Intel® 架构和 OpenStack* Juno 版本铺就线速 NFV 部署之路

执行摘要

网络功能虚拟化（NFV）小组是欧洲电信标准协会（ETSI）的一个行业规范小组（ISG），致力于定义适用于革命性网络的规范。这种革命性网络基于将整体、垂直集成、离散的应用程序转换为部署在行业标准高容量服务器上的虚拟化、可扩展应用程序的原理。

基于 Intel® 架构的服务器提供了一些功能，可以使用它们来优化和加速这些虚拟化的 NFV 应用程序的部署。一些平台中包含大量特定于虚拟化的增强，这些平台包括 Intel® Virtualization Technology（Intel® VT）for IA-32、Intel® 64 and Intel® Architecture（Intel VT-x）、Intel® Virtualization Technology（Intel® VT）for Directed I/O（Intel® VT-d）和 Intel® Virtualization Technology（Intel® VT）for Connectivity（Intel® VT-c）。基于 Intel 架构的服务器上还有一组配置功能，这些功能有助于提高 NFV 应用程序的性能，为其提供可预测特征。

充分利用增强平台感知特性简化线速 NFV 应用程序的部署，比如 SR-IOV、NUMA、大页面和 CPU 锁定等。

OpenStack* 是一个创建私有云和公有云的著名开源软件套件，可用作 ETSI-NFV 参考架构中的虚拟化基础架构管理器。为此，需要添加一些特性才能满足 NFV 应用程序的性能和可预测性需求。不必将固有的 NFV 应用程序知识嵌入到 OpenStack 中，但需要 OpenStack 扩展来提供足够的调优功能，才能部署高性能 NFV 工作负载。

参考 NFV 应用程序 Intel® Data Plane Performance Demonstrator 的实验室测试表明，通过利用增强平台感知特性，比如 Single Root I/O 虚拟化（SR-IOV）、非统一内存架构（NUMA）、大页面和 CPU 锁定，可以简化线速 NFV 应用程序的部署。

2015 年 3 月

Adrian Hoban | Przemyslaw Czesnowicz | Sean Mooney | James Chapman | Igor Shaula | Ray Kinsella | Christian Buerger | Intel 公司

目录

简介 ---------- 2
网络功能虚拟化 ---------- 2
OpenStack ---------- 2
宽带网络网关 ---------- 3
Intel® Data Plane Performance Demonstrator ---------- 3
文档范围 ---------- 4
OpenStack 扩展让 NFV 受益 ---------- 4
CPU 特性请求 ---------- 4
SR-IOV 扩展 ---------- 4
NUMA 扩展 ---------- 4
未来的 OpenStack 功能 ---------- 5
CPU 锁定 ---------- 5
大型页面 ---------- 5
I/O 感知的 NUMA 调度 ---------- 5
示例 NFV 性能结果 ---------- 5
实现最优 NFV 的 Intel® 技术 ---------- 6
针对 IA-32 和 Intel® 64 处理器的 Intel® 虚拟化技术 ---------- 6
Intel® Virtualization FlexPriority ----- 6
Intel® Virtualization FlexMigration--- 6
虚拟处理器标识符（VPID） ---------- 7
扩展页面表 ---------- 7
VMX 抢占计时器 ---------- 7
暂停退出循环 ---------- 7
针对 IA-32 和 Intel® 64 处理器的 Intel® 虚拟化技术 ---------- 7
地址转换服务 ---------- 7
大型 Intel VT-d 页面 ---------- 7
中断再映射 ---------- 7
针对连接的 Intel® 虚拟化技术 ---------- 7
虚拟机设备队列 ---------- 7
PCI-SIG SR-IOV ---------- 7
非统一内存架构 ---------- 8
CPU 锁定 ---------- 9
大页面 ---------- 9
结束语 ---------- 10
参考资料 ---------- 11

简介

本文将介绍网络功能虚拟化和一些与 OpenStack 相关的活动，它们有助于在电信环境中所使用的行业标准的、高容量的服务器上部署高性能工作负载。

网络功能虚拟化

网络功能虚拟化（NFV）小组是欧洲电信标准协会（ETSI）的一个行业规范小组（ISG）。[Ref 1] 2012 年，全球许多著名运营商一起在 ETSI 中组建了这个规范小组。这个 ISG 的作用是促进电信行业众多企业的协作，从而创建一些规范，用这些规范来加快行业标准的高容量服务器上的电信工作负载的研究、开发和部署。

该活动最终贡献了 3 本白皮书。第一本白皮书[Ref 2] 是介绍性的，为要开展的工作奠定基础。第二本[Ref 3] 和第三本[Ref 4] 白皮书从网络运营商的角度提供了对行业实现规定 NFV 目标的进度的见解。

在这些白皮书发表仅 10 个月之后，就涌现了大量规范草案，这些规范随后被更新并公开。[Ref 5] 具体地讲，NFV 性能和可移植性最佳实践规范[Ref 6] 与本文的介绍范围相关。我们确定了许多平台功能，在平台中正确配置或利用这些功能才能实现高性能的网络应用。

OpenStack

OpenStack 是一个用于创建私有云和公有云的领先开源软件套件。[Ref 7] 它所提供的大部分功能都可以划分到具有类似含义的名称下，比如云操作系统或虚拟化基础架构管理器（VIM）。OpenStack 被用于管理计算、网络和存储基础架构池。这些基础架构资源通常基于行业标准的、高容量的服务器，可以通过命令行接口（CLI）、RESTful API 或 Web 接口，根据需要为用户分区和配备这些资源。OpenStack 于 2010 年向开源社区发布，自那时起变得越来越流行，现在，它已经有了一个活跃的用户和贡献者社区。该代码是依照 Apache 2.0 许可进行发布的。

