# Next-Generation Massively Parallel Computers — Massively Parallel Computer for Particle-based Simulations

Junichiro Makino
(School of Science, University of Tokyo)

## 1 Project Organization

| | | |
|---|---|---|
| Subleader | Junichiro Makino | Associate Professor, School of Science, University of Tokyo |
| | Izumi Hachisu | Associate Professor, College of Arts and Sciences, University of Tokyo |
| | Makoto Taiji | Associate Professor, Institute for Statistical Mathematics |
| | Yoko Funato | Research Associate, College of Arts and Sciences, University of Tokyo |
| | Toshiyuki Fukushige | Research Associate, College of Arts and Sciences, University of Tokyo |
| | Koji Takahashi | JSPS research fellow, School of Science, University of Tokyo |

## 2 Research Objective

The primary goal of our project is to develop a next-generation massively-parallel computer which is specialized to particle-based simulations, and to make it possible to perform simulations which cannot be handled on general-purpose computers. We also aim to extend the application area of special-purpose computers.

It has been demonstrated that, for particle-based simulations, we can achieve much better cost-performance ratio by developing computing hardware specialized to the evaluation of the interaction between particles. To give just one example, the GRAPE-4 system[MTES97], which we completed in 1995, is the first machine for scientific simulation with the peak speed exceeding one teraflops.

In the present project, we will develop a next-generation special-purpose computer with the peak speed of sub-petaflops range, around one hundred times faster than that of GRAPE-4 system. In addition to this ultra-fast special-purpose computer, we develop a multi-purpose system based on reconfigurable logic. The total system, which we call GRAPE-6, will be a heterogeneous multicomputer which combines these two specialized system with general-purpose parallel computer.

In the following, we briefly discuss the positioning of our GRAPE-6 system compared to general-purpose massively-parallel computers, and why GRAPE-6 system can achieve the high peak speed. As we stated above, the peak speed of GRAPE-6 will be around 100 Tflops, which will be achieved by the final year of the project, FY2001. At that time, the fastest general

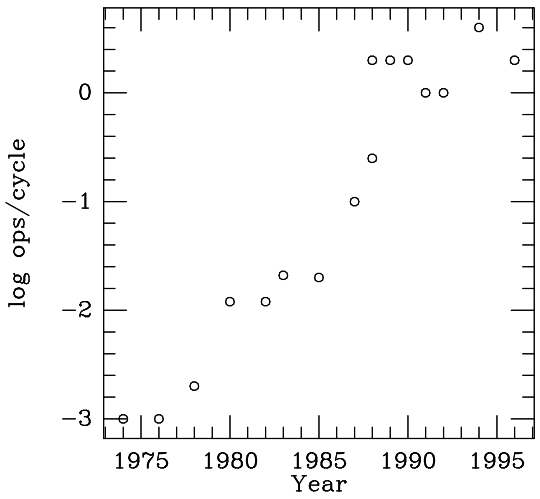


Figure 1: Maximum number of floating point operations per clock cycle for representative microprocessors

purpose computer will be the Earth Simulator with the theoretical peak speed of 40 Tflops (design goal of the sustained performance for real application is 5 Tflops), for the total cost of around 400 M USD. The ASCI project in the US has been leading in constructing large machines. In the timeframe of 2001-2002, the biggest ASCI machine will remain to be ASCI Option White with the peak speed of 16 Tflops for the cost of around 100 M USD. The hardware cost of these machines per peak speed in Mflops is around 5–10 USD/Mflops.

The total budget of the present project is around 4.5 M USD. Therefore, for our project goal of 100 Tflops, the cost performance figure is 0.045 USD/Mflops, 100-200 times better than those of general-purpose massively-parallel computers at the same time. In the case of GRAPE-4, the difference was somewhat smaller, around 50 or so.

The reason why there is such a large difference is that the fraction of the silicon real estate used to implement arithmetic logic units is small in the case of general-purpose computer. Moreover, that ratio has been decreasing as silicon semiconductor technology advances. In the 1990s, the number of transistors used in one microprocessors has increased by a factor of 20 or more. However, the number of floating point arithmetic units integrated in a microprocessor has been practically unchanged (see figure 1). In other words, the fraction of the transistors used for arithmetic units has decreased by roughly a factor of 10 in the last decade.

