OvS
Open vSwitch

Ciara Loftus

Intel Corporation

DPDK vHost User Improvements

Agenda

- DPDK vHost User Introduction/Refresh
- Time Line of DPDK vHost User in OVS
- Recent Improvements
  - NUMA Awareness
  - Client Mode & Reconnect
- Future Improvements
  - vHost User PMD
  - Zero Copy

DPDK vHost User Refresh

What is DPDK vHost User?

What is DPDK?

What is DPDK?

- Data Plane Development Kit

DPDK

DATA PLANE DEVELOPMENT KIT

What is DPDK?

- Data Plane Development Kit

DPDK
DATA PLANE DEVELOPMENT KIT

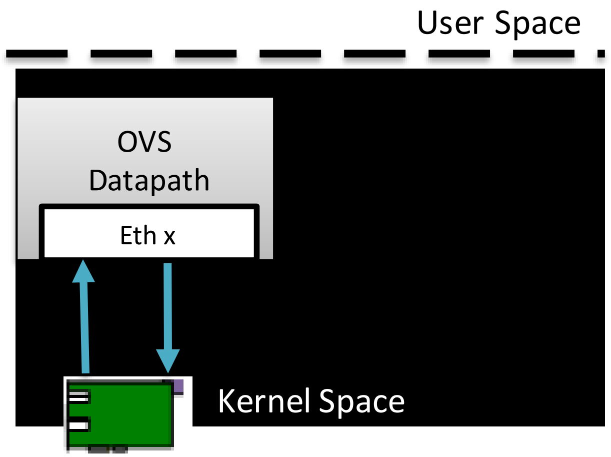
- Userspace drivers & libraries for accelerated network I/O

What is DPDK?

- Data Plane Development Kit
- Userspace drivers & libraries for accelerated network I/O
- Integrated into OVS in v2.2

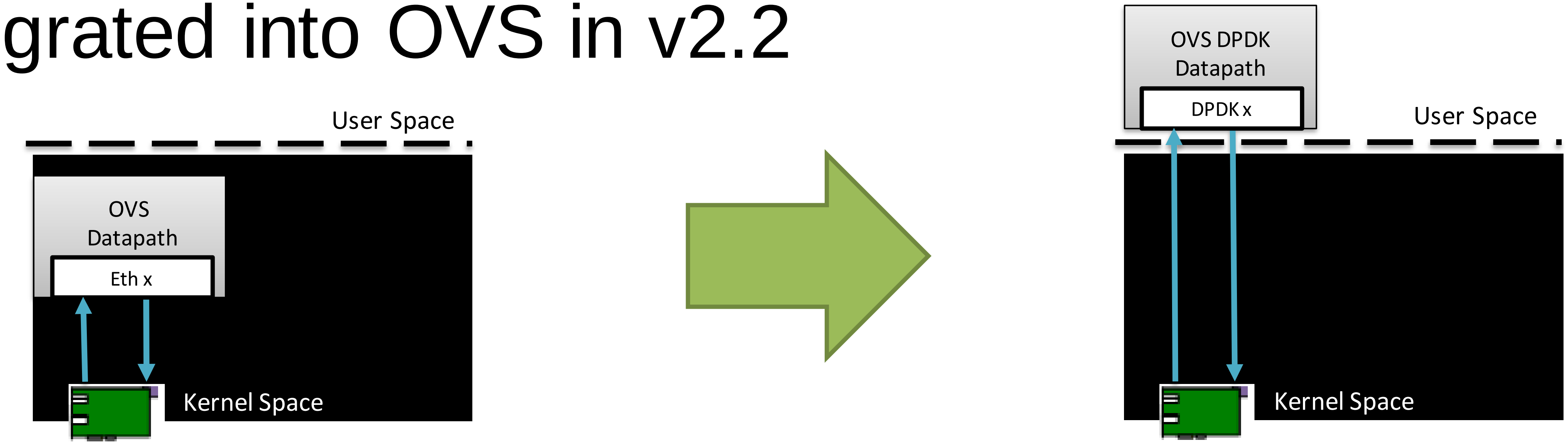
What is DPDK?

- Data Plane Development Kit
- Userspace drivers & libraries for accelerated network I/O
- Integrated into OVS in v2.2



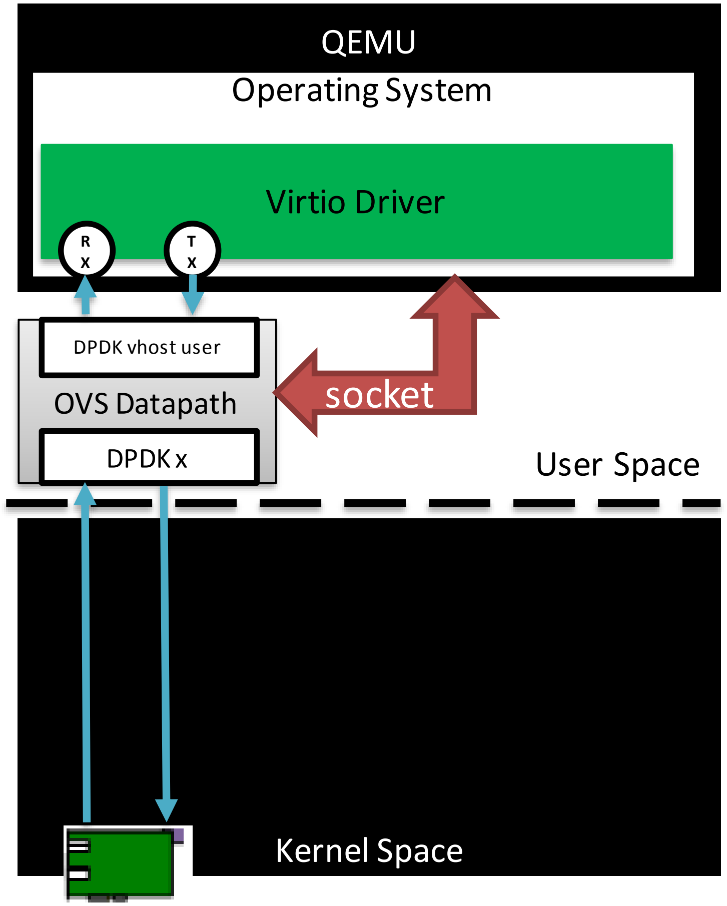
What is DPDK?

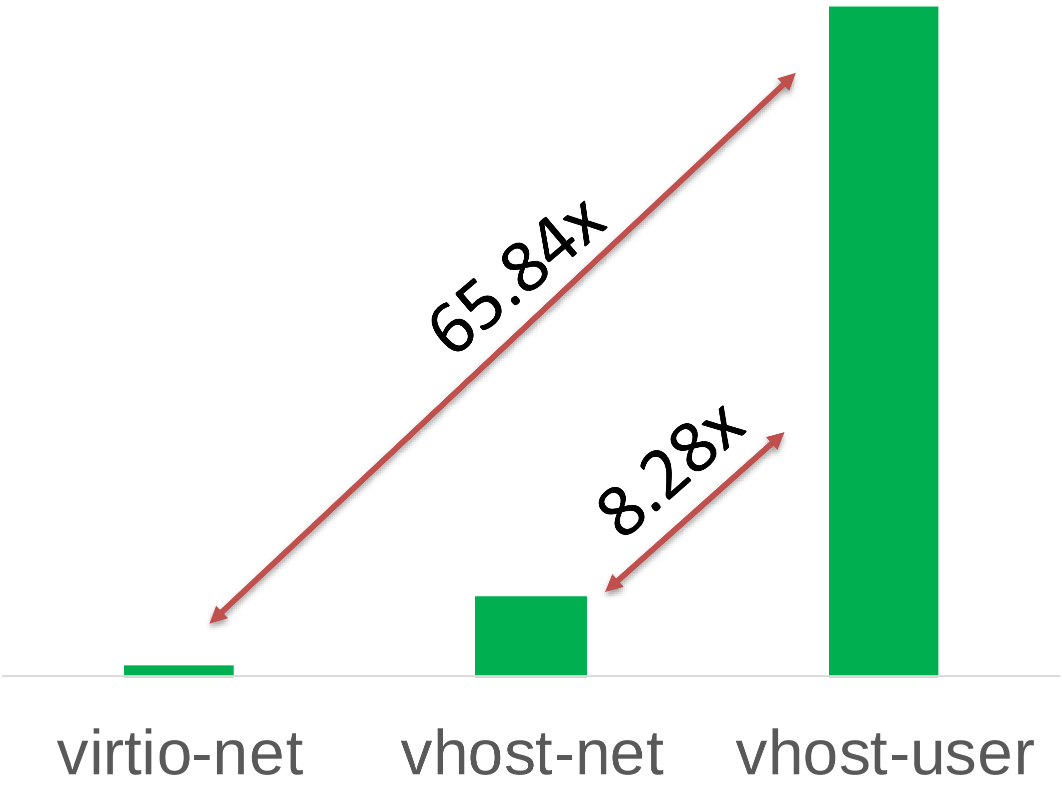
- Data Plane Development Kit
- Userspace drivers & libraries for accelerated network I/O
- Integrated into OVS in v2.2

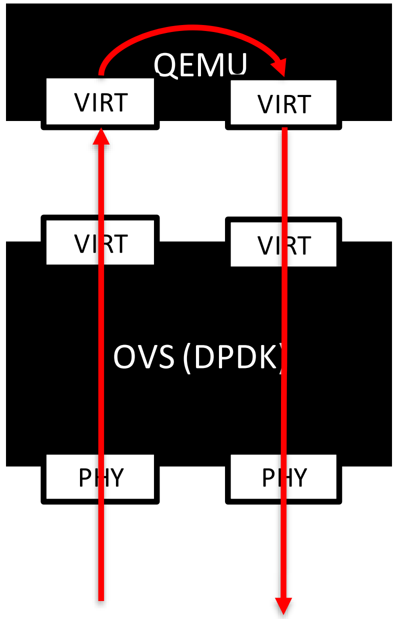


DPDK vHost User Refresh

Accelerated guest access method offered by DPDK capable of outperforming traditional methods by >8x*