OpenStack 计算服务被称为 Nova*。它负责管理 OpenStack 托管云中的所有计算基础架构。许多虚拟机管理程序驱动程序都受到支持，其中包括 QEMU*/KVM*（通过 libvirt）、Xen* 和 VMware vSphere*Hypervisor（VMware ESXi*）。Nova 包含调度功能，可用来选择哪个计算主机在运行某个特定工作负载。它根据输入参数来过滤所有可用的平台，找到一个合适的平台子集，然后根据权重例程从该子集中选择一个平台。

OpenStack 网络服务被称为 Neutron*。Neutron 被设计为一个可扩展的服务，它提供了许多不同的插件解决方案来帮助提供商执行网络管理，还提供了一个自助接口，供用户创建自己的工作负载。有一些网络模型可供使用，比如平面网络、虚拟局域网（VLAN）、VXLAN 等。IP 地址管理（IPAM）支持包括静态 IP 地址分配、动态主机配置协议（DHCP）和浮点 IP 支持，后者允许在基础架构中动态地将流量重新路由到不同的计算节点。此外，还可以管理一些高级网络服务，比如负载平衡器、防火墙、虚拟专用网络（VPN）。

还有其他一些服务，这些服务共同形成了一个 OpenStack 版本。

OpenStack 仪表板服务被称为 Horizon*。它为用户提供了通过 Web 接口管理其基础架构的能力。OpenStack 存储服务同时支持块存储和对象存储。存储支持可通过 Cinder*、Swift* 和 Ceph* 项目公开。还有一组共享服务，比如映像服务（Glance*）用于对映像执行导入、注册、创建快照等操作，身份服务（Keystone*）用作一个常见的身份验证系统和用于服务见消息授权。

宽带网络网关

宽带网络网关（BNG）是服务提供商的网络中的一个组件，用于在互联网与远程接入设备之间路由流量，这些设备包括数位用户线接入复接器（DSLAM）、电缆调制解调器终端系统（CMTS）或多服务接入节点（MSAN）等。它有时被称为宽带远程接入服务器（BRAS）。

BNG 是服务提供商的网络中的一个基于策略的服务质量执行点，是用户的链路访问协议（比如基于以太网的对等协议（PPPoE））的逻辑终端。

基于 ATM 的对等协议（PPPoA）。它通常是用户设备（远程访问客户端设备）在与互联网建立连接时寻找的第一个 IP 跃点。BNG 需要将高性能网络连接与较低的包延迟和较低的包延迟变化相结合，才能有效地运行。

应该将这个示例 BNG 视为一种协议转换应用程序。BNG 参考架构如图 2 所示。该 BNG 基于一个流水线式软件架构，在该架构中，包处理工作负载的不同部分被分开并分配给不同的线程。平衡各个流水线阶段，尽可能地减少阶段之间的等待时间。

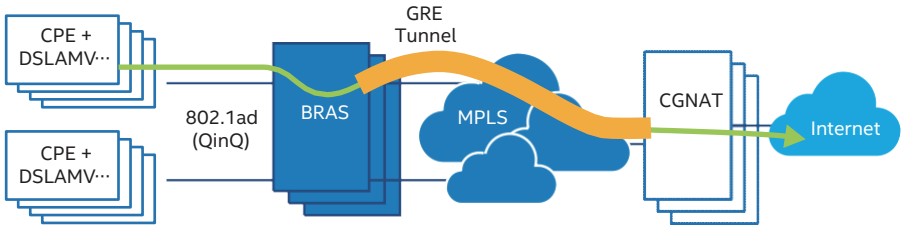

图 1. 一个服务提供商网络中的宽带网络网关部署。

Intel® Data Plane Performance Demonstrator

需要一种虚拟化网络功能（VNF）来演示如何使用 OpenStack 来配置与 VNF 的高性能网络连接，并对 VNF 本身启用高性能行为。Intel® Data Plane Performance Demonstrator（DPPD）[Ref 8] 是一个基于 DataPlane Development Kit（DPDK）v1.7 的应用程序。[Ref 9] Intel DPPD 是一个用户空间应用程序，被设计来执行基准测试，演示如何使用 DPDK 来处理与真实 BGN 流量非常类似的高密度网络流量，研究这样的工作负载下的平台性能。Intel DPPD 实现了许多典型的 BGN 功能，比如正确处理构造的 BGN 包；但是，未实现异常路径处理。该软件是依照 BSD 3-Clause 许可发布的。[Ref 10] Intel DPPD 部署在一个基于 QEMU/KVM 的虚拟环境中。

cpe 0 和 cpe 1 是连接到用户端设备（CPE）端网络的网络接口，inet 0 和 inet 1 是连接到（互联网）核心端的网络接口。对于上行链路路径，CPU 核心负载平衡器（LB）将会处理从 cpe 0 次传入的流量，将它分发给 8 个工作线程（WT）中的一个线程。完成对选中的工作线程的处理之后，流量会被定向到一个聚合线程（A），该线程将数据传输到专用的出口接口。下行链路数据路径遵循某种类似的模式。

备注：白皮书《网络功能虚拟化：使用 Linux* 和 Intel® 架构实现虚拟化 BRAS》中提供了与这个示例应用程序相关的包处理性能的详细分析。[Ref 11]

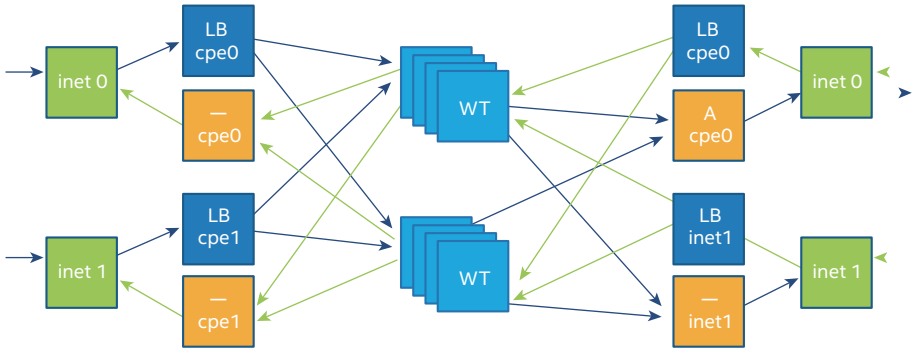

图 2. Intel® Data Plane Performance Demonstrators（BNG）软件架构。

文档范围

本文主要关注的是网络功能虚拟化，以及针对 OpenStack 的扩展如何帮助以最优方式将 NFV 工作负载部署在基于 Intel 架构的处理器上。推荐读者参阅其他特定的处理器产品文档，确定本文中提到的处理器特性是否适用于读者选定的处理器。

