托管物理机

01

托管物理机介绍

什么是托管物理机

托管物理机是一种注册物理机的方式，旨在将已有操作系统的物理机添加到OneCloud平台

添加方式

- (●) 托管
  托管：用于托管已有操作系统的服务器，托管后将同时生成物理机和裸金属记录
- ( ) 预注册
  预注册：用于预上架未配置 BMC 信息的服务器，通过预注册功能配置服务器基本信息，待服务器上电后， MAC 信息匹配即可进行注册并配置 BMC 信息等
- ( ) ISO引导注册
  ISO引导注册：用于立即注册已配置 BMC 信息的服务器，该功能不要求服务器处于 DHCP relay 网络环境，但是要求服务器支持 Redfish 功能
- ( ) PXE引导注册
  PXE引导注册：用于立即注册已配置 BMC 信息的服务器，要求服务器处于 DHCP relay 网络环境

提示：在一台或多台已安装系统的物理机中运行以下命令

```
sudo sh -c "$(curl -k -fsSL -H 'X-Auth-Token:
gAAAAABd8clwE7loj6-
6J63uuq24evYlTecs9BH3_w_xmJlLg4LnXDO5qIDS-fEq2-
lAhHOJ65szlIRXqTkB3b0KPxztMSAldame4tkfhuQxKtY_Urzio6GH0KLcN
TK-
7204XjegEnxYLtpd5oR3sy_D_yIeuq2tT6Pt8N2m5h0lUuQH_zCYrQHb0Hf
zut7BPagceJRFE7c0RzIQjmlJHLCxOZGR4T8nhcDKVmlY3R_a40V24Gdnnr
KMjunVv8euMRkgCN4p0_fh' https://10.127.10.2:8889/misc/bm-
prepare-script)"
```

点击复制

托管物理机的目标

- 在OneCloud平台创建对应的物理机和裸金属的记录，
  并且能够正确的采集物理机的信息机(CPU，内存，网卡，主板，磁盘等)

- 托管物理机的过程中尽量对原物理机的环境不产生影响

- 能够通过OneCloud平台对托管进来的物理机进行操作(开机，关机，重启，删除等)

- 能够通过OneCloud平台对托管进来的物理机进行远程连接(SOL, SSH，Java控制台)

02

托管物理机原理

在PXE引导注册的流程中，我们会先将物理机引导到ramfs中，里面运行着事先制作好的镜像(YunionOS)，然后BaremetalAgent再通过ssh的方式登录到ramfs中来采集物理机信息，上报给Region服务。

托管物理机面临的问题：

- 物理机操作系统本身无感知，不会影响到已经运行的程序
- 能够支持尽可能多的操作系统版本
- 避免对物理机操作系统环境的依赖
- 尽量复用PXE引导注册时采集物理机信息的逻辑

基于上面的这些考虑，我们想到的一个解决方案是在物理机上用容器来运行一个YunionOS

Docker:

- 需要在宿主机上安装Docker
- 我们并不需要Docker的全部功能

runC:

- 使用runC的好处是可以摆脱docker daemon
- 不用再物理机上安装docker运行环境，只需要用runC二进制来控制容器的启停

使用runC存在的问题：

- 某些系统调用低版本的内核不支持(比如prctl的某些capability)
- 依赖cgroup子系统(低版本的操作系统不会默认安装)

