OpenStack+Ceph+Docker

实践与思考

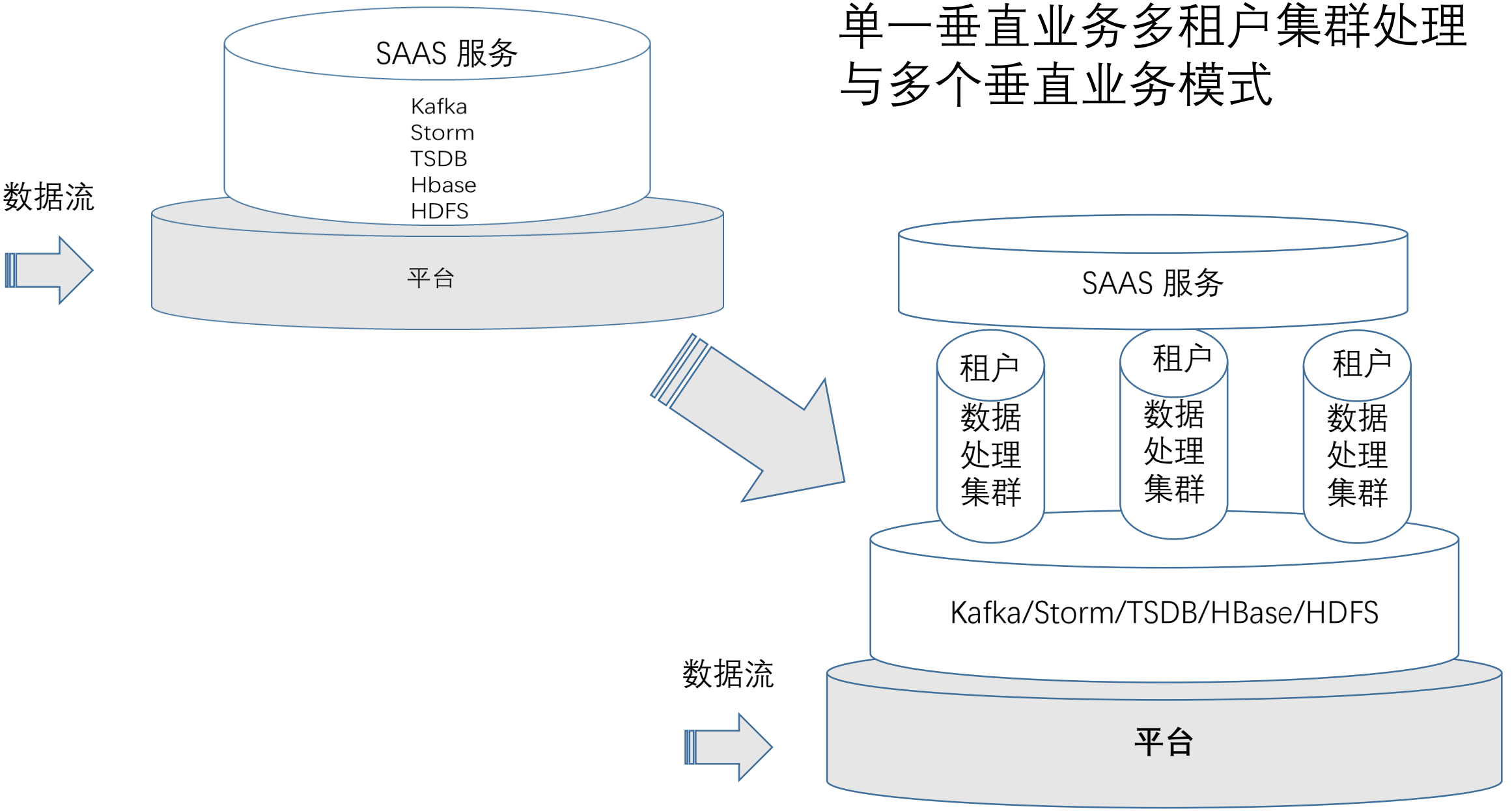
成都-子凡
OpenStack 社区
《OpenStack部署实践》作者
13618051964
153757896@qq.com

