WHITE PAPER

CloudByte™
Enterprise Storage for Virtualization

Moving metadata to Flash memory

Improve the I/O performance by effective metadata management

Keeping metadata in flash accelerates the storage access manifold for all kinds of workloads. This also enables faster RAID rebuilds in the case of drive failure.

Flash (SSD) is used to store all metadata. Metadata is normally around 10% of the overall data. However, it is highly variable depending on the nature of data and the size of the filesystem block sizes. Normally, storage systems try to keep metadata in the Cache/RAM.
It is highly impossible to keep all the metadata in the Cache for relatively larger workloads. Having the metadata in flash accelerates the storage access manifold for all kinds of workloads. This also enables faster RAID rebuilds in the case of drive failure.

Flash is used to store deduplication table. This brings out two distinct advantages. By keeping the deduplication table in flash, you free up precious RAM space for active data. Secondly, deduplication performance goes really bad when deduplication table overflows into regular spinning disk. By keeping deduplication table in flash, chance of overflow is eliminated and deduplication performance remains consistent.

Flash is used as additional READ Cache. Normally, for any dataset, only 10% of the overall data is active at any given point in time. Therefore, providing 10% as flash Cache accelerates the performance for most of the workloads.
Flash is used as WRITE Cache. In ElastiStor, WRITEs always go into flash, thereby providing a significant boost for WRITE workloads. Data gets flushed into spinning disks at regular intervals.

Scenarios where metadata on SSDs are helpful

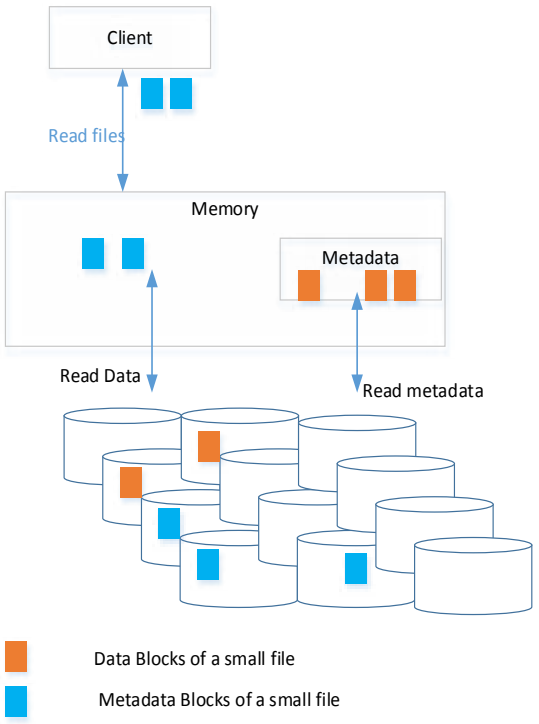
When the storage pool contains large number of small files

When designing the storage pools for archival data or backup data, the size of the storage pool is typically high.

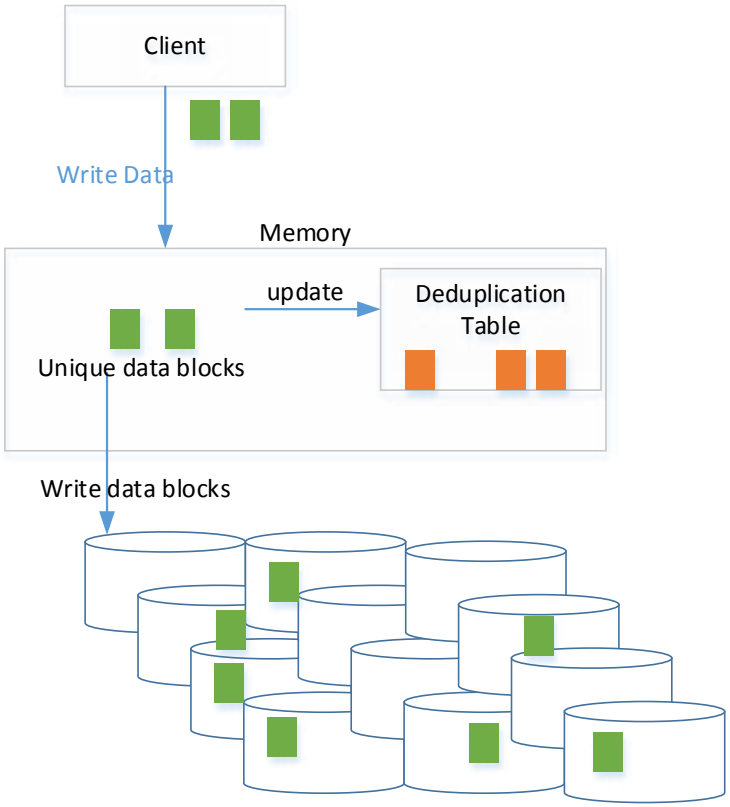
The size can range from 100TB to few Petabytes. At these sizes, it is common to design the pools using slow performance but high capacity disks such as 4TB NLSAS or 6TB NLSAS disks. The READs to the disk happen at the slowest speed.

