MCP Standard Configuration

version 1.0

Contents

Copyright notice

Preface

This documentation provides information on how to use Mirantis products to deploy cloud environments. The information is for reference purposes and is subject to change.

Intended audience

This documentation is intended for deployment engineers, system administrators and developers; it assumes that the reader is already familiar with network and cloud concepts.

Documentation history

The following table lists the released revisions of this documentation:

Revision date	Description
March 30, 2017	1.0 GA

Cloud sizes

The Mirantis Cloud Platform (MCP) enables you to deploy OpenStack environments and Kubernetes clusters of different scales. This document uses the terms small, medium, and large clouds to correspond to the number of virtual machines that you can run in your OpenStack environment or the number of pods that you can run in your Kubernetes cluster.

This section describes sizes of MCP clusters.

OpenStack environment sizes

All cloud environments require you to configure a staging environment which mimics the exact production setup and enables you to test configuration changes prior to applying them in production. However, if you have multiple clouds of the same type, one staging environment is sufficient.

Environments configured with Neutron OVS as a networking solution require additional hardware nodes called network nodes or tenant network gateway nodes.

The following table describes the number virtual machines for each scale:

OpenStack environment sizes

Environment size	Number of virtual machines	Number of compute nodes	Number of infrastructure nodes
Small	1,000	20 - 50	6
Medium	5,000	50 - 200	9 (with Neutron OVS), 12 (with Mirantis OpenContrail)
Large	10,000	200 - 500	17 (Mirantis OpenContrail only)

Kubernetes cluster sizes

Kubernetes and etcd are lightweight applications that typically do not require a lot of resources. However, each Kubernetes components scales differently and some of them may have limitations on scaling.

Based on performance testing, an optimal density of pods is 100 pods per minion node.

> Note
>
> If you run both Kubernetes clusters and OpenStack environments in one MCP installation, you do not need to configure additional infrastructure nodes. Instead, the same infrastructure nodes are used for both.

The following table describes the number of Master and Minion Kubernetes nodes for each scale:

Small cloud

Cluster size	Number of pods	Number of Minion nodes	Number of Master nodes	Number of infrastructure nodes
Small	2,000 - 5,000	20 - 50	3	3
Medium	5,000 - 20,000	50 - 200	6	3
Large	20,000 - 50,000	200 - 500	9	3

Minimum hardware requirements

When calculating hardware requirements, you need to plan for the infrastructure nodes, compute nodes, storage nodes, and, if you are installing a Kubernetes cluster, Kubernetes Master and Minion nodes.

> Note
>
> For more details about services distribution throughout the nodes, see: OpenStack environment scaling and Kubernetes cluster scaling.

The following tables list minimal hardware requirements for the corresponding nodes of the cloud:

OpenStack infrastructure nodes

Parameter	Value
CPU	2 x 12 Intel Core CPU E5-2670v3 or similar
RAM	256 GB
Disk	2 x 1.2TB SSD Intel S3710, (if available, 2TB is preferred) or similar
Network	2 x 10 GB Intel X710 dual-port NICs or similar

OpenStack compute nodes

Parameter	Value
CPU	2 x 12 Intel Core CPU E5-2670v3 or similar
RAM	256 GB
Disk	2 x 800 GB SSD Intel S3710 or similar
Network	2 x 10 GB Intel X710 dual-port NICs or similar

Ceph OSD Storage nodes

Parameter	Value
CPU	2 x 12 Intel Core CPU E5-2650v3 or similar
RAM	128 GB
Disk	• 2 x 480GB SSD Intel S3710 for journal storage, • 10 x 2 TB HDD for data storage
Network	2 x 10 GB Intel X710 dual-port NICs or similar

Kubernetes Master and Minion nodes

Parameter	Value
CPU	2 x 12 Intel Core CPU E5-2670v3 or similar
RAM	256 GB
Disk	2 x 800 GB SSD Intel S3710 or 2 x 1 TB HDD
Network	2 x 10 GB Intel X710 dual-port NICs or similar

Hardware requirements for different sizes of clouds

The following tables list minimum hardware requirements for various types of configurations.

Small cloud

Server type	Number of servers	Number of physical CPU cores per server	Memory (GB) per server	Disk (GB) per server	NICs per server
Infrastructure/control nodes	6	24	256	3000/4000	4
Compute nodes	10 - 50	24	256	1600	4
Staging Infrastructure/control nodes	6	24	256	3000/4000	4
Staging compute nodes	2-10	24	256	1600	4

Medium cloud (with Neutron OVS)

Server type	Number of servers	Number of physical CPU cores per server	Memory (GB) per server	Disk (GB) per server	NICs per server
Infrastructure/control nodes	9	24	256	3000/4000	4
Tenant network gateway nodes	3	24	256	1600	4
Compute nodes	50 - 200	24	256	1600	4
Staging Infrastructure/control nodes	9	24	256	3000/4000	4
Staging tenant network gateway nodes	3	24	256	1600	4
Staging compute nodes	2-10	24	256	1600	4

Medium cloud (with Mirantis OpenContrail)

Server type	Number of servers	Number of physical CPU cores per server	Memory (GB) per server	Disk (GB) per server	NICs per server
Infrastructure/control nodes	12	24	256	3000/4000	4
Compute nodes	50 - 200	24	256	1600	4
Staging Infrastructure/control nodes	12	24	256	3000/4000	4
Staging compute nodes	2 - 10	24	256	1600	4

