使用 Mellanox NPS-400™ 网络处理器实现 NFV 加速

白皮书

目录

1. NFV 概述

网络功能虚拟化 (NFV) 是 ETSI 正在开发的一个标准，它允许在软件中实施诸如负载平衡、防火墙和交换等网络功能，并在行业标准服务器上运行它们，从而虚拟化当前部署在专用网络硬件上的网络功能。NFV 将显著降低运营商网络的费用，同时在运营商为新客户和应用部署网络和改造网络时为他们提供更大的灵活性。加速功能让 NFV 能够提出最具性价比的点，从而导致设备制造商和运营商对 NFV 都有较高的市场接受度。

推动这一工作的运营商期望 NFV 通过在更少数量的硬件平台上进行整合来显著降低其网络的运营和资本费用。虽然网络功能的性能在 NFV 上低于针对该功能优化的专用硬件，但对于许多网络来说，该性能可以接受，并且随着服务器性能的提高和技术的升级，可并行部署许多虚拟网络功能 (VNF)，预计性能也会相应提高。基于 NFV 的网络也比专用网络设备更加灵活，在需要其他网络服务时，可以在现有运营商数据中心资源中部署其他 VNF 和服务器。

📖有关 ETSI NFV 工作的信息，请访问 http://www.etsi.org/technologies-clusters/technologies/nfv。

图 1 显示了将网络功能整合到 COTS 服务器。

图 1. NFV 整合到 COTS 服务器

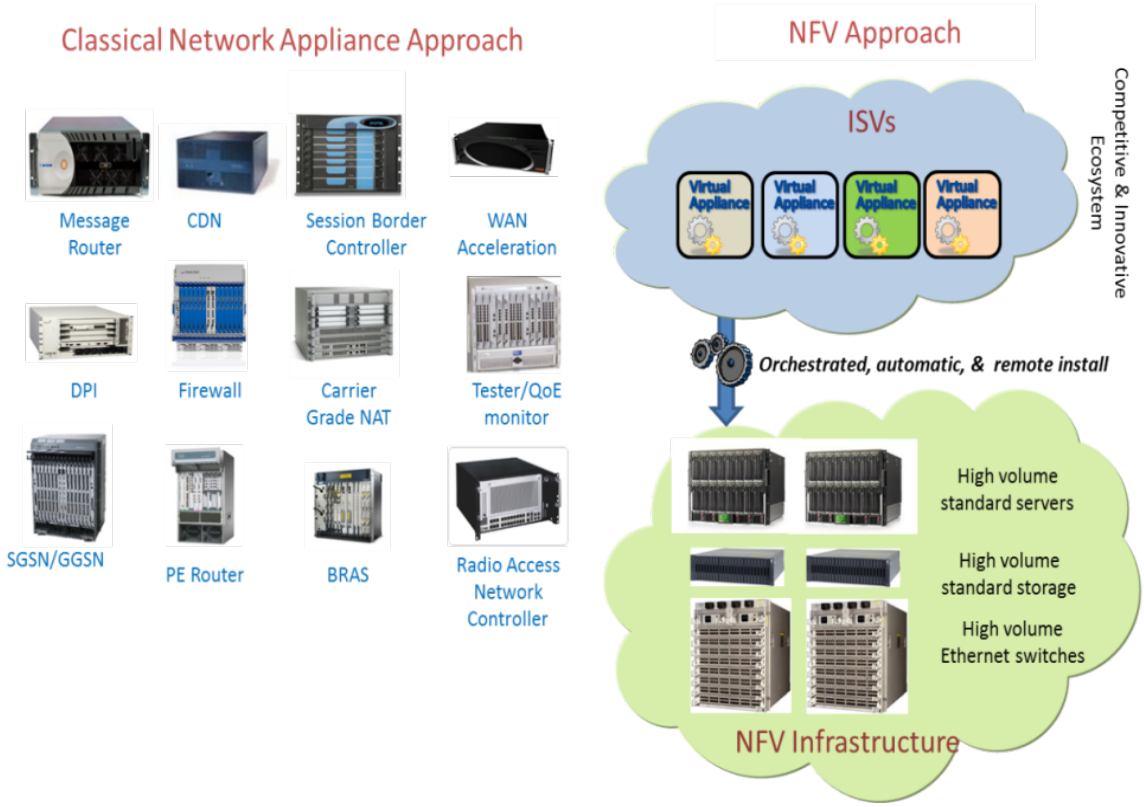


图 1 左侧显示的 16 个不同网络设备需要 16 个不同的硬件平台。随着 NFV 的出现，每个网络设备都可以虚拟化并部署为**在标准服务器和网络基础架构上运行**的虚拟网络功能 (VNF)。

