中国移动
China Mobile

中国移动NFV硬件加速策略与思考
NFV Hardware Accelaratation strategy in CMCC

中国移动研究院网络与IT技术研究所

王升

2019.06

www.10086.cn

提纲

- 硬件加速的必要性
- 加速硬件选型
- NFV硬件加速方案
- 硬件加速产业生态和开源情况
- 下一步工作

5G网络低时延、高可靠、大流量特征对网络提出更高要求

5G网络高可靠、低延时、大流量的特征以及边缘计算业务兴起对未来网络计算和转发能力提出更高要求，网络功能虚拟化后网元和VSW采用软件实现，通过消耗CPU满足计算转发成本和功耗较高，需要采用硬件加速方案，将原子化的功能单元卸载至硬件加速卡上

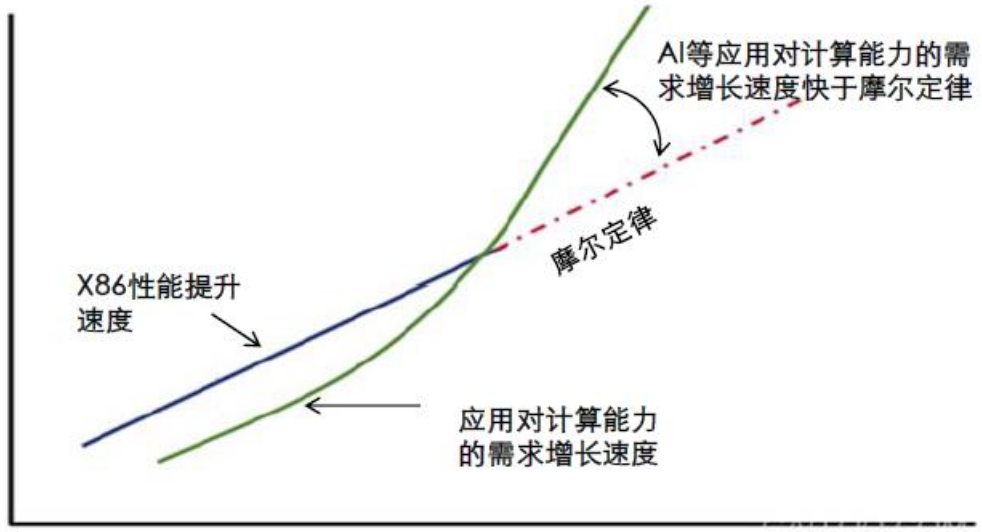
对比4G，5G网络

-支持移动数据流量高1000倍
-连接设备数高10-100倍
-尖峰数据率高10-100倍
-网络延时低10倍（抖动）
-功耗能效比高10倍
-维护成本低50倍
-更安全
-自动网络优化
-故障分析和预测
-AI、深度学习
-大数据处理

计算加速
网络加速
存储加速
图像处理加速
…

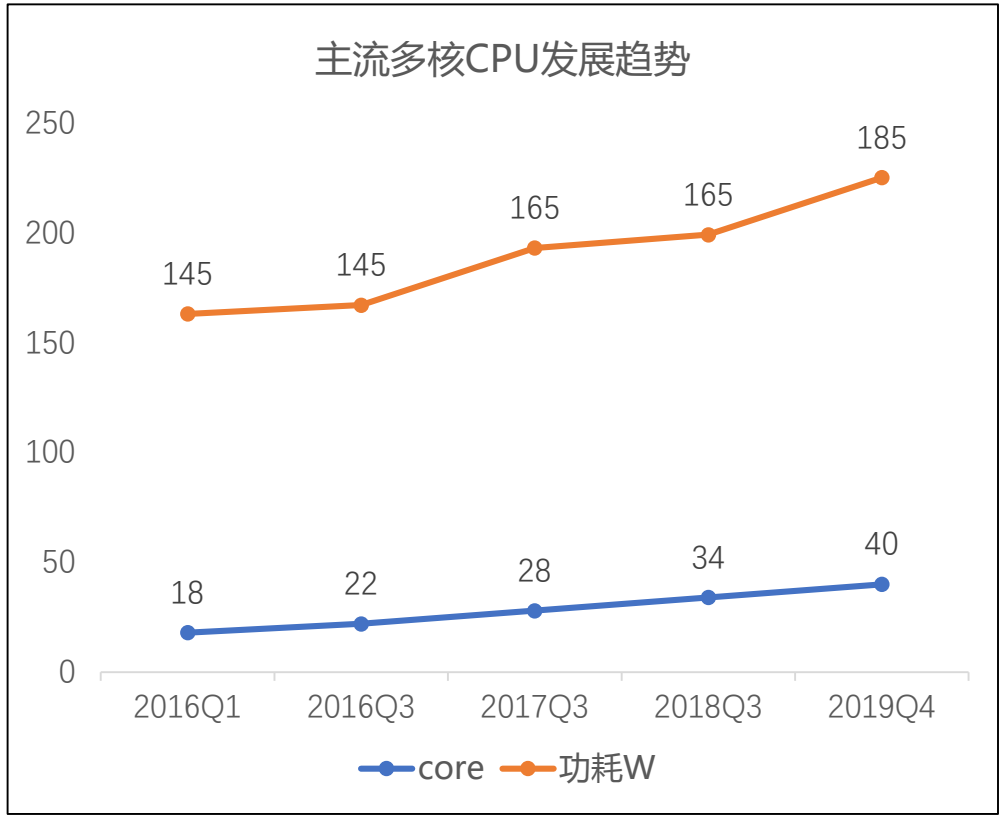
边缘云

-有限的计算资源
-有限的功耗/制冷
-有限的空间



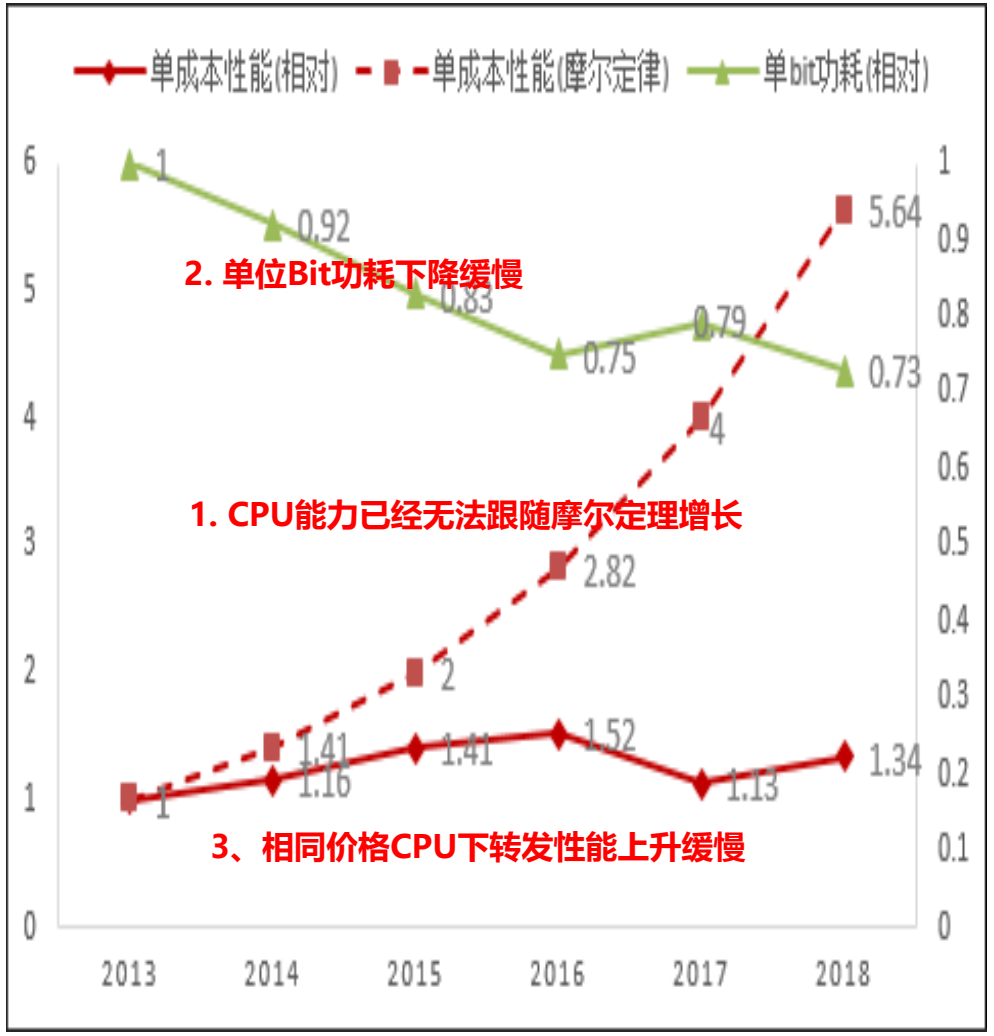
降低单bit成本：CPU堆叠不可持续

在功耗/成本恒定的情况下，CPU核数已难大幅提升，限制性能提升



2016	2017	2018	2019	2020	2021
Haswell(Grantly) 18c,145W,DDR4	Broadwell(Grantly) 22c,145W	Skylake(Purley) 28c,165W,Omni-Path,FPGA	Ice Lake(Purley) 34c,165W	Sapphire Rapids(Tinsley) 40c,185w,8CH DDR5,PCIe 4.0	Granite Rapids(Tinsley)

从产品应用看，CPU单位成本性能上升缓慢，单bit功耗下降缓慢



流量大幅增加→VNF性能大幅提升→VNF性能依赖CPU→矛盾如何解决？←CPU技术发展的摩尔定律失效

虚拟化中的软件加速只能弥补I层带来的性能下降，无法做业务加速，加速效果有限

物理形态 | 虚拟化形态

通用CPU

非虚拟化：上层应用可直接看到物理网卡，直接收发包，性能高

虚拟化(OVS)：上层应用(VM)能共享物理网卡，同时VM与物理网卡解耦，增加一层虚拟交换机实现(OVS)，需单独占用CPU的核资源，又通过内核中断方式收发报文，内核态到用户态有内存拷贝，性能较低

虚拟化(OVS+DPDK)：通过用户态进程接管网卡数据收发，“IO独占核”，每个端口分配一个核用于数据收发。轮询处理方式比中断高效，节省内存拷贝，性能相比OVS有显著提升，但虚拟交换机仍然独占CPU核，性能较弱

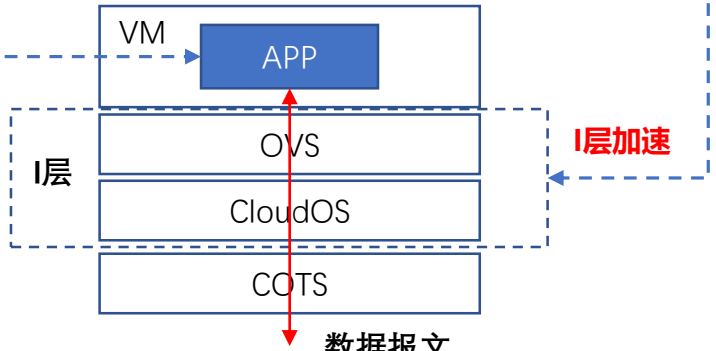
虚拟化(SR-IOV)：通过硬件设置将物理网卡映射成多个网卡供VM使用，虚拟机可以直接连接到物理网卡上，报文可跨过虚拟层直接到VM，这种方式不再需要虚拟交换功能，节省资源，但会导致VM和物理网卡无法解耦，SR-IOV在实际使用中的通信性能基本消除I层带来的性能开销

SR-IOV加速技术在实际使用中存在一系列劣势，且其通信性能已经接近上限，后续软件性能优化只能依赖上层应用逐步优化，软件优化空间不大

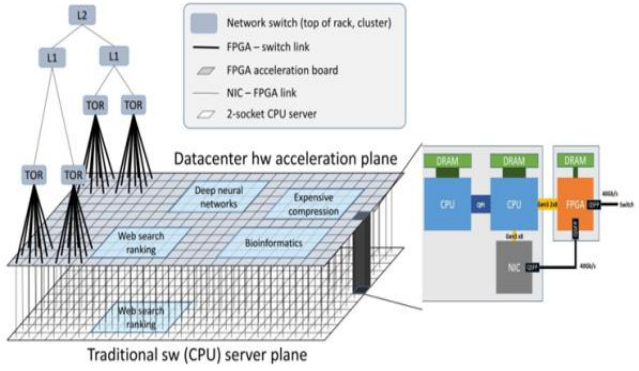
专用加速硬件

硬件加速：通过硬件卸载CPU的业务功能，从而获得更高性能，加速比高

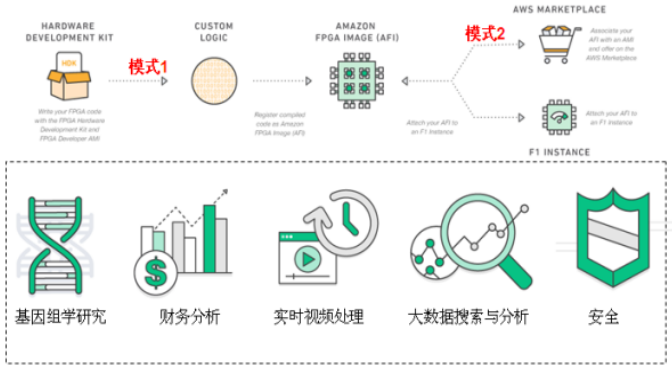
业务加速



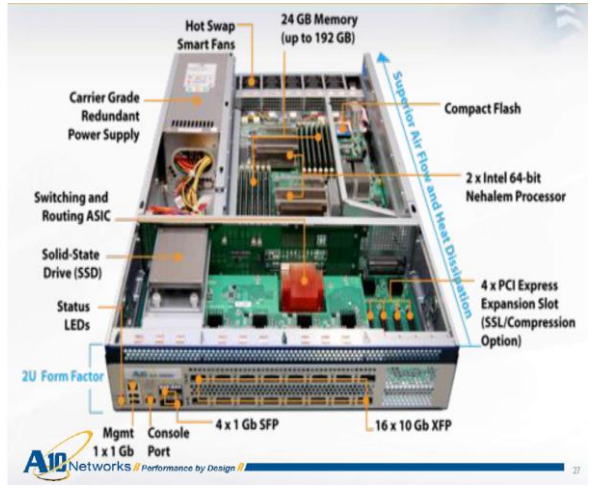
硬件加速产业应用



微软Azure硬加速方案采用FPGA进行网络加速，涵盖生物信息学，神经网络，搜索排名等方向，成为业界标杆



亚马逊AWS 16年底推出FPGA加速服务F1，作为增强EC2实例提供给第三方进行租用。



A10的一款用于数据中心的LB设备：基于COTS+加速硬件的整体结构，提升性能降低成本

- 扩展FPGA完成L4/L7转发卸载
- 扩展Crypto ASIC完成SSL功能卸载

腾讯云2017年推出国内首款高性能异构计算基础设施——FPGA云服务，利用云服务的方式将只有大型公司才能长期支付使用的FPGA服务推广到了更多企业。

阿里云 在智能加速引擎方面，采用AMD,NIVIDIA GPU加速图像处理，AI等行业应用，使用intel，Xilinx FPGA芯片,面向科学计算行业提供了良好的高速及并行性能支撑。使用mellanox智能网卡卸载OVS，实现了较高的网络性能

