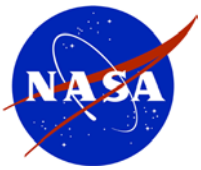





WTEC Panel Report on

HIGH-END COMPUTING
RESEARCH AND DEVELOPMENT IN JAPAN

Alvin W. Trivelpiece (Chair)
Rupak Biswas
Jack Dongarra
Peter Paul
Katherine Yelick

Final Report

World Technology Evaluation Center, Inc.

2809 Boston Street, Suite 441
Baltimore, Maryland 21224

HIGH-END COMPUTING RESEARCH AND DEVELOPMENT IN JAPAN

Sponsored by the National Science Foundation, the Department of Energy's Office of Science, the National Aeronautics and Space Administration, in cooperation with the National Coordination Office for Information Technology Research and Development of the United States Government.

Panel Members

Alvin W. Trivelpiece
14 Wade Hampton Trail
Henderson, NV 89052

Rupak Biswas
NASA Ames Research Center
Mail Stop T27A-1
Moffett Field, CA 94035

Jack Dongarra
Computer Science Department
1122 Volunteer Blvd.
University of Tennessee
Knoxville, TN 37996-3450

Peter Paul
Director's Office, MS 460
Brookhaven National Laboratory
Upton, NY 11973-5000

Katherine Yelick
777 Soda Hall
Computer Science Division
University of California at Berkeley
Berkeley, CA 94720-1776

WTEC, Inc.

WTEC provides assessments of foreign research and development in selected technologies under awards from the National Science Foundation, the Office of Naval Research, and other agencies. Formerly part of Loyola College's International Technology Research Institute, WTEC is now a separate non-profit research institute. Michael Reischman, Deputy Assistant Director for Engineering, is NSF Program Director for WTEC. Sponsors interested in international technology assessments and related studies can provide support for the program through NSF or directly through separate grants to WTEC.

WTEC's mission is to inform U.S. scientists, engineers, and policymakers of global trends in science and technology. WTEC assessments cover basic research, advanced development, and applications. Panels of typically six technical experts conduct WTEC assessments. Panelists are leading authorities in their field, technically active, and knowledgeable about U.S. and foreign research programs. As part of the assessment process, panels visit and carry out extensive discussions with foreign scientists and engineers in their labs.

The WTEC staff helps select topics, recruits expert panelists, arranges study visits to foreign laboratories, organizes workshop presentations, and finally, edits and disseminates the final reports.

WTEC Panel on

HIGH-END COMPUTING IN JAPAN

Final Report

December 2004

Alvin W. Trivelpiece (Chair)
Rupak Biswas
Jack Dongarra
Peter Paul
Katherine Yelick

This document was sponsored by the National Science Foundation (NSF) and other agencies of the U.S. Government under an award from the NSF (ENG-0104476) to the World Technology Evaluation Center, Inc. The Government has certain rights in this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Government, the authors' parent institutions, or WTEC, Inc.

ABSTRACT

This report presents the findings of a study of high-end computing (HEC) in Japan, one of a series of 60 such WTEC studies of foreign technologies. This study complements others underway at about the same time, all inspired by the achievement of the Japanese Earth Simulator (ES) in taking the lead as the world's fastest supercomputer in March, 2002. The WTEC panel gathered information by background research and a visit to 22 of the key organizations in Japan. Japan has had a broad-based strategic effort in high-performance computing over the past decade; the ES is the principal embodiment of that strategic effort. The ES has had a major impact in the Earth sciences, leading to significant advances in the field. ES is now extending its applications to other fields, including biosciences and nanotechnology. However, continued progress in large-scale high-fidelity simulation will require a significant increase in power beyond the ES. Japan has a broadly based and carefully planned but audacious program in advanced scientific simulations. The strategic attack on protein structure, cell simulations, and computational bioscience is especially noteworthy. The Protein Explorer, with its huge increase in power for molecular dynamics, could put Japan into world leadership in this area. Grid computing is a high priority for government agencies, and the funding levels for grid programs are much greater than for vector supercomputing programs like the ES. Even before the U.S. regained the lead in November, 2004 with the IBM Blue Gene /L, the WTEC panel concluded that, considering the whole spectrum of HEC, the U.S. was ahead of Japan. Follow-on efforts to the ES are underway, but U.S. leadership in HEC overall is likely to continue in the near future. However, the Japanese have a justifiable right to be proud of what has been accomplished.

World Technology Evaluation Center, Inc.

R. D. Shelton, President
Y.T. Chien, Vice President for Research
Michael DeHaemer, Vice President for Development
Geoffrey M. Holdridge, Vice President for Government Service
Roan E. Horning, Vice President for Operations
Mike Jasik, Publications Manager
Advance Work by Masanobu Miyahara

FOREWORD

> We have come to know that our ability to survive and grow as a nation to a very large degree depends upon our scientific progress. Moreover, it is not enough simply to keep abreast of the rest of the world in scientific matters. We must maintain our leadership.[1]

President Harry Truman spoke those words in 1950, in the aftermath of World War II and in the midst of the Cold War. Indeed, the scientific and engineering leadership of the United States and its allies in the twentieth century played key roles in the successful outcomes of both World War II and the Cold War, sparing the world the twin horrors of fascism and totalitarian communism, and fueling the economic prosperity that followed. Today, as the United States and its allies once again find themselves at war, President Truman's words ring as true as they did a half-century ago. The goal set out in the Truman Administration of maintaining leadership in science has remained the policy of the U.S. Government to this day: Dr. John Marburger, the Director of the Office of Science and Technology (OSTP) in the Executive Office of the President made remarks to that effect during his confirmation hearings in October 2001.[2]

The United States needs metrics for measuring its success in meeting this goal of maintaining leadership in science and technology. That is one of the reasons that the National Science Foundation (NSF) and many other agencies of the U.S. Government have supported the World Technology Evaluation Center (WTEC) and its predecessor programs for the past 20 years. While other programs have attempted to measure the international competitiveness of U.S. research by comparing funding amounts, publication statistics, or patent activity, WTEC has been the most significant public domain effort in the U.S. Government to use peer review to evaluate the status of U.S. efforts in comparison to those abroad. Since 1983, WTEC has conducted over 50 such assessments in a wide variety of fields, from advanced computing, to nanoscience and technology, to biotechnology.

The results have been extremely useful to NSF and other agencies in evaluating ongoing research programs, and in setting objectives for the future. WTEC studies also have been important in establishing new lines of communication and identifying opportunities for cooperation between U.S. researchers and their colleagues abroad, thus helping to accelerate the progress of science and technology generally within the international community. WTEC is an excellent example of cooperation and coordination among the many agencies of the U.S. Government that are involved in funding research and development: almost every WTEC study has been supported by a coalition of agencies with interests related to the particular subject at hand.

As President Truman said over 50 years ago, our very survival depends upon continued leadership in science and technology. WTEC plays a key role in determining whether the United States is meeting that challenge, and in promoting that leadership.

Michael Reischman

Deputy Assistant Director for Engineering
National Science Foundation

---

[1] Remarks by the President on May 10, 1950, on the occasion of the signing of the law that founded the National Science Foundation. *Public Papers of the Presidents* 120: p. 338.

[2] Statement of Dr. John Marburger III before the Committee on Commerce, Science, and Transporation. United States Senate. 2001. <http://www.ostp.gov/html/01_1012.html> Last accessed February 23, 2005.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

This report was prepared by WTEC, which is a non-profit research institute funded by grants from most Federal research agencies. Among other studies, WTEC has provided peer reviews by American experts of international R&D in 60 fields since 1989. In late 2003, WTEC was asked by several agencies to assess Japanese R&D in high-end computing. This report is the final product of that study.

We would like to thank our distinguished panel of experts, who are the authors of this report, for all of their efforts to bring this study to a successful conclusion. Horst Simon provided invaluable advice on the study and final report. The assistance of ATIP in helping to identify key sites was appreciated. We also are very grateful to our Japanese hosts for their generous hospitality to our panel, and to the participants in our workshop on HEC in Japan. Of course, this study would not have been possible without encouragement from our sponsor representatives: Peter Freeman, Sangtae Kim, Michael Reischman, and Deborah Young of NSF, Walt Brooks of NASA and Norm Kreisman of DOE. David Nelson and Sally Howe of NITRD helped coordinate the effort.

To make this report as current as possible, this preface will include some of the recent developments in high-end computing in the U.S. in addition to plans announced in Japan.

Dubbed by Professor Jack Dongarra a "Computenik," the ascension in early 2002 of the Japanese Earth Simulator (ES) to number one on the Top500 ratings chart for computer performance, startled many American scientists and government agencies, similar to the Soviet Union's launch of the Sputnik satellite in 1957. That event helped set up a renewed supercomputer race between the U.S. and Japan, which, three years later, has produced unexpected results for both countries. In the latest (November 2004) Top500 ranking, DOE's IBM BlueGene/L beta system overtook Japan's ES as No. 1, achieving a Linpack performance of 70.72 Tflop/s, while the Columbia system at NASA/Ames, built by SGI, gained the No. 2 spot, with a speed of 51.87 Tflop/s. The ES, first for two and a half years, dropped to No. 3 at 35.86 Tflop/s. (See Table C.1 in Appendix C)

What led the U.S. back to the top was a general determination to regain what was lost. But more important was a concerted effort by government, academia, and industry to focus on what could be done to get there—and the funding to make it happen. From the High End Computing Revitalization Task Force (HECRTF) to public-funded technology assessment, including this WTEC report, the U.S. government reached out for community advice. On the funding side the Congress pumped $165 million into the cause with the passage of the Department of Energy High-end Computing Revitalization Act of 2004. With Japan having its own strategic and commercial goals, we can't expect it to respond the same way, but we can expect a vigorous effort attaining the goal both nations share: building peta-scale machines by 2010. Already, a new government-industry consortium is being developed to plan for the post-ES world. Tables P.1 and P.2 summarize some of the recent events in both the U.S. and Japan that relate to this report.

The original emphasis of the WTEC study was on Japan's developments in general purpose machines both about the ES and beyond. What the Panel found, however, is that Japan's approach to HEC has shifted to a much broader program since the introduction of the ES, with the focus of government funding and resources expanding to cover commodity clusters, grid computing, and special purpose machines. In the last couple of years, Japanese government has begun at least two major initiatives, one on data and business computing funded under the Ministry of Economy, Trade, and Industry (METI), and the other on scientific computing – the National Research Grid Initiative (NAREGI), under the Ministry of Education, Culture, Sports, Science, and Technology (MEXT). These programs have received high priority funding from their respective agencies, as the panel has reported. However, this is not to say that Japan has given up on traditional vector-type computing. As Professor Dongarra concluded in his architecture overview of Japanese high-end computers, NEC, the manufacturer of the ES, appears to be committed to high-end vector computing. NEC believes that the development of this high-end product will spur the technology needed for the other systems. It also places emphasis on hardware continuity and sustainability of market development, as opposed to mere technology innovation in their strategy for product development. In October 2004, NEC announced the

worldwide launch of the newest SX series model, the SX-8 machine, estimated to have achieved a peak performance of 65 Tflop/s when configured with the maximum 512 units. NEC aims to achieve worldwide sales of more than 700 SX-8 units of various sizes during the next three years.

Table P.1
Recent HEC Events in the United States

New Program, Event or Activity	Description (Goal, Scope, etc.)	Outcome, Status or Impact
DOE/IBM Blue Gene/L overtook Japan's ES as No. 1 in the Top500 (ES slipped to No.3) as of November 2004; Rmax = 70.7 Tflop/s Rpeak = 91.7 Tflop/s	A follow-on to DOE's ASCI program, the Blue Gene series, combined with other HEC programs in DOE labs and vendors, is expected to usher in a transition to the next generation of HEC systems that combines the benefits of both commodity and custom designs.	A full Blue Gene/L machine is being built for the Lawrence Livermore National Lab with a peak speed of 360 Tflop/s in 2005.
DARPA's High Productivity Computing Systems (HPCS) program Phase 1: 2002 Phase 2: 2003-6 Phase 3: 2007-10	The goal of the HPCS program is to "provide a new generation of economically viable, high productivity computing systems for the national security and industrial user communities." It is expected to deliver peta-scale performance with new technologies including quantum computing.	Three consortium teams have been selected in phase 2, led by IBM ($53.3m), Cray ($43.1m), and Sun Microsystems ($49.7m). One or two of the three will be selected for phase 3 work, a full-scale engineering development effort.
The High-end Computing Revitalization Task Force (HECRTF) 2003 – an interagency group (DOD, DOE, NSF, NASA, and others), commissioned by the NSTC	The goal is to develop a strategic plan for undertaking and sustaining a robust federal HEC program to maintain U.S. leadership in science and technology. HECRTF charge from NSTC includes developing strategies for investment in core technologies; requirements for Federal HEC computing capability, capacity and accessibility; coordination in HEC procurement; and integration of HEC strategies from all government and private sectors.	A report of the HECRTF on "Federal Plan for High-End Computing" was published May 10, 2004. Included: vision, roadmaps, and recommendations for R&D strategies, and funding priorities – basis for the federal planning and management of future HEC programs.
NRC/CSTB study on "The Future of Supercomputing," commissioned by DOE in Feb 2003. Co-Chairs of the Study: Susan Graham and Marc Snir.	This study is "part of a broader initiative by the U.S. government to assess its current and future supercomputing needs and capabilities, spurred in part by the ES." Focus of the study is on the fastest and most powerful computing systems, excluding grids and networking technologies.	Final report: "Getting up to Speed: The Future of Supercomputing," Dec 2004. Included in the report are an assessment of the status of SC in the U.S. and abroad and a set of recommendations to the federal government regarding future HEC investment.

The Panel also reported on the continued strong program in the research and development of special purpose machines, a trend less in evidence in the U.S. Based on the success of several generations of the Grape (GRAvity PipE) architecture for gravitational computations in astrophysics, the Institute of Physical and Chemical Research (RIKEN) is leading an effort in the "Protein Explorer" Project to build a petaflop/s machine for molecular dynamics simulation to be completed in 2006. Dr. Peter Paul has provided in his chapter on scientific applications of Japanese high-end computing a detailed account of the Protein Explorer and its implications for future genomic and protein/cell research. Since our Panel's visit, MEXT has funded a new follow-on project, Grape-DR, for a consortium consisting of the University of Tokyo, National Institute of Information and Communications Technology, NTT Communications, National Astronomical Observatory of Japan, and RIKEN. The goal of the project is to build a high-performance system for collaborative scientific computation featuring a 2 petaflop/s computing engine and a high throughput network by the year 2008. The system would consist of Grape-DR clusters located in distant locations connected by high-speed Internet with bandwidth of 40-400 Gbps. The clusters would be powered by 32-128 PCs with a Grape-DR engine that integrates 2 million processor elements (1024 on a chip), complete with an operating

system, compiler, Web interface and distributed shared data systems. Unlike the Protein Explorer or other Grape predecessors, Grape-DR is intended for multiple applications and in networked environments. Besides the consortium members, IBM Japan participates in this project as the vendor for chip and other processor design. There is no doubt that this Grape-DR, if delivered, will provide some excitement between now and the time when both the U.S. and Japan hope to roll out peta-scale general purpose machines around 2010.

Table P.2
Recent HEC Events in Japan

New Program, Event or Activity	Description (Goal, Scope, etc.)	Outcome, Status or Impact
MEXT launched a new government-industry consortium in 2004 to develop a petaflop/s machine by about 2010. The consortium will involve NEC, Hitachi and Toshiba, as well as universities and national labs.	The project will focus on research to produce new breakthrough technologies needed for the next generation. HEC components, systems and SW. First year – concept and design; next three years for development of new technologies and a prototype system; followed by the production of a commercial version.	MEXT plans to spend ~ $20 million in fiscal 2005 for the initial design phase. In its review of agency programs, Japan's CSTP ranked as "superior" for MEXT's submission of ~$18M project on "Technologies for future supercomputing." Ministry of Finance has since approved ~15 million in its 2005 budget.
MEXT announced in May 2004 a new research project Grape-DR as a follow-on to Grape-6. Project group: U. Tokyo, RIKEN, National Institute of Information and Communications Technology, National Astronomical Lab, and NTT Communications.	The goal of the project is to develop a 2 Pflop/s computing engine and global research infrastructure that uses multi-10Gps networks by 2008. Grape-DR plans to accomplish this by integrating two million processors in high density racks with ultra high-speed switches and high-capacity storage. IBM Japan is the vendor for chip and board designs.	MEXT plans to spend ~$10M on the project over five years. The budget does not include personnel. Grape-DR works as an extension of the well known Grape series for special purpose computing, but will go beyond its architectural and application limitations.
NEC announced the launch of Supercomputer SX-8 in Oct 2004, an extended line to the predecessor on which Earth Simulator was based.	The SX-8 computer is claimed to be the world's fastest vector machine. Key features include: its CPU operates at 2 GHz and integrates the vector processor into one single chip using CMOS tech. of 90 nm; each node/unit has 8 CPUs and up to 512 nodes (4,096 CPUs) can be linked to produce a peak vector performance of 65 Tflop/s.	NEC expects to sell 700 units of the SX-8 within the next three years. Initial orders came from the U.K. (16 units, 2 Tflop/s) and Germany (64 units, 8 Tflop/s). Like its predecessor vector machines, the SX-8 is likely to deliver much higher sustained performance.

From the conclusions of this study and the recent events summarized above, we may venture a couple of observations. First, it seems that while the U.S. and Japan had different rationales for pursuing high-end computing at the outset, they now have more shared motivations for doing so. Early HEC programs in the U.S. were mainly justified by the need to meet the national security requirements imposed after the signing of the nuclear Comprehensive Test Ban Treaty (see Chapter 1). Japan, on the other hand, has always been concerned about environmental issues, leading to the development of the ES. Both countries, however, are increasingly making HEC a critical resource for solving other grand challenge problems in society: aerospace exploration, genomics and health, alternative energy sources, nanoscale science and technology, rapid response to natural and man-made threats, as well as training the next generation of computational scientists and engineers. The 2005 Guide to the NITRD Program (supplement to the President's Budget), outlined this vision clearly. A similar vision is found in the latest Japanese 2005 budget recommendations for supercomputing- and IT-related programs. Second, it is also clear that Japan's quest for supercomputer leadership has its own strategic and commercial goals, often different from ours. While both countries are aspiring to build petaflop/s-scale machines in the next five years, each is likely to take a different path to achieve it. Success will depend on many factors, including capitalizing on each country's unique strengths and overall vision. On the other hand, both countries will have to learn from each other. For the U.S.,

Japan's will and ability to achieve in high-end computing has given us a motivational goal post. To be sure, the vision of a successful DOE BlueGene program began even before the advent of the ES. Aiming to regain the supercomputer crown, however, added at the very least to the motivation for its on-time delivery of a beta system this November. Similarly, Japan in this race needs to re-assess its position to invest in a post-ES supercomputer, as is reflected in the new government-industry consortium recently announced. The race is shaping up more like a tug-of-war than a hundred meter dash. Nonetheless, it is probably fair to say that Japanese supercomputing, and the Earth Simulator in particular, has helped raise the bar in the world's quest to be the best in high-end computing. When the bar is raised, everyone stands to improve their performance.

This report covers a broad spectrum of material on the subject, so it may be useful to give a preview here. The Executive Summary was prepared by the chair, Alvin W. Trivelpiece, with input from all the panelists. The chapters in the body of this report present the panel's findings in an analytical organization by subdiscipline. Appendix A provides the biographies of the panelists. Appendix B contains the panel's individual reports on each site visited in Japan, which form a chronological or geographic organization of much of the material. Appendix C provides an update on the Top500 data, which was released in November 2004. The highlights from the U.S. workshop, held on May 25, 2004, are provided in Appendix D. A glossary is in Appendix E.

All the products of this project are available at http://www.wtec.org. The electronic color version of this report is particularly useful for figures that do not reproduce well in black and white. Also posted at this site are the slideshows from the workshop held for this project, which contain considerable additional information on R&D in Japanese HEC. Comments on this report are welcome.

Y.T. Chien

EXECUTIVE SUMMARY

ASSESSMENT OF JAPANESE RESEARCH, DEVELOPMENT, AND APPLICATION OF HIGH-END COMPUTER SYSTEMS

Alvin W. Trivelpiece

STUDY PROCESS

This report is an account of an effort to better understand High-End Computer Systems (HEC) in Japan. This effort had its origins in discussions in late 2003 among several U.S. government agencies with responsibilities for funding programs in the U.S. HEC enterprise. They include the Department of Energy Office of Science (DOE/SC), the National Science Foundation (NSF) and the National Aeronautics and Space Administration (NASA). These agencies, in cooperation with Dr. David B. Nelson, Director of the National Coordination Office for Information Technology R&D, asked the World Technology Evaluation Center (WTEC) to organize a group of scientists and engineers with appropriate knowledge to review, analyze and report on Japanese projects in high-end computing. This assessment included a study tour to Japan to allow the panel members to meet with Japanese scientists, engineers, managers and government officials involved in the support and operation of HEC at universities, national laboratories and industrial organizations.

This study complements three others underway at about the same time, all inspired by the challenge presented by the achievements of the Japanese Earth Simulator in taking the lead as the world's fastest supercomputer in March, 2002. (1) The National Academy of Sciences kicked off a study in March 2003 focused on assessing the U.S. scene. Their report, "Getting Up to Speed: The Future of Supercomputing," is available in draft form on the Web as of this writing at http://www.nap.edu/catalog/11148.html. (2) The National Science and Technology Council organized a High-End Computing Revitalization Task Force to develop a plan for a U.S. HEC program to maintain U.S. leadership in science and technology. That report, "Federal Plan for High-end Computing," (May 10, 2004) is available at http://www.itrd.gov/hecrtf-outreach. (3) Finally, Congress also commissioned a study by the JASONs on the HEC requirements for nuclear stockpile stewardship.

The WTEC Panel was organized at a kick-off meeting on January 9, 2004. It conducted a study tour in Japan from March 28 through April 3, 2004. During this time the Panel met with many individuals at the 22 institutions listed in the full report. In an effort to optimize its efforts, the Panel divided itself into two groups, assigning members to each group depending on the nature of the functions performed at each particular site. For each site, Panel members prepared a site report, which has been included in the appendix.

As part of its assignment, the Panel briefed individuals from the sponsoring agencies on May 24, 2004, and presented its preliminary findings at a public workshop on May 25. Representatives from many U.S. stakeholders were in attendance, and several representatives from Japanese organizations traveled from Japan to participate in the event. This report has tried to reflect this feedback from the workshop and on-going developments in the field while it was being prepared.

STUDY FINDINGS

In general, there is no doubt that the quality of Japanese research and development in many scientific disciplines is competitive with the world's best, as it is in high-end computing. These principal conclusions specific to this discipline are brief summaries of those given in the full report.

Conclusions on the Earth Simulator and Special Purpose Supercomputers

The ES is a superb engineering achievement and impressively led the world for about two years. At the Supercomputer 2004 conference in Pittsburgh in November, 2004, a new Top500 list was released that showed two American computers passing the ES (at 35.9 Tflop/s). These were the IBM Blue Gene/L supercomputer at 70.7 Tflop/s and the Columbia made by SGI Altix / Voltaire at 51.9 Tflop/s. Even before these latest American achievements, the WTEC panel concluded that considering the whole spectrum of HEC, the U.S. is ahead of Japan. However, Japan has a justifiable right to be proud of what it has accomplished.

There are three reasons how the Earth Simulator came about:

- At the time it was first considered for support, government funding was robust in Japan.
- Japan has regarded the climate and environment as critically important to its interests.
- Dr. Hajime Miyoshi was the key visionary and driving force behind the effort to fund and build the ES.

In the WTEC panel's visits in Japan, the hosts did not identify any plans to expand the ES itself. However, Dr. Sato, the ES director, has made a proposal for a new generation machine 10,000 times more powerful. This is conceived to be a heterogeneous machine that would be internationally funded.

A recent development is that a new Japanese consortium has been organized to build a 1 Pflop/s (1000 TFlop/s) supercomputer by around 2010. The consortium includes NEC, Toshiba, Hitachi, as well as universities and national labs. The budget for the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) for the Japanese 2005 fiscal year includes about $20 million in top priority funding that appears to be directed toward this effort.

In October, 2004 NEC announced the launch of the SX-8 supercomputer, which is based on the ES. On paper a full-blown SX-8 could achieve a peak vector performance of 65 Tflop/s.

Japan has had a broad-based strategic effort in high-performance computing over the past decade. The ES is the principal embodiment of that strategic effort. As intended, the ES has had a major impact in the Earth sciences in that it led to significant advances in the field (e.g. increased resolution, shortened turnaround time). ES is now extending its applications to other fields beyond earth sciences, including biosciences and nanotechnology. However, continued progress in large-scale high-fidelity modeling and simulation will require a significant increase in power beyond the ES.

Japan has a broadly based and carefully planned but audacious program in advanced scientific simulations. The strategic attack on protein structure, cell simulations, and computational bioscience is especially noteworthy. The Protein Explorer, with its huge increase in power for molecular dynamics, could put Japan into world leadership in this area.

The Panel concluded that the investment in software for ES is not in proportion to the investment in hardware. The software research agenda in Japan is currently skewed toward grid middleware, which overlaps cluster computing, and there are only modest research programs at the Japan Marine Science and Technology Center (JAMSTEC) for compilers, programming languages, and tools. High-performance Fortran (HPF) is much more successful in Japan than in the U.S., but there is an increasing interest in message passing interface (MPI) to enhance portability.

There is some resentment to the Earth Simulator by some research groups in Japan. Their feeling is that the ES is too expensive and drains critical resources from other essential scientific and technical programs. In

discussions with the major players in Japan, there did not yet seem to be a broad consensus for a follow-on project of similar magnitude in the near term. Some feel that distributed heterogeneous grid environments promise to satisfy HEC requirements.

Conclusions on Grids

There was more emphasis on grids than the Panel had expected. Grid computing is a high priority for government agencies, and the funding levels for grid programs are much greater than for vector supercomputing programs like the Earth Simulator. The effect of this shift in emphasis on research institutions and vendors has been dramatic, as there is currently relatively little research on supercomputing technologies or tools, but a broad interdisciplinary effort in grids. Because the boundary between grid and cluster computing is somewhat blurred, some of the research in application-level libraries and problems solving environments could easily be considered to be a high-performance computing project, rather than a grid-computing project. However, the emphasis is on commodity processors and the development of software technologies that are applicable to business and government in addition to science and engineering.

The Japanese grid agenda highlights the performance heterogeneity of systems in the grid as an opportunity, allowing applications to select the best hardware for a given part of a computation. The most aggressive use of heterogeneity combines special purpose processors like Grape with vector supercomputers, PC clusters, and shared memory workstations. These ideas have also influenced plans for an international follow-on to the Earth Simulator, which has been described by Director Sato as having exactly this type of heterogeneous architecture.

Conclusions on Government Policy Toward HEC

The Panel visited the three main government agencies involved in high-end computing: the Council on Science and Technology Policy (CSTP), MEXT, and the Ministry of Economy, Trade and Industry (METI).

CSTP is the highest policy-making body for science and technology in the Japanese government, chaired by the Prime Minister himself. It has many strategic planning functions, aimed at steering Japan's science and technology into a more competitive position in the world. Its role includes prioritization of research initiatives, like the one that led to the ES. CSTP representatives stated that proposals for follow-ons to the ES would have to come from the ministries, and the process for building such a consensus would take time.

MEXT now takes the lead in guiding and funding supercomputing through both education programs at universities and science programs at several government agencies under its jurisdiction. The ES was developed under MEXT guidance, and the Panel learned much about how it came to pass during its visit. It was proposed as a tool to promote basic research, not to promote a particular kind of computer technology. It cost about ¥60 billion, including an unexpected supplement that permitted completion in five years instead of the projected eight years. At the time of the visit, MEXT officials said that no specific plans are in place to build a follow-on to the ES. There was also no policy to subsidize the Japanese supercomputing companies. They are, instead, investing in grid computing and other long-term research areas, including quantum computing and nanotechnology.

METI is in charge of administering Japan's policies covering a broad area of economy, trade, and industry. Ten years ago, METI (then MITI) invested in supercomputing, but there were not enough applications to sustain a market. METI is no longer interested in supercomputing, but is interested in PC clusters, because of their cost effectiveness.

CHAPTER 1

INTRODUCTION

Alvin W. Trivelpiece

BACKGROUND OF THE STUDY

This project was initiated in late 2003 by the Department of Energy (DOE), the National Science Foundation (NSF) and the National Aeronautics and Space Administration (NASA) in cooperation with Dr. David B. Nelson, Director of the National Coordination Office for Information Technology R&D. These organizations asked WTEC to organize a group of scientists and engineers with appropriate knowledge to review, analyze and report on Japanese projects in high-end computing (HEC).

According to the sponsors, the purpose of this study was to gather information and disseminate it to government decision makers and the research community on the status and trends in Japanese supercomputer systems R&D in comparison to that in the United States. The panelists were to gather information on Japanese HEC R&D useful to the U.S. government in planning its own HEC R&D programs, and to compare Japanese HEC research, development, and applications activities with those in the United States. This was to focus primarily on long-term research on high-end computing in Japan, including follow-on machines to the Earth Simulator and other high-end computing architectures. As a part of this assessment of future research directions, the study was also to include a review of the development process and operational experience of the Earth Simulator (ES), including the user experience and the impact it has had on the computer science and computational science communities.

The study was to assess future trends in high-end computing in Japan in terms of how they are being affected by, and are affecting, three primary interest groups:

1. *Government agencies*. The Earth Simulator project was funded initially by three Japanese agencies: the National Space Development Agency (NASDA), the Japan Marine Science and Technology Center (JAMSTEC), and the Japan Atomic Energy Research Institute (JAERI). What other agencies will be supporting advanced computing research and applications development in the future? What is the trend in Japanese government support for high-end computing R&D? Is there a strategic plan, and if so, what is it? What is the role of the National Aerospace Laboratory (NAL) versus the ES?
2. *Computer science and computational science research communities in Japan*. What are the leading research groups, and where do they see the future of computer science and computational science going? What are the formal and informal relationships among universities, vendors and government that were utilized in developing strategies and approaches to future HEC systems and their application?
3. *Vendors*. There are three main players now – NEC, Hitachi and Fujitsu. Are there other contenders that may emerge in the future? Are these companies self-contained, or do they depend on suppliers; if the latter, who are the suppliers? Do they have alliances with key component manufacturers; if so, who are those manufacturers?

The sponsoring agencies asked WTEC to recruit a panel of six individuals who would travel to Japan and conduct a review consistent with this purpose and scope. Table 1.1 lists the panelists and other members of the delegation; short biographies are in Appendix A.

Table 1.1
WTEC Delegation

Member of Delegation	Organization
Alvin W. Trivelpiece (Chair)	Senior Consultant, Sandia Corporation
Rupak Biswas (Panelist)	NASA Ames Research Center
Jack Dongarra (Panelist)	University of Tennessee
Peter Paul (Panelist)	Brookhaven National Laboratory
Katherine Yelick (Panelist)	University of California at Berkeley
Stephen Meacham (Sponsor Rep.)	National Science Foundation
Y. T. Chien (Staff)	WTEC
Masanobu Miyahara (Staff)	WTEC

WTEC organized the initial meetings of the panel on December 10, 2003, and January 9, 2004, with sponsoring agencies to identify the individuals and sites in Japan that should be visited. WTEC also provided the advance work that gained access to the sites that the panel selected. Following these preliminary activities, the panel conducted a formal study tour in Japan from March 29 to April 3, 2004. In order to visit as many sites as practical in one week, the panel was divided into two teams. The members of each team were selected based on their experience and the activities at the site to be visited. A list of sites is given in Table 1.2, and Appendix B provides detailed reports from each site visit.

After returning from Japan, the panel presented its preliminary findings at a workshop held in the boardroom of the National Science Board in Arlington, VA, on May 25, 2004. Presentations from that workshop are posted at http://wtec.org/hec, and a summary is included here in Appendix D.

RATIONALES FOR INVESTMENT IN HEC IN THE U.S. AND JAPAN

In the U.S. much of the motivation for the development of high-end computing came from the signing of the nuclear Comprehensive Test Ban Treaty (CTBT) and the imposition of a zero-yield condition therein, which made it impossible to conduct tests to assure the reliability of the U.S. nuclear weapons stockpile. One of the results of that was the initiation of an extensive computation program at the U.S. national laboratories to replace some aspects of the process of testing with some forms of analysis. This is one of the reasons that the Advanced Super Computing Initiative (ASCI) came into being. The imperative of maintaining U.S. stockpile surety was sufficient reason to warrant and justify the costs of the acquisition of several advanced state-of-the-art supercomputers for the national security labs. While there have been other grand challenge problems, such as environmental and cryptographic concerns, cited as justification for new generations of supercomputers, none have been as compelling in the U.S. as the stockpile stewardship considerations.

In Japan no such national security justification for supercomputers was possible. However, Japan has a strong interest in environmental problems, particularly those associated with global warming from greenhouse gases. This interest led in part to a major initiative to build a supercomputer dedicated to making progress in understanding these environmental problems. The result is the Earth Simulator. This new supercomputer captured a great deal of attention around the world as it set new world records for performance, and the

Japanese have a justifiable right to be proud of this accomplishment. Although the ES has called attention to the Japanese efforts in HEC, it is only one element in an extensive effort by the Japanese government to bring about a comprehensive computational capability that is intended to ensure an ability to remain at the leading edge in various areas of science and technology, which it regards as vital to its long-term economic well-being and progress.

Table 1.2
Sites Visited in Japan

Site	Panelists	Date
Frontier Research System for Global Change (FRSGC)	Biswas, Chien, Meacham, Trivelpiece, Yelick	29 March 2004
National Institute for Fusion Science (NIFS)	Dongarra, Paul	29 March 2004
Earth Simulator Center	Biswas, Chien, Meacham, Trivelpiece, Yelick	29 March 2004
Council for Science and Technology Policy (CSTP)	Chien, Dongarra, Meacham, Trivelpiece, Yelick	30 March 2004
University of Tokyo	Biswas, Paul	30 March 2004
Japan Aerospace Exploration Agency (JAXA)	Biswas, Meacham, Paul	30 March 2004
Ministry of Economy, Trade and Industry (METI)	Chien, Dongarra, Meacham, Trivelpiece, Yelick	30 March 2004
Tokyo Institute of Technology	Biswas, Dongarra, Miyahara	31 March 2004
Fujitsu	Biswas, Dongarra, Miyahara	31 March 2004
University of Tsukuba	Chien, Meacham, Paul, Trivelpiece, Yelick	31 March 2004
High Energy Accelerator Research Organization (KEK)	Chien, Meacham, Paul, Trivelpiece, Yelick	31 March 2004
National Institute of Advanced Industrial Science and Technology, Grid Technology Research Center (AIST-GRID)	Chien, Paul, Meacham, Yelick, Trivelpiece	31 March 2004
Research Organization for Information Science and Technology (RIST)	Biswas, Dongarra, Miyahara, Yelick	1 April 2004
Institute of Physical and Chemical Research (RIKEN)	Chien, Meachma, Paul, Trivelpiece	1 April 2004
IBM	Biswas, Dongarra, Miyahara, Yelick	1 April 2004
Ministry of Education, Culture, Sports, Science and Technology (MEXT)	Chien, Dongarra, Trivelpiece, Yelick	1 April 2004
National Institute of Informatics (NII)	Chien, Meacham, Paul, Trivelpiece	1 April 2004
Hitachi, Ltd.	Biswas, Dongarra, Miyahara, Yelick	1 April 2004
NEC Corporation	Biswas, Chien, Dongarra, Yelick	2 April 2004
Japanese Atomic Energy Research Institute (JAERI)	Meacham, Paul, Trivelpiece	2 April 2004
Sony Computer Entertainment, Inc.	Biswas, Chien, Dongarra, Yelick	2 April 2004

This report documents the observations and findings a group of scientists and engineers with experience in HEC and related programmatic issues based on a one-week visit to leading laboratories, government agencies, and industrial organizations.

The delegation's arrival in Japan coincided with the beginning of the fiscal year on the first of April. This event was a prominent element in some of our discussions, because it was not just the routine beginning of a new budget year. Rather, this year’s new budget also came along with some dramatic changes in the structure of how the government funds certain universities and research institutions. The principal element was a move

toward "privatization" of institutions that have historically been funded directly by government and staffed by individuals who were direct employees of the central government. "Privatization" is used in a different context than would be understood in the U.S. The intent that underlies these moves is the desire to reduce the number of central government employees, and lessen the direct control of the central government on budget decisions that influence government-funded research and development. This is not a recently made decision, but has been part of an overall long-range Japanese government plan to focus on certain areas of science and technology as part of a strategy to provide stimulation to Japanese business enterprises. This privatization confers new freedoms on faculty and research scientists, but at the same time burdens them with some new fiscal responsibilities to control costs. This move to privatization and its effects on the long term remain to be seen, but they will have an influence on the thrust of science and technology funding generally and for HEC in particular.

The "Unofficial Version" of the "The Science and Technology Basic Plan" (2001-2005) adopted by the Japanese government, contains an element that lays out the plan for information and telecommunications. Some excerpts follow:

> In R&D in IT area, the level of Japan is considered to be superior to that in European countries and the United States, especially in mobile-phone systems, optical communication technology, and IT terminals. The United States, however, leads the world in both PCs and their related technology and in software technologies.
>
> In this area there are a great variety of needs and technologies innovating rapidly, so that Japan will promote R&D with mobility. It is also important to promote R&D concerning common technologies necessary to realize an advance IT network society in which people can use their capabilities to the maximum in a creative way through freely sending, receiving, and sharing of information. Specifically Japan will focus on the followings:
>
> - advanced network technology that enables all network activities to be performed safely, at any time, at any place, and without stress.
> - high performance computing technology that enable rapid analysis process, storage and search of a tremendous amounts of distributed information
> - human interface technology that allows everyone to enjoy the benefits of an IT society without mastering complicated equipment and feeling stress
> - device technology and software technology to support the foregoing points

PREVIEW OF REPORT

In Chapter 2 Jack Dongarra introduces the Japanese Earth Simulator design and some of its initial achievements. In Chapter 3 the role of the main Japanese government agencies in setting policies for HEC is reviewed by Alvin W. Trivelpiece. There are two chapters on various scientific applications of HEC in Japan: Chapter 4 by Rupak Biswas, and Chapter 5 by Peter Paul. Another section (Chapter 6) by Jack Dongarra reviews the supercomputer offerings of the three main Japanese companies: NEC, Fujitsu, and Hitachi. Katherine Yelick has also provided two sections: Chapter 7 on software for HEC in Japan, and Chapter 8 on grid computing in Japan.

Appendix A lists the bios of the panelists. Appendix B is a compilation of the detailed reports from the 22 sites visited. Appendix C provides an update on the Top500 data, which was released in November 2004. The highlights from the U.S. workshop, held on May 25, 2004, are provided in Appendix D. A glossary is in Appendix E.

CHAPTER 2

THE EARTH SIMULATOR

Jack Dongarra

The Earth Simulator (ES) is a high-end general-purpose parallel computer focused on global environment change problems. The goal for sustained performance of the Earth Simulator was set to 1,000 times higher than that of the most frequently used supercomputers around 1996 in the climate research field. On a series of real climate applications the Earth Simulator provides a 50-fold computational performance advantage. It represents maximum-capability computing applied to a targeted attempt at a scientific breakthrough in a specific area. It is the result of a focused, long-term, top-down Japanese design effort. Access to the ES is only possible on-site in Japan and through research collaborations with Japanese scientists. In November 2004, two American supercomputers regained the lead in the Top500 list, but until then the Earth Simulator certainly allowed higher resolution and quicker response times in its applications than those available to American scientists.

BACKGROUND

In July 1996, as part of the Global Change Prediction Plan, the promotion of research and development for the Earth Simulator plan was reported to the Science Technology Agency, based on the report titled "For Realization of the Global Change Prediction" made by the Aero-Electronics Technology Committee. In April 1997, the budget for the development of the Earth Simulator was authorized to be allocated to the National Space Development Agency of Japan (NASDA) and Power Reactor and Nuclear Fuel Development Corporation (PNC). The Earth Simulator Research and Development Center was established, with Dr. Hajime Miyoshi assigned as the director of the center. Discussions were made regarding the Earth Simulator Project, at meetings on the Earth Simulator under the Computer Science Technology Promotion Council (Chairman: Prof. Taro Matsuno of Hokkaido University), which was held six times from March to July in 1997. With a report on "Promotion of Earth Simulator Project," specific proposals were put forward to the Science and Technology Agency.

The conceptual system design of the Earth Simulator proposed by NEC Corporation was selected by bidding. Japan Atomic Energy Research Institute (JAERI) had joined the project for PNC.

Under the Computer Science Technology Promotion Council, the Earth Simulator Advisory Committee was instituted with seven members with profound knowledge in the area (Chairman: Prof. Yoshio Oyanagi of Tokyo University). From June to July in 1998, five meetings were held on the basic design. As a result, on the 24th of August it was confirmed with “The Evaluation Report for the Basic Design of the Earth Simulator.” In February 1999, the Japan Marine Science and Technology Center (JAMSTEC) joined in the Earth Simulator development project and decided to build the Earth Simulator facility in the Kanazawa ward in Yokohama, which had been the industrial experiment station of Kanagawa prefecture. Manufacturing the Earth Simulator began in March 2000, under NASDA, JAERI, and JAMSTEC. After completion, the entire operation and management of the Earth Simulator was decided to be solely charged by JAMSTEC. At the end of February in 2002, all 640 processor nodes (PNs) started up operation for initial check-up. The Earth

Simulator Research and Development Center verified and gained the sustained performance by AFES (an Atmospheric General Circulation Model for ES), recording 7.2 Tflop/s with 160 PNs, more than 1.44 times faster than the target performance, 5 Tflop/s. The Earth Simulator Center (ESC) with Director-General Dr. Tetsuya Sato began the actual operation in March 2002. On May 2, 2002 the Earth Simulator achieved sustained performance of 26.58 Tflop/s by using the AFES. The highest performance record of 35.86 Tflop/s was achieved with the Linpack Benchmark on the next day. ES was honored with the top position on the Top500 list in June 2002.

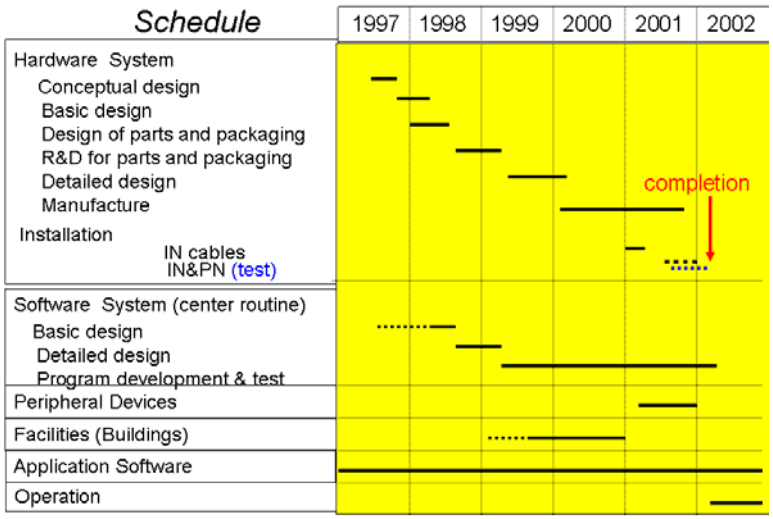


Figure 2.1. Earth Simulator Development Schedule (Courtesy JAMSTEC)

The ES was conceived, developed, and implemented by Hajime Miyoshi. Dr. Hajime Miyoshi was born in Tokyo in 1932. He graduated from the Mathematics Department of the Faculty of Science and Engineering of Waseda University in 1956. He joined the Science and Technology Agency (STA) in 1958 and later moved to the National Aerospace Laboratory (NAL) and became the Director of its Computer Center. He joined STA's Research Institute for Science & Technology (RIST) in April 1995 and became the Director of the Earth Simulator Research and Development Center in April 1997. Hajime Miyoshi is regarded as the Seymour Cray of Japan. Unlike his peers, he seldom attended conferences or gave public speeches. However, he was well known within the HEC community in Japan for his involvement in the development of the first supercomputer in Japan, the 22 Mflop/s Fujitsu FACOM230-75AU, which was installed at the National Aerospace Laboratory (NAL) in August 1977. He stayed at the forefront of HEC development in Japan. He led the development of the Numerical Wind Tunnel (NWT) at NAL. This won a Gordon Bell prize at the annual IEEE Supercomputing Conference in 1994. In 1997 he took up his post as the director of the Earth Simulator Research & Development Center (ESRDC) and led the development of the 40 Tflop/s Earth Simulator. Hajime Miyoshi had profound influence over the Japanese vendors and his work contributed much to the success of today's Japanese HEC industry. With the success of the FACOM230-75AU, Fujitsu developed their first commercial supercomputer, the VP100, which was installed at Nagoya University in the fall of 1983. The development of the Numerical Wind Tunnel helped Fujitsu to develop their VPP500 supercomputers, which were announced in 1992.

Dr. Miyoshi was appointed as Deputy Director-General in 1992 and resigned from NAL in March 1993 as he turned 60, the enforced retirement age for government employees. After he left the aerospace engineering community, he targeted the next supercomputer. He foresaw Computational Fluid Dynamics (CFD) technology achieving a matured status, and he believed CFD application could easily be done at the Numerical Wind Tunnel and following commercially based machines. He understood that global weather and ocean circulation simulation lacked the power of even the NWT. These areas were becoming more important as the global environmental issue deepened in importance. So his final target was to

Figure 2.2. Dr. Miyoshi

develop the Earth Simulator, which would serve as powerful computational engine for global environmental simulation.

Prior to the ES, global circulation simulations were made using a 100km grid width, though without coupling ocean and atmospheric models. To get quantitatively good predictions for the evaluation of environmental effects may require at least 10 km grid width or 10 times finer meshes in x, y and z directions and interactive simulation. Thus a supercomputer 1000x faster and larger than a 1995 conventional supercomputer might be required. Miyoshi investigated whether such a machine could be built in the early 2000s. His conclusion was that it could be realized if several thousand of the most advanced vector supercomputers of approximately 10 Gflop/s speed were clustered using a very high-speed network. He forecasted that extremely high-density LSI integration technology, high-speed memory and packaging technology into small-size, high-speed network (crossbar) technology, as well as an efficient operating system and Fortran compiler all could be developed within the next several years. He thought only a strong initiative project with government financial support could realize this kind of machine.

In 1997 he organized the ES Research & Development Center, which was financially supported by JAMSTEC, JAERI and NASDA, all of them belonging to the former STA, which has now been merged into the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of the Japanese government. He became the head of the ES R&D Center.

The machine was completed in February, 2002 and presently the entire system is working as an end user service. He supervised the development of NWT Fortran as the leader of NWT project and organized HPF (High Performance Fortran) Japan Extension Forum, which is used on the ES. He knew that a high-level vector/parallel language is critical for such a supercomputer.

ARCHITECTURE

The Earth Simulator is a highly parallel vector supercomputer system consisting of 640 processor nodes and an interconnection network. Each processor node is a shared memory parallel vector supercomputer, in which eight arithmetic vector processors (AP) are tightly connected to a main memory system. The vector processor has a theoretical peak of 8 Gflop/s, and main memory of 16 GB. The total system consists of 5,120 vector processors. This gives the system a theoretical peak of 40 Tflop/s and 10 TB of memory capacity.

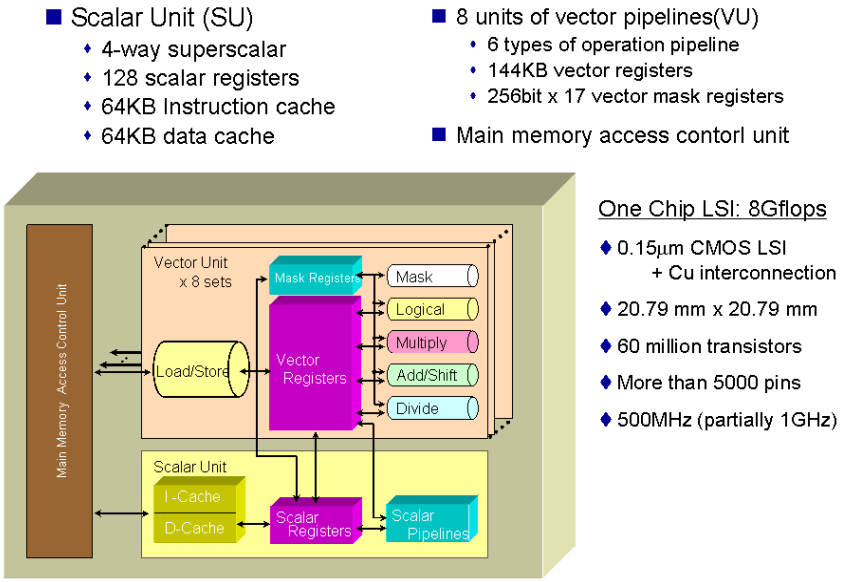


Figure 2.3. Arithmetic Processor Configuration (Courtesy JAMSTEC)

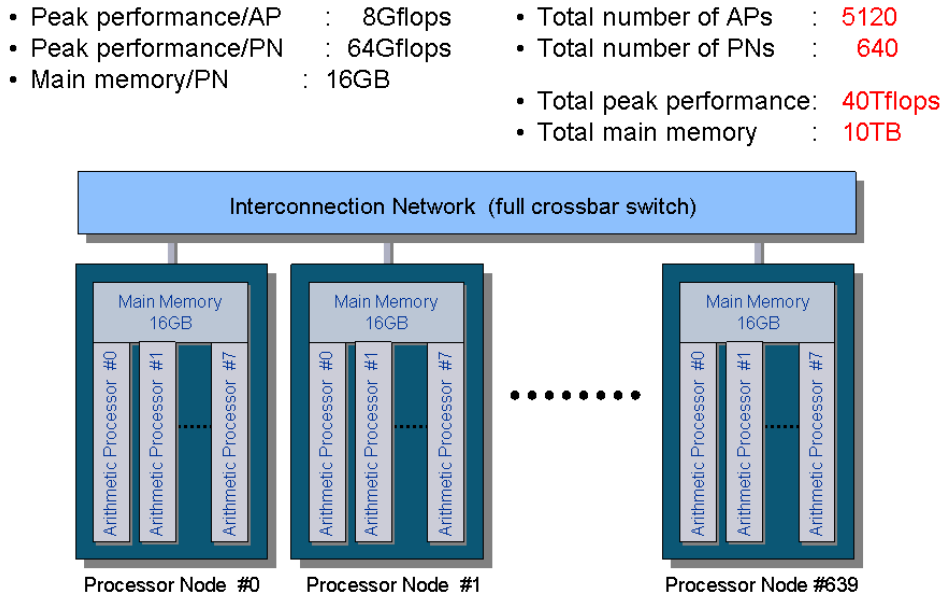


Figure 2.4. Configuration of the Earth Simulator (Courtesy JAMSTEC)

The interconnection network is a 640 by 640 non-blocking crossbar switch that connects the 640 processor nodes. The interconnection bandwidth between every two nodes is 12.3 GB/s in each direction. The aggregated switching capacity of the interconnection network is 7.87 TB/s.

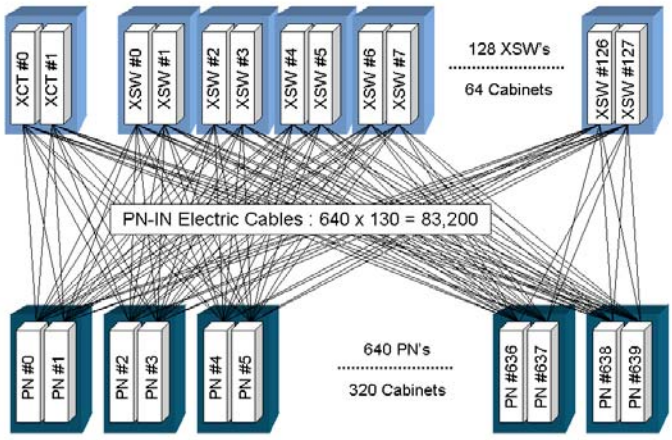


Figure 2.5. Connection between Cabinets (Courtesy JAMSTEC)

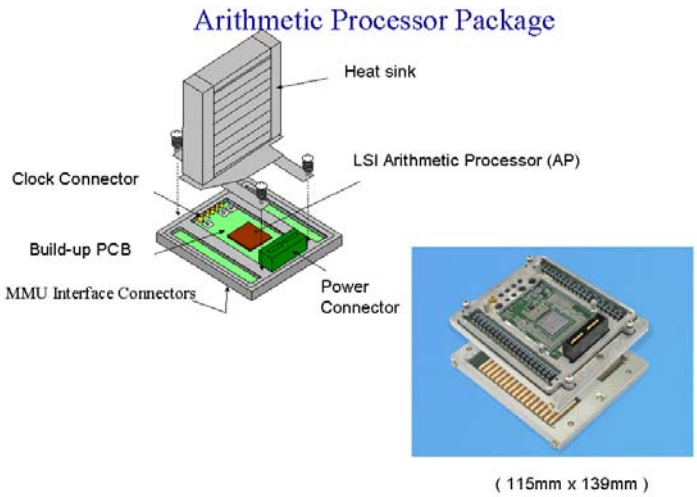


Figure 2.6. Arithmetic Processor Package (Courtesy JAMSTEC)

The vector processor is a single chip processor. The LSI uses 0.15 µm CMOS (Complementary Metal-Oxide Semiconductor) technology with copper interconnections. The vector pipeline unit operates at 1 GHz while the other components operate at 500 MHz. The processor is air-cooled and dissipates at approximately 140 Watts.

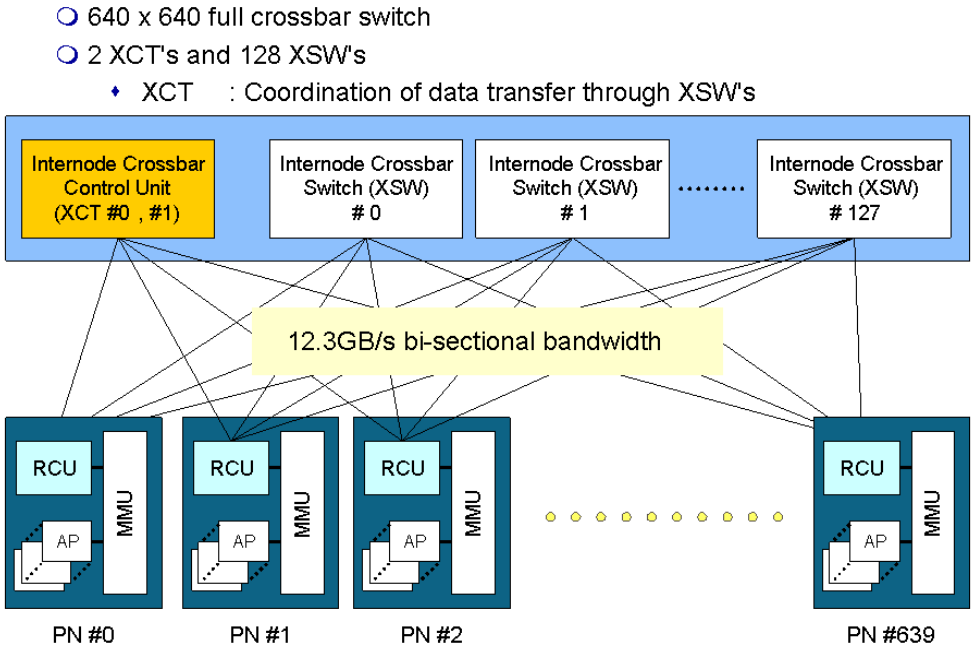


Figure 2.7. Interconnection Network (IN) (Courtesy JAMSTEC)

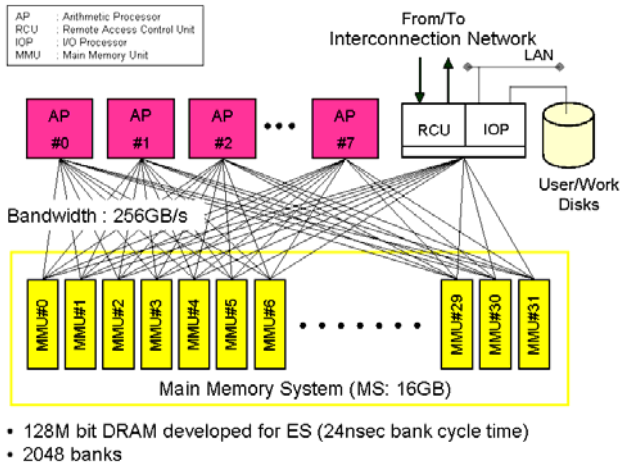


Figure 2.8. Processor Node Configuration (Courtesy JAMSTEC)

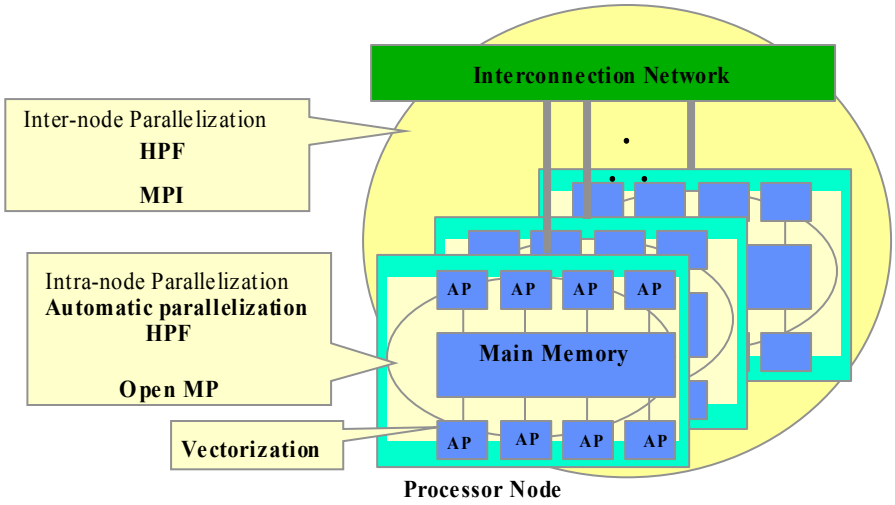


Figure 2.9. Vectorization and Parallelization (Courtesy JAMSTEC)

SOFTWARE

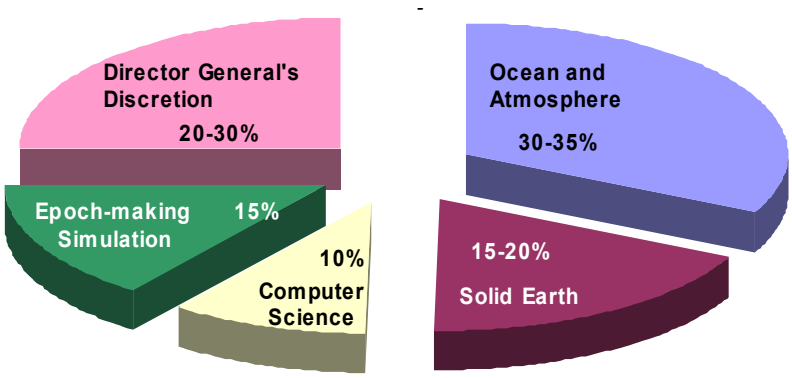
Applications

One of the strengths of the Earth Simulator project is its focus on a particular important class of problems. An attempt to understand large-scale phenomena on the earth such as global warming and El Niño events by numerical simulations is a very significant and challenging effort, in addition to the promotion of computational science and engineering. The Earth Simulator project had been started by the STA in 1997 aiming to understand and predict global environment change on Earth. The ESRDC was a joint team established by NASDA, JAERI, and JAMSTEC.

Between July 2002 and March 2003, there were only five whole system failures. Twice the problems arose with the Interconnection Network Nodes and three times with the job scheduling operation software. This is minimal trouble considering the Earth Simulator consists of 640 processor nodes and it has an enormous number of parts: 5120 vector processors, 1 million memory chips, 20,000 memory controller chips, 160,000 serial-parallel translate chips and so on. The storage system consists of RAID5 disk arrays, which helps to save user data from disk failures. Some small errors happen every day. However, practically every error is

recovered by using a correcting system or by retrying the hardware and software. The diagnosis reports are checked and preventive maintenance is done on nodes in which some recoverable errors have happened.

