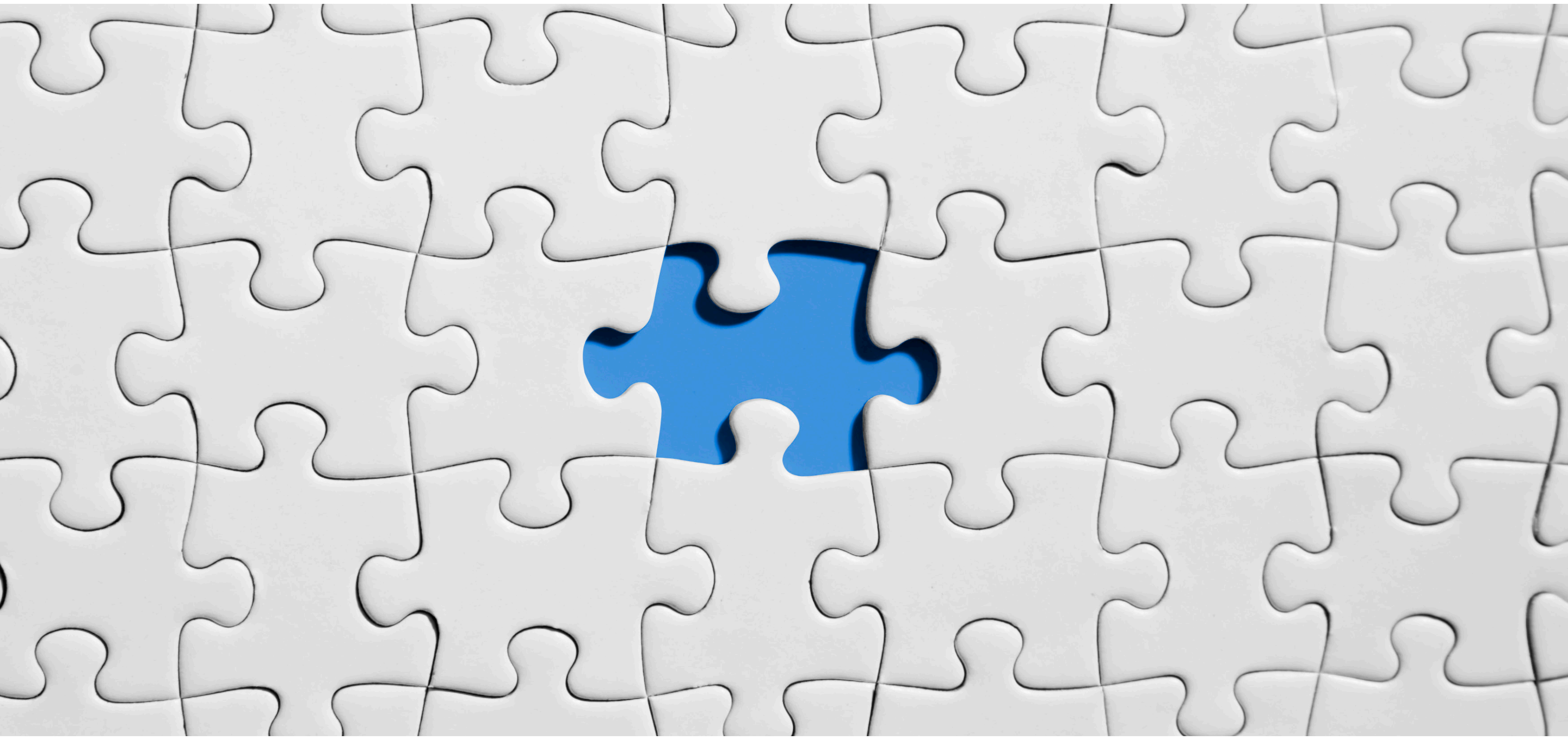
DELL Technologies

FINE-TUNING REPLICATION USING DSP



Mohammed Moueen

Anamika Kumari

Knowledge Sharing Article

DELL Technologies
Proven Professional

The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged/Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage
- Enterprise Architect

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person Classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

Table of Contents

Disclaimer: The views, processes or methodologies published in this article are those of the authors. They do not necessarily reflect Dell Technologies' views, processes or methodologies.

Preface

Data Protection has come a long way from storing data on tape to disks, from a single primary copy to multiple backup and disaster recovery copies. Still, organizations now and then face challenges while replicating their backed-up data. This article lists common challenges faced during replication of backup data for disaster recovery purposes and how Distributed Segment Processing (DSP) – a functionality of DD Boost (Dell EMC Data Domain's private protocol) – helps overcome these challenges.

DSP technology will enable deduplication on the source database or application server. This article will explain DSP architecture to understand how it works and list the advantages along with the target customers to whom this feature would be relevant. With the help of DSP technology, the true power of replication will be unlocked.

Audience

This article is for Dell Technologies sales, presales and post-sales personnel interested in acquiring high-level understanding of DSP architecture to fine-tune customer's current backup and replication process and making it more efficient.

Challenges of Replication

The below challenges are commonly faced by most organizations and have a huge negative impact on the organization's disaster recovery (DR) plan.

- **Low Throughput Performance:** Organizations often face the challenge where the system's network performance would be degraded due to low throughput rates during replication and the throughput appears to be slower than expected.

