Hierarchical Storage Management system for KEKB Computing and its Performance

A.Manabe [a], H.Fujii [a], Y.Morita [a], T.Sasaki [a], Y.Watase [a], H.Nakamura [b]

[a] *KEK, 1-1, Oho, Tsukuba, Ibaraki, 305, Japan*

[b] *Sony Corp., 4-14-1,Asahi-cho,Atsugi-shi, Kanagawa, 243, Japan*

The B factory computer system at KEK utilizes a Hierarchical Storage Management (HSM) system for keeping Mini Data Summary Tape(mini DST) data. It consists of 2 servers, 16 clients, 2.4TB RAID and 60TB tape library governed by the Sony PetaServe as a HSMS. This paper aims to show its performance issues and operational experiences.

Key words: HSM; Digital Tape Recorder; Tape Library; RAID.

1 Requirement

The B factory computer system is the system dedicated to the Belle experiment at KEK. The system will process several categories of data listed in Table.1 in each data analysis step. For the data production step (raw data to

category/requirement	capacity	bandwidth
Raw & DST data	16TB/y	$6MB/s$
MC data	10TB/y	<5MB/s
mini DST & end user data	4TB/y	$2MB/s \times 20$
programs	0.5TB	-

Table 1. Data from the Belle

DST data), though only limited programs use the category data, system must provide both huge capacity and guaranteed high bandwidth from/to storage to/from processors. Even in the case when the production processes concentrate in a small set of data, the storage system have to be able to provide the maximum bandwidth to keep up data production pace. To achieve it, rather conventional way of Direct Tape I/O(DIO) has been taken for the process step. For the user home directory where user programs and small size data are saved, DFS on Distributed Computing Environment(DCE) is adopted to give lower latency service for WAN access, expecting it smooths up international collaboration. For mini DST and end user analysis data, HSM system is planed to be used to offer to users medium(but still large) capacity, high bandwidth and easier usability. The performance requirement from the Belle group to the data category is 2MB/s access speed to 20 simultaneous user (40MB/s aggregate bandwidth) with 4TB/year storage capacity.

2 Hardware and Software

The B factory computer system has been installed on Jan. 1997 under 4 years rental contract. Tape Library 1(20TB,11drives), TL2(75TB,15drives), TL3(30TB,8drives) and TL4(30TB,8drives) and 4.2TB RAIDisks ($42 \times$ (RAID controller+100GB disks)) form the storage sub system. A part (Fig.1) of them is operated under HSM systems. The tape library in the figure corresponding to TL3 and TL4 in the whole system. Eight Sony DTF drives of GY10 in each library are connected to one Fujitsu(SUN) S-7/300U m170 (4 Ultra SPARC-I CPUs) working as HSM and NFS server. A HSM server also has 4 independent SCSI2 W/D lines each linking to a 100GB RAIDisk with RAID5 setup (8 data + 2 parity disks). Twelve S-7/300U m170 file servers each manages 100GB RAID works as HSM clients and NFS servers. The system has two complete set of this HSM subsystem. Fujitsu appropriate 'torus' shaped network so called 'APnet' connects the HSM system and Computing Servers. The theoretical maximum transfer speed of the network is stated as 200MB/s. Computing Servers are fed data from the HSM server and clients through the APnet by using NFS version 3. User workstations distributed in KEK site are connected by ATM or FDDI network to the system as well. The PetaServe by SONY is a variation of Computer Associate's OSM (Open Storage Manager) version 2 and has been selected for HSM software. The PetaServe has some gains in comparing with traditional Staging mechanisms and other HSM softwares in several points; 1) User transparent: User accesses data just the same way as accessing usual UNIX files in a disk. 2) Double buffering: During file reload from tape to disk, soon after chunks of a file are read, an application can use the read part of the file,. The rest of migration transfer is simultaneously going as the application program uses the data. 3) HSM clients (has only disks) and a server (has the tape library) can be transparently distributed in a TCP/IP network. 4) migrated files may be transferred to other HSM clients of same storage group without reloading. 5) Averaging load of migrations (data transfer disk to tape) is possible. 6) Parallel transfer by using multiple tape drives can speed up both migration and reloading. In the other hand, some limitations in using the PetaServe were found; 1) A tape library used by PetaServe could not be used for other purposes (DIO, backup, etc.), 2) A Tape library cannot be shared among more than one PetaServe server machines. This limitation might be a problem in expansion of the system.

