



Mirantis OpenStack Reference Architecture for Dell Hardware

Table of Contents

Table of Contents	
Glossary	
Trademarks	6
Overview	6
Use Case Summary	
Mirantis OpenStack	
Hardware Options	
Networking	
Key Benefits	
Mirantis OpenStack Architecture	
Mirantis OpenStack overview	
Node types	10
Infrastructure Node	10
Controller Node	10
Compute Node	
Storage Node	
High Availability	
Optional components	
Network Architecture	
Network design	
Cabling	
Rack Physical Cabling Schema	
Server Interface Configuration	
Rack hardware specification	20
Servers	
PowerEdge R630	
PowerEdge R730xd	
Recommended Server Configurations	
Switches	
Additional Rack equipment	
Disk Configuration	
Compute node	
Controller node	
Storage node	
Infractructure node	7/

Validated Configuration	25
Scaling	
Suggested Rack Configurations	
Testing	
References	27

Glossary

Term	Description		
ВМС	The baseboard management controller		
Bonding	A virtual network interface consisting of aggregated physical links		
Ceph	A distributed block store, object store, and file system		
DVR	Distributed Virtual Router – a way of configuring and running Neutron L3 agent on every Compute node in order to move load from Controller nodes. Each Compute node serves East-West routing and North-South Floating IP translation for local instances while Controller nodes serve only North-South traffic for those instances which don't have Floating IPs assigned.		
East-West traffic	Traffic within a cloud. i.e. VM to VM or tenant to tenant traffic.		
Ephemeral storage	A non-persistent disk or volume used by VM		
Floating IP	An IP address that can be instantly moved from one VM to another. Floating IP addresses enable dynamic IP address association with virtual machine instances.		
Fuel	An open source project built to deploy OpenStack. A part of OpenStack Big Tent.		
Fuel plugin	A special package which extends Fuel's functionality. For more details see https://wiki.openstack.org/wiki/Fuel/Plugins		
IPMI	Intelligent Platform Management Interface		
LACP	Link Aggregation Control Protocol - an international standard of network interfaces bonding supported by almost all vendors and operating systems. The current specification is IEEE 802.1AX-2008.		
LMA	Acronym for Logging, Monitoring, Alerting.		

LMA toolchain	A set of tools to collect and analyze logs, monitor health and status of OpenStack components, and raise alerts if needed. It consists of Elasticsearch-Kibana, InfluxDB-Grafana, and Nagios.
MOS	Mirantis OpenStack
Multi-chassis Link Aggregation	A specific configuration of network switches when two or more independent switches act as a single one from LACP standpoint.
NAT	Network address translation - a special mechanism allows hosts from one network reach out another network without having proper routing established.
North-South traffic	Traffic between cloud VMs and the rest of the network (anything outside the cloud).
os	Operating system
RA	Reference Architecture
Rally	An open source tool to perform cloud verification, benchmarking and profiling.
Shaker	A wrapper around popular network performance testing tools to measure network performance of a cloud.
Tempest	A set of integration tests to be run against a live OpenStack cloud.
TLS	Transport Layer Security protocol for securing network connections.

Trademarks

Trademarks used in this text: Dell™, the DELL logo and PowerEdge™ are trademarks of Dell, Inc. Intel® and Xeon® are registered trademarks of Intel Corporation in the U.S. and other countries. Linux® is the registered trademark of Linus Torvalds in the U.S. and 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 either to the entities claiming the marks and names, or to their products. Mirantis, Inc., disclaims any proprietary interest in trademarks and trade names other than its own.

Overview

This document describes *Mirantis OpenStack for Cloud Native Apps with Dell* — a fully-validated deployment of Mirantis OpenStack on Dell® PowerEdge® R630 and R730xd servers networked with Dell Networking S3048-ON and S4048-ON switches.

The deployment is engineered as a scalable, rack-based OpenStack dev/test/production environment for cloud-native web applications.

This Reference Architecture details all hardware and software components of *Mirantis OpenStack for Cloud Native Apps with Dell;* describes their physical integration and logical interoperation; and discusses in depth certain critical aspects of the design relevant to performance, high availability, data-loss prevention, and scaling.

