White Paper

Backup and Recovery of Large Scale VMware environments

By Mark Bowker

October, 2010

This ESG White Paper was commissioned by EMC and is distributed under license from ESG.

Contents

Introduction

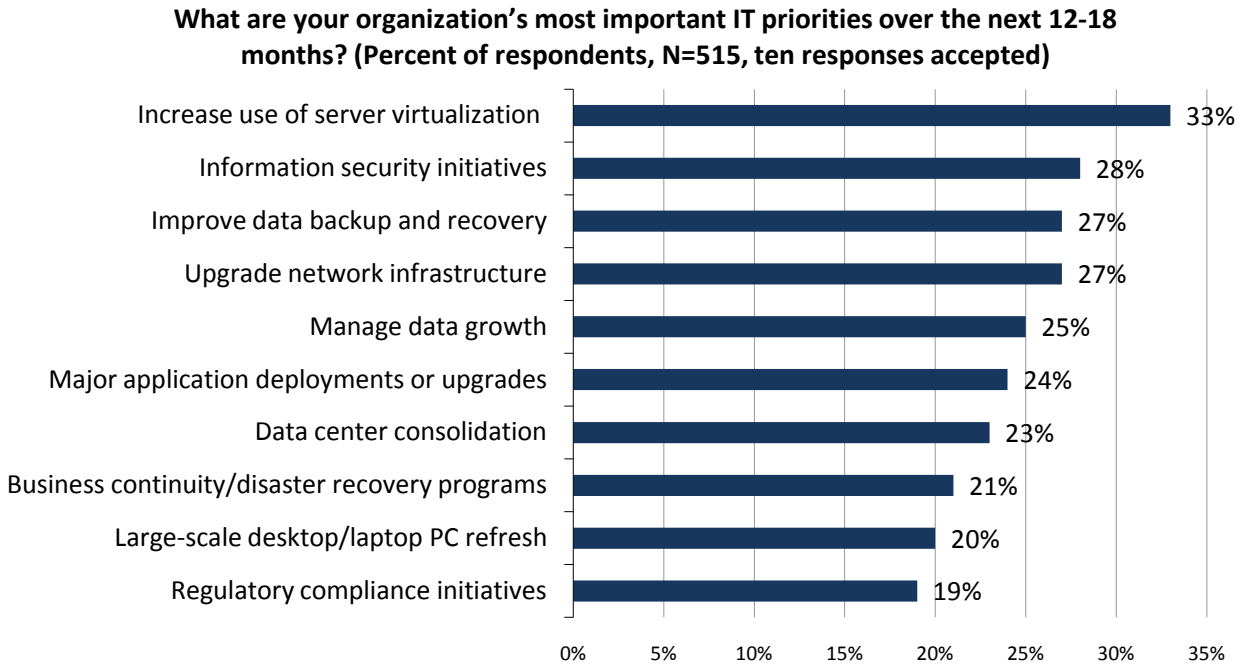
Rapid growth in the adoption of server virtualization at scale is continuing with more and more organizations deploying server virtualization as a strategy for consolidation, cost reduction, higher resource utilization, and greater efficiency. Phased adoption seems to be the standard—organizations begin with IT-based applications residing in the data center ("Stage 1") and, after they discover the benefits and become more comfortable with the technology, begin to expand virtualization deployments. As they migrate into later stages and start virtualizing user-facing applications, new challenges arise. For these applications, consolidation is not the focus; high availability and data protection rise to the top of the priority list. As a result, any task that can slow down productivity or limit scalability, such as IO-intensive backup, must be examined carefully.

Traditional backup processes aren't always the best solution in virtual deployments. They weren't built for the extreme data redundancy and hefty server loads of consolidated environments. In a traditional data center, physical servers host individual applications; server resources are significantly underutilized, but that leaves plenty of processing power for backup. In a consolidated virtual server implementation, however, multiple virtual machines (VMs) share the same physical hardware; resources are better utilized, which is cost efficient but leaves fewer resources for backup. There is also more data to back up due to redundant OS images, application profiles, and data with fewer resources to do so: an obvious bottleneck. This is exacerbated as virtual environments scale with hundreds, or even thousands, of virtual machines sharing a common resource pool. These deployments need a backup/recovery solution built to handle their needs. The best advice is to *plan* for data protection when strategizing for future virtualization growth instead of waiting to face problems in real time and risking a breakdown of virtualization goals.