OpenStack 扩展让 NFV 受益

从 NFV 角度讲，OpenStack 不需要提前知道 NFV 应用程序或它们的功能。但是，OpenStack 需要能够提供一些精心选择的高级调优功能，让服务提供商能够部署具有必要的性能特征和效率特征的 NFV。

OpenStack 特性和功能正在快速发展。这一节将介绍一些特别适用于有效部署 NFV 工作负载的特性。

CPU 特性请求

可能已经开发了一些 NFV 应用程序来专门使用某些 CPU 指令。例如，某个 VPN 设备要求编写一个高性能密码库来利用特定的指令，比如 Intel® 高级加密标准新指令（Intel® Advanced Encryption StandardNew Instructions, Intel® AES-NI）。[Ref 12] 最佳实践指南建议，在调用利用指令集扩展的代码之前， 软件开发人员应该通过 cpuid 指令检查这些扩展的可用性。

在 OpenStack Icehouse 版本中，对 Nova libvirt 驱动程序进行了更改，以便向 Nova 调度程序公开所有 CPU 指令集扩展。这与 libvirt 中的某处更改有关，该更改旨在让此数据可以通过 libvirt API 进行公开。这些更改支持通过将特定的特性请求以 extra_specs 形式添加到 Nova flavors 中，创建包含这些请求的 flavor。

在调度期间，Nova 中的 compute_capabilities_filter 将 flavor extra_specs 指定的主机 CPU 上的需求与某个主机数据库和它们各自的 CPU 特性进行比较。

SR-IOV 扩展

OpenStack Havana 版本包含对非连网设备的 SR-IOV 支持。这包括将 VF 从 PCIe 密码加速器（比如 Intel® QuickAssist Technology）分配给 VM 的能力。在 OpenStack Juno 版本中，这项功能已扩展，包含对网络设备的支持。通过这种方式，可以在物理 NIC 到 VM 之间建立最高性能的 I/O 路径。

需要执行一些步骤才能使用针对 NIC 的 SR-IOV 扩展。

- 在主机上，必须配置 NIC 的设备驱动程序来启用 NIC VF。
- 主机上的 Nova 配置文件 nova.conf 需要配置白名单，让 VF 标识符可以与 Nova 调度程序共享。
- Neutron 要求使用 SR-IOV 机制驱动程序和设置 VF 供应商和设备 ID。
- 对于 Neutron API，租户必须创建一个具有以下类型的端口：虚拟接口（VIF）类型（vif_type）macvtap 或 direct。后者意味着 VF 将直接分配给 VM。
- 从 Neutron port-create 命令返回的端口 ID 必须添加到 Nova boot 命令行。备注：在引导期间，Nova 会检查这个端口 ID 对于 Neutron 的有效性。

NUMA 扩展

在 OpenStack Juno 版本中，可以通过虚拟驱动程序 Guest NUMA 节点放置和拓扑结构扩展来添加了对 NUMA 拓扑结构的

感知能力。[Ref 13] 此特性允许租户指定其想要的 Guest NUMA 配置。Nova 调度程序使用 numa_topology_filter 进行了扩展，以帮助匹配来宾 NUMA 拓扑结构请求与主机的可用 NUMA 拓扑结构。

租户可通过基于某种 Nova flavor 的机制来指定其请求。该命令的一个示例是：

```
nova flavor-key m1.large set hw:numa_
mempolicy=strict hw:numa_cpus.
0=0,1,2,3 hw:numa_cpus.
1=4,5,6,7 hw:numa_
mem.0=1024 hw:numa_mem.1=1024
```

租户还可以选择通过一种基于映像属性的机制来指定其 Guest NUMA 拓扑结构请求。该命令的一个示例是：

```
glance image-update image_ id
–property hw_numa_ mempolicy=
strict –property hw_numa_cpus.
0=0,1,2,3 –
property hw_numa_cpus.1=4,5,6,7
–property hw_numa_mem.0=1024 –
property hw_numa_mem.1=1024
```

这些命令会导致 OpenStack 将 Guest 虚拟 CPU 1、2 和 3 映射到插槽 0（在 libvirt 术语中也称为单元 0），并将虚拟 CPU 4、5、6 和 7 映射到插槽 1。

未来的 OpenStack 功能

备注：在撰写本文时，已建议将这一节中介绍的特性包含在后续 OpenStack 版本中，比如 Kilo 和 Liberty 版本。

CPU 锁定

上面有关 OpenStack NUMA 配置的章节将介绍如何请求 Guest NUMA 拓扑结构。在提供的示例中，Guest vCPUs 0、1、2和 3 被映射到插槽 0（也称为 “NUMA 单元 0”）。但是，这并不意味着 vCPU 0 被映射到物理 CPU（pCPU）0 或 vCPU 1 被映射到 pCPU 1，等等。

主机调度程序仍能在该 NUMA 单元内灵活地移动线程。这会影响 NFV 应用程序，比如要求线程锁定到 pCPU 来遵守特定应用程序设计限制，或者要求满足特定的延迟或可预测性指标的虚拟 BGN。未来的 Open-Stack 版本预计会进行扩展，提供请求将 vCPU 锁定到 pCPU 的能力。

大型页面

OpenStack Juno 版本无法识别主机上的大型页面功能。

Intel 在 www.01.org 网站上发布了 Open-Stack Juno 版本的补丁，以启用与 Juno 功能相关的大型页面支持。[Ref 14] 对这个补丁集的更新预计会包含在未来的某个 OpenStack 版本中。Intel 还发布了一个指南 [Ref 15] 来帮助用户利用这些更改。

