Dell Red Hat Cloud Solutions Reference Architecture Guide - Version 3.0.1

Contents

Trademarks

Trademarks used in this text: Dell™, the DELL logo, Dell Precision™, OptiPlex™, Latitude™, PowerEdge™, PowerVault™, PowerConnect™, OpenManage™, EqualLogic™, KACE™, FlexAddress, Force10™, and Vostro™ are trademarks of Dell Inc. Intel®, Pentium®, Xeon®, Core®, and Celeron® are registered trademarks of Intel Corporation in the U.S. and other countries. AMD® is a registered trademark and AMD Opteron™, AMD Phenom™, and AMD Sempron™ are trademarks of Advanced Micro Devices, Inc. Microsoft®, Windows®, Windows Server®, MS-DOS®, and Windows Vista® are either trademarks or registered trademarks of Microsoft Corporation in the United States and/or other countries.

Red Hat®, Red Hat® Enterprise Linux®, the Shadowman logo, and JBoss are trademarks of Red Hat, Inc., registered in the U.S. and other countries. Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries. Novell® is a registered trademark and SUSE™ is a trademark of Novell Inc. in the United States and other countries. Oracle® is a registered trademark of Oracle Corporation and/or its affiliates. Citrix®, Xen®, XenServer® and XenMotion® are either registered trademarks or trademarks of Citrix Systems, Inc. in the United States and/or other countries. VMware®, Virtual SMP®, vMotion®, vCenter®, and vSphere® are registered trademarks or trademarks of VMware, Inc. in the United States or other countries.

DISCLAIMER: The OpenStack® Word Mark and OpenStack Logo are either registered trademarks/service marks or trademarks/service marks of the OpenStack Foundation, in the United States and other countries, and are used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation or the OpenStack community.

Other trademarks and trade names may be used in this publication to refer to either the entities claiming the marks and names or their products. Dell Inc. disclaims any proprietary interest in trademarks and trade names other than its own.

Notes, Cautions, and Warnings

A **Note** indicates important information that helps you make better use of your system.

A **Caution** indicates potential damage to hardware or loss of data if instructions are not followed.

A **Warning** indicates a potential for property damage, personal injury, or death.

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.

Glossary

BMC/IDRAC Enterprise

Baseboard management controller. An on-board microcontroller that monitors the system for critical events by communicating with various sensors on the system board and sends alerts and log events when certain parameters exceed their preset thresholds.

Bundle

A customer-orderable solution that consists of:

- All server, network, and storage hardware needed to install and operate the solution as outlined
- All necessary solution software licenses needed to install and operate the solution as outlined

Cloud Computing

See *http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf*

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Cluster

A set of computers dedicated to OpenStack that can be attached to a pair of distribution switches.

Compute Node

The hardware configuration that best supports the hypervisor server or Nova compute roles.

DevOps

Development Operations (DevOps) is an operational model for managing data centers using improved automated deployments, shortened lead times between fixes, and faster mean time to recovery. See *http://en.wikipedia.org/wiki/DevOps*

Hypervisor

Software that runs virtual machines (VMs).

LAG

Link Aggregation Group.

LOM

LAN on motherboard.

Node

One of the servers in the system.

Pod

An installation comprised of three racks, based upon server and network sizing.

SAH

The Solution Admin Host (SAH) is a physical host that supports the Infrastructure Roles needed for the cluster.

Storage Node

The hardware configuration that best supports storage functions such as Ceph.

VLT

A Virtual Link Trunk (VLT) is the combined port channel between an attached device (ToR switch) and the VLT peer switches.

VLTi

A Virtual Link Trunk Interconnect (VLTi) is an interconnect used to synchronize states between the VLT peer switches. Both endpoints must be on 10G or 40G interfaces; 1G interfaces are not supported.

Overview

This Reference Architecture focuses on helping organizations to deploy an OpenStack® based cloud.

This jointly-engineered and validated architecture details the Dell™ Red Hat® Cloud Solutions with Red Hat Enterprise Linux™ OpenStack Platform and encompasses software, hardware, operations, and integration of the solution components. The architecture provides prescriptive guidance and recommendations for compute and storage nodes, networking design and various system configurations.

OpenStack Maturity

The code base for Red Hat Enterprise Linux OpenStack Platform is evolving at a very rapid pace. Please see *https://access.redhat.com/site/support/policy/updates/OpenStack/platform* for more information.

At publication the current release of OpenStack is codenamed *Juno*. It builds upon previous releases with 342 new features to support software development, and managing data and application infrastructure, at scale. It is developed by over 1400 individuals employed by more than 133 organizations. Please see *http://www.OpenStack.org/software/juno*.

Dell and Red Hat designed this Reference Architecture to make it easy for Dell Red Hat Cloud Solutions customers to build their own operational readiness cluster and design their initial offerings, using the current releases. Dell and Red Hat provide the support and services customers need to stand up production-ready OpenStack clusters.

Hardware Options

To reduce time spent on hardware specification for an initial system this Reference Architecture offers specific choices for servers, storage and networking:

- The recommended hardware is general-purpose, and enables a wide range of configuration options.
- The recommended hardware has been optimized for infrastructure, compute, and storage roles. As noted throughout this Reference Architecture, Dell constantly adds capabilities to expand this offering.

Each of the Dell PowerEdge™ server configurations in this Reference Architecture is designed as a getting-started setup for OpenStack compute, OpenStack storage, or both simultaneously. Dell recommends starting with OpenStack software using components from this configuration because the hardware and operations processes are a flexible foundation to expand upon. By design, you can expand the Reference Architecture configuration as your cloud deployment grows, so your investment is protected.

Networking and Network Services

Network configuration is based upon using the Neutron-based options supported by the OSP code base, and not relying upon third-party drivers. This reference configuration is based upon the Neutron networking services using the ML2 drivers for `OpenVswitch` with the `vlan` option.

Additional networking can include:

- Core and layered networking capabilities
- 10GbE networking

- NIC teaming
- Redundant trunking top-of-rack (ToR) switches into core routers

This enables the solution to operate in a full production environment.

See *Network Architecture* on page 21 for guidelines. Detailed designs are available through Dell consulting services.

Taxonomy

This solution contains the core OpenStack components as delivered in the Red Hat Enterprise Linux OpenStack Platform:

- Block Storage (Cinder with Ceph and EqualLogic)
- Compute (Nova)
- Dashboard (Horizon)
- Database (Trove, available but not part of the solution)
- Data Process (Sahara, available but not part of the solution)[1]
- Identity (Keystone)
- Image Service (Glance)
- Integration Testing (Tempest, available but not part of the solution)[1]
- Networking (Neutron)
- Orchestration (Heat)
- Telemetry (Ceilometer, available but not part of the solution)[1]

Caution: Tempest information is provided in this Reference Architecture as a technology preview only. A custom Services engagement is required for installation and configuration. Before using Tempest, review the Tempest documentation at *http://docs.openstack.org/developer/tempest/*.

