UDC 681.3.06

Operating System of AP3000 Series Scalar-Type Parallel Servers

●Hiroyuki Oyake ●Yuji Iguchi ●Tsunemi Yamane

(Manuscript received May 6, 1997)

This paper outlines the control software for the AP3000 series of scalar-type parallel servers. The series provides high-performance computing power in many fields, for example, R&D, database processing, decision making support, and multimedia processing.

Each AP3000 series machine is a scalar-type parallel computer system consisting of four or more node computers that are interconnected by a high-speed communication network called AP-Net. The AP3000 series has a high processing performance to cover the higher-level models of equipment ranging from Symmetrical Multiprocessors (SMPs) to Massively Parallel Processors (MPPs). The AP3000 series contains an SMP as a node computer and has the same scalability as that of an MPP. The series has the characteristics of a cluster-type computer and has all the characteristics of typical parallel computers.

Solaris is used as the control software of the AP3000 series. By combining Solaris with operation management software products, all node computers connected to AP-Net can be operated as one system. The set of node computers can also be divided into various groups as required.

1. Introduction

A computer having multiple processors is generically called a parallel computer. Parallel computers are classified into three types: SMP[Note1], MPP[Note2], and cluster-type computers. An SMP computer has several tens of processors and a shared memory architecture. An MPP computer has several hundreds of processors and a distributed memory architecture. A cluster-type computer has several computers mutually connected by a network so that the computers can be used as a single system. Processors used in parallel computers are called processor elements (PEs) or node computers. In this paper, such processors are referred to as nodes.

Fujitsu created the MPP computer AP1000 in 1990. The AP1000, in which a specific Cell OS was adopted, was a back-end type computer that used a workstation made by SUN Microsystems as the front-end machine. From the experience gained in creating the AP1000, Fujitsu concluded that the fundamental requirements for MPP computers are hardware that enables high-speed, low-latency, and high-throughput communication between many processors, and software that makes the best use of the hardware performance.

Before creating the operating system for the AP3000 series, the successor of the AP1000, it was decided that no operating systems specific to MPP computers would be created, and instead a generic operating system with an add-on communication driver would be used. This decision was made because when an MPI[Note3] or PVM[Note4] provides high transmission rates, there are no differences between a specific OS and a generic OS for application software, which is the only beneficiary of

[Note1] Abbreviation of symmetrical multi-processor.
[Note2] Abbreviation of massively parallel processor.
[Note3] Abbreviation of message-passing interface.
[Note4] Abbreviation of parallel virtual machine. Public domain software developed by the Oak Ridge National Laboratory (USA).

parallel processing.

To help satisfy the demand for high-performance computing, the AP3000 series supports an environment in which multiple users can execute parallel processing programs at the same time. Also, to provide the high throughput required to construct a computer center system, the AP3000 series provides distributed features for nodes and enables reinforced batch processing.

In this paper, we introduce technology to exploit the computational capabilities of the AP3000 series and to operate the system.

2. Purposes of the AP3000

In the field of high-performance computation, the demand for scalar-type parallel computers as well as conventional vector-type computers is increasing. The AP3000 series are scalar-type parallel computers that support high-grade models of the SMP and MPP. The AP3000 was developed to include all the features of the SMP, MPP, and cluster-type computers.

2.1 SMP

The computational capabilities of SMPs have been significantly increased by improved processing abilities and a newly developed architecture for basic processors that enables connection of up to about 30 CPUs.

The AP3000 can connect Sun Microsystem's SMP models as nodes. Dual-CPU models such as the Ultra Enterprise[Note5)] 2 Model 2200 can be installed in a cabinet without any modification. Node computers with many CPUs such as the Ultra Enterprise 6000 will be externally connected in the future.

Although the type of nodes can be flexibly selected, there is a trade-off between the capability of a node and the size of the area affected by a failure. To localize a failure, the number of processors per node must be reduced by increasing the number of nodes.

2.2 MPP

Venture companies have looked for new markets for MPP, but the markets are not as extensive as they had expected and their investment returns are probably quite poor.

Some have blamed the poor diffusion on the small number of application programs available, while others believe it is because parallel programs are difficult to create and use. To solve these problems, the AP3000 adopts a generic OS.

In the AP3000, the nodes are connected as clusters only as viewed from inside the system. However, from the application software view, the AP3000 does exhibit some of the features of cluster-type and MPP computers.

