Lessons learned from the K computer project
---from the K computer to Exascale---

Yoshio Oyanagi[1]
Graduate School of System Informatics, Kobe University
Rokko-dai 1-1, Nada-ku, Kobe, 657-8501 Japan

E-mail: oyanagi@people.kobe-u.ac.jp

Abstract. The history of the supercomputers in Japan and the U.S. is briefly summarized and the difference between the two is discussed. The development of the K Computer project in Japan is described as compared to other PetaFlops projects. The difficulties to be solved in the Exascale computer project now being developed are discussed.

1. Brief History of Supercomputers

Since the first half of the 1980s, following the efforts to produce commercial vector computer Cray-1 in the U. S., Japanese computer vendors have been producing a number of powerful supercomputers to solve large scale scientific and engineering problems. The history of supercomputers in Japan before 1999 is described in a previous article[1]. The trends in Japan as well as in the U.S. can be summarized as follows.

1.1. Premordial age (1970s)

The vector processing was first proposed by Senzig and Smith of IBM[2]. It was implemented in the IBM2983 Array Processor (1969) as an attached processor to the I/O channel of mainframe computers. Independent vector computers were developed by two vendors: ASC of Texas Instruments (1972) and STAR-100 of Control Data Corporation (1973). These machines, although pioneering, were not commercially successful. The first commercially successful vector supercomputer was Cray-1 of Cray Research Inc. (1976).

At the same time Japanese vendors were also developing vector computers. One year after the Cray-1, Fujitsu shipped FACOM 230-75 APU (1977), which incorporated three data buffers of 256 words for vector operations. It also supported list-directed vector accesses. Hitachi and NEC made an IAP (Integrated Array Processors) for their top-of-the-line mainframe computers. The performance of these computers, however, was modest as compared to the Cray-1.

As for parallel processing, Burroughs Corporation and University of Illinois developed a pioneering parallel computer ILLIAC IV (1973). Burroughs started to develop BSP (Burroughs Scientific Processor) but the project was cancelled in 1980. It is to be noted that the ICL in U.K. developed DAP machine, which consists of one-bit processors (1979).

1.2. Vector age (1980s)

---

[1] To whom any correspondence should be addressed.

Full-fledged vector supercomputers were announced and produced in the 1980's. In the U.S., CDC produced Cyber-203 and Cyber-205 (1981). Cray Research developed Cray XMP-4 (1984), Cray-2 (1985) and Cray YMP (1988). ETA, a subsidiary of CDC, developed ETA-10 (1987). IBM developed 3090 VF (1985). There were also several minisupercomputers, such as Convex C1 (1985) and C2 (1988) or Supertek S1 (1989).

Japanese vendors shipped high performance vector computers: Hitachi's S810 (1983) and S820 (1987); Fujitsu's VP200 (1983) and VP2600 (1989); and NEC's SX-2 (1985) and SC-3 (1990). In contrast to the U.S. vector computers, Japanese machines have the following features:

1) compatibility with the mainframe computer
2) single processor with many vector pipes
3) large main memory
4) large vector registers
5) list-directed vector operation

At the same time, a considerable number of parallel venture companies emerged in the U.S. On the other hand, although there were some parallel computing research activities in Japan, they were mainly in academia and were considered as special-mission processors.

1.3 Vector parallel and commodity parallel (1990's)

Parallelism became the key trend in both vector and scalar machines in the 1990's. In the U.S., Cray Research shipped several parallel vector computers: C90 (1991), T90 (1995) and SV1 (1998). At the same time, several U.S. companies produced parallel computers using commodity processors: TMC's CM-5 (1992), Convex' SPP (1994), IBM's SP1 (1993) and SP2 (1994) and Cray Research's T3D (1993) and T3E (1996) and many others. It is to be noted that the highly parallel machines constructed in the DoE's ASCI Program were also based on commodity processors.

In contrast, the mainstream of Japanese supercomputers was based on parallel vector architecture: Hitachi's S3800 (1993); Fujitsu's NWT (Numerical Wind Tunnel, 1993) and its commercial version VPP500 (1993); and NEC's SX-4 (1995). At the same time, Japanese companies developed parallel machines with scalar processors: Hitachi's cp-pacs (1996), SR2201 (1996) and SR8000 (1998); Fujitsu's AP1000 (1994) and AP3000 (1997); NEC's Cenju-2 (1993), Cenju-3 (1994) and Cenju-4 (1997). Some of those machines were sold as a test bed rather than a production machine. The NWT, SR2201 and cp-pacs occupied top positions in some of the Top500 lists, which started in 1993.

