DPDK
DATA PLANE DEVELOPMENT KIT

THE LINUX FOUNDATION

DPDK SUMMIT CHINA 2017

主办方： intel

参与方： 腾讯云 ZTE 美团云 MEITUAN OPEN SERVICES Panabit® 太一星晨 Balance Your Networks UnitedStack有云 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 知识分享平台

DPDK DATA PLANE DEVELOPMENT KIT

THE LINUX FOUNDATION

OVS-DPDK Practices in Meituan cloud

Huai, Huang

DPDK China Summit 2017，Shanghai

主办方： intel

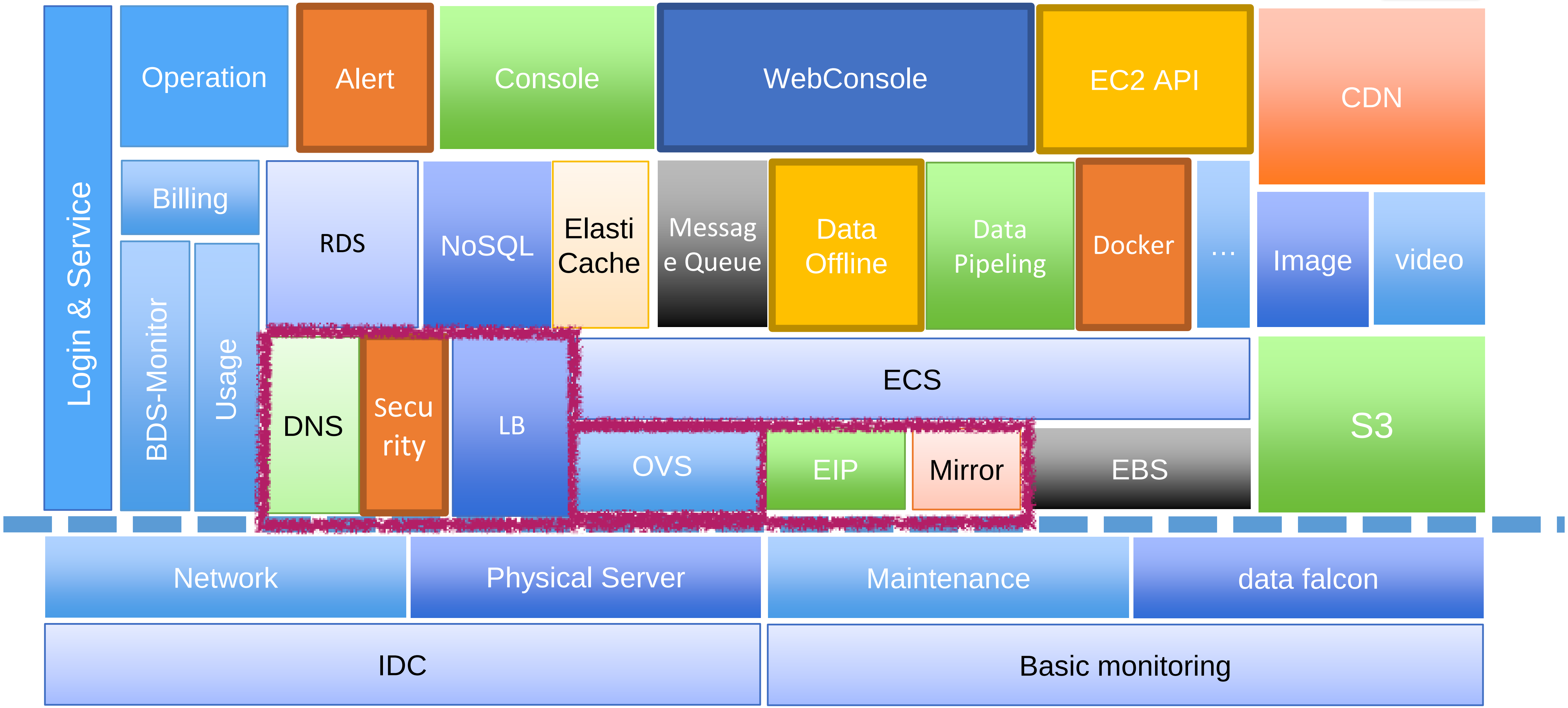
参与方： 腾讯云 ZTE 美团云 Panabit 太一星晨 UnitedStack 云杉网络

