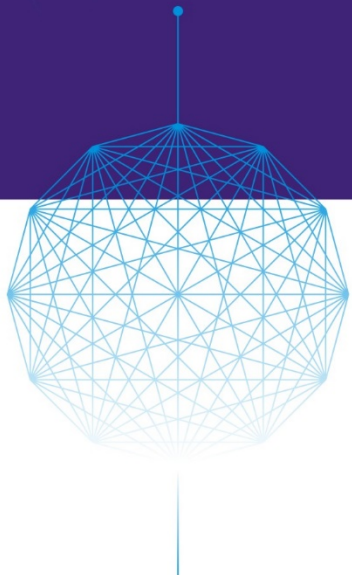







DPDK SUMMIT CHINA 2017



主办方：

参与方： 腾讯云  ZTE  美团云  Panabit®  太一星晨  UnitedStack 联合云  云杉网络 Yunshan Networks

协办方： SDNLAB 专注网络创新技术 视频支持方： IT大咖说 网络全媒平台




A BETTER VIRTIO TOWARDS NFV CLOUD

VHOST DATAPATH ACCELERATION




Cunming LIANG, Intel
Xiao WANG, Intel



DPDK China Summit 2017, Shanghai

主办方: 

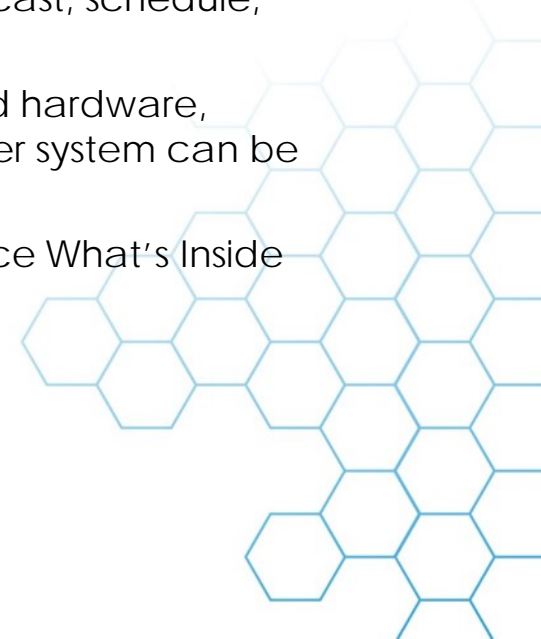
参与方:  腾讯云  ZTE  美团云  Panabit  太一星晨  云杉网络

协办方:  SDNLAB  中国网络创新技术 视频支持方:  IT大咖说



LEGAL DISCLAIMER

- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.
- Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.
- This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.
- 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.
- © 2017 Intel Corporation. Intel, the Intel logo, Intel. Experience What's Inside, and the Intel. Experience What's Inside logo 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.
- Copyright © 2017, Intel Corporation. All rights reserved.





