GTC Asia in Beijing 2011, Dec 15

AOKI Lab.

GPU-based operational Weather Model with Horizontal 500m resolution

Takayuki Aoki

Global Scientific Information and Computing Center (GSIC)

Tokyo Institute of Technology

TSUBAME 2.0

Compute Node
(2 CPUs, 3 GPUs)

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

Rack (30 nodes)

Performance: 51.0 TFLOPS
Memory: 2.03 TB

System (58 racks)

1442 nodes: 2952 CPU sockets,
4264 GPUs

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

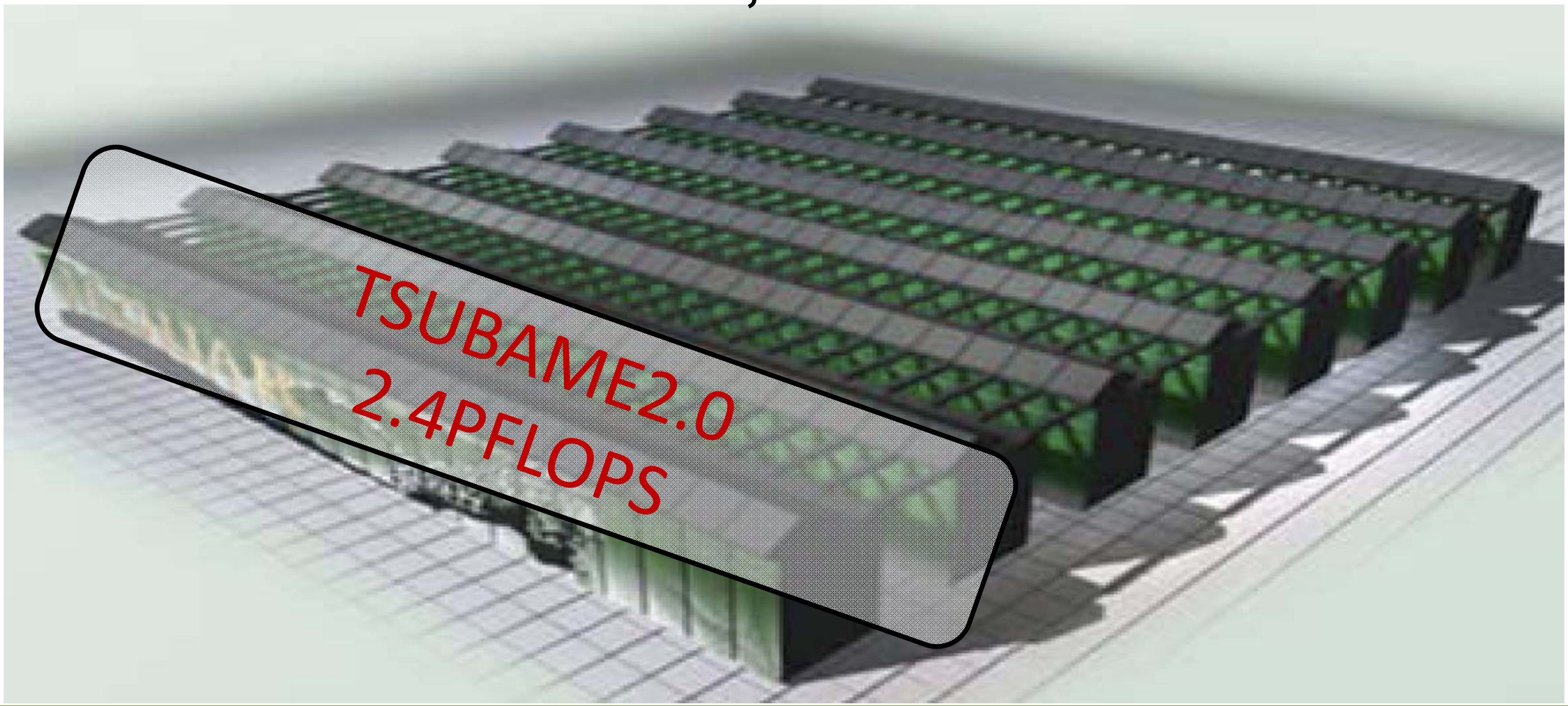
Supercomputer in the world

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

ORNL Jaguar vs Tsubame 2.0

Similar Peak Performance, 1/5 the Size and Power



THE GREEN 500™ Supercomputer in the world

AOKI Lab.

November 2011

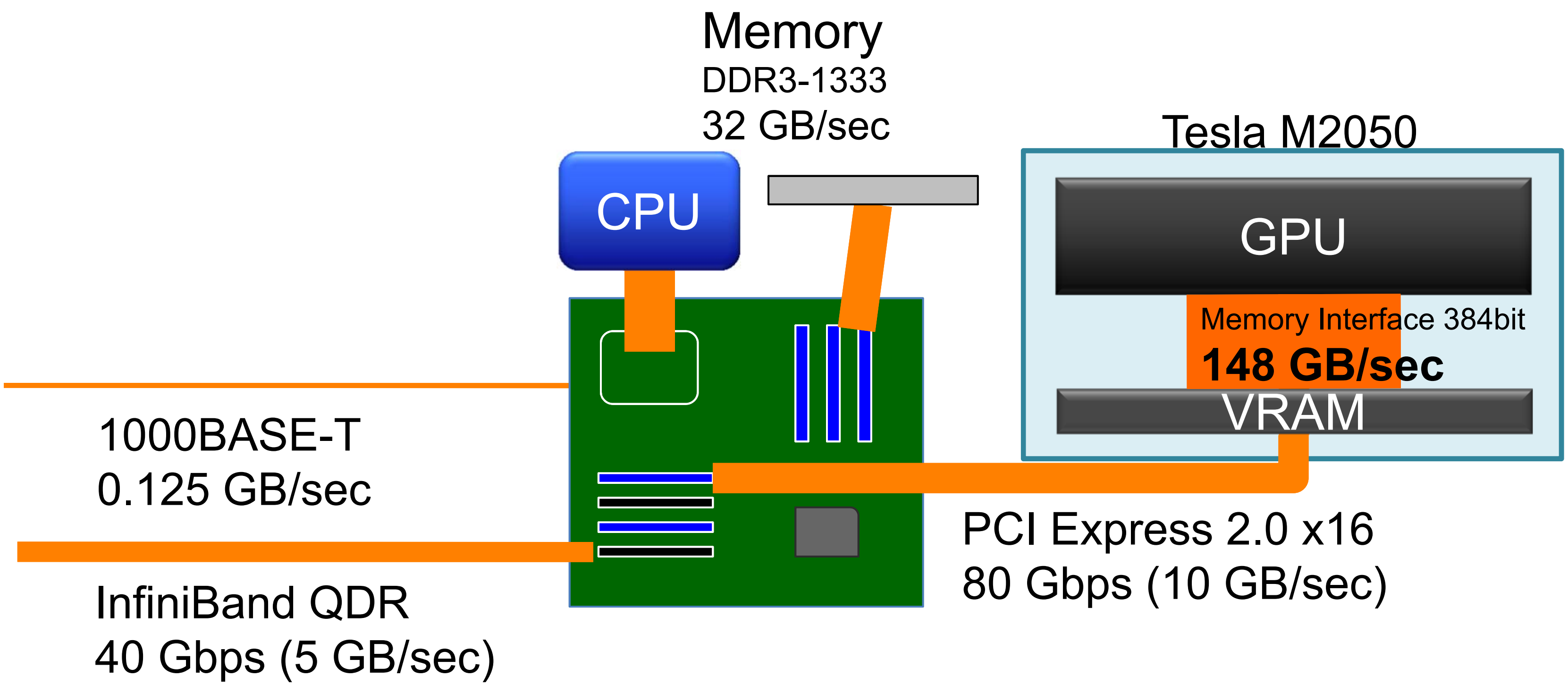
Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	2026.48	IBM - Rochester	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	85.12
2	2026.48	IBM Thomas J. Watson Research Center	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	85.12
3	1996.09	IBM - Rochester	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	170.25
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

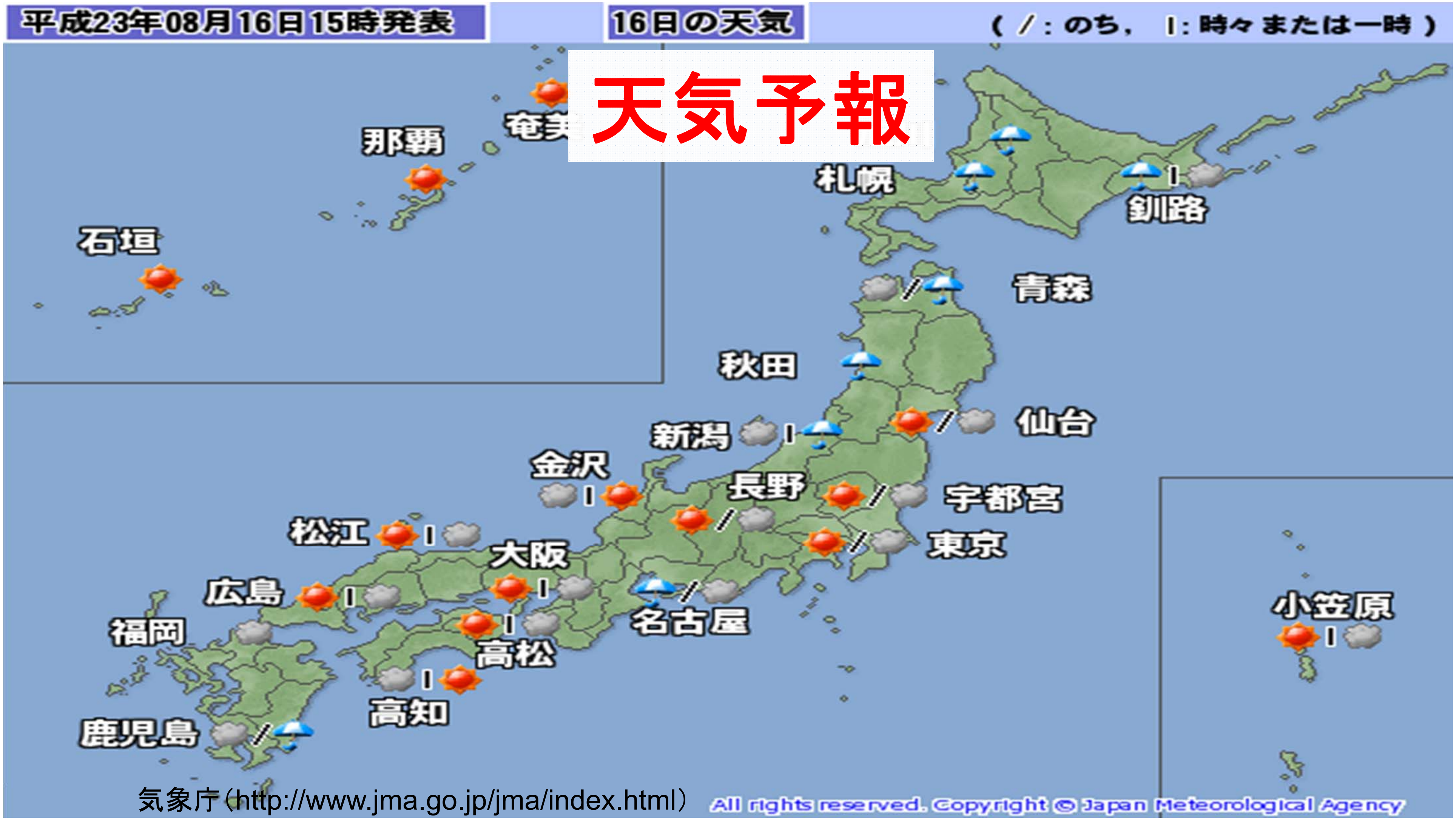
(Power Usage Effectiveness) **TSUBAME2.0 PUE = 1.2**

