

## Moving to Disk-based Backup? Seven Key Questions to Ask



IT departments of all sizes are grappling with the problem of how to backup increasing amounts of data. For many years, magnetic tape has been the easiest and most cost-effective medium to backup data, however dealing with tape is cumbersome and time consuming. And as the amount of data that organizations need to backup grows, so does the time it takes to backup that data with tape.

Fortunately, the cost of SATA drives has fallen dramatically, making backing up to disk a reality for IT departments of all sizes. Backing up to disk instead of tape eliminates the hassles of tape, provides remarkably faster backups and restores, improved security and streamlined operations. However, choosing the right disk-based backup solution can be daunting. This white paper will discuss key considerations in moving to disk-based backup, and will examine the pros and cons of popular approaches.

There are three different approaches to disk-based backup. All approaches eliminate cumbersome tape, and provide some or all of the possible benefits of backing up to disk...

- **Straight storage:** Typically SATA drives in a DAS, NAS, or SAN form factor. These are the same types of products used for primary storage.
- **Basic disk-based backup systems:** Re-packaged straight storage positioned as disk-based backup. These products may have some additional features such as simple 2x compression, but are more similar to straight storage.
- **Turnkey disk-based backup systems:** Turnkey solutions based on SATA drives with data de-duplication for efficient storage of all backup history, and with optional support for off-site disk-based retention. These systems are built specifically for use with your backup application and have a variety of important features to improve your backup operations.

With all three of the above options, you will have the ability to leverage your existing backup application infrastructure. In fact, all of the major backup application vendors have added support for backing up to disk within the last couple of revisions.

**When evaluating disk-based backup solutions, there are seven key questions to consider:**

1. Does the system allow you to keep all of your backup history on disk cost-effectively?
2. Does the system support an off-site, disk-based copy of your backup data for disaster recovery?
3. Is the system turnkey, and does it include all the required hardware and software?
4. Is the system simple to install and manage?
5. Does the system give you maximum backup and restore performance?
6. Can the vendor size it to your current needs and accommodate some growth without selling you more or less initial capacity than you need?
7. As your data grows, can the system cost-effectively scale to continue to meet your needs without disruption and complexity?

Now, let's examine each of the questions in detail.

## 1. Does the system allow you to keep all of your backup history on disk cost-effectively?

Another important thing to consider when choosing a disk-based backup solution is its ability to store your backup history. If you plan to keep only one or two weeks of full copies of your backup data onsite on disk, then you may find that buying straight SATA drives is a workable solution, despite the complexity and management overhead we will discuss later in this paper.

However, if like most companies, you plan to keep four or more weeks of full copies of your backup data onsite, SATA drives may prove to be very expensive. The reason for this is that each time you store a full copy of your backup data on standard SATA drives, it consumes the amount of space equivalent to the size of the full backup. Consider the following simple math:

- You have 1 TB of primary data on which you do a full backup on a weekly basis
- You plan to keep 10 weeks (fulls) of this data on disk
- You require RAID protection and some storage management software

If you take the 1 TB of primary data times the 10 weeks of fulls and then add 30% overhead for 4 days of incrementals, the RAID and storage software, you will see that you will need 13 TB of raw disk to meet your backup goals (1 TB X 10 weeks = 10 TB + 30% overhead = at least 13 TB). In this case, would be spending several times more money on your backup storage than on your primary storage.

A disk-based backup system that uses data de-duplication technology, such as byte-level data de-duplication combined with standard compression, can store that backup data using much less SATA drive space. Byte-level data de-duplication stores only the changes from backup to backup instead of storing full file copies, dramatically reducing the amount of disk space required. Byte-level data de-duplication takes advantage of the fact that most data in a backup has not changed since the last backup. In fact, industry estimates put the data change rate between backups at 2% at the byte level. So if 98% of the data in 2 subsequent backups is the same, why store it twice?

To reexamine the above scenario:

- You have 1 TB of primary data on which you do a full backup on a weekly basis
- You plan to keep 10 weeks (fulls) of this data on disk – with data de-duplication, the latest backup will be compressed to 500 GB while all prior backups will be reduced to 20 GB (using the 2% change rate)

With this scenario, you have a full backup of the 1 TB of primary data with the last backup compressed 2X to 500 GB, 9 additional copies of the data at 20 GB each (only the bytes that change between backups) for a total of 680 GB of space needed to store the data. Compare this to the 10 TB of space you would have needed to store the data on straight SATA drives without data de-duplication.

## 2. Does the system support an off-site, disk-based copy of your backup data for disaster recovery?

In addition to the ability to write to disk, all of the major backup applications support the ability to make a copy of the backup data to tape for long-term storage. This is called disk-to-disk-to-tape (D2D2T). What this means is that you can choose to continue your off-site tape strategy. It is important that any solution can include disk for onsite backup along with tape for offsite copies.

