International Advanced Research Workshop on High Performance Computing, Grids and Clouds 2010
June 21~June 25 2010, Cetraro, Italy

HPC Infrastructure and GPU Computing Activities in KISTI

Hongsuk Yi

hsyi@kisti.re.kr

KISTI Supercomputing Center

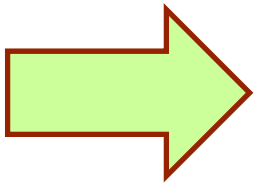
Korea Institute of Science and Technology Information

Outline

- HPC infrastructure and activities in KISTI
- Heterogeneous Computing with GPU
  - What is the scalability
  - Heterogeneous Computing with MPI+CUDA

Where is KISTI?

- KISTI is responsible for national cyber-infrastructure of Korea
  - Mission is enable Discovery through National Projects
- I will don't talk about, today
  - Grid Project (2002~2009) ~ K*Grid
  - e-Science Project (2005 ~)
  - Scientific Cloud Computing (2009~)
  - Research Network Project (2005~)



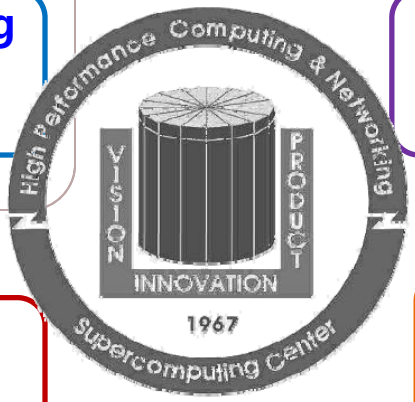
HPC Infrastructure & Multi-GPU programming

Keep securing/providing world-class supercomputing systems

Computing /Network Resource

Validate newly emerging concepts, ideas, tools, and systems
K*Grid, Scientific Cloud

Testbed Center



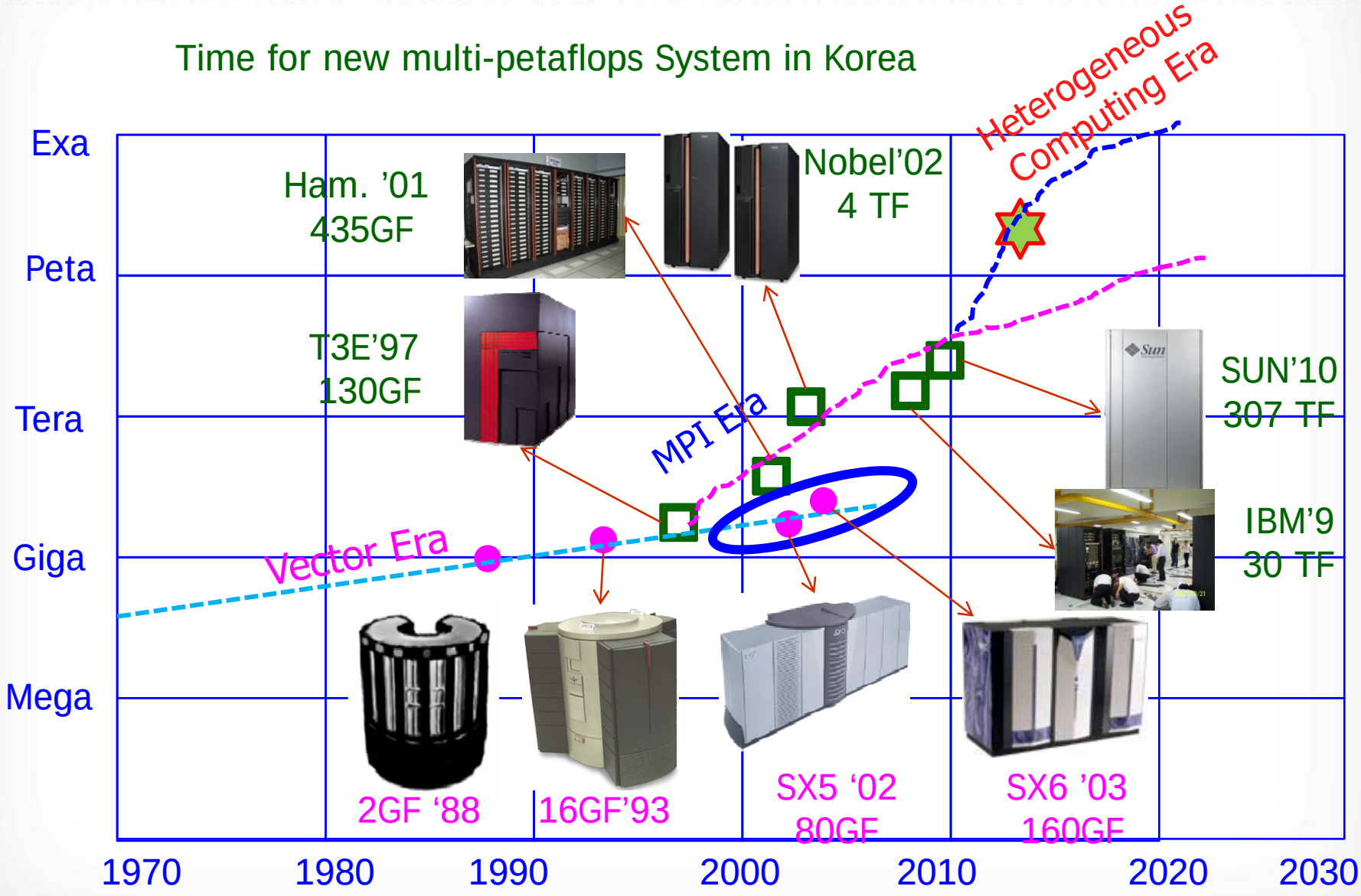
Cyber-infrastructure Environment

Help Korea research communities to be equipped with proper knowledge of Cyber-infra.