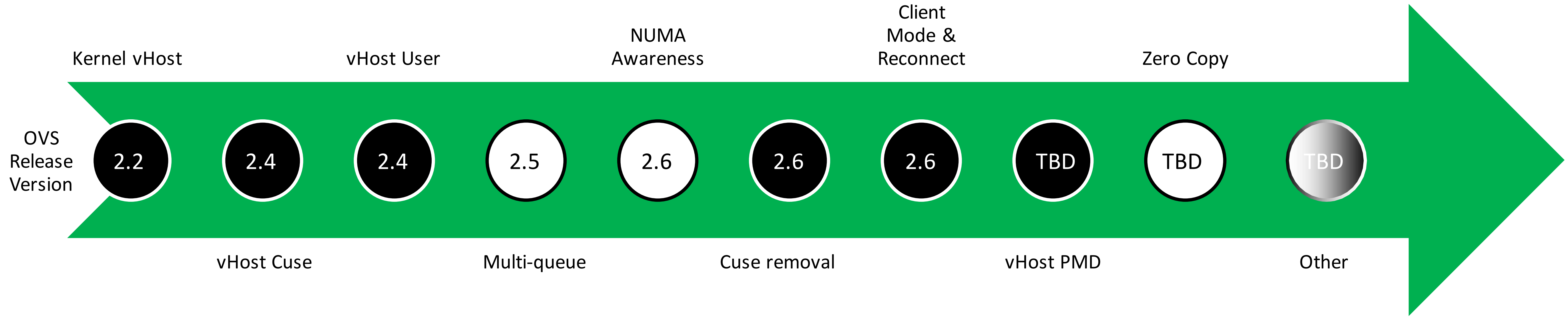




Single core, unidirectional loopback test with 64B packets

* Platform Configuration and Test Result in Backup

Timeline of vHost User in OVS



OVS Release Version

Kernel vHost 2.2

2.4 vHost Cuse

vHost User 2.4

2.5 Multi-queue

NUMA Awareness 2.6

2.6 Cuse removal

Client Mode & Reconnect 2.6

TBD vHost PMD

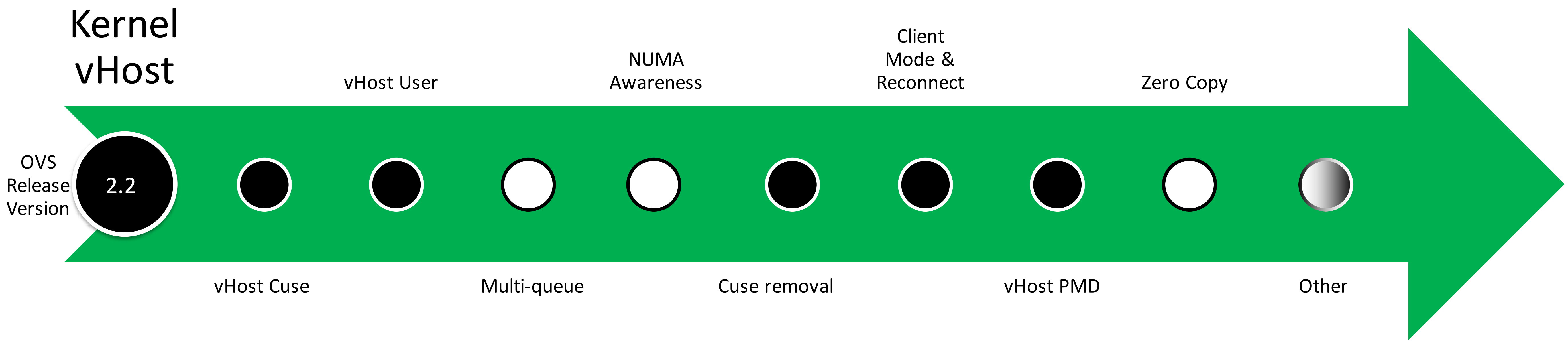
Zero Copy TBD

TBD Other

● = Functional Improvement

○ = Performance Improvement

Timeline of vHost User in OVS



Kernel vHost

OVS Release Version

2.2

vHost Cuse

vHost User

Multi-queue

NUMA Awareness

Cuse removal

Client Mode & Reconnect

vHost PMD

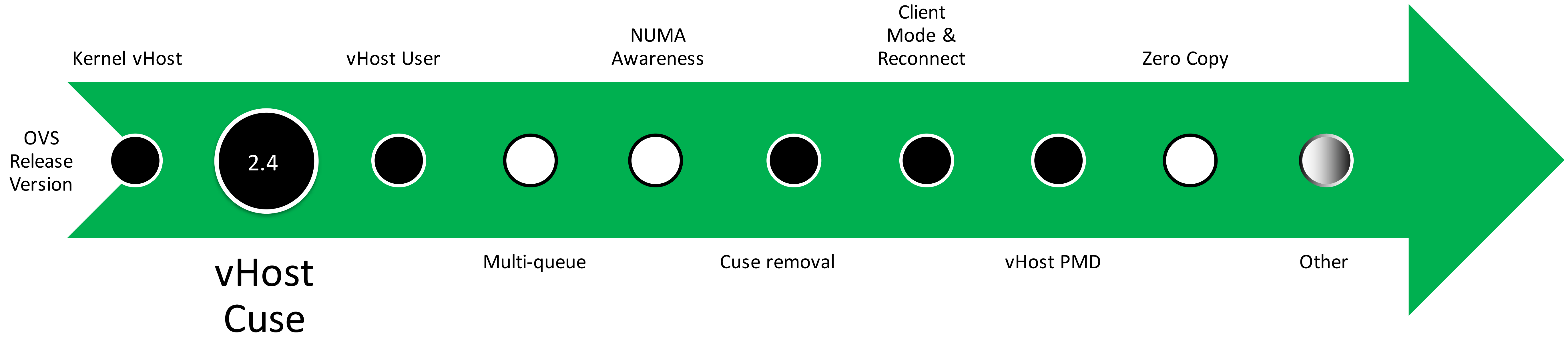
Zero Copy

Other

● = Functional Improvement

○ = Performance Improvement

Timeline of vHost User in OVS



OVS Release Version

Kernel vHost

2.4

vHost Cuse

vHost User

Multi-queue

NUMA Awareness

Cuse removal

Client Mode & Reconnect

vHost PMD

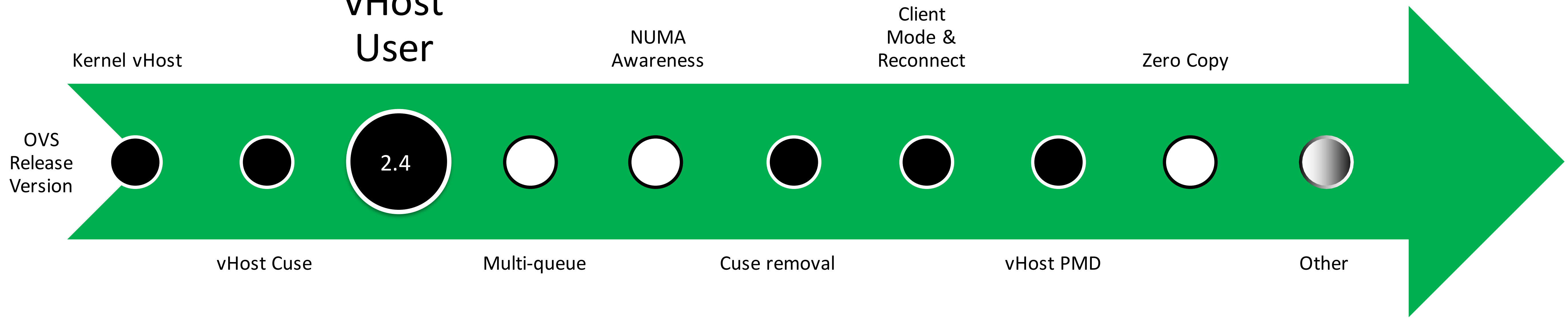
Zero Copy

Other

● = Functional Improvement

○ = Performance Improvement

Timeline of vHost User in OVS



OVS Release Version

Kernel vHost

vHost Cuse

vHost User

2.4

Multi-queue

NUMA Awareness

Cuse removal

Client Mode & Reconnect

vHost PMD

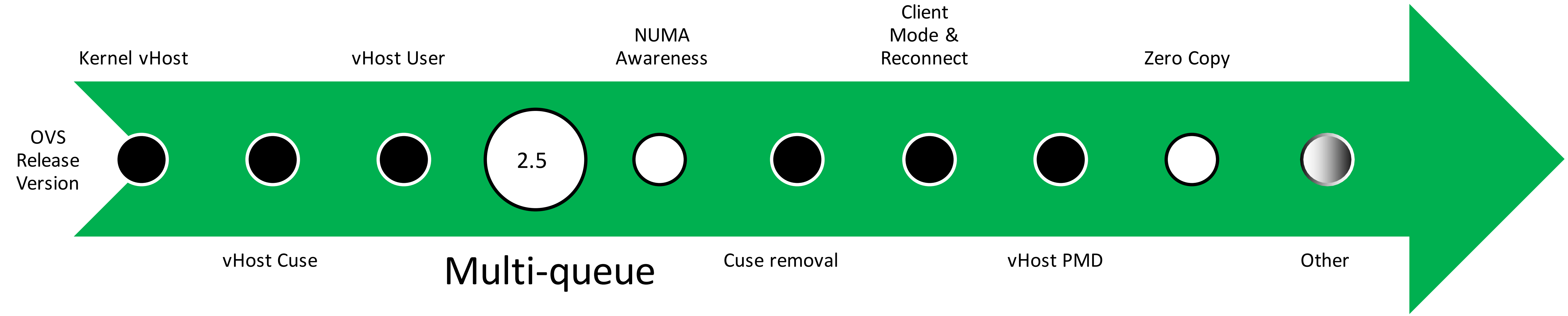
Zero Copy

Other

● = Functional Improvement

○ = Performance Improvement

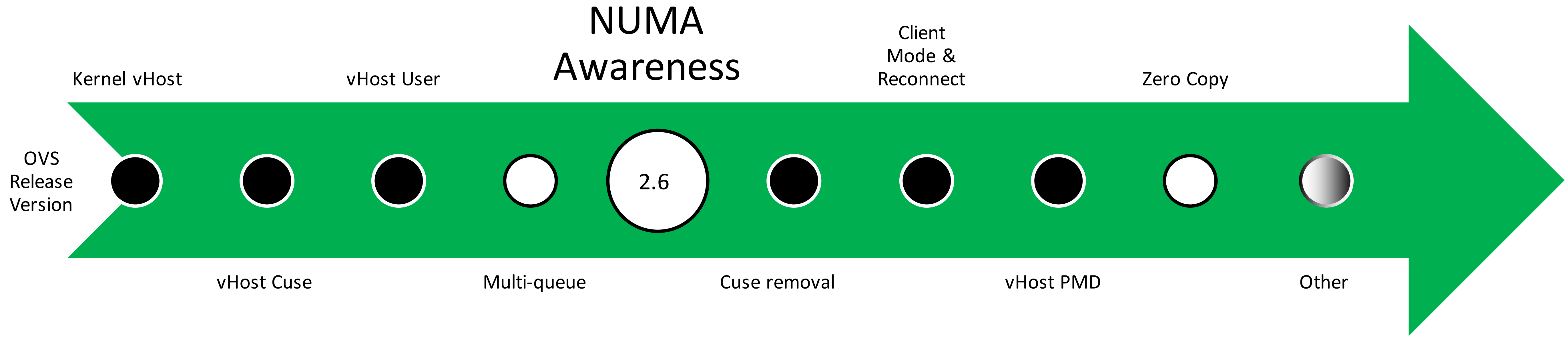
Timeline of vHost User in OVS



● = Functional Improvement

○ = Performance Improvement

Timeline of vHost User in OVS



Kernel vHost

vHost User

NUMA Awareness

Client Mode & Reconnect

Zero Copy

OVS Release Version

2.6

vHost Cuse

Multi-queue

Cuse removal

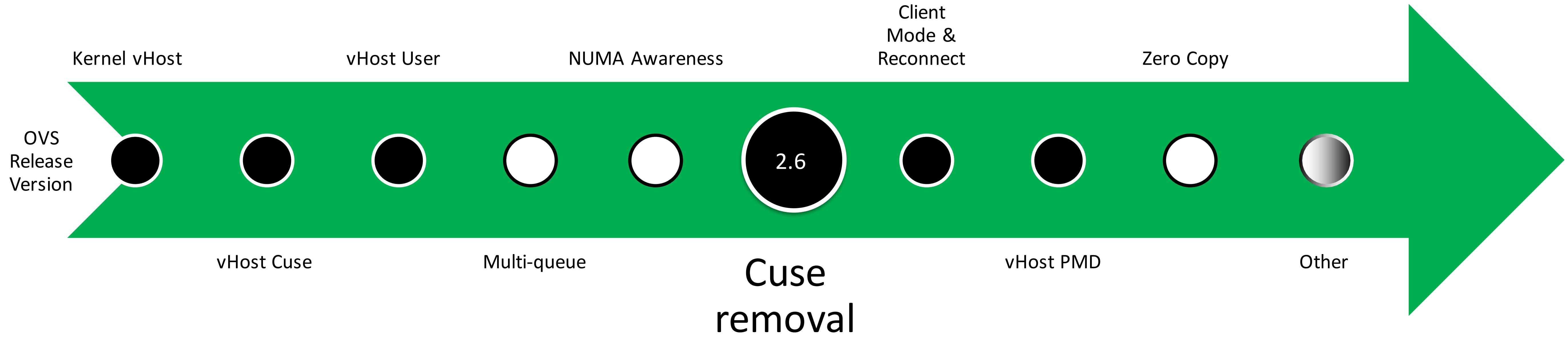
vHost PMD

Other

● = Functional Improvement

○ = Performance Improvement

Timeline of vHost User in OVS



OVS Release Version

Kernel vHost

vHost Cuse

vHost User

Multi-queue

NUMA Awareness

2.6

Cuse removal

Client Mode & Reconnect

vHost PMD

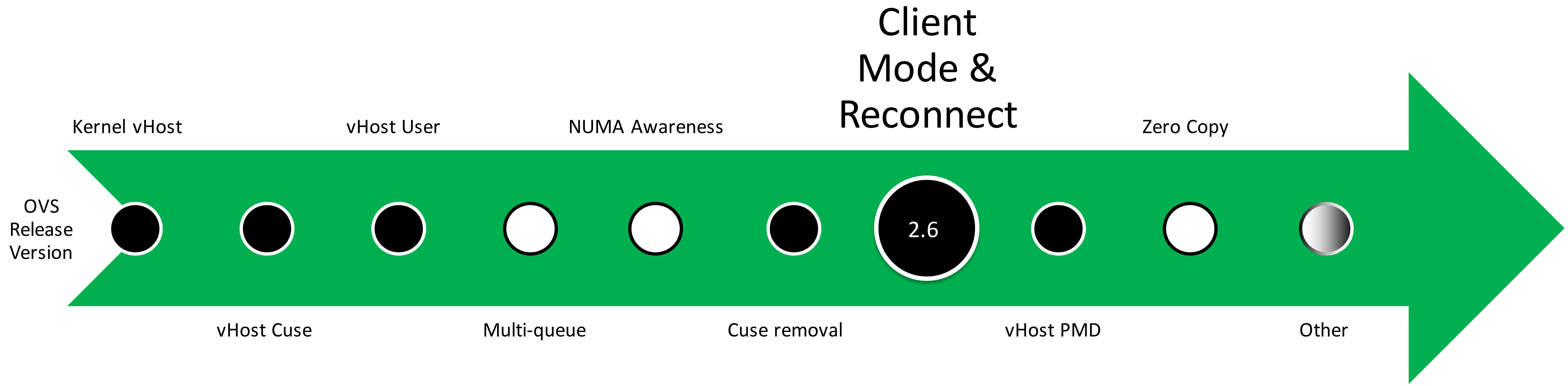
Zero Copy

Other

● = Functional Improvement

○ = Performance Improvement

Timeline of vHost User in OVS



OVS Release Version

Kernel vHost

vHost Cuse

vHost User

Multi-queue

NUMA Awareness

Cuse removal

Client Mode & Reconnect

2.6

vHost PMD

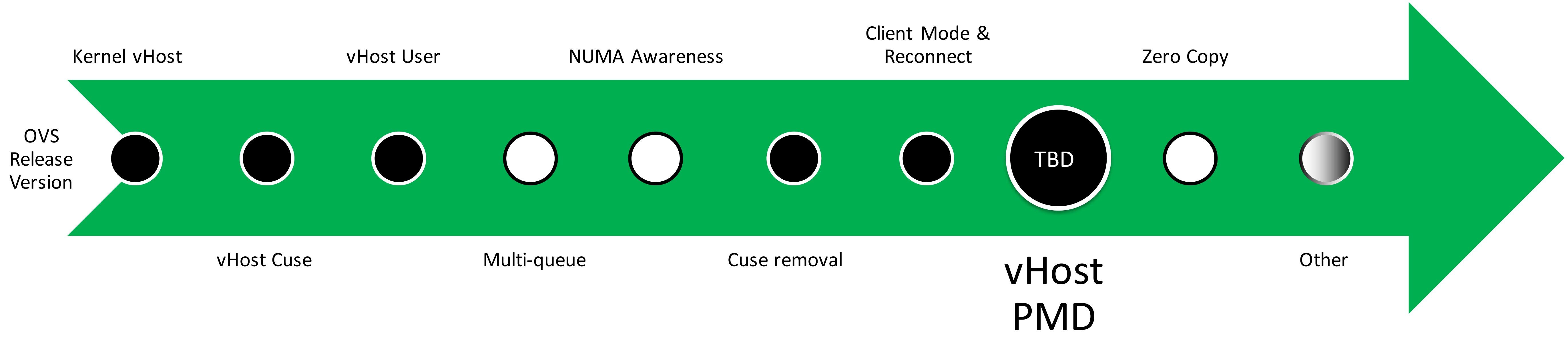
Zero Copy

Other

● = Functional Improvement

○ = Performance Improvement

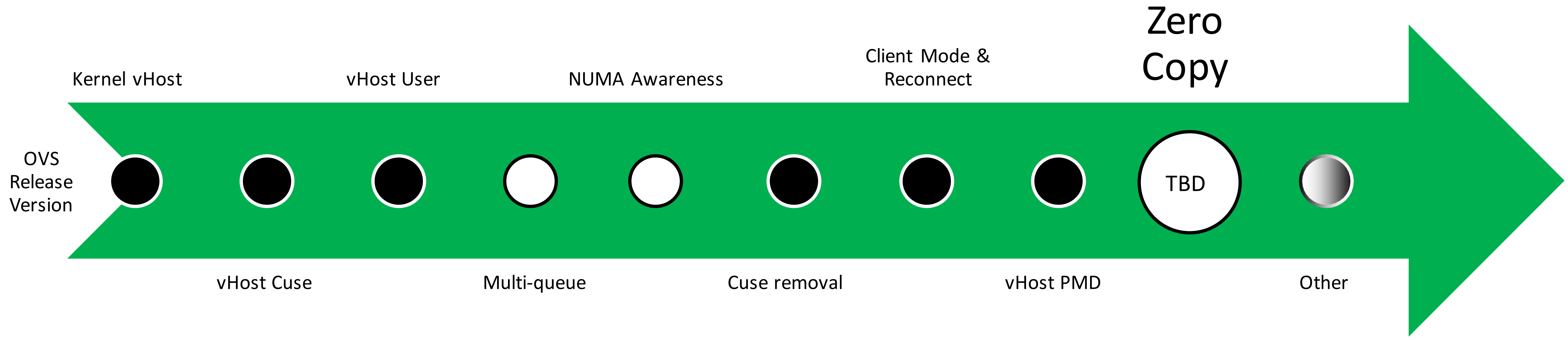
Timeline of vHost User in OVS



● = Functional Improvement

○ = Performance Improvement

Timeline of vHost User in OVS



OVS Release Version

Kernel vHost

vHost Cuse

vHost User

Multi-queue

NUMA Awareness

Cuse removal

Client Mode & Reconnect

vHost PMD

Zero Copy

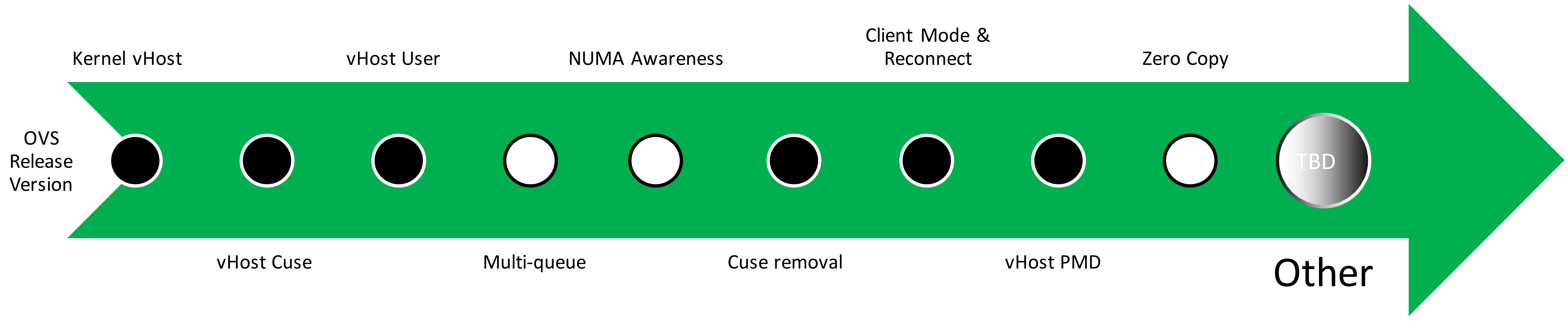
TBD

Other

● = Functional Improvement

○ = Performance Improvement

Timeline of vHost User in OVS



OVS Release Version

Kernel vHost

vHost Cuse

vHost User

Multi-queue

NUMA Awareness

Cuse removal

Client Mode & Reconnect

vHost PMD

Zero Copy

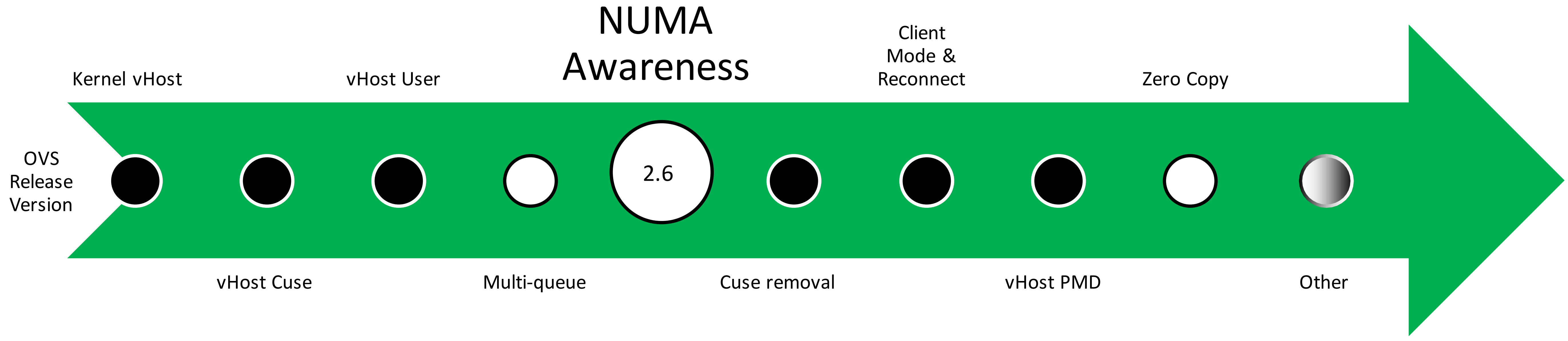
TBD

Other

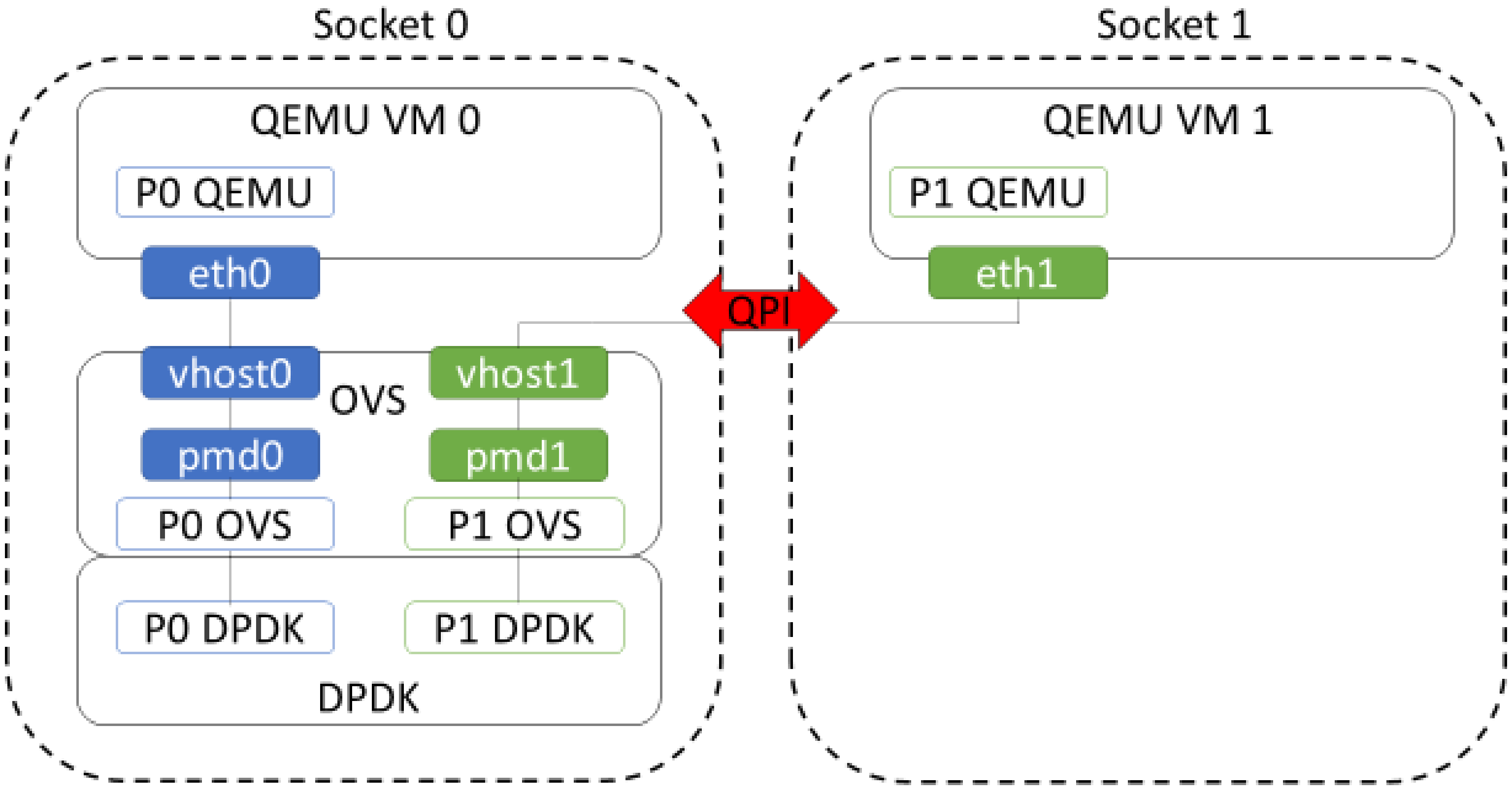
● = Functional Improvement

○ = Performance Improvement

NUMA Awareness

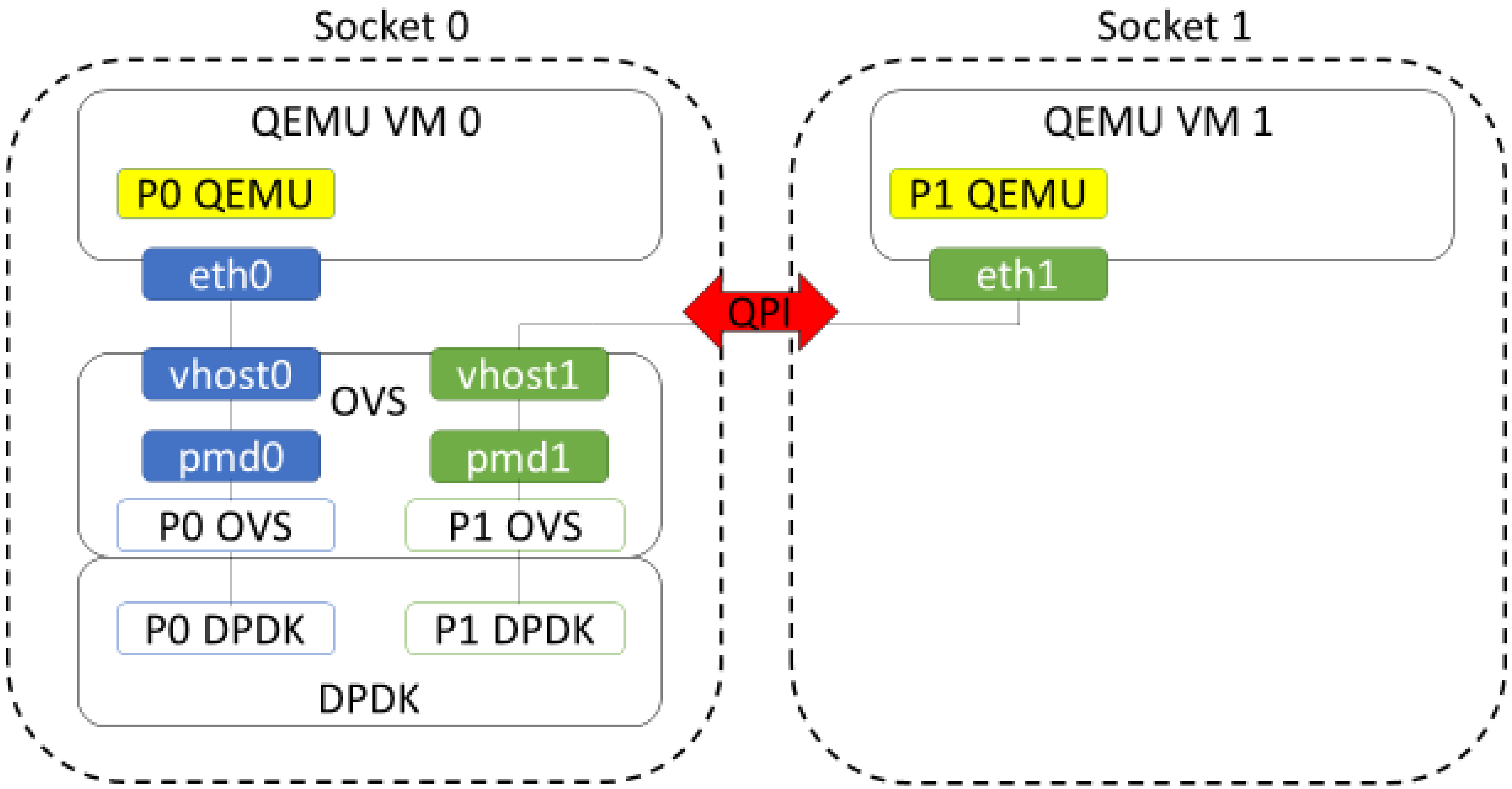


NUMA Awareness



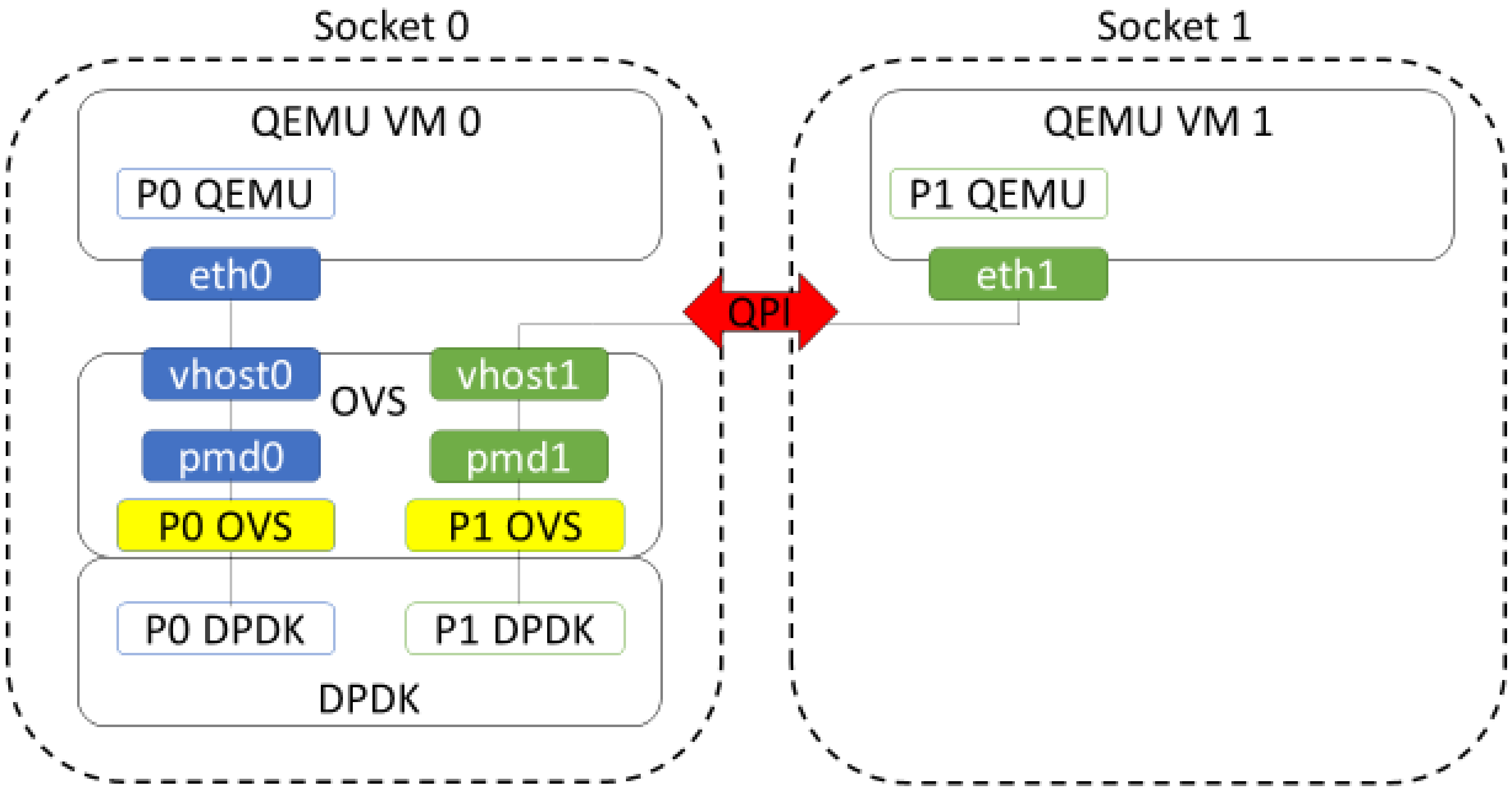
QEMU, DPDK & OVS vHost memory need to be co-located for optimal performance.

