GTC Asia in Beijing 2011, Dec 15

A 2-Petaflops Stencil Application

on GPU-rich Supercomputer TSUBAME 2.0

Takayuki Aoki

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

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)

Supercomputer in the world

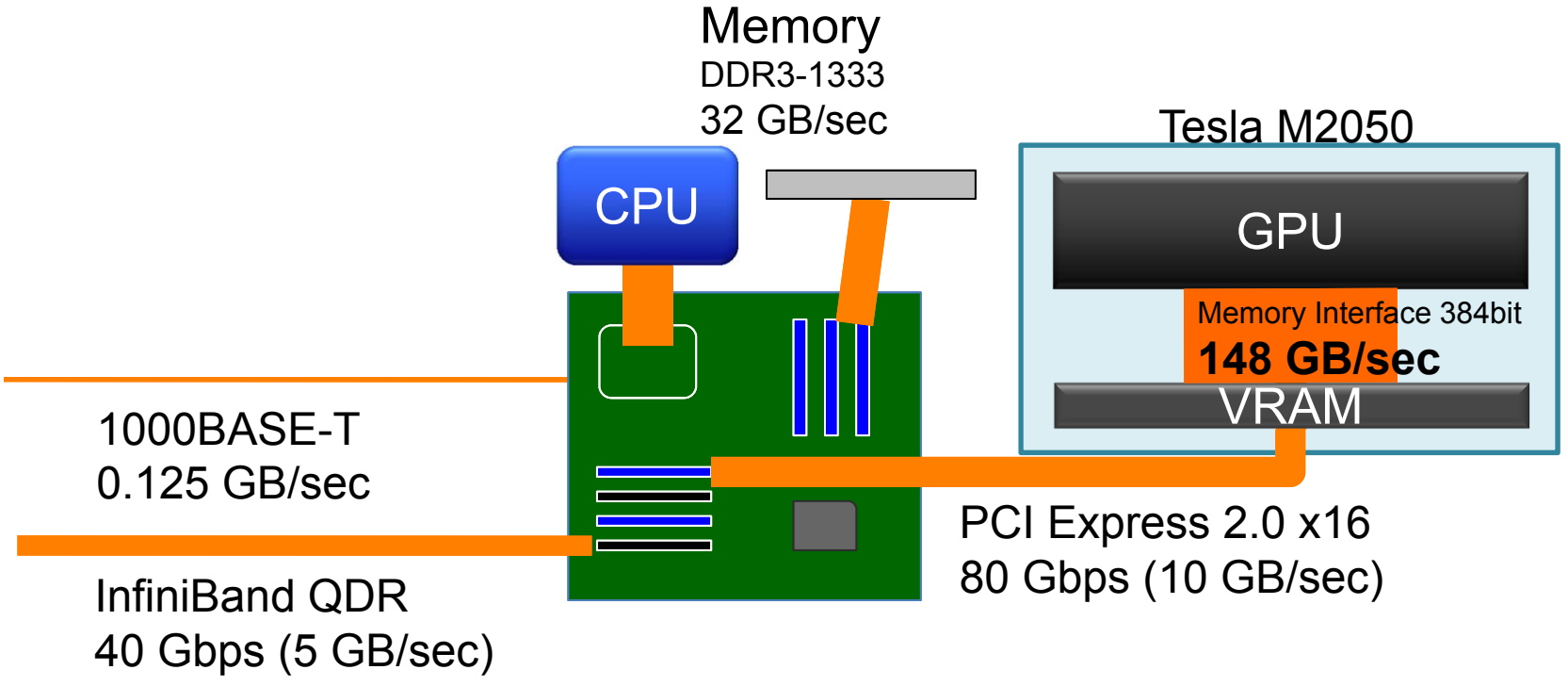
2011 November

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	120640	1271.00	2984.30	2580.0
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

Heterogeneous Computer

AOKI Lab.

- **Several Bandwidth Bottle Necks**



Next Generation

Weather Prediction

AOKI Lab.

Collaboration: Japan Meteorological Agency

Meso-scale Atmosphere Model:

Cloud Resolving Non-hydrostatic model

Compressible equation taking consideration of sound waves.