案例

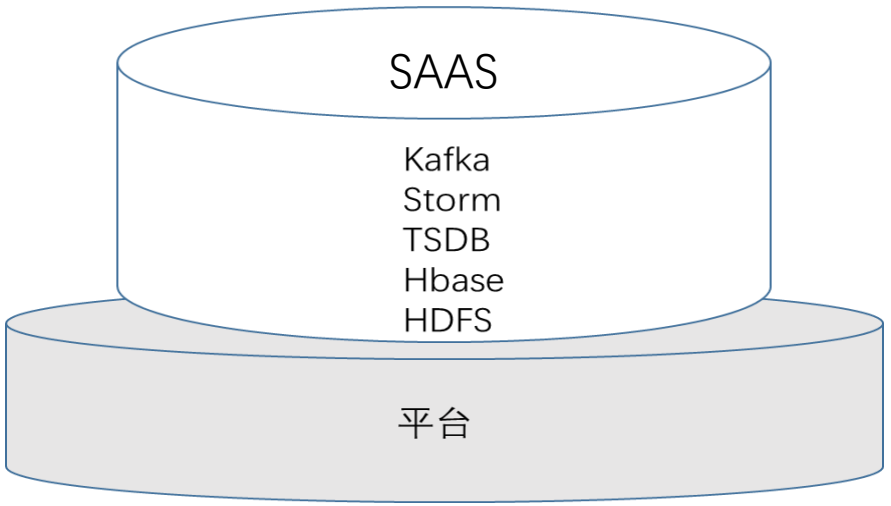
应用要求

-系统平台需要同时满足传统的Web服务及大数据处理
-数据处理采用基于Docker的kafka/Storm集群,Kafka partition 冗余处理；
-数据存储采用基于KVM虚拟机的HBase/HDFS集群, 数据存储于HDFS中
-HDFS 过期历史数据需要可靠存储，将通过hadoop/Spark作离线分析;
-一些数据处理业务运行在KVM虚拟机之中，虚拟机本身需要持久性数据卷；
-后期部分Docker业务应用需要能够加挂持久化数据卷
-SAAS 服务中Mysql 数据库读写压力非常大
-……

