

Toward a Cost-effective Cloud Storage Service

Shin-gyu Kim, Hyuck Han, Hyeonsang Eom, Heon Y. Yeom

School of Computer Science and Engineering

Seoul National University, Seoul, Korea

{sgkim, hhyuck, hseom, yeom}@dcslab.snu.ac.kr

Abstract— Base platforms for many data-intensive applications start to move onto cloud computing services. These applications inherently requires very large storage space, and append-only distributed file systems have been developed for this purpose. These distributed file systems are capable of running on commodity hardware and highly scalable. Cloud computing service providers only try to reduce management costs, but not concerning about significant up-front investment costs. Additionally, virtual machine technologies, especially Xen that most of cloud computing services use, have low I/O performance. In this paper, we present a cost-effective cloud storage service model, which are built with old PCs but shows good performance. Our experimental evaluation shows 46% better performance in Postmark benchmark.

Keywords— Cloud computing, ATA over Ethernet, Virtual machine, Network file system, I/O performance

I. INTRODUCTION

Cloud computing is a new paradigm of computing that converts privately-owned computing resources into public utility computing services. Cloud computing offers more opportunity to creative and ambitious people by lowering up-front costs to start new businesses. Many cloud computing service providers have been continuously made efforts to cut down expenses of system maintenance by developing energy-efficient and cost-effective infrastructure and platform software. In spite of these efforts, the service charge for the cloud computing is still unaffordable to the people in the developing world. In addition to the technologies reducing the maintenance costs, it is necessary to reduce a significant up-front investment to build the cloud computing service.

One of the most successful applications of cloud computing is the analysis of very large data sets. Google's MapReduce [12] and Sawzall [15], Yahoo's Hadoop [5] and Pig [13], and Microsoft's Dryad [4], DryadLINQ [18] and SCOPE [9] have been developed for this purpose. These platforms run on user-level distributed file systems such as Google File System (GFS) [11] and Hadoop File System (HDFS) [5]. These file systems are append-only, partitioned in large blocks, and replicated over several nodes. Moreover, they are designed to be built with commodity hardware and

be highly scalable. In many sites, these file systems are installed on thousands of PCs.

Most of cloud computing services are based on virtual machine (VM) technologies: Xen [6] and VMware ESX server [17], etc. Virtualized environment has good features for building cost-effective cloud computing service: server consolidation, resource provisioning, fault isolation and so on. Storage services are also built upon VM technologies. However, recent research [8] on performance of virtualization technologies shows that many of them have low I/O performance, especially Xen that most of cloud computing services use.

Recent heavy software and hardware having plentiful resources make many PCs to be out of date, much of which are good to be thrown away. Newly introduced energy-efficient hardware is also making a lot of PCs to be old ones. To recycle these old PCs, some people proposed several ways to convert old PCs into firewalls, NAT servers or private file servers [14]. These ways are not suitable to recycle hundreds or thousands of old PCs, though may be enough for a handful of old PCs.

In this paper, we propose a way to build a cost-effective cloud storage service with old PCs. Our cloud storage service uses network-attached disks (NAD) instead of direct-attached disks (DAD). Because Xen has near-native network performance, network-attached disks show better performance than DAD in many cases. In our model, the network-attached disks are connected with compute nodes through ATA over Ethernet (AoE) [10] protocol. AoE is attractive protocol for old PCs due to its lightweight and simple structure. Because our storage service uses cheap old PCs and open source software, it greatly helps cloud service providers cut down significant up-front investment to build storage services.

II. RELATED WORK

A. Virtualization Technologies

Most of cloud computing services are based upon VM technologies that provide fault isolation, server consolidation and resource provisioning. VM technologies have largely three categories: full-virtualization, para-virtualization, and virtualization at the operating system level. Full-virtualization provide complete emulation of the underlying hardware. It includes QEMU [7], KVM [1], VirtualBox [16] and VMware [17]. Those solutions can theoretically support any guest