However, if your goals are to dramatically reduce, or even eliminate your reliance on tape, then you need to select a disk-based backup system that provides you with a path to that end. That path has to include the ability to transfer your backup data to a second location that meets your disaster recovery objectives as they relate to geographic separation of your data and recovery time objectives.

In most situations, transferring your backup data to a second location will involve a wide area network (WAN). If the disk-based backup system you select does not have the appropriate technology to efficiently use your WAN, then having a second-site copy of your backup data for redundancy and disaster recovery is just not feasible. Standard replication can replicate the latest copy over the WAN, but in order to efficiently backup up all historical copies over the WAN, backup technology is required. Consider the challenge of moving 1 TB of data over the WAN weekly or during a backup window!

A disk-based backup system that is using byte-level data de-duplication for storage efficiency can also use that technique to maintain a fully redundant copy of your data across a WAN. Now, the system is only required to send the 20 GB worth of changes to the second site where the second site system can pour those changes into the prior backup to construct the most recent full. This technique allows you to keep all of your off-site retention on your disk-based backup system. Doing this reduces or eliminates the issues associated with tape and provides for much faster restores off-site as well as on-site.

With this type of disk-based backup system, you can also backup primary data located at your second site and have it sent back to the first site as a secondary copy. This results in cross-site protection for your critical business data at both sites.

### 3. Is the system turnkey, and does it include all the required hardware and software?

In implementing disk-based backup, the first option is to simply buy SATA-based storage. However, there are a variety of reasons why that may not work in your environment. One of those reasons is that buying raw disk for backup requires you to assemble the system and then manage that storage. You have to consider the type of storage it will be, the level of RAID protection you will add, and so on. You also have to determine how you will manage that storage. From creation of the initial volumes through to ongoing capacity management, you will have added unnecessary complexity while solving your current backup problems.

In addition to the above, remember the capacity requirements and resulting expense we discussed earlier. You will be paying many times more to store your backup data than you did to store your primary data.

A purpose-built disk-based backup system with compression and byte-level data de-duplication will include all of the hardware and software you need to plug it in and use it for backup. It will include:

- A simple NAS interface to which your backup application can write
- RAID 6 with a hot spare to ensure the highest reliability and the fastest recovery time from a drive failure
- Compression and byte-level data de-duplication software
- Software that makes the system self-managing
- Software that maintains historical copies offsite while only sending byte-level changes over the WAN
- A simple alerting mechanism to let you and the vendor's support organization know of anything that may need attention

When done right, a disk-based backup system will plug into your environment and silently protect all of your backup data with no ongoing management. Once fully implemented, your backups should not only take less time to complete but also take much less of your time to manage.

### 4. Is the system simple to install, and manage?

Backup is an insurance policy and does not drive your company's business results. Therefore, you should spend as little time on it as possible while still being fully protected from data loss. A disk-based backup system should require only minutes to install, configure, and be ready to accept backup data from your existing backup application. If it needs anything more than an IP address, a system name, and the creation of one or more destination NAS shares to get online, then it is too complex.

Once installed, the system should require almost no on-going management. It should worry about capacity and backup verification and simply alert you when something needs attention. A properly designed disk-based backup system should reduce the amount of time you spend managing your backups.

## 5. Does the system give you maximum backup and restore performance?

The architecture and design of a disk-based backup system will greatly affect how it performs in your network. As with a backup server, performing compression and data de-duplication on the fly will significantly slow down the backup process. Some of the systems get around this problem by providing enough disk space to receive the backup job without processing it on the fly, applying the compression and data de-duplication after the backup is complete. This ensures backup jobs run at the speed of disk but also gain the benefit of storage efficiency.

You also need to consider performance on the restore side as the various types of data de-duplication differ in approach. Data de-duplication products which use the block-level de-duplication technique must reassemble the file from small (8kb) blocks spread all over the disk causing a "restore fragmentation" slow down, whereas byte-level data de-duplication based systems keep the most recent job in it's full form, and process only one or more small change records ("deltas") for each previous version when a job is restored.

## 6. Can the vendor size it to your current needs and accommodate some growth without selling you more or less initial capacity than you need?

Obviously, you do not want to overspend for a disk-based backup option. Therefore, you should work closely with the vendor to ensure that the vendor can size the solution appropriately for your current data while allowing for reasonable near-term growth in your initial system. Many vendors want to keep things simple for themselves by carrying a limited number of configurations and as a result, customers have to buy more than they need upfront or start with a solution that is already too small for their needs. The last thing you want to do is have to go back and buy more capacity three months after the initial purchase.

Sizing a system appropriately for your environment is made easier when you have appropriately-sized building blocks available to you. If a vendor can size your system in 1 TB increments (i.e., 1 TB, 2 TB, 3 TB, 4 TB, and 5 TB), there is a high likelihood they can find the optimal capacity and price to meet your needs.

