Storage access optimization with virtual machine migration during execution of parallel data processing on a virtual machine PC cluster

Shiori Toyoshima
Ochanomizu University
2–1–1, Otsuka, Bunkyo-ku
Tokyo 112–8610, JAPAN
shiori@ogl.is.ocha.ac.jp

Saneyasu Yamaguchi
Kogakuin University
1–24–2, Nishishinjuku, Shinjuku
Tokyo 163–8677, JAPAN
sane@cc.kogakuin.ac.jp

Masato Oguchi
Ochanomizu University
2–1–1, Otsuka, Bunkyo-ku
Tokyo 112–8610, JAPAN
oguchi@computer.org

***Abstract*—We have constructed a virtual machine PC cluster that uses a virtual machine as worker nodes, and proposed a method that acquires insufficient resources dynamically from cloud computing systems while basic computation is performed on its own local clusters. Virtual machine environments enable us to manage computer resources flexibly and make use of migration. For data access, we have used iSCSI protocol which supports to access data through IP networks, and migrated virtual machine to cloud where data is stored over a high latency network. We have confirmed that the execution time becomes shorter with the migration of virtual machines even the cost of migration is taken into account, compared with the case of accessing data over a network when an I/O-intensive application is executed.**

***Keywords*-virtual machine, cloud computing, iSCSI, remote storage access, data-intensive appllication**

I. Introduction

In recent years, because of the increasing the amount of available information by means of broadband networks, IT cost has become a major problem. For this situation, cloud computing is highly expected.

Cloud computing enables to construct a system scalably according to resource usage. User can use resource as needed and it is expected to decline introduction and management cost. However, it is considered to be difficult to shift to cloud resource suddenly for who already has own cluster system or placing most systems on its own cluster even considering advantages of cloud computing. In this paper, we have constructed a virtual machine PC cluster that uses a virtual machine as worker nodes, and proposed a method that acquires insufficient resources dynamically from cloud computing systems, while basic computation is performed on its own local clusters monitoring system usage of cluster.

We have constructed a virtual machine PC cluster as a local cluster. Because we have used virtual machine as worker nodes, the migration mechanism can be introduced and it is expected to be able to use limited resources flexibility.

In this study, since we mainly assume to execute data-intensive application, we focus on storage access of the system.

II. Virtual machine PC cluster

A. Virtual machine

In recent years, one of serious problem in information system is that too many servers are introduced in a single site. It is hard to coexist some applications that process other jobs on the same machine, because it is sometimes concerned to be unstable in such a case.

As a result, when a new service is introduced, the system is expanded and new server machines should be set up, and a management procedure must be taken for each environment. In order to reduce such an introduction cost, server virtualization is effective, in which a system can be constructed as multiple computers that operate virtually on a single server machine.

Virtualization software enable us to install guest OS on top of host OS. Therefore, they have a problem of degradation of processing performance compared with the case of real machine because guest OS works as an application on host OS.

Xen [1] provides a basic platform of virtualization, on which multiple OSes operate as shown in Figure 1. Guest OS can basically access to computer resources almost directly without mediation of Host OS. Since the overhead of virtual machine is reduced, Xen can achieve higher performance which is close to the case of real machine.

Xen is used even in a business field recently, because Xen achieves remarkably high performance as open-source software. In the architecture of Xen, Virtual Machine Monitor is foundation for virtualization and virtual machines called domain are allocated on top of it. Domain0 behaves as a host OS and DomainU behaves as guest OS. Domain0 has a privilege to access physical hardware resources and to manage other domains.

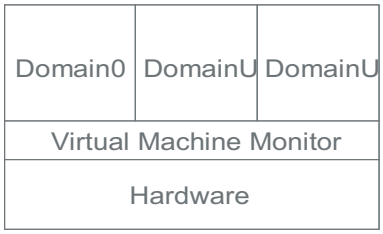

Figure 1. Architecture of Xen

B. Virtual machine PC cluster

In our cluster system, we have incorporated virtual machine in worker nodes of PC cluster and constructed a virtual machine PC cluster. As we use virtual machine as worker nodes, the migration mechanism can be introduced that migrate virtual machine to another node while maintaining the state of running applications. In addition, flexible management of infrastructure can be introduced regarding a virtual machine as a unit of resource depending on system load and service demanding.