Heterogeneous Computer

■ **Several Bandwidth Bottle Necks**



天気予報



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

Next Generation

Weather Prediction

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

Atmosphere Model

Dynamical Process:

Full 3-D Navior-Stokes Equation

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} = -\frac{1}{\rho} \nabla P - 2\Omega \times \boldsymbol{u} - \Omega \times (\Omega \times \boldsymbol{r}) + \boldsymbol{g} + \boldsymbol{F}$$

Physical Process:

Cloud Physics, Moist, Solar Radiation, Condensation, Latent heat release, Chemical Process, Boundary Layer

“Parameterization” including sin, cos, exp, …in empirical rules.

WRF GPU Computing

■ WRF (Weather Research and Forecast)

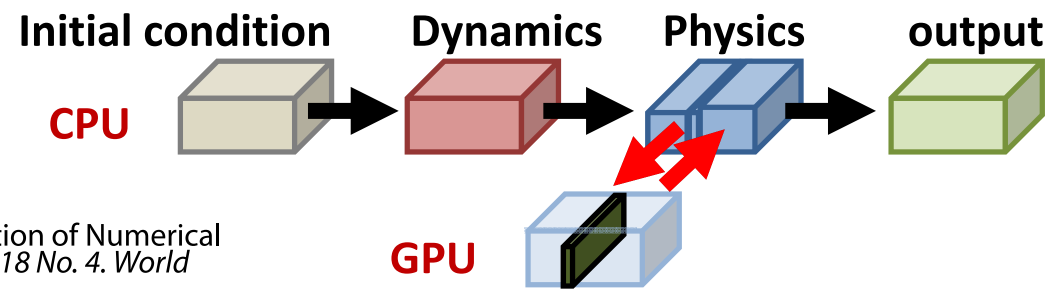
WSM5 (WRF Single Moment 5-tracer) Microphysics*

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

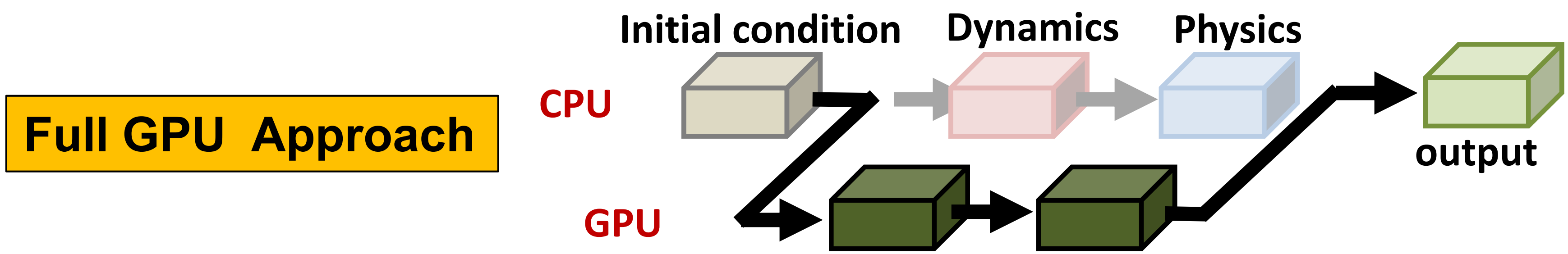
Accelerator Approach



*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.*

Full GPU Implementation: ASUCA



■ ***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.

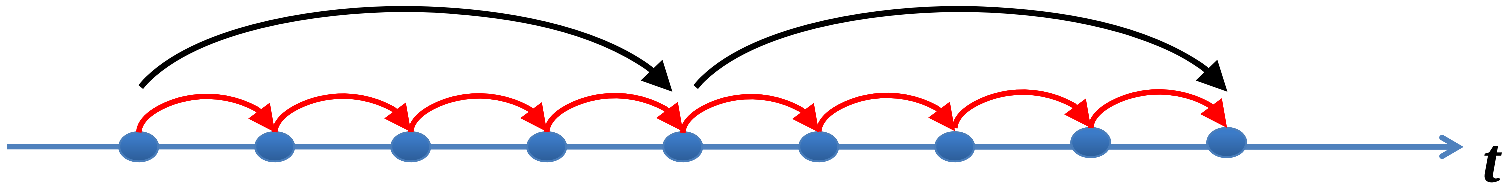
JMA-ASUCA

AOKI Lab.

- **ASUCA : developed by Japan Meteorological Agency (JMA) for the next-generation weather prediction**

Meso-scale non-hydrostatic Atmosphere Model

Time-splitting method: long time step for flow

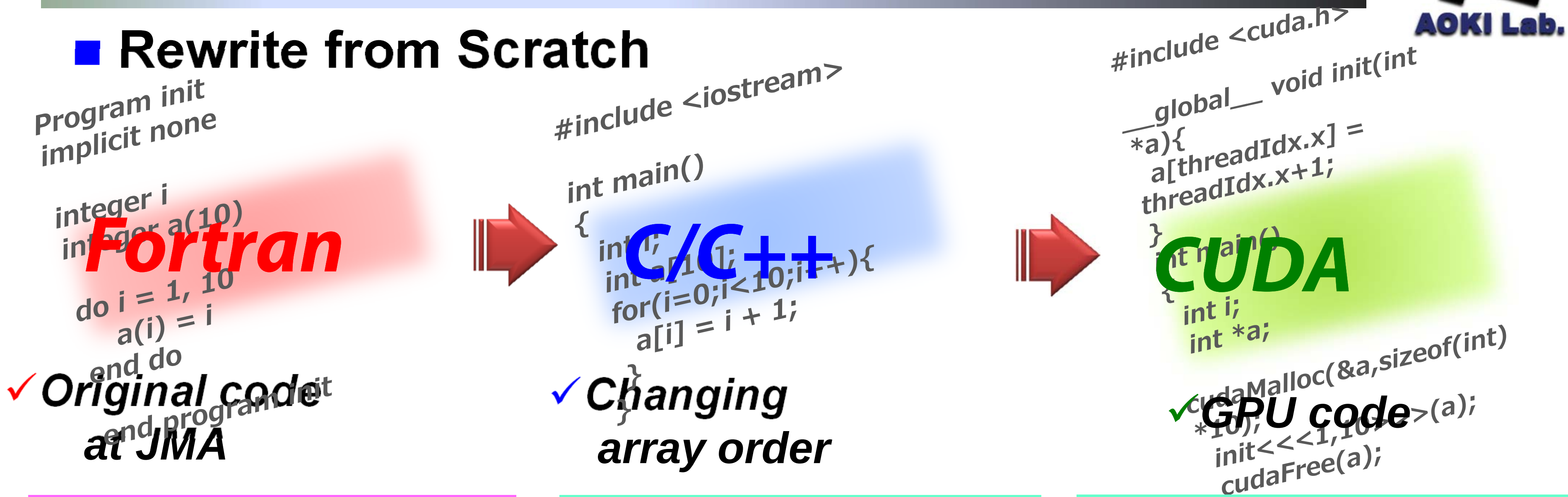


u, v (~ 100 m/s), w (~ 10 m/s) << sound velocity (~300/ms)

- ***Similar Structure as WRF***
  - ✓ ***HEVI (Horizontally explicit Vertical implicit) scheme***
  - ✓ ***Dynamical Core uses a numerical scheme with 3rd-order accuracy in time and space***
    - ***Flux-form non-hydrostatic compressible equation***
    - ***Generalized coordinate***

Entire Porting Fortran to CUDA

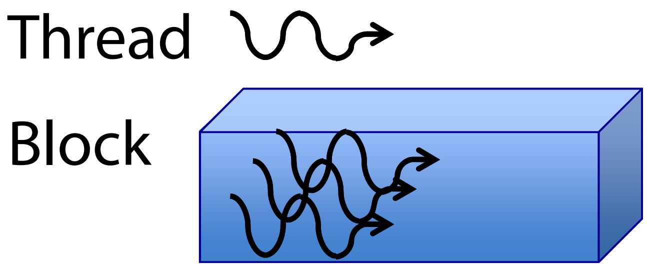
- **Rewrite from Scratch**

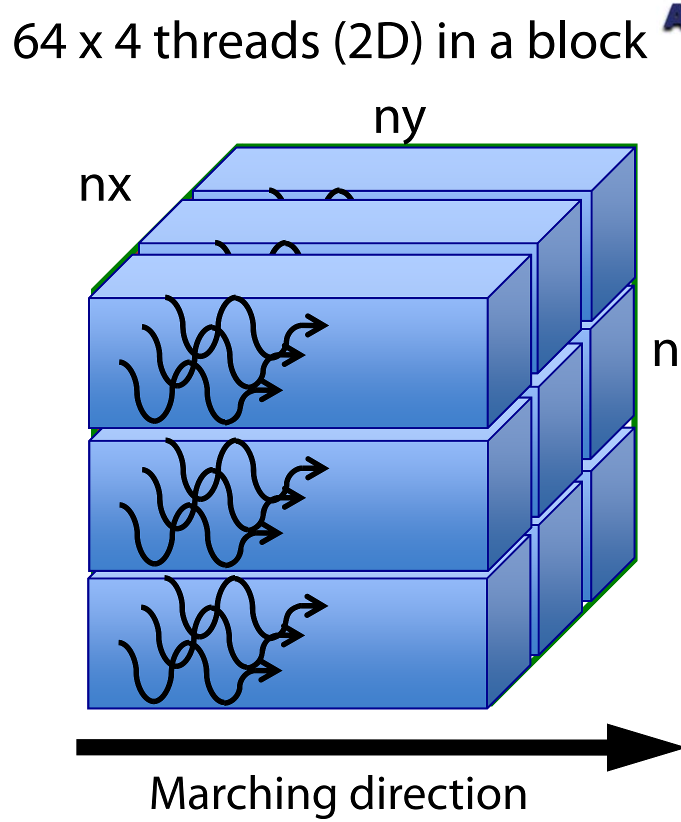


z,x,y **(k,i,j)-ordering**	*x,z,y* **(i,k,j)-ordering**	*x,z,y* **(i,k,j)-ordering**

