岁寒

(https://lvwenhan.com/) **任何事情，从现在开始做，都不晚！**

性能之殇（五）-- DPDK、SDN 与大页内存

2018-11-19 / 阅读数：17536 / 分类： 操作系统 (https://lvwenhan.com/sort/36)

上文我们说到，当今的 x86 通用微处理器已经拥有了十分强大的性能，得益于其庞大的销量，让它的价格和专用 CPU 比也有着巨大的优势，于是，软件定义一切诞生了！

软路由

说到软路由，很多人都露出了会心的微笑，因为其拥有低廉的价格、超多的功能、够用的性能和科学上网能力。现在网上能买到的软路由，其本质就是一个 x86 PC 加上多个网口，大多是基于 Linux 或 BSD 内核，使用 Intel 低端被动散热 CPU 打造出的千兆路由器，几百块就能实现千兆的性能，最重要的是拥有 QOS、多路拨号、负载均衡、防火墙、VPN 组网、科学上网等强大功能，传统路由器抛开科学上网不谈，其他功能也不是几百块就搞得定的。

软路由的弱点

软路由便宜，功能强大，但是也有弱点。它最大的弱点其实是性能：传统 *UNIX 网络栈的性能实在是不高。

软路由的 NAT 延迟比硬路由明显更大，而且几百块的软路由 NAT 性能也不够，跑到千兆都难，而几百块的硬路由跑到千兆很容易。那怎么办呢？改操作系统啊。

SDN

软件定义网络，其本质就是使用计算机科学中最常用的“虚拟机”构想，将传统由硬件实现的 交换、网关、路由、NAT 等 网络流量控制流程交由软件来统一管理：可以实现硬件不动，网络结构瞬间变化，避免了传统的停机维护调试的烦恼，也为大规模公有云计算铺平了道路。

虚拟机

虚拟机的思想自底向上完整地贯穿了计算机的每一个部分，硬件层有三个场效应管虚拟出的 SRAM、多个内存芯片虚拟出的一个“线性数组内存”，软件层有 jvm 虚拟机，PHP 虚拟机（解释器）。自然而然的，当网络成为了更大规模计算的瓶颈的时候，人们就会想，为什么网络不能虚拟呢?

OpenFlow

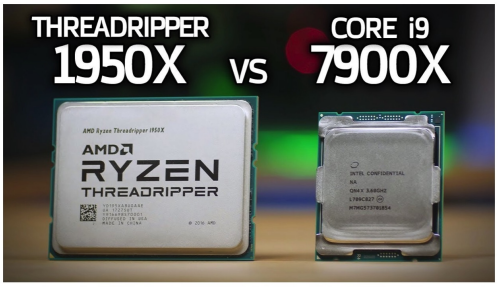
最开始，SDN 还是基于硬件来实施的。Facebook 和 Google 使用的都是 OpenFlow 协议，作用在数据链路层（使用 MAC 地址通信的那一层，也就是普通交换机工作的那一层），它可以统一管理所有网关、交换等设备，让网络架构实时地做出改变，这对这种规模的公司所拥有的巨大的数据中心非常重要。

DPDK

DPDK 是 SDN 更前沿的方向：使用 x86 通用 CPU 实现 10Gbps 甚至 40Gbps 的超高速网关（路由器）。

DPDK 是什么

Intel DPDK 全称为 Intel Data Plane Development Kit，直译为“英特尔数据平面开发工具集”，它可以摆脱 *UNIX 网络数据包处理机制的局限，实现超高速的网络包处理。



DPDK
DATA PLANE DEVELOPMENT KIT

OvS
Open vSwitch

F-Stack

DPDK 的价值

当下，一台 40G 核心网管路由器动辄数十万，而 40G 网卡也不会超过一万块，而一颗性能足够的 Intel CPU 也只需要几万块，软路由的性价比优势是巨大的。

实际上，阿里云和腾讯云也已经基于 DPDK 研发出了自用的 SDN，已经创造了很大的经济价值。

> 【DPDK峰会回顾】支撑双十一的高性能负载均衡是如何炼成的 (https://yq.aliyun.com/articles/615587) 阿里云携领先SDN能力，亮相全球网络技术盛会ONS (https://yq.aliyun.com/articles/575989) 腾讯云超高网络性能云主机揭秘 (https://cloud.tencent.com/developer/article/1006690) F-Stack 全用户态 (Kernel Bypass) 服务开发套件 (https://cloud.tencent.com/developer/article/1005179)

怎么做到的?