One of the most important problems is how to store and move massive amounts of user data. This problem has two sides. One is how to provide processing nodes with data from permanent storage systems. Another is how to retrieve data from the Earth Simulator. The ES has such a high performance level that the simulation data itself proves enormous in size. It has a tape cartridge system, which provides petabyte-class storage. However, the accessibility of the tape system is not so convenient which creates problems for users and wastes system resources. A hierarchical storage system, consisting of disks and tapes, was implemented at the ES in December 2003, which has improved this situation considerably.



Selected Projects		
Research fields	# of project in 2002	# of project in 2003
Ocean and Atmosphere	17	12
Solid Earth	8	9
Computer Science	4	2
Epoch-Making Simulation	11	11

Figure 2.10. Allocation of Computer Resources (Courtesy JAMSTEC)

USAGE

For an application to use a large number of processors on the ES a number of conditions must be satisfied.

- Number of PNs for a job must be less than or equal to 10
  - This can be extended to 512 by the application, based on other conditions being satisfied.
- Wall clock time for a job is less than or equal to 12 hours
- Number of PNs × wall clock time is less than 1536 PN*hours
- The number of PNs can be expanded if the vectorization ratio is greater than or equal to 95% and the parallelization efficiency is greater than 50%

These conditions will insure that the machine achieves relatively high efficiencies.

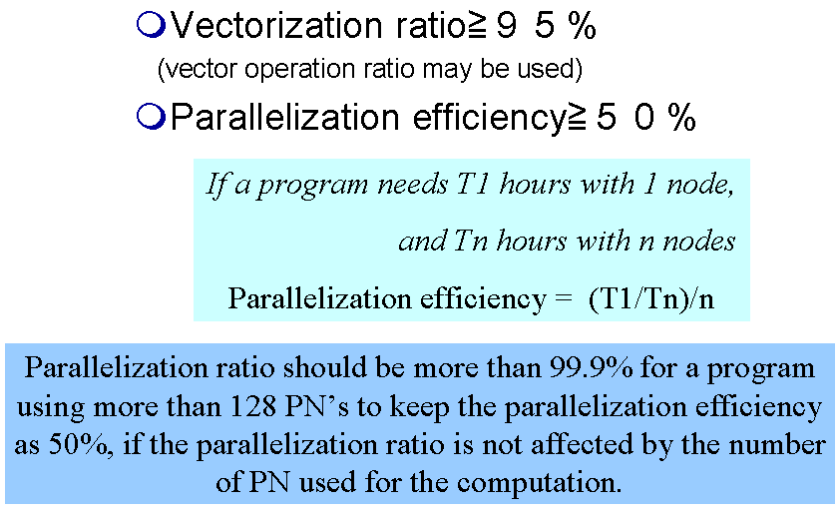


Figure 2.11. Condition to extend the PN number for a JOB (Courtesy JAMSTEC)

REFERENCES

Annual Report of the Earth Simulator Center, Outline of the Earth Simulator Project. 2003. <http://www.es.jamstec.go.jp/esc/images/annualreport2003/pdf/outline/outline.pdf> Last accessed February 23, 2005.

Habata, S., M. Yokokawa, K. Shigemune. 2003. The Earth Simulator System. *NEC Res. & Develop.*, Vol. 44, No. 1, <http://www.owlnet.rice.edu/~elec526/handouts/papers/earth-sim-nec.pdf> Last accessed February 23, 2005.

JAMSTEC, Earth Simulator Hardware. <http://www.es.jamstec.go.jp/esc/eng/ES/hardware.html> Last accessed February 25, 2005.

Sumi, A. 2003. The Earth Simulator and Its Impact for Numerical Modeling. < http://www.tokyo.rist.or.jp/rist/workshop/rome/ex-abstract/sumi.pdf> Last accessed February 25, 2005.

CHAPTER 3

POLICY CONSIDERATIONS THAT INFLUENCE HEC DEVELOPMENT IN JAPAN

Alvin W. Trivelpiece

INTRODUCTION

The previous decade has seen fundamental changes in Japanese government policy related to research and development (R&D) programs in science and technology (S&T). The source of these changes is to be found in the bursting of the Japanese economic bubble in the early 1990s. The economic downturn drastically reduced the R&D budgets of the country's high-tech industries, which in turn stifled overall S&T advancement. To reinvigorate S&T in Japan, the Diet passed the Science and Technology Basic Law in November 1995, with the stated goal of

> "achiev[ing] a higher standard of science and technology … to contribute to the development of the economy and society in Japan and to the improvement of the welfare of the nation, as well as to contribute to the progress of S&T in the world and the sustainable development of human society … [by] comprehensively and systematically promoting policies for the progress of S&T." (Science and Technology Basic Law, 1995)

The Basic Law established guidelines for, and articulated the responsibilities of, the federal and local governments and the universities in formulating, implementing, and promoting S&T policies. Furthermore, it called for the establishment of a comprehensive Science and Technology Basic Plan to promote the development of S&T in all areas of society.

The first Science and Technology Basic Plan, covering JFY1996-2000, was adopted in April 1996. The plan called for the investment of ¥17 trillion over five years to improve Japan's S&T infrastructure—effectively doubling, by 2001, the amount invested a decade earlier. In addition, the Basic Plan included provisions designed to improve S&T education and facilities. (Blanpied, 2003; Dorman, 2002)

Despite the continuing economic recession, by the end of the first Basic Plan investment had actually exceeded the amount called for at the outset. To keep the momentum going, the Diet adopted, in March 2001, the second Basic Plan for JFY2001-2005. It called for an even greater investment, ¥24 trillion, over the life of the plan—though this time it tied annual funding to the performance of Japan's Gross Domestic Product (GDP). (Blanpied, 2003) The second Basic Plan called for computer-related R&D in the following areas:

- advanced networking technologies
- high-performance computing
- human interface technologies
- simulation in materials S&T
- genetics and bio-informatics
- evaluation of R&D system performance

Thus, high-end computing can be seen as an important part of the overall S&T policy of the Japanese government. (Science and Technology Basic Plan, 2001)

Organizational Restructuring

Two significant changes to the federal government's infrastructure accompanied the second Science and Technology Basic Plan in 2001: the transformation of the universities from federal institutions to quasi-private-sector organizations, and the restructuring of the government's S&T ministries and agencies to improve coordination across organizational boundaries. Both of these changes continue to have an impact on HEC R&D in Japan.

Beginning in early January 2001, the government designated several university research institutions as Independent Administrative Institutions (IAIs). The closest analogy in the United States would be quasi-privatization, whereby the government funds some budget items and the private sector funds others. The twin goals behind the creation of the IAIs are to decrease the number of people directly employed by the Japanese national government (university professors remain government employees), and to increase the total amount of R&D money being invested.

The change in status of universities from government institutions to IAIs thus removes the legal prohibition on government employees from engaging in contracts with the private sector. In theory, professors would thus be free to seek industry funding for S&T research projects, while at the same time government agencies would be less able to exert directional influence on S&T activities by virtue of their financial support. Though the government still pays professors under the IAI arrangement, the plan anticipated that administrative costs would be reduced annually. However, an apparent disadvantage of the IAI plan is that it eliminates, or at least modifies, the established university practice of replacing leased computers with newer models every five years or so.

Concurrently with the implementation of the IAI plan, Japan's national government underwent substantial reorganization. Among the most dramatically affected ministries were those related to S&T. A new cabinet-level department, the Council for Science and Technology Policy (CSTP), was established to provide overall policy development and guidance. From a merger of the former Ministry of Education and Science (Monbusho) and the former Science and Technology Agency of Japan (STA), the Ministry of Education, Culture, Sports, Science and Technology (MEXT) was created. This organization became the government's primary S&T ministry, since prior to the reorganization Monbusho and STA were the recipients of approximately two-thirds of the government's S&T funding. (Blanpied, 2003) While at first glance it might appear unusual for a ministry to combine education with S&T, the fact that the country's supercomputers are primarily located at universities helps explain this administrative arrangement. Finally, the former Ministry of International Trade and Technology (MITI) was reconstituted as the Ministry of Economy, Trade and Industry (METI), with a primary emphasis on technology transfer and commercial applications of S&T.

This chapter looks at how these three Japanese government agencies are organized to guide computer R&D, with a particular focus on HEC. The following sections review each agency's jurisdiction and mission; discuss each agency's role in computing in general and HEC in particular; and, where possible, identify each agency's future intentions based on publicly available documents and on the information obtained during our site visits.

THE COUNCIL FOR SCIENCE AND TECHNOLOGY POLICY (CSTP)

There is no equivalent agency to CSTP in the U.S. government. Although much information is available describing the functions of the CSTP, its pivotal role in supporting and directing R&D in Japan may not be familiar to many readers.

The mission of the CSTP is to "steer S&T policies in Japan with foresight and mobility, acting as a control tower under the Prime Minister's leadership, eliminating administrative sectionalism, and steadily implementing the policies described in the Basic Plan." (Science and Technology Basic Plan, 2001) In practice, this amounts to developing the five-year Basic Plans and working with the powerful Ministry of Finance to review S&T budget requests and making recommendations on funding priorities to the Prime Minister and the Diet.

The *de jure* chair of the CSTP is the Prime Minister, while the *de facto* chair is the Minister of State for Science and Technology, who thereby has substantial leverage in negotiations with other government S&T ministers. The council body consists of seven appointed Executive Members—three permanent and four appointed for terms of two years—plus the Chief Cabinet Secretary, the President of the Science Council of Japan, and the heads of MEXT, METI, the Ministry of Finance, and the Ministry of Public Management, Home Affairs, Post and Telecommunications. (Blanpied, 2003) The seven Executive Members are distinguished individuals with experience in academic institutions and industrial organizations. The CSTP generally meets on a monthly basis; Prime Minister Koizumi usually attends their meetings. In addition to the regular members, the CSTP relies on the input of seven "expert panels" for technical evaluations. The members of these panels are drawn from the external scientific and engineering communities.

The operation of the CSTP is determined by the priorities established in the Basic Law, including:

- life sciences
- environmental sciences
- energy and infrastructure
- information and communications
- materials
- nanoscience
- manufacturing
- the "frontiers," such as the oceans and space

In the U.S., many industrial organizations have the ability to fund projects through internal research and development (IRD), while some of the national laboratories do so through laboratory-directed research and development (LDRD). These funds have certain constraints imposed on their application, but generally they can be given to an organization to allow them to emphasize programs that might lead to new projects or products. In some respects, the CSTP operates as a source of IRD and LDRD funding, in that it reviews projects proposed in the course of the regular budget cycle to determine if they are consistent with the Basic Law priorities.

Although the CSTP members may not be able to influence budgets significantly, they review and rank all project budget proposals and programs using a four-level rating system. Projects in the lowest category are not funded. The rest are ranked according to importance. When the ranking has been established and agreed upon, the Council makes its recommendations to the Prime Minister. In Japan's parliamentary government, the recommendations represent the end of the deliberation process, unlike in the U.S. where the Congress plays a key role in the final budget.

The budgetary process of the Japanese government is based in part on the expectation that the R&D funded by the government will enable academic and research institutions to implement programs and projects that will lead to improved economic circumstances for their citizens. This approach does not appear to emphasize what many would consider to be pure or basic investigator-driven studies. Even so, there are strong, well-funded programs in high-energy physics and other areas that would be categorized as quite basic or fundamental.

CSTP's Role in Computing

CSTP ranks the funding requests of ministries and IAIs for computing-related projects, as well as for all R&D projects related to S&T. Therefore, CSTP plays a critical role in determining whether individual computer R&D programs in Japan will survive, grow, retrench, or cease altogether.

CSTP views HEC as part of the broader information technology (IT) field. From that viewpoint, the CSTP has identified three important computer-related areas:

1. highly reliable IT
2. human interface systems
3. quantum computing

CSTP ranks supercomputing and HEC networks, broadband Internet, and low-power device technology as among its highest priorities for future R&D. CSTP has noted that R&D efforts in HEC are directed towards hardware, software, and middleware.

Requests for JFY2005 were announced in October 2004. For the first time, the CSTP reviewed all proposed S&T projects, whereas previously it had reviewed projects only above a specified amount (¥1 billion in 2003 and ¥2 billion in 2002). Table 3.1 lists key computer-related projects approved for funding in JFY2005 for METI, MEXT and the Ministry of Internal Affairs and Communications (MIC) (Shinohara, 2004).

Table 3.1
CSTP Rankings and Funding of Key Computer Projects

Rating*	Project	Ministry	JFY2005 Request (¥ Million)	JFY2004 Budget (¥ Million)
A	Ubiquitous Network	MIC	3,105	3,105
S	Next-Generation Backbone Network for Internet	MIC	2,000	0
B	Shift of Internet to IPv6	MIC	1,752	1,752
B	Advanced Network Identification Technologies	MIC	1,100	1,040
A	Early-Stage Caution for Computer Security	METI	1,350	0
A	Basic Software for e-Society	MEXT	1,100	1,100
A	Super High-Speed Computer Network Project (National Research Grid Initiative)	MEXT	1,950	1,950
S	Technologies for Future Supercomputing	MEXT	2,000	0
A	Business Grid Computing	METI	2,600	2,501
B	Manufacturing Technologies for Advanced Integrated Circuits, Including EUV (Extreme Ultraviolet) Light Source	MEXT	1,703	0

* = S: Very important projects to be implemented proactively

A: Important projects to be implemented

B: Some problems to be solved but efficient and effective implementation is expected

C: To be reviewed

Future Intentions

CSTP does not develop policies *per se*, but rather develops budget guidelines based on policies that are proposed "from the bottom up," i.e. by the government ministries and their constituencies themselves. The Basic Plans thus reflect a consensus vision of the government ministries and IAIs. With regard to a next-generation Earth Simulator (ES), for example, the need for such a system by a wide spectrum of R&D applications would have to be demonstrated. While CSTP is currently working with R&D organizations to develop new ideas for the next generation ES, a final decision has not been made yet. Grid computing, clusters, and a combination of different architectures are all future candidates for CSTP support.

MINISTRY OF EDUCATION, CULTURE, SPORTS, SCIENCE AND TECHNOLOGY (MEXT)

The Ministry of Education, Culture, Sports, Science and Technology (MEXT) is responsible for overseeing elementary, secondary, and higher education; developing S&T policy; and fostering sports and cultural programs. As mentioned earlier, MEXT accounts for the lion's share of government S&T funding because it inherited the former Monbusho and STA upon its creation in January 2001. MEXT has subsumed basic and applied research, while its sister ministry, METI, is oriented primarily toward developments related to commerce.

The Science and Technology Policy Bureau is one of ten administrative units within MEXT. The Bureau is responsible for a wide variety of functions, including the following: (MEXT website, 2004)

- planning and drafting basic S&T policies
- formulating R&D programs
- research evaluation
- researcher and technician training
- promotion of regional S&T programs
- increasing society's understanding of S&T
- promoting a comprehensive international research exchange policy
- promulgating safety systems for experimental nuclear reactors and radioactive isotopes

In 2002, MEXT established a Center of Excellence (COE) Program (also called the Toyama Plan, after MEXT's Minister, Atsuko Toyama). The COE Program provides competitive grants to universities in five research areas every year. MEXT allocated a total of ¥18.2 billion for the program in JFY2002. (Blanpied, 2003)

MEXT's Role in Computing

In Japan, the universities are the primary centers of HEC R&D; because MEXT's purview includes the education system, the ministry has jurisdiction over the universities. MEXT also has jurisdiction over many of the science programs that use HEC: the Japanese space program, atomic energy program, and marine science program, and others. MEXT has stated its desire to provide enough R&D funding to help ensure Japanese competitiveness in HEC. For example, the Ministry provided ¥5.9 billion to the Earth Simulator (ES) in 2004. As noted earlier in this chapter, MEXT also sought, and received, nearly ¥2 billion for the Super High-Speed Computer Network Project in JFY 2005 and ¥2 billion to develop technologies for future supercomputing. (Shinohara, 2004)

Future Intentions

The WTEC panel was told that MEXT has no specific plans to build a follow-on to the ES system, nor does it plan to fund upgrades to it. This is in part due to budget limitations. The Ministry does not intend to subsidize Japanese supercomputing companies; rather, it is investing in grid computing and long-term research areas such as quantum computing and nanotechnology and the life sciences. MEXT sees insufficient markets for supercomputers in Japan, though it recognizes the need for them at research institutions. The Ministry still has questions about the types of research problems that require supercomputers, as opposed to lower-cost clusters. Furthermore, there are no military applications to drive supercomputer development. Instead, applications are driven by efforts to model environmental, geological, and seismic activities.

In the area of software for supercomputing, MEXT is funding supercomputing software development at the University of Tokyo. MEXT is actively researching whether improved software support will allow PC clusters to be as usable as supercomputers.

MINISTRY OF ECONOMY, TRADE, AND INDUSTRY (METI)

The Ministry of Economy, Trade and Industry, as its name suggests, develops and implements government policies relating to trade, industry and the economy. Like MEXT, the Ministry was created in January 2001 by the reorganization of an existing Ministry, in this case the former Ministry of Commerce and Industry (MITI). There are seven administrative divisions within METI, namely the Minister's Secretariat and six Bureaus covering: (METI website, 2004)

- economic and industrial policy
- trade policy
- trade and economic cooperation
- industrial science and technology policy and the environment

- manufacturing industries
- commerce information policy

In addition, METI can call on the services of several *ad hoc* advisory councils when formulating specific policies. The councils consist of experts from a cross-section of applicable professional communities, including industry, finance, labor, academia, and the media. The councils issue reports that are consulted by the Ministry during policy development deliberations.

METI is the parent agency for the National Institute for Advanced Industrial Science and Technology (AIST). METI funds account for between 10 and 15 percent of the R&D budgets of AIST's institutes and centers, including the AIST Grid Technology Research Center (AIST-GRID). Funding is allocated on the basis of annual evaluations, and METI also provides bonuses to individuals whose research is deemed of sufficient merit. (Blanpied, 2003)

METI is active in a number of initiatives to strengthen S&T in Japan. For example, it provides funding support to university-based Technology Learning Organizations (TLOs), which are essentially technology incubators. University faculty who participate in TLOs are encouraged to patent and license their discoveries. METI has also launched an Industrial Cluster Program, which is intended to facilitate transfer of technologies and skills among universities, private industry, and research organizations at the prefectural level. (Blanpied 2003)

METI's Role in Computing

The WTEC panel was told that METI is no longer interested in supercomputing, citing a lack of sustainable commercial or military applications and the lack of a domestic source for microprocessors. This attitude is largely due to experience resulting from trade restrictions between Japan and the United States ten years ago. U.S. law at the time prohibited the importation of Japanese supercomputers into American markets. As a result, Japanese supercomputer manufacturers not only experienced financial losses, but also lost ground in the industry as U.S. manufacturers began introducing supercomputers with RISC processors around that time. At the time, METI also invested in scientific computing, but decided that the market could not be sustained and withdrew from further funding.

As a result, METI has chosen to emphasize grid computing, which the Japanese have adopted as a cornerstone of their efforts to provide computing resources to government R&D programs. METI has provided ¥2.6 billion for the continued development high-speed business grid computing in JFY2005, as well as funding for computer security and digital tag initiatives.

The market for high-end computers in Japan is in the range of $400 - $500 million. Even so, Japan has difficulty is selling their computers in the US as a result of "Buy America Act" restrictions in certain situations. Though METI provides financial support to the three Japanese supercomputer companies (NEC, Fujitsu, and Hitachi), it has recommended to NEC and Fujitsu that they switch to building high-reliability systems that combine multiple processors and Linux-based software.

Future Intentions

METI plans to continue supporting, through funding and policy development, the development of grid computing for commercial applications, as well as new technologies such as the digital tag (envisioned as a replacement of barcodes), consumer applications such as computerized domestic appliances, medical computer systems, computer security, and international software initiatives such as Asia OSS (open source software).

REFERENCES

Blanpied, W. 2003. The Second Science and Technology Basic Plan: a Blueprint for Japan's Science and Technology Policy. *Special Scientific Report #03-02*. <http://nsftokyo.org/ssr03-02.html> Last accessed February 25, 2005.

Dorman, C. 2002. Observations on Japanese Science and Technology by Former ONR Chief Scientist. *Report Memorandum #02-01*. <http://www.nsftokyo.org/rm02-01.html> Last accessed February 25, 2005.

Government of Japan. 1995. The Science and Technology Basic Law (Unofficial Translation). <http://www8.cao.go.jp/cstp/english/law.html> Last accessed February 25, 2005.

Government of Japan, Ministry of Economy, Trade, and Industry (METI) web site. <http://www.meti.go.jp/english/> Last accessed February 25, 2005.

Government of Japan, Ministry of Education, Sports, Science, and Technology (MEXT) web site. <http://www.mext.go.jp/english/> Last accessed February 25, 2005.

Government of Japan. 2001. Science and Technology Basic Plan (2001-2005) (Unofficial Version). <http://www8.cao.go.jp/cstp/english/plan.html> Last accessed February 25, 2005.

Shinohara, K. 2004. Report Memorandum #04-09. "Rating of S&T-related Projects - JFY2005." <http://www.nsftokyo.org/rm04-09.html> Last accessed February 25, 2005.

CHAPTER 4

SCIENTIFIC APPLICATIONS OF HIGH-END COMPUTING I

Rupak Biswas

INTRODUCTION

Scientific and engineering applications provide the ultimate requirements and justifications for pursuing increasingly more powerful supercomputers and innovative high-end computing (HEC) technologies. The promise of scientific discoveries and breakthroughs, improvements in human life on our home planet, and reductions in design cycle times are some of the driving forces behind the international computational science and engineering community's push for petaflop/s-scale machines. For example, one of the primary reasons for building the Earth Simulator (ES) was that the problem area was extremely important to Japan. At inception (and even now), the ES was and is primarily devoted to solving problems in Earth sciences such as weather, climate, and earthquakes. Similarly, the Accelerated Strategic Computing Initiative (ASCI) program in the United States was primarily conceived to assist federal defense programs shift from testing to high-fidelity simulations for the nation's nuclear stockpile stewardship program.

This chapter and the next provide an overview of the wide spectrum of scientific applications that we heard about during our weeklong visit to Japan. The report should not be considered a comprehensive survey of Japanese work in scientific applications, but is instead meant to provide a current snapshot and assessment of high-end computing research and development activity in the various scientific domains. The institutions and centers that we visited all had mission-critical requirements in large-scale computations. Most scientists and researchers that we met were able to demonstrate significant progress in their respective domains that had been enabled by the few teraflop/s of computational capability that they currently have. At the same time, future computing requirements to make even modest scientific breakthroughs were at least an order of magnitude larger. This chapter is organized into three sections: Earth science applications, computational fluid dynamics calculations for aerospace problems, and computational nanotechnology. A second chapter on HEC scientific applications (Chapter 5) covers the areas of lattice gauge simulations, plasma physics calculations, and design of advanced nuclear reactors.

EARTH SCIENCES

No discussion about Japanese HEC activities can commence without alluding to the ES. Some background information, overall system architecture, operational practices, allocation strategy, and current usage are described in Chapter 2. But since the ES is the world's fastest general-purpose computer that is primarily focused on global environment problems, we begin our report on scientific applications by presenting Japanese research and development in the Earth sciences domain. About 70% of ES cycles are devoted to ocean and atmosphere modeling, and solid Earth simulations. A total of 21 of 34 projects in 2003 are in these two areas.

In 2002, the ES created a lot of excitement in the computational science and engineering community by achieving a sustained performance of 26.58 Tflop/s (65% of peak) when running a high resolution simulation (T1279L96) using a spectral atmospheric general circulation model code, called AFES, on all 640 processor nodes (Shingu et al. 2002). This level of efficiency surpassed that of all conventional HEC simulation applications, even on vector architectures, and remains the highest standard even today. AFES has been under development since 1995 and was specifically optimized for the ES. The code is based on a global hydrostatic model, and can predict wind speeds, surface pressures, temperatures, humidity, and cloud water content.

The Japan Agency for Marine-Earth Science and Technology (JAMSTEC) manages the ES, located in the Kanazawa ward of Yokohama. JAMSTEC is an independent administrative institution under the jurisdiction of the Ministry of Education, Culture, Sports, Science, and Technology (MEXT). Most of the research and development work in Earth science applications is conducted by the Frontier Research System for Global Change (FRSGC), currently co-located with the ES. FRSGC was established in 1997 as a joint project among the National Space Development Agency (NASDA), the Japan Atomic Energy Research Institute (JAERI), and JAMSTEC to integrate process research, observations, and simulations that meet the national goal of "Prediction of Global Change." In 2003, FRSGC's management was shifted solely to JAMSTEC.

FRSGC's primary objective is to contribute to society by elucidating the mechanisms of various global environmental changes as well as making better predictions of such changes. Our planet is affected in significant ways by various anthropogenic causes (e.g. greenhouse gas effects, loss of tropical rainforests), but society also remains vulnerable to natural disasters such as earthquakes, volcanic eruptions, and abnormal weather. With this goal in mind, researchers at FRSGC are developing and simulating high-fidelity models of the atmosphere, ocean, and land. These individual elements will subsequently be integrated to model Earth as a single system, and therefore be able to make reliable seasonal to decadal predictions of various phenomena on our home planet. FRSGC's activities are classified into six broad programs, four of which conduct disciplinary research on climate variations, hydrological cycle, atmospheric composition, and ecosystem change, while two conduct cross-cutting research on global warming and integrated Earth system modeling. Each of these areas is further subdivided into more focused research groups.

Perhaps the major thrust at FRSGC is the development of an integrated Earth system model, primarily for global warming prediction. The effort, nicknamed KISSME, is part of the Kyousei2 (RR2002) project led by FRSGC with active participation of other researchers from several Japanese universities and institutions. Figure 4.1 shows a block diagram of the overall framework and sample simulation results from some components. Here, biological and chemical processes important for the global environment interact with climate changes. The integrated model adds individual component models (such as oceanic carbon cycle) to atmospheric and ocean general circulation models. Eventually, the full model will include atmosphere (climate, aerosol, chemistry), ocean (climate, biogeochemistry), and land (climate, biogeochemistry, vegetation). Atmospheric models with fine resolutions will be able to reproduce meso-scale convective systems explicitly, while high-resolution ocean models will be capable of accurately reproducing ocean eddies. The project goals also include incorporating detailed cloud microphysics into a global non-hydrostatic model, adding an explicit ice sheet model to the ocean model, and improving the integrated model's representation of the middle atmosphere to better capture stratospheric chemistry.

As part of KISSME, a significant effort is underway to develop a new high-resolution (less than 10 km horizontal, at least 100 levels vertical) global cloud-resolving model using icosahedral grids. Unlike AFES, non-hydrostatic equations are used to model the system because of very fine resolution in the horizontal direction, allowing convective motions to be simulated explicitly. The spectral transform method is also inefficient in high-resolution simulations because of its massive data transfer requirements among processors. Moreover, the icosahedral grid eliminates the singularity problem at the poles. Development of the dynamical core (without any physical processes) of the 3D global model is complete. This model, called NICAM (Non-hydrostatic Icosahedral Atmospheric Model), conserves total mass and energy, uses the finite-volume method, and can be run either in fully explicit or hybrid (explicit horizontally, implicit vertically) mode. Quantitative comparisons with AFES are extremely promising; for example, there are no significant differences in location and intensity of jet, mean eddy heat flux, etc.

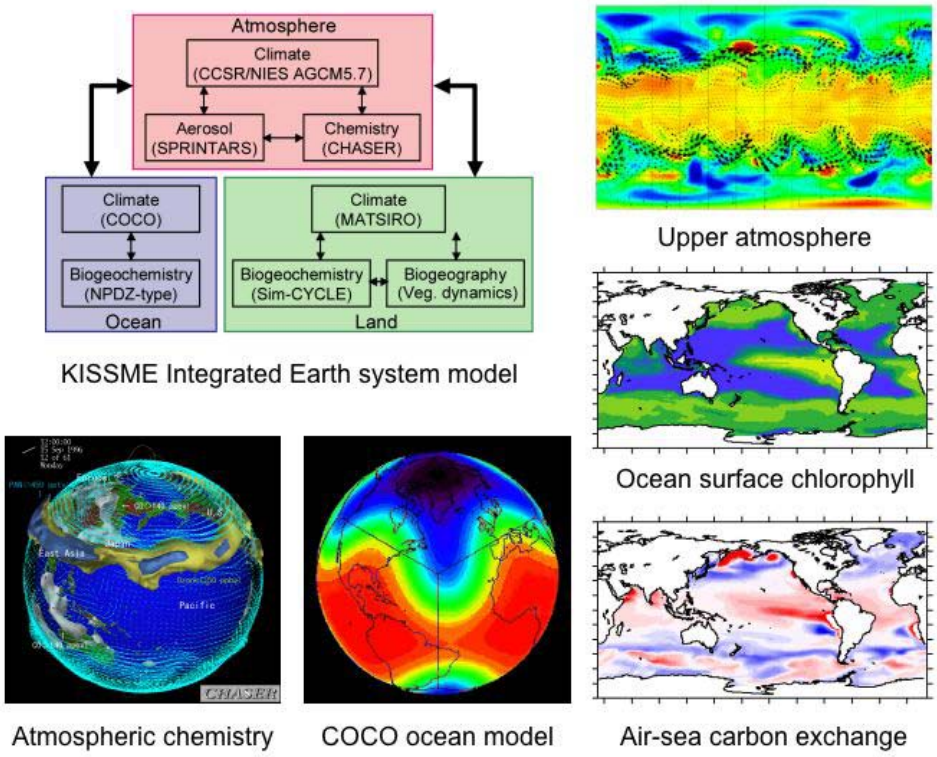


Figure 4.1. Integrated Earth system model framework and sample component results (Courtesy Taroh Matsuno, FSRGC)

In terms of parallel performance, NICAM is superior to AFES and demonstrates almost close to ideal speedup on the ES, using a fixed problem size. For instance, it requires 1.5 hours per simulation day on 320 ES processor nodes, when using a 3.5 km grid. This translates to a sustained performance of 9.75 Tflop/s (48% of peak). AFES execution time increases as $O(n^3)$ while for NICAM it only grows as $O(n^2)$, where n is the number of gridpoints. Hence, the advantage of NICAM over AFES increases as the resolution increases. NICAM is written in F90, uses MPI, and has been developed and performance-tuned by researchers over the last couple of years. NICAM will be integrated into FRSGC's next-generation atmospheric general circulation model (AGCM) and then into KISSME, along with other physical processes.

FRSGC researchers and their collaborators are also developing a high-resolution ocean model for eddy-resolving simulations, especially in high latitudes. This model, called COCO (CCSR Ocean Component model), uses a cubic grid system and focuses on efficient vectorization in order to achieve good computational performance. Energy and entropy are conserved in a high-order advection scheme. The dynamical core of the shallow water model is also incorporated. COCO has 85 levels, with a spacing of 50m at the surface, increasing to 200m near the bottom. Wind stress conditions are obtained by reanalyzing data from the European Center for Medium-range Weather Forecasting (ECMWF). Parallelization applies the domain decomposition technique on the six planes of the cubic grid.

Dr. Hajime Miyoshi, the force behind the ES, was the first Deputy Director-General of the Research Organization for Information Science and Technology (RIST), a non-profit, public service organization working for the development and utilization of computational science and engineering technology for several nationally important application areas. As a result, RIST's mission is to serve as a HEC technology catalyst by enabling large-scale simulations using the ES. In Earth sciences, RIST initiated most of the original research and parallelized various simulation models. Since 1997, at least 21 large simulation codes (10 for climate, 11 for seismology) have been co-developed by RIST, transferred to scientists, and run on the ES. RIST continues to have close interaction (in terms of code development, porting, and tuning) with people at JAMSTEC (both at FRSGC and the ES Center). Like FRSGC, RIST is also supported and funded by MEXT.

In this chapter, we selectively report work currently being conducted by RIST researchers in the Earth sciences domain. Large-scale modeling and simulation of typhoons, ocean surface temperature, clouds, and seismic wave propagation are some of the areas of interest. For instance, the JMA cloud simulation code uses one km non-hydrostatic models to successfully reproduce cloud bands extending southeast from the base of the Korean Peninsula over the Sea of Japan. RIST scientists are also using a fault model of the underground structure of southwest Japan to conduct high-fidelity simulations of seismic wave propagation. GNSS is a global non-hydrostatic simulation system that is primarily used to investigate cloud activity in tropical areas (45N to 45S). A 20 km test simulation has been completed but the ultimate target is to perform a two km simulation. CReSS (cloud-resolving storm simulator) performs large-scale high-resolution numerical simulations of clouds and meso-scale storms associated with severe weather systems, and is the only Japanese cloud model code available to the public. For example, CReSS is able to reproduce detailed structures of convective cells embedded in a rain band. The code is a hybrid of MPI, OpenMP, and vectorization; however, pure MPI (plus vectorization) has been shown to be more effective than MPI+OpenMP on the ES.

To reliably predict global Earth changes, RIST has been actively building parallel platforms where high-resolution models are integrated. The Fu-jin parallel framework couples atmospheric and ocean models to enable multi-scale simulations of salient Earth sciences features. These include downbursts and cloud models (local); oil spills, and typhoons (regional); and ocean circulation models (global). On the other hand, GeoFEM is their production-quality multi-purpose multi-physics parallel finite element framework for solid-Earth modeling. It allows Japanese Earth scientists to test various earthquake models and make hazard predictions. Sample results from these various areas are shown in Figure 4.2.

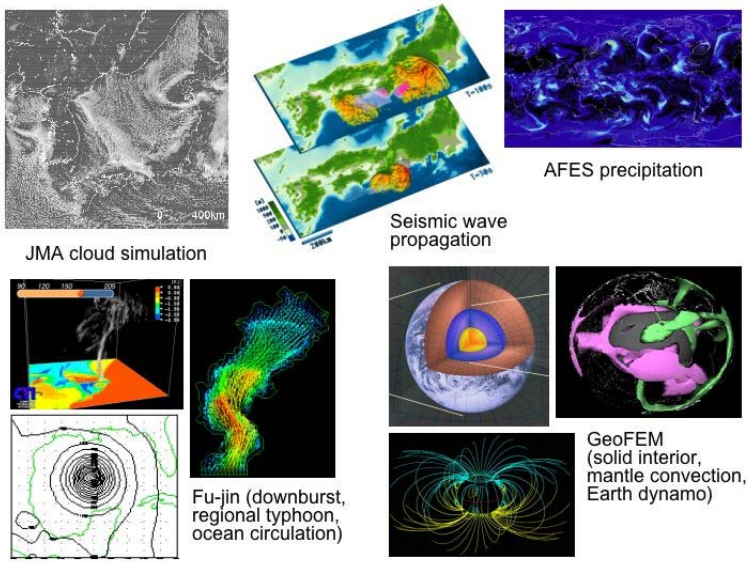


Figure 4.2. Sample Earth sciences models developed at RIST (Courtesy Hisashi Nakamura, RIST)

However, significant work in Earth sciences is also being conducted in the United States. To place the Japanese work in perspective, we provide a brief overview of related research and development activities primarily being led by the National Center for Atmospheric Research (NCAR) and the National Aeronautics and Space Administration (NASA). Note though that many other domain and computer scientists from academic institutions, national research laboratories, and government organizations are also key participants. Figure 4.3 shows visualization of simulation results from weather (DAS code), cloud (GCEM3D), and ocean and sea-ice (ECCO) modeling.

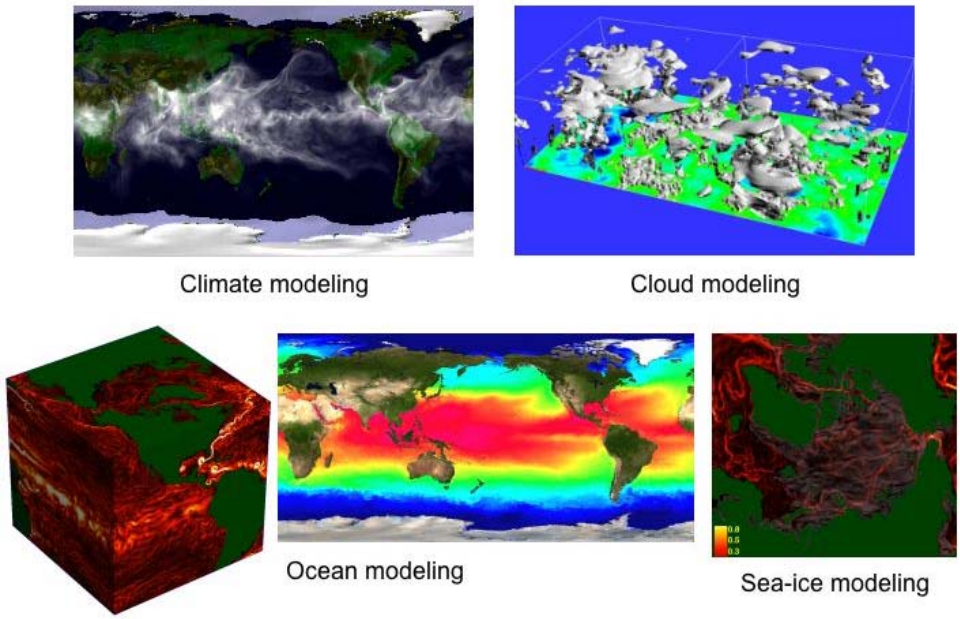


Figure 4.3. Sample simulation results from research conducted in the United States (Courtesy Robert Atlas, NASA Goddard Space Flight Center (GSFC); Wei-Kuo Tao, NASA GSFC; Dimitris Menemenlis, ECCO consortium)

Perhaps the most famous Earth sciences framework is the Community Climate System Model (CCSM), a fully coupled global climate model that enables accurate simulations of the Earth's past, present, and future climate states. The complexity of CCSM is evident from the schematic in Figure 4.4. The latest version, CCSM3, incorporates phenomena ranging from the effect of volcanic eruptions on temperature patterns to the impact of shifting sea ice on sunlight absorbed by the oceans. It integrates four atmospheric, oceanic, sea-ice, and terrestrial components: community atmosphere model (CAM), parallel ocean program (POP), community sea-ice model (CSIM), and community land model (CLM). The standard finest resolution for CAM and CLM in the production version of CCSM3 is 256 longitudinal and 128 latitudinal points (T85) in the horizontal direction with 26 vertical levels. For POP and CSIM, the finest longitudinal and latitudinal resolutions (gx1v3) are one degree and 0.3 degrees (at the equator). POP has 40 vertical levels associated with the gx1v3 resolution, with level thickness increasing monotonically from 10 m to 250 m. These are relatively coarse, hence they cannot capture many of the fine features. Runs at these resolutions also vectorize poorly because of relatively short vectors and do not scale well with an increasing processor count. Some of the individual models within CCSM are available as stand-alone versions and can be run at much higher resolutions. For example, POP has been run at 0.1 degree resolution for modeling the North Atlantic, and can simulate almost six years per day on a 512-processor 1.5 GHz Altix system. Since CAM uses a spectral method that has inherent scalability limitations, finite-volume versions of the dynamical core of the atmosphere model are being tested.

Besides CCSM, there are other multi-institutional collaborations and partnerships. For example, the Earth System Modeling Framework (ESMF) is a multi-agency effort to build a high-performance flexible software infrastructure to increase usability, portability, and interoperability in climate, weather prediction, data assimilation, and other Earth science applications. ESMF provides data structures and utilities to develop model components, and defines the architecture for composing multi-component applications. The Partnership for Advancing Interdisciplinary Global Models (PARADIGM) is another multi-institutional multi-disciplinary consortium committed to building and deploying new advanced models of ecology and biogeochemistry for understanding and predicting the future states of the oceans. The consortium for Estimating the Circulation and Climate of the Ocean (ECCO) was formed under the National Ocean Partnership Program (NOPP) to elevate ocean state estimation from experimental status to that of an operational tool for studying large-scale ocean dynamics, designing observational strategies, and examining

the ocean's role in climate variability. ECCO is currently being run at 1/6th degree resolution, and can simulate 1.5 years per day on 512 processors of an Altix.

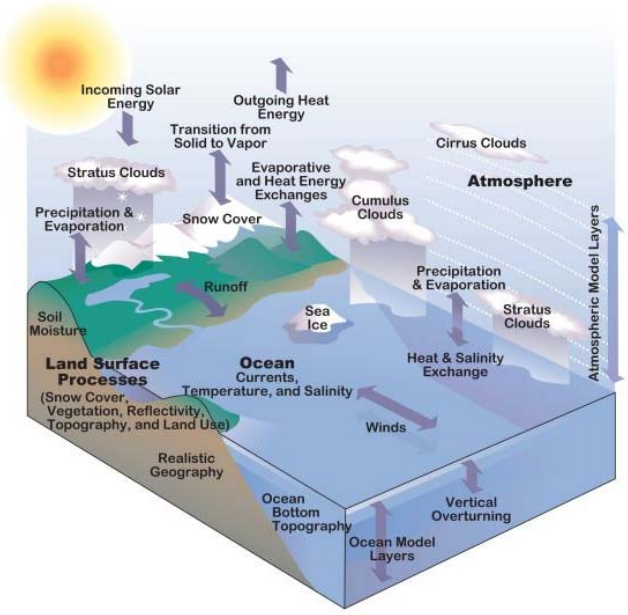


Figure 4.4. Schematic demonstrating the complexity of CCSM
(Courtesy NCAR web site, http://www.ncar.ucar.edu)

In Earth sciences, data assimilation is almost as important as developing high-resolution models. Data assimilation is the process of finding the model representation that is most consistent with the observations. NASA currently has 20 Earth-observing research satellites orbiting our planet, distributing more than 3 TB of data each day. Usually, data assimilation proceeds sequentially in time where the model organizes and propagates the information from previous observations. New data is used to modify the model state to be as consistent as possible with them as well as with past information. At a single synoptic time, the model state typically contains more information than is present in a new batch of data. It is therefore important to preserve this past knowledge while adjusting the model state to fit the new data. New research is investigating non-sequential 4D data assimilation methods.

CFD FOR AEROSPACE

Most of the Japanese research and development in computational fluid dynamics (CFD) for aerospace applications is conducted by the Japan Aerospace Exploration Agency (JAXA), under the jurisdiction of the Ministry of Education, Culture, Sports, Science, and Technology (MEXT). In many ways, JAXA is similar to NASA, albeit at a much smaller scale and with a relatively narrower scope. JAXA's Institute of Space Technology (ISTA), located at the Aerospace Research Center, performs the bulk of the leading edge R&D work. The former National Aerospace Laboratory (NAL), responsible for next-generation aviation (including aircraft, rockets, and other air and space transportation systems) and relevant enabling technologies, forms the core of ISTA. The Information Technology Center at ISTA is developing CFD techniques and software that contribute to and advance the research projects in JAXA and various Japanese industries. It also houses the second-largest supercomputer (after the ES) in all of Japan. This machine, dubbed the Numerical Simulator III (NSIII), is a 2304-processor Fujitsu PrimePower HPC2500 system, based on the 1.3 GHz SPARC64 V architecture.

Almost all of the research conducted at ISTA can be grouped into five main categories: crafting new aircraft design, designing next-generation spacecraft, enhancing aviation safety, preserving the environment, and supporting aerospace technologies. New aircraft research includes work on the next-generation supersonic

transport (SST) and the stratospheric platform airship system (SPF). ISTA is accumulating existing information and establishing new technologies for reusable space transportation systems that were verified by three flight tests: orbital reentry experiment (OREX), hypersonic flight experiment (HYFLEX), and automatic landing flight experiment (ALFLEX). To increase aviation safety, ISTA is conducting research on the safety of flight operations (such as performance, navigation, guidance and control, human factors, and weather) and on aircraft structures (such as weight reduction, crash impact, and static/fatigue loads). Preserving the environment includes both Earth and space. On Earth, ISTA is performing research to reduce jet engine noise and exhaust gas emission, and the development of technology for high-precision air-quality measurements. In space, the goal is to investigate the debris problem caused by past missions, and to prevent collisions with satellites in orbit, the International Space Station, and future spacecraft. Finally, to support aerospace technologies, ISTA conducts fundamental research on computational fluid dynamics, aircraft and spacecraft engines, structures and materials, artificial satellites, and space utilization.

The NAL, now a part of JAXA, has a long and illustrious history of numerical simulation technology research and development. They installed one of the first digital computers (Burroughs DATATRON, 1960), the first transistor computer (Hitachi 5020, 1967), and the first supercomputer (Fujitsu FACOM 230-75AP, 1977) in Japan. The last success started a close collaboration between NAL and Fujitsu that has lasted until today. In 1987, FACOM VP400, the most advanced version of Fujitsu's VP line at the time, began to run 3D Navier-Stokes CFD applications at NAL. It employed a pipeline architecture with peak vector performance of 1140 Mflop/s, and was able to complete full-configuration simulations around complete aerospace vehicles in fewer than 10 hours. Research and development of the Numerical Wind Tunnel (NWT) began as a national project in 1988 with the goal of having a 100x performance improvement over the VP400. When it was introduced in 1993, it had a peak performance of 236 Gflop/s and occupied the top spot on the November 1993 Top500 list. It was a distributed-memory vector-parallel supercomputer, and consisted of 140 VPP500 processors. In 1996, NWT was enhanced with 166 nodes and the addition of visual computer systems, and became known as Numerical Simulator II (peak of 280 Gflop/s).

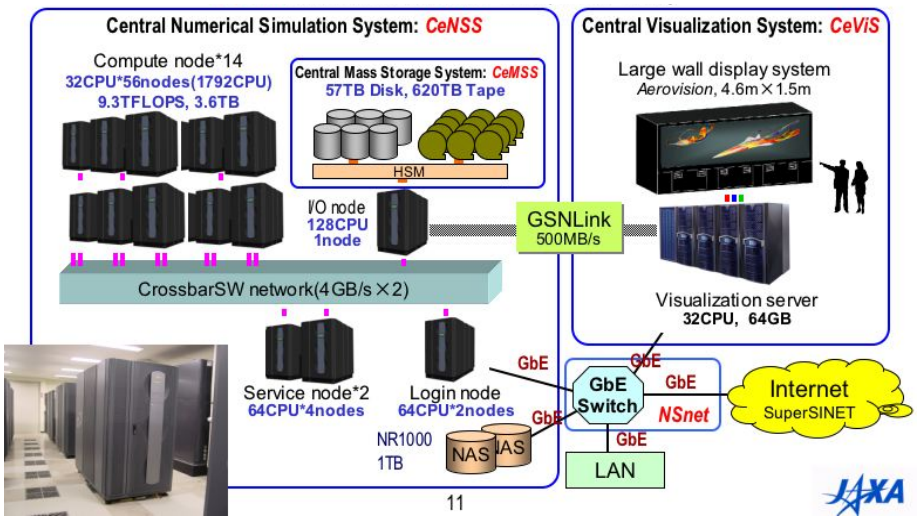


Figure 4.5. Numerical Simulator III (Courtesy Yuichi Matsuo, JAXA)

NSIII, JAXA's latest supercomputer, was introduced in 2002. This Fujitsu machine uses commodity cluster technology rather than traditional vector architecture—depicted above in Figure 4.5. The building block is the PrimePower HPC2500, with 8 CPUs (5.2 Gflop/s peak) per board and 16 boards per cabinet, connected via a 133 GB/s crossbar. The machine runs the standard 64-bit Solaris 8 operating system. The NSIII consists of the Central Numerical Simulation System (CeNSS) made up of 14 PrimePower HPC2500 cabinets, for a total of 1792 processors and a peak performance of 9.3 Tflop/s. The inter-node connection runs via a 4 GB/s bi-directional optical crossbar, where a hardware barrier provides synchronization and each node has its own data transfer unit (DTU). CeNSS has 3.6 TB of main memory, and is connected to the Central Mass Storage System (CeMSS) via a single PrimePower cabinet serving as an I/O node. CeMSS has a total capacity of 57 TB FC RAID disk space and 620 TB LTO tape library. CeNSS is connected to the Central Visualization

System (CeViS) via the I/O node and a 500 MB/s Gigabyte System Network (GSN) link. CeViS consists of a 32-processor, 64 GB SGI Onyx3400 visualization server, a 4.6m×1.5m (3320×1024 pixels) wall display, and several graphics terminals. Three additional cabinets are each partitioned into two 64-processor nodes: four of these are used as service nodes (compilation, debugging, etc.) while the remaining two are login nodes connected to the Internet. The NSIII thus has a total of 18 PrimePower cabinets (2304 processors), and is currently the only Fujitsu computer with LINPACK performance greater than 1 Tflop/s.

A hybrid programming model is used on CeNSS. Within a node, one can use either Fujitsu autoparallel or OpenMP directives (thread parallel). Across nodes, the message-passing paradigm is XPFortran or MPI (process parallel). JAXA stayed with XPFortran, instead of migrating to HPF, which has been very successful on the ES, because they already had it installed and working on the NWT since 1993. Fujitsu provided the necessary compilers, and all code transformations from NWT to CeNSS were straightforward.

Currently, CFD at JAXA is focused on engineering. Most of the work is in design confirmation, performance evaluation, and optimization of aerospace vehicles. The production codes solve Navier-Stokes equations using finite-difference/finite-volume methods on block-structured grids that are then parallelized using domain decomposition techniques. These are similar to NASA's production CFD code, called OVERFLOW. In the science arena, JAXA researchers are investigating fundamental flow physics and accumulating a giant database for basic fluid flows. However, running single CFD applications is no longer considered challenging. Hence, JAXA scientists have moved into multidisciplinary analysis such as CFD/heat/structure, CFD/acoustics, and CFD/control. We were given a detailed parallel performance presentation for four sample applications: flow-over full aircraft, combustion flow, turbulent channel flow, and helicopter rotor flow. Representative pictures for these problems are shown in Figure 4.6.

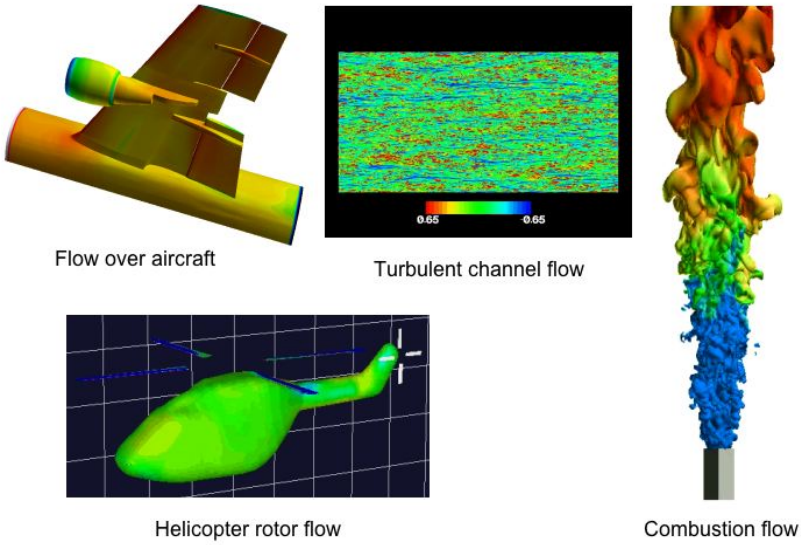


Figure 4.6. Sample CFD applications at JAXA (Courtesy Yuichi Matsuo, JAXA)

The aircraft application uses large eddy simulation (LES) and the finite-difference method to capture flow through various aircraft components. The code has moderate memory access and light communication requirements, and was originally process-parallelized via MPI. Hybrid performance on a 21M-gridpoint problem showed linear scalability to 512 processors and an efficiency of 89%.

The second application used direct numerical simulation (DNS) and the finite-volume method to solve combustion flow out of a nozzle. The code has light memory access and communication requirements, and was also originally parallelized with MPI. Again, the hybrid version on a 7M-gridpoint problem got linear performance to 512 processors and an efficiency of 85%.

The third application was to solve turbulent channel flow using DNS, FFT, and the finite-difference method. The code exhibited moderate memory access but heavy communication, and was originally parallelized using

XPFortran. The hybrid implementation demonstrated linear performance to 224 CPUs and a parallel efficiency of 68% on a 1400M-gridpoint problem, but deteriorated drastically for larger processor counts.

Finally, the fourth application simulated helicopter rotor flow using unsteady Reynolds-averaged Navier-Stokes (URANS) and the finite-difference method. The code featured indirect addressing due to interpolation, had heavy memory access and communication, and was also originally parallelized in XPFortran. The performance of the hybrid version of this application was quite poor. For a 1.5M-grid point problem, speedup was never more than 40x, even when using 512 processors. In fact, a pure XPFortran implementation performed well, showing an efficiency of 83% on 32 processors. A new RANS code, called UPACS, has recently been developed in F90, and is undergoing extensive testing and optimization.

To place the JAXA CFD effort in perspective, Figure 4.7 shows pictures of sample comparative work being currently performed at NASA. For example, highly accurate launch and ascent simulations are increasingly important as space missions and systems become more complex, expensive, and risky. Integrated high-fidelity modeling will enable simulations of failures and associated vehicle responses, while rapid turnaround will reduce design cycle times and enable rapid exploration of multiple configurations and flight regimes. Recent simulations of shuttle Columbia (STS-107) debris trajectories demonstrate NASA's current capability in solving Navier-Stokes equations with a variety of turbulence models and six-degrees-of-freedom multiple-flight bodies. A typical OVERFLOW simulation for the first second of flight requires more than 100,000 CPU hours on a 600 MHz Origin3000 machine. Modeling a complete launch environment for 30 seconds with local meso-scale weather and a three-day turnaround would require a sustained capability of more than 1 Pflop/s.

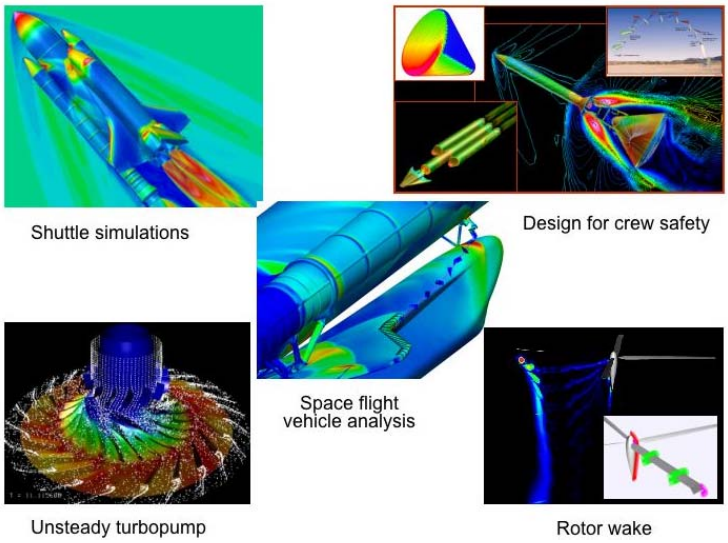


Figure 4.7. Comparative CFD work at NASA (Courtesy Stuart Rogers, Mary Livingston, Michael Aftosmis, Cetin Kiris, Roger Strawn, NASA - Ames Research Center (ARC))

Modeling unsteady flow through turbopumps is also a critical component toward providing high-fidelity design and analysis of fuel/oxidizer supply subsystem for liquid rocket propulsion. Transient phenomena at start-up and non-uniform flows that impact vibration and structural integrity are particularly interesting. Turbopump impeller/diffuser simulations for the Space Shuttle Main Engine (SSME) are routinely conducted using the INS3D code. A 34M-gridpoint problem consisting of 114 overlapping blocks requires more than 72 hours of wall-clock time on 128 processors of an Origin3000 for three impeller rotations. Simulating a six-stage turbopump for 10 revolutions with inflow/outflow piping and a three-day turnaround would easily require a 50x increase in computational power.

As expected, significant other work in CFD is underway throughout Japan at various academic and research institutions. At the Tokyo Institute of Technology, we heard a sample of their research activities. Most of the current focus is on developing highly accurate numerical schemes such as an interpolated differential

operator (IDO) and Hermite interpolation to preserve high order. The targeted applications include dendrite solidification, turbulent cavity flow, TNT explosion, blood flow, and a falling leaf. A selection of these is shown in Figure 4.8. For example, simulating the chaotic motion of falling leaves is an extremely difficult problem because of tightly coupled fluid-structure interaction, the complex shape of leaves, and their very thin structure.

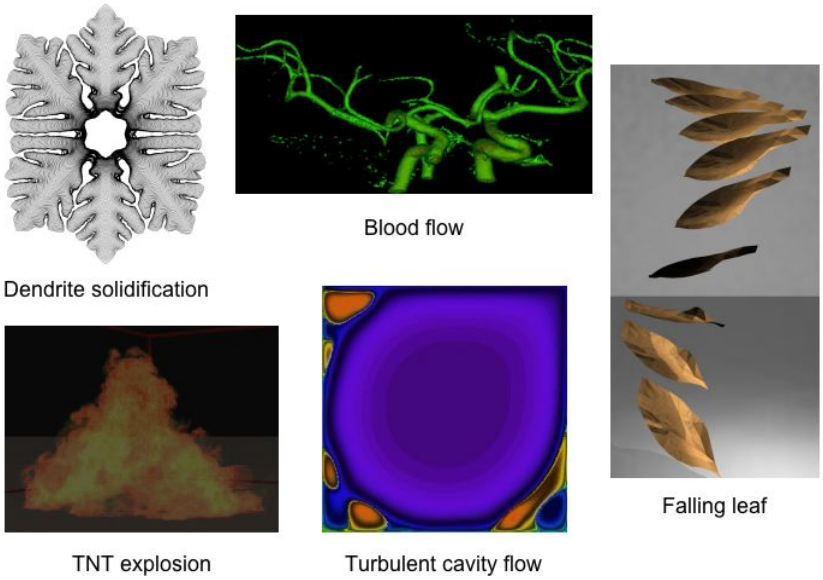


Figure 4.8. Sample CFD applications at Tokyo Institute of Technology
(Courtesy Takayuki Aoki, Tokyo Institute of Technology)

Unlike JAXA, all these simulations at the Tokyo Institute of Technology are performed on distributed grid resources at this time. These applications do not currently have huge computational requirements because the goal is to develop advanced numerical schemes. However, it is easy to envision that some of these problems (accurate human blood flow including major organs and chaotic phenomena such as falling leaves) could easily consume large amounts of computing cycles. As is described in Chapter 8 of this report, Japan is investing heavily in data and computing grids at various levels in order to link the nation's intellectual resources in a collaborative environment as well as provide seamless access to distributed data archives and computer systems.

COMPUTATIONAL NANOTECHNOLOGY

Even though the ES is still primarily devoted to solving problems in Earth sciences, the focus has lately shifted somewhat to epoch-making simulations in computational nanotechnology. Advances in nanotechnology will have a tremendous impact on almost every facet of life. Of all the sites that we visited, RIST appears to be doing the bulk of the research work in collaboration with theorists and experimentalists from other organizations both within and outside the country. Since its mission is to develop and utilize computational science and engineering technology for application domains important to Japan, RIST is now focused on nanotechnology given the recent national interest in this area. To bring in a wide spectrum of expertise, RIST has organized multi-disciplinary research groups, such as carbon nanotube simulation and high-temperature superconducting nanodevices, with industrial and academic partners. At present, RIST uses about 15% of the ES's computational power to explore simulation-driven approaches for new nanomaterial design (nanotubes, fullerenes, diamonds), self-assembly processes (fabrication, synthesis, self-healing, atomic welding), properties of nanomaterials (thermal, mechanical, electronic, magnetic), and nano-electro-mechanical systems. Unfortunately, the level of effort does not appear to be commensurate with the broad portfolio of research areas that are being targeted. The ES is ideally suited for these extremely compute-intensive applications, but even more powerful supercomputers with huge memory resources will be required to continue making scientific breakthroughs.

At RIST, we were given a brief overview of their nanotechnology research in several areas; representative pictures are shown in Figure 4.9. Japanese scientists are designing innovative nanomaterials with desirable properties, e.g. "Jungle Gym" (a three-dimensional network of nanotubes) with super hardness, and a diamond with high-temperature stability. These novel materials could potentially be used for many different applications such as sensor arrays, neural networks, and logic trees. Researchers are also simulating the evolution of nanocarbon structure by isomerization and classification with the goal of finding new structures. To discover fundamental properties of nanoscale matter through simulations, RIST scientists have computed thermal conductivities for carbon nanotubes that are 200 nm long and that contain almost 40,000 atoms. This particular simulation using a tightbinding molecular dynamics code ran for 1.4 hours at 7.1 Tflop/s on 3480 processors of the ES (25% of peak), demonstrating almost linear speedup. These are extremely impressive results and represent state-of-the-art simulations in computational nanotechnology. On the other hand, mechanical properties have been computed for nanotubes that are only 20 nm long. Finally, various nanodevices are being simulated and/or planned, such as becoming effective sources of continuous terahertz waves, nanoreactors encapsulating obstacles, and nanorobots self-assembling.