Virtualization Priorities

ESG research indicates that expanding virtualization is a top IT priority. When asked about their most important IT priorities for 2010, 33% of respondents to a recent ESG survey said they planned to increase use of server virtualization—representing the largest percentage of responses (see Figure 1).[1]

Figure 1. Most Important IT Priorities for 2010

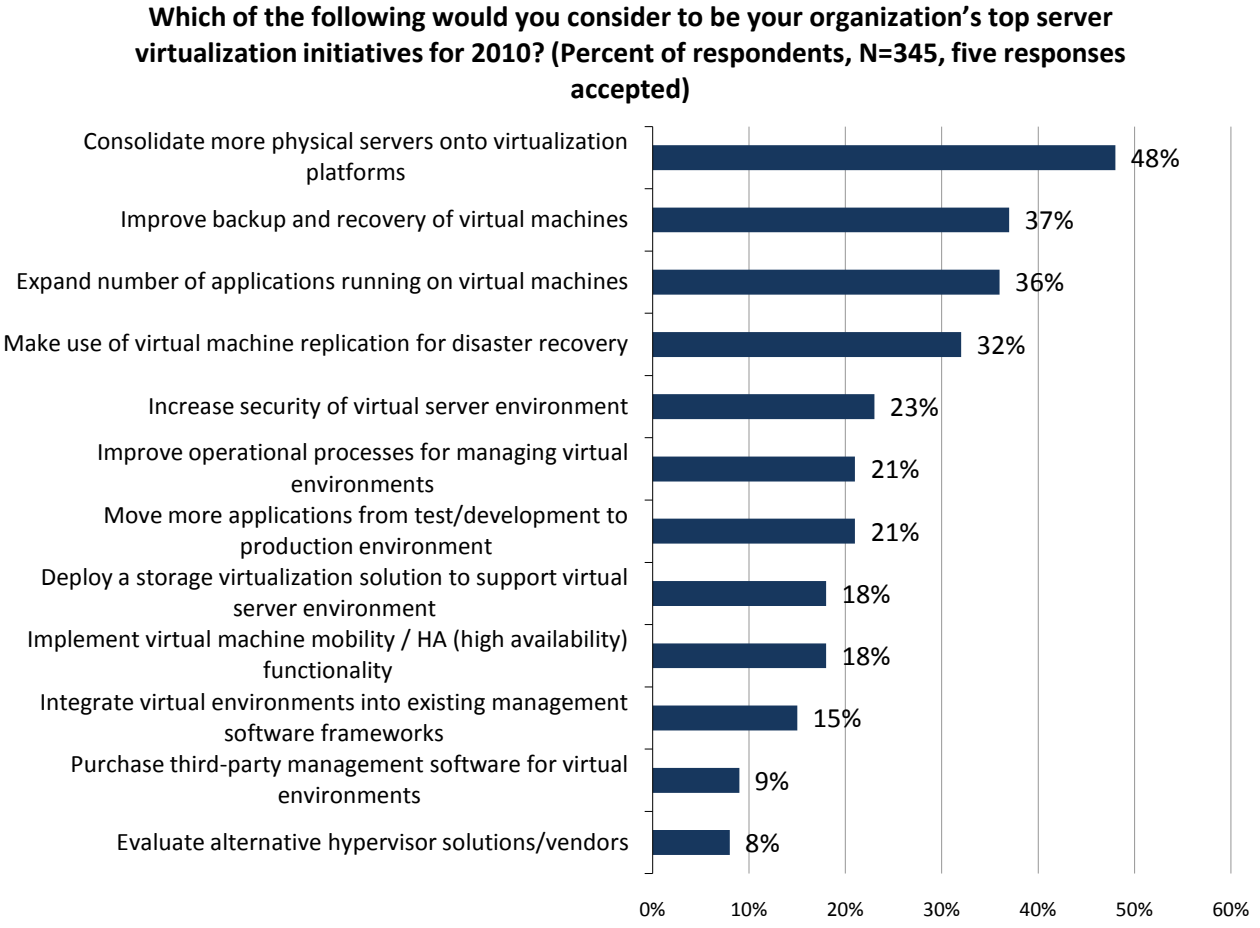


Source: Enterprise Strategy Group, 2010.

