跨地域云资源共享

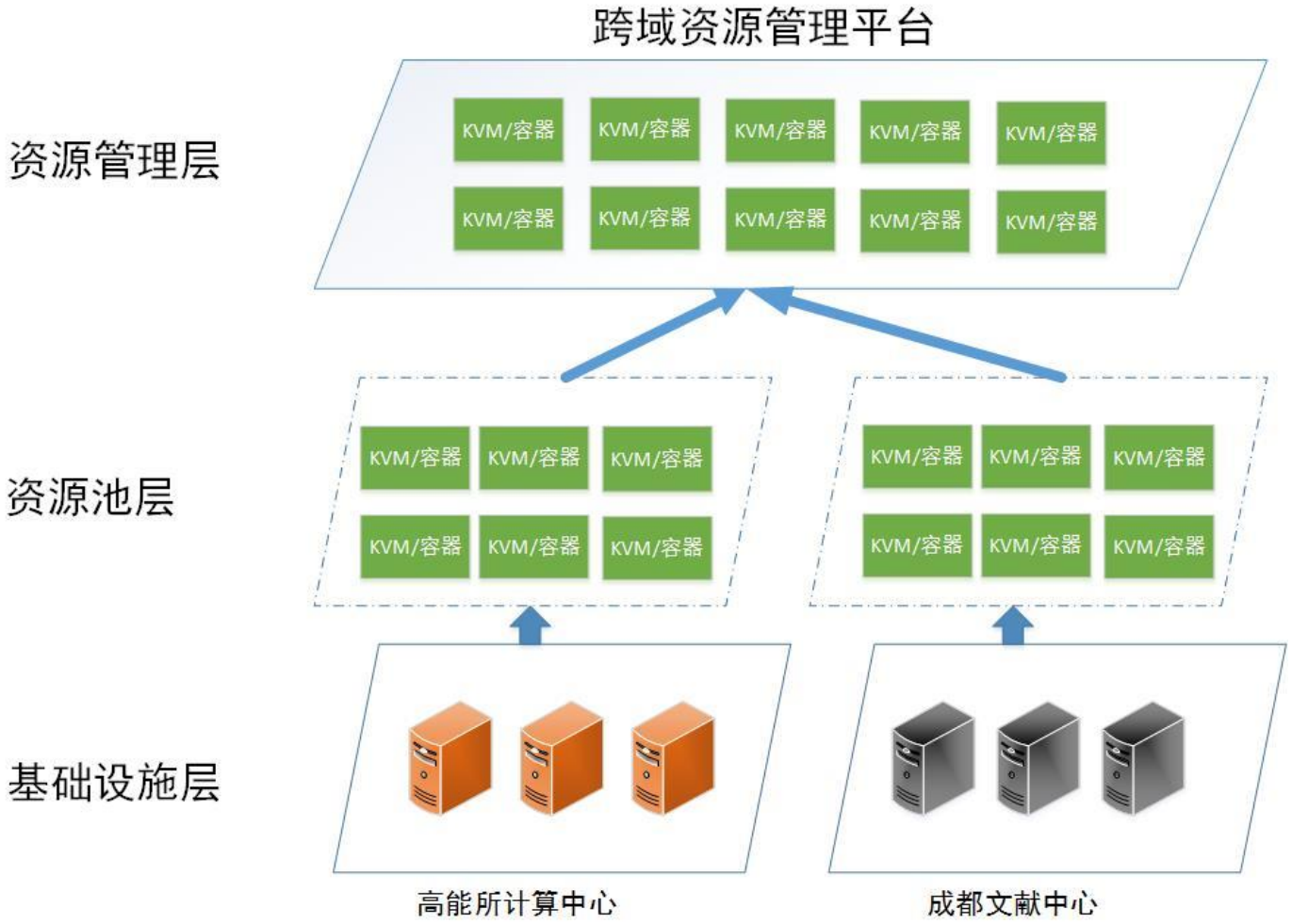
黄秋兰，李亚康
高能所/东莞
2018-06-26

提纲

- 跨地域资源共享与高能物理的计算需求
- 跨地域云资源共享的现状
- 高能物理跨域云资源共享方案及关键技术
- 总结

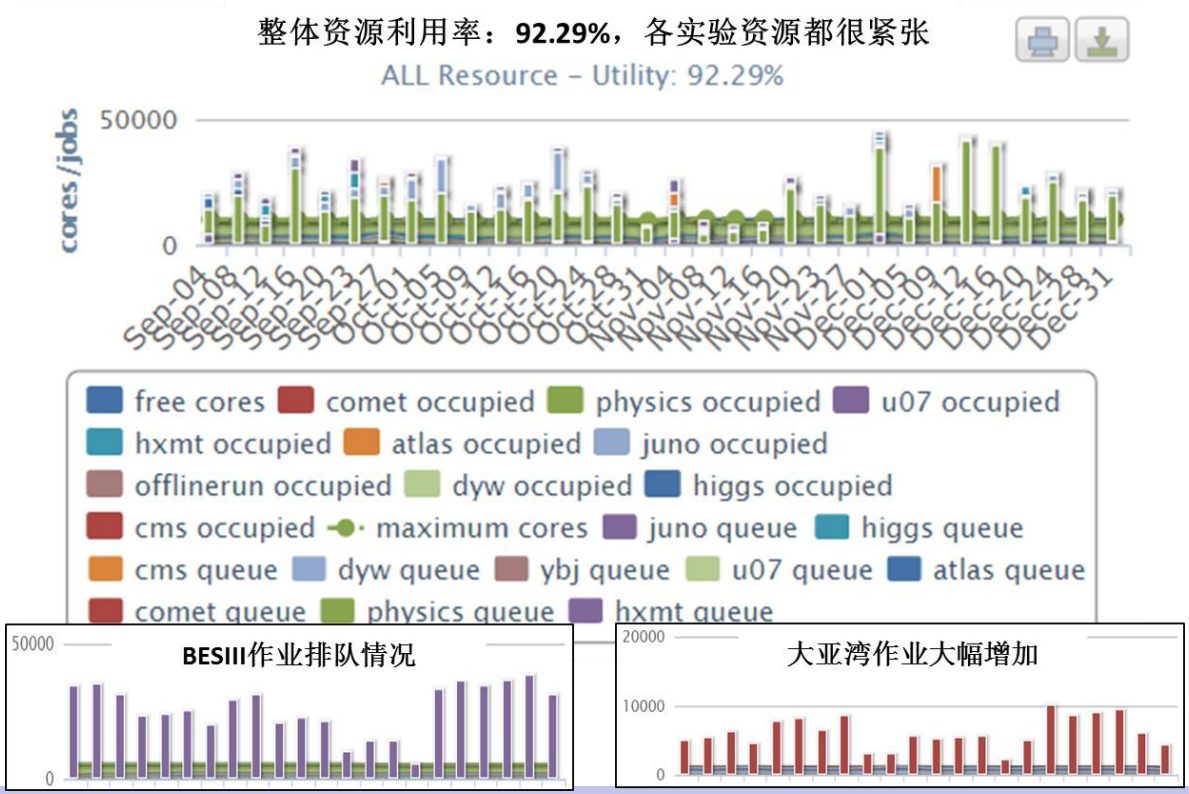
什么是跨域云资源共享？

- 跨地域云资源共享采用云联盟技术组织和融合跨地域的资源，实现资源的互联互通和高度共享
  - 实现资源的统一管理和调度
  - 通过虚拟化技术屏蔽底层基础设施的异构性，节省了维护成本
- 与传统的集群、网格计算和DIRAC分布式计算相比
  - 资源更加统一
  - 节省运维成本
  - 与应用解耦合
  - 更加透明
  - 架构相对简单


跨域资源管理平台
资源管理层
资源池层
基础设施层
KVM/容器
高能所计算中心
成都文献中心

高能物理计算需求

- 高能所计算中心支持BESIII、大亚湾、JUNO、LHAASO、CMS、Atlas等实验
  - ~15000个CPU核，11PB的磁盘空间
- 现有的单数据中心的计算资源紧张，计算集群中总出现大量作业处于排队状态



高能物理计算需求

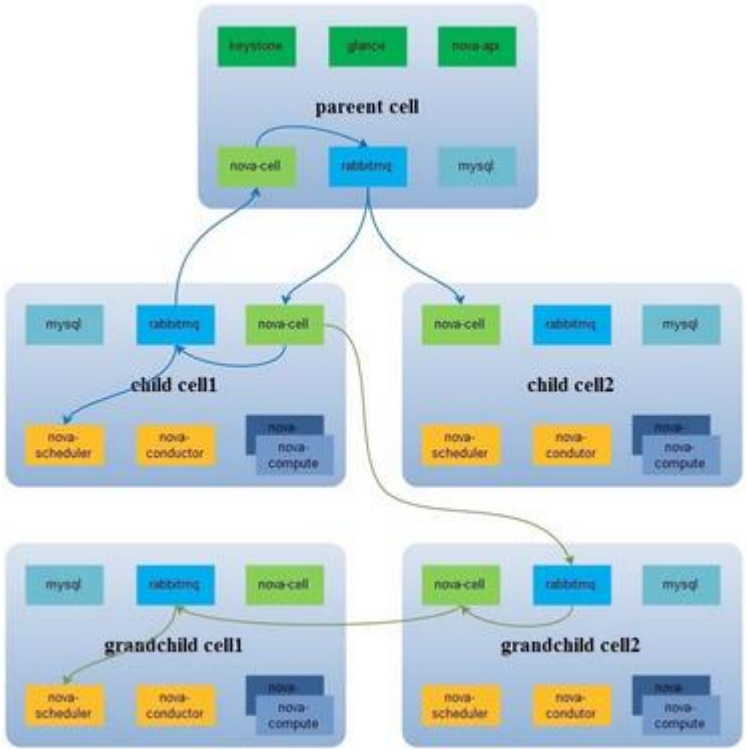
- 有必要对各实验合作组站点的资源及其他云资源进行整合
  - 扩展计算资源
  - 促进实验合作组站点间资源的共享
- 直接利用跨域资源面临的问题
  - 异地站点的系统不稳定
  - 缺少专门的技术人员，运维成本高
- 采用云联盟的方式对跨域资源进行有效的组织和管理，实现跨域资源的整合和共享
  - 采用虚拟化技术屏蔽底层细节
  - 利用Openstack对跨域异构资源进行统一管理

跨域云资源共享现状

- 单Openstack环境，主要用于解决单数据中心的扩展性问题
  - nova cell V1
  - nova cell V2
- 跨数据中心的可扩展技术，管理多个Openstack
  - Multi-region
  - Cascading
  - Tricircle

