doi: 10.32131/jes.5.20

Journal of the Earth Simulator, Volume 5, Mar. 2006, 20 – 24

MDPS: The New Mass Data Processing Storage System for the Earth Simulator

Ken'ichi Itakura

The Earth Simulator Center, Japan Agency for Marine-Earth Science and Technology,
3173-25 Showa-machi, Kanazawa, Yokohama, 236-0001, Japan
E-mail: itakura@jamstec.go.jp

(Received February 15, 2006; Revised manuscript accepted March 30, 2006)

Abstract The Earth Simulator is one of the most powerful supercomputers in the world. Its high effective performance generates unprecedentedly huge volumes of data. We therefore needed to expand its storage system. The Mass Data Storage System (MDPS), introduced in October 2003, provides a vast permanent data storage area plus high-speed input and output data transfer between the MDPS and temporary disks at processing nodes. In this paper, we discuss the mass data storage system and describe the introduced ES storage system.

Keywords: Earth Simulator, Mass data processing, Storage system, Hierarchy hybrid file system

1. Introduction

The Earth Simulator, which has been under development since 1997, came into operation at the end of February 2002[1]. It is designed to promote research into global change predictions by using computer simulations. In its first year of operation, the Earth Simulator was proven to be the most powerful supercomputer in the world on achieving 35.86 Tflops, or 87.5% of its peak system performance, according to the LINPACK benchmark[2]. Moreover, the product application programs run on the Earth Simulator with much higher effective performance. For example, a simulation product of a global atmospheric circulation model achieved 26.58 Tflops, or 64.9% of peak performance[3].

Higher effective performance generates unprecedentedly huge volumes of data, leading to a need to expand the Earth Simulator's storage system. The Mass Data Processing System (MDPS), adopted in October 2003, provides a huge permanent data stored area and high-speed input and output data transfer between the MDPS and temporary disks at processing nodes. The MDPS required not only a huge volume of storage capacity but also needed to act as a useful storage system for ES users. The original ES system had a large data storage area consisting of a cartridge tape library system and dedicated filing system software. However, this library system often suffered problems due to mechanical problems with the drives or the tape media. In the worst cases, the users lost all their data. Although there was a data backup system in the original file system software, it was not a default option and hardly anyone used it. Users employed mainly the sub-storage area. This caused some capacity problems and sometimes the wrong ES nodes were scheduled to execute user-submitted requests.

The MDPS needs to feature a total system that can cope with the huge storage capacities required allow the transfer of large volumes needed when a user application program is running on the ES nodes as well as permitting access by the user during pre- and post-processing of the simulation.

Section 2 gives an overview of the ES system. In Section 3, we discuss the requirements of a data storage system, and in Section 4, we describe the new MDPS data storage system. Section 5 lists our conclusions.

2. Earth Simulator System Overview

2.1 Cluster System

The Earth Simulator is a highly parallel vector supercomputer system consisting of 640 processor nodes (PNs) and an interconnection network (IN)[4]. Figure 1 is a model photo of the installed whole system. The Earth Simulator consists of 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 a building 65 m long and 50 m wide. The interconnection network is positioned at the center of the computer room. The area occupied by the interconnection network is approximately 180 m^2, or an area 14 m long by 13 m wide, and the Earth Simulator occupies an area of approximately 1600 m^2, or 41 m long by 40 m wide.

There are two processor nodes in each PN cabinet;

however, they are independent of each other except for sharing the same power unit. Each processor node has 8 arithmetic processors (APs) and a main memory system (MS) shared by APs. Each AP is a vector processor that can deliver 8 Gflops, and the MS is a shared memory of 16 GB. The total system thus comprises 5,120 APs and 640 MSs, and the aggregate peak vector performance and memory capacity of the Earth Simulator are 40 Tflops and 10 TB, respectively.

The IN is a huge 640 x 640 non-blocking crossbar switch linking 640 PNs. The interconnection bandwidth between every two PNs is 12.3 GB/s in each direction. The aggregate switching capacity of the interconnection network is 7.87 TB/s. Cluster systems maintain the PNs. The number of clusters is 40, and each cluster has 16 PNs. There is a Cluster Control Station (CSS) and an I/O Control Station (IOCS) in each cluster. The PNs in each cluster share a storage system. More information on the changes resulting from the newly adopted MDPS is available in Section 4. Each CCS controls 16 PNs during booting up, shutdown, and collecting system messages.

