DPDK在NFV中的应用

2018.3

NFV是网络功能虚拟化的缩写，是IT虚拟化技术和云计算技术在[illegible]信领域的应用

传统网络设备

传统网络设备，如路由器，防火墙，BRAS

强依赖于专用硬件设备
升级成本高
新业务部署慢

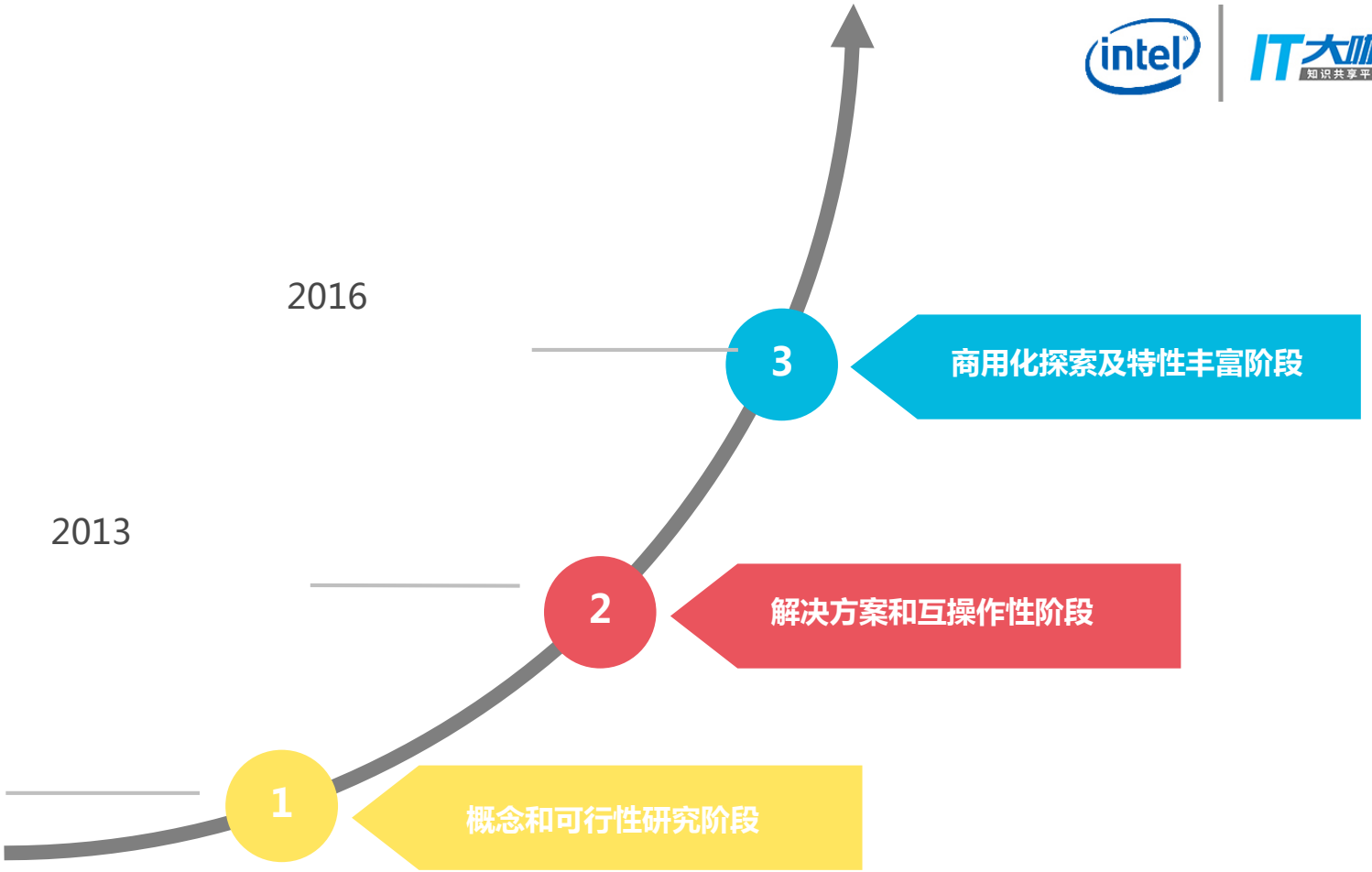
NFV-网络设备虚拟化

NFV设备，用纯软件实现网络设备功能，并运行在标准硬件服务器上。

不依赖于专用硬件设备
升级容易
新业务部署快
云化部署，可编程

H3C支持的NFV产品：VSR，vBRAS，vFW，vLB

2012年10月，借在德国召开的SDN及OpenFlow世界大会之机，13家网络运营商聚集在一起，首次发布了NFV的介绍性白皮书，第一次正式提出了NFV的构想


2016
3
商用化探索及特性丰富阶段
2013
2
解决方案和互操作性阶段
2012
1
概念和可行性研究阶段

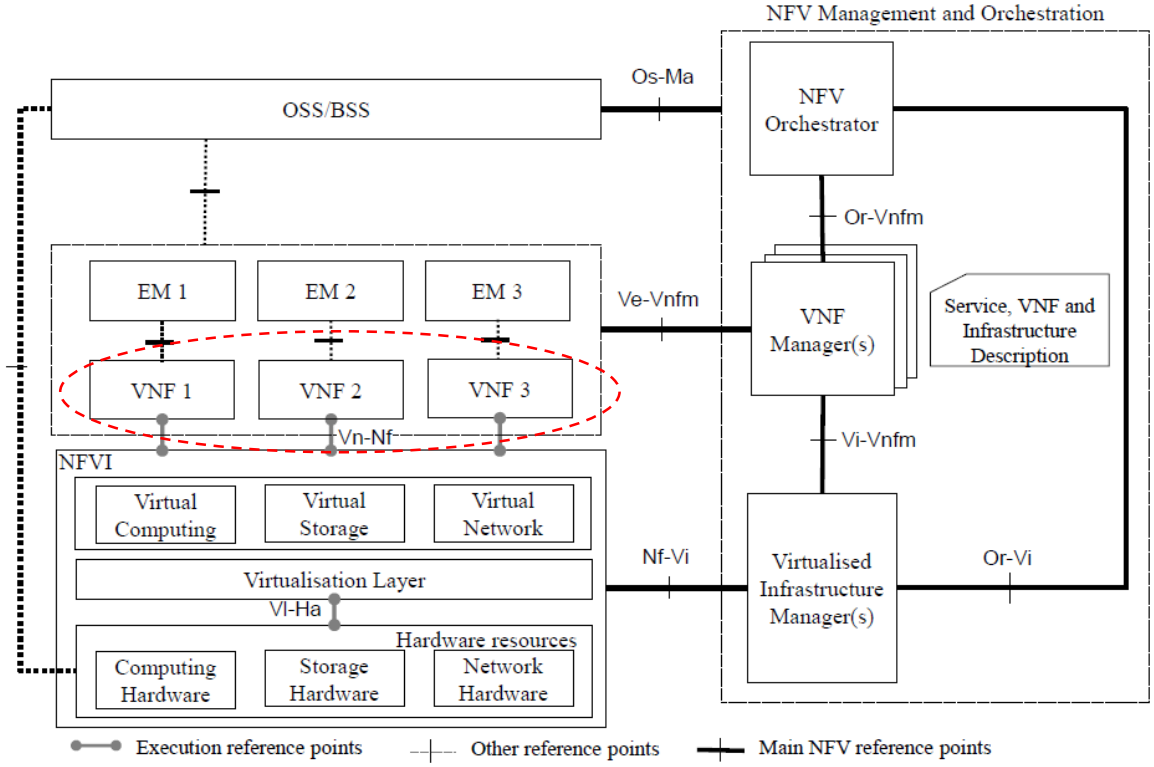
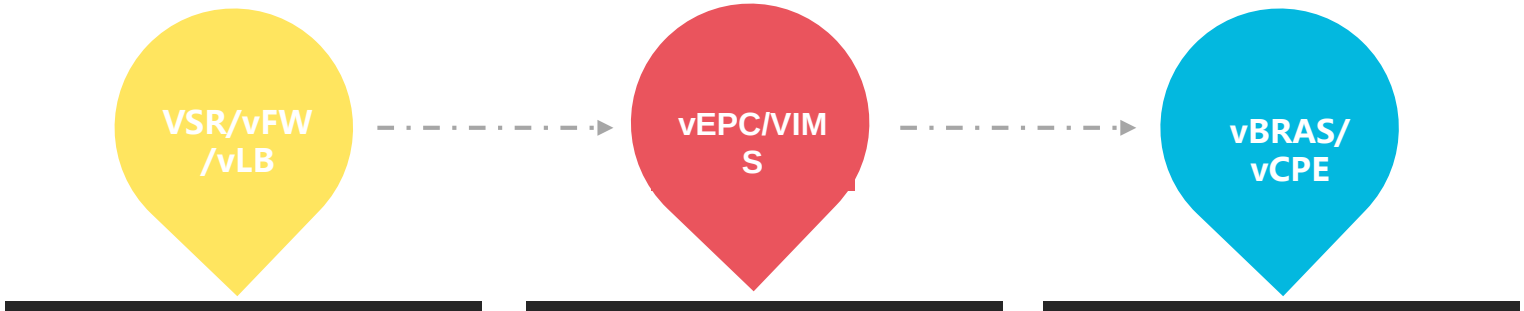


Figure 4: NFV reference architectural framework