[1] Source: ESG Research Report, *2010 IT Spending Intentions Survey*, January 2010.

Drilling down further, survey respondents were asked what they considered to be their top server virtualization initiatives for 2010. As Figure 2 shows, the top four responses were focused on expanding virtualization deployments and improving data protection, particularly backup and recovery of virtual machines.[2] As the ratio of VMs to hosts increases and more applications are virtualized, backups will take longer and consume more resources. Reducing the amount of data being backed up is essential to ensuring uncompromised production operations. ESG believes that most organizations will struggle to expand virtualization deployments without first improving virtual machine backup and recovery.

Figure 2. Top Server Virtualization Priorities for 2010



Source: Enterprise Strategy Group, 2010.

Virtualization Acceleration

There tends to be inflection points in the virtualization adoption curve where the number of virtual machines jumps from tens to hundreds to thousands. In a virtualization deployment, backup and recovery can be the limiting factors: with ten VMs, things look manageable and most organizations can get by with existing tools and processes. But when you approach 100 and then 1,000 VMs, backup and recovery becomes more daunting.

Very large virtualization deployments of thousands of VMs are currently few and far between, but their numbers are definitely growing. Fueled by the industry push toward IT as a service models and promises of agile IT services at lower costs, more and more organizations will find themselves facing these inflection points. In fact, ESG research shows a 49% increase in the percentage of virtual machines running in a production environment over the next 24 months. The question is: Will organizations consider the challenges and plan accordingly up front or wait until backups for 100 virtual machines running business applications are exceeding backup windows?

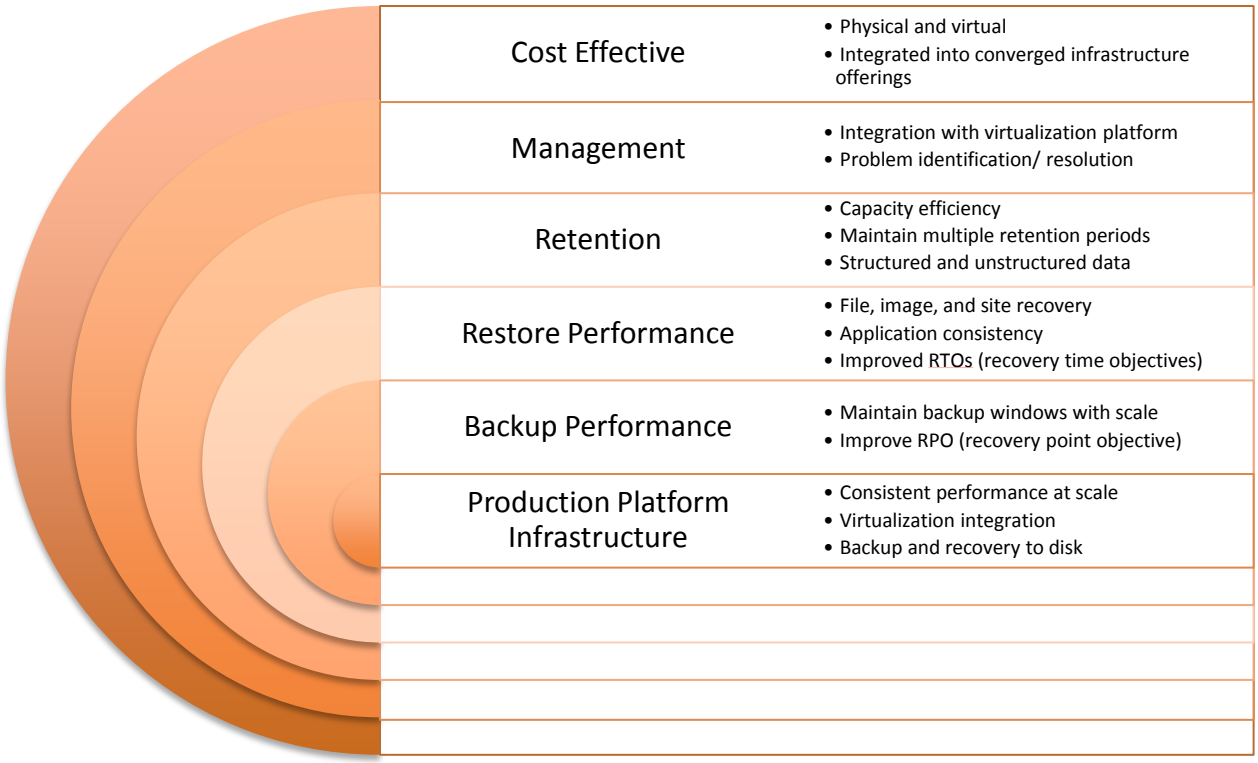
[2] Ibid.