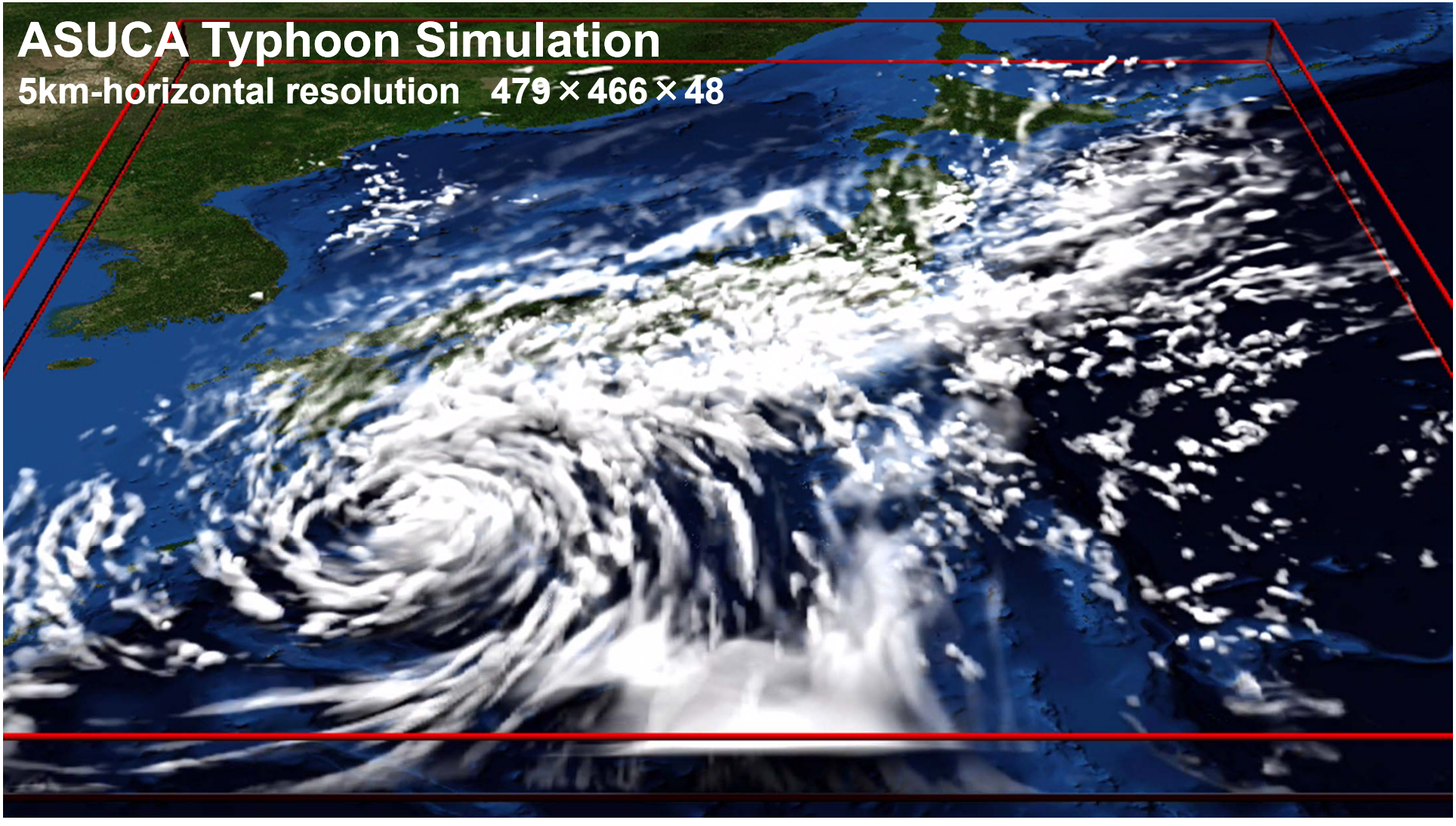
Meso-scale

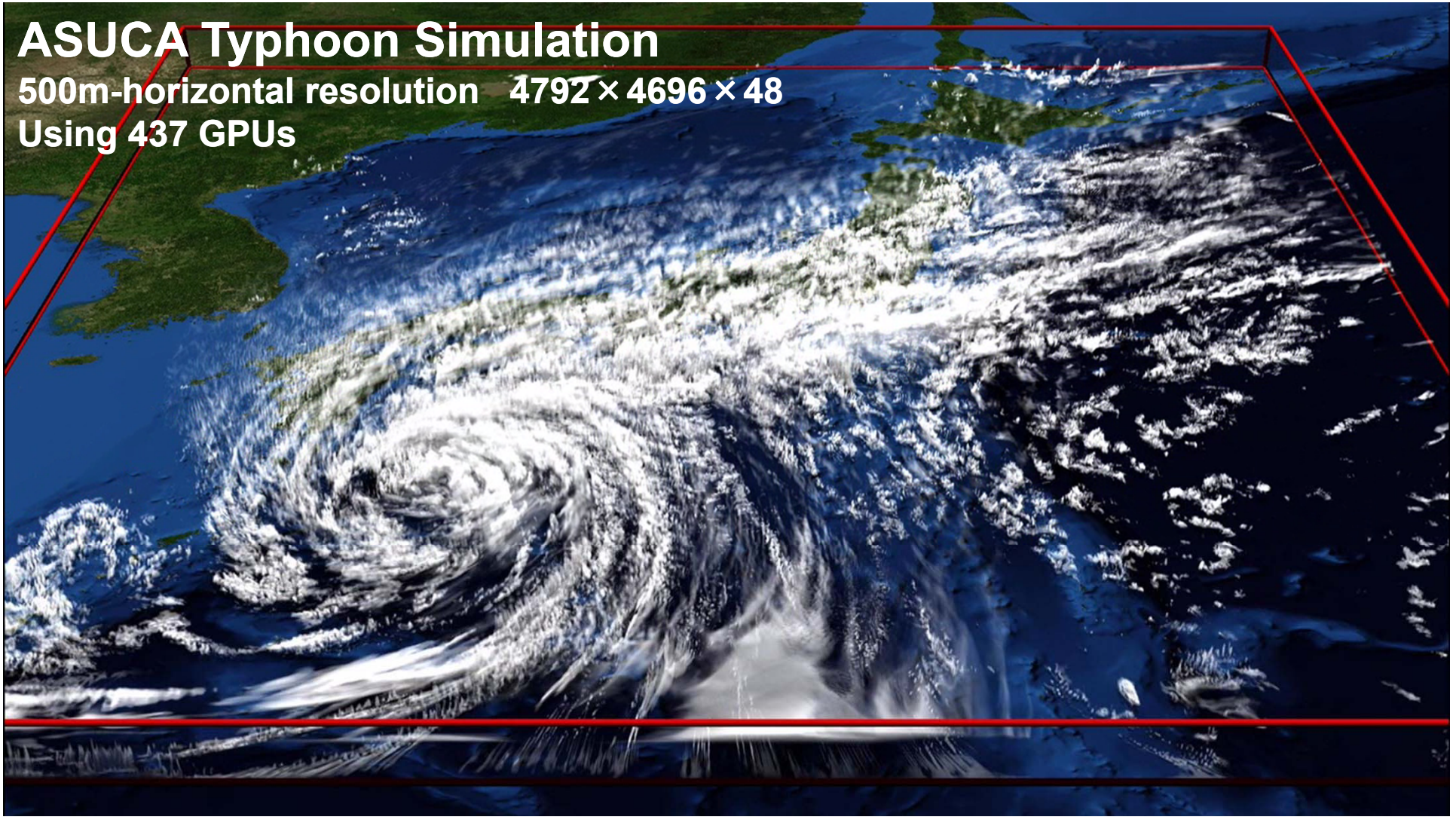
2000 km

Typhoon

a few km

Tornado, Down burst
Heavy Rain


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

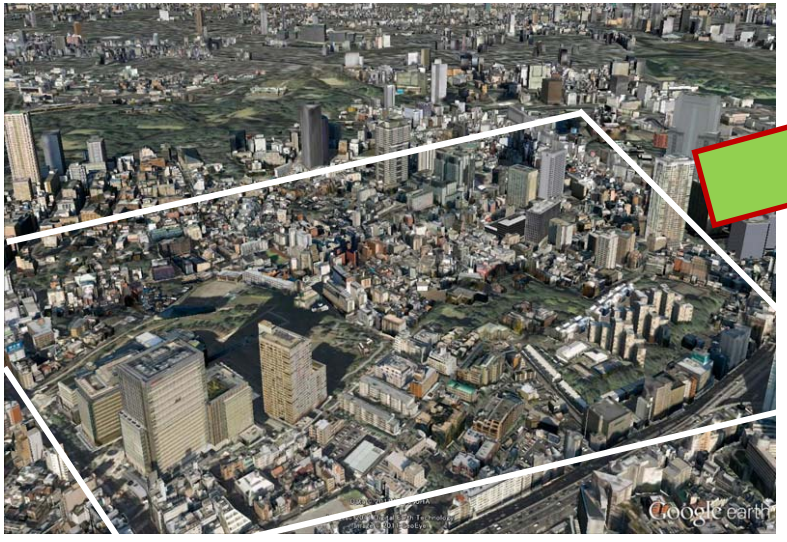

ASUCA Typhoon Simulation
500m-horizontal resolution 4792×4696×48
Using 437 GPUs

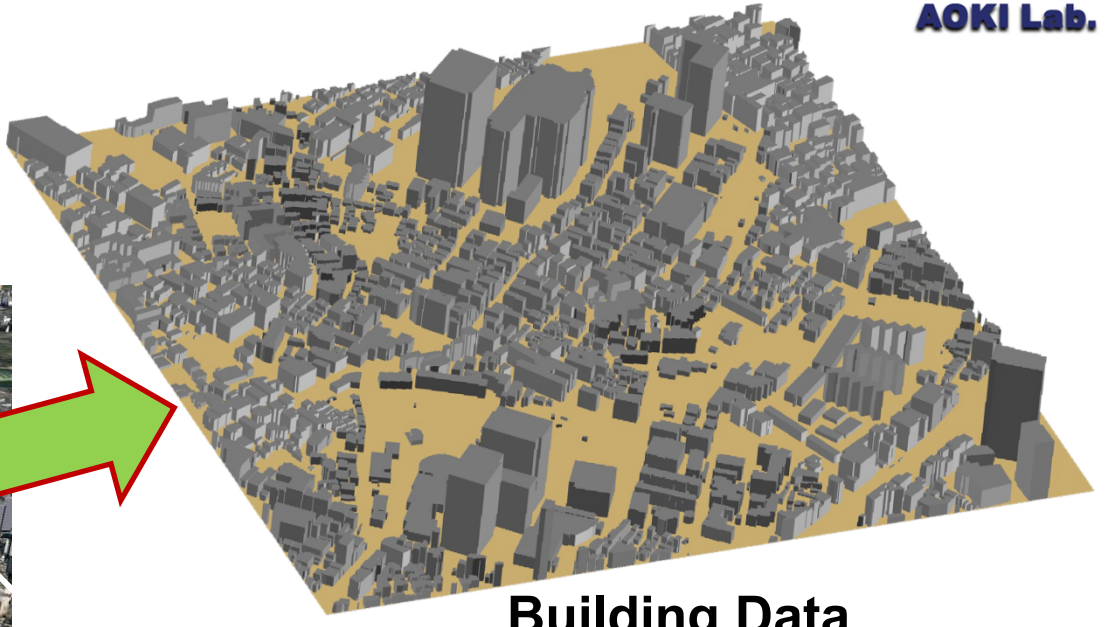
Real City Computation

AOKI Lab.

