GPUを使った
大規模並列
N体シミュレーション

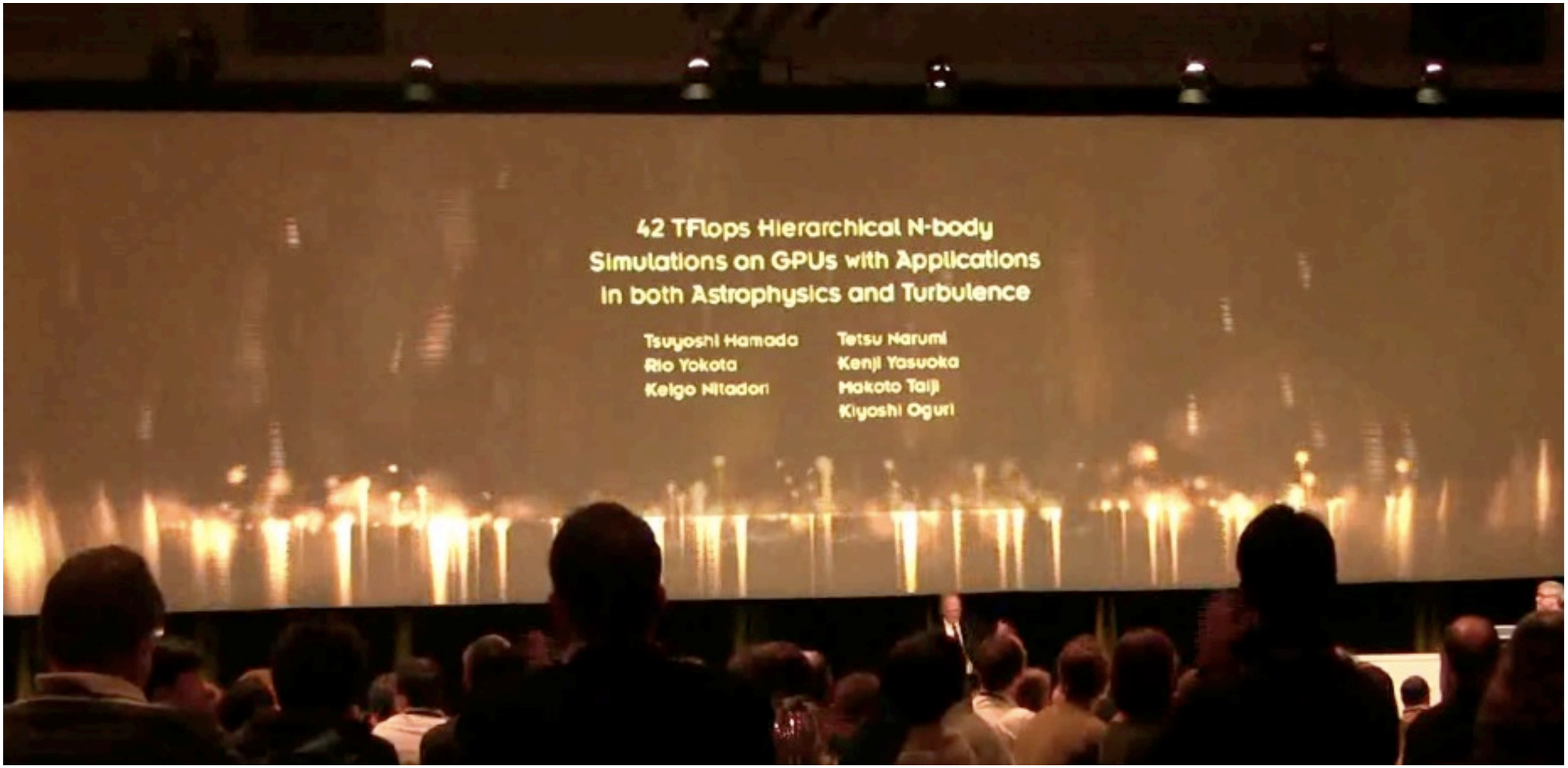
長崎大学工学部
超高速メニーコアコンピューティング研究センター
濱田 剛　（テニュアトラック助教）

Collaboration with
Rio Yokota (Bristol University)
Keigo Nitadori (RIKEN)

2009年 Gordon Bell 賞

受賞しました

受賞論文タイトル


42 TFlops Hierarchical N-body
Simulations on GPUs with Applications
in both Astrophysics and Turbulence
Tsuyoshi Hamada
Rio Yokota
Keigo Nitadori
Tetsu Narumi
Kenji Yasuoka
Makoto Taiji
Kiyoshi Oguri

今回受賞したGordon Bell賞の歴史的特徴?

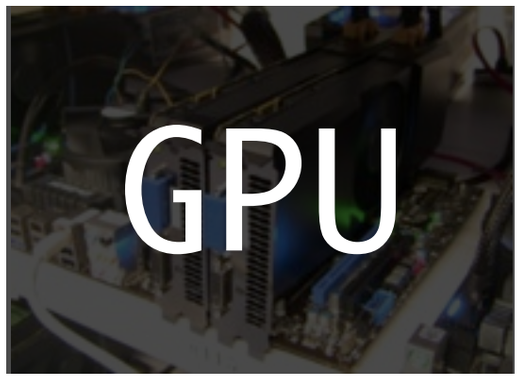
- GPUを使ったシステムが受賞した前例無し. GPUでは史上初
- 日本人の受賞者としては３年ぶり
- コスト性能部門での受賞者は過去１０年間だれもいない
  - 8年ぶり
  - ACMの正式な賞になってからは史上初

等々と歴史的に重要な出来事みたいです

以後，SCで話した内容
を時間の許す限りお話します.

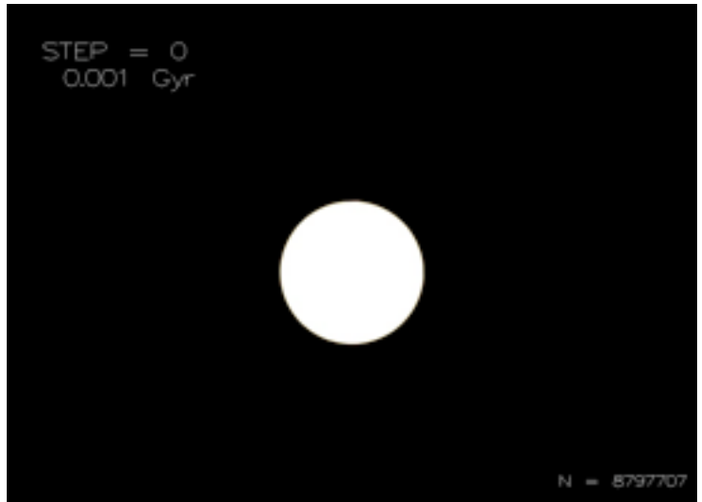
時間がなくなれば，申し訳ありませんが
SCのGB発表へ参加された方へお尋ね下
さい.

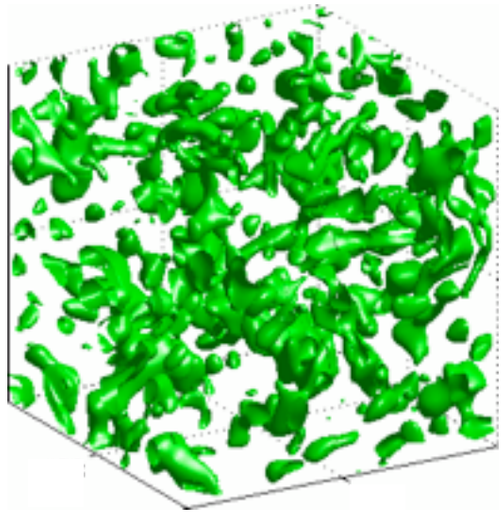
Summary



+

- Achieved Price/Performance
  - $ 7.2 / Gflops
  - 15 times better than 2006’ GB Finalist





Challenges

- To get a high efficiency on GPUs for hierarchical O(NlogN) or O(N) method (not on brute $O(N^2)$ method)