Value-adding Center

Make best use of what the center has, to create new values

History of Supercomputers in Korea

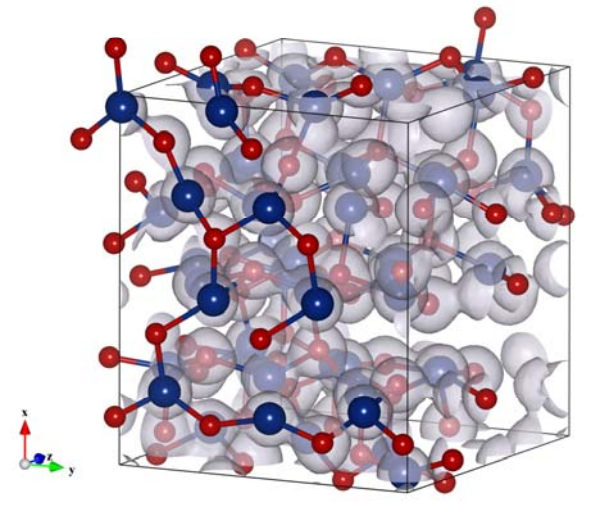


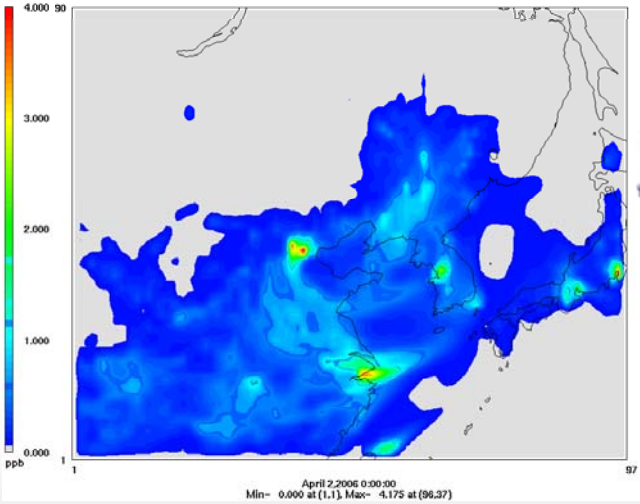
HPC ACT of Korea

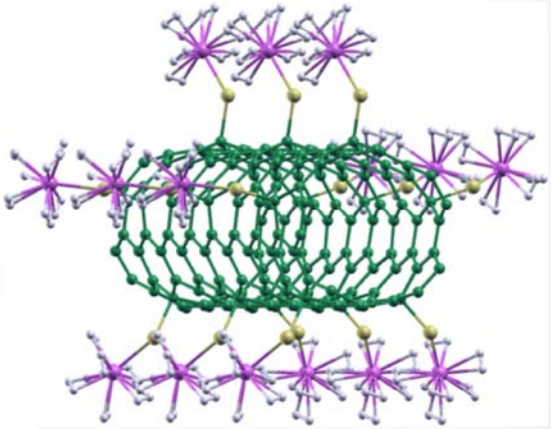
- HPC ACT has been started from 2004
  - The ACT is currently **awaiting** the approval of the National Assembly
- Purpose
  - To provide for a well coordinated national program to ensure continued Korea role in HPC and its applications by
    - ✓ improving the coordination of supercomputing resource on HPC
    - ✓ maximizing the effectiveness of the Korea's networks research (KREONET)
- We can make more contribution by expanding support for
  - National agenda research program in the field of computational science, and development of cyber-infrastructure environment, ad well as applications of extreme scale computation

NCRC

- National Core Research Center for Computational Science and Technology
  - Budget ~ 10M$
  - Sep. 2010~ (not yet completely determined)
- Application Domains
  - Energy transformation by quantum simulation
  - Migration of pollution by air including yellow sand
  - New material for Energy

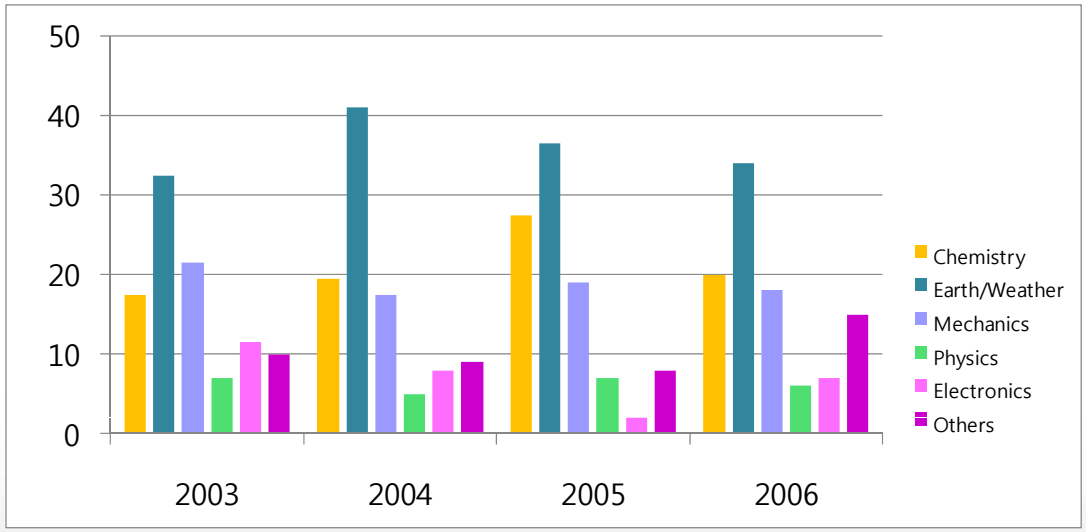






Supercomputing Resource (Tachyon)

- Tachyon-II is the 15th ranked in Top500 (June, 2010)
  - Sun Blade x6048, Intel Nehalem procs~26,232 (Memory~157 TB)
  - Peak ~ 307 Tflops (Sustained Peak 274 Tflops)
- Providing about 30% of whole computing capacity for public research in Korea
  - Users form 200 institutes in Korea
  - Utilization : 70~80%
  - Little room for large scale grand challenge problems

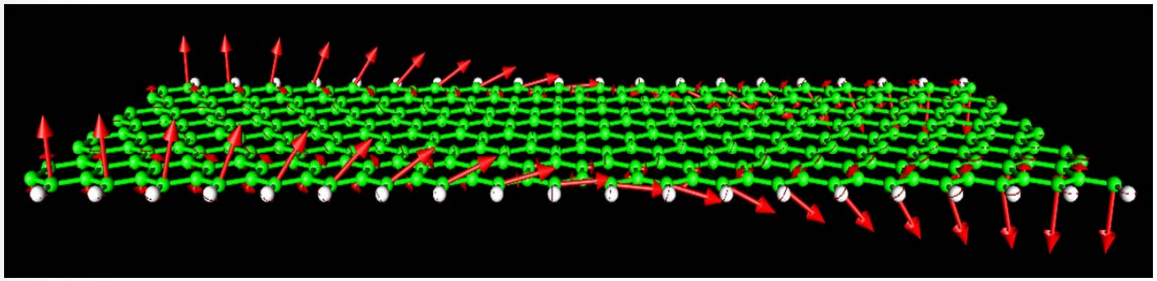


User Support and Applications

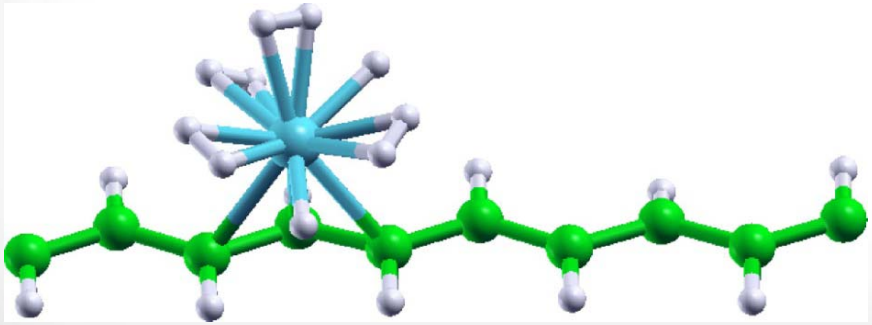
- Support code optimization and parallelization
  - We optimize and parallelize user's code.
  - Performance improvements from 5 times to more than 1,000-fold

Year	2004	2005	2006	2007	2008	2009
Optimization	6	12	13	15	20	20
Parallelization	6	8	11	15	20	25