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

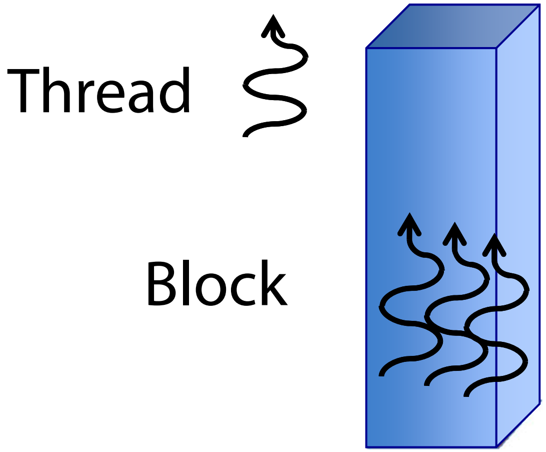
Implementation : Advection

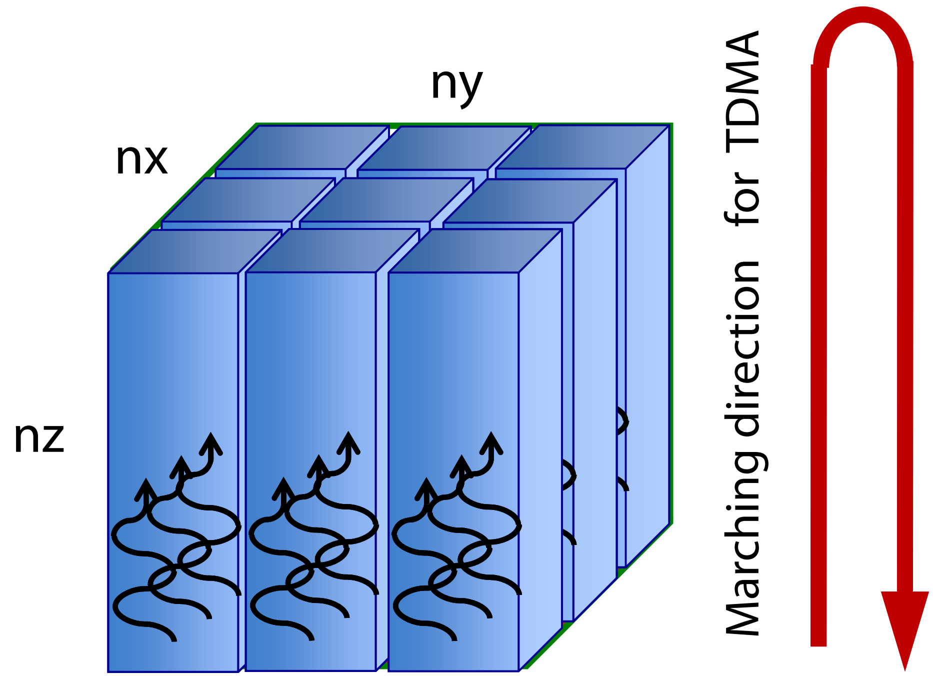




- Each thread specifies a (x, z) point、marching in y
  - ✓ Improve data transfer performance using domain decomposition

Implementation : 1D Helmholtz equation





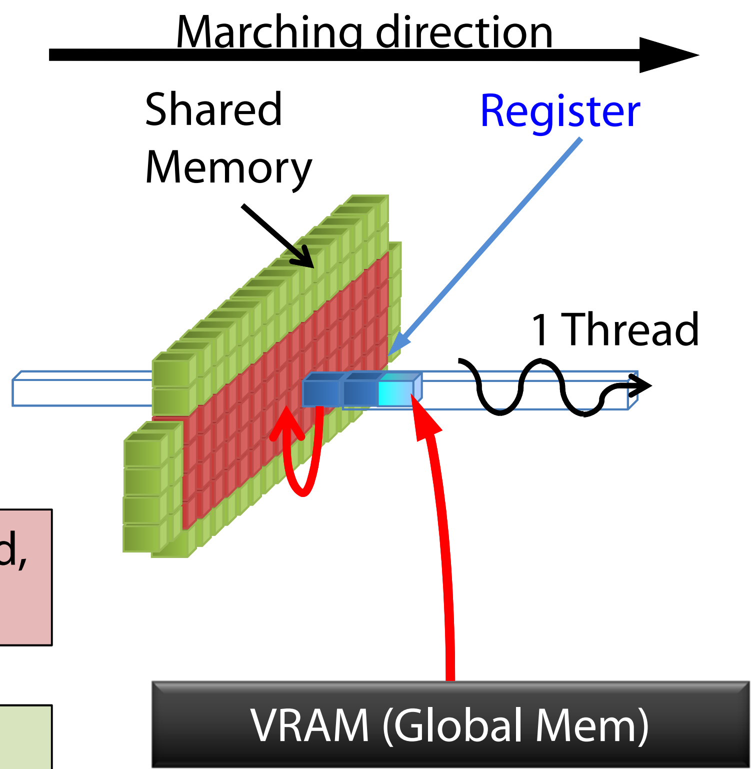
- 1D Helmholtz equation
  - ✓ Element in k depends on elements in k+/- 1
  - ⇒ marching in z direction

Using Registers in marching direction

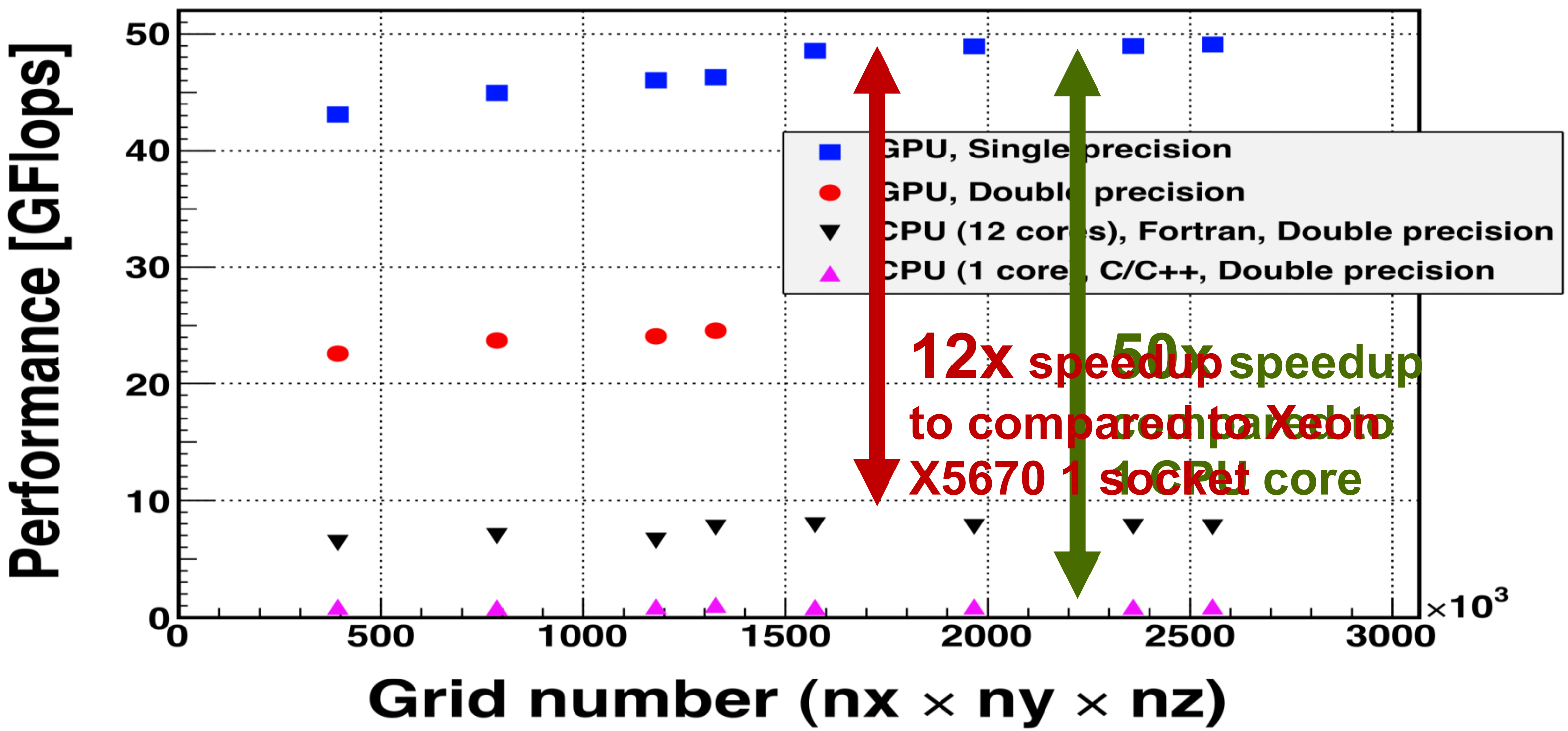
- Register
  - ✓ Access speed : 1 cycle
  - ✓ used for data not shared among threads
- Advection : 12-point stencil
  - ✓ Each thread keeps 4 y-elements in registers
  - ✓ Elements are reuse

