Feasibility Study of Future HPC Systems for Memory-Intensive Applications

Hiroaki Kobayashi

Abstract After the successful launch of K-Computer in Japan, the Japanese government started a new R&D program entitled "Feasibility Study of Future HPCI Systems." In this program, social and scientific demands for HPC in the next 5–10 years will be addressed, and HPC systems that satisfy the demands and key technologies to develop such systems will be discussed and evaluated. Currently, three system design teams get involved in this program, and this article present a HPC project entitled "Feasibility Study of Future HPC Systems for Memory Intensive Applications," which is conducted by a team of Tohoku University, JAMSTEC and NEC.

1 Introduction

Since the first peta-flop/s machine named Roadrunner became the world's first TOP500 LINPACK system in 2008, about 30 peta-flop/s systems have been installed around world; US, Germany, Italy, France, UK, Australia, Russia, China and Japan only within 4 years. Now the hot topic in the HPC community is when and where the first exascale system will become available. Although *exascale* does not exactly mean exa-flop/s, when taking a look at the trend in sustained LINPACK performance in TOP500, it takes 12 years from 1 tera-flop/s machine named ASCI RED developed in US in 1996 to 1 peta-flop/s Roadrunner in 2008 [7]. US, Europe, China and Japan started several HPC strategic programs for targeting at realization of exascale systems around 2020.

In Japan, after the successful launch of K-Computer, which was the first 10-peta flop/s LINPACK system in 2011, MEXT (Ministry of Education, Culture,

H. Kobayashi (✉)
Tohoku University, Sendai 980-8578, Japan
e-mail: koba@isc.tohoku.ac.jp

M.M. Resch et al. (eds.), *Sustained Simulation Performance 2013*,
DOI 10.1007/978-3-319-01439-5_1,

Sports, Science and Technology) organized a committee to discuss the HPC policy of Japan for the next 5- to 10-year research and development on national leading-supercomputers, and the committee decided to start a program entitled *Feasibility Study of Future HPCI systems* last year. The objectives of this program is to

- Discuss future high-end systems capable of satisfying the social and scientific demands for HPC in the next 5–10 years in Japan, and
- Investigate hardware and software technologies for developing future high-end systems available around year 2018 that satisfy the social and scientific demands.

After the review and selection of the proposals to this program, three teams, which are University of Tokyo with Fujitsu (Project Leader: Professor Yutaka Ishikawa), University of Tsukuba with Hitachi (Project Leader: Professor Mitsuhisa Sato), and Tohoku University with NEC (Project Leader: Hiroaki Kobayashi), started the feasibility studies as a 2-year national project.

In this article, we present an overview of our project entitled, "*Feasibility Study of Future HPC Systems for Memory-Intensive Applications*." Section 2 describes our system design concept with discussing target applications that would be solutions to several important issues as social and scientific demands around 2020. In addition, the basic specification and configuration of the system are presented. Section 3 summarizes the current state of the project and its future plan.

2 Design Concept of the Target System

In the last decades, microprocessors for high-end computing systems boost their flop/s rates by introducing multi- and many-core architectures into the chip design. However, their off-chip memory throughputs are not improved well, and as a result, the bytes per flop rate, B/F, which is a ratio of the memory bandwidth to the peak flop/s is decreasing as shown in Fig. 1 [6]. One exceptional case that keeps B/F high is the vector processor developed by NEC for its SX supercomputer series. The latest system of the NEC supercomputer is SX-9, and its processor provides a 102.4 Gflop/s as a single-core processor with the 256 GB/s memory interface, resulting in the significant B/F rate of 2.5, compared to 0.5 or lower B/F rates of modern microprocessors such as IBM power7 [1], Intel Xeon [3] and Fujitsu SPARC IVfx [4]. However, the vector processor is also facing the memory wall problem as the vector processor flop/s is also increasing.