1.4 Massively parallel era (2000's)

Massively parallel machines become the main trend in the 2000's. In the U.S. from the end of the 1990's, the ASCI Program (later called ASC Program) continued to build a number of massively parallel machines based on commodity processors: ASCI Red (1997), ASCI Blue Pacific (1998), ASCI Blue Mountain (1998), ASCI White (2000), ASCI Q (2002) and ASCI Purple (2005). IBM's Blue Gene/L (2004) and Cray's Red Storm (2005) later joined the ASC Program. The DARPA started HPCS (High Productivity Computing Systems) in 2002 and the NSF started the TeraGrid Project in 2001. All the machines developed in those projects were sold as commercial products to universities and laboratories as well as to industry sectors.

The biggest news in Japan in this period is the completion of NEC's Earth Simulator in 2002, which occupied the top position in five consecutive Top500 lists (2002-2004). The emergence of the huge vector supercomputer stimulated the U.S. and doubled the U.S. supercomputer budget. In contrast, Fujitsu left vector business and adopted massively parallel computers based on the Sparc64 and x86 processors.

1.5 Observations

In the last decades of the 20th century, U.S. and Japan were the major producers of supercomputers. There was, however, big difference between the policies of the two countries.

Until late 1990's, Japanese vendors focused on vector machines and users in Japan enjoyed the power of vectorization. Besides hardware, vendors provided very good vectorizing compilers, so that users were in a sense spoiled. In Japan, vendors thought that parallel machines were for specialized purposes (eg. image processing) and users dared not try to harness parallel machines in the 1980's. Although some computer scientists in Japan were interested in building parallel machines, they were not used for real applications in science and technology. Around 2000 Japan was at least ten years late in parallel processing for scientific computing as compared to the U.S.

It is to be noted that practical parallel processing for scientific computing in Japan was started by application users: qcd-pax (1989), NWT (1993), GRAPE series (1989-) and the Earth Simulator (2002).

2. Challenge to PetaFlops

2.1 PetaFlops Conferences

As early as in 1994, when the NWT took No. 1 in the Top500 list with 124 GFlops Linpack performance, a conference aiming at the PetaFlops (10^6 GFlops) was held in the U.S.[3] After several workshops, the PETAFLOPS II conference was held in Santa Barbara in 1999[4].

2.2 Reluctant Japan

At that time, there were no significant activities in Japan towards PetaFlops. The Japanese government established the IT Strategic Headquarter in 2001 and promoted high speed Internet and inexpensive Internet services, but the levelling up of supercomputers was not considered as a national project. They thought that the supercomputers may be constructed according to the needs of each field.

2.3 After the Earth Simulator

Only after the success of the Earth Simulator in 2002 the discussion started on a possible supercomputer in the PetaFlops region. The Information Science and Technology committee in the Mext (Ministry of Education, Culture, Sports, Science and Technology) started discussion on the possible measures to promote computational science and technology from August 2004.

The recommendation of this Mext committee was to promote a national project to construct a leading edge supercomputer. The Mext decided to start this project on July 25, 2005 and in October the Riken (The Institute of Physical and Chemical Research) was selected the developer of the supercomputer. The working group in the Mext identified ten killer applications: life science, astrophysics, space and aeronautics, materials, atomic energy, environment, disaster prevention, fluid dynamics, plasma and industrial design. The observation of the working group was that the multiphysics-multiscale simulation would be important in those fields and a hybrid architecture was appropriate to keep high performance in different types of computing. The author agreed on the first point that multiphysics simulation is important, but it did not mean that a hybrid architecture was suited for such simulation.

2.4 Conceptual design

The original proposal was a hybrid of three parts: scalar part (1 PetaFlops), large scale processing (vector) part (0.5 PetaFlops) and special-purpose processor part (20 PetaFlops). On September 2007, after long discussion, the Riken finalized the conceptual design of the hardware. The system consists of two parts: scalar processor part and vector processor part. The scalar processor was to be built by Fujitsu and the vector part by NEC and Hitachi. The total Linpack performance was to exceed 10 PetaFlops.

The site selection was another big issue. Out of 15 proposals, Riken finally decided to build it in Kobe.