Some data storage patterns have very large number of small files. The size of these files is so small that the metadata associated with these files are significant in size compared to the actual size of the files.

This results in generating significantly additional reads to disks in huge amounts slowing down the performance. CloudByte suggests creating such storage pools with metadata on SSDs.



When deduplication is required over large pool



Deduplication table

Unique data blocks

Deduplication works on the principle of keeping the signature of a data block in a hash table and incrementing the references to the signature when a similar data block is written into the storage. This signature hash table is typically called *deduplication table*.

The deduplication table is maintained in the memory for performance reasons. As the data size grows in the pool, the number of unique blocks also would grow and so does the size of deduplication table.

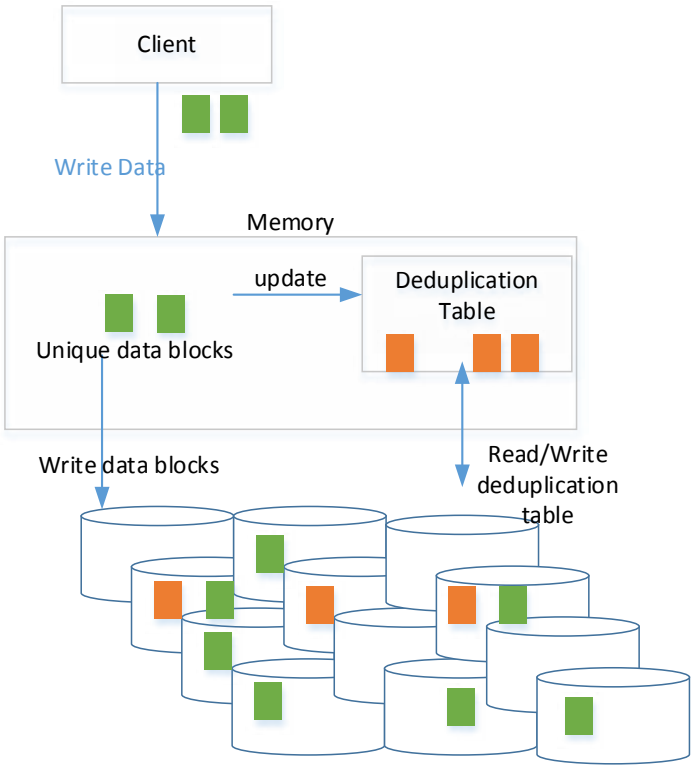
Assume the size of the deduplication table grows beyond what can be fit into the memory. The deduplication table extends into the slower HDD.

Typically, when the storage data size crosses 100TB, the deduplication table can cross 200GB. The memory cannot hold the entire deduplication data and starts extending to the slower disks.

From this point onwards, the performance of the storage can go down significantly as for each WRITE into storage, a READ and WRITE have to happen to the slower disk to update the deduplication table as shown in the following figure.

The deduplication table is considered a part of the metadata. In both the above scenarios, it can be said that, the metadata reading from the slower disks caused the performance issues.

CloudByte resolves this problem by enabling the storage pools to have the metadata on a separate SSD cache.