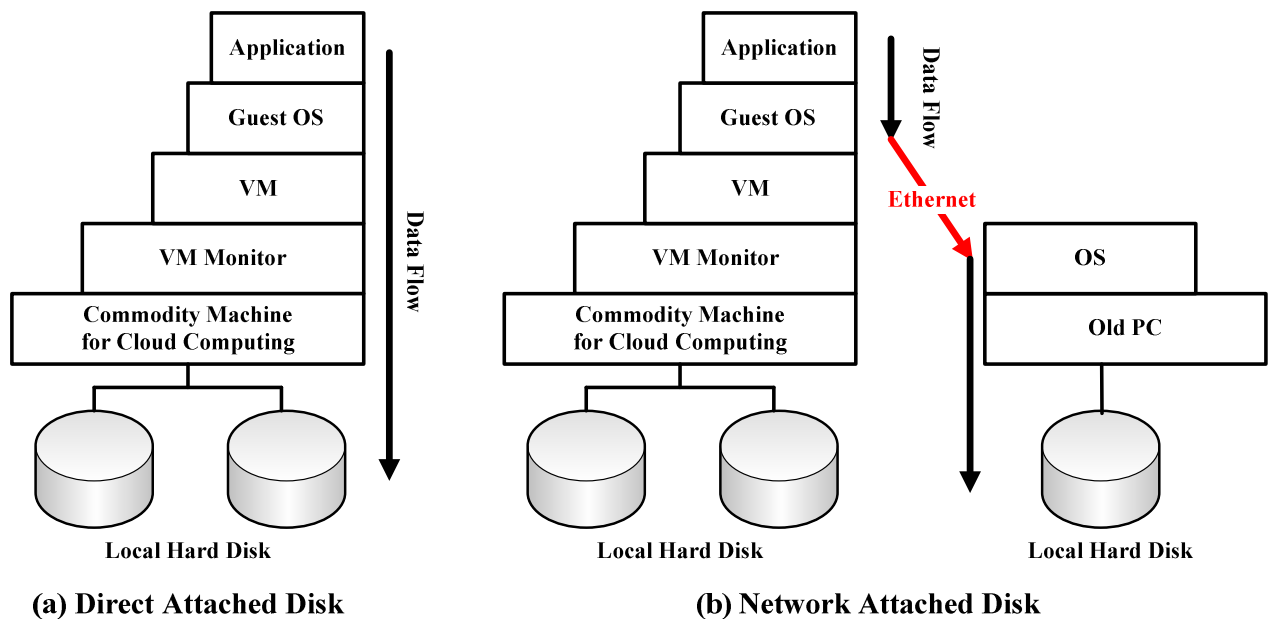


Figure 1. Two Approaches to I/O in Cloud Computing

operating system but are not efficient due to the hardware emulation.

The second VM technology, para-virtualization, is more efficient than full-virtualization technologies due to the collaboration of guest operating system with virtual machine monitor. This type of virtual machine is widely adopted to most of cloud computing services due to the good performance and moderate level of flexibility. Xen [6] is the most well-known para-virtualization solution.

The last is virtualization at the operating system level. In this kind of VM technology, guest operating systems replicate the kernel of host operating system, and there is only one kernel. This virtualization shows the best performance among three categories due to the absence of hypervisor overhead, but it lacks flexibility. Linux-VServer [2] and OpenVZ [3] are in this category.

B. ATA over Ethernet

ATA over Ethernet (AoE) [10] is one of network protocols to connect disks over network. AoE encapsulates Advanced Technology Attachment (ATA) commands inside Ethernet frames and lets them travel over an Ethernet network instead of a SATA or 40-pin ribbon cable. AoE is comprised of AoE target and AoE initiator. By using an AoE driver, AoE target looks like a block device in AoE initiator, and the AoE initiator is able to access a remote disk as if it were directly attached. AoE does not rely on network layers over Ethernet, such as TCP or IP. It is similar to Fibre Channel over Ethernet, but AoE is more lightweight due to the lack of TCP processing overhead. However, the non-routerability of AoE prevents connection over the Internet or IP networks. Thus, AoE is best suitable for old PCs within private networks. AoE target could be shared by multiple

AoE initiators, but AoE protocol does not provide any locking scheme.

III. TWO APPROACHES TO I/O IN CLOUD COMPUTING

The two classes of I/O configurations we consider in this paper are “Direct Attached Disk (DAD)” and “Network Attached Disk (NAD)” as shown Figure 1. The DAD configuration is commonly used in cloud computing services such as Amazon Elastic Compute Cloud (EC2). In the DAD configuration, an application generates I/O requests by using POSIX file system calls. Then, a guest OS that runs over a virtual machine performs file system calls and makes block device commands if necessary. The virtual machine receives the device commands and relays the commands to a virtual machine monitor. The virtual machine monitor sends the device commands to local hard disks. Even if the DAD configuration is commonly used, it is widely known that the I/O performance of DAD is pretty bad.

The NAD configuration is used for applications such as databases or file systems that require accesses to raw block level storage via commodity networks such as Ethernet. In the NAD configuration, an application also generates I/O requests by using POSIX file system calls. Then, a guest OS executes file system calls. The guest OS generates block device commands if necessary, and a specific block device sends the block device commands to an old PC. An OS that runs over the old PC receives the commands and sends the commands to its local hard disks. Amazon Elastic Block Storage (EBS) provides a storage service that is similar to NAD. Our study focuses on a cost effective and high-performance block storage service using old PCs that is connected by a commodity network.

In [8], the authors pointed out that the I/O performance of virtual machines is pretty bad. For example, performance of

Xen is less than 30% when Dbench is used for this benchmark test, while performance of Xen is close (97%) to native Linux when Netperf is used. These results imply that if disk I/O calls of Xen are translated into network I/O calls, the disk I/O performance of Xen can be greatly improved. Moreover, if old PCs of many enterprises or organizations are used for real I/O processing, remote storage devices that are composed of old PCs can be cost-effective.

Why not NFS? Some might doubt the effectiveness of the NAD configuration; there might be few differences between NFS and NAD. In fact, there are few differences in the application level viewpoint. However, NAD is independent of a file system semantic, while NFS has its own semantic that is close to but different from the POSIX semantic. Thus, NAD is particularly suited for applications such as databases or file systems that use the POSIX semantic strictly.

