# Peta-scale GPU applications on TSUBAME2.0

Global Scientific Information and Computing Center,
Tokyo Institute of Technology

Takayuki AOKI

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

# 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 |

# 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

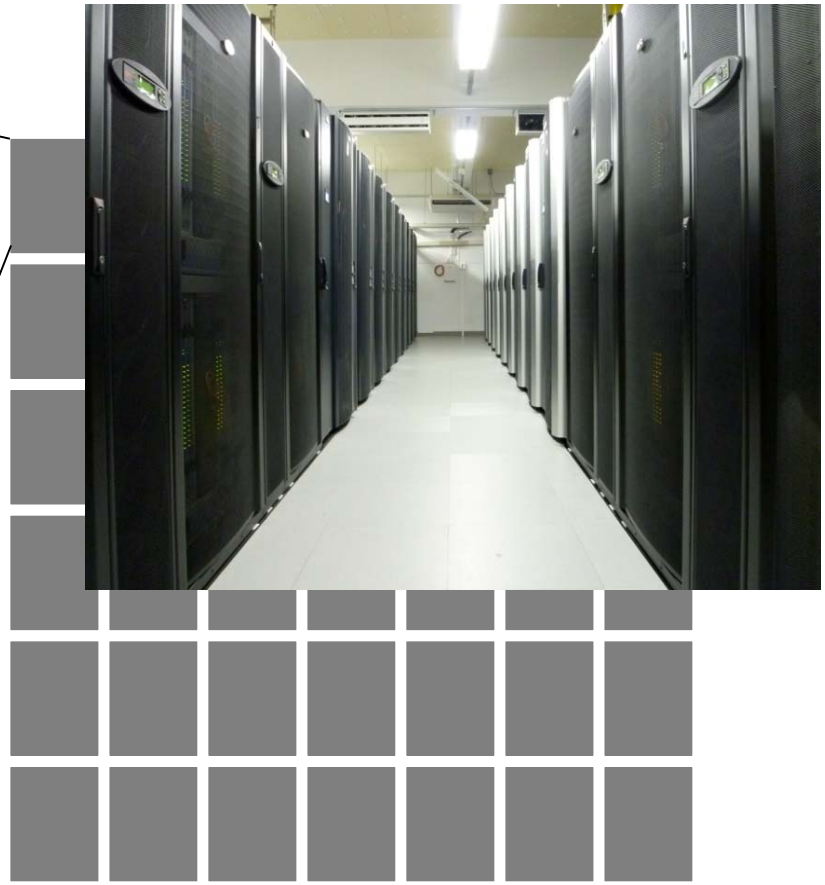
Memory: 103.9 TB

**Rack** (30 nodes)

Performance: 51.0 TFLOPS
Memory: 2.03 TB

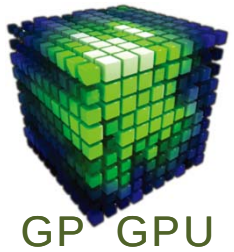
**Compute Node**
(2 CPUs, 3 GPUs)

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



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

# Details of Compute Node



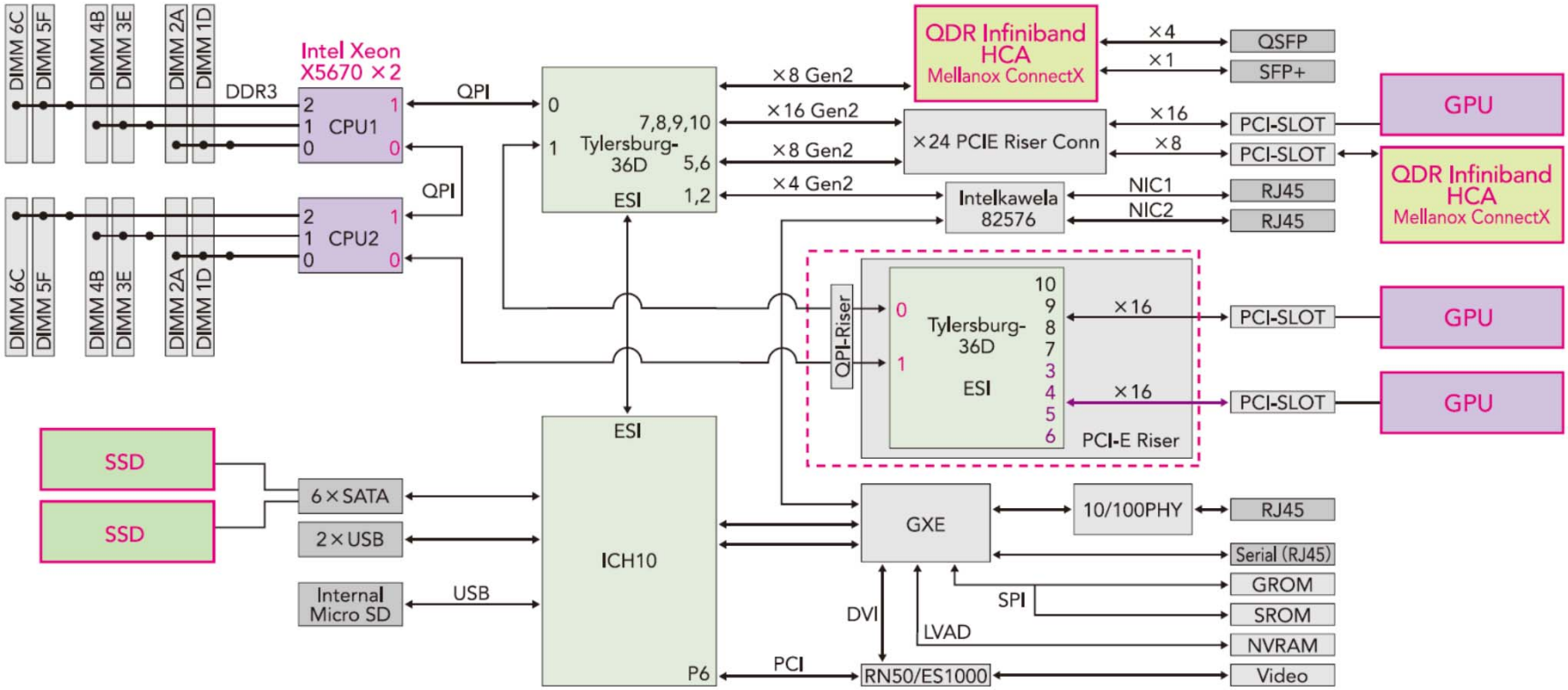

**HP ProLiant SL390s**

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

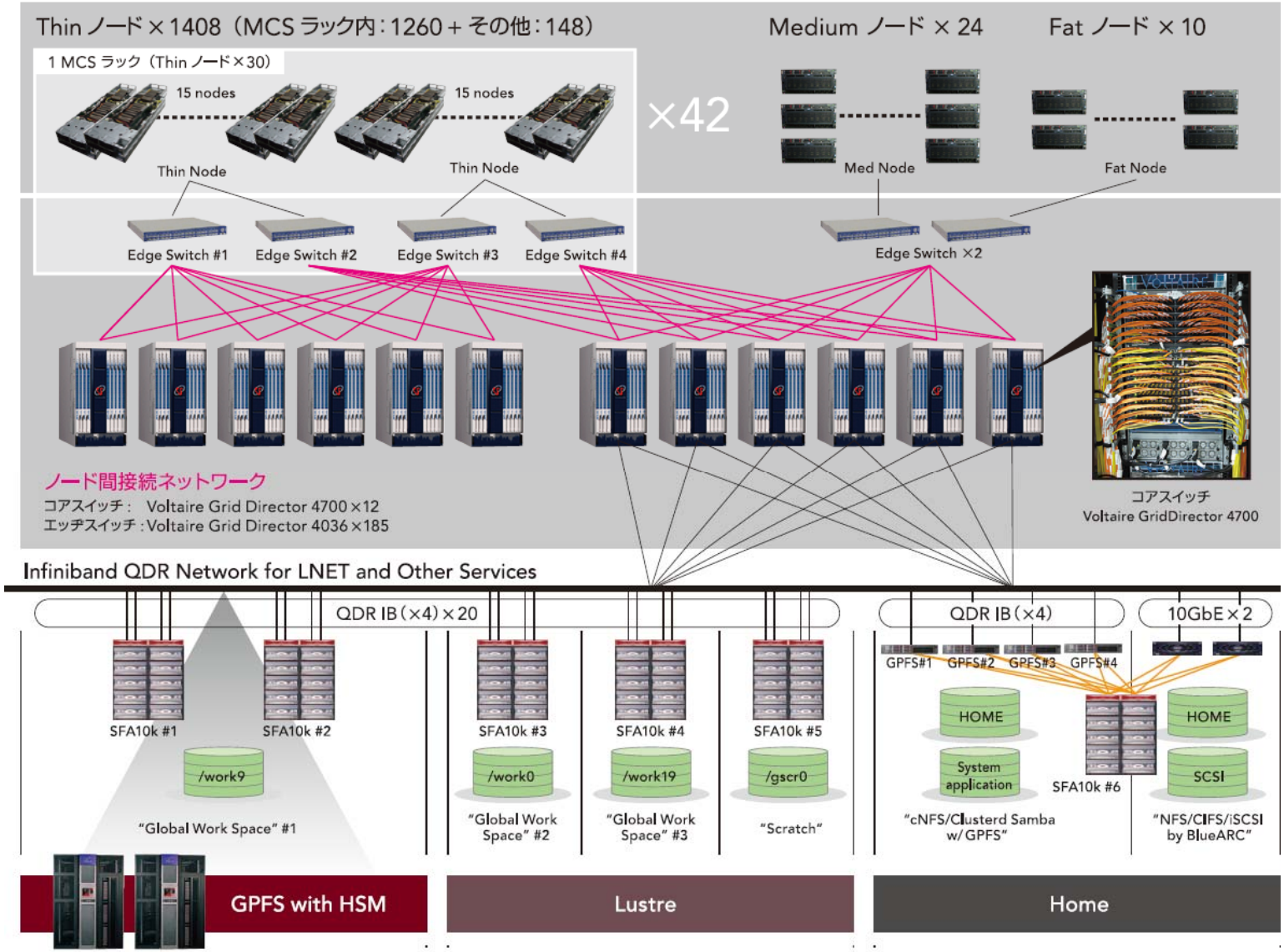
Memory : 58GB DDR3 1333MHz 一部 103GB

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




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

# High-Speed Network and Reliable Storage System



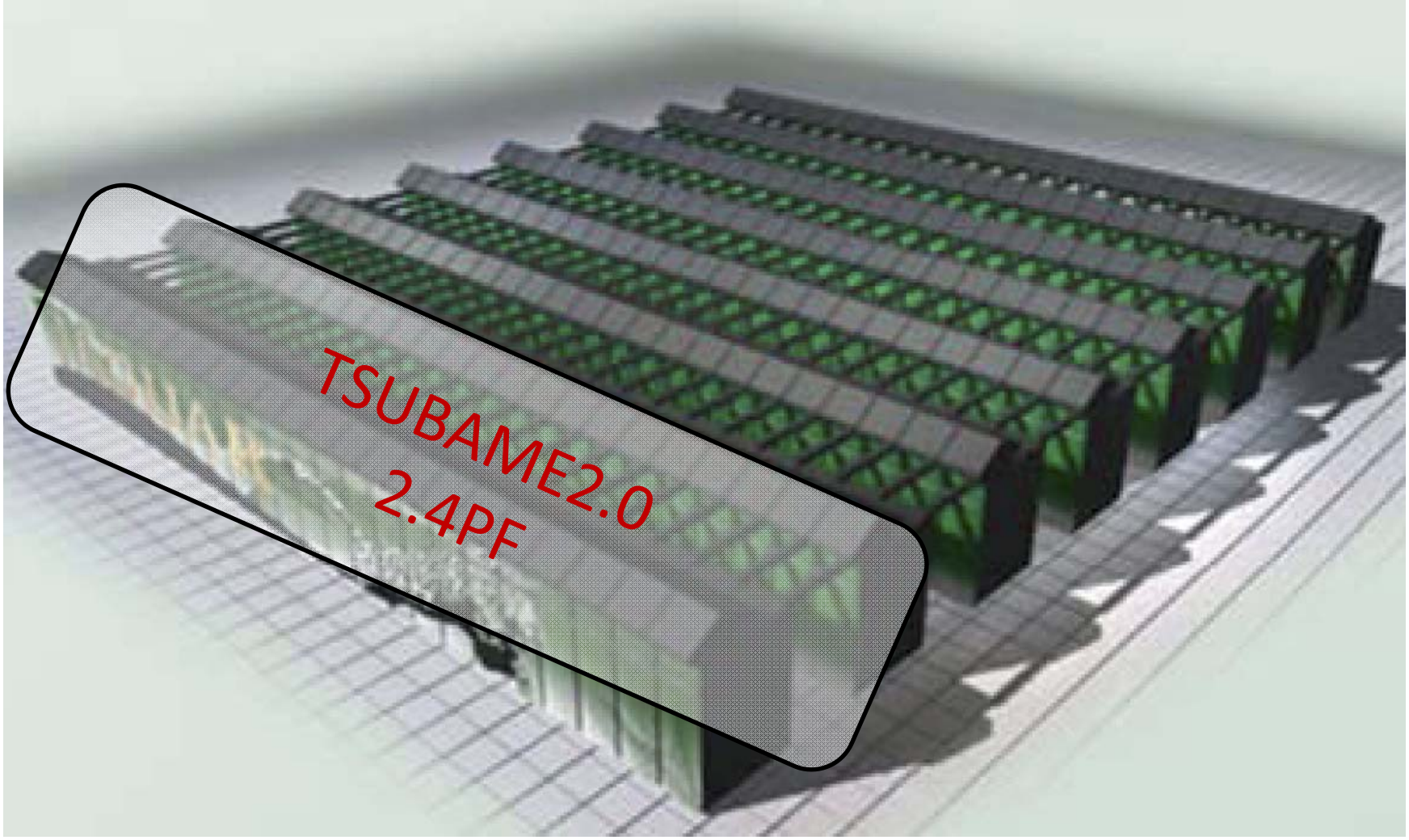

Copyright © Takayuki Aoki / Global Scientific Information and Computing Center, Tokyo Institute of Technology

# GPU M2050

# ORNL Jaguar vs Tsubame 2.0

## Similar Peak Performance, 1/5 the Size and Power

GP GPU




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

# Supercomputer in the world

November 2011

THE GREEN 500™

| Green500 Rank | MFLOPS/W | Site* | Computer* | Total Power (kW) |
|---|---|---|---|---|
| 1 | 2026.48 | IBM - Rochester | BlueGene/Q, Power BQC 16C 1.60 GHz, Custom | 85.12 |
| 2 | 2026.48 | IBM Thomas J. Watson Research Center | BlueGene/Q, Power BQC 16C 1.60 GHz, Custom | 85.12 |
| 3 | 1996.09 | IBM - Rochester | BlueGene/Q, Power BQC 16C 1.60 GHz, Custom | 170.25 |
| 4 | 1988.56 | DOE/NNSA/LLNL | BlueGene/Q, Power BQC 16C 1.60 GHz, Custom | 340.50 |
| 5 | 1689.86 | IBM Thomas J. Watson Research Center | NNSA/SC Blue Gene/Q Prototype 1 | 38.67 |
| 6 | 1378.32 | Nagasaki University | DEGIMA Cluster, Intel i5, ATI Radeon GPU, Infiniband QDR | 47.05 |
| 7 | 1266.26 | Barcelona Supercomputing Center | Bullx B505, Xeon E5649 6C 2.53GHz, Infiniband QDR, NVIDIA 2090 | 81.50 |
| 8 | 1010.11 | TGCC / GENCI | Curie Hybrid Nodes - Bullx B505, Xeon E5640 2.67 GHz, Infiniband QDR | 108.80 |
| 9 | 963.70 | Institute of Process Engineering, Chinese Academy of Sciences | Mole-8.5 Cluster, Xeon X5520 4C 2.27 GHz, Infiniband QDR, NVIDIA 2050 | 515.20 |
| 10 | 958.35 | GSIC Center, Tokyo Institute of Technology | HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows | 1243.80 |