协办方： SDNLAB 视频支持方： IT大咖说

MOS:experiences large-scale business practices

A public cloud provider based on the world's largest O2O platform

Cloud Platform Architecture



Login & Service

Operation

Alert

Console

WebConsole

EC2 API

CDN

Billing

BDS-Monitor

Usage

RDS

NoSQL

Elasti Cache

Messag e Queue

Data Offline

Data Pipeling

Docker

…

Image

video

DNS

Secu rity

LB

ECS

S3

OVS

EIP

Mirror

EBS

Network

Physical Server

Maintenance

data falcon

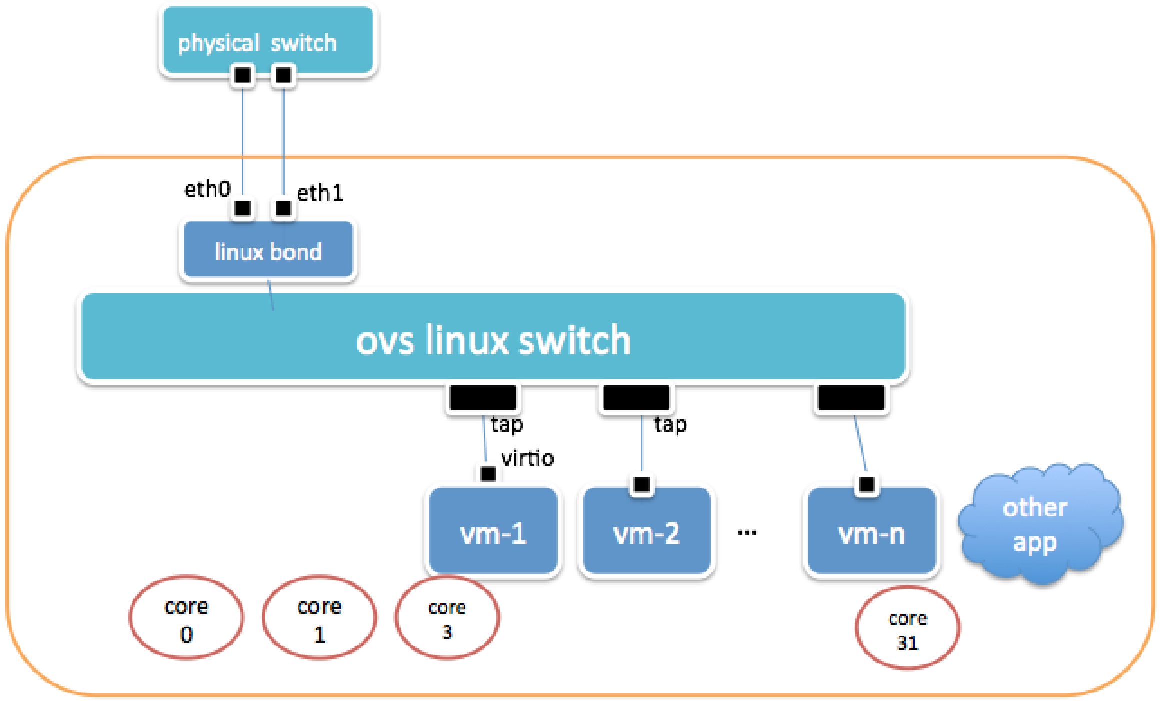
IDC

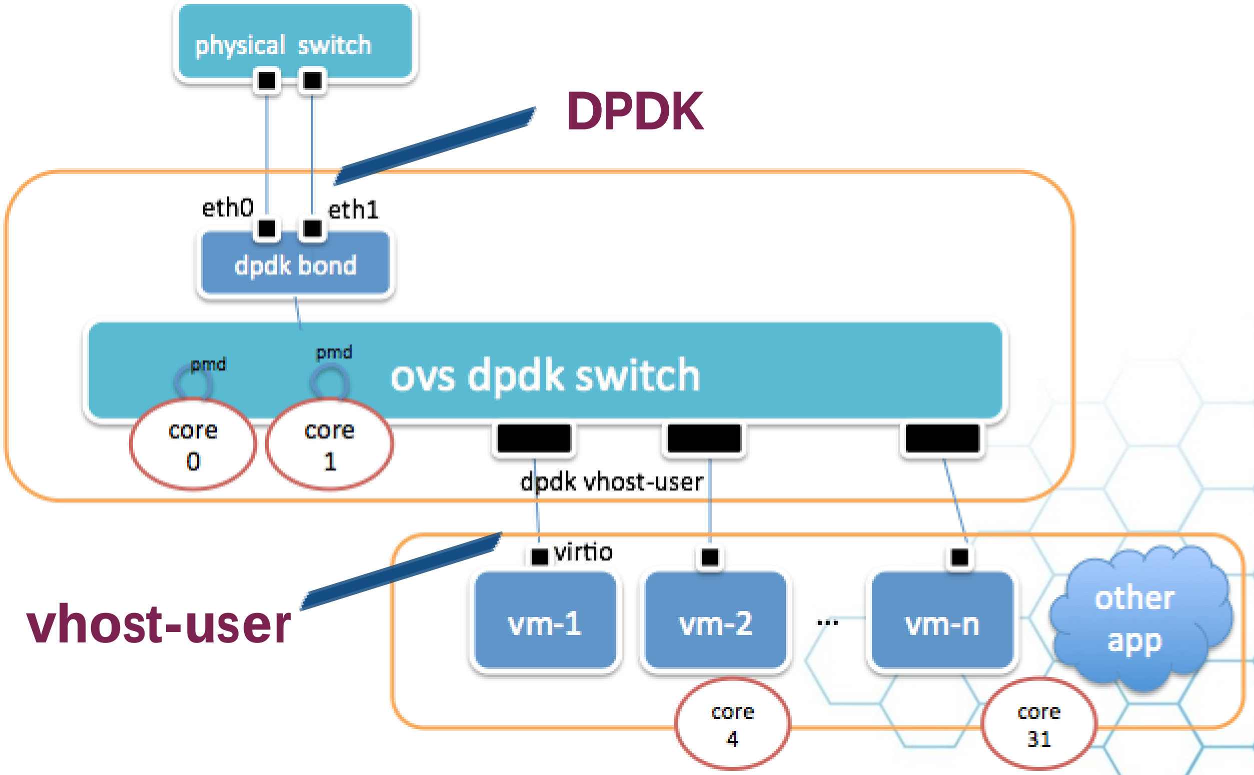
Basic monitoring

CONTETS

- Introducing OVS-DPDK
- Performance
- Our works
- Further works

OVS & OVS-DPDK


physical switch
eth0
eth1
linux bond
ovs linux switch
tap
virtio
tap
vm-1
vm-2
...
vm-n
other app
core 0
core 1
core 3
core 31

