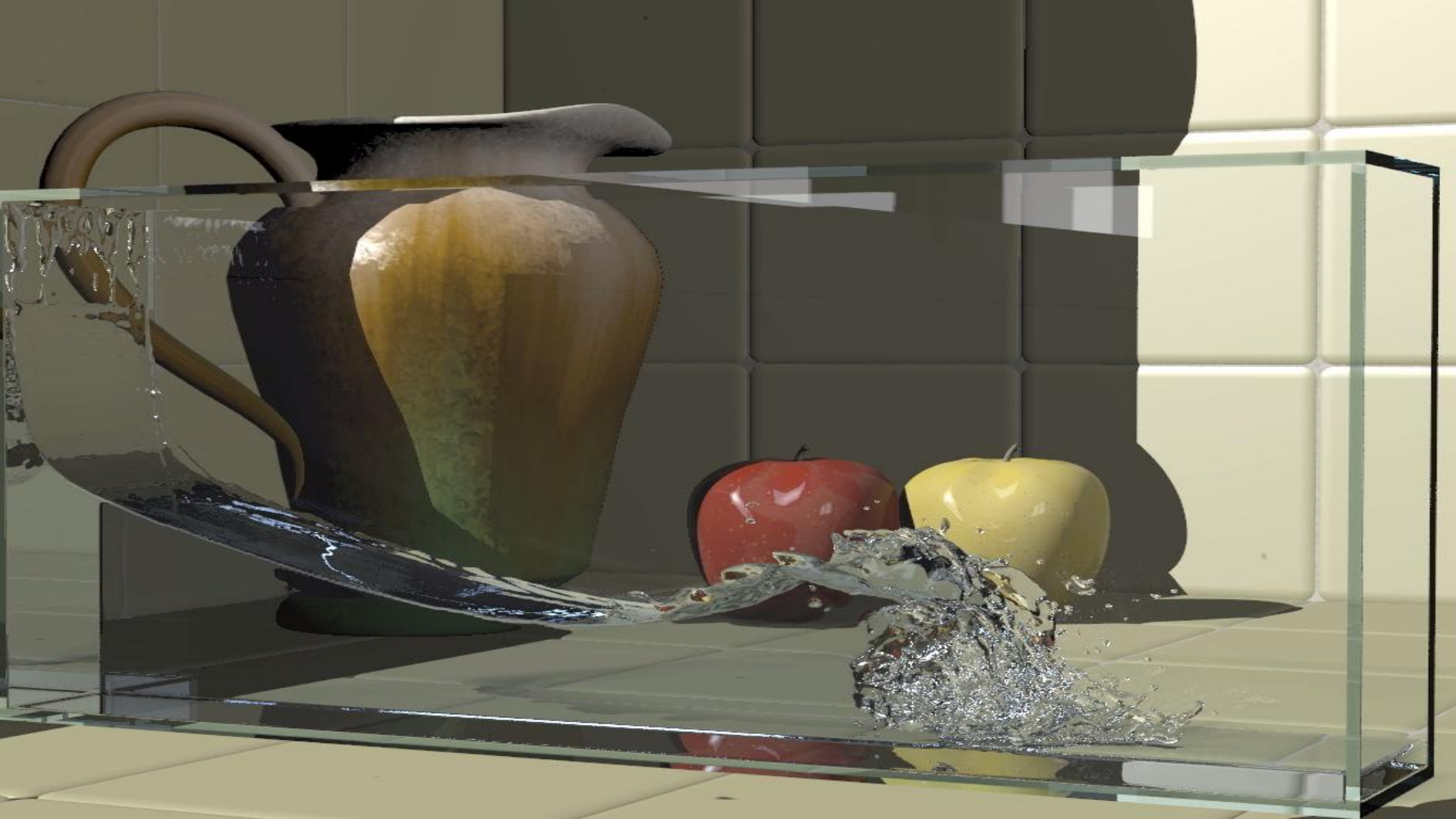
High accuracy > $10^{8\text{-}9}$ mesh points



not splash

Numerical noise and unphysical oscillation

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

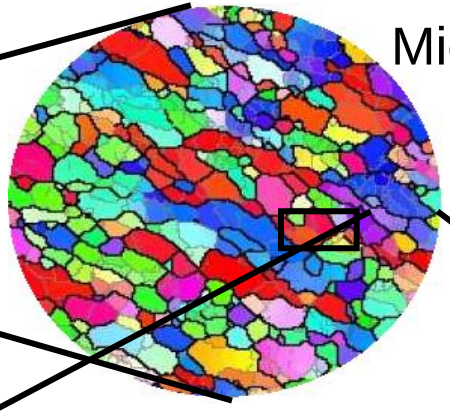


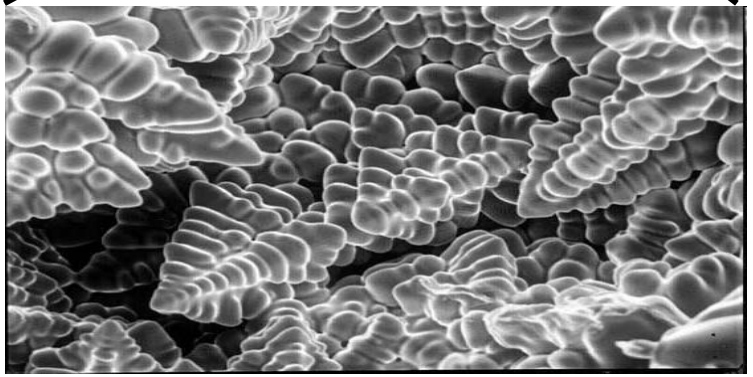
Development of New Materials

GP GPU

Mechanical Structure

Microstructure





Low-carbon society

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

Developing lightweight strengthening material by controlling microstructure

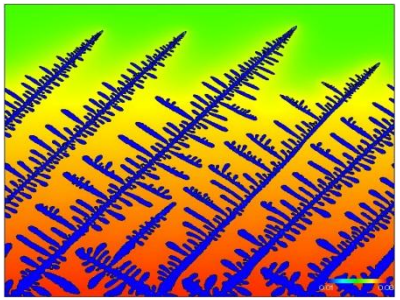
Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

