The Case for Persistent Full Clones

A better alternative to linked clones for VDI

A DeepStorage Technology Report

DeepStorage.net

Prepared for
GreenBytes

About DeepStorage

DeepStorage, LLC. is dedicated to revealing the deeper truth about storage, networking and related data center technologies to help information technology professionals deliver superior services to their users and still get home at a reasonable hour.

DeepStorage Reports are based on our hands-on testing and over 30 years of experience making technology work in the real world.

Our philosophy of real world testing means we configure systems as we expect most customers will use them thereby avoiding “Lab Queen” configurations designed to maximize benchmark performance.

This report was sponsored by GreenBytes. However, DeepStorage always retains final editorial control over our publications.

Contents

The Bottom Line

Virtual Desktop Infrastructure (VDI) has had some success reducing the cost of supporting task workers in environments like call centers, health care facilities and computer labs. In these environments, where before VDI the users typically shared desktop PCs to run a limited set of applications, the users were willing to accept non-persistent desktops that presented the same configuration each time the user logged in.

If VDI is to expand from task workers to the more demanding knowledge worker community, those VDI environments will have to provide a richer experience so users will see their desktop, with their customizations as it was when they logged out, not a generic desktop. Of course this desktop has to perform at least as well as the physical desktop it's replacing and while costing less.

Linked clones, while they are a great solution for non-persistent desktops, are problematic for persistent desktops:

- *The differencing disks grow on average 1GB/user/day eventually eating all the disk space saved by using linked clones in the first place.*
- *Separating the user profile/persona from the desktop image requires additional administrative effort.*
- *User installed applications and web browser plug-ins are deleted when the desktop images are periodically recomposed.*

Full clones allow IT to manage VDI images using the same tools and processes they've mastered over the years for physical desktops. If those clones are stored in a high performance deduplicated data store, then the users can have the full desktop experience while IT can have the security and administrative advantages of VDI.

In addition, by removing duplicate data in the shared storage, rather than VDI platform, total disk space is significantly reduced as swap files can be deduplicated along with the desktop images themselves.

GreenBytes' solid state storage-based solutions, through their high performance inline data deduplication, provide a simple path for IT departments that want to provide a high-performance, cost effective VDI experience for their users.

VDI Virtual PC options

Introduction

Virtual Desktop Infrastructure, better known as VDI, extends the hypervisor technology that allows organizations large and small to run multiple virtual servers on a single physical host to support virtual desktop PCs. By bringing desktops into the data center, VDI promises to improve the security and management of desktops while reducing the cost of supporting end users and providing access to corporate applications for mobile users including those on their own tablets and smartphones.

While VDI can improve security and improve desktop system management, shifting desktop workloads like Microsoft Word and Excel from the distributed desktop to a centralized, server-based environment introduces some complexities of its own, especially when it comes to storage.

When a large organization deploys a new fleet of desktop PCs, they first setup a single master system with the configuration and applications they want their users to run. This configuration will typically include all the organization's "universal" applications such as Microsoft Office, Acrobat Reader and the like. They'll then use disk cloning software like Symantec's Ghost to duplicate that template onto the disks of hundreds, or thousands, of additional PCs.

VDI Creates Storage Challenges

Using full copies of each disk makes sense for physical desktops where each computer has its own disk drive with 100GB or more capacity to hold the 20-40GB of operating system and applications in a typical corporate disk image. However where physical desktops store their OS and applications on the least expensive disk drives desktop vendors can buy, VDI images have to be stored on more expensive server or SAN based storage, since a disk failure in a VDI environment may affect hundreds of users where the effect of a desktop drive failure would be limited to that desktop's VDI user.

Like corporate IT departments, VDI environments such as VMware's View, also duplicate a master template; but rather than copying hard disks, VDI environments duplicate the virtual disk image files they use to store the data for each virtual computer.

VDI also challenges the performance of most storage systems. A typical desktop hard disk can perform 70-100 IO operations per second (IOPS) while PC users generate an average of 20 to 50 IOPS over the working day. A VDI server supporting just 100 users would therefore demand 2-5,000 IOPS on average, which will stress a conventional shared storage array, causing performance problems. Of course to deliver acceptable performance for our users, we have to design a system to support peak, rather than average, demand increasing the need for high performance storage. Since most organizations will have their users arrive at work and log in over a short period of time each morning the peak IO demand of this "login storm" is frequently several times average demand.

