裸机管理

需求概要	ECS 提供裸机注册、纳管、部署操作系统功能
部署方式	• license激活； • 引入新组件（例如：裸金属服务/数据库服务/...）；
涉及产品线	ECS、ECAS
产品版本	6.0.1
文档状态	Draft
产品负责人	@李向军
界面设计师	N/A
技术架构负责人	
后端负责人	@于尚斌
前端负责人	
测试负责人	@Ma Jie (Unlicensed)
JIRA Link	EAS-17736: 裸金属云产品v6.0.1开发 IMPLEMENTED
UE Link	http://4467tp.axshare.com/#id=40dvvk&p=%E5%AF%BC%E8%88%AA&g=1
UI Link	https://app.mockplus.cn/app/Da2VujCBam_/specs/design/wNxfbiyzZC-J
前置需求文档(如有)	对于功能优化类型的PRD文档，需列出前置PRD文档的Link

目录

修订记录

修改日期	修改内容

2018.12.10	v1.0，后续参考本版本撰写PRD文档

一、需求背景

现状

EasyStack目前的V5产品为加入裸机管理相应功能。

客户

客户名称	规模	销售	时间要求
铁科院	10+	丁红	
郑州商品交易所	100+	李冰	
邮政集团			

行业

金融、证券、超算中心、基因测序、企业客户（关键业务系统）

二、价值综述

客户价值

- 运维轻量化：从手工运维中解放，提升精确性和时效性；
- 稳定运行：为客户提供专属裸金属节点，独享物理服务器的稳定性能；
- 安全可靠：裸金属服务器是用户专属的计算资源，支持网络、租户隔离；

商业价值

- 赢单法宝：此功能可以在竞标的时候致胜概率增加100%；
- 云服务和解决方案快速集成：裸金属服务器基于统一的网络模型，支持自定义机型；

拓宽销售行业

三、名词定义

描述功能设计的核心产品概念和名词定义

； Execution Environment (PXE)

PXE is part of the Wired for Management (WfM) specification developed by Intel and Microsoft. The PXE enables system’s BIOS and network interface card (NIC) to bootstrap a computer from the network in place of a disk. Bootstrapping is the process by which a system loads the OS into local memory so that it can be executed by the processor. This capability of allowing a system to boot over a network simplifies server deployment and server management for administrators.

Dynamic Host Configuration Protocol (DHCP)

DHCP is a standardized networking protocol used on Internet Protocol (IP) networks for dynamically distributing network configuration parameters, such as IP addresses for interfaces and services. Using PXE, the BIOS uses DHCP to obtain an IP address for the network interface and to locate the server that stores the network bootstrap program (NBP).

Network Bootstrap Program (NBP)

NBP is equivalent to GRUB (GRand Unified Bootloader) or LILO (LInux LOader) - loaders which are traditionally used in local booting. Like the boot program in a hard drive environment, the NBP is responsible for loading the OS kernel into memory so that the OS can be bootstrapped over a network.

Trivial File Transfer Protocol (TFTP)

TFTP is a simple file transfer protocol that is generally used for automated transfer of configuration or boot files between machines in a local environment. In a PXE environment, TFTP is used to download NBP over the network using information from the DHCP server.

Intelligent Platform Management Interface (IPMI)

IPMI is a standardized computer system interface used by system administrators for out-of-band management of computer systems and monitoring of their operation. It is a method to manage systems that may be unresponsive or powered off by using only a network connection to the hardware rather than to an operating system.

裸金属节点：为裸金属主机提供计算存储能力的物理服务器，一台裸金属节点只可承载一台裸金属主机，对应计算节点。

裸金属主机：直接对用户提供计算存储能力，多用来承载大数据或数据库相关业务，可理解为构建在裸金属节点上的“云主机”；

四、使用场景

*描述用户的使用场景，需要用“拟人化”的语言去表达，要包含的干系人有参与者和使用者。

- 场景一：云管理员和普通用户可以创建、修改、删除、查询裸金属主机
- 场景二：云管理员可以配置裸金属网络方案
- 场景三：云管理员可以创建裸金属主机规格、编辑规格访问控制
- 场景四：云管理员可以注册、编辑、操作裸金属节点

五、竞品分析

可进化

社区现状

服务Release Notes

Current

- Adds support for the Intel IPMI Hardware with hardware type `intel-ipmitool`. This hardware type is same as `ipmi` hardware type with additional support of Intel Speed Select Performance Profile Technology <https://www.intel.com/content/www/us/en/architecture-and-technology/speed-select-technology-article.html>. It uses management interface `intel-ipmitool`. It supports setting the desired configuration level for Intel SST-PP.
- Adds sensor data collector to `redfish` management interface. Temperature, power, cooling and drive health metrics are collected.
- Adds option `allow_deleting_available_nodes` to control whether nodes in state `available` should be deletable (which is and stays the default).

-

Stein

- Adds option `[ansible]default_python_interpreter` to choose the python interpreter that ansible uses on managed machines.
- Adds capability to control the persistency of boot order changes during instance deployment via (i)PXE on a per-node level.
- Adds API version 1.50 which allows for the storage of an `owner` field on node objects.
- Add a new field pxe_template that can be set at driver-info level.
- Adds a new hardware type `ibmc` for HUAWEI 2288H V5, CH121 V5 series servers.

Rocky

- The new ironic configuration setting `[deploy]/default_boot_mode` allows the operator to set the default boot mode when ironic can't pick boot mode automatically based on node configuration, hardware capabilities, or bare-metal machine configuration.
- Adds `external` storage interface which is short for "externally managed". This adds logic to allow the Bare Metal service to identify when a BFV scenario is being requested based upon the configuration set for `volume targets`.

Quene

- Support for routed networks when using the `flat` network interface.
- Adds support for rescuing and unrescuing nodes
- New `xclarity` hardware type for managing Lenovo server hardware.

Pike

- Adds support for volume connectors and volume targets with new API endpoints `/v1/volume/connectors` and `/v1/volume/targets`.
- Adds possibility to attach/detach VIFs to/from active nodes.

Ocata

- Adds support for attaching and detaching network VIFs to ironic ports and port groups by using the `/v1/nodes/<node>/vifs` API endpoint that was added in API version 1.28. When attaching a VIF to a node, it is attached to the first free port group. A port group is considered free if it has no VIFs attached to any of its ports. Otherwise, only the unattached ports of this port group are available for attachment. If there are no free port groups, the first available port is used instead, where ports with `pxe_enabled` set to `True` have higher priority.
- Adds support for port groups with a new endpoint `/v1/portgroups/`.
- Adds new methods to network interfaces, which will become mandatory in Pike release:
  - `vif_list`: List attached VIF IDs for a node.
  - `vif_attach`: Attach a virtual network interface to a node.
  - `vif_detach`: Detach a virtual network interface from a node.
  - `port_changed`: Handle any actions required when a port changes.
  - `portgroup_changed`: Handle any actions required when a port group changes.
  - `get_current_vif`: Return VIF ID attached to port or port group object.
- Hardware types available in this release are:
  - `ipmi` for IPMI-compatible hardware. This type is enabled by default. Uses the `ipmitool` utility under the hood, similar to existing classic drivers `pxe_ipmitool` and `agent_ipmitool`. Supports both types of serial console: via `shellinabox` and via `socat`, both are disabled by default.
- 增加了Notification（创建事件）
- Adds support to deploy to nodes with different CPU architectures from a single conductor.
-

Newton

- 支持PXE over InfiniBand
- Added configdrive support for whole disk images for iSCSI based deploy.
- Adds out-of-band RAID management to DRAC driver using the generic RAID interface
- Exposes the `local_link_connection` and `pxe_enabled` properties of the Port resource to the REST API, raising the API maximum version to 1.19.
- A network interface is set for a node by setting the `network_interface` field for the node via the REST API. This field is available in API version 1.20 and above. Changing the network interface may only be done in the `enroll`, `inspecting`, and `manageable` states.
- Addition of the provision state target verb of `adopt` which allows an operator to move a node into an `active` state from `manageable` state, without performing a deployment operation on the node
- IPA supported iSCSI portal port customization already.
- This adds the reboot_requested option for in-band cleaning.
-

节点

- Create Node
- List Nodes
- List Nodes Detailed
- Show Node Details
- Update Node
- Delete Node

节点管理

- Validate Node
- Set Maintenance Flag
- Clear Maintenance Flag
- Set Boot Device
- Get Boot Device
- Get Supported Boot Device
- Inject NMI
- Node State Summary
- Change Node Power State
- Change Node Provision State
- Set RAID Config
- Get Console
- Start/Stop Console

虚拟网卡

- List attached VIFs of a Node
- Attach a VIF to a node
- Detach VIF from a node

网卡组

- List Portgroups

- Create Portgroup
- List Detailed Portgroups
- Show Portgroup Details
- Update a Portgroup
- Delete Portgroup

网卡

- List Ports
- Create Port
- List Detailed Ports
- Show Port Details
- Update a Port
- Delete Port