The taxonomy presented in *Figure 1: OpenStack Taxonomy* on page 10 reflects infrastructure components (shown in light green) and OpenStack-specific components (shown in red), that are under active development by the community, Dell, and Red Hat. The taxonomy reflects a DevOps perspective in that there are two sides for cloud users:

- Site-specific infrastructure
- Standards-based API (shown in pink) interactions

The standards-based APIs are the same between all OpenStack deployments, and let customers and vendor ecosystems operate across multiple clouds. The site-specific infrastructure combines open and proprietary software, Dell hardware, and operational processes to deliver cloud resources as a service.

The implementation choices for each cloud infrastructure are highly specific to the requirements of each site. Many of these choices can be standardized and automated using the tools in this Reference Architecture and by following DevOps processes. Conforming to best practices helps reduce operational risk by leveraging the accumulated experience of Dell, Red Hat and the broader OpenStack community.

[1] Available through a custom Services engagement, for customer evaluation only.

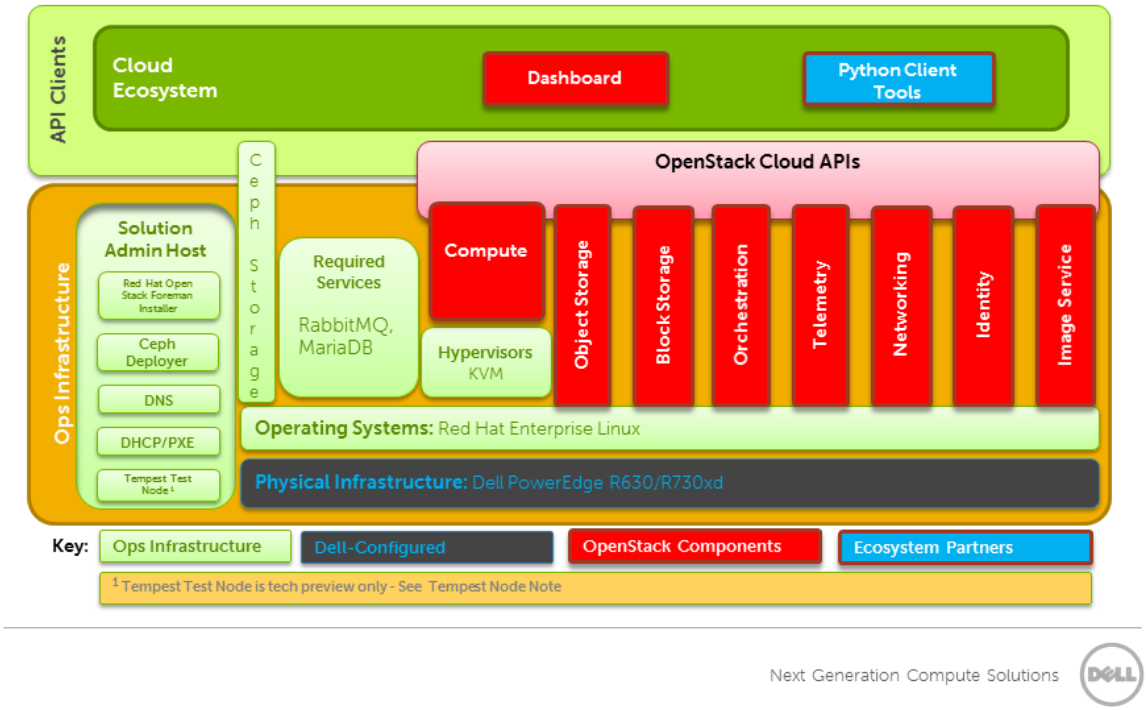

Figure 1: OpenStack Taxonomy

Red Hat Enterprise Linux OpenStack Platform 6.0

Red Hat Enterprise Linux OpenStack Platform delivers an integrated foundation to create, deploy, and scale a secure and reliable public or private OpenStack cloud. It delivers a cloud platform built from Red Hat OpenStack technology, co-engineered and integrated with Red Hat Enterprise Linux, giving you the agility to scale and quickly meet customer demands without compromising on availability, security, or performance.

Red Hat Enterprise Linux OpenStack Platform is purposely designed with the recognition of the unique dependencies OpenStack has on the underlying Linux it's installed on. Red Hat uniquely co-engineers and integrates Red Hat OpenStack technology with Red Hat Enterprise Linux Server 7.1, ensuring a stable, production-ready cloud platform. Version 6.0 boasts all the core features and functions of the community Juno release, adds some additional innovations by Red Hat, resulting in a hardened, stable cloud platform.

Key Benefits

- Co-engineered and Integrated: OpenStack depends on Linux for performance, security, hardware enablement, networking, storage, and other primary services. Red Hat Enterprise Linux OpenStack Platform delivers an OpenStack distribution with the proven performance, stability, and scalability of Red Hat Enterprise Linux, enabling you to focus on delivering the services your customers want instead of the underlying operating platform.

- Deploy with confidence, as Red Hat Enterprise Linux OpenStack Platform provides a hardened and stable branch release of OpenStack and Linux, which is supported by Red Hat for a three (3) year "production phase" life cycle, well beyond the six-month release cycle of unsupported community OpenStack. Security fixes, bug fixes, performance enhancements, and some features can be back-ported from future releases without disrupting production environments.
- Take advantage of broad application support. Red Hat Enterprise Linux running as guest virtual machines provides a stable application development platform with a broad set of certified ISV certifications, so that you can rapidly build and deploy your cloud applications.
- Avoid vendor lock-in by moving to open technologies while maintaining your existing infrastructure investments.
- Benefit from the world's largest partner ecosystem: Red Hat has assembled the world's largest ecosystem of certified partners for OpenStack compute, storage, networking, ISV software, and services for Red Hat Enterprise Linux OpenStack Platform deployments, ensuring the same level of broad support and compatibility customers enjoy today in the Red Hat Enterprise Linux ecosystem.
- Bring security to the cloud. Rely on the SELinux military-grade security and container technologies of Red Hat Enterprise Linux to prevent intrusions and protect your data when running in public or private clouds.[2]

[2] Available through a custom Services engagement, for customer evaluation only.

OpenStack Architecture

While OpenStack has many configurations and capabilities, we focus on the primary components for Red Hat Enterprise Linux® OpenStack Platform 6.0 (Juno), as Dell has defined in *Taxonomy* on page 9 above.

Note: For a complete overview of OpenStack software, visit *Red Hat OpenStack Enterprise Platform* and the *OpenStack Project*.

OpenStack Components

The following component descriptions are from the OpenStack Foundation website. Extensive documentation for the OpenStack components is available at *http://docs.openstack.org/*.

