qemu hot upgrade

北京世纪互联 罗飞

https://www.vnet.com/

Agenda

世纪互联 VNET

- background
- overview of qemu hot upgrade
- optimization and bug fix
  - some bug fix
  - support libvirt
  - fork+exec and rollback
  - interrupt compensation
  - vhost-user backend unaware of the upgrade
- TODOs

Backend

世纪互联 VNET

- online qemu needs upgrade for some reasons:
  - bug fix, security mitigations
  - launch new feature
  - version upgrade

Backend

世纪互联 VNET

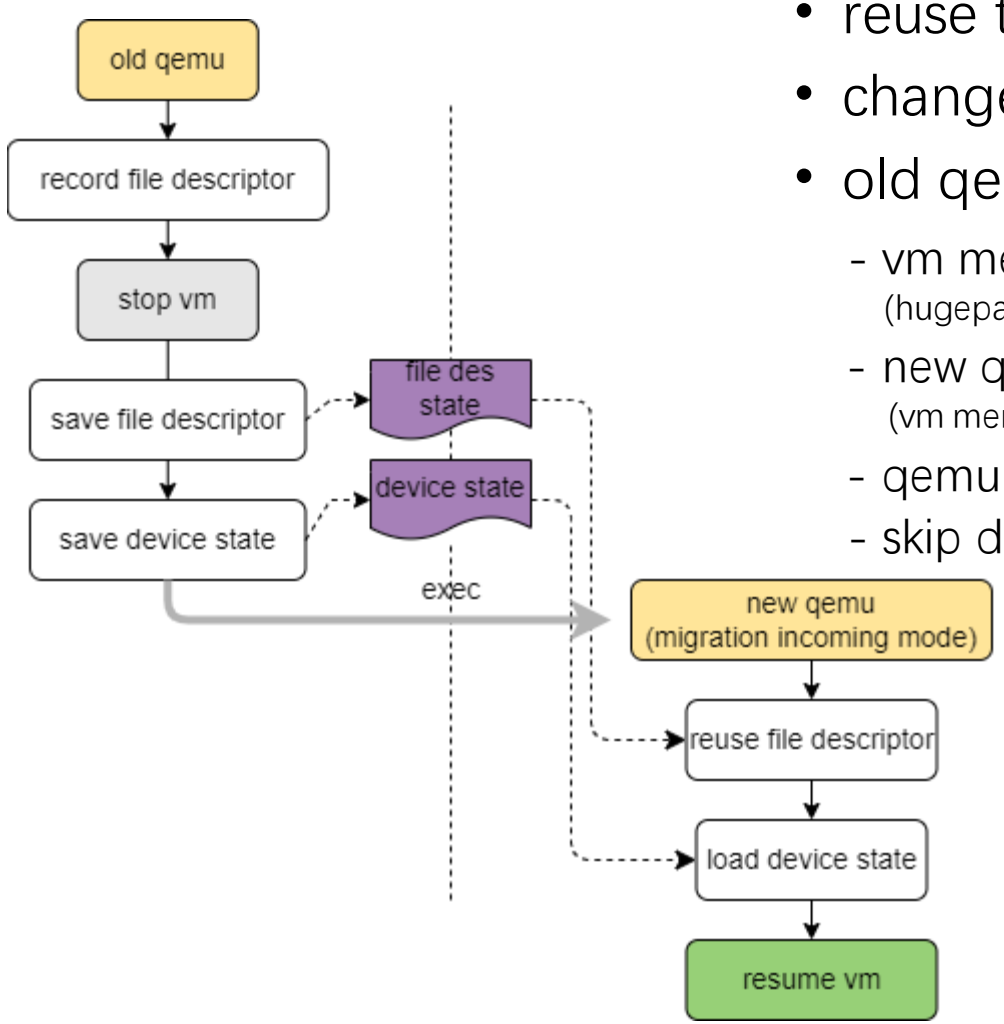
- qemu live migration(including local migration)
  - take an unacceptable long time
  - take up a lot of resources(memory, cpu...)
  - passthrough devices not supported
- qemu live patch(like libcare...)
  - only for simple bug fix

Overview of qemu hot upgrade

世纪互联 VNET

- Advantages:
  - Hot upgrade is completed in tens if milliseconds, with minimal impact on online services
  - Hot upgrade takes up very few resources
  - Support passthrough devices

Overview of qemu hot upgrade



- reuse the qemu live migation framework
- change anonymous ram to memfd
- old qemu exec new qemu binary
  - vm memory from memfd or specified file path (hugepage, epc)
  - new qemu inherit file descriptor (vm memory, socket, eventfd, vfio-device, iommu group, etc)
  - qemu device state serialized and saved to file
  - skip device reset and initialization

