Table des matières

[chapter 4: Contrail DPDK vrouter setup 2](#_Toc51017312)

[DPDK vRouter physical network interface 2](#_Toc51017313)

[DPDK vRouter CPU setup 3](#_Toc51017314)

[CPU kept for Linux Operating System 4](#_Toc51017315)

[CPU allocated to the DPDK vRouter 5](#_Toc51017316)

[CPU allocated to virtual machines 6](#_Toc51017317)

[vRouter memory setup 7](#_Toc51017318)

[Hugepage memory configuration on the compute node operating system 7](#_Toc51017319)

[Hugepage allocation for the DPDK vrouter 7](#_Toc51017320)

[DPDK physical interface rings setup 8](#_Toc51017321)

[DPDK vrouter internal queues rings setup 9](#_Toc51017322)

[DPDK virtual machine interface rings setup 11](#_Toc51017323)

[Virtual Machine vif multi-queue setup 12](#_Toc51017324)

[vRouter routing and switching object tables dimensioning parameters 13](#_Toc51017325)

# chapter 4: Contrail DPDK vrouter setup

Contrail DPDK vrouter setup mainly consist to:

* Define NIC cards to be used by the vRouter for its interconnexion with the physical network
* Define CPU resources to be allocated to the DPDK vRouter
* Define the huge pages memory to be used by the DPDK vRouter to create vrouter interface DPDK rings for physical and virtual machine network interface cards.
* Configure the number of queues of DPDK vrouter physical and virtual machine network interface cards.

Contrail vRouter DPDK specific setup is defined into /etc/sysconfig/network-scripts/ifcfg-vhost0 configuration file. Then the vrouter agent vhost0 network interface has to be restarted to get modified setup enforced:

$ sudo ifdown vhost0

$ sudo ifup vhost0

## DPDK vRouter physical network interface

Only one physical interface can be plugged onto the vif0/0 port of the vrouter. Usually for resiliency purpose, a bond interface is created to group two physical interfaces in a single entity which is plugged onto the vrouter.

Physical NICs used into the bond interface are defined in BIND\_INT parameter:

$ vi /etc/sysconfig/network-scripts/ifcfg-vhost0

DEVICE=vhost0

DEVICETYPE=vhost

TYPE=dpdk

BIND\_INT=0000:02:01.0,0000:02:02.0

Using the following command we can display PCI identifier of physical interfaces which are available onto the Linux Operating system:

$ sudo lshw -class network | grep pci@

bus info: pci@0000:02:01.0

bus info: pci@0000:02:02.0

bus info: pci@0000:03:00.0

Once the Contrail DPDK vrouter has been started, we can see the actual physical interfaces used for the underlay network interconnexion:

$ sudo docker exec contrail-vrouter-agent-dpdk /opt/contrail/bin/dpdk\_nic\_bind.py -s

Network devices using DPDK-compatible driver

============================================

0000:02:01.0 '82540EM Gigabit Ethernet Controller' drv=uio\_pci\_generic unused=e1000

0000:02:02.0 '82540EM Gigabit Ethernet Controller' drv=uio\_pci\_generic unused=e1000

Network devices using kernel driver

===================================

0000:03:00.0 'Virtio network device' if= drv=virtio-pci unused=virtio\_pci,uio\_pci\_generic

Other network devices

=====================

<none>

## DPDK vRouter CPU setup

When DPDK is used, some CPUs have to be booked for an exclusive use to the DPDK application for packet polling purpose. CPU allocation planning is a very important task to be done by the customer architecture team in charge of the virtual infrastructure.

On each compute node we have to allocate:

* Some CPUs to be kept available for the Linux Operating System
* Some CPUs for the Virtual Machines. Generally, this is the main purpose of the customer virtual infrastructure creation.
* Some CPUs for the vrouter high speed packet processing.

