ZStack 技术白皮书精选

详解 ZStack 高级功能

--裸金属服务部署实践

扫一扫二维码，获取更多技术干货吧

版权声明

摘要

大道至简·极速部署，ZStack 致力于产品化私有云和混合云。

ZStack 是新一代创新开源的云计算 IaaS 软件，由英特尔、微软、CloudStack 等世界上最早一批虚拟化工程师创建，拥有 KVM、Xen、Hyper-V 等成熟的技术背景。

ZStack 创新提出了云计算 4S 理念，即 Simple（简单）、Strong（健壮）、Smart（智能）、Scalable（弹性），通过全异步架构，无状态服务架构，无锁架构等核心技术，完美解决云计算执行效率低，系统不稳定，不能支撑高并发等问题，实现 HA 和轻量化管理。

ZStack 发起并维护着国内最大的自主开源 IaaS 社区——zstack.io，吸引了 6000 多名社区用户，对外公开的 API 超过 1000 个。基于这 1000 多个 API，用户可以自由组装出自己的私有云、混合云，甚至利用 ZStack 搭建公有云对外提供服务。

ZStack 拥有充足的知识产权储备，积极申报多项软著和专利，参与业内标准、白皮书的撰写，入选云计算行业方案目录，还通过了工信部云服务能力认证和信通院可信云认证。

ZStack 面向企业用户提供基于 IaaS 的私有云和混合云，是业内唯一一家实现产品化，并领先业内首家推出同时打通数据面和控制面无缝混合云的云服务商。选择 ZStack，用户可以官网直接下载、1 台 PC 也可上云、30 分钟完成从裸机的安装部署。

目前已有 1000 多家企业用户选择了 ZStack 云平台。

详解 ZStack 高级功能--裸金属服务部署实践

作者：ZStack 社区 秦伟

一、前言

今天我们来了解一下 ZStack 的裸金属，提到裸金属服务，很多人从字面上可能对其不是很了解，其实早在之前的私有云 OpenStack 平台，就已经推行了 Ironic 裸金属服务，而且在去年的最新 Rocky 版本中，更是对裸金属服务进行了加强。于此同时的 ZStack 在 2.6.0 版本，也推出裸金属纳管服务。那么这令人注目的裸金属服务究竟是什么呢？

首先让我们来了解一下裸金属服务的由来，近年来由于国内外云计算市场的快速发展，许多企业纷纷将自身业务迁至云端。不再将业务部署在自己自身的机房环境中，这样带来的好处就是省去了一部分的人工维护成本，转而由第三方云供应商来提供基础环境。

而且一般来说硬件资源在很多情况下是没有被充分利用的，比如我们日常在使用自己的电脑时，实质上就是在使用它的 CPU、内存、以及在硬盘上运行的操作系统等。当我们查看这些资源的使用率时，通常会发现，CPU 和内存大部分是闲置的。特别是 CPU，其利用率通常不到 10%。那么，有没有可能不让一个操作系统单独控制一台机器，而是在一台机器上安装多个操作系统，并且让它们同时地运行，把被闲置的资源利用起来呢？答案是有，相信很多人都曾经在自己的 Windows 电脑上安装 VMware workstation，并且安装了多个虚拟机，每个虚拟机都拥有自己的操作系统，它们可以同时运行，并且不互相干扰，就实现了自己硬件电脑的虚拟化，可以把一台物理服务器虚拟化为多台虚拟服务器。所以说，这种通过管理程序（VMware workstation 等）把硬件的机器、同操作系统分开的过程，就是虚拟化。当我们将业务运行在云端时，可以做到按需求选取最合适的规模，将资源的利用率使用到最大。这些资源不仅仅包括 CPU、内存、操作系统，还包括网络，ip，安全组等。

但是，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；

二、ZStack 裸金属服务概述

ZStack 作为一套产品化创新开源云计算 IaaS 平台,它可以为企业用户提供私有云和混合云服务，当我们在 ZStack 中部署裸机，用到的就是 ZStack 的高级功能-裸金属服务，即直接控制物理机进行硬件部署操作，我们一般熟知的虚拟机（宿主型）是通过 Hypervisor 来部署的。如下图，Hypervisor 是一种运行在物理服务器和操作系统之间的中间软件层,可允许多个操作系统和应用共享一套基础物理硬件，因此也可以看作是虚拟环境中的"元"操作系统，它可以协调访问服务器上的所有物理设备和虚拟机，也叫虚拟机监视器 VMM(Virtual Machine Monitor)。当服务器启动并执行 Hypervisor 时，它会给每一台虚拟机分配适量的内存、CPU、网络和磁盘，并加载所有虚拟机的客户操作系统。