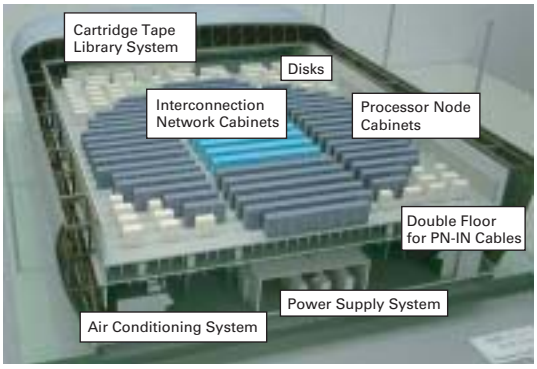


Fig. 1 A model of the ES system in a gym-like building. The building is 50 m x 65 m x 17 m and has two stories; it includes a seismic isolation system.

The operator can maintain the PNs via the CCS in remote control room. One further level, a 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 into one S-cluster and 39 L-clusters. In the S-cluster, two nodes are used for interactive use and the others are used for small-size batch jobs.

2.2 Processing Nodes

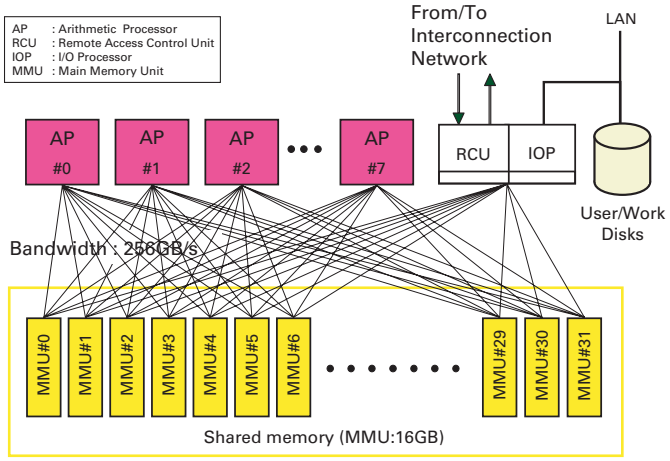
To realize a high-performance and high-efficiency computer system, three architectural features are applied to the Earth Simulator.

- 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

The processor node is a shared memory parallel vector supercomputer, in which 8 arithmetic processors which can deliver 8 Gflops are tightly connected to a main memory system with a peak performance of 64 Gflops (Fig. 2) that consists of 8 arithmetic processors, a main memory system, a Remote-access Control Unit (RCU),



ï 128Mbit developed RAM (24nsec bank cycle time)
ï 2048 banks

Fig. 2 Configuration of PN

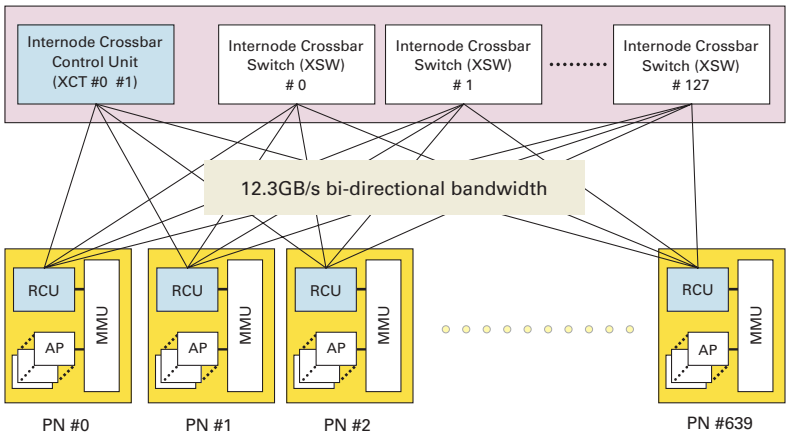


Fig. 3 Configuration of PN

and an I/O processor (IOP).

Extremely 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 8 Gflops. This highly integrated LSI is fabricated using 0.15 μm CMOS technology with copper interconnections.

The vector pipeline units on this chip operate at 1 GHz, while other parts, including external interface circuits, operate at 500 MHz. The Earth Simulator has ordinary air cooling, although the processor chip dissipates approximately 140 W, since a high-efficiency heat sink using heat pipes is adopted.

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

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

2.3 Interconnection Network

The interconnection network is a huge 640 x 640 non-blocking crossbar switch, supporting global addressing and synchronization. To realize this huge crossbar switch, a byte-slicing technique is applied. Thus, the huge 640 x 640 non-blocking crossbar switch is divided into a control unit and 128 data switch units. Each data switch unit is a one-byte type with a 640 x 640 non-blocking crossbar switch. Physically, the Earth Simulator comprises 320 PN cabinets and 64 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 a building 65 m long and 50 m wide. The interconnection network is positioned in the center of the computer room. The area occupied by the interconnection network is approximately 180 m^2, or 14 m long by 13 m wide, and the Earth Simulator occupies an area of approximately 1600 m^2, or 41 m long by 40 m wide.