NFV提出的这五年多时间里，可以看到NFV在很多地方得到了应用



NFV在互联网/云上

IT云的部署中几乎都会用到NFV，作为租户网关设备，作为防火墙

NFV在5G上

5G的部署中NFV也是重要的组成部分，要考虑物联网的时延要求

NFV在运营商接入上

运营商接入领域，要考虑原有硬件设备利旧，NFV的性价比

AT&T率先 提出Domain2.0的计划。计划2020年NFV商用达到75%，直到2017年底，AT&T宣称网络虚拟化已达55%

SKT从 2015年就开始部署了 vEPC和vIMS。2017年新部署的EPC有80%是vEPC。

三大运营商网络向NFV演进

中国联通发布CUBE-Net 2.0

中国联通在“2015中国国际信息通信展览会”对外发布了其新一代网络架构CUBE-Net 2.0白皮书，启动了“新一代网络”合作研发计划。其愿景是引入**SDN、NFV、云**等新技术，实现网络架构的重构，增强网络服务能力，降低网络运营成本，实现网络健康发展

中国移动发布NovoNet

中国移动表示，希望融合**NFV、SDN**等新技术，构建一张**资源可全局调度、能力可全面开放、容量可弹性伸缩、架构可灵活调整的新一代网络**，以适应中国移动数字化服务战略布局的发展需要，为互联网+发展奠定网络基础

