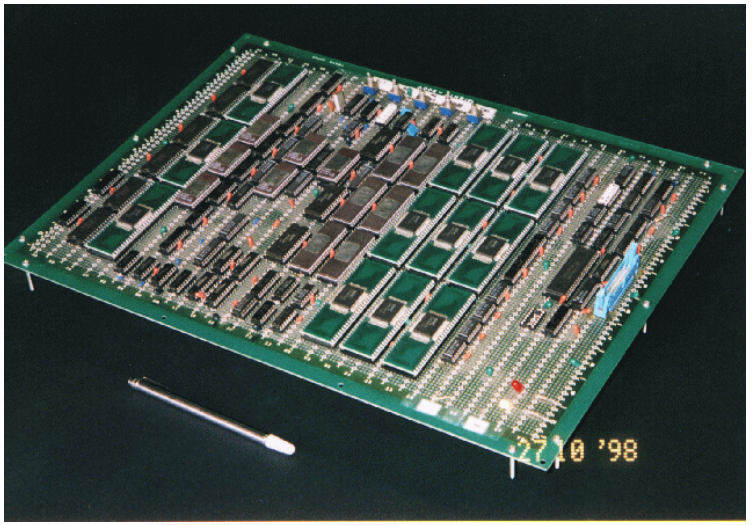
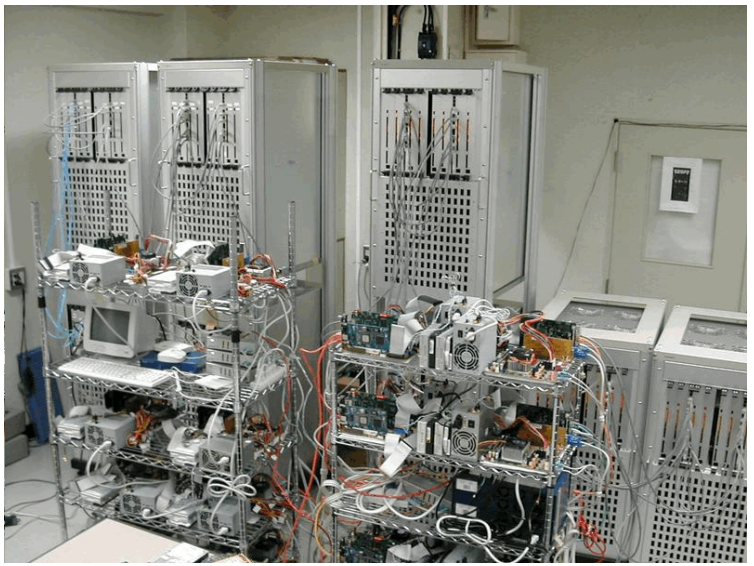
GRAPE and GRAPE-DR

Jun Makino
Center for Computational Astrophysics
and
Division Theoretical Astronomy
National Astronomical Observatory of Japan