✓ User's Application through Grand Challenge Problems



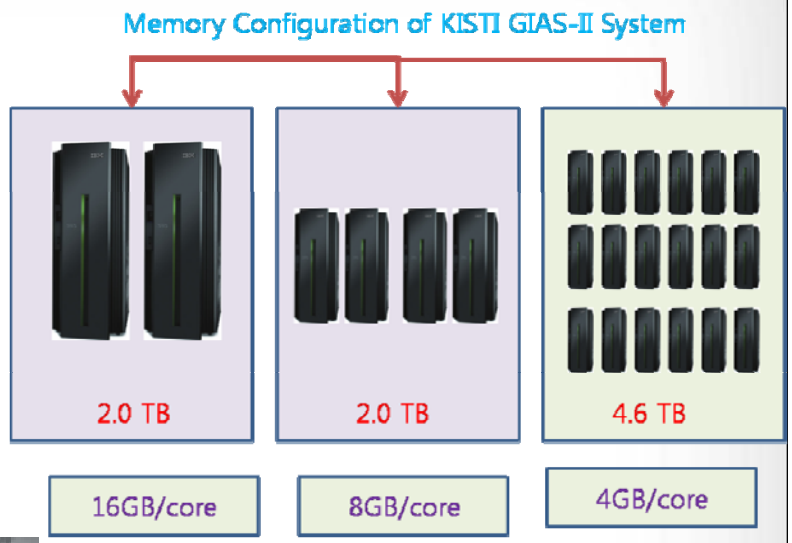
Magnetic control of edge spins
by K. S. Kim,
Nature Nanotech. 3, 408 (2008)



Hydrogen Storage Materials by J. Ihm,
Phys. Rev. Lett. 97, 056104 (2006)

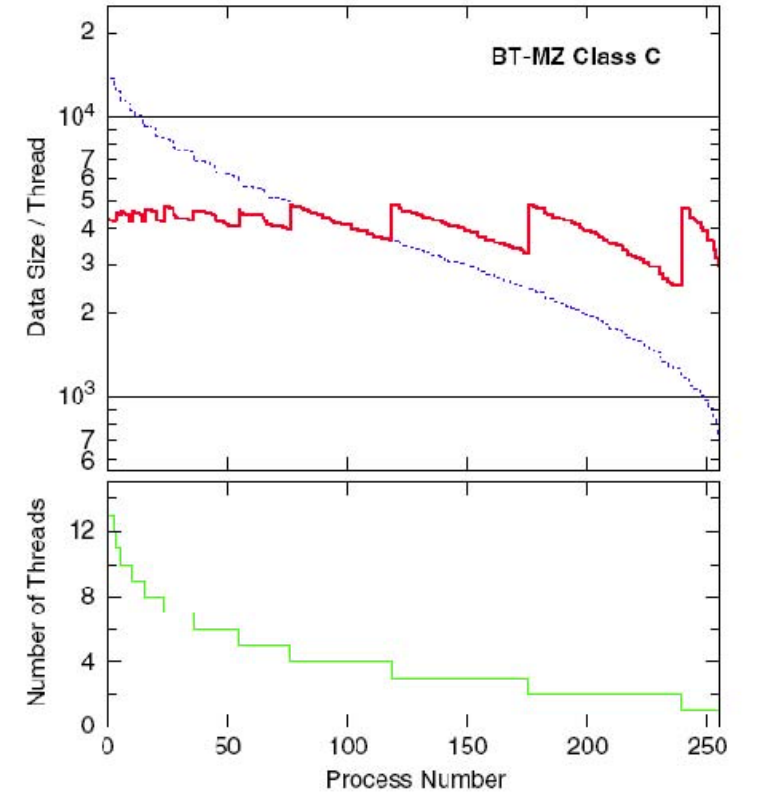
Supercomputing Resource (GAIA)

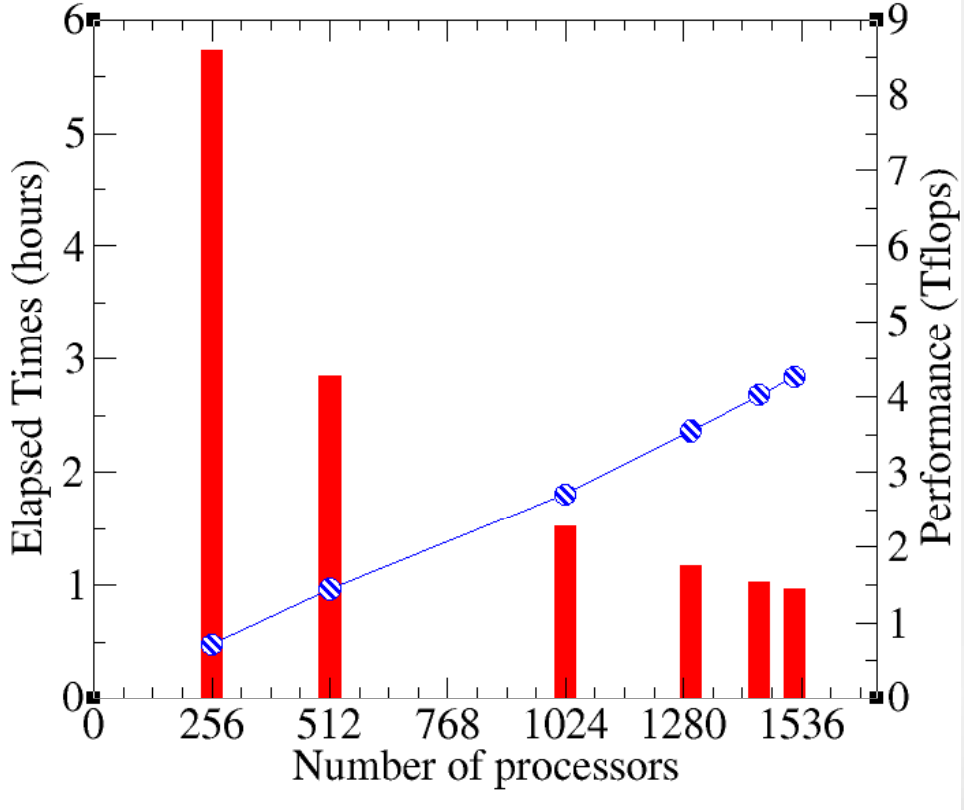
- GAIA-II is SMP Cluster
  - 393th in Top500, 2009.11
  - IBM POWER6 5 GHz, Power 595,
  - Number of Procs ~ 1,536 (64 cores/node)
  - Rpeak~30.7 Tflops (Sustained Peak 23.3 Tflops)
  - Memory (8.7 TB)



Hybrid Programming : Multi-Zone NPB

- GAIA-2 (IBM p6 5GHz)
  - ✓ Big memory ~16GB/cores
  - ✓ Bin-packing algorithm
  - ✓ The size of zones varies ~20
- BT-MZ with class F
  - ✓ Memory required ~ 5 TB
  - ✓ MPI+OpenMP Programming
  - ✓ Performance ~ 4.5TF (15%)

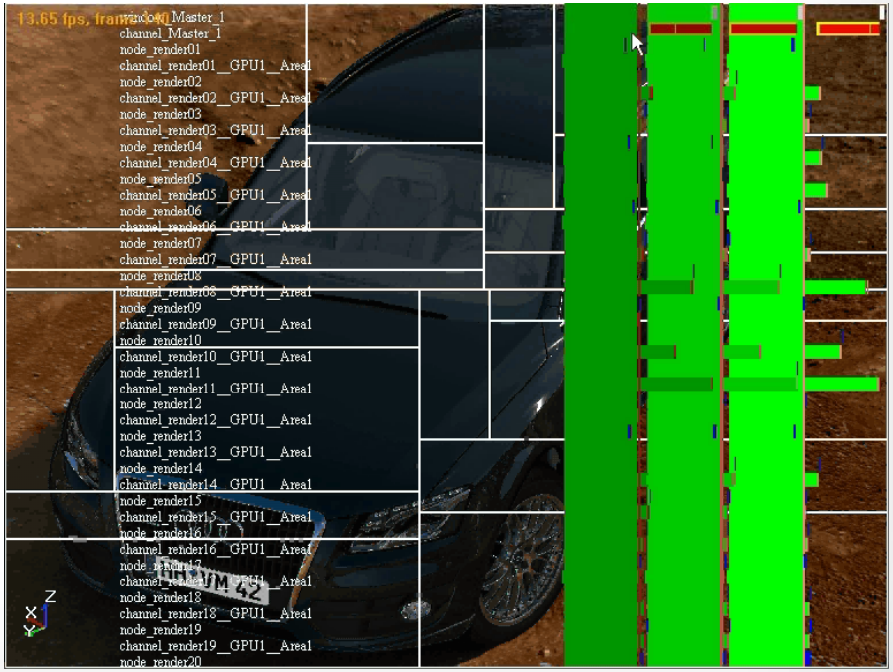




GPU computing for visualization

- All KISTI's visualization systems
  - have direct connection to GLORIAD,
  - whose bandwidth is 10 Gbps