中国电信发布CTNet2025

	近期：推进阶段	中远期：扩展阶段
网络云化	• 部分网元引入NFV • 推动CO向DC改造	• 统一全网云资源 • 实现网元DC化部署
新一代运营系统	• 引入SND控制器、网络协同与业务编排器 • 实施网络自动化配置	• 部署顶层网络协同和业务编排层 • 实现网络可编辑与按需调用

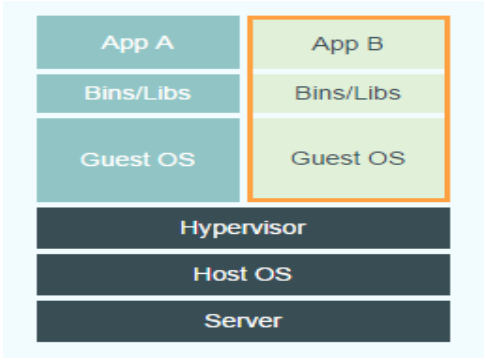
关键技术：依托**SDN+NFV+DC**，推进网络重构。
演进步骤：
近期(2016~2019)
中远期(2020~2025)

2017年
移动：制定了 关于NFV的转控分离规范，并对 多个厂家设备进行了三轮测试。
联通：部分省份进行了vBRAS商用试点，
电信：进行了多轮厂家NFV设备性能对比测试和解耦测试

从网络功能实现的角度，看这两种虚拟化技术，最大的影响就是用容器技术就必须在用户态实现网络功能。

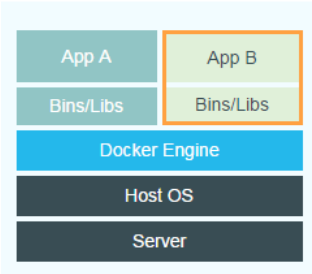
NFV虚拟化技术-VM

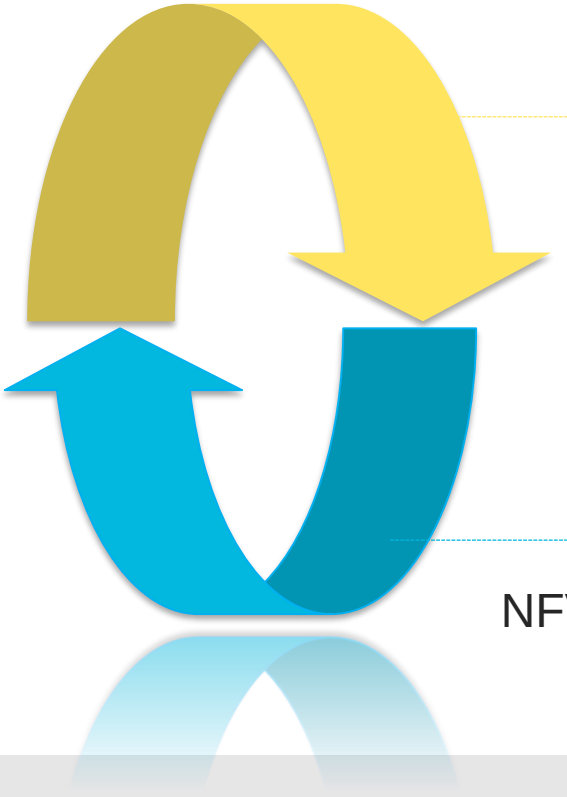
基于hypervisor
完全虚拟化
网络功能可以在内核，
也可以在用户态实现。



NFV虚拟化技术-Docker

基于Docker技术
操作系统级别的虚拟化
更轻量级，启动更快，占用资源更少
网络功能必须在用户态实现





- 关注点一：三层解耦

编排，控制，网元，虚拟化层能异厂家对接

- 关注点二： NFV的转发性能

低时延，高性能，和硬件转发相比

NFV转发性能的提升，成了NFV厂家的重要课题

DPDK的优化手段

无锁队列

大页表—减少TLBmiss

内存零拷贝

CPU编译指令优化

PMD调度模型

限制一：只支持用户态

限制二：不提供用户态协议栈，需要厂家自行开发

Test Setup -Cont.

DUT:

- Intel® Xeon® E5-2658 v4 processor,35MB L3 cache
- Super Micro* Platform (X10DRX)
- DDR4 2400 MHz, 4 x 1Rx4 registered ECC 16GB (total 64GB), 4 memory channels per socket Configuration, 1 DIMM per channel
- 1 x Intel X710-DA4-FH PCI-E Gen3X8 Quad Port Ethernet Controller (NVM: 5p04)
- 2 x Intel XL710-DA2 PCI-E Gen3x8 Dual Port 40GbE Ethernet Controller (NVM: 5p04)

IXIA* Traffic Parameters:

- Acceptable Frame Loss: 0.00001%
- Resolution: 0.1
- Traffic Duration: 20 Seconds

Software:

- BIOS version: Version: 2.0 & Date: 12/17/2015
- Operating system: Fedora 23
- Kernel version: 4.2.3-300.fc23.x86_64
- IxNetwork* : 7.40 EA
- DPDK version: 16.04
- DPDK L3fwd example application on Linux user space (LPM for route lookup)
  - .hw_ip_checksum = 0, /**< IP checksum offload enabled */
  - #define RTE_TEST_RX_DESC_DEFAULT 1024
  - #define RTE_TEST_TX_DESC_DEFAULT 1024