III. Resource management

A. Resource usage from remote site

In cloud computing, software and hardware can be used as a service across the Internet without being conscious its existence nor inner structure. And HaaS model is known in computing resouce.

Cloud computing is expected not only reduce the introduction and management cost of information systems but also increase and decrease the capacity according to the system situation when it is difficult to predict system scale in advance.

B. IP-SAN

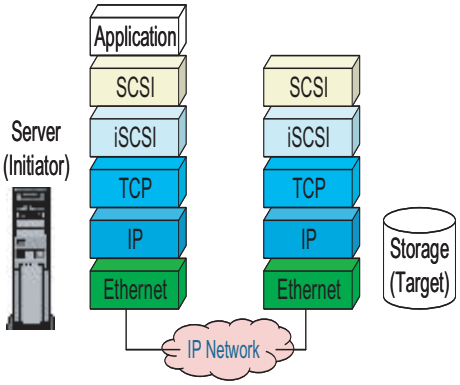

Figure 2. Configuration of iSCSI

We have used Storage area network (SAN) for storage network. SAN unifies the distributed storage held by each node, and realizes efficient practical use with central control for disk resources. IP-SAN is expected as SAN of the next generation built with IP network.

Internet Small Computer System Interface(iSCSI) [2] is the most popular protocol of IP-SAN and we can build SAN with inexpensive Ethernet and TCP/IP. In addition, iSCSI is expected for the realization of the long-distance remote access because IP network infrastructure is widely deployed and maintained in wide area networks. Therefor, it is expected to use cloud computing framework as outsourcing of computational resources, not only remote storage like data center. Figure 2 shows the layered structure of iSCSI. iSCSI encapsulates a SCSI command within a TCP/IP packet and transmits the volume of data between server (Initiator) and storage (Target). IP-SAN is expected to be used not only remote storage like data center but also outsourcing computer resources in a cloud computing framework.

C. PC cluster consolidated with IP-SAN

PC cluster consolidated with IP-SAN is introduced in [4] that consolidates back-end SAN between a node(server) and storage and front-end LAN among nodes by using iSCSI. In the case of PC cluster consolidated with IP-SAN, both the back-end SAN and front-end LAN can be unified into a single commoditized network built with TCP/IP and Ethernet by using iSCSI, as shown in Figure 3. Therefore, the reduction of network construction cost and the increase in efficiency of operational management can be achieved. However, since both back-end and front-end networks use the same network resources, it is concerned the communication packets transmitted between nodes collide with packets of storage access on the same network, and performance deteriorates as a result.

According to the result of above experiments, total performance of the system becomes CPU-bound or I/O-bound, not network-bound in these cases even though iSCSI network is consolidated.

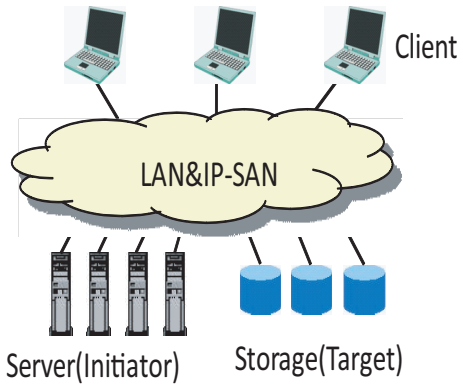

Figure 3. PC cluster consolidated with IP-SAN

IV. Basic evaluations

When users consider to employ cloud computing who possess their own cluster, they probably use the function of storage within the cloud at first, which is equal to remote backup that becomes common recently. Next, users may advance to the stage in which remote storage is accessed directly by servers, and finally to the stage servers perform calculation processing also in cloud from using cloud as remote backup.

In [3], we have constructed a system which use storage function from remote site over high-latency networks in order to examine the behavior of a virtual machine PC cluster when it uses remote resources. For a comparison with the case of using remote storage, we constructed a virtual machine PC cluster which uses storage in a local site during the execution.

