

# Evaluating Deduplication Solutions?

**YOU NEED  
TO NOT ONLY  
UNDERSTAND  
WHAT  
DEDUPLICATION  
IS, BUT MORE  
IMPORTANTLY  
KNOW THE  
VARIOUS WAYS  
THAT IT IS  
IMPLEMENTED  
AND WHAT  
EACH OF THOSE  
MEAN FOR YOUR  
ENVIRONMENT.**

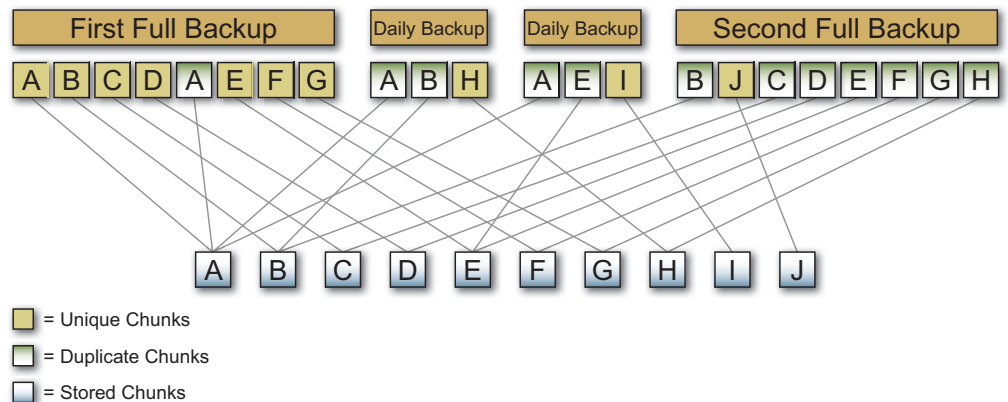
## What you Really Should Consider

Deduplication is one of the hottest technologies in the market today because of its ability to reduce costs. In order to make the best decision among the many deduplication solutions, you need to not only understand what deduplication is, but more importantly know the various ways that it is implemented and what each of those mean for your environment.

In this paper, we will explain what elements you should consider and how the right decision will make your life easier by reducing the amount of storage you have to manage, maximizing your flexibility, and shrinking your backup windows.

## What is Deduplication?

The Storage Networking Industry Association (SNIA) defines deduplication as the process of examining a data set or byte stream at the sub-file level and storing and/or sending only unique data<sup>1</sup>. With deduplication you are only storing the unique data because duplicate data is replaced with a pointer to the first occurrence of the data. Deduplication carves data into chunks. These chunks are hashed and the resulting unique identifier is compared to an index of all unique identifiers. If the identifier already exists, the data is a duplicate and is replaced with a pointer to the correct unique identifier in the index. If your backup is 100GB and 20GB is redundant or duplicate data you would only store the unique 80GB. When the unique data is then compressed, you can see even more substantial disk space savings.



Deduplication is divided into multiple types - Block Level and Byte Level. While block level only looks at the blocks, byte level deduplication does a byte-by-byte comparison of the data and compares them to the index for duplicates. Byte level deduplication is currently the only method that can guarantee full redundancy elimination.

## Single Instance Store vs. Deduplication

SNIA defines Single Instance Store (SIS) as the replacement of duplicate files or objects with a reference to a shared copy<sup>1</sup>. While Single Instance Storage is not defined as a deduplication technology by SNIA, many people confuse SIS with deduplication.

The key difference between SIS and deduplication is that SIS evaluates the data stream at the file level by looking for duplicate files while deduplication evaluates the data stream at the block or byte level. With SIS, even a small change to a file will cause it to be seen as a new and different file and be stored again. This means, that if you have a document and a user copies or renames the file, it will be seen as an entirely new file and will be stored again. With deduplication, the entire file contents will be seen as duplicate. As a result SIS delivers less space savings versus deduplication.

