Ceph硬件指导

目录

[Ceph硬件选择的一般性原则 4](#_Toc15303)

[供电模块 4](#_Toc18432)

[网络配置 5](#_Toc22119)

[服务器网络配置 5](#_Toc30382)

[机柜网络配置 6](#_Toc17749)

[跨机柜网络配置 6](#_Toc21786)

[MTU 7](#_Toc21519)

[磁盘控制器 8](#_Toc10261)

[服务器盘符保持一致 10](#_Toc26177)

[rhel操作系统中的盘符保持一致 10](#_Toc13879)

[数据盘建议做成lvm以对应盘符变化 11](#_Toc15191)

[服务高可用和数据冗余容灾设计 12](#_Toc20387)

[osd数据冗余 12](#_Toc24486)

[服务冗余 13](#_Toc24854)

[mon服务 13](#_Toc16336)

[mds服务 13](#_Toc21553)

[rgw服务 14](#_Toc14416)

[mon,rgw,mds服务部署在osd服务器上所需硬件资源 14](#_Toc20323)

[mon, rgw, mds单独物理机部署推荐硬件配置参考 15](#_Toc2634)

[mon节点 15](#_Toc14817)

[rgw节点 16](#_Toc10272)

[mds节点 16](#_Toc19592)

[osd硬件配置分为性能型，吞吐型，成本容量型 16](#_Toc8019)

[操作系统版本与Ceph版本对应关系 17](#_Toc32413)

[Ceph集群规模限制 17](#_Toc21170)

[高性能型Ceph集群配置推荐 18](#_Toc21959)

[硬件配置的目标 18](#_Toc3834)

[典型应用场景 18](#_Toc6848)

[硬件规格 18](#_Toc5438)

[联合测试产商 19](#_Toc16517)

[超微服务器sku与Ceph集群规模的大小 19](#_Toc20953)

[高吞吐量型Ceph集群配置推荐 19](#_Toc31250)

[硬件配置的目标 19](#_Toc27580)

[典型应用场景 19](#_Toc6889)

[硬件规格 19](#_Toc30546)

[联合测试产商 20](#_Toc250)

[Ceph集群规模大小与机架sku 20](#_Toc12620)

[Ceph集群规模大小与服务器sku 20](#_Toc31398)

[其他厂家服务器型号与Ceph集群 20](#_Toc2328)

[节约成本容量型Ceph集群配置推荐 21](#_Toc29167)

[硬件配置的目标 21](#_Toc16340)

[典型应用场景 21](#_Toc11797)

[硬件规格 21](#_Toc18417)

[联合测试产商 22](#_Toc18224)

[Ceph集群大小与机架级sku 22](#_Toc13035)

[Ceph集群大小服务器 SKU 22](#_Toc903)

[Ceph集群规模与其他供应商的服务器sku 22](#_Toc9659)

[一些硬件产商与红帽Ceph联合测试文档 23](#_Toc21685)

[物理机部署ceph最低硬件配置要求 23](#_Toc20982)

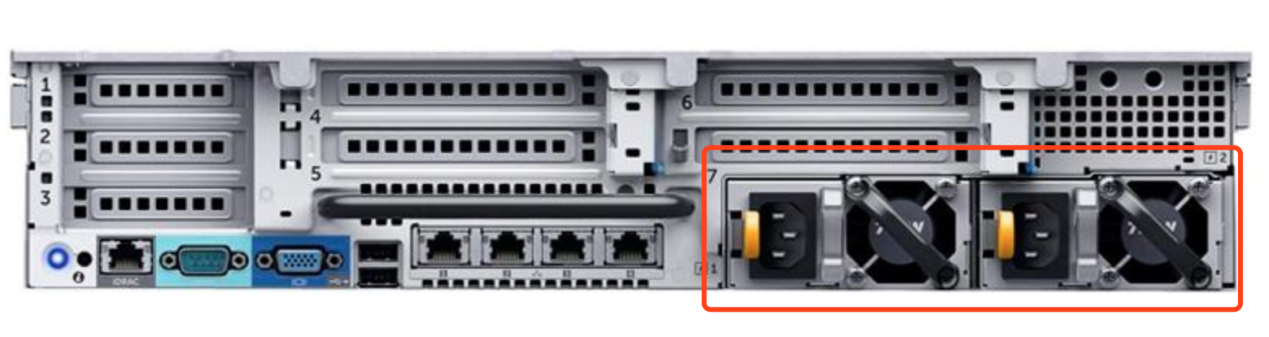
[容器化部署ceph最低硬件配置要求 25](#_Toc6873)

[dashboard硬件最低要求 27](#_Toc8365)