I/O 感知的 NUMA 调度

上面有关 OpenStack NUMA 配置的章节介绍了如何请求 Guest NUMA 拓扑结构。此配置未考虑可能向处理核心提供数据的 I/O 设备的位置。例如，如果某个应用程序要求将 vCPU 核心分配给某个特定的 NUMA 单元，但传入数据的 NIC 位于另一个 NUMA 单元内，那么应用程序的性能可能不那么理想。

Intel 在 www.01.org 网站上发布了 Open-Stack Juno 版本的补丁，以便启用与 Juno 功能相关的 I/O 感知的 NUMA 调度。[Ref 14] 对此补丁集的更新预计会包含在未来的某个 OpenStack 版本中。Intel 还发布了一个指南[Ref 15] 来帮助用户利用这些更改。

示例 NFV 性能结果

Intel DPPD 被实际应用在 Intel® 服务器主板 S2600GZ 上（代号 Grizzly Pass），BIOS 版本为 SE56600.86B.02.03.000，包含两个 Intel® 至强® 处理器 E5-2690 v2 @3.0 GHz，每个处理器上有 10 个核心（20 个线程），64 GB DDR3-1333 RAM，一个 Intel® 以太网服务器适配器 X520- SR2 和两个物理网络。

OpenStack Juno 版本被用于在该平台上使用示例 BRAS/BNG 应用程序实例化 VM。计算节点上的主机操作系统和来宾操作系统都使用了 Fedora* 20 x86_64。

在撰写本文时，Intel DPPD 尚未为了与 OpenStack Juno SR-IOV 功能进行互操作而更新，所以 SR-IOV 结果基于在 Neutron 控件外设置的 SR-IOV 连接。

大型页面功能是 OpenStack Kilo 的一个目标特性，所以此功能通过利用 www.01.org 上的 OpenStack Juno 补丁来启用。[Ref 14]

DPPD 应用程序要求将 16 个虚拟 CPU 核心分配给该应用程序。为了实现最佳性能，所有这些 vCPU 都必须来自服务器上的同一个插槽。

利用了 OpenStack NUMA 放置功能来帮助将所需的 Guest vCPU 拓扑结构部署在主机上。OpenStack Juno 功能不支持在 NUMA 配置决策中感知 I/O。这意味着无法指定 Guest 拓扑结构包含 I/O 设备位置。（有关的更多信息，请参阅 “I/O 感知的 NUMA 调度” 部分）。

为了处理该功能的缺失，我们测试和检查了测试环境，以确保平台 I/O 设备对于为了部署 VM 而选中的 NUMA 单元是本地设备。

预计未来的 OpenStack 版本将包含对在制定 NUMA 位置决策时感知 I/O 的支持。另外，OpenStack Juno 中也不支持 CPU 锁定，但计划将该功能添加到未来的版本中。对于测试设置，CPU 锁定是使用 Linux 原语来执行的。

图 3 显示了对该环境应用这些配置更改所获得的累积收益。相对改进依赖于启用特性的顺序，所以这些相对的数据点不应解释为绝对值。读者应考虑启用这 4 种特性的累积价值，以及对于 10 Gbps 以太网上的 256 字节包，如何让这个参考 NFV 工作负载实现零包丢失率的线速性能。

延迟和包延迟变化特征是性能改进的其他方面，图 3 中给显示，但与 NFV 工作负载相关。在此测试中未执行具体的度量，但可在连续的测试运行中观察到吞吐量读数的可预测性和稳定性的明显改善，尤其是在启用 CPU 锁定之后。

已有人对 DPPD 执行了一些包含延迟度量的相关研究。文章《使用 Linux* 和 Intel® 架构® 的宽带远程接入服务器中的服务质量》[Ref 16] 提供了许多专门针对延迟特征的结果。使用 OpenStack 部署的 DPPD 的延迟和 PDV 可能是未来的某篇文章的分析范围。

使用 OpenStack 部署基于 DPDK 的虚拟 BRAS
（用 10 Gbps 的百分比表示的累积收益*）

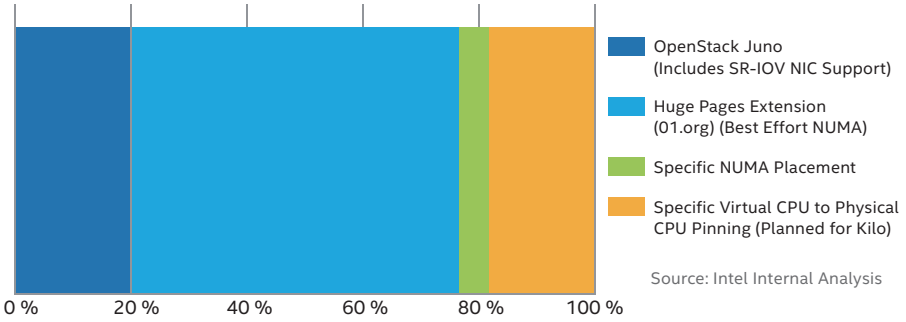

图 3. 平台优化对 Intel® Data Plane Performance Demonstrators 的累积性能影响。

实现最优 NFV 的 Intel® 技术

要在行业标准、高容量的服务器上高效地部署虚拟化、高性能的网络应用程序，可利用许多处理器和平台功能来实现显著的效果。

针对 IA-32 和 Intel® 64 处理器的 Intel® 虚拟化技术

Intel VT-x 添加了扩展来实现高性能虚拟化解决方案，帮助提高服务器利用率和降低成本。这些扩展被称为虚拟机扩展（VMX）。VMX 向指令集中引入了 10 个特定于虚拟化的指令。

