UDC 681.32.055

Hardware of AP3000 Scalar Parallel Server

●Hiroaki Ishihata ●Masanori Takahashi ●Hiroyuki Sato

(Manuscript received April 21,1997)

The AP3000 is a distributed-memory parallel server consisting of multiple workstations connected via a high-speed communication network. Each workstation (node) uses the advanced UltraSPARC CPU and the Solaris operating system. By combining the AP3000 with a remote memory copy function and by performing inter-node communication using Fujitsu's newly developed AP-Net high-speed message communication network, the AP3000 can be used as a high-performance parallel computer and a workstation cluster. The system has hardware to support single-system-image operation of a multi-node system.

1. Introduction

The high-power computer systems used in fields such as research and development, databases, decision making, and multimedia need a system architecture that can quickly and flexibly handle the requirements of various application programs.

The AP3000 parallel server consists of workstations equipped with UltraSPARC 64-bit microprocessors. These workstations are connected as node computers (referred to as nodes hereafter) through the AP-Net (Advanced Parallel Systems Network) high-speed communication network. The Solaris operating system is run on each node so that existing application software can be used. Each node can function as a single UNIX server, or nodes can be combined to form a parallel processor system for ultrahigh-speed performance.

The AP3000 was developed for users in large computer centers, CAD/CAE users, and parallel processing researchers who need to make very large numbers of calculations. In large computer centers where systems must be open, the AP3000 is superior to specific-function multiple servers in terms of easy maintenance, manageability, expandability, and network efficiency. For CAD and CAE, various application programs running on Solaris can be used without modification, and a high throughput can be obtained with simple management. For parallel processing research, the high-speed internode communication function can provide efficient parallel processing capability, and many useful tools for development and evaluation on general-purpose workstations are available. The AP3000 is also suitable for data mining and WWW server use.

This paper explains the AP3000 hardware and its features. The later sections describe the background of the development concept for the design, hardware overview, high-speed communication network AP-Net, and communication functions.

2. Background and objective of the development

Although the recent improvement in CPU performance has been surprisingly rapid, there are areas in which a single CPU is insufficient. In these areas, a parallel processing approach, in which multiple CPUs are used simultaneously, can be effective.

Fujitsu started research into parallel processing in the middle of 1980. Since then, Fujitsu has been developing and providing parallel processing computers . The AP1000, the prototype of the AP3000 series, was well received due to its inno-

vative message communication method[1] and communication network.

Most of the massively parallel computers released to date are no longer in use. This is because they were expensive and designed for parallel processing only. Particularly, to achieve high performance for parallel processing programs, the operating systems of these computers were customized for parallel processing and consequently could not run any of the wide range of software packages available for single CPU computers.

If parallel processing computers are to become popular, they must also be able to run sequential processing programs. Recently, there has been a dramatic drop in the prices of high-performance workstations, and some users now prefer to purchase low-priced workstations because of their wide range of application even though they have insufficient communication capability. However, a high-speed communication feature in a parallel processing computer can efficiently handle the throughput of sequential processing programs. The communication speed between nodes is at least 10 times higher than that of FDDI or ATM in another network. Thus, the communication facilities in a workstation environment such as TCP/IP and NFS can be speeded up, and each node can run application programs at a higher rate.

On the other hand, one problem with workstation clusters is that it is difficult to manage multiple workstations.

To simplify the management of workstation clusters, functions for collectively controlling the power and installation of each workstation are required. Conventionally, the user or SE has to integrate and set up the hardware and software of these functions, and so far there has been no system that bundles these functions.

The AP3000 is a high-throughput parallel processing computer that can also use existing application software resources. The AP3000 uses UltraSPARC workstations as nodes, and therefore can execute the wide range of software already available for these workstations. Also, it has a distributed memory architecture for a scalability that can be several times to several hundred times higher than that of a symmetrical multiprocessor (SMP).

3. Overview of the AP3000

3.1 Hardware design concept

The following four points were regarded as important in the development of the AP3000:

1) Implementation of high performance using multinode parallel processing

In parallel processing, the communication capability between nodes greatly affects performance. For a highly parallel processing performance, low-latency and high-throughput communication networks are essential. To increase the data transmission rate in parallel processing, the AP3000 uses the AP-Net[2] high-speed communication network, which is based on the techniques developed for the AP1000. To enable low-latency and high-throughput communications, the AP3000 employs a message-routing scheme similar to the one used in the AP1000. For low-latency communications, it is important not only to increase the data transmission rate in the network but also to significantly reduce the time needed to setup message communication. Therefore, in the AP3000, a user-level communication system is supported so that the message communication can be activated directly without any help from the operating system.