The customer virtual infrastructure architect is starting to define the number of CPUs to be allocated in each group described above:

![Compute Node CPU capacity planning](data:image/png;base64;base64,)

In order to get the best performance, CPUs allocated to Virtual Machines and to vRouter have to be isolated from those that are kept to Linux Operating system. CPU isolation is the first setup to be done to define the CPUs that will no more be used by the Linux Operating system. Those CPUs will be dedicated to DPDK vrouter or used by OpenStack Nova to spawn virtual machines.

Below is the CPU core topology of a 2 sockets system with 2\*12 physical cores, with hyper-threading enabled:

NUMA node0 CPU(s):

PHY cores: **0 2 4 6 8 10 12 14 16 18 20 22**

HT cores : **24 26 28 30 32 34 36 38 40 42 44 46**

NUMA node1 CPU(s):

PHY cores: **1 3 5 7 9 11 13 15 17 19 21 23**

HT cores : **25 27 29 31 33 35 37 39 41 43 45 47**

This topology will be used in the configuration examples provided in next sections.

### CPU kept for Linux Operating System

By default, all CPUs are included in the group of CPUs available for the Operating System needs. These CPUs are told: *“isolated”* because they are no more used to process all tasks. In order to get a CPU isolated several mechanisms can be used:

* remove this CPU from the “common” CPU list used to process all tasks
* change the scheduling algorithm (cooperative to preemptive)
* participate or not to interrupt processing

It is possible to remove some CPUs using ***isolcpus*** kernel parameter. This kernel parameter has to be provisioned at the system startup. GRUB configuration is updated to define *isolcpus* parameter and the system restarted.

In the example below we are keeping only CPU 0,1,24 and 25 for the Linux Operating System:

$ vi /etc/default/grub

GRUB\_CMDLINE\_LINUX="console=tty0 console=ttyS0,115200n8 crashkernel=auto rhgb quiet default\_hugepagesz=1GB hugepagesz=1G hugepages=28 iommu=pt intel\_iommu=on isolcpus=**2-23,26-47**"

$ grub2-mkconfig -o /etc/grub2.cfg

We also need to specify the CPUs that have to be used by the Linux Operating system in Systemd configuration file:

$ vi /systemd/system.conf

CPUAffinity=**0-1,24-25**

$ sudo systemctl deamon-reexec

$ sudo systemctl system.slice restart

When a RedHat Linux Operating System is used, it’s also recommended to configure tuned to get a stronger CPU isolation.

$ vi /etc/tuned/cpu-partitioning-variables.conf

**isolated\_cores=2-23,26-47**

$ sudo systemctl restart tuned

This is important to keep enough CPUs for the Linux Operating system. Not isolated CPUs are used by all tasks started and managed by the Linux Operating system, they are:

* System configuration and control tasks
* Contrail vRouter agent (SDN control plane)
* Hypervisor configuration and control tasks (Virtual Machine configuration for instance)

### CPU allocated to the DPDK vRouter

#### Packet polling and processing threads

DPDK vrouter speed is depending on the number of CPUs allocated for packet polling and processing. There is a trade-off to be found by each customer on how many CPU he will use for its own applications running on virtual machines and how many CPUs he will book for the vrouter to increase network packets processing speed.

We have first to define how many CPUs will be booked for DPDK vrouter polling and packet processing threads. We can first consider that each allocated CPU will bring 1 MPPS packets network processing speed to the vrouter. This 1 MPPS value is depending on lots of factors, mainly CPU speed, NUMA usage, packet size. It can range between 0.8 MPPS to 1.6 MPPS.

A Kernel mode vRouter is generally providing 1 MPPS packet speed in the best case, more generally Kernel mode vRouter is providing less than 1 MPPS.

With DPDK vRouter we are usually allocating 4 or 8 network packets processing CPU. We can expect around 4 to 12 MPPS packet processing performance.

A lower CPU number (1 to 3) is generally not relevant as it does not really bring a big difference with the regular Kernel mode. A higher CPU number (more than 8) is not bringing much more performance due to some side effects, especially the multi queue setup it would require on virtual machines (cf: multi queue section).