2.3 Cluster

Although most cluster-type computers consist of multiple computers, to increase the computational capabilities there are some cluster-type computers that consist of hundreds of computers, for example, the IBM SP2. The AP3000 can similarly connect a number of nodes as clusters as well. SMPs can also be connected as clusters to construct a flexible and high-speed computational server.

3. Features of the System

3.1 System configuration

The scalar parallel server AP3000 series uses a 64-bit-microprocessor UltraSparc[Note6)] and Sun Microsystems workstations as node computers. The AP3000 can connect up to 1,024 nodes through the high-speed network AP-Net. A control workstation is externally connected to AP3000 nodes by using a control network. A system control mechanism directly connected with the control workstation manages the power and provides an integrated console. **Figure 1** shows the AP3000 system structure.

Note5) A registered trademark of Sun Microsystems, Inc. (USA).

Note6) A registered trademark of SPARC International, Inc.

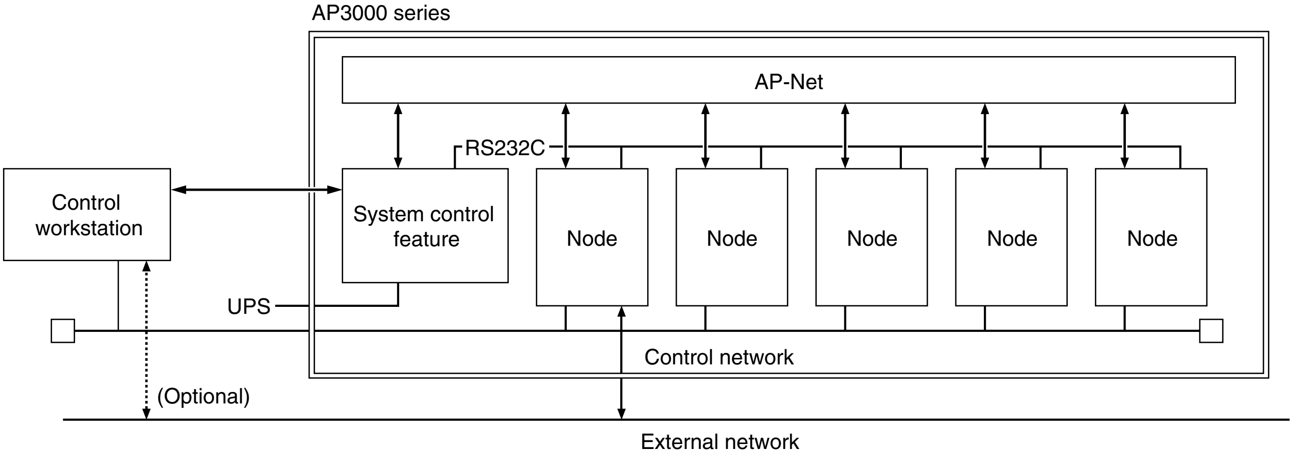


Fig.1— AP3000 system structure.

3.2 Generic OS

In the AP3000, the Solaris operating system[Note7] functions in each node. The operating system layer is a distributed processing system. To control the nodes efficiently, distributed control softwares that are commonly used on Solaris are also available on the AP3000. The administrator regards the AP3000 as a single computer, while users regard it as a system of multiple computers. This double characteristic is important because it enables popular application software to be used without modification.

3.3 High-speed parallel computation

The AP3000 supports two parallel computational modes to enable the user to choose between high-performance and multiplexed computing. The two modes are the SIMPLEX mode and SHARE mode.

In SIMPLEX mode, a single parallel-processing application program exclusively uses multiple nodes and communication paths to reduce the turn-around time. In SHARE mode, programs can be created and executed by multiple users using TCP/IP, but at a lower data transmission rate. SHARE mode is provided because, when developing or debugging parallel-processing programs, it is efficient and economical to run a highly-multiplexed program with only a few nodes or to execute multiple programs simultaneously.

The appropriate mode can be selected when programs are executed, and programs need not be modified.

3.4 Network address

The AP3000 not only enables nodes to be handled collectively as a MPP but also as individual computers like computers in a cluster system. Therefore, IP addresses are assigned to each node. A node has IP addresses for the control network and AP-Net. For practical operations, either the control network or AP-Net must provide routing information to external networks to communicate with them.

