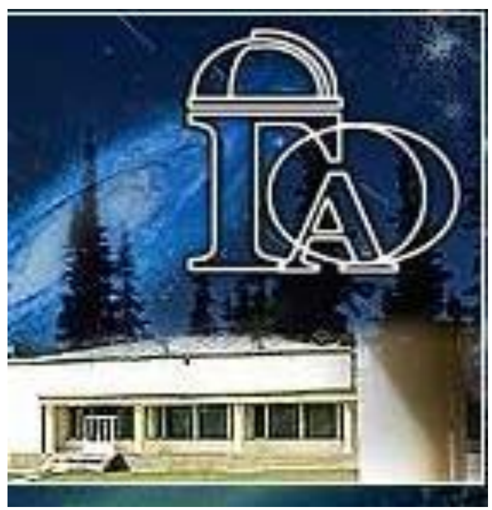
Current and future status of hardware testing with Kepler GPUs and Intel MICs.

Peter Berczik

NAOC, Chinese Academy of Sciences, Beijing
MAO, National Academy of Sciences of Ukraine, Kiev
ARI, Heidelberg University, Germany



5th China-Korea workshop on stellar dynamics and gravitational waves. Dec 12/13, 2013.

Collaborators:

-Rainer Spurzem (NAOC, Beijing; ARI, Heidelberg)

-Long Wang, Shiyan Zhong, Siyi Huang,
-Maxwell Xu Tsai, Gareth Kennedy, Shuo Li,
-Luca Naso, Changhua Li (NAOC + KIAA, Beijing)

-Alexander Veles, Igor Zinchenko (MAO, Kiev)

-Naohito Nakasato (Univ. Aizu, Japan)
-Keigo Nitadori (Univ. Tsukuba, Japan)

"Silk Road Project" – CAS, China
SFB 881 "The Milky Way System" – DFG, Germany

GPU cluster "laohu" - ZDYZ2008-2, NAOC, CAS
GPU cluster "kepler" - I/80 041-043 and I/81 396, VW
GPU cluster "golowood" – GRID/GPU, MAO, NASU

GPU Kepler Hardware

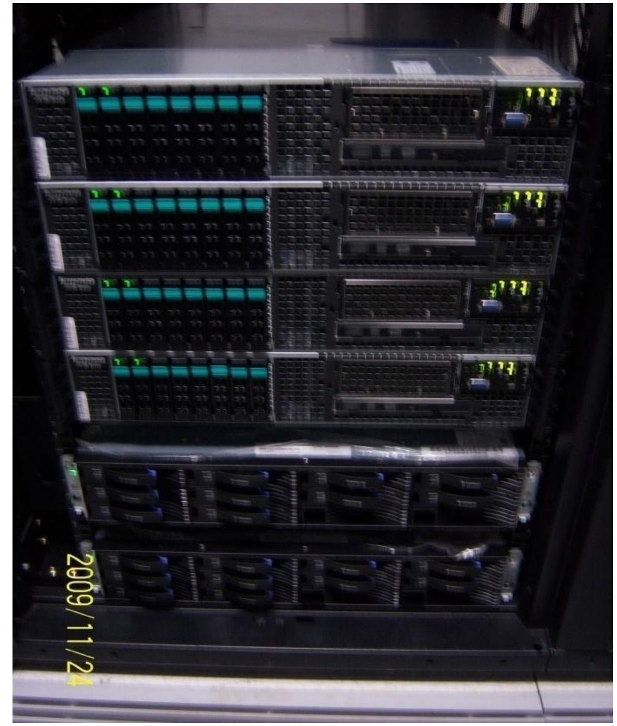
GF GTX TITAN - 2013.II.21

http://www.nvidia.com

2007: GeForce 8800 GTX, 128 SP, 768 MB
2008: GeForce 9800 GTX+, 128 SP, 512 MB
2009: GeForce GTX 280, 240 SP, 1 GB
2010: GeForce GTX 480, 480 SP, 1.5 GB
2011: GeForce GTX 580, 512 SP, 1.5 GB
2012: GeForce GTX 680, 1536 SP, 2 GB
2013: GeForce GTX TITAN, 2688 SP, 6 GB

