Tedu.cn
达内教育

集群与存储

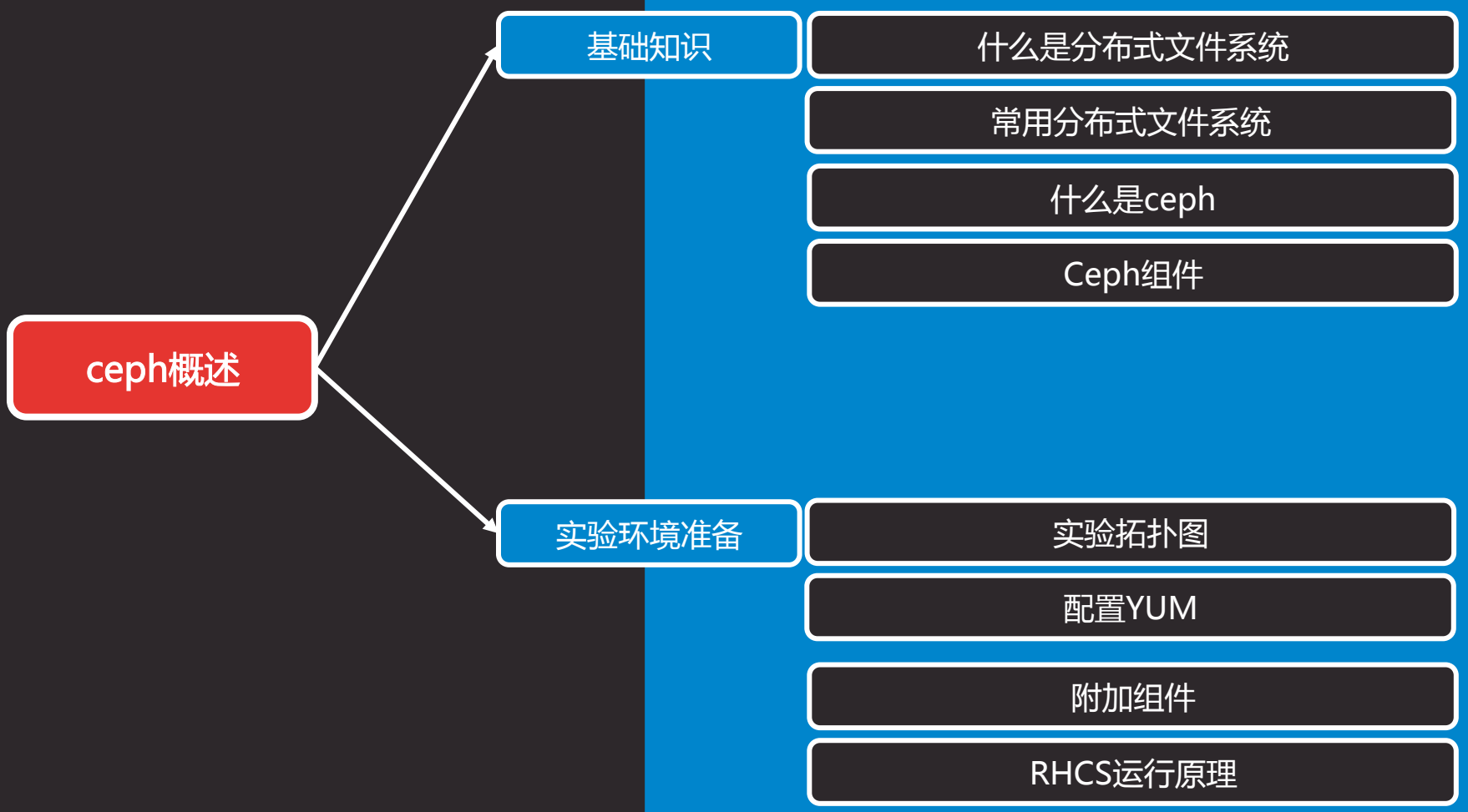
NSD CLUSTER

DAY04

内容

上午	09:00 ~ 09:30	作业讲解和回顾
	09:30 ~ 10:20	ceph概述
	10:30 ~ 11:20	部署Ceph集群
	11:30 ~ 12:20	Ceph块存储
下午	14:00 ~ 14:50	
	15:00 ~ 15:50	块存储应用案例
	16:10 ~ 17:00	
	17:10 ~ 18:00	总结和答疑

ceph概述


ceph概述
基础知识
什么是分布式文件系统
常用分布式文件系统
什么是ceph
Ceph组件
实验环境准备
实验拓扑图
配置YUM
附加组件
RHCS运行原理

Tedu.cn
达内教育

基础知识

什么是分布式文件系统

- 分布式文件系统（Distributed File System）是指文件系统管理的物理存储资源不一定直接连接在本地节点上，而是通过计算机网络与节点相连
- 分布式文件系统的设计基于客户机/服务器模式

知识讲解

常用分布式文件系统

- Lustre
- Hadoop
- FastDFS
- Ceph
- GlusterFS

知识讲解

什么是ceph

- ceph是一个分布式文件系统
- 具有高扩展、高可用、高性能的特点
- ceph可以提供对象存储、块存储、文件系统存储
- ceph可以提供PB级别的存储空间(PB→TB→GB)
  - 1024G*1024G=1048576G
- 软件定义存储(Software Defined Storage)作为存储行业的一大发展趋势，已经越来越受到市场的认可

知识讲解

ceph组件

- OSDs
  - 存储设备
- Monitors
  - 集群监控组件
- MDSs
  - 存放文件系统的元数据（对象存储和块存储不需要该组件）
- Client
  - ceph客户端