Polling Affinity for Ethernet Queues- 4x10G ports

- **2 ports – (1 Core/1 Thread /1Queue)**
  CPU1 (Core 1 SMT 0) polls port 0
  CPU1 (Core 1 SMT 0) polls port 1
  CPU1 (Core 1 SMT 0) polls port 2
  CPU1 (Core 1 SMT 0) polls port 3

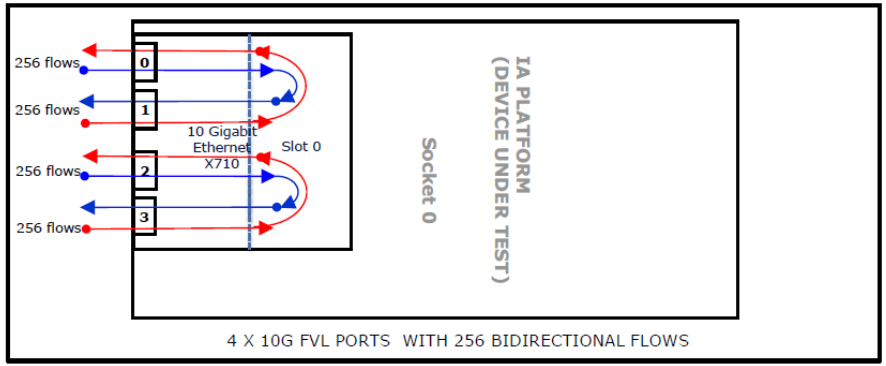
- **2 ports - (2 Core / 2 Threads/1 Queue)**
  CPU1 (Core 1 SMT 0) polls port 0
  CPU1 (Core 2 SMT 0) polls port 1
  CPU1 (Core 1 SMT 0) polls port 2
  CPU1 (Core 2 SMT 0) polls port 3

- **2 ports - (1 Core / 2 Threads/1 Queue)**
  CPU1 (Core 1 SMT 0) polls port 0
  CPU2 (Core 15 SMT 1) polls port 1
  CPU1 (Core 1 SMT 0) polls port 2
  CPU2 (Core 15 SMT 1) polls port 3

Each polling core has 100% CPU Utilization.
Remaining cores are IDLE

Flow Traffic Configuration

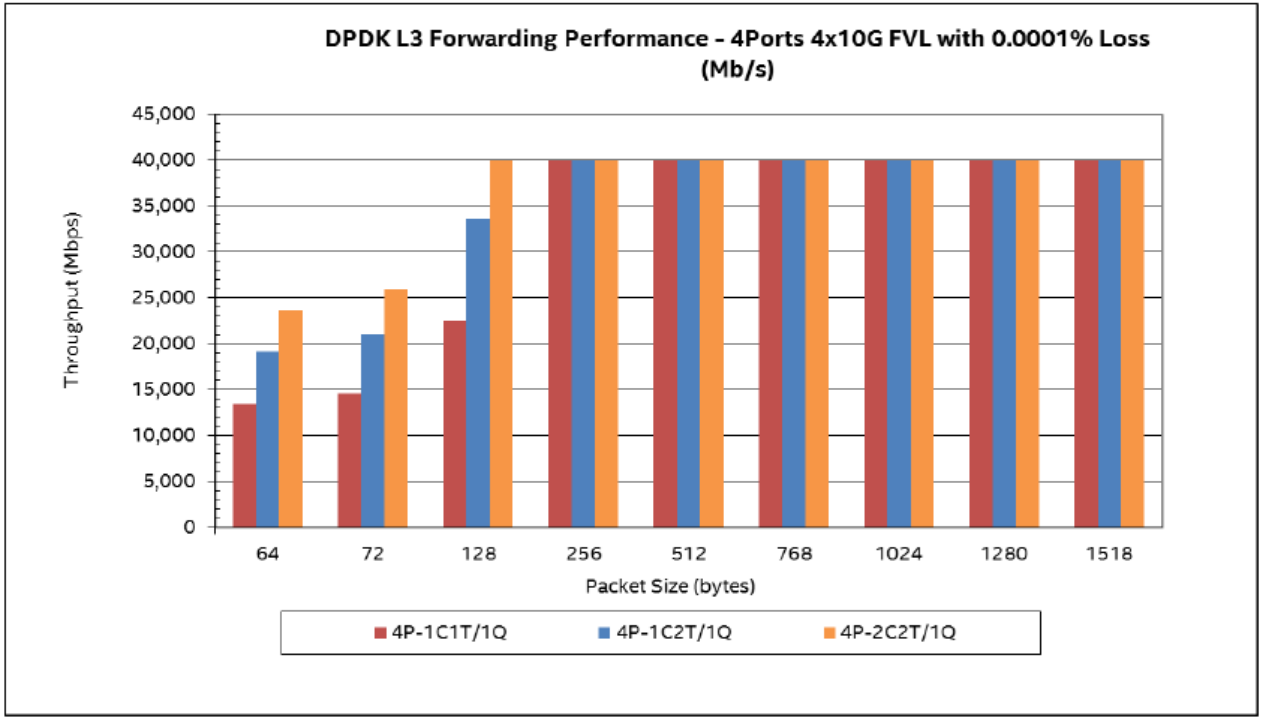
4 x10G FVL Ports



2 port configuration with 256 bi-directional flows per port

- Port 0 -> Port 1
- Port 1 -> Port 0
- Port 2 -> Port 3
- Port 3 -> Port 2

Intel® Xeon® E5-2658 v4 Processor
DPDK L3 Forwarding Performance using 10GbE



DPDK 16.04, 4端口 128字节 线速

Disclaimer: Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. Source: Intel internal testing as of July 12, 2016. See Slide 6 and Slide 7 for configuration details. For more information go to http://www.intel.com/performance

* Other names and brands may be claimed as the property of others.

