3

Recent Trends in Supercomputers

MASAO WATARI
Information and Communications Research Unit

3.1 Introduction

After the first supercomputer was developed in the 1970s as high-end computers having fastest computing capabilities, supercomputers have been steadily improving in their performance as shown in Fig. 1. According to the TOP500, the list of the world's fastest supercomputers* released in June 2001 (refer to our August, 2001 issue of Science and Technology Trends), ranking No. 1 was ASCI White, which was developed in the U.S. as part of the ASCI project and has a peak speed of 12.3 teraflops*[1]. In Japan, on the other hand, the Earth Simulator, a supercomputer with a 40 teraflops speed, is expected to start running in early 2002.

Meanwhile, driven by the growing PC market, development of general-purpose processors is vigorously pushing forward, showing remarkable performance improvements. Today's workstations and PCs have computing power equivalent to that of supercomputers ten years ago. For their cost advantages, workstations and PCs are often used even for scientific and technological computation recently.

However, supercomputers are still playing important roles as massive-scale computations are required in many basic science research fields, where supercomputer performance enhancement is the key for the area's research advancement. Supercomputer technology is vital not only for advancing a computer technology but also for providing an essential tools to advance basic science research.

Thus the environment surrounding supercomputers is changing rapidly. This report describes trends in performance and applications of supercomputers and then, discusses the future roles and prospects of supercomputers with a view of the trends in their applications

3.2 Supercomputer hardware trends

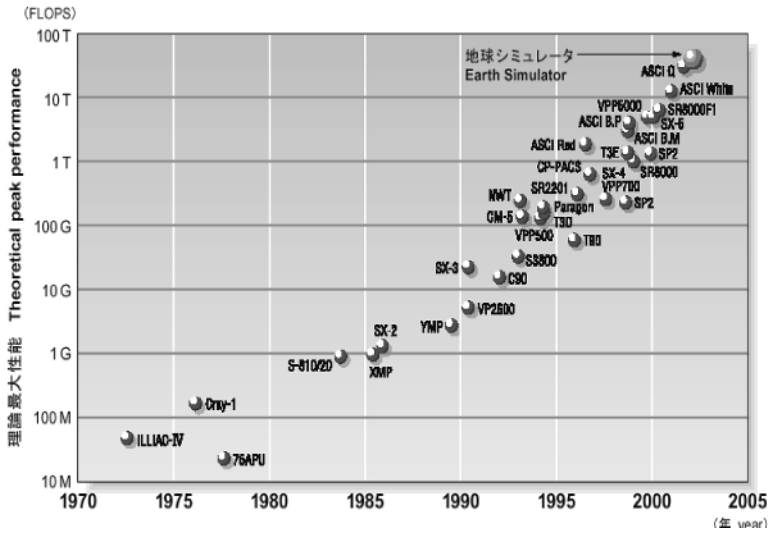
3.2.1 Supercomputer performance developments

The history of supercomputers started in 1971, when ILLIAC IV was developed at the University of Illinois, followed by the shipment of the first commercial model Cray-1 by Cray Inc. in 1976. In the early 1980s, supercomputers began their rapid development in the U.S., and Japanese makers followed. In the U.S., in the 1990s, vector supercomputers declined*[5], and scalar parallel supercomputers were developed by adopting general-purpose processors which are used for workstations and PCs. On the other hand, in Japan, performance of vector supercomputers was further enhanced, because processor speed was more important than massively parallel processing for fastest computers.

Since the supercomputer market is not very large despite high development costs for supercomputers, the market is affected by large impacts from funds granted through national projects. Described below are the latest, ongoing supercomputer development projects in Japan and the U.S.

* While supercomputers are often referred to as high performance computers (HPC) these days, the more popular term "supercomputer" is used in this report.

Figure 1: Supercomputer performance developments



Source: Earth Simulator Research and Development Center

(1) The ASCI (Accelerated Strategic Computing Initiative) project in the U.S.