知识讲解

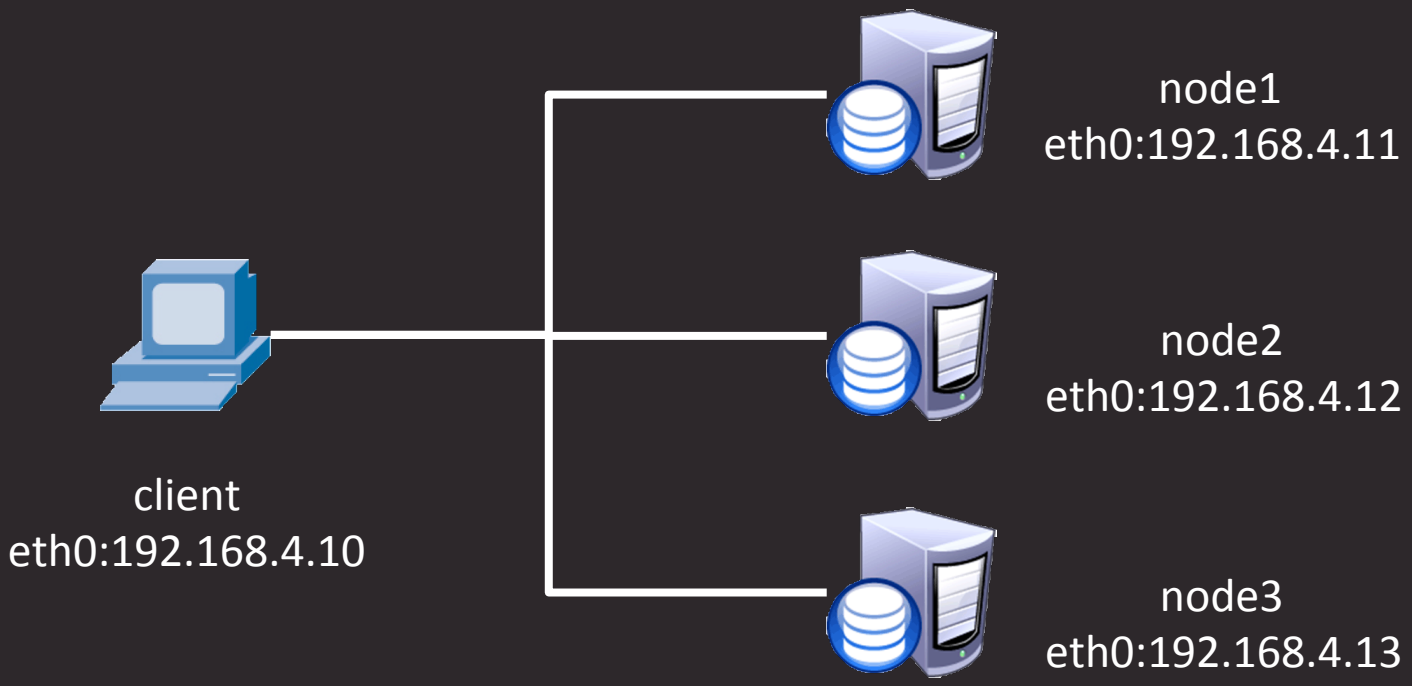
Tedu.cn
达内教育

实验环境准备

实验拓扑图

- 1台客户端虚拟机
- 3台存储集群虚拟机

知识讲解



配置YUM

- 物理机创建网络yum源服务器

```
[root@root9pc01 ~]# yum -y install vsftpd
[root@root9pc01 ~]# mkdir  /var/ftp/ceph
[root@root9pc01 ~]# mount -o loop \
rhcs2.0-rhosp9-20161113-x86_64.iso  /var/ftp/ceph
[root@root9pc01 ~]# systemctl  restart  vsftpd
```

知识讲解

配置YUM（续1）

- 虚拟机调用YUM源（下面以node1为例）

```
[root@node1 ~]# cat /etc/yum.repos.d/ceph.repo
[mon]
name=mon
baseurl=ftp://192.168.4.254/ceph/rhceph-2.0-rhel-7-x86_64/MON
gpgcheck=0
[osd]
name=osd
baseurl=ftp://192.168.4.254/ceph/rhceph-2.0-rhel-7-x86_64/OSD
gpgcheck=0
[tools]
name=tools
baseurl=ftp://192.168.4.254/ceph/rhceph-2.0-rhel-7-x86_64/Tools
gpgcheck=0
```

知识讲解

配置SSH无密钥连接

- 修改主机名

```
[root@node1 ~]# cat /etc/hosts
... ...
192.168.4.10       client
192.168.4.11       node1
192.168.4.12       node2
192.168.4.13       node3
[root@node1 ~]# for i in 10  11  12  13
> do
> scp  /etc/hosts  192.168.2.$i:/etc/
> done
```

知识讲解

配置SSH无密钥连接（续1）

知识讲解

- 非交互生成密钥对

```
[root@node1 ~]# ssh-keygen   -f /root/.ssh/id_rsa    -N ''
```

- 发布密钥到各个主机（包括自己）

```
[root@node1 ~]# for i in 10  11  12  13
> do
> ssh-copy-id  192.168.4.$i
> done
```

NTP时间同步

- 客户端创建NTP服务器

```
[root@client ~]#  yum -y install chrony
[root@client ~]#  cat /etc/chrony.conf
server 0.centos.pool.ntp.org iburst
allow 192.168.4.0/24
local stratum 10
[root@client ~]# systemctl  restart  chronyd
```

- 其他所有主机与其同步时间（下面以node1为例）

```
[root@node1 ~]#  cat /etc/chrony.conf
server 192.168.4.10   iburst
[root@node1 ~]# systemctl  restart  chronyd
```


