01.06

SmartNIC 的发展和应用

杨小康

1)华中科技大学计算机科学与技术学院数据中心课程

摘 要 在网速飞速提升、内存瓶颈突出、网络处理开销愈发显著的时代,普通网卡在网络协议处理、数据搬移、使用灵活性等方面逐渐暴露出缺陷。智能网卡,作为可编程的智能网络设备,在数据中心、科学计算领域均得到广泛关注,成为解决网络瓶颈的关键技术。在网络协议处理卸载、网络功能虚拟化、特定应用加速等应用场景中发挥着重要作用。我们可以通过智能网卡来将 OVS 操作从 CPU 卸载下来,并完成存储加速、数据加密、深度包检测和复杂路由等各种功能,将花费在处理这些工作负载上的大量的 CPU 周期返回给主机 CPU,同时解决了不同业务之间的冲突问题,大幅提升了各项业务的性能,也确保了服务器 CPU 能为应用提供最大的处理能力或者提供更多的虚拟机(VM)服务,创造更大的价值。

更快的网络速率、更加复杂的处理场景、更高昂的网络处理开销催生了智能网卡,智能网卡作为一种应用驱动的产物,在众多场景中得到应用。我看过一些关于智能网卡方面的论文了,其中有将 NF(网络功能)卸载到 SmartNIC 上的,有将 DFS(分布式文件系统)卸载到 SmartNIC 上的,也有将 SmartNIC 当做一个远端存储器的。总的来说,从四个方面来介绍 智能网卡的典型应用场景:(1)网络协议处理;(2)网络功能卸载;(3)数据中心应用;(4)科学计算应用。

关键词 智能网卡; 卸载; 数据中心; 智能网卡应用; 智能网卡发展

1 引言

在数据中心这次课程的论文讲解中, 因为某些 奇特的原因,我最终选择了ATC'21中的SKQ: Event Scheduling for Optimizing Tail Latency in a Traditional OS Kernel, 即 SKQ: 用于优化传统操 作系统内核尾部延迟的事件调度,是关于 OS 的论 文, 优化操作系统内核以此来降低尾延迟。但本来 我想选择的是另外两篇其中之一的,即 SOSP'21 上 的 Automated SmartNIC Offloading Insights for Network Functions 或者 LineFS: Efficient SmartNIC Offload of a Distributed File System with Pipeline Parallelism。这两篇论文是属于 不同的 session 的,之所以想选择这两篇论文,是 因为它们都关于了一个共同的的器件,或者是硬件 ——SmartNIC (智能网卡)。我的研究题目是基于 智能网卡的近数据处理编程模型, 因为论文相关, 所以我更喜欢相关的论文来积累知识, 还没到达触 类旁通的地步。当然我的知识积累还不是很足,所 以接下来的总结可能会有问题,希望后来的阅读者 来给予指正。接下来, 开始正式叙述。

随着网络技术、存储技术、芯片设计制造技术 的不平衡发展,目前,计算机网卡的设计面临新的 问题, 传统数据中心基于冯诺依曼架构, 所有的数 据都需要送到 CPU 进行处理, 而网络速度在 2020 年已经迈向 400 Gbps 以太网大关,并正向着更快 的 800 Gbps, 甚至 1.6 Tbps 发展, 而后摩尔时 代意味着 CPU 的频率已经趋于稳定,在这种不平 衡的现状下,使用传统的 CPU 来进行网络处理已 经显得不尽如人意,^[1]在速度上,现代 CPU 需要用 10~15 ns 来访问 L3 Cache, 而 400 Gbps 的网络 仅需 1.2 ns 便可传送 64 B 的消息; 在计算能力 上,CPU 适合于处理串行的复杂指令操作,对大量 并行的固定模式的计算并不适用; 再者, 在云环境 多租户的情况下,虚拟交换(open virtual switch, OVS)等网络功能虚拟化将占用更多的 CPU 资 源. 此外,通过 CPU 访问内存、进行数据搬移的 开销在很多应用中占据了极大的比例, 如在快速傅 里叶变换(fast Fourier transform, FFT)计算 中,数据搬移占据了 40%的开销,CPU 的增长速度 无法满足数据的爆发式增长, CPU 的处理速率已经 不能满足数据处理的要求。计算架构从以 CPU 为中 心的 Onload 模式, 向以数据为中心的 Offload 模式转变, 旁路 CPU 已经成为一种重要的解决方式。

