2015中国DPDK开发者大会

China DPDK Summit 2015

Presented By:

中国移动 China Mobile　中国电信 CHINA TELECOM　HUAWEI　intel　ZTE

基于英特尔ONP构建虚拟化的IP接入解决方案

中国电信广州院研究院

2015年4月

中国电信 CHINA TELECOM

内容

CONTENTS

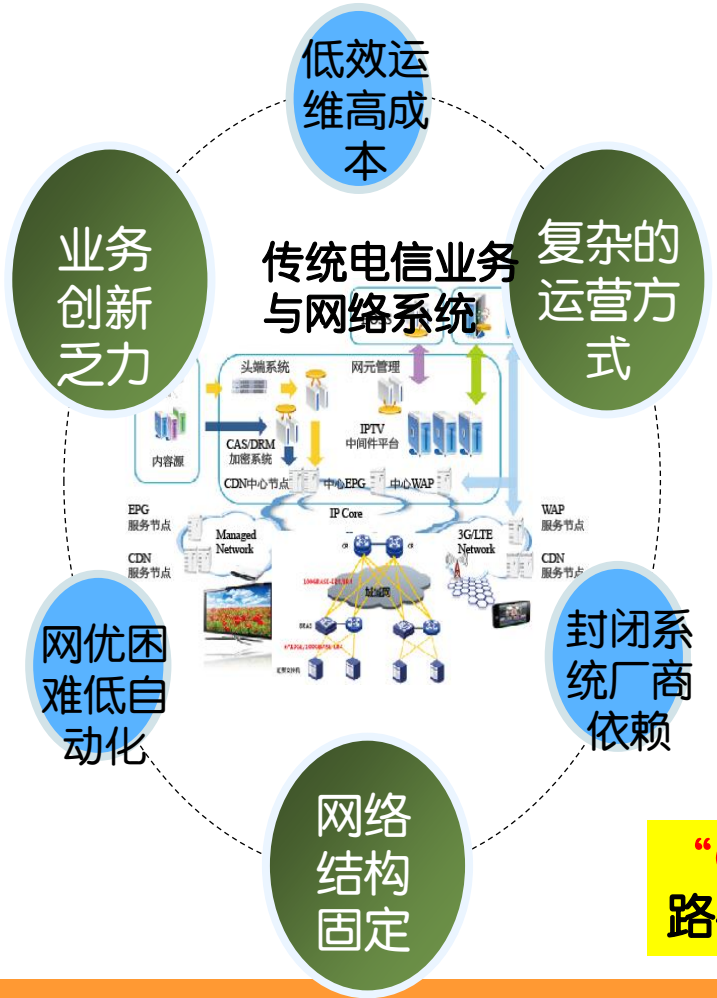
01 NFV技术发展动力

02 DPDK相关技术分析与实践

03 后续工作

SDN/NFV为电信网络升级转型带来新契机

- 传统的网络架构和产业格局，阻碍了业务创新和网络高效运营
- 需要开放的产业链、开放的网络架构，解决长期以来“运营商需求－>厂家标准－>厂商设备－>运营商测试–>部署”模式下的路径依赖



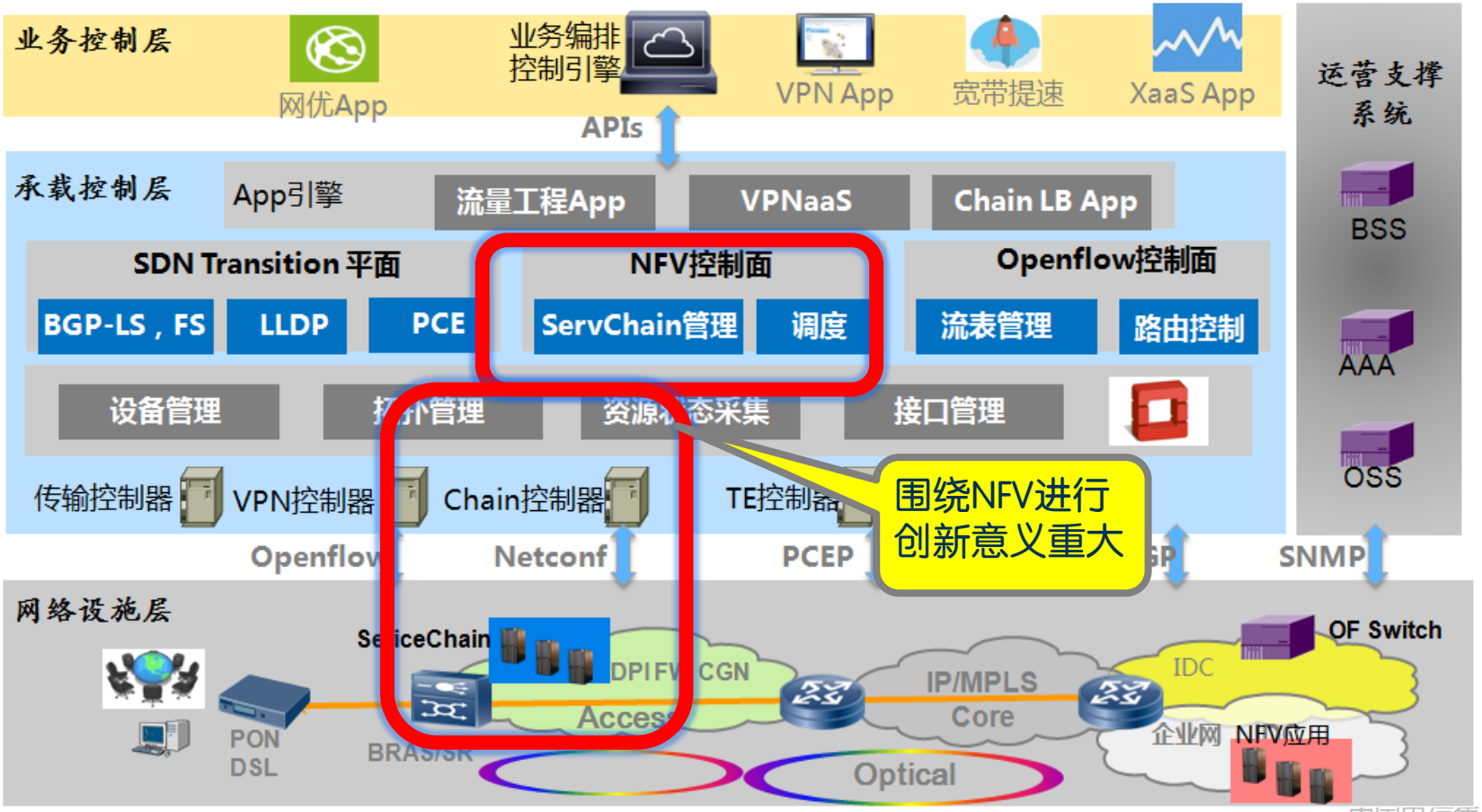
向SDN演进

- 提升资源利用率
- 快捷网络服务
- 网络部署扩展性
- 更低设备成本
- 流量全局优化
- 部署自动化

“Open+虚拟化+网元通用化”是解耦产业链中路径依赖的关键技术

NFV是运营商在SDN领域的重要布局

- NFV未来将对运营商网络架构和运营模式产生重大影响
- 业界普遍认为：基于x86通用架构，NFV可以更方便、更迅速地调整网络服务，节省成本，加快新业务的发展



NFV技术的应用场景

- **NFV目前主要用于IP RAN、移动核心网、IP edge、IDC内部/出口**
  - 硬件设备的服务器化、功能软件化是重点
  - 网络资源虚拟化相关标准、开发技术是热点
- **Service Chaining是最受关注的NFV应用，包括**
  - 固移融合：Firewalls, DPI, CGN, Load Balance, Video TransCode等等
  - 宽带IP网：Cache/CDN、VPN、vCPE、DNS
  - 移动核心网：3GPP EPC, eNodeB, S/PGW, MME, App acceleration

