裸机挂载商业存储数据盘架构设计

- Ironic Attach Volume Spec
- Problem Description
  - Concept
  - Use Cases
- Proposed Change
  - Ironic 内部逻辑
    - 离线挂载
    - 在线挂载
    - 离线卸载
    - 在线卸载
  - 前端逻辑
    - 挂载
    - 卸载
    - 云硬盘链接信息
  - 功能前置条件
  - 功能边界
  - 对其它功能影响
  - 数据安全性分析
  - 异常边界
  - Alternatives
  - API impact
  - Configuration impact
  - Deploy impact
  - Documents impact
  - Upgrade impact
- Implementation
  - Assignee(s)
  - Work Items
- References
- History

Ironic Attach Volume Spec

本文主要是描述裸金属挂载云硬盘(FC-SAN, IP-SAN)功能的设计, 包括 ironic 内部已有功能的逻辑和工作流程, 以及产品化的工作内容及产品边界等.

Problem Description

裸金属实例需要能够灵活的挂载/卸载商业存储卷(FC-SAN, IP-SAN).
关于该功能的前期调研请见 挂载 Volume 调研

Concept

- Volume connector: Ironic 内部概念, 用于描述裸金属节点作为存储客户端, 连接到服务端所需要的信息.
- Volume target: Ironic 内部概念, 用于对应一个 Cinder volume, 并用于描述服务端的连接信息. 可简单理解为挂载云硬盘时需要创建一个对应的 volume target, 卸载云硬盘时需要删除一个对应的 volume target.

Use Cases

- 管理员可以为裸金属节点配置 wwpn, wwnn, iqn 等(作为挂载存储卷的必要配置)
  - wwpn/wwnn 需要管理员手动收集, 格式为形如 `2100000e1e18ad50` 的字符串, 收集方式有:
    - 通过服务器 BMC 界面查询 HBA 设备的信息
    - 通过服务器 BIOS 中的 device manager 界面查询 HBA 设备的信息
    - 通过进入服务器系统, 通过 `systool` 工具进行查询
  - iqn 可由 node uuid 自动生成, 无需手动填写, 自动生成的格式为形如 `iqn.2017-08.org.openstack.4d98a2c0-9086-465e-8b8b-36136f231a86` 的字符串
    - iqn 允许用户定义前缀(需要说明格式), 但是 uuid 需要存在以便于保证唯一性
- 租户可以为裸金属实例挂载/卸载 FC-SAN, IP-SAN volume
  - 可以在裸金属实例列表页/详情页中挂载/卸载
  - 可以在创建裸金属实例时创建并挂载FC-SAN/IP-SAN volume
- 暂不支持 boot from volume

Proposed Change

对于上述的使用场景, 在 ironic-dashboard-api 中引入下面的操作:

- 管理员
  - 创建/编辑/修改/删除: 裸金属节点 volume connector
  - 查询 volume target 信息
- 租户
  - 查询 volume connector 信息
  - 查询 volume 信息列表
  - 查询 volume type 信息列表
  - 查询/创建/删除 volume target 信息: 用于挂载/卸载云硬盘