3.5 High availability

The AP3000 has a mechanism that makes it possible to continue operation without stopping the entire system when a hardware failure occurs.

The AP3000 provides conversational and batch services in multiple nodes. If a node drops out due to a failure, the remaining nodes continue service. In server systems similar to the AP3000, abnormal termination of an externally connected

Note7) A registered trademark of Sun Microsystems, Inc. (USA).

node causes the service to end. To avoid this, externally connected nodes must be multiplexed. When an externally connected node becomes faulty, another node must inherit the address and name made available to the outside; therefore, the system must be designed carefully.

A faulty node can be replaced without stopping the entire AP3000 system. A faulty node can be reinstalled into the system in the same way that a workstation is connected to a LAN.

For AP-Net failures, a fallback feature is provided so that the system can operate with the control network only. However, in such a case, the system must be reconfigured to reduce the system load. To replace a faulty part, the entire system must be stopped.

Automatic shutdown coordinated with a UPS[Note8)] is available for protection against failures of the commercial power supply.

4. System Software Configuration

The operating system layer of the AP3000 system software was designed to minimize modifications. Specifically, the AP-Net driver is installed in Solaris 2.5.1 and other programs are implemented in the application layer. **Figure 2** shows the structure of the AP3000 system control software.

4.1 AP-Net driver

The topology of AP-Net is the two-dimensional torus used for the AP1000 architecture.

AP-Net supports two kinds of virtual communication channels; one is the AP-Net direct communication driver used for communication between parallel processing application programs, and the other is the AP-Net stream driver used for IP. The AP-Net direct communication drivers provide high-speed communication for PVM and MPI (**Fig. 3**).

The AP-Net stream driver enables application software that uses TCP or UDP to handle AP-Net as well as general network devices. The high-speed communication made possible using the AP-Net stream driver can accelerate NFS and backup through the network without the need to modify applications. Thus, AP-Net allows more computer connections than are allowed in a LAN device.

4.2 System Controller

The AP3000 is equipped with a power controller called the system control mechanism and hardware for controlling consoles collectively.

The system controller operating on the external control workstation connected to the AP3000 is used to control the AP3000 power, activate/deactivate the nodes, and operate the node consoles through the system control mechanism. The operator can handle each node as a workstation from the control workstation while the operating system is running on each node or is in firmware mode.

The system controller helps the user to reduce the number of nodes that need to be operated in the nighttime and to isolate faulty nodes. The system controller can be linked with external devices such as the UPS. By using a link, the AP3000 can be automatically operated in computer centers.

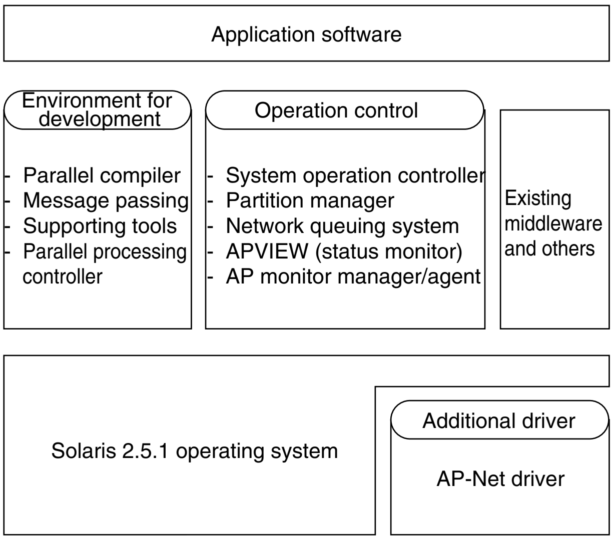

Application software
Environment for development
- Parallel compiler
- Message passing
- Supporting tools
- Parallel processing controller
Operation control
- System operation controller
- Partition manager
- Network queuing system
- APVIEW (status monitor)
- AP monitor manager/agent
Existing middleware and others
Solaris 2.5.1 operating system
Additional driver
AP-Net driver

Fig.2— System control software for AP3000.