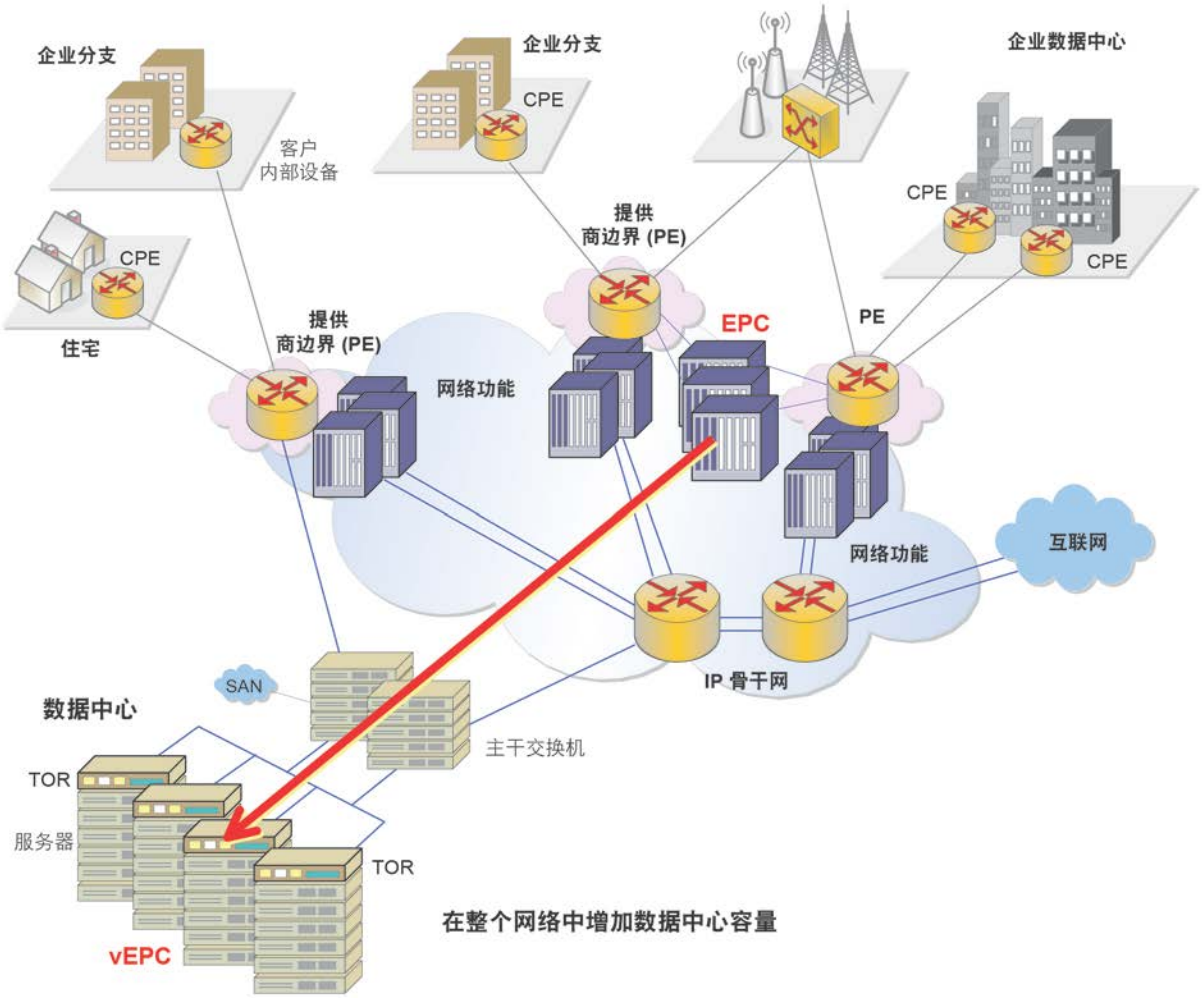
在开发 NFV 的同时，大型多租户数据中心上的工作负载也逐渐增加。新应用正在造成所使用网络带宽爆炸式增长，很大一部分是数据中心内的“东西向”流量。随着叠加 (overlay) 网络实现的多租户网络，智能正在移向网络边缘。数据包速率和每数据包处理能力的提升都会影响服务器性能。面对大家需要的这些提升，服务器性能相对于工作负载落在了后面，例如，据报告，对于平均数据包大小，10G 链路上 GRE 封装的性能为线速的 24%，而对于更小的数据包则下降到线速的 5%。

数据中心需要进行创新才能实现更高的密度、更低的成本和更低的功耗。NFV 提供了灵活的平台可用于创新，但是无法解决性能、功耗和成本问题。NFV 加速可提高性能，同时可降低功耗和成本。NFV 加速可卸载虚拟交换机 (OVS)、TCP、DPI、加密以及潜在应用特定加速基元的数据平面处理要求，从而让服务器可以集中处理控制平面。分隔和加速数据平面处理提高了整体性能，让 NFV 能够处理更大的网络、更高速度的链路以及更频繁接触的应用。

2. 虚拟网络功能 (VNF)

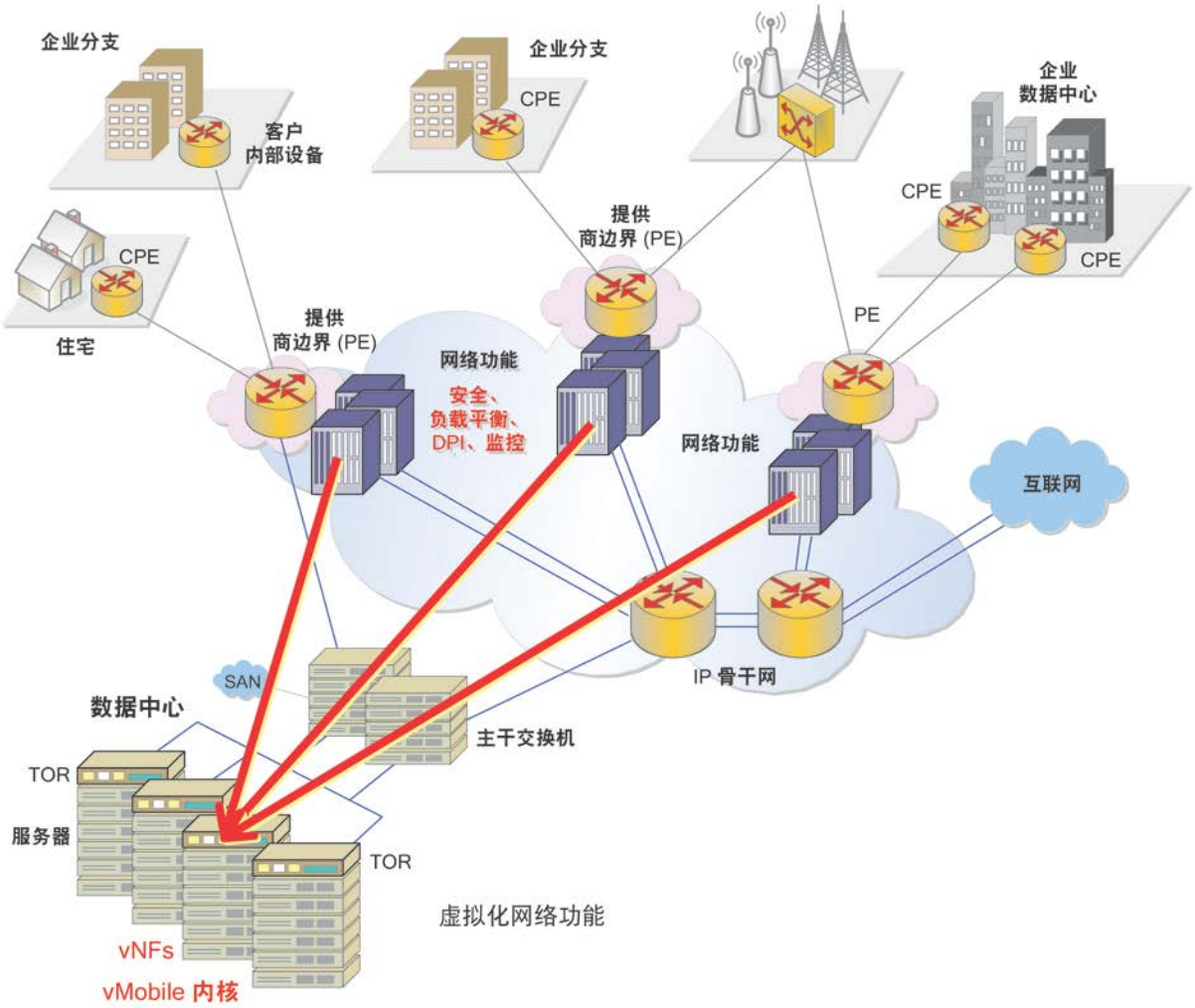
图 2 显示了运营商网络中的物理分组核心演进 (EPC) 设备如何用远程数据中心服务器上运行的 VNF 替代或扩充。