Impact on Material Science

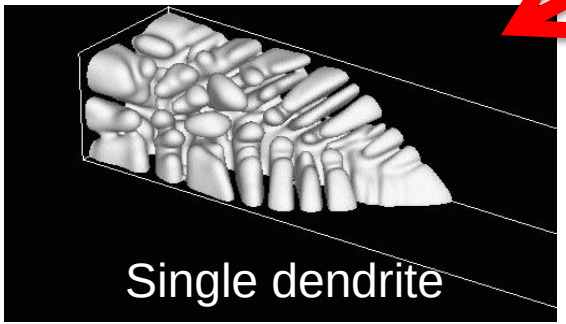
GP GPU

Previous Research

2D

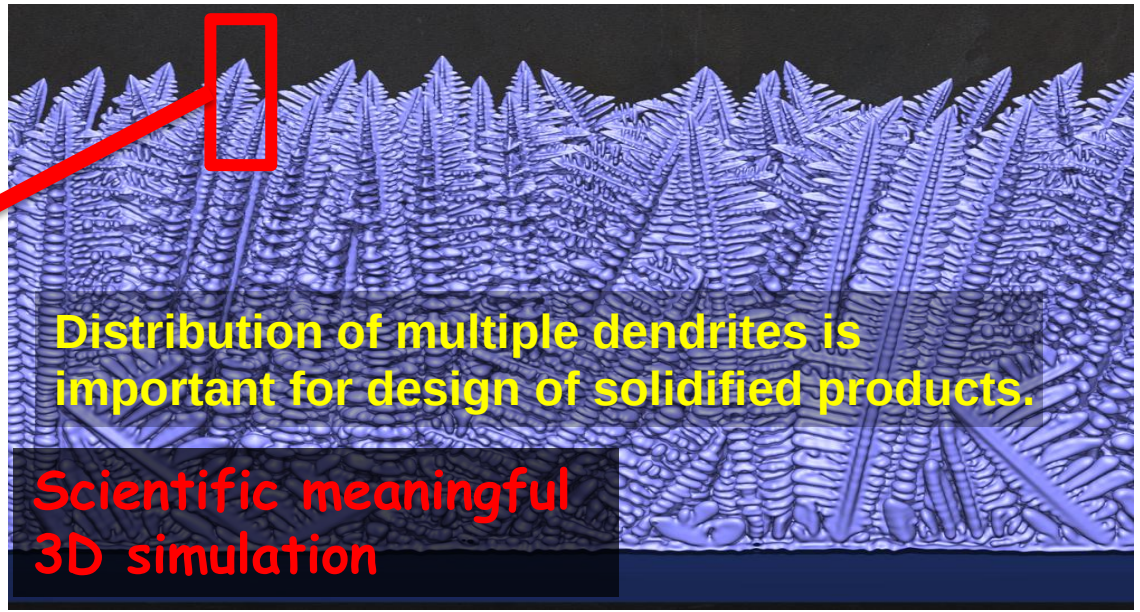


3D simple shape



Peta-scale Simulation

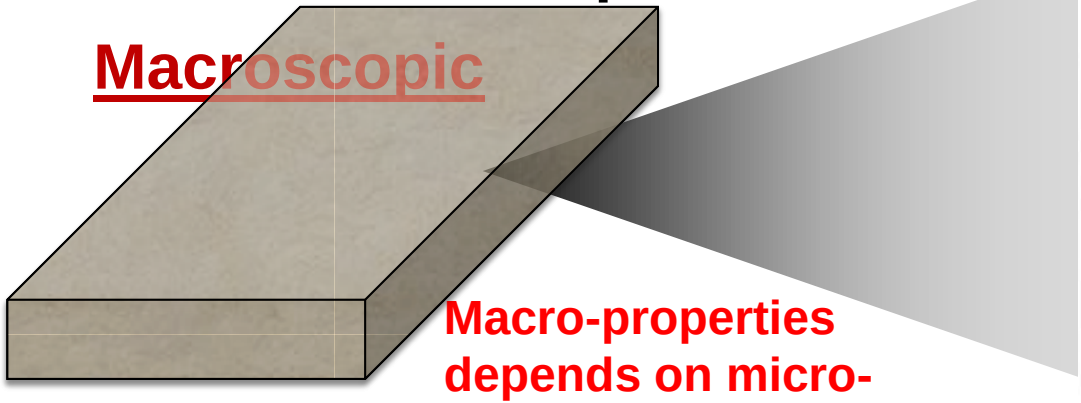
- ✓ GPU-rich Supercomputer
- ✓ Optimization for Peta-scale computing



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

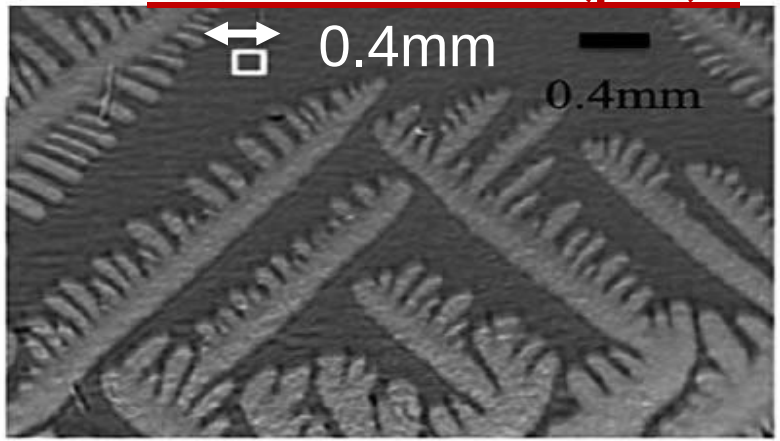
Metal Dendritic Solidification

■ Mechanical Properties of Metal



Macro-properties depends on micro-structure

Micro-structure （μm）



Experiment in Spring-8

Development of New Material Compounds

Request for large-scale simulation from microscopic view

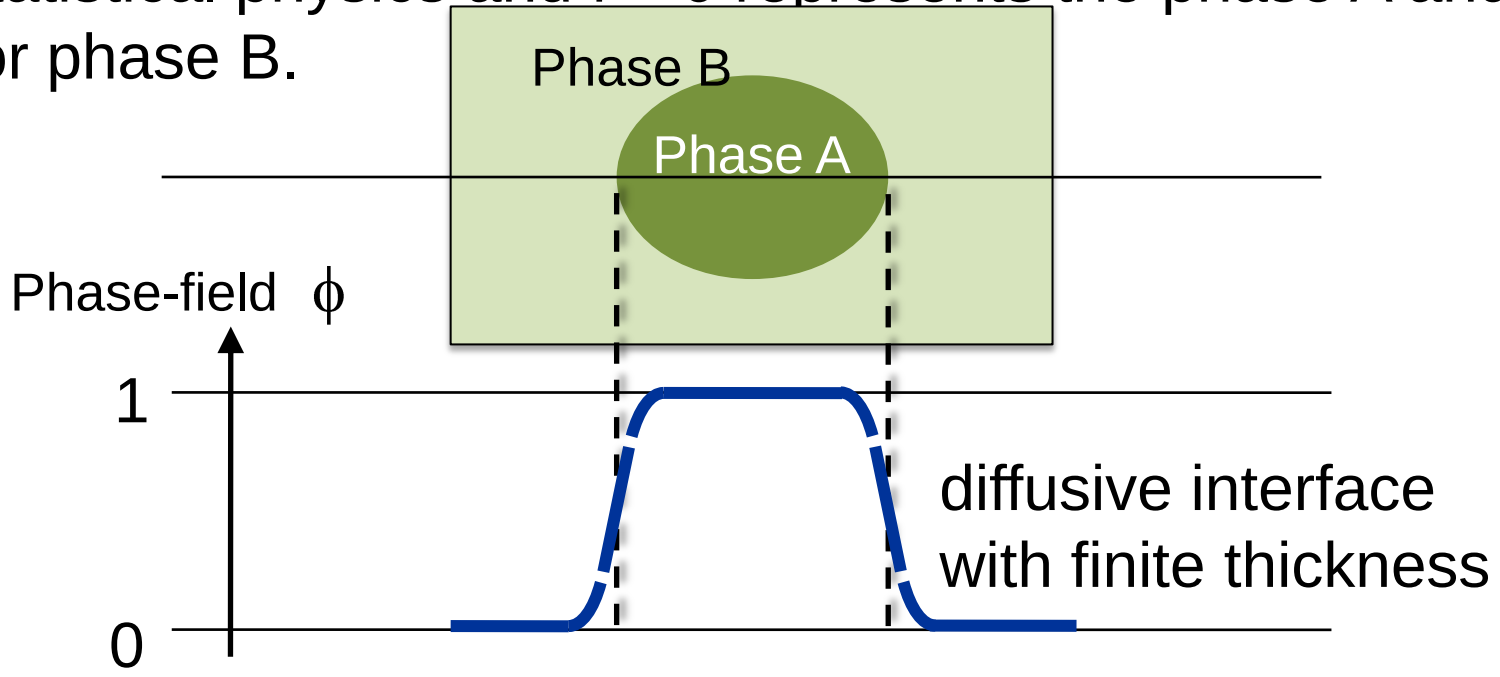
Large-scale GPU computing

Spring-8 (http://user.spring8.or.jp/sp8info/?p=17393)

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

Time evolution of the phase-field ϕ

（Allen-Cahn equation)