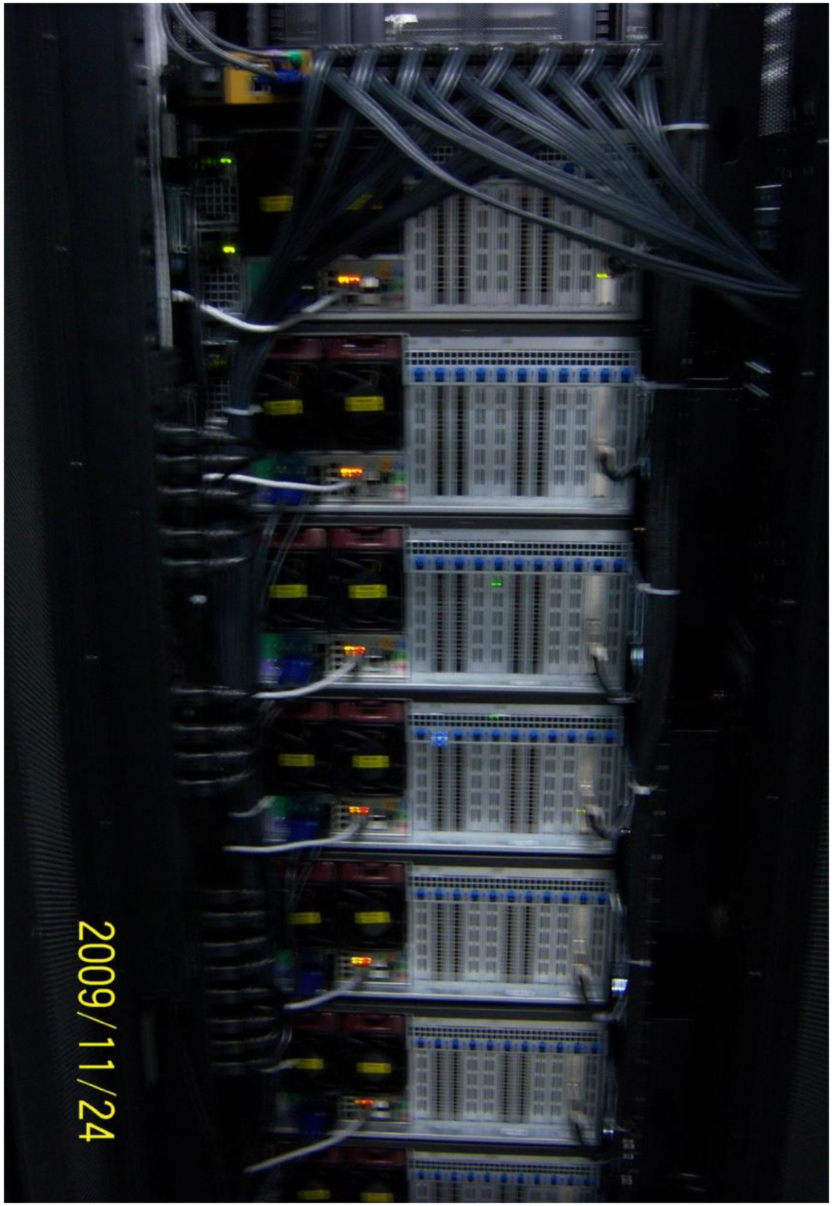
http://gpgpu.org

NAOC laohu cluster



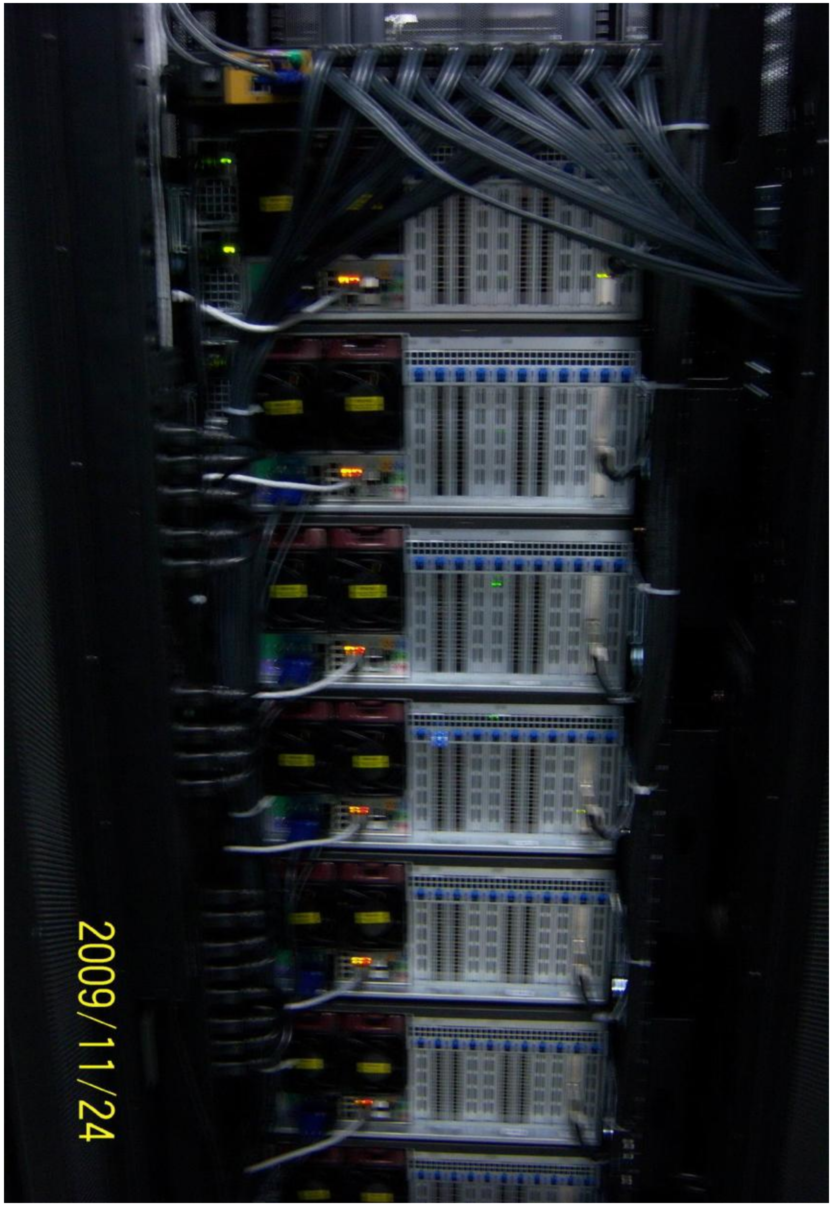
NAOC 85 node 2xC1060 cluster

- 85x 2,4 core Xeon E5520 2.33 GHz
- 85x RAM 24 GB = 2 TB
- 85x 2 = 170 TESLA C1060
- Speed: ~50 Tflops
- RAID-5: ~20 TB
- IB Network: DDR ~20 Gb/s
- Cost: ~5M CNY
- Funding: CAS NAOC



NAOC 59 K20+26 3xC1060 cluster

- 85x 2,4 core Xeon E5520 2.33 GHz
- 85x RAM 24 GB = 2 TB
- 59 = 59 KEPLER K20
- 26 x 3 = 78 TESLA C1060
- Speed: ~88 + 26 = 114 Tflops
- RAID-5: ~130 TB
- IB Network: DDR ~20 Gb/s
- Funding: CAS NAOC



Hydra GPU cluster.

Hydra GPU cluster

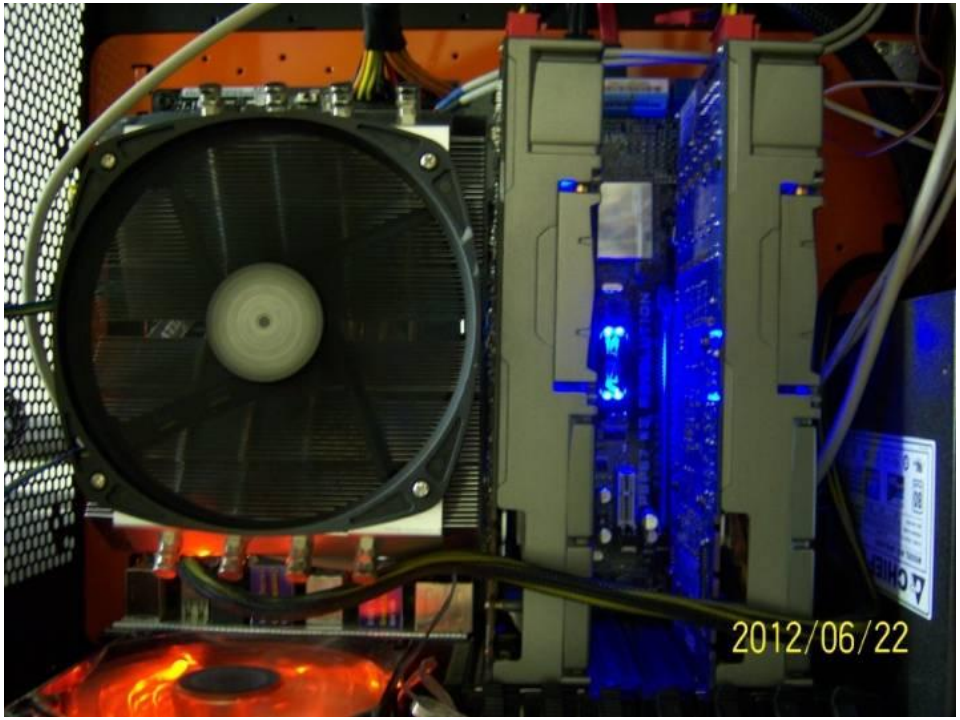
8 nodes = 8 x 4 = 32 CPU cores (@ 3.3 GHz)

8 x 16 GB = 128 GB RAM CPU memory

16 GPU GF 570 = 16 x 480 ~ 7.7k GPU threads

8 x 1.3 GB ~ 10 GB GPU device memory

since mid. 2012 operated.



Kepler GPU cluster.

Kepler GPU cluster

12 nodes = 12 x 16 = 192 CPU cores (@ 2 GHz)

12 x 64 GB = 768 GB RAM CPU memory

12 GPUs K20m = 12 x 2496 ~ 30k GPU threads

12 x 4.8 GB ~ 57 GB GPU device memory

4 x Xilinx Virtex-6 FPGA (ML 605)

since beg. 2013 operated.

