ELASTIC CLOUD STORAGE SOFTWARE ON DELL DSS 7000

Reference Architecture

ABSTRACT

This document provides a reference architecture overview of Dell EMC® Elastic Cloud Storage (ECS™) software running on Dell DSS 7000. ECS is a software-defined cloud-scale object storage platform that combines the cost advantages of commodity infrastructure with the reliability, availability and serviceability of traditional arrays. The Dell DSS 7000 is an ultra-dense, reliable and versatile storage platform which provides cost-effective commodity hardware for ECS object storage.

April 2017

WHITE PAPER

The information in this publication is provided “as is.” Dell Inc. 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 software described in this publication requires an applicable software license.

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

TABLE OF CONTENTS

INTRODUCTION

Dell EMC® Elastic Cloud Storage (ECS) is an enterprise grade, multiprotocol, simple and efficient object-based storage platform designed for next-generation applications and traditional workloads. It can be deployed as a turnkey storage appliance or software only solution architected to run on industry-standard hardware. ECS software solution allows customers to leverage commercial off the shelf hardware to reduce costs. There are two options available for customers who desire a software only solution:

- **Certified** - ECS software bundle running on certified industry standard hardware.
- **Custom** – ECS software running on hardware, operating system and tools outside of certified matrix.

The certified offering is targeted for customers needing small to large petabytes of object storage whereas custom is for big customer deployments requiring a significant amount in petabytes of object storage.

This whitepaper provides a reference architecture overview of ECS software bundle running certified hardware, the Dell DSS 7000. The Dell DSS 7000 is well suited for running ECS software. It is a low-cost, highly efficient, flexible, and scalable hardware platform.

AUDIENCE

This paper is intended for field personnel and customers who are interested in designing an object storage solution using the ECS software bundle with Dell DSS 7000. It provides an overview of the ECS software bundle, Dell DSS 7000 hardware, and related networking and services.

SCOPE

This document does not cover installation, administration, and upgrade procedures for ECS deployed on Dell DSS 7000. Its primary focus is to provide a reference architecture overview and value of deploying ECS software on industry standard hardware such as the Dell DSS 7000. Links to other related documents are provided under References section.

Updates to this document are done periodically and coincides usually with a major release or new feature and functionality change. To get the latest version of this document, please download from this link.

VALUE OF ECS ON INDUSTRY STANDARD HARDWARE

ECS Software on industry standard hardware provides several advantages and options for customers. By utilizing commercial off the shelf hardware, capital expense can be significantly reduced. It also prevents vendor lock-in allowing customers to re-purpose hardware or move to a different software defined storage vendor if required. Furthermore, it enables customers to build homogenous datacenter infrastructure with unified commodity hardware. ECS software on industry standard hardware are primarily for large enterprises and high growth potential verticals such as service providers, telecom, life sciences and others whose main uses include global content repository, web, IoT, and data analytics.

REFERENCE ARCHITECTURE OVERVIEW

The ECS software bundle running on Dell DSS 7000 has been verified and certified by Dell EMC ECS quality engineering team. It has endured the same types of rigorous testing done for the ECS Appliance. This section will describe the main components of this reference architecture which include:

- ECS Software bundled with SUSE Linux Enterprise Server 12 Service Pack 1 (SLES 12 SP1), Docker, Java Virtual Machine (JVM) and tools.
- Dell DSS 7000
- Network Switches for data and management

In order to utilize ECS software on certified Dell DSS 7000, there are minimum requirements defined Hardware and Deployment sections.

ECS SOFTWARE BUNDLE

Version 2.2.1 HF 1 is the minimum version of ECS software supported to deploy on industry standard hardware. Bundled with ECS software are the tools to verify the health of system and assist in configuration; and the infrastructure required to run ECS which include SLES 12 SP1, Docker, and JVM. Incorporating the tools and infrastructure with ECS software simplifies the install and configuration. This section will provide a quick overview on the contents of this bundle and their functions. For more in-depth details on ECS Architecture, refer to the ECS Architecture Whitepaper.

ECS SOFTWARE

ECS software was designed as a layered architecture with each layer having distinct roles and performing specific tasks or services. The software sits on top an infrastructure and hardware platform as shown in Figure 1. The layers in ECS Software include:

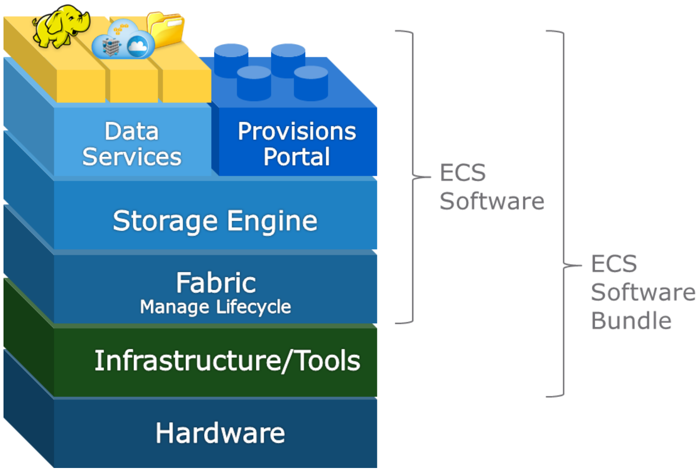
- **ECS Portal and Provisioning Services** – provides a Web-based portal for self-service, automation, reporting and management of ECS nodes. It also handles licensing, authentication, multi-tenancy, and provisioning services. A command-line interface and a set of management REST APIs are also available to manage and provision ECS.

- **Data Services** – provides services, tools and APIs to support Object (S3, Swift, CAS, Atmos), and HDFS and NFSv3 protocols.

- **Storage Engine** – responsible for storing and retrieving data, managing transactions, and protecting and replicating data. The storage engine services include:
    - **Resource service** – stores info like user, namespace, bucket, etc.
    - **Transaction service** – parses object request and reads and writes data to chunk.
    - **Index Service** – conducts file-name/data-range to chunk mapping and handles secondary indices.
    - **Chunk Management Services** – responsible for chunk information and does per chunk operations.
    - **Storage Server Management Service** – monitors the storage server and disks and re-protection of data during hardware failures.
    - **Partition Record Service** – records owner node of partition, Btree and journals.
    - **Storage Server Service (Chunk I/O) –** directs I/O operations to the disks.

- **Fabric** – provides clustering, health, software and configuration management as well as upgrade capabilities and alerting. The fabric layer has the following components to manage the overall system:
    - **Node agent** – manages host resources (disks, network, containers, etc) and system processes.
    - **Lifecyle manager** – application lifecycle management responsible for starting services, recovery, notification, and failure detection.
    - **Persistence Manager –** provides the coordination and synchronization of ECS distributed environment.
    - **Registry –** stores all the container images for ECS.
    - **Event Library** – holds the set of events occurring on the system.
    - **Hardware Manager (HWMgr)** – provides status, event information and provisioning of the hardware layer to higher level services. These services have been integrated to the Fabric Agent to support industry standard hardware.

