A Research On High Availability Mechanism Of Virtual Machine Based On Ceph

Xiaolin Yi
College of Computer Science
Beijing University of Technology
Beijing, China
e-mail: yixiaolin@bjut.edu.cn

Fan Zheng
College of Computer Science
Beijing University of Technology
Beijing, China
e-mail: zhengfan1@lenovo.com

Jie Sun
College of Computer Science
Beijing University of Technology
Beijing, China
e-mail: sunjie5@lenovo.com

Deyue Peng
Lenovo Research & Technology Group
Lenovo Group Limited
Beijing, China
e-mail: pengdy1@lenovo.com

***Abstract*—In the computer world, virtualization technology has immersed into the IT world of all levels. The use of the virtual machine increases the utilization rate of physical host resources greatly increased, but how to ensure the high-efficiency and reliable operation of the virtual machine remains an open problem.**

There are some problems in virtual machine deployment scheme, such as the consumption of network resources costs too much when machine migration occurs, the requirements of shared storage needed by virtualization and so on.

In this paper, in order to solve the above mentioned problems a high availability mechanism of virtual machine will be put forward based on the study of open source projects Ceph.

Keywords- Ceph; High Availability; virtualization

I. Introduction

With the increasing development of network and informationalized level unceasing enhancement, enterprise business growth always asked IT to expanding infrastructure, and government and companies often need to add the server to support the new application, and this inevitably leads to many servers cannot be fully used, the network management cost greatly increase, and reduce the flexibility and reliability. Resource sharing is the main points of the network and under the premise of clear division of responsibilities, using the server cluster and virtual server software can achieve stability and balanced resource utilization.

Virtualization technology is rapidly developed into the core elements of next generation data center. It is a popular method to simplify the management and share physical computing resources. With the utilization of the virtual machine platform (such as VMware, Xen), a single physical host can run multiple operating systems and different configuration of the virtual machines. Virtualization technology application improved utilization of servers, however, the resulting problems such as the reliability need to be solved urgently.

II. Field Analysis

Early deployment of server virtualization greatly save the cost and strengthen the deployment flexibility and adaptability at the same time. As virtualization software gradually mature, simple server integration (integration of 1:10) began to give way to the strategy to completely redesign server environment. The next generation of server hardware will support virtual machine and most of the software will consider virtualization, that is to say, all enterprise applications will support virtualization.

Virtual machine is more easily to deploy than traditional operating system or application. Enterprise can also run multiple virtual machines on a single server to reduce the cost. But there are some problems of the virtual machine deployment:

1) Virtual machine migrates from one physical server to another server transparently and seamlessly, this process makes management and security work more complicated. The application workload for dynamic mobile operation is difficult to ensure that the application complies with network policy. Different application server is bound to a specific VLAN, or the application of network traffic is specified different QoS priority and ACL safety by network administrator.

Virtual machine migration may encounter some technical difficulties such as disk partitions are hard coded. When we do disk virtualization, we may use different types of virtual disk devices, which can lead to changes in the disk name. For example, Xen virtual machine in the para-virtualization equipment using /dev/xvda while standard Linux para-virtualization equipment using /dev/vda. There are some scattered associations on /dev/hd * and /dev/sd * disk partition name in Linux system which is ready for migration, such as in /etc/fstab file and in some auto-started script files which are used to start the initialization file system ramfs and powered up some analytic disk devices. These associations will undoubtedly cause more overhead in the virtual machine migration.

2) Virtualization exacerbated the requirements of shared storage, because there are requirements by OS image data isolated from the underlying server hardware. For this problem, the traditional solution is to use a separate and dedicated SAN, however, it still means that fiber channel for most people. What's worse, SAN is expensive, difficult to manage, and add a new network protocol, exchange structure and management team. With the increase number of virtual machines and storage requirements, expand and manage SAN has become challenging and complicated.

3) Currently some virtual machine deployment separates the OS image and stores it in a shared storage such as a distributed file system, but it still needs to be copied to a host to start a virtual machine on this host. The process of copy image from shared storage to a specific host will not only affect the virtual machine startup time, but can take up considerable network resources.

III. Performance Analysis Of Ceph

A. Ceph Distributed File System

This article will focus on the above problems and will put forward a virtual machine deployment mechanism based on the Ceph distributed file system. Ceph [1] aims at providing a multi-petabyte capacity, high performance, reliable and scalable distributed file system. And as of late March 2010, you can now find Ceph in the mainline Linux kernel (since 2.6.34).