# Ceph硬件选择的一般性原则

## 供电模块

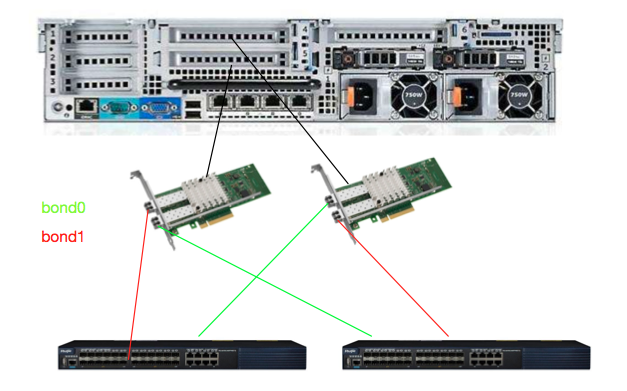
所有生产环境下的单台服务器需要满足最少2路电力供应，同时提供UPS，在发生异常掉电场景下，UPS最少能够提供30~60分钟的应急电力供应，以保障工程师在紧急情况下完成业务的转移或切换。



## 网络配置

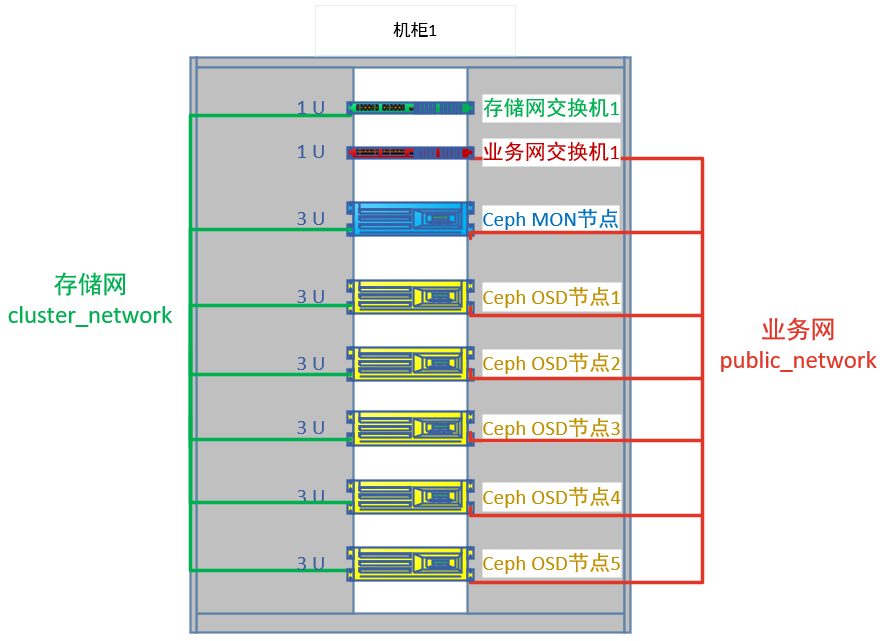
### 服务器网络配置

服务器需要同时接入业务网（public\_network）和存储网（cluster\_network） 两路网络(有安全需求可以做VLAN隔离)，每组网络推荐提供≥10Gbps的网络带宽，同时考虑到网络冗余，每路网络最好能够提供LACP Bonding(mode=4),注意服务器Bond上级联到两台交换机，避免单交换机故障导致网络中断。如下图所示