Two types of applications were executed on the experiment platform. They are data mining Hash Partitioned Apriori (HPA) and database benchmark Open Source Development Labs Database Test3 (OSDL-DBT3) [6]. HPA is a data mining application that processes massive transaction data, its main execution is calculating process. Therefore CPU load is a bottleneck of execution and HPA is a CPU-bound application. On the other hand, OSDL-DBT3 that is simplified from Transaction Processing Council Benchmark-H (TPC-H) [7] benchmark simulates a decision support system, and this is an I/O-bound application that inserts and deletes data to databases and queries are executed repeatedly.

We have inserted RTT between local site and remote site constructing remote access environment intended cloud computing. In the case of executing HPA, there is only little processing time difference between environments of using local storage and accessing remote storage. In OSDL-DBT3, on the other hand, we have confirmed that application execution time becomes longer when RTT between local site and remote site simulating cloud computing is longer to access remote storage.

According to these experiments, data-mining that is not I/O-bound like HPA can be executed with sufficiently practical performance even when the data is stored at a remote site. In contrast, in the case of I/O-intensive applications like OSDL-DBT3, we have confirmed remarkable performance decline when a virtual machine PC cluster uses storage at a remote site.

Thus in this study, we propose a technique to migrate virtual machine to a remote site that stores data if remote access cost is high, in order to achieve load balancing and optimization of storage access instead of iSCSI remote access.

Table I
Experimental setup : PCs

OS	initiator : Linux 2.6.18-53.1.14.el5(CentOS5.3)
CPU	initiator : Intel (R) Xeon(TM) 3.6GHz target : Intel (R) Xeon(TM) 3.6GHz
Main Memory	initiator(Domain0) : 2GB initiator(DomainU) : 2GB target : 4GB
iSCSI	initiator : iscsi-initiator-utils target : iSCSI-Enterprise-Target
Monitoring Tool	Ganglia

V. Experimental result and discussion

A. Experimental setup

Specification of each node of the cluster is shown in Table 1. To construct a remote access environment, we have inserted Dummynet [8] that simulates delay artificially between a local site and a remote site. A virtual machine(DomainU) has been created for each worker nodes. We have monitored a virtual machine PC cluster using a monitoring tool Ganglia [9] including iSCSI communication to access remote storage.

B. Experiment 1: Using local servers only

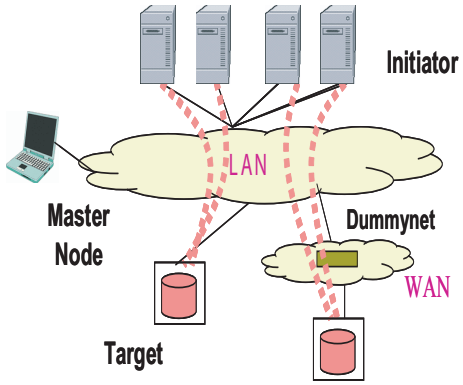

Figure 4. Experiment 1 : Experimental environment

In experiment 1, we have analyzed the behavior of a virtual machine PC cluster executing an I/O-bound application, which uses only local servers while accessing remote storage directly by the servers. We have built a virtual machine PC cluster like Figure 4.

We have compared the performance of two cases in which job is given to Domain0 and DomainU, respectively. For storage access we have used iSCSI. Two servers in local site access to local storage and other two servers access to remote storage. Assuming a remote access environment, we have inserted delay by Dummynet as RTT of remote iSCSI access is 1msec, 2msec, 4msec, 8msec, 16msec.

The execution time of OSDL-DBT3 is shown in Figure 5. According to the graph of the experiment 1, execution time is

increasing as RTT becomes longer from 1msec to 16msec in iSCSI access. Especially, severe increase in execution time is observed when delay is longer than 4msec. On the contrary, Figure 6 shows execution time of HPA with the data size of 20 megabytes which is used in the previous experiment [3].