Large cloud

Server type	Number of servers	Number of physical CPU cores per server	Memory (GB) per server	Disk (GB) per server	NICs per server
Infrastructure/control nodes	17	24	256	3000/4000	4
Compute nodes	200 - 500	24	256	1600	4
Staging Infrastructure/control nodes	17	24	256	3000/4000	4
Staging compute nodes	2 - 10	24	256	1600	4

OpenStack environment scaling

The Mirantis Cloud Platform (MCP) enables you to deploy OpenStack environments at different scales. This document defines the following sizes of environments: small, medium, and large. Each environment size requires a different number of infrastructure nodes. The Virtualized Control Plane (VCP) services are distributed among the physical infrastructure nodes for optimal performance.

This section describes VCP services distribution, as well as hardware requirements for different sizes of clouds.

OpenStack server roles

Components of the Virtualized Control Plane (VCP) have roles that define their functions. Each role can be assigned to a specific set of virtual servers which allows to adjust the number of instances with a particular role independently of other roles providing greater flexibility to the environment architecture.

The following table lists the OpenStack roles and their names in the SaltStack formulas:

OpenStack infrastructure nodes

Server role name	Server role group name in SaltStack formulas	Description
Infrastructure node	kvm	Infrastructure KVM hosts that run MCP component services as virtual machines.
Network node	gtw	Nodes that provide tenant network data plane services.
StackLight Monitoring node	mon	Servers that provide monitoring services for MCP and cluster infrastructure.
StackLight Logging Collector node	log	Servers that collect logs from MCP clusters.
DriveTrain Salt Master node	cfg	The Salt Master node that is responsible for sending commands to Salt Minion nodes.
DriveTrain / StackLight OSS node	cid	Nodes that run in StackLight OSS and DriverTrain services in containers in Docker Swarm mode.
RabbitMQ server node	msg	Nodes that run the message queue service RabbitMQ.
Database server node	dbs	Nodes that run the database cluster called Galera.

OpenStack controller nodes	ctl	Nodes that run the Virtualized Control Plane service, including the API servers and scheduler components.
OpenStack compute nodes	cmp	Nodes that run the hypervisor service and VM workloads.
Proxy node	prx	Nodes that run reverse proxy that exposes OpenStack API, dashboards, and other components externally.
Contrail Controller nodes	ntw	Nodes that run Mirantis OpenContrail controller services.
Contrail Analytics nodes	nal	Nodes that run Mirantis OpenContrail analytics services.

Services distribution across nodes

The distribution of services across physical nodes depends on their resources consumption profile and the number of compute nodes they administer.

Mirantis recommends the following distribution of services across nodes:

Distribution of services across nodes in an OpenStack environment

Service	Physical server group	Virtual	VM role group
keystone-all	kvm	yes	ctl
nova-api	kvm	yes	ctl
nova-scheduler	kvm	yes	ctl
nova-conductor	kvm	yes	ctl
nova-compute	cmp	no	N/A
neutron-server (for Neutron OVS only)	kvm	yes	ctl
neutron-dhcp-agent (for Neutron OVS only)	gtw	no	N/A
neutron-l2-agent (for Neutron OVS only)	cmp, gtw	no	N/A
neutron-l3-agent (for Neutron OVS only)	gtw	no	N/A
neutron-metadata-agent (for Neutron OVS only)	gtw	no	N/A
glance-api	kvm	yes	ctl
glance-registry	kvm	yes	ctl
cinder-api	kvm	yes	ctl
cinder-scheduler	kvm	yes	ctl

cinder-volume	kvm	yes	ctl
horizon/apache2	kvm	yes	prx
rabbitmq-server	kvm	yes	msg
mysql-server	kvm	yes	dbs
ceilometer-api	kvm	yes	mon
StackLight	kvm	yes	mon
InfluxDB	kvm	yes	mtr
Elasticsearch	kvm	yes	log
Nagios	kvm	yes	mon
DriveTrain (In Docker Swarm mode)	kvm	yes	cid
OSS Tools (In Docker Swarm mode)	kvm	yes	cid

Operating systems and versions

The following table lists the operating systems used for different roles in the Virtualized Control Plane (VCP):

Operating systems and versions

Server role name	Server role group name in the SaltStack model	Ubuntu version
Infrastructure node	kvm	xenial/16.04
Network node	gtw	xenial/16.04
StackLight monitoring node	mon	xenial/16.04
Stacklight tenant telemetry node	mtr	xenial/16.04
StackLight Logging Collector node	log	xenial/16.04
DriveTrain Salt Master node	cfg	xenial/16.04
DriveTrain StackLight OSS node	cid	xenial/16.04
RabbitMQ server node	msg	trusty/14.04
Database server node	dbs	trusty/14.04
OpenStack controller node	ctl	trusty/14.04
OpenStack compute node	cmp	xenial/16.04 (depends on whether NFV features enabled or not)
Proxy node	prx	trusty/14.04