Tokyo 六本木 Area

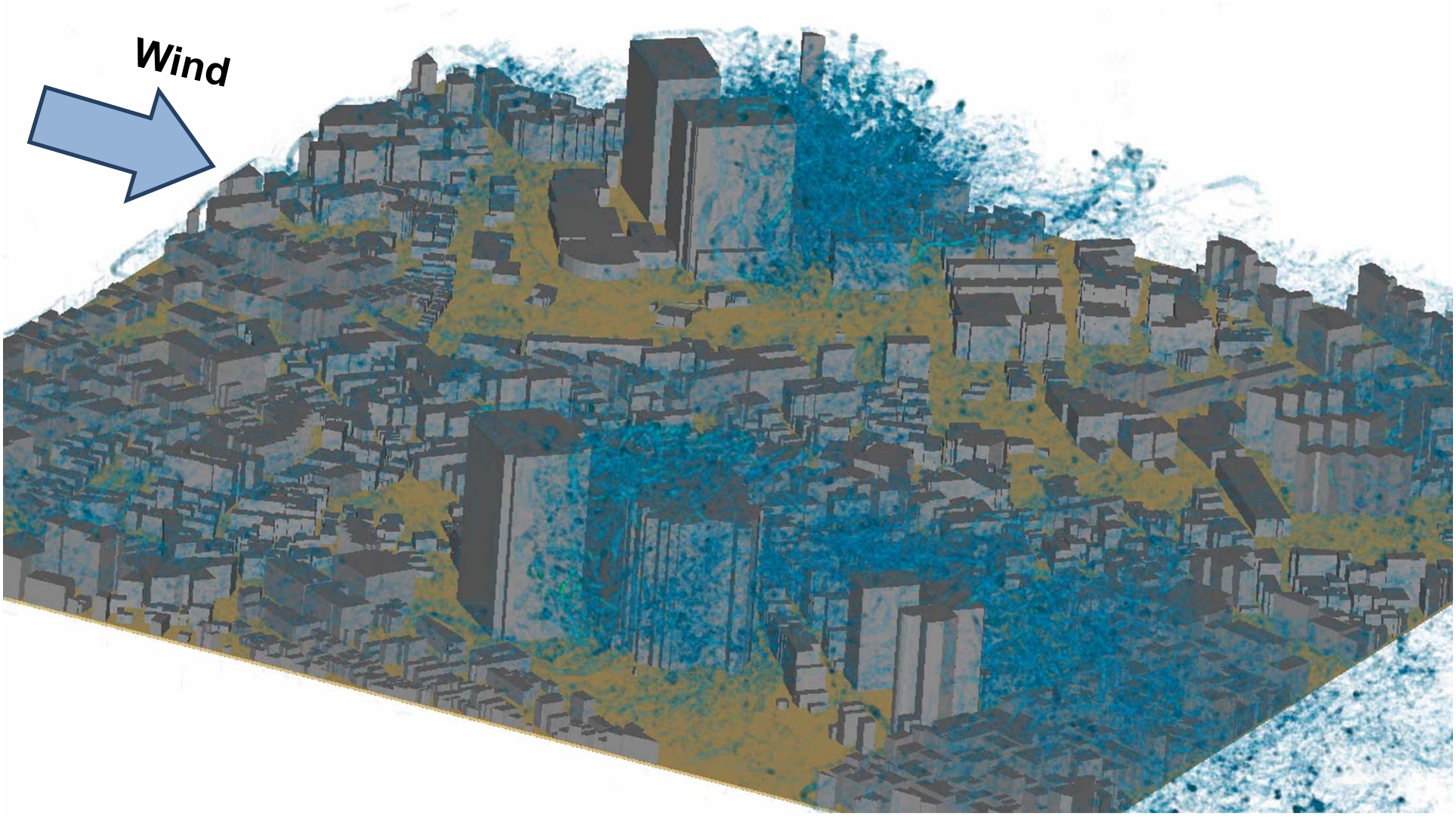
1km x 1 km



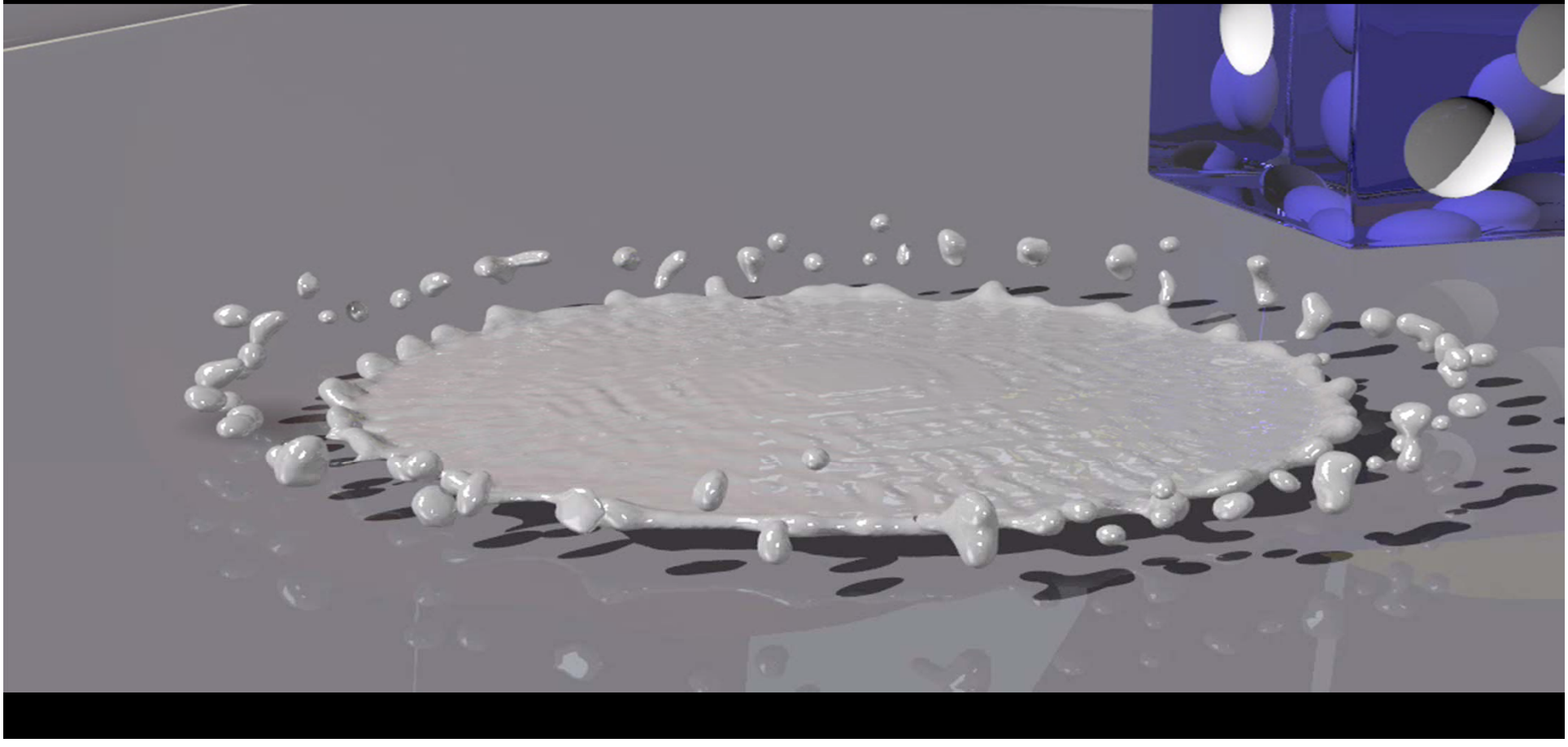


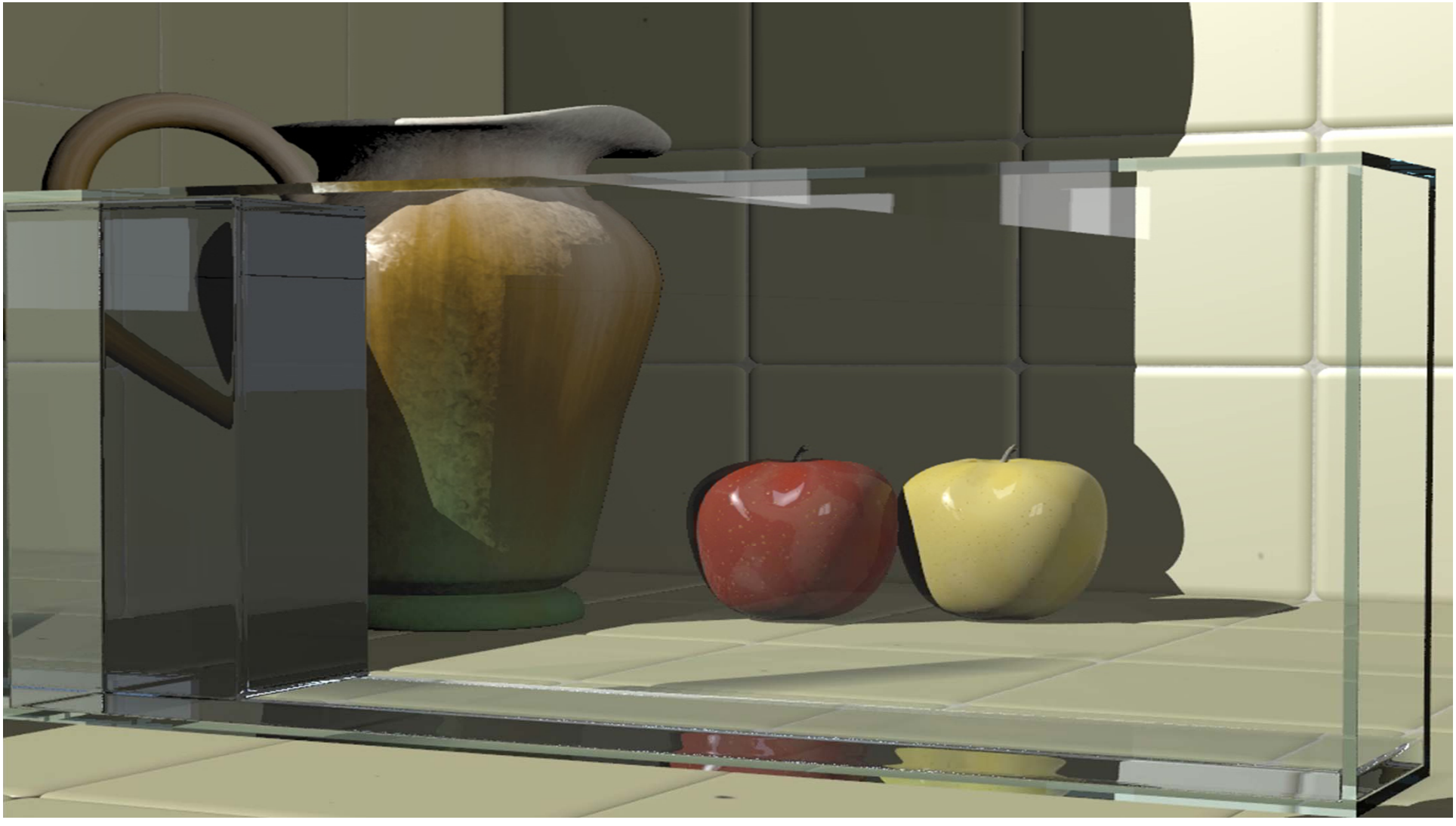
Building Data

1-m resolution 1000x1000x300


Wind

Drop on dry floor



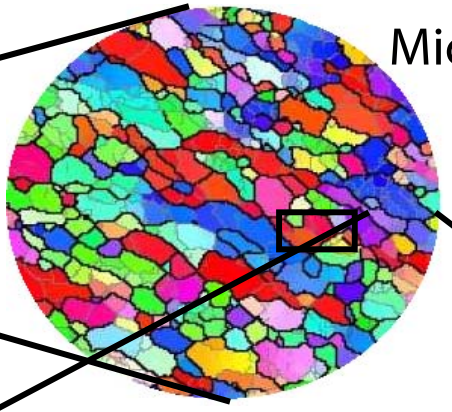


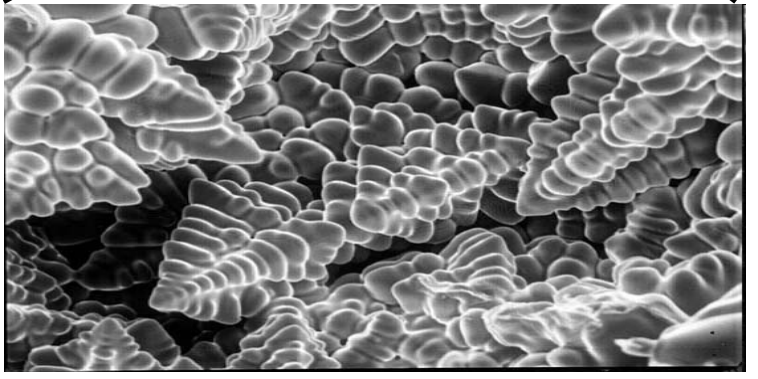
Development of New Materials

AOKI Lab.

Mechanical Structure

Microstructure





Low-carbon society

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

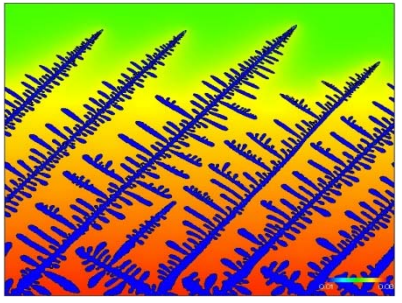
Developing lightweight strengthening material by controlling microstructure

Impact on Material Science

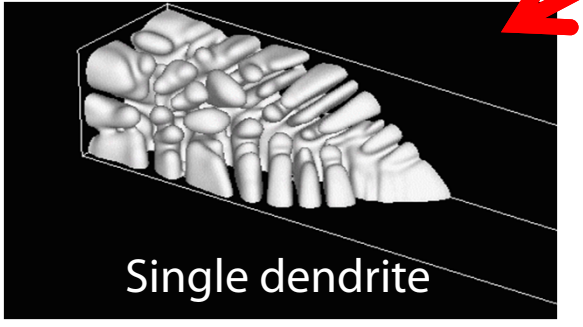
AOKI Lab.

Previous Research

2D

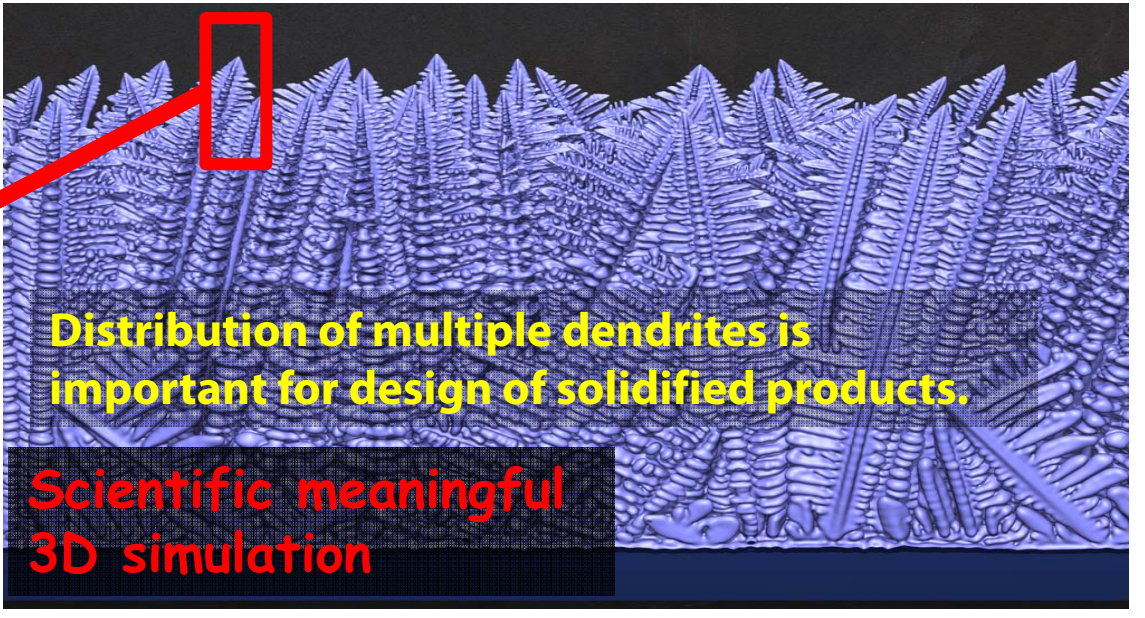


3D simple shape



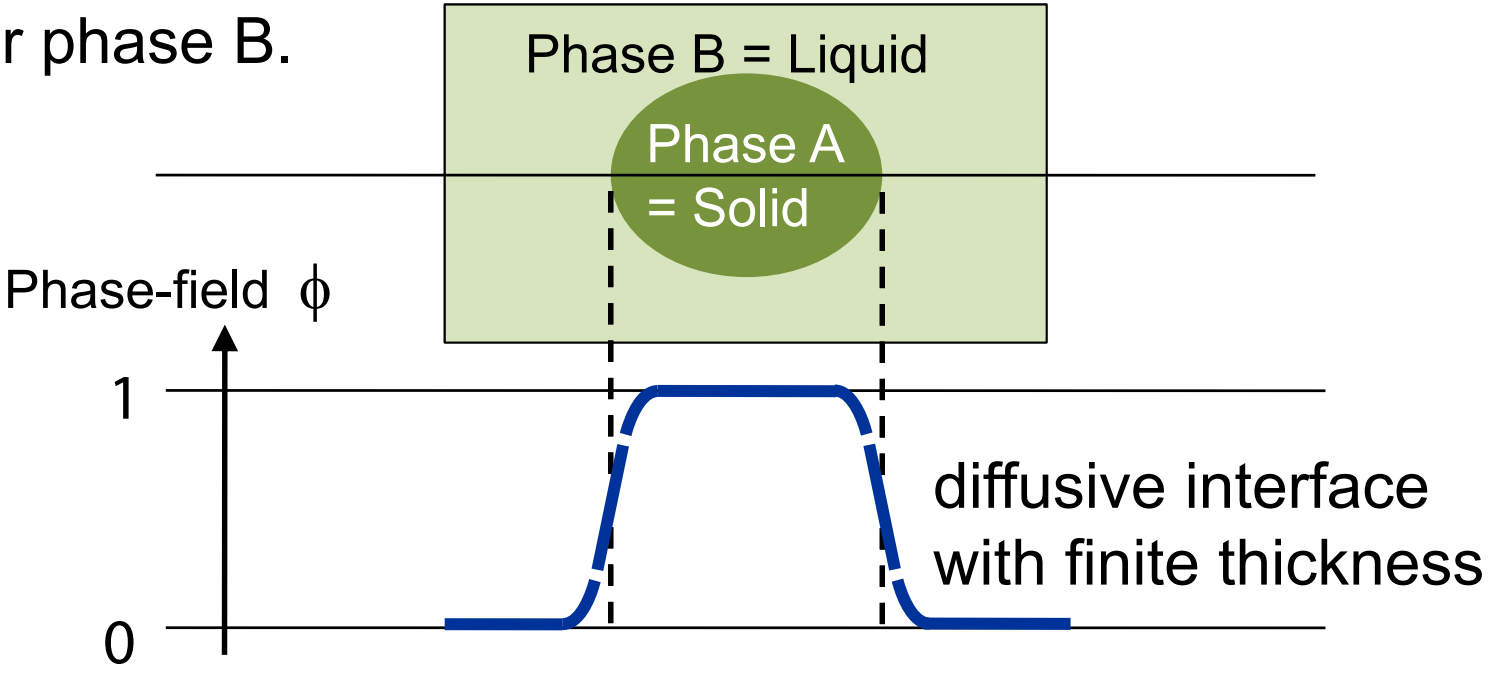
Peta-scale Simulation

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



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.