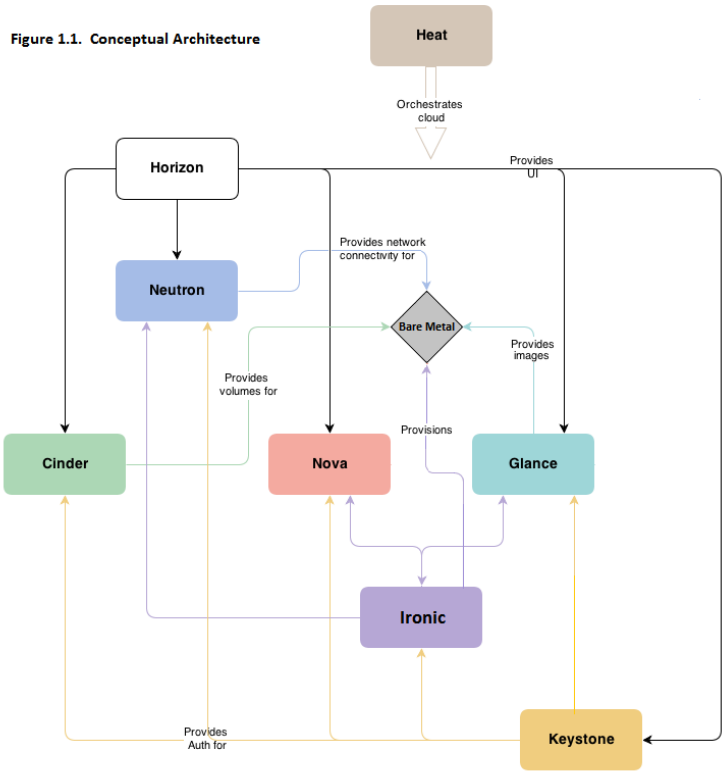
云硬盘

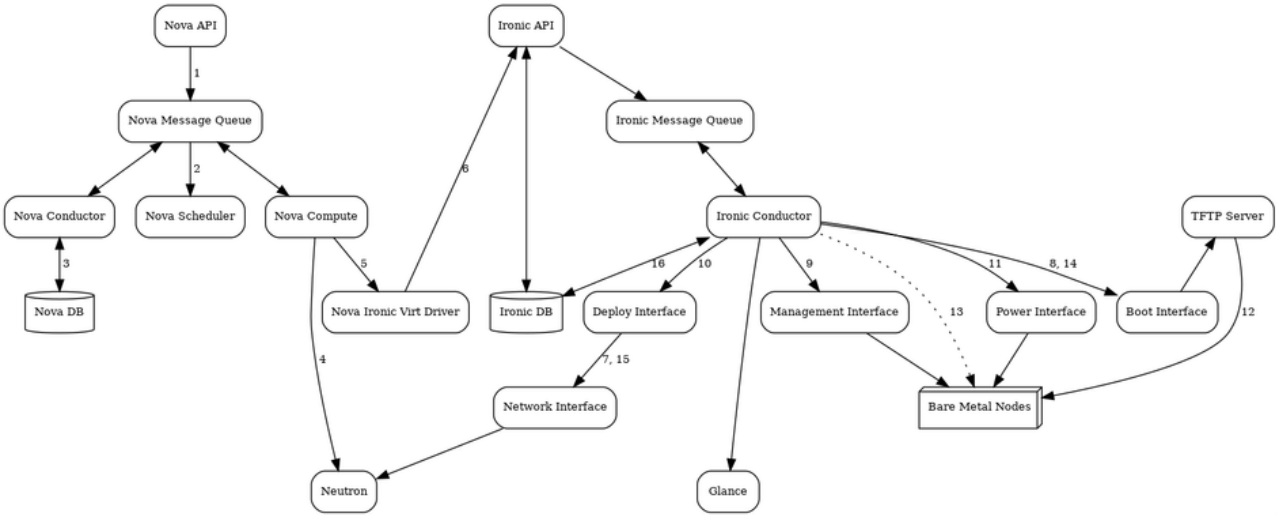
- List Links of Volume Resources
- List Volume Connectors
- Create Volume Connector
- Show Volume Connector Details
- Delete Volume Connector
- List Volume Targets
- Create Volume Target
- Show Volume Target Details
- Update a Volume Target
- Delete Volume Target

宣称支持的功能场景

- 节点硬件检测
- 节点纳管
- Raid 配置
- Bios配置
- Boot from Volume
- 多租户网络
- 端口组
- VNC
- 审计
- 安全
- WIndows镜像

部署流程





Horizon Ironic UI

Queens

- 选择动态驱动
- `Inject NMI` 支持注入NMI

Pike

- 支持开机、关机、访问VNC、支持激活/禁用 VNC（console 支持多种类型）
- 支持端口组与裸机节点的绑定关系，与端口组的管理
- 支持查看PXE状态

Ocata

- 显示节点可管理状态与清理状态
- 可管理端口是否处于PXE状态
- 可在节点初始化后更改网卡

Newton

- 更新节点状态，包括enroll，manageable，available与active
- 增加、删除节点
- 增加、删除网卡

IPv6 与 Ironic 结合

- PXE Boot

IPv6 for PXE 只支持在UEFI模式下

http://www.intel.com/support/network/sb/CS-028553.htm

- Neutron

Neutron 支持双栈，DHCP 支持双栈需要额外工作

竞品分析

AWS

裸机实例基于 Nitro 系统、一系列由 AWS 构建的硬件卸载和服务器保护组件而构建，这些组件相结合可安全地向 EC2 实例提供高性能网络和存储资源。裸机实例上的工作负载可继续利用 AWS 云的所有综合服务和功能，例如 Amazon Elastic Block Store (EBS)、Elastic Load Balancer (ELB) 和 Amazon Virtual Private Cloud (VPC)。

裸机 I3 实例自 2017 年 8 月开始投产，可为 VMware 和 AWS 经过 18 个月的联合开发和测试后发布的 VMware Cloud on AWS 服务提供支持。

Nitro Components（how nitro works https://www.youtube.com/watch?v=o9_4uGvbvnk）

The following components are part of the Nitro system:

- Nitro hypervisor - A lightweight hypervisor that manages memory and CPU allocation and delivers performance that is indistinguishable from bare metal for most workloads.
- Nitro card
  - Local NVMe storage volumes
  - Networking hardware support
  - Management
  - Monitoring
  - Security
- Nitro security chip, integrated into the motherboard

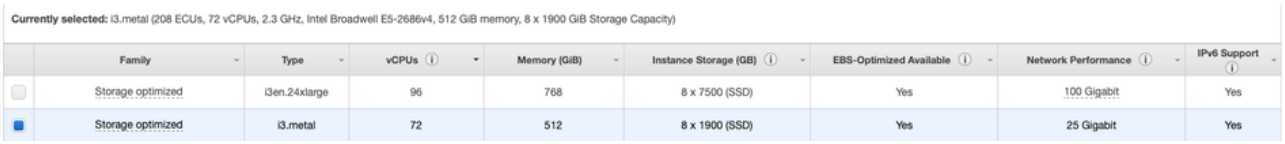
i3.metal 实例

`i3.metal` instances provide your applications with direct access to physical resources of the host server, such as processors and memory. These instances are well suited for the following:

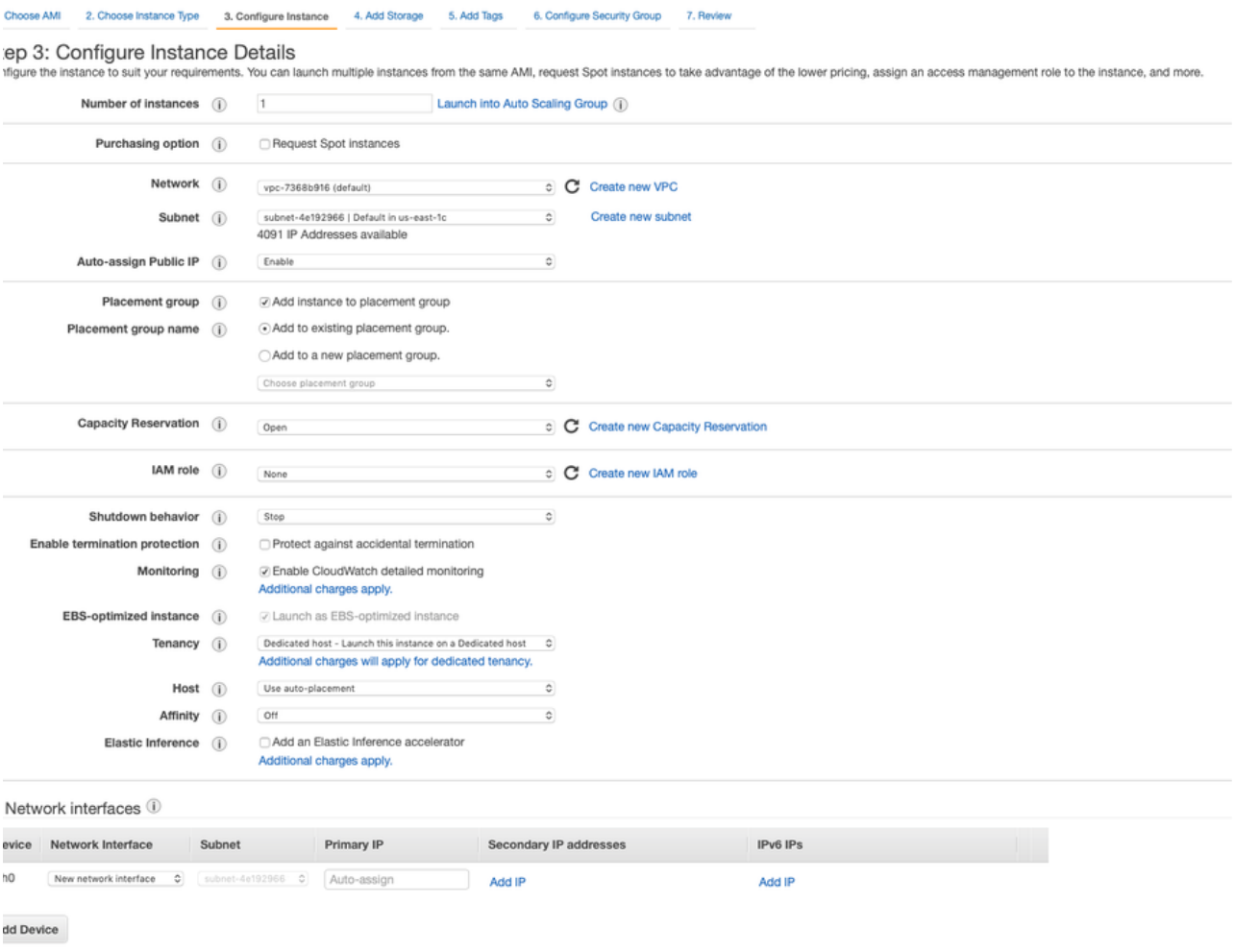
- Workloads that require access to low-level hardware features (for example, Intel VT) that are not available or fully supported in virtualized environments
- Applications that require a non-virtualized environment for licensing or support
- **Processing** – Two Intel Xeon E5-2686 v4 processors running at 2.3 GHz, with a total of 36 hyperthreaded cores (72 logical processors).
- **Memory** – 512 GiB.
- **Storage** – 15.2 terabytes of local, SSD-based NVMe storage.
- **Network** – 25 Gbps of ENA-based enhanced networking.

