LOSING MY RELIGION: VIRTUALIZATION BACKUP DOGMA, FAITH, AND FACT

CTO Series: Dr. Mark Campbell,

Chief Strategy/Technology Officer, Unitrends

CTOSERIES

INTRODUCTION

"Losing my religion" is an expression from the southern region of the United States that means being perplexed to the point of being unable to think what to do[1]. Despite the tremendous advantages that virtualization brings to IT professionals, many find themselves at their wits' end with respect to protecting ever more prevalent virtualized environments. The cacophony of competing vendor claims as well as the claims of their paid consultants only increases the confusion and attendant frustration in those simply seeking to an optimal solution for protecting their unique IT infrastructure.

There is a war of words going on right now between data protection vendors who offer data protection within the virtual machine versus those who offer data protection at the hypervisor level. These vendors illustrate specific use cases that offer advantages to their methods while ignoring those use cases in which they have a disadvantage. The best example is a vendor who had a well-respected consultant write a paper that he entitled

> "VMware and Hyper-V Backups: How VM-Level *Can Be Better* than Host-Level Backup." (italics ours.)

[1] It's also the title of a song by a band called REM that I particularly like – but that's not germane to this white paper.

This carefully written white paper thoughtfully explores specific use cases in which VM-level backups could potentially be better than host-level backups. The vendor on their web site then retitled this paper

> "VMware and Hyper-V Backups: How VM-Level *Is Better* than Host-Level." (again, italics ours.)

Given that this same consultant writes for a leading host-level vendor and writes similar papers that explore advantageous use-cases for host-level protection, it is difficult to believe that the difference here is accidental or coincidental.

In this technology brief we are going to cut through the self-serving and competing claims and examine each of the arguments in favor and against the various techniques used for virtual data protection.

VIRTUALIZATION HYPERVISOR PROTECTION ARCHITECTURES

A generic virtualization architecture is depicted in figure 1.

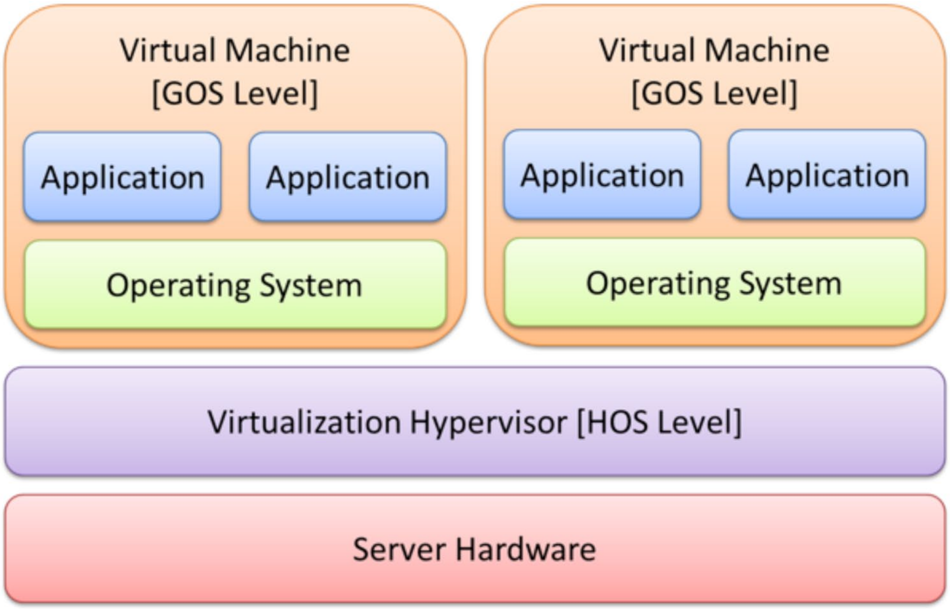
Virtual Machine [GOS Level]
Application
Application
Operating System
Virtual Machine [GOS Level]
Application
Application
Operating System
Virtualization Hypervisor [HOS Level]
Server Hardware

Figure 1: Generic Virtualization Architecture

From a data protection perspective, what's important to note here is that protection may occur at the virtualization hypervisor (called the HOS (Host Operating System) level) and within the virtual machine (called the GOS (Guest Operating System) level.) GOS-level protection is unique to the operating system and applications being protected but is independent of the virtualization hypervisor architecture and implementation.

