Dell HTSS + DX Object Storage

A Technical Supplement

Quy Ta

Dell HPC Engineering

This document is for informational purposes only and may contain typographical errors and technical inaccuracies. The content is provided as is, without express or implied warranties of any kind.

November 2011| Rev 1.0

Overview

This solution supplement describes the Dell HPC Tiered Storage Solution (HTSS) with DX Object Storage Platform (DX) integration. The Dell HTSS with DX utilize CommVault Simpana 9 Data & Information Management Software to provide a multi-tiered Hierarchical Storage Management system. The introduction of the Dell DX Object Storage provides a disk based alternative to tape for a Long Term Storage (LTS) tier. The goal is to provide a data storage technique that offers Hierarchical Storage Management (HSM) and archive solutions that automatically move data between storage tiers through the use of defined storage policies. This document describes the architecture, performance, and best practices for integrating the Dell DX Object Storage platform in such solutions.

Dell HTSS + DX technical overview

This section provides a quick summary of the technical details of the Dell HPC Tiered Storage Solution with DX Object Storage offering.

The Dell HTSS + DX offering consists of a Dell PowerVault DL2200 Disk-Based Backup Appliance and a Dell DX6000 Object Oriented Storage Platform. The DL2200 appliance serves as the HSM data manager and utilizes CommVault Simpana 9 Data & Information Management Software, but it is not in the data path for the actual HSM or archive process. This is the central point of the HSM network, where storage resources and policies are created and administered. The DX6000 Object Storage platform serves as the designated ***long-term storage*** tier in this solution. This Long-Term Storage or LTS tier allows for low cost/high capacity storage and for long term retention of data.

Dell HTSS supports the Dell NFS Storage Solution[(1)] (NSS) as the designated ***primary*** storage tier in both standalone (NSS) and high availability (NSS-HA) configurations. The NSS and NSS-HA solutions use the NFS file system on top of the Red Hat Scalable File System Add-on (based on XFS) with Dell MD PowerVault as back end storage, to provide an easy to manage, reliable, and cost-effective solution for unstructured data.

Figure 1 and 2 illustrate the HTSS + DX reference architecture with Dell NSS and NSS-HA configurations.

Figure 1: HTSS + DX reference architecture with NSS configuration

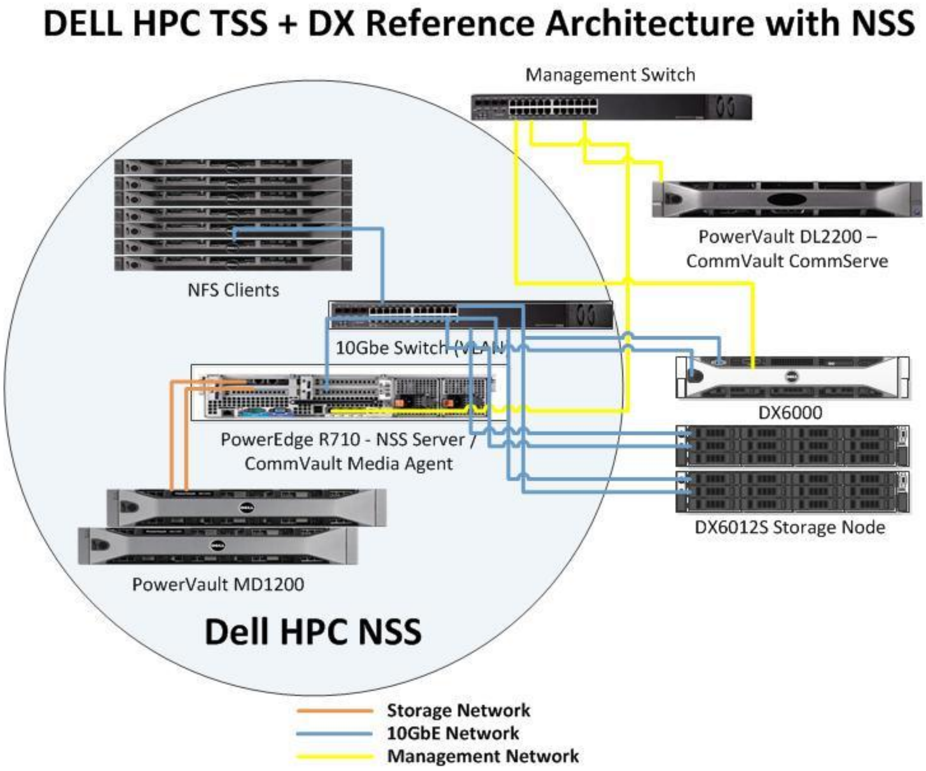
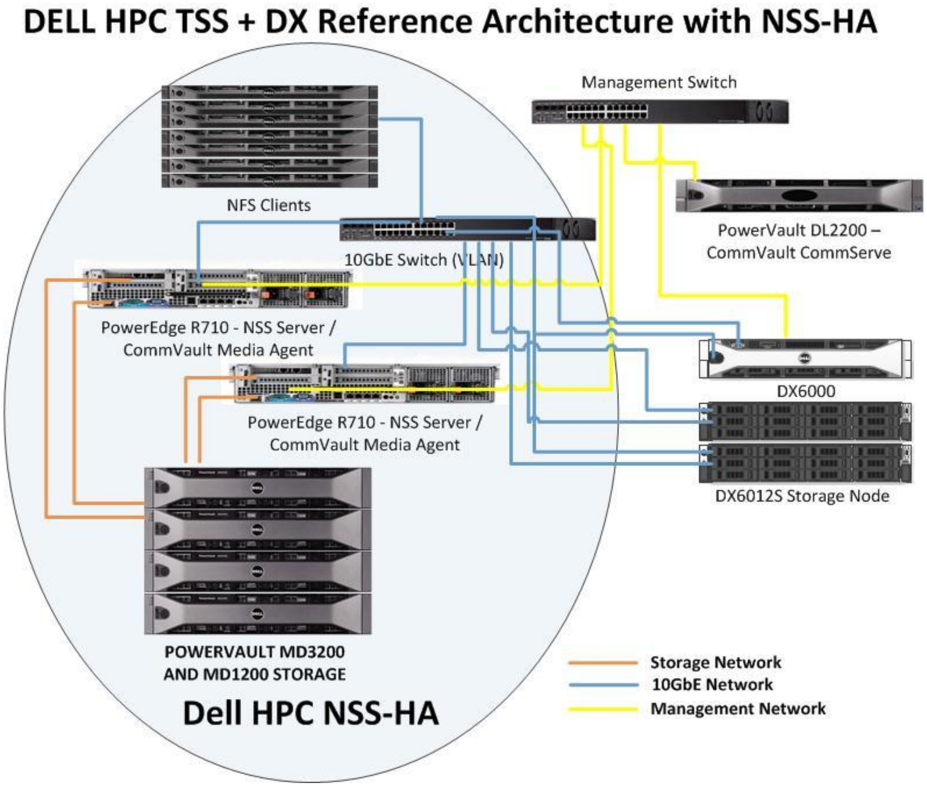


