The Performance Optimization of Lustre File System

Sun Jian ,Li Zhan-huai, Zhang Xiao
School of Computer Science
Northwestern Polytechnical University
Xi'an China
Qwert3277@163.com

Abstract—**Lustre has been widely used in mass storage systems. Its performance study also increasingly widely. During the analyzing of lustre system, this paper proposes the several important factors which can influence the system performance. From the test results we can find that: according to different Lustre application, network transmission medium, strip numbers and client numbers has a very high price for configuration.**

Keywords- Lustre; performance;optimization

I. INTRODUCTION

With the computer and related technology's rapid development, The development of high performance computer has effect from the scientific computing and the engineering calculation into commercial application and network information service areas. At the same time, the user requirement of mass data capacity and scalability improve, the network storage traditional architecture can not satisfy the need of data storage. Although NAS [1] (Network Attached Storage) is simple to operate and easy to manage, it have a limited capacity and poor scalability because of the structure of a single server. SAN [2] (Storage Area Network) system generally use the Fibre Channel (Fibre Channel) exchange equipment to connect Storage devices with application server. It has higher data transmission performance and better scalability, but it cannot share the resources between multiple platform because SAN only provides block level data service.

According to the shortcomings of the traditional network storage, the file system arises at the historic moment based on object storage framework. It combines the advantages of NAS and SAN, supports both direct access disk to improve performance and simplify the management with sharing files and metadata. At present, the mainstream of the file systems include Cluster File Systems Company Lustre [3], Panasas company Active Scale and the Chinese academy of sciences blue whale File system.

Lustre is one of the most influential and extensive application parallel file systems. Lustre is the open source, the parallel global based on object File system which is developed by the Cluster File Systems company (now be Sun acquisition). Lustre made optimization for read/write of the large files. It can provide of high performance of I/O throughput rate, global data sharing environment, independence of data storage location and redundance mechanism for the cluster system. The quickly recover service is good to meet the high performance computing cluster system requirement when the cluster's configuration is initialized or gateway failed.

In this paper, the impact elements of Lustre performance is measured, in the particular environment of network, the optimization method of Lustre system is provided by testing effect factors of the system performance. There is some reference significance for Lustre application in different environment.

II. LUSTRE FILE SYSTEM

Lustre has cluster storage architecture, it is based on open source Linux platforms (parallel) file system, provide the file system interface which is compatible with the POSIX. Lustre have two characteristics which are high scalability and high performance, support tens of thousands of client systems, PB level of storage capacity, hundreds of GB level I/O throughput. It is based on object storage technology, distributed management lock mechanism and store data of the separation of the storage solution. It provides a global namespace to avoid multiple data backup of the traditional distributed file system and concert cluster system workload.

Lustre file system is mainly composed of MDS (Metadata Server), OSS (Object Storage Servers) and the Client (Lustre Client)[4]. MDS stores the information of data description, manages namespace and target storage address. OSS provides the actual data storage and the I/O services for files; Lustre Client runs Lustre file system, and exchanges message with MDS and OSS. As shown in the figure below:

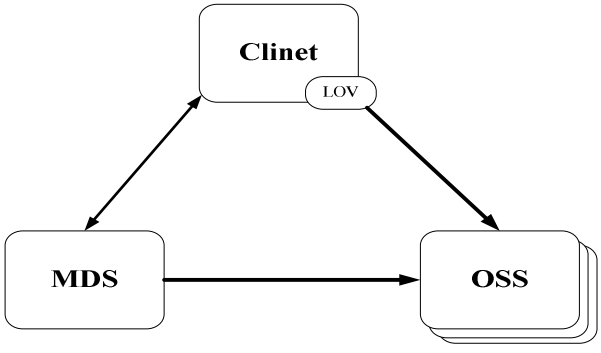

Figure 1. Architecture of Lustre

OSS (Object storage server) manages storage devices, these storage devices (Object Based Disk, OBD) can directly connect with the Disk (IDE and SCSI disks, etc) or one part of the SAN storage space[5]. OSS further classifieds management storage space volume or partition, considers the files in volume or partition as stored objects. So, we can consider the volume or partition as Object Storage Pool(Object Storage Target, OST in lustre). The Lustre file system files are stored in OST, so we can also consider OST as object storage equipment. One OSS can manage multiple OST, each OST derived a group of file object. A Lustre file system can manages hundreds of OSS. OSS and OST are not treated differently in general use. Lustre is a transparent global file system. Clients communicate with MDS and OOS through special object protocol in order to access the Lustre file. Lustre client software realizes standard file system interface, Lustre is a standard file system for client application or users, only the file it manages is not in the local but OST. In a single Lustre system, tens of thousands of the client can be installed.