- **High CPU Utilization**: Uneven workload distribution and bottlenecks often cause the CPU to be utilized at a higher rate than usual, slowing the system's capability to efficiently replicate data.

- **Exceeding Replication Window:** Every organization wants their replication window to be smaller. However, the replication window is often compromised and is exceeded due to large amounts of data and fewer bandwidth connections.

- **Proliferation of Database:** New applications often call for new supporting databases and nearly all databases have multiple copies for development, testing, standby, etc. Existing databases are getting larger. Terabyte databases used to be rare but are now increasingly becoming common.

- **Fast DR Readiness**: Even though recovery is a critical aspect for any organization, test recoveries – which should be performed regularly – are often overlooked, as they stress available resources. The challenge is that DR recoveries often take more time to recover than recovering from production site. Organizations are under pressure to make recovery quick and simple by lowering their Recovery Time Objectives (RTO's) and Recovery Point Objectives (RPO's).

Distributed Segment Processing (DSP) Overview

Introduction to DD Boost

Backup applications are a critical component of data recovery and disaster preparedness strategies. Each strategy requires a strong, simple, and flexible foundation that enables users to respond quickly and manage operations effectively.

Dell EMC Data Domain Boost (DD Boost) is a software option supported across the entire Data Domain family and enables backup servers to communicate with storage systems without the need for Data Domain to emulate tape. DD Boost is more efficient than CIFS and NFS protocols.

DD Boost software has two components:

1. DD Boost libraries that you install on each backup server.
2. The DD Boost server that runs on Data Domain systems.

The **Backup application** manages when and how backups and duplications are performed. **Admins** manage the backups and restores from a single console. The **DD Boost application** manages all files (collections of data) in the catalog, even those created by the Data Domain system. The **Data Domain system** exposes pre-made disk volumes called storage units to a DD Boost-enabled backup server. Multiple backup servers, each with the DD Boost libraries, can use the same storage unit on a Data Domain system as a storage server. Each backup server can run a different operating system, provided that the OS is supported by Data Domain and the backup application. DD Boost for backup applications enables the application to control Data Domain replication process with full catalog awareness of both the local and remote copies of the backup.

Dedupe without DSP

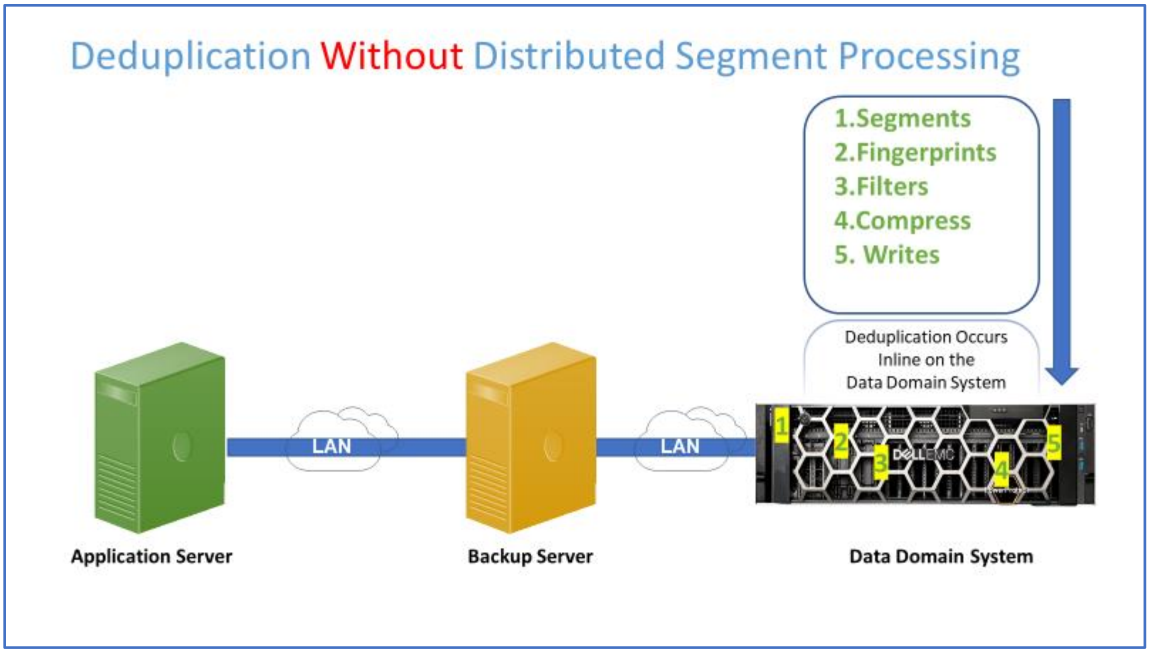


Figure 1. Deduplication without Distributed Segment Processing

Without DSP enabled, the Data Domain system performs all the functions of Segmenting, Fingerprinting, filtering, compressing and writing to disk. This can exhaust the Data Domain resources as it is performing all these processes by itself, while few of these functions can be offloaded from the Data Domain and make Data Domain resources available for other priority tasks.

Introduction to Distributed Segment Processing

Suppose you have a Data Domain system that must handle tasks from five different backup applications. Data Domain would be performing all deduplication activities by itself and worse, most of the work it does is redundant! That's because most of the files in an organization are duplicate files or files that have very negligible changes. Performing the same redundant task will over-utilize and overwhelm the Data Domain system's resources and lower the its performance. Sounds exhausting? The solution is much simpler; **Distribute the workload!**

DSP functionality of the DD Boost software distributes the deduplication process between client and server to avoid sending duplicate data to the Data Domain system.

