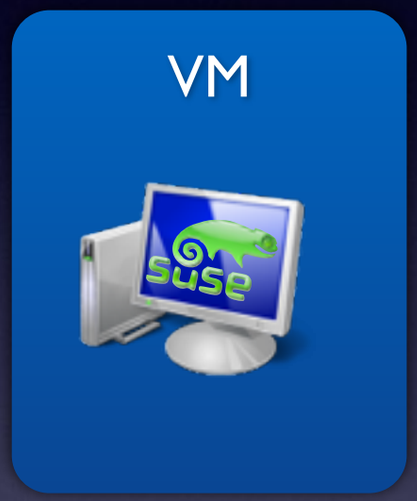
Migratable 40GBit/s Ethernet

About Me

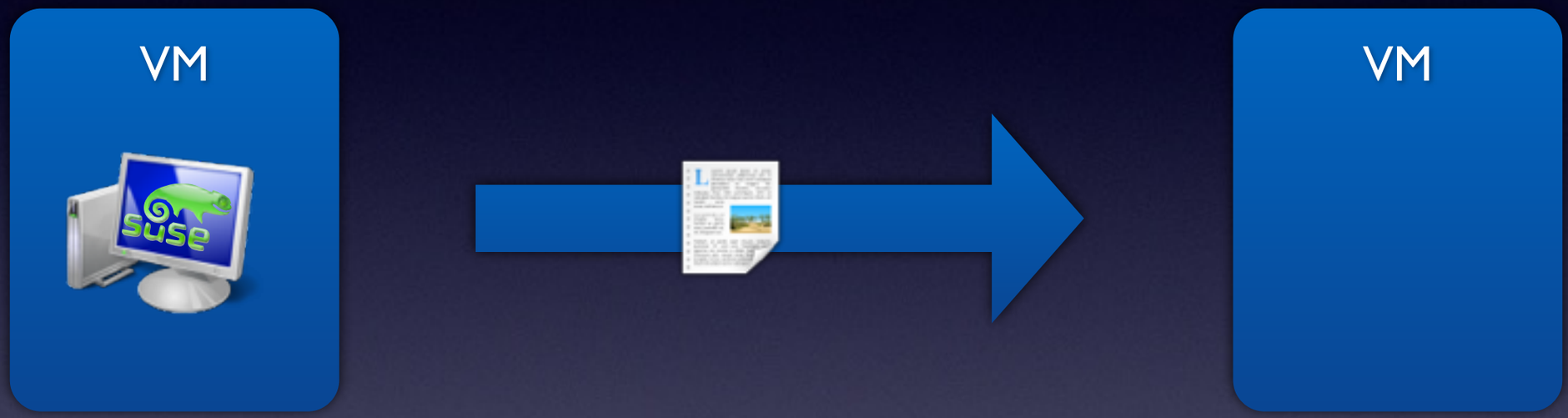
- Alexander Graf
- KVM and Qemu developer for SUSE
  - Server class PowerPC KVM port
  - S390x Qemu guest support
  - x86 Mac OS X in KVM
  - Nested SVM
  - ...

What is Live Migration

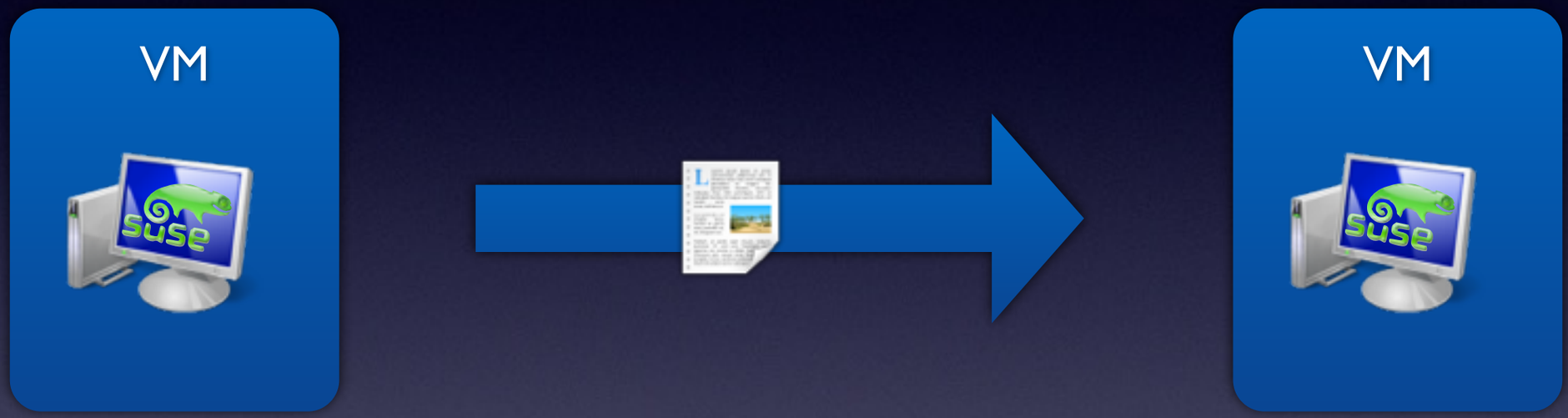
What is Live Migration



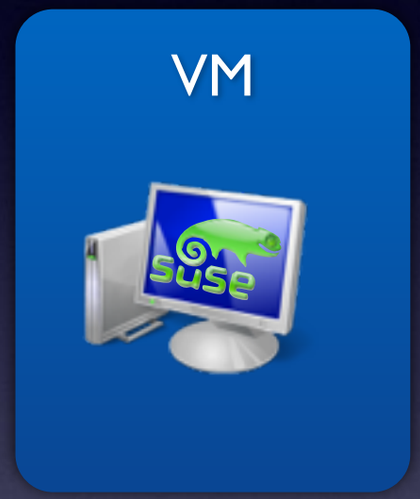
What is Live Migration



What is Live Migration

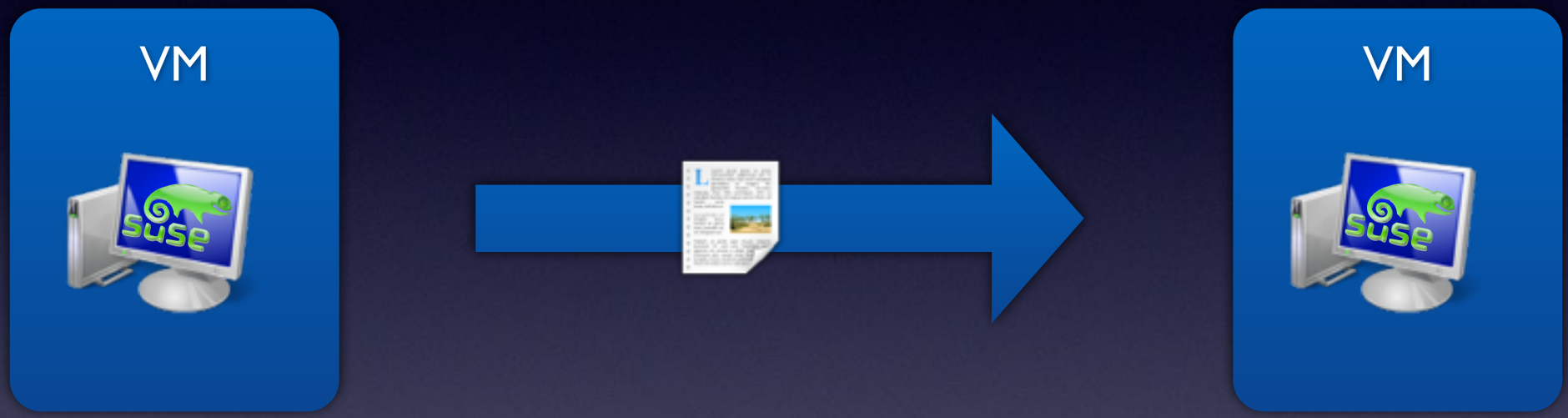


What is Live Migration

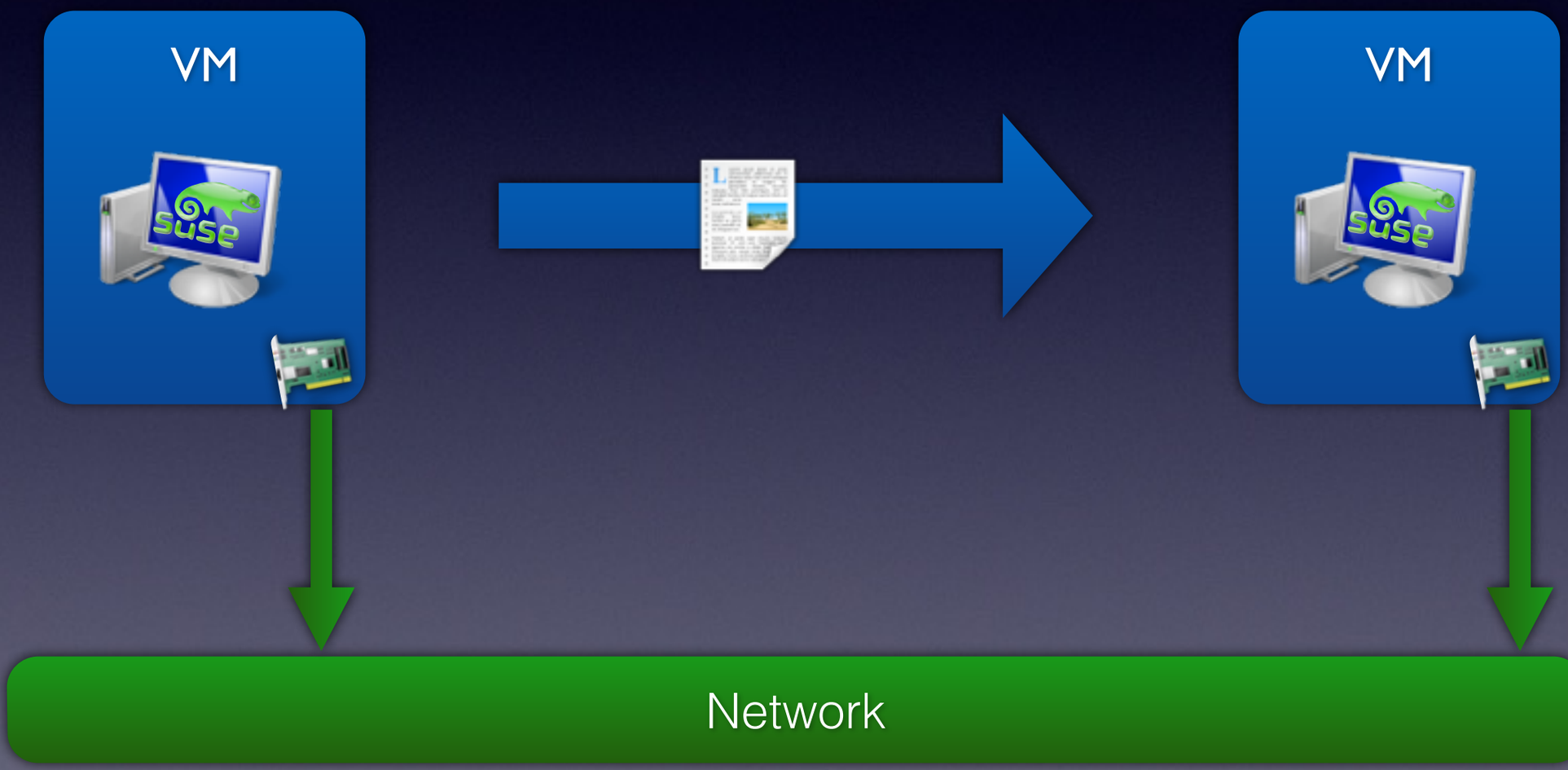


Network Live Migration

Network Live Migration

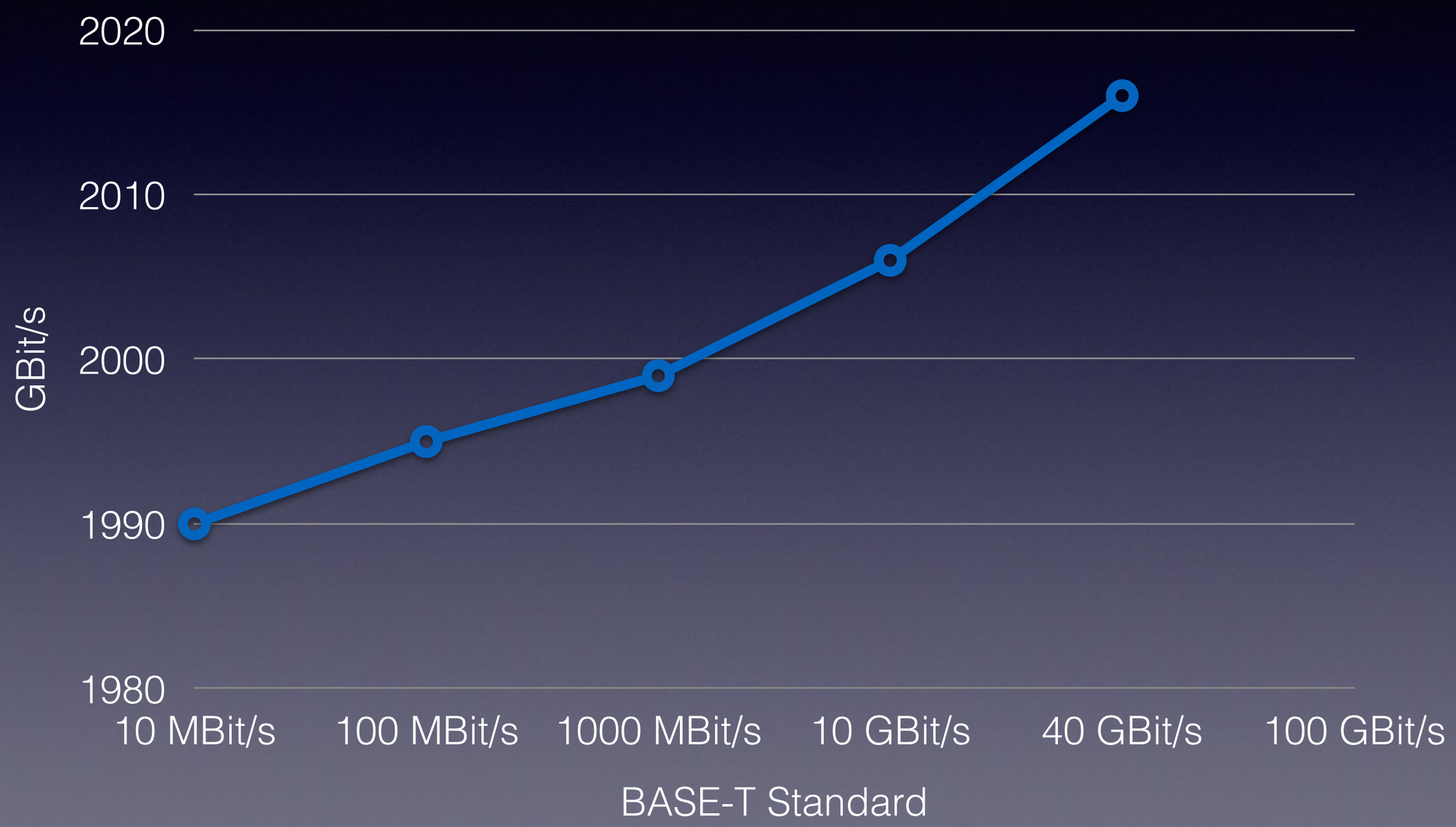


Network Live Migration

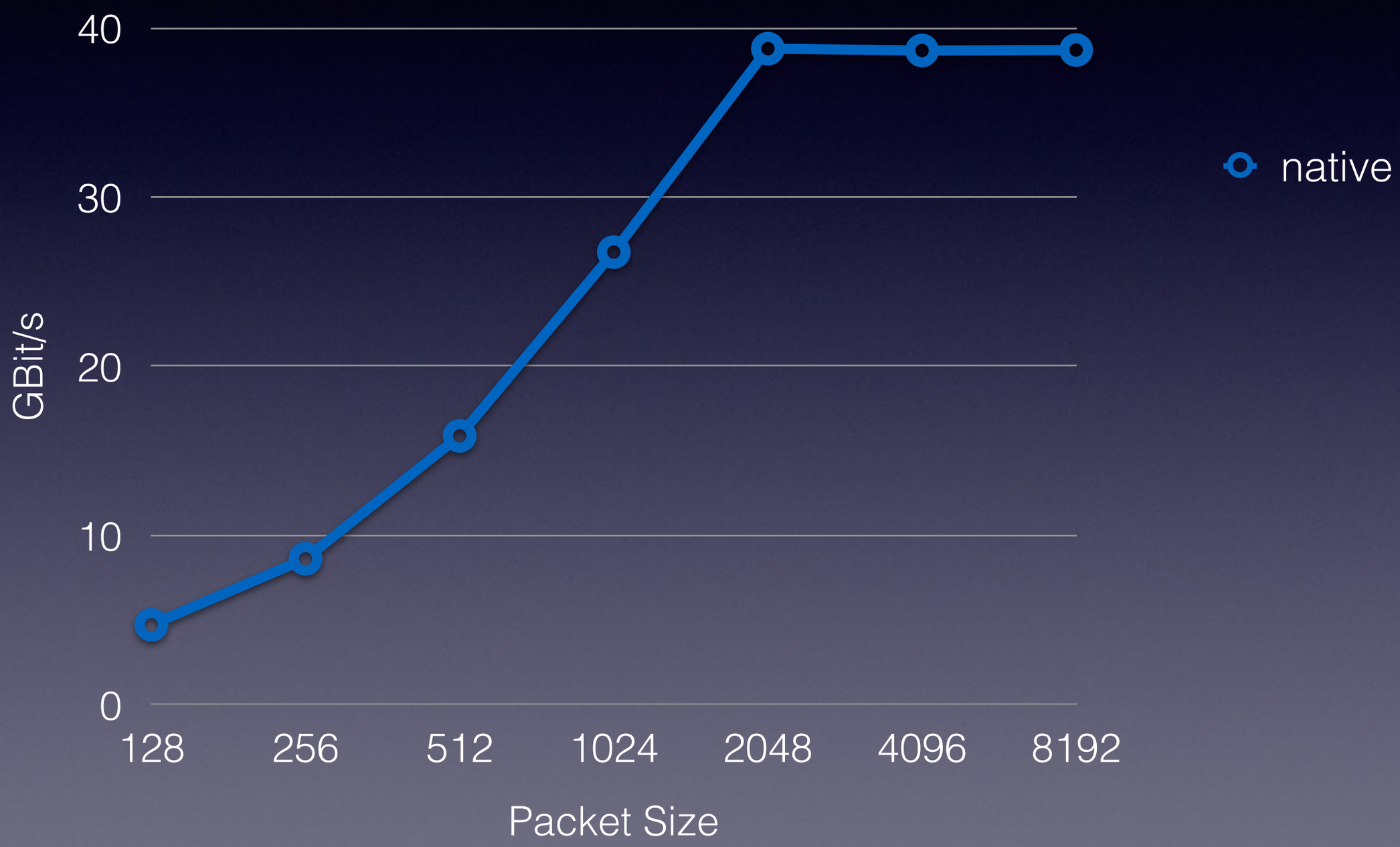


40 GBit/s

40 GBit/s

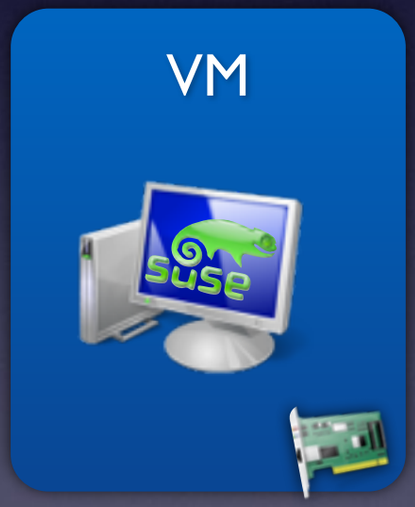

2020
2010
2000
1990
1980
GBit/s
10 MBit/s
100 MBit/s
1000 MBit/s
10 GBit/s
40 GBit/s
100 GBit/s
BASE-T Standard

