# 27. Vhost 库

Vhost库实现了一个用户空间virtio网络服务器，允许用户直接操作virtio。 换句话说，它允许用户通过VM virtio网络设备获取/发送数据包。 为了达到这个功能，一个vhost库需要实现：

访问guest内存：

对于QEMU，这是通过使用 -object memory-backend-file,share=on,... 选项实现的。 这意味着QEMU将创建一个文件作为guest RAM。 选项 share=on 允许另一个进程映射该文件，这意味着该进程可以访问这个guest RAM。  
知道关于vring所有必要的信息：

诸如可用环形存储链表的存储空间。Vhost定义了一些消息（通过Unix套接字传递）来告诉后端所有需要知道如何操作vring的信息。

## 27.1. Vhost API 概述

以下是一些关键的Vhost API函数概述：

rte\_vhost\_driver\_register(path, flags)  
此函数将vhost驱动程序注册到系统中。path 指定Unix套接字的文件路径。  
当前支持的flags包括：

* + RTE\_VHOST\_USER\_CLIENT  
    当使用该flag时，DPDK vhost-user 作为客户端。 请参阅以下说明。
  + RTE\_VHOST\_USER\_NO\_RECONNECT  
    当 DPDK vhost-user 作为客户端时，它将不断尝试连接到服务端（QEMU），知道成功。 这在以下两个情况中是非常有用的：
    - 当 QEMU 还没启动时
    - 当 QEMU 重启时(如guset OS 重启)  
      这个重新连接选项是默认启用的，但是，可以通过设置这个标志来关闭它。
  + RTE\_VHOST\_USER\_DEQUEUE\_ZERO\_COPY  
    设置此flag时将启用出队了零复制。默认情况下是禁用的。  
    在设置此标志时，需要知道以下原则：
    - 零拷贝对于小数据包（小于512）是不好的。
    - 零拷贝对VM2VM情况比较好。对于两个虚拟机之间的ipref，提升性能可能高达70%(当TSO使能时).
    - 对于VM2NIC情况，nb\_tx\_desc 必须足够小：如果未启动virtio间接特性则 <=64，否则 <= 128。  
      这是因为，当启用出队列零拷贝时，只有当相应的mbuf被释放时，客户端TX使用的vring才会被更新。 因此，nb\_tx\_desc必须足够小，以便PMD驱动程序将耗尽可用的TX描述符，并及时释放mbufs。 否则，guset TX vring将无mbuf使用。
    - Guest的内存应该使用应该使用huge page支持以获得更好的性能。最好使用1G大小的页面。  
      当启用出队零拷贝时，必须建立guest 物理地址和host物理地址之间的映射。 使用non-huge page则意味着更多的页面细分。 为了简单起见，DPDK vhost对这些段进行了线性搜索，因此，段越少，我们得到的映射就越快。 注意：将来我们可能使用树搜索来提升速度。

rte\_vhost\_driver\_set\_features(path, features)  
此函数设置vhost-user驱动支持的功能位。 vhost-user驱动可以是vhost-user net，但也可以是其他的，例如vhost-user SCSI。

rte\_vhost\_driver\_callback\_register(path, vhost\_device\_ops)  
此函数注册一组回调函数，以便在发生某些事件时让DPDK应用程序采取适当的操作。 目前支持以下事件：

* + new\_device(int vid)  
    这个回调在virtio设备准备就绪时调用，vid 是虚拟设备ID。
  + destroy\_device(int vid)  
    当virtio设备关闭时（或vhost连接中断），调用此函数处理。
  + vring\_state\_changed(int vid, uint16\_t queue\_id, int enable)  
    当特定队列的状态发生改变，如启用或禁用，将调用此回调。
  + features\_changed(int vid, uint64\_t features)  
    这个函数在feature改变时被调用。例如，VHOST\_F\_LOG\_ALL 将分别在实时迁移的开始/结束时设置/清除。

rte\_vhost\_driver\_disable/enable\_features(path, features))  
该函数禁用或启用某些功能。例如，可以使用它来禁用可合并的缓冲区和TSO功能，这两个功能默认都是启用的。

rte\_vhost\_driver\_start(path)  
这个函数触发vhost-user协商。它应该在初始化一个vhost-user驱动程序结束时被调用。

rte\_vhost\_enqueue\_burst(vid, queue\_id, pkts, count)  
传输（入队）从host到guest的 count 包。

rte\_vhost\_dequeue\_burst(vid, queue\_id, mbuf\_pool, pkts, count)  
接收（出队）来自guest的 count 包，并将它们存储在 pkts。

## 27.2. Vhost-user 实现

Vhost-user 使用Unix套接字来传递消息。这意味着DPDK vhost-user的实现具有两种角色：

DPDK vhost-user作为server：  
DPDK 将创建一个Unix套接字服务器文件，并监听来自前端的连接。  
注意，这是默认模式，也是DPDK v16.07之前的唯一模式。

DPDK vhost-user最为client：  
与服务器模式不同，此模式不会创建套接字文件; 它只是试图连接到服务器（而不是创建文件的响应）。

当DPDK vhost-user应用程序重新启动时，DPDK vhost-user将尝试再次连接到服务器。这是“重新连接”功能的工作原理。

注意：“重连” 功能需要 QEMU v2.7 及以上的版本。  
vhost支持的功能在重新启动之前和之后必须完全相同。例如，如果TSO被禁用，但是重启之后被启用了，将导致未定义的错误。

无论使用哪种模式，建立连接之后，DPDK vhost-user 都将开始接收和处理来自QEMU的vhost消息。

对于带有文件描述符的消息，文件描述符可以直接在vhost进程中使用，因为它已经被Unix套接字安装了。

当前支持的vhost 消息包括：

* VHOST\_SET\_MEM\_TABLE
* VHOST\_SET\_VRING\_KICK
* VHOST\_SET\_VRING\_CALL
* VHOST\_SET\_LOG\_FD
* VHOST\_SET\_VRING\_ERR

对于 VHOST\_SET\_MEM\_TABLE 消息，QEMU将在消息的辅助数据中为每个存储区域及其文件描述符发送信息。 文件描述符用于映射该区域。

VHOST\_SET\_VRING\_KICK 用作将vhost设备放入数据面的信号， VHOST\_GET\_VRING\_BASE 用作从数据面移除vhost设备的信号。

当套接字连接关闭，vhost将销毁设备。

## 27.3. 支持Vhost的vSwitch

有关更多vhost详细信息以及如何在vSwitch中支持vhost，请参阅《DPDK Sample Applications Guide》。