There are many reasons which prevent the increase in the number of arithmetic units per processor, but the most critical one is the limitation in the bandwidth between the processor and the main memory (so-called "memory wall"). Though we can integrate many arithmetic units, in order to use them efficiently, they need to be connected to the main memory (or at least to the cache memory of some level) with sufficient bandwidth. In the case of the main memory or off-chip cache, this implies we need fast connection between the CPU chip and memory chips.

To increase chip-to-chip communication is much harder than to increase the processing power, simply because we cannot increase the number of wires as fast as we can increase the number of transistors. When we reduce the size of transistors by a factor of two, we can use four times more transistors. However, for off-chip connection, or even for on-chip connections, we can increase the number of wires only by a factor of two. In practice, it's harder to increase the off-chip wire density to match with on-chip wire density. Also, we cannot really scale the on-chip wire size, unless all the wires are short. For long wires, RC delay can easily become much larger than the switching delay of transistors, with the present and future deep submicron technology.

The advantage of the specialize architecture such as GRAPE is that we can use almost all transistors available on chip to implement arithmetic units. As a result, we can achieve the performance orders of magnitudes higher than what is possible with general-purpose processors. In the case of GRAPE hardwares, the nature of the target physical system allows such efficient use of transistors. The target system consists of many particles interacting with long-range forces. Thus, the amount of computation is $O(N^2)$, while the amount of data is $O(N)$, where $N$ is the number of particles in the system. By designing an architecture which makes good use of this characteristics of the target system, we can use large number of transistors made available by state-of-the-arts semiconductor technology. Thus, as long as there is no drastic change in the way general-purpose computers are designed, the relative advantage of special-purpose computers such as GRAPE over general-purpose computer will keep increasing in the future.

To summarize, we believe the importance of the special-purpose computing in the research in computational science and engineering will increase in the future, primarily because the relative advantage of special-purpose approach over general-purpose approach will increase. One of the main goals of the present project is to demonstrate the relative advantage of special-purpose computers.

The second goal of our project is to construct a multi-purpose system for particle-based simulations using reconfigurable logic chips (FPGA; Field-Programmable Gate Arrays). An FPGA chip is essentially an logic LSI chip whose internal circuit design can be reprogrammed after production of the physical chip. Such a programmability is realized by a combination of logic elements by table-lookup and interconnection with programmable switches. Typically, configuration of these lookup tables and switches are stored in SRAMs, and can be downloaded while the system is in operation (while the power is up).

The use of FPGA chip as the pipeline chip for GRAPE-like system has several important advantages compared to the use of full-custom LSI chips. The most important one is the low initial development cost. The development cost of current state-of-the-arts custom logic chips exceeds 1 million USD (as of 1997-1998). On the other hand, we can buy FPGA chips which are mass-produced for less than 1,000 USD (large FPGA chips are, however, expensive, exceeding 3,000 USD per chip). The programmability also implies a shorter design cycle, since we can

correct the mistake in the logic design just by changing the data to be downloaded to the FPGA chips. Thus the debugging cycle is much less than a day. The change in the hardware design of a custom chip would take at least a few months and lots of money.

The disadvantage of FPGA chips compared to full-custom design is that they offer much less number of gates which are much slower. Thus, the advantage of FPGA chips is the flexibility, while the disadvantage is the performance. Very roughly, we can position machines based on FPGA chips between general-purpose computers and fully specialized computers such as a GRAPE for gravity.

There have been numerous projects to use FPGAs for high-performance computing. One of the pioneers, and probably most well known, is Splash [BA96]. So far, these projects have been successful only in rather limited application area such as the sequence matching for DNA. The primary reason for this limited success is that FPGAs have not been large enough to implement floating-point arithmetic units with standard 64- or even 32-bit precision. In fact, the FPGA chips which will be available in the near future will not be large and fast enough that we can implement pipeline logic such as what we use for gravity processor in our project.

In our project, we will use FPGA chips to implement the short-range particle-particle interactions which are less costly than the long-range interactions such as gravity or Coulomb interactions. There are a number of applications where such short-range interactions are important. One example is the classical molecular dynamics, where particles (atoms) interact with others through bonding force (chemical bonds), non-bonding short-range forces (van der Waals force) and long-range force (Coulomb force).