Figure 1 - High Level ECS Architecture



INFRASTRUCTURE

Bundled with ECS Software is the infrastructure components which include SUSE Linux Enterprise 12 SP1 (SLES 12 SP1) with Docker and JVM. ECS software is a java application running within several Docker containers on top an operating system. Thus, each of the nodes used for ECS is installed with a SLES 12 SP1 with Docker and JVM and each container is responsible for running a specific task. The name of the containers running and purpose are as follow:

- **object-main** – contains the resources and processes relating to the data service, storage engine, portal and provisioning services. Runs on every node in ECS.

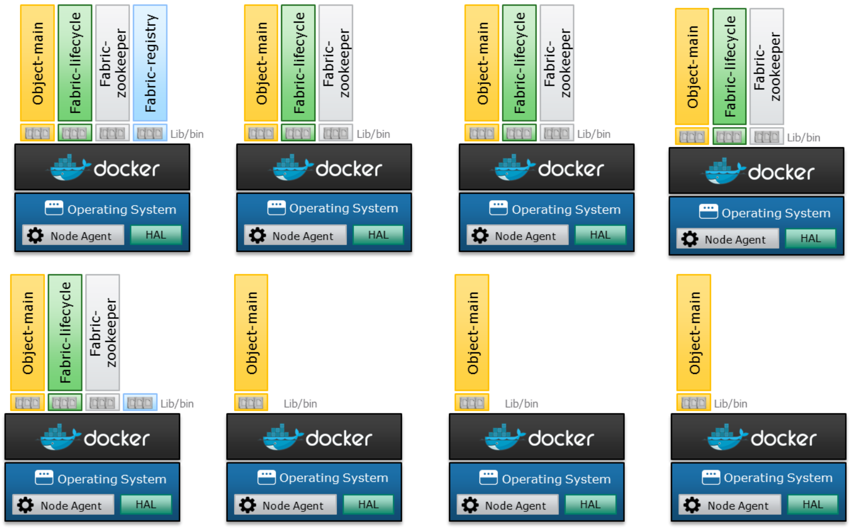
- **fabric-lifecycle** – contains the processes, information and resources required for the monitoring, configuration management and health management of the system. Depending on the number of nodes in the system, there will be an odd number of fabric-lifecycle instances running. For example, there will be three instances running on a four-node system and five instances for an eight-node system.

- **fabric-zookeeper** – centralized service for coordination and synchronization of distributed processes, configuration information, groups and naming services. It is referred to as the persistence manager and runs on odd number of nodes depending on number of nodes deployed within ECS system.

- **fabric-registry** – location or registry of the ECS images. Only one instance of this is running per ECS system.

Some of the containers listed above run on all nodes and some run on odd number of the nodes. The containers are lightweight and consist of only the runtime, system tools and libraries required to run ECS. Docker containers are individual processes sharing the same operating system and hardware resources. Figure 2 provides an example of how ECS can be deployed on an eight node system.

Figure 2 - ECS Docker containers on 8 node system example



Filesystem Layout

Each node has its own set of commodity disks. Each disk is formatted and mounted as an XFS filesystem and has a unique identifier (UUID). Data are stored in 128MB chunks and the chunks are stored in files within the filesystem. Each disk partition or filesystem is filled with files that are 10GB in size. These files are created during installation such that contiguous blocks are allocated for the files. The number of files within each disk depends on the disk size. For instance, a 1TB disk will have 100, 10GB files. The names of the files inside the disk are “0000”, “0001”, “0002”, and so forth.

TOOLS

Existing tools and libraries were enhanced and new tools were created to support ECS software deployed on industry standard hardware. These tools assist in verifying the hardware configuration, providing health and status of hardware and ECS software services, and utilities that interact with the hardware. The tools and libraries are also packaged with the ECS Software.

Fabric Command-Line Interface (FCLI)

The Fabric layer of ECS software has been improved to facilitate support for industry standard hardware. A Fabric command-line interface (fcli) is available to communicate with the node agent in the Fabric layer to diagnose issues and validate the state of the underlying hardware system. For example, Figure 3 below is an output of *" fcli disks list"* which shows cluster-wide disk allocation summary. Another example, Figure 4, provides information on agent health.

Figure 3 – Snippet Output - FCLI example – Custer Wide Disk Allocation Summary

```
admin@dallas-flamingo:~> /opt/emc/caspian/fabric/cli/bin/fcli disks list

AGENT ID                              HOSTNAME       SERVICE            DISK TYPE       GOOD  BAD SUSPECT UNKNOWN
db2eff8e-e461-4daf-9a39-baea7100a7df  dallas-flamingoobject-main         6001GB HDD        42    0       0       0
c09c3a54-7738-48d0-bc27-7a99192646a0  detroit-flamingoobject-main         6001GB HDD        43    0       0       0
46d82745-551f-459e-89ff-d0001c788c31  columbus-flamingoobject-main         6001GB HDD        43    0       0       0
a15b8644-ca68-43b3-8cd6-27f3ee6f6da7  austin-flamingoobject-main         6001GB HDD        42    0       0       0
```

Figure 4 - FCLI Example - Agent Health

```
[root@provo-beige emc]# fcli agent service.health --application object --role main
{
  "health": "GOOD",
  "status": "OK",
  "etag": 1944
}
```

For more information on fcli commands and options, refer to the Health and Troubleshooting Guide and ECS Installation guide. There is also some help information within the command itself by issuing the *"fcli –help*" or for specific help on certain sub-commands*, "fcli <sub-command> -help" (ie. fcli disks –help).*

Hardware Abstraction Layer (HAL)

HAL is a library with front-end tools used by services and utilities to interact with the hardware level. The services within ECS interact with the hardware using this library. Tools part of HAL includes *"cs_hal"* and *"hal_conf".* These tools are very useful in identifying state of the hardware. For instance, it contains information on disks and provides health of the hardware. Examples of output from these commands on the Dell DSS 7000 are displayed in the Figures 5, 6, and 7 below. More information on these commands is available in the Installation Guide and Health and Troubleshooting Guide available in ECS Product Documentation.