MODESTA-10 2010 Sept 3, 2010, Beijing, China

GRAPE-1 to GRAPE-6



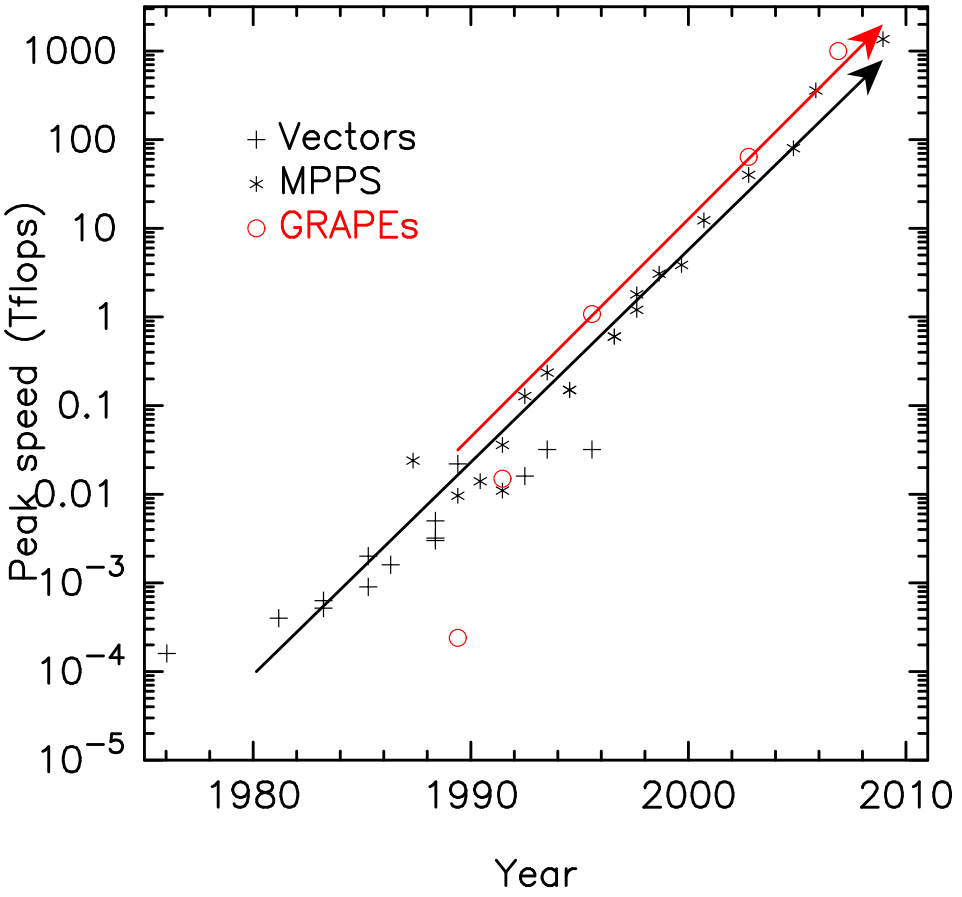


GRAPE-1: 1989, 308Mflops

GRAPE-4: 1995, 1.08Tflops

GRAPE-6: 2002, 64Tflops

Performance history



Since 1995 (GRAPE-4), GRAPE has been faster than general-purpose computers.

Development cost was around 1/100.

"Problem" with GRAPE approach

- Chip development cost has become too high.

Year	Machine	Chip initial cost	process
1992	GRAPE-4	200K$	1μm
1997	GRAPE-6	1M$	250nm
2004	GRAPE-DR	4M$	90nm
2010?	GDR2?	$>$ 10M$	45nm?

Initial cost should be 1/4 or less of the total budget.
How we can continue?

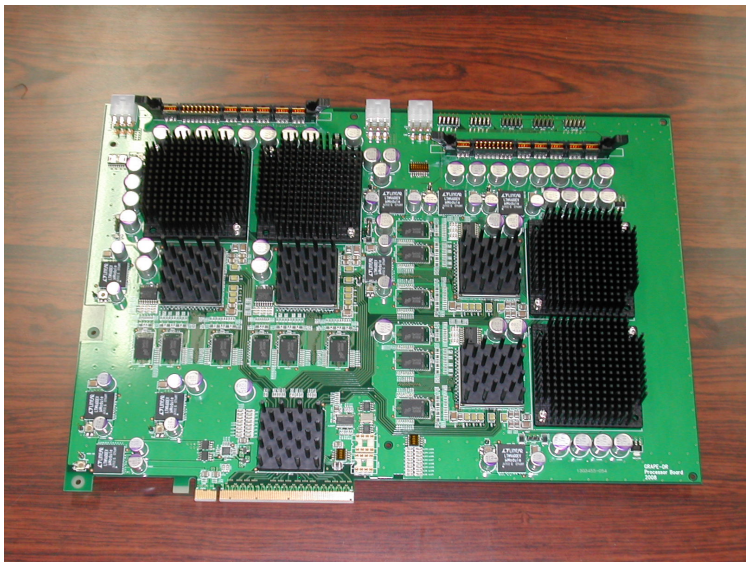
Current Generation— GRAPE-DR

- New architecture — wider application range than previous GRAPEs
- primarily to get funded
- No force pipeline. SIMD programmable processor
- "Parallel evolution" with GPUs.
- Developent: FY 2004-2008