准备存储磁盘

知识讲解

- 物理机上为每个虚拟机创建3个磁盘

```
[root@root9pc01 ~]#  cd /var/lib/libvirt/images
[root@root9pc01 ~]# qemu-img create -f qcow2 node1-vdb.vol 10G
[root@root9pc01 ~]# qemu-img create -f qcow2 node1-vdc.vol 10G
[root@root9pc01 ~]# qemu-img create -f qcow2 node1-vdd.vol 10G
[root@root9pc01 ~]# qemu-img create -f qcow2 node2-vdb.vol 10G
[root@root9pc01 ~]# qemu-img create -f qcow2 node2-vdc.vol 10G
[root@root9pc01 ~]# qemu-img create -f qcow2 node2-vdd.vol 10G
[root@root9pc01 ~]# qemu-img create -f qcow2 node3-vdb.vol 10G
[root@root9pc01 ~]# qemu-img create -f qcow2 node3-vdc.vol 10G
[root@root9pc01 ~]# qemu-img create -f qcow2 node3-vdd.vol 10G
```

- 在图形环境中为虚拟机添加磁盘

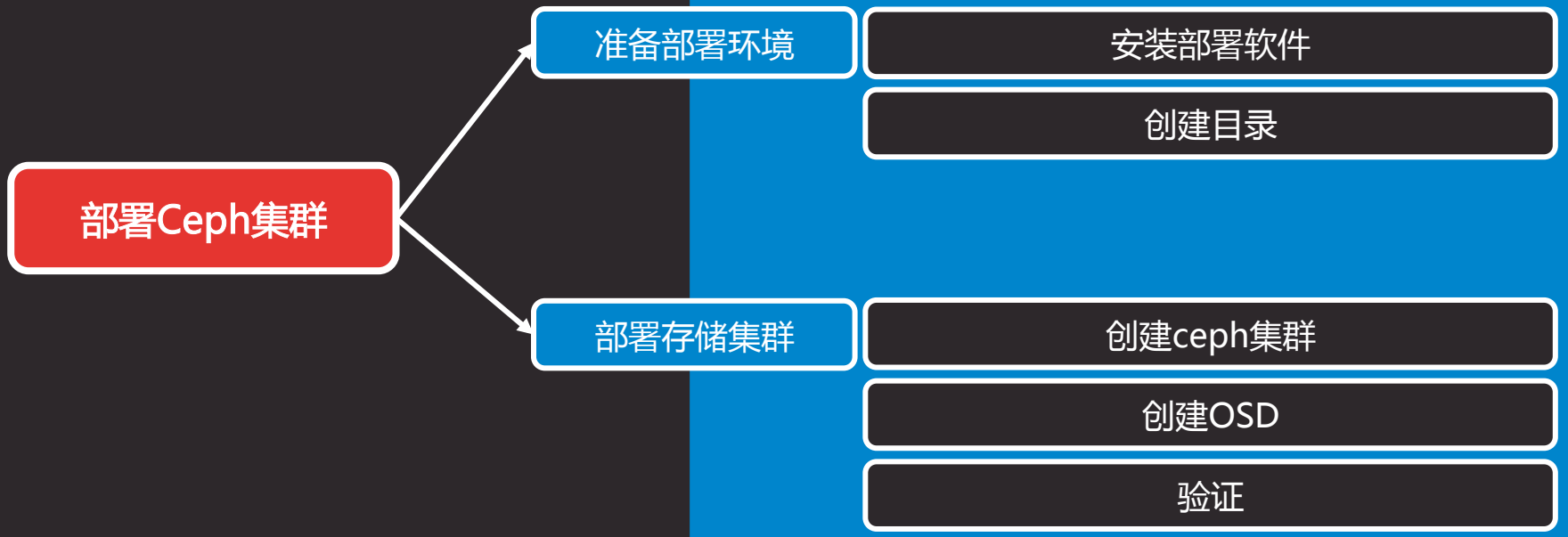
```
[root@root9pc01 ~]# virt-manager
```

案例1：实验环境

- 创建1台客户端虚拟机
- 创建3台存储集群虚拟机
- 配置主机名、IP地址、YUM源
- 修改所有主机的主机名
- 配置无密码SSH连接
- 配置NTP时间同步
- 创建虚拟机磁盘

课堂练习

部署Ceph集群



Tedu.cn
达内教育

准备部署环境

安装部署软件

- 使用node1作为部署主机

```
[root@node1 ~]#  yum -y install ceph-deploy
```

- ceph-deploy命令与子命令都支持--help查看帮助

```
[root@node1 ~]#  ceph-deploy  --help
```

知识讲解

创建目录

- 为部署工具创建目录，存放密钥与配置文件

```
[root@node1 ~]#  mkdir ceph-cluster
[root@node1 ~]#  cd ceph-cluster/
```

知识讲解

Tedu.cn
达内教育

部署存储集群

创建Ceph集群

知识讲解

- 创建ceph集群配置（所有节点都为mon）

  [root@node1 ceph-cluster]# ceph-deploy new node1 node2 node3

- 给所有节点安装ceph软件包

  [root@node1 ceph-cluster]# ceph-deploy install node1 node2 node3

- 初始化所有节点的mon服务（主机名解析必须对）

  [root@node1 ceph-cluster]# ceph-deploy mon create-initial
  //这里没有指定主机，是因为第一步创建的配置文件中已经有了，//所以要求主机名解析必须对，否则连接不到对应的主机