Data Protection at Scale

Traditional data protection models were designed for a physical infrastructure in which isolated applications were matched with hardware and backed up individually. If the backup were to interfere, only that one application would be affected. However, now that multiple virtual machines can reside on each physical server, the backup process becomes more complicated and contention for resources can cause slowdowns and interruptions to multiple applications simultaneously. In order to scale virtualization deployments effectively, traditional data protection models must be overhauled. Backup and recovery must be transformed from tape-based storage to capacity-optimized disk as tape simply cannot deliver the performance and availability new consolidated data loads require.

Scaling a virtual environment can be easier if its impact on backup and recovery is considered in advance. By proactively building a backup and recovery strategy into the overall virtualization plan, an organization can save itself some headaches.

Figure 3. Considerations for Backup and Restore at Scale in VMware Environments



Source: Enterprise Strategy Group, 2010.

VMware Enables Faster Recovery

While backup in a virtual server environment can be challenging, virtualization actually enables faster recovery in terms of provisioning and getting data back online. In a VMware implementation, a virtual machine workload is encapsulated into a single file containing the operating system, applications, and data; this file can be moved/copied anywhere. If it is copied to another server while the first is backed up, very little production time is lost because the backup no longer interferes with productivity—a VM is not dependent upon particular hardware. In addition, because the entire application stack the user needs is encapsulated in a VM, it is simpler and faster to recover. Instead of a sequential recovery process—operating system, then application, then data—everything is recovered in one (using application-specific guest OS agents) or two (using image-level backup and virtual proxy servers) steps.[3]

[3] This is described in greater detail in a previous ESG publication on EMC Avamar: ESG White Paper, *Improving Backup and Recovery for VMware vSphere Environments*, July 2010. All subsequent references to ESG publications come from this source unless otherwise stated.

In addition, advances in VMware offerings such as the vStorage API for Data Protection (VADP) and vSphere change block tracking (CBT) have greatly improved backup processes. VADP allows a live system image snapshot to be captured without impacting applications or putting too much pressure on the host server CPU. It also enables virtual (instead of physical) proxy servers. CBT, by tracking changed blocks of a VM virtual disk, enables backup and copying only of changed blocks, reducing backup time and network traffic. These improvements can make a tremendous difference in backup time and storage capacity as an environment grows.

EMC Avamar Eases Data Protection for Large Virtual Deployments

As virtualized deployments scale, it's important to know what to expect in order to plan for and avoid problems. A key part of virtualization scaling includes dealing with a mix of business applications, not just IT system workloads. The farther along in its virtualization journey an organization gets, the greater its data protection needs will be because the applications tend to be larger, more difficult to manage, and focused on different priorities. Stage 1 involves virtualizing test beds and unstructured data applications. Once an organization reaches Microsoft Exchange, SAP, Oracle databases, and desktop virtualization solutions, backup becomes much more important. Tools and solutions that can meet the service level expectations of business users are needed—applications must be fully protected, often remotely replicated, and yet remain constantly available.

