The Development of the Earth Simulator

Shinichi HABATA†, Nonmember, Mitsuo YOKOKAWA††, Nonmember and Shigemune KITAWAKI†††, Nonmember

Summary

The Earth Simulator (ES), developed by the Japanese government’s initiative “Earth Simulator project”, is a highly parallel vector supercomputer system.

In May 2002, the ES was proven to be the most powerful computer in the world by achieving 35.86 teraflops on the LINPACK benchmark and 26.58 teraflops for a global atmospheric circulation model with the spectral method. Three architectural features enable these great achievements; vector processor, shared-memory and high-bandwidth non-blocking interconnection crossbar network.

In this paper, an overview of ES, the three architectural features and the result of performance evaluation are described, in particular its hardware realization of the interconnection among 640 processor nodes.

Key words: Supercomputer, HPC (High Performance Computing), Parallel processing, Shared 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 the global change prediction by using computer simulation. One of the most important project activities was to develop a supercomputer, which could achieve at least 5 teraflops for a global atmospheric circulation model with the spectral method. After five years of research and development activities, the Earth Simulator (ES) was completed and came into operation at the end of February 2002. Within only two months from the beginning of the operational run, ES was proven as the most powerful supercomputer in the world by achieving 35.86 teraflops, or 87.5% of the system peak performance, in LINPACK benchmark. Moreover, ES achieved 26.58 teraflops, or 64.9% of system peak performance, for a global atmospheric circulation model with the spectral method. This simulation model, named AFES, was extensively optimized for ES in the “Earth Simulator project”.

2. System Overview

The Earth Simulator is a highly parallel vector supercomputer system, which consists of 640 processor nodes (PN) and interconnection network (IN) as shown in Fig.1. PN 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 8 gigaflops, and MS is a shared memory of 16GB. So the total system comprises 5120 AP’s and 640 MS’s, and the aggregate peak vector performance and memory capacity of ES are 40 teraflops and 10TB, respectively.

The interconnection network is a huge 640 x 640 non-blocking crossbar switch linking 640 PN’s. The interconnection bandwidth between every two PN’s is 12.3GB/s in bidirection. The aggregate switching capacity of the interconnection network is 7.87TB/s.

The ES operating system is based on UNIX System V with BSD features and manages ES as two-level cluster system, which called Super-Cluster System. In 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. IOCS controls file recall/migration between system disks and Mass Storage System.

The supervisor, called Super-Cluster Control Station (SCCS), manages all 40 clusters, and provides 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 S-cluster, two nodes are used for interactive use, such as compiling and debugging, and the other for small size batch jobs.

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

In L-cluster, user’s files are stored in user disks or Mass Storage System, so user’s files must be automatically recalled from user disks or Mass Storage System to system disks prior to the job execution and migrated from system disks to user disks or Mass Storage System afterwards. This recall and migration function is performed by an S-cluster node or an IOCS and called “Automated file recall and migration”, its general data movement is summarized in Fig.3.

In the ES programming environment, the MPI library (MPI/ES) and the HPF compiler (HPF/ES) are available for distributed parallel programming. MPI/ES conforms to the MPI 2.0 standard, and HPF/ES complies with the core part of HPF 2.0 and supports some features of the HPF 2.0 approved extensions and HPF/JA 1.0 extensions.

To realize 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 an AP (vectorization)
Parallel processing with MS within a PN (Shared memory parallelization)

Parallel processing among distributed PN’s via the interconnection network. (Distributed memory parallelization)

3. Processor Node

The processor node consists of 8 arithmetic processors (AP), a main memory system (MS), a Remote-access Control Unit (RCU), and an I/O processor (IOP) as shown in Fig.4. The main memory system comprises 32 main memory units (MMU), and 2048-way banking scheme is applied. Each MMU consists of a memory controller and 24 high-speed 128Mbit memory LSI’s, and 64-way banking scheme is applied. The memory capacity of MMU is 512MB. To reduce the memory access latency and access cycle time, high-speed memory device, called full-pipeline memory, is developed.

The arithmetic processor is configured with a vector unit and a scalar unit. The vector unit logically comprises 8 sets of vector pipelines synchronously operating in parallel. Each set is configured with four types of pipelines: add/shift, multiply, divide and logical operation, these four pipelines in each pipeline set operate independently (Fig.5).

The most advanced CMOS LSI technology is adopted for realizing AP, so one of the characteristic features is “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 (Fig.6). The arithmetic processor physically consists of 4 sets of vector pipelines, which operate at 1GHz, while other parts including external interface circuits at 500MHz. A physical vector pipeline set operates twice as much data as other units, so a physical vector pipeline processes two sets of logical vector pipeline data in time-sharing fashion.

In ES all packages and modules are air-cooled, though the processor chip dissipates approximately 140W, because a high-efficiency heat sink utilizing heat siphon principle is developed (Fig.7).
In addition, 500MHz source-synchronous transmission is used for the data transfer between an AP and MS to increase the memory access bandwidth. The data transfer rate between an AP and MS is 32GB/s, and the aggregate data bandwidth of MS is 256GB/s.