## What Difference Does the Deduplication Block Size Make?

When evaluating deduplication technologies, you need to look at the selection method used to identify the unique chunks. The two prevailing methods are fixed block length and variable block length. With smaller fixed block sizes, the same data stream will be divided into more chunks than if a larger block size is used. This leads to a higher percentage of duplicates being identified. With variable block length, the deduplication engine has the ability to change the block size and recognize more duplicate patterns thereby increasing the number of duplicate blocks.

WITH VARIABLE  
BLOCK  
LENGTH, THE  
DEDUPLICATION  
ENGINE HAS  
THE ABILITY TO  
CHANGE THE  
BLOCK SIZE  
AND RECOGNIZE  
MORE DUPLICATE  
PATTERNS  
THEREBY  
INCREASING  
THE NUMBER  
OF DUPLICATE  
BLOCKS.

Consider a file which has a small amount of changes made to it, such as adding a word to a sentence or removing a word. With fixed block deduplication, the rest of the file remains the same if the blocks do not line up exactly so they will be seen as unique and will not be deduplicated. With variable block length, the deduplication engine can isolate only the changed data and deduplicate the rest of the blocks, resulting in more storage savings.

In the following example where some of the words in the file were changed, you can see how fixed block deduplication will identify more blocks as unique whereas variable block sees more blocks as duplicates.

“It was a bright cold day in April, and the clocks were striking thirteen.”

**Original File**




It was	a bright	cold day	in April,	and the	clocks were	striking thirteen.
--------	----------	----------	-----------	---------	-------------	--------------------

**Fixed Block**

It was	a cold	day in	March,	and the	clocks were	showing thirteen.
--------	--------	--------	--------	---------	-------------	-------------------

**Variable Block**

It was	a	cold day in	March,	and the clocks were	showing	thirteen.
--------	---	-------------	--------	---------------------	---------	-----------

-  = Duplicate Block
-  = Unique Block
-  = Original Block

**DEDUPLICATION  
TYPICALLY IS  
PERFORMED  
INLINE OR POST-  
PROCESS. EACH  
HAS ADVANTAGES  
AND TRADE-OFFS.**

**Inline or Post Process Deduplication and Why it Matters to You**

Another technology differentiator to consider is when the deduplication processing occurs. Deduplication typically is performed inline or post-process. Each has advantages and trade-offs.

With inline deduplication, data is deduplicated as the deduplication engine receives it. Once the data is deduplicated, it is then immediately stored on disk. The advantage for inline deduplication is that it does not require any additional disk space to store the data prior to deduplication. However, inline deduplication has the following trade-offs:

- Performing the deduplication process as part of the backup will lengthen the time to complete the backup. Longer backups will elongate the backup window and can lead to degraded performance during business hours and the inability to start the next backup because the previous backup job is still running.
- Not all data deduplicates well, especially encrypted data. Inline deduplication does not allow the flexibility to leave such data in a non-deduplicated state while deduplicating the data that does deduplicate well. With inline deduplication, all the data will be deduplicated, even your encrypted data.
- Inline deduplication eliminates the flexibility to stage data before copying it to tape for offsite protection. This means that not only will every backup have to be rehydrated for restores; they must be rehydrated before being copied to tape.

- Restores are faster if performed soon after the backup is completed; however, with inline deduplication, if the data is immediately deduplicated, every restore will require rehydration or reassembly of the unique chunks into the original data stream.

With post-process deduplication, the backup is briefly placed on a disk-based staging area prior to the deduplication process. Some deduplication technologies require the entire backup to be staged before deduplication starts while others allow deduplication to start after a set amount of the data stream has been staged. This reduces the sizing requirements for the staging area while allowing the backups to complete as fast as possible. Post-process deduplication has the following advantages:

- Since deduplication is not part of the backup with post-process deduplication, backups will complete faster which shrinks your backup window.
- Enables you to leave data that does not deduplicate well in a non-deduplicated state.
- Post-process deduplication allows users to provision data on existing storage which can be up to 1/10th the cost of appliance storage.
- Restores are faster if performed soon after the backup is completed because the data has not been deduplicated yet and will not need to be rehydrated to perform the restore. Since restores typically come from the most recent backups, this enables you to speed up restores while still taking advantage of deduplication to reduce long-term storage costs.

