Use of SONY PetaSite and DTF Helical Scan Tape Drive on KEK Central Computer UNIX Cluster

Y. Morita [a], S. Yashiro [a], J. Chiba [a], Y. Watase [a], Y. Kodama [b], S. Kohno [b], N. Tanaka [b],

[a] *KEK, National Laboratory for High Energy Physics, 1-1 Oho, Tsukuba, Ibaraki 305 Japan*

[b] *Government & Public Corporation Information Systems Division, HITACHI, Ltd., Computer Group, Tsukuba Mitsui bldg. 1-6-1, Takezono, Tsukuba, Ibaraki 305 Japan*

Storing a few tens to a few hundreds of tera bytes of HEP data under distributed UNIX environment is an essential and non-trivial function for HEP computing centers. Performance issues, such as data I/O throughput and data path parallelism for multi-user environment, become critical. We have developed a UNIX tape management system which provides two functions on a single tape library: (1) HSM migration, and (2) a direct tape access method with a file index database. Twenty SONY DTF helical scan tape drives are connected to several UNIX workstation servers at sustained I/O throughput speed of 12 MB/sec. The library has total storage capacity of 20 TB. The system is being used for several KEK Proton Synchrotron fixed target experiments since January 1996.

Key words: HSM; DTF; Tape Library

1 Introduction

Typical HEP data analysis model has two distinct data access patterns. HEP raw data claims the vast majority of the data storage space. The raw data is read only a few times by limited numbers of "production" processes. After the production, the resultant physics object data (DST: Data Summary Tape) amounts a few percentage of the data storage space. DST data have to be read and analyzed frequently by many collaborators world-wide.

To store and analyze the large amount of HEP raw data, large scale storage system with high speed I/O throughput between storage and CPU is required. We have been testing several types of helical scan tape drives for this category of data access[1].

For distributing the physics object data world-wide, we also have developed a hierarchical storage management (HSM) system with a file migration capability on DCE/DFS file servers[2].

2 KEKCC Tape Library

New KEK central computer system (KEKCC) has been introduced in January 1996. The system has two SONY PetaSite tape libraries. Each library is equipped with eight or twelve SONY DTF (Digital Tape Format) helical scan tape drives. Each tape can store up to 42 GB of data.

Two PetaSite libraries are dedicated to four workstation clusters of separate research groups (*"clusters"*). *PS Cluster*, a cluster for KEK Proton Synchrotron experiment groups, has a PetaSite library of 12 TB with twelve DTF drives. The same library is shared with a cluster of another research group, *JLC Cluster*. Four DTF drives of the PS cluster and two drives of the JLC cluster are used for the HSM migration server. Another eight drives of the PS Cluster are used directly from the user processes.

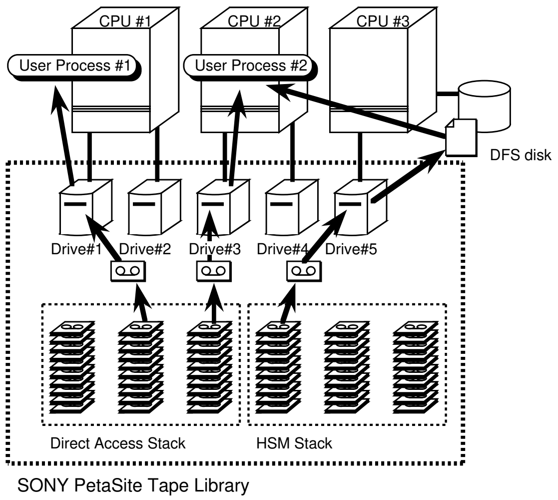

Fig. 1. Tape stacks in a single PetaSite tape library are divided into two logical groups: (1) direct access stack, and (2) HSM stack. Tapes in both stacks can be accessed from any CPU within the same cluster.

Another PetaSite system of 8 TB are connected to *Central Cluster* and *International Collaboration Cluster*. Each cluster has three DTF drives for HSM migration.

3 Storage Software

The storage management software for the KEK central computer system is a customized version of HP OmniStorage. The HSM migration function is built into the DCE/DFS file server. Detail of the HSM and DCE/DFS implementation is provided in a separate report[2].

We have also developed C and Fortran API functions to access tapes directly (Fig. 1). Each tape handling transaction is carried over TCP/IP sockets. A tape drive in the library can be connected to any workstation in the same cluster as the tape library server. Thus we have been able to build a parallel I/O system at a user process level.

4 Direct Tape Access API

Tape drives are valuable resources under multi-user environment. Drive allocation and mutual exclusion of the device file are essential requirements. We have built a mechanism to allocate a drive using the tape library server software. Only the user processes on the same workstation as the drive can be allocated, while tape library server keeps track of which drive is being used. When the tape handling request comes from a client, it controls the movement of the robot arm, load the tape into the allocated drive, then updates the file index database. The library system also checks whether the tape volume is mounted group-wise. If the drive is allocated and the tape volume is mounted in “shared” mode, any member in the same group in `/etc/group` file can access the tape volume, mount or umount the tape, and deallocate the drive.

Commands `mtalloc/mtdealloc` allocates and deallocates a tape drive. Once the drive is allocated for a given user UID, a user tape volume can be mounted or unmounted with `mtmount/mtumount` commands. C and Fortran API functions `mtopen/mtclose` opens and closes a file on the tape volume, and `mtwrite/ mtread` API functions writes and reads a data record to/from the tape volume. There are several other commands and API functions such as listing the user file catalog information or handling and reporting error status.

Tape format used in the direct access mode is designed to achieve media portability. Files on a tape volume are separated by a file mark, and all file catalog information are kept on a catalog database at the tape library server. Exported tapes can be read with a single DTF drive workstation without any tape library software. User can rebuild the file catalog information manually at any time.

5 Performance

Data I/O throughput speed of the HSM migration is strongly limited by the overhead of DCE/DFS caching disk, which is roughly 1–3 MB/sec[2]. We experienced various DCE timeout problem with HSM migration, until we modified the HSM algorithm. Nevertheless, the HSM migration with DCE/DFS environment is proven to be a viable solution for multi-user accesses under distributed environment.

The direct access method gives the highest performance of the DTF tape drive, which is roughly 12 MB/sec. Number of tape mount request was 36 times per day on average and number of file access requests was 233 times per day for January and February 1997. Tape mount time per request was 63 to 121 seconds (84.7 seconds in average).

The DTF drive is capable of high speed file mark indexing. File positioning time from the beginning of the tape to the bottom of the tape (40 GB) is roughly 160 seconds. Typical operations of file opening and closing take about 10 to 20 seconds.

6 Summary

Combination of the HSM migration and the direct tape access with a large capacity and high speed helical scan tape drive library is proven to be a viable solution for HEP data analysis environment. The system is being used for several KEK Proton Synchrotron fixed target experiment groups since January 1996.

This combination of the HSM migration and the direct tape access will also be used in the Belle offline computing environment with 150 TB mass storage system.

References

[1] H. Fujii et al, in: CHEP'95, ed. R. Shellard and T. D. Nguyen, (World Scientific, Singapore, 1995) p.36.

[2] S. Yashiro et al, in this proceedings.