创建OSD

- 所有节点准备磁盘分区（下面以node1为例）

```
[root@node1 ~]# parted /dev/vdb mklabel gpt
[root@node1 ~]# parted /dev/vdb mkpart primary 1M 50%
[root@node1 ~]# parted /dev/vdb mkpart primary 50% 100%

[root@node1 ~]# chown ceph.ceph /dev/vdb1
[root@node1 ~]# chown ceph.ceph /dev/vdb2
```

//这两个分区用来做存储服务器的日志journal盘

知识讲解

创建OSD（续1）

知识讲解

- 初始化清空磁盘数据（仅node1操作即可）

```
[root@node1 ~]# ceph-deploy disk  zap  node1:vdc   node1:vdd
[root@node1 ~]# ceph-deploy disk  zap  node2:vdc   node2:vdd
[root@node1 ~]# ceph-deploy disk  zap  node3:vdc   node3:vdd
```

- 创建OSD存储空间（仅node1操作即可）

```
[root@node1 ~]# ceph-deploy osd create node1:vdc:/dev/vdb1 node1:vdd:/dev/vdb2
//创建osd存储设备，vdc为集群提供存储空间，vdb1提供JOURNAL日志，一个存储设备对应一个日志设备，日志需要SSD，不需要很大
[root@node1 ~]# ceph-deploy osd create node2:vdc:/dev/vdb1 node2:vdd:/dev/vdb2
[root@node1 ~]# ceph-deploy osd create node3:vdc:/dev/vdb1 node3:vdd:/dev/vdb2
```

验证

- 查看集群状态

```
[root@node1 ~]#  ceph  -s
```

- 可能出现的错误
  - osd create创建OSD存储空间，如提示run 'gatherkeys'

```
[root@node1 ~]#  ceph-deploy gatherkeys node1 node2 node3
```

  - ceph -s查看状态，如果失败

```
[root@node1 ~]#  systemctl restart ceph\*.service ceph\*.target
//在所有节点，或仅在失败的节点重启服务
```

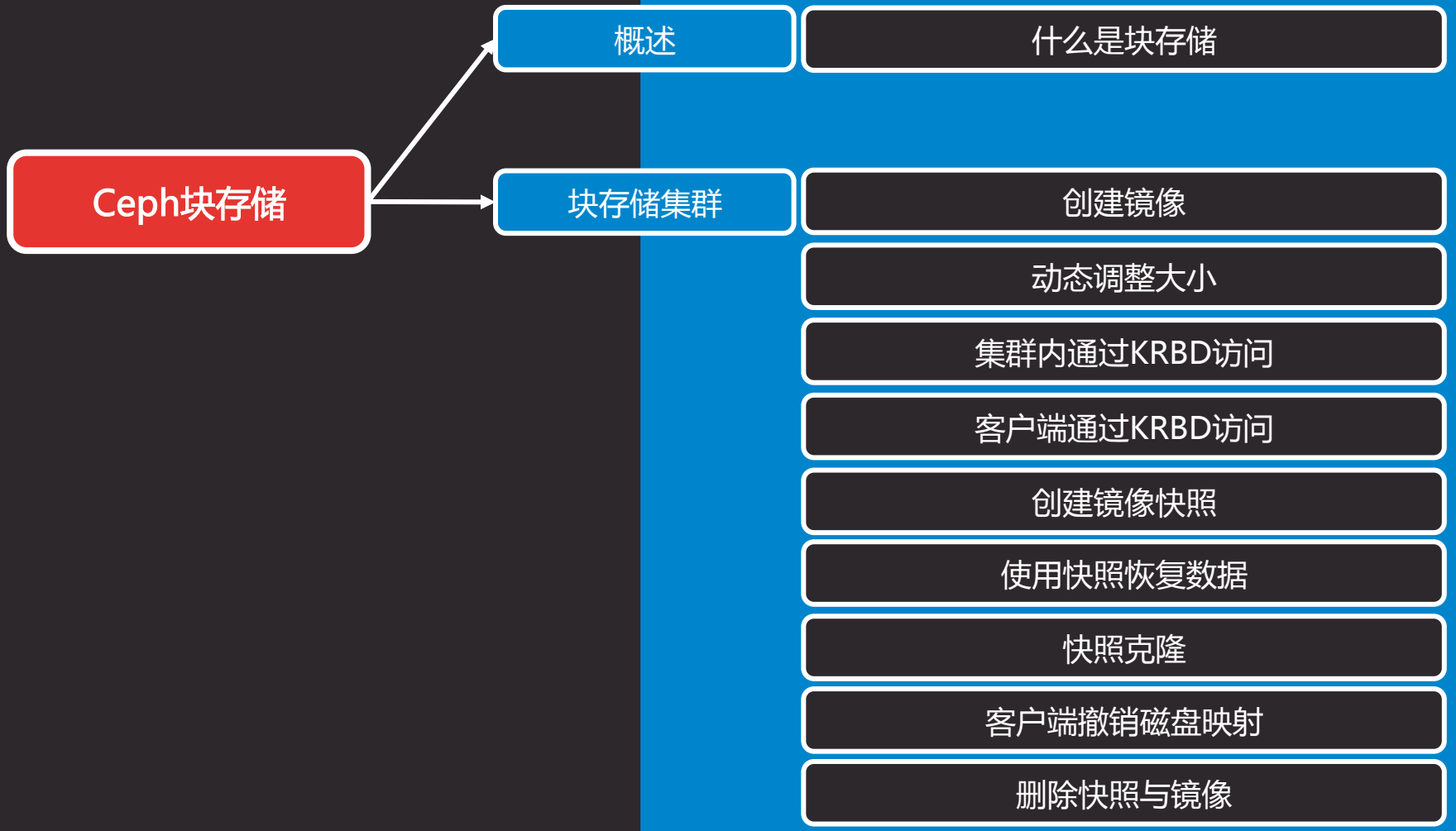
知识讲解

案例2：部署ceph集群

- 安装部署工具ceph-deploy
- 创建ceph集群
- 准备日志磁盘分区
- 创建OSD存储空间
- 查看ceph状态，验证

课堂练习

Ceph块存储


Ceph块存储
概述
什么是块存储
块存储集群
创建镜像
动态调整大小
集群内通过KRBD访问
客户端通过KRBD访问
创建镜像快照
使用快照恢复数据
快照克隆
客户端撤销磁盘映射
删除快照与镜像