[Note8)] Abbreviation for uninterruptible power source

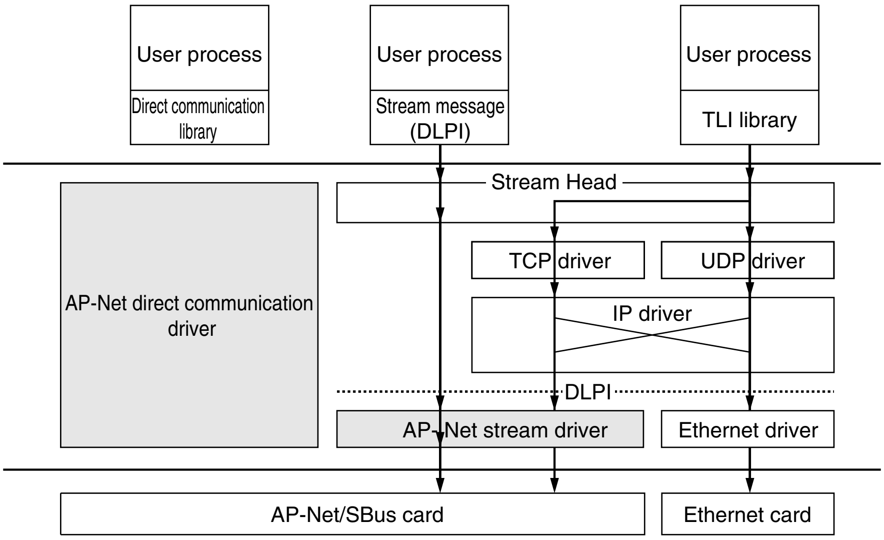


Fig.3— Structure of AP-Net device driver.

4.3 Partition manager

In some cases, regarding each node of the AP3000 as a mere group of computers does not lead to effective use of the system. Some examples of such cases are 1) when multiple users (e.g., laboratories) collectively purchase a computer system and partition the nodes and 2) when a computer center purchases a computational server for use as a throughput server during the day and for studying parallel processing during the night.

The AP3000 supports two types of frameworks to operate many nodes efficiently. One is to support use as in case 1) above by dividing the operational environment, including features such as the file server and user registration feature, into different areas. A cluster using this type of framework is called a PM cluster. The other framework is to support use as in case 2). This framework provides an operational environment that changes its configuration over time. This type of framework is called a partition. A PM cluster includes partitions.

The partition manager supports the operation of nodes under the two-layered frameworks. It centralizes the system environment settings and error recovery within the group.

For partitions, processing is divided into conversational processing and batch processing. Conversational processing partitions support execution of remote commands and remote login for low-loaded nodes. Batch processing partitions automatically distribute the load using the network queuing system to implement efficient operation.

4.4 Network queuing system

The network queuing system in the AP3000 adds the features described below to the standard UNIX batch processing system. The network queuing system improves the throughput and turn-around time.

4.4.1 Assured execution of parallel processed jobs

Multiple nodes are preliminarily assigned as execution nodes to each job queue. When a job is executed, the required nodes are automatically allocated from the preliminarily assigned nodes. Faulty nodes are not allocated. If one of the nodes becomes faulty while the job is being executed, unless otherwise specified, another node is scheduled to take its place. If there is a usable node, the node is immediately allocated to the job. Execution of jobs is assured even if a node fails.

4.4.2 Exclusively used nodes

When SIMPLEX mode is defined for a job queue, the nodes are used exclusively while the job is being executed. Therefore, in SIMPLEX mode the user does not need to worry about delays in execution or a lack of resources due to other jobs.

4.4.3 Nodes fixed to specific job queues

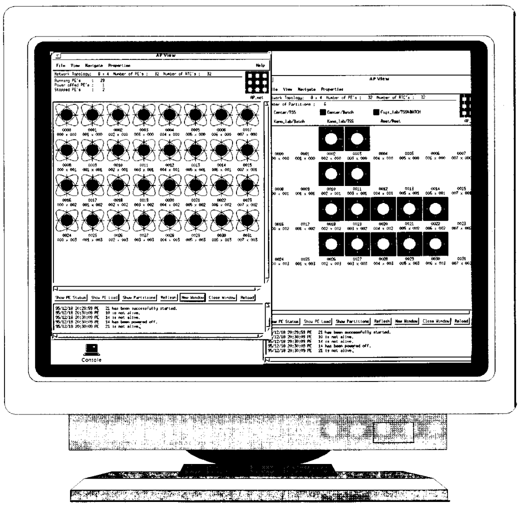
Specified nodes can be assigned to specific job queues permanently. This feature is useful for applying nodes for specific application servers.

4.4.4 Load distribution

Before execution of jobs, the least loaded of the nodes defined for a job queue is executed first to improve the throughput.