JD.京东 verizon UCLOUD flipkart.com

京东、Verizon公有云及私有云部署采用ovs卸载硬件加速，主要实现vlan/vxlan的解、封装功能，同时增强网络转发能力，释放CPU资源，简化组网

提纲

- 硬件加速的必要性
- 加速硬件选型
- NFV硬件加速方案
- 硬件加速产业生态和开源情况
- 下一步工作

加速卡硬件选型

目前市场上流行的加速芯片有多种选择。加速芯片嵌入网卡形成智能网卡是目前加速卡的主流形式。其中FPGA当前产业较为成熟，且可现场编程灵活性高；NP、SoC性价比较高，但产业成熟度有待提高；GPU主要优势为图片复杂算法处理

	FPGA	NP	SoC	GPU
产业	☆☆☆☆ 主流厂商Xilinx（美国）、Intel（美国），产业情况较热	☆ • 主流厂商均**无路标（包括**Cavium（美国）、Broadcom（美国）、EZCHIP (mellanox)、Netronome（美国**)** • 华为自研NP	☆☆☆ Mellanox (美国)/Cavium（美国）/Broadcom（美国） 多为ARM架构	☆☆☆ NVIDIA（美国，图像渲染、深度学习）、AMD（美国，图像渲染）、Intel（美国）
开发周期	☆☆☆ **0.5~1年**	☆ **1~2年**	☆☆☆ **0.5~1年**	☆ **1~2年**
生态	**好**	**封闭，较差**	**较好**	**好**
加速功能	**网络、计算加速**(网络转发、Ipsec、DPI、GTP、H-QoS、charging等)	**网络加速**(GTP、IPFWD等)	**网络、计算加速**(网络转发、transcoding、Ipsec、DPI、charging等)	**图像处理、人工智能、图像渲染、深度学习**等
编程语言	硬件描述语言，Verilog/VHDL	汇编语言，微码	SystemC	CUDA, OpenCL
接口解耦	目前DPDK和VirtIO已有相关工作	•VirtIO当前没有相关工作 •NP厂商解耦存在一定难度	VirtIO当前没有相关工作	VirtIO暂无相关工作
编排	Cyborg已部分支持	Cyborg尚无	Cyborg尚无	Cyborg已部分支持
更新	可现场编程	**NP中功能可升级** **支持热补丁** **不可重置**	**软件刷新，支持热补丁**	不可现场编程

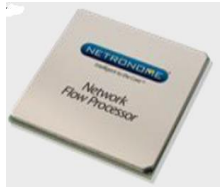
未来的加速技术多样化，需结合业务特点选择合适的硬件加速技术

由于不同加速硬件面对的加速场景不同，且加速卡片上资源有限，当前业界尚无一卡多用产品，针对不同业务加速需建设不同加速硬件资源池

CPU

功能卸载

转发加速：FPGA/NP/SOC



适于业务简单、吞吐高的数据报文转发处理，如Router、Firewall、UGW

转发业务加速资源池

业务计算加速：FPGA/SOC

具有可编程能力，适于变化不频繁业务，如数据处理

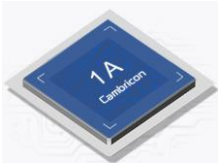
计算业务加速资源池

图形计算加速：GPU

擅于并行计算、浮点运算，整块数据流处理，如图像处理、音视频编解码

图形业务加速资源池

AI计算加速：AI芯片



一般支持业界典型AI算法，如智能摄像头、图像识别等

AI业务加速资源池

低功耗通用加速：ARM芯片

通用CPU，可以做任何业务处理，特定业务处理时性能不如专有芯片

低功耗通用计算资源池

需要加速的网元、业务和功能

中国移动 China Mobile

分析当前电信网元，网络加速需求更为突出，AI等边缘计算新业务的计算加速和存储加速需求更为突出。当前重点关注：
①OVS加速，为所有网元提供普遍的网络转发加速能力；②网元特定功能加速，如GTP；③ MEC中面向AI、图像处理的GPU加速

业务	网元/应用	网元/应用主要功能点	初步建议
EPC	SAE-GW	•L2/L3/L4层转发 •**GTP** •**DPI** •Ipsec •Charging QoS	•5G eMBB、uRLLC场景，要求网元具备大带宽、低延时、低抖动的特征，为减少CPU消耗，降低网元业务处理时延，**推荐加速** •建议：根据网元各功能的CPU消耗、产业现状，推荐首先进行**GTP**加速卸载
Volte IMS	vSBC	•transcoding •Ipsec	•SBC位于核心或地市电信云，资源相对充足，少量加速效果更改核心网资源池，略鸡肋 •建议：不加速
5G	UPF	•L2/L3/L4层转发： •DPI •Ipsec •charging QoS	•5G eMBB、uRLLC场景，要求网元具备大带宽、低延时、低抖动的特征，为减少CPU消耗，降低网元业务处理时延，**推荐加速** •UPF下沉，边缘机房资源/空间有限，需要加速以减少资源占用 •根据网元各功能的CPU消耗、产业现状，推荐首先进行**GTP**加速卸载
	MEC	DPI/IPSec/AI/视频编解码/位置服务/物联网/现实增强/本地内容优化和缓存等	除AI、AR/VR、云游戏等业务对**GPU加速**有明确需求外，其他边缘计算业务尚不明确，建议根据边缘计算具体业务需求制定加速策略
固网	BRAS-U	•L2、L3、MPLS转发 •Pppoe/ipoe隧道处理： •H-QoS •组播 •流量统计	•BRAS现状：U面尚未虚拟化，存在白盒交换机和X86加速两种方案 •X86方案优点：支持NFV、软硬解耦、快速上线与扩缩容 •建议：**进行X86加速方案研究，推动虚拟化**研究进展 •推荐加速点：H-QoS、组播、ACL等

初期推荐采用FPGA，待功能稳定后可考虑进一步引入其他低功耗低成本方案

当前电信网元加速功能下沉智能网卡尚处于初期，重点关注转发面网元GTP功能卸载，具体功能尚不稳定。考虑产业情况和灵活性，建议采用FPGA方案，可灵活修改调整。

加速功能稳定后建议可逐步固化为NP、ASIC等低功耗低成本不可编程芯片，提高硬件加速性价比

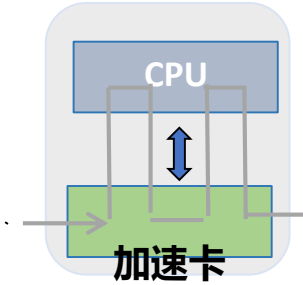
FPGA基本架构

FPGA将电路集成，通过软件定义实现具体逻辑功能，逻辑模块互相连接组成可编程模块， 通过Verilog/VHDL语言自定义电路逻辑，内部包含动态区域和静态区域：

- **静态区域**：PCIe, DMA, 以太网接口等实现
- **动态区域**（PR）：用户可编程部分，用户可在PR内开发所需加速功能

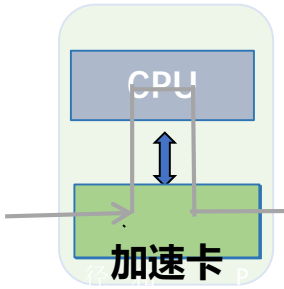
FPGA数据处理方式

Look-aside



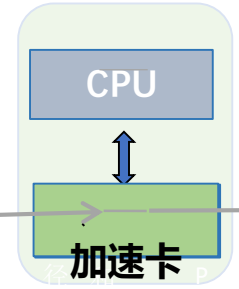
CPU与加速硬件之间交互多，占用总线带宽，影响性能

In-line



CPU与加速硬件之间交互少，加速比较高，解耦要求高

Fast-path



CPU与加速硬件之间无交互，加速效果最好，但业务逻辑完全下沉网卡对企标、网卡厂商开发能力要求高

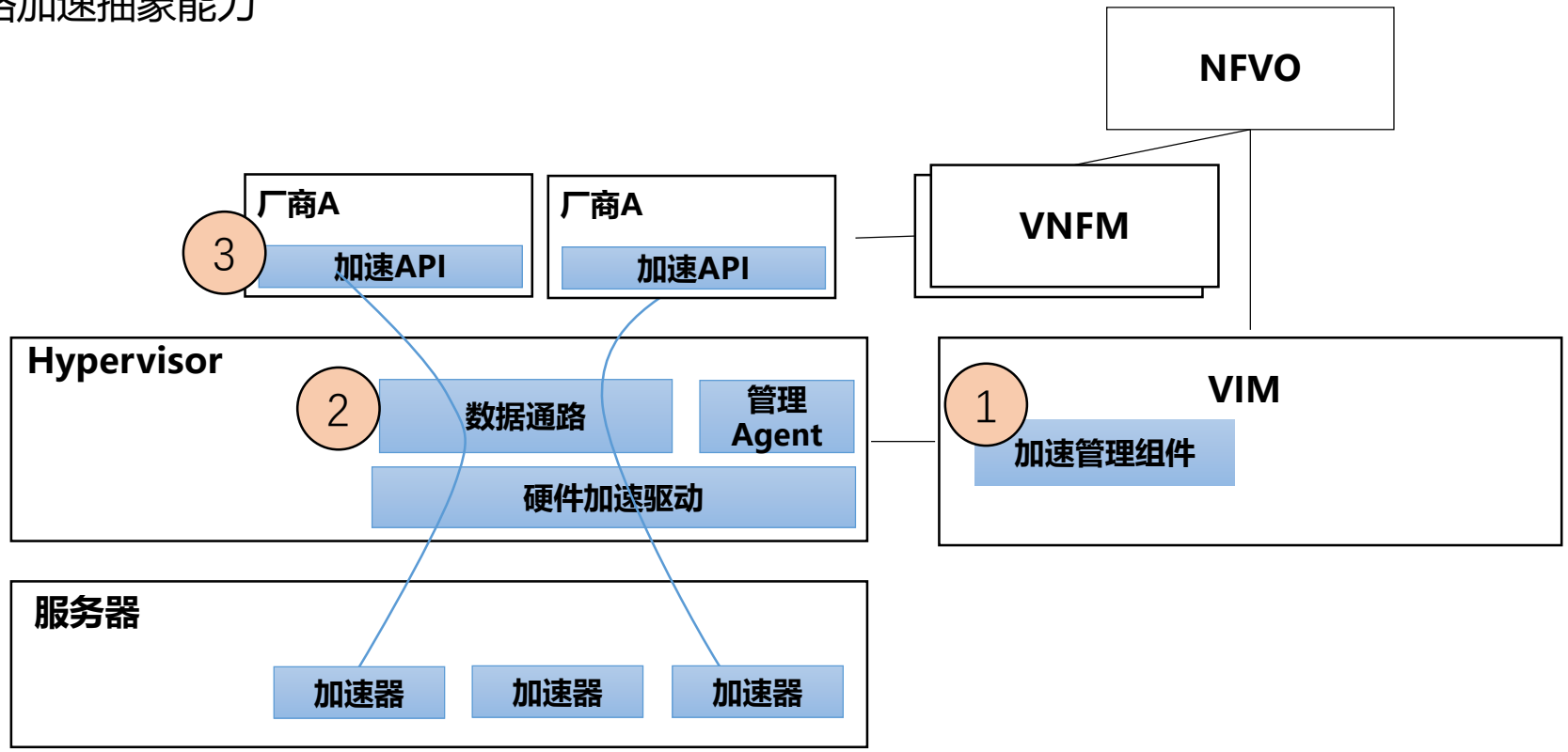
提纲

- NFV硬件加速的必要性
- 加速硬件选型
- NFV硬件加速方案
  - 加速资源管理编排方案
  - 硬件加速数据通路
  - 加速通用API和解耦方式
- 硬件加速产业生态和开源情况
- 下一步工作

