Dell Red Hat Cloud Solutions Reference Architecture Guide

A Dell Reference Architecture Guide

October 27, 2014

Contents

Tables

Figures

Trademarks

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.

Terminology and Abbreviations

Table 1: Terminology and Abbreviations

Term	Meaning
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 network gear and storage hardware needed to install 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	An operational model for managing data centers using automated deployments.
Hypervisor	Software that runs virtual machines (VMs).
Hyperscale	The ability of application and systems to scale out as demand increases. Typically found in large Cloud, Grid, and/or Distributed computing environments.
Infrastructure Node	The hardware configuration that best supports non-storage type roles, such as controller, hypervisor server, and database.
LAG	Link Aggregation Group.
LOM	LAN on motherboard.
Node	One of the servers in the system.
Pod	An installation composed of three racks, based upon server and network sizing.
Storage Admin Host	The physical host that supports the Red Hat OpenStack Manager and the Inktank Ceph Enterprise Admin.
Storage Node	The hardware configuration that best supports storage functions such as Swift and Ceph.

Overview

Approach

This Reference Architecture focuses on helping organizations to begin evaluating OpenStack® software and/or create proofs of concept. Dell and Red Hat can provide guidance for more sophisticated deployments; however, they are beyond the scope of this document. The Dell Red Hat Cloud Solutions with Red Hat Enterprise Linux™ OpenStack Platform encompasses software, hardware, operations, and integration in unique packages known as bundles. See Solution Bundles on page 25 for more information.

This Reference Architecture advocates an operational approach based upon highly automated solution deployments, using the components of the Dell Red Hat cloud solutions with RHEL OpenStack Platform. Dell believes that this operational model is the best practice for both initial cloud evaluations, and long-term maintenance of both moderate-scale and hyperscale data centers.

OpenStack Maturity

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

At publication the current release of OpenStack is codenamed "Icehouse". It builds upon previous releases with 350 new features to support software development, and managing data and application infrastructure, at scale. It is developed by over 1200 individuals employed by more than 120 organizations. Please see http://www.OpenStack.org/software/icehouse.

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

Taxonomy

This solution contains the core components of a typical Red Hat Enterprise Linux OpenStack Platform solution:

- Compute (Nova)
- Block Storage (Cinder with LVM, Ceph, EqualLogic)
- Image Service (Glance)
- Identity(Keystone)
- Dashboard (Horizon)
- Networking (Nova networking default; Neutron networking available)[1]
- Telemetry (Ceilometer)
- Orchestration (Heat)
- Object Store (Swift, available but not part of the solution[1])
- Database (Trove)

[1] Neutron networking and Swift are available via custom Services engagements, for customer evaluation only. Dell currently recommends using Nova networking for production deployments of this solution.

The taxonomy presented in Figure 1 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:

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

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.

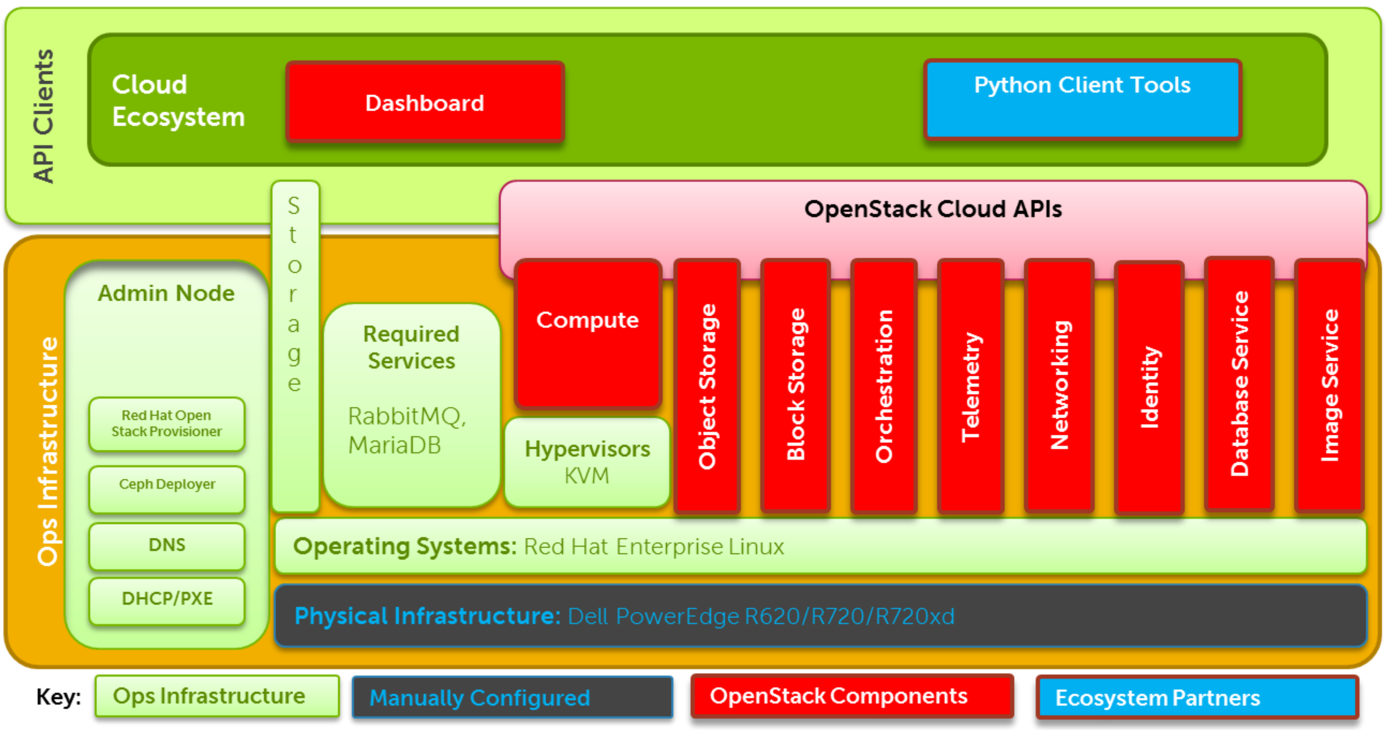


Figure 1: OpenStack Taxonomy

Hardware Options

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

- For evaluations, the recommended hardware is general-purpose, and enables a wide range of configuration options.
- For pilots, 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 repurpose the Reference Architecture configuration as your cloud deployment grows, so your investment is protected.