Enter the linked clone

VDI vendors like VMware have attempted to address the storage problems VDI creates through a technology called linked clones. Linked clones reduce the amount of disk space needed to store a large number of virtual desktops by storing only one copy of the data that they have in common.

How Linked clones work

VMware linked clones leverage VMware's redo log based snapshot technology. A pool of linked clones consists of a common replica which is a snapshot of the "Golden Master" virtual PC and the linked clones themselves. Each linked clone includes several files that differentiate this clone from both the master replica and the other linked clones in the pool. These virtual PCs are inexorably linked to the master replica and cannot run if the master replica is not available.

The virtual disk files that make up a linked clone include:

- The Differencing or delta disk – The differencing disk is the key component of the linked clone. It is seen by the linked clone as its system disk, usually the C: drive for Windows systems. The differencing disk logs all the changes between this linked clone's system disk and the master replica to which this clone is linked.
- The Internal Disk – A small virtual drive that holds identity information about the virtual PC including the SYSPREP or QUICKPREP configuration file and the password for the PC's machine account in Active Directory
- The Disposable Disk – The disposable disk for each linked clone holds temporary files that are needed only when the virtual PC is running. These files include the Windows swap file and optionally the user's temp folder and temporary internet files folder. The contents of the disposable disk are deleted when the virtual PC is shutdown to save disk space
- An optional Persistent Disk – formerly known as the user data disk, the persistent disk is presented to the clone's operating system as an additional drive letter that can be used to store user data that should persist across refreshes of the differencing disk.