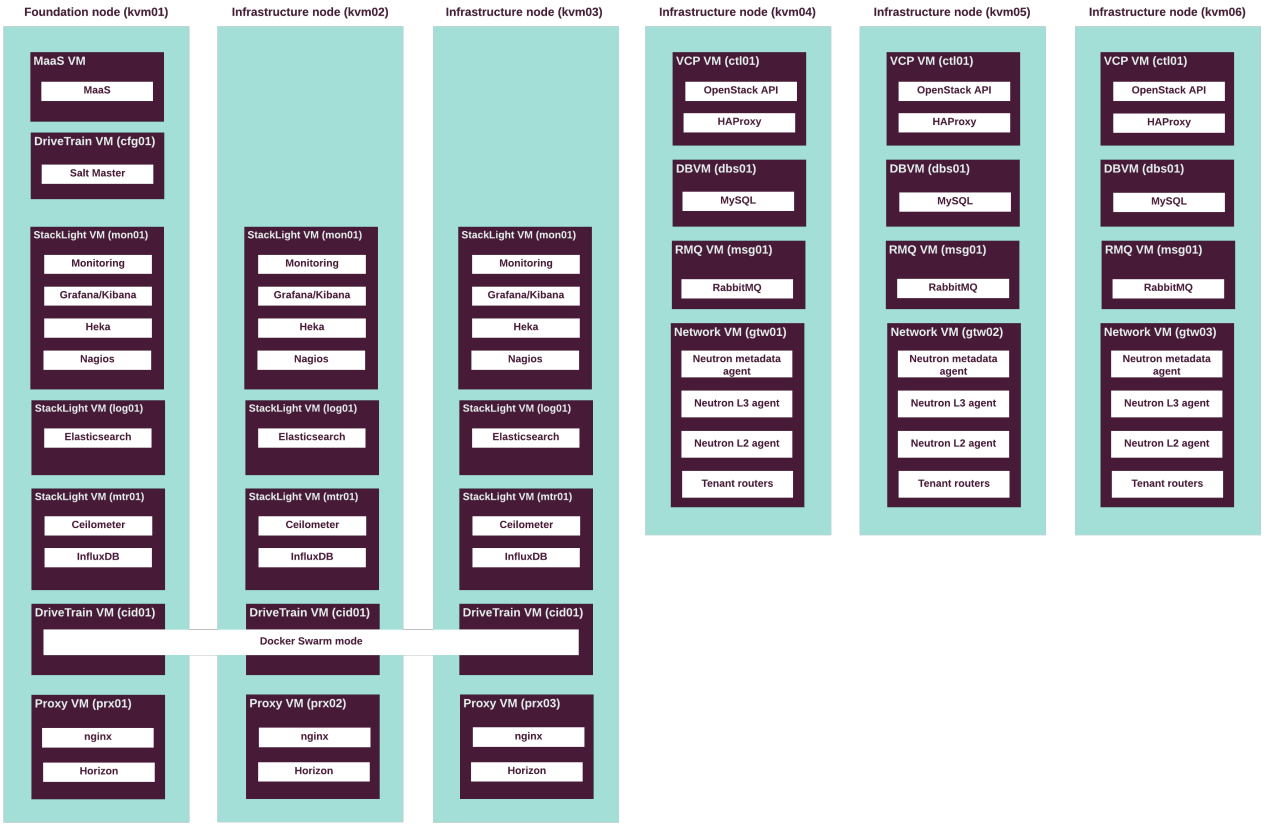
Mirantis OpenContrail Control node	ntw	trusty/14.04
Mirantis OpenContrail Analytics node	nal	trusty/14.04

OpenStack small cloud architecture

A small OpenStack cloud includes up to 50 compute nodes and requires you to have at least 6 infrastructure nodes, including additional servers for KVM infrastructure and RabbitMQ/Galera services.

A small cloud includes all the roles described in OpenStack server roles.

The following diagram describes the distribution of VCP and other services throughout the infrastructure nodes.



Note

A vCore is the number of available virtual cores considering hyper-threading and overcommit ratio. Assuming an overcommit ratio of 1, the number of vCores in a physical is roughly the number of physical cores multiplied by 1.3.

The number of nodes is the same for both Mirantis OpenContrail and Neutron OVS-based clouds. However, the roles on network nodes are different.

The following table describes the hardware nodes in MCP OpenStack, roles assigned to them, and resources per node:

Physical server roles and hardware requirements

Physical server roles	Description	# of servers	CPU vCores per server	Memory per server (GB)	SSD disk per server(GB)	# of NICs per server
kvm	Infrastructure nodes	6	32	256	3000/4000	4
cmp	Compute nodes	50 - 200	32	256	1600	4

The following table summarizes the VCP virtual machines mapped to physical servers.

Resource requirements per VCP role

Virtual server roles	Physical servers	# of instances	CPU vCores per instance	Memory (GB)	Disk space(GB)	# of vNICs
ctl	kvm04 kvm05 kvm06	3	8	32	100	2
msg	kvm04 kvm05 kvm06	3	8	64	300	2
dbs	kvm04 kvm05 kvm06	3	8	32	100	2
prx	kvm01 kvm02 kvm03	3	4	8	50	3
cfg	kvm01	1	2	8	50	2
mon	kvm01 kvm02 kvm03	3	8	16	240	2
mtr	kvm01 kvm02 kvm03	3	8	32	240	2
log	kvm01 kvm02 kvm03	3	8	32	400	2
cid	kvm01 kvm02 kvm03	3	4	32	100	2
gtw	kvm04 kvm05 kvm06	3	4	16	50	4

Note

- The gtw VM should have four separate NICs for the following interfaces: dhcp, primary, tenant, and external. It simplifies the host networking: you do not need to pass VLANs to VMs.
- The prx VM should have an additional NIC for the Proxy network.
- All other nodes should have two NICs for DHCP and Primary networks.