Through the ASCI project, which started in 1995, three laboratories under the umbrella of the Department of Energy (DOE) are working on developing the highest-performance supercomputers while providing support for hardware manufacturers.

By investing about ¥140 billion over ten years, they are aiming to realize a peak speed of 100 teraflops by 2004. In October 1996, the ASCI Red system achieved 1.8 teraflops, followed by the ASCI Blue, which achieved 3 teraflops in October 1998, and the ASCI White, which provided a 12.3 teraflop performance in June 2000. Furthermore, in 2002, ASCI Q, a supercomputer aiming at 30 teraOPS (about 24 teraflops), is scheduled to be delivered.

To achieve such high processing speeds, the ASCI project employs a large number of general-purpose processors connected in a parallel configuration to build a massively parallel scalar supercomputer. For ASCI White, as many as 8,192 processors are connected in parallel. Because of their system's massively parallel structure, emphasis is also placed on the development of software that can efficiently execute parallel processing.

(2) The Earth Simulator project in Japan

The Earth Simulator, a five-year project with a ¥40 billion budget, was started in 1997. In this project, a 40 teraflop supercomputer is being developed to elucidate and predict global climate changes such as global warming, unusual weather, and meteorological disasters. The project seeks to find solutions to global environmental problems as well as to contribute to developing measures against natural disasters by creating a "virtual Earth" model on the computer.

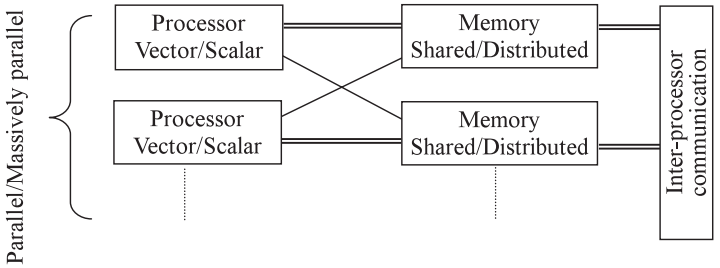
The Earth Simulator uses 5,120 vector processors connected in parallel, shared memory, and a high-speed network interface technology. Vector processors are adopted, because vector parallel computers have five to ten times the efficiency of massively parallel scalar computers, and Japan is leading the vector technology field. As of the beginning of 2002, when it starts operating, the Earth Simulator will be the supercomputer with the world's highest performance.

(3) Difference in strategies for development between Japan and the U.S.

The U.S. is focusing on increasing throughput by parallel processing rather than improving per-processor computing speed, because in the high-performance computer (HPC) market, the business sector (for transaction processing, etc.) market is larger than the scientific and technological computing market. The overall performance has been enhanced by implementing more than 1,000 general-purpose scalar processors in parallel. On the other hand, Japan is focusing on enhancing vector processors

Figure 2: Supercomputer performance factors and features

Vector processors	Designed specifically for scientific and technological computing to allow high-speed calculations.
Scalar processors	Low-priced general-purpose processors. Their performance has been remarkably enhanced.
Parallel architecture	Higher-class supercomputers having multiple processors in parallel.
Massively parallel configuration	Massively parallel configuration has more than 1,000 processors. Development of software for massively parallel processors is crucial.
Shared memory	Multiple processors shares the same memory. Memory access is high speed and a broad range of data can be accessed.
Distributed memory	Each processor has its own memory. Access to other processor's memory is made via inter-processor communication, and thus slow. PC clusters use the distributed memory system.
Inter-processor communication	High-end supercomputers adopt dedicated high-speed networks, and on the other hand PC clusters adopt general-purpose LAN.



continuously, which are more suited for scientific and technological computing. This leads to enhance computing capability with suppressing increase of number of processors in parallel. This superior vector processor technology is viewed as one of the few computer technologies that Japan has a leading edge over other countries.