Figure 5 – Snippet Output - Diagnostic Example of Single Node Disk Health

```
admin@dallas-flamingo:~> /opt/emc/hal/bin/cs_hal list disks
Disks(s):
SCSI Device Block Device Enclosure   Partition Name                      Slot Serial Number       SMART   DiskSet
----------- ------------ ----------- ----------------------------------- ---- ------------------- ------- --------
----
/dev/sg0    /dev/sda     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   11   S4D0PXK1            GOOD
/dev/sg1    /dev/sdb     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   12   S4D0PWZX            GOOD
/dev/sg2    /dev/sdc     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   13   S4D0PTPN            GOOD
/dev/sg3    /dev/sdd     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   14   S4D0JYB7            GOOD
/dev/sg4    /dev/sde     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   15   S4D0E6SE            GOOD
/dev/sg5    /dev/sdf     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   16   S4D0PS70            GOOD
/dev/sg6    /dev/sdg     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   17   S4D0PXA4            GOOD
/dev/sg7    /dev/sdh     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   18   S4D0PXJ0            GOOD
/dev/sg8    /dev/sdi     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   19   S4D0LG75            GOOD
/dev/sg9    /dev/sdj     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   20   S4D0PTCE            GOOD
/dev/sg10   /dev/sdk     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   21   S4D0JS5Q            GOOD
/dev/sg11   /dev/sdl     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   22   S4D0JQJ0            GOOD
/dev/sg12   /dev/sdm     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   23   S4D0JHLQ            GOOD
/dev/sg13   /dev/sdn     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   24   S4D0JH38            GOOD
/dev/sg14   /dev/sdo     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   25   S4D0H7B0            GOOD
/dev/sg15   /dev/sdp     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   26   S4D0JSHM            GOOD
/dev/sg16   /dev/sdq     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   27   S4D0PTL9            GOOD
/dev/sg17   /dev/sdr     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   28   S4D0AXL8            GOOD
/dev/sg18   /dev/sds     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   29   S4D0PWXT            GOOD
/dev/sg19   /dev/sdt     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   30   S4D0PX8S            GOOD
/dev/sg20   /dev/sdu     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   31   S4D0JHGJ            GOOD
/dev/sg21   /dev/sdv     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   32   S4D0PXQF            GOOD
/dev/sg22   /dev/sdw     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   33   S4D0PTEV            GOOD
/dev/sg23   /dev/sdx     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   34   S4D0K01S            GOOD
/dev/sg24   /dev/sdy     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   35   S4D0PSAL            GOOD
/dev/sg25   /dev/sdz     internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   36   S4D0LCN9            GOOD
/dev/sg26   /dev/sdaa    internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   37   S4D0L95S            GOOD
/dev/sg27   /dev/sdab    internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   38   S4D0PWW2            GOOD
/dev/sg28   /dev/sdac    internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   39   S4D0PRYG            GOOD
/dev/sg29   /dev/sdad    internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   40   S4D0PWRA            GOOD
/dev/sg30   /dev/sdae    internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   41   S4D0PTS6            GOOD
/dev/sg31   /dev/sdaf    internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   42   S4D0GAPQ            GOOD
/dev/sg32   /dev/sdag    internal    ECS:object:JTy7pHm2S1uF5CIjA2kMSA   43   S4D0H70V            GOOD
```

Figure 6 – Diagnostic Example - Single Disk Details

```
admin@dallas-flamingo:~> /opt/emc/hal/bin/cs_hal info sg30
SCSI disk       : /dev/sg30
block device    : /dev/sdae
size (via SCSI) : 5589.03 GB
size (via blk)  : 5589.03 GB
vendor          : ATA
model           : ST6000NM0024-1HT
firmware        : MA2E
SCSI id         : 0:0:41:0
S/N             : S4D0PTS6
state           : awake and running
RAID            : no
internal        : yes
system disk     : no
VM disk         : no
type            : rotational
volume count    : 1
volume          : /dev/sdae1
volume size     : 5589.03 GB
partition table : gpt
partition       : ffbf5513-48f4-42f4-af91-137cb9c0ad69
partition type  : ffbf5513-48f4-42f4-af91-137cb9c0ad69
filesystem      : ffbf5513-48f4-42f4-af91-137cb9c0ad69 (xfs; not mounted)
slot name       : 41
LED             : internal drive; no LED
SMART           : GOOD
```

Figure 7- Diagnostic Example - Single Node Details

```
admin@dallas-flamingo:~> /opt/emc/hal/bin/cs_hal info node
Node                : dallas-flamingo
BIOS date           : 04/11/2016
Bios vendor         : Dell Inc.
BIOS version        : 2.0.1
Board model         : 0RRXVV
Board S/N           : .F1ZVHB2.CN779215AC00DV.
Board vendor        : Dell Inc.
Board version       : A04
Chassis S/N         : F1ZVHB2
Chassis vendor      : Dell Inc.
Chassis model       : Dell Storage SD7000-S
System S/N          : F1ZVHB2
System vendor       : Dell Inc.
Processor count     : 40
Total memory        : 61.6869GB
Availble memory     : 51.9544GB
Total swap          : 2GB
Available swap      : 1.35392GB
Shared memory       : 0GB
Host adapter count  : 3
Net interface count : 9
Enclosure count     : 0
External disk count : 0
LED state           : not supported
```

Precheck Tool

The precheck tool verifies if the industry standard hardware meets the minimum requirements to deploy ECS software. After the operating system is installed and network configured on each node, the precheck tool is run before install of ECS Software.   In general this tool has the following capabilities:

- Collect – collect HAL inventory files from each node.
- Deploy - load utilities image on all nodes.
- Inventory – collects HAL inventory files from all nodes.
- Match – match default HAL template files to HAL inventory files.
- Precheck – run compatibility checks on specific hardware.
- Report – reports the results of prechecks.
- Topology – prepares topology file required for Fabric installer.
- Cleanup – removes utilities image on each node.

This tool will also be used to generate a HAL template (an XML file) which specifies the pieces of hardware such as disks, expander ports, etc. as attributes with values. The template simplifies HAL deployment in a big cluster environment where all nodes (groups of nodes) have similar configuration. A snippet of the HAL template is illustrated below.

Figure 8 - Snippet of HAL Template Example

```
admin@dallas-flamingo:/opt/emc/hal/etc> more hal.xml
<?xml version="1.0" encoding="UTF-8" standalone="no" ?><hal xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xsi:n
oNamespaceSchemaLocation="hal_config.xsd">
  <version>2.0</version>
  <description>Dell EMC ECS generic template</description>
  <node>
    <console port="ttyS0" speed="57600"/>
    <hbas>
      <hba device_class="0x106" device_id="0x8d62" location="internal" pci_address="0000:00:11.4"
vendor_id="0x8086"/>
      <hba device_class="0x106" device_id="0x8d02" location="internal" pci_address="0000:00:1f.2"
vendor_id="0x8086">
        <disks homogeneous="true" pid="INTEL SSDSC2BX80" rotational="false" vid="ATA">
          <disk configure="false" hasPartitions="true" sn="BTHC617301TZ800NGN" system="true"
wwn="55cd2e404c21dc6e">
            <partition configure="false" uuid="146713ac-01"/>
          </disk>
        </disks>
      </hba>
      <hba device_class="0x104" device_id="0x5d" location="internal" pci_address="0000:07:00.0"
vendor_id="0x1000">
        <disks homogeneous="true" pid="ST6000NM0024-1HT" rotational="true" vid="ATA">
          <disk configure="true" hasPartitions="true" sn="S4D0PXK1" system="false" wwn="5000c5008cd64fe8">
            <partition configure="true" uuid="185b4c18-8e46-4099-b7ad-bd96c050f1f4"/>
          </disk>
          <disk configure="true" hasPartitions="true" sn="S4D0PWZX" system="false" wwn="5000c5008cd656d4">
            <partition configure="true" uuid="85e35247-ff41-49c1-bb25-1032ffff8166"/>
          </disk>

         ..........

         ..........

        </disks>
      </hba>
    </hbas>
    <nics>
      <nic device_class="0x200" device_id="0x165f" ifname="private" label="" pci_address="0000:02:00.0"
speed="1Gb" ven
dor_id="0x14e4"/>
      <nic device_class="0x200" device_id="0x165f" ifname="unused-0" label="" pci_address="0000:02:00.1"
speed="1Gb" ve
ndor_id="0x14e4"/>

         ..........

         .........
</nics>
    <net_interfaces>
      <net_interface configure="false" is_private="false" is_public="false" name="private"
subnet="192.168.219.13/24"/>
      <net_interface configure="false" is_private="false" is_public="false" name="unused-0"/>
      <net_interface configure="false" is_private="false" is_public="false" name="unused-1"/>
</net_interfaces>
  <settings>
      <excludeInternalDrivesIfDaePresent>true</excludeInternalDrivesIfDaePresent>
    </settings></node>
</hal>
```

