



# EMC DATA DOMAIN EFFICIENCY EVALUATION



Erico Correia da Silva  
Advisory Solutions Architect  
EMC BRS – Brazil  
[ericosilva@emc.com](mailto:ericosilva@emc.com)

## Table of Contents

Introduction .....	2
Data Domain .....	3
The AutoSupport .....	3
Autosupport verification .....	4
Configuration Guidelines .....	7
CIFS and NFS .....	7
VTL.....	8
NDMP .....	9
Replication.....	9
DDBoost .....	10
Advanced Load Balancing and Link Failover - ifgroup .....	10
Backup software .....	11
Compression - The Deduplication Ratio .....	12
Choosing the best local compression algorithm .....	13
Cumulative compression ratio.....	14
Compression verification.....	14
Maintenance activities: the clean operation .....	16
Data Domain performance analysis.....	17
External performance factors .....	18
Internal performance factors .....	20
Performance verification .....	22
Conclusion .....	25
Bibliography .....	26

Disclaimer: The views, processes, or methodologies published in this article are those of the author. They do not necessarily reflect EMC Corporation's views, processes, or methodologies.

## Introduction

Efficiency is “the ratio of the effective or useful output to the total input in any system.” Thus, no matter how difficult it is to get the outcome, the amount of work, energy, and time spent by a system will dictate its efficiency while processing the inputs to get the desired results. Bringing this idea to the IT world, efficient backup systems must gather the data, catalog them, and store everything in a very short period of time, respecting the service level agreement (SLA) and lowering the total cost of ownership (TCO). In a so-called “backup system”, Data Domain<sup>®</sup> acts as the efficient storage, receiving the data from the backup software which gathers and catalogs everything.

In the realm of deduplication backup storage, Data Domain is the best-of-breed in the market and has been proved as such, over and over, since its introduction. Of course, its efficiency depends on several factors, such as how it is configured, administration activities, environmental issues, and so forth.

But how can a backup administrator improve the efficiency of Data Domain system? To answer this question, this Knowledge Sharing article will cover several points such as:

- Basic items to check in an autosupport report;
- Configuration guidelines for Ethernet-based backups (DD Boost, RMAN, CIFS, NFS) and SAN-based backups (VTL);
- Understanding and evaluating the deduplication ratio of Data Domain;
- Maintenance activities: the clean operation;
- Data Domain performance analysis.

Though every backup environment and business need is different, Data Domain has the required flexibility to achieve the most aggressive Recovery Point Objectives (RPO) and Recovery Time Objective (RTO). The first step to meet all expectations is to understand the technology, so that we can take advantage of its flexibility, then plan, configure, administer, and check the system continuously. This article will explain the various aspects involved while evaluating the health of a Data Domain system, improving the understanding of how to keep track of the technology's efficiency.

## Data Domain

A backup storage device must cover the following:

**Speed:** Due to the large amount of data managed by IT systems, sending all information to a storage device in a daily basis is a challenge. Nowadays the IT systems work 24/7, making such challenge even worse.

**Reliability:** A backup is generally the last resort to recover data after a failure or to restore information to a previous state. If the backup system (software, catalog, storage, etc.) presents a failure it definitely means data-loss, causing a potential severe impact to the business.

**Cost-effectiveness:** Storing backup data in a daily basis may require a large amount of storage space, since no matter how much the data has been changed in a day, the state of the data in that day must be recoverable, meaning that the backup system must be smart enough to keep track of the changes in a way that stored data can be combined to recreate the state in a giving point in time.

As a complete storage system ideally for backups, a Data Domain (also referenced as “DD”) has the required Speed to store the data, all data can be deduplicated in the source, reducing the amount of time needed to send the data from the source system to the Data Domain. This is a key point because the technology available so far has the media for transferring data as the lowest performance in the data path. Data Domain is complete also when talking about reliability, since its parts are redundant, the file system has a self healing capability, it has several monitoring features, and so on. Considering the cost-effectiveness Data Domain is the leader technology due to its deduplication capabilities that can reach considerably high rates.

Backup administrators have to understand the technology to get the best from a Data Domain system, raising the backup speed, assuring the reliability of the system, optimizing the deduplication feature.

## AutoSupport

By default, every day at 06:00am a Data Domain system sends to EMC a detailed report, called “autosupport” (or ASUP). The autosupport report is comprised of several command outputs and log entries, showing the system identification, configuration and state. All autosupport reports received by EMC support are saved and compiled to create additional usage information, very helpful during support activities. The list of

email recipients to get the autosupport report can be changed to add any other email address. Backup administrators can create their own tools to treat the autosupport reports and generate additional information, alerts, etc. that can help in future decisions.

### **Autosupport verification**

The autosupport report is a very strong tool to evaluate the Data Domain efficiency. The report changes from a DDOS<sup>1</sup> version to another but some basic points can be easily verified in a regular basis, avoiding issues in a proactive manner. Below is a list of areas to check in an autosupport report followed by a short explanation on how to evaluate the info:

**General Info:** Will give some identification points, version, and uptimes. Check mainly the DDOS version, uptime and filesystem uptime:

- DDOS (the DD operating system) should be the latest available or the previous release.
- Uptime will show if the system is up since the last planned reboot.
- Filesystem uptime should be the same amount of time as the system uptime. If filesystem uptime is smaller, then administrator must understand why. If filesystem has been restarted abnormally, EMC support must be contacted.

**Server Usage:** Shows the Data Domain virtual stored capacity and disk space usage. In this area the compression rates are also shown. Article (1) is a good reference to understand the outputs displayed in server usage area.

**General Status:** This autosupport area shows command outputs used to understand the status of several Data Domain system parts (hardware and software). Check this area to see memory usage, current and previous alerts, disk status, NFS and CIFS status.

- The “net show hardware” command output must be checked to identify link status, speed and duplex for Ethernet ports.
- Replication commands should be verified to understand replication configuration and day-by-day progression.
- “Enclosure show summary” and “enclosure show all” should be checked to identify enclosure connection status, temperature and other hardware items such as fans and power supplies.

---

<sup>1</sup> DDOS is the Data Domain operational system.

- “Enclosure show topology” is helpful to verify how shelves are connected. Every SAS bus should have close number of shelves (e.g. 3 busses, 3, 3 and 2 shelves). Shelves of different types (ES20, ES30) cannot be mixed in the same SAS path.
- NVRAM status has to be checked. In case of errors or if batteries are not ok, EMC support must be contacted.

**Software configuration:** This section must be checked to understand the licensed capacity and features.

**Hardware configuration:** This autosupport area shows different details about the Data Domain hardware such as head components (CPU, memory, BIOS), network configuration, disk configuration, ports and IPMI configuration.

- The “net show config” will show the ethernet interfaces configuration (“ifconfig -a”), for real ports and also for virtual interfaces if any bond has been configured.
- The “net show settings” shows interfaces and bond configuration
- “net failover” and “net aggregate” will give detail on bond and aggregations present in the system
- Disk commands are helpful to identify disk failures and SAS ports status/speed.
- Check “system show ports” to identify negotiated port speeds, especially for VTL ports.