相比之下，裸金属服务就是传统形式，直接将 OS 部署在 Hardware 上，没有 VMM 这一层的损耗，性能更加优秀。

Hosted Architecture

裸金属服务的优势不言而喻，现在的各大公有云厂商也纷纷推出了自己的裸金属服务，作为私有云的 ZStack 也不甘示弱。现在我们暂时不考虑之后裸金属部署后的性能优势，单从部署方面来说，如何像部署虚拟机一样去部署物理机呢？

ZStack 可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。它可以直接对物理机执行节点级别管理，进行物理机节点的添加、删除，进行电源管理，部署系统等操作。在完成基本的服务器上架以及相关准备工作后，(注意这里的相关准备工作,是我们是否能顺利控制裸金属设备的关键前提，后面会详细介绍)，管理员可在 UI 界面批量部署裸金属设备，部署完成后可使用裸金属设备创建裸金属主机，支持自定义安装操作系统，并对裸金属主机进行全生命周期管理。

简单来说，我们在这里可以认为裸金属服务，就是为服务器裸机安装相应的操作系统，并且获取其配置信息，最后实现对裸金属主机的生命周期控制，比如：开关机重启等操作。而且对于整个操作过程而言，前提只需要服务器主机有网络并且通电就可以。

三、ZStack 裸金属服务基本原理

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。**

裸金属管理网络拓扑所示：（官方）

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。

所以需要配置裸金属设备 IPMI 并规划 IPMI 网络：

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。

规划 IPMI 网络后，管理节点与裸金属设备的 BMC 接口可以通过 IPMI 网络连通，并且 admin 用户可在之后的 UI 界面完成所有裸金属设备的批量部署。

6. 其它网络。

支持扁平网络场景，同一个二层网络上的裸金属主机和云主机之间可互相访问，无需通过网关进行路由，需提前将裸金属设备所在的裸金属集群挂载到相应的二层网络。

四、ZStack 裸金属服务操作流程详解

此次操作流程，将管理节控制节点与 PXE 服务器部署在同一个节点，并且管理网络与部署网络为同一个网络。如果有条件，建议在生产环境中依照官方拓扑图部署。

4.1 准备工作

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）

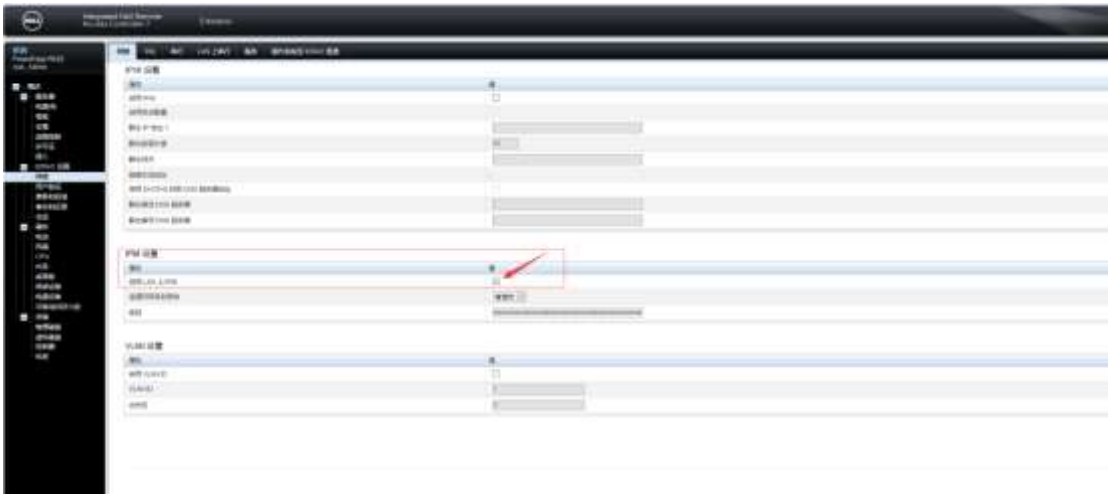
提前进入每台裸金属设备的 BIOS，确认其连接部署网络的网卡开启 PXE 功能。对于部分机型，还需确保该 PXE 网卡为首张启动网卡，或确保（启动顺位）在 PXE 网卡之前的所有网卡均关闭 PXE 功能，同时需确保裸金属设备的启动模式为 Legacy。