The strategic committee in the Mext identified five strategic fields in SPIRE (Strategic Programs for Innovative Research) on July 22, 2009: 1) Predictive bioscience, medicare and drug

design, 2) New material and new energy, 3) Earth environmental prediction for disaster prevention and mitigation, 4) Next-generation manufacturing, 5) Origin and structure of material and universe.

2.6 Outside Japan
PetaFlops was a target in various part of the world. In Los Alamos National Laboratory, the Roadrunner attain 1.026 PetaFlops in 2008 using the Cell processor as accelerators. It was the first supercomputer which exceeded 1 PetaFlops Linpack performance. In Oak Ridge National Laboratory, the Jaguar attained 1.059 PetaFlops in 2008. It is a homogeneous system like the K Computer.

In China there came several Petascale supercomputers. Nebulea in Shenzhen attained 1.271 PetaFlops with NVIDIA GPU. Tian-he 1A in Tianjin Supercomputer Center got 2.566 PetaFlops in Linpack.

2.7 Government revitalization unit
Unfortunately, NEC and Hitachi retired from the project due to bad economy on May 13, 2009. The Riken, however, decided to continue the joint development with Fujitsu to build 10 PetaFlops machine using scalar processors only on May 14, 2009.

A general election of Japan was held on August 30, 2009 for the lower house of the Diet of Japan and the opposition Democratic Party won the majority. The Democratic Party government introduced a budget screening by the Government Revitalization Unit. On Friday, November 13, 2009, the third working group of the Government Revitalization Unit examined the construction budget of the supercomputer. The budget proposal stressed that the Japanese supercomputer would win the No. 1 when it is completed. The chairperson of the working group, Ms. Lien Fang Murata, asked the Mext staff, "Why should it be No. 1 in the world?" "Is No. 2 not enough?" Another criticism was about the demise of the vector part in the project.

After some discussion the conclusion of the working group was "to freeze the project." This conclusion, although not final, met strong reactions from both academia and industry. Consequently, Government decided to overturn the conclusion in December 2009, with considerable modification of the project. It was no longer required to get the top position.

2.8 Completion
The nickname of the supercomputer "京 (Kei)" or "K" was decided on July 5, 2010 after soliciting public proposals. "京" is a Japanese numeral meaning 10^{16}. In China, however, this Chinese character is not used as a numeral.

The first eight racks were shipped to the Kobe Riken site on September 28, 2010. It took 10 months to install the full 864 racks in Kobe. In the ISC2011 conference in Hamburg in June 2011, the K Computer won the No.1 with the Linpack speed of 8.162 PetaFlops using 80% of the full system. In the SC11 conference in Seattle in November 2011, the K Computer won again the top position with 10.51 PetaFlops using the full system. The K Computer also won the No. 1 positions in four HPC Challenge benchmarks: Global HPL, Global Random Access, EP Stream per system and Global FFT. As a real application "The first principle calculation of a Silicon nanowire of 100,000 atoms on the K Computer" won the Gordon Bell prize.

The open access to the K Computer started in September 2012. Applications should be submitted and evaluated for users to have access to the K Computer. Some of the users are from the industry.

3. Challenge to ExaFlops
According to our experience of the K Computer, 10 PetaFlops is not enough for high precision simulation of complex real world systems. Next target is the ExaFlops (10^{18} Flops), which is no easy target, but there have been various efforts to approach this goal.

3.1 Challenges in the U. S.

As early as in 2004, "The Path to Extreme Supercomputing" conference was held in Santa Fe to discuss the possibility of ExaFlops.

In the SC08 in Austin, the IESP (International Exascale Software Project) started as an international collaboration of US, Japan and Europe. The target was to develop system software such as OS, compiler or middleware by international collaboration. It had meetings in Santa Fe, Paris, Tsukuba, Oxford, Mauii, San Francisco and Kobe.

In Europe, similar project started, the EESI (European Exascale Software Initiative). It also had several workshops.

3.2 Japanese efforts

The Exascale efforts in Japan began with participating in the IESP in 2008. A voluntary group of people organized SDHPC (Strategic Development of High Performance Computing) in 2010 and had ten workshops up to now.

