# STELLAR DYNAMICS ON 200 TFLOPS SPECIAL-PURPOSE COMPUTERS

J. MAKINO

*Department of Systems Science, College of Arts and Sciences, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153, Japan*

**Abstract.** We overview the GRAPE-6 project, a follow-up of the teraflops GRAPE-4 project. GRAPE-6 will be completed by 1999-2000 and its planned peak speed is 200 Tflops. Its architecture will be largely similar to that of GRAPE-4, which is a specialized hardware to calculate the gravitational interaction between particles. The improvement in the speed will come mainly from the advance in the silicon semiconductor technology. GRAPE-6 will enable us to directly simulate the evolution of star clusters with up to 1 million stars.

## 1. Introduction

In 1988, we started the development of special-purpose computers for astrophysical $N$-body problems (GRAPE; [11]). The basic idea was to build a simple and small hardware, which is designed specifically to calculate gravitational interactions between particles. This hardware operates in cooperation with a general-purpose programmable computer, which performs all other calculations such as time integration and I/O (see figure 1).

We believe this approach has been so far highly successful. The first machine, GRAPE-1 [5], achieved the speed of 240 Mflops in year 1989. GRAPE-3 [10] achieved 14 Gflops 1991. In 1995, GRAPE-4 [9] became the first computer to achieve the peak speed of 1 Tflops and sustained speed of 500 Gflops. In addition, more than 25 institutes, both within and outside Japan, now own various versions of GRAPE hardwares and many of these GRAPE hardwares are actively used. Of course, our group has been using GRAPEs, and many of our results such as gravothermal oscillations [7],

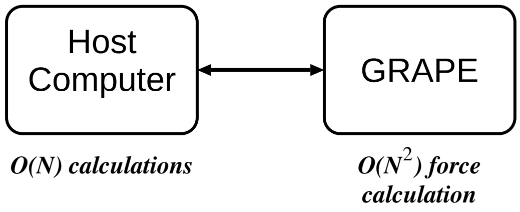


*Figure 1.* Basic structure of GRAPE

runaway growth of protoplanets [6], and formation of CDM halos [3] would have been simply impossible without GRAPE-4.

This success of GRAPE hardwares led us to investigate the possibility of the successor for GRAPE-4, namely GRAPE-6. Just after the completion of GRAPE-4 in 1995, we started to organize an international collaboration to develop and use next-generation GRAPE systems. In June 1997, the project to develop GRAPE-6 was approved by JSPS (Japan Society for Promotion of Science), as one of the projects under "Research for the Future" program. In the following, we briefly overview the GRAPE-6 project. In section 2, we discuss how we achieve the speed of sub-petaflops range. In sections 3 and 4 we discuss the architecture of the processor chip and that of the whole machine. In section 5 we briefly discuss the concept of "reconfigurable computation", which we will use to extend the application range of GRAPE architecture. In section 6 we discuss the expected scientific outcomes.

## 2. A sub-petaflops computer in year 2000

We plan to complete GRAPE-6 with the peak speed of 200 Tflops by year 2000. On the other hand, the peak speed of the fastest general-purpose computer has been increasing by a factor of 100 per decade for the last half century. This trend is expected to continue for at least next 10–20 years. Even the most ambitious project such as ASCI (Accelerated Strategic Computing Initiative) put the date to achieve 100 Tflops no sooner than 2003. So how we, just a handful of astrophysists with a limited budget, can outperform a national project with budget more than 1 billion dollars?

The answer is that we do not really outperform big projects like ASCI, since the goals are very different. The goal for ASCI machines is to provide a general-purpose high-performance computing platform, while our goal is much more limited. GRAPE-6 will be designed essentially for a single problem, namely the gravitational $N$-body problem. It will be much faster than ASCI (or any other) machine for $N$-body problem, but cannot even run a program for, say, grid-based CFD.

However, the limitation in the range of the application does not necessarily guarantee a better cost-performance. In fact, there have been a number of projects to develop special-purpose computers for various problems and only a handful of them can be regarded as success.

The basic reason for the success of GRAPE is that the problem itself allows us an extremely efficient use of available transistors. The number of transistors available on a chip increases by a factor of 100 in every decade, and the speed of the transistor increases by a factor of 10 in 10 years. Thus, if we can use all available transistors to do arithmetic operations, we can achieve the speedup by a factor of 1000 in 10 years.

In the case of general-purpose computers, the increase in the speed has been slower than this factor of 1000 in 10 years, since not all transistors are used for arithmetic operations. The fraction of the transistors used for arithmetic operations has been decreasing.