Use Case Summary

This Reference Architecture (RA) is designed for cloud native applications. It comes with the Murano catalog preloaded to support deployment of developer tools, databases, CI/CD toolchain elements (e.g., Jenkins), leading PaaS solutions (e.g., Cloud Foundry), containers and orchestrators (e.g., Docker, Kubernetes) and other solutions, including those available in the online OpenStack Community App Catalog. Neutron networking allows for self-service networking isolated by tenants while Ceph provides both block and object storage to support various application types.

Mirantis OpenStack

Mirantis is a number one contributor to OpenStack. Fuel — a core part of the Mirantis OpenStack distribution —has been included under Big Tent.

Mirantis recommends the following configuration of Mirantis OpenStack:

- Mirantis OpenStack version 7.0, based on OpenStack Kilo.
- Ubuntu 14.04 as host OS.
- The following OpenStack services deployed: Keystone, Nova, Glance, Cinder, Neutron, Horizon, Heat, Murano, Ceilometer.
- Three OpenStack Controller nodes in an HA configuration using HAproxy, Corosync/Pacemaker, MySQL Galera, RabbitMQ, MongoDB.
- A Ceph cluster provides a redundant backend for Glance, Cinder, Nova Ephemeral, and Swift-API
 - Ephemeral volumes configured to use Ceph backend to provide large volumes and support live migration
 - o 3x replication is utilized to protect data
- Secure public OpenStack API endpoints and Horizon with TLSv1.2.
- Zabbix or LMA toolchain monitoring services installed.
- Sahara (optional).

Hardware Options

For this Reference Architecture, the Dell PowerEdge R630 and R730xd servers were chosen because of their hardware options, performance, and scalability. Both platforms are powered by the Intel® Xeon® processor

E5-2600 v3 product family, which is able to utilize up to 24 DIMMS of DDR4 RAM, and supports the 12Gb PowerEdge RAID controllers (PERC9). These features, combined with a high-performance energy-efficient design, should serve well for a variety of workloads your cloud infrastructure will need to support.

The server configurations recommended are well-tested and proven with the Mirantis OpenStack for Cloud Native Apps reference design, creating a foundation that can secure your infrastructure investment and allow for easy expansion of your growing cloud.

Networking

The network layer is based on Neutron with OpenvSwitch and VLANs for tenant traffic separation because this configuration eliminates the cost of an SDN solution.

10GbE networking infrastructure plus the Neutron DVR feature enable high performance East-West traffic.

Designated NICs used for Storage replication traffic reduce pressure from tenant traffic. NIC bonding and multi-chassis LAG across multiple switches for uplinks and downlinks provide both high performance and physical redundancy for the whole network infrastructure.

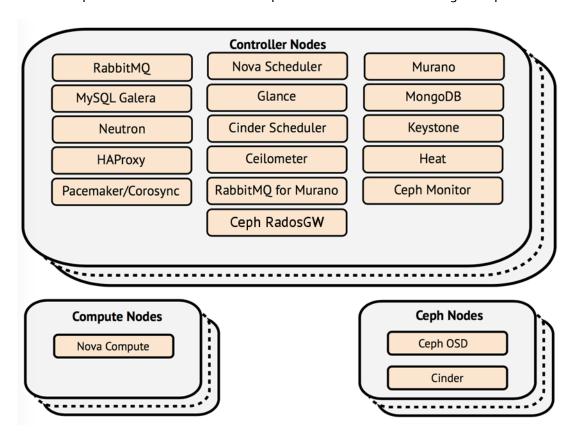
Key Benefits

- Easy to deploy integrated solution by Mirantis on Dell PowerEdge Servers.
- Flexible enough to cover a broad set of use cases.
- Based on best-of-breed, enterprise-ready Mirantis OpenStack and Dell PowerEdge Servers.

Mirantis OpenStack Architecture

Mirantis OpenStack overview

Mirantis OpenStack 7.0 is a Kilo based OpenStack distribution running on top of Ubuntu 14.04.



Mirantis OpenStack consists of the following "core" OpenStack services:

- Keystone
- Nova
- Neutron
- Cinder
- Glance

The Swift API is provided by Ceph with RadosGW.