Table 1: OpenStack Components

Function	Code Name	Description
Block Storage	Cinder	*OpenStack Block Storage* provides persistent block level storage devices for use with OpenStack compute instances. The block storage system manages the creation, attaching, and detaching of the block devices to servers. Block storage volumes are fully integrated into OpenStack Compute and the Dashboard enabling cloud users to manage their own storage needs. In addition to using simple Linux® server storage, it has unified storage support for numerous storage devices. Block storage is appropriate for performance sensitive scenarios such as database storage, expandable file systems, or providing a server with access to raw block level storage. Snapshot management provides powerful functionality for backing up data stored on block storage volumes. Snapshots can be restored or used to create a new block storage volume.
Compute/IaaS	Nova	*OpenStack Compute* is open source software designed to provision and manage large networks of virtual machines, creating a redundant and scalable cloud computing platform. It gives you the software, control panels, and APIs required to orchestrate a cloud, including running instances, managing networks, and controlling access through users and projects. OpenStack Compute strives to be both hardware and hypervisor agnostic, currently supporting a variety of standard hardware configurations and seven major hypervisors.
Dashboard/ Portal	Horizon	*OpenStack Dashboard* provides administrators and users a graphical interface to access, provision and automate cloud-based resources. The extensible design makes it easy to plug in and expose third party products and services.
Identity	Keystone	*Identity Service* provides a central directory of users mapped to the OpenStack services they can access. It acts as a common authentication system across the cloud operating system and can integrate with existing backend directory services.

Function	Code Name	Description
Integration Testing	Tempest	*Tempest* is a set of integration tests to be run against a live OpenStack cluster. Tempest includes batteries of tests for OpenStack API validation, Scenarios, and other specific tests useful in validating an OpenStack deployment. Tempest is provided as a technology preview, and is not included in Red Hat Enterprise Linux OpenStack Platform version 6.0, nor supported by the Dell Red Hat Cloud Solutions.
Networking	Neutron	*OpenStack Networking* is a pluggable, scalable and API-driven system for managing networks and IP addresses. Like other aspects of the cloud operating system, it can be used by administrators and users to increase the value of existing datacenter assets. OpenStack Networking ensures the network will not be the bottleneck or limiting factor in a cloud deployment and gives users real self-service, even over their network configurations.
Orchestration	Heat	*OpenStack Orchestration* is a template-driven engine that enables application developers to describe and automate the deployment of infrastructure. The flexible template language can specify compute, storage and networking configurations as well as detailed post-deployment activity to automate the full provisioning of infrastructure as well as services and applications. Through integration with the Telemetry service, the Orchestration engine can also perform auto-scaling of certain infrastructure elements.
Telemetry	Ceilometer	*OpenStack Telemetry* aggregates usage and performance data across the services deployed in an OpenStack cloud. This powerful capability provides visibility and insight into the usage of the cloud across dozens of data points and enables cloud operators to view metrics globally or by individual deployed resources.
Virtual Images	Glance	*OpenStack Image Service* provides discovery, registration, and delivery services for virtual disk images. The Image Service API server provides a standard REST interface for querying information about virtual disk images stored in a variety of back-end stores, including OpenStack Object Storage. Clients can register new virtual disk images with the Image Service, query for information on publicly available disk images, and use the Image Service's client library for streaming virtual disk images.

Storage Options

OpenStack has several storage services, including:

- Cinder
- Glance
- Swift [3]

Together these services provide VM with block, image, and object storage. In turn, the services employ block and object storage subsystems. Since the service design has a mechanism to replace some or all of the implementation of these services, this solution can provide alternate implementations of these services that better serve our customer's needs.

Cinder virtualizes storage enabling VMs to use persistent block storage through Nova. OpenStack consumers should write data that must exist beyond the lifecycle of the guest to Cinder volumes. The volume can be accessed afterwards by a different guest.

Glance provides images to VMs. Generally, the the images are block devices containing DVDs or virtual machines. VMs can be booted from these images or have the images attached to them.

Swift provides an object storage interface to VMs and other OpenStack consumers. Unlike block storage where the guest is provided a block device of a given format and is accessible within the cluster, object storage is not provided through the guest. Object storage is generally implemented as a HTTP/HTTPs based service, a web server. Swift in this document refers to the Swift protocol not the Swift implementation of the protocol. Client implementations within the guest or external OpenStack clients would interact with Swift without any configuration required of the guest other than providing the requisite network access. For example, VM within OpenStack can put data into Swift and later external clients could pull that data for additional processing.

As with other OpenStack services, there is a client and server components for each of these. The server component can be modified to use a particular type of storage rather than the default. For example, Cinder uses local disks as the storage back-end by default. This solution modifies the default configuration for these services.

This solution include three alternate implementations of Cinder that enable the cluster to fit many needs:

- *Ceph*
- *Dell EqualLogic Storage Arrays*
- *Cinder Multi-Backend Support*

This solution provides Ceph as an alternate implementation for Glance and Block stores.

Red Hat Ceph Storage

Ceph is a scale-out, shared-nothing, distributed, software-defined storage system. The Solution bundle includes Ceph storage. Ceph is used as the storage backend for Nova, Cinder and Glance. Ceph is used as block storage. Storage nodes run the Ceph software. Compute and controller nodes run the Ceph block client.

[3] Available through a custom Services engagement, for customer evaluation only.

Dell EqualLogic Storage Arrays

Dell EqualLogic storage arrays are designed to provide simplified deployment and administration of consolidated storage environments. Dell EqualLogic storage systems are self-optimized, utilizing embedded load-balancing technologies that react to workload demands.

The core capabilities of Dell EqualLogic storage products include comprehensive software components and host integration, which simplify administrative tasks and assist with storage management. Application-layer integration with OpenStack enables Cinder to:

- Provision and manage volumes on Dell EqualLogic storage
- Utilize SAN-based snapshots for protection capability

Cinder Multi-Backend Support

This solution can take advantage of Cinder's Multi-Backend and Multi-Instance support, using different storage types to meet the requirements that the application requires. You can meet different use cases based on performance, large data transfers, temporary storage, and other criteria.

For example, you can create a Ceph cluster with a performance group and a second storage option of a Dell EqualLogic group to support large data transfers. Then, using Cinder multi-backend, configure the virtual machine volume requirements as required.

Server Infrastructure Options

The Solution includes using the PowerEdge R630 and R730xd Server lines. The following sections describe the supported server models and configurations required. Detailed part lists and rack layouts are included in the *Dell Red Hat Cloud Solutions Build of Materials Guide*.

PowerEdge R630 Server

The PowerEdge R630 server is a hyper-dense, two-socket, 1U rack server.

With computing capability previously only seen in 2U servers, the ultra-dense PowerEdge R630 two-socket 1U rack server delivers an impressive solution for cloud solutions, virtualization environments, large business applications or transactional databases.

The PowerEdge R630 server is versatile and highly configurable for a variety of solutions, delivering the latest Intel® Xeon® processor E5-2600 v3 product family, 24 DIMMs of high-performance DDR4 memory and a broad range of local storage options.

PowerEdge R730xd Servers

The PowerEdge R730xd is an exceptionally flexible and scalable, two-socket 2U rack server that delivers high performance processing and a broad range of workload-optimized local storage possibilities, including hybrid tiering.

Designed with an incredible range of configurability, the PowerEdge R730xd meets the needs of many different workloads with the latest Intel® Xeon® processor E5-2600 v3 product family, 24 DIMMs of high-performance DDR4 memory and a broad range of local storage options.