The reason for the decrease is that the speed of a processor is limited by the degree of parallelism and the data transfer bandwidth between processor and memory [8]. Even though we can integrate hundreds of arithmetic units in a chip, it is not easy to actually use all of them. In fact, it has become very difficult to provide the memory bandwidth sufficient for a single arithmetic unit.

The equation of motion for an $N$-body problem

$$\frac{d^2\mathbf{x}_i}{dt^2} = -G\sum_{j\neq i}^{N} m_j \frac{\mathbf{x}_i - \mathbf{x}_j}{({r_{ij}}^2 + \epsilon^2)^{3/2}}, \quad (1)$$

allows us to integrate a number of arithmetic units on a chip without problems either in the way to use them or in the memory bandwidth.

For parallelism, we can use three different approaches. The first is the pipeline specialized to the force calculation. With a pipelined structure, we can use 50 or more arithmetic units to calculate the force on a single particle. The second is to parallelize these pipelines to calculate the force on different particles, and the third is to use multiple pipelines to calculate the force on one particle. We can combine all these three approaches to reduce the requirement for the memory bandwidth by many orders of magnitude compared to what is required for a general-purpose computer.

The first approach, the pipelined processor specialized to one application, has an additional advantage that the accuracy (and therefore physical size) of each arithmetic unit can be optimized for the operation it performs. Thus, if we do not need 64-bit accuracy for an operation, we can reduce the accuracy of the arithmetic unit for that operation, without sacrificing the overall accuracy. This has been a very important factor for the early success of our machines like GRAPE-1 [5] and GRAPE-3 [10].

| | GRAPE-4 | GRAPE-6 |
|---|---|---|
| Feature size | 1 $\mu$m | 0.25 $\mu$m |
| Transistor count | 400K | 7M |
| No. of pipelines | 1/3 | 6$\sim$ 8 |
| Clock cycle | 30 MHz | $\sim$ 125 MHz |
| Speed | 600 Mflops | 50 Gflops |
| Supply Voltage | 5V | 2.5-1.8V |
| Power | 5W | 10-15W |

TABLE 1. Comparison of pipeline chips for GRAPE-4 and GRAPE-6

For GRAPE-4 with relatively high accuracy and massively parallel architecture, the latter two approaches played important roles. They will be even more important for GRAPE-6.

## 3. The Pipeline Processor

How GRAPE-6 chip will look like? Here, we briefly discuss the difference between GRAPE-4 processor chip (the HARP chip) and GRAPE-6 chip. The changes are introduced to make full use of the advance of the VLSI technology.

The advance in technology has two outcomes. The first is the increase in the available number of transistors on a single chip. The HARP chip was fabricated using the 1$\mu$m technology, while the 0.25$\mu$m technology will be used for GRAPE-6. Roughly speaking, we can use 16 times more transistors. Secondly, switching delay of the transistor is improved roughly in proportional to its physical size, which we hope to give us around a factor of four increase in the clock cycle. Thus, we expect that GRAPE-6 chip will have 64 times more processing power than GRAPE-4 chip, by means of larger number of pipelines and higher clock speed.

The power consumption is still relatively low, because of the shrink in the physical size of the transistors and the reduction in the supply voltage.

The increase in the number of pipelines and clock period of the pipeline chip, however, forced us to reconsider the architecture of the chip. The pipeline chip of GRAPE-4 (HARP chip) implemented only the force calculation pipeline. GRAPE-4 has a separate pipeline to evaluate the predictor polynomials for the position of particles, so that it can be used with individual timestep algorithms (for details, see [9]). This pipeline was implemented in another chip (PROMETHEUS chip). A PROMETHEUS chip supplied the data of particles to 48 force calculation pipeline chips. The data transfer bandwidth between PROMETHEUS and HARP was 256 MB/s. The number of HARP chips connected to a PROMETHEUS chip was chosen so that we can achieve a reasonable efficiency for individual timestep algorithm. The number of particles for which the gravitational force are calculated in parallel, $N_p$, must be relatively small, since the number of particles that can be integrated in parallel is small (otherwise the individual timestep algorithm would be useless). For $N \lesssim 10^5$, $N_p$ must be less than 100.

If we use a similar architecture for GRAPE-6, a rather serious problem arises: we need a very high total memory bandwidth. With GRAPE-6, we can increase $N_p$ to around 500. Even so, the required memory bandwidth is about 40 times higher than that of GRAPE-4, since the peak speed of GRAPE-6 will be 200 times as that of GRAPE-4.