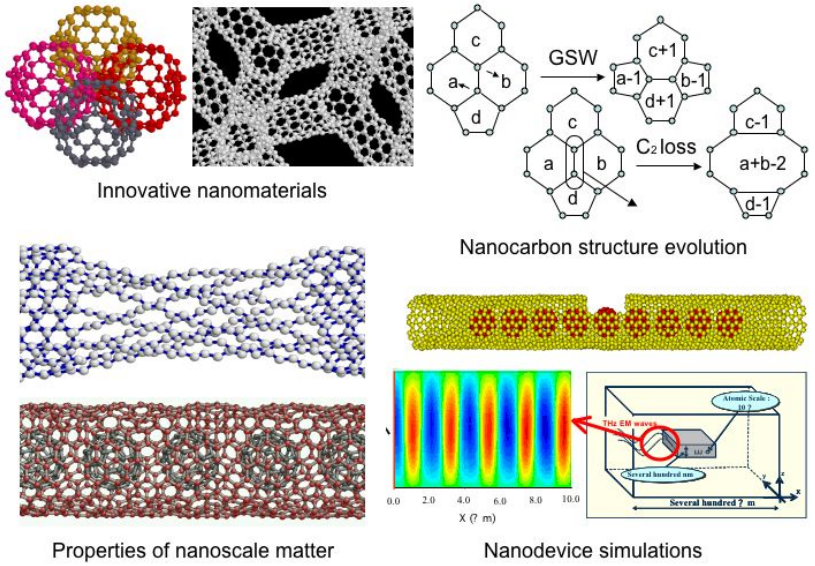


Figure 4.9. Sample nanotechnology simulations at RIST (Courtesy Hisashi Nakamura, RIST)

As expected, there is also significant activity in computational nanotechnology in the United States. Figure 4.10 shows visualization of sample comparative work. For instance, the foundation for Jungle Gym, diamond, and many other novel nanostructures is the stable y-junction that was first demonstrated computationally at NASA and was later shown to be feasible via experiments. NASA specifically has a long-term interest in several nanotechnology areas: nanomaterials for robust aeroshells and thermal protection systems; nanosensors for space vehicle health monitoring; nanomechanics for smart materials for self-repair and vehicle flight surface control; nanoelectrochemical systems for fuel and power; and nanoelectronics for more intelligent and autonomous spacecraft. U.S. researchers are also investigating the thermal properties of the peapod structure (nanotube plus fullerenes) that is similar to the RIST work on nanoreactors encapsulating obstacles. Other research areas include atomic chain electronics that may ultimately lead to the smallest switches possible; quantum electronic devices made from nanotubes; nanoscale optoelectronics lasers and detectors; and molecular sensors and machines.

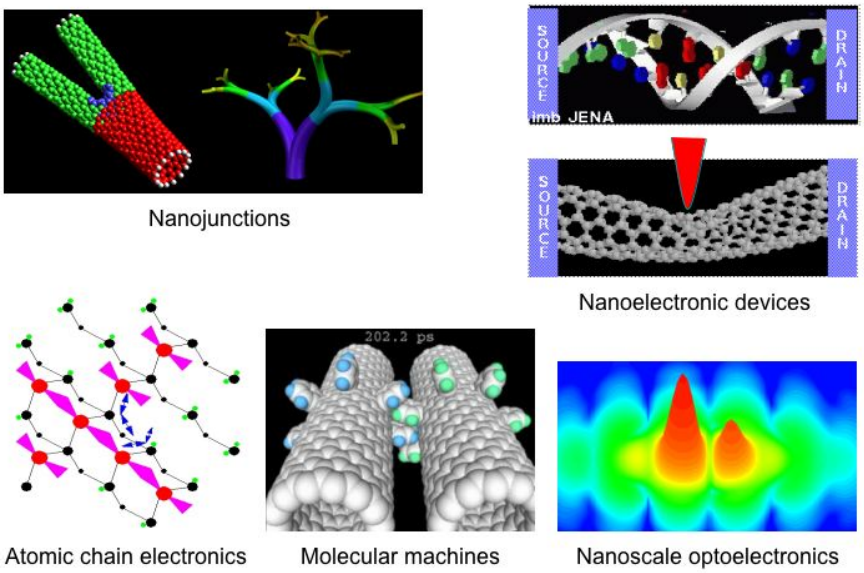


Figure 4.10. Sample computational nanotechnology simulations at NASA
(Courtesy Deepak Srivastava, Toshishige Yamada, Cun-Zheng Ning, NASA ARC)

CONCLUSIONS

There is no doubt that the quality of Japanese research and development in many scientific disciplines is competitive with the world's best. This chapter provided a selective view of their work in three application domains: Earth sciences, computational fluid dynamics for aerospace, and computational nanotechnology. We also saw ample evidence of the existence of coordinated and focused synergistic effort among government research laboratories, academic institutions, and computer companies. However, continued progress in science and engineering via large-scale high-fidelity modeling and simulation will require a significant increase in computational power beyond that currently provided by the Earth Simulator. Nevertheless, the ES has had a major impact in the Earth sciences in that it led to significant advances in the field (e.g. increased resolution, shortened turnaround time). At the same time, there is some resentment in the Japanese HEC community that it diverted funds from other HEC activities such as high-fidelity modeling, efficient numerical algorithms, and computer science research. There does not seem to be a broad consensus for a follow-on project of similar magnitude in the near term even though it is generally acknowledged that a considerable gap exists between application requirements and available HEC capability resources. The Japanese Government has its own view of what the future of HEC should be, while distributed heterogeneous grid environments promise to at least satisfy HEC capacity requirements. These aspects are discussed further in other chapters.

REFERENCES

Community Climate System Model. <http://www.ccsm.ucar.edu/> Last accessed February 25, 2005.

GeoFEM, RIST. <http://geofem.tokyo.rist.or.jp/> Last accessed February 25, 2005.

Matsuno, T. 2004. Development of Climate Models to Be Run on the Earth Simulator. <http://www.ccs.tsukuba.ac.jp/workshop/ccs-sympo04/pdf/matsuno.pdf> Last accessed February 25, 2005

Shingu, S., H. Takahara, H. Fuchigami, M. Yamada, Y. Tsuda, W. Ohfuchi, Y. Sasaki, K. Kobayashi, T. Hagiwara, S. Habata, M. Yokokawa, H. Itoh, and K. Otsuka. 2002. A 26.58 Tflops/s global atmospheric simulation with the spectral transform method on the Earth Simulator. Proceedings of SC2002, Baltimore, MD, ACM/IEEE Press, New York.

CHAPTER 5

SCIENTIFIC APPLICATIONS OF HIGH-END COMPUTING IN JAPAN II

Peter Paul

INTRODUCTION

The high-end computer centers of Japan that the team visited were all located in laboratories for which large-scale computation was mission-critical. It has been Government policy in the past to update such mission-critical computers every five to six years with more advanced capabilities, and all visited laboratories had at least a capacity at the few Tflop/s level. In several cases the Earth Simulator (ES) was used in applications outside of its core climate simulations. In this chapter we report, somewhat selectively, on examples in the use of HEC in areas where we found that the availability of HE computers had led already, or promise to lead, to significant progress. These areas were the simulation of fusion plasmas, both in a Tokamak and Stellerator, for the exploration of advanced reactor concepts, for solving the Lagrangian of quantum chromodynamics (QCD) by use of lattice gauge calculations, and for specific materials science and chemical applications. Although we learned only peripherally about HEC applications for large-scale molecular dynamics, cell functioning and the calculation of protein structures, the very significant developments that are underway in Japan using the Grape chip as an accelerator in conjunction with PC clusters are included in this chapter.

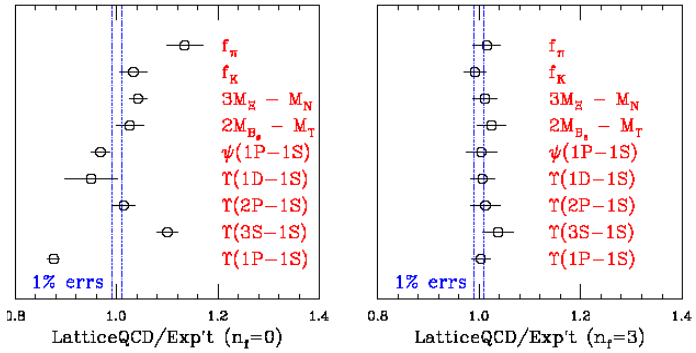
While the appearance of the ES and other advanced computers (such as the SX-7 and advanced clusters) have enabled Japan to be a leader in HEC and simulations, it is also apparent that Japan has had, over the past decade and into the future, a broad program supporting in universities the exploration of strategic algorithms and simulation challenges that will match and exploit the capabilities of the advanced computers. We include an outline of these projects in this chapter. It was clear at almost every site that Japan is investing heavily into data and computing grids at various levels. This endeavor, which involves the intellectual resources of national laboratories and universities, will have a large influence on the future of scientific computing in Japan. This aspect is discussed in another Chapter 8 of this report.

LATTICE GAUGE CALCULATIONS

Such calculations solve, in principle exactly, the fundamental Lagrangian of the strong interaction, QCD, on a space-time lattice. Quarks sit on the lattice points and interact by the exchange of gluons with their nearest neighbors. One obtains the true solution when the lattice constant is extrapolated to zero. These non-perturbative calculations address important problems in high-energy and nuclear physics. Currently the high-energy physics LGC effort is concentrated in Japan at High Energy Accelerator Research Organization (KEK) and Tsukuba University, which are geographical neighbors. They use about 80% of the total CPU time of their Hitachi SR8000-100/F1 (12 GF/s per node, 100 nodes, 1.2 Tflop/s peak) cluster computer that was installed in 2000 and may be upgraded or replaced in 2006. While this represents a sizable computational capability, it falls far short of what is needed to address questions that are raised by the experiments. The short range of the strong interaction makes it possible and profitable to design specific, very cost-effective and efficient computer architectures. In the near future the nuclear theory group at the Wako branch of the Institute of Physical and

Chemical Research (RIKEN) will enter the field with high-temperature QCD calculations using the new RIKEN-BNL machine (funded by Japan) based on the QCD on a chip (QCDOC) ASIC developed by Columbia University for that purpose. This machine will have a capability of 10 Tflop/s (peak) at a cost of $5 million and will be available early in 2005. A copy of this machine, also 10 Tflop/s, has been funded by the DOE for the U.S. LG Consortium.

The big advantage offered by a more powerful computer is the inclusion of vacuum excitations, i.e. of the so-called sea quarks, in the calculation. This can have a large benefit, as in the calculation of masses of baryons and elementary calculations, which are now obtained at the 1% level (see Figure 5.1). In nuclear physics high-temperature QCD will be crucial in understanding the evolution of QCD matter (see Fig. 5.2). The computational needs of LGC increase with the 7th power of the number of grid points. Thus even a computer with tens of Tflop/s does not allow an extension of the number of the mesh points. However one can do the previous calculations now with superior precision comparable to that of experimental data. This means that a supercomputer such as the ES does not necessarily provide a qualitative step forward in comparison with much cheaper designs. It will require computers of hundreds of Tflop/s to bring about a qualitative change.



Lattice QCD results divided by experimental results without (left panel) and with (right panel) quark vacuum polarization. These quantities were chosen because one expected them to have small systematic errors. Obtaining similar accuracy for the quantities relevant to the determination of $\bar{\rho}$ and $\bar{\eta}$ will be more challenging.

Figure 5.1. Lattice Gauge calculations of elementary masses and coupling constants demonstrating the improved accuracy obtained by including vacuum polarization. (Courtesy U.S. Lattice Gauge Consortium)

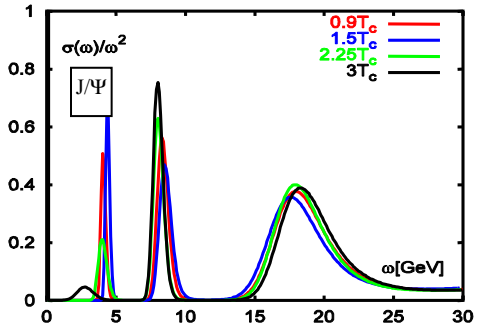


Figure 5.2. LG calculations of heavy quarkonia masses (starting with the J/ψ) as a function of Quark matter temperature. These predictions are crucial guidance for experimental search for the quark-gluon plasma and cannot be obtained in any other way. (Courtesy Peter Petreczky, BNL)

The Tsukuba and the RIKEN-BNL computers achieve an efficiency of 30% to 50%, depending on the number of mesh points. First runs on the ES achieved 25% with plans for improvement. Since QCD involves only the interaction of each mesh point with its nearest neighbor, ingenious tricks that reduce latency can bring big rewards. In the case of the Tsukuba machine the lattice parameters are distributed carefully over the individual nodes, which are arranged in a three-dimensional torus. Each node is an eight-fold symmetric multiprocessing (SMP), and communicates with adjacent nodes through message passing interface (MPI).

The goal of increasing efficiency and reducing latency is pursued even further with the QCDOC architecture of the RIKEN-BNL machine. The QCDOC chip uses a single commercial IBM Power PC chip operating at 500 MHz with a 4MB on-chip memory per node. These are arranged in a six-dimensional torus with nearest neighbor connections. (See Fig. 5.3) It uses Lattice QCD message passing (QMP) software between nodes, which can be optimized for different applications. The architecture has no switch network and in this sense is an ultra-scalable design. It is a highly parallel arrangement because each processor has 24 communication channels to its neighbors. Three computers of this type are currently in production: the first one for RIKEN-BNL, the second for the U.K. Lattice Gauge consortium, and the third for the U.S. LG consortium. The performance of this architecture for LG calculation is expected to be near 50%. However, it turns out the same architecture promises to be also very efficient (>30%) and scalable for other basic calculations, such as molecular dynamics, even though it involves 1/r (i.e. long range) potentials. QCDOC connects to a very successful design with big plans in Japan, the Grape chip (more later). It has also provided the conceptual basis for the development of the Blue/Gene series of ultra-scalable computers at IBM (which have now eclipsed the ES in speed).

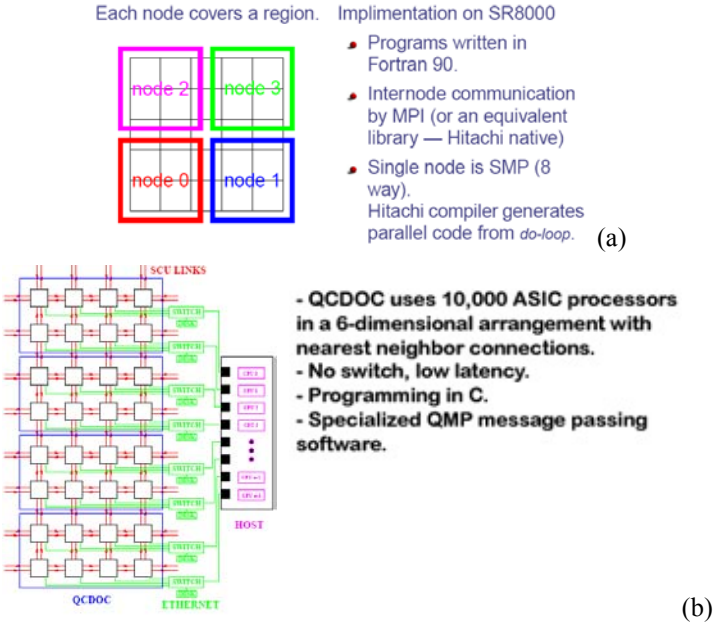


Figure 5.3. Node arrangements for LG Calculations for the SR8000 cluster (a) and for QCDOC (b). (Courtesy (a) Shoji Hashimoto, KEK, and (b) N. Christ, Columbia University)

The lattice gauge community of Japan is quite large, with about 45 senior practitioners and about 120 users overall (the U.S. LG Consortium has about 65 senior members) and both groups work very closely together. For example, while Riken has funded QCDOC, many Japanese scientists clearly prefer a PC cluster design using advanced switches, and a similar discussion is taking place in the U.S. The next few years will show which approach is more efficient.

THE PROTEIN EXPLORER

The RIKEN Laboratory has built up a very strong capacity in genomic and protein/cell research through its Genomic Science Center. Computational modeling efforts there are based in the Bioinformatics group. Among several groups that focus on modeling of cell functioning and biological databases, it includes the High-Performance Biocomputing Research Team. RIKEN has followed an imaginative and aggressive course toward truly next-generation computational capabilities based on novel concepts. The most exciting project is the Protein Explorer (PE). Our group did not specifically visit the base site for this effort at RIKEN Yokohama and did not hear about specific computational results. Nevertheless, this important development was presented at other sites and is, from its scientific impact, most notable.

This computer development, which aims at a petaflop/s machine (i.e. outclassing the ES by a factor of 25 for molecular dynamics applications) by the end of 2006, is the most ambitious development in the application of molecular dynamics for protein structure calculations. The development uses the MD-Grape-6 LSDI chip. The Grape (Gravity Pipe) chip solves the 1/r potential problem very efficiently, but is not programmable. The Grape chip has evolved in a line for high-precision machines (even numbers) and one for low-precision machines (odd numbers). The PE is then next step in the evolution from MD-Grape and MDM (Molecular Dynamics Machine). MDM at RIKEN demonstrated a 78-Tflop/s-peak performance in 2001. Grape-6 developed at Tokyo University achieved 64 Tflop/s in 2002. The potential petaflop/s performance of the PE derives from its hybrid design: A Grape accelerator board is attached to a commercial cluster. The host PC performs the bond calculations, and then transfers the coordinates of particles to the special purpose chip, which solves the particle motion under Coulomb and Van der Waals forces. This division of effort is particularly effective when the communication requirements (which scale like N, the number of particles) between host and accelerator are small compared to the computational requirements (which scale like N^2). The PE uses the MD-Grape-3 chip, which has a speed of 165 Gflop/s (operating at ~250 MHz). Six-thousand one-hundred forty-four chips achieve the nominal goal of 1 petaflop/s. The host is a 256-node cluster. The full machine will be available in 2006 at an estimated cost of $20 million. Figure 5.2 sketches the architecture of a hybrid Cluster-Grape system. We can note a parallel development in Japan of the MDM, which promises to operate at 100 Tflop/s, and may become commercially available. It remains to be seen how much of a disadvantage will be associated with the fact that the Grape chip is not programmable and thus limited to only one potential form. In May 2004 a follow-on project Grape-DR was announced at Tokyo University, which targets a goal of 2 Pflop/s in 2008, two years after the projected start of the Protein Explorer (PE). Unlike the PE, the new Grape-DR architecture aims to integrate two million processors and will be designed for multiple applications. Thus it bears some similarity to the BlueGene design philosophy, but with much higher effective throughput.

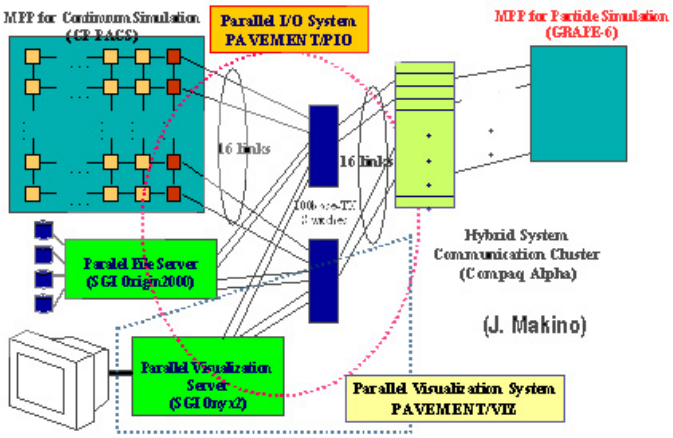


Figure 5.4. Configuration of a hybrid computer consisting of an Alpha cluster and Grape-6 accelerator boards. This hybrid architecture developed at RIKEN–Yokohama is expected to reach ~1 Tflop/s by 2006 for calculations of protein structures. (Courtesy Mako Taiji et al, Riken-Yokohama)

The PE is certainly the most determined and realistic effort to reach the huge performance levels that will be needed for the realistic modeling of reasonably large biological molecules. A new simulation algorithm (REMUCA) has been developed concurrently and first calculations have been done (C-Peptide of Ribonuclease A in water). At RIKEN's main site in Wako the plan is to soon operate the RIKEN Super Combined Cluster that will combine a 1024-node, 2048CPU PC (Intel Pentium Xenon at 3 GHz) cluster system with 20 MD boards that will operate at 1.2 Tflop/s. This compares favorably with the U.S. where successful *ab initio* predictive calculations of the folding of the smallest folded proteins have been done on clusters. Estimates for MD calculations performed on the 10 Tflop/s QCD machine indicate that high efficiency and good, predictable scaling can be expected. The IBM Blue Gene series, which is based on the QCDOC concept, is predicted to reach 184 Tflop/s (peak) by the end of this year, and both the QCDOC and the Blue/Gene L ASICs are programmable for various force fields. Clearly several options will be available within then next few years to reach well past the 100-Tflop/s levels into a new domain of simulation of protein folding and the dynamics of biological molecules and cells.

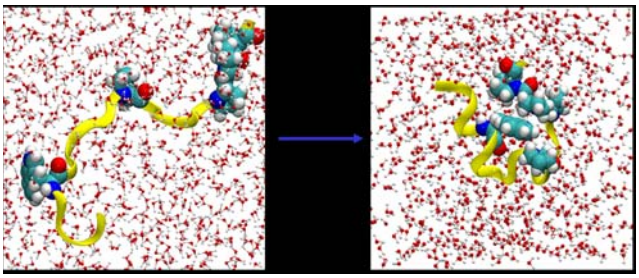


Figure 5.5. *Ab initio* calculation of the folding of the backbone (yellow) of a small protein (TRP Cage) in water. The structure prediction (done with a small cluster) was then confirmed by NMR measurements. The Protein Explorer will allow computing the folding structure of much larger proteins. (Courtesy Carlos Simmerling, Stony Brook University)

PLASMA PHYSICS CALCULATIONS

Plasma physics in Japan is centered in two large laboratories that have complementary missions for plasma fusion: the National Institute for Fusion Science (NIFS) operates and performs research on the Large Helical Device (LHD), a DC magnetic field Stellerator machine. NIFS grew out of a research center at Nagoya University and still has a very large academic connection, hosting hundreds of students. The second laboratory is the Japanese Atomic Energy Research Institute (JAERI), which has a very large mission in nuclear energy in general. At its Tokai site it operates and does research on the JT-60 Tokamak that achieved break-even plasma in 1987. JAERI's Rokkasho site will be the lead laboratory for operation of ITER if that large facility comes to Japan. Even if another continent is chosen as the ITER site, JAERI will lead a large program in plasma science and plasma-wall interaction. The Tokai site will also host J-PARC (Japanese Proton Accelerator Research Complex) (see Fig. 5.6) which will serve high-energy and nuclear physics, material science and the life sciences. These are all areas in need of high-end computational capability.

Figure 5.6. The J-PARC Facility at Tokai currently under construction. Its 3-GeV and 50-GeV high-intensity proton beams will drive a spallation neutron source and intense neutrino and meson sources. It will become operational before the next decade.

Both laboratories have modern vector processor computers: NIFS operates the SX-7/160M5, the latest member of the NEC vector processor line (The ES developed from the SX-5). JAERI uses a VPP-5000 with 64 CPUs at its Tokai site, and at Naka, an Origin 3800 system with 768 CPU's (peak of 0.8 Tflop/s), but has also started substantial use of the ES (the broadest use of the ES that the Panel encountered outside of the ES laboratory itself). Both institutions are main partners in the Fusion Grid Alliance.

Following their historic programmatic evolution, NIFS and JAERI are performing computational programs of a somewhat different character. JAERI is very mission-oriented and practical. For example, out of a total scientific and technical computational demand in 2004 of about 42 Tflop/s, 12 were requested for nuclear fission, 21 for nuclear fusion and 7.3 for photon-plasma interaction. It is projected that this will grow by 2008 to a total demand of 93 Tflop/s, with large increases in the shares for fusions and energy systems calculations. The lead program of high-performance computing in plasma fusion at JAERI is called NEXT (Numerical Experiment of Tokamaks).

This effort aims at numerical modeling of complex plasma phenomena with high accuracy and comparison with precision experiments at JET-60, presumably in preparation for ITER. (See Fig 5.7) Plasma turbulence, MHD processes and boundary effects and diverter simulations will serve to obtain a deeper quantitative understanding of the pulsed plasma of a Tokamak with the aim to simulate higher performance plasma in ITER and to predict, as well as avoid, instabilities, which can be very destructive. A typical simulation of the Jt-60 plasma with 10^8 particles, with accurate space resolution and 10,000 time steps in the plasma evolution, takes about one month on the Origin- 3200 512 CPU machine at Naka. Extrapolating this to the dimensions of ITER (twice the radius and more than twice the height of JET-60) will take several 100 Tflop/s to simulate its plasma within one day! Clearly this requirement is well beyond the capability of the ES. However, the ES has been applied to these problems with about 25% efficiency scaling well up to 4096 Gflop/s (i.e. 10% of the full capability). A non-resistive MHD calculation was already successful in predicting a previously unknown instability with a collapse time of milliseconds, which was then confirmed by experiments. Multi-scale models are needed to describe the interactions between the long-scale MHD and the small-scale motion of ions and electrons.

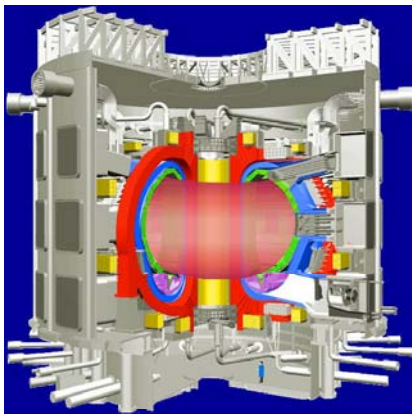


Figure 5.7. Sketch of JET-60. On the Origin 5000 512 CPU (0.5 Tflop/s) machine at JAERI it took one month to simulate the pulsed plasma, with 10^8 particles, space resolution of 160x128x128 and total step number of 10,000. ITER about doubles every dimension of the plasma volume. It is estimated that it will take 100's of Tflop/s to 1 PF/s to simulate the ITER plasma in one day.
(Courtesy Shinji Tokuda, JAERI-Naka)

A most interesting goal is the development of super solvers for eigenvalue problems with the aim to analyze the onset of MHD instabilities in real time as the Tokamak is going through its pulse. The computer could then check the stability of the plasma at any moment and either stop the field increase before a catastrophic collapse occurs or steer the parameters around a foreseen instability. Currently this computation runs with about 6% efficiency on a VPP5000, sufficient for post-analysis. A gain of about 1000 would be required to use the computer simulation in a real-time active loop.

NIFS similarly aims to couple its large experimental program with the LHD to precision simulations that will lead to a deeper understanding of the complex plasma effects that are observed in the twisted magnetic field of the steady-state Stellerator. But the NIFS simulation effort also extends to Tokamak (pulsed) plasmas. Its computational effort has the flavor of a university effort, as NIFS has about 500 graduate students involved, half from Nagoya and half from the laboratory itself. The LHD is a suitable test bed since it has achieved a hot and dense plasma, with >10 keV electron and ion temperatures and a stored energy of over 1 MJ. The SX7 /160M5 is a 1.4 Tflop/s vector processor with a very large (100 TB) common memory. It achieves 32% efficiency for three-dimensional MHD calculations. In addition to its simulation effort it has a significant effort in visualization and virtual reality, which serves to visualize the complicated flow patterns of electrons and ions in the twisted magnetic fields of LHD.

The basis for their current program starts from the observation that linear incompressible simulations of the LHD configuration give unstable results whereas experiments showed perfectly stable plasma. This led to the examination of local stabilization due to pressure flattening, the possibility of compressibility rather than incompressibility in the plasma, or the stabilizing influence of parallel sheet flows. Thus complex plasmas were studied using non-linear MHD and, especially, the development of magnetic structures in the LHD. This led to

the description of magnetic islands in the helical field of LHD. The self-healing of such magnetic islands was discovered and confirmed by experiments, and the physics of island generation became understood. The powerful computer allows linear analysis of *compressible* flow in three dimensions which gives great visual insight (see Fig. 5.8)

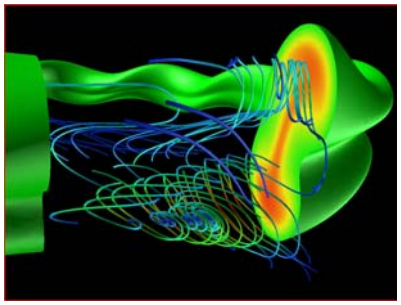


Figure 5.8. 3D nature of compressible flow structures showing spiraling streamlines inclined toward the toroidal direction. (Courtesy Masao Okamoto, NIFS, Toki)

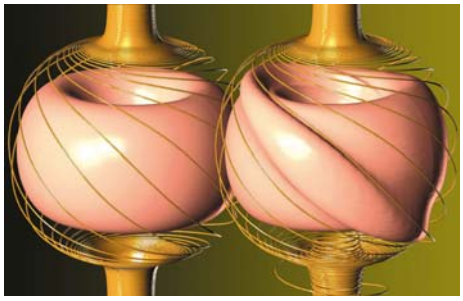


Figure 5.9. Internal reconnection events observed in spherical Tokamak simulated by MHD code. (Courtesy Masao Okamoto, NIFS, Toki)

It is clear from the pictures that one of the biggest advantages of detailed calculation is the visualization of complex plasma and particle flow in three dimensions.

A big advance in the NIFS program that flowed from the availability of a supercomputer is the development of the FORTEC code, which describes "neoclassical transport" and includes finite orbit widths (FOW) effects. The code includes particle orbits and collisions. It thus describes simultaneously the large scale magnetic and electric field, and the particle motions. The FOW describes the generation of radial electric fields, which derive from non-local effects and are very important for plasmas with steep gradient pressures. The electric shear from such radial electric fields may suppress turbulence. The computed radial electric fields are consistent with the simulations. This was then expanded to include the injection of energetic alpha particles into the burning plasma, and even pellet injection. The supercomputer allowed performing a nonlinear simulation in an open system where neutral beam injection, collisions and particle loss due to Alfven eigenmodes are all taken into account.

Based on its successful simulation of magnetic structures in LHD geometry, NIFS also did successful calculations using MHD of so-called internal reconnection events (IRE) in spherical Tokamaks that had been observed earlier. These structures are of a very strange and unexpected shape that would be difficult to visualize without the help of the computer simulation (see Figure 5.9). The computer was able to explain the time scale of strange disruptive configuration changes, identify the trigger mode, and predict the recovery of the initial spherical field configuration.

Finally, three-dimensional electromagnetic particle simulations gave important results on the dynamical evolution of current sheets in the plasma, which have a major role in the stability or instability of the plasma. Multilayer simulations in an open system allowing energy inflow and outflow for a given sheet demonstrated the creation of magnetic islands and kink instabilities. The NIFS high-end computer will make it possible to simulate macro- and micro-scale effects simultaneously and consistently.

Clearly these powerful computers have made the modeling of the complex plasma behavior and ion-electron transport both in spherical DC devices and Tokamaks a predictive science. In fact the scale of a few Tflop/s seem adequate for the study of complex plasmas, but a real-time influence, as is being achieved, e.g. in the operation of large accelerators, is beyond the capability of present high-end computers. Although NIFS had been awaiting more computational power, they considered their machine a production machine and had little interest in using the ES.

CALCULATIONS FOR ADVANCED REACTOR DESIGNS

JAERI is using high-end computers to design next-generation power reactors. An example is the simulation in detail of the two-phase flow through the fuel element bundles of an advanced light water reactor. The aim is to achieve such accuracy in modeling that design by analysis can replace physical tests. This goal of a direct numerical simulation, instead of using empirical equations, has become attainable through use of the ES with its 40-Tflop/s capability. A particular design, for which high-accuracy simulations are very important, is the so-called reduced moderation light water reactor now under design study in Japan. This reactor type retains an energetic part of the neutron spectrum in order to burn away the plutonium inside the reactor that is created in the fission process. Thus this reactor can have a conversion ratio >1 and is essentially proliferation-proof. The reduced moderation requires that cooling slots between fuel elements must be very small, typically of order 1 mm (see Fig. 5.10). Therefore the mesh spacing is very small, ~0.15 mm, and the number of mesh points is correspondingly very large, more than 100 million. The simulation calculates three-dimensional compressible/non compressible flow (for the two components, respectively) using more than 300 CPUs of the ES. The results of such simulation are quite amazing; for instance, the vapor flows downward in the region where the gap spacing between adjacent fuel rods is large. These results were then compared to experimental data.

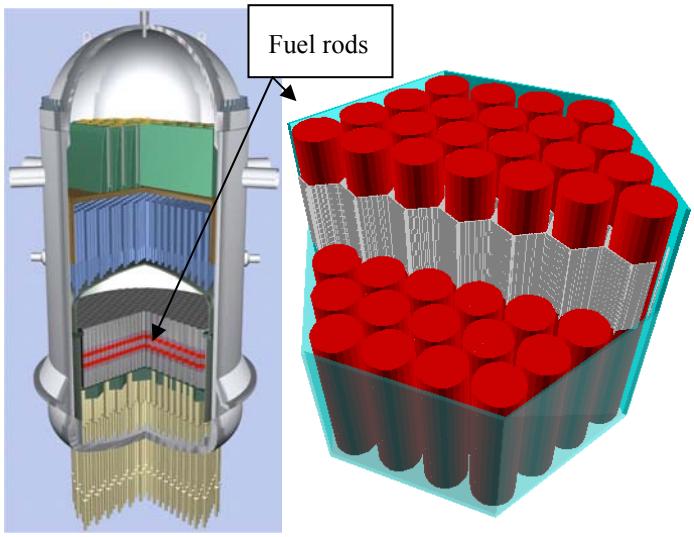


Figure 5.10. Configuration of a Reduced Moderation Light Water Reactor (*left*) which has very narrowly spaced fuel rods and the model (*right*) that was used to calculate two-phase flow with realistic boundary conditions. The tightly packed geometry is evident; number of mesh points was 2.5 billion for calculations on the ES. (Courtesy Kazuyuki Takase, JAERI-Tokai)

This is clearly a major advance in reactor design, and vector processors with their large shared memory are well suited to this application. However, because of the large number of mesh points the storage requirements are enormous. Computing flow through one fuel bundle with 217 rods will require 70 TB for one-column calculation, a full reactor design with 282 fuel bundles would require 20 PB for a full simulation.

HEC APPLICATIONS IN MATERIALS SCIENCE AND CHEMISTRY

In the broad program of HEC applied to technical problems that were presented at JAERI, a number of specific calculations were performed using the ES. These projects have considerable merit and showed the power of the ES but they are in an early, much focused phase. Thus we report on them only briefly.

One project is the *ab initio* design of new molecular ligands for the easier chemical separation of actinide nuclear waste. It involves application of relativistic density functional theory to the calculation of electron structure in actinide complexes and the use of first-principle molecular dynamics to simulate the structure of cations in aqueous solution.

Another project involved the simulation of radiation-induced defects in materials. A new parallel molecular dynamics algorithm has been developed (called the Parallel Molecular Dynamics Stencil) that makes it easy by using MPI to parallelize and expand calculations to a large scale. This code was then applied to the interaction between a dislocation core and defect clusters, and to the role of dislocations in the dynamic facture process.

Finally, the possibility was studied by simulation using the ES of employing superconducting MgB_2 for the detection of neutrons. The detected neutron, i.e. absorbed by the high-cross section boron, will produce a local heat pulse. The simulation solved the time-dependent Ginzburg-Landau equation coupled with heat diffusion and the Maxwell equations to study the time evolution of this heat pulse in the superconducting material. By parallelizing with MPI 50% of peak performance was obtained on the ES. In this process a new parallelization code was developed for the diagonalization of the huge matrix for the ground state of the strongly correlated system. The successfully diagonalized matrix had 18 billion dimensions!

COMPUTATIONAL SCIENCE FOR HEC IN JAPANESE ACADEMIC INSTITUTIONS

As is pointed out in other parts of this report, the development of advanced computers in Japan followed in a logical procession from programmatic devices to commercial computers. The numerical wind tunnel begot the VPP500, cp-pacs brought the SR2201, the ES produced the SX-6, and the Grape, and MDM may also produce a commercial machine. What is perhaps less known is the systematic progression of computational science funding cycles in Japanese academic institutions. We will list here only the most recent and current ones.

The Japanese Society for the Promotion of Science (JSPS) sponsored under its theme “Research for the Future” a program in Computational Science and Engineering (1997 – 2003, total funding $30 million), which brought together computational science and actual computing. The themes and centers (all at universities) were applications-oriented:

- Next-generation massively parallel computers
- Material science simulations for future electronics
- Protein folding from first principles
- Global; scale flow systems
- ADVENTURE Project

JST under its theme “Creative and Strategic Research” (CREST) initiated two new programs, which are still ongoing: The first is “New High Performance Information Processing,” which involves several general software projects. The second program “Innovation in Simulation Technology” funds a number of Grand Challenge-type scientific applications of HEC. The following are representative team projects (each funded for five years at $400 k/year) that were initiated in 2002:

- Multi-physics simulations by the particle method
- Hierarchical biosimulator
- Nanomaterials measurement simulator
- Basic library for large-scale simulations

Some new team projects begun in 2003 were:

- Symbolic numerical hybrid computation
- Materials design
- Simulation for radiotherapy
- Bone medical simulator
- Molecular orbit on the grid
- Heart simulator

These team projects are supplemented by postdoctorate level ($200k/year for three years) and personal ($100k/year for three years) projects. Very interesting projects among these are:

- Flight simulator for birds and insects
- Innovative single-value decomposition methods
- Correlated electron systems
- Hybrid molecular dynamics simulator
- Polymer simulation
- Relativistic molecular theory

Finally, under the Next Strategic IT Program "Development of Strategic Software," which started in 2001, the following projects were funded:

- Quantum chemistry
- Interaction between proteins and medicine
- Nano simulation
- Next-generation fluid dynamics
- Next-generation structural analysis
- Protein functions

These are a large number of strategically selected projects that are geared toward the new generation of HEC. Although each is funded at a relatively modest level (in relation to U.S. research costs) the sum adds up to a significant amount of HEC investment in universities. This thrust appears already to bear fruit in the sense that several universities (e.g. University of Tokyo and Tsukuba University) are creating new departments for computational science by combining faculty that was formerly spread out in conventional science departments. This is a significant step toward make high-end simulation truly the third research leg, complementing experimental and theoretical research.

CONCLUSIONS

Japan has a broadly based and carefully planned, but audacious program in advanced scientific simulations. The strategic attack on protein structure, cell simulations and computational bioscience is especially noteworthy, although results are still at an early stage. The Protein Explorer, with its huge increase in computing power for molecular dynamics, could put Japan into world leadership in this area. Despite the success of the ES and the dramatic achievements in cost per speed of the Grape chips, PC-based superclusters appear to have the widest appeal among the scientists of the national research labs that were visited. There appeared to be widespread reservation among scientist about investing again in a machine of the investment level and operational cost of the ES, even though it was apparent that much more computing power will be needed to attack many of the problems that were presented before the Committee.

The ES and the SX-7 at NIFS are demonstrating how much the simulation of complex processes can profit from computing power in the tens of Tflop/s range. In the next couple of years this level of computing power will become quite commonplace. However, it seems also clear that a paradigm change in simulation of physical, chemical and biological processes will require computing power of at least several hundred Tflop/s. This may come about through a new concerted effort by a government-industry alliance which seeks to produce a general-purpose 1 Pflop/s computer by around 2010. The realization that such a large step in computing capability is feasible is a most significant development.

REFERENCES

KEK, High Energy Accelerator Research Organization. 2004. <http://www.kek.jp/intra-e/index.html> Last accessed February 23, 2005.

CHAPTER 6

ARCHITECTURE OVERVIEW OF JAPANESE HIGH-PERFORMANCE COMPUTERS

Jack Dongarra

INTRODUCTION

This chapter presents an overview of Japanese high-performance computers manufactured by NEC, Fujitsu, and Hitachi. The term "high-performance computing" refers both to traditional supercomputing and to commodity-based systems, which contain components that can be purchased "over the counter." Both types of computer exploit parallel processing for performance. In Japan, as in the United States, traditional supercomputers are under heavy scrutiny, as commodity systems appear to have a better price/performance ratio.

High-performance systems can be divided into three broad categories: commodity processors connected with a commodity interconnect (or "switch"), commodity processors connected with a customized interconnect, and custom processors connected with a custom interconnect. The commodity processor/commodity interconnect system can be characterized as being more loosely coupled than the custom processor/custom interconnect. The latter has a more tightly coupled architecture and hence is more likely to obtain a higher fraction of peak performance on applications.

Table 6.1 illustrates the offerings of the three Japanese vendors in each category.

Table 6.1
Vendor Offerings by System Category

Commodity processor with commodity interconnect	NEC, Fujitsu
Commodity processor with custom interconnect	Fujitsu
Custom processor with custom interconnect	NEC, Hitachi

NEC has basically two offerings: the commodity-based TX7 series and the customized SX line. Fujitsu has two lines: a commodity-based IA-Cluster and a high performance Sparc-based system with a commodity processor and a specialized switch. Hitachi has one offering: the SR11000, based on IBM's proprietary processor and switch.

Because commodity clusters are replacing traditional high-bandwidth systems and shrinking their market, the commercial viability of traditional supercomputing architectures with vector processors and high-bandwidth memory subsystems is problematic. At least one large company in Japan, NEC, continues to be committed to

traditional parallel-vector architectures targeted for high-end scientific computing. NEC, or at least its high-end computing component, believes in the trickle-down effect. One of the strengths of NEC is its continuity in software and hardware, which stretches over 20 years.

When looking at the accumulated performance for high-performance computers in Japan, it becomes apparent that the use of high-performance computers in Japan began to decline around 1998 but picked up again in 2002 with the introduction of the Earth Simulator. Today the use of high-performance computers in Japan appears to again be in decline. See Figures 6.1 and 6.2.

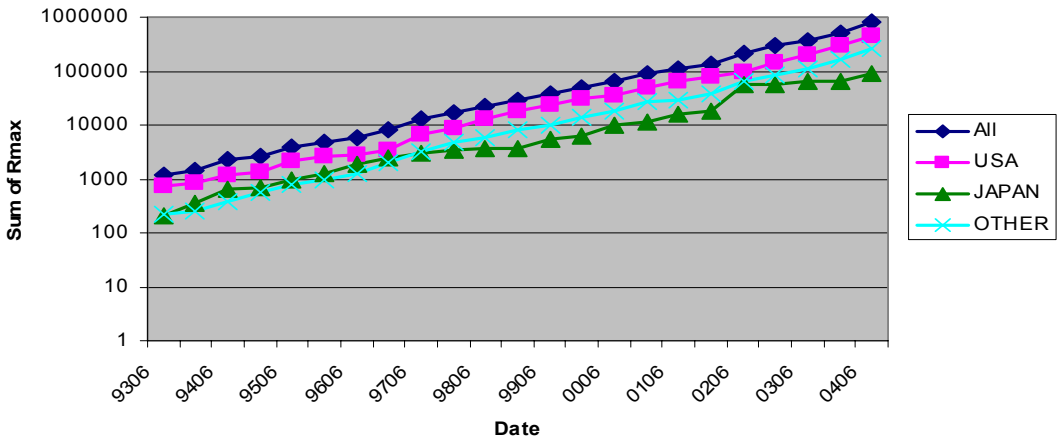


Figure 6.1. Top500 data of accumulated performance for high-performance computers in the U.S., Japan, and other countries over time. (Courtesy Top500.org)

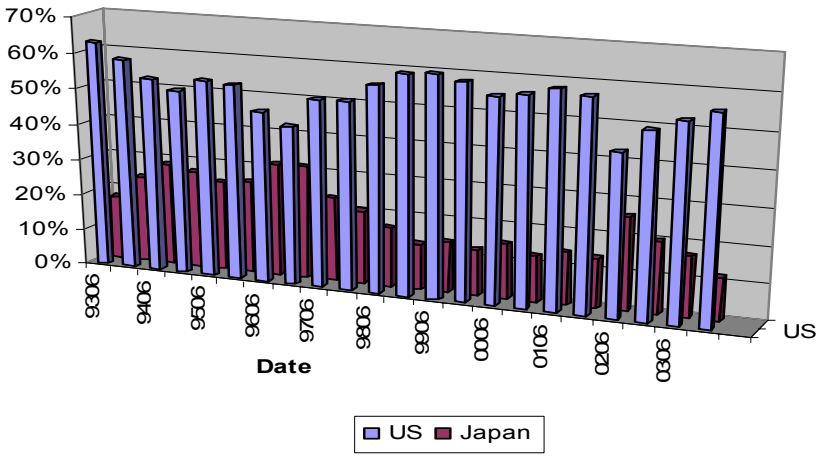


Figure 6.2. Percent of accumulated performance from the Top500 for the high-performance computers in the U.S. and Japan over time. (Courtesy Top500.org)

NEC

Background

NEC Corporation is a leading provider of Internet solutions. Through its three market-focused in-house companies, NEC Solutions, NEC Networks and NEC Electron Devices, NEC Corporation is dedicated to meeting the specialized needs of its customers in the key computer, network and electron device fields. NEC employs approximately 150,000 people worldwide. In fiscal year 2000-2001, the company saw net sales of ¥5,409 billion (approximately $43 billion).

A chart covering the history of NEC supercomputers can be found in the site report for NEC in Appendix B. Relevant product lines include high-end vector supercomputers, scalar servers and IPF servers, PC clusters, and IA workstations. All of these lines incorporate the GFS Global File System. NEC is an original equipment manufacturer (OEM) for Hewlett-Packard Company's Superdome server, and has established a working relationship with Cray Inc. to market the NEC SX system in the United States.

As mentioned earlier, NEC has two main high-performance computing product lines, the SX series and the TX-7 IPF server series. The SX series features specialized high-performance parallel-vector architecture. It is targeted for the high-end scientific computing market. The TX-7 uses commodity architecture. With this line, NEC became the first vendor to support a 16-way SMP based on IA64 Merced.

Remarks for the TX Series

NEC offers the TX-7 series in four models. This report discusses the largest two models. The TX-7 is one of several Itanium-2-based servers that have recently appeared on the market. The largest configuration presently offered is the TX-7/i9510 with 32 1.5 GHz Itanium-2 processors. Because NEC has had prior experience with Itanium servers, offering 16-processor Itanium-1 servers under the name Asuza, the TX-7 systems can be seen as the second generation.

A flat crossbar connects the processors. NEC still sells its TX-7s with the choice of processors offered by Intel, namely 1.3, 1.4, and 1.5 GHz processors with L3 caches of 3 to 6 MB depending on the clock frequency.

Unlike the other vendors employing Itanium-2 processors, NEC offers its own compilers, including an HPF compiler. This compiler is compatible with the software for the NEC SX-6, most likely because it is hardly useful on a shared-memory system like the TX-7. The software also includes MPI and OpenMP. Operating systems offered include Linux and HP-UX. The latter may be useful for migration of HP-developed applications to a TX-7. See Tables 6.2 and 6.3 and the site report for NEC in Appendix B.

Table 6.2
NEC TX-7 Series

Machine type	Shared-memory SMP system
Models	TX-7 i9010, i9510
Operating system	Linux, HP-UX (HP's Unix variant)
Connection structure	Crossbar
Compilers	Fortran 90, HPF, ANSI C, C++
Vendors information Web page	http://www.hpce.nec.com
Year of introduction	2002

Table 6.3
TX-7 System Parameters

Model	i9010	i9510
Clock cycle	1.5 GHz	1.5 GHz
Theoretical. peak performance, per proc. (64 bits)	6 Gflop/s	6 Gflop/s
Maximal performance	96 Gflop/s	192 Gflop/s
Main memory	≤ 64 GB	≤ 128 GB
Number of processors	16	32

Remarks for the SX Series

This report examines the SX-6 and SX-7 models of the SX series. Both models have a vector processor with one chip.

- SX-6: 8 Gflop/s, 16 GB, processor to memory (8*8 * 4 streams = 256GB/s from memory), between nodes 16 GB/s xbar switch
- SX-7: 8.825 Gflop/s, 256 GB, processor to memory 1130 GB/s (8.825*8 * 16 streams = 1130GB/s from memory)

Both the SX-6 and SX-7 use 0.15µm CMOS technology.

NEC SX-6

NEC offers the SX-6 series in numerous models, but most of these are simply smaller frames that house fewer processors. This report focuses exclusively on the models that are fundamentally different from each other. All models are based on the same processor, an eight-way replicated vector processor in which each set of vector pipes contains a logical, mask, add/shift, multiply, and division pipe. As multiplication and addition (but not division) can be chained, the peak performance of a pipe set at 500 MHz is 1 Gflop/s. Because of the four-way replication, a single CPU can deliver a peak performance of 8 Gflop/s. The vector units are complemented by a four-way super scalar processor; at 500 MHz this processor has a theoretical peak of 1 Gflop/s. The peak bandwidth per CPU is 32 gigabytes per second (GB/s) or 64 bytes per cycle (B/cycle). This is sufficient to ship eight 8-byte operands back or forth and just enough to feed one operand to each of the replicated pipe sets. See Tables 6.4 and 6.5.

Table 6.4
NEC SX-6 Series

Machine type	Distributed-memory multi-vector processor
Models	SX-6i, SX-6A, SX-6*x*M*y*
Operating system	Super-UX (Unix variant based on BSD V.4.3 Unix)
Connection structure	Multi-stage crossbar (see Remarks)
Compilers	Fortran 90, HPF, ANSI C, C++
Vendor's information Web page	http://www.hpce.nec.com
Year of introduction	2002

Table 6.5
SX-6 System Parameters

Model	**SX-6i**	**SX-6A**	**SX-6*x*M*y***
Clock cycle	500 MHz	500 MHz	500 MHz
Theoretical. peak performance, per proc. (64 bits)	8 Gflop/s	8 Gflop/s	8 Gflop/s
Maximal performance, single frame	8 Gflop/s	64 Gflop/s	---
Maximal performance, multi frame	---	---	8 Tflop/s
Main memory	4-8 GB	32-64 GB	≤ 8 TB
Number of processors	1	4-8	8-1024

It is interesting to note that the peak performance of a single processor has actually dropped from 10 Gflop/s in the SX-5 (the predecessor of the SX-6), to 8 Gflop/s. The reason is that the SX-6 CPU now houses on a single chip, an impressive feat, while the earlier versions of the CPU required multiple chips. The replication factor, which was 16 in the SX-5, has therefore been halved to 8 in the SX-6.

The SX-6i is the single CPU system (because of single chip implementation). It is offered as a desk-side model. A rack model is available, housing two non-connected systems.

In a single frame of the SX-6A, models fit up to 8 CPUs at the same clock frequency as the SX-6i. Internally, the CPUs in the frame are connected by a one-stage crossbar with a bandwidth of 32 GB/s/port. This is the same bandwidth as that of a single CPU system. The fully configured frame can therefore attain a peak speed of 64 Gflop/s.

In addition to these single-frame models, there are also multi-frame models (SX-6*x*M*y*) where the total number of CPUs is x = 8,...,1024 and the number of frames coupling the single-frame systems into a larger system is y = 2,...,128. SX-6 frames can be coupled in a multi-frame configuration two ways:

- A full crossbar (the IXS) that connects the various frames together at a speed of 8 GB/s for point-to-point unidirectional out-of-frame communication (1024 GB/s bi-sectional bandwidth for a maximum configuration)
- A HiPPI interface for inter-frame communication

With the IXS crossbar, the total multi-frame system is globally addressable, thus turning the system into a NUMA system. However, for performance reasons it is advisable to use the system in distributed memory mode with MPI. The HiPPI interface offers lower cost and speed.

The SX-6 uses CMOS technology, which appreciably lowers the fabrication costs and the power consumption. An HPF compiler is available for distributed computing, and an optimized MPI developed by NEC (MPI/SX) is available for message passing. OpenMP is available for shared memory parallelism.

The system attained 1484 Gflop/s, an efficiency of 97%. The size of the linear system for this result was 200,064.

NEC SX-7

On the SX-7, up to 32 CPUs are connected to a maximum 256 GB, large capacity shared memory in a single-node system. The system has realized an ultra-high data transfer speed of maximum 1130.2 GB/s between CPU and memory. This is 4.4 times faster than the existing SX-6-models. Large capacity memory of up to 16 TB can be configured in a 64-node multi-node system. The system can achieve a total data transfer speed of maximum 72 TB/s between CPU and memory. Moreover, in a multi-node system the SX-7 can achieve a maximum 18 Tflop/s of vector performance. See Table 6.6.

Table 6.6
SX-7 Specifications

I. Single Node System	
Central Processing Unit (CPU)	
Number of CPUs	4 ~32
Vector Performance	35.3 ~282.5Gflop/s
Vector Register	144k bytes x4 ~32
Scalar Register	64bits x128 x4 ~32
Main Memory Unit	
Memory Architecture	Shared Memory
Capacity	32 ~256G Bytes
Maximum Transfer Rate	1,130.2 G Bytes/Sec.
Input/Output Processor (IOP)	
Number of IOPs	1 ~4
Maximum Channel	127 channels
Maximum Transfer Rate	8 G Bytes/Sec.

II. Multi-node System	
Number of Nodes	
Central Processing Unit (CPU)	
Number of CPUs	16 ~2,048
Vector Performance	141.2G ~18,083Gflop/s
Vector Register	144k bytes x16 ~2,048
Scalar Register	64bits x128 x16 ~2,048
Main Memory Unit	
Memory Architecture	Shared/Distributed Memory
Capacity	128G ~16T Bytes
Maximum Transfer Rate	Max.72T Bytes/Sec.
Input.Output Processor (IOP)	
Number of IOPs	Max.256
Maximum Channel	Max. 8,128channels
Maximum Transfer Rate	Max. 512G Bytes/Sec.
Internode Crossbar Switch (IXS)	
Maximum Transfer Rate	Max. 512G Bytes/Sec.

Observations

NEC appears to be committed to high performance vector computing. NEC believes that the development of the high-end product will spur the technology needed for the other system components. NEC has more than 600 customers, though in the United States only the Arctic Regional Supercomputer Center uses a NEC supercomputer. NEC claims to have 15 new SX-7 systems, and has sold around 225 SX-6 systems.

One of the reasons that NEC has invested in HPF technology is that Dr. Miyoshi, the visionary behind the Earth Simulator, insisted that the Earth Simulator use HPF. NEC fully supports MPI 1 and 2. Their standard

collection of compilers is Fortran, C, and C++. They have available debugging tools from Vampir and use TotalView. Their Fortran and C compilers use different optimization strategies. Table 6.7 lists NEC machines in the Top500 as of June 2004. (New Top500 lists from Nov 2004 are available in Appendix C)

Table 6.7
NEC Machines in the Top500 (June 2004)

#	Location	Machine	Area	Country	Year Intro-duced	Linpack Perform (Gflop/s)	Procs	Peak Rate (Gflop/s)
1	Earth Simulator Center	Earth Simulator	Research	Japan	2002	35860	5120	40960
68	Meteorological Research Institute/JMA	SX-6/248M31 (typeE, 1.778ns)	Research	Japan	2004	2155	248	2232
148	DKRZ - Deutsches Klimarechenzentrum	SX-6/192M24	Research	Germany	2003	1484	192	1536
161	National Institute for Fusion Science	SX-7/160M5	Research	Japan	2003	1378	160	1412.8
184	Osaka University	SX-5/128M8 3.2ns	Academic	Japan	2001	1192	128	1280
194	Institute of Space & Astronautical Science (ISAS)	SX-6/128M16 (typeE, 1.778ns)	Research	Japan	2004	1141	128	1152
249	NEC Fuchu Plant	SX-6/128M16	Vendor	Japan	2002	982	128	1024
275	United Kingdom Meteorological Office	SX-6/120M15	Research	U.K.	2003	927.6	120	960
276	United Kingdom Meteorological Office	SX-6/120M15	Research	U.K.	2003	927.6	120	960
289	VW (Volkswagen AG)	Opteron 2.0 GHz, GigE	Industry	Germany	2004	891	360	1440
457	CBRC - Tsukuba Advanced Computing Center - TACC/AIST	Magi Cluster PIII 933 MHz	Research	Japan	2001	654	1040	970

FUJITSU

Background

Fujitsu produced Japan's first vector processor, the FACOM 230-75 APU (Array Processing Unit), which was installed at the National Aerospace Laboratory in 1977 to support list-directed vector accesses, a function it continues to serve today. AP-Fortran, an extension of the standard Fortran, was developed to derive the maximum performance from the APU hardware by including vector descriptions. The maximum performance of the APU was 22 Mflop/s in vector operations.

In July 1982, Fujitsu announced two models of the FACOM vector processor, the VP-100 and the VP-200, employing pipeline architecture and having multiple pipeline units that could operate concurrently. The maximum performance of the VP-100 and VP-200 were 285 Mflop/s and 570 Mflop/s, respectively. The first VP-200 was installed in December 1983.

In 1985, Fujitsu announced the entry model VP-50 and the top-of-the-line model VP-400, with peak vector performance of 140 Mflop/s and 1140 Mflop/s, respectively. The pipelines on the VP400 were four-way replicated. The VP-400 has since been further enhanced to give the peak vector performance of 1700 Mflop/s.

In December 1988, Fujitsu announced its VP2000 series supercomputer systems in various configurations, including uni-processor and dual scalar processor models. A year later, Fujitsu announced an enhancement of vector performance for its high-end model VP2600; this was followed by new quadruple scalar processor models in August 1990. The series supported full upward compatibility with the VP series. Fujitsu produced 10 models of the VP2000 series covering a range of vector performance from 0.5 Gflop/s to 5 Gflop/s. The Model 10 (VP2100/10, VP2200/10, VP2400/10, VP2600/10) was a uni-processor system, while the Model 20 (VP2100/20, VP2200/20, VP2400/20, VP2600/20) was a dual scalar processor system in which two scalar units could share one vector unit. The Model 40 (VP2200/40, VP2400/40) was a quadruple scalar system in which two sets of dual scalar processor systems (including the vector unit) were tightly coupled. The dual scalar processor models were introduced to increase the performance of usual programs where the busy rate of the vector unit was less than half.

In October 1992, Fujitsu announced its third-generation supercomputer, the VPP500 parallel supercomputer. The VPP500 was a distributed memory vector-parallel machine with 1.6 Gflop/s vector processor as a building block. The architecture of the VPP500 stood in sharp contrast not only to shared-memory parallel-vector processors, but also to massively parallel processors. The system scalability supported from 4 to 222 processors interconnected by a high-bandwidth crossbar network. Fujitsu extended the line of vector-parallel supercomputers to the VPP300 and VPP700 systems announced in 1995 and 1996, respectively, based on CMOS technology and air cooling.

The VPP5000 was the successor to the former VPP700/VPP700E systems (the “E” stood for “extended,” i.e., a clock cycle of 6.6 instead of 7 ns). The overall architectural changes with respect to the VPP700 series are slight. The clock cycle was halved and the vector pipes were able to deliver floating multiply-add results. With a replication factor of 16 for these vector pipes, the system could generate 32 floating-point results per clock cycle, at least in theory. In this way the VPP5000 could attain a four-fold increase in speed per processor with respect to the VPP700E.

The architecture of the VPP5000 nodes was almost identical to that of the VPP700. Each node, called a Processing Element (PE) in the system is a powerful (9.6 Gflop/s peak speed with a 3.3 ns clock) vector processor in its own right. A RISC scalar processor with a peak speed of 1.2 Gflop/s complemented the vector processor. The scalar instruction format was 64 bits wide and could cause the execution of up to four operations in parallel. Each PE has a memory of up to 16 GB while a PE communicates with its fellow PEs at a point-to-point speed of 1.6 GB/s. This communication is taken care of by separate Data Transfer Units (DTUs). To enhance the communication efficiency, the DTU had various transfer modes, including:

- contiguous
- stride
- sub array
- indirect access

The DTUs handled the translation of logical to physical PE-ids and from Logical in-PE addresses to real addresses. When synchronization was required each PE could set its corresponding bit in the Synchronization Register (SR). The value of the SR was broadcast to all PEs and synchronization had occurred if the SR had all its bits set for the relevant PEs. This method, which was comparable to the use of synchronization registers in shared-memory vector processors, proved to be much faster than synchronizing via memory. The network was a direct crossbar that should have led to an excellent throughput of the network. Contrast this arrangement with the VPP700, in which a level-2 crossbar was employed for configurations larger than 16 processors. On special order, Fujitsu could build 512 PE systems, quadrupling the maximum amount of memory and the theoretical peak performance.