NUMA Awareness



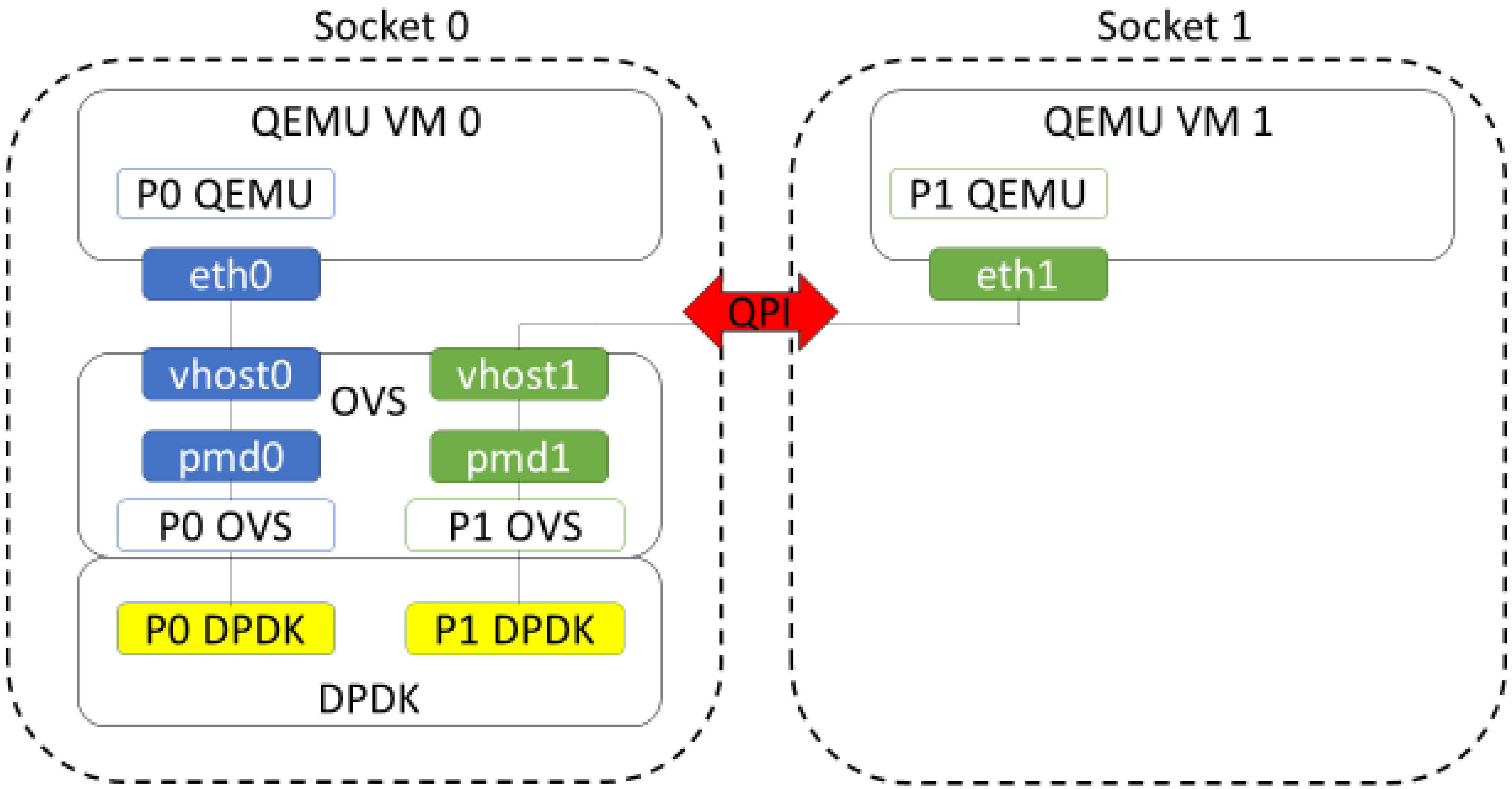
QEMU, DPDK & OVS vHost memory need to be co-located for optimal performance.

NUMA Awareness



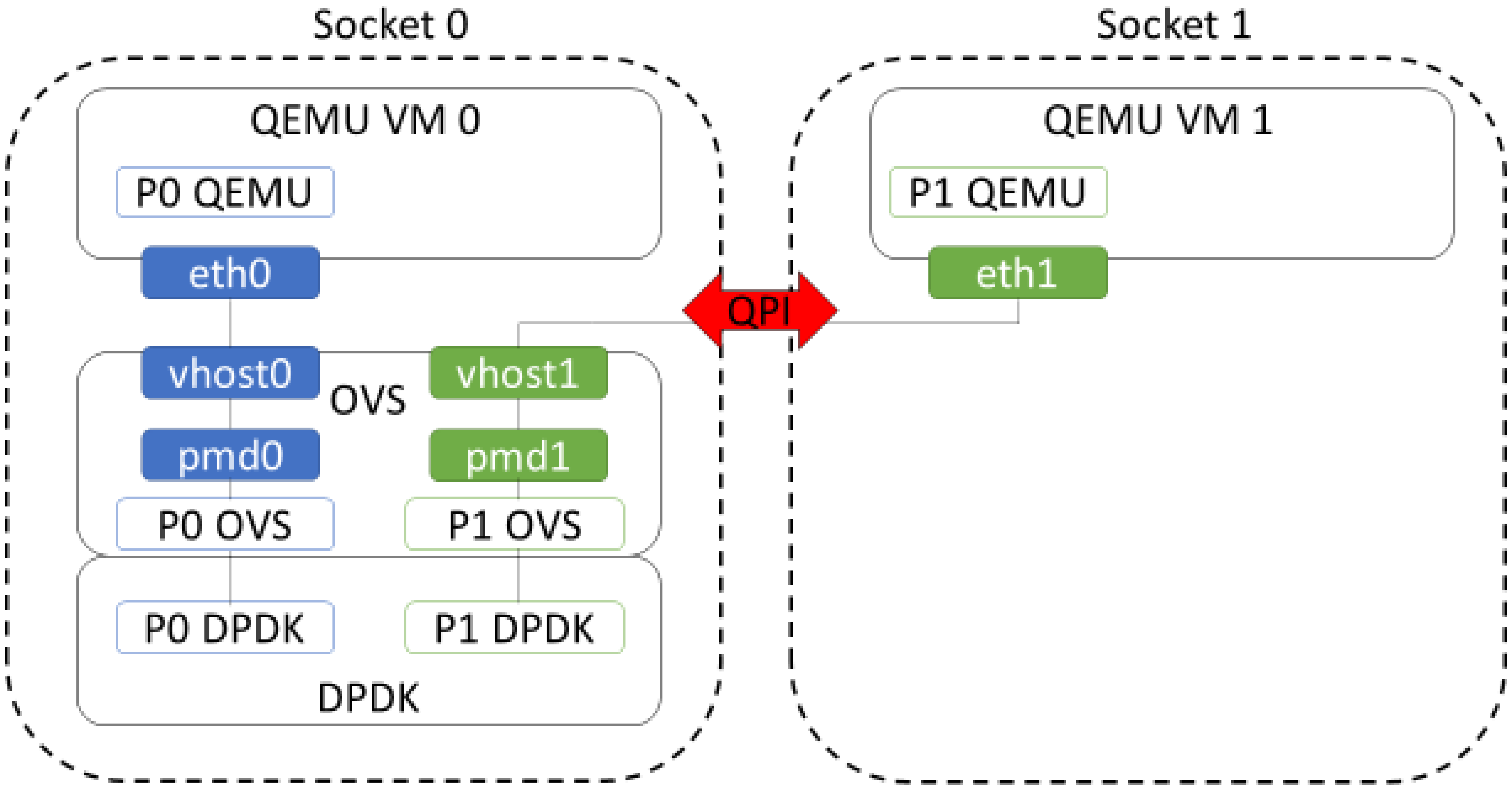
QEMU, DPDK & OVS vHost memory need to be co-located for optimal performance.

NUMA Awareness



QEMU, DPDK & OVS vHost memory need to be co-located for optimal performance.

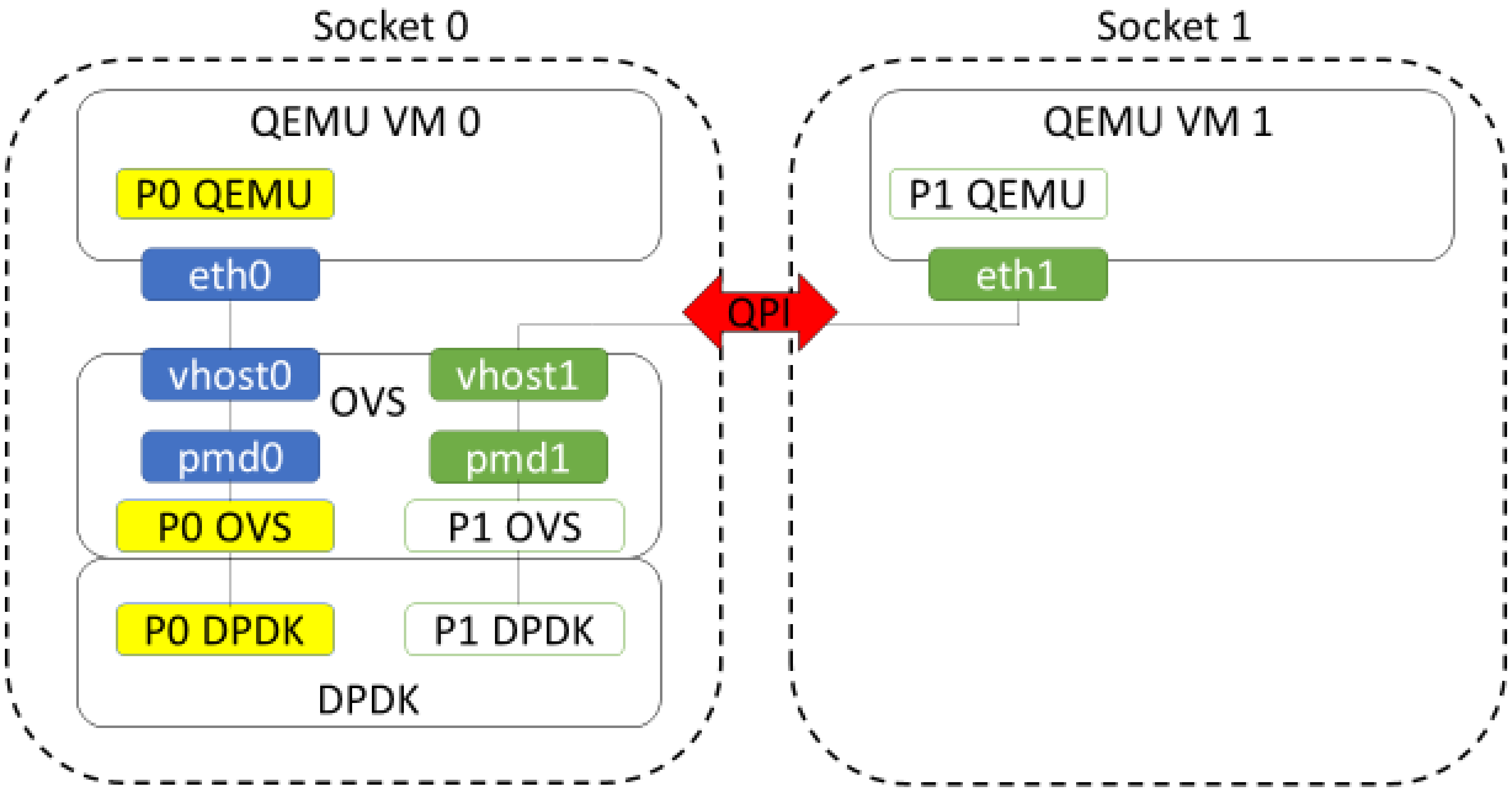
NUMA Awareness



Previous limitation:

All DPDK vHost memory must come from the same NUMA node.

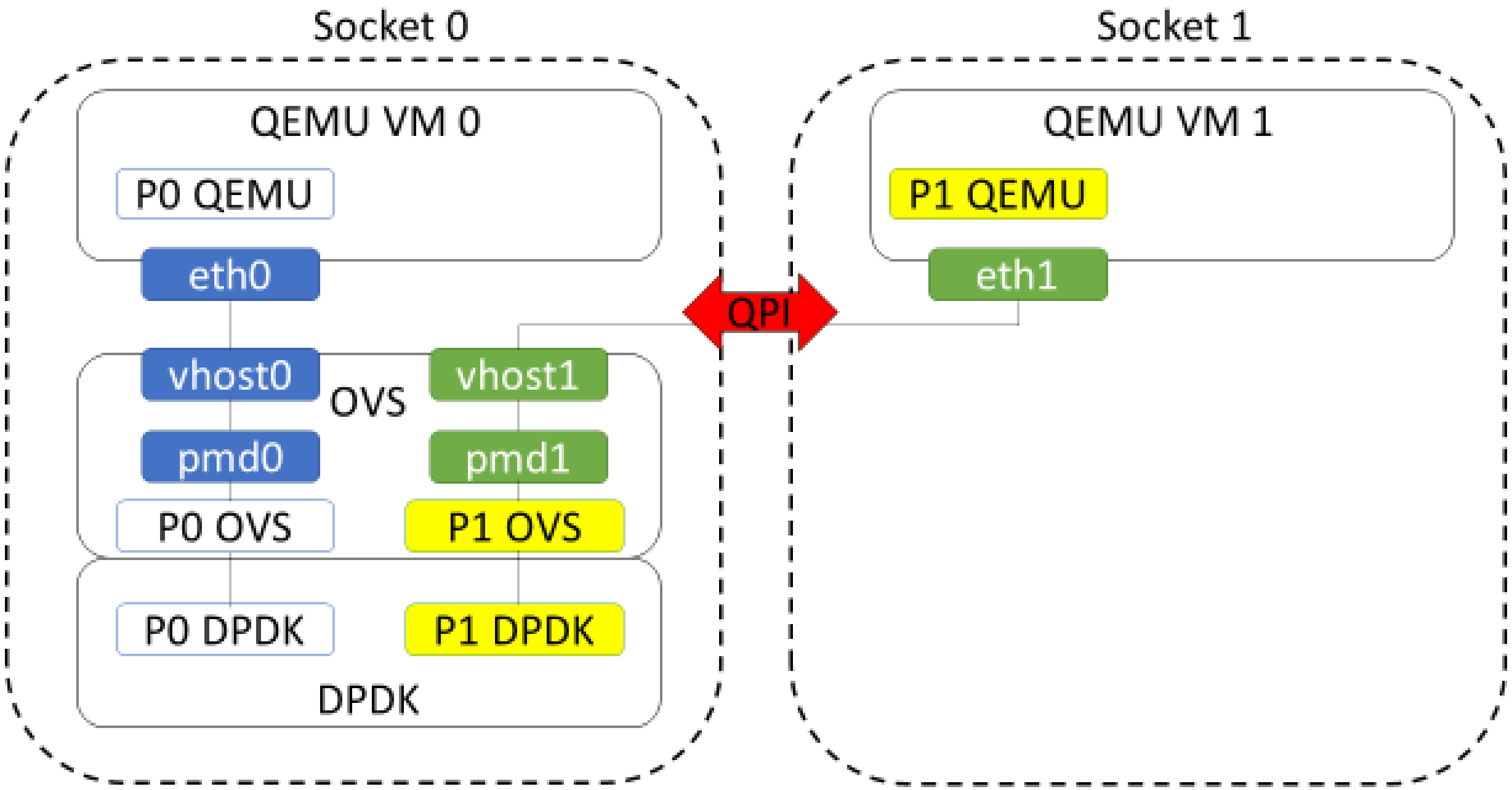
NUMA Awareness



Previous limitation:

All DPDK vHost memory must come from the same NUMA node.

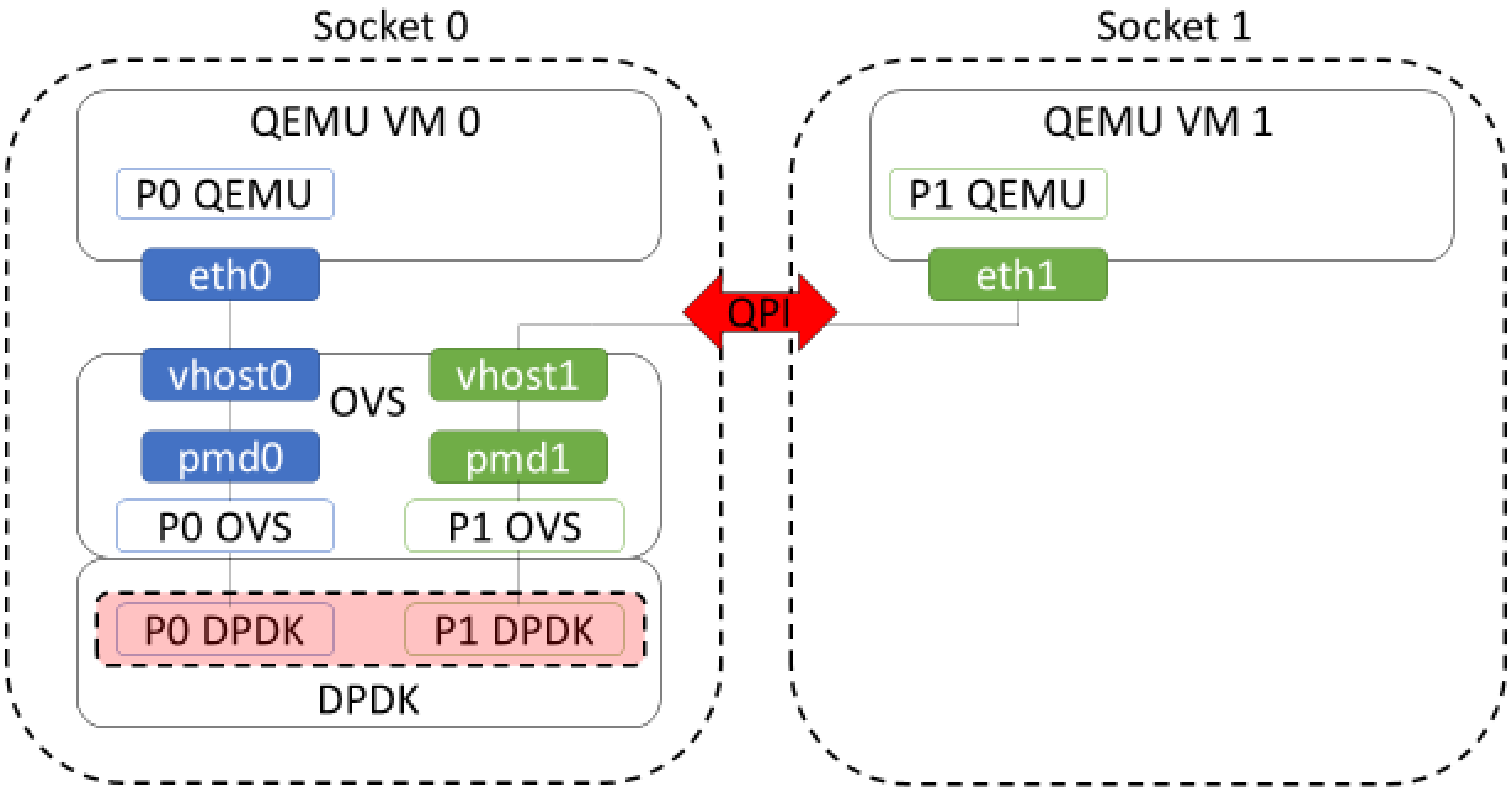
NUMA Awareness



Previous limitation:

All DPDK vHost memory must come from the same NUMA node.

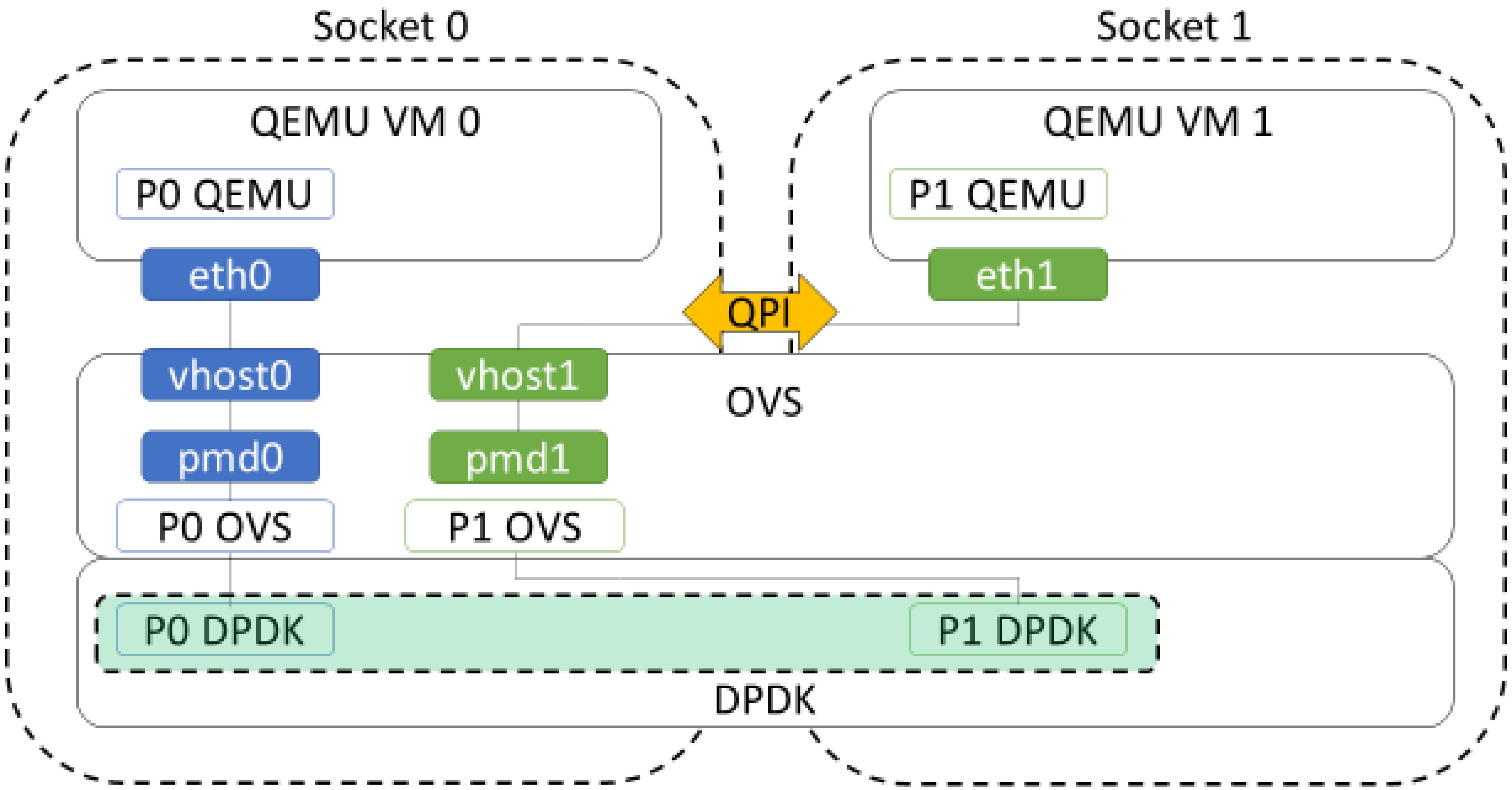
NUMA Awareness



Previous limitation:

All DPDK vHost memory must come from the same NUMA node.

