SR-IOV网卡虚拟化 使用教程

产品版本：ZStack 3.10.0

文档版本：V3.10.0

ZStack

版权声明

商标说明

ZStack商标和其他云轴科技商标均为上海云轴信息科技有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受云轴科技公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，云轴科技公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

目录

1 概述

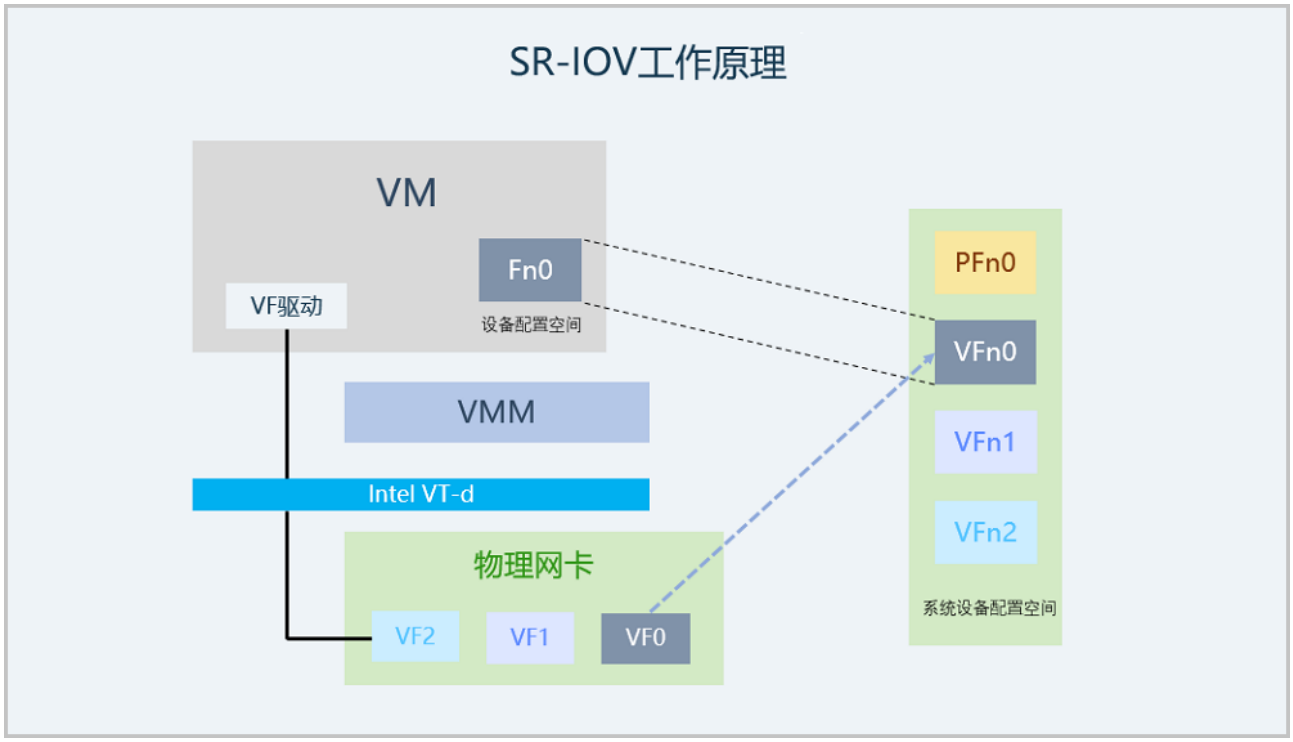
SR-IOV（Single Root I/O Virtualization）是一种基于硬件的虚拟化解决方案，它允许多个云主机高效共享PCIe设备，且同时获得与物理设备性能媲美的I/O性能，能有效提高性能和可伸缩性。

ZStack支持基于SR-IOV规范，将一张物理网卡虚拟化切割成多张VF类型网卡，直接分配给云主机使用。实现弹性灵活使用资源的同时，提高资源利用率、节约成本。相比传统的vNIC虚拟化网卡，VF网卡具有以下功能优势：

- VF网卡可直接分配给云主机，越过虚拟化层，缩短数据传输路径，使云主机获得接近物理设备的I/O性能。
- 明显减少对物理机CPU资源的消耗，即使物理机CPU压力较大，也能有效减少网络丢包，提高传输效率。

