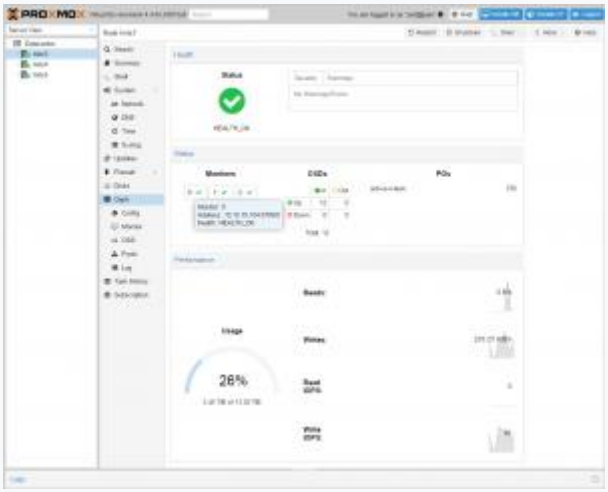
proxmox 6 搭建 ceph 服务

一、介绍



Ceph 是一个分布式对象存储和文件系统，旨在提供出色的性能，可靠性和可伸缩性 - 请参阅：http：//ceph.com。

Proxmox VE 支持 Ceph 的 RADOS 块设备用于 VM 和容器磁盘。Ceph 存储服务通常托管在外部专用存储节点上。此类存储群集总计可达数百个节点，可提供数 PB 的存储容量。

对于较小的部署，还可以直接在 Proxmox VE 节点上运行 Ceph 服务。最近的硬件具有足够的 CPU 功率和 RAM，因此可以在同一节点上运行存储服务和 VM / CT。

本文介绍如何直接在 Proxmox VE 节点上设置和运行 Ceph 存储服务。如果要安装和配置外部 Ceph 存储，请阅读 Ceph 文档。要配置外部 Ceph 存储，请按相应的 Ceph Client 一节中的说明进行操作。

优点

- 通过 Proxmox VE 上的 CLI 和 GUI 支持轻松设置和管理
- 精简配置
- 快照支持
- 自我愈合
- 没有单点故障
- 可扩展到 exabyte 级别
- 设置具有不同性能和冗余特征的池
- 数据被复制，使其具有容错能力
- 运行经济商品硬件
- 无需硬件 RAID 控制器
- 易于管理
- 开源

为什么我们需要一个新的命令行工具（pveceph）？

对于在特定 Proxmox VE 架构中的使用，我们使用 pveceph。Proxmox VE 提供分布式文件系统（pmxcfs）来存储配置文件。

我们使用它来存储 Ceph 配置。优点是所有节点都看到相同的文件，并且不需要使用 ssh / scp 复制配置数据。该工具还可以使用 Proxmox VE 设置中的其他信息。

像 ceph-deploy 这样的工具无法利用该架构。

推荐的硬件

注意： 仅使用 HBA 卡或板载控制器。Raid 控制器可能会产生极端的负面性能影响（JBOD 模式也一样）。

冗余设置至少需要三台相同的服务器。以下是我们的一个测试实验室集群的规格，包括 Proxmox VE 和 Ceph（三个节点）：

- 双 Xeon E5-2620v2,64 GB RAM，Intel S2600CP 主板，Intel RMM，Chenbro 2U 机箱，带 8 个 3.5 英寸热插拔驱动器托架，2 个固定 2.5 英寸 SSD 托架
- 用于 Ceph 流量的 10 GBit 网络（每个服务器一个双 10 Gbit Intel X540-T2，一个 10Gb 交换机 - Cisco SG350XG-2F1）
- 用于 Proxmox VE 安装的单一企业级 SSD(因为我们在那里运行 Ceph 监视器和相当多的日志)，我们每个主机使用一个 Samsung SM863 240 GB。
- 使用至少两个 SSD 作为 OSD 驱动器。您需要高质量的企业级 SSD，不要使用消费级或“PRO”消费级 SSD。在我们的测试设置中，我们有 4 个英特尔 SSD DC S3520 1.2TB，每个主机 2.5“SATA SSD 用于存储数据（OSD，无额外日志） - 此设置提供大约 14 TB 存储。通过使用 3 的冗余，您可以存储高达 4,7 TB（100％）。但要为失败的磁盘和主机做好准备，您绝不应该 100％填满您的存储空间。

- 作为一般规则，OSD 越多越好，推荐使用更快的 CPU（高 GHZ）。NVMe Express 卡也可以，例如慢速 SATA 磁盘与 SSD / NVMe 日志设备的混合。

同样，如果您期望良好的性能，请始终只使用企业级 SSD，通过在免费驱动器托架中添加更多 OSD SSD /磁盘，可以扩展存储。当然，您可以在业务增长时立即添加更多服务器，而不会中断服务并且配置更改最少。

如果您不想在同一主机上运行虚拟机和 Ceph，则只需添加更多 Proxmox VE 节点，并使用这些节点运行 guest 虚拟机，其他节点仅用于存储。

安装 Proxmox VE

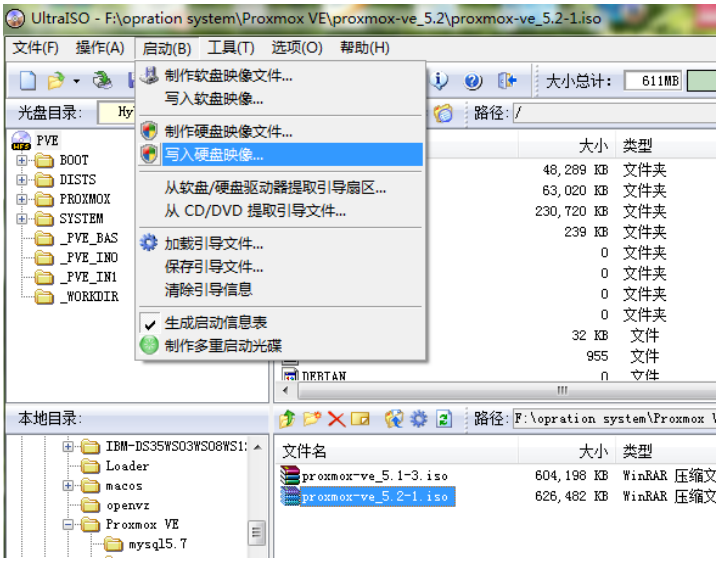
在开始使用 Ceph 之前，您需要一个具有 3 个节点（或更多）节点的 Proxmox VE 集群。我们在快速可靠的企业级 SSD 上安装 Proxmox VE，因此我们可以将所有托架用于 OSD（对象存储设备）数据。只需按照安装和 Cluster_Manager 上众所周知的说明操作即可。