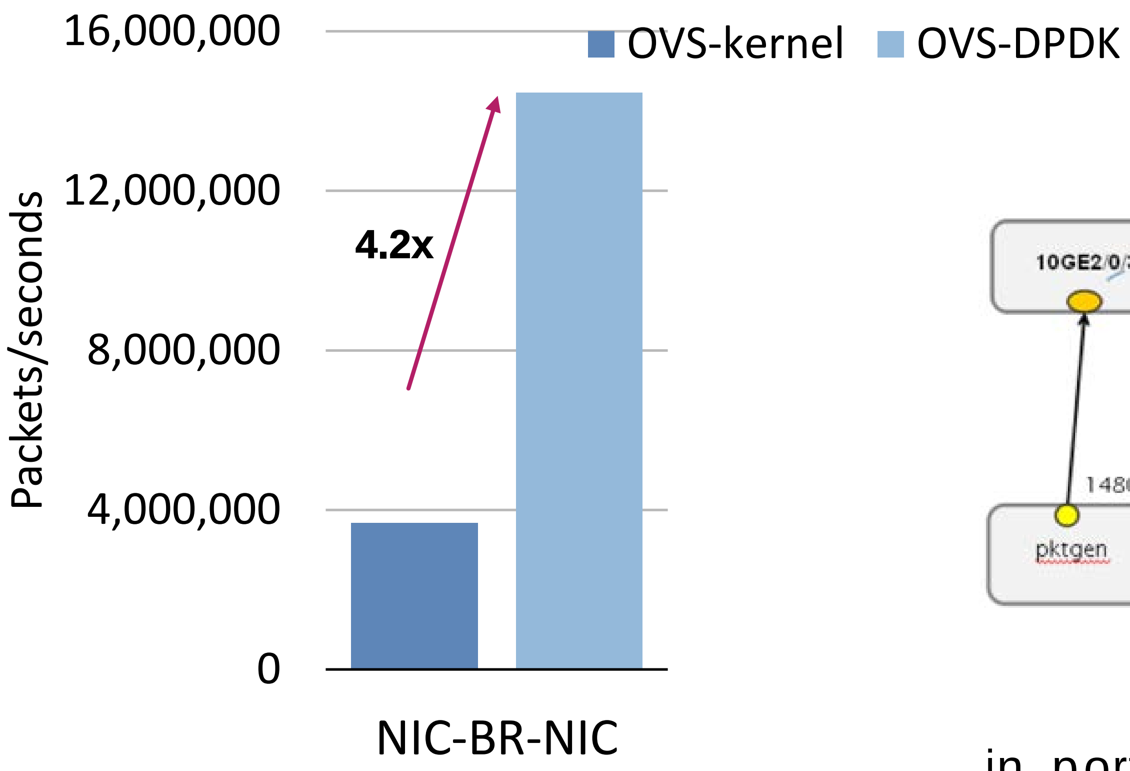

physical switch
DPDK
eth0
eth1
dpdk bond
pmd
pmd
ovs dpdk switch
core 0
core 1
dpdk vhost-user
virtio
vhost-user
vm-1
vm-2
...
vm-n
other app
core 4
core 31

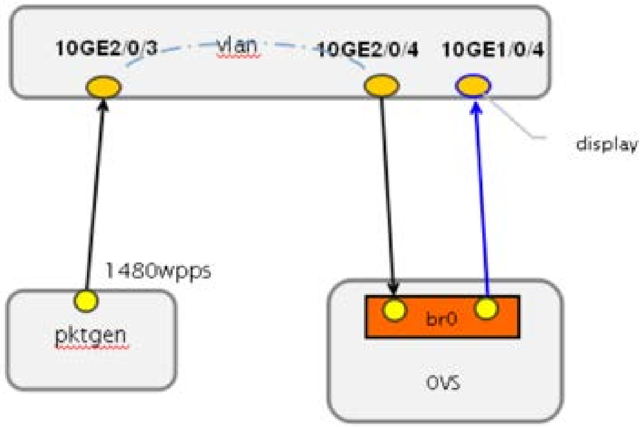
CONTETS

- OVS-DPDK
- Performance
- Our works
- Further works

NIC-BR-NIC

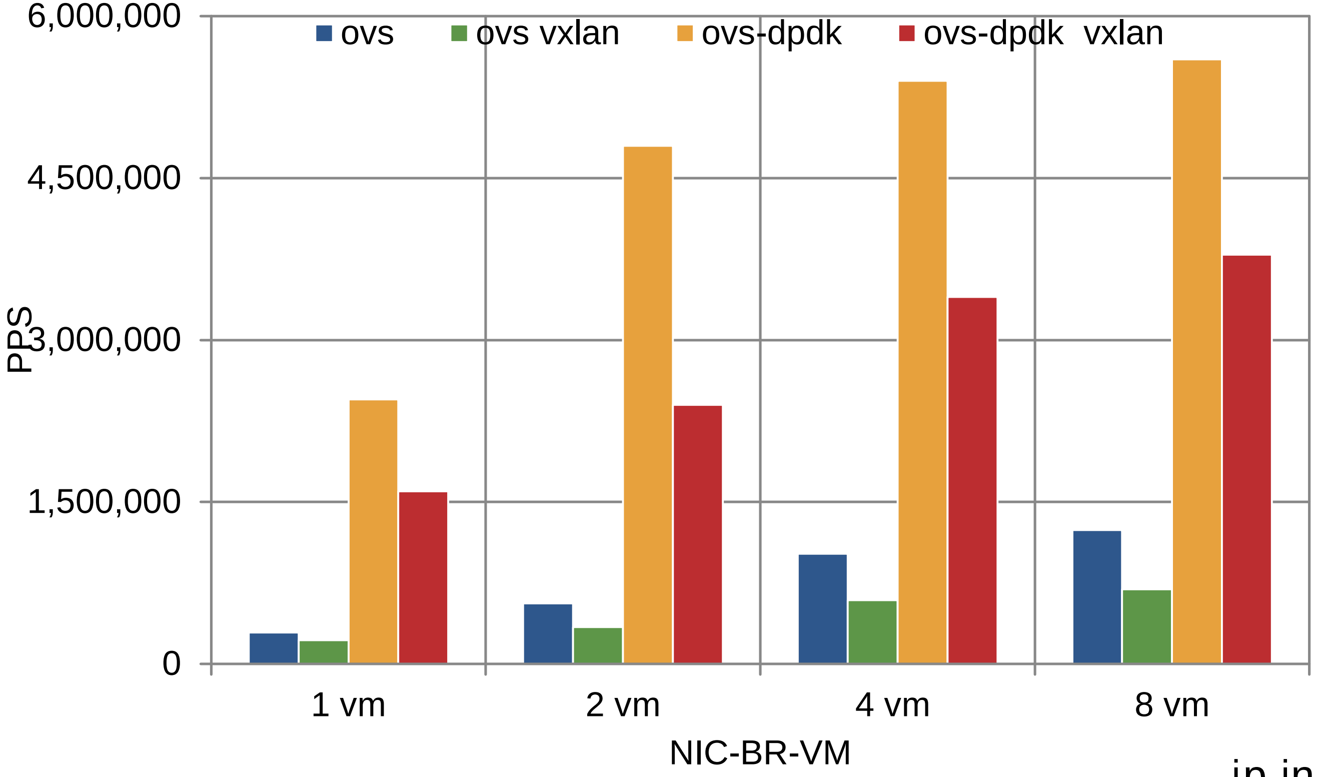
- E5-2650 v2 @ 2.6G
  - 1 core
- 128GB mem
- 10-Gigabit X540-AT2
- kernel 3.10.0
- Open vSwitch 2.4.90
- qemu 2.6.0
- 64bytes