* Performance data obtained from publicly available sources including TOP500

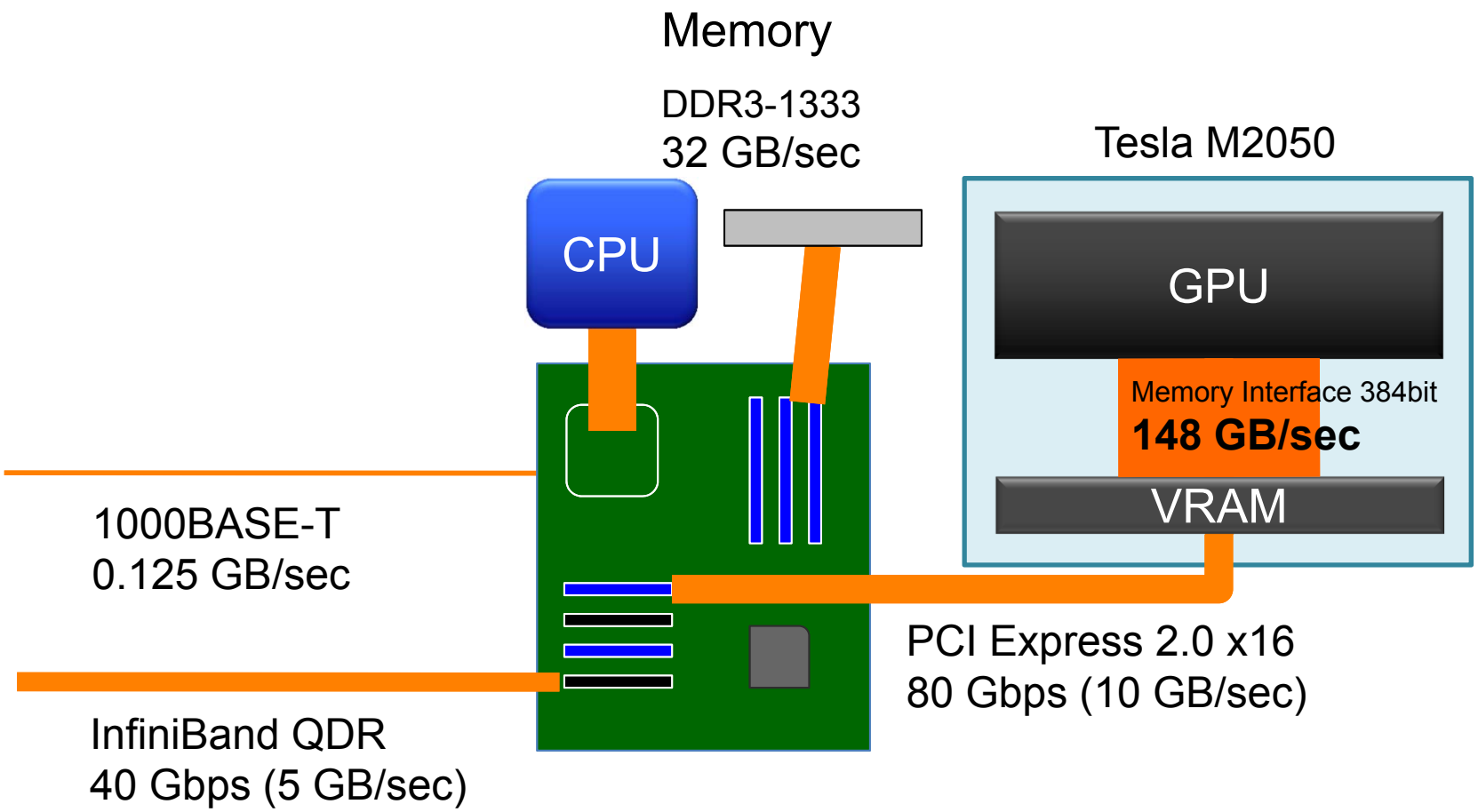
**TSUBAME2.0 PUE = 1.2** (Power Usage Effectiveness)

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

# Memory Hierarchy

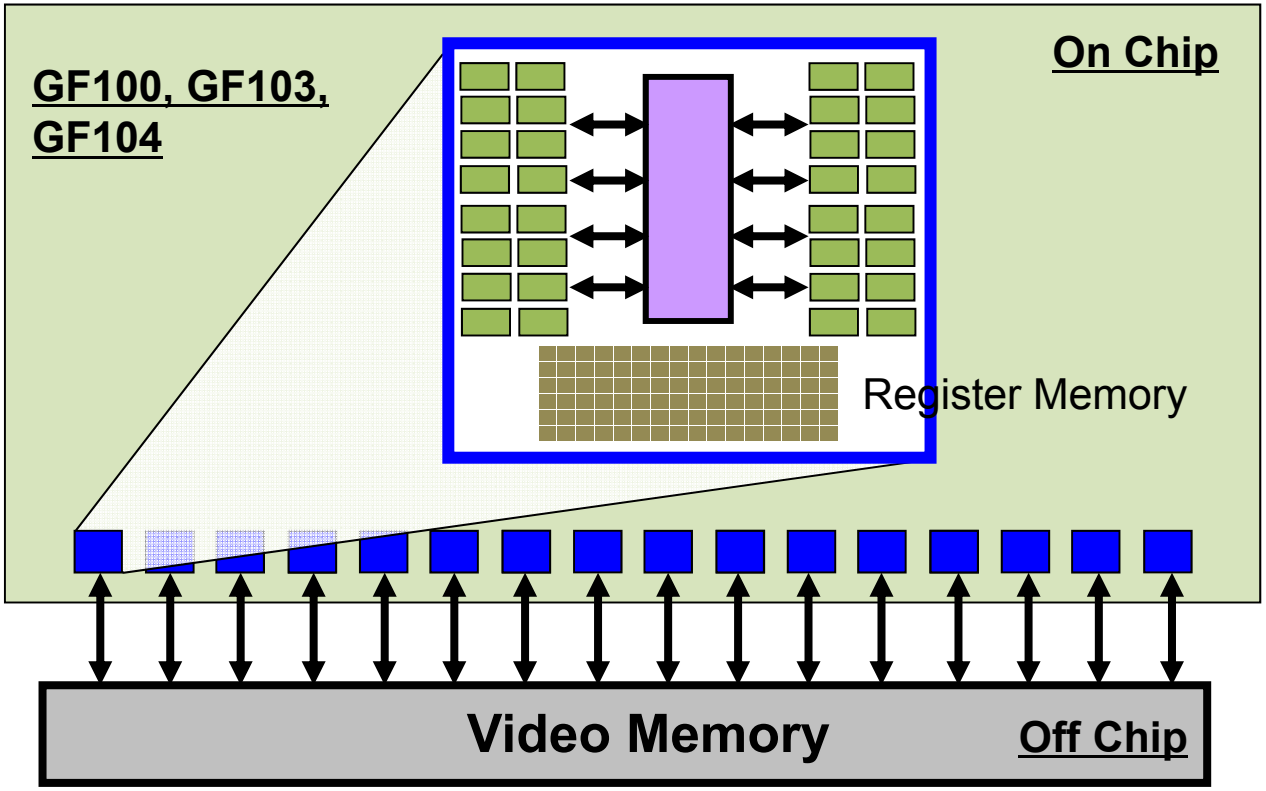
GP GPU

## ■ Several Bandwidth Bottle Necks




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

# GPU Architecture



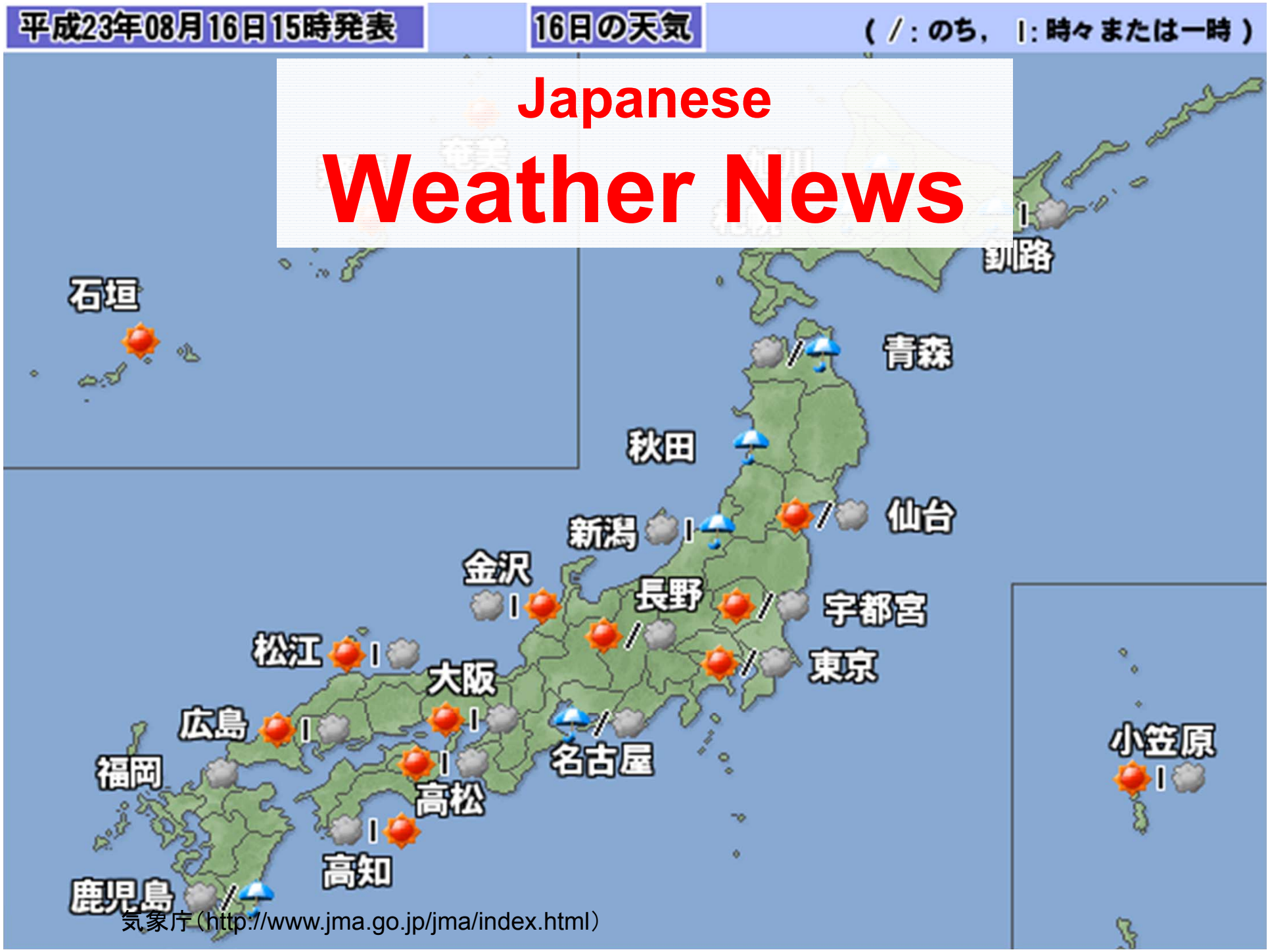

| | |
|---|---|
| Global memory | ~6GB (VRAM |
| Streaming Multiprocessor | ~16 (C2050 (GF100): 14) |
| Shared memory + L1 Cache | 64 Kbyte |
| Streaming Processor (CUDA core) | 8～48 per SM, total 512 |

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

# Large-scale memory-bound and Stencil Applications

# on TSUBAME2.0

# Japanese
# Weather News




気象庁（http://www.jma.go.jp/jma/index.html）

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** ⟷ **a few km**

**Typhoon**

**Tornado, Down burst**
**Heavy Rain**

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

# WRF GPU Computing

GP GPU

## ■ WRF (Weather Research and Forecast)

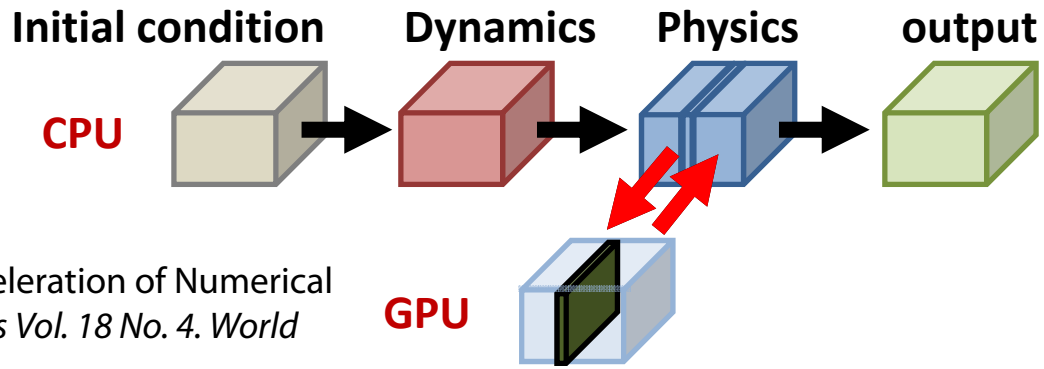
**WSM5** (WRF Single Moment 5-tracer) Microphysics*

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)

**WRF-Chem**** provides the capability to simulate chemistry and aerosols from cloud scales to regional ⇒ x 8.5 increase

**Accelerator Approach**




*J. Michalakes, and M. Vachharajani: GPU Acceleration of Numerical Weather Prediction. *Parallel Processing Letters Vol. 18 No. 4. World Scientific. Dec. 2008. pp. 531—548*

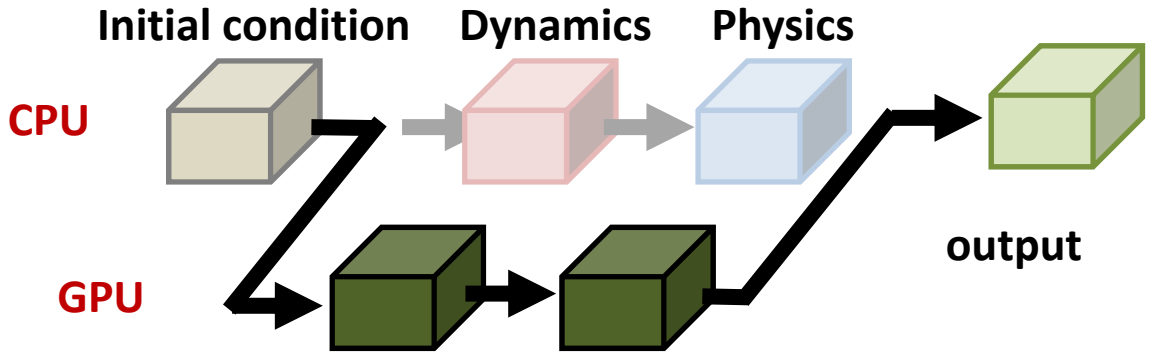
**John C. Linford, John Michalakes, Manish Vachharijani, and Adrian Sandu. Multi-core acceleration of chemical kinetics for simulation and prediction, *proceedings of the 2009 ACM/IEEE conference on supercomputing (SC'09), ACM, 2009.*

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

# Full-GPU Implementation: ASUCA

GP GPU

**Full GPU Approach**



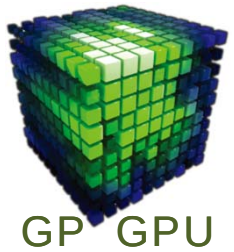

## ■*ASUCA Production Code*

- ✓ ***A next-generation high resolution weather simulation code that is being developed by Japan Meteorological Agency (JMA)***
- ✓ ***ASUCA succeeds the JMA-NHM as an operational non-hydrostatic regional model at JMA***

J. Ishida, C. Muroi, K. Kawano, Y. Kitamura, Development of a new nonhydrostatic model “ASUCA” at JMA, CAS/JSC WGNE Reserch Activities in Atomospheric and Oceanic Modelling.