Figure 2: HTSS + DX reference architecture with NSS-HA configuration


DELL HPC TSS + DX Reference Architecture with NSS-HA
Management Switch
NFS Clients
PowerVault DL2200 – CommVault CommServe
10GbE Switch (VLAN)
PowerEdge R710 - NSS Server / CommVault Media Agent
DX6000
PowerEdge R710 - NSS Server / CommVault Media Agent
DX6012S Storage Node
POWERVAULT MD3200 AND MD1200 STORAGE
Dell HPC NSS-HA
Storage Network
10GbE Network
Management Network

DX Object Storage Platform

DX6000 Storage Cluster Service Node (CSN)

The DX Storage Cluster Services Node (CSN) is an integrated services node that centralizes installation and configuration of both the network services required to run a DX Storage cluster and the software used to interface with it.

The CSN distribution is available as a collection of RPM packages that are installed with a shell script. The packages and their dependencies must be installed as the 'root' user from an attached monitor and keyboard with at least one NIC connected to the external network using the following steps.

1) Copy the distributed zip file to your Cluster Services Node and unzip it into the directory of your choice. The following example command would unzip the version of the software which was previously copied to the /tmp/csninstall directory:

```
$ sudo su -
# cd /tmp/csninstall
# unzip caringo-csn-2.0.0.zip
```

2) Install the CSN by running the self-extracting script from the directory location where the shell script was unzipped. For instance, continuing the example above from the /tmp/csninstall directory:

```
# cd caringo-csn-2.0.0
# ./caringo-csn-bundle-install.sh
```

This command will initiate installation of the CSN and its dependent packages. When the installation is complete, the following prompt will display:

```
Would you like to proceed with CSN network configuration? (yes/no):
```

Answer 'yes' to proceed with configuring the CSN. If you answer 'no', you may return to the configuration screen at a later time by running the command: `/opt/caringo/csn/bin/firstcsnboot`.

If you run the configuration at a later time, it must still be from an attached monitor and keyboard.

CSN Configuration

After installing the CSN, you will automatically be prompted to enter some minimal configuration data to configure the server on the overall network. Network settings are central to all CSN services and should be planned with care in advance by an administrator knowledgeable about the environment. The initial configuration process is only required once after the initial installation. Any necessary subsequent updates to the initial configuration parameters can be made from the CSN Console.

Several prompts will suggest a default value in brackets that can be accepted by simply pressing enter.

`Is this the Primary CSN (yes/no)? [yes]::` This prompt allows specification of whether or not the CSN is a primary or secondary CSN. The default value is yes. Administrators should take care to ensure that only a single primary CSN is configured on the internal network to prevent conflicts with both DHCP and DX Storage netboot configuration. The primary server must be configured prior to configuration of a secondary. DHCP is not started on the secondary CSN.