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

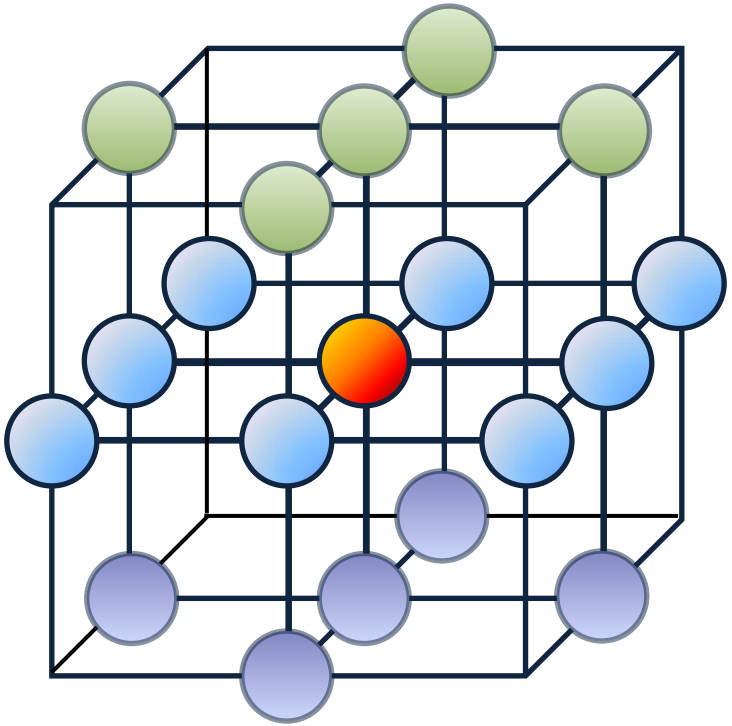
Time evolution of the condensation: c

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

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

Finite Difference Method

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



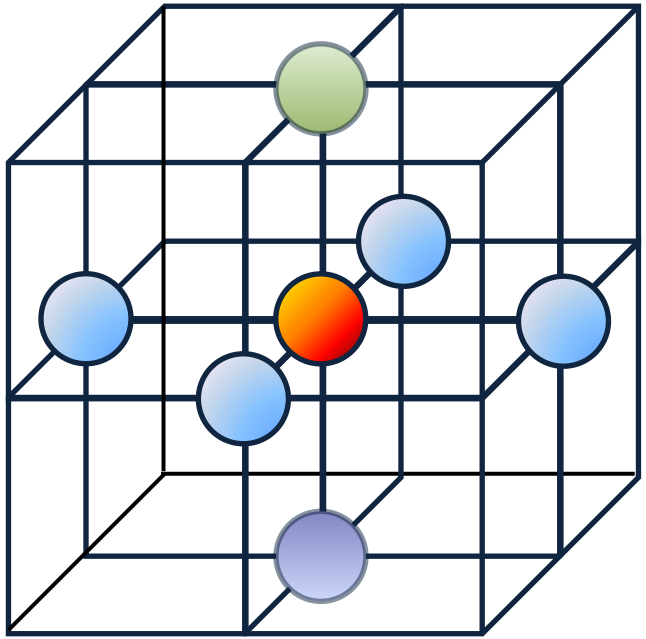
$z = k-1$

$z = k$

$z = k-1$

Condensation : $c_{i,j,k}$

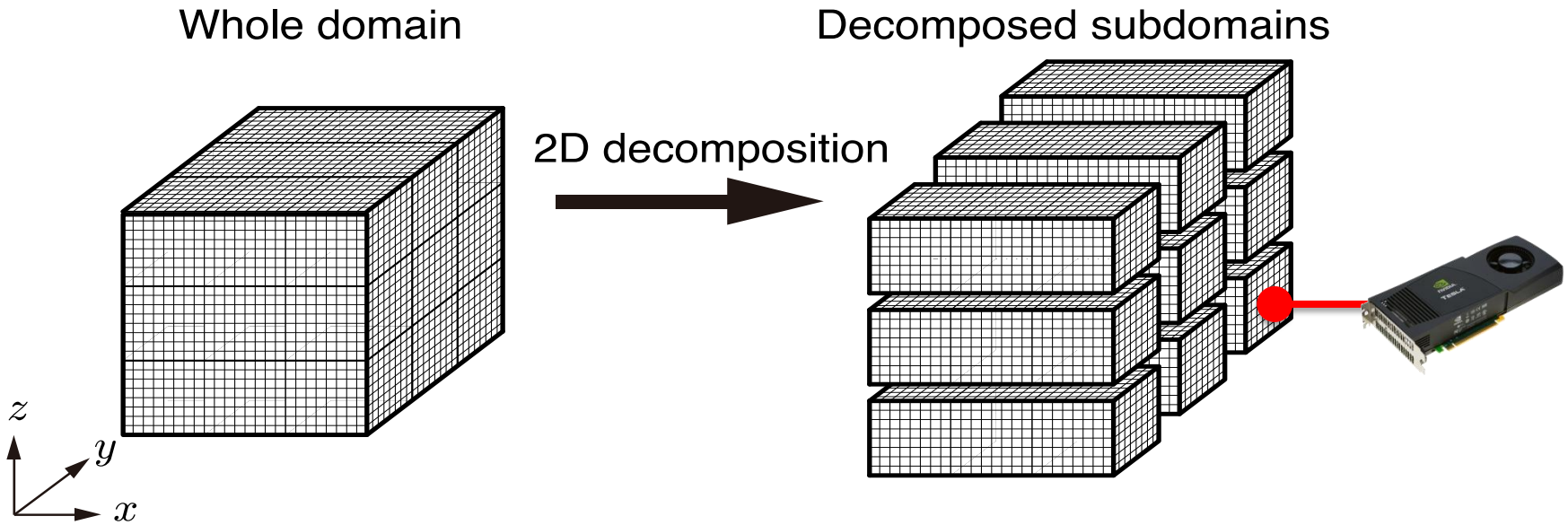
7 points to solve



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

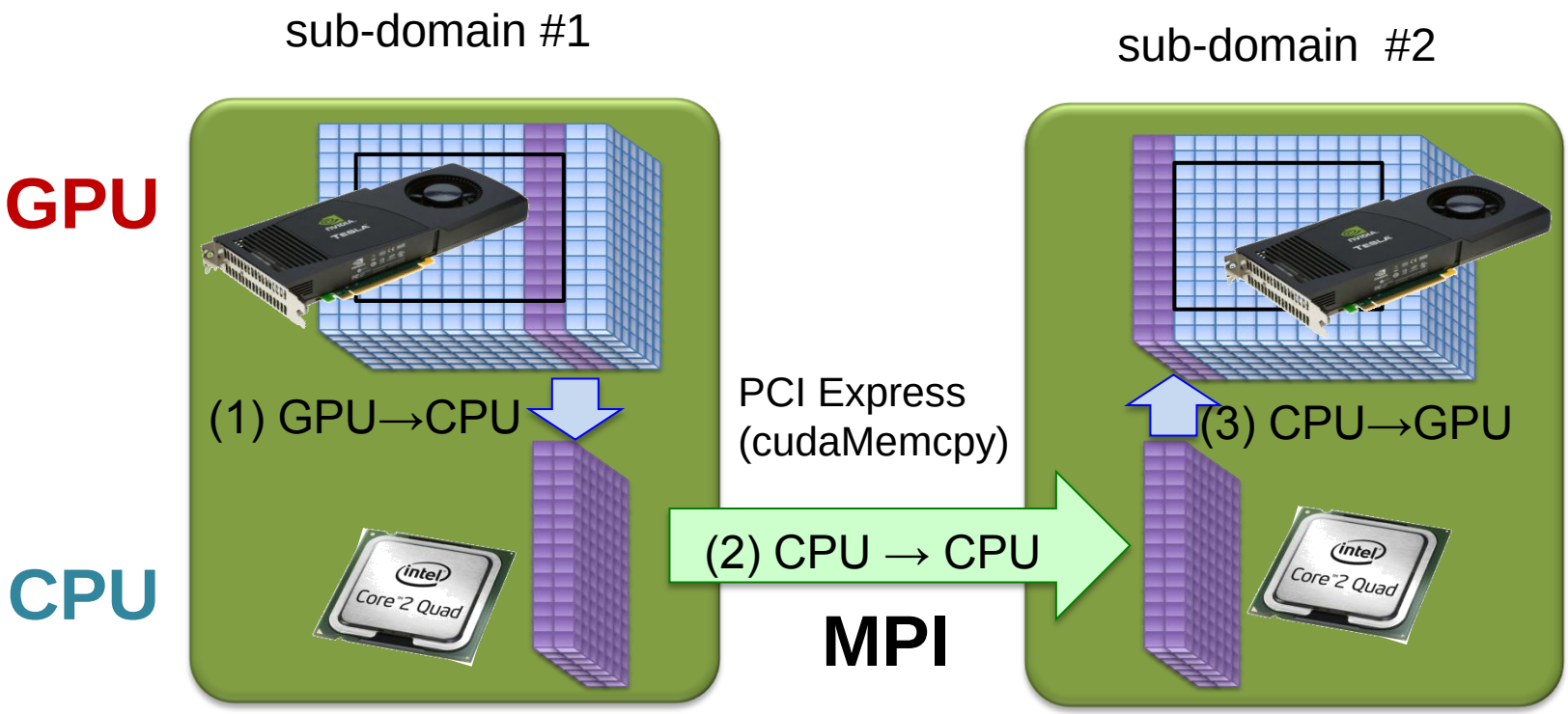
Multi-GPU Computing

- **2D Domain decomposition （in the y-, z-directions）**
  - ✓ 3D decomposition degrades GPU performance due to non-continuous memory access patterns for data exchange.