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

# Entire Porting Fortran to CUDA




- **Rewrite from Scratch**

**Fortran**

```
Program init
implicit none

integer i
integer a(10)

do i = 1, 10
   a(i) = i
end do

end program init
```

✓ *Original code at JMA*

**C/C++**

```
#include <iostream>

int main()
{
  int i;
  int a[10];
  for(i=0;i<10;i++){
    a[i] = i + 1;
  }
}
```

✓ *Changing array order*

**CUDA**

```
#include <cuda.h>

__global__ void init(int
*a){
  a[threadIdx.x] =
threadIdx.x+1;
}
int main()
{
  int i;
  int *a;

  cudaMalloc(&a,sizeof(in
t)*10);
  init<<<1,10>>>(a);
  cudaFree(a);
}
```

✓ *GPU code*

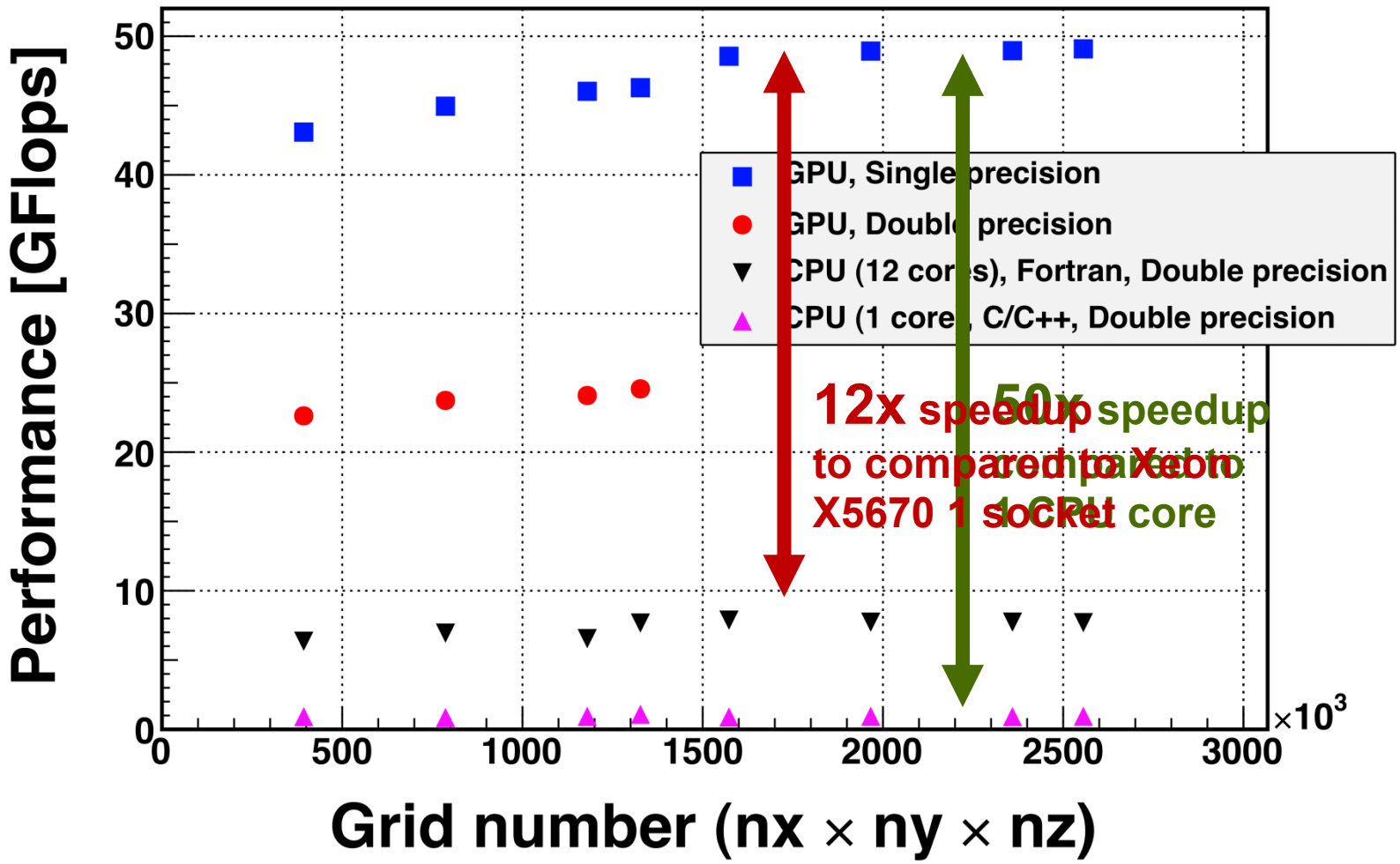
*z,x,y (k,i,j)-ordering* | *x,z,y (i,k,j)-ordering* | *x,z,y (i,k,j)-ordering*

- **1 Year by one Ph.D student**

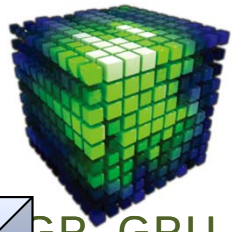
  **Introducing many optimizations, overlapping the computation with the communication, kernel fuse, re-ordering kernel, . . .**

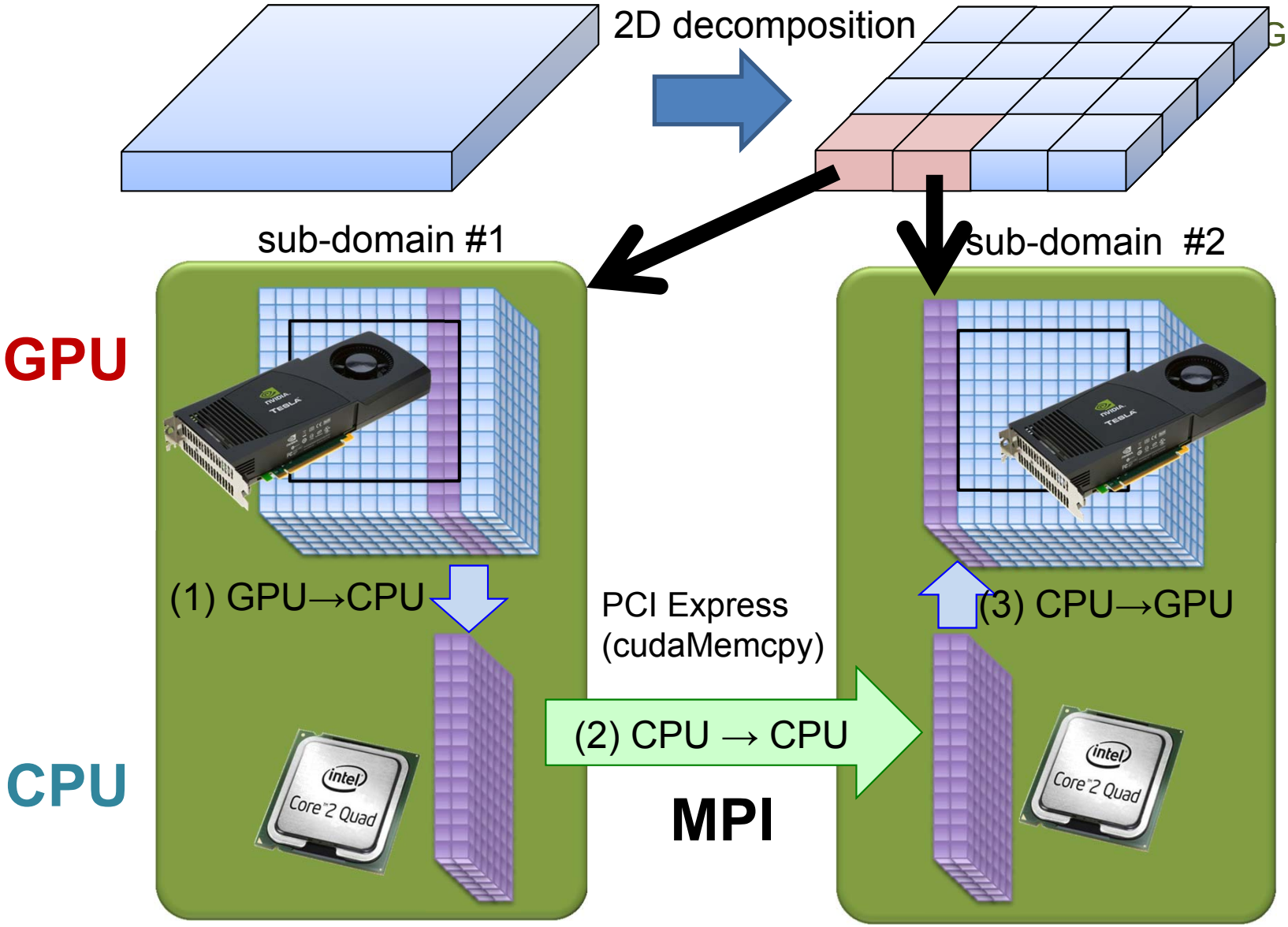
Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