φGPU current usage/results

-"Up to 700k GPU cores, Kepler, and the Exascale future for simulations of star clusters around black holes".
-10/2013, HPC-UA.
-http://adsabs.harvard.edu/abs/2013hpc..conf...52B

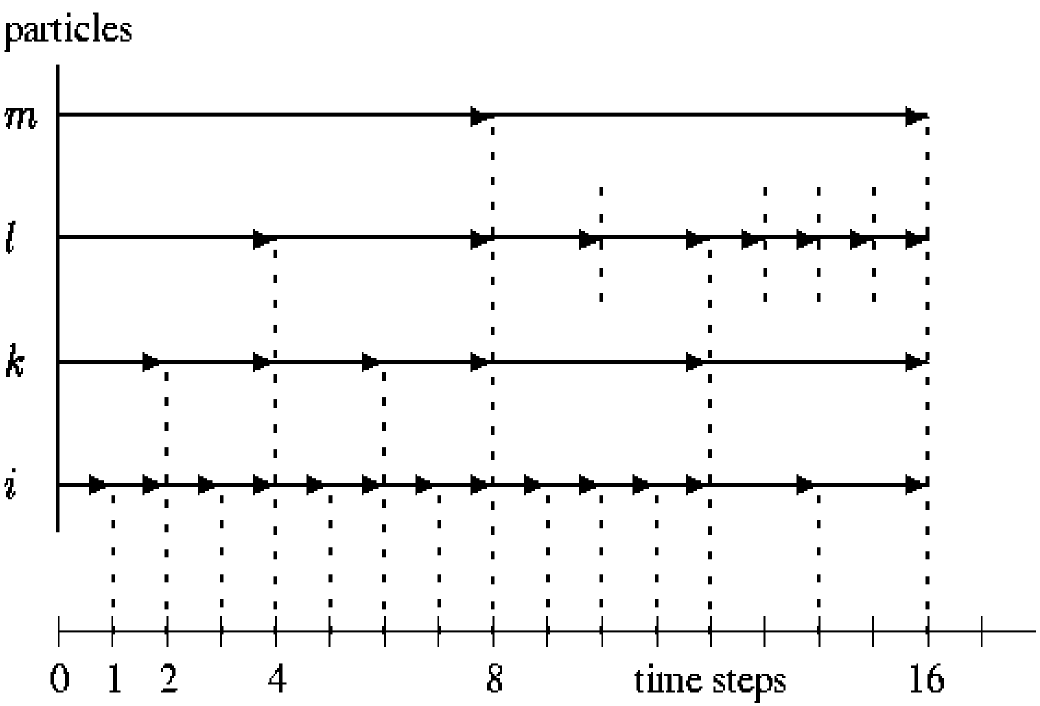
-"Supermassive Black Hole Binaries in High Performance Massively Parallel Direct N-body Simulations on Large GPU Clusters".
07/2012, ASP Conf. Proc.
-http://adsabs.harvard.edu/abs/2012ASPC..453..223S

-"High performance massively parallel direct N-body simulations on large GPU clusters".
-10/2011, HPC-UA.
-http://adsabs.harvard.edu/abs/2011hpc..conf....8B

Our own φGRAPE/GPU N-body code

Harfst et al, NewA, 12, 357 (2007) [astro-ph/0608125]

Hierarchical Individual Block Time Steps



$$\Delta t = \sqrt{\eta \frac{|\vec{a}||\vec{a}^{(2)}| + |\dot{\vec{a}}|^2}{|\dot{\vec{a}}||\vec{a}^{(3)}| + |\vec{a}^{(2)}|^2}} .$$

