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

DPDK加速无线数据核心网络(vEPC)

DPDK Fast Forwarding for Virtual EPC

陈东华
中兴 2015.04.21

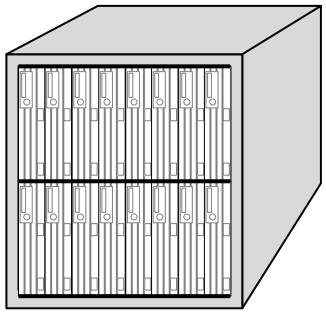
Content

- 虚拟化的转变
- DPDK
- vEPC网络使用DPDK
- 更多的DPDK解决方案

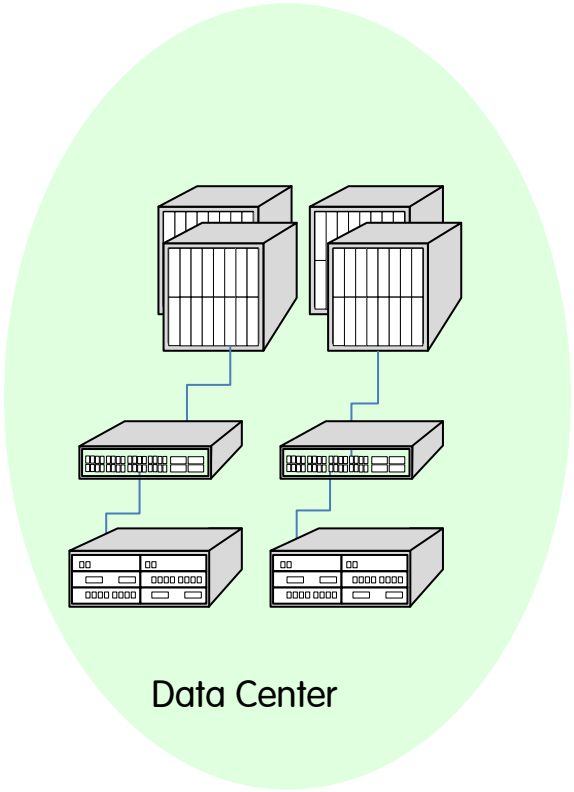
项目背景

- EPC SAE-GW 团队
- 实现NFV模型的SAE-GW应用；
- 硬件平台
  - 使用Intel XEON E5 系列CPU的通用刀片服务器
- 软件环境
  - OpenStack
  - VMM : KVM
  - OVS/SRIOV
  - VM : DPDK based SAE-GW stack