40 GBit/s

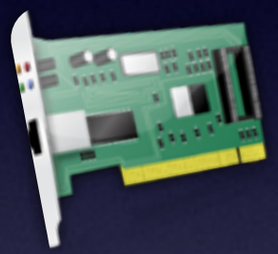

40
30
20
10
0
GBit/s
128
256
512
1024
2048
4096
8192
Packet Size
native

virtio

virtio

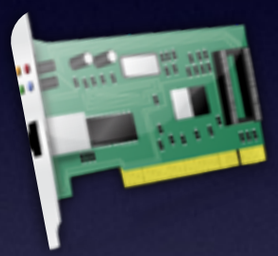


virtio



virtio-net

virtio



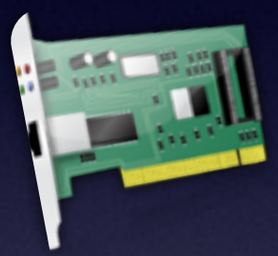
virtio-net

TX ring

RX ring

Admin ring

virtio

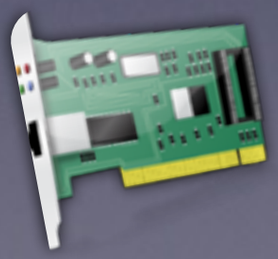


virtio-net

TX ring

RX ring

Admin ring



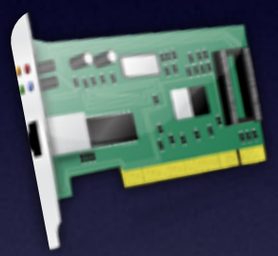
real NIC

TX ring

RX ring

Admin ring

virtio



virtio-net

TX ring

RX ring

Admin ring

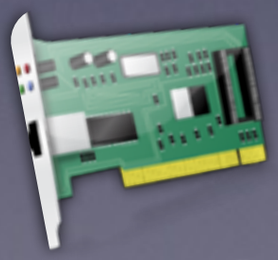
vhost

QEMU

TX ring

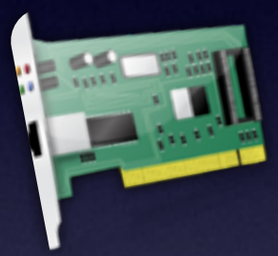
RX ring

Admin ring



real NIC

virtio



virtio-net

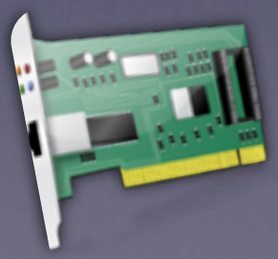
TX ring

RX ring

Admin ring

vhost

QEMU



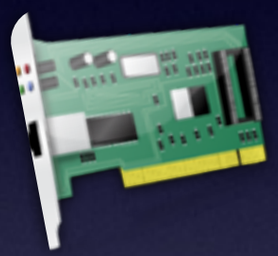
real NIC

TX ring

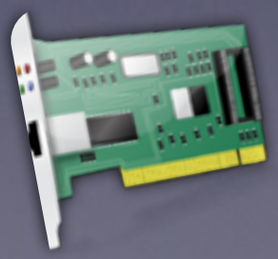
RX ring

Admin ring

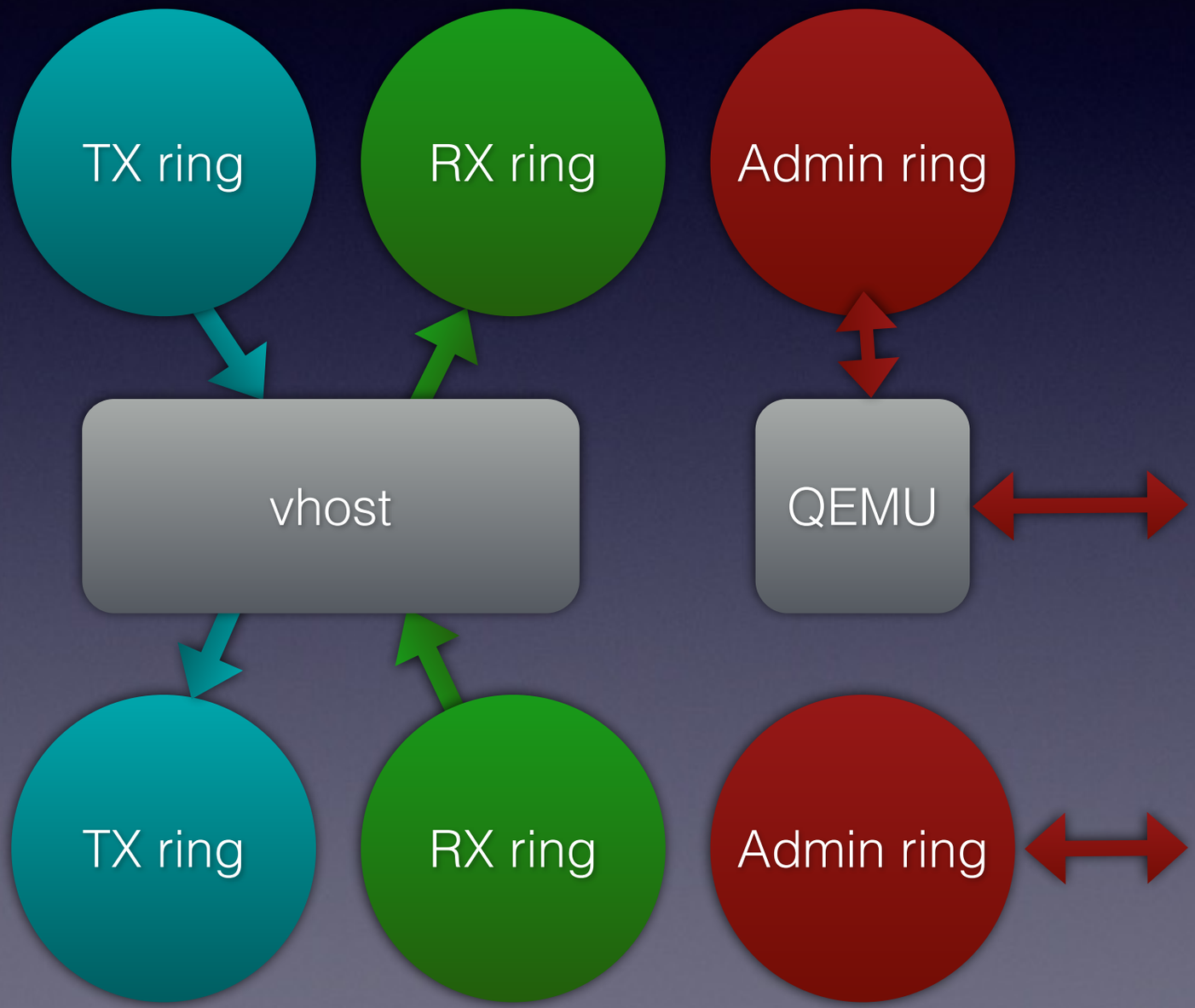
virtio



virtio-net



real NIC



virtio-net

TX ring

RX ring

vhost

real NIC

TX ring

RX ring

virtio-net

TX ring

RX ring

vhost queue driver

vhost

real NIC

TX ring

RX ring

virtio-net

TX ring

RX ring

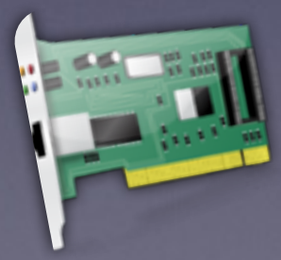
vhost queue driver

Linux tap

vhost

TX ring

RX ring



real NIC

virtio-net

TX ring

RX ring

vhost queue driver

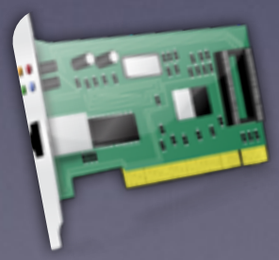
Linux tap

vhost

Linux bridge

TX ring

RX ring



real NIC

virtio-net

TX ring

RX ring

vhost queue driver

Linux tap

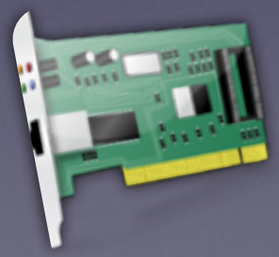
vhost

Linux bridge

Linux NIC driver

TX ring

RX ring



real NIC

virtio-net

TX ring

RX ring

vhost queue driver

Linux tap

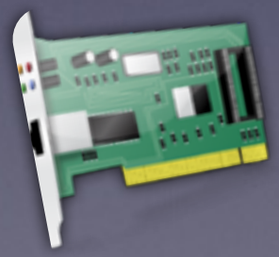
vhost

Linux bridge

Linux NIC driver

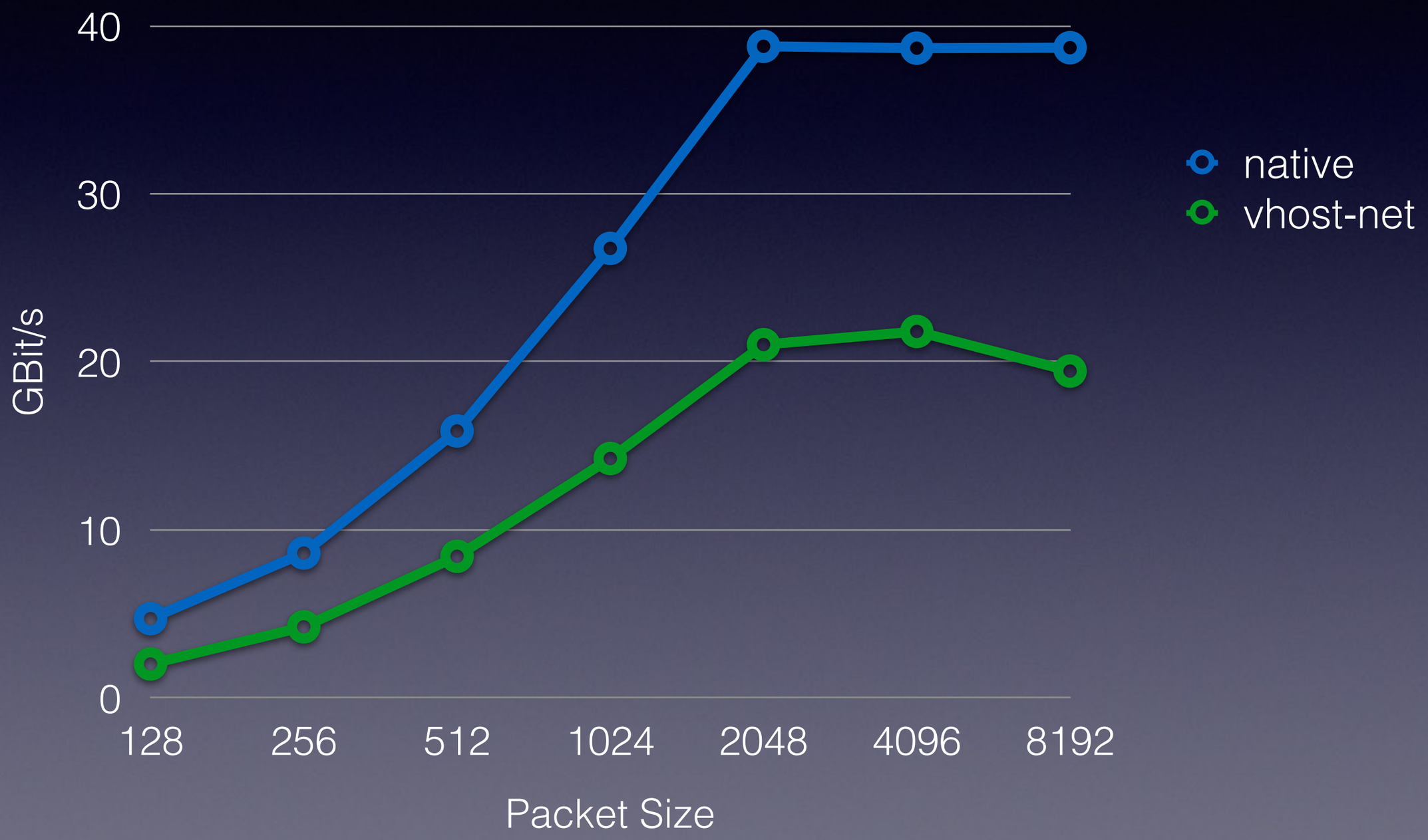
TX ring

RX ring

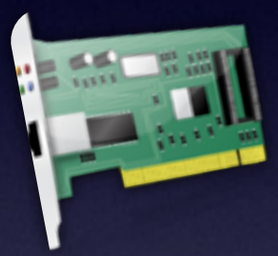


real NIC

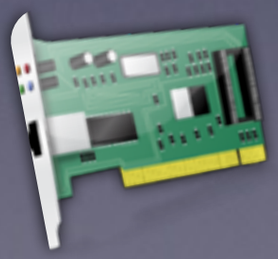
virtio


40
30
20
10
0
GBit/s
128
256
512
1024
2048
4096
8192
Packet Size
native
vhost-net

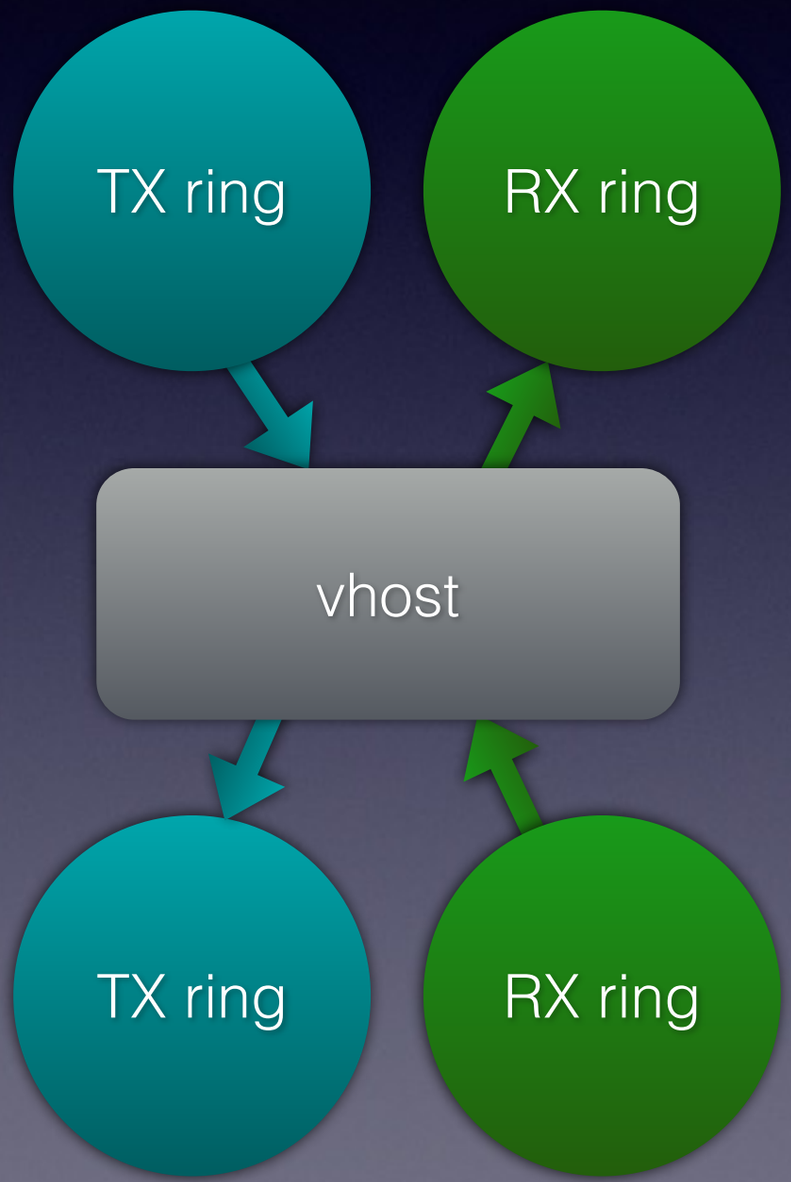
virtio



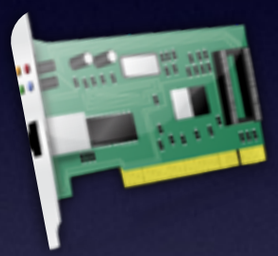
virtio-net



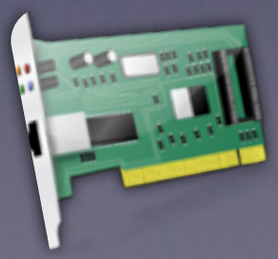
real NIC



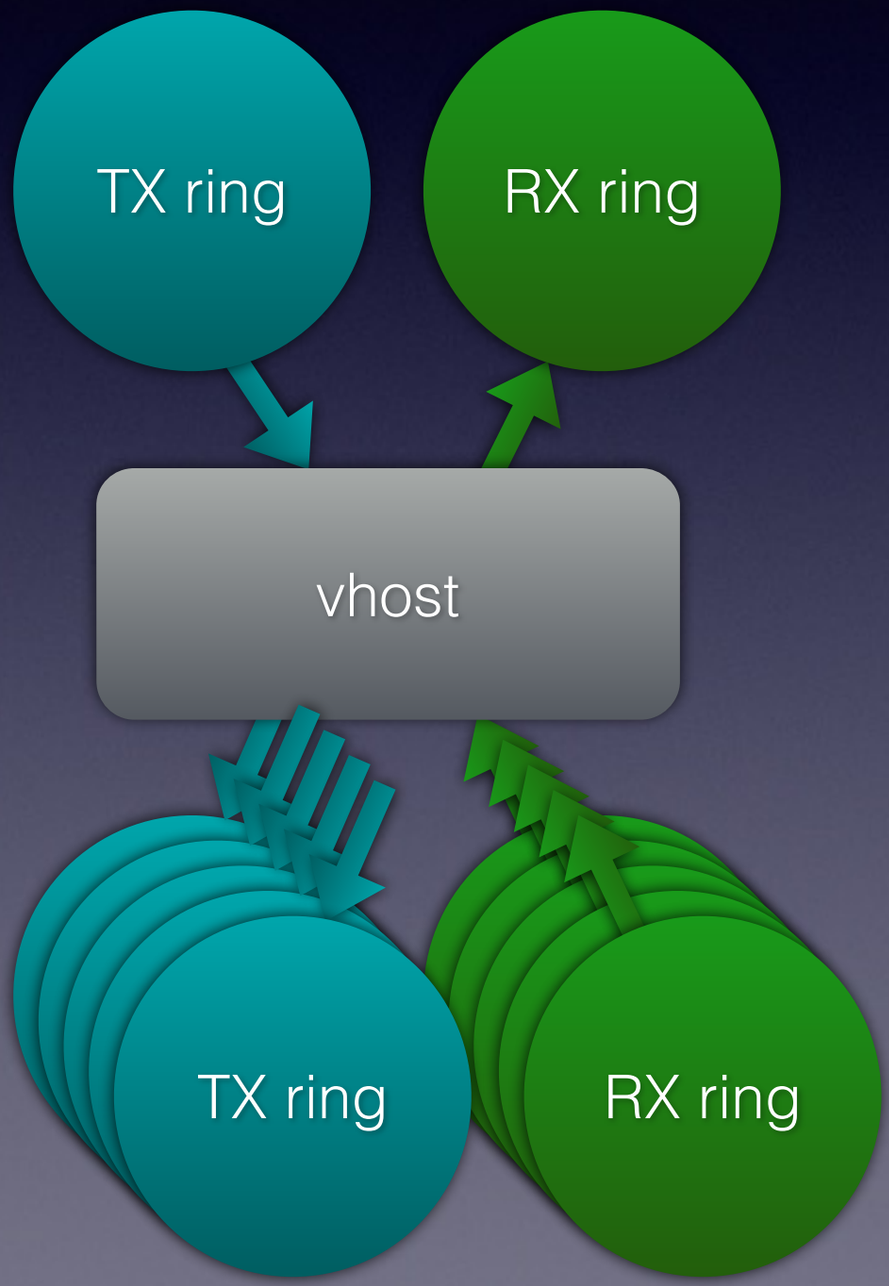
virtio



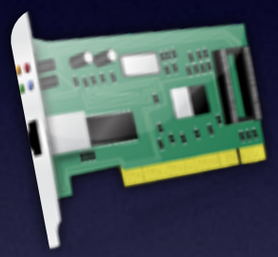
virtio-net



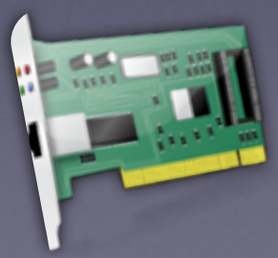
real NIC



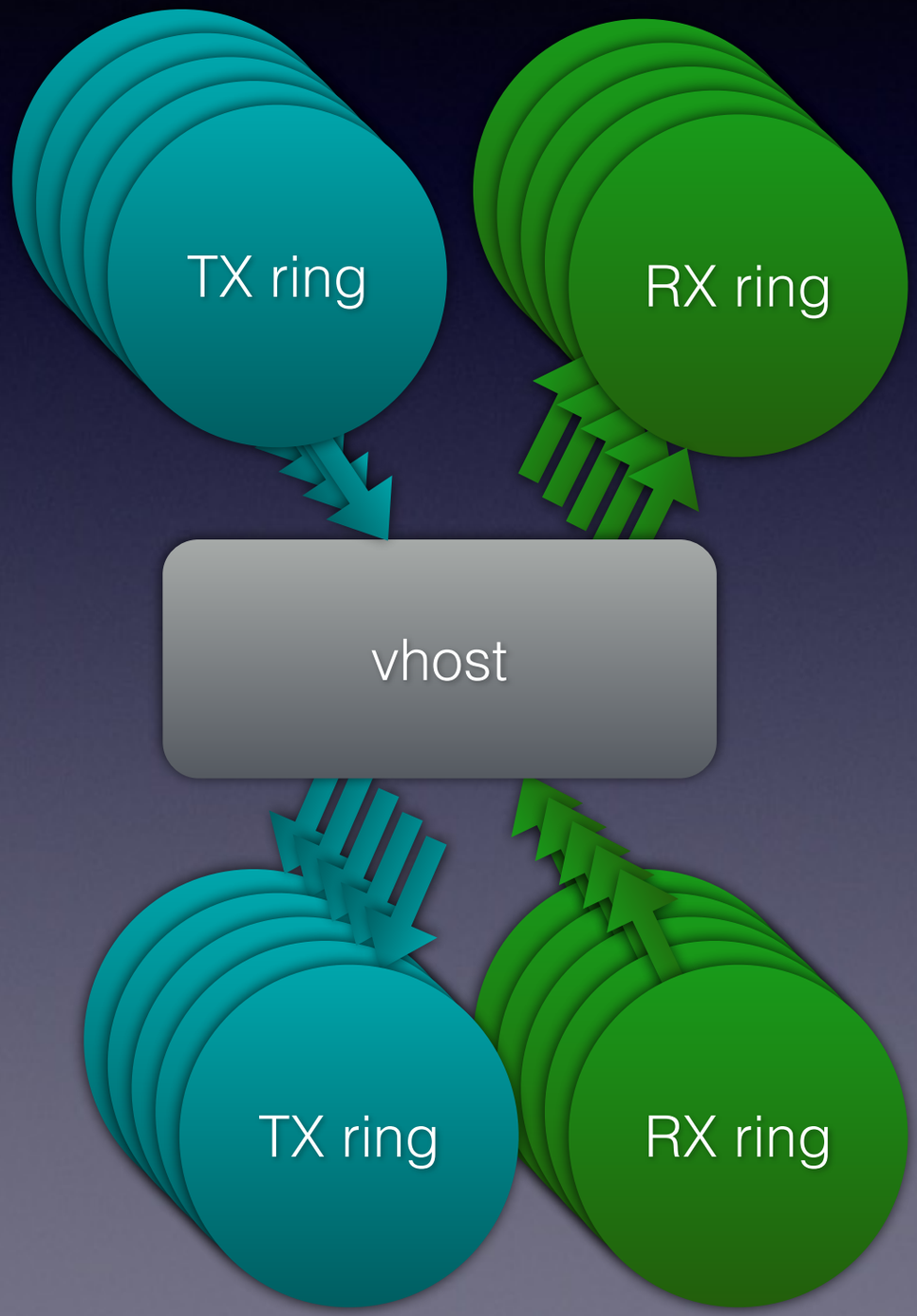
virtio



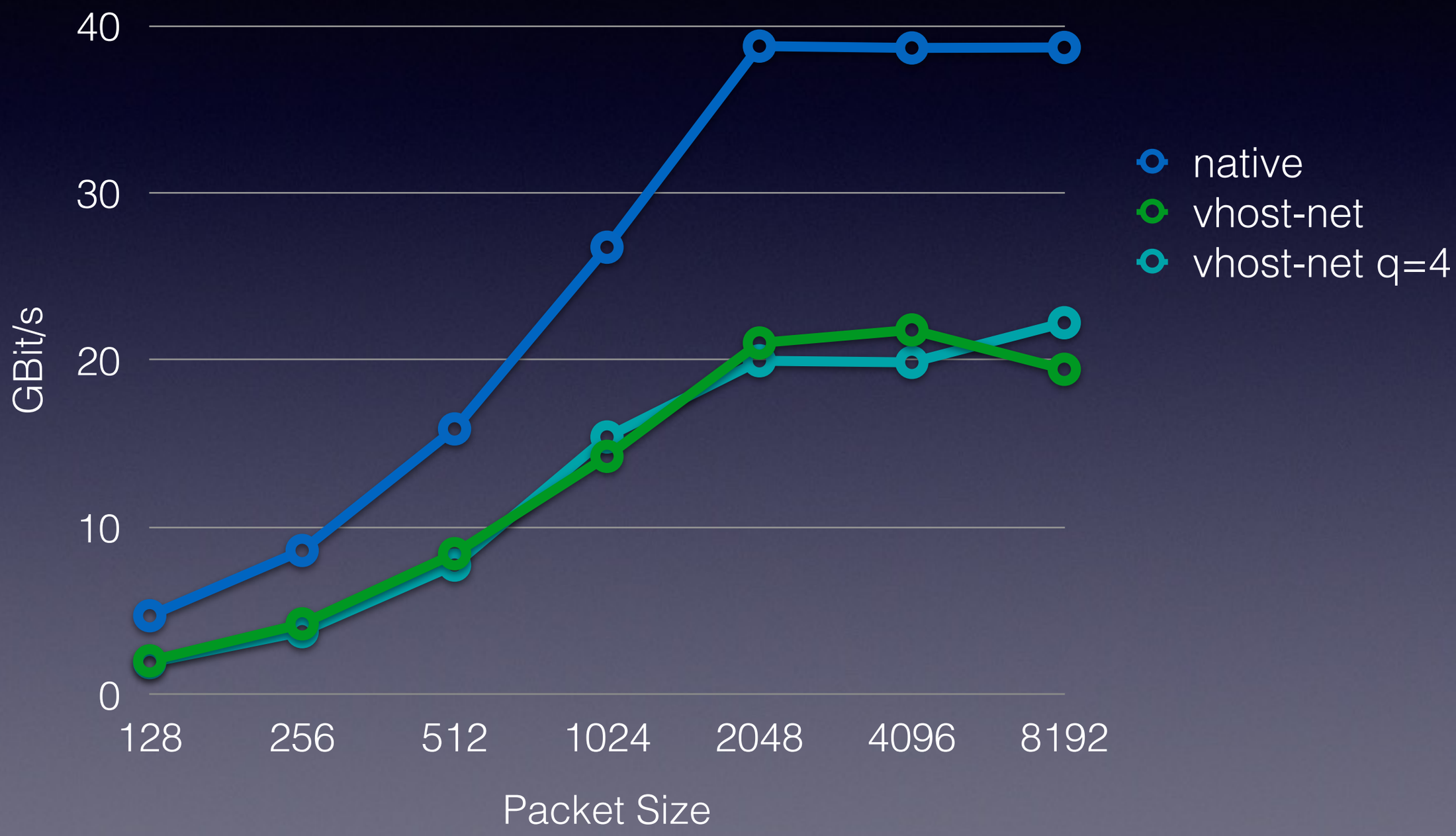
virtio-net



real NIC



virtio


40
30
20
10
0
GBit/s
128
256
512
1024
2048
4096
8192
Packet Size
native
vhost-net
vhost-net q=4