4.5 APVIEW

APVIEW is a monitoring tool based on Motif [Note9)] and is designed for the AP3000 hardware. APVIEW displays the status of power to each node, the status of power to AP-Net, the network topology, the load status (10 levels), and information about the nodes (memory capacity, number of logged-in users, and locations in the cabinet). APVIEW links with the partition manager to display how nodes are partitioned. **Figure 4** shows an example of the information displayed by APVIEW.



```
Network Topology:   8 x 4  Number of PE's :   32  Number of RTC's :   32
Running PE's     :   29
Power offed PE's :    1
Stopped PE's     :    2
```

```
95/12/18 20:29:59 PE   21 has been successfully started.
95/12/18 20:30:09 PE   10 is not alive.
95/12/18 20:30:09 PE   14 is not alive.
95/12/18 20:30:09 PE   14 has been powered off.
95/12/18 20:30:09 PE   21 is not alive.
```

```
Number of Partitions :   6
Center/TSS        Center/Batch      Fuji_lab/TSS&BATCH
Kano_lab/Batch    Kano_lab/TSS      Root/Root
```

Fig.4— Example of information displayd by APVIEW.

4.6 AP monitoring manager/agent

The AP monitoring manager/agent is an integrated console feature designed specifically for administrators to use with the AP3000. It combines two of Fujitsu's middleware products: centralized monitoring manager and Open-Eyes. Also, it contains customized data for the AP3000, so it can be used immediately after installation.

The AP monitoring manager/agent can be linked with other middleware products. For example, an operation management system can be created by linking the AP monitoring manager/agent with network management software "NetWalker" or job schedule management software "job scheduler."

4.7 Practical example

Figure 5 shows an example of an AP3000 system installed for a department and the computer center of a university.

In this example, the nodes are grouped into a PM cluster for a department and a PM cluster for the computer center, and each cluster is separated into a conversational processing partition and a batch processing partition. Grouping nodes into PM clusters enables the PM cluster to manage users. Separating clusters into partitions enables dynamic switching between nodes for conversational processing and nodes for batch processing so that the system mainly handles conversational processing during the day and batch processing at night. Conversational processing partitions can be connected to multiple external networks to distribute the load.

Although the nodes for file servers do not belong to a PM cluster, the nodes can be integrated into a PM cluster by unifying their user management settings. When file servers are being located in the AP3000, AP-Net improves the performance of the NFS.

Note9) A registered trademark of Open Software Foundation, INC.