Access GMem directly : 4 + 4 + 4 read, 1write

Using SMem and Registers : ~1 read, 1write



TSUBAME 2.0 (1 GPU)

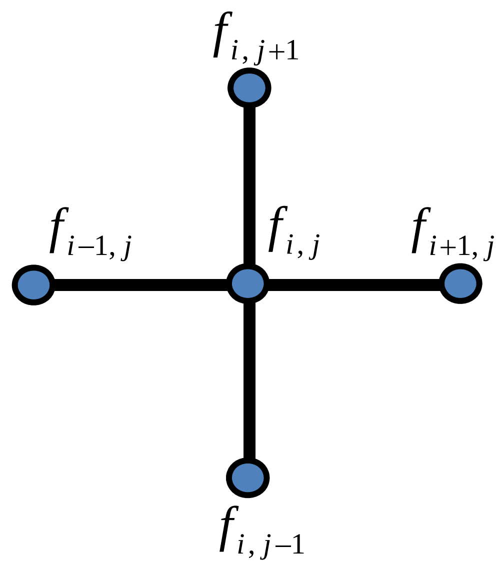


Arithmetic INTENSITY: FLOP/Byte

■ **Example: 2-dimensional diffusion Equation by FDM**

$$\frac{f_{i,j}^{n+1} - f_{i,j}^{n}}{\Delta t} = \kappa \left(\frac{f_{i+1,j}^{n} - 2 f_{i,j}^{n} + f_{i-1,j}^{n}}{\Delta x^2} + \frac{f_{i,j+1}^{n} - 2 f_{i,j}^{n} + f_{i,j-1}^{n}}{\Delta y^2} \right)$$

$$f_{i,j}^{n+1} = c_0 f_{i,j}^{n} + c_1 f_{i+1,j}^{n} + c_2 f_{i-1,j}^{n} + c_3 f_{i,j+1}^{n} + c_4 f_{i,j-1}^{n}$$



FLOP = 9
Byte = 8*(5+1) = 48 byte: read 5, write 1

FLOP/Byte = 9/48 = 0.1875 （DP）
= 9/24 = 0.375 （SP）

Achievable Performance

FLOP = number of FP operation for applications

Byte = Byte number of memory access for applications

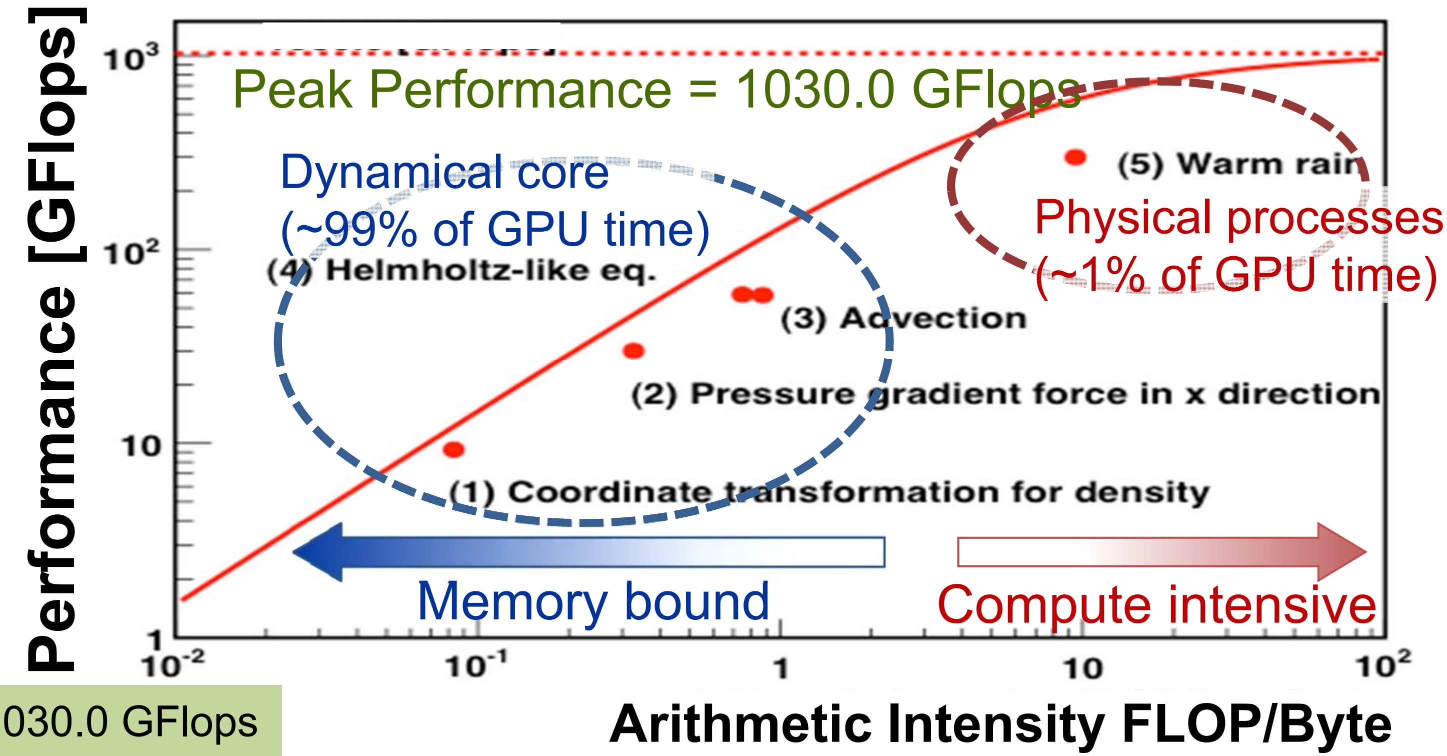
F = Peak Performance of floating point operation

