Bacula Systems™

Enterprise Data Protection

Bacula Enterprise Datasheet

Key Benefits

- First free universal deduplication solution
- Reduction of disk space
- Faster storing data process
- Universal support for deduplication storage systems
- ZFS based and supported with NetApp data ONTAP 8.0.1 and higher
- Available for subscriptions at Silver, Gold and Platinum levels

Global Endpoint Deduplication

Bacula Enterprise Global Endpoint Deduplication

Save space and money using deduplication From the source, across the network, to the storage

The name of the game

Deduplication refers to any of a number of methods for reducing the network bandwidth and storage requirements of a dataset through the elimination of redundant pieces without rendering the data unusable. In deduplication process, coupled with compression, each redundant piece of data receives a unique identifier that is used to reference it within the dataset and a virtually unlimited number of references can be created for the same piece of data.
It is popular in applications that inherently produce many copies of the same data with each copy differing only slightly from the others, or even not at all.
The most popular use of deduplication in recent years has been in the area of enterprise backups.

You said deduplication

There are many storage systems and applications on the market today which implement deduplication. All can be classified into one of two types depending on how they store their data:

- Fixed Block:
  Deduplication takes places in units of a fixed size (typically 4kB -128kB). Data must be aligned on block boundaries to dedupe
- Variable Block:
  Deduplication takes place in variable-length units anywhere from a few bytes to many gigabytes in size. Block boundaries do not exist

Deduplicable filesystems use fixed block deduplication. The optimal unit of deduplication is the record size and it varies depending on the filesystems. For example, 128kB by default for ZFS, 4kB for NetApp. Others available on request.

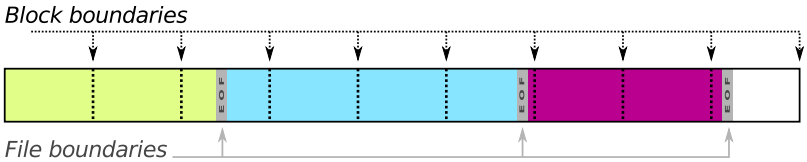
Enterprise Data Protection

Bacula Enterprise Datasheet

How do you pronounce “Backup”?

The traditional way

Traditional backup programs were designed to work with tapes. When they write to disks, they use the same format only writing to a container file instead of a tape. The Unix program tar is an example of this and so is Bacula's traditional volume format. Files are interspersed with metadata and written one after the other. File boundaries do not align with block boundaries as they do on the filesystem.

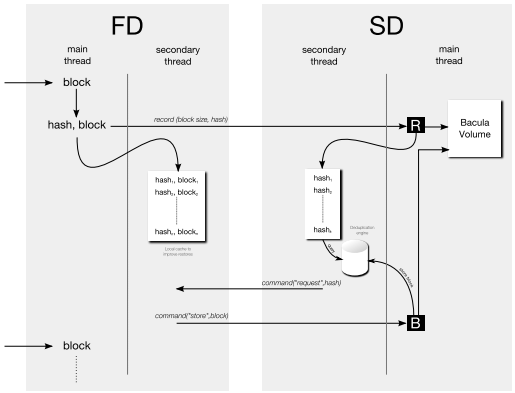


Storage without deduplication

For this reason, backup data does not typically deduplicate well on fixed-block systems. It is the main reason why variable block deduplication was invented and why there is a large market for deduplicating backup appliances.

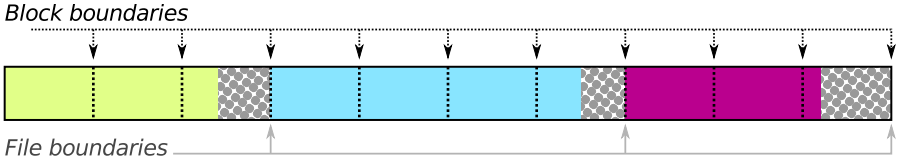
The new era: Bacula Enterprise Global Endpoint Deduplication

Global Endpoint Deduplication not only store data on disks by aligning file boundaries to the block boundary of the underlying filesystem but also optimize the bandwidth used to transfer client's data by limiting it to the strict minimum: only the missing blocks are sent over the neetwork to the storage daemon as shown on the figure below. Metadata, which does not align, is separated into a special metadata volume. Within the data volume, the space between the end of one file and the start of the next block boundary is left empty.



Client Side Deduplication overview

Since every file begins on a block boundary, redundant data within files will deduplicate well using ZFS's fixed block deduplication. This type of file is known as a sparse or holey file.