**Detailed statistics:** This section contains all statistics regarding the stored data. The characteristics of the stored data will dictate the compression rates of a Data Domain system, so this autosupport area will help the administrator to clearly see what has been stored.

- Filesystem cleaning configuration will show how often the clean process runs. This will be discussed later in this article.
- File distribution will show the age of the stored data in percentage and also the size distribution of the stored files.

**Registry** is the section where specific low level configuration settings are shown.

**Detailed network layer** will show network statistics helpful in low level TCP/IP troubleshooting.

**Detailed storage layer** section of the autosupport has several disk related information, some of them in a per disk basis, useful to identify potentially bad disks.

**Additional information** area has information such as snapshot list, weekly and daily disk space consumption, etc.

**DDBoost information:** is the autosupport section where detailed information about DDBoost related items are found. In this area the administrator should check mainly:

- Distributed-segment-processing: This option must be enabled in order to delegate part of the deduplication work to the DDBoost client, avoiding the transport of unneeded data from client to the DD system.
- Ifgroup show config: check this output to make sure that the right interfaces will be used for DDBoost backups.
- DDBoost status: make sure it is enabled.
- DDBoost connections: this autosupport part is very important to understand current DDBoost connections and the plugin version being used. Make sure that the plugin versions in place are ok per EMC Compatibility guide (2).
- DDBoost history output is useful to see how much data are being treated and how much data had to be transferred over the network.
- DDBoost file replication statistics will be useful for replication monitoring.

**Crash information** section will show filesystem and cleaning crashes. These meters have to be checked frequently. EMC support must be contacted in case they get increased or if core files for other modules are found (in the list of files) in this section.

**Messages** section can be checked for error messages clearly severe. Several messages may not indicate any severe issue, so EMC knowledge base and/or the support team should be consulted.

**System show performance:** This is one of the most important sections of the autosupport report (3). We will discuss it in detail later in this article.

The autosupport report has several other sections and, as said before, varies from one DDOS version to another. The sections not mentioned here are not less important, but discussing them deeper is out of the intent of this article.

## Configuration Guidelines

When talking about any storage system, the configuration is central to guarantee the required efficiency. For every piece of config, backup administrators have several options and their choices will directly influence the efficiency not only of the Data Domain but of the entire backup system. The next topics will depict some attention items and point to additional best practice documentation that has to be read.

### CIFS and NFS

CIFS and NFS are the simplest DD access methods available in every system and do not require any specific license. Whenever VTL or DDBoost is available, CIFS and NFS may still be used for storing secondary backup that should be available even in situations when the backup software is not available (e.g. for storing simple reports needed for a DR). Regular DDOS administration guide has the critical rules and instructions for CIFS and NFS (4).

**CIFS:** Some basic items to check are as follows:

- Set the specific interfaces that should be available for CIFS (command: “cifs option set interfaces”).
- Make sure ntp is correctly set and DD time is in sync with the Active Directory servers.
- Do not use a share of the /data/col1/backup directory. It is safer to create another share of a directory below /data/col1/backup.

Document (5) should be read for a complete discussion on CIFS configuration guidelines.

**NFS:** Some basic items to follow are:

- Tuning is different for each backup software and OS.
- Check EMC support web site for documentation specific for your platform and backup software.
- EMC recommends the use of hard mounts to ensure availability of NFS mounts after reboots.
- In case of problems, even after following (4) and the required Data Domain integration documentation, the backup administrator may follow (6) before calling EMC support for help.

## VTL

Whenever the VTL protocol is used to connect the DD system to the backup infrastructure, EMC engineering must be engaged to check the architectural scheme in order to evaluate if all components (SAN switches, HBA, OS, drivers, etc.) are fully supported and can be safely combined. The wide variety of components may drive the configuration to combinations where issues have been detected in other sites, so the qualifier process is a great tool to guarantee the maximum efficiency of a DD VTL environment.

No matter how complex or simple the VTL environment can be, the most important rule to follow for a Data Domain is dedicating one HBA channel of the hosts, specific for the Data Domain connection. The intent of this rule is to avoid several issues that can rise when you have other SAN attached systems using the same channel as the Data Domain VTL. Different Data Domain systems can share the same HBA port, but the port cannot be shared with disks, physical tapes, tape libraries, other VTL systems or any SAN attached device.

Besides the VTL Configuration Guidelines document (7), backup administrators must also check and follow the VTL Best Practices Guide (8) where several rules are described. Some of the basic rules are:

- An **initiator** must be dedicated to Data Domain (never share HBA channel with other SAN attached systems);
- Use only 1-to-1 **zoning**, use WWPN for zoning, never create zones between hosts and VTL devices that will not be used by such host;
- Create libraries with more **slots** than what you believe will be the maximum needed, as changing the number of slots is one of the activities that will require a complete reconfiguration of the VTL;
- **Changer**: refer to the backup software integration guide on which library to emulate. Netbackup requires the RESTORER-L180 library;
- **Drives**: never mix drive types, check (8) for settings such as auto-offline and auto-eject. HPUX requires specific caution, check (8) and (9).
- **Target HBAs (FC ports)**: Drives should be spread across multiple DD ports. Since backup software will dictate drive allocation, administrators should identify if it is better to assign drives to ports in sets (1,2,3,4 to port

a, 5,6,7,8 to port b) or in a round-robin fashion (drive 1 to port a, drive 2 to port b, drive 3 to port a, and so on). Secondary port should be set to none unless explicitly required for the backup software in use. Always use persistent binding.

- **Virtual tapes** should be small, averaging in size from 100 to 200GB. The amount of tapes should be initially small, enough for the backup load. New tapes should be created during operation, only if needed. Since disk space is reused when rewriting the label of the virtual tapes, make sure your backup software is configured to relabel and reuse tapes automatically. Organize virtual tapes in different Data Domain pools, using unique barcodes, try using same suffix across all pools.
- **Access Groups:** Devices on each VTL should start at LUN0. Use one initiator per access group. Avoid using configurations that share devices between hosts. Before creating an access group, administrators should read documents (8) and (10).

Before proposing a new VTL configuration, backup administrators should check EMC web site for recent versions of the documents cited here. The documents must be followed specially for bigger and complex SAN environments.

## **NDMP**

NDMP is usually configured attaching the Data Domain virtual device through a SAN connection. If so, no NDMP specific requirements are to be used, but all VTL best practices and guidelines have to be followed. Another option for NDMP is using the NDMP TapeServer Group, which is a Data Domain feature that enables the use of the ethernet infrastructure to transfer NDMP data between the file server and the DD. Refer to (4) for instructions.

## **Replication**

Replication between different sites, using WAN connection is one of the most attractive Data Domain features. In order to have the best efficiency of DD replication feature, it is important that systems should be prepared for replication since day 1. Some basic guidelines are:

- Install and configure source and target DD, before putting source system on production. This will avoid long replication backlogs to process.