图 2. 远程数据中心内的 VNF 更换运营商网络中的 EPC 设备



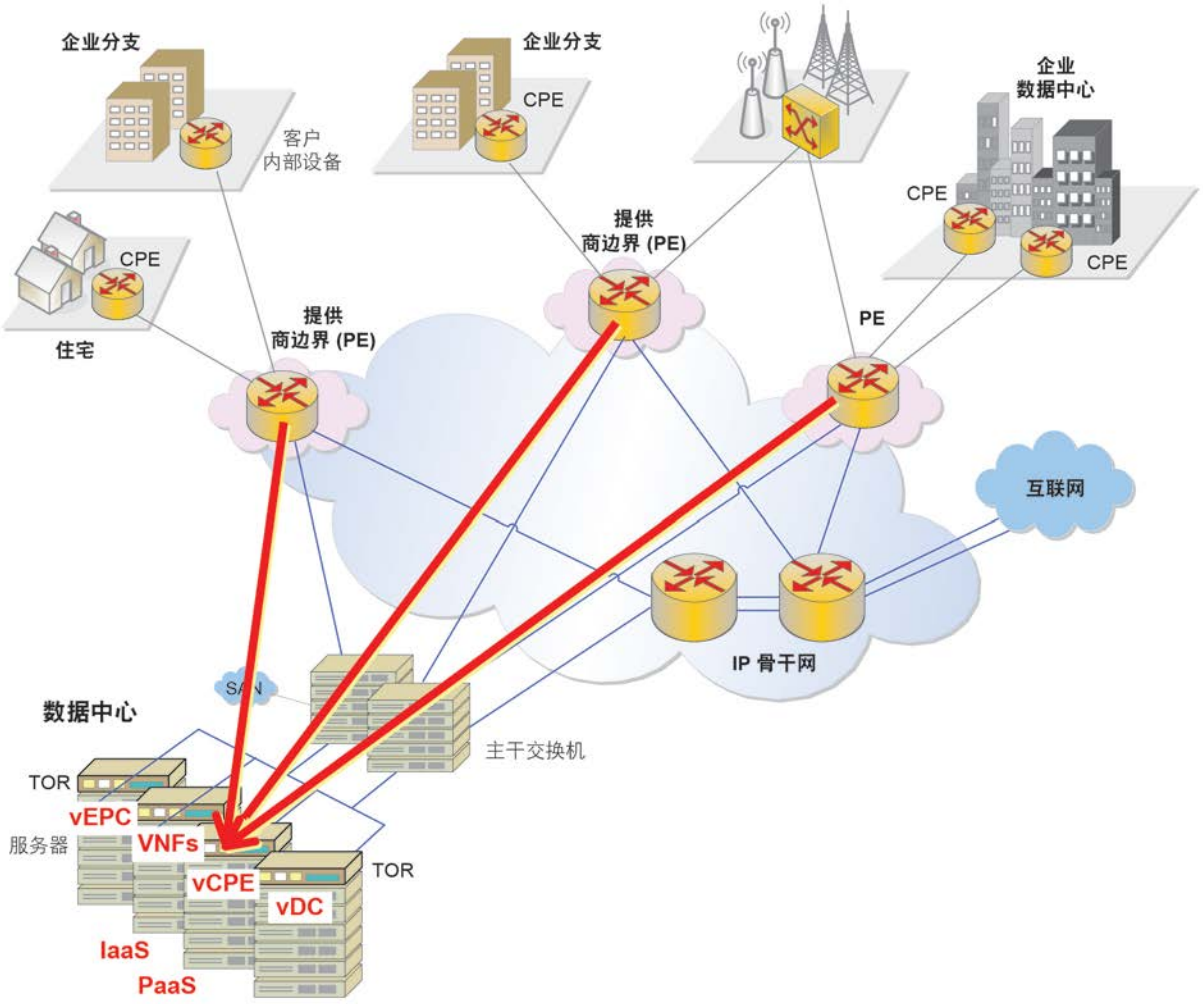
许多网络功能都可以虚拟化，包括客户内部设备、边界路由器功能甚至企业数据中心自身。随着更多的功能被虚拟化，将虚拟网络功能整合到同一个数据中心可以在 VNF 之间提供更好的通信，并提高性能。图 3 显示了用于安全、负载平衡、深度数据包检测（入侵检测或应用识别）和监控的物理网络功能的虚拟化，然后虚拟化的功能将作为数据中心中的软件 VNF 进行运行。

图 3. 网络功能的 NFV 虚拟化



任何网络功能都可以虚拟化，具体取决于所需数据速率和处理量。
图 4 显示了运营商数据中心内的其他 VNF、CPE 甚至企业数据中心的虚拟化。

图 4. 运营商数据中心内的其他 VNF、CPE 甚至企业数据中心的虚拟化



每一个物理网络功能都可以虚拟化，并作为图 4 左下角服务器中的软件 VNF 运行。对于许多 VNF 来说，服务器将位于原始网络功能附近，以避免回程流量和增加延迟。然而，如果能够在图 3 所示的远程服务器池中灵活地添加其他 VNF，那么将允许运营商更快地响应需求、新客户等各方面的变化。

从运营商网络部署中去除多个硬件平台可以大大降低运行网络的运营成本，因为它不再要求对物理网络功能、备件、卡车运输到远方的站点等提供专门的支持。但是 NFV 的主要优势是向运营商提供了更大的灵活性，从而可在无需改变物理拓扑的情况下在网络上快速配置服务。

3. NFV 挑战

现在我们讨论一下在数据中心实施 NVF 最具有挑战性的一些方面。

3.1 性能挑战

NFV 的节省和灵活性需要付出很大的代价：NFV 网络上的性能更低，功耗更高。在大多数情况下，用于虚拟网络功能的控制平面软件可在不损失任何性能的情况下移植到标准服务器。性能挑战发生在数据平面。可通过在多个虚拟机和/或内核上部署多个 VNF 来解决许多功能的 NFV 性能问题。但是，在多个虚拟机上部署后，Hypervisor 中的 vSwitch 不能有效地使用服务器和网卡中的单根 I/O 虚拟化 (SR-IOV) 功能，因为会将数据包复制到 vSwtich 缓冲区，从而导致服务器的性能降低。如果将 VNF 部署到多个内核上，那么消耗的电力将显著高于专用网络功能。在某些情况下，运行于多个内核上的 VNF 将导致不同的网络行为，例如分发网络功能时导致无序数据包交付。由多个虚拟机共享时，标准服务器提供的性能将不确定，共享处理器、缓存内存状态和 TLB 失败都会导致变化。