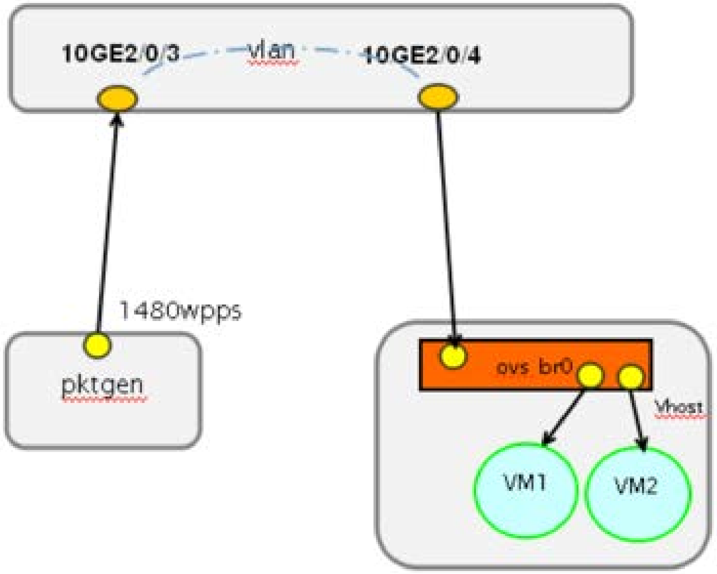




in_port=2,action=output:1

NIC-BR-VM





```
ip,in_port=2,nw_src=192.168.102.0/27 actions=output:3
ip,in_port=2,nw_src=192.168.102.32/27 actions=output:4
ip,in_port=2,nw_src=192.168.102.64/27 actions=output:5
ip,in_port=2,nw_src=192.168.102.96/27 actions=output:6
…
```

CONTETS

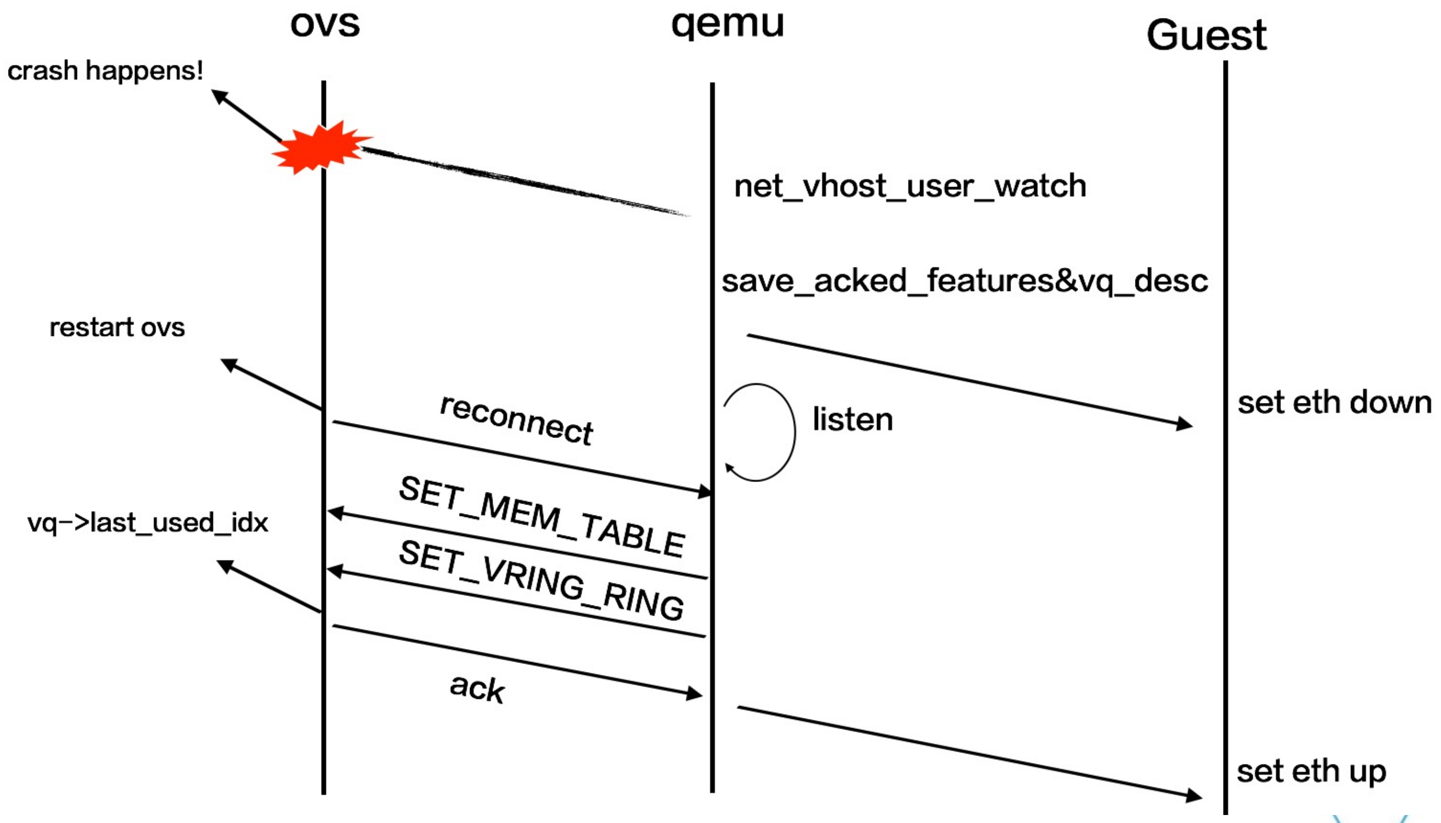
- OVS-DPDK
- Performance
- Our works
- Further

