WHITE PAPER

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 and cumbersome. As disk prices are falling, backup to disk has become a reality. However, many organizations quickly find that without compression and data de-duplication that the amount of disk 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
- Tape copy performance
- Data integrity
- Second-site system for offsite tape replacement and faster disaster recovery
- Scalability
- Customer support

Data De-duplication Method

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 and uses it as the "map" to reassemble the data into a complete and 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, slows down restores due to the time the system takes to reassemble data, and slows down tape copy times due to the time it takes to reassemble the 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.

ExaGrid's Second Generation Approach
Second generation disk-based backup systems send data directly to disk for fast performance and then perform data de-duplication. ExaGrid's disk-based backup system stores the most recent backup in its complete form in order to provide fast restores and offsite tape copy. Older backups are stored using post-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 cons to this approach.

Net Result

- ExaGrid's second generation approach achieves the fastest backups, restores, and tape copy performance along with market comparable data reduction ratios.
- The ExaGrid approach also provides superior scalability because the number 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.

Backup Performance

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 than second generation systems.

Cons

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

ExaGrid's Second Generation Approach
ExaGrid's second generation disk-based backup systems use post process data de-duplication, where the data is sent directly from the backup server to the ExaGrid. All compression and data de-duplication is processed after the backup has landed on the disk.

Pros

- Post process data de-duplication results in faster backups and shorter backup windows. The post-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

- ExaGrid's post-process data de-duplication uses more disk space than products with inline data de-duplication, and because of this, ExaGrid ships larger disks. For example, instead of shipping 250GB drives, ExaGrid may ship the next disk size up, 400GB drives. The additional disk space will still be housed in the same 3U server to maximize rack space. Additionally, ExaGrid prices its products lower than inline vendors.

Net Result
ExaGrid's second generation approach delivers:

- Up to two times the performance over first generation systems because backups write directly to disk, and then post process compression and data de-duplication is performed.
- Maximum storage in a small 3U footprint.
- 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.

ExaGrid's Second Generation Approach
ExaGrid's second generation disk-based backup system stores a complete version of the most recent backup using 2 to 1 data compression. All prior backups are reduced using ExaGrid's data de-duplication technology, which stores changes from backup to backup 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 because the most recent backup is compressed 2 to 1 and is always ready to restore.

Cons

- Combining 2 to 1 data compression along with data de-duplication for older backups provides the fastest possible restores for the most recent backups, which account for 90 percent of all restores. Restore performance for earlier versions is about the same for byte level data de-duplication and block level de-duplication.

Net Result
ExaGrid's second generation approach:

- Provides the fastest restores of recent data because the most recent backups are stored in their complete form
- Performs restores of data from earlier versions at about the same speed as 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

Tape Copy Performance

Many organizations seek a combination of disk-based backup for primary backups and tape for disaster recovery. It's important that a disk-based backup system provide high levels of offsite tape copy performance in order to free up all resources and also get the tapes offsite as quickly as possible.

First Generation Approach
With first generation systems, data is de-duplicated in line, so the disk only contains unique blocks, and a hash table is used to determine how to re-assemble the data.

Pros

- There are no strong arguments for this approach.

Cons

- This approach does not provide rapid tape copy because the data must be re-assembled from blocks before a copy can be sent to tape.

ExaGrid's Second Generation Approach
ExaGrid's second generation disk-based backup systems store the most recent backup in its complete form with 2 to 1 data compression. Because tape copies are always made from the most recent backup, this approach is fast because the most recent backup is always available in its complete form to make a tape copy.

Pros

- No additional disk is required.
- Fast tape copy performance at the lowest price.

Cons

- There are no strong arguments against this approach.

Net Result
ExaGrid's second generation approach:

- Provides the fastest tape copy performance because the most recent backups are stored in their complete form and ready to copy to tape.
- Provides the best tape copy performance at the 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 no cons to this approach.

Second Generation Systems
Second generation systems also perform checksums along all data paths, while also providing the fastest backups, restores and tape copy along with the lowest price.