于是我们基于runC改造了一个只提供mountPoint namespace 隔离的容器运行时runNS，兼容runC运行时所需的config.json文件格式和rootfs，能够在一些老版本的系统中运行。

runNS

runNS:
- bind mount
- chroot/pivot_root
- fork, exec

https://github.com/yunionio/runns

托管物理机原理介绍

首先需要在要托管的物理机上执行下面的一段脚本，启动baremetal prepare进程：

```
sudo sh -c "$(curl -k -fsSL -H 'X-Auth-Token: gAAAAABd8clwE7loj6-
6J63uuq24evY1Tecs9BH3_w_xmJlLg4LnXDO5qIDS-fEq2-
lAhHOJ65szlIRXqTkB3b0KPxztMSA1dame4tkfhuQxKtY_Urzio6GH0KLcNTK-
7204XjegEnxYLtpd5oR3sy_D_yIeuq2tT6Pt8N2m5h01UuQH_zCYrQHb0Hfzut7BPagceJRFE7c0RzIQjml
JHLCxOZGR4T8nhcDKVmlY3R_a40V24GdnnrKMjunVv8euMRkgCN4p0_fh'
https://10.127.10.2:8889/misc/bm-prepare-script)"
```

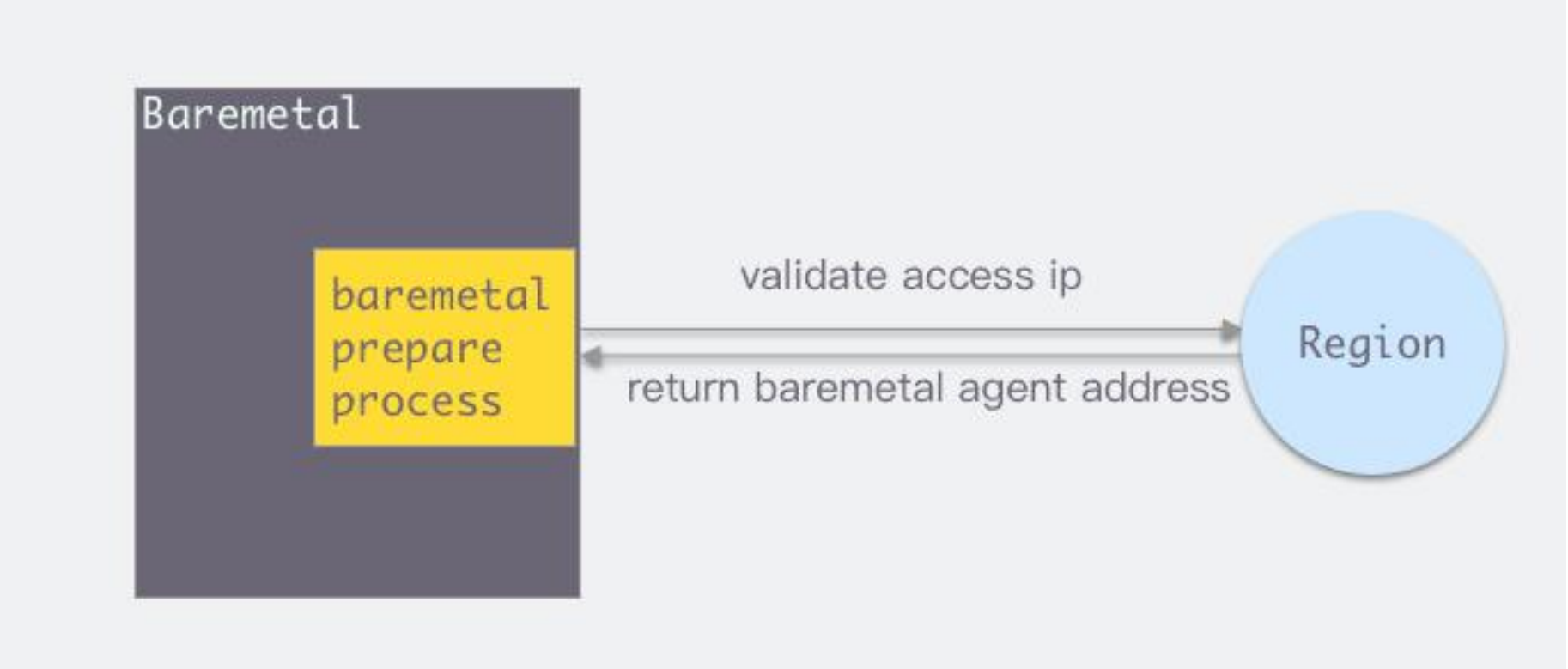
执行脚本后跟着提示输入IPMI用户名和密码，然后等待注册完成

```
[cloudroot@a13 ~]$ sudo sh -c "$(curl -k -fsSL -H 'X-Auth-Token: gAAAAABd8clwE7loj6-6J63uuq24evY1Tecs9BH
3_w_xmJlLg4LnXDO5qIDS-fEq2-lAhHOJ65szlIRXqTkB3b0KPxztMSA1dame4tkfhuQxKtY_Urzio6GH0KLcNTK-7204XjegEnxYLtp
d5oR3sy_D_yIeuq2tT6Pt8N2m5h01UuQH_zCYrQHb0Hfzut7BPagceJRFE7c0RzIQjmlJHLCxOZGR4T8nhcDKVmlY3R_a40V24GdnnrK
MjunVv8euMRkgCN4p0_fh' https://10.127.10.2:8889/misc/bm-prepare-script)"
INFO: ********** Register baremetal start ... **********
INFO: ****** Enter the IPMI username password ******
Enter the IPMI username: root
Enter the IPMI password:
Enter the IPMI password again:
```