Intel® Xeon® Gold 6152 Processor DPDK 17.02 Edge Router Aggregated Summary

Notes :

- Each VM serves 2x10 G and occupies 3 Cores 6 Threads of Socket 1.
  - Master / OS – 1 Core 2 Threads
  - Upstream – 1 Core 2 Threads
  - Downstream – 1 Core 2 Threads
- 7 VMs created with 21 Cores 42 Threads totally serving 140G

Observations :

- Each Edge Router VM instance was able to achieve 20 gigabits line rate from 512 bytes
- Aggregated system level performance reached 140 gigabits of line rate from 512 bytes with 95% CPUs (out of 22 cores involved).

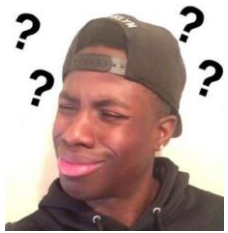
DPDK 17.02, 起7个虚机，21个core，42个threads，512字节140G

DPDK转发

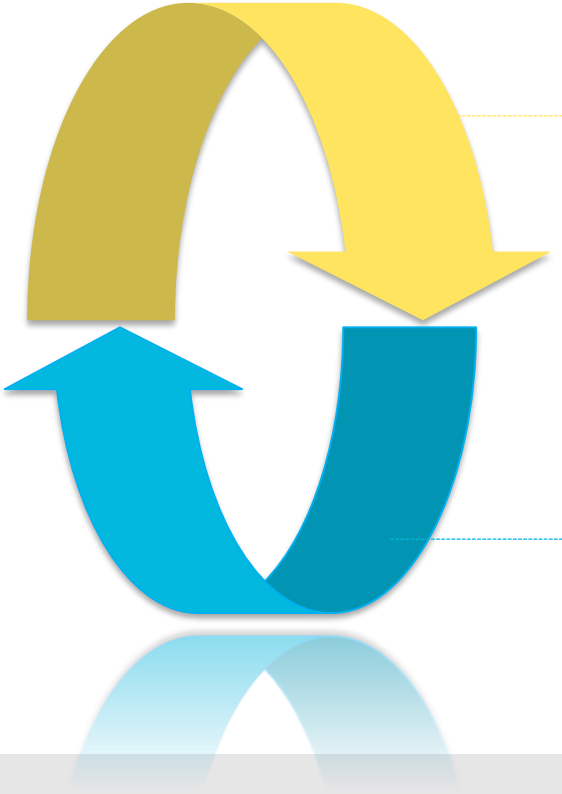
- 无锁队列
- 内存零拷贝
- PMD调度模型
- CPU指令编译优化
- 大页表减少TLBmiss
- 易于容器化