因此,能够满足高速的网络处理需要、卸载 CPU 不适合的网络处理任务、提供一定编程灵活性 的智能网络终端设备——智能网卡,应运而生,也 推动了智能网卡的高速发展。并且在协议处理、网 络功能、数据中心云服务、人工神经网络加速、科 学计算等诸多场景中发挥了重要作用。

传统网卡(Network Interface Card,简称 NIC) 是将电脑接入局域网的设备,网卡插在计算机主板 的总线插槽中,负责将用户要传递的数据转换为网 络设备能够识别的格式,通过网络介质传输。^[2]网 卡的架构通常是,一个典型的 NIC 由一个处理网络 流量的处理器、一对处理数据移动的 DMA 引擎和一 个用于缓冲区和 Doorbell 的本地存储器(通常是 RAM)组成。它支持较低链路速度,网络堆栈算法 和协议在主机端 CPU 上运行。在将数据包传输到 网络之前,主机端网卡仅执行链路层和物理层中的 无状态算法和协议。传统网卡显著依赖 CPU 算力, 且需要使用主机存储器来维持数据包之间的状态。 也就是说,传统 NIC 用于执行无状态数据包处理, 而有状态数据包处理则在主机端 CPU 上实现。^[3]

在网速飞速提升、内存瓶颈突出、网络处理开销愈发显著的时代,普通网卡在网络协议处理、数据搬移、使用灵活性等方面逐渐暴露出缺陷。智能网卡,作为可编程的智能网络设备,在数据中心、科学计算领域均得到广泛关注,成为解决网络瓶颈的关键技术。四在网络协议处理卸载、网络功能虚拟化、特定应用加速等应用场景中发挥着重要作用。

例如,在高度虚拟化的环境中,主机 CPU 需要运行 OVS(Open Virtual Switch) 相关任务,同时 CPU 还要处理存储、数据包的在线加解密或离线加解密、数据包深度检查、防火墙、复杂路由等操作,这些操作不仅消耗大量的 CPU 资源,同时由于不同业务之间的 CPU 资源争夺,导致业务的性能不能发挥到最佳。 [5] 网卡作为连接各种业务的枢纽,在网卡上对于上述业务进行加速,成为了最理想的场所。智能网卡的出现,为解决这个问题提供了新的思路,我们可以通过智能网卡来将 OVS 操作从 CPU 卸载下来,并完成存储加速、数据加密、深度包检测和复杂路由等各种功能,将花费在处理这些工作负载上的大量的 CPU 周期返回给主机 CPU,同时解

决了不同业务之间的冲突问题,大幅提升了各项业务的性能,也确保了服务器 CPU 能为应用提供最大的处理能力或者提供更多的虚拟机 (VM) 服务,创造更大的价值。^[6]

下面依据一篇论文来讲解 SmartNIC 的原理和 优势,来说明它为何成为热门,并被广泛应用。

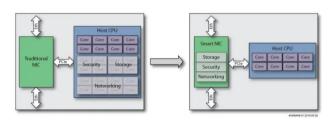


图 1 从传统网卡到智能网卡*

2 原理和优势

2.1 SmartNIC相关论文

现如今智能网卡方面的论文很多,智能网卡的 种类有很多,相应它的应用方向也很广。

我看过一些关于智能网卡方面的论文了,其中 有将 NF (网络功能) 卸载到 SmartNIC 上的,有将 DFS(分布式文件系统)卸载到 SmartNIC 上的, 也有将 SmartNIC 当做一个远端存储器的[10]。但是 在讲解智能网卡之所以能够帮助主机 CPU 进行加 速,减少 CPU 的负担,或者和其他器件一起构造 一个可靠度高的器件,我觉得还是根据 Automated SmartNIC Offloading Insights for Network Functions, 即自动化的网络功能的智能网卡卸载洞 察来讲更好,这篇论文讲的只是讲的智能网卡的一 种应用 SmartNIC offloading (卸载), 虽然它使用的 技术都是机器学习技术,但这不影响我们通过这篇 论文来讲解智能网卡之所以能够作为卸载平台,或 者构成其他器件的原因。因为它是将生成 NF (网 络功能) 卸载到 SmartNIC 的一些见解, 所以它必 然是综合考虑了智能网卡上的各种资源,要充分利 用这些资源,这些资源就是 SmartNIC 具有的优势。

2.2 从选定论文来看SmartNIC原理和优势

想要将一个 NF 从服务器卸载到智能网卡上, 需要对 SmartNIC 硬件的了解,对移植程序的仔细 手工调优,并且运用试错方法,因为我们不能轻易 地推断出迁移性能或不同移植策略的有效性。移植策略主要包括这几个方面: (1) 加速器的使用、(2) 多核扩展分析、(3) NF 状态放置、(4) 内存访问优化、(5) NF 共存建议。而这五个方面也正是SmartNIC 能够被广泛应用的原理和它们具有的优势。^[9]