Microsoft Hyper-V HOS-Level Protection

Microsoft uses a data protection architecture known as VSS (Volume Shadow Copy Service) to protect their operating systems, applications, and their virtualization. VSS at the operating system and application level is used not only by Microsoft, but by other virtualization vendors (for example, VMware) to make sure that the data being used by Microsoft operating systems and applications is in a consistent state so that recovery is insured (this is also called "quiescing.") However, Microsoft as a virtualization vendor also uses VSS at the HOS-level as well.

Compared to VMware's HOS-level protection, Microsoft's HOS-level VSS protection is a bit lower level. What this means is that data protection vendors must write software for missing functionality when offering Microsoft HOS-level virtualization protection. The most prominent example of this is CBT (Changed Block Tracking) – which is functionality that Microsoft doesn't offer within VSS but which VMware offers within its HOS-level protection architecture. From an IT administrator and user perspective, however, this isn't visible.

VMware vSphere HOS-Level Protection

After a series of mis-steps culminating in the "clunky" VCB (VMware Consolidated Backup) offering, VMware came back strong beginning in VMware vSphere 4 with its VADP (vStorage API for Data Protection) data protection architecture. VADP is the leading data protection architecture in the virtualization market today and has advanced functionality such as CBT built-in so that vendors can offer data protection with less effort.

Is there any downside to VMware's VADP? Yes. VMware limits access to their API set to only licensed versions of their hypervisor. In other words, their free (unlicensed) ESXi product doesn't support it. Thus vendors are forced to provide GOS-level protection in this case. This compares poorly to Microsoft, for example, which has no such limitation on its free Hyper-V Server 2012 or Hyper-V Server 2008 versions of its Hyper-V virtualization platform.

Citrix XenServer HOS-Level Protection

Beginning with its XenSever 5.5 release, Citrix began offering

XenServer snapshots. Snapshots provide a "point in time" disk state that can be used by data protection vendors. While the details of using snapshots vary based on the type of storage being used and are beyond the scope of this white paper, it suffices to say at a conceptual level XenSever snapshots are simply an implementation of the tried and true snapshot mechanisms other storage and virtualization vendors have used.

Because of their common heritage, quite often people get XenServer and Xen functionality confused. Note that the data protection functionality described above is a feature of XenServer, not of the Xen open source virtualization platform.

Other Hypervisors HOS-Level Protection

Most other virtualization vendors either don't offer a data protection architecture or offer an extremely limited one. In this case, it is typically recommended that protection occur at the GOS-level or via scripts that are written that will execute at the HOS-level which quiesce the virtual environment, take the virtual machines off-line, back them up, and then bring them back online.

VIRTUALIZATION PROTECTION APPROACHES

There are two basic virtualization protection approaches

- GOS (Guest Operating System)-level protection. This refers to protection within each virtual machine. The hallmark of this approach is that the protection is unaware that it's in a virtual environment and treats the virtual machine identically to a physical machine.
- HOS (Host Operating System)-level protection. This refers to protection of each virtual machine in aggregate at the virtualization hypervisor level. The hallmark of this approach is that the data protection is aware that it's in a virtual environment and thus protects a collection of virtual machines.

In the sections that follow we will discuss each of these in more detail.

GOS-Level Protection: Treating the Virtual Machine as a Physical System

GOS-level protection means that the data protection system treats the virtual machine exactly as if it were a physical system. In other words, the backup software that protects the virtual machine is typically placed within the virtual machine within the operating system just as it is in a physical environment.

Proponents of GOS-level protection are typically those vendors who have only a GOS-level solution.   GOS-level solutions typically offer more granularity than HOS-level solutions but have significant limitations concerning ease of use and automation as virtual environments dynamically grow. GOS-level solutions can also have issues protecting hypervisor-level features such as virtualization-based clustering. A detailed analysis of the various capabilities of GOS-level protection will be discussed in the next chapter.