Al-Si: Binary Alloy

AOKI Lab.

Time evolution of the phase-field ϕ

（Allen-Cahn equation)

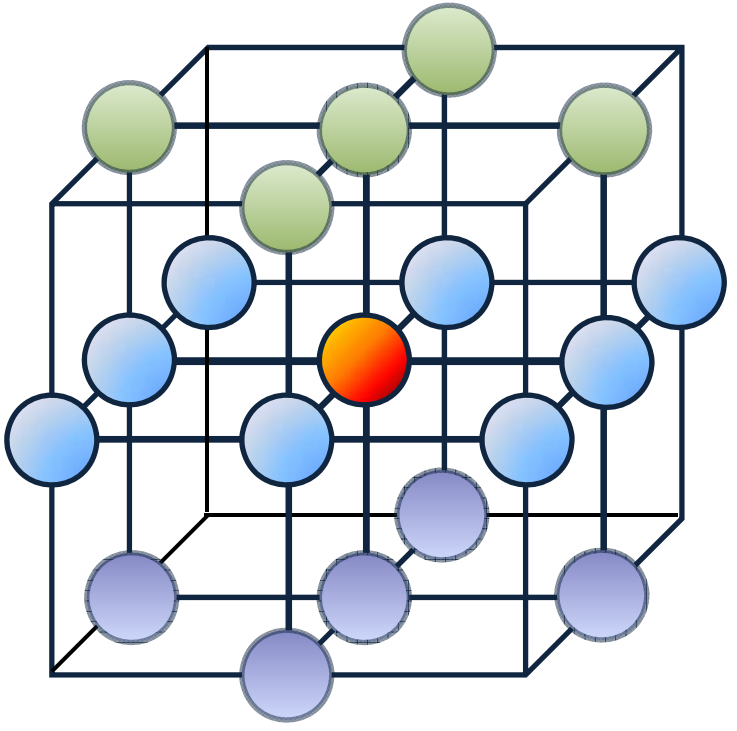
$$\frac{\partial \phi}{\partial t} = M_\phi \left[\nabla \cdot (a^2 \boldsymbol{\nabla} \phi) + \frac{\partial}{\partial x} \left(a \frac{\partial a}{\partial \phi_x} |\boldsymbol{\nabla} \phi|^2 \right) + \frac{\partial}{\partial y} \left(a \frac{\partial a}{\partial \phi_y} |\boldsymbol{\nabla} \phi|^2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