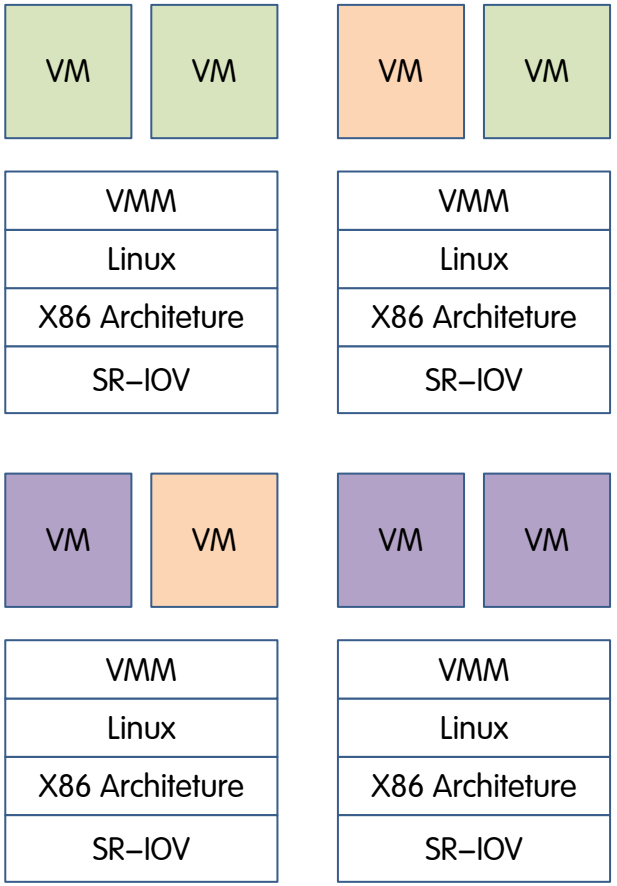
NFV虚拟化成为趋势



使用X86架构的COTS 服务器更通用；平台演进更快，通用性更强。



数据中心模式；集约，成本更优，管理便捷；

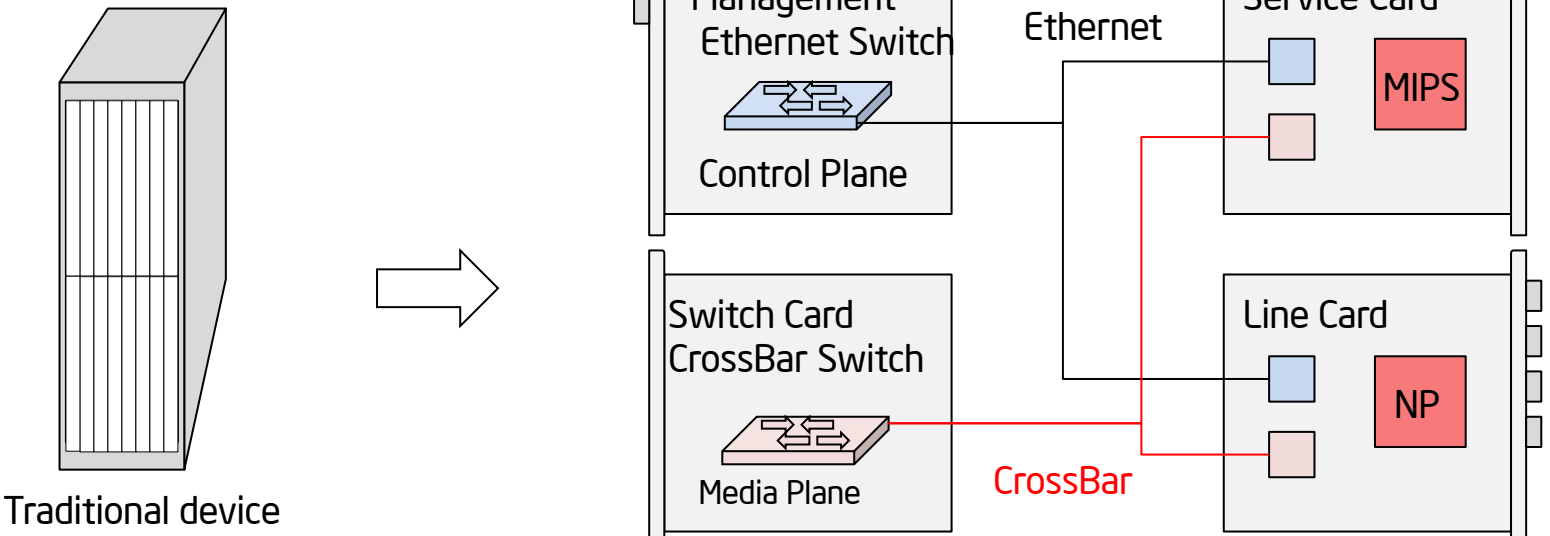


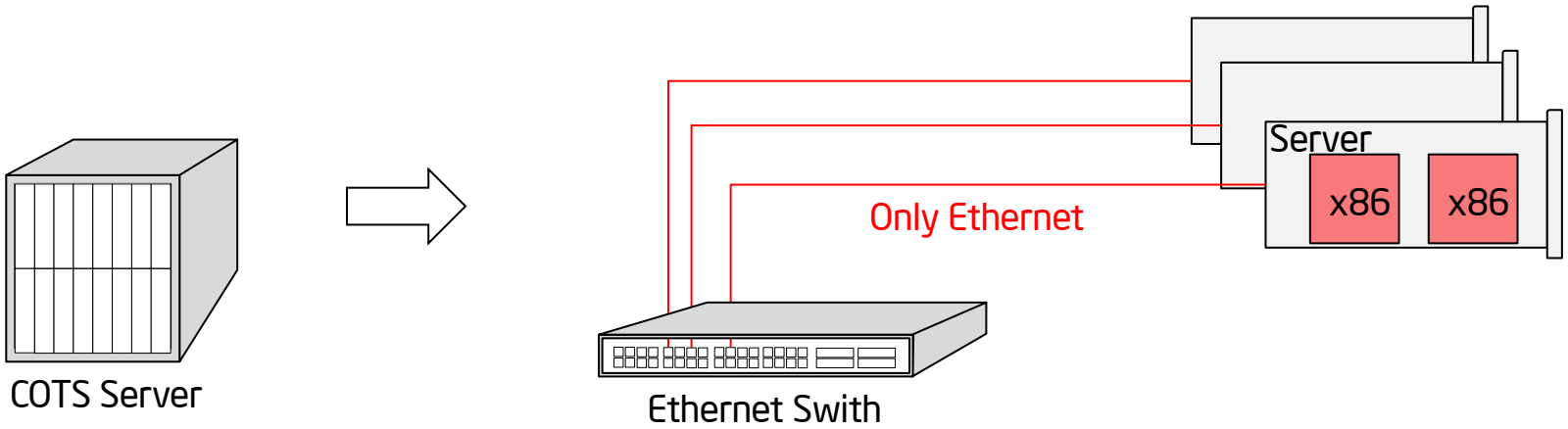
NFV模式下，虚拟机部署更灵活；能够跟随业务发展规模，合理的调整业务部署。

虚拟化的转变

- 传统网络设备：
  - CrossBar Switch
  - NP + MIPS
- NFV虚拟化：
  - Ethernet Switch
  - Intel X86
- 疑问：NFV性能是否能够和传统设备相比?
  - 传统思想：X86不善于做网络IO吞吐相关的工作；没有硬件加速，IO处理将严重消耗CPU的性能；
  - 在DPDK得到验证之前，项目充满了担忧。

虚拟化的转变：硬件对比


Management
Ethernet Switch
Control Plane
Ethernet
Service Card
MIPS
Switch Card
CrossBar Switch
Media Plane
CrossBar
Line Card
NP
Traditional device


