



Disk-Based Backup with Data De-Duplication

First Generation versus Second Generation Technology The Pros, Cons and Trade Offs.

Abstract

The question of how to effectively back up data has been plaguing IT departments for years. To date, magnetic tape has been the medium of choice because it is inexpensive and easy to transport offsite for disaster recovery purposes. But tape is also difficult to manage, unreliable, not secure (on average, does not restore 54% of the time) and cumbersome. As disk prices fall, "Backup to disk" has become an ever increasing reality. However, many organizations quickly find that without compression and data de-duplication, that the amount of disk space required to maintain backup retention is cost prohibitive.

Over the past couple years, first generation disk-based backup systems with data de-duplication technology have alleviated some of the challenges associated with backing up to disk, but these systems have their challenges as well. Newer, second-generation systems improve upon disk-based backup with de-duplication and eliminate the challenges of the first generation systems.

This paper will examine the differences between the first and second generation disk-based backup systems, and will focus on the following areas:

- Data de-duplication method
- Backup performance
- Restore performance
- Data integrity
- Second-site system for offsite tape replacement and faster disaster recovery
- Scalability
- Customer support

Data de-duplication technologies are designed to reduce the amount of data stored by eliminating redundant data. Data de-duplication is a critical component in any disk-based system because it affects backup, restore and tape copy performance, scalability and other important backup functions.

First Generation Approach to Data De-duplication

In first generation disk-based backup systems, data de-duplication is performed by breaking data into approximately 8KB blocks and then comparing the data, a method known as *block-level data de-duplication*. With this method, after the system compares the data, only the unique blocks are stored. The system keeps a hash table it uses as a "map" to reassemble the data into a usable form for restores.

Pros:

First generation disk-based backup systems can achieve data reduction rates from 10:1 to 50:1. The reduction ratio varies depending upon the type of data being de-duplicated, with the average data reduction rate is typically 20 or 25:1. Many of these systems process the block in-line or on the fly, which results in using less disk space than other approaches. However, even though these systems use less disk space, they can be more expensive than systems that use other approaches.

Cons:

Block level data de-duplication has many downsides. Most products de-duplicate data *inline* or *on the fly*. This approach slows down backups due to processing on the fly and slows down restores due to the time the system takes to reassemble data. Additionally, block level data de-duplication inhibits scalability because the hash tracking table grows very large and becomes a challenge to manage across multiple servers.

USDV's Second Generation Approach

Our Second generation disk-based backup system, stores backups using pre-process data de-duplication at the byte-level. To achieve this, backup data is compared with past backups and only the bytes that change are stored from backup to backup.

Pros

With this approach, data reduction rates from 10:1 to as much as 50:1 can be achieved. The reduction ratio varies depending upon the type of data being processed, with the average data reduction ratio typically ranging from 20 or 25:1.

Cons

There are no currently known cons to this approach.

Net Result

USDV's second generation approach achieves the fastest backups and restores currently possible.

Our approach also provides superior scalability because the numbers of segments that need to be managed are a fraction of the block level approach.

Because of this, data can easily be managed across multiple servers.

One of the biggest reasons many organizations decide to move from tape to disk-based backup is to *shorten backup windows*. Backing data up to disk is inherently faster than backing up to tape, but there also are major differences between first and second generation disk-based backup approaches to data de-duplication.

First Generation Backup Performance

In first generation systems, as the data is sent from the backup server to the disk-based backup system, the data is broken into approximately 8KB blocks and only the unique blocks are stored.

Pros

First generation disk-based backup systems use less disk space than second generation systems.

Cons

Inline data de-duplication results in longer backup times and increased backup windows.

USDV's Second Generation Approach

Our second generation disk-based backup systems use pre-process, on-site data de-duplication, where the data is de-duplicated by our resident software and then sent directly to the backup server. All compression and data de-duplication is processed before the backup has left your system, thereby saving significant costs in bandwidth as well as less time required to transmit.

Pros

Pre process data de-duplication results in faster backups and shorter backup windows. The pre-process approach can be as much as two times as fast as the older, inline method because the disk-based backup system can accept the data as fast as the backup server can push the data to it, whereas the inline method performs data de-duplication inline or "on the fly."

Cons

True, pre-process data de-duplication uses slightly more local disk space than products with inline data de-duplication,.

Net Result

USDV's second generation approach delivers:

- Significantly enhanced performance over first generation systems as everything is pre-processed and shipped directly to disk.

- Maximum storage utilization.

- Lower cost than inline vendors.

- The fastest performance and shortest backup window at the lowest price.

