Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 11, November 2014, pg.149 – 155

SURVEY ARTICLE

A SURVEY ON DE-DUPLICATION IN CLOUD COMPUTING

Priyadharsini.P[1], Dhamodran.P[2], Kavitha.M.S[3]

Department of Computer Science and Engineering, India

[1] priyadharsini.pp@gmail.com; [2] dhamodranp@gmail.com; [3] kaviktg@gmail.com

Abstract— Cloud computing is an emerging technology for providing infrastructure as a services to cloud users. The infrastructure as a service is based on virtualization where it allocates the virtual machine to user through internet. Virtual machine is a guest machine runs in the environment of host machine. VMI are used to purchase VM instances to run on virtual machine in cloud platforms. The storage of large number of VMI and provisioning remains challenging problem. In this paper, data deduplication is a method used in VMI and various data deduplication methods are available which make VMI storage and provisioning to be easy. In this, we will examine all methods, processes used in data deduplication to overcome the challenges faced in virtual machine images.

Keywords— data de-duplication, virtual machine images, virtual machines

1. INTRODUCTION

Cloud computing is the delivery of services like software, platform, infrastructure over the Internet. The infrastructure as services provides both hardware and software as a service by virtualization technology to the cloud users. Virtualization is a process of creating a virtual version of os, server, hardware, software. Virtual machine is like computer running within a computer and known as "guest" machine. VMI formats are supported by hypervisor like Xen, Kvm, VMware, Virtual box etc. The large scale VM deployment causes the burden in storage and provisioning the VM and it is overcome by data deduplication techniques. In storage of VMI leads to duplication of files in storage system. Data deduplication eliminates the redundant data in storage system which improves the utilization of storage. In the de-duplication process, redundant data is deleted only one copy or single instance of the data to be stored in the database.

2. BACKGROUND

Our research studies the effectiveness of applying deduplication to virtual machine environments.

A.VIRTUAL MACHINE IMAGES

A virtual machine image is a file which consists of virtual disk and has bootable operating system installed on it. The VMI has different format which is described below

1) RAW

The raw image format is byte-to byte copying of physical disk content into a regular file and it is supported by both KVM and Xen hypervisors.

2) SPARSE

The sparse image format constructs a mapping which is complex between the blocks in physical disks and data blocks in VMI.

3) AMI/AKI/ARI

The AMI/AKI/ARI format are supported by Amazon EC2.AMI (Amazon Machine Image) is virtual image format which is knows as raw format. AKI (Amazon Kernel Image) is a kernel file which is loaded to boot the virtual image. ARI (Amazon Ramdisk Image) is a ram disk file that is loaded to boot. A UEC (Ubuntu Enterprise Cloud) tarball is tarfile contains an AMI file, AKI file, ARI file.

4) Qcow2

The QEMU copy-on-write version 2 format use the sparse representation and it is supported by snapshots. It is supported by KVM hypervisor.

5) VMDK

Virtual Machine Disk (VMDK) is a disk image file format developed by the hypervisor called VMware for its virtual appliance of products. The virtual machines like VMware Workstation or Virtual Box are used in the virtual hard disk drive. For normal application, VMDK has the size of 2TB and later it has capacity of 62TB developed by VMware known as VMware QEMU Sphere. The other platforms which support VMDK files are Sun XVM, Virtual Box or QEMU.

6) VHD

Virtual hard disk (VHD) is a disk image file format used for storing the complete contents of a hard drive. The disk image is called a virtual machine, replicates an existing hard drive and includes all data and structural element. A Virtual Hard Disk allows multiple operating systems to reside on a single host machine. VHD formats are supported by Microsoft Virtual PC and Virtual Server. It is very easy to deployment and have backup and restore with multi-user isolation.

7) VHDX

VHDX is the version of Hyper-x which has additional features over VHD. VHDX has a large storage capacity than a VHD format. VHDX protects against the data corruption during power failures. The improved alignment of the virtual hard disk format to work well on large sector disks. Support for virtual hard disk storage capacity of up to 64 TB. The efficiency in representing data which results in smaller file size

8) OVF

Open Virtualization Format (OVF) package contains one or more image files which is defined by Distributed Management Task Force standards group. The OVF package contains virtual systems and deployed in virtual machine. It consists OVF descriptor known as XML file which have virtual machine package. The OVF package consists of disk images and has metadata for package. It is supported by the hypervisor VMware, virtual box, Oracle VM, Red Hat Enterprise Virtualization. The OVF is not designed for particular hypervisor.