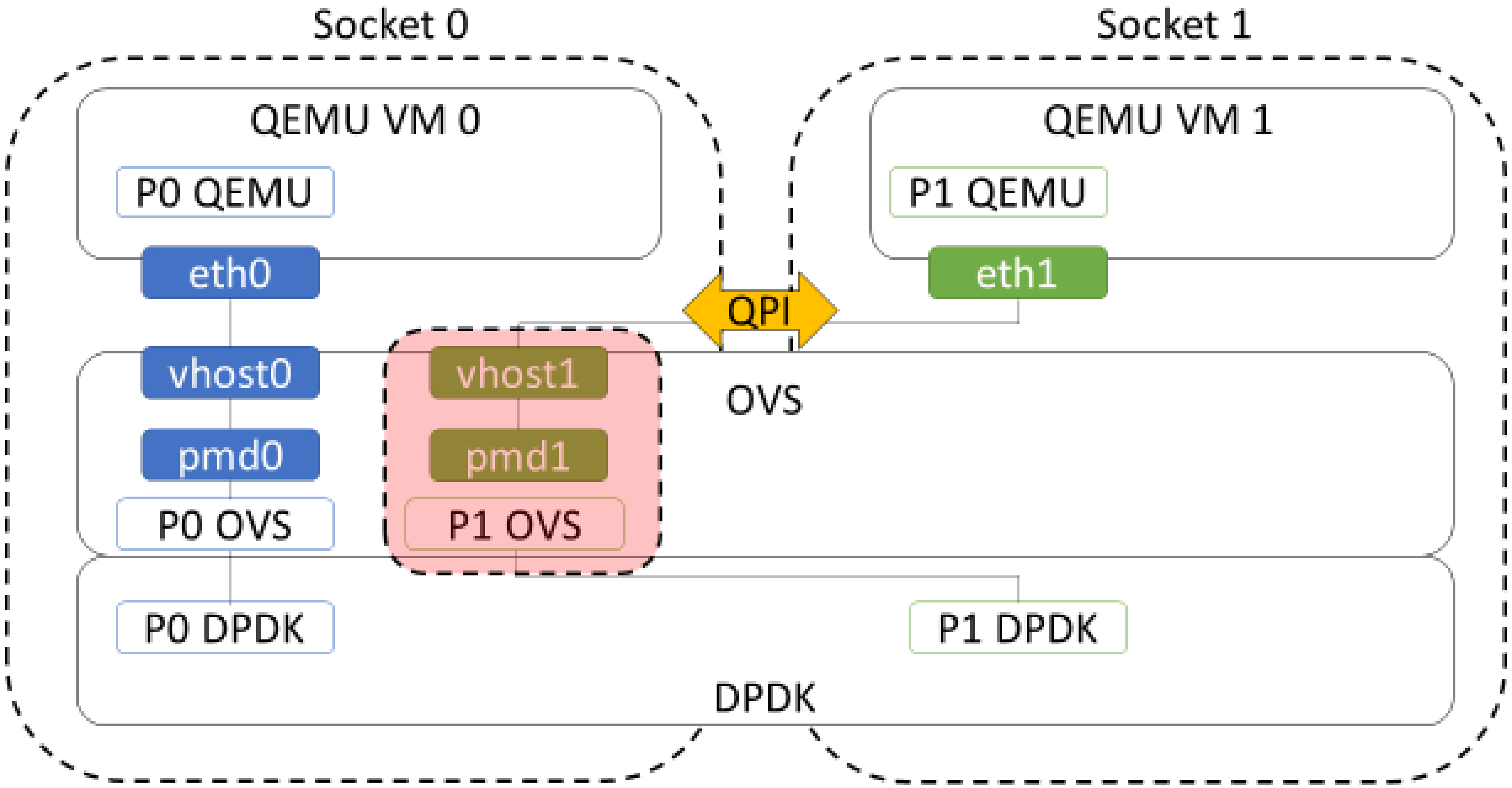
NUMA Awareness



Solution:

DPDK vHost memory relocated to correct NUMA node on VM boot.

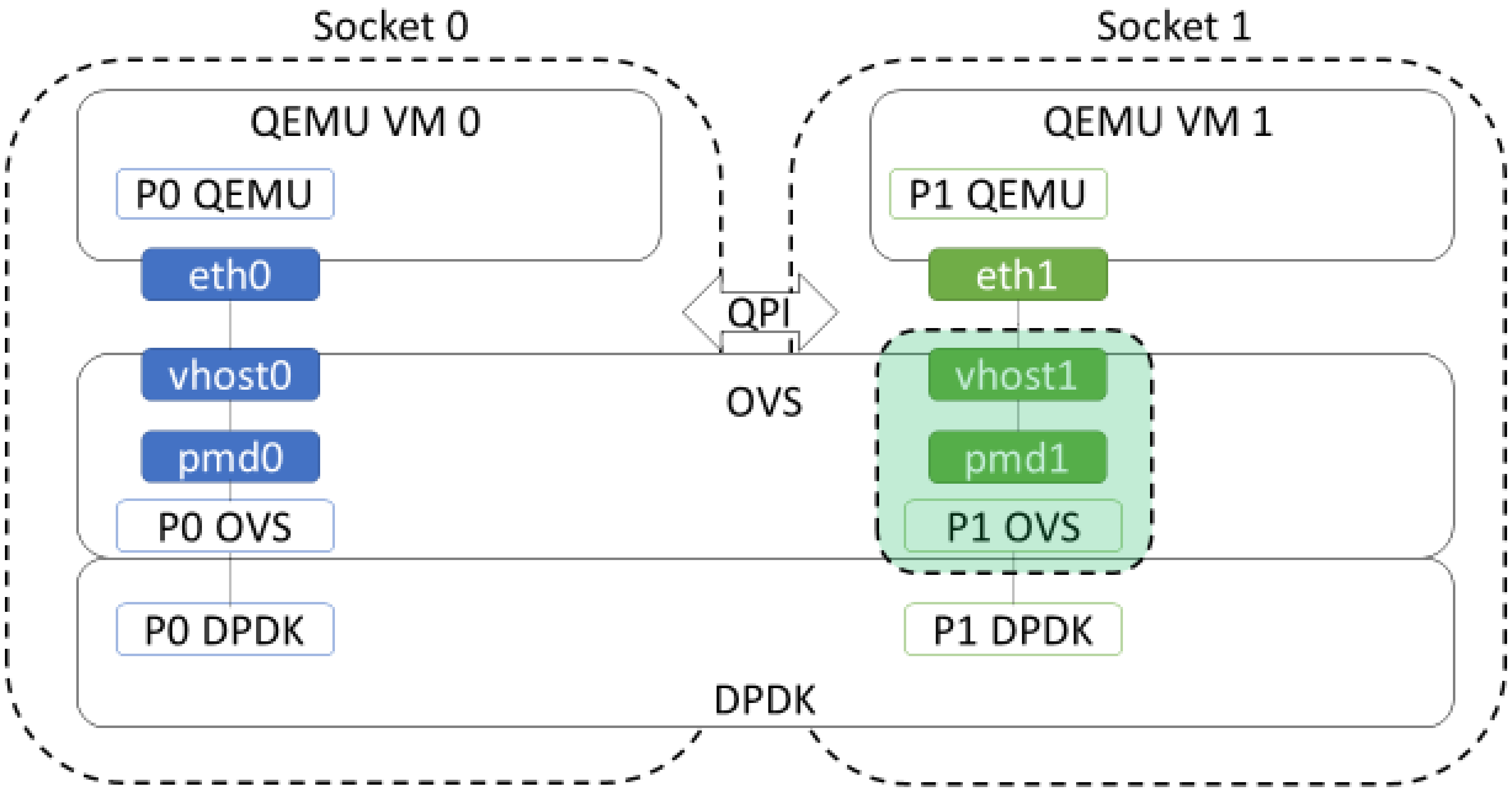
NUMA Awareness



Previous limitation:

All PMDs servicing vHost ports must come from the same NUMA node.

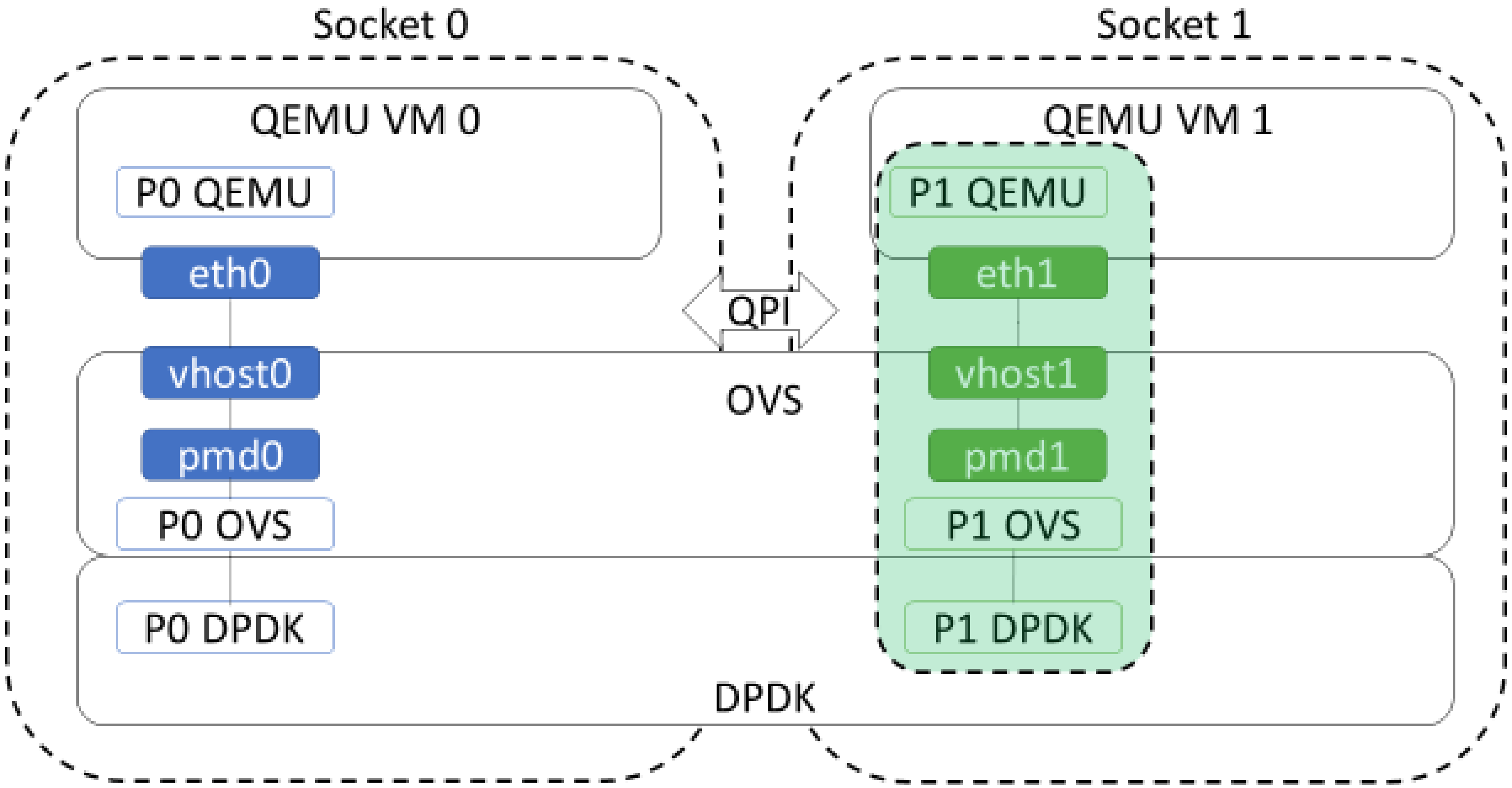
NUMA Awareness



Solution:

mbufs and servicing PMD in OVS are moved to correct NUMA during DPDK callback.

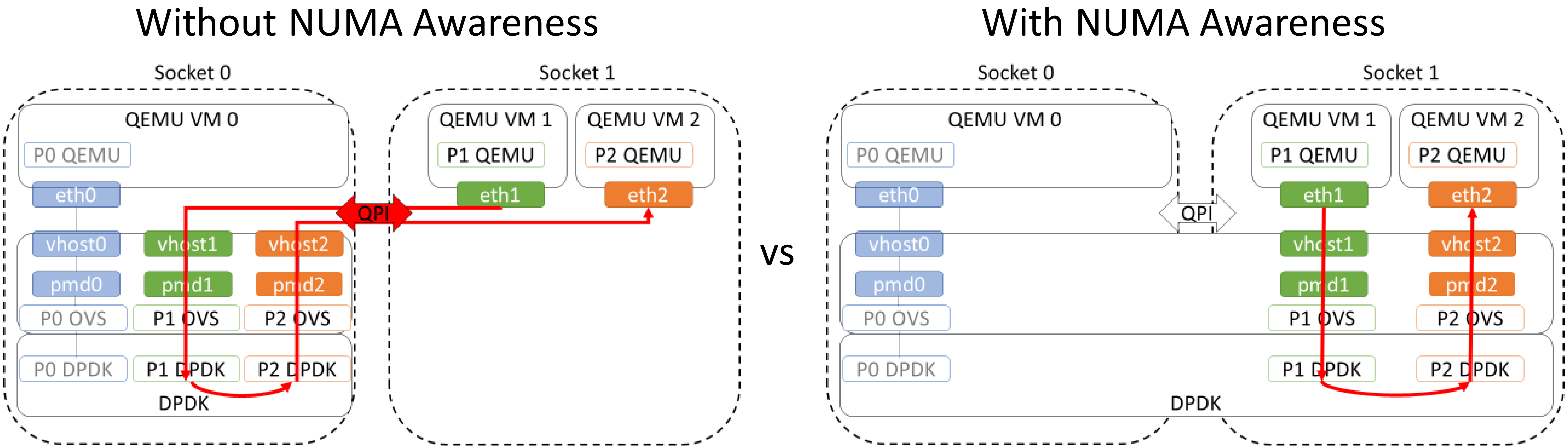
NUMA Awareness



Solution:

mbufs and servicing PMD in OVS are moved to correct NUMA during DPDK callback.

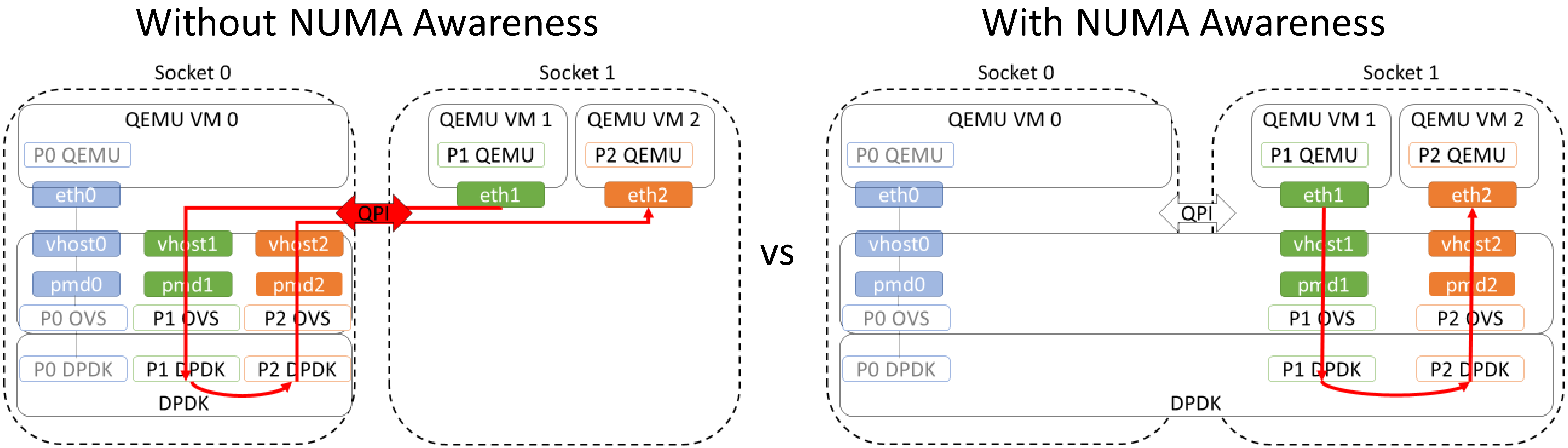
NUMA Awareness



Can achieve >50% improvement in second socket VM2VM performance*

* Platform Configuration and Test Result in Backup

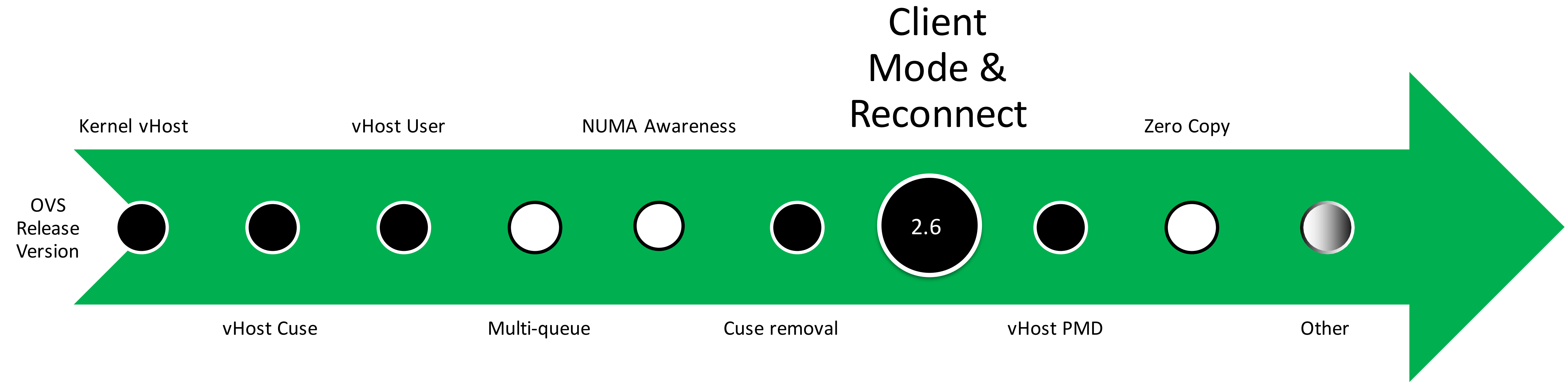
NUMA Awareness



Can achieve >50% improvement in second socket VM2VM performance*

https://software.intel.com/en-us/articles/vhost-user-numa-awareness-in-open-vswitch-with-dpdk

Client Mode & Reconnect

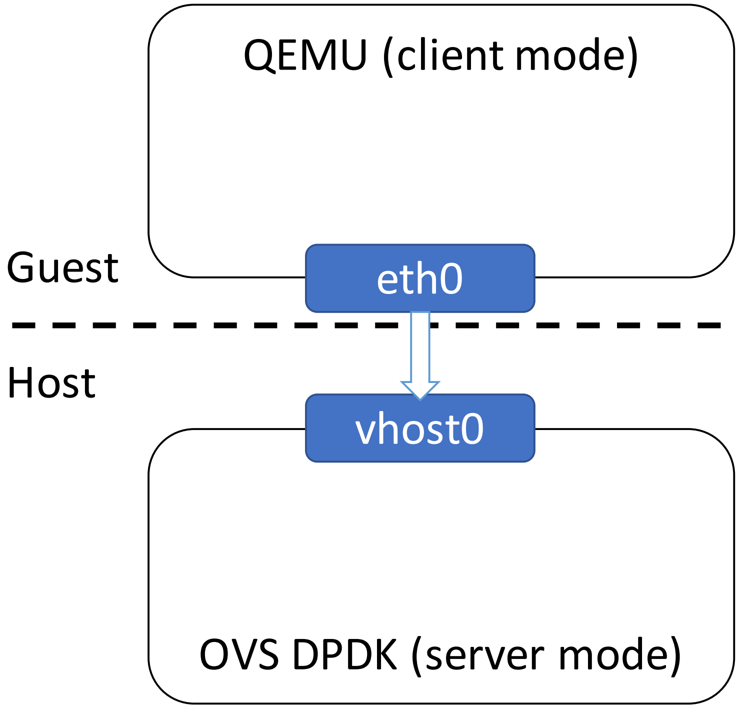


Client Mode & Reconnect

Default Mode (Server)

Previous Limitation:

VMs cannot easily regain connectivity if OVS DPDK crashes or is reset



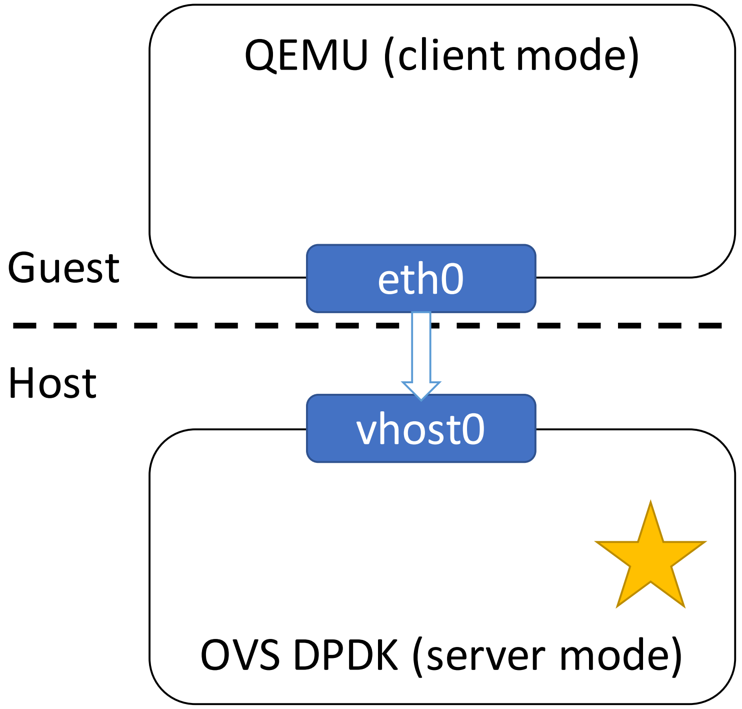
Client Mode & Reconnect

Previous Limitation:

VMs cannot easily regain connectivity if OVS DPDK crashes or is reset

OVS by default acts as the socket server

Default Mode (Server)



★ = creates/manages/destroys sockets

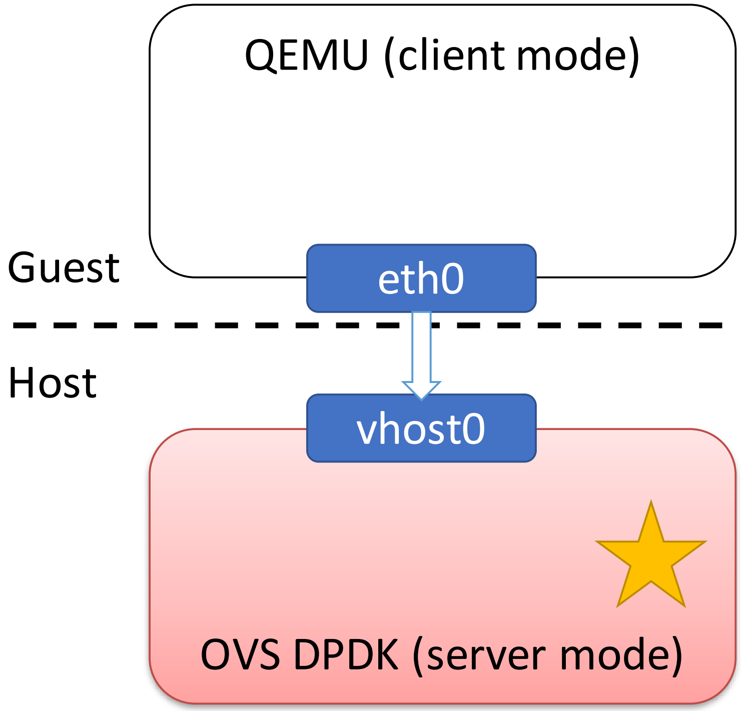
Client Mode & Reconnect

Default Mode (Server)