However, performance of a supercomputer is determined not only by the processor speed but also by the memory architecture and data transfer method between processors. There are many components which may cause high price. Also note that as every supercomputer has processors in parallel with various scale, development of efficient parallel processing method is crucial. Figure 2 summarizes supercomputers' performance factors and their features.

3.2.2 Emergence of the PC cluster system

As the computing capabilities of PC processors increase, PC clusters, which are enabled by connecting a large number of high-performance PCs via a high-speed network, have emerged. A PC cluster reaches the same performance level of a supercomputer at a lower investment because it allows massive PCs running simultaneously in high speed network. The technology has already been commercialized in the U.S, while in Japan the Real World Computing Partnership (RWCP) is working on research and development with a plan for commercialization.

As a PC cluster with 1,000 processors that achieves 550 gigaflops[*1] on the Linpak benchmark has already been realized, further performance upgrades are expected as processor performance increases. While PC clusters are not yet as fast as high-end supercomputers, their excellent cost effectiveness is worthy of attention. However, their performance is enabled only when applications allow parallel processing and limiting data transfer between PCs. Therefore, PC cluster systems will find a wide range of applications in areas where parallel processing gives an effective solution, for example, genome information processing or web search engines.

3.2.3 Dedicated computers

A processor specifically designed to perform specific computations at a high speed can contribute to building a low-cost computer that has performance equivalent to or beyond a supercomputer. For example, there is a computer with a dedicated LSI to calculate inter-particle dynamics and data matching. It can execute

celestial simulations, protein structure analysis, and other complicated calculations at a high speed.

Challenges concerning this technology are the need for specifically developed software and continuous competition with general-purpose processors whose performance continues to improve.

3.3 Grid computing: a new move in networked computing technology

As an example of widely distributed computing system over the Internet, there was a demonstration to prove the possibility of large-scale computing by gathering participating PCs' idle time through the peer-to-peer technology*2. In the case of a project led by the American Cancer Society and others with a mission to discover candidate substances to cure leukemia, 900,000 PCs took part to achieve effects equivalent to a world's top-class supercomputer (4 teraflops) in terms of provided CPU hours during a six-month period, assuming the average participating computer performance was 50 megaflops. While attracting attention as an attempt to compete with expensive supercomputers, this effort showed the limitations of such systems: a computation must be completed inside a participating PC and it must need a large number of contributors who are willing to offer their computing power for free or at a low price.

Meanwhile, a concept of providing unified computing power by connecting multiple computing centers was proposed in 1995. This scheme is called as grid computing in analogy to a power grid in which people can use energy without knowing where it is generated. It allows use of a supercomputer via a network but does not enhance the original computing power. The advantage of this technology is that scientists can form a networked community through exchanging and communizing their computing resources (computing power and data), so that their research projects can be accelerated through the use of shared software and data. There are many projects underway in the U.S. and Europe. U.S. examples include the National Technology Grid for NSF infrastructure construction, the Grid Physics Network to create a huge physics database, and the Information Power Grid to provide NASA with a seamless computing environment. EU has established its European Data Grid designed for high-energy physics, earth observation, and biotechnological research. The British government is, as part of its e-Science program, pushing forward a number of joint projects focusing on applying grid technologies, together with a project to build an infrastructure to support them. Meanwhile, the Global Grid Forum is taking the initiative in international standardization of grid technologies.

These grid technologies are not a mere application technique of supercomputers but a scheme that, in the future, will allow any person to access and use networked information resources from anywhere, anytime. They can be the next-generation Internet technology, if challenges such as resource management, security, privacy, and copyright problems are successfully overcome.

3.4 Supercomputer application fields

To provide an overview of the future needs for supercomputers, how they are currently used in major application fields, and what are their future goals will be discussed. The overview is summarized in Table 1.

(1) Prediction of Weather, Climate changes, and the Global environmental change

Today's weather forecasts are significantly improved in accuracy by using supercomputer simulation in addition to the conventional empirical method. For the further improvement, it needs a computer simulation using a finer mesh model. For example, local-level forecasts for concentrated heavy rains require the description of cumulous clouds with a resolution down to a few kilometers. If the mesh resolution is refined by one tenth, required computing power increases 1,000 times.