3.2 负载平衡挑战

NFV 可以使用 Intel® 数据平面开发套件 (DPDK) 提高确定性和性能，但是使用 DPDK 会导致更高速链路出现数据平面问题。在 10G，大多数功能都需要多个内核才能实现整个网络功能特征集的线路速度，而在 40G 或 100G 则需要许多的 VNF 和内核。然后，需要负载平衡在多个 VNF/内核上分配工作负载，但是在内核之间移动数据块可能会影响性能，因为这样会降低缓存的效力并增加 TLB 失败率。某些应用要求流中的所有数据包均由同一 VNF 进行处理。除了确保交付流中的所有数据包，负载平衡器还必须有状态。事实证明，如果处理 IP 分片，会造成额外的延迟，并且会使用更多的计算资源（尤其是负载平衡器本身就是 VNF 的时候），有状态负载平衡这时就成为一项重要的任务。甚至使用有状态负载平衡器，如果 VNF 保留数据包顺序，单流性能将限制到单个 VNF/内核的性能。大容量的单流方案是 NFV 面临的一项挑战，尤其是在必须保持数据包顺序时。

3.3 性能监控挑战

另一个 NFV 挑战是用于正确编排和服务管理的性能监控。大型 NFV 网络中的高效性能监控器必须支持有状态流表中的数百万个流，然后收集、保持并在某些情况下分析数百万个计数器。在某些应用中，除了服务器或部分网卡的精确性外，还需要提供精确的事件时间戳。

3.4 范围挑战

要支持大量订阅者，可能涉及到很大的表，如果它们超过缓存大小或影响缓存填充和弹出算法，可能会影响服务器的性能。许多网络设备均基于流，随着订阅者数量的增加，流表的大小可能会快速增大到超过标准服务器的缓存容量。

合并了 OpenFlow 交换机功能的 VNF 将具有不同大小的表和表数量，具体取决于网络中的功能和位置。具有上百万流表项的 VNF 将造成服务器的性能问题，因为服务器缓存不会大到足以有效地处理大量流。此外，个别 OpenFlow 表项可能会很大，OpenFlow 1.3 支持 40 个不同的匹配字段、IPv6 地址、MPLS、PBB 等。虽然没有一个表具有 40 个匹配字段，但是具有 12 个或更多字段的 IPv6 表非常大，如果在基于服务器的 vSwitch 中实施，会严重影响服务器的性能。

3.5 可靠性挑战

NFV 网络必须与现有的物理网络交付相同的可靠性，确定故障，监控网络服务的运行状况以及连接管理（OAM、BFD）。使用访问控制列表 (ACL) 来保护网络功能，方法是将每个数据包中的寻址和服务（端口）信息与一组规则（即 ACL） 进行比较。网络保护可以是安全保护（授予访问权限）或防止威胁（拒绝访问权限），根据网络功能的位置不同，ACL 可能会很长。为了提供网络稳定性，边界路由器可能需要检查成千上万的 ACL 规则，如果边界路由器已虚拟化，将给服务器带来巨大的工作负载。

3.6 安全挑战

对于通常驻留在数据中心外的功能的网络虚拟化，回程到 VNF 的网络流量会增加网络上的负载，并且传输延时会增加所提供服务的延迟。如果敏感数据回程返回到数据中心，或者 VNF 与运行用户应用程序的虚拟机共享服务器，则 VNF 上的安全可能会成为问题。可通过加密回程流量并使用像 VXLAN 这样的封装协议隔离来自其他用户的流量来解决回程安全问题。回程流量还会导致使用网络中的额外带宽。

为同一网络、服务器或内核上的多个租户提供安全连接和服务需要许多的相同隔离技术，例如 IEEE DCB 小组为数据中心开发的虚拟以太网端口聚合器 (VEPA)，从而允许 vSwitch 将 ACL 检查任务卸载到相邻的以太网交换机。网络和服务器基础架构必须支持来自多个供应商的多个 VNF，同时还应提供服务水平协议 (SLA) 测量，验证服务的交付以便进行计费。

3.7 流量管理挑战

在许多 VNF 的出站端，流量必须整形或以其他方式管理。有些流可能需要较大缓冲区，有些则需要低延迟。传输调度程序应该能够在用户、用户组、端口、端口上的通道等之间执行资源使用策略。资源使用策略包括优先级、公平使用带宽或缓冲区、缓冲数据包的超时等。流量管理功能可能会给服务器施加繁重的工作负载，或者由于服务器工作负载、上下文切换或在共享网卡中的排队，服务器可能无法获得整形流量所需的精确度。

3.8 虚拟化软件挑战

NFV 横向扩展意味着并行添加通常运行在多个内核上的多个 VNF，以便增加 VNF 的容量。如上所述，IP 分片对多个内核上实施的 VNF 的性能会带来严重的影响。由于 TCP 和 UDP 端口号仅在分片数据包的第一个碎片中，因此保留 5 元组状态涉及到维护有关每个流的负载平衡器信息，包括 IP 5 元组和 MAC SA/DA 和类型/长度字段。对于大多数网络功能，单个 VNF 处理流中的所有帧，并且可能会重新组合分片数据包作为其操作的一部分。如果将 IPv4 数据包碎片分配到不同的内核，由于需要在内核之间传输碎片来重组数据包，因此会影响性能。并且由于将数据块移到缓存中，从而逐出可能有用的缓存内容，此传输进一步降低了性能。如果采取捷径对数据包的传递重新排序，则网络行为将发生改变。如果在没有某种序号或其他机制保持数据包顺序的情况下跨多个 VNF 分发 IP 帧，那么就可能发生数据包重新排序。接收数据的服务器将遭遇 TCP 堆栈中的工作负载增加，或者可能会丢弃无序的帧，从而导致重新传输或网络负载增加。

有些 DPDK 网络实施会改变预期网络行为以保持较高的性能。当跨多个内核分发计算以及在某些情况下无法执行接受的功能时，可能不会保留数据包的顺序。例如，DPDK L3 示例代码让检查 IP 报文头校验和、减少 TTL 字段和重新计算 IP 校验和成为可选项，所有这些功能在较大的网络中都会影响网络稳定性。

3.9 NFV 加速

NFV 标准包括加速，它是提高网络性能的一个选项。在 NFV 基础架构小组规范中，如果网卡具有 TCP 卸载引擎 (TOE)、LSO、DPI、加密或其他加速功能，则处理器组件可以将数据包处理卸载到网卡上。基础架构网络规范包括 OpenFlow 功能，以便像负载平衡器或防火墙这样的虚拟功能可以在上游交换机中填充 OpenFlow 表，以在接受流后转发和/或处理流中的后续帧。此功能在基础架构网络中称为快速路径卸载 (FPO)。由于 OpenFlow 是一个通用机制，因此可以通过与流表项关联的指定为 OpenFlow“操作”的数据包修改，执行网络地址转换 (NAT)、隧道封装或解封等数据包处理功能。