DELL DSS 7000 HARDWARE CERTIFIED FOR ECS

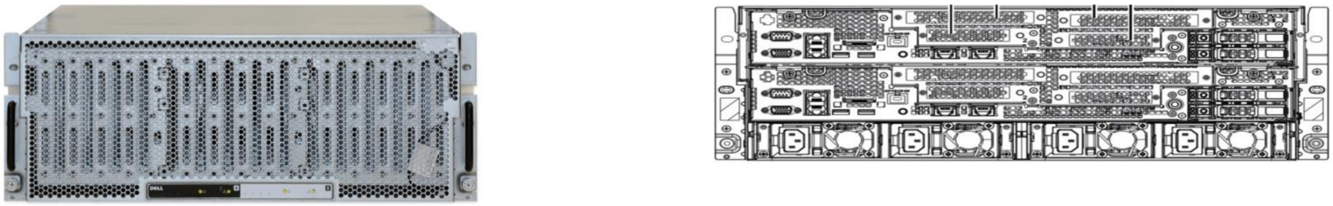
The Dell DSS 7000 is known for its ability to scale out and support exabytes of data. It is designed to be ultra-dense, highly efficient, and reliable. The performance and scale-out capabilities of the Dell DSS 7000 make it an ideal hardware platform for running ECS software. Dell EMC has validated the DSS 7000 with specific components for ECS Software. The technical specifications of the Dell DSS 7000 supported and that have been validated are listed in Table 1 below. Figure 9 shows a sample view of the Dell DSS 7000.

Table 1 - Technical Specifications of DSS 7000 Validated for ECS Software

Category	DSS 7000 Supported Technical Details	Requirements for ECS Software	Recommendations
Chassis	4 U Chassis (can hold up to two nodes)	Minimum is 5 nodes **NOTE: For DSS 7000, the minimum is 6 nodes.	2 DSS 7500 Nodes Per DSS 7000 chassis when expanding
Processors	2 x Intel® Xeon® E5-2600 v4 product family - Processor per node	Minimum 8-core per node (i.e., could be a 1 CPU 8-core or 2 CPUs 4-core)	Recommend minimum 2 CPUs per node. Each CPU has separate memory channels and having a minimum of 2 processors allows for additional memory bandwidth compared to just 1 CPU of 12-core for instance.
Memory	12 DIMMs per node:16GB/32GB DDR4 RDIMM	Minimum 64 GB per node.	
Storage Controllers	12Gb SAS and SATA RAID controller	All drives configured in JBOD mode	
Boot Volume (SSD)	2 x 2.5" hot swappable boot drives per node (rear) Currently only supports 120GB SSD due to current availability.	Minimum 400 GB SSD per node **NOTE: For DSS 7000, the minimum is 120GB SSD due to current availability.	
HDD (data)	Up to 90 x 3.5 inch hot swappable drives (45 drives/node) Supported HDD Drive Capacities: 6TB, 8TB, 10TB Requires a minimum of 30 drives per node.	Any type of 3.5" HDD supported for the DSS7000.	Recommend using the 512n and 4kN disk type since the 512e formats have additional emulation/translation which may affect performance. Also consider the following when selecting drives: • optimized dollars/GB • optimized dollars/performance
PCIe Slots	3 x8; 1x16 slots		
Embedded Networking	4 port 1Gbe Broadcom Corporation NetXtreme BCM5720 Gigabit Ethernet PCIe (4 x 1GB Ethernet LOM per) node	1 GbE management	
Network	Intel X520 10GbE Dual Port SFP+	10 GbE data (dual ports)	
Power Supply	2 x 1100W AC hot-plug redundant power supplies per node		
System Management	IPMI 2.0, BMC with vKVM support and 1 x 1 GbE management port		
Monitoring		Integrate with Dell EMC ESRS for remote debugging/monitoring	

NOTE: For compatibility of Dell DSS 7000 with ECS software, anything outside of this technical specification is not supported.

Figure 9 – Sample Front and Rear Views of Dell DSS 7000



For more information on the Dell DSS 7000, refer to the Dell website: https://www.dell.com/en-us/work/learn/extreme-scale-infrastructure and http://www.dell.com/support/home/us/en/19/product-support/product/dell-dss7000/manuals . You can also contact Dell personnel at ESI@dell.com for more information.

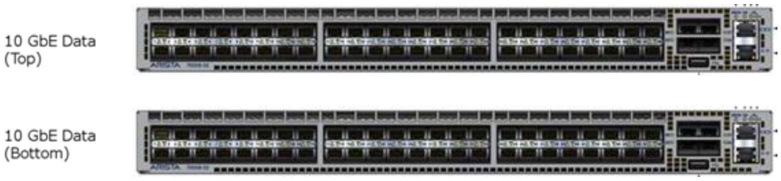
NETWORK SWITCHES

The same set of Arista switches used for the ECS Appliance will be used with the industry standard certified hardware. For more detailed information on the switches supported, please refer to the ECS Hardware and Cabling Guide. **Note:** Using a different switch for either the data and/or management switch would be considered a "custom" configuration and would need to be processed via ASD Helpdesk as a custom configuration. Two 10 GbE switches are required for data transfer and one 1 GbE switch for management.

10 GbE Switch – Data

A dual 10 GbE, 24-port or 52-port Arista switches are used for data transfer to and from customer applications as well as for internal node-to-node communication. These switches are connected to the ECS nodes in the same rack. For the two switches utilized, the Multi-Chassis Link Aggregation (MLAG) feature is employed to logically link the switches and enable active-active paths between the nodes and customer applications. This configuration results in higher bandwidth while preserving resiliency and redundancy in the data path. Any networking device supporting static LAG or IEEE 802.3ad LACP can connect to this MLAG switch pair. Finally, because the switches are configured as MLAG, these two switches appear and act as one large switch. Figure 10 displays an example of the front view of these two switches.

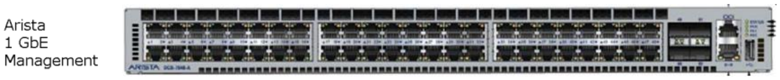
Figure 10 - Example of Arista 10 GbE Switches



1 GbE Switch - Management

The 52-port 1Gb Arista switch is used by ECS for node management and out-of-band management communication between the customer's network and the Remote Management Module (RMM) ports of the individual nodes. The main purpose of this switch is for remote management and console, install manager (PXE booting), and enables rack management and cluster wide management and provisioning. Figure 11 shows a front view of this management switch.