CPU allocated to packet polling and processing are defined into CPU\_LIST parameter. This CPU parameter can use two different syntaxes: mask or list.

Here 8 CPUs are allocated to the vRouter for packet polling and processing:

$ vi /etc/sysconfig/network-scripts/ifcfg-vhost0

**CPU\_LIST**=2,4,6,8,26,28,30,32

PS: mask for CPUs 2,4,6,8,26,28,30,32 maps to binary value:  
b0000 0000 0000 0001 0101 0100 0000 0000 0000 0001 0101 0100 (0x154000154h).

#### vrouter DPDK dataplane configuration and control threads

Two DPDK vrouter parameters are allowing to define CPUs to be allocated for dataplane control and configuration threads:

* ***DPDK\_CTRL\_THREAD\_MASK:*** defines which CPUs will be allocated for DPDK initialization setup.
* ***SERVICE\_CORE\_MASK:*** defines which CPUs will be allocated for vrouter dataplane setup (vrouter interface setup).

DPDK initialization setup is done only at vrouter startup while vRouter dataplane setup task are done at vrouter initialization and each time a new interface is plugged or remove onto the vrouter. Same CPUs can be shared for these two tasks.

Here we are allocating CPU 10 and 34 dataplane control and configuration threads:

$ vi /etc/sysconfig/network-scripts/ifcfg-vhost0

**DPDK\_CTRL\_THREAD\_MASK**=0x400000400

**SERVICE\_CORE\_MASK**=0x400000400

PS: mask for CPUs 10,34 allocated maps to binary value:  
b0000 0000 0000 0100 0000 0000 0000 0000 0000 0100 0000 0000 (0x400000400h).

### CPU allocated to virtual machines

Host compute CPUs used for user virtual machines are defined into Nova configuration file. Below, we are enforcing Nova CPU assignment:

$ openstack-config --set /etc/nova/nova.conf DEFAULT vcpu\_pin\_set 3,5,7,9,11-23,27,29,31,33,35-47

$ cat /etc/nova/nova.conf | grep vcpu\_pin\_set

vcpu\_pin\_set=3,5,7,9,11-23,27,29,31,33,35-47

In order to get these changes taken into consideration, Nova compute service has to be restarted. When OpenStack Kolla is used:

$ sudo docker restart nova\_compute

## vRouter memory setup

DPDK library need to get hugepages allocated by the Linux Operating system for vNIC rings setup. This is why, Contrail DPDK vrouter dataplane is requiring:

* Hugepage memory setup on the compute node
* Some of these huge pages to be allocated for the vrouter Physical NIC
* Some of these huge pages to be allocated for the virtual machine NICs
* Allocated hugepages to be visible from both DPDK vrouter application and Virtual Machines.

### Hugepage memory configuration on the compute node operating system

Only small hugepages (2MB) can be configured dynamically configured using sysctl. Bigger hugepages (1GB) must be configured at the system startup.

Following parameters are used:

* **default\_hugepagesz** defines which huge page size is a default (this size will appear in /proc/meminfo, and this size will be mounted by default when pagesize mounting option will not be used)
* **hugepagesz** followed by **hugepages** defines size and amount respectively and the pair can be repeated to configure different sizes of huge pages.

For instance, in order to configure 40 1GB hugepages and 40 2M hugepages at Linux System startup, we are proceeding like described hereafter and we are restarting the system:

$ vi /etc/default/grub

default\_hugepagesz=1GB hugepagesz=1G hugepages=40 hugepagesz=2M hugepages=40

$ grub2-mkconfig -o /etc/grub2.cfg

### Hugepage allocation for the DPDK vrouter

Some of the available operating system hugepages have to be allocated to the vrouter DPDK application to be used to create DPDK rings for the physical NIC.

In order the vrouter DPDK application to be able to use Linux System available hugepages, a hugetlbfs pseudo filesystem needs to be mounted. The following line needs to be added to /etc/fstab.