随着在 NFV 和 SDN 网络中出现创新，很难预测会在哪里出现变化。但是，加速需要繁重数据包处理的功能 [如高级分类、访问控制列表检查、深度数据包检测（扫描数据包内容以获取模式匹配）]，并且能够在不降低性能的情况下执行服务链，很可能就是改进的地方。在网络设备的数据平面中具有高性能的 C 型可编程网络处理器，将允许这些设备跟踪对标准的更改并对创新迅速做出响应。

4. 用于 NFV 加速的 NPS-400™ NPU 功能

NPS-400 是 Mellanox Technologies 的一个网络处理器，它能够很好地适用于 NFV 加速。与 Mellanox 的所有网络处理器一样，NPS-400 设计用于高带宽数据包处理。

下面的各个部分将介绍 NPS-400 如何在数据中心内解决实施 NVF 所面临的许多挑战。

4.1 性能

NPS-400 最少可以以 400 Gbps 处理以太网数据包（每秒 600M 数据包），并且最高可以以 960 Gbps 处理超额预订的 IO。NPS-400 上的数据包处理代码用 C 语言编写，并且在 CTOP（C 型可编程任务优化处理器）处理器阵列上使用运行到完成编程模型执行。NPS-400 的硬件线程允许以每秒 600M 数据包的处理速度进行每数据包多个表查找和多个统计计数器操作。在 NPS-400 上有 256 个 CTOP 处理器（内核），每个 CTOP 都有 16 个硬件线程可在硬件中进行快速线程调度。借助于在硬件中实施数据包顺序维护、计数器操作和搜索操作（DDR 内存和 TCAM），NPS-400 可以完美地融合处理能力、灵活性和软件指引的表查找功能，以提供 400 Gbps 的数据包处理。

4.2 负载平衡

在 NPS-400 上不需要将流固定到 CTOP 处理器，因为会在 CTOP 调度硬件中维护数据包顺序。因此，在 NPS-400 上不会将单个流的性能限制到单个内核的性能。可以将 NPS-400 用作 NFV 应用的负载平衡器，或者作为负载平衡器 VNF 的加速器。在 NPS-400 上，数据包分类不会局限于数据报文头的硬件分析，因为 C 程序可以进一步分析数据包，以构建简单或有状态的负载平衡器。对大型表的支持允许对数百万的流提供复杂的每流信息，且进行硬件调度不会占用处理器周期。

4.3 性能监控

NPS-400 的查找和统计引擎支持数百万的流，使用硬件令牌桶对每个流进行 SLA 监控。可以用简单令牌桶或耦合令牌桶来构建流量监管器，并且每流剖析可以让每个流符合不同的行业标准定义。NPS-400 还支持 1588 时间同步，从而可以为数据包或记录的数据提供精确的时间戳。可以为 CTOP 程序使用实时时钟，以便为内部事件提供精确的时间戳。

4.4 范围

NPS-400 有片上 TCAM（算法扩展到 DDR 内存）、片上 SRAM 和高达 48 GB 的外部 DDR-3 或 DDR-4 内存，从而能够为流量管理器 (TM) 查询提供大型查找表、大量计数器和巨大的数据包缓冲区。TCAM 支持算法扩展对于支持许多 vSwitch 和许多订阅者的大型 OpenFlow 表非常有用。对大型表的支持允许每流表支持数百万的流。NPS-400 的指令集可以轻松地以线速解析所有帧类型。标准控制帧在硬件中识别和解析以进行数据包调度；分配给每个数据包类型的优先级都可配置。其他寄存器允许客户让分类硬件识别其他协议类型、MAC 地址、标记和标签值，并根据需要分配优先级。

4.5 可靠性

NPS-400 支持快速故障切换到流量管理器中的备用链路，且无需更改转发数据库或路由表。NPS-400 也支持让 OAM (Y.1731) 监控与 OAM 对等的连接，并且触发快速故障切换到备用链路。内部和外部内存都有纠错码的保护，增加可靠性。

4.6 安全性

NPS-400 也具有对 AES-256 和 3DES 标准的高性能加密支持。加密的同时，NPS-400 可以使用 SHA-1、SHA-2 或 MD5 哈希函数计算身份验证 MAC。为了隔离用户组，NPS-400 使用已经过证明的 NPU 数据包编辑技术，以 400 Gbps 高效处理 VxLAN 或 NVGRE 封装和解封。叠加 (Overlay) 网络支持为多租户环境提供了安全保护。

4.7 流量管理

NPS-400 的嵌入式 TM 具有 1M 硬件队列和一个 5 级分层调度程序。每个队列和每个聚合调度程序实体都有整形器，以精确地符合出站服务水平协议。整形、优先级和加权公平队列 (WFQ) 调度的复杂调度组合可以精确地控制出站流量。

5. 代码示例

NPS-400 的内置功能是在 EZdp 库中提供的，使用添加到 CTOP 处理器的 NPU 指令来实现高性能的数据包处理。例如，CTOP 具有硬件指令可高效解析基本的 L2 和 L3 数据报文头。编译器中的内嵌 C 函数支持提供对 VLAN 标记、MPLS 标签堆栈、IPv4 报文头和 IPv6 报文头的以太网帧的单指令解析。

例如，要在单指令代码中将 IPv4 报文头解码到 **result** 结构：

```
status = ezdp_decode_ipv4 (header_ptr, size, frameSize, result);
```

header_ptr 是一个指向存储在内核本地内存中的 IPv4 报文头的指针，**size** 是 IPv4 报文头的大小，即要解码的字节数。解码指令衍生自 Mellanox NP-5 NPU（一个十多年来一直在交付高效数据包处理器的体系架构）的 TOPparse 指令集。

对数据包的各个部分执行查找操作，通常会从数据包或与数据包关联的元数据中的字段组成一个搜索键。NPS-400 在内部内存或外部 DDR 内存以及 TCAM 访问中均支持直接表、最长前缀匹配 (LPM) 和哈希表。

要在哈希表中查找一个 8 字节的键：

```
hashed_key = ezdp_hash_lookup_key 32(key.raw_data, 0, 4);
result.raw_data = ezdp_lookup_hash_entry (table,
                                          false, key_len, result_len, entry_len,
                                          hashed_key, &key, &entry, &scratchpad);
```

其中，**key** 是一个 32 字节搜索键结构，**hashed_key** 是该键的 32 位哈希值，**table** 是哈希表描述符，**false** 表示此表不是单循环哈希表，**key_len** 是搜索键的长度，**result_len** 是要接收结果的缓冲区的大小，**entry_len** 是要接收哈希表项（键加结果）的缓冲区的大小，**&entry** 是哈希表项本地内存中缓冲区的地址，**&scratchpad** 是用作临时工作区域的缓冲区。函数返回值是一个 C 型结构，其中会指示是否找到了有效项，以及结果数据的前 4 个字节是什么（如果适用）。函数的编码方式为，此结构将进入内核寄存器以高效测试查找结果。