Pros

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

Cons

- There are no cons to this approach

Net Result
ExaGrid's second generation disk-based backup system:

- Provides the same level of data integrity as first generation technology
- Provides the fastest backups, restores and tape copy performance at the lowest price

Ability to Provide a Second Site for Offsite Tape Replacement 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 offsite for disaster recovery purposes. 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.

Pros

- This approach works well for two-site configurations

Cons

- First generation systems are not as efficient in two-site configurations because the processor must be shared across two processes, de-duplication and replication, making backups slow. Additionally, 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.

ExaGrid's Second Generation Approach
ExaGrid's second generation systems are more efficient when used in a two-site configuration because only unique bytes traverse the WAN.

Pros

- This approach is WAN efficient and allows for a second system to be kept up to date at an alternate site so it is always ready to restore data.

Cons

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

Net Result
ExaGrid's second generation approach:

- Provides a WAN-efficient approach to maintain a "second," offsite system
- Provides fast restores of data from the second site
- Includes two-site replication software 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 dependent on how much the vendor charges for disk shelves.

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.

ExaGrid's Second Generation Approach
ExaGrid's second generation disk-based backup systems come packaged with servers that contain processor, memory, bandwidth and disk.

Pros

- As data grows and servers are added to the system 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

- If servers are added to the system, the cost could be higher than if only disk shelves were added. However, the ExaGrid system is the lowest cost system at all levels, from 1TB to 20TB in a single system.

Net Result
ExaGrid's second generation approach:

- Provides resources with every server to maintain performance
- Seamlessly virtualizes into ExaGrid's GRID 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 fails, some vendors have to send staff onsite to replace failed components, adding time to the process and leading to frustration.

Pros
- None

Cons
- IT staff frustration, wasted time.

ExaGrid's Second Generation Approach
Second generation systems feature customer support personnel and procedures that underline the critical nature of data protection. ExaGrid's customer support staff members are all ExaGrid employees and are based in the US, and each are responsible for named accounts and are responsible for their success. ExaGrid customer support staff members are proactive and knowledgeable about customer installations and about backup methods.

The ExaGrid system features customer replaceable components, including hot swappable drives and power supplies. The system is redundant and includes RAID6 with a spare drive and power supply set. If a drive, two simultaneous drives, or a power supply fails, the system will continue running. ExaGrid will ship a replacement component via next business day air.

Pros
- ExaGrid provides knowledgeable, proactive customer support that is among the best in the industry along with robust system with redundant, customer replaceable components for high availability.

Cons
- None

Net Result
ExaGrid's second generation approach provides:
- US-based support staff
- Proactive, knowledgeable support staff members assigned to named accounts
- Redundant and hot swappable hardware components

Conclusion

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.

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

- Data de-duplication method
- Backup performance
- Restore performance
- Tape copy performance
- Data integrity
- Availability of a second-site system
- Scalability
- Customer support

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

ExaGrid's turnkey disk-based backup system combines high quality SATA drives with byte-level delta data de-duplication, delivering a disk-based solution that is more cost effective than standard SATA drives. ExaGrid'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 standard SATA drives.

ExaGrid is easy to install and use and works seamlessly with popular backup applications, so organizations can retain their investment in existing applications. ExaGrid can be used at a primary site while maintaining tape for offsite or can be deployed as a two site solution to eliminate offsite tapes with a live data repository or for disaster recovery. When a second site is used, the cost savings are even greater because ExaGrid's byte-level data de-duplication technology moves only changes, requiring minimal WAN bandwidth.

About ExaGrid

ExaGrid is the leader in cost-effective disk-based backup solutions. A scalable system that works with existing backup applications, ExaGrid is ideal for companies looking to quickly eliminate the hassles of tape backup while reducing their existing backup windows. ExaGrid's innovative approach minimizes the amount of data to be stored by providing standard data compression for the most recent backups along with byte-level data de-duplication technology for all previous backups. Customers can deploy ExaGrid at a primary site and at a second site to supplement or eliminate offsite tapes with a live data repository or for disaster recovery.

ExaGrid Systems, Inc. | 2000 West Park Drive | Westborough, MA 01581 | 1-800-868-6985 | www.exagrid.com

Cost Effective Disk-based Backup with Data De-duplication

http://www.exagrid.com
1-800-868-6985