$ vi /etc/fstab

hugetlbfs on /dev/hugepages type hugetlbfs (rw,relatime,seclabel,pagesize=**1G**)

DPDK vrouter detects the huge pages size hugetlbfs mount point. Here, the DPDK vrouter will try to use 1GB huge pages. If no page size is specified, the DPDK vrouter is assuming 2 MB hugepages have to be used. If no available hugepages of the specified (or 2MB if not size is specified) are available, the contrail DPDK vrouter will fail to start.

Amount of huge page memory requested by the vRouter at startup for its physical NIC DPDK rings setup is specified in ***socket-mem*** parameter.

In order the vrouter to request hugepages memory on the first NUMA socket only, we are using this option with only one parameter:

--socket-mem <value>

In order the vrouter to request hugepages memory on both NUMA0 and NUMA1 socket, we are using this option with only two parameters:

--socket-mem <value>,<value>

It is important to allocate hugepage memory to all NUMA nodes that will have DPDK interfaces associated with them. If memory is not allocated on a NUMA node associated with a physical NIC or VM, they cannot be used. If you are using 2 or more ports from different NICs, it is best to ensure that these NICs are on the same CPU socket.

Here we are configuring the vRouter to request 1GB hugepages memory on both NUMA nodes:

$ vi /etc/sysconfig/network-scripts/ifcfg-vhost0

DPDK\_COMMAND\_ADDITIONAL\_ARGS="--socket-mem 1024,1024"

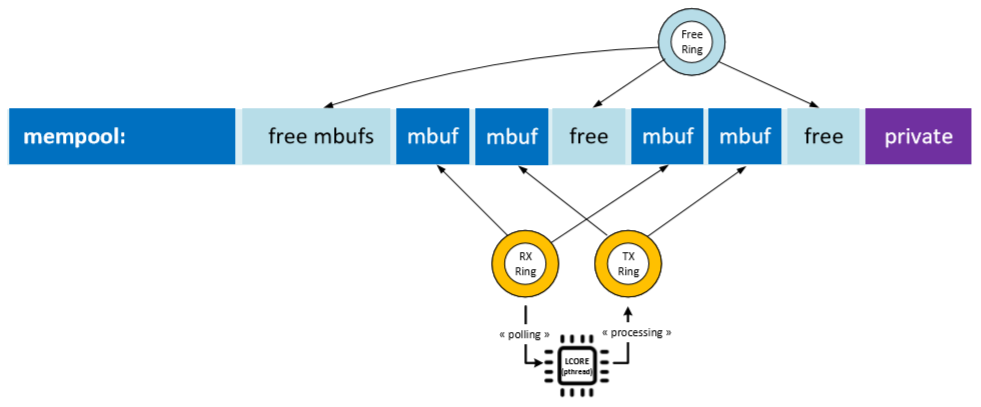
$ sudo ifdown vhost0

$ sudo ifup vhost0

### DPDK physical interface rings setup

In the previous section we’ve described how some hugepages memory is allocated to Contrail DPDK vRouter. This memory is mainly used by the DPDK vrouter to create DPDK rings for the physical interface.

Contrail DPDK vRouter will create 2 DPDK rings for each polling core (which are defined into CPU\_LIST parameter). DPDK rings are circular arrays of RX and TX descriptors that a pointing mbufs in which the packet content is stored. All mbufs for each TX/RX pair are stored into a single mempool memory area, they are representing an interface queue.



Following parameters are used for DPDK vrouter physical NIC configuration:

**--vr\_mempool\_sz** : is used to define mempool memory size. Default value is 16384.

**--dpdk\_txd\_sz** : is used to define Physical NIC TX Ring descriptor size. Default value is 256.

**--dpdk\_rxd\_sz** : is used to define Physical NIC RX Ring descriptor size. Default value is 256.

Here we are configuring the vRouter physical NIC DPDK rings with 512 RX and TX descriptors size and use 32MB mempool:

$ vi /etc/sysconfig/network-scripts/ifcfg-vhost0

DPDK\_COMMAND\_ADDITIONAL\_ARGS="--dpdk\_rxd\_sz 512 --dpdk\_txd\_sz 512 --vr\_mempool\_sz 32768"

$ sudo ifdown vhost0

$ sudo ifup vhost0

PS: Physical NIC DPDK ring size modification can lead to some unexpected side effect (packet loss). Needed mempool size depend on the configured maximum packet size (physical NIC MTU) and also the number of NIC using into the physical bond, and also the configured number of RX and TX descriptors.