The five new bare metal instances are m5.metal, m5d.metal, r5.metal, r5d.metal, and z1d.metal. M5 instances offers a balance of compute, memory, and networking resources for a broad range of workloads including web and application servers, back-end servers for enterprise applications, gaming servers, caching fleets, and app development environments. R5 instances are well suited for memory intensive applications such as high performance databases, distributed web scale in-memory caches, mid-size in-memory databases, real time big data analytics, and other enterprise applications. In addition, M5d and R5d instances have local storage, offering up to 3.6 TB of local NVMe-based SSDs. z1d instances provide both high compute performance and high memory, which is ideal for gaming, and certain relational database workloads with high per-core licensing costs.

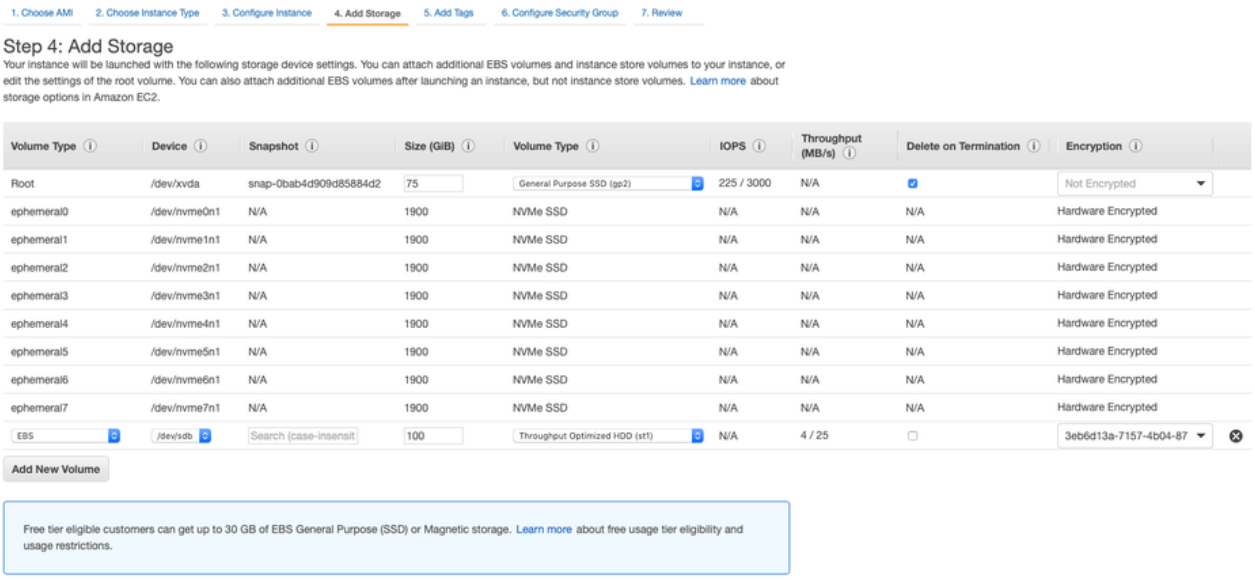
Flavor 指定对应裸机资源



完整的网络、预留资源、监控、网卡功能



可挂载EBS存储



QingCloud

青云 QingCloud 物理主机服务，提供高性能、资源独享、安全隔离的专属物理主机群组，满足各类核心应用对高性能及稳定性的需求，同时提供完整的设备管理权限及运维服务。 用户可以像使用其他云资源一样，快速、灵活的部署及管理物理主机，并可按需弹性购买。

物理主机特性：

1、承载核心业务。

独享物理隔离的硬件资源，释放极致性能，匹配 VPC 网络赋予的完整控制能力， 满足核心业务对性能、可靠性及安全合规的苛刻需求。

2、高效的资源交付。

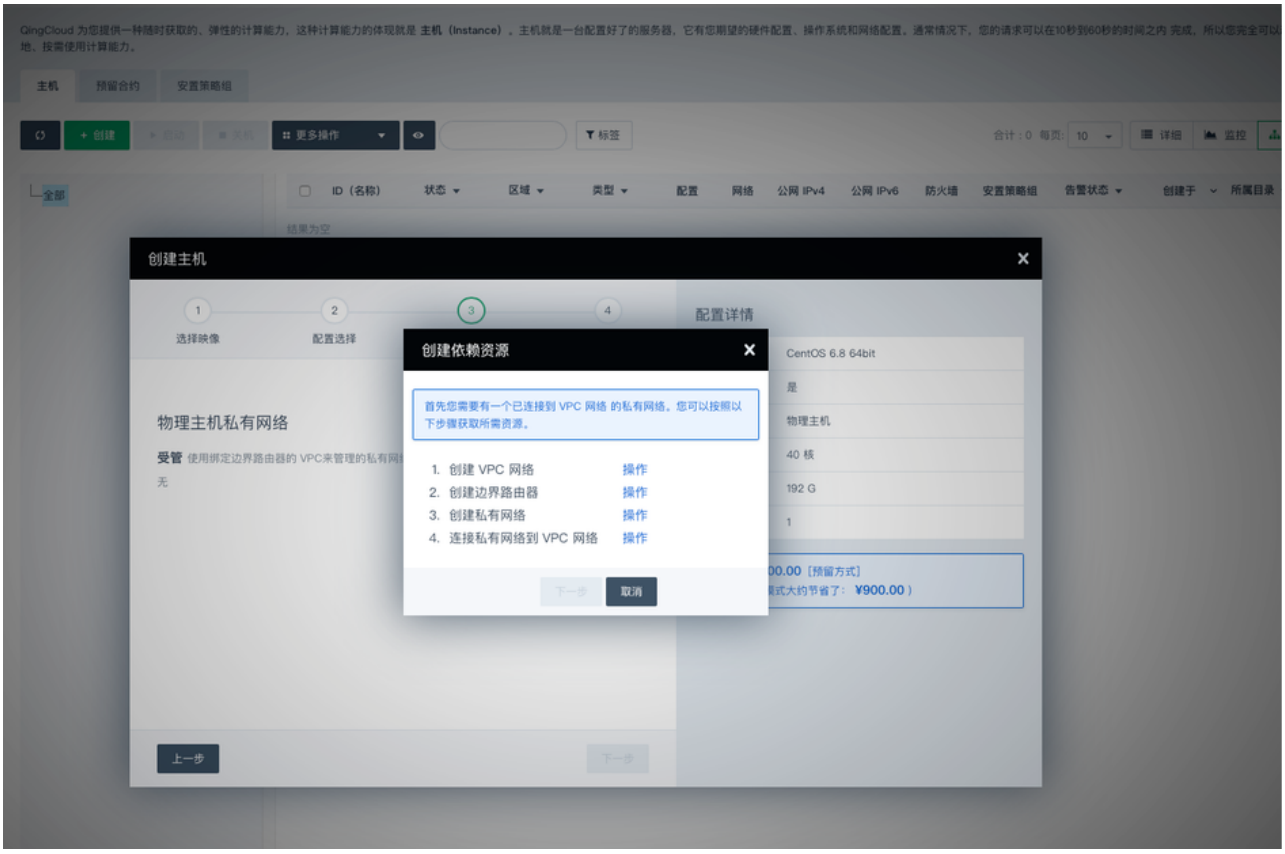
10 分钟完成 OS 部署上线，自动化配置，秒级操作响应，兼容所有虚拟主机系统镜像及用户自有镜像，助您轻松完成基础架构部署。

3、灵活简便的使用体验

VPC 直连，提供与虚拟主机（VM）一致的使用方式，支持弹性伸缩、按需秒级计费及 API 自动化管理，赋予物理主机弹性、灵活与敏捷的云端特性。

> 青云使用 zabbix 来收集物理主机的监控信息，包括 CPU 使用率、内存使用率、硬盘使用率等。 在创建主机时，青云会在物理主机上部署 zabbix agent。 创建成功后，青云会在物理主机所属的路由器上，通过 zabbix_get 收集物理主机监控信息。

物理主机可使用VPC能力



Aliyun

弹性裸金属服务器（ECS Bare Metal Instance）是一款同时兼具虚拟机弹性和物理机性能及特性的新型计算类产品，是基于阿里云完全自主研发的下一代虚拟化技术而打造的新型计算类服务器产品。与上一代虚拟化技术相比，下一代虚拟化技术的主要创新在于，不仅支持普通虚拟云服务器，而且全面支持嵌套虚拟化技术，保留了普通云服务器的资源弹性，并借助嵌套虚拟化技术保留了物理机的体验。

弹性裸金属服务器融合了物理机与云服务器的各自优势，实现超强超稳的计算能力。通过采用阿里云自主研发的虚拟化2.0技术，您的业务应用可以直接访问弹性裸金属服务器的处理器和内存，无任何虚拟化开销。弹性裸金属服务器具备物理机级别的完整处理器特性（例如，intel VT-x），以及物理机级别的资源隔离优势，特别适合上云部署传统非虚拟化场景的应用。