Regarding climate changes, it is known that global-level weather phenomena, such as the El Nino, Asian monsoon, and Aleutian low, have a large impact on the climate changes in Japan. However,

due to the limit of computing capability, the current simulation is performed on individual or regional models of the atmosphere, ocean, earth's surface, and cryosphere. If a global model or an integration model of the atmosphere and ocean can be created, the accuracy of climate change prediction is expected to dramatically increase.

Elucidating the mechanism of global climate changes such as global warming and diminishing forest is a critical issue for the entire world. This research requires massive-scale simulations by using worldwide data. Rendering with a resolution of about 10 kilometers will allow examination of model of the global warming mechanism.

The Earth Simulator, the world's fastest supercomputer that will start operation in 2002, is expected to dramatically facilitate progress in research in this area by allowing scientists to create a "virtual Earth."

(2) Bioinformatics

In the bio technology field, analysis of gene information is making rapid progress supported by supercomputers. In particular, bioinformatics, a combination of bio technology with information technology, is drawing big attention recently. Now that most of the human genome*3 has been analyzed, scientists are shifting their focus to solving protein structures. Their future targets include elucidation of cell differentiation and proliferation, development of medicine to cure malignant diseases, and analysis of biochemical reactions.

Research in bioinformatics requires massive data processing and analysis, for instance, deciphering 35,000 genes from 3 billion base pairs of the human genome, or performing three-dimensional structural analysis to identify 10,000 basic structures and functions of protein generated from genes. While most of such computations can be done through parallel processing, some complicated analyses can not run in parallel. In general, a parallelising such computation may be up to 1,000 elements. In order to use more parallel processors simultaneously, a breakthrough at parallel processing algorithms is awaited.

Protein's chemical reaction mechanism is being partly analyzed, and it can be solved through molecule chemical simulation based on quantum mechanics. Its computation volume is defined in proportion to the fourth power of the molecule size so that approximation methods are being researched in order to reduce the computation volume. Current supercomputers can simulate 100 atoms behavior during a period of nanoseconds. However, analysis of chemical reactions requires computer simulation of phenomena that last some microseconds to milliseconds, scientists are looking forward to seeing petaflop*1 computers.

(3) Materials simulation

In the field of nanotechnology, the structure and properties of materials can be derived by computer simulation. In a quantum mechanical method called First Principle simulation, specific atoms are combined inside a computer to virtually create new materials or structure for investigating new properties. Since this method does not use empirical parameters, a researcher can investigate the characteristics of completely new materials through computer simulations without the need to conduct a preliminary experiment to obtain parameter data.

The latest supercomputers can provide computing power to simulate up to 100 level atoms, standing at the entrance to the nano-scale world. A next challenge is to design a system that needs more atoms to simulate, while allowing more macroscopic analysis of properties and development of new methods for materials generation. As the computing volume increases in proportion to the third or greater power of the atomic mass, research and development to find techniques to reduce the volume is underway.

So far, due to insufficient computing power, computer simulation is limited in its applicable field and thus further development of supercomputers is looked forward greatly.

(4) Structure analysis and Fluid analysis

Structure analysis or fluid analysis is conducted through a numeric analysis where the subject is divided into fine meshes. When the number of meshes becomes huge in case of complicated form, a supercomputer is needed for its calculation.

As a structure analysis example, a car's structure is rendered on a computer to simulate a collision in

order to analyze what happens. A supercomputer can complete such collision analysis in a short time (overnight) so that result of analysis is reflected into designs. A supercomputer simulation enables to speed up the overall process of designing a car with enhanced safety.

Fluid analysis examples include simulation of strong winds blowing through tall buildings and analysis of air resistance to a car or high-speed train. Since recent high-end PCs have enough computation power to simulate simple-shaped subjects, this technique is widely used in the industry applications.

