HOST BASED STORAGE PERFORMANCE TIER FOR VMWARE TO ACCELERATE THE PERFORMANCE OF SAN BASED STORAGE APPLIANCES

Reasons for storage I/O bottleneck in virtualized data centers

In virtualized data centers storage I/O to shared storage is often a bottleneck. This is especially so if applications like relational databases, analytics, virtual desktops, search engines, and other throughput intensive or latency sensitive applications are deployed within virtual machines.

The reasons for this are listed below:

a. Disk speeds and latencies have not kept up with improvements in CPU, Memory, network and Bus speeds.
b. Storage array controllers are a bottleneck as well. Typically a single pair of redundant controllers fronts a large number of disks. And there are only a limited number of disk drives and limited amount of I/O bandwidth that such a controller pair can support.
c. Increasing adoption of virtualization at enterprise Datacenters has resulted in greater utilization of server resources, resulting in increased I/O to networked storage.
d. I/O blender - In virtualized servers, by the time sequential I/O from each VM gets through the hypervisor, it gets interspersed with I/O from other VMs on the same server, and so the I/O pattern out of the virtualized server becomes mostly random. Since disk performance deteriorates rapidly with random I/O, this reduces storage performance further.

The reason for our focus on VMware

In the case of virtualized server hardware (versus bare-metal servers), because of higher application density on each physical server, storage performance issues are more pronounced, especially for data intensive applications deployed within VMs. Since VMware has the dominant share of the virtualization market, it was obvious that VMware be our platform of choice.

Using Flash to solve the storage I/O bottleneck

Flash is the media that makes up the Solid State Drive (SSD). Flash is ideal for solving storage throughput and latency issues for random workloads, as is the case with VMware workloads. However since Flash is much more expensive than traditional Hard Drives, it needs to be used in smaller quantities to make the storage solution cost effective. Hence the need for caching software that backs slower Disk based local or shared storage with smaller quantities of faster Flash based storage.

Use of caching software with Flash

Caching software in Operating Systems has been in use for decades. We decided to apply these same time tested concepts to the relatively newer domain of virtualized servers that use networked storage, with in-server Flash as the caching media.

Virtunet Systems' Virtucache Improves Storage Performance of VMware

VirtuCache is a Kernel mode software for VMware that clusters together any in-Host SSDs installed across VMware Hosts in a VMware cluster and then caches frequently and recently used data from any SAN based primary storage appliance to this clustered pool of Host based SSDs. Subsequently, by automatically serving more and more data from in-Host SSDs, VirtuCache substantially improves storage performance for VMware from our customer's existing storage appliance, thus improving the performance of applications running within VMs and increasing the density of VMs running on each Host, without requiring an expensive upgrade to SSD based storage appliances.

Virtunet Systems, 4010 Moorpark Ave, Suite 205, San Jose, Ca 95117; info@virtunetsystems.com ; phone : (510) 315 8829

HOST BASED STORAGE PERFORMANCE TIER FOR VMWARE TO ACCELERATE THE PERFORMANCE OF SAN BASED STORAGE APPLIANCES

Accelerating Reads

All read requests from the VMs on the Host are intercepted by VirtuCache software in the VMware Kernel. VirtuCache first looks up the local SSD for this data. If the data is in the SSD, it is served to the VM from the SSD (called 'cache hit'). If the data is not in the SSD, the I/O path proceeds along its original course, and VMware retrieves the data from the backend LUN/Volume. At that point VirtuCache copies the data to the local SSD as well. Subsequently if the same data is requested again by any VM on the Host, it is now served from the local SSD, instead of from the backend storage appliance. In this way VirtuCache accelerates read operations by serving up more and more data from in-Host SSDs.