在 NPS-400 的内部 TCAM 中进行查找的示例：

```
res = ezdp_lookup_int_tcam (side, profile, key_ptr, key_len,
                            mask, &result);
```

key_ptr 是查找的搜索键，查找结果将返回到 **result** 结构中。同样，提供了内置函数以实现 DMA 操作，从而能够进行高性能封装和解封操作。NPS-400 体系架构允许帧缓冲区偏移填充，从而允许进行叠加 (overlay) 网络报文头的标准 L2 封装，而无需分配和链接其他缓冲区。

```
status = ezdp_copy_frame_data (buff_desc, offset-hdr_size,
                               template_desc, 0, hdr_size);
```

以上示例使用 DMA 引擎从模板中复制 **hdr_size** 字节数，将其附加到与缓冲区描述符 **buff_desc** 关联的缓冲区前面。对于诸如 VXLAN 等更为复杂的叠加 (overlay) 网络，可以使用 NPS-400 中的 TCP/IP 和 VXLAN 协议处理器卸载进行终止。

NPS-400 加密指令可在对 CTOP 处理器周期带来最小影响的情况下对帧进行加密、解密和哈希。

```
ezdp_encrypt (sec_handle, plain_text, cypher_text, length);
```

基于安全句柄 **sec_handle** 的内容，加密指令将使用 AES、DES、3DES 或 RC4 密码加密 **plain_text** 缓冲区，将加密的信息存储在 **cypher_text buffer** 中。消息身份验证代码 (MAC) 指令将使用 SHA-1、SHA-2 或 MD5 哈希函数计算 MAC。

```
ezdp_mac_calculation (sec_handle, buf_ptr, first_flag, last_fFlag, length);
```

安全句柄 **sec_handle** 将识别哈希函数以应用于 **buf_ptr**、**first_flag** 和 **last_flag** 指向的 CMEM 中的缓冲区，并识别缓冲区是否为第一和/或最后一个文本段，而 **length** 是 **buf_ptr** 指向的缓冲区中的数据大小。消息摘要（哈希）在与 **sec_handle** 关联的安全上下文内存中进行更新。

数据包处理完成后，数据包将由 **ezdp_send_job_to_tm** 函数排入流量管理器 (TM) 队列中。在 NPS-400 体系架构中，一个作业处理一个数据包，作业描述符包含有关作业和正在处理的数据包的信息。

```
ezdp_send_job_to_tm (&job, &job_desc, job_desc.tx_info.side, 0);
```

嵌入式 TM 是一个 5 级分层硬件调度程序，使用在每个级别均可配置的加权公平队列 (WFQ) 和优先级调度对输出端口进行整形。TM 支持 1M 队列，并且在调度层次结构的上层聚合多个队列，以符合服务水平协议 (SLA)。借助于较大的缓冲区，TM 可以为高速链路提供无损服务或缓冲。每数据包超时允许在可接受的超时值以后恢复拥塞的缓冲区或流控制的队列。

用 C 语言编码的应用特定加速可以位于网络中的任何 NPS-400 上，并且与流的数据包处理加速集成在一起。例如，NFV 快速路径卸载网络地址转换 (NAT) 函数可以转换数据包中的 TCP 端口号和 IP 地址。用户可以编写自己的加速函数，用于转换其他报文头中的其他字段。

NPS-400 的其他功能包括 OAM、1588 时间戳、深度数据包检测 (DPI) 加速、加密、解密和哈希。提供一步和二步 1588 支持，它们可以为数据包的到达或离开提供精确的时间戳，用于监控应用程序、SLA 测量值或编排。DPI 可以以 200 Gbps 的速度对成千上万的规则执行一个粗粒度筛选，加密/解密加速器可以以合计 200 Gbps 的速度加密/解密，并且可以与加密/解密活动并行以 200 Gbps 的速度计算哈希。

6. 智能 NFV ToR

对于大多数 NFV 应用，可通过将多台服务器的 vSwitch 处理转为基于 NPS-400 的架顶式 (ToR) 交换机来提高性能。NPS-400 提供封装和解封功能以访问传送虚拟网络流量的叠加 (overlay) 网络，并且将标准以太网帧传递到每台服务器上的 LOM 设备。如果 LOM 设备具有 SR-IOV 或隧道功能，则整个机箱都将获得高性能。将服务器机架的 vSwitch 整合到架顶式交换机将明显减少托管虚拟交换机的数量，同时提高服务器的性能。

在数据平面密集的 VNF（例如路由器或负载平衡器）中，数据平面处理可能会占去大多数的处理器周期并变成瓶颈。在这些情况下，卸载数据包处理将提供更高的性能，并且在某些情况下将获得正常网络行为。NFV 交换机和路由器可以卸载其大多数或全部数据平面查找操作、帧修改、转发决定和排队，同时 VNF 执行异常转发功能和控制平面处理。其他 NFV 应用可以卸载 TCP/IP 堆栈、负载平衡决定、深度数据包检测或加密处理，以便更有效地使用处理器运行其余 NFV 功能。最终结果就是 NFV 性能提高和容量变大。

服务器上 LOM 或网卡设备中的单根 IO 虚拟化 (SR-IOV) 允许在 ToR 交换机中高效地实施 vSwitch，以便让帧直接进入 VNF 缓冲区进行 DMA。在 NFV 应用中使用基于 Hypervisor 的 vSwitch 会降低 SR-IOV 加速技术的效力。此外，随着网络规模的增加，vSwitch 流和寻址表可能会变得非常大，足以影响所使用服务器内存和服务器上数据缓存内存的效力。SR-IOV 可通过消除帧复制而降低 VNF 处理的延迟，并通过提高缓存命中率来提高运行 VNF 的处理器/虚拟机的性能。

在 NPS-400 设备上使用 NFV 加速可提高性能和安全性，同时降低 NFV 应用的功耗。对于共享同一物理网络的多个租户来说，虚拟以太网端口聚合器 (VEPA) 报文头将识别虚拟网络，并且 ToR 交换机可以根据 802/1Qbg 边界虚拟桥接标准做出转发决定。通过 VEPA，同一服务器上虚拟机之间的流量可在网络上获得更好的安全性，因为 ToR 设备可以在硬件中应用 ACL 检查，然后如果 ACL 测试中编写的策略允许通信，则 hairpin 将数据包转发回同一台服务器。因为 ACL 测试通常在 NPS-400 上的 TCAM 硬件中执行，所以可以以线速检查数百万的规则。由于会对处理器带来性能影响，因此一个正常的 vSwitch 不会实施 ACL 检查，或者仅检查几个规则。

交换机、路由器或设备中的单个 NPS-400 可以以与整个服务器机架相同的速率处理大多数数据平面应用的数据包。因为基于 NPS-400 的 TOR 交换机与一台服务器所耗功率大约相同，所以使用基于 NPS-400 的加速将显著降低数据包处理的功耗。