Server
x86
x86
Only Ethernet
COTS Server
Ethernet Swith

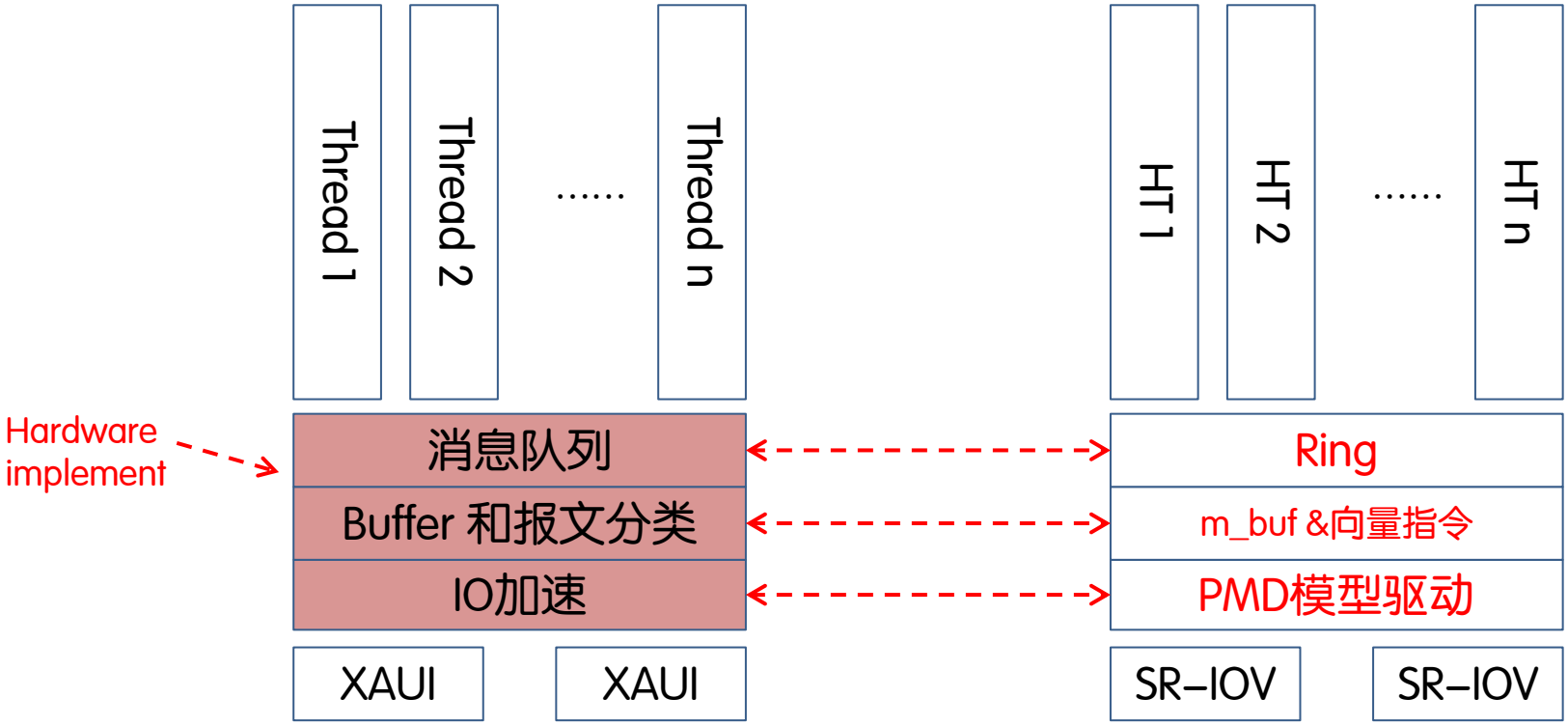
Content

- 虚拟化的转变
- DPDK
- vEPC网络使用DPDK
- 更多的DPDK解决方案

DPDK核心

- 用户态模式下的PMD Driver
  - 去除了中断影响，减少了操作系统内核的开销，消除了IO吞吐瓶颈；
  - 避免了内核态和用户态的报文拷贝；用户态下软件奔溃，不会影响系统的稳定性；
  - Intel提供的PMD驱动，充分利用指令和网卡的性能；
- HugePage和m_buf管理
  - 提供2M和1G的巨页，减少了TLB Miss，TLB Miss严重影响报文转发性能；
  - 高效的m_buf管理，能够灵活的组织报文，包括多buffer接收，分片/重组，都能够轻松应对；
- Ring
  - 无锁化的消息队列，实际验证，性能充足；
- 82599 SR-IOV NIC
  - 实现虚拟化下高速吞吐；
- Vector Instance /向量指令
  - 明显的降低内存等待开销，提升CPU的流水线效率。