virtio

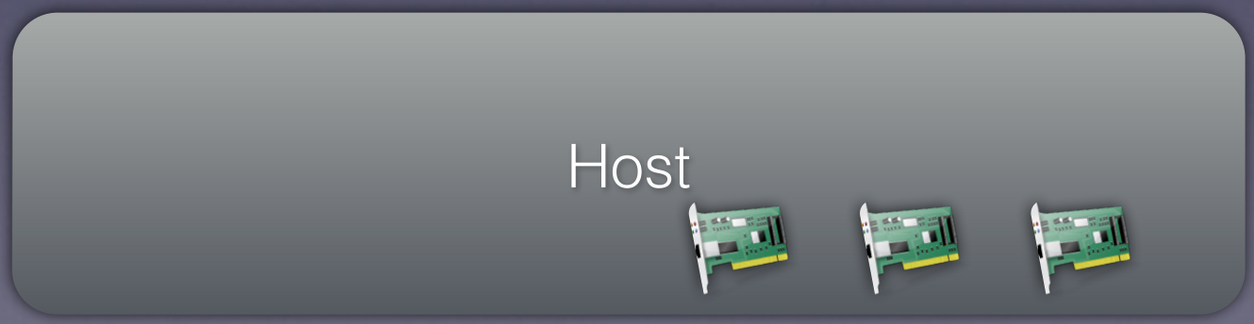
- Bridge can be accelerated with flood=off learning=off
- Requires VLAN
- Shouldn’t make a difference on direct connect

virtio

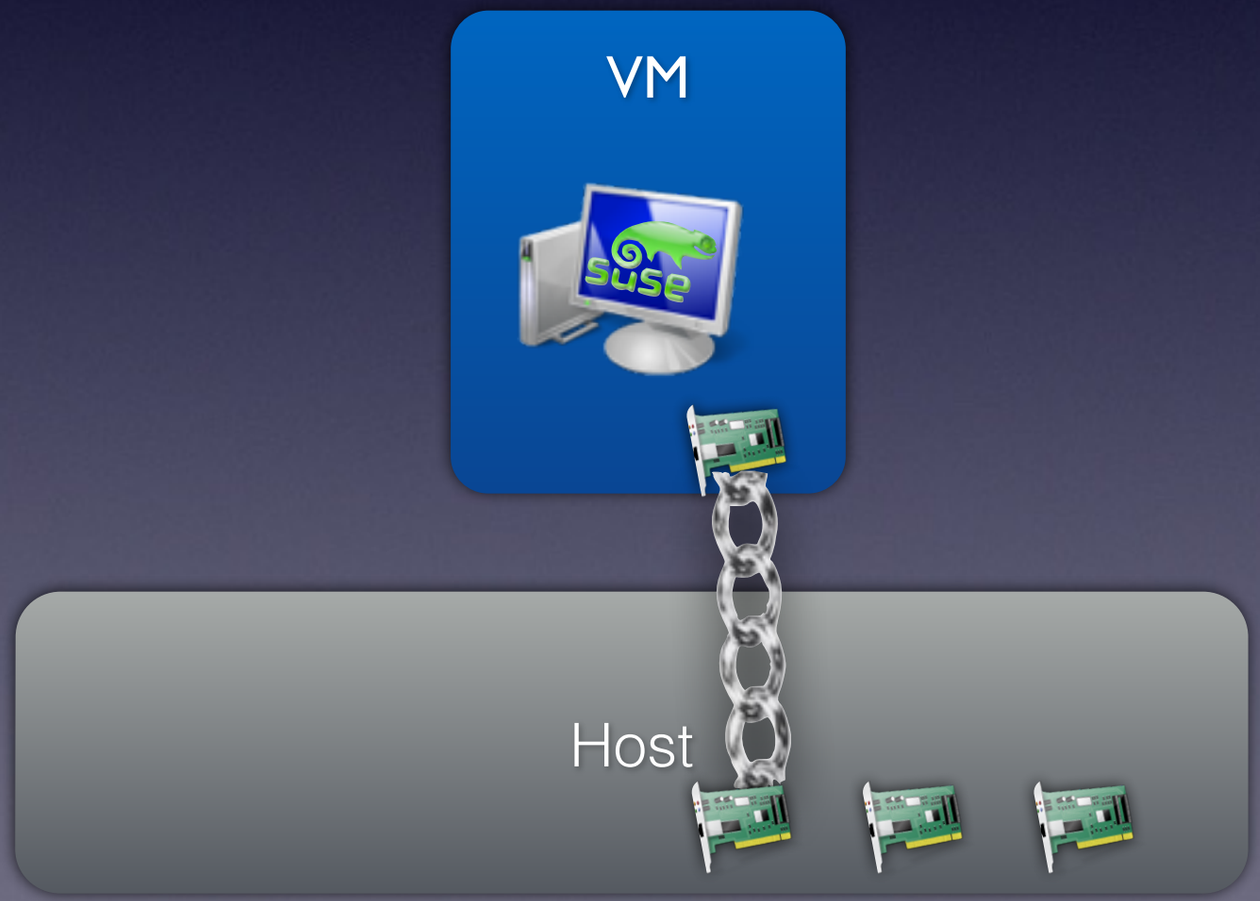
- Alternatives to bridge
  - Open vSwitch
  - rocker
  - macvtap
- Not benchmarked yet

VFIO

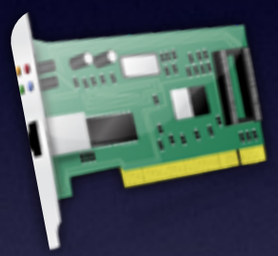
VFIO



VFIO



VFIO

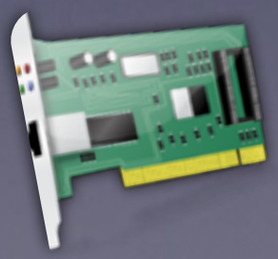


guest NIC

TX ring

RX ring

Admin ring



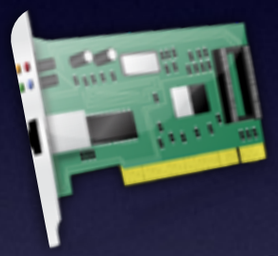
real NIC

TX ring

RX ring

Admin ring

VFIO

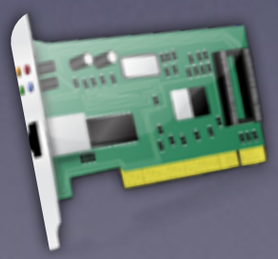


guest NIC

TX ring

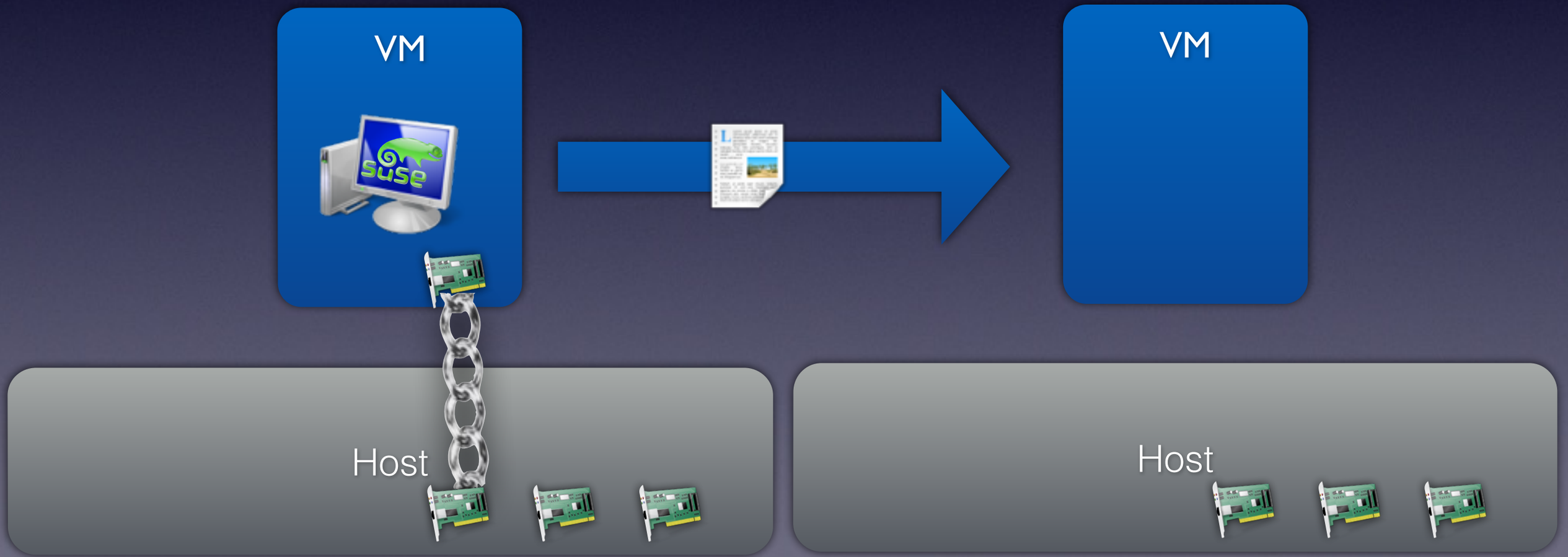
RX ring

Admin ring

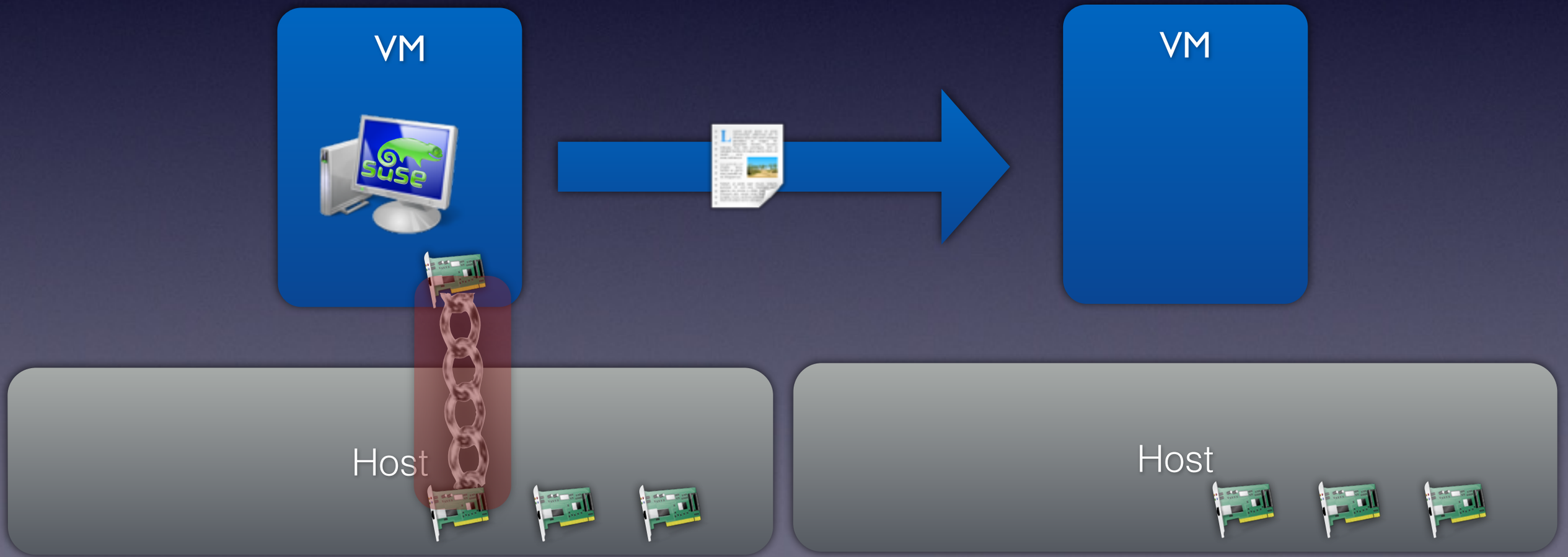


real NIC

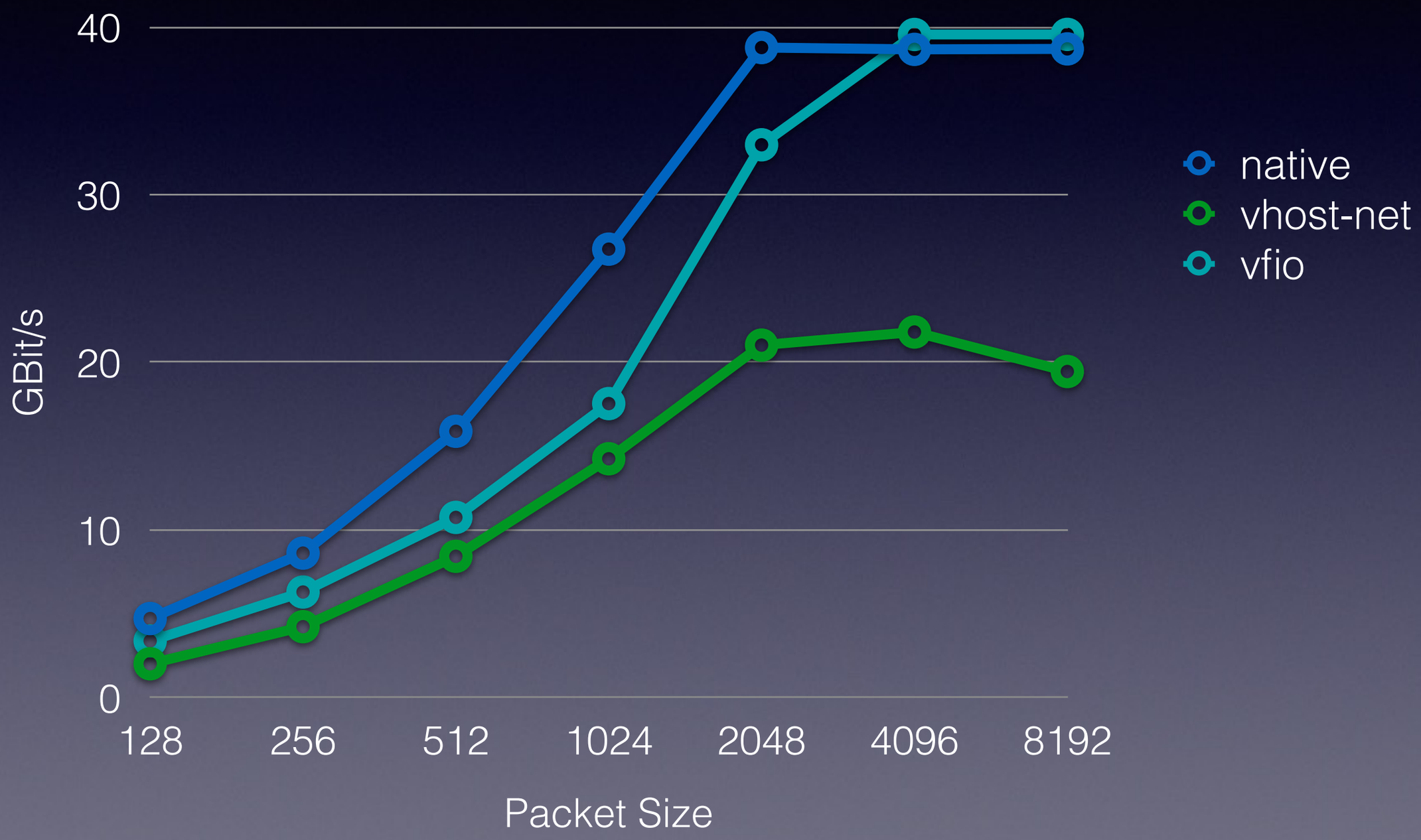
VFIO



VFIO



VFIO


40
30
20
10
0
GBit/s
128
256
512
1024
2048
4096
8192
Packet Size
native
vhost-net
vfio

VFIO

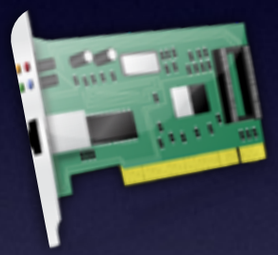
- Generic driver
- No chance for introspection