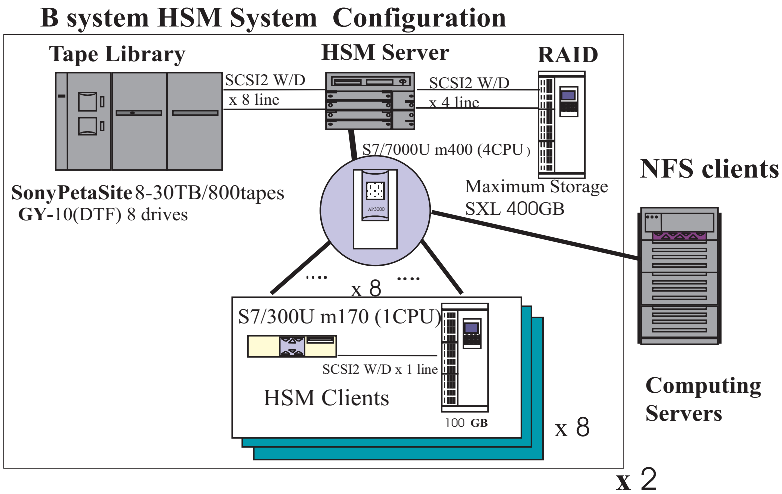

Fig. 1. Hardware configuration for HSM system

3 performance issues

In considering about HSM system performance, performance of 'migration' (disk to tape) and that of 'reloading' (tape to disk) come to mind. However, since the bandwidth required for migration could be averaged by scheduling and shadowing, migration speed is not a matter. In the other hand, reloading speed directly impacts system usability. To scale the reloading performance, we would like to introduce 'average access time' which is a average latency in data access. In the case of HSM, it could be, $T_{acc} = T_{disk_{a}ccess} + T_{tape_{a}ccess} \times Probability_{tape_{a}ccess}$ If a user access to data which is not in the disk hierarchy, user have to wait additional time (the 2nd term in the equation). The second term is considered to be a cost to pay for 'larger storage capacity'. To check feasibility of the system, model calculation of the access time was done under some assumptions ; 1) file size is 2GB. 2) tape access ratio is 0.1 (10% of whole accesses invoke reloadings). 3) 10 simultaneous access to each HSM system. Tested situations are ; Case1 (server congestion) is the case that all 10 NFS accesses from the computer servers are concentrated into one HSM server. Case2 (most probable case) is the case that 10 simultaneous accesses are distributed among 9 HSM clients and a server according to its disk capacity. A HSM server which has 4 times larger disk capacity than clients, is expected to get 4 times many average accesses. Case3(independent access) is that 9 (not 10) simultaneous accesses divided into 9 clients and server evenly (each NFS server receive one access).

Case	access time(sec)	effective speed(MB/s)
Case1 (server congestion)	2093 + 215(10%)	0.98 - 0.10
Case2 (most probable)	529 + 50(9%)	3.9 - 0.4
Case3 (independent)	263 + 32(12%)	7.6 - 0.9

Table 2. Access time.

The calculation uses system performance actually measured (Fig. 3 and 4) as input parameters and also includes a tape loading and positioning time on the tape library;
$T(sec) = 30(mediamount){+}4{+}3.5{\times}position(GB)(positioning){+}30(loading)$.
In our system, transfer speeds are concluded as

$$NFSspeed < HSMspeed < Disk, Network, Tapespeed$$

and file has such size as a tape loading time is much smaller than data transfer period. This keeps the additional access time (second term in the equation) ratio to disk access time staying the tape access ratio(10But smaller file size makes the additional access time enhanced. For example, in the case of 100MB file under the Case3, tape loading time is comparable to data transfer time and averaged access time becomes 70 % bigger incontrast to normal file systetem.

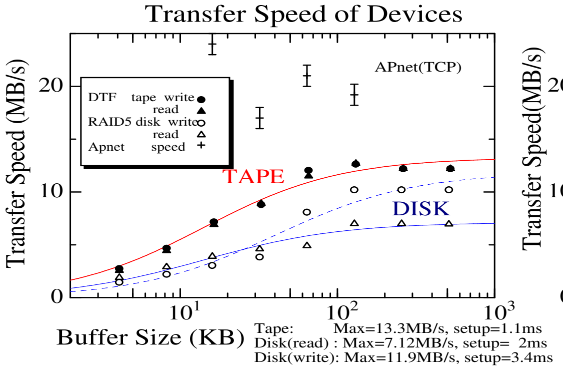

Figure 2.

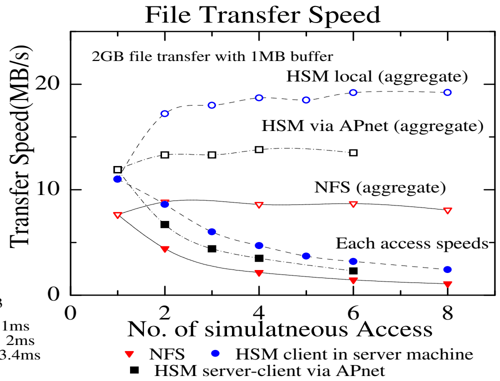

Figure 3.

4 summary

For keeping mini DST data, Sony PetaServe HSM software has been installed in B factory computer system at KEK. A model calculation shows the system performance is able to agree with the requirements of the Belle group, To fulfill the performance, combination of high speed tape drive, high speed RAID with intelligent controller and high speed networks were necessary. It was found that to lessen average access time, file size should be large enough to compensate a long loading time of helical scan tape. Further effort will be done under collaboration with Sony corporation and KEK to overcome the issues.