Performance Measurement Technique of Cloud storage system

Qinlu He
Department of Computer Science
Northwestern Polytechnic University,
Xi'an China
luluhe8848@hotmail.com

Zhanhuai Li, Lexiao Wang, Huifeng Wang, Jian Sun
Department of Computer Science
Northwestern Polytechnic University,
Xi'an China

***Abstract*—Researches on technologies about testing aggregate bandwidth of file systems in cloud storage systems. Through the memory file system, network file system, parallel file system theory analysis, according to the cloud storage system polymerization bandwidth and concept, developed to cloud storage environment file system polymerization bandwidth test software called FSPoly. In this paper, use FSpoly to luster file system testing, find reasonable test methods, and then evaluations latest development in cloud storage system file system performance by using FSPoly.**

Keywords-cloud storage, aggregate bandwidth, file system, performance evaluation

I. INTRODUCTION

With the rapid development of network and computer technology, the stored data quantity is also showing a trend of explosive growth, the problem is how to effectively, reasonable storage of the data, but these data may be a field or a unit, the lifeblood of enterprises. In order to solve this problem, storage manufacturers are working to develop high availability, high reliability, mass storage system for a time the face of such a variety of storage products, storage data center users urgent need for a storage products performance and functional differences standards, thereby providing them with a reference index. But relative to the storage amount of product variety, various manufacturers with their own criteria to evaluate your product, it may eventually lead to storage manufacturers to provide consumers with standard are not comparable. Now on the international mass storage system evaluating standards and technology still lags behind relatively, the mass storage system complexity.

At present, the existing evaluation tools and methods can not fully meet the requirements of evaluation of mass storage system, evaluation methods and tools for now most are directed to single node, and for the mass storage system, a high performance, high reliability of the system, a single node load impossible to achieve the purpose of the test. In addition, with the storage system scale and continuously improve performance, didn't think you could become a bottleneck link to now have to be reconsidered. For example, the previous storage system used by the file system is seldom considered to become the bottleneck of the storage system, but now, for the mass storage system, due to the underlying disk performance, interface performance, the performance of network equipment, the controller performance is greatly improved, and the file system mechanism and design are essential to maintain the status quo, so now need to consider whether or not it may become the bottleneck of the whole storage system.

In order to realize the mass storage system of parallel file system testing, you need to use the test tool is initiated on the target system test; but the test tool load transmission capability must be able to more than the measured system throughput. Especially for the aggregate bandwidth performance indicators have reached tens of Gbps mass storage system, the file system to achieve the aggregate bandwidth evaluation, is necessary to build the cluster approach, in order to achieve the goal of parallel test system. At the same time, in order to play a single test node transmits the load capacity, also need to use the test tool can be in a single test node to create multiple load transmitting process, to reduce the test node number. The other hand, the file system being tested may be used for different types of business, which requires test tools can be set by a different load patterns in order to achieve different business types of the target system under test conditions, aggregate bandwidth performance.

Aggregate bandwidth refers to the storage of all nodes in the system instantaneous total bandwidth provided by and for external application system, is the visible data transmission rate. Aggregate bandwidth is defined here contains the application and actual file to read and write data volume, contains no metadata and network transmission of data packets containing datagram header and all kinds of control information. The index dimension is transmitted per second Gigabit bytes(GB/s).

The existing file system testing tool for IOZone, IOMeter, Postmark, or no requirements and extensible testing ability, or the test content and aggregate bandwidth indicators do not match. In this paper, for the realization of the file system aggregate bandwidth for the parallel test, developed the file system aggregate bandwidth test tool FSPoly.

II. FSPOLY OVERVIEW

FSPoly based on client / server structure, according to the different physical location, the whole is divided into two parts: the control end and the test end; its goal is either parallel test file system aggregate bandwidth, the number of concurrent connections and other performance indicators. In accordance with the loosely coupled design reasons, to the physical distribution and the role of different, FSPoly throughout the software design is divided into a

Published by Atlantis Press, Paris, France.

communication management module, the process management module, management module, the load generation process management module, management module, parameter setting of statistical information management module, management module results a total of seven parts.