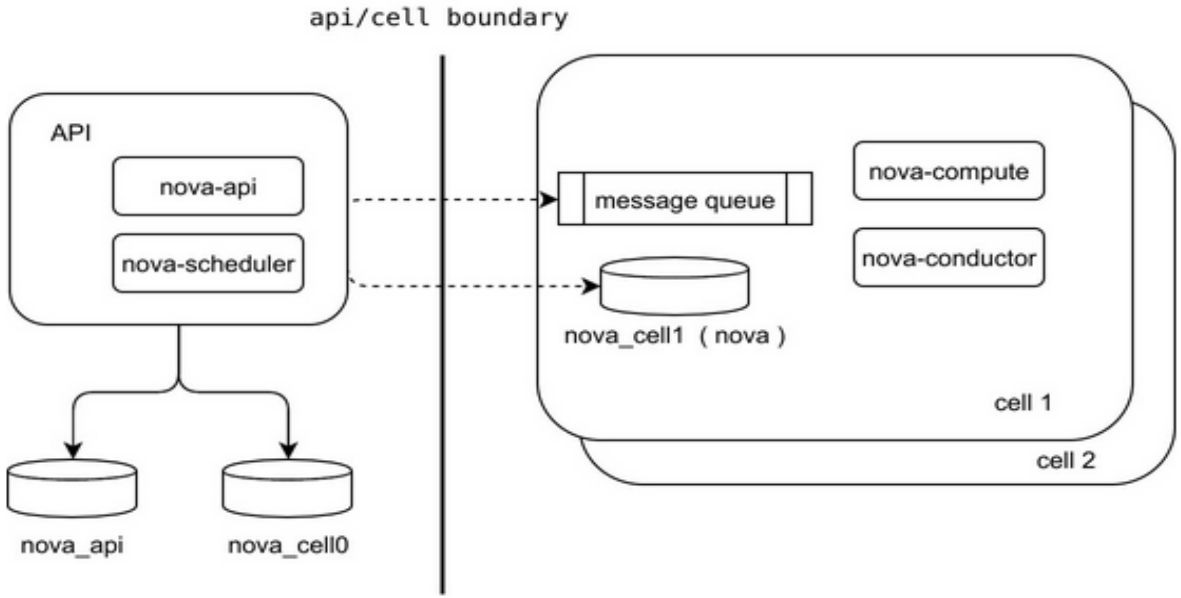
Cell V1

- 一直没有推动起来，导致该模块成熟度低，使用率低
- 不支持安全组、主机集合、可用域、特定的调度算法
- 一个nova-cells需要做两件事情的调度，既有内部的也有各个cell之间的
- 资源竞争问题
- 如果之前没使用nova-cell。重新准备使用nova-cell会比较困难
- 最顶层的那个top cell 无法扩展
- 各个层之间的数据复制



