NEW COMPACT HIERARCHICAL MASS STORAGE SYSTEM AT BELLE

REALIZING A PETA-SCALE SYSTEM WITH INEXPENSIVE IDE-RAID DISKS AND AN S-AIT TAPE LIBRARY

KATAYAMA, Nobuhiko
YOKOYAMA, Masahiko
HIBINO, Taisuke
HIGH ENERGY ACCELERATOR RESEARCH ORGANIZATION
Oho 1-1, Tsukuba-shi, 305-0003, Japan

MAKINO, Morihiro
System works Co.
Hamamatsu S building, Nishi-Asada 2-10-22, Hamamatsu-shi, 432-8045, Japan

GOTO, Kazuyuki
HIKITA, Jun
TAMURA, Koichi
Sony Corporation
Kitashinagawa 6-7-35, Shiagawa-ku, Japan

Abstract

To cope with more than one peta bytes of data for the Belle experiment, a compact hierarchical mass storage system was built using inexpensive IDE RAID systems and an SAIT tape library. The system consists of (56)150 TB of online disk systems on 8(18) file servers with 0.5(1.29) PB tertiary tape library (when extended).

THE BELLE EXPERIMENT AND ITS ONLINE DATA PROCESSING

Belle is an experiment at the KEK B-factory. Its goal is to study the origin of CP violation. It started taking data in 1999 and has accumulated more than 274 million B meson-anti B meson pairs. The KEKB accelerator produces more than 1 fb^{-1} of data in one day. Belle now collects more than one million B meson-anti B meson pairs per day. It corresponds to more than 1 TB of raw data. The raw data are written directly on Sony DTF2 tapes. Starting in fall 2003, Belle started to run a full event reconstruction program online, producing DST (data summary tape) format data at the time of the data taking.

This enables us to diagnose the Belle detector and status of electron positron collision at the KEKB accelerator in real time. For example, the instantaneous luminosity can be tuned by looking at the precise position and size of the collision point. The DST files are written at the rate of 1GB/min. Next, mu pair, Bhabha and hadron events are skimmed and written into separate files as Belle uses stream I/O model. The DST and skim files can be as large as 2 TB per day. The calibration constants of sub detectors are then generated and, if necessary, second (final) DST production is done following the updates of the calibration constants.

NEW COMPACT HSM SYSTEM

The computing model till last year used direct access tape drives for reading and writing of these files as the amount of data are so large and we could not use file systems on hard disks. However it would be much nicer if one can use an ordinary file system to keep hundreds of tera bytes of data and not worry about the disk quota and/or free space left on the disk systems. Hierarchical mass storage system can fulfill such a requirement. This poster describes how we constructed such a system using large inexpensive IDE RAID systems and an S-AIT tape library system.

The IDE RAID systems have become popular as they are extremely cost effective. For example, using 16 400GB disk, a 5+ TB disk system can be built at a fraction of the cost of a high performance fiber channel RAID system. We then combine it with a tertiary file system based on magnetic tape technology. This way, we can construct a large file system with good performance inexpensively. The Belle group, in collaboration with Sony Corporation, Sony Broadband Solutions Corporation and Systemworks Corporation, built a 500TB HSM system this March. The system consists of sixteen RAID systems on eight file servers. The RAID system consists of sixteen 300GB ATA disks and the total size of the disk systems is more than 56TB. The tape library consists of four SAIT tape drives and three cassette