Previous Limitation:

VMs cannot easily regain connectivity if OVS DPDK crashes or is reset

OVS by default acts as the socket server



★ = creates/manages/destroys sockets

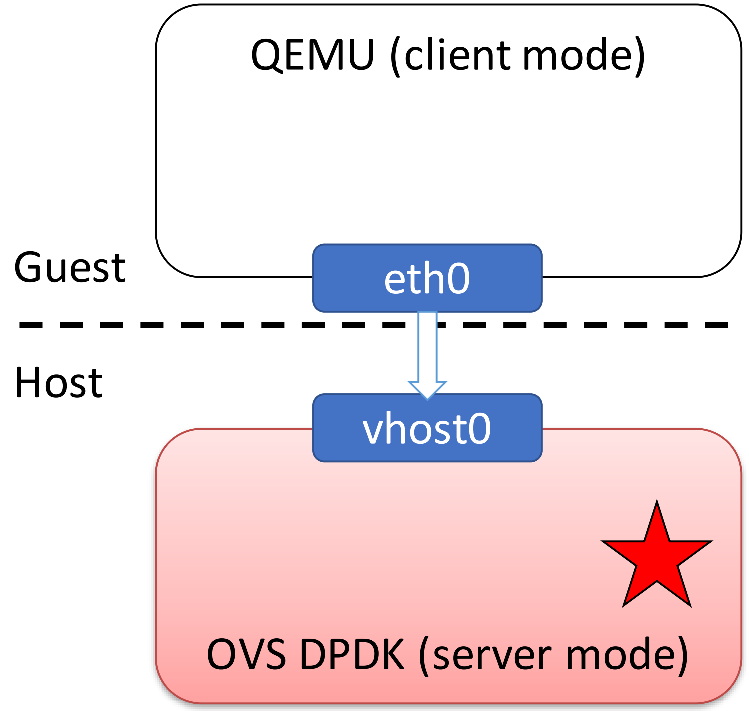
Client Mode & Reconnect

Previous Limitation:

VMs cannot easily regain connectivity if OVS DPDK crashes or is reset

OVS by default acts as the socket server

Default Mode (Server)



★ = creates/manages/destroys sockets

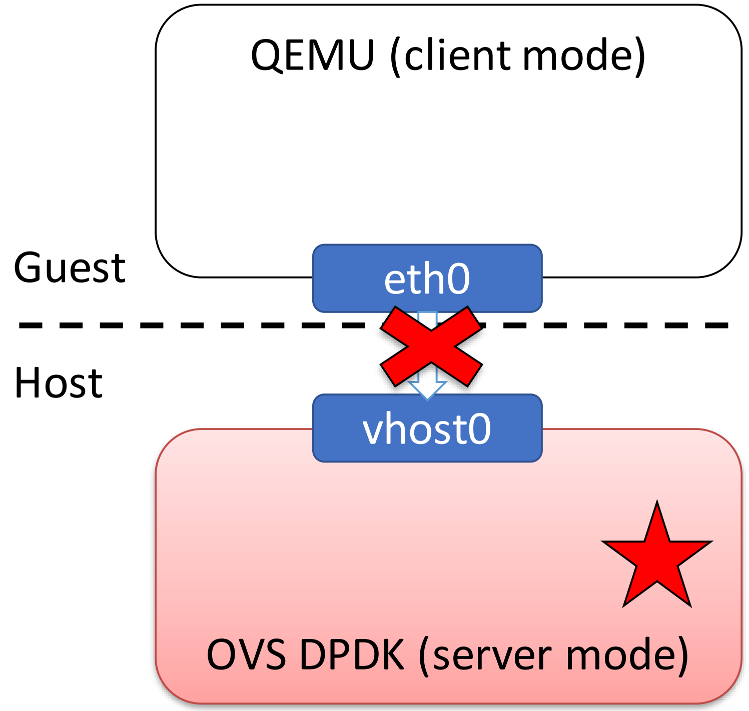
Client Mode & Reconnect

Previous Limitation:

VMs cannot easily regain connectivity if OVS DPDK crashes or is reset

OVS by default acts as the socket server

Default Mode (Server)



★ = creates/manages/destroys sockets

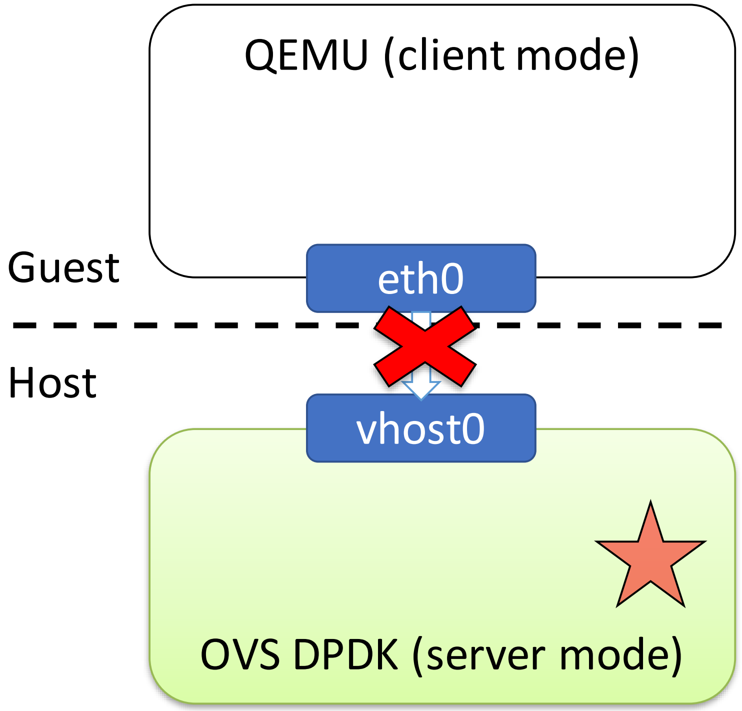
Client Mode & Reconnect

Previous Limitation:

VMs cannot easily regain connectivity if OVS DPDK crashes or is reset

OVS by default acts as the socket server

Default Mode (Server)



★ = creates/manages/destroys sockets

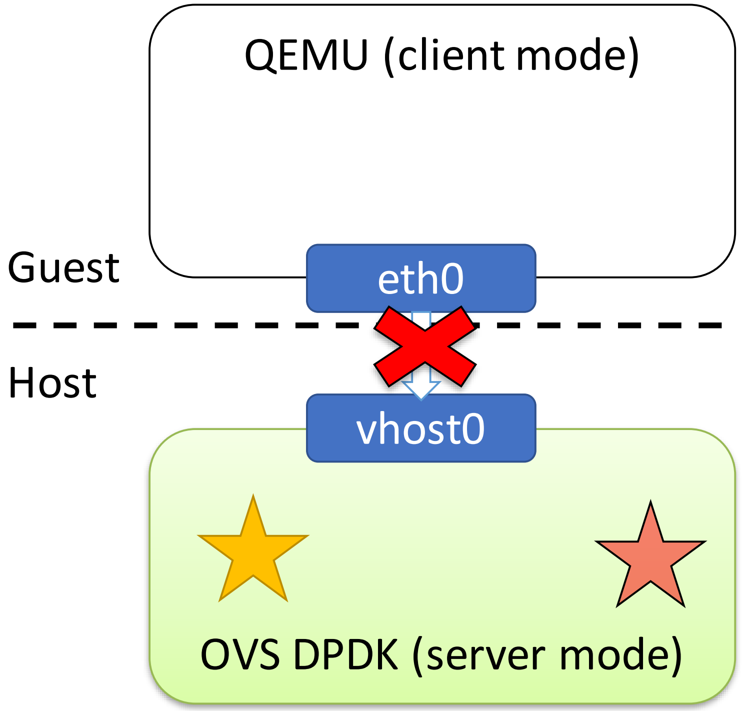
Client Mode & Reconnect

Default Mode (Server)

Previous Limitation:

VMs cannot easily regain connectivity if OVS DPDK crashes or is reset

OVS by default acts as the socket server



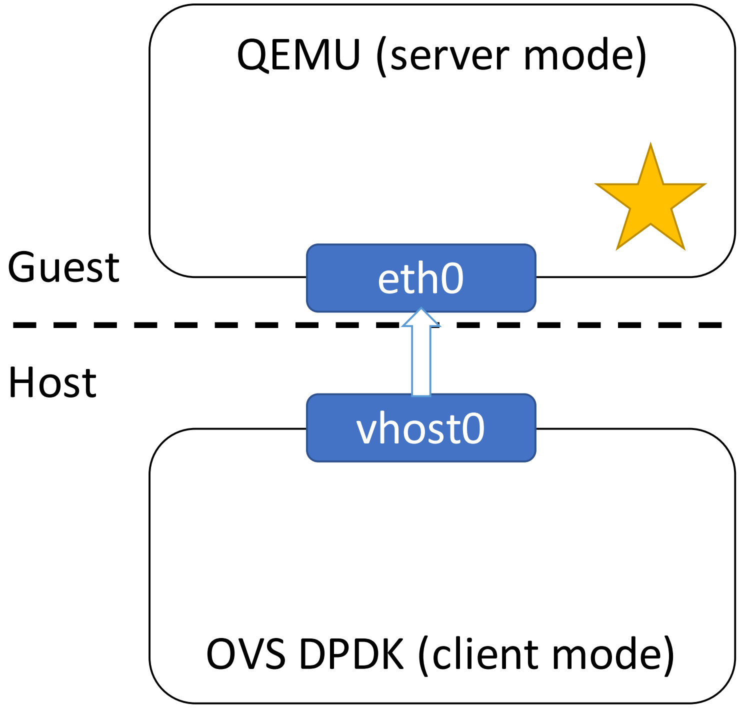
★ = creates/manages/destroys sockets

Client Mode & Reconnect

New Mode (Client)

Solution:

QEMU creates the socket and acts as the server instead



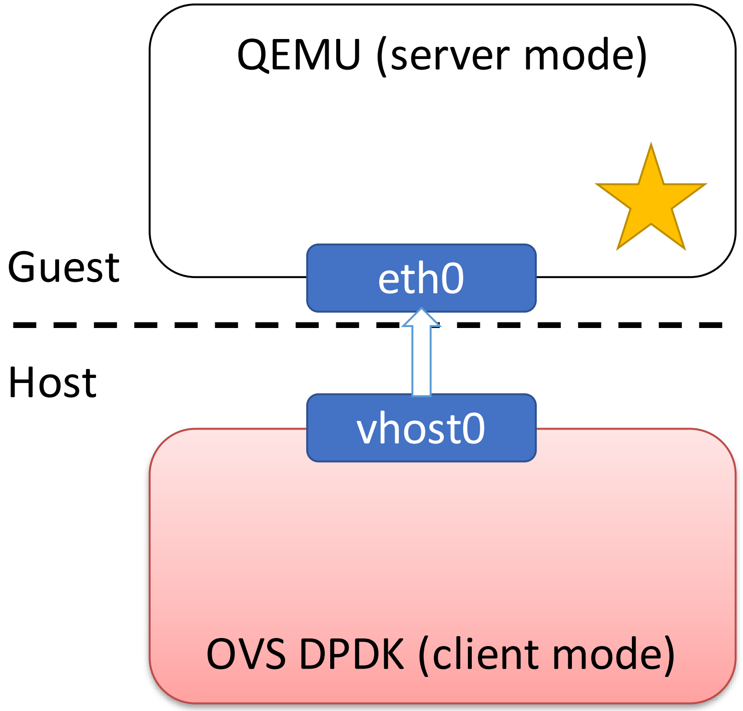
★ = creates/manages/destroys sockets

Client Mode & Reconnect

Solution:

QEMU creates the socket and acts as the server instead

New Mode (Client)



★ = creates/manages/destroys sockets

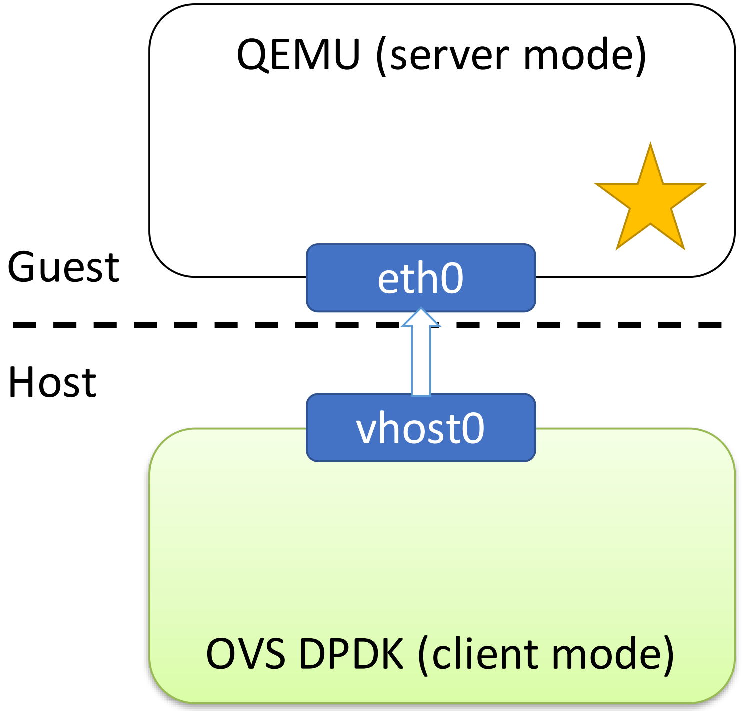
Client Mode & Reconnect

New Mode (Client)

Solution:

QEMU creates the socket and acts as the server instead

VMs can reconnect to OVS



★ = creates/manages/destroys sockets

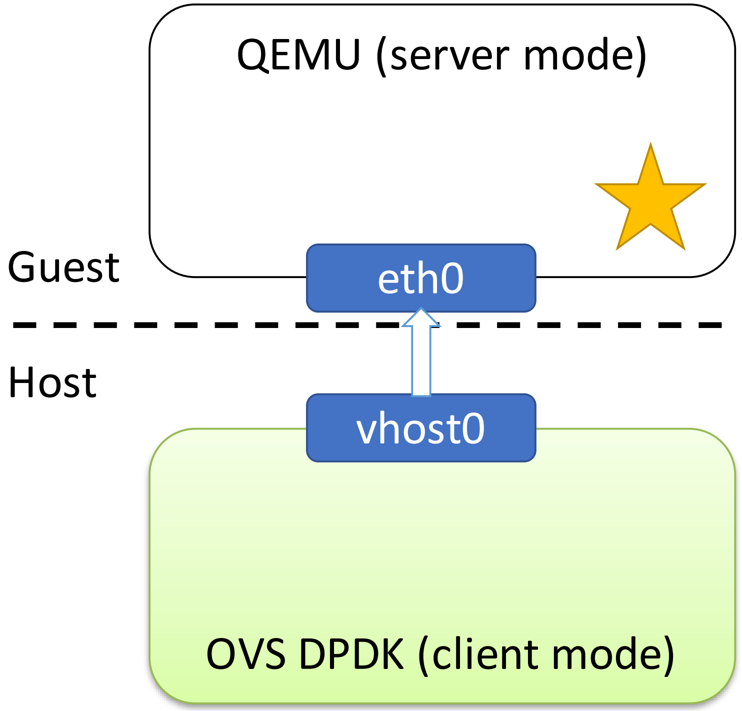
Client Mode & Reconnect

New Mode (Client)

Solution:

QEMU creates the socket and acts as the server instead

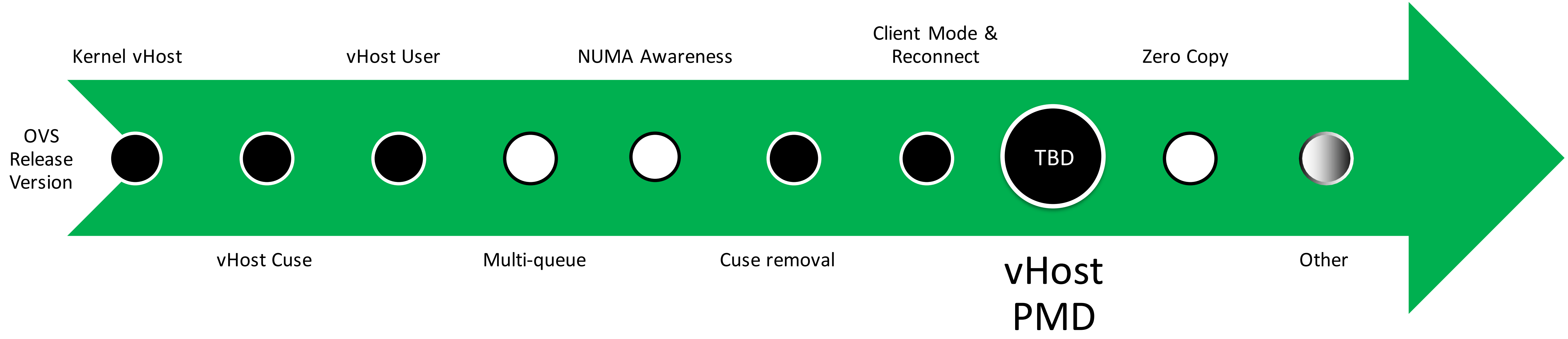
VMs can reconnect to OVS



★ = creates/manages/destroys sockets

https://software.intel.com/en-us/articles/vhost-user-client-mode-in-open-vswitch-with-dpdk

vHost PMD

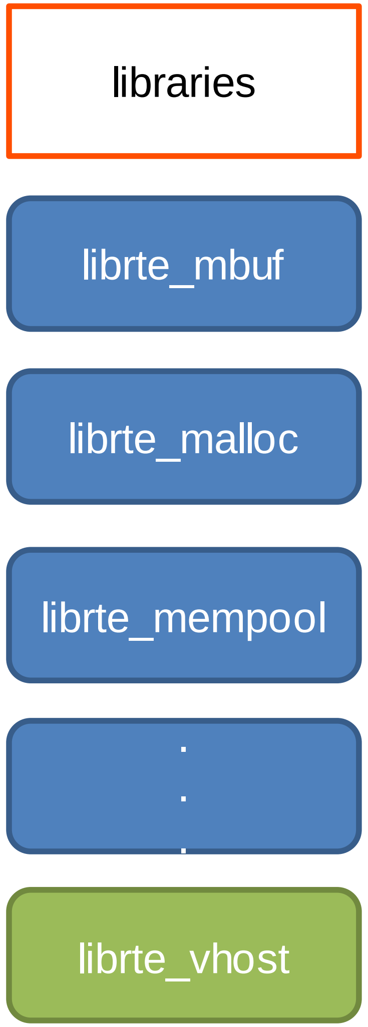

Kernel vHost
vHost User
NUMA Awareness
Client Mode & Reconnect
Zero Copy
OVS Release Version
TBD
vHost Cuse
Multi-queue
Cuse removal
vHost PMD
Other

vHost PMD

DPDK
DATA PLANE DEVELOPMENT KIT

vHost PMD

DPDK
DATA PLANE DEVELOPMENT KIT



libraries

librte_mbuf

librte_malloc

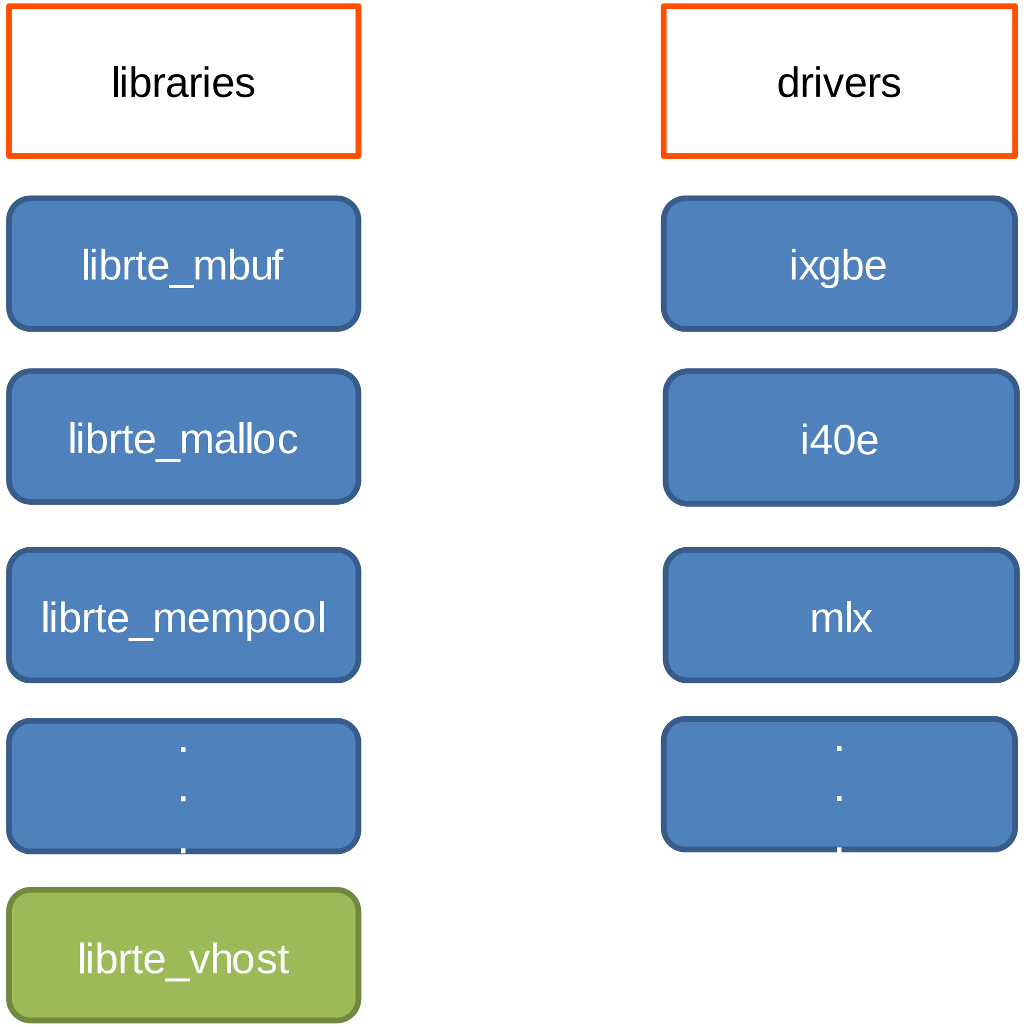
librte_mempool

.
.
.