Figure 2: PowerEdge R730xd Servers - 2.5" and 3.5" Chassis Options

Hardware Configurations

Table 2: Controller Node Hardware Configurations – PowerEdge R630

Machine Function	Solution Bundle Controller Nodes
Platform	PowerEdge R630
CPU	2 x E5-2650v3 (10-core)
RAM (Minimum)	128 GB
LOM	2 x 1Gb, 2 x Intel X520 10Gb
Add-in Network	1 x Intel X520 DP 10Gb DA/SFP+
Disk	4 x 500GB 7.2k NLSAS
Storage Controller	PERC H730
RAID	RAID 10

Table 3: Compute Node Hardware Configurations – PowerEdge R630

Machine Function	Solution Bundle Compute Nodes
Platform	PowerEdge R630v3 (8-core)
CPU	2 x E5-2650v3 (10-core)
RAM (Minimum)	128 GB
LOM	2 x 1Gb, 2 x Intel X520 10Gb
Add-in Network	1 x Intel X520 DP 10Gb DA/SFP+
Disk	6 x 600GB 10k SAS
Storage Controller	PERC H730
RAID	RAID 10

Table 4: Infrastructure Node Hardware Configurations – PowerEdge R630

Machine Function	Solution Bundle Infrastructure Nodes
Platform	PowerEdge R630
CPU	2 x E5-2630v3 (8-core)
RAM (Minimum)	32 GB
LOM	2 x 1Gb, 2 x Intel X520 10Gb
Add-in Network	1 x Intel X520 DP 10Gb DA/SFP+
Disk	4 x 500GB 7.2 NLSAS
Storage Controller	PERC H730
RAID	RAID 10

Table 5: Storage Node Hardware Configurations – PowerEdge R730xd

Machine Function	Solution Bundle Storage Nodes
Platforms	PowerEdge R730xd
CPU	2 x E5-2650v3 (10-core)
RAM (Minimum)	48 GB
LOM	1 x 1Gb, 2 x Intel X520 10Gb
Add-in Network	2 x Intel X520 DP 10Gb DA/SFP+
Disk	Flex Bay: 2 X 300GB 10K 2.5-inch (OS) Front Drives: 3 X 200GB SSD 13 x 2TB or 4TB NL SAS 7.2K 3.5-inch
Storage Controller	PERC H730
RAID	RAID 1 (operating system) pass through SSD pass through each data disk

Note: Be sure to consult your Dell account representative before changing the recommended hardware configurations.

Configuration Notes

The *Dell Red Hat Cloud Solutions Bill of Materials Guide* contains the full bill of materials (BOM) listing for the PowerEdge R630 and R730Xd server configurations.

The R630 and R730xd configurations are used with 10GbE networking. To ensure that the network is HA ready an additional network card is required in each node. Refer to the *Dell Red Hat Cloud Solutions Bill of Materials Guide* which outlines the supported cards and includes them as part of the solution.

Operational Notes

This section provides a basic overview of the following system aspects:

- *Backup/Recovery*
- *Service Layout*
- *Deployment*

Backup/Recovery

Backup and recovery have not been addressed in this configuration. The Red Hat OpenStack Manager Virtual Server, while not needed for normal operations of the services, is not redundant or backed up.

Service Layout

During the deployment each service configured by the solution needs to be on a particular hardware type.

For each server platform, two types of nodes have been designed:

- Infrastructure
- Storage

Red Hat OpenStack Manager is designed for flexibility, enabling you to try different configurations in order to find the optimal service placement for your workload. *Table 6: Node Type to Services* on page 19 presents the recommended layout of each service.

The Red Hat OpenStack Manager and the Red Hat Ceph Enterprise Admin are deployed to the Solution Admin Host as manually-configured VMs. This enables each tool to control its respective resources.

Table 6: Node Type to Services

Hardware Type	Service	Node to Deploy
Infrastructure	Ceilometer	OpenStack Controllers
Infrastructure	Cinder-scheduler	OpenStack Controllers
Infrastructure	Cinder-volume	OpenStack Controllers
Infrastructure	Database-server	OpenStack Controllers
Infrastructure	HA-Proxy (Load Balancer)	OpenStack Controllers
Infrastructure	Heat	OpenStack Controllers
Infrastructure	Keystone-server	OpenStack Controllers
Infrastructure	Neutron-server	OpenStack Controllers
Infrastructure	Nova-Controller	OpenStack Controllers
Infrastructure	Nova dashboard-server	OpenStack Controllers
Infrastructure	Nova-multi-compute	Three or more Compute Nodes
Infrastructure	Pacemaker	OpenStack Controllers

Hardware Type	Service	Node to Deploy
Infrastructure	RabbitMQ-server (Messaging)	OpenStack Controllers
Infrastructure	Red Hat OpenStack Manager	Solution Admin Host (KVM)
Infrastructure	Tempest Test Node[4]	Solution Admin Host (KVM)
Storage	Dell EqualLogic Array (optional)	Dell EqualLogic Arrays
Storage	Red Hat Ceph Enterprise Admin (Calamari)	Solution Admin Host (KVM)
Storage	Red Hat Ceph Enterprise Storage (Block)[5]	Three or more Storage Servers
Storage	Red Hat Ceph Enterprise Monitor	OpenStack Controllers

Deployment

Deployment consists of three phases:

- **Hardware Setup:**
  - Rack and stack
  - Cabling
  - Server BIOS configuration
  - Server RAID configuration
  - Switch configuration
- **Software Setup:**
  - Deploy Solution Admin Host for provisioning services
    - Deploy Calamari Node to the Solution Admin Host
    - Deploy Red Hat OpenStack Manager Virtual Server
    - Deploy Tempest Test Node to the Solution Admin Host
  - Provision all nodes with Operating Systems
  - Validate all nodes networking
  - Provision services to nodes based on their node types
- **Testing environment**

[4] Available through a custom Services engagement, for customer evaluation only.
[5] Available through a custom Services engagement, for customer evaluation only.

Network Architecture

The Dell Red Hat Cloud Solutions with Red Hat Enterprise Linux OpenStack Platform uses Dell Force10 S55 1/10-Gigabit and/or Dell Force10 S4810 10-Gigabit Ethernet switches as the top-of-rack connectivity to all OpenStack-related nodes. This Reference Architecture is used to support consistency in rapid deployments through the minimal differences in the network configuration.

Infrastructure Layouts

The network consists of the following major network infrastructure layouts:

- **Core Network Infrastructure** - The connectivity of aggregation switches to the core for external connectivity.
- **Data Network Infrastructure** - The data network consists of the server NICs, the top-of-rack (ToR) switches, and the aggregation switches.
- **Management Network Infrastructure** - The BMC management network, consisting of iDRAC ports and the out-of-band management ports of the switches, is aggregated into a 1-RU S55 switch in one of the three racks in the cluster. This 1-RU switch in turn can connect to one of the Aggregation or Core switches to create a separate network with a separate vLAN.

Network Components

The data network is primarily composed of the ToR and the aggregation switches. Configurations for 1GbE and 10GbE are included in this Reference Architecture. The following component blocks make up this network:

Server Nodes

Server connections to the network switches are setup to be highly available.

In order to create a highly-available solution, the network must be resilient to loss of a single network switch, network interface card (NIC) or bad cable. To achieve this, the network configuration uses channel bonding across the servers and switches. There are several types (or modes) of channel bonding, but only three are recommended for the Solution:

- `active-backup` (mode = 1)
- `balance-xor` (mode = 2)
- `802.3ad` or LACP (mode = 4)

Note: Other modes, such as `balance-rr` (mode=0), `broadcast` (mode=3), `balance-tlb` (mode=5), and `balance-alb` (mode=6), are not supported.