### DPDK vrouter internal queues rings setup

In some scenarios, Contrail DPDK vrouter is using a DPDK pipeline model in order to split packet polling and processing task in two different threads. When this DPDK pipeline mode is used, some internal queues are created in order to store packets that have been polled by the polling lcore thread before to be processed by the processing lcore thread.



Two parameters are used for DPDK vrouter internal queue (software rings) configuration:

**--vr\_dpdk\_tx\_ring\_sz**: is used to define forwarding lcores TX Ring descriptor size (1024 by default)

**--vr\_dpdk\_rx\_ring\_sz**: is used to define forwarding lcores RX Ring descriptor size (1024 by default).

Here we are configuring the vRouter internal rings with 2048 RX and TX descriptors:

$ vi /etc/sysconfig/network-scripts/ifcfg-vhost0

DPDK\_COMMAND\_ADDITIONAL\_ARGS="--vr\_dpdk\_rx\_ring\_sz 2048 --vr\_dpdk\_tx\_ring\_sz 2048"

$ sudo ifdown vhost0

$ sudo ifup vhost0

### DPDK virtual machine interface rings setup

Virtual machines NIC queues are not configured by Contrail vRouter, they are managed by OpenStack. By default, Nova is configuring 256 rx and tx descriptor size virtIO interfaces on the virtual machine.

Virtual machines NIC queue size is defined at OpenStack level in */etc/nova/nova.conf* configuration file. They are configured using rx\_queue\_size and tx\_queue\_size parameters.

$ cat /etc/nova/nova.conf | grep x\_queue\_size

rx\_queue\_size=512

tx\_queue\_size=512

In order to get these changes taken into consideration, Nova compute service has to be restarted. When OpenStack Kolla is used:

$ sudo docker restart nova\_compute

Virtual machine NIC and vrouter vif on which each interface is connected to are sharing the same queues (DPDK rings):

* a vrouter vif tx ring is the same as the virtual NIC rx ring it is connected to.
* a vrouter vif rx ring is the same as the virtual NIC tx ring it is connected to.

It avoids to duplicate the same information and to add processing overhead (that would be generated to manage data copy between vrouter vif and the virtual machine queues).

This is why virtual machine NIC queues have to be accessible from both vrouter and the virtual machine it belongs to. Virtual machines have to be created by the QEMU/KVM hypervisor with a specific property which is allowing them to access the host Operating system hugepages and to request hugepages allocations.

Hugepages size to be allocated by the Hypervisor to the virtual machine has to be specified with *hw:mem\_page\_size* property. The configured hugepages memory size must be the same as those used by the DPDK vrouter (defined into huge pages size hugetlbfs mount point).

Here we are configuring an OpenStack flavor named m1.large which define 1GB size hugepages in *hw:mem\_page\_size* property:

$ openstack flavor set m1.large --property hw:mem\_page\_size=1GB

Then this flavor is used at the instance creation:

$ openstack server create --flavor FLAVOR\_ID --image IMAGE\_ID INSTANCE\_NAME

PS: *hw:mem\_page\_size* property can also be defined at image level

$ openstack image set --property hw:mem\_page\_size=1GB < IMAGE\_ID>

## Virtual Machine vif multi-queue setup

As explained earlier, when supported, it is suitable to enable multiqueues on virtual machines NIC. The most suitable scenario is to configure the same number of queues on virtual NIC than the number of polling cores defined on Contrail vRouter.

So, if contrail DPDK vrouter is configured with 4 queues the best scenario is to configure 4 queues on virtual machines network inferfaces.