弹性裸金属服务器通过自研芯片和自研Hypervisor系统软件以及重新定义服务器硬件架构等软硬件和芯片技术，打造了全球领先的深度融合物理机和虚拟机特性的创新型计算架构。弹性裸金属服务器开创了一种新型的云服务器形式，它能与阿里云产品家族中的其他计算产品无缝对接，比如存储、网络、数据库等产品，完全兼容ECS云服务器实例的镜像系统，从而更多元化地结合您的业务场景进行资源构建。

弹性裸金属服务器通过技术创新实现客户价值。具体而言，弹性裸金属服务器具有以下优势：

- 用户独占计算资源

  作为一款云端弹性计算类产品，弹性裸金属服务器超越了当前时代下物理机级的性能和隔离性，使您独占计算资源，无虚拟化性能开销和特性损失。在CPU规格选择上支持8核、32核、96核等多个规格，并支持超高主频实例。以8核产品为例，弹性裸金属服务器实例支持超高主频至3.7 GHz ~ 4.1 GHz，与同类产品相比，它可以让游戏以及金融类业务获得更好的性能和更快的响应。
- 加密计算

  在安全性方面，弹性裸金属服务器除了具备物理隔离特性外，为了更好地保障您云上数据的安全性，弹性裸金属服务器采用了芯片级可信执行环境（Intel® SGX），能确保加密数据只能在安全可信的环境中计算。这种芯片级的硬件安全保障相当于为您云上的数据提供了一个保险箱功能，您可以自己掌控数据加密和密钥保护的全部流程。详情请参见 安装SGX。
- 兼容多种专有云

  弹性裸金属服务器可以进一步解决客户对高性能计算的强需求，更好地帮助客户搭建新型混合云。弹性裸金属服务器不仅具有虚拟机的灵活性和弹性，同时具备物理机的一切特性和优势，因此也具备再次虚拟化的能力，线下的专有云均可无缝平移到阿里云上，而不用担心嵌套虚拟化带来的性能开销，为客户上云提供一种新途径。
- 异构指令集处理器支持

  弹性裸金属服务器采用阿里云完全自主研发的虚拟化2.0技术，零成本支持ARM等其他指令集处理器。

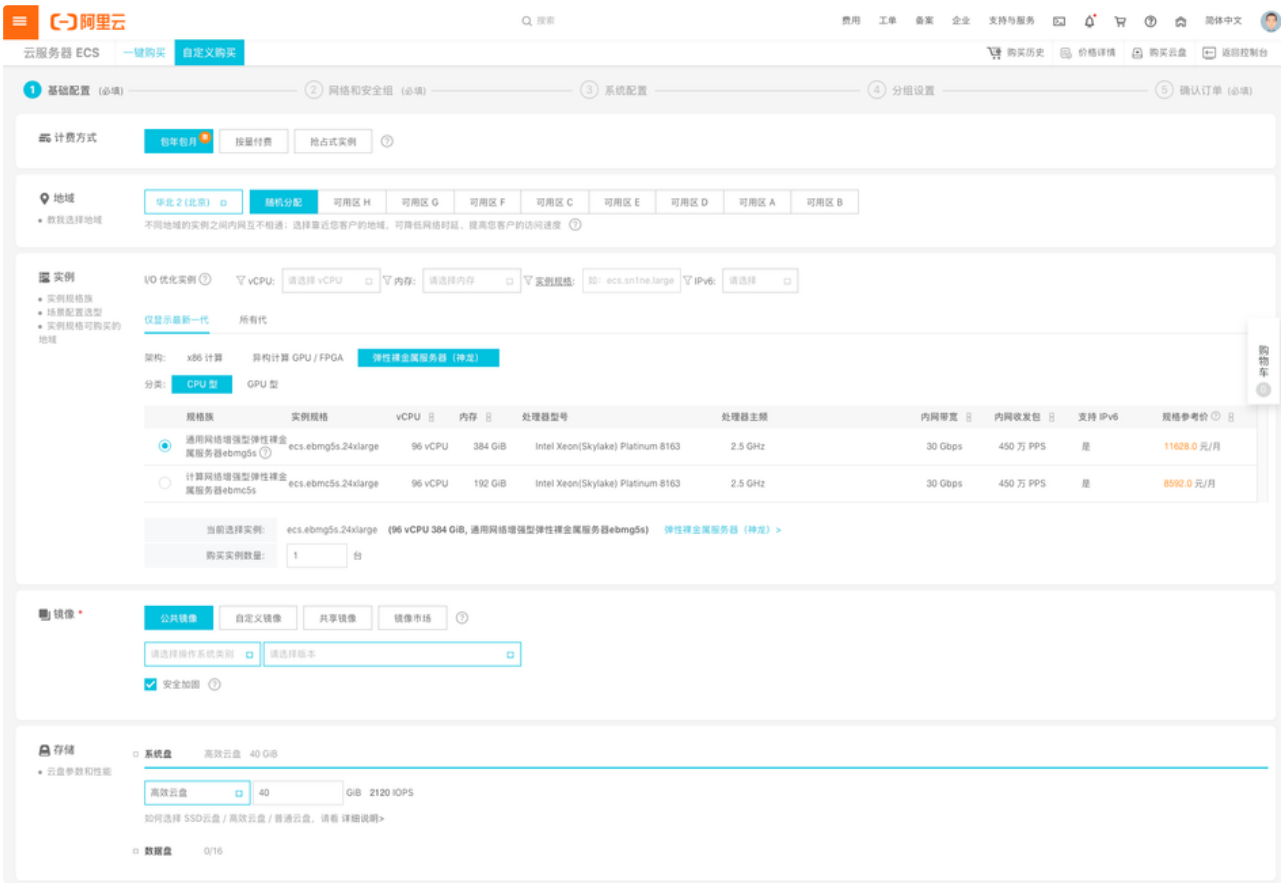
> ⚠ 使用弹性裸金属服务器时，请注意：
> - 目前不支持规格变配。
> - 当弹性裸金属服务器发生硬件故障时，支持故障转移，数据都保留在云盘中。

在配置参数时，您需要注意以下几点：

- **地域**：目前只能选择 **华东2可用区D**、**华北2可用区C**、**华东1可用区G**和**华南1可用区D**。
- **实例**：可以选择ebmhfg5、ebmc4和ebmg5。规格族的详细信息，请参见实例规格族。
- **镜像**：只支持部分公共镜像，如下表所示。 **操作系统类别镜像**Linux
  - CentOS 7.2/7.3/7.4/6.9/6.8 64位
  - Ubuntu 14.04/16.04 64位
  - Debian 8.9/9.2 64位
  - OpenSUE 42.3 64位
  - SUSE Linux Enterprise Server 12 SP2 64位
  - Aliyun Linux 17.1 64位

  Windows
  - 2016 数据中心版 64 位中文版
  - 2016 数据中心版 64 位英文版
  - 2012 R2 数据中心版 64 位中文版
  - 2012 R2 数据中心版 64 位英文版

- **存储**：弹性裸金属服务器支持最多挂载16块数据盘。您可以在这里添加数据盘，也可以在实例创建成功后再单独创建并挂载数据盘。具体操作，请参见创建按量付费云盘和挂载云盘。



功能分类	功能	弹性裸金属服务器	物理机	虚拟机
运维自动化	分钟级交付	Y	N	Y
计算	无性能损失	Y	Y	N
	无特性损失	Y	Y	N
	资源无争抢	Y	Y	N
存储	完全兼容ECS云盘系统	Y	N	Y
	使用云盘（系统盘）启动	Y	N	Y
	系统盘快速重置	Y	N	Y
	使用云服务器ECS的镜像	Y	N	Y
	物理机和虚拟机之间相互冷迁移	Y	N	Y
	免操作系统安装	Y	N	Y
	免本地RAID，提供更高云盘数据保护	Y	N	Y

网络	完全兼容ECS VPC网络	Y	N	Y
	完全兼容ECS经典网络	Y	N	Y
	物理机集群和虚拟机集群间VPC无通信瓶颈	Y	N	Y
管控	完全兼容ECS现有管控系统	Y	N	Y
	VNC等用户体验和虚拟机保持一致	Y	N	Y
	带外网络安全	Y	N	N/A

Zstack

并不是所有业务都适合在云端虚拟机上运行的，比如一些高性能的计算任务，如果运行在虚拟机上，就达不到在物理机上的效果。于是就需要裸金属服务，简单来说，裸金属服务就是为应用提供专属的物理服务器，保障核心应用的高性能和稳定性。ZStack早在2.6.0版本，在高级功能中以单独的功能模块形式，推出了裸金属服务。支持自定义安装操作系统，并提供裸金属主机的全生命周期管理。裸金属服务在以下几个方面拥有巨大优势：

1、高性能计算；

2、无法使用虚拟化的计算任务；

3、数据库主机；

4、单租户、专用硬件、安全性、可靠性以及其它需求。

裸金属管理服务的基本原理是：**PXE服务器提供DHCP服务和TFTP服务，指示多台裸金属设备由PXE网卡启动并分配动态IP，裸金属设备从PXE服务器中下载相关软件包，用于裸金属主机的系统安装。**

1. 管理节点与管理网络（Management Node）：需提前规划管理网络，要求镜像仓库、PXE服务器均与管理节点连通。管理节点作为安装系统的物理主机，提供ZStack的UI管理、云平台部署功能。一般是安装ZStackiso镜像的主机，通过前端的dashboard界面，进行图形化管理。

2. 镜像仓库：也位于管理网络网段之下，为裸机（可认为没有安装操作系统的新机器）提供多种操作系统镜像文件。在ZStack中，镜像支持本地与URL导入。

3. PXE(preboot execute environment，预启动执行环境)，支持通过网络从远端服务器下载映像，并由此支持通过网络启动操作系统，在启动过程中，终端要求服务器分配IP地址，再用TFTP服务协议下载一个启动软件包到本机内存中执行，由这个启动软件包完成终端(客户端)基本软件设置，从而引导预先安装在服务器中的终端操作系统。PXE可以引导多种操作系统。