HPA is a parallel application that parallelizes association rule mining which is based on Apriori algorithm using a hash function. Although HPA is data mining that processes huge amount of transaction data, since it includes heavy CPU processing, it is not I/O-bound. Therefore the difference of execution time is small even in the case of longer delay (Figure 6). On the contrary, OSDL-DBT3 is an I/O-intensive application that executes continuous access to database. Remote storage access has caused I/O-bound in the application execution, and therefore significant execution time difference was observed as the delay becomes longer (Figure 5).

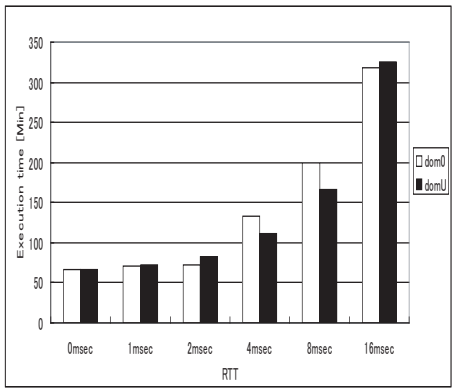

Figure 5. Experiment 1 : Execution time of OSDL-DBT3

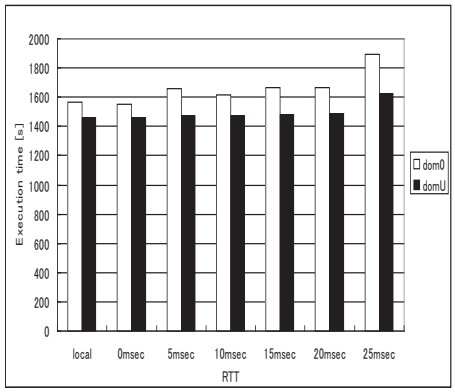

Figure 6. Experiment 1 : Execution time of HPA

Figure 7 and 8 show CPU utilization and Figure 10 and 11 show memory utilization of iSCSI Initiator and Target when OSDL-DBT3 is executed. Figure 9 shows CPU utilization of HPA. Monitoring results in HPA show that almost 100 % CPU is used. Compared with the result of HPA, when OSDL-DBT3 is executed, CPU utilization still has a margin.

In memory utilization, when executing OSDL-DBT3, almost 100 % was consumed in Initiator and Target including cache. From this monitoring result, Initiator uses Target memory as a cache space as well as its own disk cache.

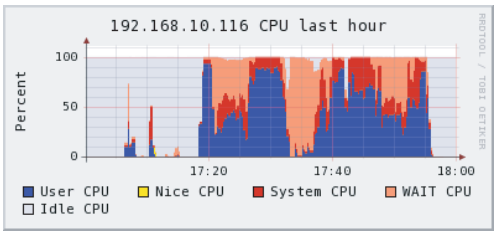

Figure 7. CPU utilization of initiator (OSDL-DBT3)

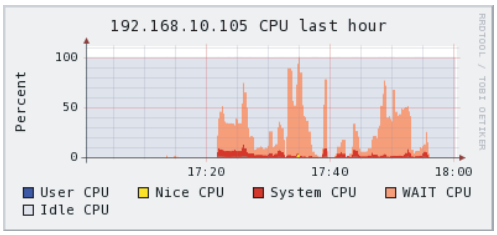

Figure 8. CPU utilization of target (OSDL-DBT3)

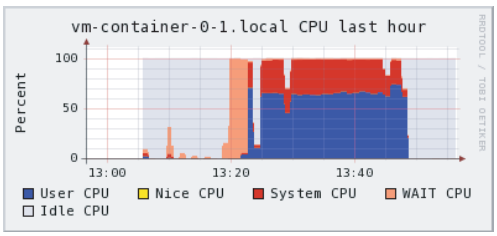

Figure 9. CPU utilization of initiator (HPA)

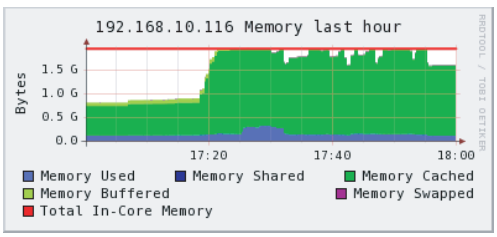

Figure 10. Memory utilization of initiator (OSDL-DBT3)

