Collocated Deduplication with Fault Isolation for Virtual Machine Snapshot Backup

Wei Zhang, Michael Agun, Tao Yang
Department of Computer Science
University of California, Santa Barbara, CA 93106

ABSTRACT

A cloud environment that hosts a large number of virtual machines (VMs) has a high storage demand for frequent backup of system image snapshots. Deduplication of data blocks can lead a big reduction of redundant blocks when their signatures are identical. However it is expensive and less fault-resilient to perform a global deduplication using signatures and let a data block share by many virtual machines. This paper studies a VM-centric scheme which collocates a lightweight backup service with other cloud services in a cluster and it integrates multiple duplicate detection strategies that localize deduplication as much as possible within each virtual machine. It also organizes the write of small data chunks into large file system blocks so that each underlying file block is associated with one VM for most of cases. Our analysis shows that this VM centric scheme can provide better fault tolerance while using a small amount of computing and storage resource. This paper provides a comparative evaluation of this scheme in accomplishing a high deduplication efficiency while sustaining a good backup throughput.

1. INTRODUCTION

In a cluster-based cloud environment, each physical machine runs a number of virtual machines as instances of a guest operating system and their virtual hard disks are represented as virtual disk image files in the host operating system. Frequent snapshot backup of virtual disk images can increase the service reliability. For example, the Aliyun cloud, which is the largest cloud service provider by Alibaba in China, automatically conducts the backup of virtual disk images to all active users every day. The cost of supporting a large number of concurrent backup streams is high because of the huge storage demand. Using a separate backup service with full deduplication support [7, 11] can effectively identify and remove content duplicates among snapshots, but such a solution can be expensive. There is also a large amount of network traffic to transfer data from the host machines to the backup facility before duplicates are removed.

This paper seeks for a low-cost architecture option that collocates a backup service with other cloud services and uses a minimum amount of resources. We also consider the fact that after deduplication most data chunks are shared by several to many virtual machines. Failure of shared data chunks can have a catastrophic effect and many snapshots of virtual machines would be affected. The previous work in deduplication focuses on the efficiency and approximation of finger print comparison, and has not addressed fault tolerance together with deduplication. Thus we also seek deduplication options that yield better fault isolation. Another issue considered is that that deletion of old snapshots compete for computing resource as well. That is because data dependence created deletion of old snapshots competes for computing resources as well so deletion needs to be considered in system design. The additional complexity for deletions come from the dependencies created by multiple links to the same problem, so some solution is needed to determine blocks with no dependent snapshots after deletion. especially when VMs can migrate around in the cloud.

The paper studies and evaluates an integrated approach which uses multiple duplicate detection strategies based on version detection, inner VM duplicate search, and controlled cross-VM comparison. This approach is VM centric by localizing duplicate detection within each VM and by packaging data chunks from the same VM into a file system block as much as possible. By narrowing duplicate sharing within a small percent of data chunks, this scheme can afford to allocate extra replicas of these shared chunks for better fault resilience. Localization also brings the benefits of greater ability to exploit parallelism so backup operations can run simultaneously without a central bottleneck. This VM-centric solution uses a small amount of memory while delivering a reasonable deduplication efficiency. We have developed a prototype system that runs a cluster of Linux machines with Xen. The backup storage uses a standard distributed file system with data replication and block packaging.

The rest of this paper is organized as follows. Section **??** reviews background and related work. Section **??** discusses the design options for snapshot backup with a VM-centric approach. Section **??** analyzes the benefit of our approach for fault isolation. Section 5 describes our system architecture and implementation details. Section **??** is our experimental evaluation that compare with the other approaches. Section **??** concludes this paper.

2. BACKGROUND AND RELATED WORK

At a cloud cluster node, each instance of a guest operating system runs on a virtual machine, accessing virtual hard disks represented as virtual disk image files in the host operating system. For VM snapshot backup, file-level semantics are normally not provided. Snapshot operations take place at the virtual device driver level, which means no fine-grained file system metadata can be used to determine the changed data.

The previous work for storage backup has extensively studied data deduplication techniques that can eliminate redundancy globally among different files from different users. Backup systems have been developed to use content fingerprints to identify duplicate content [7, **?**]. Offline deduplication is used in [**?**, **?**] to remove previously written duplicates during idle time. Several techniques have been proposed to speedup searching of duplicate fingerprints. For example, the data domain method [11] uses an in-memory Bloom filter and a prefetching cache for data chunks which may be accessed. An improvement to this work with parallelization is in [10, **?**]. The approximation techniques are studied in [2, 4, **?**] to reduce memory requirements with the tradeoff of a reduced deduplication ratio.

Additional inline deduplication techniques are studied in [5, 4, **?**]. All of the above approaches have focused on optimization of deduplication efficiency, and none of them have considered the impact of deduplication on fault tolerance in the cluster-based environment that we have considered in this paper. We will describe the motivation of using the cluster-based approach for running the backup service and then present our solution with fault isolation.

3. DESIGN CONSIDERATION AND OPTIONS

As discussed earlier, collocating the backup service on the existing cloud cluster avoids the extra cost to acquire a dedicated backup facility and reduces the network bandwidth consumption in transferring the un-deduplicated raw data for backup. Figure 1 illustrates the cluster architecture where each physical runs a backup service and a distributed file system (DFS) [**?**, **?**] serves a backup store for the snapshots. The previous study shows that deduplication can compress the backup copies effectively in a 10:1 or even 15:1 rage. Therefore the portion of space in a cluster allocated for snapshots of data should not dominate the cluster storage usage.

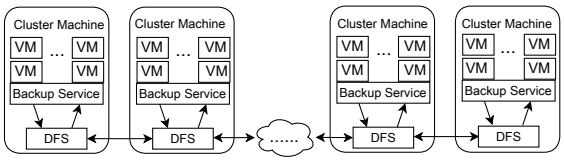


Figure 1: Collocated VM Backup System.

We discuss the design considerations as follows.

- *Deduplication localization, sharing minimization, and fault tolerance.*

  Because a data chunk is compared with signatures collected from other VMs during the deduplication process, only one copy of duplicates is stored in the backup storage and this artificially creates data dependency among different VM users. Content sharing via deduplication affects fault isolation since machine failures happen at daily basis in a large-scale cloud and loss of a small number of shared data chunks can cause the unavailability of snapshots for a large number of virtual machines. Localizing the impact of deduplication can increase fault isolation and resilience. Thus from the fault tolerance point of view, duplicate sharing among multiple VMs is discourage. Another disadvantage of sharing is that it complicates snapshot deletion. The mark-sweep approach [**?**] is effective for deletion, and its main cost is to count if a data chunk is still shared by other snapshots. Thus our goal is to localize deduplication and minimize sharing, while we seek for a tradeoff since there are a significant number of duplicates cross VMs.

- **Packaging data chunks as file system blocks.** Since the file block size in the Hadoop and GFS is uniform and large with 64MB as a default setting, the content chunk in a deduplication system is of nonuniform size with 4KB or 8KB on average. We need to build an intermediate layer that supports large snapshot writing with append operations and infrequent snapshot access. Packaging that maps data chunks to file system blocks can create data dependence among VMs since a file block can be shared even more VMs. Thus we need to consider a minimum association of a fie system block to VMs in the packaging process.

Because of collocation of this snapshot service with other existing cloud services, cloud providers wish that the backup service only consumes small resources with a minimal impact to the existing cloud services. The key usage of resource for backup is memory for storing and comparing the fingerprints. We will consider the approximation techniques with less memory consumption studied in [2, 4] along with the fault isolation consideration discussed below.