B. Hypervisors

The hypervisors are also known as virtual machine monitor used to create, runs and monitor the virtual machine. Each virtual machine is known as guest machine and in which hypervisor and virtual machine are running is known as host machine. The hypervisors manages the execution of guest operating system. It has two types of hypervisors, they are

1) Type 1 hypervisor

Type 1 hypervisor are known as bare metal or native hypervisors runs directly on the host machine hardware to control the hardware and to manage guest operating systems. It is supported by Oracle VM Server, XenServer, VMware ESX/ESXi and Microsoft Hyper-v.

2) Type 2 hypervisor

Type 2 hypervisor are known as hosted hypervisors run within a operating system environment. The guest operating system runs above the hardware and it is supported by VMware Workstations and Virtual Box

3. ISSUES IN VMI STORAGE

The main issue in large scale VMI deployment and provisioning is storage consumption where existing system solve the issue by SAN cluster deduplication. The SAN is very expensive and does not satisfy the large scale deployment of VMI. The Virtual Machine provisioning is done with deploying and structuring the Virtual Machine Images before VM hosting. During this VM hosting storage is not efficiency due to data duplication and there is revenue loss for cloud vendors and cause fault tolerance problem. Data duplication makes number of duplicate copies of same data in storage or datacenter leads to low efficiency .To overcome this issue, the specialized data compression technique called data deduplication is used for eliminating duplicate copies of repeating data.

4. Data de-duplication

The technique data de-duplication is used to store single instance of redundant data and eliminates the duplicate data in datacenter. It is used to decrease the size of datacenter and reduce the replications of data that were duplicated on cloud. The de-duplication process helps to remove any block or file that are not unique and store in smaller group of blocks

The basic steps for data de.duplication process are.

- The files are converted into small segments.
- Then new and existing data are checked for redundancy
- Metadata are updated and segments are compressed.
- Duplicate data are deleted and check the data integrity.

They are two methods used to break the file into segments, called, fixed size chunking and variable size chunking. The fixed size chunking will splits the original file into blocks in same size. The variable size chunking is done with Rabin fingerprint on file content and it also detects the boundaries inside the file. Comparing both VMI formats mostly use fixed size chunking which is good in de-duplication.

Advantages,

- reduced storage
- efficient volume replication
- scalability
- IO performance

A. Data de-duplication process

1) Offline data de-duplication In offline data de-duplication, the de-duplication process is carried out after storing the data in storage disk or datacenter.

2) Online data de-duplication Online data de-duplication, the de-depulication process is carried out before storing the data in storage disk or datacenter.

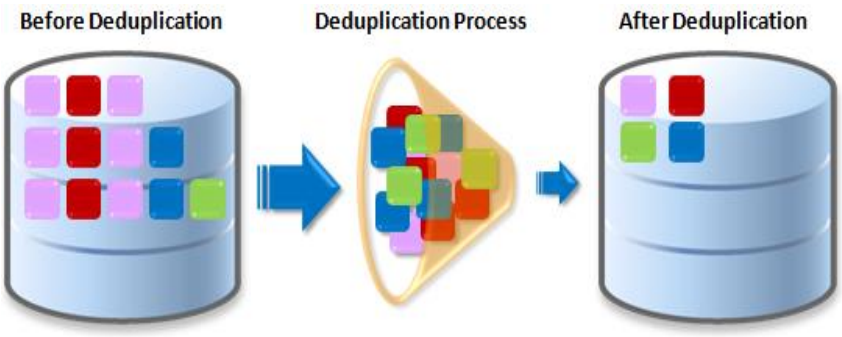

Fig1: Data-deduplication process

B. *Data-deduplication types*

1) *Target-based deduplication*

Target based deduplication is used to eliminate the duplicate copies from data after transmission to backup. it will reduce only the storage space and will not reduce the size of data. it mainly used to save the storage space and does not save the bandwidth.