The OpenStack Controller nodes can only be used with two of the channel bonding modes:

- `active-backup` (mode = 1)

- `balance-xor` (mode = 2 - solution default)

The OpenStack Compute nodes, Ceph Storage nodes, and Solution Admin Host can use any of the three channel bonding modes:

- `active-backup` (mode = 1)
- `balance-xor` (mode = 2)
- `802.3ad` or LACP (mode = 4 - solution default)

For all nodes, the endpoints are terminated to switch ports, that have been configured for the particular channel bonding mode, across two Dell Force S4810s that are built up with a VLTi across them. The configuration settings are explained in greater detail in the *Dell Red Hat Cloud Solutions Deployment Guide*.

Table 7: Channel Bonding Modes Supported

	Channel Bonding Type		
Node Type	**active-backup (Mode 1)**	**balance-xor (mode 2)**	**802.3ad (LACP mode 4)**
Solution Admin Host	Yes	Yes	Yes (solution default)
OpenStack Controller Nodes	Yes	Yes (solution default)	No
OpenStack Compute Nodes	Yes	Yes	Yes (solution default)
Ceph Storage Nodes	Yes	Yes	Yes (solution default)

A single port is an option; it is not used in the solution. The need to eliminate single points of failure is taken into consideration as part of the design, and this option has been eliminated wherever possible.

Please contact your sales representative to find out if there other options available.

Access Switch or Top of Rack (ToR)

The servers connect to ToR switches. Typically there are two in each rack. The switches recommended by Dell are:

- **1GbE Connectivity** - Dell Force10 S55
- **10GbE Connectivity** - Dell Force10 S4810

The 10GbE configuration utilizes Force10 S4810 switches as the ToR switches. Dell recommends this pair of switches run Virtual Link Trunking (VLT) for HA. This feature enables the servers to terminate their LAG interfaces into two different switches instead of one. This enables active-active bandwidth utilization. This feature provides redundancy within the rack if one switch fails, or needs maintenance. The uplink to the aggregation pair is 80Gb, using a LAG from each ToR switch. This is achieved by using two 40Gb interfaces in a LAG connecting to the aggregation pair. Therefore, a collective bandwidth of 160Gb is available from each rack.

Each rack is managed as a separate entity from a switching perspective, and ToR switches connect only to the aggregation switches.

Check with your Dell Sales Representative for other options.

Aggregation Switches

For a deployment from one to three racks of 10G servers, Dell recommends the Dell Force10 S4810 as the aggregation switch. It is both 10GbE and 40GbE capable.

The 40GbE interfaces on the S4810 could be converted into four 10GbE interfaces, thereby converting this switch into 64 10GbE-capable ports. ToR switches connect to aggregate switches via uplinks of 10GbE interfaces from the ToR Force10 S4810 to the Force10 S4810.

Dell's recommended architecture uses Virtual Link Trunking (VLT) between the two Force10 S4810 switches in each rack and then aggregation to a core switch. This feature enables a multi-chassis LAG from the ToR switches in each rack. The stacks in each rack can divide their links between this pair for switches to achieve powerful active-active forwarding, while using full bandwidth capability, with no requirement for spanning tree. Running 40GbE Ethernet switches, like the Dell Force10 Z9000, in aggregation can achieve a scale of up to hundreds of 1G deployed nodes.

For the 10G server deployment, Dell's recommendation depends upon:

- The scale at which the rack layouts are planned
- Required future scaling

When designing a large deployment, Dell recommends the Force10 S4810 for aggregation for smaller scale and the Force10 Z9000 for larger deployments. The Force10 Z9000 is a 32-port, 40G high-capacity switch. It can aggregate up to 15 racks of high-density PowerEdge R630 and R730xd servers. The rack-to-rack bandwidth needed in OpenStack would be most suitably handled by a 40G-capable, non-blocking switch. The Force10 Z9000 can provide a cumulative bandwidth of 1.5TB of throughput at line-rate traffic from every port.

Core

The aggregation layer could itself be the network core in many cases, but otherwise it would connect to a larger core. Details on this topic are beyond the scope of this document.

Layer-2 and Layer-3 Switching

The layer-2 and layer-3 boundaries are separated at the aggregation layer.

The Reference Architecture uses layer-2 as the reference up to the aggregation layer. That is why VLT is used on the aggregation switches. The Red Hat Foreman OpenStack Manager requires a layer-2 domain in order to provision servers.

The three network links - Provisioning, Storage, and Management - can have uplinks to a gateway device. The Provisioning network can use the Red Hat Foreman OpenStack Manager as a proxy for pulling packages from a subscription server, or a gateway can be added. The EqualLogic arrays on the Storage network may need access:

- From metrics and monitoring tools
- To enable management and updates

There are many tools for OOB management for the iDRAC, by simply adding the gateway to the network and updating the iDRAC these tools can be used.

The OpenStack Controllers are connected to a gateway device, usually a router or firewall. This device will handle routing for all networks external to the cluster. The required networks are:

- The floating IP range used by virtual machines
- A network for all external RESTful API and Graphical User Interface access

vLANs

This Reference Architecture implements at a minimum seven (7) separate Layer 2 vLANs:

- **Management/Out of Band (OOB) Network** - iDRAC connections can be routed to an external network. All OpenStack HA Controllers need direct access to this for IPMI operations.
- **Internal Network for Tenants** - Sets up the backend network for Nova and the VMs to use.
- **External Network for API, Graphical User Interfaces and Tenants** - Sets up a Layer 3 connection to a router external to the cluster. All external traffic to the cluster travels over this link. Sets up the front network for routable traffic to individual VMs, access to the RESTful API, and the Horizon GUI.

Depending upon the network configuration these networks may be shared, or routed as needed. Access to the External API Network by the Tempest Test Node is required.

- **Provisioning Network vLAN** - Connects a NIC from all nodes into the fabric, used for setup and provisioning of the OpenStack servers.
- **Private API Network Cluster Management vLAN** - Used for communication between OpenStack Controllers, Tempest Test Node, and Compute nodes for RESTful API and cluster communications.
- **Storage Network vLAN** - Used by all the nodes for the data plane reads/writes to communicate to OpenStack Storage, setup, and provisioning of the Ceph storage cluster, and when included, the EqualLogic arrays.
- **Storage Clustering Network vLAN** - Used by all Storage nodes for replication and data checks (Ceph clustering).

Out of Band Management Network

The management network of all the servers and switches is aggregated into a Dell Force10 S55 switch that is located in each rack of up to 3 Racks or a pod. It uplinks on a 10G link to the S4810 switches.

The Out of Band (OOB) Management network is used for several functions:

- The highly available servers, the controllers, use it for a heartbeat to ensure the servers are all up and functioning
- The highly available software uses it to reboot and partition servers
- When an uplink to a router is added and the iDRACs configured to use it as a gateway, there are tools for monitoring the servers and gather metrics on them. These are out of scope of this solution.