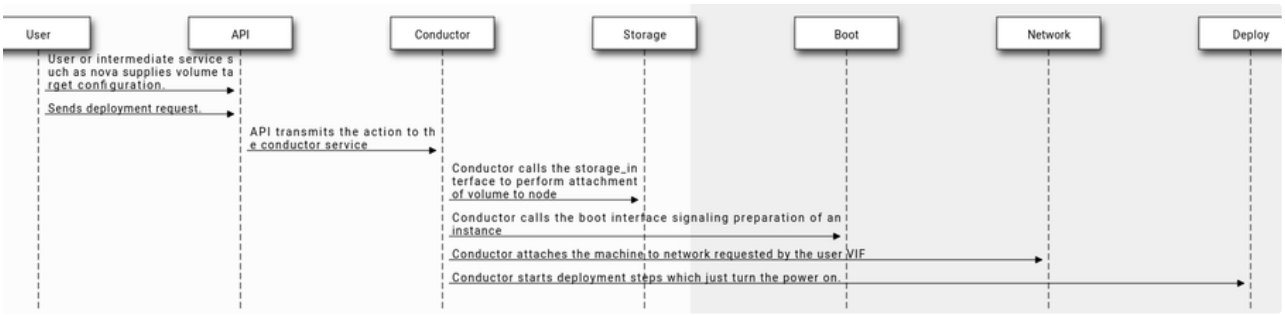
在 ironic 中引入新的 API

- 在线挂载 volume
- 在线卸载 volume

Ironic 内部逻辑

这里主要用于描述 Ironic 内部逻辑, 即管理员/租户如何通过非界面操作进行 Volume 的挂载/卸载

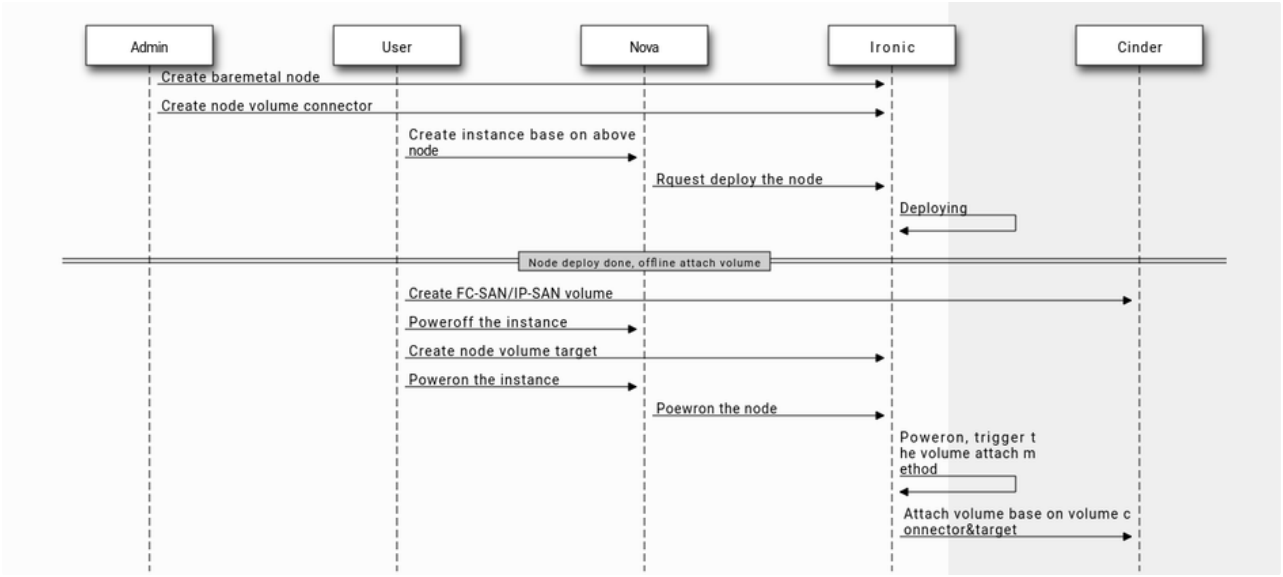
对于挂载 volume, 社区的 ironic 内部的功能实现是 boot-from-volume Boot From Volume — ironic 23.1.1.dev87 documentation :



上图主要是表明 ironic boot-from-volume 的大概流程, 但本文只需要挂载 volume, 不需要从卷启动, 故需将之产品化:

离线挂载

离线挂载, 即裸金属实例的电源为关闭时的挂载操作

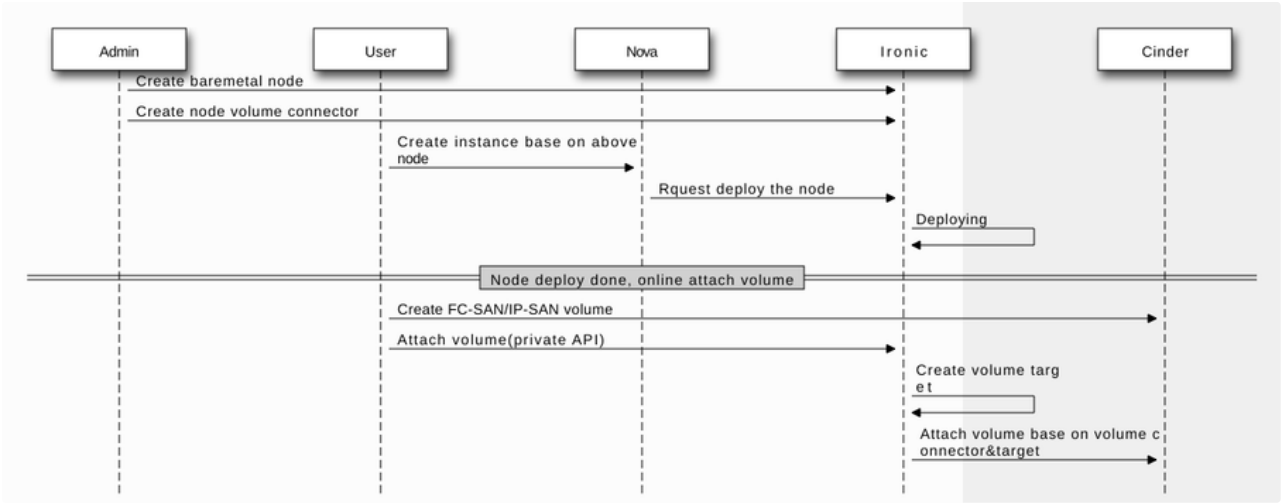


可得离线挂载的流程为:

- 管理员注册裸金属节点, 并为裸金属节点创建 `volume connector`
- 用户从该裸金属节点创建裸金属实例, 且裸金属实例创建完毕
- 用户创建一个 FC-SAN/IP-SAN 卷, 需要与上述的 `volume connector` 类型匹配, 即 `volume connector` 是 FC-SAN, 就创建 FC-SAN 的 volume
- 用户关闭裸金属实例电源
- 用户为该裸金属 **节点** 创建一个 `volume target`, 用于指向上述创建的 volume
- 用户发送启动裸金属实例请求, 用于触发 ironic 内部挂载 volume 的操作
  - 在裸金属节点启动过程中, ironic 会向 cinder 发送请求, 将 volume 挂载到裸金属节点上. 需要注意, 在进入系统后, 用户还需要手动配置并扫描 SAN 块设备

在线挂载

在线挂载, 即裸金属实例的电源为开启时的挂载操作



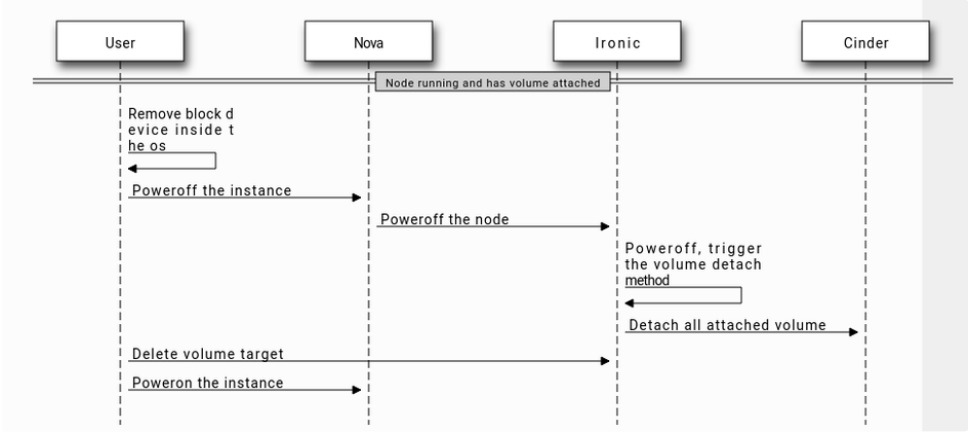
可得在线挂载的流程为:

- 管理员注册裸金属节点, 并为裸金属节点创建 `volume connector`
- 用户从该裸金属节点创建裸金属实例, 且裸金属实例创建完毕
- 用户创建一个 FC-SAN/IP-SAN 卷, 需要与上述的 `volume connector` 类型匹配, 即 `volume connector` 是 FC-SAN, 就创建 FC-SAN 的 volume
- 用户向 ironic 发送挂载 volume 的请求, **注意该 ironic API 为 easystack 新增 API**
- Ironic 向 cinder 发送挂载请求并完成挂载操作
  - 需要注意, 进入系统后, 用户还需要手动配置并扫描 SAN 设备

注意不同于 nova instance 的 volume 的挂载, 即在平台上挂载 volume 后就能在 虚拟机内部看到对应的块设备. 裸金属 Ironic 调用 Cinder 进行 volume 挂载, 从平台角度来看只是建立了裸金属实例和 volume 的逻辑对应关系, 从 SAN 存储来看只是在服务端录入了客户端(即 volume connector)的信息. 用户还需要进入系统内部, 进行客户端的配置和扫描, 才能够 看到对应的块设备.