HOS-Level Protection: Treating the Virtual Machine as a Collection of Virtual Machines

HOS-level protection means that the data protection system identifies the physical system upon which the virtualization hypervisor resides and treats it as a virtual host. In other words, the backup software that protects each virtual machine is typically placed within the hypervisor. Another name for this approach is "virtualization aware", i.e., the backup software is aware of virtualization and protects one or more virtual machines within the overall virtualization environment.

Proponents of HOS-level protection are typically those vendors who have only an HOS-level solution. HOS-level solutions typically offer better overall ease of use and automation but have limitations concerning granularity. A detailed analysis of the various capabilities of GOS-level protection will be discussed in the next chapter.

DOGMA, FAITH, AND FACT: DO YOU HAVE TO PICK A VIRTUAL BACKUP RELIGION?

Faith is a set of personal beliefs while dogma is a set of beliefs that are compulsory for a particular sect. What in the world does dogma, faith, and fact have to do with virtualization and data protection? In theory, absolutely nothing – most IT professionals strive to be fact-based and largely succeed. However, what vendors often do is attempt to sway potential buyers through emotion. Catchy advertising appealing to virtualization lovers or cloud lovers[2] are attempts to sway IT professionals away from facts through an emotional appeal.

I've noticed that backup vendors are beginning to do the same thing with respect to the method by which they approach the data protection

[2] I always wanted to see a backup vendor create a marketing campaign targeted to puppy lovers; to me, it makes as much sense as most advertising in our space.

market. If a vendor offers only GOS-level protection, then that is offered by the vendor as the end-all be-all solution for everything – and HOS-level protection causes everything from bad breath to global warming. Conversely, if a vendor offers only HOS-level protection, then that is offered by the vendor as the ultimate panacea – and of course GOS-level protection becomes the root cause of male pattern baldness as well as the national debt.

In this chapter we're going to explore the primary advantages and disadvantages of GOS-level protection versus HOS-level protection. We're going to do so without the hype and without the dogma by focusing on the fact-based capabilities of each.

Agent-Based and Agentless Virtual Data Protection

Before we begin to analyze capabilities in detail, let's quickly discuss the trend toward using the terms "agent-based" and "agentless" data protection. The terms "agent-based" and "agentless" with respect to virtual data protection is relatively meaningless despite the incredible number of white papers, e-books, podcasts, and keynote speeches expounding on the evils or one and the beneficence of another. Why? The reason is because most modern virtualization protection uses agents of some kind to protect virtual environments.

A backup agent is software that resides within the virtual machine (the GOS) that helps protect the virtual machine. Both GOS- and HOS-level protection schemes tend to use backup agents to differing degrees. As we've discussed previously, GOS-level protection treats each virtual machine as if it were a physical machine – thus the backup agent in GOS-level protection is completely and utterly responsible for backing up all of the data.   However, in HOS-level protection the backup agent doesn't work alone but rather in concert with the GOS-level protection software to protect the virtual machine.

So when you see the terms agent-based and agentless, just substitute GOS-level protection and HOS-level protection for those terms and you'll be one step ahead of getting past industry jargon and to the unvarnished truth.

Ease of Use

Ease of use is a two-edged sword. GOS-level protection can be and often is easier to use for a single virtual machine. Unfortunately, GOS-level protection doesn't tend to scale to virtual infrastructure and does not handle multiple virtual machines very well. The paradigm of treating virtual machines the same as physical machines has a tendency to void all of the advantages that virtualization consolidation of physical servers brings to the table.

Many leading GOS-level protection vendors don't allow scheduling of virtual machines at one time which drives your operational involvement, and thus your TCO (Total Cost of Ownership), higher. In addition, many leading GOS-level protection vendors don't integrate with aggregation points such as vCenter thus driving your TCO higher.

Since we're evaluating ease of use within virtual environments here, the advantage goes to HOS-level protection. It can be argued that GOS-level protection offers more flexibility in that both virtual and physical environments can be managed – this is discussed later this chapter.

