# 裸金属云产品架构设计

## 问题描述

产品支持裸金属部署

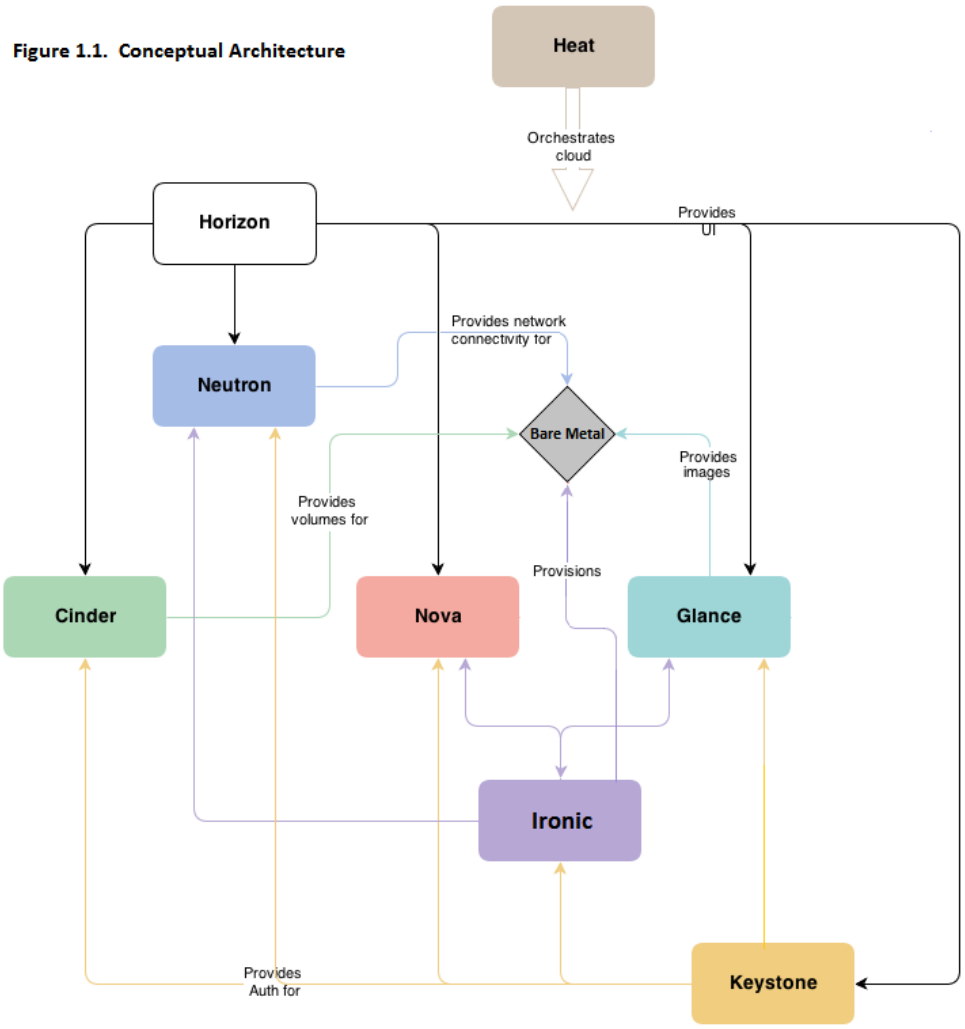
产品需求PRD: 裸机管理

## 方案提议

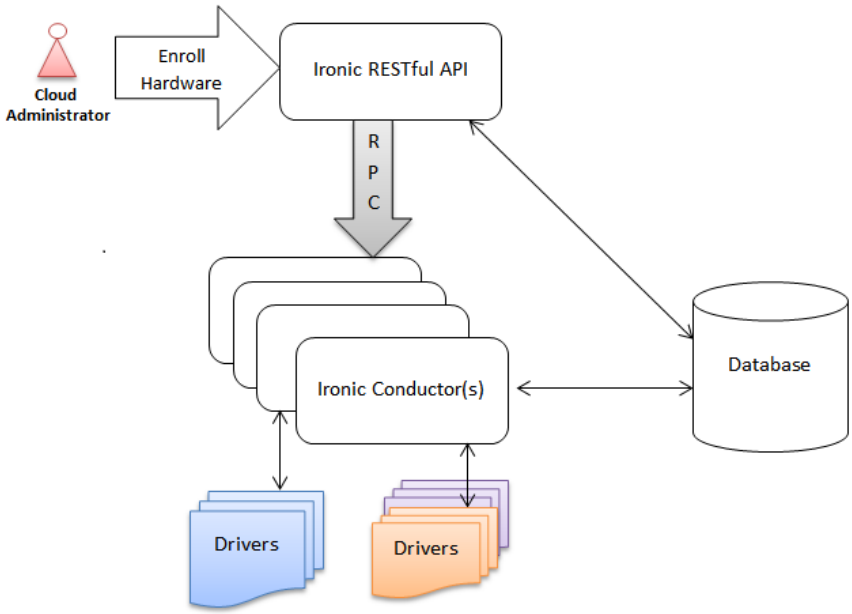
1 总体架构

采用OpenStack Ironic组件实现裸金属功能

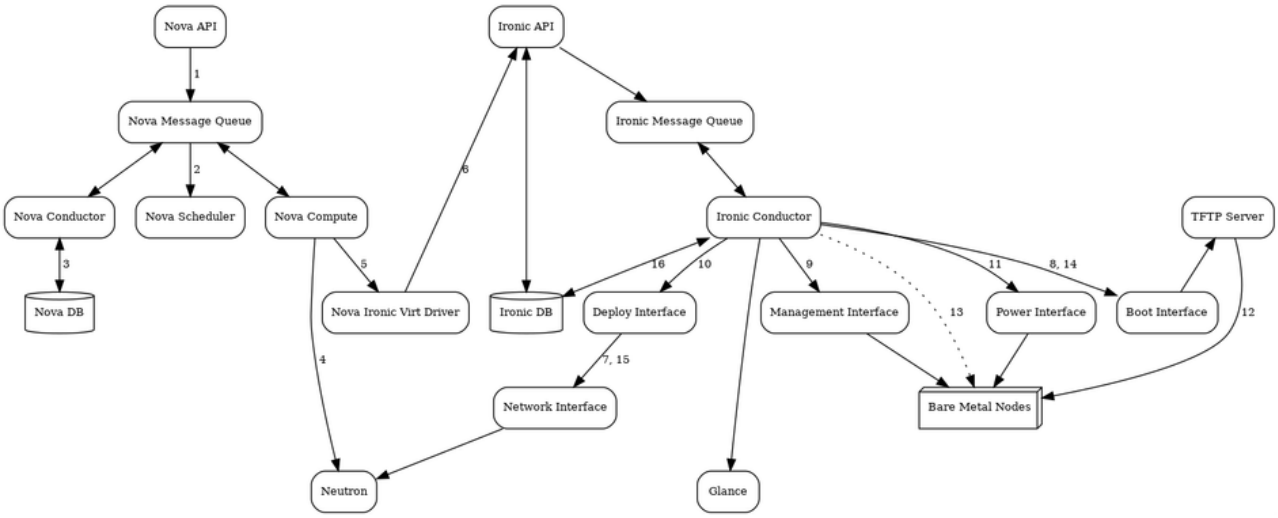
Ironic服务于其他组件关系图如下:




Ironic组件架构如下:




Ironic在部署裸机时简要任务流程如下:




部署裸机的流程如下：

1. nova api接收http rest request，处理请求，通过rpc再传给conductor，再传给scheduler。
2. scheduler按传进来的flavor和image进行调度，再把请求发给和ironic对接的nova-compute上。至此，和虚机的流程是一样的。
3. compute声明占用资源。
4. nova-compute在neutron租户网络中创建port。
5. 调用nova-compute在virt层调用ironic_driver spawn，和ironic通信，做验证和node数据更新。
6. nova-compute调用ironic-api执行部署操作。
7. ironic-conductor调用neutron设置node-port映射的neutron-port的PXE/TFTP属性。
8. ironic-conductor为裸机实例生成pxelinux的配置文件，拉取裸机deploy镜像，放置在tftp server上。
9. ironic-conductor设置裸机从网络启动。
10. ironic-conductor缓存裸机部署所需的部署镜像和用户镜像。
11. ironic-conductor控制裸机加电。