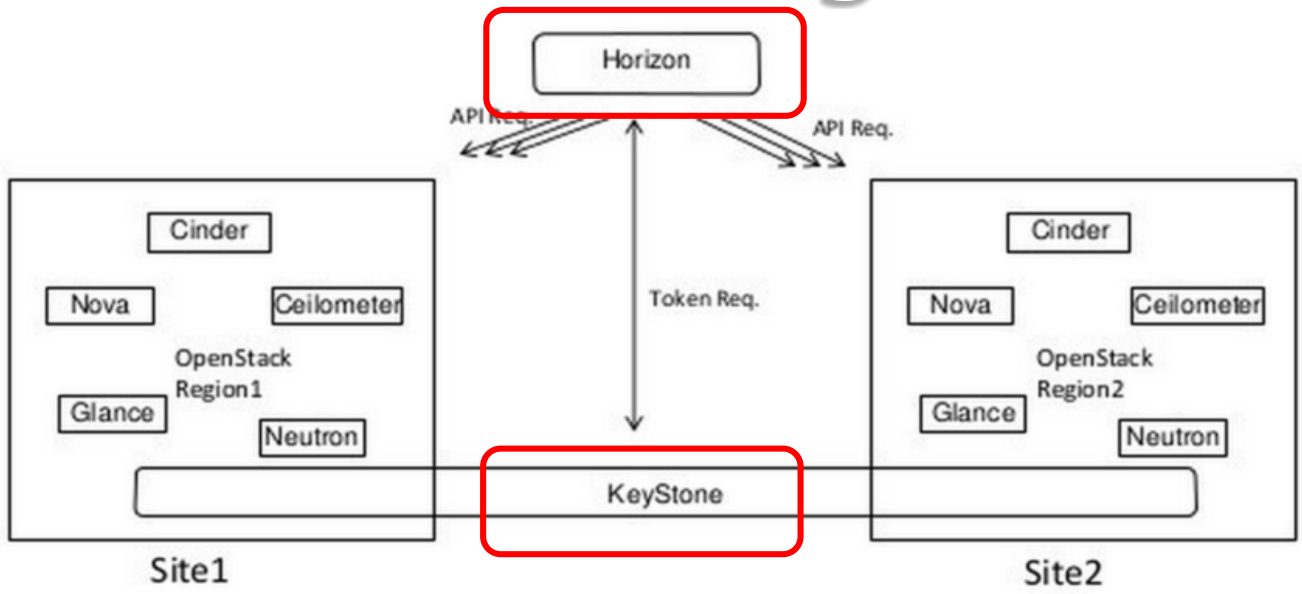
CERN采用的V1方案

Cell V2



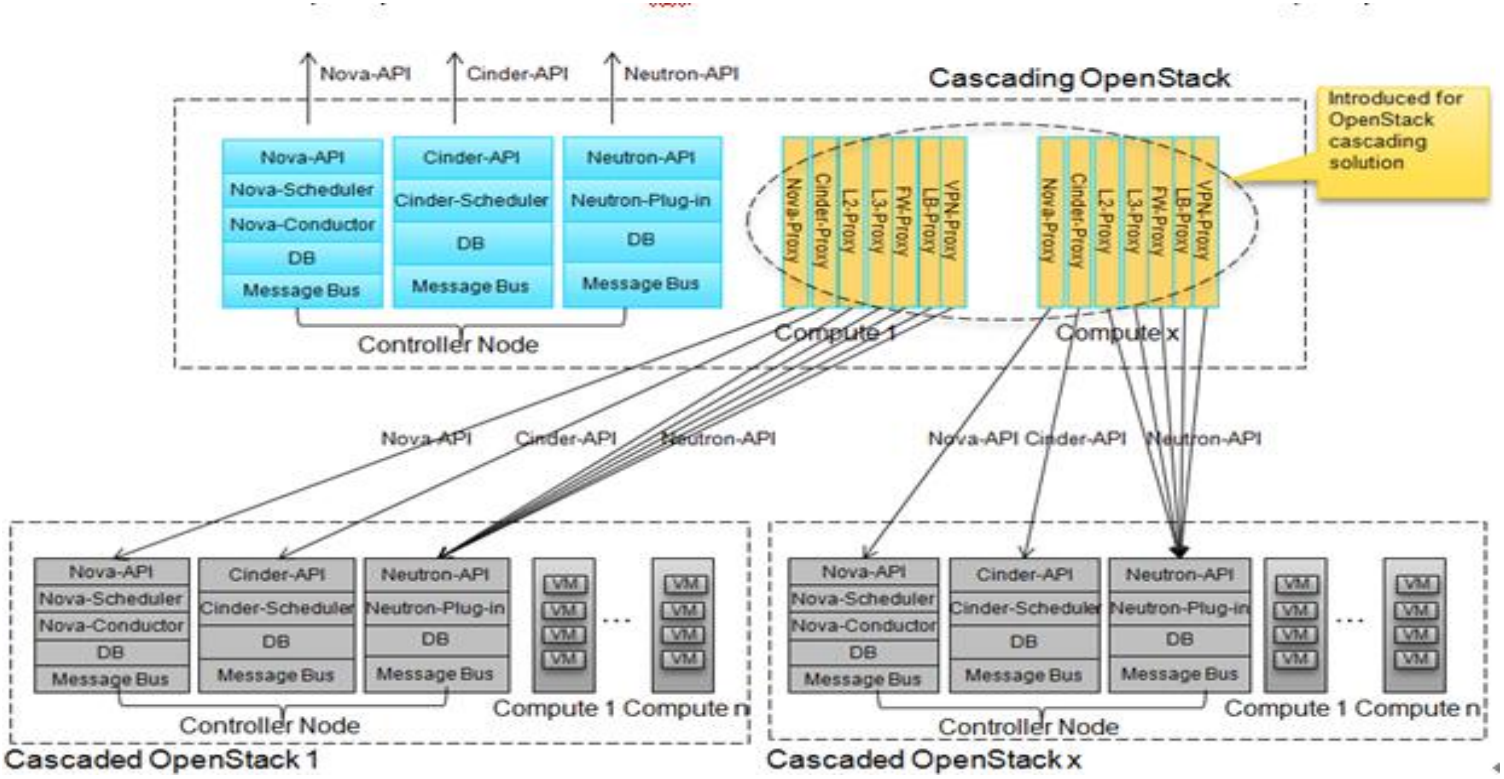
- 所有的 cell 变成一个扁平架构 。 比之前的多层父子架构要简化很多
- api 上面服务会直接连接 cell 的 MQ 和 DB, 所以不需要类似 nova-cell 这样的额外服务存在 ，性能上也会有极大的提升

Multi-region



- Region用来区分地理位置上分开的两openstack节点，region之间完全隔离
- 用户可以选择离自己更近的region来使用里面的资源
- 多个regions之间共享同一个keystone 和 dashboard
- 优点
  - 部署简单，keystone中直接指明不同region就可以实现
  - 逻辑概念清晰，并且已经在生产中大量的使用
- 缺点
  - openstack 节点之间相互隔离，不能互访
  - 不同版本的openstack部署会有问题