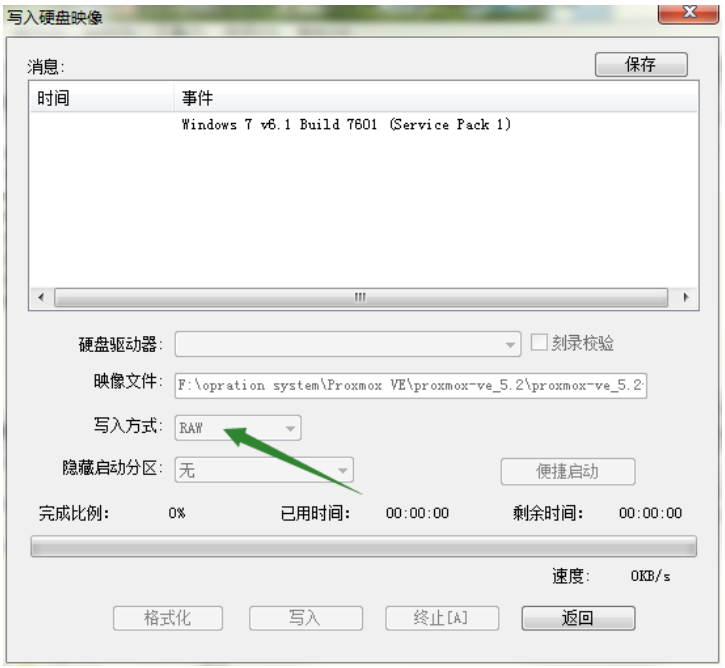
注意：

如果在 SSD 上安装，请使用 ext4（在安装 ISO 的引导提示符下，您可以指定参数，例如“linux ext4 swapsize = 4”）。

二、Proxmox VE 安装步骤

1、proxmox.com 下载 proxmox VE iso，用 UltraISO 写入 U 盘。





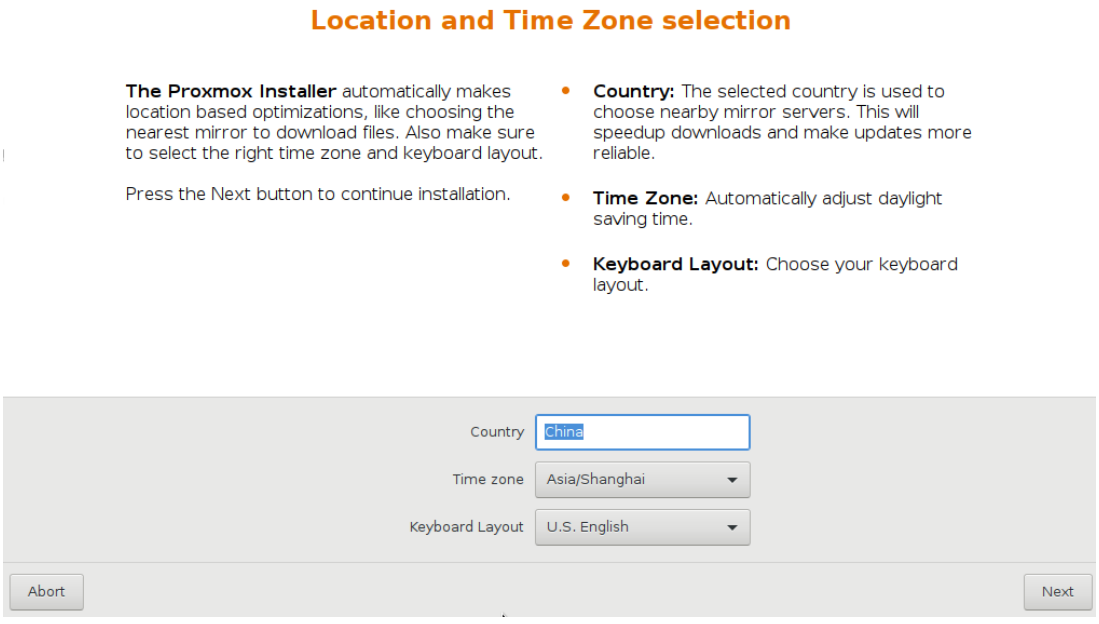
写入方式要选择 RAW，会清空 U 盘数据，请注意。

2、服务器 bios 打开虚拟化相关设置，U 盘启动，安装 proxmox

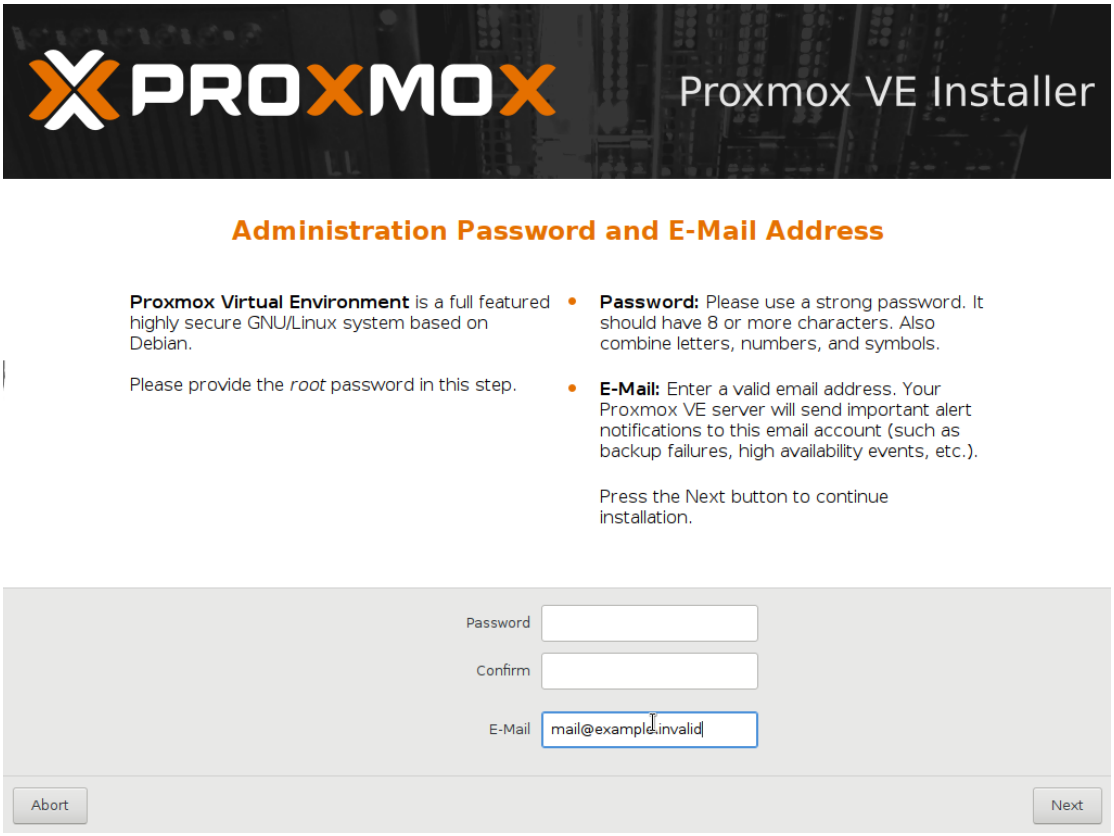

Proxmox VE 5.2 (iso release 1) - http://www.proxmox.com/
PROXMOX
Welcome to Proxmox Virtual Environment
Install Proxmox VE
Install Proxmox VE (Debug mode)
Rescue Boot
Test memory