Visualization Computer		
Total number of nodes		±150
CPU	# of CPU cores	800+
Total memory		3.5+ TB
GPU	Model	NVIDIA Quadro FX 5600
	# of GPUs	96+
Network	Interconnection	20 Gbps
	External network	160+ Gbps

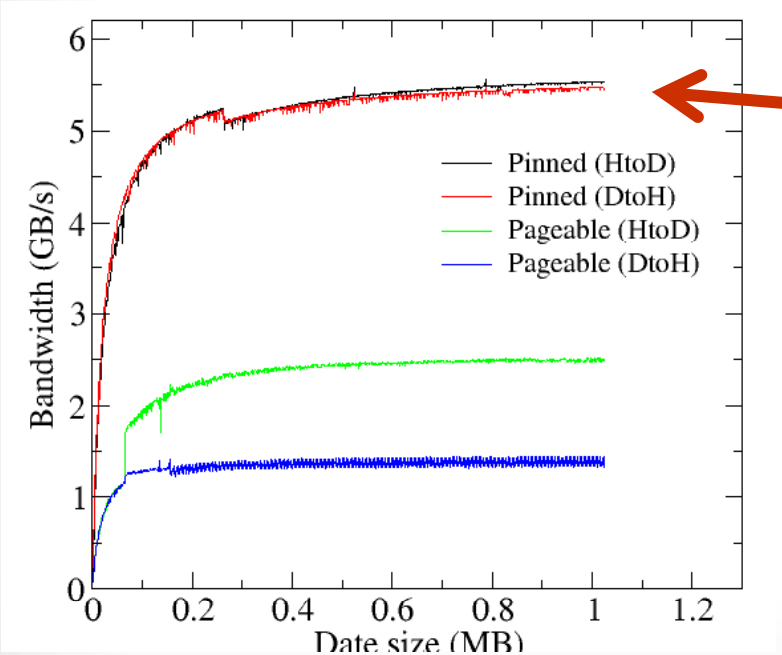


GPU Computing Activities

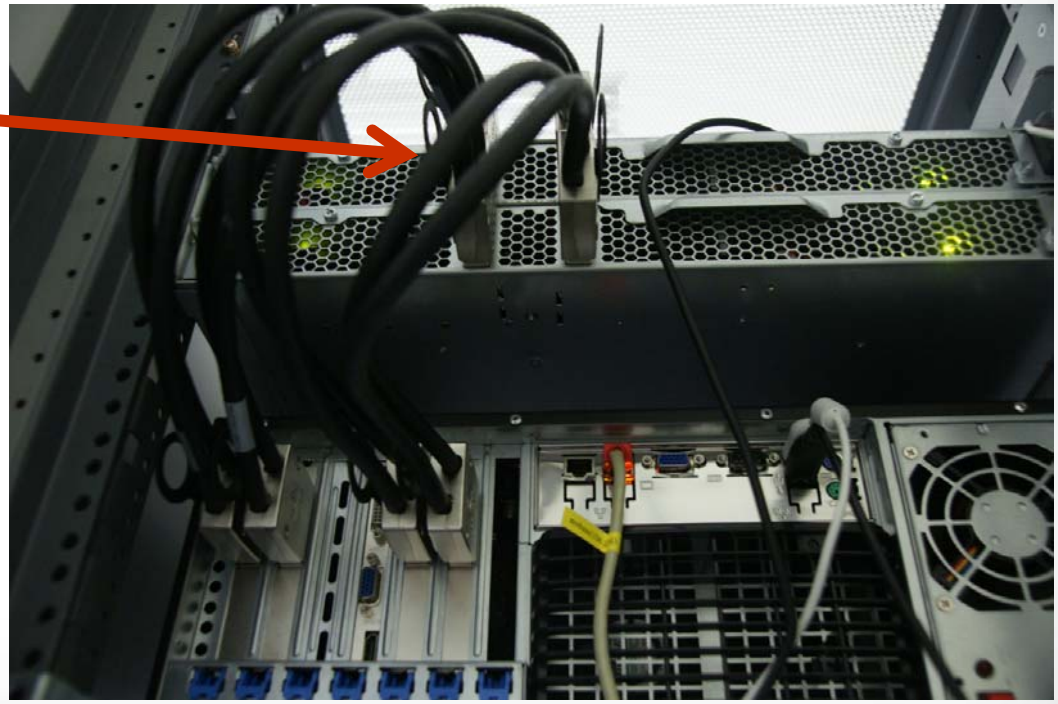
- KSCSE (Korea Society for Computational Sciences and Engineering)
  - Establish as a new computing society in 2009
  - Support GPU computing Forum and workshop
    - ✓ 200 participants, two days, May 2010, Seoul
  - Open for international collaboration on the extreme scale computing

Heterogeneous Computing Testbed

- Heterogeneous Computing System refer to system
  - that use a variety of different types of computational units.
  - A computational unit could be a GPU, co-processor, FPGA
- KISTI Heterogeneous GPU Testbed
  - NVIDIA 2* S1070 (8 GPUs), D870, GTX280



