Study of Fast Data Access Based on Hierarchical Storage for EAST Tokamak

F. Wang[1], *Member IEEE*, Y. Chen[1], S. Li[1], Y. Wang[1], X.Y. Sun[1], F. Yang[2]

***Abstract*–The Experimental Advanced Superconducting Tokamak is a large fusion research device which aims at long-term and high parameters plasma operation. A distributed and continuous data acquisition system based on PXI/PXI Express technology has been developed for EAST. At present the system has more than 50 data acquisition units and more than 2000 raw channels, the maximum data throughput is about 5GBytes/s. How to access the large amount of data as fast as possible becomes a big problem. So we planned to design a special hierarchical data storage system for EAST. The system is composed of 4 storage tiers. The tier1 is based on PCIe SSD storage which provide the data access cache. The tier2 is the local SAS raid on MDSplus server cluster for real-time data collection from the data acquisition units. The tier3 is Lustre storage which stored all the current campaign data and can provide fast parallel data access for data processing. The tier4 is NAS storage for historic data achieving. A schedule program is designed to control the data flow between tier1 and tier2/tier3/tier4. The system has been designed and will be adopted in the next campaign and the system details will be given in the paper.**

I. Introduction

The Experimental Advanced Superconducting Tokamak (EAST) is a large fusion research device which aims at long-term and high parameters plasma operation. A distributed and continuous data acquisition (DAQ) system based on PXI/PXI Express technology has been developed for EAST. At present the system has more than 50 data acquisition units and more than 2000 raw channels, the maximum data throughput is about 5GBytes/s. In case of long-term discharge mode, the raw data of one shot is more than 1TBytes, and the total data of one campaign is more than 100TBytes [1]-[5]. How to access the large amount of data as fast as possible becomes a big problem.

At present all the acquired data except video/image are stored into MDSplus database which is a set of software tools for data acquisition and storage and a methodology for management of complex scientific data [6]. There are several MDSplus trees for EAST as listed below.

- Diagnostic raw data;
- Diagnostic data of 1K/s frequency reduction;
- Engineering system data;
- Analysis and calculation data;
- EFIT calculation and simulation data;
- Plasma control system data;

These MDSplus trees are organized as distributed mode and accessed in different servers. The distributed system can accelerated data access in general. However in case of long-pulse discharge, access to the large amount of data is still slow for client tools.

At the same time the data access frequency of different shots and signals are different, some data are accessed frequently by all users while some data are only access by a few persons. There are currently commercial tiered storage production to treat these issue, but they are very expensive and difficult to be customized according to the data characteristic of fusion experimental.

So we planned to construct a special hierarchical data storage system for EAST tokamak.

II. System Architecture

According to the requirement and current status of EAST data acquisition and management, a new data storage system architecture has been designed as shown in Fig.1. The whole data storage system can be divided into 4 tiers.

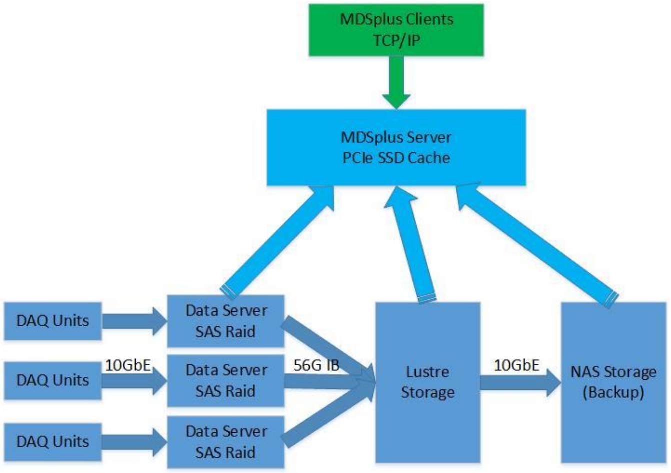

Fig. 1 System architecture

Manuscript received May 30, 2016. This work is supported by the National Magnetic Confinement Fusion Science Program of China (Grant No: 2014GB103000) and National Natural Science Foundation of China (Grant No: 11405205).

F. Wang is with the Institute of Plasma Physics, Chinese Academy of Sciences, 350 Shushanhu Road, Hefei, Anhui 230031, P. R. China (phone: 86-551-65592311; e-mail: wangfeng@ ipp.ac.cn).

Y. Chen, S. Li, Y. Wang and X. Y. Sun are with the Institute of Plasma Physics, Chinese Academy of Sciences, 350 Shushanhu Road, Hefei, Anhui 230031, P. R. China.

