The Earth Simulator System

By Shinichi HABATA,* Mitsuo YOKOKAWA† and Shigemune KITAWAKI‡

ABSTRACT The Earth Simulator, developed by the Japanese government's initiative "Earth Simulator Project," is a highly parallel vector supercomputer system that consists of 640 processor nodes and interconnection network. The processor node is a shared memory parallel vector supercomputer, in which 8 vector processors that can deliver 8GFLOPS are tightly connected to a shared memory with a peak performance of 64GFLOPS. The interconnection network is a huge non-blocking crossbar switch linking 640 processor nodes and supports for global addressing and synchronization. The aggregate peak vector performance of the Earth Simulator is 40TFLOPS, and the intercommunication bandwidth between every two processor nodes is 12.3GB/s in each direction. The aggregate switching capacity of the interconnection network is 7.87TB/s. To realize a high-performance and high-efficiency computer system, three architectural features are applied in the Earth Simulator; vector processor, shared-memory and high-bandwidth non-blocking interconnection crossbar network. The Earth Simulator achieved 35.86TFLOPS, or 87.5% of peak performance of the system, in LINPACK benchmark, and has been proven as the most powerful supercomputer in the world. It also achieved 26.58TFLOPS, or 64.9% of peak performance of the system, for a global atmospheric circulation model with the spectral method. This record-breaking sustained performance makes this innovative system a very effective scientific tool for providing solutions to the sustainable development of humankind and its symbiosis with the planet earth.

KEYWORDS **Supercomputer, HPC (High Performance Computing), Parallel processing, Shared memory, Distributed memory, Vector processor, Crossbar network**

1. INTRODUCTION

The Japanese government's initiative "Earth Simulator project" started in 1997. Its target was to promote research for global change predictions by using computer simulation. One of the most important project activities was to develop a supercomputer at least 1,000 times as powerful as the fastest one available in the beginning of "Earth Simulator project." After five years of research and development activities, the Earth Simulator was completed and came into operation at the end of February 2002. Within only two months from the beginning of its operational run, the Earth Simulator was proven to be the most powerful supercomputer in the world by achieving 35.86TFLOPS, or 87.5% of peak performance of the system, in LINPACK benchmark.

It also achieved 26.58TFLOPS, or 64.9% of peak performance of the system, for a global atmospheric circulation model with the spectral method. This simulation model was also developed in the "Earth Simulator project."

*Computers Division

†National Institute of Advanced Industrial Science and Technology (AIST)

‡Earth Simulator Center, Japan Marine Science and Technology Center

2. SYSTEM OVERVIEW

The Earth Simulator is a highly parallel vector supercomputer system consisting of 640 processor nodes (PN) and interconnection network (IN) (Fig. 1). Each processor node is a shared memory parallel vector supercomputer, in which 8 arithmetic processors (AP) are tightly connected to a main memory system (MS). AP is a vector processor, which can deliver 8GFLOPS, and MS is a shared memory of 16GB. The total system thus comprises 5,120 APs

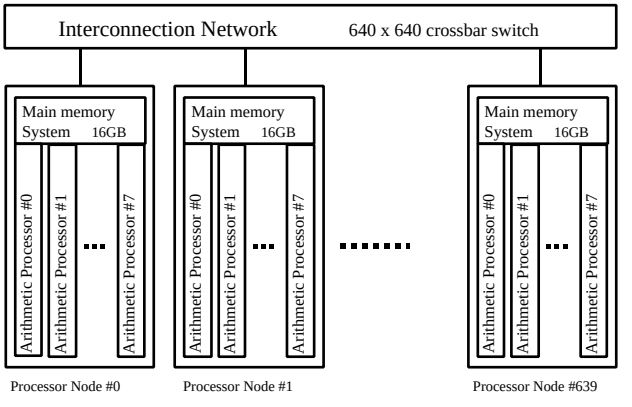

Fig. 1 General configuration of the Earth Simulator.

and 640 MSs, and the aggregate peak vector performance and memory capacity of the Earth Simulator are 40TFLOPS and 10TB, respectively.

The interconnection network is a huge 640 × 640 non-blocking crossbar switch linking 640 PNs. The interconnection bandwidth between every two PNs is 12.3GB/s in each direction. The aggregate switching capacity of the interconnection network is 7.87TB/s.