The Chip

Sample chip delivered May 2006
90nm TSMC, Worst case 65W@500MHz

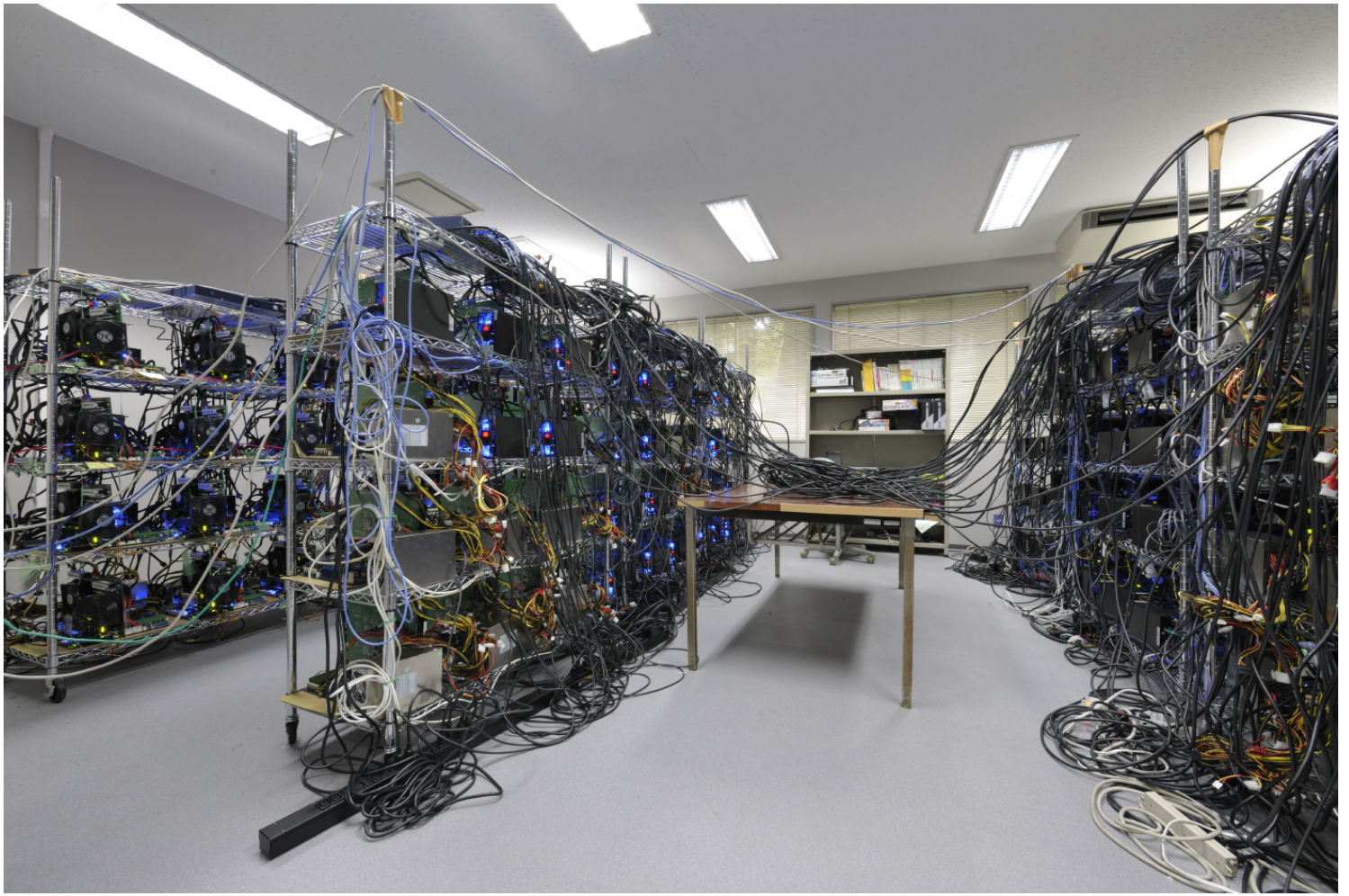
Processor board



PCIe x16 (Gen 1) interface
Altera Arria GX as DRAM controller/communication interface

- Around 200W power consumption
- Not quite running at 500MHz yet... (FPGA design not optimized yet)
- 819Gflops DP peak (400MHz clock)
- Available from K&F Computing Research (www.kfcr.jp)

GRAPE-DR cluster system



OpenMP-like compiler

Goose compiler (Kawai 2009)