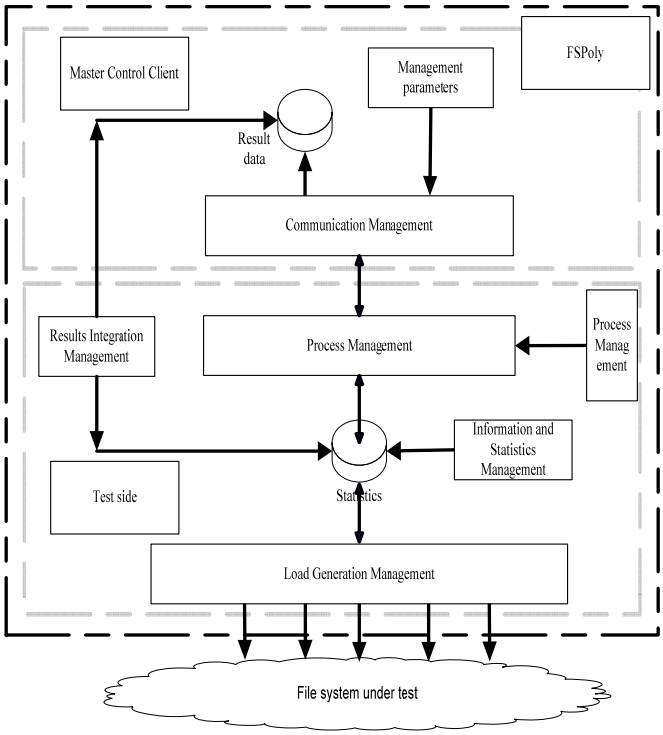

Figure 1. FSPoly structure

Figure 1 shows the entire FSPoly structure. On the whole it into total control end and a test terminal. Master control end by a single control nodes; the test ends by single or multiple initiated the actual load nodes. General control nodes running the master control program is responsible to the test terminal of each node sends the control command, access rules, control of the whole testing process; test end each node will run a" housekeeping process", this process is responsible for receiving control end command, and is responsible for the load transmitting processes throughout the life cycle, and send the results to the general control terminal.

III. LUSTRE PARALLEL FILE SYSTEM AGGREGATE BANDWIDTH TEST

This section will achieve the widely used Lustre parallel file system to be tested, the file system is widely used in storage system, support PB storage capacity. The latest version has support for IB network environment. This section of Lustre testing is also based on the IB network environment.

A. The test environment and related configuration

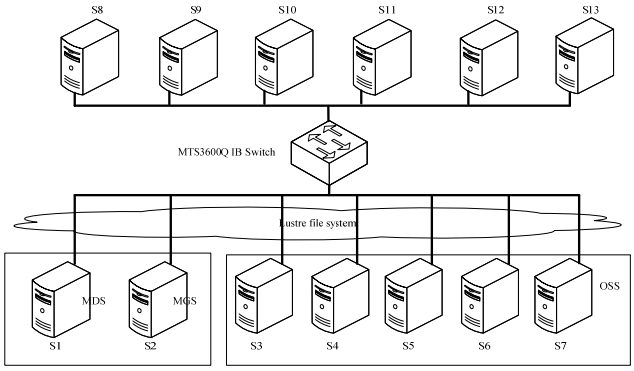

Figure 2 Lustre topology figure

(1)hardware configuration information. Models for Mellanox MTS3600Q, port check rate to 40Gbps InfiniBand switch a. The S1to S13Inspur AS300N storage server, its configuration information: data width of 64, frequency of 1333M HZ DDR316GB Dual Intel memory; Nehalm5502 Duo Memory dedicated processor (LGA 1366) two, frequency of 1.87G HZ,4MB Cache, alignment value is 64 bytes,133M HZ FSB; each server installation model of the MHQH19-XTC Mellanox ConnectX4X QDR InfiniBand HCA card a, one-way speed of 40Gbps; servers and switch between using the IB connecting line models for MCD4Q26C-007, the one-way speed of 40Gbps. Node S7for NFS server, its are stored using a3 hard disk (15.7k rpm,3.0Gbps SAS speed rate interface 300G Cheetah RAID0hard disk).

(2)software information. The entire test using a test tool FSPoly. FSPoly control end run on the installed Windows XP system of PC machine,5-15did not give this node; FSPoly test end run on S8to S13node. Each node have installed RedHat Enterprise Linux 5.3system, kernel for 2.6.18-128.el5xen; and OFED-1.5.1IB driver.

(3) S1 to S7Lustre-1.8.2 was installed on. Where S1 is a metadata server, the rear end is stored as 4disk consisting of RAID0; S2metadata management server, the rear end is stored as 3disk consisting of RAID0; S3 to S7for object storage servers, the rear end is stored as 7disc RAID5. All hard disk parameters: speed 15.7KRPM, interface speed for 3.0Gbps SAS interface, the capacity of the 300GB Cheetah hard disk.

B. test method

The purpose of this section is to measure in the current circumstances, Lustre parallel file system bandwidth maximum polymerization. Access mode choice of sequential read and write sequence two, from the four aspects (Lustre own strip size, read and write the block size, the number of concurrent connections and client number) gradually measures the aggregate bandwidth optimal value. Specific test steps are as follows:

The first step uses a single client node, setting different Lustre strip size, measured the aggregate bandwidth to reach the optimal value of Lustre bandwidth size.