The Ceph [2] ecosystem can be broadly divided into four components: clients, metadata servers, an object storage cluster, and finally the cluster monitors. Clients perform metadata operations (to identify the location of data) using the metadata servers. The metadata servers manage the location of data and also where to store new data. Note that metadata is stored in the storage cluster (as indicated by "Metadata I/O"). Actual file I/O occurs between the client and object storage cluster. In this way, higher-level POSIX functions (such as open, close, and rename) are managed through the metadata servers, whereas POSIX functions (such as read and write) are managed directly through the object storage cluster.

Ceph file system [3] stripes a file onto predictably named objects and utilizes pseudo random distribution algorithm called CRUSH (Controlled Replication Under Scalable Hashing) [4] of to distribute each object to the various nodes. CRUSH uses crushmap (a data structure describes physical resources through hierarchy) and data placement rule to decide which OSD to store an object. Given a single integer input value x, CRUSH will output an ordered list R of n distinct storage targets. Thus, no matter read or write, Ceph will utilize echo nodes' resources such as CPU and memory to calculate the location of an object. When the system increases or reduces osd nodes, data can be live migration and this also utilizes osd's CPU and memory so that make full use of the intelligence of a single node. There are 3 features of Ceph:

1) Decoupled Data and Metadata: Ceph maximizes the separation of file metadata management from the storage of file data. Metadata operations are collectively managed by a metadata server cluster, while clients interact directly with OSDs to perform file I/O. With the utilization of CRUSH, any party can calculate (rather than look up) the name and location of objects comprising a file's contents, eliminating the need to maintain and distribute object lists, simplifying the design of the system, and reducing the metadata cluster workload.

2) Dynamic Distributed Metadata Management: Because file system metadata operations make up as much as half of typical file system workloads, effective metadata management is critical to overall system performance. Ceph utilizes a novel metadata cluster architecture based on Dynamic Subtree Partitioning that adaptively and intelligently distributes responsibility for managing the file system directory hierarchy among tens or even hundreds of MDSs. A hierarchical partition preserves locality in each MDS's workload, facilitating efficient updates and aggressive prefetching to improve performance for common workloads.

3) Reliable Autonomic Distributed Object Storage [5]: RADOS is one of the core part of the Ceph. It can be used as a file system because it has three components of Ceph: OSD, MON and Clients. Ceph assumes that device failures are frequent and expected, and large volumes of data are created, moved, and deleted. And RADOS allows Ceph to more effectively leverage the intelligence (CPU and memory) present on each OSD to achieve reliable, highly available object storage with linear scaling.

B. High Availability

High availability [6] is a system design approach and associated service implementation that ensures a prearranged level of operational performance will be met during a contractual measurement period. There are 2 kinds of solutions to achieve high availability. One is to use redundant hardware resource and another is software control to make full use of existing hardware resources. The former solutions provide a high level of availability, but it's too expensive. So most applications choose the latter solution. Almost all of the computer system supplier can offer such products which can avoid a single point of failure (that is a component failure, will lead to paralysis of the whole system's security operation) to ensure scalability and high availability.

The availability of computer systems is measured by the system reliability and maintainability. And availability is measured by MTTF. MTTF is the length of time a device or other product is expected to last in operation. The higher the reliability of the system is, the longer MTTF will be. Maintainability is measured by MTTR. Mean time to recovery (MTTR) is the average time that a device will take to recover from any failure. The better maintainability of the system is, the MTTR will be. The availability of computer system are defined as MTTF / + MTTR (MTTF) * 100%. Thus, the availability of computer system is defined as the percentage of the system to maintain normal running time.

IV. High Available Mechanism Of Virtual Machine Based On Ceph

Cloud computing platform offers the personal service of IaaS. It is designed for large-scale computing tasks, for instance to provide the computing cluster for Big Data analytics. These computing tasks is lax for the timely response and ability of user interaction, it is more incline to the throughput of cluster. When doing the actual data analytics, the virtual computing cluster of cloud computing platform is required to have the high availability to run without interruption. The cloud computing platform must offer a mechanism of high availability to ensure the tasks on the machine can continue to be processed on other compute node or cluster when the machine is down or meets some errors.

To meet these requirements, this paper proposed and designed a high available mechanism of virtual machine based on Ceph. Now, most of the high available mechanism of virtual machine is based on the special storage, for example IPSAN and more expensive FCSAN [7]. It will take higher costs of device, deployment, maintenance and service. And the special storage will also limit the storage capacity. To meet the strong requirement of Big Data that the ability of storage is unlimited, this cloud computing mechanism selects the Ceph storage.