Other examples include smoothed particle hydrodynamics (SPH), element-free Galerkin method and other "gridless" methods to handle continuous systems. In these methods, partial differential equations are discretized through distribution and interaction of particles. This approach allows us to achieve adaptivity and parallelization relatively easily, compared to traditional finite-difference or finite-element methods based on regular or irregular grids. On the other hand, the gridless nature of these methods implies inherently higher calculation cost compared to grid-based methods. The reconfigurable computer for particle-based simulation would offer a cost-effective solution for these methods.

## 3 Overview of Research Plan

We first develop a massively-parallel special-purpose system for gravity and Coulomb interaction. The peak speed of this part will be around 100 Tflops. For this system, we will develop a custom LSI processor chip. At the same time, we will develop a reconfigurable multi-purpose machine with the overall architecture specialized for the evaluation of particle-particle interaction, but using reconfigurable FPGA chips as the processing element. The speed of this reconfigurable system will depend strongly on the type of applications, but for typical applications with relatively low accuracy requirement it would offer around 1 Tflops. The total system will connect

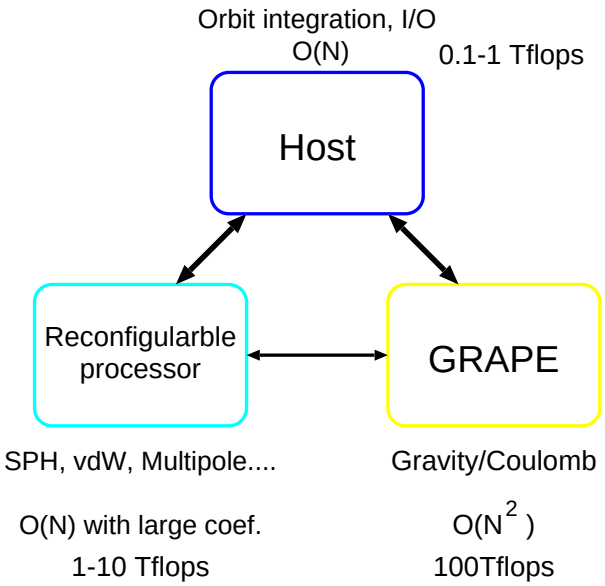


Figure 2: The architecture of the whole system

these two specialized systems with general-purpose frontend (Figure 2). We will apply this machine to astrophysical $N$-body simulations, classical MD (molecular dynamics) simulations, and other applications such as particle-based hydrodynamical simulations such as SPH.

In the following, we give a brief overview of the gravity/Coulomb processor and reconfigurable multi-purpose system.

### 3.1 Gravity/Coulomb processor

The current goal of the peak speed of the gravity/Coulomb processor is 100 Tflops. In order to achieve this performance, we will use around 3,000 processor chips with the peak speed of around 33 Gflops each. The processor chip is manufactured by 0.25$\mu$m process and integrates six pipeline processors for gravity/Coulomb interaction on chip. It operates at the clock speed of around 92 MHz, and achieved the peak speed of 31 Gflops.

These 3,000 processor chips are organized as follows. First, four chips with their local memories are mounted on what we call "processor modules", with the size of 130 mm by 90 mm (figure 3).

Eight of these processor modules are then mounted to the processor board shown in the left panel of figure 4. Eight of these processor boards are connected to the host computer through a two-stage tree network made of three network boards shown in the right panel of figure 4.

These boards are housed in a card cage shown in figure 5. We plan to build 12 to 16 racks in total, depending on the budget we will be awarded in FY2001.