内核转发

- **无锁队列**
- 内核零拷贝
- PMD调度模型
- CPU指令编译优化
- ？ 内核中使用，效果甚微
- 容器化需要用户态程序，无法支持



NFV采用DPDK的驱动力



- 容器技术的发展

DOCKER的开源,技术逐渐成熟,在互联网业务中得到很多应用. NFV领域也把容器技术包含到虚拟化技术里面

- NFV性能提升的需求

运营商关注的重点

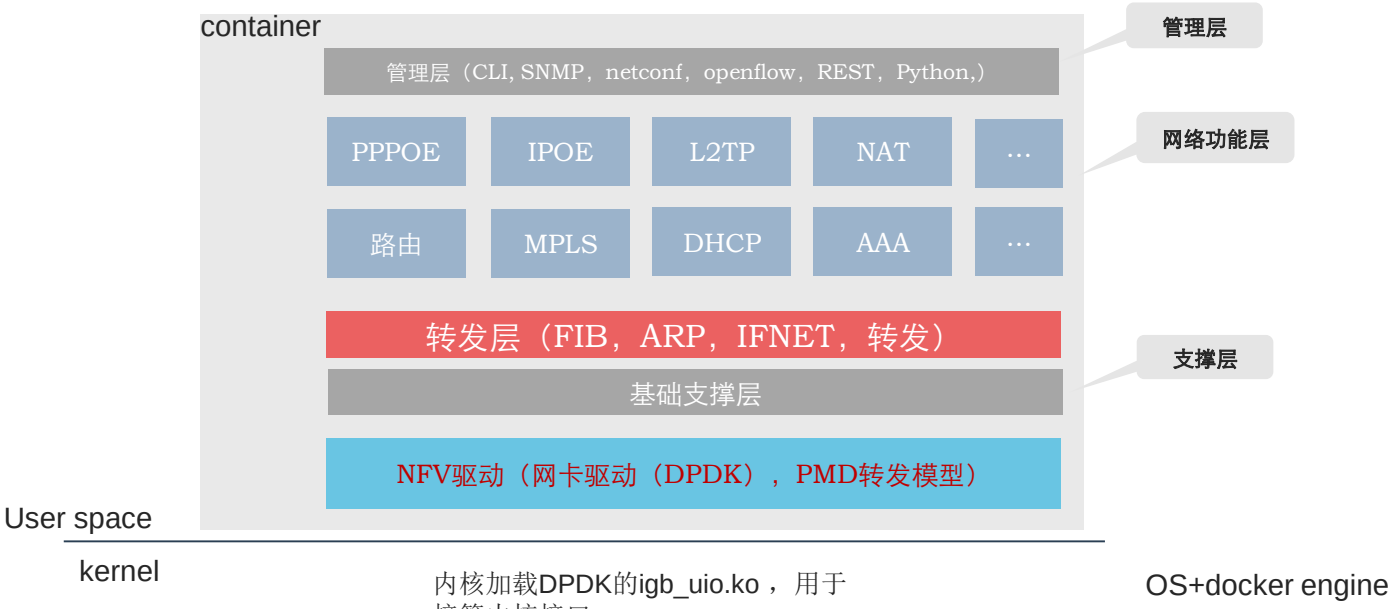
实现用户态协议栈

DPDK不支持用户态协议栈，需要厂家自行支持。
内核实现的协议栈，需要移植到用户态，以配合DPDK转发。
另一种选择：转发报文在用户态走DPDK转发，本机报文下内核，维持原有流程。

移植DPDK驱动

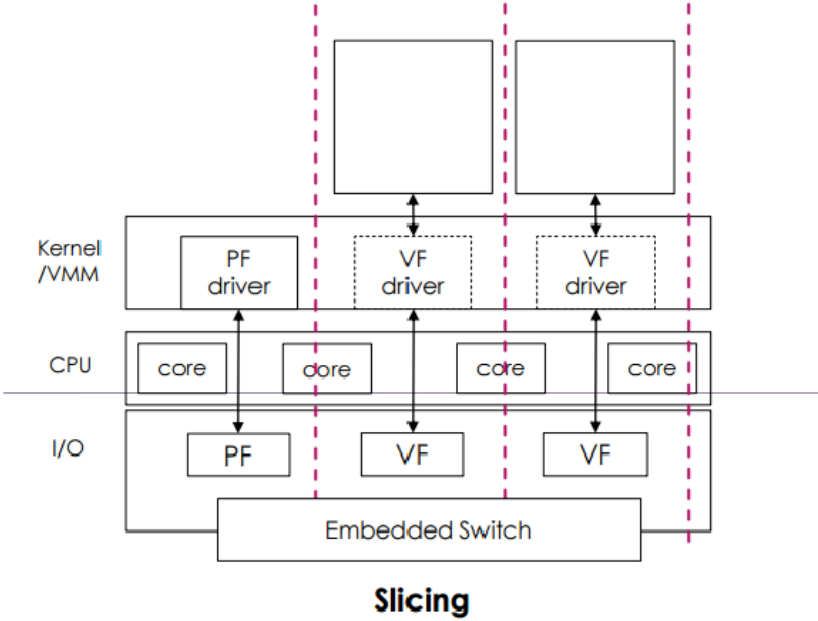
H3C的NFV基于Comware平台，移植DPDK驱动需要做适配。
支持PF口，VF口，Virtio口。

NFV+DPDK结构图


container
管理层（CLI, SNMP，netconf，openflow，REST，Python,）
PPPOE
IPOE
L2TP
NAT
…
路由
MPLS
DHCP
AAA
…
转发层（FIB，ARP，IFNET，转发）
基础支撑层
NFV驱动（网卡驱动（DPDK），PMD转发模型）
管理层
网络功能层
支撑层
User space
kernel
内核加载DPDK的igb_uio.ko，用于
接管内核接口
OS+docker engine