Intel® Virtualization FlexPriority

本地高级可编程中断控制器（APIC）有一个寄存器用来跟踪当前运行的进程优先级，这个寄存器叫做任务优先级寄存器（TPR）。需要使用 VMM 来跟踪 Guest 向虚拟化 TPR 写入的数据，从而了解何时可以安全地向 VM 注入中断。这意味着每次 Guest 向虚拟化的 TPR 写入数据时（通常发生在任务上下文切换时），就会生成一个 VM退出。Intel® VT FlexPriority 支持功能消除了对 VMM 跟踪向虚拟化 TPR 寄存器的写入操作的需求，从而减少了 VM 退出的数量。

Intel® Virtualization FlexMigration

CPUID 指令集用于报告某个处理器的特性集。在初始化应用程序时， 通常使用 CPUID 指令来检查特性是否可用。要支持实时迁移，也就是宕机时间极少、无需重新启动 VM 的迁移），目标处理器必须支持与源处理器相同的特性集。Intel® VT FlexMigration 使 VMM 能够虚拟化 CPUID指令。这使 VMM 能够控制向 Guest 公开哪些特性，从而使 VMM 能够管理一个处理器池中公开的特性。不同的处理器可能具有不同的特性集，但 VMM 可以通过虚拟化的 CPUID 指令让它们看起来对 Guest 是相同的。

虚拟处理器标识符（VPID）

MMU 拥有一个转换后援缓冲器（Translation Look-aside Buffer, TLB）来缓存从虚拟到物理内存的地址转换。TLB 使用一个虚拟处理器 ID（VPID）字段进行了扩展。该字段使与 VM 相关的地址转换条目能够在上下文切换时持久保存在 TLB 中，这消除了在上下文切换时重新填充 TLB 的需求，使得无需执行以前关联的 TLB 擦除操作就可以执行 VM 上下文切换。

扩展页面表

扩展页面表（Extended Page Tables, EPT）是 MMU 的一个特性。EPT 特性使虚拟管理器无需 VM 更新捕获到页面表中（一个称为页面表影子的特性），从而提高了性能。EPT 在 VMM 中维护一个二级页面表来描述 Guest 物理地址到主机物理地址的映射，从而取代了页面表影子的功能。

VMX 抢占计时器

VMX 抢占计时器（VMX Preemption Timer）是一个额外的平台计时器，它使虚拟机管理程序能够在特定时间量后抢先执行 VM。这样就无需使用另一个平台计时器来方便上下文切换。

暂停退出循环

暂停退出循环（Pause Loop Exiting, PLE）是一个硬件特性，它会在检测到自旋锁循环持续时间过期事件时强制 VM 退出。这使 VMM 能够调度另一个 vCPU，有助于减少锁持有者抢占的影响。

针对 IA-32 和 Intel® 64 处理器的 Intel® 虚拟化技术

地址转换服务

外围组件互联特殊兴趣小组（Peripheral Component Interconnect Special Interest Group, PCI-SIG）定义了地址转换服务（ATS）规范，使用它作为 I/O 虚拟化（IOV）规范的一部分。ATS 指定了一组事务，PCI Express 组件可使用这些事务来支持 I/O 虚拟化，让 PCIe* 设备能够直接向 VM 内存执行直接内存存取（DMA）。

Intel VT-d 是硬件中的一种实现，可为虚拟环境中的 DMA 事务执行必要的地址转换。实质上，Guest 物理地址与主机物理地址之间的这种转换可以帮助确保 DMA 事务对正确的物理页面应用保护机制，预防与其他 VM相关的内存页面受到影响。

大型 Intel VT-d 页面

大型 Intel VT-d 页面（Large Intel VT-d Pages）特性支持在内部 VT-d 页面表中使用 2 MB 和 1 GB 的大页面。在 I/O 内存管理单元（IOMMU）中利用这些大页面表条目，有助于减少 I/O 转换查询缓冲区（IOTLB）的数据丢失，从而有助于提高网络性能，尤其对于小数据包。（有关更多信息，请参阅：“大页面”节。）

中断再映射

中断再映射支持对分配给 VM 的 CPU 的中断进行路由和隔离。再映射硬件会通过跟踪中断发起者 ID 来执行隔离，也可以缓存常用的再映射结构来提高性能。

针对连接的 Intel® 虚拟化技术

Intel VT-c 改善了网络吞吐量，降低了 CPU 利用率并减少了系统延迟。此技术是 Intel 在以太网服务器适配器方面的一个特性，比如 Intel® 82599 10 Gigabit Ethernet 控制器。

虚拟机设备队列

虚拟机设备队列（VMDQ）是 Intel 以太网服务器适配器的一个特性，它基于 VLAN 和目标 MAC 地址来将以太网帧过滤到特定于 VM 的队列中，从而提高网络性能。然后，这个特定于 VM 的队列可与特定的核心关联，提高缓存命中率和总体性能。

PCI-SIG SR-IOV

PCI-SIG SR-IOV 允许对一个 Intel® 以太网服务器适配器端口进行分区，这个适配器也被称为物理功能（PF）。这个 PF 是一种完整的 PCIe 功能，将 SR-IOV 扩展功能（用于配置和管理 SR-IOV 功能）包含在多个 VF 中。这些 VF 是轻量级的 PCIe 功能，包含移动数据所必要的资源，但最小化了配置资源集。可以将它们分配给 VM，每个 VF 拥有自己的带宽配额。它们提供了进入 VM 的高性能、低延迟的数据路径。

下一页的图 4 显示了一个基于 Intel 架构的平台上的某种 SR-IOV 实现的总体视图。图中描绘了如何使用 SR-IOV 绕过虚拟机管理程序，提供进入 VM 的高性能、低延迟路径。

