誉天教育
YU TIAN EDU

Linux云计算架构师涨薪班

Ceph安装与扩容

学习目标

- 生产环境组件规划
- 生产环境硬件配置
- ceph部署方式及版本
- cephadm简介及原理
- 红帽RHCS5.0部署
- 开源ceph集群部署
- ceph功能组件开放端口
- 使用容器管理ceph集群
- cephadm扩展集群节点
- cephadm扩容集群容量

生产环境组件的最少规划

- 至少三个MON节点
- 至少三个OSD节点（每节点可以有多个OSD进程）
- 至少两个MGR节点
- 如果使用CephFS，至少需要两个配置完全相同的MDS节点
- 如果使用Ceph RADOSGW,则至少需要两个RGW节点

容器硬件配置推荐

容器名	CPU	RAM	磁盘空间	网络
ceph-osd-container	1	5GB	整个硬盘	2x 10 GB 以太网 NIC
ceph-mon-container	1	3GB	10GB	2x 1GB 以太网 NIC
ceph-mgr-container	1	3GB	\	2x 1GB 以太网 NIC
ceph-radosgw-container	1	1GB	5GB	1x 1GB 以太网 NIC
ceph-mds-container	1	3GB	2GB	2x 1GB 以太网 NIC
红帽建议在生产环境中使用 10 Gb 以太网部署 Red Hat Ceph Storage。1 GB 以太网网络不适用于生产环境的存储集群。				

OSD要求

- - 设备不得有分区。
- - 设备不得处于 LVM 状态
- - 设备不得挂载
- - 设备不得包含文件系统
- - 设备不得包含 Ceph BlueStore OSD
    - ceph的存储驱动由原来的filestore调整为bluestore
- - 设备必须大于 5 GB

选择硬件时可能出现的常见错误

- 仍然使用旧的、性能较差的硬件用于 Ceph。
- 在同一个池中使用不同的硬件。
- 使用 1Gbps 网络而不是 10Gbps 或更快的网络。
- 没有正确设置公共和集群网络。
- 使用 RAID 作为数据保护
- 在选中驱动器时只考虑了价格而没有考虑性能或吞吐量。
- 当用例需要 SSD 日志时，在 OSD 数据驱动器上进行日志。
- 磁盘控制器的吞吐量不足。

Ceph的部署方式

- 纯手动部署
  - 部署极其复杂
  - 便于初学者理解具体的工作机制
- ceph-deploy
  - ceph早期部署方式
- ceph-ansible
  - RHCS3的部署方式
- cephadm
  - ceph容器化部署方式
  - ceph官方和红帽推荐

ceph的版本

版本	主版本号	初始发行时间	停止维护时间
Quincy	17	2022-04-19	2024-06-01
Pacific	16	2021-03-31	2023-06-01
Octopus	15	2020-03-23	2022-06-01
Nautilus	14	2019-03-19	2021-06-30
Mimic	13	2018-06-01	2020-07-22
Luminous	12	2017-08-01	2020-03-01
Kraken	11	2017-01-01	2017-08-01
Jewel	10	2016-04-01	2018-07-01

ceph从Nautilus版本（14.2.0）开始，每年都会有一个新的稳定版发行，每年的新版本都会起一个新的名称（例如，“Mimic”）和一个主版本号（例如，13 代表 Mimic，因为“M”是字母表的第 13 个字母）