The second step of setting the aggregate bandwidth to achieve when the optimal values of the relevant parameters, gradually changed to read and write the block size, measured the aggregate bandwidth optimal read block size.

The third set of the above steps optimal values for the parameters measured, gradually increasing the number of concurrent connections, measured the aggregate bandwidth performance of the optimal number of concurrent connections of the parameters.

The fourth step in order to increase the number of Lustre client, measured the aggregate bandwidth index optimal when the client number, and finally obtain the environment Lustre parallel file system aggregate bandwidth optimal value.

C. Test results and analysis

Figure 3 shows the different parameters on Lustre file system aggregate bandwidth influence curve. As can be seen from the graph, the sequential read aggregate bandwidth than the writing sequence of aggregate bandwidth performance. And, for the sequential read operation, the client number on the aggregate bandwidth performance impact is obvious; the client number is 6, its value is reached, is about 1000MB / s.

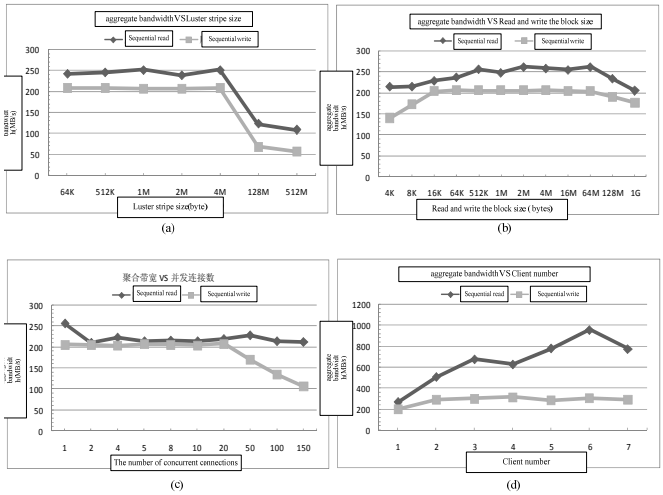

Figure 3 Lustre aggregate bandwidth test results

IV. MASS STORAGE SYSTEM AGGREGATE BANDWIDTH TEST

This section uses FSPoly on developing mass storage system for aggregate bandwidth test, and gives the test environment and test results.

A. Test environment and related configuration

As shown in Figure 4, the measured mass storage system internal topology. Map used in the back-end storage by five eight controller IB disk array is provided, each IB array configuration : the inside of a disk array controller and the disk frame body using dual port 4 channel SAS card (rate: 3.0Gbps) interconnection; array is used in hard disk for all 2TB size SATA hard disk, the rotating speed for 7200rpm, a five array is configured with 496 hard disks; use from 12 disc RAID0 configuration, and as the back-end storage is provided to each OSS and MDS; the server and between arrays using the IB network interconnection and hanging. Map used in the IB switch to the 324 port, the one-way speed of 40Gbps single port. Graph in which each server for all the wave of AS3000 storage server, wherein the 74 client memory size for all 1GB, all other server memory size for all 32GB. Each server and other related configuration: two CPU, each CPU Quad, which includes 8 CPU core; models for MHQH19-XTC HCA card a, the one-way speed of 40Gbps, for IB network. The server is used by the operating system kernel for CentOS5.4,: 2.6.18-164. The entire storage system using the file system is the latest development of the CapFS parallel file system.

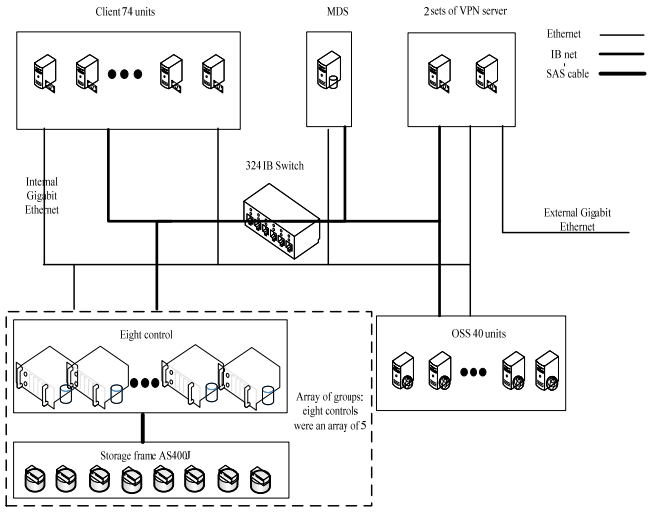

Figure 4 the measured mass storage system topology diagram

B. Test Results

Use the test tool FSPoly, from the 74 client to the tested storage system initiated polymerization of bandwidth testing. Test parameter is set to 2GB: file size (memory for a client two times), for each client to create 3 processes, to read and write the block size is set to 4KB, 512KB, 1MB, 2MB, the sequential read test. The test results as shown in figure 5.