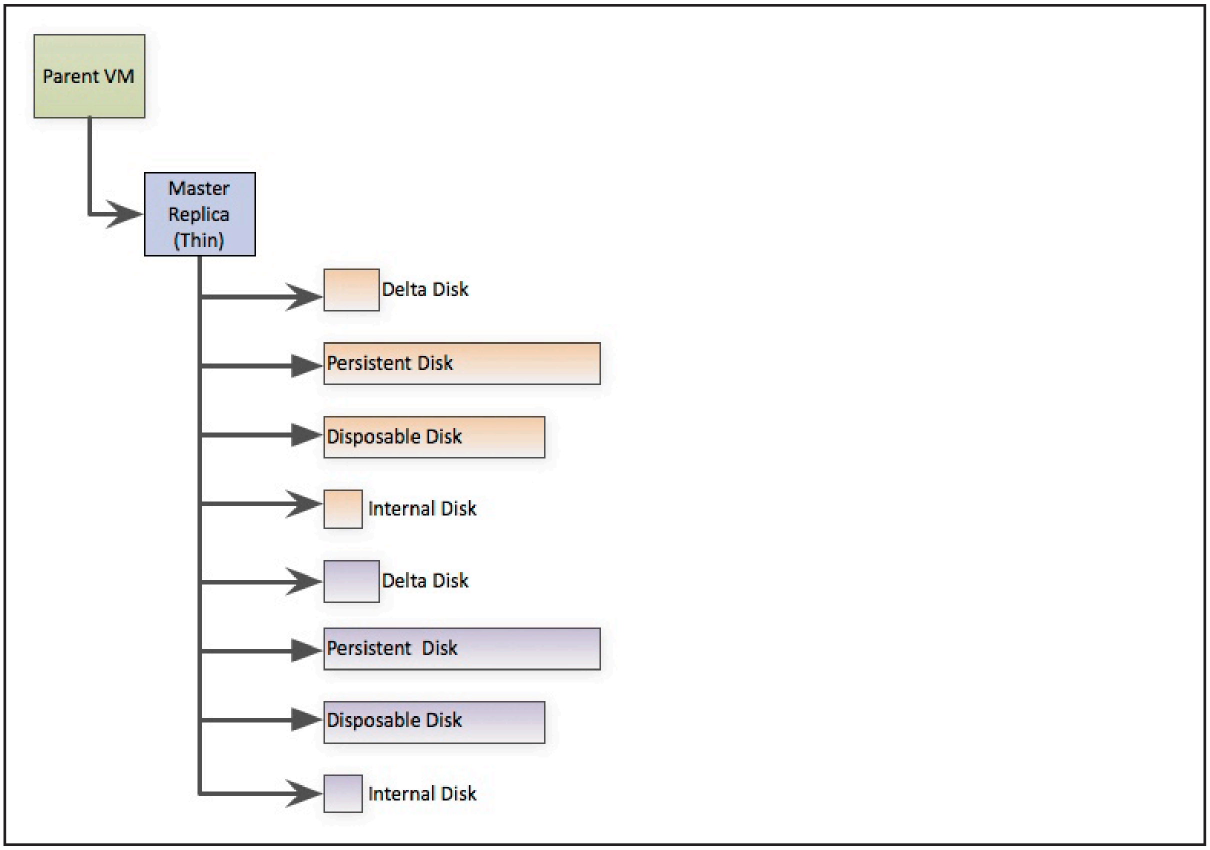

Figure 1. Linked Clone Files

Linked clones, since they're derived from vSphere snapshot technology, are in reality based on redo logs that accumulate the changes between the linked clone and its parent replica. These block level redo logs accumulate block level changes to the system disk growing constantly. Most significantly, linked clones grow as new data is written to them but don't shrink when files are deleted.

The Windows NTFS file system is particularly profligate in its use of disk space. As with most file systems when files are deleted from an NTFS volume the system simply removes the file's entry from the directory and flags the data blocks the file occupies as over-writeable. When the system creates a new file, it will exhaust the free space on the disk before overwriting the data blocks occupied by deleted files in order to allow deleted file recovery.

In VMware's international deployment of View the linked clones disks grow as much as 1GB per VM per week.

This combination of NTFS and a redo log based virtual disk means that the redo logs will continue to grow over time as Windows updates the registry and other system files writing new versions of updated files to free space on the virtual disk.

In today's computing environment, data is constantly being written to the workstation's system disk. While persona management tools like roaming profiles and folder redirection can keep the user's My Documents folder, wallpaper and custom dictionary on a network file share or the clone's persistent disk the clone's system disk will still be written to on a regular basis.

The two most significant sources of this data growth are temporary files and security updates of various kinds. While it would be reasonable to expect applications to write their temporary files to the designated *TEMP* folder, which could be easily redirected to the temporary disk and deleted each time the clone PC is shutdown, that's not always the case. When a user opens a file they received as an email attachment in Microsoft Outlook, for example, it saves that file in the hidden A*ppData\Local* folder, which is not part of the user's profile and may not be redirected.

Security updates include not just the Windows operating system security patches Microsoft distributes on the second Tuesday of each month but also, anti-virus definition updates and the constant flow of security updates to applications like Oracle Java , Adobe Acrobat Reader and Flash.

This constant flow of data to the system disk means that even in well managed VDI environments linked clones can be expected to grow by 1GB per clone per week.

Linked Clones and Storage Performance

Storing the common operating system and application data for all the virtual PCs in a linked clone pool in the master replica not only reduces the amount of disk space the pool occupies but also concentrates much of the storage I/O to the master replica. This is especially true of the massive burst during a boot or login storm.

System administrators can take advantage of this I/O concentration by placing the master replicas for their linked clone pools on an SSD or other high performance storage device. Hybrid storage vendors that use flash memory as a write through or read-only cache will similarly emphasize how their systems accelerate VDI workstation boot times.

While flash memory read caches, and VMware's View Storage Accelerator which uses system RAM as a read cache, can address the boot storm issue they have a much smaller impact on VDI system performance after 10AM as the level of write traffic increases dramatically once the virtual PCs are up and running.

Non-persistent (stateless) and Persistent (stateful) Desktops

Many of the first successful VDI implementations were in environments like call centers, airport ticket counters and hospitals. In addition to high transaction volumes, and therefore a requirement for high availability, these environments have several similarities. Each workstation in a college computer lab, airport or hospital is used by several different users over different shifts and those users run a small number of applications to perform their jobs. While a typical office user will run an email client, spreadsheet, word processor and of course web browser on a daily basis the workstation at an airline ticket counter will run just the passenger check-in application.

Since the users of these shared workstations just run this limited set of applications, one way to make sure that they run reliably day after day is to return each workstation to a known state when a user logs in or logs out. To return each virtual PC to that known state for these non-persistent images, VMware View deletes the temporary disk for the linked clone and refreshes the clone with a new, essentially blank, differencing disk

While refreshing each workstation to its starting condition is a good thing for reliability, that reset reverses any changes to the workstation's state the users make while they're working. This type of non-persistent machine is advisable in shared workstation environments most office users find non-persistent workstations to be confining.