7. 智能边界路由器设计

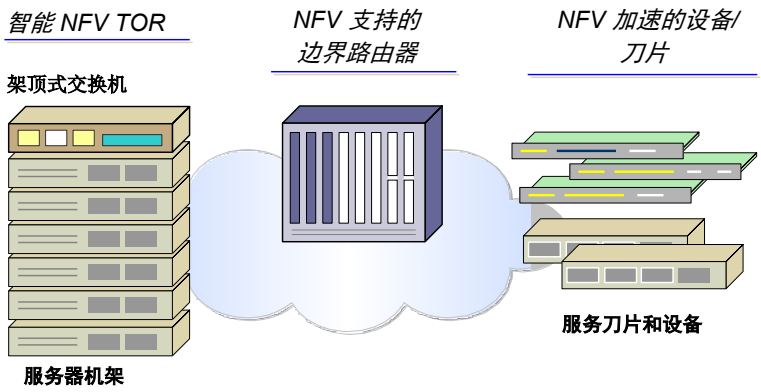
进入数据中心的互联网流量通常由边界路由器进行处理，边界路由器在路由器线卡或服务卡上具有数据包处理功能。例如，边界路由器将为流提供 VXLAN 叠加 (overlay) 网络报文头，从而让数据包可以安全地通过多租户数据中心网络。对于 NFV 来说，边界路由器可以执行深度数据包分类并将数据包映射到流和 NFV 服务链，同时跨许多分布式 VNF 动态地对流进行负载平衡。许多这类设备都使用 OpenFlow 表和操作，并且可以提供 NFV 加速作为 OpenFlow 表处理的一部分（请参见以上有关 NFVI 网络加速的讨论）。在边界路由器中使用 NPS-400 进行数据包处理将为 NFV 加速提供一个灵活的平台，并且允许在新加速功能中进行创新。对于使用单 NPS-400 的 2x100G 线卡或者 6x40G 线卡，NPS-400 的数据包处理功能提供全双工线速处理。也可以使用 NPS-400 构建更高密度的超额预订线卡。边界路由器执行大量的数据包处理，并且 NPS-400 允许使用具有算法扩展的片上 TCAM 对数百万的规则进行线速访问控制列表检查，还允许对隧道和叠加 (overlay) 网络进行封装和解封，以及终止安全连接。NPS-400 还支持数百万个流的 OAM、1588 和流量监管，从而允许高服务交付以及与现有网络管理协议的集成，同时预防攻击。

不管是入口还是出口流量，NPS-400 流量管理器都改进了服务交付。在入口，可以使用分层调度和整形来聚合经过受限链路或设备的流并进行整形，从而保护网络和服务器不会过载。在出口，TM 的大量队列、整形器和聚合实体让各个流和/或用户组符合 SLA，从而避免由于城域网络中的流量监管功能而出现丢包现象。

8. 智能加速设备设计

NFV 加速设备可以与内核或主干交换机相邻放置，对许多服务器机架或多个数据中心提供 NFV 加速服务。该设备是一个具有 NFV 加速功能的纯数据包处理器，它实施 NFV 基础架构小组规范中的加速基元，从而允许设备加速多个供应商的多个 VNF 的 NFV 处理。例如，请考虑使用 FPO 加速的 NFV 负载平衡器，除非网络交换机不支持 FPO。设备可以作为“单臂路由器”连接到主干交换机，在进行初始流检测和负载平衡决定后，将在设备中加速处理流中的流量。图 5 显示了部署在 ToR 交换机、边界路由器或作为设备或服务刀片部署的加速。

图 5. 各种智能设备中的 NPS-400 部署



可以将 NFV 加速资源池部署为刀片机箱中的多个服务刀片，或者多个设备箱。请注意，单个 NPS-400 或许能为许多 VNF 提供一个加速器池，例如，构建于 NPS-400 上的一台 ToR 交换机能够加速服务器机架中的所有 VNF。

9. 白盒设计

随着基于 SDN 和 NFV 的新网络体系架构作为运营商和数据中心网络的前进之路出现，白盒系统的相关发展势头变得越来越猛。与知名供应商提供的专有网络设备完全不同，白盒系统源自制造现成硬件的非品牌制造商 (ODM)。白盒的特定网络功能通过下载到白盒的软件进行确定，通常由应用程序软件供应商提供。可以将白盒部署为服务器、架顶式交换机、网络设备等。随着 SDN 和 NFV 对迁移到基于软件的虚拟网络功能的推动，也为白盒释放出了大量机会，可以在白盒上运行或加速虚拟（软件）网络功能。由于强调 SDN 和 NFV 可降低运营支出和资本支出的承诺，白盒网络进一步压低了成本，并且促进了硬件和软件供应商生态系统的形成。

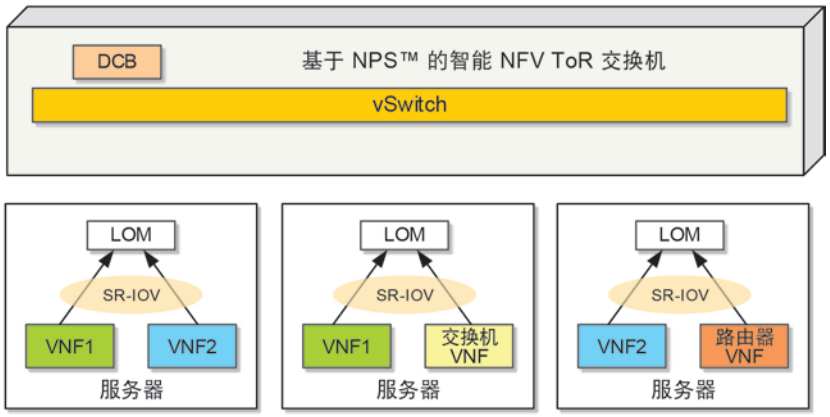
能否迁移到白盒网络的关键在于白盒核心处使用的强大商业芯片的供应情况。当市场提供众多商业芯片时，很明显像 NPS-400 这样能够以更高的吞吐量提供更丰富功能的芯片能够支持更加广泛的虚拟功能，可以将这些功能部署到白盒中并且给网络运营商带来更大的价值。

相对于通常基于不可编程交换芯片（功能有限或者 CPU 的性能有限）的其他系统，在白盒系统的核心使用 Mellanox 的 NPS-400 为 ODM 供应商和软件应用程序供应商提供了更大的优势。NPS-400 提供了独特的 NPU 性能与 CPU 功能组合。400 Gb/s 的高吞吐量、C 语言编程、Linux® 操作系统和 7 层数据包处理功能，既可加速数据平面处理，又可以让虚拟功能发挥出史无前例的功能和性能水平。这些功能包括交换、路由、高级分类、流量管理、ACL（访问控制列表）、有状态流表、负载平衡、安全（防火墙、IPsec 和 SSL VPN）、DPI（深度数据包检测）、网络监控、应用识别、订阅者管理、TCP 终止和网络叠加 (overlay) 终止。