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

GPU-to-GPU Communication



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

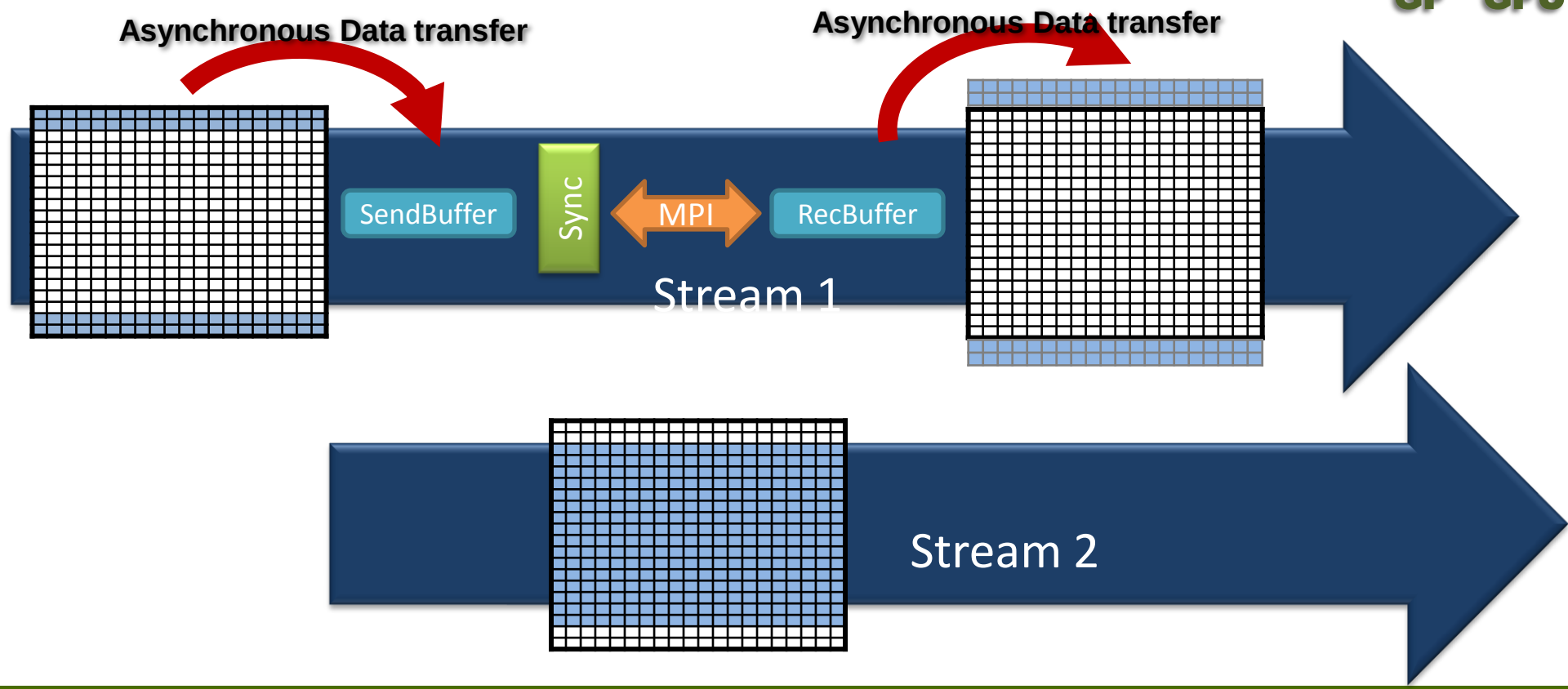
Multi-GPU Optimization

Typical explicit time integration

1. **GPU-only Method (without overlapping)**
2. **Hybrid-YZ Method**
3. **Hybrid-Y Method**

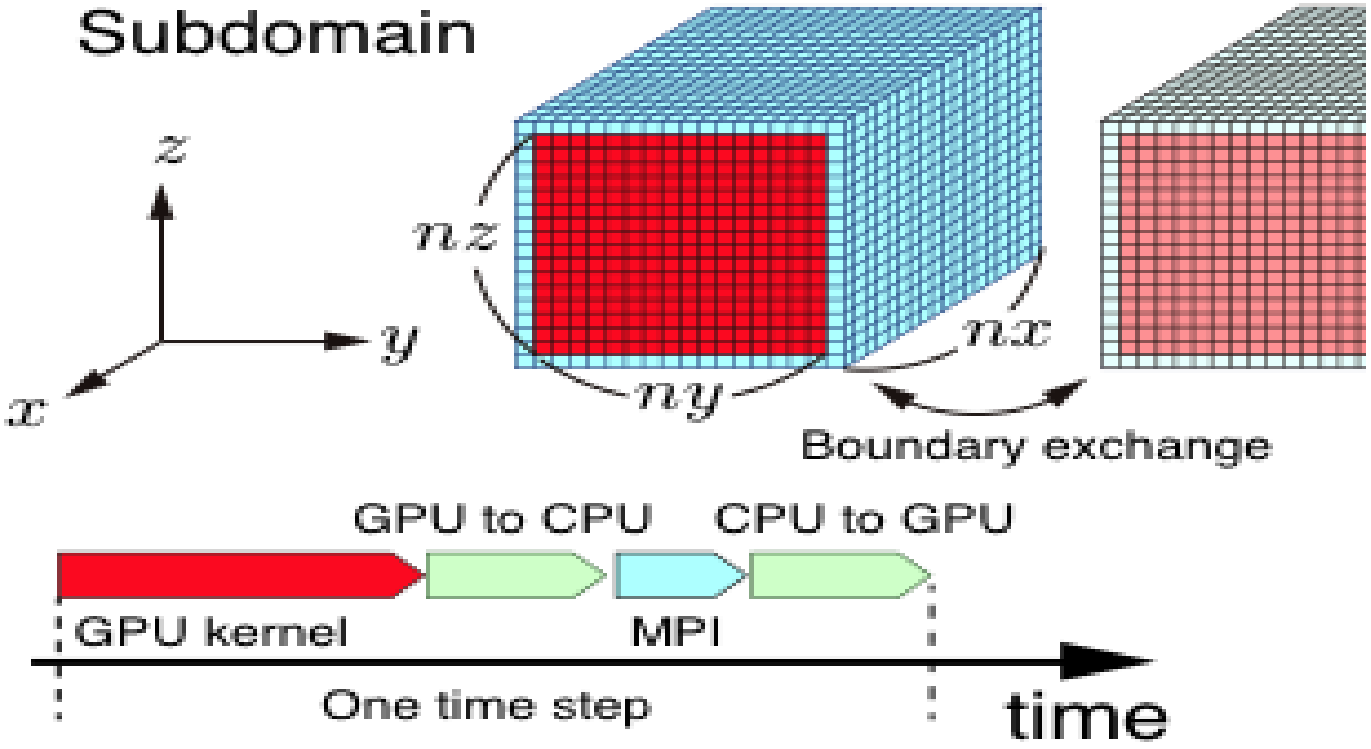
Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

Overlapping between Computation and Communication



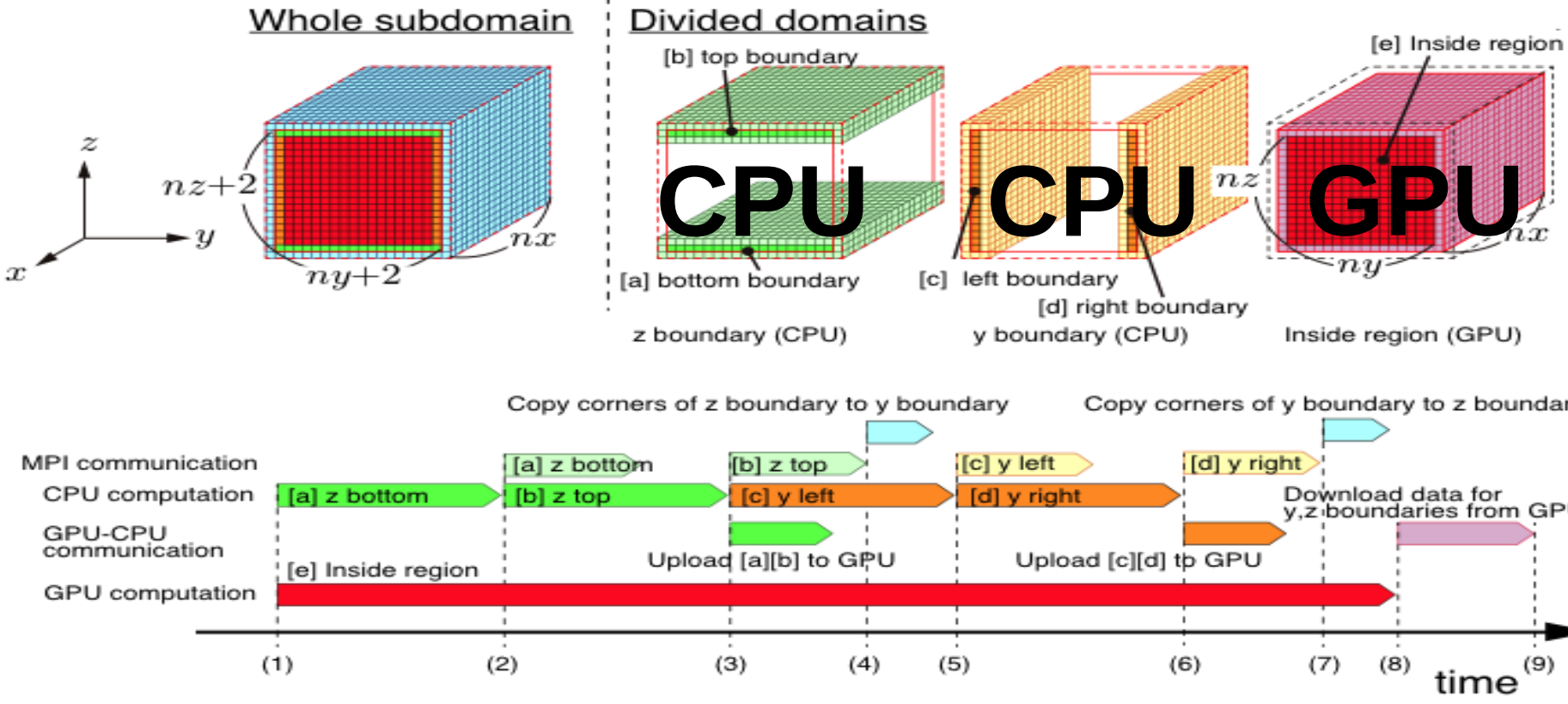
Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