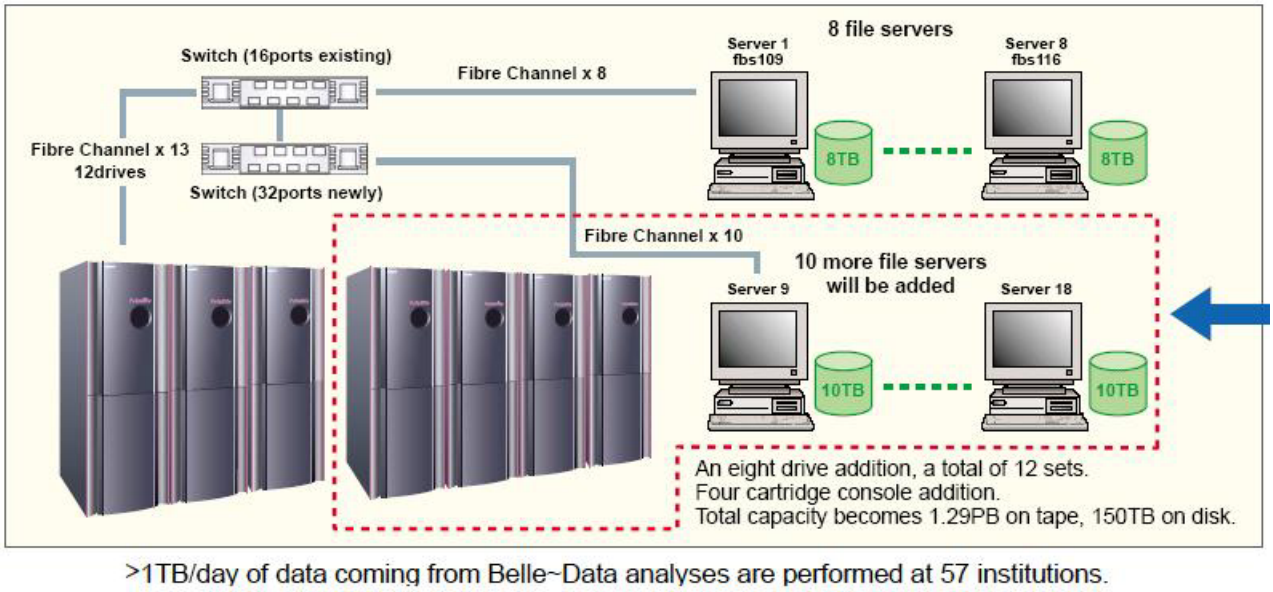


Figure 1: The schematic picture of the new compact HSM system. The system was build at the end of 2003. The components inside the red dashed line are the extension being installed by the end of 2004.

consoles that can hold 1000 tapes. Each tape can hold 500GB uncompressed, totaling 500 TB.

The maximum tape read/write speed is 30MB/s. The four drives and eight file servers are connected via a sixteen port fiber channel switch. The operating system of the file servers is based on Redhat Linux version 7.3 with extensions in order to manage the hierarchical mass storage system, PetaServe. PetaServe uses XFS file system as one of the unique features of XFS is called the Data Management API (DMAPI) through which a hierarchical mass storage system can be implemented in between file systems and the kernel. The system consists of physically independent 32 disk partitions on eight servers with total storage capacity of 500 TB when the tape library is filled with tapes. We stored most of mini-DST files that contain 274 M BBbar pairs as well as more than one billion e+e- to qqbar events in this system. More than 100 users used these files before summer conferences from more than 200 dual CPU Xeon compute servers under LSF. For performance reasons, we do not use NFS but use a simple protocol to read and write on disks on remote node. File servers had, at the busiest time in the worst case, more than 50 connections per server. Each file server is connected to the network via a GbE line. We have not measured the aggregate speed of the data transfer but each connection can exceed the transfer speed of 8MB/s. The limitations comes from the physics analysis and decompression inside the Belle software to read input files.

Although there were number of disk failures and power outages we have not lost data at all. We have stored close to 200 TB of data in four months since the start of the operation of the system. The system seems quite reliable and is easy to maintain. We do distribute data among 32 disk partitions on eight servers by hand, making two copies of each data files so that the accesses to data are balanced. The system worked more than expected and we decided to triple the capacity this year; We extend the capacity of the tape library to 1.29PB and the disk space, 150TB by adding eight S-AIT tape drives and four cassette consoles and 20 IDE-RAID systems, each of which consists of 16 400GB SATA disk made by Hitachi GST, on ten file servers, and a 32 port fiber channel switch connecting them. The new HSM system is quite compact. The entire system can be fit in eight consecutive 19-inch rack space (1.29PB Sony SAIT Peta Site) and four 19-inch racks (18 2U file servers and 36 3U RAID-systems). The floor space, including the working area is less than 30 m^2 for both tape library and the file servers. The electric power consumption is roughly 20KVA. The total system cost including the new extension and tape media for 800TB is less than 1.5 million Euros.

Using this system, we plan to migrate all data kept on DTF(2) tapes. At the beginning of the Belle experiment, the raw data were written on DIR tapes which hold 100GB per tape. We copied the raw data on DIR tapes onto DTF2 tapes. We now have more than 2500 DTF tapes (40GB each) and 4500 DTF2 tapes (200GB each). The total amount of data exceeds one peta bytes. These include copies (duplication) of raw data in the raw data tapes and in the DST tapes. For fire safety, we keep the DST tapes in a separate storage space at KEK site. We plan to copy all data on DTF(2) tapes to the new HSM system. This operation requires 20TB/day of organized data movement. 20TB/day corresponds to 200MB/s data transfer rate. Since the DTF tape library system is located

at the Computing Research Center and the SAIT HSM system in the Belle experimental hall which are 1.5km apart, we installed 10Gbps link between the two sites.

FEATURES OF SAIT DRIVE TECHNOLOGY

SAIT introduces a tape-based data storage technology platform that in its first generation delivers the industry's highest capacity tape drive, storing up to 500GB of uncompressed data on a single-reel, half-inch tape cartridge and featuring a sustained native transfer rate of up to 30MB per second uncompressed.