IV. EXPERIMENTAL EVALUATION

We evaluated two approaches to I/O by running Postmark benchmark. Postmark was performed on two PCs. One is equipped with dual Intel Xeon quad-core CPU 2.0GHz and 4 GB RAM, and the other with dual AMD Opteron dual-core CPU 1.8GHz and 8 GB RAM. The two PCs have same model of disks and runs Linux with 2.6.24 kernel. Benchmark program was run in Xen 3.2 at Intel machine, and the size of RAM is limited to 1 GB to clarify the difference in performance between DaD and AoE-attached disk. The two PCs are connected through a switched 1 Gbps Ethernet LAN. We used the vBlade 2.0 software for AoE target that runs in native Linux environment.

Postmark is designed to evaluate the performance of file servers for applications such as e-mail, netnews, and web-based commerce. This kind of workload consists of enormous numbers of relatively short-lived files. Thus, Postmark shows the performance for random and small I/O. In this experiment, the number of subdirectories was 100, and other parameters were remained unchanged. Nine experiments were performed by increasing the number of files and the number of transactions.

Table 1 shows the results of Postmark benchmark for DaD and AoE-attached disks. Overall, AoE-attached disk outperforms DaD by 46.1% in average. This result means that AoE-attached disk has better performance than DaD.

In [8], the authors showed an performance of I/O operations of Xen without physically stressing DaD. The performance was 30% of native Linux, and it degraded to 20% of native Linux with physical access to DaD. However, the network performance of Xen was near the performance of native Linux. The reason for good result of AoE-attached disk comes from the fact that the overhead of accessing DaD was replaced by the network operation.

V. CONCLUSION

We presented a way to build a cost-effective cloud storage service with old PCs. Our storage model uses AoE-attached disk instead of direct attached disk, and reduces significant up-front investment costs to build cloud storage services. We

Table 1. Results of Postmark Benchmark

(a) Number of Files = 100000

	Number of Transactions		
	30000	60000	90000
DaD (sec)	147.4	247.2	448.8
AoE (sec)	107.6	127.0	262.4

(b) Number of Files = 200000

	Number of Transactions		
	30000	60000	90000
DaD (sec)	350.0	678.4	988.2
AoE (sec)	188.2	266.6	487.8

(c) Number of Files = 300000

	Number of Transactions		
	30000	60000	90000
DaD (sec)	541.0	898.6	1262.6
AoE (sec)	292.6	492.2	641.8

showed that AoE-attached disk has better performance than DaD by Postmark benchmark. AoE-attached disk outperforms DaD by 46.1% in average. Currently, our model does not concern about energy consumption. This issue is remained as our future work.

ACKNOWLEDGMENT

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency (NIPA-2009-C1090-0902-0006)).

REFERENCES

- [1] Kvm (kernel-based virtual machine). <http://www.linux-kvm.org>.
- [2] Linux-vserver. <http://linux-vserver.org/>.
- [3] Openvz. <http://www.openvz.org>.
- [4] Dryad: distributed data-parallel programs from sequential building blocks. In *EuroSys '07: Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007*, 2007.
- [5] Apache. *Hadoop*. <http://hadoop.apache.org/>.
- [6] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield. Xen and the art of virtualization. In *SOSP '03: Proceedings of the nineteenth ACM symposium on Operating systems principles*, pages 164–177, New York, NY, USA, 2003. ACM.
- [7] F. Bellard. Qemu, a fast and portable dynamic translator. In *ATEC '05: Proceedings of the annual conference on USENIX Annual Technical Conference*, 2005.
- [8] F. L. Camargos, G. Girard, and B. des Ligneris. Virtualization of linux servers: a comparative study. In *Linux Symposium*, 2008.
- [9] R. Chaiken, B. Jenkins, P.-A. Larson, B. Ramsey, D. Shakib, S. Weaver, and J. Zhou. Scope: easy and efficient parallel processing of massive data sets. *PVLDB*, 1(2):1265–1276, 2008.
- [10] Coraid. ATA over Ethernet. <http://www.coraid.com/TECHNOLOGY/AoE-Protocol-Definition>.
- [11] S. Ghemawat, H. Gobioff, and S.-T. Leung. The google file system. In *SOSP*, 2003.

- [12] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, 2004.
- [13] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. Pig latin: a not-so-foreign language for data processing. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, New York, NY, USA, 2008. ACM.
- [14] pcmag.com. 10 things to do with old pcs. <http://www.pcmag.com/article2/0,2817,1744926,00.asp>.
- [15] Rob Pike and Sean Dorward and Robert Griesemer and Sean Quinlan. Interpreting the data: Parallel analysis with sawzall. *Scientific Programming Journal*, 2005.
- [16] I. Sun Microsystems. Virtualbox. <http://www.virtualbox.org>.
- [17] VMware. VMware esx server. <http://www.vmware.com/products/esx/>.
- [18] Y. Yu, M. Isard, D. Fetterly, M. Budiu, U. Erlingsson, P. K. Gunda, and J. Currey. DryadLINQ: A system for general-purpose distributed data-parallel computing using a high-level language. In *OSDI*, 2008.