Accelerating Writes

All writes from VMs on the Host are written to the local SSD without synchronously writing to the backend storage appliance. By writing to the in-server SSD, writes are substantially accelerated, however the fact that we are not synchronously committing the writes to the backend storage appliance introduces the risk of data loss/corruption in case the local Host or SSD were to fail. To guard against this possibility, VirtuCache protects the local cache by replicating/mirroring the writes across Hosts in a VMware cluster.

Syncing 'Dirty' Writes to backend storage

Dirty Writes are writes on the local SSD cache that have not yet been synced with backend storage. VirtuCache has a background task that continuously syncs Dirty Writes to the backend SAN based storage. VirtuCache adjusts the speed and frequency at which Dirty Writes are synced based on the latencies exhibited by the SAN and appliance, so as not to choke the SAN by trying to sync to the backend appliance quickly. Also, at no point in time will more than a few minutes of Dirty Writes be stored on the local SSD. This is to avoid large amounts of Dirty Writes following the VM during a vMotion.

Cache Replication to protect against local Host or SSD failure

One of the main benefits of clustering SSDs across VMware Hosts is being able to mirror the cache across VMware Hosts in a distributed fashion. The administrator specifies the number (0, 1, or 2) of copies of cache that need to be kept for each local cache in the cluster. The number of copies indicates the maximum number of node failures that can be sustained before there is data loss in the cluster. If a customer chooses to keep, say, 2 copies of cache for each local Host based cache, VirtuCache automatically replicates the dirty writes across two SSDs on two VMware Hosts in the same cluster. We default to using the vMotion network for such replication. However a separate network can be configured as well. Reads are not replicated since the backend storage appliance is always consistent as far as reads go. In the event of a Host or SSD failure, VirtuCache syncs the backup copy of the dirty write cache from another Host to the backend storage appliance.

Flow control to prevent write intensive VMs from taking over the SSD

Since the SSD capacity deployed within VMware Hosts is typically a small percentage of the total LUN capacity of the backend storage appliance, care needs to be taken to prevent write intensive VMs from taking over the entire SSD. VirtuCache allows bursty writes from VMs to be written to the SSD at native SSD write speeds without synchronously syncing the data to the backend disk. However for prolonged write intensive activity from VMs, VirtuCache's flow control feature throttles back the write speeds to the SSD. This helps ensure fair allocation of SSD capacity to other VMs on the Host and ensures orderly de-staging of writes from the SSD to the backend LUN.

HOST BASED STORAGE PERFORMANCE TIER FOR VMWARE TO ACCELERATE THE PERFORMANCE OF SAN BASED STORAGE APPLIANCES

Keeping the cache 'fresh'

We use a combination of Least Recently Used (LRU) and First-in-First-Out (FIFO) algorithms to replace less frequently used older data with newer data in cache, much like how traditional Operating Systems have been using these algorithms for Disk-to-Memory caching.

Virtucache Differentiators Versus Other Host Side Caching Vendors

VirtuCache is the highest performance Host side SAN acceleration solution on the market. It also has the lowest management overhead. In the sections below, we explain the reasons for both:

1. High Performance

a. We accelerate both read and write operations. Most of our competition accelerates only reads. The fact that our competition does not accelerate writes means that in their case even small amount of writes choke the read operations that are behind the writes on the same thread. Hence with their software, not only are writes not accelerated but even read latency and throughput is not improved to the extent that we do.

b. VirtuCache is installed in the Hypervisor layer of VMware. There are no Userspace or VM level software components that are in the I/O path. We are unique in the VMware space because we intercept block I/O requests and make all caching decisions from within the VMware Kernel. The fact that there is no storage IO traversing the Userspace-Kernelspace boundary makes our software much higher performance than our competition who require a virtual appliance per Host or agents installed in VMs.

c. Ours is a block level caching software. Block level caching is faster than file level caching, since it works at a lower level in the software stack than the file system.

d. We have shown 5-20X IO acceleration with TPC-C, TPC-H, Sqlio, Iometer, and Fio benchmark tests, and 3-7X improvement in VM densities at customers. The faster the in-Host SSD and slower the SAN, the higher is the performance multiplier with VirtuCache.

