



Red Hat Virtualization 4.0

部署 SR-IOV 时需要考虑的硬件因素

在 Red Hat Virtualization 中部署 SR-IOV 时需要考虑的硬件因素

Red Hat Virtualization Documentation TeamRed Hat

Red Hat Virtualization 4.0 部署 SR-IOV 时需要考虑的硬件因素

在 Red Hat Virtualization 中部署 SR-IOV 时需要考虑的硬件因素

Red Hat Virtualization Documentation Team
Red Hat Customer Content Services
rhev-docs@redhat.com

法律通告

Copyright © 2017 Red Hat, Inc.

The text of and illustrations in this document are licensed by Red Hat under a Creative Commons Attribution–Share Alike 3.0 Unported license ("CC-BY-SA"). An explanation of CC-BY-SA is available at

<http://creativecommons.org/licenses/by-sa/3.0/>

. In accordance with CC-BY-SA, if you distribute this document or an adaptation of it, you must provide the URL for the original version.

Red Hat, as the licensor of this document, waives the right to enforce, and agrees not to assert, Section 4d of CC-BY-SA to the fullest extent permitted by applicable law.

Red Hat, Red Hat Enterprise Linux, the Shadowman logo, JBoss, OpenShift, Fedora, the Infinity logo, and RHCE are trademarks of Red Hat, Inc., registered in the United States and other countries.

Linux ® is the registered trademark of Linus Torvalds in the United States and other countries.

Java ® is a registered trademark of Oracle and/or its affiliates.

XFS ® is a trademark of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries.

MySQL ® is a registered trademark of MySQL AB in the United States, the European Union and other countries.

Node.js ® is an official trademark of Joyent. Red Hat Software Collections is not formally related to or endorsed by the official Joyent Node.js open source or commercial project.

The OpenStack ® Word Mark and OpenStack logo are either registered trademarks/service marks or trademarks/service marks of the OpenStack Foundation, in the United States and other countries and are used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation, or the OpenStack community.

All other trademarks are the property of their respective owners.

摘要

本指南介绍了在 Red Hat Enterprise Linux 中部署 SR-IOV 时需要考虑的硬件因素，以及在 Red Hat Virtualization 中的设备分配信息。

目录

1. 简介 2

 1.1. 部署 SR-IOV 时需要考虑的硬件因素概述 2

2. 使用设备分配功能需要格外考虑的硬件因素 2

 2.1. 设备分配功能所需的硬件考虑因素 3

1. 简介

SR-IOV (Single Root I/O Virtualization - 单根 I/O 虚拟化) 是一个硬件架构系统, 它可以使一个单一的 PCI Express (PCIe) 端点作为多个独立的设备使用。这是通过使用两个 PCIe function - PF (Physical Function - 物理 function) 和 VF (Virtual Function - 虚拟 function) 实现的。

PF 是包括了 SR-IOV 功能的传统 PCIe function, 它具有对 PCIe 设备的完全配置和控制功能 (包括数据移动)。每个 PCIe 设备可以有一个到 8 个独立的 PF。

VF 是一组简化的 PCIe function, 它包括了数据移动所需的资源, 以及一组最小化的配置资源。每个 PF 上都可以创建多个 VF, 而每个 PF 所支持的 VF 数量可以不同。但是, 设备所允许的 VF 总量是由 PCIe 设备的厂商决定的, 不同设备对 VF 总量的限制可能会有所不同。

PCIe 通过 ARI (Alternative Routing ID Interpretation) 实现对大量 VF 的支持。ARI 会重新解析 PCIe 头数据中的设备号项中的数据, 达到可以支持多于 8 个 function 的目的。这需要 PCIe 设备和它的上一级端口 (root 端口或 switch) 都支持 ARI。

系统固件 (BIOS 或 UEFI) 会为 PCIe 拓扑分配包括内存、I/O 端口、PCI 总线号范围在内的资源。因此, 系统固件需要支持并启动对 SR-IOV 的支持。

1.1. 部署 SR-IOV 时需要考虑的硬件因素概述

- ✦ 固件 (BIOS 或 UEFI) 需要支持 SR-IOV。检查扩展是否被默认启用, 如果没有启用, 则需要手工启用它。这和启用虚拟化扩展 (VT-d 或 AMD-Vi) 类似。请参阅厂商的手册来获得相关信息。
- ✦ root 端口或 PCIe 设备的上一级端口 (PCIe switch) 需要支持 ARI。
- ✦ PCIe 设备需要支持 SR-IOV。

