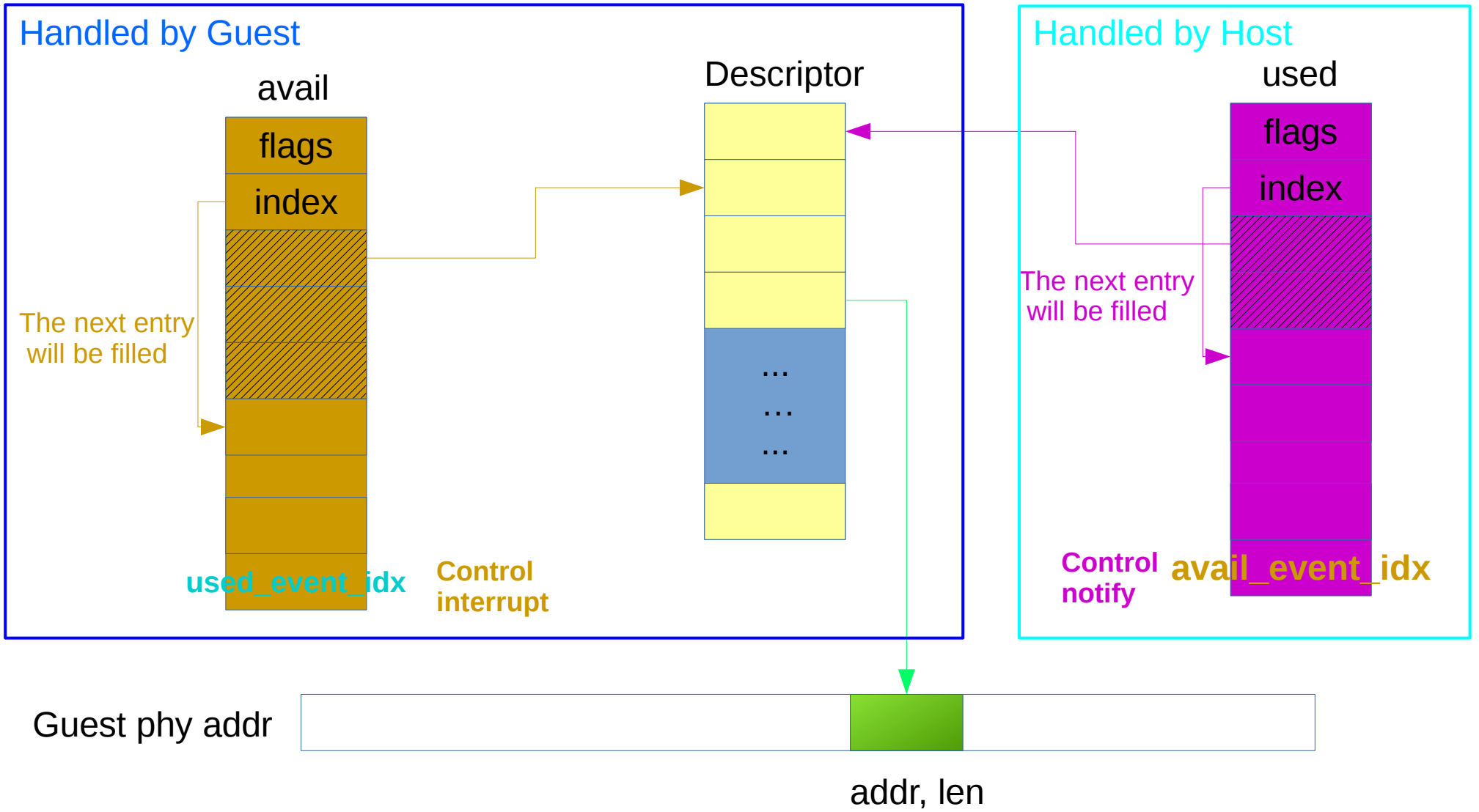
Vhost and vhost-user

Xiao Guangrong
<guangrong.xiao@linux.intel.com>

Index

- Vring
- Vhost
- Vhost-user

Vring


Handled by Guest
avail
flags
index
The next entry
will be filled
used_event_idx
Control
interrupt
Descriptor
...
...
...
Handled by Host
used
flags
index
The next entry
will be filled
Control
notify
avail_event_idx
Guest phy addr
addr, len

Vring interrupt

- Send by host to tell guest new data has been put into 'used queue'
- Controlled by 'flags' or 'used_event_idx' in 'valid queue'
  - If the feature, VIRTIO_F_EVENT_IDX, is enabled, interrupt is triggered if ‘index’ in the 'used' queue reaches used_event_idx + 1.
    - set used_event_idx = index in 'used' queue will tiger interruption after host fills data next time.
    - Set used_event_idx to the first entry where we fetched data out in used queue to disable the interruption (as the data number filled by host can not be more than the vqueue size).
  - Otherwise, if VIRTQ_AVAIL_F_NO_INTERRUPT on flags is cleared, interrupt is triggered whenever new data is written in host side.
- Linux Kernel uses NAPI to suppress interrupt
  - If less then NAPI_POLL_WEIGHT (64) packages is received, NAPI will be turned off and use interrupt instead.