Advantage: HOS-level protection

Automation and Inclusion

The ability to automatically detect new virtual machines and include in a pre-defined existing schedule is present from every major HOS-level data protection vendor but is lacking in most if not all GOS-level data protection products. The reason for this of course is that HOS-level protection protects at the host level where virtual machine creation (and deletion, for that matter) can be easily detected.

Because HOS-level protection can include newly created virtual machines, the advantage with respect to automation and inclusion easily goes to HOS-level protection.

Advantage: HOS-level protection

Heterogeneity

By its very nature, GOS-level protection is tied to the operating system and applications being protected. HOS-level protection protects at the level of the virtualization infrastructure and thus is abstracted from the operating system and application.

The fact that HOS-level protection is abstracted from the underlying

GOS-level operating system and application means that HOS-level protection has a significant advantage in heterogeneity versus GOS-level protection.

Advantage: HOS-level protection

Granularity

Granularity refers to the ability to select some object not to backup. That object can be files, directories, volumes, database, or any other object. GOS-level protection typically offers the maximum degree of granularity because those objects are seen natively through the specific operating system, file system, and application-oriented backup software.

HOS-level protection typically offers granular protection at the virtual machine level or at the virtual hard drive level. Since this can be offered by the GOS-level backup software by enabling/disabling it, GOS-level protection if implemented properly has an advantage over HOS-level protection.

One thing to be aware of is that there are block-level implementations of GOS-level protection that have many if not all of the same drawbacks as HOS-level protection. Beware if your vendor uses terms like "data is data."

Advantage: GOS-level protection (with a caveat that some block-based GOS-level vendors offer only limited granularity.)

NAS Protection

HOS-level protection does not typically offer protection of NASs in an environment as flexibly as GOS-based solutions. NAS protection in HOS-level protection tends to be restricted to only specifically configured NAS at the host level. GOS-level protection can typically offer NAS protection either through the servers that are attached to the NAS or via direct attachment.

The one caveat here is to insure that your GOS-level protection vendor offers these capabilities. There are GOS-level protection vendors using kernel block-based drivers who offer no way of doing NAS-based protection. If your GOS-level vendor starts talking about "Smart Agents", or "block-based filter drivers", or "image-level protection" then make sure to investigate this if NAS is important to you.

Nevertheless, correctly and flexibly implemented GOS-level protection offers a significant advantage over HOS-level protection with respect to protecting NASs.

Advantage: GOS-level protection (with a caveat that some block-based GOS-level vendors offer no NAS protection.)

Microsoft Hyper-V CSV (Cluster Shared Volume) Support

A CSV is a volume that is simultaneously available to directly read from and write to by all notes of a Microsoft failover cluster. CSVs enable an IT administrator to reroute data over the network in the event that a node loses its path(s) to the shared storage array.

HOS-level protection typically works with virtualization constructs such as CSV. GOS-level protection is implemented at a level lower than the virtualization infrastructure and thus depending upon implementation will not support virtualization constructs such as CSV. This is the case for one major vendor who released and then had to de-release CSV support (AppAssure.) Note that in "theory" GOS-level protection could support CSV; in practice it's proven a bit difficult for some vendors.

Advantage: HOS-level protection

VMware vCenter and VMware HA Support

VMware vCenter Server provides centralized management of vSphere virtual infrastructure. vCenter is an aggregation mechanism that allows IT administrators to gain centralized control and visibility into their vSphere deployment, provide proactive management, and manage multiple vCenter Server instances.

Data protection products that operate at the HOS-level are able to work directly with VMware vCenter and handle situations such as a virtual machine migrating from one ESX/ESXi host to another.

HOS-level protection typically works with virtualization constructs such as VMware vCenter Server. GOS-level protection is implemented at a level lower than the virtualization infrastructure and thus don't typically support virtualization constructs such as VMware vCenter Server.   Note however that GOS-level protection will work with other virtualization constructs, such as vMotion, that are implemented "underneath" the GOS such that GOS-level protection does not have to be aware of their existence.