The operating system manages the Earth Simulator as a two-level cluster system, called the Super-Cluster System. In the Super-Cluster System, 640 nodes are grouped into 40 clusters (Fig. 2). Each cluster consists of 16 processor nodes, a Cluster Control Station (CCS), an I/O Control Station (IOCS) and system disks. Each CCS controls 16 processor nodes and an IOCS. The IOCS is in charge of data transfer between the system disk and Mass Storage System, and transfers user file data from the Mass Storage System to a system disk, and the processing result from a system disk to the Mass Storage System.

The supervisor, called the Super-Cluster Control Station (SCCS), manages all 40 clusters, and provides a Single System Image (SSI) operational environment. For efficient resource management and job control, 40 clusters are classified as one S-cluster and 39 L-clusters. In the S-cluster, two nodes are used for interactive use and another for small-size batch jobs.

User disks are connected only to S-cluster nodes, and used for storing user files. The aggregate capacities of system disks and user disks are 415TB and 225TB, respectively. The Mass Storage System is a cartridge tape library system, and its capacity is more than 1.5PB.

To realize a high-performance and high-efficiency computer system, three architectural features are applied in the Earth Simulator:

- Vector processor
- Shared memory
- High-bandwidth and non-blocking interconnection crossbar network

From the standpoint of parallel programming, three levels of parallelizing paradigms are provided to gain high-sustained performance:

- Vector processing on a processor
- Parallel processing with shared memory within a node
- Parallel processing among distributed nodes via the interconnection network.

3. PROCESSOR NODE

The processor node is a shared memory parallel vector supercomputer, in which 8 arithmetic processors which can deliver 8GFLOPS are tightly connected to a main memory system with a peak

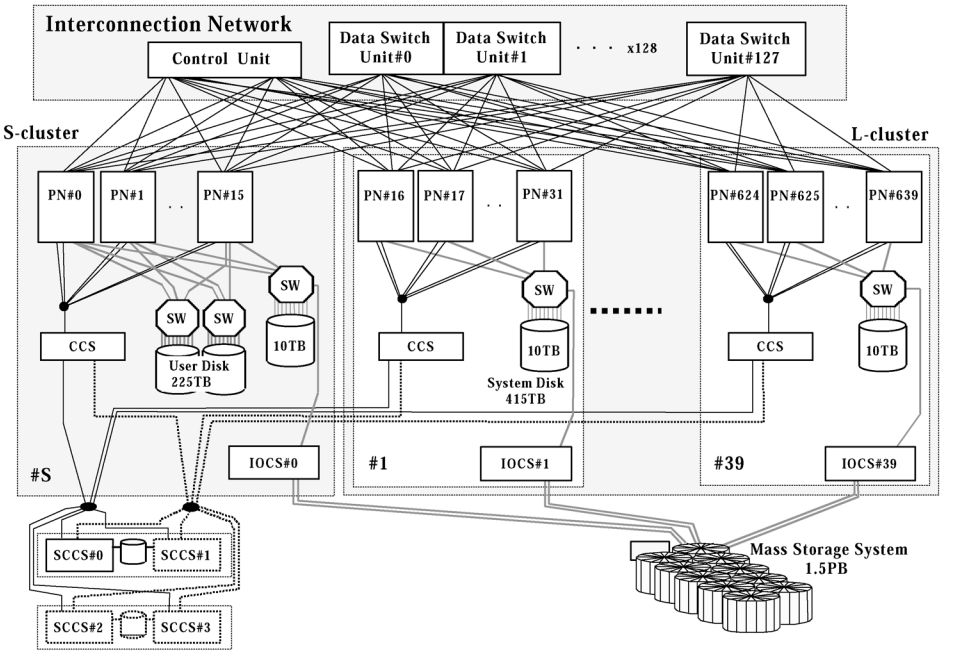

Fig. 2 System configuration of the Earth Simulator.

performance of 64GFLOPS (Fig. 3). It consists of 8 arithmetic processors, a main memory system, a Remote-access Control Unit (RCU), and an I/O processor (IOP).