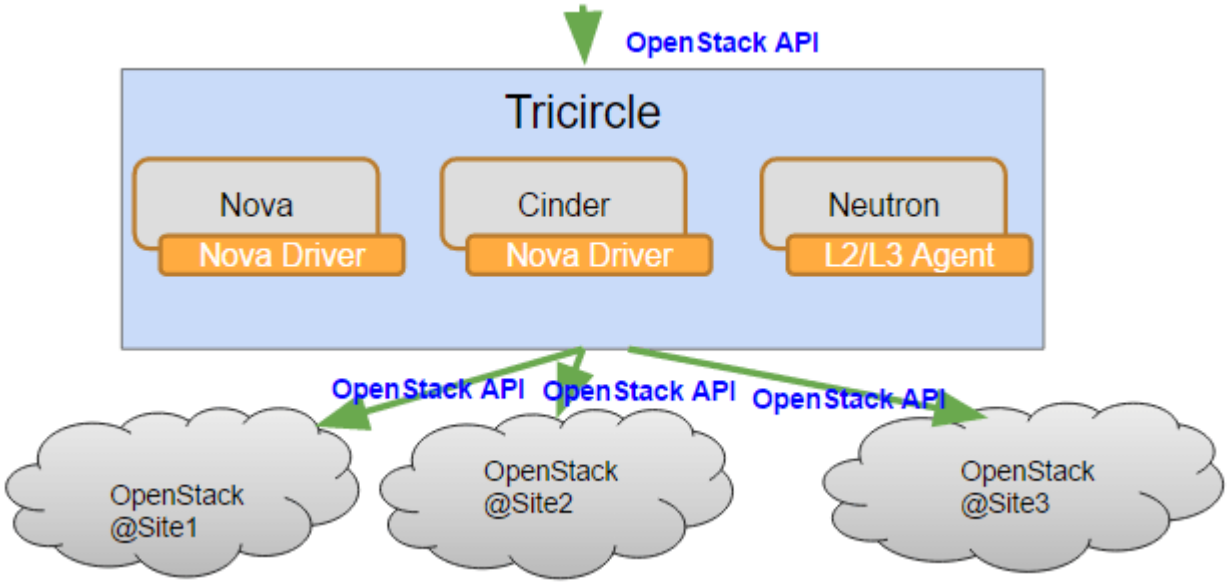
cascading



缺点

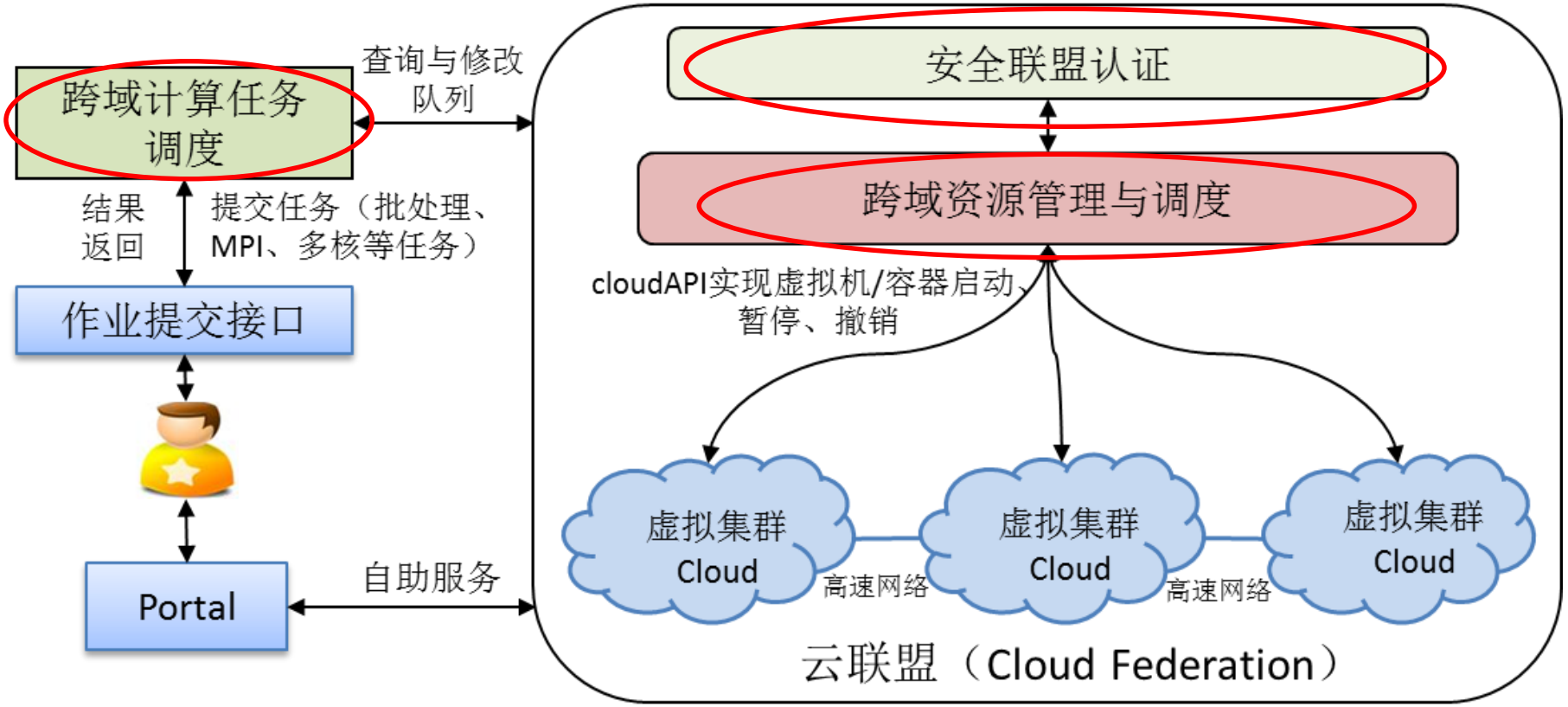
- 现在只支持到J版本
- 对于neutron的支持存在很多的问题
- Cascading openstack 的api 管理属于有状态的请求，所以扩展会不方便