With post-process deduplication, the only real trade-off is that it requires additional disk space for the staging area. The size of the staging area will depend upon how long data will remain in the staging area awaiting deduplication and how much data will not be deduplicated. The shorter the amount of time the data stays in the staging area, the smaller the area must be.

### **What is the Difference Between Source and Target-Side Deduplication?**

In addition to looking at when the deduplication occurs, you need to consider where deduplication is performed. There are two places where deduplication happens: either on the source/client or on the target/storage.

Source-side deduplication typically uses a deduplication engine that is located on the client that will perform the process of hashing the data and check for duplicates with a centrally located deduplication index, which is typically located on the backup server or media server. Duplicates that are found are not transmitted across the wire to the deduplication storage, but the unique blocks will be transmitted. The advantage of source side deduplication is that it reduces network contention because less data is sent over the network since it is only sending the unique data. While in some scenarios this can be beneficial, you also have to look at the drawbacks of this type of deduplication. By running source-side deduplication you are adding hashing, which is a processor intensive algorithm, to your client. This process will run on every client regardless of the computing power of the client. This means that clients that are already overloaded will become even more overloaded and this could have very unexpected results for your applications as well as slowing down the backups and lengthening the backup window.

Target-side deduplication removes the deduplication process from the client and runs the deduplication at the target/storage. The advantage here is that you do not need to worry about having clients with enough horsepower to handle the hashing algorithm. With target-side deduplication, the hashing happens at the target, and then the index is checked. The trade-off is that more data is going to be sent over the network, so if you have a network that is already congested, this method will add additional congestion.

whitepaper  
IN ADDITION  
TO LOOKING  
AT WHEN THE  
DEDUPLICATION  
OCCURS,  
YOU NEED TO  
CONSIDER WHERE  
DEDUPLICATION IS  
PERFORMED.

The other factor to consider with source vs. target-side deduplication is that target-side deduplication is better suited for environments where there are larger amounts of data. Many industry experts have said that while source-side deduplication is ideal for remote sites with smaller amounts of data to be deduplicated, target side deduplication fits better into environments where there are larger amounts of data that require deduplication.

Different deduplication technology vendors have created different solutions that may mix and match the when and where. For example, you may have a solution that does inline deduplication starting at the source, while others may do post-processing deduplication starting at the target. When evaluating deduplication solutions that best meet your needs, be sure to not only consider when the deduplication process is happening, but also where deduplication is occurring.

### **What Do Deduplication Ratios Really Mean?**

We have discussed the different types of deduplication, how they work, and some of the differences. Next, we are going to look at the results of deduplication and how these should be taken into consideration when evaluating deduplication solutions.

Deduplication is typically reported in a ratio as ratio: 1 or ratio X. For example 12:1 or 12X. The ratio is calculated as ratio = bytes in / bytes out. Ratios can be viewed as the data capacity of a system divided by its used storage capacity. If 500GB of data only consumes 50GB of storage, the deduplication ratio is 10:1.

Each vendor in the deduplication market has their own set of tests, and this leads to a different ratio for the vendors. This is where comparisons start to become problematic, as each test uses a different set of data, as well as a different set of assumptions. When dealing with deduplication ratios you have to understand that their significance is based on certain things:

- Ratios are only meaningful if they are compared when using the same set of assumptions
- Even low deduplication ratios provide significant space savings
- Higher ratios yield marginally less space reduction

Deduplication Ratio	Space Reduction Percentage
2:1	1/2 = 50%
5:1	4/5 = 80%
10:1	9/10 = 90%
12:1	11/12 = 91.67%
15:1	14/15 = 93.33%
20:1	19/20 = 95%
30:1	29/30 = 96.67%

As you can see above, once you hit the 10:1 deduplication ratio, the actual amount of disk savings does not increase significantly. If one vendor claims 20:1 ratio and the other claims 12:1 ratio, the two numbers may sound drastically different, but when you look at the actual amount of disk savings you find that there is less than 5% difference. This is why deduplication products must be evaluated on factors other than deduplication ratios.