Restore Performance

In evaluating any backup system, it's critical to consider restore performance. The faster data is restored; the sooner users can get back to work.

First Generation Restore Performance

First generation systems use inline data de-duplication to reduce data, so the disk only contains unique blocks of data and a hash table used to determine how to reassemble the data in the event of a restore.

Pros

There are no strong arguments for this approach.

Cons

Restores are slow because the data must be reassembled from blocks before the restore can be completed.

USDV's Second Generation Approach

Our second generation disk-based backup system stores one complete version of the backup using high level data compression. All other backups are reduced due to USDV's data de-duplication technology, which stores changes from backup to backup, at a byte level, instead of storing full file copies.

Pros

Because 90 percent of all restores come from the most recent backup, this approach provides the fastest possible restores.

Cons

Restore performance for earlier versions is faster for byte level data level de-duplication, than for block level de-duplication.

Net Result

USDV's second generation approach:

- Provides the fastest restores of data

- Performs restores of data from earlier versions faster than other data de-duplication methods

- Requires no additional disk because the disk requirements are factored into the disk requirements for post process data de-deduplication

- Provides the best restore performance and lowest price

Data Integrity

One of the main reasons organizations decide to move backups from tape to disk is to provide a greater level of data integrity. Simply put, when data needs to be restored, it's critical that the data is *valid and available* to be restored.

First Generation Approach

First generation disk-based backup systems perform checksums along all data paths to ensure restorability.

Pros

Provides high levels of data integrity because data is checksummed to ensure that files can be restored when needed.

Cons

There are currently no cons to this approach.

Second Generation Systems

Second generation systems also perform checksums along all data paths, while **also** providing the fastest backups and restores. All with virtually NO data degradation or loss and does it at the lowest price.

Pros

Provides high levels of data integrity because data is checksummed to ensure that files can be restored when needed. And, it is multi-optimized for high speed performance, both up and down.

Cons

There are no cons to this approach

Net Result

USDV's second generation disk-based backup system:

- Provides the same level of data integrity as first generation technology

- Provides the fastest backups and restores at the lowest price and the lowest cost in IT time

Ability to Provide a Second Site for Remote Redundancy and Faster Disaster Recovery

For organizations that want to significantly reduce or eliminate tape, systems with data de-duplication provide the ability to have a second disk-based backup system at an additional remote offsite facility for faster recovery purposes, specially in a disaster. Data de-duplication makes two-site systems efficient because only changed data is moved across the WAN, making transmission of data extremely efficient and allowing the second site to be kept up to date.

First Generation Systems

In systems with inline or on the fly data de-duplication, only the unique blocks of data traverse the WAN. This approach is WAN-efficient and allows for a second system to be kept up to date at an alternate site, its costs can still be prohibitive.

Pros

This approach works well for two-site configurations

Cons

First generation systems are not as efficient in multiple-site configurations because the processor must be shared across two processes, de-duplication and replication, making backups slow. Also, restore times are slow when performed at the second site because the backup data must be reassembled.

Many first-generation systems have additional charges for replication software, so the overall system price is significantly more.

USDV's Second Generation Approach

Our second generation systems are more efficient when used in a multiple-site configuration because only unique bytes traverse the WAN.

Pros

This approach is WAN efficient and allows for multiple systems to be kept up to date at an alternate site so it is always ready to restore data. This is also why USDV provides each Client with a FREE redundant, remote backup. (Our way of saying Thank you.)

Cons

Depending upon a number of variables, systems with post-process data de-duplication may experience a slight delay (ms) in synchronization. However, this is a small price to pay because backups and restores are significantly faster with second generation systems.

Net Result

Our second generation approach:

Provides a WAN-efficient way to maintain a "second," offsite system

In the event of disaster it provides fast restores of data from the second site, as well as the first

Includes two-site replication at no additional charge

Scalability

Many IT shops report that their data grows by 20 percent to as much as 50 percent a year, so scalability is a critically important feature of any disk-based backup system. Scalability must be seamless and cost effective while keeping the backup window as short as possible.

First Generation Approach

In first generation systems, the primary architecture consists of a head server with processor, memory, bandwidth and disk. Additional capacity is added on as storage capacity only.

Pros

Scaling first generation systems can be cost-effective, but it is totally dependent on both hard and soft space costs.

Cons

Block-level data is spread across the fixed processor and memory, resulting in performance that degrades as data grows. Additional storage can be added, but processors and memory cannot, so the backup window grows as the amount of backup data increases.