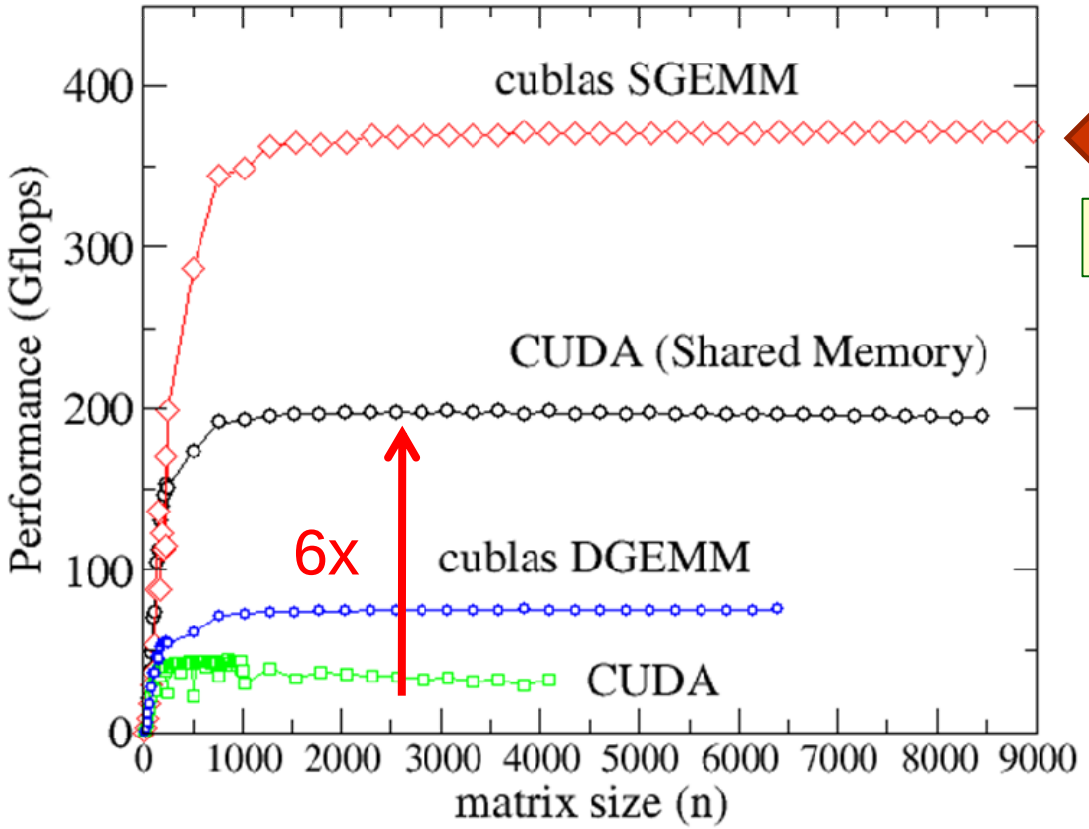
PCI-e : 16x



Performance Benefit of GPU Computing

- The ratio of operations to elements transfer : O(N)
  - Matrix-Matrix Multiplication
    - ✓ Operation N^3, Transfer $3xN^2$, ,Scaling O(N)
  - Matrix-Matrix Addition
    - ✓ Operation N^2, Transfer $3xN^2$, Scaling O(1)

NVIDA GTX280 (Peak 933 GHz)



40% Efficiency

Matrix-Matrix Multiplication
C= AxB

Image Compression Using SVD

- SVD is an important **factoriz ation** of matrix
  - with many applications in signal p rocessing and statistics.
  - culaSgeSVD() by using CULA
- RBG full color
  - 2048x2048 total 4,194,304 pixels

Original	12,288 KB
1 Rank	12 KB
10 Rank	120 KB
50 Rank	600 KB
80 Rank	960 KB
100 Rank	1,200 KB

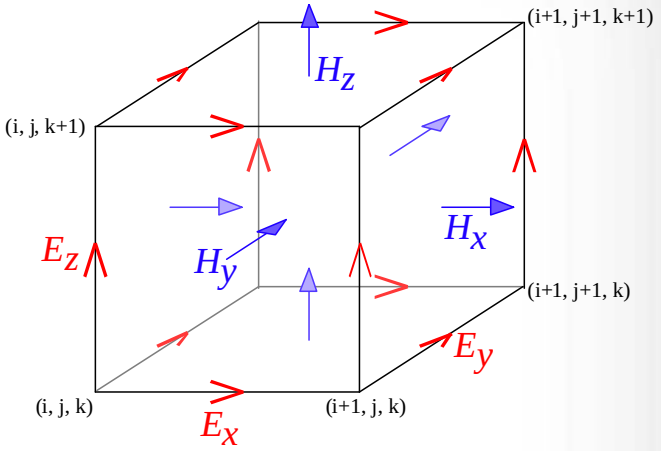
Original 1 Rank

10 Rank 50 Rank

80 Rank 100 Rank

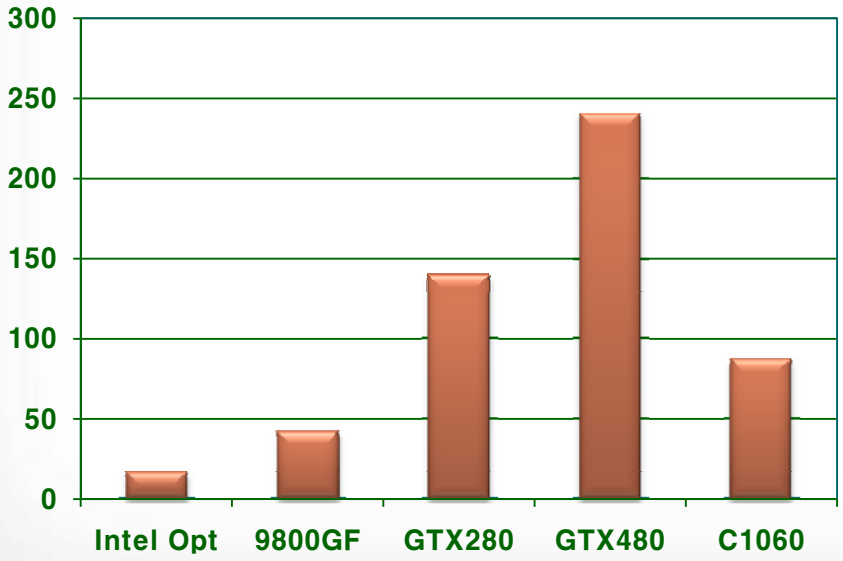
3D FDTD on GPU

- Finite-Difference Time-Domain
  - Divide both space and time into discrete grids
- 3D FDTD Benchmark Results
  - Memory ~ 988 MB, Grid~300x300x240

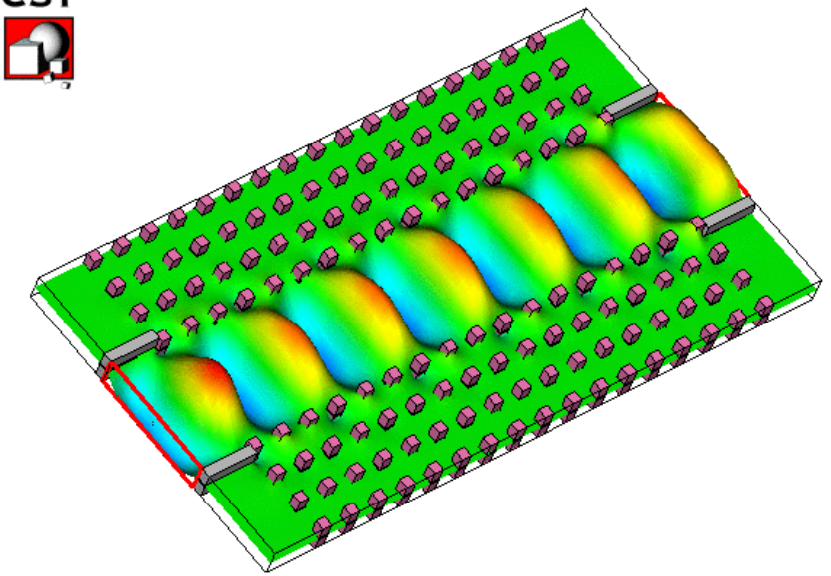


$$\frac{\partial \mathbf{H}}{\partial t} = -\frac{1}{\mu}\vec{\nabla} \times \mathbf{E} \qquad \frac{\partial \mathbf{E}}{\partial t} = \frac{1}{\varepsilon}\vec{\nabla} \times \mathbf{H} - \frac{\mathbf{J}}{\varepsilon}$$

By Q. Park and K. Kim, Korea Univ.