Half of the NIC ports on this system will be bonded and assigned to the external network. The following questions configure the external network:

```
      Enter the CSN IP address []:
```

This parameter requires entry of an external IP address for the CSN node. The entered address must be in a valid w.x.y.z format and must not already be in use on the network.

`Enter the cluster IP address. This IP address will remain with the Primary CSN in the event of a CSN failover []:` This parameter requires entry of an external IP address for the CSN cluster address. This well-known address remains with the primary CSN in the event of a failover, meaning the cluster can always be reached at this address. The entered address must be in a valid w.x.y.z format and must not already be in use on the network.

`Enter the subnet mask [255.255.255.0]:` This parameter requires entry of the subnet mask for the external network that corresponds with the entered IP address. The default is 255.255.255.0.

`Enter the gateway IP address []:` This parameter requires entry of the gateway associated with the entered IP address for the external interface.

Half of the NIC ports on this system will be bonded and assigned to the
internal network. The following questions configure the internal network:

```
      Enter the network address, e.g. 192.168.100.0 (small network),
      192.168.0.0, 172.20.0.0 (large network) []:
```

This parameter allows specification of the network interface that should be used for the internal network. Enter an interface in the format of 192.168.100.0 to use a small interface that will support 128 DX Storage nodes. Enter an interface in the format of 192.168.0.0 or 172.20.0.0 to use a larger network that will support much larger DX Storage clusters. The entered interface will be divided between the CSN(s) and privileged applications on the internal network and the DX Storage nodes. The initial configuration process automatically creates multiple alias IP addresses on the internal network for use by various system services and reserves similar IP addresses for a Secondary CSN.

`Enter a list of IP addresses (separated by spaces) for external name servers [8.8.8.8 8.8.4.4]:` This parameter allows specification of one or more DNS servers for the external interface. Entries must be separated by spaces. Publicly available name servers have been defaulted.

`Enter a list of IP addresses or server names (separated by a space) for external time servers [0.pool.ntp.org 1.pool.ntp.org 2.pool.ntp.org]:` This parameter allows specification of one or more NTP servers for the external interface. The defaults are public NTP servers.

`Enter a unique storage cluster name. This name cannot be changed once assigned. A fully qualified domain name is recommended []:` This parameter is used to populate the name of the storage cluster on the admin console as well as in stream metadata for all streams written to the local cluster. The CSN also uses this name to detect all the nodes participating in the cluster. For all of these purposes, the name must be unique. An IANA fully qualified domain name is recommended.

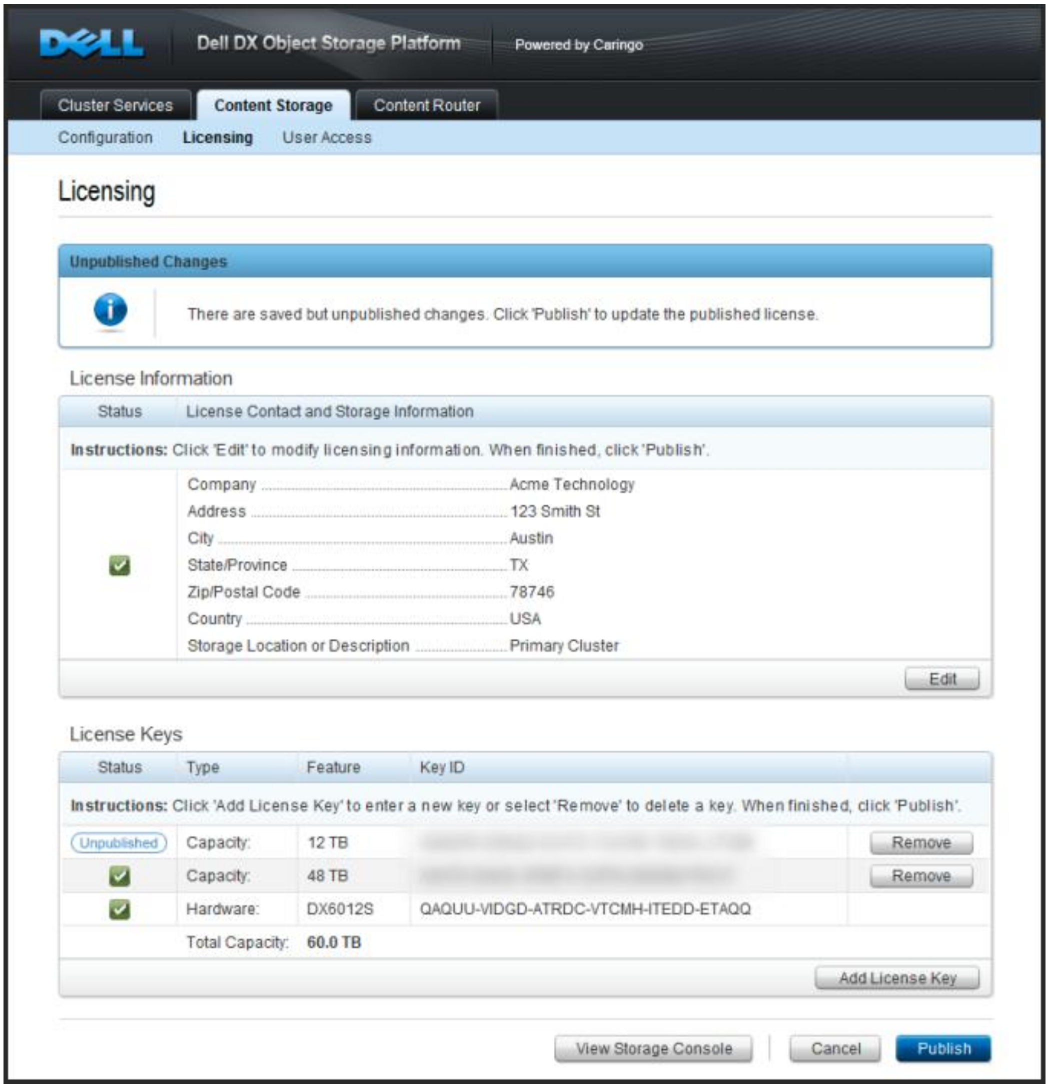
`Are these values correct (yes/no)?` This last step allows you to review the values entered for all prompts before submitting them. Answering yes will allow the initial configuration process to proceed with network and service configuration, resulting in a fully functional CSN. Answering no to the final initial configuration prompt will restart the initial configuration script at the first prompt, with the previously entered values populated.