The memory bandwidth is a key factor to exploit system peak performance and make simulation much more efficient and productive runs. As the B/F is decreasing, it becomes more difficult to feed the necessary data to plenty of arithmetic units on a chip, resulting in lowering the processing efficiency. Figure 2 shows the attainable performance of applications on several modern microprocessors as a function of application B/F rates. An application B/F rate is the memory access intensity of an application, and shows the amount of memory access in bytes per one floating operation in its highest cost kernel. In the figure, "*Roof Lines*" of processors are depicted.

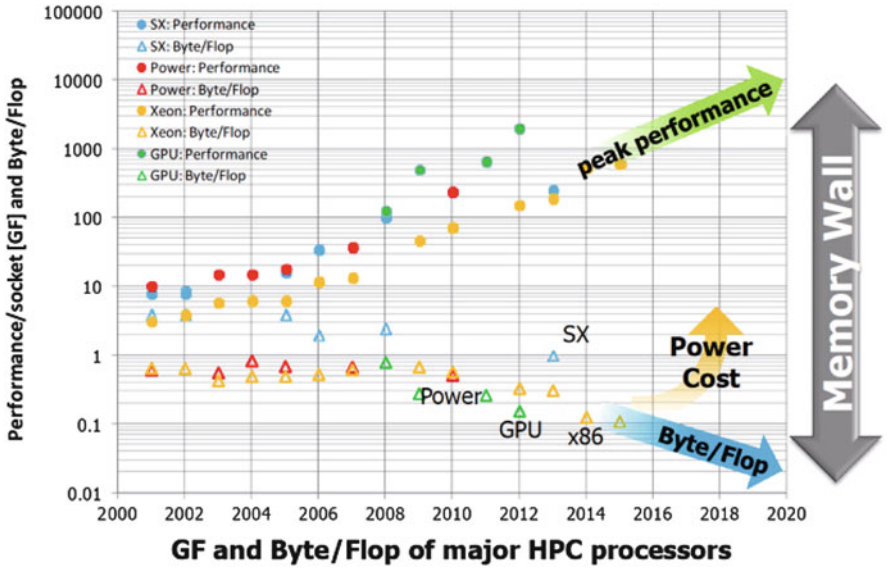


Fig. 1 Trend in processor peak flop/s and B/F

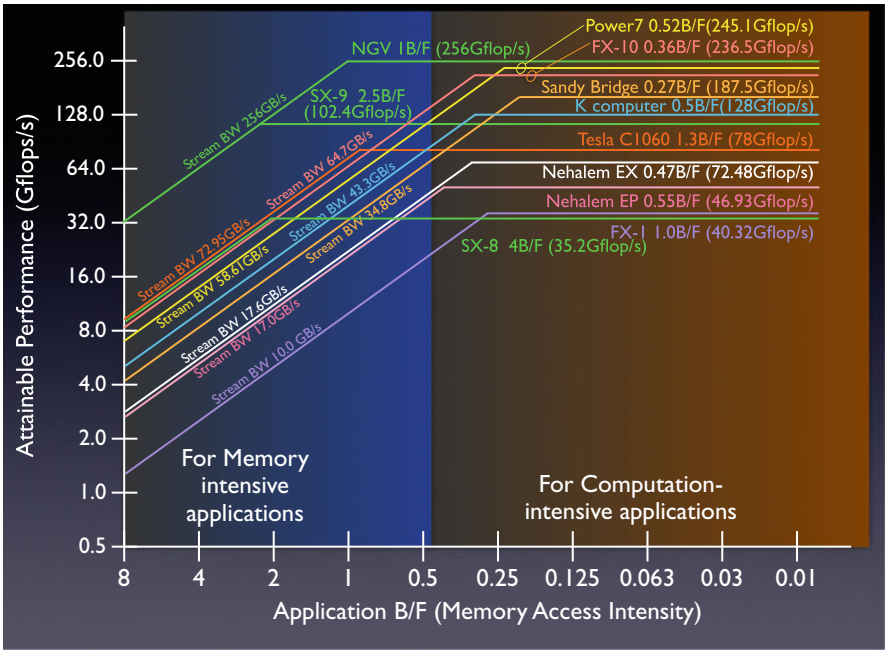


Fig. 2 Roofline models of modern microprocessors

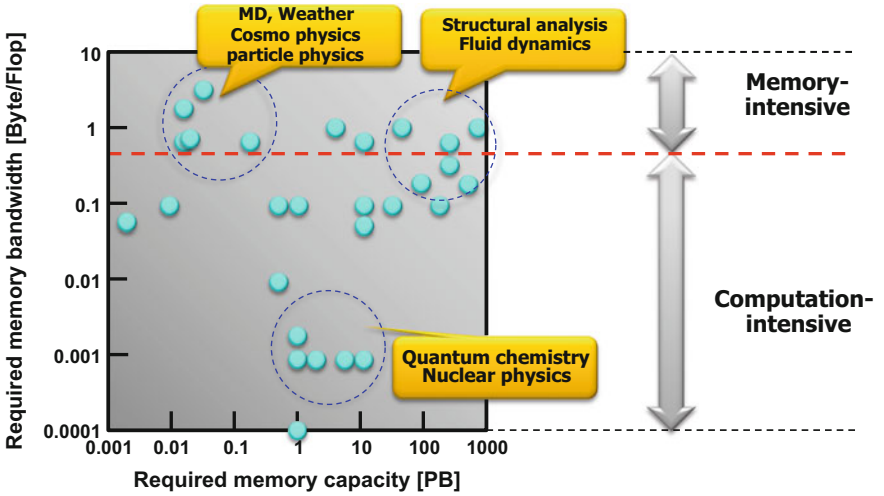


Fig. 3 Memory requirements of applications

A flat roof line shows the upper limit of attainable performance of an application defined by the peak performance of a processor on which the application runs, and a slant roof line is given by the limitation of memory bandwidth of the processor when the application B/F is larger than the processor B/F. As the figure suggests, the order of the processors regarding attainable performance in the case of memory-intensive applications is quite different from that in the computation-intensive applications. For memory-intensive applications, the attainable performance is far from peak-performance of processors when their B/F rates do not match the application B/F rates.

Unlike LINPACK, sustained performance of many practical applications is memory-limited, because these applications need a lot of data during operations in their high cost kernels. According to the application development roadmap report summarized in Japan in 2012 [5], many important applications in the wide variety of science and engineering fields need 0.5 B/F or more, as shown in Fig. 3. Therefore, if we continue to develop high-end computing systems by concentrating on increasing flop/s rates, simply targeting toward exa-flop/s in 2020, rather than memory bandwidth, their applicable area are getting limited, i.e., there will be a high probability that only few % of peak performance of exa-flop/s would be effective in the execution of practical exascale applications, and lots of arithmetic unites end up wasted during their execution.

Based on the above observation, in our project, we are much more focusing on device and architectural technologies to keep high B/F in the design of future HPC systems for exascale computing of important applications available around 2020. In particular, we will explore the design space of the future HPC systems to make them more flop/s-efficient and applicable to the wide fields of science and