OpenStack medium cloud architecture with Neutron OVS DVR/non-DVR

A medium OpenStack cloud includes up to 200 compute nodes and requires you to have at least 9 infrastructure nodes, including additional servers for KVM infrastructure and RabbitMQ/Galera services. If you use Neutron OVS as a networking solution, bare-metal network nodes are installed to accommodate network traffic as opposed to virtualized controllers in the case of Mirantis OpenContrail.

A medium cloud includes all roles described in OpenStack server roles.

The following diagram describes the distribution of VCP and other services throughout the infrastructure nodes.

> Note
>
> A vCore is the number of available virtual cores considering hyper-threading and overcommit ratio. Assuming an overcommit ratio of 1, the number of vCores in a physical is roughly the number of physical cores multiplied by 1.3.

The following table describes the hardware nodes in MCP OpenStack, roles assigned to them, and resources per node:

Physical server roles and hardware requirements

Physical server roles	Description	# of servers	CPU vCores per server	Memory per server (GB)	SSD disk per server(GB)	# of NICs per server
kvm	Infrastructure nodes	9	32	256	3000/4000	4
gtw	Tenant network gateway nodes	3	32	256	1600	4
cmp	Compute nodes	50 - 200	32	256	1600	4

The following table summarizes the VCP virtual machines mapped to physical servers.

Resource requirements per VCP role

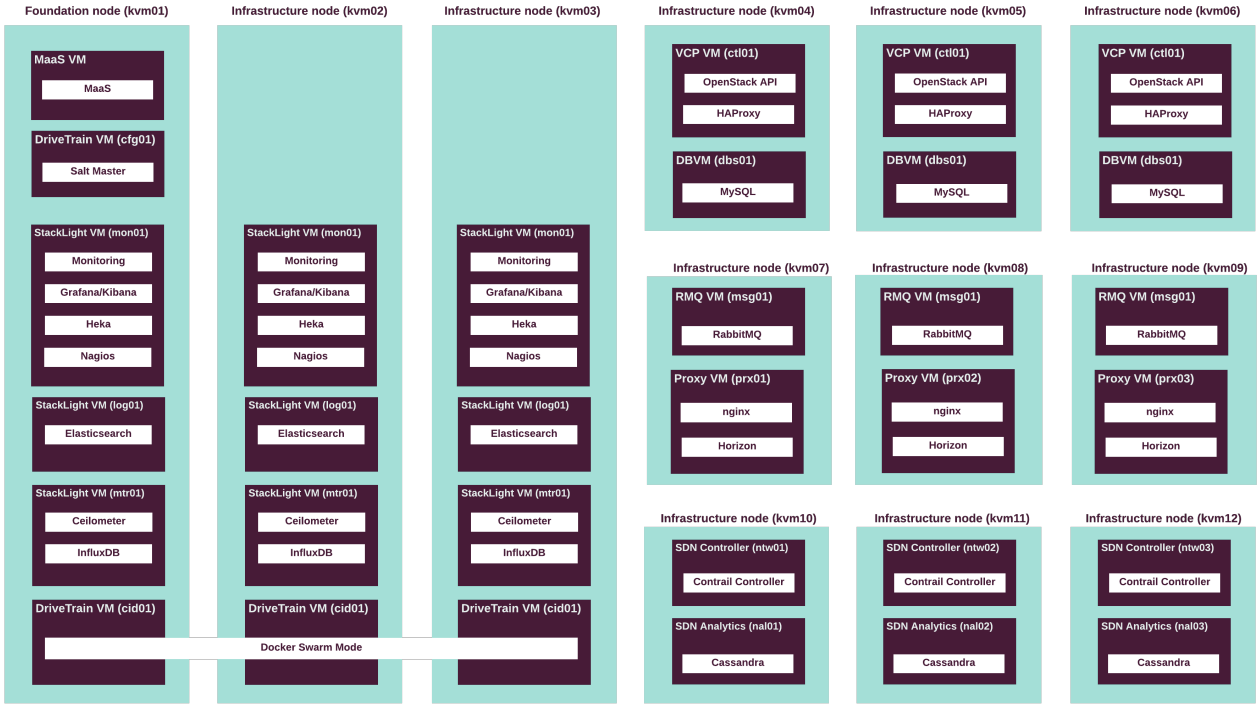
Virtual server roles	Physical servers	# of instances	CPU vCores per instance	Memory (GB)	Disk space(GB)	# of vNICs
ctl	kvm04 kvm05 kvm06	3	24	64	100	2
msg	kvm07 kvm08 kvm09	3	24	64	300	2
dbs	kvm04 kvm05 kvm06	3	8	32	100	2
prx	kvm07 kvm08 kvm09	3	4	16	100	3
cfg	kvm01	1	2	8	50	2
mon	kvm01 kvm02 kvm03	3	4	16	240	2
mtr	kvm01 kvm02 kvm03	3	10	48	1000	2
log	kvm01 kvm02 kvm03	3	16	48	2000	2
cid	kvm01 kvm02 kvm03	3	4	32	200	2

OpenStack medium cloud architecture with Mirantis OpenContrail

A medium OpenStack cloud includes up to 200 compute nodes and requires you to have at least 12 infrastructure nodes, including additional servers for KVM infrastructure and RabbitMQ/Galera services. If you use Mirantis OpenContrail as a networking solution, a virtualized OpenContrail Controller is installed instead of bare metal network nodes as in the case of Neutron OVS.

