High Performance File System for Supercomputing Environment

H.Nishino, S.Naka, K.Ikumi
NEC Corporation
W.Leslie
HNSX Supercomputers, INC.

Abstract:

SUPER-UX is a high performance UNIX operating system for NEC SX-3/SX-X supercomputer and has many enhanced facilities to make available the maximum SX-3/SX-X computing power (peek speed of 22GFLOPS) to users. This paper presents the SUPER-UX file system, a strong point of SUPER-UX, which is based on the standard UNIX file system to satisfy scientific computing environment requirements for I/O performance and facilities.

1. Introduction

UNIX is the only standard operating system which supports full range of computers from personal computer to supercomputer and is widely used in various kinds of field, including R&D, OA and FA. Furthermore, customer demand maker it is necessary to support UNIX as a supercomputer OS. However, there are many challenges in adapting environment the original UNIX design and facilities to the supercomputer environment as follows:

- tightly coupled multiprocessor environment,
- high speed I/O,
- batch processing facility,
- enhanced operation management facilities,
- enhanced security facility, etc

SUPER-UX is based on AT&T UNIX SystemV R3.1 and includes many useful facilities from 4.3BSD (network facilities, utilities etc.). Additionally, it is enhanced with facilities and performance to meet the challenges described above and deliver the maximum computing power and resources of SX-3/SX-X[1] to users.

This paper presents the overview of SX-3/SX-X hardware and SUPER-UX in section 2 and 3 respectively. Then describes the background, implementation method, virtual volume concept, SFS(Supercomputer File System) and high speed I/O techniques(IAS; Intelligent I/O Accelerator Subsystem) of the SUPER-UX file system in section 4. Several future directions of SUPER-UX are mentioned in section 5.

UNIX is a registered trademark of AT&T.

2.Overview of SX-3/SX-X Hardware

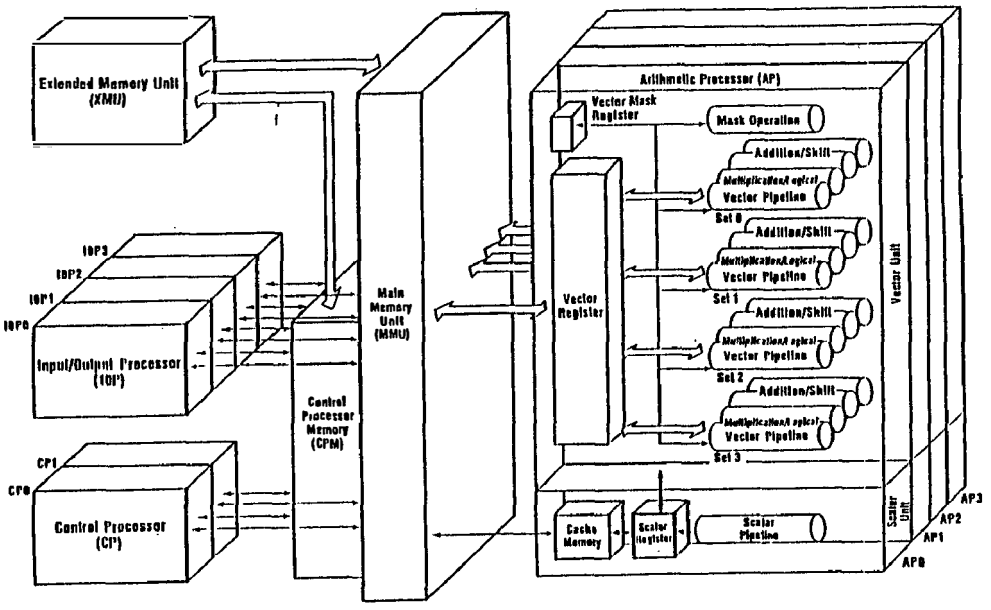

Figure 1. SX-3/SX-X Hardware System Configuration

Figure 1 shows SX-3/SX-X hardware system configuration. The arithmetic processor(AP) achieves the fast machine cycle time of 2.9nsec based on basic hardware technologies such as VLSI(20,000gates, with gate switching speed of 70 picoseconds), and high density packaging employing liquid cooling. Each AP can provide vector peek performance of 5.5GFOLPS when utilizes 16 vector pipelines operate simultaneously. There are 144KB vector registers to minimize temporary memory storage traffic. Additionally, a multiprocessor configuration which consists of 4APs (maximum configuration) can provide 22GFLOPS.