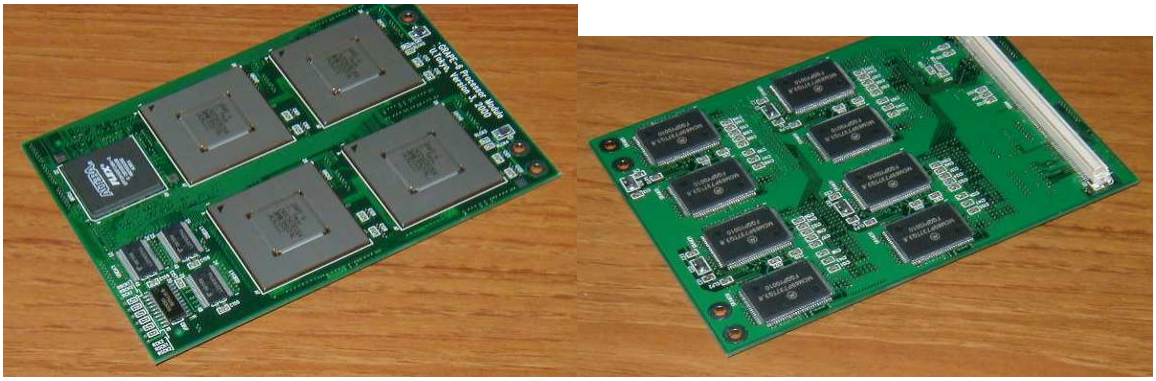

Figure 3: The top (left) and bottom (right) views of the GRAPE-6 processor modules. Four large chips on the top side are the processor chips. Eight chips on the bottom side are memory chips.

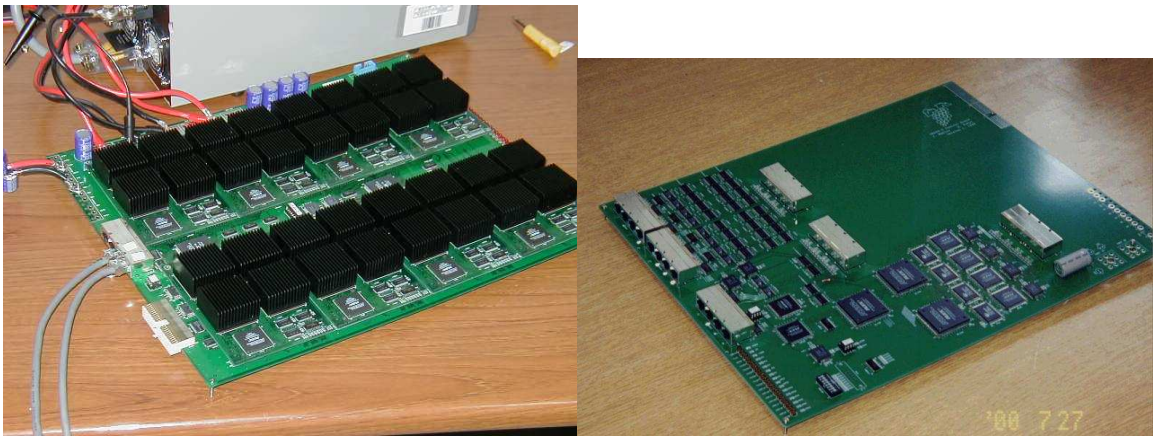

Figure 4: Processor board (left) and Network board (right).

Figure 5: A card cage with four processor boards installed.

### 3.2 Multi-purpose system

For this part, it is difficult to specify a definite goal for the peak speed, since the performance varies depending on the applications, in particular the required accuracy. For the multi-purpose reconfigurable processor we will develop in our project, we place the following goals:

- For the processor chip, we will use the FPGAs with nominal gate count of 200K to 500K. The target clock speed for typical applications will be 50 to 75 MHz.
- The total number of the chips in the system will be 500 to 1000.

When we require standard 32-bit floating-point accuracy, the speed of one chip will be around 1 Gflops, and that of the total system 1 Tflops. This is the worst case performance and if the required accuracy is lower we can achieve much higher performance.

## 4 Research Achievements in FY2000

This FY, we have achieved most of what have been originally planned when we started the project, though there has been some delays in achieving the quantitative goals, mainly due to the reduction of the research budget for this FY. The achievements in this year are summarized as follows:

- Completion of processor board and network board for gravity/Coulomb processor

Figure 6: Prototype system with eight processor boards

.

- Detailed design of multi-purpose system
- Study of new method to accelerate Barnes-Hut treecode with special-purpose computer
- 2000 Gordon Bell Prize for performance

In the following, we briefly describe each of them.

### 4.1 Processor board and network board for gravity/Coulomb processor

The most important achievement of this year is that we have completed the design and the testing of the prototype system consisting of multiple processor boards and network boards. Figure 5 is the prototype system with 4 processor boards and one network board. This system uses our first prototype version of the processor board with the maximum capacity of 16 processor chips. Thus, the theoretical peak speed of this prototype processor board is about 500 Gflops, and the peak speed of the system shown in figure 5 is 2 Tflops. We have successfully the operation with up to eight boards (figure 6), for the peak speed of 4 Tflops.