  Treecode, FMM

- using large amount of GPUs 547,200=240*3*760 FP units

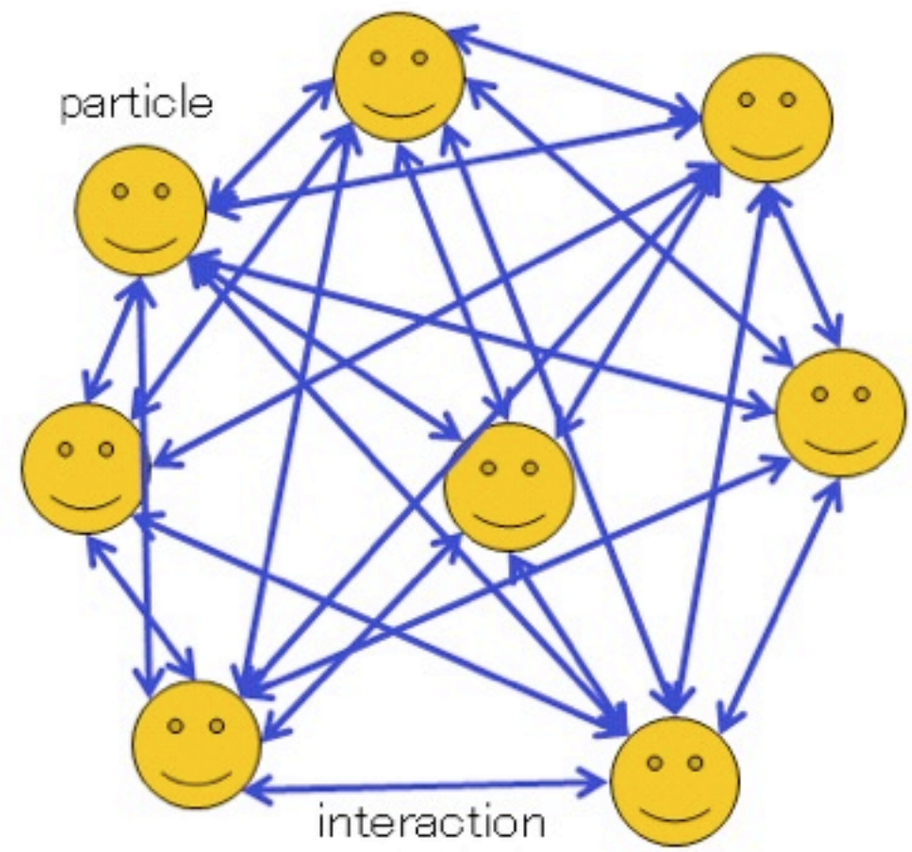
  Multiple walk

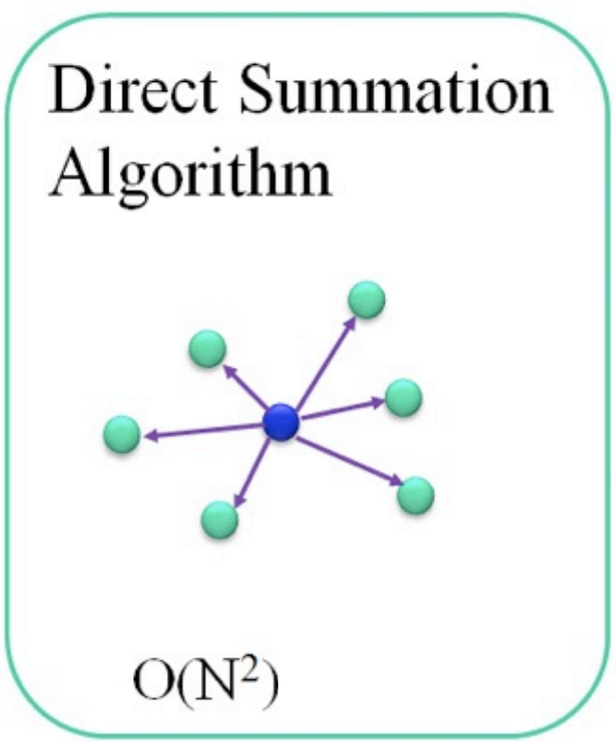
- To get a good scalability on commodity network (GbE)

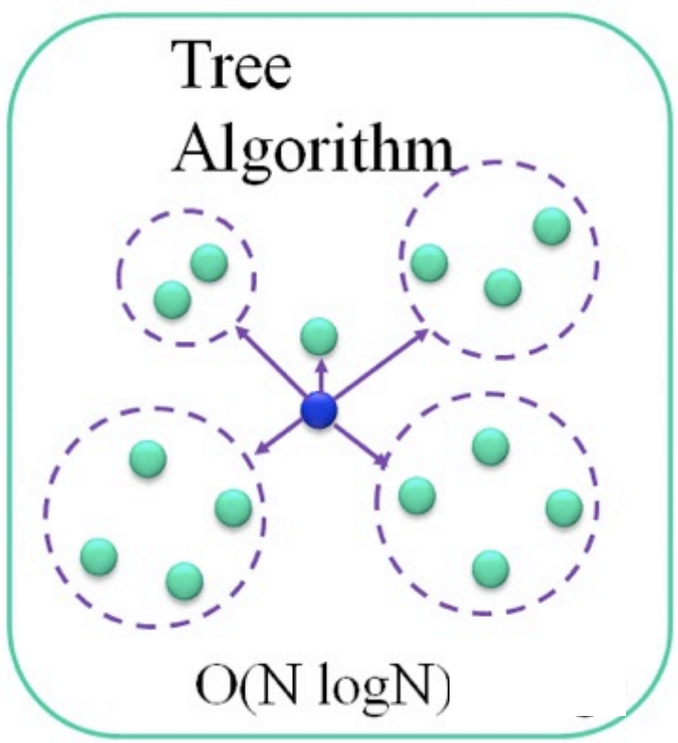
  Delegated Alltoallv

N-body simulation

- Particles interact with each other
  - Stars, Galaxies, Atoms, etc.
- Computational cost
  - Direct sum - $O(N^2)$
  - Tree algorithm - O(NlogN)
  - Fast Multipole Method - O(N)







Applications of N-body

Poisson

$$\nabla^2 u = -f$$

Astrophysics

$$\nabla^2 \phi = 4\pi G M$$

Electrostatics

$$\nabla^2 \phi = -\frac{q}{\epsilon_0}$$

Fluid Mechanics

$$\nabla^2 p = -\nabla \cdot \{\mathbf{u} \cdot (\nabla \mathbf{u})\}$$

$$\nabla^2 \mathbf{u} = -\nabla \times \boldsymbol{\omega}$$

Helmholtz

$$\nabla^2 u + k^2 u = -f$$

Acoustics

$$\frac{1}{c^2}\frac{\partial^2 \phi}{\partial t^2} = \nabla^2 \phi$$

Electromagnetics

$$\mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = \nabla^2 \mathbf{E}$$

$$\mu_0 \epsilon_0 \frac{\partial^2 \mathbf{H}}{\partial t^2} = \nabla^2 \mathbf{H}$$

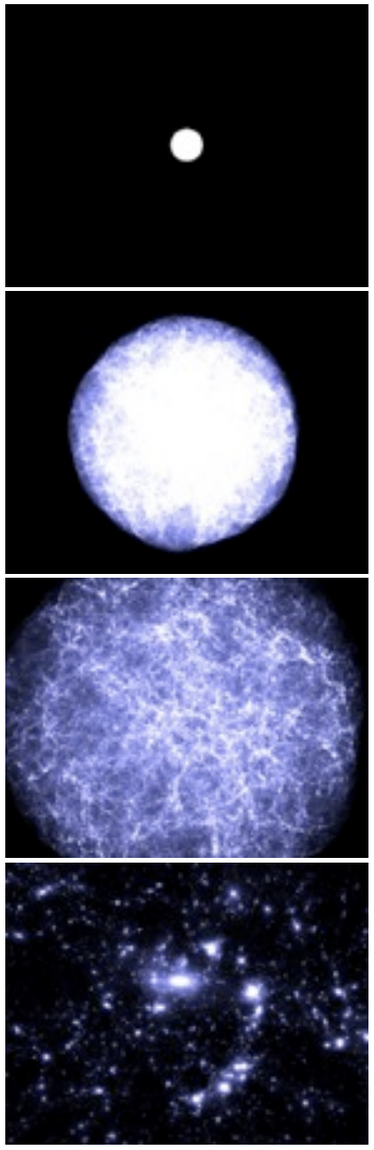
Quantum Mechanics

$$\frac{1}{c^2}\frac{\partial^2 \phi}{\partial t^2} = \nabla^2 \phi - \frac{m^2 c^2}{\hbar^2}\phi$$

Run-1: Astrophysics

- Standard cold dark-matter cosmological simulation

$$\frac{d^2\vec{r}_i}{dt^2} = \sum_{j \neq i} -\frac{G m_j \vec{r}_{ij}}{r_{ij}^3}$$



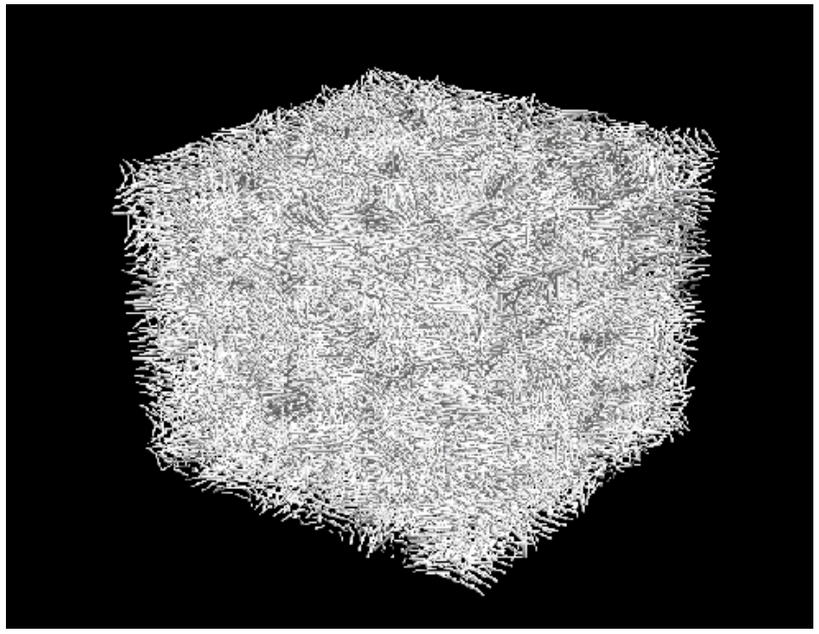
- Barnes & Hut tree method
- 760 MPI processes
  - 19 x 8 x 5 decomposition
- 4.5 Giga particles
  - 4,497,841,079 particles

Run-2: CFD

- Isotropic Turbulence vortex particle simulation

$$\frac{d\vec{x}_i}{dt} = \sum_{j \neq i} \frac{\boldsymbol{\alpha}_j \times \vec{r}_{ij}}{4\pi r_{ij}^3} g_\sigma$$

- Greengard & Rokhlin FMM
- 512 MPI processes
- 512x512x512 grids
  - 134,217,728 particles



