An Object-based Storage Controller Based on Switch Fabric

Shuibing He, Dan Feng

Key Laboratory of Data Storage Systems, Ministry of Education of China
School of Computer, Huazhong University of Science and Technology, Wuhan, China
E-mail: hesbingxq@163.com, dfeng@hust.edu.cn

Abstract

Object-based Storage Controller (OSC) plays a decisive role in the performance of the whole Object-based Storage Systems (OBSS). A new OSC based on switch fabric proposed in this paper. The new architecture aims to improve the OSC's disk I/O performance. Through parallel data transfer using two independent high-speed PCI-X buses attached to the switch fabric, the possibility that a single PCI bus will become a bottle neck is avoided. Theoretical analysis indicates that 24 SATA (serial ATA) disks can operate concurrently in this architecture at sustained rates of 65 MB/s. The traditional architecture would permit only 16 SATA devices.

1. Introduction

Object-Based Storage Systems (OBSS) has led to a new wave in the next storage technology [1].The OBSS consists of the Metadata Server (MDS), the Object-Based Storage Controllers (OSCs), and clients. The OSC includes CPU,memory, network interface and disk interface[2]. It takes charges of the intelligent object storage management, the heavy network communication, and secure access mechanism. The OSC is so important that it plays a decisive role in the performance of the whole OBSS.

A few researches on OSC have been done. IBM Haifa Labs implemented an OSC prototype: ObjectStone[3], to achieve high performance ,it runs on a Linux server. Though this OSC based on a server has high performance, it has an expensive cost. Contrastly, the traditional OSC usually is based on a PC platform. For example, the Object Storage Target (OST) in Lustre File system which acts as an OSC is implemented with general-purpose PC [4].The OSC in Panasas is StorageBlade[5],it uses 1.2GHz Intel Celeron CPU and 2 SATA disks as its hardware platform.

In the traditional OSCs based on PC platform, disk controllers, network interface controller, and other devices usually are attached to the single IO bus which normally is the PCI bus. As a result, if the number of the disks operating concurrently reaches to a limit, the PCI bus will become the performance bottleneck. To solve this problem, a new OSC architecture based on switch fabric is presented in this paper.

2. The Design of OSC Based on Switch Fabric

Figure 1 illustrates the new architecture of the OSC. The Intel 80314 is the interconnection core of the new OSC. The 80200 is the chief processing unit which has maximal frequency of 733 MHz. The DIMM SDRAM with the bandwidth 200*64Mbit/s is connected to the 80314 as system main memory. Two Gigabit Ethernet PHY Transceivers connect to the 80314's two integrated MAC ports. Connected to the 80314 by PCI-X bus, the four Intel 31244 serial ATA controllers realize the communication between host and disk storage.

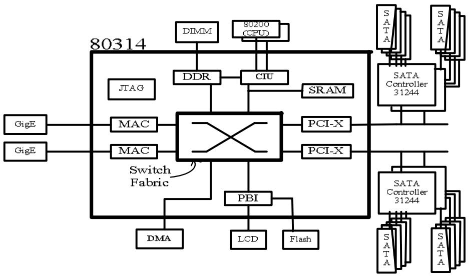

Figure 1. The new OSC architecture based on switch fabric

The significant characteristic of the 80314 is that it is designed as a fabric-centric, any-port-to-any-port bridge. It uses an internal switch fabric and supports concurrent transactions from any interface to any other interface [6]. Through parallel data transfer using two independent high-speed PCI-X buses attached to the switch fabric, we can improve the OSC disk I/O performance.

3. The I/O Performance Analysis

We use Object striping technology to enhance the disk performance and we only consider the course that data is transferred from disks to memory. In this instance, let t1,t2, t3, and t4 respectively denotes the disk service time, the time in which block is transferred from disk to 31244 SATA controller, from 31244 SATA controller to Intel 80314, from Intel 80314 to memory.

To get the best disk performance, we assume that all disks work with a sustained rate and all disks start their n^{th} stripe at the same time. We let N denotes the number of the Serial ATA disks attached to the PCI-X, L denotes the object length (bytes), and B denotes the stripe size (bytes), and object is striped across the N disks like RAID 0.

Now assume that T denotes the whole object transfer time, $R_{m,}$ $R_{p,}$ R_s and R_d denote the data transfer rate of memory, PCI-X, serial ATA bus and disk, and P denotes the number of the stripes in the object. Then t1 =B/R_d, t2 =B/R_s, t3 =B/R_p,t4 =$B/R_{m,}$ and P=$\left\lceil \frac{L}{B} \right\rceil$.

International Conference on Networking, Architecture, and Storage (NAS 2007)
0-7695-2927-5/07 $25.00 © 2007 IEEE

In the traditional OSC with only one PCI-X, the PCI-X bus will saturate if N*t3 is bigger than t1. Consequently, T changes according to N. Suppose M $=\left\lfloor \frac{t1}{t3} \right\rfloor = \left\lfloor \frac{R_P}{R_d} \right\rfloor$, therefore

1 if N$\leq$M, the utilization of PCI-X is smaller than 100%, with the increase of N, the utilization will be further improved .Now assume that $a_{tra}^{n}=\left\lceil \frac{P}{N} \right\rceil$ and b_{tra}^{n} =P-(a_{tra}^{n} -1)*N,

T $= a_{tra}^{n}$ *t1+t2+ b_{tra}^{n} *t3+t4 (1)

2 if N>M, PCI-X is fully used, adding more disks to the system has no help to reduce the object transfer time as the PCI-X become the new bottleneck of the I/O chain.