Most knowledge workers have long taken the concept of PC as *personal* computer seriously. While IT professionals may dream of interchangeable, stateless, PCs the users insist on being able to change their wallpaper, maintain their own custom dictionaries and in other ways modify their PCs to meet their requirements. While persona management tools could allow these changes in a non-persistent desktop environment today's users will generally insist on persistent desktops.

Non-Persistent desktop problems

A few of the problems with non-persistent desktops in a typical office environment include:

- Users cannot install their own applications, browser plug-ins and other pieces of code.
- Some applications use attributes of the desktop system as part of their license enforcement. They may require consistent MAC addresses or install a license file in a particular folder on the desktop's system disk. When non-persistent clones are reset these licensing attributes may be destroyed
- Since users cannot install applications corporate IT must maintain a separate pool of clones for each group of users that use a different set of applications. This can easily get out of hand with IT having to maintain a large number of master images and linked clone pools.
- Despite all efforts by IT some users will save files to the C: drive and those files will be lost

Some of these problems can be addressed though application virtualization and streaming technologies such as VMware's ThinApp but these technologies introduce complexities of their own. Even when they work for 95% of an organization's applications there are always a few apps that simply won't run in a stream and must be installed on the workstation.

Persistent linked clone management issues:Trouble with the recomposition commission

Advocates of linked clone environments frequently argue that it simplifies the application of patches and other updates. Rather than installing the patches and updates on each virtual PC the system administrator can update the master replica for the pool.

This not only ensures that all the virtual PCs in the pool actually have the latest updates installed but since the updates were applied to the master only one copy of the updates will be stored as part of the master replica. Once the patches have been applied to the master the pool can be recomposed linking the user's virtual PC to the new master replica.

Periodic recomposition is also promoted as the solution to the constant growth of the differencing disks on linked clones. As the linked clones grow over time the disk space savings from using linked clones diminish. If a typical virtual desktop uses 26GB of space and each linked clone grows by 1GB per week the linked clone environment will actually consume more disk space than a simple full clone environment after just 6 months.

However recomposition is not a panacea. To recompose a linked clone pool the administrator has to perform a series of steps:

1. Apply the security and other updates to the master virtual machine

2. Create a new snapshot from the master virtual machine

3. Recompose the user desktops

These simple steps assume that an organization has only one linked clone pool and that the pool is being recomposed only to install simple security patches. In reality many organizations have several linked clone pools to support different user communities with different application sets. Should an organization's administrators want to push out a new version of say an internally developed expense reporting application the system administrators will have to install and test that application on the master virtual machine for each of their linked clone pools which could take days or weeks.

Because the recomposition process requires this manual intervention, system administrators typically only recompose their users desktops periodically, commonly in the few days following Patch Tuesday. Since VMware View can only recompose workstations that are logged out, administrators must choose between forcing users to log out at say 10PM, which could be a problem for organizations using VDI to support mobile workers, or to have the recomposition process wait until the user has logged out of each station before recomposing it. The cumulative delays of waiting for a periodic recomposition and then waiting for a user to log out leaves the workstations vulnerable to attack while they're still running the old image.

The biggest problem with periodically recomposing linked clones is that the recomposition process deletes the differencing disk from the clones. While user's personas and data can be preserved on the optional persistent disk or a networked file share changes the user made to the system disk, like user installed applications, will be lost.

Even worse the shortcuts to user installed applications will remain in the user's start menu or on the user's desktop so users' will click the shortcuts and then call the helpdesk when application fails to load.

Periodically recomposing desktops, while it may address the disk space problems created by constantly growing linked clones, it effectively reduces persistent linked clones to being only temporarily persistent with their state persisting only from one recomposition to the next. Therefore organizations that have adopted persistent linked clones to support user's work patterns have to choose between constantly growing linked clones and unhappy users.

If the constantly growing differencing disks of persistent linked clones will eventually grow to eliminate the initial disk savings are persistent linked clones worth the trouble?

Recomposing linked clones clears their differencing disks reducing them from persistent to temporarily persistent desktops.