The most advanced hardware technologies are adopted in the Earth Simulator development. One of the main features is a "One chip Vector Processor" with a peak performance of 8GFLOPS. This highly integrated LSI is fabricated by using 0.15μm CMOS technology with copper interconnection. The vector pipeline units on the chip operate at 1GHz, while other parts including external interface circuits operate at 500MHz. The Earth Simulator has ordinary air cooling, although the processor chip dissipates approximately 140W, because a high-efficiency heat sink using heat pipe is adopted.

A high-speed main memory device was also developed for reducing the memory access latency and access cycle time. In addition, 500MHz source synchronous transmission is used for the data transfer between the processor and main memory to increase the memory bandwidth. The data transfer rate between the processor and main memory is 32GB/s, and the aggregate bandwidth of the main memory is 256GB/s.

Two levels of parallel programming paradigms are provided within a processor node:

- Vector processing on a processor
- Parallel processing with shared memory

4. INTERCONNECTION NETWORK

The interconnection network is a huge 640 × 640 non-blocking crossbar switch, supporting global addressing and synchronization.

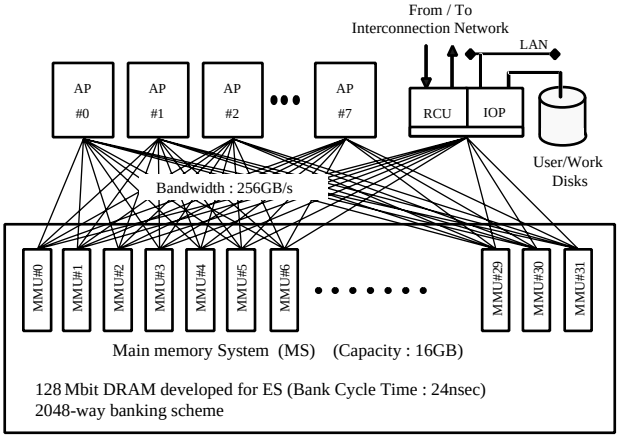

Fig. 3 Processor node configuration.

To realize such a huge crossbar switch, a byte-slicing technique is applied. Thus, the huge 640 × 640 non-blocking crossbar switch is divided into a control unit and 128 data switch units as shown in Fig. 4. Each data switch unit is a one-byte width 640 × 640 non-blocking crossbar switch.

Physically, the Earth Simulator comprises 320 PN cabinets and 65 IN cabinets. Each PN cabinet contains two processor nodes, and the 65 IN cabinets contain the interconnection network. These PN and IN cabinets are installed in the building which is 65m long and 50m wide (Fig. 5). The interconnection network is positioned in the center of the computer room. The area occupied by the interconnection network is approximately 180m^2, or 14m long and 13m wide, and the Earth Simulator occupies an area of approximately 1,600m^2, or 41m long and 40m wide.

No supervisor exists in the interconnection network. The control unit and 128 data switch units operate asynchronously, so a processor node controls the total sequence of inter-node communication. For

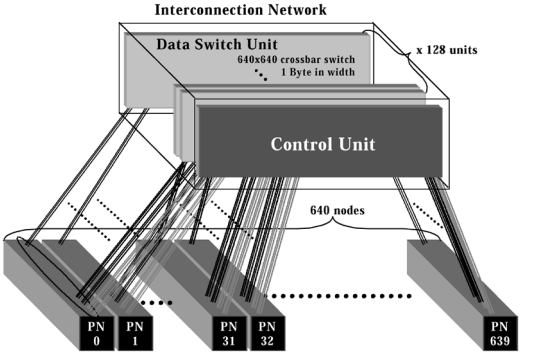

Fig. 4 Interconnection network configuration.

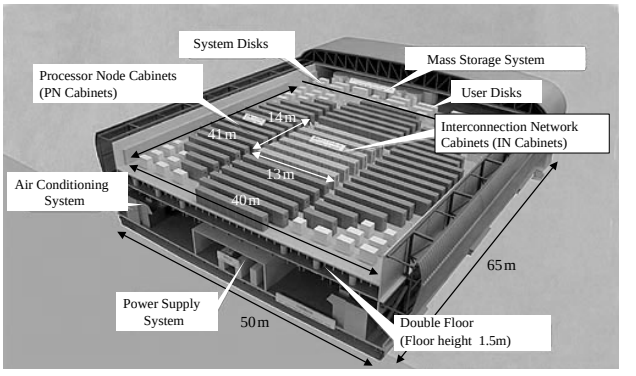

Fig. 5 Bird's-eye view of the Earth Simulator building.

example, the sequence of data transfer from node A to node B is shown in Fig. 6.

