GPU スパコンにおける大規模 HPC アプリケーションのスケーラビリティ

— 格子系流体計算 —

Takayuki Aoki

Global Scientific Information and Computing Center
Tokyo Institute of Technology

GPGPU

General-Purpose Graphics Processing Unit

■ **Major differences from Previous Accelerators**

ClearSpeed, Grape, , ,

High Memory Bandwidth
suitable for wide variety of applications

Consumer Product
inexpensive

Software Development Environment

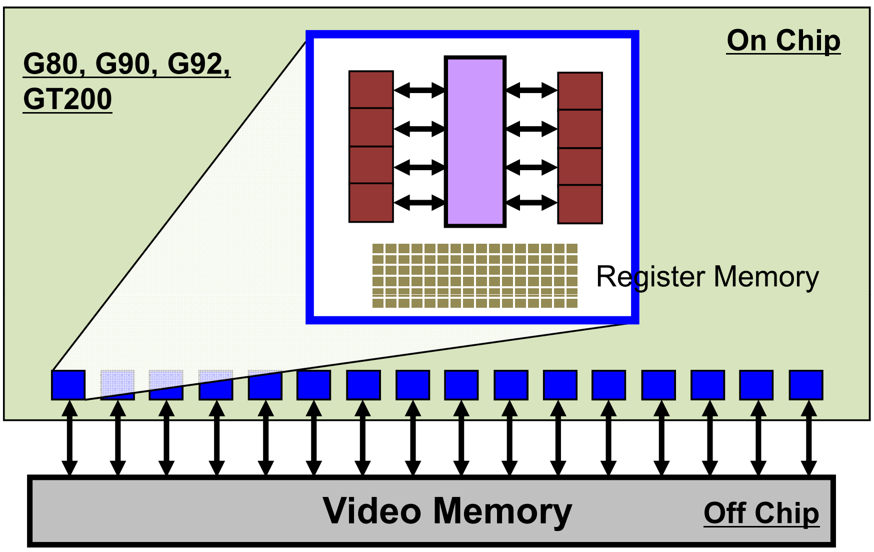
CUDA, Open CL

NVIDIA GPU

		Intel Core i7 Extreme	GeForce GTX 285	GeForce GTX 480 Fermi
GPU	Peak Performance [GFlops]	**102.4**	**708***,**1063**	**1300**
	Number of Processor	4	240	480
	Core Clock [MHz]	3200	1476	1400
Memory	Bandwidth[GB/s]	**32**	**159**	**177**
	Memory Interface [bit]	64	512	384
	Memory Data Clock [MHz]	1333 (DDR3)	2484 (GDDR3)	(GDDR5)
	Capacity [GB]	-----	2046	1500
B/F	Bandwidth/Performance	**0.312**	**0.149**	**0.136**

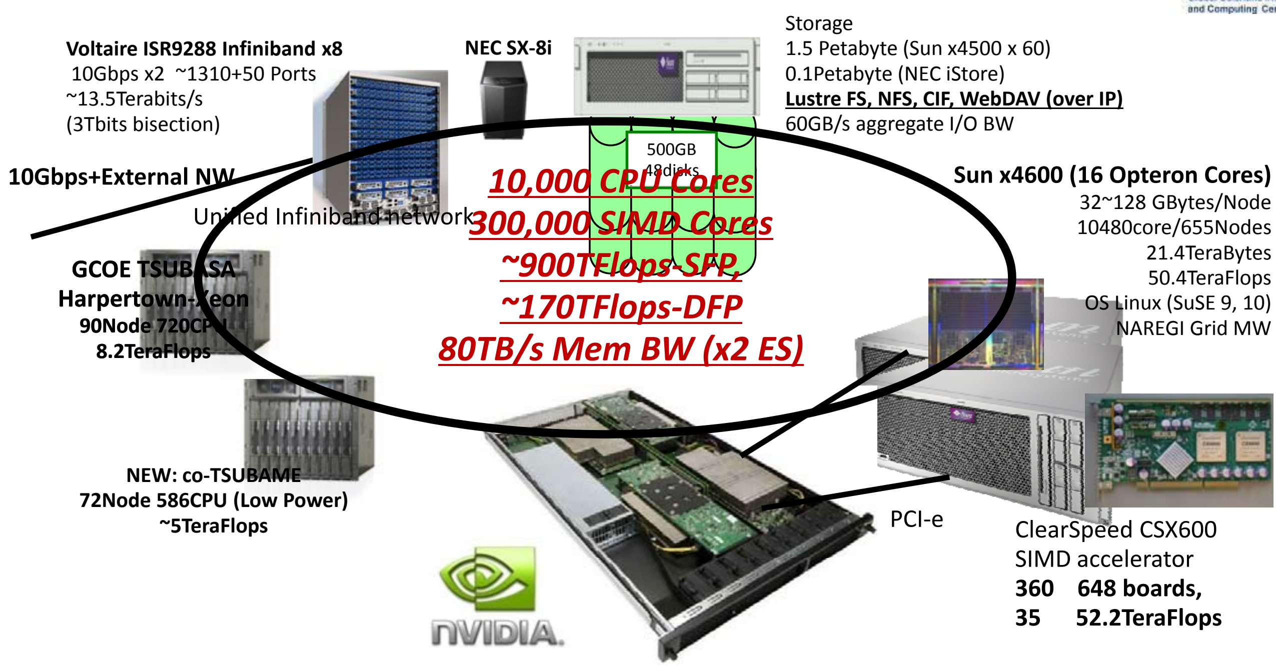
Peak Power : 183W

NVIDIA GPU Architecture



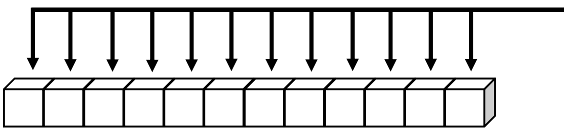
Global memory	~4GB (VRAM
Streaming Multiprocessor	~30 (GTX280(GT200))
Shared memory	16 Kbyte
Streaming Processor	8 per SM, total 240

Tokyo Tech TSUBAME 1.2

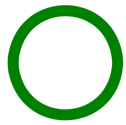


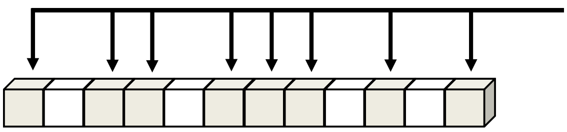
170-Nvidia Tesla S1070: total 680 GPU cards
High Performance in Many BW-Intensive Apps
10% power increase over TSUBAME 1.0 (130TF SFP / 80TF DFP)

Memory Access

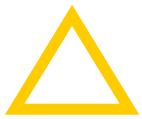


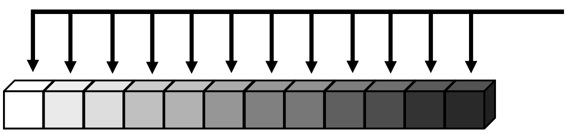
Continuous Access
FDM (Finite Difference)
Coalescing



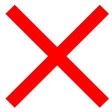


Random (Indirect) Access
FEM (Finite Element)
A[i] = B[IP[i]] + B[IP[i-1]];





Data Dependency
A[i] = A[i-1] + A[i-2]*C;

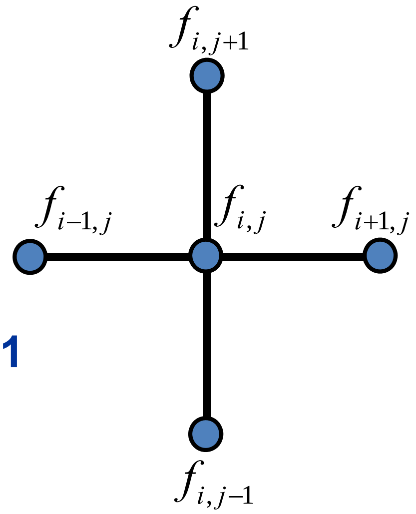


Compute Intensive or Memory Bound ?

■ **2-dimensional diffusion Equation**

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

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



FLOP = 9

Byte = 4*6 = 24 byte : read 5, write 1

FLOP/Byte = 9/24 = 0.375

Arithmetic INTENSITY: FLOP/Byte

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

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

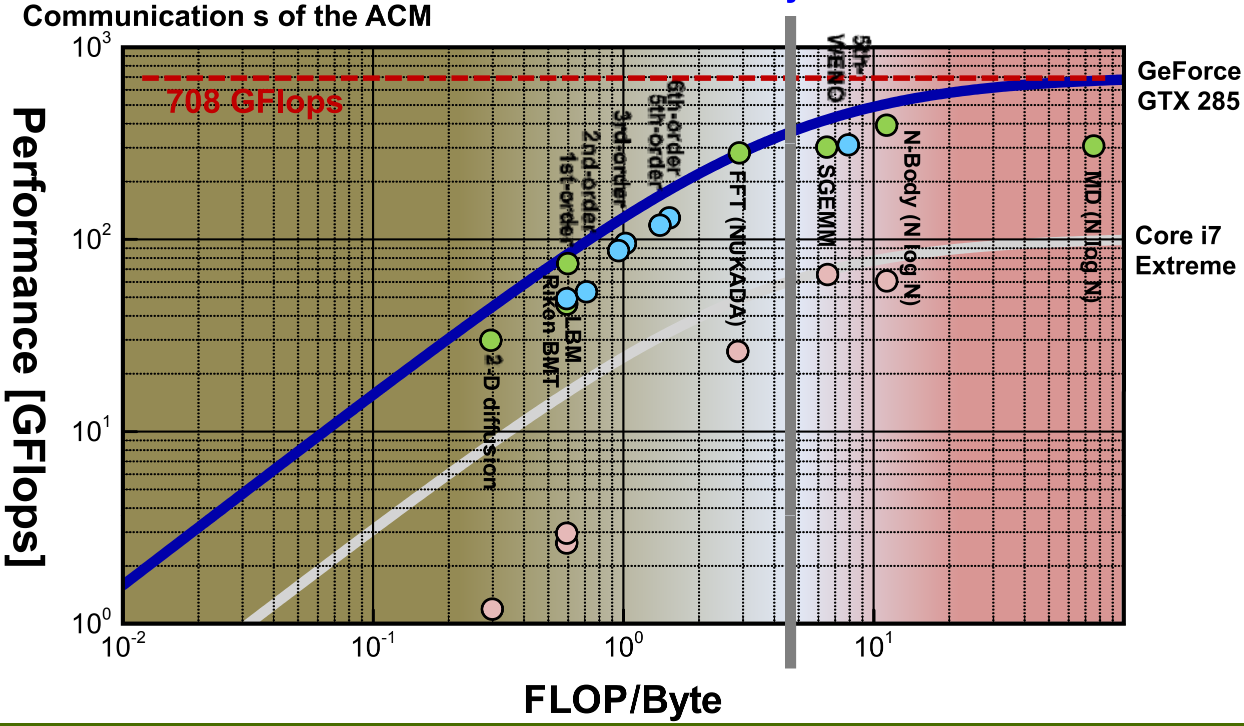
$$= \frac{\text{FLOP/Byte}}{\text{FLOP/Byte} + \text{F/B} + \alpha} \text{F}$$

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

Application Performances

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

FLOP/Byte = F/B



Classification of CFD

Compressible fluid analysis

Supersonic flow, Acoustic wave, Explosion, Shock wave, . . .