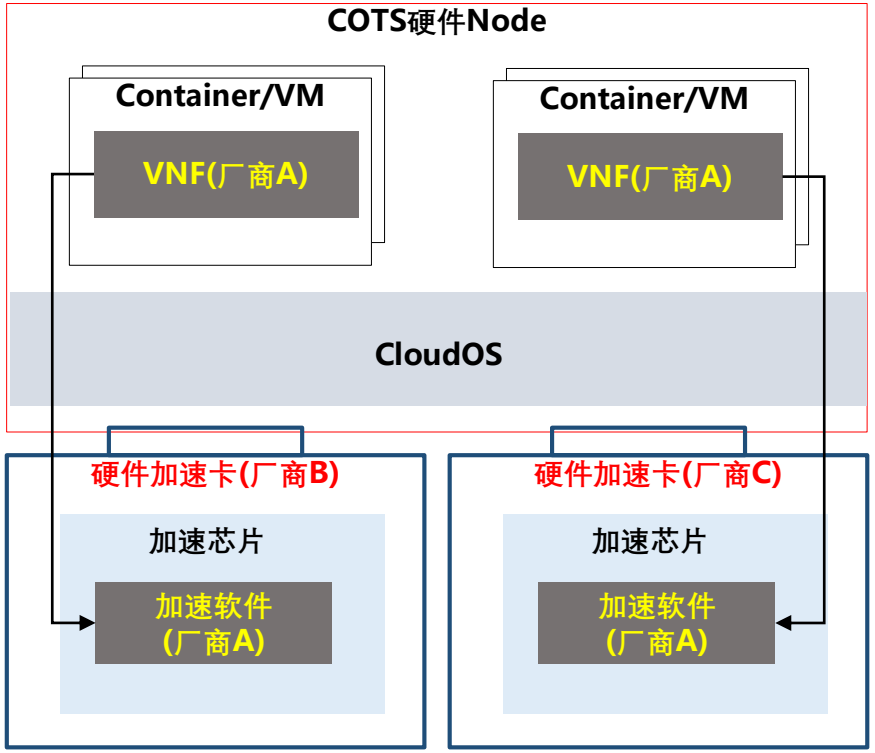
NFV硬件加速方案关键问题

硬件加速资源被NFV网元调用需要解决以下几个问题

1. 加速资源的编排，确保通过NFVO和VIM可以看到并选择适当的加速资源供VNF使用，OpenStack Cyborg目前已提供部分功能
2. 通用的软硬通道，解决VNF软件与加速硬件间数据通路。目前较为成熟方案为VirtIO和SR-IOV
3. 通用的加速API，解决VNF与加速硬件解耦问题，确保加速硬件提供能力充分抽象。目前DPDK已提供部分网络加速抽象能力



通用性：加速卡的解耦是加速资源池化的基础



解耦方式一：软硬解耦

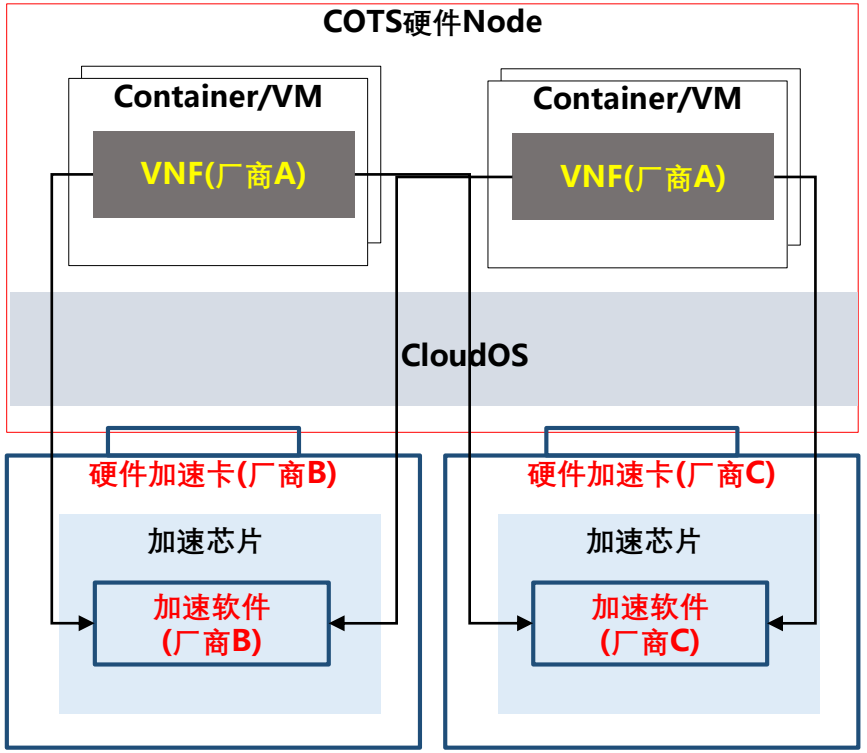
➢ 加速软件与VNF软件同厂商，与加速硬件异厂商

优点：

- 不需统一VNF和加速软件之间接口

不足：

- 需针对不同芯片开发不同软件，需匹配合理的加速硬件采购方案(如厂商A分别提供针对厂商B和C的两套软件，则A可在两个卡上运行；或者A只提供针对厂商B的软件，则只能在B厂商卡上运行)
- 加速软件与加速硬件不能统筹设计，难以做到高性价比
- 只能做到软硬解耦，VNF软件无法使用异厂商加速软件能力
- 采购模式变化，需要软硬一起采，不便于硬件资源池的统筹建设



解耦方式二：软软解耦

➢ 加速软件与VNF软件异厂商，与加速硬件同厂商

优点：

- 软件厂商不需在不同硬件上重复开发
- 加速软件与加速硬件能统筹设计，易做到高性价比
- 软软解耦，VNF可使用异厂商加速能力

不足：

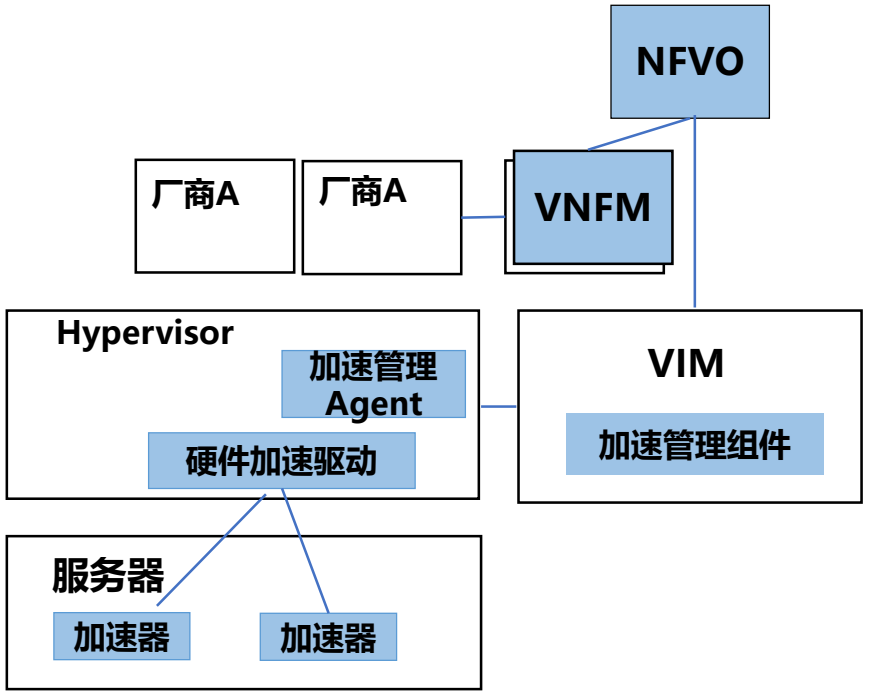
- VNF软件与加速软件间的接口需标准化，难度大、周期长
- 涉及到VNF业务逻辑卸载到加速卡，对加速卡硬件厂商开发能力要求高
- 需要对VNF进行改造，上层VNF升级可能需要与加速卡加速软件升级同步，拉长上线周

提纲

- NFV硬件加速的必要性
- 加速功能分类
- 加速硬件选型
- NFV硬件加速方案
  - 加速资源管理编排方案
  - 硬件加速数据通路
  - 加速通用API和解耦方式
- 硬件加速产业生态和开源情况
- 下一步工作

加速管理编排方案

当前，在NFV加速管理编排虚拟机方面，需要MANO、OpenStack和加速器协同完成，共同打通管理编排流程。



NFVO/VNFM对于加速资源管理的流程和接口要求

- NFVO要求能够获取VIM上的加速资源信息
- VNFD中有对加速资源的描述信息
- VNFM能够解析VNFD中的加速资源信息
- 其他要求（可靠性、安全和兼容性要求）

Hypervisor对于加速资源的管理和兼容性要求

- 加速管理agent对加速硬件的纳管要求
- 加速镜像管理要求
- 加速数据的标准化（包括加速镜像名称，镜像UUID ,设备商，版本号，驱动，驱动版本等）
- 其他要求（可靠性、安全和兼容性要求）

VIM对于加速资源的管理和接口要求

- Cyborg相关组件的管理、调度和其他组件的协同要求
- VIM北向接口加速相关原生接口使用要求
- VIM需要将普通网卡与加速器信息区别上报要求
- 其他要求（可靠性、安全和兼容性要求）

加速编排管理——OpenStack Cyborg（1/2）

中国移动 China Mobile

当前OpenStack Cyborg提供的功能尚不能满足中国移动的加速管理编排方案需求，迫切的需要厂商积极参与cyborg项目，实现管理功能，推动其成熟； MANO需要建立并完善相应功能，目前厂商并没有相应实现，可先通过企标推动实现；最终实现加速编排管理流程打通。

类别	加速管理编排需求	Cyborg实现情况（截止到R版本）	需要推动和加强的工作
NFVO/VNFM对于加速资源的管理要求	1.NFVO要求能够获取VIM上的加速资源信息	不属于Cyborg范围，cyborg只提供对上的接口，暴露相关接口和信息	需要MANO做相统一要求和实现，协同加速管理流程的端到端实现。
	2.VNFD中要有对加速资源的描述信息		
	3.VNFM具有解析相应加速字段的能力		
虚拟层对于加速资源的管理和兼容性(Hypervisor & vim)	1. 对加速网卡的纳管要求	目前实现的是FPGA 的纳管，其他加速硬件并未有实质进展	需要考虑更多硬件的纳管，目前相对比较成熟的是FPGA
	2.加速镜像（FPGA）管理要求	仅最基本功能实现，有些问题处理方案暂无，很不成熟，包括与Openstack nova等其他组件的交互等等。	1.很不成熟，主要表现与Openstack组件互操作还欠缺很多，需要推动厂商充分参与Cyborg项目 2. 加速卡信息需与普通网卡信息区别上报
	3.加速数据的标准化问题	已经标准化	暂无
	4.性能和安全问题，比如迁移后，加速资源的使用问题	未考虑	Cyborg基本功能尚未实现，需要从中国移动自身的需求上推动完备性考量
	5.告警	不涉及	可视为虚拟层告警中的一部分

加速编排管理——OpenStack Cyborg（2/2）

中国移动 China Mobile

当前中国移动的加速管理编排方案需求和Cyborg提供的功能还有一定差距，迫切的需要厂商积极参与cyborg项目，实现管理功能，推动其成熟； MANO需要建立并完善相应功能，目前厂商并没有相应实现，可先通过企标推动实现；最终实现加速编排管理流程打通。

类别	加速管理编排需求	Cyborg实现情况（截止到R版本）	需要推动和加强的工作
VIM对于加速资源的管理和接口要求	1. Cyborg相关组件的管理、调度和其他组件的协同要求	仅最基本功能实现。与其他组件的交互还未实现，包括与Nova的虚机挂载交互和加速资源信息上报交互。	1.很不成熟，主要表现与Openstack组件互操作还欠缺很多，需要推动厂商充分参与Cyborg项目
	2. VIM北向接口加速相关原生接口使用要求	没有问题	推动Cyborg成熟
	3. VIM需要将普通网卡与加速器信息区别上报要求	没有问题	只需要vim北向接口区别上报即可
	4.其他要求（可靠性、安全和兼容性要求）	不涉及	Cyborg基本功能尚未实现，需要从中国移动自身的需求上推动完备性考量

提纲

- NFV硬件加速的必要性
- 加速功能分类
- 加速硬件选型
- NFV硬件加速方案
  - 加速资源管理编排方案
  - 硬件加速数据通路
  - 加速通用API和解耦方式
- 硬件加速产业生态和开源情况
- 下一步工作

硬件加速数据通路

网元到加速网卡的数据通路可以选择SR-IoV和VirtIO。其中Virtio便于热迁移，且使用virtio FE或DPDK可实现加速能力抽象。有利于网元和加速硬件解耦

建议对当前virtio已较为成熟支持的数据通路卸载能力（如ovs卸载）优先采用virtio；对目前暂不成熟的（如GTP卸载），可暂时先采用SR-IOV， 同时积极推动VirtIO和DPDK开源和标准化工作

	SR-IOV	virtio
驱动	• VM需适配VF驱动 • 虚拟层需适配PF驱动 • VNF与虚拟层解耦后可能存在版本不匹配	通用驱动
功能	仅支持二层转发	支持SDN vtep、流镜像等功能
热迁移	不支持	支持
资源消耗	不占用额外资源	2个核（如果virtio BE也卸载至智能网卡则不需要）
端口数	受限（一般为64）	无限制