Time evolution of the condensation: c

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

Finite Difference Method

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



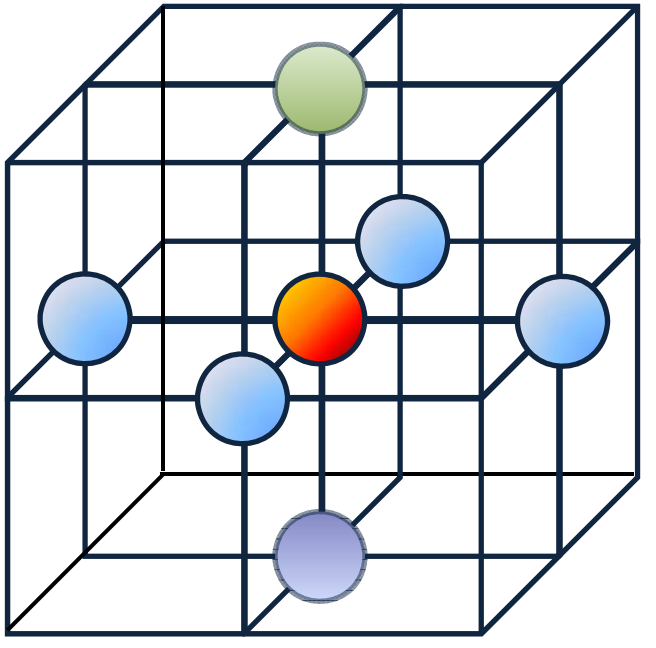
$z = k-1$

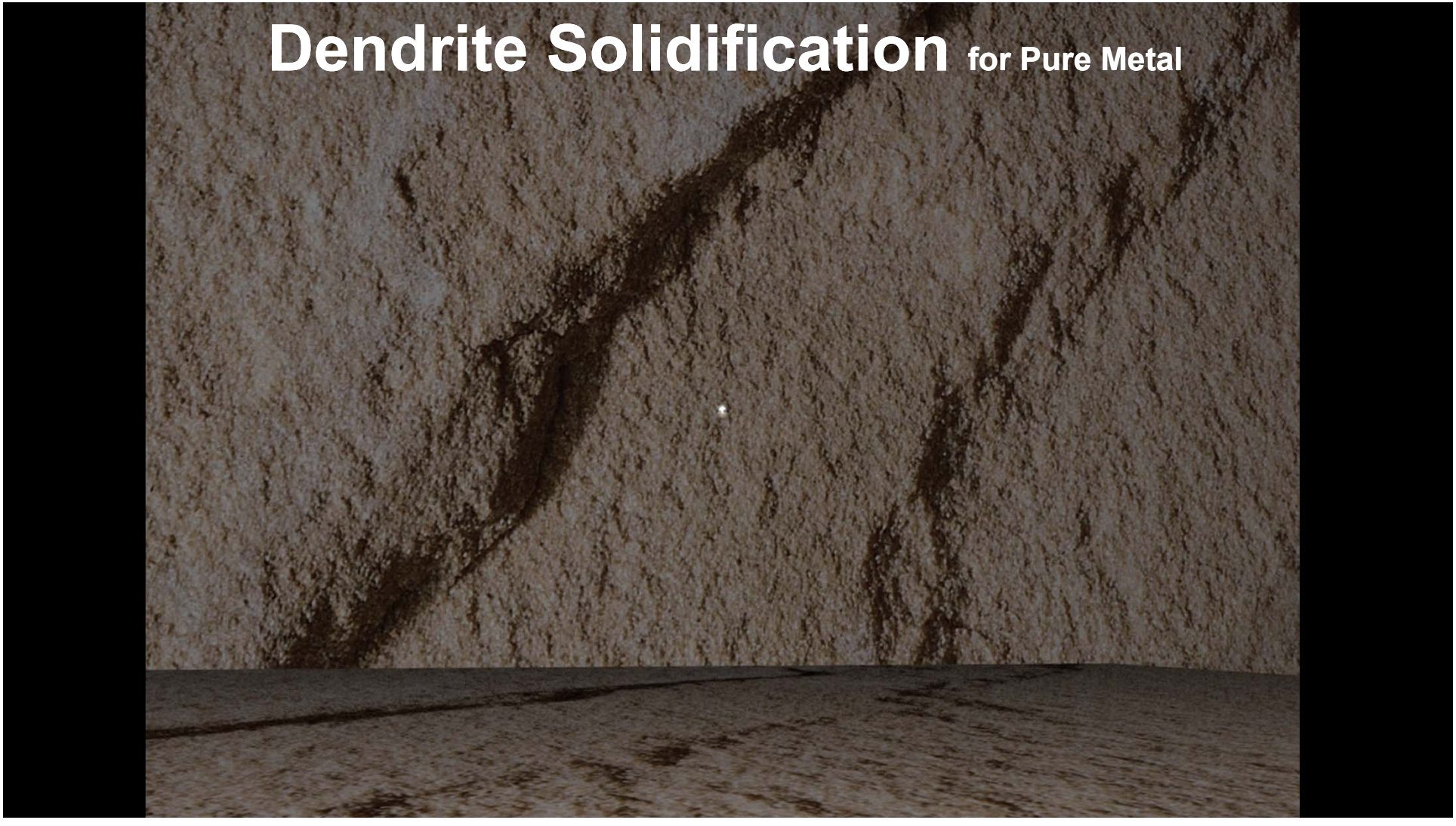
$z = k$

$z = k-1$

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

7 points to solve

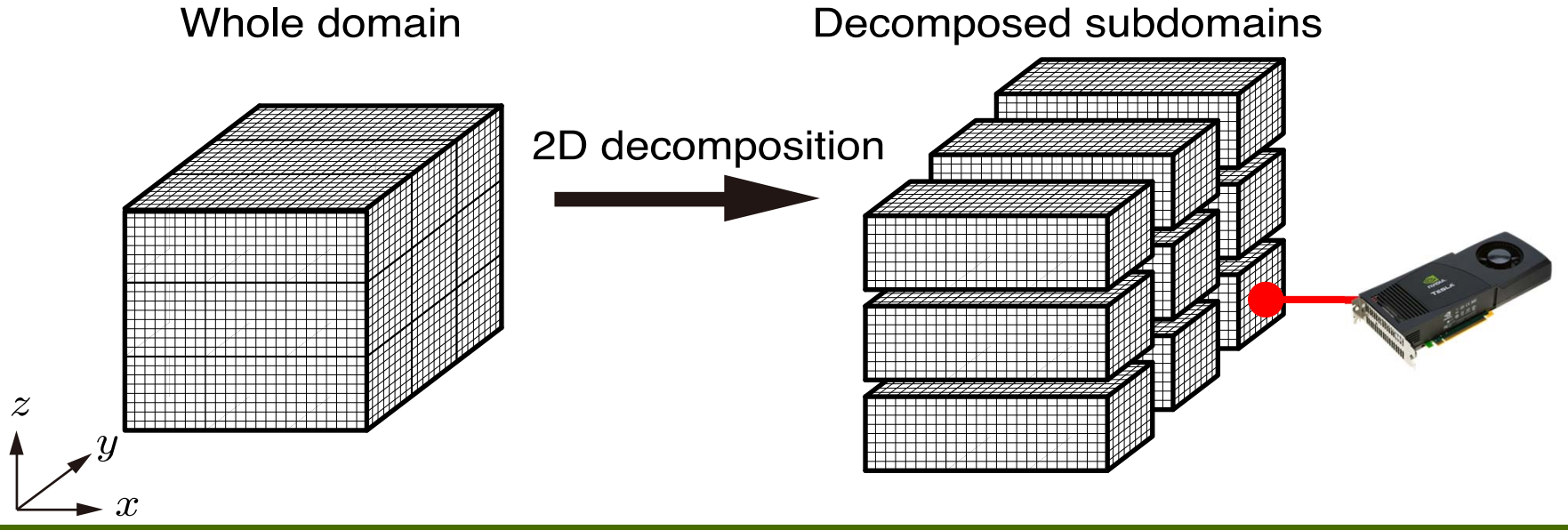



Dendrite Solidification for Pure Metal

Multi-GPU Computing

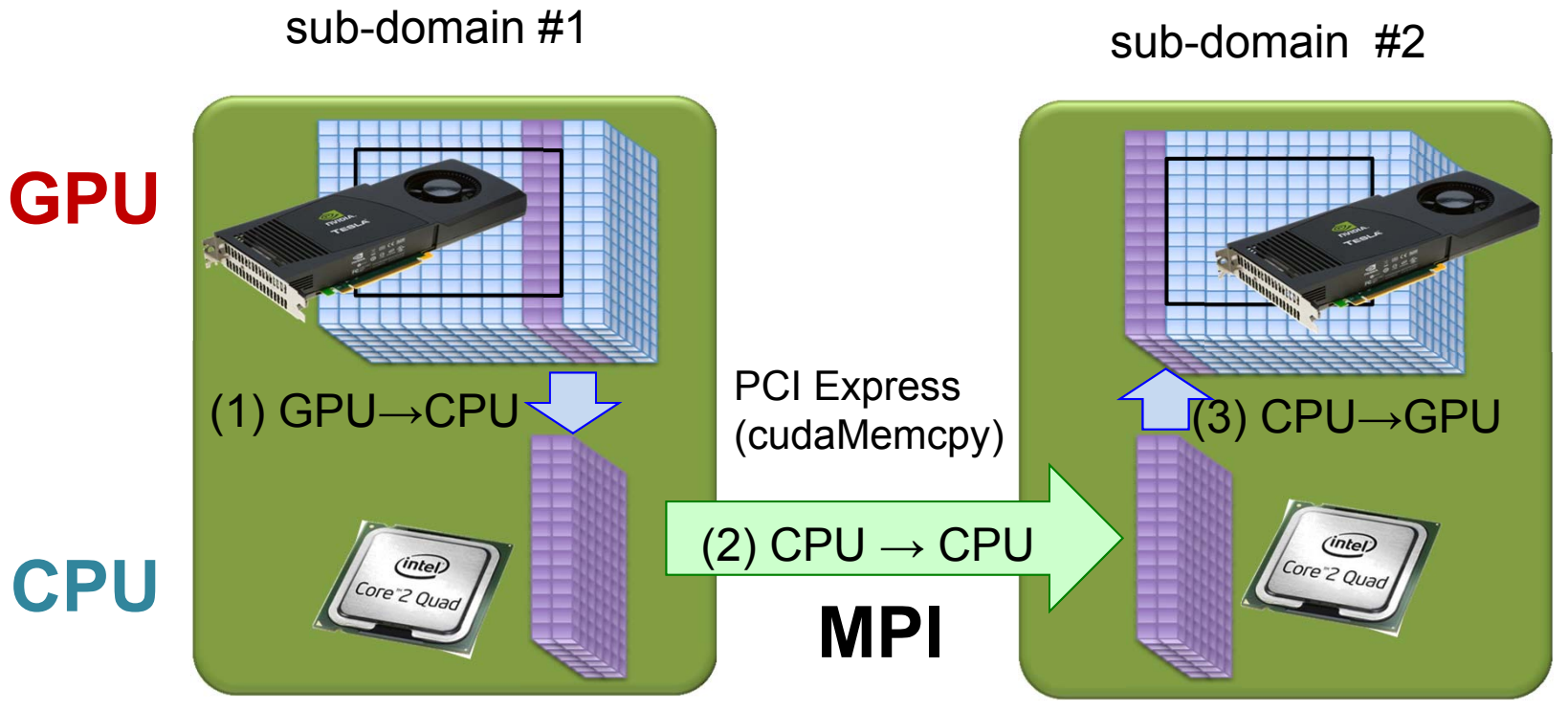
AOKI Lab.

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

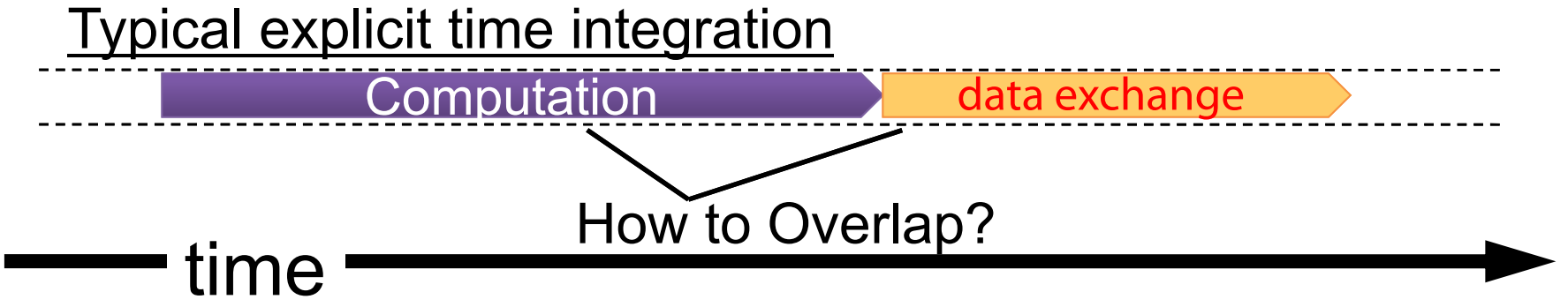


GPU-to-GPU Communication

AOKI Lab.

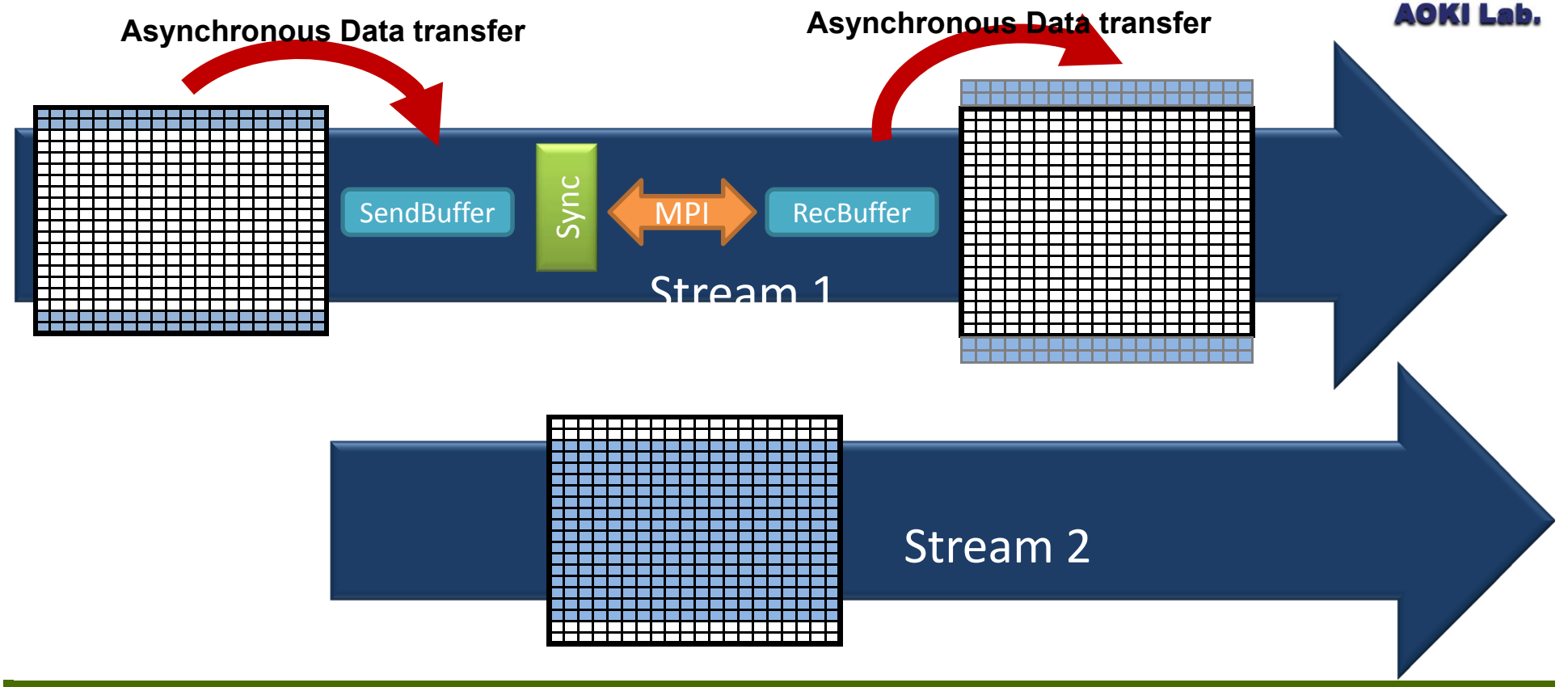


Multi-GPU Optimization

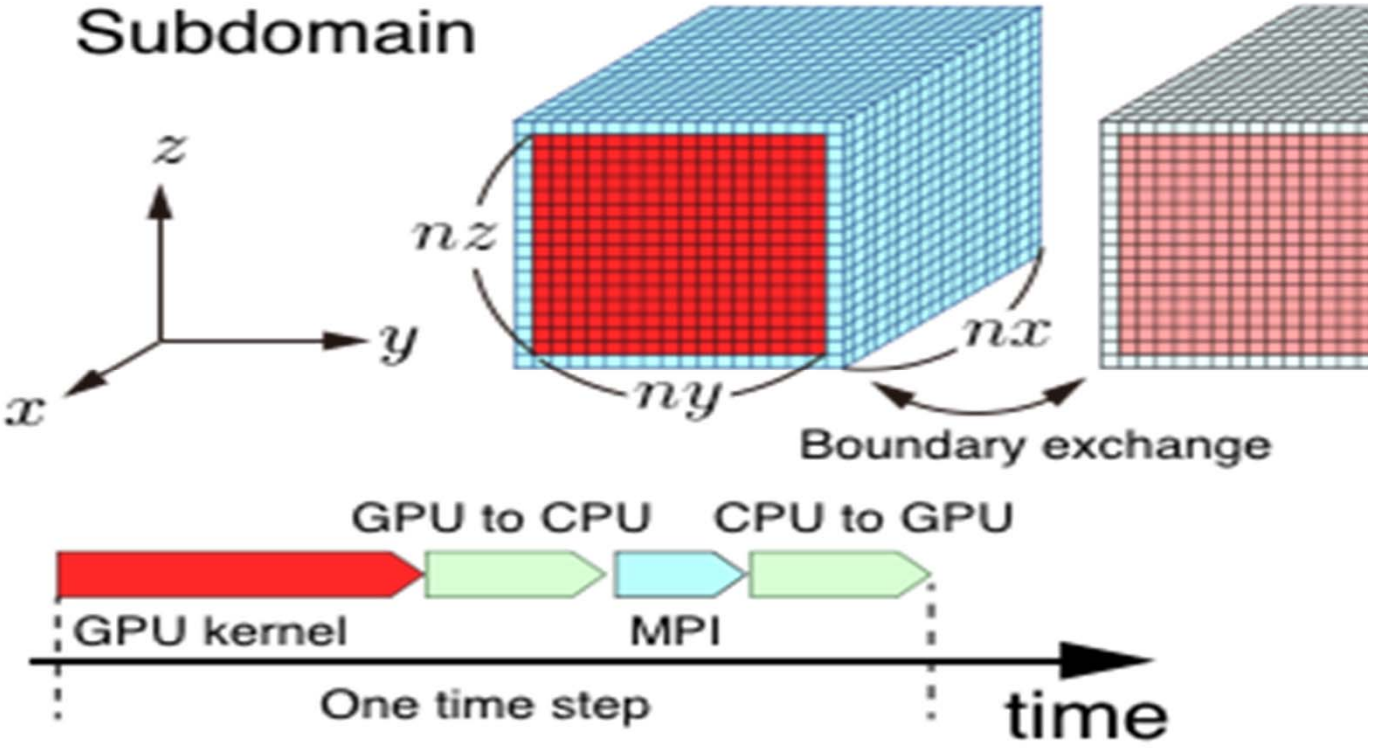


1. **GPU-only Method (without overlapping)**
2. **Hybrid-YZ Method**
3. **Hybrid-Y Method**

