裸机多计算可用区优化架构设计文档

Review 记录信息

时间	**参与人员**	**主要议题**	**后续review计划**

问题描述

目前裸机的多计算可用区及相关调度均由前端实现, 在大规模的场景下, 性能和实时性等不满足要求. 并且无法做到 nova-compute-ironic 组与 ironic-conductor 组的对应, 后端的 nova/ironic 间数据同步可能会出现非预期情景.

方案提议

方案设计

1. ironic-conductor-init 按照节点 label 生成 conductor_group 配置(现有逻辑)
2. ark-ironic 中 nova-compute-ironic statefulset 服务的生成方式与 ironic-conductor 相似, 通过 `.Values.pod.deploy.ironic_conductor.name` 来生成若干个三副本的 statefulset, 命名规则与 ironic-conductor 类似, 形如 `nova-compute-ironic-{conductor_group_name}`
3. nova-compute-ironic-init 获取 pod 名称, 并以中间的 `{conductor_group_name}` 作为配置项 `partition_key` 的值. 同时获取该 statefulset 的副本数, 按照形如 `nova-compute-ironic-{conductor_group_name}-0, nova-compute-ironic-{conductor_group_name}-1` 的形式作为配置项 `peer_list` 的值.
4. 注册/编辑裸金属节点时, 可用区和节点组的选择有所联动
   a. 注册节点到空(无nova-compute-ironic服务)可用区: 允许选择节点组, 但可选择的节点组需要未与可用区对应
   b. 注册节点到非空((有nova-compute-ironic服务))可用区: 不允许选择节点组, 自动切换到可用区对应的节点组上
   c. 编辑节点到空(无nova-compute-ironic服务)可用区: 需要重新选择节点组, 可选择的节点组需要未与可用区对应
   d. 编辑节点到非空((有nova-compute-ironic服务))可用区: 不允许选择节点组, 自动切换到可用区对应的节点组上
5. 可用区管理界面中, 允许向可用区添加 nova-compute-ironic 服务, 并具有相同 `partition_key` 配置(即 conductor_group 名称)的 nova-compute-ironic 服务应当同时只能在一个可用区中.

其它可选方案

方案	**方案描述**	**代价**	**代表厂商**
前端实现	注册裸金属节点时选择预先创建的可用区(可为空), 并保存在节点的 extra_info 中. 创建实例时, 在前端获取所有节点的 extra_info, 解析出对应的可用区列表以供用户选择, 并且按照 flavor/image 的硬件要求过滤出合适的节点, 按照升序排列. 在向后端发送请求时, 同时发送节点列表前列的节点 UUID, 后端通过指定节点创建的方式来创建实例.	1. 前端调度, 大规模的场景下, 性能和实时性等不满足要求 2. 无法做到 nova-compute-ironic 组与 ironic-conductor 组的对应, 后端的 nova/ironic 间数据同步可能会出现非预期情景	EasyStack 目前实现策略

竞品类似方案

华为云

公有云有类似功能, 但无功能实现的说明. 私有云未见类似功能描述


裸金属服务器 B… 户指南.pc
17 6 2022, 02:46 上午


华为云Stack 8.1… 01.pdf
17 6 2022, 02:44 上午

可升级影响

可随云产品进行升级, 待讨论问题见文末

稳定可靠性影响

无影响。

性能影响

无影响

安装部署影响

无影响。

API影响

1. 获取 conductor group 信息的 API 接口需要返回 nova-compute-ironic 服务与 conductor group 的关系
2. 获取裸金属节点信息的 API 接口需要返回对应的可用区信息

前端界面影响

1. 可用区管理界面中, 允许向可用区添加 nova-compute-ironic 服务, 并具有相同 partition_key 配置(即conductor_group 名称)的 nova-compute-ironic 服务应当同时只能在一个可用区中
2. 注册/编辑裸金属节点时, 可用区和节点组的选择有所联动, 见上述方案设计描述

安全性影响

无影响

文档影响

用户手册，API文档，技术白皮书，运维手册内容补充

监控 日志 告警方案

无

License 方案

无

其它影响

无

实现

主要实现成员

王亚，翟元杰

JIRA 任务 ID

测试

功能测试

多可用区+多 conductor group 下调度功能测试

性能测试

并发创建实例的测试

压力测试

无

场景测试

无

组件间依赖

依赖组件：

nova

参考链接

待讨论确认的内容

既有环境的升级:

- 单可用区对应单 conductor group: 将 nova-compute-ironic-{conductor_group} 服务加入到对应的可用区中
- 多可用区对应多 conductor group, 且可用区与 conductor group 一一对应: 将 nova-compute-ironic-{conductor_group} 服务加入到对应的可用区中
- 单可用区对应多 conductor group: 将 nova-compute-ironic-{conductor_group} 服务加入到对应的可用区中
- 多可用区对应单 conductor group: 无法支持该模式, 需要更新 license 以生成多个对应的 conductor group.

- 升级后会创建出对应的 nova-compute-ironic-{conductor_group} 服务, 但是处于运行中的裸金属节点在 nova compute_nodes 表中依然使用原有的 nova-compute-ironic-{conductor_group} 服务, 将导致裸金属实例的重建/删除等操作失败, 对此需要升级时启动一个 job, 用于更新 nova 数据库中 instances 表中裸金属实例的 `host`, `launched_on` 字段
- 升级后创建对应的 job, 将 nova-compute-ironic-{conductor_group} 服务加入到对应的可用区中
- ~~需要修改调度策略~~, 无需修改, 创建裸金属实例, 向 nova 传参时, 传入 `os:scheduler_hints.query` 参数, 其中指定相关的调度策略. 此外需要在 nova ironic driver 中将节点的 CPU 类型/启动方式/商业存储类型同步到 nova 的 compute_nodes 表中存储以便于后面的调度使用.
  - 目前前端的可选择因素:
    - 可用区(无需修改)
    - CPU 架构(无需修改)
    - CPU 类型
    - 启动方式(BIOS/uEFI)
    - Flavor(无需修改)
      - CPU
      - memory
      - disk
    - Image(无需修改)
      - disk
    - 挂载商业存储
    - 指定节点
- 需要修改 nova-scheduler 中的权重计算, 以实现裸金属资源的最小匹配
  - 是否只考虑 CPU/Memory 资源, Disk 资源不予以考虑?
- 现有环境将conductor_group当做网络az使用，比如邮储环境已经存在hx和wl的conductor_group，升级需要考虑兼容
- 当有多个计算可用区时，由于每个可用区需要有一个conductor_group对应，而每个conductor_group对应3副本ironic-conductor和nova-compute-ironic服务，节点资源需求会比较大
- 从计算可用区与conductor_group的对应逻辑来看，计算可用区应该是云产品部署前规划好的，如果需要增加可用区，则需要扩容相应服务，这一点从灵活性上比云主机计算可用区有所限制

修订记录

当前版本	主要修改人	主要签署人	修改时间	描述
v0.1	王亚		2022.6.28	初稿