For an accurate analysis of real world phenomena, coupled simulation*[4], in which structure analysis and fluid analysis can be simultaneously performed, is often required for the use of a supercomputer. As examples, coupled simulation are required for design of an aircraft wing (fluid mechanics + structural mechanics), analysis of engine combustion (heat + chemical reaction + fluid dynamics + structural mechanics), and simulation of blood flow in a blood vessel or a heart (fluid mechanics + structural mechanics). Analysis of fluid noises such as those generated by winds against pantographs or by a train entering into a tunnel requires more than one billion grids. Its computation can be done by a supercomputer equivalent to the Earth Simulator. In general a vector supercomputer works more efficiently for fluid analysis. In the case of simple fluid analysis without requiring many computations among remote grids, parallelizing processors is expected to be up to 10,000.

(5) Celestial mechanics and Elementary particle physics

The secrets of the universe, such as formation of the Galaxy from fluctuation in the early universe, and creation of a black hole and planets in the solar system, are researching through computer simulation. By modeling the universe as multiparticle gravitational interaction, scientists are carrying out simulation on a supercomputer or a dedicated computer having high-speed gravity computation capability.

On the other hand, in the area of basic physics that deals with atomic nuclei and elementary particles, supercomputers are used to simulate their behavior in each of a vast number of microscopic grids, to elucidate new properties of atomic nuclei and elementary particles.

Whether the subject is the universe or the atom, researchers use a supercomputer for simulating basic theoretical models to understand phenomena that are extremely difficult to observe. The supercomputer is an essential research tool for them.

(6) Nuclear fusion

Simulation by supercomputers plays an important role in the study of fusion energy, which is expected to become a future energy source. Since there is no complete established theory for fusion plasmas, scientists must analyze simultaneously both particle and fluid level phenomena for a collection of a huge number of particles. They are now building the theory of plasma on the basis of their knowledge obtained through experiments and computer simulations. Simulating behavior of within seconds of several tens of millions of particles takes a few days for a supercomputer. For realizing complete simulation of plasma, electron analysis should be taken into account in addition to the current simulation. To meet this end, computing power will be required as tens of thousand times than the current supercomputer.

(7) Data mining

Data mining technology can derive to find useful knowledge from large volumes of data. It is based on learning models or language processing models developed through the research of artificial intelligence. When data volume is enormous, a high-speed computer is required. For example, the Web search engines collect ever-changing Web site information over the Internet in order to find user required web site by keywords. A supercomputer is used to perform large-scale similarity calculations to figure out. It is rather easy to parallelize the calculations for this kind of applications. Small-scale data mining can be handled by high-performance PCs.

(8) Economic forecasting and Financial engineering

It is widely known that mathematical theories are applied for developing models for stock and

derivative trading. Also, there are many studies that attempt to provide macro-economic forecasts by simulating interactions among many economic factors on a computer. However, the complexity of economies makes it extremely difficult to verify and justify their theoretical models. In addition, economy is heavily affected by electronic transactions systems worldwide, an event is transferred very quickly, easily causing an unexpected chain reaction.

While leading securities firms used to perform simulations on a supercomputer, their demand for supercomputers has weakened as they began to question the cost efficiency of investing a large

