Barracuda

Barracuda Backup Deduplication

White Paper

Abstract

Data protection technologies play a critical role in organizations of all sizes, but they present a number of challenges in optimizing their operation. Barracuda Backup provides a fully-integrated, capacity-optimized solution that is simple to deploy yet robust and secure. Barracuda deduplication, an inherent capability of all Barracuda Backup products, lets organizations significantly reduce capacity needs, bandwidth requirements, and backup costs. For organizations protecting multiple sites, Barracuda's global deduplication and cloud storage technology help distributed networks stay protected while reducing the backup storage footprint.

Backup Storage Challenges

For as long as there has been data, there have been efforts to protect it, even as it has grown relentlessly. While data growth challenges are not new, the pace of growth has been accelerating rapidly and many organizations have attempted to solve their growing storage and retention needs by traditional means of backing up and storing on external media (e.g., tape, SAN/NAS). Such practices are often complex, time consuming, and prone to human and hardware failure. Using standard external media to store backups does not solve the storage problem effectively and ends up costing organizations more in the long run than investing in a more efficient solution at the outset. Capacity optimization plays an integral role in any compelling backup platform.

Beyond capacity constraints, many organizations cannot centralize their entire IT infrastructure, leaving remote sites either with IT staffing overhead or without sufficient, or any, protection. Tape and other external media do not scale effectively across sites and require additional intervention from local/remote IT staff to ensure consistent backups and to allow for growth. Organizations deploying site-to-site connections tend to back up the remote sites to a main data center, saturating bandwidth and causing long backup and restore delays. Technology like deduplication that can significantly reduce the amount of data needed to transmit can both avoid overloading corporate networks and reduce backup/restore times.

What is Deduplication?

Deduplication is a process that breaks down files and other data into "chunks" and uses a tracking database to ensure that only a single copy of that chunk is stored across all backup data. For subsequent client backups, incoming data is compared against the tracking database to determine which chunks have been protected and only transfers and stores unique chunks. For example, if five different servers are backing up data to a Barracuda appliance and a file chunk is found that exists on all five of those servers, only a single copy of the chunk is actually stored on the appliance, with small pointers tracking how that chunk should be rehydrated (recompiled) across all five devices during a restore. The tracking database ensures that these chunks are kept until all backups referencing a given chunk have been deleted. Since only the unique portion of data is stored by the server, there is a significant reduction in capacity needs. During restores, a file is rehydrated based on the information contained in the tracking database, then sent to the destination for recovery.

Barracuda Deduplication

There are different types of deduplication technology provided by suppliers and those differences can impact results for IT administrators. An older, more simplistic deduplication technology called post-process must wait for a backup job to finish writing to disk before initiating the deduplication and replication process, extending the time until data is fully protected. This increases the load on local systems as data must be addressed on disk three times before it can be replicated (written as backup data, read for deduplication, written as deduplicted data). Since post-process deduplication requires duplicate landing capacity, for the backup dataset landing and to store the deduplicated data, it drives inefficient capacity utilization. In some cases, vendors that rely on post-process deduplication must offload storage to external media such a tape or disk, in order to meet data retention requirements.

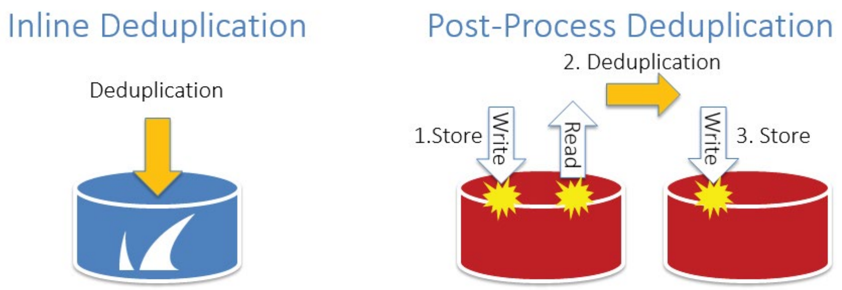
Instead of post-process, Barracuda Backup created its inline deduplication (Figure 1) technology. With inline deduplication, the appliance performs deduplication in one step as the data is ingested, eliminating the need for the superfluous landing-space capacity required by slower two-step post-process deduplication. Barracuda inline deduplication helps organizations save money by eliminating the need for a larger disk array dedicated to holding ingested data before deduplication can begin. This deduplication method also can help reduce risk of lost data by accelerating time-to-backup processing and full replication since data is queued for replication as the backup job is being processed.

Deploying a Barracuda Backup appliance significantly improves DR readiness by reducing time to get data offsite through inline deduplication and instant replication. Because data is deduplicated inline, it can be ready for replication more quickly than if it had to be processed after the backup process fully completed. And because there is no need to ingest the entire data set prior to replication commencing, data can be moved offsite as it is backed up and deduplicated, providing faster offsite protection.