On the requirement of high available for computing cluster, the mechanism will store the images of virtual machine on Ceph, and do a special optimization for configuration. Because the Ceph can make virtual share storage on logic, it will make the images have high availability and support the migration and live migration for virtual machine. By viewing the storage state, computing node state and virtual machine state of cloud computing platform storage and getting the timely data of all nodes, it can test the fault of compute node and virtual machine quickly and take a timely dispatcher for fault virtual machine to ensure the running of computing cluster.

A. Efficient Access For Storage

After writing the images of virtual machine to Ceph RBD, the compute nodes will map and access images directly by using custom rbd agreement. The rbd agreement is supported by rbd kernel module. Because it is connected with kernel, the update of source code is not very timely. Along with the development of Ceph and the promoting of actual requirement, more and more special requirement have been found. It will be very difficult for official agency to maintain the source code of timely and efficient access agreement for storage. To depth customize the rbd access agreement for storage, and to support the selection for virtual machine will be realized so that it can meet the special requirement for image of virtual machine and get the max promotion of performance. In the storage device, it will offer two levels of storage; they are the normal device and efficient device. The normal device is used to provide large storage capacity by using the universal SATA equipment or SAS equipment. The efficient device uses SSD in small-scale to offer greater access efficiency. It will provide the API of device selection to support the function that users can select whole cluster or some subset of the cluster to deploy on special storage device. The users can transfer their selection to the cloud computing platform by using this API. The cloud computing platform will make a final decision according to the user selections and the load of system. By using this grading method for application and refine the user scenarios to get the promotion of performance.

B. Failure Detection And Task Restart

Cloud computing platform need a well-functional monitor mechanism, Ganglia is chosen after serious consideration. Ganglia is a widely used cluster monitoring software which supplies flexible customization ability. Using Ganglia to monitoring computing nodes can detect node failures as soon as possible. And then failure information is passed to scheduler of our cloud computing platform. We customized OpenNebula as our scheduler, whose core algorithm is modified to support immediately scheduling a group of virtual machines. OpenNebula records relationship between computing nodes and virtual machines running on them. With this information and failure nodes, OpenNebula can get the virtual machines which need to be rescheduled. Based on this information, OpenNebula can reschedule the victim of failure nodes to active computing nodes.

There is another situation in which computing node is active but the virtual machine monitor software fails, and then virtual machine running on it isn't connectable. We can also use Ganglia to monitor these key virtualization daemons, restart is taken when daemon failure is detected and the failure is recoverable. In this situation, daemon restart will work and no scheduling is needed. But there is a more complicated situation, which is certain virtual machine may fail when computing node and virtualization daemons are both fine. Ganglia is used again to monitor all running virtual machines. But it's limited to Linux related virtual machines, as we can customize these virtual machines to insert a Ganglia monitor proxy.

Storing virtual machine images into Ceph storage cluster brings great flexibility to scheduler. Many failure situations can be satisfied when using this mechanism. Based on this fundamental mechanism, we build virtual computing clusters for big data analysis tasks. And two levels of services are supplied: task discontinuous service and task continuous service. For task discontinuous service, our virtual computing cluster only supplies simple virtual machine high availability, which is implemented by failure detection and rescheduling to maintain steady compute ability. When failed virtual machine is restarted, interrupted tasks are not handled. To use this service, tasks themselves need to have certain failure adaptive ability. But as data is stored into Ceph too, user can specify data synchronization frequency to keep data stored in disk fresh. So tasks are not need to start over totally, only a little part will need to be re-computed.

C. Split-Brain Prevention And Performance

Split-brain is a term in computer jargon, based on an analogy with the medical Split-brain syndrome. It indicates data or availability inconsistencies originating from the maintenance of two separate data sets with overlap in scope, either because of servers in a network design, or a failure condition based on servers not communicating and synchronizing their data to each other. Virtual machine split-brain is a common error in virtualization systems nowadays. Split-brain can bring terrible results, virtual machine image can be corrupted and data can be confused. By putting virtual machine images into Ceph storage cluster and customizing Ceph source code, we can have a robust split-brain prevention solution. Storage systems demand strong data safety and consistency, so as Ceph. Ceph use Paxos protocol and distributed locking subsystem to maintain data consistency. By adding virtual machine image's access information into core data structure, and sharing it with virtual machine scheduler, scheduler can discover split-brain and stop scheduling that virtual machine.