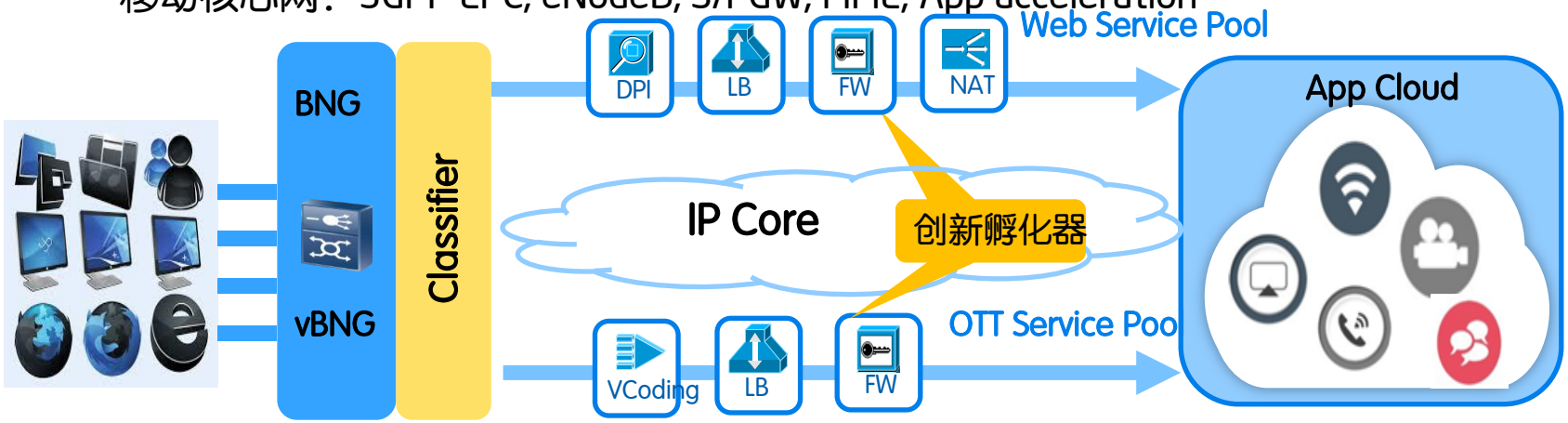
宽带网络的核心竞争力

↑

智能边缘 ServiceChain

↑

创新型业务孵化器



传统IP Edge设备面临的问题

- 智能边缘是城域网接入控制的L3终结点和业务POP点，负责宽带用户拨号接入、专线用户接入、组播以及L2/L3VPN承载，其能力直接决定了城域网智能化水平

- 目前BRAS设备业务功能少、新业务板卡开发/部署迟缓
  - 内置DPI：5年未实现产品实用化
  - CGN：花费3–5年研制与部署
- 部署缺乏灵活，多厂家无法兼容
  - 厂家间板卡无法共用

严重阻碍宽带业务快速发展

- 性能横向扩展能力不足、业务单一、新业务部署难
  - 单板卡性能受限，DPI仅达25%的整机性能
- 不具备承载云化资源、云IDC等新业务能力

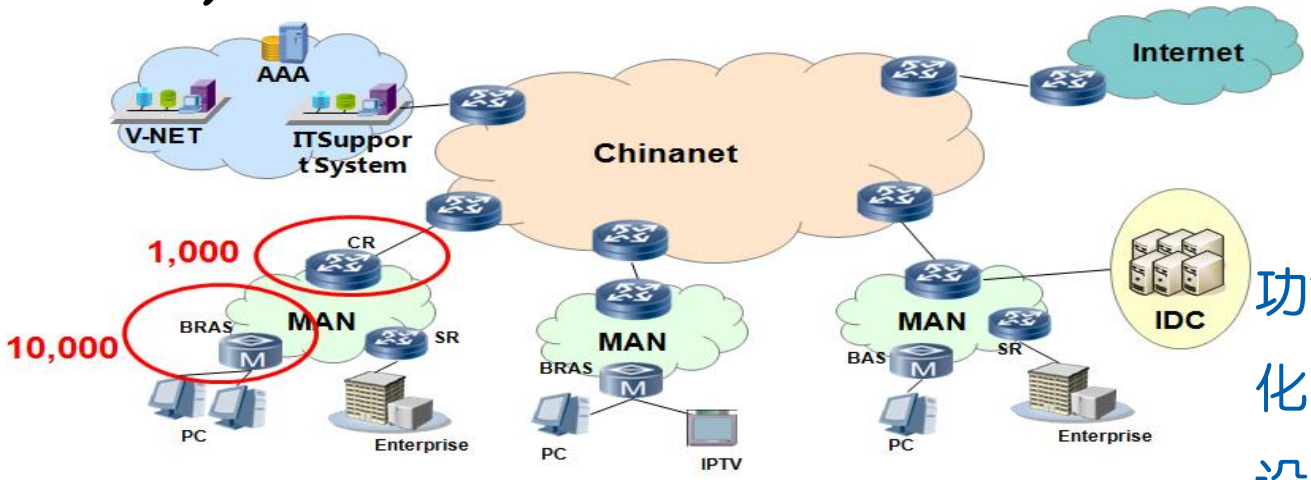
功能项	项目
路由处理能力	OSPF
	ISIS
	BGP
用户接入能力	PPPOE接入速率
	PPPOE并发用户数

功能项	项目
PPPoE用户接入	PPPoE上下线
	用户接入认证
	PPPoE用户计费
用户QoS	用户带宽限速(CAR
	用户带宽调整（COA）

功能项	项目
CGN	CGN管理
DPI	DPI管理
FW	FW管理
其它	

IP Edge向NFV的演进思路

- IP edge设备在IP城域网中数量庞大，系统容量方面与X86服务器转发能力接近（40G–400G）



功能软件化、虚拟化

设备通用化

控制转发分离

- 演进技术路线

运营商对NFV技术的需求

- NFV Controller+编排器：对物理/虚拟化资源池进行调度与编排，提供服务开放和应用接口
- NFV基础设施：提供一致的设备环境支撑VNF运行、高性能灵活转发能力、Chaining能力