- If install of target DD is not at the beginning of the production work, make sure that it will be possible to install target DD on the same site as the source and keep both of them together until they get synchronized. When both systems get in sync, target DD can then be moved to remote location.
- Check replication lag before adding replication contexts, the more replication contexts used, the more are the streams used for replication. Since each Data Domain system has a limit of concurrent streams, using more replication contexts than needed may impact the overall performance.
- Low bandwidth optimization should be used only for very slow network connections, because this will require more resources from both source and target systems.
- Make sure the desired ethernet port is being used for replication and double check routing configuration on both source and target Data Domains.
- Before choosing between MTree and directory/pool replication, backup administrators should read (4) and (11) to understand the differences.

### **DDBoost**

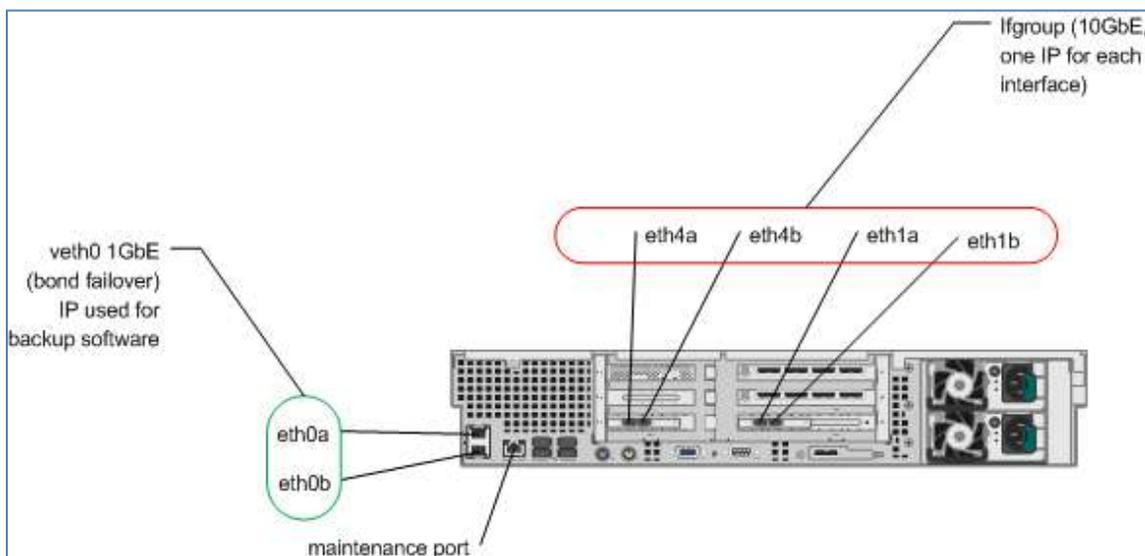
DDBoost is the best access mode when the media to be used is Ethernet. Since part of the deduplication process can be delegated to the source host, one of the first concerns is how big the overhead will be. This most of the time cannot be accurately predicted but tests and field experience show that DDBoost overhead on the host side is much smaller than the overhead needed for required non deduplicated I/O. So machines sized for regular backups can be safely used for DDBoost backups (12).

### **Advanced Load Balancing and Link Failover - ifgroup**

Ifgroups is the DDOS/DDBoost feature that makes possible to use more than one interface for backup and restores, leveraging load-balance and failover of the available NICs. This kind of results can be accomplished by using other protocols such as LACP, but ifgroup is the recommended option. Some other points to follow are:

- Choose ifconfig to aggregate multiple network interfaces whenever possible. With DDOS 5.2 Boost configuration can range from very simple configurations to more complex multiple-ifconfig configurations.

- If NIC aggregation and/or failover is needed on the host side, the use of some other NIC aggregation approach is a must. Choose the link aggregation approach (e.g. LACP, etherchannel, etc.) based on the host OS and available network infrastructure. It is important to keep in mind that DDOS ifgroup feature acts only on the DD side, so other parts of the data path must be sized accordingly.
- Do not use more than one ifgroup unless explicitly required by the network infrastructure (e.g. DD interfaces being part of multiple VLANs).
- Do not mix interfaces of different speeds in the same interface groups.
- All ifgroup interfaces should be part of the same subnet.
- For configurations comprised of both 1GbE and 10GbE interfaces, choose to combine two 1GbE in a virtual interface (failover bond), present such virtual interface to the backup software, but put only the 10GbE interfaces in the ifgroup. With this approach transactions are negotiated using redundant cheaper connections and all 10GbE are fully used for backup/restore I/O (Figure 1).



**Figure 1: DDBoost 1GbE and 10GbE connections example (DD890, 4x10GbE ports)**

Backup administrators must fully understand the DDBoost administration guide (13) before proposing DDBoost configuration changes.

## Backup software

All supported backup software products have their own set of recommendations depending on the used OS and protocol in place to send data to the Data Domain. For

example using NFS to send data to the DD requires a different set of NetWorker configurations than what would be needed if VTL were the protocol. Same way, if we are talking about NetWorker over Solaris, the settings may be different to the ones used for Linux, and so on. Backup administrators must consult the available integration documentation found in EMC Data Domain website. For NetWorker, the “NetWorker Data Domain Integration Guide” (14) must be checked before any configuration. A deeper discussion, about the integration of each backup software, is out of the scope of this article.

## Compression - The Deduplication Ratio

Data Domain systems deduplicate the data inline, during the write process, so data are written to disk already deduplicated (15). Because of its SISL algorithm (Stream-Informed Segment Layout) Data Domain systems storing process is CPU centric, meaning that write speed will rise as the CPUs get more powerful, and not necessarily depending on the amount of disk spindles available.

Data Domain compression is divided as follows:

Local Compression: The process of reducing the data footprint by using compression algorithms such as gzip or LZ.

Global Compression: The actual deduplication, where backup footprint is reduced by eliminating redundant data.

To increase compression efficiency, local compression should be guaranteed by preferring data that can be locally compressed (compacted) easily. Storing data that were already compressed (e.g. zip files, mp3 files, images) will result in a bad local compression ratio, thus causing a bad impact in the overall compression ratio. Similar impact happens if administrators choose to store data that is always unique, (e.g. Oracle archives); this will drop the global compression ratio. Other causes for bad compression rates are:

- **Pre-encrypted data:** backup software can be told to encrypt data before sending to the storage system, by doing this chances are that data sent over will not be the same if compared to the previous backup, even if the data didn't change.
- **Tape markers not set:** Backup applications can insert periodic markers into the backup stream for their control. These markers positions may

vary from backup to backup reducing the compression algorithm efficiency. Data Domain filesystem should be configured to efficiently identify these markers. The “auto” marker type of DDFS<sup>2</sup> will work fine in most of the cases.

- **Multiplexing:** Specially for VTL configuration, backup administrators should make sure that the backup software be configured without multiplexing. A backup software that sends data to real tape will have to multiplex data (combine several input streams in one output stream) in order to get better writing performance. For a Data Domain system, backup configuration must be done as to avoid data multiplexing, otherwise, if data gets multiplexed differently from day to day, the compress ratio may be impacted.