VFIO-i40evf

VFIO-i40evf

VFIO-i40evf

VFIO-i40evf

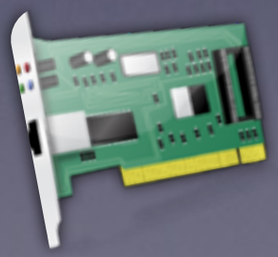


guest NIC

TX ring

RX ring

Admin ring



real NIC

VFIO-i40evf

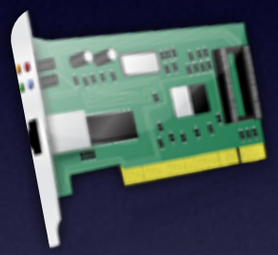
Admin ring

VFIO-i40evf

Admin ring

- Set MAC addresses
- Set RX/TX locations
- Set RX/TX properties
- …

VFIO-i40evf

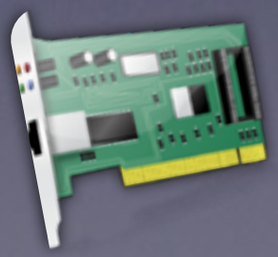


guest NIC

TX ring

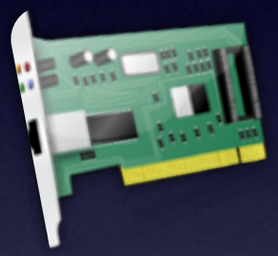
RX ring

Admin ring



real NIC

VFIO-i40evf

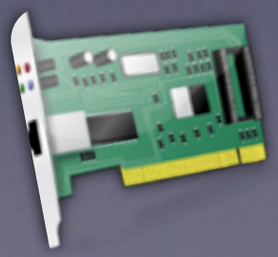


guest NIC

Admin ring

TX ring

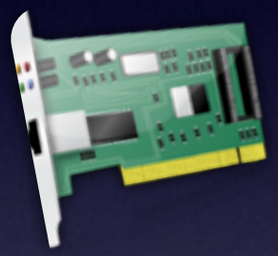
RX ring



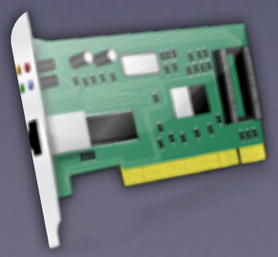
real NIC

Admin ring

VFIO-i40evf



guest NIC



real NIC

TX ring

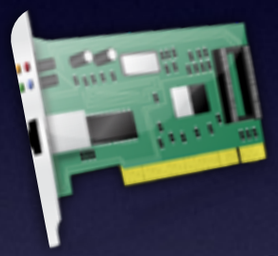
RX ring

Admin ring

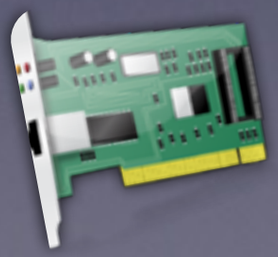
QEMU

Admin ring

VFIO-i40evf



guest NIC



real NIC

TX ring

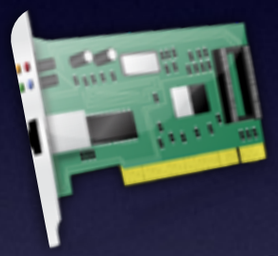
RX ring

Admin ring

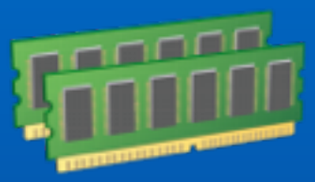
QEMU

Admin ring

VFIO-i40evf



guest NIC



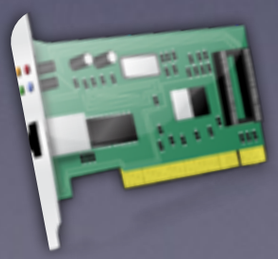