Table 1: Supercomputer application fields and their features

Supercomputer application field	Current applications	Future visions	Computing requirements	Needs
Prediction of Weather and Global environmental change, Geological prospecting	Weather forecasts based on atmospheric models and empirical parameters; Prediction of oil reserves based on geological studies	Elucidating global warming; Predicting unusual weather and other climate changes; Forecasting local-level weather such as concentrated heavy rains.	Vector processing work better on climate computations. Large-scale computation for a fine mesh is required because a local phenomenon affects the overall results.	Mesh refinement is needed to improve the accuracy of prediction. If the grid resolution is refined one tenth, computing power requires 1,000 times (equivalent to a teraflop computer).
Bioinformatics	Gene analysis; Protein structure analysis	New medicine development from the genome data; Customized remedies; Protein's chemical reaction by using quantum theories (computational chemistry).	Calculation of matching, energy and so on sometimes allow for parallel processing, but at times require more complicated computations.	A 1,000-fold increase will enable only microsecond-level analysis. An additional 1,000-fold upgrade is needed for chemical reaction simulation.
Materials simulation	Analysis of structure and properties of materials through first principle simulation on a 100-atom level	Designing and creating new materials and nanostructures; Analyzing their functions and properties (computational physics).	Complete theoretical computation is possible only in a small area. The computing volume is the third or greater power of the atomic mass.	Shared memory systems are suitable as access to a wide range of data is required. The demand for upgrading computing power is strong.
Structure analysis and Fluid analysis	Virtual experiment of car crashes; Analysis of air resistance to aircraft and many other industrial applications	Coupled simulation to solve multiple theoretical models in a unified manner. e.g., engine combustion and biodynamic simulation.	High-performance PCs or workstations can be used for simple structure / fluid simulation.	A petaflop computer is needed for coupled simulation. Easier 3D data entry is called for.
Celestial mechanics and Elementary-particle / nuclear physics	Simulation of Galaxy formation and elementary- particle / nuclear simulation	Finding the secrets of the universe such as formation of the Galaxy and creation of a black hole	Gravity computations for large-scale particles require a teraflop or faster computer.	Dedicated computers are sometimes used.
Nuclear fusion simulation	Simulation of nuclear fusion plasma based on theoretical models and the knowledge obtained through experiments	Gaining energy from nuclear fusion by real-time control based on theories and simulations.	Macroscopic fluid analysis and microscopic inter-particle dynamics are simultaneously calculated.	Complete simulation integrated with electron analysis requires tens of thousands times the computing power.
Data mining	Web search engines; Analysis of customer information or product popularity	Extracting knowledge (meaning, characteristics) from large volumes of data.	Large-scale data calculation based on leaning models and language processing models.	HPCs are required when the data volume is large. Parallel processing is applicable.
Economic forecasts and Financial engineering	Macro economic forecasts; Stock quotation projection	Establishing superior models to forecast economic fluctuations by large-scale complex simulation.	Calculating various interactions based on mathematical theories and probability models.	Price performance is essential for computer simulation. Many users moved to PC systems.

sum of money into a supercomputer when enhanced-performance PCs are widely available.

3.5 Computer simulation trends

In industries, supercomputers are typically used for simulations of structure or fluid analysis in the many fields such as machinery, civil engineering, construction, and electronics. They often contribute to the production design process. A small-scale analysis can be run even on a high-performance PC. In the fields of biotechnology, chemistry, materials and energy, a supercomputer usually serves the research and development section for simulations but not yet for product design.

The capabilities of computer simulation are expanding as supercomputer performance improves. In particular, supercomputers are indispensable when studying something hard to experiment with (high temperature/pressure) or to obtain (new materials), something dangerous to handle (collisions, poisons), and something difficult to observe (elementary particles, atoms, and molecules). Conventional simulation systems provide only simple idealized simulation based on a single phase or steady state model, the next developmental goal for the computer simulations is to enable more realistic or accurate simulation such as coupled simulation, in which a multi scale process is simulated in a unified manner. To achieve this goal, further upgrading of supercomputer performance is anxiously anticipated.

3.6 Software development trends

Software developed for supercomputers is partly dependent on supercomputer types. Science and technological computations often use vector processing, and programs for vector processors are easier to develop while providing efficient operations. However, software developed for vector processors does not deliver full performance on scalar processors. In the U.S., where vector computers have become unavailable, many applications are being rewritten for massively parallel scalar computers.

On the other hand, a parallel computer can achieve its full performance only on a program designed for parallel processing. Thus there are many researches in software for parallel processing such as parallel compilers and parallel communication processing programs that support parallel computing. While the U.S. started working on parallel architecture early, Japan is generally behind in this field.