The VPP5000U was one of the few single-processor vector processors offered by Fujitsu. It was simply a single-processor version of the VPP5000, of course without the network and data transfer extensions that are required in the VPP5000.

The Fortran compiler that came with the VPP5000 had extensions that enabled data decomposition by compiler directives. This evaded, in many cases, having to restructure the code. The directives were different from those as defined in the High-Performance Fortran Proposal but it should be easy to adapt them. Furthermore, it is possible to define parallel regions, barriers, etc., via directives, while there are several intrinsic functions to enquire about the number of processors and to execute POST/WAIT commands. Furthermore, a message passing programming style is possible by using the available PVM or MPI communication libraries.

Remarks for the Primepower Series

Today, Fujitsu has abandoned vector computing and has turned to cluster-based technology. Their new system, the Primepower HPC2500, is based on Sparc architecture with 8 CPUs per board (5.2 Gflop/s peak per processor at 1.3 GHz), 16 boards per node. Nodes are connected with a 8.3 GB/s connection to a crossbar (133 GB/s) and then connected through a 4 GB/s X 4 optical crossbar interconnect. This allows up to 128 nodes (16,384 processors). The complete system would peak at 85 Tflop/s and 64 TB of memory. Though Fujitsu is using the Sparc architecture, they have built their own version of the chip. Fujitsu uses Solaris as the operating system. Parallelnavi, a compiler for parallel programs on Primepower, is based on Solaris. See the site report for Fujitsu in *Appendix B* for additional information on the Primepower HPC2500 architecture.

The other high-end system is cluster-based. It uses Intel Pentium processors connected with Infiniband (8Gb/s X 2 ports) connected on the 133 MHz 64 bit PCI-X bus or Myrinet 2000 (4Gb/s). The cluster software is based on SCore, which is similar to the Scyld cluster operating system. SCore was developed as part of the Japanese Real World Computing Project (RWCP). See the site report for Fujitsu in Appendix B.

Fujitsu's vector architecture VPP system had a 300 MHz clock and as a result had weak scalar performance compared to commodity processors, like the Primepower. The VPP saw 30% peak performance on average for applications, while the Primepower sees about 10% peak performance on average. The difference can easily be made up in the cost of the systems. The VPP is 10 times the cost of the Primepower system.

Future versions of the HPC2500 will use the new Sparc chip 2 GHz by the end of the year. See Tables 6.8 and 6.9 and the site report for Fujitsu in Appendix B.

Table 6.8
Fujitsu/Siemens Primepower Series

Machine type	RISC-based shared-memory multi-processor
Models	Primepower 1500, 2500
Operating system	Solaris (Sun's Unix variant)
Connection structure	Crossbar
Compilers	Parallel Fortran 90, C, C++
Vendors information web page:	http://primepower.fujitsu.com/en/index.html
Year of introduction	2002

Table 6.9
Primepower System Parameters

Model	Primepower 1500	Primepower 2500
Clock cycle	1.35 GHz	1.3 GHz
Theoretical peak performance, per proc. (64 bits)	2.7 Gflop/s	5.2 Gflop/s
Maximal performance	86.4 Gflop/s	666 Gflop/s
Main memory		
Memory/node	≤ 4 GB	≤ 4 GB
Memory/maximal	≤ 128 GB	≤ 512 GB
Number of processors	4—32	8—128
Communication bandwidth		
Point-to-point	---	---
Aggregate	---	133 GB/s

Observations

Today the National Aerospace Laboratory of Japan has a 2304 processor Primepower 2500 system based on the Sparc 1.3 GHz. This is the only Fujitsu computer on the Top500 list that goes over 1 Tflop/s. Table 6.10 lists the other current Fujitsu systems.

Table 6.10
Fujitsu System Installations

Location	System
Japan Aerospace Exploration Agency	Primepower 128CPU x 14 (Computer Cabinets)
Japan Atomic Energy Research Institute (ITBL Computer System)	Primepower 128CPU x 4 + 64CPU
Kyoto University	Primepower 128CPU x 11 + 64CPU
Kyoto University (Radio Science Center for Space and Atmosphere)	Primepower 128CPU + 32CPU
Kyoto University (Grid System)	Primepower 96CPU
Nagoya University (Grid System)	Primepower 32CPU x 2
National Astronomical Observatory of Japan (SUBARU Telescope System)	Primepower 128CPU x 2
Japan Nuclear Cycle Development Institute	Primepower 128CPU x 3
Institute of Physical and Chemical Research (RIKEN)	IA-Cluster (Xeon 2048CPU) with Infiniband & Myrinet
National Institute of Informatics - (NAREGI System)	IA-Cluster (Xeon 256CPU) with Infiniband Primepower 64CPU
Tokyo University (The Institute of Medical Science)	IA-Cluster (Xeon 64CPU) with Myrinet Primepower 26CPU x 2
Osaka University (Institute of Protein Research)	IA-Cluster (Xeon 160CPU) with Infiniband

In many respects this machine is very similar to the Sun Microsystems Fire 3800-15K. The processors are 64-bit Fujitsu implementations of Sun's Sparc processors, called Sparc 64 V processors, and are completely compatible with the Sun products. Also the interconnection of the processors in the Primepower systems is similar to that of the Fire 3800-15K: a crossbar that connects all processors at the same footing, i.e., *not* a NUMA machine.

For the Top500, a cluster of 18 fully configured Primepower 2500s was used to solve a linear system of order N= 658,800. This yielded a performance of 5.4 Tflop/s with an efficiency level of 45% on 2,304 processors. See Table 6.11.

Table 6.11
Fujitsu Machines in the Top500 (June 2004)

#	Location	Machine	Area	Year Intro-duced	Linpack Perform (Gflop/s)	Procs	Peak Rate (Gflop/s)
7	Institute of Physical and Chemical Res. (RIKEN)	RIKEN Super Combined Cluster	Research	2004	8728	2048	12534
22	National Aerospace Laboratory of Japan	Primepower HPC2500 (1.3 GHz)	Research	2002	5406	2304	11980
24	Kyoto University	Primepower HPC2500 (1.56 GHz)	Academic	2004	4552	1472	9185
393	University of Tsukuba	VPP5000/80	Research	2001	730	80	768

HITACHI

Background

Hitachi was founded in 1910 as an electrical repair shop and quickly grew to encompass the manufacture of electric motors, appliances, and ancillary equipment. In 1959 the company built its first transistor-based electronic computer, Today, Hitachi's various divisions manufacture power and industrial systems, electronics devices, digital media and consumer products, and information and telecommunications systems. This last division is responsible for nearly 20% of the company's revenue. Currently Hitachi Ltd. has six corporate labs in Japan, five in the United States, and four in Europe. The company has over 5,000 people engaged in R&D. In 2002, it spent nearly ¥318 billion on R&D across all business areas.

Twenty years ago, the company entered the high performance computing field. Hitachi's main areas of focus were hardware, operating systems, compilers, and parallelizing techniques for application programs. Hitachi produced the HITAC M-180 IAP (Integrated Array Processor) in 1978, the M-200H IAP in 1979, and the M-280H IAP in 1982. The following year they introduced Japan's first vector machine, the S810. The S820 followed in 1987, possessing a peak single CPU vector performance of 3 Gflop/s. Refer to the site report for Hitachi in Appendix B for a flowchart summary of Hitachi's HEC developments over time.

Hitachi developed the SR2201 in 1996 and the SR8000 two years later. Both are RISC parallel machines based on pseudo-vector processing, or PVP. PVP, developed as part of a collaborative research effort between Hitachi and the University of Tsukuba, generates instructions that process the data referenced in a loop in one of the following ways:

- The data is loaded beforehand in a floating-point register and the data loading is completed while the loop that references the data is performing calculations from previous iterations. This is called *preload optimizing*.

- The data is transferred beforehand onto memory cache and the transfer to the cache memory is completed while the loop that references the data is performing calculations from previous iterations. This is called *prefetch optimizing*.

PVP offers a performance increase over RISC processors. Generally, a RISC processor machine has a cache memory between the processor and the main memory for high-speed data transmission to the processor, which thereby increases the performance. For many numerical calculations cache memory gets in the way of accessing large arrays of data and can lead to a loss of performance. PVP allows higher-speed transmission of data from the memory to the processor; for operations on long vectors, one does not incur the detrimental effects of cache misses that often ruin the performance of RISC processors, unless code is carefully blocked and unrolled.

The PVP used in the SR2201 was first developed for the CP-PACS machine at the Center for Computational Physics, University of Tsukuba. The most recent machine in the family is the SR11000, delivered in 2004, which also uses PVP. Refer to the site report for Hitachi in *Appendix B* for a diagrammatic explanation of PVP.

Remarks for the SR8000

The SR8000 is the third generation of distributed-memory parallel systems of Hitachi. It is designed to replace its direct predecessor, the SR2201, as well as the late top-vector processor, the S-3800. The basic node processor is a 2.22—4 ns clock PowerPC node with major enhancements from Hitachi, such as hardware barrier synchronization and PVP. The SR8000 features both preload and prefetch optimizing.

The peak performance per basic processor, or IP, can be attained with two simultaneous multiply/add instructions resulting in a speed of 1 Gflop/s on the SR8000. However, eight basic processors are coupled to form one processing node, all addressing a common part of the memory. For the user this node is the basic computing entity with a peak speed of 8 Gflop/s.

Hitachi refers to this node configuration as COMPAS (for *C*o-*o*perative *M*icro-*P*rocessors in single *A*ddress *S*pace). In most of these systems, the individual processors in a cluster node are not accessible to the user. Every node also contains a system processor (SP) that performs system tasks while also managing communications with other nodes and with a range of I/O devices.

The SR8000 has a multi-dimensional crossbar with a bi-directional link speed of 1 GB/s. From 4—8 nodes the cross-section of the network is 1 hop. For configurations 16—64 it is 2 hops and from 128-node systems on, it is 3 hops.

The E1 and F1 models are in almost every respect equal to the basic SR8000 model. However, the clock cycles for these models are 3.3 and 2.66 ns, respectively. Furthermore, the E1, F1, and G1 models can house twice the amount of memory per node, and their maximum configurations can be extended to 512 processors. These factors make them, at the time of this writing, theoretically the most powerful systems available commercially. Hitachi claims a bandwidth of 1.2 GB/s for the network in the E1 model, with a bandwidth of 1 GB/s for the basic SR8000 and the F1. By contrast, the G1 model has a bandwidth of 1.6 GB/s.

The following software products are supported in addition to those already mentioned above: PVM, MPI, ScaLAPACK, and BLAS. In addition, Hitachi offers numerical libraries such as NAG and IMSL. For the SR8000 models, MPI, PVM, and HPF are all available. See Tables 6.12 and 6.13.

Table 6.12
Hitachi SR8000 System

Machine type	RISC-based distributed memory multi-processor
Models	SR8000, SR8000 E1, SR8000 F1, SR8000 G1
Operating system	HI-UX/MPP (Micro kernel Mach 3.0)
Connection structure	Mult-dimensional crossbar
Compilers	Fortran 77, Fortran 90, Parallel Fortran, HPF, C, C++
Vendors information Web page	www.hitachi.co.jp/Prod/comp/hpc/eng/sr81e.html
Year of introduction	Original system in 1998; E1 and F1 in 1999; G1 in 2000

Table 6.13
SR8000 System Parameters

Model	SR8000	SR8000 E1	SR8000 F1	SR8000 G1
Clock cycle	250 MHz	300 MHz	375 MHz	450 MHz
Theoretical. peak performance, per node (64 bits)	8 Gflop/s	9.6 Gflop/s	12 Gflop/s	14.4 Gflop/s
Maximal performance	1 Tflop/s	4.9 Tflop/s	6.1 Tflop/s	7.3 Tflop/s
Main memory				
Memory/node	≤8 GB	≤ 16 GB	≤ 16 GB	≤ 16 GB
Memory/maximal	≤ 1 TB	≤ 8 TB	≤ 8 TB	≤ 8 TB
Number of processors	4—128	4—512	4—512	4—512
Communication bandwidth	1 GB/s	1.2 GB/s	1 GB/s	1.6 GB/s

An SR8000 configured in 144-node G1 (450 MHz) mode obtained an observed speed of 1709 Gflop/s out of 2074, reaching an efficiency of 82% for the solution of a 141,000 full linear system. A 168-node 375 MHz F1 model achieved 1635 out of 2016 Gflop/s, an efficiency of 82%. On a single node of this processor, a speed of over 6.2 was measured in solving a full linear system, while a speed of 4.1 Gflop/s was measured in solving a full symmetric eigenvalue problem of order 5000.

Hitachi developed the processor chip for the SR8000 using an extension of the PowerPC architecture. Hitachi originally intended to use the processor widely, but today it is only used in the company's supercomputer line. This is in part because their chip design was optimized for the HEC arena and has no second level cache memory. The level 1 cache is 128 KB with 128 registers for PVP features. The latency from memory ranges from 100 to several hundred cycles.

The cache uses a write-through policy. Conflicts in cache are resolved in one of two ways:

- For sequential access, the compiler generates prefetches
- For irregular and strided accesses, the compiler generates preloads

Remarks for the SR11000

The Super Technical Server SR11000 Model H1 can be fitted with anywhere from four to 256 nodes, each of which is equipped with 16 - 1.7GHz IBM Power4+ processors. Each nodes achieves a theoretical computation performance of 108.8Gflop/s. This is approximately four times the performance of the predecessor SR8000 series. The architecture of the SR11000 is in many ways similar to the SR8000:

- 16-way SMP node
- 256 MB cache per processor
- High memory bandwidth SMP
- PVP equipped
- COMPAS for providing parallelization of loops within a node
- High-speed internode network
- AIX operating system
- No hardware preload for compiler
- No hardware control for barrier
- Nodes connected by IBM's High-Performance Switch™
- 2 to 6 links per processor (or planes)
- AIX with cluster system management

Unlike the SR8000, the SR11000 does not have a preload feature. Instead it relies on a prefetch controlled by software and hardware. LINPACK efficiency is at 82% on the SR8000 and 80% (target) on the SR11000. In comparison with the IBM p690 system using a Power4 processor, the SR11000 has 6 planes per 16 processors to the IBM's 8 planes per 32 processors.

With regard to compilers and libraries for the SR11000, Hitachi offers an optimized Fortran 90 compiler, and optimized C. Though Hitachi uses IBM's VisualAge C++, their compiler effort is separate from IBM's efforts. They are focusing their efforts on developing an automatic parallelizing compiler, rather than on co-Array Fortran or other languages. Hitachi has no plans for HPF because customers found the performance for HPF to be too low for their needs.

Hitachi intends to maintain the ratio between network speed and node performance to about 10:1. They expected to accomplish this by using 16 processors per node to improve memory bandwidth to processor ratio. Though this improves the network to processor ratio, it uses 6 planes rather than 8. With a 100 GFlop/s node, a communication rate of 10 GB/node is desired. See Table 6.14.

Table 6.14
SR11000 Model H1 System Parameters

System	Number of nodes	4	8	16	32	64	128	256
	Peak performance	435GF	870GF	1.74TF	3.48TF	6.96	13.9TF	27.8TF
	Maximum total	256GB	512GB	1TB	2TB	4TB	8TB	16TB
	Inter-node transfer speed	4GB/s (in each direction) x 2						
		8GB/s (in each direction) x 2						
		12GB/s (in each direction) x 2						
	External interface	Ultra SCS13, Fibre Channel (2Gbps), GB-Ether						
Node	Peak performance	108.8 Gflop/s						
	Memory capacity	32GB/64GB						
	Maximum	8GB/s						

Observations

At this point, Hitachi has three customers for the SR 11000, which tops out at 7 Tflop/s and whose largest system stands at 64 nodes. All three are part of the Ministry of Education, Culture, Sports, Science and Technology (MEXT). The Okazaki Institute for Molecular Science already possesses a 50 node machine. Though not yet announced, the National Institute for Material Science in Tsukuba will be acquiring a 64 node machine. The Institute for Statistical Mathematics has plans for four nodes.

The University of Tokyo, which has a long history of using Hitachi machines, may well buy the SR11000. The company's current close collaboration with IBM will continue, though they may reconsider alternatives at a later time.

Table 6.15 lists the Hitachi SR8000 machines in the Top500 as of June 2004.

Table 6.15
Hitachi SR8000 Machines in the Top500 (June 2004)

#	Location	Machine	Area	Country	Year Intro-duced	Linpack Perform (Gflop/s)	Procs	Peak Rate (Gflop/s)
122	University of Tokyo	SR8000/MPP	Academic	Japan	2001	1709.1	1152	2074
127	Leibniz Rechenzentrum	SR8000-F1/168	Academic	Germany	2002	1653	168	2016
280	High Energy Accelerator Research Organization /KEK	SR8000-F1/100	Research	Japan	2000	917	100	1200
295	University of Tokyo	SR8000/128	Academic	Japan	1999	873	128	1024
333	Institute for Materials Research/Tohoku University	SR8000-G1/64	Academic	Japan	2001	790.7	64	921.6
416	Japan Meteorological Agency	SR8000-E1/80	Research	Japan	2000	691.3	80	768

REFERENCES

Hitachi Global History: 1910-1959. <http://www.hitachi.com/about/history/1910_1959/index.html> Last accessed February 25, 2005.

Parallelnavi: A Development and Job Execution Environment for Parallel Programs on Primepower [Abstract], *Fujitsu Magazine*, v52 n1, <http://magazine.fujitsu.com/vol52-1/v52n1a-e.html> Last accessed February 25, 2005.

Top 500 Supercomputer Sites. 2004. <http://top500.org/> Last accessed February 25, 2005.

CHAPTER 7

SOFTWARE FOR HIGH-END COMPUTING

Katherine Yelick

BACKGROUND

The Earth Simulator (ES) is generally known for its physical characteristics and hardware performance, but the announcement of some of the earliest results on the system came with an equally surprising software story: some of the large, multi-teraflop/s codes were written in high-performance Fortran (HPF), a language that had been almost entirely abandoned within the United States. Just as the success of the hardware system is attributable to a focused and sustained design and development effort, so too is the success of HPF. The original visionary of the ES, Dr. Hajime Miyoshi, envisioned a machine that was not only a major leap in performance relative to previous machines, but was also easier to program. He therefore made HPF a required part of the initial requirements for the machine.

The use of HPF was the major innovation discussed in software for supercomputers, and in the case of the Earth Simulator, it was entirely an industrial effort by NEC. A consortium of industrial and academic members also had a role in the Japanese HPF effort to design their own variation in the language, which is described below. It was striking that there were few, if any, academic efforts on high-performance numerical libraries, problem-solving environments, or application frameworks for the ES or other vector supercomputers. Instead, all of these research problems were represented in the grid computing projects, often for cluster computers that formed computational and storage resources on the grid. Grid computing research is described in another chapter.

Application performance results on the ES demonstrate that the ES has a scalable operating system, collective communication, and a high-performance Fast Fourier Transform. While each of these might be part of a research effort in a grid environment, for the ES they were considered a standard part of the system that was the vendor's responsibility to provide. Thus, the cost of the ES includes a much higher level of software support than one would expect when purchasing a cluster. There was surprisingly little in the form of open source software running on the machine.

The remainder of this chapter will therefore focus on the HPF effort in Japan, beginning with a brief summary of high-performance programming languages to set the stage for further discussion. It then describes the Japanese extensions to HPF, followed by application performance results on the ES. Finally, it summarizes the impact of this effort and looks ahead towards the future of HPF in Japan, which is less promising.

HIGH-PERFORMANCE PROGRAMMING OVERVIEW

High-performance programming models include both integrated languages, designed specifically for a high-end computing environment, and libraries that are combined with standard languages such as Fortran, C or C++, where the parallelism constructs are provided in the library. HPF is an example of an integrated language, which requires its own specialized compiler, whereas the Message Passing Interface (MPI) is the most common library

extension used in high-end computing today. OpenMP is somewhere between a fully integrated language and a library, involving compiler directives and runtime library support. A fourth model of parallel machine use is to hide the parallelism from the application programmer entirely by having a sophisticated compiler automatically convert sequential loops into parallel code. While this latter model is the most attractive for programmers, it has proven impractical for machines of this scale and for applications that are as large and complex as those used on high-end machines today.

HPF falls into the general category of data parallel language. It is an extension of the Fortran 90 language in which programmers express fined-grained parallelism over arrays. For example, a simple array statement such as A = B + C may expresses a parallel addition operation followed by parallel assignment, all for arrays that are spread across processors. HPF allows the programmer explicit control over the layout of arrays across processors, which the compiler then uses to partition the work across processors. The HPF language design effort began in 1991, and a *de facto* standard (HPF v1.0) was distributed in 1993. In the United States there was a concerted effort to implement and optimize HPF in the mid-90s, both in industry and in academia, but nearly all of that activity has been abandoned, in part because the initial performance results were disappointing.

MPI is a library of communication primitives based primarily on two-sided (send/receive) communication that is relatively easy to implement, since it does not require its own compiler support. It therefore runs on parallel machines across a spectrum of cost, size, and speed, giving the programmer precise control over both data layout and computation distribution, which are important for performance. The disadvantage is that the two-sided communication model can be cumbersome for some programs and that there is significant up-front investment in parallelism, hurdles some users never overcome.

Many of the high-end machines in use today have multiple levels of parallelism, with a set of nodes connected by a network at the top level, shared memory parallelism within the nodes, and either superscalar or vector parallelism within each processor in the node. This hierarchy is reflected in the programming models, with MPI used between nodes, and either MPI (implemented on shared memory), OpenMP, or threads within shared memory. At the lowest level, an automatic vectorizing compiler is often used on vector architecture, or code scheduling is done on a superscalar processor to enable hardware parallelism. In Japan the shift from vector supercomputers to cluster architectures has made vectorization less important and MPI more important, although there is a strong interest in trying to retain the programming models from vector machines on current and future hardware.

SUPERCOMPUTER VENDOR SOFTWARE

NEC has a significant ongoing investment in HPF. They participated in the international HPF language definition effort and also on the Japanese HPF/JA design effort, which is described below. They also support MPI versions 1 and 2, and vectorizing compilers for Fortran, C, and C++. Their compilers use different strategies for C/C++ when compared to Fortran, and as a result they see much better vector performance from Fortran code than from C/C++. They also provide debugging and performance tools such as TotalView and Vampir. Starting with their SX-3 generation of processors, they use Unix as the operating system on their machines.

The operating system and compilers for the ES were provided entirely by NEC, although there are some ongoing research efforts at the ES facility that involve work in these areas. The machine has several programming models available to users, including message passing or HPF between processing nodes. While NEC supports all the more widely used MPI and OpenMP, it is also still investing in HPF and automatic parallelization, as illustrated in Figure 2.9 in Chapter 2. The MPI implementation has a latency of 5.6 microseconds and achieves a maximum bandwidth of 11.8 GBytes/sec, out of a hardware peak of 16 GBytes/sec. Within a processing node, the eight application processors communicate by shared memory, and the parallelism can be expressed using HPF or OpenMP, or the parallelization may be left to an automatic parallelizing compiler.

Fujitsu has a similar set of programming models, strongly influenced by their former vector product line. Fujitsu developed their own parallel data decomposition language, which is similar to HPF, but is unique to Fujitsu machines. In addition to the layout directives, this VPP Fortran language has parallel regions, barriers, and semaphores, and intrinsic functions to find the number of processors. The PrimePower systems are non-vector machines, but they support a similar set of tools, including an XP Fortran compiler, which is a port of the VPP Fortran compiler. The PrimePower systems also support MPI and PVM for communicating between nodes and they run a cluster operating system based on Score, which was developed as part of the Real World Computing Project.

The JAXA site is an interesting example of one of Fujitsu's major customers. JAXA has a long history of buying Fujitsu machines, to the extent that their working relationship with Fujitsu is more important than the architecture or software characteristics. Rather than prioritizing portability of their application codes across all architectures, they place a higher emphasis on backward compatibility with the software they wrote on the VPP vector line. Thus, their codes are written in Fujitsu's XP Fortran rather than the somewhat more portable HPF or much more portable MPI. This strong allegiance between vendors and costumers is common in Japan.

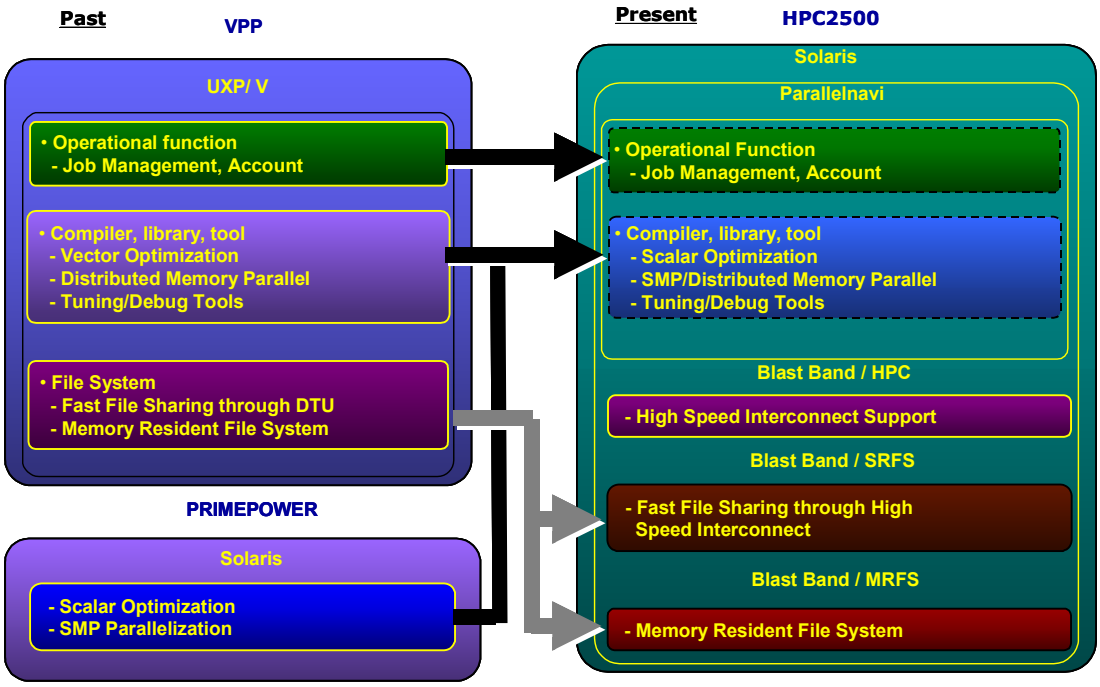
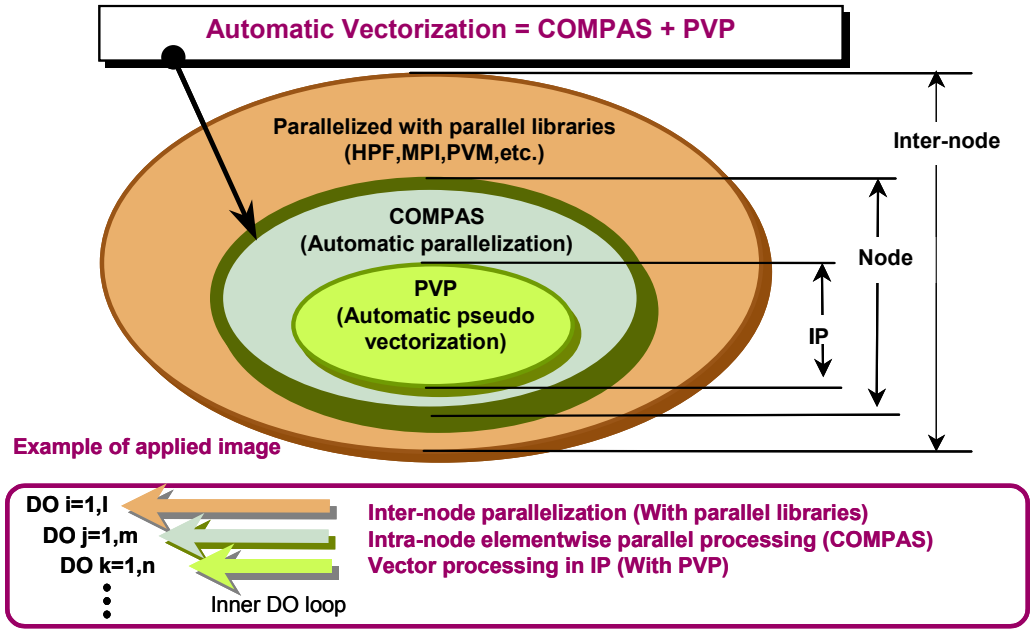


Figure 7.1. Programming models on Fujitsu machines (Courtesy Fujitsu)

Hitachi's supported programming models are similar to those of Fujitsu, because they have abandoned true vector processors in favor of the IBM Power processor line through a partnership with IBM. Also unwilling to give up the vector programming model on which their customers' code is based, Hitachi has pursued a line of pseudo-vectorization, with varying levels of hardware support to make vectorization easier. The current Power4+-based product has little hardware support for vectors, but continues to use the vector programming model within processing nodes. In addition to automatic vectorization, Hitachi continues to develop automatic parallelization techniques, which are also useful within a node. For the higher level, coarse-grained parallelism, they have offered MPI, PVM, and HPF, although the HPF effort is no longer active.



PVP: Pseudo Vector Processing
COMPAS: CO-operative Micro-Processors in single Address Space
IP: Instruction Processor

Figure 7.2. Parallel programming models for Hitachi machines (Courtesy Hitachi)

HPF/JA

The HPF effort in Japan involved significant compiler development efforts at the three main supercomputing vendors. In addition, though, some early applications experience was used to extend the language to make it more useful. HPF 2.0 extends the original HPF language for irregular applications, while HPF/JA is an orthogonal extension primarily altered for performance on applications with regular data structures. HPF/JA was developed by a Japanese consortium that included the three major vendors: NEC, Hitachi, and Fujitsu. NEC has an additional set of extension designs for the ES, with the resulting language called HPF/ES. Figure 7.3 gives an overview of the features in each of these languages.

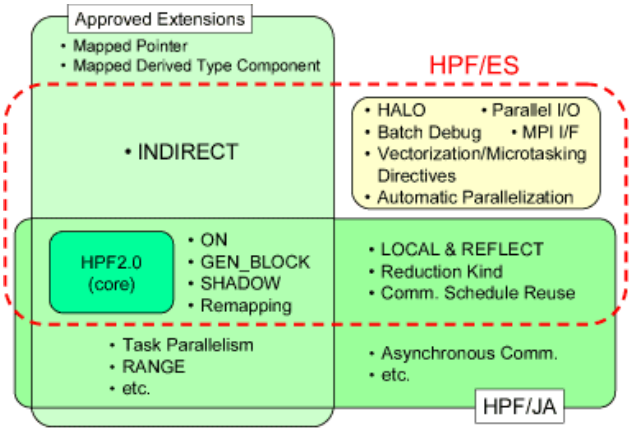


Figure 7.3. HPF/JA, HPF/ES, and HPF 2.0 extensions

Some of the key extensions in HPF/ES and HPF/JA include:

- REFLECT: a data layout used for nearest-neighbor computation.

- LOCAL: an assertion that no communication is needed within a given scope, thereby enabling better compiler optimizations with less sophisticated analysis.
- Extended ON HOME: partition computation replication which is useful in some problems if it avoids more expensive communication.
- HALO regions for updating ghost cells on unstructured meshed.

Of more general interest than the details of the language extension or particular performance results is the evidence of online parallel language activity in Japan, which lasted much longer than the HPF effort in the United States.

HPF ON THE EARTH SIMULATOR

The HPF language was used in at least two major application developments on the ES with remarkable results. The first is a plasma simulation code, IMPACT3D, which uses a Total Variation Diminishing (TVD) scheme programmed in HPF/EX. This code achieved 14.9 Tflop/s on a 512-node execution on the ES, which corresponds to 45% of the peak. This application earned the Gordon Bell Award for language in SC2002. Figure 7.4 shows visualizations of IMPACT3D results.

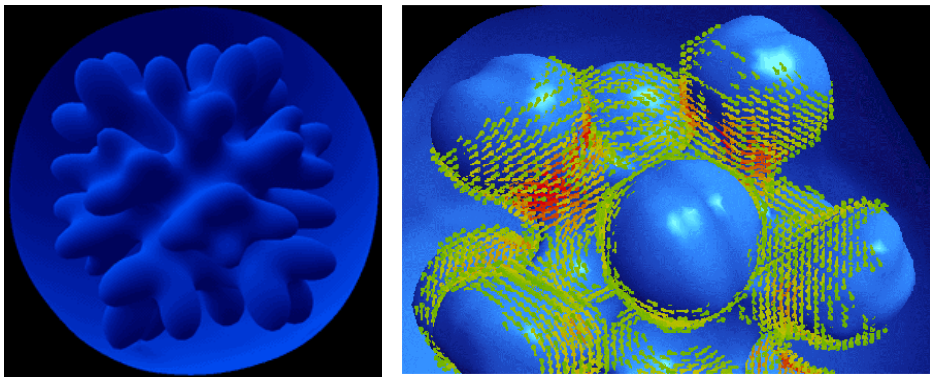


Figure 7.4. Visualization results from the IMPACT3D plasma code. The left picture shows the isovalue surfaces of density at the maximum compression of imploding targets in laser fusion; the right shows the pusher-fuel contact surface and vorticity.

The second major HPF/ES result is using the PFES code, an Oceanic General Circulation Model based on the Princeton Ocean Model. This code achieved 9.85 Tflop/s with 376 nodes (41% of the peak performance of the machine), using entirely automatic parallelization and vectorization within a node. This same code runs at 10.5 Tflop/s on the same number of nodes if some explicit parallelism is used within a node.

CONCLUSIONS

The software research agenda in Japan is currently skewed towards grid, which overlaps with cluster computing, and there are only modest research programs at the Japan Marine Science and Technology Center (JAMESTEC) for compilers, programming languages, and tools. HPF was much more successful in Japan than in the United States, probably due to a variety of factors:

- A reluctance of high-performance computing consumers to change programming languages, which leads to a demand for familiar vector programming models even on non-vector machines
- A good match between vector architectures and the fine-grained parallelism available in HPF. Message passing models are still used on these Japanese machines, but they are not required in all applications, and even when they are used, the extremely high-end vector/SMP compute nodes mean that the scale of parallelism at the MPI can be smaller relative to a commodity cluster.

- Longer sustained industrial projects to develop high-performance compilers, probably due to customer demands.
- Vision of the ES that included easier programming through HPF.

In spite of this relative success of HPF in Japan, especially on the ES, interest in the language has declined significantly. MPI is now the dominant model for programming most Japanese supercomputers today. The reason given by some vendors is that customers want portability of their code across machines. HPF does not run as well on non-vector hardware, and the vector hardware has become prohibitively expensive for at least two of the three Japanese supercomputing vendors. The investment in automatic parallelizing and vectorization compiles still seems to have paid off, because it allows the programmer to discover a smaller degree of parallelism at the coarse-grain level.

There was less work in supercomputing software than expected, which is explained by the interest in grid computing, which will be covered in the next chapter.

REFERENCES

PRIMEPOWER Benchmark Achievements – Fujitsu Siemens Computers.
<http://www.fujitsu-siemens.com/products/unix_servers/benchmarks.html> Last accessed February 23, 2005.

CHAPTER 8

GRID COMPUTING IN JAPAN

Katherine Yelick

BACKGROUND AND MOTIVATION

Within the international high-end computing community, Japan is currently best known for the Earth Simulator (ES), yet the Japanese research community is much more focused on grid and cluster computing than on such vector supercomputers. The reason for this interest in grids can be attributed to the combination of new research challenges on the grid, a shift in government funding from supercomputing to grids, and the low cost of entry into the grid computing arena when compared to supercomputing.

A precise definition of grid computing is difficult to find, and the definitions have shifted somewhat over time as the grid community refines their understanding of the technical challenges and opportunities. Foster, Kesselman, and Tuecke define the *grid problem* as "flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources---what we refer to as virtual organizations." The grid terminology comes from the analogy to the electric power grid, where one views computing as a public utility that one plugs into and pays for on demand. The grid computing models vary from a scenario in which large computations are performed by harvesting unused cycles on desktop machines throughout the system, e.g., the SETI@home model, to more modest proposals in which large computing and storage facilities provide major services, and users submit requests to the grid of machines, rather than to an individual machine on which they have a personal account. The key technical challenges in grid computing are authentication, authorization, resource access, and resource discovery.

In Japan, cluster computing is viewed as part of the grid problem, since these provide some of the major resources on the grid. Heterogeneous computing is also considered a key technical challenge, with the view that full application involves a set of different algorithms with different needs, and the computation may move around the grid as it runs to access the resources with the best fit to the demands of the algorithm.

Some of the government agencies, in particular METI, view grid computing as an enabler of an electronic society, and phrases like e-business, e-government, e-science and even e-Japan are popular. The e-Japan project, which has 103 subprojects, aims to make Japan the world's strongest IT nation by 2005. Part of the motivation for this line of work is a general push toward a "knowledge-emergent society," where everyone can utilize information technology. In 2001, Japanese Internet usage was at the lowest level among major industrial nations, and the government identified four strategies to address the problem:

1. Ultra high-speed network infrastructure: This gives all individuals access to information technology, with the network being a primary component of a computational grid.
2. Facilitate electronic commerce: METI funds research projects in grids for business applications.
3. Realize electronic government: By law there is no secret government information in Japan, and the idea of electronic government is to engage the citizens in government activities while also making the government more efficient through information sharing across agencies.

4. Nurturing high quality human resources: Training and support of students, researchers, etc. Whereas supercomputing may be viewed as an activity for a small elite group of scientists, grid computing is envisioned as a resource for all of society.

OVERVIEW OF GRID EFFORTS

The Japanese effort in grid computing research includes data and business grids, in addition to the computational grids used for large science and engineering applications. The nature of the grid makes these projects distributed and collaborative, so some researchers and institutions are involved in more than one of the projects described.

The foundation of the Japanese grid effort is SINet, a network infrastructure that covers much of Japan. The architecture is shown in Figure 8.1. The backbone is SuperSINet, a 10 Gbits/sec photonic network. There are Gigabit Ethernet bridges for peer connections and more than 6,000 km of dark fiber and over 100 e-e lambdas. SINet is connected internationally with a 5 Gbits/sec line.

One of the major grid computing efforts in Japan to use the SINet infrastructure is the National Research Grid Initiative (NAREGI), which is a five year project (FY2003-FY2007) funded by MEXT. The funding level was about ¥2B (close to $19 million US) in 2003. NAREGI is a collaboration of national laboratories, universities, and industry. One of the primary goals is develop grid software, and to perform basic research on grid middleware as it is developed. NAREGI is building a prototype for scientific grid infrastructure in Japan and is providing a testbed for grid computing research, with over 100 Tflop/s expected in 2007. The Phase I version of this testbed contains a 10 Tflop/s cluster at the Computational Nanoscience Center at the Institute for Molecular Science (IMS), another 5 Tflop/s at the Center for Grid R&D at NII, and a 17 Tflop/s system of their own. There are also smaller clusters at the other institutions, which include Kyushi University, Kyoto University, Hohoku University, AIST, KEK and ISSP.

NAREGI is emphasizing computational nanotechnology as one of the applications of a high-end grid computing environment, and they are performing some simulations over SINet. NAREGI researchers are active in international collaborations with researchers from the U.S., Europe, and the Asian Pacific region, and they also contribute to standards activities, such as the Global Grid Forum (GGF).

The IT-Based Laboratory (ITBL) is an applications-oriented grid effort involving several of the national laboratories: NAL, RIKEN, NIED, NIMS, JST, and JAERI. It has 21 connections to SINet and is looking at applications of grid technology, including mechanical simulation, computational biology, material science, environmental modeling, cell simulation, aerodynamics, and earthquake engineering. The first step in was to connect these laboratories using the SuperSINet network. On top of these networks ITBL is creating “Virtual Research Environments,” which is a set of computational platforms and software that are grid-enabling laboratory applications. Researchers without a Virtual Research Environment can easily share information across institutions and across disciplines. ITBL has 523 users at 30 institutions sharing the computer resources of 12 institutions. The years 2001 to 2003 were spent on development and infrastructure; the years 2003 to 2005 are for practice and expansion. ITBL is a grand design with the goal of increasing efficiency by sharing resources (computers, staff expertise) and increasing performance (combination of computers, parallel and pseudo-scalable systems, shared expertise). ITBL software efforts must address security issues with a firewall across the system, communication between heterogeneous computers and the Task Mapping Editor (TME, visual work flow).

The National Institute of Advanced Industrial Science and Technology (AIST) has a large grid effort called the Grid Technology Research Center (GTRC). There are six areas of research at GTRC: programming tools, high-speed networking, computer clusters as computer resources on the grid, international demonstration and verification experiments, application demonstrations, and practical issues in grid deployment including security and reliability.

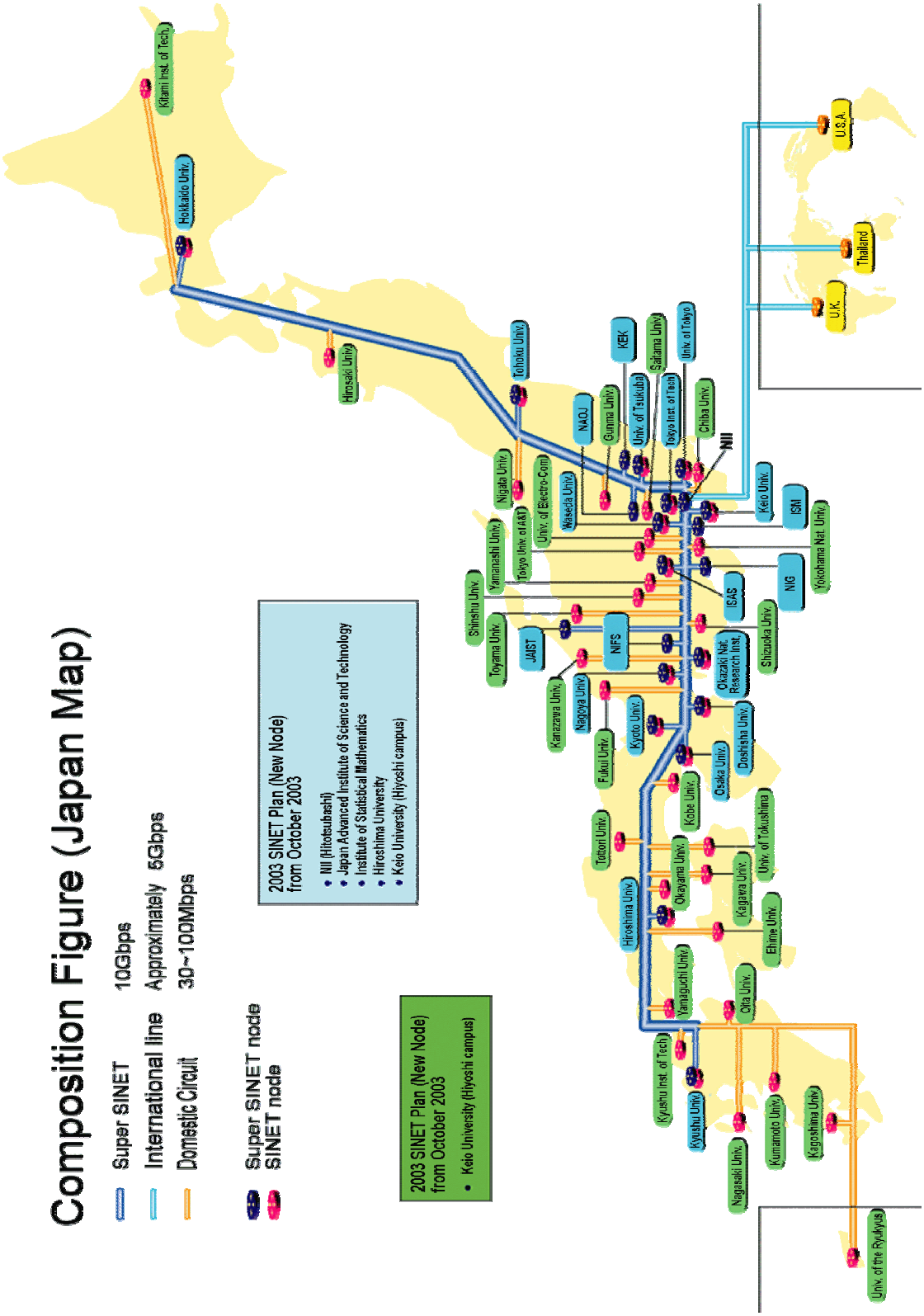


Figure 8.1. SuperSINET, an all-optical research network in Japan (Courtesy SINET)

At the Tokyo Institute of Technology, the Global Scientific Information and Computing Center (GSIC) does research in low-power, high-performance clustering technology and grid middleware, both in collaboration with researchers from other institutions. GSIC also has a major grid prototype called the "campus grid," which they use for experimental work.

GRID HARDWARE

Many of the research institutions in Japan have clusters that can be used for both middleware and applications development. A hallmark of most of the large clusters is their heterogeneity, which is going to be a likely characteristic of any grid computing environment. The heterogeneity in some cases involves different PC processor types and networking details, but in the extreme it may include special purpose hardware such as Grape processors for n-body calculations, vector supercomputers, and PCs all in a single logical grid facility, albeit distributed. This section highlights some of the larger hardware facilities for grid research.

AIST houses a large computing resource called the Supercluster in a large new building that was not yet occupied (except by the Supercluster) at the time of WTEC's visit. The purpose of the Supercluster is to provide a testbed for grid computing research. One of the research issues with grid computing is how to design software that works in a heterogeneous environment, spreading computations or storage across a system with a mixture of hardware, systems software, and performance characteristics. The Supercluster was intentionally designed to be heterogeneous, mixing processor types, networks, clock rates, and memory capacity, so software designed for the full system will naturally be heterogeneous.

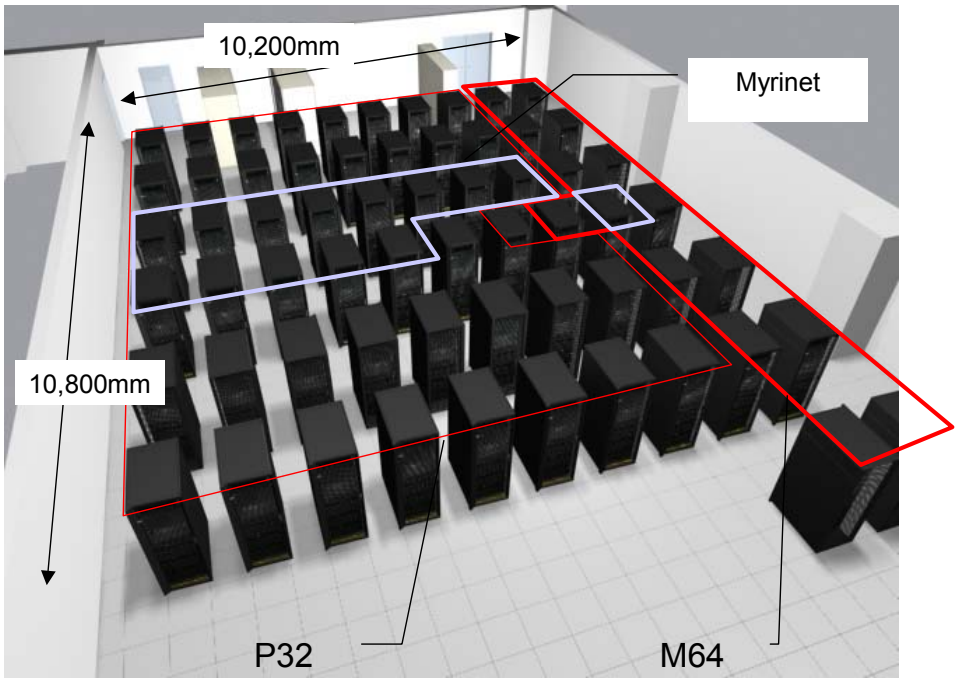


Figure 8.2. Supercluster at AIST/GTRC (Courtesy AIST)

A picture of the AIST Supercluster is shown above in Figure 8.2. It has a theoretical hardware peak of 14.6 Tflop/s and contains three clusters:

- P32: An 8.6 Tflop/s IBM eServer325 is a set of 128 dual processor Opteron nodes with Myrinet 2000 interconnect. The processors have a 2.0 GHz clock and 6GB of memory each.
- M64: A 2.7 Tflop/s Intel Tiger 4 cluster with 131 quad processor Itanium nodes with Myrinet 2000 interconnect. The processors are 1.3GHz Itanium (Madison) processors, each with 16GB of memory.

- F32: A 3.1 Tflop/s Linux Networx cluster with 256 dual Xeon nodes connected by Gigabit Ethernet. The processors are 3.06 GHz Xeons with 2GB of memory.

The cost to build this Supercluster was approximately $20 million, and it has an ongoing maintenance cost of about $1.4M annually.

The GSIC campus grid consists of 800 CPUs (400 servers) spread over 13 locations, and two Titech campuses, which are connected via Super TITANET (1-4 Gbit/sec). The system has a total peak performance of 1.2 Tflop/s and a total storage capacity of 25 Tbytes. It includes 752 CPUs (376 servers) of Express5800/BladeServers and was Japan's first grid computing system using blade technology. Blade technology is important because it achieves high-performance, space and thermal efficiency, as well as high price/performance ratio using commodity technology. The thermal and space efficiency of blades is especially important because the grid nodes are located in several remote departments and are often subject to severe space and thermal constraints.

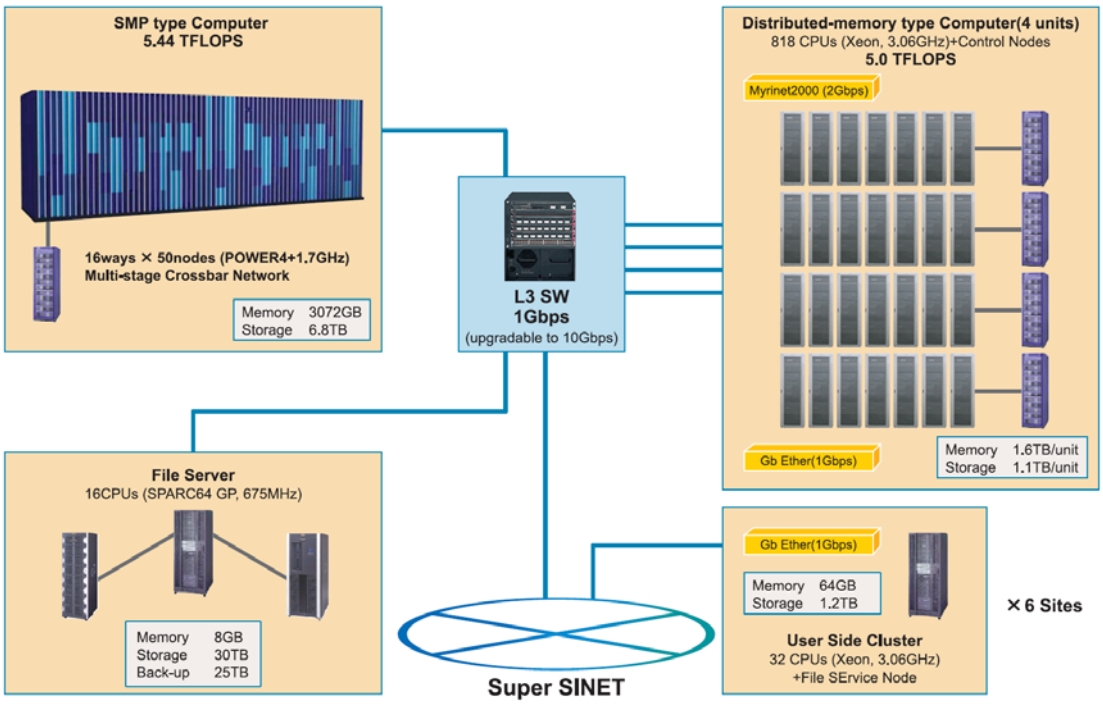


Figure 8.3. The IMS System for grid R&D, particularly for nanotechnology (Courtesy NAREGI)

Figures 8.3 and 8.4 show two additional grid prototypes, both exhibiting a degree of heterogeneity. The NII cluster is a testbed for building grid middleware and other general grid software. It contains a mixture of scalar and super-scalar processors types, such as Power4, Itanium-3, SPARC64, and Xeons. The IMS cluster is targeted towards nanotechnology, and contains shared and distributed memory multiprocessors, and is built from Itanium-2, Xeon, and Power4 machines.

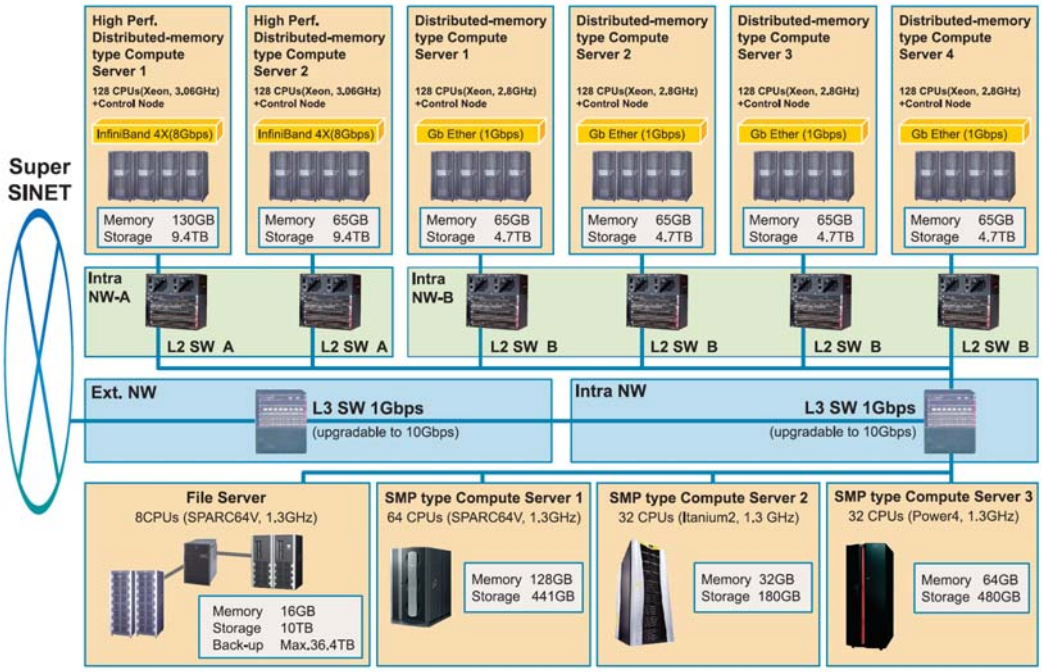


Figure 8.4. The NII heterogeneous cluster for developing grid software (Courtesy NAREGI)

GRID MIDDLEWARE

One of the challenges in grid computing is making them usable by application programmers. Currently, programming on the most general heterogeneous grid is significantly harder than writing an MPI program for a supercomputer. The goal of grid middleware and tools is to hide the complexity of the grid system, including any hardware heterogeneity, performance aberrations due to resource sharing, job control, and system management.

Central to NAREGI's goals and their five-year plan is research and development of a scaleable grid software environment that can support a variety of real world applications with distributed computing resources. The strategic approach NAREGI is taking to accomplish this complex task is one of "divide and conquer" by having six working groups (work packages), each focusing on a thematic area. Figure 8.5 shows a picture of these work packages, which are defined as follows.

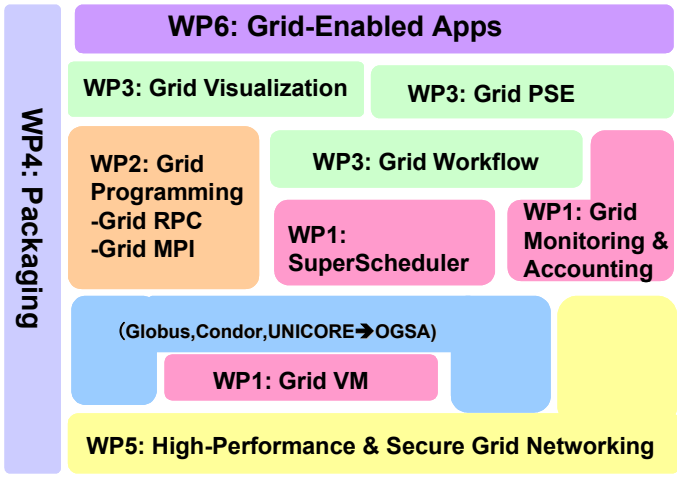


Figure 8.5. Work Packages in the NAREGI software system for grids (Courtesy NAREGI)

- WP-1: Lower and Middle-Tier Software for Resource Management (Matsuoka from Titech, Kohno from ECU, and Aida from Titech). This layer provides interoperability across low-level grid layers developed elsewhere, including Unicore, Condor, and Globus. It contains a Meta-scheduler to assign jobs to the grid based on its current workload and the needs of the job, a grid information service that does user and job auditing and accounting, and a virtual machine layer that supports co-scheduling and virtualization of the processor space.
- WP-2: Grid Programming Middleware (Sekiguchi and Ishikawa from AIST). This layer provides a basic programming mechanism for users, namely an implementation of GridRPC (a remote procedure call mechanism for the grid, which is an API defined by the Global Grid Forum). Ninf-G is a reference implementation of GridRPC developed at AIST, which is used by several groups outside of Japan. This layer also provides implementation of GridMPI, which allows users to run their MPI applications on the grid.
- WP-3: User-Level Grid Tools & PSEs (Miura from NII, Sato from Tsukuba-u, and Kawata from Utsunomiya-u). This level includes grid workflow tools, which allow programmers to describe the communication and computation workloads in their applications at a high level. It also contains visualization tools and Problem Solving Environments (PSEs), which are domain-specific software frameworks. Grid PSE Builder from AIST is a software environment that can be used to build web services, grid-based portal sites and PSEs. It provides key facilities for grid programming, such as user authentication, resource selection, job scheduling, job execution, monitoring, and collection of accounting information.
- WP-4: Packaging and Configuration Management (Miura from NII). This level coordinates with the first to provide autonomous configuration management as well as a testing infrastructure.
- WP-5: Networking, Security & User Management (Shimojo from Osaka University, Oie from Kyushu Tech., and Imase from Osaka University). This level performs traffic measurement on SuperSINET and optimal QoS Routing based on user policies and network measurements. It also has robust TCP/IP for grids.
- WP-6: Grid-enabling tools for Nanoscience Applications (Aoyagi from Kyushu University). This is the application-specific layer that is adapted for each new applications domain that is supported on the grid.

The grid middleware research in Japan is recognized internationally, and collaborates regularly on standards committees. One highlight from the AIST Grid effort is the Grid Datafarm, a project to build a parallel file system on top of the grid so that users can easily access their files from any location on the grid. A team of AIST researchers won a “bandwidth challenge” award at SC2002 using this Grid Datafarm infrastructure.

The Personal Power Plant (P3) is grid middleware that allows for sharing of resources, including computing power, storage and I/O. The model is similar to that of SETI@home, but uses two-way sharing so that participants may contribute and receive resources. A future goal is to integrate millions of personal information applications, such as PCs, PDAs, and mobile phones.

GRID APPLICATIONS

Many applications of grid computing technology are either complete or under active development in Japan. This includes a Quantum Chemistry Grid at AIST, which is a Problem Solving Environment for quantum chemistry. The system combines several applications (e.g., Gaussian and GAMESS) into a single environment, and optimizes performance of the system by keeping a database of performance information on past calculations. The system also serves as a portal to Gaussian, an electronic structure program, and GAMESS, a program for general *ab initio* quantum chemistry. Both Gaussian and GAMESS were developed elsewhere. AIST is also working on a weather forecasting system designed to predict short- and mid-term global climate change. Forecasting involves averaging across several independent predictions, and the volume of computational capacity in a grid environment allows a much larger number of these independent calculations to be done than on a single cluster. The weather forecasting grid runs on 11 clusters that are part of an international grid system called the Asia Pacific Grid (ApGrid). The system uses the Ninf-G middleware and can be used from a web interface that was built with the PSE Builder, as well as vector and parallel machines. There are several other active applications projects associated with application science,

such as a radioactive source estimation problem on the grid, a fusion grid, a grid for KEK. Some of these are described in Chapters 4 and 5 of this report.