世纪互联
VNET

Optimization and bug fix

Digital Economy
Infrastructure Service Provider

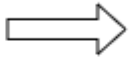
Optimization--support libvirt

- add qemu hot upgrade mode
- keep connection with libvirtd

```
<domain type='kvm'>
  <name>qemuliveupgrade</name>
  <uuid>166c19f3-e1d8-490a-8048-9753313b2a57</uuid>
  <liveupgrade enabled='yes'/>
  <memory unit='KiB'>33554432</memory>
  <currentMemory unit='KiB'>33554432</currentMemory>
  <memoryBacking>
    <hugepages/>
  </memoryBacking>
```

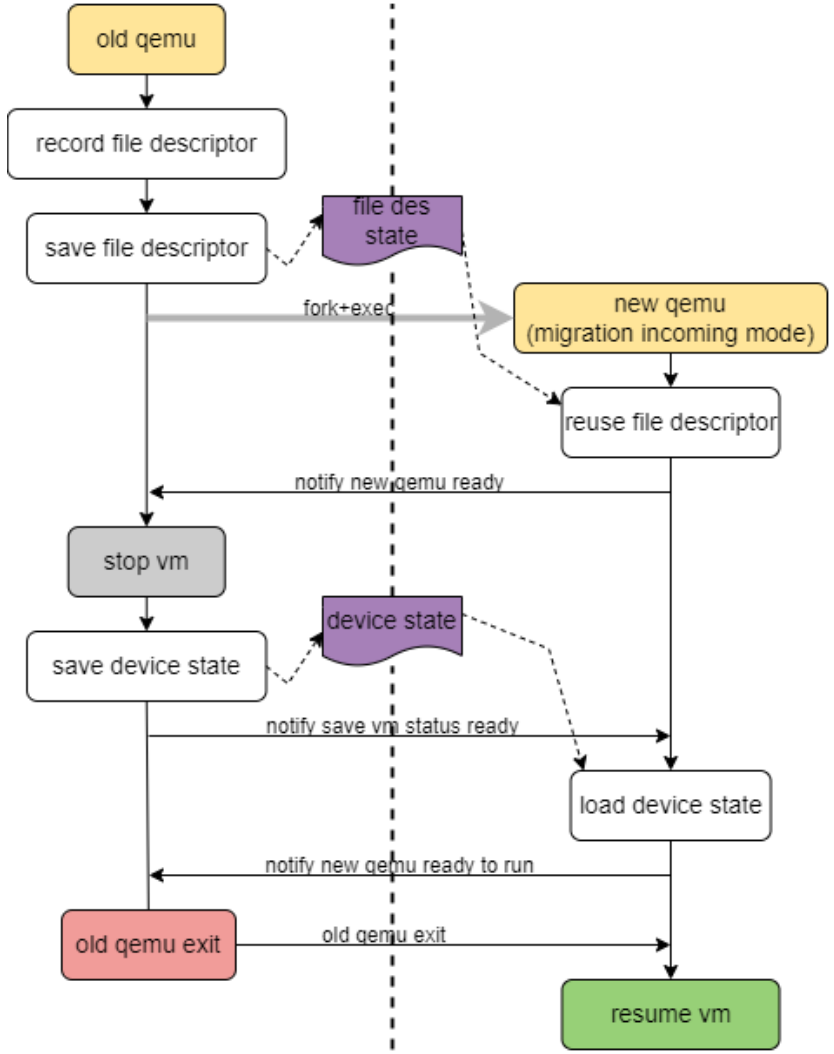
```
#qemu-system-$arch ... -migrate-mode-enable cpr-exec
QEMU 6.2.0 monitor - type 'help' ...
(qemu) info status
VM status: running

(qemu) migrate_set_parameter mode cpr_exec
(qemu) migrate_set_parameter cpr-exec-args
    qemu-system-$arch ... -incoming defer
(qemu) migrate -d file:/tmp/qemu.sav
QEMU 6.2.1 monitor - type 'help' ...
(qemu) info status
VM status: paused (inmigrate)
(qemu) migrate_incoming file:/tmp/qemu.sav
(qemu) info status
VM status: running
```