librte_vhost

vHost PMD

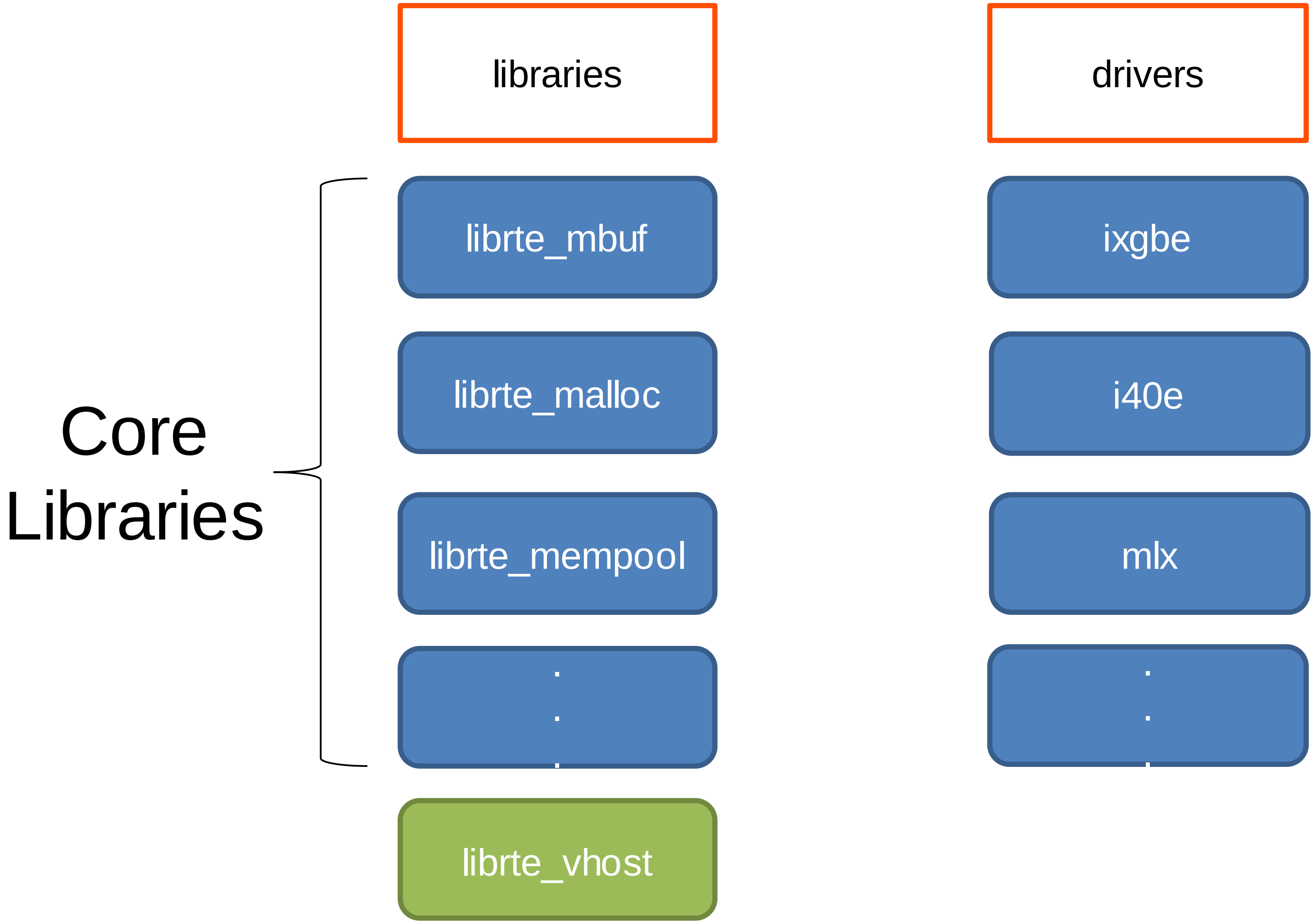
DPDK
DATA PLANE DEVELOPMENT KIT



libraries	drivers
librte_mbuf	ixgbe
librte_malloc	i40e
librte_mempool	mlx
⋮	⋮
librte_vhost	

vHost PMD

DPDK
DATA PLANE DEVELOPMENT KIT



Core Libraries

libraries

librte_mbuf

librte_malloc

librte_mempool

. . .

librte_vhost

drivers

ixgbe

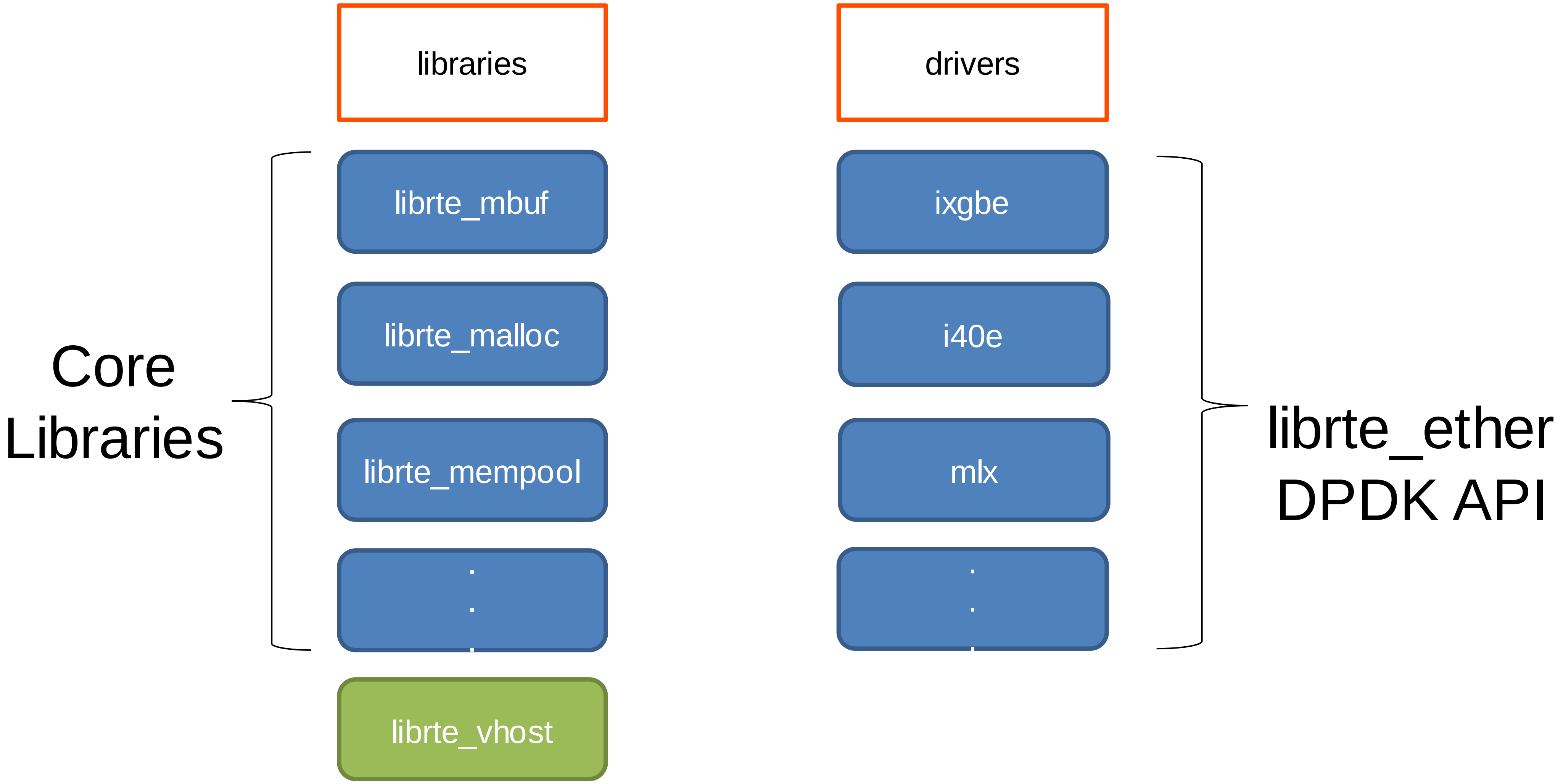
i40e

mlx

. . .

vHost PMD

DPDK
DATA PLANE DEVELOPMENT KIT



libraries

drivers

Core Libraries

librte_mbuf

librte_malloc

librte_mempool

. . .

librte_vhost

ixgbe

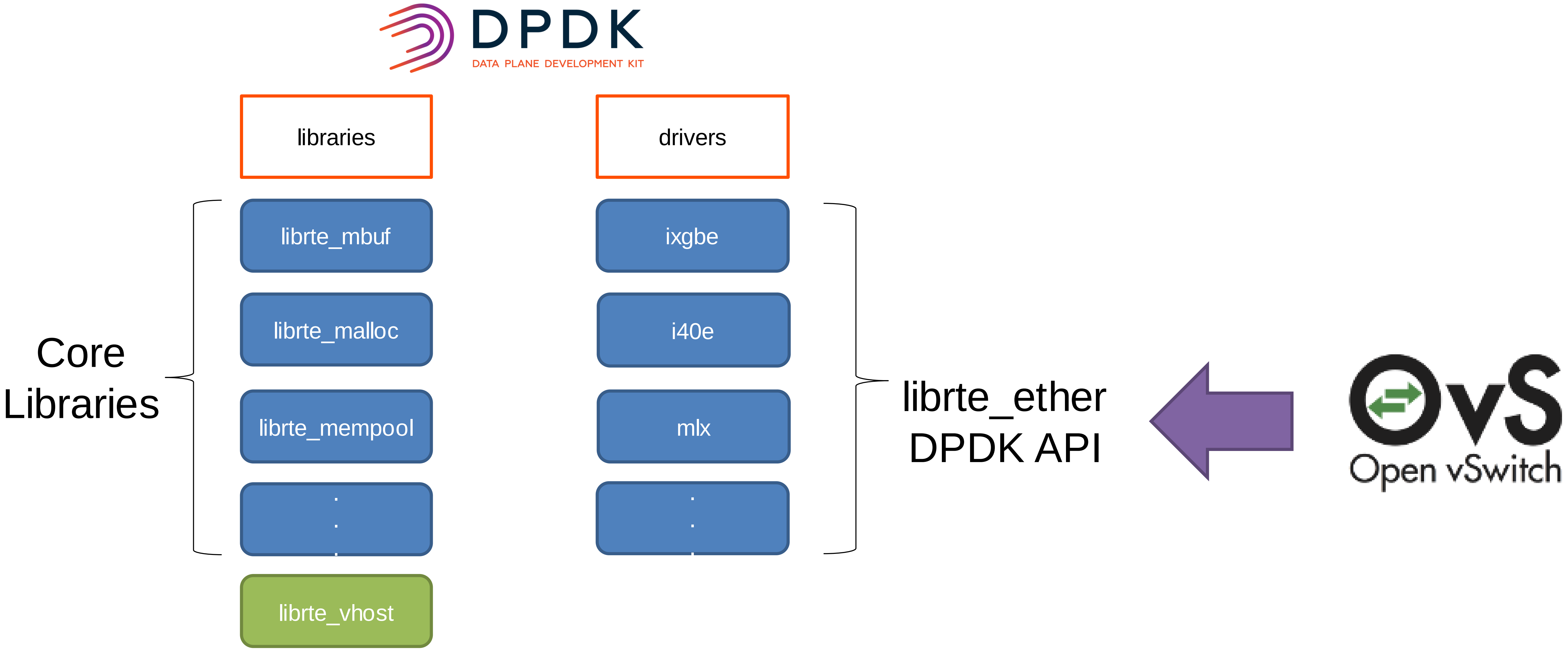
i40e

mlx

. . .

librte_ether DPDK API

vHost PMD



DPDK
DATA PLANE DEVELOPMENT KIT

libraries

drivers

Core Libraries

librte_mbuf

librte_malloc

librte_mempool

. . .

librte_vhost

ixgbe

i40e

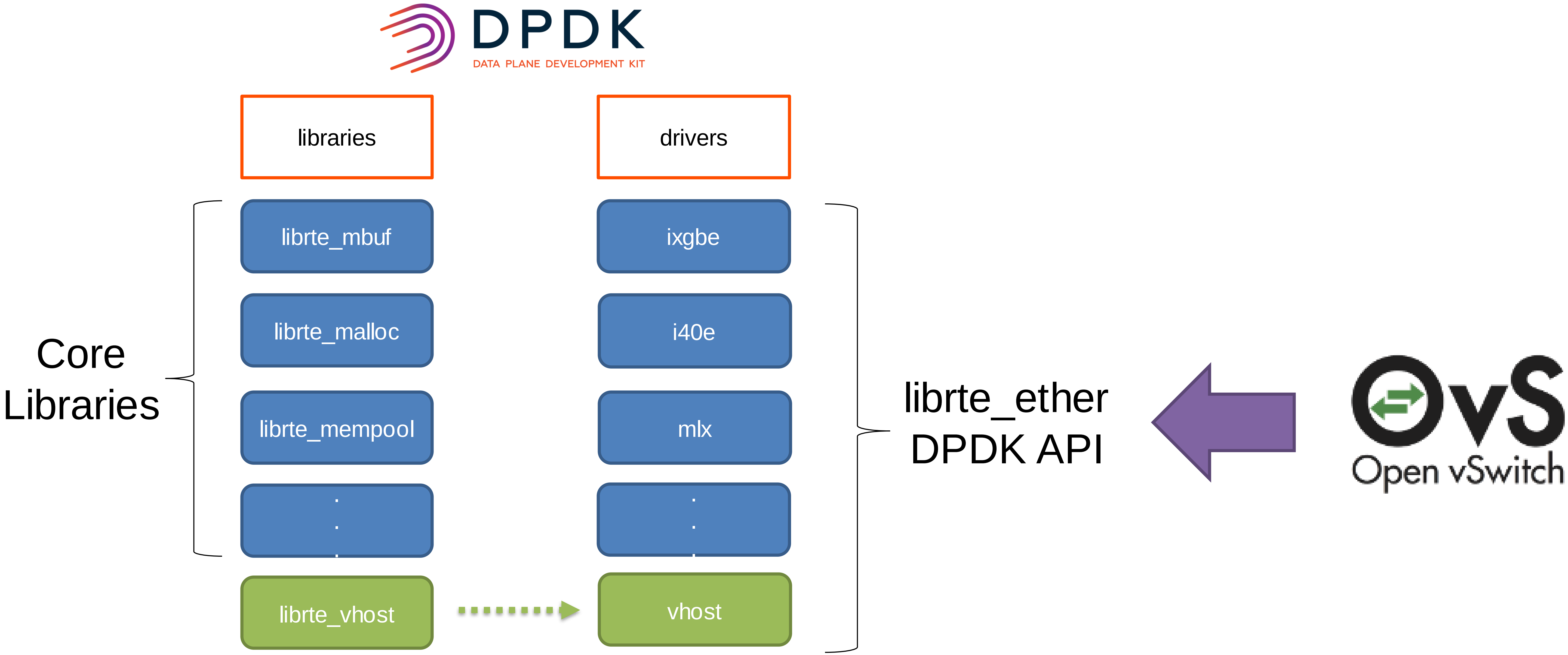
mlx

. . .

librte_ether DPDK API

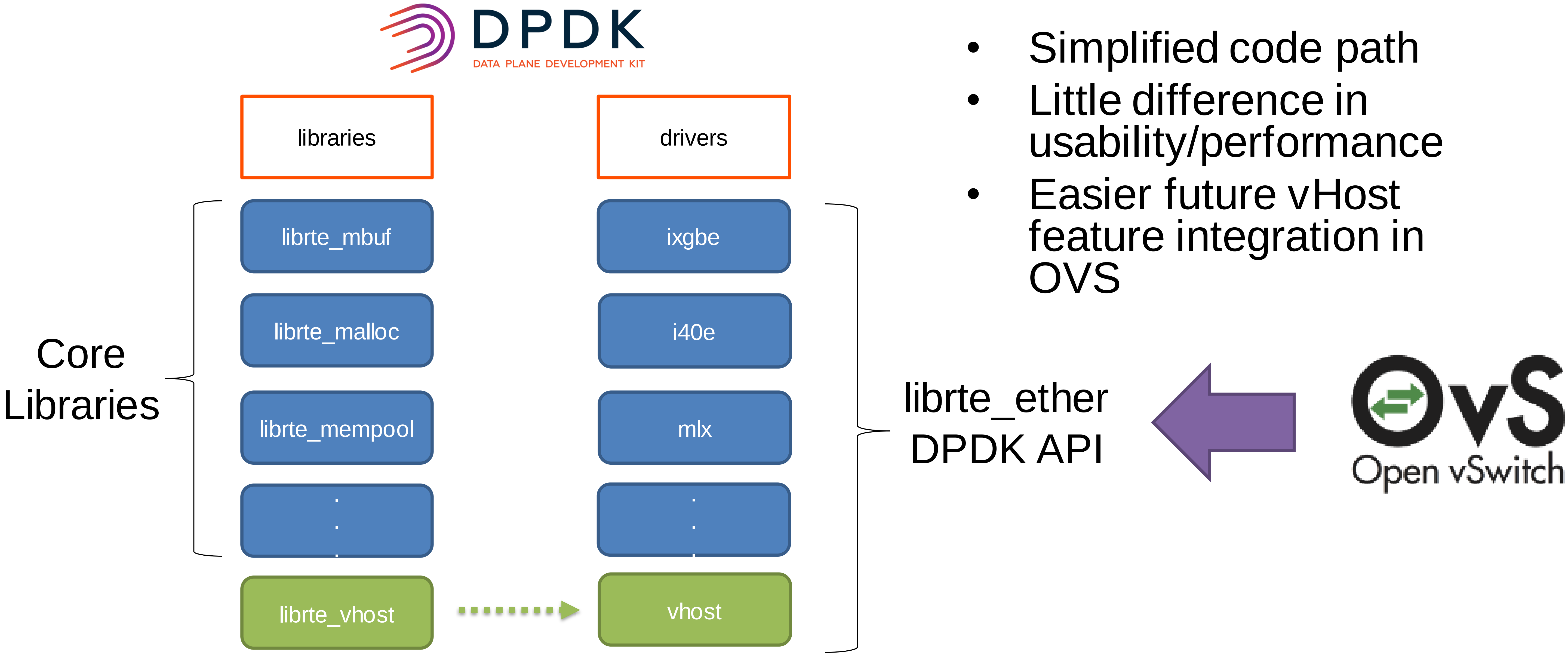
Open vSwitch

vHost PMD



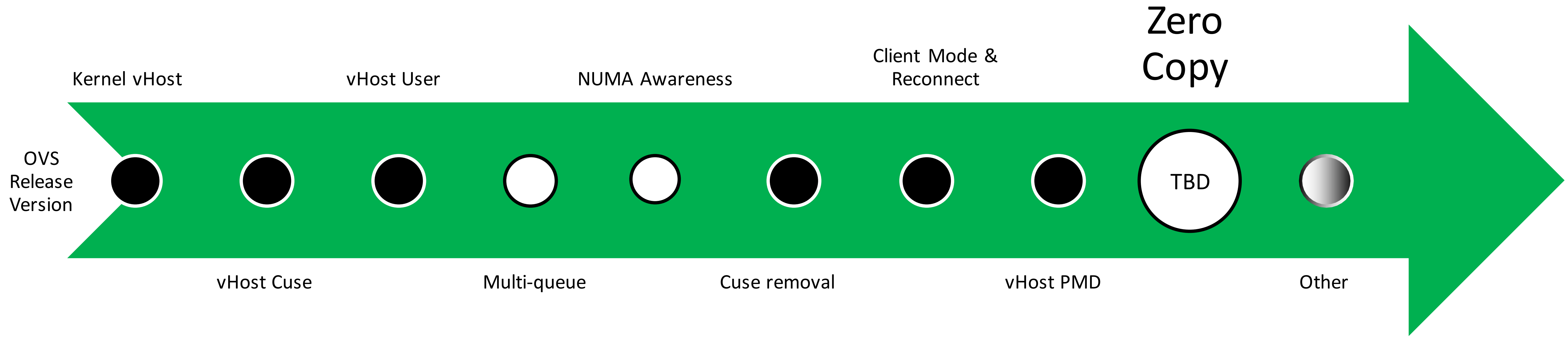
DPDK
DATA PLANE DEVELOPMENT KIT
libraries
drivers
Core Libraries
librte_mbuf
librte_malloc
librte_mempool
. . .
ixgbe
i40e
mlx
. . .
librte_vhost
vhost
librte_ether
DPDK API
Open vSwitch

vHost PMD



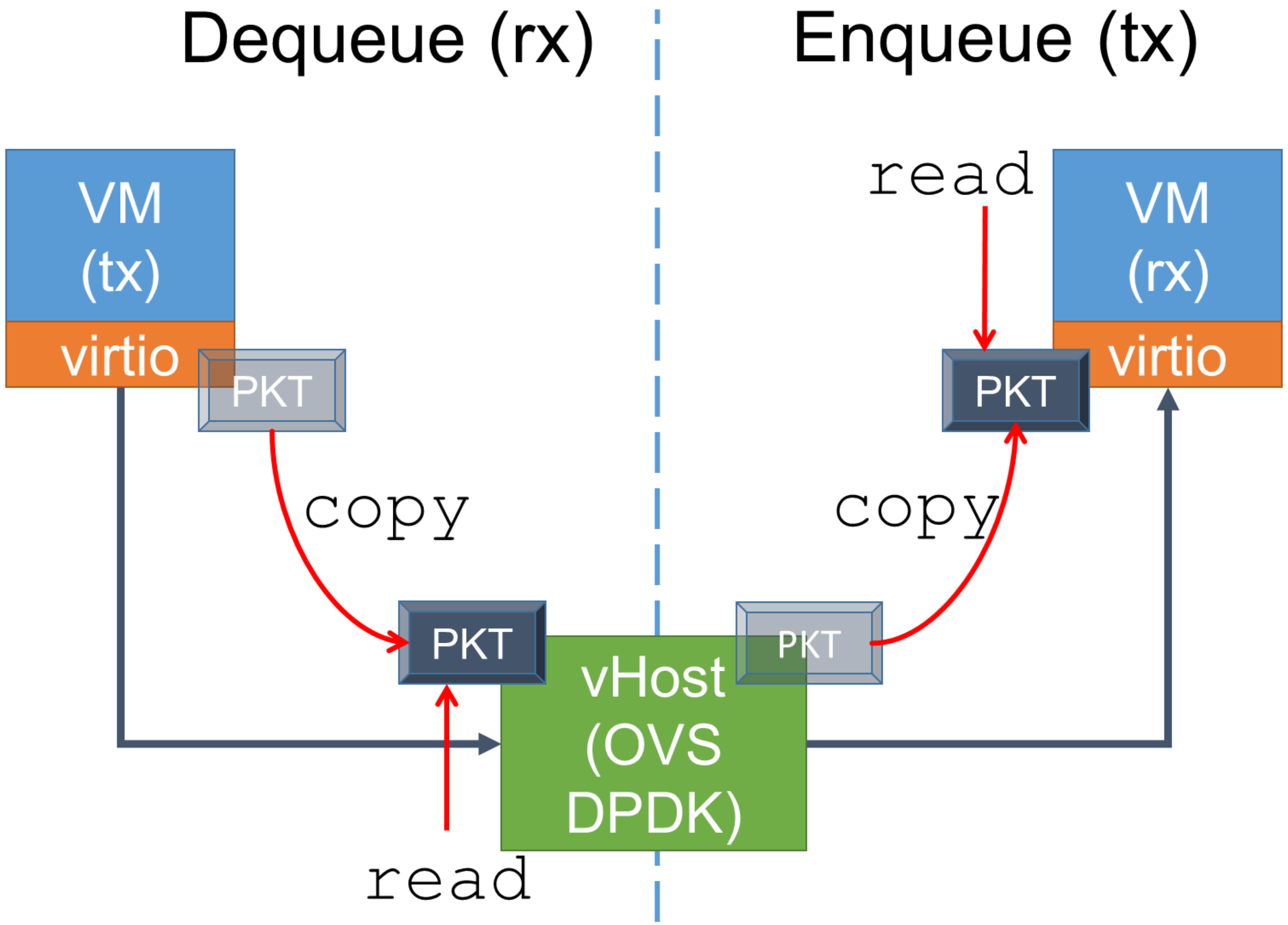
- Simplified code path
- Little difference in usability/performance
- Easier future vHost feature integration in OVS