3.7 Conclusion

Application of supercomputers is more often seen in the area of basic scientific research in addition to industrial fields. In particular, the latest supercomputers are contributing to basic scientific research in producing new developments. For example, as a new field Bioinformatics is emerged combining biotechnology with information technology in the U.S. Supercomputers are playing an essential role in Bioinformatics. Venture businesses are being fostered aiming at even launching a new industry.

Also supercomputers help predict global environmental changes, promoting embodiment of an enriched society. Simulation by supercomputers is a basic technology crucial for both science and industry as a means to quickly and safely solve complicated phenomena occurring in various areas including natural science, engineering, and socioeconomics.

Demand for supercomputers is still strong. Especially in simulation for biotechnology and physics, petaflop supercomputers, that are 100 to 1,000 times faster than the current super-computers, are awaited. So far, supercomputer performance has been upgraded 10-fold in every four to five years. Assuming this trend continues, a petaflop computer is expected to come out by 2010. Of course extensive development of relevant technologies must continue in order to realize this projection. In the U.S., discussion is already taking place about a plan to develop a petaflop computer. This is because they consider supercomputers not only as a key technology for science but also as a defense technology for the nation. Also they recognize that the market is not large enough to provide sufficient funds for development of supercomputers and thus government support is required for nurturing

supercomputer technologies.

Japan maintains unique technologies for supercomputer hardware, which are different from those of the U.S. This superiority should be taken into account when we discuss our strategy for the next-generation supercomputer. Aside from the hardware aspect, application-oriented approaches to identify demands for and uses of supercomputers should also be considered. We must realize that if supercomputers become an exclusive technology of the U.S, it may affect not only the computer field but also the basic science research field.

Japan also has been failed to acknowledge the importance of software, posing a major challenge for us. Dr. Pople won the Nobel Prize for his development of the famous molecule chemistry software called Gaussian, together with Dr. Kohn who established the theory for it. In Japan, where software is not recognized as highly valued, a researcher in science cannot gain a doctor's degree by new software development alone. This makes us realize the need for adding a new point of view to our research evaluation system.

Glossary

*1 gigaflops, teraflops, and petaflops
Giga (G) represents 10^9, tera (T) 10^{12}, and peta (P) 10^{15}. Flop, which stands for floating-point operations per second, is an index to indicate the computing power of a computer. Linpack, a program to solve liner equations, is often used for benchmark tests.

*2 peer-to-peer technology
A technology that enables individuals on the Internet to directly exchange information each other. By using this technology, many computers can be interconnected to share their computing power and files through network.

*3 human genome
The entire human genetic information that provides the basic blueprint of human life. There exist 3 billion base pairs in DNS contained in chromosomes within a cell nucleus, and genes are said to account for 3% of them.

*4 coupled simulation
Also known as multi-disciplinary or multiscale simulation, this term refers to simulating multiple theoretical models in a unified manner. Examples include blood flow simulation, which analyzes fluid and structures at the same time, and fracture analysis, which is simultaneously finding solutions for both the microscopic quantum mechanics theory and macroscopic classical physics.

*5 a decline of vector supercomputers in the U.S.
In the U.S., where mainframe manufacturers did not make supercomputers, specialized producers such as Cray led the supercomputer market as well as the technology. Japanese computer manufacturers, who entered the market in the 1980s, over took their U.S. counterpart by the late 1980s because of their total technological strength that covered even the semiconductor field. In the early 1990s, as mainframes were downsized, no additional companies moved into the vector supercomputer market, while Cray, which was suffering from unstable operation, lost their competitive edge in technology development. At the start of anti-dumping duties in 1996, top-end vector supercomputers became unavailable in the U.S. Under the duties, which were lifted in May 2001, U.S. computer makers focused on development of massively parallel scalar supercomputers.

(Original Japanese version: published in October 2001)