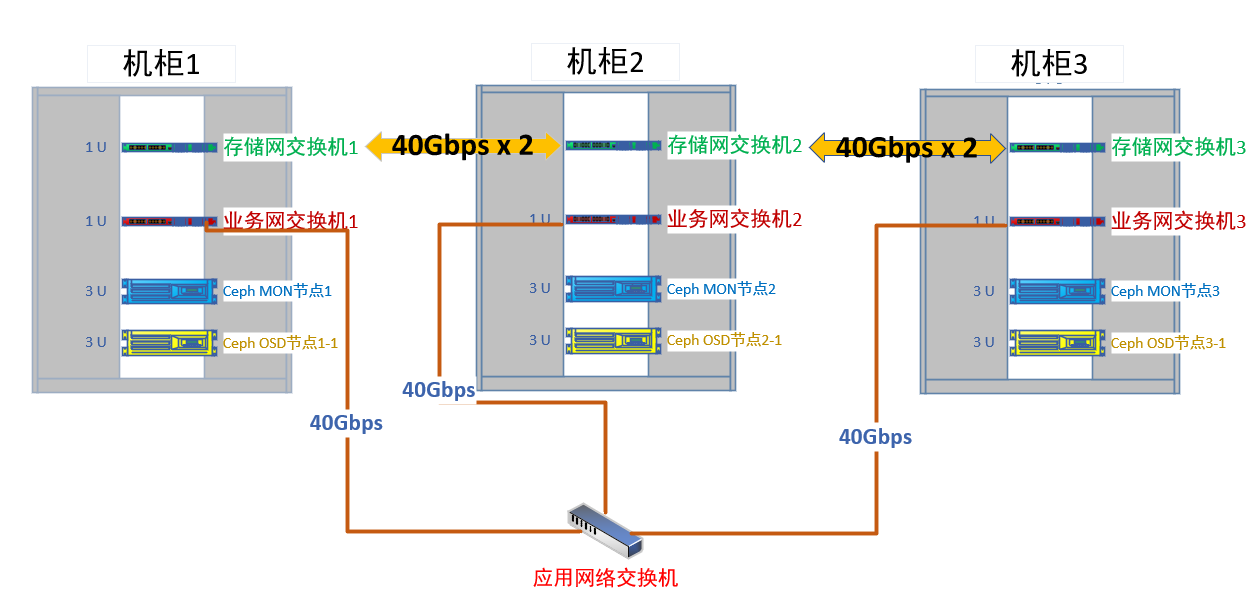
### 机柜网络配置

单机柜内每台节点需要同时接入存储网和业务网交换机，具体网络接线图如下



### 跨机柜网络配置

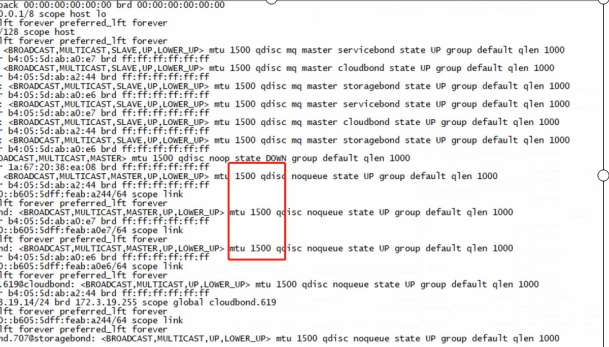
在单个Ceph集群内，跨机柜之间的交换机级联需要满足最少≥80Gbps (40Gbps x 2)的网络带宽，以满足单个存储集群内部的数据同步，同时单个机柜对外最少提供≥40Gbps业务网络访问。如下图所示。



### MTU

建议使用9000字节的MTU 。设置MTU为9000可以获得显著的性能增益。这需要在客户端和Ceph集群之间的路径上的所有交换机和主机上设置（一般客户端是openstack node/openshift node，如果客户端不接受设置MTU=9000，那么则保持所有链路上MTU为默认值1500）。如果使用Nexus系列作为网络交换机，由于Nexus 3k系列最大支持的MTU为9216，建议在Nexus侧设置MTU为9216。建议在服务器端设置MTU 9000。