Figure 11 - Example of Arista 1 GbE Switch



In addition to Arista, there is now support for Cisco 52 port 1 GbE switch for management. This switch is meant to support customers who have strict Cisco only requirements. It is available only for new racks and is not supported to replace Arista management switches in existing racks. Configuration file will be pre-loaded in manufacturing and will still be under control of Dell EMC personnel. ECS 3.0 is the minimum to support the Cisco management switch; however, patches are required to be installed until ECS 3.1 is released. Due to the patch requirements for ECS 3.0, racks shipped from manufacturing with Cisco switches will not have the operating system pre-loaded on the nodes.   Figure 12 shows front view of a Cisco 1 GbE management switch**.**

Figure 12 - Example of Cisco 1 GbE Switch

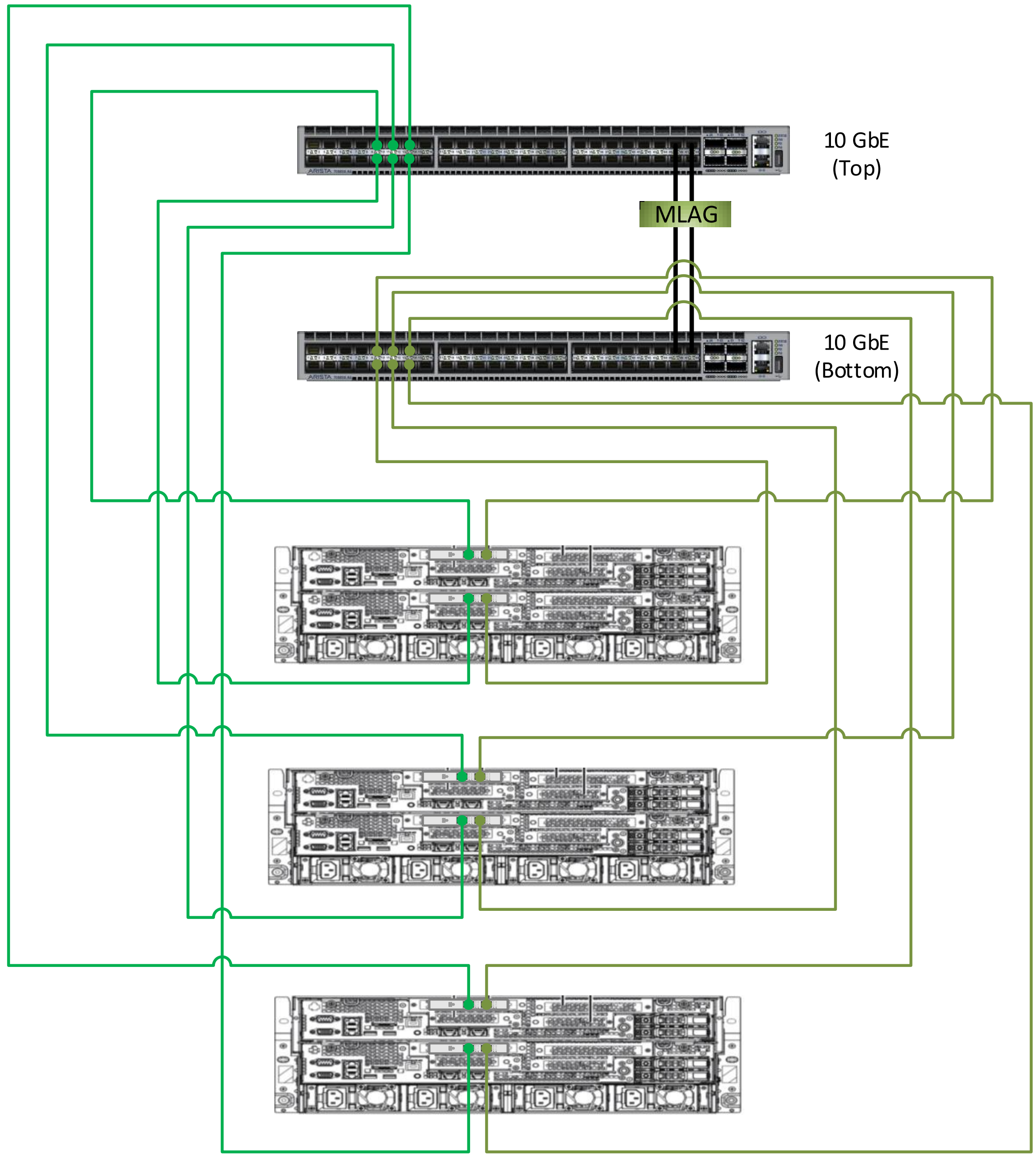
Node Network Connectivity

Each node has two 10 GbE ports, which appear to the outside world as one port via NIC bonding. Each 10 GbE node port connects to one port in the 10 GbE data switch pair . The 10 GbE data switches in each rack will be connected to a customer-provided switch or backplane. Thus the data traffic will flow through the 10 GbE network. These public ports on the ECS nodes get their IP addresses from the customer's network, either statically or via a DHCP server. Customer applications connect to ECS by using the 10 GbE public IP addresses of the ECS nodes.

The 1 GbE management port on each node connects to an appropriate port in the 1 GbE management switch and has a private address of 192.168.219.X. Each physical unit also has a connection between its RMM port (1 GbE on the Dell DSS 7000) and a port in the 1 GbE switch, which in turn has access to a customer's network to provide out-of-band management of the servers. To enable access for the RMM ports to the customer's network, ports 51 and/or 52 of 1 GbE management switch are linked to the customer's network directly. The RMM port is used by Dell EMC field service personnel for monitoring, troubleshooting and installation. You can expand an ECS rack by linking one or more racks to an existing rack also via ports 51 and 52 of the management switch. The 1 GbE management switches in the racks are used for serially connecting the racks.