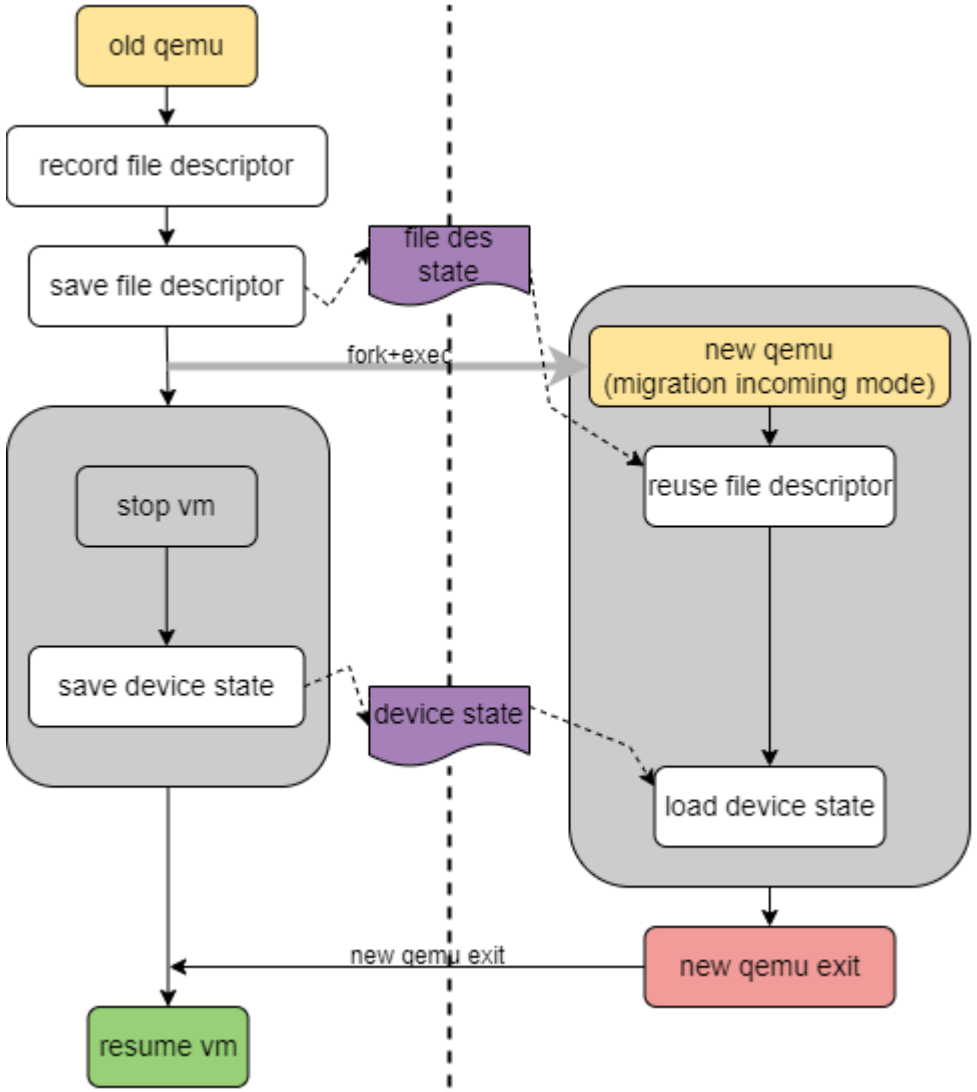
```
virsh start xxx.xml
virsh migrate-liveupgrade xxx
```