# TSUBAME 2.0 (1 GPU)

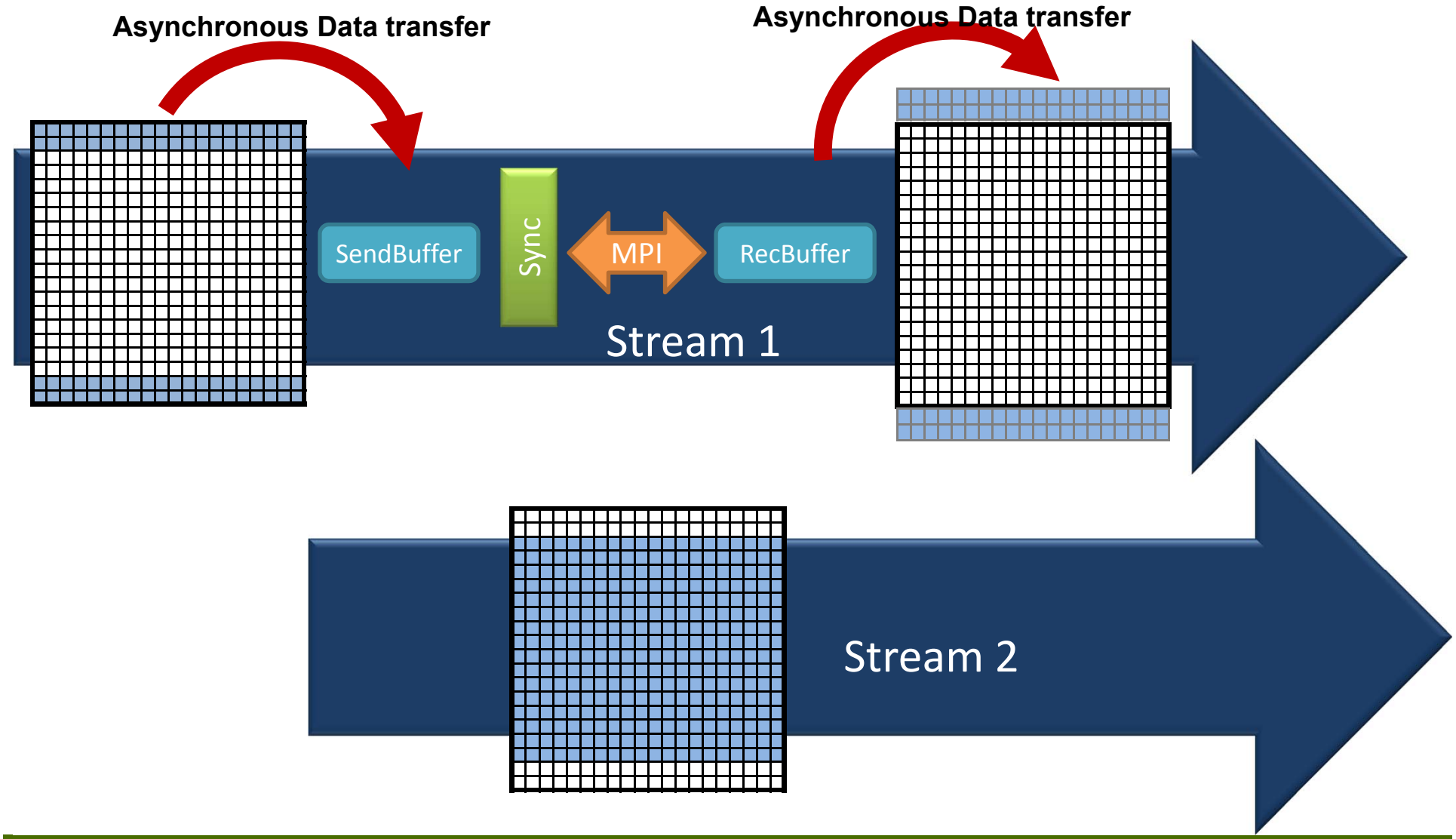


# Multi-GPU : Domain decomposition

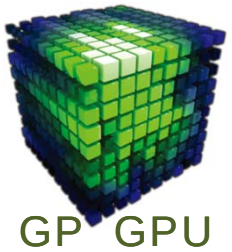


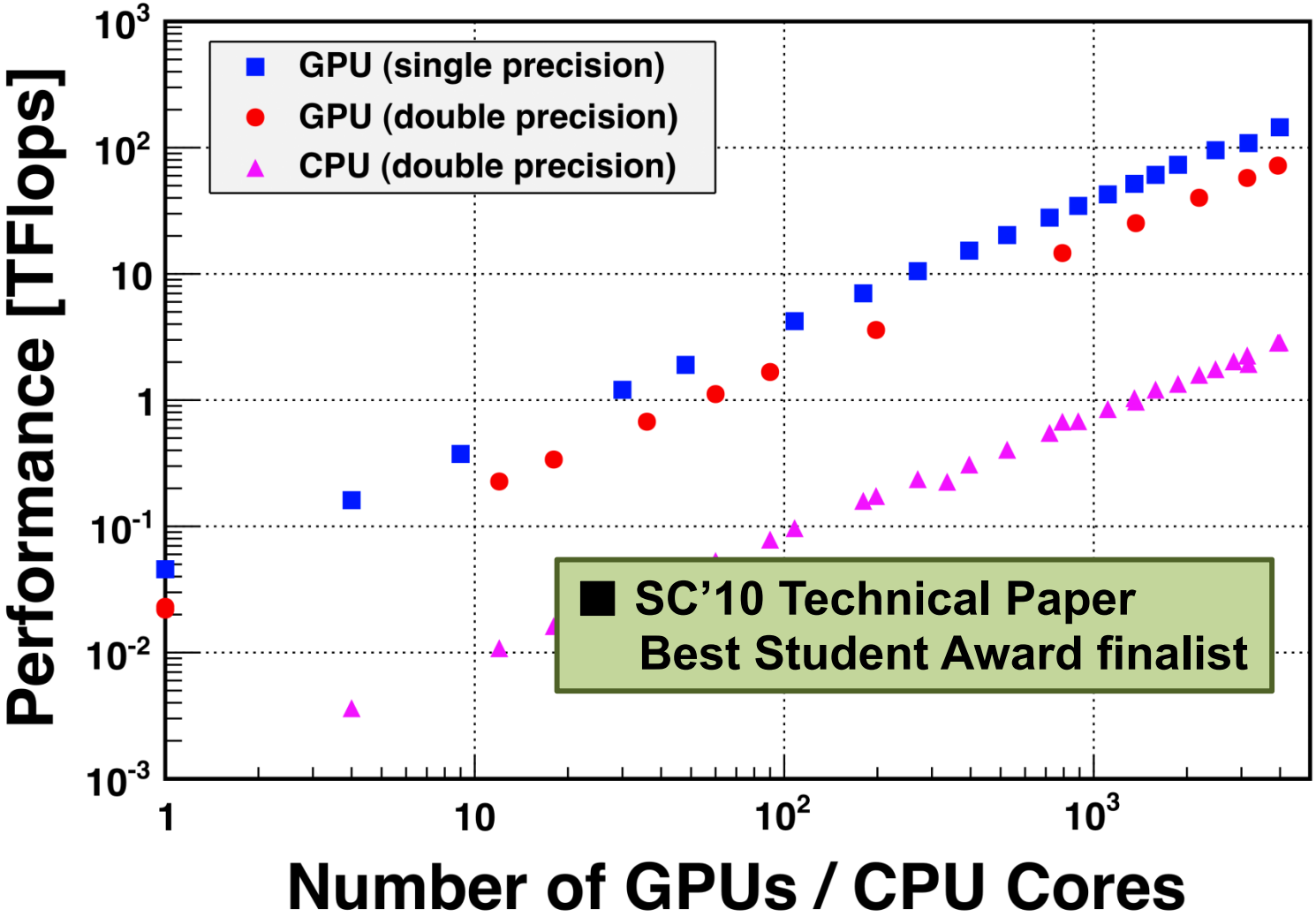


# Overlapping between Computation and Communication



# TSUBAME 2.0 Weak Scaling




**145.0 Tflops**
Single precision

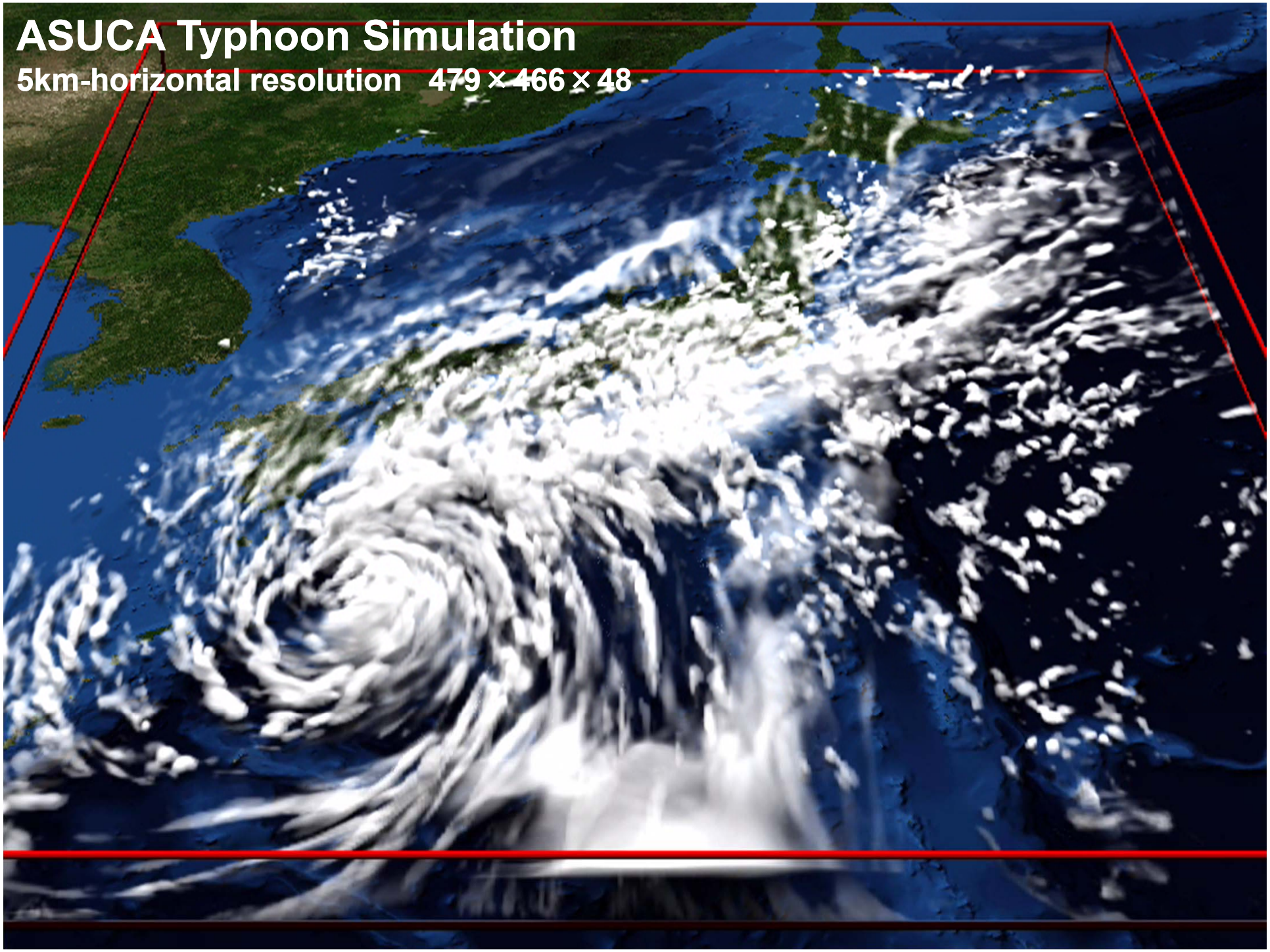
**76.1 Tflops**
Doublele precision

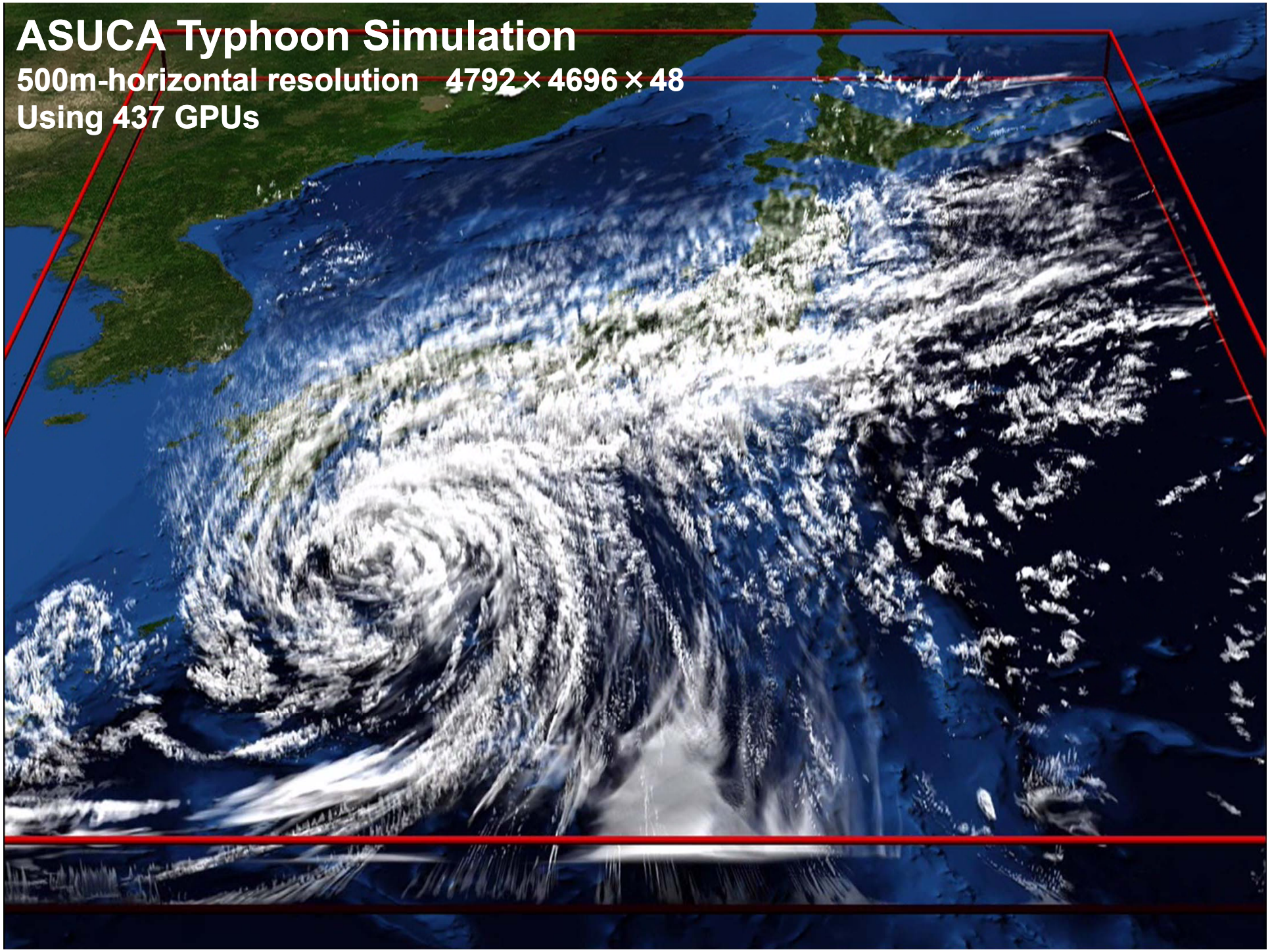
**Fermi core Tesla M2050**

**3990 GPU**

■ **SC'10 Technical Paper**
**Best Student Award finalist**

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

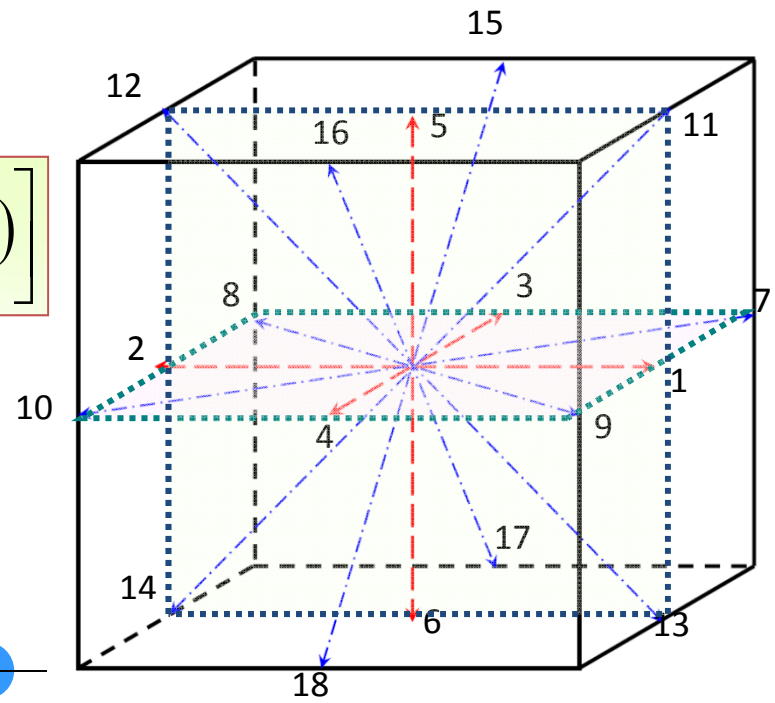

ASUCA Typhoon Simulation
5km-horizontal resolution 479 × 466 × 48


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

# Lattice Boltzmann Method

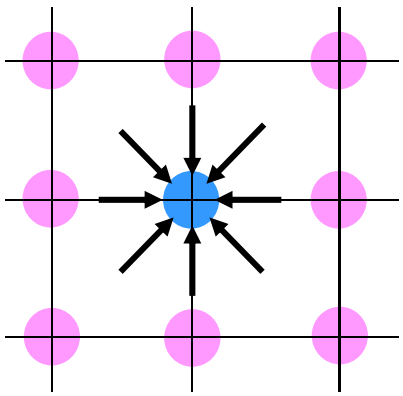
$$\frac{\partial f_i}{\partial t} + \mathbf{e}_i \cdot \nabla f_i = -\frac{1}{\lambda}\left(f_i - f_i^{eq}\right)$$

$$f_i^{eq} = \rho w_i \left[1 + \frac{3}{c^2}(\mathbf{e}_i \cdot \mathbf{u}) + \frac{9}{2c^4}(\mathbf{e}_i \cdot \mathbf{u})^2 - \frac{3}{2c^2}(\mathbf{u} \cdot \mathbf{u})\right]$$

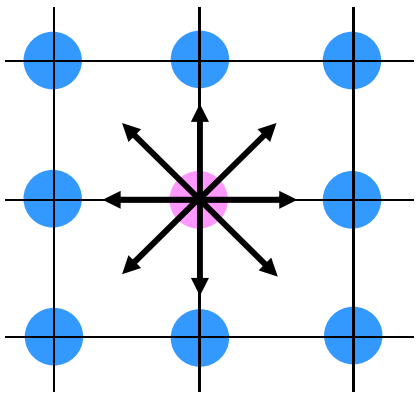



**Strongly Memory Bound Problem:**

**Collision step:**



**Streaming step:**



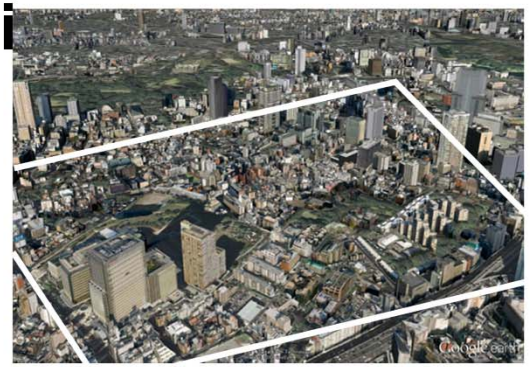
***i*** is the value in the direction of *ith* discrete velocity
$e_i$ is the discrete velocity set;
$w_i$ is the weighting factor
$c$ is the particle velocity
**u** is the macroscopic velocity