Question 1: stability

- stability problem
  - OVS-DPDK, DPDP, qemu
  - fix ovs, dpdk, qemu 16 bugs
- debug status
  - rte resource: rte_heap
  - vhost: virtqueue, fd_set, reconnect_list, virtio_dev…
  - NIC: link status, bond status…

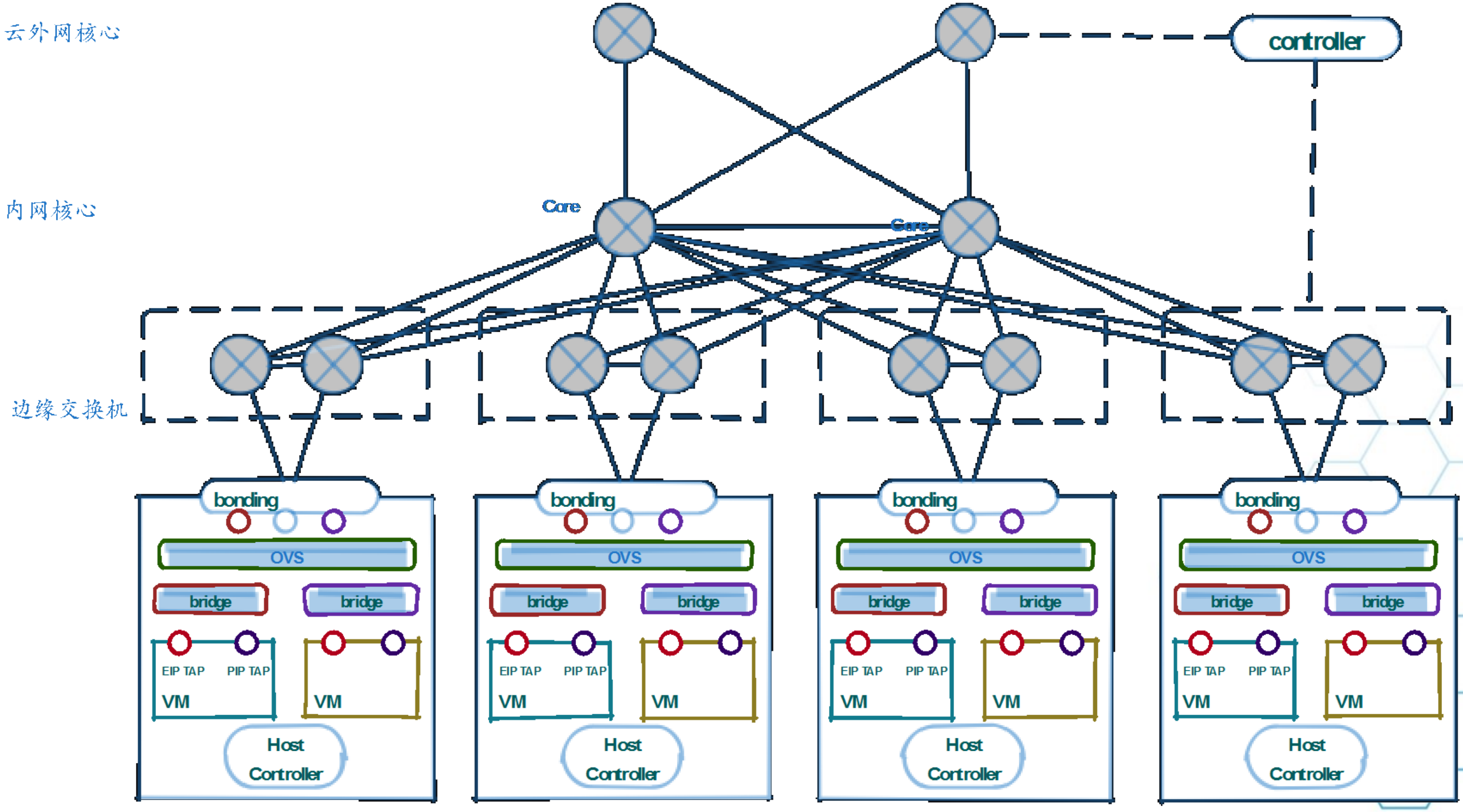
Question 2: Function Defect

- raise
  - vhost recovery
  - live migration
  - windows vm run
- finished by Intel



