The Future of Supercomputing

Junichiro Makino[1]

[1] Department of Astronomy, University of Tokyo

The advance of the supercomputer technology has been the main driving force for the advance of the computational science. However, the direction of the evolution of high-performance computers for the last fifteen years, and its likely direction for the next ten years or so, seem to have some serious problems. In this talk, I'd like to give an overview of the evolution of computers for high-performance computing, from the point of view of one of the designers of special-purpose computers for astrophysical N-body simulations, GRAPE.

In the simplest, and probably most practical, definition, supercomputers are simply the fastest computers of their time. In 1960s, that really meant just a fast computer, such as CDC 6600/7600. In 1970s, vector processors such as Cray-1 becomes the representative of supercomputers. This period corresponds to the time at which it became possible to make just one fully pipelined floating-point multiplier unit for reasonable cost. In 1980s, parallel vector architectures such as that o Cray X(Y)/MP or parallel pipeline vector architectures as that of Hitachi, Fujitsu, NEC appeared. At the same time, many startup companies made a wide variety of parallel computers, some based on commodity microprocessors but mostly based on custom-designed processors. In 1990s, vector architecture started to decline, and almost all companies of parallel computers went out of business. Practically speaking the only architectures still viable by now (2005-6) are PC clusters and clusters based on scalar microprocessors. Parallel vector machines still exist, but do not dominate the HPC market.

The reason for the decline of the vector architecture and parallel computers other than PC clusters is quite simple. Their cost-performance (C/P) ratio is orders of magnitude worse than that of PCs. That was not so 30 years ago. In 1975, Cray-1 offered C/P about 60 times better than that of PDP-11. In 2005, the C/P of SX-8 is about 100 times worse than that of a PC with a Pentium D processor chip. In 1975, Cray-1 was better than other computers for almost any program, even if it was not vectorized, and if vectorized well it would be even better. In 2005, as far as C/P is concerned, even for a 100% vectorized and parallelized program, SX-8 is worse than a PC if the efficiency of a PC is better than 1%.

The difference in C/P comes from the efficiency of the use of transistors. In 1975, the vector architecture was more efficient in transistor usage. In 2005, it is not. Vector architecture was good when the total number of transistors for a computer was 1 million. If the number of transistors is enough to make millions of arithmetic units, we need something else. The architecture of PCs is optimized for the technology of around 1990, when one chip could integrate 1 million transistors.

However, PC clusters do have two serious problems, from the viewpoint of the users. One is that the intercommunication network is slow and therefore the development of programs tend to be very difficult. The second problem is that its architecture was good 15 years ago, but not at all good now. Now, less than 0.1% of transistors on a processor chip is used for arithmetic operation. Clearly, a new architecture, which allows us more efficient use of transistors. is necessary.

An extremely important lesson that the past history of supercomputers teaches us is that such new architectures generally did not come from computer manufacturers or computer scientists, but from computational scientists, or users. Two important examples are distributed-memory parallel processors based on microprocessors (several QCD machines and PAX computers and VPP in Japan.), and PC clusters. In other words, the ideas from users will determine the future of supercomputing. I will discuss several ideas which might help the advance of supercomputing in the next decade.