### 4.2 Detailed design of the multi-purpose system

We completed the design of the multi-purpose system. In short, the physical design of the multi-purpose system will be exactly the same as that of the gravity/Coulomb processor, except for the processor module. The processor module of the gravity/Coulomb processor houses 4 GRAPE-6 custom processor chips and memory chips. Instead of the GRAPE-6 processor chips, we will use FPGA chips from Altera (ACEX 2K series) as the processor chips for the processor module of multi-purpose system. The architecture of the FPGA module is similar to the internal design of the GRAPE-6 chip, with memory interface and host interface implemented on a specialized FPGA chip (controller chip) and pipeline units implemented to multiple FPGA chips on module. The sample of FPGA module will be delivered to us in spring 2001.

### 4.3 Study of new method to accelerate Barnes-Hut treecode with special-purpose computer

We have developed the pseudo-particle multipole method ($P^2M^2$)[Mak99], a novel method to implement high order multipole expansion using particles. As we have already described in previous reports, $P^2M^2$ allows us simple and efficient implementation of both the Barnes-Hut tree algorithm and FMM (fast multipole method), which allow us to reduce the calculation cost of long-range interaction from $O(N^2)$ to $O(N \log N)$ or at least theoretically $O(N)$. In practice, the calculation cost of these methods, in particular when high accuracy is required, is still high. A important advantage of $P^2M^2$ is that it allows us to use special-purpose hardwares designed to calculate particle-particle interactions to evaluate higher order multipole expansions. In other words, we can use GRAPE to implement high-order tree algorithm.

One limitation of the present combination of $P^2M^2$and GRAPE is that there is no efficient way to implement locally adaptive timesteps (sometimes called as individual timestep). In many astronomical systems such as star clusters and galaxies, the orbital timescale of particles (stars) in them varies by many orders of magnitudes, and simple algorithms in which all particles share the same timestep cannot be used. However, it has been difficult to combine tree or FMM with individual timestep, and no algorithm has been known to combine individual timestep, tree algorithm and GRAPE.

We devised a new algorithm, which allows us to combine all three in an efficient manner. It requires small modification of the GRAPE architecture, The overall structure of GRAPE-6 would need no change. The processor module need some change, and we will test the feasibility of this algorithm with FPGA-based system before we actually develop a dedicated system for tree algorithm.

### 4.4 2000 Gordon Bell Prize for performance

We are awarded this year's Gordon Bell Prize for performance for the simulation of black holes in the core of a galaxy performed on the eight-board system (figure 6). The number of particles used

was 768k, and the speed achieved was 1.349 Tflops. This is the first time that the performance number for Gordon Bell Prize exceeded 1 Tflops. The Gordon Bell Prize for performance is given to the calculation with highest sustained speed for real scientific applications. In other words, our eight-board prototype system is already recognized as the machine with the highest sustained speed for scientific applications.

# 5 Future Plan

The next FY, FY2001, is the final year for the present project. The goal for FY 2001 is to complete gravity/Coulomb system, to mass produce the FPGA modules for multi-purpose system, and finally to connect them to general-purpose MPP in Tsukuba.

# 6 Presentation and Publication

## 6.1 International Conferences


[1] J. Makino, *Direct Simulation of Dense Stellar Systems with GRAPE-6*, Star 2000: Dynamics of Star Clusters and the Milky Way, Heidelberg, Germany, March, 2000

[2] T. Fukushige, A. Kawai, and J. Makino, *Star Clusters Simulations Using GRAPE-5*, Star 2000: Dynamics of Star Clusters and the Milky Way, Heidelberg, Germany, March, 2000

[3] Y. Funato and J. Makino, *On the disruption time-scale of sub star systems*, Star 2000: Dynamics of Star Clusters and the Milky Way, Heidelberg, Germany, March, 2000

[4] Koji Takahashi, *The Dynamical Evolution of Globular Clusters and the Stellar Initial Mass Function*, Star 2000: Dynamics of Star Clusters and the Milky Way, Heidelberg, Germany, March 2000

[5] S. Portegies Zwart, S. L. W. McMillan, J. Makino and P. Hut, *Dissection of Open Star Clusters*, Star 2000: Dynamics of Star Clusters and the Milky Way, Heidelberg, Germany, March 2000

[6] J. Makino, *N-body simulations on GRAPE-6*, Stellar Collisions, Mergers and their Consequences, New York, USA, May, 2000.