待要托管的物理机注册成功后，将会返回物理机在云平台的id

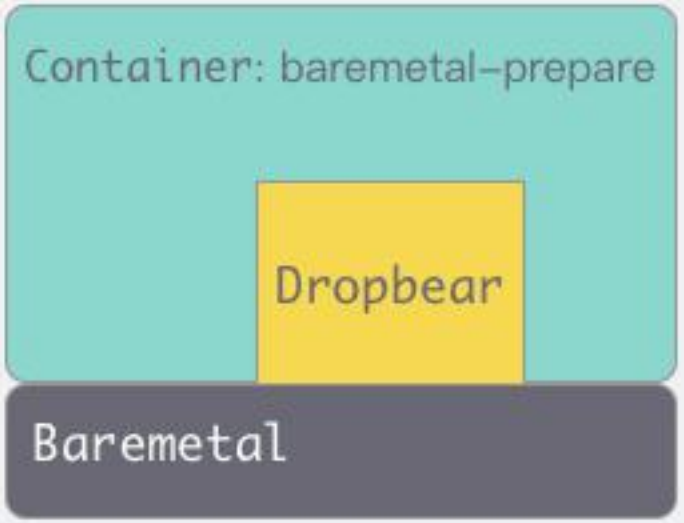
```
INFO: ********** Register baremetal start ... **********
INFO: ****** Enter the IPMI username password ******
Enter the IPMI username: root
Enter the IPMI password:
Enter the IPMI password again:
INFO: baremetal agent: https://10.127.10.2:8879
passwd: no record of root in /etc/shadow, using /etc/passwd
Changing password for root
New password:
Retype password:
passwd: password for root changed by root
Generating key, this may take a while...
Failed moving key file to /etc/dropbear/dropbear_rsa_host_key: File exists
Exited: Failed to generate key.

INFO: Prepare SUCCESS waiting register, It takes a few minutes...
INFO: baremetal instance id: 1924d08f-79e6-4bc9-8162-b20a87810517
```

首先baremetal prepare进程会访问Region服务，获取BaremetalAgent服务的地址，并且同时校验物理机的ip地址是否在OneCloud平台管理的ip子网中