2) *Source-based deduplication*

Source based deduplication is performed before the transmission of data to target backup. Source deduplication will increase both bandwidth and storage efficiency. Source deduplication will backs up only the unique data and never sent the replicate data through network. and it is used only for large amount of data.

C. Data-deduplication levels

1) File-level deduplication

In this deduplication method, the duplicate files are identified if they have same hash value and is performed over single file. It requires less processing power since files' hash numbers are relatively easy to generate.

Advantage:

- If any change is made in a file it makes to save the whole file again in file level deduplication.
- In file level deduplication indexes are small, and so it takes less time for computational when it identifies the duplicate copies.

2) Block-level deduplication

Block level deduplication is performed over blocks. It first divides files into blocks and stores only a single copy of each block. It could either use fixed-sized blocks or variable-sized chunks.

Advantage:

- Block level deduplication can eliminate or delete the small redundant chunk of data when compared to whole file.
- Each and every file system can use same deduplication algorithm in block level deduplication..

3) Byte-level deduplication

Byte-level deduplication is a form of block-level deduplication that understands the content, or "semantics", of the data. Byte-level deduplication understands the content of the data and the system can more efficiently deduplicate the bytes within the data stream that is being deduplicated.

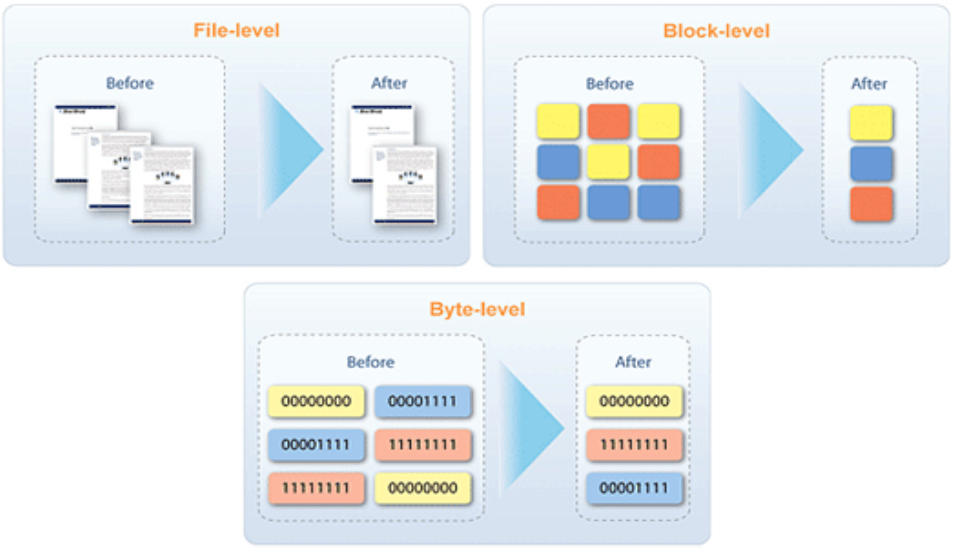

Fig2: Deduplication levels

5. Deduplication storage system

There are several deduplication storage systems which is designed for eliminating duplicate copies in storage system. The similarity between them is just that all are data de-duplication based storage system.

A. LiveDFS:

Live Deduplication File System enables deduplication storage of VM images in a open source cloud which is implemented under low cost .LiveDFS can save up to 40% of storage space for VM images. the implementation of LiveDFS prototype as a storage layer in cloud based on open stack and it is based on inline deduplication. It is designed based on commodity hardware and os. It is applicable only to single storage partition.

B. Venti

Venti is a new approach to archival storage, and it is building block for storage applications such as logical, physical backup and snapshots. It is based on block-level network storage and enforces the write-once policy. Venti is not applicable for different block size if data is shifted within file or application. Authenticating to venti server, client can read any blocks by single root fingerprint.

C. HYDRAstor

HYDRAstor is scalable secondary storage solution system consists of a back-end architecture as a grid of storage nodes built around a distributed hash table. The data blocks are in a DAG (Directed acyclic graph). It has the problem in delivering a value to end user and it saves the bandwidth when deduplication is moved to proxy server.

D. Liquid

Liquid is scalable deduplication file system for VM images which is designed for large scale VM deployment. Liquid propose

- Instant cloning of VMI
- On-demand fetching
- Copy-on-write.