以下为错误案例

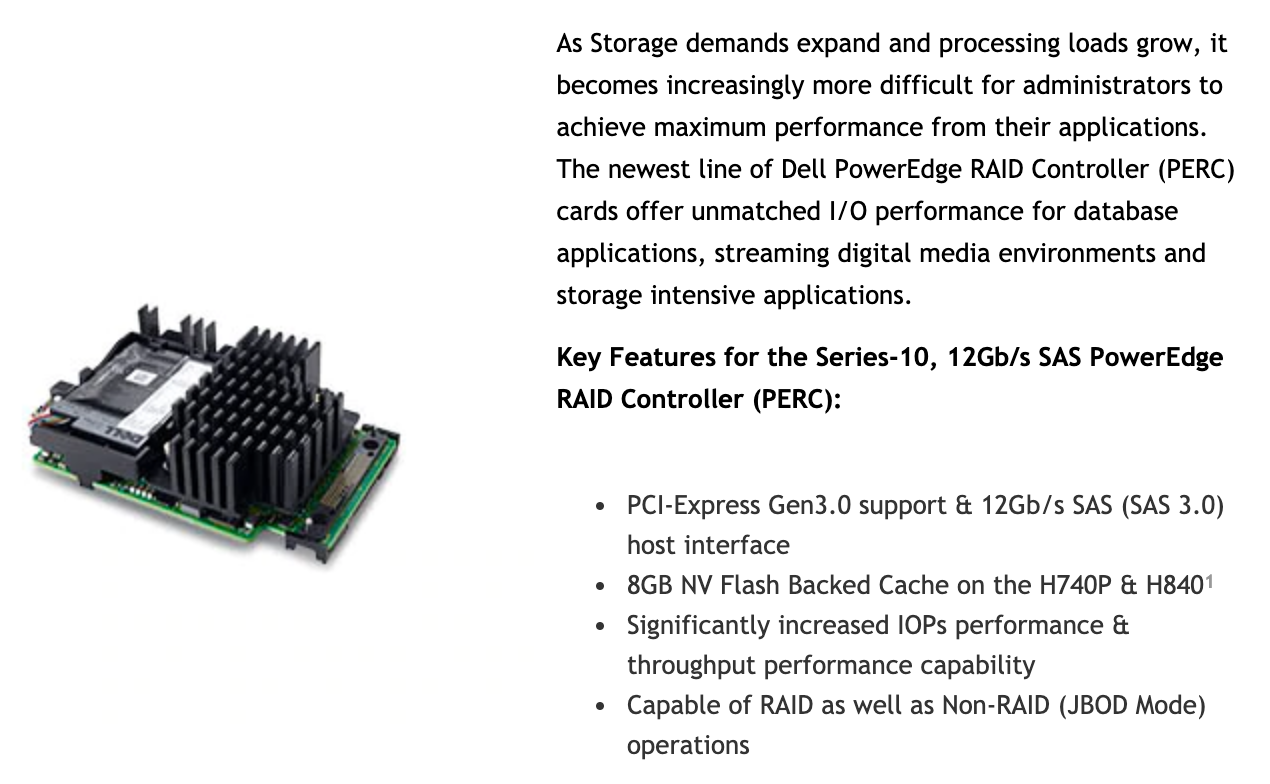


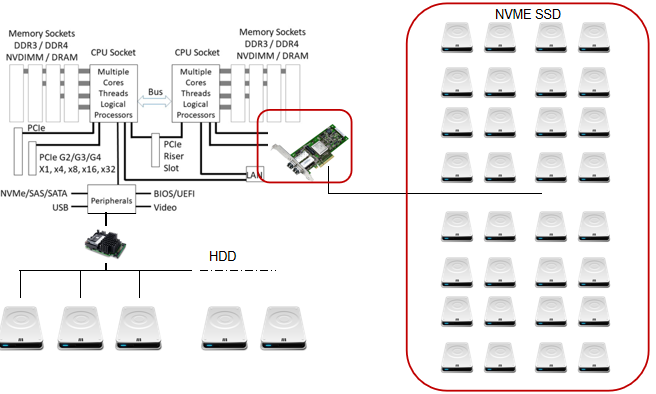
## 磁盘控制器

磁盘控制器需要足够的吞吐量，一般服务器产商会根据服务器大小和盘位配置合适吞吐量的raid卡。在ceph的服务器配置中，如果raid卡支持jbod模式，优先采用jbod(non-raid，直通模式)，如果不支持jbod模式，设置硬盘为raid0，并设置raid卡缓存为write-through模式。

磁盘不建议使用smr硬盘



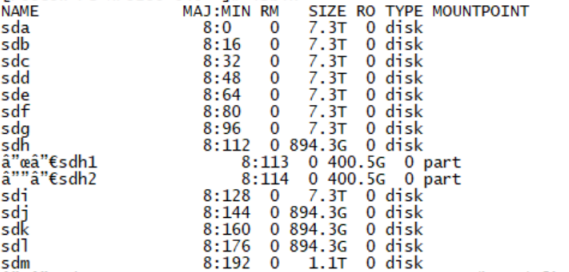


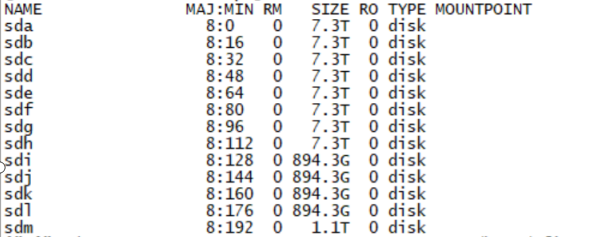
****

**服务器盘符保持一致**

### rhel操作系统中的盘符保持一致

盘符不一致会增加实施部署运维的难度，所以在服务器上线的时候我们需要验收，保持服务器盘符的一致，以下为服务器盘符不一致的案例

****

****

### 数据盘建议做成lvm以对应盘符变化

服务器拔盘后重启操作系统，linux的正常处理方式是把盘符顺延，这样就会导致osd找不到对应的盘符，如果在磁盘上做逻辑卷lvm就可以避免这样的问题，一个osd对应一个lv，lv对应磁盘的uuid。

操作系统盘设计

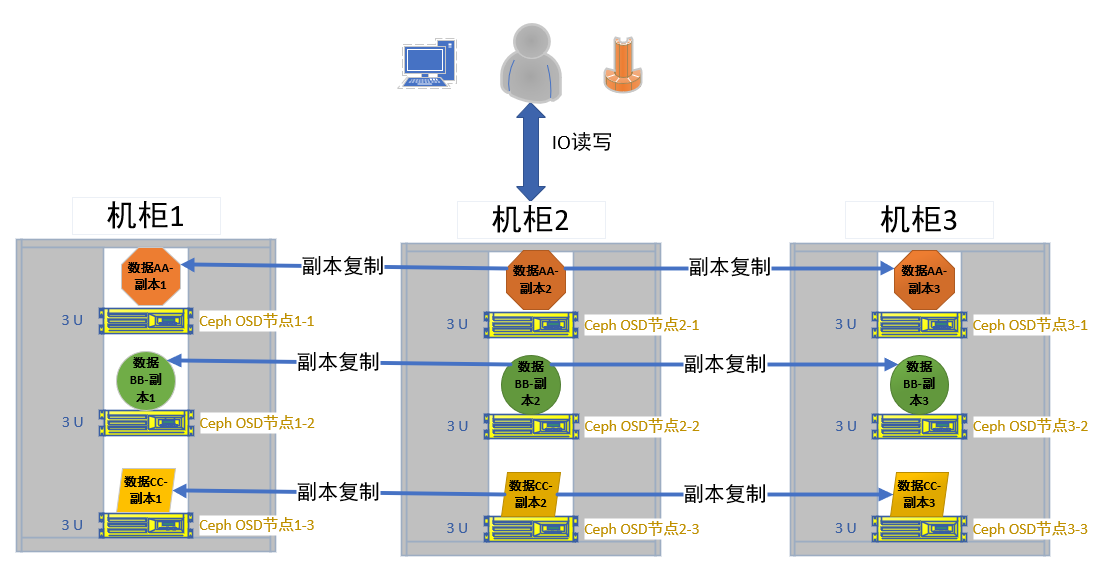
推荐使用SAS磁盘，操作系统盘做raid1，提供高可用性，当一个磁盘失效，操作系统能直接切换到备份磁盘上读写。