One of the main challenges and opportunities being pursued by the Japanese grid project is that of heterogeneity. While heterogeneous computing is a challenge from the software perspective, it has potential advantages in performance and resource utilization by allowing different parts of a simulation to run on hardware best suited to that phase. This idea is key to many of the Japanese grid application projects. Figures 8.6 through 8.8 show some examples. The first two are from ITBL, and use a mixture of vector and scalar machines, using vectors where they are most effective and scalar processors elsewhere. Figure 8.8 shows a nanotechnology simulation that uses a combination of shared memory and clustering.

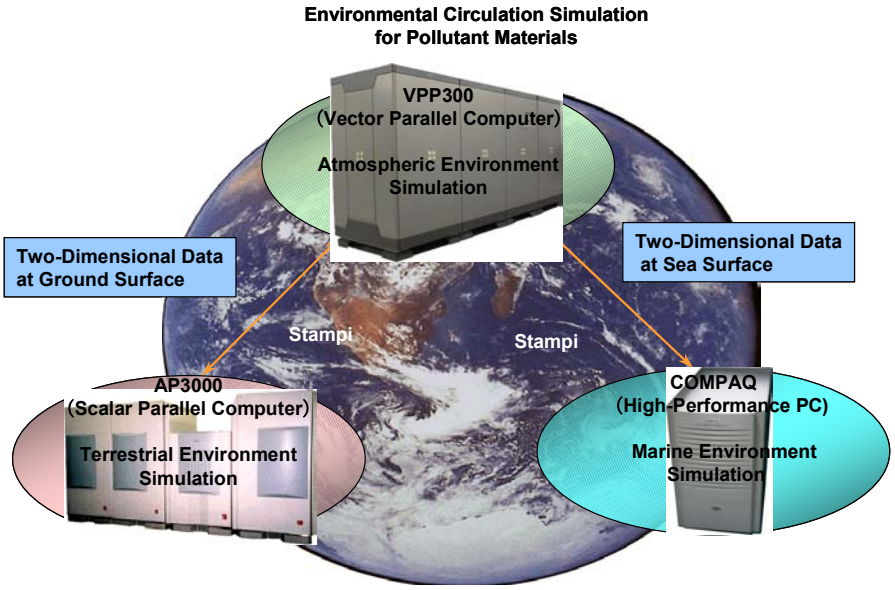


Figure 8.6. Using a heterogeneous grid for environmental circulation simulations (Courtesy ITBL)

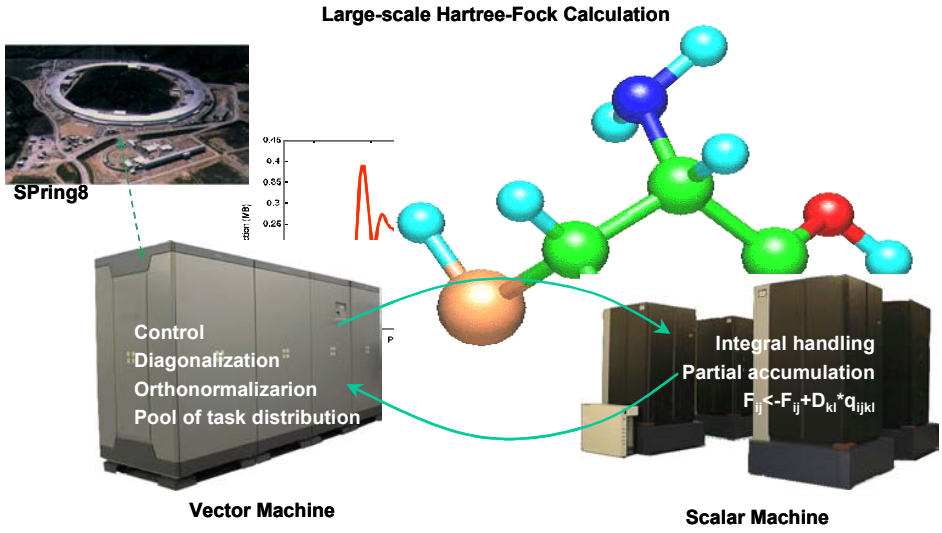


Figure 8.7. Use of a heterogeneous grid for Hartree-Fock calculation (Courtesy ITBL)

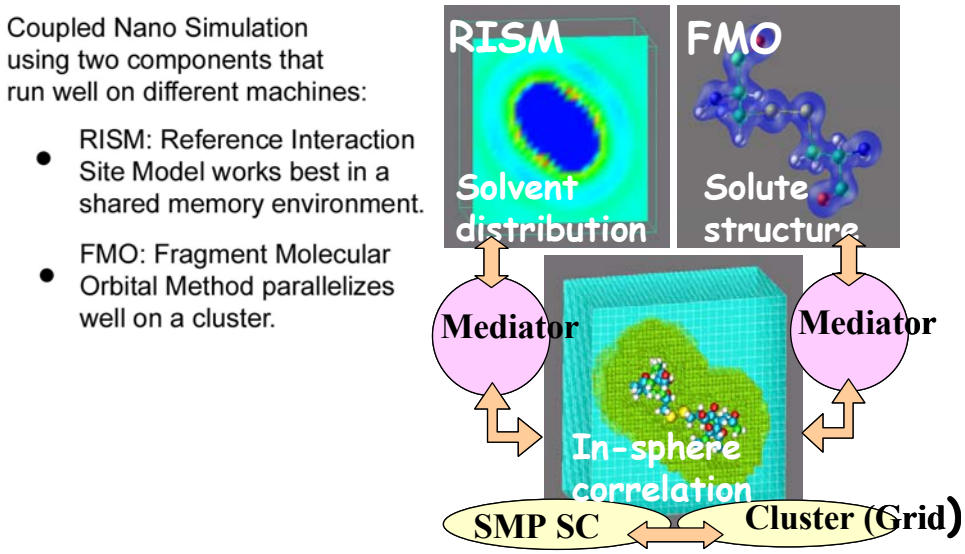


Figure 8.8. Use of a heterogeneous grid for nanotechnology simulation (Courtesy NAREGI)

Figure 8.8 shows a nanotechnology simulation that uses a combination of shared memory and clustering technology. Figure 8.9 is the most extreme case for heterogeneity, because it includes special-purpose Grape processors, vector processors, and parallel clusters.

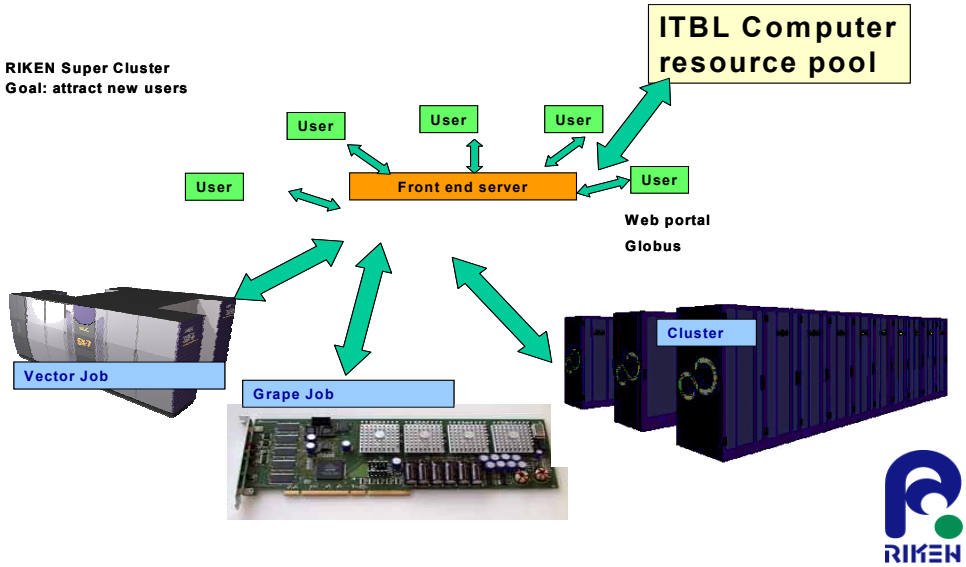


Figure 8.9. The RIKEN grid uses Grape boards, clusters, and vector processors (Courtesy RIKEN)

CONCLUSIONS

The grid computing program in Japan has a broad range of research topics for low-power hardware to heterogeneous grid-based applications. Grid computing is a high priority for government agencies, and the funding levels for grid programs are much greater than for vector supercomputing programs like the Earth Simulator. The effect of this shift on research institutions and vendors has been dramatic, as there is currently little or no research in supercomputing technology and tools, but a broad interdisciplinary effort in grids. Because the boundary between grid and cluster computing is somewhat blurred, some of the research in application-level libraries and Problem Solving Environments could easily be considered a high-performance

computing project, rather than a grid-computing project. However, the emphasis is on commodity processors and the development of software technologies that are applicable to business and government in addition to science and engineering.

The grid program has produced middleware tools that are used outside of Japan, and the area has been and should continue to be a fruitful area for international collaborations. As with earlier efforts in supercomputing hardware and software, the Japanese researchers are making a concerted effort to address all aspects of grid computing, and have a large group of application researchers from many different domains involved. This is somewhat surprising given the relative immaturity of grid programming tools, but reflects the kind of single-minded agenda that made the earlier supercomputing efforts successful.

One of the key technical challenges in a grid environment is heterogeneity, which affect all aspects of the programming environment from the availability of compilers and libraries down to storage formats for arrays and numeric types. The Japanese grid agenda highlights the performance heterogeneity of systems in the grid as an opportunity, allowing applications to select the best hardware for a given part of a computation. The most aggressive use of heterogeneity combines special purpose processors like Grape with vector supercomputers, PC clusters, and shared memory workstations. These ideas have also influenced plans for an international follow-on to the Earth Simulator, which has been described by Director Sato as having exactly this type of heterogeneous architecture.

REFERENCES

Information Technology Based Laboratory. 2004. <http://www.itbl.riken.go.jp/> Last accessed February 23, 2005.

Miura, K. 2003. “NAREGI Project Overview.” <http://www.naregi.org/papers/data/SC2003Miura.pdf> Last accessed February 23, 2005.

National Research Grid Initiative. <http://www.naregi.org/index_e.html> Last accessed February 23, 2005.

RIKEN Super Combined Cluster System. <http://accc.riken.jp/E/rscc_e/index_e.html> Last accessed February 23, 2005.

Science Information Network (SINET). <http://www.sinet.ad.jp/english/> Last accessed February 23, 2005.

APPENDIX A. PANELIST BIOGRAPHIES

Alvin W. Trivelpiece
(Panel Chair)

Since May 2000, Alvin Trivelpiece has been a consultant to Sandia National Laboratories.

From January 1989 through March 2000 he served as the Director of Oak Ridge National Laboratory (ORNL). In January 1996, he was appointed President of Lockheed Martin Energy Research Corporation, the managing and operating contractor for ORNL. At ORNL, he was responsible for programs that included applied research and engineering development in the support of the Department of Energy's fusion, fission, conservation, and fossil energy technology programs and basic scientific research in selected areas of the physical and life sciences. As director of ORNL, he supervised a staff of over 5,000 and managed a budget of more than $500 million.

Trivelpiece served as the executive officer of the American Association for the Advancement of Science (AAAS) from April 1987 to January 1989. As the executive officer of the country's leading general science organization, he was responsible for all of the Association's activities and programs and served as publisher of *Science*, the Association's weekly journal.

He came to the AAAS from the U.S. Department of Energy, where he served as the director of the Office of Energy Research from 1981 to 1987. From 1978 to 1981, Trivelpiece was corporate vice president at Science Applications, Inc., in La Jolla, California, and from 1976 to 1978 he was vice president for engineering and research at Maxwell Laboratories in San Diego, California.

Trivelpiece was a professor of physics at the University of Maryland from 1966 to 1976 and was a professor at the University of California, Berkeley, in the Department of Electrical Engineering from 1959 to 1966. While on leave from the University of Maryland, from 1973 to 1975, he served with the U.S. Atomic Energy Commission as assistant director for research in the Division of Controlled Thermonuclear Research.

A native Californian, he received his B.S. degree from California Polytechnic State University in 1953, and his Master's (in 1955) and Ph.D. degree (in 1958) from the California Institute of Technology.

He was a member of the Board of Directors of Bausch & Lomb, Inc from 1989 to 2001, and of Charles River Laboratories from 1992 to 1999. He was the Head of the "1986 U.S. Delegation on Peaceful Uses of Atomic Energy to the USSR." He was a member of the National Research Council's Committee on Science and Technology Policy Aspects of Selected Social and Economic Issues in Russia, and a member (2000 – 2002) of the National Academy of Sciences Committee on the Technical Aspects of the Comprehensive Nuclear Test Ban Treaty.

His research has focused on plasma physics, controlled thermonuclear research, and particle accelerators. He has been granted several patents on accelerators and microwave devices and is the author or co-author of many papers and two books.

He serves as an advisor to government agencies. He is a fellow of the AAAS, the American Physical Society, and the Institute of Electrical and Electronics Engineers, and is a member of the American Nuclear Society, the American Association of University Professors, Tau Beta Pi, and Sigma Xi.

Rupak Biswas

Rupak Biswas received his Bachelor of Science (Honors) in Physics (1982) and his Bachelor of Technology in Computer Science (1985), both from the University of Calcutta, India, and his Master's (1988) and Ph.D. (1991) in Computer Science from Rensselaer Polytechnic Institute, Troy, NY. He has been at the National Aeronautics and Space Administration (NASA) Ames Research Center, Moffett Field, CA, since 1991, and is currently a Senior Computer Scientist in the NASA Advanced Supercomputing Division. He is the Group Leader of about 25 scientists in the Algorithms, Tools, and Architectures Group that performs research in computer science technology for high-performance scientific computing. The group's goal is to advance the state-of-the-art in parallel and distributed computational performance for key NASA algorithms, applications, and workloads. He is also the Level 3 Manager for High-end Computing Research and the Level 2 Planning Lead for Advanced Computing Architectures and Technologies under the CICT Program of the NASA Office of Aerospace Technology.

He was awarded a NASA Excellence Award in 1993 for his work on developing an automatic mesh adaptation procedure for three-dimensional unstructured meshes for problems in computational fluid dynamics. His paper that compared and analyzed parallel performance of a dynamic irregular application using different programming paradigms on various architectural platforms won the Best Paper Award at the SC'99 conference. He was a co-recipient of the 2001 NASA Group Achievement Award as a member of the Information Power Grid Group, which made outstanding contributions to the Agency's mission.

He has published more than 100 technical papers in archival journals and major peer-reviewed conferences, given numerous invited talks, and edited several journal special issues. His research interests include parallel and distributed processing, high-end computing, innovative computer architectures, parallel adaptive finite element methods, performance evaluation and modeling, and helicopter aerodynamics and acoustics. He has also served as a Program Committee member for many national and international conferences, and was a member of the NITRD-sponsored High-End Computing Revitalization Task Force.

Jack Dongarra

Jack Dongarra received a Bachelor of Science in Mathematics from Chicago State University in 1972 and a Master's of Science in Computer Science from the Illinois Institute of Technology in 1973. He received his Ph.D. in Applied Mathematics from the University of New Mexico in 1980. He worked at the Argonne National Laboratory until 1989, becoming a senior scientist. He now holds an appointment as University Distinguished Professor of Computer Science in the Computer Science Department at the University of Tennessee and ranks as one of the Distinguished Research Staff in the Computer Science and Mathematics Division at Oak Ridge National Laboratory (ORNL), and is also an Adjunct Professor in the Computer Science Department at Rice University. He is the director of the Innovative Computing Laboratory at the University of Tennessee, which has a staff of 50 people doing research in the area of high-performance computing. He is also the director of the Center for Information Technology Research at the University of Tennessee, which coordinates and facilitates IT research efforts at the University.

Dongarra specializes in numerical algorithms in linear algebra, parallel computing, the use of advanced-computer architectures, programming methodology, and tools for parallel computers. His research includes the development, testing and documentation of high-quality mathematical software. He has contributed to the design and implementation of the following open source software packages and systems: EISPACK, LINPACK, the BLAS, LAPACK, ScaLAPACK, Netlib, PVM, MPI, NetSolve, Top500, ATLAS, and PAPI. He has published approximately 200 articles, papers, reports and technical memoranda and he is coauthor of several books. He is a Fellow of the AAAS, ACM, and the IEEE and a member of the National Academy of Engineering.

Peter Paul

Peter Paul has served as Brookhaven National Laboratory's Deputy Director for Science and Technology since March 1998, when Brookhaven Science Associates took over management of the Laboratory. From October 2001 until March 2003, Paul also served as Interim Laboratory Director during a crucial time in the development of several new initiatives and facilities at Brookhaven.

After Peter Paul received a Ph.D. in experimental nuclear physics from the University of Freiburg in 1959, he spent seven years at Stanford University (1960-7). He joined the faculty of Stony Brook University's (USB) Department of Physics in 1967. He became Distinguished Service Professor in 1992, and he served as Chair of the Physics Department from 1986 to 1990, and from 1996 to 1998. As Director of USB's Nuclear Structure facility, Paul developed and constructed the university's superconducting heavy ion linear accelerator. Paul has served as Brookhaven's Deputy Director for Science and Technology since March 1998, when BSA took over the Laboratory's management.

A Fellow of the American Physical Society, Paul won the Alexander von Humboldt Senior Scientist Award in 1983. He was a member of the DOE/National Science Foundation Nuclear Science Advisory Committee from 1980 to 1983 and served as chair of the committee from 1989 to 1992. He is the author of about 170 refereed articles in nuclear and accelerator science journals.

Katherine Yelick

Katherine Yelick is a professor in the EECS Department at the University of California at Berkeley. Her research in high-performance computing addresses parallel programming languages, compiler analyses for explicitly parallel code, and optimization techniques for communication and memory system architectures. Much of her work has addressed the problems of programming irregular applications on parallel machines. Her parallel language and compiler projects include the Split-C and UPC languages, which are parallel extensions of C, and the Titanium language, a high-performance scientific computing language based on Java. She also led the compiler effort for the Berkeley IRAM project, a single chip system that combines vector processing computing in a low-power Processor-in-Memory chip, and the Sparsity code generation system for automatic tuning of sparse matrix kernels. She currently leads the UPC team at Lawrence

Berkeley National Laboratory and co-leads the Titanium and BeBOP (Berkeley Benchmarking and Optimization) teams at the University of California, Berkeley.

Yelick received her Bachelor's, Master's, and Ph.D. degrees from the Massachusetts Institute of Technology, where she worked on parallel programming methods and automatic theorem proving. She won the George M. Sprowls Award for an outstanding Ph.D. dissertation at MIT.

APPENDIX B. SITE REPORTS

Site: **AIST-GRID: National Institute of Advanced Industrial Science and Technology, Grid Technology Research Center**
Tsukuba Central-2, 1-1-1 Umezono, Tsukuba,
Ibaraki 305-8568
Phone: +81-29-861-5877
http://www.gtrc.aist.go.jp/en/index.html

Date Visited: March 31, 2004

WTEC Attendees: K. Yelick (Report author), A. Trivelpiece, P. Paul, S. Meacham, Y.T. Chien

Hosts: Mr. Kunihiro Kitano, Deputy Director, International Division, AIST,
Dr. Satoshi Sekiguchi, Director, Grid Technology Research Center, AIST

OVERVIEW

The National Institute for Advanced Industrial Science and Technology (AIST) has been in existence for 150 years, but it was the Agency of Industrial Science and Technology prior to 2001. In 2001 it was made an Independent Administrative Institution (IAI), and gained its current name. As an IAI, it receives funding from the government to the institution as a whole and is in charge of managing its own budget, rather than having the government fund individual projects within the institution. The current president of AIST can be hired from industry or other areas, while under the old model the president had to have come from the government. The conversion of government institutions to IAIs was a frequent topic of WTEC visits to other institutions, because the universities were being converted during the week of the group's visit, but AIST had gone through this conversion process three years earlier.

The reason for the reorganization was to reduce the number of government officials. It is not clear whether this actually happened, because people like Mr. Kitano are still government employees even though he works for an independent institution. AIST is now conceptually independent of the government, and they can set their own agendas. They expect a synergistic affect and are better able to do interdisciplinary work.

There are three types of research organizations in AIST: *research institutes*, which are permanent organizations that perform mid- to long-term research in which research ideas are proposed in a bottom-up manner; *research centers*, which are temporary (less than seven-year) organizations that are spun off from the research institutes, operated by top-down management, and designed to conduct pioneering or strategic research; *research laboratories,* which are used to promote the formation of new research fields or research centers. There are 21 research institutes in AIST, 32 research centers, and eight research laboratories. There was some discussion of traditional Japanese society, in which seniority is very important, and the research institutes follow this model with the head of the institution selected based on seniority. In a research center, on the other hand, anyone with a good idea can write a proposal and possibly work on it, so quality of ideas drives the funding and status within the organization.

The old AIST had 15 research institutes, eight in Tsukuba and seven elsewhere. All of them have been reorganized as part of the new AIST, a single Independent Administrative Institution with nine sites throughout Japan.

RESEARCH DIRECTIONS

The most challenging and highest level goals for AIST to perform are the kind of long-term research that requires government support, and that is meant to enhance international competitiveness, create new industry, and provide technology infrastructure and maintenance. The three major research directions are:

1. *Environment and Energy*: This includes research in chemical risk management, power electronics, photoreaction control, explosion and safety, environmental management, green technology, and energy utilization. It does not include nuclear energy research.
2. *Resources Measurement and Standard Geo-Science*: This includes research in deep geological environments, active fault research, marine resources and environments, and metrology.
3. *Life Science, IT, Materials/Nano, and Manufacturing Machinery*: Specific research projects in the Life sciences include: Computational Biology, Bio Informatics, Tissue Engineering, Bioelectronics, and Genomics. In Information Technology, it includes research in: Advanced Semiconductors, Near-Field Optics, Cyber Assist, Digital Human, and Grid Technology. In the area of material science and nanotechnology, projects include: Advanced Carbon Materials, Macromolecular Technology, Synergy Materials, Smart Structures, Nanoarchitectonics, Advanced Nanoscale Manufacturing, Digital Manufacturing, and Diamonds.

AIST employs roughly nine thousand people, divided into permanent staff, visitors, and students, as shown in the table below. The breakdown of staff into research areas (as of April 2003) is: 23% in Environment, 22% in Nanotechnology, 18% in Information Technology, 13% in Life Sciences, 12% in Geosciences, and 11% in Measurements and Standards.

Table B.1
AIST Personnel Numbers

Staff	Tenured researchers	2,073
	Fixed-term researchers	302
	Part-time technical staff	1,182
	Administrative staff	718
	Part-time administrative assistants	617
	Total staff	4,892
Visiting Researchers	Postdoctoral researchers (Domestic and Overseas)	465
	Researchers from private sector	1,999
	Overseas researchers	637
Students from University		900
Total personnel		8,893

The budget for AIST is predominantly government funded from the subsidy as an IAI, but they also acquire some direct research funds from METI, MEXT and other government agencies, in addition to some funding from industry. In FY2003, the total budget was ¥92 billion (B), with ¥68B from the subsidy, ¥18B from commissioned research, ¥4B from a facilities management grant, and ¥2B from other miscellaneous support.

The reward system within AIST includes both publications and patents. AIST encourages the formation of startup companies to pursue commercialization of technologies developed at AIST. AIST does not provide any of the startup funding directly, but will help in finding funding sources. Intellectual property agreements are made up for specific projects, since there are many researchers at AIST who are visiting from industry. The default IP is that AIST keeps ownership, but in practice most of the time an exclusive license is given to the company.

GRID TECHNOLOGY RESEARCH CENTER

AIST has a large effort in grid computing at the GTRC, founded in January 2002. Dr. Satoshi Sekiguchi is the director of this center. He gave a presentation of GTRC and led much of the discussion. There are six areas of research at GTRC: programming tools, high-speed networking, compute clusters as computer

resources on the grid, international demonstration and verification experiments, application demonstrations, and practical issues in grid deployment including security and reliability.

One of the challenges in grid computing is making them usable by application programmers. GTRC has several projects within the programming tools area.

- Ninf-G is programming middleware that provides a remote procedure call (RPC) mechanism across the grid. Ninf-G is a reference implementation of the GridRPC specification, an interface that is part of an international standardization effort for grid middleware. Ninf-G is also compatible with the Globus Toolkit, developed in the U.S.
- Grid MPI is a new internal design of an implementation of the Message Passing Interface (MPI), which is commonly used in writing large-scale parallel applications. By providing an implementation of MPI in a grid environment, application programmers with MPI-based software are more likely to experiment with a grid environment.
- Grid PSE Builder is a software environment that can be used to build web services, grid-based portal sites and Problem Solving Environments (PSEs). It provides key facilities for grid programming, such as user authentication, resource selection, job scheduling, job execution, monitoring, and collection of accounting information.
- Grid Datafarm is a project to build a parallel file system on top of the grid so that users can easily access their files from any location on the grid. A team of AIST researchers won a "bandwidth challenge" award at SC2002 using this Grid Datafarm infrastructure.
- The Personal Power Plant (P3) is grid middleware that allows for sharing of resources, including computing power, storage and I/O. The model is similar to that of SETI@home, but uses two-way sharing so that participants may contribute and receive resources. A future goal is to integrate millions of personal information applications, such as PCs, PDAs, and mobile phones.

AIST houses a large computing resource called the Supercluster in a large new building that was not yet occupied (except by the Supercluster) at the time of WTEC's visit. The purpose of the Supercluster is to provide a testbed for grid computing research. One of the research issues with grid computing is how to design software that works in a heterogeneous environment, spreading computations or storage across a system with a mixture of hardware, systems software, and performance characteristics. The Supercluster was intentionally designed to be heterogeneous, mixing processor types, networks, clock rates, and memory capacity, so software designed for the full system will naturally be heterogeneous.

A picture of the Supercluster is shown in Figure 8.2. It has a theoretical hardware peak of 14.6 Tflop/s and contains three clusters:

- *P32*: An 8.6 Tflop/s IBM eServer325 is a set of 128 dual processor Opteron nodes with Myrinet 2000 interconnect. The processors have a 2.0 GHz clock and 6GB of memory each.
- *M64*: A 2.7 Tflop/s Intel Tiger 4 cluster with 131 quad processor Itanium nodes with Myrinet 2000 interconnect. The processors are 1.3GHz Itanium (Madison) processors, each with 16GB of memory.
- *F32*: A 3.1 Tflop/s Linux Networx cluster with 256 dual Xeon nodes connected by Gigabit Ethernet. The processors are 3.06 GHz Xeons with 2GB of memory.

The cost to build this Supercluster was approximately $20M, and it has an ongoing maintenance cost of about $1.4M annually.

Several applications of grid computing technology are under active development. These include a Quantum Chemistry Grid, which is a problem-solving environment for quantum chemistry. The system combines several applications (e.g., Gaussian and GAMESS) into a single environment, and optimizes performance of the system by keeping a database of performance information on past calculations. The system also serves as a portal to Gaussian, an electronic structure program, and GAMESS, a program for general *ab initio* quantum chemistry. Both Gaussian and GAMESS were developed elsewhere.

A second grid application described by Dr. Sekiguchi is a weather forecasting system designed to predict short- and mid-term global climate change. Forecasting involves averaging across several independent predictions, and the volume of computational capacity in a grid environment allows a much larger number of these independent calculations to be done than on a single cluster. The weather forecasting grid runs on 11 clusters that are part of an international grid system called the Asia Pacific Grid (ApGrid). The system uses the Ninf-G middleware and can be used from a web interface that was built with the PSE Builder.

Dr. Sekiguchi referred to applications in biology and medicine, including REXMC, a replica exchange Monte Carlo method used for molecular dynamics simulations. REXMC uses two levels of parallelism: at the coarse level, the system generates multiple copies of molecules and assigns a random temperature to each, with infrequent communication between processors to exchange temperatures; within each molecular simulation a parallel *ab initio* calculation is done.

OBSERVATIONS

AIST/GTRC has an impressive program in grid computing, especially in the area of grid middleware. Software such as Ninf-G appears to widely used throughout the grid community. The hardware resources were also impressive, with a huge price performance advantage over a custom system like the Earth Simulator. However, this particular Supercluster was very new, and there were not yet large calculations running against which one could compare effective performance of the two classes of systems. GTRC also has a significant involvement in applications of grid computing, which are mostly collaborative efforts that leverage projects both within Japan and in the international community.

Site: **Council for Science and Technology Policy (CSTP)**
Cabinet Office, Government of Japan
3-1-1 Kasumigaseki, Chiyoda-ku, Tokyo
http://www8.cao.go.jp/cstp/english/s&tmain-e.html

Date Visited: March 30, 2004

WTEC Attendees: Y.T. Chien (Report author), J. Dongarra, S. Meacham, A. Trivelpiece, K. Yelick

Hosts: Dr. Hiroyuki Abe, Council member,
Mr. Masanobu Oyama, Council member,
Mr. Shigeyuki Kubota, Counselor (Information & Telecommunications Policies)

BACKGROUND

For the past decade, Japan has been undertaking major efforts in streamlining government, with the goal of making the various ministries and their constituencies work more efficiently and effectively. In the area of science and technology, these efforts are reflected in the enactment of the Science and Technology Basic Law in 1995 [1] and its subsequent adoption and implementation of the basic plans [2]. These actions represent some of the most visible signs of Japan's reform and restructuring of science and technology in recent years.

Behind these actions, and central to their successful outcomes, is the new Council for Science and Technology Policy (CSTP). Established in January 2001 within the Cabinet Office, the CSTP acts as an advisory to the government, but has become the *de facto* policy maker and implementer for S&T plans and programs. The Council is chaired by the Prime Minister and consists of seven appointed Executive Members from the science and engineering communities, the Chief Cabinet secretary, five cabinet members who head the four ministries related to Science and Technology and the Ministry of Finance (MOF), and the President of the Science Council of Japan. The CSTP has many functions, aimed at steering Japan's S&T into a more competitive position in the world. Its influence in S&T is best reflected in at least two important ways. First, it develops five-year S&T plans serving as the foundation and goal posts for the various ministries and their programs. Second, by working with the powerful Ministry of Finance, it initiates the S&T budgeting process and makes recommendations on funding priorities, which are often endorsed by the Prime Minister and the Diet. In the words of a Japan S&T observer [6], implementing such a strategic approach to a Japanese S&T budget would have been exceedingly difficult, if not impossible, prior to the creation of the CSTP. In many ways, it is now a key ingredient of Japan's goal towards a coherent and effective S&T policy and its implementation, long sought after by government leaders and the general public.Our visit to the CSTP was graciously hosted by two of the Executive members, Dr. Abe (former President of Tohoku University) and Mr. Oyama (former Senior Executive Vice President and Director of Toshiba Corporation). It provided a rare opportunity for the panel to exchanges views on S&T matters in general, and high-end computing (HEC) in particular, with two of the top leaders in Japan.

GENERAL FUNCTIONS AND ACCOMPLISHMENTS

Prior to our visit, the panel submitted to CSTP a set of questions designed to help the hosts understand the panel's primary areas of interest and the main issues of concern to the WTEC's study project. These questions were roughly divided into two parts. The first deal with the roles of CSTP in Japanese government, how those roles are being fulfilled, and the major impact of CSTP on Japanese S&T so far. The second group of questions was related to the specific area of high-end computing and its relationship to information technology and other high-priority areas of Japan's second basic plan (2001-2005). Dr. Abe led the discussion by addressing the issues raised in the first group of questions.

Using a set of handouts, Dr. Abe explained in some detail the mission and the general organization of the CSTP. He emphasized that one of the functions of CSTP is to develop strategic S&T plans for the Cabinet Office as part of the annual and long-term budgeting processes. The CSTP is instrumental in setting program priorities for the Prime Minister and the National Diet. For example, in the current fiscal year, the top priority areas are life sciences, nanotechnology and materials, environmental sciences, and information and communications technology. Next to this group is a set of secondary priorities in several areas such as marine science, manufacturing, etc. These priorities help guide the individual ministries to develop and submit their budgets for consideration by the Cabinet. Based on these submissions, the CSTP makes overall prioritization of the projects, using a four-level ranking scale (S-A-B-C, from the highest to the lowest) for funding recommendations to the Ministry of Finance. While MOF still negotiates with each ministry on its budget, CSTP's recommendations are largely followed and eventually accepted by the Diet almost always in their entirety.

Asked by the panel by what means CSTP evaluates these projects in determining their significance and priority, Dr. Abe indicated that the Council meets regularly; monthly on policy issues and weekly on other matters. It also has seven "expert panels" for technical evaluations with members drawn from the external scientific and engineering communities. Dr. Abe was more modest in addressing the accomplishments and impacts of the Council. However, one of the handouts [3] reveals a more detailed picture of how the CSTP functions to "act as a control tower to implement S&T policy," "steer S&T with foresight and mobility," and "integrate S&T with humanities and society." Working with the ministries and the external communities, the CSTP in its short existence has issued a large number of strategic documents aimed at promoting S&T. In the current fiscal year, as a result of CSTP's initiatives, the budget bills passed by the Diet followed its recommendations and the R&D expenditure in general account increased by 4.4%, a much larger increase compared to that of the total general account (which increased by only 0.1%).

Information Technology and High-End Computing

Mr. Oyama then led a discussion that addressed the second group of our questions, concerning CSTP's roles in high-end computing. He first pointed out that HEC is not an isolated field from the rest of information technology (IT). From that viewpoint CSTP feels that there are three important areas in the broad IT context at this juncture: highly reliable IT, human interface, and quantum computing. Supercomputing and HEC networks (along with broadband Internet and low-power device technology) certainly rank among the highest priorities for future research and development. Currently, R&D efforts in HEC are directed towards both hardware and software, including middleware. The two key agencies for such efforts are MEXT (Ministry of Education, Culture, Sports, Science, and Technology) and METI (Ministry of Economy, Trade, and Industry). For example, METI and MEXT are funding business grid computing for ¥2.5 billion and high-speed grid computing for ¥2 billion, respectively, in FY04. MEXT's Earth Simulator (ES) project funding is at ¥5.9 billion this year.

Asked by the panel whether the CSTP has a policy or plan in place (or in the works) for the future of Japan's HEC or the next generation Earth Simulator (ES), Mr. Oyama commented that such policy or plan is usually developed from the bottom up, namely from the ministries such as MEXT and METI and their constituencies. CSTP's role is again to develop a consensus vision, set the tone for the future, and issue guidelines for the budget process when it comes to that. Mr. Oyama further commented that the need for a next generation ES will have to be demonstrated in many different applications (e.g., biology, manufacturing, etc.) that require ultra high-end computing (at the petaflop/s level) for their further progress. He said that currently the CSTP is working with the R&D communities to develop new ideas for the next generation of ES, but nothing is finalized. Grid computing, clusters, and possibly a combination of different architectures are all among the candidates for future considerations.

OBSERVATIONS

It is probably not an exaggeration to suggest that CSTP is a very influential office in the Japanese S&T hierarchy. This influence is best evidenced in the area of budgetary development and control across ministries, which is the bread and butter for agencies in any national government. In a budget document [4]

given to the panel during our visit, there is a complete list of new project initiatives for FY04 submitted by various S&T ministries, with their designations of priorities as ranked by CSTP (S-A-B-C) in October 2003. CSTP's prioritized list has since been forwarded to the Ministry of Finance for funding mark-up. Of the some 200 major projects in all categories, MOF agreed with all but one or two of CSTP's recommendations, according to a follow-up document provided by the NSF/Tokyo Office in February 2004 [5]. This level of policy coherence is hardly achievable unless there is a highly effective working model and coordination process firmly in place.

In addressing our questions concerning high-end computing, both Dr. Abe and Mr. Oyama were cautious in predicting the direction of future developments in Japan. They were both quick to point out that any possible plan for a next-generation initiative would have to come from the ministries. However, both were also hinting that CSTP is developing new ideas with inputs from the research community, a slow, deliberate process that is not obvious to us, but has apparently worked before in the context of the Earth Simulator.

Finally, our visit cannot escape the temptation to draw a superficial comparison between Japan's CSTP and the counterparts in the United States. In many ways, the CSTP assumes many of the functions that also exist in the National Science and Technology Council (NSTC), the Office of Science and Technology Policy (OSTP), and the National Coordination Offices (NCOs for special initiatives) in the U.S. Smaller and perhaps more disciplined, the CSTP is more influential on budgets as noted before. As a relatively new establishment, the CSTP has undoubtedly drawn on the many experiences from its counterparts, especially the NCOs on interagency coordination. The challenges are similar and some solutions are clearly transferable across national borders.

REFERENCES

Blanpied, W. 2003. "The Second Science and Technology Basic Plan: a Blueprint for Japan's Science and Technology Policy." NSF/Tokyo special scientific report 03-02.

"Council for Science and Technology Policy," Publication of the Cabinet Office, Government of Japan, 2004.

"JFY2004 S&T-related budgets by projects," NSF/Tokyo Report Memorandum 04-03, February 20, 2004 (English translation).

"Rating of S&T-related projects – JFY 2004" (Japanese version); NSF/Tokyo Report Memorandum 03-10, October 2003 (English translation).

"The Science and Technology Basic Law (unofficial translation)." <http://www8.cao.go.jp/cstp/english/law.html> Last accessed February 23, 2005.

"The Science and Technology Basic Plan (2001 – 2005)," Government of Japan publication (unofficial translation), March 31, 2001.

Site: **Earth Simulator Center**
Japan Marine Science and Technology Center (JAMSTEC)
3173-25 Showa-machi, Kanazawa-ku
Yokohama-City, Kanagawa 236-0001, Japan
http://www.es.jamstec.go.jp/

Date Visited: March 29, 2004

WTEC Attendees: K. Yelick (Report author), R. Biswas, A. Trivelpiece, S. Meacham, Y.T. Chien

Hosts: Dr. Kunihiko Watanabe, Simulation Science & Technology Research Program, Program Director
Hiroshi Hirano, Executive Assistant to the Director-General
Toshiaki Shimada, System Technology Group and System Development Group, Group Leader
Shigemune Kitawaki, Program Development and Support Group, Group Leader

BACKGROUND

Dr. Watanabe began with an overview of the Earth Simulator Research and Development Center, which was developed as part of an effort to understand earth systems, in particular climate and earthquakes. The Center was established in 1997, the year that the Kyoto Protocol was adopted by the United Nations Framework Convention on Climate Change. Prior to this, in 1995, the Intergovernmental Panel on Climate Change (IPCC) had published a report stating that if global warning continued at its current rate through the 21st century, the average temperature would rise by about 2 degrees Celsius, and the sea level would rise by about 50 cm. It was also in 1995 that Japan was hit with the Kobe Earthquake, which gave citizens a real interest in understanding natural disasters. The Center was established by three organizations---the Japan Marine Science and Technology Center, the National Space Development Agency of Japan, and the Japan Atomic Energy Research Institute, and it was funded by the former Science and Technology Agency of Japan, which has since been reorganized and renamed to the Ministry of Education, Culture, Sports, Science and Technology (MEXT). The Earth Simulator Center was established in 2001 for the Operation of the Earth Simulator (ES) and for promoting research activities using ES.

MANAGEMENT OF THE EARTH SIMULATOR CENTER

The Earth Simulator Center is managed by the Director General, Dr. Tetsuya Sato, along with two committees, a Selection Committee and a Mission Definition Committee. The Mission Definition Committee helps set the general direction of the Center. The committee is made up of about 20 people, most of who are from outside the Center, such as University faculty and reporters. The Selection Committee determines which scientist may use the Earth Simulator (ES) and for how long. This committee is made up of about 10 people, most of whom are scientists.

The maintenance costs on ES are about $30M per year, with about $7M for power. The machine consumes 6-7 MWatts. In response to a question about upgrading the system in place, for example, upgrading the processors to SX-7 nodes, the response was that there were no plans for a simple upgrade, and no government funds were available for such an activity. The researchers at the center did talk about the possibility of a new machine, but did not give any details.

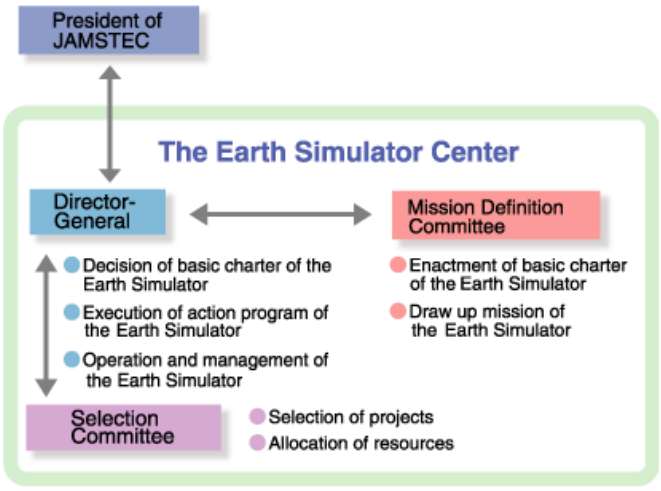


Figure B.1. The Earth Simulator Management (Source: Annual Report of the Earth Simulator Center)

CONSTRUCTION OF THE EARTH SIMULATOR

Dr. Hajime Miyoshi started the Earth Simulator Research and Development Center (ESRDC) in 1997 and led the development of the Earth Simulator (ES). Dr. Miyoshi had a vision of the machine architecture, its software, and its use in applications, and oversaw many aspects of the design and construction, but he passed away shortly before the machine was completed.

Construction began in 1999 and was completed in early 2002. Four months, starting from February 2001, was spent on the work of laying 83,200 network cables to connect the various processors. The total length of cables came to about 2,400 km, which would be enough to connect the northernmost and southernmost tips of Japan. The delivery and installation of 320 node cabinets (each of which would store two computer nodes) and 65 cabinets for storing the interconnection network devices began in the summer of 2001; final adjustments and test operations were conducted from the last half of 2001 into 2002. Figure 2.1 in Chapter 2 summarizes the construction schedule of the system.

HARDWARE OVERVIEW

The Earth Simulator was built by NEC and uses a combination of parallel and vector processing technologies. The system has 640 processing nodes, each of which contains eight arithmetic processors, giving a total of 5120 arithmetic processors. Each arithmetic processor has a peak performance of 8 Gflop/s, giving an overall system peak of 40 Tflop/s, which was five times faster than the next fastest general purpose computer at the time it was completed. Each processing node contains 16 GBytes of memory, given the entire system 10 TBytes. The network is a full crossbar, with a peak network bi-directional bandwidth of 12.3 GBytes/sec. Figure 2.4 in Chapter 2 shows the overall system design.

The clock rate on the processors is 500 MHz. There is a scalar processor that is a four-way superscalar with a peak of 1 Gflop/s, and a vector processor with a peak of 8 Gflop/s. Each vector processor has 32 GBytes/sec of memory bandwidth, while the scalar processors have only 4 GBytes/sec. Each scalar processor has a data cache and instruction cache that are each 64 Kbytes.

ES is housed in a 65 by 50 meter building, which is significantly smaller than was expected during the initial planning stages for the machine. The computer system building contains storage and computational nodes on the top floor, with 2 computational nodes per cabinet, a floor containing the crossbar interconnect cables, and a lower floor with air conditioning units and power supplies. The entire system consumes about 7 million watts at a cost of about $7 million per year.

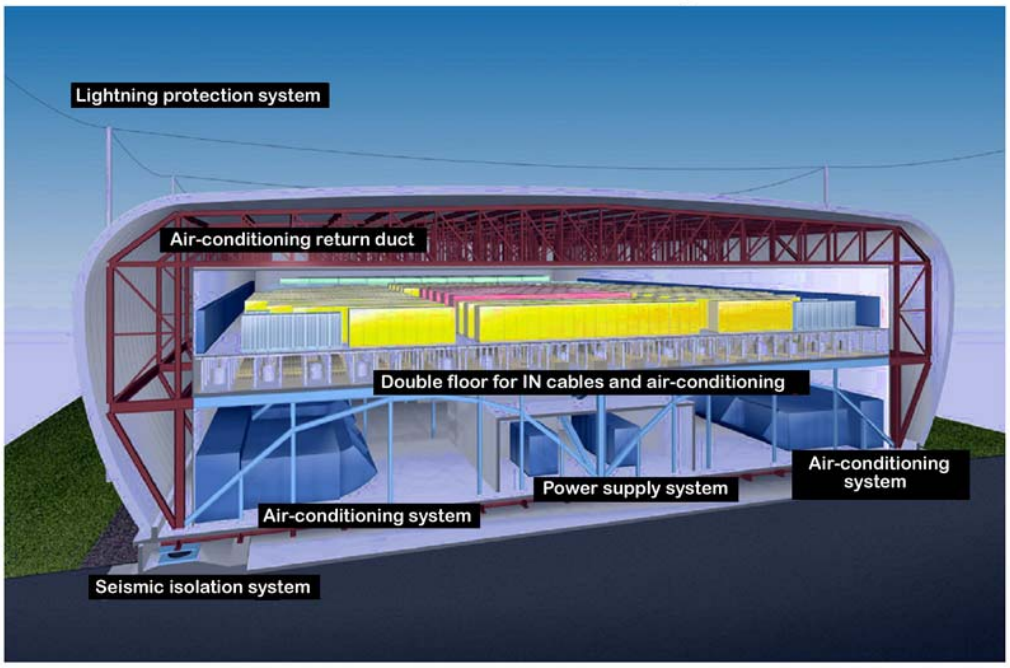


Figure B.2. The ES facility. (Courtesy the Earth Simulator Center)

The machine room and building have several features designed to protect the machine. There is a seismic isolation system to reduce the effect of earthquakes and three layers of Faraday shields to protect the machine from lightening and other electrostatic discharges: 1) Faraday shield of the outer walls by metallic plates; 2) Faraday shield of the computer room by metallic gauze (meshes); 3) Faraday shield of the double floor by metallic gauze and plates. There is an additional lightning protection system with groundings that independent of the shields and lighting within the machine room is done by light tubes to reduce electrical noise.

SOFTWARE OVERVIEW

The system software on ES was provided entirely by NEC. This includes the operating system and compilers. The machine has several programming models available to users, including message passing (MPI) or High-Performance Fortran (HPF) between processing nodes. The MPI implementation has a latency of 5.6 microseconds and achieves a maximum bandwidth of 11.8 GBytes/sec, out of a hardware peak of 12.3 GBytes/sec. Within a processing node, the eight arithmetic processors communicate by shared memory, and the parallelism can be expressed using MPI, HPF or OpenMP, or the parallelization may be left to an automatic parallelizing compiler, as can be seen in Figure 2.9 in Chapter 2.

During the WTEC visit there was no presentation specifically addressing programming languages or libraries. There were few staff members who supported programming and optimization (including vectorization, parallelization) and also promoted the use of HPF to the scientists using ES. The expectation among the hosts was that the next system after ES would look at languages and OS as well as hardware and algorithms.

USING THE EARTH SIMULATOR

The system is not available for remote login through the Internet, so users have to travel to the Center, which is 40 km south of Tokyo, to use the machine. There are plans to add a 2.5 Gb/s connection for external users, which would connect to the national Super SCINet network, in October 2004. There are also strict requirements on the use of the machine when running with a large number of processors. Only after

demonstrating adequate vectorization and parallel scalability are users given an opportunity to run on a large number of processors. Specifically, 95% execution time must be vector code and the parallel efficiency must be at least 50% before the node count can be increased.

There are about 200 users of ES split among over 30 projects. The complete list of projects is included at the end of this report. About 30-35% of machine time is allocated to ocean and atmospheric modeling, while 15-20% is for solid earth modeling and another 15% is for very large and novel "Epoch-making" simulations. Computer Science research uses about 10%, and the Director General's discretionary allocation is 20-30%. A complete list of Japanese projects is shown below. The non-discretionary projects are competed once per year, and in the past year there was some consolidation of projects. For example, within the turbulence community there had been several research teams vying for time on the machine, and whichever group received the allocation had a big advantage over the other groups in publishing papers. This community has been has now been consolidated into a single allocation.

International users are all involved in collaborative projects, which make up part of the Director's discretionary account. There are eight international collaborations from five different countries. Although the machine can only be used while visiting the Center, the data can be taken back to home institutions.

APPLICATIONS

There are several large applications running on ES. One of the first results on the machine was a simulation of the AFES climate model, which scales to ~28 Tflop/s. The following figure shows the results for a simulation of precipitation using a T639L24 mesh, which corresponds to 640*3 mesh points around the equator and 24 vertical levels.

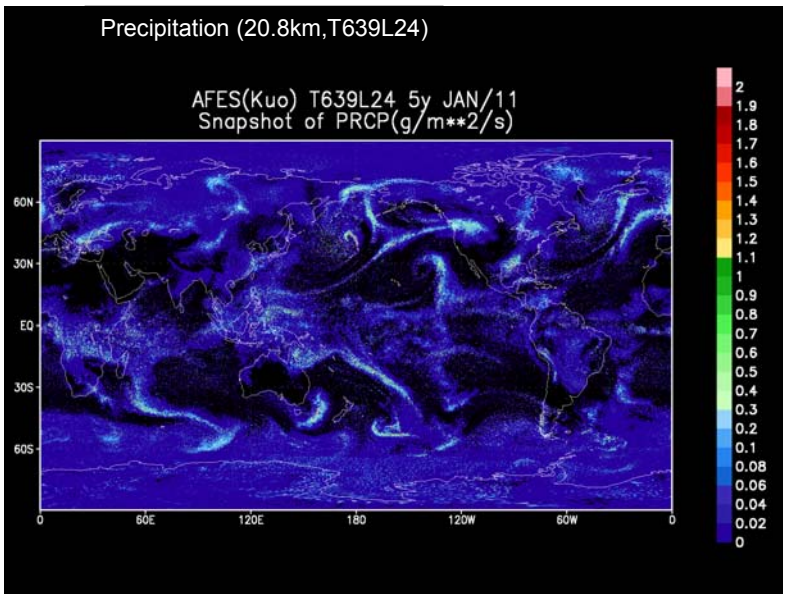


Figure B.3. Results for a simulation of precipitation using a T639L24 mesh.

Another application area with increasing emphasis is nanotechnology. This includes the design of innovative nano-materials with desired properties, such as a super hard diamond with a Jungle Gym structure, and discovering thermal and mechanical properties of nano-scale materials. The simulation of a 200 nm carbon nanotube with 40K atoms, runs at 7.1 Tflop/s on 3480 CPUs of ES, for example.

Most of the application efforts on ES are done by, or in collaboration with, researchers outside the Earth Simulator Center. The Frontier Research System for Global Change (FRSGC), which is located in an attached building, is responsible for much of the climate research, and the Research Organization for Information Science and Technology (RIST) performs applications research nanotechnology, quantum

chemistry, nuclear fusion, as well as additional work on climate modeling. There are also some projects on computational biology on ES. Although there are computational needs, large scale simulation codes large enough to take advantage of ES have not been developed yet.

OBSERVATIONS

ES (both machine and facility) is an impressive feat of engineering design and implementation. The machine has been shown to get high performance on a large set of applications in climate modeling, nano science, earthquake modeling and other areas. The allocation policy ensures that there are no significant negative results on the machine performance, since only problems that vectorize and scale well are allowed to use a large configuration. The main scientific result enabled by ES has been the ability to simulate larger problems with finer meshes in a variety of areas. These can be difficult to quantify because of the differences in algorithms and software used on different machines. Within the climate modeling community there is a detailed plan for running on ES and other machines, which includes a comparison of running times. The data, available from: http://www.cgd.ucar.edu/ccr/ipcc/, gives the following comparisons:

- *NERSC IBM SP3*: 1 simulated year per compute day on 112 processors
- *IBM SP4*: 1.8 - 2 simulated years per compute day on 96 processors
- *IBM SP4*: 3 simulated years per compute day on 192 processors
- *ES*: 40 simulated years per compute day (number of processors not given)

There were no specific plans for a follow-on machine discussed during the visit, and no funding for a more modest upgrade of the current machine. However, subsequent talks by Director Sato have described plans for an international facility with orders of magnitude more computational power than the ES.

PROJECTS ON ES

This list does not include international collaborations, which are part of the Director's discretionary account.

Table B.2
Projects on the Earth Simulator

Ocean and Atmosphere (12)		
Development of Super High-Resolution Atmospheric and Oceanic General Circulation Models on Quasi-Uniform Grids	Yukio Tanaka	JAMSTEC, FRSGC
Atmospheric Composition Change and its Climate Impact Studied by Global and Regional Chemical Transport Models	Hajime Akimoto	JAMSTEC, FRSGC
Development of High-Resolution Cloud-resolving Regional Model and its Application to Research on Mesoscale Meteorology	Fujio Kimura	JAMSTEC,FRSGC
Process Studies and Seasonal Prediction Experiment Using Coupled General Circulation Model	Toshio Yamagata	JAMSTEC, FRSGC
Future Climate Change Projection using a High-Resolution Coupled General Circulation Model	Akimasa Sumi	Center for Climate System Research, University of Tokyo

Table B.2 Projects on the Earth Simulator		
Development of High-Resolution Atmosphere-Ocean couple model and Global Warming Prediction	Kouki Maruyama	Central Research Institute of Electric Power Industry
Research on Development of Global Climate Model with Remarkably High Resolution and Climate Model with Cloud Resolution	Takashi Aoki	Japanese Meteorological Research Institute
Research Development of 4-Domensional Data Assimilation System using a Coupled Climate Model and Construction of Reanalysis Datasets for Initialization	Toshiyuki Awaji	JAMSTEC, FRSGC
Development of Integrated Earth System Model for Global Change Prediction	Taro Matsuno	JAMSTEC, FRSGC
Parameterization of Turbulent Diffusivity in the Deep Ocean	Toshiyuki Hibiya	Graduate School of Science, University of Tokyo
Mechanism and Predictability of Atmospheric and Oceanic Variations Induced by Interactions Between Large-Scale Field and Meso-Scale Phenomenon	Wataru Ofuchi	JAMSTEC, Earth Simulator Center
Development of Holistic Simulation Codes on Non-Hydrostatic Atmosphere-Ocean Coupled System	Keiko Takahashi	JAMSTEC, Earth Simulator Center
Solid Earth (9)		
Global Elastic Response Simulation	Sejii Tsuboi	JAMSTEC, IFREE
Simulation Study on the Generation and Distortion Process of the Geomagnetic Field in Earth-Like Conditions	Yozo Hamano	JAMSTEC, IFREE, Graduate School of Science, University of Tokyo
Numerical Simulation of the Mantel Convection	Yoshio Fukao	JAMSTEC, IFREE
Predictive Simulation of Crustal Activity in and Around Japan	Mitsuhiro Matsuura	Graduate School of Science, University of Tokyo
Numerical Simulation of Seismic Wave Propagation and Strong Ground Motions in 3D Heterogeneous Media	Takashi Furumura	Earthquake Research Institute University of Tokyo
Simulation of Earthquake Generation Process in a Complex System of Faults	Kazuro Hirahara	Graduate School of Environmental Studies, Nagoya University
Development of Solid Earth Simulation Platform	Hiroshi Okuda	Graduate School of Engineering, University of Tokyo

Table B.2
Projects on the Earth Simulator

Simulator Experiments of Physical Properties of Earth's Materials	Mitsuhiro Toriumi	JAMSTEC, Earth Simulator Center
Dynamics of Core-Mantle Coupled System	Akira Kageyama	JAMSTEC, Earth Simulator Center
Computer Science (2)		
Design and Implementation of Parallel Numerical Computing Library for Multi-Node Environment of the Earth Simulator	Ken'ichi Itakura	JAMSTEC, Earth Simulator Center
Performance Evaluation of Large-scale Parallel Simulation Codes and Designing new Language Features on the HPF (High Performance Fortran) Data-Parallel Programming Environment	Yasuo Okabe	Academic Center for Computing and Media Studies, Kyoto University
Epoch-Making Simulation (11)		
Numerical Simulation of Rocket Engine Internal Flows	Hiroshi Miyajima	NASDA
Large-scale Simulation on the Properties of Carbon-Nanotube	Kazuo Minami	RIST
Development of the Next-Generation Computational Solid Mechanics Simulator for a Virtual Demonstration Test	Ryuji Shioya	Graduate School of Engineering Kyushu University
Study of the Standard Model of Elementary Particles on the Lattice with the Earth Simulator	Akira Ukawa	Center of Computational Physics, University of Tsukuba
Large-Scale Simulation for a Tera Hz Resonance Superconductor Device	Masashi Tachiki	National Institute for Material Science
Geospace Environment Simulator	Yoshiharu Omura	Kyotot University
Particle Modeling for Complex Multi-Phase Systems with Internal Structures Using DEM	Hide Sakaguchi	RIST
Development of Transferable Materials Information and Knowledge Base for Computational Materials Science	Shuhei Ohnishi	Collaborative Activities for Material Science Programs (CAMP) Group
Cosmic Structure Formation and Dynamics	Ryoji Matsumoto	Chiba University
Bio-Simulation	Nobuhiro Go	Forum on the Bio-Simulation
Large Scale Simulation on Atomic Research	Hiroshi Okuda	Atomic Energy Society of

Table B.2
Projects on the Earth Simulator

		Japan
Sub-Theme Under Large-Scale Simulation of Atomic Research (9)		
Large-Scale Numerical Simulations on Complicated Thermal-Hydraulics in Nuclear Cores with Direct Analysis Methods	Kazuyaki Takase	JAERI
First Principles Molecular Dynamics Simulation of Solution	Masaru Hirata	JAERI
Direction Numerical Simulations of Fundamental Turbulent Flows with the Largest Grid Numbers in the World and its Application to Modeling for Engineering Turbulent Flows	Chuichi Arakawa	Center for Promotion of Computational Science and Engineering (CCSE), JAERI
Research on Structure Formation of Plasmas Dominated by Hierarchical Dynamics	Yasuaki Kishimoto	JAERI
Large-Scale Parallel Fluid Simulations for Spellation Type Mercury Target Adopted in the Project of High-Intensity Proton Accelerator	Chuichi Arakawa	CCSE, JAERI
Studies for Novel Superconducting Properties and Neutron Detection Applications by Superconductor Nano-Fabrication Techniques	Masahiko Machida	JAERI
Electronic and Atomistic Simulations on the Irradiation Induced Property Changes and Fracture in Materials	Hideo Kaburaki	JAERI
Large-Scale Simulations on the Irradiation Induced Property Changes and Fracture in Materials	Hiroshi Okuda	Graduate School of Engineering, University of Tokyo
First-Principles Molecular Dynamics Simulation of Oxide Layers for Radiation Tolerant SiC Devices	Atumi Miyashita	JAERI

REFERENCES

Annual Report of the Earth Simulator Center, Outline of the Earth Simulator Project, <http://www.es.jamstec.go.jp/esc/images/annualreport2003/pdf/outline/outline.pdf> Last accessed February 23, 2005.

Habata, S., M. Yokokawa, S. Kitawaki, Shigemune. 2003. The Earth Simulator System. *NEC Res. & Develop.*, Vol. 44, No. 1. <http://www.owlnet.rice.edu/~elec526/handouts/papers/earth-sim-nec.pdf> Last accessed February 23, 2005.

JAMSTEC, Earth Simulator Hardware. <http://www.es.jamstec.go.jp/esc/eng/ES/hardware.html> Last accessed February 23, 2005.

Sumi, A. 2003. "The Earth Simulator and Its Impact for Numerical Modeling." <http://www.tokyo.rist.or.jp/rist/workshop/rome/ex-abstract/sumi.pdf> Last accessed Febrary 23, 2005.

Site: **Frontier Research System for Global Change (FRSGC)**
JAMSTEC Yokohama Research Institute for Earth and Sciences
3173-25 Showa-machi, Kanazawa-ku, Yokohama-City, Kanagawa 236-0001
http://www.jamstec.go.jp/frsgc

Date Visited: March 29, 2004

WTEC Attendees: R. Biswas (Report author), A. Trivelpiece, K. Yelick, S. Meacham, Y.T. Chien

Hosts: Taroh Matsuno, Director-General,
Hirofumi Tomita,
Michio Kawamiya,
Eri Ota, Assistant to Dr. Matsuno

BACKGROUND

Under the former Science and Technology Agency (STA) of the Japanese Government, the Subcommittee on Earth Science and Technology published in July 1996 a review titled "Toward the Realization of Global Change Prediction." The report highlighted the need to integrate research and development using one system, incorporating process research, observations, and simulations. Based on this review and the "Common Agenda for Cooperation in Global Perspective," former U.S. Vice President Al Gore and former Japanese Prime Minister Ryutaro Hashimoto agreed in early 1997 to form the International Arctic Research Center (IARC) and the International Pacific Research Center (IPRC) as U.S.-Japan centers for cooperation in global change research and prediction. In October 1997, the Frontier Research System for Global Change (FRSGC) was established as a joint project between the National Space Development Agency (NASDA) and the Japan Marine Science and Technology Center (JAMSTEC) to implement process research to meet the goal of "Prediction of Global Change." At that time, IPRC at the University of Hawaii and IARC at the University of Alaska were also established.