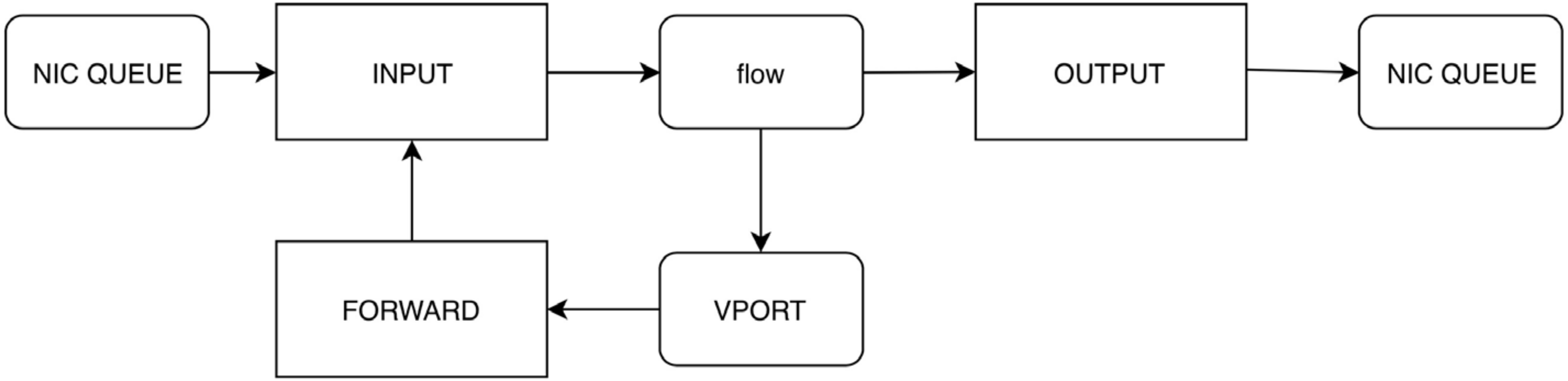
Problem 3: Compatibility

- DPDK-BOND port
- dump
- QoS



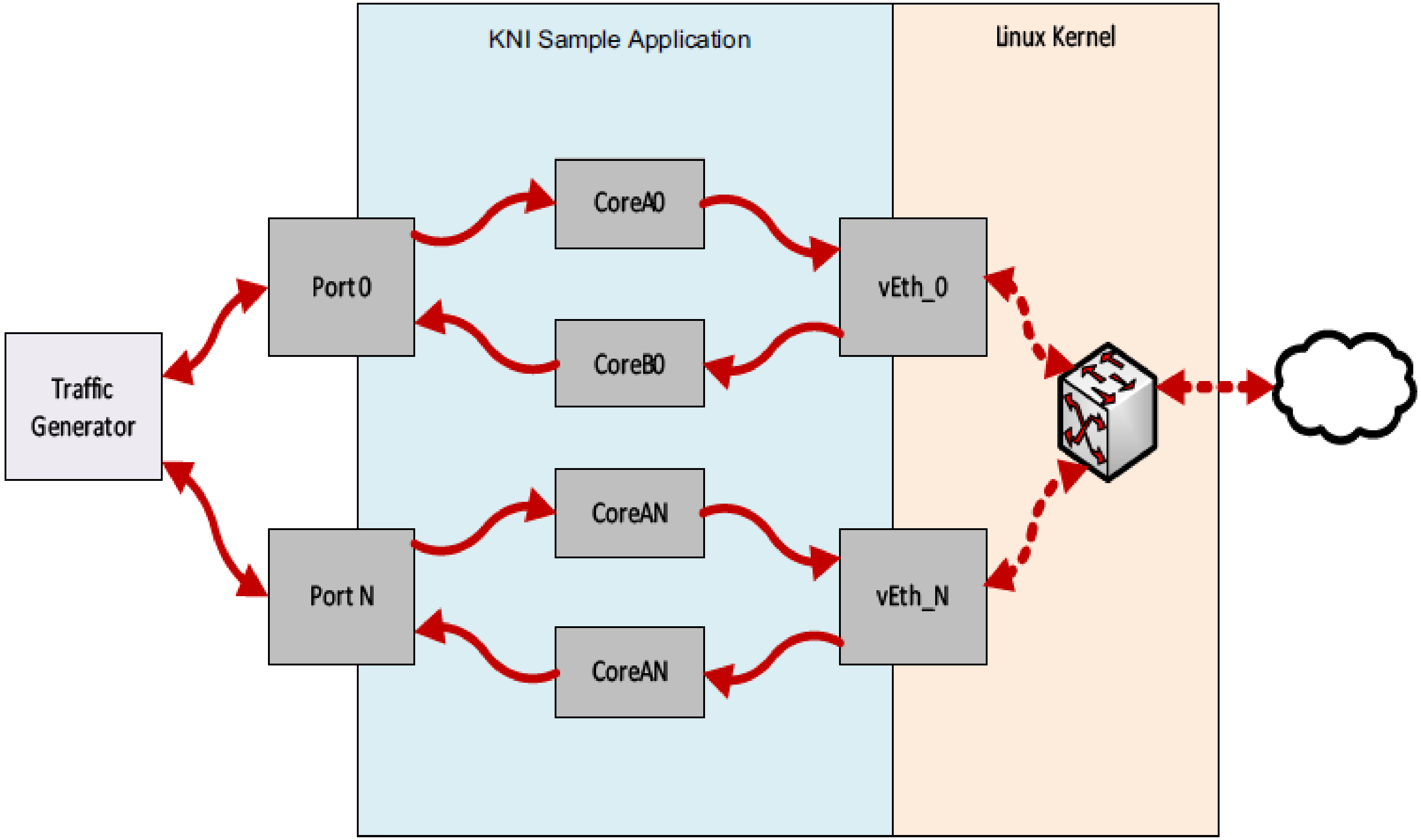
Problem 3: Compatibility

- dump
- ovs-filter framework
  - filter hook
  - load/unload module
- kernel module transplant
  - L4 session