仅有MANO是不够的

NFV

X86平台线速转发问题

- 端口密度受限、单网卡转发能力?
- HOST与VM中I/O性能瓶颈
- vSwitch的性能问题
- Service Chain负载均衡

热点技术

- MANO
- Service Chain
- X86性能优化
- vApp模块

重点关注

- Chain基础设施、控制器
- 开放的vApp模块
- 转发平面灵活适配

转发灵活性问题

- 如何应对大量的封装转化
- 如何适配未来新定义封装
- 能否实现Encap的可编程能力

内容

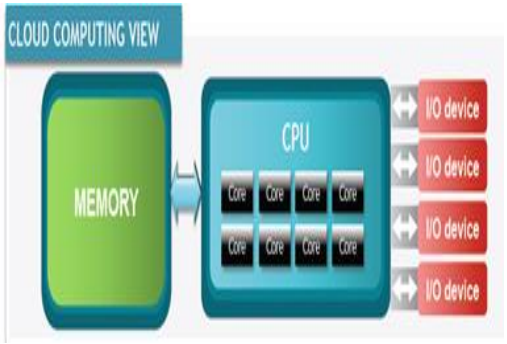
CONTENTS

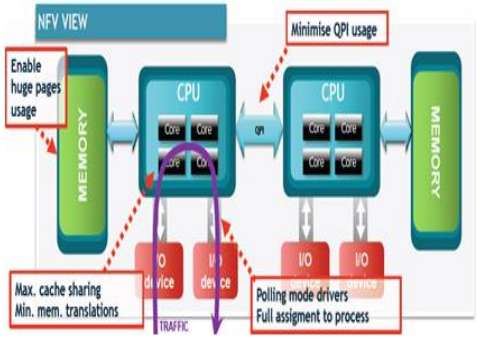
01 NFV技术发展动力

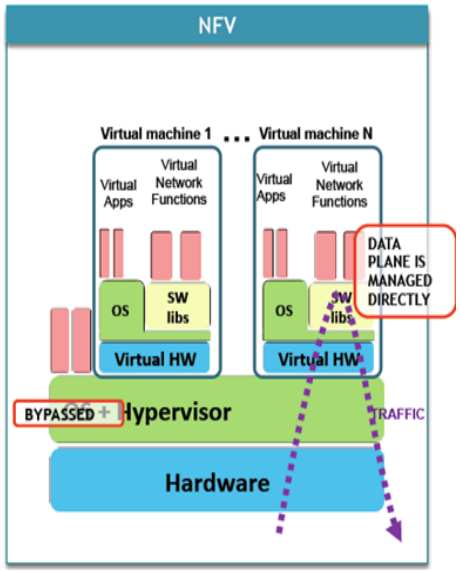
02 DPDK相关技术分析与实践

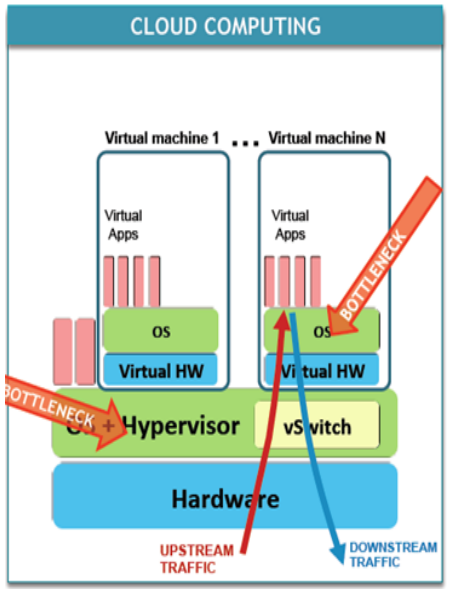
03 后续工作

NFV在x86环境的应用特点









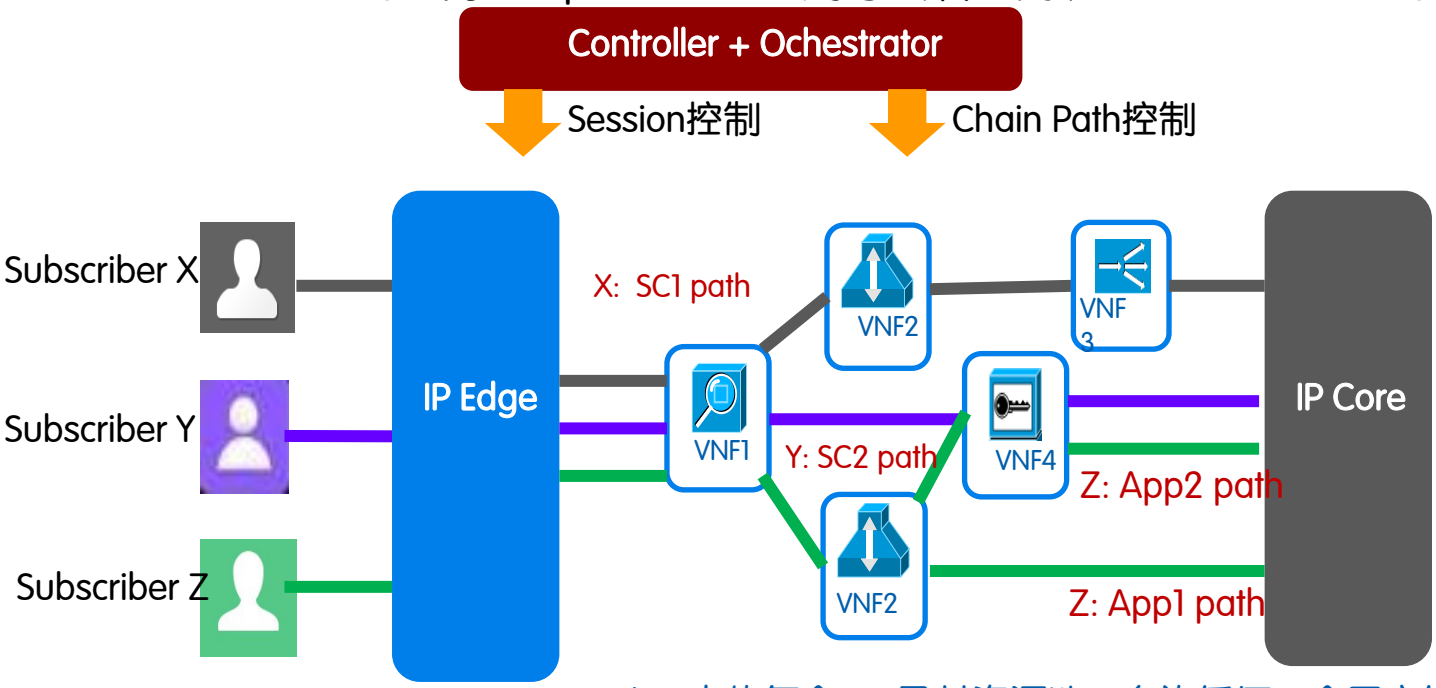
OpenMano，by Diego Lopez, Telefonica, NFVRG, 2015 Mar, IETF 92th

- 云计算场景
  - 各类VM对资源的使用是平等的，共享Core、存储、I/O
- NFV场景
  - 为提升处理能力，需要优化I/O，消除从物理nic到VM中TCP/IP协议栈的处理瓶颈
  - 跨CPU QPI会带来性能瓶颈，需要考虑亲和性：CPU、NUMA和I/O

- HOST优化：需要DPDK类加速技术旁路hypervisor和kernel，优化vSwitch性能
- VM/VNF优化：需要对虚拟HW中I/O与kernel优化，VNF针对DPDK API进行移植

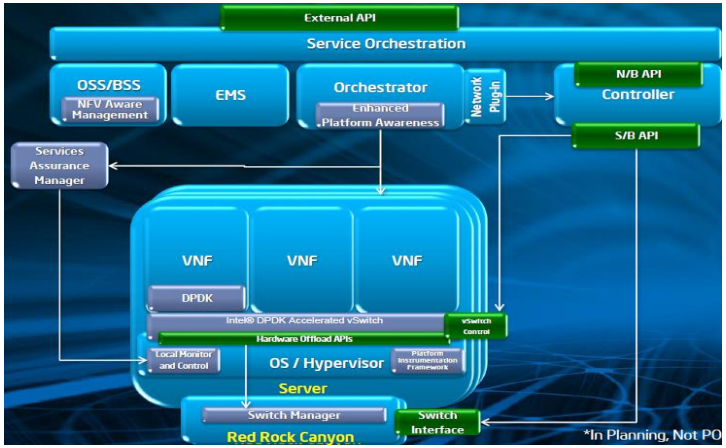
Service Chain在IP Edge中的流量特征

- IP边缘设备采用NFV技术的转发特征
  - 高性能、海量用户Session级控制：单节点数万用户session，40G–400G吞吐量，数百万Session
  - Subscriber Session与Tenant流量的区别：对VNF的时延、带宽要求更高
  - Service chain的业务VNF path：VNF组网与负载均衡、VNF直连WAN时的路由问题



- IP Edge 中的每个VNF及其资源池，允许任何一个用户使用
- VNF及其管理配置，必须Subscriber Session Aware

英特尔ONP参考设计平台的优势



对NFV虚拟化提供了有力支持

- 支持NFV的IA架构
- IA CPU 及其虚拟化VT技术
- 开源的DPDK/SR-IOV的加速技术
- 高性能NIC ASIC、加速Chipset / SoC
- 开放的网络软件、商业网络协议栈