1. GPU-only method



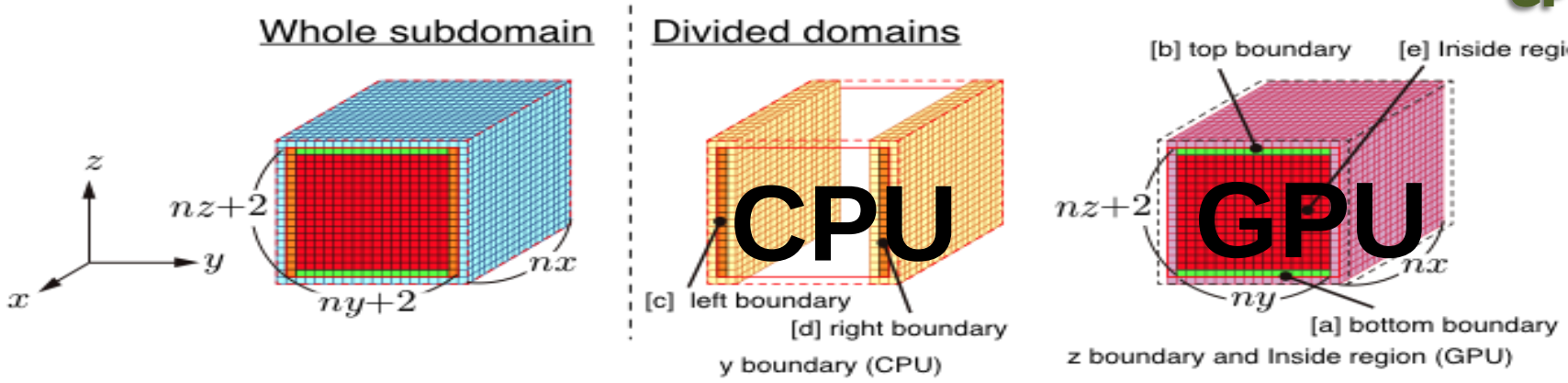
Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