Other possible causes for bad compression rates are data with high change rate (>10%), too many small files and random writes (not sequential writes). **Error! Reference source not found.** shows likely causes for bad compression rates.

Global Compression	Local Compression	Likely cause
Low (2x to 4x)	Low (1x to 1.5x)	Precompressed or encrypted data
Low (1x – 2x)	High (>2x)	Unique but compressible data (e.g. archive logs)
Low (2x to 5x)	High (>1.5x)	Markers not detected or high data change rate or stream multiplexing
High (>10x)	Low (<1.5x)	Backups of the same compressed and/or encrypted data. This is uncommon.

Table 1: Likely causes of low compression

### Choosing the best local compression algorithm

The Data Domain default local compression algorithm is lz and is the best choice for most of environments, although this may not be the best choice for every environment so backup administrators should be prepared to choose the best algorithm before putting

<sup>2</sup> DDFS is the Data Domain filesystem

the DD system on production. Every algorithm has its pros and cons. The choice may be based on the following (16):

- In general the order of compression is as follows:
  - lz < gzfast < gz
- The rough difference is:
  - lz to gzfast gives ~15% better compression and consumes 2x CPU
  - lz to gz gives ~30% better compression and consumes 5x CPU
  - gzfast to gz gives ~10-15% better compression

Changing the compression algorithm will only affect new data, that's why it is not recommended to change it on a production DD system without talking to EMC support previously.

### **Cumulative compression ratio**

Individual compression ratios are not added up linearly. On article (16) EMC engineering explains:

*“Suppose the compression on the first full backup is 2x and that on the second full backup is 20x. The cumulative compression is not (2+20)/2 or 11x, but  $2/(1/2+1/20)$  or 3.64x. In general, lower compression ratios have more impact than higher ones on the cumulative compression ratio.”*

The formula for calculating the cumulative compression rate is:

$$C = (\text{total logical size})/(\text{total space used})$$

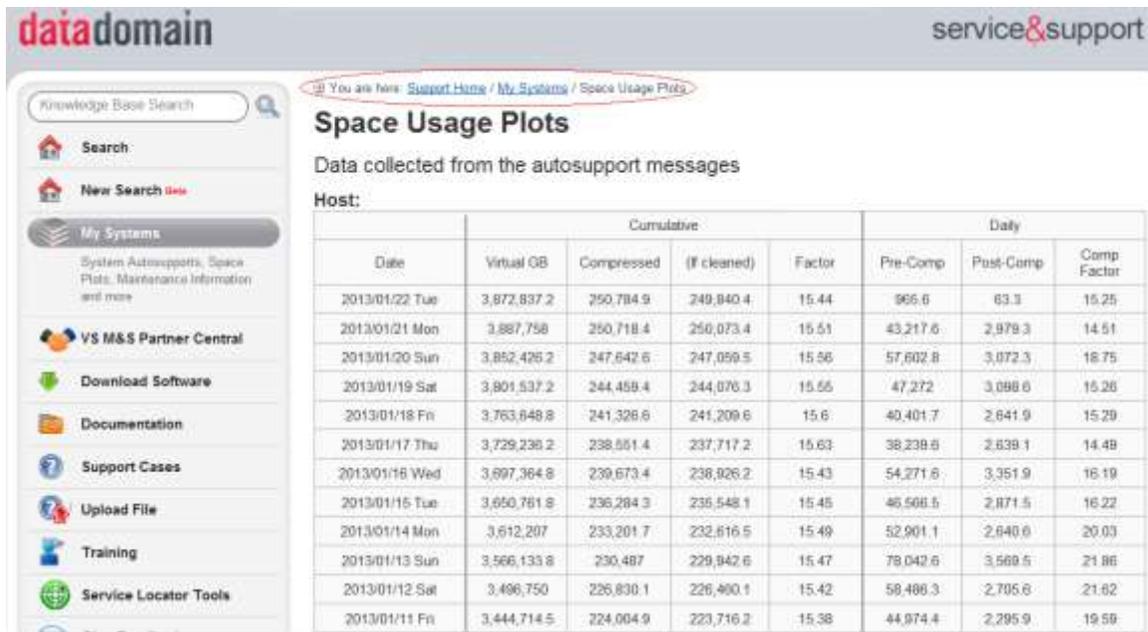
$$\text{total logical size} = s_1 + s_2 + \dots + s_k$$

$$\text{total space used} = s_1/c_1 + s_2/c_2 + \dots + s_k/c_k$$

Document (16) should be read for a detailed discussion.

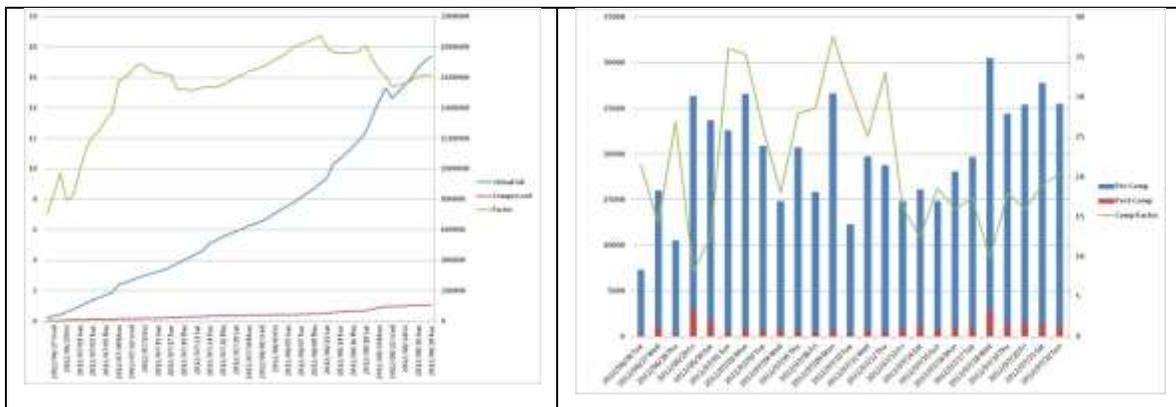
### **Compression verification**

Using the Data Domain website (my.datadomain.com or “support.emc.com -> Data Domain”), it is possible to get historical data about compression. After logging in to my.datadomain.com,(17) go to “My Systems”, select the Data Domain system (by hostname or serial number), click on the system, then “View space plot”. At the bottom of the space plot page, there is a link “View detailed tabular data”. Click on this link to get a table of the data compiled from the autosupport reports sent to EMC over the time. Figure 2 shows the tabular data that can be taken from my.datadomain.com (Support Home / My Systems / Space Usage Plots).



**Figure 2: Space usage tabular view - my.datadomain.com**

Using the data taken from table shown in Figure 2 backup administrator can easily create charts using excel or other tools, Figure 3 shows the charts created (cumulative on the left, daily on the right). These charts are similar to the ones available on the Data Domain Enterprise Manager GUI.



