WIND RIVER

以多核技术优化网络性能

提要

在网络通信设备的评测中，性能和成本一直是最关键的要素。然而，性能包括了多个方面，包括吞吐能力、时延和CPU占用率等。即便对于容量小于1GB的系统，可预测的响应时间和可用于运行应用的CPU周期都至关重要。多核的出现为性能和成本的优化带来了机遇。在多个内核之间高效率地分布网络通信功能，系统就可以实现比前一代产品更高的吞吐能力、更低的CPU占用率、更小的设备尺寸和更低的成本。本文描述了操作系统、网络协议栈和多核的有效集成将会给网络通信行业带来怎样的变革。

更高性能带来的挑战

根据摩尔定理，处理器的能力每两年就要翻一倍[1]。由此对所有使用微处理器的设备产生了广泛的影响，其中当然包括网络通信设备。性能更强大的终端节点能够更快地处理数据，这也是网络通信带宽需求不断增长的动力。在过去的15年内，局域网（LAN）的传输速率已经增长了1000多倍。

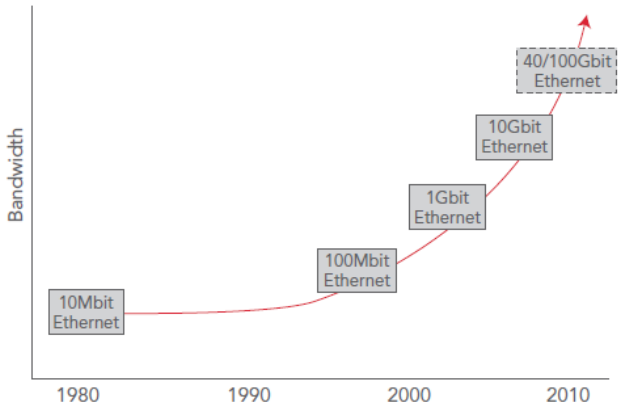

图1：局域网数据速率的演进

广域网（WAN）的数据速率虽然没有局域网速率那么快，但是也呈现出指数级的增长。

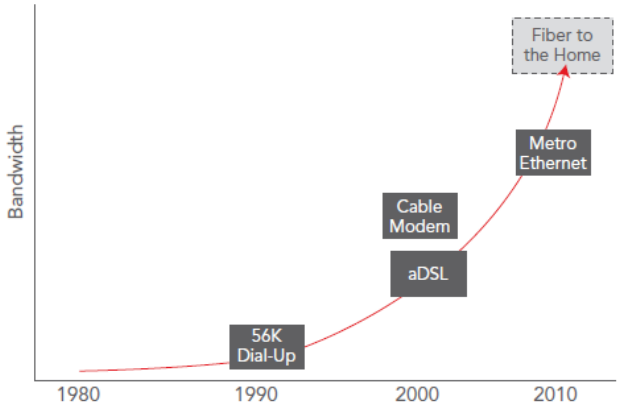

图2：广域网数据速率演进

发展最快的网络通信技术是无线通信。由于不再受到线缆连接的牵绊，无线通信为用户互连提供了极大的便利性。

有人争辩说，多媒体内容的需求已经存在多时，早就在等待相应的网络通信技术来支持了。视频和音频文件不仅体积比纯文本数据文件大得多，而且对时间延迟更为敏感。

语音、视频和数据的融合需要更精密的通信设备以满足低时延的需求。如今的家庭网关需要实现互联网接入、VoIP语音通信和视频流等多种业务混合一体的处理能力。

同样，像苹果iPhone这类手机设备中也融合了语音、数据、音乐、互联网访问和视频等多种功能，而且把这些功能放在了更小的设备中。

所有这些发展趋势都需要高带宽、低时延的网络通信技术来实现，不论是对于终端用户设备，而且包括各种接入、汇聚和核心部件。设备制造商面临的挑战是在开发周期缩短、产品利润空间压缩的压力下，为市场提供更高性能的平台。解决所有这些需求和挑战需要全新的解决方案。多核就是解决方案。