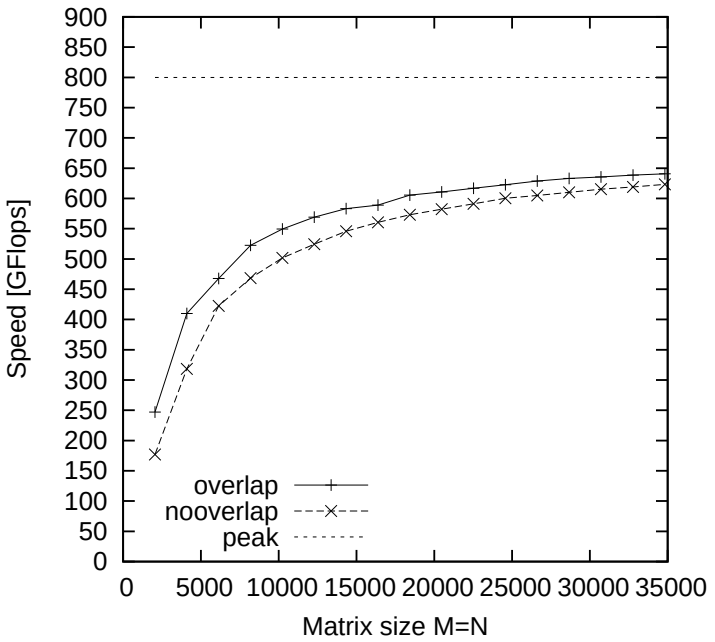
```
#pragma goose parallel for icnt(i) jcnt(j) res (a[i][0..2])
    for (i = 0; i < ni; i++) {
        for (j = 0; j < nj; j++) {
            double r2 = eps2[i];
            for (k = 0; k < 3; k++) dx[k] = x[j][k] - x[i][k];
            for (k = 0; k < 3; k++) r2 += dx[k]*dx[k];
            rinv = rsqrt(r2);
            mf = m[j]*rinv*rinv*rinv;
            for (k = 0; k < 3; k++)  a[i][k] += mf * dx[k];
        }
    }
```

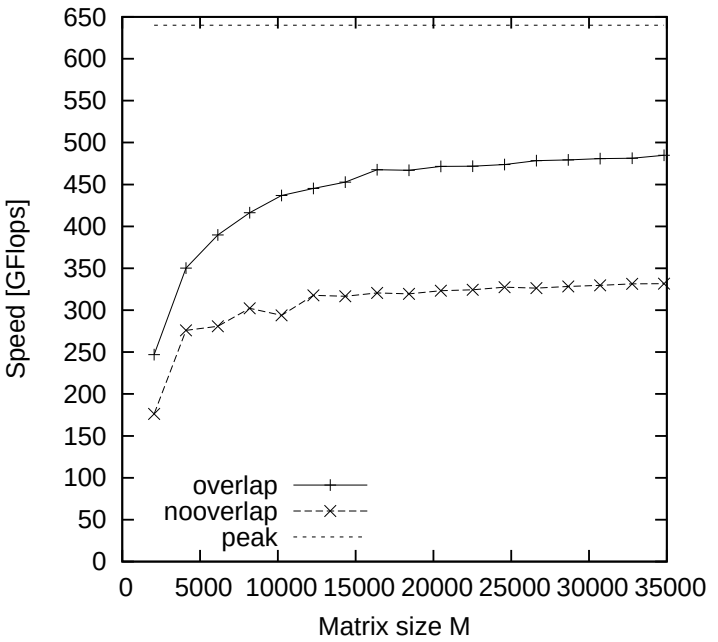
Generates code for single- and double-loops
(Translates to Nakasato's language)