DPDK 使用自研的数据链路层（MAC地址）和网络层（ip地址）处理功能（协议栈），抛弃操作系统（Linux，BSD 等）提供的网络处理功能（协议栈），直接接管物理网卡，在用户态处理数据包，并且配合大页内存和 NUMA 等技术，大幅提升了网络性能。有论文做过实测，10G 网卡使用 Linux 网络协议栈只能跑到 2G 多，而 DPDK 分分钟跑满。

用户态网络栈

上篇文章我们已经说到，Unix 进程在网络数据包过来的时候，要进行一次上下文切换，需要分别读写一次内存，当系统网络栈处理完数据把数据交给用户态的进程如 Nginx 去处理还会出现一次上下文切换，还要分别读写一次内存。天寿啦，一共 1200 个 CPU 周期呀，太浪费了。

而用户态协议栈的意思就是把这块网卡完全交给一个位于用户态的进程去处理，CPU 看待这个网卡就像一个假肢一样，这个网卡数据包过来的时候也不会引发系统中断了，不会有上下文切换，一切都如丝般顺滑。当然，实现起来难度不小，因为 Linux 还是分时系统，一不小心就把 CPU 时间占完了，所以需要小心地处理阻塞和缓存问题。

NUMA

NUMA 来源于 AMD Opteron 微架构，其特点是将 CPU 直接和某几根内存使用总线电路连接在一起，这样 CPU 在读取自己拥有的内存的时候就会很快，代价就是读取别 U 的内存的时候就会比较慢。这个技术伴随着服务器 CPU 核心数越来越多，内存总量越来越大的趋势下诞生的，因为传统的模型中不仅带宽不足，而且极易被抢占，效率下降的厉害。

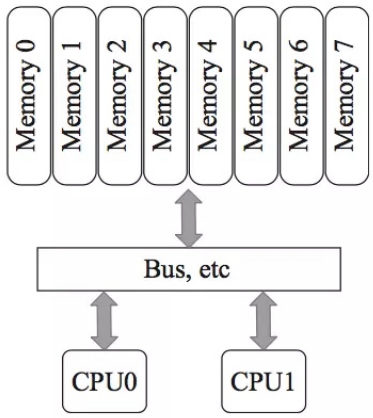


图 2-14　SMP 系统示意图

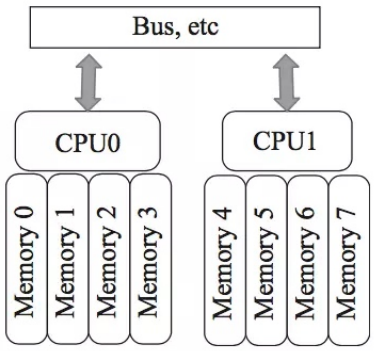


图 2-15　NUMA 系统示意图

> NUMA 利用的就是电子计算机（图灵机 + 冯·诺依曼架构）天生就带的涡轮：局部性。 涡轮：汽车发动机加上涡轮，可以让动力大增油耗降低

细说大页内存

内存分页

为了实现虚拟内存管理机制，前人们发明了内存分页机制。这个技术诞生的时候，内存分页的默认大小是 4KB，而到了今天，绝大多数操作系统还是用的这个数字，但是内存的容量已经增长了不知道多少倍了。

TLB miss

TLB（Translation Lookaside Buffers）转换检测缓冲区，是内存控制器中为增虚拟地址到物理地址的翻译速度而设立的一组电子元件，最近十几年已经随着内存控制器被集成到了 CPU 内部，每颗 CPU 的 TLB 都有固定的长度。

如果缓存未命中（TLB miss），则要付出 20-30 个 CPU 周期的带价。假设应用程序需要 2MB 的内存，如果操作系统以 4KB 作为分页的单位，则需要 512 个页面，进而在 TLB 中需要 512 个表项，同时也需要 512 个页表项，操作系统需要经历至少 512 次 TLB Miss 和 512 次缺页中断才能将 2MB 应用程序空间全部映射到物理内存；然而，当操作系统采用 2MB 作为分页的基本单位时，只需要一次 TLB Miss 和一次缺页中断，就可以为 2MB 的应用程序空间建立虚实映射，并在运行过程中无需再经历 TLB Miss 和缺页中断。