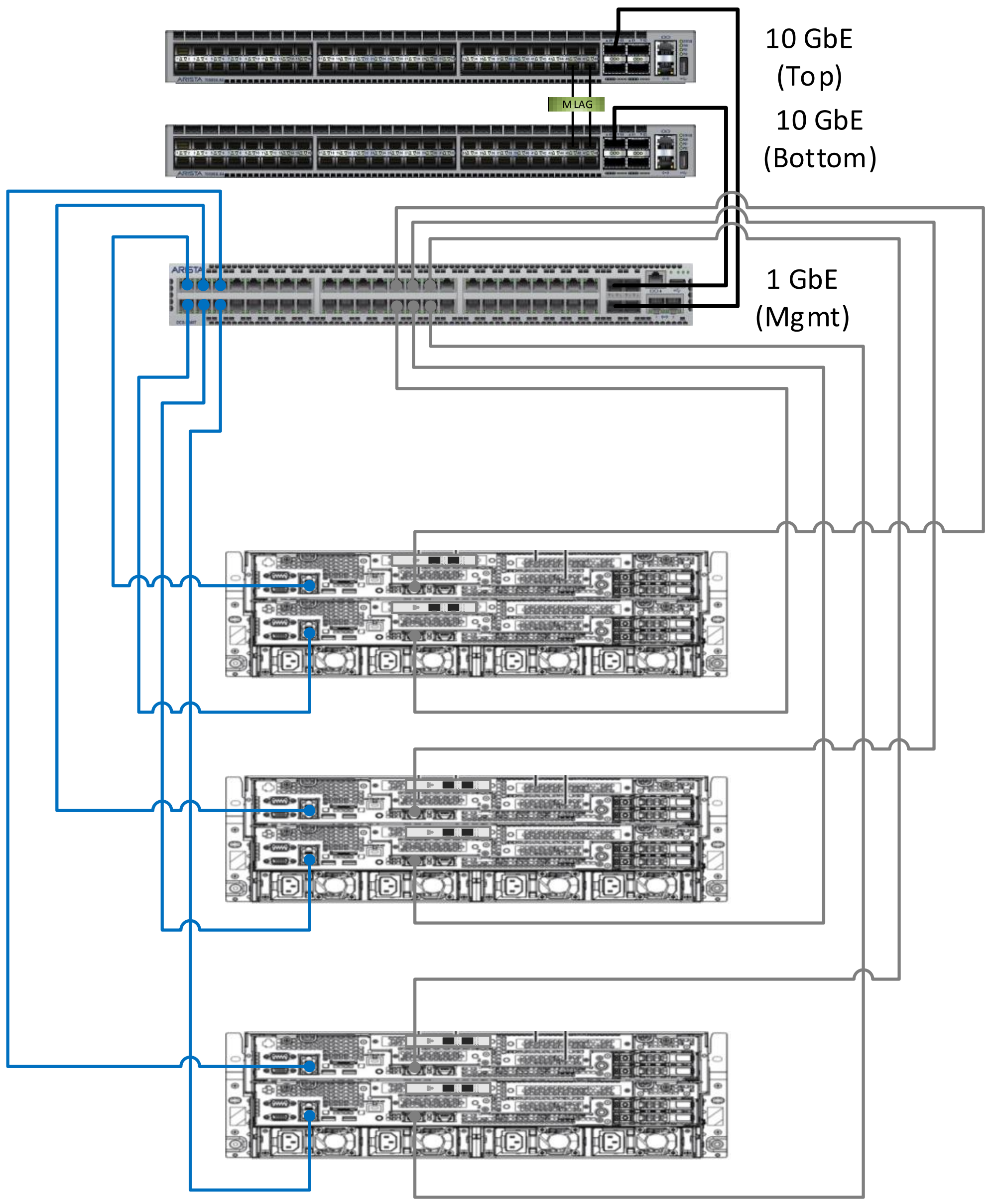
Figures 13 and 14 depict the 10 GbE and 1 GbE network cabling with 4 (2-Node) Dell DSS 7000. The example of the 1 GbE network below illustrates the connections of the out of band management ports (1 GbE port – NIC 3) on each node and management 1 GbE ports (NIC 2) on each node connected to the management switch. For more detailed information on the switches supported, please refer to the ECS Hardware and Cabling Guide. As a best practice, when physically connecting nodes to the switches, do so in an ordered and sequential fashion. For instance on the management switch, node 1 should connect to port 1, node 2 to port 2 and so on. For additional details on ECS networking, refer to the ECS Networking and Best Practices whitepaper.

Figure 13 - Example of 10 GbE Network Cabling for 3 (2-Node) Dell DSS 7000



Three (2-Node) Dell DSS7000

Figure 14 - Example of 1 GbE Network Cabling for 3 (2-Node) Dell DSS 7000



DEPLOYMENT

ECS can be deployed on a single or multi-site (replicated) configuration. In a multi-site replicated environment, different hardware platform can be used in the other replicated sites. For instance, industry standard hardware such as the Dell DSS 7000 can be in one site and the replicated site can be an ECS Appliance. As previously mentioned, there are minimum configurations required to run ECS Software on industry standard hardware. Once the hardware is selected and setup, ECS software functions in the same fashion as if it was running on an ECS appliance. This section discusses the configurations to install ECS Software bundle on the Dell DSS 7000. It will also discuss the customer provided infrastructure needed.

CONFIGURATIONS

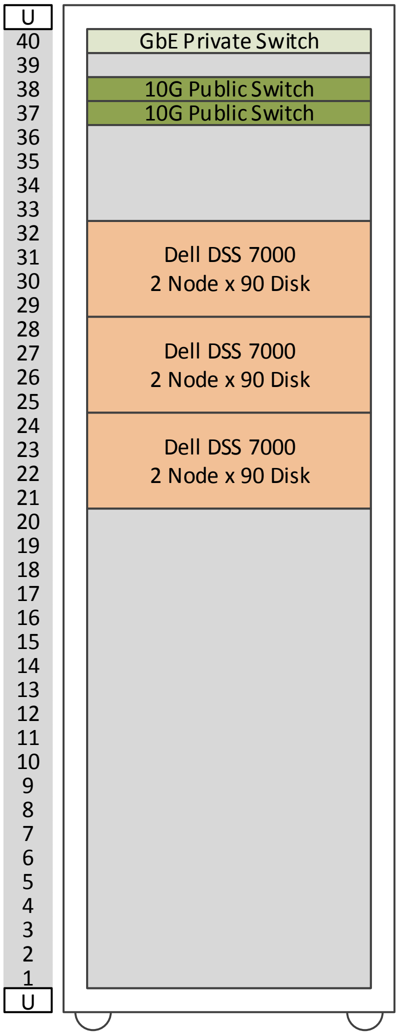
There are flexible entry points with rapid scalability to petabytes and exabytes of data. ECS scales linearly both in capacity and performance by just adding additional nodes with disks to your environment – with minimal impact to the business. The basic hardware required to run ECS software on DSS 7000 certified industry standard hardware include at minimum six nodes with data disks, two 10

Gbe switches for data, and a single 1 Gbe management switch. **NOTE**: As mentioned previously in the Network Switches section, the switches supported for the certified hardware is the Arista switches for data and Arista or Cisco for management. Table 2 highlights the minimum and maximum hardware components and capacities for Dell DSS 7000. The capacity calculations for minimum use the 6 TB and 10 TB for maximum. This table is derived based on supported hardware component defined for DSS 7000. Figure 15 provides a sample rack configuration for Dell DSS 7000 with switches.