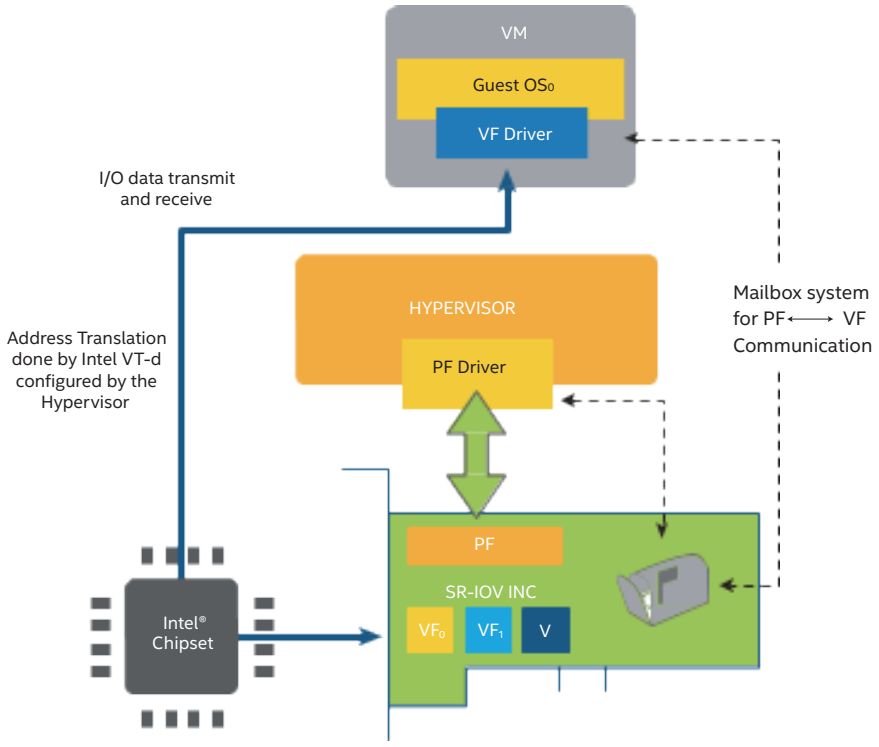

图 4. Single Root I/O 虚拟化的总体架构。

图 5 中描绘了一种 NUMA 拓扑结构。绿色箭头表示最高性能的内存访问路径，在该路径中，一个 CPU 核心上执行的进程访问本地附加的内存。橙色箭头表示从处理器核心到 “远程” 内存的一条次优的内存访问路径。

使用基于 NUMA 的设计，有助于实现极高的内存性能特征。但是，要以最高效的方式利用平台中的资源，必须考虑内存分配策略以及进程/线程与处理器的相联。内存分配策略通常可使用操作系统 API 进行控制。可告诉操作系统向一个线程分配位于执行该线程的处理器核心本地或一个指定核心本地的内存。使用的软件架构利用内存管理线程来为系统中其他工作线程分配内存时，后一种方法特别有用。下一节将介绍核心关联考虑因素。

非统一内存架构

在统一内存架构（UMA）设计中，系统中的每个核心都拥有相同的内存性能特征，比如延迟和吞吐量，这通常是由于共享与内存的相互连接所导致的。

许多现代多处理器系统中（比如 Intel® 至强® 处理器）通常使用 NUMA 设计。借助此模型，处理器核心能够访问本地附加的内存，表现出更高性能的内存访问特征。处理器核心也能够通过共享相互连接来远程访问挂载在另一个处理器核心的本地内存。

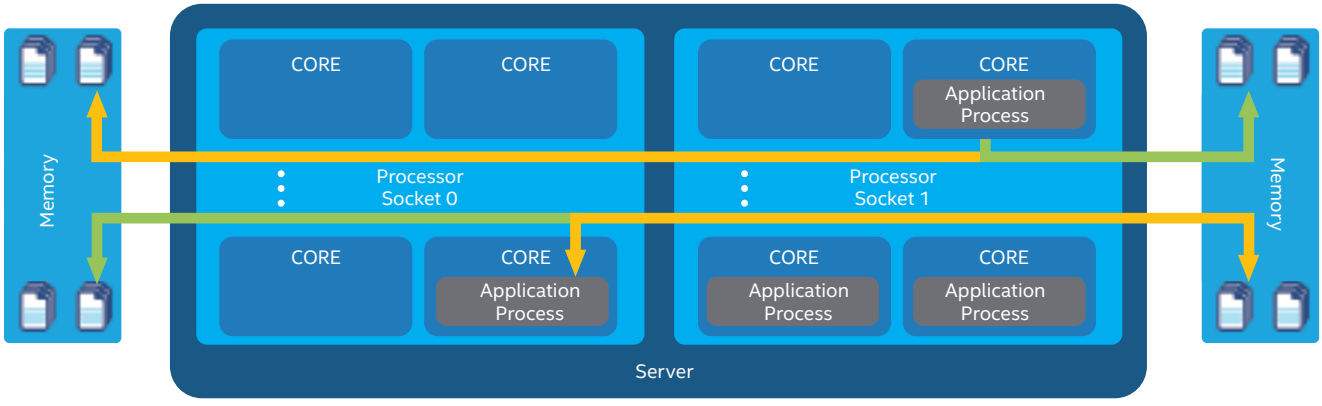

图 5. Non-Uniform Memory Architecture。

CPU 锁定

操作系统有一个调度程序，负责将处理器核心上的时间分段分配给系统中运行的线程。在多核处理器中，调度程序一般拥有在处理器核心之间转移线程的灵活性和能力。调度程序通过这种方式帮助系统中的所有核心对工作负载进行负载平衡。在基于 NUMA 的系统的上下文中，这种能力可能意味着：在一个处理器核心上执行且能够访问本地内存的线程，在随后的时间片段可以在另一个核心上执行，访问相同的内存位置，但相对该处理器核心而言，内存现在位于一个远程位置。

CPU 锁定技术支持为进程/线程配置与一个或多个核心的关联，通过配置 CPU 关联，调度程序现在只能将线程调度到一个指定的核心上执行。在 NUMA 配置中，如果为一个线程请求了特定的 NUMA 内存，此 CPU 关联设置会帮助确保内存保持在该线程的本地。图 6 中的逻辑视图显示了如何可锁定到特定 CPU 核心和内存的线程分配到这些核心的本地。