概述

什么是块存储

- 单机块设备
  - 光盘
  - 磁盘
- 分布式块存储
  - Ceph
  - Cinder

知识讲解

什么是块存储（续1）

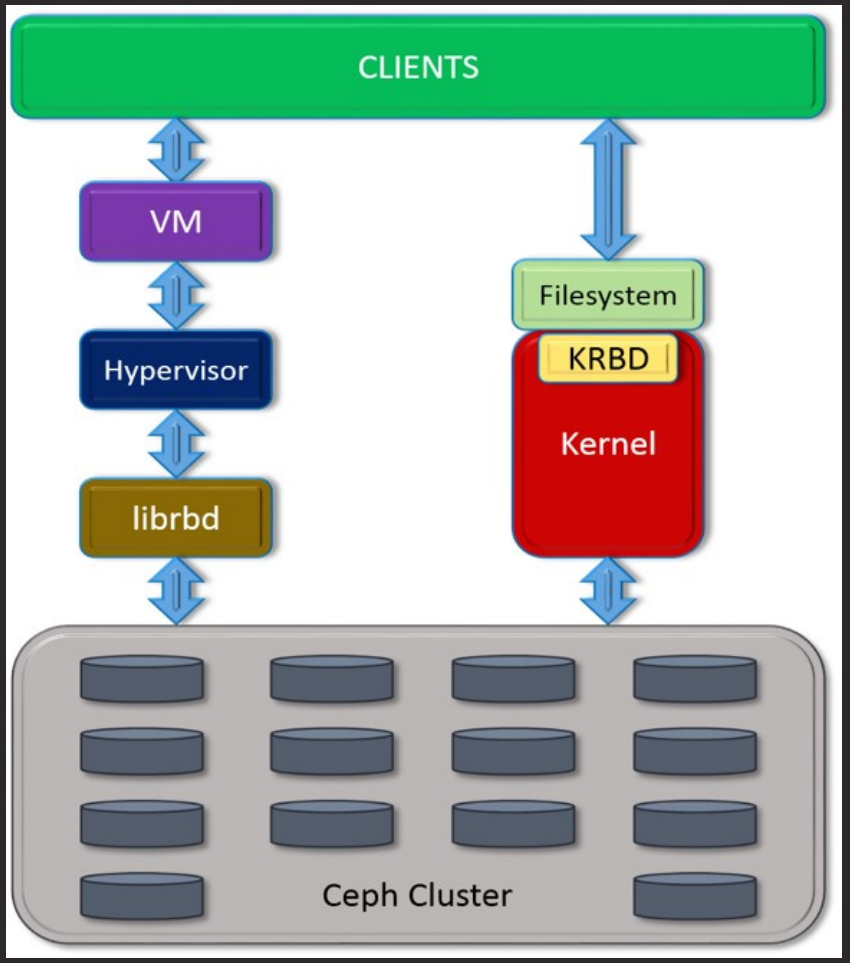
- Ceph块设备也叫做RADOS块设备
  - RADOS block device：RBD

- RBD驱动已经很好的集成在了Linux内核中
- RBD提供了企业功能，如快照、COW克隆等等
- RBD还支持内存缓存，从而能够大大提高性能

知识讲解

什么是块存储（续2）

- Linux内核可用直接访问Ceph块存储
- KVM可用借助于librbd访问



知识讲解

Tedu.cn
达内教育

块存储集群

创建镜像

知识讲解

- 查看存储池（默认有一个rbd池）

```
[root@node1 ~]# ceph osd lspools
0 rbd,
```

- 创建镜像、查看镜像

```
[root@node1 ~]# rbd create demo-image --image-feature  layering --size 10G
[root@node1 ~]# rbd create rbd/image --image-feature  layering --size 10G
[root@node1 ~]# rbd list
[root@node1 ~]# rbd info demo-image
rbd image 'demo-image':
        size 10240 MB in 2560 objects
        order 22 (4096 kB objects)
        block_name_prefix: rbd_data.d3aa2ae8944a
        format: 2
        features: layering
```

动态调整大小

- 缩小容量

```
[root@node1 ~]# rbd resize --size 7G image --allow-shrink
[root@node1 ~]# rbd info image
```

- 扩容容量

```
[root@node1 ~]# rbd resize --size 15G image
[root@node1 ~]# rbd info image
```

知识讲解

集群内通过KRBD访问

知识讲解

- 将镜像映射为本地磁盘

```
[root@node1 ~]# rbd map demo-image
/dev/rbd0
[root@node1 ~]# lsblk
… …
rbd0          251:0    0   10G  0 disk
```

- 接下来，格式化了！

```
[root@node1 ~]# mkfs.xfs /dev/rbd0
[root@node1 ~]# mount  /dev/rbd0  /mnt
```

客户端通过KRBD访问

知识讲解

- 客户端需要安装ceph-common软件包
- 拷贝配置文件（否则不知道集群在哪）
- 拷贝连接密钥（否则无连接权限）

```
[root@client ~]# yum -y  install ceph-common
[root@client ~]# scp 192.168.4.11:/etc/ceph/ceph.conf  /etc/ceph/
[root@client ~]# scp 192.168.4.11:/etc/ceph/ceph.client.admin.keyring \
/etc/ceph/
```