请参阅相关硬件厂商的文档来确认所使用的硬件满足这些要求。

使用 `lspci -v` 命令可以输出已在系统上安装的 PCI 设备信息。

2. 使用设备分配功能需要格外考虑的硬件因素

设备分配 (device assignment) 功能提供了把一个虚拟客户机直接分配到一个 PCIe 设备的能力, 从而使客户机获得完全访问的能力, 以及近乎于原生的性能。当此功能和 SR-IOV 一起使用时, 一个虚拟客户机将可以直接分配到一个 VF。这意味着, 多个虚拟客户机可以分配到一个 PCIe 设备中的多个 VF。

把虚拟机直接分配到 PCIe 设备并不需要启用 SR-IOV, 而设备分配也不是创建 VF 的唯一应用。但是, 一起使用这两个功能会为系统带来好处。如果使用它们, 则需要考虑额外的硬件因素。

设备分配功能需要 CPU 和固件中的 IOMMU (I/O Memory Management Unit - I/O 内存管理单元) 的支持。IOMMU 会在 I/O 虚拟地址 (IOVA) 和物理内存地址间进行转换, 从而使虚拟客户机使用客户机的物理地址对设备进行编程, IOMMU 将会把这些地址转换为主机的物理地址。

IOMMU 组是由一组设备组成的一个最小的 IOMMU “颗粒”, 它与系统中的其它 IOMMU 组隔离开。通过 IOMMU 组, IOMMU 就可以在进行 IOMMU 组外设备和 IOMMU 内设备的 DMA (Direct Memory Access - 直接内存访问) 操作时, 区分出与 IOMMU 组相关的操作。

分离虚拟客户机操作和 PCIe 设备的 VF 是设备分配功能的基础。PCIe 中定义的 ACS (Access Control Service - 访问控制服务) 和服务器规格是实现 IOMMU 组隔离的硬件标准。如果没有原生的 ACS (或硬件厂商提供的相应功能), IOMMU 组中的任何有问题的设备都会产生一个潜在的风险: 在 IOMMU 保护外的 function 间暴露点到点 (peer-to-peer) 的 DMA。这就相当于扩展了 IOMMU 组来包括那些缺少恰当隔离保护

的 function。

另外，我们还推荐服务器的根端口（root port）也要提供原生的 ACS 支持。否则，在这些端口上安装的设备会被分为一个组。根端口有两个不同类型，一个是基于处理器（northbridge）的根端口，另一个是基于控制器中心（southbridge）的根端口。如果设备分配功能和 SR-IOV 一起使用，虚拟客户机被分配到 VF，则以上所提到的端口都需要支持 ACS 和 ARI。

Intel 的 Xeon Processor E5 系列、Xeon Processor E7 系列和高端桌面处理器都包括基于处理器的根端口上的原生 ACS 支持。

在通常情况下，Intel 设备不包括基于控制器中心的根设备上的原生 ACS 支持。但是，Red Hat Enterprise Linux 7.2 内核提供了在 X99、X79、5 系列到 9 系列芯片组的这些根端口上实现与 ACS 的功能相应的隔离功能。

在安装 PCIe 设备时，为了确保根端口支持 ACS，请参阅硬件厂商的相关文档。

另外，在 I/O 拓扑中的所有 PCIe 交换机和网桥也需要支持 ACS，否则，它可能会扩展 IOMMU 组。

2.1. 设备分配功能所需的硬件考虑因素

- ✧ CPU 需要支持 IOMMU（如 VT-d 或 AMD-Vi）。IBM POWER8 默认支持 IOMMU。
- ✧ 固件需要支持 IOMMU。
- ✧ 使用的 CPU root 端口需要支持 ACS（或与 ACS 相应的功能）。
- ✧ PCIe 设备需要支持 ACS（或与 ACS 相应的功能）。
- ✧ 另外，还推荐 PCIe 设备和 root 端口间的所有 PCIe 交换机和网桥都需要支持 ACS。如果一个交换机不支持 ACS，这个交换机后面的所有设备共享相同的 IOMMU 组，则只能分配到相同的虚拟机。

请参阅相关硬件厂商的文档来确认所使用的硬件满足这些要求。

使用 **lspci -v** 命令可以输出已在系统上安装的 PCI 设备信息。