III. PERFORMANCE OPTIMIZATION FOR LUSTRE

Through the above analysis, it is not difficult to find that Lustre file system performance could fully exertion mainly depends on three aspects: the transport network performance, Lustre file system itself and the client application configuration. In order to optimize Lustre and exertion it best performance, the above three aspects will be further analysis.

A. The Test Environment

On the test environment, Inspur AS300N servers are used to be test node and Lustre nodes. The specific configuration as follows: four core Intel (R) Xeon (R) E5502, 16 GB memory; Mellanox ConnectX 4X QDR InfiniBand network card or gigabit Ethernet card; Linux 2.6.18-164.11.1. El5_lustre. 1.8.2 operating system; LSI Logic/Symbios Logic Mega SAS RAID card; Cheetah 15.7 k 300 G hard disk.

The total control point runs Windows XP operating system, FSPoly[6] which is developed as polymerization bandwidth testing tool by northwestern polytechnical university. In order to avoid the influence of the file system cache, , we set the size of the file as 32GB for reading and writing which is double times larger than memory.

Lustre usually RAID (Redundant Array of Independent Disks) technology is used to protect the storage data prevent data missing caused by fault. the RAID card of test environment supports 12 hard drive hot swap, the organization RAID forms are RAID 0/1/5/6/10/30.During the test, RAID5 is chosen as data storage organization combining with the characteristics of Lustre and specific experimental environment.

B. Network Architecture Influences The Performance Of Lustre

Lustre improves the read and write performance of the whole system through multiple reading of OSS, however, if the network transmission performance is too low, the advantage of Lustre still can't be exertion. In setting up the environment, the gigabit Ethernet structure and InfiniBand architecture were respectively tested and the results are shown in the figure below:

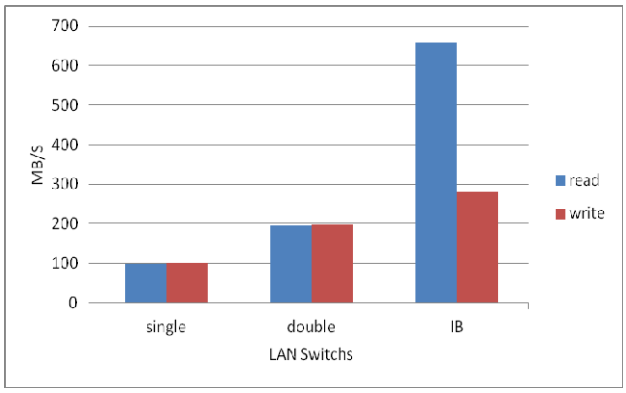

Figure 2. Network architecture influences theperformance of Lustre

The above single GbE test result is basic standard with the officials 110 MB/S of conclusion. From the above results we can be find that write and read performance of the system arise greatly with the improvement of network performance. The InfiniBand network architecture get the best result, the following performance tests are working in InfiniBand network.

C. Configuration Influences The Performance Of Lustre

The Lustre configuration involves two aspects: strip number (that is, the number of OST) and how to strip. The two aspects are key to realize Lustre concurrent I/O.

The strip number needs to be appointed according to the test configuration. In the specific process, there are three related factors: stripe count, stripe size and start-ost.

Stripe-count set to -1,that is Strips are division to all OST, ensure the maximization parallel I/O. Start-ost set to-1 which don't specify the start of the OST ,it ensure the OST load balance; Stripe-size configuration is according to specific test set purpose, the minimum is 64 KB, and it must be multiples of 64 KB.