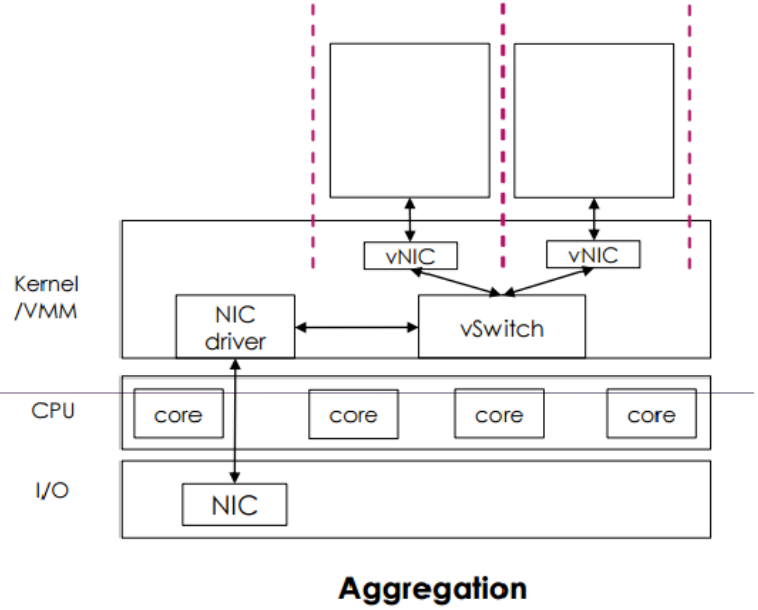
DPDK在Docker中的网络接口：

NFV需要多个网络接口，也需要各种类型的接口。DPDK提供了多种接口支持。



Slicing

可以获得更好的性能（SRIOV）

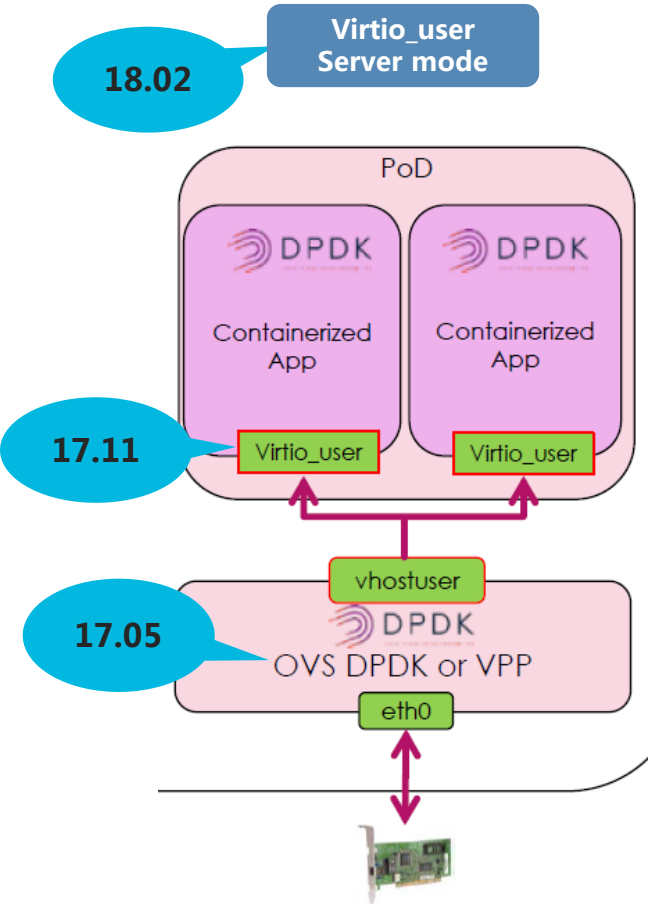


Aggregation

可以用于功能调试或性能要求不高的场景（Virtio）

DPDK调试中遇到的问题

- DPDK版本选择问题：Intel长期维护版本17.11(一年)。有些新功能可能不会合入这个分支。需要使用者长期自己维护和跟踪升级。
- 支持X710网卡驱动的版本需要和固件firmware版本配套,否则可能流分类不均网卡达不到线速，影响吞吐。
- DPDK对某些非Intel生产的网卡支持会有问题，比如HP加工的82599网卡，读DMA会失败。商用时会遇到兼容问题。
- 编译环境问题：需要在编译时选择运行环境，一次编译出的版本不能运行于多种CPU环境。和NFV乃至容器的一次编译，多次运行的思想不一致。



- 1.支持Virtio网卡驱动需要OVS+DPDK做中间交换，OVS目前只支持DPDK17.05

Virtio user侧需要使用DPDK17.11版本。
OVS DPDK的vhost可以兼容17.11的Virtio。目前可以使用。

- 2. 原来DPDK侧Virtio user不支持server mode，

和OVS建立连接时，只能把OVS DPDK作为server端使用
OVS升级或故障会影响NFV，要重新建连。
DPDK18.02开发了Virtio user支持server mode，解决了这个问题。
不过需要使用者自己把代码从18.02移植到17.11上。

DPDK调试中遇到的问题

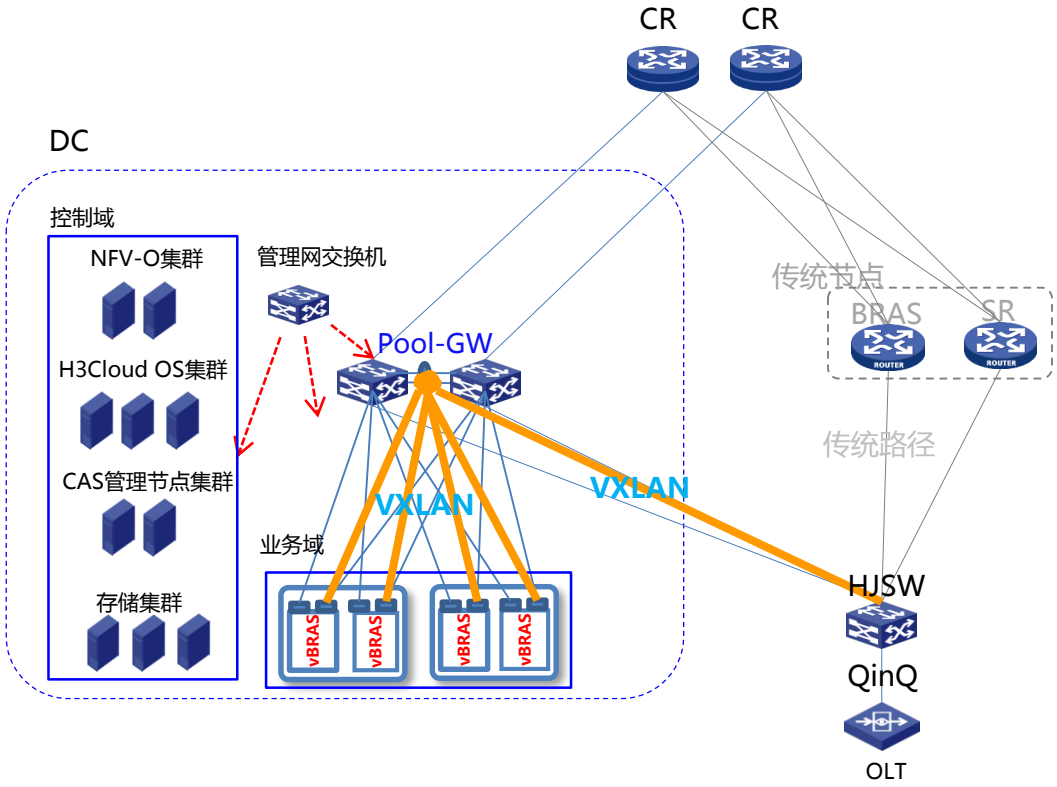
- Virtio_user作为server场景，容器启动依赖于ovs dpdk上的client端口：当ovs dpdk上加上dpdk vhostuserclient端接口时，才会正常初始化virtio-user接口；
- 接口RSS功能，需要增加新类型时周期比较长，且要升级firmware。商用时不方便。且VF口的RSS需要在PF口上配置。
- Host上的82599 PF状态改变，对应容器中的DPDK VF口不能感知状态变化
- CPU占用率问题：PMD模式显示CPU占有率通常是100%，不能真正体现CPU忙闲程度。