Storage with Global Endpoint Deduplication

Enterprise Data Protection

Bacula Enterprise Datasheet

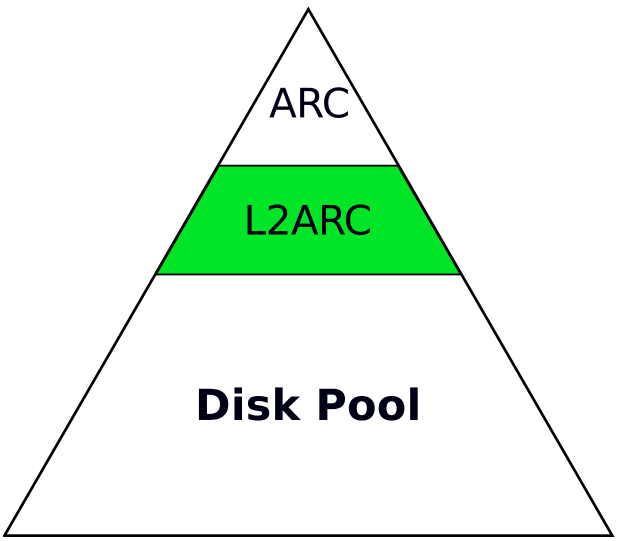
What to dedupe?

The following data types deduplicate well:

- Files that change constantly but are only appended to large logfiles
- Large files that change daily but only in small amounts
  - Monolithic Databases
  - Some types of email boxes
- Identical files that appear in backups from many clients
  - Operating system data from virtual machines
  - Email attachments with multiple recipients

Global Endpoint Deduplication are limited only by ZFS itself:

- Data will not dedupe across zpools
- Dedupe metadata is stored in the ARC/L2ARC
- Only 1/4 of the ARC/L2ARC is reserved for metadata
- Large dedupe repositories will require a large ARC/L2ARC



ZFS "ARC" cache level 2

Sizing

How big does my L2ARC need to be?

It is tempting to start with the total amount of primary data to be backed-up when calculating the size of the L2ARC but the space taken by holes in the **Bacula** volumes needs to be considered too. This must be subtracted when trying to estimate the total amount of primary data that can be backed-up and deduped using an L2ARC of a given size. One way to think about this is to picture the storage of data inside a deduplication volume in terms of full and partially full blocks. It is the number of these blocks that affects the size ot the L2ARC, not the amount of data they contain.

- Files smaller than the block size will consist of one partially full block
- Files larger than the block size will consist of one or more full blocks and usually end with one partially full block

Deduplication sizing – important parameters

- The amount of data to deduplicate
- The block size used for deduplication
- The average percent full per-block vs empty space
- The percentage of the L2ARC reserved for meta data

Examples of the impact of changing the parameters

Primary data	100 TB
ZFS record size	128 kB
Average block fill percentage	50%
Retention period	90 days

- A typical situation with default values

L2ARC metadata percentage	25%
Daily percent of data changed	2%
L2ARC size needed	560 GB
L2ARC as percentage of primary data	0.547%

- Changing daily percentage: 2% ⇒ **5%**

L2ARC metadata percentage	25%
L2ARC size needed	1 110 GB
L2ARC as percentage of primary data	1.074%

- Changing L2ARC metadata percentage: 25% ⇒ **50%**

Daily percent of data changed	5%
L2ARC size needed	550 GB
L2ARC as percentage of primary data	0.537%

Enterprise Data Protection

Bacula Enterprise Datasheet

How to size?

Bacula Systems sizing tool

Accurate sizing is difficult in practice. Oversizing and using conservative estimates is recommended.
To help in sizing your infrastructure for deduplication, **Bacula Systems** provides you an online tool available at http://www.baculasystems.com/deduplication-sizing-calculator-2

Sizing Recommendations

- Install as much RAM as possible (ARC)
- Use only SSDs for your L2ARC
- Create a much larger L2ARC than you think you need

How sizing impacts your costs?

Current storage pricing trends bode well for ZFS deduplication. Solid State Drive (SSD) performance continues to increase and prices have come down significantly in the past years making large L2ARCs economically feasible. The combination of fast *I/O processor* (IOP) performance and large capacity is essential to maintain performance as the amount of data stored in the filesystem increases.

A unique feature from Bacula Systems

- No other backup software stores data this way (patent pending)
- **Bacula Systems** helps you overcome your scaling challenges
- Raising the record size limit brings a major positive impact to your storage costs

Global Endpoint Deduplication is supported with:

- Nexenta Systems OpenStorage Appliances
- NetApp Data ONTAP 8.0.1 and higher
- Oracle / Sun ZFS Storage Appliances
- Whitebear Solutions WBSAirback Appliances
- ZFS on Linux (64-bit only)

To know more about Global Endpoint Deduplication, please contact us

Check our training sessions' dates and venues here

To get a free assesment of your current Backup & Restore infrastructure, contact us

To try Bacula Enterprise Edition with its Global Edition Deduplication feature, click here