A deduplication scheme compares the fingerprints of the current snapshot with its parent snapshot and also other snapshots in the entire cloud without consideration of . We call this as the VM-oblivious (VO) approach. In designing and selecting a duplication algorithm, we have considered the following options.

- **Version-based change detection**. VM snapshots can be backed up incrementally by identifying file blocks that have changed from the previous version of the snapshot [3, **?**, **?**]. Active snapshots will contain all the information needed to restore the virtual disk and when deleting a snapshot, data referenced by other snapshots

are removed. While full signature comparison can deliver additional reduction [4, **?**, **?**], content change detection falls into a VM-centric approach since deduplication is localizable. We are seeking for additional optimization to improve duplication efficiency and design the association of data chunks with the underlying file blocks so that file block sharing among VMs is minimized.

- **sampled Index**. One alternative approach to reducing the use of memory space is to use a sampled index with prefetching, proposed by Guo and Efstathopoulos[4]. An evaluation with our test data shows that using a sampled index can achieve a high deduplication efficiency with a single machine setup. One key problem is that the algorithm is VM oblivious and not easy to be adopted for VM-centric. Another problem is to design a distributed version in deduplicating large bodies of data. To use a distributed memory version of the sampled index, every deduplicate request may access a remote machine for index lookup, which incurs a significant overhead.

  Another possibility is to use this approach partially in our VM centric solution by indexing the most popular data chunks. For a small set of popular data chunks, the prefetching strategy used in the sampled index will not work well because the spatial locality is limited among popular data chunks. In our test data, on average the number of consecutive data chunks is 7, which is too small to be effective for index sampling.

- **Stateless Data Routing**. Another approach for duplicate comparison is to use a content-based hash partitioning algorithm called stateless data routing [**?**] that divides the deduplication work with an approximation, which is similar to Exreme Binning[2]. Again such an approach is VM-oblivious while each request does incur a network routing latency, and then additional overhead for a bloomer filter based disk lookup [**?**] or a disk index access [2].

With these considerations in mind, we study a VM-centric approach (called VC) for a backup service co-hosted in the existing set of machines and resource usage is friendly to the existing applications. We will first discuss and analyze the integration of the VM-centric deduplication strategies with fault isolation, and then present an architecture and implementation design with deletion support.

4. VM-CENTRIC SNAPSHOT DEDUPLICATION

Our VC design has the following objectives:

- 1) Localize the deduplication and data blocking within each VM as much as possible so that any failure of a file system block mainly affects the associated VM. Localizing the snapshot data deduplication within a VM improves the system by increasing data independence between different VM backups, simplifying snapshot management and statistics collection, and facilitating parallel execution of snapshot operations.

- 2) Maintain a competitive deduplication effiency while being resource-friendly to other existing cloud services. Cross-VM duplication can be desirable since there are many cloud images use widely-used software and libraries and their data blocks are duplicate. As the result, different VMs tend to backup large amount of highly similar data.

- 3) Mininimize the number of data chunks shared among VMs and add extra replicas to incrase fault resilience for such data chunks.

4.1 Key VC Strategies

- **Local duplicate search.** We start with the standard dirty bit approach in a coarse grain segment level and use the Xen virtual device driver to implement a feature called "changed block tracking" for the storage device and the dirty bit setting is maintained in a coarse grain level we call it a segment. In our implementation, the segment size is 2MB. Since every write for a segment will touch a dirty bit, the device driver maintains dirty bits in memory and cannot afford a small segment size.

  It should be noted that dirtybit tracking is supported or can be easily implemented in many major virtualization solution vendors. The VMWare hypervisor has an API to let external backup application know the changed areas since last backup. Xen doesn't directly support it, However, their open-source architecture allows anyone to extend the device driver, thus enabling changed block tracking. We implement dirty bit tracking this way in Alibaba's platform. The Microsoft SDK provides an API that allows external applications to monitor the VM's I/O traffic, therefore changed block tracking can be implemented externally.

  Since the best deduplication uses a nonuniform chunk size in the average of 4K or 8K [**?**], we conduct additional local inner-VM deduplication by comparing chunk signatures within a dirty segment from its parent snapshot. We load the chunk fingerprints in the corresponding segment from the parent and perform fingerprint matching for further inner-VM deduplication. The amount of memory for maintaining those fingerprints is small, as we only load one segment at a time. For example, with a 2MB segment, there are about 500 fingerprints to compare.

- **Cross-VM deduplication with popular chunks and replication support** This step accomplishes the standard global fingerprint comparison as conducted in the

previous work [**?**]. One key observation is that the inner deduplication has removed many of the duplicates. There are fewer deduplication opportunities across VMs while the memory and network consumption for global comparison is more expensive. Thus our approximation is that the global fingerprint comparison only searches for the top k most popular items. This dataset is called PDS (common data set). The popularity of a chunk is the number of data chunks from different VMs that are duplicates of this chunk after the inner VM deduplication. This number can be computed periodically on a weekly basis. Once the popularity of all data chunks is collected, the system only maintains the top k most popular chunk signatures in a distribued shared memory.

Since k is relatively small and these top k chunks are shared among multiple VMs, we can afford to provide extra replicas for these popular chunks to enhance the fault resilience.

- **VM-centric file system block management.** When a chunk is not detected as a duplicate to any existing chunk, this chunk will be written to the file system. Since the backend file system typically uses a large block size such as 64MB, each VM will accumulate small local chunks. We manage this accumulation process using an append-store scheme and discuss this in details in Section 5. The system allows all machines conduct the backup in parallel in different machines, and each machine conducts the backup of one VM at a time, and thus only requires a write buffer for one VM.

  PDS chunks are stored in a separate append-store instance. In this way, each file block for non-PDS chunks is associated with one VM and does not contain any PDS chunks.

4.2 Impact on deduplication efficiency

We analyze how the choice of value k impacts the deduplication efficiency. The analysis is based on the charactristics of the VM snapshot traces studied from Alibaba's production user data. Our previous study shows that the popularity of data chunks after inner VM deduplication follows a Zipf-like distribution[**?**] and its exponent α is ranged between 0.65 and 0.7. [**?**]. Table 1 lists parameters used in this analysis.

[Need to find a place to put these numbers in: Total number of chunks in 350 snapshots: 1,546,635,485. Total number of chunks after localized dedup: 283,121,924. Total number of unique chunks: 87,692,682.]

As summarized in Table 1, let c be the total number of data chunks. c_u be the total number of fingerprints in the global index after complete deduplication, and f_i be the frequency for the ith most popular fingerprint. By Zipf-like

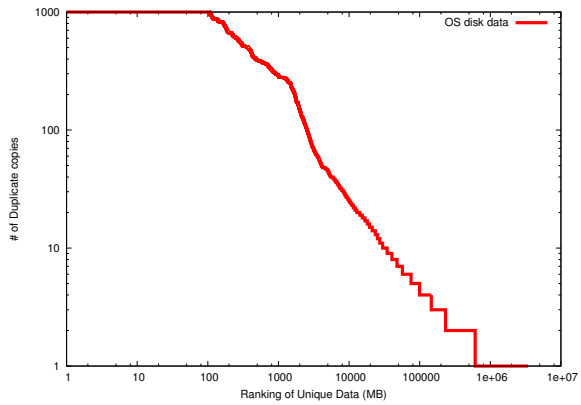


Figure 2: Duplicate frequency versus chunk ranking in a log scale.