Networking and Network Services

This Reference Architecture uses a starter configuration for networking, which provides access to OpenStack capabilities while reducing the initial complexity during evaluation. After a review of the OpenStack Nova-Network and Neutron networking options, Dell has determined that Nova-Network is the best fit for initial deployments. Nova-Network's advantages are that it:

- Spreads the network load across all nova servers
- Keeps all routing and higher functions on the nodes themselves
- Eliminates the Network Controller as a single point of failure

For a production system, additional networking and redundancy configurations are required. This Reference Architecture supports adding these components as the system grows. This includes:

- Core and layered networking capabilities
- 10GbE networking
- NIC teaming
- Redundant trunking top-of-rack (ToR) switches into core routers

See Network Architecture on page 22 for guidelines. Detailed designs are available through Dell consulting services.

OpenStack Architecture

While OpenStack has many configurations and capabilities, we focus on the primary components for Red Hat Enterprise Linux® OpenStack Platform 5.0 (Icehouse), as Dell has defined in the above taxonomy.

For a complete overview of OpenStack software, visit Red Hat OpenStack Enterprise Platform (https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux_OpenStack_Platform/) and www.OpenStack.org.

OpenStack Components

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

Table 2: OpenStack Components

Function	Code Name	Description
Identity	Keystone	http://www.OpenStack.org/software/OpenStack-shared-services/ 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.
Dashboard/ Portal	Horizon	http://www.OpenStack.org/software/OpenStack-dashboard/ The 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.
Object Storage	Swift	http://www.OpenStack.org/software/OpenStack-storage/ OpenStack Object Storage (Swift) is open source software for creating redundant, scalable object storage using clusters of standardized servers to store petabytes of accessible data. It is not a file system or real-time data storage system, but rather a long-term storage system for a more permanent type of static data that can be retrieved, leveraged, and then updated if necessary. Primary examples of data that best fit this type of storage model are virtual machine images, photo storage, email storage, and backup archiving. Having no central "brain" or master point of control provides greater scalability, redundancy, and permanence. Objects are written to multiple hardware devices in the data center, with the OpenStack software responsible for ensuring data replication and integrity across the cluster. Storage clusters can scale horizontally by adding new nodes. Should a node fail, OpenStack works to replicate its content from other active nodes. Because OpenStack uses software logic to ensure data replication and distribution across different devices, inexpensive commodity hard drives and servers can be used in lieu of more expensive equipment.

Function	Code Name	Description
Compute/IaaS	Nova	http://www.OpenStack.org/software/OpenStack-compute/ 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.
Virtual Images	Glance	http://www.OpenStack.org/software/OpenStack-shared-services/ OpenStack Image Service (Glance) 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.
Block Storage	Cinder	http://www.OpenStack.org/software/OpenStack-storage/ OpenStack 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.
Networking	Neutron	http://www.OpenStack.org/software/OpenStack-networking/ 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.
Telemetry	Ceilometer	https://www.OpenStack.org/software/OpenStack-shared-services/ The OpenStack Telemetry service 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.

Function	Code Name	Description
Orchestration	Heat	https://www.OpenStack.org/software/OpenStack-shared-services/ 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.
Database Service	Trove	http://www.openstack.org/software/openstack-shared-services Designed to run entirely on OpenStack, the service has the goal of enabling users to quickly and easily utilize the features of a relational database without the burden of handling complex administrative tasks. Cloud users and database administrators can provision and manage multiple database instances as needed. Initially, the service will focus on providing resource isolation at high performance while automating complex administrative tasks including deployment, configuration, patching, backups, restores, and monitoring.

Red Hat Enterprise Linux OpenStack Platform 5.0

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 for 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 5 is built from the combination of Red Hat Enterprise Linux 7 and the latest Red Hat OpenStack technology, which is an enterprise-hardened version of the community Icehouse release. This version boasts all the community core features and functions, as well as some additional innovations by Red Hat.

Specifically, Red Hat Enterprise Linux OpenStack Platform 5 includes the latest features and functionality found in the Icehouse compute, networking, and storage services. In addition, version 5 includes improved support for VMware's vCenter driver for seamless side-by-side integration with existing VMware infrastructure, and easing the challenges of future migrations to OpenStack. As well, Red Hat will be supporting dual versions of Red Hat Enterprise Linux OpenStack Platform, supporting installation on both Red Hat Enterprise Linux 6 as well as the next generation operating system, Red Hat Enterprise Linux 7. Finally, in an effort to provide a more seamless infrastructure to customers, it is integrated with additional Red Hat infrastructure tools, including Red Hat Storage to provide optional object, block, and image storage services. Red Hat CloudForms for unified operations management, and Red Hat Enterprise Virtualization for customers who are looking for an alternative scale-up virtual infrastructure.

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 three (3) year life cycles, 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.
- Carrier-grade platform: Red Hat Enterprise Linux OpenStack Platform offers a carrier-grade, massively scalable platform, ideally suited for the telecommunications industry and their need for Network Functions Virtualization (NFV).

Storage Options

Cinder has introduced several backend options that can be used with the OpenStack solution. Dell's solutions include three options to enable the cluster to fit many needs:

- Ceph
- Dell EqualLogic Storage Arrays
- Logical Volume Manager (LVM)

Ceph

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

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

Logical Volume Manager (LVM)

Logical Volume Manager (LVM) is a basic service provided by the kernel that can be used to create a volume that is used in conjunction with the iSCSI drivers to create volumes and mount them to Virtual Machines. Cinder can use this method when other options are unavailable.