**Figure 3: Data Domain usage charts**

Cumulative usage charts should be used to identify period of times where compression ratio changed its behavior. Cumulative usage chart is also useful to compare source and target Data Domain usage when replication is managed by the backup software. Administrators may quickly identify any rough difference between source and target, what would indicate a missing Netbackup storage lifecycle policy, for example. The daily usage chart is useful to verify, for example, weekly backups that are dropping the compression rates.

Two other useful charts about compression can be created with data of a specific day. This can be done using the “system show performance” output, found also in the autosupport report. The “system show performance” output gives you the global compression and the local compression over the day (gcomp and lcomp columns).

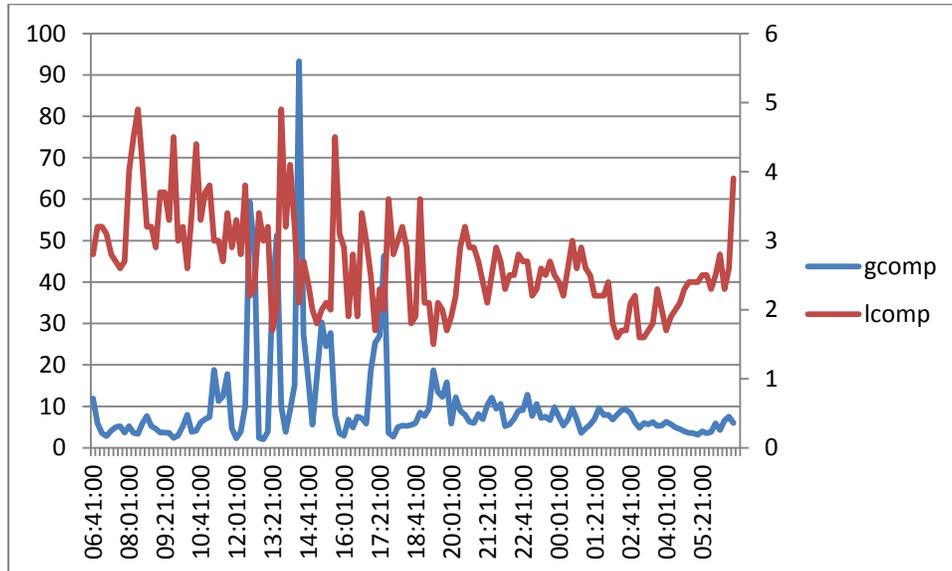


Figure 4: Global/Local compression chart

Backup administrator can use chart shown in Figure 4 to identify if during the day there are periods where global or local compression is not good. Based in these numbers **Error! Reference source not found.** can be used as a starting point to understand what backup is hurting the compression rates.

## Maintenance activities: the clean operation

For a Data Domain system, disk space is not recovered right after a file deletion. The clean operation is the process, usually run in a weekly basis, where Data Domain systems recover disk space in use by expired data. By default the clean operation runs every Tuesday, at 6 am, with a 50% of CPU throttle. The default should not be changed to a more frequent, more aggressive, configuration unless indicated by EMC support. Document (18) should be read for a complete discussion but some attention points about the filesystem clean operation are:

- Unless indicated by EMC support, do not change the cleaning schedule to have more frequent cleaning cycles because this will decrease Data Domain efficiency (compression and performance);

- Change cleaning day and throttle with caution. This must be done according to the backup/restore schedule since the cleaning process may impact Data Domain overall performance;
- Directory replication may impact the cleaning efficiency since data segments will not be removed from source before they get replicated. In case of severe replication lags, source Data Domain system may get full because segments are not being replicated in a timely fashion.
- Clean operation will not recover disk space associated with snapshots that are not expired. Check the existence of snapshots, whenever a clean operation does not recover the expected amount.

Keep in mind that the cleaning process results in a bigger overhead on systems approaching full capacity. A good practice is getting in touch with an EMC representative to expand system capacity whenever the system reaches 80% of disk space usage.

## Data Domain performance analysis

Data Domain performance (I/O speed) depends on several factors that can be divided in two groups:

**External:** comprised of those factors related to the backup environment and data path. This is composed by every configuration item, hardware or software, from the host/storage where the data resides, passing through the Storage Node host configuration, considering also every network component (either SAN or LAN) and their configuration.

**Internal:** are those factors related to the Data Domain itself. These internal factors are, for example, the DD model, DDOS version, local compression algorithm, replication configuration, encryption, number of streams in parallel. In a typical backup environment it would be only the Data Domain systems (the orange area of Figure 5, except the WAN infrastructure).