Past Approaches (treecode in Gordon Bell)

- Massively-parallel system
  - Warren et al. (1997, winner)
- Dedicated hardware, single node
  - Kawai et al. (1999, Winner)
- FPGA, single node
  - Kawai et al. (2006, finalist)
- GPU, massively-parallel
  - Hamada et al. (This work)

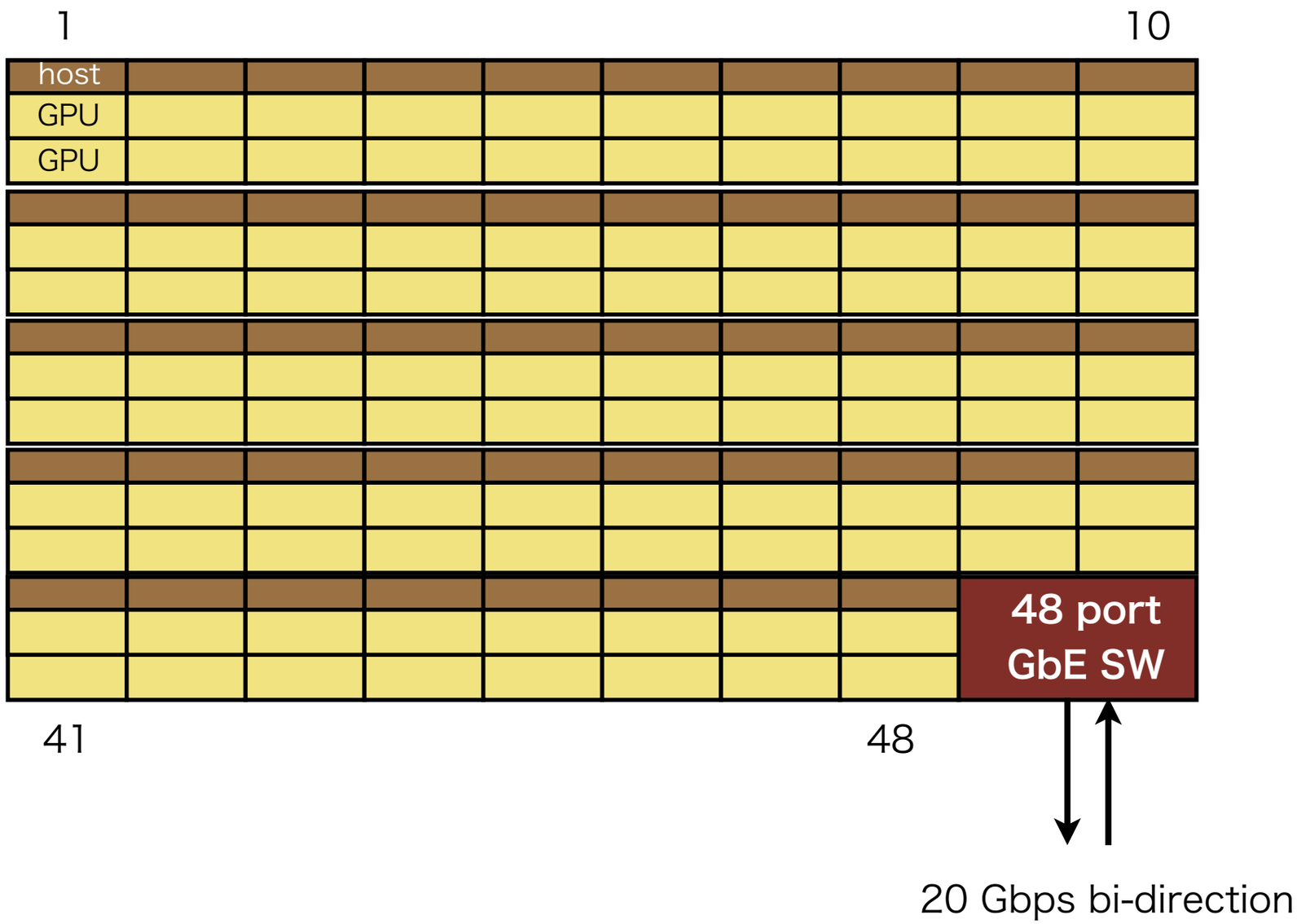
Hardware Configuration

- 190 nodes of GPU cluster
  - 760 NVIDIA GT200 chips
- Host
  - Core i7 920
  - 12GB DDR3
  - MSI X58 pro-E
- GPU
  - Dual GeForce GTX 295 (single PCB) per node
- Network
  - 48 port Netgear GS748TS x 4

Price

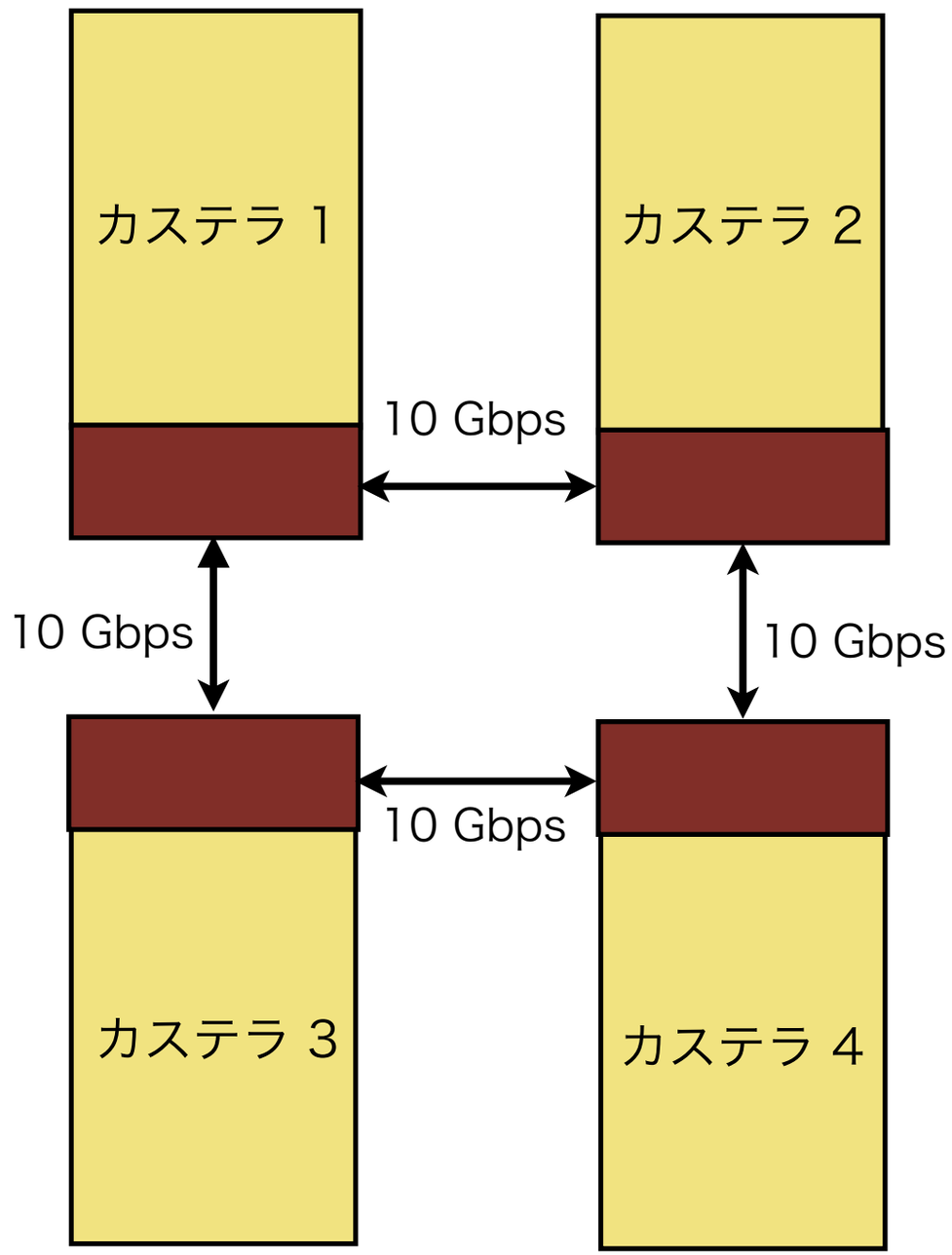
- Host: $205,805
- GPU: $201,446
- Network: $6,933
- Total: $414,185

カステラ(CASTERA: CAScadable TERA flops unit)