TX ring

RX ring

Admin ring

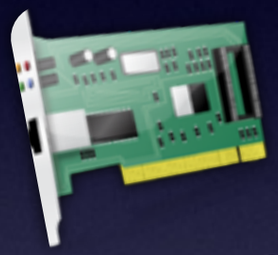
QEMU

Admin ring

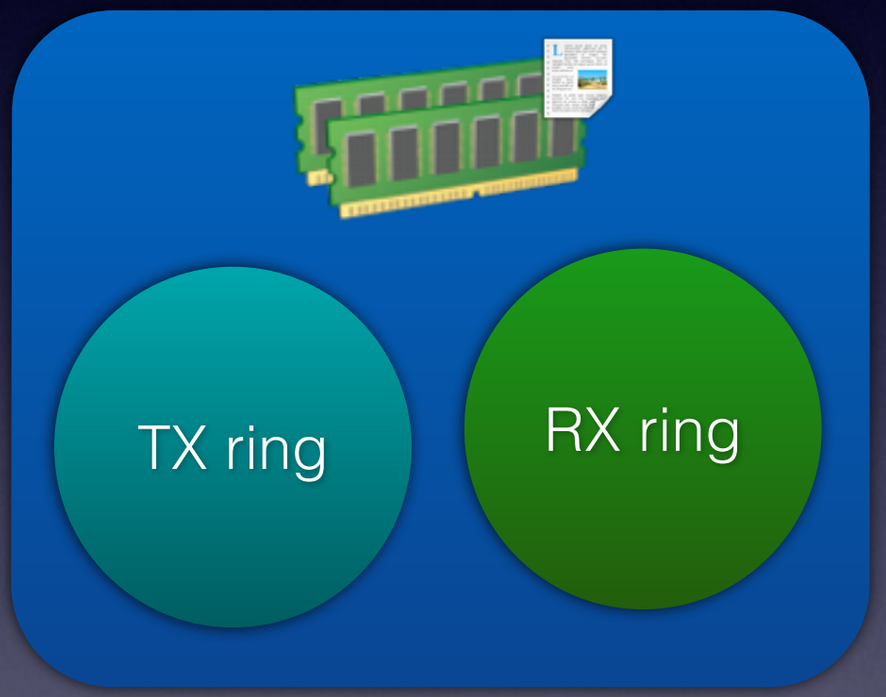


real NIC

VFIO-i40evf



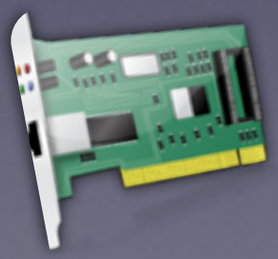
guest NIC



Admin ring

QEMU

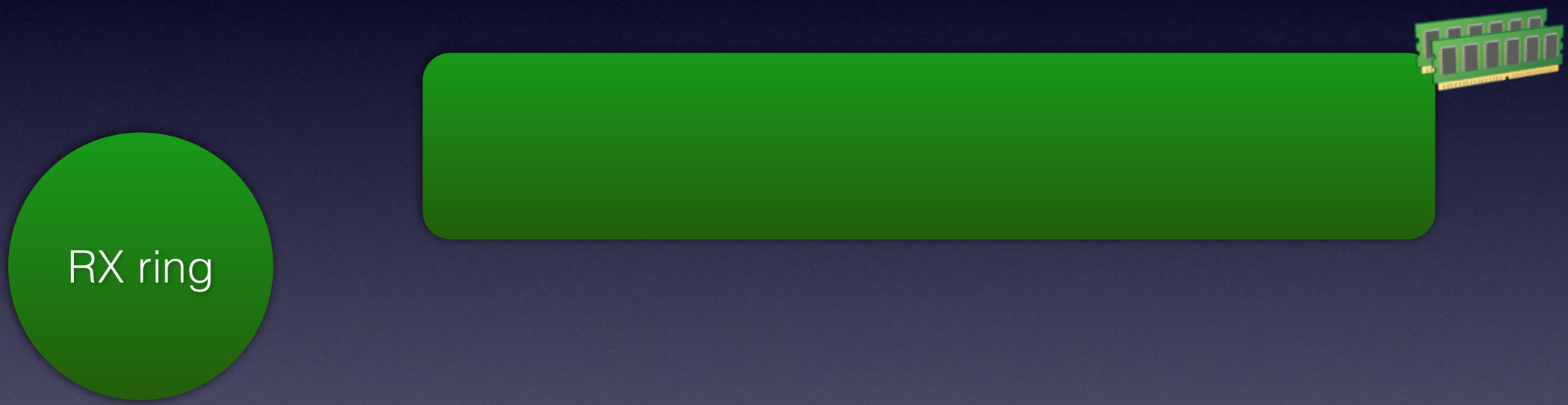
Admin ring



real NIC

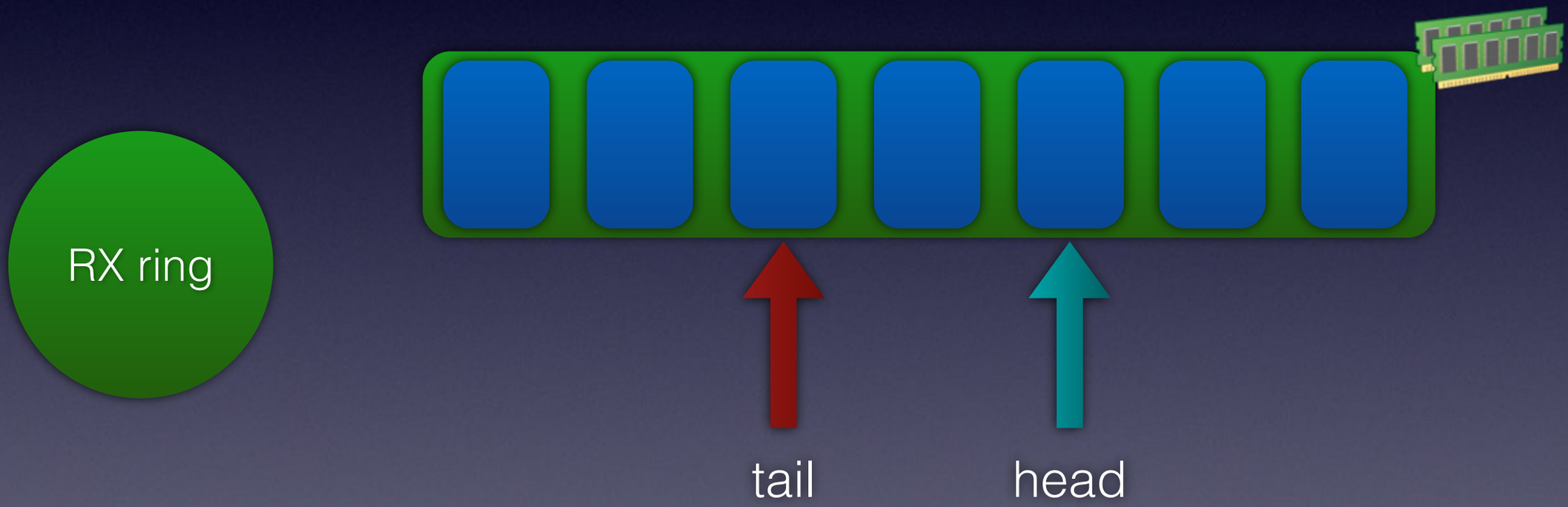
VFIO-i40evf

VFIO-i40evf

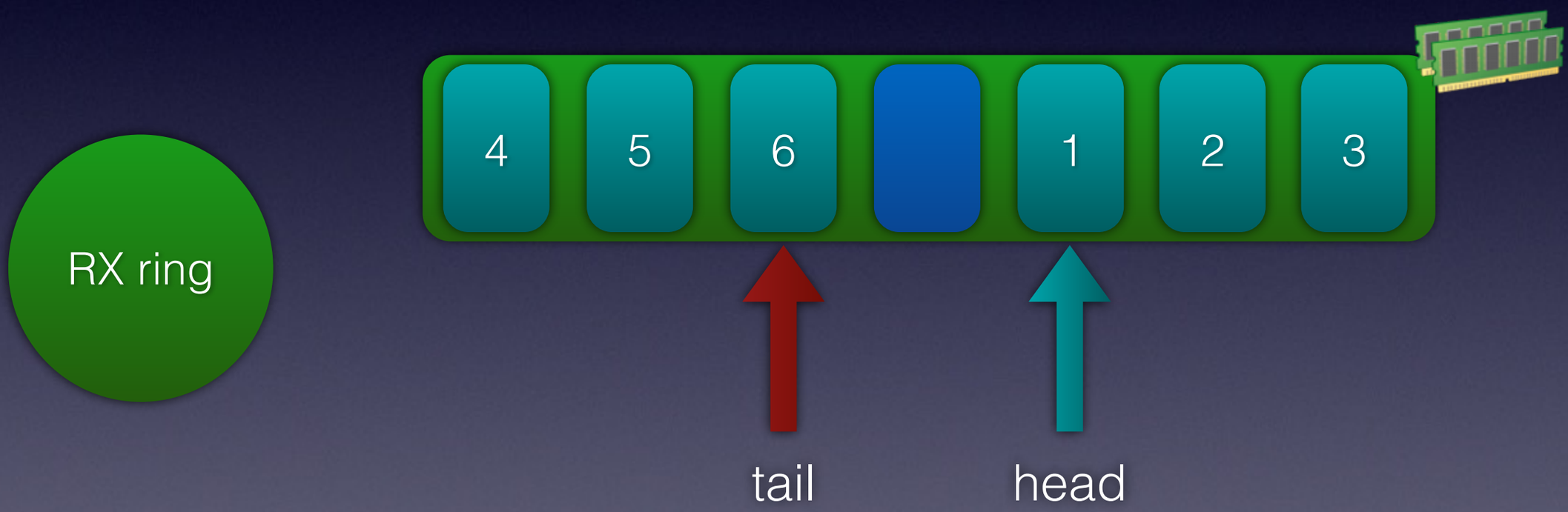


VFIO-i40evf

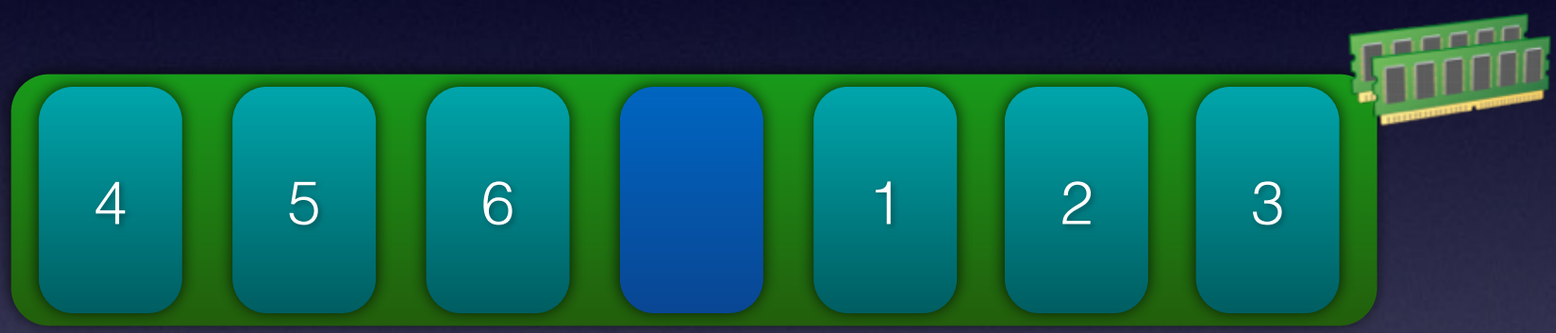
VFIO-i40evf



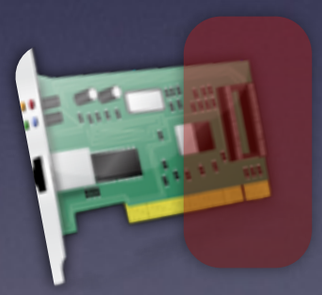
VFIO-i40evf



VFIO-i40evf



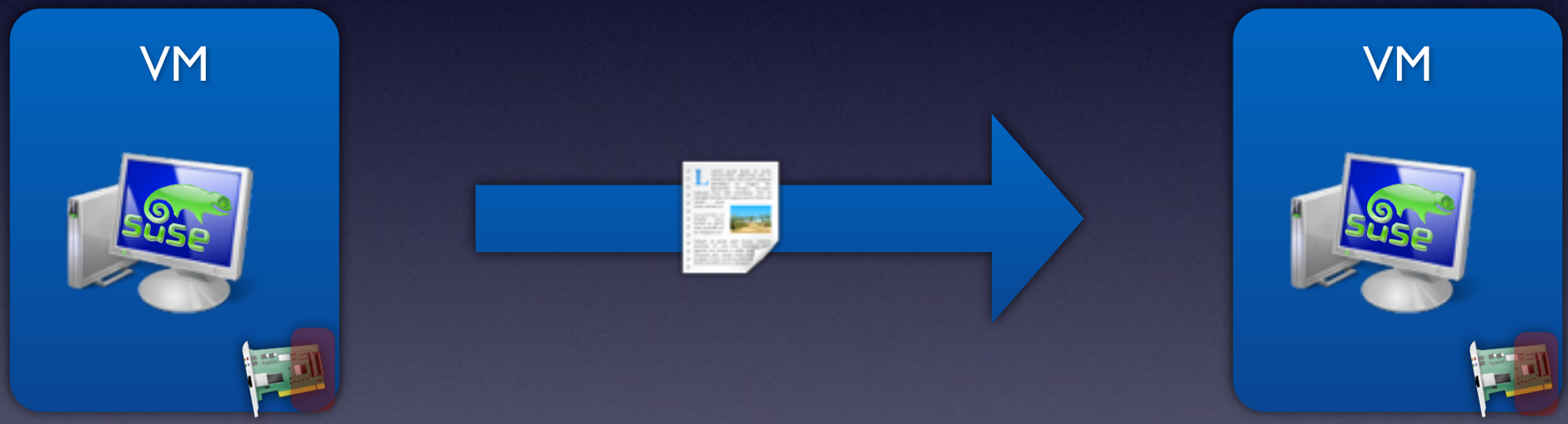
RX ring



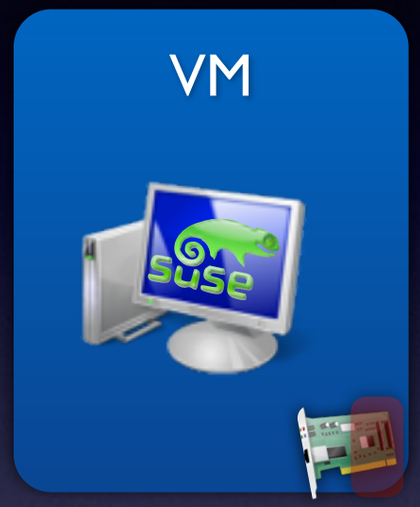
tail

head

VFIO-i40evf

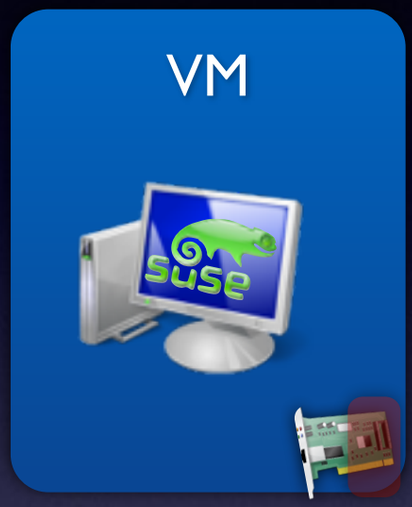


VFIO-i40evf



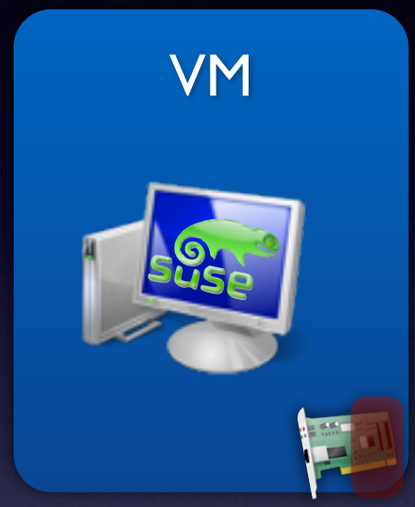
RX ring

VFIO-i40evf



RX ring

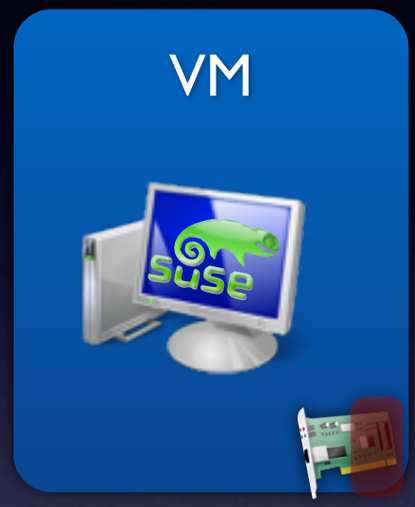
VFIO-i40evf



RX ring

head

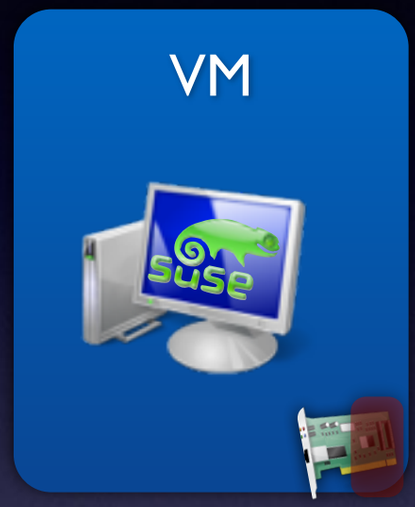
VFIO-i40evf



RX ring

head

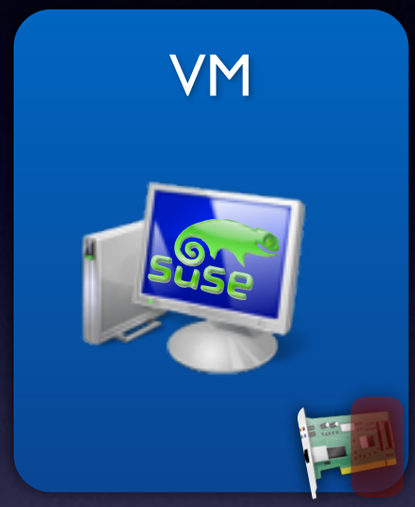
VFIO-i40evf



RX ring

head

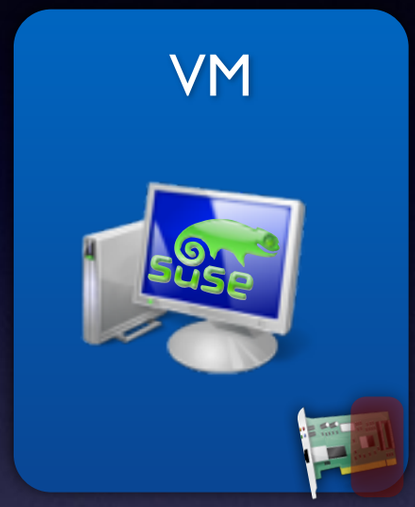