DPDK vs MIPS 架构

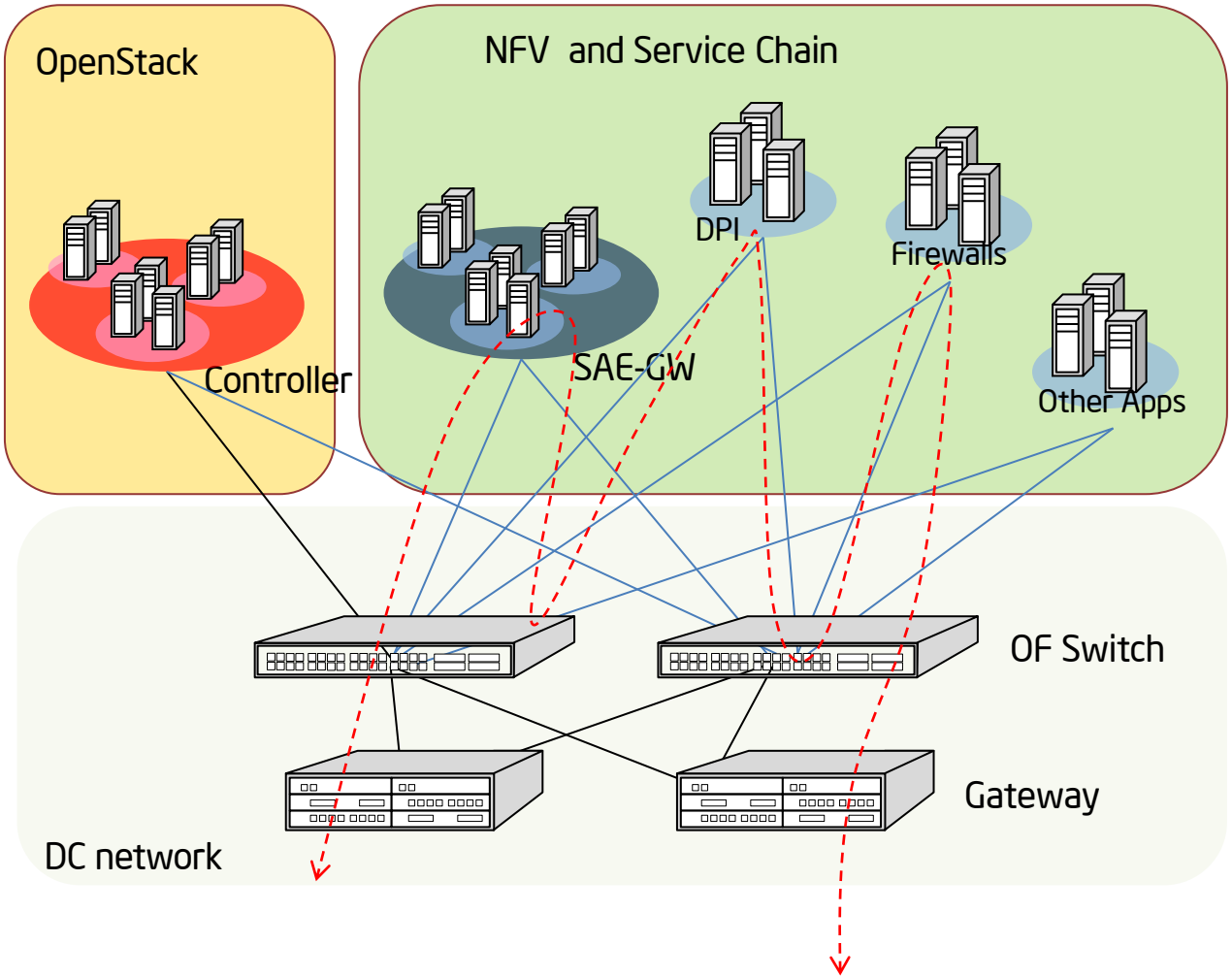


- DPDK提供了通讯协议处理最为关键的部分，具备了报文转发的整体解决方案。
- 相比较MIPS架构，DPDK软件解决方案，具备更强的灵活性。尤其是在报文分流层面，规则定义更方便，不受硬件限制；Ring队列灵活性更强，没有数量和队列长度饿限制。
- 使用软件化的解决方案，摆脱了业务对特定硬件的依赖，同时，软件方案具备了更强的扩展能力