- 映射镜像到本地磁盘

```
[root@client ~]# rbd map image
[root@client ~]#  lsblk
[root@client ~]# rbd showmapped
id pool image snap device
0  rbd  image -    /dev/rbd0
```

客户端通过KRBD访问（续1）

- 客户端格式化、挂载分区

```
[root@client ~]# mkfs.xfs /dev/rbd0
[root@client ~]# mount /dev/rbd0 /mnt/
[root@client ~]# echo "test" > /mnt/test.txt
```

知识讲解

创建镜像快照

知识讲解

- 查看镜像快照

```
[root@node1 ~]# rbd snap ls image
```

- 创建镜像快照

```
[root@node1 ~]# rbd snap create image --snap image-snap1
[root@node1 ~]# rbd snap ls image
SNAPID NAME          SIZE
     4 image-snap1 15360 MB
```

- 注意：快照使用COW技术，对大数据快照速度会很快！

使用快照恢复数据

知识讲解

- 删除客户端写入的测试文件

```
[root@client ~]# rm  -rf   /mnt/test.txt
```

- 还原快照

```
[root@node1 ~]# rbd snap rollback image --snap image-snap1
```

- 客户端重新挂载分区

```
[root@client ~]# umount  /mnt
[root@client ~]# mount /dev/rbd0 /mnt/
[root@client ~]# ls  /mnt
```

快照克隆

- 如果想从快照恢复出来一个新的镜像，则可以使用克隆
- 注意，克隆前，需要对快照进行<保护>操作
- 被保护的快照无法删除，取消保护(unprotect)

知识讲解

```
[root@node1 ~]#  rbd snap protect image --snap image-snap1
[root@node1 ~]#  rbd snap rm image --snap image-snap1    //会失败

[root@node1 ~]# rbd clone \
image --snap image-snap1 image-clone --image-feature layering

//使用image的快照image-snap1克隆一个新的image-clone镜像
```

快照克隆（续1）

- 查看克隆镜像与父镜像快照的关系

```
[root@node1 ~]#  rbd info image-clone
rbd image 'image-clone':
        size 15360 MB in 3840 objects
        order 22 (4096 kB objects)
        block_name_prefix: rbd_data.d3f53d1b58ba
        format: 2
        features: layering
        flags:
        parent: rbd/image@image-snap1
```

知识讲解

快照克隆（续2）

- 克隆镜像很多数据都来自于快照链
- 如果希望克隆镜像可以独立工作，就需要将父快照中的数据，全部拷贝一份，但比较耗时！！！

```
[root@node1 ~]#  rbd flatten image-clone
[root@node1 ~]#  rbd info image-clone
rbd image 'image-clone':
        size 15360 MB in 3840 objects
        order 22 (4096 kB objects)
        block_name_prefix: rbd_data.d3f53d1b58ba
        format: 2
        features: layering
        flags:
//注意，父快照信息没了！
```

知识讲解

客户端撤销磁盘映射

- umount挂载点

  [root@client ~]# umount /mnt

- 取消RBD磁盘映射

  [root@client ~]# rbd showmapped
  id pool image snap device
  0 rbd image - /dev/rbd0
  //语法格式:
  [root@client ~]# rbd unmap /dev/rbd/{poolname}/{imagename}