Performance and Tuning example

- **HPL (LU-decomposition)**
- **Gravity**

Matrix-multiplication performance



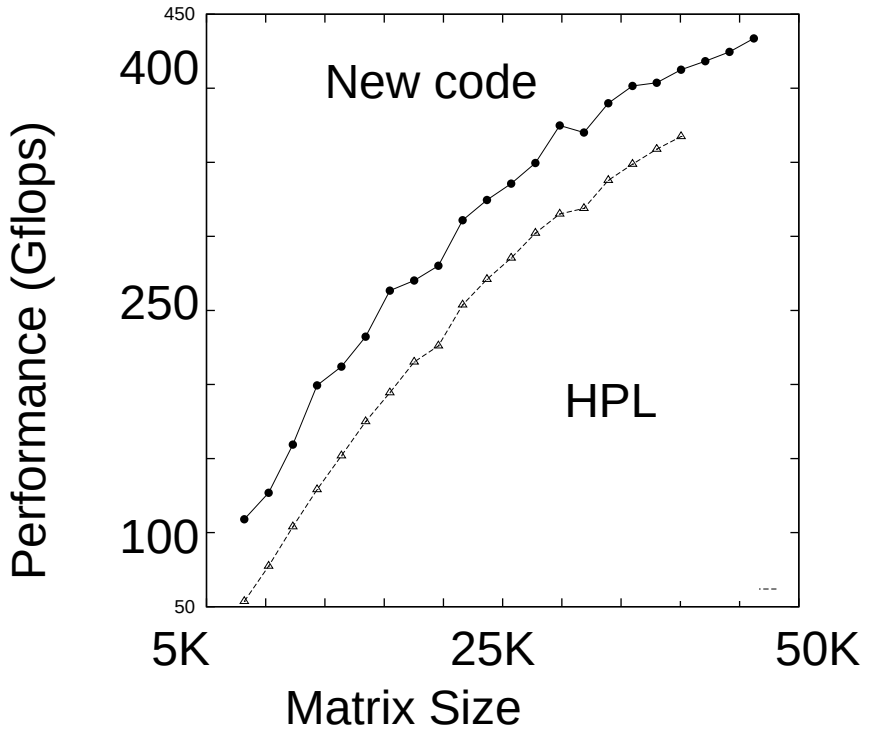


M=N, K=2048, 640 Gflops **N=K=2048, 450 Gflops**

FASTEST single-card performance on the planet.

(Fermi: 3-400Gflops?)

LU-decomposition performance



Speed in Gflops as function of Matrix size
Top: new code
Bottom: HPL 1.04a
430 Gflops (54% of theoretical peak) for N=50K

Little Green 500, June 2010

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	815.43	National Astronomical Observatory of Japan	GRAPE-DR accelerator Cluster, Infiniband	28.67
2	773.38	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
2	773.38	Universitaet Regensburg	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
2	773.38	Universitaet Wuppertal	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
5	536.24	Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw	BladeCenter QS22 Cluster, PowerXCell 8i 4.0 Ghz, Infiniband	34.63

#1: GRAPE-DR, #2: QPACE: German QCD machine
#9: NVIDIA Fermi

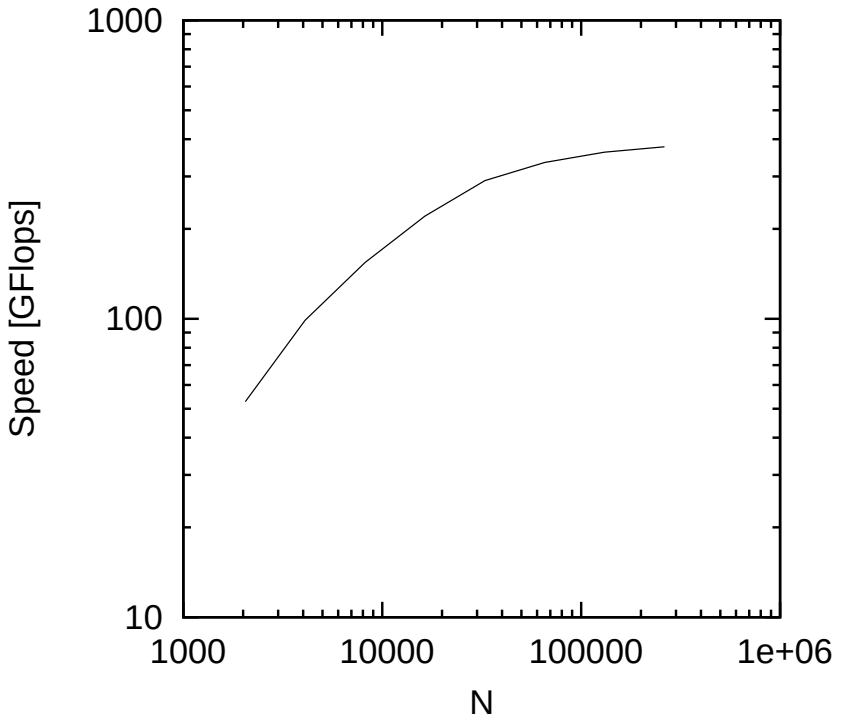
HPL (parallel LU)

- **Everything done for single-node LU-decomposition**
- **Both column- and row-wise communication hidden**
- **TRSM further modified: calculate UT^{-1} instead of $T^{-1}U$**
- **More or less working, still lots of room for tuning**

N=240K, 64 nodes: 24Tflops/29KW
x2 performance compared to HPL 1.04a
815Mflops/W: #1 in Little Green500 list

Gravity kernel performance

(Performance of individual timestep code not much different)



Assembly code (which I wrote) is not very optimized yet... Should reach at least 600 Gflops after rewrite.