distribution, $f_i = \frac{f_1}{i^\alpha}$. Since $\sum_{i=1}^{c_u} f_i = c$,

$$f_1 \sum_{i=1}^{c_u} \frac{1}{i^\alpha} = c.$$

Given $\alpha < 1$, f_1 can be approximated with integration:

$$f_1 = \frac{c(1-\alpha)}{c_u^{1-\alpha}}. \tag{1}$$

The k most popular fingerprints can cover the following number of chunks after inner VM deduplication:

$$f_1 \sum_{i=1}^{k} \frac{1}{i^\alpha} \approx f_1 \int_1^k \frac{1}{x^\alpha} dx \approx f_1 \frac{k^{1-\alpha}}{1-\alpha}.$$

Deduplication efficiency of VC using top k popular chunks is the percentage of duplicates that can be detected:

$$\frac{c(1-\delta) + f_1 \frac{k^{1-\alpha}}{1-\alpha}}{c(1-\delta) + \delta c - c_u} = \frac{(1-\delta) + \delta(\frac{k}{c_u})^{1-\alpha}}{1 - \frac{c_u}{c}}. \tag{2}$$

Let p be the number of physical machines in the cluster, m be the memory on each node used by the popular index, E be the size of an index entry, D be the amount of unique data on each physical machine, and C be the average chunks size. We store the popular index using a distributed shared memory hashtable such as MemCacheD. Then k and c_u can be expressed as: $k = p * m/E$, and $c_u = p * D/E$.

The overall deduplication efficiency of VC is

$$\frac{(1-\delta) + \delta(\frac{m*C}{D*E})^{1-\alpha}}{1 - \frac{c_u}{c}}.$$

where $(\frac{m*C}{D*E})^{1-\alpha}$ represents the percentage of the remaining chunks detected as duplicates after inner VM deduplication. Figure 3 shows measured PDS coverage in our 35 VM dataset. You can see from the graph that the coverage actually increases as the datasize increases, which indicates that α increases with the datasize (a fixed-α model would predict constant coverage). This makes the PDS even more effective, as the coverage of the PDS increases as more data is added.

c	the total amount of data chunks
c_u	the total amount of unique fingerprints after inner VM deduplication
f_i	the frequency for the ith most popular fingerprint
δ	the percentage of duplicates detected in inner VM deduplication
σ	the number of unique non-PDS chunks over the number of the PDS chunks.
p	the number of machines in the cluster
D	the amount of unique data on each machine
C	the average data chunk size. Our setting is 4K.
s	the average size of file system blocks in the distributed file system. The default is 64MB.
m	memory size on each node used by VC
E	the size of an popular data index entry
N_1	the average number of non-PDS file system blocks in a VM
N_2	the average number of PDS file system blocks in a VM
N_o	the average number of file system blocks in a VM for VO

Table 1: Modeling parameters and symbols.

As illustrated in Figure **??***I am not sure where this reference should point, but I don't think it is pointing to the right figure*, when the number of machines at each cluster increases, the number of total VMs increases. Then k increases since more memory is available to host the popular chunks index. But for each physical machine, the number of VMs remains the same, and thus D is a constant. Then the overall deduplication efficiency of VC remains a constant.

4.3 Storage Space and Impact on Fault Tolerance

The replication degree of the backup storage is r for regular file blocks and $r = 3$ is a typical setting in the distributed file system [**?**, **?**]. In the VC approach, a special replica degree r_c used for PDS blocks where $r_c > r$. Notice that the ratio of non-PDS data size vs PDS data size for each VM is

$$\sigma = \frac{c * \delta(1 - (\frac{k}{c_u})^{1-\alpha})}{k}.$$

Thus storage cost for VO with full deduplication is $c_u * r$ and for VC, it is

$$k * r_c + k * \sigma * r.$$

In our experiment with Alibaba data, the ratio σ is 162. Thus allocation of extra replicas for PDS only introduces a small amount of extra space cost. Figure **??** shows the storage cost ratio of VC and VO when r=3, and r_c varies from 3 to 10. The result shows that the storage cost for adding extra replication for PDS is insignificant.

Next we compare the impact of losing d machines to the

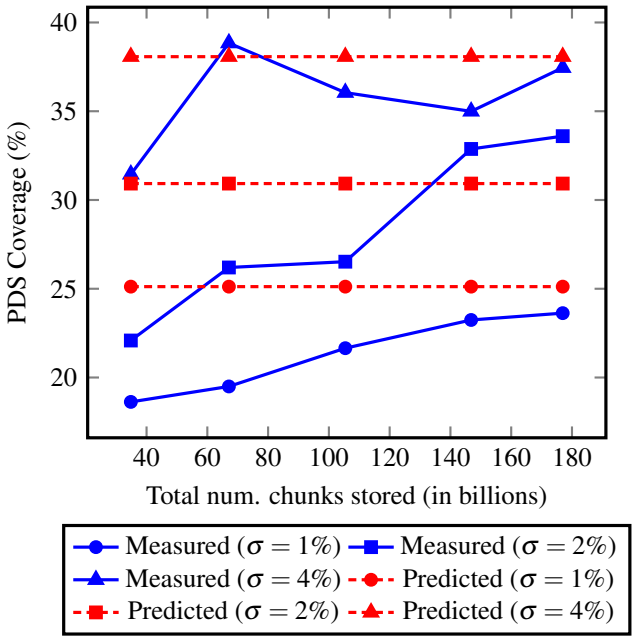


Figure 3: fixed-alpha predicted vs. actual PDS coverage as data size increases.

the VC and VO approaches.

In characterizing the reliability of VM backups in our model, we consider the likely hood that a file system block fails, given some number of storage machine failures. Every time a filesystem block fails, we say that we have lost data for that virtual machine, so it is no longer available. In reality it may be only one snapshot that is affected, but it is the user who must decide which snapshots are important, so we consider the worst case. We use filesystem blocks rather than a deduplication data chunk as our unit of failure because the DFS keeps filesystem blocks as its base unit of storage.

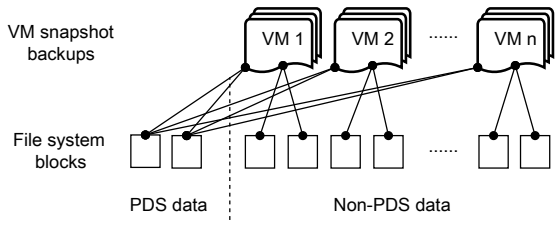
To compute the probablity of losing snapshots of a virtual machine, we estimate the number of file system blocks per VM in each apporach. We can build a bipartite graph representing the association from unique file system blocks to their corresponding VMs. An association edge is drawn from a file block to a VM if this file is used by this VM. For VC, each VM has an averge number of N_1 file system blocks for non-PDS data. It also refers an average of N_2 file system blocks for PDS data For VO, each VM has an average of N_o file system blocks and let V_o be the averge number of VMs shared by each file system block. Figure **??** illustates the bipartite association.