At the completion of a successful initial configuration, the CSN will immediately reboot the server to initialize all services. When the node comes back up all network services will be configured and available including SNMP, syslog, DHCP, DNS, NTP, and firewall. Additionally, the CSN Console will be available, the SCSP Proxy will be configured and started and the DX Content Router Publisher will be configured and started.

Configuring Object Storage Cluster Licensing

Subsequent to rebooting after configuration and prior to booting any storage cluster nodes, an aggregate license file comprised of any capacity keys included with your hardware or sent by your hardware manufacturer must be entered and published via the administrative CSN Console. The console allows web-based configuration of all CSN services after the initial network configuration. Please reference subsequent chapters for a full overview of console capabilities.

To access the console initially for license publication, enter the following address:http://<CSNExternalIP>:8090 . You will be required to authenticate prior to being granted access. The username for the console is 'admin' and the default password is 'dell'. Once authenticated, click on the Licensing link under the Content Storage tab to access the licensing interface.



Licensing

Unpublished Changes

There are saved but unpublished changes. Click 'Publish' to update the published license.

License Information

Status	License Contact and Storage Information
	Instructions: Click 'Edit' to modify licensing information. When finished, click 'Publish'.
✓	Company .... Acme Technology Address .... 123 Smith St City .... Austin State/Province .... TX Zip/Postal Code .... 78746 Country .... USA Storage Location or Description .... Primary Cluster

License Keys

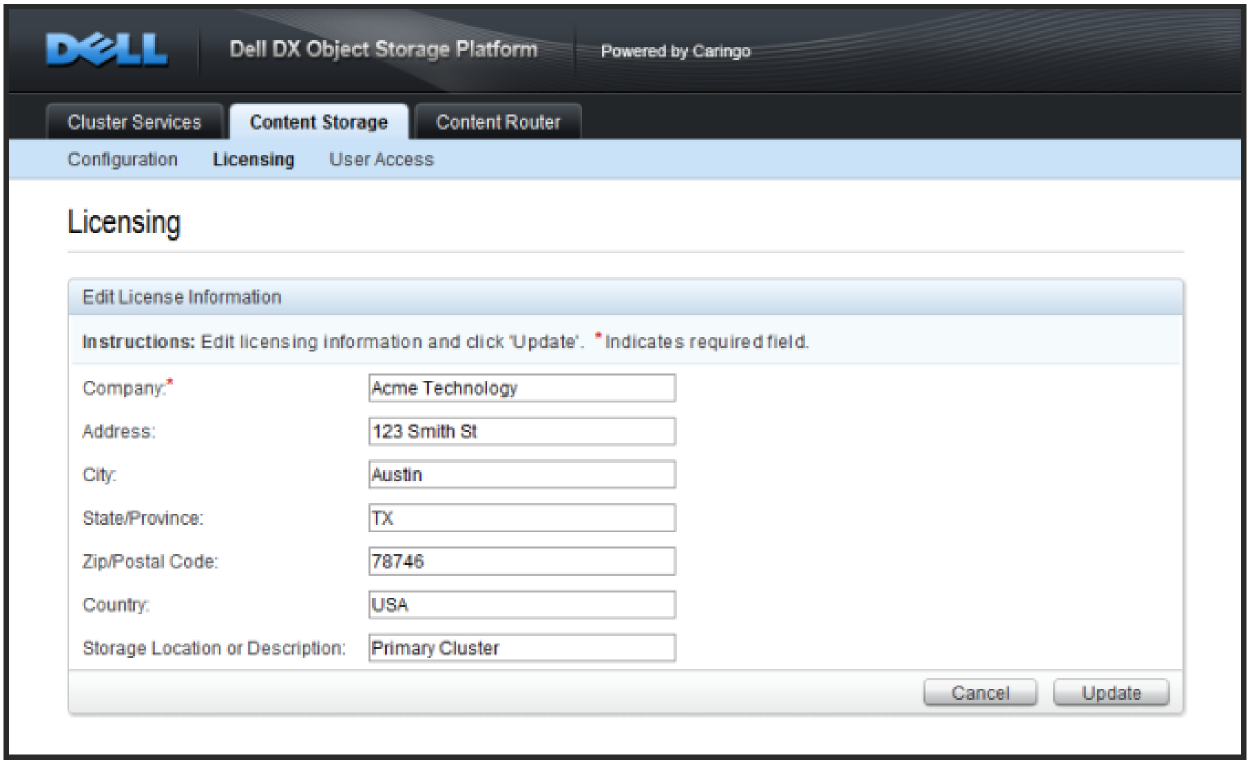
Instructions: Click 'Add License Key' to enter a new key or select 'Remove' to delete a key. When finished, click 'Publish'.