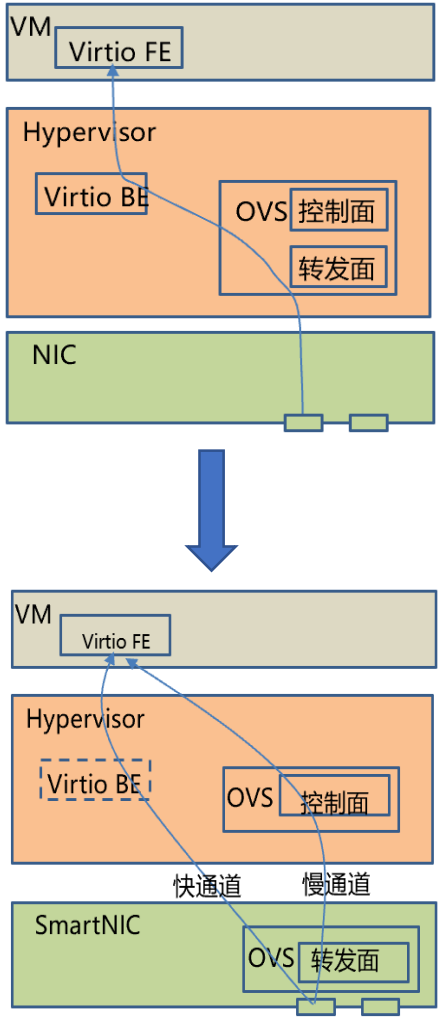
提纲

- NFV硬件加速的必要性
- 加速功能分类
- 加速硬件选型
- NFV硬件加速方案
  - 加速资源管理编排方案
  - 硬件加速数据通路
  - **加速通用API和解耦方式**
    - **OVS硬件加速卸载**
    - **网元硬件加速**
    - **GPU加速**
- **硬件加速产业生态和开源情况**
- **下一步工作**

OVS卸载介绍

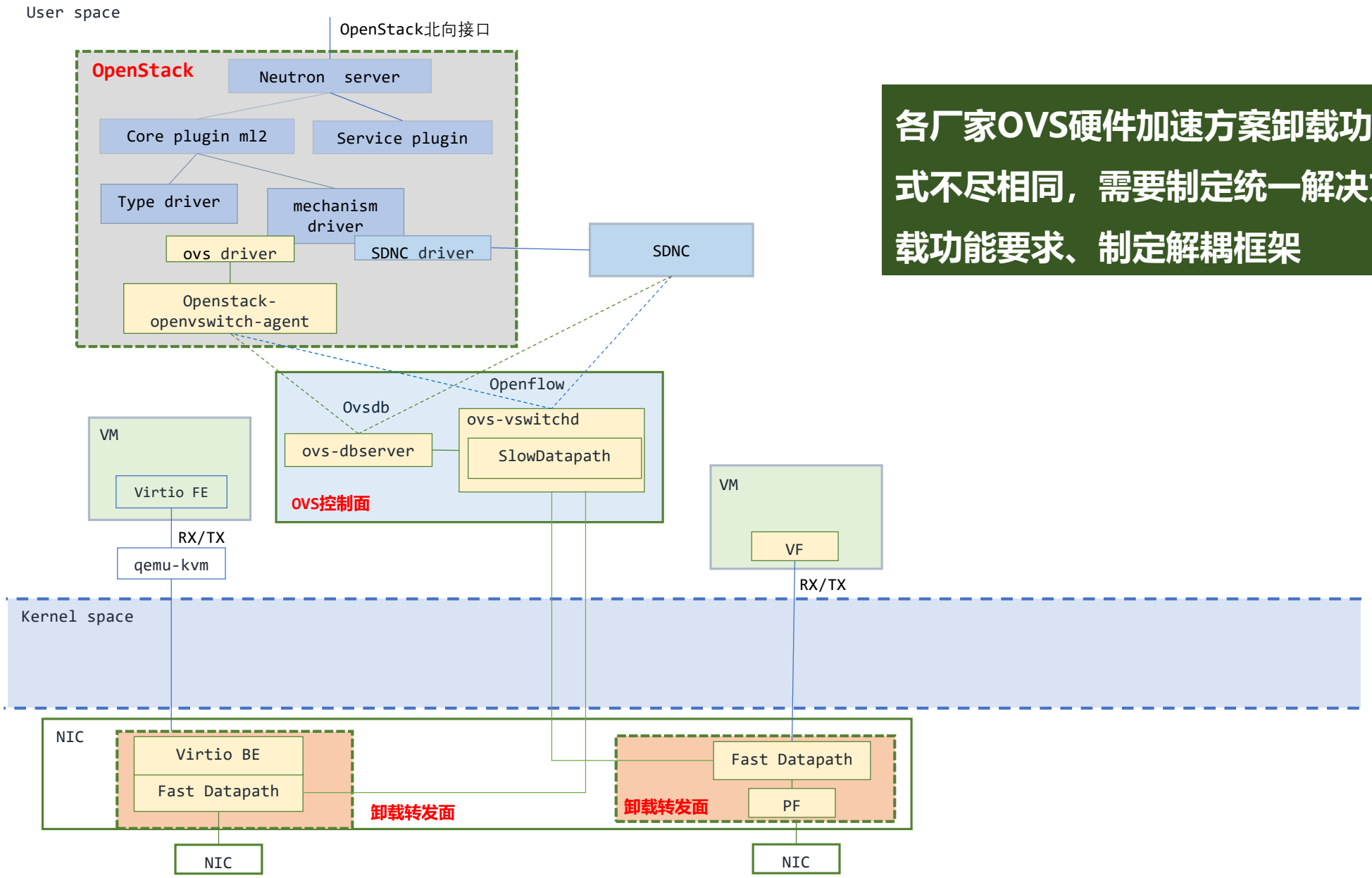
虚拟化场景下，当前数据面转发采用SR-IoV和OVS-DPDK软件处理两种方案。其中SR-IoV在驱动绑定，灵活性（热迁移）和与SDN组网方面存在技术劣势，OVS软加速需消耗CPU计算能力，性能较低。

Ovs卸载方案：将ovs转发面卸载至智能网卡，控制面仍在host运行



	SR-IOV	OVS/OVS+DPDK	OVS硬件加速
驱动	• VM需适配VF驱动 • 虚拟层需适配PF驱动 • VNF与虚拟层解耦后可能存在版本不匹配	通用驱动	通用驱动
功能	• 仅支持二层转发，不支持SDN vtep点（需要专用TOR进行VXLANID转换） • 组播模式下MAC混杂性能差 • 不支持安全组	支持SDN vtep、流镜、安全组像等功能	支持SDN vtep、流镜像、安全组等功能
热迁移	不支持	支持	支持
资源消耗	不占用额外资源	**占用大量CPU核**	占用少量资源
端口数	受限（最大64）	无限制	无限制
性能	近线速	小包场景性能较差 多流场景下转发能力线性下降	高于OVS软加速，随着版本优化，接近SRIOV性能

OVS硬件加速架构



各厂家OVS硬件加速方案卸载功能模块、解耦方式不尽相同，需要制定统一解决方案，标准化卸载功能要求、制定解耦框架

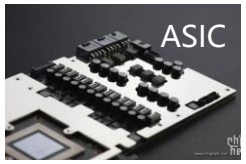
OVS硬件加速现状

当前业界积极进行OVS硬件加速技术的研究，中兴、华为、mellanox、Intel、曙光、netronome 等均有相应的OVS加速方案，部分厂家产品已在微软、腾讯、阿里等互联网公司应用，所用硬件包含FPGA、NP、ASIC，卸载方式也不尽相同，需要制定统一解决方案，推动生态发展。

intel
NETRONOME
The Flow Processing Company
ZTE中兴
Mellanox TECHNOLOGIES
HUAWEI
FiberHome
lenovo
Chelsio Communications Accelerate
中科曙光 Sugon

厂家

加速卡形态	选型厂商
FPGA	中兴、曙光联想
ASIC	Mellanox，华为
NP	netronome



加速硬件

OVS硬件加速解耦方案

方案	解耦方式	概述	优点	缺点
1	不解耦	虚拟层、加速网卡、服务器一家	• 无解耦难度，一体机方案	• 不符合解耦策略 • 采购模式变化，网卡与服务器绑定，软硬需要一起采
★ ★ 2	虚拟层&OVS功能&加速网卡 服务器	虚拟层携带OVS加速卡，与服务器解耦	• 仅需与服务器解耦，解耦难度小	• 爱立信无加速卡方案 • 采购模式变化：软硬绑定，需虚拟层与加速网卡一起采购
3	虚拟层&OVS功能 加速网卡&服务器	虚拟层携带OVS功能，加速卡加载OVS加速功能	• 纯软硬解耦	• 虚拟层的OVS的加速代码需要加载到加速卡，适配难度大，且OVS加速功能属通用功能，由加速卡做更合适
4	虚拟层&OVS控制面功能 OVS转发面&加速网卡&服务器	虚拟层携带OVS控制面功能，加速卡负责OVS转发面加速	• 需制定控制面与转发面标准接口即可 • 较贴合NFV架构	• OVS控制面/数据面接口业界无标准，厂家私有实现多，标准化困难 • 一个网络功能实现需要虚拟层厂商与网卡厂商共同开发，集成复杂，故障点较多，定位困难
★ ★ ★ 5	虚拟层 OVS功能&加速网卡&服务器	由服务器厂商提供加速网卡及OVS功能	• 服务器网卡一体，解决二三层网络转发工作，更贴合NFV架构 • OVS功能均由加速卡厂家实现，无需制定协议级接口，该解耦方式较易实现	• 加速卡需对虚拟层做内核版本等适配工作

从采购流程，开发敏捷，部署快速，运维简便等方面来看，建议采用方案5，即ovs与hypervisor解耦方案，由加速网卡厂商提供ovs功能，虚拟层厂商集成加速网卡ovs。

提纲

- NFV硬件加速的必要性
- 加速功能分类
- 加速硬件选型
- NFV硬件加速方案
  - 加速资源管理编排方案
  - 硬件加速数据通路
  - 加速通用API和解耦方式
    - OVS硬件加速卸载
    - 网元硬件加速
    - GPU加速
- **硬件加速产业生态和开源情况**
- **下一步工作**

转发面U面引入硬件加速的优势

转发面U面引入硬件加速能够提高单服务器吞吐量，降低处理时延、抖动和丢包率，减少各站址U面服务器部署数量，从而降低对机房空间、功耗、散热要求， 也可降低Capex和Opex成本。

CPU计算资源瓶颈使C2服务器模型承载转发面U面业务时造成50G/100G网卡转发资源浪费

基于现有控制面服务器模型，1个物理核可达约1G吞吐量，40个物理核可达约40G吞吐量，50G/100G网卡带宽资源浪费

	纯软	带加速卡
CPU型号	6138	6138
网卡	2块双口25G	2块100G
吞吐量	54Gbps	176Gbps
时延	79.8us	4.42us
丢包率	1.13×10^{-6}	3.71×10^{-10}

空间、功耗要求降低

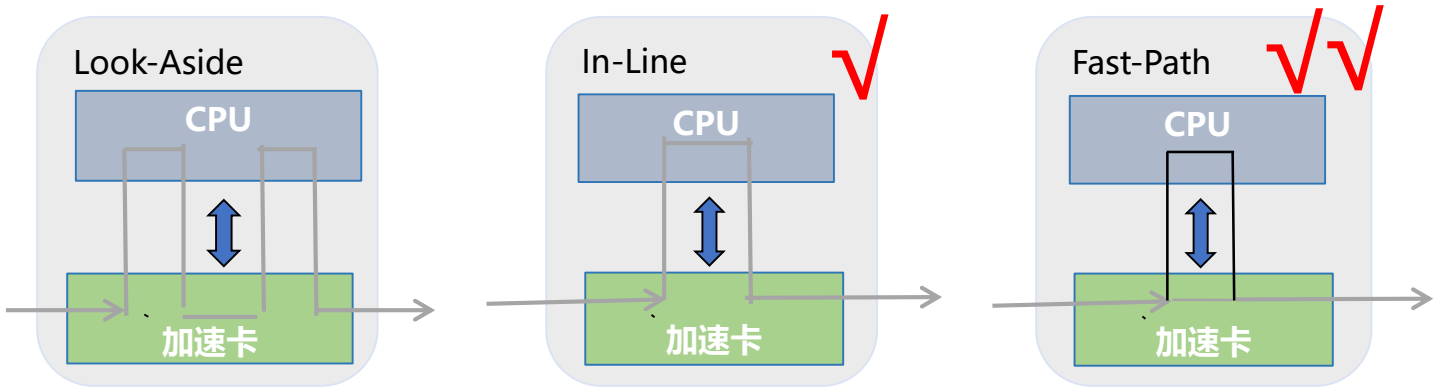
基于现有测试数据，加速卡引入较CPU更新性能提升更高，对空间、功耗要求明显降低。

	C2(5118)	C2(6138)	C2(6138)+加速卡	一体机
计算功耗(W)	105/400	125/420	200/495	-/3980
网卡(G)	25	100(2块双口25)	200(2块100)	-
吞吐量(Gbps)	22.7	54	176	315
高(U)	2	2	2	14
每G功耗(W/Gbps)	17.6	7.8	2.8	12.6
每U吞吐(Gbps/U)	11.3	27	88	22.5

转发面U面引入智能网卡加速

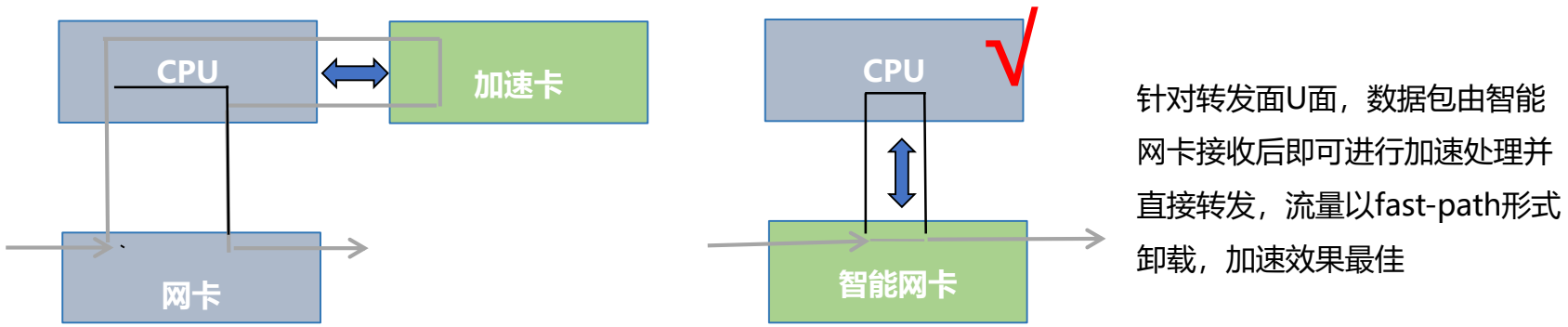
为使加速性能最优，转发面U面引入加速硬件形态选择智能网卡，卸载功能选择满足卸载方式为In-Line或Fast-Path。