可以概括认为ZStack的PXE服务器包含二大功能：其一就是DHCP服务（指示多台裸金属设备由PXE网卡启动并分配动态IP），其二就是TFTP服务（裸金属设备从PXE服务器中下载相关软件包，用于裸金属主机的系统安装）。

4. 部署网络，确保裸金属设备的PXE网卡与PXE服务器的DHCP监听网卡通过部署网络连通。可以说就是安装操作系统用的，它的独立性适用于生产环境(优先独立配置)，也可以以管理网络作为部署网络。

5. IPMI网络，确保管理节点与裸金属设备的BMC接口通过IPMI网络连通。IPMI的核心是BMC，即基板管理控制器，其并不依赖于服务器的处理器、BIOS或操作系统来工作，是一个**单独运行的无代理管理子系统**，只要有BMC与IPMI固件（运行在ROM里的只读程序）其便可开始工作，BMC通常是一个安装在服务器主板上的独立板卡。在工作时，所有的IPMI功能都是向BMC发送命令来完成的。

实现裸金属设备的**带外控制**（通过不同的物理通道传送管理控制信息和数据信息，两者完全独立，互不影响。），要求裸金属设备配备BMC接口（现在一般都有），并提前为每台裸金属设备配置好IPMI地址、端口、用户名和密码。

正因为IPMI的独立性，我们在进行裸机操作时，可以对其进行控制。当裸机安装完成，操作系统正常使用时才进行数据信息处理。正如上面所说的，通过不同的物理通道传送管理控制信息和数据信息。如下图（来自网络）：IPMI接口与服务器一般网络接口在不同位置。

为保证批量部署裸金属设备的顺利进行，需提前做好以下准备工作：

1. 手动安装管理节点，并安装相应许可证；即需要先安装好ZStack环境，并保证在ZStack环境中可以使用裸金属服务。

2. 在镜像仓库中准备若干ISO镜像，用于裸金属主机的系统安装。（此处的镜像服务器单独部署，镜像BIOS模式为legacy）

3. 进入裸金属设备BIOS启用PXE（可以自己进入裸金属设备BIOS开启）

4. 规划部署网络

要求PXE服务器的DHCP监听网卡是一个独立的、有IP地址的网卡，对外提供稳定的DHCP服务。

5. 配置裸金属设备IPMI并规划IPMI网络

提前规划IPMI网络，确保管理节点与裸金属设备的BMC接口通过IPMI网络连通。

这样通过IPMI网络，admin就可在UI界面完成所有裸金属设备的批量部署；并且管理节点可远程控制裸金属设备的开关机、网络启动、磁盘启动等行为。

6. 其它网络（可选）

如果裸金属主机需要与云虚拟主机进行交互的话。可以在一个扁平网络下，设置二类主机互通。

准备工作完成后，admin可登录管理节点界面(ZStack的dashboard界面)，进行接下来的操作。

- **创建裸金属集群，为裸金属设备提供单独的集群管理（和云主机区分开来）。**
  - 裸金属集群可以为裸金属设备提供单独的集群管理。注意：一个裸金属集群只允许挂载一个部署服务器。

- **创建部署服务器，为裸金属设备提供PXE服务和控制台代理服务。**
  - 本次与管理节点合并，但独立部署PXE服务器，可以满足多管理节点物理机高可用场景需求，且避免单点故障，大幅提升部署效率。然后将部署服务器挂载到裸金属集群中。
  - DHCP服务（为裸金属设备由PXE网卡启动并分配动态IP），TFTP服务（裸金属设备从PXE服务器中下载相关软件包，用于裸金属主机的操作系统安装）。

- **添加裸金属设备**
  - 裸金属设备：就是待安装操作系统的裸金属服务器，通过BMC接口以及IPMI配置进行唯一识别。
  - 需要填写IPMI网络，这样管理节点可远程控制裸金属设备的开关机、网络启动、磁盘启动等行为。

- **创建裸金属主机，进行自定义安装操作系统。**

  裸金属主机：即已安装操作系统的裸金属服务器，裸金属设备部署完成后可用于创建裸金属主机。创建界面如下，需要注意的是裸金属主机创建完成后会自动重启，然后根据所选镜像开始安装操作系统

需要注意的是，自动重启时，主机已经安装好操作系统，此时的启动应该从硬盘启动，而不是之前的网卡启动，可以登录裸金属设备的控制台，设置第一个引导设备为磁盘驱动器，确保主机从正确的地方启动，否则有可能导致主机无限重启。

重启完成后，如下图所示，主机处于正常运行状态。

在部署服务器上，可以看到镜像的缓存位置以及此过程中DHCP服务与TFTP服务。这里的部署服务器就相当于一个PxeServer。裸金属主机会发送DHCP广播请求，然后DHCP服务器向主机提供可用的IP地址并告知主机TFTP服务器的地址，之后TFTP向客户机提供内核，驱动及引导文件，最后通过TFTP获得安装文件，而安装时的参数由cfg文件来提供。

HUAWEI CLOUD STACK（核心：快速发放：5分钟一台、虚拟机互通：租户间安全隔离、支持共享盘：320TB共享盘）

BMS（Bare Metal Server），即裸金属服务器，为租户提供专属的物理服务器，拥有卓越的计算性能，能够同时满足核心应用场景对高性能及稳定性的需求，并且可以和VPC等其他云服务灵活的结合使用，综合了传统托管主机的稳定性与云上资源高度弹性的优势。

https://support.huawei.com/enterprise/zh/doc/EDOC1100062360 （6.5.0）

https://support.huawei.com/enterprise/zh/doc/EDOC1100091821/32efef08 （6.5.1）

BMS服务主要提供以下功能：

- 支持管理BMS的生命周期，包括创建、删除、查询、开启、关闭和重启BMS。
- 支持BMS之间，以及BMS与弹性云服务器之间的网络互通。
- 支持通过公共镜像创建BMS。
- 创建BMS时，支持选择虚拟私有云、添加网卡、添加高速网卡、绑定弹性IP、挂载云硬盘，并选择是否共享该硬盘等。
- 创建BMS后，支持绑定与解绑弹性IP、挂载与卸载云硬盘。

裸金属服务器具有以下技术优势：

- 混合部署，灵活组网

  裸金属服务器在可用分区内，内网互通。通过VPC实现与外部资源的互通，同时可以结合ECS等服务混合部署、灵活组网，满足用户多种复杂场景的不同诉求。
- 稳定可靠，性能卓越

  为租户提供专属的裸金属服务器，独享物理服务器的稳定性能，充分满足对高性能、稳定性以及数据安全和监管的业务诉求。
- 高吞吐、低时延

  为租户提供同一可用分区内裸金属服务器之间高吞吐、低时延网络，带宽最高可达到10Gbits，时延低至25μs。可应用于要求网络有高吞吐或低时延的场景。
- 裸金属服务VHAhttps://support.huawei.com/enterprise/zh/doc/EDOC1100091820/8d94b12d

应用场景

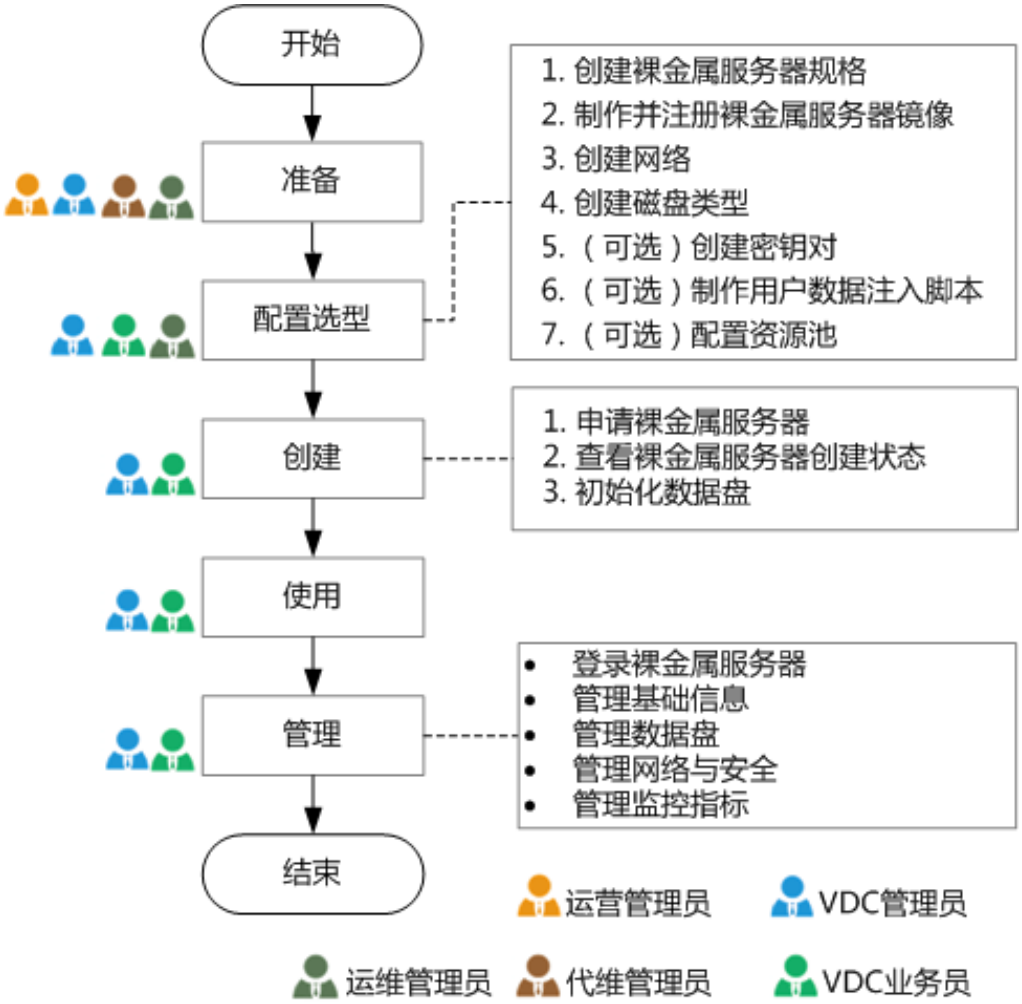
- 安全和监管高要求场景

  金融、证券等行业对业务部署的合规性要求，以及某些客户对数据安全的苛刻要求，只能采用物理服务器部署，确保资源独享、数据隔离和可监管可追溯。
- 高性能计算场景

  超算中心、基因测序、图形渲染等高性能计算场景，处理数据量大，对计算性能、稳定性、实时性等性能要求很高，无法承担虚拟化带来的性能损耗和超线程等影响。
- 核心数据库场景