カステラ 1個分の写真

System Configuration



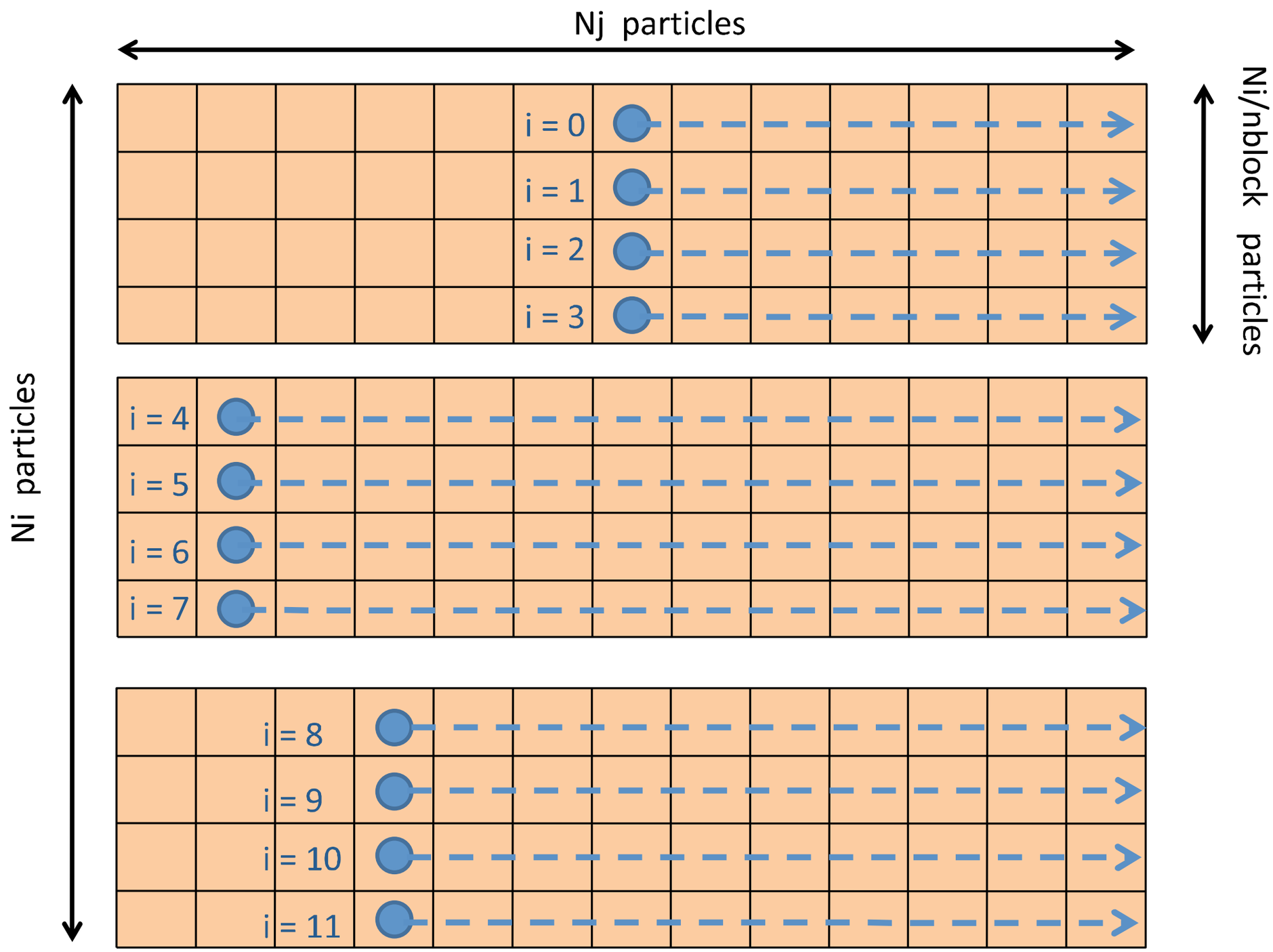
Malfunctioned GPUs

- Some GPUs crashed during calculations or returned wrong results

Parallel implementation

- Domain decomposition
  - 3-dimensional recursive multi-section
- Force from other domain
  - Exchanging LET (local essential tree)
- FMM also uses similar method

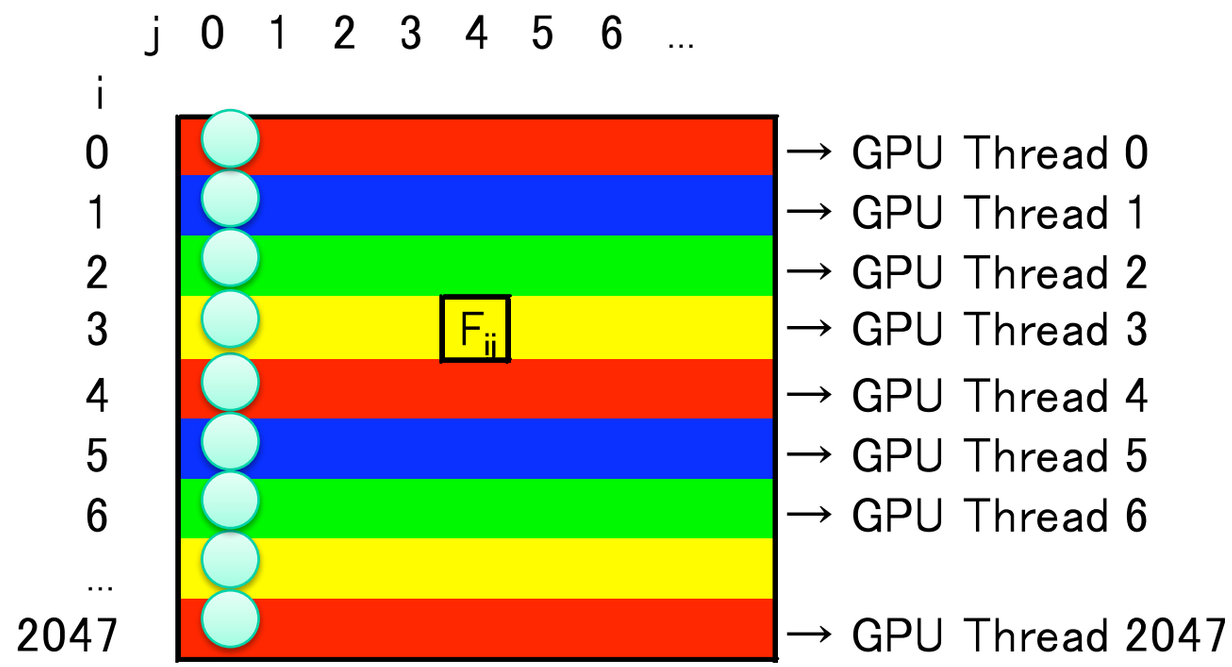
GPU implementation for $O(N^2)$



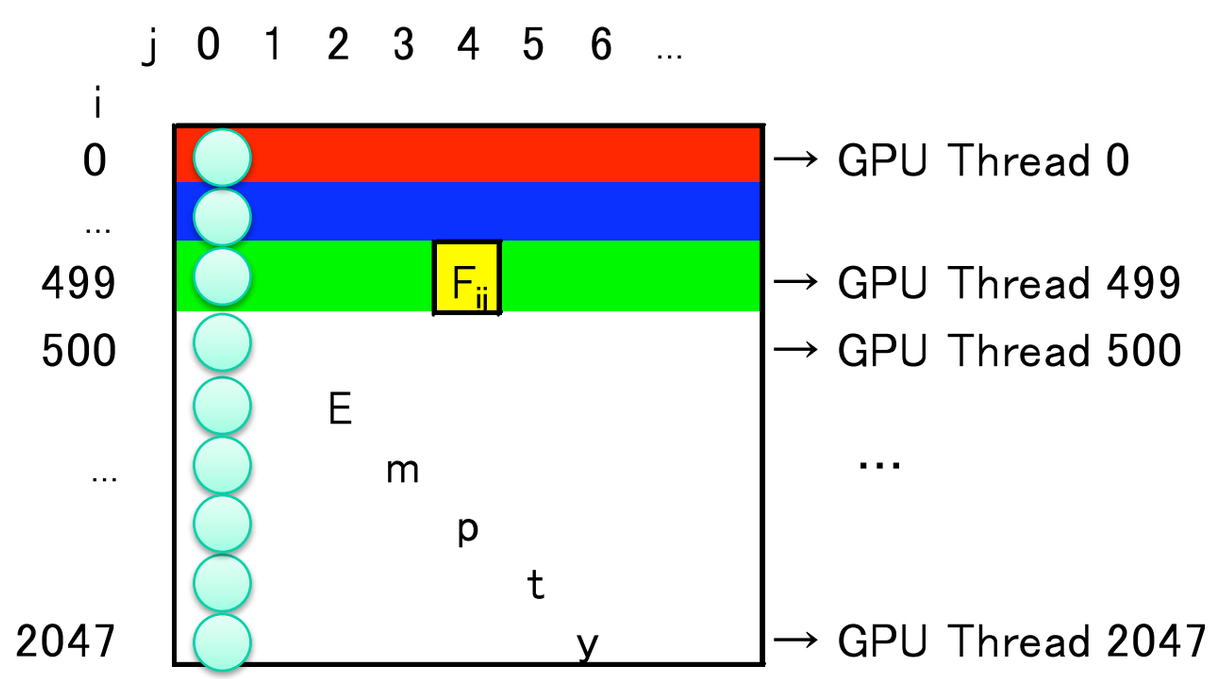
Hamada et al 2007, Belleman 2007, Nyland 2007, etc

Naive treecode implementation

Direct sum.