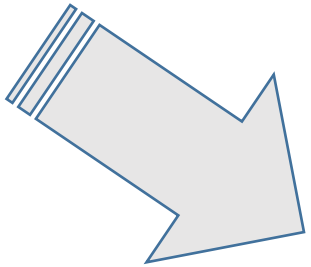
业务弹性要求：单一SAAS服务会演变为多租户处理的复杂结构

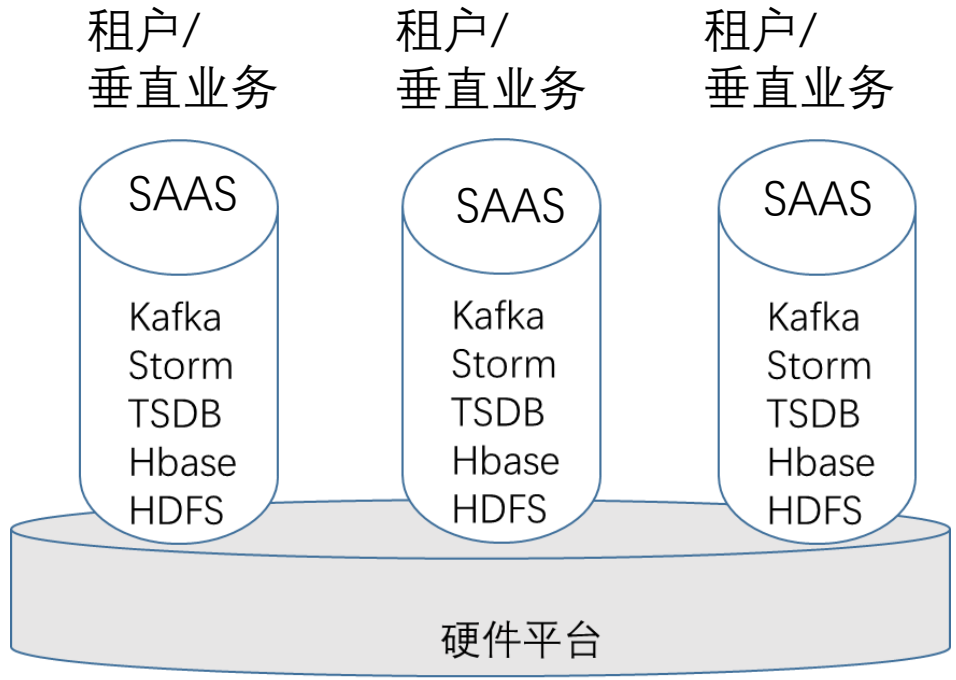


业务弹性要求：单一SAAS服务会演变为多租户处理的复杂结构



单一垂直业务多租户集群处理与多个垂直业务模式





- 不同业务间数据完全隔离
- 同一业务内不同用户间数据与处理完全隔离

解决方案一

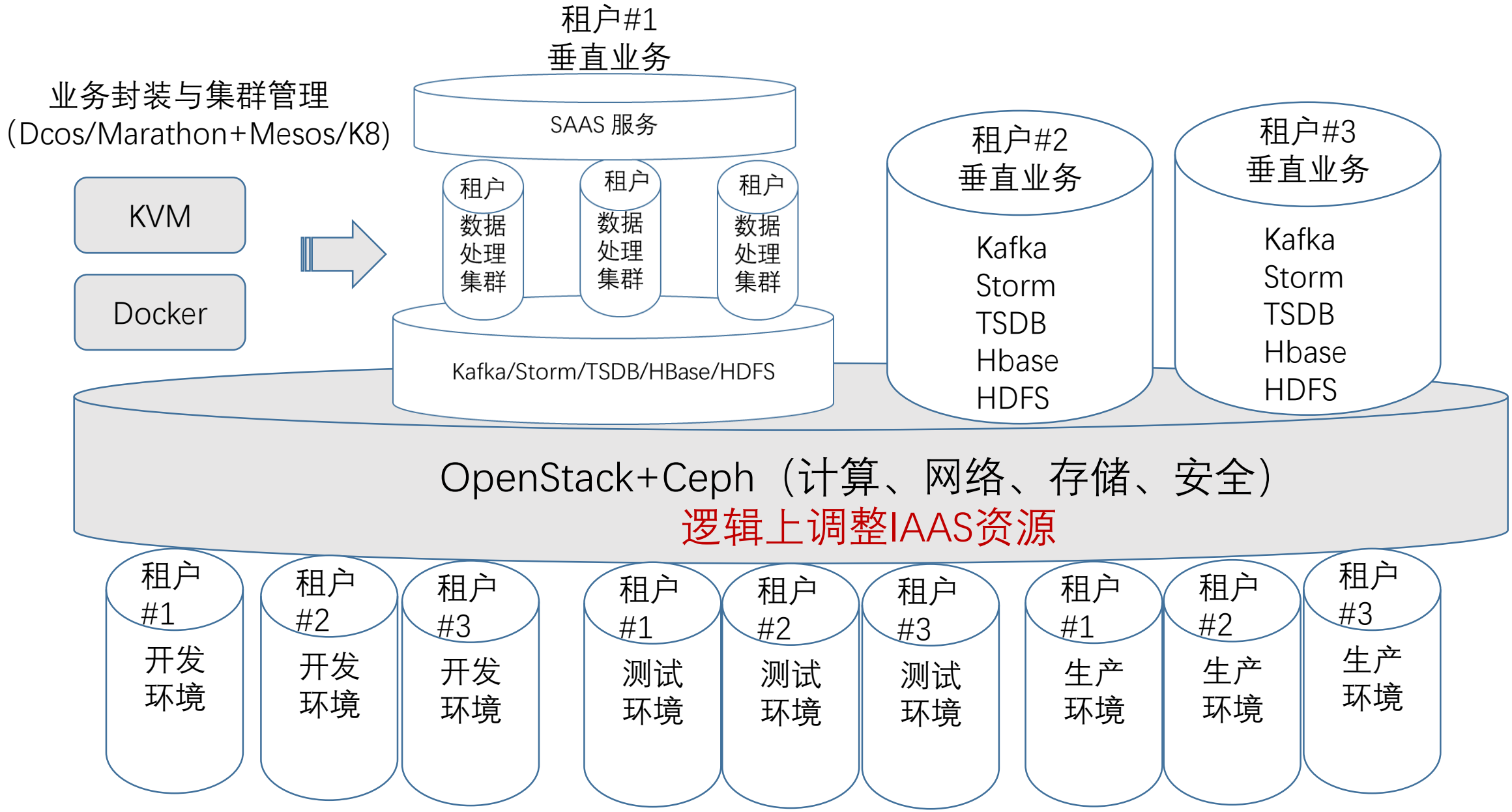
物理服务

物理服务

直接在物理机中跑各种业务性能最好，但还有以下问题需要考虑：

1. 如何满足业务发展的不确定性对平台的要求？
2. 如何延续撑已有的KVM虚拟机业务应用系统？
3. 面对业务的变化与不确定性，服务器、网络、存储与安全等IAAS资源的管理难道回到手工调节时代？
4. 如果处理HDFS层面对多租户间数据与处理的完全隔离？
5. ……

解决方案二：OpenStack+Ceph+Docker的弹性云架构



OpenStack: 以租户为基础的多垂直业务资源管理

项目

	名称	描述	项目ID
☐	FormalTenant	生产组项目	04ee5967308c45ceb23e46941f43e60c
☐	DevTenant	开发组项目	0b5b63c0c9ea415b97ac251633d0a5c8
☐	serviceTenant	service	34104028403c4b9a973303f1d516f09b
☐	adminTenant	admin	9754422ee9bd46a4ae6e5d723621cf65
☐	TestTenant	测试组项目	a544f9a4c6014fb7aecdd361639a6939