High-accurate numerical methods:

T. Aoki, Comp. Phys. Comm., Vol.102, No.1-3, 132-146 (1997)
Y. Imai, T. Aoki and K. Takizawa, J. Comp. Phys., Vol. 227, Issue 4, 2263-2285 (2008)
K. Kato, T. Aoki, M. Yoshida, et. al., Int. J. Numerical Methods in Fluids, Vol.51, 1335-1353 (2006)
Y. Imai, T. Aoki, J. Comp. Phys., Vol.217, 453-472 (2006)
Y. Imai, T. Aoki, J. Comp. Phys., Vol.215, 81-97 (2006)

Incompressible fluid analysis

- Most of flow phenomena in our daily life,
- Turbulent flow, Multi-phase flow, Reacting flow, . .
- Semi-implicit Time Integration → Poisson Solver

Rayleigh-Taylor Instability

GP GPU

Y. Imai, T. Aoki and K. Takizawa, J. Comp. Phys., Vol. 227, Issue 4, 2263-2285 (2008)

IDO-CF Scheme

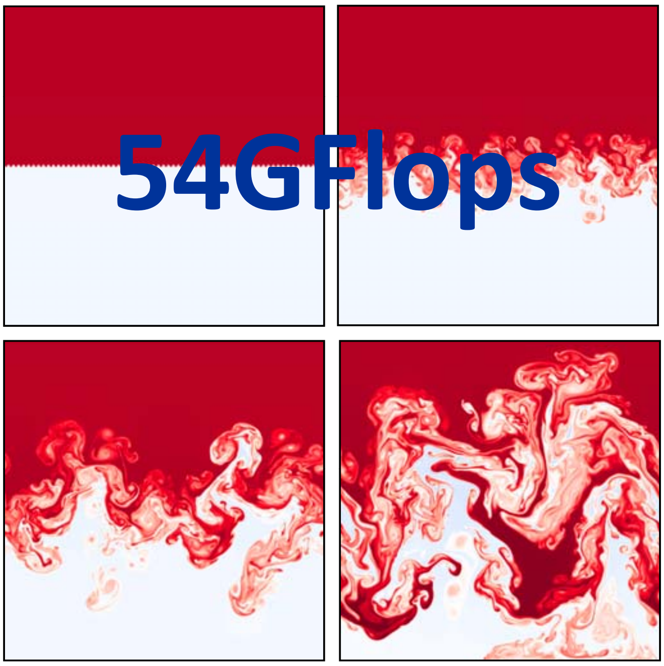
Euler equation:

$$\frac{\partial \mathbf{Q}}{\partial t} + \frac{\partial \mathbf{E}}{\partial x} + \frac{\partial \mathbf{F}}{\partial y} = 0$$

$$\boldsymbol{Q} = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ e \end{bmatrix} \quad \boldsymbol{E} = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ eu + pu \end{bmatrix} \quad \boldsymbol{F} = \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ ev + pv \end{bmatrix}$$

Heavy fluid lays on light fluid and unstable.

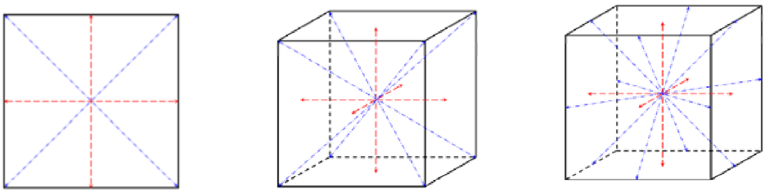
512 x 512



Lattice Boltzmann Method

GP GPU

Discretizing Boltzmann-BGK equation in the momentum space using a finite set of velocities $\{e_i\}$



D2Q9 D3Q15 D3Q19

D2Q9
D3Q15
D3Q19...

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

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

$\boldsymbol{i}$ 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

Lattice Boltzmann Method

GP GPU

Further discretization in physical space x, and time t

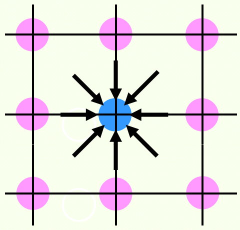
$$f_i(\mathbf{x}+\mathbf{e}_i\Delta t, t+\Delta t) - f_i(\mathbf{x},t) = -\frac{1}{\tau}\left(f_i(\mathbf{x},t) - f_i^{eq}(\mathbf{x},t)\right)$$

$$\tau = \lambda / \delta t$$

LBM code

Collision step:

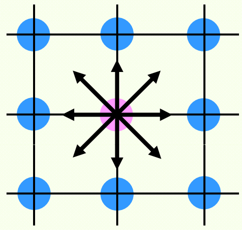
$$\bar{f}_i(\mathbf{x},t) = f_i(\mathbf{x},t) - \frac{1}{\tau}\left(f_i(\mathbf{x},t) - f_i^{eq}(\mathbf{x},t)\right)$$



Purely local !!

Streaming step:

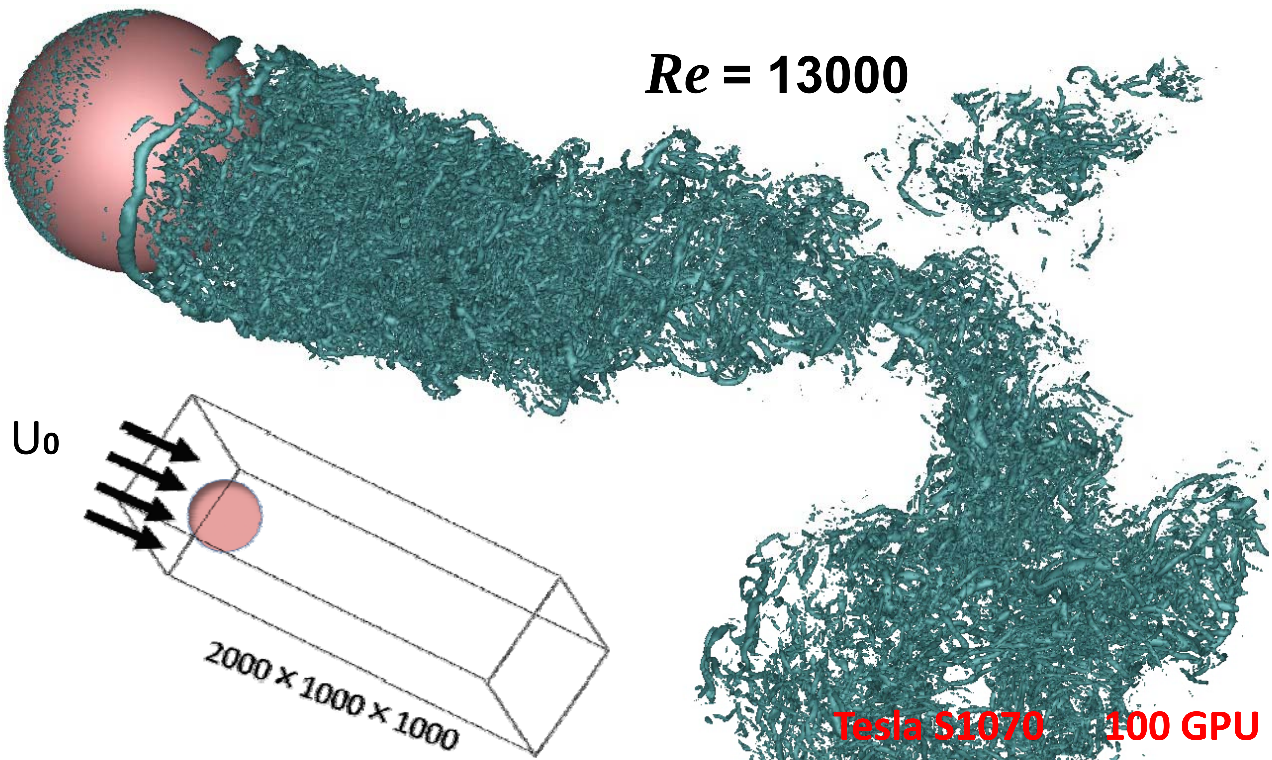
$$f_i(\mathbf{x}+\mathbf{e}_i\Delta t, t+\Delta t) = \bar{f}_i(\mathbf{x},t)$$



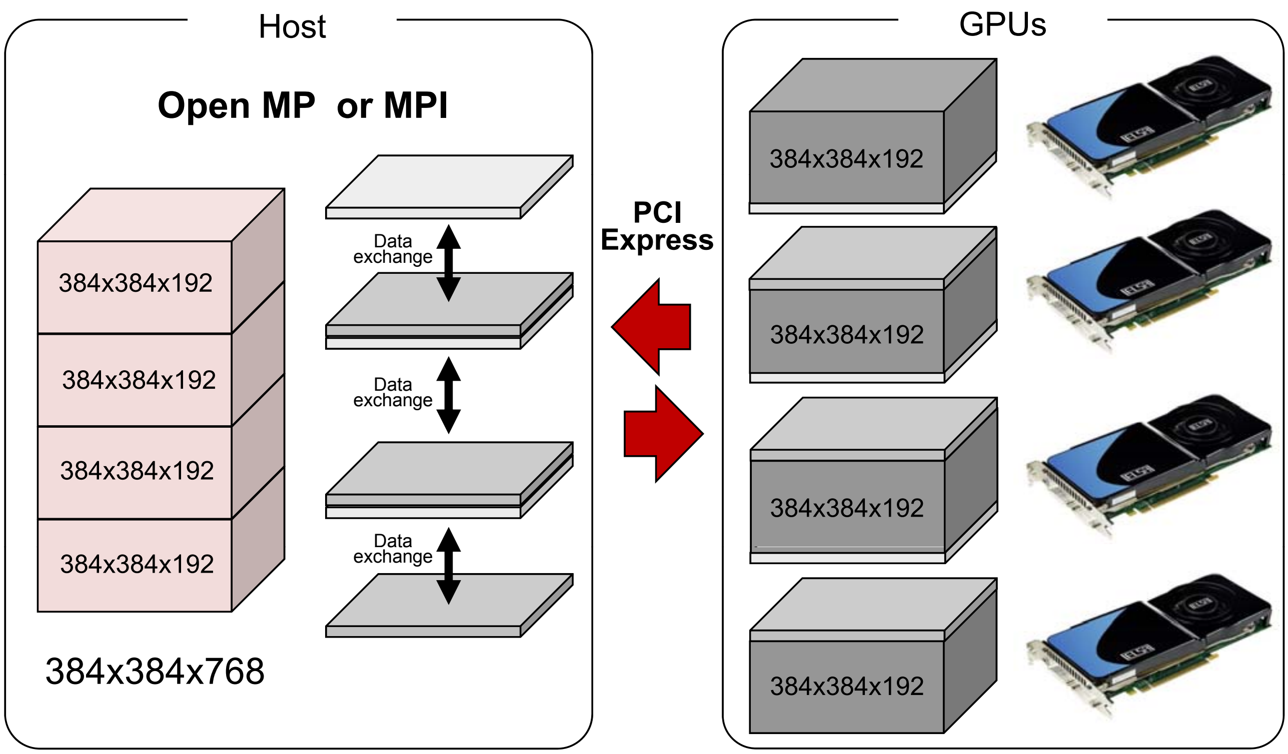
Uniform data shifting
Little computational effort !!