DSP provides the following benefits:

- DSP potentially lowers network traffic generation because the DD Boost Library sends only unique data to a Data Domain system. In general, greater the redundancy in the data set, greater the saved network bandwidth to the Data Domain system.
- With DSP, DD Boost Library uses 24 MB of memory for every file backed up and DD Boost 5.7 with high availability (HA) uses 128 MB of memory for every file backed up.

Dedupe with DSP

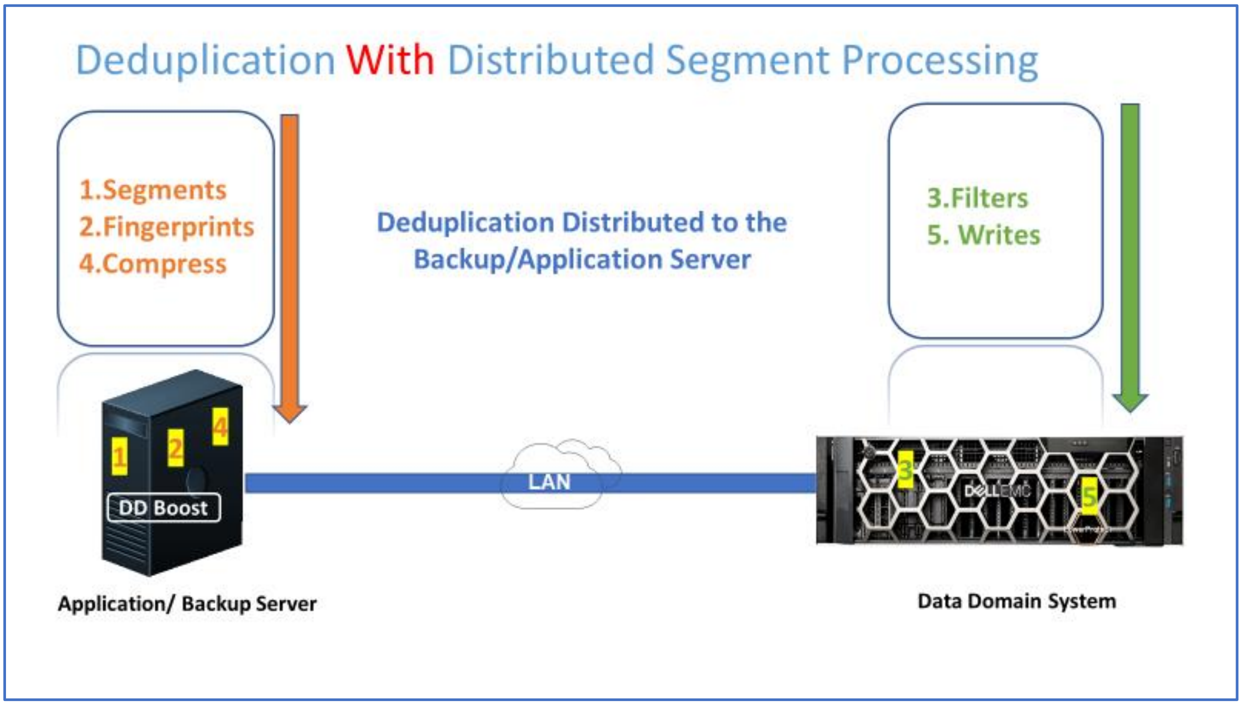

Deduplication With Distributed Segment Processing
1.Segments
2.Fingerprints
4.Compress
Deduplication Distributed to the Backup/Application Server
3.Filters
5. Writes
1
2
4
DD Boost
LAN
3
5
Application/ Backup Server
Data Domain System

Figure 2. Deduplication with Distributed Segment Processing

DSP shares deduplication duties with the backup host.

With DSP enabled, the backup host performs these functions:

- ✓ Segments the data
- ✓ Creates fingerprints of segment data and sends them to the Data Domain system
- ✓ Optionally compresses data
- ✓ Sends only the requested unique data segments to the Data Domain system

With DSP enabled, the Data Domain system performs these functions:

- ✓ Filters the fingerprints and requests data not previously stored
- ✓ Records references (pointers) to previously stored data and writes new data

This can save you both time and money by using your current infrastructure and unlocking its true efficiency!

DSP is enabled by default on systems initially installed with DD OS release 5.2 or higher. On system upgrades from DD OS release 5.0.x/5.1.x up to DD OS release 5.2, DSP remains in its previous state.

DD Boost can operate with DSP either enabled or disabled. DSP must be enabled or disabled on a per-system basis. Individual backup clients cannot be configured differently than the Data Domain system. DSP cannot be disabled on an extended retention Data Domain series system.

Architecture and Process

Let's detail how DSP works between our Backup clients or application server and the Data Domain.

