Design and evaluation of a hybrid storage system in HEP environment

Qi Xu¹,², Yaodong Cheng² and Gang Chen²
¹University of Chinese Academy of Sciences, Beijing, China
²IHEP computing center, 19B Yuquan Road, Beijing 100049, China

E-mail: xuq@ihep.ac.cn

Abstract. Nowadays, the High Energy Physics experiments produce a large amount of data. These data are stored in mass storage systems which need to balance the cost, performance and manageability. In this paper, a hybrid storage system including SSDs (Solid-state Drive) and HDDs (Hard Disk Drive) is designed to accelerate data analysis and maintain a low cost. The performance of accessing files is a decisive factor for the HEP computing system. A new deployment model of Hybrid Storage System in High Energy Physics is proposed which is proved to have higher I/O performance. The detailed evaluation methods and the evaluations about SSD/HDD ratio, and the size of the logic block are also given. In all evaluations, sequential-read, sequential-write, random-read and random-write are all tested to get the comprehensive results. The results show the Hybrid Storage System has good performance in some fields such as accessing big files in HEP.

1. Introduction
The scale of high energy physics experiment is expanding such as the Large Hadron Collider (LHC) in Europe, Beijing Electron Positron Collider II (BEPC II), the neutrino experiment in Daya Bay, the cosmic ray experiment in Yangbajing massive experimental data are generated [1]. At the same time many HEP experiments are under constructed such as the second phase of the Daya Bay neutrino detector experiment, and more experimental data will be processed.

In the face of mass amounts of experimental data, efficient storage device is essential in the system, and the performance, capacity, reliability and the cost-performance should be considered to design the storage system. At present, hard disk drives, tapes and solid-state drives are used as common storage devices. Among these devices tapes are used to store the cold data which should be analyzed for a long period, hard disk drives are usually used in normal computing devices and solid-state drives are usually used for storing hot data because of their has high I/O performance and high cost [2]. Therefore the way of combining the hard disk drive (HDD) and the solid-state drive (SSD) efficiently can help to get a cheap and efficient hybrid storage device and provide a better storage environment.

2. The architecture of hybrid storage system in High Energy Physics computing system

2.1. Typical architecture in High Energy Physics computing system
Computing in high energy physics is a typical data-intensive computing and there are mainly large files to be analyzed. The essence of computing in the high energy physics is data mining and different
files are analyzed by different computing nodes simultaneously without needs for communications with each other [3]. Therefore, the characteristics of high energy physics computing are data concurrency and high throughput, so it does not need some parallel computing technologies such as MPI. Based on the characteristics, cluster computing systems are widely used in high energy physics, and the computing and storage are separated. The typical structure of storage system is shown in Figure 1.

![Figure 1. The structure of typical storage system in HEP.](image)

Data processing in HEP includes three types: simulation, reconstruction and physical analysis. When analysis tasks are running on computing nodes, most of task requests distribute between 256KB to 4MB and 65% of the offsets between two consecutive requests are between 1MB to 4MB, which indicate the way of accessing files is skipping among big data blocks [4]. Therefore, the performance of computing and I/O system in high energy physics will be improved by the technology of caching big data blocks.

2.2. Hybrid storage system in High Energy Physics computing system

Nowadays, there are two popular kinds of hard drives SAS and SATA. The speed of reading and writing in SATA is slow and SAS is usually composed of multiple disks [5]. Now, the price of flash-based storage stays high. Aimed at this situation, a hybrid storage system is designed to provide a perfect storage environment, as shown in Figure 2.

![Figure 2. The structure of hybrid storage system in HEP.](image)

A hybrid storage layer is designed between the I/O server and the physical storage device, the hybrid storage layer uses the technology called Flashcache [6]. Flashcache works as a module of Linux kernel which is dynamically loaded without modifying the kernel when it is used, and it can keep the original data layout on the disk [7]. As a universal model, the solid-state drive and the mechanical disk are integrated into one block device which has improved I/O performance while it is still cheap. The solid state disk works as a cache for HDD to accelerate the I/O speed and each virtual block corresponds to each logical address of the disk, which makes no difference for upper-level file
systems to access the virtual device and a disk device. The design of the hybrid storage technology is based on the Linux Device Mapper Framework (DM) which is located between the virtual file system layer (VFS) and the block device layer, as shown in Figure 3. There are a lot of plug-in modules based on the DM framework such as LVM2 (Linux Volume Version 2), EVMS (Enterprise Volume System), DMraid (Device Mapper Raid Tool) and so on [8].