Displaying 5 items

OpenStack: 以租户为基础的虚拟机管理

实例

云主机名称 ▾ | Filter | 筛选 | 启动云主机 | 终止实例 | More Actions ▾

云主机名称	镜像名称	IP 地址	配置	值对	状态	可用域	任务	电源状态	从创建以来	Actions
cluster260-01-hadoop260-nn-01-001	sahara-kilo-vanilla-2.6-centos-6.6-UP-A2.qcow2	192.168.11.34	sahara-III-80	sahara	运行中	AG1	无	运行中	3 日，19 小时	创建快照 ▾
cluster260-01-hadoop260-dn-01-003	sahara-kilo-vanilla-2.6-centos-6.6-UP-A2.qcow2	192.168.11.33	sahara-III-80	sahara	运行中	AG1	无	运行中	3 日，19 小时	创建快照 ▾
cluster260-01-hadoop260-dn-01-002	sahara-kilo-vanilla-2.6-centos-6.6-UP-A2.qcow2	192.168.11.32	sahara-III-80	sahara	运行中	AG1	无	运行中	3 日，19 小时	创建快照 ▾
cluster260-01-hadoop260-dn-01-001	sahara-kilo-vanilla-2.6-centos-6.6-UP-A2.qcow2	192.168.11.31	sahara-III-80	sahara	运行中	AG1	无	运行中	3 日，19 小时	创建快照 ▾
OBD_OL_UPDATE	ubuntu_14.04_affs	192.168.32.40	m4.4G_50G	tongfang-user	运行中	AG2	无	运行中	2 周	创建快照 ▾
hadoop3	hadoop3/hadoop3.qcow2	192.168.32.32	m8.8G_310G	tongfang-user	运行中	AG2	无	运行中	2 周，5 日	创建快照 ▾
hadoop2	hadoop2/hadoop2.qcow2	192.168.32.31	m8.8G_310G	tongfang-user	运行中	AG2	无	运行中	2 周，5 日	创建快照 ▾
hadoop1	hadoop1/hadoop1.qcow2	192.168.32.30	m8.8G_310G	tongfang-user	运行中	AG2	无	运行中	2 周，5 日	创建快照 ▾

OpenStack: 以租户为基础的存储卷管理

云硬盘

云硬盘 | 云硬盘快照 | 卷备份

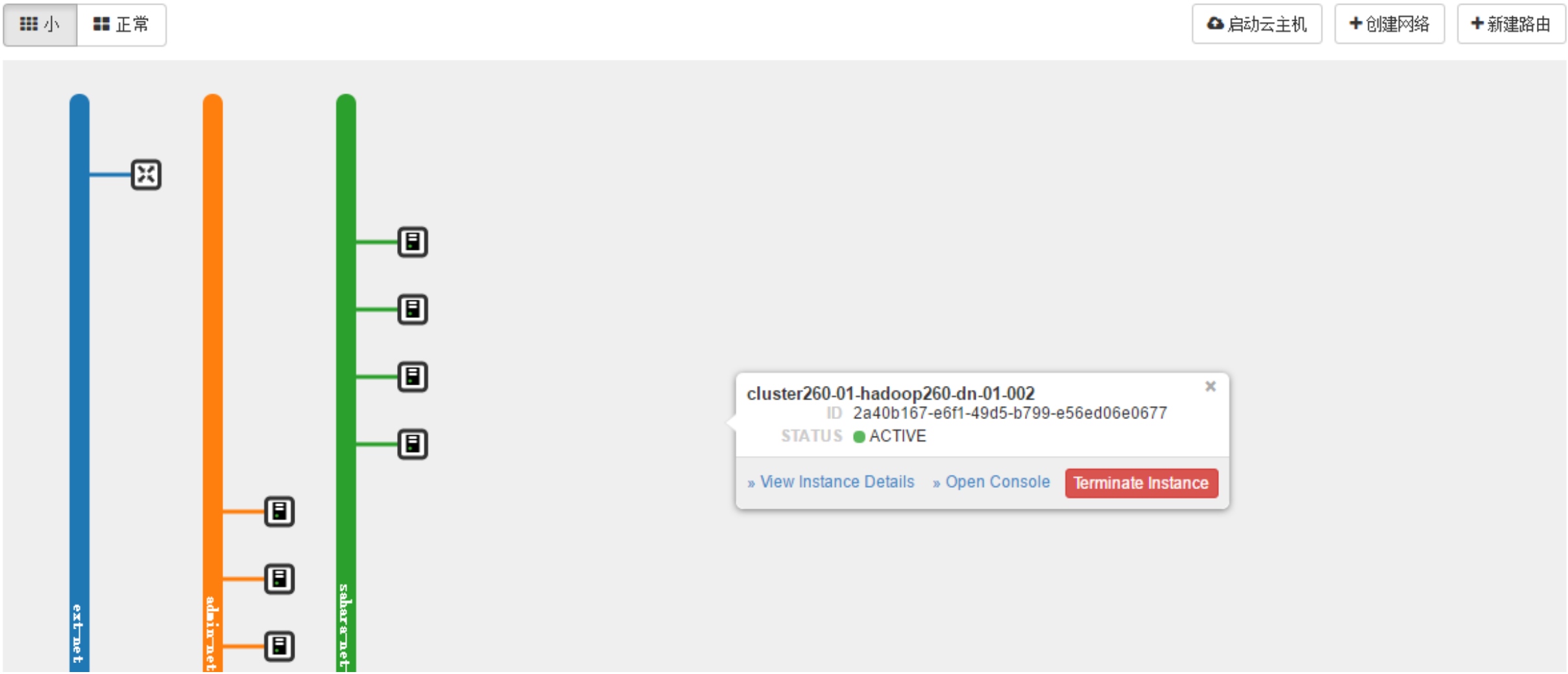
Filter

名称	描述	配置	状态	类型	连接到
hadoop03	-	800GB	正在使用	Ceph	在设备/dev/vdd上连接到hadoop3
hadoop02	-	800GB	正在使用	Ceph	在设备/dev/vdd上连接到hadoop2
hadoop01	-	800GB	正在使用	Ceph	在设备/dev/vdd上连接到hadoop1
volume_cluster11-hadoop-dn-11-001_1	-	10GB	可用配额	Ceph	
volume_cluster11-hadoop-nn-11-001_1	-	10GB	可用配额	Ceph	