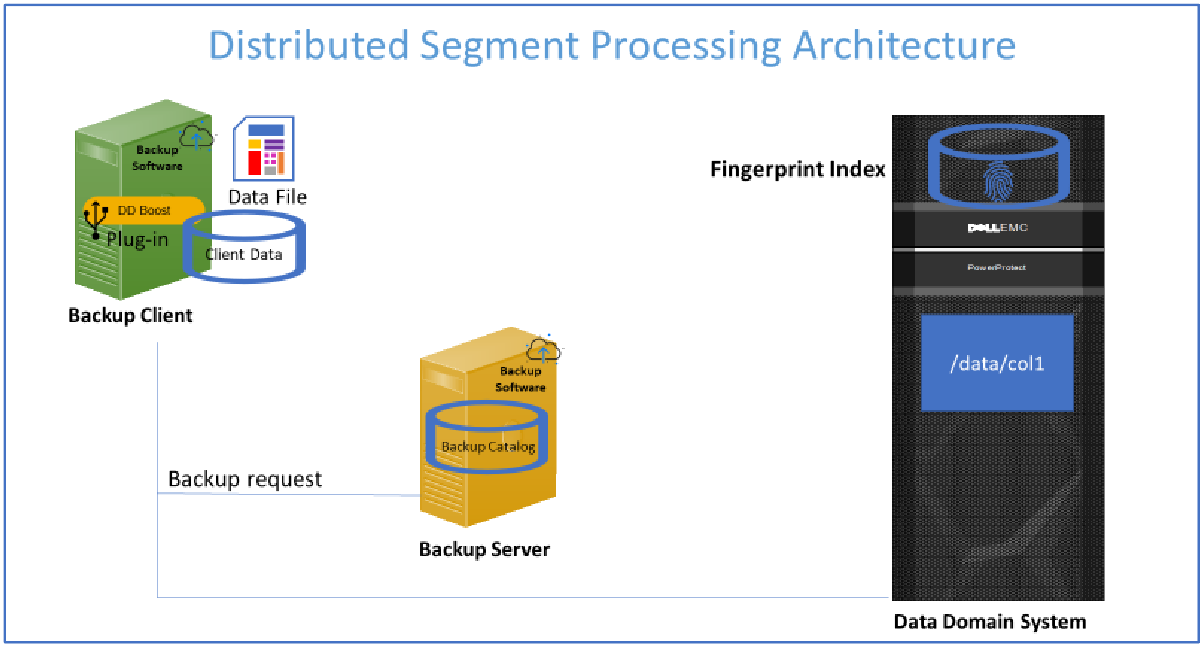


Figure 3. DSP Architecture Overview

Figure 3 depicts a production environment that has a **backup client** that has client data and a **DD Boost plug-in** installed in the client's backup software. It also includes a **backup server** that will have the configuration files, scheduling and meta-data details in backup catalog and a **Data Domain** that will provide the storage in /data/col1 filesystem. As part of its filesystem it also records the metadata, known as **fingerprints**, of all data that is being backed up represented as fingerprint logo. It also has a master catalog of all fingerprints called **Fingerprint Index** that's stored on the Data Domain. The below process goes through how DSP works along with Data Domain Stream Informed Segment Layout (SISL) architecture.

Process:

1. Backup software makes a request to the backup client to perform a backup and backup software will pass the information to the DD Boost plug-in, resulting in a data file that we want to backup.

2. Before the data file is sent across the network to be stored on the Data Domain, we would want to analyze it and determine what parts of the file are already on the Data Domain.

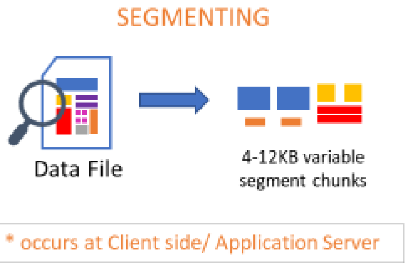

SEGMENTING
Data File
4-12KB variable segment chunks
* occurs at Client side/ Application Server

Figure 4. Segmenting

3. The data file is broken into parts from the DD Boost code on client. Data Domain refers to these as **segments**, data chunks that vary in size from 4k to 12k. Each file that is broken into parts is always broken up the same way every time when it is processed by the Data Domain. This way we can match it against all the segments and fingerprints backed up on the Data Domain.

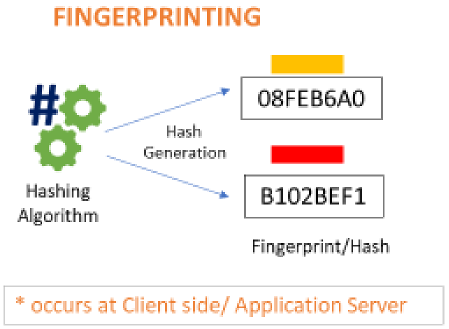


Figure 5. Fingerprinting

4. The client does not know which of these segments are already on the Data Domain or not, so it will then generate a fingerprint using **SHA-1 Hashing Algorithm** and send it across the network to the Data Domain.

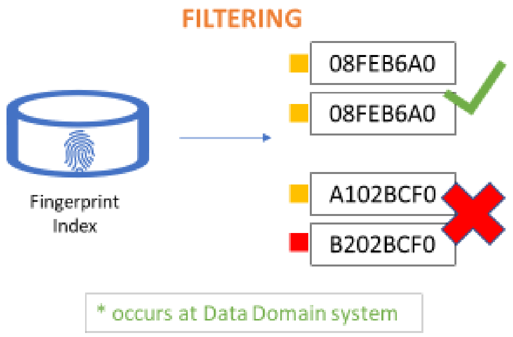


Figure 6. Filtering

5. Data Domain will check against its stored fingerprints in the fingerprint index to see if it has stored this piece of data or not.