An SX-3/SX-X provides maximum of 2GB main memory unit(MMU) with 1024 way interleaved configuration. An extended memory unit(XMU) with 16GB capacity and 2.75GB/sec bandwidth contributes to the improvement of data supply capability to APs.

Data Control Processor(DCP) with maximum 256MB data control processor memory(DCPM) provides effective I/O data buffering, and data transfer control between peripheral storage equipments and XMU as backdoor staging and destaging in cooperation with the input/output processor(IOP).

3.Overview of SUPER-UX

Figure 2 illustrates an ideal supercomputer UNIX. Figure 3 illustrates the first version of SUPER-UX, derived from figure 2 through synthetic analysis and judgment of urgency, technical importance, and timeliness. The major enhanced facilities of SUPER-UX are as follows.

- It was necessary to implement 64-bit UNIX to support full memory access for future models with greater than 32-bit address space and the performance of SX-3/SX-X which is a 64-bit work machine. Modification for 64-bit have been achieved throughout the kernel and program linkage convention has also been expanded.
- The memory management facility of SUPER-UX was rewritten from virtual memory management algorithm of SystemV to real memory management. SX-3/SX-X hardware does not provide the mechanisms for virtual memory control.
- Some other enhancements (multiprocessor support, batch scheduling, system freeze/restart etc.) were embedded in the Kernel.
- To provide SX-3/SX-X as a compute server in distributed environment, high speed LANs (Ethernet and HYPERchannel-DX), TCP/IP and many network utilities, NQS(Network Queuing System), and NFS(Network File System) are supported.
- Vectorizing and FORTRAN and C compilers are supported to provide maximum speed of SX-3/SX-X vector and scaler capabilities to users.

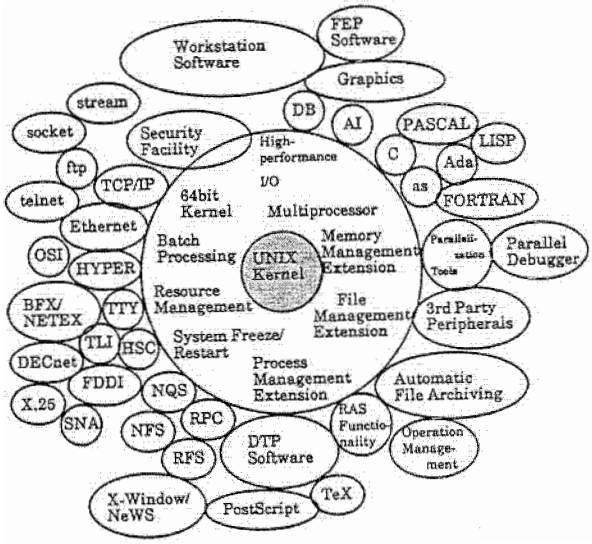

Figure2. Structure of Supercomputer UNIX

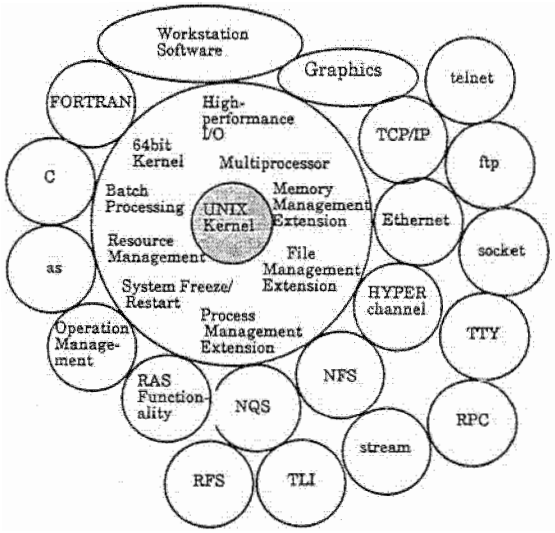

Figure3. Overview of SUPER-UX

4.SUPER-UX FILE SYSTEM

This section describes the SUPER-UX file system which is enhanced beyond the standard UNIX file system. The purposes of these enhancements are to satisfy I/O performance and functional requirements for a supercomputer OS.