Dell OpenSwitch Solution

In addition to the Dell switch-based Reference Architecture, Dell provides an open standard that enables you to choose other brands and configurations of switches for your OpenStack environment.

You are expected to ensure that the switches conform to these requirements, and that they are configured according to this Reference Architecture's guidelines. The following list of requirements will enable other brands of switches to properly operate with Dell's required tools and configurations:

- Support for IEEE 802.1Q vLAN traffic and port tagging
- Support using one untagged and multiple tagged vLANs on the same port
- Ability to provide a minimum of 170 Gigabit Ethernet ports in a non-blocking configuration within the Provisioning vLAN
    - Configuration can be a single switch or a combination of stacked switches to meet the additional requirements
- The ability to create Link Aggregation Groups (LAGs) with a minimum of two physical links in each LAG
- If multiple switches are stacked:
    - The ability to create a LAG across stacked switches
    - Full-bisection bandwidth
    - Support for vLANs to be available across all switches in the stack
- 250,000 packets-per-second capability per switch
- A managed switch that supports SSH and serial line configuration
- SNMP v3 support

Solution Bundle

This core architecture provides prescriptive guidance and recommendations jointly engineered by Dell and Red Hat for deploying Red Hat Enterprise Linux OpenStack Platform 6 with Dell infrastructure

Our aims are to:

- Provide practical system design guidance and recommended configurations
- Develop tools to use with OpenStack for day-to-day usage and management
- Develop networking configurations capable of supporting a production system

The development of this architecture builds upon the experience and engineering skills of Dell and Red Hat and encapsulates best practices developed in numerous real world deployments. The designs and configurations in this architecture have been tested in Dell and Red Hat labs to verify system functionality and operational robustness.

Solution Bundle

This does not preclude the need to become familiar with OpenStack administration and day-to-day management. This bundle consists of the components described in *Figure 3: Solution Bundle and Ceph Cluster* on page 26.

Given a virtual machine with the 2 cores, 4GB of memory, and 40GB ephemeral drive, you can expect to run around 90 virtual machines with a 1.5 oversubscription of CPU cores. At 120 virtual machines, you will have two-to-one CPU core oversubscription, still have ephemeral storage under subscribed, and memory just starting to be oversubscribed.

As the Solution Bundle is designed for the beginning of a production environment, key OpenStack services are made highly available (HA) by clustering the OpenStack Controller nodes. The networking is based on 10Gbe bonds for data networks and the network switches are setup for HA. The Out of Band Management network is not HA and is 1GbE. Please review and discuss with your Sales Team the specifics.

Larger Configurations

Clusters larger than the three (3) racks of a Solution bundle, and/or Production cluster, must be designed, sized, and configured based on your requirements. Please work with your Dell sales representative to properly architect Production cluster deployments.

Solution Bundle Rack Layout

The Solution bundle includes three (3) storage nodes. These are set up in a Ceph cluster, which is tied into Cinder, Glance, and Nova.

See *Table 2: Controller Node Hardware Configurations – PowerEdge R630* on page 17 and *Table 5: Storage Node Hardware Configurations – PowerEdge R730xd* on page 18 for Hardware Configurations. The Solution Bundle includes:

- Node 1: R630 Solution Admin Host with the Red Hat OpenStack Manager Installed
- Nodes 2 - 4: R630 OpenStack Controllers
- Nodes 5 - 7 R630 Nova Compute Nodes
- Nodes 8 - 10: R730xd Storage Nodes

- EqualLogic Storage Array is optional for additional storage
- Network Switches: Two (2) Force10 S4810, and one (1) Force10 S55

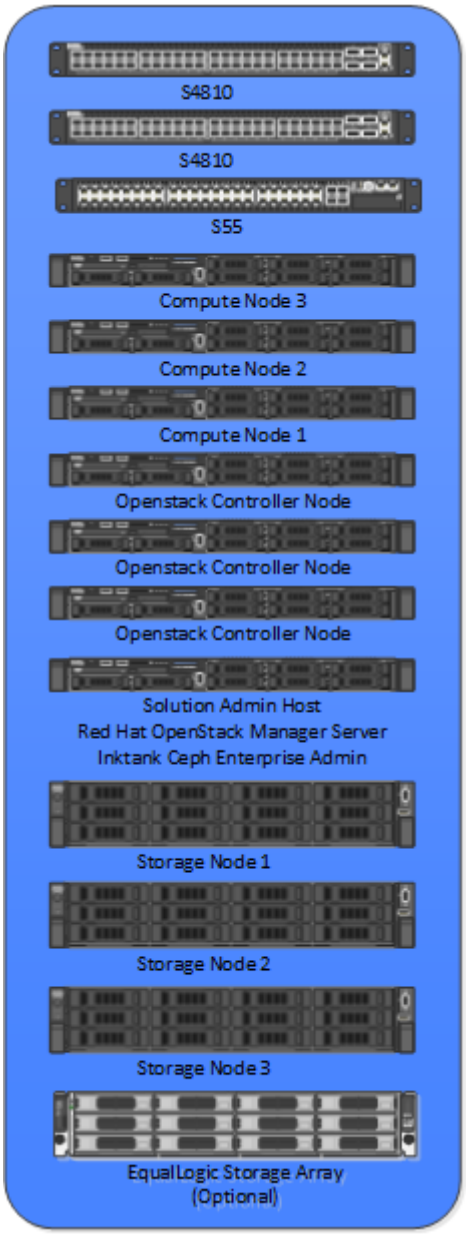

Figure 3: Solution Bundle and Ceph Cluster

The Ceph cluster provides data protection through replication, block device cloning, and snapshots. By default, the data is striped across the entire cluster, with three replicas of each data entity. The number of Storage nodes in a single cluster can scale to hundreds of nodes and many petabytes in size.

Ceph considers the physical placement (position) of Storage nodes within defined fault domains (i.e., rack, row, and data center) when deciding how data is replicated. This reduces the probability that a given failure results in the loss of more than one data replica.

There are two services in the Ceph storage cluster:

- **Object Storage Daemon (OSD)** - Running on Storage nodes, the OSD serves data to the Ceph clients from disks on the Storage nodes. Generally, there is one OSD process per disk drive.
- **Monitor (MON)** - Running on Controller nodes, the MON process is used by the Ceph clients and internal Ceph processes to determine the composition of the cluster and where data is located.

There should be a minimum of three MON processes for the Ceph cluster. The total number of MON processes should be odd.

Note: If MON processes on Controller nodes become a bottleneck, then additional MON processes can be added to the cluster by using dedicated machines, or by starting MON processes on Storage Nodes. A custom Services engagement can be arranged; contact your Dell representative for assistance.

The Storage Network vLAN is described in the Ceph documentation as the public network. The Storage Cluster Network vLAN is described in the Ceph documentation as the cluster network.