B = Peak Memory Bandwidth

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

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

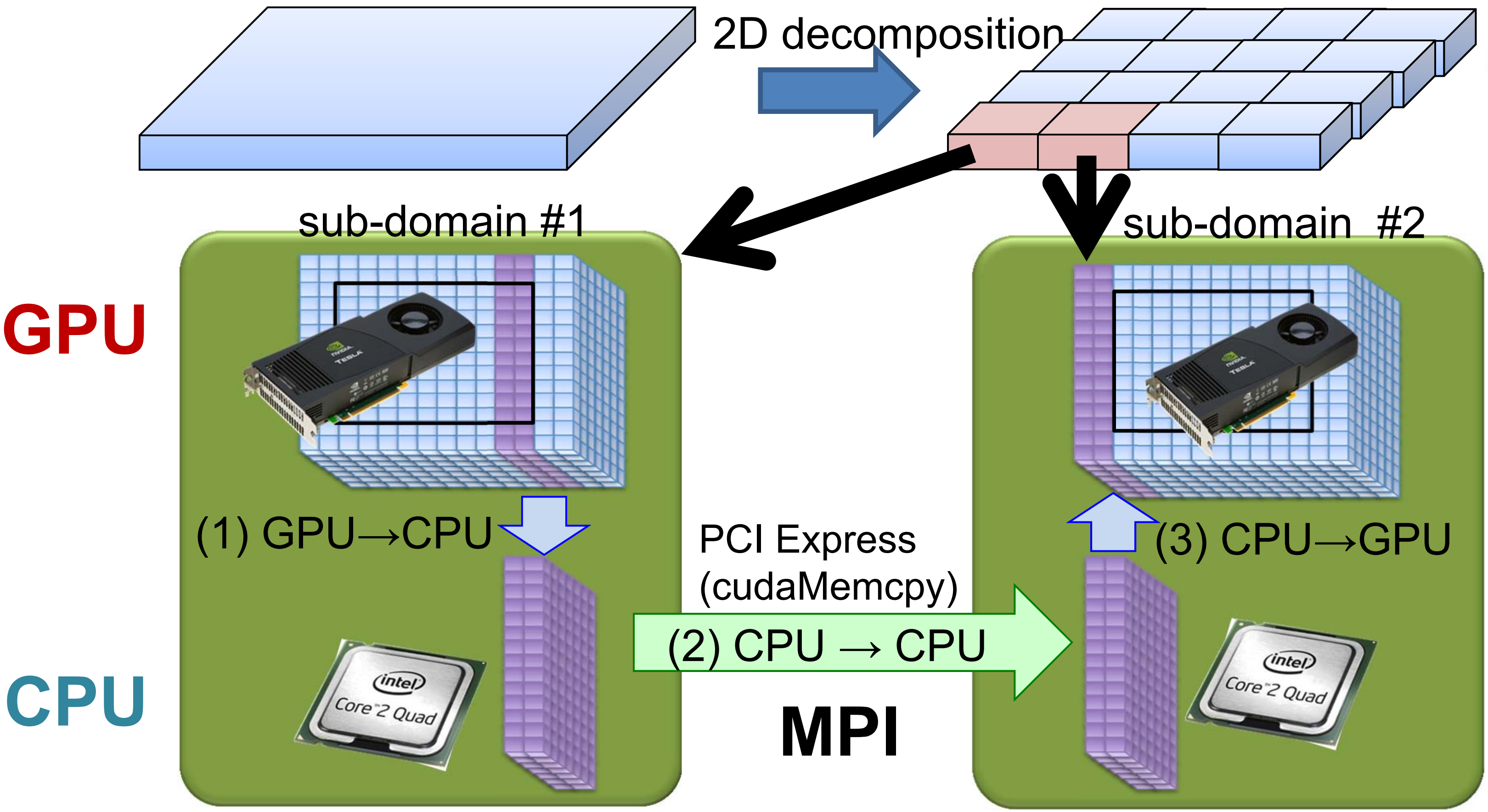
Performance of 5 kernels



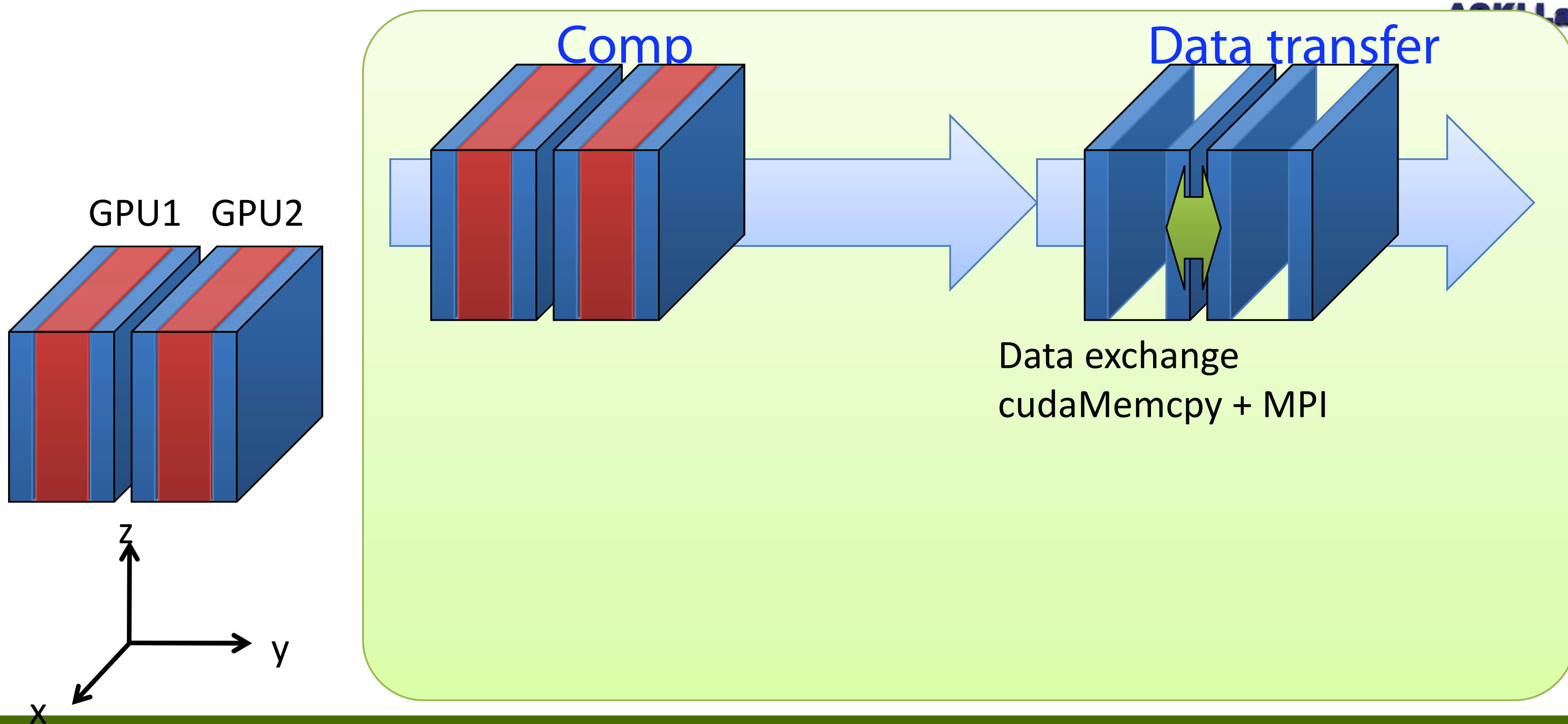
F_{peak} = 1030.0 GFlops

B_{peak} = 150.0 GByte/s

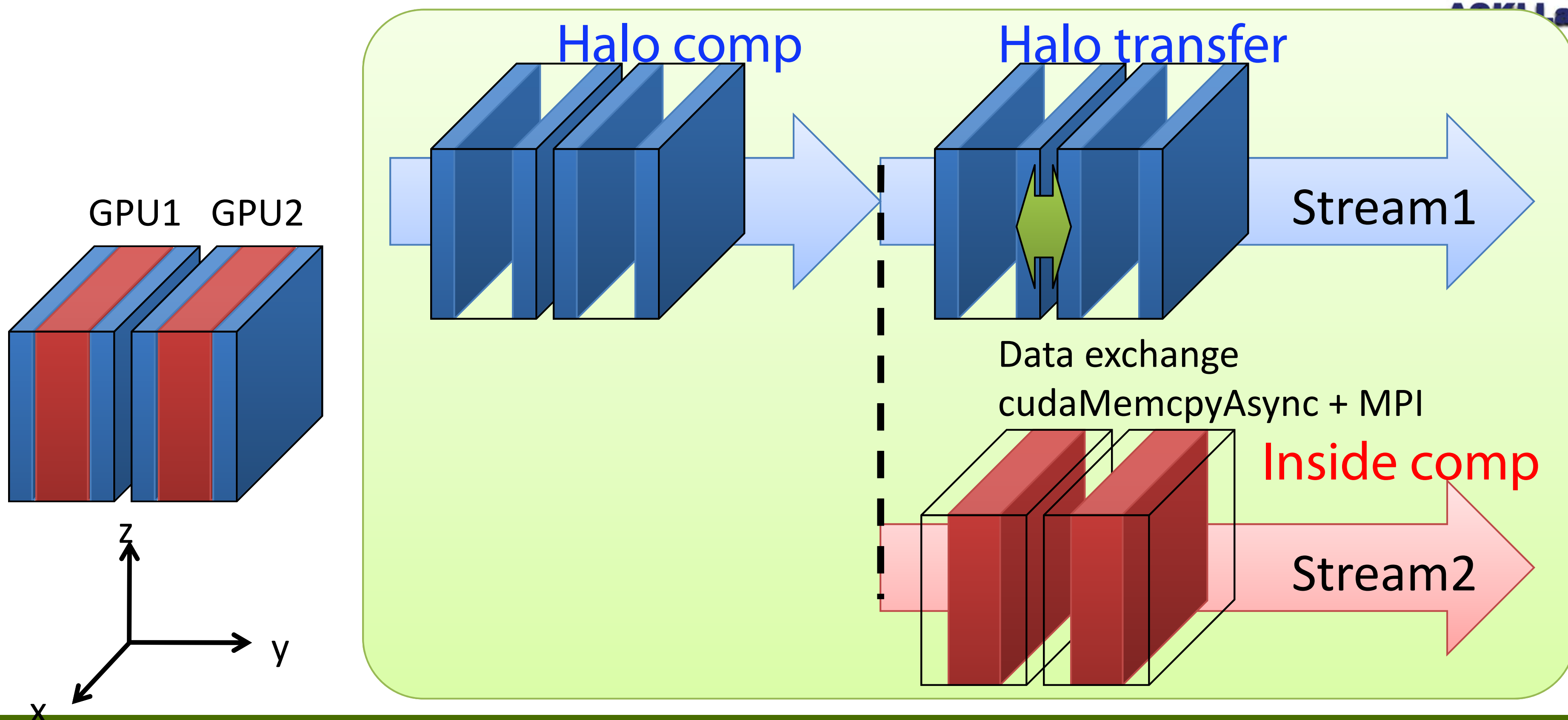
Multi-GPU : Domain decomposition