1) Strip size influences the performance of Lustre:

In the test environment, five OSS are chosen and one OST mount on each OSS, the test results are as follows:

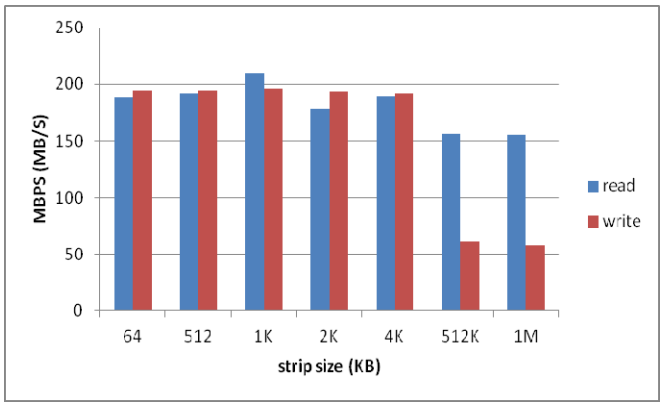

Figure 3. Strip size influences the performance of Lustre

From the results, the Lustre of small strip size could better than large on existing the performance of the Lustre. When the strip size is too large, more I/O works on one OST, it causes I/O waiting which affect system simultaneity.

2) Strip number influences the performance of Lustre:

Strip number is another important factor of the performance, According to above test results, we set strip size to 1MB and other test conditions remain unchanged, the experimental results under different Strip are as follows:

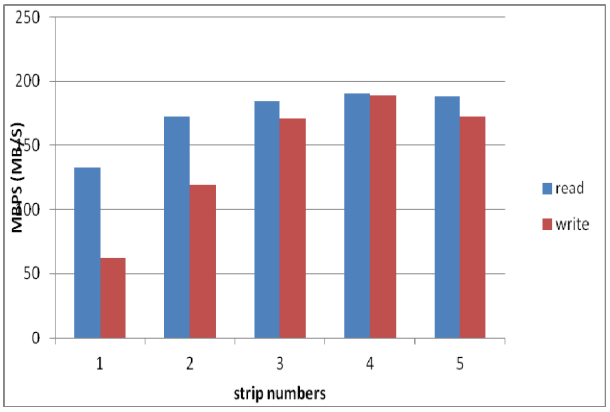

Figure 4. Strip number influences the performance of Lustre

When Strip number is 1, all of the data reading occur at one OST whitout paiallel working,it is lowest performance.With the increase in the number of OST, Lustre performance gradually enhanced, in the current environment, OST set to 5, we get the best effect.

3) Client configuration influences the performance of Lustre:

The client configuration influences the performance of the system mainly from three aspects: processes number of a single client, the size of R/w block , the client number.

a) Processes number of a single client influences the performance of Lustre

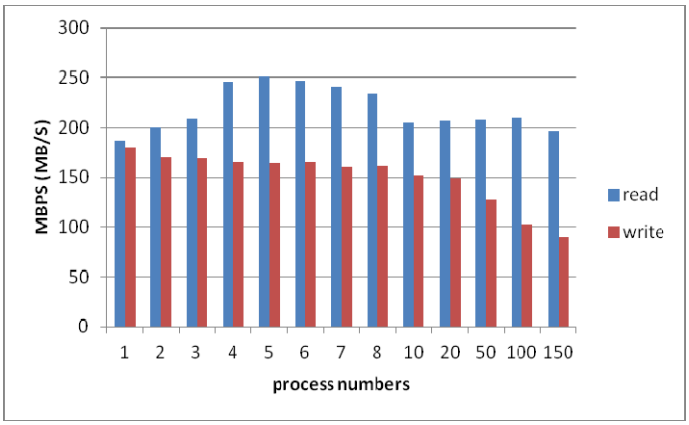

Figure 5. Processes number of a single client influences the performance

We can see from the above graph, reading and writing performance arises with the increase number of processes. When it add to a certain degree, the increasing number of processes, reduce the performance of the system.This means that process by the switch between process is waiting for started to influence the performance of the system. In the current environment, processes number set to 5, the system performance is the best.