USDV's Second Generation Approach

Our second generation disk-based backup systems by Dell come packaged with servers that contain processor, memory, bandwidth and disk.

Pros

As data grows and servers are added for additional capacity, processor, memory, bandwidth and disk are also added so performance remains the same.

Data is automatically load balanced among servers to ensure that data is evenly distributed.

Cons

When servers are added to the system the cost can be higher than if only NAS boxes were added. However, the USDV system is the lowest cost system at all levels, from 5GB to 20TB in a single, continuously scalable system.

Net Result

USDV's second generation approach:

- Provides resources with every server to maintain performance

- Seamlessly virtualizes into Storage Architecture

- Automatically load balances data

- Is lower priced than systems that can only add disk capacity without additional processor and memory resources

Customer support

Backing up data is a daily function for most IT departments. Occasionally, IT departments have issues with backup equipment and need to get quality customer support from equipment vendors. Customer support is a key component in any IT purchase decision, and is particularly important in the backup arena.

First Generation Approach

With first generation disk-based backup systems, when hardware or software fails, often lengthy and complex remote efforts are required, including significant tech time.

Pros

None

Cons

IT staff frustration, wasted time.

USDV's Second Generation Approach

Second generation systems feature customer support personnel and procedures that underline the critical nature of data protection. USDV's customer support staff members are all corporate employees and are based at or in the specific facility where your data resides. Each is responsible for named accounts and are responsible for their success. USDV customer support personnel try to be proactive and are knowledgeable about our software and systems as well as customer installations and backup methods.

USDV systems feature fully replaceable components, including hot swappable drives and power supplies. The systems are multiply redundant and include RAID6 with a spare drives and power supply sets. If a drive, two simultaneous drives, or a power supply fails, the system will continue running.

Pros

USDV provides knowledgeable, proactive customer support that is among the best in the industry along with robust systems with redundant, secure operations/ Additionally, we were the first to provide a "Live Tech Chat" service as well as phone based, call back support, all at no cost.

Cons

None

Net Result

USDV's second generation approach provides:

Local US or local foreign based support staffs. NOT Contractors.

Proactive, knowledgeable support staff members assigned to named accounts

Redundant and hot swappable hardware components and Free Software as needed.

(Never additional Licenses.)

Choosing a Provider

Selecting a disk-based backup system can be a daunting challenge. Many systems on the market today promise faster backups and restore performance, but by digging a little deeper and discovering the pros and cons of both first and second generation backup systems, organizations can find a solution that greatly improves backup and restore operations. And, does it affordably.

When choosing a disk-based backup system, it is critical to evaluate:

Data de-duplication method

Backup performance

Restore performance

Data integrity

Availability of a second-site system

Scalability

Customer support

USDV's second generation disk-based backup systems address all of the above issues and more. By evaluating the pros, cons and tradeoffs of first and second generation disk-based backup systems, IT managers can confidently choose the right solution to streamline backup operations.

Intelligent Data Protection

USDV's disk-based backup system combines high quality, high reliability, multi functional software with everything needed for a total backup solution. We use byte-level delta data de-duplication, delivering a disk-based solution that is more cost effective than standard services. USDV's byte-level delta de-duplication technology stores only the changes from backup to backup instead of storing full file copies, reducing the amount of disk space needed by 10 to 50:1, or more resulting in a solution that is 25 to 30% the cost of other services.

USDV Backup Pro is easy to install and use and works seamlessly with popular programs such as Exchange and Outlook, so organizations can retain their investment in existing applications. USDV can be used as a primary site while maintaining tape for local storage or it can be deployed as a multi site solution to eliminate offsite tapes and to provide a live data repository or for disaster recovery. Remember, we also provide a FREE Redundant copy in a second, remote facility for added protection and faster disaster based restores. Also, cost savings are even greater because USDV's byte-level data de-duplication technology moves only changes, requiring minimal bandwidth.

About US DataVault

USDV is a 9 year old leader in cost-effective disk-based backup solutions. A scalable system that works with existing applications, USDV is ideal for companies looking to quickly eliminate the hassles of tape backup while reducing their existing backup windows AND costs. USDV's innovative approach minimizes the amount of data to be stored by providing high level data compression and encryption along with byte-level data de-duplication technology for all backups. Customers can deploy USDV at a primary site and at a second site to supplement or eliminate offsite tapes with a live data repository or for disaster recovery. Think of it, one system for ALL your remote office/back office operations.

US DataVault, Inc. | Box 33 | Nashville, TN 37202-0033 | 1-615-596-4898 | www.usdatavault.com