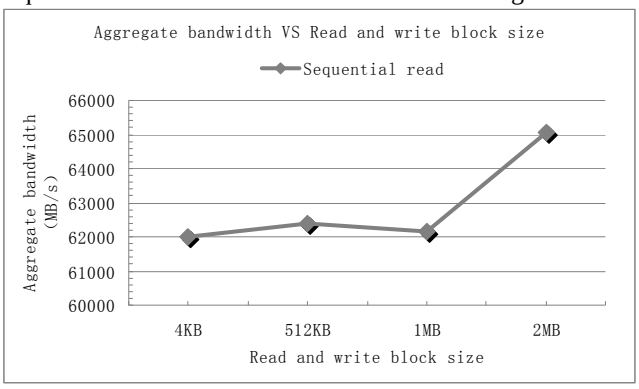

Figure 5 Mass Storage System aggregate bandwidth test results

As shown in Figure 14, storage system aggregate bandwidth with read and write block size has a significant change, in a data block size 4K ~ 1M change trend is very small, and where the data block for the 1M ~ 2M appear obvious upward trend, the measured mass storage system

aggregate bandwidth in the setting of access rules under the conditions of polymerization, storage system the bandwidth can be achieved in 65000MB / s.

V. CONCLUSIONS AND FUTURE WORK

This paper introduces the latest development by the aggregate bandwidth test software FSPoly began the effectiveness of the verification FSPoly on the basis of test results, based on the parameters set in turn proved to find the optimal value of the aggregate bandwidth of the test method is reasonable. Ultimately for the latest mass storage system developed for evaluating aggregate bandwidth.

At present a variety of parallel file system, the manufacturers have their own standards and preferences, while the parallel file system, and there is no uniform standard of evaluation, evaluation technology is lagging behind the benchmark. The complexity of parallel file system, leading to the job more difficult. Therefore, a parallel file system benchmark in all aspects of the evaluation study, was extremely urgent.

ACKNOWLEDGMENTS

This piece of work was Supported by the NPU Fundamental Research Foundation under Grant No.JC20110227,the Industry Innovation Foundation under Grant No.2011BAH04B05 and the National Natural Science Foundation of China under Grant No.61033007. Many people contributed ideas to the design of Fsploy, Most of all Mr Lexiao Wang Made a great contribution to Fsploy design. I would especially like to thank him who participated in the Design Team.

REFERENCES

[1] Hilbert M, Priscila López. The World's Technological Capacity to Store, Communicate, and Compute Information[J]. Science, 2011, 332 (6025): 60.

[2] Cisco System Inc. Cisco Global Cloud Index: Forecast and Methodology, 2010-2015[EB/OL]. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.pdf.2011.11.30.

[3] Nitin Agrawal, Andrea C. Arpaci-Dusseau and Remzi H. Arpaci-Dusseau. Generating Realistic Impressions for File-System Benchmarking. In FAST 2009.

[4] N.Agrawal, W.J.Bolosky, J.R.Douceur, and J.R.Lorch. A Five-Year Study of File-System Metadata. In FAST'07, San Jose, CA, February 2007.

[5] Joseph L N, Mohamed F M, David H C D. Pantheon : Exascale File System Search for Scientific Computing. In: Proceedings of the 23rd International Conference on Scientific and Statistical Database Management. Portland, 2011: 461-469

[6] SUNDARARAMAN, S., SUBRAMANIAN, S., RAJIMWALE, A., ARPACI-DUSSEAU, A. C., ARPACIDUSSEAU, R. H., AND SWIFT, M. M. Membrane: Operating system support for restartable file systems. In Proceedings of the USENIX Conference on File and Storage Technologies (FAST) (2010).

[7] O. Khan, R. Burns, J. S. Plank, and C. Huang. In search of I/O-optimal recovery from disk failures. In Workshop on Hot Topics in Storage Systems, 2011.

[8] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. J. Wainwright, and K. Ramchandran. Network coding for distributed storage systems. IEEE Trans. Inf. Theor., 56(9):4539–4551, September 2010.

[9] O. Riva, Q. Yin, D. Juric, E. Ucan, and T. Roscoe. Policy expressivity in the Anzere personal cloud. In 2nd ACM Symposium on Cloud Computing (SOCC), Cascais, Portugal, 2011.