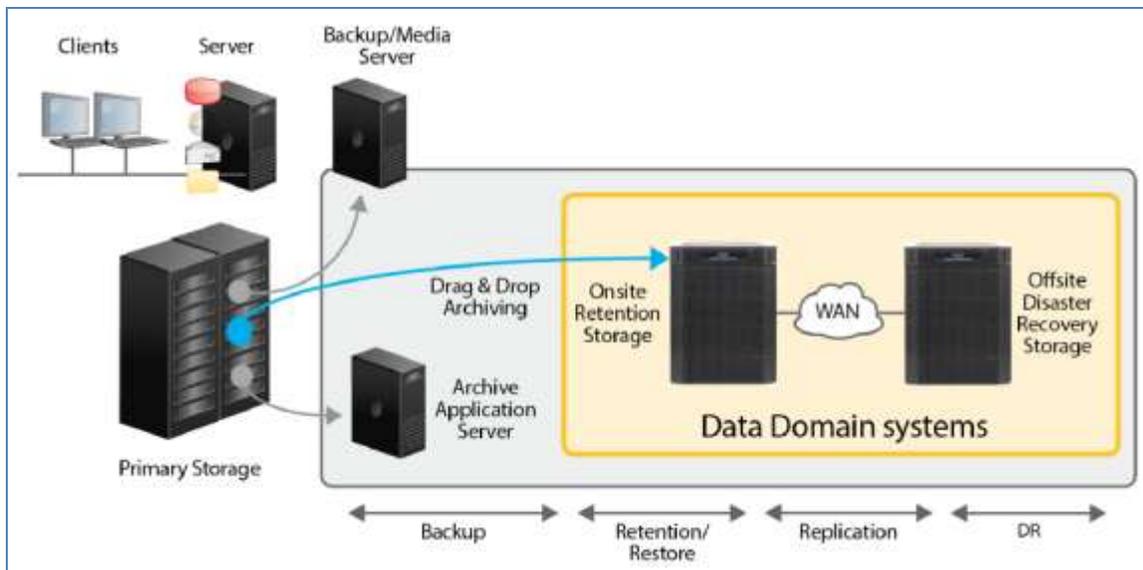
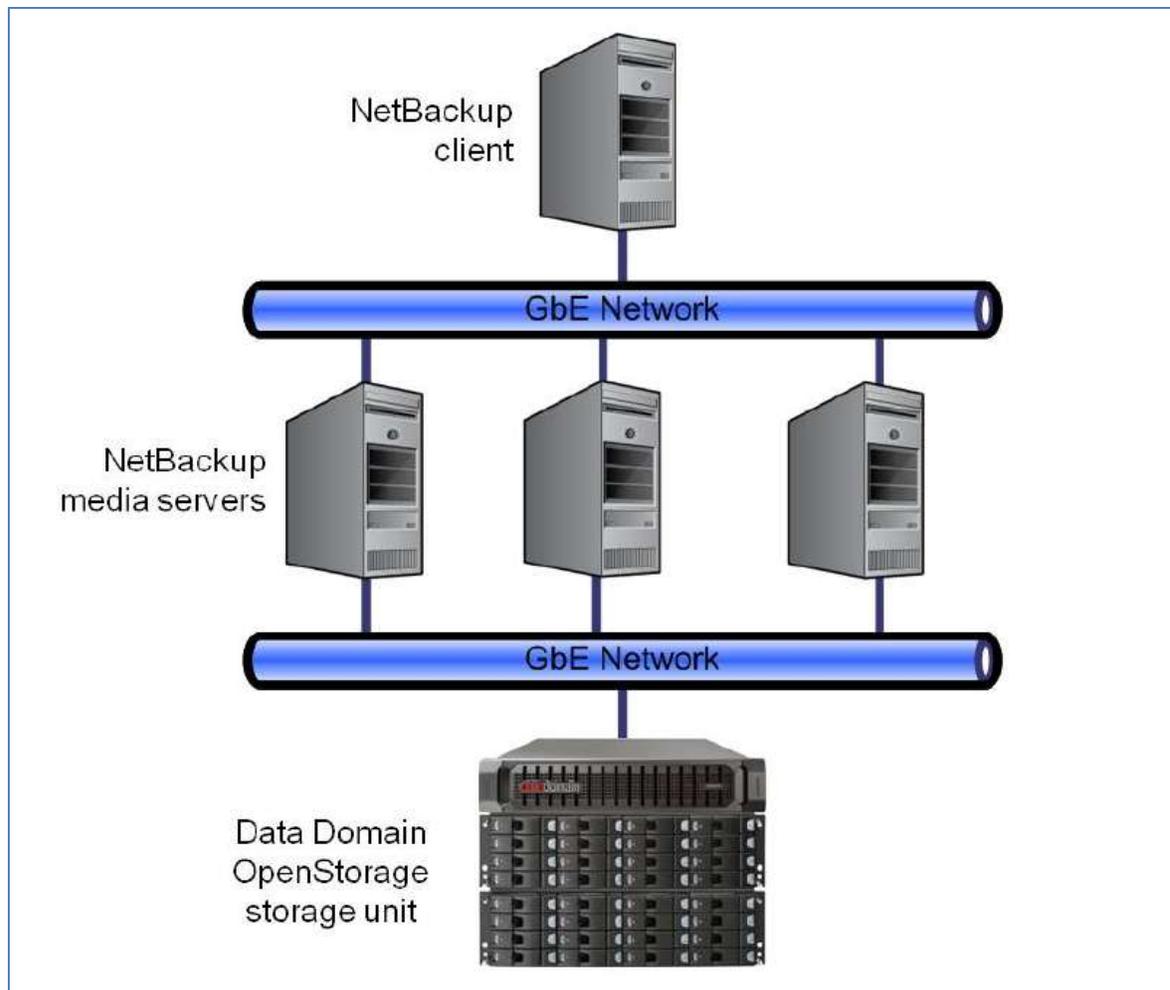


Figure 5: Typical backup environment

### External performance factors

In general these are the factors that constrain the backup performance. Some of these factors impact any other technology present in the backup environment. The network, for instance, if not well designed, configured, and sized will impact not only the Data Domain but other applications using the infrastructure. A good idea is to research how other applications and/or storage systems are performing. E performance factors to keep in mind are:

**The network:** Data Domain systems rely on good, well-tuned networks, either SAN, WAN, or LAN. A good practice is to segregate the connectivity between Data Domain systems and their clients in a separate network. If it is not possible to have a separate physical network, a good alternative is having a separate VLAN (or VSAN) just for the Data Domain traffic (19).



**Figure 6: Segregated network for Data Domain traffic**

Backup administrators should test the I/O performance of the network to understand speed limits of the network infrastructure, for this iperf is a suitable tool (20). In case of bad network performance, backup administrators should tune the network accordingly, using considerations specific for the OS and backup software in use. EMC knowledge base has several documents in this subject (21)(22), but the support of the OS and backup software must be contacted too. Backup administrators should also check other network factors such as:

- **Latency:** using ping, latency will be the RTT divided by 2;
- **Host resolution:** forward and reverse lookup may be delayed by DNS related issues. This can be checked by running nslookup from both client and DD. Use hosts files whenever possible to avoid name resolution delays;

- **Routes:** the more complex the network topology is, the greater the likelihood of specific routes will be needed. Use traceroute/tracepath to understand interfaces being used. The use of a wrong interface will dramatically drop the backup performance. Use the iostat command to check which interface is being used.
- **Network quality:** no matter how fast a network is, the performance may not be good if there is a high amount of packet losses and/or packet retransmissions. Administrators can use ping, iperf and netstat to verify network quality.
- **Jumbo frames:** The use of jumbo frames (23) may improve network performance, depending on the data being transferred. Not only interfaces on both ends (DD and its clients) but also all other network components must be configured for jumbo frames. Maximum MTU size for DDOS is 9014. The MTU size must be the lowest supported along the data path.

Besides the network related factors, other commonly found external factors are:

- **Clients directly connected:** Overall performance will depend on how powerful are the clients and/or how many clients are connected to the Data Domain (where client means “the host directly connected to the Data Domain”). It doesn’t matter if, for example, a Data Domain has a 10GbE interface if hosts can only push 2GbE over the network, throughput will not reach the 10GbE limit.
- **Clients connected to Storage Nodes (media servers):** If hosts (where the data reside) are not directly connected to the Data Domain systems, administrators must make sure that the I/O speed (over the network between hosts and Storage Nodes) will be fast enough to accomplish the desired backup speed.