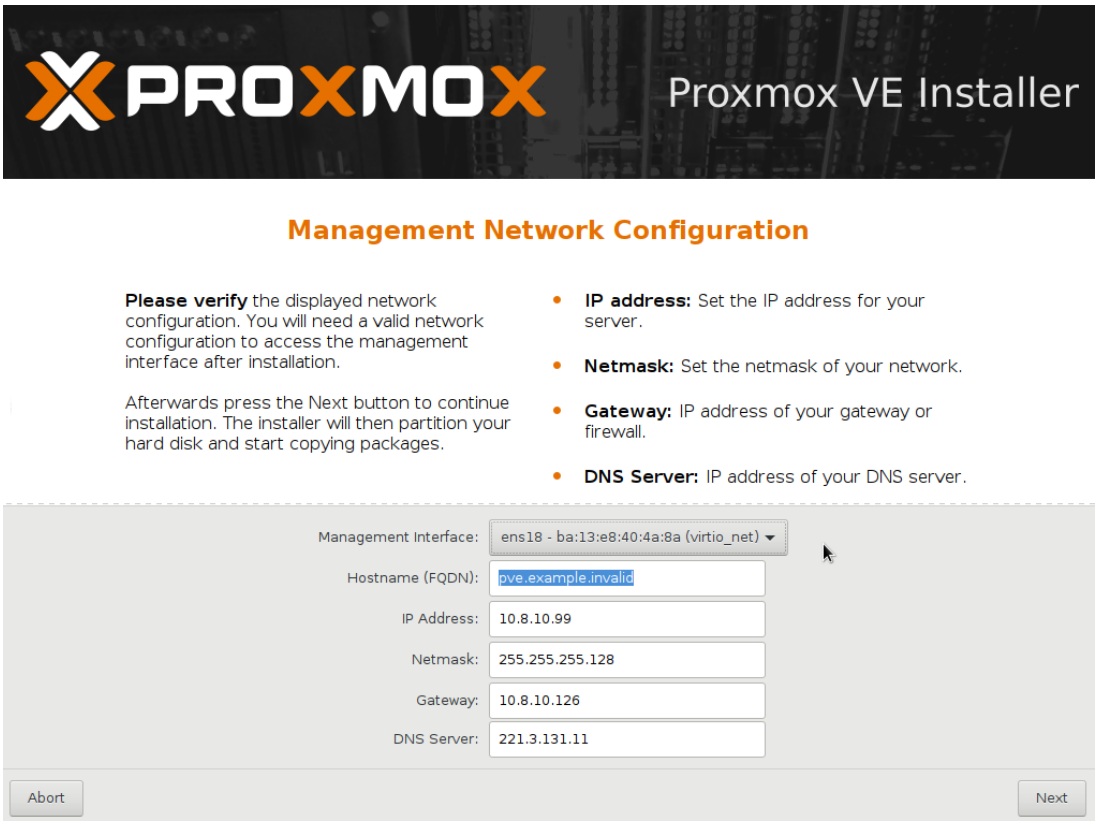

PROXMOX
Proxmox VE Installer
END USER LICENSE AGREEMENT (EULA)
END USER LICENSE AGREEMENT (EULA) FOR PROXMOX VIRTUAL ENVIRONMENT (PROXMOX VE)
By using Proxmox VE software you agree that you accept this EULA, and that you have read and understand the terms and conditions. This also applies for individuals acting on behalf of entities. This EULA does not provide any rights to Support Subscriptions Services as software maintance, updates and support. Please review the Support Subscriptions Agreements for these terms and conditions. The EULA applies to any version of Proxmox VE and any related update, source code and structure (the Programs), regardless of the the delivery mechanism.
1. License. Proxmox Server Solutions GmbH (Proxmox) grants to you a perpetual, worldwide license to the Programs pursuant to the GNU Affero General Public License V3. The license agreement for each component is located in the software component's source code and permits you to run, copy, modify, and redistribute the software component (certain obligations in some cases), both in source code and binary code forms, with the exception of certain binary only firmware components and the Proxmox images (e.g. Proxmox logo). The license rights for the binary only firmware components are located within the components. This EULA pertains solely to the Programs and does not limit your rights under, or grant you rights that supersede, the license terms of any particular component.
2. Limited Warranty. The Programs and the components are provided and licensed "as is" without warranty of any kind, expressed or implied, including the implied warranties of merchantability, non-infringement or fitness for a particular purpose. Neither Proxmox nor its affiliates warrants that the functions contained in the Programs will meet your requirements or that the operation of the Programs will be entirely error free, appear or perform precisely as described in the accompanying documentation, or comply with regulatory requirements.
3. Limitation of Liability. To the maximum extent permitted under applicable law, under no
Abort
I agree