Tricircle

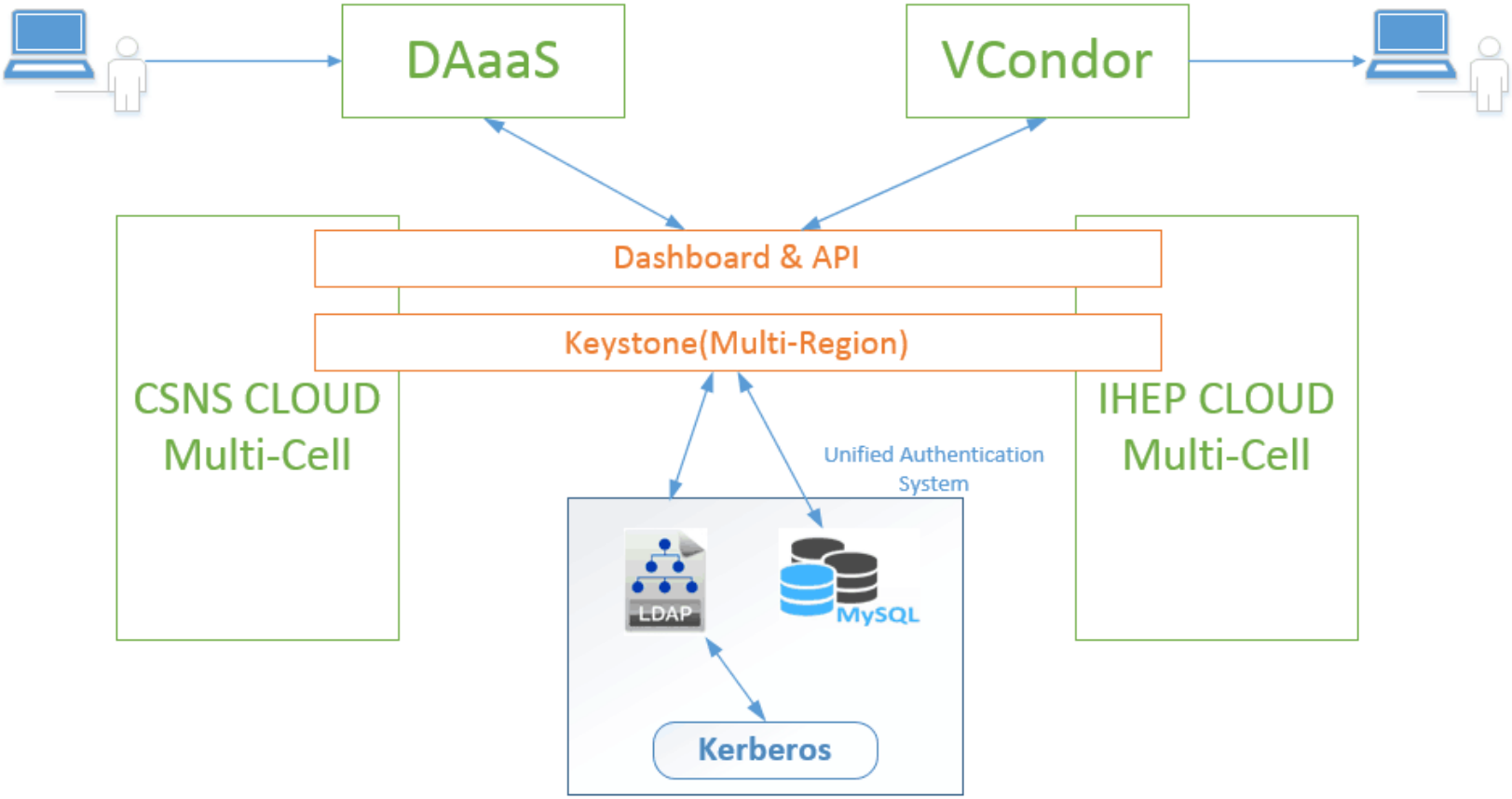


- 上层的openstack也是通过 openstack api和底层openstack互通
- Tricircle使用 stateless 机制，更利于扩展节点
- 实现机制Tricircle 更加类似 nova-cell的机制，不过nova-cell是只扩展了nova，而Tricircle则是所有的组件都进行了扩展