1) Node A requests the control unit to reserve a data path from node A to node B, and the control unit reserves the data path, then replies to node A.
2) Node A begins data transfer to node B.
3) Node B receives all the data, then sends the data transfer completion code to node A.

In the Earth Simulator, 83,200 pairs of 1.25GHz serial transmission through copper cable are used for realizing the aggregate switching capacity of 7.87TB/s, and 130 pairs are used for connecting a processor node and the interconnection network. Thus, the error occurrence rate cannot be ignored for realizing stable inter-node communication. To resolve this error occurrence rate problem, ECC codes are added to the transfer data as shown in Fig. 7. Thus, a receiver node detects the occurrence of intermittent inter-node communication failure by checking ECC codes, and the error byte data can almost always be corrected by RCU within the receiver node.

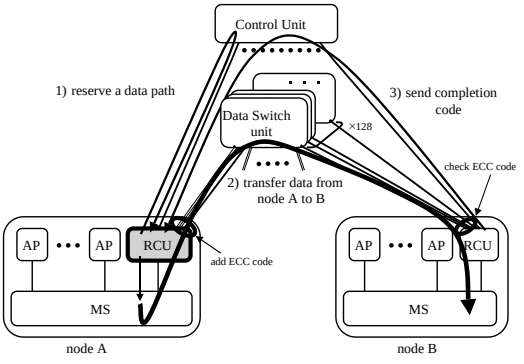

Fig. 6 Inter-node communication mechanism.

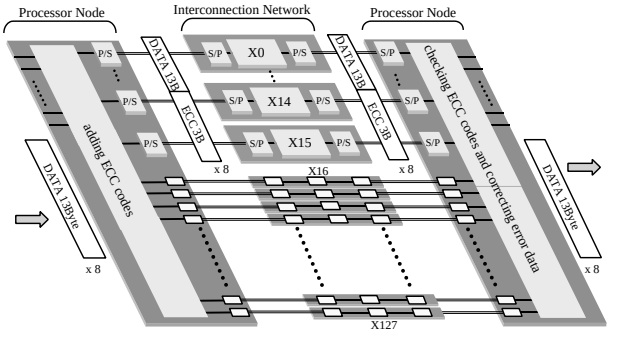

Fig. 7 Inter-node interface with ECC codes.

ECC codes are also used for recovering from a continuous inter-node communication failure resulting from a data switch unit malfunction. In this case the error byte data are continuously corrected by RCU within any receiver node until the broken data switch unit is repaired.

To realize high-speed parallel processing synchronization among nodes, a special feature is introduced. Counters within the interconnection network's control unit, called Global Barrier Counter (GBC), and flag registers within a processor node, called Global Barrier Flag (GBF), are used. This barrier synchronization mechanism is shown in Fig. 8.

1) The master node sets the number of nodes used for the parallel program into GBC within the IN's control unit.
2) The control unit resets all GBF's of the nodes used for the program.
3) The node, on which task completes, decrements GBC within the control unit, and repeats to check GBF until GBF is asserted.
4) When GBC = 0, the control unit asserts all GBF's of the nodes used for the program.
5) All the nodes begin to process the next tasks.

Therefore the barrier synchronization time is constantly less than 3.5μsec, the number of nodes varying from 2 to 512, as shown in Fig. 9.

5. PERFORMANCE

Several performances of the interconnection network and LINPACK benchmark have been measured.

First, the interconnection bandwidth between

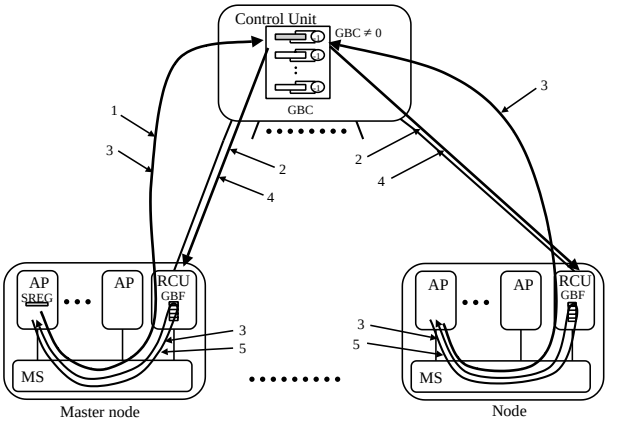

Fig. 8 Barrier synchronization mechanism using GBC.