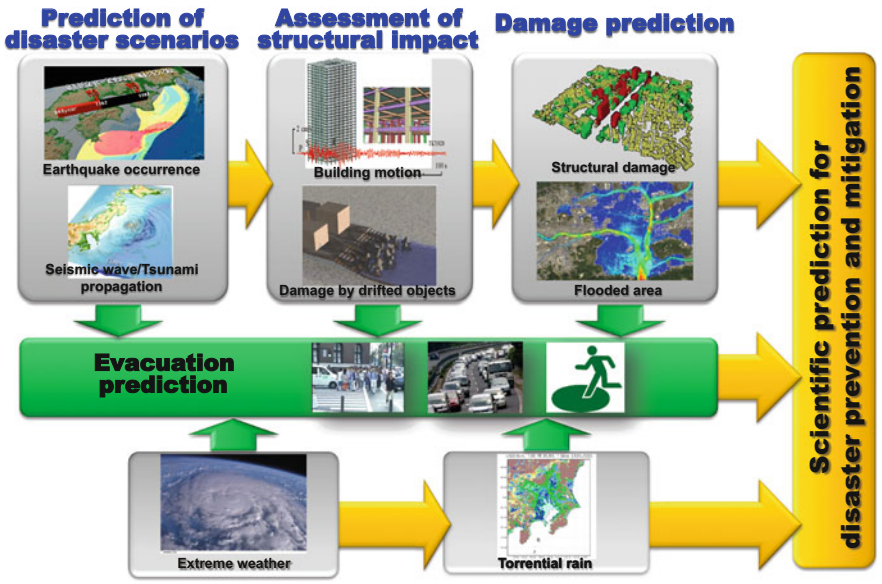


Fig. 4 Simulation model of compound disaster

engineering, especially satisfying social demands for realizing the safe and comfortable society with the HPC technology. To this end, we bring many leading application researchers into the project for co-design of architectures and applications, in order to introduce applications viewpoints into the system design, and make the system highly optimized and friendly in both performance per power/energy consumption and programming, respectively.

Although many applications need higher B/F rates in their programs, important memory-intensive applications come from earth sciences and engineering fields. Especially, after the great East-Japan earthquake in 2011, there is a growing demand for the prevention and mitigation of natural disaster by large earthquakes and tsunami in Japan. Therefore, we study the simulation for dealing with compound disaster, which consists of an underground structure model, an strong motion model, an tsunami model, and a whole city model, in order to provide useful information about damage of lands and building, and evacuation guidance after large earthquakes and tsunami as shown in Fig. 4. In this compound disaster simulation model, the weather model that simulates typhoon, concentrated heavy rains, and tornado is also combined as another type of disaster sources that needs to be considered.

In addition to natural disaster analysis as memory-intensive applications, we also address engineering simulations as our memory-intensive applications, in particular, high-resolution airplane models for digital flight and multi-physics models for reliable and efficient turbo machineries for power plant systems, which need exascale computing as shown in Fig. 5. For these target applications, we investigate their fundamental simulation models available in 2020, and estimate B/F of their