如图 *1: SR-IOV工作原理*所示：

图 1: SR-IOV工作原理



2 注意事项

ZStack云平台使用SR-IOV功能需要注意以下情况：

- 使用SR-IOV功能前须严格确保准备工作全部完成，否则此功能无法正常使用。详情请参考准备工作章节。
- 若VF网卡已分配给云主机使用，请勿卸载物理网卡驱动，否则可能导致VF网卡强行回收。
- 若物理网卡已配置bond，继续使用SR-IOV功能可能导致VF与vNIC网卡相互通信受影响，推荐使用方式如下：
  - 推荐针对单个物理网卡配置bond，并继续使用SR-IOV功能。
  - 多个物理网卡配置bond时，推荐仅对其中一个物理网卡进行SR-IOV切割。
- VF网卡不支持QoS功能。
- 使用启用SR-IOV功能的三层网络创建的云主机，暂时不支持使用以下网络服务：
  - 使用公有网络/扁平网络创建的云主机，其VF网卡不支持使用安全组和弹性IP网络服务。
  - 使用云路由网络/VPC网络创建的云主机，其VF网卡不支持使用安全组网络服务。
- 运行状态且加载VF网卡的云主机不支持迁移、存储迁移操作。须停用云主机或卸载VF网卡才能执行这些操作。
- 停用云主机将自动释放VF网卡；启用云主机时重新获取，若无可用VF网卡，将导致启用云主机失败。

3 准备工作

ZStack云平台使用SR-IOV功能需要准备以下工作：

- 确保物理网卡支持SR-IOV功能，并安装到物理机主板。例如：支持SR-IOV功能的Intel网卡列表*Intel官方文档*。
- 确保该物理机BIOS已开启Intel VT-d / AMD IOMMU配置和SR-IOV配置。
- 确保物理机已安装物理网卡（PF网卡）驱动；确保已获取到VF网卡对应的驱动。

  注: 相关网卡驱动以及安装方法请联系网卡提供厂商获取帮助。

  - PF网卡驱动需安装到相关物理机，保证物理网卡可被物理机识别，且能正常进行SR-IOV切割。
  - VF网卡驱动需安装到相关云主机，保证VF网卡被云主机识别且正常工作。
- 确保物理机CPU支持Interrupt Remapping。在物理机执行以下脚本，查看CPU是否支持Interrupt Remapping：

```
[root@localhost ~]# cat interrupt_remapping_check.sh
#!/bin/sh
if [ $(dmesg | grep ecap | wc -l) -eq 0 ]; then
  echo "No interrupt remapping support found"
  exit 1
fi

for i in $(dmesg | grep ecap | awk '{print $NF}'); do
  if [ $(( (0x$i & 0xf) >> 3 )) -ne 1 ]; then
    echo "Interrupt remapping not supported"
    exit 1
  fi
done
```

若物理机CPU不支持Interrupt Remapping，须执行以下命令进行配置：

```
[root@localhost ~]# echo "options vfio_iommu_type1 allow_unsafe_interrupts=1" > /etc/modprobe
.d/iommu_unsafe_interrupts.conf
```

4 典型使用流程

背景信息

SR-IOV网卡虚拟化的典型使用流程如下：

1. 物理机启用IOMMU设置；
2. SR-IOV切割物理网卡；
3. 部署启用SR-IOV的网络环境；
4. 创建云主机并加载VF网卡；
5. 为云主机安装VF网卡驱动。

使用SR-IOV网卡前，请务必确保所有准备工作已完成且准确无误。以下详细介绍SR-IOV功能的操作步骤：

操作步骤

1. 物理机启用IOMMU设置

   确保物理机BIOS已开启Intel VT-d / AMD IOMMU配置和SR-IOV配置的前提下，在ZStack云平台开启物理机IOMMU设置。

   - 新添加物理机：在**硬件设施** > **物理机**界面添加物理机过程，勾选**扫描物理机IOMMU设置**配置，添加物理机的同时开启IOMMU设置。如图 *2:* 新添加物理机并启用*IOMMU*设置所示：