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

1. Node A requests the control unit to reserve a data path from node A to node B. 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 1.25 GHz serial transmissions through copper cable are used for realizing the aggregate switching capacity of 7.87 TB/s, and 130 pairs are used for connecting the processor nodes with the interconnection network. Thus, to achieve reliable inter-node communication, the error occurrence rate cannot be ignored. To resolve the error occurrence rate problem, error correcting codes (ECCs) are added to the transfer data. Thus, a receiver node detects the occurrence of intermittent inter-node communication failure by checking the ECCs. The error byte data can almost always be corrected by the RCU within the receiver node.

ECCs 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 the RCU within any receiver node until the faulty data switch unit is repaired.

To realize high-speed parallel processing synchroniza-

tion among nodes, a special feature is adopted. Counters within the interconnection network's control unit, called Global Barrier Counters (GBCs), and flag registers within a processor node, called Global Barrier Flags (GBFs), are used. The barrier synchronization time is therefore consistently less than 3.5 μs, with the number of nodes varying from 2 to 512.

3. Requirements of a huge data storage system

3.1 Original System

Figure 5 shows the original storage system used from starting ES operation. There are three partitions for storing user data. Two partitions are managed by the S cluster system and formatted in a standard UNIX file system. However, the third partition is named "/L". The data in the /L partition are stored in a Cartridge Tape Library (CTL) in the original file system format. In the past, users needed to access files in /L using their own commands and to copy between the /L and S cluster file systems. The /L data storage area was not used in the early stages of ES operating history. Most of the user data was stored in the S cluster partitions. In the latter part of 2002, some projects began to need to handle mass simulation data. These large volumes of data were stored in /L. However, accessing /L files was difficult. Frequently used files were then stored in S cluster partitions.

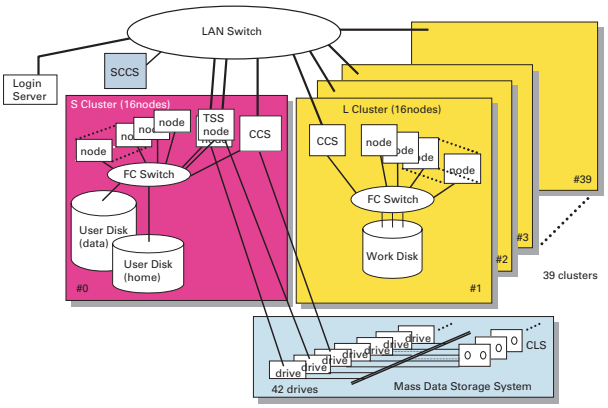


Fig. 4 The Original Permanent Storage System

Executing a job required data transfer from the permanent area to the executing nodes before running the program, followed by the reverse process. The original job scheduler supported pre/post file transfer between /L and executing the node's file systems[5].

On the other hand, data transfer between S cluster partitions and executing nodes takes place just before or after executing a user job script, since it uses IN and execution node CPU power. The job scheduler allocates the starting time and location of executing node based on the job script, in which the user indicates elapsed execution time, number of nodes and disk capacity per node.

3.2 Problems and Solutions

The problems with the original storage system are listed below.

- Cartridge Tape Library-stored mass volume data has low reliability.
- Accessing files in the CTL, which have no directory, is confusing to people accustomed to using a conventional UNIX file system.
- Data transfer between S cluster partitions and L cluster nodes is not separated by running allocation.