As our cloud computing platform mainly supplies massive virtual compute ability, it's not for daily office usage. So real time respond ability is not needed. We focus on the whole compute ability and throughput. Putting virtual machine images into distributed storage system may be slower than local disks. There may be some performance discount. But we can make it better by using better hardware, for example, we use 10000M storage Ethernet. And as Ceph support customizable stripe mechanism, we provide a set of stripe specifications for different circumstances and supply encapsulation API to user for better performance. We also deeply customized Ceph's snapshot functionality for better support of failure detection and task restart, failure recovery is eased and re-computation is decreased.

V. Experiment

The virtual machine will be started on host and block device. The configuration of machine is 2 CPUs; 4G RAM and 20G ROM. The image of machine is opensuse11.4. The host machine is named vm-host and the chunk machine is named vm-block.

A. Start Up

The vm-host machine will be started on the OpenNebula management platform. Because the image of virtual machine needs to move to vm-host location from the share storage location, it will spend about 5 minutes to transfer (a image size is 30G). When to start the vm-block machine, the vm-block spend 10 seconds because the image of virtual machine is pre-assigned. Even if there is not pre-assigned work, it will spend few time because the image is stored in Ceph file system and it will be split to many object to store on different OSD node, the copy process will be parallel so that can offer a high performance.

B. Lmbench Test Tool

Lmbench is an easy and portable test tool for UNIX/POSIX. It accords with the ANSI/C standard and be used to estimate the comprehensive performance of system. Its test content contains the IO of file, the operation of memory, the cost of process creation, and the network performance and so on. Table I is to show the runtime of processor and the process.

Null call: simple system call (get the process number).

Null I/O: simple IO operation (the average value of read and write of null)

Stat: the operation of getting the state of document

Open clos: open then close the document immediately

Slct tcp: network test

Sig inst: configuration of signal

Sig hndl: catch the process signal.

Fork proc: to exit directly after the Fork process

Exec proc: to execute the ‘execve’ after the Fork process

Sh proc: to execute the ‘shell’ after the Fork process

Table II shows context switch times between different numbers of processes with different working set sizes, 2p/16K presents 2 processes handle 16K data parallel.

From the above experiment data, it can be seen that there is not obvious difference on processor operation, context switching, local communication bandwidth and others between two types of virtual machine.

C. IO Test By Using Iozone

Because the virtual disk of vm-host is the physical disk of host, the performance of IO will better than vm-block. Fig.1 and Fig.2 show the test results of read and write by using iozone on two virtual machines. Then use the result that the vm-block data to divide the vm-host data as the difference of IO performance between on vm-block and vm-host. In the figure, the X axis indicates the size of file (from 64KB to 512MB); the Y axis indicates the record size (from 4k to 64k) and the Z axis indicates the percentage. Most of the performance number is between in 0.8 and 1.1, so the IO performance is acceptable.

TABLE I. PROCESSOR, PROCESSES - TIMES IN MICROSECONDS - SMALLER IS BETTER

Host	OS	Mhz	null call	null I/O	stat	open clos	slct TCP	sig inst	sig hndl	fork proc	exec proc	sh proc
vm-block	Linux 2.6.37	2770	0.05	0.10	1.00	1.94	2.28	0.14	0.89	312	964	3372
		2770	0.05	0.11	1.06	1.86	2.33	0.14	0.89	165	485	2663
		2770	0.05	0.12	1.06	1.75	2.31	0.14	0.93	184	590	2743
vm-host		2770	0.05	0.11	1.09	1.78	2.48	0.14	0.90	306	917	3368
		2770	0.05	0.11	1.03	1.86	2.31	0.15	0.94	295	581	2061
		2770	0.05	0.11	1.05	1.75	2.31	0.14	0.92	297	930	3356

TABLE II. CONTEXT SWITCHING - TIMES IN MICROSECONDS - SMALLER IS BETTER

Host	OS	2p/0K ctxsw	2p/16K ctxsw	2p/64K ctxsw
vm-block	Linux 2.6.37	1.3900	1.5500	1.9900
		1.3700	1.6600	2.0500
		1.5000	1.5400	2.0400
vm-host		1.6000	1.5000	1.8100
		1.4100	1.6900	1.6800
		1.3800	1.4000	1.7500

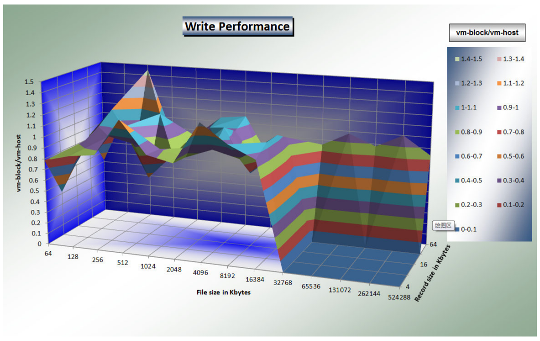

Figure 1. Write test.