Comparing post-process and Barracuda inline deduplication:

- Cost: Because post-process deduplication requires a landing space before data can be turned into a deduplicated state, dependent on the size of your data set, a larger and more expensive device is often required. With the Barracuda inline solution, a larger device is not necessary.
- Time: Post-process deduplication must wait until the backup job is finished before it can start the deduplication process and then replicate the data. Although this may appear to accelerate the backup process, in fact, data is not yet fully protected because the deduplication and replication process have yet to be completed. This serial processing and multi-stage activity can lead to significant delays to full data protection compared to real-time inline deduplication with simultaneous replication.
- Risk: With post-process deduplication, time to DR readiness is extended and any failures in three-step post-process activity can lead to a corrupt data set. Should a network problem occur or a site go down, despite the backup job reported as complete, data can be lost.

Figure 1. Target Deduplication



Barracuda's Deduplication Methodology

Barracuda Backup leverages three-stage deduplication to minimize bandwidth, optimize capacity usage, and reduce processing overhead of backups: Source, Target, and Global.

- **Source Deduplication**, where local data is deduplicated at the source and sent to the Barracuda Backup appliance in deduplicated form, minimizing LAN bandwidth and data sent to the local server. Also called client-side deduplication.
- **Target Deduplication**, where data is deduplicated directly on the backup appliance across sources, minimizing the amount of data that needs to be cached and replicated.
- **Global Deduplication**, where data is deduplicated across all local servers that have been replicated to a central appliance or cloud.

Source:

Source deduplication is implemented through the Barracuda Backup Agent. During its installation, a small database is created on the server to keep track of data chunks so only unique data seen by the agent is compressed and sent to the appliance for processing, reducing network traffic and the backup window. (Figure 2).

Figure 2. Source Deduplication

For VMware backups, Barracuda leverages VMware's vStorage APIs for Data Protection (VADP) to back up virtual disks. With VADP, Barracuda can use Changed Block Tracking (CBT) to send only unique chunks to the Barracuda Backup appliance (Figure 3).

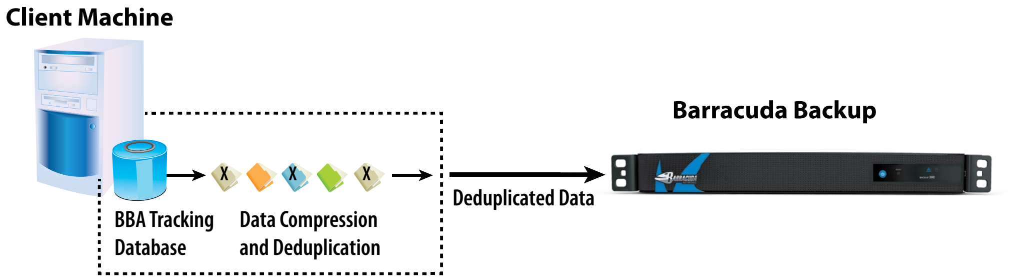
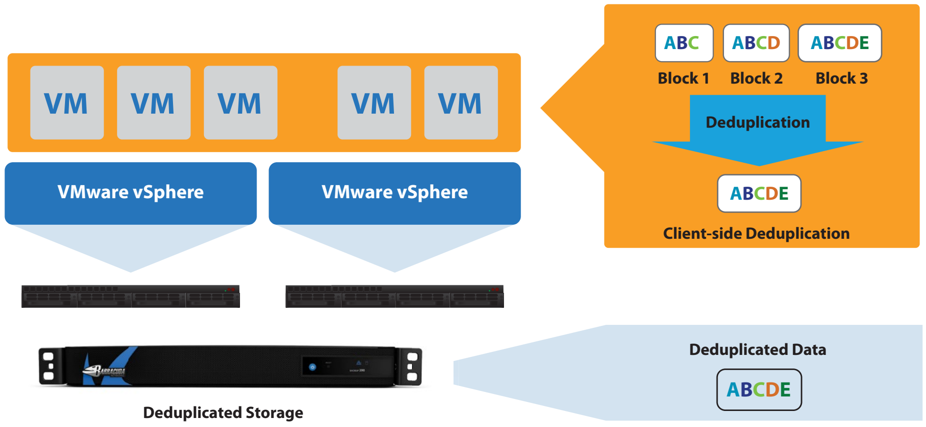