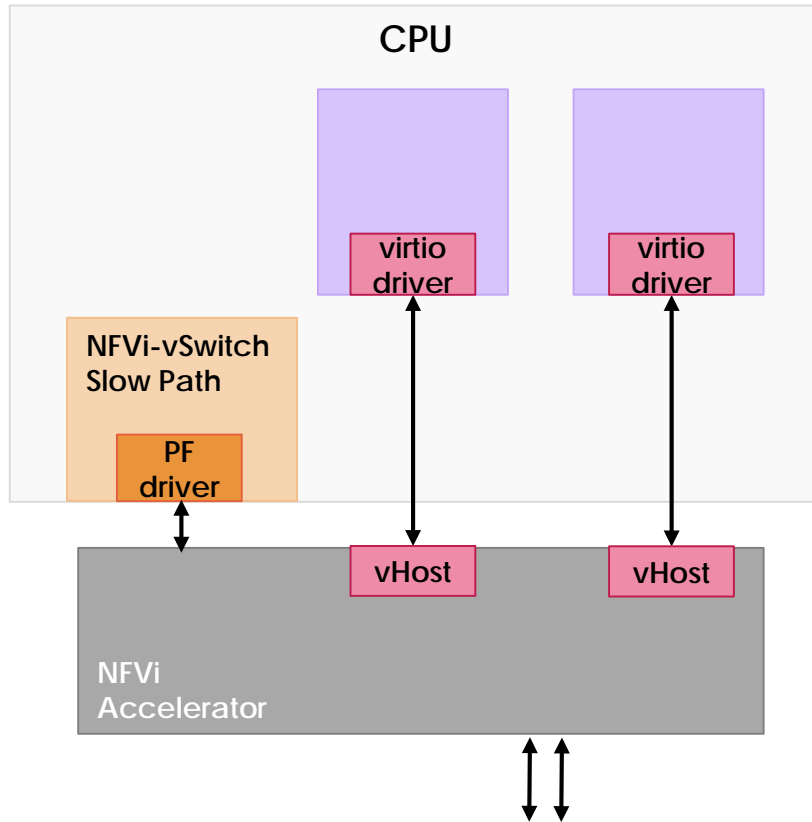
Agenda

- ▶ Problems towards NFV Cloud
- ▶ New Model of Direct I/O
- ▶ vHost Data Path Acceleration
 - ▶ Under the Hood
 - ▶ DPDK High Level Design
 - ▶ HW Prerequisites
 - ▶ Live-migration for Stock VM
- ▶ Remaining Challenge
- ▶ Status & WIP
- ▶ Key Takeaway





Problems towards NFV Cloud

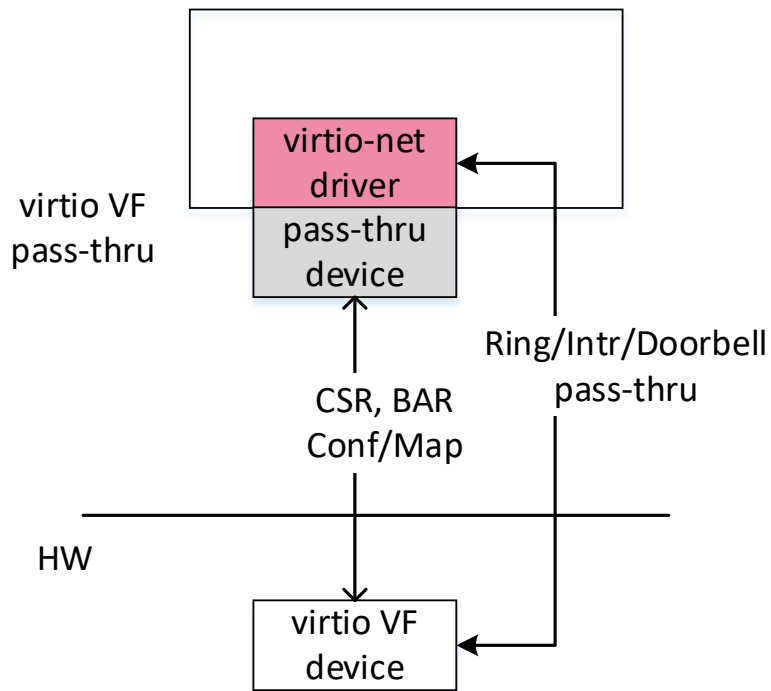


- ▶ vswitch/virtio is **well recognized** by cloud networking
- ▶ **Accelerator** is used to address **higher performance**
- ▶ SR-IOV **device pass-thru** represents for **fast I/O**
- ▶ Device specific **VF lacks** a few **cloud characteristics**
- ▶ Zero-copy buffer swap costs **unpredictable # of CPU**
- ▶ Other **direct I/O** approach besides device pass-thru?
- ▶ Para-virtualized device **w/ HW acceleration**, how?