In addition to saving disk space using linked clones appears to have some advantages when it comes to storage performance. Separating the common data from workstation unique data concentrates storage access so administrators to can either manually tier their storage placing the frequently accessed master replica on SSDs or other high performance storage. This IO concentration can also improve performance in hybrid storage systems that use flash memory as caches large enough to hold the master replicas for all the linked pools on a storage system due to the high IO density master replicas create.

Linked clones also have a performance downside as each disk access to a linked clone requires vSphere to access the clone's redo log metadata to determine if the block being accessed is stored in the linked clone or in the master replica. Only then can it retrieve the data from the master replica or linked clone.

Persistent full clones

Full clones on the other hand simply store their data directly in their VMDK files directly without the complications of redo logs.

The biggest advantage of using full clones is that administrators can use the same tools and workflows to manage full VDI clones that they use to manage their physical desktops. Even if an organization's ultimate goal is to convert to a 100% VDI environment the conversion will take a minimum of several months during which both physical and virtual workstations will need to be maintained.

Rather than applying patches and application updates to master replicas and periodically recomposing desktops administrators can use automated patch management and software distribution systems like LANDesk or Microsoft's WSUS/SCCM to apply patches at user login or even in real time which would also reduce the time to which they are exposed to security vulnerabilities for which patches have just been issued.

The biggest problem full clones present is that one hundred 40 gigabyte full clones will consume 4TB of valuable shared storage space. Luckily a new generation of storage systems can help eliminate this problem.

How modern storage addresses VDI issues

While most of us continue to think of a storage system as the disk arrays that dominated the datacenter in the ‘90s and 2000s a new generation of storage systems has emerged over the past few years that take advantage of major developments in both solid state storage and processor technology. By integrating non-volatile flash memory with sophisticated software running on the latest processors these systems can provide a level of performance and storage efficiency well beyond even the high end systems of yesteryear.

Solid state storage for IOPs

Storage performance is critical to the success of a VDI initiative. Not only does the concentration of IO requests from hundreds of users onto a single storage system require that storage system to deliver consistently high performance with low latency, but now that all of our users are sharing a single storage system, they're susceptible to the noisy neighbor problem. In a standard desktop environment if a user starts a disk intensive operation, like a search of all the documents in their My Documents folder, it will tie up their local hard disk for several minutes. That same operation in a VDI environment whose storage system is already working close to its maximum capacity will slow performance for all the other users as well.

Studies have shown that almost half of all VDI initiatives stall at the point where they transition from pilot to production because of storage related issues.

Following the inexorable march to higher transistor densities described by Moore's law Intel and AMD release a new generation of server processors every eighteen months or so. Since each generation of processors is somewhere between two and six times as fast as the previous generation today's servers are thirty or more time faster than the servers of just a decade ago. Sadly over that same decade the disk drive industry, while they have managed to increase capacity from 40GB to 4TB per drive hasn't made disk drives significantly faster.

> Moore's Law: The number of transistors in an integrated circuit doubles every 2 years

Just as the widening gulf between processor and disk drive performance was reaching Grand Canyonesque proportions server and VDI virtualization increased the demand servers placed on their storage systems. As multiple virtual servers accessed their data from the same volume their requests were multiplexed together randomizing what were requests for adjacent data.

Luckily the semiconductor industry came to our rescue in the form of flash memory. Flash is a non-volatile form of solid state memory. Since it retains data even when powered off flash memory could, through the use of an appropriate controller chip, be combined into solid state disks (SSDs) that could replace those old fashioned spinning disks but since they had no moving parts could respond to requests for data scattered at random across their capacity 100 more times as quickly as a disk drive could.

By using solid state storage virtualization administrators can greatly increase the performance of their storage systems. Unfortunately as a wise man once said there's no such thing as a free lunch and the extraordinary performance of solid state storage comes at a significant cost.

Solid state storage can cost as much as twenty times as much as old fashioned disk when we look at storage costs on a capacity or $/GB basis. The good news is that while solid state storage is much more expensive than spinning disks on a capacity basis it's so much faster than spinning disks that solid state storage is actually significantly less expensive than spinning disks when we look at the cost of performance in $/IOP.

Deduplication advantages

As solid state storage started entering the data centers of corporate America it became clear that if there were some way to squeeze more data into a solid state disk, and therefore reduce the effective cost per GB of storage, flash would be an attractive storage medium for workloads like VDI. Luckily data deduplication, a technology that identifies duplicate blocks of data in a data store and only stores one copy, fit the bill.