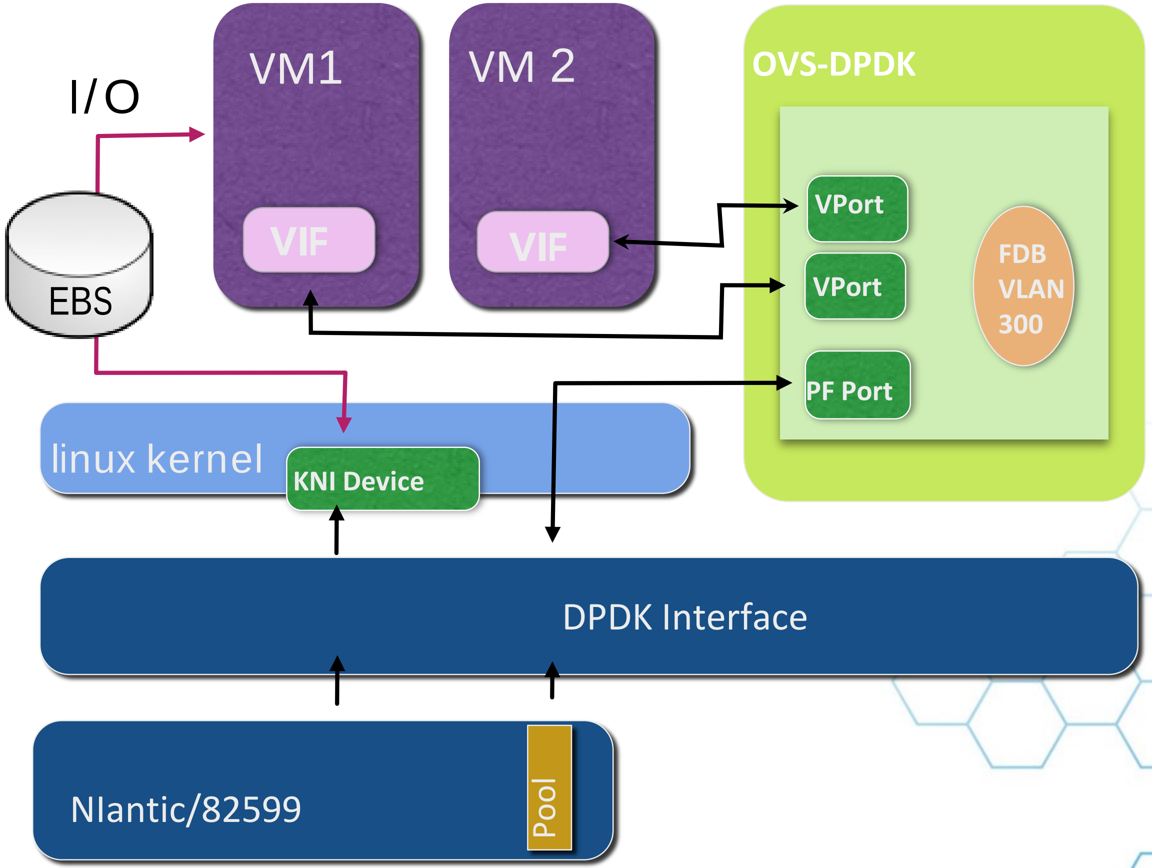
Problem 4: LOCAL capability

- DPDK current method
  - TAP
  - KNI
  - 2-3Gbps big packets



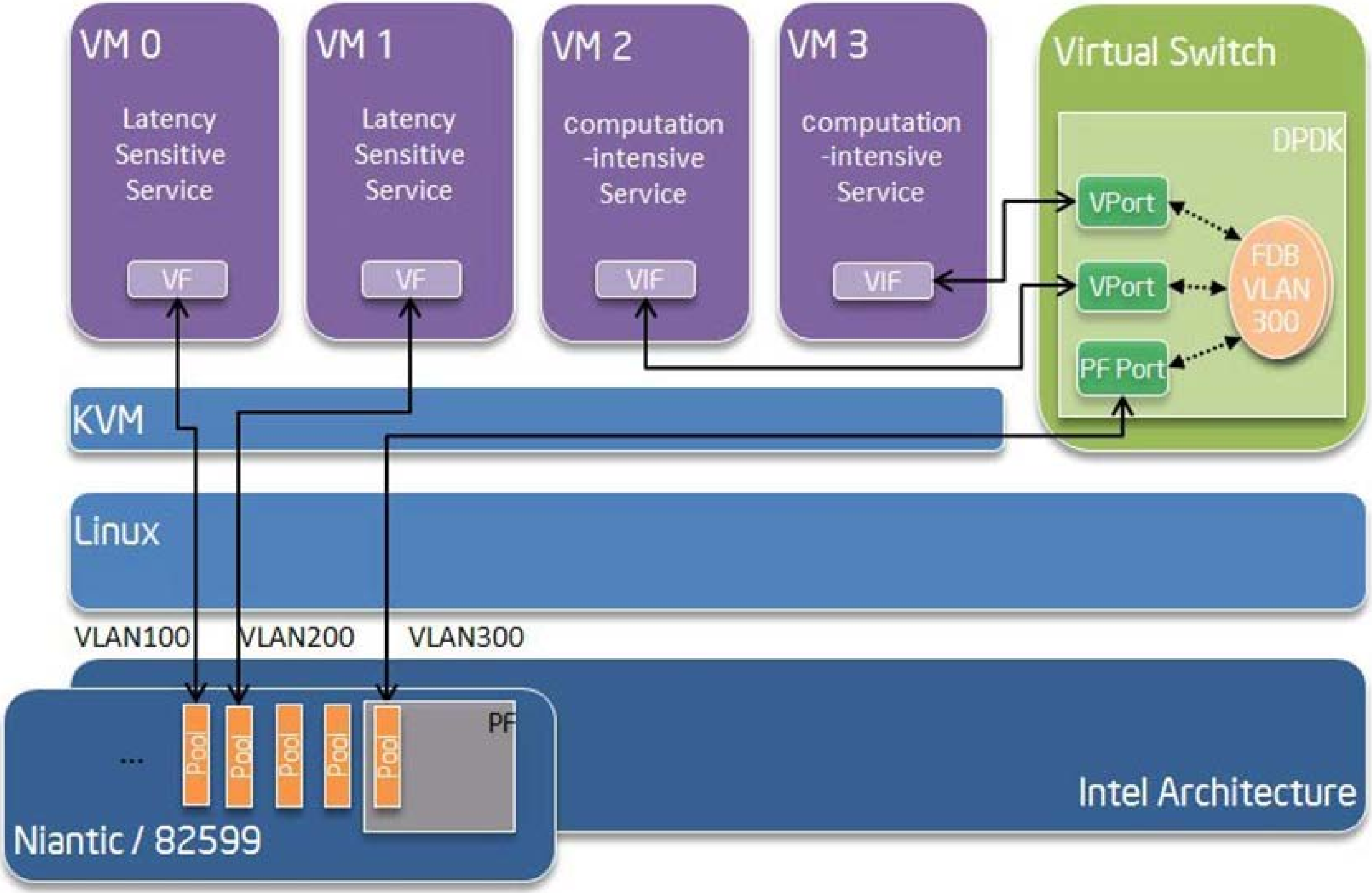
Problem 4: LOCAL capability

- Local port
  - forward bottlenecks
- Case.
  - Distributed file system（EBS）
  - I/O -> network
  - Offline calculation，Data collection