     图 2: 新添加物理机并启用IOMMU设置

- 已添加物理机：在物理机详情页，启用**IOMMU启用状态**配置，针对已添加物理机开启IOMMU设置，重启物理机生效。如图 *3: 已添加物理机启用IOMMU设置*所示：

图 3: 已添加物理机启用IOMMU设置

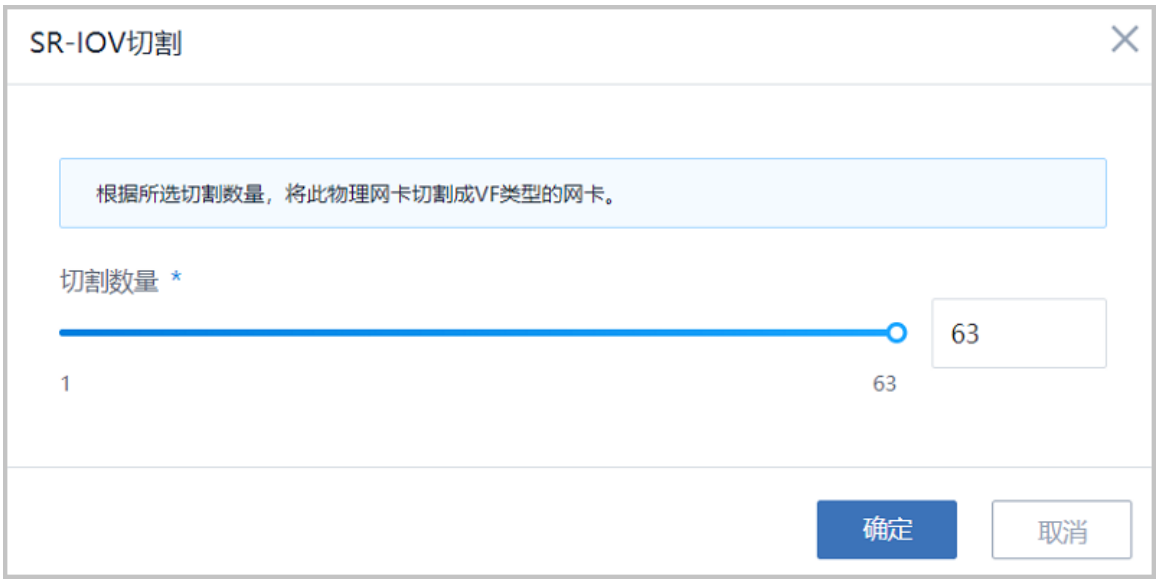
注: 物理机开启IOMMU设置后，还需在物理机详情页确保**IOMMU就绪状态**为可用，否则也无法正常使用SR-IOV功能。若IOMMU启用状态为启用，但IOMMU就绪状态不可用，可能以下原因：

- 首次开启IOMMU设置，但未重启物理机，请手动重启物理机。
- 物理机配置错误，请进入物理机BIOS并开启Intel VT-d / AMD IOMMU配置。

2. SR-IOV切割物理网卡

确保准备工作完成以及IOMMU状态正常时，即可在物理机详情页的**外接设备**页面，选中某个可虚拟化的物理网卡并点击**操作** > **虚拟化切割**按钮，将物理网卡切割成指定数量的VF网卡。如图 *4: SR-IOV切割*所示：

图 4: SR-IOV切割



物理网卡SR-IOV切割后，可在物理机下详情页查看使用情况，如图 *5: VF网卡使用情况*所示：

图 5: VF网卡使用情况

注: 点击**操作** > **SR-IOV还原**按钮支持将VF网卡还原成物理网卡。此时，当前物理网卡切割成的VF网卡正在被云主机使用，SR-IOV还原将同时从云主机卸载相关网卡，请谨慎操作。

3. 部署启用SR-IOV的网络环境

确保二层网络使用的物理网卡已进行SR-IOV切割，即可部署启用SR-IOV的网络环境，包括以下几个步骤：

1. 创建启用SR-IOV功能的二层网络：创建L2NoVlanNetwork、L2VlanNetwork类型的二层网络时，可选择是否启用SR-IOV。启用后，该二层网络下的所有三层网络将支持启用SR-IOV。