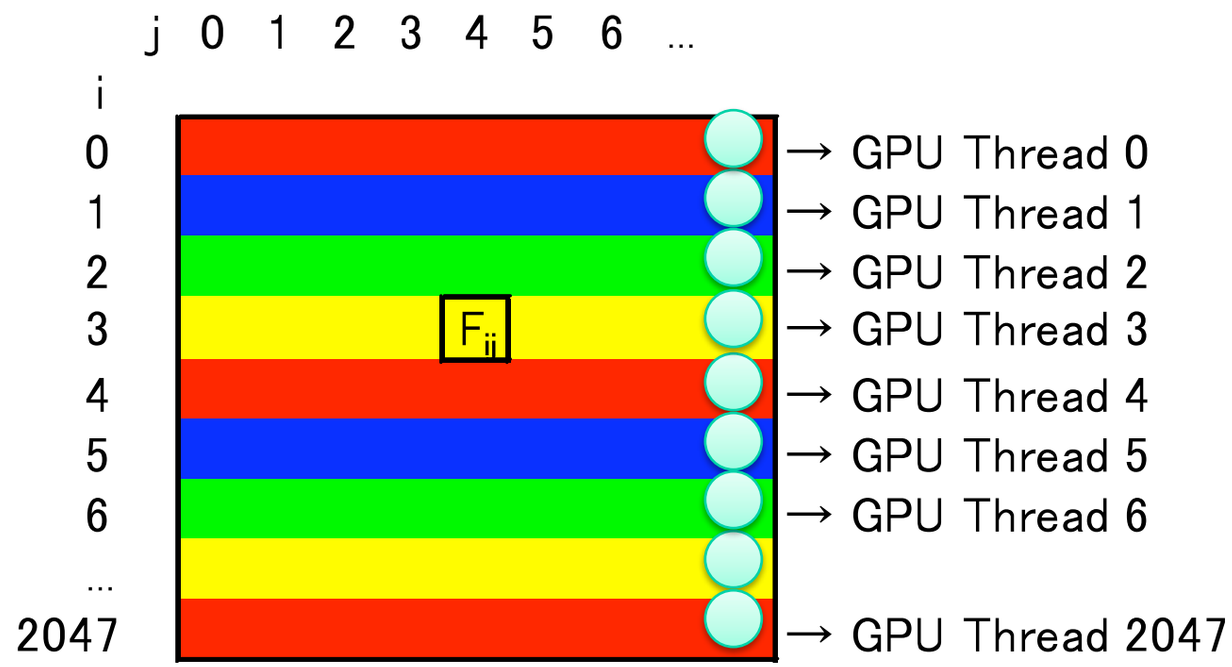


Treecode

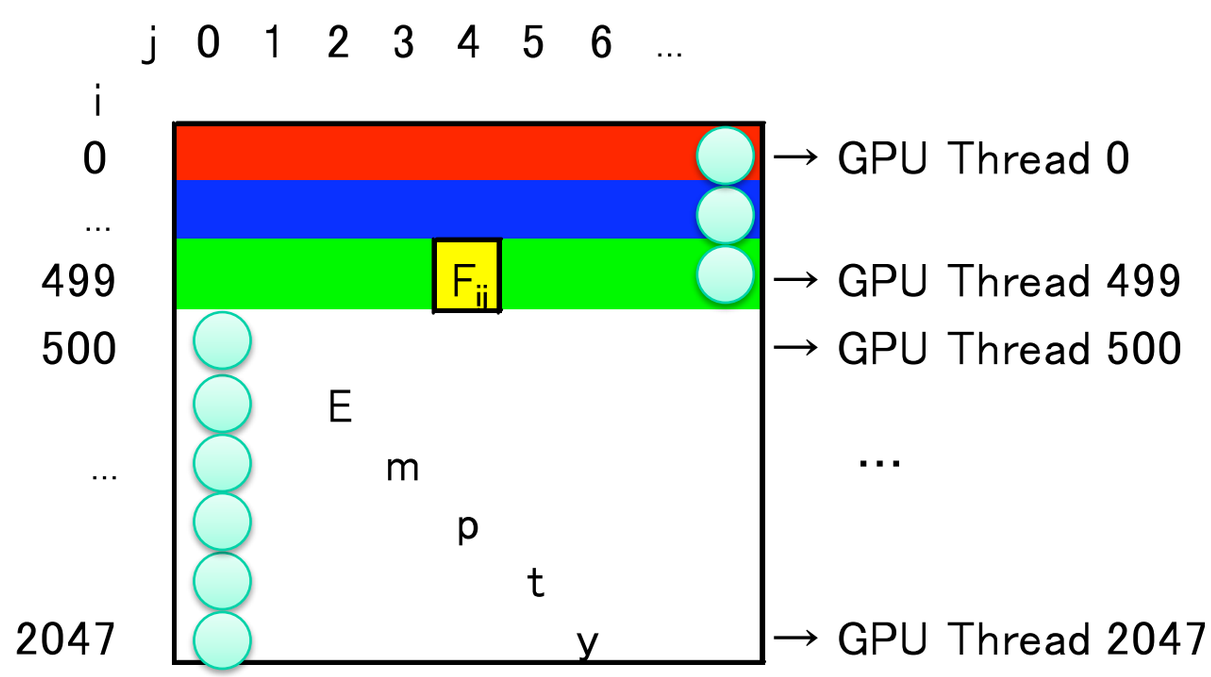


Naive treecode implementation

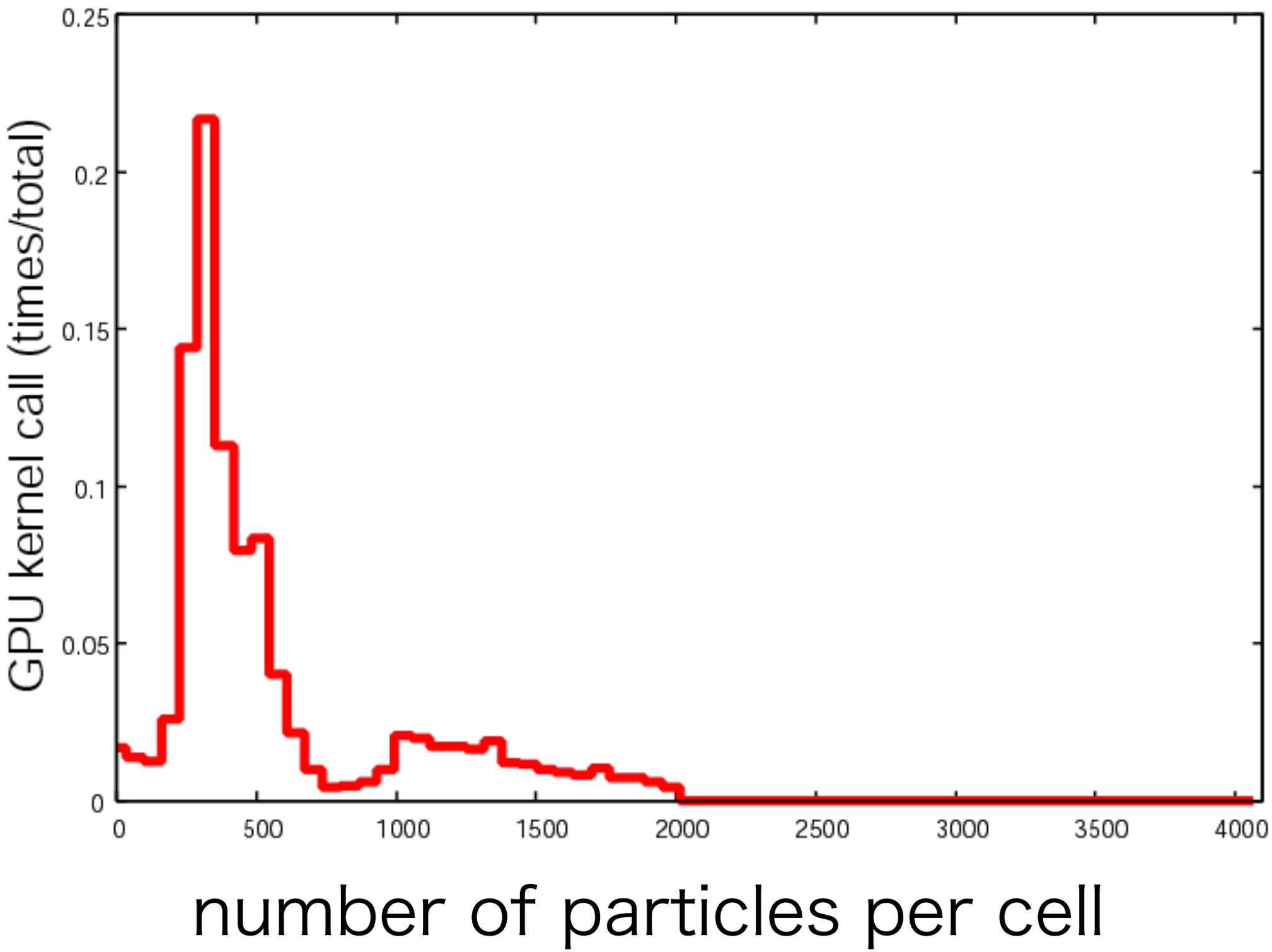
Direct sum.