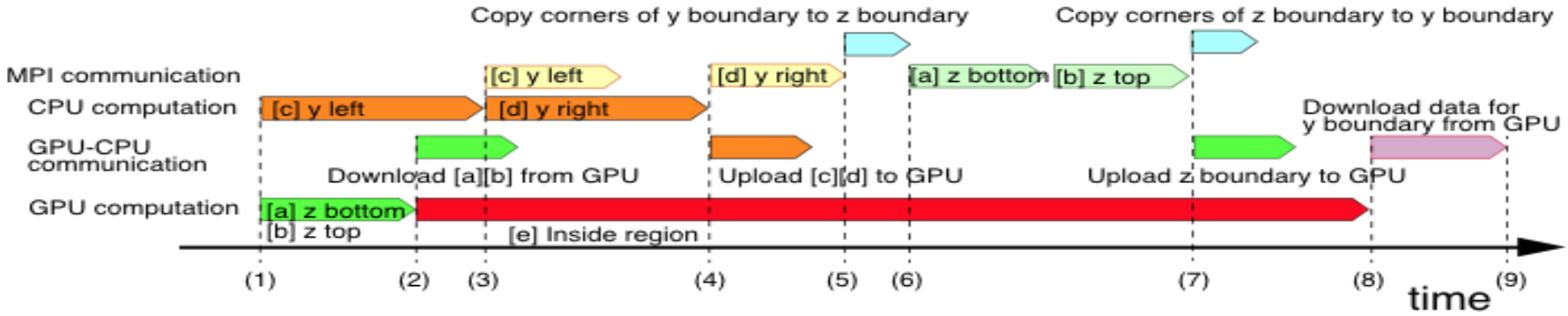
2. Hybrid-YZ method



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

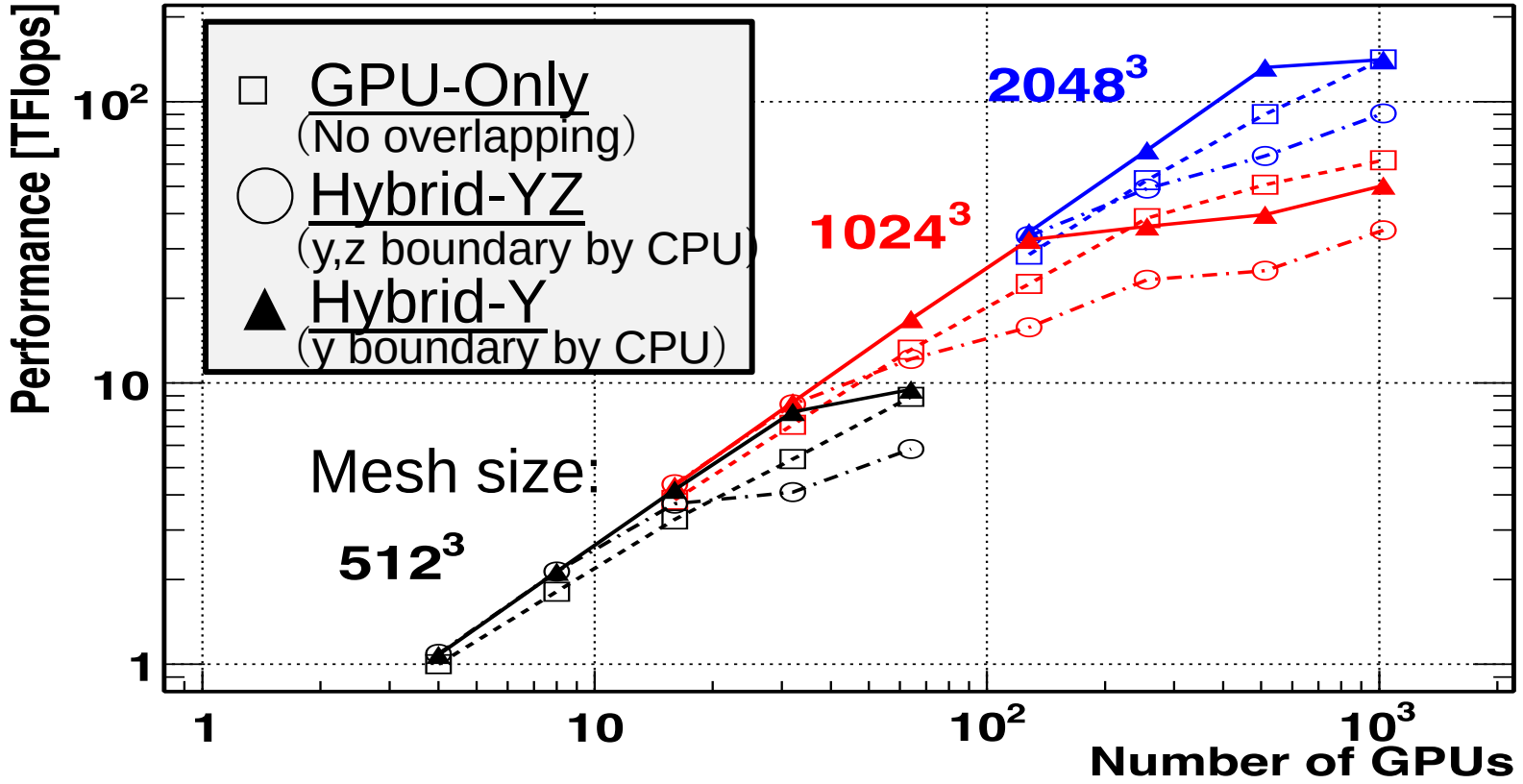
3. Hybrid-Y method





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

Strong Scalability

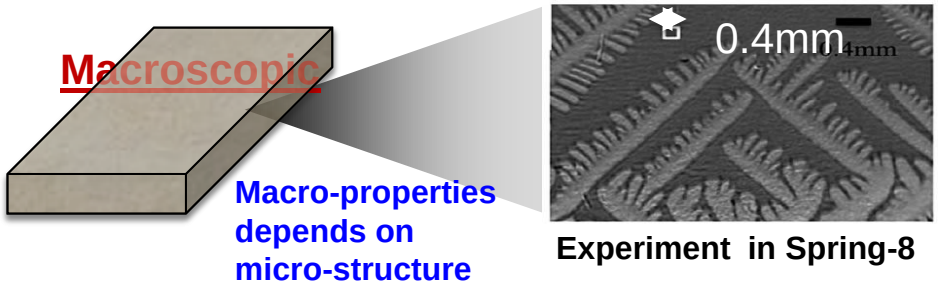


Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

Dendritic Solidification

GP GPU

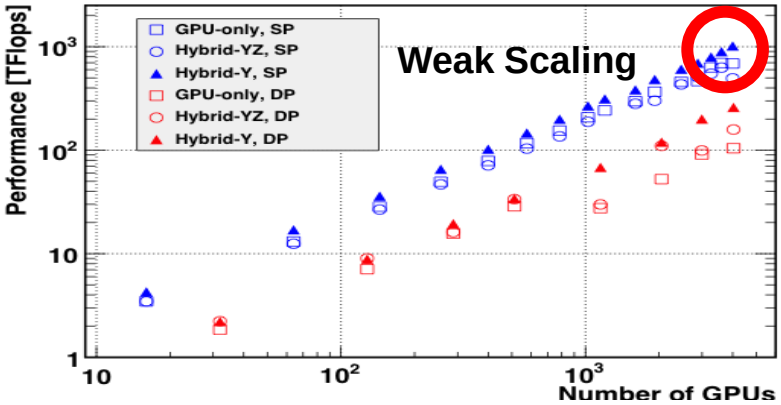
Micro-structure（μm）

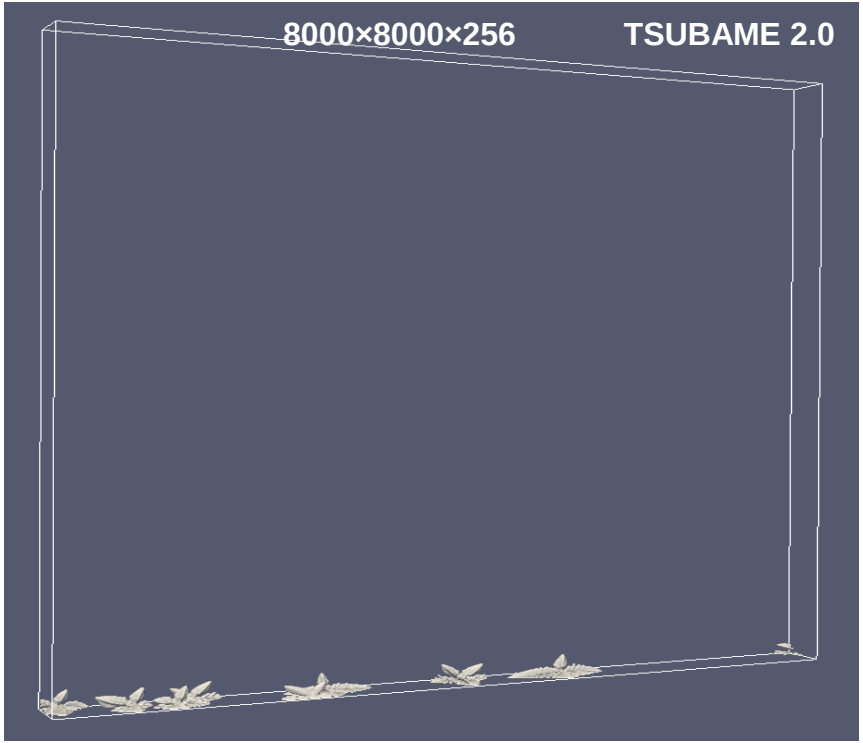


Macro-properties depends on micro-structure

Experiment in Spring-8

1.017 PFLOPS





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

Further Tuning from 1 PFlops to 2PFlops