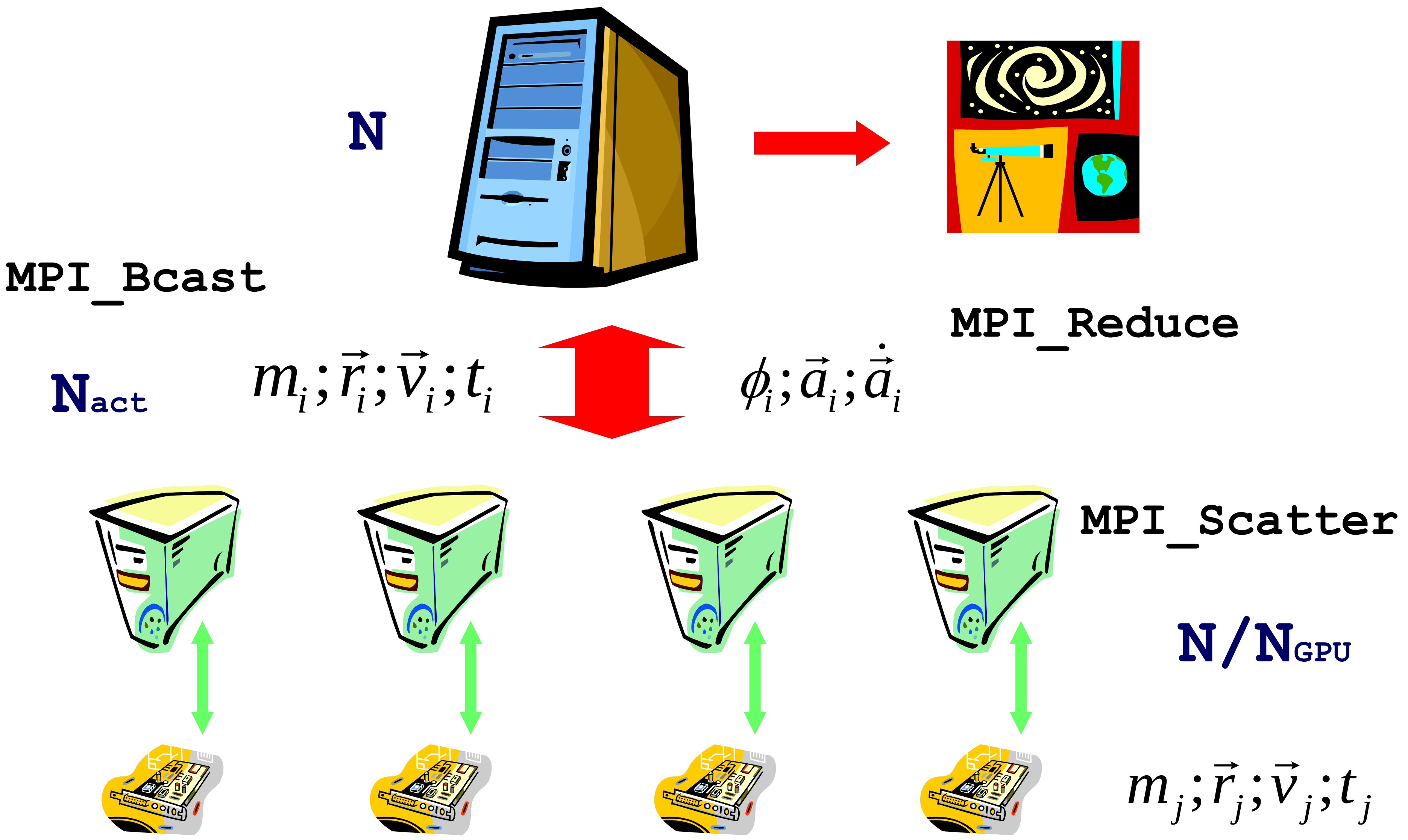
4th order Hermite scheme

$$\frac{d^2 \vec{r}_i}{dt^2} = \vec{a}_i$$

ftp://ftp.mao.kiev.ua/pub/berczik/phi-GRAPE/

ftp://ftp.mao.kiev.ua/pub/berczik/phi-GPU/

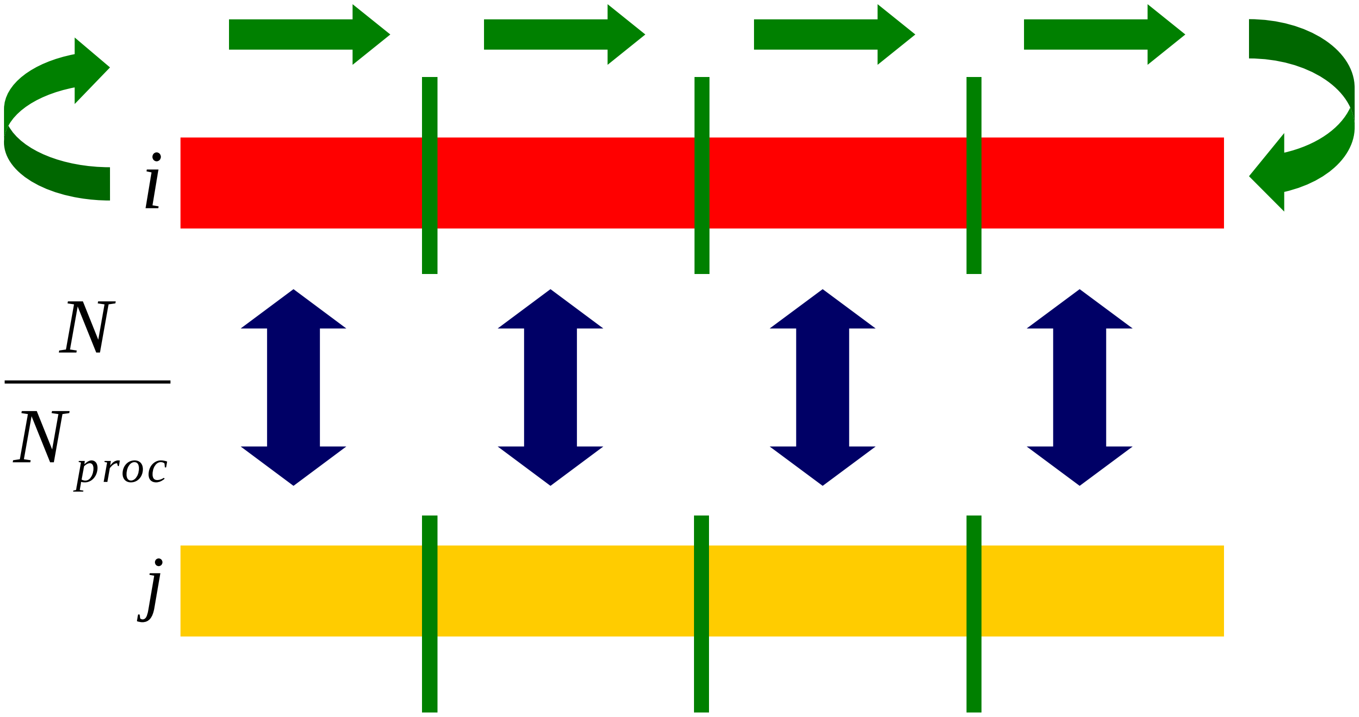
Parallel code on the cluster



Basic idea of any parallel N-body code

$$i, j - particle$$

Some communication scheme...



$$N_{loc} = \frac{N}{N_{proc}}$$

Award 2011

INTERNATIONAL SUPERCOMPUTING CONFERENCE

TITLE

Astrophysical Particle Simulations with Large Custom GPU Clusters on Three Continents

AUTHORS

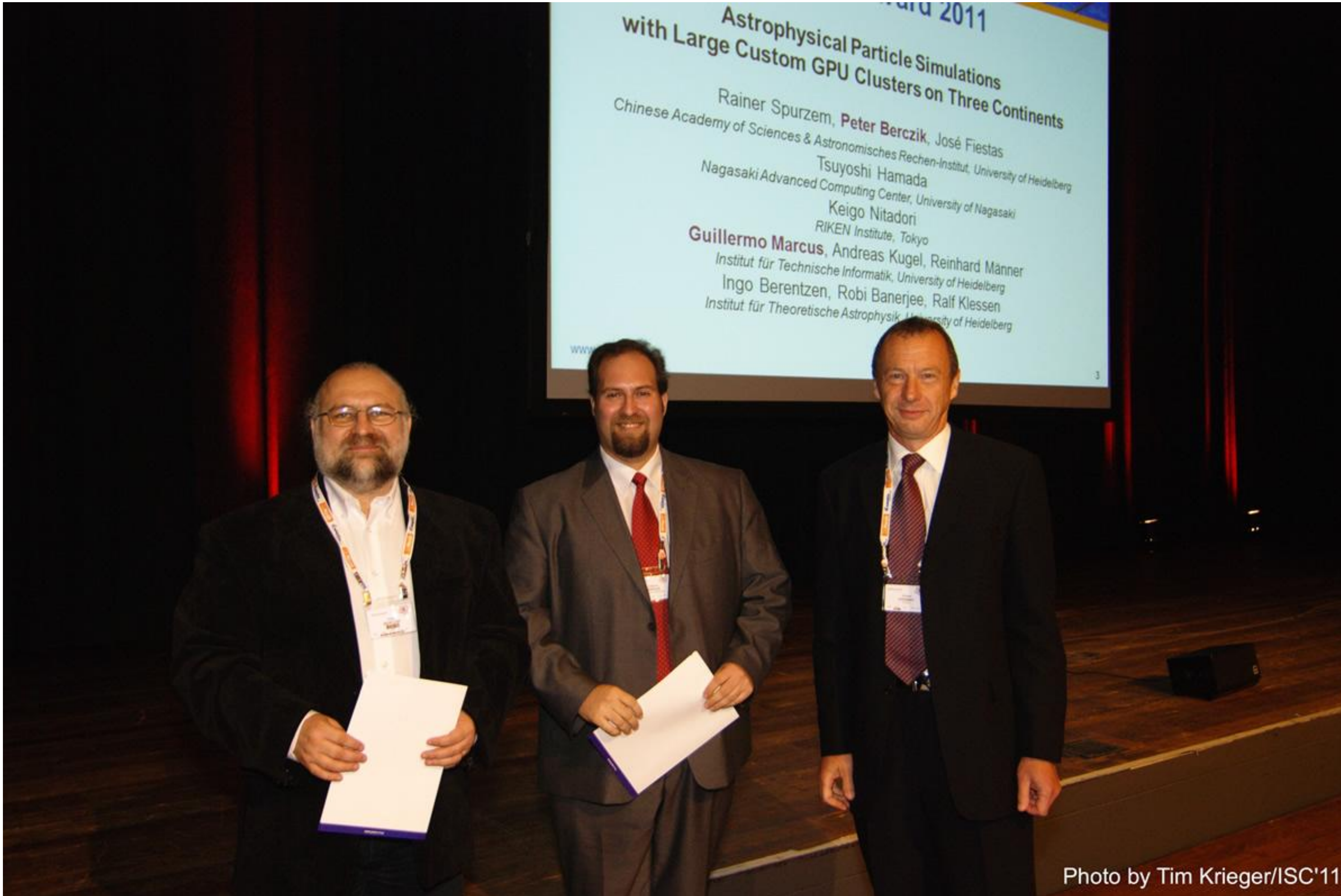
R. Spurzem, P. Berczik, T. Hamada, K. Nitadori, G. Marcus, A. Kugel, R. Manner, I. Berentzen, J. Fiestas, R. Banerjee and R. Klessen