Displaying 5 items

OpenStack: 以租户为基础的网络管理

网络拓扑



OpenStack: 以租户为基础的安全管理

管理安全组规则： default (ee4bd6a1-b616-4e2e-bdad-cb46d7979688)

+添加规则　✖删除规则

	方向	以太网类型（EtherType）	IP协议	端口范围	远端IP前缀	远端安全组	Actions
☐	出口	IPv4	任何	任何	0.0.0.0/0	-	删除规则
☐	出口	IPv6	任何	任何	::/0	-	删除规则
☐	入口	IPv4	任何	任何	0.0.0.0/0	-	删除规则
☐	入口	IPv6	任何	任何	-	default	删除规则
☐	入口	IPv4	任何	任何	-	default	删除规则
☐	入口	IPv4	ICMP	任何	0.0.0.0/0	-	删除规则
☐	入口	IPv4	TCP	任何	0.0.0.0/0	-	删除规则
☐	出口	IPv4	TCP	22 (SSH)	0.0.0.0/0	-	删除规则
☐	入口	IPv4	TCP	22 (SSH)	0.0.0.0/0	-	删除规则

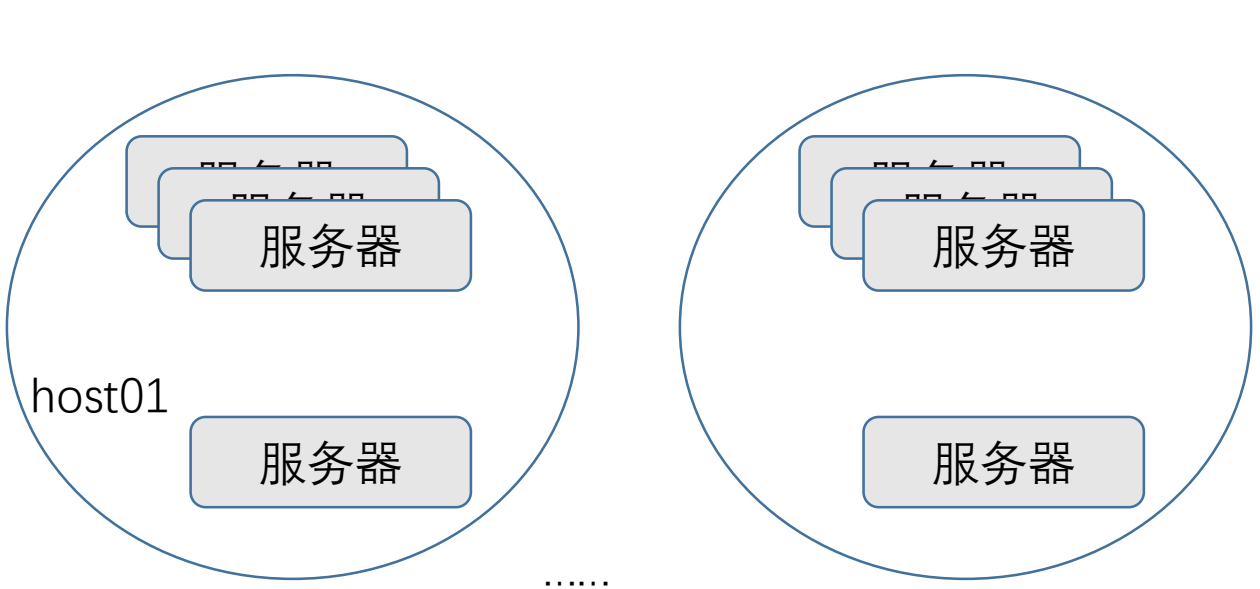
多垂直业务的性能保障

通常，我们需要业务需要按着我们预期的性能来运行
此时我们可以通过OpenStack的Aggregate与Zone 来切分服务器资源

通过这个办法，我们能够有效地保障PAAS层面、不同业务的
DCOS/Marathon+Mesos/Kubernetes集群的稳定处理能力