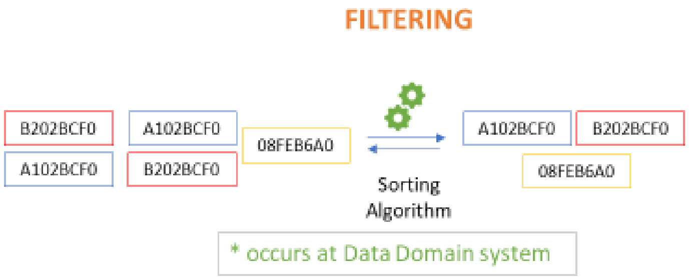


Figure 7. Filter using sorting algorithm

6. The segment then uses a **Summary Array Vector** to do the sorting in memory. The summary array vector is based on a math principle called bloom filter.
7. Once it determines that this segment has been stored and seen before it will respond to the client that the fingerprint is already on the Data Domain. Therefore, it will make a record of the fingerprint (Pointers) that was part of the backup but will not have to actually send the whole segment across the network.
8. Next, it processes another chunk of the segment, generates the hash and sends it to the Data Domain. The Data Domain checks if it has this piece of segment or not; in this case there is no match. When we don't have the hash in the Data Domain, it lets the client know that it needs this important segment.

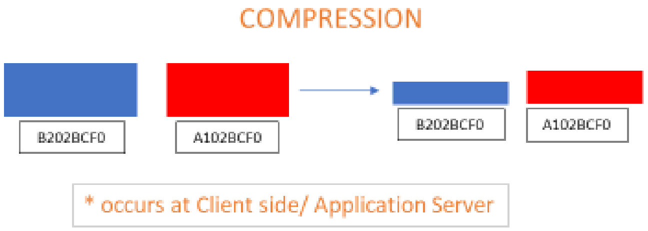


Figure 8. Compression

9. The segment is compressed, and client will send the compressed segment across the network to the Data Domain and the hash will get updated in the Data Domain record as well. The second compression it performs is called the **Local Compression**. The compression or sorting against all the fingerprints on the Data Domain is usually referred to as **Global Compression**.

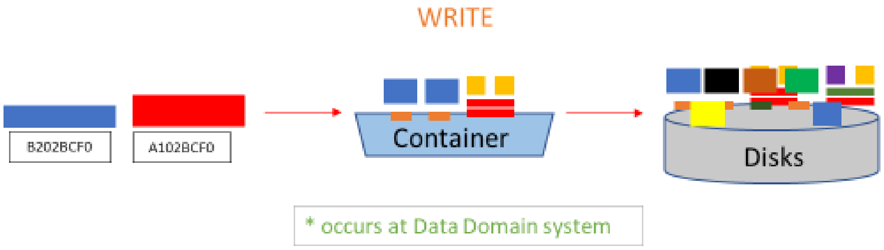


Figure 9. Write to Disk

10. Overall, we get an effect of deduplicating the data before we send it across the network because we are only sending those segments that aren't already on the Data Domain.