### **Internal performance factors**

The internal performance factors are central when talking about Data Domain efficiency because they should be considered and worked by the backup administrators at the beginning, during the Data Domain implementation, and in a daily basis during the normal operation. The main internal factors are:

- **DD Model:** during the pre-sales phase the EMC technical consultant should be informed about how the Data Domain will be used. The backup administrator should provide detailed information on what kind of data will be backed up, how the backup schedule is, how long the data will be retained, what DD features should be available, access methods that will be used, network technology available, OS of the hosts that will be connected to the Data Domain, backup software versions involved, backup software licensing, and so on. All this information will be critical while choosing the appropriate DD model. Data Domain models differs from each other especially on how their capacity can be expanded and what performance they may achieve (number of simultaneous streams).
- **Encryption:** whenever encryption has to be used, the appropriated DDOS version must be in use to guarantee a good performance. In older DDOS versions the encryption used to require more resources, so upgrading the DDOS may improve the DD performance if it is possible to start using new encryption technologies available.
- **MTrees:** DDOS 5.2 supports up to 100 MTrees, although the best practice is using simultaneously no more than 14 MTrees for backup (excluding MTrees used only for replication)
- **Local compression:** as mentioned before on page 13, the use of CPU will vary depending on the chosen local compression algorithm. This is not true for DDBoost when DSP<sup>3</sup> is enabled, since this feature will delegate the compress overhead to the DDBoost client.
- **Simultaneous Streams:** Every Data Domain model has a different limit of simultaneous streams. Document (24) has a table of concurrent streams limits for DDOS 5.2. In order to achieve the optimal performance, the number of concurrent streams must be close to the DD model in average. If this limit of streams is exceeded, a decrease in performance may occur (25). The table in document (24) has also the hard limits of concurrent streams which, if exceeded, will result in an extremely low performance and a severe impact in compression rates. Administrators should read document (25) to get familiar with the stream count best practices.

---

<sup>3</sup> DSP: Distributed Segment Processing

## Performance verification

Backup administrators can check Data Domain performance in a daily basis using “system show performance” command (3), run through the command line or taken from the autosupport report. Document (3) should be read for a good comprehension of the “system show performance” command output. Following are some points to check with this command:

Proc/Send/Recv: DDFS processing consists basically of three steps: receiving a request over the network, processing the request, sending the reply back to the request originator.

A high proc indicates that the Data Domain system is working to its limit, which is good. A low proc indicates that the system is not being stressed enough and there must be an external factor acting as a bottleneck.

If send/recv is high, chances are that the host connected to the Data Domain system is overloaded and/or there is a network related issue making the system wait for send/recv transactions to be transmitted.

If proc, recv, and send are all low, we can conclude that the Data Domain system is lightly loaded. If this occurs, administrators should figure out if Data Domain clients are really able to send enough data to be stored.

Other variables that can be taken from the “system show performance” output are shown in **Error! Reference source not found..**

Administrators can create charts, for example, to evaluate throughput and number of concurrent streams along the day (Figure 7).

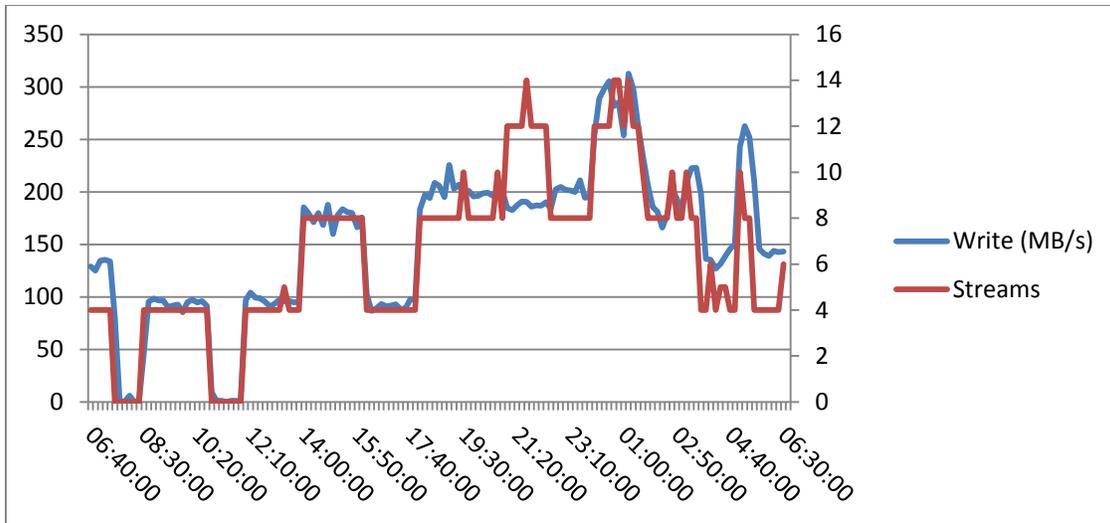


Figure 7: "system show performance" chart example

Throughput	
Read	read throughput from the DDR
Write	write throughput to the DDR
Repl Network	replication network throughput in to and out of the DDR
Repl Pre-comp	replication pre-compressed (logical) throughput into and out of the DDR (always zero for collection replication)
Protocol	
ops/s	Operations per second.
load %	Load percentage (pending ops/total RPC ops *100).
data(MB/s) in/out	Protocol throughput. Amount of data the filesystem can read from and write to the kernel socket buffer.
wait(ms/MB) in/out	Time taken to send and receive 1MB of data from the filesystem to kernel socket buffer.
Utilization	
proc	percent time spent processing requests
rcv	percent time spent receiving requests over the network
send	percent time spent sending requests over the network
idle	percent idle
Compression	
gcomp	global compression rate
lcomp	local compression rate
Streams	
rd	active read streams
wr	active write streams
r+	reopened read file streams in the past 30 seconds
w+	reopened write file streams in the past 30 seconds
Utilization	
cpu	average and max CPU utilization; the CPU ID of the most loaded CPU is shown in the brackets
disk	max (highest) disk utilization over all disks; the disk ID of the most loaded disk is shown in the brackets (may not correspond to the ID used by 'disk show' commands)

**Table 2: "system show performance" main variables**

## **Conclusion**

Every backup environment is different and, because of that, every Data Domain should be carefully configured, following the known best practices. Not only configured, but monitored and reconfigured if needed. This is not a trivial duty but can be successfully accomplished by taking advantage of the available documentation, knowledge base articles, and EMC technical assistance. On the other hand, it is not a simple task to summarize the wide range of available knowledge base articles, in order to create a basis for the skill set required to guarantee satisfactory efficiency of Data Domain, thus having an efficient backup system.

This article describes several aspects to be considered when working with a Data Domain system, and summarizes in its references a list of very important articles. However, one should not avoid the use of EMC's knowledge base and support assistance in activities needed to achieve the best level of efficiency.