2. Minimal Administrative Overhead

a. We think that automatic caching of data with no human involvement is the ideal way to do caching. This has been validated by the fact that most operating systems have been successfully caching Disk data to Memory in an automated fashion for the past few decades.

Similarly we automatically cache VM level storage IO and use time tested LRU and FIFO algorithms to keep the cache fresh.

On the other hand, there are caching solutions on the market that let administrators define caching policies at the application and file level. In theory, such granularity for defining caching policies looks impressive, however it becomes a huge management burden for the server administrator to intelligently assign caching policies on a per VM and per application basis, even in a small data center with hundreds of VMs. The server administrator will need to understand the applications deployed within VMs, their relative importance and I/O patterns, for devising intelligent caching policies. Also application level caching in many cloud service

provider environments is not possible, since VMs are deployed by end customers themselves and the service provider has no visibility within guest VMs. Unlike our competition, we do not need a dedicated virtual appliance VM per VMware server, or any software installed within guest VMs.

b. Again unlike our competition, we do not need storage for each VM reconfigured. We do not need VMs or the Host restarted. VirtuCache operation is also seamless to the customer's existing SAN based storage architecture, in the sense that end users and applications running within VMs do not realize that the data is being read from and written to the local SSD, instead of the backend storage appliance.

c. In a virtualized data center, important applications are typically isolated within VMs. Despite being in the Kernel we can correlate I/O to VMs. Consequently, VirtuCache can accelerate specific applications by enabling caching for only the VMs that the applications are installed in.

d. I/O de-duplication – In Linked Clone deployments, we cache parent VM blocks that are repeatedly being requested by multiple desktop VMs only once, thus conserving SSD usage. Since OS and application installation files are duplicated across virtual machines in VMware Horizon View VDI deployments, this feature results in high cache hit ratios by using relatively small SSD capacities.

e. Without any administrative overhead, we automatically support VMware's advanced features – VAAI, Snapshots, DRS, Linked Clones, vMotion, High Availability, and Storage vMotion.

3. VMware Certified

One of the biggest challenges in developing Kernel mode software for Host side storage acceleration for VMware is to do it in such a way so as to be able to get the software blessed by VMware, which meant using publicly available APIs from VMware. Since VMware has no publicly available APIs for Kernel mode Host side SAN caching use case, most of our competition has taken the easier route of developing such software using a VM based approach, either requiring a VM per Host (Virtual Storage Appliance) or requiring agents in guest VMs.

To satisfy VMware's certification requirement, we developed our Kernel mode caching driver as a Path Selection Plug-in (PSP) for VMware Native Multipath Plug-in (NMP), using VMware's publicly available Multi-Pathing Plug-In framework.

This does require that the storage vendor supplied MPP be replaced with VMware's own NMP. For the majority of customers this is not a concern since VMware NMP is the default MPP used by them. In customer situations where the customer insists on using the storage vendor provided MPP (most popular of which is EMC's PowerPath), we have a parallel implementation of our solution which is also a Kernel mode software solution that is below the MPP layer and it can coexist with storage vendor provided MPPs.

VirtuCache is now certified as Partner Verified and Supported as attested by the below link.
http://kb.vmware.com/kb/2116221

Summary

Our main design goals were to make VirtuCache high performance, zero overhead, and VMware certified.