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

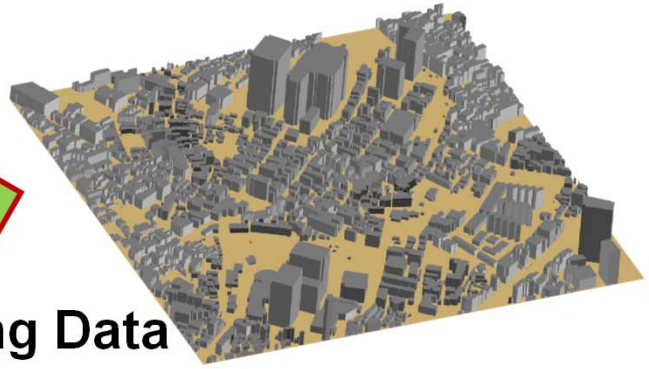
# Real City Atmosphere

**Tokyo Roppongi Area**

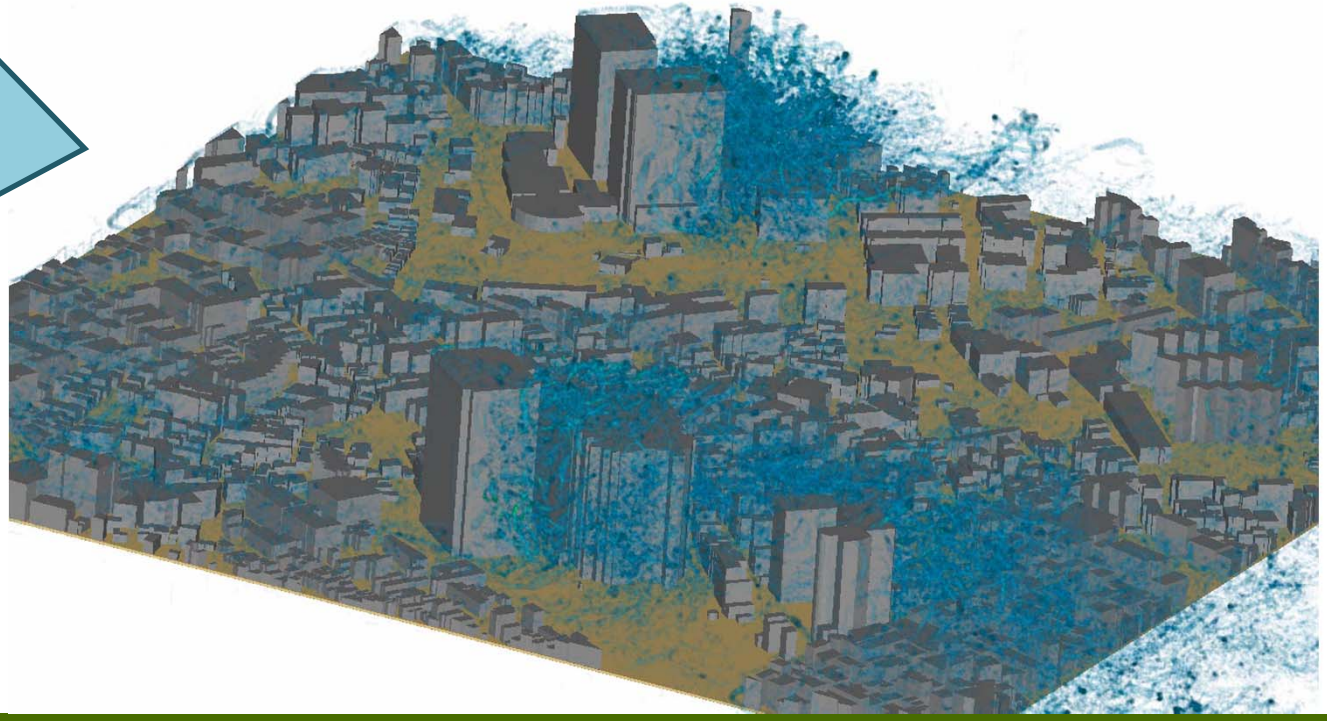
**8 km x 4 km**



**Building Data**



***Wind***



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

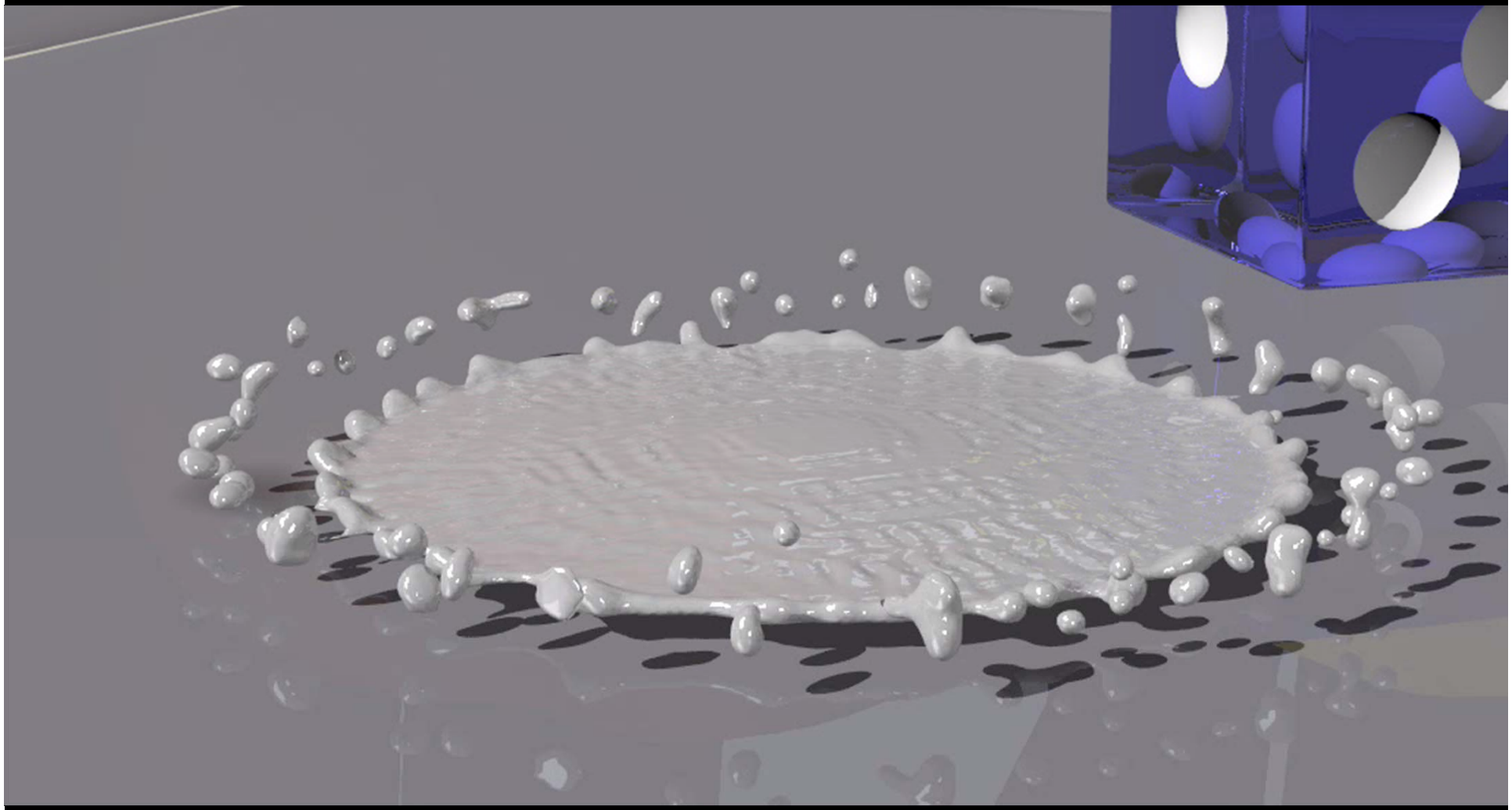
# Water-air mixing

# Two-Phase Flows

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

# Milk Crown

# Drop on dry floor



# Two-Phase Flow Simulation

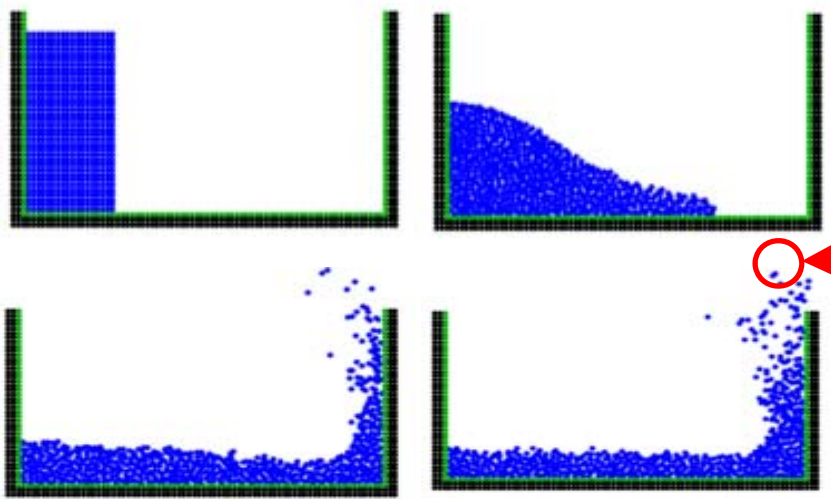
GP GPU

**Particle Method**
**ex. SPH**

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

**Mesh Method (Surface Capture)**

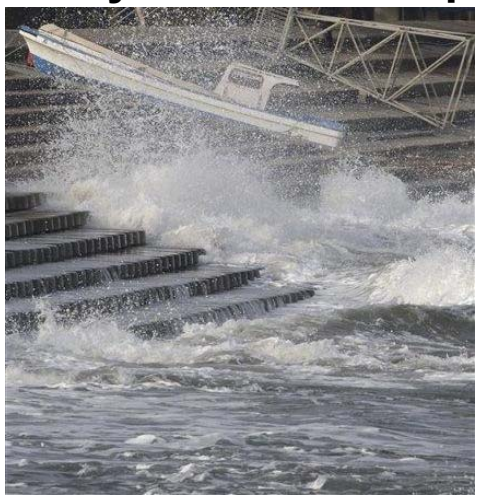
- Navier-Stokes solver : Fractional Step
- Time integration : 3rd TVD Runge-Kutta
- Advection term : 5th WENO
- Diffusion term : 4th FD
- Poisson : MG-BiCGstab
- Surface tension : CSF model
- Surface capture : CLSVOF(THINC + Level-Set)



not splash

Numerical noise and unphysical oscillation

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



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

# Level-Set method (LSM)

GP GPU

The Level-Set methods (LSM) use the signed distance function to capture the interface. The interface is represented by the zero-level set (zero-contour).

$\phi$ **: Level-Set function(distance function)**

$H$ : Heaviside function

$$\begin{cases} H(\phi) = \dfrac{1}{2} & \phi > \varepsilon \\ H(\phi) = \dfrac{1}{2}\left(\dfrac{\phi}{\varepsilon} + \dfrac{1}{\pi}\sin\left(\dfrac{\pi\phi}{\varepsilon}\right)\right) & |\phi| \le \varepsilon \\ H(\phi) = -\dfrac{1}{2} & \phi < -\varepsilon \end{cases}$$

Re-initialization for Level-Set function

$$\frac{\partial \phi}{\partial \tau} = sgn(\phi)\left(1 - |\nabla \phi|\right)$$

Advantage : Curvature calculation, Interface boundary

Drawback : Volume conservation

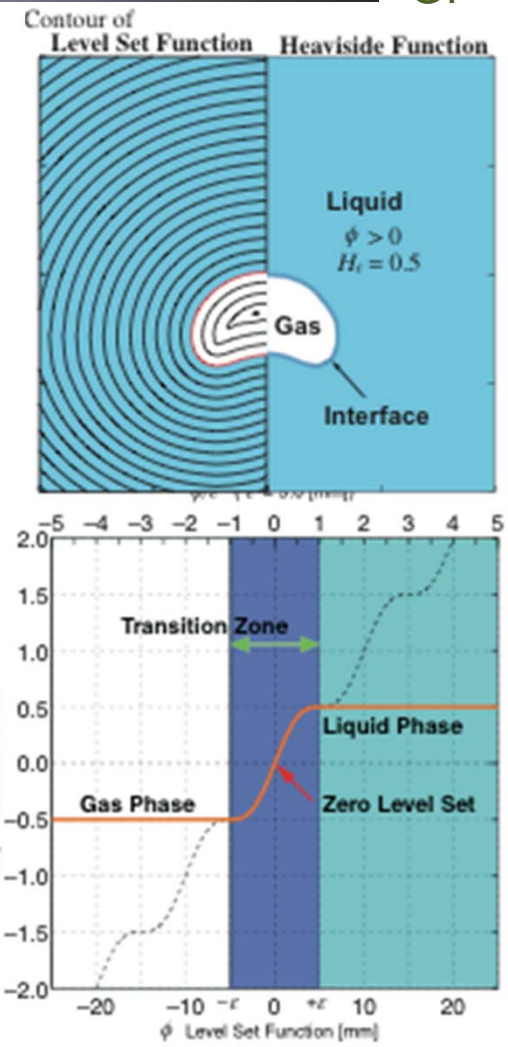



Fig. Takehiro Himeno, et. Al., JSME, 65-635,B(1999),pp2333-2340

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

# Anti-diffusive Interface Capture

GP GPU

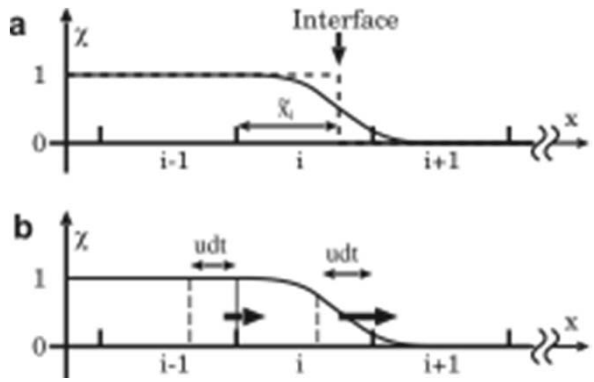
## THINC (tangent of hyperbola for interface capturing) Scheme

[ Xiao, etal, Int. J. Numer. Meth. Fluid. 48(2005)1023 ]

- VOF(volume of fluid) type interface capturing method
- Flux from tangent of hyperbola function
- Semi-Lagrangian time integration

$$F_i(x) = \frac{1}{2}\left(1 + \alpha \tanh\left(\beta\left(\frac{x - x_{i-1/2}}{\Delta x} - \tilde{x}_i\right)\right)\right)$$

$$\alpha = \begin{cases} 1 & (\text{if} \quad n_x > 0) \\ -1 & (\text{if} \quad n_x \leq 0) \end{cases}$$




- 1D implementation can be applied to 2D & 3D → Simple

$$Fl_{x,i+1/2} = -\int_{x_{i+1/2}}^{x_{i+1/2}-u_{i+1/2}\Delta t} F_{up}(x)\ dx \qquad up = \begin{cases} i & (\text{if} \quad u_{i+1/2} > 0) \\ i+1 & (\text{if} \quad u_{i+1/2} \leq 0) \end{cases}$$

- Finite Volume like usage
  - ∗ THINC is the method how to compute flux

→ 3 krenel (x, y, z) can be fused to 1 kernel. Merit in memory R/W

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

# Continuous Surface Force (CSF) model by Brackbill, Kothe and Zemach (1991)

The interfacial surface force is transformed to a volume force in the region near the interface via a delta function

Curvature

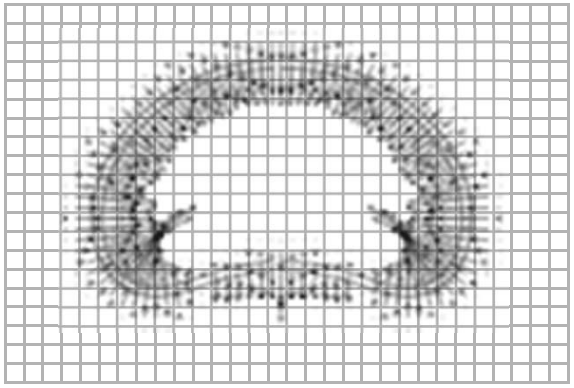
Surface tension force

$$\boxed{\mathbf{F}_S = \sigma\ \kappa\ \mathbf{n}}$$

Normal vector

$$\kappa = -\nabla \cdot \mathbf{n} = -\nabla \cdot \frac{\nabla \phi}{|\nabla \phi|}$$

$$\boxed{\mathbf{F}_S = \sigma\ \kappa\ \delta(\phi)\ \nabla \phi}$$

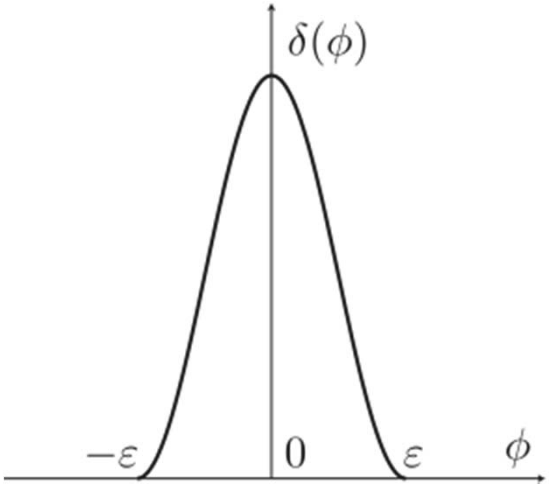


Surface tension represented by volume force

Approximate delta function

$$\delta(\phi) = \frac{\partial H(\phi)}{\partial \phi} = \frac{1}{2}\left(\frac{1}{\varepsilon} + \frac{1}{\varepsilon}\cos\left(\frac{\pi\phi}{\varepsilon}\right)\right)$$

$$\int_{-\varepsilon}^{\varepsilon} \delta(\phi)\ d\phi = 1$$




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

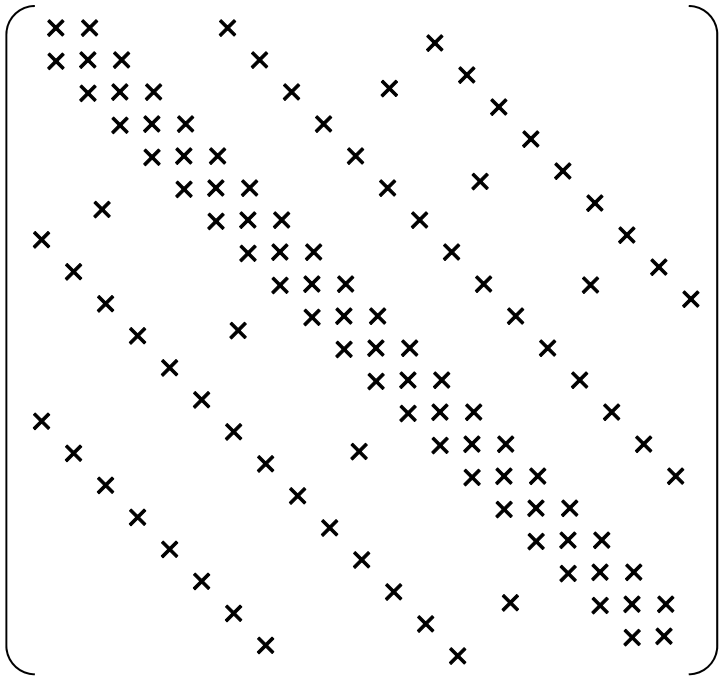
# Sparse Matrix Solver

$$\mathbf{A}x = b \qquad \textit{for} \quad \nabla \cdot \left( \frac{1}{\rho} \nabla p \right) = \frac{\nabla \cdot \boldsymbol{u}}{\Delta t}$$

**Krylov sub-space methods:**
**CG, BiCGStab, GMRes, , ,**

**Pre-conditioner:**
**Incomplete Cholesky,**
**ILU, MG, AMG,**
**Block Diagonal Jacobi**

**Non-zero Packing:**
**CRS → ELL, JDL**



# BiCGStab + (A)MG

GP GPU

**Collaboration with**
**Mizuho Information & Research Institute**

**Set** $k = 0 \quad \boldsymbol{r}_0 = \boldsymbol{p}_0 = \boldsymbol{M}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0)$

**for** $\boldsymbol{k = 0; k < N; k{+}{+};}$

$$\alpha_k = \frac{(\boldsymbol{r}_0, \boldsymbol{r}_k)}{(\boldsymbol{r}_0, \boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{p}_k)} \qquad \boldsymbol{q}_k = \boldsymbol{r}_k - \alpha_k \boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{p}_k \qquad \omega_k = \frac{(\boldsymbol{q}_k, \boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{q}_k)}{(\boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{q}_k, \boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{q}_k)}$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k + \omega_k \boldsymbol{q}_k$$

$$\boldsymbol{r}_{k+1} = \boldsymbol{q}_k - \omega_k \boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{q}_k$$

**if** $(\boldsymbol{r}_{k+1}, \boldsymbol{r}_{k+1}) < \varepsilon^2(\boldsymbol{b}, \boldsymbol{b})$ **exit;**

$$\beta_k = \frac{(\boldsymbol{r}_0, \boldsymbol{r}_{k+1})}{\omega_k(\boldsymbol{r}_0, \boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{p}_k)}$$

$$\boldsymbol{p}_{k+1} = \boldsymbol{r}_{k+1} + \beta_k(\boldsymbol{p}_k - \omega_k \boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{p}_k)$$