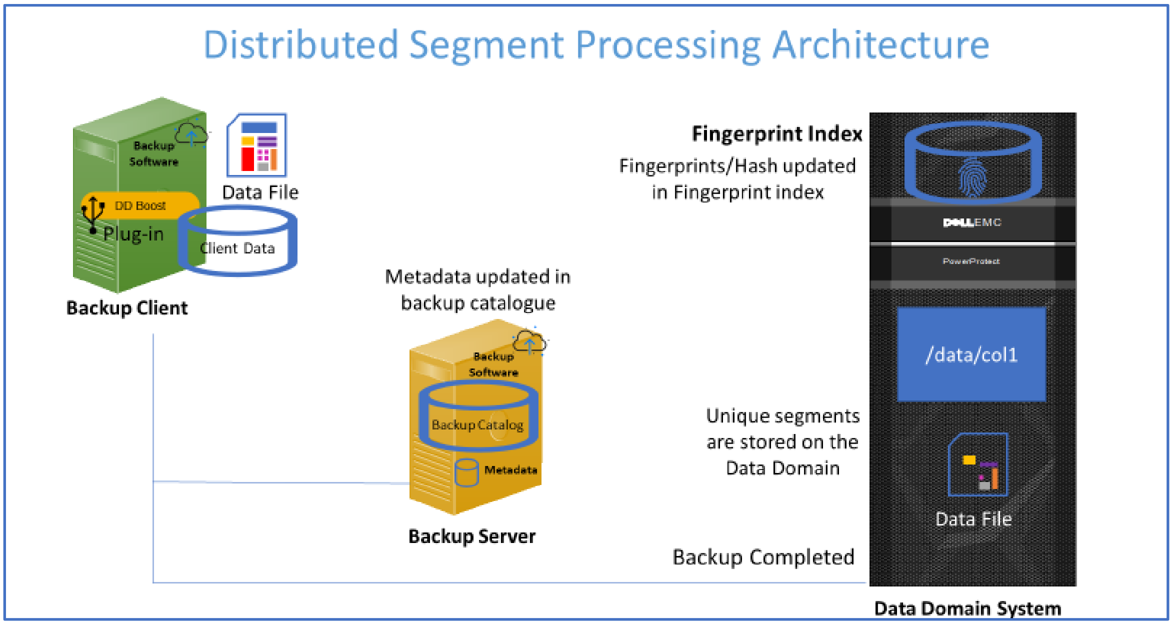


Figure 10. Backup completed using DSP

11. During this process the backup server will record that this file has been backed up as part of the scheduled backup.

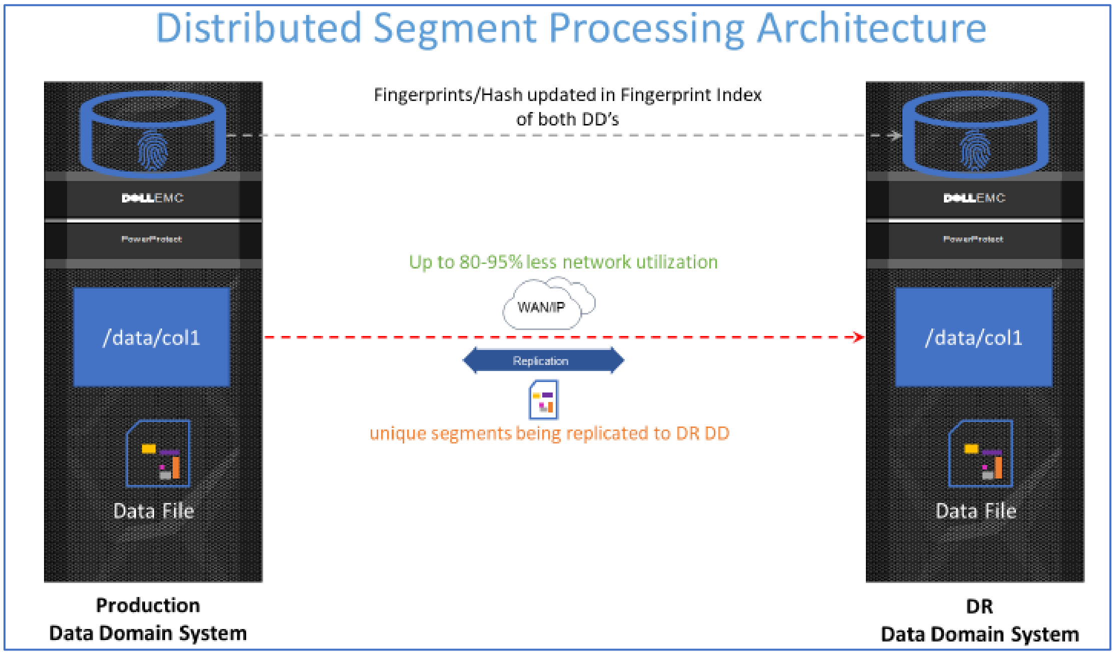

Distributed Segment Processing Architecture
Fingerprints/Hash updated in Fingerprint Index of both DD's
DELL EMC
PowerProtect
Up to 80-95% less network utilization
WAN/IP
Replication
unique segments being replicated to DR DD
/data/col1
Data File
Production Data Domain System
DR Data Domain System

Figure 11. Replication using DSP

12. Once the backup is completed, the Production Data Domain system will replicate the data to the Data Domain system in the DR site for disaster recovery. This is where we notice the

benefit of using DSP for replication. Since the majority of the task is completed in the production site, the replication will occur fast as only the unique segments are being written to the DR Data Domain. This reduces the storage space required on the DR Data Domain and minimizes WAN bandwidth use, thus lowering the RPO and RTO.

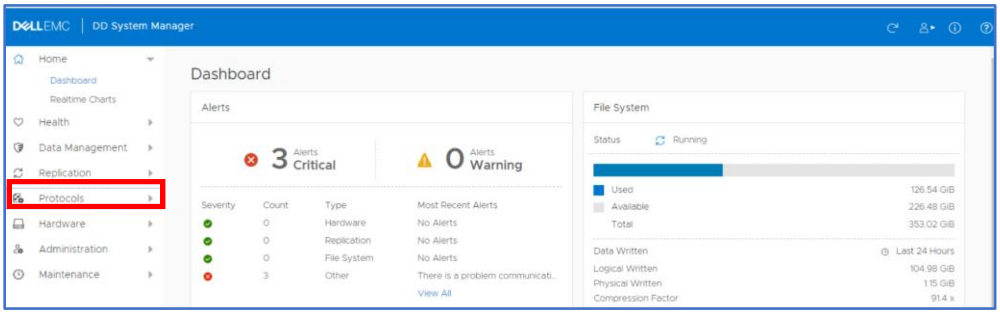
Configuring Distributed Segment Processing in Data Domain

The DSP option is configured on the Data Domain system and applies to all the backup servers and the DD Boost libraries installed on them. You can choose to enable or disable DSP when you send backup data to a Data Domain system using DD Boost software.