大页面

MMU TLB 中的大页面支持功能可帮助 TLB 缓存更大范围的内存转换。内存地址转换请求涉及的部分步骤如图 7 所示。在此示例中，使用小页面表条目（4 KB）导致 TLB 中涵盖的内存地址空间更少。这可能导致更多次的 TLB 丢失。一次 TLB 丢失会导致一次内存访问，以读取页面表来获取请求的地址转换条目。大量的 TLB 丢失会增加内存访问次数。这可能导致需要内存地址转换的应用程序的性能欠佳。

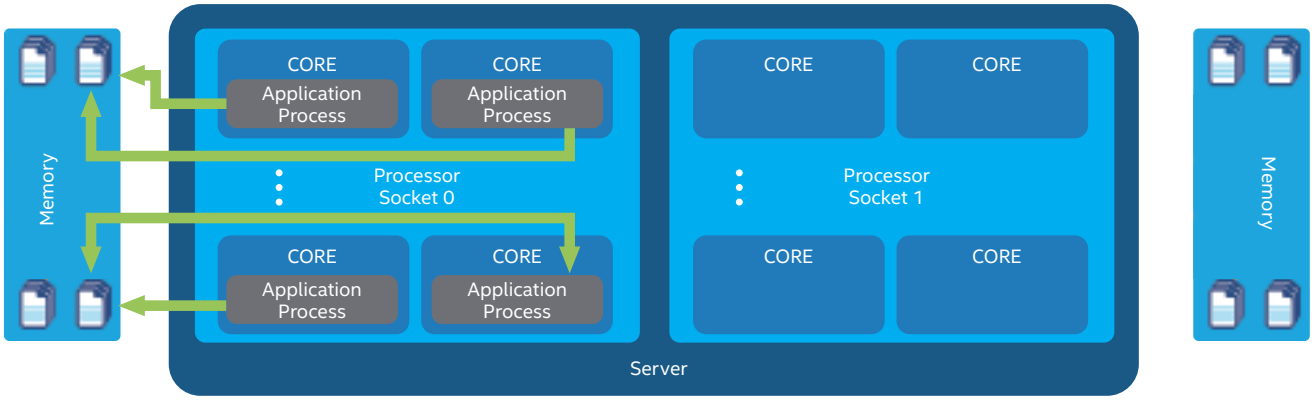

图 6. CPU 与非统一内存架构的最优位置的关联。

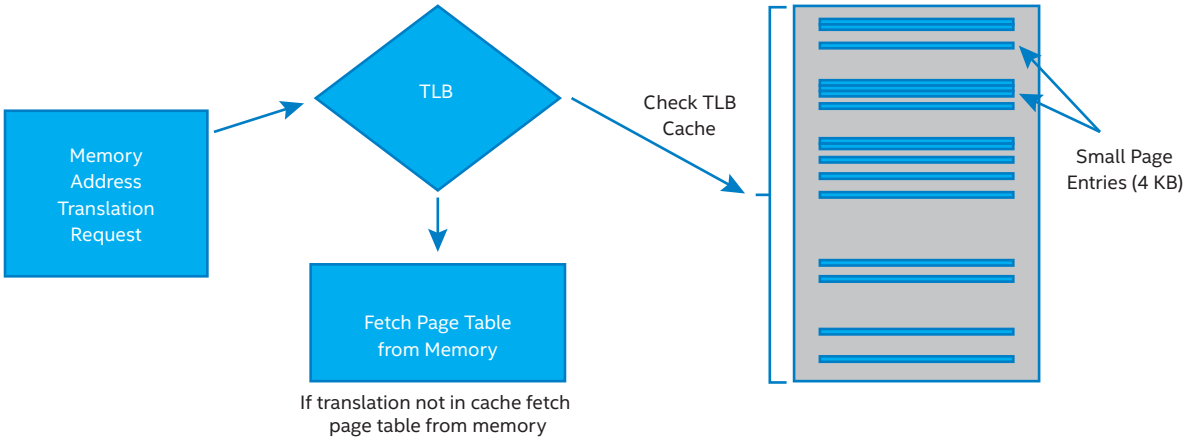

图 7. 包含小页面表条目的内存地址转换的概念视图。

图 8 中使用了大页面表条目。在此示例中，1 GB 的页面表条目大小会导致 TLB 中的内存空间大得多，从而有助于最大程度地减小 TLB 缓存丢失的负担。

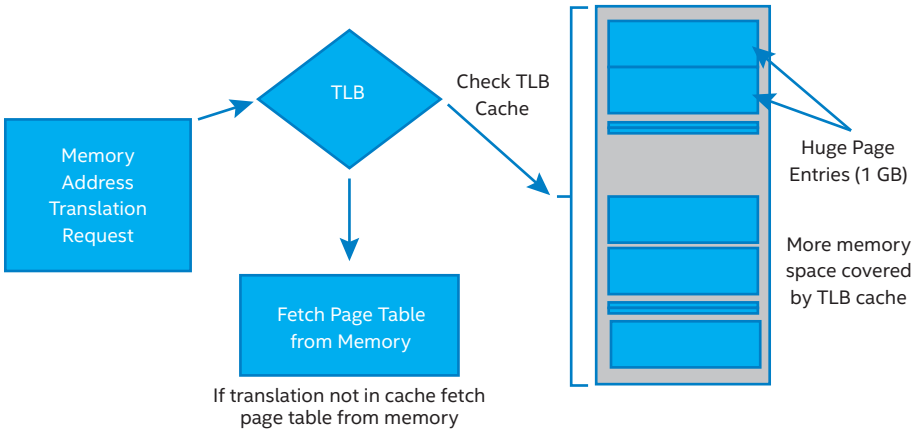

图 8. 包含大页面表条目的内存地址转换的概念视图。

结束语

ETSI-NFV ISG 正在大力开发规范来引导网络变革，对行业标准机构和开源软件堆栈施加影响。

基于 Intel 架构的平台提供了大量功能，这些功能可以在部署具有高性能需求的虚拟化工作负载时带来效益。Intel® VT-x、Intel® VT-d 和 Intel® VT-c 功能的结合使用，为虚拟化应用程序提供了优秀的体验。将这些功能与平台中广泛的调优工具相结合，可以帮助实现难以置信的工作负载性能。