卸载方式



Look-aside模式，数据包需要在CPU与加速卡之间多次转发，占用PCIe资源，影响加速性能。

加速硬件形态

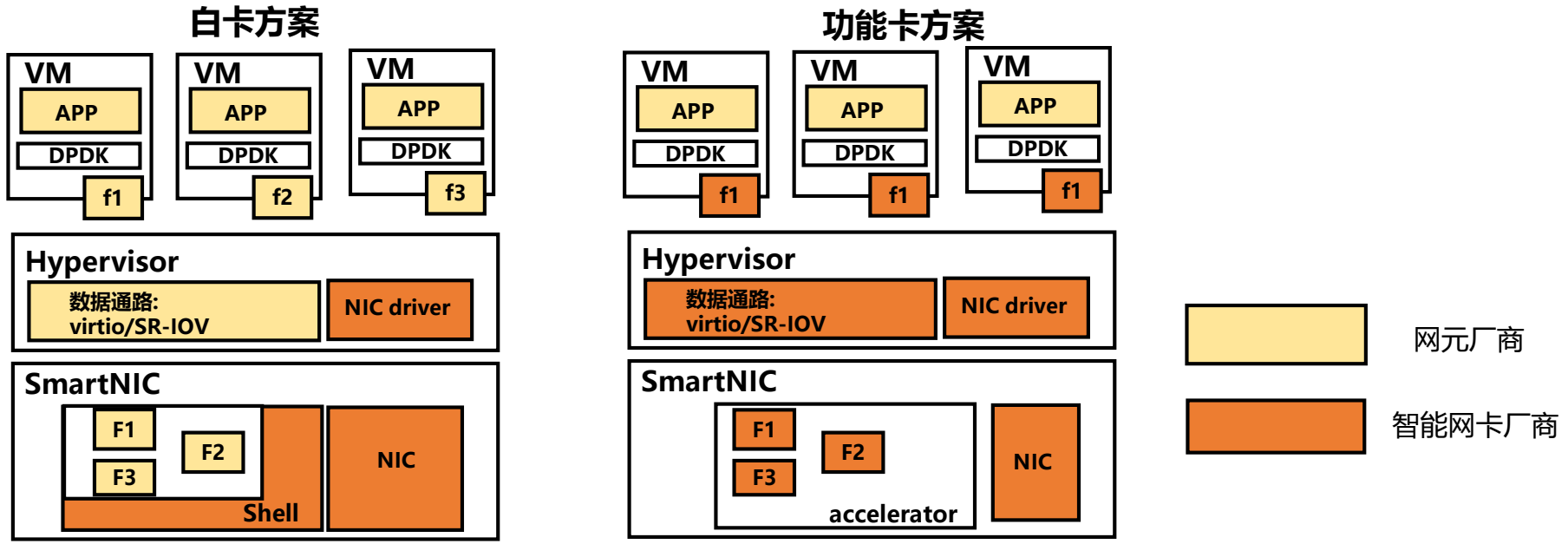


针对转发面U面，数据包由智能网卡接收后即可进行加速处理并直接转发，流量以fast-path形式卸载，加速效果最佳

网元加速功能与智能网卡的解耦方案

网元业务加速存在白卡和功能卡两种解耦方案。对可以通过运营商标准化的方案采用功能卡，对厂商产品差异较大的采用白卡。后续可考虑运营商定制加速网卡的方案，逐步实现加速网卡收敛，降低网元开发和加速网卡标准化复杂度

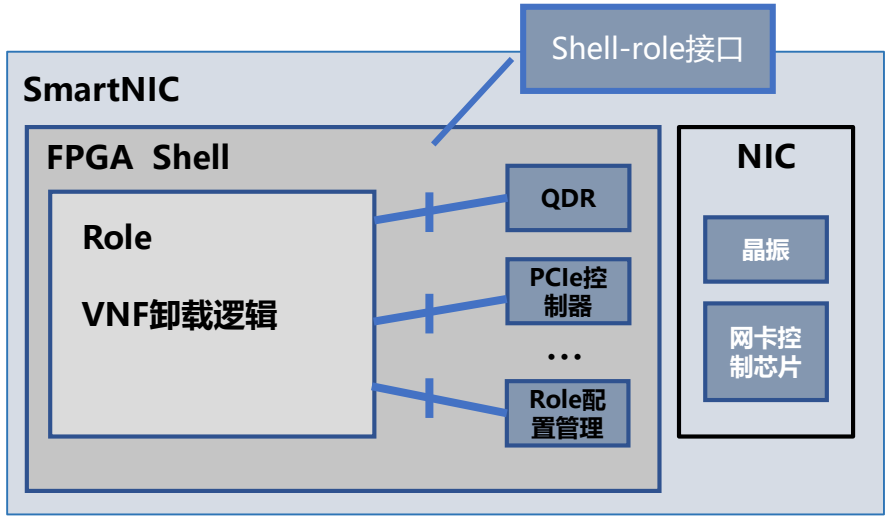
	白卡方案	功能卡方案
网元开发复杂度	高	低
加速设备可选择范围	FPGA、SoC等可现场编程加速设备	FPGA等可现场编程设备及NP、ASIC等不可现场编程设备
解耦规范制定复杂度	较低 需明确加速卡的具体厂商、型号、使用接口、分区方式等	较高 需统一加速功能的所有参数、调用方式



转发面U面硬件加速方案-白卡方案

硬件资源池提供加速卡资源，由网元厂家自主烧写卸载功能。

	FPGA √	SoC
产业	Xilinx（美国）、Intel（美国），产业情况较热	Mellanox （以色列）/Cavium （美国）等；多为ARM架构（英国）
技术特点	可编程，硬件描述语言，Verilog	可编程，软件开发语言
基础原理	SR&PR：PR隔离实现加速功能，PR可动态划分 适合大吞吐量、高并发处理	规格由核数、主频、内存等决定，ARM上可安装开源Linux 擅长控制逻辑，简单重复处理
编排	Cyborg已有	Cyborg尚无
更新	擦写重烧，可灵活部署加速功能	软件刷新，支持热补丁