For example, let's say you are currently backing up 2.5 TB of data in your weekly fulls. A system rated for 3 TB of primary data would easily fit your existing data and allow for 20% growth before requiring you to add capacity. If the vendor did not have a 3 TB option and the next available size was 5 TB, you would be buying twice the capacity you need at the outset or trying to fit into a system that is too small or provides no room for growth.

## 7. As your data grows, can the system cost-effectively scale to continue to meet your needs without disruption and complexity?

Data growth is simply a fact of life. Most companies are going to see moderate to significant growth of their primary data year over year. So, in addition to buying the right sized solution at the outset, you need to fully understand how the vendor plans to support your ongoing data growth. With straight SATA drives, the answer is painfully obvious — you will constantly buy more SATA drives. You will manage how you configure, load balance, and administer the capacity. And most importantly, the expense will grow right along with your data.

However, if you are considering a system that utilizes data de-duplication, then there are additional considerations:

1. Because data de-duplication does require some CPU and memory to maintain performance, how do you ensure the system continues to perform as you add capacity?
2. Can you add capacity in increments that match your needs or do you have to overbuy each time?
3. What is required to install and configure the additional capacity?
4. When you add capacity to your system, do you continue to get the full benefit of your data de-duplication?

### The last two questions are critical to ask.

Remember, your initial system is nearly full. With an implementation that does not scale, you may find that you have to start subdividing and reconfiguring your backup jobs to use the little remaining space on the initial system and redirect the rest of your backup jobs to the new system. You now have essentially two separate systems to manage. You have to watch and manage the available capacity of each.

In reality, the above implementation does not handle your data growth at all, and you are simply buying another separate system. This approach leads to problems that are not that different from the headaches caused by having pockets of isolated, direct-attached storage for your primary data.

The other problem with the above approach is that the new system does not contain any of the backup or data de-duplication history from the original system. Generally, disk-based backup systems with data de-duplication deliver the most significant benefit after weeks of history. With new, isolated capacity you are back to square one on data de-duplication.

To avoid these problems, look for a system where added capacity virtualizes into a single pool of storage for your backup jobs. In this type of architecture, each time you add a server for additional capacity it communicates with the previously installed servers. A product with this architecture can:

- Support servers that simply plug in and create a virtual pool of capacity along with your initial servers with little or no configuration
- Automatically relocate data to available space while requiring no change to the backup job configuration
- Provide a simple utility that allows you to migrate some of your data to the new capacity if you wish to do so
- Take advantage of the backup and data de-duplication history stored on the initial capacity because it is all part of the same virtual pool
- Automatically manage overall capacity for the whole system without your intervention

## Conclusion

There are a number of choices available for disk-based backup. It is important to understand:

- Your retention goals on-site – how much backup data you want to keep on disk onsite?
- Do you plan today or in the future to use disk-based backup with a system at a second site to augment or replace your off-site tape strategy?
- Is it a turnkey system?
- Is it simple to install and manage?
- The performance of the system for backup and restore
- Can the vendor size the system appropriately when you purchase?
- How will the system handle your on-going data growth? Is it a scalable architecture or will you be adding separate, isolated systems that you have to manage?

Fully understanding these issues will help you and your organization determine the right approach to disk-based backup for your unique situation.

## Disk-based Backup Check List

- Uses existing backup server – no change to current backup application
- All backups are controlled by your existing backup server
- All restores are controlled by your existing backup server
- 100% turnkey including all hardware and software
- Includes management software
- Includes RAID6 with a spare SATA drive
- Cost-efficient with compression and byte-level data de-duplication
- Full hardware maintenance program – next business day replacement of all hardware
- Allows for appropriate initial sizing of system
- Allows for long term scaling as data grows
- Allows for disk onsite with tape offsite
- Allows for disk onsite with disk offsite, sending only the byte-level changes over the WAN
- Allows for only certain data to be kept at the second site
- Set up time is less than 30 minutes – plug the disk-based backup system into the backup server via Ethernet and point existing backup jobs to simple NAS shares

## About ExaGrid

ExaGrid® offers a turnkey appliance that works in conjunction with your existing backup applications and is 30 percent the cost of standard SATA storage. ExaGrid provides next generation byte-level data de-duplication technology, which stores only byte-level changes for each version instead of storing full file copies. This unique approach reduces the amount of disk space needed to at least 20 to 1, resulting in a significant cost savings over standard SATA storage.

Compliments of:



**3835R E. Thousand Oaks Blvd. #315**  
**Westlake Village, CA 91362**  
**Tel 877.230.2837 ~ Fax 805.435.2500 ~ [www.ESS-Direct.com](http://www.ESS-Direct.com)**