AFFILIATION

Chinese Academy of Sciences & University of Heidelberg

We congratulate

Hamburg, June 20, 2011


Astrophysical Particle Simulations
with Large Custom GPU Clusters on Three Continents
Rainer Spurzem, Peter Berczik, José Fiestas
Chinese Academy of Sciences & Astronomisches Rechen-Institut, University of Heidelberg
Tsuyoshi Hamada
Nagasaki Advanced Computing Center, University of Nagasaki
Keigo Nitadori
RIKEN Institute, Tokyo
Guillermo Marcus, Andreas Kugel, Reinhard Männer
Institut für Technische Informatik, University of Heidelberg
Ingo Berentzen, Robi Banerjee, Ralf Klessen
Institut für Theoretische Astrophysik
Photo by Tim Krieger/ISC'11

Parallel code on the cluster

$$\Delta T_{total} = \underline{\Delta T_{host}} + \underline{\Delta T_{GPU}} + \underline{\Delta T_{comm}} + \underline{\Delta T_{MPI}}$$

$$\Delta T_{MPI} \propto (\tau_{lat} + N_{act}) \cdot \log(N_{GPU})$$

$$\Delta T_{GPU} \propto N \cdot \frac{N_{act}}{N_{GPU}}$$

Parallel code on the cluster

• active part. scan:	$O(N_{act} \log(N_{act}))$	$+T_{host}$
• all part. prediction:	$O(N/N_{GPU})$	$+T_{host}$
• ''j'' part. send. to GPU:	$O(N/N_{GPU})$	$+T_{comm}$
• ''i'' part. send. to GPU:	$O(N_{act})$	$+T_{comm}$
• ''force'' determ. on GPU:	$O(N\ N_{act}/N_{GPU})$	$+T_{GPU}$
• receive the ''force'':	$O(N_{act})$	$+T_{comm}$
• MPI global comm.:	$O((\tau_{lat}+N_{act}) \log(N_{GPU}))$	$+T_{MPI}$
• corr. for ''i'' part.:	$O(N_{act})$	$+T_{host}$

φGPU current usage/results

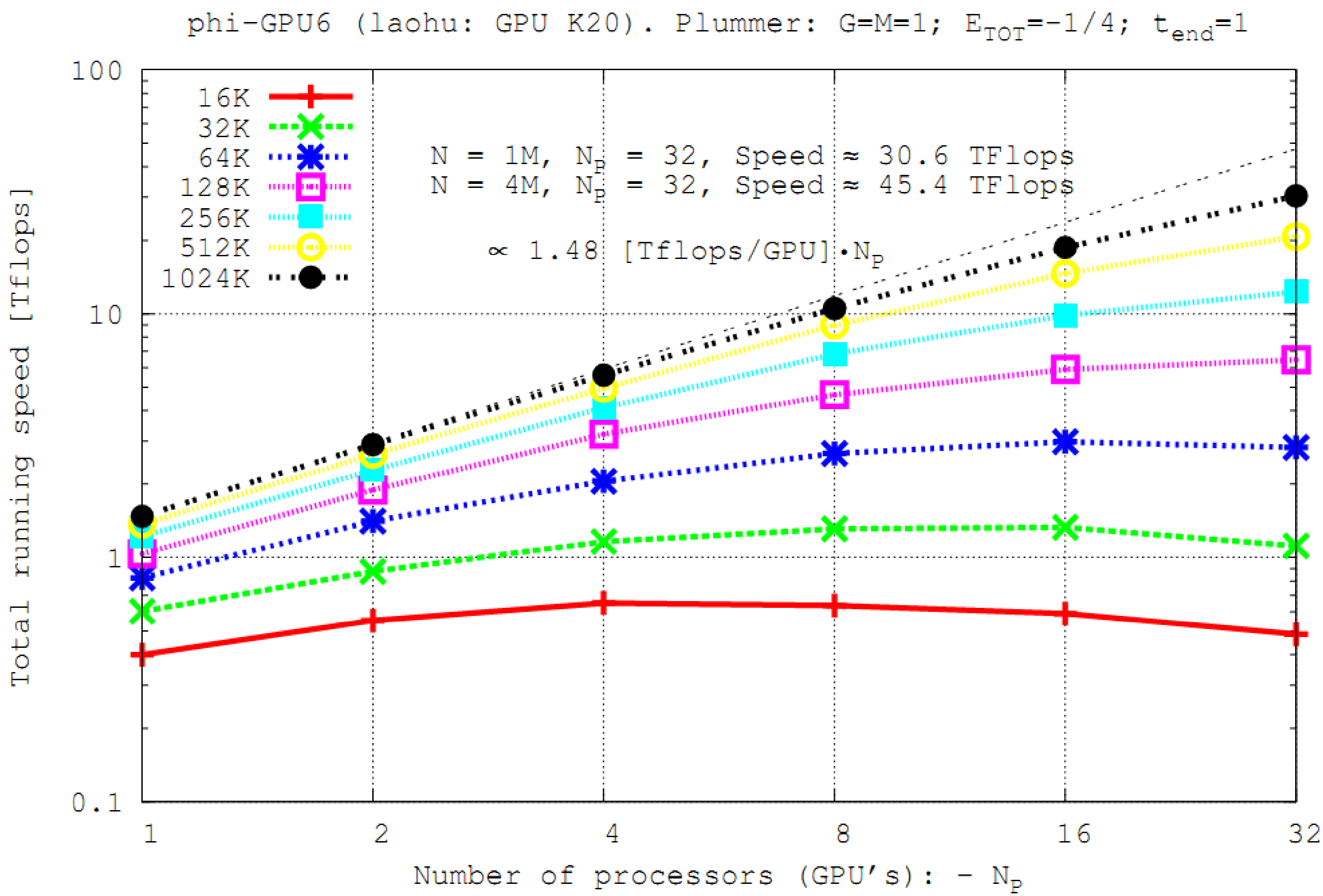


Figure 1. Speed performance with mixed (fp32 + fp64) precision of the φ-GPU 6^{th} order scheme on the K20 GPU cards. The lines with different symbols presents the different particle numbers.