离线卸载

下面再看一下卸载的流程, 注意这里描述的是离线卸载, 即需要关闭裸金属节点电源:

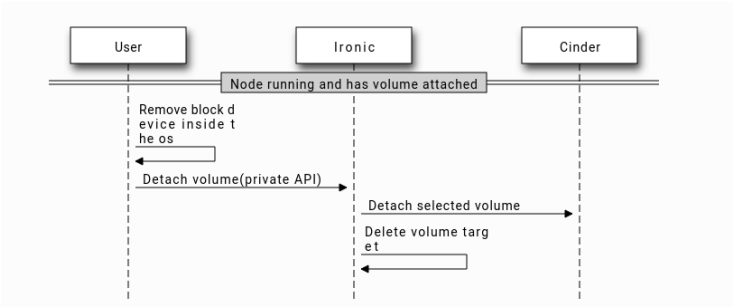


可得离线卸载的流程为:

- 用户在系统中移除对该 SAN 块设备的使用, 例如 `/etc/fstab`
- 用户发送关闭裸金属实例请求, 用于触发 ironic 内部卸载 volume 的操作
- 用户删除 volume 对应的 `volume target`
- 用户发送开启裸金属实例请求

在线卸载

在线卸载, 即裸金属实例的电源为开启时的卸载操作



可得在线卸载流程为:

- 用户在裸金属系统内部卸载 SAN 块设备
- 用户向 ironic 发送卸载 volume 的请求, **注意该 ironic API 为 easystack 新增 API**
- Ironic 向 cinder 发送卸载请求并完成卸载操作

综上所述, 在 ironic 内部, 我们将会新增两个 API, 用于在线挂载/卸载 volume. 对于使用者(即直接使用 ironic 的对象)而言, 可以在裸金属节点电源为关机时, 使用上述的离线挂载/卸载方式, 在裸金属节点电源状态为开机时, 使用上述的在线挂载/卸载方式.

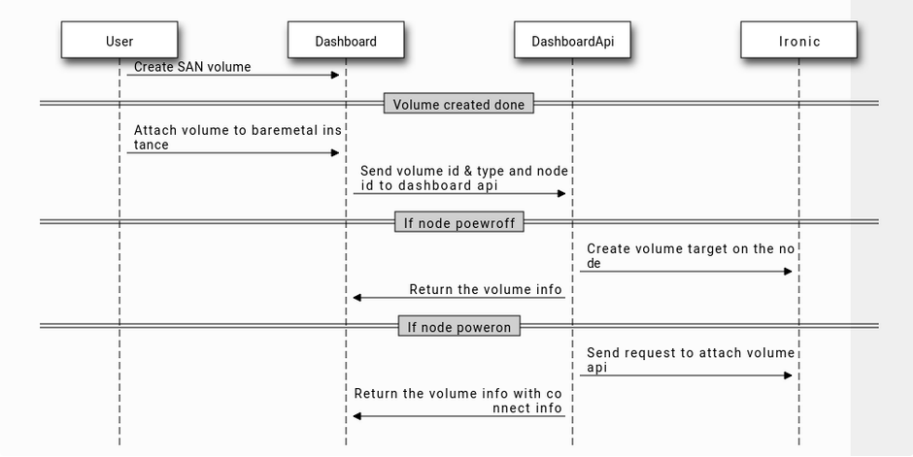
前端逻辑

此处主要是描述 ironic-dashboard-api, ironic-dashboard 上的操作逻辑

- 前端获取可挂载的 volume 列表时, 需要根据节点 `volume connector` 筛选合适的 volume 并返回
  - 从 roller 中获取商业存储的 `backend name` && `protocol type`
  - 根据上述信息可筛选出对应的 `volume type`
  - 根据 `volume type` 筛选出合适的 volume