In mid-2001, FRSGC was relocated from Tokyo to Yokohama to be closer to the Earth Simulator. In April 2003, FRSGC's management was shifted solely to JAMSTEC, six months before NASDA was merged with the Institute of Space and Astronautical Science (ISAS) and the National Aerospace Laboratory (NAL) to form the Japan Aerospace Exploration Agency (JAXA). With the beginning of their new fiscal year on April 1, 2004, JAMSTEC was merged with the Ocean Research Institute of the University of Tokyo to form an independent administrative institution called the Japan Agency for Marine-Earth Science and Technology. This new organization is currently under the jurisdiction of the Ministry of Education, Culture, Sports, Science, and Technology (MEXT).

We were hosted by Taroh Matsuno, the FRSGC Director-General. Incidentally, Dr. Matsuno was the Chairperson of the STA Subcommittee whose review back in 1996 initiated the process of establishing FRSGC. It is a 20-year project, divided into two 10-year phases. Interim evaluation is carried out every five years by reporting progress, and indicating a future research direction. Currently, FRSGC has about 180 researchers, of which a third are invited domain experts from overseas. In FY2002, their budget was ¥31.5 billion.

FRSGC PRESENTATIONS

FRSGC Director-General Taroh Matsuno first gave an overview of the organization's current activities. He began by observing that our planet has recently begun to be affected in significant ways by various human activities (e.g. El Niño, greenhouse gas effects, disappearance of tropical rainforests). In addition, society remains vulnerable to natural disasters such as earthquakes, volcanic eruptions, and abnormal weather. He stressed that FRSGC's primary objective was to contribute to society by elucidating the mechanisms of various global natural changes as well as making better predictions of such changes. With this goal in mind, researchers at FRSGC are developing and simulating high-fidelity models of the atmosphere, ocean, and

land. These individual elements will subsequently be integrated to model Earth as a single system, and therefore be able to make reliable predictions of various phenomena on our home planet.

FRSGC's activities are classified into six research programs: Climate Variations, Hydrological Cycle, Atmospheric Composition, Ecosystem Change, Global Warming, and Integrated Modeling. Each of these areas is further subdivided into more focused research groups.

Hirofumi Tomita presented his group's work on the development of a new global cloud-resolving model using icosahedral grids. The development of a dynamical core of the 3D global model is complete. This model, called NICAM (Non-hydrostatic Icosahedral Atmospheric Model), requires 1.5 hours per simulation day on 2560 processors of the Earth Simulator (ES) with a 3.5 km grid. NICAM will be integrated into FRSGC's next-generation atmospheric general circulation model (AGCM) for climate studies. The goal is to run AGCM with super-high resolutions such as 5 km or less in the horizontal directions and 100 m in the vertical direction. The code for NICAM is written in F90, uses MPI, and has been developed and performance-tuned by researchers over the last couple of years. Comparisons with AFES, a global atmospheric circulation code based on the spectral model that was optimized for the ES, are extremely promising.

The next presentation was given by Michio Kawamiya on the development of an integrated Earth system model for global warming prediction. Here, biological and chemical processes important for the global environment interact with climate changes. The integrated model adds individual component models (such as oceanic carbon cycle) to atmospheric and ocean general circulation models. Eventually, the model will include atmosphere (climate, aerosol, chemistry), ocean (climate, biogeochemistry), and land (climate, biogeochemistry, vegetation). Fine-resolution atmospheric models will be able to reproduce meso-scale convective systems explicitly, while high-resolution ocean models will be capable of accurately reproducing ocean eddies.

REMARKS

The research work being conducted by FRSGC is very impressive. They have a large coordinated effort to understand global natural changes, and then to make better predictions of such changes. They have an excellent visitor program whereby experts from outside Japan visit FRSGC and work with resident scientists. Their international cooperation extends to the U.S., Australia, Canada, and the European Union. Under a cooperative agreement, information is exchanged with the International Research Institute for Climate Prediction in the U.S. The proximity to the Earth Simulator is an added benefit.

Site: **Fujitsu Headquarters**
Shiodome City Center Bldg.
1-5-2 Higashi-Shimbashi, Minato-ku, Tokyo 105-7123
http://us.fujitsu.com/home/

Date visited: March 31, 2004

WTEC Attendees: J. Dongarra (Report author), R. Biswas, M. Miyahara

Hosts: Motoi Okuda, General Manager Computational Science and Engineering Center,
Email: m.okuda@jp.fujitsu.com
Takayuki Hoshiya, Project Manager Software
Ken Miura, Fellow
Koh Hotta, Director, Core Technologies Dept. Software (compiler project)
Shi-ichi Ichikawa, Director HPC, Computational Science and Engineering
Yuji Oinaga, Chief Scientist, Server Systems

PRESENTATION

Motoi Okuda, General Manager Computational Science and Engineering Center presented the current state of Fujitsu's high performance computing efforts. He described the platform, technology, the latest sites using high-performance computing, and their vision for the future. Please refer to Chapter 6 for detailed background information on Fujitsu's past and current high-performance computing efforts.

Today the National Aerospace Laboratory of Japan has a 2304 processor Primepower HPC2500 system based on the Sparc 1.3 GHz. This is the only Fujitsu computer on the Top500 list that goes over 1 Tflop/s. The figure below illustrates the architecture of the Primepower HPC2500.

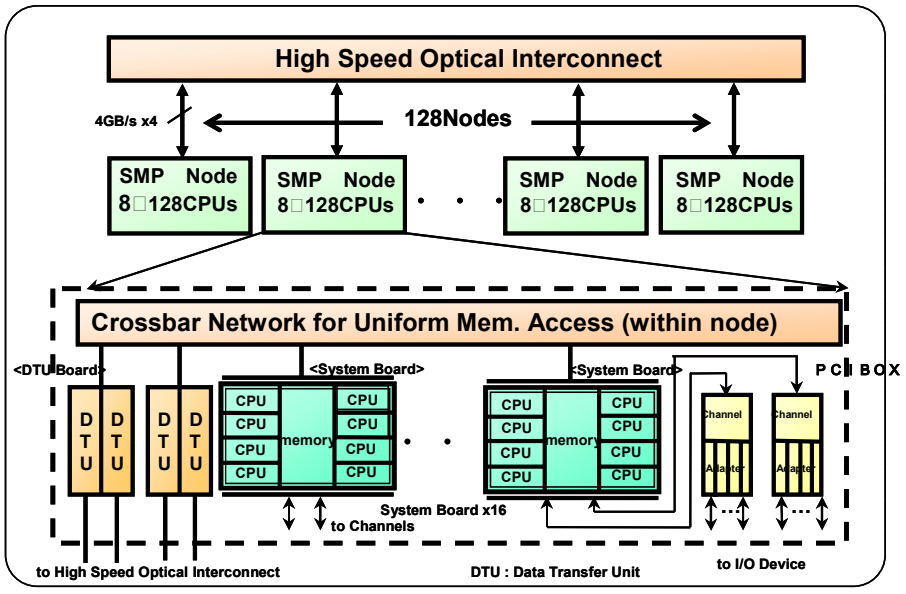


Figure B.4. PrimePower HPC2500: Architecture (Courtesy Fujitsu)

The Fujitsu VPP system (vector architecture) had a 300 MHz clock and as a result had weak scalar performance compared to commodity processors. The VPP saw 30% peak performance on average for applications, while the Primepower sees about 10% peak performance on average. The difference can easily be made up in the cost of the systems. The VPP was 10 times the cost of the Primepower system. Future versions of the HPC2500 will use the new Sparc chip 2 GHz by the end of the year.

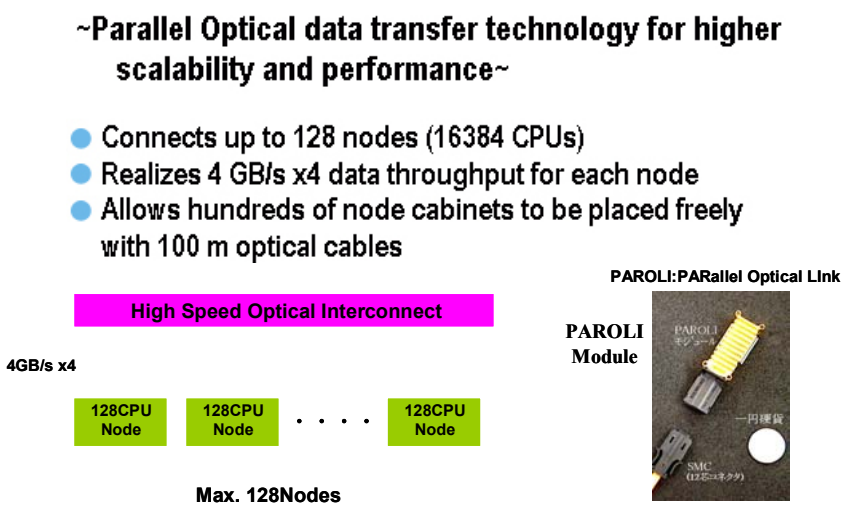


Figure B.5. PrimePower HPC2500: Interconnect (Courtesy Fujitsu)

REMARKS

In many respects this machine is very similar to the SUN Fire 3800-15K. The processors are 64-bit Fujitsu implementations of SUN's SPARC processors, called SPARC 64 V processors, and they are completely compatible with SUN products. Also the interconnection of the processors in the Primepower systems is like the one in the Fire 3800-15K: a crossbar that connects all processors at the same footing, i.e., it is *not* a NUMA machine. The figures below and above illustrate additional features of the Primepower HPC2500.

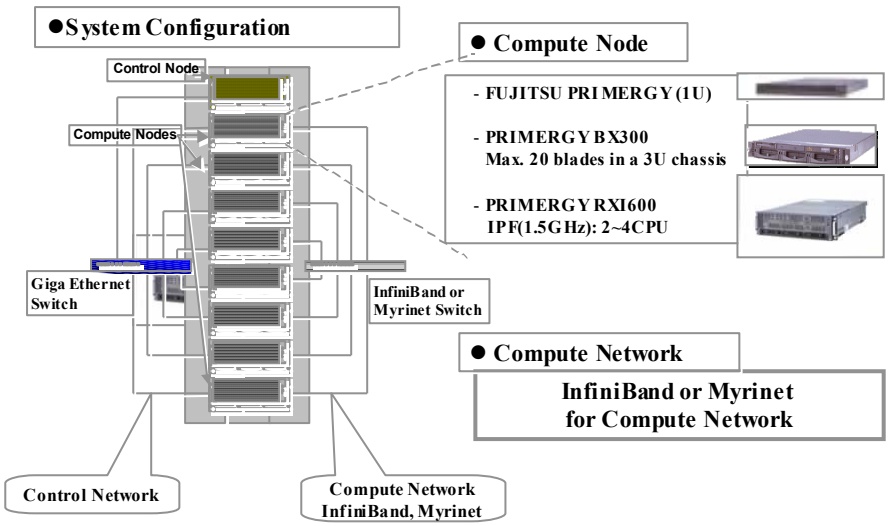


Figure B.6. IA-Cluster: System Configuration (Courtesy Fujitsu)

Site: **Hitachi, Ltd.**
Harmonious Center of Competency
Shinagawa East One Tower 13F, 2-16-1 Kounan,
Minato-Ku, Tokyo, 108-0075 Japan
http://www.hitachi.co.jp/Prod/comp/hpc/index.html

Date visited: April 1, 2004

WTEC Attendees: J. Dongarra (Report author), R. Biswas, K. Yelick, M. Miyahara

Hosts: Yasuhiro Inagami, General Manager HPC Business, Enterprise Server Division
Yoshiro Aihara, Senior Manager HPC, Enterprise Server Division,
Satomi Hasegawa, Senior Engineer HPC, Enterprise Server Division
Fujio Fujita, Chief Engineer OS division Department, Software Division
Nobuhiro Ioki, Deputy Department Manager Language Processor Dept., Software Division
Naonobu Sukegawa, Senior Researcher Platform Systems, Central Research Laboratory

BACKGROUND

Hitachi was founded 1910. Today it has a total of 340,000 employees, $68B in sales and $85B in assets.

Some important areas for Hitachi are:

- Power and Industrial systems
- Electronics Devices
- Digital Media and Consumer Products
- Info and Tele Systems: 19% of revenue

Hitachi has six corporate labs in Japan; five in USA; four in Europe. The figure below illustrates the development of high-performance computing at Hitachi.

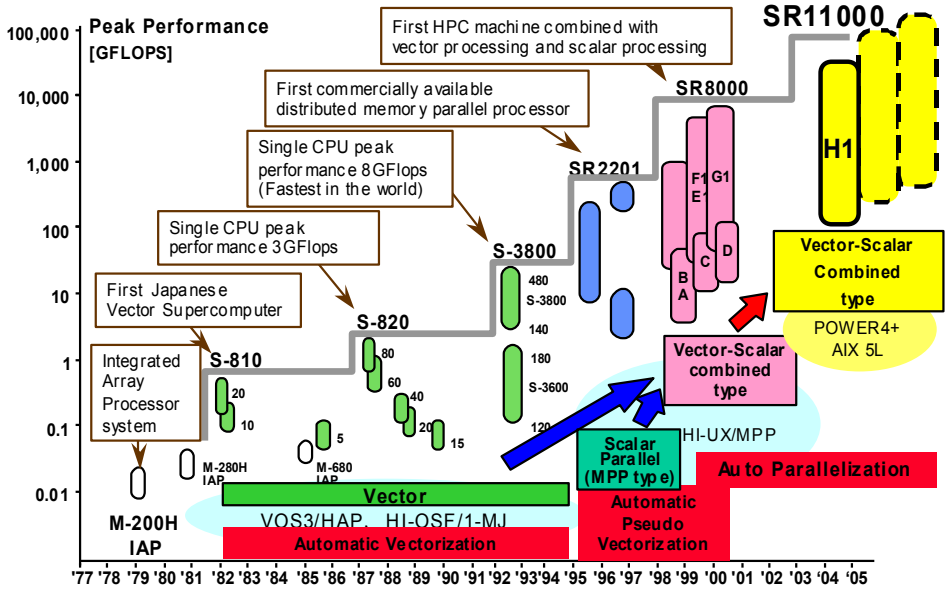


Figure B.7. Hitachi's HPC System (Today based on IBM Components) (Courtesy Hitachi)

The SR8000 is the third generation of distributed-memory parallel systems at Hitachi. It is to replace both its direct predecessor, the SR2201 and the late top-vector processor, the S-3800. Figure 7.3 in Chapter 7 illustrates the architecture of parallel vector processing (PVP).

The Super Technical Server SR11000 Model H1 can have between four and 256 nodes, each of which is equipped with 16 - 1.7GHz IBM Power4+ processors and achieves a theoretical computation performance of 108.8Gflop/s per node, about four times the performance of its predecessor SR8000 Series. The architecture is very similar to the SR8000.

- 16-way SMP node
- 256 MB cache per processor
- High memory bandwidth SMP
- PVP equipped
- COMPAS for providing parallelization of loops within a node
- High-speed internode network
- AIX operating system
- No hardware preload for compiler
- No hardware control for barrier
- Nodes connected by IBM's High-Performance Switch™
- 2 to 6 links per processor (or planes)
- AIX with cluster system management

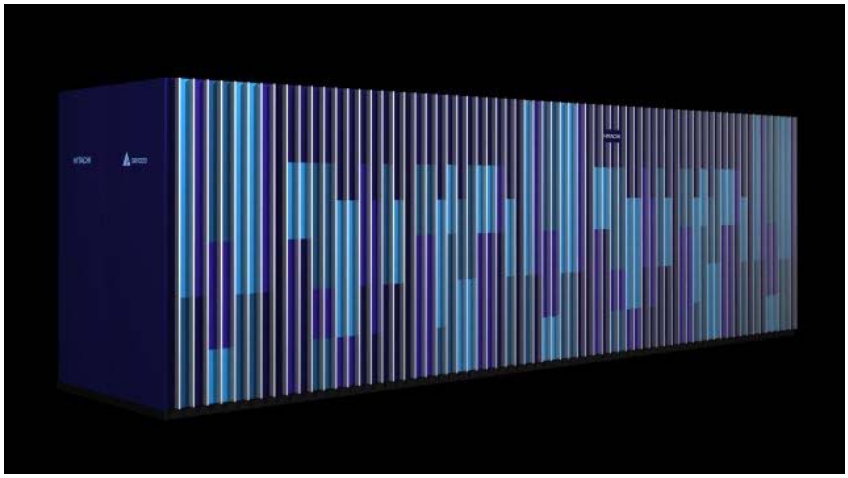


Figure B.8. High performance/area /power: 109Gflop/s node, Multi-stage Crossbar Network (4GB/s, 8GB/s, 12GB/s), A cabinet with 1.0m x 1.5m (41"x61") footprint contains 8 nodes (870Gflop/s), 30kW/cabinet (32kVA @3phase-200VAC); Flexible system configuration: From 4 to 256 nodes (0.4 to 28 Tflop/s); High-density packaging. (Courtesy Hitachi)

Table B.3
SR11000 Model H1 Specifications

		SR11000 model H1						
System	**Number of nodes**	**4**	**8**	**16**	**32**	**64**	**128**	**256**
	Peak performance	435GF	870GF	1.74TF	3.48TF	6.96TF	13.9TF	27.8TF
	Maximum total	256GB	512GB	1TB	2TB	4TB	8TB	16TB
	Inter-node Transfer speed	4GB/s (in each direction) x 2						
		8GB/s (in each direction) x 2						
		12GB/s (in each direction) x 2						
	External interface	Ultra SCS13, Fibre Channel (2Gbps), GB-Ether						
Node	Peak performance	108.8Gflop/s						
	Memory capacity	32GB/64GB						
	Maximum	8GB/s						

FUTURE PLANS

At this point they have three customers for the SR 11000, at 7 Tflop/s, the largest system with 64 nodes

- Institute for Molecular Science (50 nodes)
- National Institute Material Science
- Institute of Statistical Mathematics (part of MEXT) (4 nodes)

The University of Tokyo, Japan Meteorological Agency, and other similar agencies and organizations have a long history of using Hitachi machines.

Site: **IBM**
IBM-Japan Headquarters,
3-2-12 Roppongi, Minato-ku,
Tokyo 106-8711
JAPAN
http://www.ibm.com/jp/

Date Visited: April 1, 2004

WTEC Attendees: K. Yelick (Report author), J. Dongarra, R. Biswas, M. Miyahara

Host: Mr. Gary L. Lancaster, Vice President, pSeries, IBM Asia Pacific
Dr. Kazuo Iwano, Director, Emerging Business, IBM Japan, Ltd.
Dr. Hikaru Samukawa, Researcher, IBM Japan, Ltd.
Mr. Motoyuki Suzuki, Manager, Governmental Programs, IBM Japan, Ltd.

BACKGROUND

IBM has a substantial presence in Asia, with around 20% of its revenue on the continent and of that, around 70% is in Japan. One of IBM's largest research labs is in Japan, and there are large development labs in Yamato in Kanagawa close to the Tokyo area. There are about 20,000 employees in Japan.

The WTEC committee's visit to IBM was by necessity and in matter of fact different from most of the other visits. In particular, there were no formal presentations by IBM about their research or product lines in high-end computing. The committee members are all familiar with the strategies and technologies from IBM through executive and technical contacts and briefings in the United States. Instead, the discussion provided the visiting team with some perspective on the views on the high-end computing market from within Japan.

THE JAPANESE MARKET

There are several partnerships between American and Japanese vendors in high-end computing and in other markets. IBM has mainly partnered with Hitachi, HP with NEC, and Sun with Fujitsu. This makes the U.S.-developed products more attractive within the Japanese market and allows the Japanese vendors to offer a broader product line than would be possible on their own.

In the view of the IBM persons we met, the high-end computing field has been through a period of technical continuity, where dollars per megaflop dominated other issues, but it was about to go through regeneration as evidenced by machines like IBM's BlueGene/L. They also referred to the "Power Everywhere" meeting that had taken place the previous day, which announced IBM's plans for a more open hardware design model for the Power series processors, which allows processors to be specialized or reconfigured to a particular application domain. The program involves some new and existing partnerships, including Sony, Hitachi, and Nintendo, and was described as a new model that was much more promising than a "return to vectors." Another point is that the Power PC 970 with VMX is provided with a vector instruction set.

Price performance continues to be a major concern in the market, and pushes users toward more widely used architectures. The automobile manufactures, in particular, have moved away from vector machines towards lower-cost clusters. As another example of a system purchased for low cost, IBM cited the Weta Digital cluster system, which is used for graphics processing, often for special effects in movies. (The largest of these are two 1755-processor clusters with Pentium processors and gigabit ethernet; they run Linux along with Pixar's Renderman software and are at positions 43 and 47 on the Top500 list.) The operation-based cluster at a national university in Japan getting on the Top500 is another excellent example. Interest continues across Asia in machines using IBM's Power architecture including examples at the Korea Institute for Science and Technology (KISIT), number 45 on the Top500. There is also interest in the Asian high-end

computing market in machines like BlueGene/L, although production of the machine has not ramped up to the point where it can be sold in large numbers.

GRID COMPUTING IN JAPAN

Supercomputing is not the dominant driving force in the Japanese technology marketplace, whereas e-business and e-government, driven by improvements in pervasive digital communication, are. The interest is in grid computing, with the focus on digital communications driven by computing, and on using technology to "bring government, industry, and citizens closer together." R&D funding from the Japanese government is generally in everyday computing rather than high-end computing. On the computational side, the vision is to have parallel compute engines for hundreds to millions of users. These would be backend servers that provide integrated software systems, not just hardware.

THE EARTH SIMULATOR

There was some discussion on the role and significance of the Earth Simulator (ES) within Japan. Although there was a feeling that the cost was too high, the system also had significant PR value. The two key applications of the ES were climate modeling and earthquakes, with earthquake modeling being of perhaps even higher interest for the average Japanese citizen. The future public vision, for Tokyo for example, is to transition from a horizontal to a vertical landscape in order to provide a more viable and enjoyable urban living environment. However, Japan regularly experiences earthquakes. Therefore, the general public has a direct understanding of the need for a scientific understanding of the behavior of the Earth both from micro and macro points of view. Also, ES can be justified against a catastrophic event, such as earthquakes in Japan (or California) or terrorism or nuclear weapons safety in the U.S. The ES system is viewed as an isolated system with very specific applications, not a general service to the high-end market as a whole. It was suggested that an alternative approach of investing an equal amount of money in more peak Mflop/s using superscalar processors with a focus on improving software technology would yield a machine with a higher impact on both science and technology. Popular engineering software systems in automotive design, such as Gaussian and NASTRAN, are available on machines like the ES. However, their use is moving rapidly toward commodity clusters. Investments in improving the software tools required to improve massively parallel computing would have had the added benefit of making the ES project more widely beneficial to the economy at large.

OBSERVATIONS

While some of the analysis is specific to IBM's view of high-end computing technology, this visit highlighted some of the key aspects of the HPC in Japan. In particular, there is much more interest in grid computing than in supercomputing in Japan, which started a year or two after interest in grid in the U.S. picked up. Major investments like the ES require a driving application to sell the idea to the public. For ES it was earth systems modeling, with earthquakes and climate being the two key application areas. IBM's plans for processors follows IBM's strong commitment to open standards, parallel computing and the belief that applications, software and total systems technology are important attributes for advancing high-end computing. Successful high-end products will be built around Power processors that are adaptable or reconfigurable to support a wide variety of high-performance applications with a single processor line.

Site: **Japanese Atomic Energy Research Institute (JAERI) in Tokai**
2-4 Shirakata shirane, Tokai-mura
Naga-gun Ibaraki-ken 319-1195
http://www.jaeri.go.jp/english/

Date Visited: April 2, 2004

WTEC Attendees: P. Paul (Report author), A. Trivelpiece, S. Meacham

Hosts: Dr. Yoshiaki Kato, Execute Director at JAERI-Tokai
Dr. Toshio Hirayama, Deputy Director of Center for Promotion of Computational Science and Engineering (CSSE) at Tokai and group leader of R&D in computer science
Dr. Norihiro Nakajima, Group Leader for computer science R&D

BACKGROUND

JAERI is a very large and broadly based research organization focused on nuclear energy through fission reactors and fusion. Its main site at Tokai (the focus of this visit) is dedicated to nuclear energy systems, materials science, environmental science and nuclear safety. It is currently the construction site of the billion dollar J-PARC project (60 GeV proton synchrotron: spallation neutron source, source of copious beams of muons, kaons and neutrinos) to be completed by 2008. The Naka Fusion Research site, which works on Tokamak fusion, is host for the the JT-60 Tokamak and has the lead on the Japanese component of the ITER project. The Nakasaki Radiation Chemistry Research site and the Oarai site work on reactor technology; at the Kansai Research site JAERI works on photon sources. HAERI also operates, together with RIKEN, the Spring-8 synchrotron radiation facility. The MUTSU site operates a nuclear ship. JAERI has a total staff of about 2,400, with a total budget of approximately $1.2 billion.

JAERI has a Center for Promotion of Computational Science & Engineering (CCSE) and, separately, an Earth Simulator R&D Center. CCSE is headquartered in Tokyo/Ueno.

The focus of this visit was CSSE in Tokai and at the Tokyo Headquarters, and the Earth Simulator Center. At the Tokai site we were given a broad overview of the applications of CCSE and ES capabilities to science and engineering problems.

THE CCSE

Dr. Nakajima gave an overview of the Center for the Promotion of Computational Science and Engineering. Its headquarters is at the Tokyo site. Its annual budget is ~$50 million. Its mission is "to establish basic technologies for processing advance[d] parallel computing and advance[d] scientific and technological frontiers." In 2003 the demand for computing time was 80% for advanced photon simulations (see later) and nuclear fusion, with less than 10% going to computing science (CS). It is projected that in 2008 almost 50% will go to fusion, and 25% each to photon simulations and energy systems, while a tiny fraction will go to CS.

The computing power of JAERI is widely distributed: The major (>100 processors) computers are an SR8000F (160 CPU, general purpose) at Tokai, an Origin 3800 (769 CPU, scalar parallel) at Naka, an SC/ES40 (908 CPU scalar parallel) and a Prime Power (576 CPU, scalar parallel, dedicated to the ITBL) at Kansai. These are all connected to the Tokai site. Kansai, Tokyo and Tokai are scheduled to be connected to the SuperSINET with 1Gbps each. JAERI has a large program in computations for various fields using its supercomputers and the ES, some of the work done by CSSE staff and some by departments of JAERI research departments (see below). JAERI has already used 300 CPUs of the ES.

Dr. Nakajima presented the CCSE participation in the ITBL (IT Based Laboratory) national project. ITBL ties in with SuperSINET and the e-Japan project (103 subprojects), which aims to make Japan the world's largest IT nation by 2005. ITBL is a grand design with the goal of increasing efficiency by sharing resources (computers, staff expertise) and increasing performance (combinations of computers, parallel and pseudo scalable systems, shared expertise). The ITBL project has 523 users at 30 institutions, sharing the computer resources of 12 institutions. The years 2001 to 2003 were spent on development and infrastructure; the years 2003 to 2005 are for practice and expansion. Six major organizations (NIMS, JAXA, NIED, JAERI, RIKEN, JST) lead the project. The main backbone will be the SuprSINET with a speed of 19 Gbps. There are seven major applications (materials science, data bases, vibration and corruption simulation, aerodynamics, cell simulation, life sciences, numerical environments) and 21 connections into the network with a total capability of 7 Tflop/s (not including the ES). ITBL software efforts must address a firewall across the system, communication between heterogeneous computers and the Task Mapping Editor (TME, visual work flow). Much of the infrastructure software has already been done. This is an impressive effort and JAERI appears to be heavily involved.

JAERI HIGH-END COMPUTING

A large number of large-scale calculations across the research spectrum of JAERI were presented, almost all trying a first use of the ES.

1. Molecular design of new ligands for actinide separation (Masaru Hirata, Tokai Material Science) uses relativistic density functional theory on the VPP5000 ITBL computer at JAERI to understand better 5f electron orbits, and first-principle (rather than classical) molecular dynamics on the ES. Much larger systems can be calculated than before.
2. Computation in plasma physics for magnetic fusion research: Shinji Tokuda, Naka Fusion Research Establishment addressed the NEXT project (Numerical Experiment of Tokamaks: particle simulation for plasma turbulence MHD simulation), grid computing in fusion research, and development of a fast solver of the eigenvalue problem in MHD stability. The goal is prediction of turbulence transport in a Tokamak, such as ITER, by nonlinear gyrokinetic particle simulation. It will take several 100 Tflop/s or even Pflop/s to simulate the ITER plasma within one day. A comparison of these calculations on the JAERI Origin 3800 system (800 scalar PEs, 1Gflop/s/PE, total peak 0.8 Tflop/s) yields 25% efficiency up to 512 PEs = 512 Gflop/s; for the ES (5210 vector PEs, 8Gflop/s/PE, 40 Tflop/s total) it stands at 26% efficiency for 512 PEs = 4096 Gflop/s, without saturation in either case. These calculations identified previously unknown instabilities. This leads to the suggestion that a sufficiently powerful computer in a control loop with a Tokamak could determine when the plasma is approaching instability and perform online corrections. The Fusion Grid with universities in Japan has started and the main computing resources will come through ITBL.
3. Two-phase flow simulations for an advanced light water reactor (Kazuyuki Takase, Nuclear Energy Systems, Tokai) use the ES two-phase flow analysis in a reactor core. The code is parallelized using MPI and was able to use 300 CPUs of the ES. A full core simulation took 20 Pb of memory.
4. Parallel Molecular Dynamics Stencil (PMDS, Futoshi Shimizy, CSSE) developed at JAERI is used for defect simulations in materials with a 100nm cube for 10^6 time steps = 1ns. Visualization is done with ATOMEYE from Ohio State U. PMDS is written in C using MPI for parallelization. The code divided the cell into subcells, with each processor tracking all the particles within every subcell. However, atoms may interact with atoms in another cell.
5. Advanced photon simulations with a Tflop/s/TB parallel computer (Mitsuru Yamagiwa, Advanced Photon Research Center, lasers and FELs), which computes laser-gas interaction, which produces plasma. The result is that the radiation pressure can produce very energetic ions.
6. Large-scale numerical simulations for superconducting devices (Masahiko Machida, CCSE) solved the time-dependent Ginzburg-Landau equations on the ES. A prototype test code, including Maxwell's equations using MPI, produced an excellent efficiency of 18.4 Tflop/s (56% of peak)/ 512 nodes. HPF produced 96% of the MPI efficiency for 128 nodes. They also use the ES for parallel exact diagonalization for strongly correlated electron systems. Using the Lanczos method they cussed in

diagonalizing an 18-billion dimensional matrix with 33% peak efficiency (22.7 Tflop/s). This matrix requires 1.1 Tb of memory.

OBSERVATIONS

The drive at JAERI toward the use of the Earth Simulator for the solution of practical problems other than climate research is impressive. They obtain excellent efficiencies for a wide range of scientific and engineering problems on the ES. They seem less interested in their own next computer than in tying themselves, through the ITBL and SuperSINET, to all computers in the national research system. They are hard at work on the ITBL and are beginning to use it. They foresee less a grid application where a problem is distributed among several computers than moving a given problem to the most appropriate computer.

Site: **Japan Aerospace Exploration Agency (JAXA)**
Information Technology Center
Institute of Space Technology and Aeronautics
7-44-1 Jindaiji-Higashi, Chofu-shi, Tokyo 182-8522
http://www.jaxa.jp/index_e.html

Date Visited: March 30, 2004

WTEC Attendees: R. Biswas (Report author), P. Paul, S. Meacham, A. Aono (interpreter)

Hosts: Toshiyuki Iwamiya, Director
Takashi Nakamura, Chief Manager
Naoki Hirose, Senior Research Scientist
Yuichi Matsuo, Manager

BACKGROUND

On October 1, 2003, three Japanese space research and development organizations, the Institute of Space and Astronautical Science (ISAS), National Aerospace Laboratory (NAL), and National Space Development Agency (NASDA) were merged to form one independent administrative institution known as the Japan Aerospace Exploration Agency (JAXA). ISAS was originally responsible for space and planetary research, and operated like an academic institution; NAL was independent, and focused on research and development for next-generation aviation; NASDA developed launch vehicles such as the International Space Station with funding from the Japanese Government. It is believed that this consolidation will enable a sustained and coordinated approach to space exploration in Japan, from fundamental research through development to practical applications and missions. JAXA has been placed under the jurisdiction of the Ministry of Education, Culture, Sports, Science, and Technology (MEXT).

JAXA has several field centers within Japan, as well as offices overseas. The Washington, DC, office serves as its liaison with the National Aeronautics and Space Administration (NASA) and National Oceanic and the Atmospheric Administration (NOAA) headquarters. Its office in Los Angeles covers the western part of the U.S., and serves as the liaison with the NASA Jet Propulsion Laboratory. Similarly, its Houston office covers the southern U.S., and liaisons with the NASA Johnson Space Center. JAXA's fourth office is at NASA Kennedy Space Center, and is primarily responsible for coordinating with the U.S. on the International Space Station.

The Aerospace Research Center (ARC), where JAXA's Institute of Space Technology and Aeronautics (ISTA) is located, performs leading-edge research and development in aerospace technology. NAL forms the core part of ISTA; the remainder consists of NASDA's Tsukuba Space Center. NAL was originally established as the National Aeronautical Laboratory in July 1955, but assumed its final name with the addition of the Aerospace Division in April 1963. It pursues research on aircraft, rockets, and other air and space transportation systems, including relevant enabling technology. NAL is also responsible for developing, maintaining, and enhancing large-scale test facilities (e.g. hypersonic and low-speed wind tunnels, altitude and ramjet engine test centers, a flight simulator, a multipurpose aviation laboratory), and making them available to other related organizations (universities, laboratories, and companies). The Information Technology Center (ITC) of JAXA/ISTA that we visited at ARC in Chofu-shi is developing CFD technologies and software that contribute to and advance the research projects in JAXA and various Japanese industries. It also houses the second-largest supercomputer (after the Earth Simulator) in all of Japan. This machine is a 2304-processor Fujitsu PrimePower HPC2500 system, based on the 1.3 GHz SPARC64 V architecture.

Almost all of the research conducted at ISTA can be grouped into five main categories: crafting new aircraft design, designing next-generation spacecraft, enhancing aviation safety, preserving the environment, and supporting aerospace technologies. New aircraft research includes work on the next-generation supersonic transport (SST) and the stratospheric platform airship system (SPF). ISTA is accumulating existing

information and establishing new technologies for reusable space transportation systems that were verified by three flight tests: orbital reentry experiment (OREX), hypersonic flight experiment (HYFLEX), and automatic landing flight experiment (ALFLEX). To increase aviation safety, ISTA is conducting research on the safety of flight operations (such as performance, navigation, guidance and control, human factors, and weather) and on aircraft structures (such as weight reduction, crash impact, and static/fatigue loads). Preserving the environment includes both Earth and space. On Earth, ISTA is performing research to reduce jet engine noise and exhaust gas emission, and the development of technology for high-precision air-quality measurements. In space, the goal is to investigate the debris problem caused by past missions, and to prevent collisions with satellites in orbit, the International Space Station, and future spacecraft. Finally, to support aerospace technologies, ISTA conducts fundamental research on computational fluid dynamics (CFD), aircraft and spacecraft engines, structures and materials, artificial satellites, and space utilization.

Shuichiro Yamanouchi is the current President of JAXA. As of FY2003, it had a regular staff of 1,772 and a six-month budget of ¥98.46B. Before the merger, NASDA had over 60% of the personnel and over 70% of the budget, while the remainder was split almost equally between ISAS and NAL. Susumu Toda is the Director-General of ISTA. The ITC has a budget of about $16M per year and a staff of approximately 25, of 10 of whom are permanent.

JAXA PRESENTATION

Yuichi Matsuo, Manager of ITC, made a presentation that focused on JAXA's new computing system (Numerical Simulator III) and its parallel performance on a wide range of applications. He began by giving an overview of the long and illustrious history of numerical simulation technology research and development at NAL since the time it was established more than 30 years ago. NAL installed the Burroughs DATATRON digital computer in 1960, and the Hitachi 5020 (the first transistor computer in Japan) in 1967. In the mid 1970s, NAL and Fujitsu jointly developed the first Japanese supercomputer, FACOM 230-75AP, which became operational in February 1977 (the Cray-1 was built in 1976). It supported list-directed vector accesses; so AP-Fortran had to be developed to derive the maximum performance from the Array Processing Unit (APU) by including vector descriptions. The peak performance of the APU was 22 Mflop/s. This was the first vector processor in Japan, and it started a close collaboration between NAL and Fujitsu that has lasted till today. However, since there was no interest in vector processing from other research institutions and university computer centers, only one unit of the FACOM 230-75AP was ever built.

In 1987, FACOM VP400, the most advanced version of Fujitsu's VP line at the time, began to run 3D Navier-Stokes CFD applications at NAL. It was able to complete full-configuration simulations around complete aerospace vehicles in fewer than 10 hours. The VP400 employed a pipeline architecture with a peak vector performance of 1140 Mflop/s. Research and development of the Numerical Wind Tunnel (NWT) began as a national project in 1988 with the goal of having a 100x performance improvement over the VP400. In late 1988, Fujitsu launched its VP2000 series supercomputer system, and announced the high-end VP2600 (with enhanced vector performance) in December 1989. When the NWT was introduced in February 1993, it had a peak performance of 236 Gflop/s and occupied the top spot on the Top500 list (November 1993). It was a distributed-memory vector-parallel supercomputer, and consisted of 140 VPP500 processors. In 1996, NWT was enhanced with 166 nodes and the addition of visual computer systems, and became known as Numerical Simulator II (peak of 280 Gflop/s).

The latest supercomputer, dubbed Numerical Simulator III (NSIII), was introduced in 2002. This machine, also built by Fujitsu, uses commodity cluster technology rather than traditional vector architecture. The building block of the NSIII is the PrimePower HPC2500, a 1.3 GHz SPARC64 V architecture (Fujitsu proprietary chip that is completely compatible with the SUN products) with 8 CPUs (5.2 Gflop/s peak) per board and 16 boards per cabinet. A cabinet is the physical unit in terms of the hardware and is a 128-way SMP with 665.6 Gflop/s peak and 256 GB shared memory. A cabinet can be partitioned into two or four nodes, which is the logical unit at the operating system level. Within a cabinet, the 16 boards are connected via a 133 GB/s crossbar. The machine runs the standard 64-bit Solaris 8 operating system.

The NSIII consists of the Central Numerical Simulation System (CeNSS), which is the computing subsystem and has 14 PrimePower HPC2500 cabinets, for a total of 1792 processors and a peak performance of 9.3 Tflop/s. The inter-node connection is via a 4 GB/s bi-directional optical crossbar, where a hardware barrier provides synchronization and each node has its own data transfer unit (DTU) (similar to a network interface card). CeNSS has 3.6 TB of main memory, and is connected to the Central Mass Storage System (CeMSS) via a single PrimePower cabinet serving as an I/O node. CeMSS has a total capacity of 57 TB FC RAID disk space and 620 TB LTO tape library. CeNSS is connected to the Central Visualization System (CeViS) via the I/O node and a 500 MB/s Gigabyte System Network (GSN) link. CeViS consists of a 32-processor, 64 GB SGI Onyx3400 visualization server, a 4.6m×1.5m (3320×1024 pixels) wall display, and several graphics terminals. There are plans to upgrade CeViS, which is more than three years old and not very popular with scientists. Three additional cabinets are each partitioned into two 64-processor nodes: four of these are used as service nodes (compilation, debugging, etc.) while the remaining two are login nodes connected to the Internet. The NSIII thus has a total of 18 PrimePower cabinets (2304 processors), and is currently the only Fujitsu computer with LINPACK performance greater than 1 Tflop/s.

NSIII COMPUTER FACILITY

A hybrid programming model is used on CeNSS. Within a node, one can use either Fujitsu autoparallel or OpenMP directives (thread parallel). Across nodes, the message passing paradigm is XPFortran or MPI (process parallel). JAXA stayed with XPFortran, instead of migrating to HPF, which has been very successful on the Earth Simulator, because they already had it installed and working on the NWT since 1993. Fujitsu provided the necessary compilers, and all code transformations from NWT to CeNSS were straightforward.

Table B.4
Comparison between NSII and NSIII

	NS II (NWT)	NS III (CeNSS)
Number of Nodes	166	56
CPUs per Node	1	32
Number of Processors	166	1792
Peak (Gflop/s)	282	9318
Total Memory (GB)	42.5	3584
GB/Gflop/s	0.151	0.385
Memory Bandwidth (GB/s)	6.6 per PE	133 per 128p
Interconnect Bandwidth (GB/s)	0.421 ×2	4 ×2
Disk Storage (TB)	0.3	57

The presentation included parallel performance results for four applications: flow over full aircraft, combustion flow, turbulent channel flow, and helicopter rotor flow. The aircraft application used the large Eddy simulation (LES) technique, had moderate memory access and light communication requirements, and was originally process-parallelized via MPI. Hybrid performance on a 21M-gridpoint problem showed linear scalability to 960 processors. The combustion application used the direct numerical simulation (DNS) technique, had light memory access and communication requirements, and was also originally parallelized with MPI. Again, the hybrid version on a 7M-gridpoint problem showed linear performance to 512 processors. The turbulent channel flow was simulated using DNS and FFT, had moderate memory access but heavy communication, and was originally parallelized using XPFortran. The hybrid implementation demonstrated linear performance to 256 CPUs on a 1400M-gridpoint problem, but deteriorated for larger processor counts. Finally, the fourth application simulated helicopter rotor flow using unsteady Reynolds-averaged Navier-Stokes (URANS), featured indirect addressing due to interpolation, had heavy memory access and communication, and was also originally parallelized in XPFortran. The performance of the hybrid

version of this application was not very good. A new RANS code, called UPACS, has recently been developed in F90, and is undergoing extensive testing and optimization.

ANSWERS TO OUR QUESTIONS

In the area of programming paradigms, Yuichi Matsuo responded that they have not seen any major innovation in the past several years. However, he believes that since the future of massively parallel HEC systems depends on the efficiency and usability of programming paradigms, there may be some breakthroughs in the near future. They also question the viability of MPI for very large parallel systems, and admit that HPF has several advantages. NAL lost interest in HPF because of its lack of portability, and instead focused on XPFortran, a similar data-parallel language with which they were familiar. OpenMP is beginning to gain acceptance among users, so an evaluation of hybrid programming paradigms is currently underway. There is also very little R&D activity on programming tools (such as automatic parallelization, debugging, and performance analysis) in Japan. However, there is some on-going work in developing post-processing tools for visualizing large data sets. Less than 5% of the total budget is spent on software development.

One likely innovation is the research and development of scheduling systems based on local environments and operation policies. The goal is to increase resource utilization and throughput. At NAL, research on job schedulers has been conducted since the time of the NWT, and work continues under the new JAXA organization. There is some concern about the reliability of open source software and the challenge of dealing with frequent version upgrades.

Dr. Matsuo believes that petascale systems (consisting of more than 10,000 processors) have not yet been considered seriously in Japan. At JAXA, even the 2304-processor system poses many challenges. Scientists and engineers at ITC are still trying to maximize effective usage and sustained performance for real applications. He thinks that hardware has been taking the lead in Japan so far; therefore, the development of software and systems technologies has fallen behind.

Regarding computer architectures, Dr. Matsuo thinks that national projects like the Earth Simulator may choose the option that will provide maximal sustained performance as long as there is adequate funding. Therefore, vector architectures may continue to be selected if such projects are funded. However, under today's severe price competition, he doubts the viability of vector computing. He is also not aware of any major Japanese thrust on designing innovative architectures (such as processor-in-memory, and streaming technology). Power requirement is also a major constraint since the price of electricity in Japan is fairly high. At this time, ITC has 3 MW of electric power available.

At JAXA, computational aeroacoustics, aviation weather, and cosmic science are considered the next major application areas. Japan as a whole may put more emphasis on supporting nanotechnology, computational biology, protein structure studies, and materials research. Applications directly related to improving the quality of life or the development of the national economy and industry are very important. Space science is not as well established as it is in the U.S.. Military and security affairs also have relatively low priority. JAXA realizes that there will always be capability and capacity requirements, and a roadmap for CFD applications is under construction.

REMARKS

It was surprising to see Fujitsu abandon their traditional vector architectures in favor of commodity cluster technology. Even more surprising was the importance that people and organizations in Japan place on personal and corporate relationships. This was evident when NAL gave up on vector machines (with which they had been extremely successful) to continue their longstanding partnership with Fujitsu. They did go through a procurement process before acquiring the PrimePower HPC2500, with the main objective to satisfy user requirements. Several people at JAXA/ISTA reiterated that Fujitsu has been a good partner, and that they always did their best to meet the scientists' demands. However, they also had technical and economic

reasons to switch from the VPP series to PrimePower. The VPP systems had a slower clock; hence scalar performance was poor compared to commodity processors. But the VPP machines achieved a higher percentage of peak performance (on average, 30% compared to 10% for the PrimePower); a difference that is easily made up in the cost of the systems (the PrimePower costs about a tenth of the VPP). Other reasons, according to ITC personnel, for moving from vector to commodity cluster, were power requirements and procurement/maintenance costs. ISTA's next acquisition is slated for 2006-07, but it is not clear what that machine will be.

The ITC folks also remarked that in 1997 although the Japan Marine Science and Technology Center (JAMSTEC), Japan Atomic Energy Research Institute (JAERI), and NASDA were brought together by Hajime Miyoshi to develop the Earth Simulator (ES), JAXA is not supporting the ES at this time. NAL has a long track record in HEC, but scientists must write research proposals to get time on the ES. This is surprising given that Mr. Miyoshi spent more that 33 years at NAL, retiring as Deputy Director-General in March 1993. He was a visionary, had a passion for supercomputing, and was able to convince the right people at the right time (Kyoto Protocol) to fund and collaborate on the ES project, but currently there is no strong HEC leadership in Japan. As a result, the people we met at ISTA believe that there is no definite long-term HEC plan and commitment in Japan, and a follow-on to the ES seems unlikely. Furthermore, big social changes are underway to privatize universities and supercomputing centers starting April 1, 2004, in order to make the Japanese Government more efficient. Flexible management is an advantage of any independent agency, but it must be operated properly and efficiently in order to survive and fulfill its mission.

REFERENCES

JAXA. 2003. Japan Aerospace Exploration Agency (Brochure).

NAL Research Progress 2002~2003. 2003. National Aerospace Laboratory. ISSN 1340-5977.

Numerical Simulator. 2003. Progress in Numerical Simulation and History of Computer System at JAXA Information Technology Center (Brochure).

Site: **High Energy Accelerator Research Organization (KEK)**
1-1 Oho, Tsukuba,
1-2 Ibariki 305 0801
http://www.kek.jp/intra-e/

Date Visited: March 31, 2004

WTEC Attendees: P. Paul (Report author), A. Trivelpiece, K. Yelick, S. Meacham, Y. Chien

Hosts: Dr. Yoshiuki Watase, Director, KEK Computing Research Center
Professor Shoji. Hashimoto, head Computational Physics Group
Dr. Nobu Katayama

BACKGROUND

KEK is the high-energy physics research center of Japan. It runs the Belle B factory, which has the world's best luminosity. Its designers won the Wilson Prize this year. It is also leading the world in long-base line neutrino studies. KEK heads the large J-PARC (Japanese Proton Accelerator Research Project) construction project at Tokai that will be dedicated to materials and life sciences (spallation neutron source) and nuclear physics (neutrino source). KEK is the lead laboratory for the proposed Global Linear Collider project. In addition it has efforts in materials science and protein studies. It was established in its present form and location in 1971, has a total staff of ~700 and an annual budget of ~$330 million provided by MEXT.

The Computing Research Center is part of the Applied Research Laboratory. It serves the general computing and IT needs of KEK, and develops software for the research programs.

The focus of this visit was to learn about recent decisions at CRC and plans for the future.

CRC COMPUTERS

Dr. Watase gave a brief overview of the CRC. It operates the Central Data Analysis System, the B-factory Computer System and the Supercomputing System. It collaborates with the Japanese Tier-1 center at University of Tokyo for the ATLAS detector project at the LHC.

Dr. Katayama presented the Belle computing system. It consists of an extensive PC farm from Compaq and Fujitsu with a huge data storage system (250 TB/year right now, more needed after Belle upgrade). The PC farm is being upgraded continuously, most recently with 120 2CPU 3.2 GHz Xeon boards from Fujitsu. Two people manage this PC farm. Belle is working on the connectivity of the SuperSINET to connect to several universities for data processing.

Dr. Hashimoto described the Supercomputer at KEK: a Hitachi SR8000/100model F1 system with a main memory of 448 GB, installed in 2000. It has 100 nodes, each with 12 Gflop/s at peak speed. The total 1.2 Tflop/s peak performance makes it presently #9 in Japan. The machine has about 100 users, with ~80% spent on lattice gauge calculations, for which the machine is 30% efficient. The language is Fortran 90, which is best for the Hitachi compiler (provides parallelized code).

OBSERVATIONS

They have access to about 1% of the Earth Simulator time. However the ES is not very useful (~30% efficiency) because the lattice gauge program structure is not well matched to the ES. The view was that the ES was an expensive machine to build and to maintain (at ~$30 million/year).

KEK would like to have a larger computer, preferably a cluster for massive amounts of data analysis. Computing at KEK is driven by specific science, whereas the nearby National Institute of Advanced

Industrial Science and Technology and the University of Tsukuba Institute for Information Sciences and Electronics with a Center for Computational Sciences are both national drivers in computational science.

Site: **Ministry of Economy, Trade and Industry (METI)**
1-3-1 Kasumigaseki, Chiyoda-ku, Tokyo 100-8901
Tel: +81-3-3501-6944, Fax: +81-3-3580-2769
http://www.meti.go.jp/english/index.html

Date Visited: March 30, 2004

WTEC Attendees: K. Yelick (Report author), A. Trivelpiece, J. Dongarra, S. Meacham, Y.T. Chien

Hosts: Mr. Hidetaka Fukuda, Director, Information and Communication Electronics Division Commerce and Information Policy Bureau
Mr. Hidehiro Yajima, Deputy Directory, Information and Communication Electronics Division Commerce and Information Policy Bureau

OVERVIEW

The Ministry of Economy, Trade and Industry (METI) is in charge of administering Japan's policies covering a broad area of economy, trade and industry. METI was created in 2001 as a reorganization of the Ministry of Commerce and Industry (MITI), which was established in 1946. There are six bureaus within METI: Economic and Industrial Policy Bureau; Trade Policy Bureau; Trade and Economic Cooperation Bureau; Industrial Science and Technology Policy and Environment Bureau; Manufacturing Industries Bureau; and Commerce Information Policy Bureau. There are also several councils set up as advisory bodies to METI, made up of experts from industry, finance, labor, academia and media. The WTEC group met with the members of the Commerce and Information Policy Bureau.

ELECTRONICS INDUSTRY

Ten years ago, Mr. Fukuda was the Deputy Director in charge of supercomputing. It was a difficult time in trade relations between the U.S. and Japanese, and although U.S. researchers wanted to buy Japanese supercomputers, the Japanese vendors had little market access due to government restrictions. At that point, Hitachi, NEC, Fujitsu had mainframe products and did not recognize the change that was occurring as RISC processors from HP, Sun, and others started to take over the mainframe business.

Today we have the IBM Power architectures, HP PA RISC, Sun Micro systems SPARC, along with Intel and AMD processors. There are no players in the microprocessor market in Japan. Ten years ago METI invested in scientific computing, but there were not enough applications to sustain a market. METI is no longer interested in supercomputing.

The Japanese envy the U.S. venture capital model for investing in industry. There is nothing similar from Japanese investors. There is investment from U.S. venture capital firms into Japan, but it is for services, not for high-tech. While technology investments spilled over into private sector in the U.S. that did not happen in Japan.

SUPERCOMPUTING

On the subject of the Earth Simulator, Mr. Fukuda said that even if the government invested money in supercomputing, there would be no commercial applications. The ES people have explored commercial applications (e.g., Toyota), but the machine is too big to be useful. The chip design in ES is unique and proprietary, not a commodity, which makes it expensive given such a small market.

There was some discussion of possible applications areas for supercomputing, including oil exploration, automobile design, and aerospace. The METI hosts emphasized they are not interested in huge supercomputers, but are interested in PC clusters, because of their cost effectiveness. There was some

discussion about ease of use of PC clusters, and the general conclusion was that usability depended on the application domain.

The supercomputing market in Japan reaches $400-500M, and while the Japanese companies compete well with U.S. supercomputing vendors for the European market and others, they cannot compete in the U.S. There is not a sufficient market, therefore, to sustain three Japanese companies building supercomputers. METI provides financial incentives to industry, including NEC ($30M over three years), Fujitsu ($50M) and Hitachi. METI recommended to NEC and Fujitsu that they change their hardware plans due to the cost of developing new chips. Their specific proposal was to build high-reliability systems in the spirit of Tandem machine by combining multiple Intel or AMD processors, using software based on Linux.

Mr. Fukuda expressed more enthusiastic support for grid computing, with applications in e-business and e-government. They see this technology as a way to foster better communication between the government and individuals, for example. The table below summarizes METI's funding in information technology, which shows that the business grid is a high priority, surpassed only by their program for creating digital tags that may replace bar codes.

Table B.5
METI's Funding in Information Technology

	JFY2004 Budget	JFY2004 Request	JFY2003 Budget
	¥ Million	¥ Million	¥ Million
Digital tag: diffusion	3,000	3,500	0
IT in medical field	1,494	1,962	0
Electric resource development	527	880	0
Illegal access prevention	674	1,000	0
Software engineering by industry-university collaboration	1,482	2,750	0
Business grid computing	2,602	2,801	2,797

OBSERVATIONS

METI's role within the government is to set policies and make investments that will encourage commercial growth, so their vision of possible applications of supercomputing were limited to commercial applications rather than scientific ones. Although METI had supported scientific computing in the past, their lack of current interest in it was quite striking. There are several reasons for this, some of which are unique to Japan. The first is the lack of commercial supercomputing applications, especially for very high-end machines like the ES. In particular, Japanese automobile manufacturers are not interested in using the ES, because it is a larger machine than they need. Second, there is no microprocessor line coming from Japan, so it is difficult to justify investing in high-end computing technology based on an argument that some of the technology would trickle down to computer systems with larger sales volumes. Third, the lack of a defense industry within Japan removes one of the drivers of high-end computing technology that exists in other countries.

While METI officials were not interested in technology specifically designed for supercomputing, such as vectors, there was some interest in PC clusters and even more in grid technology. PC clusters were viewed as a viable option for high-end computing, due to cost performance, and there was little appreciation for any limitations of this approach. The interest in grid computing was motivated by business and government applications, rather than scientific ones, which is consistent with METI's agenda.

Site: **Ministry of Education, Culture, Sports, Science and Technology (MEXT)**
2-5-1 Marunouchi, Chiyoda-ku, Tokyo, 100-8959
Tel: +81-3-6734-4061, Fax: +81-3-6734-4077
http://www.mext.go.jp/english/index.htm

Date Visited: April 1, 2004

WTEC Attendees: K. Yelick (Report author), A. Trivelpiece, J. Dongarra, Y.T. Chien

Hosts: Mr. Tsuyoshi Maruyama, Deputy Director-General, Research Promotion Bureau, MEXT
Mr. Harumasa Miura, Director, Information Division, Research Promotion Bureau, MEXT

BACKGROUND

The Ministry of Education, Culture, Sports, Science and Technology (MEXT) is responsible for several activities within the Japanese government, including elementary, secondary, and higher education, science and technology policy, and programs in sports and culture. The WTEC committee met with members of the Research Promotion Bureau within MEXT. The Research Promotion Bureau is responsible for promoting fundamental research and development in life sciences, information science and technology, nanotechnology and materials, and quantum and radiation research. It also formulates policies to promote scientific research and utilization of research results and business-academic-public sector cooperation.

The Earth Simulator (ES) Program

Mr. Tsuyoshi Maruyama was the Director of the Ocean and Earth Division of the Science and Technology Agency (STA) of Japan at the time the Earth Simulator was first conceived. The ES was conceived as a tool to promote basic research, not to promote a particular computer technology. A Global Change Research program was envisioned at that time as having three components:

1. Modeling
2. Measurement tools (satellites, buoys, etc.)
3. Simulation (The Earth Simulator)

Around the time the ES program was first proposed, there were also discussions about global change between the Japanese and U.S. Governments involving the International Pacific Research Center (IPRC), the National Aeronautic and Space Administration (NASA), the National Atmospheric and Oceanic Administration (NOAA), National Science Foundation (NSF), Science and Technology Agency (STA, from Japan). They identified three centers for global chance research that should be tied together in a kind of international consortium, one in Hawaii, one in Alaska (the IARC), and one in Japan (the ES). There was no formal structure, however, and no international funding. The group from these government agencies discussed the possible access to the ES machine by U.S. researchers. Today U.S. researchers can use the machine in collaboration with Japanese researchers, although there is no Internet access to the machine, even for the other two centers identified above.

The cost of the Earth Simulator was about ¥60 billion. It was considered a very risky project, as there was no assurance at the time the project was started that it could be accomplished. They received a 50% "supplementary" budget to complete the project, which allowed them to complete the project in five years. Without the supplement, the project would have taken approximately eight years to complete.

FINANCIAL STRUCTURE OF JAPANESE RESEARCH ORGANIZATIONS

In April 2004 a new fiscal year started (two days after the WTEC visit), and at that time there was a major change in the funding of research organizations, including universities, in Japan. Some of the research organizations had already changed to this new model. Under the old model, funding was given directly from the government to particular research activities within a university, so the government had direct control over how money was spent. Under the new model, the funding goes to each university, and the university manages the budget. In addition, there will be a 2-3% cut in managerial expenses (the equivalent of "overhead" in the U.S. system) each year. (In response to a question about whether the 2-3% cut was on top of inflation, the MEXT people responded that there had been almost no inflation in Japan recently, so this was not an issue.) The goals of the new model are twofold: 1) to save money, and 2) to give the universities more autonomy from the government.

Budget ideas can be proposed to the Council of Science and Technology Policy (CSTP) and they will give MEXT a prioritization of those requests. In addition, under the new funding model, universities may start new projects by cutting old ones, since they control their own money. Some research organizations may also have some form of seed-money funding (similar to the LDRD program in the DOE labs in the U.S.) to fund new projects based on some kind of tax on existing ones.

All 89 universities switched to the model on April 1, 2004. This includes the seven supercomputing centers, each of which has a computing facility purchased under a procurement system. In the future, they may not be able to introduce bigger computers, due to budget limitations. As a cost-cutting measure, the lifetime of each computer that is purchased is getting longer, so it is also possible that the centers will move toward PC clusters to get better performance for their money.

FUTURE PRIORITIES IN COMPUTING

When asked about follow-on projects to ES, the MEXT officials said there were no specific plans to build such a system, nor were their funds to upgrade the system. There was also no policy to subsidize the Japanese supercomputing companies. They are, instead, investing in grid computing and other long-term research areas, including quantum computing and nanotechnology.

- *Grid Computing*: MEXT is putting ¥2 billion per year into grid research through 2007. This money covers both equipment and personnel, and includes major projects such as the National Research Grid Initiative (NAREGI) at the National Institute of Informatics (http://www.nii.ac.jp/) and several universities and laboratories.
- *Quantum Computing*: The quantum computing activities involve industry and universities. For example, NEC just made a new quantum-computing device.
- *Nanotechnology*: There is an investment in nanotechnology from MEXT, including a joint workshop with NSF.
- *Life sciences*: MEXT is funding work in genomics, medical research, and other areas in the life sciences.

There was a discussion of the four classes of machines in the U.S.: clusters, supercomputers (including vectors), special purpose machines, and grids. Mr. Maruyama indicated that there was insufficient market in supercomputers, although they were still needed at research institutions like JAMSTEC; PC clusters provide good cost performance, but a discussion is needed about their usefulness. In response to a question about whether the government needs to help the supercomputer vendors, Mr. Maruyama made several points:

- There are no plans for a follow-on machine to the ES, given budget limitations within the government.
- There were still questions about what research problems need supercomputers, as opposed to lower-cost solutions.
- There are no military applications in Japan to drive the need for supercomputing, so the answer may be different in Japan than in the U.S. The application drivers in Japan are things like global warming and earthquake modeling.

In the area of software for supercomputing, MEXT is funding some work at the University of Tokyo. One of the goals of the MEXT agenda in general is to see if PC clusters can be made as usable as supercomputers through better software support.