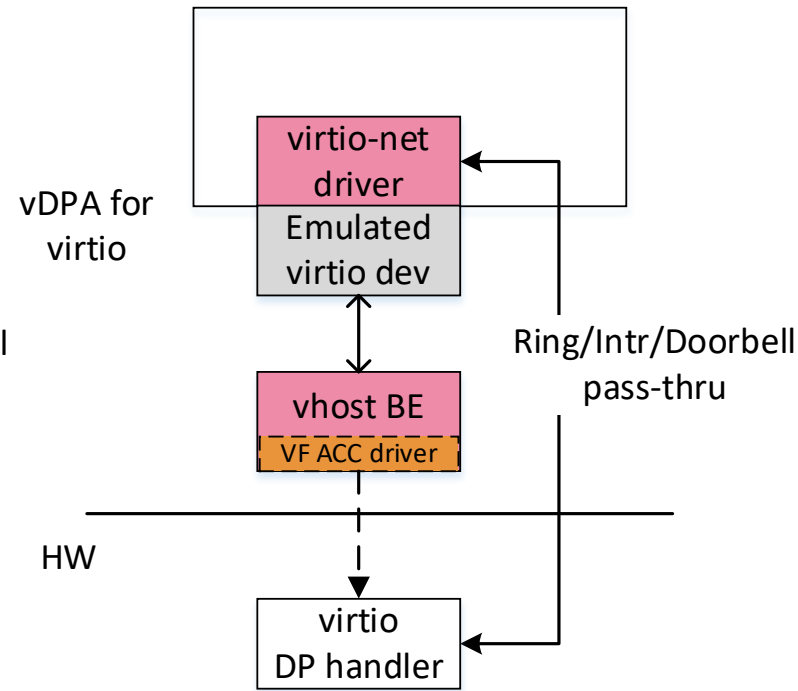
Unspecific Accelerator
 SR-IOV Like Performance
 Friendly Live-migration
 Stock VMs Support



New Model of Direct I/O



VIRTIO Device Pass-thru



vHOST Data Path Acc.

▶ **Key Objective**

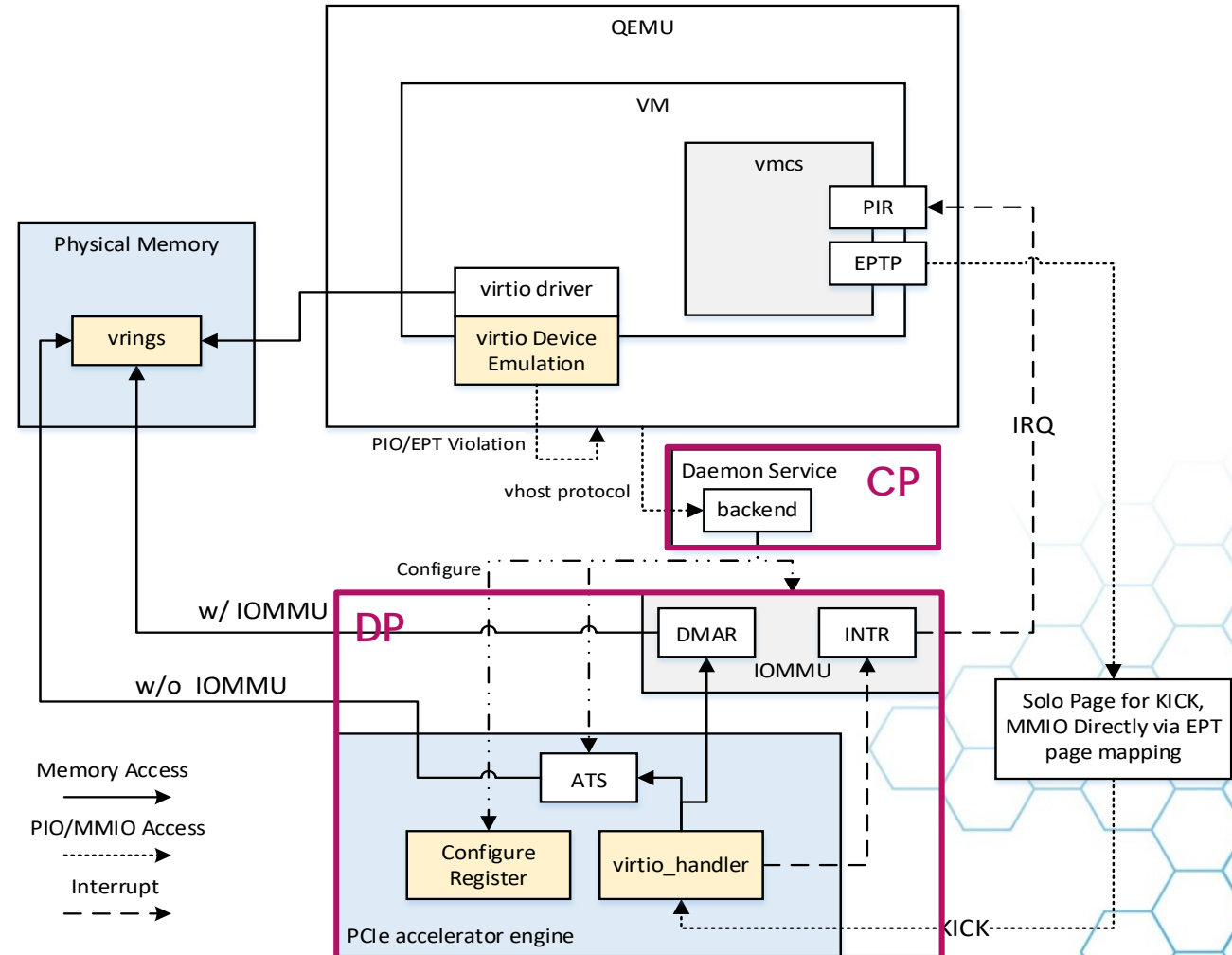
- Follow Spec.
- SR-IOV like performance
- Friendly Live-migration Support
- Support stock VMs
- ▶ **Good-enough** pass-thru
- ▶ **Para-virtualized** device w/ accelerator
- ▶ DPDK will **support both** model
- ▶ 2017' Q2 Prototype Finished