12. 裸机开机后，从neutron的dhcp服务上获取ip、获取tftp地址，并从tftp上获取boot deploy ramdisk。
13. boot deploy ramdisk运行，里面装有ironic python agent, agent和ironic conductor通信callback，将本地磁盘以iscsi的方式挂给控制结点(iSCSI部署方式)。
14. 控制裸机的pxe配置,切换成从硬盘启动。
15. 如果启动网络多租户功能的话,将裸机的网络从部署网络切换到租户网络。
16. 部署完成,裸机状态变成active。

裸金属部署涉及关键技术

> PXE/iPXE/DHCP/NBP/TFTP/IPMI

2 产品边界确认

硬件类型边界

| Hardware | Support |
| --- | --- |
| ipmi | Y |
| redfish | PLAN |
| xclarity | N |
| idrac | N |
| ilo | N |
| ilo5 | N |
| irmc | N |
| intel-ipmi | N |

ipmi, 支持IPMI 1.5/2.0硬件

redfish, 支持兼容Redfish硬件

ibmc, Huawei V5 series rack server such as 2288H V5, CH121 V5

xclarity ,联想IMM 2.0 and IMM 3.0 管理的服务器

idrac , Dell EMC服务器, 兼容IPMI和Redfish

ilo, HPE ProLiant Gen8 and Gen9 systems

ilo5, HPE ProLiant Gen10 and later systems, 兼容IPMI和Redfish

irmc, FUJITSU PRIMERGY via ServerView Common Command Interface (SCCI)

intel-ipmi, IPMI额外支持intel SST

- 多租户硬件交换机边界

  SwitchTypeSupportCisco 300-series switchesVLANYCisco IOS switchesVLANYHuawei switchesVLANYOpenVSwitchVLANYArista EOSVLANYDell Force10VLANYDell PowerConnectVLANYBrocade ICX (FastIron)VLANYRuijie switchesVLANYHPE 5900 Series switchesVLANYJuniper Junos OS switchesVLANYSDN

  - 

  PLAN
- 平台边界

  ArchSupportX86_64Yaarch64Y

- Firmware边界

  FirmwareSupportBIOSYUEFIY

- 启动类型边界

  BootSupportPXEYiPXEPLAN

- Target Distribution边界
  - Windows

  Target OSARCHSupportWindows Server 2012 R2X86_64YWindows Server 2016X86_64Y
  - Linux

  Target OSARCHSupportCentos 6, 7X86_64YCentos 7.6ARMYDebian 8 (“jessie”)X86_64YFedora 30, 31X86_64YRHEL 6, 7X86_64YUbuntu 12.04 (“precise”), 14.04 (“trusty”)X86_64YGentooX86_64YopenSUSE Leap 42.3, 15.0, 15.1 and Tumbleweed (opensuse-minimal only)X86_64Y