OpenStack下的计算资源管理

将服务器资源切分到不同的聚合组/可用区内



服务器聚合组#1
Servergroup01

服务器聚合组#2
Servergroup02

同一集群管理下
多个业务分别在不同的聚合组中，
统一管理又互不影响

同一集群管理下
研发、测试与生产环境并存又互不影响

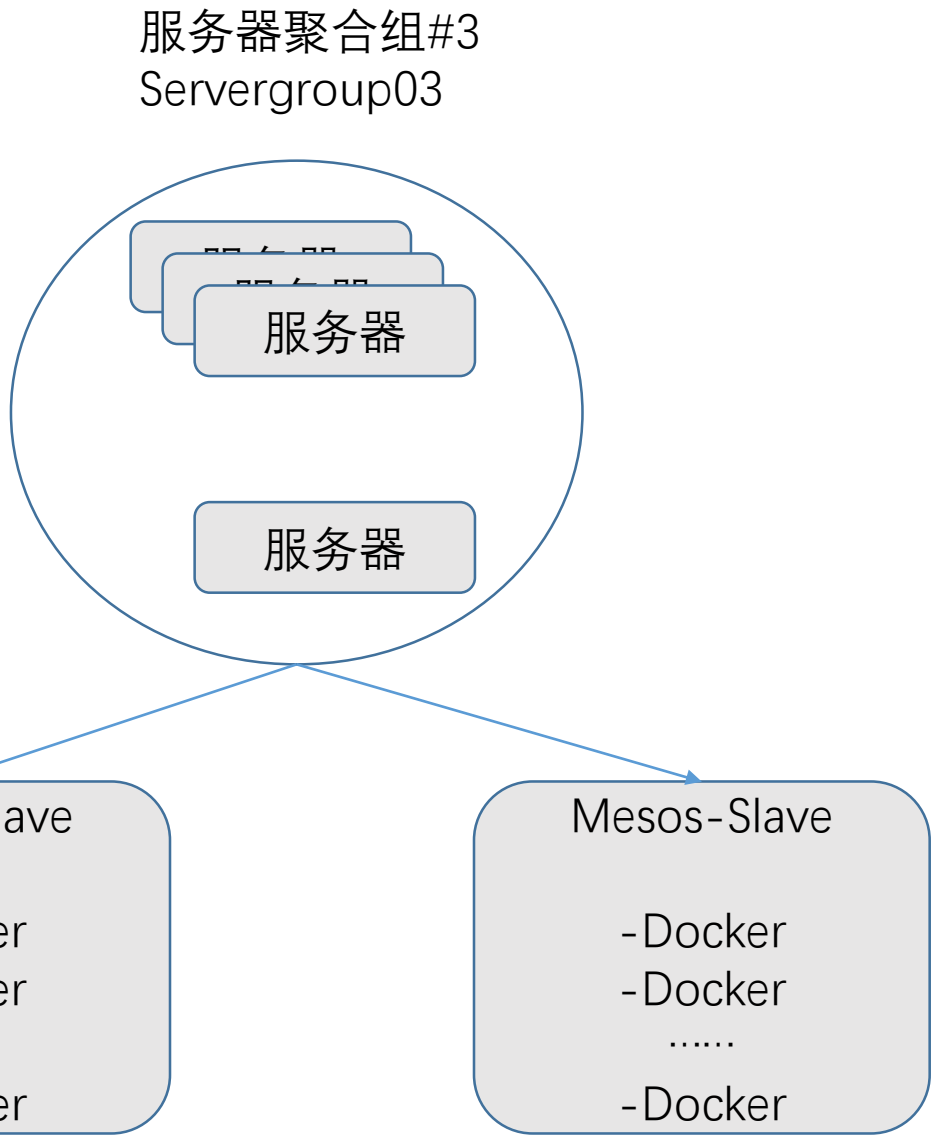
同一集群管理下
多个Mesos/K8S集群并存

nova aggregate-create/list
nova aggregate-add-host servergroup01 host01

OpenStack下的计算资源管理

如何确保Marathon+Mesos+docker模式下恒定的计算能力？

一个物理机聚合组中，每个物理机仅仅运行一个大的KVM 虚拟机，
将组中的服务器计算资源全部交给Mesos来管理使用：



物理机=KVM虚拟机
CoreOS/Ubuntu/Centos

界限

OpenStack+Docker 如何融合？

OpenStack社区提供的方法

- Nova-Docker
- Heat+Docker
- Magnum

纷杂现象

OpenStack是IAAS层面的利器，但它想进入Docker的应用管理世界；

Docker是应用层面的利器，但它想进入IAAS管理层：
-Flocker
-Network
……
未来是OpenStack还是Docker?
互联网产生后，报纸消失了吗？
边界在哪里？