  [root@client ~]# rbd unmap /dev/rbd/rbd/image

知识讲解

删除快照与镜像

- 删除快照（确保快照未被保护）

  [root@node1 ~]# rbd snap rm image --snap image-snap

- 删除镜像

  [root@node1 ~]# rbd list
  [root@node1 ~]# rbd rm image

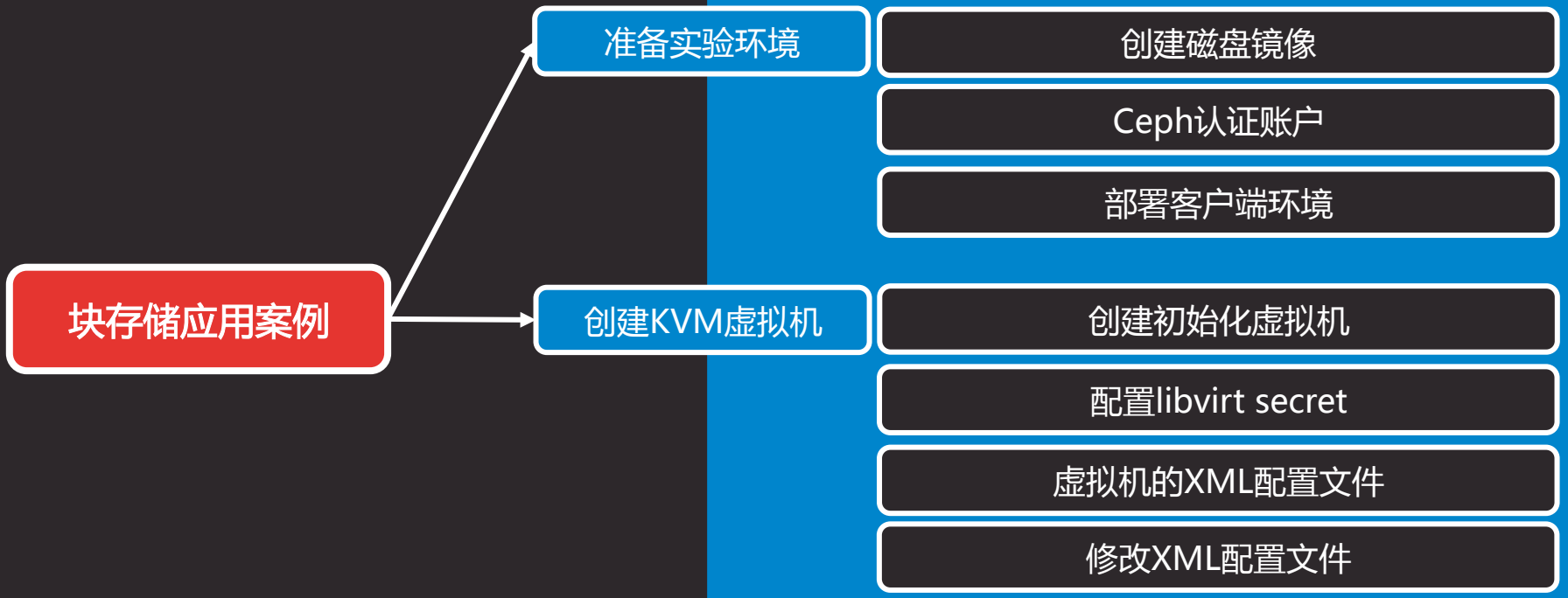
知识讲解

案例3：创建Ceph块存储

- 创建块存储镜像
- 客户端映射镜像
- 创建镜像快照
- 使用快照还原数据
- 使用快照克隆镜像
- 删除快照与镜像

课堂练习

块存储应用案例


块存储应用案例
准备实验环境
创建磁盘镜像
Ceph认证账户
部署客户端环境
创建KVM虚拟机
创建初始化虚拟机
配置libvirt secret
虚拟机的XML配置文件
修改XML配置文件

Tedu.cn
达内教育

准备实验环境

创建磁盘镜像

- 为虚拟机创建磁盘镜像

```
[root@node1 ~]# rbd create vm1-image --image-feature  layering --size 10G
[root@node1 ~]# rbd create vm2-image --image-feature  layering --size 10G
```

- 查看镜像

```
[root@node1 ~]# rbd  list
[root@node1 ~]# rbd  info  vm1-image
[root@node1 ~]# qemu-img  info  rbd:rbd/vm1-image
image: rbd:rbd/vm1-image
file format: raw
virtual size: 10G (10737418240 bytes)
disk size: unavailable
```

知识讲解

Ceph认证账户

- Ceph默认开启用户认证，客户端需要账户才可以访问
  - 默认账户名称为client.admin，key是账户的密钥
  - 可以使用ceph auth添加新账户（案例我们使用默认账户）

知识讲解

```
[root@node1 ~]# cat /etc/ceph/ceph.conf          //配置文件
[global]
mon_initial_members = node1, node2, node3
mon_host = 192.168.2.10,192.168.2.20,192.168.2.30
auth_cluster_required = cephx                    //开启认证
auth_service_required = cephx                    //开启认证
auth_client_required = cephx                     //开启认证
[root@node1 ~]# cat /etc/ceph/ceph.client.admin.keyring   //账户文件
[client.admin]
        key = AQBTsdRapUxBKRAANXtteNUyoEmQHveb75bISg==
```

部署客户端环境

知识讲解

- 注意：这里使用真实机当客户端！！！
- 客户端需要安装ceph-common软件包
- 拷贝配置文件（否则不知道集群在哪）
- 拷贝连接密钥（否则无连接权限）

```
[root@room9pc01 ~]# yum -y  install ceph-common
[root@room9pc01 ~]# scp 192.168.4.11:/etc/ceph/ceph.conf  /etc/ceph/
[root@room9pc01 ~]# scp 192.168.4.11:/etc/ceph/ceph.client.admin.keyring \
/etc/ceph/
```