Expanded virtualization results in net growth of storage volumes. Daily incremental and weekly full backups create a lot of duplicate data as each VM's backup job includes OS, application, and file data. How do you continue to fully protect data and still minimize the CPU, bandwidth, and storage needed? Scaling creates more opportunities for resource contention as virtual workloads compete for limited, shared resources. So, the smaller the backup load, the better.

EMC Avamar delivers highly efficient, deduplicated backup and recovery, which is a must for rapidly expanding virtual server environments. Because reliable data protection is imperative for a virtualization strategy, Avamar can actually help accelerate VMware adoption as it meets all the requirements for optimizing backup and recovery in a virtual environment.

Management

EMC Avamar supports both guest- and image-level backups with support for VMware's vStorage API for Data Protection; minimizes the backup load; enables cost-efficiency in management, licensing, bandwidth, and storage; and leverages the full hypervisor feature set. In addition, Avamar is integrated with VMware vSphere, so it can be monitored and managed via vCenter. As a result, IT can monitor backup and recovery operations in the Activity Monitor and view virtual machine protection policies ("guest," "image," and "none") as well as the date and time of the most recent backup. This makes it very easy for operators to identify newly created VMs that may not be adequately protected. From vCenter, virtual machines may be added and backup policies defined for them.

For image-level backup, Avamar's architecture leverages a proxy pool that scales up and down based on backup and restore demands. This allows the backup job to be sent to the first available proxy server as opposed to waiting for a dedicated proxy server. There are no hard assignments of virtual machines to proxy servers or mapping of storage LUNs to proxy servers.

These features:

- Help improve backup performance
- Scale up with performance upon demand to help quickly onboard new applications
- Scale out to meet the needs of large VMware environments
- Are tightly integrated with the management platform for rapid change integration

The enabling technology for Avamar is source-based deduplication. Instead of copying all data during every daily full backup, Avamar only copies new, unique, sub-file, variable-length data segments. Compared to traditional full backup methods, Avamar can reduce the daily impact on virtual and physical infrastructure. Traditional backup

software moves approximately 200% of primary backup data weekly; Avamar moves far less over that same period. Avamar's global data deduplication can eliminate the cost of storing and moving redundant data across thousands of geographically separate VMs.

The implications for resource contention, network bandwidth, and storage requirements are obvious: deduplication at the source reduces all of these factors. What makes it so important in a growing virtual environment is that it enables greater levels of server consolidation and actually makes expansion possible for many organizations. Without it, they suffer backup bottlenecks, bandwidth problems, contention for server resources, and missed backup windows. IT can also benefit from minimizing the resources required for the backup and restore process and confidently predict utilization based on application workload without having to factor in the impacts of backups and restores on the virtualization platform.

What makes Avamar an efficient, client-based data deduplication engine is its intelligent algorithms for sub-file, variable-length data segments. Other solutions that use fixed-block or fixed-length data segments to deduplicate can be fooled by logical data shifts, such as inserting data into the beginning of a file. Avamar's algorithms quickly determine logical boundary points and redundant data segments. Because it so dramatically reduces the amount of data copied, organizations can retain full backups on disk for longer periods of time, making data more quickly recoverable. Avamar also provides single-step recovery, completely eliminating the tasks of restoring a previous full backup plus subsequent incremental backups. Avamar's support for VMware's change block tracking, coupled with its own inherent capability to reconcile changed data into recoverable backup images, gives it a recovery performance advantage over most traditional approaches. In addition, Avamar can offer granular, file-level recovery from image-level backups by opening the image and leveraging its own integrated file system to present the directory structure to the administrator. Finally, with Avamar, administrators have the flexibility to recover data to the originating virtual machine or to create, provision, and recover data to a brand new virtual machine—all managed from within the Avamar user interface. For disaster recovery, Avamar provides secure, efficient, virtual-to-virtual or virtual-to-physical replication.

Avamar itself scales using a grid architecture, offering linear increases in both performance and capacity as nodes are added. Each node increases CPU, memory, IO, and capacity for the entire grid.

