

Five Key Considerations for Evaluating the Scalability of Disk-based Backup Solutions

December 2006



Data growth is a fact of life for IT departments. As your business grows, so does the amount of data it generates, and the amount of storage capacity needed to properly retain the data cascades out of control. Organizations of all sizes face the dilemma of how to backup increasing amounts of data while reducing the hassles of traditional tape-based systems. Fortunately, the cost of SATA disk has fallen in recent years, making disk-based backup systems affordable for companies of all sizes. In addition to reducing or even eliminating reliance on tape, disk provides many advantages over tape, including faster and more reliable backups and restores.

In selecting a disk-based backup solution, it's critical to consider the scalability of each prospective solution to ensure that it will meet the needs of your organization now and into the future. This paper outlines the five key questions you should ask when evaluating the scalability of any disk-based backup solution.

1. Can the system size to your current requirements and accommodate near-term data growth without requiring you to purchase more initial capacity than you need?
2. Does the system enable you to keep all of your backup history and retention on disk cost-effectively?
3. As your data grows, can you cost-effectively add capacity to the system without disruption or added complexity?
4. Does the system maintain leading backup and restore performance as capacity is added?
5. Does the system's core architecture scale along with your data?

Selecting a disk-based backup solution can be daunting, however most systems fall into one of three categories:

- Straight storage solutions: general purpose storage, such as SATA-based disk arrays
- Purpose-built backup storage systems: straight storage products that are sold with supporting features such as 2:1 data compression and packaged as backup appliances
- Disk-based backup systems with data de-duplication: turnkey solutions that incorporate advanced data de-duplication technologies and enable you to store all backup history and retention in a small amount of disk space

1. Can the system size to your current requirements and accommodate near-term data growth without requiring you to purchase more initial capacity than you need?

Any system you choose should be able to be sized appropriately to handle the amount of data that needs to be backed up today while allowing for reasonable growth in the near-term. However, many vendors offer only a limited number of configurations, and you should work with a vendor who can assure that the system isn't already too small or more than you need at the time of installation.

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, there is a higher likelihood that the system can be sized appropriately for your needs today and into the future.

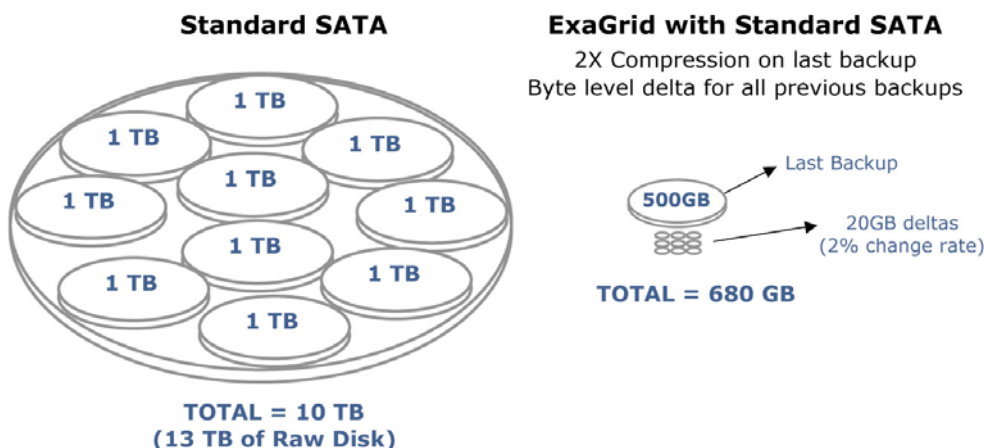
For example, if you are currently backing up 2.5 TB of data in your weekly fulls, a system rated for 3 TB of primary data would easily accommodate your existing data and allow for 20 percent growth before requiring additional capacity. If the vendor does not offer a 3 TB option and the next available size is 5 TB, you will be required to purchase a system that provides twice the capacity needed at the outset.

2. Does the system enable you to keep all of your backup history on disk cost-effectively?

One benefit of disk-based backup systems is the ability to store all of your backup history and retention on disk cost-effectively. If you plan to keep only one to two full copies of your backup data on disk, then you may find that buying a standard SATA-based storage solution is a cost-effective and workable solution, despite the complexity and management overhead associated with straight storage.

However, if like most companies, you plan to keep four or more full copies of your backup data, then straight SATA storage will prove to be very expensive. To properly size the system, you must thoroughly consider your future retention needs. For example, if you plan to only retain two full copies of your backups, and then decide to keep four or more copies at a later date, you will need to purchase much more capacity. In a case where four or more copies of backups are to be kept on disk, a solution that uses data de-duplication to reduce the amount of disk space required will significantly keep costs down.

Consider the simple diagram below. A disk-based backup system without data de-duplication takes over 13 TB to store ten full backups of 1 TB of data (10 TB for data plus 30 percent overhead). By using data de-duplication technology, such as byte-level data de-duplication combined with last backup compression, the 10 TB of data can be stored in only 680 GB of space. The result is an affordable disk-based backup solution that can store all of your backup history.