2) High throughput for existing application programs

The AP3000 uses existing workstations as nodes without any modification so that application software resources that have not been modified for parallel processing can be handled. To handle application programs for distributed processing at a high rate, communication interfaces that are fast and compatible with standard local area networks are required. Therefore, operations such as access to files using NFS or IP (Internet Protocol) communication are routed via AP-Net to accelerate the data transmission rate.

3) Easy system control and management

Large systems (systems with 100 or more workstations) are difficult to manage. Therefore, system managers must be provided with facilities that enable them to simultaneously control the power of all nodes, install nodes, monitor the operating status, and perform other tasks. Also, functions to control the system automatically according to operating schedules should be supported.

4) Support of increases in the number of nodes and I/O channels

The AP3000 must support easy extension of workstation clusters and high-speed communication for parallel processing. For AP-Net, a two-dimensional torus network with high expandability and scalability is used to support from 4 to 1,024 nodes. Also, the AP3000 should be able to quickly adapt to upgrades in workstation (node) performance.

3.2 System configuration

Figure 1 shows the AP3000 system configuration. The control workstation is connected to nodes through the control network (Ethernet)[Note 1)] and AP-Net. The control workstation commands the entire AP3000 system and functions as the remote console and installation server for each node. The system control (SYSCNTL) unit controls power and relays console messages transferred between nodes and the control WS to manage the entire system and to install new nodes.

Figure 2 shows the hardware configuration of the nodes. Table 1 shows the AP3000 specifications, and Table 2 shows the specifications of nodes in the AP3000. **Figure 3** shows a photograph of the AP3000. AP-Net is installed in the network cabinet(s) (small cabinet on the right in Fig. 3), and nodes are installed in the node cabinet(s) (large cabinet on the left). A node cabinet can contain up to eight nodes. A network cabinet can be connected to up to eight node cabinets, enabling a system to have up to 64 nodes. A system is expanded by installing network cabinets for additional nodes. The maximum number of network cabinets is 16, and therefore an AP3000 system can have a maximum of 1,024 nodes.

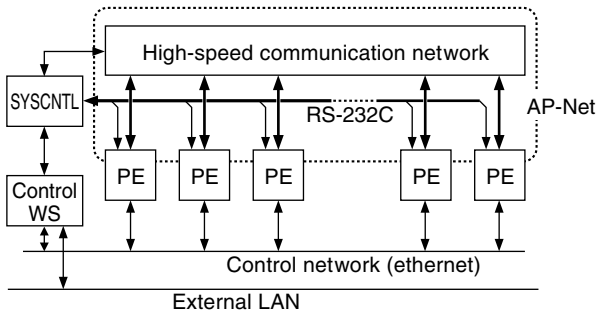

Fig. 1— AP3000 system configuration.

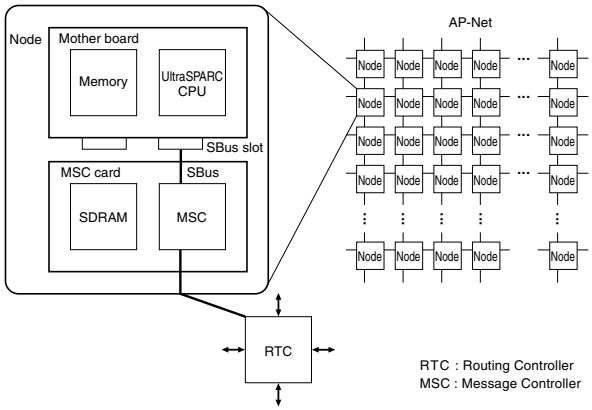

Fig. 2— Hardware configuration of nodes.

Fig. 3— AP3000 external view.

Note 1) Registered trademark of Fuji Xerox

Table 1. AP3000 specifications

Number of nodes	4~1024
Node type	Three types (U140, U170, and U200)
Memory capacity	128 Mbytes to 2 Tbytes
Built-in hard drive	8.4 Gbytes to 4.2 Tbytes
Internal network	AP-Net (200 Mbytes/s × bidirectional)
Connectable external network	Ethernet, Fast Ethernet, FDDI, ATM, and other standard network
Connectable peripheral device	Disk-array device, tape library device, and other standard device
Operating system	Solaris

3.3 Features of AP3000

The main features of the AP3000 are as follows:

1) The latest UltraSPARC workstation models can be used. Each node can support Fast/Wide SCSI, Ultra SCSI, 10BASE-T, 100Base-TX, FDDI, ATM, Fibre Channel, and other external interfaces.
2) AP-Net enables high-speed, high-throughput, and low-latency communications using hardware message handling features and wormhole-routing. These features provide not only a high peak-performance hardware but also high effective performance for real application software.
3) CPUs, memory, and I/O channels are scalably expandable from the basic four-node configuration to the maximum configuration of 1,024 nodes.
4) The operator can centrally control the power of all nodes, install nodes, and monitor the operating status. Also, the system can be operated automatically according to schedules.