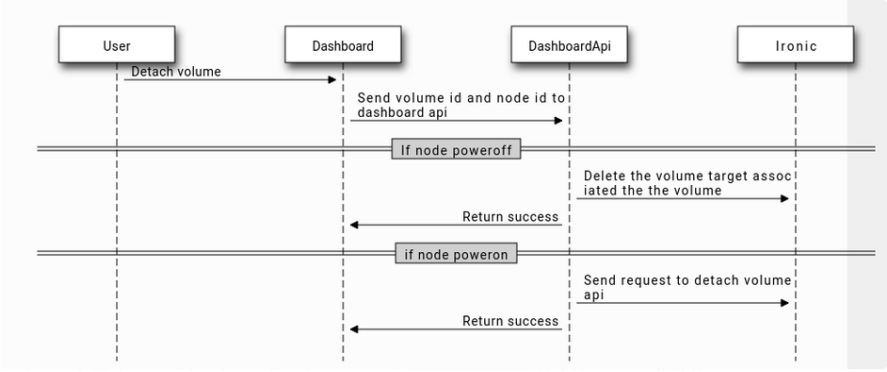
挂载

假定实例已经创建, 那么对于挂载:



卸载

对于卸载:



从上面两段描述可以看出, 对于用户而言, volume 与 `volume target` 基本等价:

- 当用户尝试挂载 volume 时, `volume target` 由 ironic-dashboard-api 自动创建, 并将 `volume target` 中的连接信息及volume详情返回到 ironic-dashboard
- 当用户尝试卸载 volume 时, `volume target` 由 ironic-dashboard-api 自动删除

云硬盘链接信息

`ironic-dashboard-api` 返回的 `volume target` 信息中, 将会包含链接相关的信息, 用于描述裸金属实例系统如何连接到存储服务端.

- 对于 IP-SAN, 该信息应当类似于

```
    {
      "target_discovered": false,
      "encrypted": false,
      "ironic_volume_uuid": "1cc36994-56a3-4d68-a717-179fa877d9b1",
      "qos_specs": null,
      "target_iqn": "iqn.1986-03.com.ibm:2145.v3700.node1",
      "target_portal": "192.168.110.71:3260",
      "volume_id": "1cc36994-56a3-4d68-a717-179fa877d9b1",
      "target_lun": 0,
      "access_mode": "rw"
    }
```

- 对于 FC-SAN, 该信息应当类似于

```
    {
      "initiator_target_map": {
        "2100000e1e18ad50": [
          "500507680306FBAC",
          "500507680306FBAD"
        ]
      },
      "target_discovered": false,
      "encrypted": false,
      "ironic_volume_uuid": "687419d1-1625-4fe1-97bd-1ce3a51d0ee8",
      "qos_specs": null,
      "volume_id": "687419d1-1625-4fe1-97bd-1ce3a51d0ee8",
      "target_lun": 0,
      "access_mode": "rw",
      "target_wwn": [
        "500507680306FBAC",
        "500507680306FBAD"
      ]
    }
```

将这两个信息合并, 可以得到三个需要展示给用户的条目:

- `target_id` : 对应 FC-SAN 中的 `initiator_target_map` 键, 对应 IP-SAN 中的 `target_iqn` . 用于描述服务端的 ID
- `target_portal` : 对应 FC-SAN 中的 `initiator_target_map` 键, 对应 IP-SAN 中的 `target_portal` . 用于描述服务端的登录地址
- `target_lun` : 对应 FC-SAN 及 IP-SAN 中的 `target_lun` . 用于描述该 volume 在服务端的 lun id

功能前置条件

- 基于 foundation 602 + cinder 602 + nova 604
- FC-SAN/IP-SAN 存储被平台 cinder 服务接管

- 对于 FC-SAN, 需要提前知晓裸金属节点 HBA 卡的 WWPN/WWNN
- 对于 IP-SAN, 裸金属节点需要有网络能够与 IP-SAN 通信
  - 用户手册说明

功能边界

- 支持主流 Linux/Windows 系统, 需要系统镜像内置 HBA 卡驱动及 iSCSI initiator 程序
  - CentOS
  - Kylin
  - Redflag
  - Windows server
- 支持在线挂载/卸载云硬盘
- 支持创建实例时挂载云硬盘（依赖计算云产品升级）
- multipath: 用户手册说明配置方式
- **测试验证**
  - 云硬盘在线扩容
  - 云硬盘更新状态
  - 云硬盘重置挂载状态
  - 云硬盘创建快照