OpenStack 是一个创建和管理私有和公有云基础架构的著名开源软件套件。为了适合部署高性能 NFV 工作负载，OpenStack 需要一些额外的特性来启用增强平台感知功能，释放 NFV 应用程序的性能潜力。

使用 Intel DPPD 作为参考 NFV 应用程序，表明通过利用 SR-IOV 连接、NUMA 感知 Guest 操作系统配置、大页面表配置和 CPU 锁定，可以帮助实现 10 Gbps 的线速性能。

其中一些特性已添加到 OpenStack Juno 版本中。Intel 和 OpenStack 社区的其他成员正在努力向后续 OpenStack 版本积极贡献其他 NFV 支持功能。

参考资料

1. 欧洲电信标准协会网络功能虚拟化小组：
   http://www.etsi.org/technologies-clusters/technologies/nfv
2. 《网络功能虚拟化 — 简介》白皮书：http://portal.etsi.org/NFV/NFV_White_Paper.pdf
3. 《网络功能虚拟化 — 更新》白皮书，网络运营商对行业进步的看法：
   http://portal.etsi.org/NFV/NFV_White_Paper2.pdf
4. 《网络功能虚拟化》— 第三本白皮书，网络运营商对行业进步的看法：
   http://portal.etsi.org/Portals/0/TBpages/NFV/Docs/NFV_White_Paper3.pdf
5. ETSI NFV 规范：http://www.etsi.org/technologies-clusters/technologies/nfv
6. 网络功能虚拟化（NFV）；NFV 性能和可移植性最佳实践
   http://www.etsi.org/deliver/etsi_gs/NFV-PER/001_099/001/01.01.01_60/gs_NFV-PER001v010101p.pdf
7. OpenStack，www.OpenStack.org
8. Intel® Data Plane Performance Demonstrators：https://01.org/intel-data-plane-performance-demonstrators
9. Data Plane Development Kit：www.dpdk.org
10. BSD 3-Clause 许可：http://opensource.org/licenses/BSD-3-Clause
11. 网络功能虚拟化：《使用 Linux* 和 Intel® 架构实现虚拟化 BRAS》白皮书：
    http://networkbuilders.intel.com/docs/Network_Builders_RA_vBRAS_Final.pdf
12. 使用 Intel® 高级加密标准新指令和 PCLMULQDQ 显著提高 Linux 上的 IPSec 性能：http://www.intel.com/content/www/us/en/intelligent-systems/wireless-infrastructure/aes-ipsec- performance-linux-paper.html
13. 虚拟驱动程序 Guest NUMA 节点放置和拓扑结构：
    https://blueprints.launchpad.net/nova/+spec/virt-driver-numa-placement
14. 针对 OpenStack Juno 的大页面和 I/O NUMA 调度补丁：
    https://01.org/sites/default/files/page/OpenStack_ovdk.l.0.2-907.zip
15. 具有 Intel® 架构的 OpenStack* 网络的入门指南：
    https://01.org/sites/default/files/page/accelerating_openstack_networking_with_intel_architecture_rev008.pdf
16. 具有 Linux* 和 Intel® 架构® 的宽带远程接入服务器中的服务质量：
    https://networkbuilders.intel.com/docs/Network_Builders_RA_NFV_QoS_Aug2014.pdf

作者

Adrian Hoban、Przemyslaw Czesnowicz、Sean Mooney、James Chapman、Igor Shaula、Ray Kinsella 和 Christian Buerger 是 Intel 公司网络平台小组的软件架构师和工程师。他们专门研究网络功能虚拟化和 OpenStack 等开源组件中的软件定义网络扩展。

致谢

文本的完成离不开开发 DPDK 和 DPPD 的团队的不懈努力和创新。

缩略语

API	应用编程接口
APIC	高级可编程中断控制器
ATS	地址转换服务
BNG	宽带网络网关
BRAS	宽带远程接入服务器
CLI	命令行接口
CPE	用户端设备
CPU	中央处理单元
DHCP	动态主机配置协议
DMA	直接内存存取
DSLAM	数位用户线接入复接器
EPT	扩展页面表
ETSI	欧洲电信标准协会
FW	防火墙
IOMMU	输入/输出内存管理单元
IOTLB	输入/输出转换外援缓冲器
IP	Internet 协议
IPAM	Internet 协议地址管理
ISG	行业规范小组
KVM	内核虚拟机
LAN	局域网
LB	负载平衡器
MMU	内存管理单元
MSAN	多服务接入节点
NFV	网络功能虚拟化
NIC	网卡
NUMA	非统一内存架构
OPNFV	开放网络功能虚拟化平台
OS	操作系统
PCI-SIG	外围组件互联特殊兴趣小组
pCPU	物理中央处理单元
PLE	暂停退出循环
PPP	点对点协议
PPPoA	基于 ATM 的点对点协议
PPPoE	基于以太网的点对点协议
QEMU	硬件虚拟化软件
QoS	服务质量
SR-IOV	Single Root I/O 虚拟化
RT	路由线程
SDN	软件定义网络
TLB	转换外援缓冲器
TPR	任务优先级寄存器
UMA	统一内存架构
vCPU	虚拟中央处理单元
VF	虚拟功能
VIF	虚拟接口功能
VIM	虚拟化基础架构管理器
VLAN	虚拟局域网
VM	虚拟机
VMDQ	虚拟机设备队列
VMM	虚拟机监视器
VMX	虚拟机扩展
VNF	虚拟化网络功能
VPID	虚拟处理器标识符
VPN	虚拟专用网络
VT	虚拟化技术
VXLAN	虚拟可扩展 LAN
WT	工作线程

关键词

业务流程，OpenStack，网络功能虚拟化，软件定义网络

在美国印刷　0315/JL/MESH/PDF　请回收再利用　332169-001US