VFIO-i40evf



RX ring

head

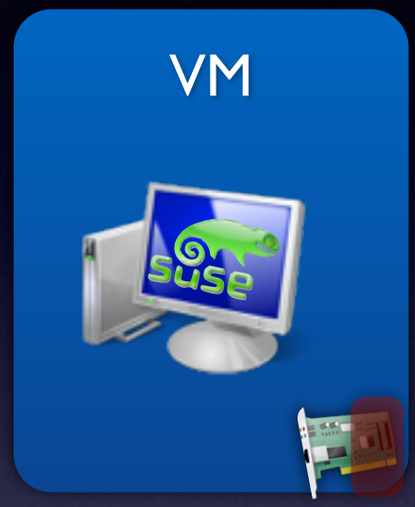
VFIO-i40evf



RX ring

head

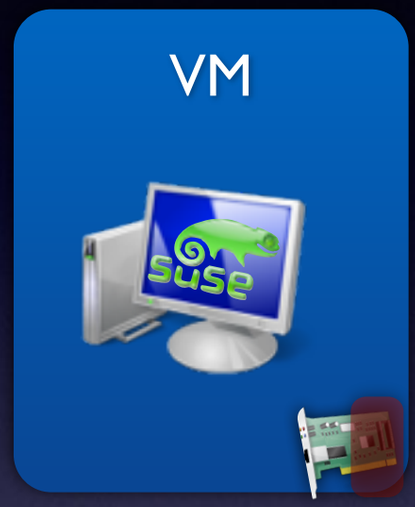
VFIO-i40evf

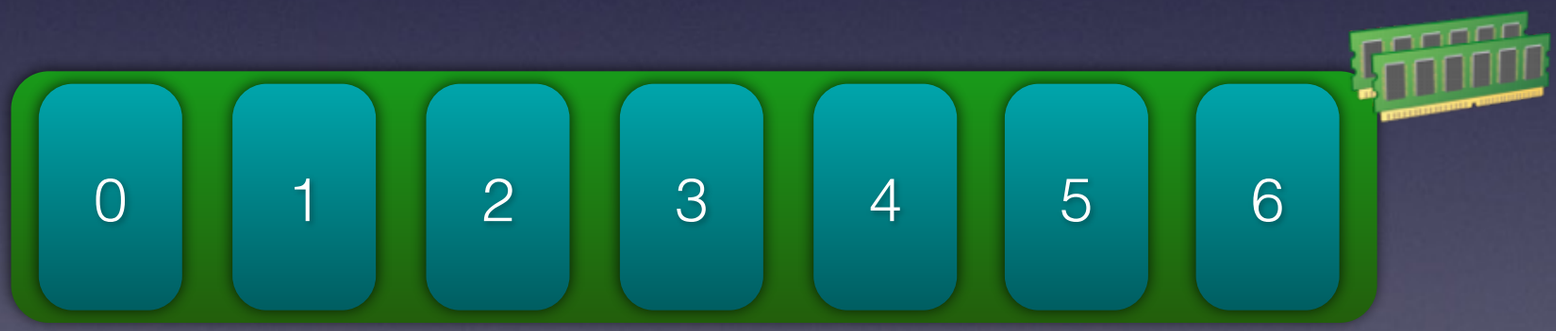


RX ring

head

VFIO-i40evf



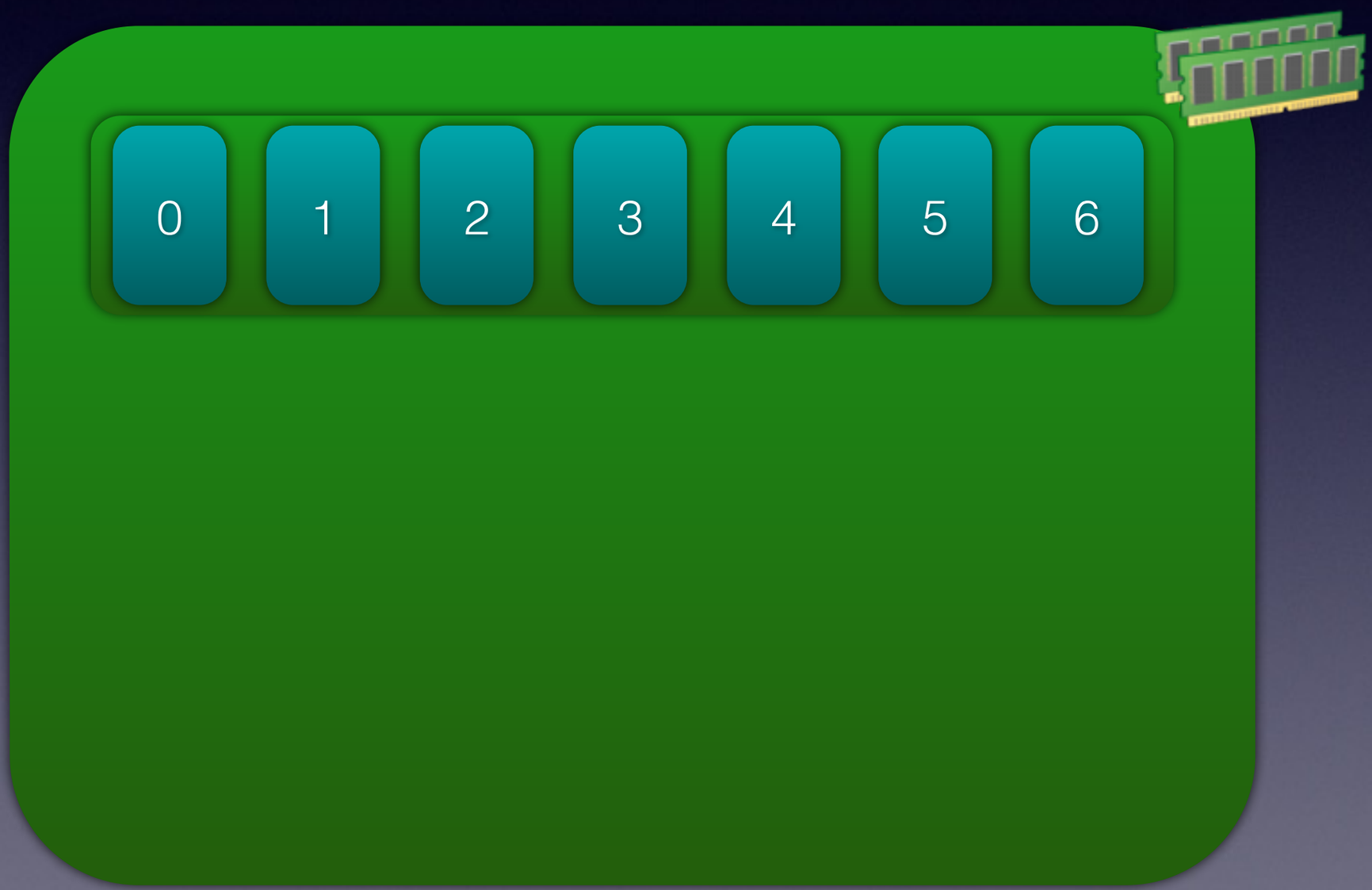


RX ring

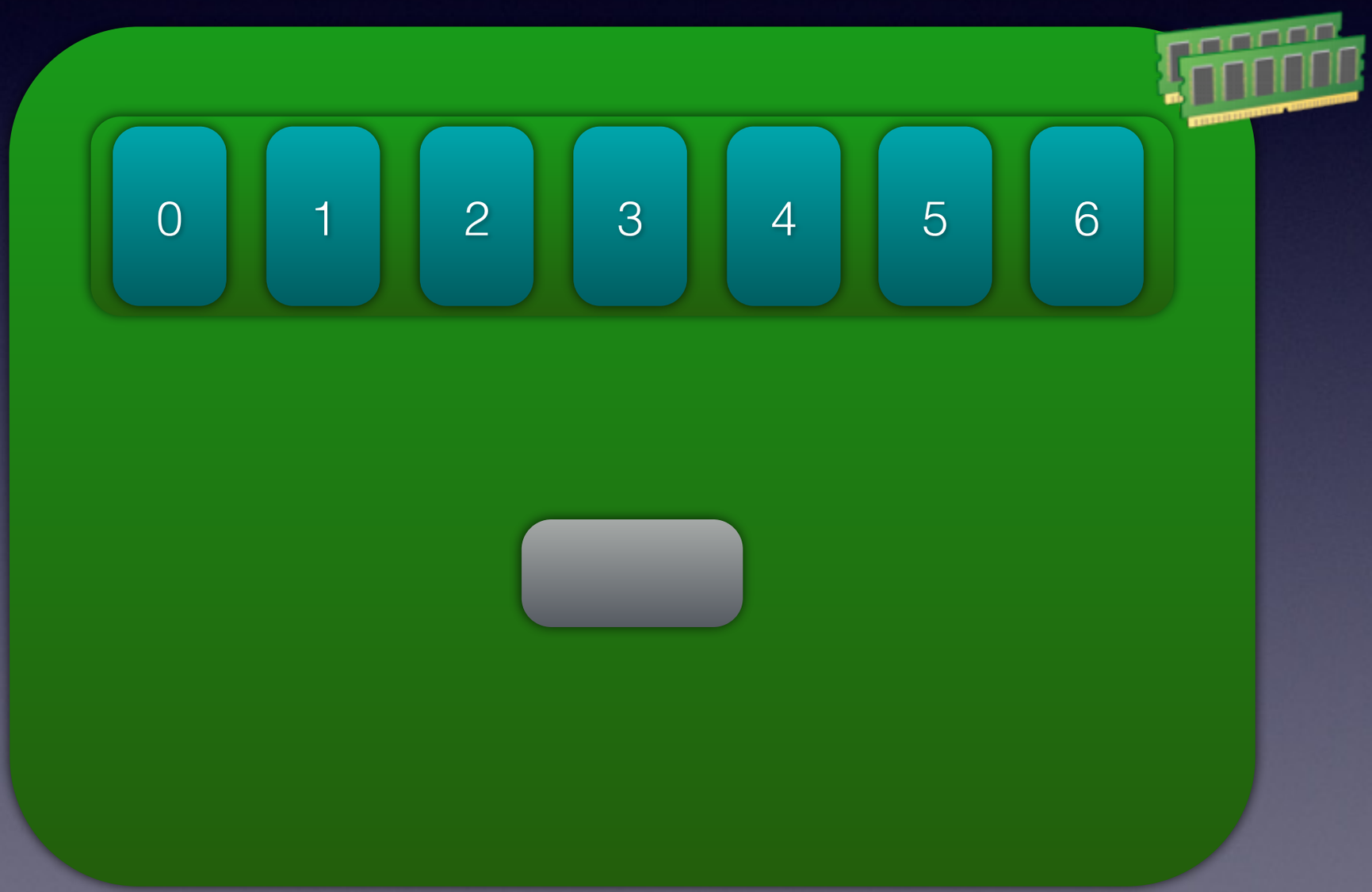
head

tail

DMA



DMA

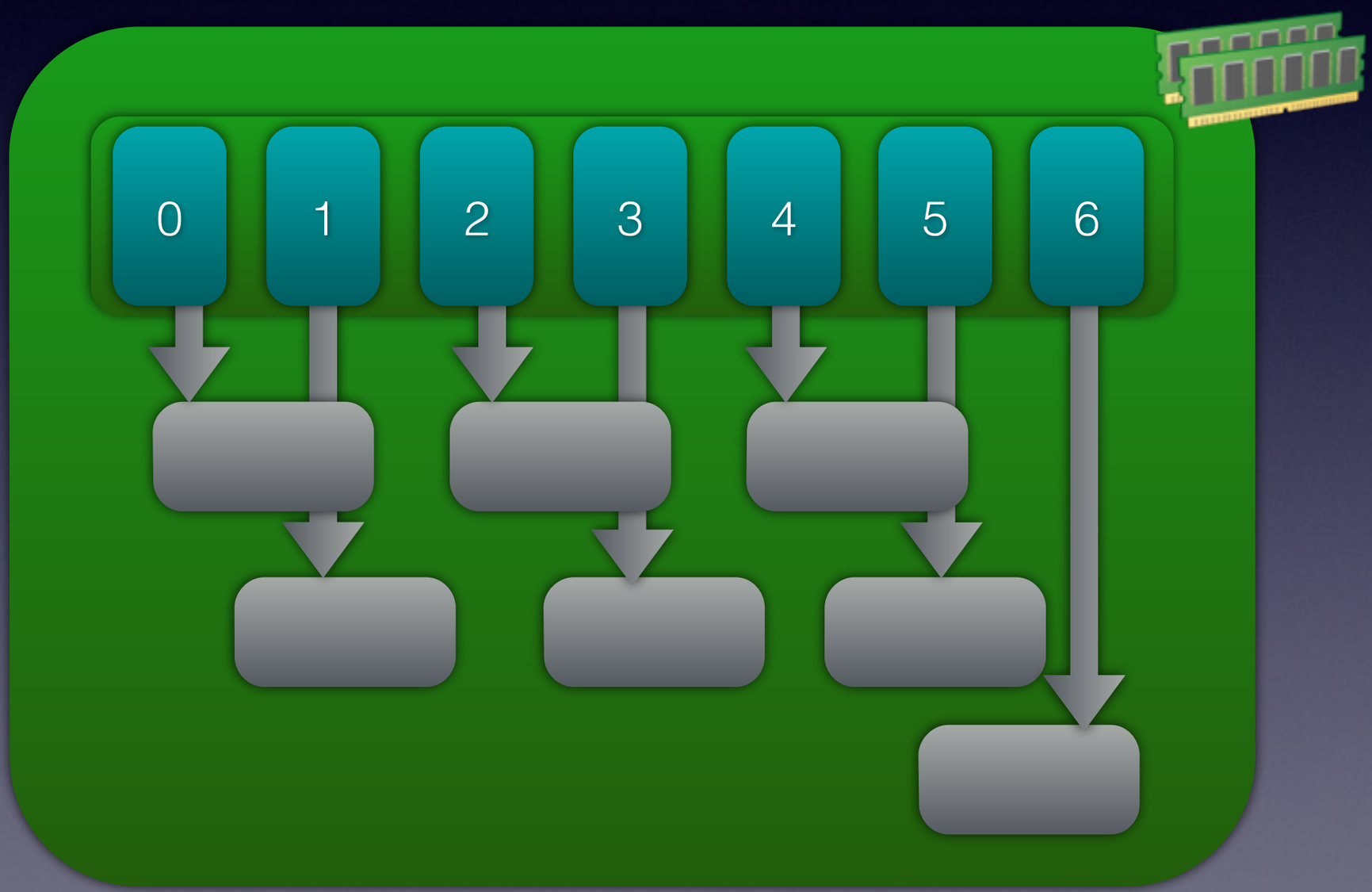

0
1
2
3
4
5
6


345085-0024 BKDHS
CPU
1070338467725-30
AML-0092847-88
9376510375 4H-ALLP4

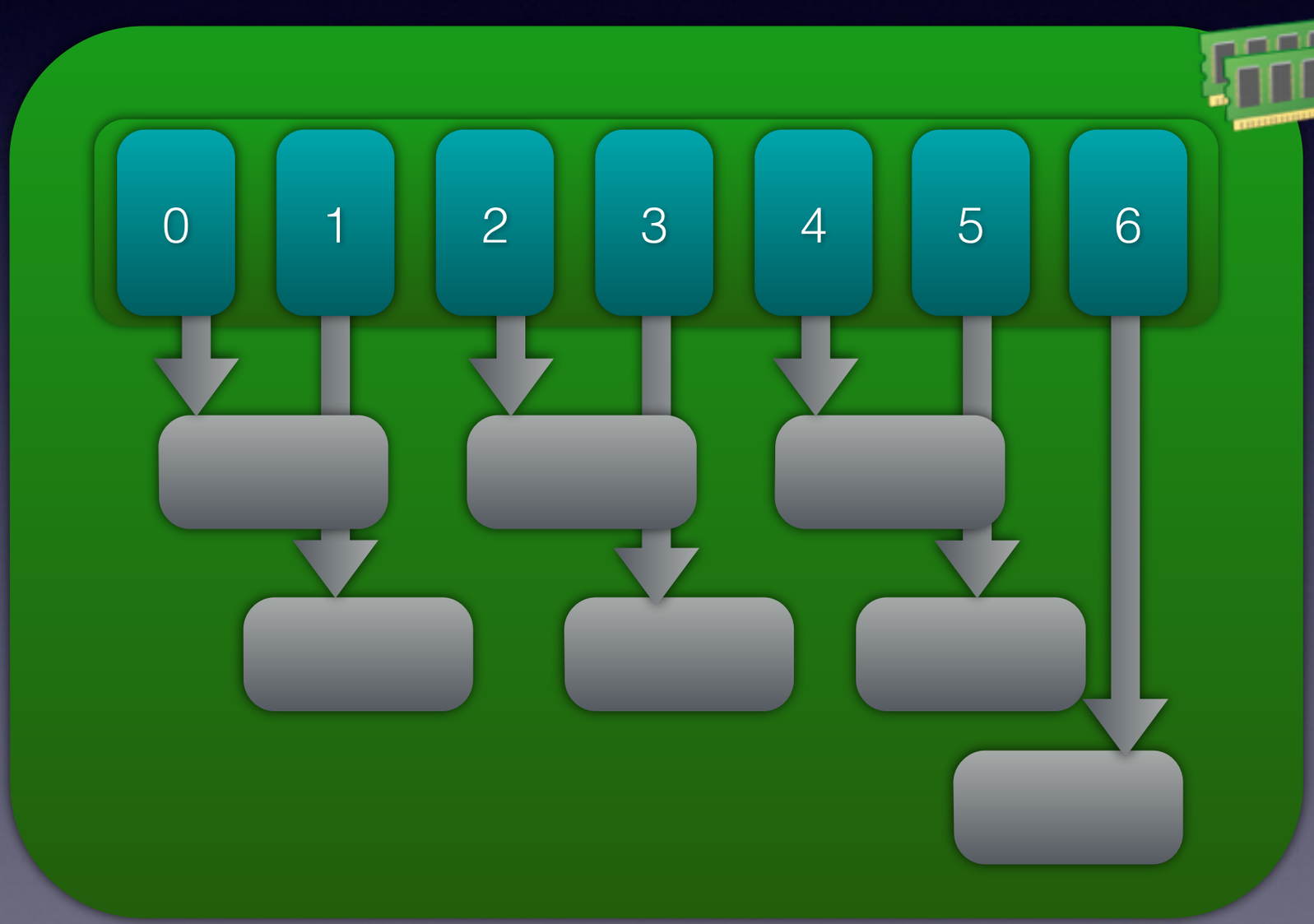
DMA


0
1
2
3
4
5
6
345085-0024 BKDH5
CPU
1070338467725-30
AML-0092847-88
9376510375 4H-ALLP4

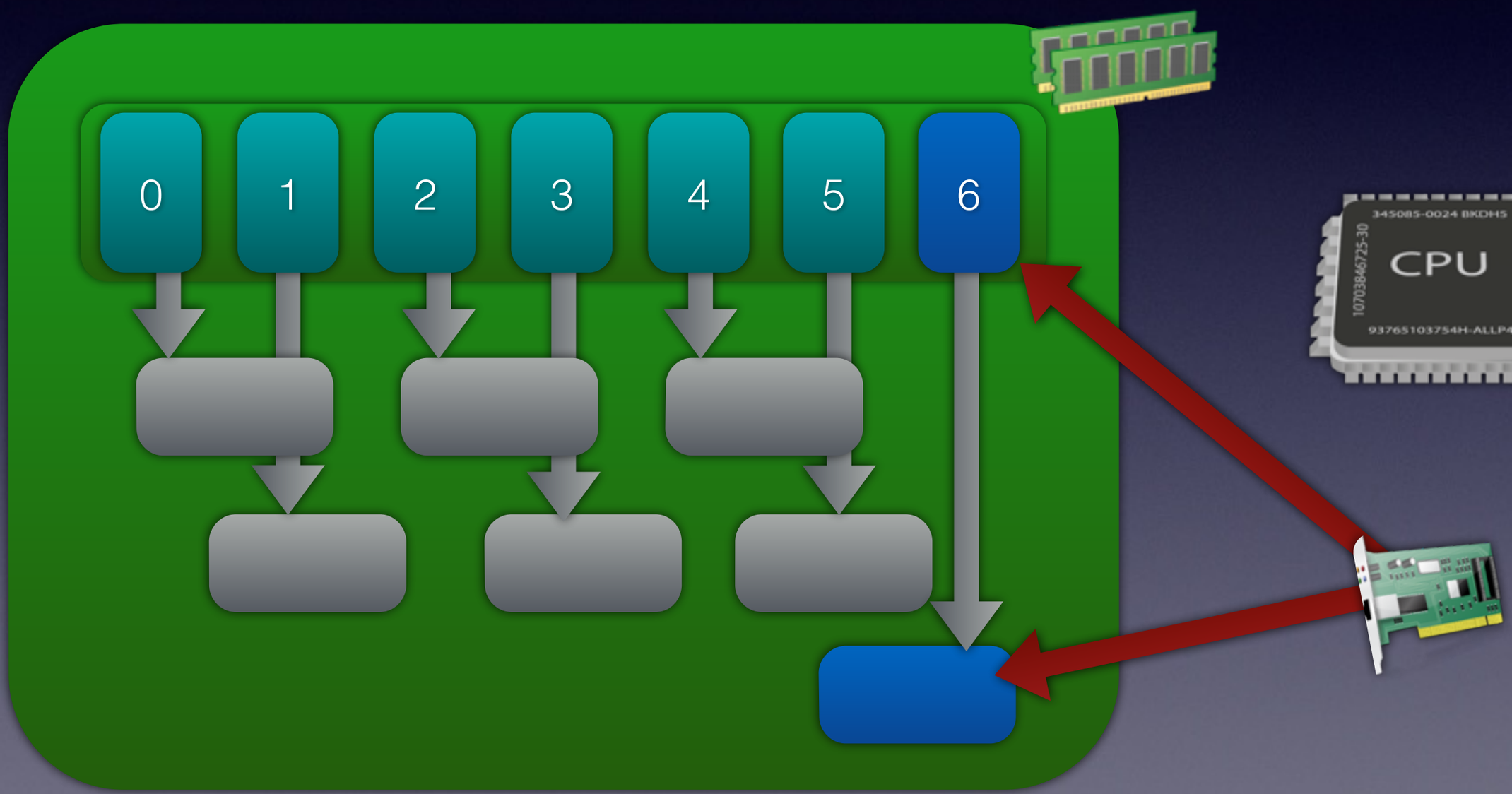
DMA



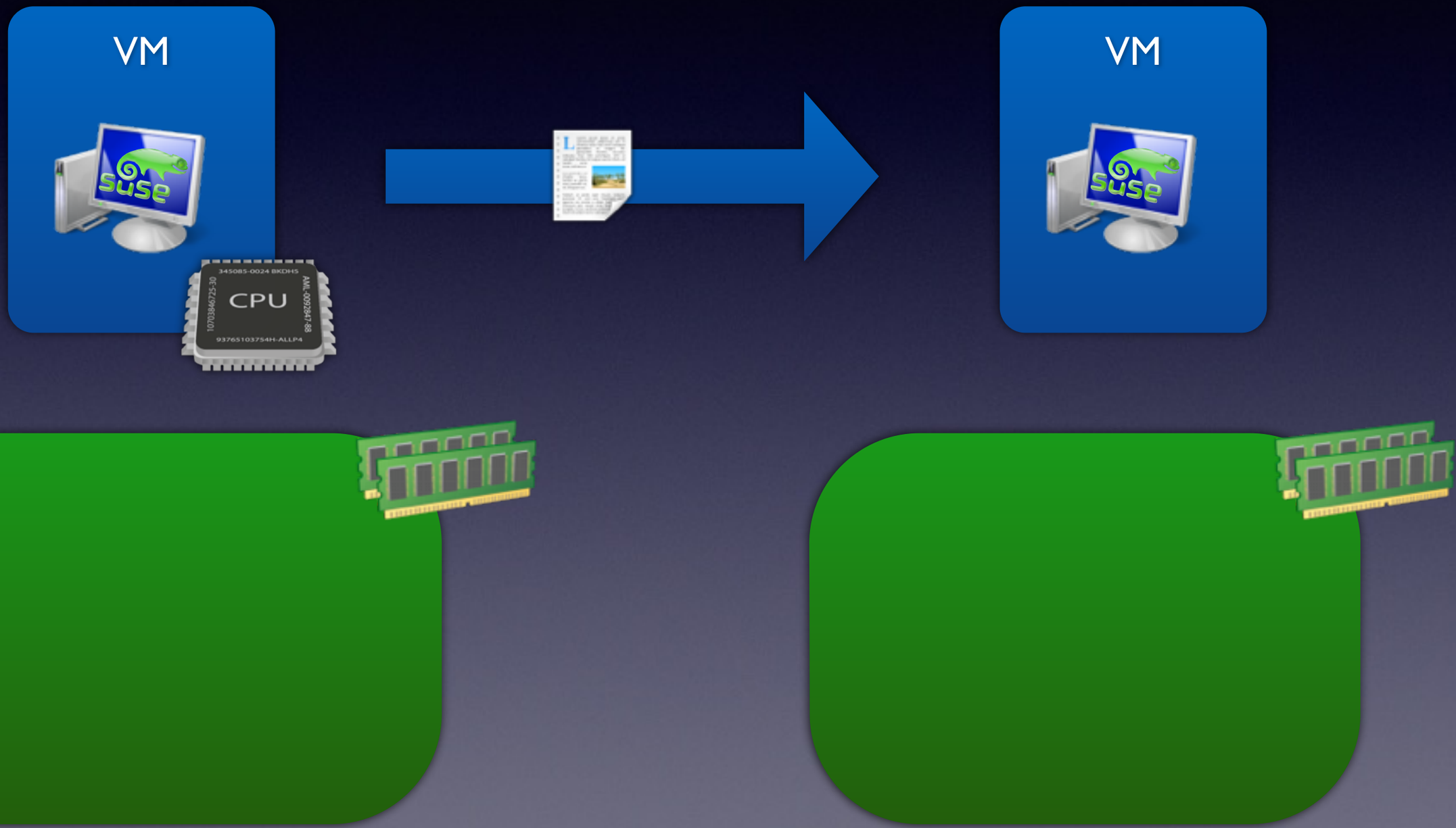
DMA



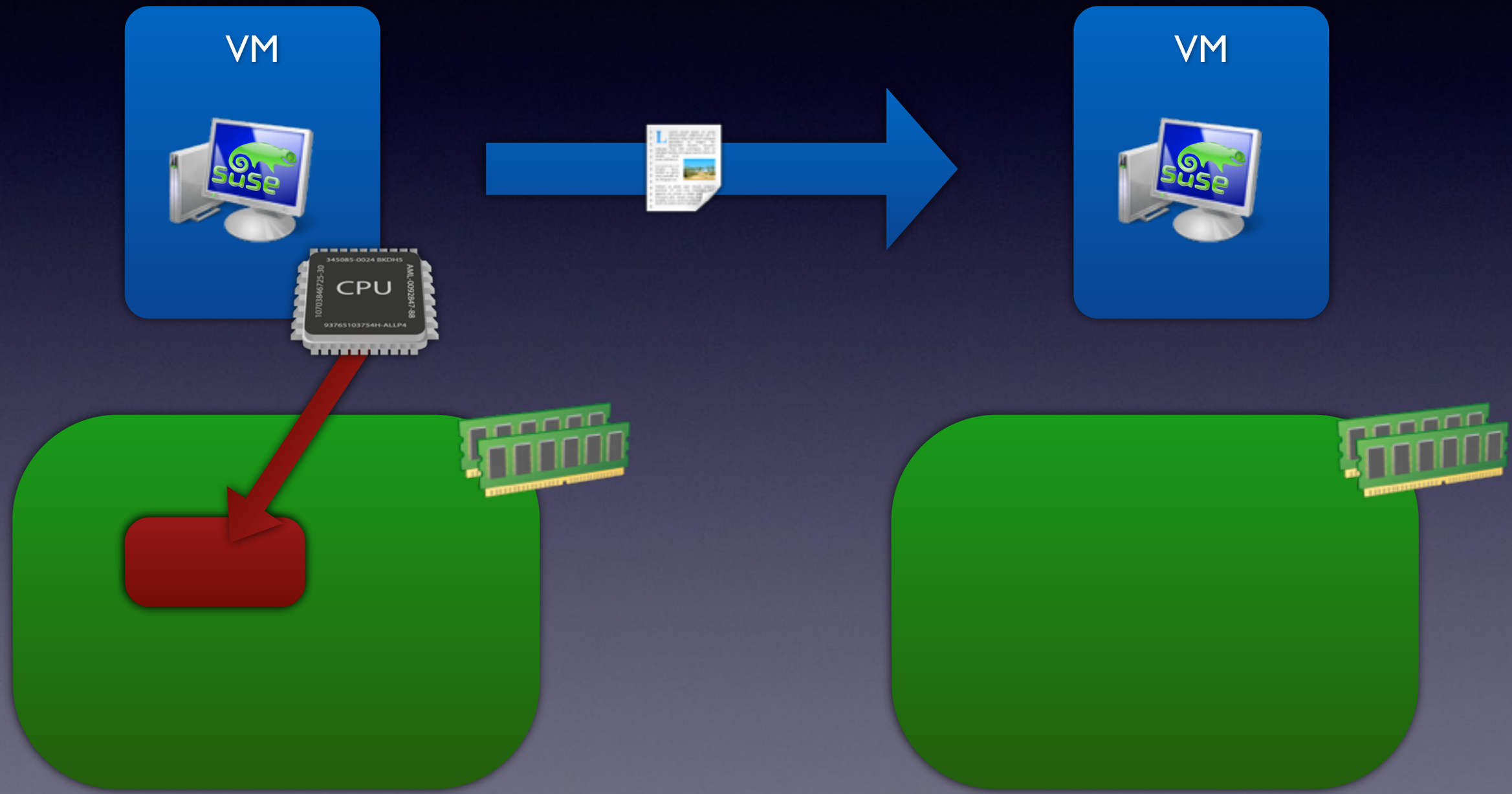
DMA


0
1
2
3
4
5
6
CPU
345085-0024 BKDHS
AML-0092847-88
937651037S4H-ALLP4
1070384672S-30

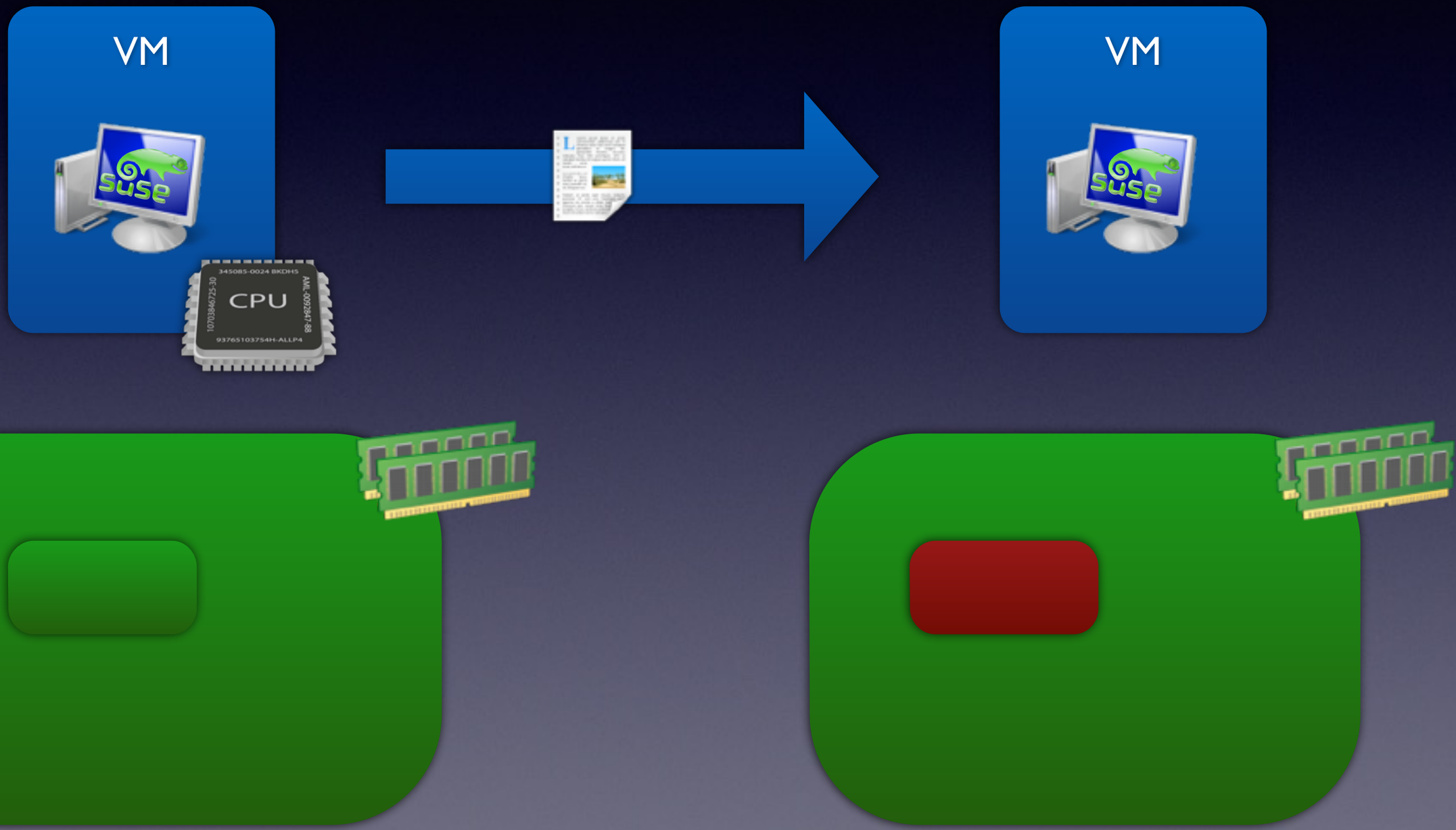
VFIO-i40evf



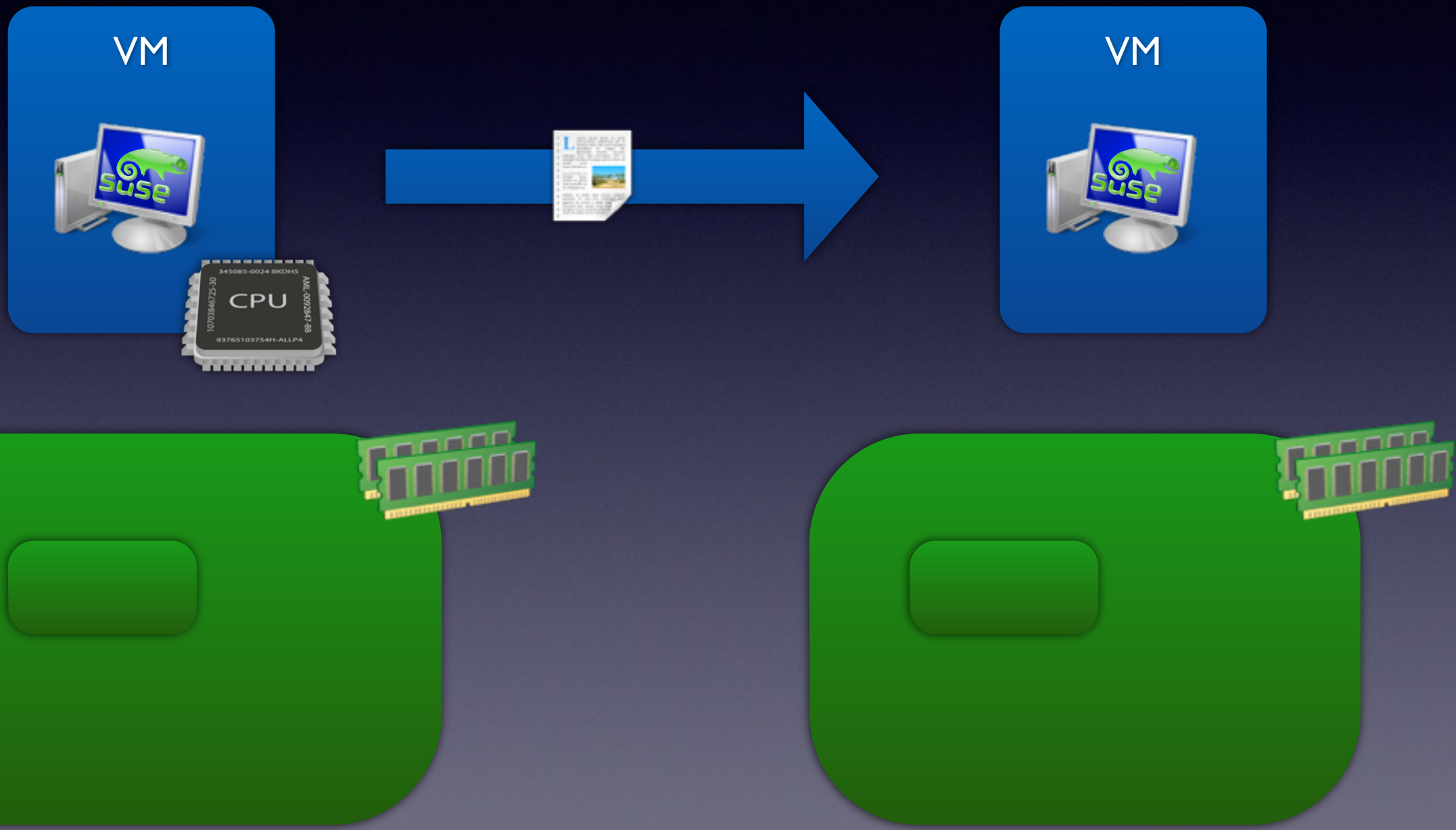
VFIO-i40evf



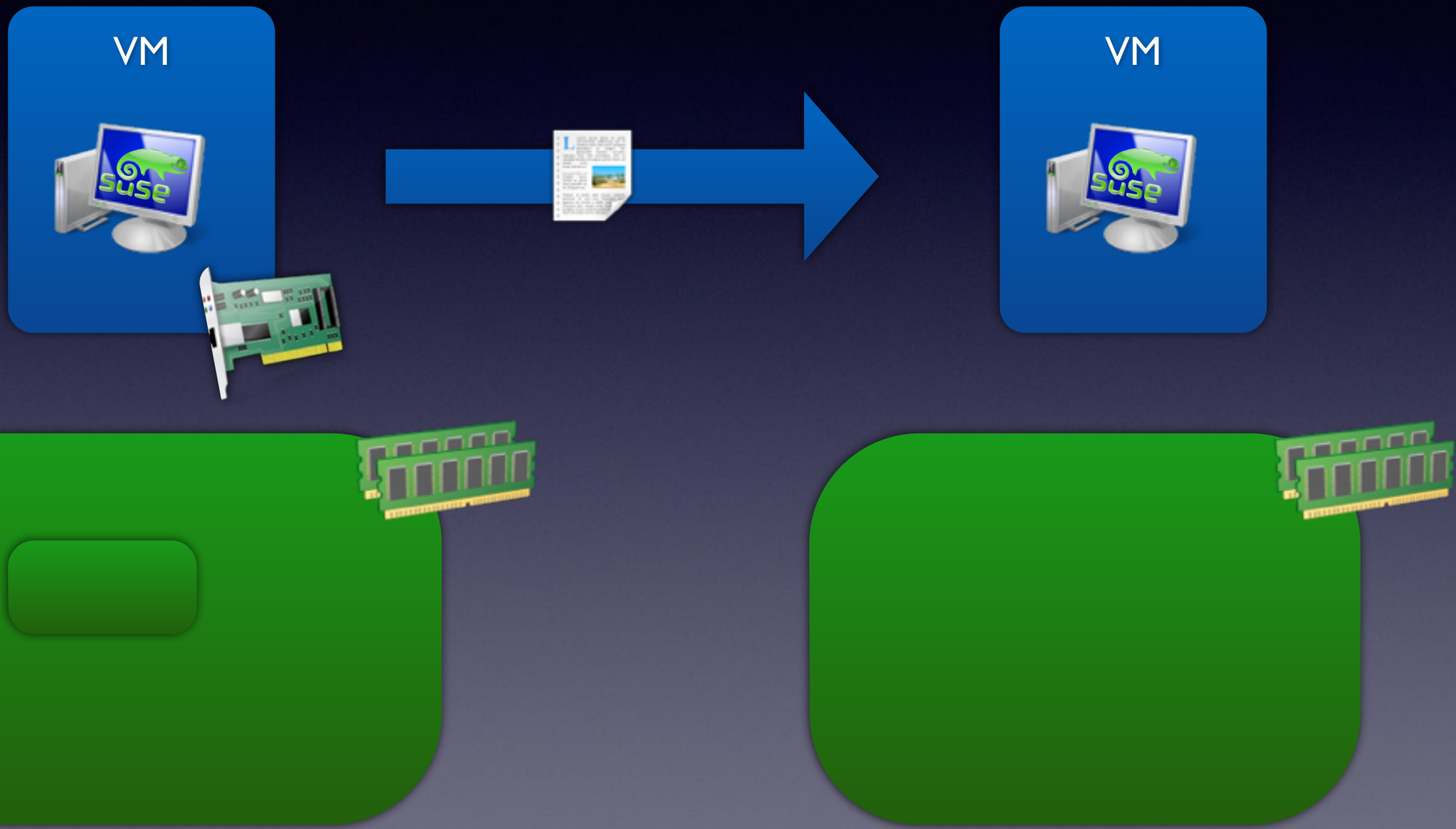
VFIO-i40evf



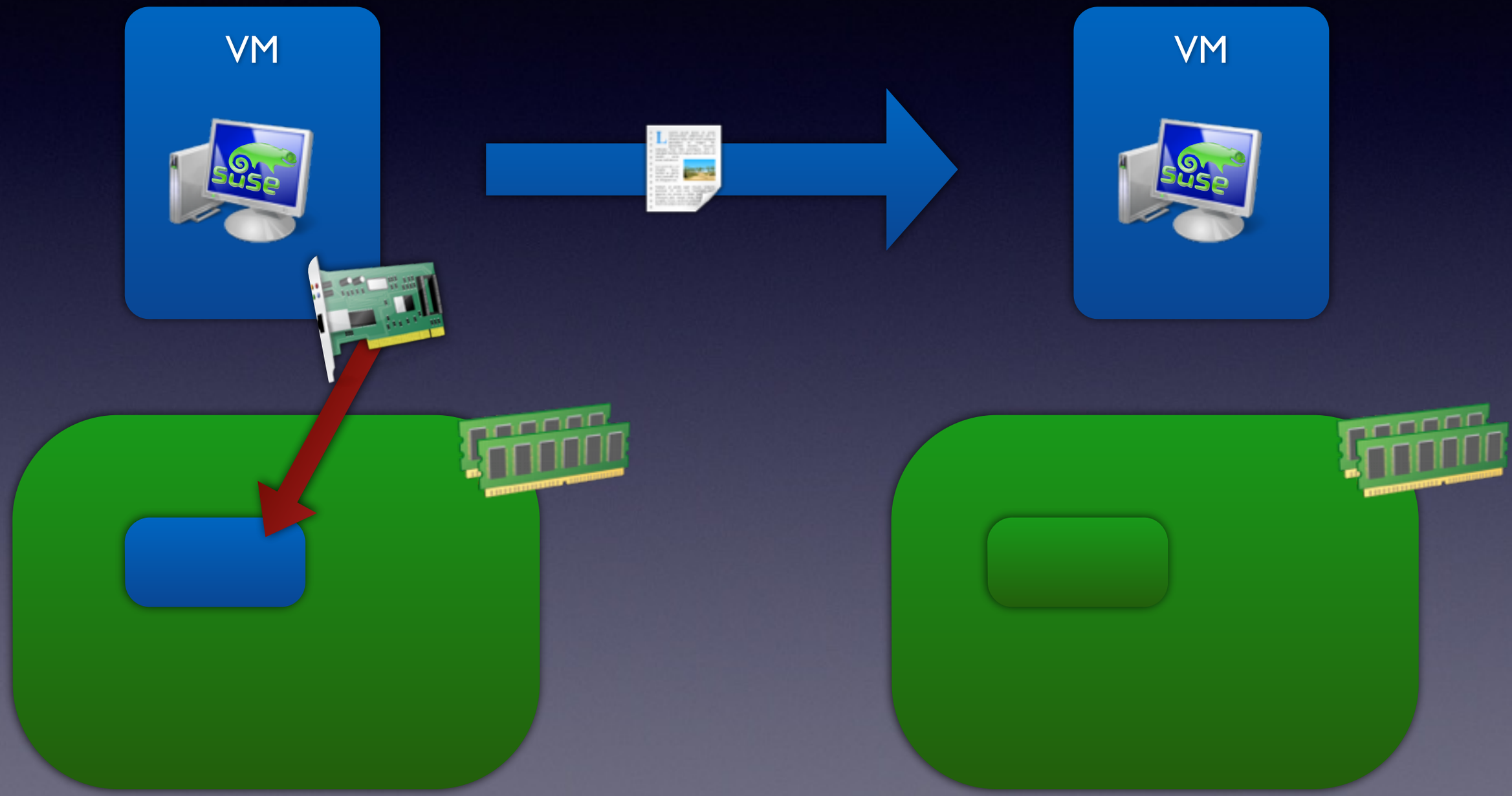
VFIO-i40evf