- 功能边界

  功能Support裸机注册Y裸机管理Y虚拟网卡Y端口管理Y端口组PLAN启动云盘PLAN

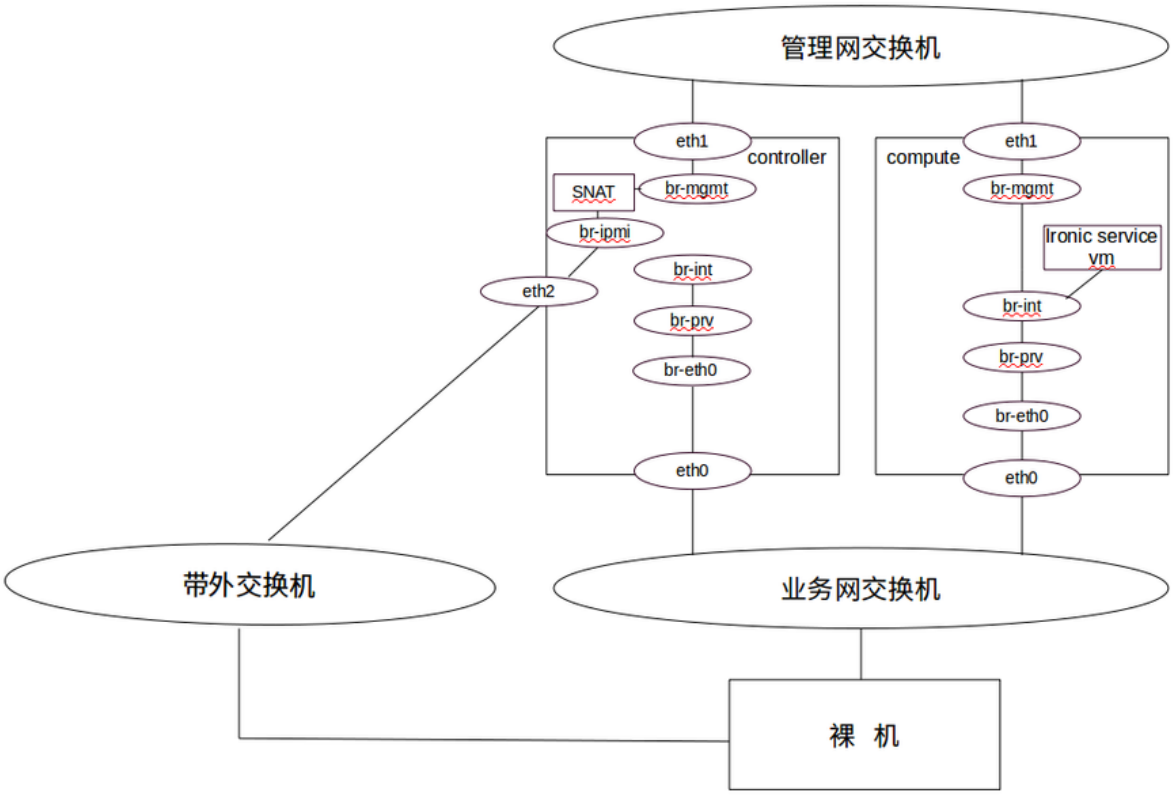
3 方案设计

- 网络拓扑

Ironic逻辑上需要以下网络:

> 管理网：ironic 服务rpc， 与其他openstack组件交互，对应br-mgmt。
>
> 带外网：ironic与裸机BMC之间的通讯网络，例如IPMI， 对应br-ipmi。
>
> 租户网：裸机部署完成后，使用的租户网络，对应br-prv。
>
> 发现网：ironic inspector自动裸机发现注册网络，如果不用自动注册功能则不需要，对应br-inspector。
>
> 部署网：部署裸机时候，下载ramdisk/kernel，ipa执行与ironic-api的lookup和heartbeat网络，可与清理网合并。
>
> 清理网：移除裸机部署时候，下载ramdisk/kernel， ipa执行与ironic-api的lookup和heartbeat网络，可与部署网合并。