x.y.z

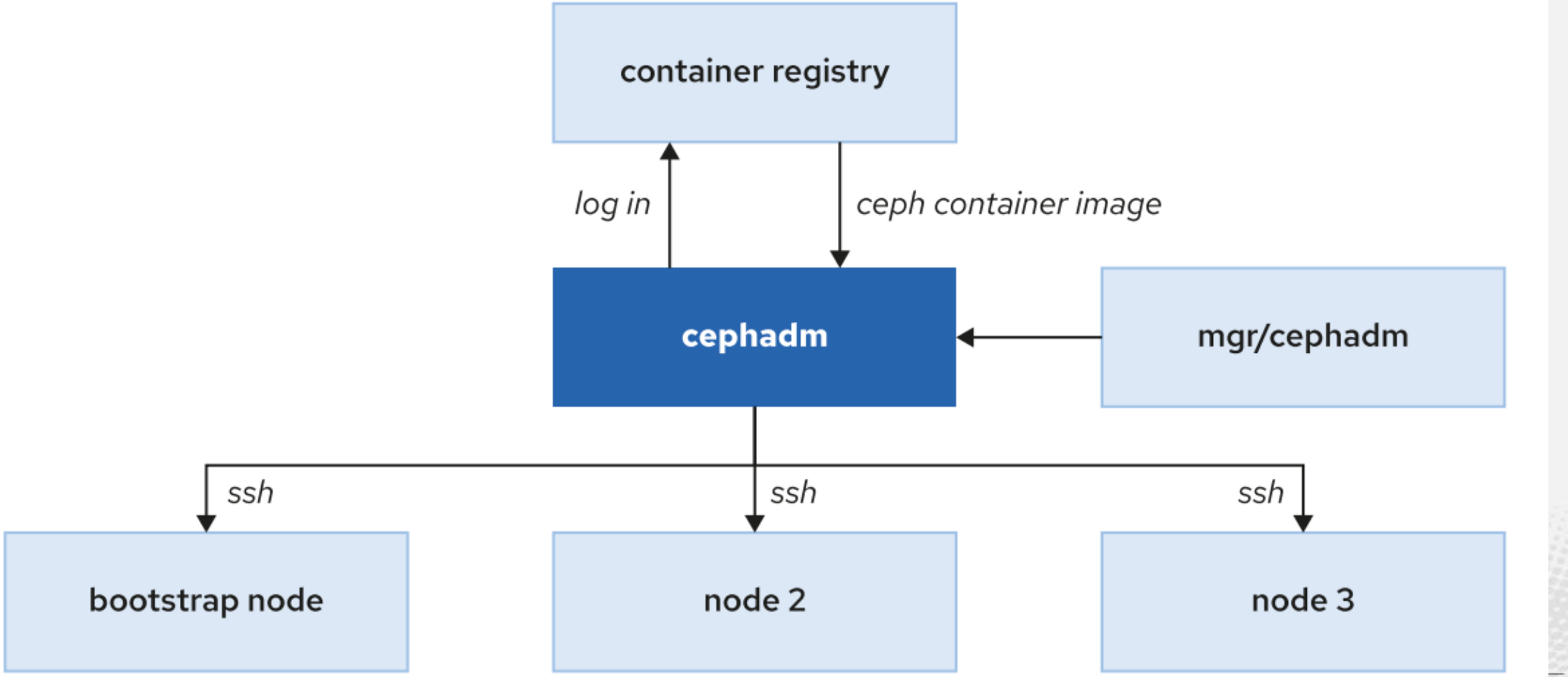
x: 发布周期
y:发布类型0 开发 1 测试 2 稳定
z:次版本号

rhcs5
cephadm
rhcs4
rhcs3
rhcs2

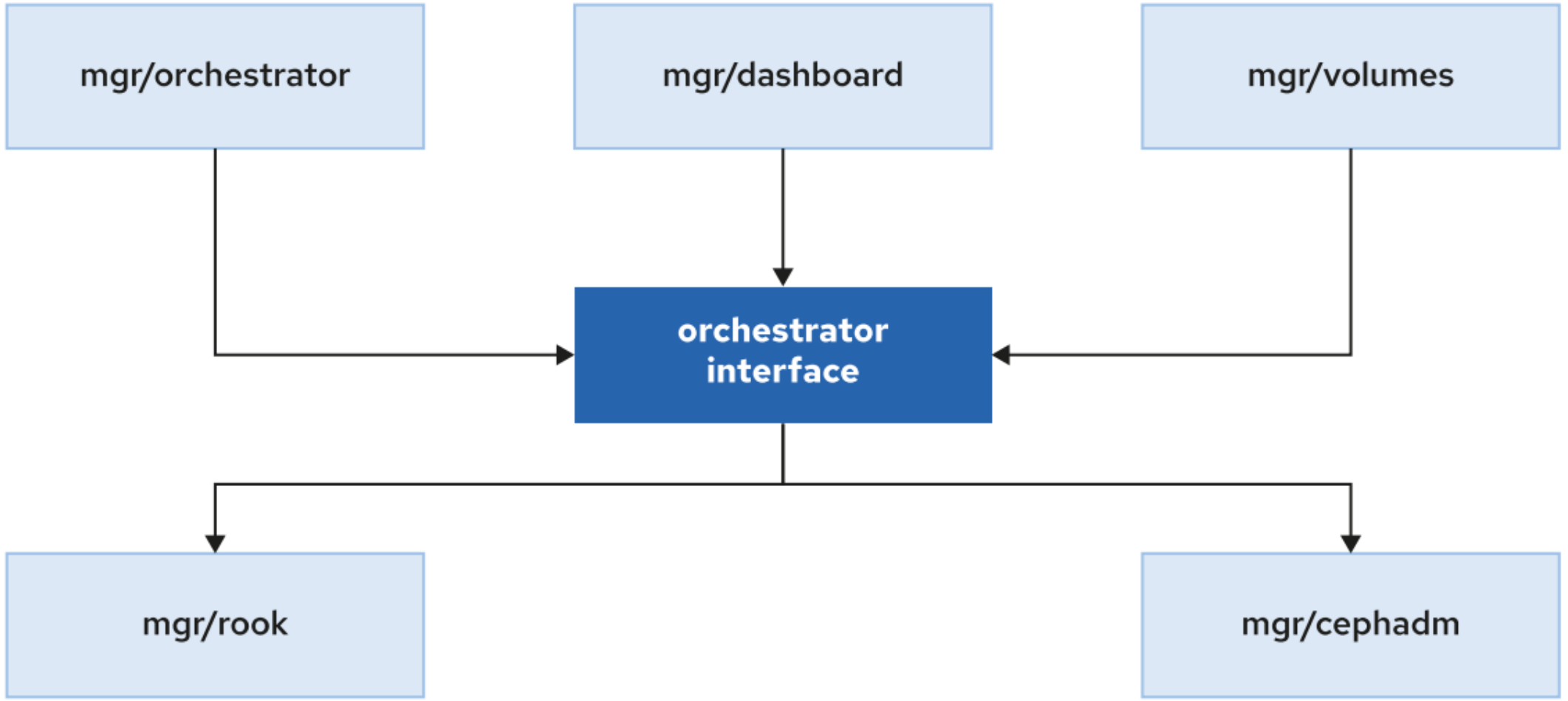
cephadm介绍

- cephadm由 cephadm （主程序）和 cephadm orchestrator （编排器）
- Cephadm 的必要部署条件:
  - cephadm（脚本） podman/docker python3 chrony或NTP
- cephadm 命令在 Ceph 提供的管理容器内运行 bash shell。使用 cephadm shell 执行最初的集群部 署任务以及集群安装和运行后的集群管理任务
- cephadm orchestrator为编排器 ceph-mgr 模块提供了一个命令行界面，与外部编排服务交互。编排器旨在用于协调必须在存储集群中的多个节点和服务之间协作执行的配置更改。

cephadm组件交互



cephadm编排器



RHCS5.0部署

- 摧毁已经存在的集群
  - [student@workstation ~]$ lab start deploy-deploy
- 将serverc作为引导节点
  - yum install cephadm-ansible
- 执行节点预配cephadm-preflight.yml
  - ✓ 编写ansible主机清单 /usr/share/cephadm-ansible/hosts
  - ✓ ansible-playbook -i hosts \ cephadm-preflight.yml --extra-vars “ceph_origin=“ 执行预配
- 使用cephadm进行部署
  - [root@serverc ceph]# cephadm bootstrap --mon-ip=172.25.250.12 \ --apply-spec=initial-config-primary-cluster.yaml \ --initial-dashboard-password=redhat \ --dashboard-password-noupdate \ --allow-fqdn-hostname \ --registry-url=registry.lab.example.com \ --registry-username=registry \ --registry-password=redhat

开源pacific版ceph部署

- 执行节点预配
  - 配置主机名和IP的映射关系
  - 关闭selinux和firewalld防火墙
  - 配置时间同步且所有节点chronyd服务开机自启
- 获取指定版本的cephadm
  - wget https://github.com/ceph/ceph/raw/pacific/src/cephadm/cephadm获取 cephadm
  - chmod +x cephadm 添加执行权限
  - ./cephadm add-repo --release pacific 添加指定版本的ceph源
  - ./cephadm install 安装cephadm工具
  - ./cephadm install ceph-common 安装ceph的客户端工具(可选)