选择硬盘

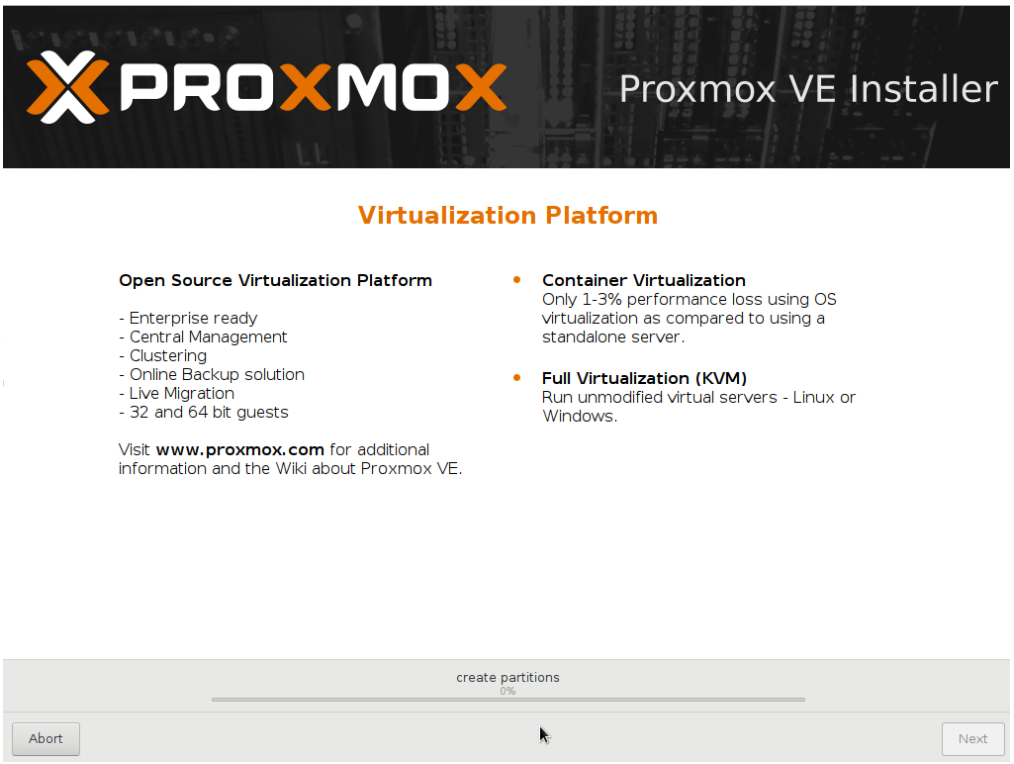


选择时区、键盘布局，然后输入密码和联系邮箱。





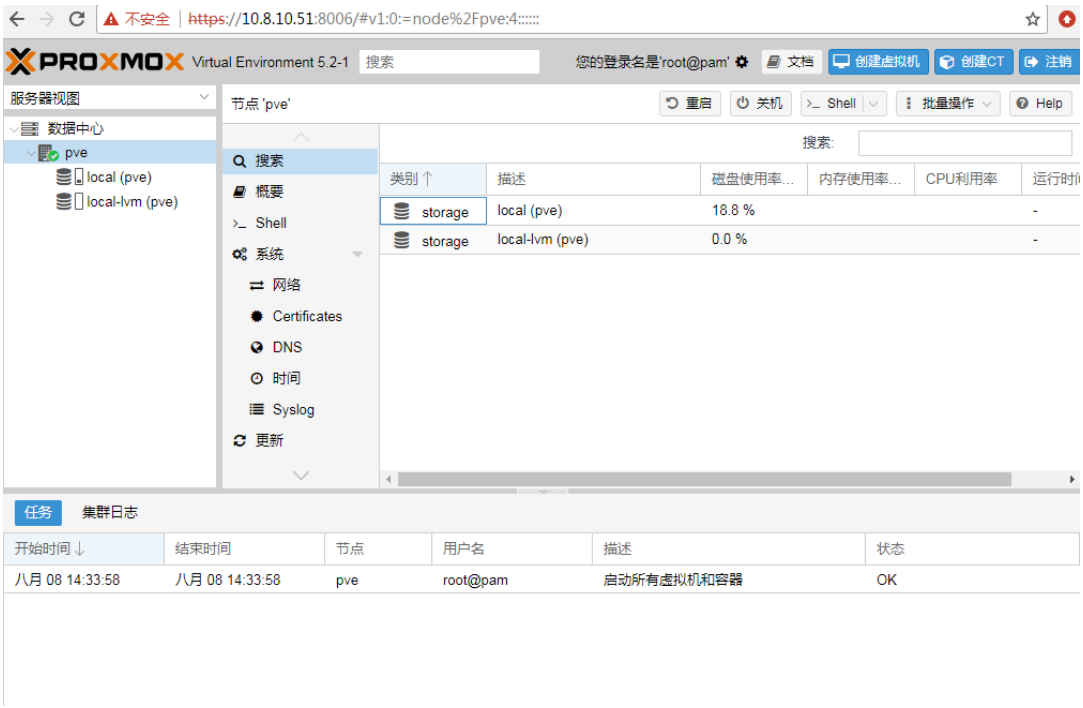
输入 IP 地址，这里选择网卡，IBM 服务器会是管理口，一定要换。

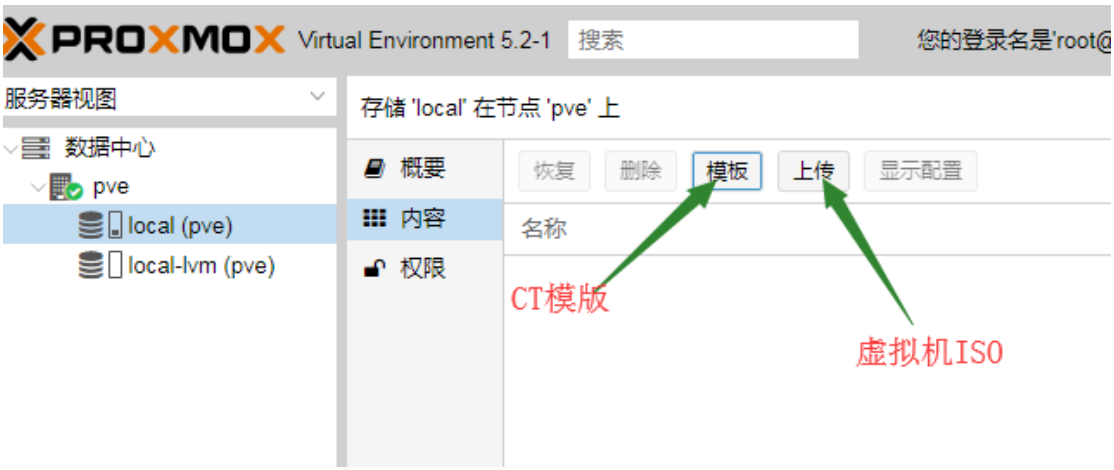


安装完成，重启。

浏览器打开

https://ip:8006 即可访问。





在 local 这里上传 iso 镜像，就可以新建虚拟机。模版中选择一个模版下载就可以新建 CT（容器）。

三、Ceph on Proxmox VE 6

在 Proxmox VE 6 中，Ceph 版本是 Nautilus，稳定和可用于生产环境。

1、环境

硬件：四台 proxmox6 虚拟机 ceph1 ceph2 ceph3 client

每台机器两块硬盘： /dev/vda, /dev/vdb

每台机器两块网卡：ens18 ， ens19

ens18 为系统网络，ens19 为 ceph 专用网卡。

软件：proxmox 6.0-4 内核: 5.0.15

2、安装 proxmox

每台服务器都安装 proxmox。proxmox 的安装参考上一节的内容。

安装完成登录服务器。

a、升级系统补丁

如果没有订阅，修改源为

#vi /etc/apt/sources.list.d/pve-enterprise.list

deb http://download.proxmox.com/debian/pve buster pve-no-subscription

#apt-get update&&apt-get upgrade

b、时间同步 Proxmox VE 默认使用 systemd-timesyncd 作为 NTP 客户端，并默认配置使用一组互联网服务器作为时间源。大部分情况下，系统安装后即可自动实现时钟同步。Ceph 对时间同步要求比较严格，添加自己的时间服务器,这里添加阿里云和苹果的时间服务器

#vi /etc/systemd/timesyncd.conf

[Time]

Servers=ntp1.aliyun.com time.apple.com

重启时钟同步服务后（systemctl restart systemd-timesyncd.service），你可以查看日志以验证新配置的 NTP 服务器是否已被启用（journalctl --since -1h -u systemd-timesyncd）。