**loop end**

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

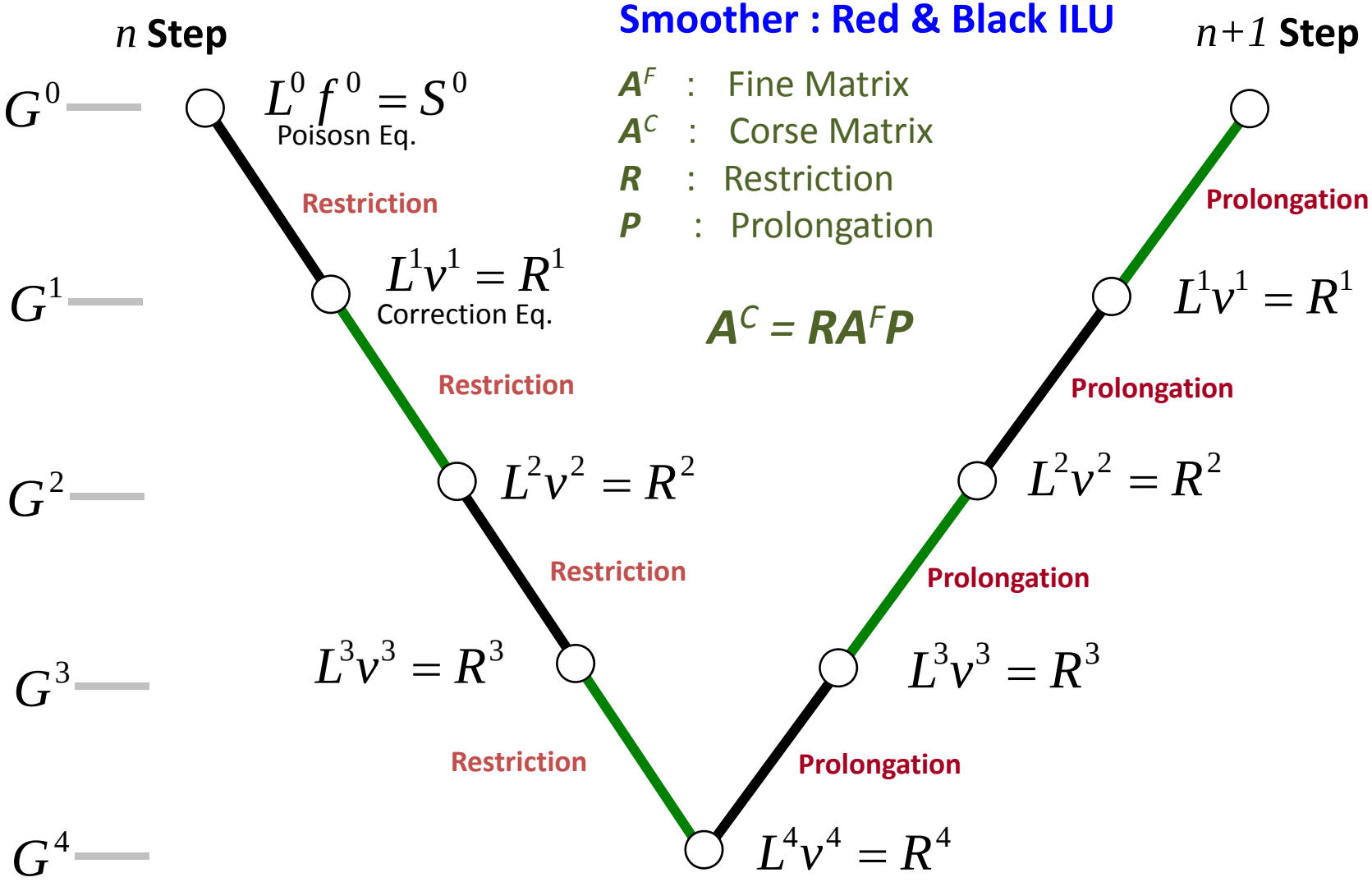
# MG V-Cycle

GP GPU

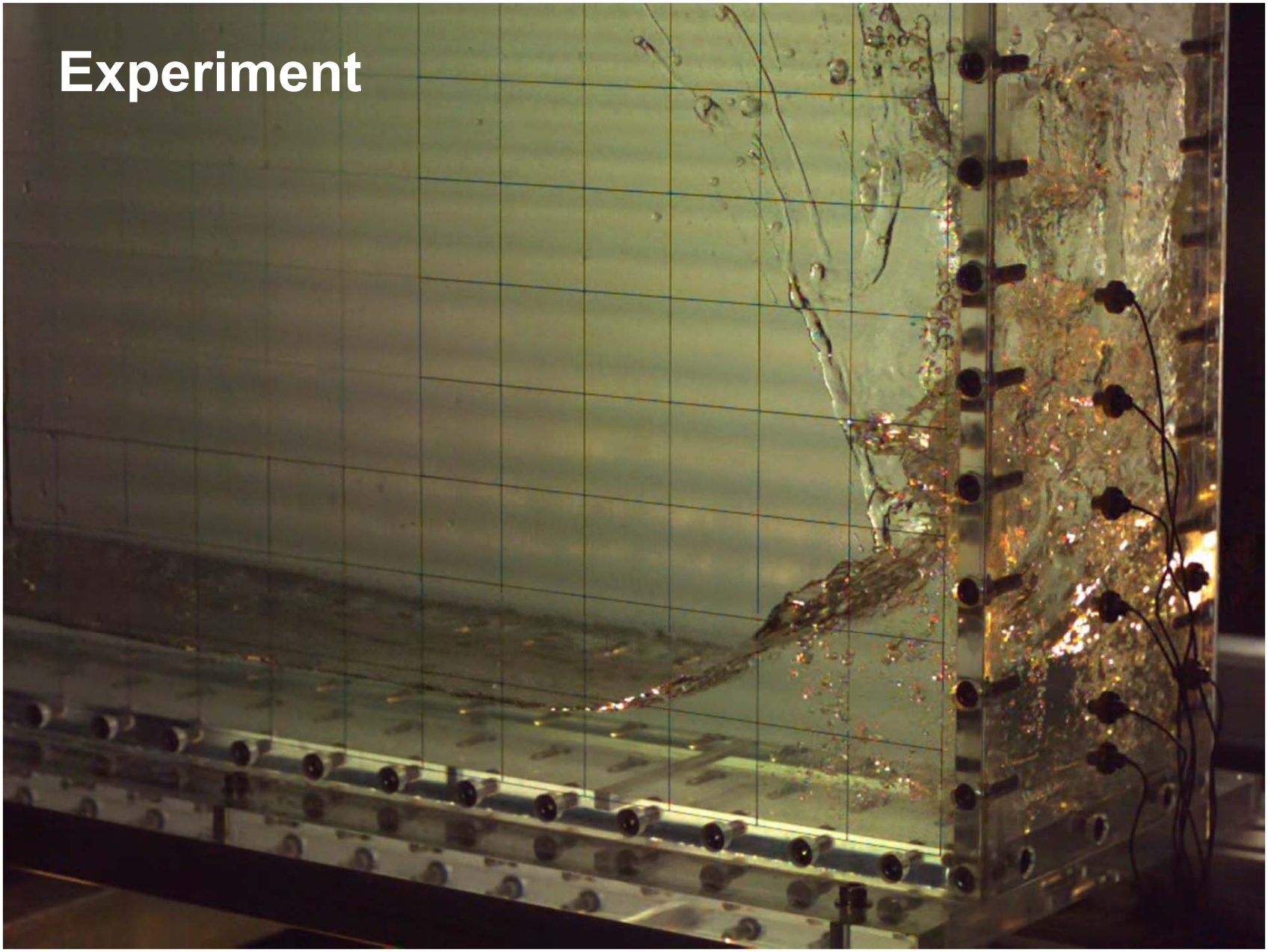
**Smoother : Red & Black ILU**

$A^F$ : Fine Matrix
$A^C$ : Corse Matrix
$R$ : Restriction
$P$ : Prolongation

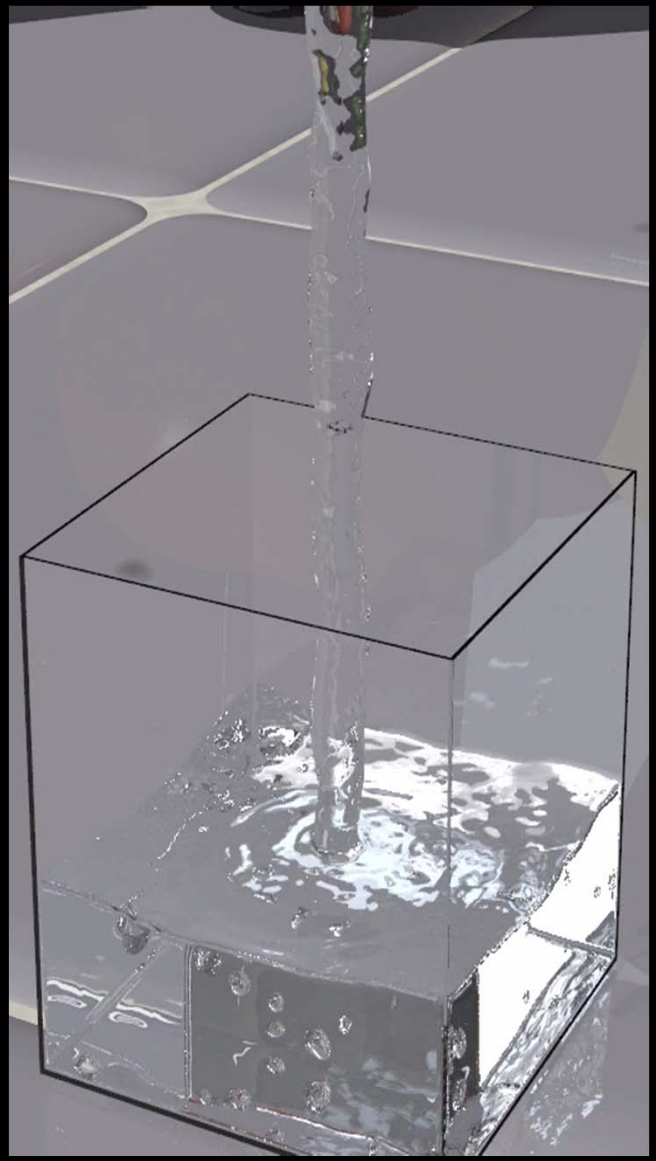
$$A^C = RA^FP$$

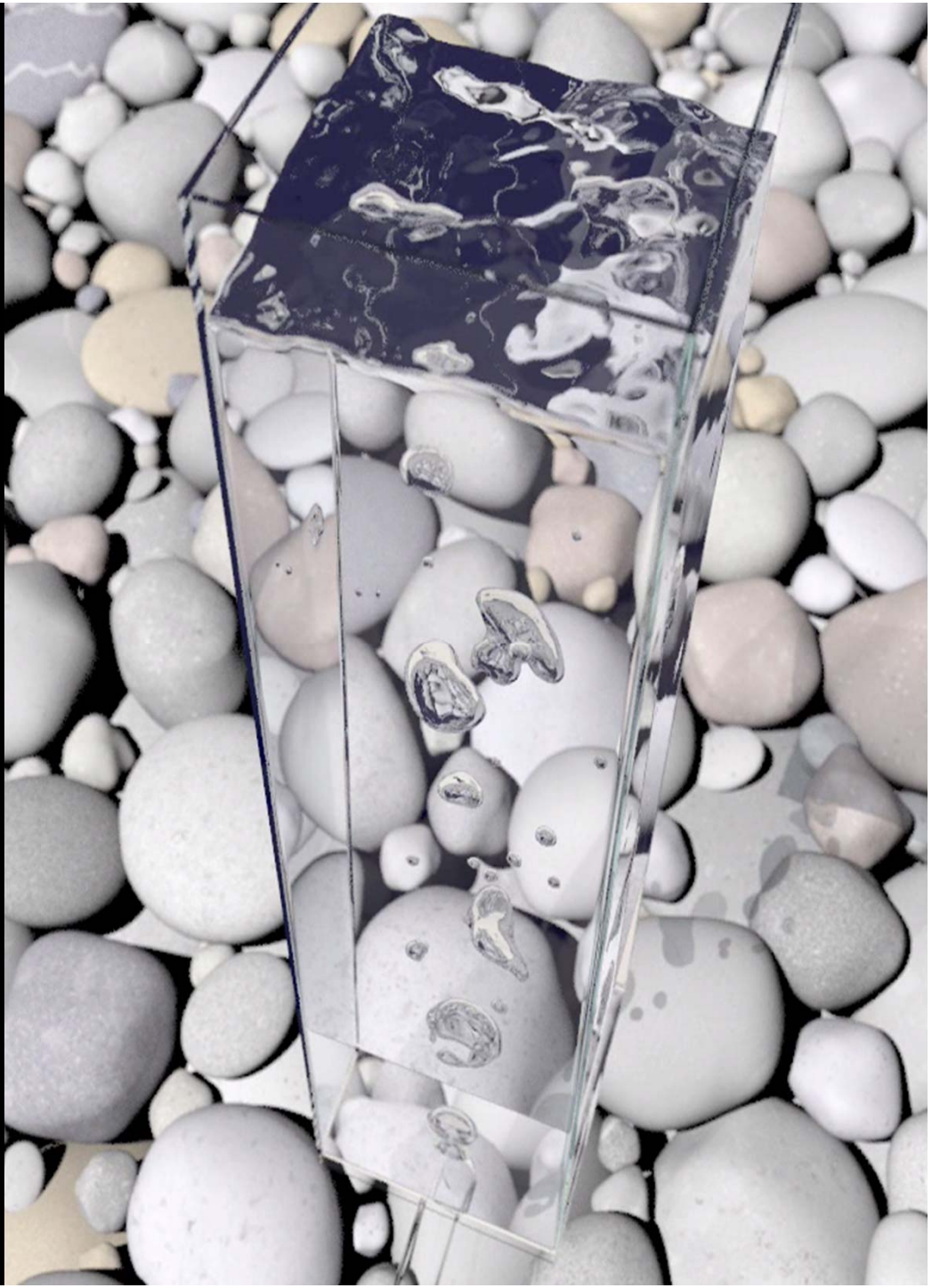



Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology




Collaboration: Prof. Hu and Dr. Sueyoshi, RIAM, Kyusyu University





# Industrial Appl. Steering Oil

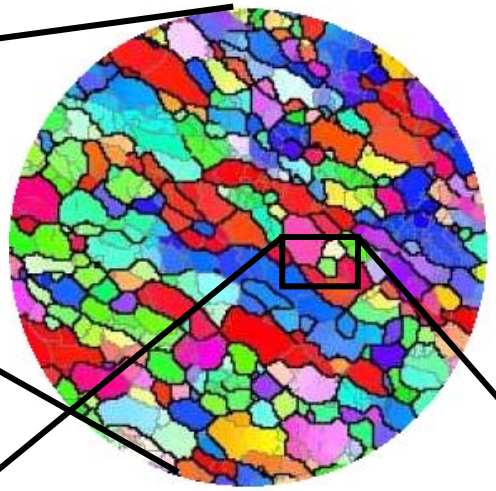
Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