基于ONP平台构建NFV基础设施

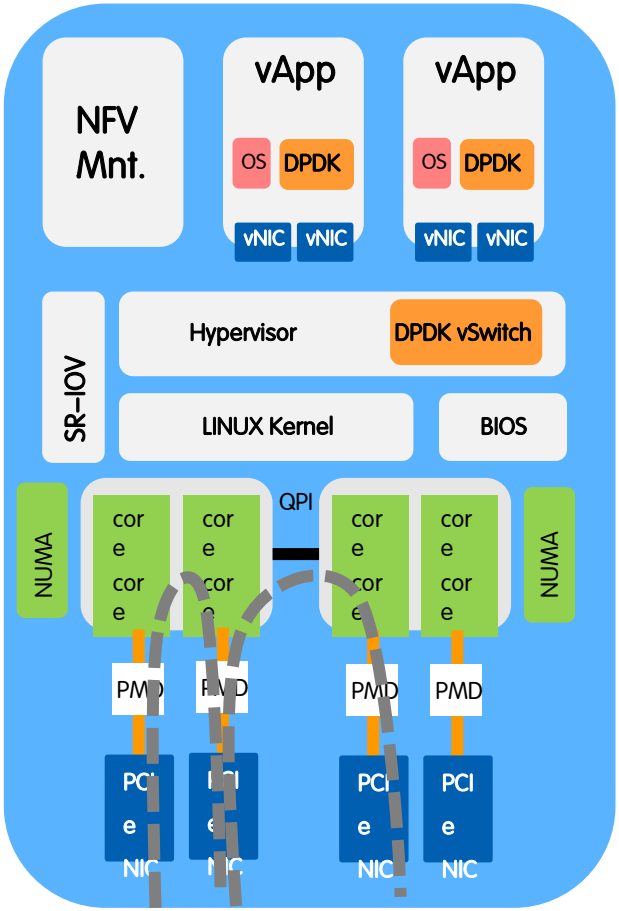
- 开放的硬件交换机参考设计
- ONP服务器参考设计
- DPDK的I/O加速、Open vSwitch加速能力
- 极大提升Service chain中end-to-end转发能力

可以帮助运营商快速构建各类NFV原型应用

DPDK 技术应用中的典型问题

- **单服务器DPDK性能影响因素**
  - CPU版本、CPU主频的性能差异
  - CPU core分配方案：OVS/OVDK/VM
  - Huge Pages尺寸的影响、BIOS配置
  - 跨QPI与NUMA通信问题
  - 流量中的flow个数、vSwitch流表条目

- **DPDK在NFV中应用带来的相关问题**
  - 不同厂商机型、配置项性能的稳定性、一致性
  - 大量VNF需要基于DPDK进行移植
  - DPDK接管I/O后，L2/L3的路由与拓扑管理
  - Service Chain中对业务流量的调度方法



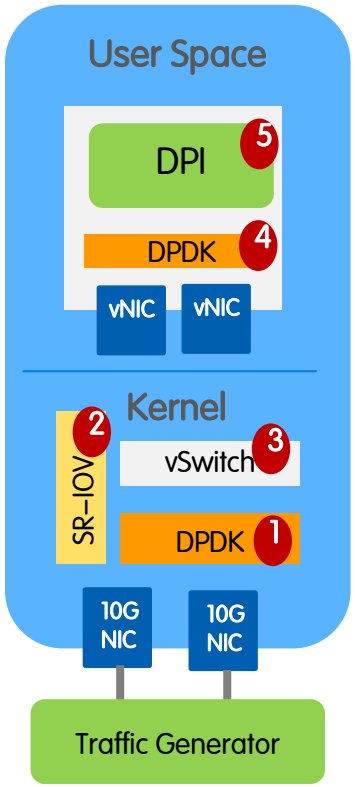
https://git.opendaylight.org/gerrit/#/c/18404/

DPDK性能测试部分结果

环境与配置

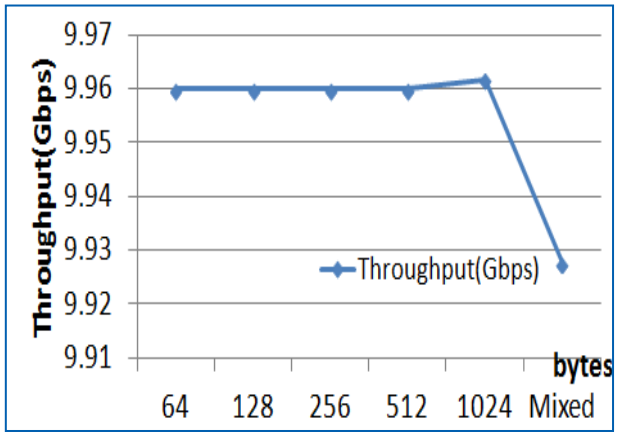
配置： 2 Xeon(R) CPU E5-2699 v3, 2.30GHz，18核/cpu, 64G DDR4, 82599 NIC