A medium cloud includes all roles described in OpenStack server roles.

The following diagram describes the distribution of VCP and other services throughout the infrastructure nodes.



Note

A vCore is the number of available virtual cores considering hyper-threading and overcommit ratio. Assuming an overcommit ratio of 1, the number of vCores in a physical is roughly the number of physical cores multiplied by 1.3.

The following table describes the hardware nodes in MCP OpenStack, roles assigned to them, and resources per node:

Physical server roles and hardware requirements

Physical server roles	Description	# of servers	CPU vCores per server	Memory per server (GB)	SSD disk per server(GB)	# of NICs per server
kvm	Infrastructure nodes	12	32	256	3000/4000	4
cmp	Compute nodes	50 - 200	32	256	1600	4

The following table summarizes the VCP virtual machines mapped to physical servers.

Resource requirements per VCP role

Virtual server roles	Physical servers	# of instances	CPU vCores per instance	Memory (GB)	Disk space(GB)	# of vNICs
ctl	kvm04 kvm05 kvm06	3	24	64	100	2
msg	kvm07 kvm08 kvm09	3	24	64	300	2
dbs	kvm04 kvm05 kvm06	3	8	32	100	2
prx	kvm07 kvm08 kvm09	3	4	16	100	3
cfg	kvm01	1	2	8	50	2
mon	kvm01 kvm02 kvm03	3	4	16	240	2
mtr	kvm01 kvm02 kvm03	3	10	48	1000	2
log	kvm01 kvm02 kvm03	3	16	48	2000	2
cid	kvm01 kvm02 kvm03	3	4	32	200	2
ntw	kvm10 kvm11 kvm12	3	16	64	100	2
nal	kvm10 kvm11 kvm12	3	24	96	1200	2

OpenStack large cloud architecture

A large OpenStack cloud includes up to 500 compute nodes and requires you to have at least 17 infrastructure nodes, including dedicated bare-metal servers for RabbitMQ/Galera, OpenStack API services, and database servers. Use Mirantis OpenContrail as a networking solution for large OpenStack clouds. Neutron OVS is not recommended.

A large cloud includes all roles described in OpenStack server roles.

The following diagram describes the distribution of VCP and other services throughout the infrastructure nodes.

Note

A vCore is the number of available virtual cores considering hyper-threading and overcommit ratio. Assuming an overcommit ratio of 1, the number of vCores in a physical is roughly the number of physical cores multiplied by 1.3.

The following table describes the hardware nodes in MCP OpenStack, roles assigned to them, and resources per node:

Physical server roles and hardware requirements

Physical server roles	Description	# of servers	CPU vCores per server	Memory per server (GB)	SSD disk per server(GB)	# of NICs per server
kvm	Infrastructure nodes	17	32	256	3000/4000	4
cmp	Compute nodes	200 - 500	32	256	1600	4

The following table summarizes the VCP virtual machines mapped to physical servers.

Resource requirements per VCP role

Virtual server roles	Physical servers	# of instances	CPU vCores per instance	Memory (GB)	Disk space(GB)	# of vNICs
ctl	kvm04 kvm05 kvm06 kvm07 kvm08	5	24	128	100	2
msg	kvm12 kvm13 kvm14	3	32	196	300	2
dbs	kvm09 kvm10 kvm11	3	24	64	1000	2
prx	kvm09 kvm10 kvm11	3	8	64	100	3
cfg	kvm01	1	4	16	50	2
mon	kvm01 kvm02 kvm03	3	8	64	240	2
mtr	kvm01 kvm02 kvm03	3	16	128	1000	2
log	kvm01 kvm02 kvm03	3	16	64	2000	2
cid	kvm01 kvm02 kvm03	3	4	32	300	2
ntw	kvm15 kvm16 kvm17	3	8	64	100	2
nal	kvm15 kvm16 kvm17	3	24	128	2000	2

Kubernetes cluster scaling

As described in Kubernetes cluster sizes, depending on the anticipated number of pods, your Kubernetes cluster may be a small, medium, or large scale deployment. A different number of Kubernetes Worker nodes is required for each size of cluster.

This section describes the services distribution across hardware nodes for different sizes of Kubernetes clusters.

Small Kubernetes cluster architecture

A small Kubernetes cluster includes 2,000 - 5,000 pods spread across roughly 20 - 50 physical Kubernetes Worker nodes and 3 Kubernetes Master nodes. Mirantis recommends separating the Kubernetes control plane that includes etcd and the Kubernetes Master node components from Kubernetes workloads.