![Figure 3. Hybrid storage module in the kernel.](Diagram)

The data on the solid state drive and the hard disk drive are divided into blocks, and the solid state disk is divided into the metadata area and the cache data area. By default, every 512 blocks are grouped into a group and the data blocks are mapped to the cache blocks by group reflection. The address of a data block (the ID of the data block on the disk) in sectors is mapped to the cache group directly, then it is mapped to the only cache block through the linear hash detection.

3. The benchmarks and optimizations of hybrid storage system

3.1. Benchmarks

3.1.1. Test environment. The test environment including hardware and software is introduced below.

**Hardware**
- CPU: Intel(R) Xeon(R) E5-2407 v2 2.4GHz*8
- Memory: 8GB DDR3 1600MHz*4
- Motherboard: Dell 0K7WRR
- HDD: HDD--Seagate ES.3 ST4000NM0033-9ZM170 (7200rpm, 4TB, 128MB)
- SSD: Samsung SSD 850 PRO S25UNSAG401352F (256GB)

**Software**
- Operating system: CentOS 7.0 x64 Linux / Linux version 3.10.0-229.1.2.el7.x86_64
- Test tool: fio-2.2.8[9]

There are mounts of distributed storage nodes in high energy physics computing environment and one of them is chosen to make the test. In each test bellow, only one factor changes and the others are fixed to show the effect. The cache block size and the number of concurrent processes as the system parameters can be set independently.

3.1.2. Performance comparison of storage devices. Firstly, the I/O performance of three types of storage devices (Solid State Drive (SSD), Hard Disk Drive (HDD) and Hybrid Storage Drive) is tested. The test results are shown in Figure 4-1 to Figure 4-4.
3.1.3. Hybrid ratio test. Tests about the hybrid ratio are made in Figure 5-1 to Figure 5-4. In these tests the size of the HDD is fixed at 5GB and the size the SSD is adjusted between 512MB and 5GB. The IOPS of the hybrid storage system with 5GB SSD is much higher than others, so the ordinate use logarithm of 2 as unit.
3.1.4. Cache block size test. As mentioned above the basic scheduling unit in hybrid storage system is cache block and most operations in the system are for cache blocks. The size of cache blocks has significant impact on the performance. Several sizes of cache blocks are used and their performance are tested, the results are shown in Figure 6-1 to Figure 6-4.

![Figure 6-1](image1.png)  Sequential-read with Different Cache Block.

![Figure 6-2](image2.png)  Sequential-write with Different Cache Block.

![Figure 6-3](image3.png)  Random-read with Different Cache Block.

![Figure 6-4](image4.png)  Random-write with Different Cache Block.

3.2. Analysis of the results
According to the results in 3.1.2, the performance of the hard disk drive is relatively stable and lower, while the solid state drive shows higher performance in reading and writing in the case of multiple processes. The hybrid storage drive shows excellent performance in writing and reading which is similar with the solid state drive.

In the chapter 3.1.3, the size of the solid state drive is adjusted from 512MB to 5GB and the size of the hard disk drive is kept as 5GB, then the performance of the hybrid storage system is tested. The performance of the system decreases drastically when the hybrid ratio decreases. That is because the cache misses occur what resulted in the replacement of the cache block and the access to the hard disk directly.

The cache block is the basic unit of the storage, when the cache block is larger the data granularity will be larger, I/O scheduling will be reduced and IOPS will be lower. But when the cache block size is too small, the CPU will be very busy and the speed of I/O will decrease. Generally, the size of the cache block is better from 4KB to 8KB.

The cache block size, the number of concurrent processes and the hybrid ratio are all the system parameters which can be set independently and they will not affect the others.

3.3. System optimizations
A variable named VOF is defined to indicate the cost performance by Formula 1. IOPS\textsubscript{x} represents the IOPS of the system with xGB SSDs, IOPS\textsubscript{0} represents the IOPS of the system without SSDs and Price\textsubscript{x} represents the cost of xGB SSDs.

\[
VOF = \frac{IOPS\textsubscript{x} - IOPS\textsubscript{0}}{Price\textsubscript{x}} \tag{1}
\]

In this case the I/O performance of the hard disk is taken as the standard value and the difference between the hybrid storage device and the hard disk drive is calculated. Taking the random-write as an example, the result is shown in Table 1. Besides the performance the cost is also considered. According to market research it is found that the price of the solid state drive increases 1.5 to 3.0 times when the capacity doubles. This price growth is also considered into VOF.