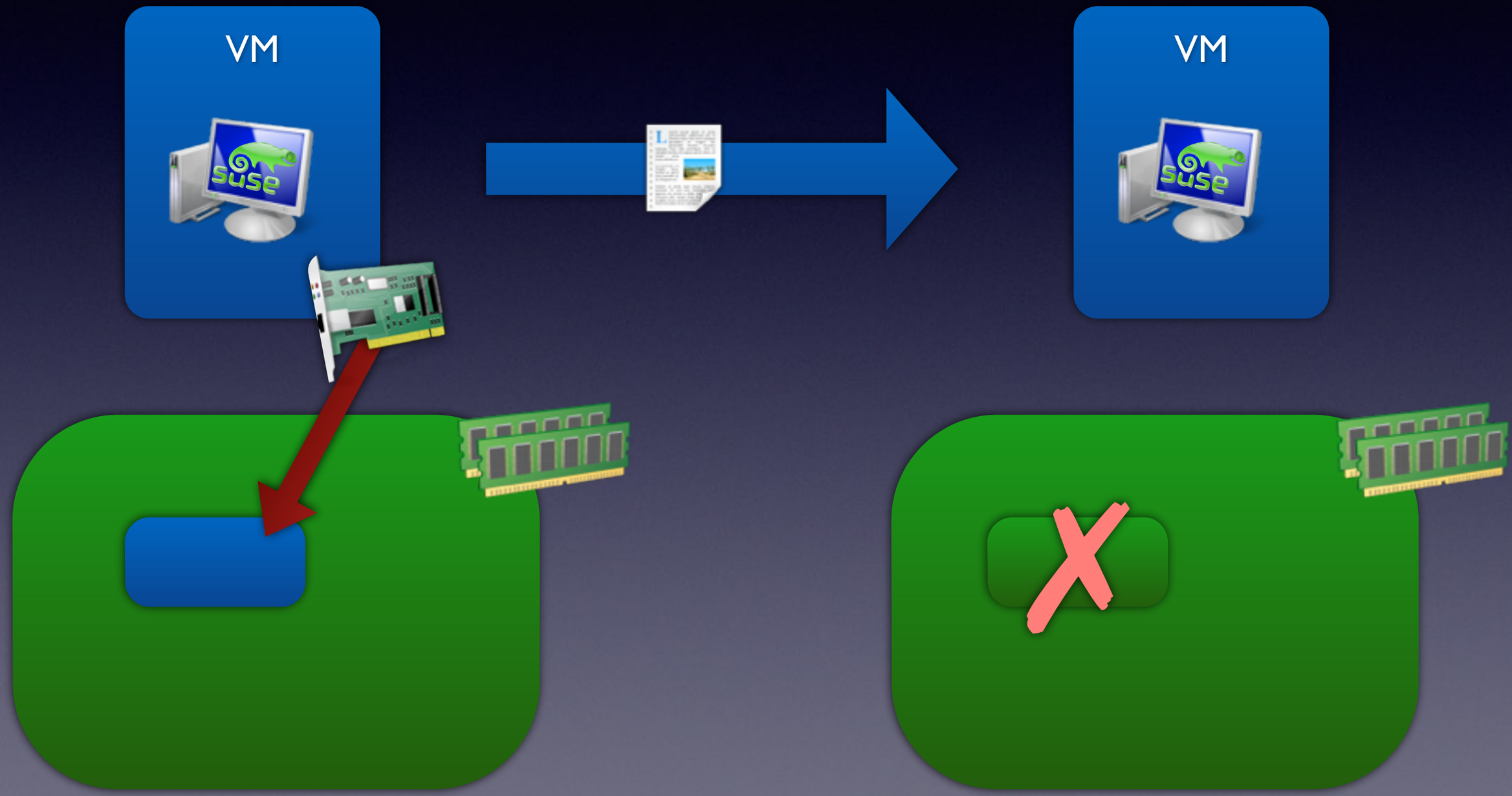
VFIO-i40evf



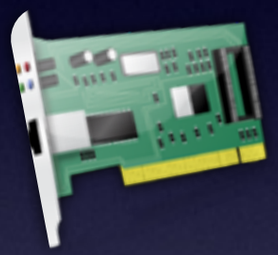
VFIO-i40evf



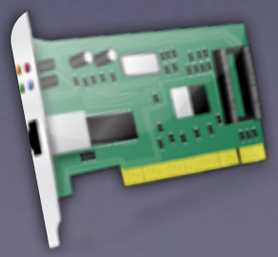
VFIO-i40evf



VFIO-i40evf



guest NIC



real NIC

TX ring

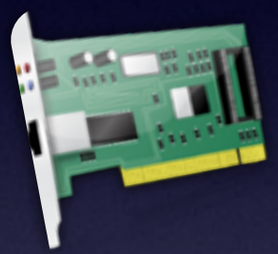
RX ring

Admin ring

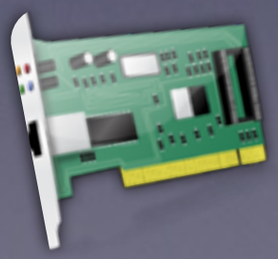
QEMU

Admin ring

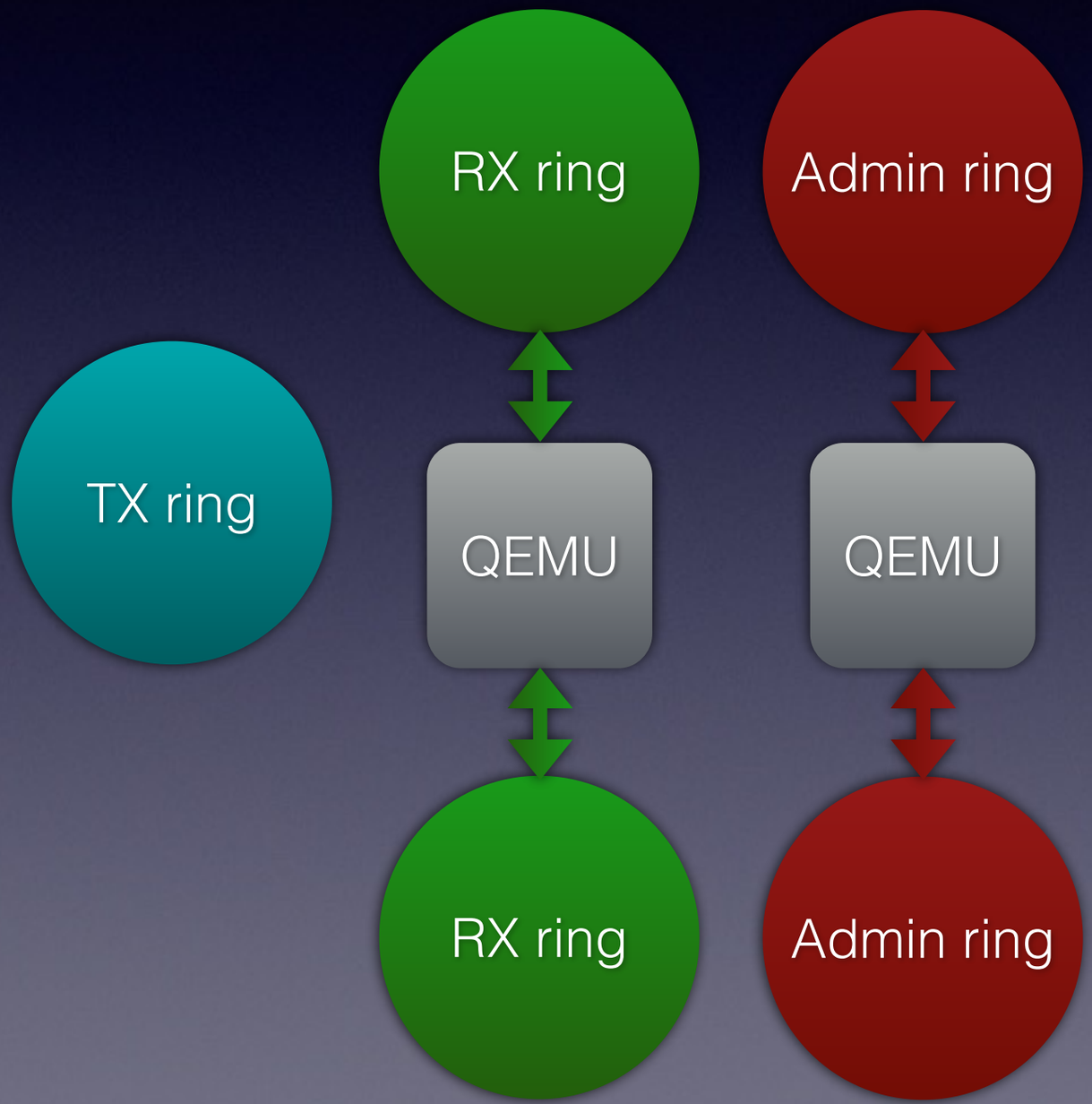
VFIO-i40evf



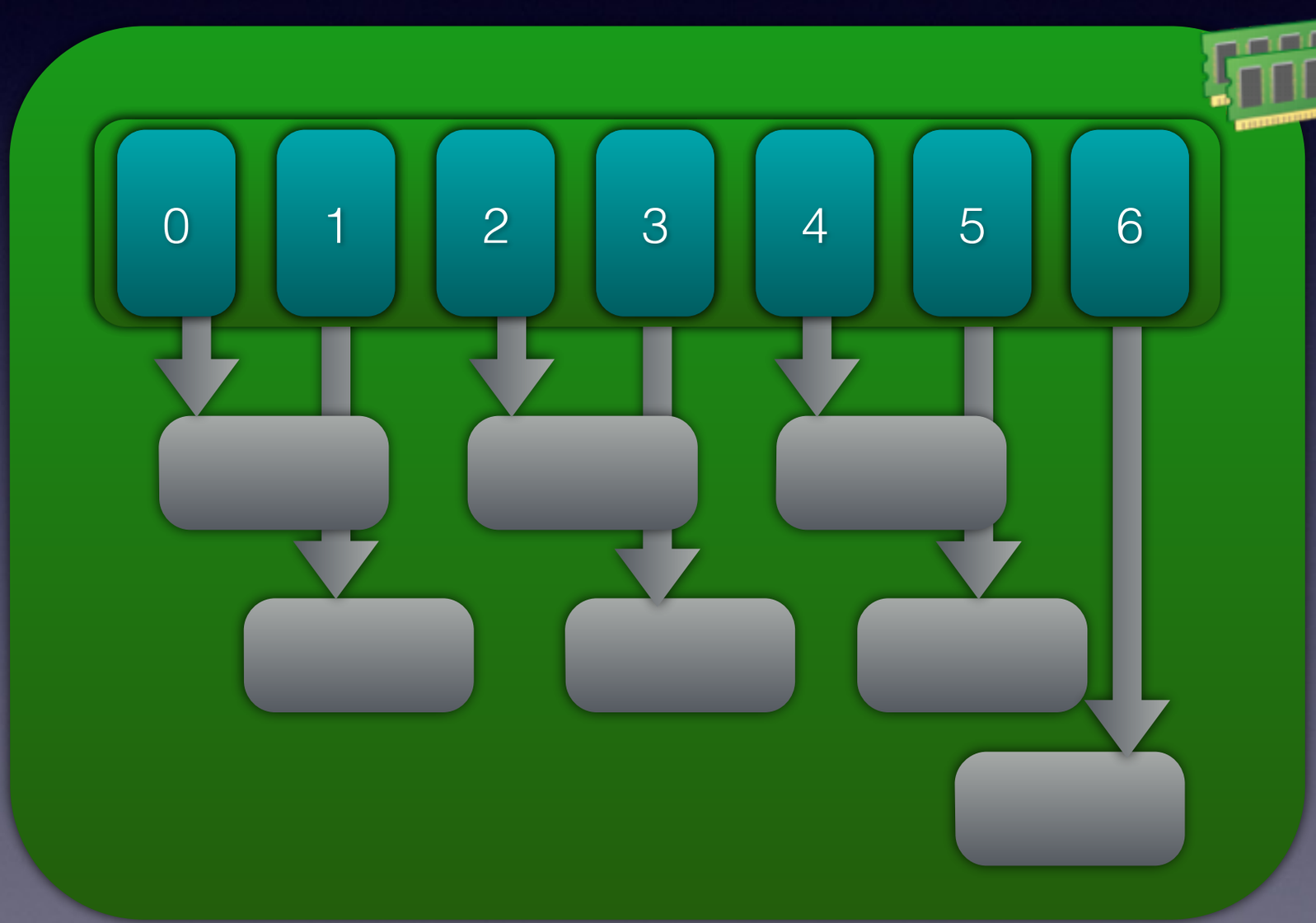
guest NIC



real NIC

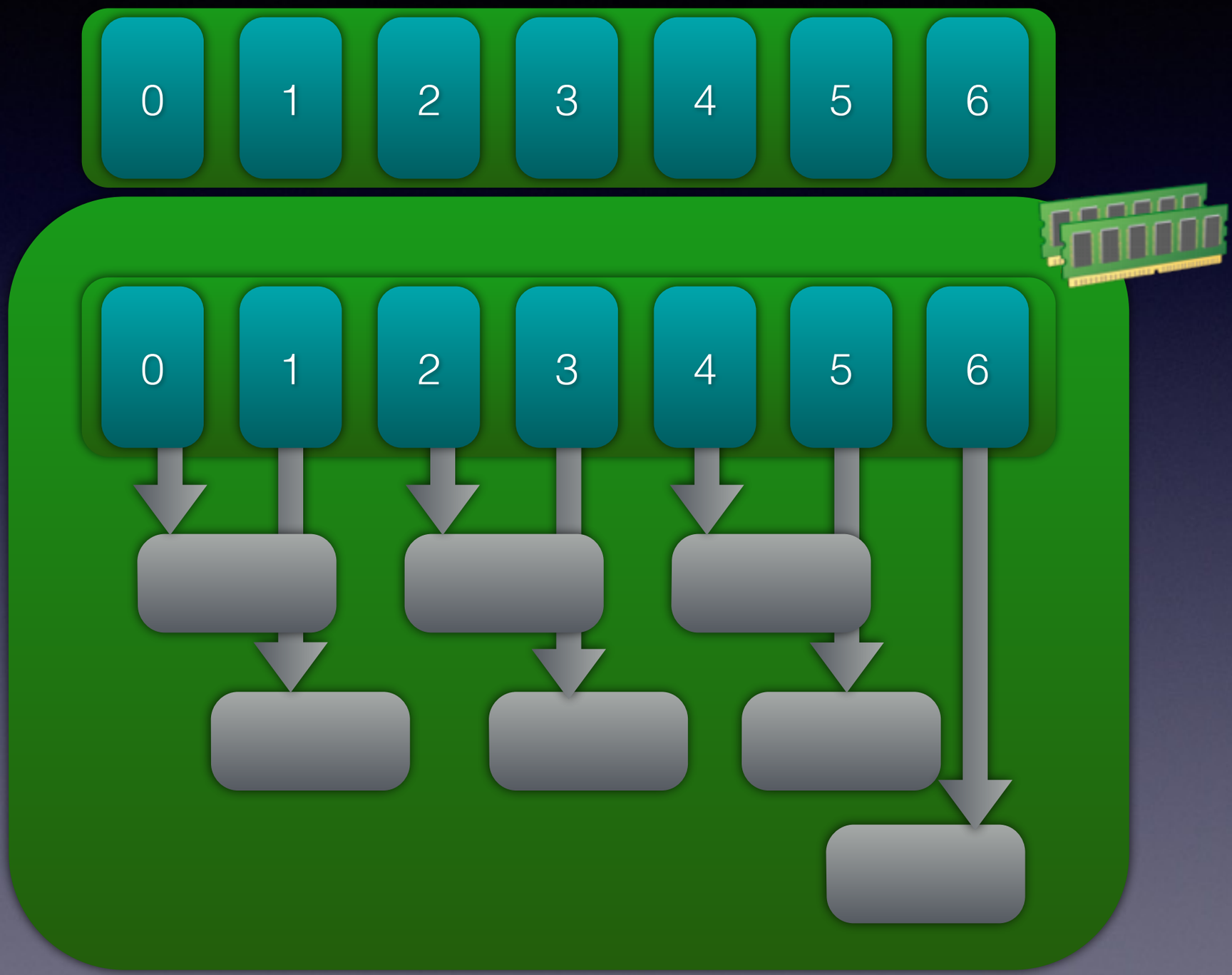


VFIO-i40evf



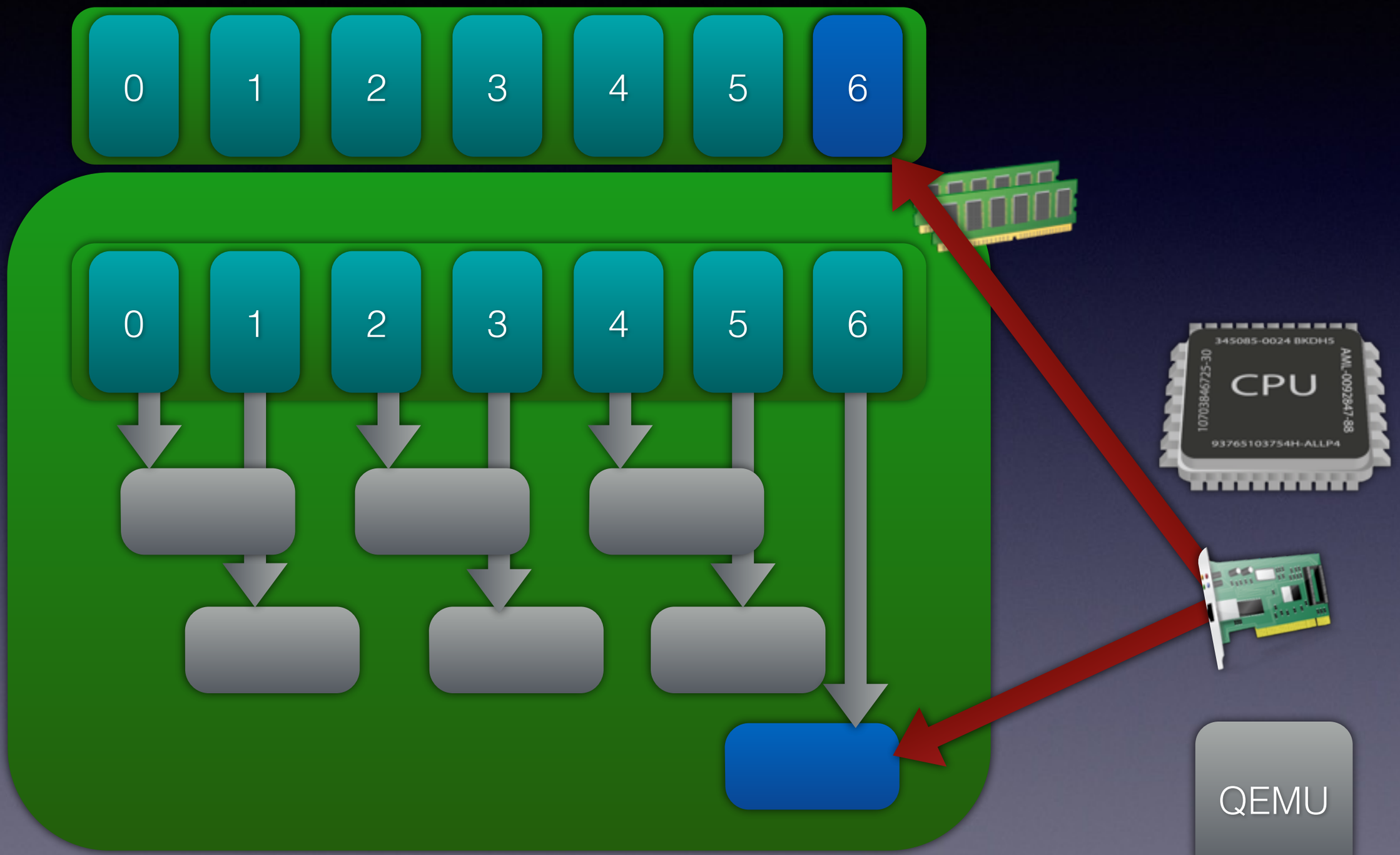
QEMU

VFIO-i40evf

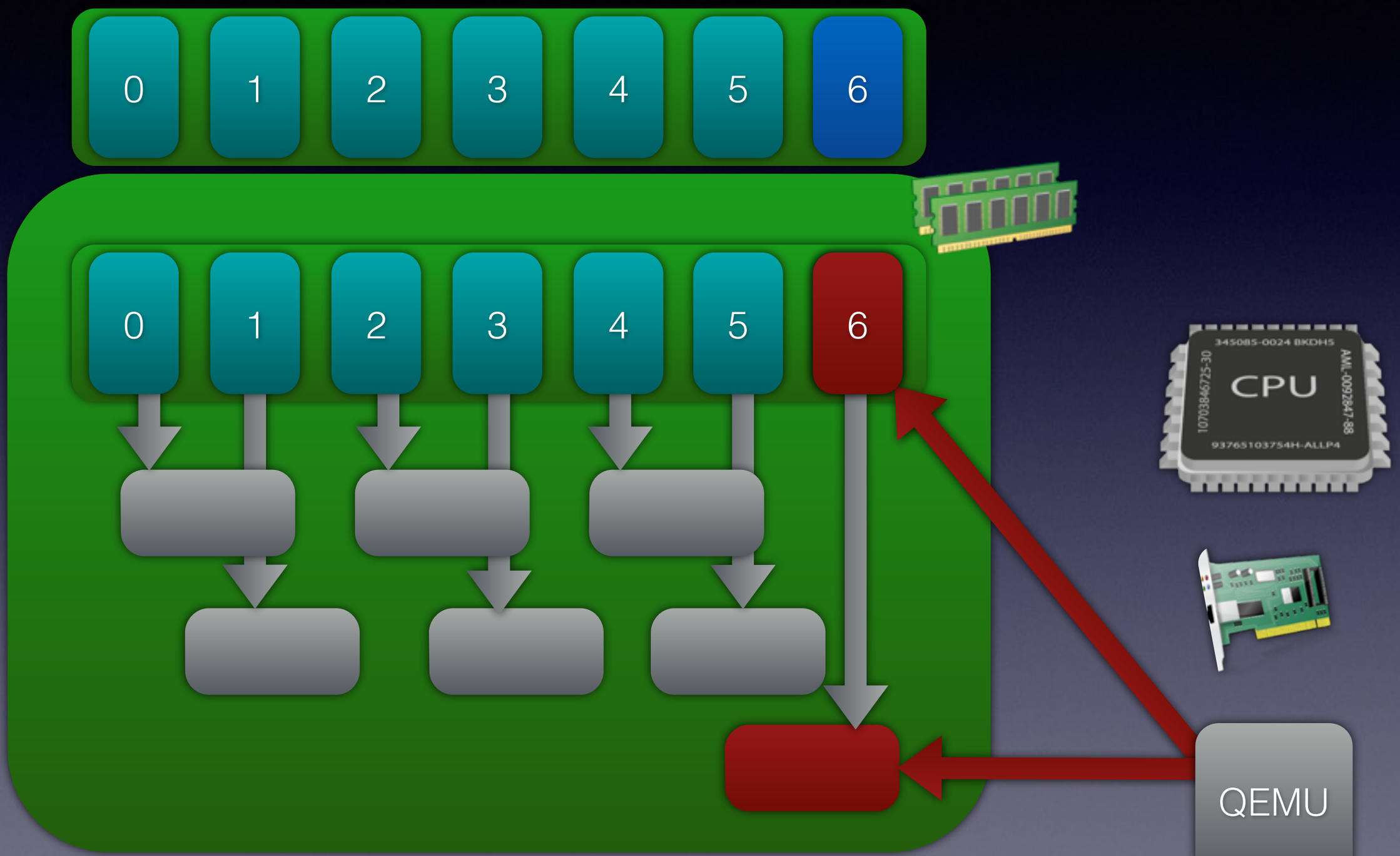


QEMU

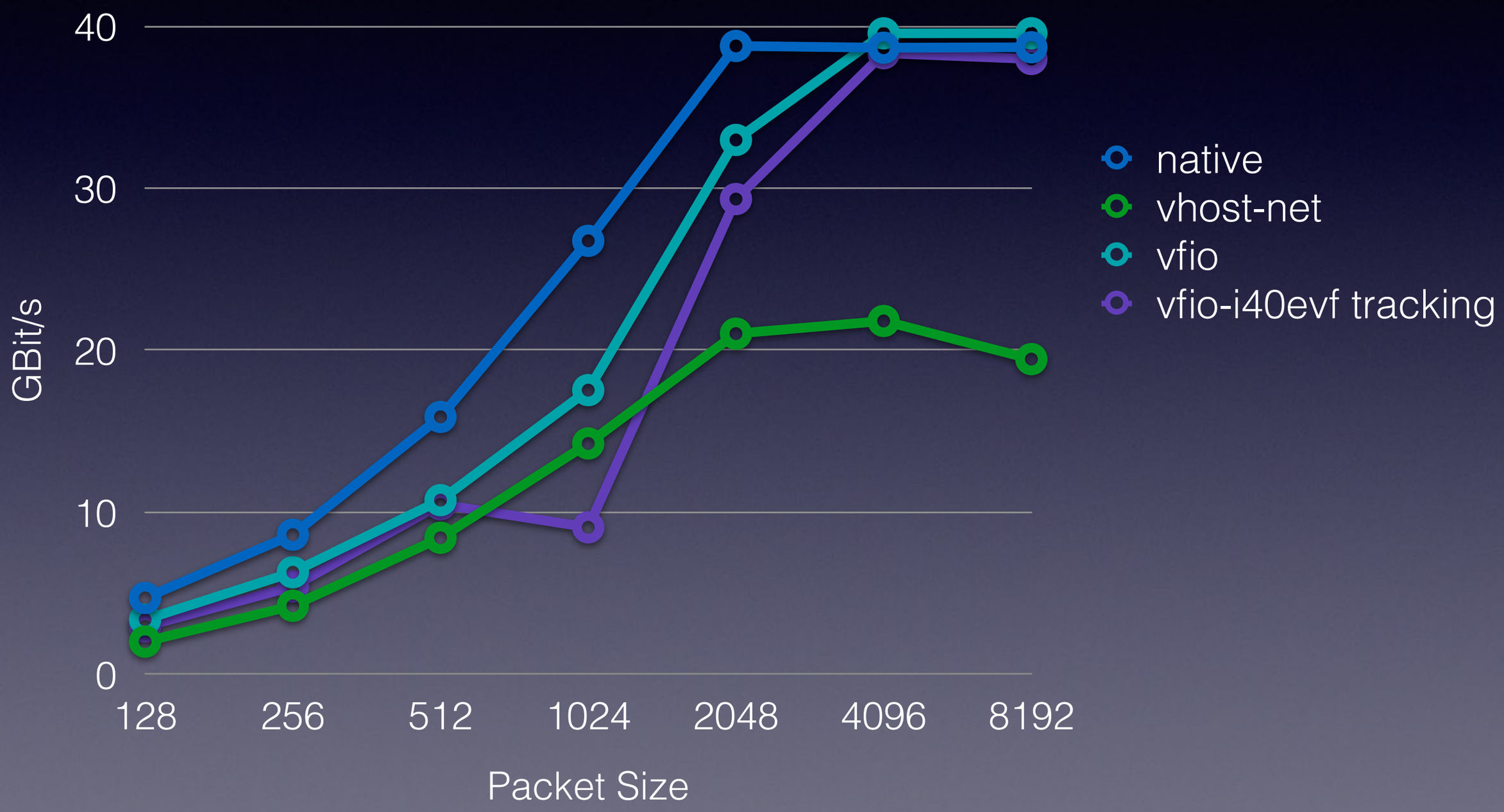
VFIO-i40evf


0
1
2
3
4
5
6
0
1
2
3
4
5
6
345085-0024 BKDHS
CPU
1070338467725-30
AML-0092847-88
9376510375 4H-ALLP4
QEMU

VFIO-i40evf


0
1
2
3
4
5
6
0
1
2
3
4
5
6
CPU
345085-0024 BKDHS
AML-0092847-88
937651037S4H-ALLP4
QEMU

VFIO-i40evf

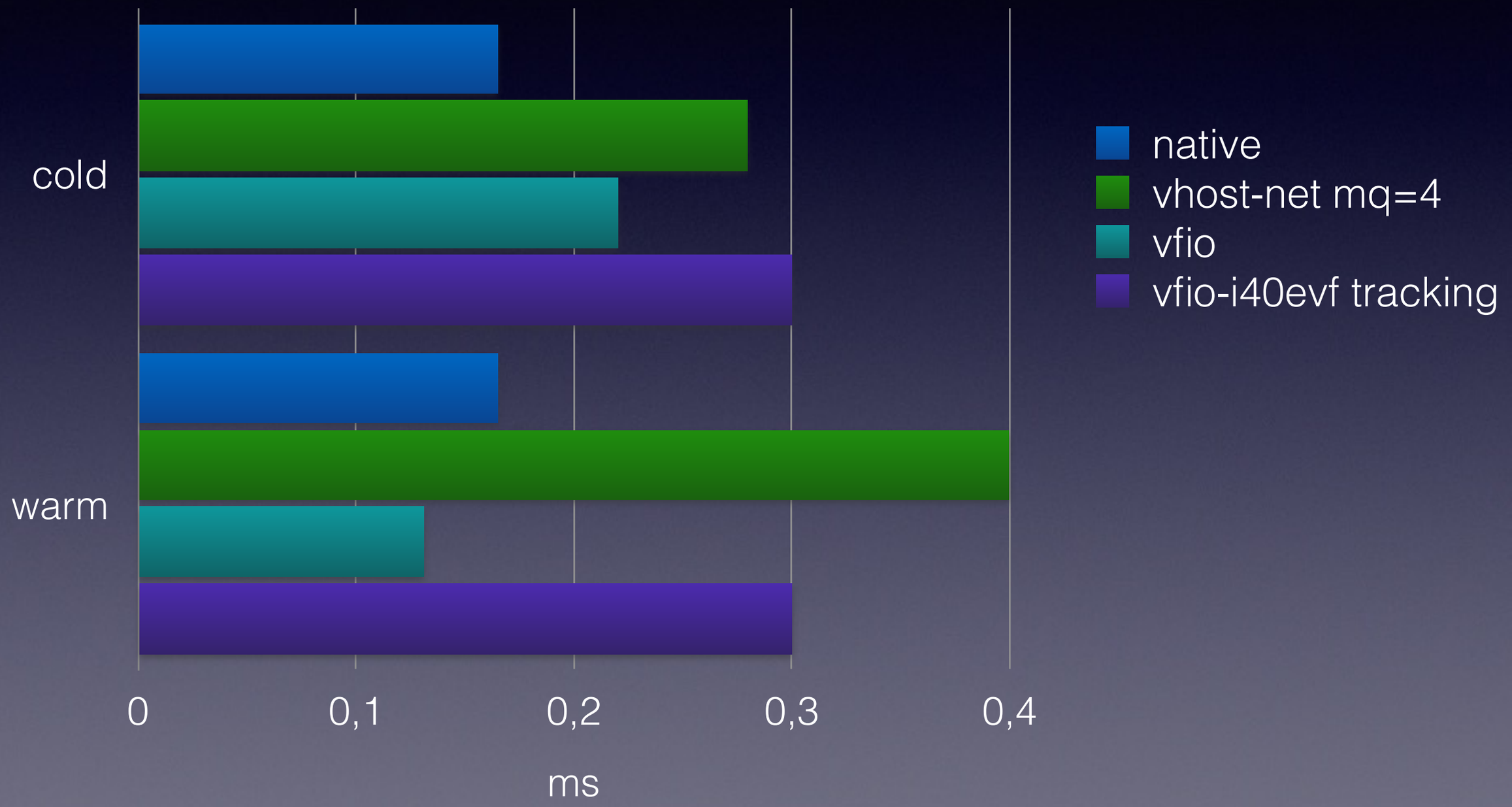

40
30
20
10
0
GBit/s
128
256
512
1024
2048
4096
8192
Packet Size
native
vhost-net
vfio
vfio-i40evf tracking