  有些客户要求其关键的数据库业务不能部署在虚拟机上，而是必须通过资源专享、网络隔离、性能有保障的物理服务器承载。

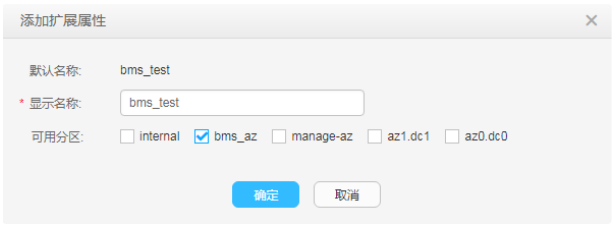
操作流程



当创建了裸金属服务器后，需了解裸金属服务器的运行状态。通过在裸金属服务器中安装监控Agent，自动收集裸金属服务器的CPU、内存、磁盘以及网络使用情况等监控指标，以便及时了解裸金属服务器运行状态。

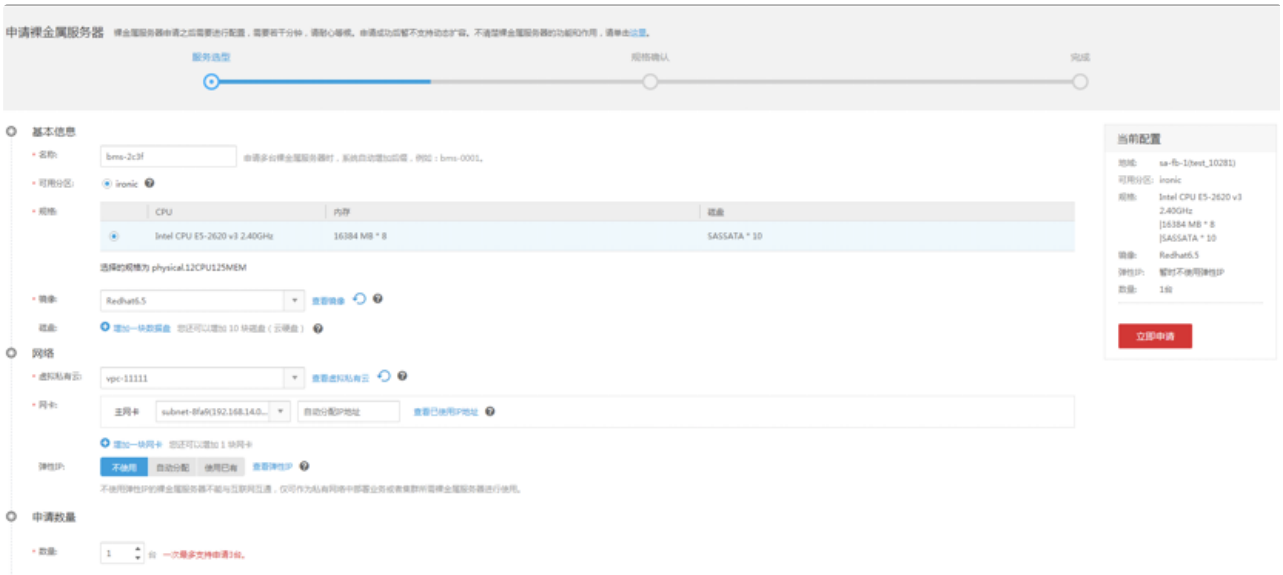
创建Flavor

控制可用域范围

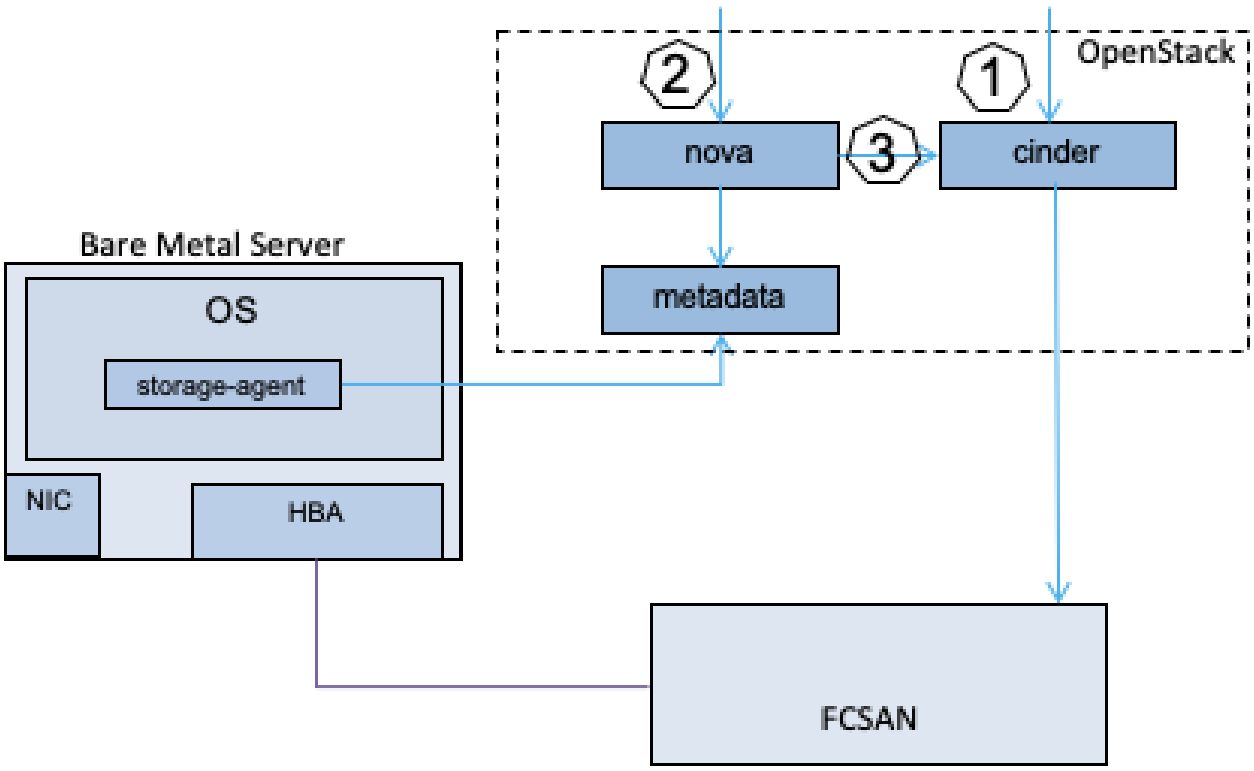


创建物理主机


* 项目
* 虚拟私有云
vpc-57ef
新建虚拟私有云
* 安全组
default (1b64faa3-b39f-4282-9000-e...
新建安全组
* 网卡
主网卡
VPC子网
subnet-57fa(192.168....
自动分配IP
查看已使用IP地址
增加一块网卡
您还可以增加 4 块网卡
高速网卡
增加一块高速网卡
您还可以增加 2 块高速网卡
新建高速网络
弹性IP
不使用
自动分配IP
使用已有
新建弹性IP


申请裸金属服务器
基本信息
网络
申请数量
当前配置
立即申请

支持对接使用商业存储


OpenStack
2
1
nova
3
cinder
metadata
Bare Metal Server
OS
storage-agent
NIC
HBA
FCSAN

Cinder 接口申请商业存储云硬盘，调用Nova接口将云硬盘挂载给裸机，Nova调用Cinder查询到的卷信息后，将挂载卷信息存到Metadata中，裸机的Storage-agent 读取Metadata挂载卷信息，自动完成扫描挂载卷操作。

ARM 架构：

支持采用Taishan2280 V2服务器作为裸金属服务器节点

ARM裸金属服务器部署方式支持：

- 全栈ARM服务器部署
- 管理节点采用X86服务器，业务节点采用ARM服务器部署
- 管理节点采用ARM服务器，业务节点采用X86服务器部署

ARM裸金属服务器AZ支持对接的存储类型：

- 支持对接FC SAN
- 支持采用iSCSI方式 ,对接FusionStorage

ARM裸金属服务器GuestOS 支持：

CentOS 7.6、Euler OS 2.8 、后续规划兼容、Ubuntu 18.04、中标麒麟V7u6

约束与限制

- TOR交换机采用华为自有产品
- BMS仅支持挂载SCSI模式的磁盘。（VHA只支持FCSAN）
- 无论是挂载共享盘还是非共享盘，均要求实例必须与云硬盘位于同一个可用分区。
- 不支持挂载给已经过期的实例。
- 不支持挂载给已经软删除的实例。
- 不支持挂载给关机状态的实例。
- 禁止修改网络相关的配置，否则可能导致无法连接裸金属服务器。
- 如果您需要对操作系统进行升级或打补丁，请从云服务商处获取相应的OS文件。
- 裸金属服务目前只支持从现有的操作系统进行升级或打补丁操作，不支持对已有的裸金属服务器重装操作系统