Deduplication table

Unique data blocks

Moving metadata to flash helps faster RAID rebuilds

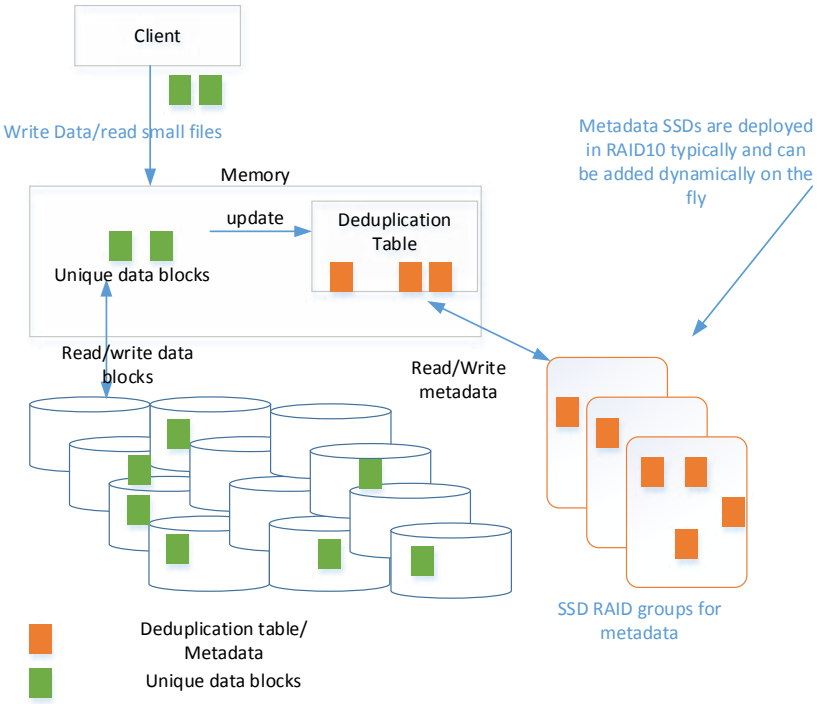
If storage pools contain slower and larger capacity disks, RAID rebuilds take longer time. This is one of the deterrent while choosing higher capacity disks such as 7.2K RPM 6TB disks.

RAID rebuild involves scanning of complete metadata from the pool. This is to identify the blocks that have to be recovered from the lost disk. For a fully populated storage pools (in the worst case), this can run into several days.

With metadata on a separate group of SSDs, metadata scanning is faster and the RAID rebuild time is cut down by more than 50 percent.

With metadata on a separate group of SSDs, metadata scanning is faster and the RAID rebuild time is cut down by more than 50 percent.

Option to have metadata on the SSD



CloudByte ElastiStor has the option to create storage pools with the storage pool metadata on a separate SSD cache. This cache is called metadata cache or metadata accelerator.

As metadata cache is very critical to the viability of the storage pool, the SSD cache is stored in RAID mirrors.

Metadata cache can be started with fewer RAID mirror groups and can be added dynamically on the fly.

Other types of RAIDs such as RAID50, RAID60 are also supported.

How much metadata cache is needed ?

Typically one to two percent of the storage pool capacity.

This may go up to four percent. For a storage pool size of 100TB, the SSD metadata cache needs to be sized upto 200GB. The actual metadata size depends the size of the files and also in the case of deduplication it depends on the number of unique blocks.

If the percentage of unique blocks is high (greater than 50 percent), deduplication is not advisable and can result in negative economic results.

Moving metadata to flash: benefit summary

You realize the following benefits by placing the metadata on a separate set of SSDs:

- Higher READ performance when the work load consists of large number of smaller files

- Higher WRITE and READ performance when the deduplication is in effect on pool sizes of greater than 100TB

- Faster RAID rebuilds and hence can take advantages of higher capacity disks available in the market

WHY CLOUDBYTE ELASTISTOR

CloudByte ElastiStor Appliance (ESA) extends server virtualization to the storage component. With ESA, one can spin out Virtual Storage Machines (VSM) like VMs in the server world. VSMs abstract all the resources of the storage system such as CPU, disk I/O, network, and cache – since storage is not just disks.
The storage performance can be moved between VSMs on the fly without application disruptions; true software defined storage. Each storage volume under a VSM guarantees performance at a granular level, such as 2345 IOPS or 234 IOPS.
This solution can scale across multiple data centers by adding additional appliances. The complete solution can be managed with a single global management console.

ABOUT CLOUDBYTE

CloudByte is the leading provider of enterprise storage for the virtual environment. Its patented software defined storage architecture solves storageI/O contentions in the virtual environment and provides granular storage performance guarantees for each application. Established in 2011 and managed by technology executives from companies such as NetApp, EMC, LSI, and Novell, CloudByte is headquartered in the Silicon Valley and has a development center in India. CloudByte is venture-backed by Fidelity Worldwide Investment, Nexus Venture Partners, and Kae Capital.

CloudByte™
Enterprise Storage for Virtualization

info@cloudbyte.com | www.cloudbyte.com

20863 Stevens Creek Boulevard, Suite 530, Cupertino, CA 95014, USA | +1-855-380-BYTE (2983)
Plot No. 2799 & 2800, Srinidhi Bldg, 3rd Floor, 27th Main, Sector – 1, HSR Layout, Bangalore 560102, India +(91)-80-2258-2804