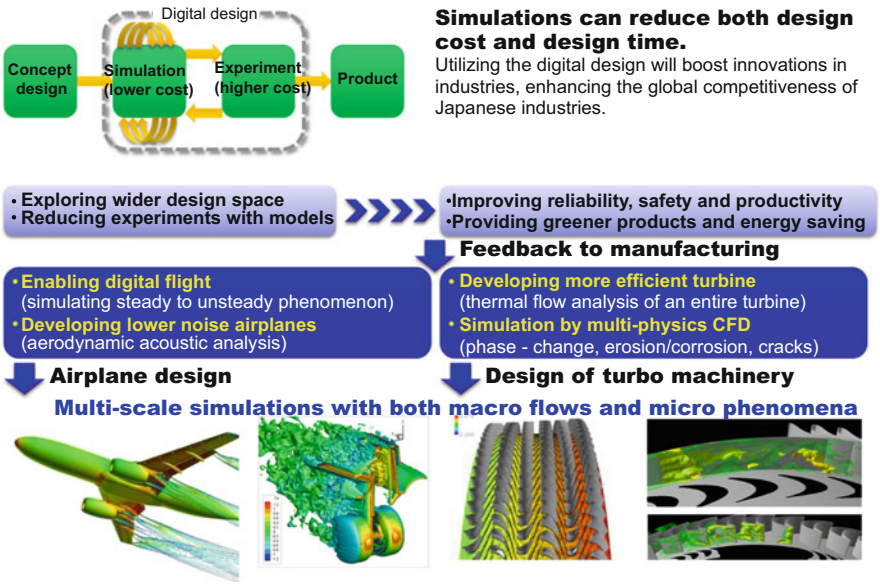


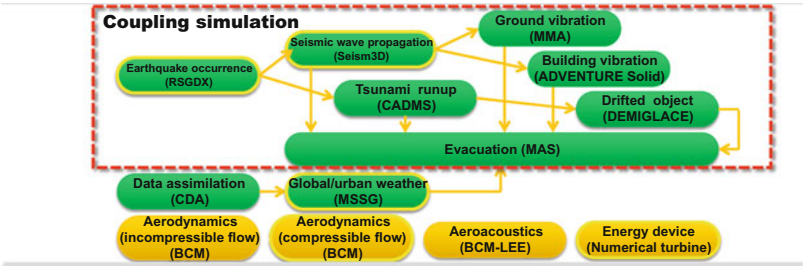
Fig. 5 Advanced digital design of airplanes and turbo machineries

high cost kernels, the total amount of the required computation, memory capacity, and the time limitation to the solution, for the productivity of simulation as basic requirements when considering the design of the system architecture for future HPC systems whose design target is around 2018. Table 1 summarizes requirements of the applications. In the top of the table, applications examined in this project are presented, and the simulation flow in the dotted box shows compound disaster analysis. In the bottom of the table, there are two categories for applications: one is for applications that need very high-resolution models for their single run. The other is for applications each of which is a collection of several runs of moderate resolution models with different simulation parameters. The table suggests that all the applications for disaster analysis and engineering product design need two or more B/F rates in their high cost kernels, and one to hundreds exascale floating point operations should be processed less than 10 h down to in 30 min for their simulation productivity. They also need up to 14 PB for memory space.

After the careful review of these application requirements for exascale computing, we reached the following basic design concepts:

- Design an advanced vector multi-core architecture optimized not only for long vectors but also for short vectors with indirect memory access,
- Design a larger computing node connected to larger shared memory at higher B/F rate, compared to conventional microprocessors and their successors projected, in order to keep the number of parallel processing nodes required by target applications as low as possible,

Table 1 Requirements of target applications for the design of the HPC system



Purpose	Application	B/F	# of floating point operations ($x10^{18}$)	required memory size	Max. allowed runtime (hr)
Super-high resolution single run	RSGDX	4.50	230	1.5 PB	10
	Seism3D	2.14	160	9.6 PB	2
	MSSG	4.00	720	175 TB	6
	BCM	5.47	1	13.6 TB	0.5
High-resolution ensemble run	Comprehensive disaster prevention	2∼5	100	4 PB	2∼3
	Numerical turbine	2.33	140	163.5 TB	20

- Design a advanced memory subsystem using innovative 2.5D silicon interposer and 3D-die stacking technologies to satisfy the requirements of B/F rates of 2 or more of target applications,
- Design a network system with better local communications, while keeping the latency of global communication as low as possible,
- Design a hierarchical storage and I/O subsystem that satisfies the requirements for data assimilation in disaster analysis, and
- Design a system software that provides the standard programming environment, i.e., LINUX-based system software, but it should be customized for exploiting the potential of the advanced vector architecture without sacrificing the standard programming environment.

Figure 6 shows a block diagram of the system designed based on these concepts. The basic specification of the system is as follows:

- Performance/Socket
  - >1 Tflop/s of a peak performance
  - >1 TB/s of a memory bandwidth
  - >128 GB of memory capacity
- Performance/Node
  - Up to 4 sockets/node
  - >4 Tflop/s of a node peak performance
  - Up to 1 TB of the shared memory