多核性能

在过去几十年里，处理器能力每两年就翻一倍。然而，近几年处理器速率上升曲线开始变得平缓，这是由于受到发热和功耗等因素限制，无法再通过增加晶体管数量来提升处理器性能。

但是，多核处理提供了新思路。通过并发地使用多个内核，处理性能可以进一步提升以满足高性能的需求。

全球领先的各大处理器芯片厂商都开始推出多核芯片，在单个芯片内集成了多个处理内核。通过非常快速的任务间数据交换，虚拟内核或线程可以进一步细分内核资源。

多核处理器芯片的性能依据时钟速率和内核数量而不同。目前已经有16—32个内核的处理器芯片。这些芯片中大多数都集成了网络处理功能，减小了由传统网络协议软件所造成的时延。

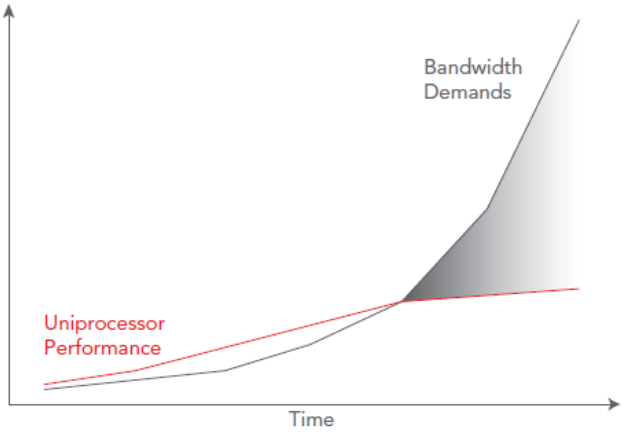

图3：单个处理器性能差距

多核处理方法

多核软件可以由多种模式实现[2]。在采用对称对处理（SMP）模式实现的系统中，在运行操作系统和任务时，多个内核基本上是可以互换的。有一种SMP采用了联姻（affinity）或CPU预留技术来指定任务与某个内核的绑定，由此使其变成较为高效的专用处理器。

非对称对处理（AMP）通常是指运行着多个操作系统的架构。Supervised AMP采用了虚拟化技术对各种处理单元进行抽象，例如内存、内核或设备等。

为了发挥新型芯片的优势，必需设计出新的软件。一种常见的误解是，为单核处理器环境编写的软件在多核处理器环境下自然能够运行得更快更好。让我们以机器人为例，在装配生产线上经常使用机器人手臂来搬动箱子。当采用单核处理器运行时，每分钟能够搬运12个箱子。如果同样的系统和软件以SMP模式采用多核处理器运行的话，机器人手臂并不会运行得更快，每分钟仍然只能搬运12个箱子。但是，如果将软件面向多核处理器技术进行重新编写，系统就能够使用更多的处理能力去执行其他的任务。例如，如果在上述机器人控制多核系统中，将第二个处理器用于控制另一个机器人手臂，并且与第一个手臂交叉配合，可以实现每分钟搬运24个箱子，使生产效率加倍。此外，第二个处理器还可以用于控制传送带、遥感探测反馈或分担第一个处理器上的任务负载，将生产效率提高到每分钟搬运15个箱子。从这个例子中显而易见，仅仅在硬件方面向多核处理技术的转变不会自动地提升性能，必需通过内核间的相互配合和交互，才能将充分发挥多核处理器的性能。

以多核处理网络协议

网络通信协议栈中最常见的协议就是IP和TCP。这些协议在互联网技术中广泛使用，实际上还用于所有使用网络连接功能的行业和产品。同样，网络通信设备也必须支持一套通用的网络协议，由此支持保持所有行业实现互连的基础架构。传统方式下，这些协议被当做统一处理数据包的单一体栈。