Overlapping between Computation and Communication

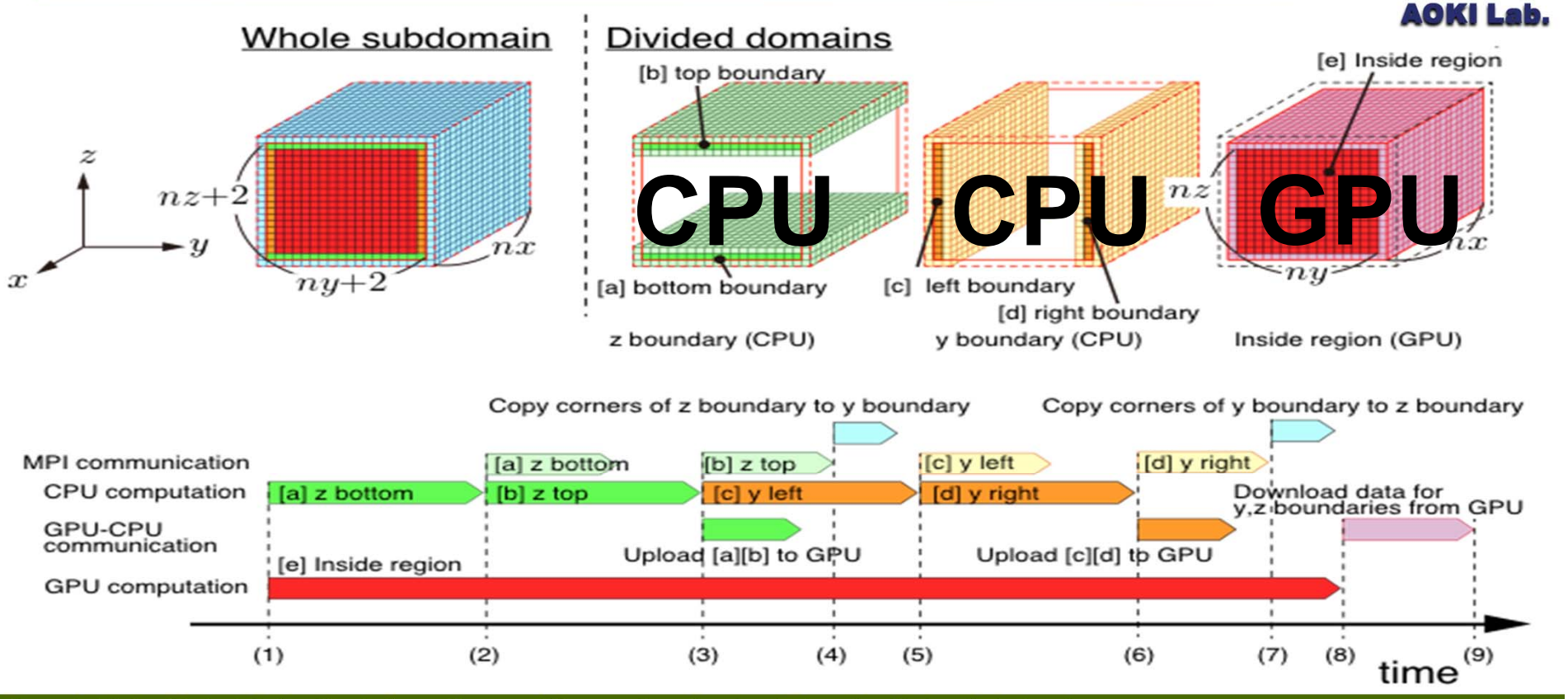


1. GPU-only method

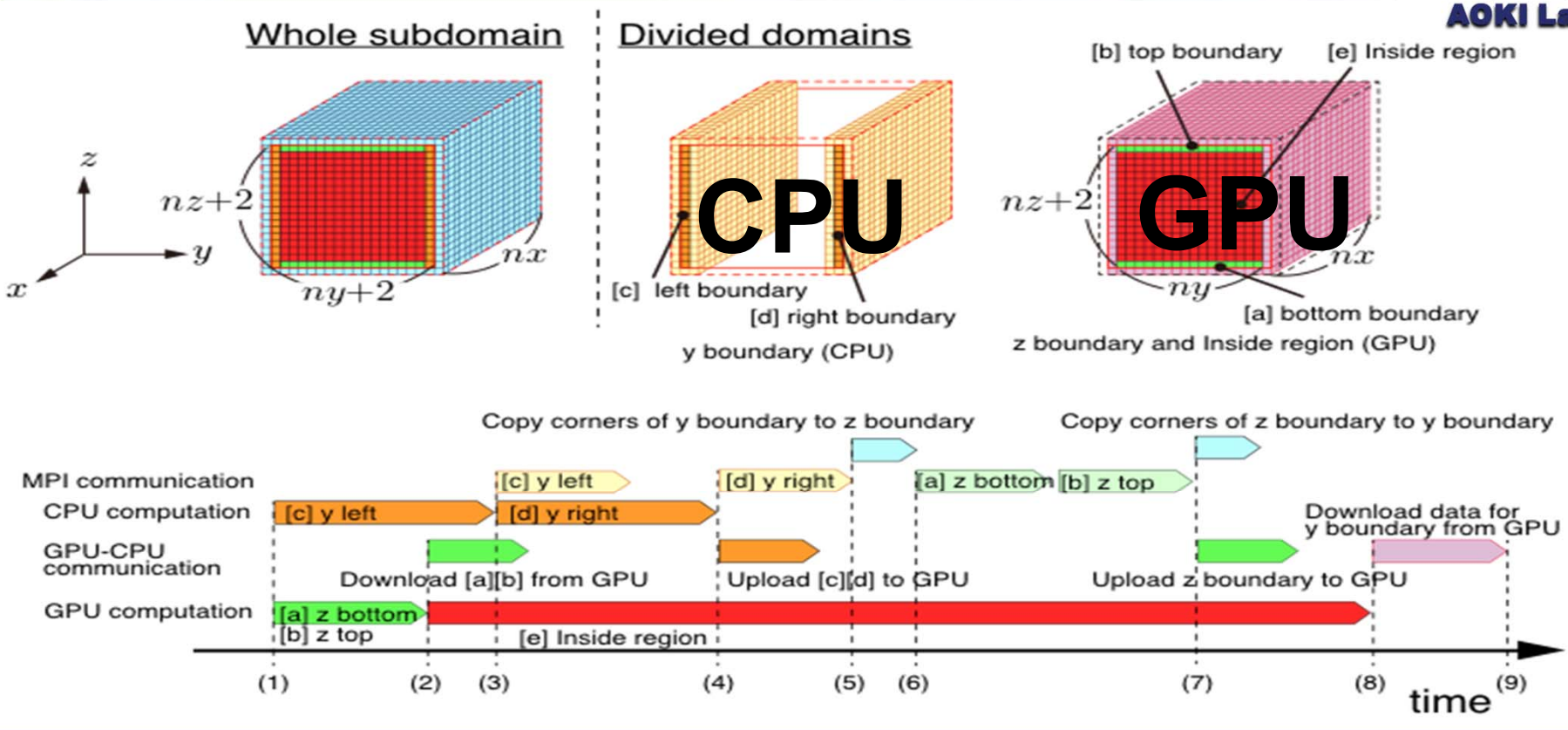


2. Hybrid-YZ method

AOKI Lab.