Incompressible Turbulent Flow
behind a spherical body

GP GPU

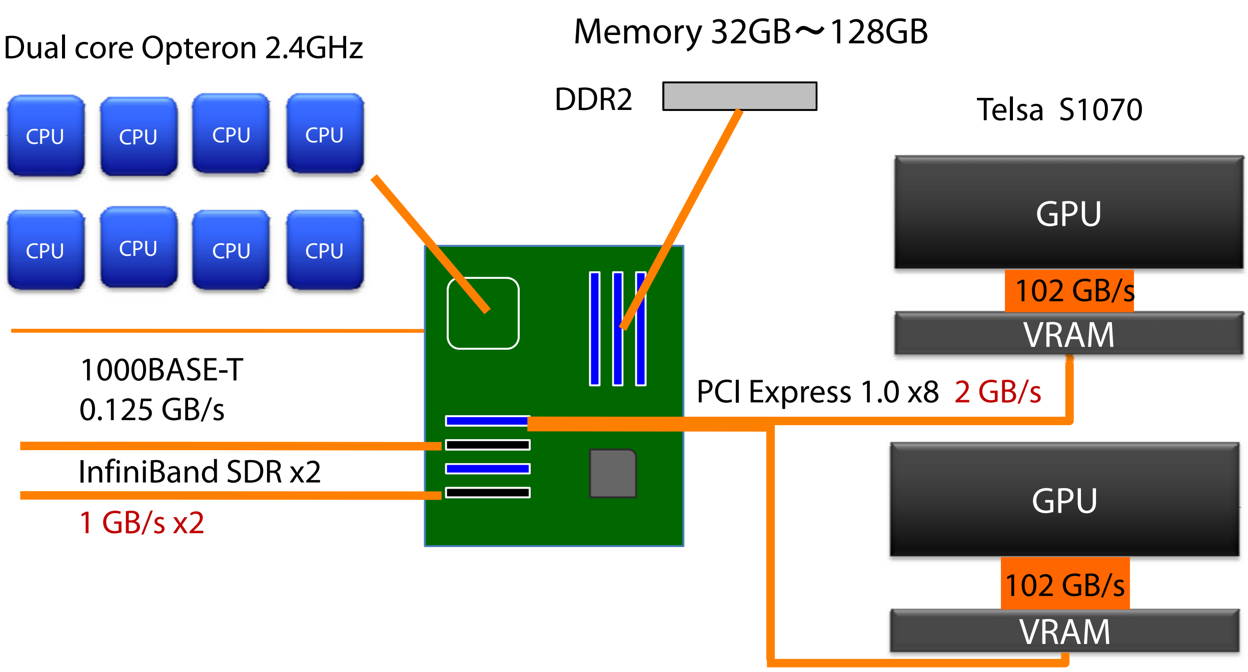


1-D domain decomposition and GPU-to-GPU communication

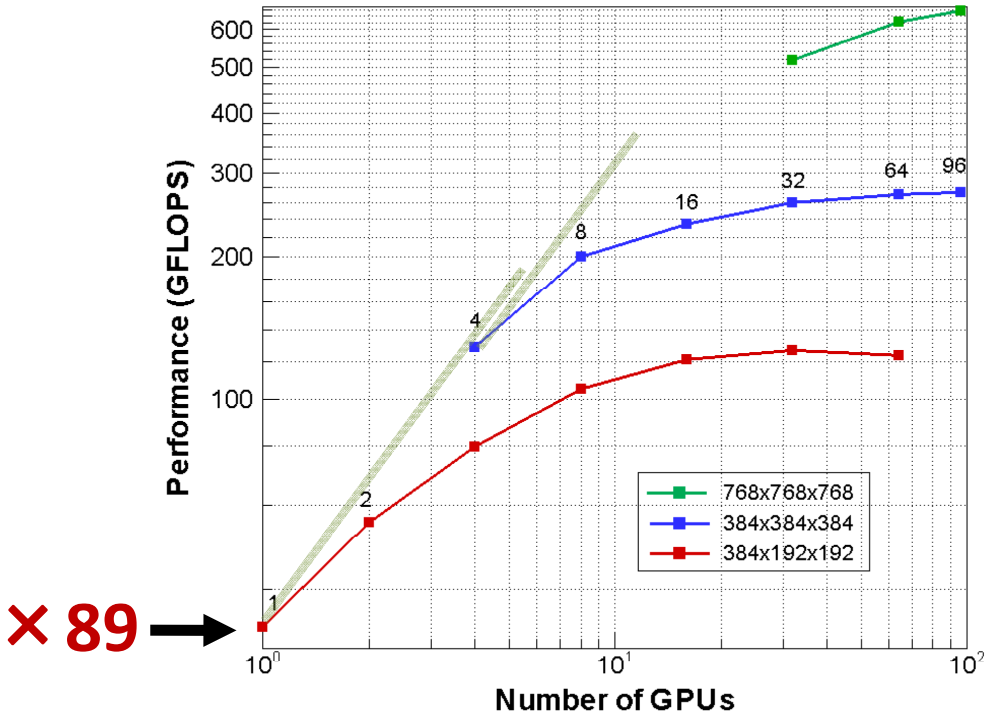


TSUBAME 1.2 node detail

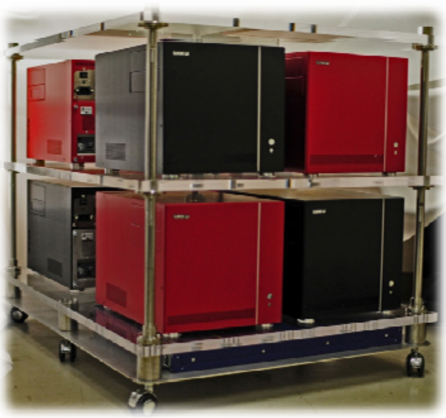
SunFire X4600

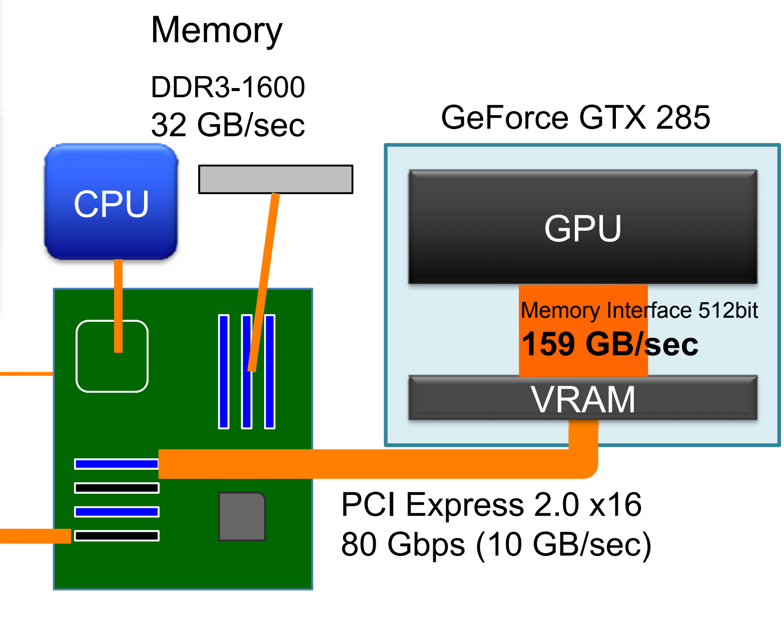


Multi-GPU Scalability



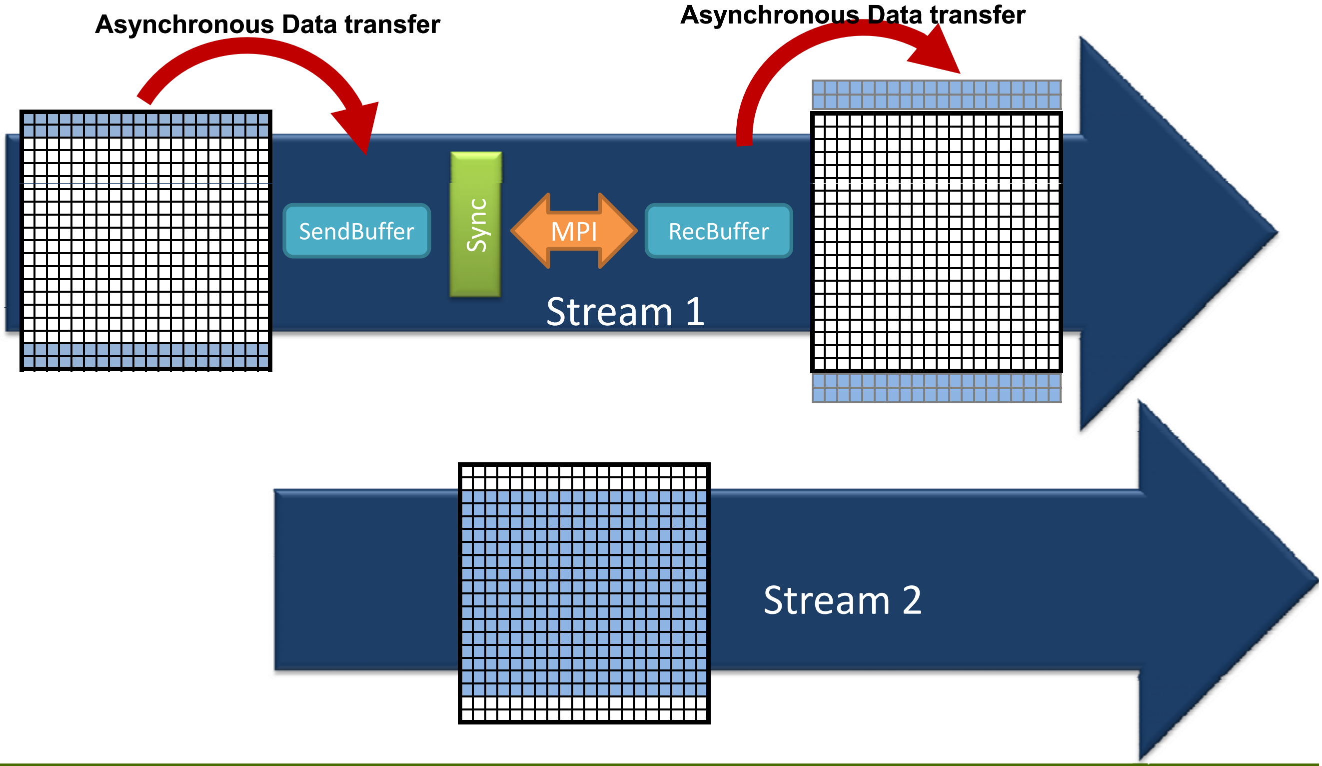
Lab. Machine (RED & BLACK)