In VC, each non-PDS file system block is associated with one VM while PDS file system blocks are shared among VMs (at most V VMs). Thus,

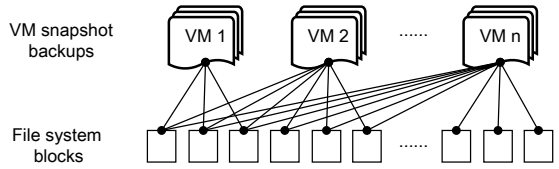
$$V * N_1 * s = pD\frac{\delta}{\delta+1} \text{ and } V * N_2 * s \leq pD\frac{1}{\delta+1} * V.$$

For the VO approach,

$$V * N_o * s = pDV_o.$$



(a) Sharing of data under VM-oblivious dedup model



(b) Sharing of data under VM-centric dedup model

Figure 4: Difference of sharing data under VO and VC approaches

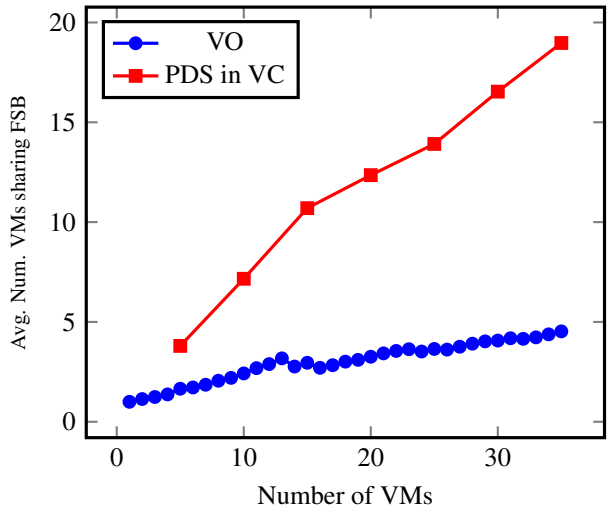


Figure 5: Measured Average number of VMs sharing a 64MB FSB with global dedup (VO), and in a 2% PDS for VC.

Then

$$N_1 = \frac{pD\delta}{Vs(\delta+1)},\ N_2 \leq \frac{pD}{s(\delta+1)},\ \text{and}\ N_o = \frac{pDV_o}{sV}.$$

Since each file block (with default size $s = 64MB$) contains many chunks (on average 4KB), each file block contains the hot low-level chunks shared by many VMs, and it also contains rare chunks which are not shared. Figure 5 shows the number of VMs sharing by each file block. In our experiment, we find that $V_o \approx 0.2V$ when backing up VMs one by one. We can observe in Figure 6 that $N_1 + N_2 < N_o$. This is likely because the PDS FSBs tightly pack data used by many VMs, which decreases the overall number of FSBs required to backup a VM. If the backup for multiple VMs is conducted concurrently, there would be more VMs shared by each file block on average.

Figure 5 shows the average number of VMs sharing a filesystem block (FSB) as VMs are added. Though we don't expect the linear trend to continue indefinitely for much larger datasets, it should continue to increase, which has important implications on VM backup availability in the precense of failures. Below we show the significant impact the number of links has on VM backup reliability.

The snapshot availability of is a VM is the likelyhood that there is no data loss for all its file blocks. With replication degree r, the likelyhood of a file block block is the probability that all of its replicas appear in d failed machines. Namely, $\binom{d}{r}/\binom{p}{r}$. This can be seen in Figure 7 for a 100 machine cluster for 3 different replication factors.

When there are $r \leq d < r_c$ machines failed and then there is no PDS data loss, the snapshot avaiability of a VM in the VC approach is and is

$$(1 - \frac{\binom{d}{r}}{\binom{p}{r}})^{N_1}.$$

When $r_c \leq d$, both non-PDS and PDS file blocks in VC can have a loss. The snapshot availability of a VM in the VC approach is

$$(1 - \frac{\binom{d}{r}}{\binom{p}{r}})^{N_1} * (1 - \frac{\binom{d}{r_c}}{\binom{p}{r_c}})^{N_2}.$$

The snapshot availability of a VM in the VO approach is

$$(1 - \frac{\binom{d}{r}}{\binom{p}{r}})^{N_o}.$$

Figure 8 shows the reliability of VM backups as storage nodes fail for different amounts of sharing of filesystem blocks in the index. We show 5 and 20 for the VO and VC block links counts because those are what we measured in our small dataset, and chose higher numbers to account for much larger datasets. We chose to use a very optimistic setting for VO (20), and the absolute worst case amount of sharing for VC in a 100 machine cluster with 25 VMs per machine (2500). Our results show that even the VC worst case provides better VM reliability that VO. As you can see from the graph, even small changes in the number of links to VO blocks has a significant impact on reliability, while even a 50 times increase in the VC block link counts only slightly affects VC reliability. The key factor placed is that $N_1 + N_2 < N_o$, caused by the fact that the VM-centric approach localizes deduplication and packs data blocks for one VM as much as possible. The extra replicaton for PDS blocks also signficantly increases the snapshot availability even when a PDS file block is shared by every VM.

Figure 9 shows the advantages of increasing the replication factor for PDS blocks. the numbers in the graph are only meant to be relative, as we assume higher PDS block links than we have measured in our dataset (to account for a much larger body of data). It is easy to see though that increasing

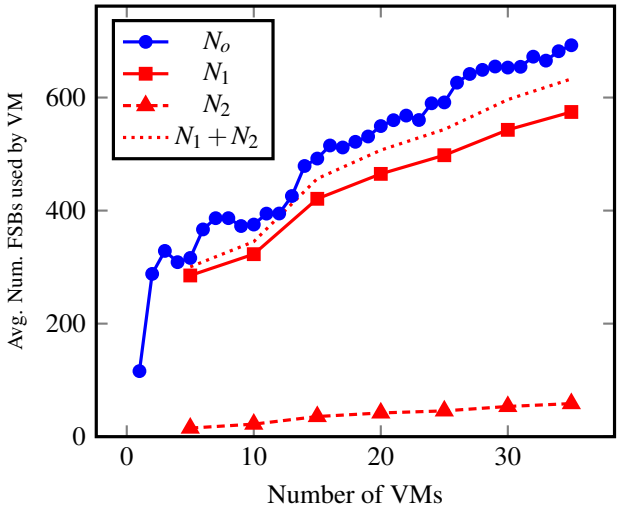


Figure 6: Measured Average number of 64MB FSBs used by a single VM. For VC both the number of PDS and Non-PDS FSBs used are shown.

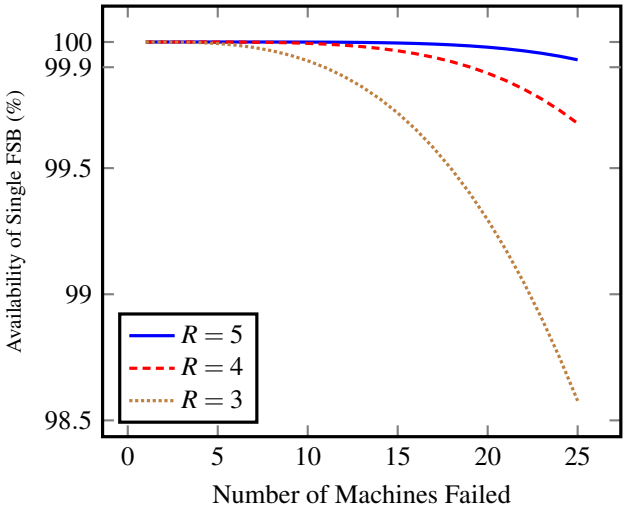


Figure 7: Availability of Individual FSBs in 100 Machine Cluster with different replication factors.

Nodes Failed	Availability(%)		
	$R=3$	$R=4$	$R=5$
1	100	100	100
2	100	100	100
3	99.999	100	100
5	99.993	99.9999	99.999998
10	99.926	99.995	99.999
20	99.295	99.876	99.979

Table 2: Availability in a 100 machine cluster of a single FSB in the global data store with different replication factors

the replication of just the PDS blocks (which are the most popular blocks) can have a positive impact on the overall reliablity of the VM backups. These figures together show the advantages of the VM-centric model, and the advantages that separating the replication factor for popular blocks can have on reliability.