/var目录建议独立分区，并且大小推荐300G，这样在故障分析的时候有足够的空间可以调高日志等级。

# 服务高可用和数据冗余容灾设计

## osd数据冗余

数据的冗余通过Ceph内置的Crush算法，实现一份数据同时写入到多个节点(推荐3副本)，同时考虑到单个机柜可能会出现整柜故障(一般是电力或者网络),因此数据需要分布到多个机柜，实现最少机柜级别的数据冗余。

****

## 服务冗余

单个Ceph集群一般由MON、RGW/MDS、OSD这几个角色组成，不同的角色组合对应不同的存储服务类型，具体对应入下表所示。

|  |  |  |
| --- | --- | --- |
| **服务类型** | **服务角色** | **备注** |
| 块存储服务 | OSD、MON | 单机群MON 满足最少3个即可，OSD数量控制在2500以内为宜.(其他类型的服务相同) |
| 对象存储服务 | OSD、MON、RGW | RGW服务的高可用需要借助SLB，单集群配置≥3个。 |
| 文件存储服务 | OSD、MON、MDS | 多个MDS组成Active-Standby模式，单集群配置≥2个。 |

考虑到单机柜可能出现故障，因此每一种服务类型都需要部署在多个机柜中，实现最小都能满足机柜级别的容灾，推荐以单集群最少要分布在3个机柜上。

### mon服务

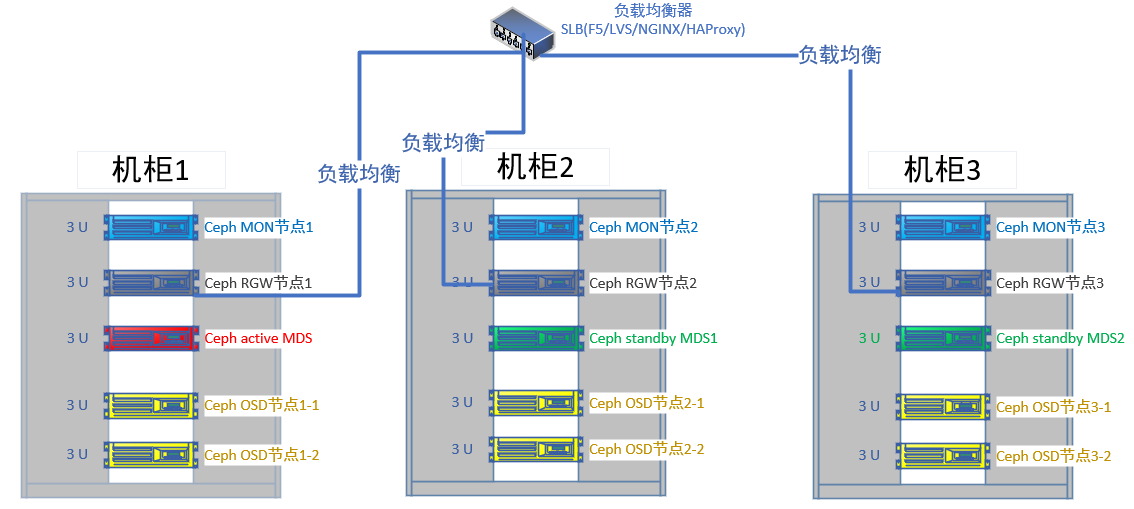
MON之间需要运行Paxos算法选举leader(奇数个节点)，同时要满足MON服务的容灾，因此最少要3个以上的节点，如果OSD数量超过500个，建议采用SSD用来存储MON的数据(加速底层LevelDB/RocksDB的compact)。

### mds服务

MDS目前有Active-Active和Active-Standby两种模式的高可用方案，比较推荐使用Active-Standby，因此最少需要2个以上节点。

### rgw服务

RGW服务对外提供HTTP接口，并且每台RGW相对独立，原生的Ceph没有实现其服务的高可用，因此需要借助外部的4层或者7层负载均衡设备，因此RGW的高可用需要借助SLB这边的服务来实现。