CST



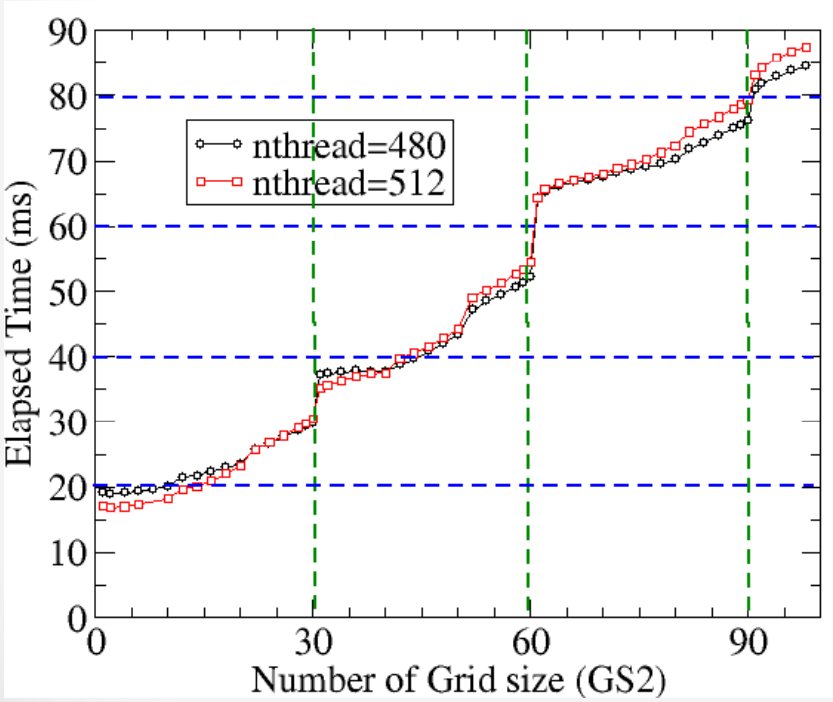
RNG and Monte Carlo Algorithm

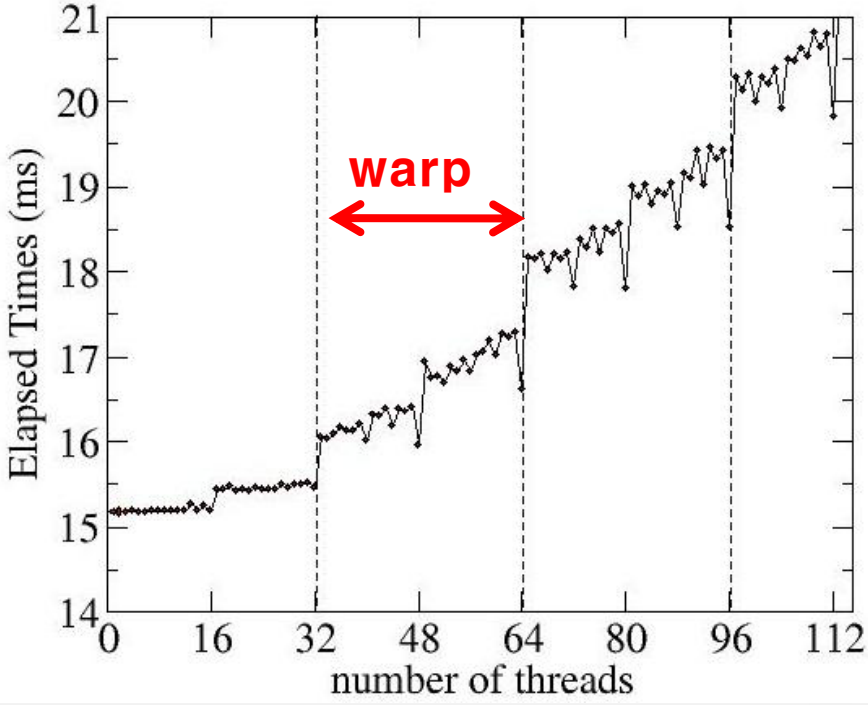
- Ising Model Model and probability weight

$$E = -\sum_{<i,j>=1}^{L} J_{ij} s_i s_j - H \sum_{i=1}^{L} s_i \qquad p = \frac{1}{Z(T)} exp[-E_\alpha / k_B T]$$

✓ 1000 MC steps and 512 threads per block independent of block number and (tx=512, ty=1,by=1), bx

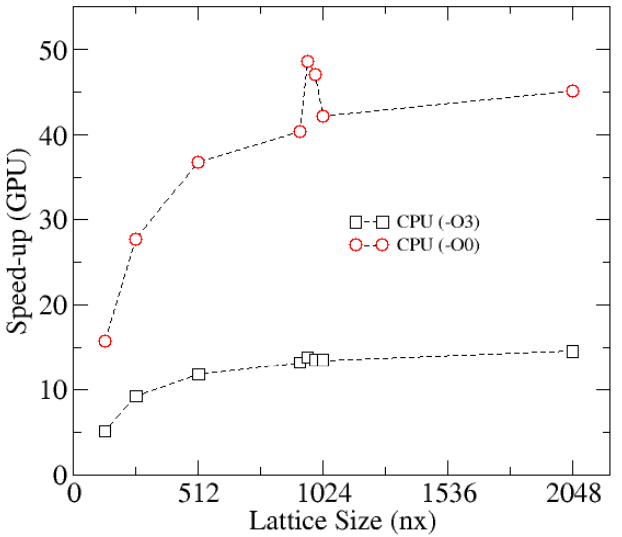
✓ Size of the warp ~32 and the number of thread block ~ 30