D. Virtual Machine Live Migration

Use the live migration feature of the OpenNebula management platform to test the time needed for testing live migration of virtual machine. Live migration time is related to network bandwidth, the performance of the physical host and virtual machine configuration. Network connection based on 100 MBPS Ethernet in this experiment, the migration process takes about 400 seconds. When test the migration of virtual machine started on block devices, due to the virtual machine startup based on Ceph, virtual machine live migration time decreases greatly by the migration part of the file system greatly decreases, and you just need to map the block device to the target host.

E. *Reliability Test*

When virtual machine is running, in order to simulate a scene that a physical disk is down, pull out a piece of disk deliberately. Then use command "dd(eg:dd if=/dev/zero of=testfile bs=1M count=2000)" to write file to disk uninterruptedly. Take this test for 10 times and compute the average. Compare the normal results and the result in the scene that a disk is down, the former is 67.7389s and the latter is 68.4632s. This indicates that damage to a single disk is not affected for the use of virtual machine.

VI. CONCLUSION

This paper summarizes the status of the virtual machine deployment, the analysis of existing problems, and puts forward a kind of virtual machine high availability mechanism based on the Ceph distributed file system. Using features of Ceph, by demonstrating experiment, the mechanism can effectively shorten the time needed for the virtual machine migration, reduce the network resources to virtual machine migration occupancy; Shorten the virtual machine startup time;

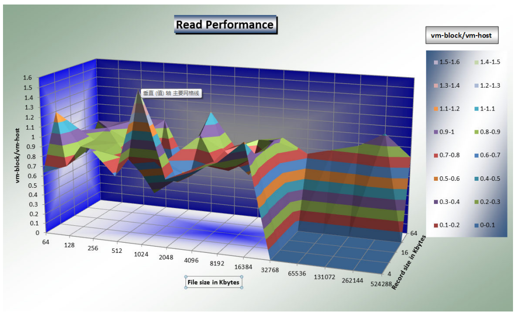

Figure 2. Read test.

And solved the problem that traditional virtual machine deployment plans require expensive equipment.

Contrast the solution that image file stored in a file system and that image file stored in the Ceph RBD, the former solution access to the data through the file system interface, while the latter one uses a special piece of equipment access interface, generates shorter access path.

Using Ceph RBD block device to store image, the most direct way is to write image file content section byte by byte into the RBD equipment. On the basis of this simple scheme, in order to guarantee the high availability of image storage, virtual machine images for RBD specially created a series of Ceph RBD resource pools, each resource pool has a fixed number of data backup. Then the upper according to the specific needs of the business, stores virtual machine images which require high reliability into the more copies of data backup resource pool, while store those don't need very high reliability into the less copies of data backup resource pool. In order to obtain better performance, we can optimize the image format. While increasing SSD and other high-speed access equipment is a good way to improve the performance of the Ceph itself.

REFERENCES

[1] M.Tim Jones, "Ceph: a linux petabyte-scale distributed file system," IBM developer works, 2010.5

[2] Ceph Official documentation [OL].http://ceph.com/docs/master/

[3] Weil, S. A., Brandt, S. A., Miller, E. L., Long, D. D., & Maltzahn, C. (2006, November). Ceph: A scalable, high-performance distributed file system. In Proceedings of the 7th symposium on Operating systems design and implementation (pp. 307-320). USENIX Association.

[4] Weil, S. A., Brandt, S. A., Miller, E. L., & Maltzahn, C. (2006, November). CRUSH: Controlled, scalable, decentralized placement of replicated data. In Proceedings of the 2006 ACM/IEEE conference on Supercomputing (p. 122). ACM.

[5] Leung, Sage A. Weil Andrew W., and Scott A. Brandt Carlos Maltzahn. "RADOS: A Scalable, Reliable Storage Service for Petabyte-scale Storage Clusters."

[6] Floyd Piedad, Michael Hawkins (2001). High Availability: Design, Techniques, and Processes. Prentice Hall

[7] Date C, Spoor R, Peddemors A, et al. Survey of Technologies for Wide Area Distributed Storage[J]. 2010.