T= a_{tra}^{y} *t1+t2+ b_{tra}^{y} *t3+t4 (2)

In above formulas, $a_{tra}^{y}=\left\lceil \frac{P}{M} \right\rceil$ and b_{tra}^{y} =P-(a_{tra}^{y} -1)*M.

Assume that B=16KB, R_m=200*64Mbit/s,R_p=133*64Mbit/s,R_s=150MB/s, and R_d=65MB/s (the sustained data rate of a serial ATA disk) , then M=16. That’s to say, if N is bigger than 16, PCI-X will saturate and some disks will be idle in the transfer.

In the new OSC, we can further improve the performance using the parallel transfer on the two PCI-X buses. Since the maximum bandwidth of memory is 200*64Mbit/s, we can only make each PCI-X run at the rate of 100* 64Mbit/s. Here, suppose that N serial disks connect to each PCI-X bus.

According to the previous discussion, it is easy to evaluate the whole object transfer time as follows:

1 if N$\leq$M, assume that $a_{new}^{1n}=\left\lceil \frac{L}{2B*N} \right\rceil$, $b_{new}^{1n} = \left\lceil \frac{L}{2B} \right\rceil$ -(a_{new}^{1n} -1)*N, $a_{new}^{2n} = \left\lfloor \frac{L}{2B*N} \right\rfloor$, $b_{new}^{2n} = \left\lfloor \frac{L}{2B} \right\rfloor$ -(a_{new}^{2n} -1)*N, then :

T =max { a_{new}^{1n} *t1+t2+ b_{new}^{1n} *t3+t4, a_{new}^{2n} *t1+t2+ b_{new}^{2n} *t3+2t4} (3)

2 if N>M, we can get the object transfer time :

T =max { a_{new}^{1y} *t1+t2+ b_{new}^{1y} *t3+t4, a_{new}^{2y} *t1+t2+ b_{new}^{2y} *t3+2t4} (4)

In above formulas, $a_{new}^{1y}=\left\lceil \frac{L}{2B*M} \right\rceil$, $b_{new}^{1y} = \left\lceil \frac{L}{2B} \right\rceil$ -(a_{new}^{1y} -1)*M, $a_{new}^{2y} = \left\lfloor \frac{L}{2B*M} \right\rfloor$, $b_{new}^{2y} = \left\lfloor \frac{L}{2B} \right\rfloor$ - (a_{new}^{2y} -1)*M.

When R_p is 100*64Mbit/s, other data transfer rates and B are the same to that in traditional OSC, we can calculate that M=12 for each PCI-X. Namely, the new OSC supports parallel data transfer of 24 serial ATA disks.

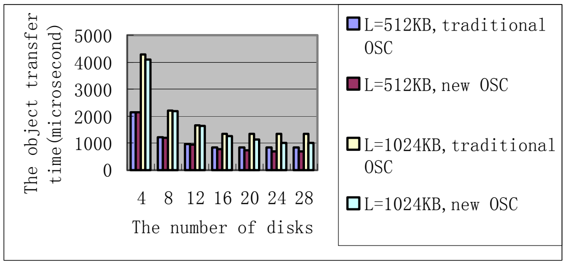

Figure 2. The comparison of the object transfer time

We can see from Figure 2 that the performance in the new OSC is better than that in traditional architecture. The advantage is obvious when N is between 16 and 24. The reduced object transfer time is maximal when N is 24. Furthermore, with the increase of L, more considerable time can be reduced.

4. Conclusions

In this paper, a new OSC which has the Intel 80314 companion chip with switch fabric as its core components is proposed. Through parallel data transfer using two independent high-speed PCI-X buses attached to the switch fabric, the possibility that a single PCI bus will become a bottle neck is avoided. Theoretical analysis indicates that 24 SATA (serial ATA) disks can operate concurrently in this architecture at sustained rates of 65 MB/s. The traditional architecture would permit only 16 SATA devices.

5. Acknowledgements

This work was supported by the National Basic Research Program of China (973 Program) under Grant No.2004CB318201, and the National Science Foundation of China No.60603048.

6. References

[1] M.Mesnier,G.R.Ganger,E.Riedel. Object-based Storage[J] .IEEE Communications Magazine,2003，Vol.41, Issue 8:84-90

[2] PANASAS WHITE PAPER: Object Storage Architecture: Defining a new generation of storage systems built on distributed, intelligent storage devices. 2003

[3] M. M. Factor, K. Meth, D. Naor, et al. Object Storage: The Future Building Block for Storage Systems. In: Proceedings of the 2nd International IEEE Symposium on Mass Storage Systems and Technologies. 2005. 119~123

[4] Peter J Braam. The Lustre Storage Architecture. Cluster File Systems, Inc. Whiter Paper. http://www.clusterfs.com. 2004

[5] Panasas Inc. Object Storage Architecture. White Paper. http://www.panasas.com/ objectbased_mgnt.html

[6] Intel Corp: Intel GW80314 I/O companion Chip Datasheet,2004