开源pacific版ceph部署

- 在引导节点上部署单节点集群
  - cephadm bootstrap --mon-ip 172.17.0.81 --allow-fqdn-hostname --initial-dashboard-user admin --initial-dashboard-password redhat --dashboard-password-noupdate
    - bootstrap 引导集群
    - --mon-ip 172.17.0.81 指定mon节点地址
    - --allow-fqdn-hostname 使用主机名作为dashboard地址
    - --initial-dashboard-user admin 指定dashboard用户名为admin
    - --initial-dashboard-password 指定dashboard密码为redhat
    - --dashboard-password-noupdate 首次登陆dashboard无需更改密码

各节点容器的作用

- 各容器的功能作用
  - ceph-mon mon组件
  - ceph-mgr mgr组件
  - ceph-mds mds组件
  - ceph-osd osd组件
  - ceph-crash mgr用来收集守护进程的崩溃信息
  - prometheus 普罗米修斯监控的主程序
  - node-exporter 普罗米修斯监控程序 Linux收集端，windows使用WMI-exporter
  - alertmanager 普罗米修斯 报警管理器

ceph开放端口

服务名称	端口	描述
监控器 (MON)	6789/TCP (msgr)，3300/TCP (msgr2)	Ceph 集群内的通信
OSD	6800-7300/TCP	每个 OSD 使用这个范围中的三个端口：一个用于通过公共网络与客户端和 MON 通信，一个用于通过集群网络或公共网络（如果前者不存在）发送数据到其他 OSD，另外一个用于通过集群网络或公共网络（如果前者不存在）交换心跳数据包。
元数据服务器 (MDS)	6800-7300/TCP	与 Ceph 元数据服务器通信
控制面板/管理器 (MGR)	8443/TCP	通过 SSL 与 Ceph 管理器控制面板通信
管理器 RESTful 模块	8003/TCP	通过 SSL 与 Ceph 管理器 RESTful 模块通信
管理器 Prometheus 模块	9283/TCP	与 Ceph 管理器 Prometheus 插件通信

ceph开放端口

Prometheus Alertmanager	9093/TCP	与 Prometheus Alertmanager 服务通信
Prometheus 节点导出器	9100/TCP	与 Prometheus 节点导出器守护进程通信
Grafana 服务器	3000/TCP	与 Grafana 服务通信
Ceph 对象网关 (RGW)	80/TCP	与 Ceph RADOSGW 通信。如果 `client.rgw` 配置部分为空， `cephadm` 会使用默认端口 `80`。
Ceph iSCSI 网关	9287/TCP	与 Ceph iSCSI 网关通信

开源ceph部署

- 给单节点集群添加OSD磁盘
  - cephadm shell 进入到ceph操作环境
  - ceph orch daemon add osd node1.example.com:/dev/sdb 添加指定主机的单块磁盘为osd
  - ceph orch daemon add osd node1.example.com:/dev/sdc
  - ceph orch daemon add osd node1.example.com:/dev/sdd
  - ceph orch apply osd --all-available-devices 在所有主机上将未使用的磁盘部署为osd
- 给集群添加主机
  - ceph cephadm get-pub-key > ~/ceph.pub 获取集群公钥并保存到当前路径下ceph.pub
  - ssh-copy-id -f -i ~/ceph.pub root@node2 拷贝公钥到node2
  - ceph orch host add node2.example.com 需要写完整的主机名

添加ceph节点角色

- 关闭集群组件自扩展
  - ceph orch apply mon --unmanaged=true　关闭mon自动扩展
- 添加mgr节点
    - ceph orch daemon add mgr --placement=node2.example.com
- 添加mon节点
    - ceph orch daemon add mon --placement=node2.example.com

移除服务和节点

- 移除节点上的指定服务
  - ceph orch ps 查询各个节点上所运行的服务
  - ceph orch daemon rm mgr.node4.ifomqg 移除node4节点上的mgr服务
- 移除OSD
  - ceph osd metadata osd.6
  - ceph orch daemon stop osd.6
  - ceph orch daemon rm osd.6 --force
  - ceph osd rm 6 osd序号
  - ceph orch device zap ceph03.example.com /dev/sdd --force 擦除磁盘数据
  - ceph osd crush rm osd.6 删除cursh的osd映射
- 移除主机
  - ceph orch host rm node4.example.com 移除主机

使用标签部署服务

- 为主机添加标签
  - ceph orch host label add node3.example.com mgr
  - ceph orch host label add node4.example.com mgr
- 根据标签批量部署
  - ceph orch apply mgr --placement="label:mgr"

誉天教育
YU TIAN EDU

Thank you