The following "optional" OpenStack services are installed in Mirantis OpenStack:

- Horizon
- Heat
- Murano
- Ceilometer

All services except Ceilometer are using local MySQL as a database backend.

Ceilometer uses MongoDB.

All services are using RabbitMQ as a messaging queue service.

Ceph provides the Swift API via RadosGW and is also used as a backend for Cinder and Glance.

Nova Ephemeral storage may use local disks or Ceph. The local disk gives you better performance but Ceph enables live-migration for instances with ephemeral storage used. Placement of Ephemeral storage to local disks is considered as technical preview.

To learn more, please see the description of the Mirantis OpenStack architecture at https://docs.mirantis.com/openstack/fuel/fuel-7.0/reference-architecture.html.

Node types

Servers used in this Reference Architecture will serve as one of these node types: Infrastructure, Controller, Compute, or Storage.

Infrastructure Node

The Infrastructure node is an Ubuntu 14.04-based node which carries two virtual appliances:

- Fuel Master node an OpenStack deployment tool.
- ◆ Cloud Validation node a set of OpenStack post-deployment validation tools including Tempest and Rally.

Controller Node

The Controller node is a control plane component of a cloud which incorporates all core OpenStack infrastructure services such as MySQL, RabbitMQ, HAproxy, OpenStack APIs, Horizon, and MongoDB. This node is not used to run VMs.

Compute Node

The Compute node is a hypervisor component of a cloud which runs virtual instances.

Storage Node

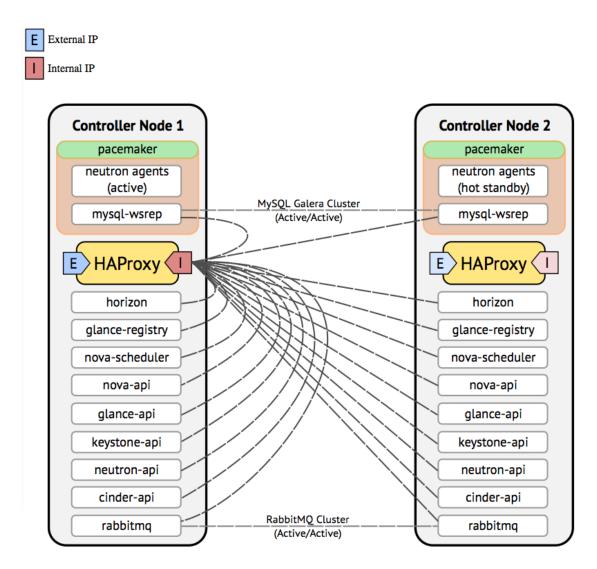
The storage node is a component of an OpenStack environment which keeps and replicates all user data stored in your cloud including object and block storage. Ceph is used as a storage back end.

High Availability

The High Availability model implemented in Mirantis OpenStack (MOS) is well described in the official documentation: https://docs.mirantis.com/openstack/fuel/fuel-7.0/reference-architecture.html#multi-node-with-ha-deployment, but here is a short description.

To protect core OpenStack services from failure of a Controller node, control plane components are duplicated on multiple Controller nodes. It's possible to deploy 3, 5, 7 or any other odd number of Controller nodes and add more nodes after the initial deployment (but keeping an odd total number) to distribute load and increase the redundancy level. Pacemaker/Corosync is used to manage critical infrastructure parts when failure occurs. Because of the quorum majority voting algorithm used in Corosync, a cluster of 'N' Controller nodes may survive when 'N/2-1' nodes go down. For 3 nodes it's 1 node down, for 5 nodes it's 2 nodes, and so on.

Almost all services are working in active-active mode and load is balanced by HAproxy listening at a virtual IP address. Openstack components support active/active mode natively, as do RabbitMQ and MongoDB. Galera Cluster is used to protect MySQL.



From the networking perspective, high availability is upheld by redundant and bonded connections to two ToR switches. That means if one cable gets cut or an entire switch goes down, a cloud will survive seamlessly with half its normal network bandwidth still available.