C. Experiment 2: experiment including server migration

The results of experiment 1 have shown that performance deteriorates when accessing remote storage over high letency network when an I/O-intensive application is executed. Thus

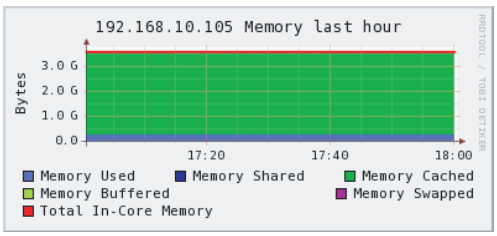

Figure 11. Memory utilization of target (OSDL-DBT3)

we migrate a local virtual machine to a remote site where data is stored. In experiment 2, we have built a virtual machine PC cluster as shown in Figure 12 and 13 that contains six servers (Initiator) and two iSCSI storage(Target). Four servers and a storage are located in the local site, two servers and a storage are located in the remote site. Assuming remote iSCSI access, we have inserted 1msec, 2msec, 4msec, 8msec, 16msec RTT by Dummynet between the local site and the remote site as the same with experiment 1 (Figure 13).

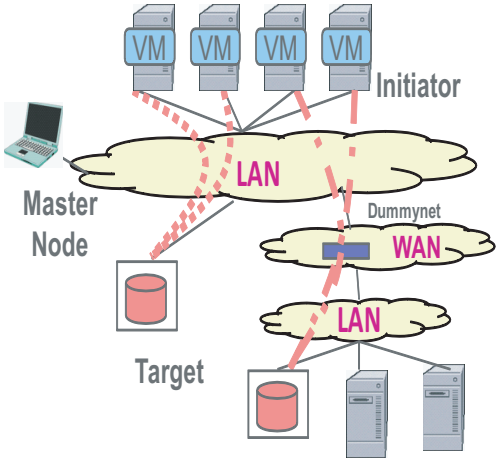

Figure 12. Experiment 2: Experimental environment of before migration

Because we have observed in the result of experiment 1 that performance to access remote storage from a local server degrades when we execute an I/O-intensive application, we migrate the local virtual machine to the remote site so that it executes the application and accesses storage on the remote site where data exist.

First, Figure 14 shows migration time of a virtual machine from a local site to a remote site on each RTT. The result is 21 seconds when RTT is from 0msec to 4msec, and 52 seconds when RTT is 16msec that is the longest RTT measured in this study.

Figure 15 shows total amount of time that sums up migration time to migrate a virtual machine to a remote site and execution time of OSDL-DBT3 at the remote site. Figure 15 includes DomainU execution time of experiment 1 for

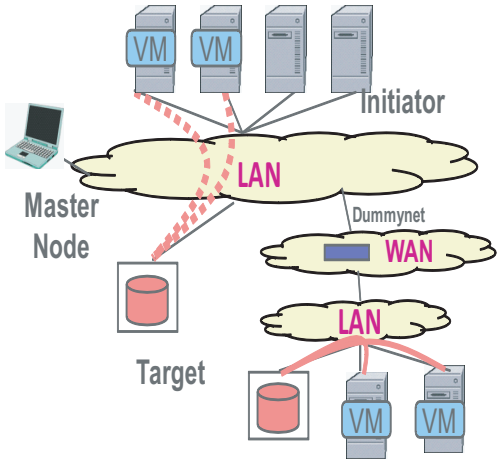

Figure 13. Experiment 2: Experimental environment of after migration

comparison. From this figure, the execution of experiment 2 is faster than that of experiment 1 when RTT is long. We can execute an application without delay on iSCSI by means of migration of virtual machines to a remote site. Therefore, as RTT becomes longer, it is effective to migrate a server to a remote site where storage is located.