Advantage: HOS-level protection

Validation

Validation refers to various techniques for ensuring that a backup is recoverable. Both HOS and GOS-level data protection vendors typically offer various types of audited instant system recovery techniques. In additional, GOS-level protection vendors who offer application-aware protection typically offer various types of integrity checking that go beyond the "data is data" mantra of block-only protection.

Advantage: Tied (vendor implementation is the critical deciding factor with respect to validation)

Backup Performance

File-based backup allows the greatest degree of flexibility and granularity; however, this flexibility and granularity comes at the cost of slower backup performance. Particularly on Windows-based file systems, the read speed of the file system can be much slower than block-based backup. Thus block-based backup tends to be faster than file-based backup.

HOS-level backups are always block-based. GOS-level backups can be either file-level, block-level, or both.

Make sure if you use block-based backups with applications that your application vendor supports it. For more information see "Application Support" later in this chapter.

Advantage: Tied (both GOS- and HOS-level backup if performed at a block level is faster than file-level backup.)

RPO: Recovery Point Objective

RPO is the maximum amount of data, measured in time, which you can afford to lose. Thus if you have an RPO of 5 minutes, that means you can afford to lose up to 5 minutes of data.

Vendors typically talk about RPO as a monolithic single entity for an IT infrastructure. But the truth is that RPO should be based on the discrete entities that constitute your IT infrastructure. There are going to be certain physical servers, virtual machines, applications, volumes, directories, or even files that are more critical than others.

Note that in the simplest case, you could just decide your entire IT infrastructure has an RPO of 1 minute (for example) and use HOS-level or GOS-level block protection. The problem tends to be the change rate of your data in your environment. If you have a low change rate and you know you'll always have a low change rate, then this non-granular approach to RPO works fine. But as your change rate increases, you find that you're not able to capture and transfer all of the changed data at once – and even if you are it takes a toll on your production environment. Thus it's best to optimize this through granular application-aware backup.

The ability to granularly define how often you backup an entity up is key to achieving an optimal RPO throughout your IT infrastructure.

Advantage: GOS-level protection (with the caveat that you're using a non-block-based application-aware backup and not just application-aware recovery)

RTO: Recovery Time Objective

RTO is the maximum amount of time that it takes to recover data, a system, or your entire environment. As long as your data protection solution offers modern instant recovery technology, your RTO will be measured in minutes whether you're using HOS- or GOS-level protection.

If you need an RTO in seconds, you should look into a virtualization-based high availability solution to augment your data protection solution.

Advantage: Tied

Application Support

HOS-level vendors tend to tout application-aware recovery – the reason is that they do not offer application-aware backup. GOS-level vendors who offer only block-level backup tend to do the same. The best application-level data protection solution is one which is not just application-recovery aware but also application-backup aware. Application-aware backup allows IT administrators the ability to change backup policies and to recover with a finer degree of granularity.

Also be aware that there are applications vendors which specify they will support you in your recovery attempts only if you use application-aware backup. For example, Microsoft specifically notes in their system requirements for Exchange 2010:

> *Some hypervisors include features for taking snapshots of virtual machines. Virtual machine snapshots capture the state of a virtual machine while it's running. This feature enables you to take multiple snapshots of a virtual machine and then revert the virtual machine to any of the previous states by applying a snapshot to the virtual machine. However, virtual machine snapshots aren't application aware, and using them can have unintended and unexpected consequences for a server application that maintains state data, such as Exchange. As a result, making virtual machine snapshots of an Exchange guest virtual machine isn't supported.*

A good conversation regarding this, with references to the Microsoft documentation, may be found at http://communities.vmware.com/message/2035867.

Note that this could be applied to either HOS- or GOS-level block backup. Check with your application vendor for more information.

Advantage: GOS-level protection (if application-aware backup and recovery are offered.)

Distributed Application Support (e.g., SharePoint Farms)

HOS-level protection is agnostic to distributed application support. GOS-level protection at the block level is also agnostic to distributed application support. Unfortunately, protecting distributed applications requires a backup solution that understands how to synchronize distributed applications. Thus an application-aware backup scheme must exist.