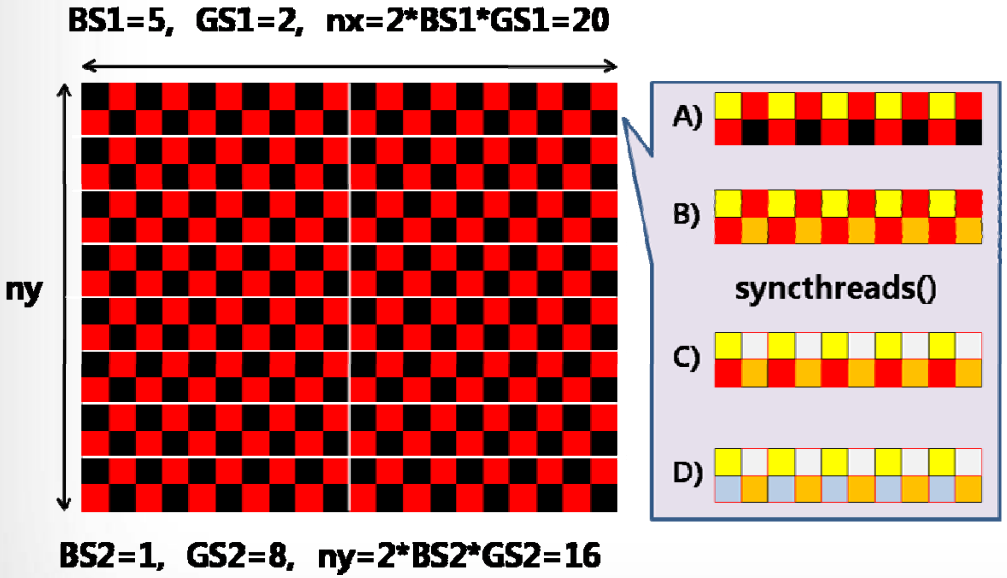


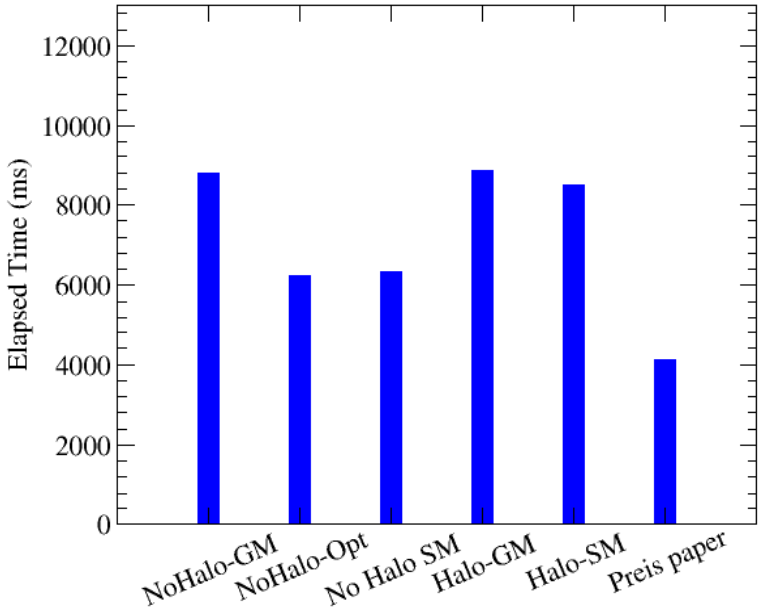
Performance of Ising Model on a GPU

- Use the checkerboard decomposition
  - to avoid the read/write conflicts
  - the spin field and the seed values of the RNGs
  - Divide the spin on the lattice into blocks on the GPU.



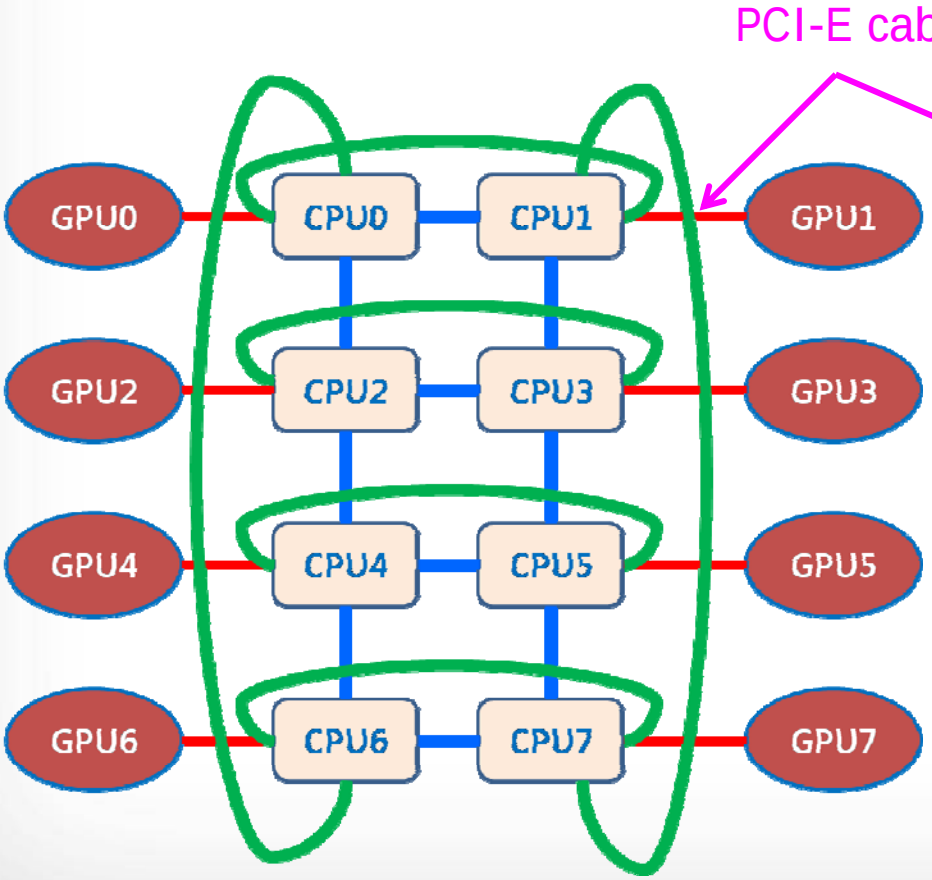
Example for a (20x16) lattice

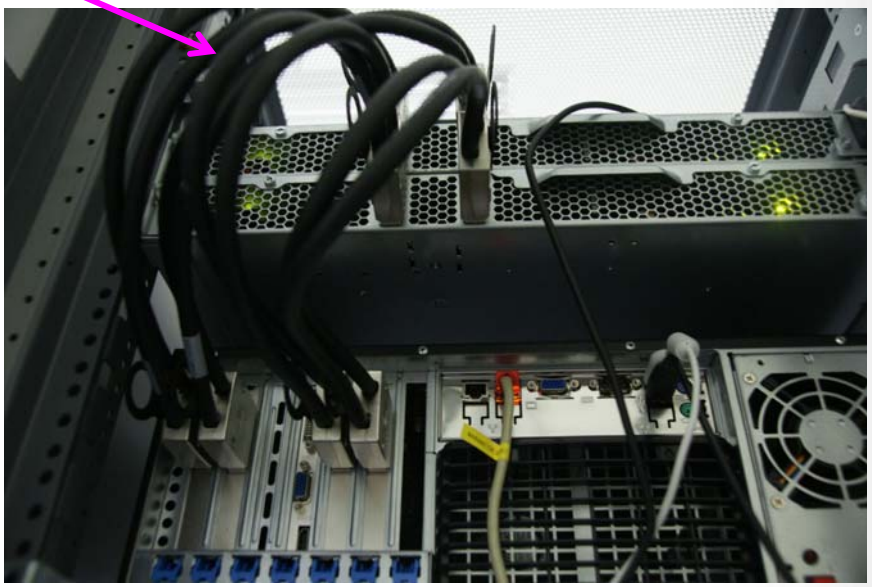




MPI Virtual Topology + CUDA Model

- VT is weak scaling problem
  - ✓ The problem size grows in direct proportion to the num. of cores
  - ✓ Using PBC with MPI_Sendrecv()
  - ✓ Intel Nehalem 8 cores + Nvidia Tesla C1060*8GPUs