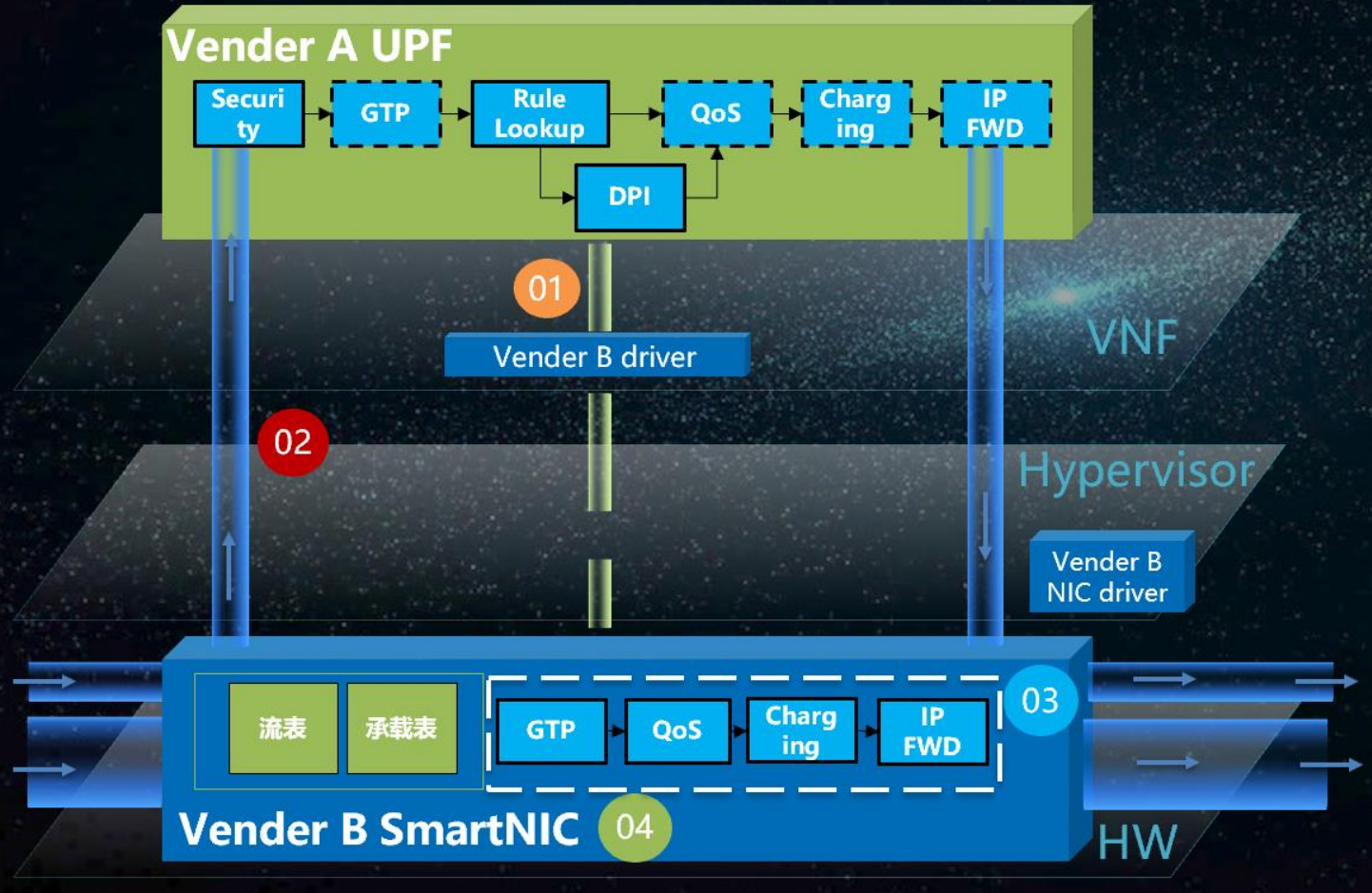


通过定义shell-role接口，shell屏蔽板卡物理接口差异，VNF写入Role的卸载逻辑与加速硬件中板卡电路设计解耦。

中国移动5G UPF边缘开放硬件加速平台方案

业务卸载，提高转发性能

API标准化，实现软硬解耦



01 • 统一功能接口

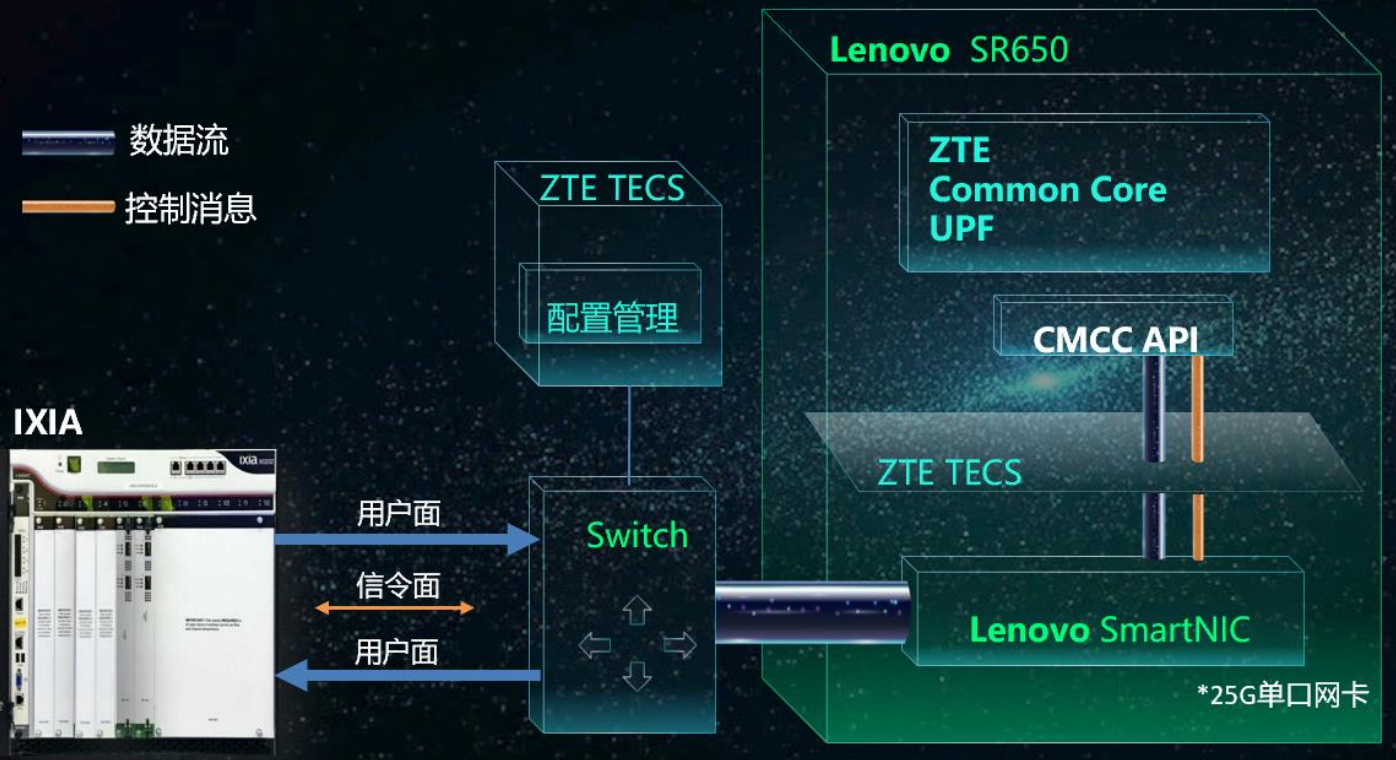
02 • 统一卸载流程

03 • 统一卸载功能

04 • 统一硬件规格

60% 流量卸载

中国移动5G UPF开放硬件加速平台演示方案



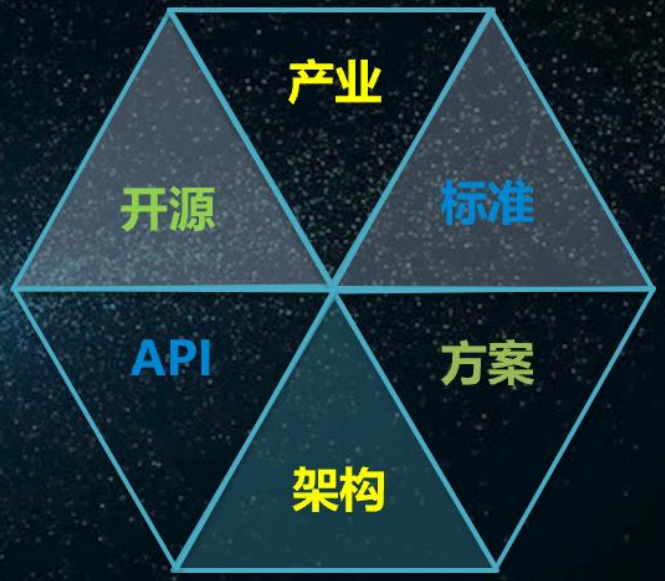
大带宽* 23Gbps

超低时延 6μs

丢包率 10^{-6}

CPU占用率 15%

中国移动转发面网元硬件加速技术白皮书发布

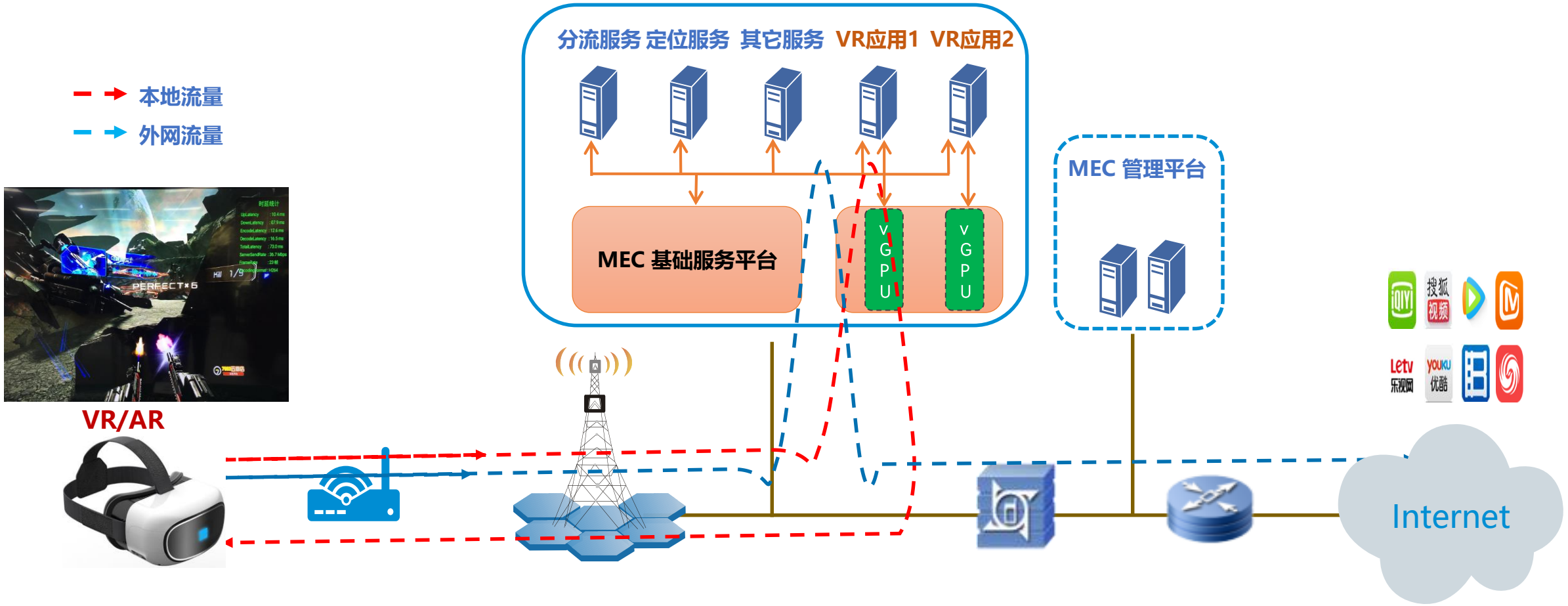


愿与业界携手共建开放硬件加速平台，助力5G发展

提纲

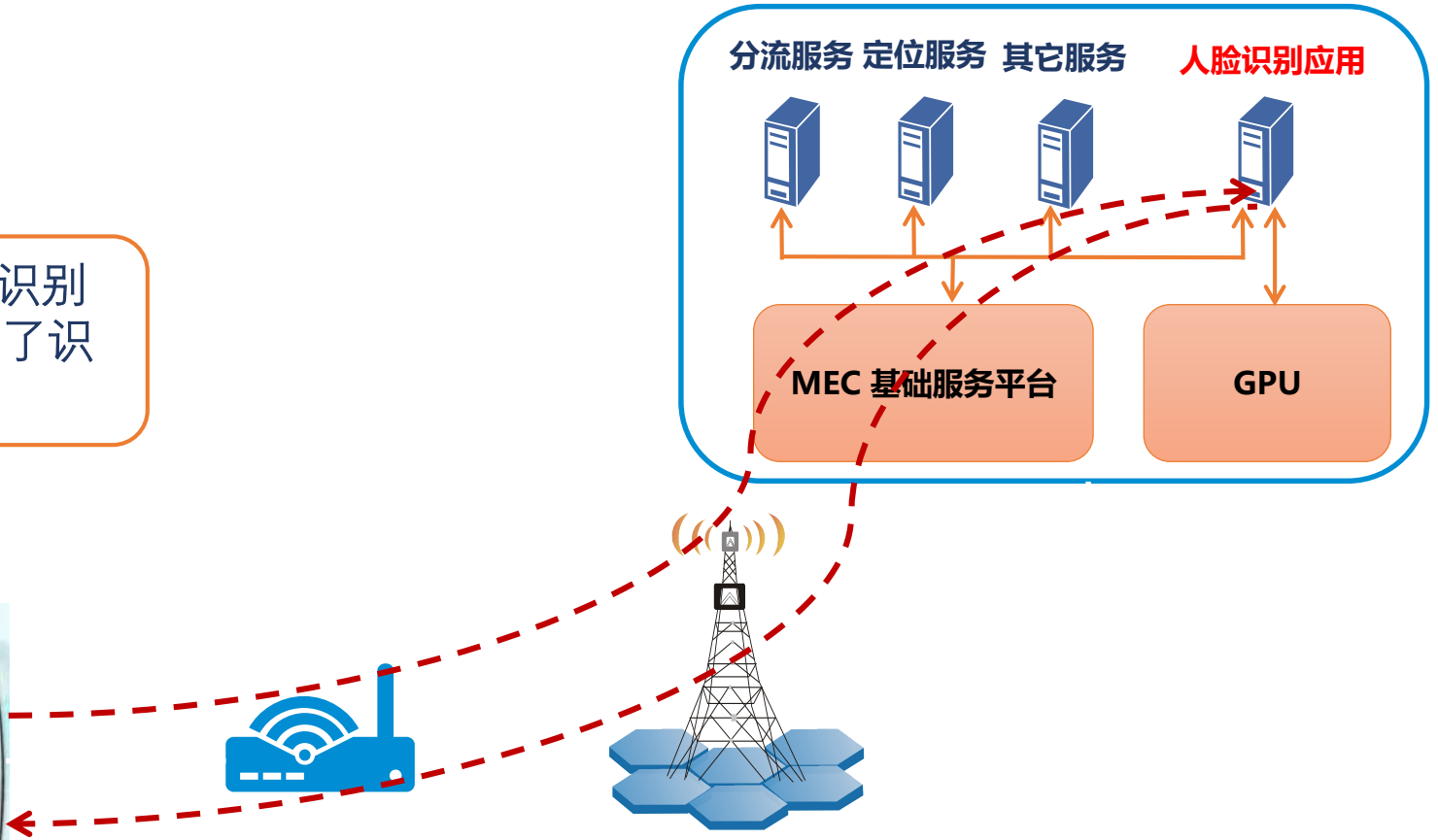
- NFV硬件加速的必要性
- 加速功能分类
- 加速硬件选型
- NFV硬件加速方案
  - 加速资源管理编排方案
  - 硬件加速数据通路
  - 加速通用API和解耦方式
    - OVS硬件加速卸载
    - 网元硬件加速
    - GPU加速
- **硬件加速产业生态和开源情况**
- **下一步工作**

vGPU方式实现VR游戏应用



GPU透传方式实现人脸识别

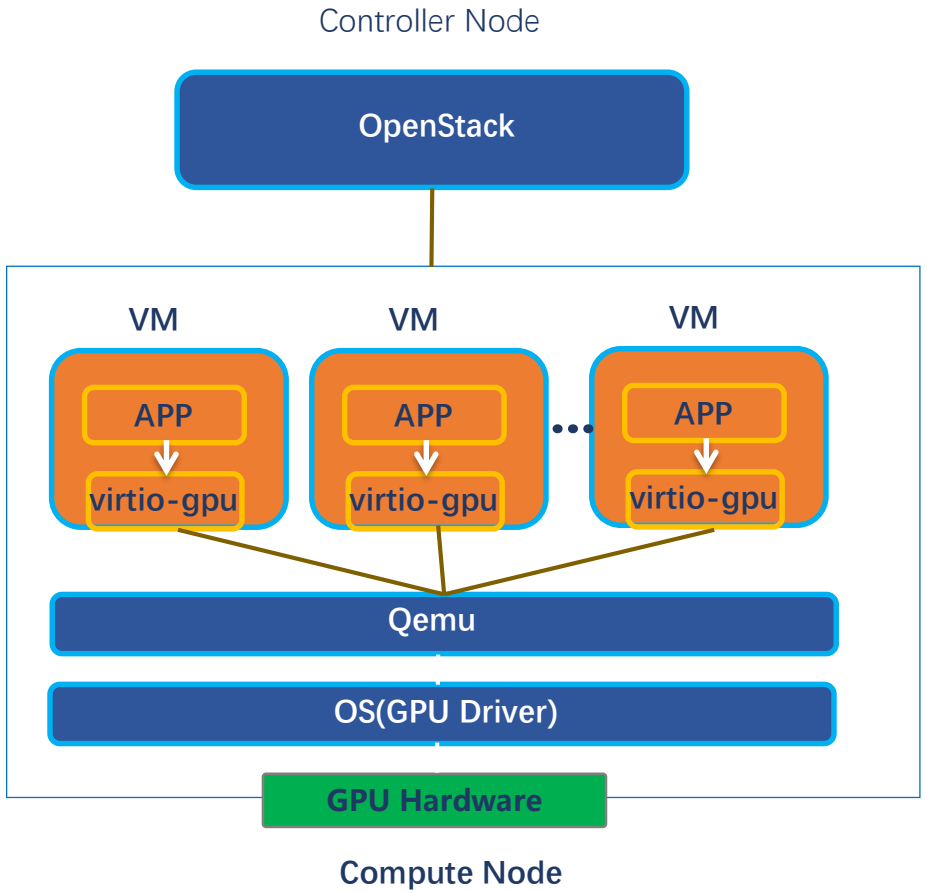
GPU资源透传给人脸识别应用独享，极大提高了识别速度和准确率



GPU使用方式

使用方式	典型应用	说明	特点
GPU共享	机顶盒业务	虚拟机采用虚拟GPU设备，多个虚拟机共用一个物理GPU	成本低，资源利用率高，性能低
GPU透传	高并发计算，深度学习，大数据，图形工作站	GPU卡直接透传给单个虚拟机使用	成本高，资源利用率低，性能高
vGPU(SR-IOV)	高并发计算，深度学习，大数据，图形工作站	GPU卡虚拟成多个vGPU，透传给虚拟机使用	成本高，资源利用率高，性能高