It also has some minor performance overhead

6. Working mechanism

The data-deduplication is processes which can be explained with some of the files that can be easily understand. We have three files namely image.txt, images.txt, img.txt stored in database. When image.txt is stored first data deduplication breaks the whole files into segments as 1 2 3 4. The files can be split into more segments also based on deduplication algorithm. For reconstruction, add hash identifier to all the segments so now all the segments get stored separately in database. Next the second file images.txt is stored in database and again the file gets breaks into four segments as 1 2 3 4. These segments are same as image .txt file segments. Now de-duplication system will delete the copy of images.txt and will not store. It provides a link to last stored segments. When another file img.txt is to be store in database, then the system will breaks the file into segments. The img.txt breaks the segments as 5 2 3 4. The 5 is new segments and 2 3 4 are already store in the database. So now, system will store only 5^{th} part. Then it will provide the link to another part. Finally only five segments of data will be stored in place of 12 blocks. So the de-duplication clearly shows that it reduces the storage space. In this if we didn't apply de-duplication 7MB of space will be waste. So we save totally 7MB memory space by using data de-duplication and provide link to another segment.

7. Performance metrics

A. Storage Space:

When de-duplicated segments are saved on the cloud storage, storage space is reduced. The storage space is saved is by deduplication method is test using two files, one original file and other duplicate file. If two files are saved to system the data- deduplication method is applied to both of the files and after data-duplication if it finds one file is original and other is copy of original file then only original file is saved and copy of file is removed. So it saves the memory space. Here is an performance chart shown that clearly how the data de-duplication reduce the overall size of the storage system.

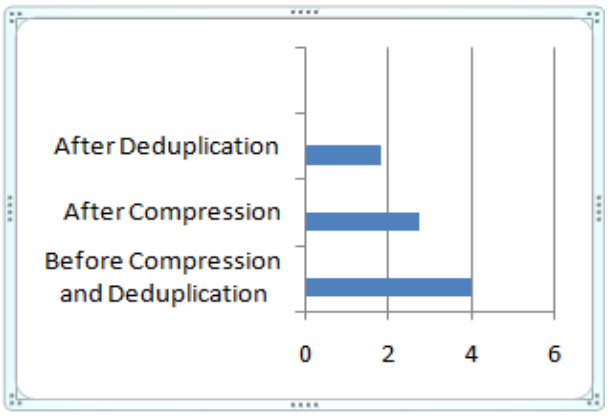

Fig3: Performance chart

The chart show file size before compression and deduplication is 4mb. After compression it reduce the size to 2.5mb and after applying deduplication method in removing redundant data the file size is reduced to 1.5 mb. So by applying deduplication in storage it will reduce the storage space and have good efficiency.

To improve the storage utilization data deduplication techniques is used for eliminating duplicate copies of repeating data.

8. Conclusion

Deduplication is an efficient approach to reduce storage demands in environments with large numbers of VM disk images. It included all detail about data de-duplication and methods to achieve it. The deduplication of VM disk images can save more of the space required to store the operating system and application environment .In future, deduplication of data storage in order to reduce the amount of drives spinning.

References

1. C. Ng, M. Ma, T. Wong, P. Lee, and J. Lui, ‘‘Live Deduplication Storage of Virtual Machine Images in an Open-Source Cloud,’’ in Proc. Middleware, 2011, pp. 81-100
2. S. Quinlan and S. Dorward, ‘‘Venti: A New Approach to Archival Storage,’’ in Proc. FAST Conf. File Storage Technol., 2002, vol. 4, p. 7.
3. C.Dubnicki, l.Gryz, l.Heldt, M.Kaczmarczyk, W.Kilian, P.Strzelczak, j.Szczepkowski, c.ungureanu,” Hydrastor: A Scalable Secondary Storage” in proc 7th cong.file sorage technol.,, 2009.
4. http://www.starwindsoftware.com/file-level-vs-block-level-vs-byte-level-deduplication
5. http://blog.unitrends.com/backup-and-file-level-versus-block-level-deduplication/
6. http://documentation.commvault.com/hds/release_8_0_0/books_online_1/english_us/features/single_instance/single_instance.htm#How_does_Deduplication_work