Overlapping between Computation and Communication

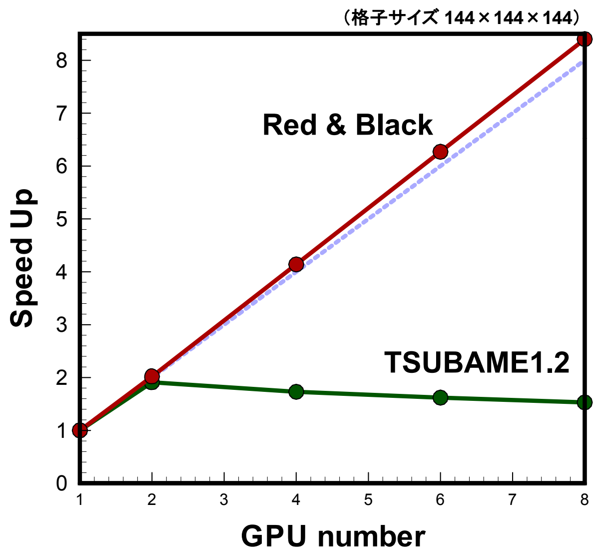
GP GPU



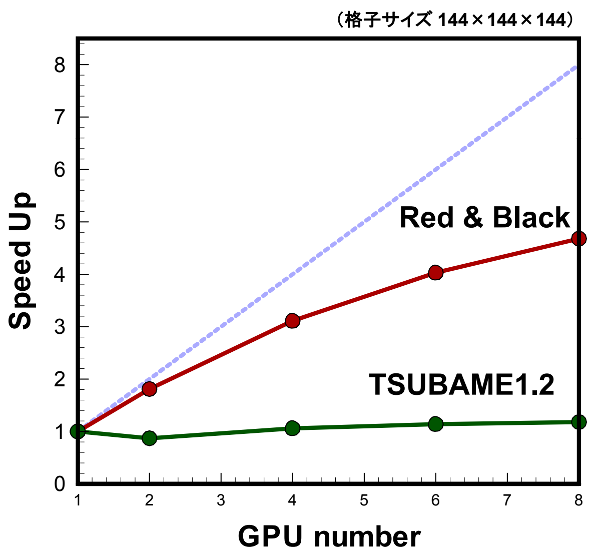
TSUBAME 1.2との性能比較

GP GPU

通信と計算のオーバーラップ法
の導入



通信と計算のオーバーラップ法
の導入なし



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

Atmosphere Model

GP GPU

Dynamic Process:

Full 3-D Navior-Stokes Equation

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

Physics Process:

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

So called "Parameterization" including many empirical rules.

WRF GPU Computing

GP GPU

WRF (Weather Research and Forecast)

Community Code developed by NCAR, NCEP, OU, NOAA/FSL, AFWA

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 scales

⇒ x 8.5 increase

•Michalakes, J. 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.*

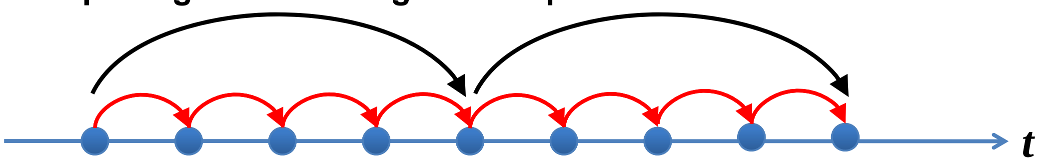
WRF versus ASUCA

GP GPU

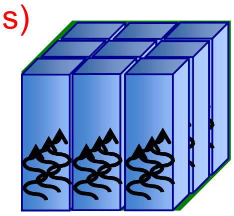
ASUCA : developed by Japan Meteorological Agency (JMA) for the next-generation weather prediction

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)



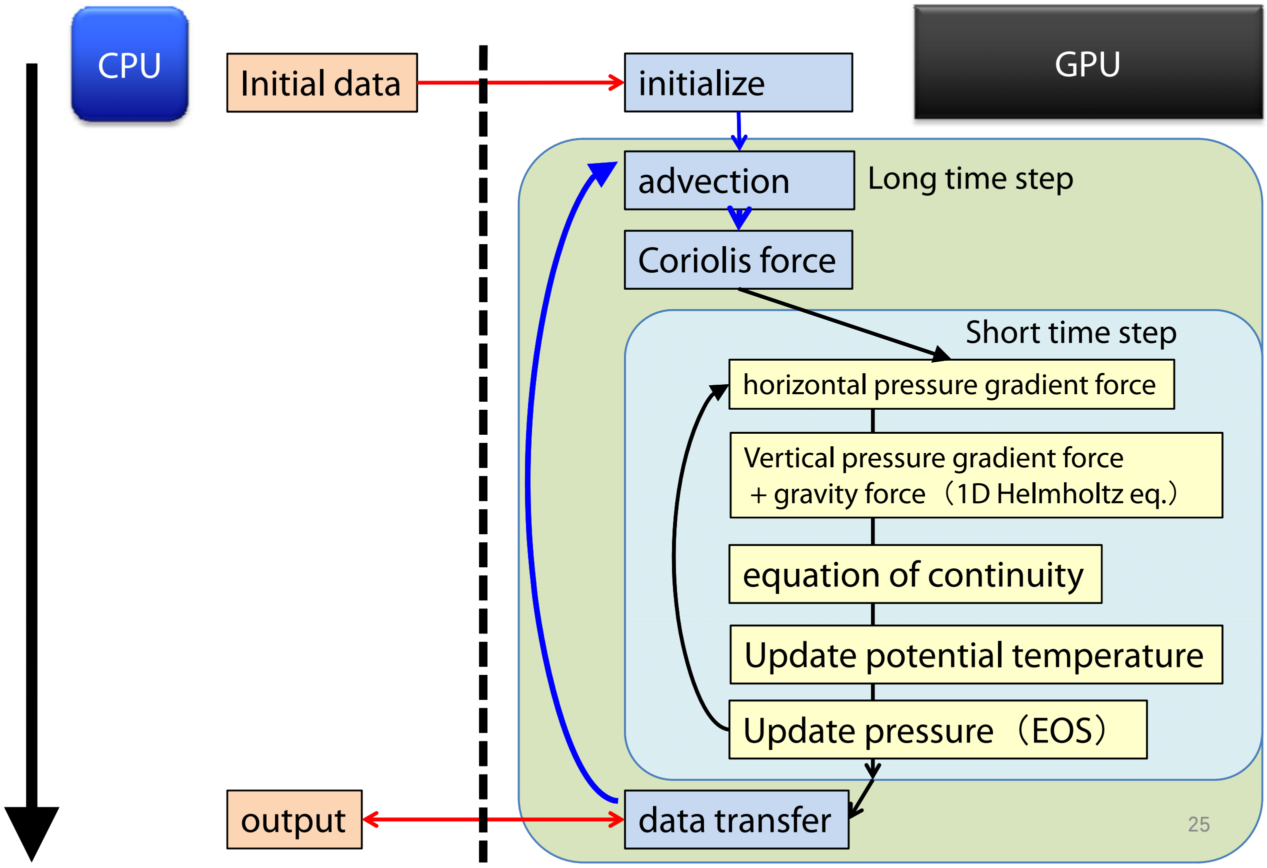
HEVI (Horizontally explicit Vertical implicit) scheme