Zero Copy


Kernel vHost
vHost User
NUMA Awareness
Client Mode & Reconnect
Zero Copy
OVS Release Version
TBD
vHost Cuse
Multi-queue
Cuse removal
vHost PMD
Other

Zero Copy

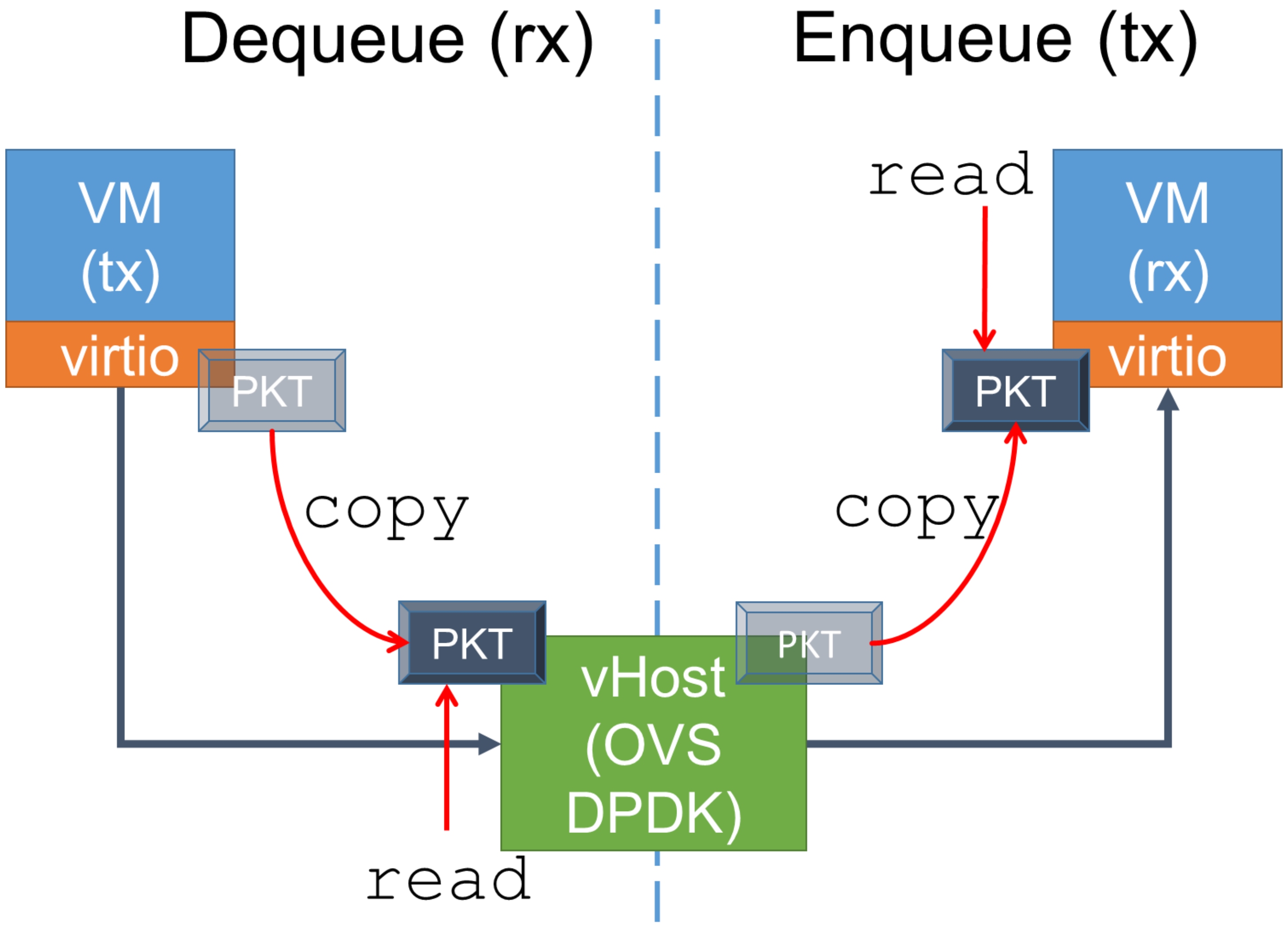
DPDK 16.11 performance improvement



Zero Copy

DPDK 16.11 performance improvement

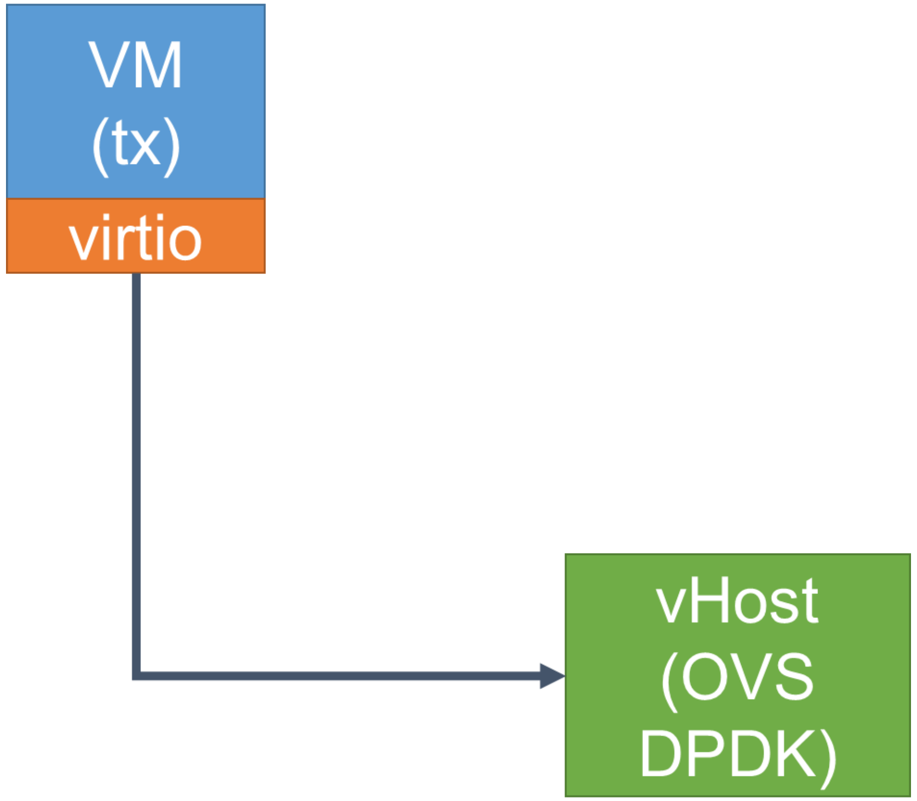
Both dequeue (rx) and enqueue (tx) paths usually incur a copy.



Zero Copy

Dequeue path involves copying a packet from the VM to the host

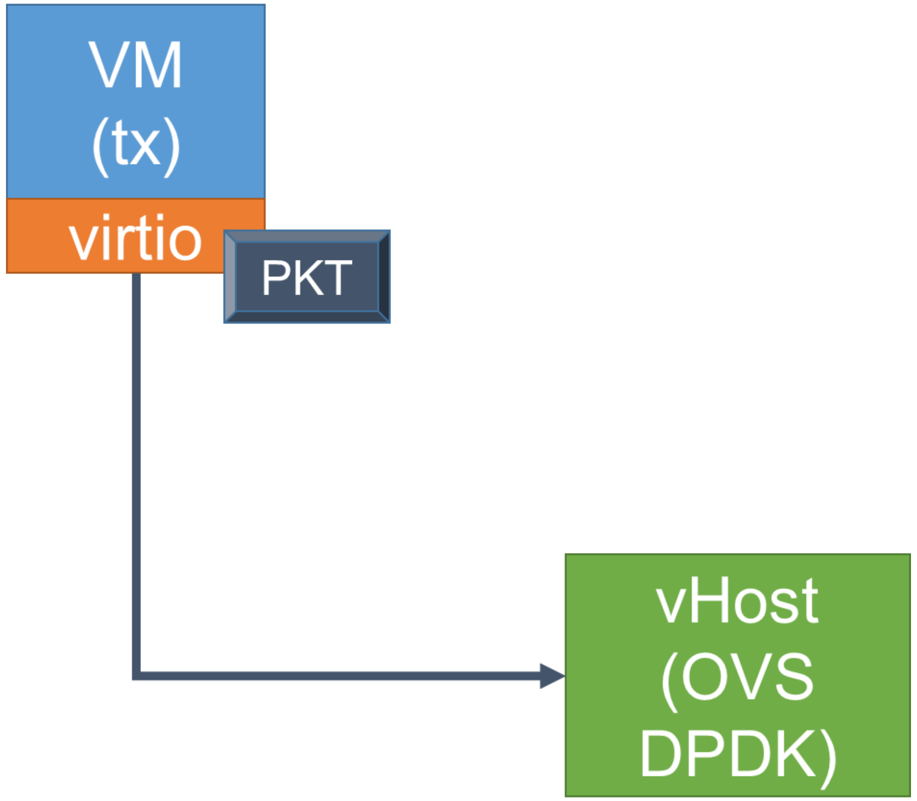
Dequeue (rx)



Zero Copy

Dequeue path involves copying a packet from the VM to the host

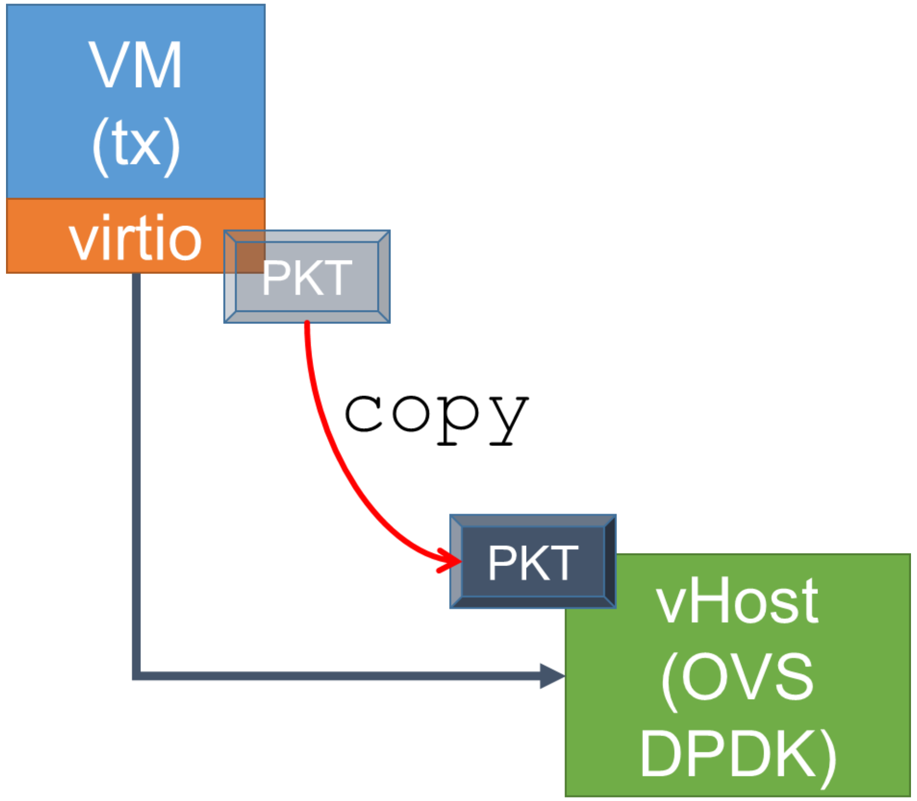
Dequeue (rx)



Zero Copy

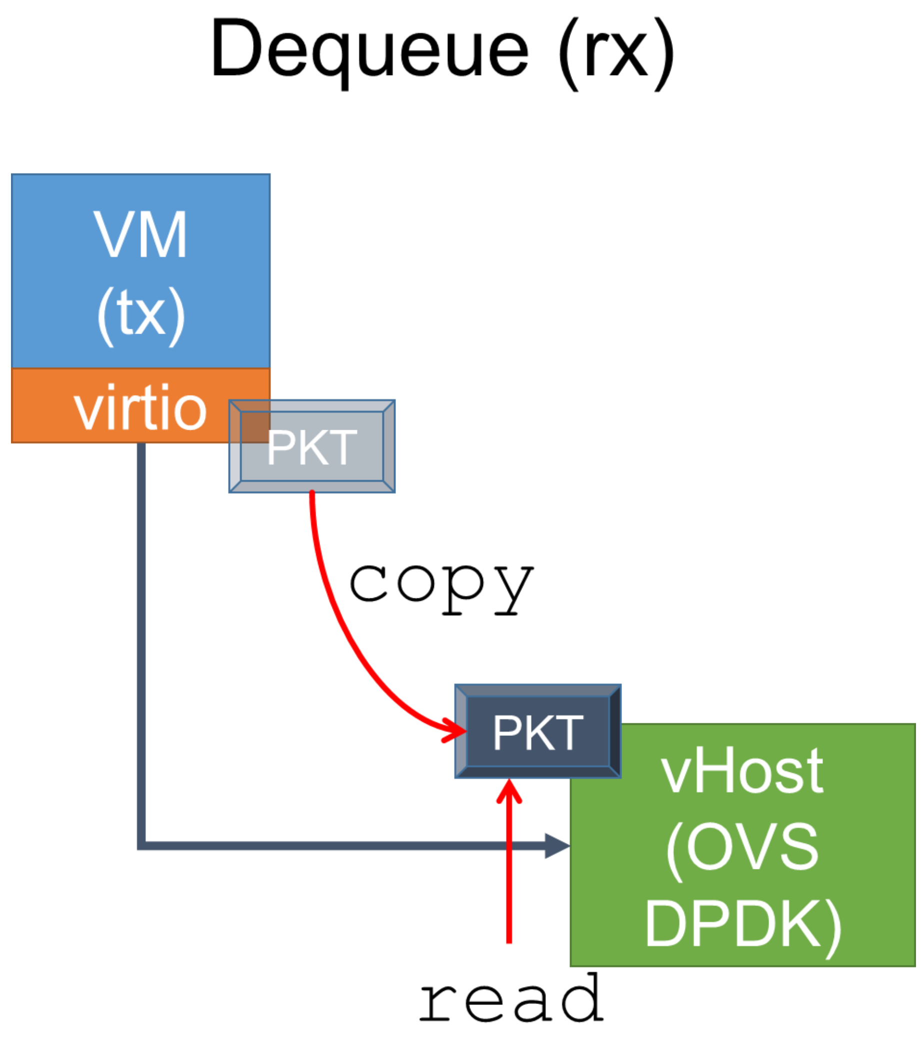
Dequeue path involves copying a packet from the VM to the host

Dequeue (rx)