5. ARCHITECTURE AND IMPLEMENTATION DETAILS

Our system runs on a cluster of Linux machines with Xen-based VMs. A distributed file system (DFS) manages the physical disk storage and we use QFS [**?**]. All data needed for VM services, such as virtual disk images used by runtime VMs, and snapshot data for backup purposes, reside in this distributed file system. One physical node hosts tens of VMs, each of which access its virtual machine disk image through the virtual block device driver (called TapDisk[9] in Xen).

5.1 Components of a cluster node

As depicted in Figure **??**, there are four key service components running on each cluster node for supporting backup and deduplication: 1) a virtual block device driver, 2) a snapshot deduplication component, 3) an append store client to store and access snapshot data, and 4) a CDS client to support CDS index access. We will further discuss our deduplication scheme in Section **??**.

We use the virtual device driver in Xen that employs a bitmap to track the changes that have been made to virtual disk. When the VM issue a disk write, the bits corresponding to the segments that covers the modified disk region are set, thus letting snapshot deduplication component knows these segments must be checked during snapshot backup. After the snapshot backup is finished, snapshot deduplication component acknowledges the driver to resume the dirty-bits map to a clean state. Every bit in the bitmap represents a fix-sized (2MB) region called *segment*, indicates whether the segment is modified since last backup. Hence we could treat segment as the basic unit in snapshot backup similar to file in normal backup: a snapshot could share a segment with previous snapshot it is not changed. As a standard practice, segments are further divided into variable-sized chunks (average 4KB) using content-based chunking algorithm, which brings the opportunity of fine-grained deduplication by allowing data sharing between segments.

The representation of each snapshot has a two-level index data structure. The snapshot meta data (called snapshot recipe) contains a list of segments, each of which contains segment metadata of its chunks (called segment recipe). In a snapshot recipes or a segment recipe, the data structures includes reference pointers to the actual data location.

5.2 A VM-centric snapshot store for backup data

We build a snapshot storage on the top of a distributed file system. Following the VM-centric idea for the purpose

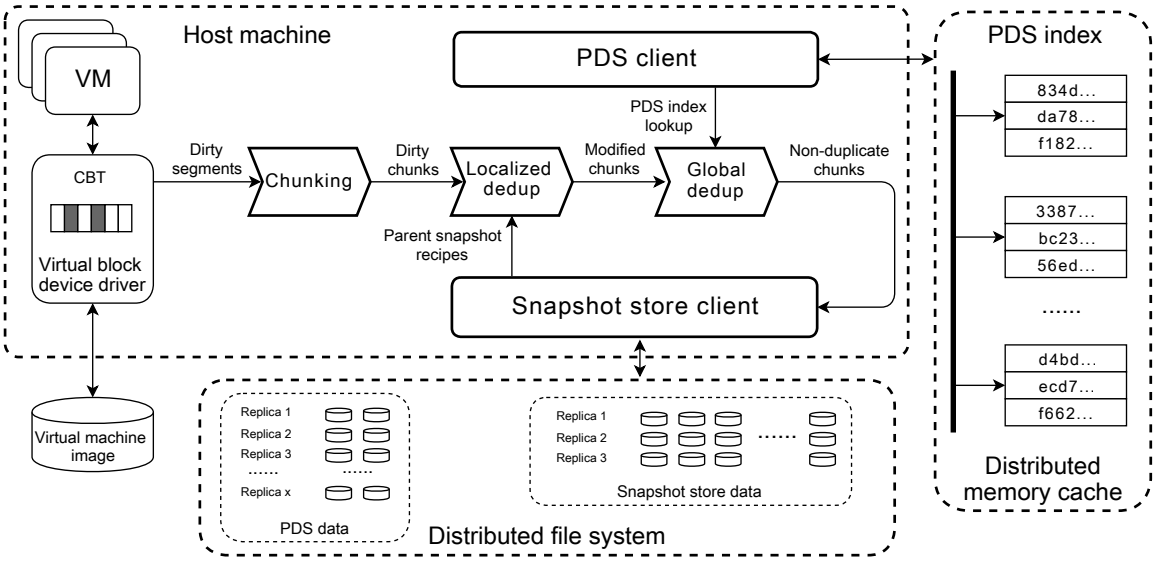


Figure 11: System Architecture

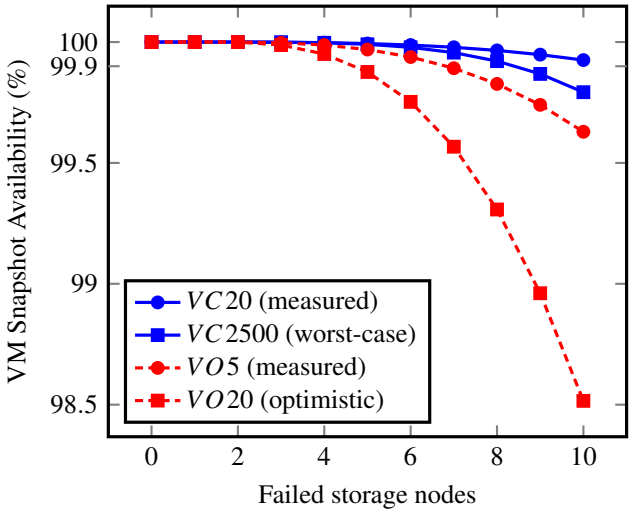


Figure 8: Availability of VM backups as nodes fail for VO and VC models for varying degrees of block sharing (i.e. average number of VMs that use a block). Non-PDS replication fixed at 3 and PDS replication increased to 4. Graph shown for 100 machine cluster.

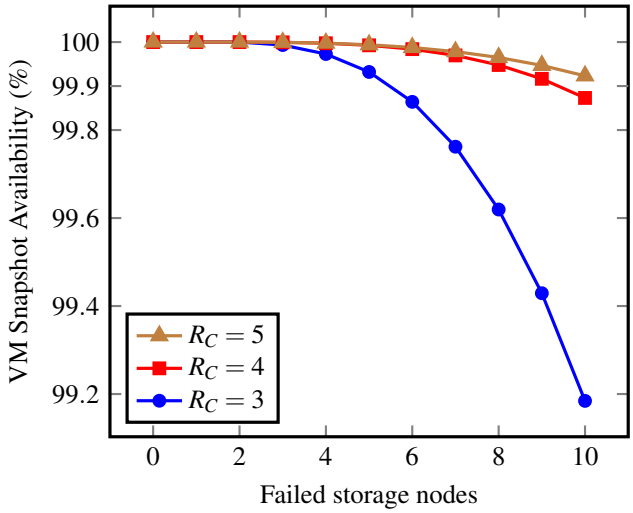


Figure 9: Availability of VM backups as nodes fail in the VC model for different PDS Replication factors (Non-PDS replication fixed at 3, and average PDS block links set to 1000)

of fault isolation, each VM has its own snapshot store, containing new data chunks which are considered to be non-duplicates. There is also a special store containing all PDS chunks shared among different VMs. The replication degree of file blocks of snapshot stores in the underlying file system Extra replication of this store is added as discussed in Section **??**. As shown in Fig.12, we explain the data structure of snapshot stores as follows.

- The PDS snapshot contains a set of commonly used data chunks and is accessed by its offset and size in the corresponding DSF file. The PDS index uses the offset and size as a reference in its index structure.

- Each non-PDS snapshot store is divided into a set of containers and each of them is approximately 1GB. The reason we divide the snapshot into containers is to simplify the compact process conducted periodically. There are chunks deleted without any reference from other snapshots and to reclaim the space of these useless chunks, a compaction routine can work on one container at a time and copy used data chunks to another container.