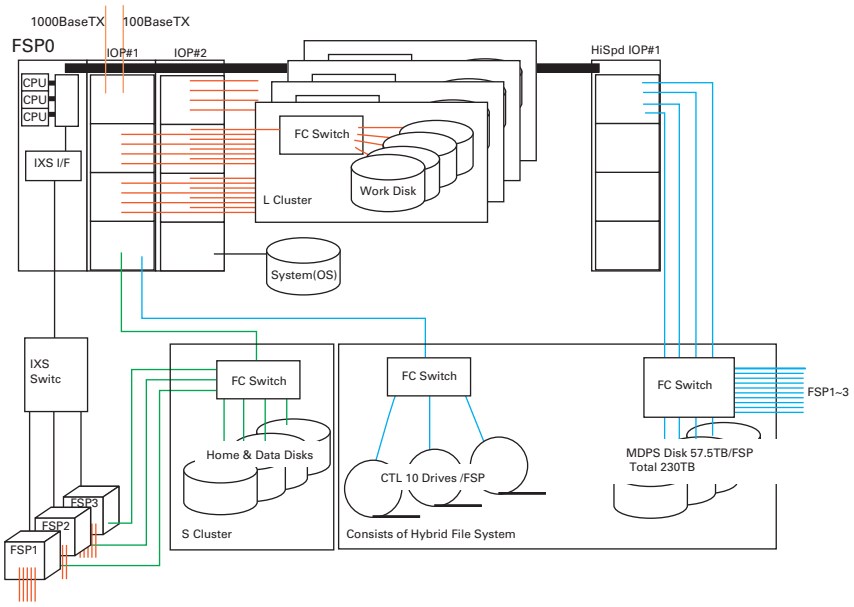


Fig. 5 MDPS Storage System

One of the solutions to the low reliability and unfamiliar CTL is a hierarchy file system. This consists of an HDD as front main storage and CTL as back storage. Users can access the files using a conventional UNIX file system. The management system controls the HDD capacity, and if more space is needed, can swap files or parts of files between the HDD and CTL. To stop pre/post data transfer, one large data file system needs to be shared by all L nodes. However, it is difficult to maintain high throughput. One of the more realistic solutions is the unification of data transfer method between permanent data storage and the temporary type in working nodes.

4. MDPS

We now describe the clusters of File Service Processors (FSPs). One of the roles of FSPs is as a controller of the hybrid file system's hierarchy. A new 5 HDD RAID system (240 TB) and existing CTLs (1.5 PB) are connected to the FSPs. They are called MDPS disks. In this hybrid file system, all files are stored in the CTL in duplex. When the user reads or writes them, files are automatically swapped if necessary. The second of role of the FSPs is to transfer data between permanent data storage and temporary storage in working nodes. FSPs connect not only MDPS disks but also S disk directories. All data transfer is therefore controlled by the FSPs.

The FSPs consist of 4 nodes which have a high I/O throughput, and their OS are close to those on ES nodes. So the ES nodes are connected to HD storage by fiber channels, and the storage is shared by nodes in a cluster. The GFS system provides directory access to node storage from the FSP[6].

There are twenty-five 1-Gbps fiber channel I/Fs and two 2 Gbps versions. Ten 1-Gbps lines are connected to ten clusters and another ten lines are connected to other clusters for backup. System disks, the CTL and the S Cluster disks are connected by two, two, and one, respectively. Two 2 Gbps lines are connected to MDPS HDD.

5. Conclusion

In this paper, we have described the MDPS system, a new permanent storage system for the Earth Simulator system. The ultra-high processing power characteristic of the ES generates huge volumes of data. Handling this data is an extremely difficult and important problem when operating a supercomputer. The MDPS provides high and interactive accessibility from the login node and combines high capacity with good cost performance. The FSPs controls all of the permanent storage. The file transfers at the pre/post stage of NQS user requests are more efficient and the running scheduling is more precise.

(This article is reviewed by Dr. Horst D. Simon.)

References

[1] Mitsuo Yokokawa, "Present Status of Development of the Earth Simulator", Innovative Architecture for Future Generation High-Performance Processors and Systems (IWIA '01), January 18-19, 2001.

[2] http://www.top500.org

[3] Satoru Shingu, Hiroshi Takahara, Hiromitsu Fuchigami, Masayuki Yamada, Yoshinori Tsuda, Wataru Ohfuchi, Yuji Sasaki, Kazuo Kobayashi, Takashi Hagiwara, Shinichi Habata, Mitsuo Yokokawa, Hiroyuki Itoh and Kiyoshi Otsuka, "A 26.58 Tflops Global Atmospheric Simulation using the Spectral Transform Method on the Earth Simulator", SC2002, Nov. 16-22, 2002.

[4] Shinichi Habata, Mitsuo Yokokawa and Shigemune Kitawaki, "The Earth Simulator System", NEC Research & Development, Vol. 44, No. 1, Jan. 2003

[5] Atsuya Uno, Tetsuo Aoyagi and Keiji Tani, "Job Scheduling on the Earth Simulator", NEC Research & Development, Vol. 44, No. 1, Jan. 2003

[6] Atsuhisa Ohtani, Hiroshi Aono and Hiroko Tomaru, "The Earth Simulator System", NEC Research & Development, Vol. 44, No. 1, Jan. 2003