**ONCE YOU  
 HIT THE 10:1  
 DEDUPLICATION  
 RATIO, THE  
 ACTUAL AMOUNT  
 OF DISK  
 SAVINGS DOES  
 NOT INCREASE  
 SIGNIFICANTLY.**

**white**

## How Long Should I Retain Deduplicated Data?

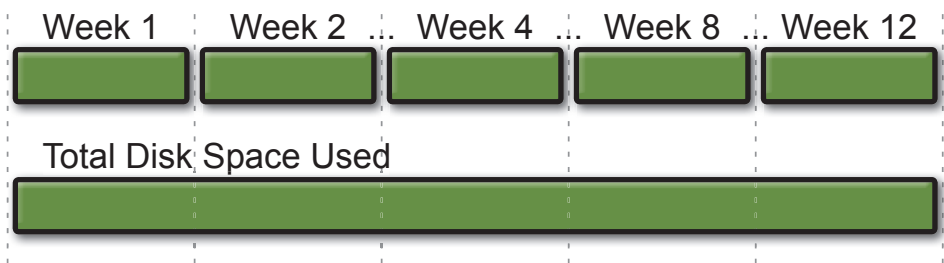
A final issue to be reviewed when evaluating deduplication technologies is deciding how long you are retaining data. The more data that is examined, the greater the likelihood that duplicate data will be found which will increase your disk space savings.

For example, the initial Full Backup that you deduplicate will only be deduplicated against itself and will result in a small amount of reduction in the storage footprint depending upon the type of data that is being deduplicated. When the Full Backup for Week 2 is performed, only the unique data that has been updated or added since Week 1 will be stored. When the Full Backup for Week 4 is performed, the chunks will be compared against all the unique data for Week 1, 2, and 3 which increases the changes that duplicate chunks will be found.

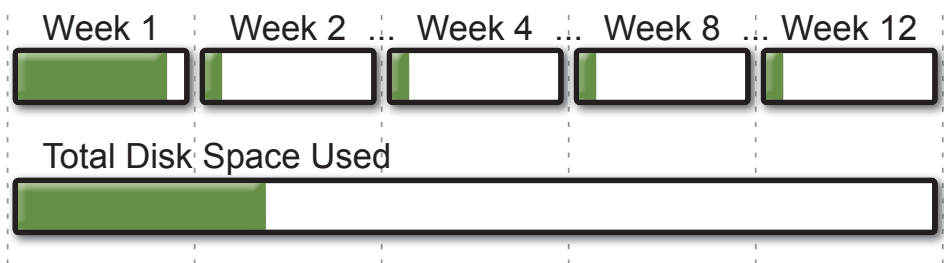
When you deduplicate your backups, each additional week of backups can be retained for a decreasing amount of additional disk space. This allows you to store even more backups on the same amount of disk-based storage for a longer period and virtually eliminates the need to restore from offsite storage unless there is complete site failure.

**THE MORE DATA THAT IS EXAMINED, THE GREATER THE LIKELIHOOD THAT DUPLICATE DATA WILL BE FOUND WHICH WILL INCREASE YOUR DISK SPACE SAVINGS.**

### Without Deduplication



### With Deduplication



## What Should I Consider in My Deduplication Decision?

Your goal(s) for your deduplication solution will influence which deduplication technologies you should evaluate. Following are some typical deduplication solution goals and considerations.