## mon,rgw,mds服务部署在osd服务器上所需硬件资源

所有服务必须容器化部署

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 服务类型 | CPU | 内存 | 磁盘 | 网卡 |
| ceph-mon-container | 1x AMD64 or Intel64 cpu core | 5G | 使用操作系统/var分区50G | 2x 10GB Ethernet NICs |
| ceph-mgr-container | 1x AMD64 or Intel64 cpu core | 3G | 使用操作系统/var分区20G | 2x 10GB Ethernet NICs |
| ceph-radosgw-container | 1x AMD64 or Intel64 cpu core | 5G | 使用操作系统/var分区20G | 2x 10GB Ethernet NICs |
| ceph-mds-container | 1x AMD64 or Intel64 cpu core | 5G  可以配置读缓存mds\_cache\_memory\_limit 。根据内存大小配置一般为50G~200G | 使用操作系统/var分区20G | 2x 10GB Ethernet NICs |

## mon, rgw, mds单独物理机部署推荐硬件配置参考

### mon节点

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 机器类型 | CPU | 内存 | 磁盘 | 网卡 | 电源 | 备注 |
| 1U-机架式 | Intel E5-2623 v3 3.0GHz单路 | ECC 32GB | 系统盘:500GB 15K RPM SAS \*2 Raid1 | 10G SFP+ 双端口 | 热插拔双电源 | 单集群500个OSD以内 |
| 1U-机架式 | Intel E5-2623 v3 3.0GHz  双路 | ECC 64GB | 系统盘:500GB 固态盘 MLC \*2 Raid 1 | 10G SFP+ 双端口 | 热插拔双电源 | 单集群OSD数量2500 以内 |

### rgw节点

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 机器类型 | CPU | 内存 | 磁盘 | 网卡 | 电源 |
| 1U-机架式 | ≥ Intel E5-2623 v3 3.0GHz 双路 | ECC 128GB | 系统盘:500GB 15K RPM SAS \*2 Raid1 | 10G SFP+ 双端口 | 热插拔双电源 |

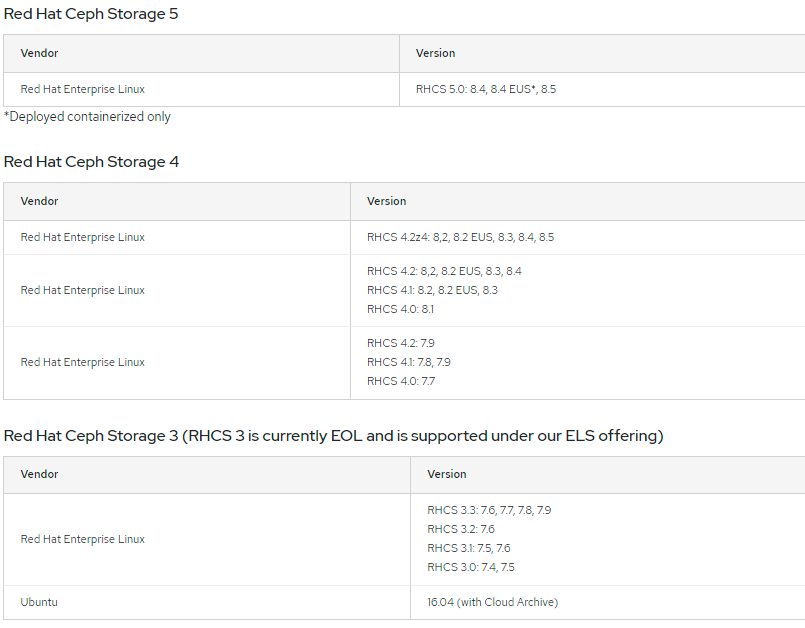
### mds节点

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 机器类型 | CPU | 内存 | 磁盘 | 网卡 | 电源 |
| 1U-机架式 | ≥ Intel E5-2623 v3 3.0GHz双路 | ECC 128GB | 系统盘**:5**00GB 15K RPM SAS \*2 Raid1 | 10G SFP+ 双端口 | 热插拔双电源 |

## osd硬件配置分为性能型，吞吐型，成本容量型

后续章节根据业务需求，做具体的数据硬件推荐

# 操作系统版本与Ceph版本对应关系



# Ceph集群规模限制

磁盘最大不超过12TB

每个数据节点不超过36个osd

集群不超过2500个osd(如果需要达到5000个osd，需要红帽技术支持做review)

集群用作对象存储和文件存储， ssd加速分区的大小为osd盘大小的4%

集群用作块存储，ssd加速分区的大小为osd盘大小的1%

经过红帽测试验证过的Ceph对象存储最大100PB

经过红帽测试验证过的Ceph块存储最大10PB

经过红帽测试验证过的Ceph文件存储最大1PB

# 高性能型Ceph集群配置推荐

## 硬件配置的目标

每GB最高的IOPS

## 典型应用场景

IO密集型场景，比如数据库（mysql postgresql MariaDB ）

## 硬件规格

CPU：每个 NVMe SSD 6 个核，假设 CPU 为 2 GHz。(非NVME的SSD，每个SSD OSD对应2个核)