Table 3 Selecting your Storage Option

Storage Option	When to use
LVM	Use only for single-node evaluations and Proof of Concept. This is available only in the Dell PoC bundle. Not recommended for data requiring high availability, fault resiliency, spanning multiple nodes or scale-out model with redundancy.
Ceph	Use for low-cost, scale-out storage, ideal as volume backend for Nova VMs and also an object store Perfect for Dev/Test, back-up and storage-as-a-service use cases. Not recommended for very high-performance and database applications.
Dell EqualLogic	Use for higher-performance block storage backend requirements for Nova VM's and application data. Ideal for customers who already have EQL arrays and want to utilize them with OpenStack

Cinder Multi-Backend Support

In this Solution, it is now possible to take advantage of Cinder's Multi-Backend and Multi-Instance support and use different storage types to meet the requirements that the application requires. With this you can meet different use cases based on performance, large data transfers, temporary storage, and others. An example would be to create a Ceph cluster to have a performance group and a 2nd storage option of a Dell EqualLogic group to support large data transfers, then using Cinder multi-backend, configuring the Virtual Machine Volume requirements as required.

Server Infrastructure Options

The Solution includes using the PowerEdge R620, R720 and R720xd Server lines. The following sections describe the supported server models and configurations required. Detailed part lists and rack layouts are included in a Build of Materials Guide.

PowerEdge R620 Server

The PowerEdge R620 server is a hyper-dense, two-socket, 1U rack server. This server has a large memory footprint and impressive I/O options that make it an exceptional platform for space-sensitive environments. With its hyper-dense memory (up to 768GB in a 1U form factor) and the Intel® Xeon® processor E5-2600 or E5-2600 v2 product families, the R620 is perfect for virtualization, high-performance computing (HPC), and workgroup collaboration applications.

PowerEdge R720 and R720xd Server

The PowerEdge R720 and R720xd servers are Dell's 12G PowerEdge mainstream 2-socket 2U rack servers. They are designed to deliver the most competitive feature set, best performance, and best value. In this generation, Dell offers a large storage footprint, best-in-class I/O capabilities, and more advanced management features. The PowerEdge R720 and R720xd are technically similar, except that the R720xd has a backplane that can accommodate up to 12 3.5" or 24 2.5" drives.

Figure 2: PowerEdge 720xd Server

Advantages of the PowerEdge R720 over the PowerEdge R620 to meet the application needs include:

- Additional drive slots
- Additional PCI slots
- A larger selection of processors

Dell recommends that you use the PowerEdge R620 for Infrastructure and Compute nodes.

Hardware Configurations

Table 4: Compute and Infrastructure Node Hardware Configurations – PowerEdge R620 and R720

Machine Function	PoC Bundle Nodes	Pilot Bundle Infrastructure and Compute Nodes	Pilot Bundle Optional Compute Node
Platform	PowerEdge R620	PowerEdge R620	PowerEdge R720
CPU	2 x E5-2650v2 (8-core)	2 x E5-2670v2 (10-core)	2 x E5-2670v2 (10-core)
RAM (Minimum)	128 GB	128 GB	128 GB
LOM	4 x 1GbE	4 x 1GbE	4 x 1GbE
Add In Network	<none>	2 x Intel X520 DP 10Gb DA/SFP+	2 x Intel X520 DP 10Gb DA/SFP+
DISK	6 x 600GB 7.2K SATA 2.5 inch	6 x 1TB 7.2K SAS 2.5-inch	8 x 1TB 7.2K SAS 3.5-inch
Storage Controller	PERC H710	PERC H710	PERC H710
RAID	RAID 10	RAID 10	RAID 10

Table 5: Storage Node Hardware Configurations – PowerEdge R720xd

Machine Function	Pilot Bundle Performance Storage Nodes
Platform	PowerEdge R720xd
CPU	2 x E5-2650v2 (8-core)
RAM (Minimum)	32 GB
LOM	4 x 1GbE
Add In Network	2 x Intel X520 DP 10Gb DA/SFP+
DISK	Flex Bay: 2 X 300GB 10K 2.5-inch (OS) Front Drives: 2 X 200GB SSD 10 x 2TB or 4TB NL SAS 7.2K 3.5-inch
Storage Controller	PERC H710p
RAID	RAID 1 OS RAID 0 SSD RAID 0 per data disk

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 R620, R720 and R720Xd server configurations.

The R620, R720 and R720xd configurations can be used with 10GbE networking. To use 10GbE networking support, an additional network card is required in each node. Refer to the *Dell Red Hat Cloud Solutions Bill of Materials Guide* for the details on the supported card.

Operational Notes

Backup/Recovery

Since the system is designed for exploration that could later be extended to a production stage, 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 and 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 presents the recommended layout of each service.

The Red Hat OpenStack Manager and the Inktank 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 Upon
Infrastructure	Red Hat OpenStack Manager	Solution Admin Host (KVM)
Infrastructure	Database-server	OpenStack Controllers
Infrastructure	RabbitMQ-server (Messaging)	OpenStack Controllers
Infrastructure	Keystone-server	OpenStack Controllers
Infrastructure	Cinder-scheduler	OpenStack Controllers
Infrastructure	Cinder-volume	OpenStack Controllers
Infrastructure	Neutron-server(optional)	Neutron Controllers (Optional)
Infrastructure	Nova-Controller	OpenStack Controllers
Infrastructure	Nova-multi-compute	1 or more Compute Nodes
Infrastructure	Nova dashboard-server	OpenStack Controllers
Infrastructure	Ceilometer	OpenStack Controllers
Infrastructure	Heat	OpenStack Controllers
Infrastructure	Nova-Network	Same nodes as Nova-multi-compute
Infrastructure	HA-Proxy (Load Balancer)	OpenStack Controllers
Infrastructure	Pacemaker	OpenStack Controllers
Storage	Swift Object Store (optional)	3 or more Storage Servers
Storage	LVM Block Storage (optional)	For PoC only OpenStack Controller
Storage	Ceph Enterprise (optional)	3 or more Storage Servers
Storage	Ceph Monitor (optional)	3 or more across the Storage Nodes in the Pilot

		with no HA
Storage	Inktank Ceph Enterprise Admin (Calamari)	Solution Admin Host (KVM)
Storage	Dell EqualLogic Array (optional)	Dell EqualLogic Arrays

Deployment

Deployment consists of three phases:

- Hardware Setup:
  - Rack and stack
  - Cabling
  - BIOS configuration
  - RAID configuration
  - Switch configuration
- Software Setup:
  - Deploy Solution Admin Host for provisioning services
    - Deploy Ceph Service VM to the Solution Admin Host
    - Deploy Red Hat OpenStack Manager Virtual Server
  - Deploy controller node(s)
  - Deploy storage services
  - Deploy compute nodes
- Testing environment

Network Architecture

The Dell Red Hat cloud solutions with RHEL OpenStack Platform use 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.