方法： RFC 2544

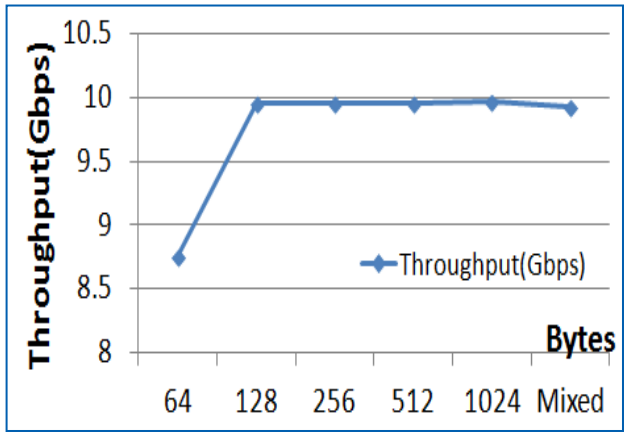


60	120	240	576	1500
45%	23.50 %	15%	11%	5.50%

Host + DPDK 1

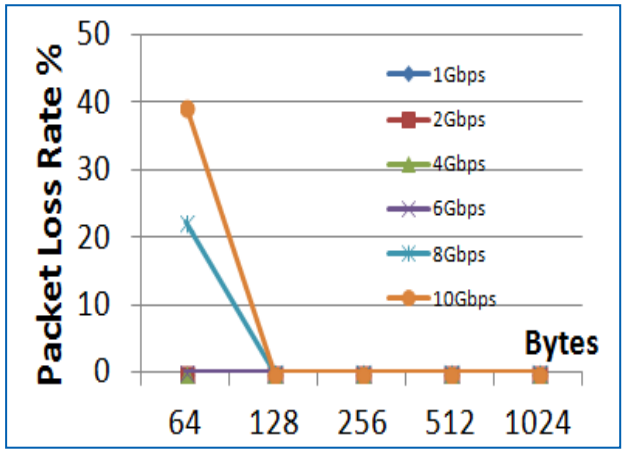


Host + DPDK+跨NUMA 1

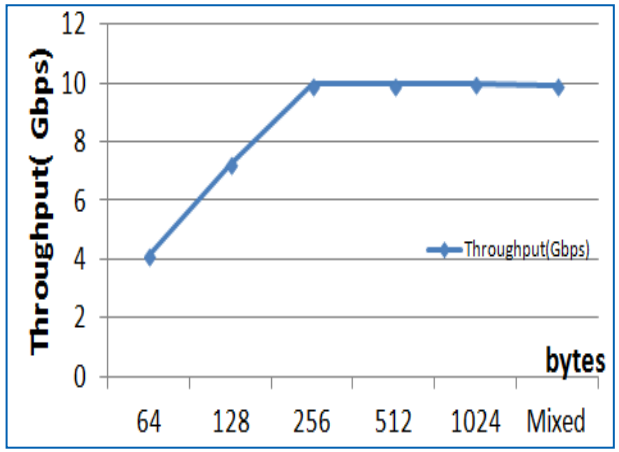


Host + OVDK 1 3

OVDK 4cores， 8192*2m hugepages



SRIOV+DPDK+DPI(VM) 2 4 5



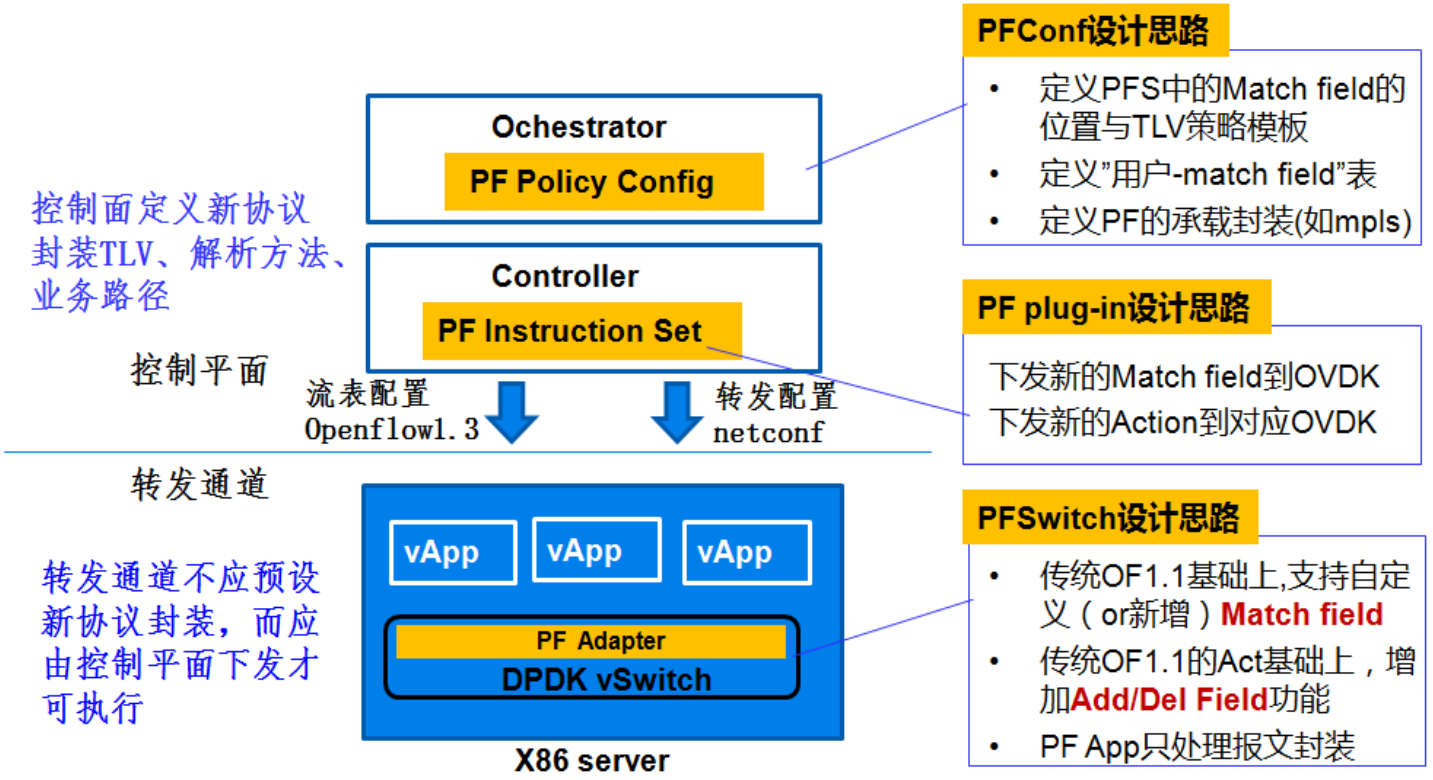
（非正式测试结果，部分参数仍待优化中）

基于DPDK建立NFV基础设施参考模型

- DPDK的实用性分析
  - X86内部转发延迟可以忽略不计，一般延迟均<50us，极端拥塞时的延迟不超过1ms
  - 非全部64B小包，在无vSwitch时，各类DPDK转发方案基本接近线速（9.9G上下）
  - 100%的64B小包无损转发并无实用价值，也要看混合包长（Mixed）测试结果
  - **NFV步入实用化的标志：DPDK技术可以保证“物理NIC – 》VNF” 转发速率达到70%线速**

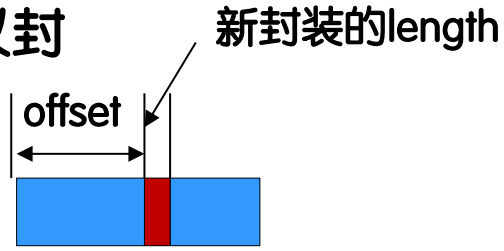
基于DPDK实现可编程转发架构（PFA）

- **可编程转发架构（PFA）：由中国电信提出，它是为提升网络封装灵活性而提出的一种SDN转发技术，类似的技术包括PIF 和PoF**
  - PFA支持抽象的网络转发行为“定义+执行”架构，支持任意格式的报文封装和按需转发
  - PFA可以基于物理Switch实现，也可以基于各类vSwitch实现，目前已基于OVDK



基于DPDK实现可编程转发架构（PFA） *Cont.*

- **需要在Match Field中支持试验类型的协议封装**
  - 识别某类已经定义的新封装
  - 新封装的定义由controller传入参数



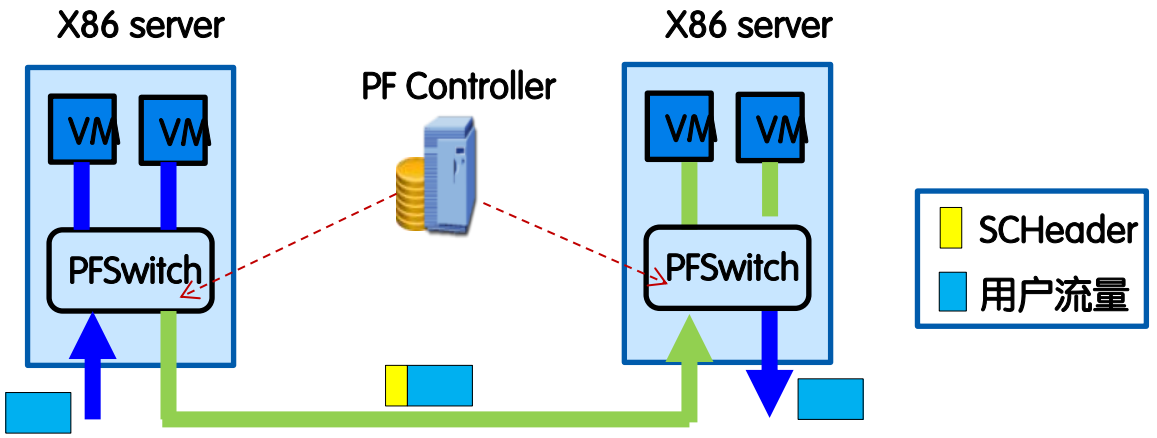
- **开源OVDK基础上，Act Set(OF1.1)/Instruction(OF1.3)中增加3类操作**
  - Add Field(offset, type, length, value)：允许OVS添加新字段的能力
  - Set Field (offset, type, value)： 修改新封装的内容（如果为OF1.3，无须新增）
  - Del field(offset, type)：移除某项新封装

Match Field	Action	说明
10.10.1.0/24	Add Field (100, TEST_META, 64, 0x25)	将10段用户添加试验封装
TEST_META=0x3C	OFPAT_OUTPUT	后续端口发现TEST_META匹配项，跳转到指定端口转发