Avamar offers equivalent benefits to both physical and virtual environments and supports all major proprietary and open-source operating systems. It includes support for major database and messaging applications from IBM, Microsoft, and Oracle as well as NDMP for NAS filers. Avamar's interoperability enhances its scalability; organizations are not restricted as they expand their virtual server environments.

Virtual Computing Environment Coalition Vblocks

Virtualization across the data center creates the need for tighter integration among servers, networks, and storage. This is even more the case in large scale virtualized environments. As a result, more and more organizations and service providers are using converged infrastructure units, such as the Vblock from the Virtual Computing Environment (VCE) coalition (VMware, Cisco, and EMC), to deliver IT services. These pre-integrated bundles of compute, network, and storage resources offer a different consumption model for IT services that can dramatically simplify delivery and accommodate massive scaling of virtualized environments. With Vblocks, IT gains visibility into the complete infrastructure and can more easily maintain performance, troubleshoot, and proactively respond to problems. Capacity can also be served up much faster to keep up with rapid growth. The massive scale of virtualization results not only in data growth, but also in greater power consumption and management complexity; Vblocks reduce power consumption and simplify management. As server virtualization deployments focus on SAP, Oracle, and Windows workloads, Vblocks can help speed deployment and time to value.

Vblocks can contain thousands of virtual machines with a mix of application workloads and data sizes. The high-end configuration, Vblock 2, is in intended for 3,000 to 6,000 VMs while the mid-sized Vblock 1 is designed for 800 to 3,000 VMs. All Vblocks are built with VMware vSphere 4, Cisco compute and networking equipment, and EMC storage. The three industry leaders have gone to great lengths to provide integrated tools, services, and support to ensure that IT services are delivered fast without hiccups and bottlenecks. However, if not considered in advance,

the data protection needs of these Vblocks could cause performance problems—exactly what the VCE vendors have worked so hard to avoid.

Avamar fits well into this type of scenario. Pre-integrated units consolidate a lot of IT resources; Avamar can reduce the amount of network bandwidth and storage Vblocks need for backup. Avamar also adds another benefit: business users often need to be convinced to allow their production applications to be virtualized to build comfort in the virtualized platform. They have similar concerns about converged units such as the Vblock as they are wary of change and its potential impact on productivity. Avamar, with its ability to eliminate backup resource contention, bottlenecks, and performance problems, can actually help demonstrate to business owners the value of the Vblock infrastructure as it has no negative impact on application performance in large scale environments. Whether the "infrastructure-as-a-service" is coming from an internal IT department or a service provider, Avamar can help deliver guaranteed performance, availability, and data protection that build user confidence.

The Bigger Truth

Backup redesign is a must if customers are to reach their virtualization goals—without it, scaling the virtual server environment is nearly impossible. Next-generation backup applications must align with the next generation of IT by doing more than just escalating server virtualization; they need to support resource and management consolidation, infrastructure-as-a-service, "cloud" services, and pre-integrated infrastructure bundles. Backup and recovery processes built for the physical world will not provide the required data protection, application performance, or storage optimization needed in these new environments. Only with a backup process that is built for this kind of infrastructure can organizations reap the operational benefits and ROI of large-scale deployments.

VMware recognized this and, as a result, introduced APIs for data protection and change block tracking to keep data protection of larger systems from being a bottleneck. The combination of VMware enhancements with Avamar's architecture creates an exceptional solution. With Avamar, data is captured, transferred, and stored more efficiently because it takes advantage of vSphere enhancements. Avamar adds another efficiency dimension by taking innovative approaches to economize and streamline backup and recovery processes. It drives higher VM to ESX server ratios, supports all methods available (including guest and image backup plus remote office backup), and optimizes capacity.

When it comes to data protection, IT organizations can either proactively build a solution such as Avamar into their virtualization plans or wait for problems to emerge and then fix the backup process *and* restore user confidence in the virtual infrastructure at the same time. In standard environments, a backup and recovery solution based on deduplication is extremely helpful; in a very large virtualized infrastructure, it's an absolute necessity.

Enterprise Strategy Group | **Getting to the bigger truth.**

20 Asylum Street | Milford, MA 01757 | Tel:508.482.0188 Fax: 508.482.0218 | www.enterprisestrategygroup.com