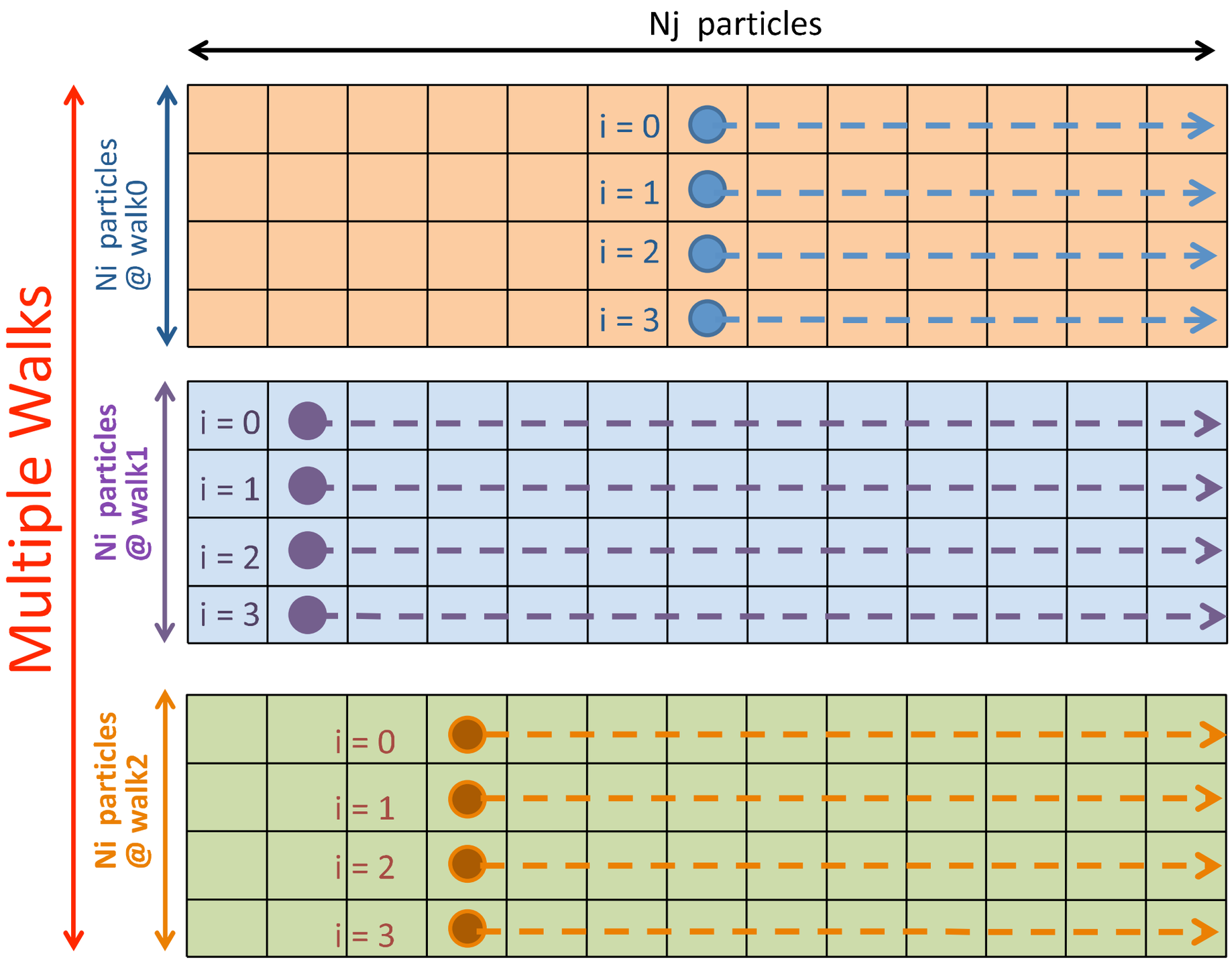
Treecode



Number of working threads



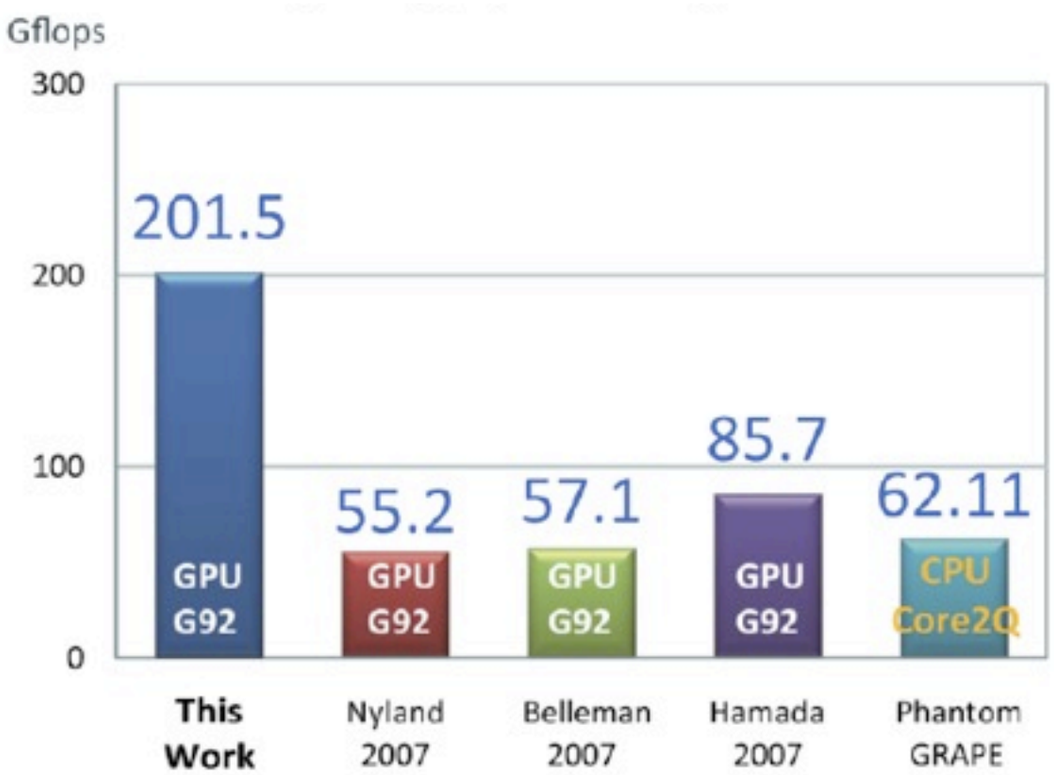
Multiple walks



No Reduction phase

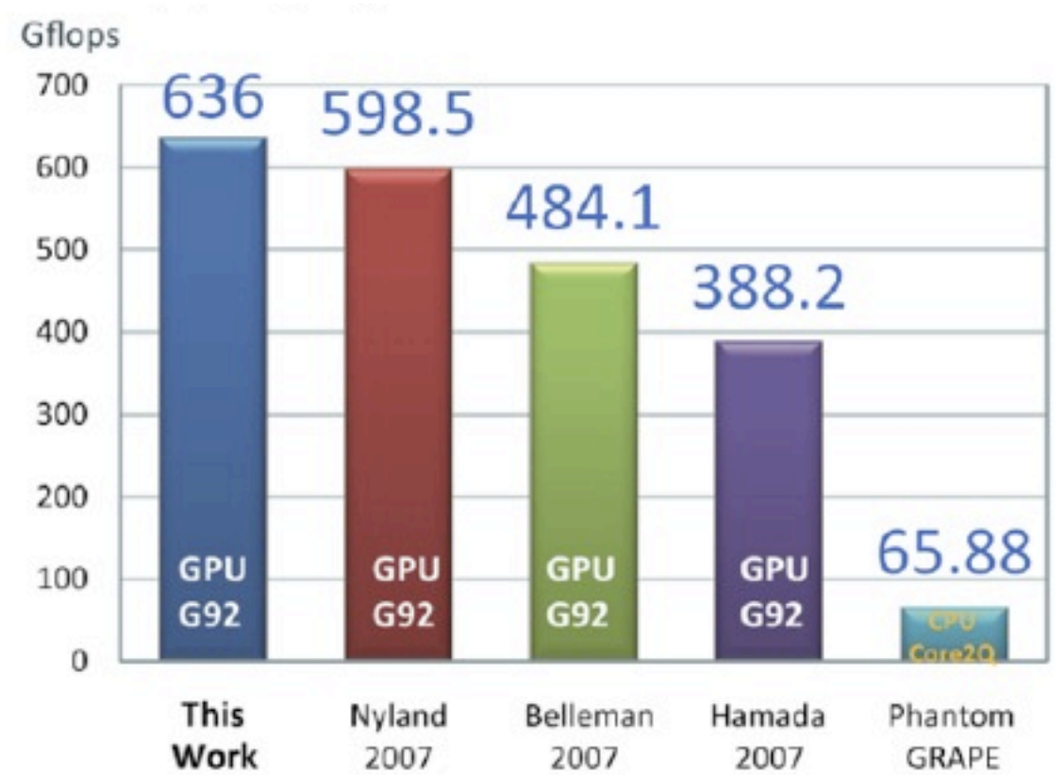
Flops on GPUs (single node)