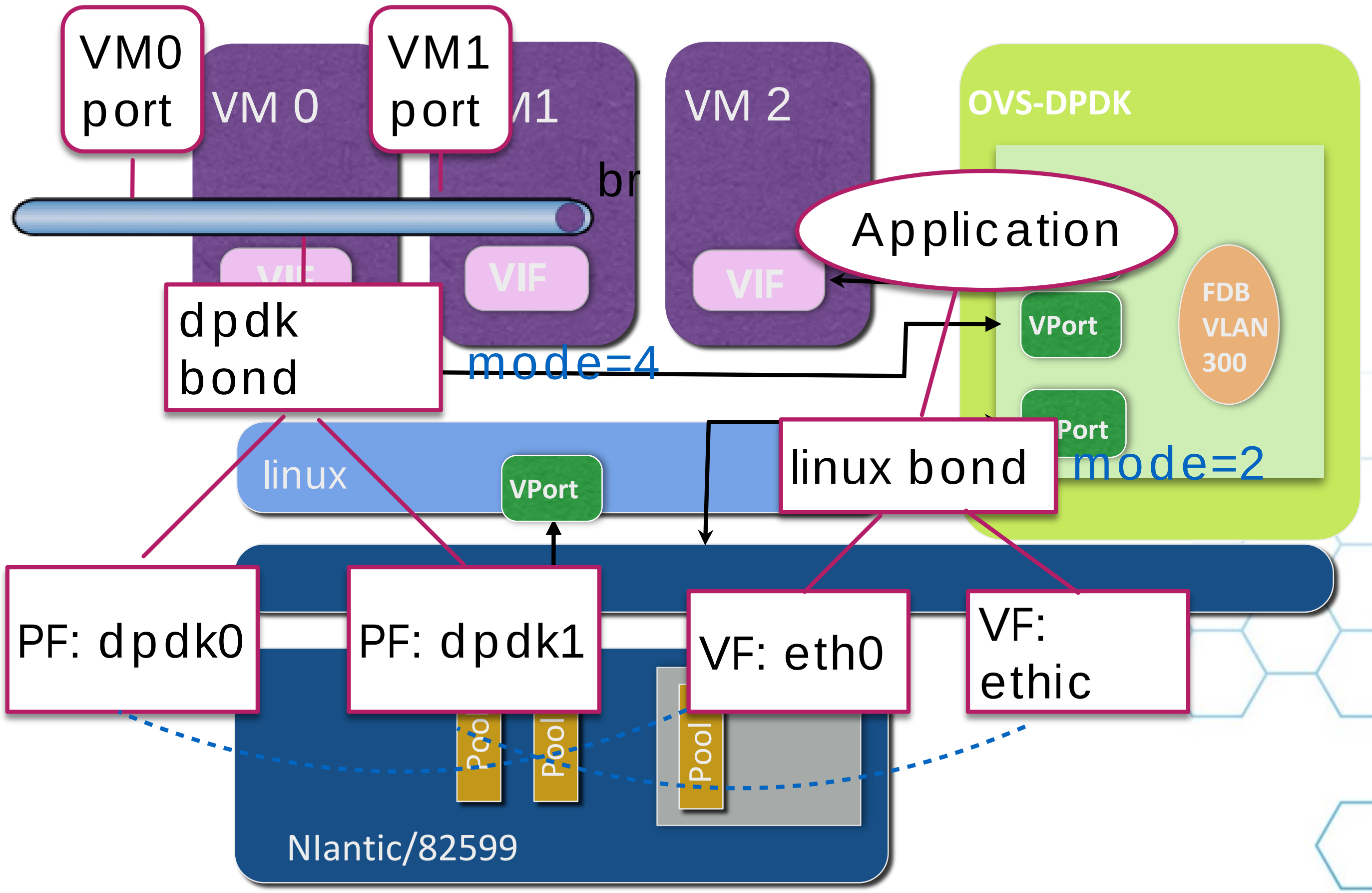
SR-IOV Approach

- SRIOV
  - PF：DPDK NIC
  - VF ：VM
  - eswitch
- Switch
  - controll
  - user isolation
  - live migration



LOCAL capability

- SR-IOV method
  - linux PF + DPDK VF
  - DPDK PF + linux VF
- CHANGES
  - DPDK VF code
  - ixgbevf code
  - PF、VF queue set
  - bond4



CONTETS

- OVS-DPDK
- Performance
- Our works
- Further

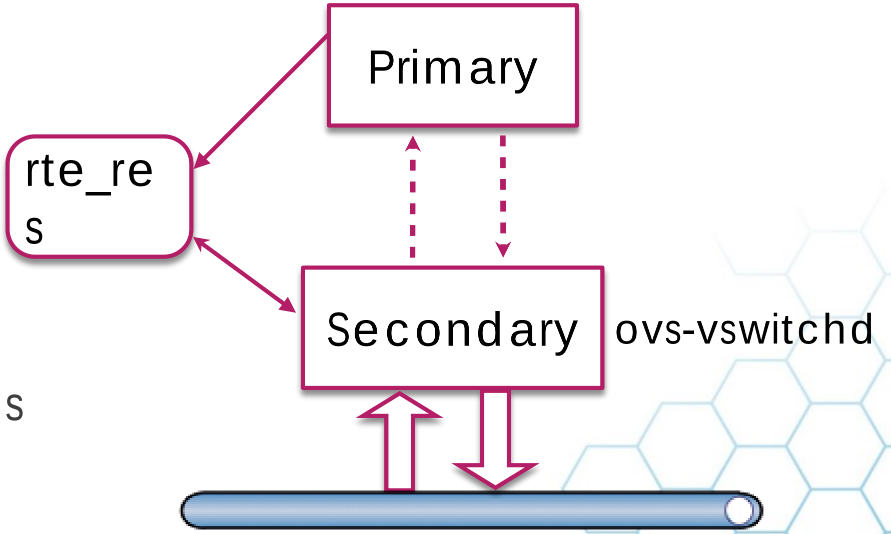
Problem 5: OVS -> OVS-DPDK

- old OVS update OVS-DPDK
  - Method 1：native update
  - Method 2: offline migration
  - question：VM need shutdown
- live migration
  - vhost-kernel vm -> vhost-user vm

Problem 6: Hot Update

- Question
  - ovs-vswtichd restart 2-3min
- last time section
  - hugepage init
  - nic port restart
    - link auto negotiation
    - bond4 handshake
  - vhost-user port reinit

- step 1:
  - 2+min -> 10- s
- step 2:
  - 10- s -> 2- s



Solution.

Thanks!!

huanghuai@meituan.com

欢迎关注**DPDK开源社区**