In addition, dedicate separate physical servers for shared storage services GlusterFS and Ceph. Running these components on the same physical servers as the control plane components may result in insufficient cycles for etcd cluster to stay synced and allow the kubelet agent to report their statuses reliably.

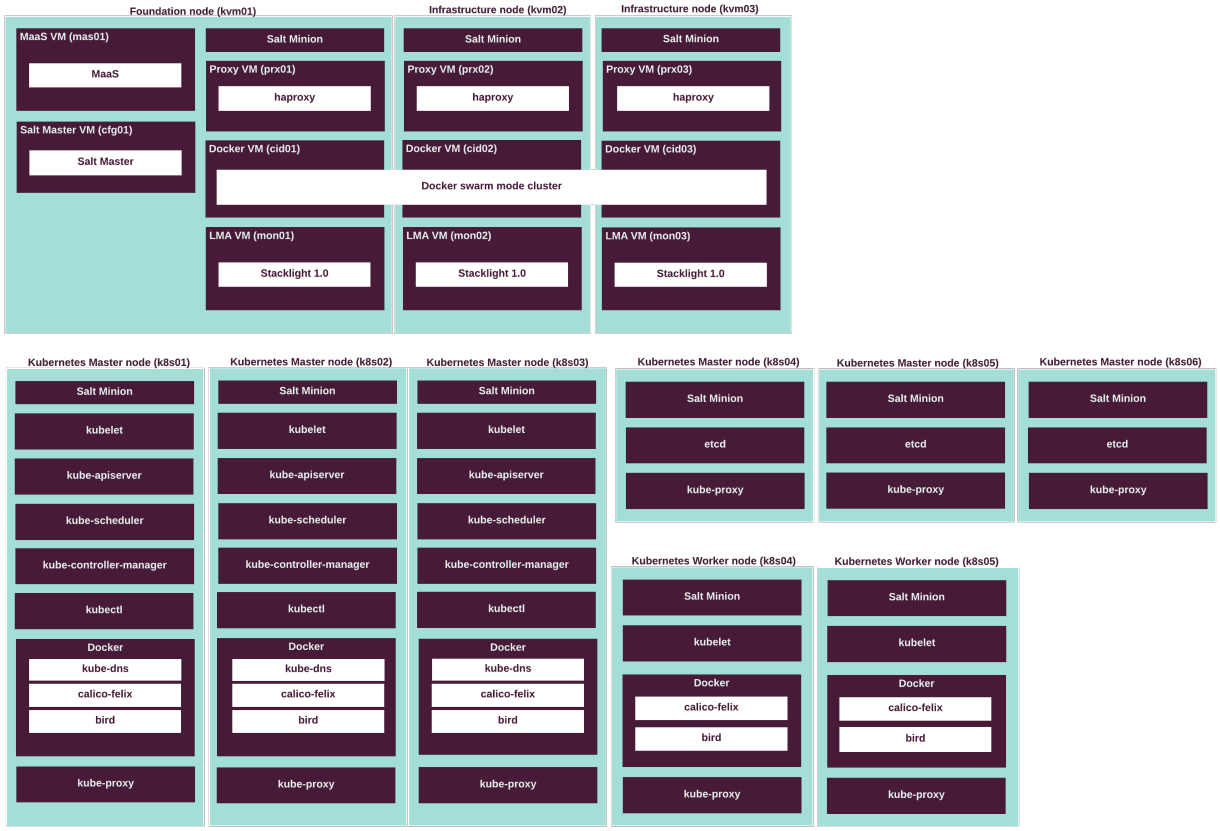
The following diagram displays the layout of services per physical node for a small Kubernetes cluster.

Medium Kubernetes cluster architecture

A medium Kubernetes cluster includes 5,000 - 20,000 pods spread across roughly 50 - 200 physical Kubernetes Worker nodes and 6 Kubernetes Master nodes. Mirantis recommends separating the Kubernetes control plane that includes etcd and the Kubernetes Master node components from Kubernetes workloads. While Kubernetes components can run on the same host, run etcd on dedicated servers as the etcd workload increases due to constant recording and checking pod and kubelet statuses.

You can place shared storage services GlusterFS and Ceph on the same physical nodes as Kubernetes control plane components.

The following diagram displays the layout of services per physical node for a medium Kubernetes cluster.



Large Kubernetes cluster architecture

A large Kubernetes cluster includes 20,000 - 50,000 pods spread across roughly 200 - 500 physical Kubernetes Worker nodes and 9 Kubernetes Master nodes. Mirantis recommends to separate the Kubernetes control plane that includes etcd and Kubernetes Master node components from the Kubernetes workloads.

> Note
>
> While Kubernetes components can run on the same host, run etcd on dedicated servers as the etcd workload increases due to constant recording and checking pod and kubelet statuses.

In addition, Mirantis recommends placing kube-scheduler separately from the rest of the Kubernetes control plane components due to the high number of Kubernetes pod turnover and rescheduling, kube-scheduler requires more resources than other Kubernetes components.

You can place shared storage services GlusterFS and Ceph on the same physical nodes as Kubernetes control plane components.

The following diagram displays the layout of services per physical node for a large Kubernetes cluster.