   注: 二层网络启用SR-IOV功能需要注意以下情况：

   - VXLAN类型的二层网络暂不支持使用SR-IOV功能。
   - 二层网络使用的物理网卡未进行SR-IOV切割，即使勾选启用SR-IOV按钮，SR-IOV功能并不能生效。
   - 若二层网络已创建，可在二层网络详情页修改SR-IOV启用状态。

   如图 *6:* 创建二层网络所示，创建L2NoVlanNetwork、L2VlanNetwork类型的二层网络时，勾选**启用SR-IOV**配置。

   图 6: 创建二层网络

2. 创建支持SR-IOV功能的三层网络：创建三层网络（公有网络/扁平网络/云路由/VPC）时须加载启用SR-IOV的二层网络，该三层网络将继承SR-IOV属性，可自定义选择是否启用SR-IOV功能。

   如图 7: 创建三层网络所示，创建三层网络时，选择已启用SR-IOV功能的二层网络。

   图 7: 创建三层网络

4. 创建云主机并加载VF网卡

启用SR-IOV的网络环境部署完成后，即可使用此三层网络创建云主机并加载VF网卡。

如图 *8:* 创建云主机所示，创建云主机选择支持SR-IOV功能的三层网络，并勾选**启用SR-IOV**按钮。

图 8: 创建云主机

注: 创建加载VF网卡的云主机，需要注意以下情况：

- 同一云主机支持加载多个VF网卡，且VF网卡和vNIC网卡支持互联互通。
- 若SR-IOV按钮置灰，可能因为该三层网络加载的二层网络不支持。
- 若VF网卡数量不足，勾选启用SR-IOV按钮，将导致创建云主机失败。
- 已有云主机可通过加载网卡操作追加VF网卡，在云主机详情页的配置信息页面的网卡列**加载网卡**即可。
- 停止状态的云主机支持将VF网卡切换为vNIC类型，在云主机详情页的配置信息页面的网卡列**设置网卡类型**即可。

如图 *9:* 云主机详情页所示，云主机详情页可查看网卡类型。

图 9: 云主机详情页

5. 为云主机安装VF网卡驱动

   相关云主机须安装VF网卡驱动，才能保证VF网卡被云主机识别且正常工作。VF网卡驱动获取以及安装方法请联系网卡提供厂商获取帮助。

后续操作

至此，SR-IOV网卡虚拟化的典型使用流程介绍完毕。

5 典型应用场景

ZStack云平台凭借简单、健壮、弹性、智能的优势，帮助用户快速上云。但受传统虚拟化系统技术限制，Hypervisor或VMM软件层面消耗了大量资源和时间，导致PCIe设备的性能优势无法彻底发挥。

为了消除这一软件瓶颈，ZStack支持SR-IOV功能，多个云主机可以高效共享物理网卡设备，获得与物理设备性能媲美的I/O性能，同时又能减少对物理机CPU资源的消耗。可适用于网络NFV、云游戏、视频流（UDP）等典型应用场景。

5.1 网络功能虚拟化（NFV）

随着云计算、虚拟化等技术逐渐成熟，电信行业也对传统网络领域的架构进行了虚拟化变革，采用NFV设备加速完成软件化转变，方便网络设备中的应用程序可以进行大规模虚拟化部署，以节省成本并提高灵活性，有效提升竞争力。

但这些应用程序对网络的吞吐、转发、数据包处理等能力有极高要求，需要使用高性能的虚拟网络才能发挥作用。因此，具备SR-IOV功能的网卡凭借其成熟性、可虚拟化等特点在NFV设备中得到大量应用，使得数据中心能够以较低的成本获得高性能、易伸缩的弹性网络管理能力。

5.2 云游戏

随着宽带网络的发展，以及移动终端设备的普及，将游戏计算至于云端，客户端仅仅负责显示与控制的游戏模式也悄然开始流行。对于一些实时性能要求较高的游戏来说，网络包转发、瞬间吞吐能力、延迟稳定性等网络性能能必不可少，否则可能出现画面掉帧、操作延时等情况，严重影响游戏体验。

这种云游戏模式下，可以借助物理网卡稳定的网络性能，通过SR-IOV技术将VF直接分配给云主机使用。云主机虚拟网卡的流量直接发送给VF，减去中间的桥接网卡或openswitch等软交换机，明显减少网卡包量的损耗，为用户提供更好的游戏体验。