### Table 1. Difference of IOPS in Random-write Test

| The Size of The Solid State Drive in The Hybrid Storage System | IOPS | The Difference |
|---------------------------------------------------------------|------|----------------|
| 0MB                                                           | 471  | 0              |
| 512MB                                                         | 545  | 74             |
| 1GB                                                           | 632  | 161            |
| 2GB                                                           | 836  | 563            |
| 3GB                                                           | 1297 | 826            |
| 4GB                                                           | 2513 | 2042           |
| 5GB                                                           | 9619 | 9148           |

According to the test results in chapter 3.1, it is shown that the optimal configuration in the computing environment is that the hybrid ratio of hard disk drive and solid state drive is 1.43:1, the number of concurrent processes is 11 and the size of the cache block is 8KB. To get the appropriate size of SSDs used in the system, the cache block size is set to 8KB and the number of concurrent processes is set to 11 as independent parameters of the system. The ratio of IOPS difference and price difference is used to evaluate the cost performance. The growth rate of VOF is shown in Figure 7 and the polynomial is taken to fit the data point in Formula 2.

\[
y = 10.133x^4 - 117.83x^3 + 482.52x^2 - 795.34x + 502.75 \tag{2}
\]

**Figure 7. The Change of VOF.**

In Formula 2 it is shown that the performance of the hybrid storage system reaches high magnitude when the solid state drive is about 3.5 GB. At the point the solid state drive gets efficient use and the VOF of hybrid storage system is higher. It is a scaled test for the actual computing environment in high energy physics. The storage in high energy physics computing environment is usually at PB level and the system is configured with the corresponding parameters from the test to get a better performance. The configuration parameters are applied to optimize the local computing environment, then jobs are tested to show the actual I/O performance, as shown in Figure 8 and Figure 9. In Figure 8, Server.HDD represents the server with hard disk drives, Server.HSD represents the server with the hybrid storage system and R-read represents random-read, S-read represents sequential-read, R-write represents random-write, S-write represents sequential-write. In Figure 9, Single.Job represents only one job runs on the server, Multi.Jobs represents multiple jobs run on the server.
4. The deficiencies of the hybrid storage system

There are two deficiencies of the hybrid storage system. Firstly the management of the metadata takes synchronous update and batch update. The waiting time for each update increases the updating time. On this point it can be changed to asynchronous update to speed the processing of the metadata. Secondly it needs the I/O parallelization instead of serial execution to get a better I/O performance. It is because there is a mutex lock in the whole cache and it is too large for frequent I/O. Data sets can be considered and one mutex lock should be applied in each data set to provide better parallel capability.

5. Conclusion

Nowadays, large amounts of experimental data and high concurrency of job files are significant challenges in storage systems and the traditional storage system with hard disk drives cannot meet the requirements of the storage. The hybrid storage system can provide high I/O performance in high energy physics computing environment with low cost. Although the hybrid storage system has some deficiencies, it brings a lot of performance benefit. The principle of the hybrid storage technology has been analyzed, the I/O performance of it has been tested and the factors affecting the performance are analyzed. The solid state drive becomes popular and the architecture of storage system may be innovated in the future. But for now, the hybrid storage technology still plays an important role in storage systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 11575223), National key Research Program of China “Scientific Big Data Management System” (No. 2016YFB1000604), the Chinese NSF grants “Research on fine grained Event Service for the BESIII offline software and its scheduling mechanism” (No. 11675201) and “Research on BESIII offline software and scheduling mechanism on desktop grid” (No.11405195).

References

[1] WLCG - Worldwide LHC Computing Grid, http://lcg.web.cern.ch/LCG,2013.7.
[2] Cabrera L, Long D D E. Swift: Using Distributed Disk Striping to Provide High I/O Data Rates. Computing Systems, 1991, 4(4): 402-441.
[3] Cheng Y D, Shi J Y, Chen G. A survey of High Energy Physics Computing System[J]. e-Science Technology & Application, 2014, 31: 1189-1194.
[4] Cheng Y D, Shi J Y, Chen G. Design and Optimization of Storage System in HEP Computing Environment[J]. Computer Science, 2015, 42(01):54-58.
[5] Mituzas D. Flashcache at Facebook From 2010 to 2013 and beyond[EB/OL]. [2013-10-9]. https://www.facebook.com/notes/facebook-engineering.
[6] Facebook/Flashcache. https://github.com/facebook/flashcache/.
[7] More A, Ganjewar P. Dynamic Cache Resizing in Flashcache[J]. Advances in Intelligent Systems and Computing, 2015, 327:537-544.
[8] Ao Q Y. Analysis of the Storage Technology Principles: Basedon the Linux Kernel 2.6, 2011. 363~476.
[9] Freecode/fio. http://freecode.com/projects/fio/.