Content

- 虚拟化的转变
- DPDK
- **vEPC网络使用DPDK**
- 更多的DPDK解决方案

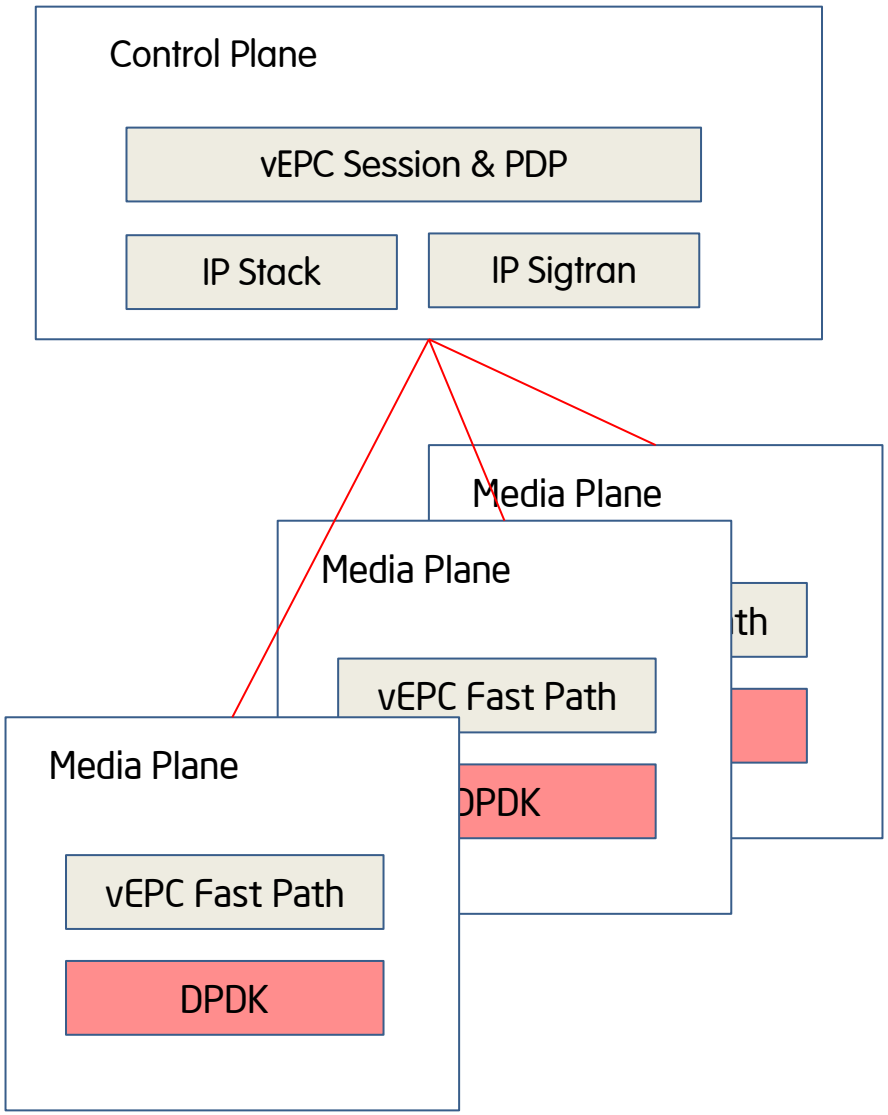
NFV模型下的SAE-GW 业务演进趋势



- NFV之后的EPC网络

1. 物理网络层次减少；物理层只有数据中心的交换机和网关；
2. 逻辑实体增加；各种业务，都通过NFV的模型部署在数据中心里面；
3. 业务链形成，并且趋向于智能化；
4. 问题：越来越多的网络设备，需要通过COTS服务器实现，服务器的性能面临着考验。
5. 趋势：DPDK在所有需要性能的业务之中，都值得推荐。
6. 现状：在NFV的发展过程中，客户和设备商具备了共识，DPDK是NFV的核心组件，NFV环境中缺省就提供部署。

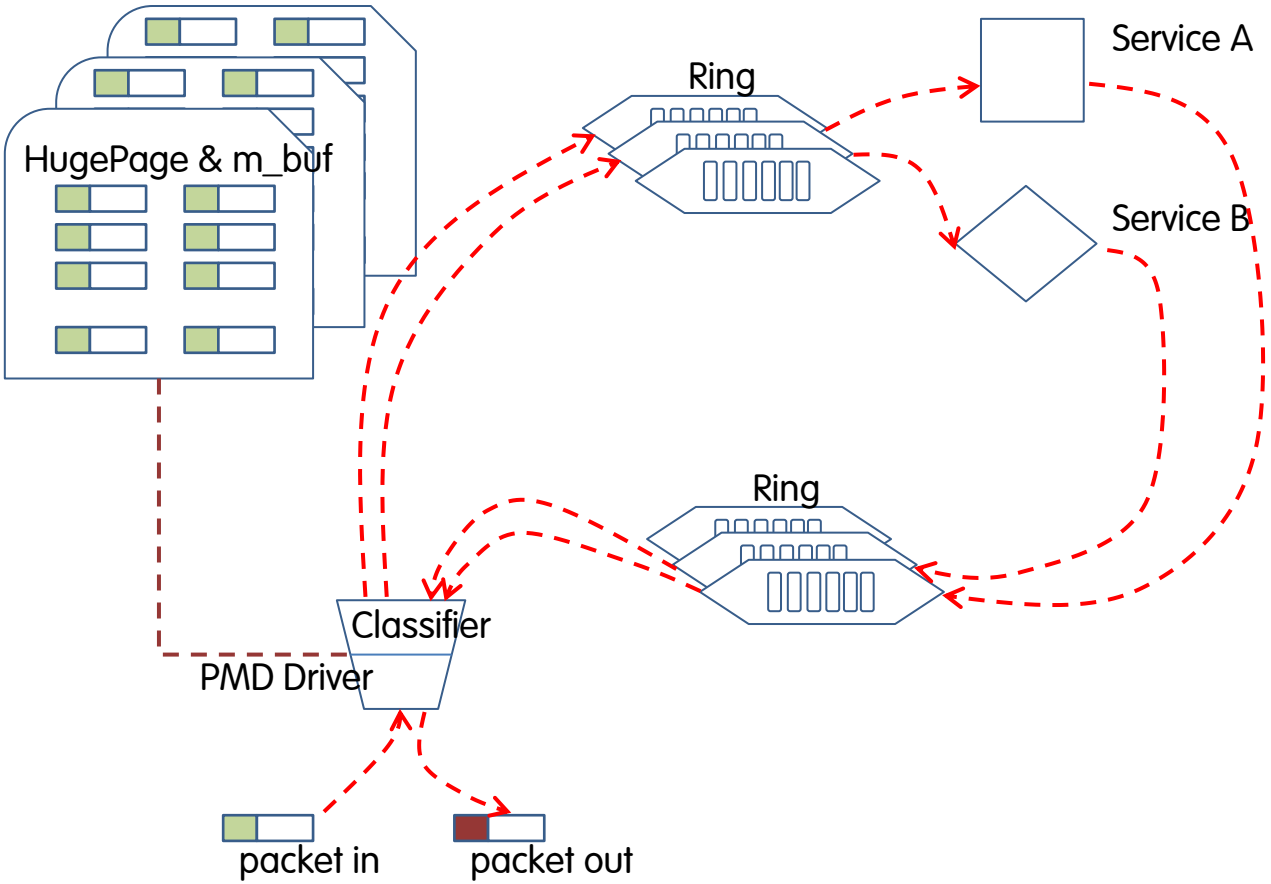
SAE-GW业务中的DPDK：快速通道基础



- vEPC业务特点

1. 控制面趋向于集中；对于控制面应用，DPDK作为建议部署方式，可以显著减少业务在IO层面的开销。在用户态具备完整协议栈的时候，推荐部署；
2. 媒体面趋向于分散；DPDK在媒体面高速处理中，是作为核心组件存在的，不可或缺；
3. 参考指标：DPDK提供的二层转发能力，单个CPU核模型下达到9Mpps吞吐；