This Reference Architecture implements at a minimum six (6), separate vLANs:

- **Management/Out of Band (OOB) network —** iDRAC connections can be routed to external network. All OpenStack HA controllers need direct access to this for IPMI operations.
- **Nova Network Private vLAN–** Sets up the backend network for nova and the VM's to use
- **Nova Network Public vLAN –** Sets up the front network for routable traffic to individual VMs
- **Provisioning Network vLAN**—Connects all nodes NICs into the fabric used for setup and provisioning of the servers.
- **Private API Network Cluster Management vLAN**— Used for communication between OS controllers and nodes for RESTFUL API and Cluster Heartbeat.
- **Public API Network Access vLAN** – Sets up access to the RESTFUL API, and the Horizon GUI.
- **Storage Network vLAN–** Used by all the nodes for data plane writes/reads to communicate to OS storage (not in POC)
- **Storage Clustering Network vLAN –** Used by all the storage nodes for replication and data checks (for Ceph Clustering)

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 can be one of four possible configurations:

1. Active-Active LAG in load-balance bond formation
2. Active-Backup in failover/failback formation
3. Active-Active round robin based on gratuitous ARP
4. Single port

In the first case the connectivity on the switch side must be in a LAG (or port-channel). In cases 2 and 3, Dell recommends that you do the configuration as a LAG but the ports should still be part of the same layer-2 domain. In some cases all members of the LAG connect to a single ToR switch. In others the LAG splits into two ToR switches. This is an optional setup as OpenStack has redundancy built into the application.

The teaming configuration that Dell recommends is *transmit-tlb* (mode = 5). This configuration setting is explained in greater deal in the *Dell Red Hat Cloud Solutions Deployment Guide*. Please contact your sales representative for a copy of the deployment guide.

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 the Force10 S55 for 1GbE and S4810 for 10G connectivity.

The 10GbE configuration utilizes Force10 S4810 switches as the ToR switches. Dell recommends for HA this pair of switches run Virtual Link Trunking (VLT). 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 80 Gb, using a LAG from each ToR switch. This is achieved by using two 40G interfaces in a LAG connecting to the aggregation pair. Therefore, a collective bandwidth of 160G 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.

Aggregation Switches

For a deployment of three to nine racks of 10G servers (12 racks max), 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 aggregation. 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 R620, R720 and R720xd 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

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. Red Hat Foreman OpenStack Manager requires a layer-2 domain in order to provision servers.

The three optional links (Provisioning, Storage, and Management) represent 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
- To be used by EqualLogic Host Integration toolset

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.

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.

Dell Open Switch 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. The customer is expected to ensure that the switches conform to these requirements, and that they are configured according this RA's guidelines. The following list of requirements will enable other brands of switches to properly operate with the tools and configurations in the *Dell Red Hat Cloud Solutions Reference Architecture*:

- 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 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 Bundles

In setting up deployments, Dell has found that there are two distinct user types. The first user type is the person or team that wants to use OpenStack in an environment that starts to duplicate production, with only enough hardware to prove the concepts that have been proposed. This team is looking to:

- Develop tools to use with OpenStack for day-to-day usage and management
- Develop applications that require actual hardware as the base
- Determine and investigate what networking issues will be seen in a production system

Dell recommends for this user type a PoC Bundle of 5 Servers and networking.

The second user type is the team that has already identified OpenStack as the platform of choice, has begun building tools to support OpenStack, and has applications ready for deployment. They are usually ready to design the first production implementation, with networking, Operating System and management. Dell recommends for the second user type a Pilot Bundle.

PoC Bundle

The PoC Bundle has been created for the customer who wants to start learning OpenStack and developing applications in a Proof of Concept environment. This bundle provides enough infrastructures to exercise OpenStack administration tasks, network configuration, application development, and to prove how to execute on a deployment plan.

Given a virtual machine with the characteristics of 2 cores, 4GB of memory and 40GB of ephemeral hard disk, you can expect to run around 70 virtual machines with a 1.5 oversubscription of CPU cores. At 90 virtual machines, you will have 2 to 1 CPU core oversubscription and still have ephemeral storage under subscribed and memory just starting to be oversubscribed. This solution only uses LVM storage on the OpenStack Controller node.

PoC Bundle Network Configuration

The PoC bundle configuration logical networking will look like Figure 3: PoC Networking Layout with no network connections requiring 802.1q vLAN tagging. The bundle includes a Dell Force10 S55 switch to which all nodes will be connected. This bundle is configured with NO network redundancy. The rack layout is similar to Figure 4: PoC Bundle. Please note that all nodes are stacked together, with a single switch above them.