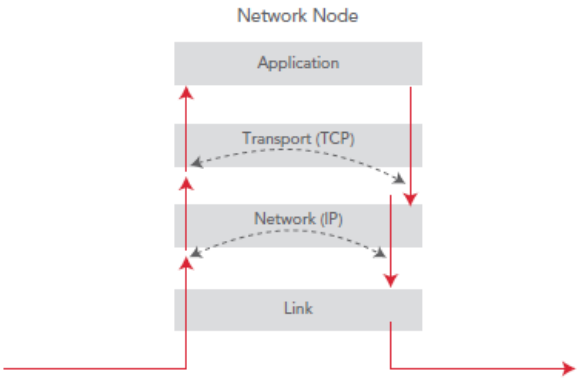

图4：传统的单一体数据包处理

数据包进入单一体栈后的处理步骤可以想象为一个过程状态机。数据包首先被物理链路层（通常是以太网）接口接收，然后排队进入网络层（IP层）。IP层负责确定数据包目的地是是本机还是需要被继续转发。此外，作为IPsec或IKE等安全协议的一部分，IP层还需要完成对应的数据包加密操作。如果数据包的目的地就是本机，那么它将跳过本层而被转到上一层协议（通常是TCP或UDP）进行处理。更多的安全功能需要通过安全套接字层（SSL）进行处理。如果数据封包还是需要送往本机，那么它将被转到更高的应用层协议，包括FTP、

SMTP、Telnet和HTTP等。在单核系统甚至纯SMP系统中，所有这些网络处理过程都需要竞争处理器工作周期。

单一的分布式网络协议栈架构并不能完全反应所有的应用场景。例如，同样是不断重复执行的步骤，当用于包转发时和用于数据包就是不同的。交换机和网关在第三层存在区别，交换机负责在接口和web服务器间转发数据包，而网关负责接收（或终止）数据请求并返回HTML数据的页面。对网络协议栈的优化需要针对不同的应用场景采用不同的架构。

另外，系统设计的选择和确定必须根据协议栈的哪一层进行分发。如果大多数的连接需要在第四层（TCP）进行管理，最好采用能够跨处理器内核分配多TCP实例的架构。然而，多实例必然带来协议复杂度的增加，需要内核间更多的交互，从而导致时延的增加和内存带宽的受限。这个问题在带宽为1G或2G的情况下可能还不明显，但随着系统吞吐带宽的进一步扩展，它将成为严重的制约因素之一。成本/收益分析必须综合考虑整个系统的目标

在交换机、路由器和网关等高性能网络通信设备中，绝大多数网络数据包的转发都发生在IP层。作为一种高频次重复的任务，数据包转发功能尤其适于采用多核技术来实现。首先，需要将IP协议中实现转发功能的代码与地址建立逻辑代码进行分离。当接收数据包中的地址是第一次出现时，将地址记录在地址数据表格中。由于此项任务只需执行一次，数据表的维护可以采用多用途内核中的“慢速通道（slow path）”，它的功能就像是操作系统中的传统网络协议栈处理器。“快速通道（fast path）”处理器（或转发内核）仅仅需要检测接收数据包的目的地地址，并且查找其缓存数据表。如果地址被找到，快速通道处理器快速确定对应的输出端口，并将数据包及时送入转发队列中。

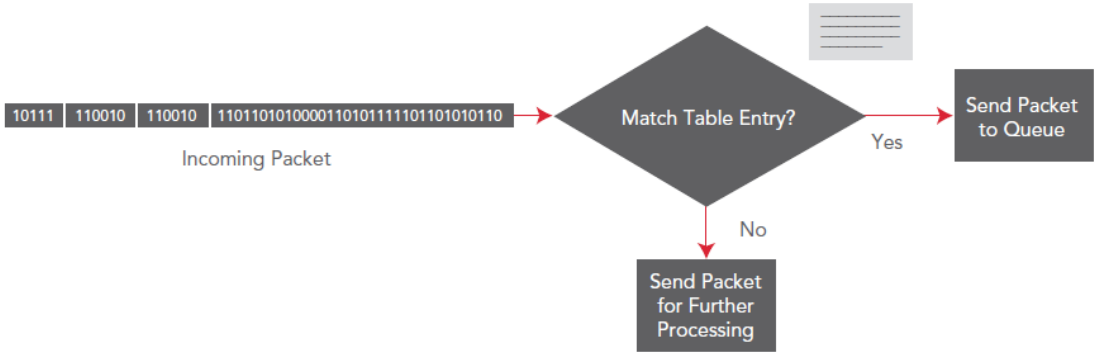

图5：数据包处理逻辑状态

通过这种方式实现的IP数据包分类算法适于各种高效的第三层网络功能，包括包转发、网络地址翻译（NAT）、访问控制列表、IPsec和其他加密数据功能等。某些处理器芯片具有针对数据加密的专用引擎，可以与核心CPU并发运行。