Optimization--fork+exec and rollback



- change exec mode to fork+exec
- keep vm running during new qemu initialization

Optimization--fork+exec and rollback



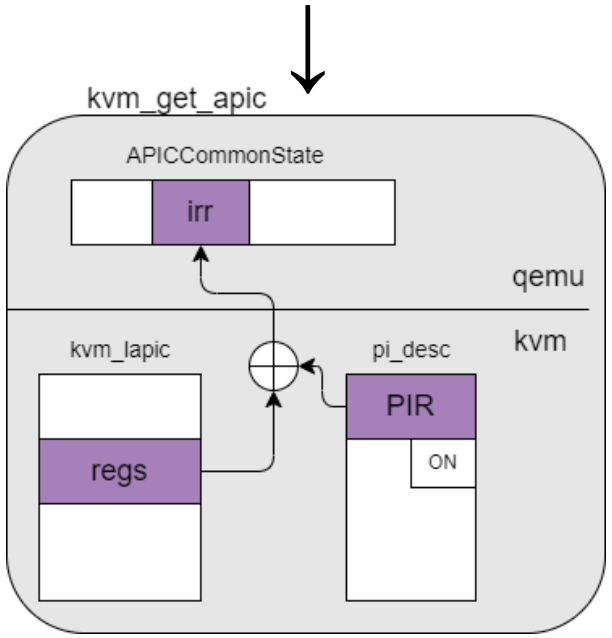
- old qemu resume vm when new qemu exit

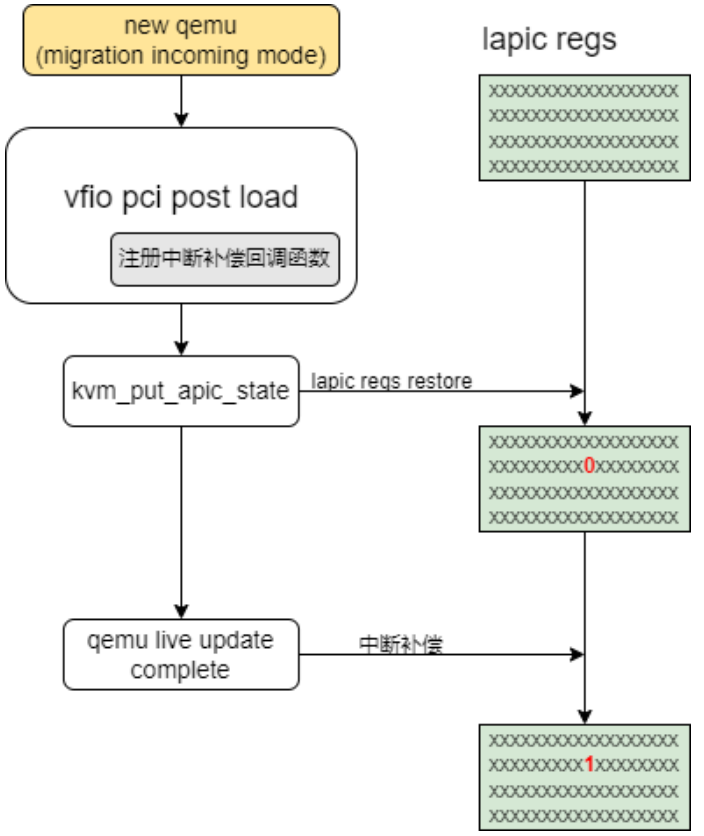
Optimization--interrupt compensation

- passthrough devices may lose interrupts

  (NVIDIA GeForce RTX 2080)

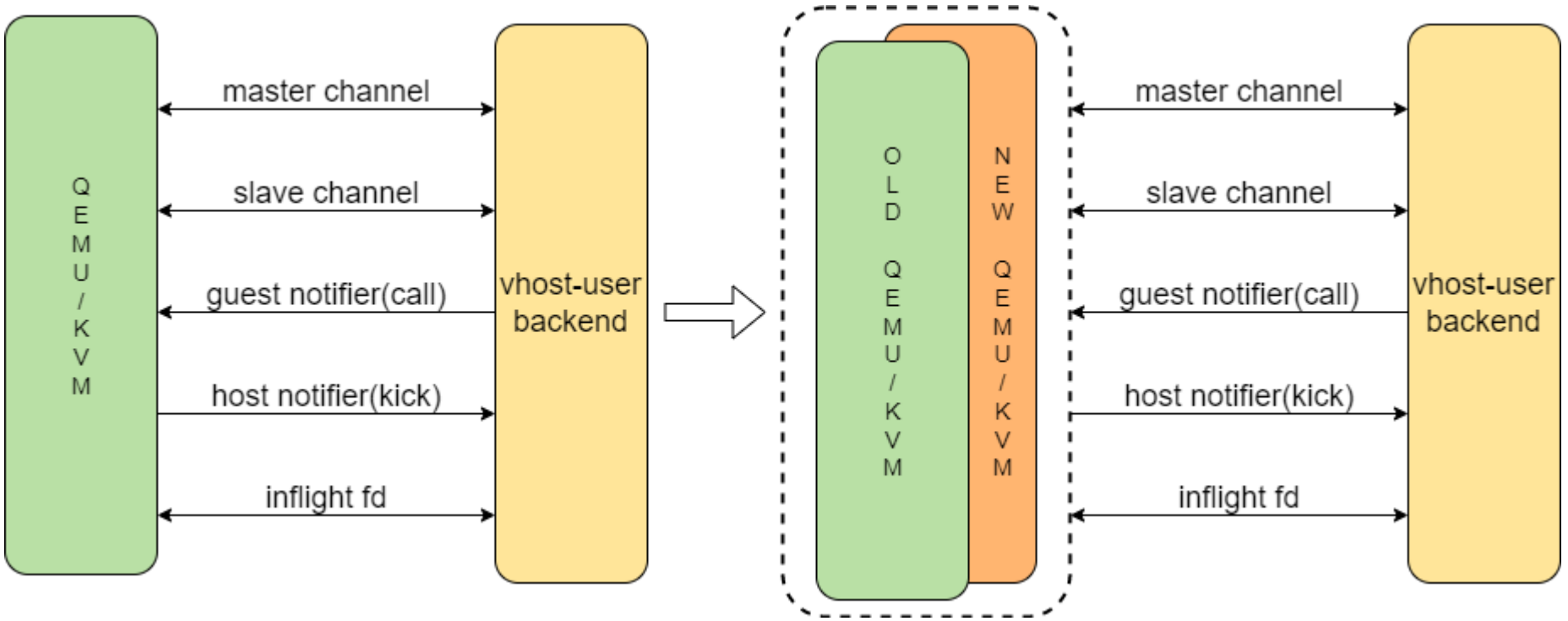
```
@@ -70,7 +70,7 @@ kvm_get_apic(kapic): index: 544: 0x0, 0x
 kvm_get_apic(kapic): index: 552: 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0
 kvm_get_apic(kapic): index: 560: 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0
 kvm_get_apic(kapic): index: 568: 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0
-kvm_get_apic(kapic): index: 576: 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0
+kvm_get_apic(kapic): index: 576: 0x0, 0x0, 0x8, 0x0, 0x0, 0x0, 0x0, 0x0
 kvm_get_apic(kapic): index: 584: 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0
 kvm_get_apic(kapic): index: 592: 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0
 kvm_get_apic(kapic): index: 600: 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0
```