Another low-level hardware feature that provides redundancy is RAID1 for disks on Infrastructure and Storage nodes to keep an operating system intact in case of a disk failure. Also it's highly recommended to use two independent hot-plug power supplies for each hardware piece. This helps protect components from power supply failures.

Ceph provides redundancy for stored data. The recommended replication number is 3 which means at a given time there are 3 copies of the same data spread across Ceph nodes.

Optional components

There are two possible ways to use MongoDB with Ceilometer: install the database locally or externally. Both options are available in Mirantis OpenStack. When installing MongoDB locally we recommend using at least 3 separate servers to form a MongoDB cluster. Placing MongoDB onto Controller nodes may cause resource starvation for key services such as MySQL or RabbitMQ, which may lead to severe issues with the cloud as a whole. The best approach is to use dedicated nodes for MongoDB; either external ones or nodes deployed as part of MOS.

For validation purposes, we put MongoDB on the Controllers, but give it a separate physical disk on each node. Each controller node contains two SSDs: the first one for the OS plus other services and the second one for MongoDB alone.

Sahara is a service used to deploy Big Data (i.e. Hadoop) clusters on top of OpenStack that may be installed on demand. Sahara wasn't used during the validation.

By default, Mirantis OpenStack protects public OpenStack APIs and Horizon with TLSv1.2 using self-signed or user-provided certificates, but it's possible to disable this feature if necessary. For validation purposes, we used self-signed certificates.

Mirantis OpenStack may be extended via Fuel plugins to provide two validated options to monitor your cloud: Zabbix or LMA toolchain (Elasticsearch/Kibana, InfluxDB/Grafana and Nagios). Neither was used during this validation.

Network Architecture

The underlying network consists of OpenStack control plane, data plane, and BMC Management. A pair of ToR S4048-ON switches in each rack forms a Virtual Link Trunking (VLT) group and provides for OpenStack control and data plane traffic and uplinks. VLT allows servers' bonded NICs and rack uplinks to be connected to both switches (opposite to traditional LACP), thus every connection is secured against physical link or entire switch failure.

S3048-ON serves BMC Management network 1Gbps connections, including Dell iDRAC, switch management traffic, and OpenStack Provisioning traffic.

Ceph IO and monitoring traffic goes on the OpenStack Management network and is secured with LACP.

It's recommended to run Ceph replication traffic separately and secure it with LACP along with the OpenStack Management network. We did this during validation.

Depending on intended load, you may or may not elect to run the OpenStack Private network over LACP. Alternatively, you may wish to enable LACP for some Compute nodes but not others, providing different SLAs for these loads. We ran the Private network over LACP during validation.

Network design

The networking model of Mirantis OpenStack requires the following networks to be set up:

Network	VLAN IDs1
BMC Management networks:	
Mirantis OpenStack Admin/PXE (OpenStack Provisioning) Network	120
Out of Band IPMI network	100
OpenStack Control plane networks:	
OpenStack Management and Ceph Public network	140
OpenStack Public network	160
OpenStack Ceph Replication network	180
Out of Band IPMI network	100
OpenStack data plane network:	
OpenStack Private network	200-1000

To learn more about Mirantis OpenStack's networking model see https://docs.mirantis.com/openstack/fuel-7.0/reference-architecture.html// network-architecture

We recommend deploying Mirantis OpenStack with Neutron DVR enabled. It runs a standalone router on every Compute node. It routes East-West and North-South traffic of Floating IPs for local instances, dramatically decreasing load to Controller nodes, which serve only North-South NAT traffic in this case. To learn more about DVR see https://docs.mirantis.com/openstack/fuel/fuel-7.0/reference-architecture.html#neutron-with-dvr

We had DVR deployed during validation.

¹ These VLAN numbers are given for reference.

² Note: Each Storage node has two network cards with two ports each. Each bonded interface comprises two ports.