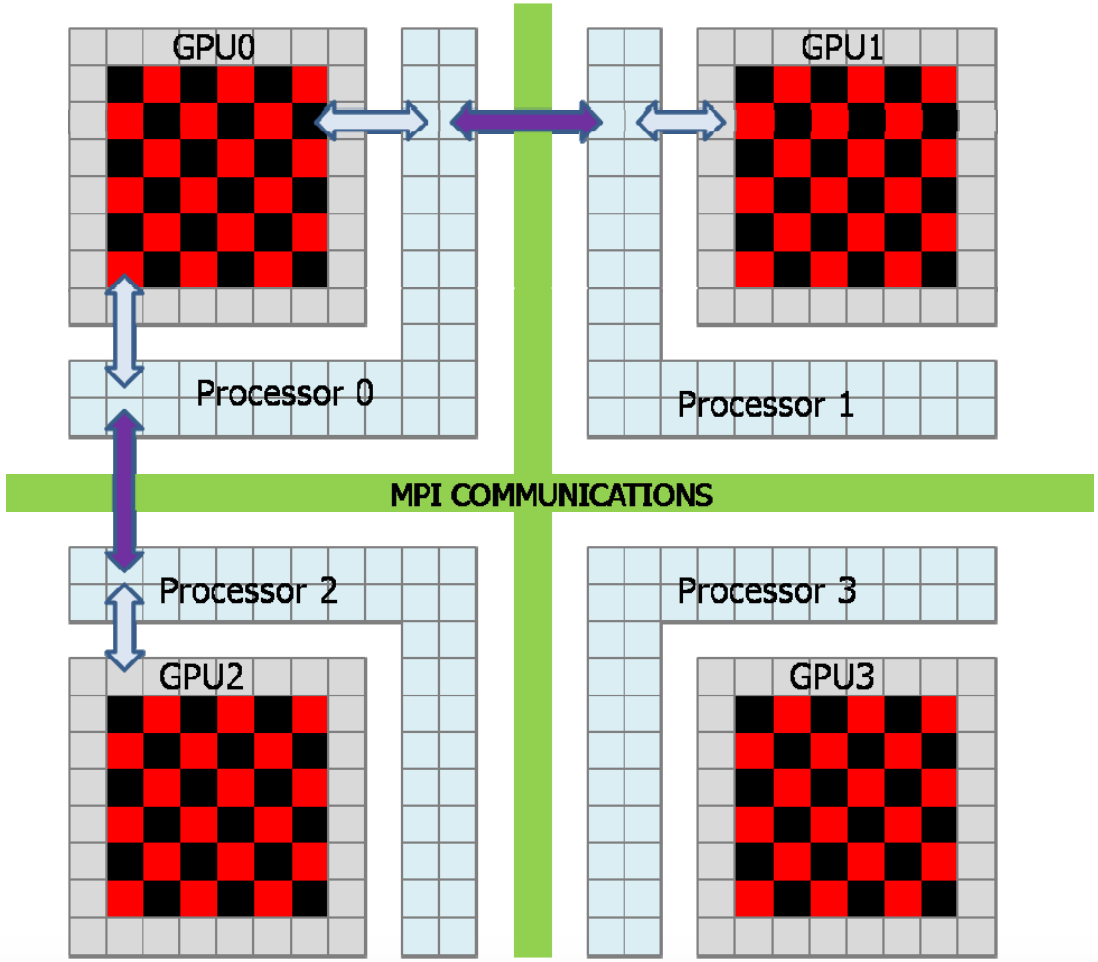




Heterogeneous Communication Pattern

PCIe InfiniBand PCIe

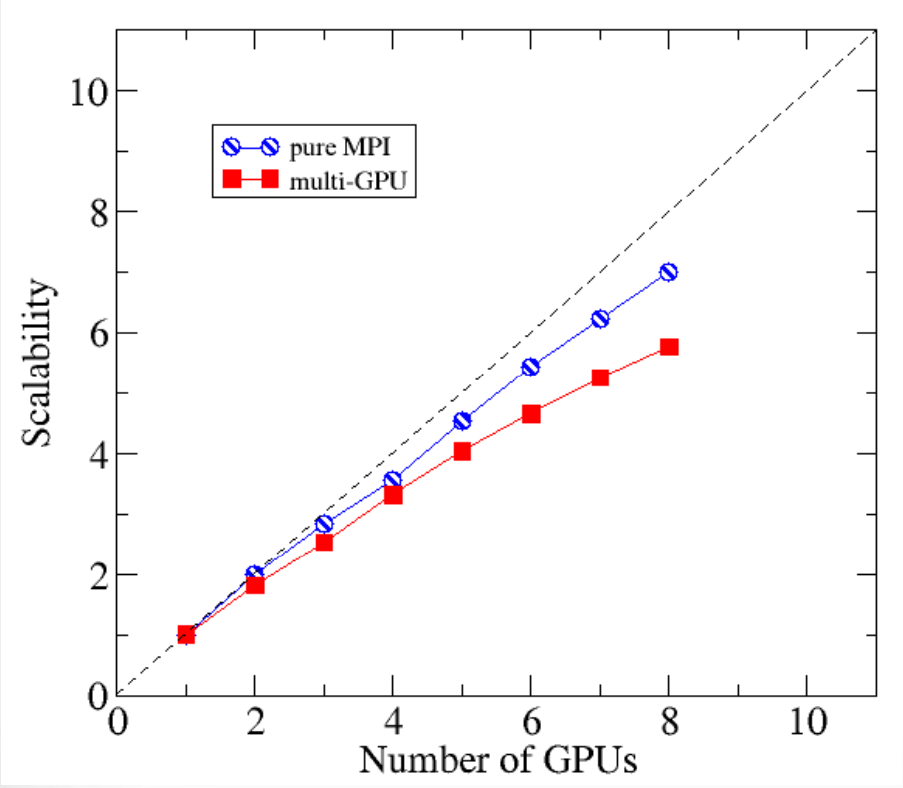
GPU0 --------→ CPU1 ---→ CPU2 -------> GPU2

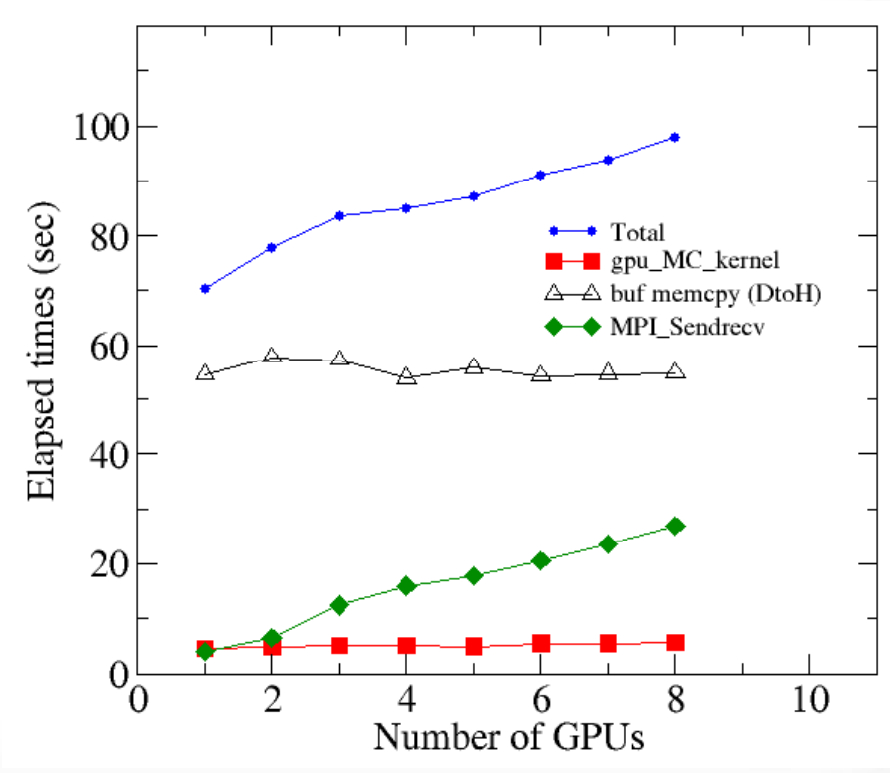


There is an extra cudaMemcopy() involved in message passing

GPU scaling Issues

- Achieving good scaling is more difficult with GPUs
  - The kernels are much faster so the MPI communication becomes a larger faction of the overall execution time





Summary

- HPC ACT of Korea is in progress
  - The act is awaiting the approval of the Korea assembly
  - Time for heterogeneous petaflops system in Korea
  - Consider too many things, power, space, user's ability of porting
- In the MPI+CUDA model, achieving good scaling is more difficult than pure MPI since
  - the kernels are still faster on the GPU
  - There is an another communication over head between CPU and GPUs

Q&A

Thank You!