Status	Type	Feature	Key ID	
Unpublished	Capacity:	12 TB	[redacted]	Remove
✓	Capacity:	48 TB	[redacted]	Remove
✓	Hardware:	DX6012S	QAQUU-VIDGD-ATRDC-VTCMH-ITEDD-ETAQQ	
	Total Capacity:	**60.0 TB**		

After installation, a license shell is defaulted into the licensing interface. If published as is, this license will provide 0 TB of capacity for storing content. To correctly publish a license the following high-level steps are required:

1. Enter License Contact and Storage Information
2. Enter capacity keys one at a time
3. After all keys have been entered, publish the license

To add or update this information, click the 'Edit' button under the License Contact and Storage Information section of the interface and enter the desired company name, address and cluster description information. Only the company name is required but populating the address and description fields is recommended to assist with easily distinguishing licenses from multiple locations. When all the desired information has been entered, click the 'Update' button to save the changes.



Licensing

Edit License Information

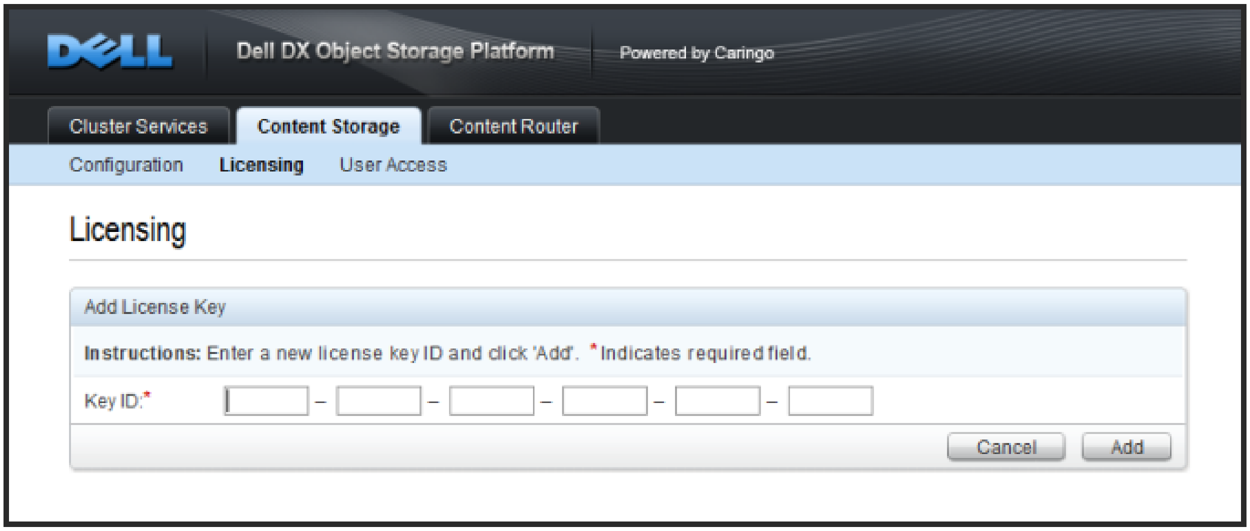
Instructions: Edit licensing information and click 'Update'. * Indicates required field.

Company:*	Acme Technology
Address:	123 Smith St
City:	Austin
State/Province:	TX
Zip/Postal Code:	78746
Country:	USA
Storage Location or Description:	Primary Cluster

The license interface supports entry of two different types of license keys: hardware keys and capacity keys. Hardware keys are used to determine the validity of the host hardware for each storage node at boot. Unauthorized hardware will log a critical error followed by a forced shutdown. Hardware keys are updated infrequently and, in most installs, pre-populated so additional entry is not required. Once entered hardware keys cannot be deleted. Administrators who receive hardware validation errors for storage nodes that are believed to be authorized should contact their support representative.