4. Communication architecture

This section describes the communication architecture of the AP3000 (see Fig. 2). The MSC card (message controller) is connected to each node through SBus (I/O bus) for connection with AP-Net. The MSC card consists of a message controller LSI (MSC) and a buffer memory. The MSC includes DMA controllers for data transfer with AP-Net.

4.1 AP-Net

AP-Net supports a two-dimensional torus topology and consists of routing controller LSIs (RTCs) that route messages. AP-Net features the following:

1) High data-transmission rate of 200 Mbytes/s through a port

AP-Net transfers 16-bit data in parallel for a transmission rate of 200 Mbytes/s. The routing method is static routing; messages are first routed in the direction of the X axis and then in the direction of the Y axis.

2) Wormhole-routing

Wormhole-routing[3] divides data (messages) to be transferred into small pieces called flits, each of which consists of several bytes. The routing nodes transfer messages in flit form. Flits in the header determine the routing path route of the message, and the subsequent data is routed over the same path.

3) Dual virtual communication channel

The AP-Net hardware supports virtual communication paths called channels so that data can be independently transferred between nodes using dual channels. One of the dual channels is used for IP communication by the system, and the other is used by the user for parallel processing application software. Each channel has different logical communication paths to prevent deadlocking and to handle request messages and response messages separately. Furthermore, each path is duplicated to avoid deadlocking in the torus topology. Thus, there are eight logical communication paths on a physical communication path.

4) Internode barrier synchronization

In a parallel processing system, although the nodes operate independently, the entire system must keep step by achieving barrier synchronization. The AP3000 achieves internode barrier synchronization by distributing and collecting synchronization messages on the network. The RTC hardware distributes and collects synchronization messages.

5) Reliability, availability, and serviceability

Table 2. Node specifications

Node type	U 140	U 170	U 200
Number of CPUs	1	1	1~2
Element	UltraSPARC (64-bit, 143 MHz)	UltraSPARC (64-bit, 167 MHz)	UltraSPARC (64-bit, 200 MHz)
Cache memory	Internal : 32 Kbytes External : 512 Kbytes	Internal : 32 Kbytes External : 512 Kbytes	Internal : 32 Kbytes/CPU External : 1 Mbytes/CPU
Memory	32 Mbytes to 1 Gbyte	64 Mbytes to 1 Gbyte	64 Mbytes to 2 Gbytes
Built-in hard drive	2.1 Gbytes to 4.2 Gbytes	2.1 Gbytes to 4.2 Gbytes	4.2 Gbytes
Number of SBus slots usable	2	2	3

To quickly alert the user to the occurrence of errors in the network, the monitoring processor in SYSCNTL looks for errors in RTCs. When an error occurs, information about the error is immediately posted to the control workstation. RTCs check to which nodes messages are posted. If an attempt is made to send a message to a node outside the initially defined groups or an incorrect message is received from an external source, the RTC returns an error.

4.2 Communication interface

The MSC is hardware used for inter-node communication. It has a communication controller with two channels for the system and two channels for the user. In addition to the conventional SEND and RECV, which transfer messages via send/receive buffers, the MSC also supports direct access to remote node memory, PUT/GET[4], and CSI (compare and swap instruction) and FOP (fetch and operation) functions.

SEND transmits data in local memory to the specified node. The transmitted data is written in the message-reception buffer in the receiver. The buffer is controlled by the hardware.

PUT copies data contained in the local node (hereafter referred to as local memory) into the specified node memory (hereafter referred to as remote memory). GET copies data contained in remote memory into local memory.

Figure 4 shows how PUT and GET handle data. PUT transfers to AP-Net the specified amount of data in node A starting from address src_addr. Node B writes the data posted from node A to the area beginning at the specified address dst_addr. The MSC in node A updates (increments) the src_flag after the data is transferred. The CPU monitors the src_flag value to find out when data transfer is completed.