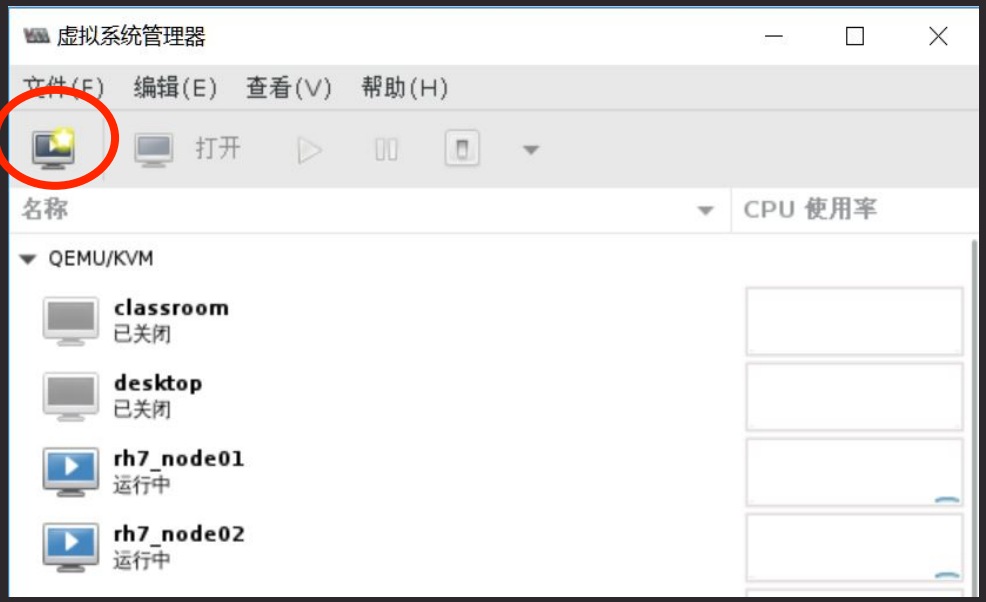
Tedu.cn
达内教育

创建KVM虚拟机

创建初始化虚拟机

- 使用virt-manager创建2台普通的KVM虚拟机
  - 这里以1个虚拟机为例

[root@room9pc01 ~]# virt-manager



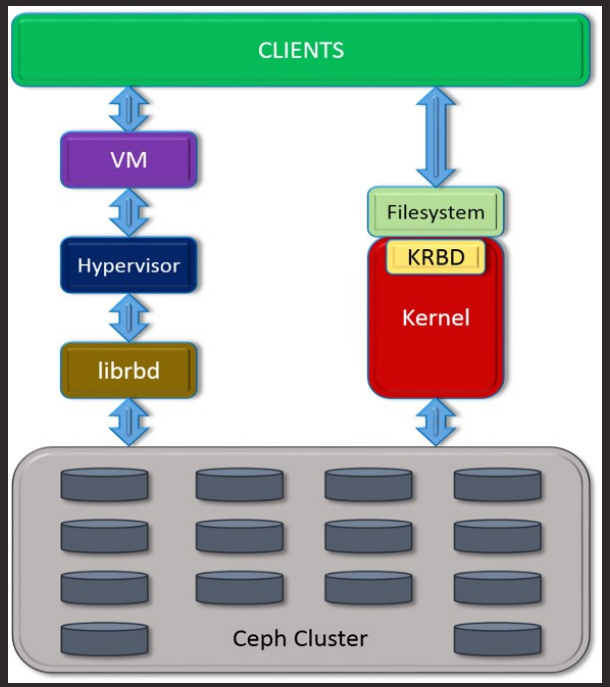
知识讲解

创建初始化虚拟机（续1）

- 创建虚拟机后，不着急启动虚拟机（关闭虚拟机）

配置libvirt secret

- KVM虚拟机需要使用librbd才可以访问ceph集群
- Librbd访问ceph又需要账户认证
- 所以这里，我们需要给libvirt设置账户信息



知识讲解

配置libvirt secret（续1）

- 编写账户信息文件（真实机操作）

```
[root@room9pc01 ~]# vim secret.xml          //新建临时文件，内容如下
<secret ephemeral='no' private='no'>
      <usage type='ceph'>
            <name>client.admin secret</name>
      </usage>
</secret>
```

- 使用XML配置文件创建secret

```
[root@room9pc01 ~]# virsh secret-define --file secret.xml
733f0fd1-e3d6-4c25-a69f-6681fc19802b
//随机的UUID，这个UUID对应的有账户信息
```

知识讲解

配置libvirt secret（续2）

知识讲解

- 编写账户信息文件（真实机操作）

```
[root@room9pc01 ~]# ceph auth get-key client.admin
//获取client.admin的key，或者直接查看密钥文件
[root@room9pc01 ~]# cat /etc/ceph/ceph.client.admin.keyring
```

- 设置secret，添加账户的密钥

```
[root@room9pc01 virsh secret-set-value \
--secret 733f0fd1-e3d6-4c25-a69f-6681fc19802b \
--base64 AQBTsdRapUxBKRAANXtteNUyoEmQHveb75bISg
//这里secret后面是之前创建的secret的UUID
//base64后面是client.admin账户的密码
//现在secret中既有账户信息又有密钥信息
```

虚拟机的XML配置文件

- 每个虚拟机都会有一个XML配置文件，包括：
  - 虚拟机的名称、内存、CPU、磁盘、网卡等信息

知识讲解

```
[root@room9pc01 ~]# vim /etc/libvirt/qemu/vm1.xml
//修改前内容如下
<disk type='file' device='disk'>
    <driver name='qemu' type='qcow2'/>
    <source file='/var/lib/libvirt/images/vm1.qcow2'/>
    <target dev='vda' bus='virtio'/>
    <address type='pci' domain='0x0000' bus='0x00'
slot='0x07' function='0x0'/>
  </disk>
```

修改XML配置文件

- 不推荐直接使用vim修改配置文件
- 推荐使用virsh edit修改配置文件

```
[root@room9pc01 virsh edit vm1          //vm1为虚拟机名称
<disk type='file' device='disk'>
    <driver name='network' type='raw'/>
    <auth username='libvirt'>
    <secret type='ceph' uuid='3b8b0c5c-bebc-4fc2-9137-0e3deb61dc8b'/>
    </auth>
    <source protocol='rbd' name='rbd/vm1'>
      <host name='192.168.4.11' port='6789'/>
   </source>
   <target dev='vda' bus='virtio'/>
    <address type='pci' domain='0x0000' bus='0x00' slot='0x07' function='0x0'/>
   </disk>
```

知识讲解

修改XML配置文件（续1）

- 关键词说明

```
<secret type='ceph' uuid='3b8b0c5c-bebc-4fc2-9137-0e3deb61dc8b’/>
//这里的uuid就是secret的uuid，有client.admin账户和密钥信息
    <source protocol='rbd' name='rbd/vm1'>
        <host name='192.168.4.11' port='6789’/>
    </source>
//这里说明使用账户连接哪台ceph主机和端口，访问哪个池和镜像
    <target dev='vda' bus='virtio’/>
//这里说明，将获取的镜像，设置为虚拟机的vda磁盘
```

知识讲解

案例4：块存储应用案例

- Ceph创建块存储镜像
- 客户端安装部署ceph软件
- 客户端部署虚拟机
- 客户端创建secret
- 设置虚拟机配置文件，调用ceph存储

课堂练习

Tedu.cn
达内教育

总结和答疑