vDPA: Under the Hood

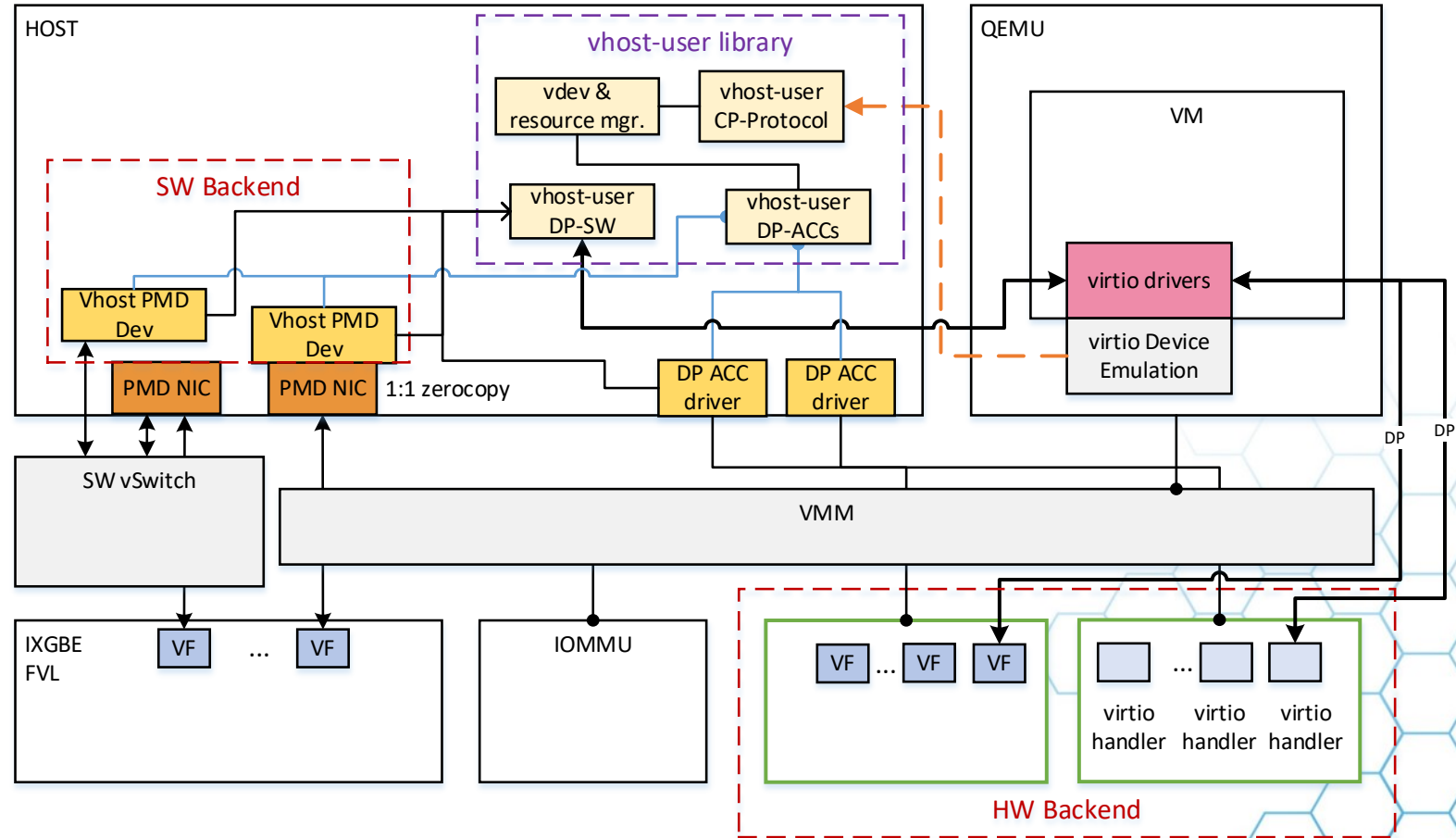
- ▶ Device emulated by QEMU
- ▶ Decompose DP/CP on Backend
 - ▶ DP: DMA, INTERRUPT, DOORBELL
 - ▶ CP: vhost Protocol, DP configure
- ▶ IOVA Translation by IOMMU/ATS
- ▶ PI/EPT Mapping for INT/DOORBELL
- ▶ Selective DP Acceleration Engine
- ▶ Available SW DP Fallback
- ▶ Compatible Live-Migration
- ▶ Minimum HW Prerequisites