There was also some discussion of human resource issues in supercomputing and scientific fields in general. In the Japanese educational system there are many talented students, but most quit with a master's degree and do not complete a Ph.D. At the moment, Japan has 100,000 international students, who are much more likely to join a Ph.D. program in Japan or in the United States.

BUDGETING PROCESS

Budgets prepared for MEXT are based on bottom-up budgeting and sent to the Minister of Finance. Parliament approves that budget. Parliament does not modify the budget, out of tradition, so any negotiations are done in advance. Social Security and Science/Technology are getting budget increases this year, while all other budgets are decreasing. The Science and Technology spending plan is divided into two five-year terms. They are currently in the second of those terms. In the first term, they had a target budget of ¥17 trillion, and spent slightly above that level at ¥17.6 trillion. The second term goal is ¥24 trillion, although it will be difficult to meet that goal because of the downturn in the Japanese economy. The table below summarizes the Information Technology projects funded by MEXT (or in one case proposed for funding).

Table B.6
Information Technology Projects Funding by MEXT

	JFY2004 Budget ¥ Million	JFY2004 Request ¥ Million	JFY2003 Budget ¥ Million
E-science project	3,500	4,508	1,505
E-society software	1,100	1,202	1,202
Digital archive of intelligent assets	500	1,007	0
National research grid initiative	1,950	3,202	2,002
Advanced computing for simulation	0	2,000	0
EUV light source	1,140	1,200	1,200
IT program	3,500	4,508	3,005

Japan spends 0.8% of its total GDP on Research and Development, which is comparable to that of other countries. The percentage coming from the private sector (0.35% of the GDP) is the highest among other nations.

OBSERVATIONS

The two most significant items that came up during the visit were the change in the university funding model and general budgetary concerns due to the economic downturn. It was clear that there was no plan, at least at the government level, for a follow-on to the Earth Simulator, and as the funding levels for the projects in the table show, grid computing is currently a much higher priority than scientific simulation.

Site: **National Institute of Informatics (NII)**
National Research Grid Initiative (NAREGI)
Jinbocho Mitsui Bldg., 14F
1-105, Jinbocho, Kanda, Chiyoda-ku, Tokyo
http://www.naregi.org/index_e.html

Date Visited: April 1, 2004

WTEC Attendees: Y.T. Chien (Report author), S. Meacham, P. Paul, A. Trivelpiece

Hosts: Dr. Kenichi Miura, Project Leader, NAREGI and Professor of NII
Dr. Satoshi Matsuoka, Deputy Leader, NAREGI and Professor of Tokyo Institute of Technology

BACKGROUND

Grid computing is an important element of Japan's recent push towards superiority in high-end computing. In recent years, there have been several major grid-related research projects at the national level, designed to promote supercomputing capabilities through the use of computer networks. The National Research Grid Initiative (NAREGI) is the latest and perhaps the most ambitious. It was launched to promote collaborative research and development across academia, industry and government. Funded by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), NAREGI's overall goal is to provide high-performance computing for scientific communities using high-speed networks and scaleable middleware. Its broad vision is to contribute to the technical advances in grid computing and thereby help improve Japan's international competitiveness and economic developments.

NAREGI is a five-year program (2003-2007) anchored in the National Institute of Informatics (NII), itself a relatively new organization (formerly NACSIS – National Center for Science Information Systems), known for its role in providing ultra high-speed computational and networking capabilities (Super SINET) and information resources for academic research. Our hosts were Dr. Kenichi Miura, Project Leader for NAREGI who is a Professor at NII and Dr. Satoshi Matsuoka, Deputy Leader and a Professor at the Tokyo Institute of Technology. Both of them are well known for their work in high-end computing in and outside of Japan.

Project Goals

Our visit to the NAREGI project took place at the NII headquarters in downtown Tokyo. The 2+ hour visit consisted of a brief overview given by Dr. Matsuoka [1], followed by discussions. Dr. Matsuoka in his presentation pointed out several important facts about NAREGI. First, NAREGI is really only one year old, having just completed the acquisition of computer resources in the last fiscal year. The project received ¥2B (~$17million) from MEXT in 2003, with additional $47 million for the experimental testbed currently under development. One of the key features of this project is its collaborative mandate that brings universities, national laboratories, and industries into a research consortium with shared goals and complementary expertise and resources. There are four general goals:

1. To develop a grid software system (R&D in grid middleware and upper layer) as the prototype for future grid infrastructure in scientific research in Japan
2. To provide a testbed to prove that the high-end grid computing environment can achieve 100+Tflop/s by 2007. This capability will be demonstrated in nanoscience research over the Super SINET
3. To participate in international collaboration for grid computing research (U.S. Europe, Asian Pacific)
4. To contribute to standardization activities, e.g., within the GGF (Global Grid Forum)

FIVE-YEAR PLAN AND RESEARCH FRAMEWORK

NAREGI has developed a five-year plan and a comprehensive R&D framework designed to achieve its ambitious goals. The five-year plan, not discussed at Dr. Matsuoka's overview, is contained in a separate NII publication handed out during our visit [2]. The plan divides the five-year project into four phases: research and development (2003-2004), evaluation (2005), refinement, improvement & re-evaluation (2006-2007), and utilization (2008 and beyond). In each phase, research is organized into three layers of efforts: computer resources (e.g., acquisition, installation, integration of equipment for the testbed), grid software R&D (e.g., middleware development, application support for grid environment), and verification of the grid system (e.g., in selected application areas). Upon the completion of research in 2007, NAREGI will help transition the technologies from academic research organizations to the private sector, according to the plan.

To make the plan work effectively, NAREGI places a heavy emphasis on collaboration among universities, national labs, and industries. AIST, Titech, Osaka U., Kyushu U., Utsunomiya U., national supercomputing centers, vendors, as well as a consortium for promotion of grid applications in industry are some of the organizations participating in various joint R&D efforts. The Super SINET, an all-optical network also run by the NII, serves as the backbone of the NAREGI grid infrastructure. According to Dr. Miura, the Center for Grid Research in NII is officially the site where NAREGI carries out its work. However, nearly all of the research funds go to various research partners and vendors for carrying out specific tasks of the project.

GRID SOFTWARE ENVIRONMENT

Central to NAREGI's goals and the five-year plan is research and development of a scaleable grid software environment that can support a variety of real world applications with distributed computing resources. The strategic approach NAREGI is taking to accomplish this complex task is one of "divide and conquer" by having six working groups (work packages), each focusing on a thematic area. For example, Dr. Matsuoka leads a group of researchers focusing on lower and middle-tier middleware for resource management. It basically covers most of the grid's administrative functions such as scheduling, accounting, auditing, and user-oriented information services. The goal here is not to develop entirely new software but to build on existing international efforts such as Unicore, Condor and Globus, with innovative pieces to enhance functionality and to ensure interoperability. Similarly Dr. Miura leads two working groups, one on user-level grid tools and Problem Solving Environments (PSE), and another on packaging and configuration management. Other working groups focus on grid programming models, network measurement and security, and grid-enabling tools for applications. These working groups constitute the technical workhorses of the project. The overall objective for the NAREGI's software environment is for the grid to be widely and easily usable. The software product and services from these work packages must easily enable the execution, linkage, and coordination of the applications, computational modules, data, and other resources interconnected by the network (Super SINET).

An Application Focus on Nanoscience

NAREGI's research is driven by an overriding focus on "Nanoscience and Technology Applications." Working with the Computational Nanoscience Center of the Institute for Molecular Science (IMS) and the Institute for Solid State Physics, AIST, and several universities (Tohoku, Kyoto, TiTech, Osaka, Kyushu) and industries (Materials, Nano-scale devices), NAREGI is developing a grid testbed in which to develop and evaluate the grid middleware tailored to the computational needs of nanoscience research and applications. The various local computer clusters and computational grids are being connected (Super SINET) to achieve an expected 17 Tflop/s performance in Phase 1 of the experiment. A variety of research topics and groups are being targeted: functional nano-molecules (carbon nanotubes, fullerene, etc.), nano-molecule assembly (bio-molecules, etc.), magnetic properties, electronic structure, molecular system design, and nano-system simulation.

While it is still too early to assess the significance of this testbed, NAREGI's focus on nanoscience application marks a major Japanese effort in harnessing high-end computing technologies with a national

imperative other than environmental sciences (as in the case of the Earth Simulator). This is a bold and risky strategy, but perhaps necessary one in order to move forward beyond the Earth Simulator mentality.

RELATIONSHIP TO OTHER PROJECTS

As discussed, NAREGI is one of the several grid projects funded by the Japanese government. When asked to comment on the relationship between NAREGI and the ITBL (Information Technology Based Laboratory) project, also visited by the Panel earlier, Dr. Miura indicated that ITBL was a production-oriented program, driven by special purpose applications. NAREGI, on the other hand, is more R&D oriented. ITBL was launched purely as a domestic project, but NAREGI has a significant international dimension and intends to be a contributor to grid technology across scientific and national borders.

Dr. Miura also explained that there is a separate Business Grid project funded under the Ministry of Economy, Trade, and Industry (METI). Business Grid is more data-oriented, aimed at developing grid technologies to enable business transactions and commerce applications. NAREGI, however, is entirely designed to support high-end scientific applications.

Observations – Implications for HEC

NAREGI is clearly a major grid project, aimed at R&D in grid middleware and related technology. It is also a serious national initiative on high-end computing as indicated in NAREGI's goal to achieve a target performance of 100 Tflop/s by 2007. In a couple of years, even with a moving target, it could be a serious contender for a spot in the top 20 among the Top500 LINPACK performers [3] The initiative appears to be well-funded with a first year budget of $17 million (plus $47 million in computer resources). Over the five-year project's lifetime, the total budget could reach roughly $130 million, a sizable sum in the HEC context (about 25% of the Earth Simulator's budget for its development).

Another important distinct feature of NAREGI is its emphasis on software research and applications, a departure from traditional HEC R&D in Japan. Japan's superior capabilities in consumer electronics and manufacturing make computer hardware a natural focus in pursuing a leadership role in HEC, often at the expense of software. By setting their sights high on software and middleware tailored to complex, heterogeneous computing environments, NAREGI announces to the world that it intends to be a serious contender as a software innovator. Dr. Miura and Dr. Matsuoka are two strong leaders, well-trained and respected in HEC, but they clearly have their work cut out for them in the coming years.

REFERENCES

An overview of the National Research Grid Initiative (NAREGI), slide presentation by the host; an updated version is also available as presentation at the National Research Grid Initiative of Japan (NAREGI) HPC Asia 2004 Workshop, July 21, 2004.

"National Research Grid Initiative: Grid software infrastructure research and development." 2003. National Institute of Informatics, October 2003. Handout at visit.

Top 500 Supercomputer Sites. 2004, <http://www.top500.org> Last accessed February 23, 2005.

Site: **NEC Corporation**
5-7-1 Shiba, Minato-ku, Tokyo 108-8001
http://www.nec.com

Date visited: April 2, 2004

WTEC Attendees: J. Dongarra (Report author), R. Biswas, K. Yelick, Y.T. Chien

Hosts: Tadashi Watanabe, Vice President, Computer Platform Business Unit,
Fifteen members from Computers Division, 1st Computers Software Division, and HPC Marketing Promotion Division

BACKGROUND

NEC was established in 1899. NEC Corporation is a leading provider of high-value solutions in IT, networks, and electron devices dedicated to meeting the specialized needs of its customers. NEC Corporation employs approximately 143,000 people worldwide and saw net sales of ¥4.9 trillion (approx. US$47 billion) in fiscal year 2003-2004. For further information, please visit the NEC home page at: http://www.nec.com.

The figure below is a timeline history of high-performance computers at NEC. For high-performance computing NEC has two main product lines:

- SX series
  - Specialized high-performance parallel-vector architecture
  - Targeted at the high end of scientific computing
- TX7 IPF server series
  - Commodity architecture
  - First vender who supports a 16-way SMP based on IA64 Merced

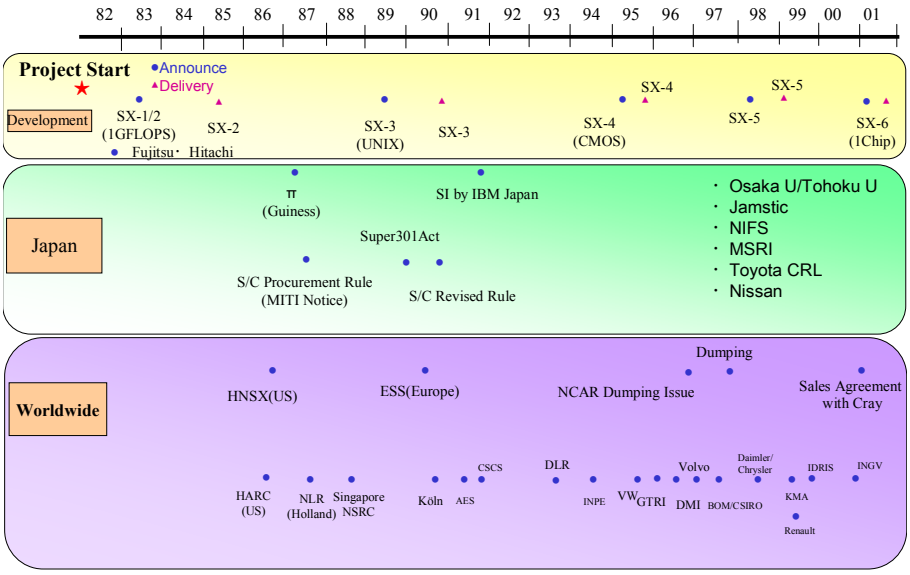


Figure B.9. History of NEC Supercomputers (Courtesy NEC)

REMARKS ON THE TX SERIES

The TX7 series is offered in four models, of which we only discussed the two largest. The TX7 is another of the Itanium-2-based servers that recently appeared on the market. The largest configuration presently offered is the TX7/i9510 with 32 1.5 GHz Itanium-2 processors. NEC already had some experience with Itanium servers offering 16-processor Itanium-1 servers under the name AzusA. So, the TX7 systems can be seen as a second generation. The processors are connected by a flat crossbar. NEC still sells its TX7s with the choice of processors that Intel has: 1.3, 1.4, and 1.5 GHz processors with L3 caches of 3—6 MB depending on the clock frequency. The figure below illustrates some of the features of the TX7.

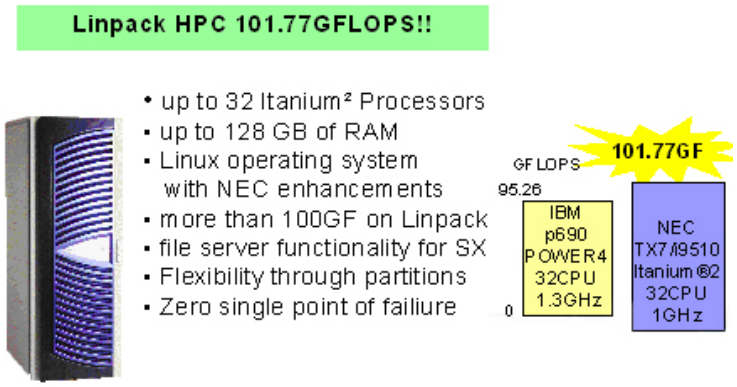


Figure B.10. TX7 Itanium² Server (Courtesy NEC)

Unlike the other vendors that employ the Itanium-2 processors, NEC offers its own compilers including an HPF compiler which is probably available for compatibility with the software for the NEC SX-6 because it is hardly useful on a shared-memory system like the TX7. The software also includes MPI and OpenMP. Apart from Linux HP-UX is also offered as an operating system that may be useful for migration of HP-developed applications to a TX7.

Measured Performances

Results for a 24-frame SX-6/192M24 with 192 processors attained 1,484 Gflop/s, an efficiency of 97%. The size of the linear system for this result was 200,064.

NEC SX-7

The main feature of the SX-7 is that up to 32 CPUs are connected to a maximum 256 GB large capacity shared memory in a single-node system. The ultra-high data transfer speed of maximum 1130.2 GB/s between CPU and memory has also been realized. This is 4.4 times faster than the existing SX-6-models. In a 64-node multi-node system, large capacity memory of up to 16 TB can be configured, and the total data transfer of maximum 72 TB/s speed between CPU and memory can be achieved. Moreover, SX-7 achieves a maximum 18 Tflop/s of vector performance in a multi-node system.

Both have a vector processor with one chip.

- *SX-6*: 8 Gflop/s, 64 GB, processor to memory (8*8 * 4 streams = 256GB/s from memory), between nodes 16 GB/s xbar switch
- *SX-7*: 8.825 Gflop/s, 256 GB, processor to memory 1130 GB/s (8.825*8 * 16 streams = 1130GB/s from memory)

They both use 0.15µm CMOS technology

Remarks on the SX series

The SX-6 series is offered in numerous models but most of these are just smaller frames that house a smaller amount of the same processors. We only discuss the essentially different models here. All models are based on the same processor, an eight-way replicated vector processor where each set of vector pipes contains a logical, mask, add/shift, multiply, and division pipe. As multiplication and addition can be chained (but not division) the peak performance of a pipe set at 500 MHz is 1 Gflop/s. Because of the eight-way replication a single CPU can deliver a peak performance of 8 Gflop/s. The vector units are complemented by a scalar processor that is four-way super scalar and at 500 MHz has a theoretical peak of 1 Gflop/s. The peak bandwidth per CPU is 32 GB/s or 64 B/cycle. This is sufficient to ship eight 8-byte operands back or forth and just enough to feed one operand to each of the replicated pipe sets.

It is interesting to note that the peak performance of a single processor actually has dropped from 10 Gflop/s in the SX-5, the predecessor of the SX-6, to 8 Gflop/s. The reason is that the SX-6 CPU is now housed on a single chip, an impressive feat, where in the former versions of the CPU multiple chips were always required. The replication factor, which was 16 in the SX-5, had to therefore be halved to eight.

The SX-6i is the single CPU system, but because of single-chip implementation, is offered as a desk side model. Also a rack model is available that enables housing two systems in a rack but there is no connection between the systems.

In a single frame of the SX-6A, models fit up to 8 CPUs at the same clock frequency as the SX-6i. Internally the CPUs in the frame are connected by a one-stage crossbar with the same bandwidth as that of a single CPU system: 32 GB/s/port. The fully configured frame can therefore attain a peak speed of 64 Gflop/s.

In addition, there are multi-frame models (SX-6/*x*M*y*) where $x = 8,...,1024$ is the total number of CPUs and $y = 2,...,128$ is the number of frames coupling the single-frame systems into a larger system. There are two ways to couple the SX-6 frames in a multi-frame configuration: NEC provides a full crossbar, the so-called IXS crossbar to connect the various frames together at a speed of 8 GB/s for point-to-point unidirectional out-of-frame communication (1024 GB/s bi-sectional bandwidth for a maximum configuration). For the IXS crossbar solution, the total multi-frame system is globally addressable, turning the system into a NUMA system. However, for performance reasons it is advised to use the system in distributed memory mode with MPI.

The technology used is CMOS. This lowers the fabrication costs and the power consumption appreciably.

For distributed computing there is an HPF compiler, and for message passing, an optimized MPI (MPI/SX) is available (NEC did their own version of MPI). In addition, for shared memory parallelism, OpenMP is available.

OBSERVATIONS

The NEC group we spoke with was committed to high-performance vector computing. NEC believes that the development of the high-end vector computers is a technology driver for other components. They have sold more than 600 SX systems. Within the States only the Artic Regional Supercomputer Center has their product.

Dr. Miyoshi, the visionary for the Earth Simulator (ES), insisted that the ES effort use HPF. This is one of the reasons NEC has invested in HPF technology. They fully support MPI 1 and 2. There, the standard collection of compilers is Fortran, C, and C++. They have debugging tools available from Vampir and use TotalView. Their Fortran and C compilers use different optimization strategies. Starting with the NEC SX-3, Unix was used.

REFERENCES

Top 500 Supercomputer Sites. <http://top500.org> Last accessed February 23, 2005.

Site: **National Institute for Fusion Science (NIFS)**
322-6 Oroshi-Cho, Toki (near Nagoya)
Gifu, 509-5292 Japan
http://www.nifs.ac.jp/

Date Visited: March 29, 2004

WTEC Attendees: P. Paul (Report author), J. Dongarra

Hosts: Professor Masao Okamoto, Director, Theory and Computer Simulation Center
Professor Shigeru Sudo, Deputy Director-General of NIFS
Professor Seiji Ishiguro, staff scientist
Professor Rituko Horiuchi, staff scientist

BACKGROUND

NIFS is one of two major fusion research centers in Japan, the other being JAERI. Its mission is to operate the Large Helical Device (LHD), a Stellerator-type fusion device with superconducting magnetic field coils, and to study the characteristics and behavior of high-performance plasma. The Theory and Computer Simulation Center (TCSC) operates an NEC SX7/160 M5 computer for the simulation of complex plasma effects in Stellerators, Tokamaks and in geosciences problems. NIFS also has a large post-graduate education mission, involving about 50 graduate students at NIFS itself (Sokendai) and in connection with Nagoya University.

NIFS used to be funded by Monbusho (whereas JAERI was the STA-funded laboratory). Now funding for NIFS comes from the combined agency Ministry of Education, Culture, Sports, Science and Technology (MEXT). NIFS has a staff of about 250 and an annual budget of about $100 million. The institute is grouped into the Department of Engineering and Technical Services, Department of the LHD Project, Theory and Computer Simulation Center, Fusion Engineering Research Center, Data and Planning Center, Safety and Environmental Research Center, and Computer Center.

As part of the general reorganization of national research facilities and national universities beginning with the new fiscal year on April 1, 2004, NIFS will become an autonomous institution. It will also join a group of four other laboratories to form a five-lab integrated unit. Thus, at the time of our visit, there was substantial uncertainty about the future relationship with the government and with the budget flow.

The focus of this visit was the TCSC and its components, including the Super Computer System.

Presentation of the LHD

Professor Sudo gave an overview of the LHD. Its mission is to produce high-performance plasma. It produced its first plasma in 1998 and has since reached electron and ion temperatures of 10 keV, stored energies of over 1 MJ and reached maximal confinement times of 360 msec (at lesser temperatures). Its data provides the basis for the simulations performed by the TCSC. Since the field configuration has no axial symmetry the theoretical simulations come from three-dimensional solutions of the Magneto-hydrodynamic equations. The LHD is a large facility with a huge array of data taking stations, all in very new buildings

Presentation of the TCSC

Masao Okamoto gave an overview over the mission, work, and accomplishments of TCSC. He has been director for only about a year (he replaced Prof. Tetsuya Sato who now is the director of the Earth Simulator). The mission of TCSC centers on fusion science, complex phenomena (such as plasma turbulence) and visualization/virtual reality. The new NEC SX-7/160M5 computer is used to solve the MHD equations in three dimensions and to include particle flow. A three-dimensional code FORTEC that was developed ~20 years ago is used to compute neoclassical transport with high accuracy. This code was

successfully vectorized to run on the new NEC computer. A toroidal gyrokinetic Vlasov code is under development for Tokamak simulations. The amount and accuracy of simulations (when compared to experiments) of complex plasma phenomena, including a burning plasma and particle transport, was impressive. Japan will push multilayer model simulations to understand complex plasma phenomena.

Presentation of Super Computer System

Professor Ishiguro gave an overview of the super computer system, followed by a tour of the facility. TCSC has a staff of about 20, with 13 dedicated to simulations. TCSC obtained its most recent supercomputer only about 15 months ago. The parallel-vector processor SX-7/160 M5 has a total memory of 1280 Gbytes, a peak performance of 1.412 Tflop/s, with five nodes and 32 PE/node. Memory per node is 256 GBytes, peak performance per PE is 8.83 Gflop/s with a data transport rate between nodes of 8 GBytes/s. Its LINPACK performance is 1.378 Tflop/s = 97.54%. The computer is leased, at an annual cost of ~$10 million. With its capabilities it rates #81 on the Top500 list.

Software uses mainly high-performance Fortran (HPF), with only a few applications using MPI. For its 3D MHD simulation code with periodic boundary conditions, it performs with 32% efficiency (452.2 Gflop/s) after NEC optimized the code. Its plasma Particle-In-Cell (PIC) simulation code with periodic boundary conditions tracks 8.6 million particles with 15.8% efficiency (224 Gflop/s), after the program was optimized over three to four months. For Monte Carlo simulations with a parallelized numbers generator it runs 32 million particles through 8000 steps using five nodes, 160 PEs with 25% efficiency (358 Gflop/s).

The computer is mainly used by in-house staff and graduate students from Sokendai and Nagoya University as a capability machine. Most of the time there are only a few jobs running concurrently. About 20% of the time the entire machine is used for one job only. Outside users are screened by a collaboration committee. Speed is more important than memory, for their applications. The vector architecture is preferred for their needs, with the large common memory being an essential point. Their previous machine was also from NEC (an SX-5); however, they wonder if the next machine would be another parallel-vector processor or rather, because of cost and space consideration, a parallel-scalar machine. There was no interest in clusters.

Under normal conditions they expect a new machine in five to six years. However, with the imminent reorganization they are uncertain whether this tradition will be maintained. However, we heard at several other meetings in Japan that labs for whom high-end computation is mission critical can expect new computers at the five-six year interval rate. Since the machine is leased renewal does not require a large budget bump-up.

A live demonstration was given of the virtual reality system CompleXcope based on a “CAVE” using the software and hardware from the University of Illinois. This was an impressive visualization of plasma conditions and particle flow in the LHD.

General Observations

NIFS is a large and modern laboratory with an excellent supercomputer facility and an outstanding computational and theoretical staff. The simulations of complex plasma behavior and particle flow were impressive. Together with the University of Nagoya, NIFS for the most part appears a self-sufficient institution. The laboratory produces its own software, almost exclusively using HPF. There is no interest in C, which is a fundamental difference with the U.S., where HPF has been largely abandoned. They did not use the Earth Simulator. There seemed little interest in grid computing although TCSC will undoubtedly be connected to the all-Japan backbone SINET. They have not yet started seriously to define their next supercomputer. It will have to be a capability machine with higher speed. Although they prefer vector architecture, there was doubt that that would be practical and cost-effective. There was no discussion of a national strategy regarding high-end computing. Over the next year NIFS will define its role as a partner in the new five-institution consortium within the general reorganization of Japanese research institutions. The choice of the next computer will probably depend on the details of NIFS’s position within that consortium, since all funds will flow from the consortium, rather than directly from the government to NIFS as it did in the past.

Site: **Institute of Physical and Chemical Research (RIKEN)**
Advanced Center for Computing and Communication (ACCC)
2-1 Hirosawa, Wako, Saitama-ken 351-0198, Japan
http://www.riken.go.jp/engn/

Date Visited: April 1, 2004

WTEC Attendees: P. Paul (Report author), A. Trivelpiece, S. Meacham, Y.T. Chien

Host: Dr. Ryutaro Himeno, Director of ACCC
Dr. Yorinao Inoue, Executive Director, RIKEN
Dr. Hideto En'yo, Head of Radiation Laboratory

BACKGROUND

RIKEN is the oldest and perhaps the most prestigious research organization in Japan. Its main site at Wako was built in 1963. Famous names, such as Nishina, Yukawa and Tomonaga, are associated with RIKEN. Today its research at the Wako site is grouped into four main units: Discovery Research Institute (pecial research projects, nuclear physics); Frontier Research System (carefully selected frontier research groups), Brain Science Institute (since 1997); and Advanced Center for Computing and Communication (ACCC). In addition RIKEN has several branch sites: Tsukuba Institute (life sciences) Harima Institute (operates the Spring-8 synchrotron light source, together with JAERI, high-throughput protein factory), RIKEN-BNL Research Center (High-energy nuclear physics), Yokohama Institute (plant science, genomics, allergy and immunology) and the Kobe Institute (developmental biology)

Much of the recent research thrust of RIKEN is in the life sciences, with material science second. Its total staff numbers approximately 2,400 and the budget is about $800 million. It has over 2,000 visitors a year and over 1,000 students. In 2003 as part of the general reorganization of Japanese science labs and universities, RIKEN became an Independent Administrative Institution. Its Government funding comes from MEXT.

The focus of this visit was to discuss strategic plans for computing at the ACCC at RIKEN.

PRESENTATION BY DR. HIMENO

The ACCC has an annual budget of $14 million. It used to operate a Fujitsu VPP700E/160 vector machine, which is completely inadequate today. In particular, applications have shifted from vector-type to scalar type, such as bioinformatics and computational chemistry. Thus in 2000 ACCC acquired RICE (RIKEN PC Cluster Environment), a Pentium 4/64 (1.7 GHz) cluster from Fujitsu (PRIMERGY CL460J). It contained 10 racks at 8 CPUs/rack, and a Myrinet 2000 network that had a LINUX operating system and achieved a peak performance of 218 Gflop/s with parallelization by domain decomposition. Based on these results, and despite the concern over the large costs of large-scale PC clusters, ACCC decided to make such a cluster their next main computer system using MyrinetXP (2 Gb/s) and InfiBand (8 Gb/s) as network connectors. However, they saw a need for a second computer for programs that are not parallelized and that need more than 2 Gb shared memory.

Thus they arrived at their current system: It consists of 2048 CPUs (Pentium Xeon 3.06 GHz) on 1024 nodes (2 CPU/board) and has a peak performance of 12 Tflop/s. It has an aggregate memory of 3T and HDDD capacity of 140 TBB that are divided into five units: 512 nodes are combined in a large cluster for large jobs. These are connected by an 8 Gb/s InfiniBand network. The other 512 nodes are divided equally into groups of 128 nodes each internally connected with a 2 Gb/s Myrinet. These are intended for experimental data processing, bioinformatics, interactive use, and computational chemistry. The latter sub-cluster is connected to 20 Grape accelerator boards optimized for molecular dynamics. These work at 64 Gflop/s per board or 1.2 Tflop/s total. We were shown the installed system bought from Fujitsu.

The system is completed with a new NEC SX-7/32 vector machine with a peak performance of 283 Gflop/s and a memory capacity of 256 GB for large-memory jobs.

This system is optimized to serve a diverse range of users. An outside user will send this job to the system, which will rout it to the computer system that is most appropriate for it.

Dr. Himeno suggested that this concept could expand further by specialization among laboratories: One specializes on clusters, another on a vector machine, a third on a scalar machine, etc. These could be connected by the national computing grid (which is in preparation) and each job could then be automatically routed to the computer type best suited to the problem. The demands on the grid would be quite low since the entire job is done in one place.

Observations

Such a grid scheme would obviously be very advantageous to RIKEN with its far-flung institutes. RIKEN has been very adventurous in developing specialized chips for special jobs. The Yokohama Institute is developing a MDGrape-3 (a.k.a. Protein Explorer) at 1 Pflop/s for protein modeling and it will be completed in 2006. The Wako Institute already runs a MD-Grape-2 at 78 Tflop/s for molecular dynamics. RIKEN has paid for the development of the QCDOC chip at Columbia University specialized for lattice gauge calculations, which turns out to be also very good for MD, and is getting a 10 Tflop/s (peak) version of QCDOC at RIKEN-BNL.

Five national research laboratories, the largest being RIKEN, JAERI (Japanese Atomic Energy Research Institute) and JAXA (Japanese Aerospace Exploration Agency), are collaborating in the ITBL (Information Technology Based Laboratory) project funded by JST (now replaced by MEXT at a rate of $10 million/year for five years), to develop middleware and other software for a large, fast user grid. This project is in its third year. It will make use of the SuperSINET and will connect the laboratories, universities and industry.

RIKEN is clearly enthusiastic about clusters but felt that 1,000 CPUs was the top number that could be connected with the currently available network. There was little discussion of the Earth Simulator, although it could, of course, play the role of the vector processor in the nationally distributed effort.

Site: **Research Organization for Information Science and Technology (RIST)**
2-2-54 Nakameguro, Meguro-ku, Tokyo 153-0061
http://www.tokyo.rist.or.jp

Date Visited: April 1, 2004

WTEC Attendees: R. Biswas (Report author), J. Dongarra, K. Yelick, M. Miyahara

Hosts: Hisashi Nakamura, Director, Division of Computational Science and Technology
Akio Kitsunezaki, Executive Director
Syogo Tejima, Researcher, Nanotechnology
Mikio Iizuka, Researcher, Applied Nano Devices
Atsushi Miyauchi, Researcher, Quantum Calculations
Kazuo Minami, Researcher, Nuclear Fusion
Takashi Arakawa, Researcher, Climate Modeling
Takamichi Arakawa, Researcher, HPC Software Technology
Kouji Makino, Researcher
Giri Prabhakar, Researcher

BACKGROUND

The Research Organization for Information Science and Technology (RIST) is a non-profit, public service organization working for the development and utilization of computational science and engineering technology for several nationally important application areas. RIST's predecessor, the Nuclear Energy Data Center (NEDAC), was founded in 1981 in Tokai-mura to promote the development and utilization of nuclear computer codes. The scope of work was later expanded to incorporate the support of integrated computational science research and development through cooperative interactions with other national research organizations. This ranged from Earth science projects such as climate and weather modeling to computational nanotechnology. RIST was established on April 1, 1995, under the strong support of the former Science and Technology Agency (STA) of the Japanese government, while the former NEDAC was changed to RIST Tokai. RIST is currently under the jurisdiction of the Ministry of Education, Culture, Sports, Science, and Technology (MEXT).

Hajime Miyoshi, the force behind the Earth Simulator (ES), was the first Deputy Director-General of RIST. While there, Dr. Miyoshi realized that even powerful supercomputers such as the Numerical Wind Tunnel (NWT) at the National Aerospace Laboratory (NAL) were incapable of satisfying the computational requirements of global weather and ocean circulation simulations. These areas were becoming important at that time as global environmental issues captured international headlines (Kyoto Protocol). He therefore decided to develop the ES to serve as a powerful compute engine for Earth science simulations. In 1997, he organized the ES Research and Development Center with financial support from Japan Marine Science and Technology Center (JAMSTEC), Japan Atomic Energy Research Institute (JAERI), and National Space Development Agency (NASDA), all of which were under the STA. He left RIST in April 1997 to become the Director of the ES R&D Center.

Currently, RIST employs about 100 people. We visited the Division of Computational Science and Technology in RIST-Tokyo, and were hosted by their Director, Hisashi Nakamura. Their budget is $3-5M per year and they have a staff of about 30.

RIST PRESENTATIONS

Hisashi Nakamura, Director of the Division of Computational Science and Technology, provided an overview of RIST, its history, and current activities. RIST's mission has been to serve as a catalyst or starter for computational science and engineering efforts ranging from climate modeling to nanotechnology with the

primary focus on large-scale simulations using the Earth Simulator (ES). RIST played an important role in the development of the ES. Between 1995 and 1997, not only did Hajime Miyoshi propose a conceptual design of the ES while at RIST, but other researchers also collaborated with Fujitsu and NEC on parallel operating systems development. Since 1997, at least 21 large simulation codes (10 for climate, 11 for seismology) have been developed, transferred to scientists, and run on the ES. Currently, such Earth science applications use up 70 % of the time allotted to the ES.

At present, RIST uses about 15% of the ES's time to conduct research on cutting-edge technology. One of the major goals is to find a new path for nanotechnology, given the recent Japanese interest in this area. RIST is primarily looking at simulation-driven approaches for new material design, self-assembly processes, properties of nanomaterials, nano electro mechanical systems (NEMS), etc. For this reason, they have organized multi-disciplinary research groups, such as the carbon nanotube (CNT) simulation group (with Prof. Morinobu Endo at Shinshu University, and other academic and industrial partners). H. Nakamura cited a computational nanotechnology application that ran for 1.4 hours at 7.1 Tflop/s on 3480 processors of the ES, demonstrating almost linear speedup. This is more that 18,000 times faster than a single 2.5 GHz Pentium 4. RIST continues to have close interaction (in terms of code development, porting, and tuning) with people at the ES Center and Frontier Research System for Global Change (FRSGC), both of which are located at JAMSTEC. They have extremely low overhead; furthermore, using the ES is free to RIST researchers. The organization was much larger during 1997-2002; presently they are supporting ES's operation of Earth science codes and demonstrating its capabilities (in areas such as nanocarbons, nanobio, quantum calculations, nuclear fusion, etc.). RIST believes that modeling and simulation drive new science and technology.

H. Nakamura concluded his presentation by giving some examples of large-scale modeling and simulation in Earth sciences: typhoons, ocean surface temperature, clouds, and seismic wave propagation. For instance, he discussed the JMA cloud simulation code that uses 1km non-hydrostatic (NHS) models to successfully reproduce cloud bands extending southeast from the base of the Korean Peninsula over the Sea of Japan. Several cloud streets and other clouds were also very accurately simulated. He showed a fault model of the underground structure of southwest Japan, and a simulation of seismic wave propagation. There is on-going work on coupled atmospheric and ocean models (called Fu-jin); and multi-scale simulations incorporating features such as downbursts and NHS cloud models (local), fine clouds, oil spills, and typhoons (regional), and ocean circulation models (global). He discussed GeoFEM: their multi-purpose/multi-physics parallel finite element framework for solid Earth modeling. This software is downloadable directly from the RIST site.

After the overview talk, several RIST researchers gave brief presentations on specific research topics. The first presentation was by Syogo Tejima who described his work in nanotechnology simulations. Nanotechnology is now a major application area in Japan as it is in much of the world. In FY2004, almost 15% of the ES cycles are being devoted to computational nanotechnology. This work is focused on carbon nanotubes and fullerenes as novel materials and candidates for potential applications. There are three primary objectives with this work: design of innovative nanomaterials with certain desired properties; obtaining fundamental properties in nanoscale matter; and nano-applications such as nanoreactor encapsulating obstacles.

The second talk, by Kouji Makino, was on quantum combinational chemistry simulations. Here the objective is to find new nanocarbon structures by isomerization using generalized Stone-Wales (GSW) rearrangement and C_2 loss. Isomerization is extremely compute intensive, so is the follow-on classification step. The maximum computational load through k steps is $(6n)^k$. Since k is typically larger than 20 and assuming n>200, the simulation clearly needs very large-scale computers. The long-term goal is to find a path to make new nanodevices and nanorobots from the many generated nanostructures by self-organization. There is no parallel code at this time; the researchers are just beginning to develop these algorithms.

Mikio Iizuka then presented his work on nanodevice simulations, currently focused on those of high-temperature superconductors (HTS). The goal is to develop effective sources of continuous terahertz waves, which have many applications in biotechnology and information sciences. A large multi-scale parameter space has to be found for the best operating conditions of the device. At present, one run of a 2D model

requires about 12 hours on 20 nodes of the ES. A realistic 3D model will easily require more than 100x the computational resources available on the ES. Simulations of HTS nanodevices have revealed a new mechanism of terahertz wave emission and that the frequency is tunable by altering the applied current. M. Iizuka believes that the ES enabled both theorists and experimentalists to develop HTS nanodevices, but bigger and faster computers are required to make progress. Advances are also required in the software arena, particularly reliable and robust tools for load balancing, interprocedural dependence analysis, high-level vectorization, and performance analysis. Future research will focus on quantum electromagnetic simulations with the goal of discovering new phenomena (such as low-energy collective and elementary excitation) in substances when bombarded with terahertz waves.

The fourth talk by Atsushi Miyauchi was on quantum calculations. He gave an overview of why quantum computing is interesting: quantum parallelism (coexistence in many worlds), reversible operations (no heat generation or energy consumption), and no cloning (impossible to copy or forge). Fundamental algorithms range from simple estimation (Deutsch-Jozsa oracle) through Fourier transforms to versatile, multi-purpose search algorithms. Potential applications can be found in the fields of physics, cryptography, communication, theoretical computer science, and many others. RIST activity in this area is minimal; current focus is on applications to physical problems (quantum many-body problems and/or mesoscopic quantum phenomena in quest of exotic materials with superconductivity, superfluidity, etc.), new high-level language design for quantum computers, and extending non-standard quantum computing methods (adiabatic, continuous variable, and geometric).

Kazuo Minami then presented his work on performance optimization for nuclear fusion. His particle code had two performance bottlenecks: a recurrence process in an electrical current calculation, and severe workload imbalance. He improved parallel performance (by a factor of three) by using a reordering method that removed the recurrence relationship, and by using multi-level parallelism for better load balance. His eigenvalue calculation code showed poor performance in the block tridiagonal matrix solver routine. He improved performance by parallelizing the cyclic reduction method and by increasing the ratio of arithmetic operations to loads/stores. The resulting code achieves 4 Gflop/s (50% of peak) on one node of the SX-6 (building block of the ES). K. Minami is also optimizing other codes for superconductivity calculations, nanodevice simulations, and molecular dynamics (MD) simulations. The MD code achieved 7.1 Tflop/s on 435 nodes of the ES (almost 26% of peak).

The next talk was by Takashi Arakawa who presented RIST's activities in weather and climate simulations. He claimed that RIST had initiated most of the original research in this area and parallelized various models, examples being Fu-jin (a parallel framework for coupled atmospheric and ocean models), GNSS (global non-hydrostatic simulation system), CReSS (cloud resolving storm simulator), and coupled climate model for predicting global warming. The last three were ported to the ES and are being used regularly by Earth scientists. GNSS is a semi-global (45N to 45S) code that investigates cloud activity in tropical areas. A 20km simulation test run has been completed but the ultimate target is to perform a 2km simulation. CReSS performs large-scale numerical simulations of clouds and mesoscale storms associated with severe weather systems, and is the only Japanese cloud model open to the public. The code is a hybrid of MPI, OpenMP, and vectorization; however, pure MPI (plus vectorization) is more effective than MPI+OpenMP on the ES.

The final presentation was by Takamichi Arakawa on high-performance computing (HPC) environments. The objective is to provide HPC middleware (HPC-MW) between hardware and systems software on one side, and simulation, analysis and other tools used by scientists and engineers on the other. HPC-MW is a library-based infrastructure for developing optimized and reliable simulation codes that make efficient use of the finite element method. Examples of HPC-MW capabilities include matrix assembly, linear solvers, adaptive mesh refinement, dynamic load balancing and graph partitioning, volume rendering and visualization, and multi-component coupling. It will make hardware and software details transparent to users so that code development, reliability, porting, tuning, and maintenance become more efficient and error-free.

ANSWERS TO OUR QUESTIONS

RIST did not specifically answer our list of questions on programming paradigms, program development tools, systems software, hardware architectures, and applications; but did allude to them during their presentations. One of their most important observations was that vector architectures, like the ES, are most suitable for RIST applications for two reasons: they return much higher sustained performance (as a percentage of peak), and most RIST researchers are more familiar with vector codes.

REMARKS

The RIST personnel were very proud and supportive of the ES as they very well should be. After all, Hajime Miyoshi started the design of the ES while he was at RIST. They believe that the ES has led to new scientific discoveries, even though we saw little concrete evidence. They also think that it has raised the hopes and interests of a younger generation of scientists. One interesting comment made by Hisashi Nakamura, the Director, was that RIST's use of the ES is free. This may not remain true after the Japanese government's massive privatization efforts began on April 1, 2004. RIST is a fairly big user of the ES and they do not have any in-house supercomputers. On the other hand, the ES R&D Center is being asked to gradually acquire its own funding.

It was also surprising to hear that RIST has no collaboration with JAXA. Nakamura does not know if there will be a follow-on to the ES and, if so, where the funding will come from. However, from the users' perspective, a considerable gap exists between available resources and application requirements. There are many examples of applications with 100 Tflop/s compute requirements; in fact, computational nanotechnology applications will certainly need petaflop/s-scale computing power. Most scientific applications also require large random access memories that provide high memory bandwidth. Nakamura therefore believes that the high-end computing community needs an ES2 (with computational power at least a factor of 10 times that of the ES). For their part, RIST will continue to do the research and development in computational science and engineering. In his opinion, grids will never be an alternative solution for tera-or petaflop/s-scale capability computing; however, they may be suitable for problems (such as parameter studies and ensemble calculations) that have high-capacity computing requirements.

Site: **Sony Computer Entertainment, Inc.**
2-6-21, Minamiaoyama
Minato-ku, Tokyo, 107-00062 Japan

Date Visited: April 2, 2004

WTEC Attendees: K. Yelick (Report author), J. Dongarra, R. Biswas, Y.T. Chien

Hosts: Mr. Masayuki Chatani, Corporate Executive & CTO and Senior VP for R&D Division
Mr. Teiji Yutake, Vice President of Software Platform Development Department
Mr. Tsutomu Horikawa, Director of Software Platform Development Department
Ms. Nanako Kato, Corporate Communications Department

BACKGROUND

Sony Computer Entertainment Inc. (SCEI) was established in 1993 as a joint venture between Sony Corporation and Sony Music Entertainment Inc. SCEI develops, distributes and markets the PlayStation® (PS1) game console and PlayStation2® (PS2) entertainment system. It also develops and publishes software for both platforms and functions as headquarters for the subsidiary companies. The PlayStation game console project, started in 1990, introduced 3D graphics into the game market and is the number-one selling video game system, with close to 100 million units shipped by December 2003. The PlayStation2 was introduced in 2002 and had shipped 70 million units as of January 2004. The President and CEO of SCEI is Ken Kutaragi, who led the original PlayStation team in 1990. SCEI research and development activities center around the architecture of their four main product lines: PS1, PS2, PSX (adds high-speed user interface), and PSP/NextGen.

The company has a capitalization of ¥1.93 billion. Its fiscal profile for 2003 is as follows:

- Sales and operating revenue: ¥1,015.7 billion
- Operating income: ¥90.0 billion
- Net income: ¥50.7 billion

PLAYSTATION2 HARDWARE

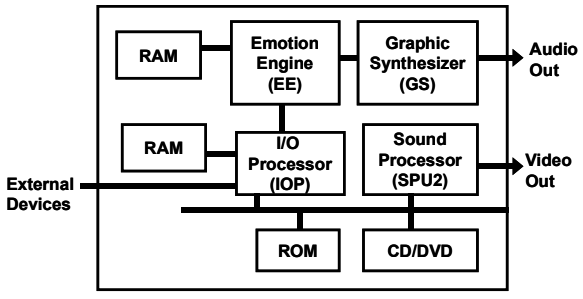


Figure B.11. PlayStation2 block diagram (Steffen 2003)

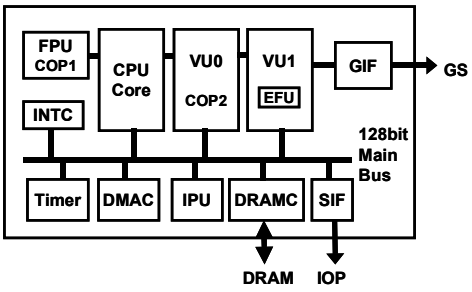


Figure B.12. Emotion Engine block diagram (Steffen 2003)

A block diagram for the PlayStation2 hardware is shown in the left figure. The hardware includes five main components: the Emotion Engine processors, a Graphics Synthesizer, the IO Processor, a Dynamic Sound Processor, and a DVD/CD ROM disc system. The system is specialized for game applications, including support for 3D graphics and audio, with the Emotion Engine chip being the most general-purpose component.

The CPU is a MIPS processor with two coprocessors: the floating point unit (FPU) and one of the vector units (VU0). The second vector unit (VU1) is used with the GIF for geometry processing, and is not directly tied to the CPU through a coprocessor interface. Each of the two vector units contains four FMACs and one FDIV unit for 32-bit floating point. The four FMACs are fully pipelined, so each can complete a fused multiply add instruction (two FP Ops) per cycle in the stead state. The latency on the FDIV unit is seven cycles, and it is not pipelined, so each FDIV can perform 1/7th of a FP Op per cycle. They are used as four-wide SIMD processors, so four FMACs can be completed every cycle. The two VUs contain identical micro architectures, except that the VU1 has an additional piece of hardware known as the Elementary Functional Unit (EFU), which contains another FMAC and FDIV unit. Thus the theoretical peak is broken down as follows:

- 1 FMAC unit has a peak of 2 FP Ops/cycle *300 MHz = .6 Gflop/s
- 1 FDIV has a peak of 1/7 FP Ops/cycle * 300 MHz = .04 Gflop/s
- VU0 can execute 4 FMACs and 1/7 of an FDIV per cycle = (4 * .6 + 0.04) Gflop/s = 2.44 Gflop/s
- VU1 has an additional MAC and FDIV unit: (2.44 + .6 (MAC) + .04 (FDIV)) = 3.08 Gflop/s
- FPU CPU core has a 1 FMAC and 1FDIV: .64 Gflop/s
- Total for FPU + VU0 + VU1: 2.44 + 3.08 + .64 = 6.16 Gflop/s

Observations

A more realistic peak for a code dominated by multiply-accumulate instructions is: 10 MACs * 300 MHz = 6 Gflop/s. Even this would be difficult to achieve in practice, since it requires careful scheduling of the two vector units, the CPU's floating point unit, and the EFU's floating point unit. Using only the FMACs on the two VUs, the peak is 4.8 Gflop/s. As noted above, VU0 can be programmed as a coprocessor to the main CPU, essentially using the instruction set extensions for the CPU. To utilize both VUs in VLIW mode, however, separate code must be downloaded to control their execution.

The Emotion Engine demonstrates the widely held belief that high floating-point performance is possible even in a low-cost system. However, the PS2 as currently built has several limitations with respect to general purpose computing. There is no double-precision (64-bit) floating-point support, and the 32-bit support that is available does not meet the IEEE 754 spec. The 32 MB of memory is not expandable, and the achieved memory bandwidth is 2.4 GB/s, which is roughly 0.4 Bytes/flop. The PCM/CIA bus is only 16bits wide, the 100 MB/s Ethernet is too slow for most computations.

PLAYSTATION2 SOFTWARE

The PlayStation2 can be programmed using one of two development kits. The programming environment originally provided for the system, the DTL-T10000, costs about $10,000 US and is a workstation version of the PlayStation2. Purchase of such a system also involves signing an agreement with Sony that is designed specifically for game developers; this involves a licensing agreement for the sale of games and protection of Sony's intellectual property. The most commercially sensitive part of the system is the DVD/CD encoding tools. The DTL-T10000 runs a 64-bit Linux kernel and comes with Sony's software development tools (called SDEVTC), which consists of a compiler, linker, debugger, assembler, etc. An IDE runs with Windows front-ends. The package also comes with an extensive set of libraries for graphics, sound, networking and device handling, and tools for developing and editing graphics and audio content. It also has hardware probes used for performance-tuning small computational kernels by measuring hardware level activity within the Emotion Engine. Developers can also purchase a CD/DVD Emulator DTL-T14000 that uses a hard disk to test a PlayStation2 CD or DVD image.

The second development platform, the "Linux for Playstation2" system, has many of the same software tools, but for a relatively small investment in hardware. It contains a keyboard, mouse, and internal 40 GB hard drive, along with a connection to support an external display. It is designed as an add-on for an existing PlayStation2 system, rather than a full workstation version of the game hardware. It also has an Ethernet adaptor and Linux distribution that runs on the machine. This Linux development kit currently sells for about

$100 US. The Linux distribution comes with a gecko compiler for the MIPS processor. Programming of the VUs and other hardware is done using an assembly-level interface, using manuals provided in the kit. This development environment does not include support for programming the DVD/CD controllers.

Observations

The programming environment tools and libraries, like the hardware, are specialized for game development. In particular, effective use of the VUs (especially VU1) requires a low-level programming effort, unless one is using the graphics/audio libraries provided in the development kit. There has been some interest in trying to use the PlayStation2 for scientific computing, which involved the purchase of a Linux development kit for each node of a PlayStation2 cluster. So far, the best reported performance result is 1 Gflop/s on a 28x28 dense matrix multiply [Steffen 2003].

REFERENCES

Kunimatsu, A, N. Ide, T. Sato, Y. Endo, H. Murakami, T. Kamei, M. Hirano, M. Oka, M. Ohba, T. Yutaka, T. Okada. 1999. 5.5 Gflop/s Vector Units for Emotion Synthesis. *Hot Chips.* http://www.hotchips.org/archive/hc11/hc11pres_pdf/hc99.s3.1.Kunimatsu.pdf

Kunimatsu, A, N. Ide, T. Sato, Y. Endo, H. Murakami, T. Kamei, M. Hirano, M. Oka, M. Ohba, T. Yutaka, T. Okada. 1999. Designing and Programming the Emotion Engine. *IEEE Micro*: 20-28.

Oka, M. and M. Suzuoki. 1999. Vector Unit Architecture for Emotion Synthesis. *IEEE Micro:* 40-47.

Steffen, C. 2003. Scientific Computing on the Sony PlayStation 2. <http://arrakis.ncsa.uiuc.edu/ps2/> Last accessed February 23, 2005.

Site: **University of Tokyo**
Department of Computer Science
Hongo 7-3-1, Bunkyo-ku,
Tokyo, 113-8656, Japan
http://www.i.u-tokyo.ac.jp/cs/cs-e.htm

Date Visited: March 30, 2004

WTEC Attendees: P. Paul (Report author), R. Biswas

Host: Professor Yoshio Oyanagi, Department chair, in the department office at the U. of Tokyo

BACKGROUND

Professor Oyanagi has been involved centrally in the development and planning of high-end computing (HEC) in Japan. In the 1980s he started lattice gauge calculations in Japan with Iwasaki, Fukugita, Ukawa and others. He then went through the development of several early generations of high-end computers. In 1998 he chaired the interim evaluation committee for the Earth Simulator (ES). Since 1992 he has been seminally engaged in the development of the Grape chip that is the basis of the new Pflop/s Protein Simulator at the Yokohama Center of RIKEN (which won the Gordon-Bell Prize this year). Grape development has now moved out from his department.

As a sign of the increasing weight that is being given to computational science in Japanese universities, the University of Tokyo has just formed a new Graduate School for Information Science and Technology. The Department of Computer Science is one of five departments in this school. During our visit the University was in spring recess, thus Professor Oyanagi was our sole contact.

The focus of this visit was to hear from Professor Oyanagi about the strategic planning for high-end computing in Japan, and his assessment of the role that ES has played and is playing in these plans.

PRESENTATION BY PROFESSOR OYANAGI

The history of supercomputing in Japan over the past 20 years is interesting and filled with accomplishments. It started around 1980 with the lattice gauge project WCDPAX that spawned the highly specialized Grape architecture, and continued with the MITI supercomputer project and the Fifth Generation project that ended in the mid 1990s. Since 1983 the Japanese Government eagerly installed high-performance machines in national laboratories and at seven University Supercomputer Centers. Many computers came and went during this period but the overall trend has been a transition from vector computers to MPPs (Multi Processor), with the exception of the NEC SX6/7 machines. Japan caught on late to the MPPs but this switch is now in full swing. It is notable that no Japanese computer maker failed during this transition. This was due to the fact that Japanese HEC manufacturers (NEC, Fujitsu, Hitachi) were/are also large chipmakers. Japanese vector machines were designed as extensions of mainframe computers with easy usability as key.

The government encouraged new architecture with special applications and commercialized them. Examples are: the Numerical Wind Tunnel that led to the VPP500; the Earth Simulator that led to the SX-6 and Grape, MDM (Molecular Dynamics Machine) and eHP. However, over the past ten years Japan has fallen behind. In 1994, 12 of the top 20 machines were Japanese, by 1997 (after ASVI started in the U.S.) only five were in that grouping; by Feb. 2003 only two machines were left. However, the next Top500 list will contain several big cluster machines from Japan.

He sees the future for hardware divided into two groups:

- General purpose machines from the established manufacturers: NEC with its SX6/7 line may continue with vector machines, Fujitsu will go with clusters (VPP, Prime Power), Hitachi will use pseudo-vector MPPs to the SR11000.
- Specialized research machines such as Grape, MDM, EHPC and hybrid multicomputer systems.

In parallel with the hardware support there has been a continuing program for software/applications development over the past five years. Japanese Society for the Promotion of Science (JSPS) sponsored the program "Computational Science and Engineering, from 1997 to 2003" (a total investment of over $30 million). JST supported the ACT-JST projects (a series of three-year projects at $500k each) from 1998 to 2004. In 2001 under the leadership of Professor Tanaka the JST CREST project was begun in support of new high-performance information processing; in 2002 under Professor Doi the JST Strategic and Creative Research Initiative was begun with the theme of "Innovation in Simulation Technology." In 2001 Ministry of Education, Science and Technology (MEXT) started several strategic IT programs: "Development of Strategic Software" headed by Professor Kobayashi, and "Information Science to Deepen the IT" headed by Professor Anzai.

These new programs are an indication that pressure to fund information science and technology in academia has begun to pay off. High-performance resources are now broadly available to Japanese researchers and research computers are influencing industry. In 2000 the IT Strategic Headquarters, headed by the Prime Minister, and the IT Strategic Commission, were formed. IT Strategic focuses on e-commerce, information security, IC card and the digital government. Its backbone will be the 10 Gbit of the SuperSINET, which runs down the spine of Japan. A lot of activity across Japanese institutions addresses the need for this grid for computing and data transfer.

Some Details on Hardware mentioned above

Numerical Wind Tunnel

Dedicated to computational fluid dynamics, built by NAL and Fujitsu, completed by 1993; 140 vector processors with 1.7 Gflop/s, distributed memory and single stage cross bar connection; technology as used in VPP500 machines and its followers.

Cp-pacs

Dedicated to computational physics, developed by U. Tsukuba and Hitachi, completed in 1996; 2048 processors with 600 Gflop/s peak pseudo-vector processing and three-dimensional hyper-crossbar interconnection; technology was used in the SR2201 machine.

Earth Simulator

Dedicated to atmosphere, ocean and materials simulation, manufactured by NEC, 640 nodes with 40 Tflop/s peak; 8 PR per node and 16 GB/node for a total of 10 TB; single stage Crossbar network. In 1998 the Interim Evaluation Committee chaired by Professor Oyanagi ascertained that the design would reach 40 Tflop/s peak and 5 Tflop/s sustained for atmospheric calculations. Some of this technology went into SX-7. Hitachi SR 11000 is based on Power 5 chip from IBM.

Some Details on (mostly) Academic Software Development

JSPS Research for the Future

This is an NSF-type program responding to bottom-up proposals. It lasted from 1997 to 2003 with a total budget of over $30 million.

- Next Generation massively parallel computers at U. Tokyo and U. Tsukuba
- Materials Science Simulation for Future Electronics at ISSP, U. Tokyo
- Global scale Flow Systems at Nagoya U. and Tokyo IT

- ADVENTURE project at U. Tokyo

ACT-JST Advanced Computational Science and Technology

JST has a top-down approach to projects. From 1998 to 2004 each three-year project was budgeted at ~$500k. In 2001, it won six awards in materials science, four in bioscience, four in environmental science, four in earth/space science, three in network development.

JST CREST Program

Begun in 2001, Projects started in 2002. Software certification (Kinoshita, U. Tokyo); Dependable systems (Sakai u. Tokyo); Multimedia coding (Toraichi, U. Tsukuba); Quantum Computing (Muto, Hokkaido U.)

JST "Strategic and Creative Researches"

The following are team projects set at $400k/project over five years.

- Multiphysics simulator by particle method (Tsukagoshi, U. Tokyo, awarded in 2002)
- Hierarchical biosimulator (Doi, Nagoya U., 2002)
- Nanomaterials measurement simulator (Qwatanabe, Tokyo, 2002)
- Advanced radiotherapy (Saito, JAERI, 2002)
- Basic library for large-scale simulations (Nishida, Tokyo, 2002)
- Symbolic numerical hybrid computation (Anai, Fujitsu, 2003)
- Materials Design (Ishida, Tohoku U., 2003)
- Simulation for radiation therapy (Sasaki, KEK, 2003)
- Bone medical simulator (Takani, Osaka U., 2003)
- Molecular orbit on a grid (Nagashioma, AIST, 2003)
- Heart simulator (Hisada, U. Tokyo, 2003)

GENERAL COMMENTS

This highly informative discussion traced a planning curve over the past 25 years until now through the academic computational science sector. However, professor Oyanagi pointed out that computational science still needs to push hard in Japan to be recognized as a new modality of science. Computational science should no longer be regarded as a service but as a research field in its own right

It was emphasized that Japanese researchers prefer not to use complicated machines, which explains the widespread use of Fortran. It was impressive that the Grape-based Protein Simulator is already working and will reach 1 Pflop/s by the end of next year! The commercial version of Grape is the MDM (Molecular Dynamics Machine) and that is also available. Was the Earth Simulator a success? Professor Oyanagi responded: Technically yes, but it is too expensive to serve as the model for the next step. Vector machines may die out. The next-generation supercomputers should be spherical clusters. Several big clusters will come online in Japan soon. For example, the National Institute for Advanced Industrial Science and Technology (AIST) has just installed a 556-processor Evolocity E- II cluster in AIST's Grid Technology Research Center, to join the AIST 11 Tflop/s Supercluster, which is connected to another system that forms Japan's largest distributed computing grid. AIST is interviewed separately in this report.