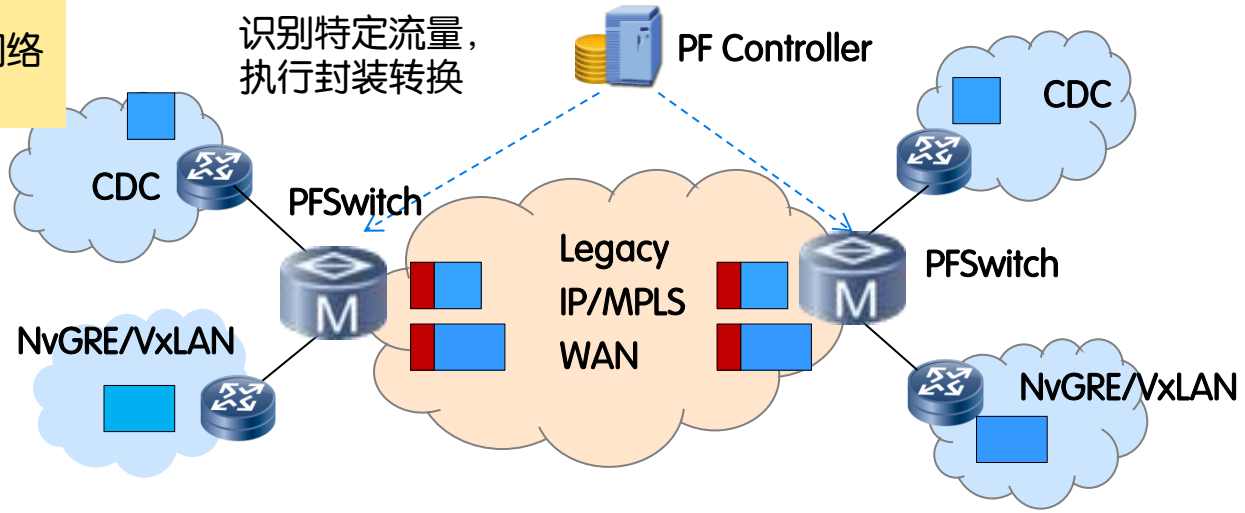
后期将随着DPDK向OVS– netdev版本迁移，支持标准OVS操作及其扩展

PF Switch应用场景

Service Chaining
- 在Service Chaining中提供Header(NSH)功能
- 基于NSH配置App的转发路径置量

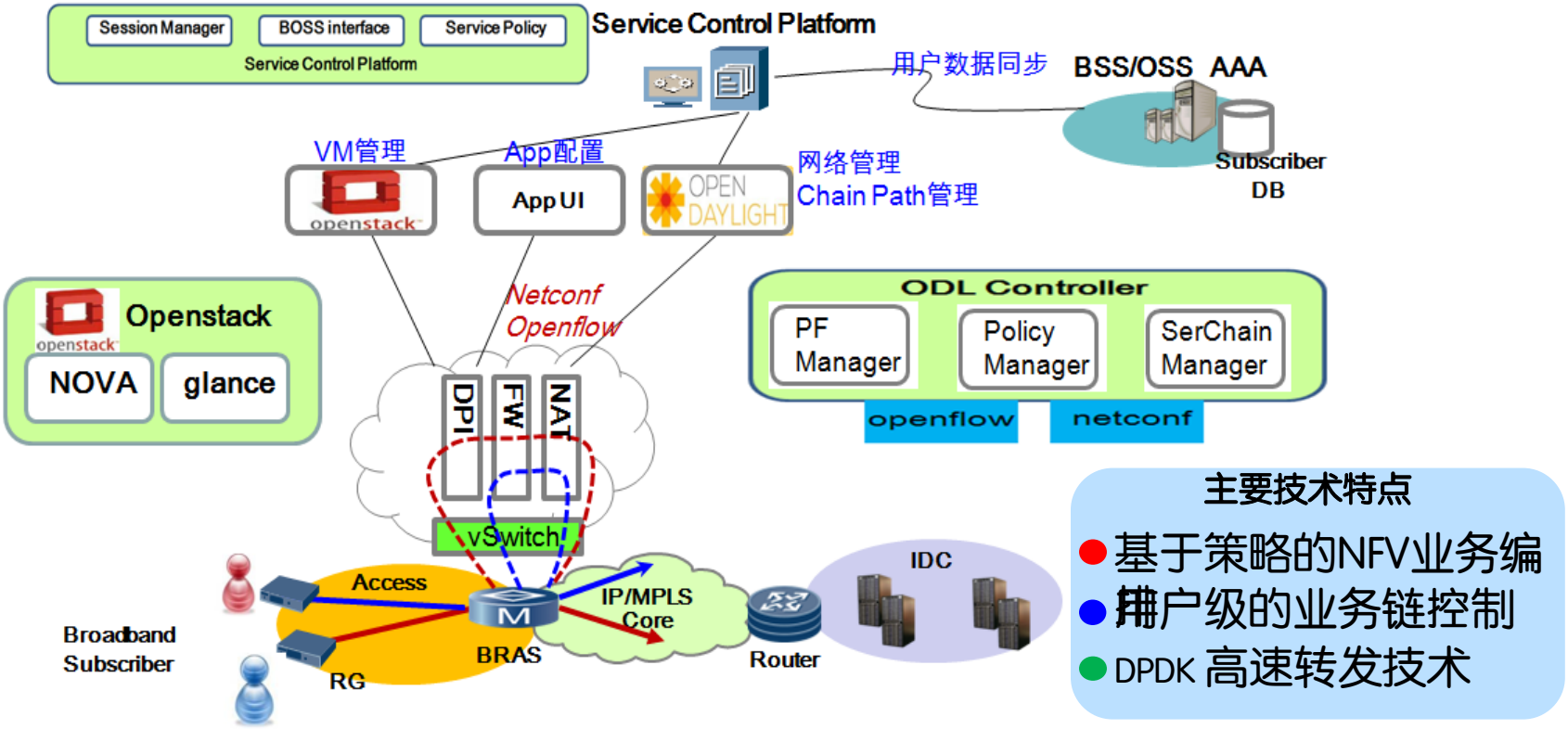


Service Function Proxy
- 传统IP网络与新型SDN网络间的流量封装格式转换



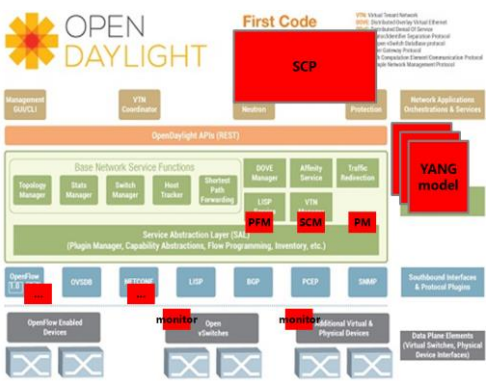
IP Edge Service Chaining总体设计方案

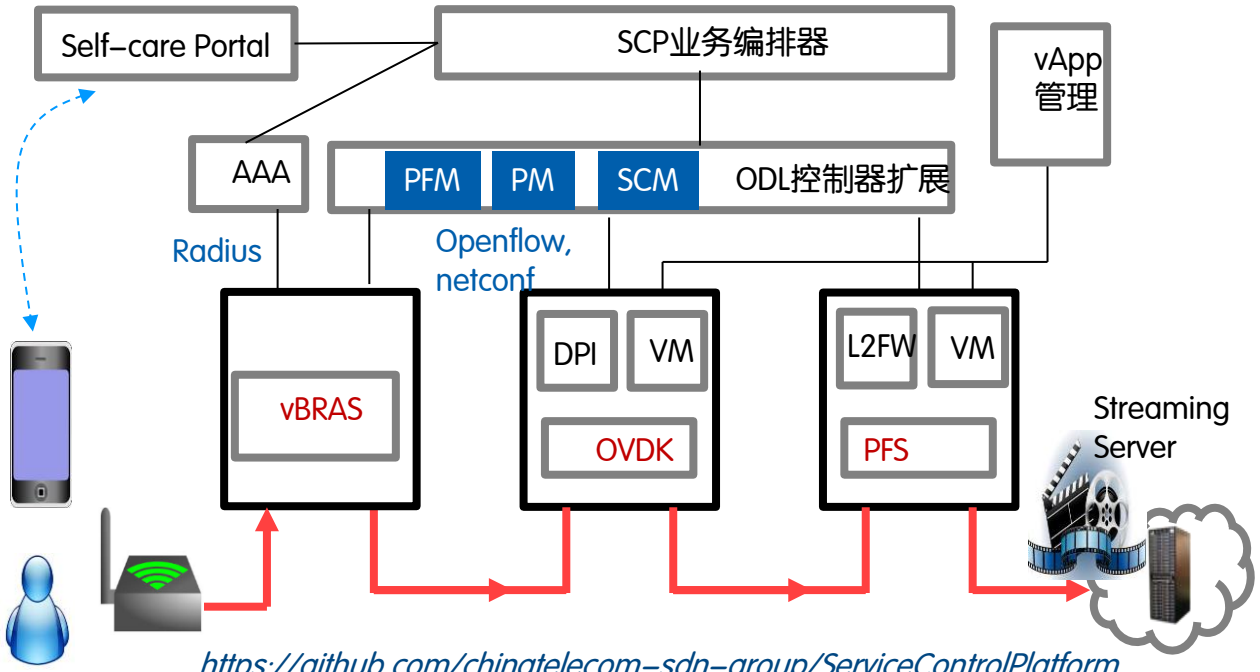
- 基于ONP对高性能NFV优化、IP Edge灵活转发、虚拟化管理等多项技术进行集成
  - SC基础设施管理： “App+服务器+网络”虚拟化协同管理技术，基础设施配置
  - SC业务配置工具：业务路径规划、网络配置模型化技术、拓扑管理、调度策略
  - 转发策略与路由管理：流量调度、路由策略、RIB/FIB配置、策略库、拓扑自动维护
  - SC性能监测：包括App性能监测、自动化负载均衡、业务保护



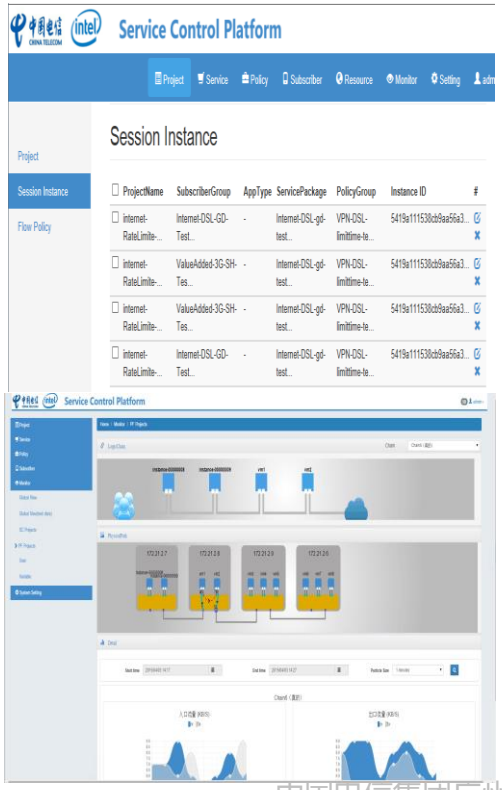
CT与Intel的vBNG联合创新（MWC 2015/ IDF 2015）