Horizontal resolution ~ 1 km

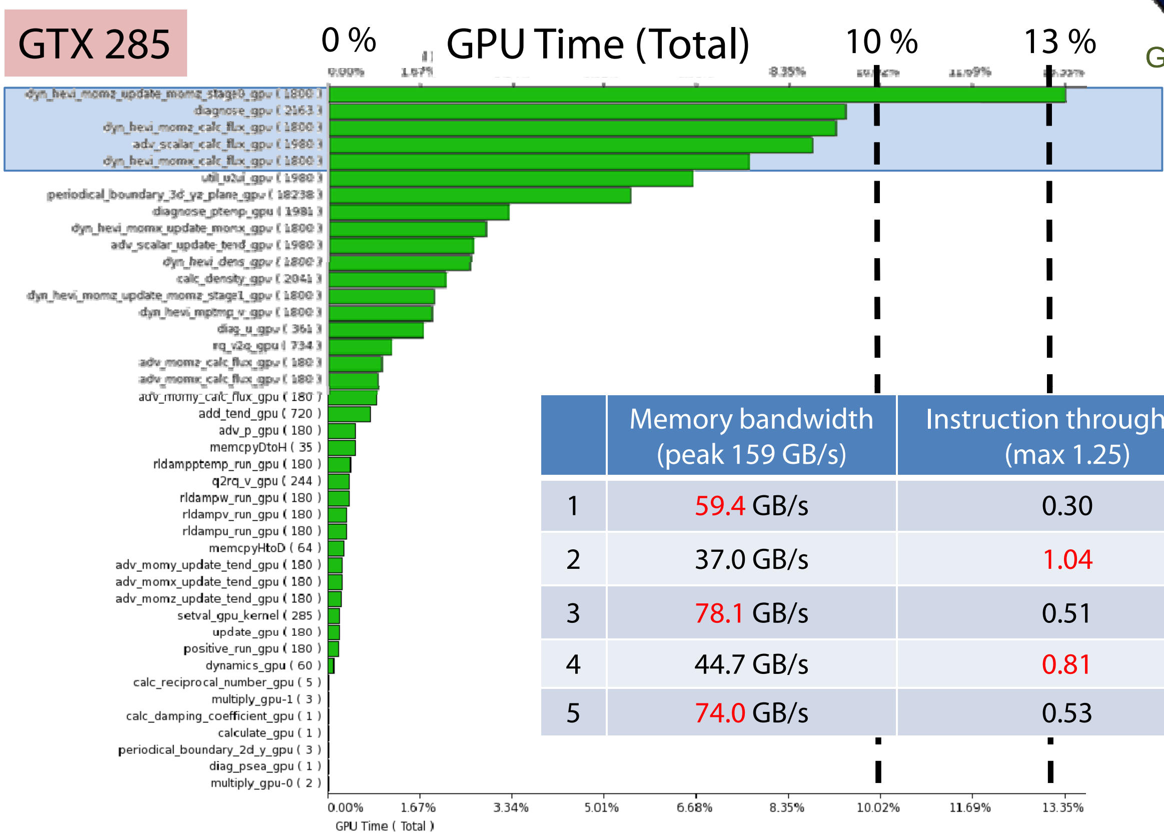
Vertical resolution ~ 100 m

1-D Helmholtz equation (like Poisson eq.) ⇨ sequential process

ASUCAにおける計算の流れ



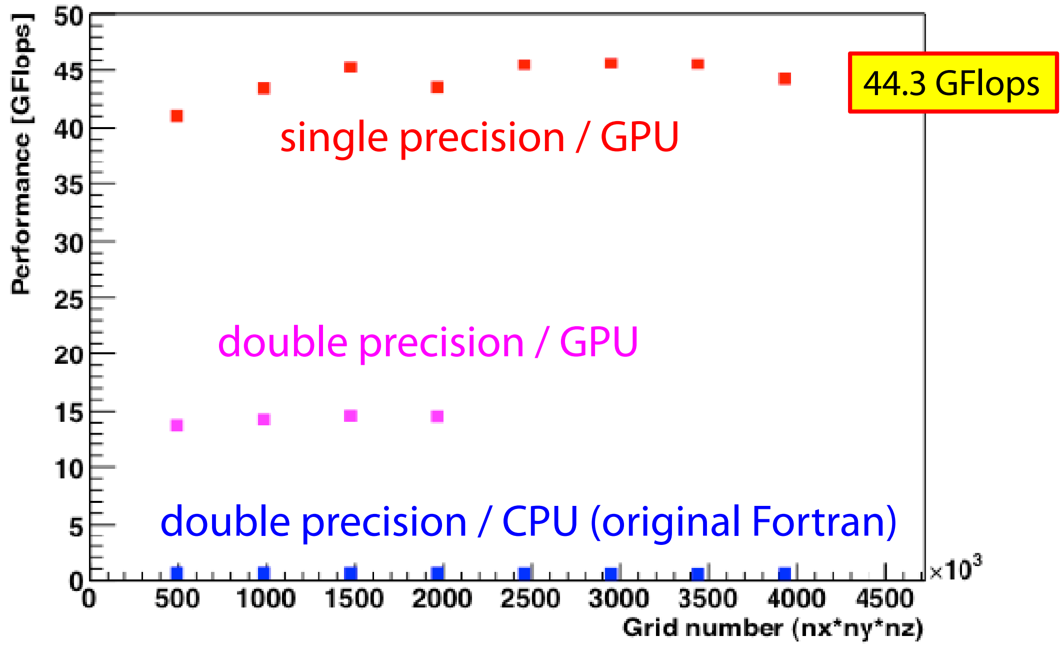
Profile of functions in ASUCA



	Memory bandwidth (peak 159 GB/s)	Instruction throughput (max 1.25)
1	59.4 GB/s	0.30
2	37.0 GB/s	1.04
3	78.1 GB/s	0.51
4	44.7 GB/s	0.81
5	74.0 GB/s	0.53