Capacity keys are more commonly entered than hardware keys. Printed capacity keys are delivered in conjunction with a hardware shipment. The capacity increment for an individual capacity key usually corresponds to the capacity of an individual storage node. Each key represents a preallocated amount of license space from 3TB to 480 TB. The sum of all entered capacity keys determines the total amount of available licensed capacity for the cluster as a whole. The total capacity for all keys is displayed at the bottom of the license key grid for easy validation.

Both types of license keys are represented by an encoded string broken down into 6 groups of 5 characters each, separated by dashes (XXXXX-XXXXX-XXXXX-XXXXX-XXXXX-XXXXX). To enter a new key, click the 'Add License Key' button on the main licensing page and enter the capacity key exactly as printed. When complete, click the 'Add' button to submit the key. Invalid or mistyped keys will result in an error on the screen. To proceed, either correct the key entry or click 'Cancel' to return to the main licensing page. Valid keys will be reflected on the main licensing page after successful entry.



Licensing

Add License Key

Instructions: Enter a new license key ID and click 'Add'. *Indicates required field.

Key ID:* _____ – _____ – _____ – _____ – _____ – _____

Entering hardware and capacity keys in the licensing interface will save the keys to the CSN's local storage. To create a single license file that can be used by the storage cluster, administrators must publish all the saved changes. The licensing interface denotes any saved but not published keys with an 'Unpublished' icon in the 1st column to visually assist with determining when a license needs to be published. To Publish all saved changes, click the blue 'Publish' button at the bottom of the page. This will create a signed license file with all saved changes and update the license file location the storage cluster monitors for license updates. Once the license file has been created after the initial install, the storage nodes can be booted.

Booting DX Storage Nodes

Once the CSN has been configured and rebooted and a DX Storage license has been published with licensed capacity, DX Storage nodes may be powered on, on the internal network. DX Storage nodes require no additional pre-configuration other than to ensure they are configured in BIOS to network boot. As long as that's the case the nodes will automatically receive an IP address from the CSN as well as the pre-configured CAStor software and cluster configuration.

Admins may wish to first ensure that the CSN has had adequate time to sync with a reliable NTP source as a precautionary step prior to booting the nodes. To do this simply type `'ntptrace'` at a system command line. The returned output should contain a stratum value of `'15'` or lower indicating that the CSN's local time server is less than 15 layers separated from a reliable source and is therefore reliable enough to serve as an NTP server for the DX Storage on the internal network.

CommVault Agents on DX6000 Cluster Service Node

You will need to install the CommVault Media Agent driver on the DX6000 CSN and configure a `“Disk Device”` based storage resource in the CommCell for the DX Object Storage resource.

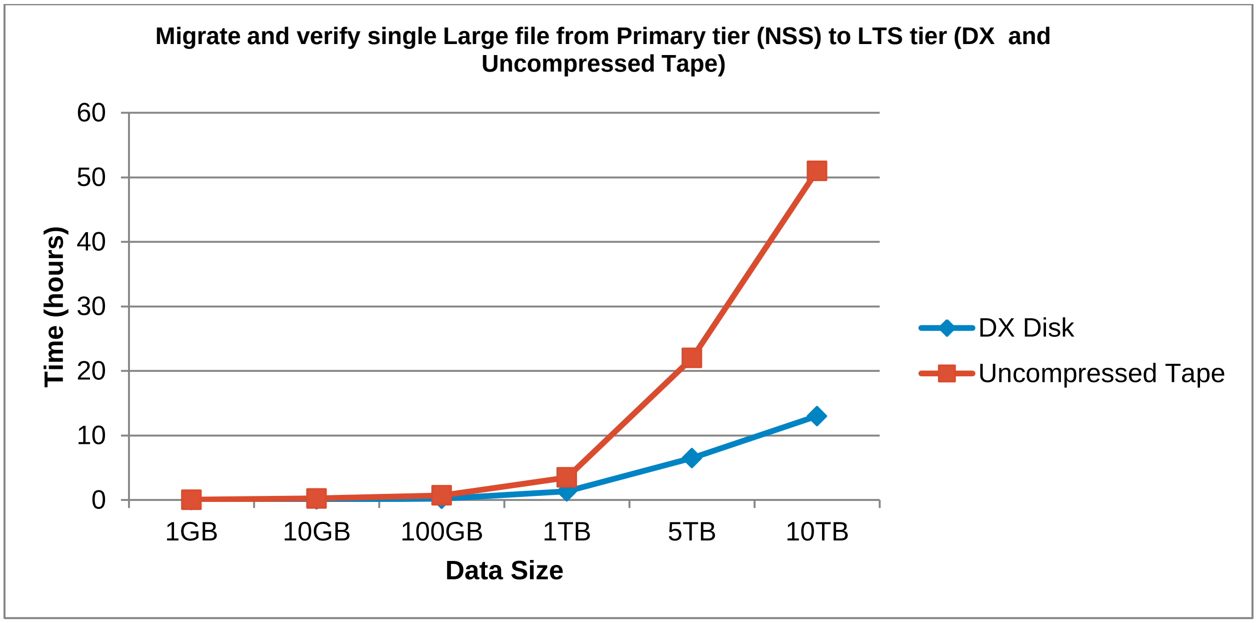
Please reference the "`Installing CommVault agent:`" and "`Configuring storage resources`" section of the HTSS whitepaper as well as your CommVault documentation for detailed install and configuration procedures and examples. [3]