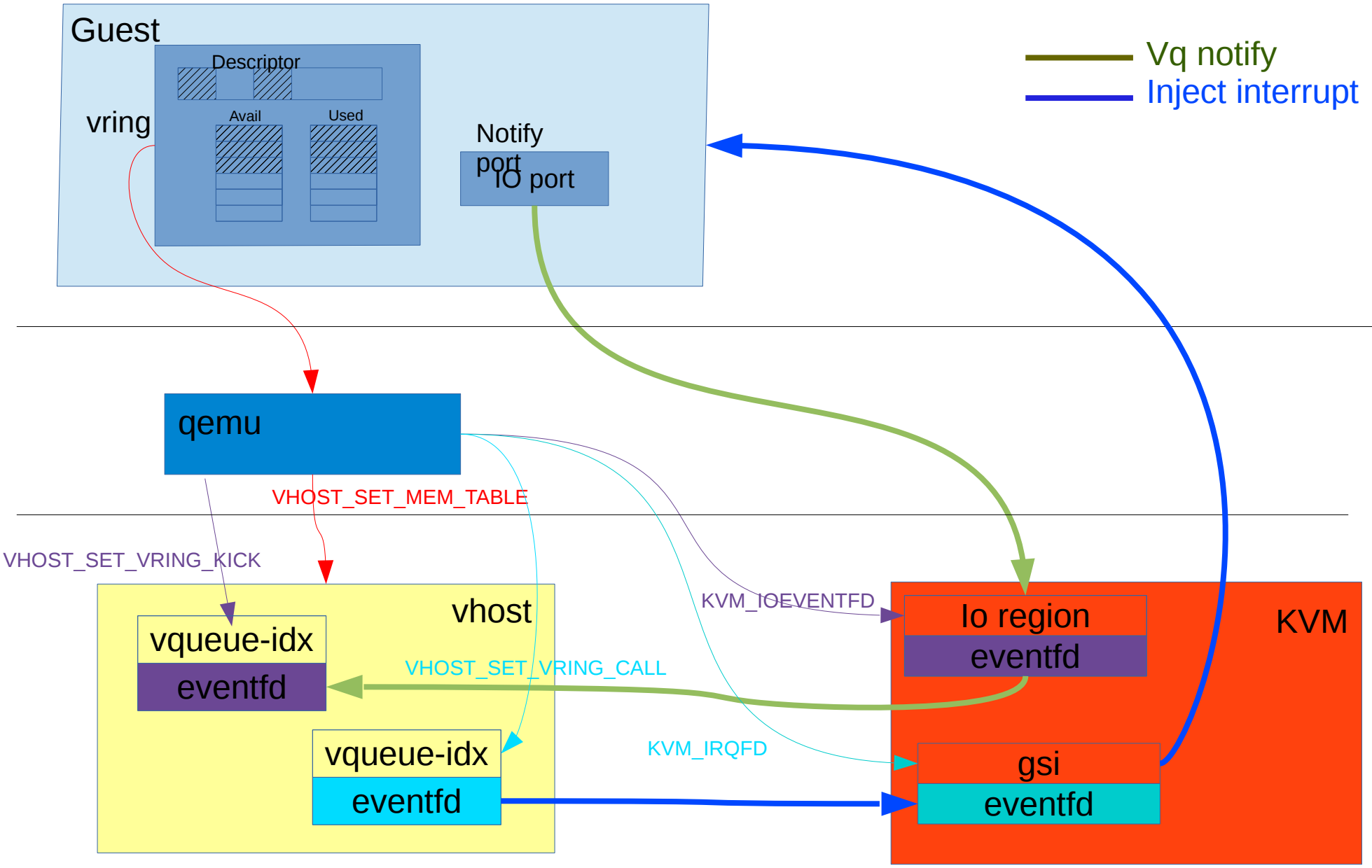
Vring notify

- Send by guest to notify host that the data in avail queue is ready.
- It is controlled by the the similar way of interruption
  - The 'flags' in 'used' queue or 'avail_event_idx' in 'used' queue.
- The notification is suppressed:
  - It is disabled when host is handling a notification. After all received data is drained out it is enabled again.