GPU共享



方案：

- 虚机内部采用virtio-gpu设备，通过半虚拟化驱动和物理驱动进行交互，多个虚拟机排队共享硬件GPU资源，提升资源使用率

不足：

- virtio-gpu和硬件驱动之间需要进行调优，提升性能
- Nvidia的GPU驱动相对封闭，对virtio框架无法支持，当前以AMD的GPU为主

适用场景

- 对性能要求低的GPU应用，例如机顶盒相关应用

GPU透传

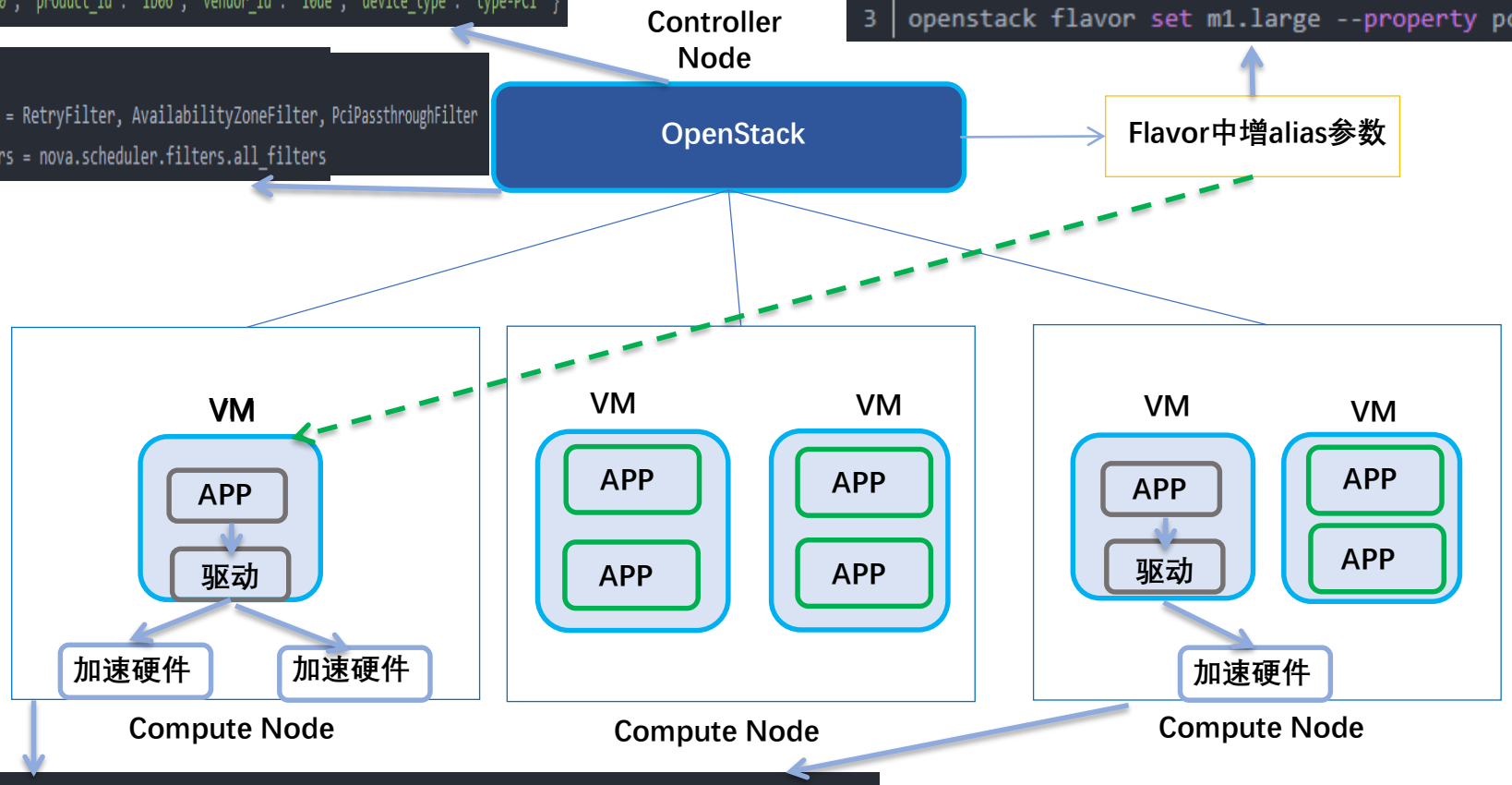
```
[pci]

alias = { "name": "nvidia1080", "product_id": "1b06", "vendor_id": "10de", "device_type": "type-PCI" }
```

```
[DEFAULT]
scheduler_default_filters = RetryFilter, AvailabilityZoneFilter, PciPassthroughFilter
scheduler_available_filters = nova.scheduler.filters.all_filters
```

```
openstack flavor create --public --ram 2048 --disk 20 --vcpus 2 m1.large

openstack flavor set m1.large --property pci_passthrough:alias='nvidia1080:2'
```




```
[pci]

passthrough_whitelist = { "vendor_id": "10de", "product_id": "1b06" }

alias = {
        "name": "nvidia1080",
        "product_id": "1b06",
        "vendor_id": "10de",
        "device_type": "type-PCI"
}
```

方案：

- GPU卡透传给某个虚拟机使用，虚拟机完全使用该GPU卡，提升此虚机性能

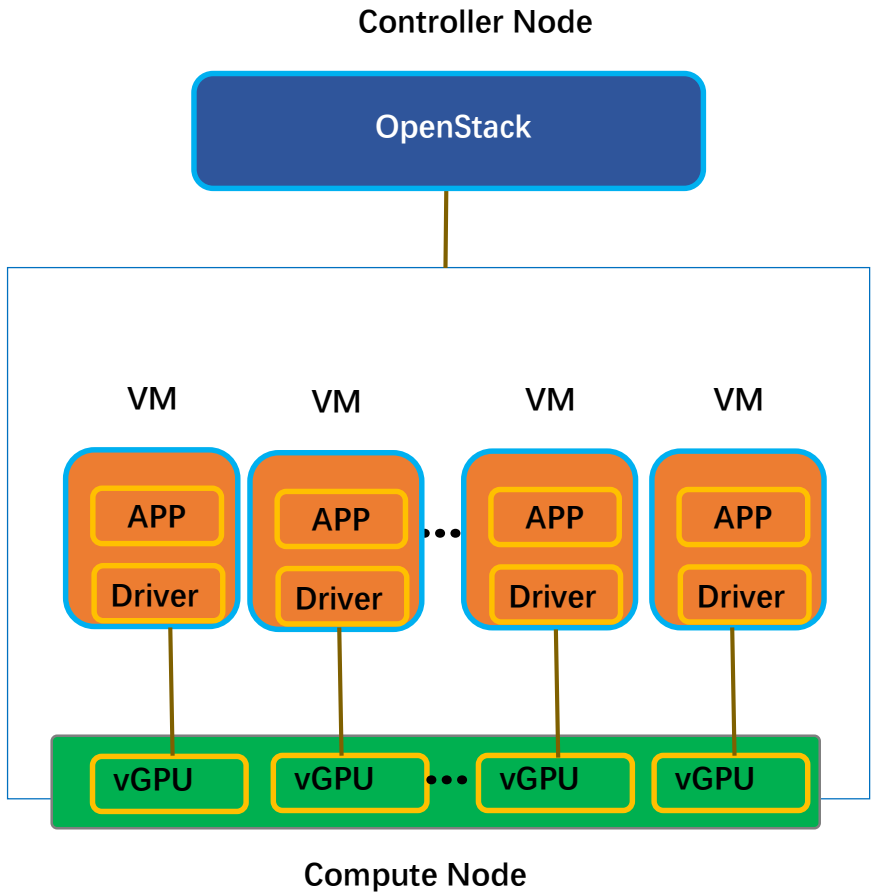
不足：

- GPU只能给一个虚拟机使用，同节点其他虚拟机无法使用，GPU使用效率低

适用场景

- 对GPU性能要求较高且使用固定的场景

vGPU方式



方案：

- GPU虚拟化成多个vGPU，vGPU透传给虚拟机使用

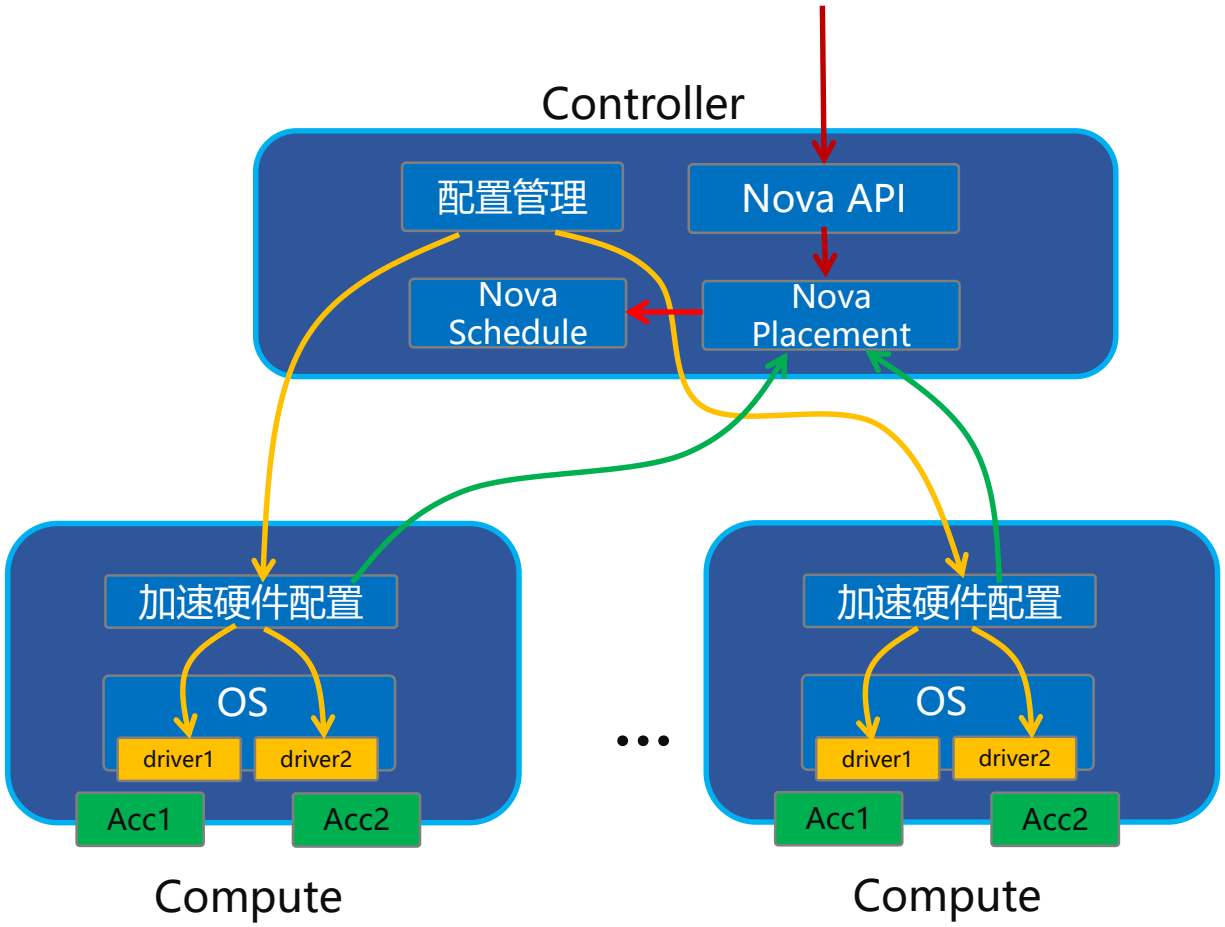
不足：

- 每个计算节点仅支持设置一种模式（Nvidia的GPU卡）
- 每个虚拟机仅支持配置一个vGPU子卡
- 主机操作系统要求在Cent OS7.5以上

适用场景

- 对GPU性能要求较高且同一主机多个虚拟机都需要GPU功能

vGPU硬件管理一阶段



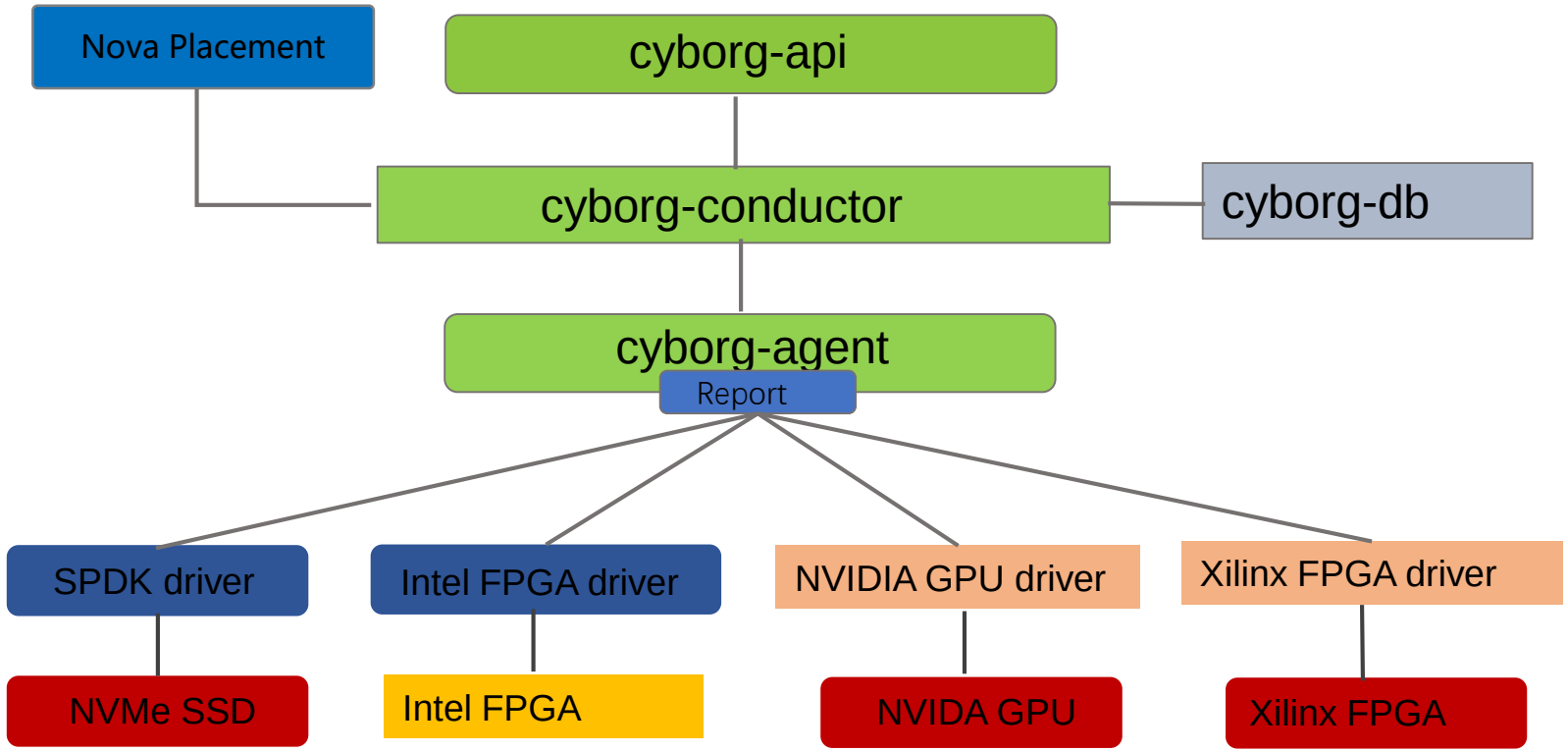
方案介绍

1. OS安装时预先包含GPU硬件的驱动，检测到硬件时自动进行驱动加载
2. 通过配置管理界面进行配置，每个主机上驻留配置模块完成加速硬件的配置，并上报能力至Nova Placement
3. 虚拟机部署时在flavor中设置需要的加速硬件资源，通过Nova Placement过滤到有加速硬件的主机然后通过 Nova Schedule选择进行虚拟机部署