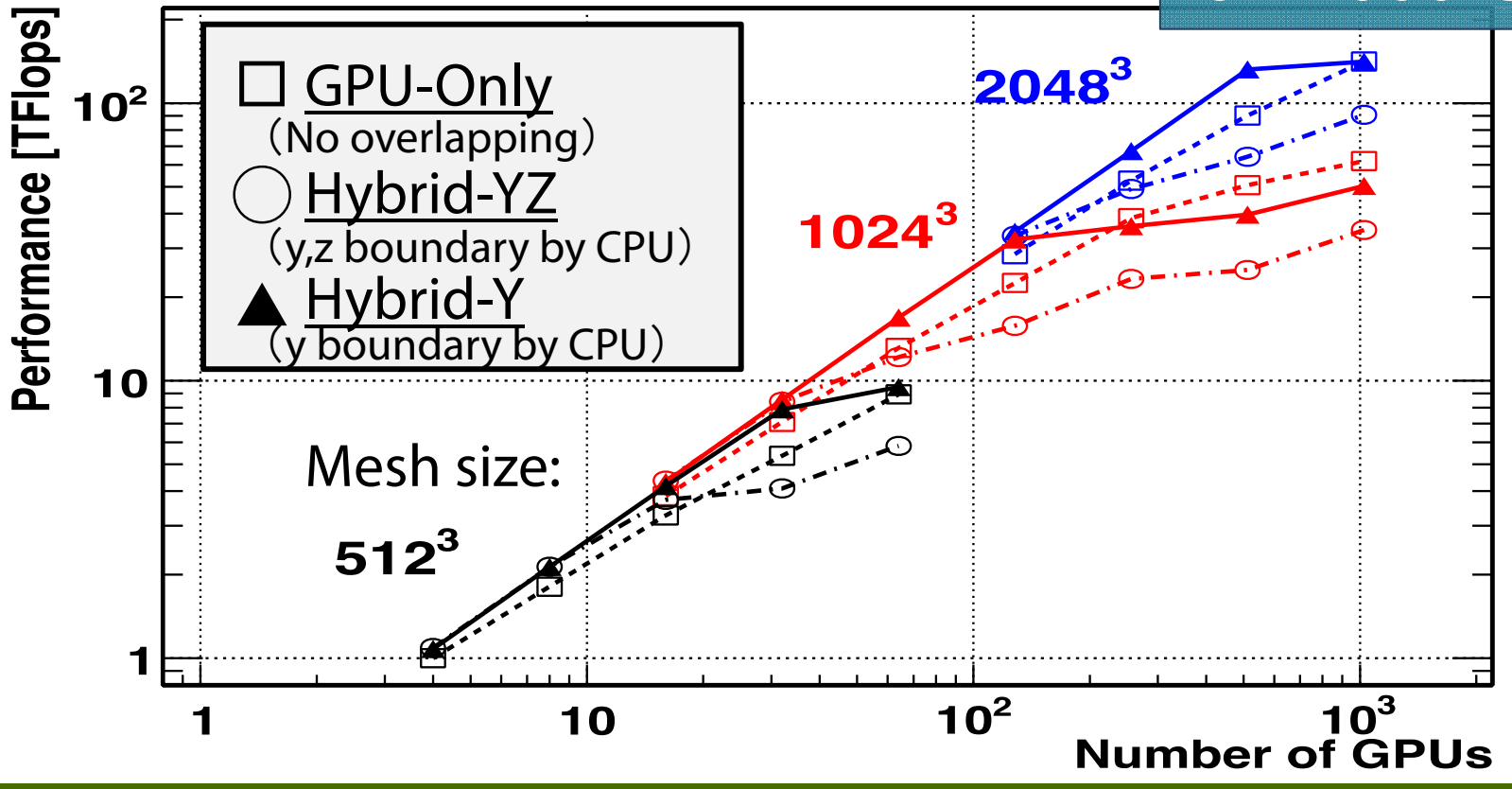


3. Hybrid-Y method



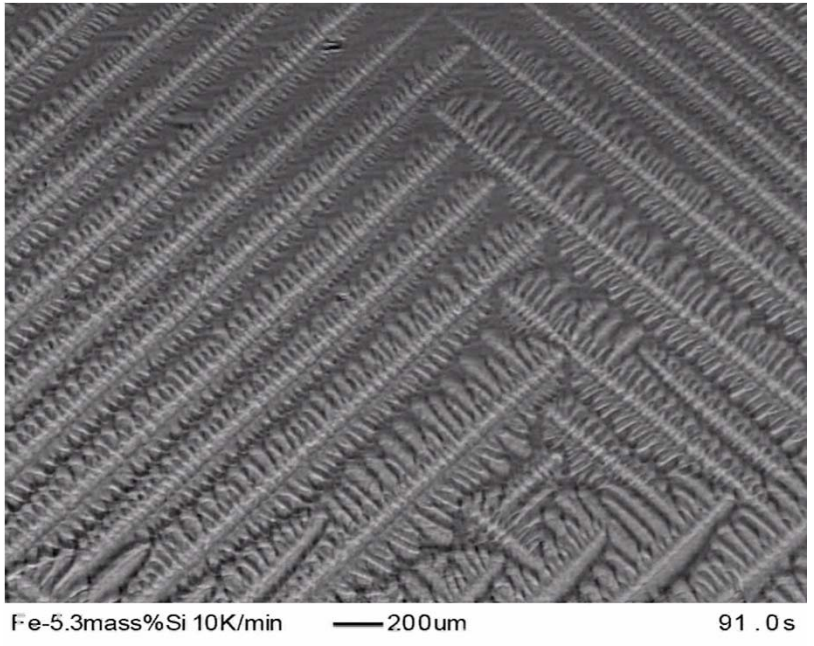
Strong Scalabilit

1.017 Pflops on 4000 GPUs

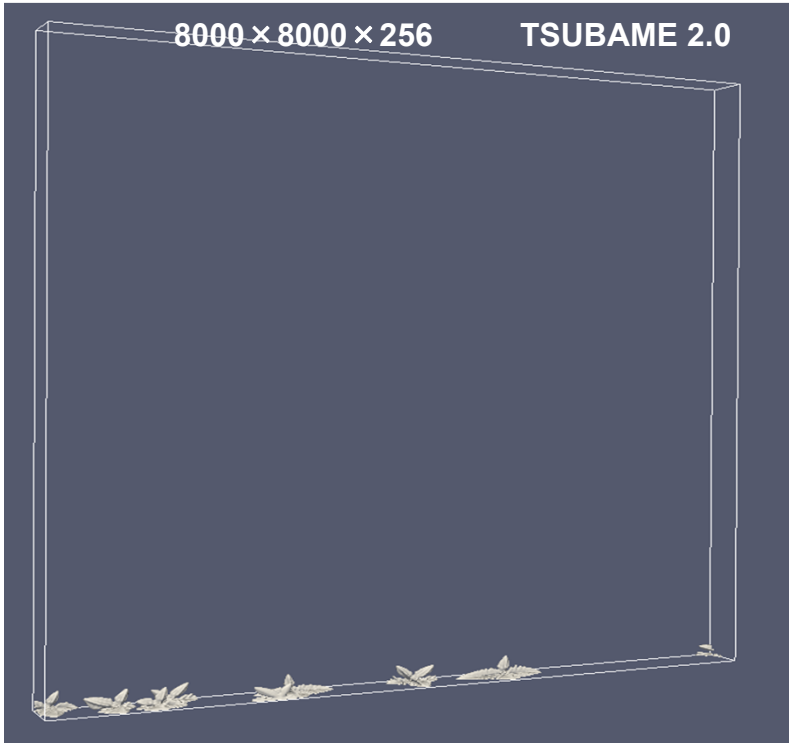


Comparison with Experiment

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



Further Tuning from 1 PFlops to 2PFlops

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.