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

Five Key Tunings

A. **CUDA 4.0** is used instead of CUDA 3.2.

B. Compiled with **–prec-div=false** and **–ftz=true**

- ✓ -prec-div=false uses a faster approximation division.
- ✓ -ftz=true flushes denormalized numbers to zero.

C. Host memory is allocated by **valloc**

- ✓ Allocated memory by **valloc** is aligned on a page boundary.
- ✓ Performance of MPI transfer is improved.

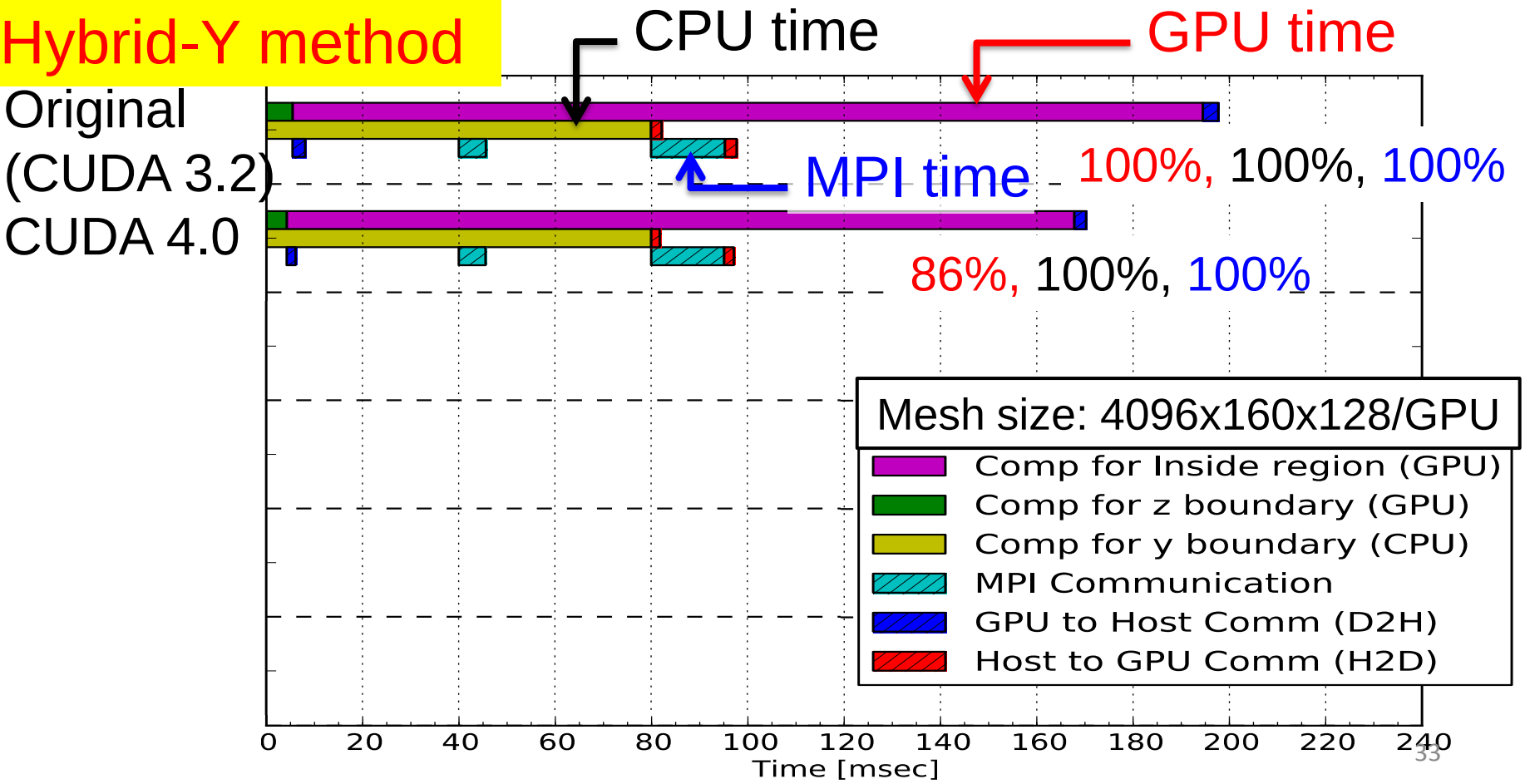
D. Compiled as **32-bit app.** instead of 64-bit app.

E. **SSE** is used for CPU computation.

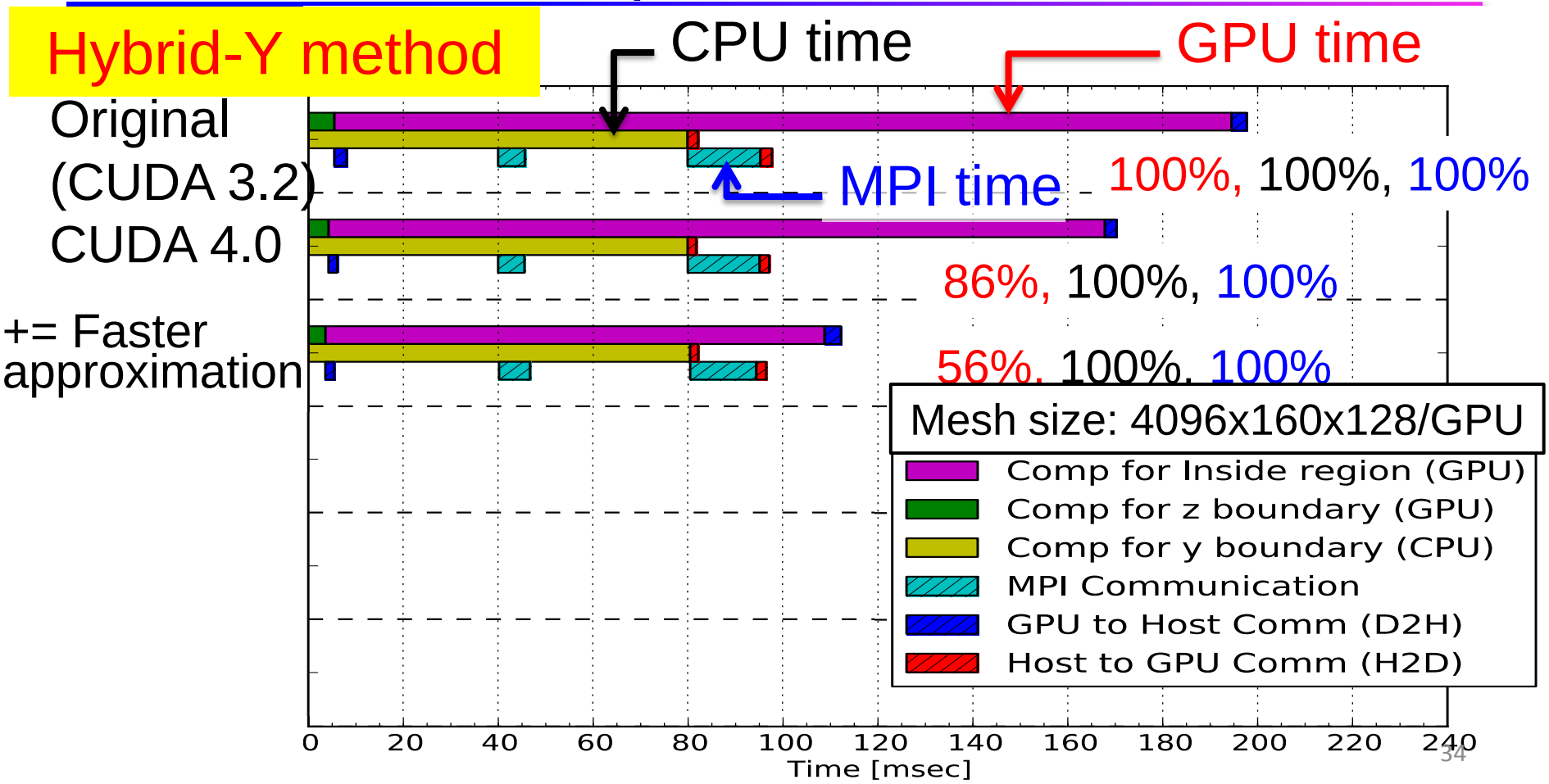
Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