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

A system's scalability is key to a successful implementation, and it's important to fully understand how any system will support your ongoing data growth. With straight SATA storage, the answer is simple: you will buy more SATA storage. You will also manage how to configure, load balance, and administer the capacity. You will have to configure additional volumes and redirect most or all of the backup jobs to the new capacity. And most importantly, the expense will grow right along with your data.

If you are leaning toward a system that uses data de-duplication, then it is important to understand how to add capacity to the system as your data grows, when it should be added, and how disruptive it will be to the existing configuration. Additionally, you must consider whether the newly expanded system appears as another isolated pool of storage or whether it joins as a single virtual pool.

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 the backup jobs to use the little remaining space on the initial system and redirect the rest of the backup jobs to the new system. There are now essentially two separate systems to manage, and the available capacity of each must be monitored and managed.

In reality, this type of implementation does not handle data growth effectively: it is simply another separate system. This approach leads to a scenario that is not that different from the headaches caused by having pockets of isolated, direct-attached storage for primary data. The systems become separate, totally unrelated islands of storage for your backup data.

The other problem with this 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 when repeated data can be found and eliminated. Therefore, it's important to determine how the data is managed between your original and newly expanded capacity.

To avoid these problems, look for a system where added capacity virtualizes into a single pool of storage for all backup jobs. In this type of architecture, each time a server is added 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, ensuring that all available space is used as efficiently as possible
- Provide a simple utility that allows the migration of some of the data to the new capacity so the initial load balancing may be controlled

- 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, reducing the storage required and the cost
- Automatically manage overall capacity for the whole system without intervention so that added data does not mean added management time

4. Does the system maintain leading backup and restore performance as you add capacity?

Systems that use data de-duplication do require some CPU cycles and memory to quickly perform the task of efficiently storing data. The CPU and memory requirements increase as the amount of data that is being processed increases. Therefore, any scalability strategy should not only add drive space but also should add processing power and memory. When combining an implementation where all newly-added capacity virtualizes into a common pool of storage and the added capacity comes with additional CPU and memory, the result is a system where performance scales with capacity.

5. Does the vendor's core architecture scale along with your data?

In addition to system scalability, you should also ensure that the system's core architecture can scale to meet your current and future backup needs. With systems that use data de-duplication, the choice of data de-duplication method is a major differentiator. Consider the two most common methods of data de-duplication employed today:

- Data de-duplication, which reads blocks of data as it is written and stores only unique blocks
- Byte-level data de-duplication, which compares like files and stores only the differences

For data de-duplication to be effective, it breaks data into small "chunk sizes" so that it can get a maximum number of like segments. These chunk sizes are generally around 8 kb in size. Data de-duplication keeps pointers to these unique segments in a hash table. Over time, problems can develop, including:

- The hash tables get too large and create a limit on how much data can be handled
- Hash collisions may occur where two different segments result in the same hash
- Restores become very fragmented.

Byte-level data de-duplication performs its work by comparing versions of data to itself. For example, if a 100 kb PowerPoint presentation is backed up, and then the next week, the same file is backed up with only 2 kb of changed data, byte-level data de-duplication handles the file more effectively. Byte-level data de-duplication technology compares the two files, determines they are nearly identical, and stores the most recent copy and the byte-level changes required to reconstruct the previous copy. Because byte-level data de-duplication reduces the data by comparing like files and storing only the changes, there is no reliance on creating very small chunks of data with no use of ever-growing hash tables needed to reconstruct the data at a later

time. Also, byte-level data de-duplication stores data in larger 100 MB segments. When restoring files, there is no need to reassemble them from tiny chunks scattered throughout the disk, so there is no “restore fragmentation” problem. In fact, the most recent backup is stored in its entirety using only simple compression, and most restores are from the most recent backup. With byte-level data de-duplication, your restores will be significantly faster.

Conclusion

There are a number of choices available for disk-based backup. It is extremely important to understand how a system will scale as your data grows. Make sure you cover the five key considerations in your analysis:

1. Can the system size to your current needs, including near-term growth, without including more initial capacity than you need?
2. Does the system allow you to keep all of your backup history and retention on disk cost-effectively?
3. As your data grows, can you cost-effectively add capacity to the system without disruption or added complexity?
4. Does the system maintain leading backup and restore performance as you add capacity?
5. Does the system’s core architecture scale along with your data?

A system that does not grow with your needs will prove to be expensive and time-consuming down the road. If you keep these considerations in mind, you will select a system that will serve you well for years to come.

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 disk. ExaGrid provides next generation byte-level data de-duplication, 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.

ExaGrid Systems, Inc.

2000 West Park Drive
Westboro, MA 01581

1 800.868.6985
www.exagrid.com