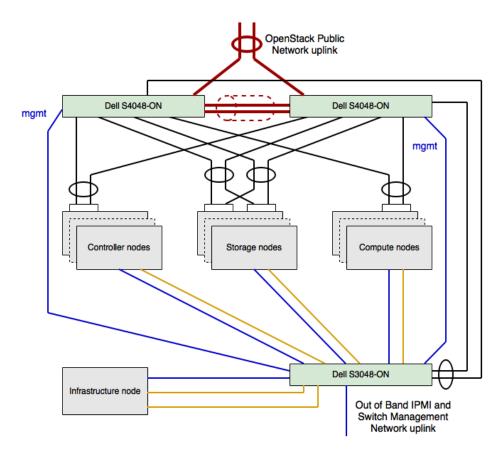
The following subnets are used by Mirantis OpenStack by default.

Network	Subnet Details
Mirantis OpenStack PXE Network	172.16.1.0/24*
OpenStack Management and Ceph Public Network	172.17.0.0/24*
OpenStack Public Network	Depends on requirements. Minimum /26 network is recommended. To calculate the required size of Public Network see https://docs.mirantis.com/openstack/fuel/fuel-7.0/reference-architecture.html#public-and-floating-ip-address-requirements
OpenStack Private Network	For validation purposes a single tenant network will be created during the deployment with the network address 192.168.122.0/24* It's possible to remove the network after the validation is complete.
OpenStack Ceph Replication Network	172.18.0.0/24*
Out of Band IPMI and Switch Management Network	No specific requirements

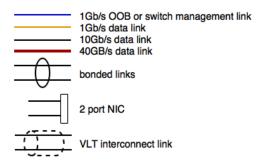
^{*} This network may be fully isolated inside a cloud and doesn't need to be routed to a customer's network.

Cabling

Rack Physical Cabling Schema



Legend:



Server Interface Configuration

Controller Node

Interface	1st 1G	1st & 2nd 10G bond0	IPMI	
VLANs	120 (untagged)	Active/Active LACP	100 (untagged)	
		140, 160, 180, 200- 1000 (tagged)		

Compute Node

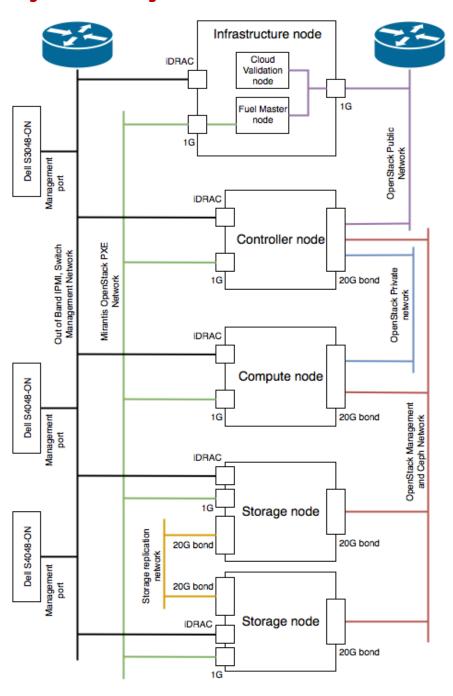
Interface	1st 1G	1st & 2nd 10G bond0	IPMI
VLANs	120 (untagged)	Active/Active LACP	100 (untagged)
		140, 160, 180, 200- 1000 (tagged)	

Storage Node

Interface	1st 1G	1st & 3nd 10G bond0 ²	2rd & 4th 10G bond1 ²	IPMI
VLANs	120 (untagged)	Active/Active LACP 140, 160, 200-1000 (tagged)	Active/Active LACP 180 (tagged)	100 (untagged)

² Note: Storage node has two network cards with two ports each. And each bonded interface consists of two ports from different physical network cards to provide protection against network card failure.

Logical Networking Schema



Rack hardware specification

Servers

This Reference Architecture covers Dell PowerEdge R630 and R730xd server configurations but it's possible to use other server configurations. Please contact your Dell sales representative for available options.

PowerEdge R630

The PowerEdge R630 two-socket rack server delivers uncompromising density and productivity. Part of the 13th generation of PowerEdge servers, the R630 is ideal for virtualization. The processor and memory density with up to 24 DIMMs of DDR4 RAM provides great memory bandwidth.