定价规则

私有云定价模式 目前市面上多采用：产品（软硬件）+服务模式，价格直观体现在：

1. license：控制裸机功能开启与否，可通过设置基础版功能套餐与高级版功能套餐进行灵活定价，例如：基础版包含裸机管理及正常使用的基础功能，高级版包含日志分析、多种登录方式、监控等功能，基础版和高级版的功能可依照功能列表和业务场景设定并制定价格；
2. 订阅：控制裸机功能的可用时长，依据不同订阅套餐差异化计费，套餐差异体现在如下几点定价因素：

- 节点数量能力（平台能力+物理设备）：限制客户能够使用的节点数量，提供默认裸机套餐，同时提供管理和使用更多节点的能力，增加节点需要单独收费，用户可以自己购买设备或购买我们的设备。例如：标准裸机套餐提供管理3个标准裸机节点的能力，如果客户需要使用更多节点，需要付出相应的费用，具体增加节点数量与价格的对应关系待整理；
- 节点性能（物理设备）：默认裸机套餐采用的标准裸机节点（具体设备型号待定），我们也提供的裸机设备目录供用户选择，用户可根据需要选择更高性能的物理设备。性能越高，价格越贵（设备与价格的关系待整理）；
- 服务 ：

标准规划与部署服务：标准套餐包含标准化的规划与部署，提供标准网络规划、镜像服务器部署、pxe服务器部署服务；

定制化规划与部署：在裸机设备目录内，根据客户的实际业务场景，提供专门的解决方案，并进行裸机集群规划、相关服务器部署、定制化镜像制作等服务。

高级开发定制：依据客户的特殊硬件需求进行深入开发，追加支持裸机设备目录外的设备，此项价格应极高；

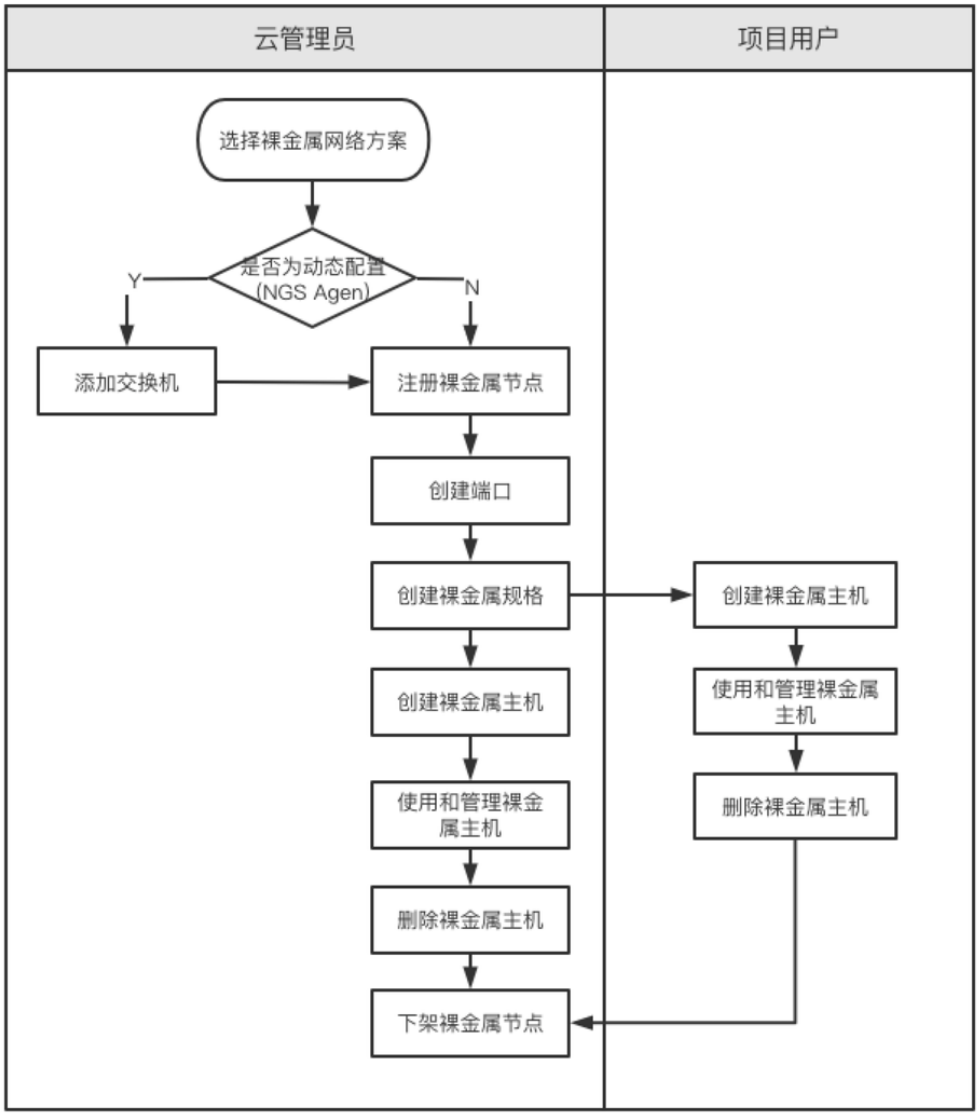
后期运维服务：提供故障分析与处理服务，帮助客户解决生产中遇到的问题。

定价规则可采用以下几种：

1. License买断：功能套餐费用（一次收费，定价要高）+部署服务费用（按次收费）+后期运维服务费用（按时长收费）；
2. 灵活订阅（客户自采购）：裸机套餐费用（节点数量能力费用）+部署服务费用（按次收费）+后期运维服务费用（按时长收费）；
3. 灵活订阅（本公司采购）：裸机套餐价费用（节点数量能力费用+裸机节点费用（受数量与性能影响））+部署服务费用（按次收费）+后期运维服务费用（按时长收费）；
4. 灵活订阅高级定制：裸机套餐费用（节点数量能力费用+裸机节点价费用（受数量与性能影响））+部署服务费用（按次收费）+高级开发定制费用+后期运维服务费用（按时长收费）；

六、业务流程

使用“泳道图”描述其中的干系人角色和业务流程。



七、功能性需求

需求模块	需求要素	排期	备注

裸机网络配置	提供三种网络配置方案：动态配置、预配置、后配置	一期	动态配置方案提供两种驱动方式：NGS Agent和SDN
	添加/编辑/删除交换机	一期	只适用于动态配置网络方案（NGS Agent）
裸机资源池管理	裸金属节点作为独立的AZ	一期	用户需要能感知和使用一AZ
	支持一云多芯	一期	支持在同一个AZ中可纳管和部署X86和ARM机器
	同一区域内包含裸金属节点和其他服务器	一期	
	支持裸金属节点OS的自动化安装	一期	
	批量注册（或称作发现）裸金属节点	一期	曾有需求提出自动注册功能: ERS-156: Ironic物理机资源自动注册 IMPLEMENTED 批量注册的最大数量待调研
	网卡bond	一期	支持两种bond模式：主备（bond1）和动态链路聚合（bond4）
	注册时通过可视化界面设置磁盘raid	二期	RAID放入二期，一期暂不支持； 陕西农信曾提出批量注册与信息导出需求 https://easystack.atlassian.net/browse/ERS-231
	编辑裸金属节点的基本信息	一期	名称、可用区、元数据等
	下架裸机	一期	
	裸机开机/关机/重启	一期	通过IPMI操作
	维护/恢复	一期	
	HBA卡管理	二期及以后	目前暂未获取相应需求
	创建/编辑/删除端口	一期	
	端口组	一期	创建、编辑、删除、添加移除端口
	机框	二期及以后	目前暂未获取相应需求
	~~VNC/控制台登录~~	~~无此需求~~	
	导出裸机列表	一期	陕西农信曾提出批量注册与信息导出需求 https://easystack.atlassian.net/browse/ERS-231
	裸金属节点集群管理	二期及以后	需要支持把裸金属节点作为统一的集群进行管理

裸金属主机调度	通过列表展示裸金属主机类型的信息	一期	
	创建裸金属主机规格	一期	
	编辑裸金属主机规格	一期	
	删除裸金属主机规格	一期	
裸金属主机生命周期管理	支持列表展示裸金属主机相关信息	一期	
	支持详细展示逻辑云主机相关信息	一期	
	支持创建（批量创建）裸金属主机	一期	
	能够根据规格条件选择合适的裸金属服务器	一期	
	编辑裸金属主机的基本信息	一期	名称
	启动	一期	
	关机	一期	
	重启	一期	
	删除	一期	
	重建	一期	
	恢复	二期及以后	
	控制台登录	一期	支持IPMI硬件类型，可见jira EAS-65610: 裸机支持console控制台 已完成
	虚拟网卡管理	一期	
	支持本地盘作为数据盘	一期	当前仅能使用本地盘
	支持挂载/卸载其他存储设备作为数据盘	二期及以后	
	系统盘支持使用本地盘	一期	
	支持从cinder启动裸金属主机	二期	iPXE启动方式的系统盘可以云盘CINDER启动
	支持指定存储设备大小和类型	二期及以后	一期可能只有一种存储类型即其本地存储，不支持指定存储类型（云硬盘的flavor）和云硬盘容量
	自动化安装所需的裸金属主机操作系统	一期	操作系统至少支持基于centOS7.6的红旗操作系统