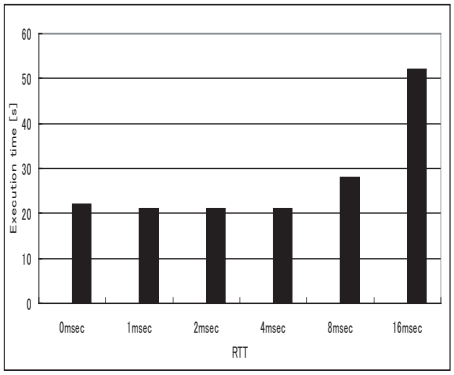

Figure 14. Experiment 2: Migration time

VI. Conclusion

In this study, we have executed database benchmark OSDL-DBT3 which is considered to perform frequent I/O access and analyzed the behavior of a virtual machine PC cluster including iSCSI remote access. In experiment 1, we have experimented that servers access to remote storage directly during runtime. We have confirmed as RTT is longer the execution time becomes longer in iSCSI remote access because remote storage access. Since remote access is bottleneck of execution of the application, in experiment 2, we have migrated a virtual machine on a local site to a remote site and OSDL-DBT3 is executed with storage access and directly in the remote site. As a result, total amount of

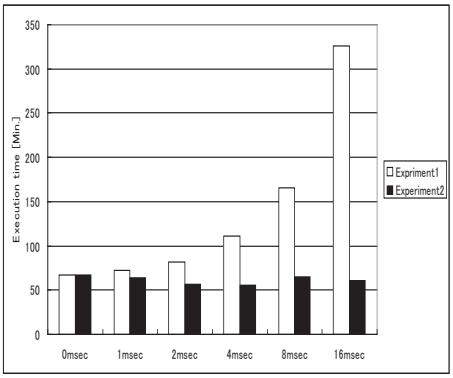

Figure 15. Experiment 2: Execution time

migration and execution time is shorter than experiment 1 as RTT is longer. We have confirmed that our method is effective to relocate a local server to a remote site when it takes longer to access to a remote storage.

As a future work, we will build a system of load balancing dynamically if the load is heavy which migrates virtual machine to a remote site automatically, and analyze its behavior.

REFERENCES

[1] Xen : http://www.xen.org/

[2] iSCSI RFC: http://www.ietf.org/rfc/rfc3722.txt

[3] Shiori Toyoshima, Saneyasu Yamaguchi, Masato Oguchi: ”Analyzing performance of storage access optimization with virtual machine migration,” CPSY, Vol.109, No.296, CPSY2009-37, pp.13-18, Kyoto, Nobember 2009.

[4] Asuka Hara, Kikuko Kamisaka, Saneyasu Yamaguchi, and Masato Oguchi: ”Analyzing Characteristics of PC Cluster Consolidated with IP-SAN using Data-Intensive Applications,” In Proc. 20th IASTED International Conference on Parallel and Distributed Computing and Systems (PDCS2008), No.631-042, November 2008.

[5] Masato Oguchi and Masaru Kitsuregawa:”Using Available Remote Memory Dynamically for Parallels Data Mining Application on ATM-Connected PC Cluster”,14th IEEE International Parallel and Distributed Processing Symposium (IPDP2000), pp.411-420, May 2000.

[6] OSDL-DBT3:http://ldn.linuxfoundation.org/

[7] TPC-H:http://www.tpc.org/tpch/

[8] Luigi Rizzo:”Dummynt”http://info.iet.unipi.it/ luigi/dummynet/

[9] Ganglia Monitoring System:http://www.ganglia.info/

[10] Aravind Menon,Alan L.Cox,Willy Zwaenepoel: ”Optimizing Network Virtualization in Xen”, 2006 USENIX Annual Technical Conference, pp.15-28, May 2006.

[11] Jose Renato Santos,Yoshio Turner,G.(John)Janakiraman,Ian Pratt:”Bridging the Gap between Software and Hardware Techniques for I/O Virtualization”, 2008 USENIX Annual Technical Conference, pp.29-42, June 2008.

[12] Tanimura Yusuke, Ogawa Hirotaka, Nakada Hidemoto, Tanaka Yoshio, Sekiguchi Satoshi: ”Comparison of Methods for Providing An IP Storage to A Virtual Cluster System”, IPSJ SIG Notes 2007, pp.109-114, March 2007.