虽然道路有点曲折，但是未来可期

使用DPDK带来的好处

- DPDK屏蔽了不同网卡驱动对链路层的差异，新增网卡类型很方便
- RSS功能很有用 ，硬件直接分流，可以节省软件CPU负担。
- 可以使用DPDK的全部优化手段
- Intel强大的科研团队，支持团队，不断的优化转发性能
- 性能提升!

NFV应用 - vBRAS资源池全业务部署



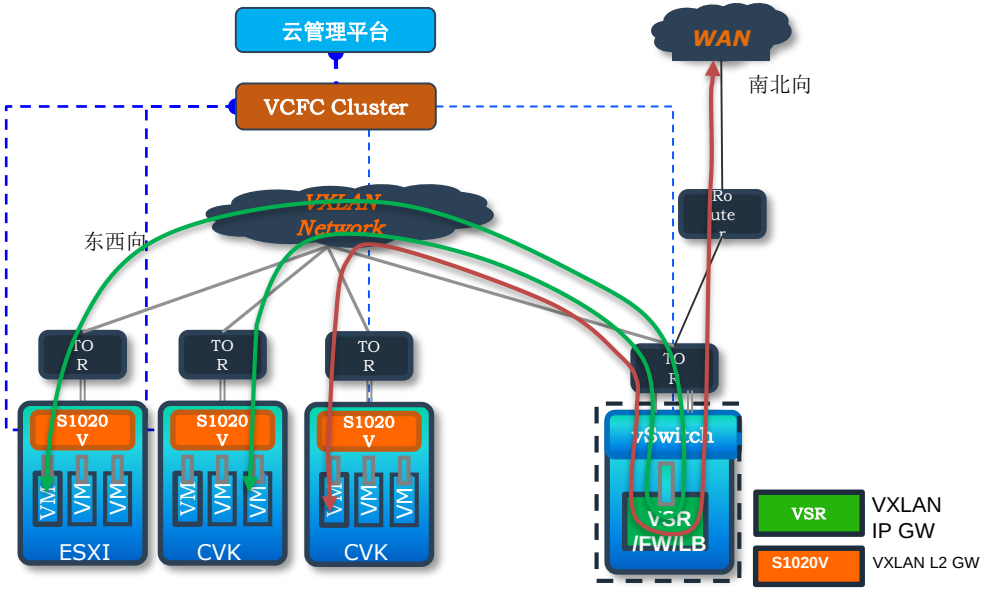
1、vBRAS在DC内。可以是1：1热备，可以是N：1资源池冷备

2、vBRAS为容器形态，支持IPOE，PPPOE，Portal，L2TP，全业务支持

3、故障管理：vBRAS控制器在设备故障后，自动拉起新设备，同步配置。

4、数据报文通过VxLAN隧道上送vBRAS。

NFV应用 - VSR公有云



1、租户流量很低时，一个物理服务器上可以起多个NFV。

2、租户流量大时，比如专线用户，一个物理服务器上只能起一个NFV，且需要不断提高性能。

无论哪种应用，性能提升都是非常重要的。

01 NFV简介

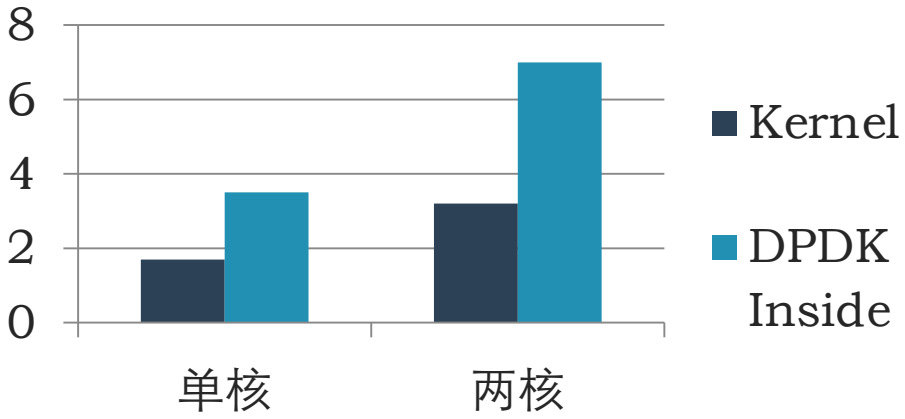
02 DPDK的优势

03 DPDK在NFV中的应用

03 性能对比

应用了DPDK后，单核转发性能可以翻倍。

XEON GOLD处理器，开睿频，开超线程，可以获得较好性能
XEON E5V4处理器
XEON E5V3处理器，单核性能最高，但核数少（10核）



Thanks !

新华三集团
www.h3c.com