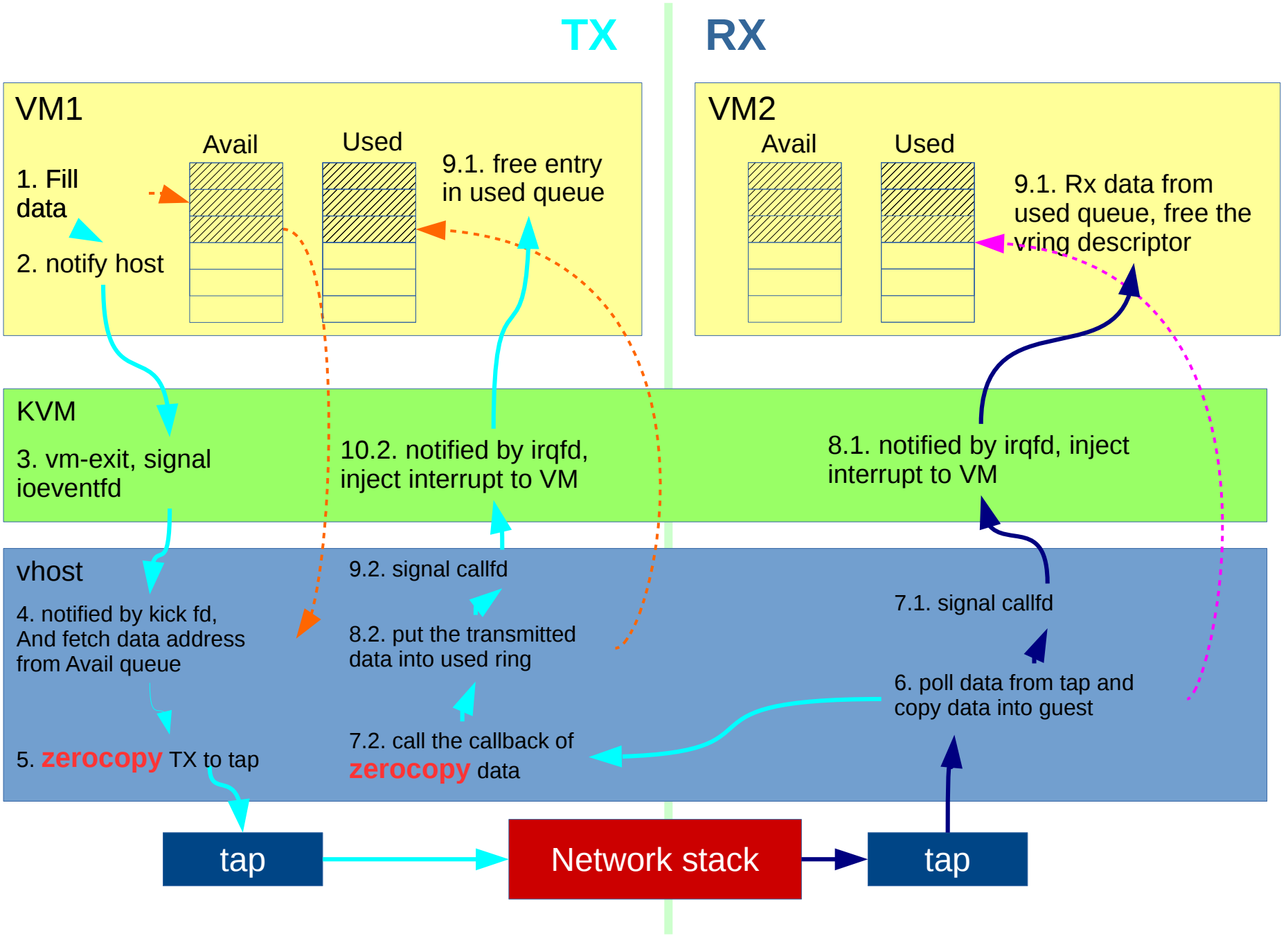
vhost

- Put the datapath into kernel side to reduce context switch
- Sequences
  - a) VHOST_SET_OWNER: set the owner of vhost fd.
  - b) VHOST_SET_MEM_TABLE: tell the guest memory info to vhost.
  - c) VHOST_SET_VRING_NUM: set the vqueue size
  - d) VHOST_SET_VRING_BASE: the base position of avail queue.
  - e) VHOST_SET_VRING_ADDR: the address of descriptor, avail queue, used queue.
  - f) VHOST_SET_VRING_CALL: set the eventfd to inject interrupt to guest
  - g) VHOST_NET_SET_BACKEND: set backed socket or tap fd.
  - h) VHOST_SET_VRING_KICK: set the evntfd to monitor vqueue notify signal.

vhost


Guest
Descriptor
vring
Avail
Used
Notify port
IO port
Vq notify
Inject interrupt
qemu
VHOST_SET_MEM_TABLE
VHOST_SET_VRING_KICK
vhost
vqueue-idx
eventfd
vqueue-idx
eventfd
KVM_IOEVENTFD
VHOST_SET_VRING_CALL
KVM_IRQFD
Io region
eventfd
gsi
eventfd
KVM