PowerEdge R730xd

The incredible versatility of the PowerEdge R730 server delivers outstanding functionality in just 2U of rack space. With the Intel® Xeon® processor E5-2600 v3 product family and up to 24 DIMMs of DDR4 RAM, the R730 has the processing cycles and threads necessary to deliver more, larger and higher-performing storage for virtual machines. Highly scalable storage, with up to sixteen 12Gb SAS drives and the high-performance 12Gb PowerEdge RAID Controller (PERC9), can greatly accelerate data access for your virtualized environment.

Recommended Server Configurations

	Controller Node	Compute Node	Storage Node	Infrastructure Node
Server Model	R630	R630	R730xd	R630
Sockets	2	2	2	2
Cores per socket	8	8	6	8
RAM Configuration	256 GB	256 GB	256 GB	256 GB
Storage Configuration	2x 400GB SSD disks	1x 400GB SATA disk	4x 200GB SATA Intel® DCS3710 20x 1.2TB SAS 2 x 1.2TB SAS FlexBay	2x 400GB SSD disks
Networking Configuration	2x 1G and 2x 10G	2x 1G and 2x 10G	2x 1G and 4x 10G (2 cards with 2 ports each)	2x 1G and 2x 10G

Switches

Switch Model	Quantity
Dell Networking S3048-ON	1
Dell Networking S4048-ON	2

Additional Rack equipment

You may order additional equipment such as Ethernet-managed PDUs and wire-management panels. Please contact your Dell representative.

Disk Configuration

The following configuration is the recommended minimum in terms of storage requirements.

Node	RAID type	Disks Number	Disk Type	Purpose
Controller	None	1	SSD	Operating system
	None	1	SSD	MongoDB
Compute	None	1	SSD	Operating system
Storage	RAID1	2	10k SAS(front)	Operating system
	JBOD	20	10k SAS	CEPH OSDs
	JBOD	4	SSD	CEPH Journal
Infrastructure	RAID1	2	SSD	Operating System

Compute node

Each Compute node requires at least one disk with average speed characteristics, for OS deployment. We recommend using at least a 400GB disk for this purpose. RAID1 is optional.

Ephemeral storage type is a cloud-wide configuration option that selects a backend type for Nova Ephemeral volumes. If Ephemeral storage on local disks is chosen, it's better to put that storage on a separate disk, which may be SAS or even SSD depending on the SLA required for cloud applications. These disks may be put into RAID0 or RAID1 to utilize the RAID controller's cache and gain even more performance. It's possible to use different disk types in different groups of Compute nodes, increasing flexibility.

Even with local Ephemeral storage, it's still possible to live-migrate instances which don't have Ephemeral storage attached.

We used 1x 400GB SSD for OS during validation. Ceph is used for Ephemeral storage.

Controller node

Controller nodes require at least one disk with good speed characteristics for OS and OpenStack services. It's recommended to use a 400GB SSD for these purposes. RAID1 is optional, but preferable.

If MongoDB is running on Controller nodes, we recommend using a second disk for the database. We suggest using an SSD with at least 200GB for a cloud with 30 Compute nodes. Actual MongoDB load depends on Ceilometer configuration. Please contact your Mirantis representative for detailed information.

For validation purposes 2x 400GB SSDs were used: one for OS and OpenStack services, and another for MongoDB. For RAID1, at least 3 disks should be used: 2x SSDs in RAID1 for OS and the third SSD with no RAID for MongoDB.

Storage node

Storage nodes require at least one disk with average speed characteristics for the OS. We recommend using at least a 200GB disk for this purpose. RAID1 is optional but preferable if you have configured a small number of Storage nodes.

To speed up Ceph we recommend putting the Ceph journal on SSD disks. The amount and size of required SSDs is roughly calculated like this: 40GB SSD for every 1-1.2TB SAS disk, and a single SSD may be used with up to 5 SAS disks. The above formula calculates 2x 200GB SSDs for 10 1.2TB SAS disks, and so on.

Taking all the above into account, the next configuration for Dell R730xd turns out to be:

2x of SAS located in the FlexBay are set up as RAID1.
4x SATA 200GB Intel DC S3710 SSDs and 20x 1.2TB SAS disks.

We used this exact configuration during validation.

Infrastructure node

Infrastructure nodes require at least one disk with average speed characteristics for OS deployment. We recommend using at least a 400GB disk for this purpose. RAID1 is optional, but preferable. Otherwise, the Fuel Master node must be backed up.