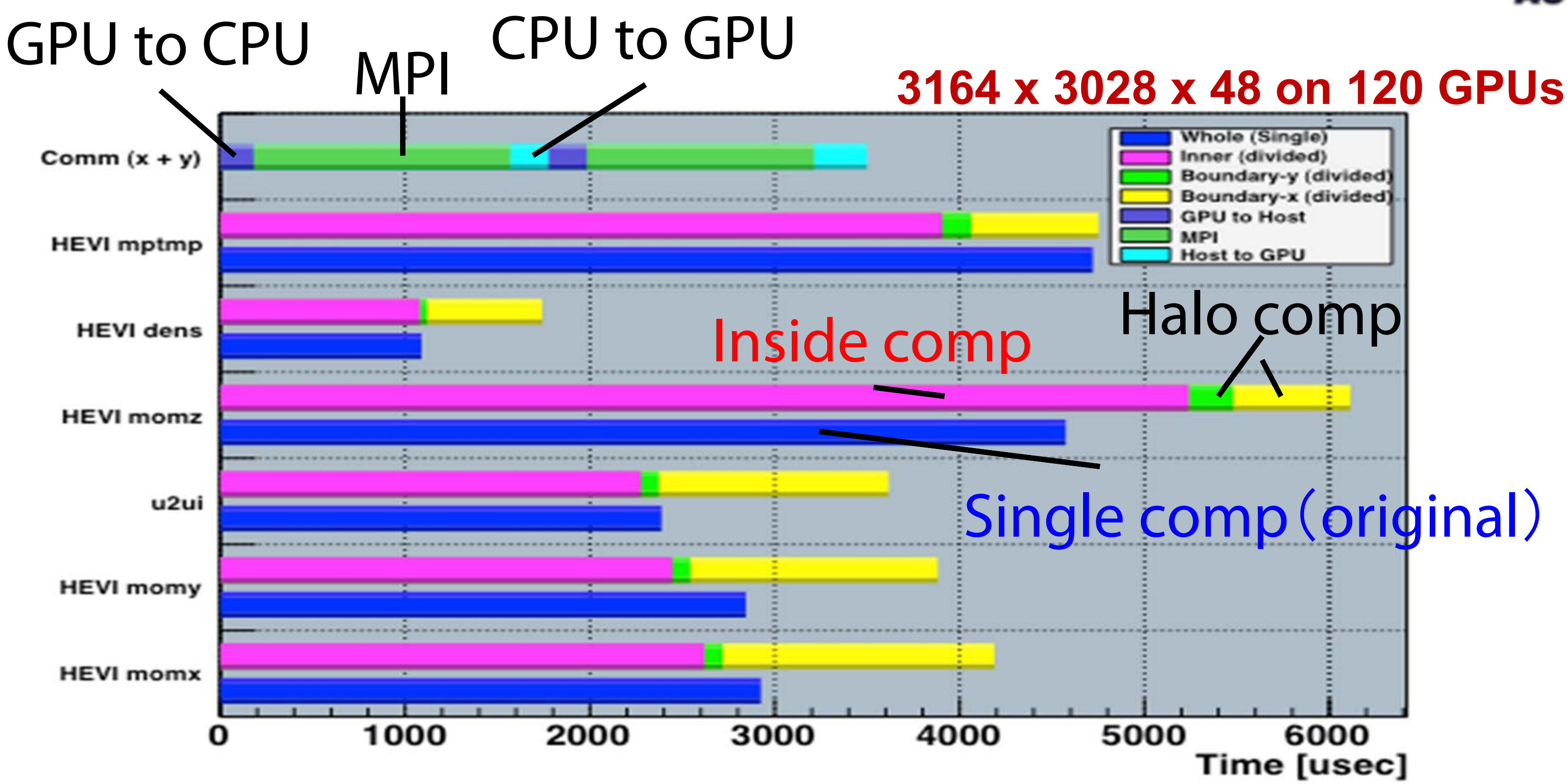
Computation and Communication



Overlapping communication with computation

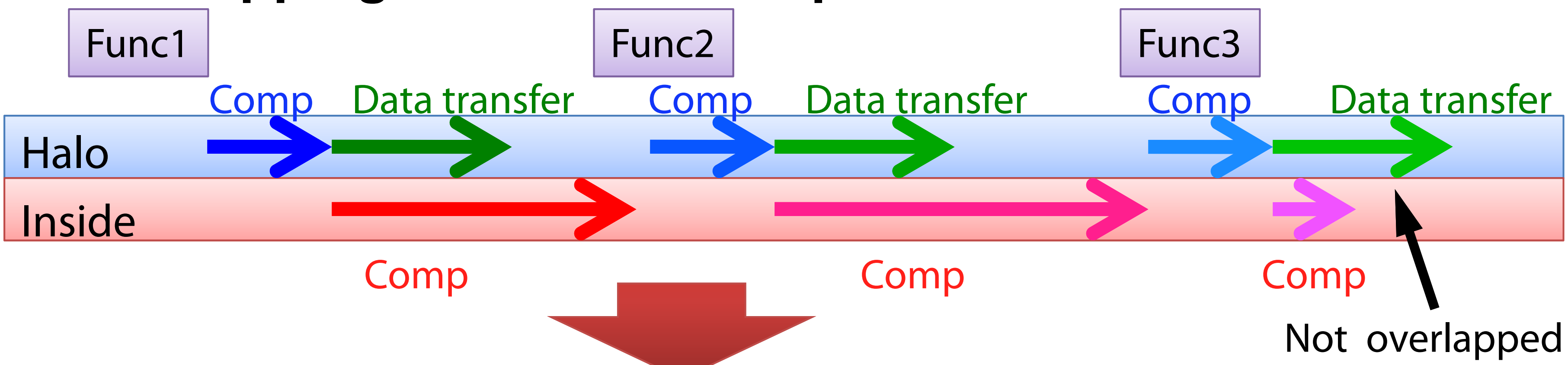


Breakdown of the GPU Kernel Functions in Time Loop

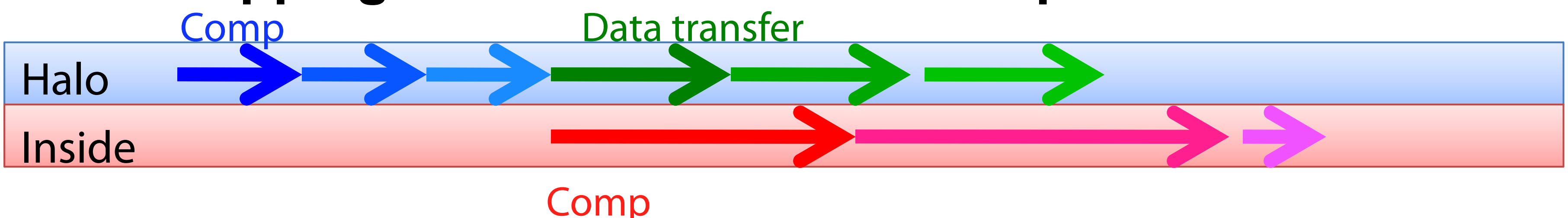


Re-ordering the communication and computation

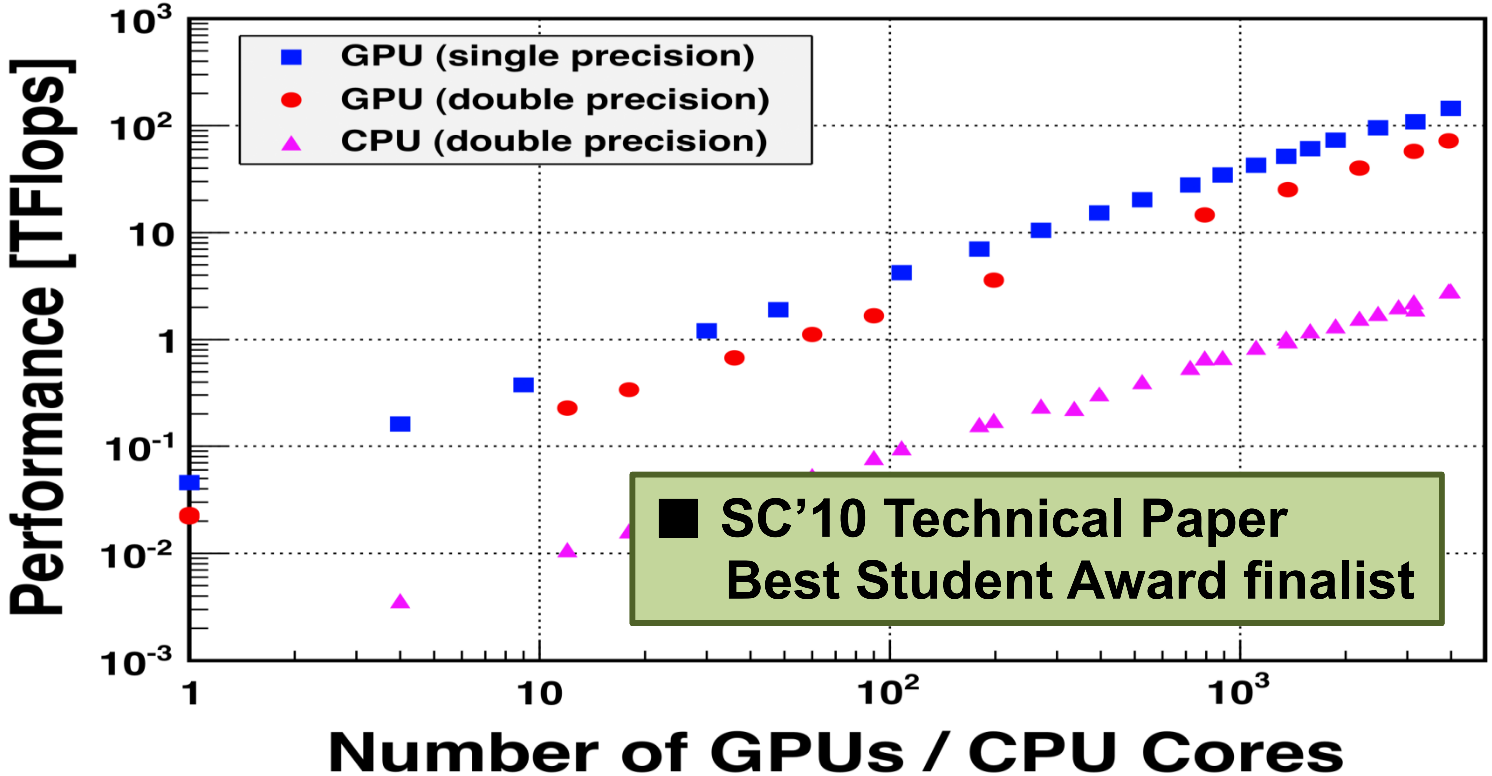
- **Overlapping comm and comp in each function**