注：如果设置了时间同步后，还是报时间问题，可以修改 ceph 配置文件在 global 中加入

mon clock drift allowed = 2

mon clock drift warn backoff = 30

c、配置网络（可以在界面上操作）

安装系统时为服务器配置 ip，例如

ceph1：10.8.10.51

ceph2：10.8.10.52

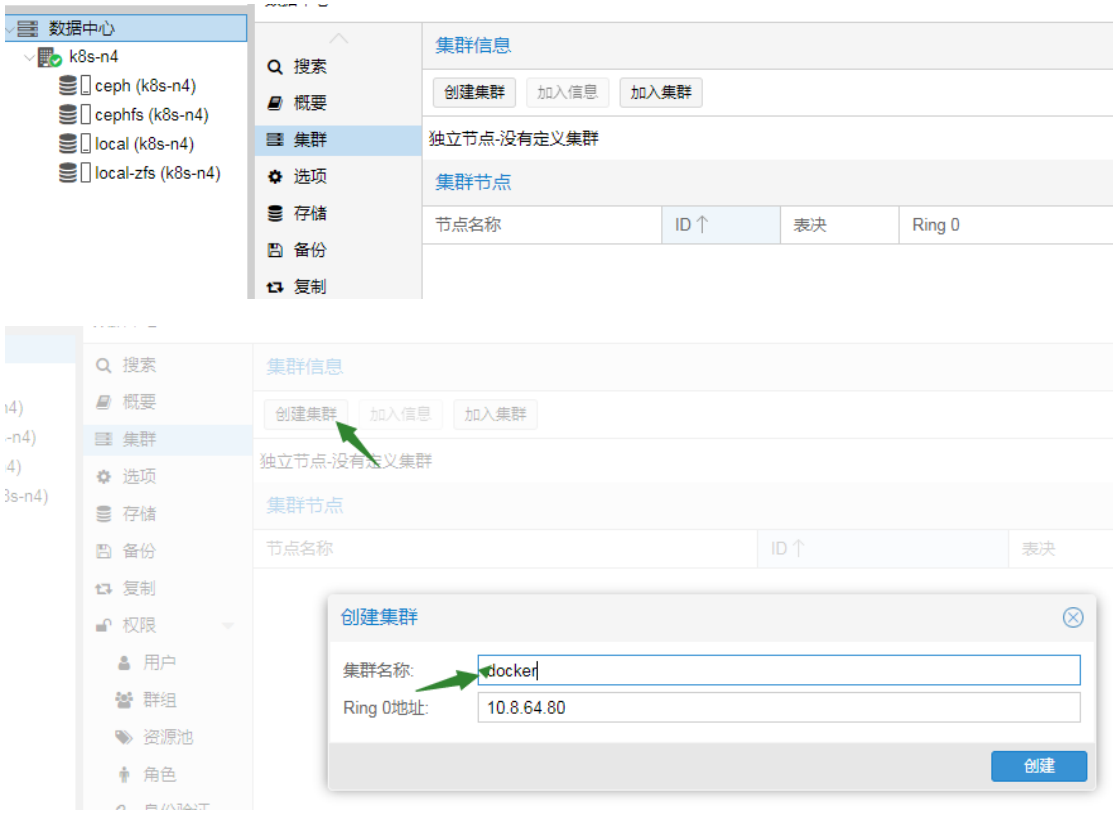
ceph3：10.8.10.53

client： 10.8.10.54

分界线，以下操作可以在 web 界面中完成，也可以用以下命令行完成

————————————————————————————————

3、创建集群



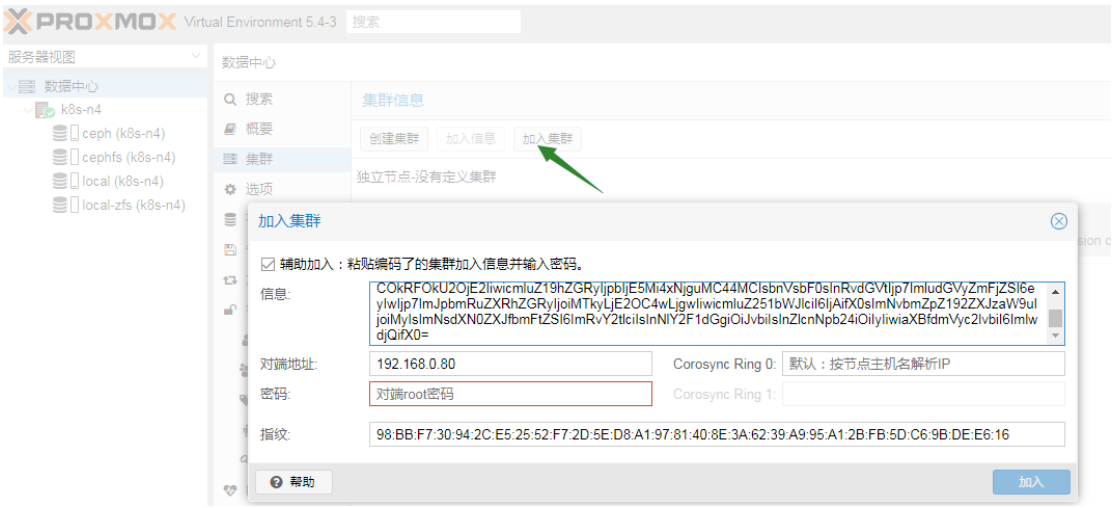
数据中心中集群，进行集群创建。输入名称即可。网络地址为可选，

如果要用独立的网卡来管理集群，可以输入单独的网络地址。

创建完成。

点击加入信息，拷贝加入信息。

其他服务器点击加入集群，在信息中粘贴拷贝的信息。



输入服务器密码，即可加入集群。剩下的服务器同样的操作加入集群。

也可以命令行进行集群创建

各服务器执行命令：

ceph1：

#pvecm create xxx(集群名称)

ceph2：

#pvecm add 10.8.10.51

ceph3：

#pvecm add 10.8.10.51

创建完成后，可以用 pvecm status 查看集群状态。client 这台服务器只是使用 ceph。可以加入集群，也可以不加入。

可以用浏览器访问 https://10.8.10.51:8006 来管理集群，(任意节点都可以管理集群)

4、安装 Ceph

在界面中设置网卡 ip(界面修改完,需要重启动服务器才生效)。或者命令行进行编辑

增加以下内容(每台服务器都配置,建议和系统在不同的网段。)

ceph1:

#vi /etc/network/interfaces

auto ens19

iface ens19 inet static

address 192.168.10.51

netmask 255.255.255.0

ceph2:

#vi /etc/network/interfaces

auto ens19

iface ens19 inet static

address 192.168.10.52

netmask 255.255.255.0

ceph3:

#vi /etc/network/interfaces

auto ens19

iface ens19 inet static

address 192.168.10.53

netmask 255.255.255.0

client:

#vi /etc/network/interfaces

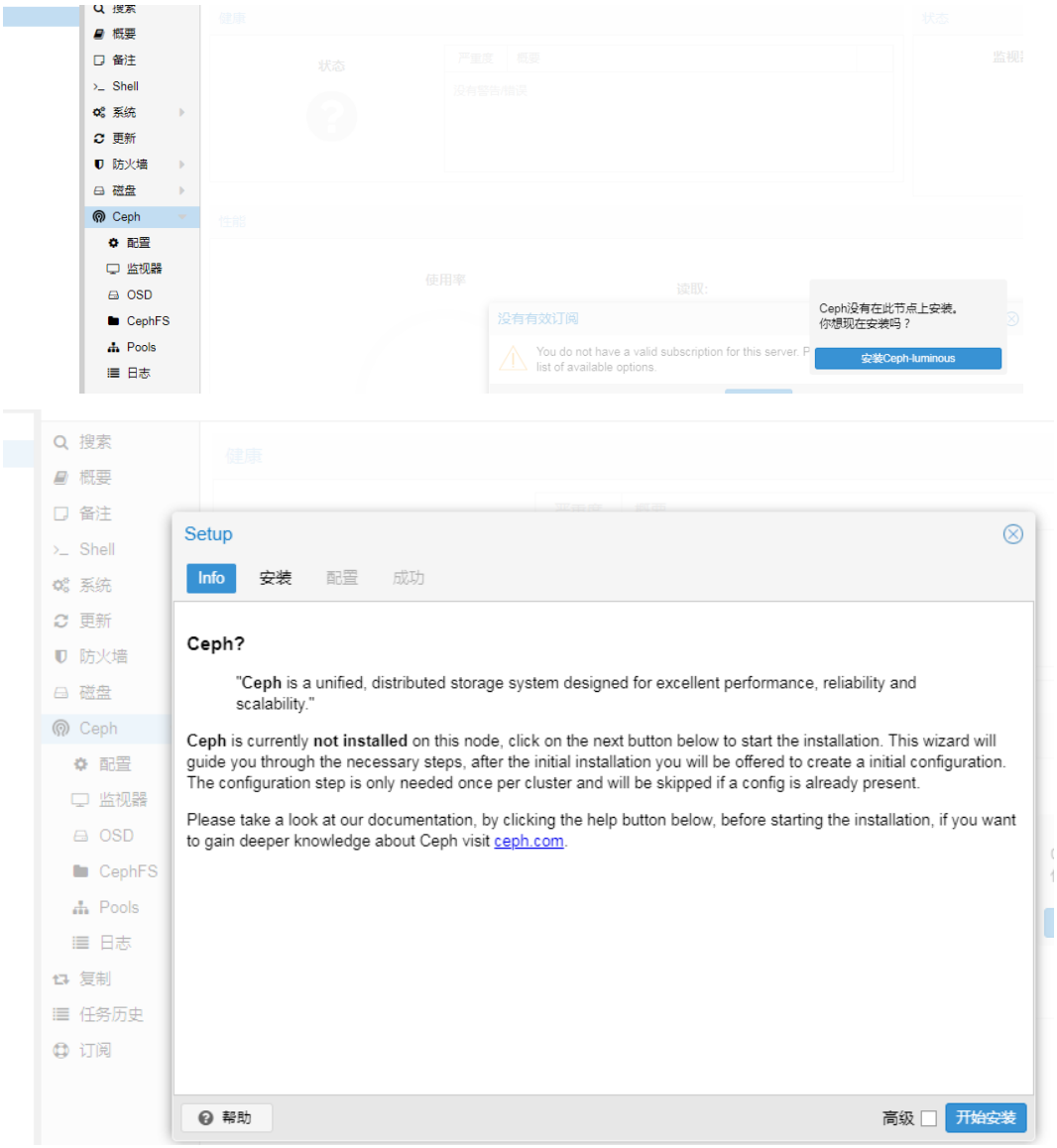
auto ens19

iface ens19 inet static

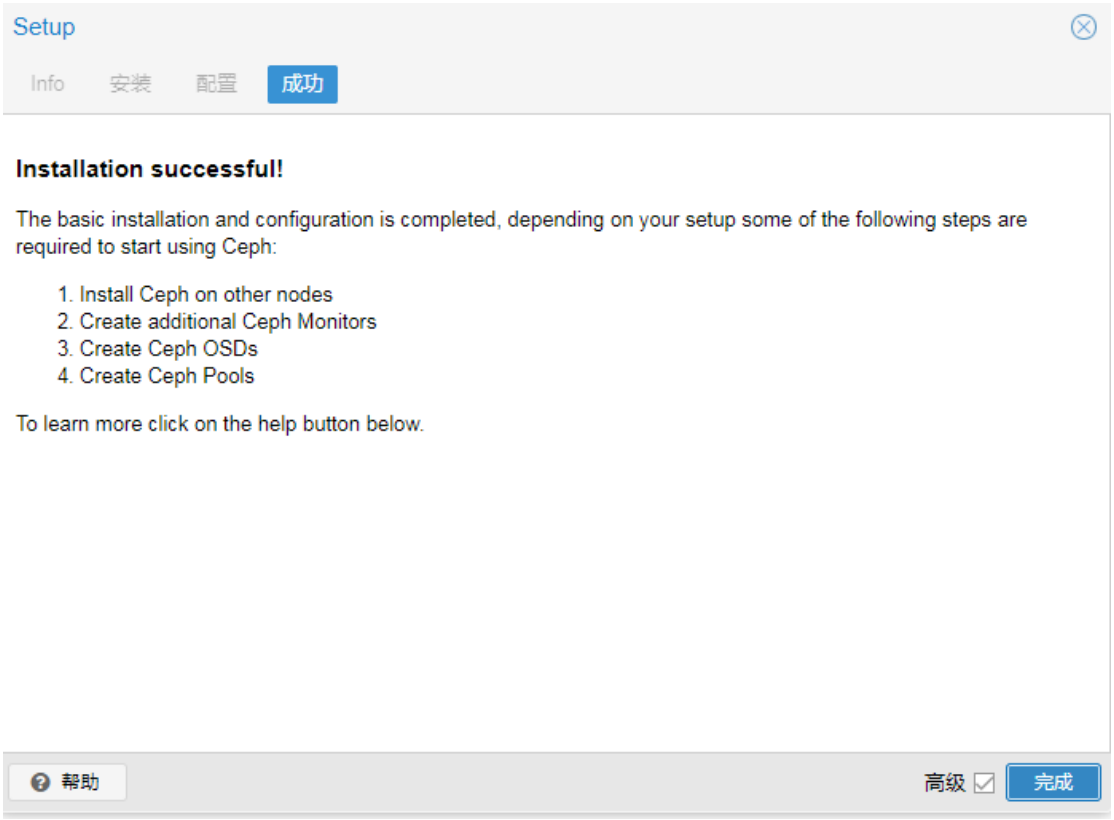
address 192.168.10.54

netmask 255.255.255.0

配置完成重启服务器，点击 ceph，会弹出安装提示，可以按提示进行安装。



点击安装，安装完成后，选择配置。



命令行安装方式：

ceph1、ceph2、ceph3 都运行安装命令

#pveceph install

每一台服务都需要安装，如果下载太慢，可以安装好一台后，复制安装文件到其他服务器进行安装。

复制安装好 ceph 的服务器/var/cache/apt/archives 下*.deb 到本机，然后 dpkg -i *deb 即可。

5、初始化 Ceph 配置

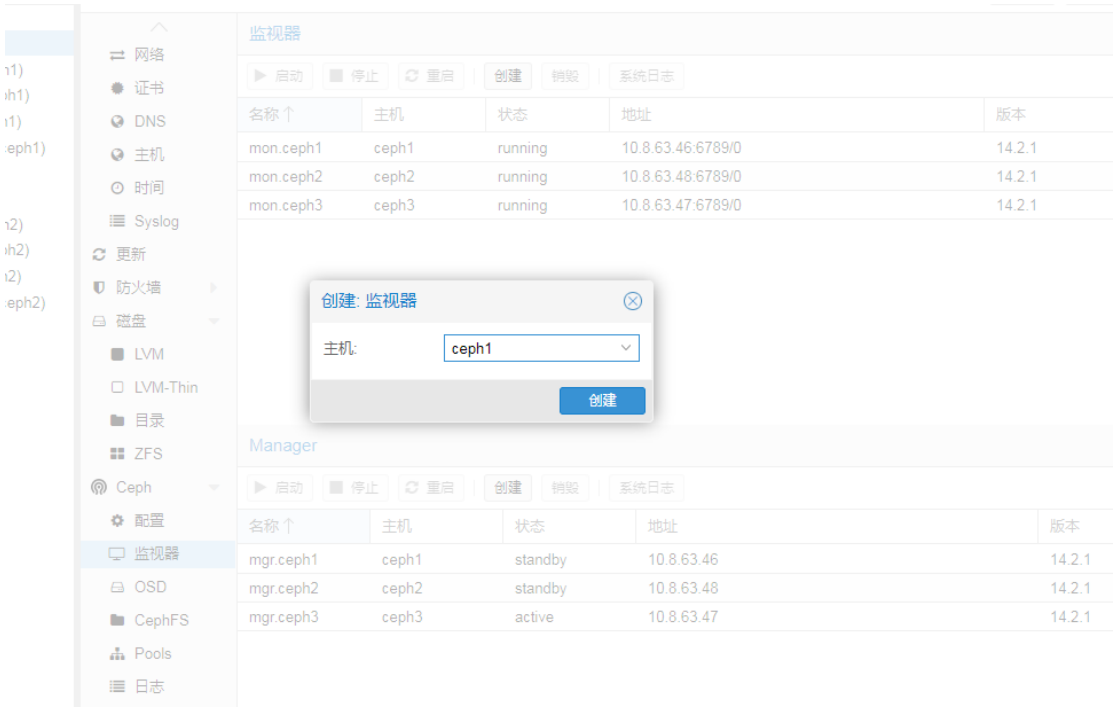
ceph 三个节点的任意一台服务器运行

#pveceph init --network 192.168.10.0/24

注意：新版可以把公共网络(ceph 存储)和集群网络（迁移、克隆）分

开。可以加参数--cluster-network

6、创建监控



命令行创建方式：

ceph 三个节点都运行

#pveceph　createmon

#pveceph　createmgr

注意：luminous 版本以上必须要运行 mgr，一个 ceph 集群中最少需要三个 mon，在创建 mon 的同时系统会自动创建 mgr，在没有创建 mon 的服务器上需要单独创建 mgr。

7、创建 osd

ceph 三个节点都运行

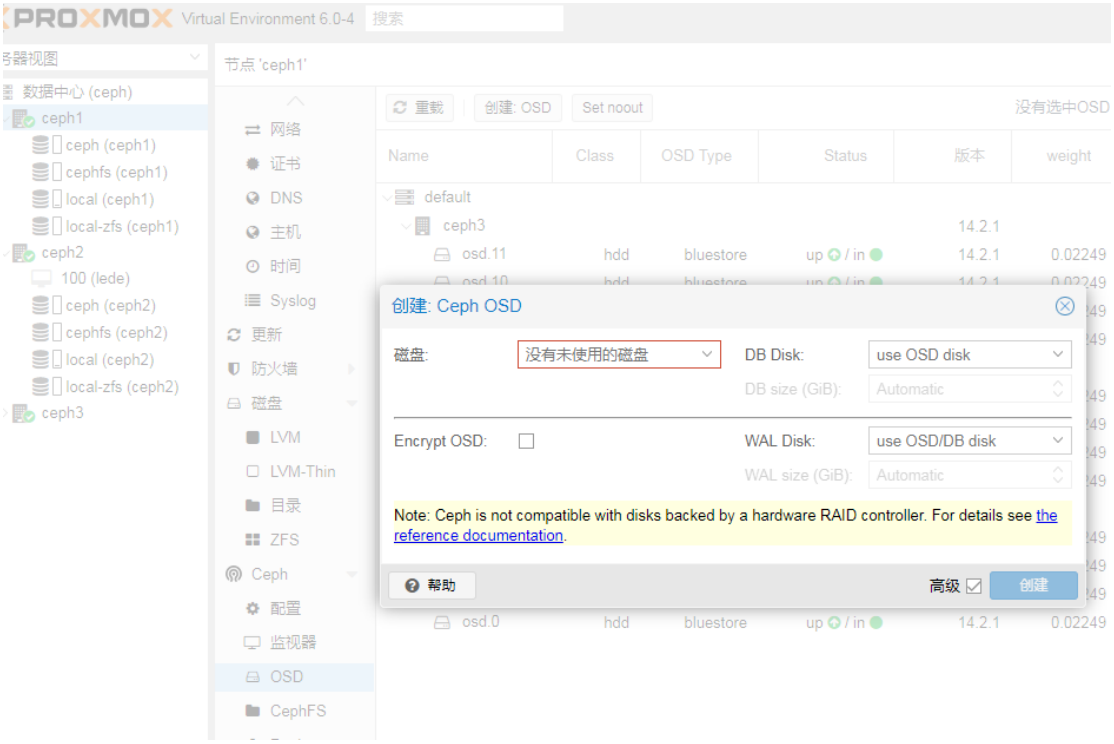
#pveceph createosd /dev/sd[x]