虽然用于实现这种快速通道式转发的代码相对更简单，但高达数G比特的接口产生的高数据流量速率还是会对处理器内核带来巨大的压力。采用更多的转发内核可以部分地缓解高吞吐量带来的问题，但必需结合采用其他的系统设计技术，才能避免系统性能不会因为这些瓶颈而达到上限。

多核设计考虑的因素

基于多内核设计的分布式网络协议栈能够极大地提升网络通信设备的系统性能，但必须考虑相关因素。当多个内核共享内存时，某一时刻只有一个内核能够更新信息，其他内核必须等待访问权。这种互锁安全机制一般通过软件使用semaphores、spinlock和mutual exclusion等技术实现。由于多核处理器必须相互等待才能完成共享内存中的数据结构修改，随着内核数量的增加，也会带来更多的冲突。这就是系统吞吐量性能曲线在某个时间点出现下拐或趋平时所呈现的情况。

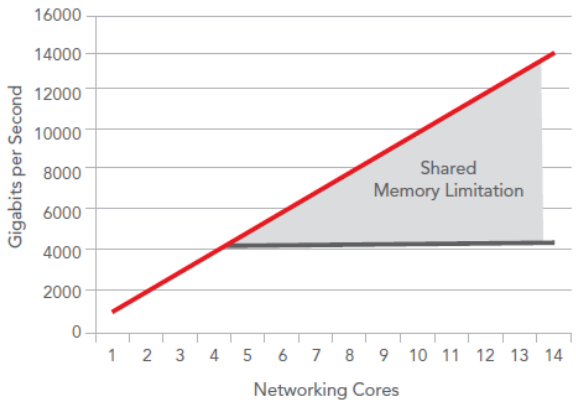

图6：共享内存带来的带宽限制

因此，对多核架构代码设计时，必须仔细考虑，才能将互锁带来的影响降到最小。如果可能，应该尽量应用芯片中无互锁代码的功能特性，以避免软件性能极限。

缓存（Caching）与哈希（Hashing）技术

几十年来，指令缓存技术（Instruction caching）不断被研究和改善，用于实现更高效的数据处理。高频次重复的任务很适合采用多核负载分流技术，但是如果指令缓存没有得到很好地优化，也很容易造成性能的反常下降。让我们设想一个每秒能够处理140万个数据包的IP转发内核。单单一个缓存的缺失就会导致处理循环需要多花费15%的时钟周期，从而造成处理能力的相应下降。调校代码以匹配指令缓存，这项工作是非常精细繁杂的，但是高效地使用缓存技术能够为网络协议处理的性能带来极大地提升。

同样，我们也需要高效地处理各种数据表结构，包括端口映射、地址翻译、流量信息和安全关联等，从而避免性能曲线的下拐。哈希（Hashing）技术是存取这些信息的高效手段，但是其前提是必须分配足够的空间以避免冲突发生。如果两个数据表入口都控制相同的哈希值，那么数据访问效率将会降低。因此，为了应付预期的数据流量速率，系统必须预先就进行正确地设计和配置，才能避免效率的下降。

多核性能获益

在网络通信协议栈中应用多核可以从两方面获益。一方面，也是最明显的方面，是提升吞吐能力。吞吐能力的度量单位通常是“包/秒”或者“兆比特/秒”。通常，我们常误认为千兆以太网（Gigabit Ethernet）能够完全支持每秒1千兆比特的数据吞吐量。实际上，一部分带宽会被数据包间隙或头部开销等所占据。