	支持多网卡及虚拟网卡管理	一期	SRIOV暂不影响此功能 最大网卡数量待研发调研和反馈
	支持绑定公网IP	一期	
监控和告警	支持统计裸金属资源池CPU，内存，磁盘分配情况和总量；	二期及以后	期待裸金属节点支持和计算节点无差别的的监控和告警-->需要额外agent，调研中 期待裸金属主机实现和普通云主机无差别的监控和告警-->需要额外agent，调研中 期待裸机资源池实现和普通虚拟资源池无差别的监控参数-->需要额外agent，调研中 （以上由李向军帮忙调研）-->8.3向军反馈，由于客户环境限制，为保证安全性，无法打通管理网与实际业务网，即使安装了相关agent，也无法实现监控数据的传输，因此一期暂不实现监控告警相关功能
	支持裸金属节点硬件系统事件的采集，包括硬件错误事件，并支持服务器系统事件转化为告警信息。	二期及以后	
	支持针对裸金属节点进行统一性能监控（CPU使用率、内存使用率等）和告警；	二期及以后	
	支持针对裸金属主机进行统一性能监控（CPU使用率、内存使用率等）和告警；	二期及以后	待确认能否通过BMC的IPMI直接获取部分监控信息。-->只能获取到风扇转速、CPU电压数据，无法获取CPU内存等的使用率。
存储资源池	支持使用本地存储	一期	目前只支持本地存储，不支持挂载外部存储设备 邮政需求：计算虚拟化节点只访问一个IP SAN存储池，而裸机可以访问两个存储池的资源，但需保留灵活控制能访问存储池的能力。
	支持多台裸金属节点使用同一集中SAN存储资源池	二期及以后	
	支持同时对接IP SAN和FC SAN	二期及以后	
	支持使用多个存储资源池	二期及以后	
多租户	支持多租户	一期	
负载均衡器	支持裸金属主机作为负载均衡资源池的资源	一期	
网络	支持VLAN	一期	
	支持VXLAN	二期及以后	
	裸机网络与云主机网络互通	一期	
	支持对接SDN硬件	一期	
数据导出	导出裸机相应监控数据	二期及以后	
	导出裸金属节点列表	一期	

	导出裸金属主机列表	一期	
计费	支持裸金属主机单独定价	二期	向军：裸机计费和虚拟机类似，计费只要引入一种裸机的计费类型。 具体方案等待调研但具体计费策略待定，邮政不需要计费
日志/操作审计	记录用户通过界面或后台对裸机的操作审计信息	一期	
编排部署	支持审批	二期及以后	暂不开发

八、非功能性需求

*升级与进化需求

- 升级需求：裸机管理目前仍存在很大的功能迭代空间，在今后的版本也会不断迭代新的功能，裸机作为云基础设施的重要组成部分，需要实现平滑升级。裸机服务可包装作为独立的子服务，支持作为独立的子服务进行版本升级。
- 进化需求：裸机管理目前兼容的硬件设备和功能边界优先，存在一些二期及以后要完成的功能要素，在进行相应设计和开发时，应考虑到可进化与平滑，保证现有设计代码可复用。

*能力需求

- License 需求：需要激活ECS解决方案架构的License，才能具备裸机管理功能
- 版本管理需求：裸机管理可作为独立的子服务，不同版本的裸机管理功能依赖于不同的平台版本、CPU架构、部署形态，需要注意之间的对应关闭，便于子服务的升级和统一管理。此外，待确认对OEM版本的影响
- 安装部署扩容能力需求：裸机管理功能需要深入讨论平台网络架构的调整及存储对接相关的问题，目前可通过现有的发现节点把裸机节点当做普通的计算节点使用（但这并非值得提倡的使用方式）。安装部署裸机节点需要实现可灵活发现注册节点，可在线扩容节点。
- 高可用需求：需要实现裸机服务的高可用和裸机部署架构的高可用。
- 计费需求：需要计费，按照资源用量，资源类型计费，需深入考虑定价模式，目前可采用云主机的计费策略（即针对裸金属主机类型计费）。

接口需求

- API接口：需要符合OpenAPI的规范
- 硬件接口

性能需求

- 时间特性要求：创建裸金属主机时间？
- 存储性能要求：...
- 网络性能要求：...

安全性需求

- 密码：密码相关校验和强度和云主机保持一致
- 容灾：一期暂不支持容灾
- 备份：一期暂不支持备份

权限需求

- 权限：需要增加相应功能权限，并考虑不同角色的用户的权限范围。

权限：

一级功能	二级功能	三级功能	用户类型		
			云管理员	部门管理员	项目管理员/项目用户
裸机网络配置	编辑网络配置方案		√	×	×
	添加/编辑/删除交换机		√	×	×
裸机节点	裸机信息展示	端口创建、编辑、删除	√	×	×
		端口组创建、编辑、删除、管理端口	√	×	×
	注册与批量注册		√	×	×
	编辑裸金属节点		√	×	×
	删除裸机		√	×	×
	开机		√	×	×
	关机		√	×	×
	重启		√	×	×
	维护		√	×	×
	取消维护		√	×	×
	导出裸金属节点列表		√	×	×
裸金属主机类型	裸金属主机类型的信息展示		√	×	×
	创建裸金属主机类型		√	×	×
	编辑裸金属主机类型		√	×	×
	删除裸金属主机类型		√	×	×
裸金属主机	裸金属主机信息展示		√	√仅已加入项目的	√仅已加入项目的
	创建（批量创建）裸金属主机		√	√仅已加入项目的	√仅已加入项目的
	编辑裸金属主机		√仅编辑自己的	√仅已加入项目的	√仅已加入项目的
	开机		√	√仅已加入项目的	√仅已加入项目的

	关机		√	√仅已加入项目的	√仅已加入项目的
	重启		√	√仅已加入项目的	√仅已加入项目的
	重建		√	√仅已加入项目的	√仅已加入项目的
	删除		√	√仅已加入项目的	√仅已加入项目的
	控制台登录		√	√仅已加入项目的	√仅已加入项目的
	虚拟网卡管理		√	√仅已加入项目的	√仅已加入项目的
	挂载盘管理		√	√仅已加入项目的	√仅已加入项目的
	网卡管理		√仅编辑自己的	√仅已加入项目的	√仅已加入项目的
	弹性IP管理		√	√仅已加入项目的	√仅已加入项目的
	导出裸金属主机列表		√	√仅已加入项目的	√仅已加入项目的
~~监控~~	~~监控信息查看与导出~~		~~√~~	~~√仅已加入项目的~~	~~√仅已加入项目的~~
~~告警~~	~~告警~~		~~√~~	~~√仅已加入项目的~~	~~√仅已加入项目的~~
负载均衡器	作为负载均衡器的资源		√	√仅已加入项目的	√仅已加入项目的
计费管理	设置计费项		√	×	×
日志/操作审计	查看和导出日志/操作审计		√	√仅已加入项目的	√仅已加入项目的

- 多区域：所有区域都需要支持启用裸机管理功能
- 配额：可通过配额控制各部门和项目可使用的裸金属云主机数量和裸金属节点数量。

*可运维性需求

- 日志：完整记录全部日志
- 监控：实现裸机监控和裸金属云主机监控（二期及以后）
- 告警：实现裸金属服务告警功能

硬件要求：

- 硬件配置要求

一期暂不支持第三方商业存储，存储仅支持本地盘

*号需求重点考虑

九、产品边界与限制

功能边界

- 架构：支持一云多芯，同一节点下可同时纳管X86和ARM机器；
- 镜像：分为部署镜像和业务镜像。必须支持基于centos6.7的红旗操作系统，其他镜像期待支持：Windows镜像（Microsoft Windows Server操作系统已启用的支持版本）、Linux发行版操作系统（RHEL/CentOS系列、Debian/Ubuntu系列、SUSE/openSUSE系列等）
- 批量注册（即纳管）：支持批量，一次最多注册数量暂无限制；支持同时注册X86和ARM架构节点
- 多层网络：支持VLAN，VXLAN的支持依赖对接SDN；
- 部署网络：不支持安装裸金属云产品后修改部署网络
- 网卡数量：最大网卡数量受到裸金属服务器物理网卡数目的限制
- IPV4/IPV6：支持；
- 硬盘多挂载：暂不支持硬盘挂载
- 登录方式：

1. 裸金属云主机：控制台

- 信息展示：

1.裸机

基本属性：显示裸机当前的注册状态、电源状态、名称、规格（包括CPU核数、内存大小、本地盘大小）、CPU架构、CPU型号、IPMI配置、创建的裸金属云主机、裸机设备UUID等信息；

硬件配置：显示裸机已配置的网卡信息列表、磁盘信息；

审计：查看此裸机的相关操作

2.裸金属云主机

基本属性：展示裸金属云主机当前的状态、名称、规格（包括CPU核数、内存大小、本地盘大小）、CPU架构、CPU型号、IP、UUID、相应的裸金属节点、使用的镜像。

配置信息：展示裸金属云主机已配置的网卡信息列表、磁盘信息列表。

审计日志：查看此裸金属云主机的相关操作记录。

性能边界

- 并发创建主机：？

规模与容量边界

- 目前需要支持多大规模数量的裸金属节点：？

十、附录

参考资料

- 裸机服务安装指南（T版）：https://docs.openstack.org/ironic/train/install/index.html
- 裸机管理员指南（T版）：https://docs.openstack.org/ironic/train/admin/index.html
- 为Bare Metal服务创建用户镜像（T版）：https://docs.openstack.org/ironic/train/install/creating-images.html

- 设置裸机服务的驱动程序（T版）：https://docs.openstack.org/ironic/train/install/setup-drivers.html
- openstackIronic 服务安装部署指南（最新版本）：https://docs.openstack.org/ironic/latest/install/refarch/common.html#components
- 基于Ironic实现X86裸机自动化装机实践与优化（民生银行）：http://www.talkwithtrend.com/Article/243747

附表1:

- 来自社区的ironic支持的设备

交换机硬件设备支持表		
Switch	Type	Support
Cisco 300-series switches	VLAN	Y
Cisco IOS switches	VLAN	Y
Huawei switches	VLAN	Y
OpenVSwitch	VLAN	Y
Arista EOS	VLAN	Y
Dell Force10	VLAN	Y
Dell PowerConnect	VLAN	Y
Brocade ICX (FastIron)	VLAN	Y
Ruijie switches	VLAN	Y
HPE 5900 Series switches	VLAN	Y
Juniper Junos OS switches	VLAN	Y
SDN		Y