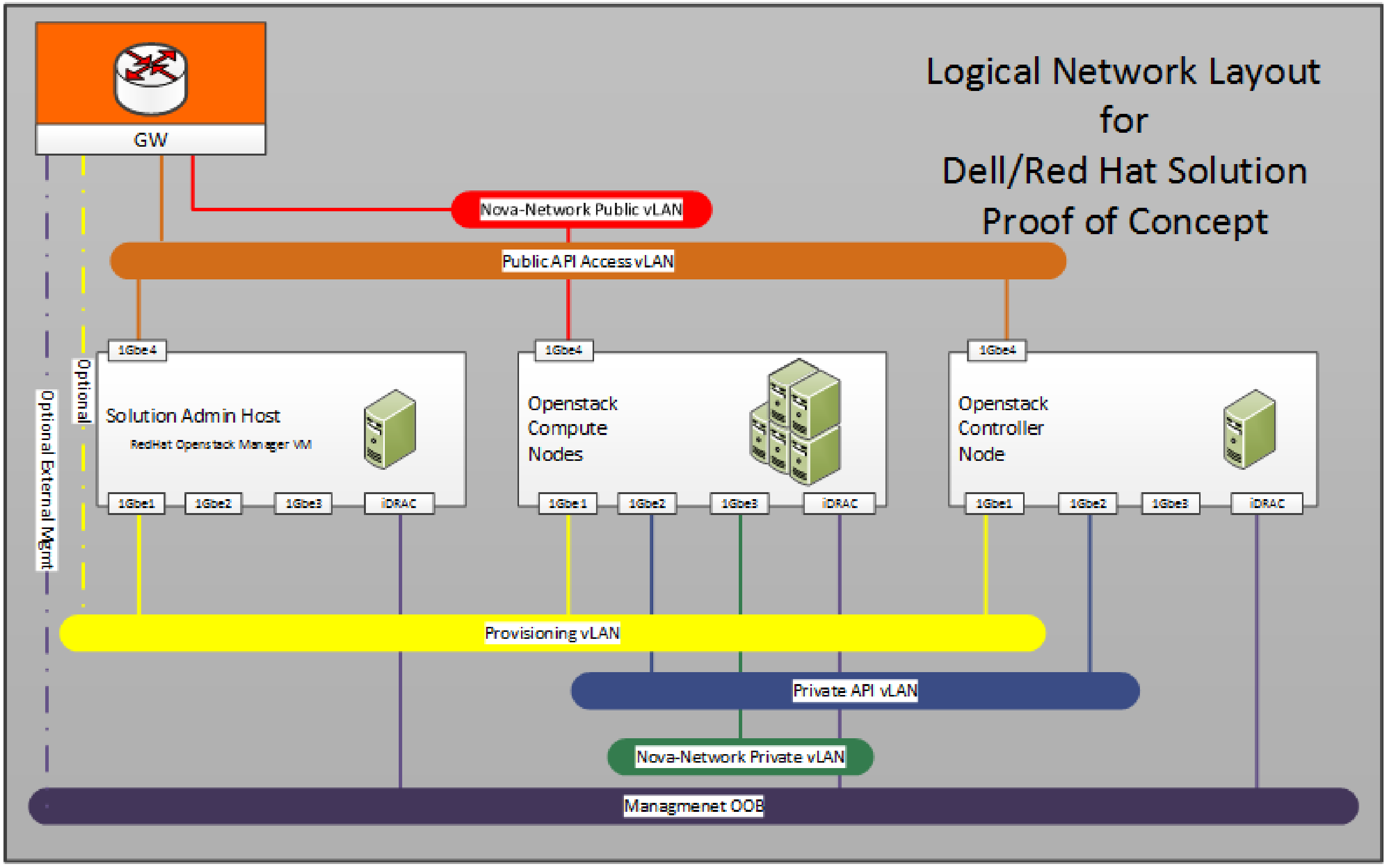
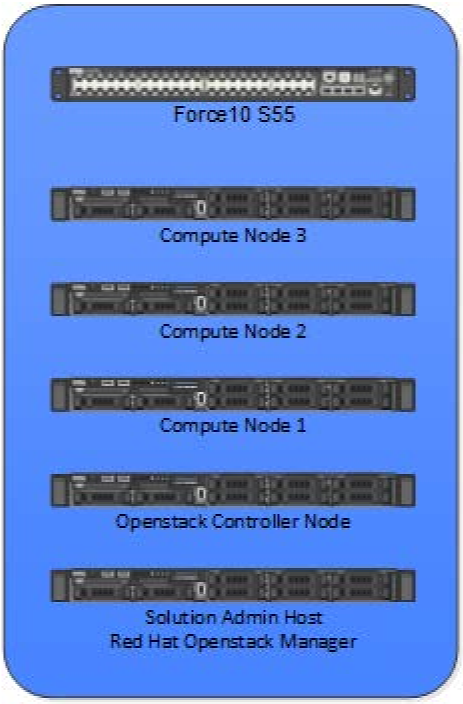


Figure 3: PoC Networking Layout

POC Bundle Hardware Layout

All nodes are R620 nodes with Dual 8 Core Xeon Processors, 4TB raw (2TB usable) drive space, 128 GB of memory and 1 Gbe Ethernet:



PoC Bundle

See Table 4 and Table 5 for Hardware Configurations

- Node 1: Solution Admin Host - Red Hat OpenStack Manager Server
- Node 2: OpenStack Controller
- Nodes 3-5: Nova Compute Nodes

Figure 4: PoC Bundle

Pilot Bundle

The Pilot Bundle is designed for pilot environments with storage. This bundle is for customers who want entry-level testing of an application, and the beginning of a production environment. This does not preclude the need to become familiar with OpenStack administration and day to day management. These bundles consist of the components described in Figure 5: Pilot Bundle on page 30.

Given a virtual machine with the same characteristics as above, you can expect to run around 170 virtual machines with a 1.5 oversubscription of CPU cores. At 228 virtual machines, you will have 2 to 1 CPU core oversubscription, still have ephemeral storage under subscribed, and memory just starting to be oversubscribed.

As the Pilot 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. Options for non-HA clusters are available, please review and discuss with your Sales Team the specifics to see if a non-HA bundle is appropriate.

Expansion of the Pilot Bundles

The Pilot bundles can be expanded using the compute bundle or the storage bundle. Using this one could expand up to twenty (20) servers per rack and/or 30 U's. (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.

Rack 1

Base Pilot with Ceph or EqualLogic:

- Solution Admin Host
- 3 Controller Nodes
- 5 Nova Compute Nodes
- 3 Storage Nodes or EqualLogic Arrays

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 8 R620 Nova computes

Or

- Up to 7 R720xd Storage nodes or EqualLogic Arrays

You can use a combination of the two options that does not exceed a total of twenty (20) servers or thirty (30) U's.

Rack 2

- 2 Z9000 or S4810 Aggregation switches depending on your load requirements
- 2 S4810 TOR
- 1 S55 Mgmt.

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

Other Nodes

- Add up to 19 R620 Nova Computes

Or

- Add up to 14 R720xd Storage Nodes or EqualLogic Arrays

You can use a combination of the two options that does not exceed a total of twenty (20) servers or thirty (30) U's.

Rack 3

- 2 S4810 TOR
- 1 S55 Mgmt.

To split HA across the racks, you can move a controller from Rack 1 to Rack 2 (one controller per rack).