4.1 Development Background

At the beginning of design phase, an I/O analysis of many scientific computing applications (mainly FORTRAN jobs) running on supercomputer was performed. The following items were identified as supercomputer I/O characteristics.

① The number of files which are used depends on each application. Dozens of files are used normally and some applications use more than 60 files.

② The size of files also varies with each application. There are files from a few KB to a few GB. Data storage of dozens GB to a few TB file size need to be supported in the near future.

③ I/O operation size varied from a few B to hundreds MB.

④ The I/O frequency is inconsistent. A program's execution divides into two parts. One is the I/O-bound part, with I/O operations requested continuously, and the other part is the CPU-bound part in which computation is proceeded.

⑤ The file may be accessed both sequentially and randomly.

⑥ Some application areas, processes large numbers of unreliable and unknown format (seismic processing is an example).

Thus, the necessary conditions which a SUPER-UX file system should satisfy are as follows.

① Giant files can be created.(lagger than a device)

② Data transfer between CPU and peripheral storage equipments, which should achieve approximately peak hardware performance.

③ Easy dynamic file creation and deletion.

④ Automatic and unlimited expansion.

⑤ Both sequential and direct access on the same file.

⑥ Flexible magnetic tape access.

⑦ Transparent file migration from other computer systems.

⑧ Linking special file formats to peripheral storage equipment type should be avoided.

4.2 Implementation Method

This section explains the implementation method of SUPER-UX file management facility which satisfy conditions described in section 4.1.

The conditions ③,④ and ⑤ in 4.1 are advantages that current UNIX file management presents, and these contribute to flexible and clear file system construction. Furthermore, ⑥ and ⑧ can be satisfied based on the raw I/O facility provided as a standard UNIX file management facility.

Consequently, to inherit these advantages in SUPER-UX, enhanced facilities for high speed I/O is implemented under the UNIX standard file management facility transparently and separately.

Standard UNIX cannot satisfy the condition ①. Therefore, the new file management facility specific for supercomputing environment is developed to satisfy ① (giant file creation requirement). This new file management facility is called SFS (Supercomputer File System) and located at the same functional layer of S5FS, the standard UNIX file management facility. An I/O request to SFS or S5FS file is distinguished by FSS(File System Switch), the original facility of SystemV R3. The enhanced facilities for high speed I/O condition is integrated to the functional layer which is called IAS(Intelligent I/O Accelerator Subsystem). IAS is located at the lower layer to SFS and S5FS and shared by both file management facilities.

Figure 4 illustrates the implementation method of the SUPER-UX file system described in this section. To satisfy the condition ⑦ (requirement for transparent file migration between SUPER-UX and other systems), NFS(Network File System) is also supported.