10. 软件体系架构

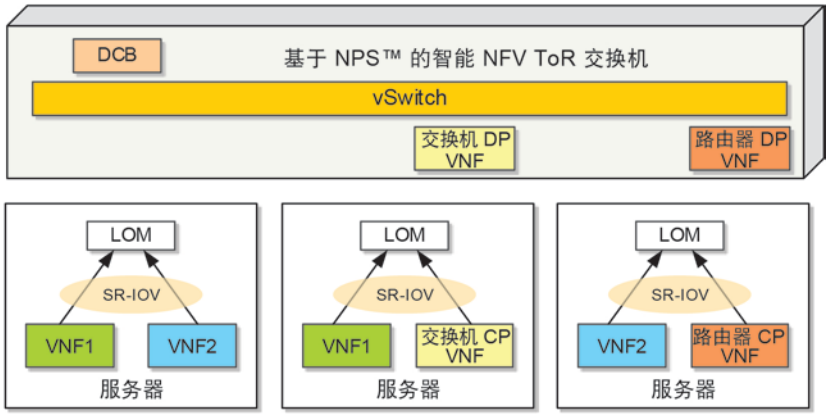
图 6 显示了基于 NPS-400 的 ToR 交换机软件体系架构，在该体系架构中，针对 VNF 机架，使用的是 vSwitch。如果物理服务器上的 LOM 设备能够进行 SR-IOV 或隧道传输，则 vSwitch 将与 VNF 直接通信，从而避免必须在 Hypervisor 中缓冲数据包。当只有 VNF 使用数据包数据时，SR-IOV 传输可避免缓存污染，即使用 Hypervisor 缓冲区内容填充处理器缓存。更好的缓存性能可提高服务器性能，并且卸载 vSwitch 处理可以让更多的周期用于驻留在处理器中的 VNF。为机箱准备单个 vSwitch 可减少数据中心内的交换机数目，从而降低管理负载并使网络的响应速度更快。

图 6. 智能 NFV ToR vSwitch 的软件体系架构



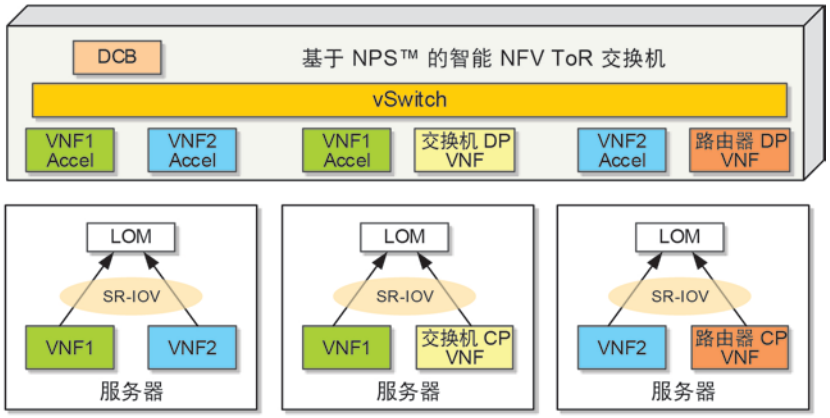
也可以加速其他 VNF，请注意，图 6 显示了一个交换机 VNF (L2) 和一个路由器 VNF (L3)，这两个虚拟设备都使用 NFV 基础架构中的 vSwitch。这两个 VNF 有应用程序的数据平面和控制平面部分，并且可将数据平面功能卸载到 NPS-400，如图 7 所示。

图 7. 在智能 NFV ToR vSwitch 中软件卸载到 NPS-400



数据平面的 NFV 加速基元是 OpenFlow 表项和操作。交换机 DP VNF 可能包括具有 ACL 检查的 VEPA 处理。数据平面上的大多数流量均由 ToR NPS-400 处理，并根据 OpenFlow 表项转发至出口端口，且不会干扰控制平面。如果检测到控制平面流量，则通过 vSwitch 将数据包发送到控制平面，从而使用 SR-IOV 传输最大程度降低对服务器缓存的影响。通过卸载交换机和路由器 VNF 的数据平面处理，并且通过避免缓存性能的降级，再一次提高了服务器的性能。以后可能会出现其他类型的加速基元，例如，以后的一组 OpenFlow 操作可能会允许应用识别 VNF 将正则表达式测试填充到 OpenFlow 交换机。如果图 7 中绿颜色的 VNF 是应用识别 VNF 的两个实例，那么图 8 显示了如何与数据平面的硬件加速器一样，将正则表达式检查部署到基于 NPS-400 的同一 ToR 交换机。

图 8. 在同一 ToR 交换机上添加正则表达式



具有应用识别的 VNF 可以使用一组深度数据包检测规则（正则表达式），请求 ToR 物理交换机处理链路上的所有数据包，并且仅将需要其他处理的数据包转发给防病毒 VNF。使用加速设备可通过给 NFV 服务器卸载来提高系统性能，并且可通过在加速器中处理大量数据包，让交付给用户的服务更可预测，从而能够更快地做出响应并减少网络工作负载。

11. 结论

Mellanox NPS-400 加速的 VNF 既能够交付高网络性能，同时又能够为最苛求的数据平面应用提供灵活的平台。NPS-400 是一个网络处理器 (NPU)，它能够卸载数据平面处理，并且可通过让更多的处理器周期可用于 VNF 并且凭借更好的缓存利用率提高这些周期的性能，提高基于标准服务器（具有通用 CPU）的 VNF 的性能。VNF 让处理器更多地用于控制平面和异常处理，让数据平面数据包的延迟更短，并且物理网络可通过避免某些回程流量而获得更好的链路利用率。由于通过向网络基础架构添加 NPS-400 加速了许多服务器的性能，因此降低了每个数据包的整体功耗。在交付全部网络服务的背景下，NPS-400 可在保持数据包顺序的同时交付高性能的数据包处理，准确地进行流量监管和整形，并且可以缓冲大量数据，以确保无损服务，保留统计信息或镜像流量。与 DPDK 不同，NPS-400 上的 NFV 加速可以位于网络中的最佳地点，可以是 ToR 交换机、边缘路由器或加速设备，具体取决于应用和工作负载。由于在数据平面中保留着传统网络行为，因此在大范围的 NFV 部署中，将 NPS-400 加速的 VNF 集成到现有网络将更加容易，并且不再需要在网络完整性、性能或功耗上做出妥协。

北京市朝阳区望京东园七区保利国际广场 T1 15 层
Tel: 010-5789 2000
www.mellanox.com