# Development New Materials

GP GPU

Mechanical Structure

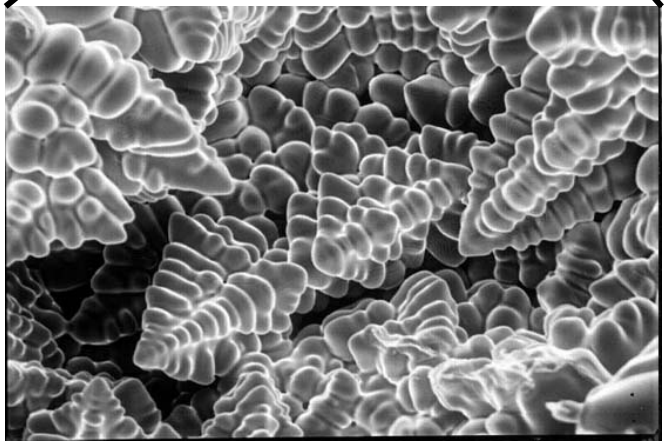
Material Microstructure



**Low-carbon society**

Improvement of fuel efficiency by reducing the weight of transportation

Developing lightweight strengthening material by controlling microstructure

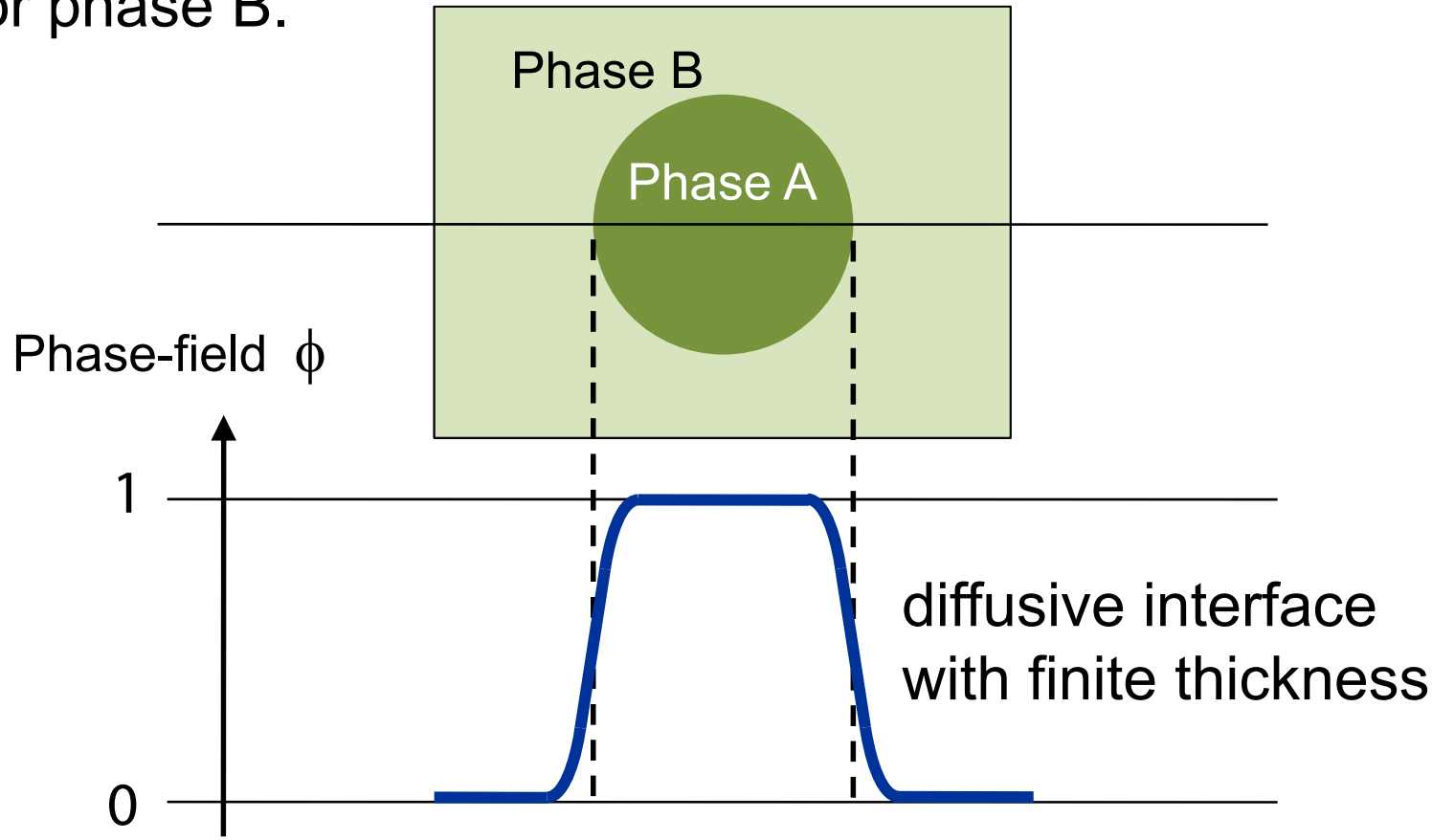


Dendritic Growth

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

# 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.



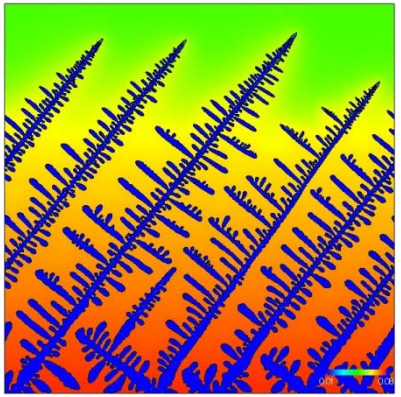

Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

# Requirement for Peta-scale Computing

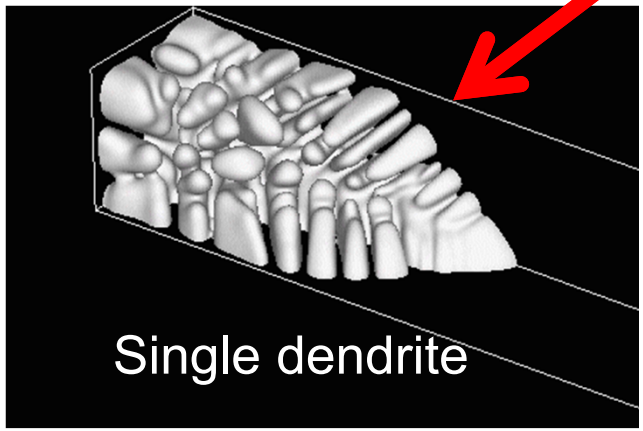
GP GPU

## Previous works

**2D computation**

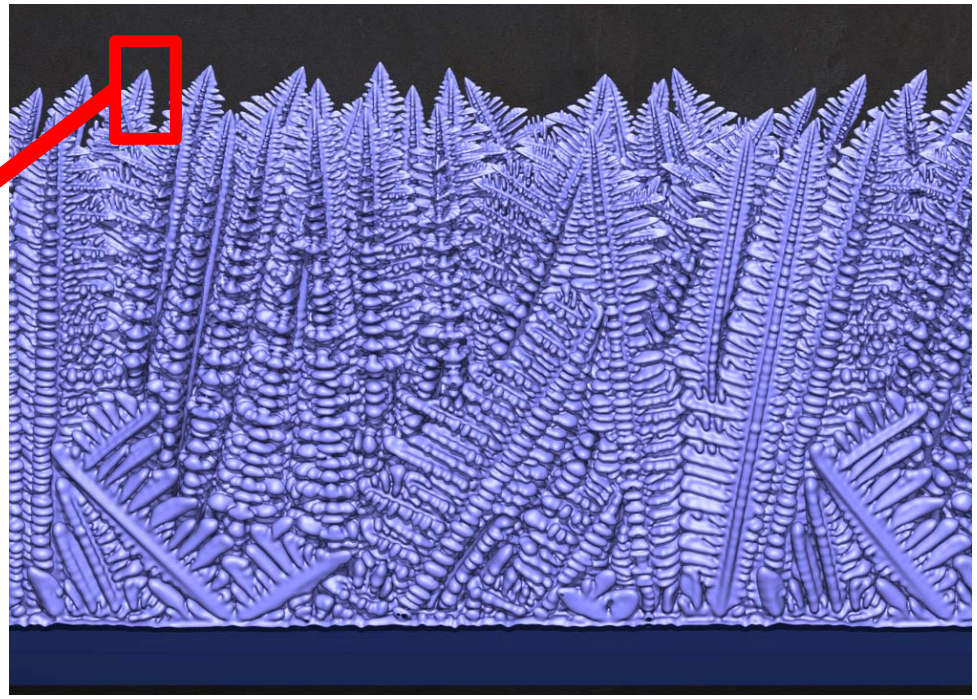


**3D computation**




**× 1000 large-scale computation**

**on TSUBAME 2.0**

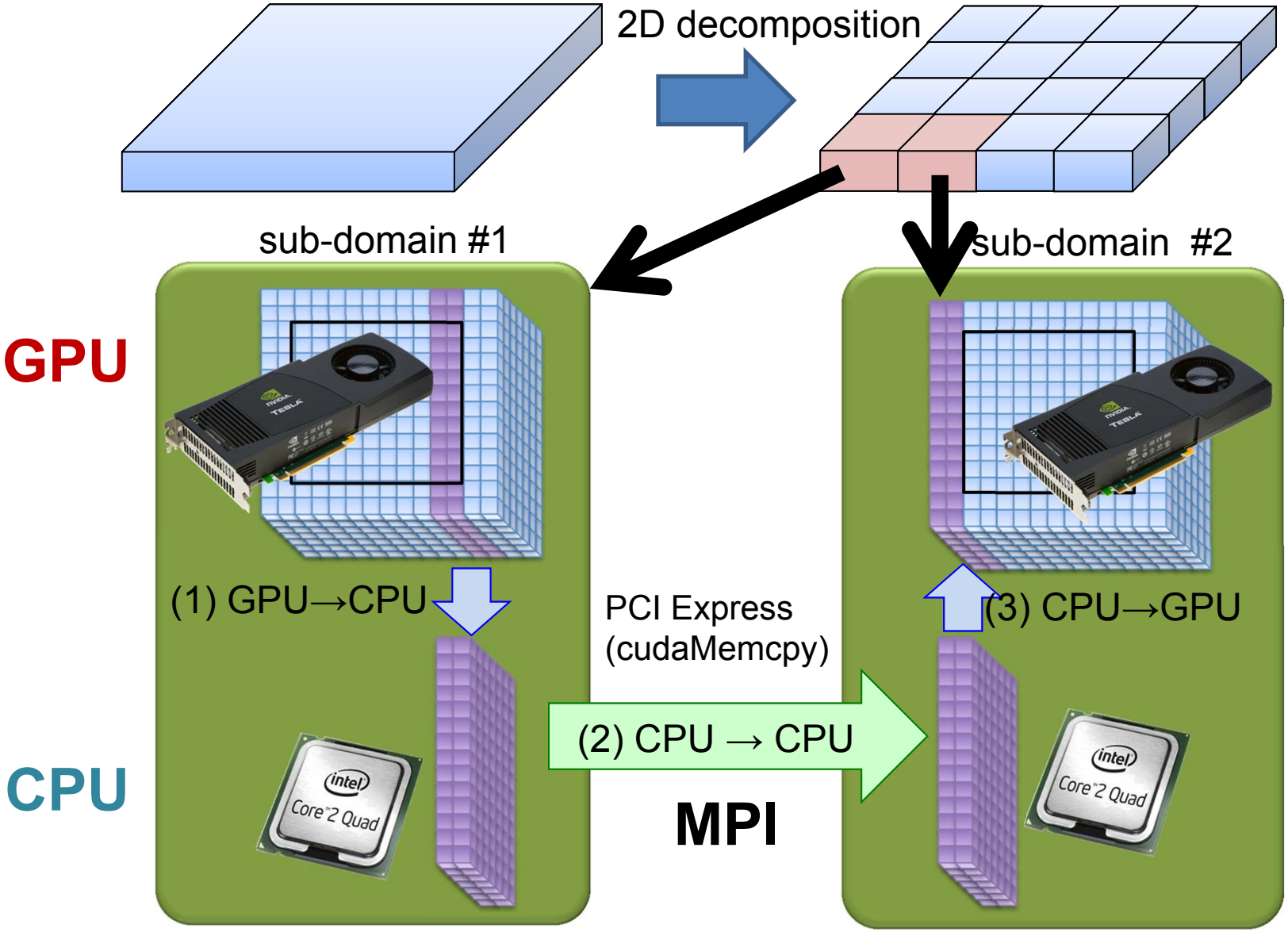


← ~ mm scale →

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

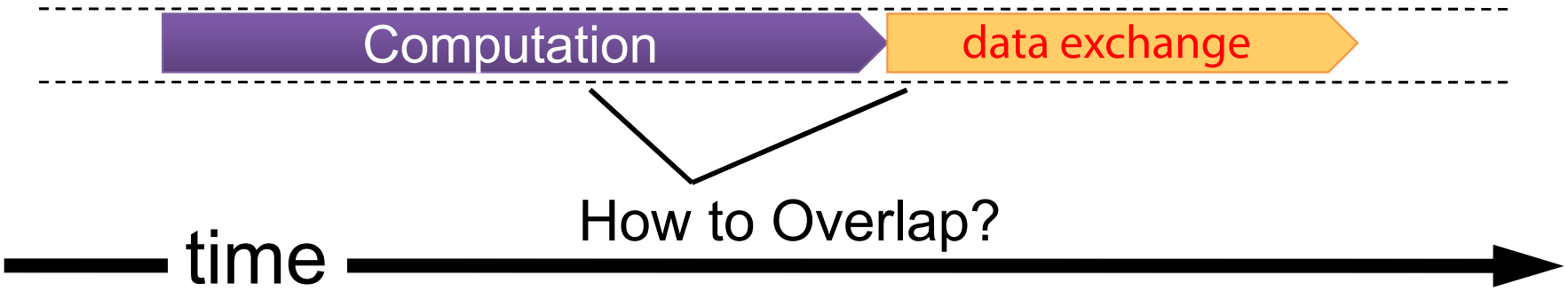
# Multi-GPU : Domain decomposition

GP GPU



# Multi-GPU Optimization

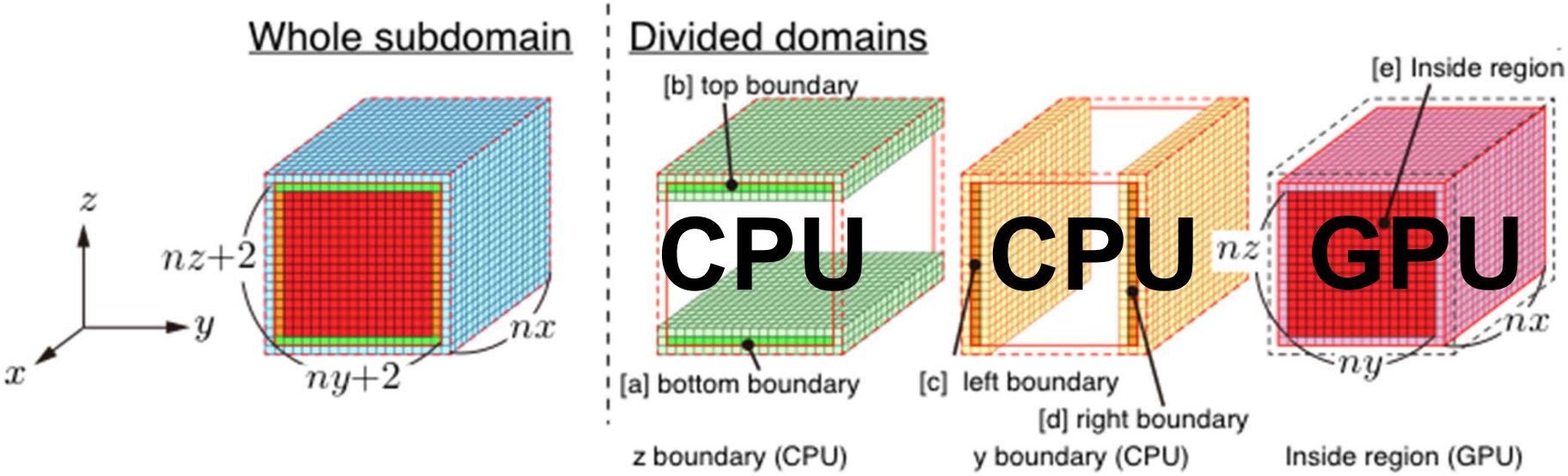
Typical explicit time integration

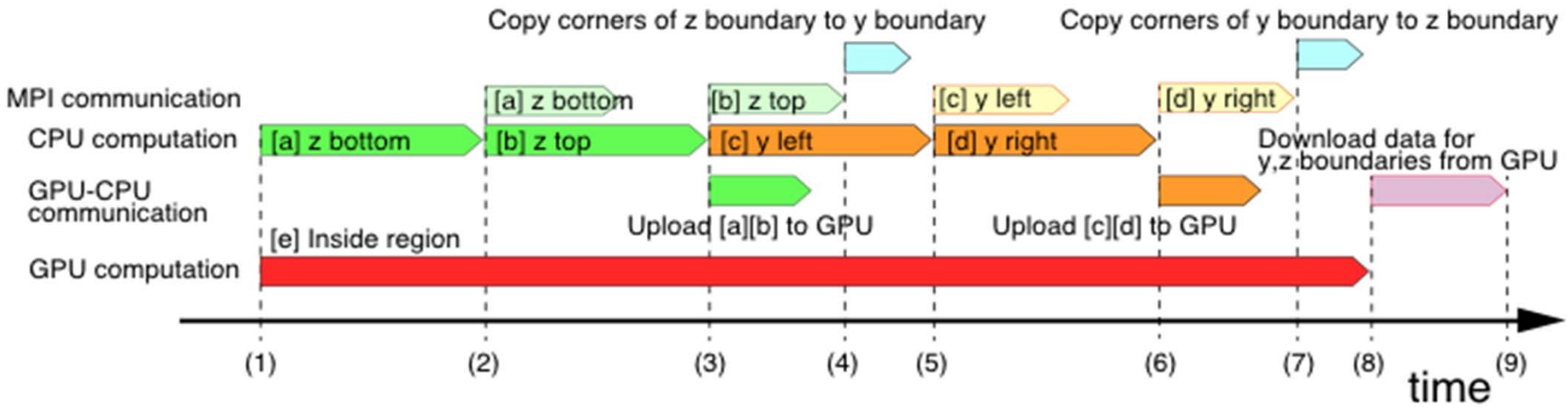



1. **GPU-only Method (without overlapping)**
2. **Hybrid-YZ Method**
3. **Hybrid-Y Method**