There are a number of different approaches to achieve this high memory bandwidth. We analyzed several of them and reached the conclusion that a tightly-coupled memory-processor chip pair is, at present, most cost-effective solution. If we can integrate the memory and pipelines into a single chip, it would be an even better solution. However, as of late 1997, the logic-memory integration still has too large impact on the density and performance. In a few years, advance in the process technology might make the logic-memory integration a practical option.

If we place memory chips and a pipeline chip physically close, it is not very difficult to achieve a high bandwidth. We can use high clock frequency without much problem. Of course, we have to use advanced packaging technologies such as MCM (multi-chip module), which was used in GRAPE-4. Figure 2 shows an example design. Here, a multi-chip module contains two GRAPE-6 chips, each connected to two SSRAM (Synchronous Static Random Access Memory) chips. For the connection between memory and pipeline chips, we use a high speed (125 MHz) clock, to achieve the data transfer speed of around 1.2 GB/s. Two GRAPE-6 chips share a common I/O port, through which they are connected to the communication path to the host. For the I/O port, we plan to use one port with 64-bit width and 25 MHz data rate (200MB/s). The physical wire length for the I/O port will be considerably longer than that of the port to the memory. Therefore it is important to keep both the clock frequency and number of wires (*i.e.*, the I/O bandwidth) as low as possible.

## 4. The overall architecture

The total system will consist of 4096 chips, which will be organized into 16 clusters each with 8 processor boards. Each board will carry 32 processor chips. Total power consumption will be around 40-50 kW. Thus, both the power consumption and the physical size of GRAPE-6 would be several

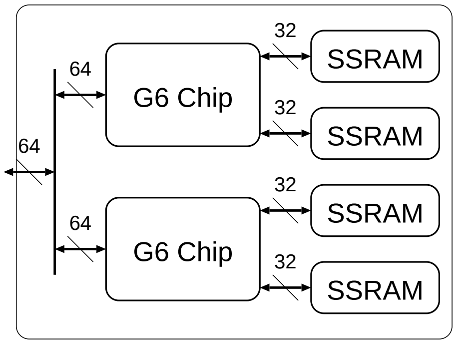


*Figure 2.* The GRAPE-6 processor module

time larger than those of GRAPE-4 (4 clusters, 10 kW). The difference in size is essentially to deal with the increased heat dissipation.

Let us now consider the communication network. The design goal of GRAPE-6 is to achieve a reasonable performance for the simulation of globular clusters with $N = 5 \times 10^5$ at the full-size configuration. This implies that the data transfer speed between GRAPE and host must be at least 2 GB/s, and that $N_p$ must be less than 1000 at the maximum.

At the level of the cluster, different clusters calculate the forces on different particles, but from the same set of particles. Therefore, all clusters must have a same image of $j$ particles (particles stored in the off-chip memory). This implies we need different requirements for the data transfer for $j$ particles and that for $i$ particles. For $j$ particles, data must be broadcasted and the transfer speed to each box must be around 1.6 GB/s. For $i$ particles, different clusters require different data, and necessary bandwidth is around 100 MB/s for a cluster.

Within one cluster, different chips calculate the forces on the same set of particles, but from different particles. Thus, the partial forces calculated on pipeline chips must be summed up through a reduction network. Also, data for $i$ particles must be broadcasted. For $j$ particle, each cluster should have around 1.6 GB/s. Therefore, each board must accept around 200 MB/s of data stream.

Figure 3 shows one possible design. In this figure, 16 clusters are connected to the host computer through two separate networks. One is the broadcast link (B-net), which connects host and other 16 clusters in a one-dimensional, one-directional communication path. The other is a star-type network which directly connects each cluster to the host computer (S-net). B-net is used to broadcast the data to be written to the particle memory, while S-net is used to read/write the registers in the pipeline chip.

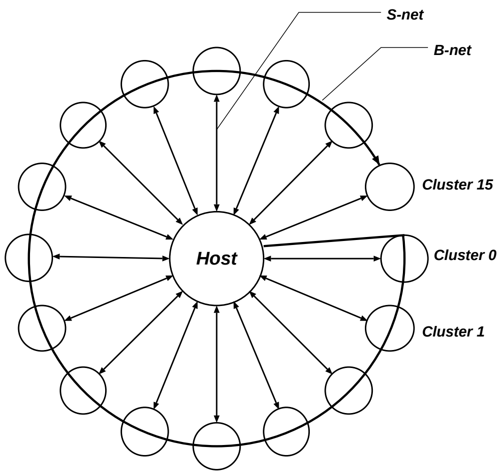


*Figure 3.* The overall structure of GRAPE-6 system

For the link between clusters and/or boards, we plan to use physical layer chipset of FibreChannel or SCI, which can operate at the speed of 100 MB/s.