A special distribution of Ceph is used in this solution: Red Hat Ceph Enterprise v1.2.3, which also includes the Calamari Ceph cluster management client. The Red Hat Ceph Enterprise management client also includes Ceph troubleshooting and servicing tools and utilities. Red Hat Ceph Enterprise is installed on a virtual machine that runs on the Solution Admin Host (SAH). Note that:

- The SAH must have access to the Controller and Storage nodes through the Private API Access vLAN in order to manage Ceph; and for the monitoring process on all Storage nodes to return status and performance telemetry.
- The Controller nodes must have access to the Storage nodes through the Storage Network vLAN in order for the MON processes on the Controller nodes to be able to query the Ceph MON processes, for the cluster state and configuration.
- The Compute nodes must have access to the Storage nodes through the Storage Network vLAN in order for the Ceph client on that node to interact with the storage nodes, OSDs, and the Ceph MON processes.
- The Storage nodes must have access to the Storage Network vLAN, as previously stated, and to the Storage Cluster Network vLAN.

Solution Bundle Network Configuration

The network for this Solution has been designed to support production-ready servers with a highly available network configuration.

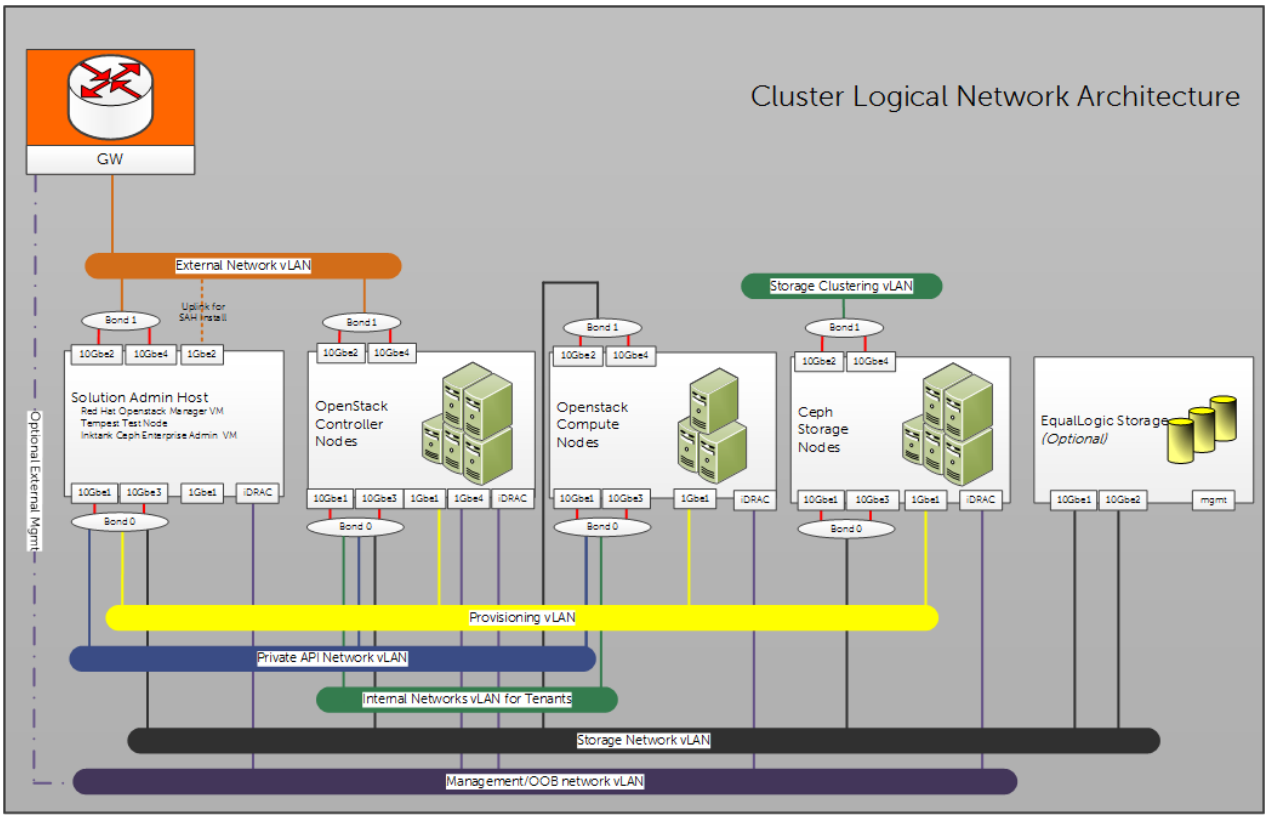

Figure 4: Cluster Network Logical Architecture

The node type will determine how the switches are configured in delivering the different networks. *Table 8: Node Type to Network 802.1q Tagging* on page 28 outlines the networks to the node type. The Management/OOB network is used by the Cluster Software to manage the OpenStack Controllers; therefore, they are the only ones that need direct connections. All iDRACs are plugged into this network without using tagging.

Table 8: Node Type to Network 802.1q Tagging

Network	Solution Admin Host	OpenStack Controller	OpenStack Compute	Ceph Storage	EqualLogic Array
Provisioning	Connected - Tagged	Connected - Untagged	Connected - Untagged	Connected - Untagged	Not Connected
External Network	Connected - Untagged[6]	Connected - Untagged	Not Connected	Not Connected	Not Connected
Private API Network Cluster Management	Connected - Tagged	Connected - Tagged	Connected - Tagged	Not Connected	Not Connected
Internal Tenant Networks	Not Connected	Connected - Tagged	Connected - Tagged	Not Connected	Not Connected
Storage Network	Connected - Tagged	Connected - Tagged	Connected - Untagged	Connected - Untagged	Connected - Untagged

[6] The 1Gbe port is used for provisioning the SAH during its installation only and can be disconnected afterwards.

Network	Solution Admin Host	OpenStack Controller	OpenStack Compute	Ceph Storage	EqualLogic Array
Storage Clustering	Not Connected	Not Connected	Not Connected	Connected - Untagged	Not Connected
Management/ OOB	Not Connected	Connected - Untagged	Not Connected	Not Connected	Not Connected
iDRAC physical connection to the Management/ OOB	Connected - Untagged	Connected - Untagged	Connected - Untagged	Connected - Untagged	Not Connected

Solution Bundle with Dell EqualLogic Storage added

The Solution Bundle with Dell EqualLogic Storage has the same characteristics; the only change is the Storage backend software is setup to use both Ceph and EqualLogic. The storage node servers are supplemented with one or more Dell EqualLogic arrays.

The Solution bundle shown has a single Dell EqualLogic Storage Array; this can be one or more depending on your Application and Storage needs. Prior to ordering your Sales Representative will work to find the proper configuration for your needs.

The controller nodes will use the Storage Network vLAN to access the Storage Pools created on the Storage Group for creation, deletion, and snapshots. The compute nodes must have access to the storage nodes through the Storage Network vLAN in order for the iSCSI driver on that node to interact with the volumes associated to Virtual Machines hosted by that node. EqualLogic Arrays are connected to the Storage Networking vLAN untagged only and all other nodes use the same layout as in *Table 8: Node Type to Network 802.1q Tagging* on page 28.

Solution Bundle: Solution Admin Host (SAH) Networking

The Solution Admin Host has internal bridged networks for the Virtual Machines.