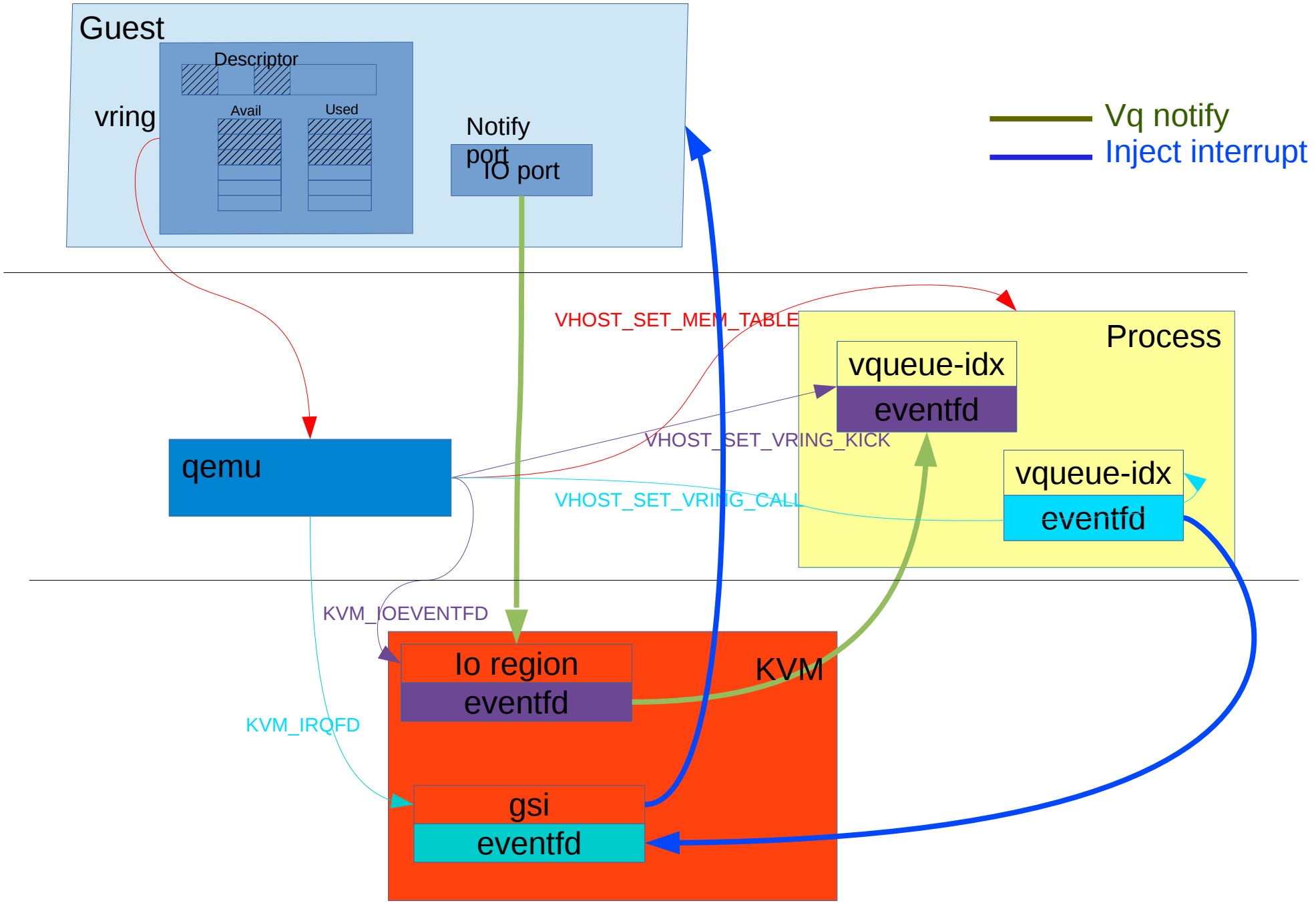
Vhost data flow



Vhost-user

- Similar with vhost but replace the vhost module with a dedicated process.
- It's build on shared memory. So that the dedicated process can access all memory of the VM.

Vhost-user


Guest
Descriptor
vring
Avail
Used
Notify port
IO port
Vq notify
Inject interrupt
VHOST_SET_MEM_TABLE
Process
vqueue-idx
eventfd
VHOST_SET_VRING_KICK
qemu
vqueue-idx
eventfd
VHOST_SET_VRING_CALL
KVM_IOEVENTFD
Io region
eventfd
KVM
KVM_IRQFD
gsi
eventfd

Questions? :)