- 国内首个IP智能边缘层的业务链方案，实现基于用户session的业务控制
- 提供了一种NFV控制器PoC和NFV性能优化方案，部分突破x86性能瓶颈
  - 基于DPDK的高性能datapath，ovdk交换机
  - 业务链管理与智能化调度方案：智能算法、session控制
  - 可编程转发（PFA）：PFC+PFS共同解决承载的可扩展性问题
- 近期将提供NFV性能优化规范，部分代码已公开至github和ODL社区



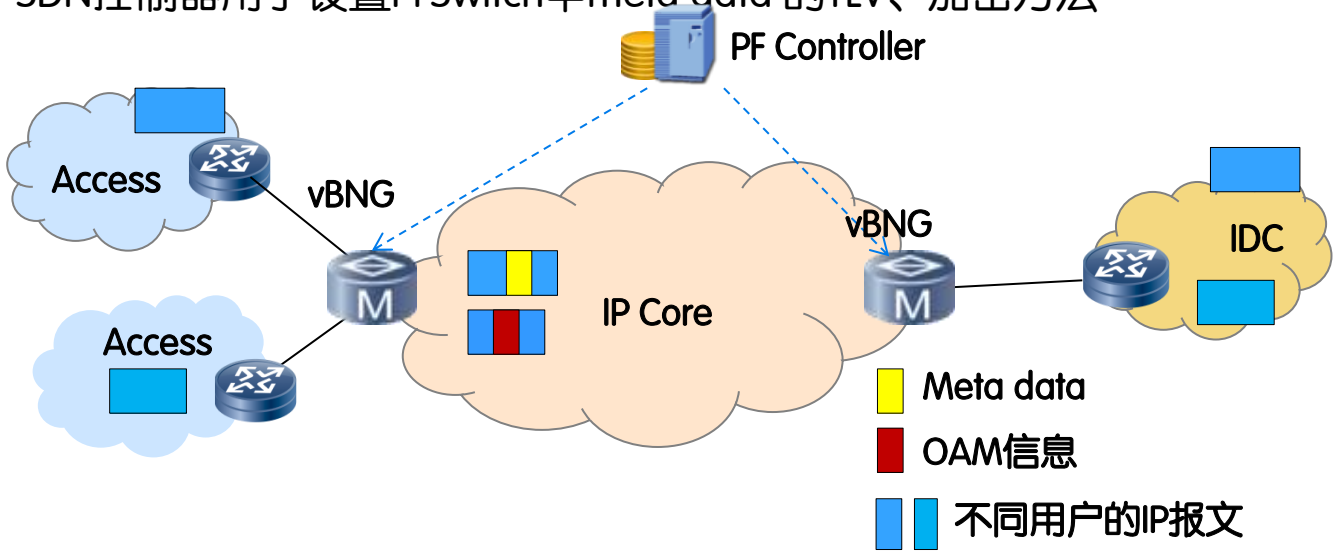


https://github.com/chinatelecom-sdn-group/ServiceControlPlatform



应用场景一： 基于meta data的应用优化承载

- 在IP报文中加载meta data、OAM信息，可沿传统IP网络传递到OTT网络
  - PFSwitch可以将在报文中插入/移去meta data（线路信息、用户信息、App信息）、OAM信息
  - SDN控制器用于设置PFSwitch中meta data 的TLV、加密方法

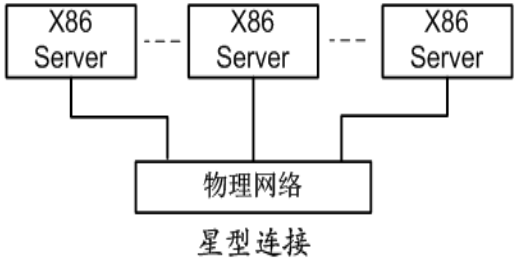


vBNG功能：
- Access侧：对特定用户流量的特定位置插入meta data，或者将meta data信息映射到MPLS标签中
- IDC侧：解析meta信息，重定向至指定位置