Figure 2: Four file servers with eight 16 300GB IDE RAID systems

First generation SAIT drives incorporated into automation solutions will provide uncompressed capacities ranging from 10TB in a space-efficient configuration to more than 500TB in a 1,000-cartridge freestanding library.

SAIT technology utilizes Advanced Metal Evaporated (AME) media, which combines high capacity with high durability.

AME creates a recording layer of nearly 100% magnetic material, in contrast to conventional tape technologies that may have less than 50% magnetic material.

Higher density and capacity is the result, as well as a smoother tape surface that significantly prolongs head and media life. SAIT leverages Sony's R&D investments, as well as the field-proven AIT recording technology, but scales this into a half-inch cartridge and full-height 5.25-inch mechanism.

Based on advanced helical-scan recording technology, which is known for its high data density, outstanding data transfer performance, as well as outstanding reliability and durability advantages, the SAIT technology platform has a forward-looking roadmap of four defined generations that have the potential to scale to up to four terabytes of uncompressed capacity in a single cartridge.

PETASERVE

PetaServe is commercial hierarchical storage management software from Sony. Belle has been using it since 1999 on Solaris. It now runs on Linux (RH7.3 based) using Data Management API of XFS developed by SGI. When a drive and a file server are connected via FC switch, the data on disk can directly be written to tape.

Files are migrated from the disk if they are not accessed for long time when more files are written onto the disk or files that had been migrated are being read (and therefore restored). With fiber channel network, the file server can directly write data onto tape as if the tape drive is locally attached to the server.

The system administrator can tune various migration policies so that the system fits to migration pattern of the users. The user can issue migrate commands if he/she knows the file he/she has written may not be accessed for long time. It can be setup so that every night, newly written files on disks are migrated while keeping a copy on disk (called shadowing). The minimum size of the files which can be migrated and the size of the beginning of the files which remain on disk even when the rest of the files are migrated can be set by the administrators.

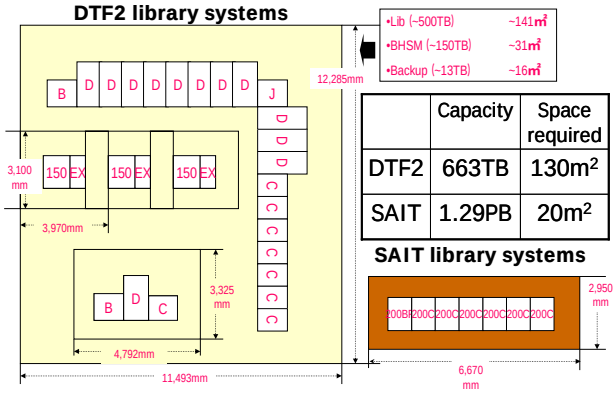
The backup system, PetaBack, works intelligently with PetaServe; In particular, in the next version, if the file has been shadowed (one copy on tape, one copy on disk), it will not be backed up, saving space and time. So far, it's working extremely well. Some disadvantages of the HSM system are; as it is not a single rooted file system, the files must be distributed by hand among disk partitions (now 32, in three months many more as there is 2TB limit…). As the front-end disk by itself is a file system and not cache, usage as a scratch disk space may waste space on tape although it has an automatic garbage collection mechanism.

Belle sets the minimum size and the size on disk to be 8MB as we traditionally use 4MB fixed length buffers for reading and writing data from the stream. A backup solution, PetaBack, operates in conjunction with PetaServe. In particular, the next version of PetaBack is supposed not to back up files that are migrated by PetaServe, but back up stubs of the migrated files. This will greatly reduce the time and tapes it takes to backup the 150TB disk system.

Figure 3: Sony SAIT PetaSite tape library. Seven tape drives, control PC, 16 port fibre channel switch are in the left most console. The left-most console holds 200 tapes (100TB). Each of the right two consoles hold 400 tapes (200TB).

FLOOR SPACE REQUIREMENT

The system requires less than 30m^2 of space including workspace. Space required for the new SAIT HSM system (1.29PB tape, 150TB disk) is six times less than the DTF2 HSM system of 663TB (tape) capacity.



	Capacity	Space required
DTF2	663TB	130m^2
SAIT	1.29PB	20m^2

Figure 12: Floor space required with the new SAIT HSM system (1.29PB tape, 150TB disk) is six times less than the DTF2 HSM system of 663TB (tape) capacity

CONCLUSIONS

We have built a compact HSM system using inexpensive IDE RAID system and SAIT PetaSite tape library using PetaServe HSM software. The original system of 500TB tape + 56TB disk system works very well and we are extending it to have the capacity of 1.29PB tape and 150TB disk.