treecode



multiple walks

brute force

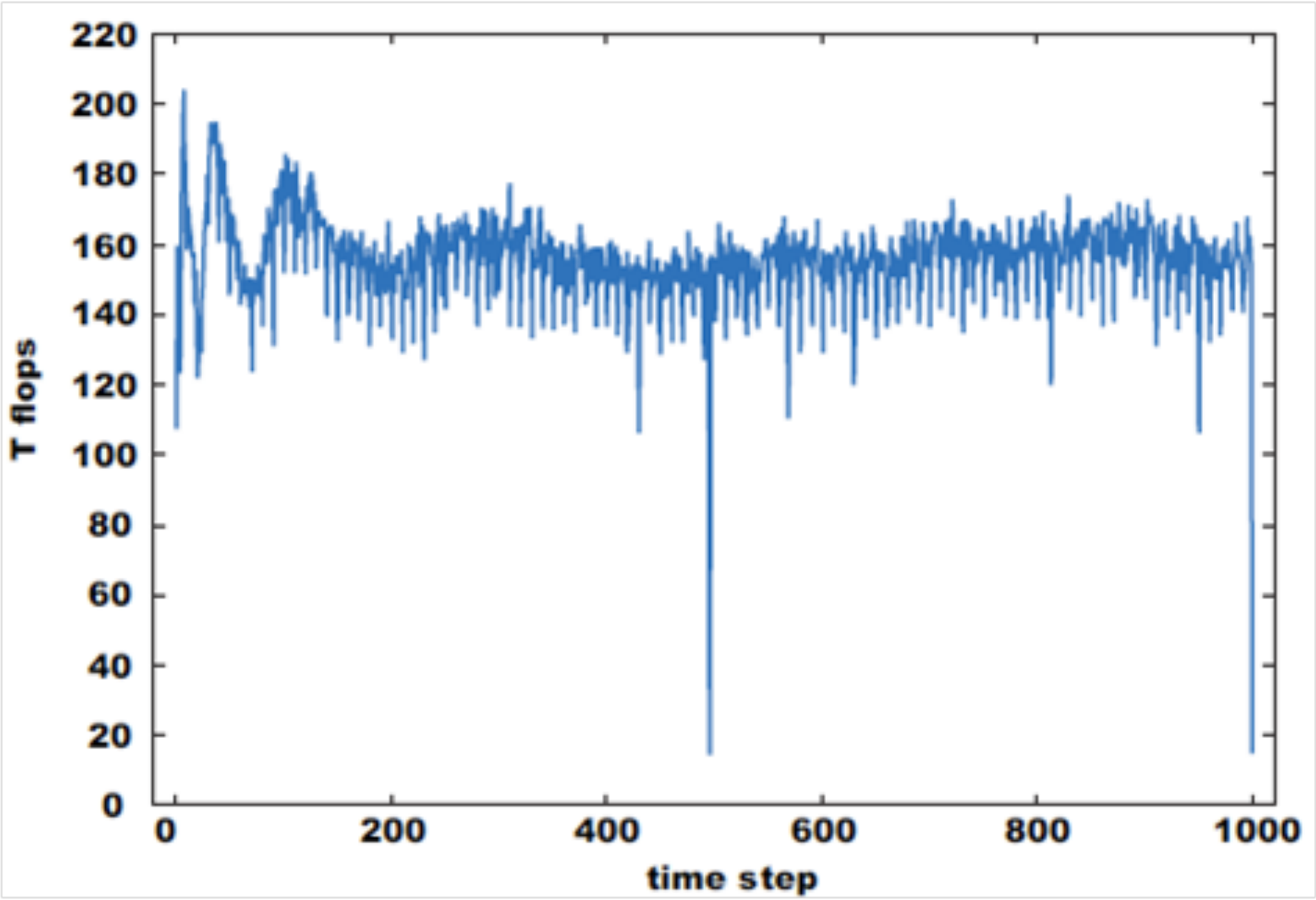


Parallel Performance

- Total # of particles
  - 4,497,841,079 particles
- Total # of interactions
  - 4.28e+16
- Wall-clock time
  - 10294 sec
  - time/step : 10.3 sec
- Particle advances per second
  - 436,950,861 particles step/sec

Raw (uncorrected) performance

- 158 Tflops average
  - Flop count per interaction : 38



Performance correction I

- What is “correct” performance comparable with conventional machines?
- 158 Tflops (raw/uncorrected)
- Correction factor : 0.55
  - (list length for CPU) / (list length for GPU)
  - 5232 / 9496 = 0.55
- Corrected
  - 158 * 0.55 = 87 Tflops

ng	nlist	optimal for
4	5232	CPU(scalar)
102	6546	CPU(SSE)
388	9496	GPU

Performance correction II

- 83.1 Tflops (using 20 counting)
- Correction factor : 0.69
  - (list length for CPU) / (list length for GPU)
  - 6546 / 9496 = 0.69
- Corrected
  - 83.1 * 0.69 = 57.3 Tflops

ng	nlist	optimal for
4	5232	CPU(scalar)
102	6546	CPU(SSE)
388	9496	GPU

Cost performance

- Total cost : $414,185
- Performance : 57.3 Tflops

- Performance/Price : 138 Mflops/$
- Price /Performance: $ 7.2/Gflops

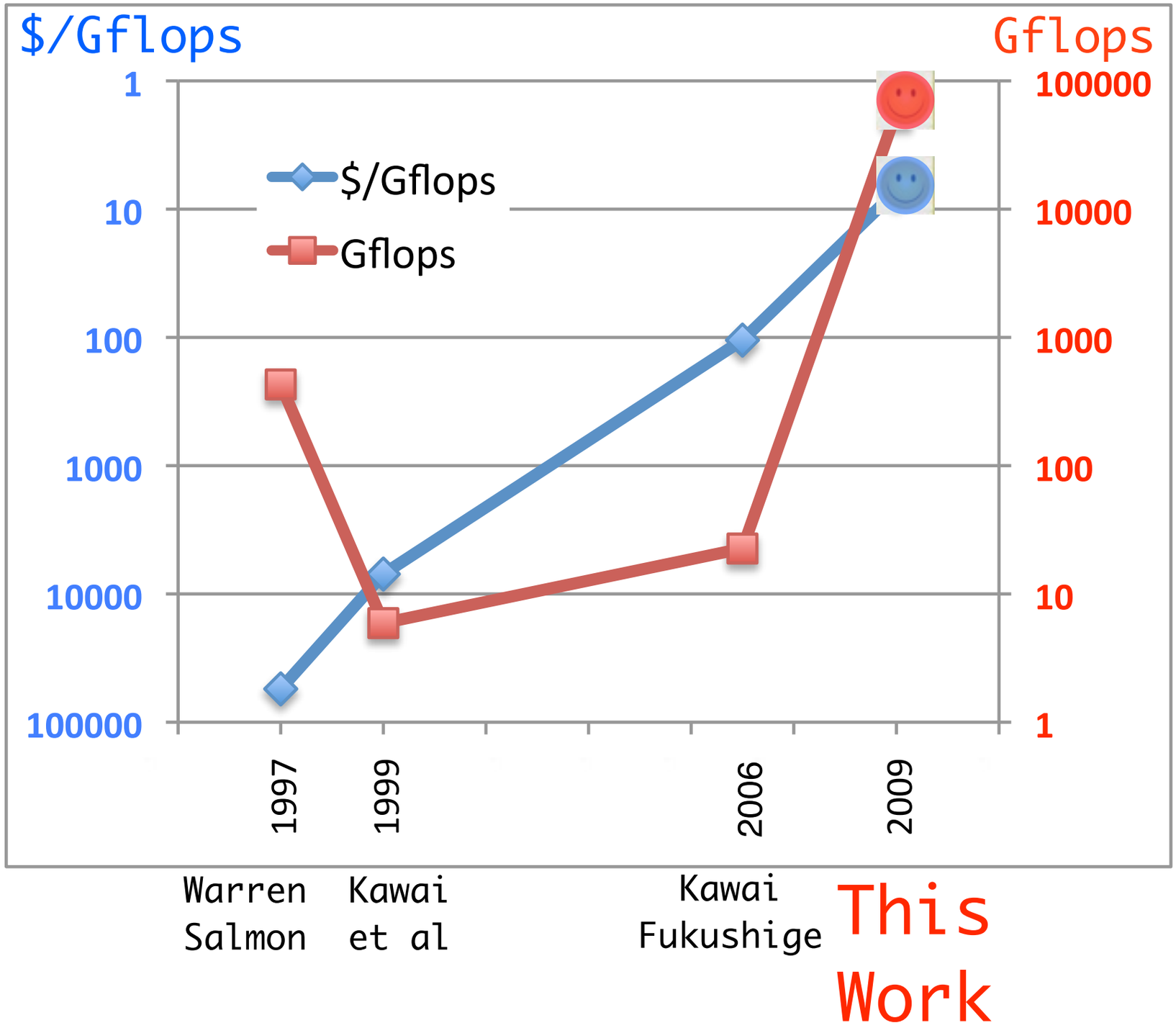
Performance per Watt

- Energy consumption : 286.23 kWh
- Computation time : 10,294 sec

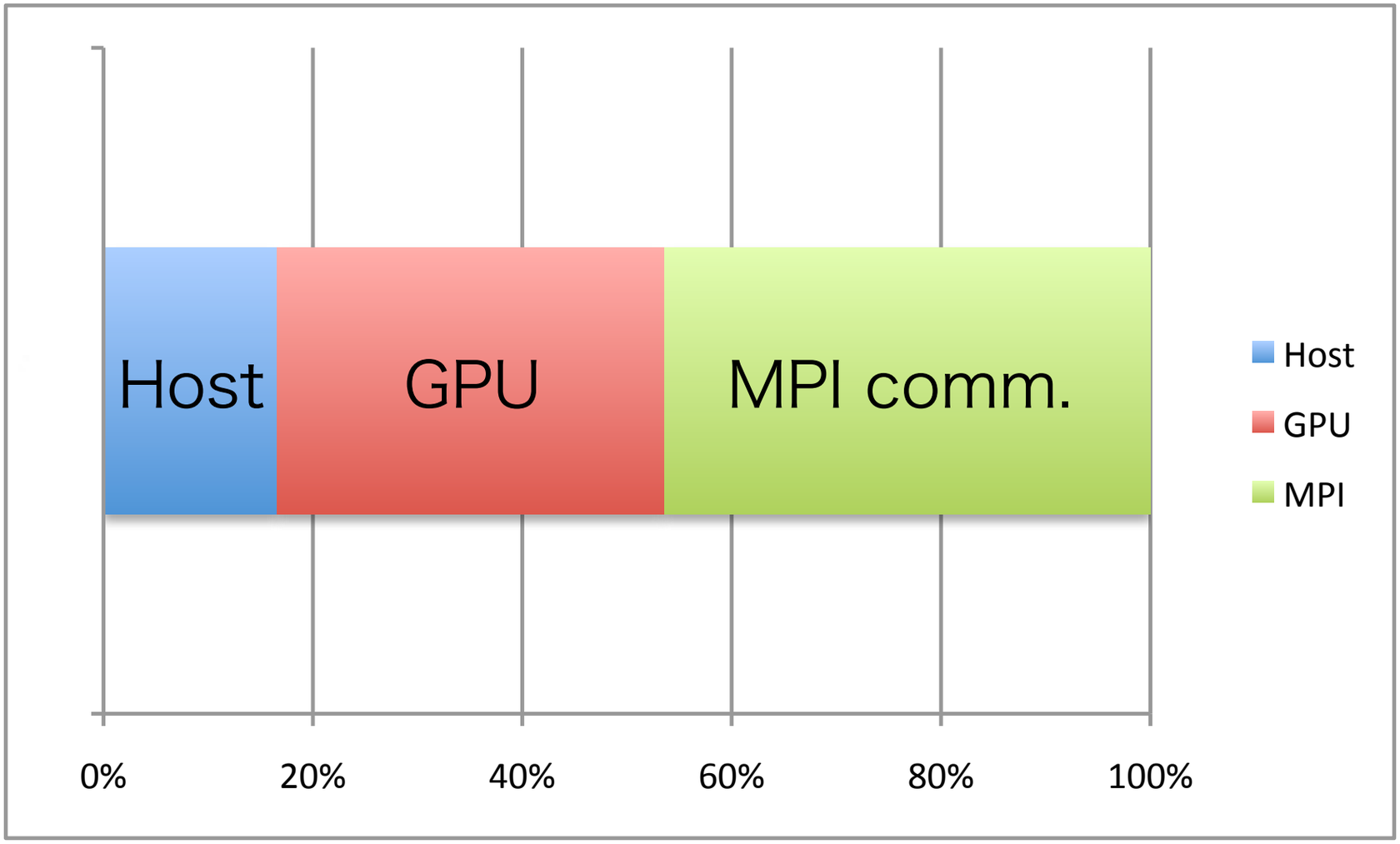
- Power consumption: 10294/286.23 = 100.1 kW
- Performance/Watt : 573 Mflops/W