Breakdown of Elapsed Time

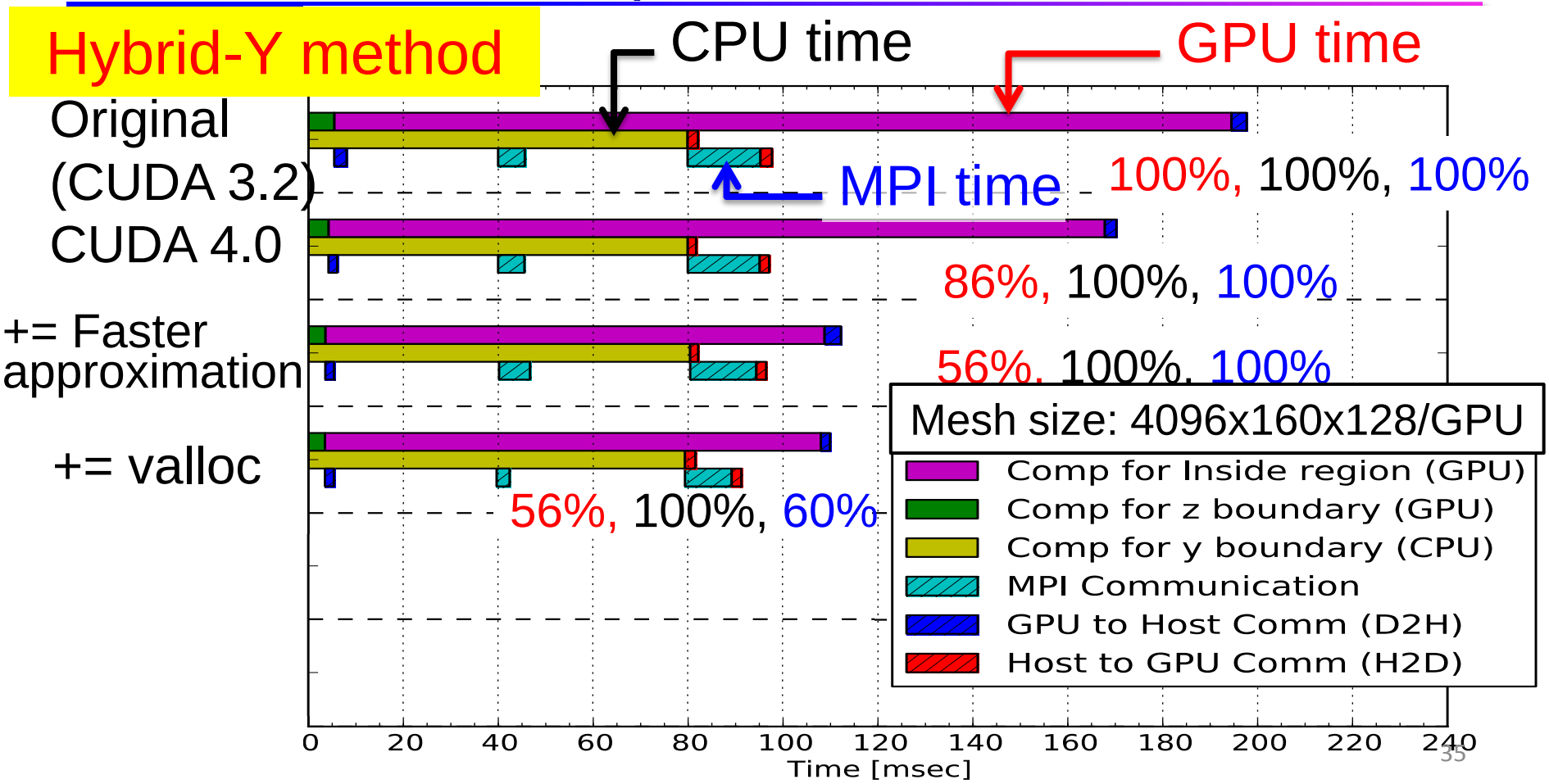
Hybrid-Y method



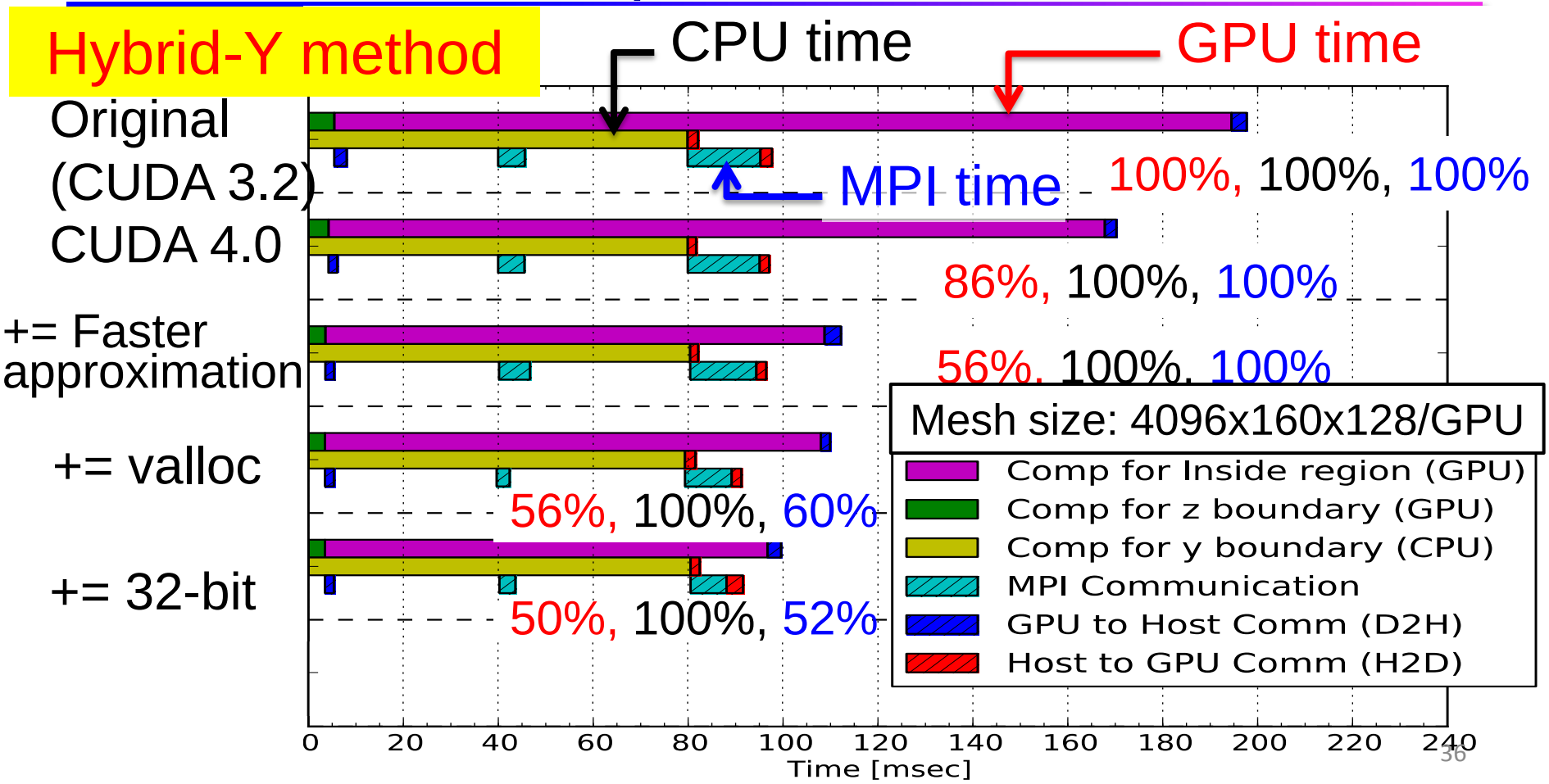
Breakdown of Elapsed Time



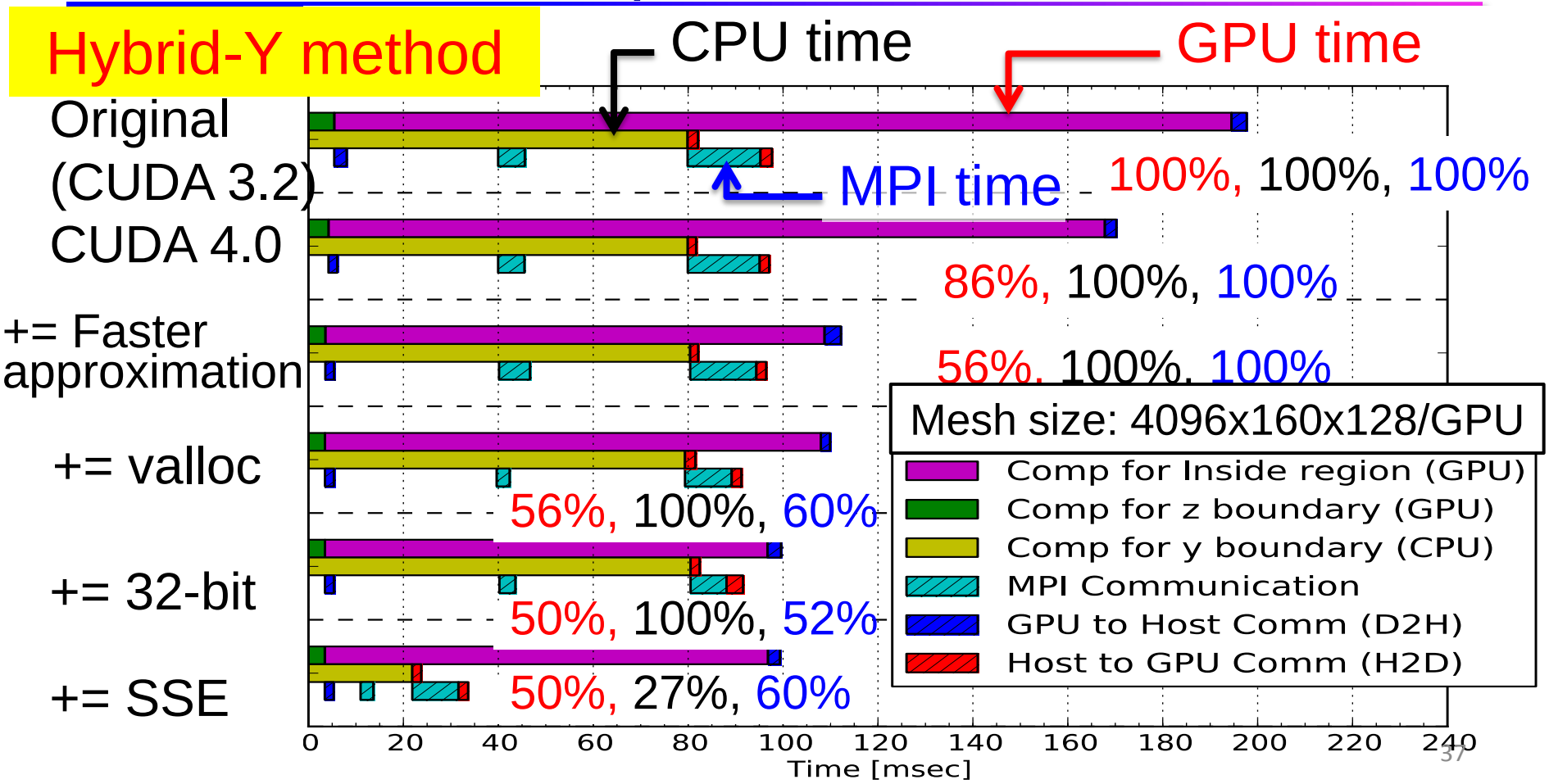
Breakdown of Elapsed Time



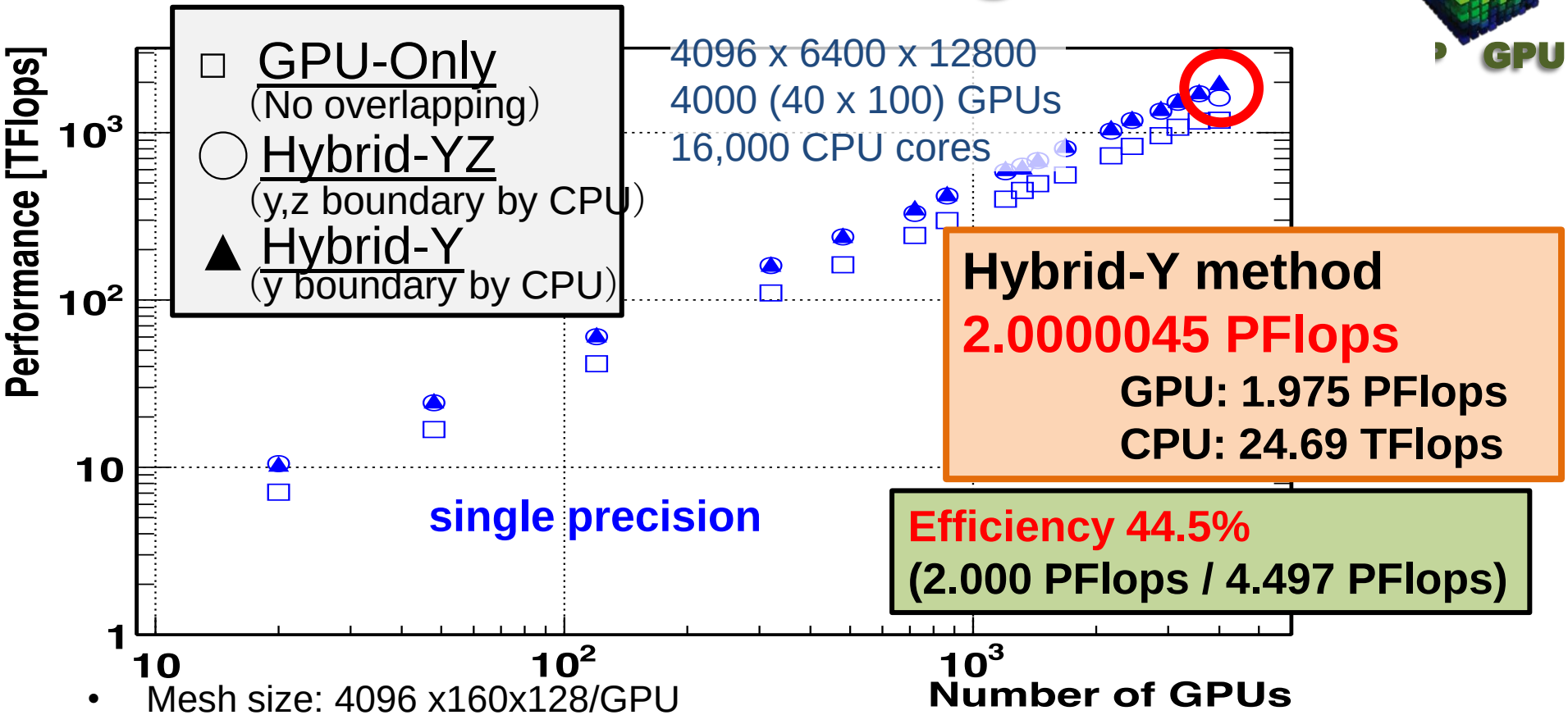
Breakdown of Elapsed Time



Breakdown of Elapsed Time



Weak scaling



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

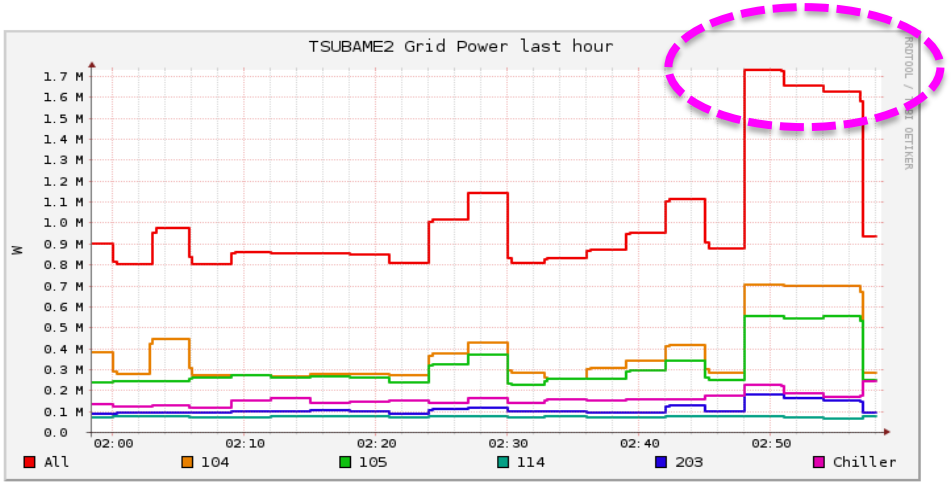
Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

Power Efficiency

- The power consumption by application 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**

~1.36 MW

Simulation results by much less electric power than before.



Ref. **Linpack**

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

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

SUMMARY

- 2-Petaflops performance has been achieved for the Phase-Field simulation on GPU-based supercomputer TSUBAME 2.0

- Extremely high-performance 44.5 % of the peak for a mesh-based practical application

- Green computing : much less electrical power to get meaningful application results

Gordon Bell Award
Finalist !
SC'11 in Seattle

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

Thank you

for your kind attention

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