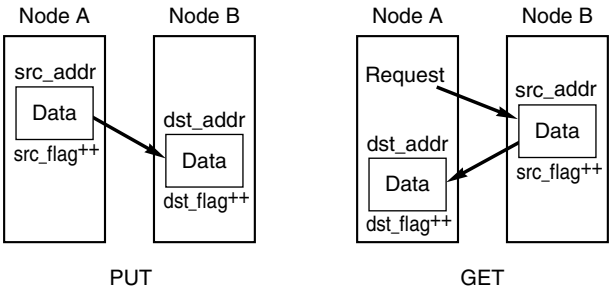

Fig. 4— Communication using PUT / GET operation.

GET is the counter operation of PUT. Node A specifies address src_addr, which is the starting address of the required data in node B. Then node A posts GET. After receiving GET, the MSC in node B posts the required data to node A. Node A receives the data transferred from node B. Flags src_flag and dst_flag are updated to indicate the end of data transfer.

PUT/GET provides effective communication when the data to be transferred between nodes is determined in advance; this is because, unlike the case for SEND/RECV, there is no need to copy data in the receiver.

CSI and FOP are used for exclusive access to remote memory. These features can be used for exclusive control of a database.

4.3 User interface

The MSC has a feature for queuing data transmission instructions such as PUT, GET, and SEND. This feature enables data transmission requests to be separately processed so that calculation and communication can be overlapped.

The dual-channel communication controller in an MSC is controlled by a system communication device driver and user-level communication device driver. The system communication device driver is installed in the Solaris OS stream driver to enable IP communication. User-level communication is implemented through a communication library used for direct access to the MSC communication hardware.

Users can handle high-speed communication using standard message-passing libraries such as MPI and PVM.

5. Conclusion

This paper outlined the parallel server AP3000 series. The features of the AP3000 can be summarized as follows:

1) The AP3000 is a scalable parallel server with distributed memory.
2) It has the latest 64-bit UltraSPARC architecture.
3) It uses the high-speed network AP-Net for communication between nodes.
4) It has a supervisory feature to handle multiple nodes as a single system.

The AP3000 system can be flexibly expanded according to the amount of data to be processed by using the architecture that connects the workstations to the high-speed network. AP3000 users can use new workstations immediately after they have been released. Therefore, the users can use a latest and higher-performance system.

Although AP-Net has good communication performance, the effective communication performance is limited by the performance of SBus, which is the I/O bus to which the MSC is connected. A better interface will be supported in the future to improve performance. Furthermore, to enable various modifications and to meet various demands of users, for example, the installation of various nodes, the AP-Net currently used in the AP3000 communication network will be connected to large servers connected next to the AP3000.

References

1) T. Shimizu, T. Horie, and H. Ishihata: Low-latency message communication support for the AP1000. The 19th Annual International Symposium on Computer Architecture, pp.288-297, 1992.
2) O. Shiraki et al.: AP-Net: Advanced high-performance network for scalable parallel server. Hot Interconnects IV, 1996
3) W. J. Dally and C. L. Seitz: Deadlock-free message routing in multiprocessor interconnection networks. IEEE Transactions on Computers, 36, t, pp.547-553 (1987).
4) K. Hayashi et al.: AP1000+ : Architectural support of put/get interface for parallelizing compiler. Architectural Support for Programming Languages and Operating Systems (ASPLOS VI), ACM, 1994.

Hiroaki Ishihata received the B.S. and Dr degrees in Electrical Engineering from Waseda University, Tokyo, Japan in 1980 and 1996, respectively. He joined Fujitsu Laboratories Ltd., Kawasaki, and Fujitsu Ltd., Kawasaki in 1980 and 1994, respectively. He has been engaged in research and development of parallel computer architecture. He received the Motooka Award in 1992 from the Motooka Memorial Association and the IEICE outstanding paper award in 1993 from the Information and Communication Engineers (IEICE) of Japan.

Hiroyuki Sato received the B.S. and M.S. degrees in Electrical Engineering from Waseda University, Tokyo, Japan in 1975 and 1977, respectively. He joined Fujitsu Laboratories, Ltd., and Fujitsu, Ltd. in 1977 and 1994, respectively. He has been engaged in research and development of hardware and software for parallel computers. He is a member of the IEEE and the Information Processing Society of Japan.

Masanori Takahashi received the B.E. degree in Telecommunication Engineering from the University of Tokai, Kanagawa, Japan, in 1976. He joined Fujitsu Ltd., Kawasaki in 1976 and has been engaged in development of large-scale general purpose computers and supercomputers.