Performance

This section describes the performance results of job process times for HSM migration and persistent data recall events. The test methodology utilizes data sets of various sizes as well as the number of file(s) in job, to illustrate the effect of overall data set size vs. the number of files within a migration job. The test is conducted using the DX Object Storage Platform as the disk based LTS tier. The same test is then repeated using the PowerVault ML6000 Tape Library as the tape based LTS tier. Results are then charted for comparison. In each of the migration and recall jobs, a storage policy was used to trigger job start and was also configured to perform data verification to assure data integrity of migrated data. [3]

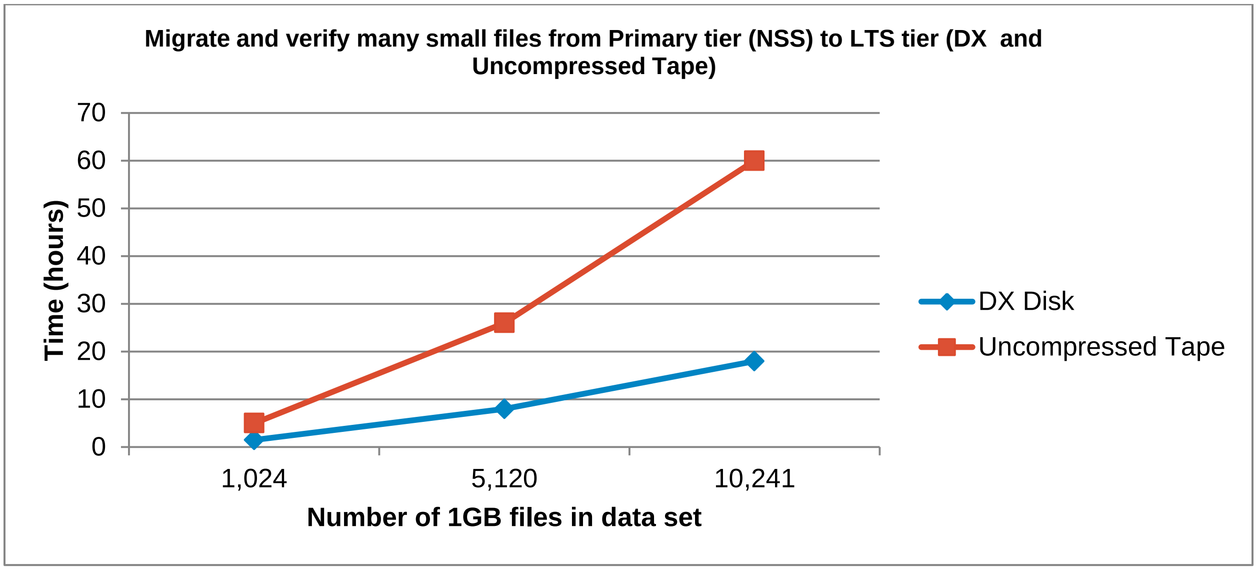
The first test involves a single file of various sizes to measure how long it took to migrate the data to the LTS and leave behind a file stub as well as verify the transaction.

Single large file



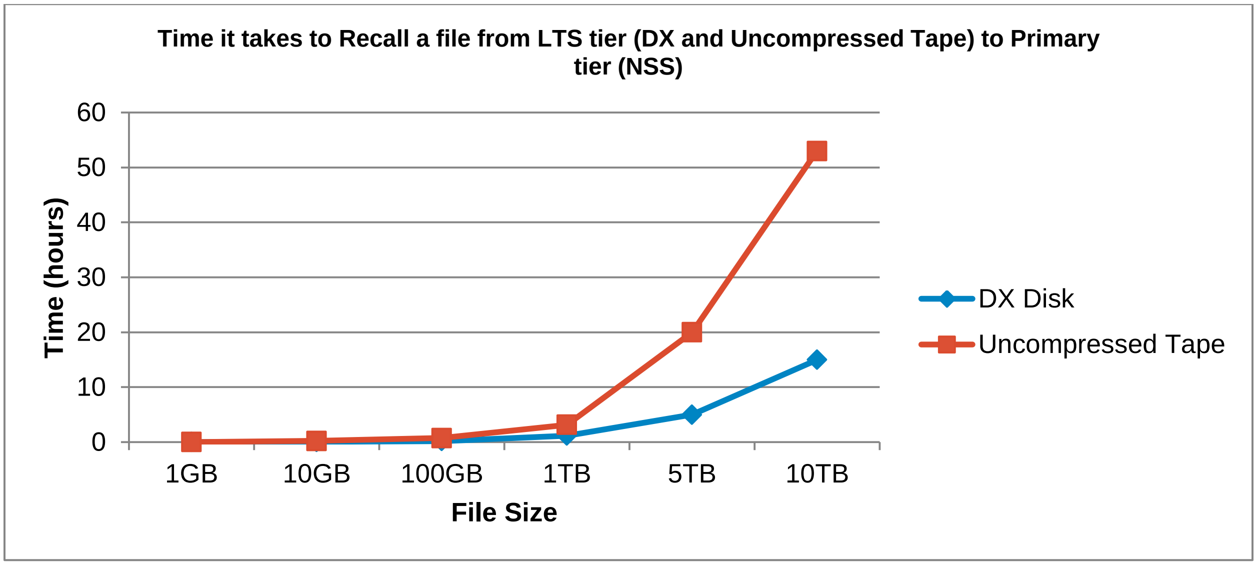
The second test uses a large number of files in an HSM migration for a larger total data size.