## 5. Extension

Currently, a number of people are using GRAPE systems for SPH simulations (see, *e.g.*, [12]). In these simulations, GRAPE is used to calculate gravity and to construct the list of neighbor particles. The host handles the SPH interaction between neighbors. The calculation of SPH interaction consumes fairly large fraction of the total CPU time.

If SPH interaction can also be handled by some specialized hardware like GRAPE, we can achieve further speedup. The speedup we can achieve is not very large, typically around a factor of 10 or less. This is because the calculation cost of SPH interaction is still $O(N)$ and not as large as that of gravity. On the other hand, this fact implies that we do not need a very fast hardware.

In GRAPE-6 project, we will try a relatively new approach, so-called "reconfigurable computing" [1], to accelerate SPH and similar applications. An alternative is to develop a hardware specialized to SPH [13]. However, whether the high initial cost of a custom LSI can be justified by a relatively modest speedup is not clear. The basic idea of "reconfigurable computing" is to use a programmable LSI chip (field-programmable gate arrays or FPGA) to implement (part of ) applications. Currently, FPGAs with nominal gate

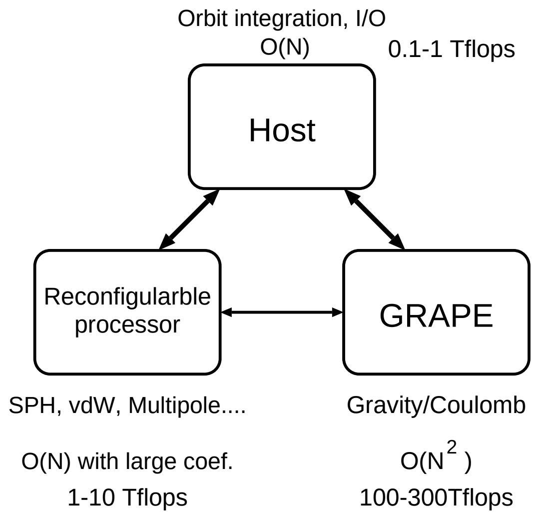


*Figure 4.* Extended GRAPE architecture

count of 100,000 are available. This number is about a factor of 100 smaller than that for a full custom LSI, but may be sufficient to implement single pipeline for SPH calculation.

The readers interested in FPGA and reconfigurable computing are referred to [1]. The bottom line is that we may be able to use them as pipeline processors more flexible than hardwired GRAPE pipelines and at the same time to achieve price performance better than that of programmable general-purpose computers. Of course, it is also true that reconfigurable computing is neither as flexible as general-purpose computer, nor as efficient as GRAPE. Thus, it cannot directly compete with either of them. However, for the part of computation which is relatively time consuming, but much less so compared to gravitational force calculation, reconfigurable computing would offer an ideal solution.

Thus, GRAPE-6 might become a heterogeneous computer with three, not two, components (figure 4). We may be able to use the reconfigurable part for various applications, such as the calculation of van der Waals force in molecular dynamics and evaluation and shifting of spherical harmonics in the fast multipole method.

## 6. Science with GRAPE-6

Table 2 gives a rough estimate of the maximum number of particles we will be able to use on both the general-purpose computers and GRAPE-6 at year 2000, based on the numbers at present. We here assumed that on both

| Problem Area | 1996 GPC | 1996 GRAPE-4 | 2000 GPC | 2000 GRAPE-6 | GRAPE-6 Time |
|---|---|---|---|---|---|
| Planet Formation | $5 \times 10^3$ | $5 \times 10^4$ | $5 \times 10^4$ | $10^6$ | 7 years |
| Globular Cluster | $10^4$ | $5\times10^4$ | $5\times10^4$ | $5\times10^5$ | 10 years |
| Black Hole Binary in Galactic Center | $10^5$ | $10^6$ | $10^6$ | $3\times10^7$ | 10 years |
| Galaxy Evolution | $10^6$ | $2\times10^6$ | $3\times10^7$ | $10^8$ | 3 years |
| Large Scale Structure & Galaxy Formation | $4\times10^6$ | — | $10^8$ | $3\times10^8$ | 3 years |

TABLE 2. Particle Number of Largest Simulation Feasible and GRAPE-6 Time Advantage

general-purpose computers and on GRAPE, we use most efficient algorithm. For collisional systems, we can expect very large gain, even if we take into account the difference in the algorithm used. For collisionless calculations, the gain is not very large, mainly because the size of the calculation is not limited by the calculation speed but by the amount of the memory available. For collisionless calculations, however, the calculation speed of GRAPE will be much faster.