Configuring DSP using GUI

Below are the steps to configure DSP on your Data Domain system:

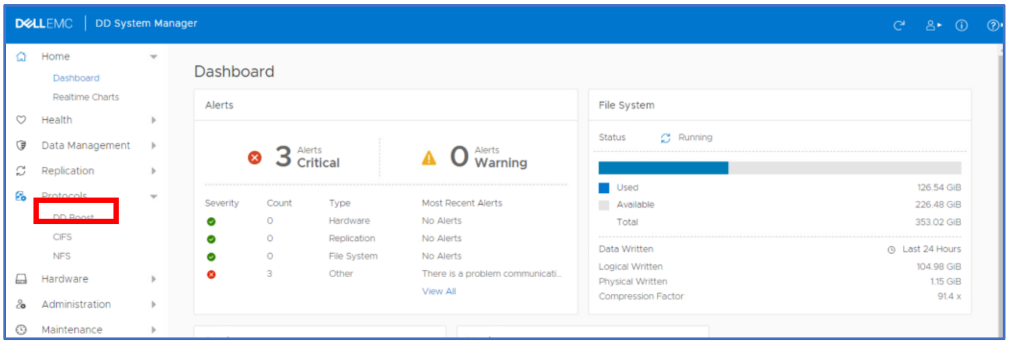
1. In DD System Manager, select Protocols.



Severity	Count	Type	Most Recent Alerts
●	0	Hardware	No Alerts
●	0	Replication	No Alerts
●	0	File System	No Alerts
●	3	Other	There is a problem communicati...

Figure 12. Selecting Protocols

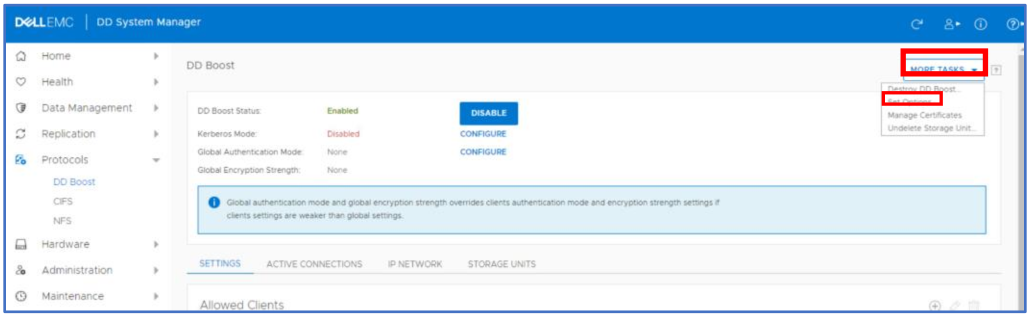
2. Expand the Protocols Tab > select DD Boost.



Severity	Count	Type	Most Recent Alerts
●	0	Hardware	No Alerts
●	0	Replication	No Alerts
●	0	File System	No Alerts
●	3	Other	There is a problem communicati...

Figure 13. Selecting DD Boost

3. Select the Option DD Boost > More Tasks in the right-hand corner and click on Set Options.

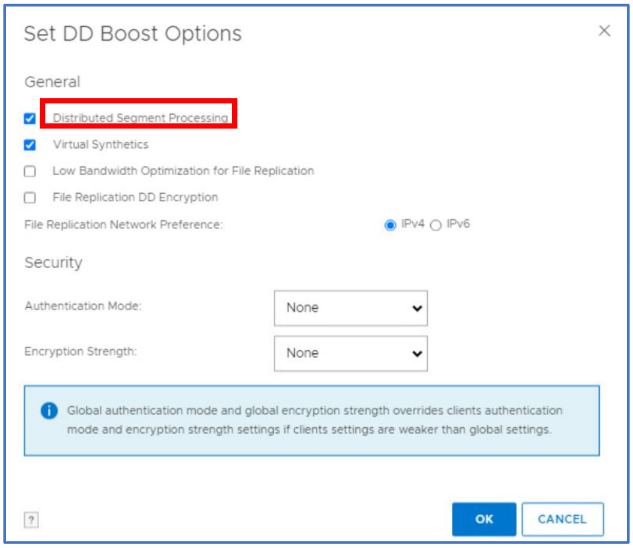


DD Boost Status: Enabled
Kerberos Mode: Disabled
Global Authentication Mode: None
Global Encryption Strength: None

Global authentication mode and global encryption strength overrides clients authentication mode and encryption strength settings if clients settings are weaker than global settings.

Figure 14. Selecting More Tasks> Set Options

4. In Set Options > select Distributed Segment Processing and Click OK.



☑ Distributed Segment Processing
☑ Virtual Synthetics
☐ Low Bandwidth Optimization for File Replication
☐ File Replication DD Encryption

File Replication Network Preference: ● IPv4 ○ IPv6

Authentication Mode: None

Encryption Strength: None

Global authentication mode and global encryption strength overrides clients authentication mode and encryption strength settings if clients settings are weaker than global settings.

Figure 15. Enabling Distributed Segment Processing

Configuring DSP using CLI

To enable or disable distributed segment processing:

```
# ddboost option set distributed-segment-processing {enabled |
disabled}
```

This option is enabled by default, but verify that it is enabled before using Data Domain Boost backups:

```
# ddboost option show
```

Enabling or disabling the distributed segment processing option does not require a restart of the Data Domain file system. Distributed segment processing is supported with version 2.2 or later of the DD Boost libraries communicating with a Data Domain system that is running DD OS 4.8 or later.

The True Value of Distributed Segment Processing

Now that we have seen the functionality of DSP and how it works, let us go through how the above challenges mentioned at the beginning of the article are being resolved using DSP.

- ✓ DSP has a very **easy implementation** as it is included with the DD Boost plug-in software.

- ✓ By distributing the deduplication process, **higher ingest rates** are achieved. (Though it should be noted that a Data Domain system is no slouch when it comes to ingest speed. However, next to recoverability, speed is always important in a backup and recovery environment.)