云硬盘连接器操作矩阵:

裸金属节点\操作	查询	创建	更新	删除	备注
状态可用, 电源关机	Y	Y	Y	Y	
状态可用, 电源开机	Y	Y	N	N	
状态运行中, 电源开机	Y	Y	N	N	
状态运行中, 电源关机	Y	Y	Y	Y	
状态部署中, 电源开机或关机	Y	N	N	N	
状态清理中, 电源开机或关机	Y	N	N	N	
状态故障, 电源开机或关机	Y	N	N	N	
无状态或其它状态	Y	N	N	N	

裸金属主机挂载/卸载云硬盘操作矩阵:

裸金属主机\操作	挂载	卸载	备注
状态运行中	Y	Y	
状态关机	Y	Y	
状态创建中	N	N	
状态正在开机或正在关机	N	N	
状态错误	N	N	

其它状态	N	N	

对其它功能影响

- 开启裸金属实例电源: Ironic 会在节点上电时挂载 `volume target` 对应的 volume, 故需要保证对应的 cinder volume 存在且可用
- 创建实例时挂载云硬盘或重建有云硬盘的裸金属实例: Ironic 会在节点上电时挂载 `volume target` 对应的 volume, 故 ironic-python-agent 需要在部署系统前识别并剔除 SAN 云硬盘
- 删除裸金属实例时会将已挂载的云硬盘自动卸载. 即使选择了清理, 云硬盘也不会进入清理流程
- 关机裸金属实例时, ironic 会将已挂载的云硬盘卸载; 开启一个关机的裸金属实例时, ironic 会将云硬盘挂载上
- 裸金属主机重启时不会卸载云硬盘
- 裸金属节点重置状态时，会将已挂载的云硬盘卸载
- 裸金属主机关机再开机, 或重启后, 是否需要系统内重新扫盘取决于系统配置, 比如对于 FC-SAN 存储, linux 内核驱动会在系统启动时自动扫盘; 对于 IP-SAN 存储, 则需要配置 iSCSI 客户端开机自启动进行扫盘(此外还需要保证开机时网络配置服务自启, 以保证网络畅通)。
- 云硬盘界面现状：云硬盘挂载弹窗显示挂载到云主机，并且云主机列表中不包含裸金属主机；已挂载到裸金属主机的硬盘卸载弹窗里显示云主机，并且可以点击卸载操作，但卸载操作不成功。如果要支持云硬盘界面给裸金属主机挂载和卸载云硬盘，需要cinder云产品进行适配开发。