RAM： 16 GB 基线，加上每个 OSD 5 GB。

网络： 10 Gb 以太网对应每 2 个 OSD。

OSD 介质：高性能、高耐用性的企业级 NVMe SSD。

OSD：每个 NVMe SSD 两个osd。

Bluestore WAL/DB：高性能、高耐用性的企业 NVMe SSD，与 OSD 共存。

磁盘控制器：原生 PCIe 总线。

## 联合测试产商

### 超微服务器sku与Ceph集群规模的大小

| **供应商** | **小型 (250TB)** | **中等 (1PB)** | **大 (2PB+)** |
| --- | --- | --- | --- |
| supermicro | SYS-5038MR-OSD006P | 不适用 | 不适用 |
| [超微服务器的详细信息](https://www.supermicro.com/zh_cn/solutions/red-hat-ceph) | | | |

# 高吞吐量型Ceph集群配置推荐

## 硬件配置的目标

每TB最高的MBps

## 典型应用场景

视频，音频，图片，流媒体

## 硬件规格

CPU：每个 HDD 0.5 个核心，假设 CPU 为 2 GHz。

RAM： 16 GB 基线，加上每个 OSD 5 GB。

网络：每 12 个 OSD 10 GbE，每个用于客户端和面向集群的网络。

OSD 媒体： 7200 RPM 企业级sata/sas硬盘。

OSD：每个 HDD 一个。

Bluestore WAL/DB：SATA SSD。

主机总线适配器 (HBA)：硬盘直通(JBOD)。

## 联合测试产商

### Ceph集群规模大小与机架sku

| **供应商** | **小型 (250TB)** | **中等 (1PB)** | **大 (2PB+)** |
| --- | --- | --- | --- |
| supermicro | SRS-42E112-Ceph-03 | SRS-42E136-Ceph-03 | SRS-42E136-Ceph-03 |

### Ceph集群规模大小与服务器sku

| **供应商** | **小型 (250TB)** | **中等 (1PB)** | **大 (2PB+)** |
| --- | --- | --- | --- |
| supermicro | SSG-6028R-OSD072P | SSG-6048-OSD216P | SSG-6048-OSD216P |
| QCT | QxStor RCT-200 | QxStor RCT-400 | QxStor RCT-400 |
| [QCT：QxStor Red Hat Ceph 存储版](https://go.qct.io/solutions/" \l "specifications) | | | |

### 其他厂家服务器型号与Ceph集群

| **供应商** | **小型 (250TB)** | **中等 (1PB)** | **大 (2PB+)** |
| --- | --- | --- | --- |
| 戴尔 | PowerEdge R730XD | DSS 7000 ，双节点 | DSS 7000，双节点 |
| 思科 | UCS C240 M4 | UCS C3260 [[c]](https://access.redhat.com/documentation/en-us/red_hat_ceph_storage/4/html/hardware_guide/server-and-rack-solutions_hw" \l "ftn.rhc3260) | UCS C3260 [[d]](https://access.redhat.com/documentation/en-us/red_hat_ceph_storage/4/html/hardware_guide/server-and-rack-solutions_hw" \l "ftn.c3260) |
| 联想 | System x3650 M5 | System x3650 M5 | 不适用 |
| [[a]](https://access.redhat.com/documentation/en-us/red_hat_ceph_storage/4/html/hardware_guide/server-and-rack-solutions_hw" \l "r730xd)有关详细信息， 请参阅[适用于 Red Hat Ceph 存储的 Dell PowerEdge R730xd 性能和选型指南 - Dell Red Hat 技术白皮书](http://en.community.dell.com/techcenter/cloud/m/dell_cloud_resources/20442913/download)。  [[c]](https://access.redhat.com/documentation/en-us/red_hat_ceph_storage/4/html/hardware_guide/server-and-rack-solutions_hw" \l "rhc3260)有关详细信息， 请参阅[红帽 Ceph 存储硬件参考架构](https://www.redhat.com/en/resources/resources-red-hat-ceph-storage-hardware-selection-guide-html)。  [[d]](https://access.redhat.com/documentation/en-us/red_hat_ceph_storage/4/html/hardware_guide/server-and-rack-solutions_hw" \l "c3260) 详见UCS [C3260](https://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-c3260-rack-server/index.html) | | | |

# 节约成本容量型Ceph集群配置推荐

## 硬件配置的目标

每TB最低的成本

## 典型应用场景

文件对象归档，数据备份，日志库

## 硬件规格

CPU：每个 HDD 0.5 个内核，假设 CPU 为 2 GHz。

RAM： 16 GB 基线，加上每个 OSD 5 GB。

网络：每 12 个 OSD 10 GbE。

OSD 媒体： 7200 RPM 企业级硬盘。

OSD：每个 HDD 一个。

Bluestore WAL/DB：位于 HDD 上。

HBA： JBOD。

## 联合测试产商

### Ceph集群大小与机架级sku

| **供应商** | **小型 (250TB)** | **中等 (1PB)** | **大 (2PB+)** |
| --- | --- | --- | --- |
| supermicro | 不适用 | SRS-42E136-Ceph-03 | SRS-42E172-Ceph-03 |