vDPA: DPDK High Level Design

- ▶ DPDK **vhost-user** library
 - ▶ **CP-Protocol**, communicate channel with QEMU
 - ▶ **vdev Mgr.**, virtual device and resource management
 - ▶ **DP-ACCs**, vhost data path abstraction layer
 - ▶ **DP-SW**: SW vhost data path
- ▶ DP-ACC engine providers drive the accelerators which can be either **PCIe based** or **non-PCIe based**
- ▶ PMD and Port Representor Driver of DP-ACC can **leverage DP-SW** library to build **SW vhost data path**





vDPA: HW Prerequisites

- ▶ Ring Layout Follows the virtio Spec. (**MUST**)
- ▶ Ring Feature Capability Awareness (**MUST**)
- ▶ R/W vring index status (**MUST**)
 - ▶ BAR configure register: R/W 16bits index register (last_used_idx) per vring
 - ▶ last_used_idx is the HW internal status of used vring
- ▶ Log dirty pages (**MUST**, note: will be addressed by Vt-d)
 - ▶ BAR configure register
 - ▶ 64bits register for log memory base address
 - ▶ 64bits register for log memory size
 - ▶ 1bit register to enable logging
- ▶ Kick RARP: w/ VIRTIO_NET_F_GUEST_ANNOUNCE, no need for HW to trigger the RARP