φGPU current usage/results

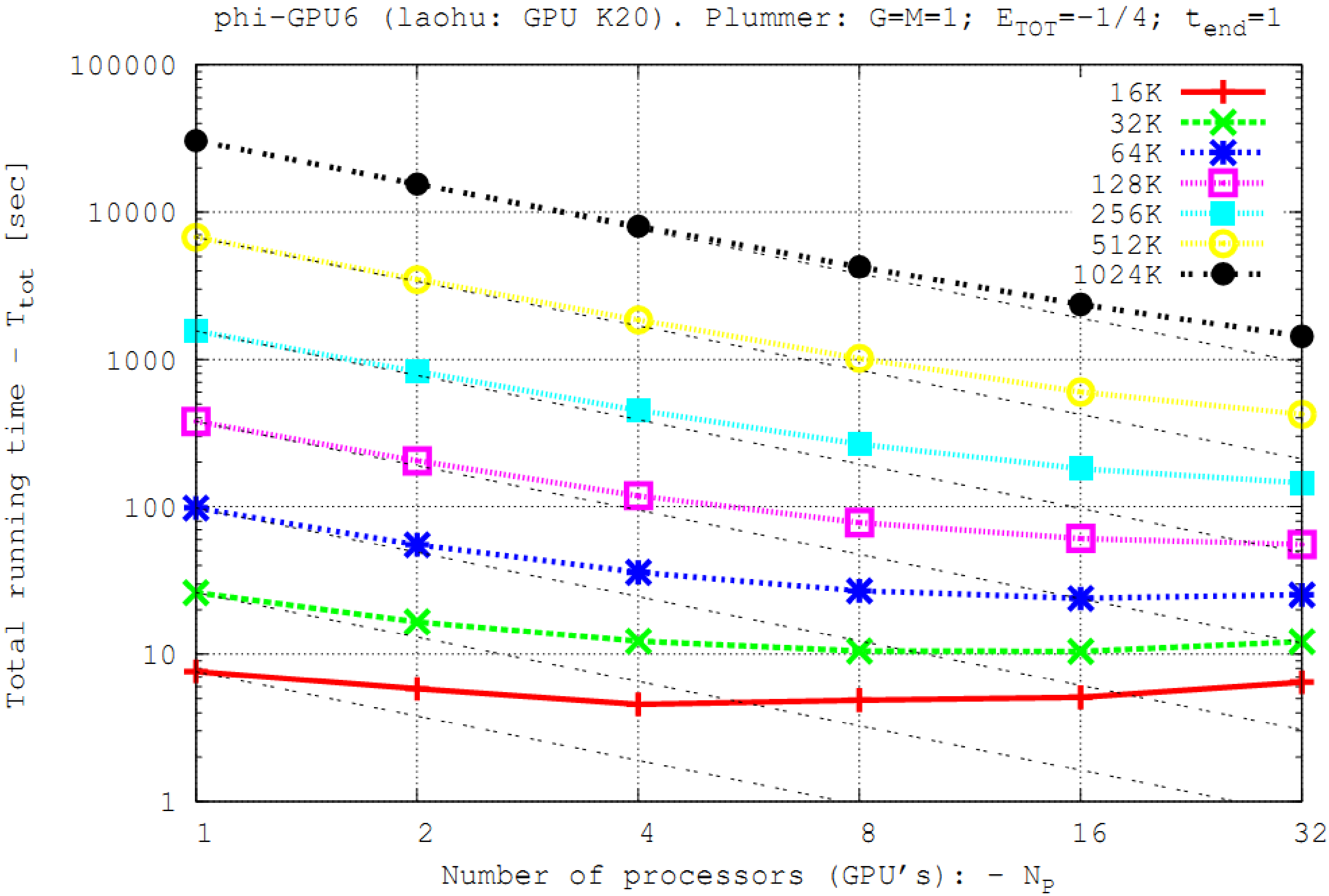
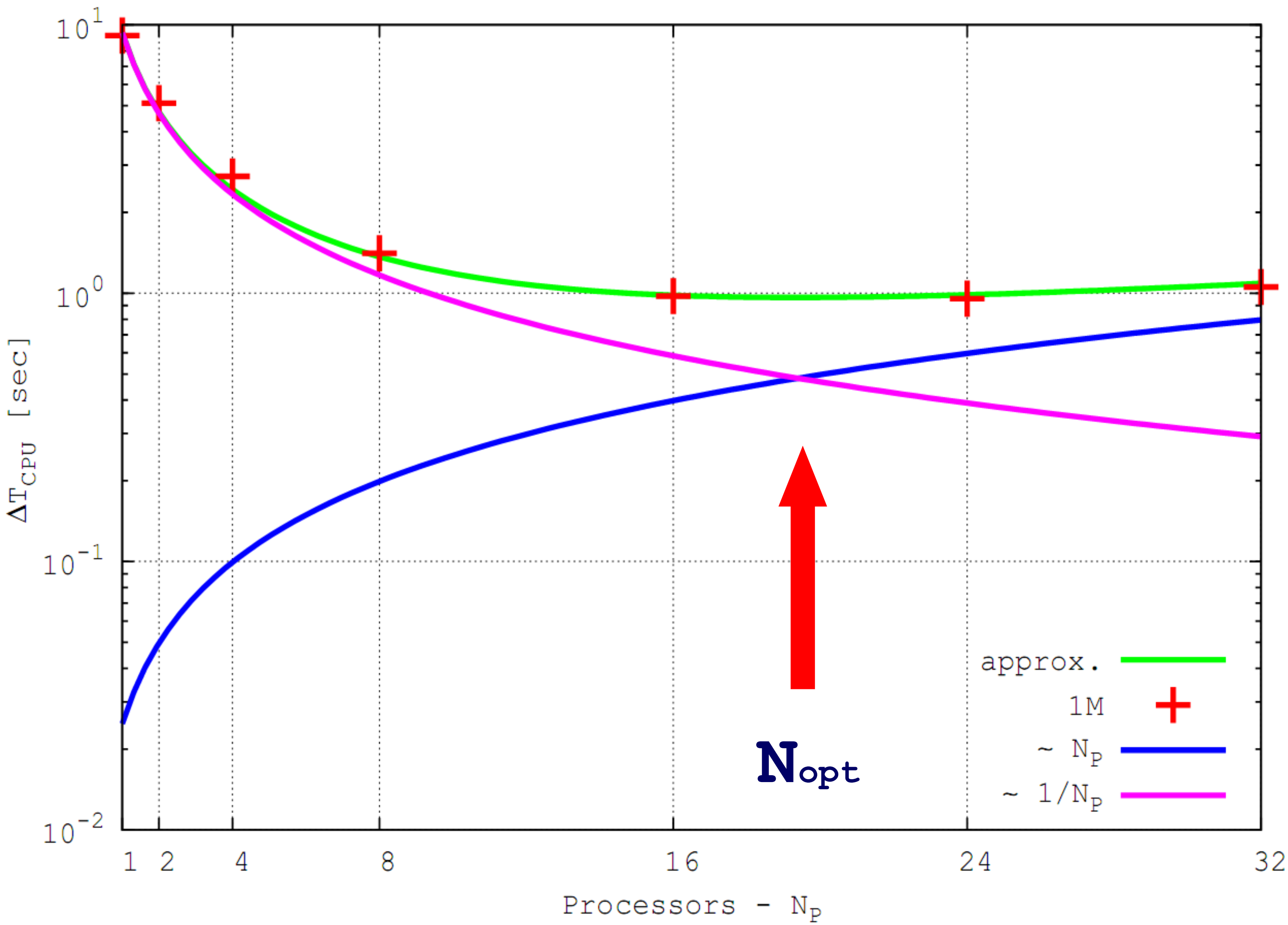


Figure 2. Total wall clock time of 1 time unit integration with the φ-GPU 6^{th} order scheme on the K20 GPU cards. The lines with different symbols presents the different particle numbers.