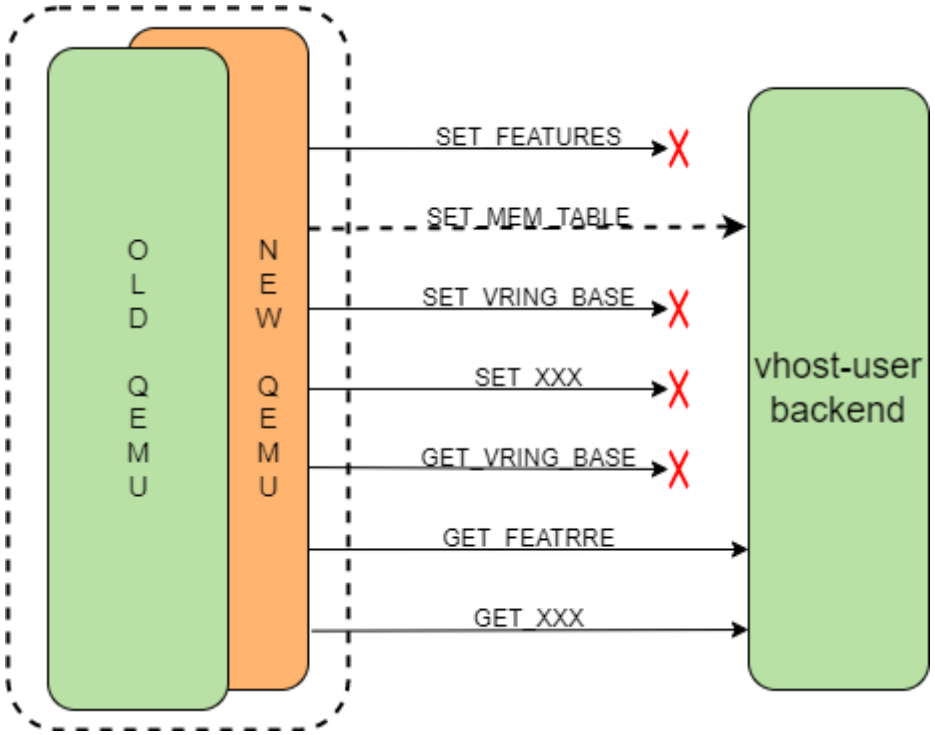


Optimization--transparent to vhost-user backend

- inherit the fds between qemu and vhost user backend
  (problem: new qemu vhost user command may cause connection broken)



Optimization--transparent to vhost-user backend



- skip almost all set commands and some get commands
- skip the cleanup of notifier and inflight fd when old qemu exit

Optimization--others

世纪互联 VNET

- support sandbox
- hostmem-file
- vm with passthrough GPU device crash
- vhost-user backend

TODOs

世纪互联 VNET

- sandbox spawn denied
- socket(SO_REUSEADDR->SO_REUSEPORT)
- fixed address map
- more scenes(mdev, ...)
- inherit vm fd, vcpu fd?
- save and load device state concurrently like live migration?
- kvm module live upgrade

Submit to euler community

世纪互联 VNET

- Pull Request

  qemu branch: qemu-6.2.0-hotupgrade

  https://gitee.com/openeuler/qemu/pulls/569

世纪互联
VNET

THANKS