我的看法是：还给他们自由……

朴实无华的融合方案
-Docker run in VM

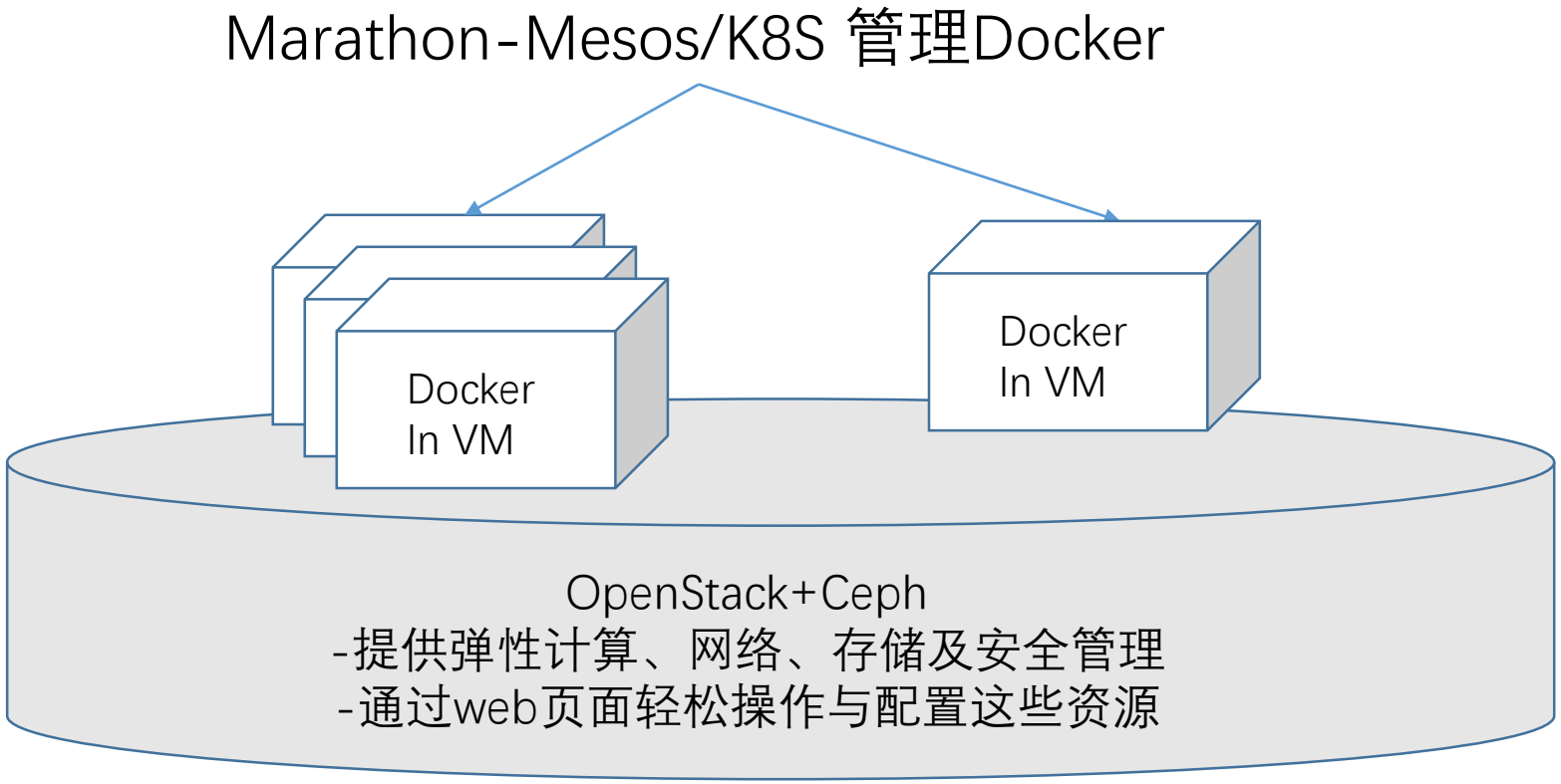
将Docker 跑在虚拟机之中：

- **虚拟机与IAAS交由OpenStack统一管理**
- **应用封装到Docker中由Mesos/K8s 管理**

额外益处：

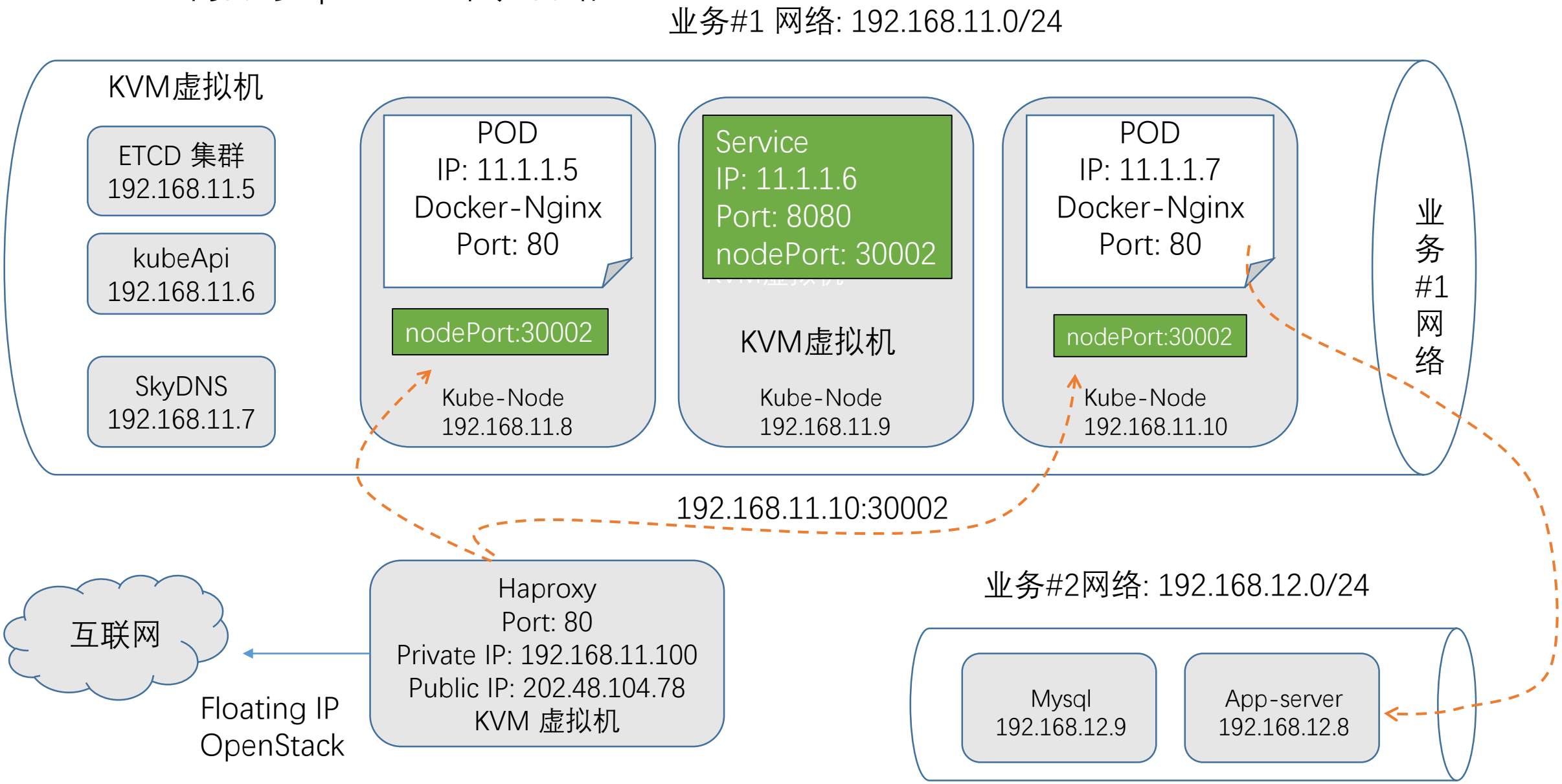
- Docker引导起虚拟机内核崩溃不影响整个物理机

- Docker安全性提高



Kubernetes 内部网络与 OpenStack 租户网络

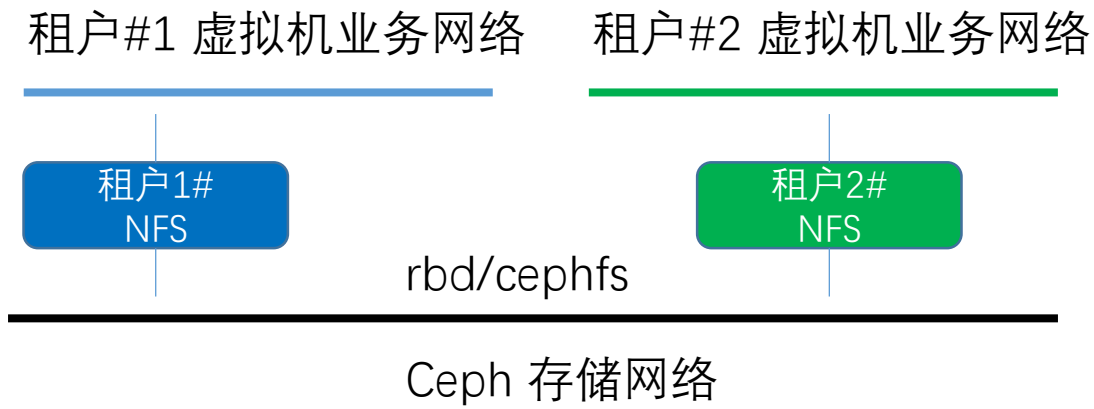
Kubernetes 内网与OpenStack 租户网络



Docker 微服务对共享文件系统的依赖

K8S RC 使用共享NFS：

```
apiVersion: v1
    kind: ReplicationController
    metadata:
      name: node-manager
      labels:
        name: node-manager
    spec:
      replicas: 1
      selector:
        name: node-manager
      template:
        metadata:
          labels:
            name: node-manager
        spec:
          containers:
          -
            name: node-manager
            image: image/node-manager
            volumeMounts:
              - name: nfs
                mountPath: "/usr/share/nginx/html"
          volumes:
            - name: nfs
              nfs:
                server: 192.168.13.13
                path: "/mnt/nfsshare"
```



独立 Docker数据持久化

1. Ceph给VM提供Volume，Docker在vm中启动后带入

2. Ceph直接给裸Docker 提供volume

rbd-docker-plugin插件

创建一个用于支持Docker的20G ceph卷：
rbd-docker-plugin --create --user=docker --pool=docker &

在docker虚拟机启动项中加入--volume选项及参数：
docker run -it -p 50001:22 --volume-driver=rbd --volume foo:/mnt/foo ubuntu-1404-romi bash

登录到虚拟机中查看信息，通过df命令发现多了一个/dev/rbd1的磁盘卷：
root@66593760b2f7:/# df
Filesystem   1K-blocks   Used Available Use% Mounted on

/dev/rbd1   20961280   33344 20927936   1% /mnt/foo

思考

融合OpenStack与Docker社区的精华
有效融合而不是互斥

成都-子凡

13618051964

153757896@qq.com