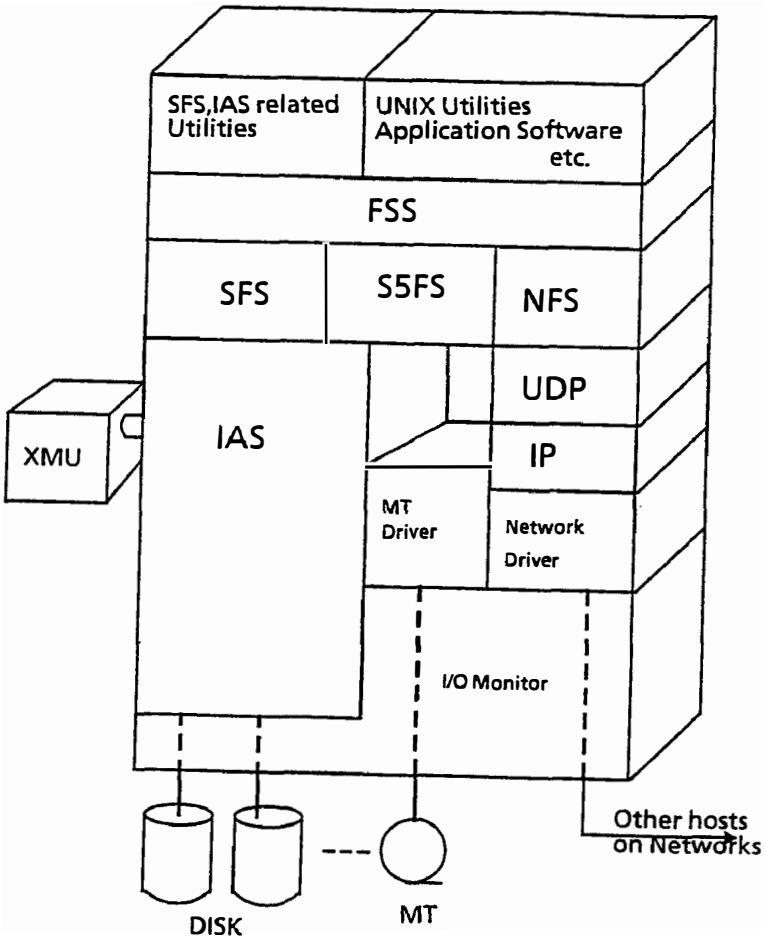

Figure4. Structure of SUPER-UX File System

4.3 Virtual Volume Concept

To integrate supercomputer file system performance and facilities, the virtual volume concept is introduced.

The virtual volume is a storage equipment which satisfies both high speed I/O and giant file system requirements and has the following characteristics.

- Virtual volume equipment may consist of XMUs and magnetic disks.
- The virtual volume equipment is configurable based on performance and reliability requirements from system administrator.
- The equipment is handled as a special file by all application programs and utilities, and this is the same view as other peripheral storage equipments (XMU, DISK, MT etc.).
- The equipment may be shared by SFS and S5FS.
- Various forms of virtual volume equipment structure can be introduced. One-to-one correspondence with a real equipment is the simplest example (equivalent to original UNIX file system configuration) and a virtual volume consists of more than one real equipments is another example (parallel disks or multivolume disks).

Figure 5 represents the concept of virtual volume equipment.

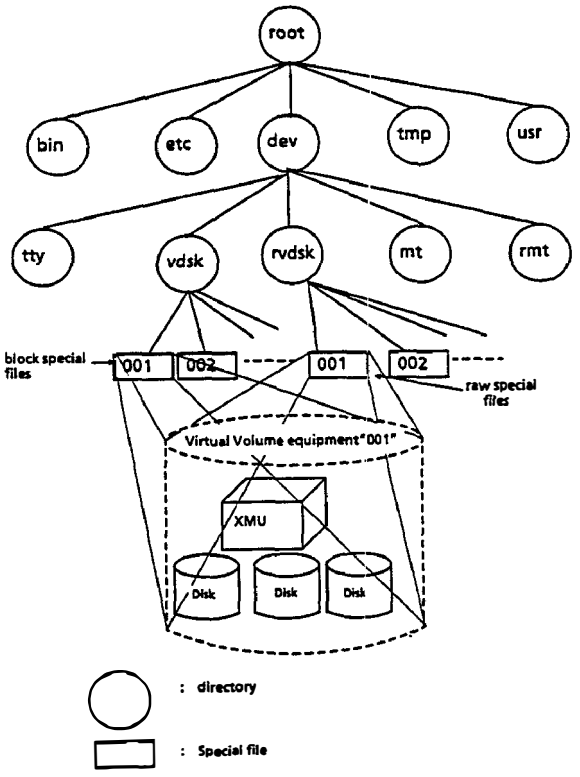

Figure5. Virtual Volume Equipment

4.4 SFS (Supercomputer File System)

SFS is an enhance file management facility based on S5FS to satisfy a giant file system requirement.

The characteristics of SFS are follows.