As an official effort, the Mext started two working groups for the Exascale computing in July 2011. They published two reports[5] in March 2012: 1) Application working group white paper, and 2) HPCI technology roadmap white paper. The latter included architecture, system software, programming and numerical library. The co-design of application and architecture is important. It is proposed to interface application and architecture on a two-dimensional map of relative memory bandwidth (B/Flop) as the x-axis and relative main memory capacity (B/Flops) as the y-axis. We first thought that in the Exascale region each application might require different architecture. We finally found that relatively universal computer can cover considerable area of applications. We may need, however, other architectures for high bandwidth jobs or big data processing.

In February 2012, the Mext set up Next Generation HPCI Working Group (chaired by Oyanagi) for two years. This working group published an interim report[6] in June 2013. The recommendation is that the government should build one flagship machine to cover relatively wide area of applications. It should be supplemented by other leading machines which have different characteristics. The final report will be published in March 2014.

At the same time, the Mext funded four feasibility study projects for two years, one for application and three for architecture. The application team is to find and estimate a number of social and scientific problems to be solved only by Exascale computing. Three architectures to be considered in the feasibility study are: 1) homogeneous scalar processor, 2) accelerator, and 3) vector processor.

In August 2013, the Mext submitted a budget proposal of 30 million US dolars for the Exascale conceptual design in the 2014 fiscal year (April 2014 – March 2013) to the Ministry of Finance. The total budget for seven years is estimated roughly 1 billion US dollars.

3.3 Technological issues

Very few are thinking of so-called disruptive technologies such as superconducting processor or quantum computing to build Exascale systems. The majority is going to use CMOS semiconductor technology along the future trend. Even using 10 nm or less, there are lots of technological issued to be addressed by our development.

3.3.1 Power wall

From the economic situation, the power consumption of the future Exascale computer should be at the same level of the current K Computer, i.e. around 20 MW. This means we should develop 100 times higher Flops/W than the K Computer. Even considering future progress of semiconductor technology, it is not an easy task. The problem is that more energy is lost in data move than calculations. We should develop algorithms to reduce data moving, not to reduce numerical operations.

3.3.2 Memory wall

The nominal Flops may increase thanks to the large number of transistors in a chip, but the bandwidth between processor and memory is limited due to the number of pins of a chip. There are number of proposals to address this issue. One idea is to increase the bandwidth by optical connection or 3D

stacking of chips. Other idea is to put the on-chip memory in the processor. A cache.is a kind of on-chip buffer memory but is not addressable and controlled by the OS. Addressable on-chip memory is proposed but then the programming would be very difficult without clever compilers.

3.3.3 Reliability wall

Since the number of circuit elements is enormous in ExaFlops, the failure rate may be proportional to it. Fault tolerance in hardware, OS, interconnection and software will be crucially important.

3.3.4 Programming wall

The number of processes or threads will be huge and the memory hierarchy deep. It would be very difficult to write programs on such a machine. Computing efficiency, that is the ratio of measured Flops to the peak Flops is important. In the case of the K computer, it is not easy to exceed 10%. Some users stay below 1%. The situation might be worse for Exascale computers.

4. Conclusion

We have seen that due to the success of vector computers in the 1980's in Japan, parallel processing was behind the U.S. and Europe. We note that practical parallel computer in Japan (NWT, cp-pacs, the Earth Simulator etc.) were built in the initiative of application users.

Although the K Computer project met with strong head wind in the early stage, it is completed and is working stably and producing scientific outcomes. Japan is catching up with the U.S. and Europe in parallel computing.

We are planning to build Exascale supercomputers around 2020. Difference applications require different architecture in terms of B/Flop and B/Flops. Since many applications strongly demand such machines, we are making efforts to acquire taxpayers' support to build Exascale machines.

References

[1] Oyanagi Y 1999 *Parallel Computing* **25** 1545 (Further references therein)

[2] Sentig D N and Smith R V 1965 *AFIPS Proc. FJCC* **27** 117

[3] *Proc. Enabling Technologies for Petaflops Computing (Pasadena, U.S.A., 22-24 February 1994)* ed. T Sterling, P Messina and P H Smith (the MIT Press)

[4] *Proc. PETAFLOPS II, Second Conference on Enabling Technologies for Peta(fl)ops Computing (Santa Barbara, U.S.A. 15-19 February 1999)* ed. P Messina, T Sterling and P Smith

[5] http://open-supercomputer.org/wp-content/uploads/2012/03/FutureHPCI-Report.pdf (in Japanese)

[6] http://www.mext.go.jp/b_menu/shingi/chousa/shinkou/028/gaiyou/1337595.htm (in Japanese)