VFIO-i40evf

- Live Migration works
- No memory overcommit
- Performance identical to VFIO in normal operation
- Throughput close to VFIO during migration

Latency

Latency


cold
warm
0
0,1
0,2
0,3
0,4
ms
native
vhost-net mq=4
vfio
vfio-i40evf tracking

Other ideas

Hardware aware vhost

virtio-net

TX ring

RX ring

vhost queue driver

Linux tap

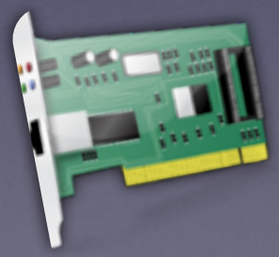
vhost

Linux bridge

Linux NIC driver

TX ring

RX ring



real NIC

virtio-net

TX ring

RX ring

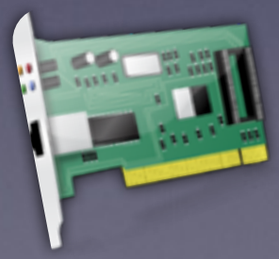
vhost queue driver

vhost

Linux NIC driver

TX ring

RX ring



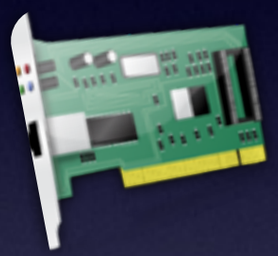
real NIC

Hardware aware vhost

- Vhost talks directly to NIC driver
- NIC driver allocates queues for Vhost
- No guest exposure to real NIC

Virtio Queue Format Extension

Virtio Queue Format Extension

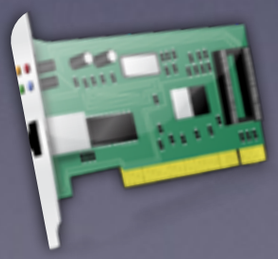


virtio-net

TX ring

RX ring

Admin ring



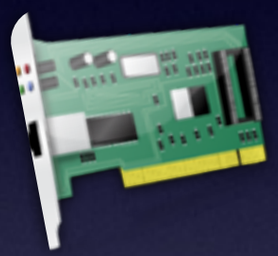
real NIC

TX ring

RX ring

Admin ring

Virtio Queue Format Extension



virtio-net

TX ring

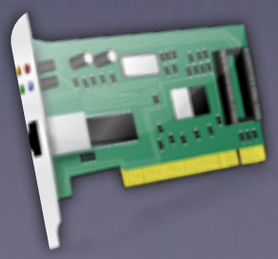
RX ring

Admin ring

TX ring

RX ring

Admin ring



real NIC

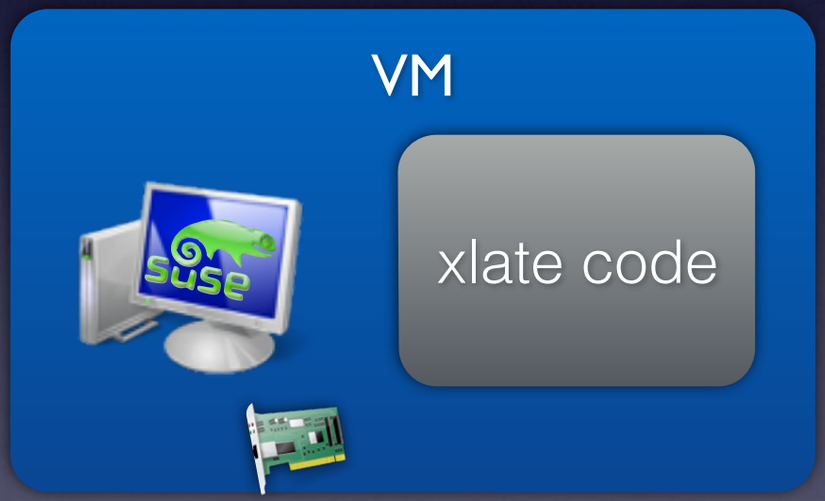
Virtio Queue Format Extension

- Guest sees virtio device with additional BAR
- New BAR provides direct access to HW queues

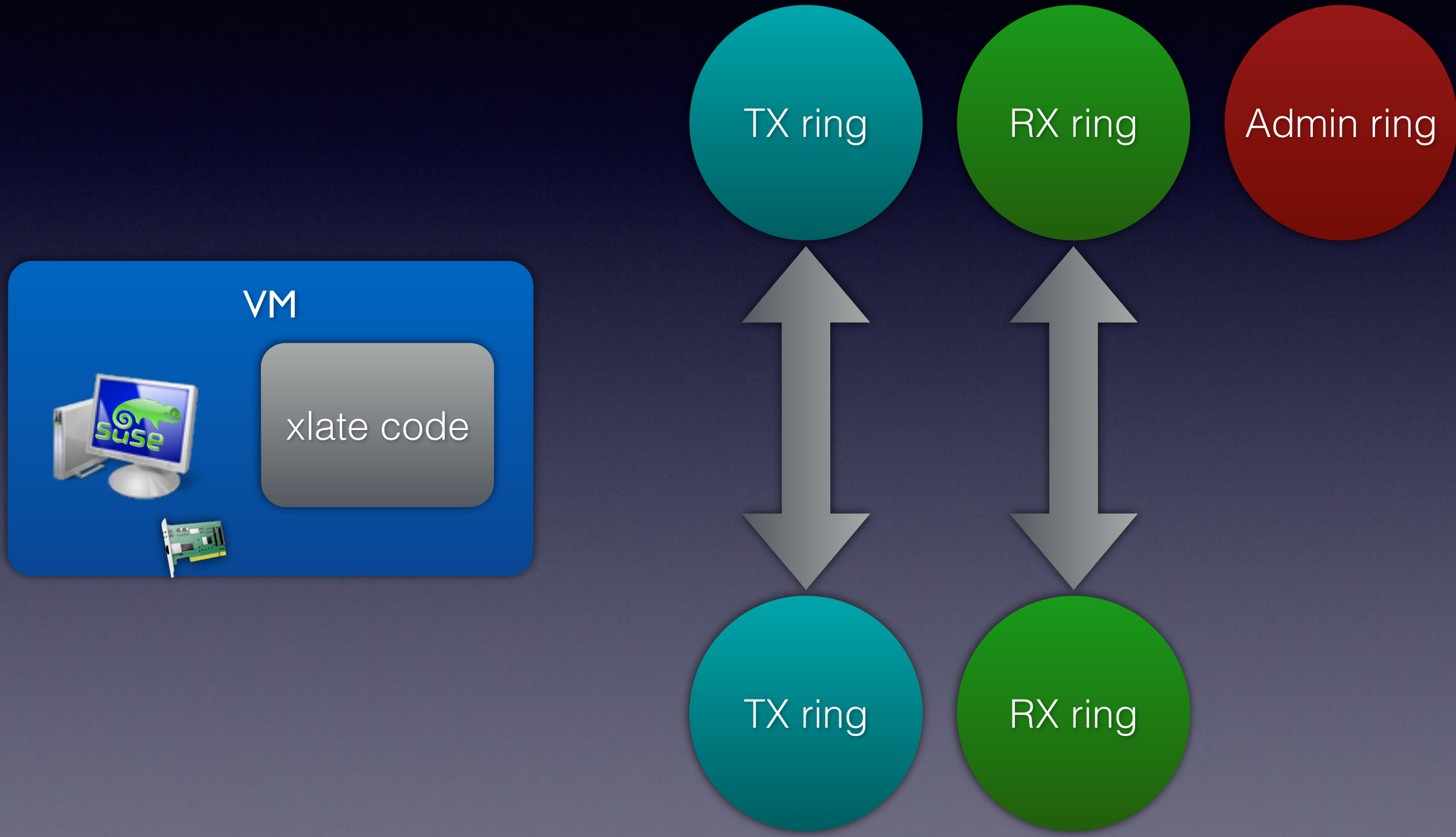
Guest injected driver

Guest injected driver

Guest injected driver



Guest injected driver

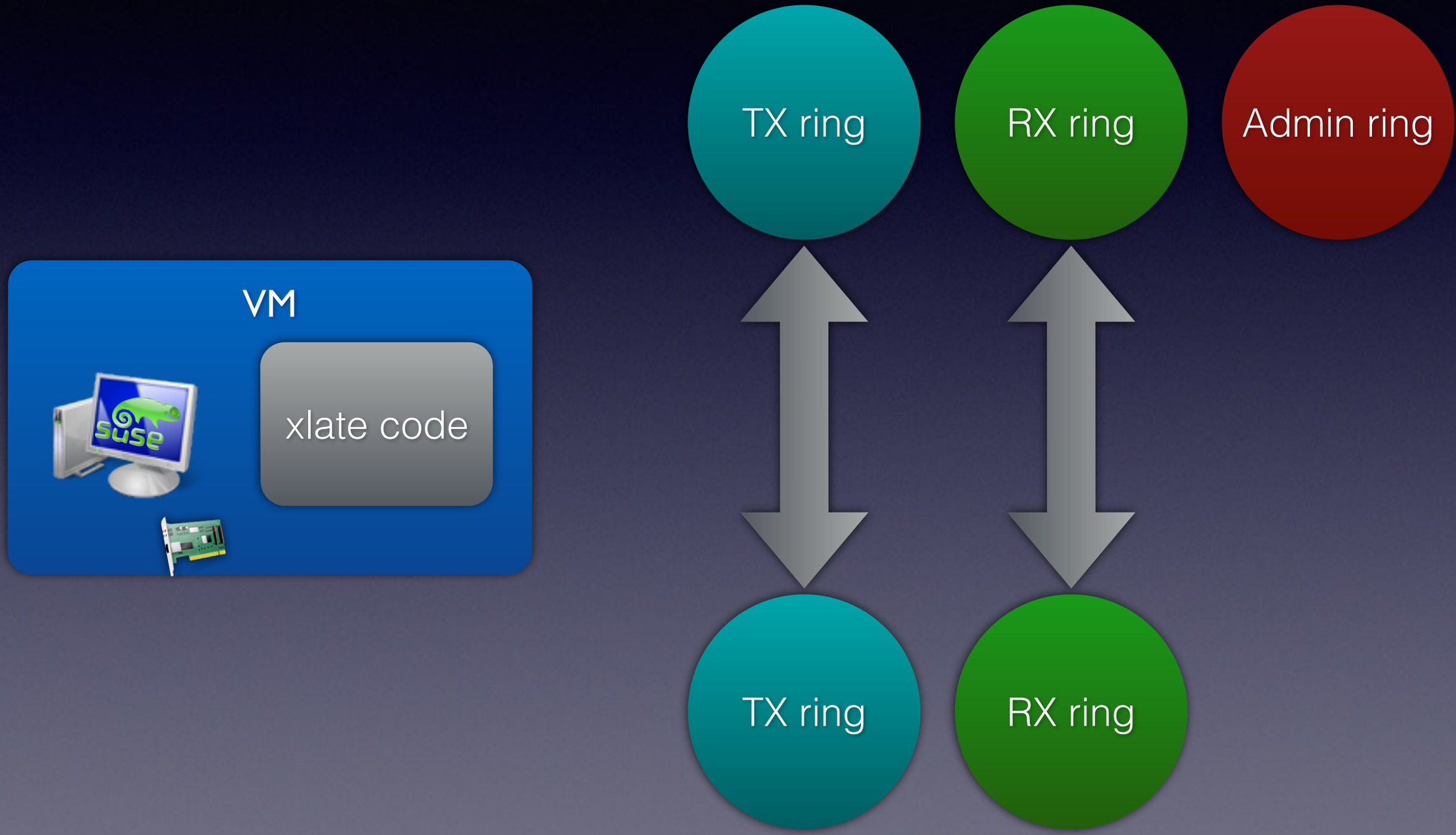


Guest injected driver

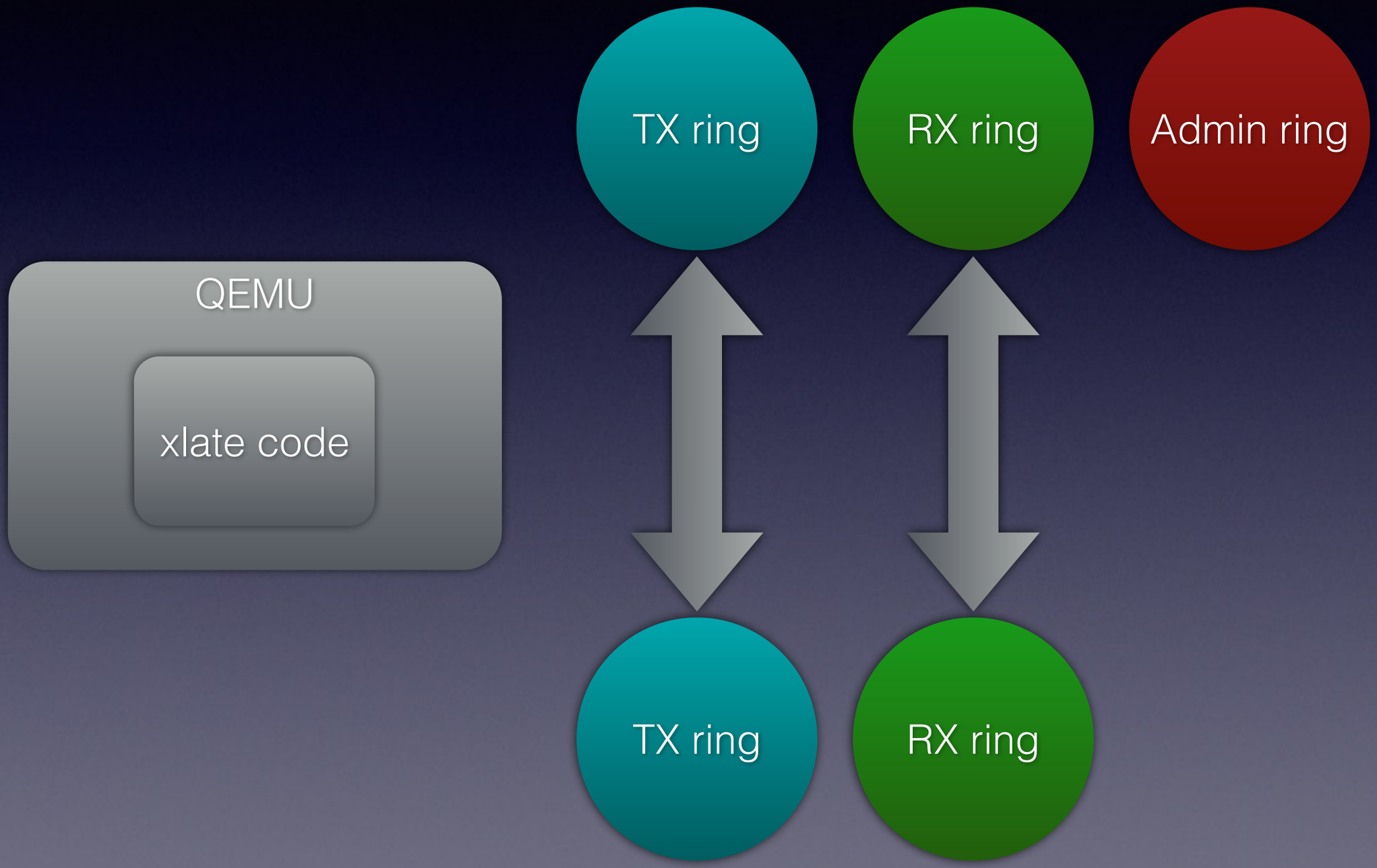
- Guest sees extended virtio device
- Code to convert hw queues to virtio queues comes from hypervisor
- Benefits:
  - Less code in privileged context
  - Faster IRQ path

QEMU based virtio queue conversion

QEMU based virtio queue conversion



QEMU based virtio queue conversion



QEMU based virtio queue conversion

- QEMU allocates real hardware
- QEMU has specific HW/virtio conversion driver
- Guest only sees virtio
- Probably bad for latency

Conclusions

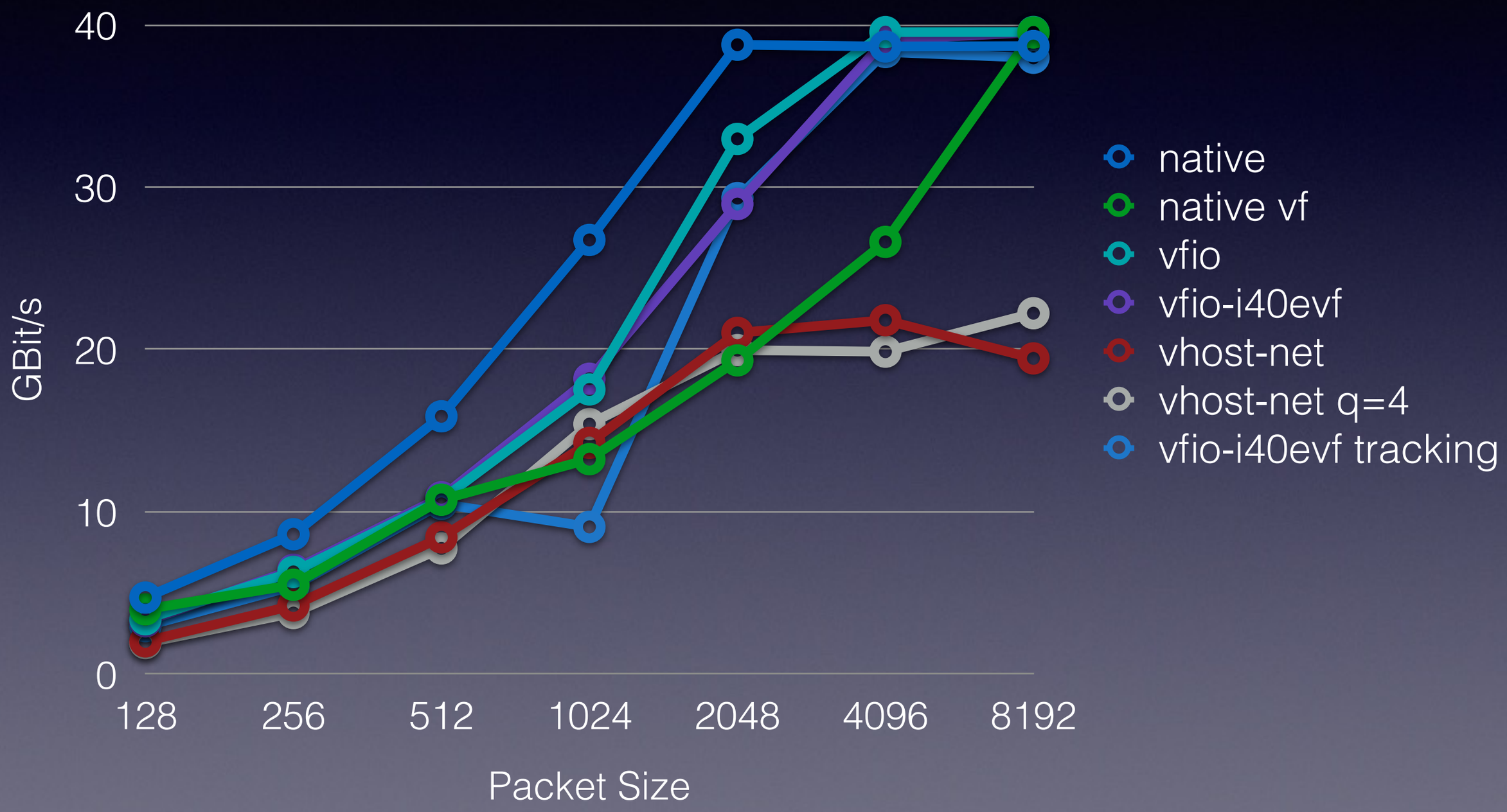
- Virtio is fast enough for most cases
- Room for improvement exists
- Live migration of real hardware works

Thank You

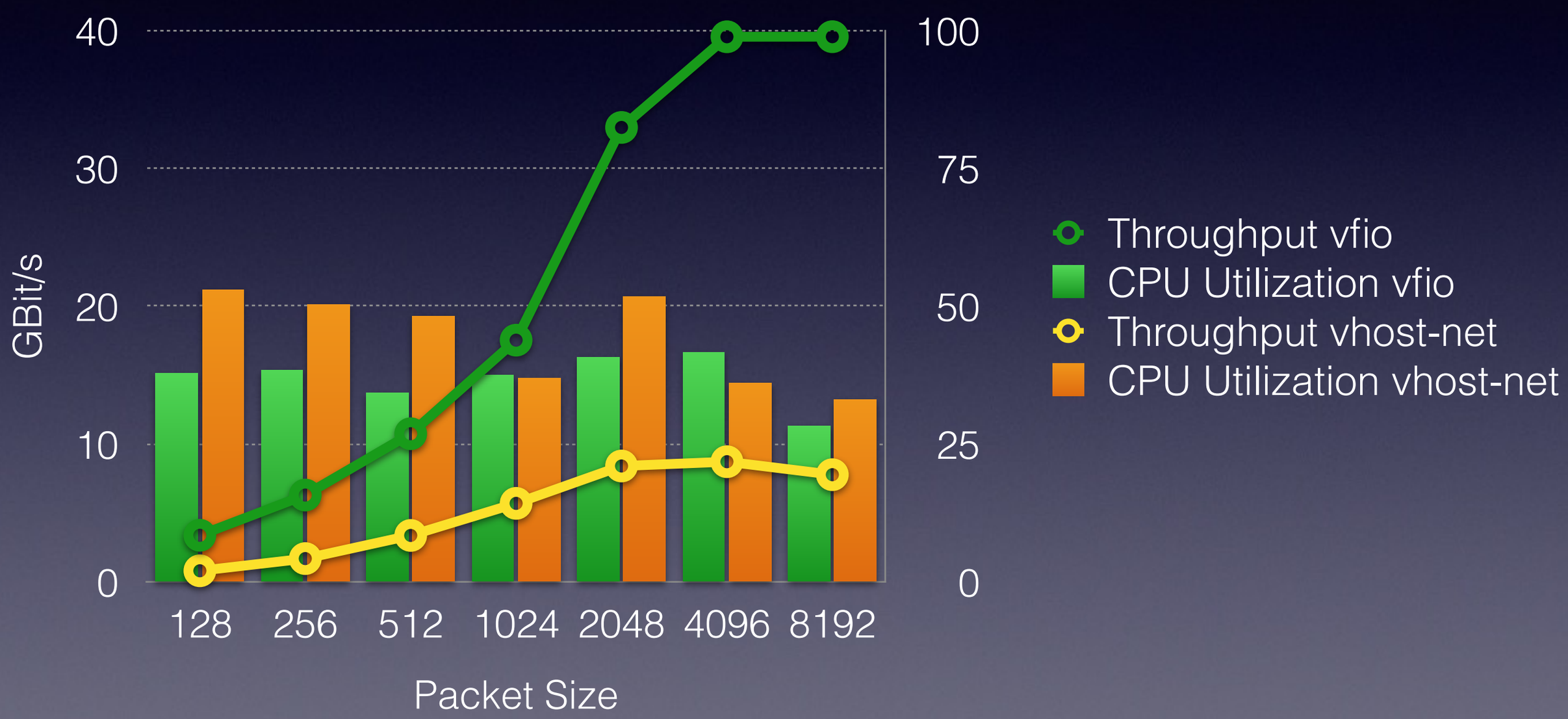
git://github.com/agraf/qemu.git vfio-i40vf

Performance Backup

Throughput


40
30
20
10
0
GBit/s
128
256
512
1024
2048
4096
8192
Packet Size
native
native vf
vfio
vfio-i40evf
vhost-net
vhost-net q=4
vfio-i40evf tracking

Throughput vs CPU Utilization



Throughput vs Latency