- In addition to "block", the original file allocation unit on peripheral storage equipments, a "cluster" is introduced as a new file allocation unit. The cluster length can be assigned by the system administrator. Accordingly, contiguous allocation of large volume of data can be achieved by assigning a large volume of data can be achieved by assigning a long size as the cluster length (In the case of disk, a cluster length could equal cylinder size).

- Furthermore, a "staging unit" is introduced as the new I/O unit to realize collective I/O of large volume of data on a cluster in addition to "block", the original I/O unit. Staging unit length can also be controlled by system administrator. Data length of collective I/O operation can vary from a staging unit length (as a minimum length) to a cluster length (as a maximum length). In the case of disk, for example, if a staging unit length is equivalent to a track size, each I/O operation can be performed as single or multi-track I/O.
- Large volume date transfers occur frequently on SFS files. Consequently, UNIX buffer cache can be bypassed, because whole capacity and caching unit of the buffer cache are very small and will be a bottleneck of data transfer.

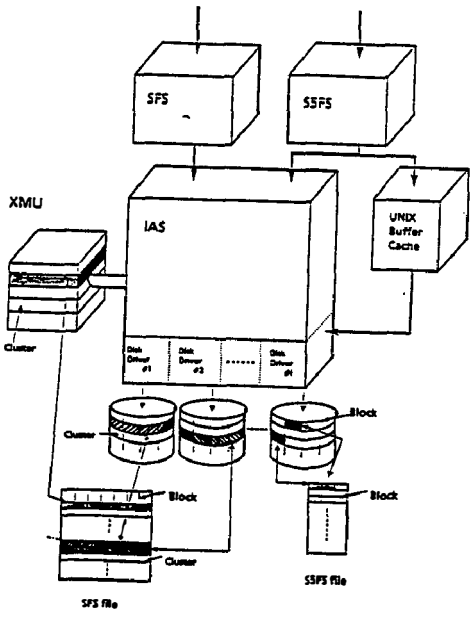

Figure6. Structure of SFS File

4.5 High Speed I/O techniques (IAS)

IAS(Intelligent I/O Accelerator Subsystem) has responsibility to provide virtual volume equipment for SFS and S5FS and contains the high speed I/O techniques.

This section mainly describes Virtual Volume Cache Control Facility (VVCCF) which is one of the high speed I/O techniques introduced on IAS. VVCCF can provide two-layer caching mechanism which consists of memory and XMU for disk-based virtual volume equipment. I/O requests for large data volume from SFS are cached by the staging unit. In addition to reducing disk I/O requests, VVCCF realizes a few hundred to a few thousand times faster I/O transfer speed than disk I/O when cache hit occurs.

The basic algorithm of cache control in figure 7 is as follows.

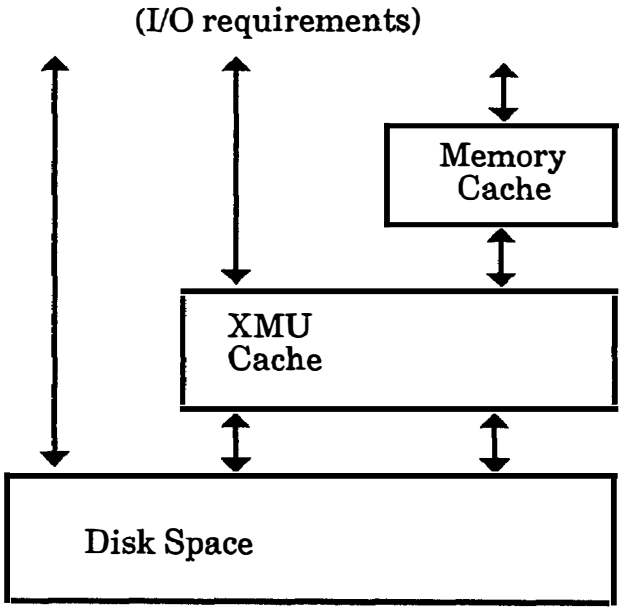

Figure7. Structure of Virtual Volume Cache

- Caching to the memory is done by write-through (the data written in the memory cache is written to the XMU cache or disk simultaneously).
- Caching to the XMU is done by write-back (Writing data to disk is delayed until the of that XMU cache area is required).