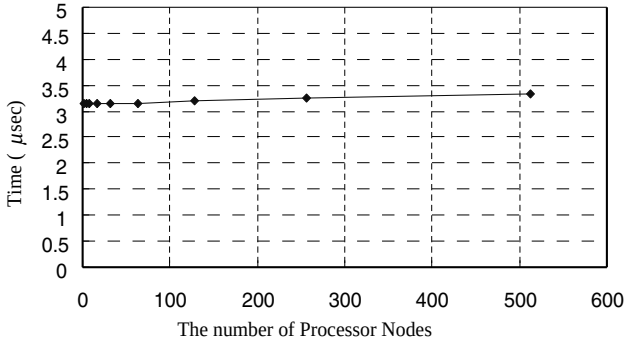

Fig. 9 Scalability of MPI_Barrier.

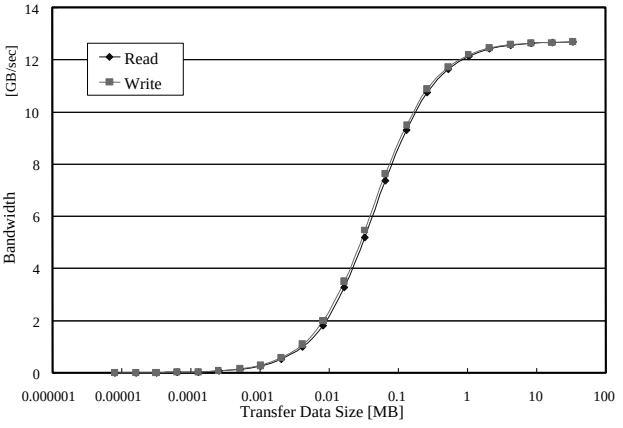

Fig. 10 Inter-node communication bandwidth.

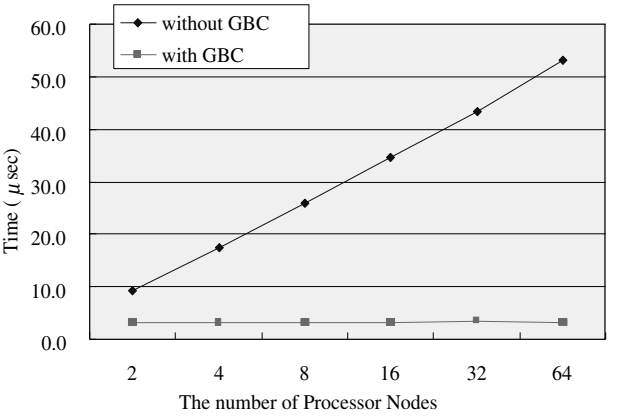

Fig. 11 Scalability of MPI_Barrier.

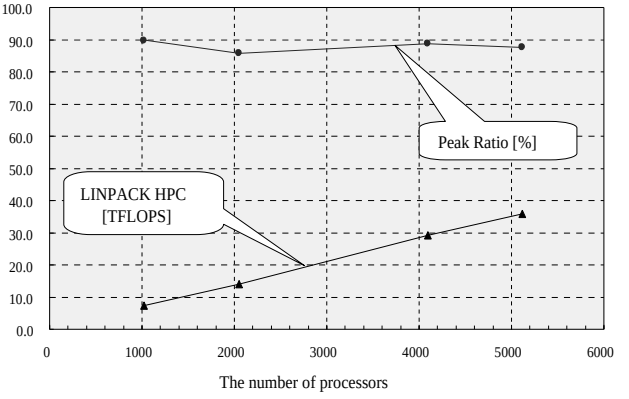

Fig. 12 LINPACK performance and peak ratio.

every two nodes is shown in Fig. 10. The horizontal axis shows data size, and the vertical axis shows bandwidth. Because the interconnection network is a single-stage crossbar switch, this performance is always achieved for two-node communication.