使用DPDK部署SAE-GW快速通道的实际模型



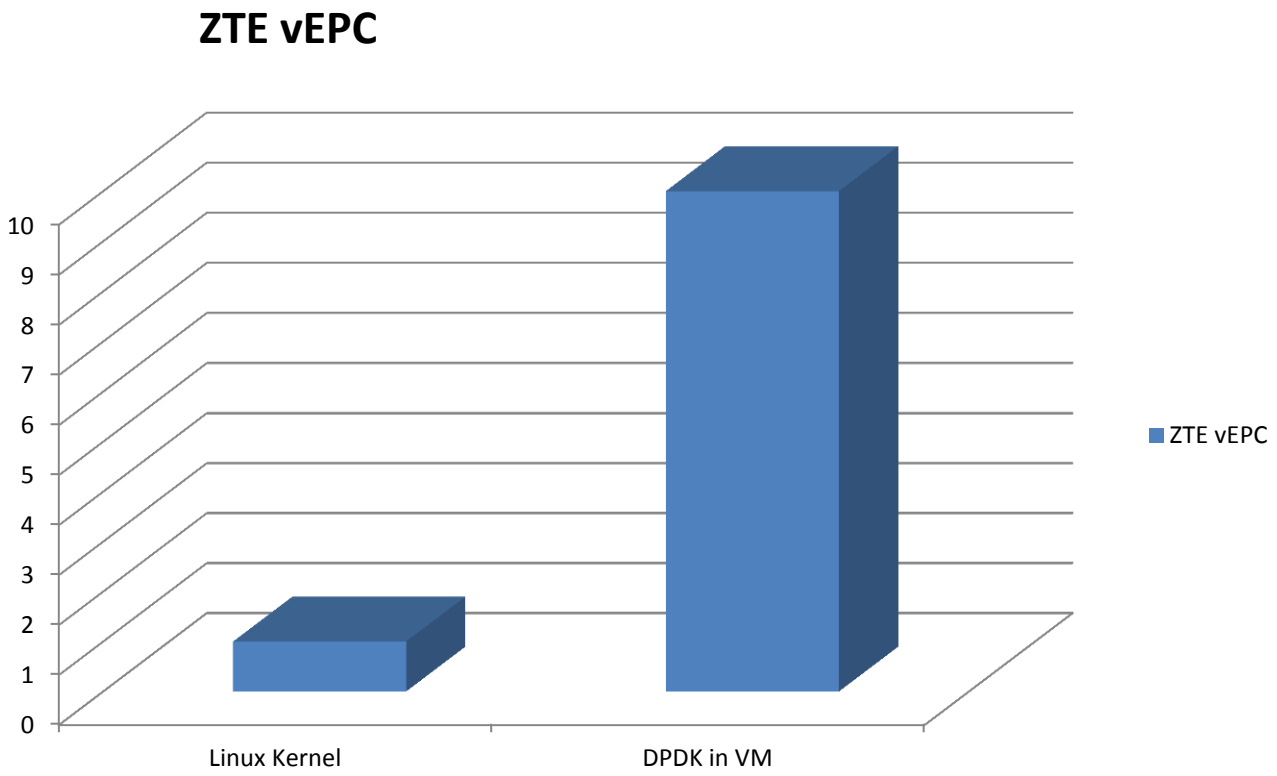
- vEPC使用DPDK部署业务

1. 报文通过PMD Driver接收，选择向量指令驱动，能够提供额外的20%性能提升；
2. DPDK提供了基础算法，高性能的HASH算法，在报文分类流程中作用巨大；
3. HugePage and m-buf机制，方便对报文进行管理；m_buf的数目可以灵活定义；即便是巨帧，也能轻松处理；降低了报文管理的复杂度。
4. Ring机制，在实际使用中完全没有瓶颈；即便是多生产，多消费模型，也统计不到冲突带来的延迟；
5. 报文分类后，进入不同的业务模块并行处理；
6. 简单的模型，提供了充足的性能；便于维护。

向量指令的应用

- SAE-GW快速通道，充分享受向量指令带来的好处；
  - **Intel DPDK提供了完整的2层转发和3层转发样例，其中有向量指令的应用，可以作为开发的参考样例；**
  - 向量指令能够在一条指令中，完成128bit/256bit的数据读写。等同于4/8个int类型变量的读写操作；原本，这需要4/8条指令。
  - 在解析报文的时候，作用最为明显：
  - 业务通常需要解析报文的Ethernet头部，IP头部，TCP/UDP头部，以往在提取报文字段时，最快的方式就是按照8字节对齐模式，逐个提取报文字段，并从8字节中抽取出关键字来。这种方式编写困难，尤其是要保证字节对齐；
  - 向量指令提供了封装函数，不需要考虑字节对齐，代码编写更加方便，操作效率再次提高；
  - 效果：采用8字节对齐读写报文，相比逐个字段解析，可以提升1.5倍性能；采用向量指令，可以进一步提升性能。