GP GPU

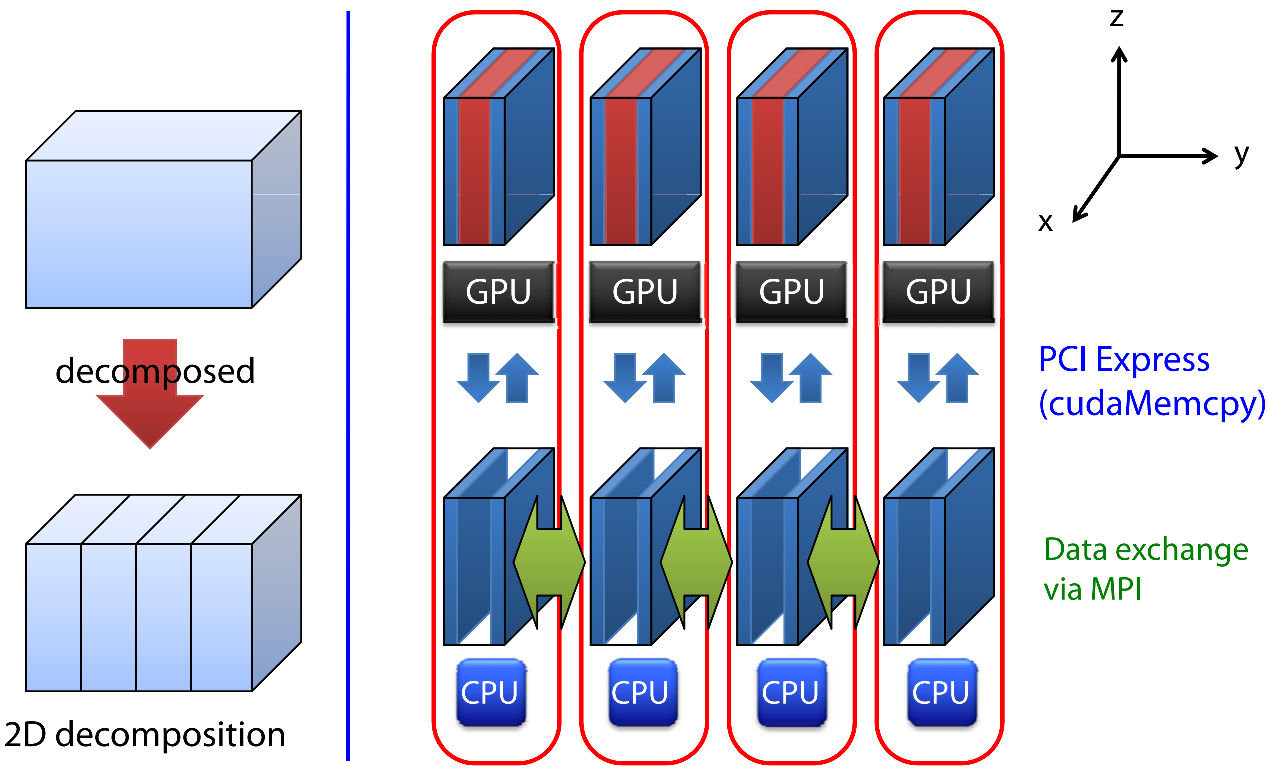
Single GPU Performance



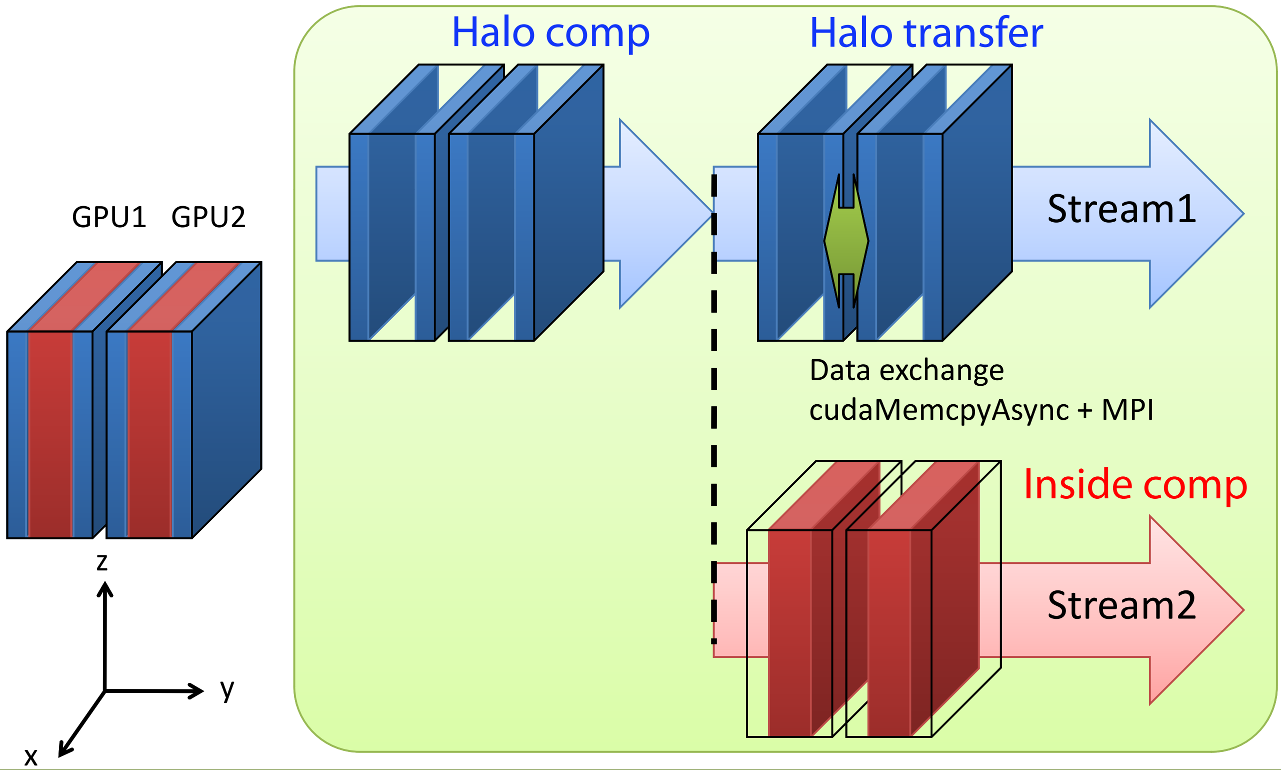
320 x 256 x 64

Mountain Wave Test
NVIDIA Tesla S1070 card

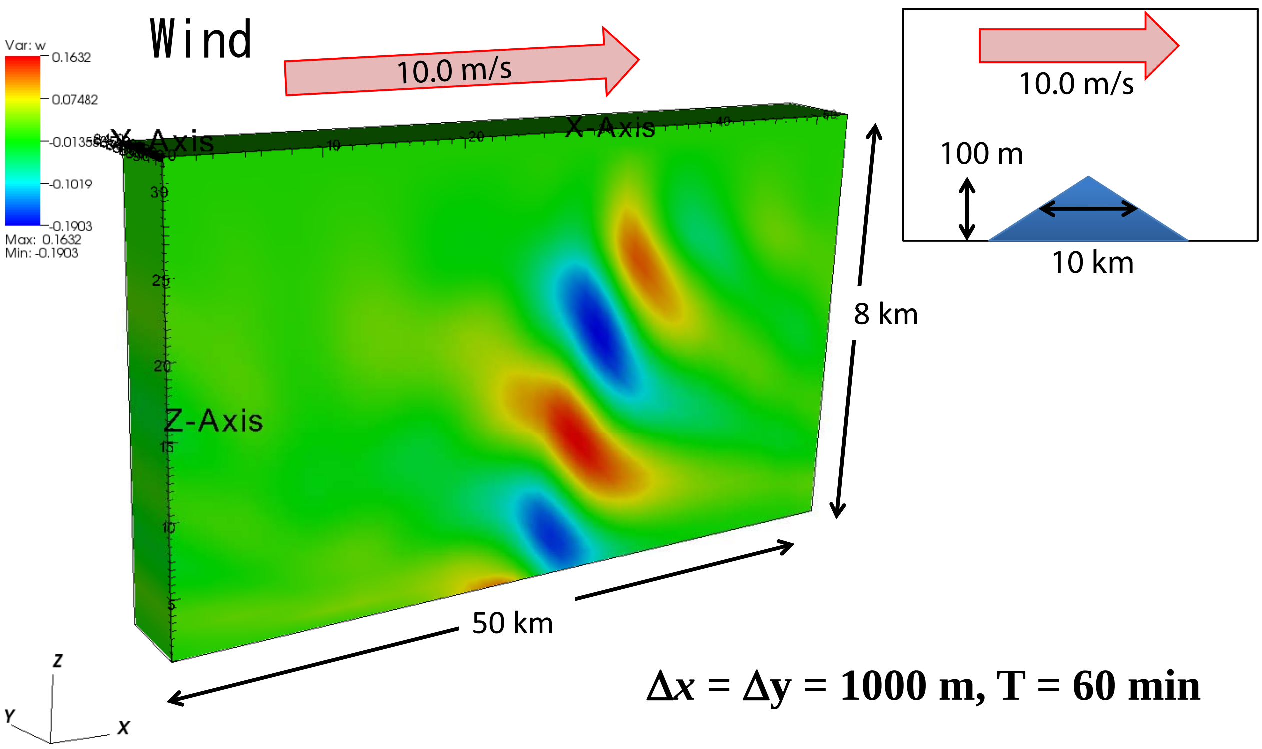
Multi-GPU : Domain decomposition



Overlapping communication with computation



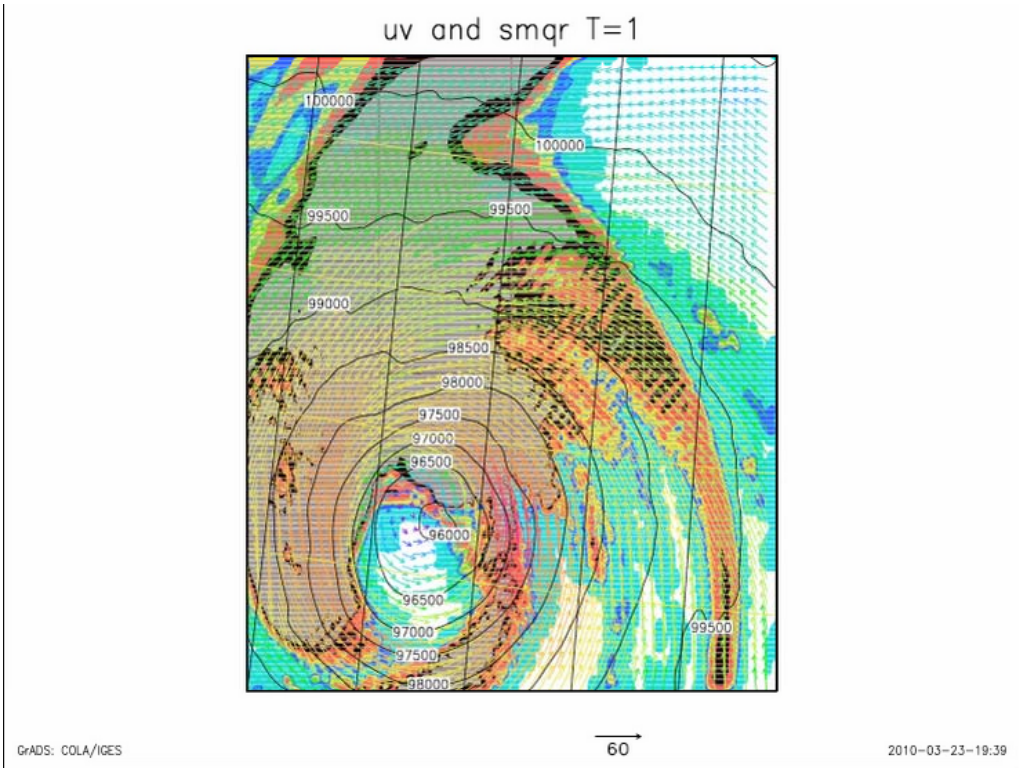
Benchmark Test
Mountain Wave



$\Delta x = \Delta y = 1000$ m, T = 60 min

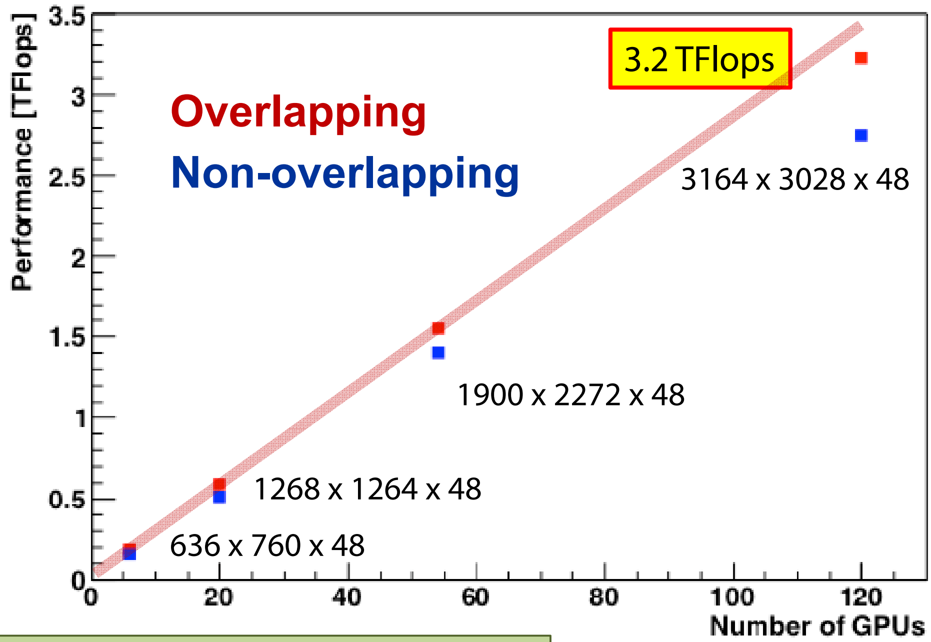
ASUCA Example: Typhoon

GP GPU



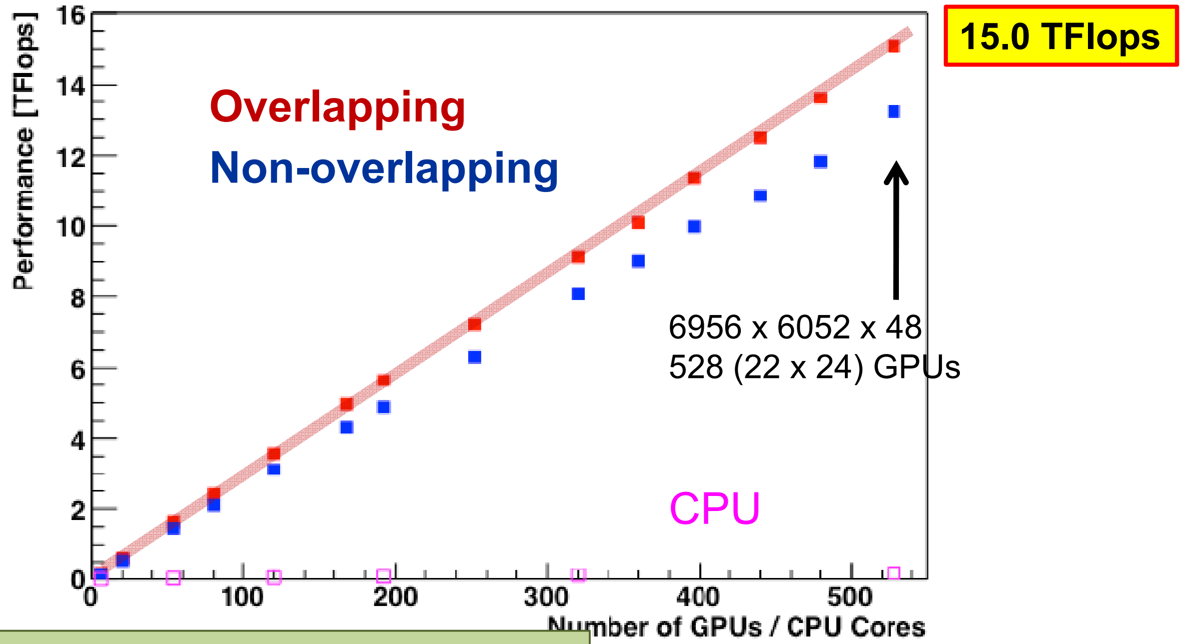
Multi-GPU ASUCA Performance

GP GPU



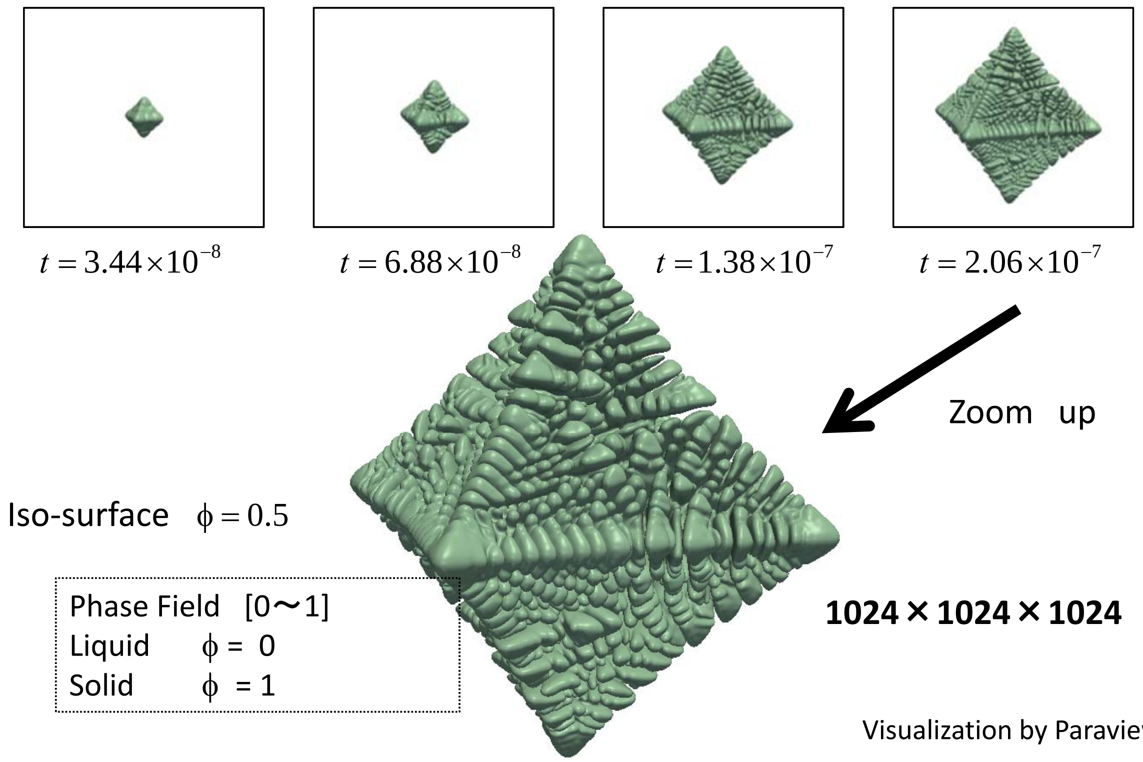
Mountain Wave Test in single precision
NVIDIA Tesla S1070 on TSUBAME

Multi-GPU ASUCA Performance



Mountain Wave Test in single precision
NVIDIA Tesla S1070 on TSUBAME

Dendrite Solidification



Visualization by Paraview 3.6

Dendrite Solidification

Allen-Cahn Equation based on Phase Field Model

$$\frac{\partial \phi}{\partial t} = M \left[\frac{\partial}{\partial x}\left(\varepsilon^2 \frac{\partial \phi}{\partial x} + \varepsilon \frac{\partial \varepsilon}{\partial \phi_x} |\nabla \phi|^2 \right) + \frac{\partial}{\partial y}\left(\varepsilon^2 \frac{\partial \phi}{\partial y} + \varepsilon \frac{\partial \varepsilon}{\partial \phi_y} |\nabla \phi|^2 \right) + \frac{\partial}{\partial z}\left(\varepsilon^2 \frac{\partial \phi}{\partial z} + \varepsilon \frac{\partial \varepsilon}{\partial \phi_z} |\nabla \phi|^2 \right) + 4W\phi(1-\phi)\left\{ \phi - \frac{1}{2} + \beta + a\chi \right\} \right]$$

$$\beta = -\frac{15L}{2W}\frac{T - T_m}{T_m}\phi(1-\phi) \qquad \varepsilon = \bar{\varepsilon}\left(1 - 3\gamma + 4\gamma \frac{\phi_x{}^4 + \phi_y{}^4 + \phi_z{}^4}{|\nabla \phi|^4} \right)$$

Thermal Conduction

$$\frac{\partial T}{\partial t} = k\nabla^2 T + \frac{L}{C} 30\phi^2(1-\phi)^2 \frac{\partial \phi}{\partial t}$$

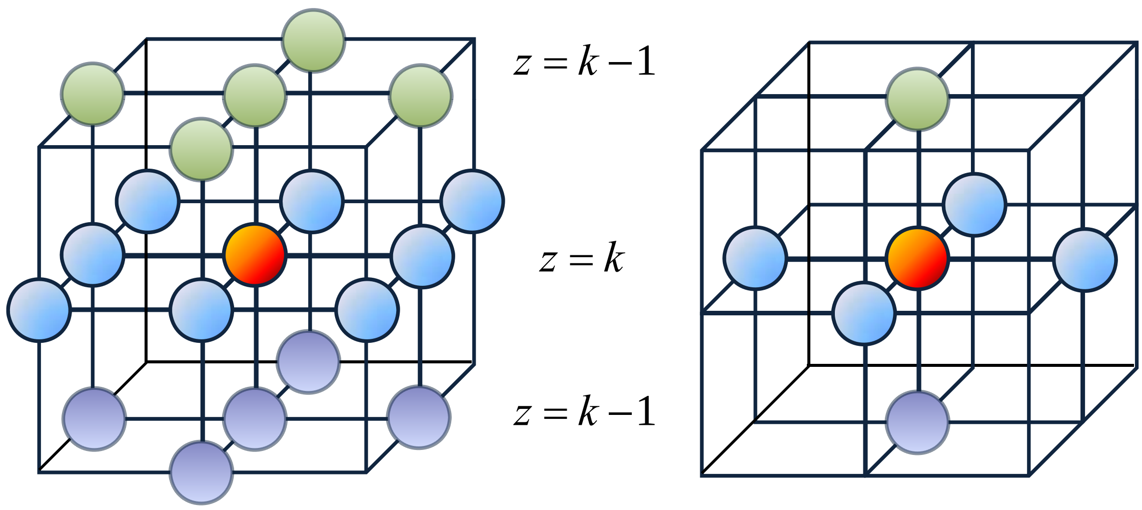
↑ Introduction of non-isotropic surface energy

Second-order Finite Difference Method and 1st-order Euler Time Integration

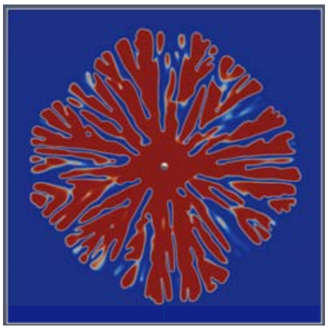
Stencil Access

Phase Field ϕ — Temperature t

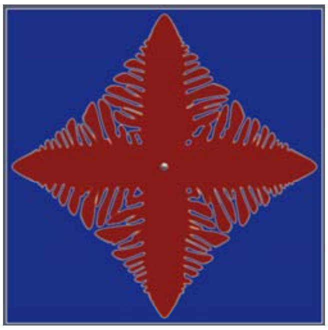
19 points to solve $\phi_{i,j,k}$ — 7 points to solve $t_{i,j,k}$