高能物理跨地域云资源共享

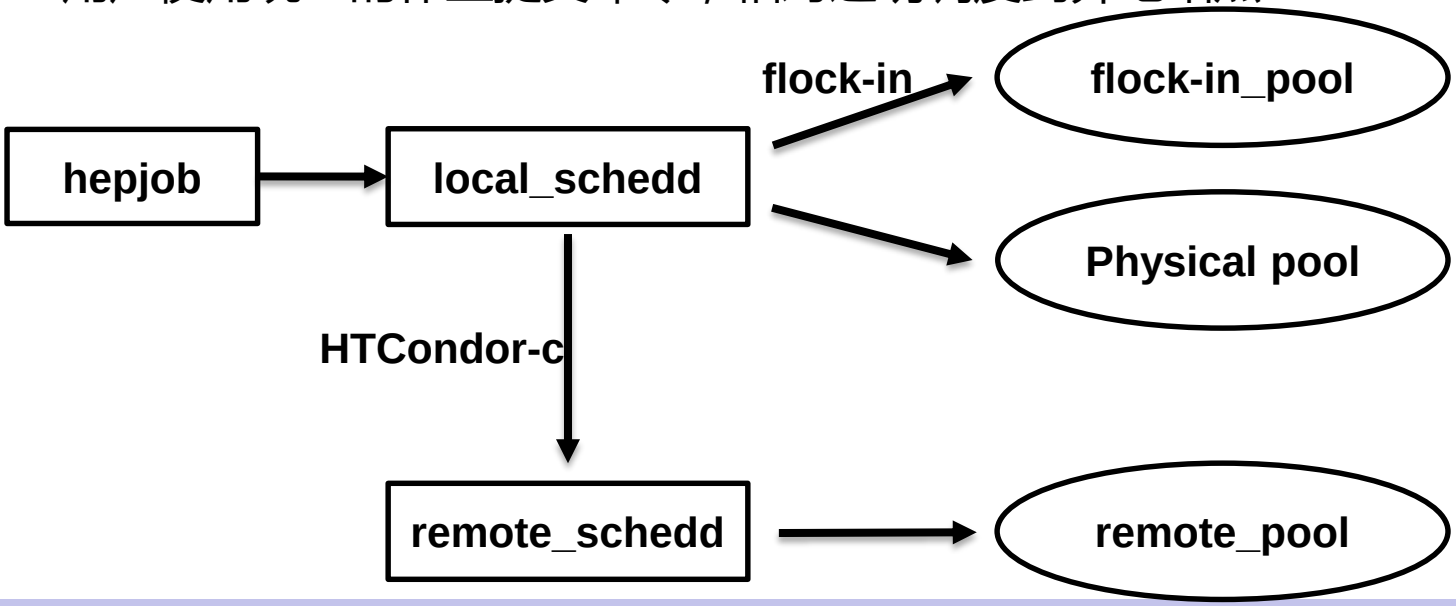


基于Openstack的跨域云资源共享架构



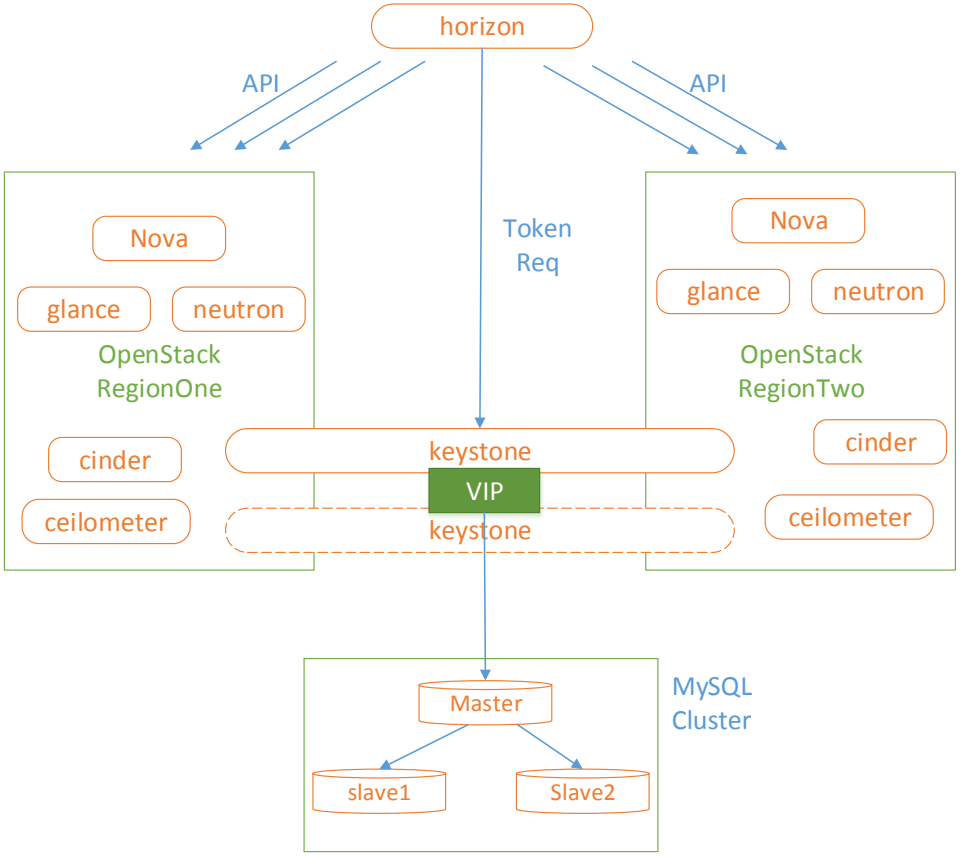
相关技术

- 跨地域资源管理和调度
  - Openstack的multi-region技术
- 联盟认证
  - 统一认证入口
  - 用户信息在ldap和 Openstack的mysql中
- 作业跨域调度
  - 基于HTCondor-C的跨域调度
  - 用户使用统一的作业提交命令，相对透明调度到异地站点