Table 2 - Minimum and Maximum Hardware Specifications for Dell DSS 7000

SPECIFICATION	MINIMUM	MAXIMUM
DSS 7500 Nodes per cluster	6	No Known Limit
Disks per DSS 7500 node	30	45
Disk Capacities available for DSS 7500	6 TB	10 TB
Raw Capacity per node	180 TB (Using 6 TB x 30 disk for calculation)	450 TB (Using 10 TB x 45 disk for calculation)
Raw Capacity per rack (assumes 9 DSS 7000 that can be housed in 40U rack with the switches)	3.2 PB	8.1 PB
Memory per node	64 GB	Up to 12 DIMM Slots of size 16 GB or 32 GB DDR RDIMMs can be configured
Cores per node	8-core	18-core

Figure 15 – Sample Rack Configuration for Dell DSS 7000



BUILD OF MATERIALS (BOM)

Table 3 below provides a sample BOM for a ***single*** DSS 7000 enclosure with 2 DSS 7500 nodes (45x8TB disks per node) with no rack for reference. **Note:** The nodes inside the DSS 7000 chassis is referred to as the DSS 7500. Based on your requirements this BOM will change in the sense of number of CPUs, number of cores, number of drives, drive capacities, amount of memory, etc. Supported network switches as described the Network Switches section are not included in this BOM and would need to be ordered as well through your Dell EMC sales representative.

Table 3 - Sample BOM for a single DSS 7000 enclosure with 2 DSS 7500 nodes (45x8TB Disks per node), no rack

Quantity	Description
1	DSS7000 Enclosure with up to 90 Hard Drives
1	DSS7500, Compute Node
1	DSS7500, Compute Node
1	DSS7000 Enclosure Chassis with up to 90 Hard Drives
1	PowerEdge DSS7500, V2 Dual Compute Node, Mgmt Port
1	PowerEdge DSS7500, V2 Dual Compute Node, Mgmt Port
1	US Order
1	US Order
1	US Order
1	Intel Xeon E5-2650 v4 2.2GHz,30M Cache,9.60GT/s QPI,Turbo,HT,12C/24T (105W) Max Mem 2400MHz
1	Intel Xeon E5-2650 v4 2.2GHz,30M Cache,9.60GT/s QPI,Turbo,HT,12C/24T (105W) Max Mem 2400MHz
1	Intel Xeon E5-2650 v4 2.2GHz,30M Cache,9.60GT/s QPI,Turbo,HT,12C/24T (105W) Max Mem 2400MHz
1	Intel Xeon E5-2650 v4 2.2GHz,30M Cache,9.60GT/s QPI,Turbo,HT,12C/24T (105W) Max Mem 2400MHz
1	DSS7000 Shipping DAO
1	Performance Optimized
1	Performance Optimized
1	DSS7/7500 DIMM Filler Blanks
1	DSS7/7500 DIMM Filler Blanks
4	16GB RDIMM, 2400MT/s, Dual Rank, x8 Data Width
4	16GB RDIMM, 2400MT/s, Dual Rank, x8 Data Width
1	2400MT/s RDIMMs
1	2400MT/s RDIMMs
1	DSS7500 Heatsink and Shroud
1	DSS7500 Heatsink and Shroud
1	Consolidated Shipping
1	Fan Test required for first node only
1	Performance BIOS Settings
1	Performance BIOS Settings
45	8TB 7.2K SATA,6Gbps,512e,Internal 3.5 inch Hard Drive
45	8TB 7.2K SATA,6Gbps,512e,Internal 3.5 inch Hard Drive
2	120GB,Solid State SATA,6Gbps,Internal 2.5 Inch, Intel S3510 Boot Hard Drive**
2	120GB,Solid State SATA,6Gbps,Internal 2.5 Inch, Intel S3510 Boot Hard Drive**
1	LSI9361-8I Controller
1	LSI9361-8I Controller
1	Redundant 1100W Power Supply
1	Redundant 1100W Power Supply
1	No Trusted Platform Module

1	No Trusted Platform Module
4	C13 to C14, PDU Style, 12 AMP, 6.5 Feet (2m) Power Cord, North America
1	Intel X520,DP,10G,SFP+ Adapter, Full Height Half Length for Slot 2
1	Intel X520,DP,10G,SFP+ Adapter, Full Height Half Length for Slot 2
1	Intel X520,DP,10G,SFP+ Adapter, Low Profile for Slot 1
1	Intel X520,DP,10G,SFP+ Adapter, Low Profile for Slot 1
1	No Operating System
1	No Operating System
1	System ordered as part of Multipack order
1	System ordered as part of Multipack order
1	PowerEdge DSS7000 Rails + Cable Management Arm,4U,Toolless, v2 for DAO
1	C1 Configuration for quantity 30 or 45 HDD, SAS or SATA
1	C1 Configuration for quantity 30 or 45 HDD, SAS or SATA
1	UEFI BIOS
1	UEFI BIOS

EXPANSION

When adding nodes and drives to current deployment, there are some rules and best practices to adhere to.

Rules include:

- Add drives up to what is supported for the node.
- Add a drive per node across all nodes evenly. For instance, if desiring to add 2 drives, add 2 drives on all nodes in the current deployment.
- For DSS 7000, DSS 7500 nodes must have the same number of drives as current nodes at a time. Recommended to add 2 DSS 7500 nodes at a time.
- Follow minimum disks and node rules supported as indicated in “Table 1”.

Best practices include:

- Add drives of same type and capacity.
- Best practice is not to wait until the storage platform is completely “full” before adding drives/nodes. A reasonable storage utilization threshold is 70% taking consideration daily ingest rate and expected order, delivery and integration time of added drives/nodes.

CUSTOMER PROVIDED INFRASTRUCTURE

In order to be able to deploy ECS, certain customer provided infrastructure requirements need to be reachable by the ECS system. A list of required and optional components includes:

- **Authentication Providers** – users (system admin, namespace admin and object users) can be authenticated using Active Directory or LDAP or Keystone
- **DNS Server –** Domain Name server or forwarder
- **NTP Server** – Network Time Protocol server. Please refer to the NTP best practices for guidance on optimum configuration
- **SMTP Server** – (optional) Simple Mail Transfer Protocol Server is used for sending reports from the ECS rack.
- **DHCP server –** only if assigning IP addresses via DHCP

- **Load Balancer** - (optional but highly recommended) evenly distributes loads across all data services nodes. Load balancers can use simple algorithms such as random choice or round robin. More sophisticated load balancers may take additional factors into account, such as a server's reported load, response times, up/down status, number of active connections, geographic location and so on. The customer is responsible for implementing load balancers; customers have several options including Manual IP allocation, DNS Round Robin, Client-Side Load Balancing, Load Balancer Appliances, and Geographic Load Balancers. The following are brief descriptions of each of those methods:

    - **Manual IP Allocation -** Data node IP addresses are manually distributed to applications. This is not recommended because it does not evenly distribute loads between the nodes and does not provide any fault-tolerance if a node fails.

    - **DNS Round-Robin -** With DNS Round-Robin, a DNS name is created for ECS and includes all of the IP addresses for the data nodes. The DNS server will randomly return the IP addresses when queried and provide some pseudo-load balancing. This generally does not provide fault-tolerance because you would need to remove the IP addresses from DNS to keep them out of rotation. Even after removing them, there is generally some TTL (time-to-live) issues where there is a delay to propagate the removal. Also, some operating systems like Windows will cache DNS lookups and can cause "stickiness," where a client keeps binding to the same IP address, reducing the amount of load distribution to the data nodes.