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

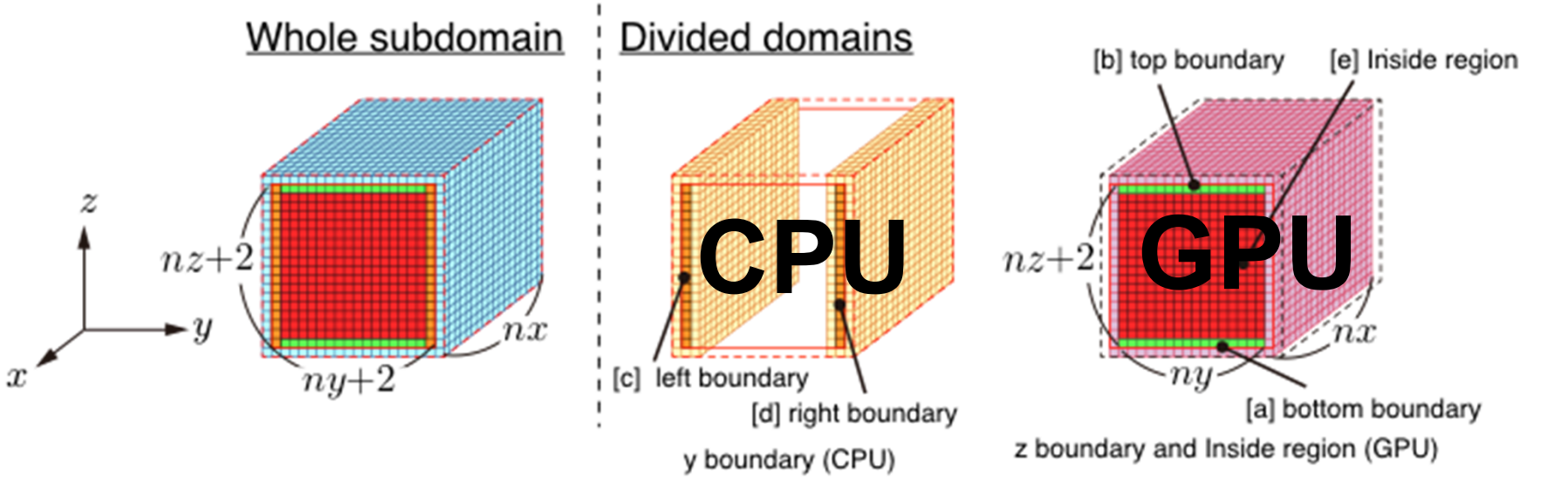
# 2. Hybrid-YZ method

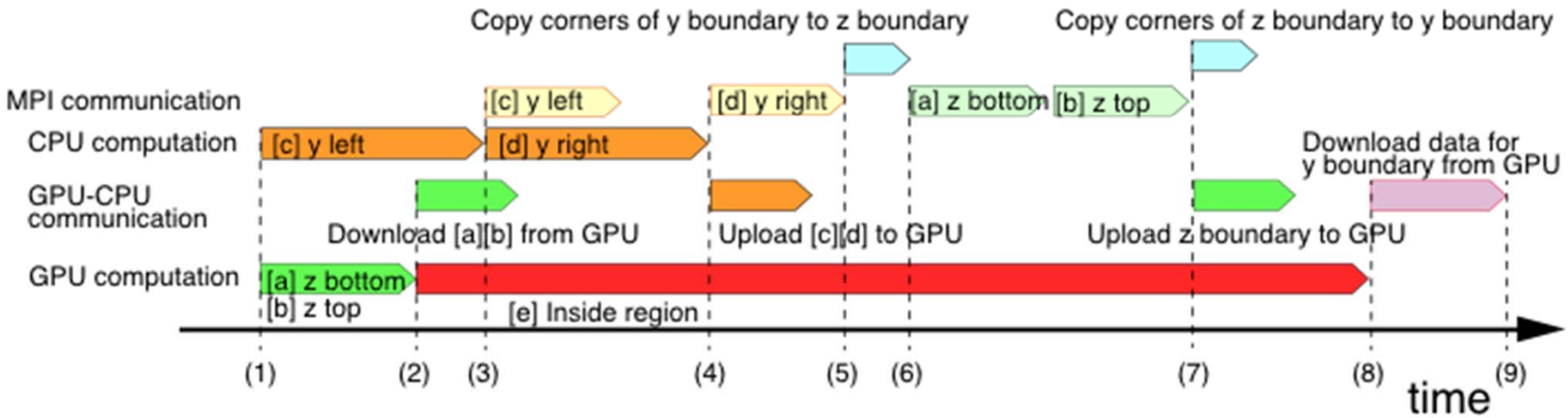



Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

# 3. Hybrid-Y method

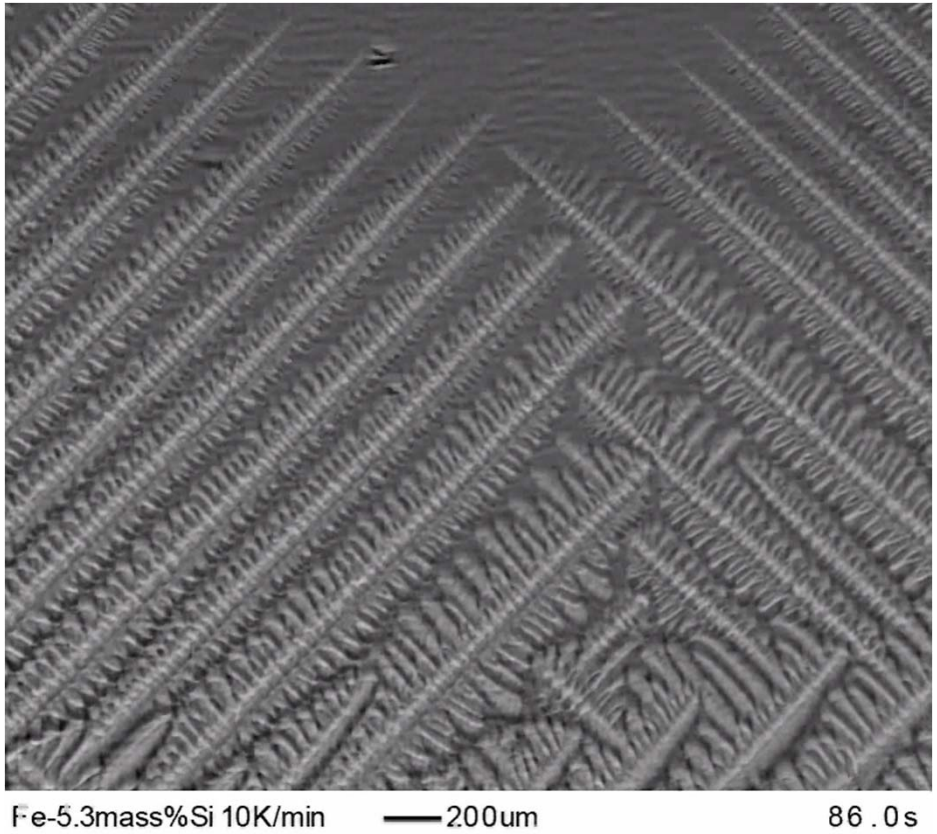



Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

# Comparison with Experiment

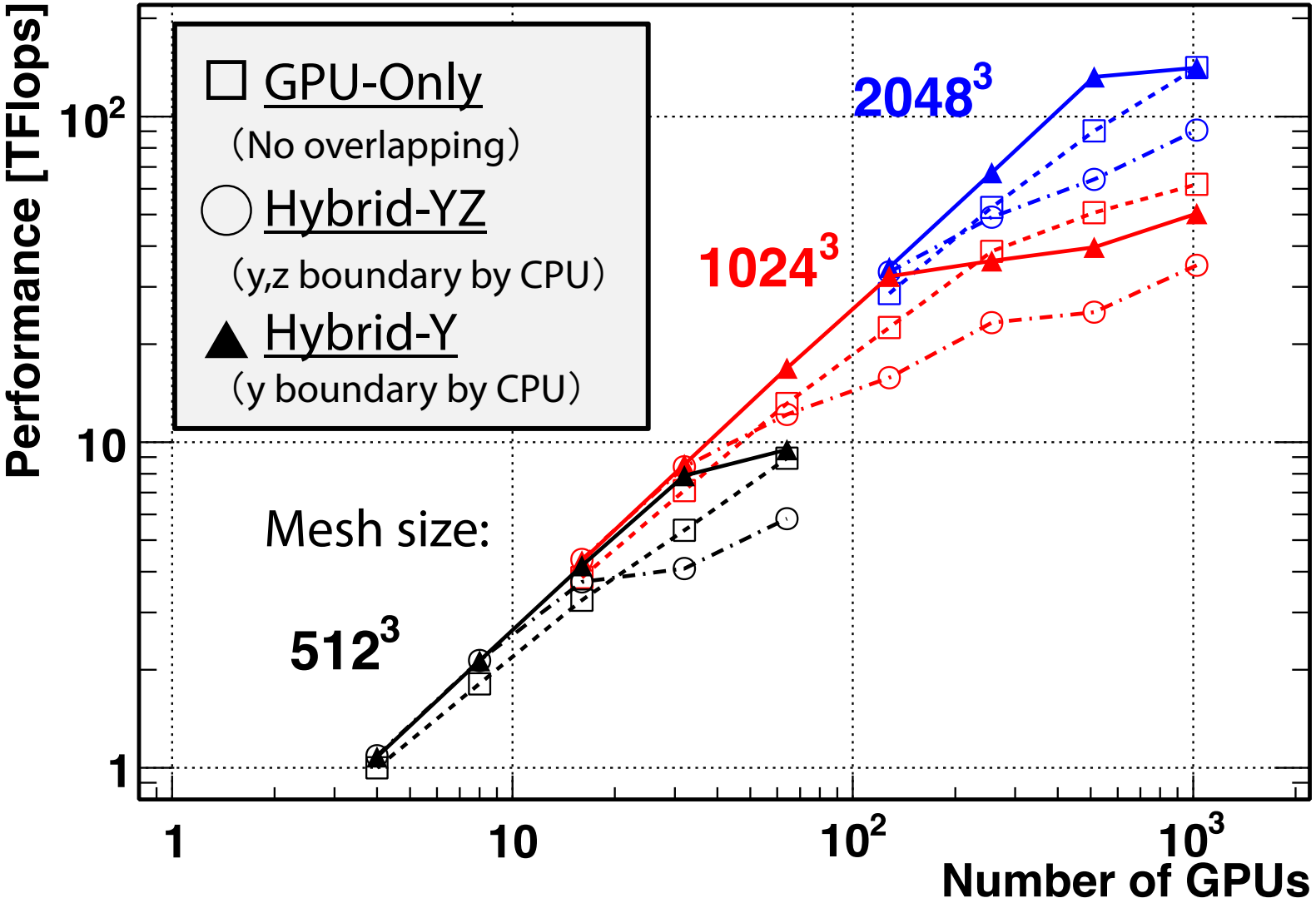
**Observation:**
X-ray imaging of Solidification of a binary alloy at Spring-8 in Japan by Prof. Yasuda (Osaka University in Japan)



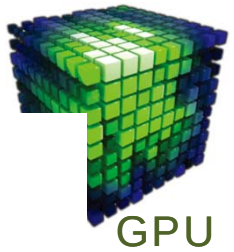

**Phase-field simulation**

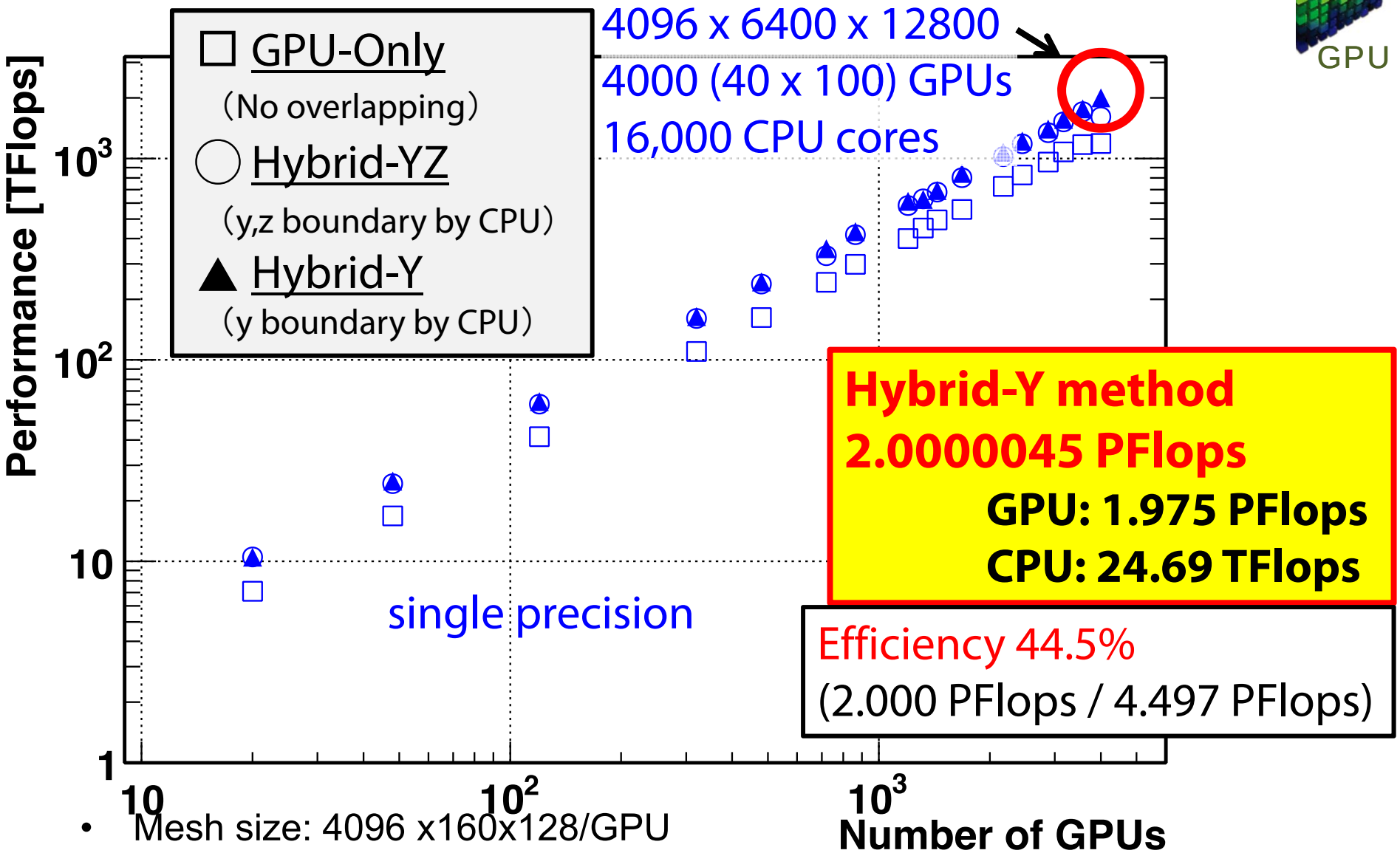
Copyright © Global Scientific Information and Computing Center, Tokyo Institute of Technology

# Strong Scalability



# 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*

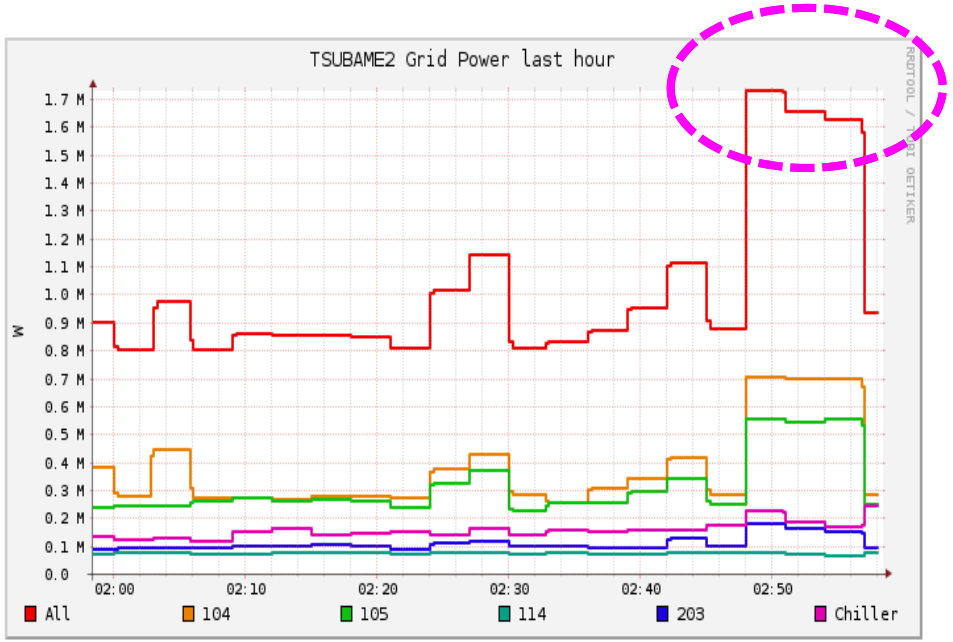
IEEE COMPUTER SOCIETY

# Power Efficiency

- 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 electrical 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

- **Large-scale GPU applications successfully running on TSUBAME2.0**
  - Meso-scale weather prediction
  - Lattice Boltzmann Method
  - Shallow  Water TSUNAMI simulation
  - Gas-Liquid Two-phase flow
  - Phase-Field model for developing new materials

- **GPU Green computing : much less electrical power to get application results.**

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