缺陷：

1. NFVO无法感知加速能力，因此部署时指定到对应的VIM进行部署，且无法获取VIM中加速硬件使用情况
2. 无法进行配额管理等操作，只有部署失败时才能感知资源不足

vGPU硬件管理二阶段



方案介绍

1.采用Cyborg管理加速硬件，完成硬件的发现、资源上报，资源管理等功能

2.硬件的驱动加载和配置功能也需要由Cyborg来完成

3.虚拟机部署时在flavor中设置需要的加速硬件资源，通过Nova Placement过滤到有加速硬件的主机然后通过 Nova Scheduler选择进行虚拟机部署

提纲

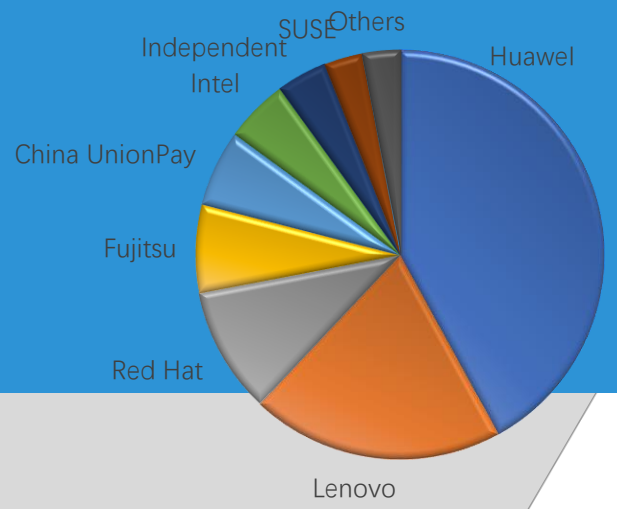
- NFV硬件加速的必要性
- 加速功能分类
- 加速硬件选型
- NFV硬件加速方案
  - 加速资源管理编排方案
  - 硬件加速数据通路
  - 加速通用API和解耦方式
    - OVS硬件加速卸载
    - 网元硬件加速
    - GPU加速
- **硬件加速产业生态和开源情况**
- **下一步工作**

加速编排管理——OpenStack Cyborg

cyborg距离可商用尚有很大距离，迫切的需要运营商和厂商共同推进cyborg成熟、打通NFV加速管理编排流程，实现硬件加速的管理编排。

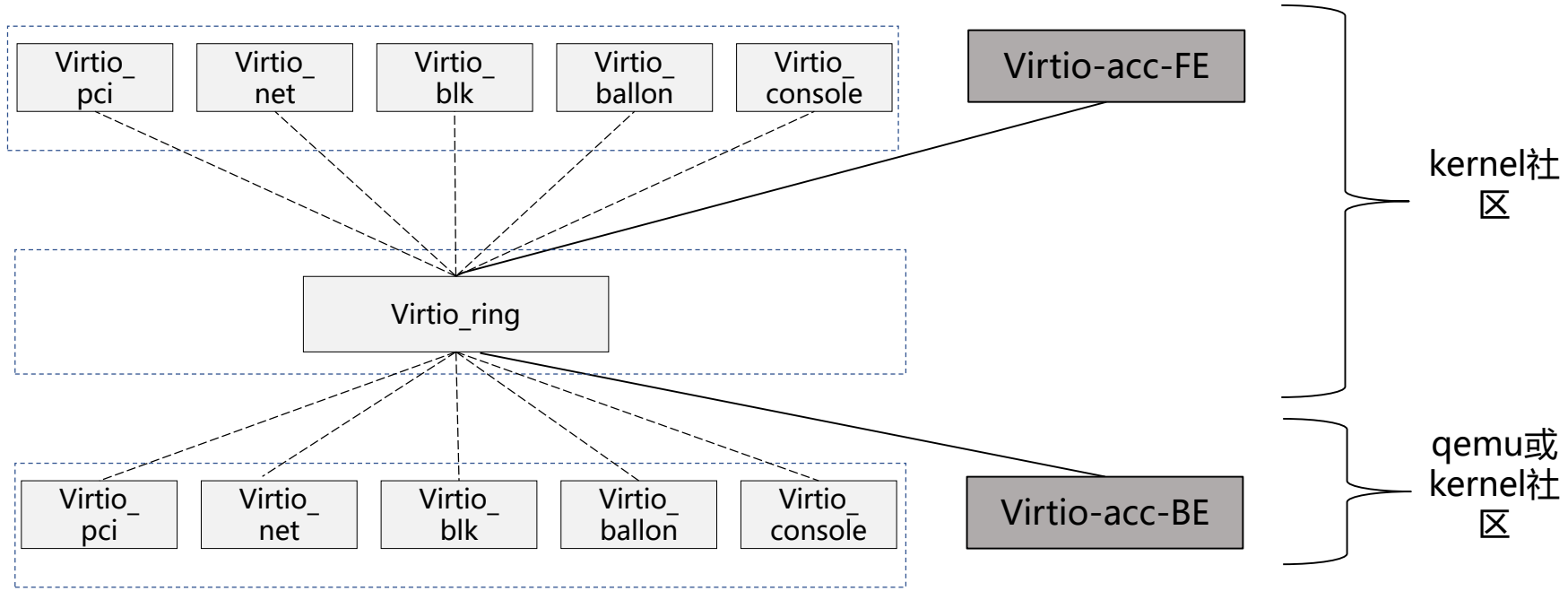
cyborg

- 是加速管理面唯一开源项目，厂商关注度很高，业界影响力很高，但实际参与率少（主要参与投入的公司华为、Intel、联想、红帽、Nokia、ZTE、Xilinx，截止R版，但每个公司仅投入2-4人，贡献统计参见饼状图）；
- 目前很不成熟，功能实现缓慢，已经发布的版本Q、R、S版本实现功能很少，T版本（2019年10月）路标继续打通与NOVA的交互，且目前仅关注FPGA、GPU，**距离可真正商用尚有很大距离。**
- **急需厂商深入投入，共同推动cyborg成熟。**



加速数据通路——Virtio

virtio框架修订涉及virtio、kernel和qemu社区，新增加速通道的周期约为3年，较为缓慢。一方面应考虑使用当前已有路径（如virtio-net）实现加速功能，一方面推动社区关注加速的需求并加快新增通道的支持



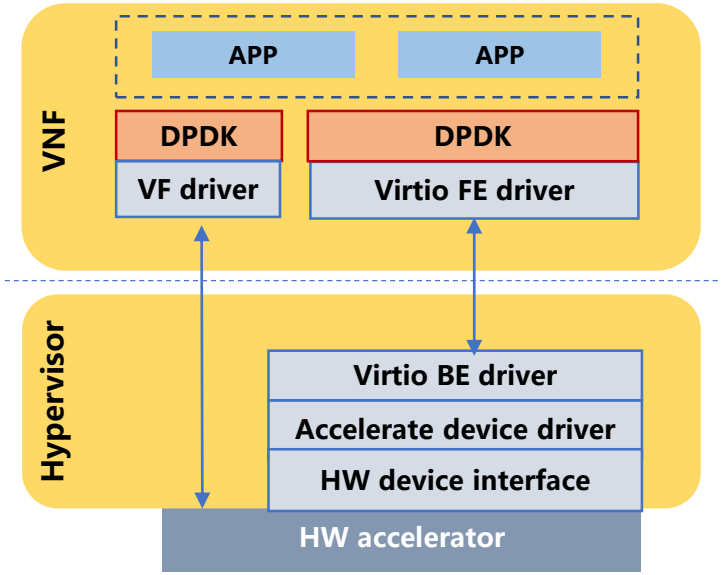
OASIS的virtio社区是virtio协议的标准组织，定义了virtio公共部分，virtio blk、virtio net等具体设备驱动的实现都会用到该协议。

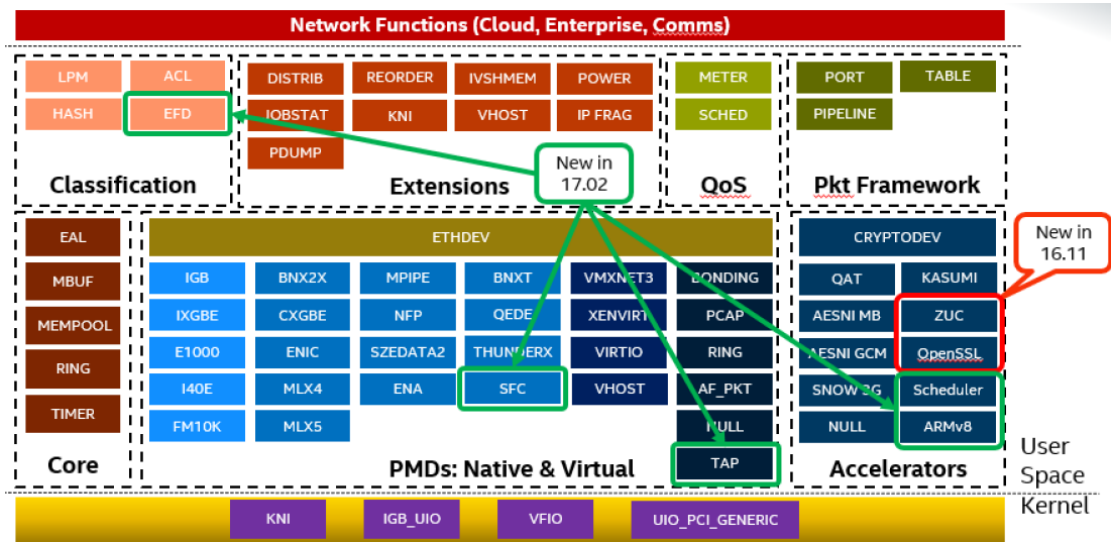
Virtio框架主要维护人员：

- Michael S. Tsirkin <mst@redhat.com>
- Jason Wang <jasowang@redhat.com>
- 虚拟化负责人列表：virtualization@lists.linux-foundation.org

加速API——DPDK

DPDK(Data Plane Development Kit)是一套基于通用硬件加速数据包处理工作的软件lib库和驱动。可以实现VNF调用加速功能的接口定义或封装。目前DPDK内对网络加速已有相关工作。需进一步推动细化和加速抽象层的开源





intel CAVIUM arm Mellanox TECHNOLOGIES XILINX ALL PROGRAMMABLE HUAWEI ZTE Alcatel·Lucent ERICSSON BROADCOM

vSwitches
VPP
OvS Open vSwitch
BESS
Lagopus

DPDK in OS Distros
redhat. Version 7.1 +
FreeBSD Version 10.1 +
ubuntu Version 15.10 +
WIND Version 6 +
CentOS Version 7.1 +
fedora Version 22 +

vRouters
OPENCONTRAIL
CloudRouter
VPP

Packet Generators
TRex
Pktgen
MoonGen
Ostinato

Storage
Storage Performance Development Kit

TCP/IP Stacks
mTCP
TLDK & VPP
Seastar
LWIP DPDK

+ Many more

加速集成和测试——OPNFV Rocket

中国移动于2018年推动OPNFV Rocket项目成立，旨在以此项目聚集社区和行业力量，统一需求，推动数据面加速API实现和开源。厂商参与度较高，参与厂商有华为、中兴、联想、INTEL、NOKIA、ARM、MELLANOX和WINDDRIVER等，该项目目前处于需求讨论阶段。

Rocket 目标

- 定义数据面加速API，屏蔽底层硬件差异，以实现软硬解耦
- 加速API需求文档和加速器文档
- 加速器测试
- 提供其他加速问题的解决方法

Rocket 版本计划

H版本发布计划

- 网元加速需求文档
- 通用加速API定义

Step 1:

- OVS社区测试
- GTP卸载需求定义和GPT加速API定义

中国移动硬件加速时间计划

- 完成OVS卸载方案实验室测试，并与SR-IOV、软件OVS加速方案进行对比
- Hypervisor与OVS硬件加速网卡不解耦方案基本成熟，在下半年边缘商用试点中验证

- OVS硬件加速方案商用试点测试完成
- GPU加速方案商用试点测试完成
- GTP卸载接口规范完成，推动业界产品成熟，准备试验网测试

- 启动GTP等功能卸载试验网测试
- 研究IPSEC,Checksum等卸载API的制定

- GTP等功能加速卸载方案基本成熟，启动商用试点
- 完成其他功能卸载的技术规范与方案的制定，推动产业成熟，开展实验室测试

谢谢！

Email：
wangshengyjy@chinamobile.com