In the following, we just briefly discuss the possible scientific outcomes. For globular clusters, we will be able to honestly simulate small globular clusters without cutting corners. Thus, we will be able to address the timescale problem [4], which arises if we try to model globular clusters with number of particles less than that of real clusters. We will be able to understand the reason why the result of Fokker-Planck calculations and that of $N$-body calculations were wildly different [2], and to get solid understanding of the evolution of globular cluster under the effect of galactic tidal field and stellar evolution.

For planet formation, increased number of particles implies we can model a wider range in the radius with better initial mass resolution. Current simulations with GRAPE-4 are limited to very narrow range in the radius, and therefore cannot address, for example, the mechanism in which the separation of present-day planets are determined. Large scale simulations with GRAPE-6 will allow us to directly investigate the final stage of the planet formation.

Of course, what we learned from the advance in the simulation technique, and in particular from the use of GRAPE-4, is that the increase in the number of particles provided us not only quantitative improvements, but also qualitatively new results which were not anticipated beforehand. We certainly expect similar new results from GRAPE-6.

## 7. Budget and Timetable

GRAPE-6 project is a five-year project with total budget of around 500 M JYE. The plan is to develop the processor chip by 1998 and small prototype by 1999, and full-scale system by year 2000. We plan to make "small" systems (8-16 processor chips; 0.5-1 Tflops) available to many institutes. They will serve as main computing workhorses for large scale simulation of self-gravitating systems.

I'm most grateful to Daiichiro Sugimoto, who has been leading the GRAPE project since its start at 1988. I'd also like to thank to all of those who participated in GRAPE project.

## References

1. Buell D., Arnold J.M. and Kleinfelder W. (1996) *Splash 2: FPGAs in a Custom Computing Machine*. IEEE Comp. Soc. Press, Los Alamitos, CA.
2. Fukushige T. and Heggie D. C. (1995) Pre-collapse evolution of galactic globular clusters. *MNRAS* 276: 206–218.
3. Fukushige T. and Makino J. (1997) On the origin of cusps in dark matter halos. *ApJL* 477: L9–12.
4. Heggie D. C. (1996) Statistics of small-$n$ simulations. In Hut P. and Makino J. (eds) *Dynamical Evolution of Star Clusters; Proceedings of IAU Symposium 174*, pages 131–140. Kluwer, Amsterdum.
5. Ito T., Makino J., Ebisuzaki T., and Sugimoto D. (1990) A special-purpose N-body machine grape-1. *Computer Physics Communications* 60: 187–194.
6. Kokubo E. and Ida S. (1997) Oligarchic growth of protoplanets. submitted to Icarus.
7. Makino J. (1996) Postcollapse evolution of globular clusters. *ApJ* 471: 796+.
8. McCulpin J. D. (1995) Memory bandwidth and machine balance in current high performance computers. *Newsletter of IEEE Computer Architecture Technical Committee* 1: 19–25.
9. Makino J., Taiji M., Ebisuzaki T., and Sugimoto D. (1997) Grape-4: A massively parallel special-purpose computer for collisional n-body simulations. *ApJ* 480: 432+.
10. Okumura S. K., Makino J., Ebisuzaki T., Fukushige T., Ito T., Sugimoto D., Hashimoto E., Tomida K., and Miyakawa N. (1993) Highly parallelized special-purpose computer, GRAPE-3. *PASJ* 45: 329–338.
11. Sugimoto D., Chikada Y., Makino J., Ito T., Ebisuzaki T., and Umemura M. (1990) A special-purpose computer for gravitational many-body problems. *Nature* 345: 33–35.
12. Steinmetz M. (1996) GRAPESPH: cosmological smoothed particle hydrodynamics simulations with the special-purpose hardware grape. *MNRAS* 278: 1005–1017.
13. Yokono Y., Ogasawara R., Takeuchi T., Inutsuka S., Miyama S. M., and Chikada Y. (1996) Development of special-purpose computer for cosmic hydrodynamics by SPH. In Tomisaka K. (ed) *Numerical Astrophysics Using Supercomputers*. National Astronomical Observatory, Japan.