Data deduplication was first used to reduce the size of backup repositories as multiple nightly backups of a data center full of Windows or Linux servers contains a lot of duplicate data. VDI environments, especially one using full clones similarly presents a target rich environment with lots of duplicate data.

Reduced capacity requirements

Data duplication's most obvious advantage is the reduction in the amount of space needed to store any given set of data. While linked clones can significantly reduce the amount of space needed to support a set of VDI users still stores separate replica for each pool of VDI stations and as we've already seen the linked clones grow over time necessitating periodic recomposition to maintain the savings.

A deduplicated data store will use significantly less storage space to hold VDI clones regardless of whether they're linked or full clones. As operating system patches, antivirus definition files or other updates are applied to 2 or more clones the deduplication system they'll only be stored once regardless of whether the clones are members of the same pool or not.

Inline and Post Process Deduplication

Deduplication can be performed either in real time, generally known as inline deduplication, or as a periodic process, known as post process deduplication. An inline deduplication system examines each block of data it's asked to store and determines if it's already holding a block that contains the same data. If the system doesn't already hold a block with the same data as the new block it writes the block to its back end storage like any other storage system would. If it already has the data the new block represents it creates a pointer in its metadata, pointing to the block that contains that data instead of storing it.

Post process systems periodically run a process that searches the data in the repository to identify duplicate blocks. When duplicate blocks are found the system updates its metadata to point to one block and marks the other(s) as overwriteable.

While there are arguments in favor of using post process deduplication in backup environments we at DeepStorage believe that inline deduplication is preferable for primary storage applications like VDI hosting.

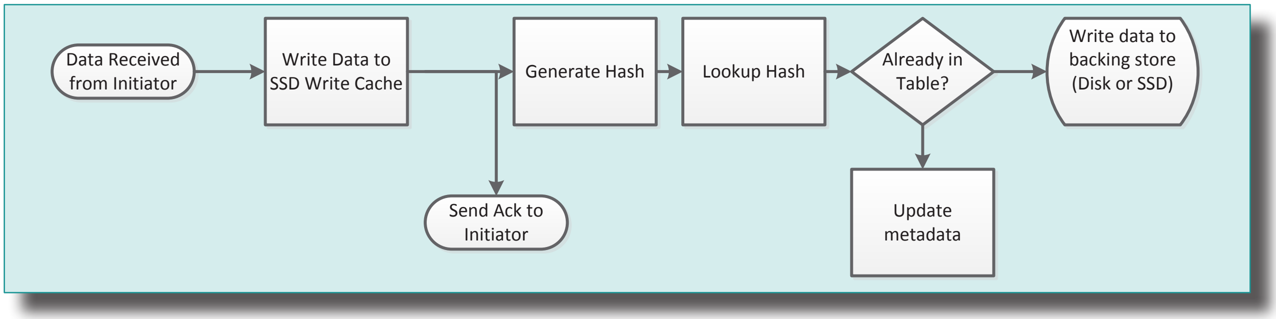
Data Received from Initiator
Write Data to SSD Write Cache
Generate Hash
Lookup Hash
Already in Table?
Write data to backing store (Disk or SSD)
Send Ack to Initiator
Update metadata

Figure 2. The Inline Deduplication Process

Deduplication and performance

Many system administrators believe that deduplicated storage systems are slower than equivalent systems that don't use deduplication. It's true that disk based storage systems, like those used for backup data, are slowed down by deduplication. This is primarily because spreading the blocks that make up large backup files across all the disks in a deduplicated repository means changes reading that file from a sequential process, which disk drives can handle well, into a more random process that requires a lot more head motion. Solid state disks can perform random I/O as quickly as they perform sequential I/O eliminating this issue.

Modern storage systems combining deduplication and solid state storage have demonstrated that they can be significantly faster than even high end storage systems based on spinning disks.

Improved cache utilization

While flash memory is about 1,000 times faster than spinning disk drives, it's about 1,000 times slower than DRAM. As a result even all solid state storage systems use DRAM as a cache holding the most frequently accessed data. Just as deduplicating systems use less disk space to store any given set of data they also use less cache memory to store the most frequently accessed data, like common OS files.

This allows the remaining cache memory to be used to hold slightly less frequently accessed data improving overall system performance.