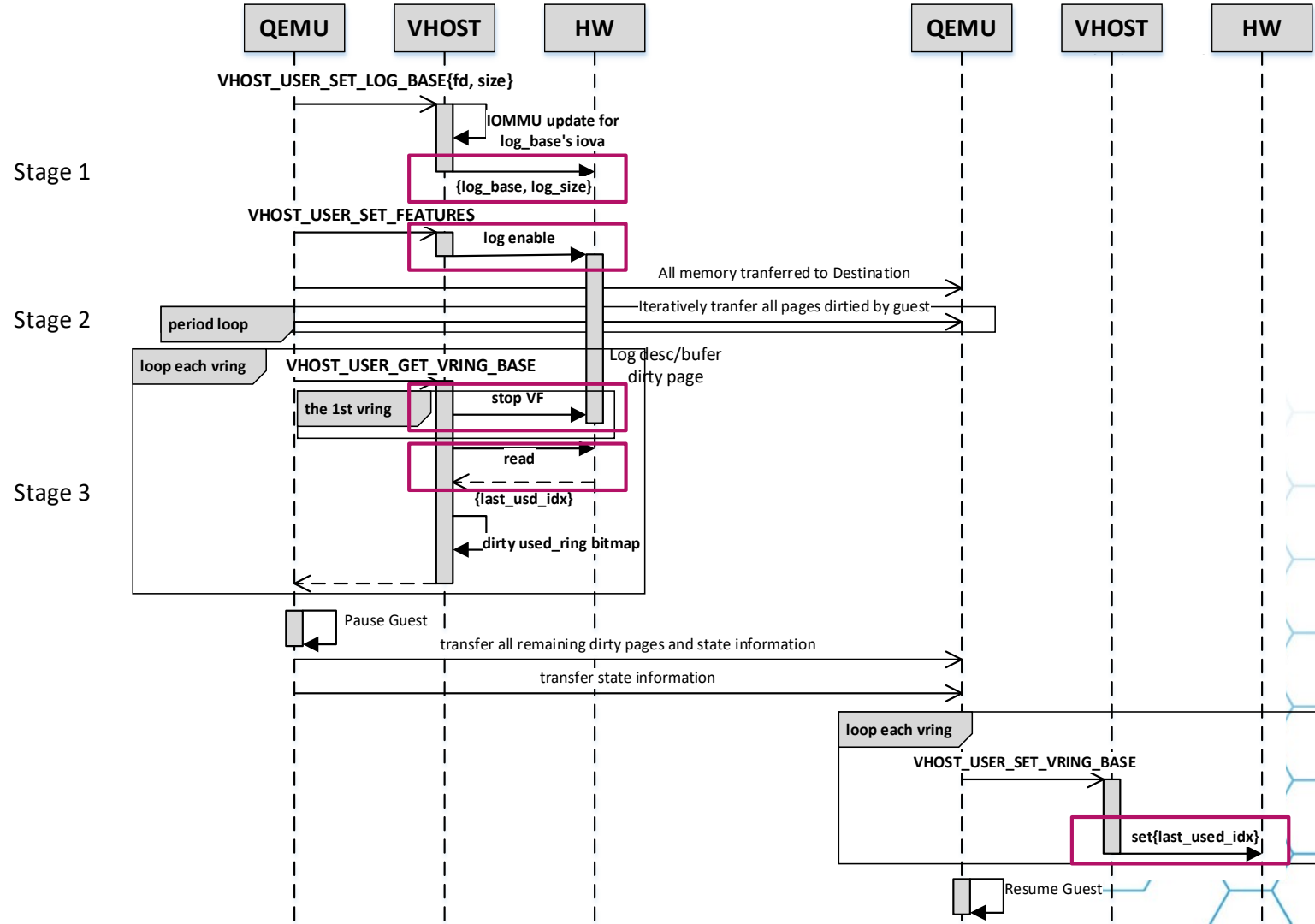




vDPA: Live-Migration Support *Source*

Destination

- ▶ **Compatible** with SW backend
 - ▶ Dirty Page Logging
 - ▶ VRING state report/restore
 - ▶ Kick RARP (alternative)
- ▶ Be possible to **transparently** upgrade/live-migrate **stock VM** to a new **platform w/ accelerator** in the backend
- ▶ **Challenge** remains for **bus overhead** of **small size transaction** for the dirty page logging



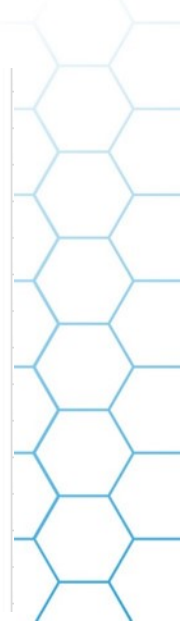
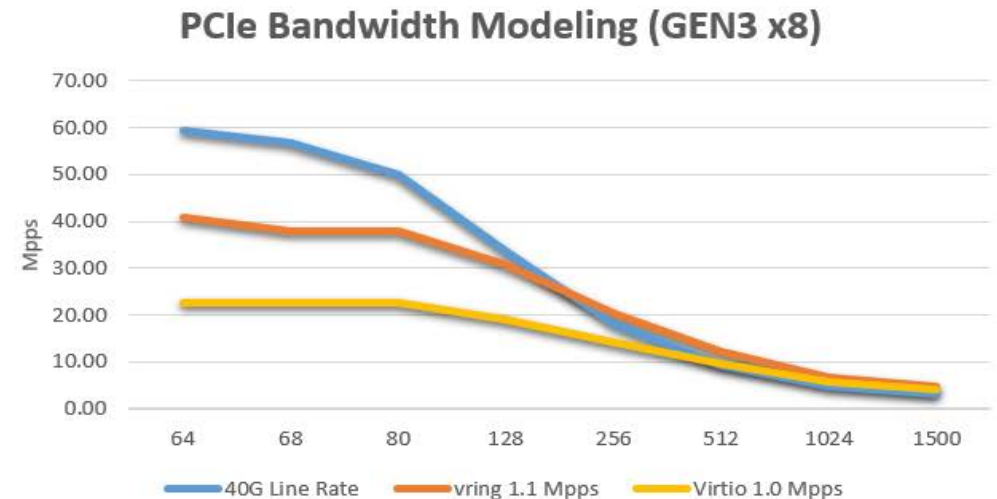
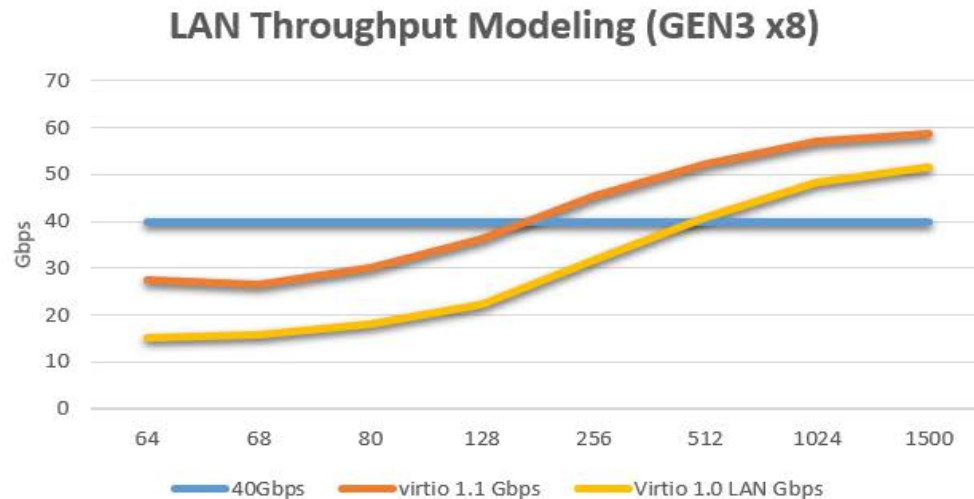