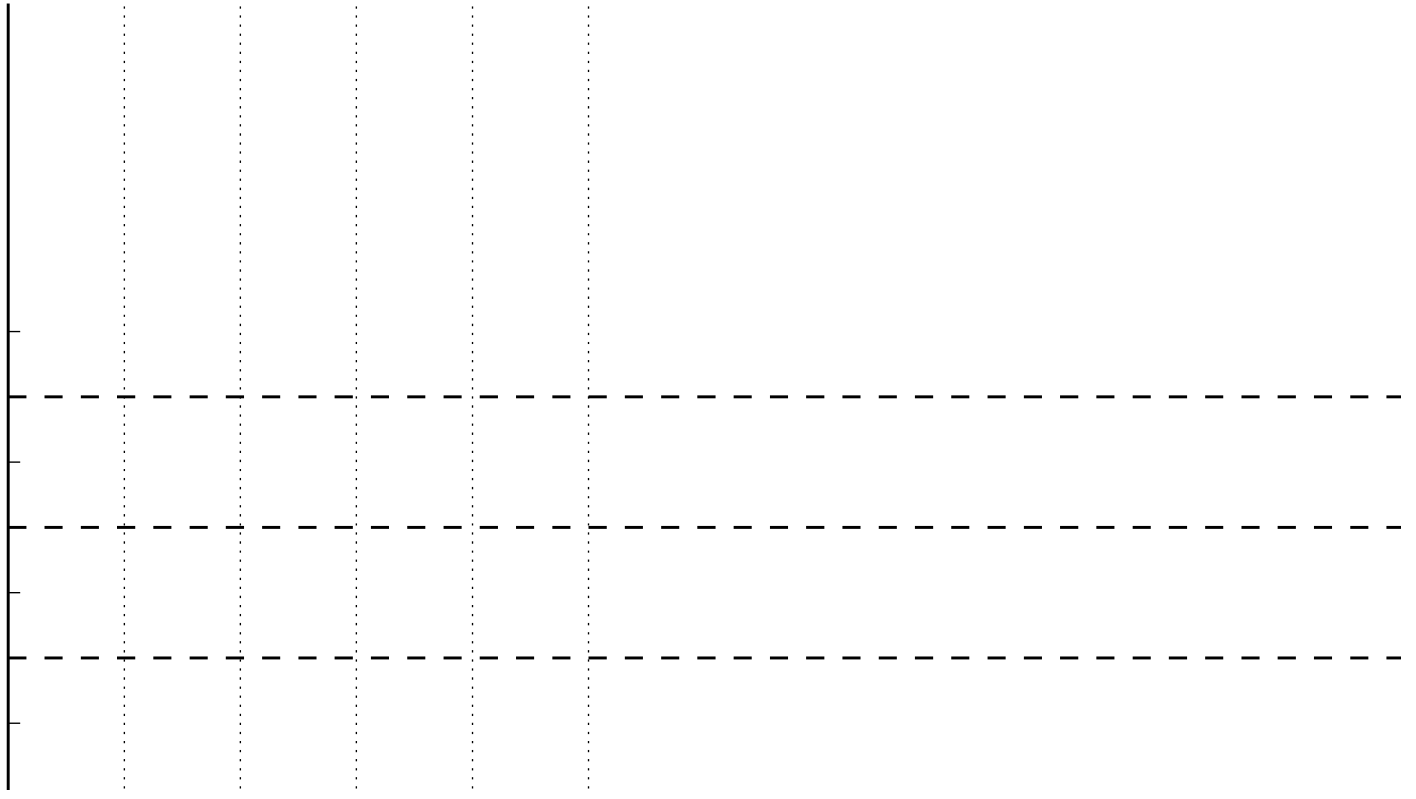
Breakdown of Elapsed Time

Hyb

Orig

(CUI

CUD



Breakdown of Elapsed Time

Hyb

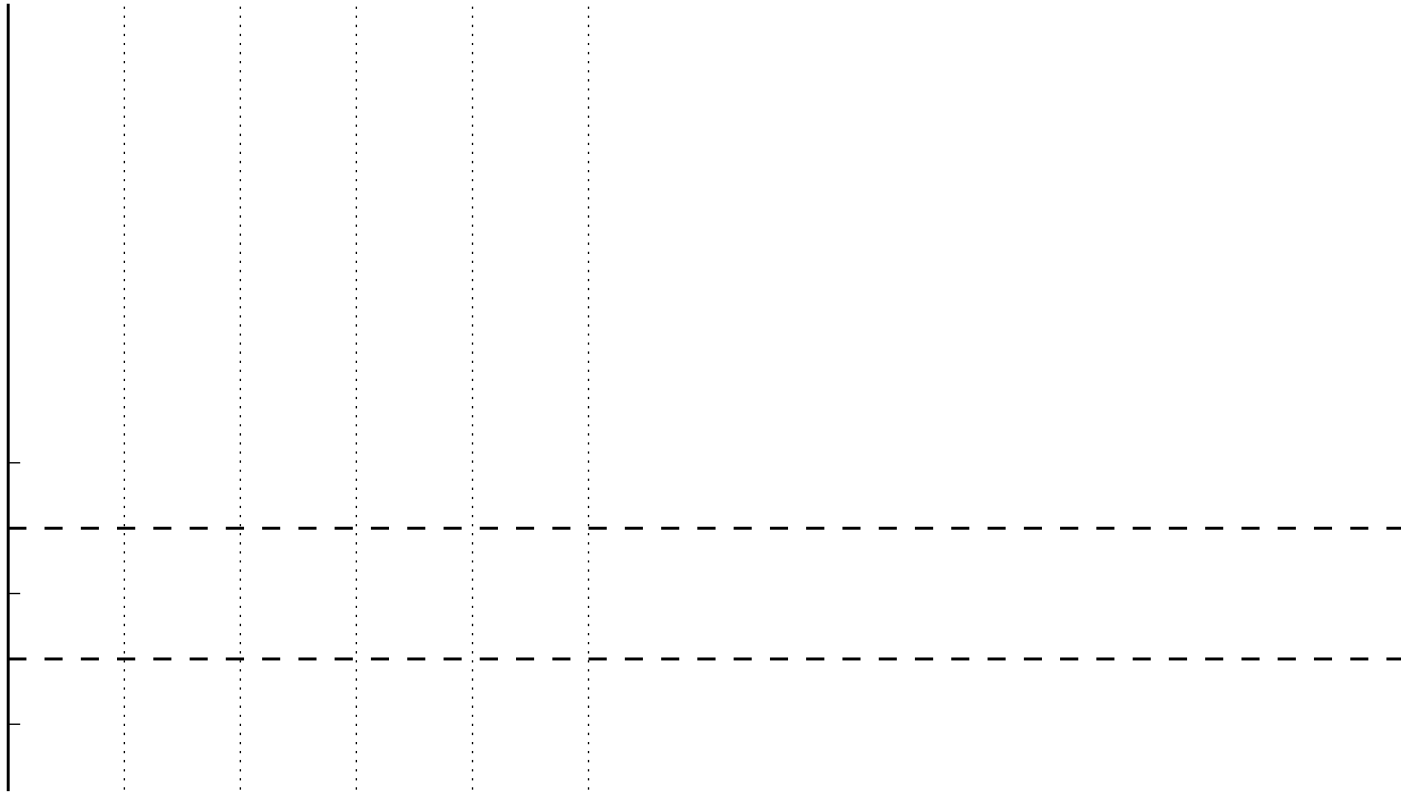
Orig

(CUI

CUD

+= Fas

approx



Breakdown of Elapsed Time

Hyb

Orig

(CUI

CUD

+= Fas

approx

+= \

Breakdown of Elapsed Time

Hyb

Orig

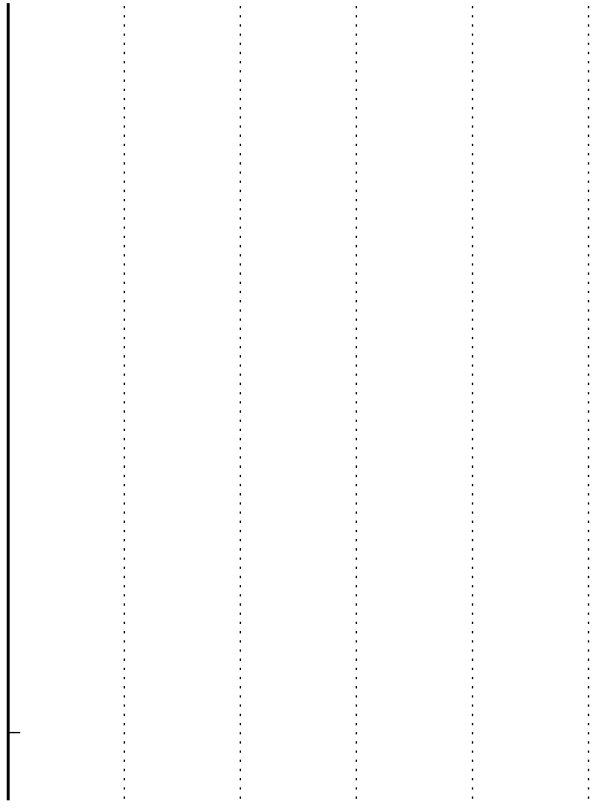
(CUI

CUD

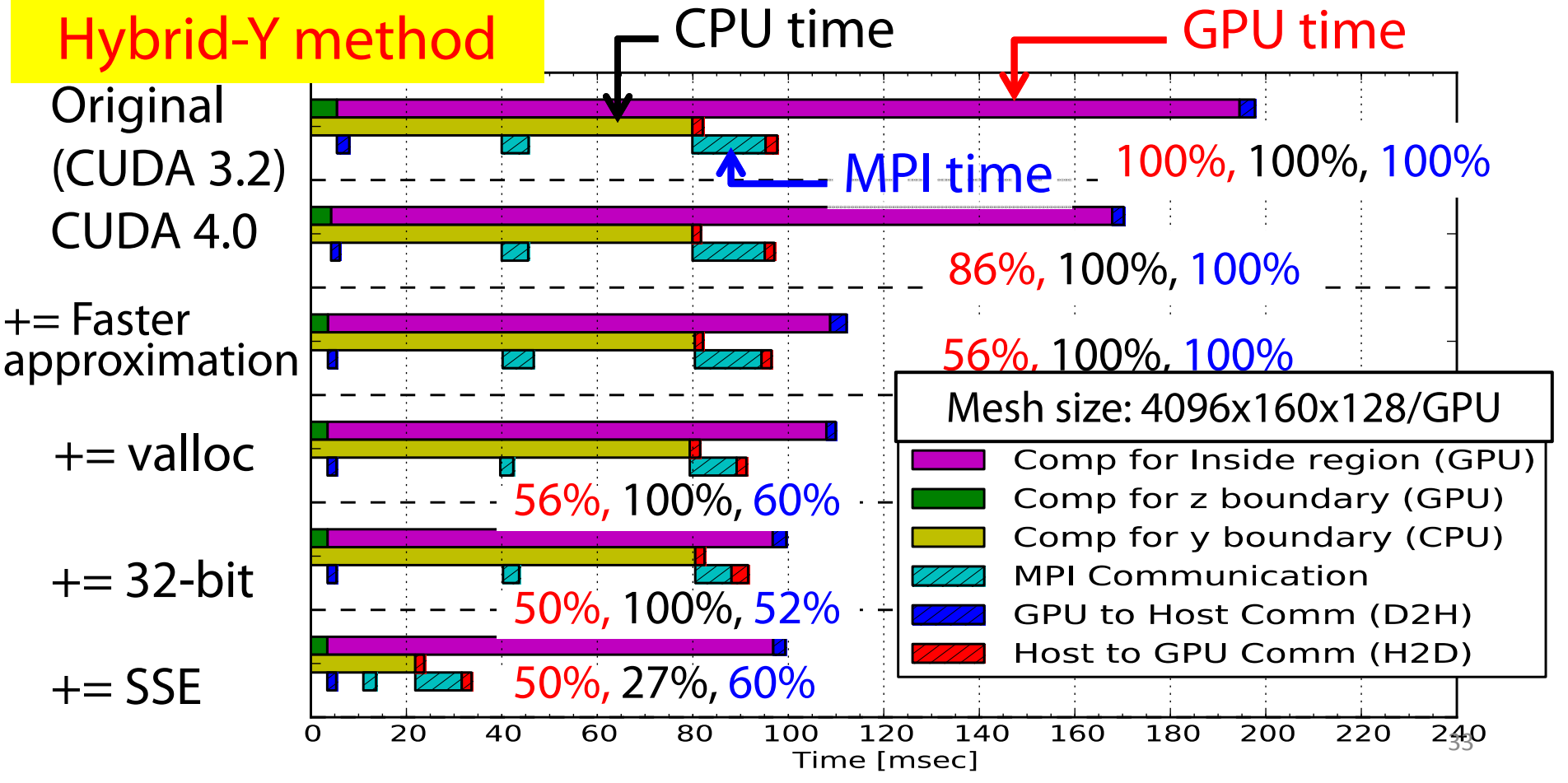
+= Fas
approx

+= \

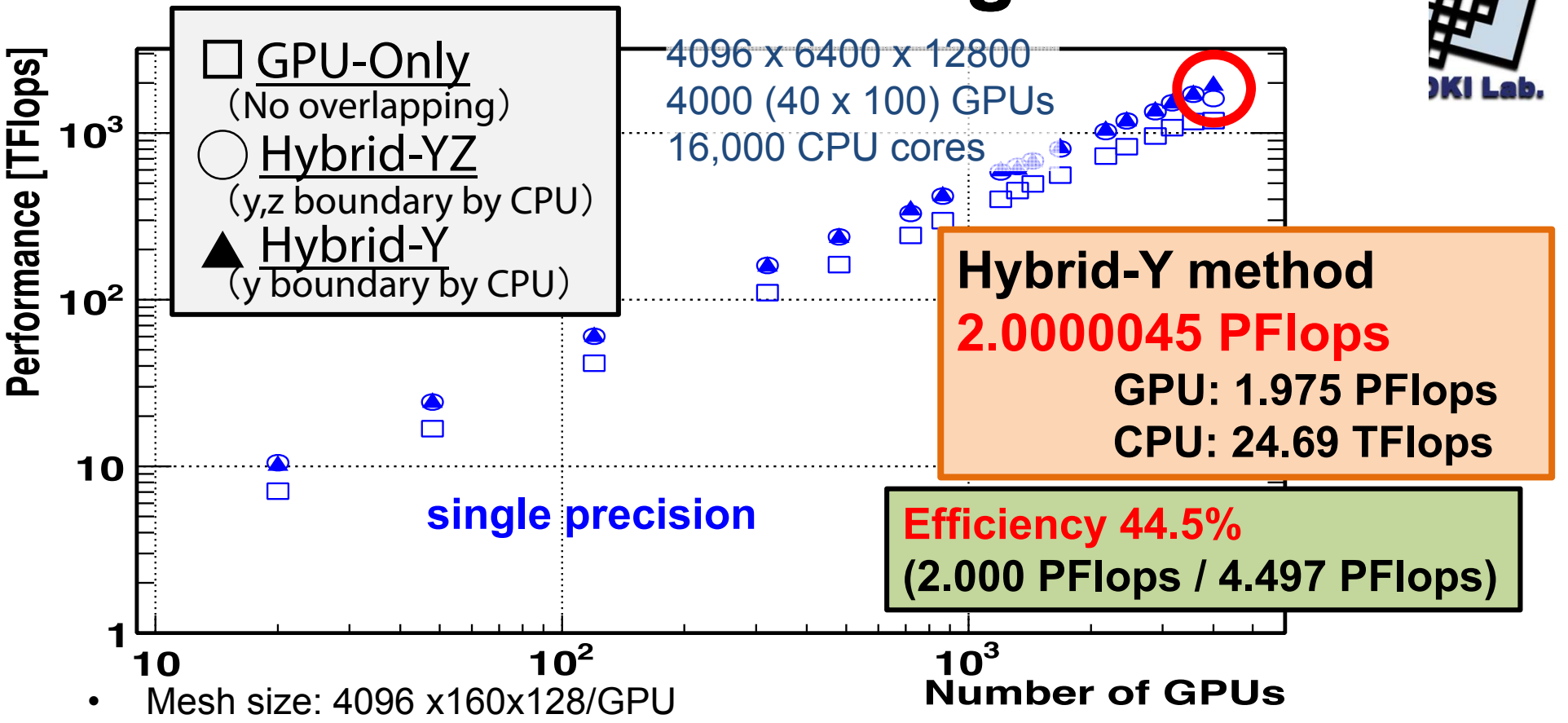
+= 3



Breakdown of Elapsed Time



Weak scaling



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


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
acm
Scott Lathrop
SC11 Conference Chair
Thom H. Dunning, Jr.
Gordon Bell Chair
IEEE COMPUTER SOCIETY

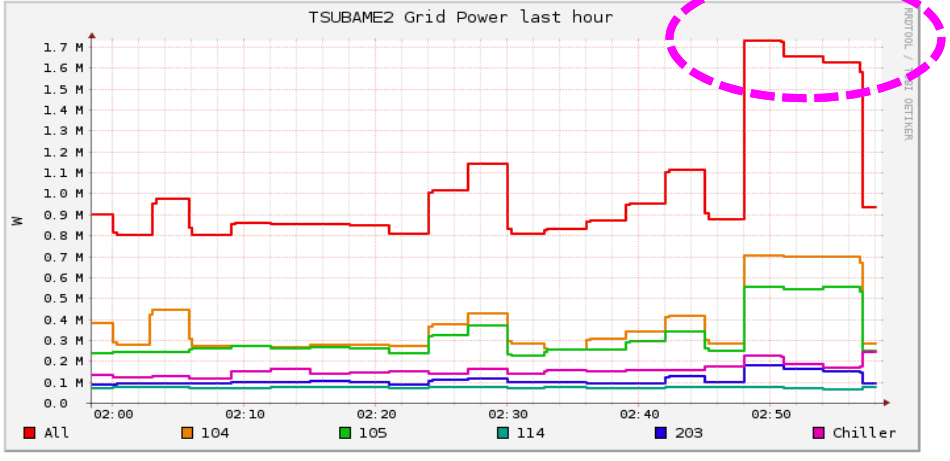
Power Efficiency

AOKI Lab.

- 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

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

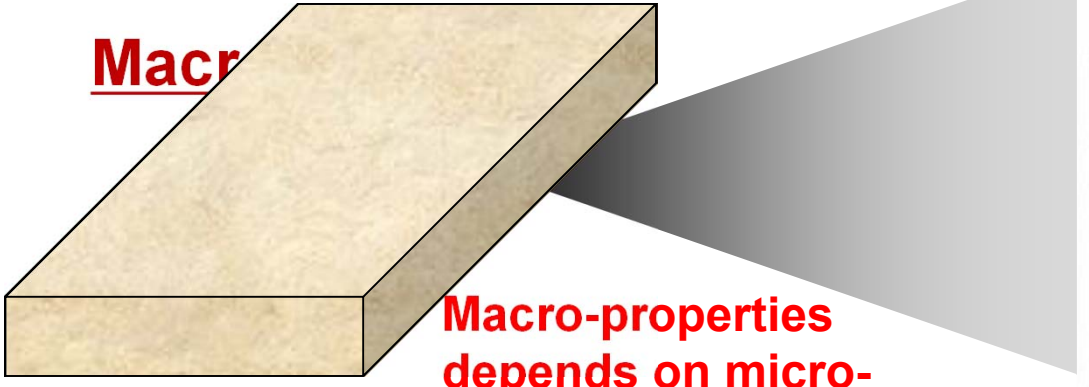
Thank you

for your kind attention

Metal Dendritic Solidification

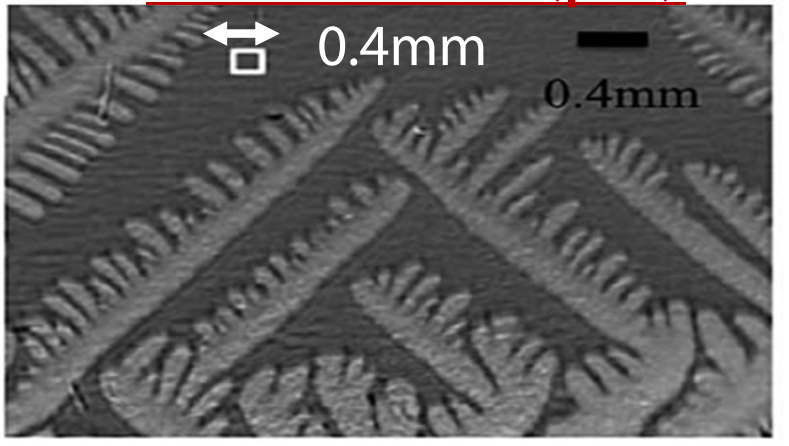
AOKI Lab.

■ 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)