Consequently, data inconsistency between the memory cache and the XMU cache or disk does not occur.

File integrity can be maintained when the system crashes and hence it is possible to organize a reliable file system.

Figure 8 shows the position of VVCCF. VVCCF is the uppermost layer in IAS and is called by Virtual Volume Driver(VVD) which provide virtual volume equipment to SFS and S5FS. Other high speed I/O techniques are embedded in the lower layers of VVCCF. These techniques are removable and are able to respond to every sites' administrator requirements for performance and facilities because of the modular design.

Figure 9 represents the internal structure of VVCCF. VVCCF consists of Virtual Volume Cache Controller(VVCC) and Virtual Volume Cache Control Daemon(VVCCD).

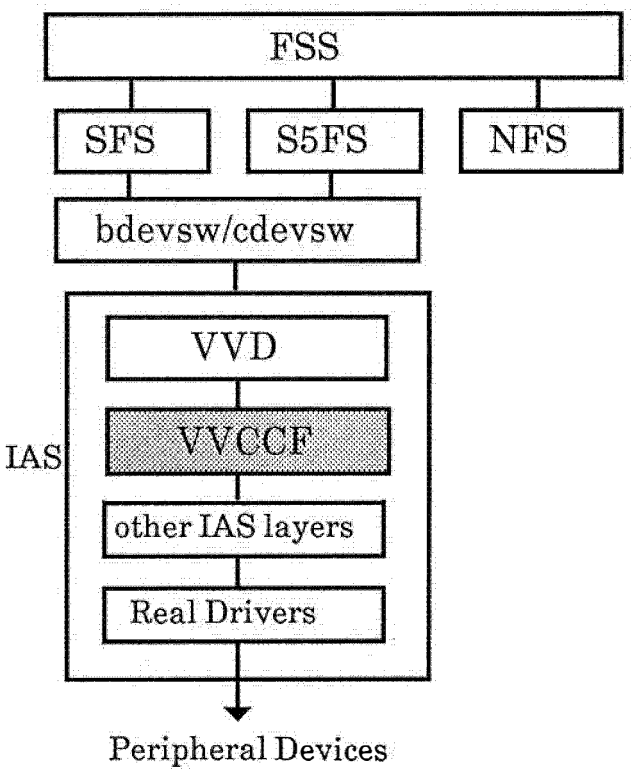

Figure 8. Location of Virtual Volume Cache Control Facility

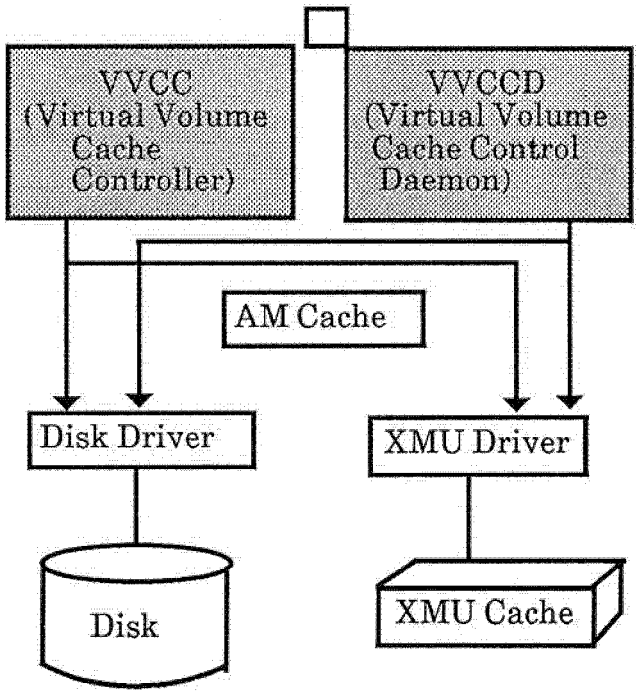

Figure 9. Internal Structure of Virtual Cache Control Facility

(1)VVCC

VVCC is called by VVD and carries out the following cache control algorithm by making use of the disk driver and the XMU driver. VVCC manages the caches in the following manner.