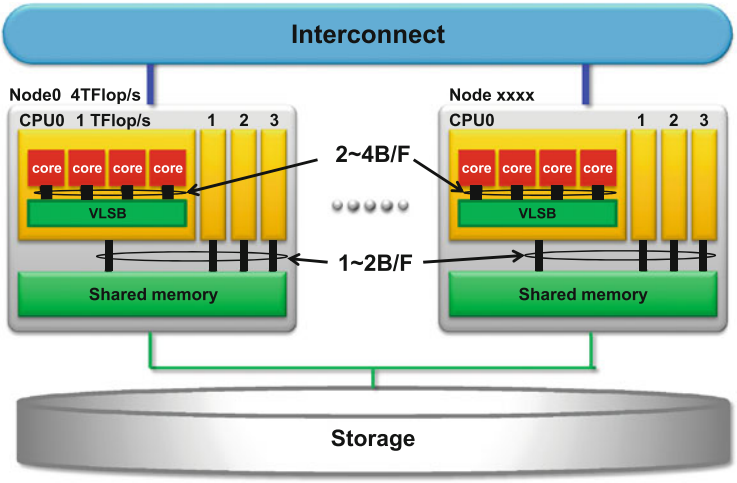


Fig. 6 Conceptual design of the target system

The system scale will be decided so as to satisfy the requirements regarding *the time to solution* of individual applications after the performance estimation.

The design concepts presented here are very aggressive and challenging, and we understand that there several obstacles that should be solved before going into the production of the above system mainly due to cost and power budget in addition to technological issues, but application designers, architecture designers and device designer in our project work tightly together to make innovation happen for realizing highly productive HPC systems that will be expected to became available around 2018!

3 Summary

This article describes our on-going project for design and development of future HPC systems, especially suited for memory-intensive applications. Our design is based on the vector architecture, but many advanced technologies for efficient processing of short vectors with indirect memory references will be introduced. New device technologies such as 2.5D interposer and 3D die-stacking for the design of the memory subsystem, which make the memory subsystem high-bandwidth and low-power, will aggressively be applied to the system design. For the true holistic collaboration among applications, architectures and device technologies, we bring leading researchers and engineers in individual fields nation-wide into the project. Now we are finalizing the details of the architecture and estimating its performance by using target applications scaled to exa-level computing, which satisfy the social demands and make science and engineering breakthrough happen.

Our design philosophy is to *waking up plenty of sleeping floating point units on a chip by improving memory performance for practical applications, not optimized for LINPACK*. We believe the brute force to exascale computing by simply increasing the number of simple cores on a chip does not make sense any more for practical applications in the wide variety of science and engineering. So far, LINPACK of TOP 500 drives the development of high end systems as their performance measure, however, many people are now aware of the limitation of LINPACK for the evaluation of productivity of high-end computing systems [2] We think that innovative memory system design would be a key to realization of highly productive HPC systems in the next 5–10 years, and we believe that our approach will open up the new way to exascale computing. Let' make the supercomputer for the rest of us happen, not for LINPACK only!

Acknowledgements Many colleagues get involved in this project, and great thanks go to Dr. Y. Kaneda and Dr. K. Watanabe of JAMSTEC (Japan Agency for Marine-Earth Science and Technology) as co-leaders of the application group, Professor M. Koyanagi of Tohoku University as the leader of the 2.5D/3D device group, and Ms. Y. Hashimoto of NEC as the leader of the NEC application, system and device design group. This project is supported by MEXT.

References

1. M. Floyd et al. Harnessing the Adaptive Energy Management Features of the POWER7 chip. In *HOT Chips 2010*, 2010.
2. M. Flynn et al. Moving from petaflops to petadata. *Communications of the ACM*, 56:39–42, 2013.
3. Oded Lempel. 2nd Generation Intel* Core* Processor Family: Intel Core i7, i5 and i3. In *HOT Chips 2011*, 2011.
4. Takumi Maruyama. SPARC64(TM) Viiifx: Fujitsu's New Generation Octo Core Processor for PETA Scale Computing. In *HOT Chips 2009*, 2009.
5. MEXT HPC Task Force. (in Japanese) Report on Application R&D Roadmap for Exascale Computing. 2012.
6. S. Momose. Next Generation Vector Supercomputer for Providing Higher Sustained Performance. In *COOL Chips 2013*, 2013.
7. Top 500 Supercomputer Sites. http://www.spec.org/.