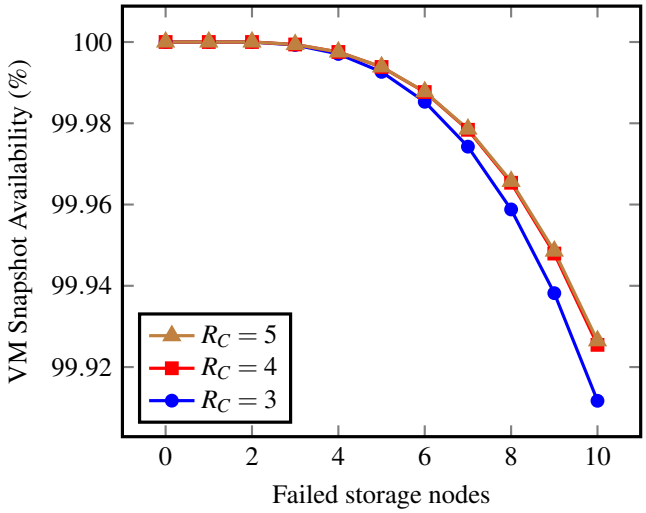


Figure 10: Availability of VM backups as nodes fail in the VC model for different PDS Replication factors (Non-PDS replication fixed at 3, and average PDS block links set to 20)

– Each container is divided into a set of chunk data groups. Each group is composed of a set of data chunks and is the basic unit for the snapshot in data access and retrieval. In writing a chunk group, data chunks in this group are compressed together and then stored. When accessing a particular chunk, its chunk group is retrieved from the disk storage and uncompressed. Given the high spatial locality in snapshot data accessing [**?**, **?**], retrieval of data chunks by group naturally fits in the prefetching scheme to speedup snapshot access. A typical chunk group contains 100 to 1000 chunks, with an average size of 200-600KB. Chunk grouping also reduces the container index size as we discuss below. Given the average chunk size 4KB, the index size for a 1GB container reduces from 10MB to 100KB when the chunk group size is 100.

– Each data container is represented by three data files in the DFS: 1) the container data file holds the actual content of data chunks, 2) the container index file is responsible for translating a data reference into its location within a container, and 3) a chunk deletion log file saving all the deletion requests within the container. A VM snapshot store typically has a small number of containers because each container is fairly large with an average size of 1GB, it maintains a limited number of snapshots (e.g. 10 in the Alibaba case), and new snapshot data chunks can be effectively compressed in chunk groups in addition to deduplication.

– We maintain a chunk counter and assign the current number as a chunk ID (called CID) within this container as a reference of a new chunk added to a container. Since data chunks are always appended to the snapshot store, a CID is monotonically increasing. A data chunk reference stored in the index of snapshot recipes is composed of two parts: a container ID (2 bytes) and CID (6 bytes). Once a snapshot is to be accessed access, the receipt in this snapshot will point to either a data chunk in the PDS or a reference number with a container ID and CID. With a container ID, the corresponding container index file is accessed and the chunk group is identified using this CID. Once this chunk group is loaded to memory, its header contains the exact offset of the corresponding chunk and the content is then accessed from the memory buffer.

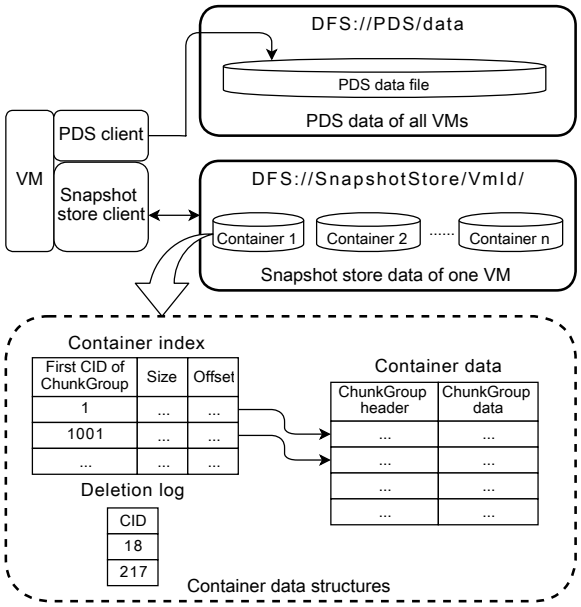


Figure 12: Data structure of VM snapshot stores.

The snapshot store supports three API calls.

- *Put(data)* places data chunk into the snapshot store and returns a reference to be stored in the recipe metadata of a snapshot.

  The write requests to append a data chunk to a VM store are accumulated in the client side. When the number reaches the group size g, the snapshot store client compresses the accumulated chunk group, adds a chunk group index in the beginning, and then append the header and data to the corresponding VM file, Then a new container index entry is also created and is written the corresponding container index file.

  The writing of PDS data chunks is conducted periodically when there is a new PDS calculation. Since the

PDS dataset is small, a new PDS file is created during the periodical update.

- *Get(reference).* The fetch operation for the PDS data chunk is straightforward since each reference contains the offset and size within the PDS underlying file. We also maintain a small data cache for the PDS data service to speedup the process.

  To read a non-PDS chunk data by a reference, the snapshot store client first loads the corresponding VM's container index file specified by the container ID, then searches the chunk group that covers the chunk by the group CID range. After that, it reads the whole chunk group from DFS, decompresses it, seeks the exact chunk data specified by the CID. Finally, the client updates its internal chunk data cache with the newly loaded content to anticipate future sequential reads.

- *Delete(reference).* A data chunk can be deleted when a snapshot expires or gets deleted explicitly by a user. We will discuss the snapshot deletion shortly in the following subsection. When deletion requests issued for a specific container, those requests are logged into the container's deletion log file initially and thus a lazy deletion strategy is exercised. Once CIDs appear in the deletion log, they will not be referenced by any future snapshot and can be safely deleted when needed. Periodically, the snapshot store picks those containers with an excessive number of deletion requests to compact and reclaim the corresponding disk space. During compaction, the snapshot store creates a new container (with the same container ID) to replace the existing one. This is done by sequentially scan the old container, copying all the chunks that are not found in the deletion log to the new container, creating new chunk groups and indices. However, every chunk's CID is plainly copied rather than re-generated. This does not affect the sorted order of CIDs in new container, but just leaving holes in CID values. As the result, all data references stored in upper level recipes are unaffected, and the data reading process is as efficient as before.

5.3 VM-centric Approximate Snapshot Deletion with Leak Repair

In a busy VM cluster, snapshot deletions are as frequent as snapshot creations. The VM-centric snapshot storage design simplifies the deletion process since we need to look for the useless chunks within each VM when deleting a snapshot. The PDS data chunks are commonly shared all VMs and we donot consider their reference counting during snapshot deletion. The selection of PDS data chunks is updated periodically independent of snapshot deletion process.

While we can use the standard mark-sweep technique [**?**], it takes time to conduct this process every time there is a snapshot deletion request. In the case of Alibaba, snapshot backup is conducted automatically and there are about 10 snapshot stored for every user. Then when there is a new snapshot created every day, there will be a snapshot expired everyday to maintain a balanced storage use. Given a large number of snapshot deletion requests, we are seeking a fast solution with a very low resource usage to delete snapshots with an approximation.

With this in mind, we develop an *approximate* deletion strategy to trade deletion accuracy for speed and resource usage. Our method sacrifices a tiny percentage of storage leakage to effectively identify unused chunks in $O(n)$ speed, with n being the logical number of non-PDS chunks to be deleted from a VM snapshot store. The algorithm contains three aspects.

- **Computation for snapshot fingerprint summary.** Every time there is a new snapshot created, we compute a bloom-filter fingerprint summary for all non-PDS chunks used in this snapshot.