Site: **Tokyo Institute of Technology**
Global Scientific Information and Computing Center (GSIC)
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8550
http://www.gsic.titech.ac.jp/English/index.html

Date Visited: March 31, 2004

WTEC Attendees: R. Biswas (Report author), J. Dongarra, M. Miyahara

Hosts: Yoshinori Sakai, Director, GSIC
Satoshi Matsuoka, Professor
Takayuki Aoki, Professor

BACKGROUND

The Tokyo Institute of Technology (TiTech) was formally established in 1929, but its roots can be traced as far back as 1881 when the Tokyo Vocational School was founded at Kuramae by the Ministry of Education, Science and Culture. As of May 2003, TiTech had more than 1,100 faculty members and close to 10,000 students. We visited the Global Scientific Information and Computing Center (GSIC) that was established on April 1, 2001, with the merger of the Computer Center and the International Cooperation Center for Science and Technology. The role of GSIC is to support research and education by leveraging leading-edge information technologies, meeting the requirements of its users, and promoting an active and cooperative exchange of knowledge among various organizations both inside and outside of Japan. GSIC employs 12 full-time and visiting professors for two information divisions (Information Infrastructure and Advanced Computing and Education Infrastructure) dedicated to the cooperative advancement of research and development. GSIC also has a Global Scientific Collaboration division that promotes joint international research projects with seven full-time and visiting professors.

Incidentally, our site visit coincided with the last day of the Japanese fiscal year, after which all of the national universities, including TiTech, were to be transformed into independent administrative institutions (IAI). This gives the organizations greater freedom to pursue their own goals but makes them more responsible and accountable for their actions. It remains to be seen how the academic institutions adapt to and take advantage of this new environment in the coming years.

We were hosted by Dr. Yoshinori Sakai, the Director of GSIC. Prof. Satoshi Matsuoka, the lead of the Problem Solving Environment Group, gave a detailed presentation of TiTech's role in Japanese grid activities. He also took us on a tour of GSIC and his computer laboratory. Finally, Prof. Takayuki Aoki gave a description of his group's research efforts to develop accurate numerical schemes and presented the results of simulations performed on TiTech's grid resources.

GRID ACTIVITIES

Prof. Matsuoka is a respected expert in grid technologies not only within Japan but also internationally. He gave an overview of his group's R&D efforts to construct and manage the next generation of a distributed heterogeneous computational infrastructure for conducting large-scale computational science simulations, in addition to collaborating with other institutions. This includes the parallelization of high-level multi-disciplinary applications, parallel and distributed software and frameworks, and commodity PC clustering.

The computer systems at GSIC include a 16-processor NEC SX-5 (128 Gflop/s, 96 GB memory), a 256-processor SGI Origin2000 (200 Gflop/s, 256 GB memory), and a 64-processor HP/Compaq GS320 AlphaServer (128 Gflop/s, 64 GB memory). A SGI Onyx2 4-IR2 with four graphics pipelines helps visualize the simulation results generated by these computer systems. All systems are heavily utilized (99% for the SX-5), but 4- to 8-processor Gaussian runs consume 50% of all compute cycles. Only a handful of users run bigger jobs. All three machines disappeared from the Top500 list in November 2002, yet almost three years

of a six-year contract valued at approximately $35M still remain. Prof. Matsuoka noted that the aggregate computational power at all Japanese university computer centers is much less than that of the Earth Simulator. He believes that this is insufficient to meet the computational science requirements of academic researchers who need a large but diverse set of resources.

To achieve this objective, GSIC is managing a campus grid at TiTech called SuperTITANET (Super Tokyo Institute of Technology Academic Network), which ties together many mid-sized clusters consisting of low-price commodity processors distributed over the entire campus. This includes the semi-production TiTech grid clusters and servers (816 processors, 1.26 Tflop/s), cluster projects such as the COE-IBM Data Cluster (150 processors, 600 Gflop/s), and experimental research clusters such as Presto III (600 processors, 1.6 Tflop/s). The goal is not only to enable large-scale parallel computations but to also explore how to integrate a diverse set of users and their applications. The individual university campus grids throughout Japan are linked together using SuperSINET.

Prof. Matsuoka confessed that significant work still remains to have a robust production grid. On the user side, scientists are not used to exploiting distributed resources, have little expertise with grid middleware, and do not yet have a handle on how to couple applications on a grid. On the operations side, most system administrators do not yet have the skills to manage a large number of distributed clusters and to facilitate/enforce campus-wide security. On top of all this, there are many open research issues in the general grid arena itself.

HIGH ORDER NUMERICAL SCHEMES

Prof. Aoki then gave a talk describing his group's work on developing highly accurate numerical schemes such as an interpolated differential operator (IDO) and Hermite interpolation to preserve high order. The targeted applications include dendrite solidification, turbulent cavity flow, a TNT explosion, blood flow, and a falling leaf. For example, simulating the chaotic motion of falling leaves is an extremely difficult problem because of tightly coupled fluid-structure interaction, the complex shape of leaves, and their very thin structure. To solve this problem, he uses a Cartesian grid with a special treatment of the cut-cells describing the irregular boundary. However, none of these applications currently have huge computational resource requirements because the near-term goal is to develop advanced numerical schemes. All simulations are performed on TiTech's grid resources at this time.

Prof. Aoki is also collaborating with Recherche Prevision Numerique (RPN) of Environment Canada to develop a grid-based global environment simulator. This would combine GSIC's expertise in grid computing and IDO high order numerical schemes with RPN's mesoscale atmosphere model MC2. RPN is responsible for the research and development of the modeling component of the Numerical Weather Prediction (NWP) system for several Canadian meteorological organizations.

REMARKS

The people we met at TiTech were clearly more interested in grid computing than traditional tightly-coupled supercomputing. This is not to insinuate that they do not believe in the need for supercomputing but rather highlights their conviction that distributed heterogeneous computational resources can meet the high-end computing requirements of scientists and engineers. Nevertheless, grids have other advantages, such as enabling distributed human collaboration and providing remote access to data archives, specialized instruments, sensor networks, and large-scale supercomputers.

The interest in grid technology is also motivated by business applications, consumer desires, and government policies. Clearly, the largest Japanese computer vendors (except perhaps NEC) are more interested in grids from a commercial perspective and give scientific simulations a much lower priority. At present, policy makers in the relevant Japanese ministries are more interested in PC clusters than large supercomputers like the Earth Simulator to further e-business and e-governance to reduce government overhead, benefit industry, and reach out to citizens.

Site: **University of Tsukuba**
Center for Computational Physics (CCP)
(Center for Computational Sciences, as of April 1, 2004)
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573
http://www.rccp.tsukuba.ac.jp/ccp-top.html

Date visited: March 31, 2004

WTEC Attendees: Y.T. Chien (Report author), S. Meacham, P. Paul, A. Trivelpiece, K. Yelick

Hosts: Dr. Akira Ukawa, Director of CCP and Professor, CCP and Graduate School of Physics
Dr. Taisuke Boku, Associate Professor of CCP and Graduate School of Computer Science
Dr. Mitsuhisa Sato, Professor of CCP and Graduate School of Computer Science

BACKGROUND

WTEC's visit coincided with the Center's last day of business in its present organization as the Center for Computational Physics (CCP). As of April 1, CCP will become the Center for Computational Sciences, carrying a new name and a new undertaking to expand its mission of research clearly beyond computational physics. This event was also closely in sync with the transition of the University of Tsukuba's status, as with all of Japan's national universities, into an independent administrative organization, a recurrent theme and discussion in many of the sites we visited in Japan during the week.

The Center was founded in 1992 as an inter-university facility, open to all researchers of universities and other research institutions across Japan. Its main objective has been to carry out research and development of high-performance computer systems suitable for work in a variety of areas, including particle physics, condensed matter physics, astrophysics, and biophysics. The emphasis on linking computation with physics is reflected in the Center's organization and the distribution of human resources. While computer science is one of the five research divisions, roughly half of the 11 faculty members are computer scientists. Much of the work in CCP for the last decade or so has been the result of close collaboration between physicists and computer scientists.

Our visit was graciously hosted by Dr. Ukawa, director of CCP, a physicist and two faculty members from the School of Computer Science: Professor Boku and Professor Sato.

HEC THEN AND NOW

The centerpiece of CCP's research facility is a massively parallel computer system, called CP-PACS (Computational Physics by Parallel Array Computer System), developed in-house by the joint efforts between physicists and computer scientists, in collaboration with Hitachi, Ltd. The system is an MIMD (Multiple Instruction Multiple Data) parallel computer with a theoretical peak speed of 614 Mflop/s and a distributed memory of 128Gb. It consisted of 2048 (currently 2172) processing units for parallel floating point calculations and 128 I/O units for distributed input/output processing. The processing units are connected in an 8x17x16 three-dimensional array by a Hyper-Crossbar network. These elements, CPUs, network and I/O devices, are well-balanced so as to support the high-performance capability of the system for parallel processing. The CP-PACS system started operation in March 1996. In November 1996, the system made it to the top of the Top500 list in LINPACK performance.

In the ensuing years, CCP has been pursuing research and development of next-generation massively parallel computers for computational physics. In 1997, a new program to help promote that work by the JSPS (Japan Society for the Promotion of Science, now a unit of MEXT - the Ministry of Education, Culture, Sports, Science and Technology) selected CCP as one of the two projects (the other being at the University of

Tokyo) to focus on the next-generation parallel computers for computational physics. The CCP chose two themes for this research. One is on the architecture of an ultra high-performance processor with on-chip memory, aimed at minimizing the traditional bottleneck of the bus bandwidth between main memory and processors. Another key component is the development of highly flexible and high-bandwidth parallel I/O and visualization systems. This research project ended in 2002, but some of the work is continuing to this day.

HETEROGENEOUS MULTI-COMPUTER SYSTEMS (HMCS)

The early efforts of CP-PACS, coupled with the results of the more recent five-year, JSPS project on parallel computers, have led the Center to put in place a prototype facility with mixed architectural features. This new facility seems to have better positioned the Center to expand its research in and beyond physics. HMCS (Heterogeneous Multi-computer System) is a framework of computer resources that combines vector machines, MPPs, clusters, and special purpose machines. The prototype system at present consists of an MPP (the PC-PACs and eight Grape-6 (6 Tflop/s) special purpose machines) for the simulation work in astrophysics research, more specifically, galaxy formation under UV background radiation.

In replying to the Panel's questions on the future HEC needs in computational physics, Dr. Ukawa said that 10 to 20 Tflop/s is needed in the near term and 100 Tflop/s will be needed in the not too distant future. If cost is not a limiting factor, vector architecture is probably still the first choice, and clusters the next. Even grid technologies should be considered, as they may provide not only high-performance computational power, but also large-scale data-intensive applications for researchers to share data and software. The HMCS frame can conceptually accommodate grids as well.

NEW DIRECTIONS AFTER RE-ORGANIZATION

Interestingly, the real highlight of this visit is in fact not what we learned about the Center's past, but its future as of the next day, April 1, 2004. Dr. Ukawa in his overview presentation began his remarks by describing the important changes taking place. First, the name of the Center will be changed to the "Center for Computational Sciences." The scope of research will be greatly expanded to include other sciences as well as to facilitate how collaborative research is done with broader computer science disciplines. The new organization will have six divisions: Particle- and astro-physics; materials and life sciences; earth and life environments; high-performance systems; computation and informatics; and collaborative research. The addition of a division on computation and informatics seems to place greater emphasis in the new Center on large-scale data-centered research, a common theme in computational biology, for example. The new Center will have about 30 faculty members in various disciplines. Its annual budget is approximately $5 million.

OBSERVATIONS

In reply to a question by Professor Yelick, Dr. Ukawa said that the Director of the new Center has control over these positions and works with various departments to fill them in the best interest of the joint faculty research efforts between the Center and the participating disciplines. Such control is apparently an important driver for cross-disciplinary research at the Center across the Tsukuba campus. In contrast, directors of inter-disciplinary research centers in the United States hardly have such leverage at all, according to Professor Yelick (UC Berkeley), and other panelists.

QUESTIONS AND ANSWERS

Our visit concluded with a session, led by Professor Boku, in which he went over in some detail the written answers to a list of questions the Panel submitted ahead of its visit. The answers, meticulously prepared by our hosts and other faculty members, elaborate on various aspects of the Center research specifically and the high-end computing issues in general. They range from programming paradigms to system software (large-scale, fault-tolerant) to HEC architectures to underlying technologies to future applications, including their

candid assessments on and the lessons learned from the Earth Simulator. The text, which contains the entirety of the questions and answers, follows.

Programming Paradigms:

- What innovations in programming paradigms are under development for future massively parallel HEC systems?

  Two level paradigms – easy utilization of shared memory such as OpenMP, and traditional message passing. MPI will remain like Fortran[, which] is still alive.

- Are you looking beyond traditional message passing libraries such as MPI?

  The basic system may still remain on message passing system, but [an easier] programming style is desired for simple description of data parallel on distributed memory system.

- What about HPF, UPC, and shared memory programming language like OpenMP?

  OpenMP is a good extension both for Fortran and C. There will be very few users of HPF, but we have to learn something from the experience of HPF, [such as] ways to describe data distribution easily.

- What about hybrid paradigms that combine message passing, shared memory, and even vectorization?

  In [the] next decade, it will be necessary to exploit multiple levels of parallelism for maximum performance. Automatic parallelization or distribution, software DSM requires more research and development.

- What do you see as the most important areas for future research and development in programming models for high-performance machines?

  We have to treat massively parallel elements in a program. For large-scale message passing programs, more sophisticated and simple[r] styles or models are required.

System Software:

- What innovations in compilers, dynamic libraries, file systems, and other systems software are being developed by Japanese research that will have the greatest impact, especially for petascale systems consisting of 10,000 or more processors?

  One of the representative system[s] software from Japan is Score developed by RWCP[, which] still continues to research in PC cluster Consortium. This is cluster middleware to manage large-scale (>1000 processors) [of a] high-end cluster system. On Score, various system software such as the OpenMP compiler, software DSM, etc., has been developed.

- What are your most significant recent achievements in the area of developing fault-tolerant software, and what do you see in the future?

  It is still on going, but we are developing a fault-tolerant interconnection network system for clusters. It is based on multiple links of commodity network[s] such as GbE, and we utilize these links both for wide bandwidth and fault-tolerancy.

Architectures:

- Do you believe that vector architectures will remain the core technology for the next generation of Japanese HEC machines? If so, how do you envision maintaining your leadership for the next 5-10 years? If not, what other directions will you be moving in and what other architectures are you working on?

We don't think vector architecture is the only [answer] for HEC, but many important applications will require it. We think various sizes and levels of HEC machines from PC-cluster to [an] ES-like high-end machine have to be used according to the problem size and characteristics.

We are currently very interested in how we can tolerate QCD calculation with [a] cluster solution [in the] 10 Tflop/s-[range].

- Is there a major thrust on designing innovation architectures such as processor-in-memory, streaming, FPGAs, hardware multithreading, percolation, etc?

  We are currently researching on-chip memory architecture with very simple model[s] to exploit well-scheduled data movement between [the] processor and off-chip memory. This architecture is named SCIMA (Software Controlled Integrated Memory Architecture). SCIMA performs much better than traditional cache architecture especially on large-scale scientific calculation[s] where we can analyze and control the data movement.

- Since some of these concepts are being tested in the U.S., do you plan to delve into these areas or are you looking at other ideas? Do you have a strategy for beating Moore's Law?

  We are not sure that our way can break Moore's Law, but within a limited memory bandwidth, we have to consider how to reuse the data on [the] on-chip memory, how to maintain [their] lifetime within limited capacity, etc.

- Do you design a computer for a broad range of applications or one specialized for a class of applications?

  In our center, we concentrated on several fields of computational physics problems. In this sense, our target is relatively limited. But it doesn't mean that we need a special class of machines because the methodologies for these fields (currently, lattice QCD, astrophysics, condensed matter physics) differ.

- Given that scientific computing may require petaflop or multi-petaflop machines within a decade or less, what do you anticipate the power requirements to be? Will the choice of architecture take power consumption into consideration?

  Power consumption is a very serious problem in [the] near future. It affects both power itself and cooling. [A] power-aware approach is needed not only on architecture or hardware but also on software.

- What approaches have you taken to address the growing gap between memory system performance and processor performance, and what techniques do you view as most promising for the future?

  As answered above, we are researching on-chip memory architecture. Software-controlled data movement seems troublesome for application users, but we are also developing a compiler for our architecture. Data reusability and efficiency [for] data transfer are very important issues on limited-memory bandwidth.

- What criteria do you use to determine if a system is well-balanced, in terms of network, processing, and memory performance? How well-balanced are your systems (or others) by this measure?

  In our current MPP system (CP-PACS with 2172 CPUs), all these elements are very well-balanced. Since it is an old machine, the absolute power is small, but a processor node has 300 Mflop/s CPU, 1.2 GB/s memory bandwidth and 300MB/s network bandwidth.

- What were the main technical innovations of your hardware development over the past few years?

  On CP-PACS, pseudo-vector processing [architecture] was introduced. [Though the implementation of the non-pure] vector feature[s] [was successful], [it does require very expensive] CPU and memory. The balance of super-scalar execution and memory latency hiding was very good.

- How do you see capability, clusters, and grids fitting together in the future?

  Clusters will take [a] very important role in the future. [Clusters in the] 10 Tflop/s-class may be used for most medium-class job[s]. The future of grids depends on how wide and fast network[s are] in [the] near future. Currently, we can solve only some master-slave (worker) style problems on grids, but actual parallel job-distribution will be able to [run].

Underlying Technology:

- What CPU technology do you currently use and expect to use for a future petaflop machine? E.G. Bipolar, CMOS, other?

  CMOS, because [there are] power consumption problem[s] on ultra-large scale system[s].

- What communication technology do you use now and plan to use in the future for a petaflop machine? This is for communication within a chip (maybe metallurgy?), chip-to-chip, chip to module, module-to-module, and computer-to-computer. What switching technology will be used for a high[-]bandwidth optical communication link to petaflop machine?

  Optical communication is necessary for petaflop computing. But [a] fully optical switch is not yet on stage. A parallel (not so wide) link (trunk) of very high[-]speed serial link is one of the candidates.

Applications:

- What are the biggest driving applications in addition to Earth sciences? Is space sciences the next big thing?

  Particle physics, nanotechnology simulation, biotechnology simulation.

- Do you have a clear idea of the computational requirements in your major thrust areas for the next 5-10 years? And do you have a roadmap of how you are going to meet those requirements (either by stand-alone supercomputer capability systems or distributed grid capacity systems)?

  From [the] viewpoint of cost-effectiveness, [an] ES-like system is [very] expensive. One of the short path[s] to petaflop[/s] computing is a hybrid system of special purpose and general purpose parallel processing systems.

- What has been the science impact of the Earth Simulator? Has it enabled you to make new discoveries? And how has that affected the Earth sciences community?

  This will be the first lattice QCD calculation without any approximation of the underlying theory.

- In the U.S., computational science is increasingly being recognized as the third “leg” of science in addition to theory and experimentation. Do you see a similar development in Japan?

  Yes. From now, it is important to see “What’s beyond the simulation?” or “What’s given by [the] simulation?”

- What lessons did you learn from the Earth Simulator project? What impact did it have on your applications, on computational science in Japan in general?

  Lattice QCD computation is characterized by vector computation (linear solver of a large sparse matrix) and parallelism (regular 4-dim space-time mesh). So ES with its vector-parallel architecture is good and useful for lattice QCD.

- What architecture do you think is most suitable for your application? Why?

  Suitable architecture may be vectors. But to achieve [a good balance between the CPU and network/memory bandwidth, a large-scale parallel vector is doubtful. We need wide-bandwidth but not a

flexible communication pattern on [the] network. A cluster with [a] powerful nearest-communication feature is suitable for QCD.

- What are the current bottlenecks for your application on HEC systems? What is being done to reduce them?

  Network performance. We need more bandwidth. [The problem is] not so severe on latency.

- Are you collaborating with computer scientists/applied mathematicians in order to make progress in computational science?

  [That] is just what we are doing here [at the] Center for Computational Physics.

- Who (which agency) is funding computer acquisitions for your application discipline?

  MEXT.

- There is a strong tradition in Japan to build special-purpose machines such as MD-Grape or QCDPACS. Why do you think this is? Do you expect more special-purpose machines to be built in the future? For which applications?

  [C]ost-effectiveness for special purpose machines is very high. But in [the] next generation simulation, we think a hybrid system is a key issue to make special and general purpose machine complementary. Not just a simple special-purpose system.

APPENDIX C: RECENT CHANGES IN THE TOP500 LIST

In November 2004, the 24th Top500 list was unveiled at the Supercomputer Conference (SC2004) in Pittsburgh, PA. The latest list, released as this study was going to press, revealed significant shifts in the top ranks of the high-end computing world since the list was last updated in June 2004 (see Table C.1). Most importantly for this study, two U.S. computers have wrested the top spot from the Earth Simulator (ES), which has held on to the position for 2 1/2 years. The impact of this development on high-end computing in Japan, as well as for the supercomputer industry as a whole, will likely be significant.

Table C.1
Top500 as of November 2004 (#1 – #20, inclusive)

#	Computer and Processors	Site	Year	Previous Rank	R_{max} R_{peak} (in Tflop/s)
1	*BlueGene/L DD2 beta-System* 0.7 GHz PowerPC 440	IBM/DOE, United States	2004	*	70.72 91.75
2	*Columbia* 1.5 GHz SGI Altix, Voltaire Infiniband	NASA Ames Research Center/NAS, United States	2004	*	51.87 60.96
3	*Earth Simulator*	Earth Simulator Center, Japan	2002	1	35.86 40.96
4	*MareNostrum* eServer BladeCenter JS20 (2.2 GHz PowerPC 970), Myrinet	Barcelona Supercomputer Center, Spain	2004	*	20.53 31.36
5	*Thunder* 1.4 GHz Intel Itanium2 Tiger 4, Quadrics	Lawrence Livermore National Laboratory, United States	2004	2	19.94 22.94
6	*ASCI Q* 1.25 GHz AlphaServer SC45	Los Alamos National Laboratory, United States	2002	3	13.88 20.48
7	*System X* 2.3 GHz 1100 Dual, Apple Xserve/Mellanox Infiniband 4X/Cisco GigE	Virginia Tech, United States	2004	3**	12.25 20.24
8	*BlueGene/L DD1 Prototype* 0.5 GHz PowerPC 440 w/Custom	IBM-Rochester, United States	2004	4	11.68 16.38
9	*eServer pSeries 655* 1.7 GHz Power4+	Naval Oceanographic Office (NAVOCEANO), United States	2004	*	10.31 20.02
10	*Tungsten*	NCSA, United States	2003	5	9.82

Table C.1
Top500 as of November 2004 (#1 – #20, inclusive)

#	Computer and Processors	Site	Year	Previous Rank	R_{max} R_{peak} (in Tflop/s)
	PowerEdge 1750, 3.06 GHz P4 Xeon, Myrinet				15.30
11	*eServer pSeries 690* 1.9 GHz Power4+	ECMWF, United Kingdom	2004	6	9.24 16.54
12	*eServer pSeries 690* 1.9 GHz Power4+	ECMWF, United Kingdom	2004	18***	9.24 16.54
13	*John Von Neumann* LNX Cluster, 3.4 GHz Xeon, Myrinet	US Army Research Laboratory (ARL), United States	2004	*	8.77 13.93
14	*RIKEN Super Combined Cluster*	Institute of Physical and Chemical Research (RIKEN), Japan	2004	7	8.73 12.53
15	*BlueGene/L DD2 Prototype* 0.7 GHz PowerPC 440	IBM-Thomas Watson Research Center, United States	2003	8	8.66 11.47
16	*Mpp2* 1.5 GHz Integrity rx2600 Itanium2, Quadrics	Pacific Northwest National Laboratory, United States	2003	9	8.63 11.62
17	*Dawning 4000A* 2.2 GHz Opteron, Myrinet	Shanghai Supercomputer Center, China	2003	10	8.06 11.26
18	*Lightning* 2 GHz Opteron, Myrinet	Los Alamos National Laboratory, United States	2003	11	8.05 11.26
19	*MCR Linux Cluster* 2.4 GHz Xeon, Quadrics	Lawrence Livermore National Laboratory, United States	2002	12	7.63 11.06
20	*ASCI White* 375 MHz SP Power3	Lawrence Livermore National Laboratory, United States	2000	13	7.30 12.29

* = Debut in Top500

** = Rank from November 2003; not listed in June 2004 Top500

*** = Formerly of HPCx, United Kingdom

IBM's BlueGene/L DD2 beta-System, built for the Lawrence Livermore National Laboratory's new Terascale Simulation Facility and currently being tested at the IBM facility in Rochester, NY, claimed the

top spot. BlueGene/L turned in an Rpeak of 91.75 Tflop/s and Rmax of 70.72 Tflop/s, 1.92 and 2.24 times, respectively, the Rmax and Rpeak of the ES. BlueGene/L is a scalable system of cellular design with a theoretical peak Linpack performance of 367 Tflop/s. Each node consists of single compute IBM ASIC (application-specific integrated circuit) and SDRAM-DDR memory chips; nodes are connected via 3D torus, combining tree, and barrier networks.

NASA's SGI Altix-powered Columbia took second place with an Rmax of 51.87 Tflop/s and an Rpeak of 60.96 Tflop/s, which are respectively 1.45 and 1.49 times as fast as the ES. Located at NASA's Advanced Supercomputing (NAS) facility, Columbia is an integrated cluster of 20 Altix 512-processor systems. It is the largest Linux-based shared-memory computer system in the world. Built in only 120 days, Columbia is designed to support NASA's Visions for Space Exploration mission and will be made available to the larger national science and engineering community. Columbia is based on the SGI NUMAflex™ architecture with 10,249 Intel Itanium® 2 processors, a Linux operating system, and 20 TB total memory.

The 24th Top500 list revealed other shake-ups in the upper tier of the list as well. Three of the top five machines were brand new to the list. In addition to BlueGene/L and Columbia, the new IBM BladeCenter JS20 based MareNostrum system at the Barcelona Supercomputing Center debuted at #4, right below the ES. MareNostrum is now the most powerful computer in Europe. The list also highlights the continued increase in performance of the top 10 machines; only one of the systems in the top 10 has a Linpack performance of less than 10 Tflop/s.

Table C.2 lists NEC, Fujitsu, and Hitachi computers in the Top500. Their rankings in the current and previous Top500 lists are provided for the purposes of comparison. While four new Japanese machines have placed in the Top500 (two each from NEC and Hitachi), an equal number have dropped off the list (one each from NEC and Fujitsu and two from Hitachi), resulting in no net gain in presence for these Japanese manufacturers, whereas 20 new systems from the United States were added to the list. Furthermore, the relative positions of previously listed NEC, Fujitsu, and Hitachi machines have all dropped, some of them considerably.

Table C.2
NEC, Fujitsu, and Hitachi Machines in the Top500 (by November 2004 ranking)

Location	Machine	Manu-facturer	Area	Rank (Nov. 2004)	Rank (June 2004)	Linpack Perform (Gflop/s)	Peak Rate (Gflop/s)
Earth Simulator Center	Earth Simulator	NEC	Research	3	1	35860	40960
Institute of Physical and Chemical Res. (RIKEN)	RIKEN Super Combined Cluster	Fujitsu	Research	14	7	8728	12534
National Aerospace Laboratory of Japan	Primepower HPC2500 (1.3 GHz)	Fujitsu	Research	32	22	5406	11980
Kyoto University	Primepower HPC2500 (1.56 GHz)	Fujitsu	Academic	33	24	4552	9185
National Institute for Materials Science	SR11000-H1/56	Hitachi	Research	59	—	3319	6093
Institute for Molecular Science	SR11000-H1/50	Hitachi	Research	69	—	2909	5440

Table C.2
NEC, Fujitsu, and Hitachi Machines in the Top500 (by November 2004 ranking)

Location	Machine	Manu-facturer	Area	Rank (Nov. 2004)	Rank (June 2004)	Linpack Perform (Gflop/s)	Peak Rate (Gflop/s)
Meteorological Research Institute/JMA	SX-6/248M31 (typeE, 1.778ns)	NEC	Research	88	68	2155	2232
DaimlerChrysler	Opteron 2.0 GHz, GigE	NEC	Industry	154	—	1778	3584
University of Tokyo	SR8000/MPP	Hitachi	Academic	163	122	1709.1	2074
Leibniz Rechenzentrum	SR8000-F1/168	Hitachi	Academic	168	127	1653	2016
DKRZ - Deutsches Klimarechenzentrum	SX-6/192M24	NEC	Research	198	148	1484	1536
National Institute for Fusion Science	SX-7/160M5	NEC	Research	217	161	1378	1412.8
Osaka University	SX-5/128M8 3.2ns	NEC	Academic	291	184	1192	1280
Institute of Space & Astronautical Science (ISAS)	SX-6/128M16 (typeE, 1.778ns)	NEC	Research	309	194	1141	1152
Bureau of Meteorology / CSIRO HPCCC, Australia	SX-6/144M18	NEC	Research	317	—	1130	1152
NEC Fuchu Plant	SX-6/128M16	NEC	Vendor	411	249	982	1024
United Kingdom Meteorological Office	SX-6/120M15	NEC	Research	459	275	927.6	960
United Kingdom Meteorological Office	SX-6/120M15	NEC	Research	460	276	927.6	960
High Energy Accelerator Research Organization /KEK	SR8000-F1/100	Hitachi	Research	464	280	917	1200
VW (Volkswagen AG)	Opteron 2.0 GHz, GigE	NEC	Industry	475	289	891	1440
University of Tokyo	SR8000/128	Hitachi	Academic	483	295	873	1024
Institute for Materials Research/Tohoku University	SR8000-G1/64	Hitachi	Academic	—	333	790.7	921.6
University of Tsukuba	VPP5000/80	Fujitsu	Research	—	393	730	768
Japan Meteorological Agency	SR8000-E1/80	Hitachi	Research	—	416	691.3	768
CBRC - Tsukuba Advanced Computing Center - TACC/AIST	Magi Cluster PIII 933 MHz	NEC	Research	—	457	654	970

REFERENCES

Dunbar, J., editor. 2004. Columbia's Construction. <http://www.nas.nasa.gov/About/Projects/Columbia/columbia_build.html> Last accessed February 23, 2005.

Dunbar, J., editor. 2004. NAS Computing Resources – Columbia Supercomputer. <http://www.nas.nasa.gov/Resources/Systems/columbia.html> Last accessed February 23, 2005.

"Facts on BlueGene/L," distributed at SC2002, Baltimore, Maryland, November 16–22, 2002, <http://www.llnl.gov/asci/platforms/bluegenel/images/bluegenel_brochure.pdf> Last accessed February 23, 2005.

Highlights from Top500 List for November 2004. 2004. <http://www.top500.org/lists/2004/11/trends.php> Last accessed February 23, 2005.

Top500 Current List and Supercomputer Site Profiles. 2004. <http://www.top500.org/lists/2004/11/> Last accessed February 23, 2005.

APPENDIX D: HIGHLIGHTS FROM THE U.S. HEC WORKSHOP

INTRODUCTION

After the Earth Simulator, what's next for the manufacturers of high-end computing hardware and software in Japan? The WTEC Panel on High-End Computing (HEC) in Japan was convened on May 25, 2004, in part to answer this question. The members of the panel began by gathering and analyzing information from a variety of sources:

- Their own knowledge and insights;
- Reviews of published and gray literature;
- Correspondence with colleagues; and
- Site visits in Japan resulting in detailed site visit reports.

Following the site visits, the panelists convened in Arlington, VA, for a workshop where they presented their preliminary findings to a public audience of sponsors, colleagues from the United States, and hosts of the sites visited in Japan. This chapter summarizes the panel's presentations at that workshop. The panelists then incorporated the comments provided by the workshop participants into the chapters that appear in this report.

A detailed introduction to the WTEC study, along with biographies of the panelists and viewgraphs presented at the workshop, are available at the HEC study homepage, hosted by WTEC at http://www.wtec.org/hec.

SUMMARIES OF EXPERT PRESENTATIONS

Context of the HEC Study

Dr. R. D. Shelton

Dr. Shelton opened the workshop by explaining the context of the high-end computing study as one of a series of 60 international research assessments conducted by WTEC since 1989. Dr. Shelton pointed out that WTEC has completed more studies of this type than any other U. S. organization, and he discussed the expertise of WTEC staff in this area. He also discussed WTEC's process for organizing and conducting science and technology (S&T) assessments, and explained how WTEC disseminates the results to a broad audience through print and electronic media.

Dr. Shelton outlined the purposes and value of S&T assessments in general. He also reviewed why hosts for WTEC teams are traditionally cooperative, citing the scientific tradition of sharing research results and the benefits to the hosts resulting from technology transfer and professional collaboration.

Executive Summary

Al Trivelpiece

Dr. Trivelpiece began by summarizing the panel's conclusions about the Earth Simulator (ES). He reviewed the history of the ES and pointed out that while the ES is a superb engineering achievement and will likely remain the leader for some time, the United States is actually currently ahead of Japan despite the latter's broad-based strategic effort. The emphasis in Japan on grid computing had surprised the committee.

The key policy-making role of the Council for Science and Technology Policy (CSTP) was reviewed, as was the ways that government projects have encouraged the development of commercial architectures, in particular the systems that were developed as a result of the Numerical Wind Tunnel, cp-pacs, and the ES.

The recent policy of spinning off university research facilities as Independent Administrative Institutions (IAI) is a major change in the approach to government funding. The committee learned that Japan is, for the most part, abandoning vector supercomputing architectures with high-bandwidth memory subsystems because they are not seen to be commercially viable. High-bandwidth systems are being replaced by commodity clusters.

Introduction to the Study

Al Trivelpiece

As Chair of the WTEC panel on HEC in Japan, Dr. Trivelpiece reviewed the purpose and scope of the study. He discussed the process used by the panel to gather information on current and future Japanese HEC research and applications in the public, private, and academic sectors, and then compared those findings to HEC efforts in the United States. Of particular interest to the panel were the development process, operational experience, impact, and follow-ons to the ES, and the identification of areas of potential cooperation between the two countries.

After reviewing the names and qualifications of the panel members, Dr. Trivelpiece outlined the panel's travel agenda and discussed highlights from the site visits that occurred between March 29 and April 2, 2004.

The Earth Simulator

Jack Dongarra

Dr. Dongarra reviewed the history, design, and performance of the ES ultra high-speed parallel supercomputer, pointing out that since its completion in early 2002, the ES has remained in the top slot of the Top500 and is expected to remain there for at least another year. Significant statistics about the design, costs, and performance of the ES were highlighted, as were the unique characteristics of the building that houses the ES and its support systems. The architecture and performance characteristics of the ES and related software were extensively detailed (see Figures D.1 and D.2).

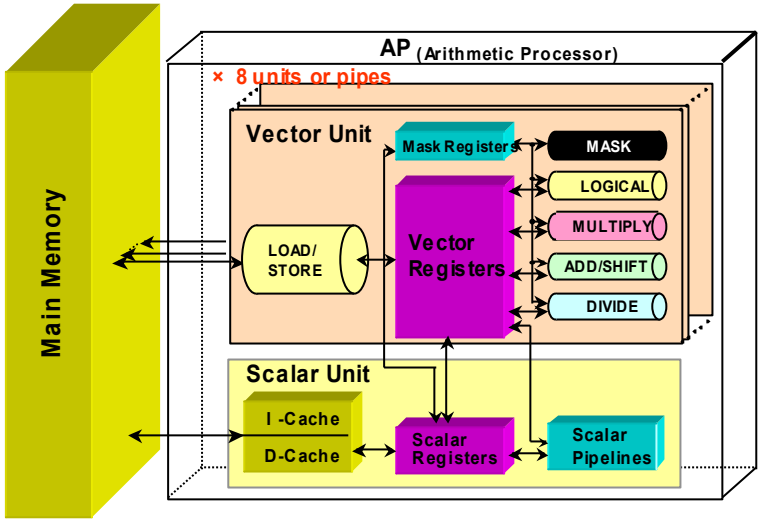


Figure D.1. Block diagram of arithmetic processor (single chip). Bandwidth from vector registers to/from memory (32 GB/s) (Courtesy JAMSTEC)

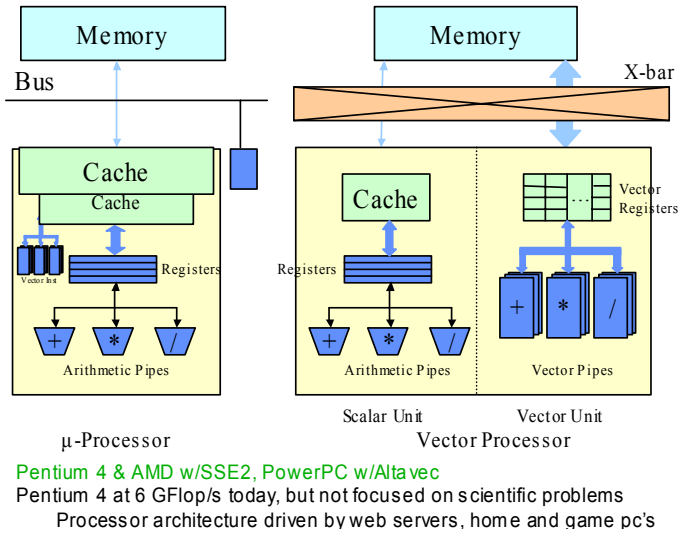


Figure D.2. Scalar vs. vector processors (Courtesy JAMSTEC)

Dr. Dongarra also profiled the life and career of Dr. Hajime Miyoshi, often referred to as "the Seymour Cray of Japan" for his pioneering role in the development of HEC there. The development of the SX series of processors by NEC – the ES hardware system vendor – was also reviewed. Dr. Dongarra concluded that the ES was a singular accomplishment resulting from the unique confluence of robust government funding, an important problem area, and the impetus of Dr. Miyoshi.

Scientific Applications I

Rupak Biswas

Dr. Biswas reviewed selected scientific applications of HEC at five Japanese agencies, grouped into the following broad categories: computational fluid dynamics for aerospace problems at the Japan Aerospace Exploration Agency (JAXA) and the Tokyo Institute of Technology (TiTech); earth science calculations at the ES and the Frontier Research System for Global Change (FRSGC); and finally Computational Nanotechnology at the Research Organization for Information Science & Technology (RIST). Brief overviews of the history and mission of each agency were followed by highlights of notable HEC applications past and present (see Figure D.3). Comparisons and contrasts were made with comparable projects in the United States.

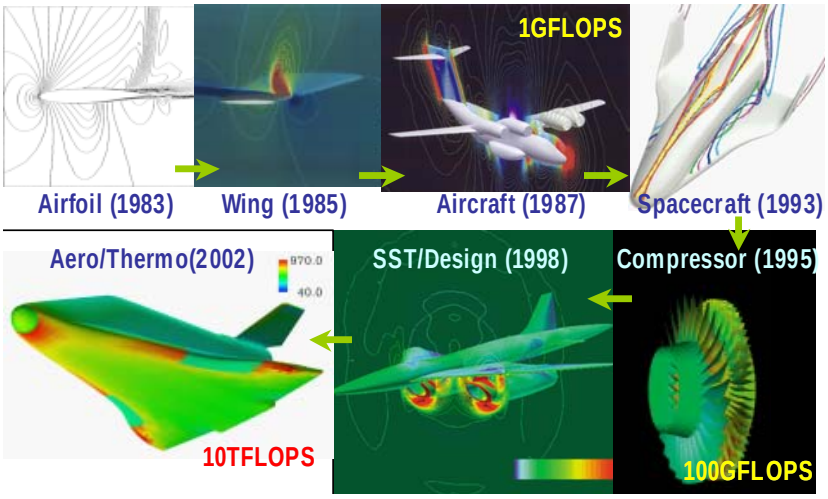


Figure D.3. Evolution of CFD projects at JAXA (Courtesy JAXA)

Dr. Biswas observed that while the ES has led to significant advances in HEC, it has also engendered resentment over the diversion of funds from other HEC activities to pay for it. Furthermore, as yet there is no consensus for a follow-on program. While the quality of research and development in the various scientific applications discussed is competitive with the world's best, continued progress will require a significant increase in computational power.

Scientific Applications II

Peter Paul

Dr. Paul continued the review of scientific applications of HEC, focusing on six Japanese organizations grouped into the following broad categories: lattice gauge calculations at the National Laboratory for High Energy Physics (KEK), Tsukuba University, and the Institute of Physical and Chemical Research-Brookhaven National Laboratory Research Center (RIKEN-BNL); the Protein Explorer at RIKEN-Yokohama; plasma fusion at the National Institute for Fusion Science (NIFS) and the Japan Atomic Energy Research Institute (JAERI); and reactor cooling and materials defects at JAERI. Dr. Paul also reviewed the University of Tokyo's plans for HEC.

Dr. Paul affirmed many of the observations of previous speakers, noting that Japan is actively concentrating on grid computing software and superclusters. He noted that Japanese scientific institutions have also implemented a notable strategic approach to proteomics, cell sciences, and biosciences.

Architecture

Jack Dongarra

Dr. Dongarra compared and contrasted the HEC architectures used by NEC, Fujitsu, and Hitachi. He also reviewed special purpose architectures such as Grape-6 and the Emotion Engine in the Sony PlayStation 2. He noted that current system architectures fall into one of three broad categories: commodity processors with commodity interconnect clusters, commodity processors with custom interconnects, and custom processors with custom interconnects. Examples of systems in each category were highlighted. For example, the Pseudo Vector Processor (PVP) architecture of Hitachi's SR11000 was profiled in the context of that firm's history of HEC development (see Figures D.4 and B.7 in Appendix B).

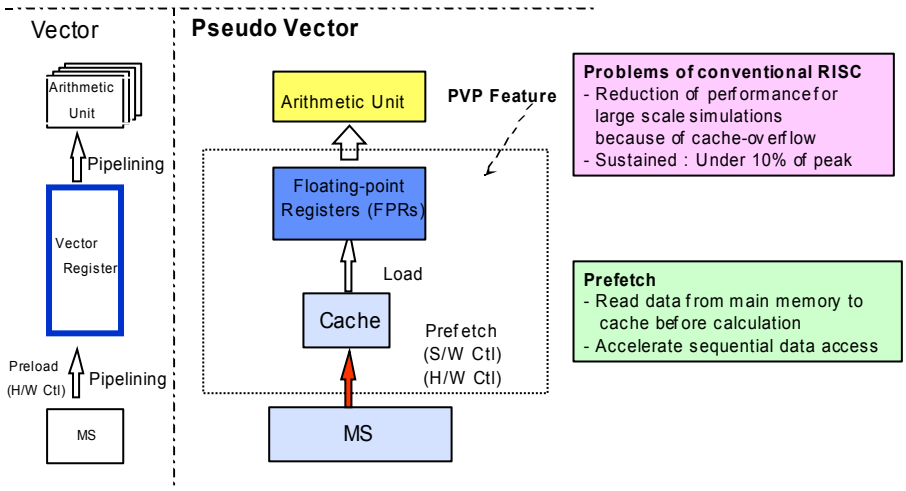


Figure D.4. Pseudo Vector Processing in the Hitachi SR11000 (Courtesy Hitachi)

Dr. Dongarra discussed the diminishing commercial viability of traditional supercomputing architectures that employ vector processors and high-bandwidth memory subsystems. He pointed out that commodity clusters are increasingly replacing traditional high-bandwidth systems in most scientific applications.

Software & Grids

Katherine Yelick

Dr. Yelick reviewed the languages, compilers, libraries, and other software developed by Fujitsu, Hitachi, and NEC for high-end computing. Past and future developments high-performance Fortran (HPF) and the HPF/JA extensions were particularly focused on. While NEC has an ongoing investment in HPF and HPF/JA, particularly in reference to supporting the ES (see Figure 7.1 in Chapter 7), Hitachi has no plans for HPF and Fujitsu software uses data decomposition extensions that are similar to, but distinct from, HPF. There is a more sustained effort for HPF in Japan than in the United States due in large part to the ES, though there is less interest today due to portability and performance issues. MPI is the dominant model for internode communication, while MPI and automatic vectorization is used for communication within nodes.

With regard to grid computing, Dr. Yelick began by reviewing how e-business, e-government, and science have acted as motivating factors for grid computing in Japan. A nationwide grid computing initiative involves the development of an ultra high-speed network infrastructure that will share information across agency and societal boundaries, and that will also nurture human resources of high quality. The architecture and capabilities of specific grid projects, including Super-SINET, the National Research Grid Initiative (NAREGI) (see Figure C.7), the Campus Grid at the Tokyo Institute of Technology, the Grid Technology Research Center at AIST, and the Information Technology Based Lab (ITBL), were discussed. Dr. Yelick also reviewed grid applications at various laboratories and universities, including AIST's VizGrid, Osaka University's BioGrid, and the Japan Virtual Observatory (JVO).

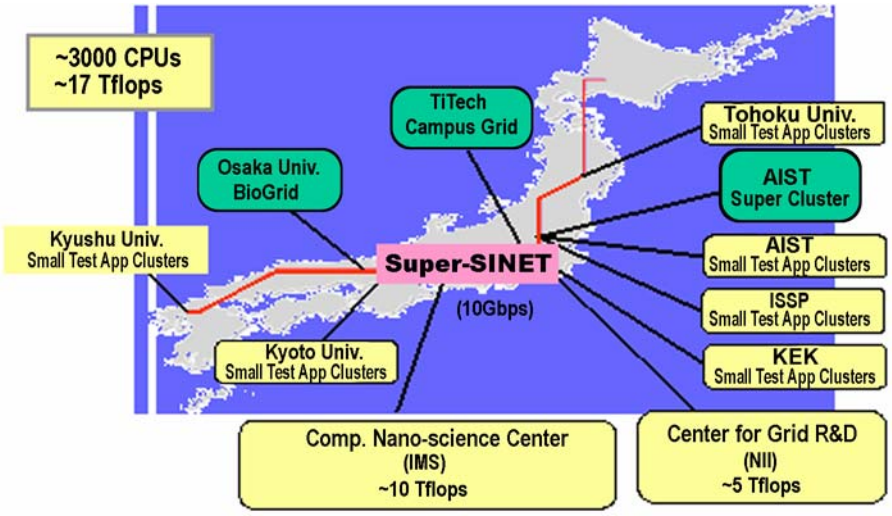


Figure D.5. Structure of the NAREGI Phase I testbed (Courtesy NAREGI)

Dr. Yelick compared and contrasted the sources and levels of funding for the various grid projects, noting that the emphasis on grids in Japan was higher than expected by the WTEC panel. Grids have received high levels of government support, application involvement, and tools. Grids are in place, or are being developed, for a wide range of computational, data, and business applications. Japan's research contributions include cluster computing and grid middleware. Furthermore, Japanese institutions are heavily involved in international collaborations.

High-End Computing Revitalization Task Force

Dave Nelson

Dr. Nelson began by thanking the WTEC panel for its hard work and successful study. The study satisfied the requirement for a comprehensive review of the activities and plans of Japanese government agencies, HEC users, and vendors. While the study revealed a few surprises – namely the changes in government

funding patterns, the singularity of the ES, and the eclipse of vector computing by grid and cluster computing – it also affirmed many trends that have already been spotted. The context for the WTEC study and its role in helping to determine an overall strategy for HEC in the United States were also reviewed.

Dr. Nelson provided an overview of the history, structure, and goals of the High-End Computing Revitalization Task Force (HECRTF) and the current and predicted roles of HEC in the United States. User and agency views on HEC revealed that the research pipeline is dry; furthermore, there are concerns over the adequacy of the industrial base, the small market, protracted development times, and the maintenance of systems and software. Dr. Nelson cited the need for vendors to obtain consistent acquisition metrics and discussed the need for interagency cooperation in developing acquisition strategies.

WORKSHOP PARTICIPANTS

Table D.1
Workshop Participants

Name	Organization	Location
Cecil Uyehara		Bethesda, MD
Chenita Keene' Amey		Washington, DC
Christian Austin-Hollands		Washington, DC
Edward M. Williams	AFOSR	Arlington, VA
Dr. Thomas Beutner	AFOSR/NA	Arlington, VA
David K. Kahaner	Asian Technology Information Program	Albuquerque, NM
Peter Paul	Brookhaven National Laboratory	Upton, NY
Praveen Chaudhari	Brookhaven National Laboratory	Upton, NY
Bryan Allinson	Carnegie Mellon University	Pittsburgh, PA
Susan Fratkin	CASC	
Craig T. LeVan	Council on Competitiveness	Washington, DC
Suzy Tichenor	Council on Competitiveness	Washington, DC
Mark Guiton	Cray Inc.	Arlington, VA
Mark H. Crawford	DOC	Annandale, VA
Timothy Miles	DOC	Washington, DC
Jon Boyens	DOC	Washington, DC
Mike Kane	DOC/NOAA	Washington, DC
John Grosh	DOD	Washington, DC
Maureen Raley	DOD (DTSA/TDD)	Washington, DC
Dale Koelling	DOE	Washington, DC
Jeffrey Mandula	DOE	Germantown, MD
Norm Kreisman	DOE	Washington, DC
Altaf H. Carim	DOE	Washington, DC
Subodh Tripathee	Forum for Information Technology	Nepal
Tetsuro Uruno	Fujitsu Limited	Japan
F. Brett Berlin	George Mason University/SCS	Alexandria, VA
Yolanda L. Comedy	IBM	Washington, DC
Elisabeth M.C. Lutanie	Institute of Physics	Arlington, VA
Sally Howe	ITRD	Arlington, VA

Table D.1
Workshop Participants

Name	Organization	Location
Dave Nelson	ITRD	Arlington, VA
Norihiro Nakajima	Japan Atomic Energy Research Institute	Japan
Akintunde Michael Yinka	Japan Science and Technology Agency	Japan
Eiichiro Watanabe	Japan Science and Technology Agency	Japan
Kuhiniko Niwa	Japan Science and Technology Agency	Japan
Horst D. Simon	Lawrence Berkeley National Laboratory	Berkeley, CA
V.J. Benocraitis	Loyola College	Baltimore, MD
Richard Brown	Loyola College	Baltimore, MD
Christopher Moore	Moore Research Corporation	Fairfax, VA
Jan S Aikins	NASA / Ames	Moffett Field, CA
Rupak Biswas	NASA / Ames	Moffett Field, CA
W. Phil Webster	NASA Goddard Space Flight Center	Greenbelt, MD
Kenichi Miura	National Institute of Informatics	Japan
Peter M. Lyster	National Institutes of Health	Bethesda, MD
Cynthia A. Patterson	National Research Council	Washington, DC
Mark F Corcoran	NEC ATCC (NECSAM) Supercomputing	Cincinnati, OH
Ujahar S. Bhachu	NRC	Germantown, MD
David Hart	NSF	Arlington, VA
Vicky Bookes	NSF	Arlington, VA
Danielle Kriz	NSF	Arlington, VA
Celeste M. Rohlfing	NSF / Chemistry	Arlington, VA
Kamal Abdali	NSF / CISE	Arlington, VA
Sang Kim	NSF / CISE	Arlington, VA
Yunku Yuh	NSF / CISE	Arlington, VA
Ken Chong	NSF / ENG	Arlington, VA
Michael M. Reischman	NSF / ENG	Arlington,VA
Steve Meacham	NSF / GEO	Arlington,VA
George Wilson	NSF / OD	Arlington, VA
Gregory Martin	NSF / OD	Arlington, VA
Patrick G. Sullivan, Jr.	Office of the Secretary of Defense	Washington, DC
Bob Sorensen	Office of Transnational Issues	Washington, DC
Krystal Hathaway	ONR	Arlington, VA
George Gamota	S&T Management Associates	Lexington, MA
Al Trivelpiece	Sandia	Henderson, NV
Dolores Shaffer	STA	
Robert Chadduck	The National Archives	College Park, MD
John G. Dardis	U.S. Department of State	Washington, DC
Katherine Yelick	University of California	Berkeley, CA

Table D.1
Workshop Participants

Name	Organization	Location
Jack Dongarra	University of Tennessee	Knoxville, TN
Stephen A. Roberts	VeriSign, Inc.	Sterling, VA
Masanobu Miyahara	WTEC	Japan
Halyna Paikoush	WTEC	Baltimore, MD
Michael DeHaemer	WTEC	Baltimore, MD
R.D. Shelton	WTEC	Baltimore, MD
Roan Horning	WTEC	Baltimore, MD
Y.T. Chien	WTEC	Baltimore, MD
Thomas J. Bartolucci	WTEC / NNCO	Arlington, VA
Geoffrey Holdridge	WTEC / NNCO	Arlington, VA

APPENDIX E. GLOSSARY

ab initio	From first principles
AFES	An atmospheric general circulation model called AFES (AGCM for Earth Simulator) that was developed and optimized for the architecture of the Earth Simulator (ES)
AGCM	Atmospheric general circulation model for the Earth Simulator (see AFES)
AIST	National Institute of Advanced Industrial Science and Technology (Japan)
ALFLEX	Automatic landing flight experiment
Altix	A series of servers and supercomputers from SGI, using Itanium-2 processors
ApGrid	A partnership for grid computing in the Asia-Pacific region
APU	Array processing unit
ARC	Aerospace Research Center (Japan). JAXA's Institute of Space Technology and Aeronautics is located here.
ARC	Ames Research Center (US)
ASCI	Advanced Super Computing Initiative (US)
CAM	Community atmosphere model
CCSM	Community climate system model
CeMSS	Central mass storage system
CeNSS	Central numerical simulation system
CFD	Computational fluid dynamics
CLM	Community land model
CMOS	Complementary metal-oxide semiconductor
COCO	A high-resolution ocean model for eddy-resolving simulations, especially in high latitudes
COE	Center of excellence (Japan)
CReSS	Cloud-resolving storm simulator
CSIM	Community sea-ice model
CSTP	Council for Science and Technology Policy (Japan)
CTBT	Comprehensive Test Ban Treaty (US)
DARPA	Defense Advanced Research Projects Agency (US)
de facto	In reality or fact, actually

de jure	According to law, right
Diagonalization	Changing a square matrix to diagonal form (with all non-zero elements on the principal diagonal)
DNS	Direct numerical simulation
DOE	Department of Energy (US)
DTU	Data transfer unit
ECCO	Estimating the Circulation and Climate of the Ocean
ECMWF	European Center for Medium-range Weather Forecasting
Eigenvalue	The factor by which a linear transformation multiplies one of its eigenvectors
ESMF	Earth System Modeling Framework
ESRDC	Earth Simulator Research & Development Center (Japan)
FRSGC	Frontier Research Center for Global Change (Japan)
FOW	Finite orbit widths
GAMESS	A program for general ab initio quantum chemistry.
GBytes/sec	10^9 bytes per second
GCEM3D	Goddard Cumulus Ensemble Model in 3D
GDP	Gross Domestic Product
GeoFEM	Multi-purpose/multi-physics parallel finite element simulator/platform for solid Earth
Gflops/s	Gigaflops or billions of floating point operations per second
GNSS	Global non-hydrostatic simulation system that is primarily used to investigate cloud activity in tropical areas
Grape chip	Gravity Pipe accelerator chip
Grid Datafarm	A project to build a parallel file system on top of the grid so that users can easily access their files from any location on the grid
GSFC	Goddard Space Flight Center (US)
GSIC	Global Scientific Information and Computing Center (Japan)
GTRC	Grid Technology Research Center (Japan)
HECRTF	High End Computing Revitalization Task Force (US)
Hermite interpolation	A family of polynomials used to interpolate between data points
HPF	High-performance Fortran

HYFLEX	Hypersonic flight experiment
IAI	Independent administrative institutions
IAP	Integrated array processor
Icosahedral	A polyhedron having 20 faces
IDO	Interpolated differential operator
IRD	Internal research and development
IRE	Internal reconnection events
Isomerization	To cause to change into an isomeric form.
ISTA	Institute of Space Technology and Aeronautics (Japan)
ITBL	IT-based laboratory
JAERI	Japan Atomic Energy Research Institute (Japan)
JAMSTEC	Japan Marine Science and Technology Center (Japan)
JASONs	An elite group of scientific advisors who provide the federal government with largely classified analyses on defense and arms controls issues
J-PARC	Japanese Proton Accelerator Research Complex (Japan)
JSPS	Japanese Society for the Promotion of Science (Japan)
Jungle Gym	A three-dimensional network of nanotubes
KEK	High Energy Accelerator Research Organization (Japan)
KISSME	An effort at the Frontier Research Center for Global Change (FRSGC) to develop an integrated Earth system model, primarily for global warming prediction (Japan)
Lagrangian	The Lagrangian of a system is defined as L = T-V, where T is the total kinetic energy and V is the total potential energy.
LES	Large eddy simulation
LHD	Large helical device
LDRD	Laboratory-directed research and development
MDM	Molecular dynamics machine
meso-scale	A medium scale simulation, in between micro scale and conventional
METI	Ministry of Economy, Trade and Industry (Japan)
MEXT	Ministry of Education, Culture, Sports, Science and Technology (Japan)
MIC	Ministry of Internal Affairs and Communications (Japan)

MPI	Message passing interface
Myrinet Networx	Myrinet is a packet-communication and switching technology that is widely used to interconnect clusters. Linux Networx is a also a switching technology for clusters
NAL	National Aerospace Laboratory (Japan)
NAREGI	National Research Grid Initiative (Japan)
NASA	National Aeronautics and Space Administration (US)
NASDA	National Space Development Agency, now a part of JAXA (Japan)
NCAR	National Center for Atmospheric Research (US)
NEXT	Numerical Experiment of Tokamaks
Ninf-G	A global grid computing infrastructure from AIST
NICAM	Non-hydrostatic icosahedral atmospheric model
NIFS	National Institute for Fusion Science (Japan)
NOAA	National Atmospheric and Oceanic Administration (US)
NOPP	National Ocean Partnership Program (US)
NSF	National Science Foundation (US)
NSIII	Numerical Simulator III (Japan)
NWT	Numerical wind tunnel (Japan)
OEM	Original equipment manufacturer
OpenMP	An application program interface (API) that may be used to explicitly direct multi-threaded, shared memory parallelism
Opteron	A 64-bit microprocessor from AMD
OREX	Orbital reentry experiment
PARADIGM	Partnership for Advancing Interdisciplinary Global Models
PCs	Personal computers
PDA	Portable digital assistant
PE	Processing element
Pflop/s	10^{15} floating point operations per second
PNC	Power Reactor and Nuclear Fuel Development Corporation (Japan)
PNs	Processor nodes
POP	Parallel ocean program

Prefetch optimizing	The data is transferred beforehand onto memory cache and the transfer to the cache memory is completed while the loop that references the data is performing calculations from previous iterations.
PSEs	Problem solving environments
QCD	Quantum chromodynamics
QCDOC	QCDOC on a chip
QMP	Lattice QCD message passing
RWCP	Real World Computing Project (Japan)
RIKEN	Institute of Physical and Chemical Research (Japan)
RIST	Research Institute for Science & Technology (Japan)
ScaLAPACK	Dense and band matrix software developed by a consortium of U.S. labs
SETI@home	Search for extraterrestrial intelligence. A massive grid effort that uses home PCs to search for patterns in signals received by radio telescopes. The organizers claim that the overall effort can achieve up to 15 Tflop/s at a cost of about $500K.
SPF	The stratospheric platform airship system
SSME	Space shuttle main engine
SST	Supersonic transport
STA	Science and Technology Agency, now a part of the Ministry of Education (Japan)
Stellerator	Another toroidal structure used to contain plasma for fusion research. In a stellarator the helical lines of force are produced by a series of coils which may themselves be helical in shape. But no current is induced in the plasma. (See Tokamak)
SMP	Symmetric multiprocessing
Supercluster	A large computing resource that provides a testbed for grid computing research
SuperSINet	An all-optical research network in Japan
Tflop/s	10^{12} floating point operations per second
Tightbinding	Tight-binding molecular dynamics (TBMD) provides an efficient method for calculating properties of materials
TLOs	Technology Learning Organizations
TME	Task mapping editor (visual work flow).

Tokamak	A toroidal structure used to contain plasma for fusion research. A toroidal field is created by a series of coils evenly spaced around the torus-shaped reactor, and the poloidal field is created by a strong electric current flowing through the plasma. (See stellerator)
TotalView	A software debugger product of Etnus LLC
TVD	Total variation diminishing
URANS	Unsteady Reynolds-averaged Navier-Stokes
Vectorization	Converting program code to take better advantage of the parallelism in a supercomputer.