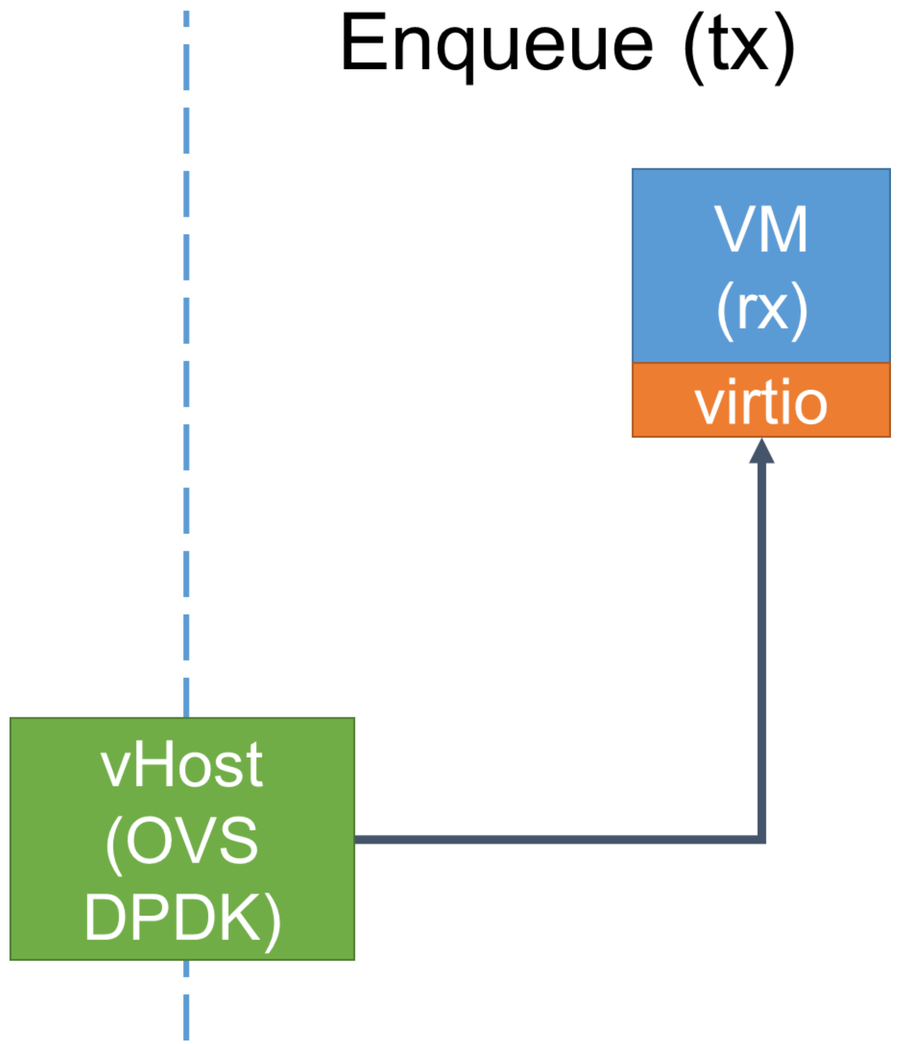
Zero Copy

Dequeue path involves copying a packet from the VM to the host



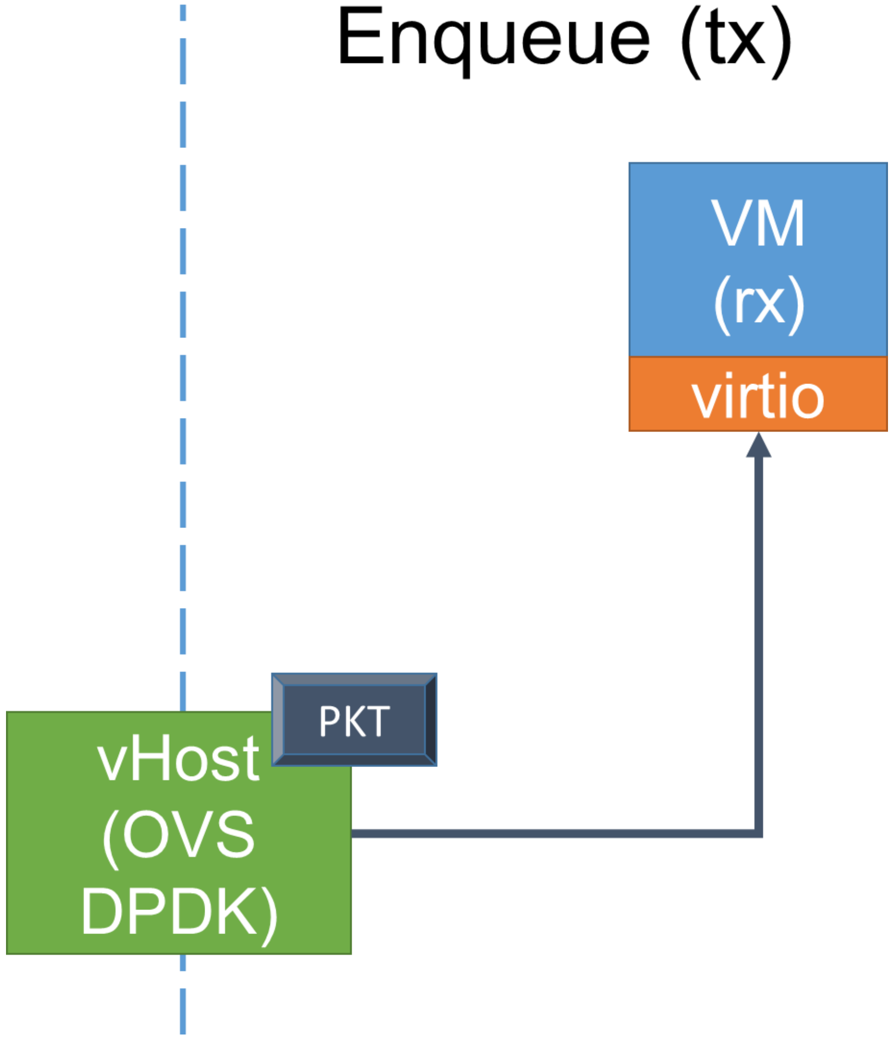
Zero Copy

Enqueue path involves coping a packet from the host to the VM



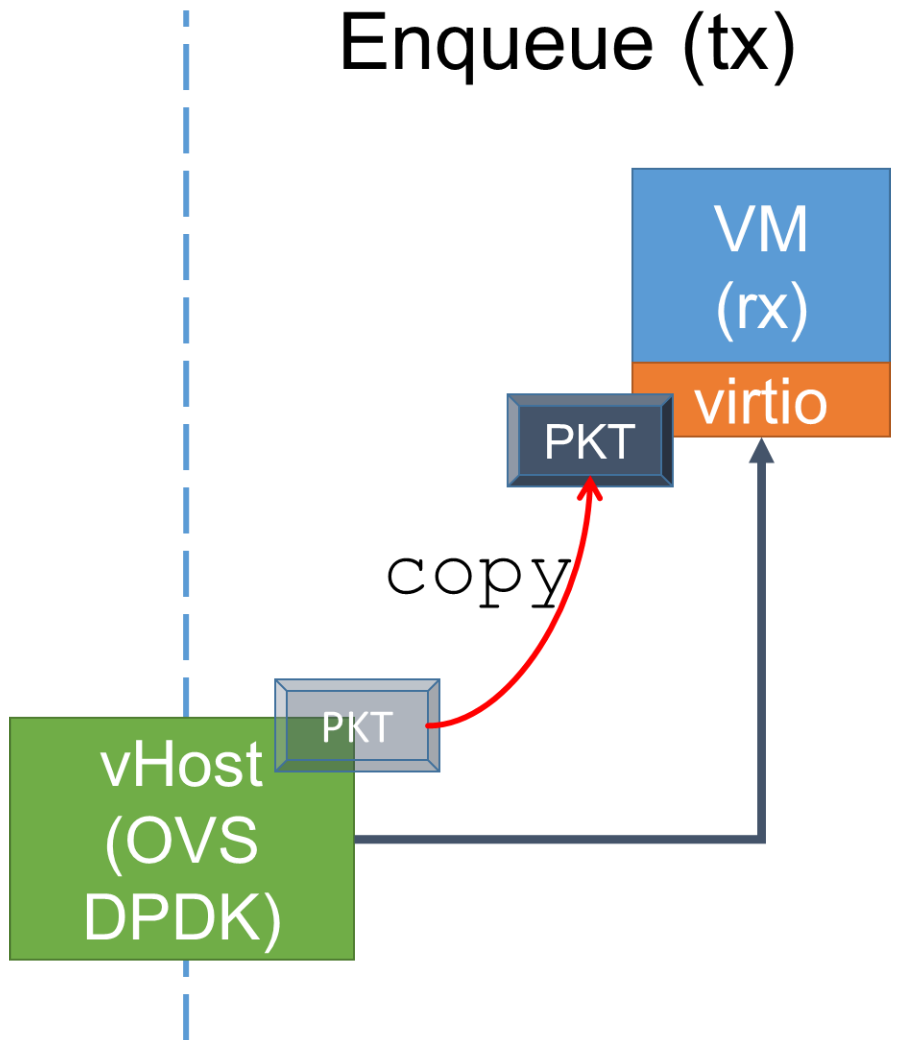
Zero Copy

Enqueue path involves coping a packet from the host to the VM



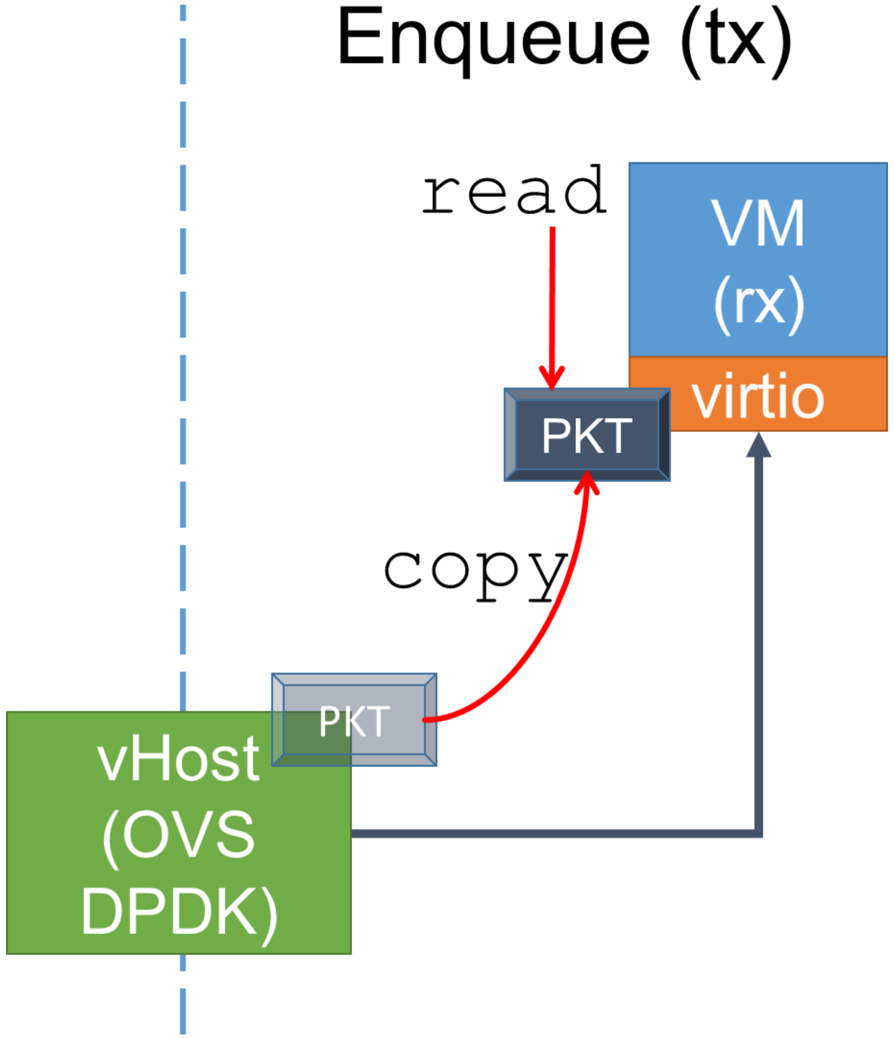
Zero Copy

Enqueue path involves coping a packet from the host to the VM



Zero Copy

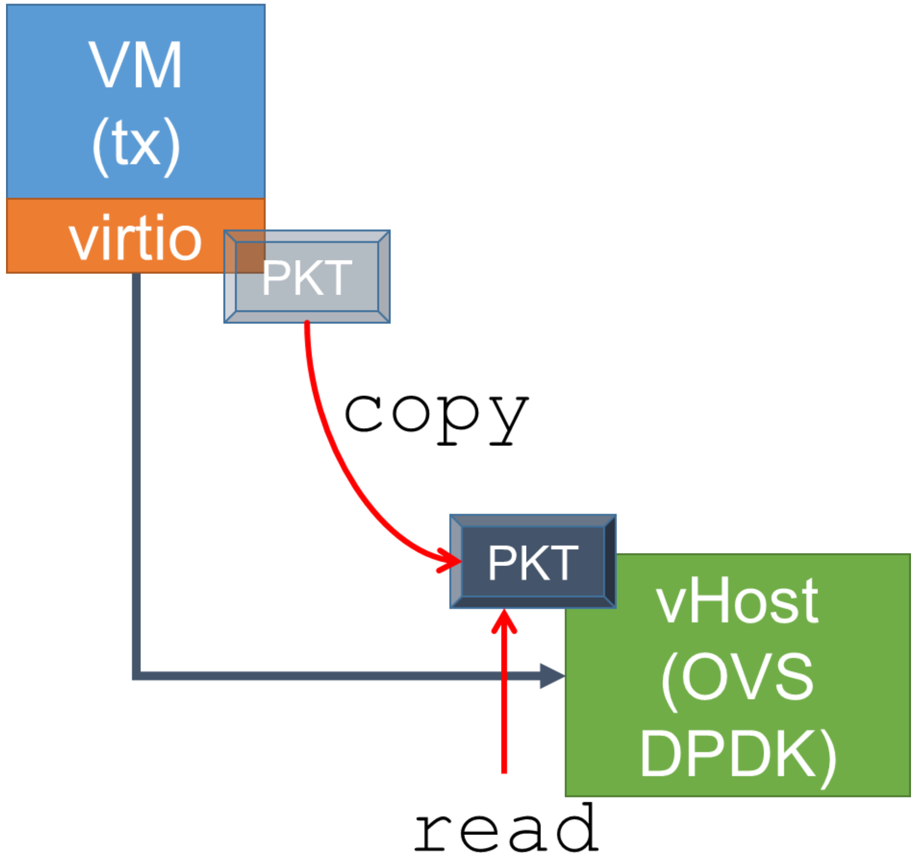
Enqueue path involves coping a packet from the host to the VM



Zero Copy

Zero copy is possible for dequeue if the mbuf references the virtio descriptor buffer **directly**.

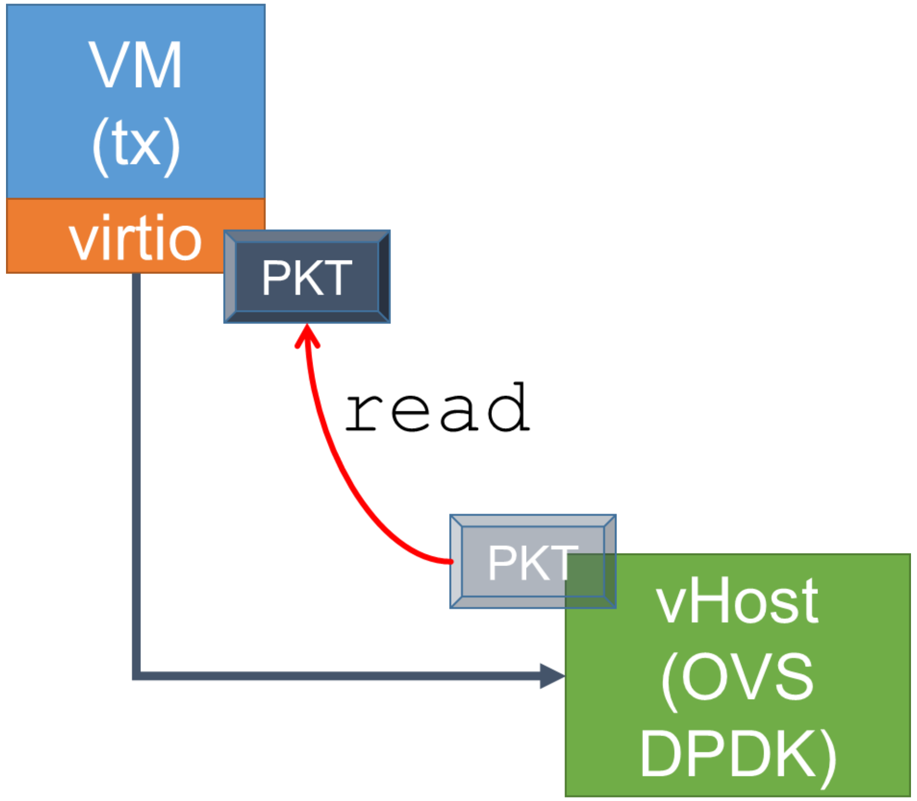
Dequeue (rx)



Zero Copy

Zero copy is possible for dequeue if the mbuf references the virtio descriptor buffer **directly**.

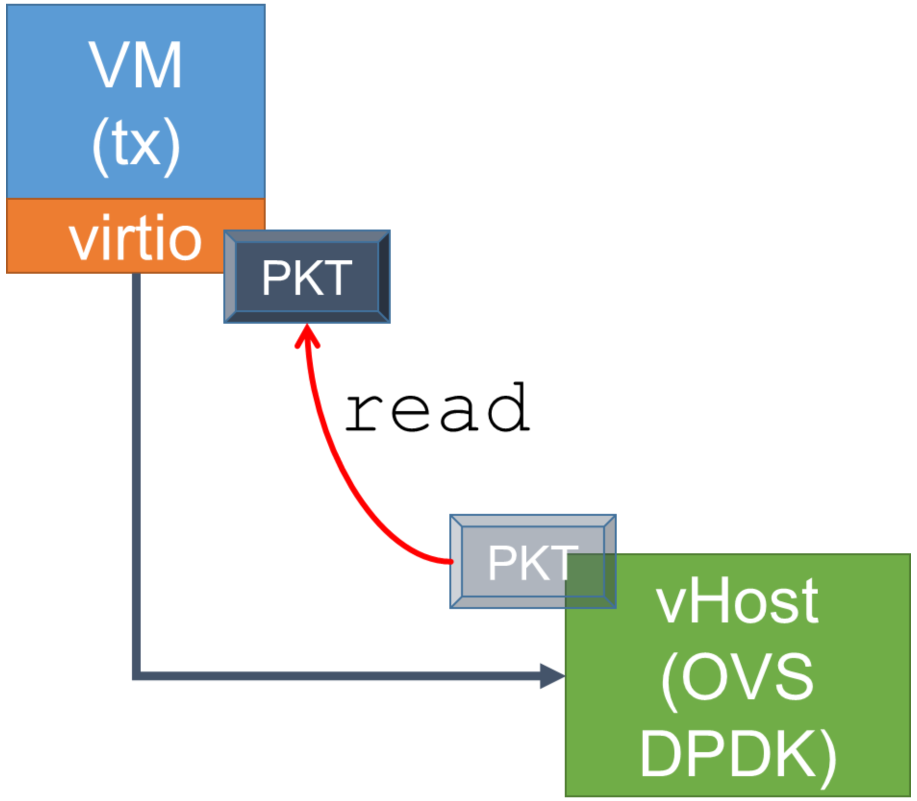
Dequeue (rx)



Zero Copy

Not suitable for small packet sizes (~ < 512B)

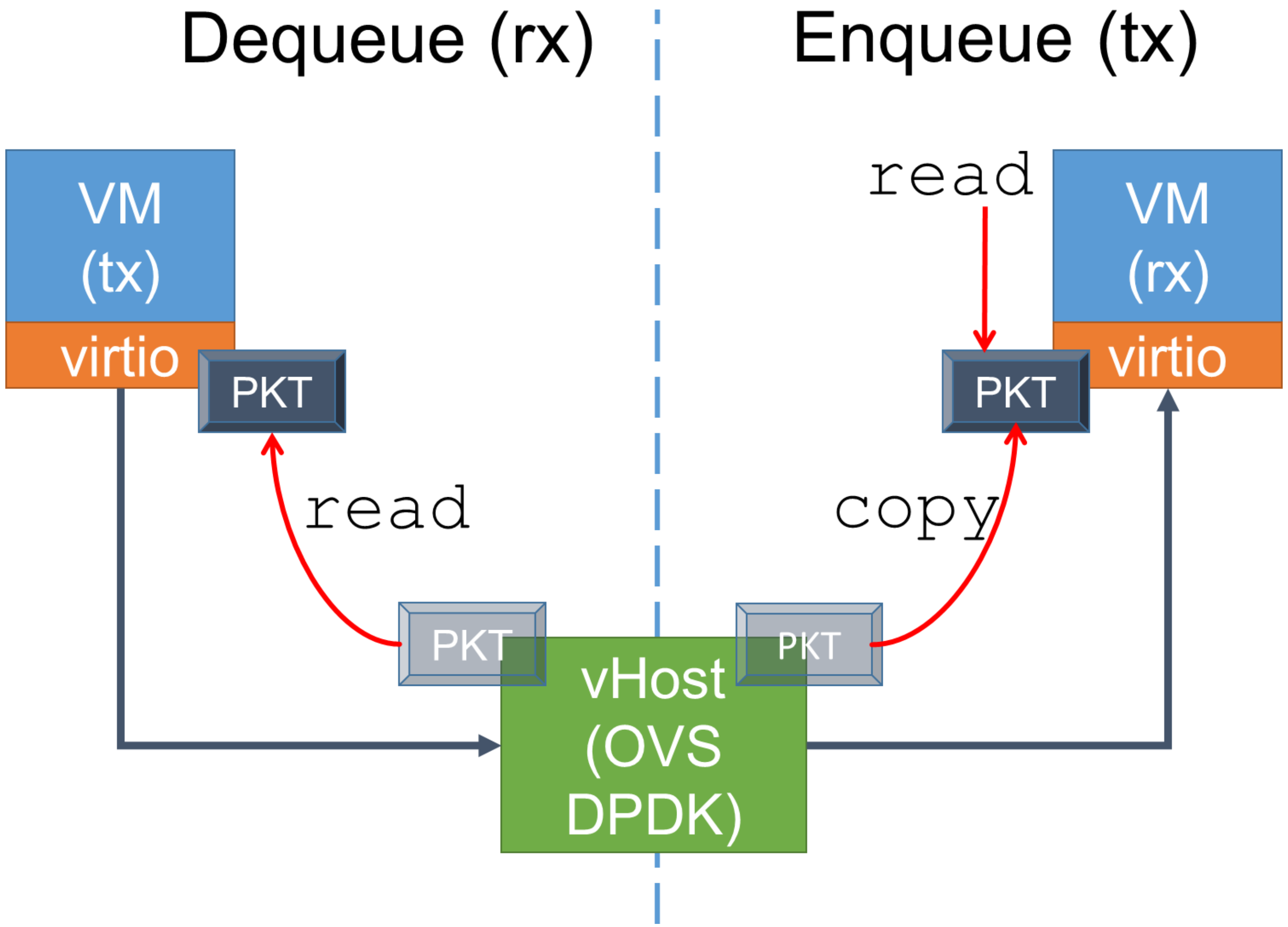
Dequeue (rx)



Zero Copy

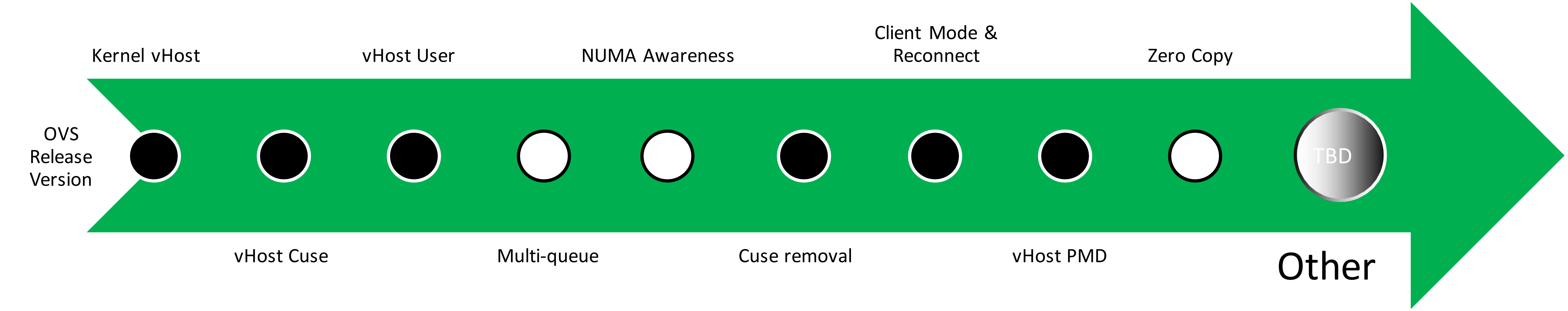
Can achieve >15% increase in throughput for 1518B packets for this use case*

(vHost ⇨ OVS-DPDK ⇨ vHost)



* Platform Configuration and Test Result in Backup

Other Future Improvements



● = Functional Improvement

○ = Performance Improvement

Other Future Improvements

- Virtio User (16.11)
  - New “PMD”
  - Method of using vHost User in containers

Other Future Improvements

- Virtio User (16.11)
  - New “PMD”
  - Method of using vHost User in containers
- Mergeable buffers path improvement (16.11)

Other Future Improvements

- Virtio User (16.11)
  - New “PMD”
  - Method of using vHost User in containers
- Mergeable buffers path improvement (16.11)
- vHost PCI (POC)
  - VM2VM path performance enhancement
  - vHost vEth pair

Conclusion

- Since it’s introduction to OVS in 2015, many incremental improvements to DPDK vHost User have been added.

- Many more improvements to look forward to.

OvS

Open vSwitch

Questions?

OvS

Open vSwitch

BACKUP

Legal Disclaimer

General Disclaimer:

© Copyright 2016 Intel Corporation. All rights reserved. Intel, the Intel logo, Intel Inside, the Intel Inside logo, Intel. Experience What's Inside are trademarks of Intel. Corporation in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others.

Technology Disclaimer:

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at [intel.com].

Performance Disclaimers (include only the relevant ones):

Cost reduction scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

Platform Configuration & Test Results

Item	Description
Server Platform	Intel® Server Board S2600WTT (Formerly Wildcat Pass) 2 x 1GbE integrated LAN ports Two processors per platform
Chipset	Intel® C610/X99 series chipset (Formerly Wellsburg)
Processor	Intel® Xeon® Processor E5-2695 v3 (Formerly Haswell) Speed and power: 2.30 GHz, 120 W Cache: 35 MB per processor Cores: 14 cores, 28 hyper-threaded cores per processor for 56 total hyper-threaded cores QPI: 9.6 GT/s Memory types: DDR4-1600/1866/2133, Reference: http://ark.intel.com/products/81057/Intel-Xeon-Processor-E5-2695-v3-35M-Cache-2_30-GHz
Memory	Micron 16 GB 1Rx4 PC4-2133MHz, 16 GB per channel, 8 Channels
NICs	2 x Intel® Ethernet CAN X710 Adapter (Total: 4 x 10GbE ports) (Formerly Fortville)
BIOS	Version: SE5C610.86B.01.01.0008.021120151325 Date: 02/11/2015
OS	Fedora 22
Software	DPDK - v2.2.0, OVS – v2.5.0 pre-release (commit 522aca), QEMU – 2.3.0, Linux kernel – 4.0.6-300.fc22.x86_64

Guest Access Method	Packets per Second
virtio-net	51131
vhost-net	406515
vhost-user	3366374

Platform Configuration & Test Results

Item	Description
Server Platform	Intel® Server Board S2600WTT (Formerly Wildcat Pass) 2 x 1GbE integrated LAN ports Two processors per platform
Chipset	Intel® C610/X99 series chipset (Formerly Wellsburg)
Processor	Intel® Xeon® Processor E5-2695 v3 (Formerly Haswell) Speed and power: 2.30 GHz, 120 W Cache: 35 MB per processor Cores: 14 cores, 28 hyper-threaded cores per processor for 56 total hyper-threaded cores QPI: 9.6 GT/s Memory types: DDR4-1600/1866/2133, Reference: http://ark.intel.com/products/81057/Intel-Xeon-Processor-E5-2695-v3-35M-Cache-2_30-GHz
Memory	Micron 16 GB 1Rx4 PC4-2133MHz, 16 GB per channel, 8 Channels
NICs	2 x Intel® Ethernet CAN X710 Adapter (Total: 4 x 10GbE ports) (Formerly Fortville)
BIOS	Version: SE5C610.86B.01.01.0008.021120151325 Date: 02/11/2015
OS	Fedora 22
Software	DPDK – v16.07, OVS – v2.6.0 (commit 136e425df951), QEMU – 2.7.0, Linux kernel – 4.2.8-200.fc22.x86_64

	Packets per Second
Without NUMA Awareness	2545945
With NUMA Awareness	3831019

Platform Configuration & Test Results

Item	Description
Server Platform	Intel® Server Board S2600WTT (Formerly Wildcat Pass) 2 x 1GbE integrated LAN ports Two processors per platform
Chipset	Intel® C610/X99 series chipset (Formerly Wellsburg)
Processor	Intel® Xeon® Processor E5-2695 v3 (Formerly Haswell) Speed and power: 2.30 GHz, 120 W Cache: 35 MB per processor Cores: 14 cores, 28 hyper-threaded cores per processor for 56 total hyper-threaded cores QPI: 9.6 GT/s Memory types: DDR4-1600/1866/2133, Reference: http://ark.intel.com/products/81057/Intel-Xeon-Processor-E5-2695-v3-35M-Cache-2_30-GHz
Memory	Micron 16 GB 1Rx4 PC4-2133MHz, 16 GB per channel, 8 Channels
NICs	2 x Intel® Ethernet CAN X710 Adapter (Total: 4 x 10GbE ports) (Formerly Fortville)
BIOS	Version: SE5C610.86B.01.01.0008.021120151325 Date: 02/11/2015
OS	Fedora 22
Software	DPDK – v16.11-rc2, OVS – v2.6.0 (commit 136e425df951, patched to enable feature), QEMU – 2.7.0, Linux kernel – 4.2.8-200.fc22.x86_64

	Packets per Second
Without zero copy	2094554
With zero copy	2415784