- **Approximate deletion with fast summary comparison.** Then when there is a snapshot deletion, we use the snapshot fingerprint summary to identify if chunks to be deleted from one snapshot are still used by other snapshots. This is done approximately and quickly by comparing the fingerprints of deleted chunks with the fingerprint bloomer-filter summary of other live snapshots. Since the number of live snapshots is limited for each VM (e.g. 10 in the Alibaba's production system), the complexity of comparison is acceptable. However, there is a small false-positive ratio which would identify unused data as in use, resulting in temporary storage leakage.

- **Periodic repair of leakage** Since there are certain unused chunks which are not deleted during the approximated deletion, the leakage repair is conducted periodically. In this phase we load the metadata of all containers from a VM snapshot store, mark if any chunk is used. For unused chunks within this VM, such chunks are to be deleted. This process is identical to the previously-developed mark-and-sweep techniques except that we only need to conduct a VM-specific mark-and-sweep operation.

We estimate how often the leak repair needs to be conducted. Let L be the amount of temporary storage leakage caused by our approximate snapshot deletion, H be the average size of non-PDS snapshot data that should be deleted. ε be the false-positive rate of the merged bloom filter. R be the number of runs using the approximate deletion before a full leak repair is conducted. Then $L = R * \varepsilon * H$.

When ratio L/H exceed a threshold τ, then:

$$\frac{L}{H} = R * \varepsilon > \tau \Rightarrow R > \frac{\tau}{\varepsilon}.$$

For example, when $\varepsilon = 0.01$ and $\tau = 3$, we would expect deletion phase-2 be triggered once for every $T/P_{bl} = 300$

runs of fast deletion. When one machine hosts 25 VMs and there is one snapshot deletion per day per VM, there would be only one full leak repair scheduled for every 12 days, which is sufficiently small to prevent the heavy I/O workload of the mark-sweep process.

6. EVALUATION

We have implemented and evaluated a prototype of our VO scheme on a Linux cluster of 8-core AMD FX-8120 at 3.1 GHz with 16 GB RAM. Our implementation is based on Alibaba's Xen cloud platform [1, **?**]. Each machine is equipped with a a distributed file system (QFS) running in the cluster manages six 3TB disks with default replication degree set to 3. The CDS replication is set to 5. Objectives of our evaluation are: 1) Study the deduplicate detection and removal effectiveness of the VC approach and compare with an alternative VO design. 2) Examine the impacts of VC for fault isolation and tolerance. 3) Evaluate the backup throughput performance VC for a large number of VMs. 4) Assess the resource usage of VC in terms of memory, storage, and network bandwidth.

We will also compare our VC approach with with a VO approach using stateless routing with binning (SRB). SRB executes a distributed deduplication by routeing a data chunk to a machine [**?**] using the chunk hashing function discussed in [**?**]. Within such a machine, the data chunk is compared with a local index.

6.1 Settings

We have performed a trace-driven study using a 1323 VM dataset collected from 100 Alibaba Aliyun's cloud nodes [**?**]. The production environment tested has about 1000 machines with 25 VMs on each machine. For each VM, the system keeps 10 automatically-backed snapshots in the storage while a user may instruct extra snapshots to be saved. The backup of VM snapshots is completed within a few hours every night. Based on our study of its production data, each VM has about 40GB of storage data usage on average including OS and user data disk. All data are divided into 2 MB fix-sized segments and each segment is divided into variable-sized content chunks [6, 8] with an average size of 4KB. The signature for variable-sized blocks is computed using their SHA-1 hash. Popularity of data blocks are collected through global counting and the top 1% will fall into CDS, as discussed in Section **??**.

Since it's impossible to perform large scale analysis without affecting the VM performance, we sampled two data sets from real user VMs to measure the effectiveness of our deduplication scheme. Dataset1 is used study the detail impact of 3-level deduplication process, it compose of 35 VMs from 7 popular OSes: Debian, Ubuntu, Redhat, CentOS, Win2003 32bit, win2003 64 bit and win2008 64 bit. For each OS, 5 VMs are chosen, and every VM come with 10 full snapshots of it OS and data disk. The overall data size for this 700 full snapshots is 17.6 TB.

Dataset2 contains the first snapshots of 1323 VMs' data disks from a small cluster with 100 nodes. Since inner-VM deduplication is not involved in the first snapshot, this data set helps us to study the CDS deduplication against user-related data. The overall size of dataset2 is 23.5 TB.

6.2 Deduplication Efficiency

Figure 13 shows the deduplication efficiency for SRB and VC. We define deduplication efficiency to be the percent of duplicate chunks which are detected and deduplicated. The figure also compares several PDS sizes chosen for VC. "$\sigma = 2\%$" means that we allocate the space of distributed shared memory to accommodate 2% of data chunks shared among VMs, as defined in Table 1. Namely With $\sigma = 2\%$, the deduplication efficiency can reach over 90% while with $\sigma = 4\%$, the deduplication efficiency can reach over 92%. The loss of efficiency in VC is caused by the restriction of the total physical memory available in the cluster to support the PDS. The loss of efficiency in SRB is caused by the routing of data chunks which restricts the search scope for global comparison. In all cases, VC provides similar or better deduplication efficiency than SRB.

This result shows VC can remove a competitive amount of duplicates. In general, our experiments show that dirty-bit detection at the segment level can reduce the data size to about 23% of original data, which leads about 77% reduction. Local search within a segment can further reduce the data size to about 18.5% of original size, namely it delivers additional 4.5% reduction. The PDS-based cross-VM detection with $\sigma = 2\%$ can reduce the size further to 8% of original size, namely it delivers additional 10.5% reduction.

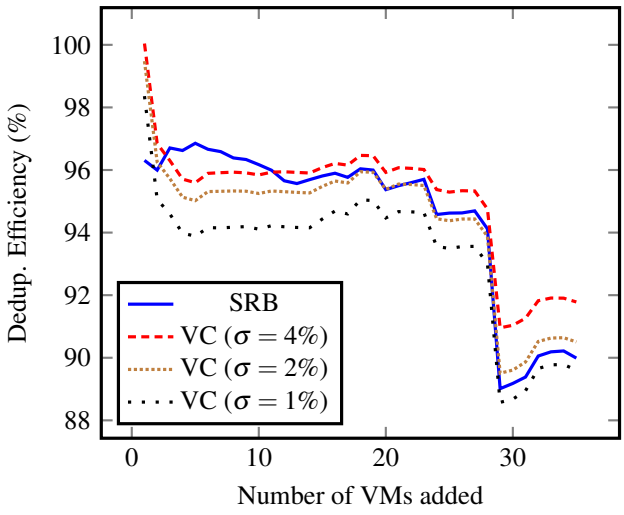


Figure 13: Deduplication Efficiency ($efficiency = \frac{d-du}{d-du_{complete}}$) **comparison between SRB and VC.**

6.3 Resource Usage

Storage cost of replication.

PDS replication degree	Total space (GB)
3	3065
4	3072
5	3079
6	3086
7	3093

Table 3: Storage space cost under different PDS replication degree

Table 3 shows the total storage space required by VC as the PDS replication degree is changed. The increase in storage cost is minimal because the PDS makes up a very small percent of the total storage space, but the increase in reliability can be much more significant (see Figure 9).

6.4 Processing Time and Throughput

Figure 14 shows the average processing time of each segment using VC and SRM in handling our dataset. It has a breakdown of time for reading data and updating the metadata, network latency to visit a remote machine, and index access for finger printer comparison. SRB has a higher index access and fingerprint comparison because once a chunk is routed to a machine, it relies on this local machine to access its index (often on the disk) and perform comparison. VC is faster for index access because it conducts in-memory local search first and then accesses the PDS which is on distributed shared memory. SRB spends slighter more time in reading data and update because it also updates the on-disk local meta data index. Overall speaking, VC is faster than SRB while data reading dominates the processing time.