巨大的性能改观



- 在E5-2670v2， 2.8GHz CPU上做了对比验证；
- vEPC使用DPDK后，相比Linux本身，性能有10倍以上的提升。
- 从性能角度来看，NFV模型的SAE-GW已经满足商用部署要求了。

回顾：传统设备和 NFV+DPDK

- 在高性能设备的开发选择上，不需要犹豫
  - 传统设备存在的问题：对NP芯片，专有芯片有很强的依赖，芯片的专有架构，芯片稳定性，都会对部署产生很大的影响。一个芯片的bug，解决周期超过2年，甚至无法解决；
  - X86架构使用更广泛，稳定性更好，性能提升更快；
  - NFV让业务软件化，业务演进速度更快；
- DPDK开源社区，确保DPDK的发展和演进，消除了DPDK使用的疑虑；
  - 更好的稳定性；
  - 更少的漏洞；
  - 更快的演进；

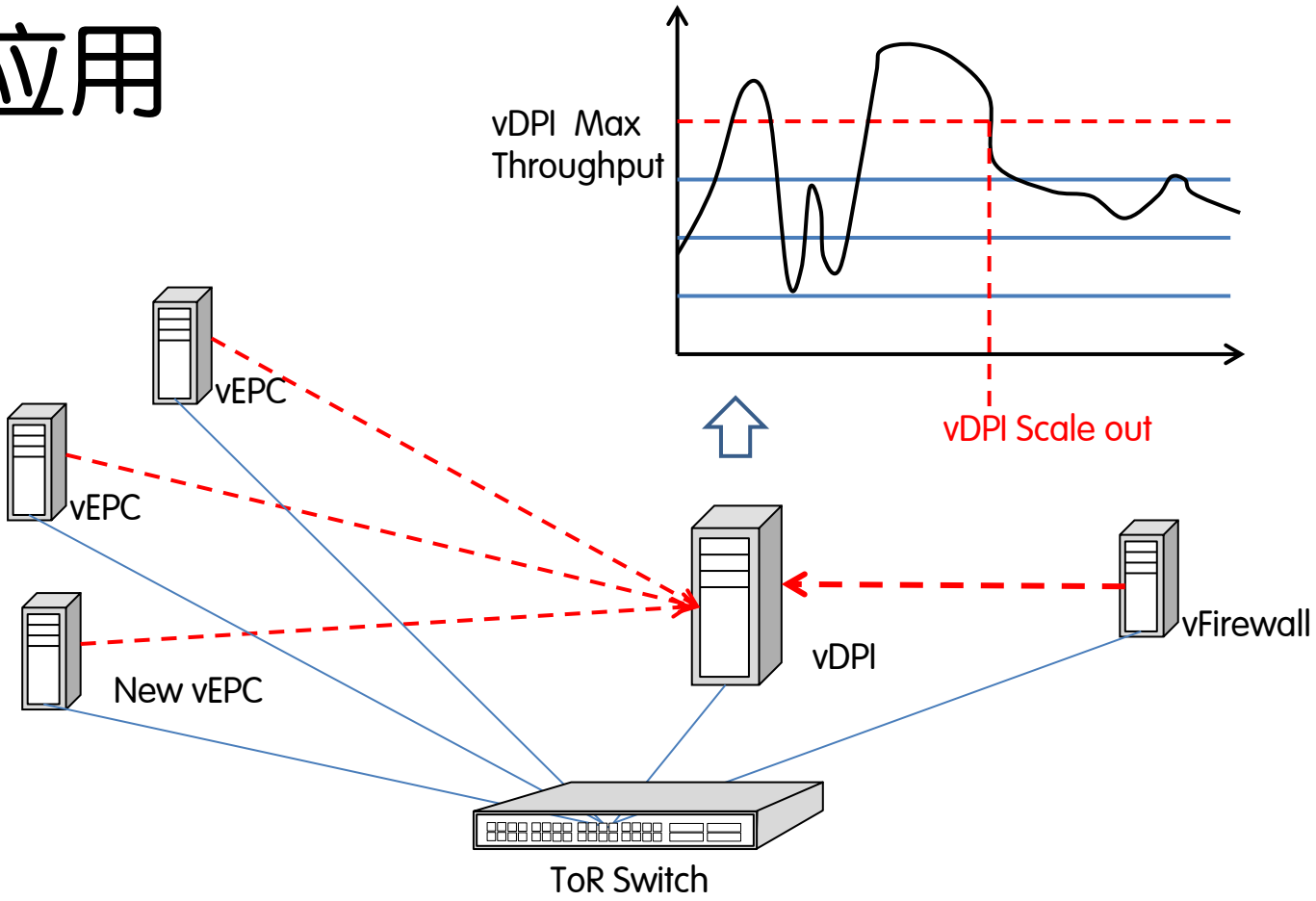
Content

- 虚拟化的转变
- DPDK
- vEPC网络使用DPDK
- 性能
- **更多的DPDK解决方案**

更多的DPDK解决方案展望

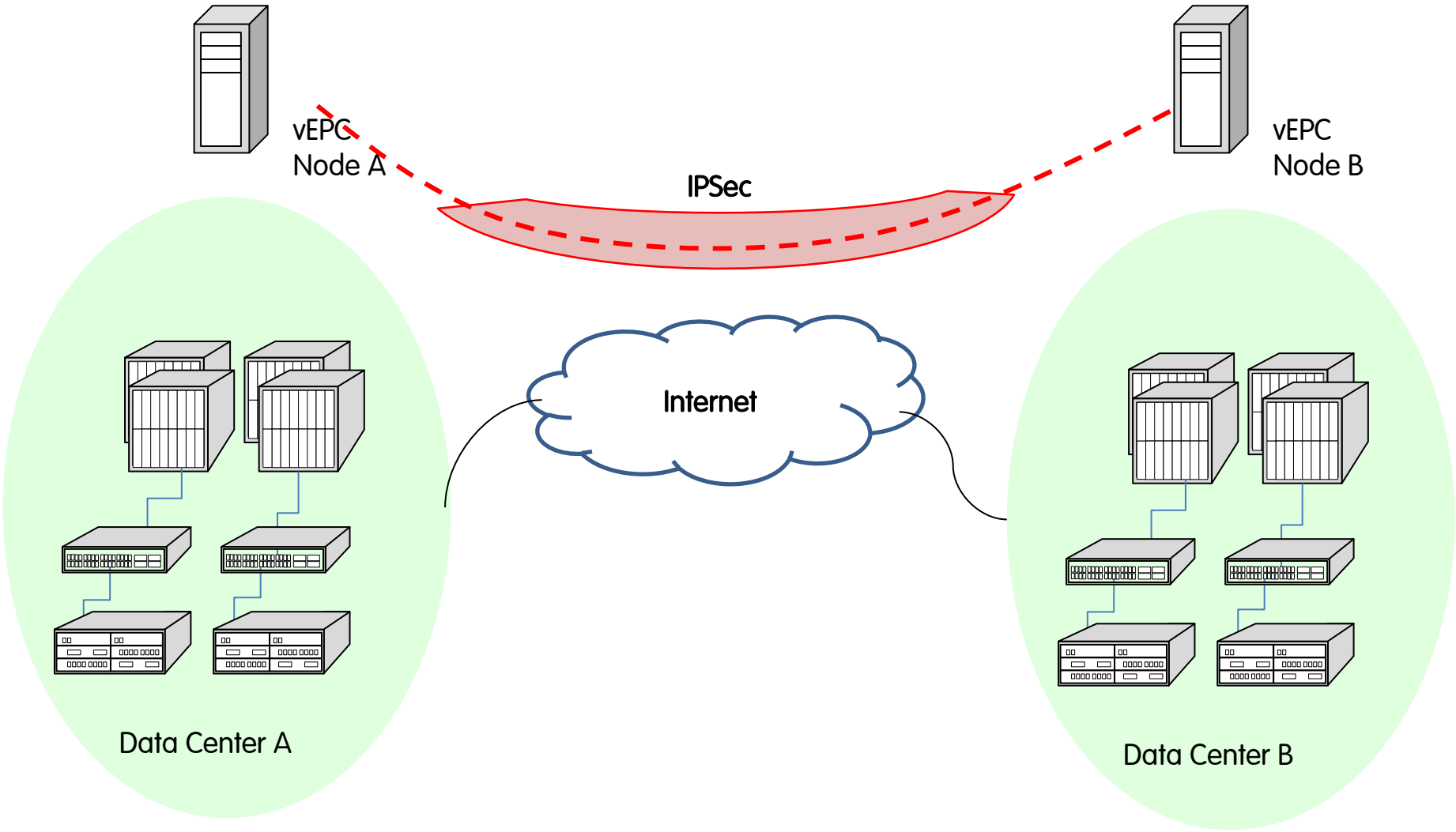
- 高性能下的QoS应用
- 高性能下IPsec安全应用
- 40Gbps高速吞吐

QoS 应用