- 要求物理机和Region服务之间的网络能通
- 物理机的ip在OneCloud管理的子网中

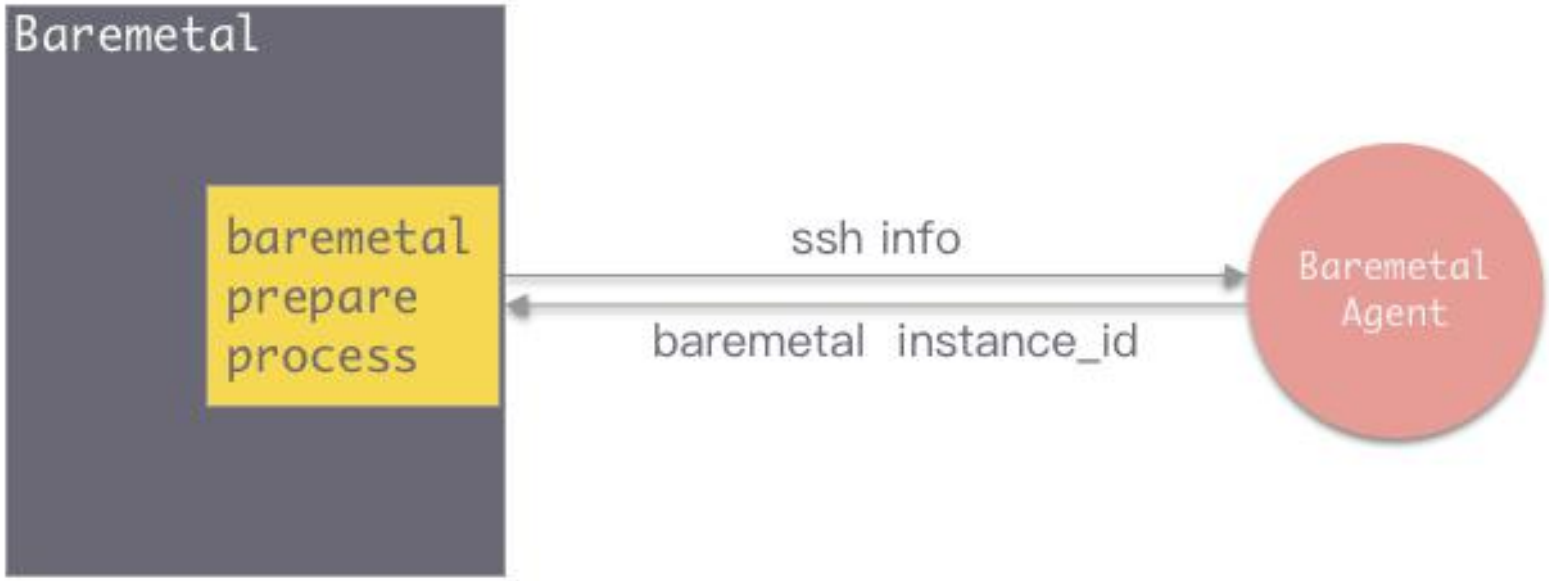
请求Region完成后接下baremetal prepare进程将在物理机上启动baremetal-prepare容器