I/O elimination for common data updates like patches

While data deduplication's most obvious advantage is that it reduces the amount of disk space used, inline deduplication systems also reduce the amount of IO that has to be performed on the actual storage. When the second, or 200th, workstation applies an operating system patch the inline deduplication engine only has to write updates to its metadata it doesn't have to write to the actual disks or SSDs.

Since any given set of disks or SSDs can only perform so many IOPS eliminating these write requests allows them to respond to other storage requests more quickly. In addition flash based SSDs have limited write endurance, that is any given block on the SSD can only be erased and written over a limited number of times. By eliminating duplicate writes systems using inline data deduplication extend SSD lives.

Swap file reduction

Modern operating systems including Windows and VMware's ESXi hypervisor implement virtual memory which swaps less frequently used memory pages out to disk to let the applications think the computer has more memory available than it really does. To support virtual memory, and to ensure there is disk space available when it's needed most systems preallocate a swap file.

VMware goes one step further, recognizing that most computers don't use all of their memory all the time, they empowered the ESXi hypervisor to over-commit it's memory resources so a host can run a set of virtual machines where the total memory allocated to all the VMs exceeds the memory in the server. As long as the amount of memory used by all those machines is less than the memory in the server all the VMs get the memory they need and everything's honky dory. If however, the total amount of memory the VMs use exceeds the amount of physical memory the ESXi host has to page to disk.

Whenever a virtual machine is powered up on a VMware ESXi host the host creates a .vswp file that's as large as the amount of memory allocated to that VM less the amount that's reserved, and will never be swapped out.

A typical VDI implementation with Windows 7 VMs that have 4GB of allocated memory with 1 GB reserved would need:

- 2GB for the Windows swapfile
- 3GB for the ESXi .vswp file

For a total of 5GB of disk for each virtual workstation; even a modest 500 seat VDI farm would need 2.5TB of additional disk for these swapfiles using conventional storage. Since the swapfiles are frequently unused, and even when they are used, can contain duplicate data, this 2.5TB of idle space will now require a modest amount of actual storage when deduplicated.

Full clones and modern storage: Two great tastes that taste great together

As we've seen for environments where user satisfaction is an important measure of the success of a VDI environment, it's important for the VDI architect to ensure that their user's virtual PCs maintain their state for long periods of time. While linked clones can be persistent over time the constant growth of linked clones eventually eliminates the disk space savings that were the reason we went with linked clones in the first place.

Since modern storage systems can be even more efficient in their use of disk space with full clones than conventional storage can be with linked clones we can use full clones and manage them with the same tools and workflows we use for physical systems.

We strongly recommend that those organizations that haven't yet adopted comprehensive desktop management through tools like Dell's KACE and/or LANDesk integrate these tools into their new virtual desktops. Giving system administrators the ability to automatically push patches and new applications to their users can significantly reduce the operational cost of supporting a large number of user workstations.

VAAI accelerated clone creation

While deduplication is no longer a rare feature on primary storage systems those storage systems that integrate more tightly with the host's hypervisor through VAAI or ODX have some significant advantages.

Without VAAI support the clone copy operation is performed by a host computer reading the template and writing a new copy of the data to the storage system. A deduplicating storage system would then have to break that incoming stream into chunks and check each chunk to see if it's new or duplicate data. Since the clone is almost all duplicate data it may not write much data to its storage but the CPU will be very busy.

System administrators are rightfully concerned about the load on the server(s) creating the clones and the storage network as this process takes place.

If however the storage system supports the **Clone Blocks** primitive that's included in VMware's VAAI (vStorage APIs for Array Integration) and the master template is on the same storage system as the clones the entire process can be offloaded to the storage system. In reality creating a new clone on a deduplicated data store is just a metadata operation as new pointers to the existing data are created. The process of creating a pool of 100 desktops, which could take many minutes for linked clones, can take as little as 2-3 minutes by simply updating the metadata to represent new linked clones. On the best of today's storage systems it actually takes longer to update the vCenter database to include the new clones than it takes for the storage system to create them.

GreenBytes' solutions for VDI

GreenBytes has been a pioneer in the application of both data deduplication and solid state disks to the storage needs of today's data center. This combination not only reduces the effective cost of solid state storage to the point where it is competitive with high performance disks, but also allows the two technologies to address each other's limitations.