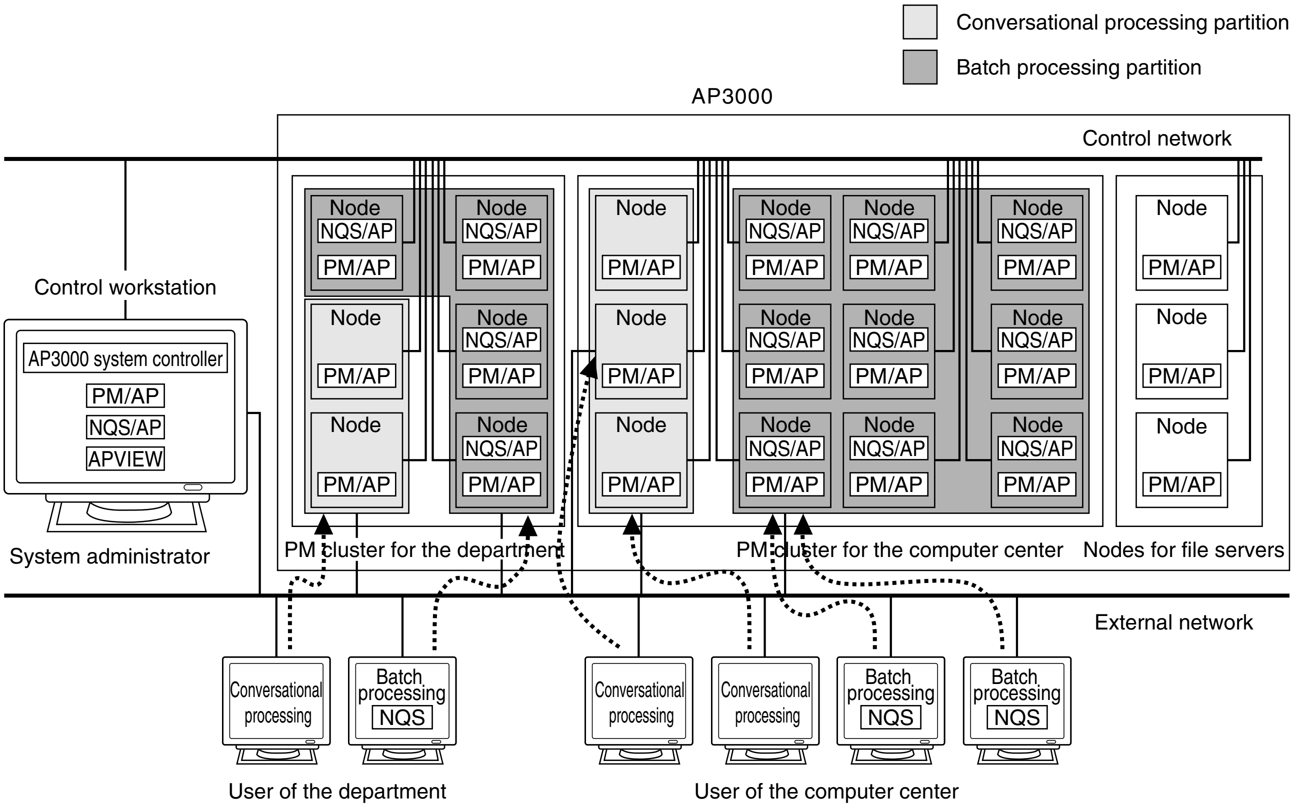


Fig.5— Example of system structure for computer center of university.

5. Developments in the Future

The AP3000, with MPPs and general-purpose cluster computers unified and common techniques and parts integrated, is only the beginning result of trials for areas that require very high computational powers. In the future, the system will be refined by providing the following.

5.1 Support of faster communication

It is important for parallel computers to support high-speed communication between nodes. To make the best use of AP-Net, support of an interface faster than SBus is urgently needed.

5.2 Reduction of network addresses

IP addresses are assigned to all nodes for both parallel application software and distributed application software. However, a new framework must be created to reduce the amount of IP resources required.

5.3 Support of high-speed file access

Although parallel file systems are effective for high-speed file access, there is a problem with compatibility with standard file systems. A faster file access system that is compatible with standard file systems must be developed.

5.4 High reliability and availability

To improve the overall reliability of the system, features that provide a high reliability must be developed. Most platforms already provide, or are scheduled to provide, high-reliability features that use cluster configurations. However, the AP3000 must be especially reliable because of its large size.

5.5 High-level operational design support system

The AP3000 consists of multiple computers, so we can expect more closely integrated features for constructing and operating the system to be developed. To assist in this development, new operation design and operation support tools must be created.

5.6 Distributed computing using different types of computers

To make use of Fujitsu's considerable skills and experience, the AP3000 is based on Sun Microsystems workstations. Also, to satisfy various requirements for systems in the future, our techniques will need to be improved to create heterogeneous systems consisting of other equipment, for example, PC servers.

6. Conclusion

By manufacturing the AP3000, we proved that a large computational server can be created by combining components available in the market. Thanks to the availability of low-priced, high-performance microprocessors, we can continue to provide comparatively low-priced, high-performance computers. Also, we are planning to reinforce the functions for large-sized servers by applying the computational capabilities of the AP3000 to new devices such as WWW servers and OLAP servers.

Hiroyuki Oyake received the Associate degree in Electrical Engineering from Fukushima National College of Technology, Fukushima, Japan in 1979.
He joined Fujitsu Ltd., Kawasaki in 1979, where he has been engaged in research and development of online DB/DC subsystems, fault-tolerant operating systems, and UNIX operating systems for mainframes, supercomputers, and parallel computers.

Tsunemi Yamane received the B.S. degree in Chemistry from Kyoto University, Kyoto, Japan in 1972.
He joined Fujitsu Ltd., Kawasaki in 1973, where he has been engaged in research and development of operating systems for mainframe computers and parallel computers.

Yuji Iguchi graduated from Yoshiwara Technical High School, Shizuoka, Japan in 1981.
He joined Fujitsu Ltd., Kawasaki in 1981, where he has been engaged in research and development of UNIX operating systems for mainframes, minicomputers, supercomputers, and parallel computers.