The advantage here goes to GOS-level protection if and only if application-aware backup exists.

Advantage: GOS-level protection (with the caveat that application agents exists.)

Reporting

GOS-level reporting can be performed at a per-object level and thus is much more granular than HOS-level reporting with respect to files, directories, volumes, databases, and other application objects. So if you're looking for sub-virtual machine object level reporting, GOS-level protection is superior.

However, GOS-level protection is blind to the virtualization infrastructure; thus GOS-level protection can't report on virtual machines and aggregations of virtual machines.

Advantage: Tied (depends on the type of reporting you want.)

Archiving

Archiving refers to making tertiary copies of data that are packaged together and transferred to some physically separate storage device. An easier way to think about this is that backup should be designed primarily for recovery and short-term retention while archiving should be designed primarily for longer-term retention. Archiving should be integrated into a data protection system and offer both rotational archiving strategies (e.g., disk and tape) as well as fixed archiving strategies (e.g., NAS, SAN, cloud.)

Neither GOS- nor HOS-level data protection confers an inherent advantage over the other in terms of archiving; instead, the implementation of archiving by the vendor is critically important.

Advantage: Tied (depends upon vendor implementation.)

Replication

In terms of replication, both GOS- and HOS-level data protection can do a fine job of transferring data. What's more important is the method by which replication is performed. The two things to watch out for with respect to replication are

- Primary/primary versus primary/secondary replication. Primary/primary replication means that both backup and replication occurs from the primary (or live) data. This means that replication will contend with user data access and possibly with backup as well. Primary/secondary replication means that backup occurs from the primary (or live) data and replication occurs from backup (secondary) data. Primary/

secondary replication is almost always recommended. This is independent of HOS- and GOS-level data protection.

- The granularity of objects that can be selected (or unselected) for replication. WAN bandwidth tends to be the primary issue with respect to replication – thus the ability to select what will and won't be replicated is critically important. GOS-level data protection tends to be more flexible and granular with respect to replication granularity.

Advantage: Tied with a nod toward granularity if WAN bandwidth is important (depends upon vendor implementation.)

Licensing and Pricing

Regarding licensing and pricing, the key questions to ask are

- ***Is more than one type of licensing available?*** Typically at least two are preferable. That way if you're IT infrastructure represents an end-case for a licensing methodology (for example, 1 server with a petabyte or 100 servers with only 5GB each to protect) you are more likely to have a choice as to which you'll use.
- ***What types of licensing are available?*** Typically it's better if both a resource-based licensing scheme (based on sockets, servers, applications, and the like) as well as a capacity-based licensing scheme (based on terabytes) are available. That way you can choose the licensing that best matches your environment and your anticipated growth in the future.

There is no specific advantage of GOS- versus HOS-level protection with respect to licensing and pricing because this is more a function of the vendor than it is whether GOS- or HOS-level protection is used.

Advantage: Tied

VMware Free (Unlicensed) ESXi Support

As noted previously, VMware free (unlicensed) ESXi doesn't support the use of VADP – thus HOS-level protection is not possible. GOS-level protection, where each virtual machine is treated as a physical machine, is the only option possible.

Advantage: GOS-level protection.

VMware RDM (Raw Device Mapping) in Physical Compatibility Mode

VMware's VADP doesn't support VMware RDM in physical compatibility mode. Thus GOS-level protection must be used to protect this particular configuration. Note that GOS-level protection that is block-based will not work in this situation; only a more flexible approach will work.

Advantage: GOS-level protection (if not implemented as block-level backup)

CONCLUSION

There's only one easy answer in terms of virtualization data protection – and that's to reject dogma and embrace facts in order to find the optimal data protection solution for your unique IT infrastructure. Vendors shouldn't ask you to conform to their data protection offerings; instead they should prove to you how their solution can flexibly adapt to your existing IT infrastructure as well as enabling you to respond to future needs and requirements of your business and your users in an agile manner.