Traditional deduplication has a not entirely deserved reputation for reducing system performance. One major contributor to this is that when deduplicated data is stored on spinning disks it will result in slower sequential read performance. Since a deduplicated data store holds chunks of data organized by their content or when they were written, not what files they were contained in, a long sequential read of a file will cause the system to reassemble the file from its constituent chunks which are scattered across the repository. This randomized I/O will cause a lot of head motion slowing the read. Since SSDs perform random I/O as quickly as they perform sequential I/O, storing the deduplicated data on SSDs eliminates this bottleneck.

The Achilles Heel of the flash memory that's used in SSDs is that it has limited write endurance. By reducing the amount of data that's written to the flash, data deduplication extends the life of the SSDs, allowing vendors to use MLC flash, which has lower write endurance than the much more expensive SLC flash, and still be sure that their products will provide sufficient endurance even under heavy workloads like VDI.

The IO Offload Engine

GreenBytes' flagship *IO Offload Engine* allows system administrators to shift high IOPS generated by workloads like VDI off of their existing SAN storage systems onto an appliance designed specifically to support them. Each IO Offload Engine has sufficient solid state storage to support up to a few thousand VDI users, assuming 30GB full clones and the usual degree of data redundancy.

The IO Offload Engine runs GreenBytes' patented, "zero latency inline deduplication" and stores its deduplicated data in a log-based data structure that further reduces the number of writes to the SSDs, extending their life.

Integrating the IO Offload Engine into a VDI environment is simple. It looks to VMware's ESXi or another hypervisor as a block storage device accessed via iSCSI or Fibre Channel. Each IO Offload Engine has redundant controllers for a full high availability solution.

GreenBytes vIO, the virtual IO Offload Engine

While the IO Offload Engine can relieve an organization's existing storage infrastructure of the load presented by thousands of VDI users, no product is perfect for every use case. An IO Offload Engine would be overkill for a remote office with fifty or a hundred users, but we still want to provide those users with the same rich, persistent VDI experience.

vIO implements GreenBytes' high performance inline deduplication as a virtual storage appliance using server-attached SSDs, PCIe flash cards or even a slice of storage from an all flash array.

Once the vIO virtual machine is installed, it takes ownership of the server-mounted SSD and presents the new deduplicated data repository via iSCSI or NFS. The new, deduplicated data store can support up to a few hundred VDI users. Like the IO Offload Engine, vIO can use replication to protect the user's desktops from server or SSD failures.

Compared to all solid state arrays

Today's all flash storage systems have been designed primarily to address high storage performance requirements for database applications of one sort or another. Like drag racers these systems are designed simply to go fast and are typically lacking such common storage management functions as snapshots and replication. In fact some vendors are even peddling all solid state systems that don't even have redundant controllers promoting performance ahead of reliability.

Most all solid state array vendors have specifically avoided data deduplication because they're aware that users link deduplication with a performance penalty, a misconception we discussed earlier in this paper, and because database data, unlike VDI images, doesn't deduplicate well.

As a result a typical all flash array can provide the performance VDI requires but since they don't deduplicate data they're much too expensive on a dollars per gigabyte basis for all but the richest organizations. As a result most of these vendors use non-persistent linked clones in their examples when promoting their products for VDI.

By deduplicating the VDI images, both the IO Offload Engine and vIO provide the performance of an all solid state system without the complications of linked clones.

Compared to Hybrid arrays

Since the GreenBytes IO Offload Engine uses all solid state disks, it provides more consistent performance than hybrid arrays can. Hybrid arrays use two basic techniques to manage their flash and spinning disk storage pools.

Most hybrid systems use automated tiering which determines which data chunks are being accessed most frequently. They then periodically migrate the most active, or hot, chunks to their flash pool while simultaneously demoting less active blocks to their spinning disk pool.

Other hybrid systems use their flash pools as a cache copying, rather than moving, the active data to the flash pool. Since data can be copied to the cache as its read or written by the servers cache systems can respond to changes in the data being accessed more quickly than tiering systems.

The problem with hybrid storage is that while it can deliver quite impressive IOP levels and average latency any time the data a user is trying to access is not currently in flash it has to be retrieved from the spinning disk pool. As a result a typical hybrid system will deliver data in 200µs from flash but an access to the disk pool will take 10ms or 500 times as long.

By using flash optimized deduplication, GreenBytes systems can be cost competitive with hybrid storage systems and still deliver sub millisecond latency 100% of the time.