数据安全性分析

卸载云硬盘：

1、裸金属主机运行中卸载云硬盘

可能出现云硬盘正在IO状态导致云硬盘卸载失败，正在写入数据落盘失败，再次卸载成功不会对已存储数据造成影响

2、裸金属主机/节点关机时卸载云硬盘

卸载成功与否不会对已存储数据造成影响

3、裸金属主机关机状态挂载并卸载云硬盘

挂载/卸载成功与否不会对已存储数据造成影响

4、裸金属主机删除时卸载云硬盘

卸载成功与否不会对已存储数据造成影响

5、裸金属主机重建时卸载云硬盘

卸载成功与否不会对已存储数据造成影响

6、裸金属主机创建失败时卸载云硬盘

卸载成功与否不会对已存储数据造成影响

7、裸金属节点重置状态时卸载云硬盘

卸载成功与否不会对已存储数据造成影响

删除云硬盘：

裸金属侧不会触发删除云硬盘操作

清理云硬盘：

裸金属侧不会触发清理云硬盘操作

异常边界

1、创建裸金属主机时挂盘失败，会导致裸金属主机创建失败，裸金属主机进入错误状态，裸金属节点进入故障状态，创建的云硬盘不受影响

恢复方法：可以删除裸金属主机，看裸金属节点是否恢复可用状态，如果没有恢复，可以重置节点状态恢复。

2、裸金属主机开机时挂载盘失败，会导致开机失败，裸金属主机保持关机状态，云硬盘不受影响

恢复方法：处理完挂载失败问题，再次开机。

3、裸金属主机关机时卸载盘失败，不会导致关机失败，只会有warning日志，裸金属主机进入关机状态，云硬盘保持挂载状态

恢复方法：手动卸载云硬盘。

4、裸金属主机删除时卸载盘失败，裸金属主机会删除，裸金属节点进入故障状态，云硬盘不受影响

恢复方法：重置节点状态，会卸载云硬盘，云硬盘恢复未挂载状态。

5、重建裸金属主机时挂盘失败，会导致裸金属主机重建失败，进入错误状态，裸金属节点进入故障状态，云硬盘不受影响

恢复方法：可以删除裸金属主机，看裸金属节点是否恢复可用状态，如果没有恢复，可以重置节点状态恢复。

6、重置节点状态时卸载盘失败，会导致重置节点状态失败，云硬盘不受影响

恢复方法：处理完卸载失败问题，再次重置节点状态。

7、裸金属服务器更换商业存储类型，更换hba卡或IP-SAN网卡

恢复方法：如果裸金属节点上没有裸金属主机，步骤为：1、裸金属节点关机；2、更换设备；3、更新云硬盘连接器信息。如果裸金属节点上有裸金属主机，步骤为：1、裸金属主机关机（期间会自动卸载云硬盘）；2、更换设备；3、更新云硬盘连接器信息；4、裸金属主机开机（期间会自动挂载云硬盘）。

8、挂载云硬盘的裸金属主机，服务器直接关机或掉电，挂载的云硬盘不会卸载

恢复方法：服务器开机或上电进入运行状态，即可恢复正常使用。

Alternatives

可通过智能网卡进行块设备的挂载/卸载, 但考虑到设备的专用性, 在目前版本暂不予已支持.

API impact

详细 API 文档见:

- 裸金属v6.1.1 API文档
- 裸金属v6.1.1 Django API文档

`volume connector` 的操作与裸金属节点状态的矩阵见文档 裸机支持挂载云硬盘需求文档

Configuration impact

Ironic 服务需要添加如下默认配置项::

```
[DEFAULT]
default_storage_interface = cinder
enabled_storage_interfaces = cinder,noop

[cinder]
auth_type = password
auth_url = http://keystone-api.openstack.svc.cluster.local:80/v3
password = BBDwDvfk
project_domain_name = Default
project_name = service
region_name = RegionOne
user_domain_name = Default
username = cinder
```

Deploy impact

None

Documents impact

- 管理员手册中需要指出界面功能使用
- 管理员手册中需要指出 HBA wwpn 收集方式
- 用户手册中需要指出界面功能使用
- 用户手册中需要指出如何在 OS 中扫盘&挂载
  - 需要指出挂载了 SAN 块存储后, 原盘符名称可能会在重启后有变化
- 用户手册中需要指出在线卸载云盘的安全流程

Upgrade impact

- 603保持一台裸机仅支持挂载一种类型商业存储的限制(即仅支持 FC-SAN 或仅支持 IP-SAN)，后面版本可以通过升级放开限制

  处理办法：通过修改代码逻辑放开限制，升级生效。
- 存量裸金属节点如何添加云硬盘连接器

  处理办法：存量裸金属节点可以创建云硬盘连接器，并通过界面操作挂载商业存储盘，裸金属节点规模较大时，通过后台脚本创建云硬盘连接器。
- 存量裸机系统里如何装入并配置hba卡驱动，multipath 多路径软件，iSCSI 客户端软件

  处理办法：对于已运行业务裸机，可以给裸金属节点创建云硬盘连接器，并在裸机系统内手动安装和配置驱动和软件；对于无业务运行裸机，给节点创建云硬盘连接器，并用支持的裸机镜像重建裸机。
- 平台的裸金属部署镜像如何升级

  处理办法：通过云产品升级。
- 存量裸机手动挂载的商存盘如何纳管

  处理办法：不能纳管。

Implementation

Assignee(s)

Primary assignee:

ya.wang@easystack.cn

Work Items

References

- 挂载 Volume 调研
- Boot From Volume — ironic 23.1.1.dev87 documentation
- EAS-85553: [API文档]裸机支持挂载商业存储数据盘 已完成

History

.. list-table:: Revisions

:header-rows: 1

-
  - Release Name
  - Description
-
  - 6.0.3
  - Introduced