- ✓ It **reduces CPU usage** on the Data Domain system. As a result, CPU in the Data Domain system can be used for other tasks, such as replication and cleaning. This will incur 20-40% lower CPU overhead on the media server as well.

- ✓ **Reduce amount of data** sent across the network since we are sending the fingerprints ahead of the actual data.

- ✓ Since less data would be sent across the network, you can use the existing 1GbE networks for replication traffic **and avoid spending** on new 10GbE network infrastructure.

- ✓ **Shrinks the replication window** considerably because it's not necessary to send all segments, only unique data not present on the Data Domain. The Data Domain sorts out what data is already a duplicate.

- ✓ **Save disk space** on the Data Domain. As In-line, variable block dedupe occurs, we are storing on a smaller footprint on the disks of the Data Domain. This is very effective especially for large databases as they continue to grow larger every day.

- ✓ Failed backup/replication tasks perform much **faster on retries**.

- ✓ DSP provides higher aggregate in throughput which makes recoveries much faster and lowers RTO's and RPO's, therefore, providing **faster DR readiness.**

Example: The efficiencies of EMC BOOST Distributed Segment Processing

The real-life customer example below exhibits how DD Boost and DSP can reduce bandwidth use. Figure 16 shows the load on a 10G Ethernet link to a Data Domain before and after enablement of Distributed Segment Processing. DSP was enabled on 29-05-2013 during the day. The difference was so significant that the customer's network department warned them that backups had stopped working!

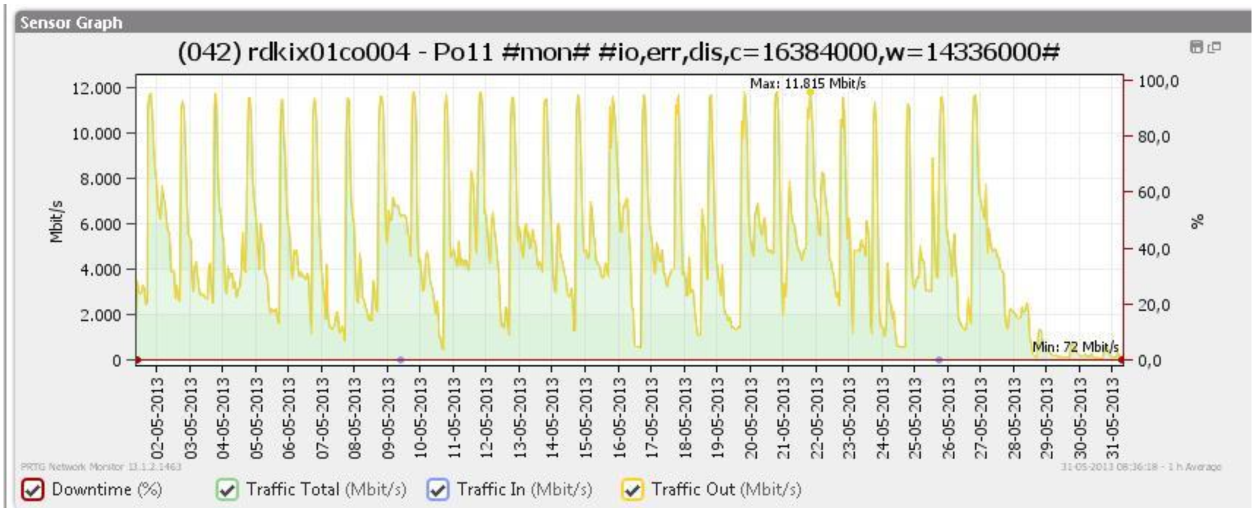


Figure 16. Efficiency of distributed segment processing

The graph shows Mbit/sec but it is in GB/sec. Notice how much the bandwidth was reduced on the very first day of enabling DSP. This gives us confidence in the power of distributed segment processing.

Conclusion

Workload distribution has always brought fruitful results. Backup and replication data tend to have a lot of duplicate data because we capture data from different clients, and we don't know what type of data we might lose or must restore. There is so much duplicate data captured every day that does not have to be sent across the network or stored on disks NOW! Backups and replication go hand in hand; an efficient backup process affects replication too, as most of the burden is reduced during the deduplication process on the client side making replication efficient.

References

https://www.delltechnologies.com/en-us/collaterals/unauth/technical-guides-support-information/products/networking-4/docu85192.pdf

https://infohub.delltechnologies.com/l/deployment-guide-dell-emc-ready-bundle-for-sap-with-unity-storage-1/configuring-data-domain-data-protection#:~:text=DD%20Boost%20distributes%20parts%20of,is%20transferred%20over%20the%20network.

https://www.mass.dk/netbackup-quick-hints/the-efficiencies-of-emc-boost-distributed-segment-processing/

https://slideplayer.com/slide/5814234/

https://www.fondazionecrui.it/wp-content/uploads/2017/09/h11755-business-value-dd-boost-wp.pdf

https://dell.sabacloud.com/Saba/Web_spf/PRODTNT091/app/shared;spf-url=common%2Flearningeventdetail%2Fcurra000000000009651

Dell Technologies believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED "AS IS." DELL TECHNOLOGIES MAKES NO RESPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying and distribution of any Dell Technologies software described in this publication requires an applicable software license.