封包头：　　8字节
数据包间隙：　　12字节/20字节
　　20字节

在采用64字节以太网帧的1Gb/秒链路中，在线路上实际发送的数据为20+64=84字节。也就是说，20/84 = 23.8%的线路容量（即238Mbps）被用于头部开销，剩下76.2 %（即762Mbps）用于传送数据。当数据帧大小增加时，它们的发送次数可以降低，从而开销占据带宽的比例也随之下降。下表显示了各种帧大小情况下的最大帧速率和可用数据量。

Frame Size	Throughput	FPS
64	762Mbps	1,488,000
128	865Mbps	844,595
256	928Mbps	452,900
512	962Mbps	234,962
1024	981Mbps	119,732
1280	985Mbps	96,154
1518	987Mbps	81,274

表1：以Mbps和帧、每秒(FPS)衡量1Gb以太网理论最大有效载荷

常用的测量基准还有包转发（指一个接口接收到数据包并转发至另一接口）和包终止（指数据包被处理并终止）。根据被测试网络设备的类型，某些测试手段可能更适合于测量设备性能。

图7展示了通过基于风河VxWorks 6.7平台的500MHz Cavium OCTEON 3860测试的IPv4数据包转发速率。

Frame Size	Traditional IP Forwarding	Multicore Network Acceleration	Percent Increase
64	18	762	4133%
128	34	865	2444%
256	64	928	1350%
512	127	962	657%
1024	277	981	254%
1280	343	985	187%
1518	410	987	141%

Measured with VxWorks 6.7 EAR release on Cavium 3860, 500MHz, two cores/ one forwarder; VxWorks core is 99% to 100% idle during test

图7：采用多核加速的网络数据包处理

除了网络吞吐能力以外，系统的整体性能还依赖于可供其他系统任务使用的处理能力。如果所有的资源都被耗在包处理方面，那么即使最简单的系统管理任务也可能导致资源耗尽和性能下降。也就是说，就算这些设备在测试过程中能够获得期望的吞吐速率，但前文所描述的瓶颈在也会非常明显地表现出来。为了让网络负载分流真正发挥明显的效益，必须让负载分流算法对核心操作系统的依赖性降低到最小的程度。

结论

多核芯片提供了实现快速包处理的全新方案。通过在单个芯片上整合多个处理内核，网络设备可以被设计得体积更小、功耗更低、成本更低而性能更高。然而，要获得可扩展的线速性能，必需细致地考虑硬件和软件架构。硬件加速特性可以节省宝贵的处理器周期，应当尽量采用。有效地使用缓存队列和哈希数据表，是实现快速路径性能最大化的关键。对称多处理模式非常有效，但也必需仔细设计，确保网络通信任务以高度并行化的方式执行。采用专用内核来完成负载分流任务将会非常高效，达到满意的线速性能和可扩展性。

注：

1. 摩尔定律宣称，集成电路中可以置入的晶体管数量将会每两年翻一倍。这个定律已经被证明适用于性能、容量和成本等许多相关领域。

2. Device Software Optimization for Concurrent and Consecutive Systems, Wind River, http://windriver.com/whitepapers/.

3. Achieving Business Goals with Wind River’s Multicore Solution, Wind River, http://windriver.com/whitepapers/.

Wind River 就在您身边

北京代表处	北京市朝阳区望京中环南路9号望京大厦B座18层	邮编：100102	电话：010-84777100	传真：010-64398189
上海代表处	上海市西藏路585号新金桥广场3-H,I,J室	邮编：200003	电话：021-63585586/87/89/90	传真：021-63585591
深圳代表处	深圳市福田区车公庙天安数码时代大厦A座606室	邮编：518040	电话：0755-25333408/3418/4508/4518	传真：0755-25334318
西安代表处	西安市高新区科技二路68号西安软件园秦风阁H103	邮编：710075	电话：029-87607208	传真：029-87607209
成都代表处	成都市高新区天府软件园二期D7 14层	邮编：610041	电话：028-65318000	传真：028-65319983

关于风河更多内容请访问：http://www.windriver.com.cn　　Email：inquiries-ap-china@windriver.com

WIND RIVER