Second, the effectiveness of the Global Barrier Counter is summarized in Fig. 11. The horizontal axis shows the number of processor nodes, and the Vertical axis shows MPI_barrier synchronization time. The line labeled "with GBC" shows the barrier synchronization time using GBC, and "without GBC" shows the software barrier synchronization. By using GBC feature, MPI_barrier synchronization time is constantly less than 3.5μsec. On the other hand, the software barrier synchronization time increases, or is proportional to the number of nodes.

The performance of LINPACK benchmark is measured to verify that the Earth Simulator is a high-performance and high-efficiency computer system. The result is shown in Fig. 12, the number of processors varying from 1,024 to 5,120. The ratio of peak performance is more than 85%, and LINPACK performance is proportional to the number of processors.

6. CONCLUSION

An overview of the Earth Simulator system has been given in this paper, focusing on its three architectural features and hardware realization, especially the details of the interconnection network.

Three architectural features (i.e.: vector processor, shared memory and high-bandwidth non-blocking interconnection crossbar network) provide three levels of parallel programming paradigms; vector processing on a processor, parallel processing with shared memory within a node and parallel processing among distributed nodes via the interconnection network. The Earth Simulator was thus proven to be the most powerful supercomputer, achieving 35.86TFLOPS, or 87.5% of peak performance of the system, in

LINPACK benchmark, and 26.58TFLOPS, or 64.9% of peak performance of the system, for a global atmospheric circulation model with the spectral method.

This record-breaking sustained performance make this innovative system a very effective scientific tool for providing solutions for the sustainable development of humankind and its symbiosis with the planet earth.

ACKNOWLEDGMENTS

The authors would like to offer their sincere condolences for the late Dr. Hajime Miyoshi, who initiated the Earth Simulator project with outstanding leadership.

The authors would also like to thank all members who were engaged in the Earth Simulator project, and especially H. Uehara and J. Li for performance measurement.

REFERENCES

[1] M. Yokokawa, S. Shingu, et al., "Performance Estimation of the Earth Simulator," Towards Teracomputing, Proc. of 8th ECMWF Workshop, pp.34-53, World Scientific, 1998.

[2] K. Yoshida and S. Shingu, "Research and development of the Earth Simulator," Proc. of 9th ECMWF Workshop, pp.1-13, World Scientific, 2000.

[3] M. Yokokawa, S. Shingu, et al., "A 26.58 TFLOPS Global Atmospheric Simulation with the Spectral Transform Method on the Earth Simulator," SC2002, 2002.

Received November 1, 2002

* * * * * * * * * * * * * * * *

Shinichi HABATA received his B.E. degree in electrical engineering and M.E. degree in information engineering from Hokkaido University in 1980 and 1982, respectively. He joined NEC Corporation in 1982, and is now Manager of the 4th Engineering Department, Computers Division. He is engaged in the development of supercomputer hardware.

Mr. Habata is a member of the Information Processing Society of Japan.

Mitsuo YOKOKAWA received his B.S. and M.S. degrees in mathematics from Tsukuba University in 1982 and 1984, respectively. He also received a D.E. degree in computer engineering from Tsukuba University in 1991. He joined the Japan Atomic Energy Research Institute in 1984, and was engaged in high performance computation in nuclear engineering. He was engaged in the development of the Earth Simulator from 1996 to 2002. He has joined the National Institute of Advanced Industrial Science and Technology (AIST) in 2002, and is Deputy Director of the Grid Technology Research Center of AIST. He was a visiting researcher of the Theory Center at Cornell University in the US from 1994 to 1995.

Dr. Yokokawa is a member of the Information Processing Society of Japan and the Japan Society for Industrial and Applied Mathematics.

Shigemune KITAWAKI received his B.S. degree and M.S. degree in mathematical engineering from Kyoto University in 1966 and 1968, respectively. He joined NEC Corporation in 1968. He was then involved in developing compilers for NEC's computers. He joined Earth Simulator project in 1998, and is now Group leader of "User Support Group" of the Earth Simulator Center, Japan Marine Science and Technology Center.

Mr. Kitawaki belongs to the Information Processing Society of Japan and the Association for Computing Machinery.

* * * * * * * * * * * * * * * *