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。

4.2 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。

创建界面如下图：创建完成后，默认启动。

创建完成如下图：

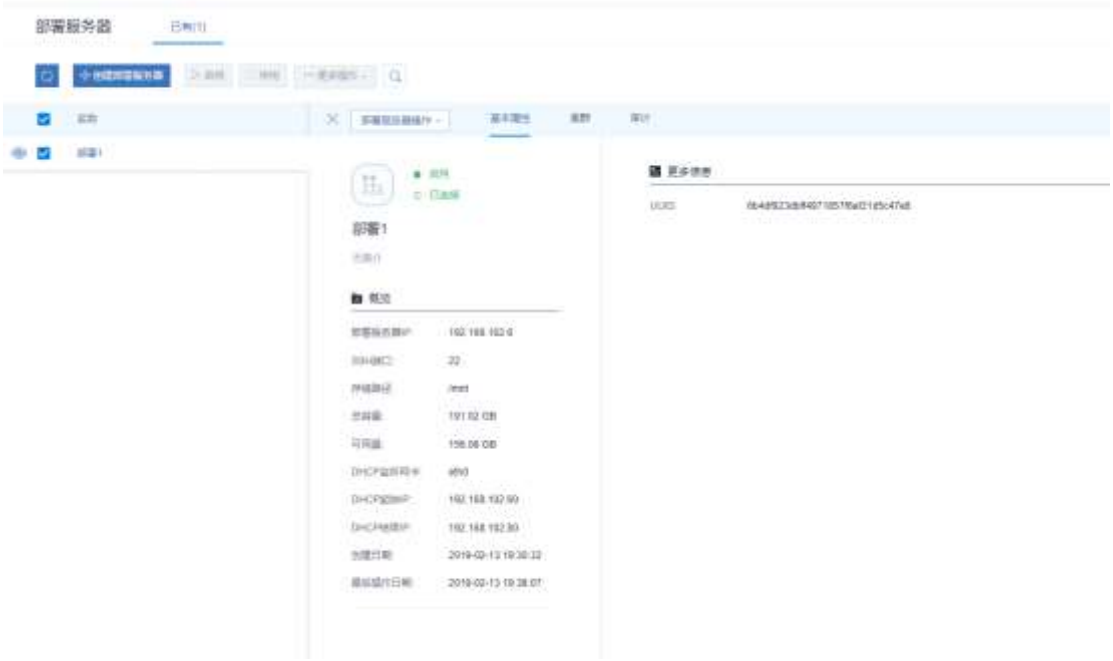
4.3 创建部署服务器，为裸金属设备提供 PXE 服务和控制台代理服务。

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。

如下图所示：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）。

创建完成后如下图所示：

同时，点击部署服务器可看到属性信息：



4.4 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。创建如下图：

创建完成后，如下图所示，可以看到已经获取到了硬件信息。

此时这里可以打开控制台，直接跳转至该裸金属设备的 IPMI 管理界面（登录界面），输入之前已配置好的 IPMI 用户名和 IPMI 密码，即可从 ZStack 界面跳转登录。

4.5 创建裸金属主机，进行自定义安装操作系统。

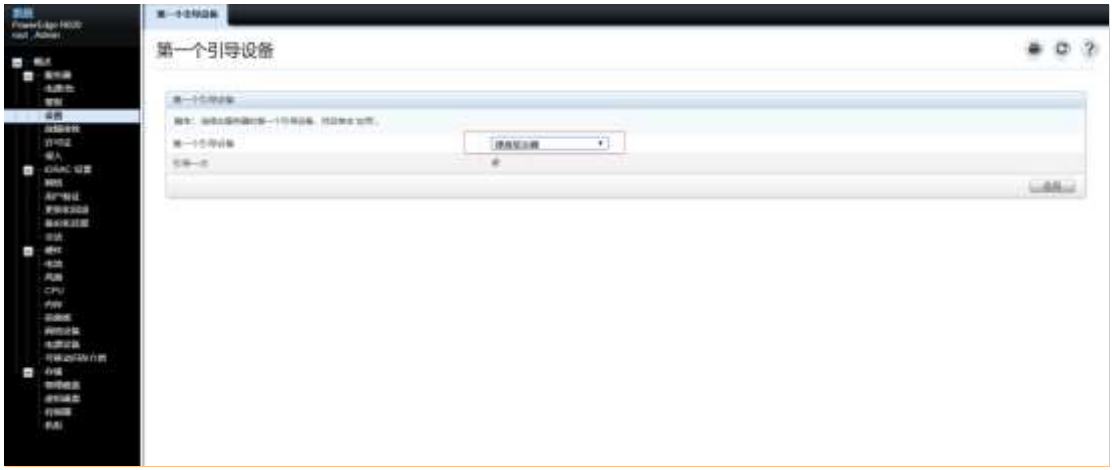
裸金属主机：即已安装操作系统的裸金属服务器，裸金属设备部署完成后可用于创建裸金属主机。创建界面如下，需要注意的是裸金属主机创建完成后会自动重启，然后根据所选镜像开始安装操作系统；