The Solution Admin Host is physically connected to the following networks:

1. **Public Network** - used for:
   a. Inbound Access -
      a. HTTP/HTTPS access to the OpenStack Foreman Installer Node
      b. HTTP/HTTPS access to the Calamari Node
      c. Optional - SSH Access to the OpenStack Foreman Installer Node and Calamari Node
   b. Outbound Access
      a. HTTP/HTTPS access for Ceph and Red Hat Updates
      b. Used by the Tempest Test Node to run test using the OpenStack public API
2. **Provision Network** - Used by the OpenStack Foreman Installer Node to service DHCP to all hosts, provision each host, and act a proxy for the updates host.
3. **Private API Network** - Used by the Tempest Test Node to run tests against the OpenStack private API
4. **Storage Network** - Used by the Calamari Node to provision, monitor and manage the Ceph Cluster.

Figure 5: Solution Admin Host Internal Network Fabric on page 30 displays how the networks are bridged inside the Solution Admin Host. As the Provision, Private API, and Storage networks come in on the same physical interface, 802.1q tagging is configured on the SAH and corresponding switch ports. This Reference Architecture does not cover any security aspects, so the appropriate Network and Security teams should be involved before connecting any machine to the externally accessible networks.

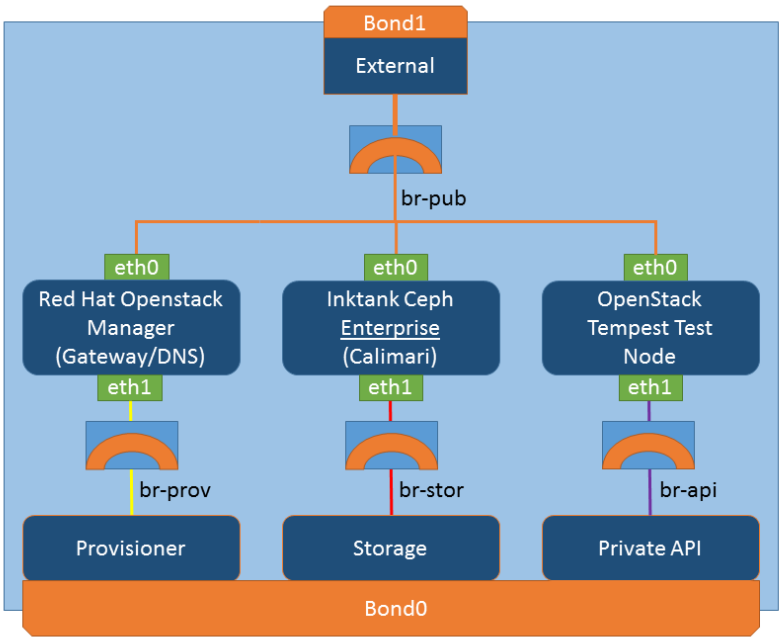

Figure 5: Solution Admin Host Internal Network Fabric

Solution Bundle Expansion

The Solution bundle can be expanded by adding Compute nodes, Storage nodes, or EqualLogic Arrays. Using this, one could expand up to twenty (20) servers per rack and/or 30 Us (Infrastructure, Compute, and Storage combined).

Expanding beyond the first rack will require the addition of aggregation network switches as described in the networking section, additional TOR and management switches in each rack, and the appropriate power and cooling. Expansion beyond a total of three (3) racks will need to be designed and configured based on your requirements. Please work with your Sales Representative to properly architect these large cluster deployments.

Note: When expanding the cluster, the Controller nodes can be expanded to more systems, but the expansion must be done in odd numbers only. For other expansion details, please speak with your Sales Representative.

Rack 1

Base Solution with Ceph, optionally EqualLogic:

- Solution Admin Host
- 3 Controller Nodes
- 3 Nova Compute Nodes
- 3 Storage Nodes

- Optional Dell EqualLogic Storage Arrays can be added

This configuration consists of a total of 15 U's, and a total of 12 servers (allowing up to 8 more servers in Rack 1). In rack 1 you could add up to either:

- 8 R630 Nova Compute Nodes, or
- 7 R730xd Storage Nodes or
- Dell EqualLogic Arrays

Note: You can use a combination of the three options that does not exceed a total of 20 servers or 30 rack units.

Rack 2

- 2 Z9000 or S4810 aggregation switches, depending upon your load requirements
- 2 S4810 TOR Switches
- 1 S55 Management Switch

Note: To split HA across the racks, you can move one or two Controllers from Rack 1 to Rack 2.

For additional nodes, you can add up to either:

- 19 R630 Nova Compute Nodes, or
- 14 R730xd Storage Nodes or EqualLogic Arrays, or
- Dell EqualLogic Arrays

Note: You can use a combination of the three options that does not exceed a total of 20 servers or 30 rack units.

Rack 3

- 2 S4810 TOR Switches
- 1 S55 Management Switch

Note: To split HA across the racks, you can move a Controller from Rack 1 or Rack 2 to Rack 3, giving one Controller per rack.

For additional nodes, you can add up to either:

- 19 R630 Nova Compute Nodes, or
- 14 R730xd Storage Nodes, or
- Dell EqualLogic Arrays

Note: You can use a combination of the three options that does not exceed a total of 20 servers or 30 rack units.

Moving the Controllers is not documented as part of the solution.

Update History

The following changes have been made to this guide:

Version 1

First Reference Architecture for the Dell Red Hat Cloud Solutions with Red Hat Enterprise Linux™ OpenStack Platform

Version 2

Update to support:

- Red Hat OpenStack Provisioning 5 Icehouse
- R620
- Ceph
- Cinder Multi-Backend and Multi-Instance
- Dell EqualLogic
- HA

Version 3

Updated as follows:

- New network diagrams
- Support for HA-only clusters
- Support for up to three (3) racks of equipment
- Renamed *Admin Node* to *Solution Admin Host*
- Added Virtual Servers to support Provisioning nodes
- Added optional Gateways for Provisioning/Storage/Management networks

Version 4

Updated as follows:

- New network diagram
- Added support for PowerEdge R630 and R730xd
- Removed support for PoweEdge R620, R720, and R720xd
- Added support for OpenStack Neutron
- Removed support for Nova-Network
- Removed the POC from the Solution
- Updated to Juno
- Updated to RHEL OSP 6.0
- Standardized terminology for platforms and nodes

Update 1

- New Network Diagram

    - SAH changes
        - 1Gbe to extrenral
        - Private API to bond
- Solution Admin Diagram
    - Added Private API for Tempest Test Node
    - Added Tempest Test Node
- Solution Admin Diagram Text
    - Added Private API for Tempest Test Node
    - Added Tempest Test Node to Public API Outbound
- Server Networking
    - Added discussion about the new bond modes and what is used by solution
    - Added table outline all modes and what can be used where.
        - Indicated the "solution default" on all modes

Update 2

- Removed support for Ceph Object Storage (Ceph Block Storage remains in the Solution)

References

Additional information can be obtained at *http://www.dell.com/openstack* or by e-mailing *openstack@dell.com*.

If you need additional services or implementation help, please contact your Dell sales representative.

To Learn More

For more information on the Dell Red Hat Cloud Solutions with Red Hat Enterprise Linux™ OpenStack Platform visit *http://www.dell.com/openstack*.