Next-Generation GRAPE

Question:
Any reason to continue hardware development?

- **GPUs are fast, and getting faster**
- **FPGAs are also growing in size and speed**
- **Custom ASICs practically impossible to make**

Next-Generation GRAPE

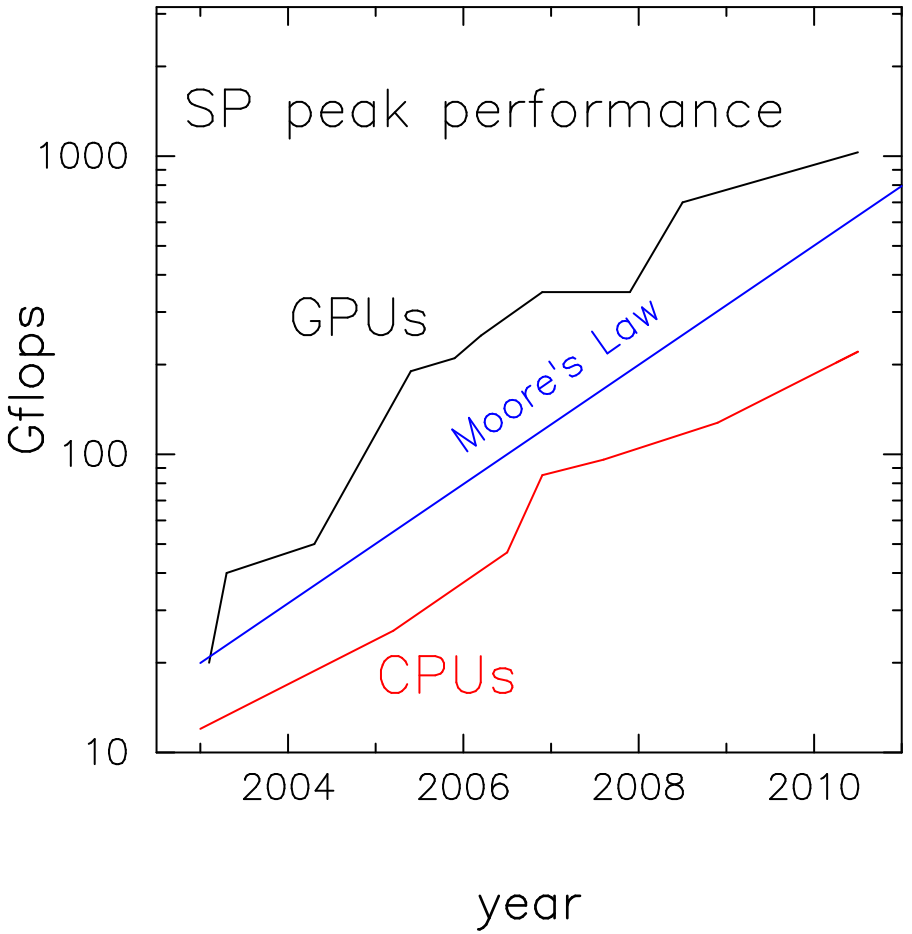
Question:
Any reason to continue hardware development?

- **GPUs are fast, and getting faster**
- **FPGAs are also growing in size and speed**
- **Custom ASICs practically impossible to make**

Answer?

- **GPU speed improvement might have slowed down**
- **FPGAs are becoming far too expensive**
- **Power consumption might become most critical**
- **Somewhat cheaper way to make custom chips**

GPU speed improvement slowing down?