    - **Physical or Virtual load balancing-** This option is the most common approach to load balancing. In this mode, an appliance (hardware or software) receives the HTTP request and forwards it on to the data nodes. The appliance keeps track of the state of all of the data nodes (up/down, # of connections) and can intelligently distribute load amongst the nodes. Generally, the appliance will proactively "health check" the node (e.g. GET/?ping on the S3 head) to ensure the node is up and available. If the node becomes unavailable it will immediately be removed from rotation until it passes a health check. Another advantage to this kind of load balancing is SSL termination. You can install the SSL certificate on the load balancer and have the load balancer handle the SSL negotiation. The connection between the load balancer and the data node is then unencrypted. This reduces the load on the data nodes because they do not have to handle the CPU-intensive task of SSL negotiation.

    - **Geographic load balancing -** Geographic load balancing takes Physical or Virtual Load Balancing one step further: it adds load balancing into the DNS infrastructure. When the DNS lookups occur, they are routed via an "NS" record in DNS to delegate the lookups to a load balancing appliance like the Riverbed SteelApp. The load balancer can then use Geo-IP or some other mechanism to determine which site to route the client to. If a site is detected to be down, the site can be removed quickly from DNS and traffic will be routed to surviving sites.

  Available for reference is ECS with HAProxy Load Balancer Deployment Reference Guide which provides information and examples how to implement HAProxy, an open-source and free load balancer software, with ECS.

VALIDATION TESTING

Tests were conducted to validate and certify Dell EMC-approved industry standard hardware with ECS Software Bundle. These sets of tests are also used on the ECS Appliance and exercise the entire integrated system to the fullest to verify that there are no issues in using ECS Software Bundle with the hardware. The set of tests run include:

- Installation – tests install of ECS software.
- Extend – extend an ECS cluster by adding more nodes.
- Upgrade – upgrade cluster to a newer version of ECS Software.
- Geo - clusters spans multiple physical locations.
- Load testing – puts load on ECS running on the infrastructure and hardware for numerous hours.
- Serviceability and service procedure testing - node shutdown, disk replacement and node replacement tests.
- Mixed mode testing - validates ECS running on varying hardware at each site, ie. Dell DSS 7000 hardware and ECS appliance deployed in a geo-replicated environment.

Since varying certified hardware and models can be used for the ECS software, a precheck tool was developed to check that industry standard hardware meets the minimum specifications as discussed in the Tools section. Also, the precheck tool creates the configurations files needed to install ECS Software.

SIZING

Sizing and configuration for your particular use case is determined during preliminary qualification and engagement with Dell EMC sales and support personnel. Generally your capacity and performance requirements would define the number of Dell DSS 7000 needed. However, here is a list of considerations when sizing:

- **Capacity and Ingest rate** –net data capacity needed and expected rate of growth (i.e. how many objects will be created each day, what is the average size of each object, etc.)
- **Performance** – dependent on your application needs for performance (i.e. any known throughput or latency expectations, how many hosts/devices will be creating this data, etc.).
- **Erasure coding scheme** – ECS utilizes a 12+4 or 10+2 erasure coding schemes. The 10+2 erasure coding is mostly used for cold archive use case.
- **Number of Replication sites**– replicating to 3 or more sites reduces ECS storage overhead and must be considered when sizing.

The sizing and capacity tool will eventually be updated for industry standard hardware to assist in sizing. Performance results will also be available soon to further provide information on how many servers would be ideal to meet your storage requirements.

ORDERING AND COST

Ordering and quoting of ECS Software is available via ASD Helpdesk. The licensing model of ECS Software is based on the amount of storage deployed per site of the customer. The cost consists of the software license plus cost of premium support and professional services support.

SUPPORT

Support for ECS Software Bundle on industry standard hardware is a combination of both customer and Dell EMC support. The installation is done by both customer and Dell EMC personnel. After all ECS Software Bundle has been successfully installed, then hardware issues are handled by customer and ECS software and infrastructure such as operating system and Docker container issues are handled by Dell EMC support with customer assistance.

INSTALL SUPPORT

Installation of certified hardware is a collaborated engagement between customer and Dell EMC personnel (ECS Professional and Engineering services (as needed)). The steps of engagement and process of install include the following:

1. **Customer Engagement and Qualification** - preliminary discussions between customer and Dell EMC field or professional services personnel to gather requirements, use cases, current infrastructure, and other information.

2. **Node Preparation**

   a. Setup hardware and networking

   b. Install operating system and Docker

   c. Setup or use customer provided infrastructure - DNS, NTP, authentication providers, load balancers, etc, password-less ssh (Customer)

3. **Validation of Hardware platform - Precheck**

   a. ECS precheck tool is run to assess the deployed environment and hardware

   b. If precheck PASSES “clean”, proceed to Step 4 - Install.

   c. If precheck FAILS, remediate environment and escalate accordingly to the appropriate engineering teams (i.e. ASD/CSE or ASD/Engineering).

   d. Resolve precheck issues with assistance from ECS Engineering (i.e. ASD/CSE, ASD/Engineering).

4. **Install of ECS Software**– deploy ECS layers & services – HAL, Fabric, and Object similarly to the same way the ECS Appliance is done.

5. **Provisioning** – provision storage and ready for use

GENERAL SUPPORT

After ECS Software Bundle is up and running on the hardware and system has been provisioned, any issues encountered will follow the normal process of reporting bugs and issues (i.e. file SR). General support depending on type of issue for certified hardware is shown in Table 4.

Table 4 - General Support for Certified Hardware

Type	Support
Hardware Maintenance	Customer
Hardware Issue Resolution	Customer
Operating System Patch Deployment and Management	Dell EMC
Operating System Issue Resolution	Dell EMC
ECS Sofware Installation and Upgrade Using ECS Tools	Dell EMC
ECS Storage Engine and Access Protocol – Issue Resolution	Dell EMC

CONCLUSION

ECS Software Bundle on Dell DSS 7000 offers options to customers who prefer to reduce their capital expense by utilizing commercial off the shelf hardware, eliminate vendor lock-in and/or build a homogenous datacenter with unified commodity hardware. The Dell DSS 7000 is an outstanding hardware platform for ECS Software and is well known for their performance, capacity and scale-out capabilities. ECS software, tools and libraries have been enhanced and improved to accommodate for industry standard hardware. With the assistance of Dell EMC personnel, the hardware and software configuration best suited for your use case and environment can be installed.

REFERENCES

- **ECS Product Documentation**
  - ECS Architecture Whitepaper
    - http://www.emc.com/collateral/white-papers/h14071-ecs-architectural-guide-wp.pdf
  - ECS product documentation sites
    - https://support.emc.com/products/37254_ECS-Appliance-/Documentation/

- **Dell Product Documentation**
  - Dell DSS 7000 Manuals
    - http://www.dell.com/support/home/us/en/19/product-support/product/dell-dss7000/manuals
  - Dell Extreme Scale Infrastructure
    - https://www.dell.com/en-us/work/learn/extreme-scale-infrastructure