Remaining Challenges: Bus Overhead

- ▶ Reducing bus overhead for Logging dirty page
 - ▶ PCIe based: **coarse-grained** logging
 - ▶ Ideally logging the Dirty bits in IOMMU (long term)
 - ▶ It' **not a problem** for **memory based** accelerator
- ▶ Reducing bus overhead for Ring manipulation
 - ▶ VIRTIO v1.1 New Ring Layout [\[1\]](#)[\[2\]](#)
 - ▶ Simple modeling shows **lower bus overhead**

Not in Perfect Stage, but **manageable** !

[1]: <https://lists.oasis-open.org/archives/virtio-dev/201702/msg00010.html>
 [2]: <https://lists.oasis-open.org/archives/virtio-dev/201702/msg00035.html>





Status & Working in Progress

- ▶ 2017 Q1~Q2 PoC [DONE]
- ▶ 2017 Q2 shared in DPDK Monthly Virtio Community Call [DONE]
- ▶ 2017'Q2 Finish v1.1 experimental prototype in DPDK ^[1] [DONE]
- ▶ 2017 Q3 Feedback Collection from Early Trial [WIP]
- ▶ 2017 Q3/Q4 v1.1 ring layout optimization, proposal, PoC [WIP]
- ▶ 17.08/17.11 DPDK vDPA framework RFC patch [WIP]
- ▶ 17'Q4 QEMU patch for virtio direct I/O support [WIP]
 - ▶ INTR/Doorbell Mapping
- ▶ 17'Q4 Kernel RFC patch for vDPA

Para-virtualized device w/ HW acceleration is coming.
Welcome on board!



Acknowledgement

- ▶ Zhihong Wang
- ▶ Tiwei Bie
- ▶ Jianfeng Tan
- ▶ Heqing Zhu
- ▶ Yuanhan Liu
- ▶ Amnon Ilan
- ▶ Franck Baudin
- ▶ Martin Roberts
- ▶ Dan Daly
- ▶ Gerald Rogers
- ▶ Roger Chien





Key Takeaway

- ▶ What is vDPA? -- **v**Host **D**ata **P**ath **A**cceleration
- ▶ New approach of Direct I/O: **small granularity** data path **pass-thru**
- ▶ Target to next-gen **para-virtualized** device **w/ accelerator**
- ▶ Key benefits
 - ▶ 'SR-IOV' like performance w/ compatible live-migration support
 - ▶ Transparently upgrade stock VM to enhanced platform w/ very small set of HW prerequisites
- ▶ Remaining **challenges** are **manageable**
- ▶ Welcome for any feedback/contribution





Thanks!!



Contacts:
cunming.liang@intel.com
zhihong.wang@intel.com