We used 2x 400GB SSDs in RAID1 during validation.

Validated Configuration

Here are the validated configurations summed up:

- Hardware configuration
 - O Dell PowerEdge R630 is used for Infrastructure, Controller, and Compute nodes
 - O Dell PowerEdge R730xd is used for Storage nodes
 - O 1 Infrastructure node
 - 2x 400GB Intel® SSD disks built into RAID1
 - 1x 1GB Intel® NIC to carry BMC networks
 - 1x 1GB Intel® NIC to carry Public network
 - O 3 Controller nodes
 - 2x 400GB Intel® SSD disks, one for OS and another one for MongoDB
 - 2ports 10GB Intel® NIC bonded into LACP to carry control and data plane networks
 - 1x 1GB Intel® NICs to carry BMC networks
 - O 3 Storage nodes
 - 2x of SAS located at FlexBay are built into RAID1 for OS
 - 4x SATA 200GB Intel® DCS3710 SSDs for Ceph journals
 - 20x 1.2TB SAS disks for Ceph OSDs
 - 2x 2ports Intel® 10GB NICs bonded into two LACP bonds to carry control-plane and Ceph replication traffic separately
 - 1x 1GB Intel® NICs to carry BMC networks
 - O 3 Compute nodes
 - 400GB SSD disk one for OS
 - 2ports 10GB NIC bonded into LACP to carry control and data plane networks
 - 1x 1GB NICs to carry BMC networks
 - O Dell Networking S3048-ON and Dell Networking S4048-ON are used.
 - O Cabling and logical network configuration are as described in corresponding sections.
 - O Tenant traffic segmentation is done by VLANs and DVR is used to accelerate East-West communication
- Software configuration
 - O Ceilometer is installed along with local MongoDB which is situated on Controller nodes
 - O Ceph with replication factor 3 is used for Cinder, Glance, Swift, and Nova Ephemeral storage
 - O TLSv1.2 with self-signed certificate is enabled for public Openstack API and Horizon
 - O No Sahara, Zabbix, or LMA toolchain is installed

Scaling

Our preliminary tests show that three bare-metal Controller nodes may serve up to 100 Compute nodes in the majority of use cases. The recommended Compute and Storage nodes ratio is 4 to 1 but this may be adjusted for a particular use case. Below, we offer a method for calculating how many nodes you need to host a given number of VMs. It's possible to connect multiple racks into one cloud. In this case the second and other racks would have the same configuration except for the Controller nodes. It's possible to put all 3 controllers into a single rack or distribute them across multiple racks to have better resiliency.

Multiple racks may be connected to each other and to the uplink in multiple ways.

- One is to use a dedicated aggregation switch such as Dell S6000-ON. This is the highest-performance, but most expensive approach.
- Another way is to connect the first rack to uplink and the next racks to the previous ones so they will form a chain. This way is less expensive but also less performant and redundant.

Contact your Mirantis or Dell representative to discuss options.

Suggested Rack Configurations

	Single rack configurations					
	tiny	tiny+	small	small+	medium	medium+
Controller Nodes	3	3	3	3	3	3
Compute Nodes	6	8	12	16	20	24
Storage Nodes	3	3	3	4	5	6
Estimated Maximum Number of VMs*	192	256	384	512	640	768
Power Footprint**	5kW	6kW	7kw	8.5kW	10.5kW	12kW

*The following parameters were used for estimating the number of VMs supported by each configuration. No RAM oversubscription and a 4:1 vCPU to Hyperthreaded CPU oversubscription ratio were used for the calculations.

- 4 x Virtual CPU's
- 8GB RAM
- 60GB Storage

**Use the Dell ESSA (Dell Energy Smart Solution Advisor) to plan your hardware and calculate its power consumption http://www.dell.com/learn/us/en/04/campaigns/config_calculator.

Testing

The post-deployment cloud validation is performed in a semi-automatic way using such tools as Tempest, Rally, Shaker, and others.

References

- Mirantis OpenStack 7.0 Documentation: https://docs.mirantis.com/openstack/fuel/fuel-7.0/