Parallel code on cluster

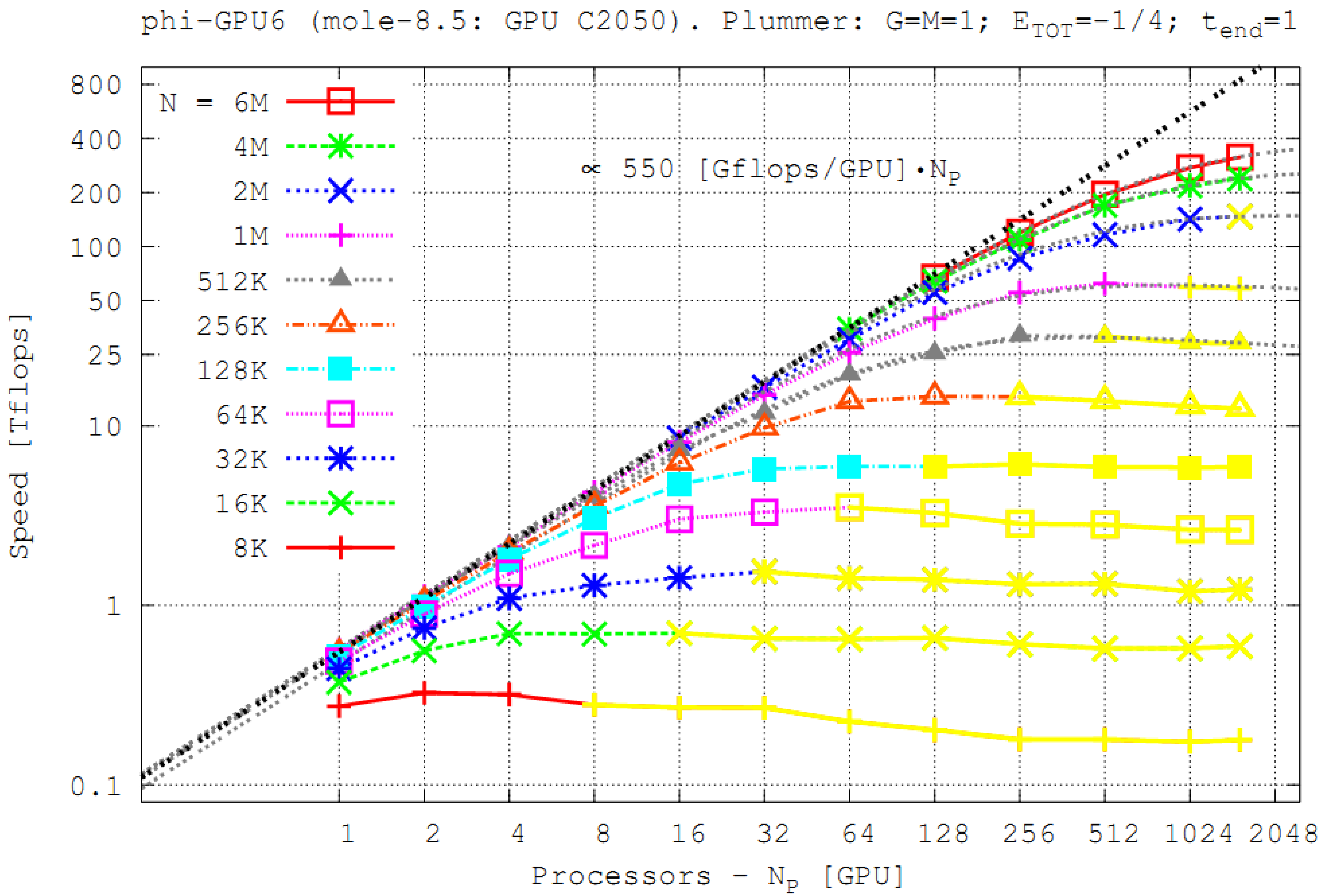


Parallel code on cluster

$$P = \frac{N_{operation}}{\Delta T_{total}} = \frac{\gamma \cdot N_{act} \cdot N}{\Delta T_{total}}$$

$$P = \frac{\gamma \cdot N_{act} \cdot N}{\alpha \cdot N_{act} \cdot N / N_{GPU} + \beta \cdot (\tau_{lat} + N_{act}) \cdot \log(N_{GPU})}$$

φGPU Hermite results

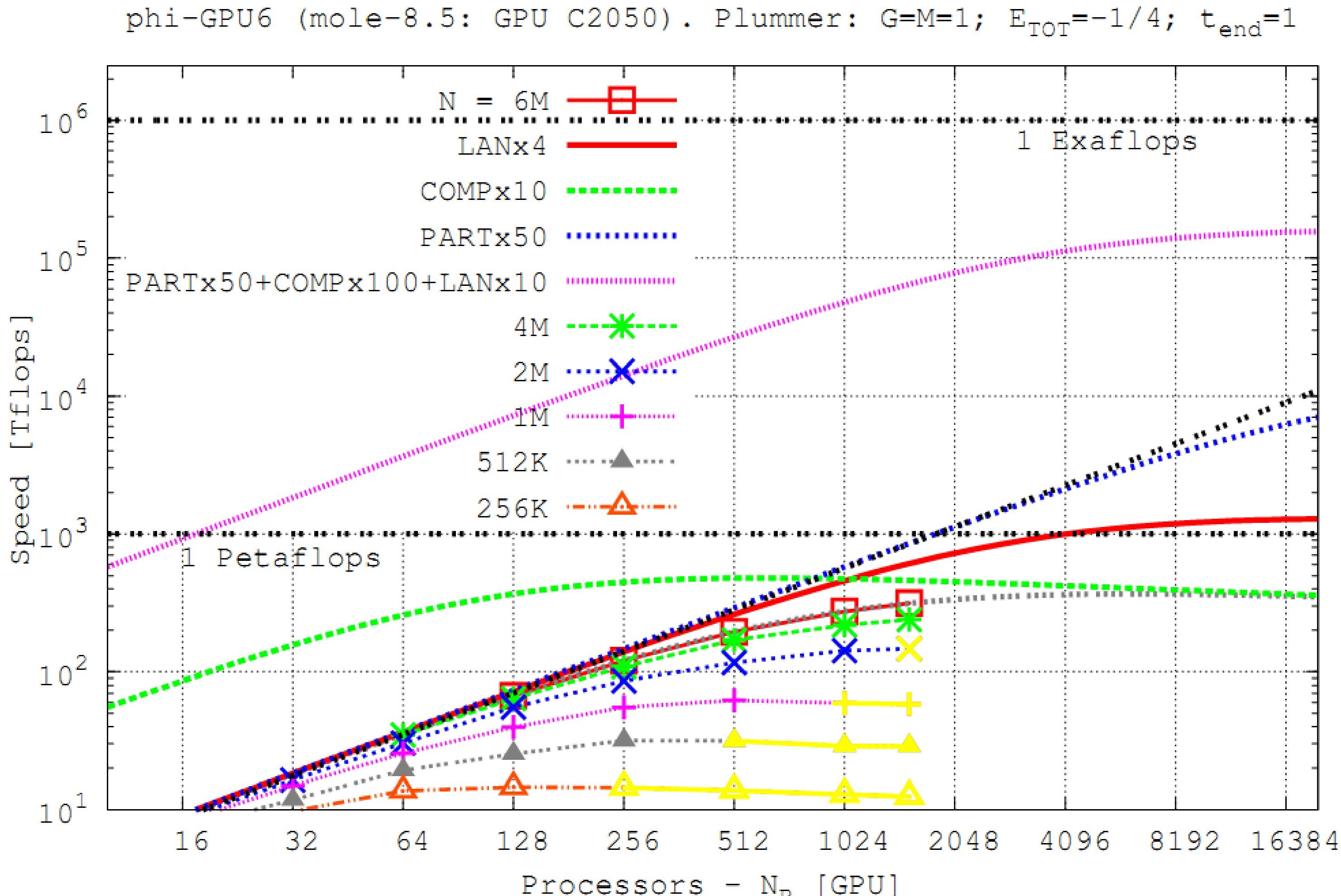


Berczik et al. 2011.04.