- Cache replacement is based on LRU(Least Recently Used) control. Each staging unit is managed by a prioritized multi-level freelist which is prioritized by expected cache hit ratio values, to control "hold time" for cached data (figure 10).
- Memory cache is bypassed to avoid decrease of cache hit ratio in the following I/O operations because of memory cache flush when the size of an I/O exceeds the cluster length.
- Selection of I/O paths (whether memory cache and XMU cache are used) is based on access patterns and I/O length, cache usage situation, and other criteria.

VVCC does not perform data migration between caches and disk directly to reduce the overhead of I/O processing. VVCC only enters migration requests in a special queue (called migration queue). Actual migrations are performed by the VVCCD.

(2)VVCCD

VVCCD is activated at a fixed time interval and manages data migrations between memory cache and XMU cache, and caches and disk, according to requests from the migration queue. VVCCD performs the following three migration processes.

- Write back operation from XMU cache to disk.
- Prefetch operation when read requests of contiguous stating units in a cluster are issued.
- Replace memory migration date from memory cache to XMU for cache data which has high hit ratio.

Other high speed I/O techniques embedded on the lower layers of VVCCF in figure 8 are as follows.

- parallel disk I/O

  This facility improves I/O performance a few times faster than I/O to a single disk. One file is divided and allocated on a few disk drives and IO requests for that file are divided with the pieces of the I/O request are executed simultaneously [2].
- Dynamic allocation within cluster

  For small files allocated by cluster unit on disk drives, internal fragmentation (low space utilization within allocation unit) increases. To satisfy both high speed I/O and effective space usage dynamic allocation within cluster is provided.

Data is transferred by cluster to XMU, where the small files forming a group can be quickly individually accessed at XMU transfer rate.

4.6 Other Facilities

To organize and maintain SFS file system, all of standard UNIX utilities related to file management (mkfs, fsck etc.) are available. Users can handle SFS file system the same as S5FS file system.

Additionally, the new utilities, related to IAS, are provided to create virtual volumes, and gather and edit various kinds of account information.

5.Future Directions

The following items are future directions of the SUPER-UX file system project.

(1) Evaluation and verification of high speed I/O/.
When the paper is written, VVCCF facility described in this paper is being evaluated by a simulator.

(2) Functional expansion of supercomputer OS.
Based on extended virtual volume concepts, it is possible to provide timely enhancements towards the ideal supercomputer UNIX illustrated in figure 2. For examples, archiving, parallel tape, etc.

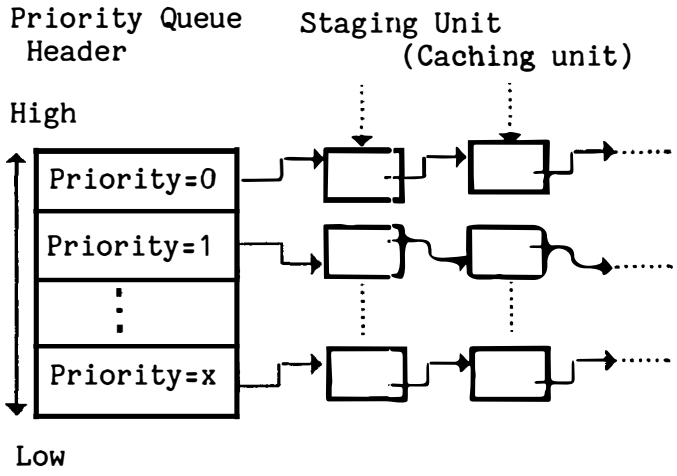

Figure 10. Free List of Virtual Volume Cache

References

[1] Watanabe,T., Matsumoto,H. Tannenbaum,p., : "Hardware Technology and Architecture of NEC Supercomputer SX-3/SX-X System", Proceedings of Supercomputing '89.

[2] Poston,A. : "A High Performance File System for UNIX," Proceedings of Workshop on UNIX and Supercomputers, 1989, pp.215-226.