Other Nodes

- Add up to 19 R620 Nova Computes

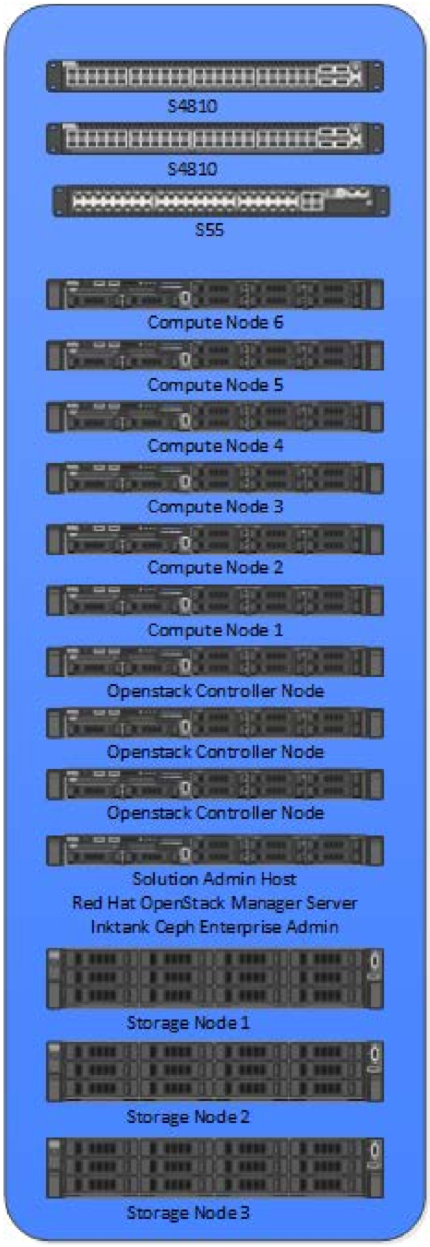
Or

- Add up to 14 R720xd Storage Nodes or EqualLogic Arrays

You can use a combination of the two options that does not exceed a total of twenty (20) servers or thirty (30) U's.

Pilot Bundle with Ceph Storage

The Pilot bundle has 3 storage nodes with it, on these you can setup a Ceph cluster, which can be tied into Cinder, Glance, and Nova. With Ceph you will need to plan out the different services.



Pilot Bundle with Ceph Storage

See Table 4 and Table 5 for Hardware Configurations

- Node 1 Solution Admin Host with the Red Hat OpenStack Manager Installed R620
- Node 2 - 4: OpenStack Controllers R620
- Nodes 5-10 Nova Compute Nodes R620 Pictured (or optionally R720)
- Nodes 11 – 13 Storage Nodes- These are the R720xd nodes added to this bundle to bring in 3 Storage nodes.
- Network Switches: Two (2) Force 10 S4810 and one (1) Force 10 S55

Figure 5: Pilot Bundle with Ceph Storage

The Ceph cluster provides data protection through replication, block device cloning, and snapshots. By default, the data is striped across the cluster. 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 of storage nodes when deciding how data is replicated. By defining fault domains, like a rack, row, and data center, and defining the Ceph position of the nodes and disks, Ceph using the position in the decision for data replication, thereby reducing the probability that a given failure results is the loss of more than one data replica.

There are two services in the Ceph storage cluster, the OSD and the MON (monitor). The OSD services data to the Ceph clients from disks on the storage nodes. Generally, there is one OSD process per disk drive. All OSDs run on the storage nodes. The MON process is used by the Ceph clients and other Ceph processes to determine the composition of the cluster and where data is located. There should be a minimum of three of these MON processes for the Ceph cluster. The MON processes are installed on all OpenStack controller nodes.

If the load of the clients querying the cluster causes the MON processes to take an inordinate amount of processing, then additional MON processes can be added to the cluster using dedicated machines, then retiring the MON processes co-located on either the Storage Node or OpenStack Controller nodes.

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

A special distribution of Ceph is used in this solution: Ceph Enterprise v1.2. That distribution uses the Firefly release of Ceph (v 0.81). Ceph Enterprise also includes the Calamari Ceph cluster management client. Ceph Enterprise is installed on a virtual machine that runs on the Solution Admin Host. The Solution Admin Host also includes Ceph troubleshooting and servicing tools and utilities.

The Solution Admin Host must have access to the controller, compute and storage nodes through the Private API Access vLAN in order to manage Ceph and for the monitoring process on the storage node 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 Ceph clients on the controller nodes to be able to query the Ceph MON processes for the state and configuration of the cluster.

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 and the OSDs and the Ceph MON processes.

The storage nodes must have access to the Storage Network as previously stated and to the Storage Cluster Network vLAN.