如果使用用过的磁盘创建 osd，需要以下命令删除分区表、引导扇区或 OSD 配置：

```
ceph-volume lvm zap /dev/sd[X] --destroy
```

从 Ceph Kraken 版本开始，引入了一种新的 Ceph OSD 存储类型，即所谓的 Bluestore。这是自 Ceph Luminous 版本之后创建 OSD 时的默认设置。

可以在每一个节点的以下位置创建 osd。



Block.db 和 block.wal

如果要为 OSD 使用单独的 DB / WAL 设备，可以通过 -db_dev 和 -wal_dev 选项指定它。如果没有单独指定，WAL 将与 DB 一起放置。

#pveceph createosd /dev/sd[X] -db_dev /dev/sd[Y] -wal_dev /dev/sd[Z]

您可以直接选择 分别使用-db_size 和-wal_size 参数的大小。如果没

有给出它们，将使用以下值（按顺序）：

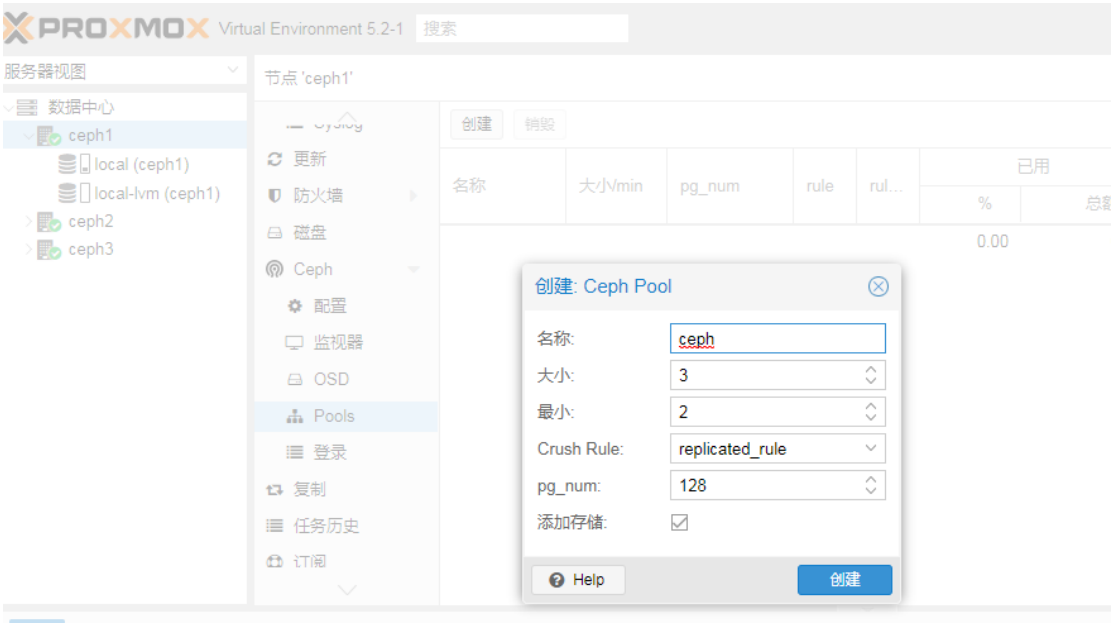
- ceph 配置中 bluestore_block_ {db，wal} _size ...
  - ...database，section [*osd*]
  - ...database，*section* [global]
  - ...file，section [*osd*]
  - ...file，section [global]
- OSD 大小的 10%（DB）/ 1%（WAL）

注意:6.0 有个 bug,这个命令行参数和界面设置都无法自定义 db 和 wal 的大小,需要在 ceph.conf 中添加参数。

8、创建存储池。直接在界面中完成。

存储池需要计算 pg_num。计算公式为 osd 数 x 100 ÷ 副本数（即图中的大小那个选项框）÷ pool 数（如果只建一个池就是 1）

结果必须取最接近该数的 2 的幂



本例，只建一个 pool，那么 pg_num=3×100÷3÷1=100

最接近 100 的 2 的幂是 128。

如果选择了添加到存储这个框,系统会自动为服务器挂载 ceph 盘。

这里可以添加一个 test 的 pool 来简单测试读写速度。

写入速度: rados -p test bench 10 write --no-cleanup

读取速度: rados -p test bench 10 seq

这里的 pg_num 如果设置有问题，后期可以直接修改参数。

ceph -s 查看状态。pool1 为创建的 pool 名

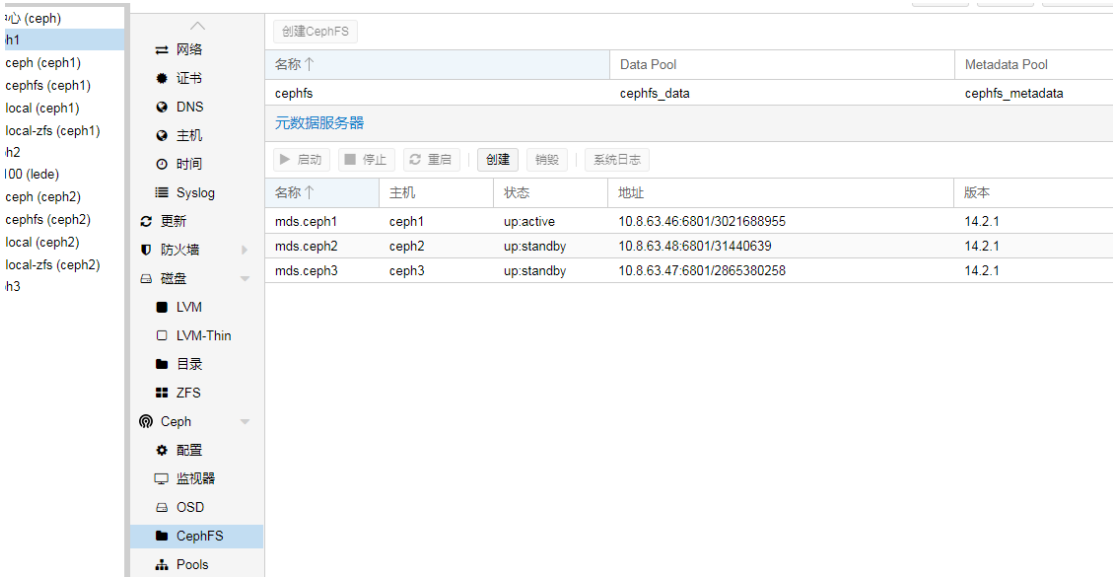
```
ceph  osd  pool  get  pool1  pg_num
ceph  osd  pool  get  pool1  pgp_num
```

查询完成，可以用以下命令设置

```
ceph  osd  pool  set  pool1  pg_num 256
ceph  osd  pool  set  pool1  pgp_num 256
```