### **Maximum Disk Space Savings**

- Deduplication offers more disk space savings than Single Instance Store.
- Variable block deduplication typically provides better deduplication ratios than fixed block deduplication.
- Retaining deduplicated data longer will allow you to store even more backups on the same amount of disk-based storage for a longer period.
- Near inline or inline deduplication reduces disk space requirements since it does not require a large staging area or a staging area at all.
- Source-side deduplication can increase the disk space savings by examining a centralized index for uniqueness.

### **Maximum Flexibility**

- Variable block deduplication enables the technology to adjust to the data stream to find more duplicates.
- Post-process deduplication offers the ability to leave data that does not deduplicate well in a non-deduplicated state to ensure that you are not using valuable time and processing power on data that will not benefit from deduplication.
- With post-process deduplication, restores are faster when performed soon after the backup is completed because the data has not been deduplicated yet and will not need to be rehydrated to perform the restore. Since restores typically come from the most recent backups, this provides the flexibility to speed up restores while still taking advantage of deduplication to reduce long-term storage costs.
- Post-process deduplication allows users to provision data on existing storage which can be up to 1/10th the cost of appliance storage.

### **Shrink Backup Windows**

- Post-process deduplication is not part of the actual backup and can be scheduled to occur outside the backup window.
- With target-side deduplication, deduplication is also performed independent of the backup job and can ensure deduplication does not unnecessarily elongate backup windows.

## **NetVault® SmartDisk™ Advantage**

While there are many deduplication vendors in the market, only one vendor, BakBone Software stepped up to the plate and currently offers a Open Data Protection Platform (ODP) that extends users' data protection environments with the vision of providing an affordable and long-term disk-based storage solution that can remain in place regardless of who or where the data is coming from. This gives you the opportunity to consolidate your data protection solutions using the same deduplication solution, thereby reducing costs and providing more flexibility with regard to vendors.

With the initial release of NetVault: SmartDisk, the first product based upon ODP, there is seamless integration with NetVault: Backup which enables users to utilize NetVault: SmartDisk for disk-based backup and deduplication needs within an enterprise class, easy-to-use data protection suite.

NetVault: SmartDisk's deduplication option uses byte-level deduplication with a variable block size. The deduplication process for NetVault: SmartDisk happens post-process so that it does not interfere with the backup window and allows backup administrators to schedule the deduplication process to occur when they need it. NetVault: SmartDisk also allows for near-line deduplication by allowing the deduplication process to start before the entire backup is done for those who want to start the process earlier.

**YOUR GOAL(S)  
FOR YOUR  
DEDUPLICATION  
SOLUTION WILL  
INFLUENCE WHICH  
DEDUPLICATION  
TECHNOLOGIES  
YOU SHOULD  
EVALUATE.**

**white**

Since all data does not deduplicate well, NetVault: SmartDisk's integration with NetVault:Backup offers job-level deduplication allows you to create different selection sets, select which data you want to deduplicate and which data you do not want to deduplicate. This helps to ensure that you do not deduplicate data that does not lend itself to deduplication, such as encrypted files. The ability to create these different selection sets allows you to create the best data protection scenario for your environment and get the most benefit from the deduplication solution without having to waste valuable space and processing time on items that will see little benefit from deduplication.

In order to provide you with a solution that can be integrated into your environment, NetVault: SmartDisk is hardware agnostic enabling it to work on most file systems. This means that there is no need to acquire specific drives or expensive appliances in order to create a data protection strategy that includes deduplication. NetVault: SmartDisk also offers the ability to easily add additional file system paths to NetVault: SmartDisk Storage Pools reducing costs by deferring storage expenditures into new budget periods and ensuring that storage does not sit unused. In addition NetVault: SmartDisk has heterogeneous platform support supporting the most popular operating system platforms in the market.

NetVault: SmartDisk was designed from the ground up to give you more choices and maximize your investment. With NetVault: SmartDisk, you can deploy multiple NetVault: SmartDisk Instances to improve load balancing and performance, place the deduplication solution where you will see the most benefit such as in remote sites or at DR sites to help ease the burden of disaster recovery.