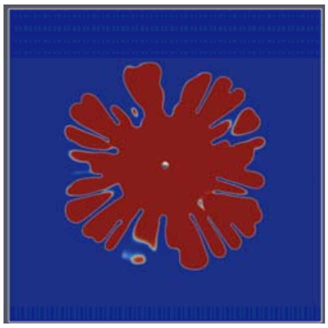
Dependence on Parameters



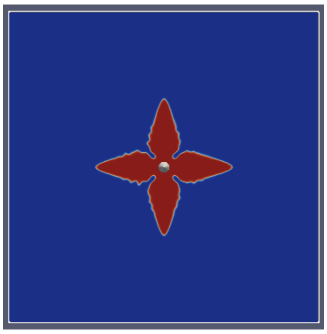
γ = 0.015, A=0.01



γ = 0.075, A=0.01



γ = 0.015, A=0.01, Δt=half



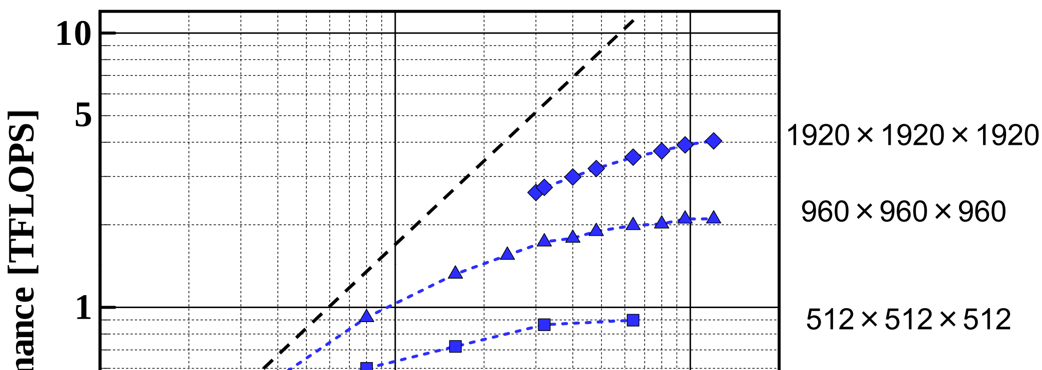
γ = 0.1, A=0.01

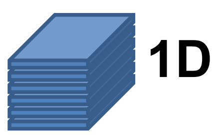
γ = 0.075, A=0.01

γ = 0.075, A=0.005

Multi-GPU Peformance

w/o overlapping

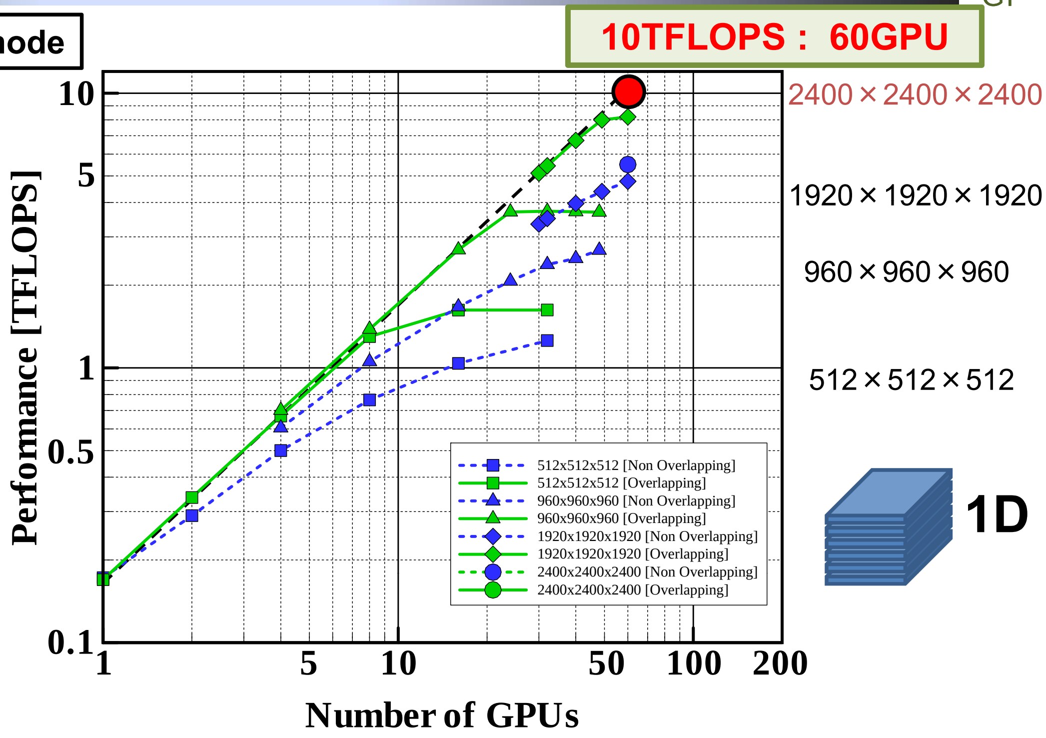




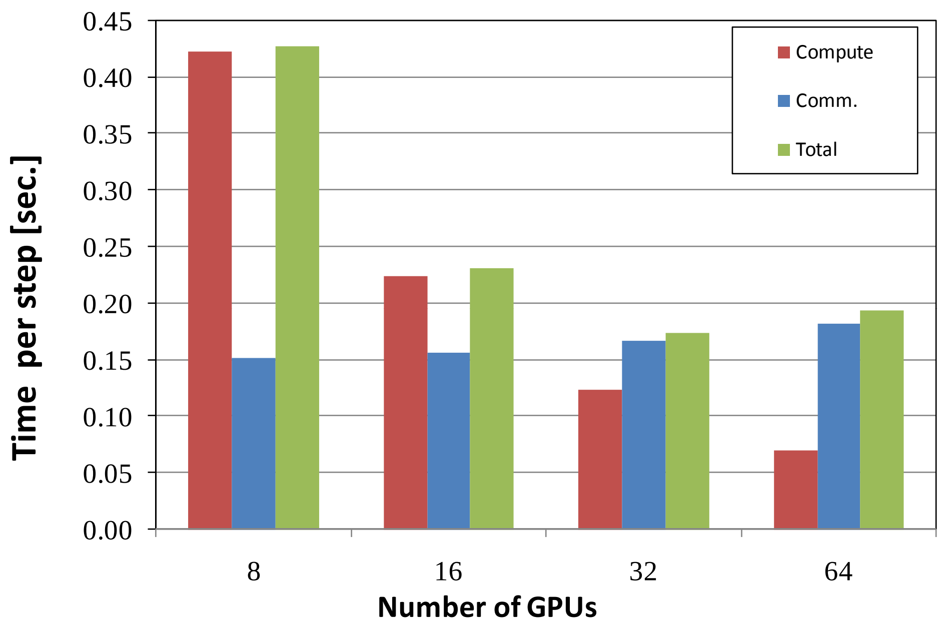
Multi-GPU Peformance

w/o overlapping

1 GPU/node



Breakdown of Overlapping Case

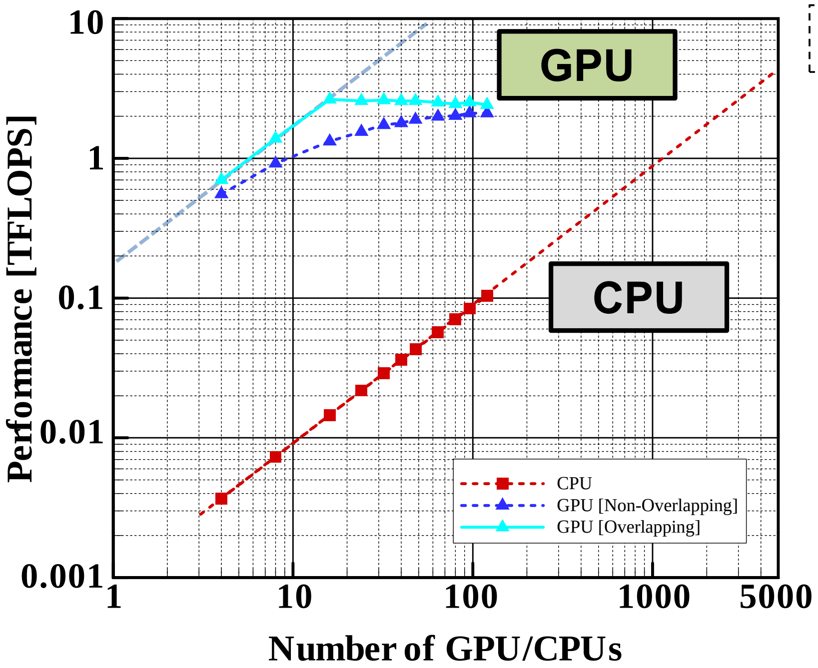


Grid Number:
1024x1024x1024

Machine:
TSUBAME
Tesla S1070
Infiniband
(hpc1tes2 Queue)

- Computational time becomes short for more GPU numbers
- Communication time is almost same
- The communication time can not be hidden for more than 32 GPUs

Direct Comparison : CPU/GPU



960^3 mesh

Tesla S1070

60 GPU

10000 CPU

TSUBAME 1.0

Electric Power Consumption

for Multiple-GPU Applications

CFD Applications

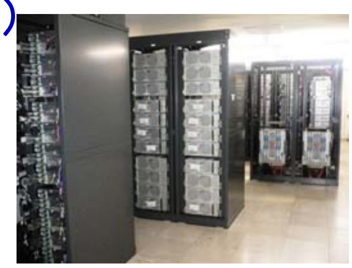
MAX 200W/GPU

Comparing to TSUBAME Opteron CPU 1 core,

Tsunami: x100 （10 GPU = 1000 CPU）

Lattice Boltzmann: ～ x100

Phase Field Model: x170



TSUBAME ～ 1MW/10000CPU

When we assume the acceleration is typically x100,

1 GPU < 200W for our applications

100 GPU: 20 kW

10000 CPU core of TSUBAME : 1MW (1000kW)

1/50 Ultra low Power

Multi-GPU Summary

GP GPU

Some CFD applications show good strong scalability up to 32 GPUs in the case of TSUBAME.

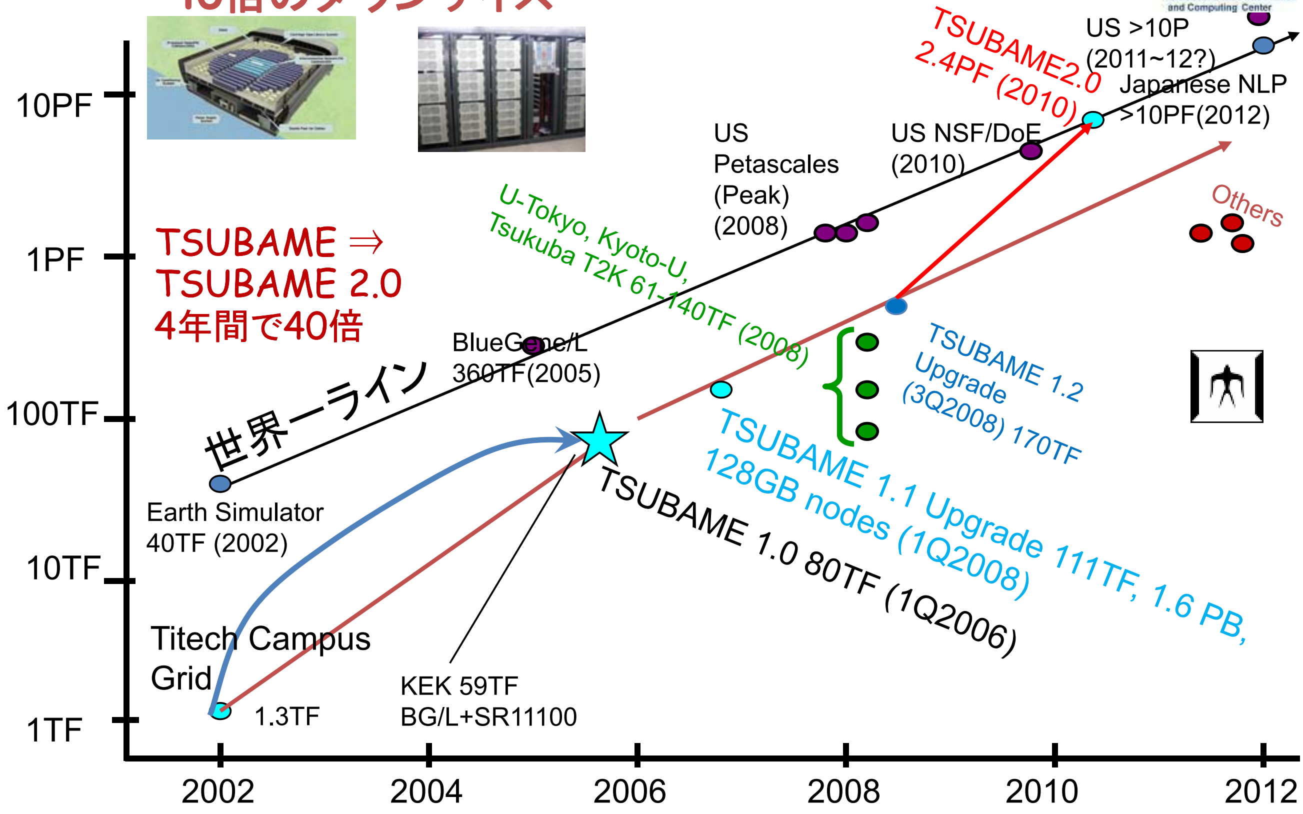
The balance between computation and communication performance becomes bad because of the high GPU performance.

In order to achieve high performance for multi-GPU application, the overlapping technique between computation and communication is very important.

Be careful for GPU-to-CPU data transfer (cudaMemcpy) and CPU-to-CPU data transfer (MPI library).

TSUBAME2.0 世界トップクラスへ

地球シミュレータ ⇒ TSUBAME 4年間
40倍のダウンサイズ



TSUBAME 共同利用

他大学や公的研究機関の研究者の「学術利用」

有償・成果公

民間企業の方の「産業利用」

トライアルユース

有償・成果公開

有償・成果非公開

その他の組織による社会的貢献のための「社会貢献利用」

有償・成果公開

有償・成果非公開

有償利用は随時申請受付中！！

成果公開　10万円 / 1口

成果非公開　40万円 / 1口

TSUBAME共同利用サービス

http://www.gsic.titech.ac.jp/kyodou/

複数GPUを試してみることのできる絶好の機会です。ご利用をお待ちしています。

GSIC

GPU コンピューティング研究会

GP GPU

東京工業大学・学術国際情報センターの中の研究会活動

活動内容:

- 月に一度の GPU (CUDA) 講習会
- 2～3ヵ月に一度の講演会
- 一年に一度のシンポジウム(予定)
- 個別のユーザーサポート(可能な範囲で)
- メーリングリストを通じた情報交換

東工大の教員・学生だけでなく、他大学・研究機関、民間企業の方も参加(入会)可能。基本的に無料。

http://gpu-computing.gsic.titech.ac.jp/