Two levels of parallel programming paradigms are provided within a PN:

- Vector processing on an AP (vectorization)
- Parallel processing with MS (Shared memory parallelization)

4. Interconnection Network

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

To realize such a huge crossbar switch, byte-slicing technique is applied. So a huge 640 x 640 non-blocking crossbar switch is divided into a control unit and 128 data switch units as shown in Fig.8. Each data switch unit is one-byte width 640 x 640 non-blocking crossbar switch.

Inter-node connection between a processor node and interconnection network is realized by using 1.25GHz serial transmission through copper cable. 130 pairs of shielded electric cables with differential four-core quad are used for connecting a processor node and the interconnection network (IN), so for linking 640 nodes to IN, 83200 pairs used, and the aggregate switching capacity of IN is 7.87TB/s.

83200 pairs of 1.25GHz serial transmission are used for interconnection among 640 PN's, so the error occurrence rate can't be ignored for realizing stable inter-node communication. To resolve this error occurrence rate problem, ECC codes are added into transfer data as shown in Fig.9. A receiver node detects the occurrence of intermittent inter-node communication failure by checking ECC codes, and almost always the error byte data can be corrected by RCU within the receiver node.

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.

No supervisor exists in the interconnection network, the control unit and 128 data switch units asynchronously operate in parallel, so a processor node controls the total sequence of inter-node communication. For example, the sequence of data transfer from node A to node B is shown in Fig.10.

- Node A requires the control unit to reserve a data path from node A to node B, and the control unit reserves the data path, then replies for node A.
- Node A begins data transfer to node B.
Node B receives all the data, then sends the data transfer completion code to node A.

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 for this barrier synchronization, and its mechanism is shown in Fig.11.

Master node sets the number of nodes used for the parallel program into GBC within the IN’s control unit.

- The control unit resets all GBF’s of the nodes used for the program.
- The node, on which task completes, decrements GBC within the control unit, and repeats to check GBF until GBF is asserted.
• When GBC=0, the control unit asserts all GBF’s of the nodes used for the program.
• All the nodes begin to process next tasks.

5. System Installation

Physically, the Earth Simulator comprises 320 PN cabinets and 65 IN ones (Fig.12). Each PN cabinet contains two processor nodes, and the interconnection network control unit and 128 data switch units are contained in 65 IN cabinets. One IN cabinet contains the control unit, and each of the other contains two data switch units. These PN and IN cabinets are installed in the building, which is 65m long and 50m wide as shown in Fig.13. The interconnection network is positioned in the center of computer room. The area occupied by the interconnection network is approximately 180 m², or 14m long and 13m wide, and the Earth Simulator occupies the area of approximately 1600 m², or 41m long and 40m wide. The setting of 83200 inter-node communication cables (approximately 2,400km in total) is one of the most difficult problems to realize ES. So the cable...
route and setting order are determined by using the customized routing program, which derived from the automatic router for the wiring board. Fig.15 shows the result of cable setting simulation. Using this simulator, cable length and height of installed cables were estimated by seeking optimal cable routes and by changing order of installation of each cable. As a result, the inter-node cable installation was completed in three months. One-tenth scale model of the double floor was constructed for evaluating the order of cable setting and the height of cable piled in double floor (Fig.16).

6. Performance

Several performances of the interconnection network and LINPACK benchmark are measured.

First, the interconnection bandwidth of point-to-point data transfer is shown in Fig.16. The horizontal axis shows data size and the vertical axis shows bandwidth. Because the interconnection network is a single-stage crossbar switch, always this performance is achieved for point-to-point data transfer.

The result of MPI_barrier call time comparison between software barrier synchronization and barrier synchronization using GBC is shown in Fig.19. The horizontal axis shows the number of processor nodes, and the vertical axis shows MPI_barrier call time. The line labeled “with GBC” shows MPI_barrier call time using GBC, and “without GBC” shows the software barrier synchronization. By using GBC feature, MPI_barrier call time is constantly less than 3.3 µsec. On the other hand, the software barrier synchronization time increases from 9 µsec to 53 µsec, proportional to the number of nodes.

The performance of LINPACK benchmark is measured to verify the Earth Simulator is high performance and high efficiency computer system. The result is shown in Fig.20, varying the number of processors from 1024 to 5120. The ratio of peak performance is more than 85%, and LINPACK performance is proportional to the number of arithmetic processors.

The result of scalability measurement of AFES, a global atmospheric model with the spectral method, is shown in Fig.21. The simulation with 80, 160, and 320
arithmetic processors are based on the parallelism with MPI, while those with 640, 1280, 2560, and 5120 arithmetic processors are for MPI-based parallelization in combination with 8 intra-node microtasking processes.

7. Conclusion

An overview of the Earth Simulator system is described in this paper, in particular its three architectural features and hardware realization, especially the detail of the interconnection network.

Three architectural features, or vector processor, shared memory and high-bandwidth non-blocking interconnection crossbar network, provide three-level 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. Therefore the Earth Simulator was proven as the most powerful supercomputer by achieving 35.86 TFLOPS, or 87.5% of peak performance of the system, in LINPACK benchmark, and 26.58 TFLOPS, or 64.9% of peak performance of the system, for a global atmospheric circulation model with the spectral method.

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

Acknowledgments

The authors would like to offer sincere condolences to late Dr. Hajime Miyoshi who initiated the Earth Simulator project with outstanding leadership. The authors also would like to thank all members who were engaged in the Earth Simulator project, and specially thanks to H. Uehara, J. Li for performance measurement.

References