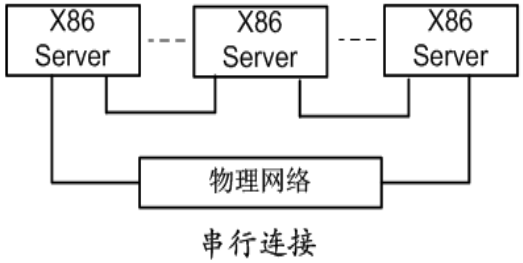
PF Controller功能：
- 配置Meta data在PFS中的TLV
- 管理配置PFS流量路径和流向

应用场景二：Service Chain流量均衡调度方法

- Servic Chain优化调度：基于CPU、Mem、Path Bandwidth的资源优化调度
- Service chain异构资源调度：属于NP-hard问题，需要设计相关Heuristic算法



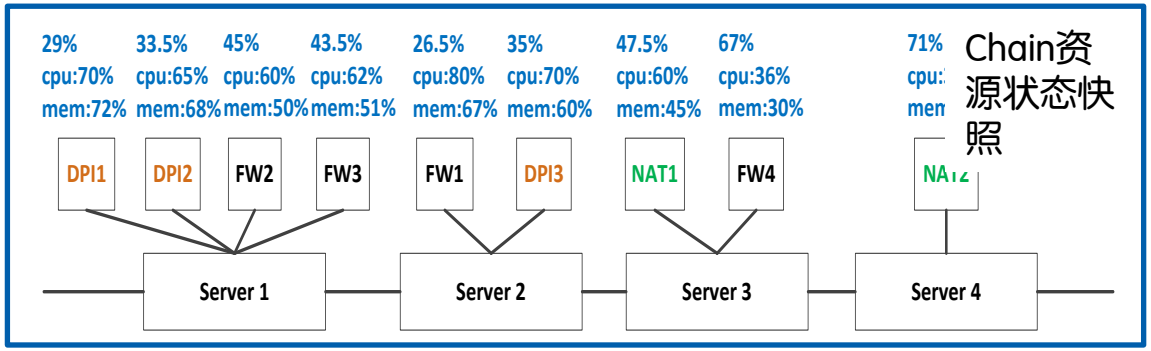
星型连接

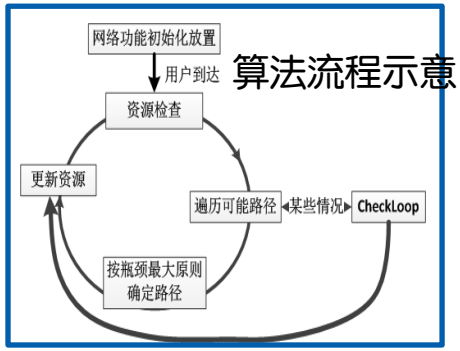


串行连接

不同场景需要的chain拓扑，各自的成本代价也不同，需要不同的LB调度策略

- 目前对串型Service chain的资源混合调度提供了算法，并基于ODL的SFC架构提供代码实现

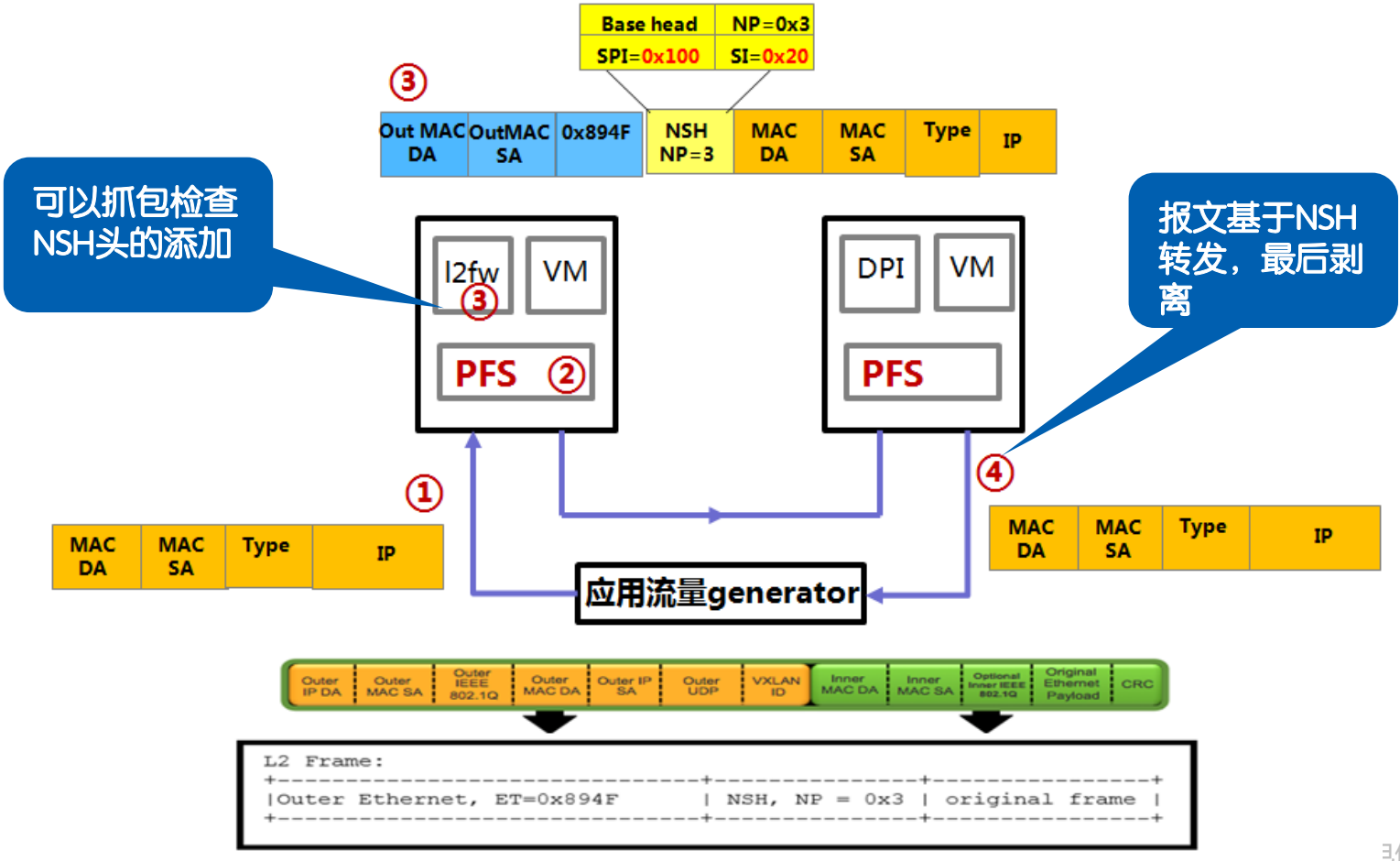




https://git.opendaylight.org/gerrit/#/c/18404/

应用场景三：基于Service Chain Header转发

- 编排器/控制器可以快速定义一个SC Header，并下发封装/转发策略和流表，至对应转发设备
- 通过PF vSwitch可以实现IETF定义的各类Service Chain Header支持，如：NSH



后续工作

- 深化DPDK相关的NFV性能提升技术研究，继续完善相关PoC
  - 关注Service Chain端到端性能提升技术
  - 关注海量流表下的转发性能优化问题
- 加强DPDK技术的推广应用，提供NFV应用规范和配置模板
  - 联合合作伙伴，公布技术白皮书与测试报告
  - 提供技术指导意见、测试技术规范、分场景应用配置模板
  - 推动IETF标准化、ODL等开源社区项目参与工作
- 诚邀业界同仁共建基于DPDK的高性能VNF生态，鼓励技术合作与技术互补
  - 推进基于DPDK的VNF性能优化技术方案、协商制定测试方案与技术规范
  - 积极参与SDN产业联盟等产业活动

中国电信
CHINA TELECOM

Thanks