Many small files for large total size data set



The third test that was done is actually the inverse of the first test. That is, it measures the amount of time to recall the data from the LTS to the primary storage.

Single file recall



Best practices for using Dell DX Object Storage in HTSS

DX Object Storage Cluster

The DX Object Storage cluster consists of at least one DX6000CSN (Cluster Services Node) and two or more Storage Nodes (ex. DX6012s). Data i/o and info requests are handled between these nodes through UDP and IP multicast protocol. It is strongly advised to have these nodes on a private switch or vlan from public nodes/traffic as this could have significant impact on your overall network performance.

Switches

- Disable link aggregation configuration. DX Object Storage Nodes bond the system NIC ports in balanced-alb modes.
- If using jumbo frames (more than 1500 bytes payload), you need to increase the networkMTU (maximum transmission unit) parameter in the **cluster.cfg** file to be the same as the jumbo frames default payload (9000 bytes). The **cluster.cfg** file is located at **/var/opt/caringo/netboot/content/cluster.cfg**.
- Do not use "super jumbo" frames. If the networkMTU exceeds the default value of jumbo frames (typically 9,000 bytes payload), the neworkMTU is the value that will be used. In general, networkMTU should not exceed the jumbo frames default, as it typically degrades performance.
- Disable storm control. Storm control monitors the incoming broadcast traffic and/or unknown unicast traffic to compare it with the level that you specify. If it exceeds the specified level, packets for the controlled traffic types are dropped. This affects transmissions to the primary access storage node.
- Consider trunking ports to aggregate throughput and increase performance.
- If the cluster uses multiple switches, disable spanning tree protocol if the switches are not trunked. If switches are trunked, enable spanning tree protocol and port fast on the data-intensive ports.
- Disable Flow Control and similar controls for quality of service or traffic shaping. Switches with larger buffers, typically are better utilized by having Flow Control disabled. On switches with less buffer, enabling Flow Control helps prevent packets from being dropped and the resulting latency. However, using Jumbo Frames and increasing the networkMTU (see above) should compensate for the smaller buffer and not using Flow Control.
- Disable IGMP snooping only on the VLAN or dedicated switch that contains the cluster's multicast traffic. If IGMP cannot be disabled on that VLAN, then it must be disabled for the entire switch.

Conclusion

The Dell HPC Tiered Storage Solution with DX Object Storage is available with deployment services and full hardware and software support from Dell and CommVault Systems, Inc. This document provides a recipe on architecting, deploying, and tuning such a solution. The guidelines include hardware and software information along with configuration steps and best practices to make it easy to deploy and manage.

References

1) Dell NFS Storage Solution for HPC (NSS)

http://content.dell.com/us/en/enterprise/d/hpcc/storage-dell-nss

2) CommVault Simpana 9 Data & Information Management Software

   http://www.commvault.com/simpana.html

3) CommVault Simpana 9 Books Online

   http://documentation.commvault.com/commvault/release_9_0_0/books_online_1/default.htm

4) Dell PowerVault DL2200 System Documentation

   http://support.dell.com/support/edocs/stor-sys/pvdl2200/en/index.htm

5) Dell DX Object Storage Platform manuals

   http://support.dell.com/support/edocs/systems/DX6000/en/index.htm

Tools used to test and generate data sets

IOzone

The IOzone tool was used to generate data sets used in verification. It was also used to establish a baseline with the NSS/NSS-HA configurations. It measure sequential read and write throughput (MB/sec) as well as random read and write I/O operations per second (IOPS). You can download IOzone http://www.iozone.org. Version 3.353 was used for these tests and installed on both the NFS servers and compute nodes used.

dd

The Linux dd utility was used to generate data sets used in verification as well as configuration baseline. DD is part of the coreutils package and the version native to RHEL6.1, dd (coreutils) 5.97, was used.