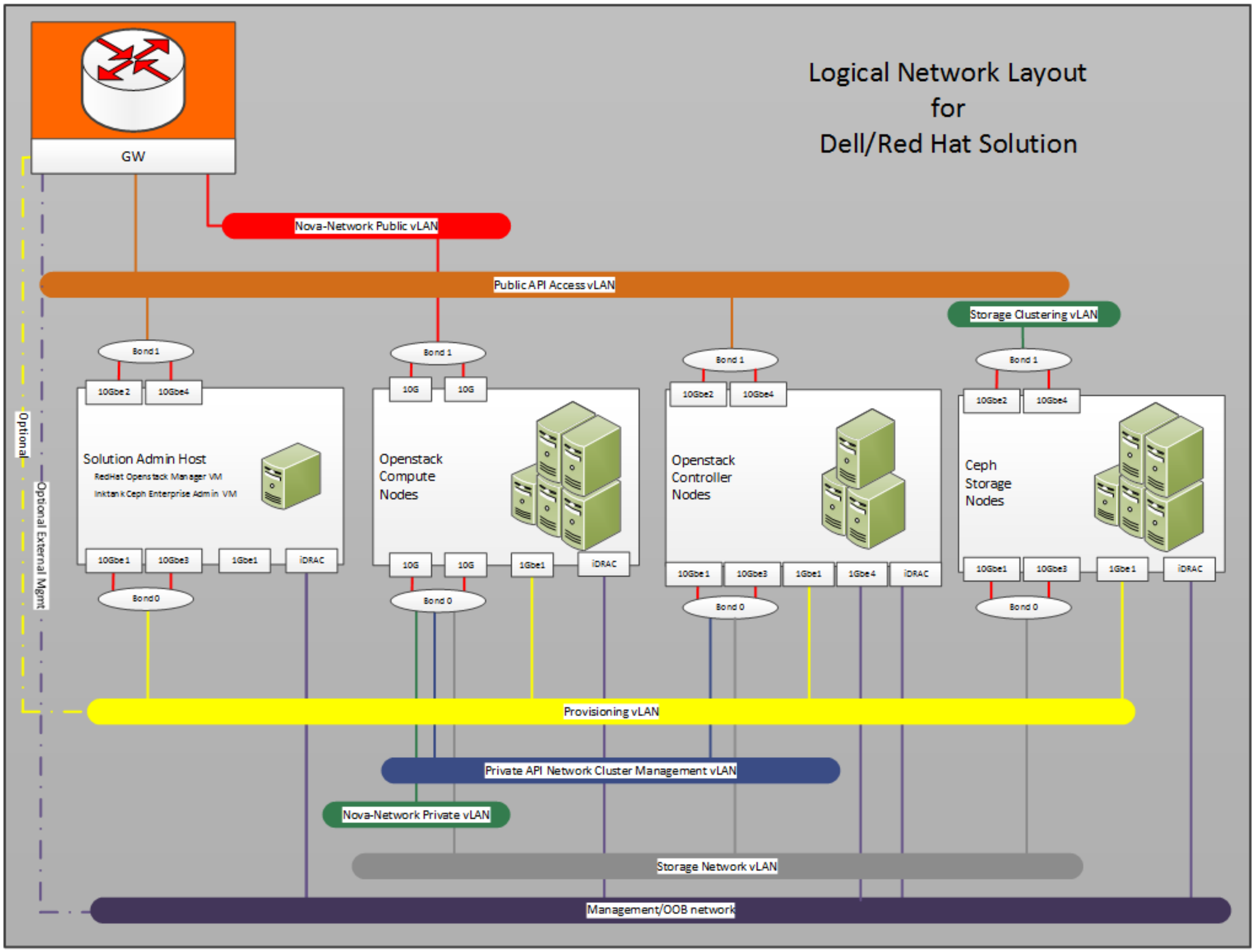


Figure 6: Pilot Logical Network with Ceph

Pilot Bundle with Ceph Network Configuration

The networks for this Pilot and Larger bundles have been designed to support production-ready servers with a highly available network configuration. The rack layout consists of two Dell Force 10 S4810 switches as the ToR, and one Dell Force 10 S55 as management (a second Force 10 S55 should be considered for the Heartbeat to be truly redundant). The switches take advantage of Virtual Link Trunking (VLT). Using additional Dell Force 10 features, the ports can be configured to support one or many virtual LANs; enabling the nodes to communicate between themselves, and allowing network segregation of tenants utilizing vLANs.

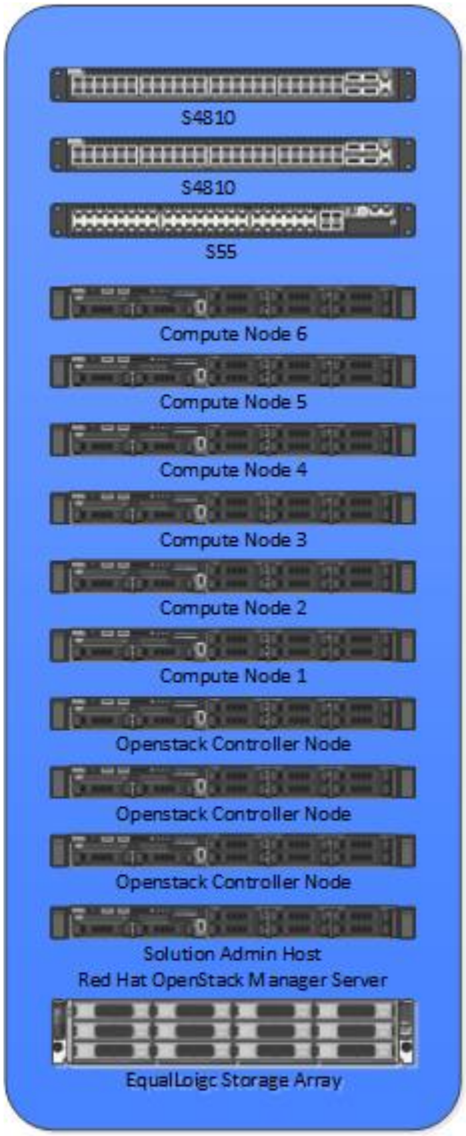
The Server function will determine how the switches are configured in delivering the different networks. **Table** 7 **Node Type to Network 802.1q Tagging** outlines the networks to the server function. The Management/OOB network is used by the Cluster Software to manage the OpenStack Controllers, thus they are the only ones that need direct connections. All iDRAC are plugged into this network without using tagging.

Table 7 Node Type to Network 802.1q Tagging				
	Solution Admin Host	**OpenStack Controller**	**OpenStack Compute**	**Ceph Storage**
Provisioning	Not Tagged	Not Tagged	Not Tagged	Not Tagged
Nova-Network Public	Not Connected	Not Connected	Not Tagged	Not Connected
Public API Access	No Tagging	No Tagging	Not Connected	Not Connected
Private API Network Cluster Mgmt	Not Connected	Tagged	Tagged	Not Connected
Nova-Network Private	Not Connected	Not Connected	Tagged	Not Connected
Storage Network	Not Connected	Tagged	Tagged	Not Tagged
Storage Clustering	Not Connected	Not Connected	Not Connected	Not Tagged
Management/OOB	Not Connected	Not Tagged	Not Connected	Not Connected

Pilot Bundle with Dell EqualLogic Storage

The Pilot Bundle with Dell EqualLogic Storage has the same characteristics as the Pilot Bundle with Ceph Storage; the only change is the Storage backend software, hardware and networking. The storage node servers are replaced with one or more Dell EqualLogic arrays and the networks are configured to support the Dell EqualLogic array(s).

The Pilot bundle shown here 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.