Figure 3. Changed Block Tracking

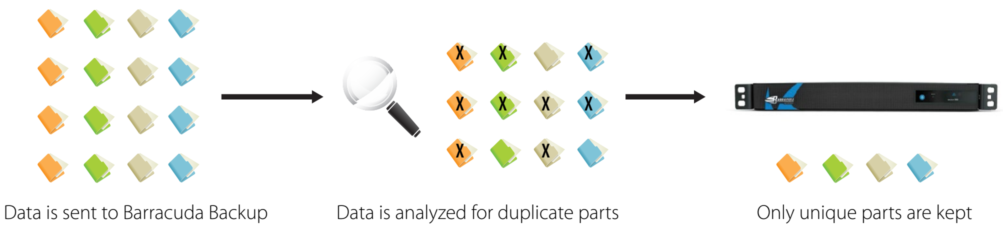
For Microsoft Hyper-V backups, the Barracuda Backup Agent reduces the backup window by deduplicating the VHD files on the host server to minimize the amount of data sent to the backup appliance.

Target:



Target deduplication occurs on the Barracuda Backup appliance to eliminate redundancy across all local agents and minimize the amount of local cache and cloud capacity needed to store backups. Organizations backing up SAN or NAS filers (Figure 4), VMware, and file share often cannot use source-based deduplication, leaving target deduplication as the primary methodology.

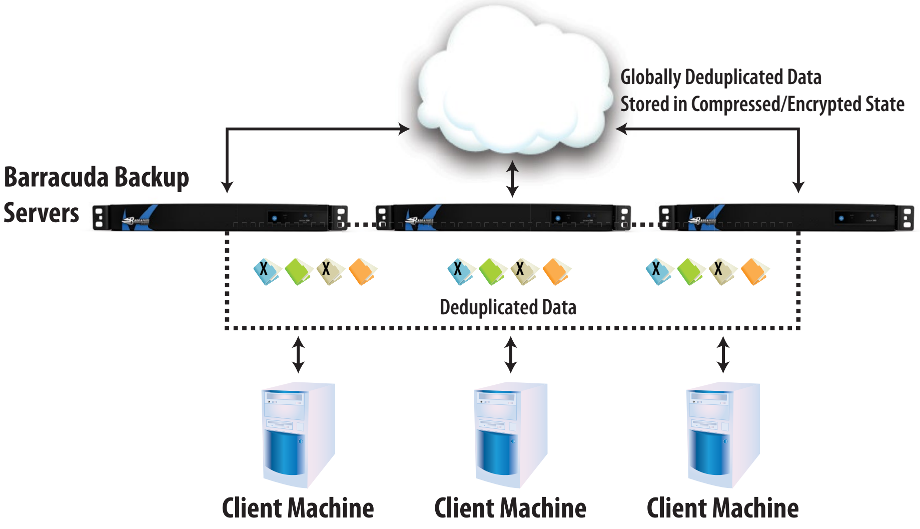
Figure 4. Target Deduplication.


Data is sent to Barracuda Backup
Data is analyzed for duplicate parts
Only unique parts are kept

Global:

Global deduplication is implemented on either an appliance used as a central replication target or in the cloud, eliminating redundancy across backup appliances throughout worldwide infrastructure and allowing organizations to reduce capacity needed to store backup data in a compressed and encrypted state (Figure 5).

Figure 5. Global Deduplication



Barracuda Backup leverages three-stage deduplication for multiple data sources as described in the following table:

Data Source Type	Backup Method	Deduplication Method
Microsoft Exchange Server Microsoft SQL Server Microsoft Hyper-V Microsoft Windows Microsoft Active Directory Microsoft SharePoint	Barracuda Windows Backup Agent	Source, Target, and Global
Lotus Domino Server	Windows Volume Shadow Copy Service (VSS)	Source, Target, and Global
Linux Systems	Barracuda Linux Backup Agent	Source, Target, and Global
Novell Open Enterprise Server 2 SP2.0+	Barracuda Linux Backup Agent	Source, Target, and Global
Mac OS X	Barracuda Macintosh Backup Agent	Source, Target, and Global
Unix	CIFS/SSHFS	Target, and Global
Network Addressable Storage	CIFS/Access Control Lists (ACL)	Target, and Global
VMware Server and Guests	VDAP with Changed Block Tracking (CBT)	Source, Target, and Global

Barracuda Deduplication Implementation

Barracuda Backup provides application-aware, variable length, block level inline deduplication for maximum data reduction and minimum capacity needs, reducing storage footprint, bandwidth requirements, and backup/restore times. For standard data sets backed up over time, users could see 20-50x data reduction, on average, from the three-stage deduplication process.