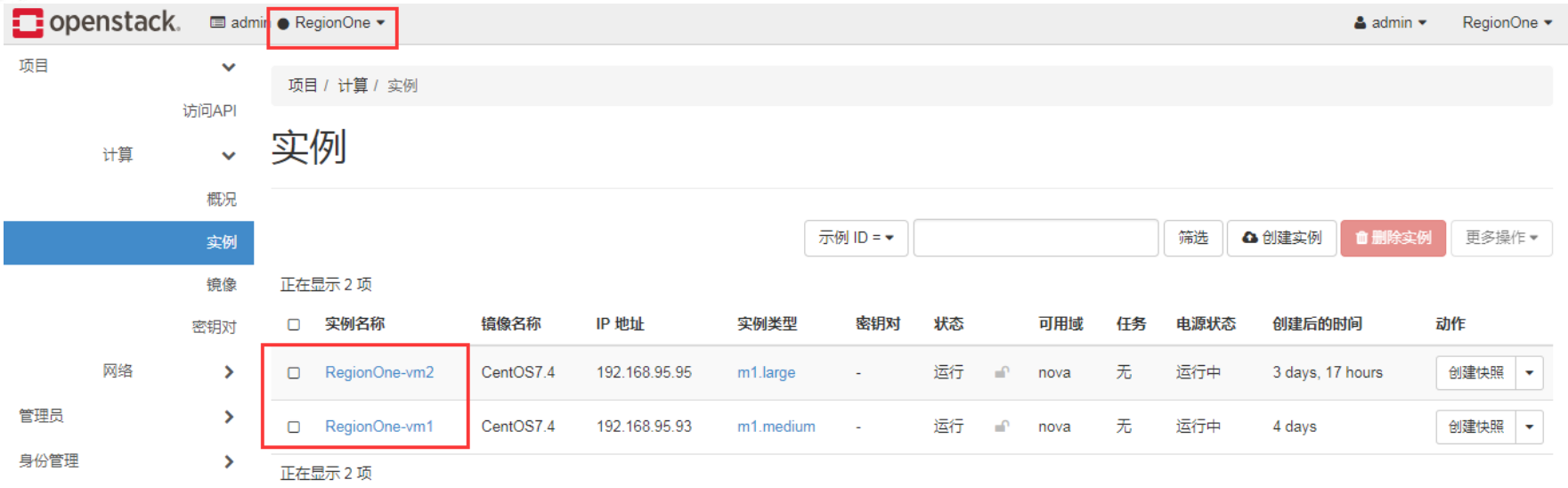
北京-东莞的Multi-Region部署

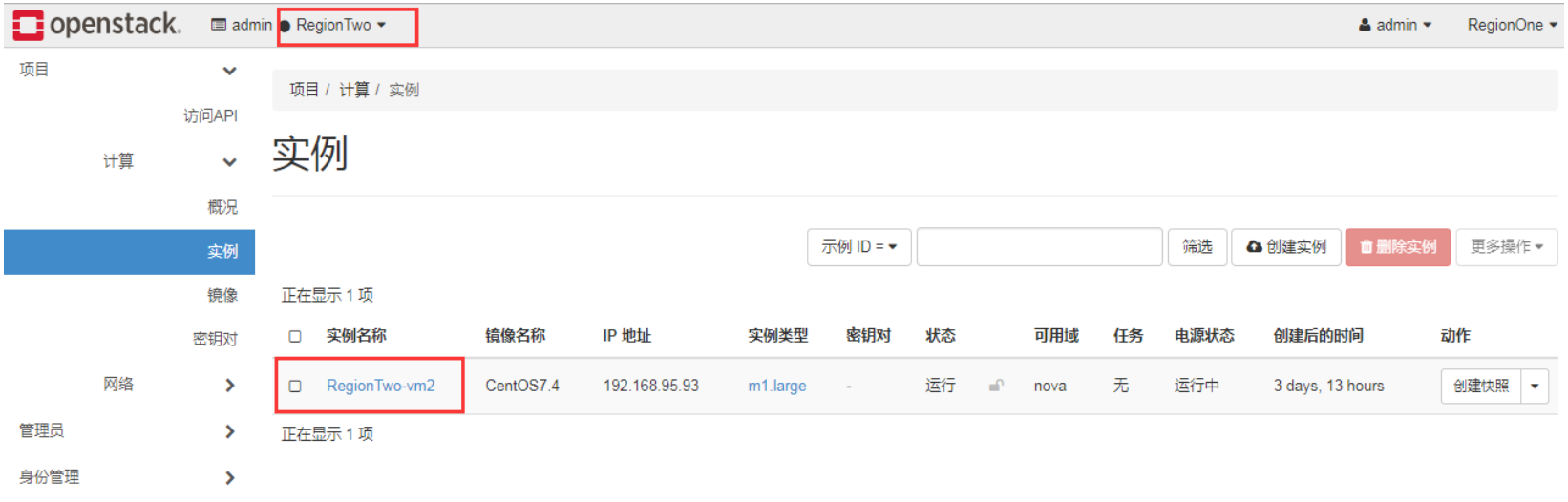
- Horizon
  - 提供web服务
  - Region切换
- Keystone
  - 集中存储所有站点的endpoint
  - HA（pacemaker）
  - 通过VIP提供认证
- MySQL集群
  - 1主2从
  - 保证数据高可靠性
- 集群上所有Service的认证（auth_url，auth_uri）都指向keystone的VIP