Pilot Bundle with EqualLogic Storage

See Table 4 and Table 5 for Hardware Configurations

- Node 1 Solution Admin Host with the Red Hat OpenStack Manager Installed R620
- Node 2 - 4: OpenStack Controllers R620
- Nodes 5-10 Nova Compute Nodes R620 Pictured (or optionally R720)
- EqualLogic Storage Array(s) – See Sales Representative to find the proper array for your applications.
- Network Switches: Two (2) Force 10 S4810 and one (1) Force 10 S55

Figure 7: Pilot Bundle with Dell EqualLogic Storage

Pilot Bundle with EqualLogic Storage Network Layout

The OpenStack controllers hosting Cinder and/or Glance servers and compute nodes must have access to the Dell EqualLogic Storage Group through the Storage network vLAN. The controller nodes will use the network 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** 7 **Node Type to Network 802.1q Tagging**

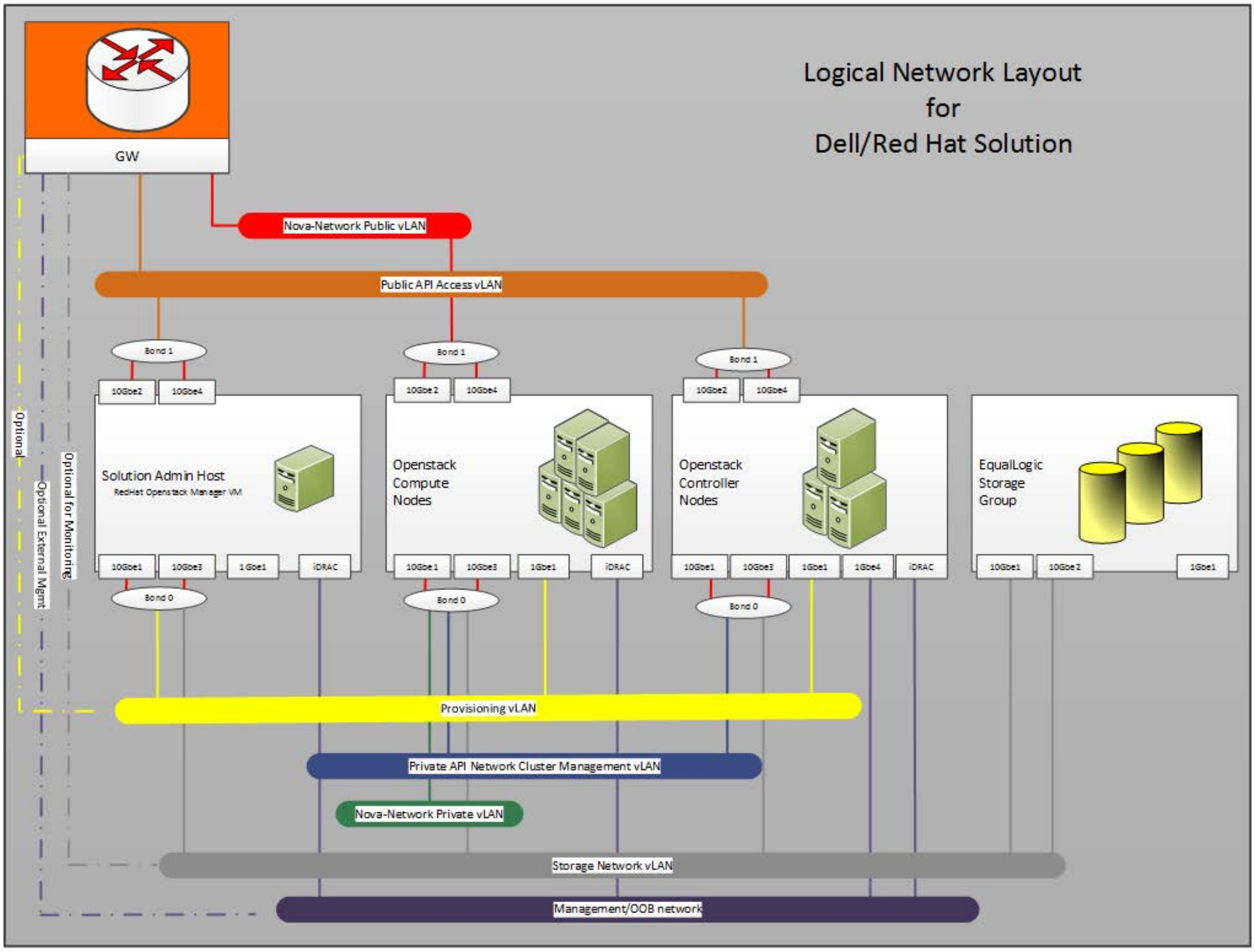


Figure 8 Logical Network with EqualLogic Storage

Production Bundle

Clusters larger than the 3 Racks of a Pilot bundle and/or Production Cluster must be designed, sized and configured based on your requirements. Please work with your Sales Representative to properly architect Production cluster deployments.

Update History

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 with:

- New Network Diagrams
- Support of HA only clusters
- Support of up to three racks of equipment
- Renamed Admin Node to Solution Admin Host
- Added Virtual Servers for the support of Provisioning nodes
- Added optional Gateways for Provisioning/Storage/Management networks

Getting Help

Contacting Dell

For customers in the United States, call 800-WWW-DELL (800-999-3355).

If you do not have an active Internet connection, you can find contact information on your purchase invoice, packing slip, bill, or Dell product catalog.

Dell provides several online and telephone-based support and service options. Availability varies by country and product, and some services may not be available in your area. To contact Dell for sales, technical support, or customer service issues:

- Visit support.dell.com.
- Click your country/region at the bottom of the page. For a full listing of country/region, click **All**.
- Click **All Support** from the **Support** menu.
- Select the appropriate service or support link based on your need.

Choose the method of contacting Dell that is convenient for you.

References

For more information about deploying OpenStack in your solution portfolio, see:

- Red Hat® Enterprise Linux® OpenStack® Platform Deployment Guide
- Inktank Ceph Enterprise

To Learn More

For more information on the ell Red Hat Cloud Solutions with Red Hat Enterprise Linux™ OpenStack Platform, visit:

www.dell.com/OpenStack