The length of backup data chunks used in deduplication is based on the type and size of the file. Each chunk is then given three unique hashes (digital fingerprints): MD5sum, SHA1, and size of the file. Each hash is unique for each chuck and is stored in a database by the Barracuda Backup Agent running on the local server along with a database on the local appliance. As the backup runs, each calculated hash value is compared to the values of those of chunks already processed, and if the value is unique, the chunk is transmitted to the appliance. For hash values already seen, only a small pointer is sent to the appliance. Once the data is added to the local Barracuda Backup appliance, the hashes are compared again across all agents. If duplicate entries are found, the appliance stores a single copy of the data on local appliance and makes note that it has been backed up, and can be restored to any server requesting the hash.

The following information should be observed for optimal performance of the Barracuda Backup Agent:

Deduplication Database Sizing:

The local deduplication database is a small portion of the total file system size. It increases linearly with total stored deduplicated data. When sizing, plan for roughly 1-3 GB of database size per TB of stored deduplicated data; e.g., 2 TB of deduplicated data equates to a 4 GB deduplication database.

Processor Utilization

The Barracuda Backup Agent can increase processor usage during a backup, because it uses the client machine's CPU in the source deduplication process along with compression. There is no limit set by the local Barracuda Backup Agent to limit the client machine's processor resources during backup and restore.

Memory Utilization

The Barracuda Backup Agent will increase system memory usage during a backup. The agent uses the client machine's memory to store backup data before sending the compressing and sending the hashes to the Barracuda Backup appliance. The agent uses up to 512 MB of memory during the backup process to store data chunks, allowing it to quickly walk the file system.

The Barracuda Solution vs. Other Deduplication Methodologies

Products without change detection and deduplication back up the same data repeatedly, creating long backup windows and excessive storage and bandwidth requirements. Barracuda Backup deduplication technology reduces backup time and allows customers to replicate their backups nightly. Barracuda integrates three different types of deduplication options allowing a robust yet simple-to-use solution. This section examines specific characteristics underlying Barracuda's technology.

Fixed Block vs. Variable Block

Fixed block deduplication is the simplest method of deduplication. Fixed block examines specific chunks of a given size of the dataset being backed up. Since the chunk size never changes, fixed block deduplication uses a limited amount of CPU and disk processing. However, reduction is limited since a predefined block misses duplicate data on certain data sets compared to more advanced types of deduplication.

Variable block, application-aware deduplication is an advanced method that looks at the data set/application being backed up, and increases or decreases the block size for optimal results. Since the chunk size changes based on data being backed up, additional CPU and disk resources are needed to accomplish deduplication but data reduction is maximized. Barracuda Backup uses variable block deduplication for all three stages. Barracuda's advanced variable block deduplication analyzes the data type and chunk size, setting a block size to obtain the greatest level of deduplication without taxing CPU and disk processing in the process.

Because Barracuda Backup is a hardware appliance, it can provide variable block deduplication without loading the CPU and disk resources, optimizing the underlying hardware and software in the backup appliance for this chunk method for maximum data ingest rates.

Software Deduplication vs. Hardware Deduplication

Software Deduplication

Software deduplication is usually an add-on/plug-in provided by a software vendor to help reduce the storage footprint needed to store the backup data to media. Software deduplication is often used to supplement compression before writing to a media set. With software deduplication, organizations must follow strict hardware requirements for their deployment due to the additional overhead deduplication has on an environment. While software deduplication is often cheaper or included with backup software, the software is only a fraction of the actual backup deployment, often causing organizations to overlook the broader requirements for the heavy-handed software deduplication model.

Hardware Deduplication

Hardware deduplication is a dedicated storage device that acts as a target for backup software or has its own integrated software solution. This solution has become a more common form of deduplication than software for most medium and enterprise-level organizations due to its efficiency and performance advantages. Hardware deduplication is used to offload the additional resources needed to deduplicate data. The hardware appliance works by offloading intense CPU and storage processing related to deduplication, eliminating the need for multiple devices to protect the environment.

Conclusion

Barracuda Backup deduplication simplifies data protection and reduces overhead, media, and network costs. Barracuda's three-stage, inline, variable length deduplication solution enables efficient long-term storage of protected servers while reducing backup time. With Barracuda deduplication, organizations can protect remote offices with limited resources from a central location. Organizations wishing to protect virtual environments can meet their Recovery Point Objective (RPO) while reducing their storage footprint by leveraging CBT on both VMware and Hyper-V, storing only the necessary changes across the entire data set.

About Barracuda Networks, Inc.

Protecting users, applications, and data for more than 150,000 organizations worldwide, Barracuda Networks has developed a global reputation as the go-to leader for powerful, easy-to-use, affordable IT solutions. The company's proven customer-centric business model focuses on delivering high-value, subscription-based IT solutions for security and data protection. For additional information, please visit www.barracuda.com.

Barracuda

Barracuda Networks
3175 S. Winchester Boulevard
Campbell, CA 95008
United States
1-408-342-5400
1-888-268-4772 (US & Canada)
www.barracuda.com
info@barracuda.com