In OpenStack, in order to get virtual machines configured with multiqueues, we have to enable multiqueue support on virtual instance image. It can be done with following command:

$ openstack image set --property hw\_vif\_multiqueue\_enabled="true" < IMAGE\_ID>

Then this image is used at the instance creation:

$ openstack server create --flavor <FLAVOR\_ID> --image <IMAGE\_ID> <INSTANCE\_NAME>

Then, we an instance is started with multiqueue vif property enabled, each interface is automatically configured with several queues. The number of queues to be enabled on each interface is automatically defined by Nova.

If the compute host (hypervisor node running qemu/kvm) is running Linux Kernel 3.X, the number of queues configured on the virtual machine interface is the same as the number of virtual CPUs configured on the virtual machine but can’t exceed 8 queues. It means for a virtual machine configured with 10 vCPUs, all its virtual network interface cards will be configured with 8 queues when multi queue is enabled.

If the compute host (hypervisor node running qemu/kvm) is running Linux Kernel 4.X, the number of queues configured on the virtual machine interface is the same as the number of virtual CPUs configured on the virtual machine but can’t exceed 256 queues. It means for a virtual machine configured with 10 vCPUs, all its virtual network interface cards will be configured with 10 queues when multi queue is enabled.

As explain earlier, Contrail vRouter is not able to process packets generated by connected virtual network interface cards configured with more queues than the number of CPU defined into its CPU\_LIST (number of polling and processing cores defined on Contrail vRouter).

Consequently, a Contrail vRouter configured with only 4 polling and processing cores won’t be able to collect a Virtual Machine configured with 10 vCPU with vif multiqueue property enable is connected to.

One of the following changes has to be performed:

* disable multiqueue on the Virtual Machine
* add more polling and processing cores on the vRouter (increase to 8 cores instead of only 4)
* decrease the number of queues configured by Nova on the virtual machine

Unfortunately, no mechanism is provided by Nova to specify a specific value for the number of queues to be enabled on virtual machine network interface. Only the described algorithm is proposed.

In order to decrease the number of queues on the virtual machine network interfaces we have to run ethtool command inside this Virtual Machine. For instance, below we are configuring 4 logical queues on eth0 vNIC :

$ sudo ethtool -L eth0 combined 4

Virtual Machine initialization script has to be modified to automatically decrease the default value defined by Nova for the number of queues configured on its network interfaces to a lower one.

This is why, the most efficient setup today is to use Linux Kernel 3.X on OpenStack compute node running QEMU/KVM and to configure 8 CPU into the CPU\_LIST of the Contrail DPDK vRouter.

## vRouter routing and switching object tables dimensioning parameters

Some parameters supported as well as DPDK vrouter than Kernel one are allowing to define the size of internal objects tables. They are:

* **--vr\_flow\_entries**: maximum flow entries (default is 512K)
* **--vr\_oflow\_entries**: maximum overflow entries (default is 8K)
* **--vr\_bridge\_entries**: maximum bridge entries (default is 256K)
* **--vr\_bridge\_oentries**: maximum bridge overflow entries (default is 0)
* **--vr\_mpls\_labels**: maximum MPLS labels used in the node (default is 5K)
* **--vr\_nexthops**: maximum next hops in the node (defaut is 512K)
* **--vr\_vrfs**: maximum VRFs supported in the node (default is 4096)
* **--vr\_interfaces**: maximum interfaces that can be created (default is 4352)

In order to override their default values, we can configured an updated value using DPDK\_COMMAND\_ADDITIONAL\_ARGS parameter defined in vhost0 DPDK vrouter configuration file.

For instance; we can decrease the nexthops table size to 32K instead of 512K configured by default:

$ vi /etc/sysconfig/network-scripts/ifcfg-vhost0

DPDK\_COMMAND\_ADDITIONAL\_ARGS=”--vr\_nexthops=32768”

$ sudo ifdown vhost0

$ sudo ifup vhost0

All these parameters could increase vRouter performances but could also have a bad impact when not properly configured.