大页内存

大页内存 HugePage 是一种非常有效的减少 TLB miss 的方式，让我们来进行一个简单的计算。

2013 年发布的 Intel Haswell i7-4770 是当年的民用旗舰 CPU，其在使用 64 位 Windows 系统时，可以提供 1024 长度的 TLB (https://www.7-cpu.com/cpu/Haswell.html)，如果内存页的大小是 4KB，那么总缓存内存容量为 4MB，如果内存页的大小是 2MB，那么总缓存内存容量为 2GB。显然后者的 TLB miss 概率会低得多。

DPDK 支持 1G 的内存分页配置，这种模式下，一次性缓存的内存容量高达 1TB，绝对够用了。

不过大页内存的效果没有理论上那么惊人，DPDK 实测有 10%~15% 的性能提升，原因依旧是那个天生就带的涡轮：局部性。

WRITTEN BY

JohnLui (https://github.com/johnlui)

程序员，Swift Contributor

相关日志：

性能之殇（六）-- 现代计算机最亲密的伙伴：局部性与乐观 (https://lvwenhan.com/操作系统/497.html)

性能之殇（四）-- Unix 进程模型的局限 (https://lvwenhan.com/操作系统/495.html)

性能之殇（七）-- 分布式计算、超级计算机与神经网络共同的瓶颈 (https://lvwenhan.com/操作系统/498.html)

> 软件工程师需要了解的网络知识：从铜线到HTTP（五）—— HTTP 和 HTTPS (https://lvwenhan.com/操作系统/489.html)
>
> 软件工程师需要了解的网络知识：从铜线到HTTP（四）—— TCP 和路由器 (https://lvwenhan.com/操作系统/488.html)

标签: 性能之殇 (https://lvwenhan.com/tag/%E6%80%A7%E8%83%BD%E4%B9%8B%E6%AE%87) 性能 (https://lvwenhan.com/tag/%E6%80%A7%E8%83%BD) DPDK (https://lvwenhan.com/tag/DPDK) SDN (https://lvwenhan.com/tag/SDN) 大页内存 (https://lvwenhan.com/tag/%E5%A4%A7%E9%A1%B5%E5%86%85%E5%AD%98) HugePage (https://lvwenhan.com/tag/HugePage)

← 性能之殇（六）-- 现代计算机最亲密的伙伴：局部性与乐观 (https://lvwenhan.com/操作系统/497.html)
性能之殇（四）-- Unix 进程模型的局限 → (https://lvwenhan.com/操作系统/495.html)

评论：

elmelundsvej

2021-02-18 18:21

DPDK是Intel公司主导开源的技术产品。我想知道Linux基金会对此是什么态度？如果Linux基金会的大神经们决定的在kernel中颠覆性的引入bypass机制，那么你觉得DPDK还会有前途吗？
补充说明，无法应用操作系统安全机制是在实际工程中工程师采用DPDK编写应用程序时面临的重要问题。其次，工程师们必须采用非常艰难的方式协调采用DPDK开发的应用程序与CPU为其他设备处理系统软中断之间的优先级问题，而且还要打破应用与CPU的亲缘性，这些无疑让小型的软件公司无法凭借现有的工程技术人员快速实现功能。其使用门槛不亚于采用传统的netfilter，甚至是基于eBPF的XDP技术。

回复

Even

2020-12-10 09:25

现有的软路由有支持DPDK的么?

回复

Matrix (http://www.chuancn.cn)

2020-01-12 15:50

😍😍

作者大牛，有件事情很好奇啊，现在的软路由有基于linux（DPDK）的也有openwrt的；此处为什么linux的性能没有Openwrt的强悍呢？还有就是如果要做软路由的相关开发要怎么入手啊？需要用到那些技术栈啊？

回复

JohnLui (https://lvwenhan.com/)

2020-05-22 09:55

@Matrix：OpenWRT 就是 Linux，至于 DPDK 只是一个应用软件，你说的这几个东西没法这么对比🤔🤔

回复

yan

2018-12-12 23:08

hugepage的问题是不是一旦出现swap, 性能就会很糟糕，因为比如1GB page swap下去是很耗时的。

但是只要内存容量足够，关闭swap, hugepage还是很有优势的。

回复

发表评论：

昵称

邮件地址 (选填)

个人主页 (选填)

发表评论