- NFV环境，QoS问题更加突出，东西向流量增加，ToR交换机QoS能力不足
- 弹性机制，导致虚拟节点经常面临承受流量波动，直到新的虚拟节点被创建出来，在大流量环境，会出现短时间内服务质量下降；这将成为NFV应用中的常态。
- DPDK提供了QoS解决方案；

Security Problem



- vEPC节点可以跨数据中心部署，需要具备高性能的IPsec隧道，保护业务核心数据。
- 高性能的IPsec方案也即将验证；

40Gbps 高速吞吐

- 单刀片40Gbps吞吐，新的挑战；
  - 随着单用户无线传输速率的提升，NFV的网关设备面临更大的挑战，单刀片40Gbps的应用很快就会到来；
  - 单刀片实现40Gbps吞吐，需要的不仅仅是技巧；
  - 使用DPDK的部署方式：
    - 网卡多队列；
    - 合理的资源规划：CPU，内存，TLB页表等等；
    - 精简的处理流程：流程的腐化是致命的，技巧无法弥补；
    - 协调的指令编排。例如：多个报文一起读取，将比单次读取一个报文更高效；
- 对于40Gbps高速吞吐方案充满信心；

总结

- 在NFV的发展过程中，DPDK发挥了巨大的作用；
- 选择NFV，选择DPDK：
  - 摆脱了对专有硬件的依赖；
  - 令人满意的网络处理性能；

Thanks