1536 * 448 = 688k

Figure 5. Speed performance with mixed (fp32 + fp64) precision of the φ-GPU 6^{th} order scheme on the C2050 GPU cards. The lines with different symbols presents the different particle numbers.

φGPU Hermite results



"TH-1" cluster

Figure 6. Speed performance prognosis with mixed (fp32 + fp64) precision of the φ-GPU 6^{th} order scheme on the C2050 GPU cards. The lines with different symbols presents the different particle numbers. The dotted horizontal lines shows the 1 Petaflops & 1 Exaflops levels.

Intel MIC **Hardware**

Intel® Xeon Phi™ Coprocessor Family Reference Table

SKU #	Form Factor, Thermal	Peak Double Precision	Max # of Cores	Clock Speed (GHz)	GDDR5 Memory Speeds (GT/s)	Peak Memory BW	Memory Capacity (GB)	Total Cache (MB)	Board TDP (Watts)	Process
SE10P (special edition)	PCIe Card, Passively Cooled	1073 GF	61	1.1	5.5	352	8	30.5	300	22nm
SE10X (special edition)	PCIe Card, No Thermal Solution	1073 GF	61	1.1	5.5	352	8	30.5	300	
5110P	PCIe Card, Passively Cooled	1011 GF	60	1.053	5.0	320	8	30	225	
3100 Series	PCIe Card, Actively Cooled	>1 TF	Disclosed at 3100 series launch (H1'13)		5.0	240	6	28.5	300	
	PCIe Card, Passively Cooled	> 1 TF			5.0	240	6	28.5	300	

PCIe Card, Actively Cooled

PCIe Card, Passively Cooled

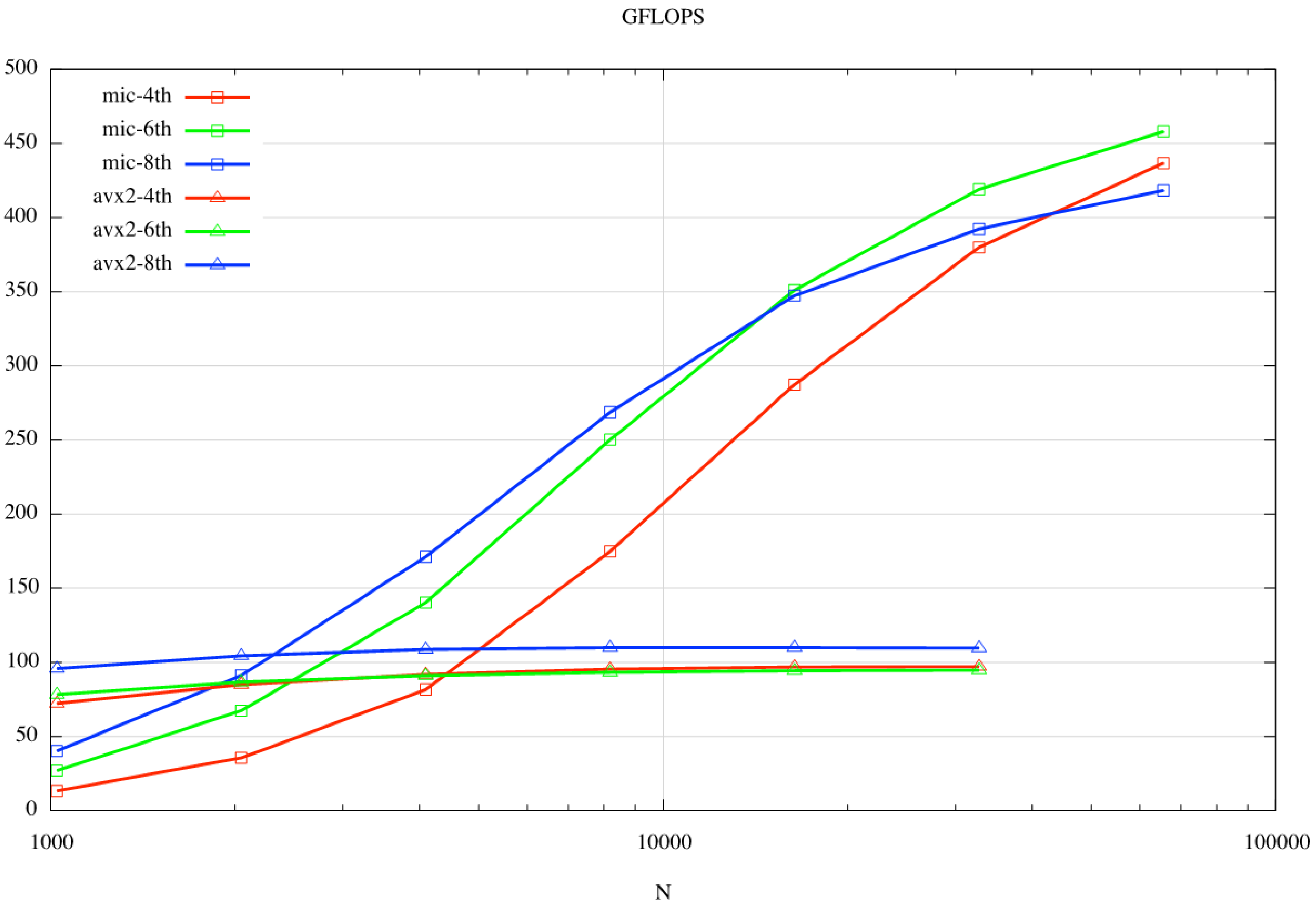
Intel MIC Hardware

INSPUR, NAOC - 2013.XI.26

icpc … "-mmic" … 61 x 4 = 244 x 1.1 GHz omp cores !!!

Full fp64 !!!

φGPU Hermite results



Conclusions…

Our massively parallel codes (φGPU and NBODY6++GPU), which use MPI parallelization as well as acceleration by many GPU's, scale well on large numbers of cores.

They both run very well with no sign of saturation e.g. by communication on the new Kepler K20 GPU accelerator, reaching almost 1.5 Tflop/s per GPU with 2496 cores.

These codes are currently being used for astrophysical research on galactic nuclei, requiring large particle resolution.

With realistic technical improvements of GPU hardware and network speed we expect to reach ~0.2 Exaflop/s speed for N=300M particles.

Thank you for your attention…