[7] S. Portegies Zwart, J. Makino, S. L. W. McMillan, and P. Hut, *Runaway Collisions in Star Clusters*, Stellar Collisions, Mergers and their Consequences, New York, USA, May, 2000.

[8] J. Makino, *GRAPE Project: special-purpose computers for many-body simulations*, Formal Methods for HW/SW design for Grand Challenge Scientific Applications, Ischia, Italy, June, 2000.

[9] T. Fukushige and J. Makino, *Structure of Dark Matter Halo through Hierarchical Clustering*, Constructing the Universe with Clusters of Galaxies, Paris, France, June, 2000

[10] T. Sensui, Y. Funato, and J. Makino, *Evolution of Clusters of Galaxies: Growth of Common Halos* (poster presentation), The Physics of Galaxy Formation, Tsukuba, Japan, July, 2000

[11] T. Nakano and J. Makino, *On the Cusps around Central Black Holes in Elliptical Galaxies*, The Physics of Galaxy Formation, Tsukuba University, Japan, July, 2000

[12] T. Fukushige, *The Structure of Cold Dark Matter Halos* VC3: Victria Computational Cosmology Conference, Victoria, Canada, August, 2000

[13] University of Tokyo, GRAPE project, Research Exhibit, SC2000, Dallas, USA, November, 2000

[14] J. Makino, T. Fukushige and M. Koga, *A 1.349 Tflops simulation of black holes in a galactic center on GRAPE-6*, SC2000, Dallas, USA, November, 2000.

### 6.2 Domestic Conferences

[1] T. Fukushige and J. Makino, *Structure of Dark Matter Halos*, Annual Meeting of the Astronomical Society of Japan, April 2000.

[2] J. Makino, M. Koga, A. Kawai, T. Fukushige and M. Taiji, *Processor board for GRAPE-6*, Annual Meeting of the Astronomical Society of Japan, April 2000.

[3] Y. Funato,T. Sensui and J. Makino, *On the growth rate of the common halo in a cluster of galaxy*, Annual Meeting of the Astronomical Society of Japan, April, 2000

[4] T. Sensui, Y. Funato, and J. Makino, *Dynamical Evolution of Galaxies in a Cluster Environment*, Annual Meeting of the Astronomical Society of Japan, April, 2000

[5] K. Takahashi and T. Sensui, Growth of Common Halos in Clusters of Galaxies: Fokker–Planck Models, Annual Meeting of the Astronomical Society of Japan, October 2000.

[6] J. Makino, M. Koga, T. Fukushige and M. Taiji, Current Status of GRAPE-6, Annual Meeting of the Astronomical Society of Japan, October 2000.

### 6.3 Papers

[1] T. Fukushige and D. C. Heggie, *The Time Scale of Escape from Star Clusters*, MNRAS, 318, 753-761 (2000)

[2] Koji Takahashi, Simon F. Portegies Zwart, *The Evolution of Globular Clusters in the Galaxy*, ApJ, 535, 759-775 (2000)

[3] Koji Takahashi, Hyung Mok Lee, *Evolution of multimass globular clusters in the Galactic tidal field with the effects of velocity anisotropy*, MNRAS, 316, 671-683 (2000)

[4] T. Hamada, T. Fukushige, A. Kawai and J. Makino, *PROGRAPE-1: A Programmable, Multi-Purpose Computer for Many-Body Simulations*, PASJ, 52, 943-954 (2000)

[5] A. Kawai, T. Fukushige, J. Makino and M. Taiji, *GRAPE-5: A Special-Purpose Computer for N-Body Simulations* PASJ, 52, 659-676 (2000)

### 6.4 Review articles etc.

[1] K. Takahashi, *Disruption of Globular Clusters in the Galaxy*, Parity, No. 5, 2000, pp.4-10 (in Japanese).

## References


[BA96] Buell D. and Arnold J.M.and Kleinfelder W. (1996) *Splash 2: FPGAs in a Custom Computing Machine*. IEEE Comp. Soc. Press, Los Alamitos, CA.

[Mak99] Makino J. (1999) Yet another fast multipole method without multipoles — pseudo-particle multipole method. *Journal of Computational Physics* 151: 910–920.

[MTES97] Makino J., Taiji M., Ebisuzaki T., and Sugimoto D. (1997) Grape-4: A massively parallel special-purpose computer for collisional n-body simulations. *The Astrophysical Journal* 480: 432–446.