### Ceph集群大小服务器 SKU

| **供应商** | **小型 (250TB)** | **中等 (1PB)** | **大 (2PB+)** |
| --- | --- | --- | --- |
| supermicro | 不适用 | SSG-6048R-OSD216P | SSD-6048R-OSD360P |
| QCT | 不适用 | QxStor RCC-400 | QxStor RCC-400 |
| 详情请参阅[Supermicro 的 Ceph 整体解决方案](https://www.supermicro.com/solutions/datasheet_Ceph.pdf)。 | | | |

### Ceph集群规模与其他供应商的服务器sku

| **供应商** | **小型 (250TB)** | **中等 (1PB)** | **大 (2PB+)** |
| --- | --- | --- | --- |
| 戴尔 | 不适用 | DSS 7000，双节点 | DSS 7000，双节点 |
| 思科 | 不适用 | UCS C3260 | UCS C3260 |
| 联想 | 不适用 | system x3650 M5 | 不适用 |

### 一些硬件产商与红帽Ceph联合测试文档

[三星 NVMe SSD 上的红帽 Ceph 存储](https://www.samsung.com/semiconductor/global.semi.static/SamsungNVMeSSDs_Red_Hat_Ceph_Storage_CS-0.pdf)

[在红帽 Ceph 存储上部署 MySQL 数据库](https://www.redhat.com/cms/managed-files/st-ceph-storage-mysql-refarch-technology-detail-inc0448222-201609-en.pdf)

[面向云的英特尔® 数据中心模块 – 采用红帽 Ceph 存储的红帽 OpenStack 平台](https://www.intel.com/content/dam/www/public/us/en/documents/reference-architectures/redhat-cloud-blocks-reference-architecture.pdf)

[QCT 服务器上的红帽 Ceph 存储](https://www.redhat.com/en/files/resources/st-ceph-qct-servers-inc0347490-en.pdf)

[配备英特尔处理器和 SSD 的服务器上的红帽 Ceph 存储](https://www.redhat.com/cms/managed-files/st-rhcs-intel-config-guide-technology-detail-inc0406282-201606-en.pdf)

[红帽数据服务解决方案](https://www.redhat.com/cms/managed-files/st-10-billion-objects-overview-f25947wg-202011-en.pdf)

# 物理机部署ceph最低硬件配置要求

| **程序** | **标准** | **最低推荐** |
| --- | --- | --- |
| ceph-osd | 处理器 | 1x AMD64 or Intel 64 |
| 内存 | 5 GB |
| 操作系统磁盘 | 每台主机 1 个操作系统磁盘 |
| 硬盘 | 每个守护进程 1 个硬盘 |
| 网络 | 2x 1GB Ethernet NICs |
| ceph-mon | 处理器 | 1x AMD64 or Intel 64 |
| 内存 | 1GB |
| 磁盘空间 | 15 GB |
| 网络 | 2x 1GB Ethernet NICs |
| ceph-mgr | 处理器 | 1x AMD64 or Intel 64 |
| 内存 | 1 GB |
| 网络 | 2x 1GB Ethernet NICs |
| ceph-radosgw | 处理器 | 1x AMD64 or Intel 64 |
| 内存 | 1 GB |
| 磁盘空间 | 5 GB |
| 网络 | 2x 1GB Ethernet NICs |
| ceph-mds | 处理器 | 1x AMD64 or Intel 64 |
| 内存 | 2 GB |
| 磁盘空间 | 2GB |
| 网络 | 2x 1GB Ethernet NICs |

# 容器化部署ceph最低硬件配置要求

| **程序** | **标准** | **最低推荐** |
| --- | --- | --- |
| ceph-osd-container | 处理器 | 1x AMD64 or Intel 64 |
| 内存 | 5 GB |
| 操作系统磁盘 | 每台主机 1 个操作系统磁盘 |
| 硬盘 | 每个 OSD 容器 对应一个硬盘 |
| 网络 | 2x 1GB Ethernet NICs |
| ceph-mon-container | 处理器 | 1x AMD64 or Intel 64 |
| 内存 | 3 GB |
| 磁盘空间 | 10 GB |
| 网络 | 2x 1GB Ethernet NICs |
| ceph-mgr-container | 处理器 | 1x AMD64 or Intel 64 |
| 内存 | 3 GB |
| 网络 | 2x 1GB Ethernet NICs |
| ceph-radosgw-container | 处理器 | 1x AMD64 or Intel 64 |
| 内存 | 1 GB |
| 磁盘空间 | 5 GB |
| 网络 | 2x 1GB Ethernet NICs |
| ceph-mds-container | 处理器 | 1x AMD64 or Intel 64 |
| 内存 | 3 GB |
| 磁盘空间 | 2 GB |
| 网络 | 2x 1GB Ethernet NICs |

# dashboard硬件最低要求

2.5 GHz 或更高频率的 4 核处理器

8 GB 内存

50 GB 硬盘驱动器

1个千兆以太网网络接口