F. Yang is with the Department of Computer Science, Anhui Medical University, Hefei, Anhui 230032, P. R. China (email: fyang@ipp.ac.cn).

A. Tier1

Tier1 is the main MDSplus server which is installed with PCIe Solid-State Disk (SSD). When client access the data, server will check cache firstly, if data are not found tier2/tier3/tier4 storage will be searched by order. There is only cache data on tier1 while tier2/tier3/tier4 are real data storage. The detailed information about tier1 is given in the next chapter.

B. Tier2

Tier2 is the data servers for new data collecting from DAQ units. The acquired data are transferred to these servers continuously during discharge. There are 6 high performance data servers DELL R730XD which is installed with 12x6TB disks and the available capacity is more than 50TB. The total volume of the 6 servers are more than 300TB which is enough for the new data storage of one year discharge experiment. The continuous data streaming on each servers are more than 1GB/s, and 10GbE interface is also installed so that the data bandwidth of each server is more than 1GB/s and the total bandwidth is about 6GB/s, which is satisfied with the requirement of the diagnostic data acquisition.

C. Tier3

Tier3 is the primary data storage which is designed based on InfiniBand networking and Lustre file system, for the purpose of fast data access and parallel computing. The storage is shared with the High Performance Computing Cluster and other calculation servers. There are 144x6TB SAS drives and the total available capacity after RAID6 is about 600TB. The write speed of single thread is about 1GBytes/s and the maximum data throughput is about 8GBytes/s [7]-[10]. The collected data of tier2 servers are transferred asynchronously to tier3 by InfiniBand network while there is no discharge experiment.

D. Tier4

Tier4 is the data backup storage designed based on Network Attached Storage. We use a high performance data server installed CentOS Linux act as NAS backup server, which is connected by SAS to DELL MD3460 enclosure and MD3060e expansion. There are 120x6TB SAS drives and the total available capacity is about 600TB. The maximum data throughput is more than 1GBytes/s. All the data on tier3 are backup to tier3 automatically using rsync service.

III. System Design

Since MDSplus is database structure so the Input/Output Per Second (IOPS) of data storage is very important for data access speed. The performance of MDSplus on NFS storage is much lower than local SAS raid storage, even 10GbE parallel NAS is used. System RAM is the best cache media however it is very expensive and the capacity is also limited compared with the data size of EAST, then we have to consider some cost performance solution. Now the PCIe SSD technology has been widely used for good performance and the price is also not much high. So we choose PCIe SSD as the data cache.

A. PCIe SSD

Compared with the traditional drives, SSD has very good IOPS value, especially PCIe SSD. The following table shows the comparison results where the Lustre is the current Lustre storage system on EAST.

Table I. IOPS comparison (4KB)

Disk & Storage	IOPS
SATA disk	50~100
SAS disk	100~150
FC disk	100~150
SATA SSD	50K~100K
PCIe SSD	100K~500K
Lustre single	1K~2K
Lustre parallel	100K~200K

From the results we can see the IOPS of one PCIe SSD is much larger than the whole EAST Lustre storage system.

B. Hardware Configuration

There are many PCIe SSD and the price are much different, and finally we choose the Intel PCIe SSD DC P3700 according to the good performance features [11].

- Sequence Read/Write: 2800/2000MB/s
- Sequence Latency Read/Write: 20/20us
- 4k IOPS Read/Write: 460K/175K
- Endurance Rating: 62PBW
- 2TBytes

We build our cache system using a high performance server DELL R730 as flowing configuration.

- E5-2643 v3x2
- 256GB RAM
- SATA SSD
- 56Gb InfiniBand
- 10GbE

At most 4xP3700 board can be inserted into the PCIe slots and the total available cache size is about 8TB.

C. Data Flow

The primary data flow schema is shown in Fig. 2.

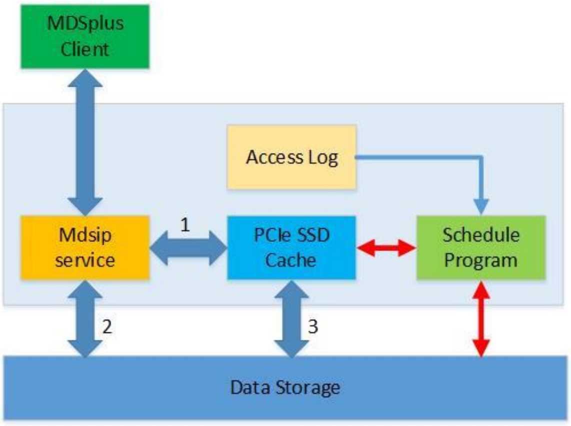

Fig. 2 Data flow schema

When Mdsplus client send the data request to mdsip service, the following steps are performed.