Ceph hardware requirements

Mirantis recommends to use the Ceph cluster as primary storage solution for all types of ephemeral and persistent storage. A Ceph cluster that is built in conjunction with MCP must be designed to accommodate:

- Capacity requirements
- Performance requirements
- Operational requirements

This section describes differently sized clouds that use the same Ceph building blocks.

Ceph cluster considerations

When planning storage for your cloud you must consider performance, capacity, and operational requirements that affect the efficiency of your MCP environment. The minimum number of Ceph cluster nodes for a small OpenStack environment is nine.

> Note
>
> This section provides simplified calculations for your reference. Each Ceph cluster must be evaluated by a Mirantis Solution Architect.

Capacity

When planning capacity for your Ceph cluster, consider the following:

- Total usable capacity

  The existing amount of data plus the expected increase of data volume over the projected life of the cluster.

- Data protection (replication)

  Typically, for persistent storage a factor of 3 is recommended, while for ephemeral storage a factor of 2 is sufficient.

- Cluster overhead

  To ensure cluster integrity, Ceph stops writing if the cluster is 90% full. Therefore, you need to plan accordingly.

- Administrative overhead

  To catch spikes in cluster usage or unexpected increases in data volume, an additional 10-15% of the raw capacity should be set aside.

The following table describes an example of capacity calculation:

Example calculation

Parameter	Value
Current capacity persistent	500 TB
Expected growth over 3 years	300 TB
Required usable capacity	800 TB
Replication factor for all pools	3
Raw capacity	2.4 PB
With 10% cluster internal reserve	2.64 PB
With operational reserve of 15%	3.03 PB
Total cluster capacity	3 PB

Performance

When planning performance for your Ceph cluster, consider the following:

- Raw performance capability of the storage devices. For example, a SATA hard drive provides 150 IOPS for 4k blocks.
- Ceph read IOPS performance. Calculate using the following formula:

  number of raw read IOPS per device X number of storage devices X 80%
- Ceph write IOPS performance. Calculate using the following formula:

  number of raw write IOPS per device X number of storage devices / replication factor X 65%
- Ratio between reads and writes. Perform a rough calculation using the following formula:

  read IOPS X % reads + write IOPS X % writes

> Note
>
> Do not use this formula for a Ceph cluster that is based on SSDs only. Contact Mirantis for evaluation.

Overall sizing

When you have both performance and capacity requirements, scale the cluster size to the higher requirement. For example, if a Ceph cluster requires 10 nodes for capacity and 20 nodes to meet performance requirements, size the cluster to 20 nodes.

Operational requirements

- A minimum of 8 Ceph OSD nodes is recommended to ensure node failure does not impact cluster performance.
- Mirantis does not recommend using servers with excessive number of disks, such as more than 36 disks.

- All OSD nodes should be configured equally.
- If you use multiple availability zones (AZ), the number of nodes should be evenly divisible by the number of AZ.

The expected number of IOPS a storage device can carry, as well as its throughput, depends on the type of device. For example, a hard disk may be rated for 150 IOPS and 75 MB/s. These numbers are complementary because IOPS are measured with very small files while throughput is typically measured with big files.

Read IOPS and write IOPS differ depending on the device. Typically, considering typical usage patterns helps determining how many read and write IOPS the cluster must provide. A ratio of 70/30 is fairly common for many types of clusters. The cluster size must also be considered, as the maximum number of write IOPS a cluster can push is divided by the cluster size. Furthermore, Ceph can not guarantee the full IOPS numbers a device could theoretically provide, because the numbers are typically measured under testing environments, which the Ceph cluster cannot offer and also because of the OSD and network overhead.

You can calculate estimated read IOPS by multiplying the read IOPS number for the device type and multiplying with the number of devices, and then multiplying by ~0.8. Write IOPS are calculated as the (device IOPS * number of devices * 0.65) / cluster size. If the cluster size for the pools is different, an average can be used. If the number of devices is required, the respective formulas can be solved for the device number instead.

Ceph cluster sizes

When sizing a Ceph cluster, you must consider the number of drives needed for capacity and the number of drives required to accommodate performance requirements. You must also consider the largest number of drives that ensure all requirements are met. Allocate 0.5 CPU and 4 GB of RAM per Ceph Object Storage Device (OSD). Mirantis does not recommend using RAID. Instead, all drives must be available to the Ceph cluster individually. If the RAID controller does not support JBOD mode, you can configure an individual RAID0 for each device. The Ceph monitor can run in virtual machines on the infrastructure nodes and use little resources.

The following table describes an example of 20-disk chassis 2U nodes hardware configuration.

Ceph cluster nodes hardware requirements

Parameter	Value
CPU	2 x 8-core CPU
RAM	128 GB
Disk	20 x 2.5" TB devices, 4 x 2.5" 200 GB write-optimized SSDs
Boot drives	2 x 128 GB Disk on Module (DOM)

The following table provides an example of input parameters for a Ceph cluster calculation

Example of input parameters

Parameter	Value
Virtual instance size	40 GB
Read IOPS	14
Read to write IOPS ratio	70/30
Number of availability zones	3

For 50 compute nodes, 1,000 instances

Number of OSD nodes: 9, 20-disk 2U chassis

This configuration provides 120TB of raw storage and with cluster size of 3 and 60% used initially, the initial amount of data should not exceed 72TB. Expected read IOPS for this cluster is approximately 18300 and write IOPS 4400, or 14000 IOPS in a 70/30 pattern.