9、cephFS

Ceph Luminous 版本之后，支持了 cephFS，直接界面操作，先添加元数据，然后创建 cephfs，勾选添加到存储，自动添加到服务器。



10、其他机器挂载 ceph

这里只举例 proxmox 的服务器。

a、网络

您需要增加一块网卡并配置为存储网络 ip 段，本文中的 192.168.10.54 ，并把这块网卡的网线接入 ceph 服务器的存储网络中。

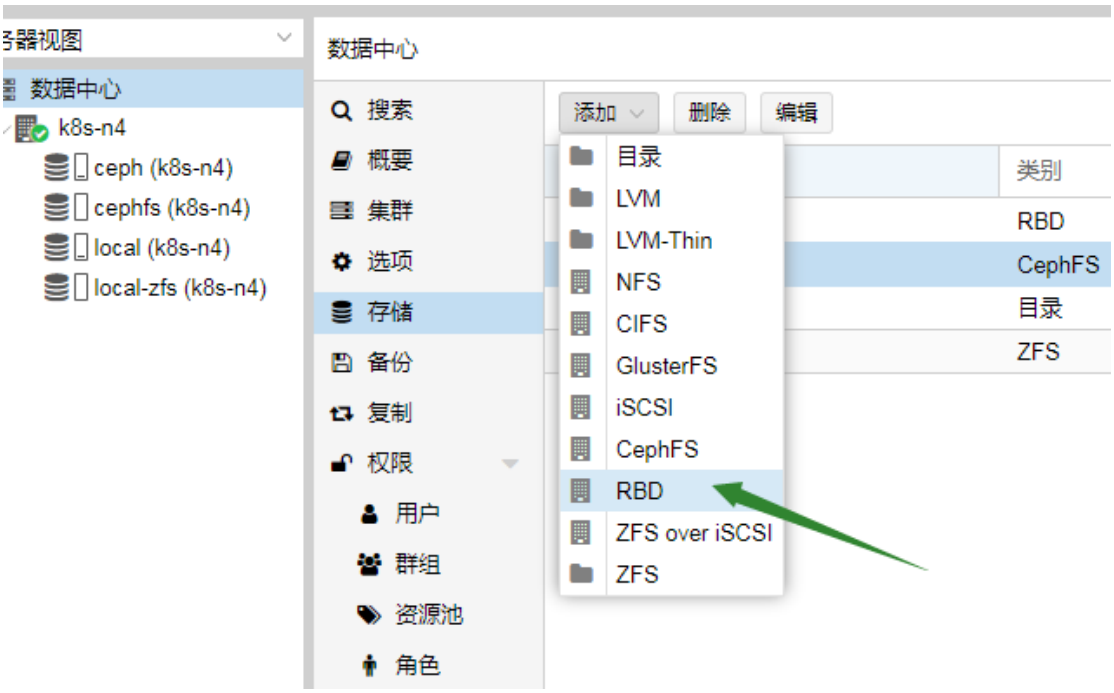
b、key

您还需要将密钥环复制到预定义的位置。

请注意，文件名必须是存储 ID + .keyring 存储 ID 是/etc/pve/storage.cfg 中'rbd：'之后的表达式，它是当前示例中的 ceph-vm

#mkdir /etc/pve/priv/ceph

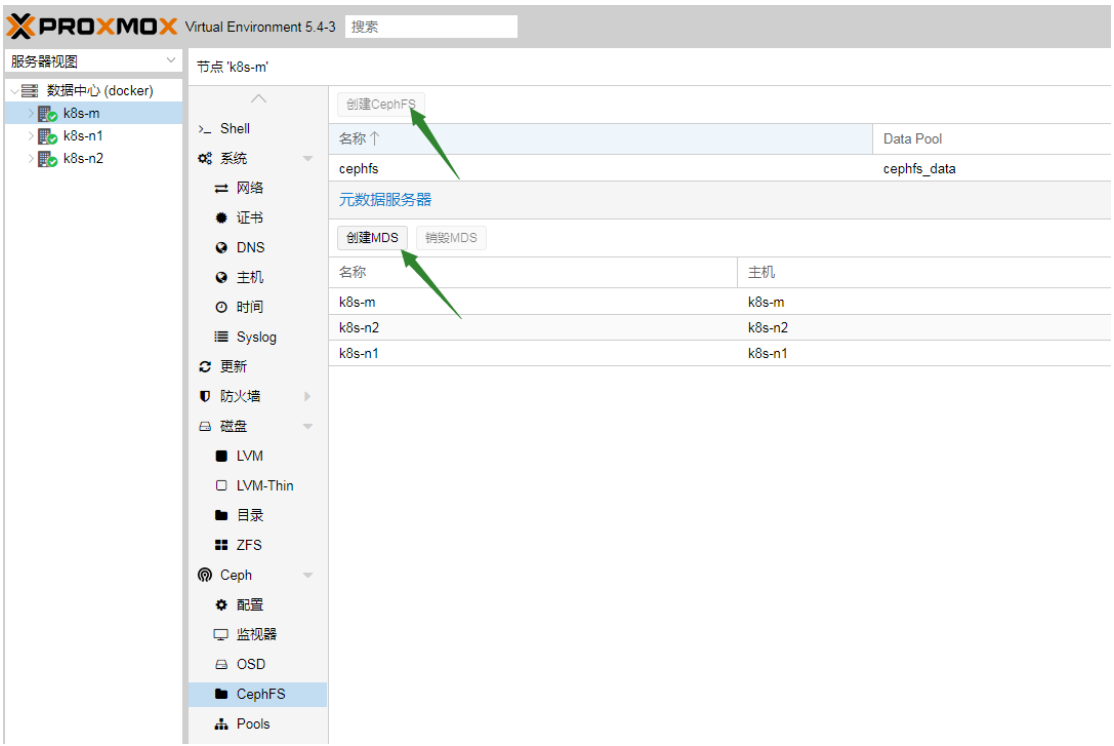
#scp root@10.8.10.51:/etc/pve/priv/ceph.client.admin.keyring /etc/pve/priv/ceph/ceph.keyring



ID：就是存储 id，资源池是 ceph 服务器的存储池名称，monitor 是 ceph 服务器的监视器的 ip，中间用空格分开。用户名默认 admin，内容，虚拟机选择磁盘映像，容器的选容器。

在新增加的存储中概要页能够显示容量，即表示添加成功。如果不能识别容量，请重复该步骤，并检查是否有内容出错。

5.3 版本后，增加了 cephfs 的功能。



创建 cephfs，先创建 MDS，创建完成后，创建 CephFS。集群内的集群创建后就会自动挂载。

如果挂载失败，可以编辑/etc/pve/storage.cfg 手动添加。

```
rbd: ceph
        content rootdir,images
        krbd 0
        pool ceph

cephfs: cephfs
        path /mnt/pve/cephfs
        content iso,vztmpl,backup
        maxfiles 1
```

11、其他机器添加 cephfs

1、 ceph client 必须文件安装。

/sbin/mount.ceph 这个文件。解决办法有两种：

a、从别的 server 拷贝这个文件;

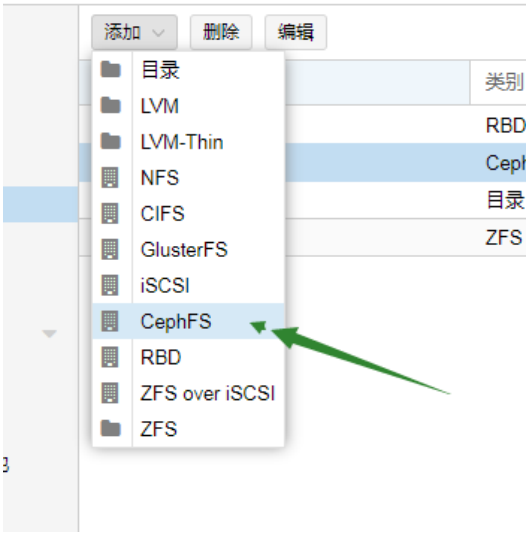
scp root@10.8.64.80:/sbin/mount.ceph /sbin/ceph/mount.ceph

b、安装 ceph-common；

apt-get install ceph-common

2、复制 secret 文件（client key）

scp root@10.8.64.80:/etc/pve/priv/ceph.client.admin.keyring /etc/pve/priv/ceph/cephfs.secret



id 是 ceph 服务器创建 cephfs 的名称，monitor(s)选择 ceph 服务器的 ip，中间用空格分开。内容，选择 dump，iso，模版，片段等。添加,即可完成。如果报错，可以直接编辑/etc/pve/storage.cfg 文件

```
rbd: ceph
        content images,rootdir
        krbd 0
        monhost 10.8.64.80 10.8.64.81 10.8.64.82
        pool ceph
        username admin

cephfs: cephfs
        path /mnt/pve/cephfs
        content iso,vztmpl,backup,snippets
        maxfiles 1
        monhost 10.8.64.80 10.8.64.81 10.8.64.82
        username admin
```

这里显示磁盘空间,为成功，失败请重新添加。