The NetVault family of products has a reputation for providing a solution that is not only enterprise class but easy to use. With NetVault: SmartDisk, we have continued that tradition by providing a deduplication solution that is easy to use and reduces the level of storage expertise required to perform deduplicated disk-based backups. This allows you to spend more time on the areas that you truly need to.

### **What Did We Learn?**

In conclusion, we saw that deduplication analyzed data at the block or byte level while Single Instance Store examined at the file level and learned that smaller, variable block selection size means more duplicates will be found for greater space savings. We observed that, while inline processing saves some space, it is not as fast or flexible as post-processing deduplication. Source side deduplication reduces network loads, but significantly increases the client processing workloads when compared with target-side deduplication. We noted that, while deduplication ratios may sound dramatically different, the actual difference in space savings is marginally smaller after about 10X. Finally, we found that the longer data was retained, the greater is the likelihood that duplicate data will be found.

We saw that NetVault: SmartDisk can be the right decision for you because it reduces the amount of storage you have to manage, maximizes your flexibility, and shrinks your backup windows.

But, don't just take our word, get more information and try the NetVault products out for yourself by taking advantage of the following resources:

#### **Learn More:**

- [www.bakbone.com/smartdisk](http://www.bakbone.com/smartdisk)
- [www.bakbone.com/backup](http://www.bakbone.com/backup)

#### **Evaluate NetVault: SmartDisk and NetVault: Backup**

- [www.bakbone.com/downloads](http://www.bakbone.com/downloads)

#### **Contact BakBone:**

- [www.bakbone.com/howtobuy](http://www.bakbone.com/howtobuy)

**THE RIGHT  
DECISION WILL  
MAKE YOUR  
LIFE EASIER BY  
REDUCING THE  
AMOUNT OF  
STORAGE YOU  
HAVE TO MANAGE,  
MAXIMIZING YOUR  
FLEXIBILITY,  
AND SHRINKING  
YOUR BACKUP  
WINDOWS.**



<sup>1</sup> [http://www.snia.org/forums/dmf/knowledge/white\\_papers\\_and\\_reports/Understanding\\_Data\\_Deduplication\\_Ratios-20080718.pdf](http://www.snia.org/forums/dmf/knowledge/white_papers_and_reports/Understanding_Data_Deduplication_Ratios-20080718.pdf)

©1999-2010 BakBone®, BakBone Software®, NetVault®, Application Plugin Module™, BakBone logo®, Integrated Data Protection™, NetVault: SmartDisk™, Asempra®, FASTRecover™, ColdSpark® and SparkEngine™ are all trademarks or registered trademarks of BakBone Software, Inc., in the United States and/or in other countries. All other brands, products or service names are or may be trademarks, registered trademarks or service marks of, and used to identify, products or services of their respective owners.

whitepaper

**BakBone Global Headquarters**

9540 Towne Centre Drive, Suite 100  
San Diego, CA 92121  
Toll Free Phone: 877-939-2663  
Phone: 858-450-9009  
Fax: 858-450-9929  
Email: [sales@bakbone.com](mailto:sales@bakbone.com)

**Asia Pacific Headquarters**

Shinjuku Dai-ichi-Seimei Bldg. 11th Floor  
2-7-1 Nishi Shinjuku, Shinjuku-ku  
Tokyo, Japan 163-0711  
Phone: 81-3-5908-3511  
Fax: 81-3-5908-3512  
Email: [sales@bakbone.co.jp](mailto:sales@bakbone.co.jp)

**Europe Headquarters**

100 Longwater Avenue  
Green Park  
Reading  
RG2 6GP  
United Kingdom  
Phone: 44 (0)1189-224-800  
Fax: 44 (0)1189-224-899  
Email: [sales\\_europe@bakbone.com](mailto:sales_europe@bakbone.com)