b) The size of R/w block influences the performance of Lustre

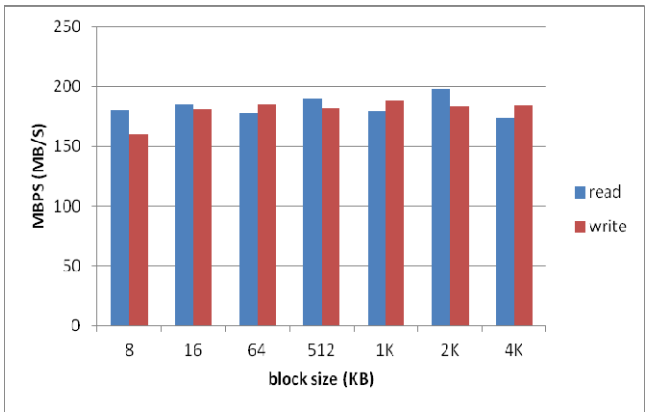

Figure 6. The size of R/w block influences the performance of Lustre

From the results we can see that performance is low when the size of R/w block is small, Because most of the reading and writing complete on OST without simultaneity. with the increase of reading and writing block siez, I/O simultaneity gradually rise, reading and writing performance increase slowly, when reading and writing block size in 64 KB ~ 4 MB, performance of reading and writing is in a stable state.

c) Client numbers influences the performance of Lustre

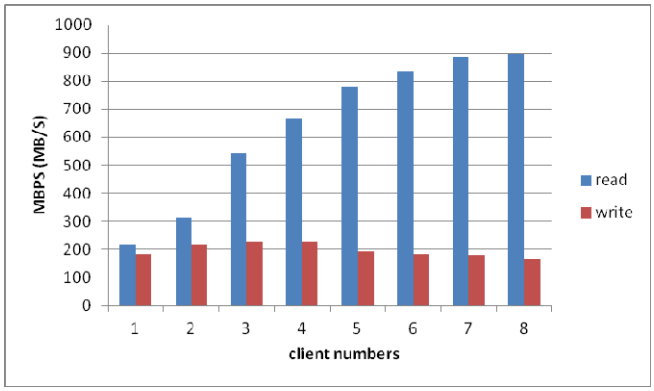

Figure 7. Client numbers influences the performance of Lustre

with the client increase, the system's read performance increase obviously, But the write performance does not obvious effect.

IV. CONCLUSION

InfiniBand network framework application expresses the advantage of Lustre file system. This means that high-speed network transfer rate is good for the advantage of Lustre file system. In the specific process of configuration, making storage server in load balance mode is the prerequisite to reach the highest performance of the system. Smaller size of strip

Lustre can get better polymerization system performance; The strip number also has greatly affect to system performance, if the back-end storage and network performance is not the bottleneck, The strip number has good expansibility for system performance. The increase of processes number is good to improve performance, but not the more the better, its size should be tested repeatedly based on the specific environment; the size of the block of reading and writing influence on the performance weakly. The client number is good to improvement the performance of the system which has good expansibility.

From the test results we can find that: according to different Lustre application, network transmission medium, strip numbers and client numbers has a very high price for configuration.

The development of mass storage pushes the application of Lustre file system, in cluster system, how to work with the bottom of the storage for optimization Lustre and how to explore Lustre file system organization of store data will be our future direction of the research.

ACKNOWLEDGMENT

The research work is supported by the National High Technology Research and Development Program of China(863 Program)(2009AA01A404) and the National Natural Science Foundation of China (Grant No.60970070), Low energy storage equipment development and industrialization

(2011BAH04B05).

REFERENCES

[1] David H.Gehring ,Beat Schilbach. Network Attached Storage ,2004.7.9

[2] Steve Chidlow,UNIVERSITY of LEEDS. JISC Technology and Standards Watch Report: Storage Area Networks , 2003.11

[3] Lustre http://wiki.lustre.org/images/0/09/821-0035_v1.3.pdf

[4] S. Cochrane, K. Kutzer, and L. Mcintosh, "Solving the HPC I/O Bottleneck: Sun™ Lustre™ Storage System, " Sun BluePrints™ Online, Sun Microsystems, 2009.

[5] L. Mandel, F. Plateau, and M. Pouzet. Lucy-n: a n-Synchronous Extension of Lustre. In Proc. of MPC, Québec, Canada, June 2010.

[6] Lexiao Wang, A Parallel Strategy for Measuring Network Bandwidth Cyber-Enabled2010