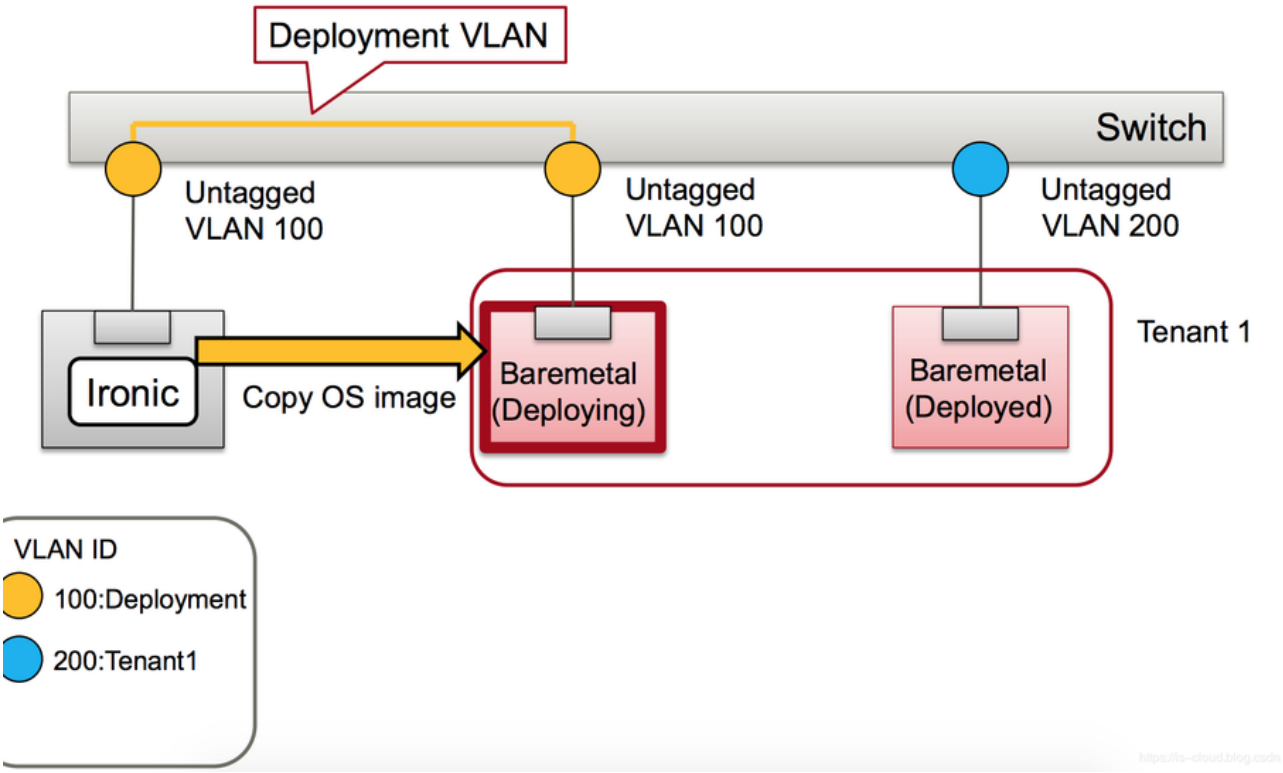
当Ironic以子服务方式部署，并且裸机使用多租户网络时，网络拓扑如下：



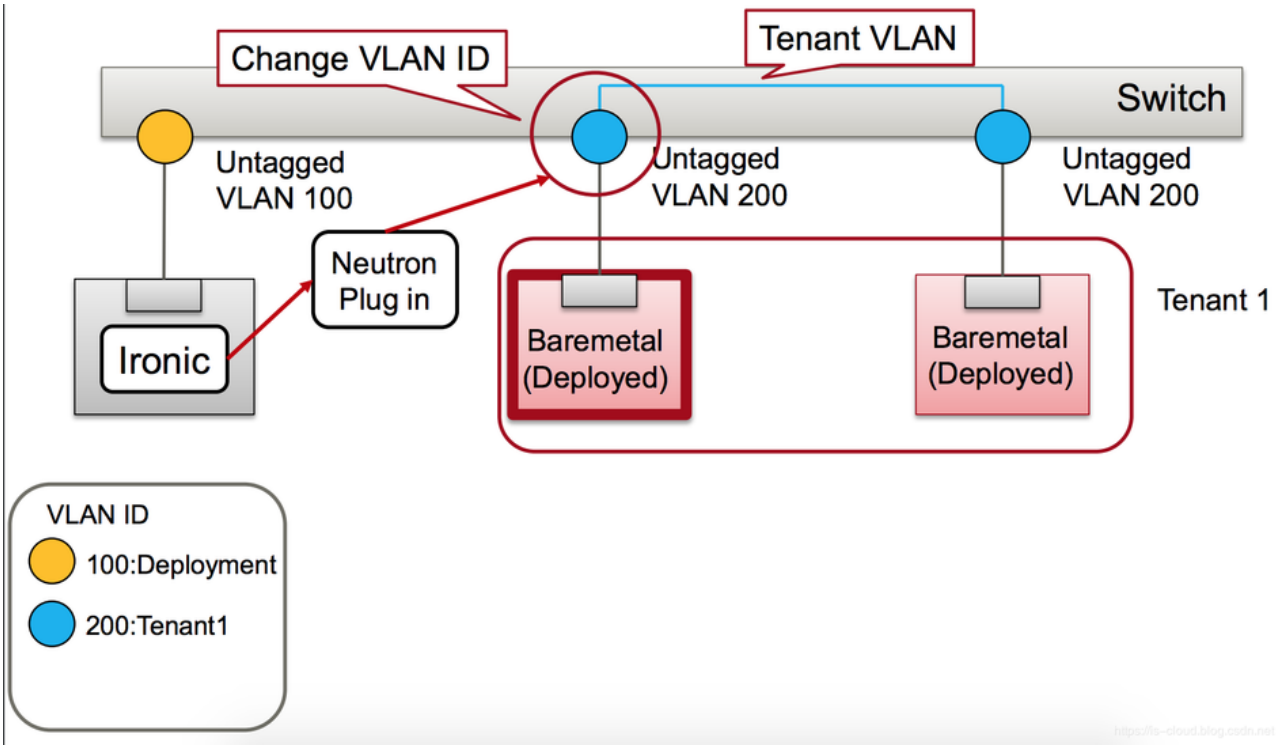

- Ironic子服务激活时在service租户下创建vlan网络，并通过br-int连接到Ironic服务虚机上，Ironic使用该网络作为裸机部署和清理网络。
- 通过控制节点SNAT在br-mgmt和br-ipmi之间转换，Ironic可以与裸机BMC之间进行通信。

多租户网络切换示意:

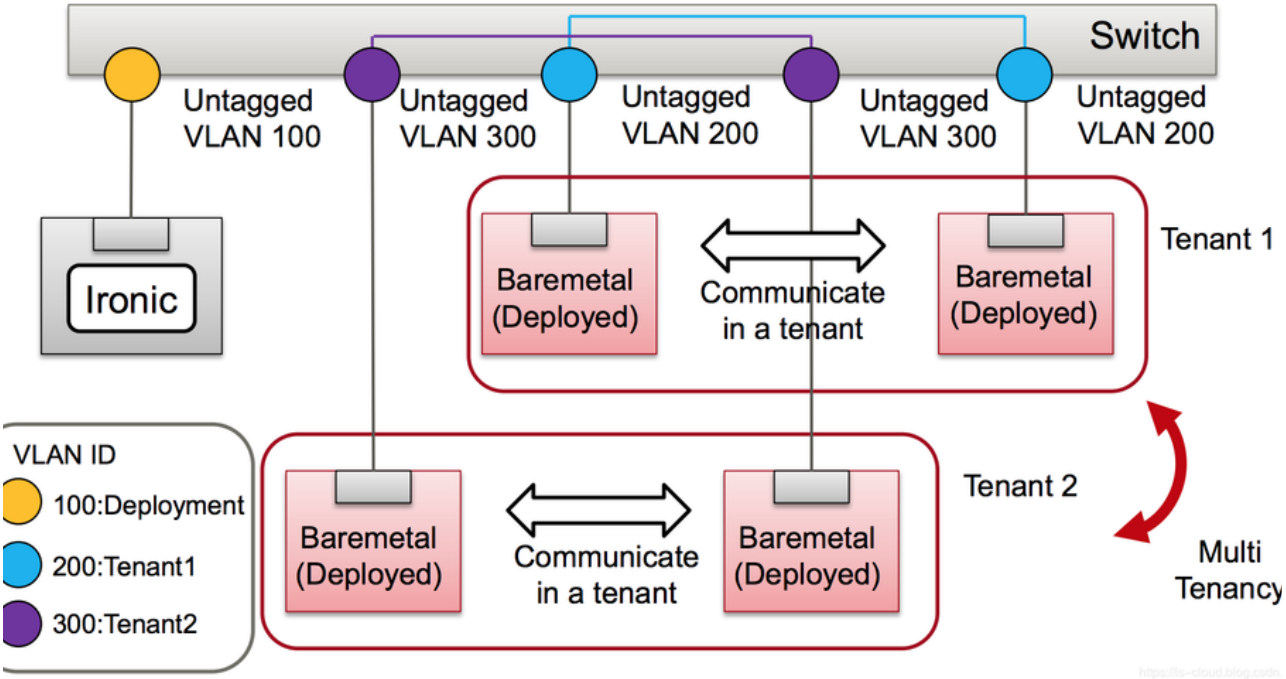
部署裸机时，裸机的网络端口在部署网络中，vlan tag与部署网络相同。




部署完成后，有neutron-generic-switch-agent调整裸机的网络端口到租户网络中。

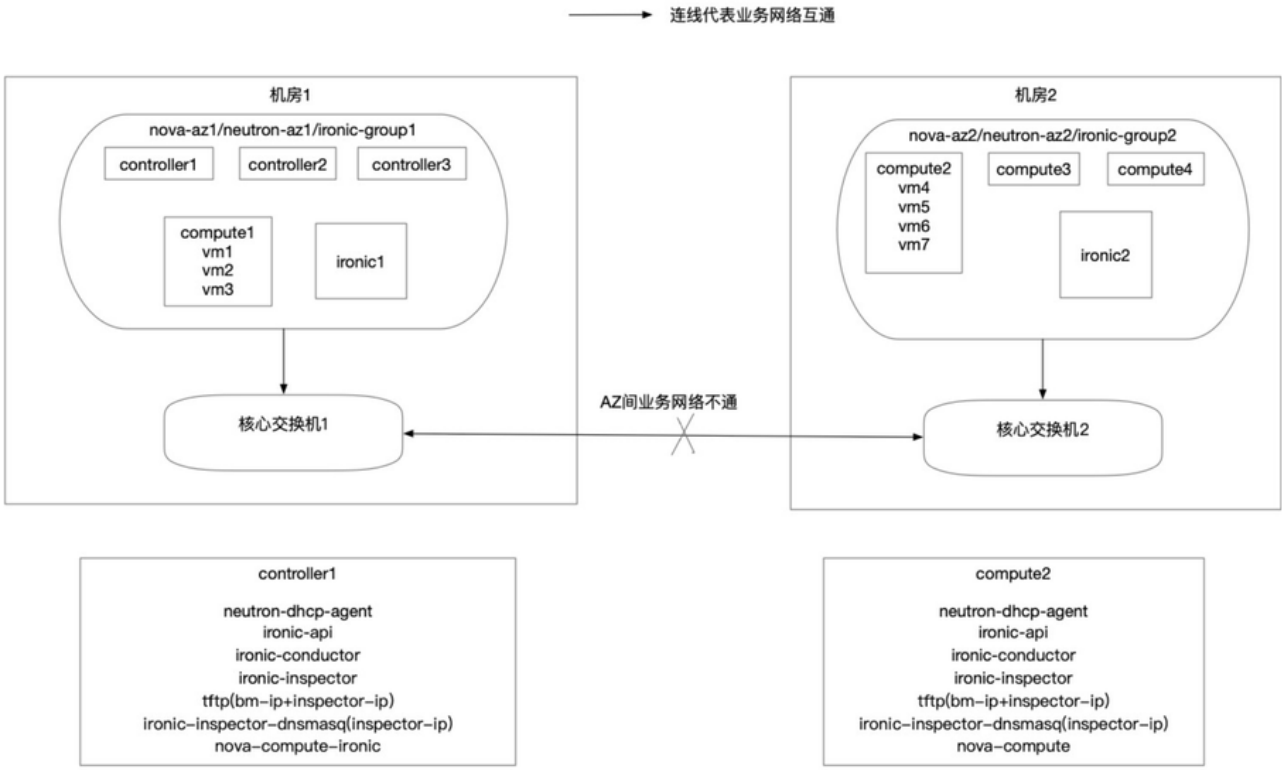



最终多租户示意:

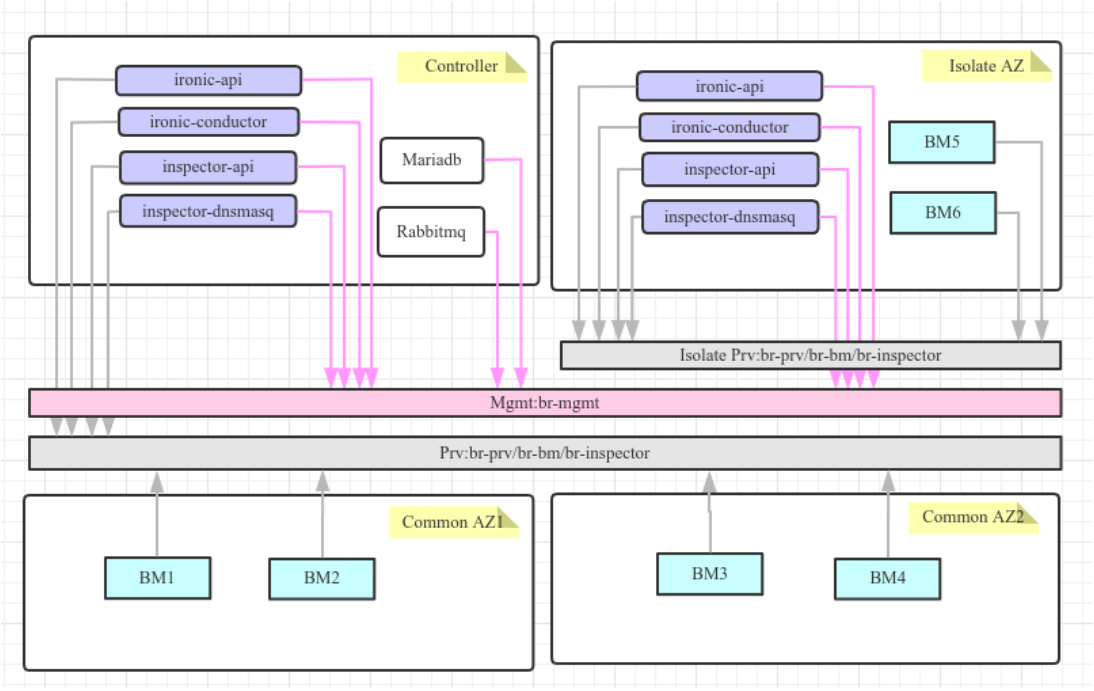



- 网络AZ

当启动网络多az时, 需要在每个隔离的生产网az内启动一组ironic(api/conductor/inspector/dnsmasq), 用于管控对应az内的裸金属服务器




从Ironic组件角度看




通过配置ironic-conductor和az的映射关系

```
[conductor]
conductor_group = az

```

在注册裸金属节点时,传入conductor_group az参数

```
openstack baremetal node set \
  --conductor-group "az" <uuid>

```

- 监控计费
  - 计费：通过flavor和ceilometer notification计费，兼容虚机场景。
  - 监控
    - 带外网络设备获取监控数据. 优点是简单无侵入；缺点是监控数据不足。
    - User Image内置 Agent方式. 优点是监控数据丰富；缺点是Agent定制,有操作系统兼容性/侵入性问题,且依赖网络,租户隔离场景无法实现。
- ECP集成

  根据标准 https://docs.easystack.cn/pages/viewpage.action?pageId=7614207 集成裸金属子服务
  - license控制裸金属子服务开启：按照子服务标准集成裸金属服务，license通过配置项控制开启子服务。
  - 裸金属子服务部署：裸金属服务和计算服务之间存在耦合服务nova_compute_ironic，与裸金属服务共同作为子服务部署。
  - 裸金属子服务支持多区域：裸金属子服务按照多区域的开发方式封装支持多区域的client。
- **其它可选方案**

暂无。

- **竞品类似方案**

暂无。

- **重要图表**
- **可升级影响**

通过ark和cube升级，无特殊升级方案和脚本。

- **稳定可靠性影响**
- **性能影响**

无影响

- **安装部署影响**

影响平台网络拓扑

- **API影响**

新增baremetal api

- **前端界面影响**

新增裸机管理页面

- **安全性影响**

无影响

- **文档影响**
- **其它影响**
  标准化部署单neutron az中ironic-api pod限制api worker数量为8，三副本pod一共24个api worker可以提供24个ironic api并发处理需求，当并发请求超过24个时会出现请求排队直至前24个请求中有完成的请求才会处理后续请求。

## 实现

- 主要实现成员

于尚斌 潘广超 王亚

- JIRA 任务 ID

JIRA Link: EAS-17736: 裸金属云产品v6.0.1开发 IMPLEMENTED

## 测试

- 自动化场景测试：

待补充。

- 单元测试

## 组件间依赖

## 待讨论确认的内容

## 修改历史

| 当前版本 | 主要修改人 | 主要签署人 | 修改时间 | 描述 |
| --- | --- | --- | --- | --- |
| v0.1 | 张玉军 | 张玉军 | 2020-04-10 | 初始版本 |
| v0.2 | 于尚斌 | 于尚斌 | 2020-08-20 | 第一次修订 |