已经完成北京和东莞两个站点的部署

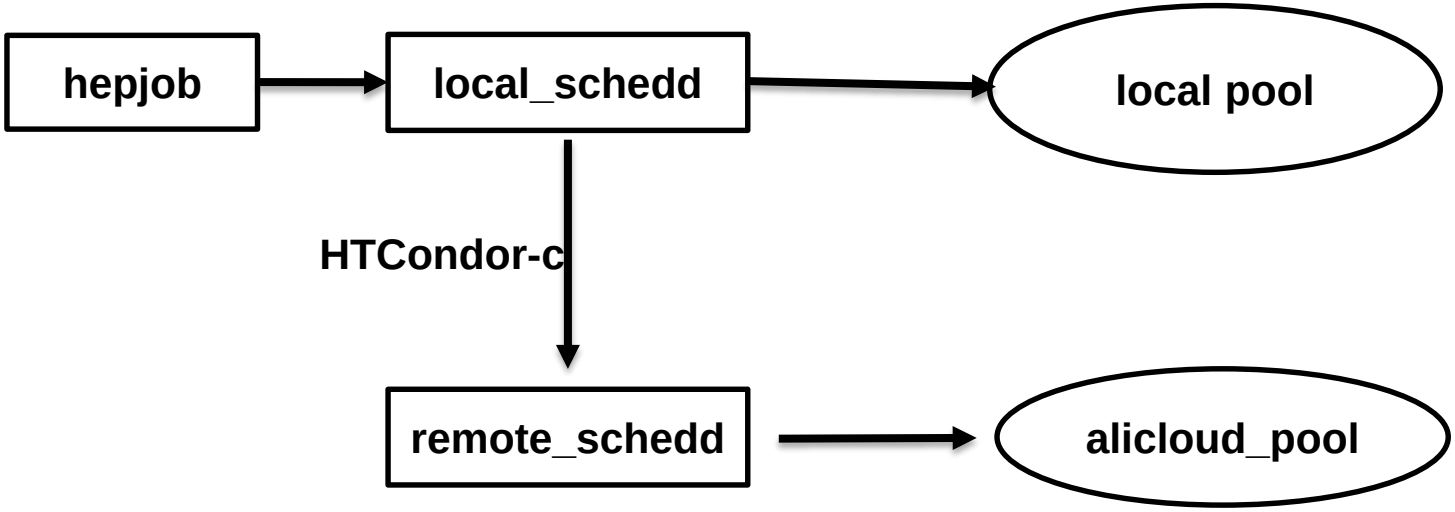
跨域云资源管理界面


openstack
admin
RegionOne
admin
RegionOne
项目
访问API
计算
概况
实例
镜像
密钥对
网络
管理员
身份管理
项目 / 计算 / 实例
实例
示例 ID =
筛选
创建实例
删除实例
更多操作
正在显示 2 项
实例名称 镜像名称 IP 地址 实例类型 密钥对 状态 可用域 任务 电源状态 创建后的时间 动作
RegionOne-vm2 CentOS7.4 192.168.95.95 m1.large - 运行 nova 无 运行中 3 days, 17 hours 创建快照
RegionOne-vm1 CentOS7.4 192.168.95.93 m1.medium - 运行 nova 无 运行中 4 days 创建快照
正在显示 2 项


openstack
admin
RegionTwo
admin
RegionOne
项目
访问API
计算
概况
实例
镜像
密钥对
网络
管理员
身份管理
项目 / 计算 / 实例
实例
示例 ID =
筛选
创建实例
删除实例
更多操作
正在显示 1 项
实例名称 镜像名称 IP 地址 实例类型 密钥对 状态 可用域 任务 电源状态 创建后的时间 动作
RegionTwo-vm2 CentOS7.4 192.168.95.93 m1.large - 运行 nova 无 运行中 3 days, 13 hours 创建快照
正在显示 1 项

阿里云资源的整合

- 已完成对阿里云资源整合的测试
  - 支持LHAASO实验
  - 开发alicloud api对阿里云资源的远程控制
  - 基于HTCondor-C的跨域调度



小结

- 实现跨域资源的整合和共享是缓解高能物理计算资源紧张的有效手段
- 通过虚拟化的技术可以大大节省运维成本
- 基于Openstack的Multi-region是可行的方案，但对异构的云资源（宝德云、商业云等）如何更好的整合还需进一步研究
  - 从作业调度层面进行整合
  - 设计和实现松耦合的资源管理
- 跨域的数据联盟
  - 高能物理的数据量大
  - 数据访问带宽

谢谢大家！

Q&A?