- **Overlapping overall comms and comps in functions**



TSUBAME 2.0 Weak Scaling

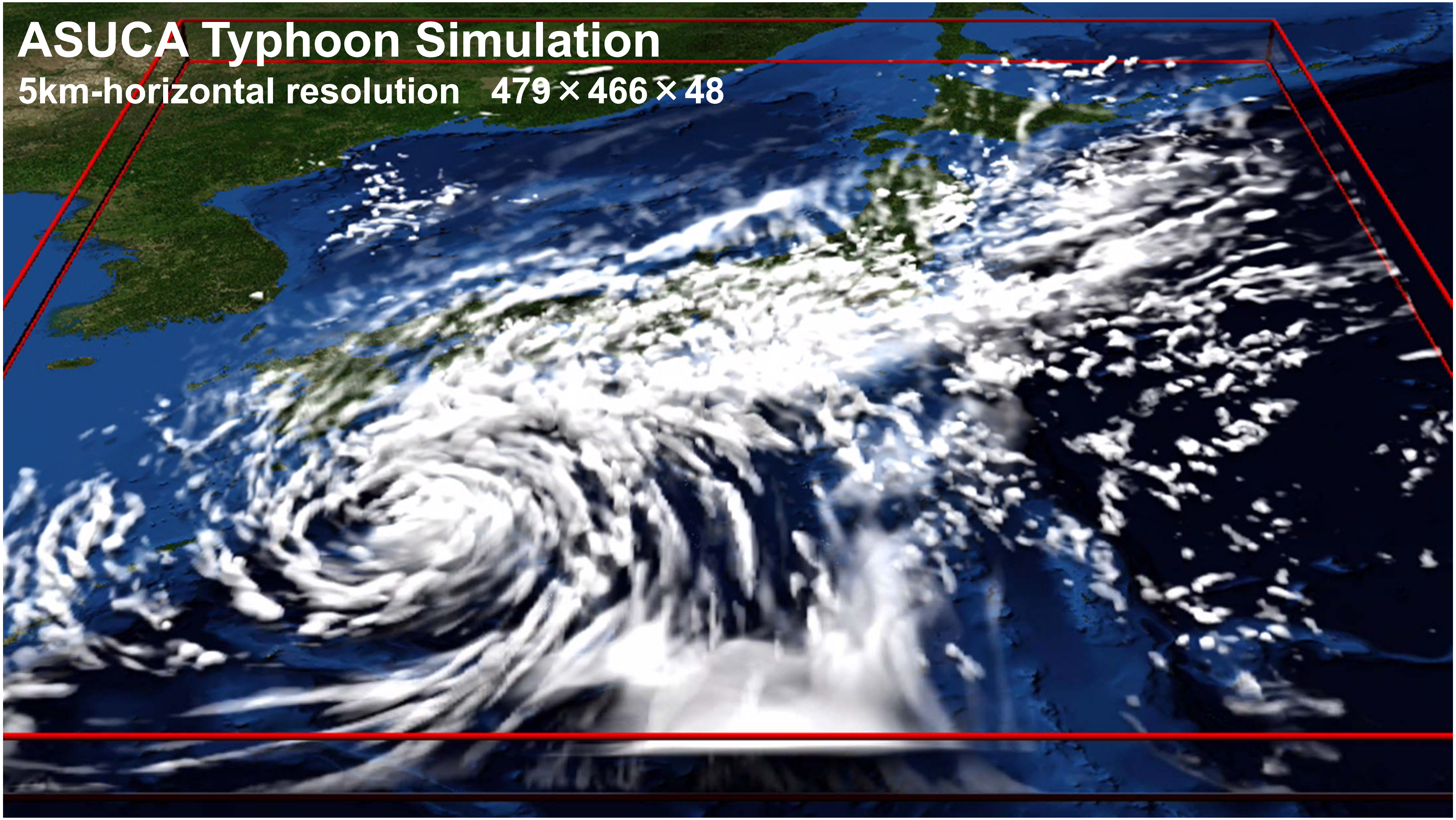


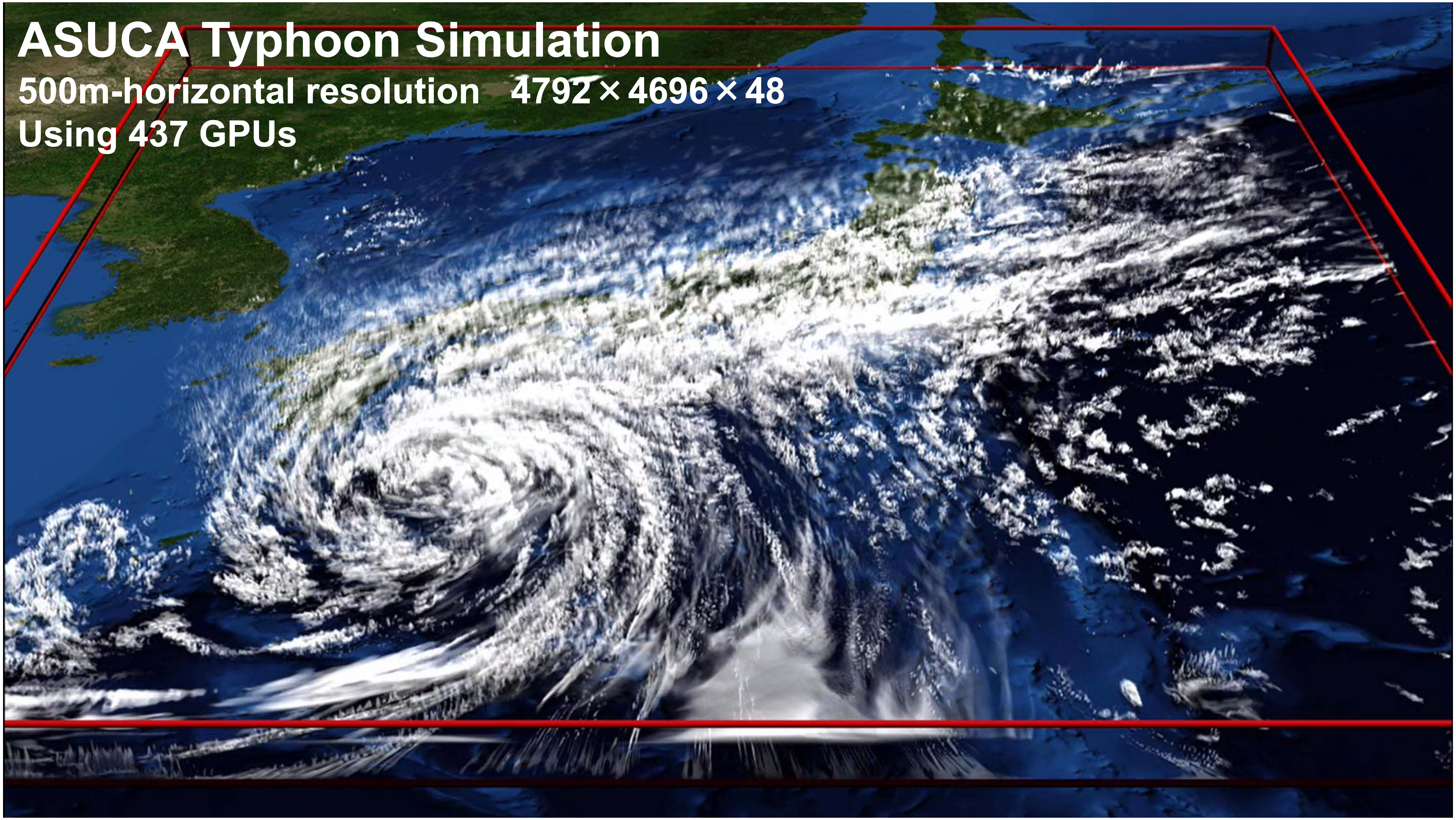
145.0 Tflops
Single precision

76.1 Tflops
Doublele precision

Fermi core Tesla M2050

3990 GPU

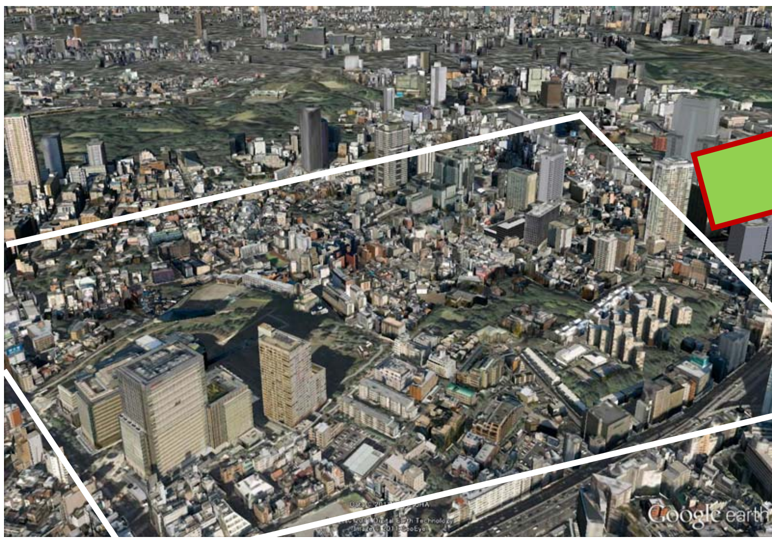

ASUCA Typhoon Simulation
5km-horizontal resolution 479×466×48

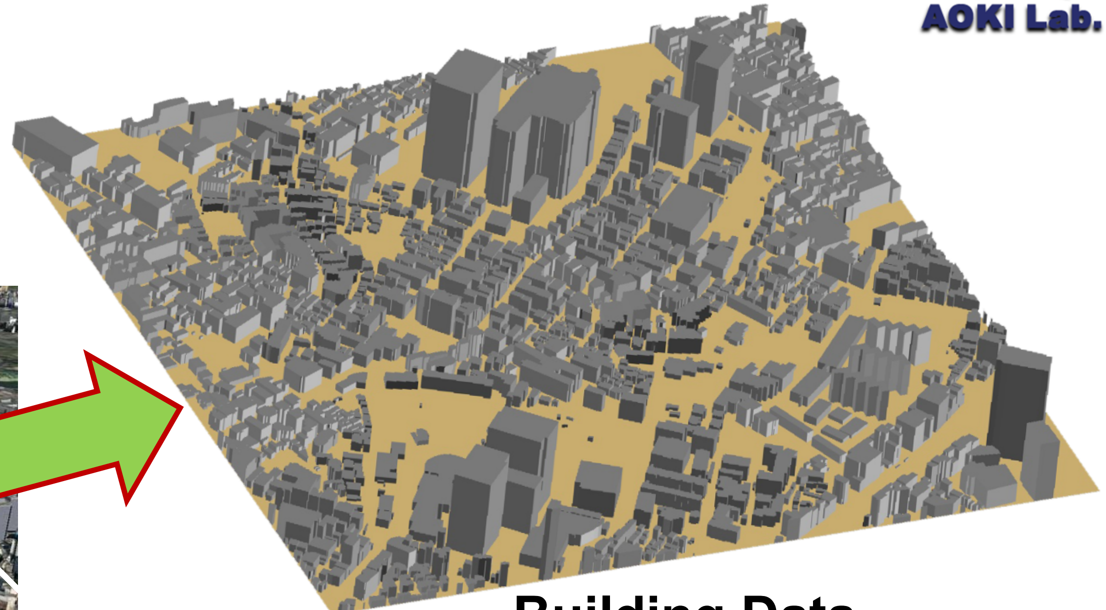

ASUCA Typhoon Simulation
500m-horizontal resolution 4792×4696×48
Using 437 GPUs

Real City Computation

Tokyo 六本木 Area

1km x 1 km





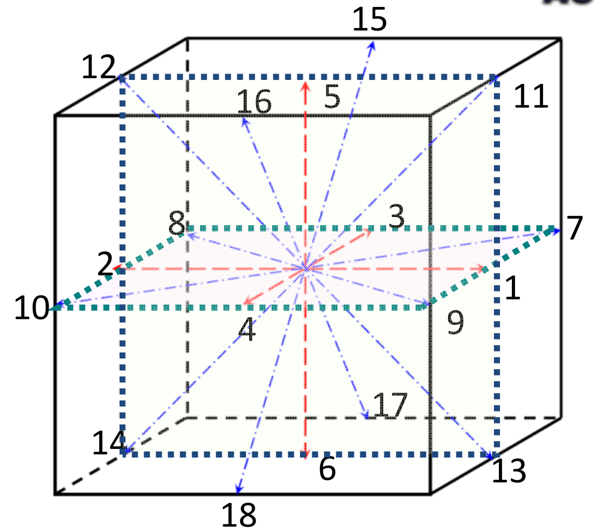
Building Data

1-m resolution 1000x1000x256

Lattice Boltzmann Method

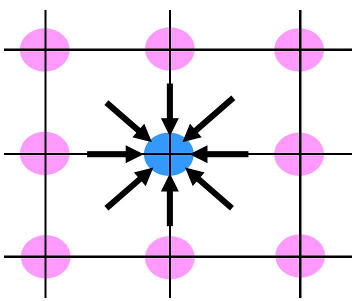
$$\frac{\partial f_i}{\partial t} + \mathbf{e}_i \cdot \nabla f_i = -\frac{1}{\lambda}\left(f_i - f_i^{eq}\right)$$

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

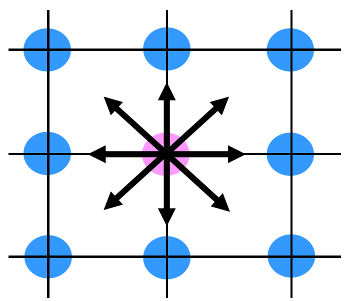


Strongly Memory Bound Problem:

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
$\mathbf{u}$ is the macroscopic velocity

LES for Turbulent Flows

Filtered Navier–Stokes equation

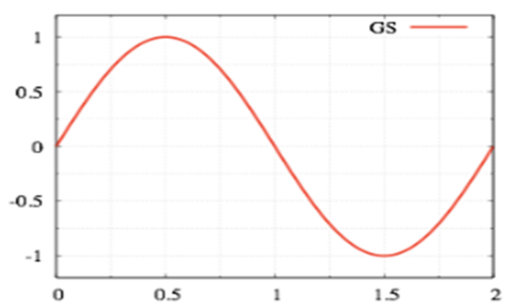
$$u = \underline{\bar{u}} + \underline{u'}$$

Energy dissipation

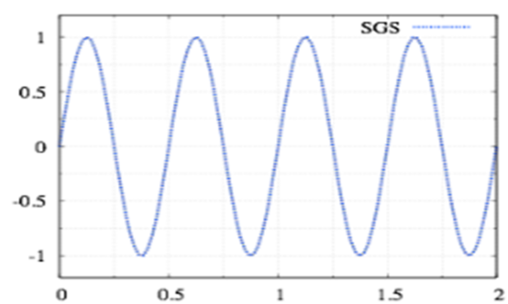
$$\frac{\partial \bar{u}_i}{\partial t} + \frac{\partial \bar{u}_i \bar{u}_j}{\partial x_j} = -\frac{\partial \bar{p}}{\partial x_i} + \frac{1}{Re}\frac{\partial^2 \bar{u}_i}{\partial x_j^2} - \boxed{\frac{\partial \tau_{ij}}{\partial x_j}} \qquad \frac{\partial \bar{u}_j}{\partial x_j} = 0$$

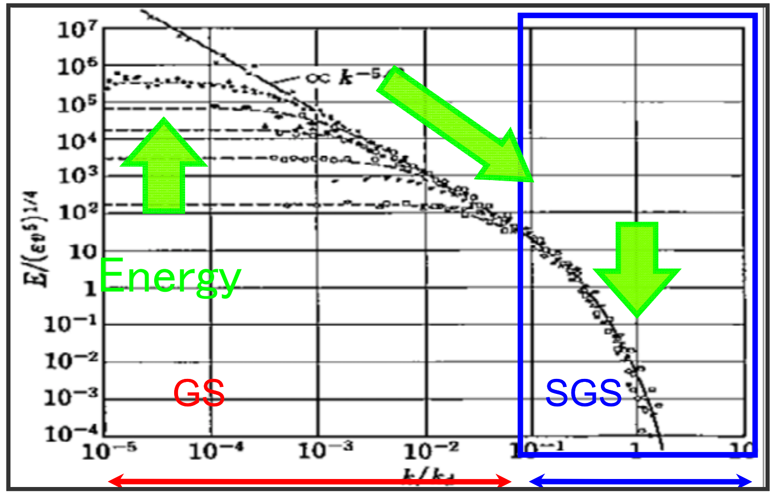
$$\boxed{\begin{aligned}\tau_{ij} &= \overline{u_i u_j} - \bar{u}_i \bar{u}_j \\ &= -2\nu_{SGS} S_{ij}\end{aligned}} \qquad S_{ij} = \frac{1}{2}\left(\frac{\partial u_j}{\partial x_i} + \frac{\partial u_i}{\partial x_j}\right)$$

GS



SGS





LES modeling

Smagorinsky model

$$\tau_{ij} = -2\nu_{SGS} S_{ij}$$

$$\nu_{SGS} = C\triangle^2|S| \qquad \boxed{C : const}$$

○ Simple

△ inaccurate for the flow with wall boundary

△ emperical tuning for the constant model coefficient

Dynamic Smagorinsky model

$$\nu_{SGS} = C\triangle^2|S|$$

$$\boxed{C = \frac{< L_{ij}L_{ij} >}{< M_{ij}M_{ij} >}} \qquad \begin{aligned} L_{ij} &= \widehat{\bar{u}_i\bar{u}_j} - \hat{\bar{u}}_i\hat{\bar{u}}_j \\ M_{ij} &= 2\bar{\triangle}^2\widehat{|\bar{S}|\bar{S}_{ij}} - 2\hat{\bar{\triangle}}^2|\hat{\bar{S}}|\hat{\bar{S}}_{ij} \end{aligned}$$

○ applicable to wall boundary

△ complicated calculation

△ average process over the wide area

→ not available for complex shaped body

→ not suitable for large-scale problem

Coherent-Structure Smagorinsky model

*H.Kobayashi, Phys. Fluids.17, (2005).

$$\nu_{SGS} = \underline{C}\triangle^2|S|$$

→ model coefficient determined by the second invariant of the velocity gradient tensor

$$\boxed{C = C_1|F_{CS}|^{3/2}}$$

$$\boxed{F_{CS} = \frac{Q}{E}} \qquad Q = -\frac{1}{2}\frac{\partial \bar{u}_j}{\partial x_i}\frac{\partial \bar{u}_i}{\partial x_j} \qquad E = -\frac{1}{2}\left(\frac{\partial \bar{u}_j}{\partial x_i}\right)^2$$

$$(-1 < F_{CS} < 1)$$

△ model coefficient

○ applicable to wall boundary

○ model coefficient is locally determined.

LES modeling on LBM

AOKI Lab.

Turbulence model :

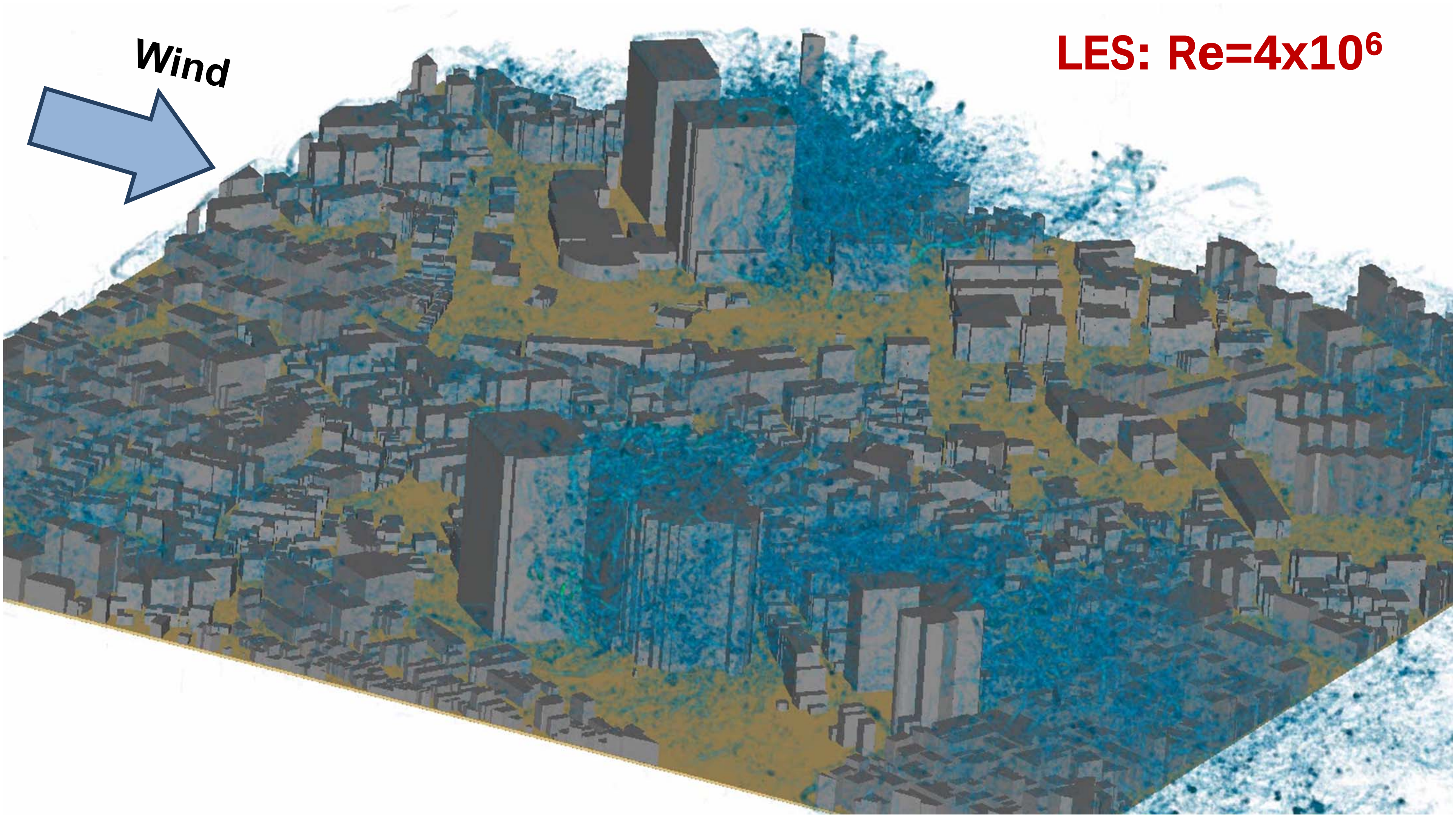
$$\nu_* = \nu_0 + \nu_t = \frac{1}{3}\left(\tau_* - \frac{1}{2}\right)c^2\delta_t = \frac{1}{3}\left(\tau_0 + \tau_t - \frac{1}{2}\right)c^2\delta_t, \quad \nu_t := \frac{1}{3}\tau_t c^2 \delta_t,$$

Molecular viscosity + eddy viscosity

$$\nu_t = (C_S \Delta_x)^2 \bar{S}$$

Smagorinsky model subgrid closure $C_S = 0.22$

$$\bar{S}_{ij} = \frac{1}{2}(\partial_j \bar{u}_i + \partial_i \bar{u}_j) \quad \bar{S} = \sqrt{2 \sum_{i,j} \bar{S}_{ij} \bar{S}_{ij}}$$


Wind
LES: Re=4x10^6

SUMMARY

- ASUCA (JMA operational weather prediction model) has been successfully implemented on GPU supercomputer TSUBAME 2.0.
- 500m-resolution model runs on 437 GPU, i.e, only 10 % of TSUBAME 2.0.
- A showcase of Large-scale GPU applications
  - ･Lattice Boltzmann Method
  - ･LES with CSM SGS model

Thank you

for your kind attention