the No.1 of Green 500(Jun 2006) : 536 Mflops/W

Astrophysical Simulation in GB History



Profile of 760 GPU run



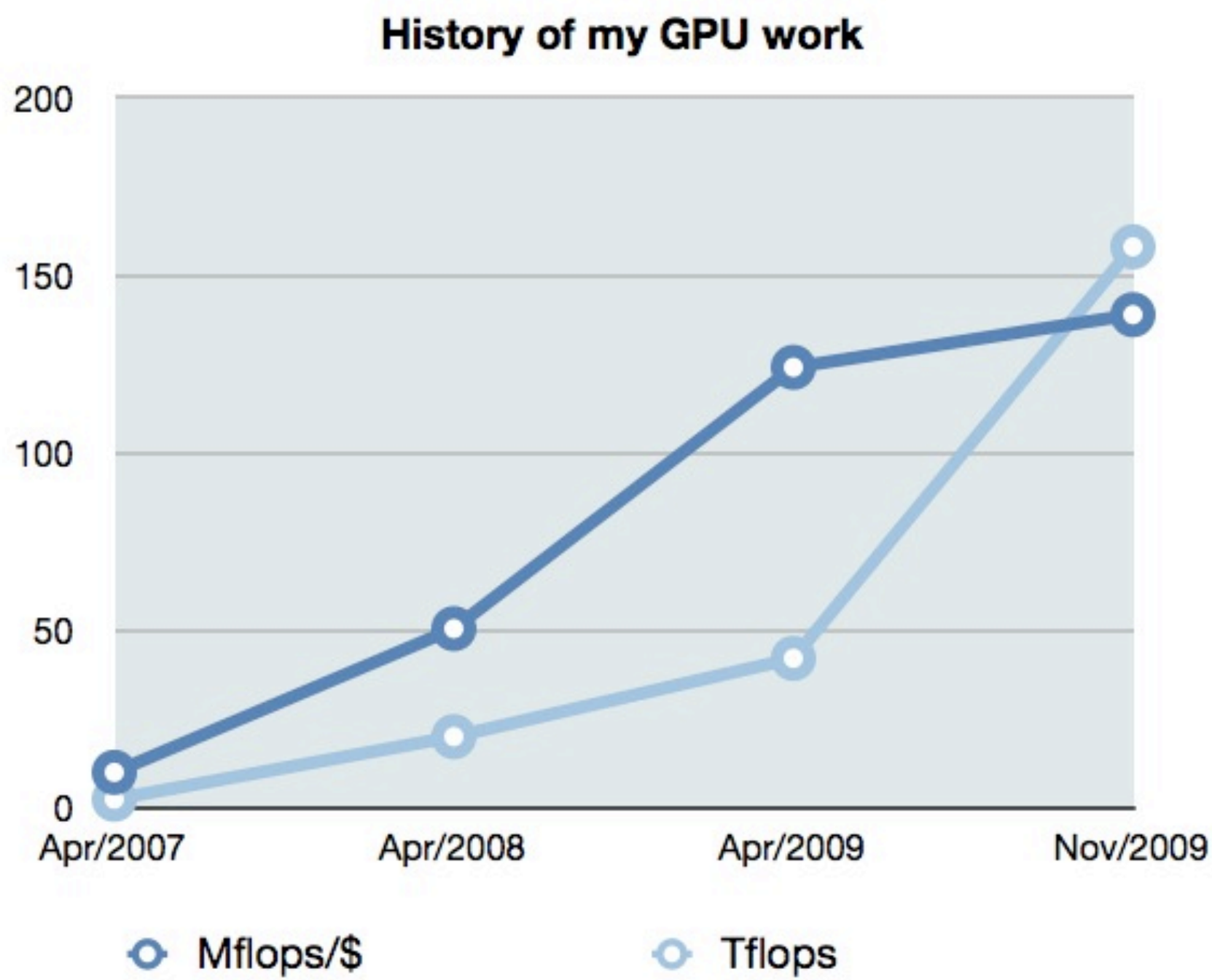
Comparison to other works

- GreeM - Fine tuned treePM for X86(SSE) cluter/MPP
  - Ishiyama, Fukushige, Makino. 2009
  - Cray XT4
  - Particle advanced per second (per core) : 43,000

- Our treecode run
  - Particle advanced per second (190 node) : 436,950,861
  - Particle advanced per second (per node) : 2,299,741

- Our result is
  - equivalent to Cray XT4 10,000 cores !
  - 50 times faster (1 GPU node vs Cray XT4 1 core)
  - 10 times faster (190 GPU node vs Cray XT4 1k cores)

The History


History of my GPU work
200
150
100
50
0
Apr/2007
Apr/2008
Apr/2009
Nov/2009
Mflops/$
Tflops

Mar 2007

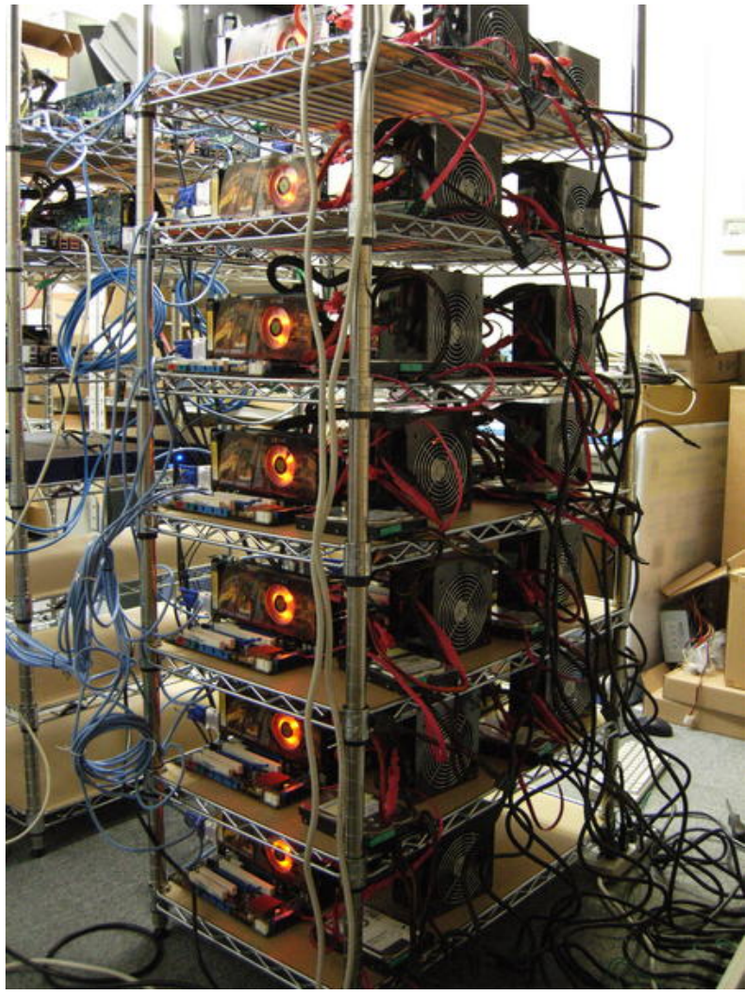
Host: Xeon 3.0 GHz .... x 40

GPU: GeForce 8800GTX .... x 40

~ 2 Tflops in cosmology sim.

History of our GPU cluster

Mar 2008



Host: Core2Quad 2.4 GHz .... x 32

GPU: GeForce 8800GT .... x 32

History of our GPU cluster

April 2008

Host: Core2Quad 2.4 GHz .... x 128
GPU: GeForce 8800GTS .... x 128

~ 20 Tflops in cosmology sim.

History of our GPU cluster

Nov 2008

Host: Core2Quad 2.4 GHz .... x 128
GPU: GeForce 8800GTS .... x 128
GPU: GeForce 9800GTX+ .... x 128

~ 40 Tflops in cosmology sim.

History of our GPU cluster

Aug 2009

Power supply 600A -> 2000A

Host: Core2Quad 2.4 GHz .... x 166
GPU: GeForce 9800GTX+ .... x 256
GPU: GeForce GTX295 .... x 33

Never give up

- challenge, challenge, challenge


出島
DEGIMA
cluster


出島
DEGIMA
cluster