Clear "slowing down" after 2006 (after G80)

Reason: shift to more general-purpose architecture

Discrete GPU market is eaten up by unified chipsets and unified CPU+GPU

But: HPC market is not large enough to support complex chip development

FPGA

"Field Programmable Gate Array"

- **"Programmable" hardware**
- **"Future of computing" for the last two decades....**
- **Telecommunication market needs: large and fast chips (very expensive)**

Structured ASIC

- **Something between FPGA and ASIC**
- **eASIC: 90nm (Fujitsu) and 45nm (Chartered) products.**
- **Compared to FPGA:**
  - **3x size**
  - **1/10 chip unit price**
  - **non-zero initial cost**
- **Compared to ASIC:**
  - **1/10 size and 1/2 clock speed**
  - **1/3 chip unit price**
  - **1/100 initial cost ($>$ 10M USD vs $\sim$ 100K)**

GRAPEs with eASIC

- Completed an experimental design of a programmable processor for quadruple-precision arithmetic. 6PEs in nominal 2.5Mgates.
- Started designing low-accuracy GRAPE hardware with 7.4Mgates chip.

Summary of planned specs:

- around 8-bit relative precision
- $\sim$100 pipelines, 300-400 MHz, 2-3Tflops/chip
- small power consumption: single PCIe card can house $\sim$8 chips (10 Tflops, 50W in total)

Will this be competitive?

Rule of thumb for a special-purpose computer project:

> **Price-performance goal should be more than 100 times better than that of a PC available when you start the project.**

— x 10 for 5 year development time
— x 10 for 5 year lifetime

Compared to CPU: Okay
Compared to GPU: ??? (Okay for electricity)

Will this be competitive?

Rule of thumb for a special-purpose computer project:

> **Price-performance goal should be more than 100 times better than that of a PC available when you start the project.**

— x 10 for 5 year development time
— x 10 for 5 year lifetime

Compared to CPU: Okay
Compared to GPU: ??? (Okay for electricity)

Will GPUs exist 10 years from now?

Tree-Direct hybrid

BRIDGE Hamiltonian (Fujii et al 2007)

$$H = H_\alpha + H_\beta,$$

$$H_\alpha = -\sum_{i<j}^{N_G} \frac{G m_{G,i} m_{G,j}}{r_{GG,ij}} - \sum_{i=1}^{N_G} \sum_{j=1}^{N_{SC}} \frac{G m_{G,i} m_{C,j}}{r_{GS,ij}},$$

$$H_\beta = \sum_{i=1}^{N_G} \frac{p_{G,i}^2}{2m_{G,i}} + \sum_{i=1}^{N_{SC}} \frac{p_{C,i}^2}{2m_{C,i}} - \sum_{i<j}^{N_{SC}} \frac{G m_{C,i} m_{C,j}}{r_{CC,ij}},$$

Separate internal motion (or potential) of star cluster from parent galaxy (and interaction with it)

PPPT

Oshino et al (in prep)
PPPT (Particle-Particle, Particle-Tree) Hamiltonian

$$H = H_{Hard} + H_{Soft},$$
$$H_{Hard} = \sum_{i=1}^{N} \left(\frac{p_i^2}{2m_i} - \frac{G m_i m_\odot}{r_i} \right) - \sum_{i<j}^{N} \frac{G m_i m_j}{r_{ij}} W(r_{ij}),$$
$$H_{Soft} = \sum_{i<j}^{N} \frac{G m_i m_j}{r_{ij}} (1 - W(r_{ij})),$$

Separate near field and far field (cutoff could depend on particle mass)

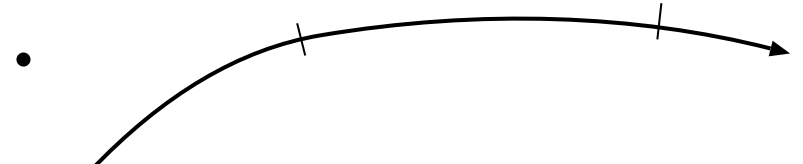
PPPT example run

Planetesimal run (earth region 10^4 particles, $10^{-10} M_\odot$ particles)

Good enough for planet formation

Okay for star cluster?

Limit of individual timestep algorithm



Summary

- **GRAPEs, special-purpose computer for gravitational N-body system, have been providing 10x - 100x more computational power compared to general-purpose supercompuers.**
- **GRAPE-DR, with programmable processors, has wider application range than traditional GRAPEs.**
- **Peak speed of a GRAPE-DR card with 4 chips is 800 Gflops (DP).**
- **DGEMM performance 640 Gflops,**
  LU decomposition $>$ 400Gflops
- **Achieved the best performance per W (Top 1 in the Little Green 500 list, 815Mflops/W)**
- **Accelerators require new algorithms, not just porting and tuning**