## Bibliography

1. **EMC.** How To Confirm Compression Rates. *my.datadomain.com*. [Online] 07 08, 2011. [Cited: 01 18, 2013.]  
[https://my.datadomain.com/download/kb/appliance/confirming\\_compression\\_rate.html](https://my.datadomain.com/download/kb/appliance/confirming_compression_rate.html).
2. —. Data Domain DDBoost compatibility guide.
3. —. System Show Performance FAQ. *my.datadomain.com*. [Online] 01 18, 2013. [Cited: 01 22, 2013.]  
[https://my.datadomain.com/download/kb/all/SystemShowPerformance\\_FAQ\\_Internal.html](https://my.datadomain.com/download/kb/all/SystemShowPerformance_FAQ_Internal.html).
4. —. *DD OS 5.2 Administration Guide*. Santa Clara : EMC, 2012.
5. —. *CIFS and Data Domain Systems*. Santa Clara : EMC, 2011.
6. —. Accessing the DDR with NFS. *my.datadomain.com*. [Online] 03 14, 2012. [Cited: 01 18, 2013.] [https://my.datadomain.com/download/kb/rp/accessing\\_ddr\\_nfs\\_rp.html](https://my.datadomain.com/download/kb/rp/accessing_ddr_nfs_rp.html).
7. —. *Integrating the Data Domain Operating System 5.2 VTL with a Storage Area Network (for Open Systems)*. Santa Clara : EMC, 2012.
8. —. VTL Best Practices Guide. *my.datadomain.com*. [Online] 05 24, 2012. [Cited: 09 24, 2012.] [https://my.datadomain.com/download/kb/all/VTL\\_Best\\_Practices\\_Guide.html](https://my.datadomain.com/download/kb/all/VTL_Best_Practices_Guide.html).
9. —. VTL Knowledge Base Articles by Component Listing. *my.datadomain.com*. [Online] 09 11, 2012. [Cited: 09 24, 2012.]  
[https://my.datadomain.com/download/kb/all/VTL\\_KB\\_Listing.html](https://my.datadomain.com/download/kb/all/VTL_KB_Listing.html).
10. —. Access Groups Best Practices. *my.datadomain.com*. [Online] 12 04, 2012. [Cited: 01 18, 2013.]  
[https://my.datadomain.com/download/kb/all/Access\\_Groups\\_Best\\_Practices.html](https://my.datadomain.com/download/kb/all/Access_Groups_Best_Practices.html).
11. —. MTree Replication Frequently Asked Questions. *my.datadomain.com*. [Online] 12 05, 2012. [Cited: 01 21, 2013.]  
[https://my.datadomain.com/download/kb/all/Replication\\_Mtree\\_FAQ.html](https://my.datadomain.com/download/kb/all/Replication_Mtree_FAQ.html).
12. —. EMC Data Domain Boost Software. *powerlink.emc.com*. [Online] 05 01, 2012. [Cited: 01 18, 2013.] <http://powerlink.emc.com>.
13. —. *Data Domain Boost for OpenStorage Administration Guide 2.5*. Santa Clara : EMC, 2012.
14. —. *EMC® NetWorker® Data Domain® Deduplication Devices Release 8.0 Integration Guide*. Santa Clara : EMC, 2012. 300-013-562.
15. —. EMC Data Domain Systems. *powerlink.emc.com*. [Online] 12 01, 2012. [Cited: 01 18, 2013.] <http://powerlink.emc.com>.

16. —. DataDomain Restorer & Compression: Frequently Asked Questions. *my.datadomain.com*. [Online] 07 18, 2008. [Cited: 01 18, 2013.]  
[https://my.datadomain.com/download/kb/appliance/compression\\_faq.html](https://my.datadomain.com/download/kb/appliance/compression_faq.html).
17. —. How to Get Historical System Usage. *my.datadomain.com*. [Online] 01 22, 2013. [Cited: 01 22, 2013.]  
[https://my.datadomain.com/download/kb/all/How\\_to\\_Get\\_Historical\\_System\\_Usage.html](https://my.datadomain.com/download/kb/all/How_to_Get_Historical_System_Usage.html)  
.
18. —. Scheduling Cleaning on a Data Domain System. *my.datadomain.com*. [Online] 10 03, 2012. [Cited: 01 22, 2013.]  
[https://my.datadomain.com/download/kb/appliance/scheduling\\_cleaning-ddr.html](https://my.datadomain.com/download/kb/appliance/scheduling_cleaning-ddr.html).
19. —. EMC Data Domain Boost for Symantec NetBackup OpenStorage - Best Practices Planning. *powerlink.emc.com*. [Online] 07 01, 2010. [Cited: 01 18, 2013.]  
<http://powerlink.emc.com>.
20. —. Testing Network Bandwidth Using the net iperf Command. *my.datadomain.com*. [Online] 05 10, 2012. [Cited: 01 18, 2013.]  
[https://my.datadomain.com/download/kb/appliance/using\\_iperf\\_command.html](https://my.datadomain.com/download/kb/appliance/using_iperf_command.html).
21. —. Network Performance Troubleshooting. *my.datadomain.com*. [Online] 12 06, 2012. [Cited: 01 18, 2013.]  
[https://my.datadomain.com/download/kb/all/Network\\_Performance\\_Troubleshooting.html](https://my.datadomain.com/download/kb/all/Network_Performance_Troubleshooting.html)  
l.
22. —. Tuning Windows Media Servers for Performance. *my.datadomain.com*. [Online] 03 26, 2012. [Cited: 01 22, 2013.]  
[https://my.datadomain.com/download/kb/all/Tuning\\_Windows\\_Media\\_Servers.html](https://my.datadomain.com/download/kb/all/Tuning_Windows_Media_Servers.html).
23. —. Jumbo Frames Definition. *my.datadomain.com*. [Online] 03 10, 2011. [Cited: 01 18, 2013.]  
[https://my.datadomain.com/download/glossary/Jumbo\\_Frames\\_Definition.html](https://my.datadomain.com/download/glossary/Jumbo_Frames_Definition.html).
24. —. Supported Stream Counts for Data Domain OS 5.2. *my.datadomain.com*. [Online] 03 12, 2012. [Cited: 01 22, 2013.]  
[https://my.datadomain.com/download/kb/all/StreamCountTable\\_DDOS5.2.html](https://my.datadomain.com/download/kb/all/StreamCountTable_DDOS5.2.html).
25. —. Stream Counts Frequently Asked Questions. *my.datadomain.com*. [Online] 11 29, 2012. [Cited: 01 22, 2013.]  
[https://my.datadomain.com/download/kb/all/StreamCounts\\_FAQ\\_Internal.html](https://my.datadomain.com/download/kb/all/StreamCounts_FAQ_Internal.html).

26. —. Best Practices for Directory and Pool Replication. *my.datadomain.com*. [Online] 07 08, 2011. [Cited: 01 18, 2013.]

<https://my.datadomain.com/download/kb/all/ReplicationBestPractices.html>.

27. —. Jumbo Frames Usage for DDBoost. *my.datadomain.com*. [Online] 12 20, 2012. [Cited: 01 18, 2013.]

[https://my.datadomain.com/download/kb/all/Jumbo\\_Frame\\_Usage\\_for\\_DDBoost.html](https://my.datadomain.com/download/kb/all/Jumbo_Frame_Usage_for_DDBoost.html).

28. *NFS Server Tuning*. EMC, 2012.

## **Author's Biography**

Erico is a Solutions Architect working for EMC since 2000, supporting the products of EMC BRS portfolio. He possesses a Master's degree in engineering, specializing in object oriented development and parallel processing. Erico has worked with BRS delivery/support teams for 13 years, previously worked with IT support for 4 years, and as a C/C++ developer for 6 years. He is a certified EMC Proven Professional with certifications for NetWorker, NAS, Network Management, and Cloud Infrastructure.

EMC 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." EMC CORPORATION MAKES NO REPRESENTATIONS 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 EMC software described in this publication requires an applicable software license.