> Note
>
> In this case performance is the driving factor, and so the capacity is greater than required.

For 300 compute nodes, 6,000 instances

Number of OSD nodes: 54, 36-disks chassis

The cost per node is low compared to the cost of the storage devices and with a larger number of nodes failure of one node is proportionally less critical. A separate replication network is recommended.

For 500 compute nodes, 10,000 instances

Number of OSD nodes: 60, 36-disks chassis

You may consider to use a larger chassis. A separate replication network is required.

The following table summarizes the number of storage nodes required per cloud size that provides the mentioned above performance characteristics:

Ceph storage nodes per cloud size

Cloud size	Number of compute nodes	Number of virtual machines	Number of storage nodes
Small	10 - 50	1,000	9 (20-disk chassis)
Medium	50 - 200	5,000	54 (20-disk chassis)
Large	200 - 500	10,000	60 (36-disk chassis)

StackLight hardware requirements

StackLight is the Logging, Metering, and Alerting (LMA) toolchain that enables you to monitor the health of your nodes, services, and clusters, as well as business KPIs.

StackLight logging

StackLight for MCP includes Elasticsearch and Kibana as components of the logging system. They have the following requirements:

Elasticsearch and Kibana requirements for StackLight

Requirement	Comments
Disk space	Elasticsearch requires at least 15 GB of disk space for the system and 10 GB for the logs. Mirantis recommends installing the Elasticsearch database on a dedicated disk partition. The size of the partition depends on many factors including the size of the deployment, the retention period, and the log level. Logging at the DEBUG level requires 10 times more space than logging at the INFO level. According to StackLight test results, a medium size OpenStack deployment of 200 compute nodes requires a 500 GB disk space on every Elasticsearch node with three infrastructure nodes and a retention period of one month.
Hardware specification	The hardware specification required for Elasticsearch depends on the size of the deployment and other factors like the retention period and logging activity. A small size OpenStack deployment requires a quad-core CPU with 4 GB of RAM and a fast disk of 500-1000 IOPS. For larger deployments, Mirantis recommends having 8 GB of RAM or 50% of the available memory up to 32 GB maximum for the JVM heap.

The Log Collector agent heka runs on every node of an MCP cluster. It has low resource requirements and should not be counted against the available resources capacity on those nodes.

StackLight metering

StackLight for MCP includes InfluxDB and Grafana as components of the metering system. They have the following requirements:

InfluxDB and Grafana requirements for StackLight

Requirement	Comments
Disk space	InfluxDB requires at least 15 GB of disk space for the system and 10 GB for the logs. Mirantis recommends installing the InfluxDB database on a dedicated disk partition. The size of the partition depends on many factors including the size of the deployment and the retention period. According to StackLight test results, a medium size OpenStack deployment of 200 compute nodes requires a 100 GB disk space on every InfluxDB node with three infrastructure nodes and a retention period of one month.
Hardware specification	The hardware specifications required for InfluxDB and Grafana depend on the size of the deployment and other factors like the retention period. A small size OpenStack deployment requires at least a quad-core CPU with 8 GB of RAM and a fast disk of 500-1000 IOPS. See the InfluxDB Hardware Sizing Guide for additional sizing information.

The metering agents collectd and heka require low resources and run on each node of the MCP cluster being monitored.

StackLight alerting

StackLight for MCP includes either Sensu or Nagios as the main components of the alerting system. They are placed on the monitoring nodes (VMs) along with the remote Collector and Aggregator services.

In addition to the requirements described in StackLight metering, the StackLight alerting system has the following capacity requirements:

The StackLight alerting system requirements

Requirement	Comments
Disk space	The alerting system requires at least 15 GB of disk space for the system, 10 GB for the logs, and 20 GB for either Sensu or Nagios themselves.
Hardware specification	A small size OpenStack deployment requires at least a quad-core CPU with 8 GB of RAM and a fast disk.

Seealso

- Logging, metering, and alerting planning in the MCP Reference Architecture
- Install StackLight in the MCP Deployment Guide

NFV considerations

Network function virtualization (NFV) is an important factor to consider while planning hardware capacity for the Mirantis Cloud Platform. Mirantis recommends separating nodes that support NFV from other nodes to reserve them as Data Plane nodes that use network virtualization functions.

The following types of the Data Plane nodes use NFV:

1. Compute nodes that can run virtual machines with hardware-assisted Virtualized Networking Functions (VNF) in terms of the OPNFV architecture.
2. Networking nodes that provide gateways and routers to the OVS-based tenant networks using network virtualization functions.

The following table describes compatibility of NFV features for different MCP deployments.

NFV for MCP compatibility matrix

Type	Host OS	Kernel	Huge pages	DPDK	SR-IOV	NUMA	CPU pi nning	Multiq ueue
OVS	Xenial	4.8	Yes	No	Yes	Yes	Yes	Yes
Kernel vRouter	Xenial	4.8	Yes	No	Yes	Yes	Yes	Yes
DPDK vRouter	Trusty	4.4	Yes	Yes	No	Yes	Yes	No (version 3.2)
DPDK OVS	Xenial	4.8	Yes	Yes	No	Yes	Yes	Yes