HOST BASED STORAGE PERFORMANCE TIER FOR VMWARE
TO ACCELERATE THE PERFORMANCE OF SAN BASED STORAGE APPLIANCES

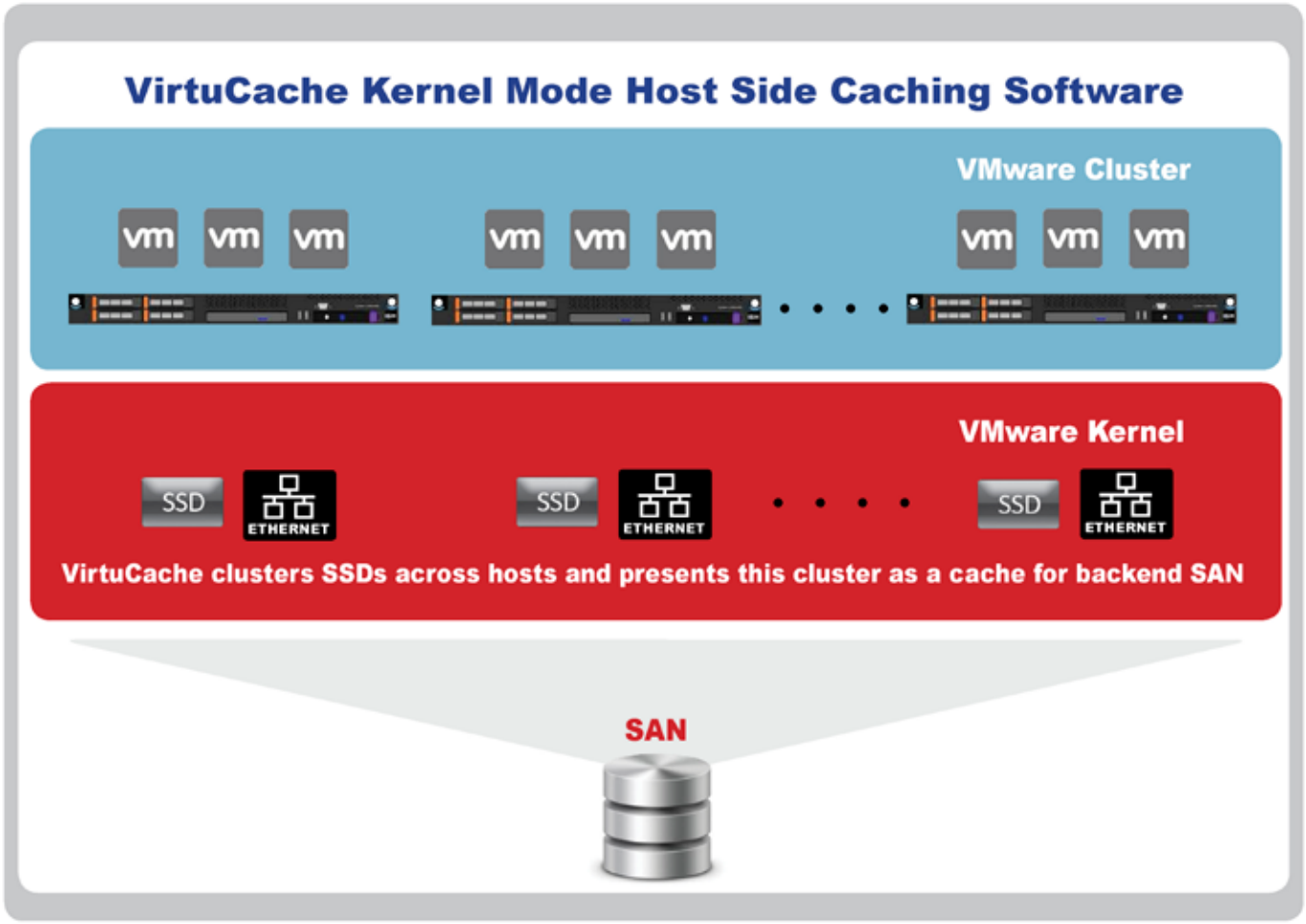

Figure 1: VirtuCache deployment across a VMware cluster

VirtuCache consists of two software components – Driver software installed in the Hypervisor layer of VMware vSphere Hosts that need to be accelerated, and one VirtuCache Management VM per vCenter to manage all the VirtuCache Driver software instances installed in Hosts managed by that vCenter instance.