创建过程中，裸金属主机的状态会暂时显示为部署中。

这时我们需要打开控制台，进入系统安装界面，手动进行相关配置。如下图：

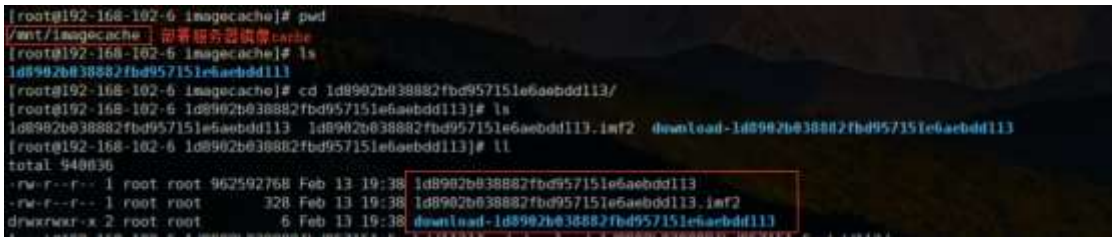
部署完成后，裸金属主机自动重启，就绪状态显示为已部署。

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。



```
[root@192-168-102-6 download-1d8902b038882fbd957151e6aebdd113]# cd /var/lib/zstack/baremetal/vsftpd/
[root@192-168-102-6 vsftpd]# ls
vsftpd.conf
[root@192-168-102-6 vsftpd]# cat vsftpd.conf
anonymous_enable=YES
anon_root=/var/lib/zstack/baremetal/ftp/
local_enable=YES
write_enable=YES
local_umask=022
dirmessage_enable=YES
connect_from_port_20=YES
listen=NO
listen_ipv6=YES
pam_service_name=vsftpd
userlist_enable=YES
tcp_wrappers=YES
xferlog_enable=YES
xferlog_std_format=YES
xferlog_file=/var/log/zstack/baremetal/vsftpd.log
[root@192-168-102-6 vsftpd]# cd ..
[root@192-168-102-6 baremetal]# ls
dnsmasq  ftp  noVNC  noVNC.tar.gz  tftpboot  vsftpd
[root@192-168-102-6 baremetal]#
```

安装完成时，登录裸金属主机，可以看到 cfg 配置参数文件：

```
[root@localhost ~]# ll
total 8
-rw-------. 1 root root 3351 Feb 13 19:51 anaconda-ks.cfg
-rw-------. 1 root root 2476 Feb 13 19:51 original-ks.cfg
[root@localhost ~]# cat anaconda-ks.cfg
#version=DEVEL
# System authorization information
auth --enableshadow --passalgo=sha512
# Use network installation
url --url="ftp://192.168.102.6/1d8902b038882fbd957151e6aebdd113"
# Use graphical install
graphical
firstboot --disable
ignoredisk --only-use=sda
# Keyboard layouts
# old format: keyboard us
# new format:
keyboard --vckeymap=us --xlayouts='us'
# System language
lang en_US.UTF-8

# Network information
network  --bootproto=static --device=74:86:7a:e0:07:6c --gateway=192.168.102.6 --ip=192.168.102.64 --nameserver=223.5.5.5 --netmask=255.255.255.0 --activate
network  --hostname=localhost.localdomain
# Reboot after installation
reboot
# Root password
rootpw --iscrypted $6$HWorT6BP$4U9kdZBBOor8BBGwLtaeYQRMn/.DOp3lPw/.EJ5ZZlBDo3DkaHnyb.qazqrN4FxBCyeuDcBQNB8OWY0K8SQ1
# SELinux configuration
selinux --disabled
```

4.6 对裸金属主机进行全生命周期管理。

总结：

由上分析，ZStack 裸金属管理服务具有以下功能优势：首先可以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；其次在操作过程中的各个服务可以进行独

立部署，比如：PXE 服务器，可满足多管理节点物理机高可用场景需求，彻底避免 DHCP 冲突，由于每个裸金属集群均可挂载独立 的 PXE 服务器，避免单点故障，大幅提升部署效率，以及镜像仓库的独立部署。同时管理员可在 UI 界面上批量添加裸金属设备，包括：手动添加和模板文件导入两种方式，支持批量添加 IPMI 地址，高效部署裸金属集群，提升运维效率，而且支持自定义安装操作系统。最后裸金属主机并不是独立的，它还支持扁平网络场景，同一个二层网络上的裸金属主机和云主机之间可互相访问，不需要通过网关进行路由，可以与企业自身业务紧密联合，发挥裸金属主机的优势。

借用 ZStack 官网上看到的一句话来说一下私有云裸金属的未来：随着虚拟机技术的日趋成熟，虚拟机所带来的性能损耗会越来越少，一些基于性能考虑而选择裸金属的需要未来可能会越来越少。但在一些特殊场合，针对一些特殊设备如龙芯或其他不能虚拟化的设备中，云平台以裸金属形式纳管这些设备一定时间内还会长期存在。针对这种形式的裸金属设备，提供通用的管控接口，智能调度和状态监控也许是未来私有云裸金属管理方面发展的重点。