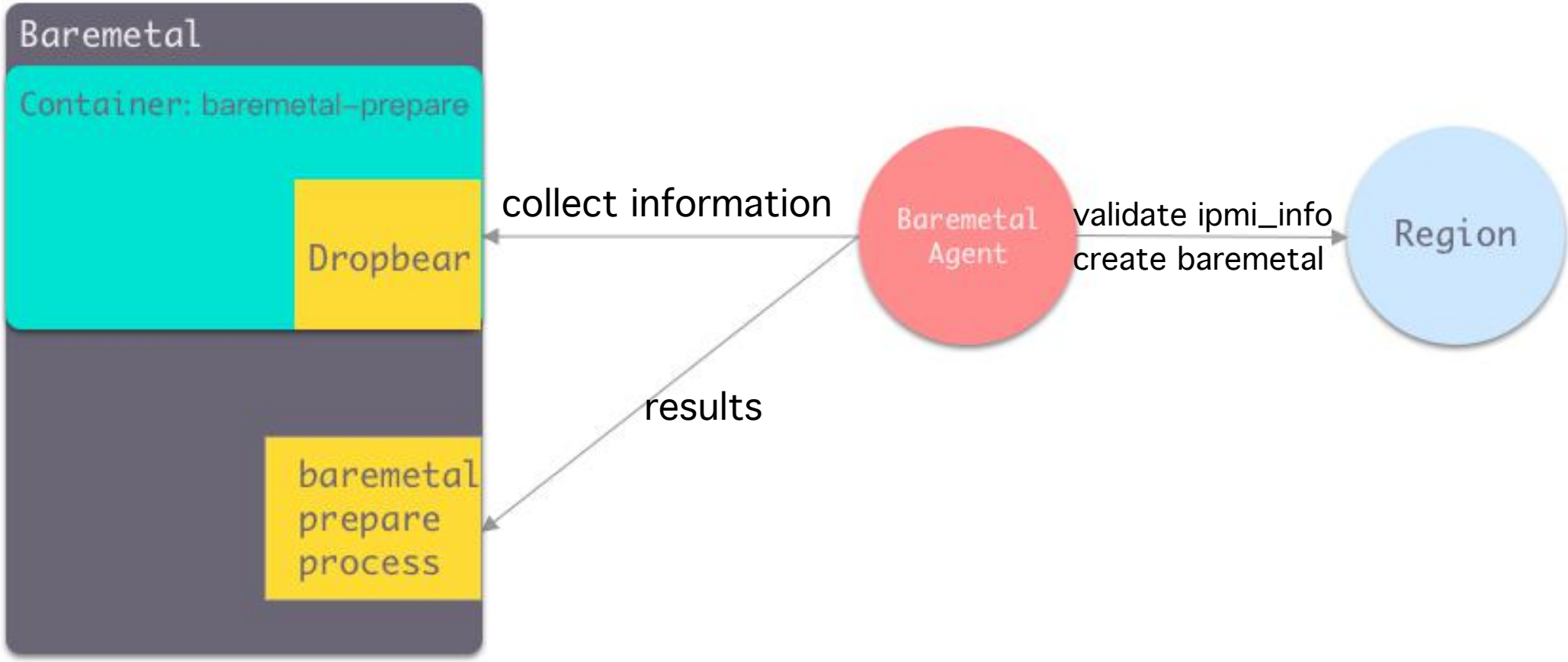
这个容器的镜像是前面提到的YunionOS,里面运行着一个Dropbear进程(一个轻量级的SSH),允许BaremetalAgent通过ssh访问到容器内部来采集信息。这个容器挂载了宿主机的/dev, /sys, /proc等这些目录。

- dropbear监听端口是从2222开始找第一个可用的端口，需要保证BaremetalAgent到这个端口能通

等Dropbear进程启动成功并监听到某个端口之后，请求BaremetalAgent开始注册流程，期望返回Baremetal Instance的ID



BaremetalAgent先通过ssh先获取到物理机的IPMI的ip地址，然后校验IPMI的ip是否在云平台的ip子网中



校验成功后创建baremetal的实例，然后就开始做剩余的信息采集工作(CPU，内存，网卡，主板，磁盘等)。

待BaremetalAgent采集信息完毕，将物理机的状态置为running后就可以进行操作。

- 这里的物理机显示已分配的裸金属记录是我们伪造的，用来防止物理机被再次调度

03

总结

总结

实现原理：

- 原理上来说我们是在物理机内创造一个隔离环境(YunionOS)，能够通过ssh登录到系统采集信息
- 设计实现上遵守的原则是尽量减少对物理机的影响和依赖

注意事项：

- 需要在OneCloud平台先准备好物理机相关的子网
- 控制节点和物理机直接的网络能够互通（http/https/ssh)
- 托管进来的物理机会创建一条伪造的裸金属服务器的记录，用来防止被再次调度

云联万维
YUNION

Thanks
Q&A

云联万维
YUNION

请填写此调查问卷，协助我们把活动办得更好、更符合您的需求，提交后请在签到台领取纪念品一份！

携手同行 共赢未来

云联万维 YUNION | 智能多云领导者