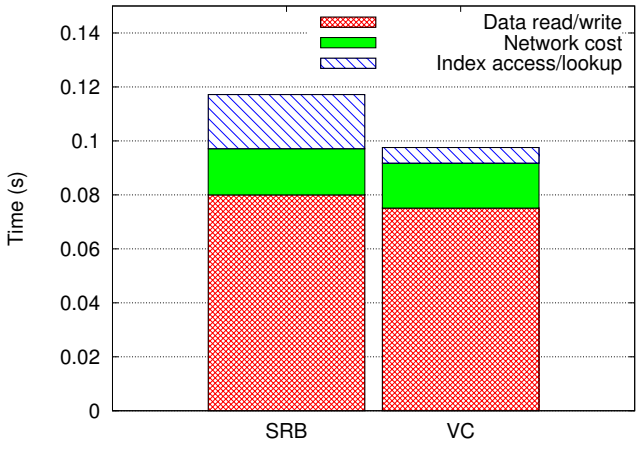


Figure 14: Time to backup a dirty segment under SRB and VC approach

Figure 15 reports the average bacup time for a VM with VC with a varying number of PDS size. Again it shows that data reading dominates the backup process, our deduplication procedure only takes a small fraction of the total backup

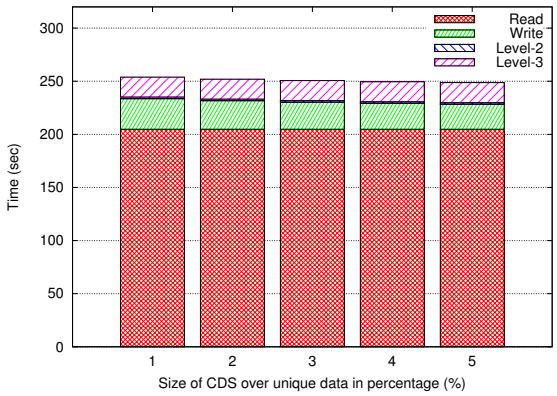


Figure 15: Backup times for varying CDS sizes

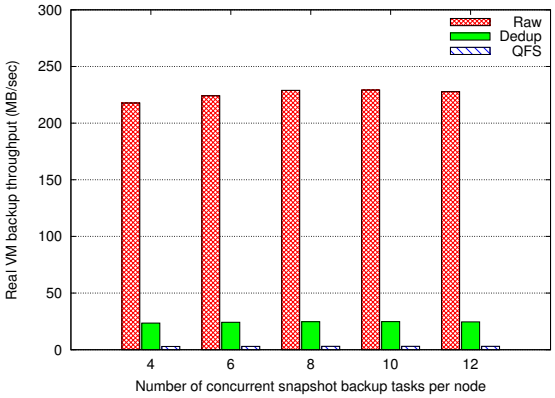
time. The change of δ does not significantly affects the overall time and PDS lookup takes a small amount of time.

Figure 16 shows the node throughput when all machine nodes run the backup procedure in parallel. To begin, on each node we let 4 VMs writing snapshot concurrently, and gradually increase number of VMs to 12 to saturate our system capability. We observed the per-node throughput peaked at 2700 MB/s when writing 10 VM snapshots in parallel, which is far beyond our QFS file system capability. The reason behind it is our efficient deduplication architecture and compression greatly reduce the amount of data that need to be written to the file system. The main bottleneck here is our QFS only manages one disk per node, making it inefficient to utilize the benefits of parallel disk access. We expect our architecture can perform even better in production cluster which normally has more than ten disks on each node.

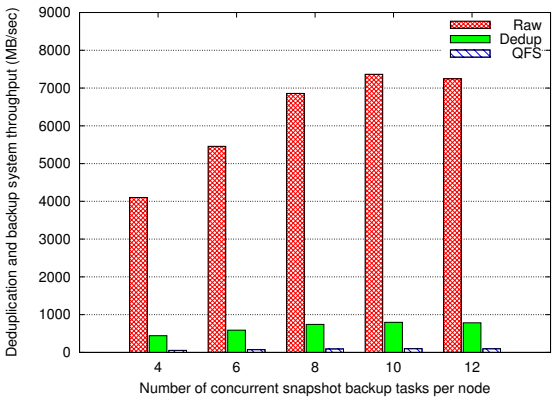
7. CONCLUSION

In this paper we propose a VM-centric deduplication scheme for snapshot backup in VM cloud for maximizing fault isolation and tolerance. Inner-VM deduplication localizes backup data dependency and exposes more parallelism while cross-VM deduplication with a small common data set effectively covers a large amount of duplicated data. The scheme organizes the write of small data chunks into large file system blocks so that each underlying file block is associated with one VM for most of cases. Our solution accomplishes the majority of potential global deduplication saving while still meets stringent cloud resources requirement. Our analysis shows that this VM centric scheme can provide better fault tolerance while using a small amount of computing and storage resource.

[Talk about more what we learn from Evaluation] Evaluation using real user's VM data shows our solution can accomplish 75% of what complete global deduplication can do. Compare to today's widely-used snapshot technique, our scheme reduces almost two-third of snapshot storage cost. Finally, our scheme uses a very small amount of memory on each node, and leaves room for additional optimization we



(a) Real VM backup performance



(b) Deduplication and storage system performance

Figure 16: Throughput per-node with concurrent snapshot backup tasks

are further studying.

8. REFERENCES

[1] Aliyun Inc. http://www.aliyun.com.

[2] D. Bhagwat, K. Eshghi, D. D. E. Long, and M. Lillibridge. Extreme Binning: Scalable, parallel deduplication for chunk-based file backup. In *Modeling, Analysis Simulation of Computer and Telecommunication Systems, 2009. MASCOTS '09. IEEE International Symposium on*, pages 1–9, 2009.

[3] A. T. Clements, I. Ahmad, M. Vilayannur, and J. Li. Decentralized deduplication in SAN cluster file systems. page 8, June 2009.

[4] F. Guo and P. Efstathopoulos. Building a high-performance deduplication system. page 25, June 2011.

[5] M. Lillibridge, K. Eshghi, D. Bhagwat, V. Deolalikar, G. Trezis, and P. Camble. Sparse Indexing: Large Scale, Inline Deduplication Using Sampling and Locality. In *FAST*, pages 111–123, 2009.

[6] U. Manber. Finding similar files in a large file system. In *Proceedings of the USENIX Winter 1994 Technical Conference*, pages 1–10, 1994.

[7] S. Quinlan and S. Dorward. Venti: A New Approach to Archival Storage. In *FAST '02: Proceedings of the Conference on File and Storage Technologies*, pages 89–101, Berkeley, CA, USA, 2002. USENIX Association.

[8] M. O. Rabin. Fingerprinting by random polynomials. Technical Report TR-CSE-03-01, Center for Research in Computing Technology, Harvard University, 1981.

[9] A. Warfield, S. Hand, K. Fraser, and T. Deegan. Facilitating the development of soft devices. page 22, Apr. 2005.

[10] J. Wei, H. Jiang, K. Zhou, and D. Feng. MAD2: A scalable high-throughput exact deduplication approach for network backup services. In *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*, pages 1–14, May 2010.

[11] B. Zhu, K. Li, and H. Patterson. Avoiding the disk bottleneck in the data domain deduplication file system. In *FAST'08: Proceedings of the 6th USENIX Conference on File and Storage Technologies*, pages 1–14, Berkeley, CA, USA, 2008. USENIX Association.