Step1: Firstly, mdsip service will search the PCIe SSD cache for the data;

Step2: Secondly, if data are not found, mdsip service will check the other data storage including the server local raid, Lustre storage and NAS backup storage, and the access information will be logged automatically;

Step3: The schedule program will monitor the mdsip and analysis the access log automatically, then manage what data should be cached in the PCIe SSD.

D. Schedule Program

What kind of data should be cached in PCIe SSD is important, so we developed a schedule program to manage the data flow automatically between tier1 and tier2/tier3/tier4, and four types of data will be cached in the schedule as following.

(1) The newest data. This is the data of recent several days. At present the data of the last 5 days will be stored and the total amount is about 2TB. This algorithm is static configuration.

(2) The most important data. This is the data from the important diagnostic system which is important for discharge experiment such as the plasma current, temperature, and density. About 2TB volume will be reserved for these data. This algorithm is also static configuration.

(3) The most frequent data. This is the data accessed most frequently by client tools. It is dynamic procedure and the primary implementation steps are described below.

- The access information of all clients are logged;
- The log files are analyzed in real-time;
- The access info database is built from the analysis;
- The cache is refreshed according to the database;

(4) The high priority data. It is dynamic and static procedure and the primary implementation steps are described below.

- The high priority user database is built;
- The access information of the users are monitored;
- The access action and behavior is analyzed;
- The cache is refreshed according to the user preference;

The schedule program is developed by GNU C language and running as system daemon to monitor the cached data. The data schedule algorithm (1) and (2) have been developed, while (3) and (4) are still under developing.

IV. Summary

To get fast data access to EAST experiment data, a new hierarchical data storage system has been designed which includes 4 tiers including PCIe SSD cache, server local raid, Lustre storage, and NAS backup storage. The hot data can flow between tier1 and tier2/tier3/tier4 controlled by the schedule program. At present the primary schedule algorithm have been implemented and will be adopted in the next campaign of EAST tokamak.

In the future more PCIe SSD boards will be added into the cache system to promote the data access speed, and more advanced schedule algorithm will be implemented according to the requirements.

Acknowledgment

This work is supported by the National Magnetic Confinement Fusion Science Program of China (Grant No: 2014GB103000) and National Natural Science Foundation of China (Grant No: 11405205). We would like to thank all the EAST Team members for their support.

References

[1] B.J.Xiao, Q.P.Yuan, et al., “Recent plasma control progress on EAST”, Fusion Engineering and Design, Vol. 87, p1887-1890, 2012.

[2] Liu Ying, Li Guiming, Zhu Yingfei, Li Shi, “New developments of the EAST data system”, Fusion Engineering and Design Vol. 86, p151-154, 2011.

[3] Wang Feng, Li Guiming, Li Shi, Zhu Yingfei and Wang Yong, “A Continuous Data Acquisition System Based on CompactPCI for EAST Tokamak”, IEEE TRANSACTIONS ON NUCLEAR SCIENCE, Vol. 57, Issues 2, p. 669-672, 2010.

[4] Li Shi, Luo Jiarong, Wu Yichun, Li Guiming, Wang Feng, Member IEEE, Wang Yong, “Continuous and real-time data acquisition embedded system for EAST”, IEEE TRANSACTIONS ON NUCLEAR SCIENCE, Vol. 57, Issues 2, pp. 696-699, 2010.

[5] F. Yang, B. J. Xiao, “A web based MDSplus data analysis and visualization system for EAST”, Fusion Engineering and Design, Volume 87, Issue 12, p2161-2165, 2012.

[6] MDSplus, http://www.mdsplus.org

[7] F. Wang, S. Li, Y. Chen, F. Yang, “The design of data storage system based on Lustre for EAST”, Fusion Engineering and Design, April, 2016.

[8] InfiniBand, http://www.infinibandta.org

[9] Lustre, http://lustre.org

[10] Data Direct Networks (DDN), SFA7700 The Hybrid Flash Storage Appliance with Application Acceleration to Maximize Big Data Insights, 2014.

[11] Intel Corporation, https://www-ssl.intel.com/content/www/us/en/solid-state-drives/solid-state-drives-dc-p3700-series.html