5.3 视频流

随着Internet的发展，多媒体信息在网上的传输越来越重要，为了保证传输速度，基于UDP协议的视频流被广泛使用。对于一些需要实时传输视频的场景（直播、视频会议等），网络包转发、瞬间吞吐能力、延迟稳定性等网络性能能必不可少，否则可能出现画面卡顿、视频分辨率降低等情况，严重影响视频观感。

众所周知，UDP协议虽然能提高传输速度、极大缩短传输时间，但无法保证数据可靠性。网络波动、网络丢包等直接影响视频质量，欲速则不达。具备SR-IOV功能的网卡凭借其成熟性、可虚拟化等特点在实时视频场景能够发挥独特优势，保证网络性能的同时，还能降低CPU压力，有效减少网络丢包，提高传输效率，完美弥补UDP协议可靠性问题，可谓相当契合。

术语表

区域（Zone）

ZStack中最大的一个资源定义，包括集群、二层网络、主存储等资源。

集群（Cluster）

一个集群是类似物理主机（Host）组成的逻辑组。在同一个集群中的物理主机必须安装相同的操作系统（虚拟机管理程序，Hypervisor），拥有相同的二层网络连接，可以访问相同的主存储。在实际的数据中心，一个集群通常对应一个机架（Rack）。

管理节点（Management Node）

安装系统的物理主机，提供UI管理、云平台部署功能。

计算节点（Compute Node）

也称之为物理主机（或物理机），为云主机实例提供计算、网络、存储等资源的物理主机。

主存储（Primary Storage）

用于存储云主机磁盘文件的存储服务器。支持本地存储、NFS、 Ceph、Shared Mount Point、Shared Block类型。

镜像服务器（Backup Storage）

也称之为备份存储服务器，主要用于保存镜像模板文件。建议单独部署镜像服务器。支持ImageStore、Sftp（社区版）、Ceph类型。

镜像仓库（Image Store）

镜像服务器的一种类型，可以为正在运行的云主机快速创建镜像，高效管理云主机镜像的版本变迁以及发布，实现快速上传、下载镜像，镜像快照，以及导出镜像的操作。

云主机（VM Instance）

运行在物理机上的虚拟机实例，具有独立的IP地址，可以访问公共网络，运行应用服务。

镜像（Image）

云主机或云盘使用的镜像模板文件，镜像模板包括系统云盘镜像和数据云盘镜像。

云盘（Volume）

云主机的数据盘，给云主机提供额外的存储空间，共享云盘可挂载到一个或多个云主机共同使用。

计算规格（Instance Offering）

启动云主机涉及到的CPU数量、内存、网络设置等规格定义。

云盘规格（Disk Offering）

创建云盘容量大小的规格定义。

二层网络（L2 Network）

二层网络对应于一个二层广播域，进行二层相关的隔离。一般用物理网络的设备名称标识。

三层网络（L3 Network）

云主机使用的网络配置，包括IP地址范围、网关、DNS等。

公有网络（Public Network）

由因特网信息中心分配的公有IP地址或者可以连接到外部互联网的IP地址。

私有网络（Private Network）

云主机连接和使用的内部网络。

L2NoVlanNetwork

物理主机的网络连接不采用Vlan设置。

L2VlanNetwork

物理主机节点的网络连接采用Vlan设置，Vlan需要在交换机端提前进行设置。

VXLAN网络池（VXLAN Network Pool）

VXLAN网络中的 Underlay 网络，一个 VXLAN 网络池可以创建多个 VXLAN Overlay 网络（即 VXLAN 网络），这些 Overlay 网络运行在同一组 Underlay 网络设施上。

VXLAN网络（VXLAN）

使用 VXLAN 协议封装的二层网络，单个 VXLAN 网络需从属于一个大的 VXLAN 网络池，不同 VXLAN 网络间相互二层隔离。

云路由（vRouter）

云路由通过定制的Linux云主机来实现的多种网络服务。

安全组（Security Group）

针对云主机进行第三层网络的防火墙控制，对IP地址、网络包类型或网络包流向等可以设置不同的安全规则。

弹性IP（EIP）

公有网络接入到私有网络的IP地址。

快照（Snapshot）

某一时间点某一磁盘的数据状态文件。包括手动快照和自动快照两种类型。