首先我们来看第一个方面,加速器的使用,因 为 smartnic 有专门的加速器,某些包处理算法(例如 CRC 校验和、最长前缀匹配) 将受益于 SmartNIC 中专用的 ASIC 加速器。因为这些专门的加速器, 某些本来在服务器端运行的代码被卸载到 SmartNIC 时,不仅减少了将数据从存储器复制到 CPU 的步骤, 而且也能够在特定的加速器中得到加 速。第二个方面是,多核,即 SmartNIC 使用多核 并行来提高数据包处理性能,增加 NF 的 core 数通 常会导致更高的吞吐量,但超过某一点后,由于内 存子系统的争用,吞吐量会趋于稳定。这样的争用 也会导致数据包处理延迟随着核数的增加而增加。 此外,根据内存/计算强度的级别,NF 程序将在不 同的核心计数处遇到瓶颈。所以总的来说, 在未到 达临界点之前,核数越多,性能越好。第三点,是 状态放置, 因为与在 x86 系统中, 所有状态都在 DRAM 中维护,缓存子系统是硬件管理的,而 SmartNIC 平台具有更复杂的层次结构,有些缓存是 硬件管理的,但也有一些是软件管理的。此外,如 果没有 OS 或完全的 libc 支持,有状态 NFs 就不能 轻易地依赖熟悉的 malloc 或 brk 调用来动态分配内 存。类似的限制也适用于回收调用。所以我们需要 确认在哪里为每个有状态数据结构分配内存。不同 的状态放置会导致非常不同的结果, 所以我们要仔 细考虑这一点。第四点是内存访问优化,这一点其 实在 pc 也有体现,但是这里我们依据相同的原理 将该操作移植到智能网卡中来,为了减少对内存的 访问次数,选择将经常一起访问的存放在一起,再 根据经常需要访问的数据总和大小来确定内存访 问大小,最终目的都是为了最小化内存延迟。最后 一点, NF 共存建议, 这一点实际上可以从 NF 中推 出,即不单单限于 NF,而是多个可以同时共存在 一个智能网卡上的卸载,应用设置都放置在一起, 但是如果共存的程序彼此不友好, 它们将经历显著 的性能下降。

这五个方面说出来了一些智能网卡之所以被 广泛应用的原因和其优势,但是实际上这只是其中 一部分内容,还有的如 RDMA 引擎等。并且智能 网卡是可以根据特定领域的任务来进行定制的,或选择特定的智能网卡,程序也都会从这种特定中得到更大的收获。文件排版采用 MS Word。

2.3 总结

智能网卡的雏形是微软亚洲研究院在 2014 年 提出的基于现场可编程门阵列 (field programmable gate array, FPGA)的 Catapult设计,一种用于加速大规模数据中心服 务的可重构网络,并在后续的一系列研究中逐渐发展。智能网卡(SmartNIC),也称智能网络适配器,除了能完成标准网卡所具有的网络传输功能之外,还提供内置的可编程、可配置的硬件加速引擎,提升应用的性能和大幅降低 CPU 在通信中的消耗,为应用提供更多的 CPU 资源。这些都是它的优势,也是它能够被各大数据中心大规模使用,被各个实验室研究的原因。

3 研究进展

这么说吧,毕竟我们不是专门搞硬件的,而且设计一个智能网卡对于我们太过于困难,所以在这一部分,即研究进展的部分,主要研究 SmartNIC 的应用的研究进展,即智能网卡被应用到哪些方面。

更快的网络速率、更加复杂的处理场景、更高昂的网络处理开销催生了智能网卡,智能网卡作为一种应用驱动的产物,在众多场景中得到应用。先总的说一下,从四个方面来介绍智能网卡的典型应用场景:(1)网络协议处理;(2)网络功能卸载;(3)数据中心应用;(4)科学计算应用。当然还有其他方面,这里只是以这四个方面为例。

3.1 网络协议处理

智能网卡作为一种具备一定编程能力的网卡, 其最基础的功能就是快速地处理网络协议,提供高 效的网络 I/O。在网络协议种类方面,传统的网卡 多是仅支持一种网络,标准的以太网,或者标准的 Infiniband,或者自定义的网络协议。部分智能网 卡(如 Mellanox 网 卡)则 可 以 根 据 用 户 设 置 , 兼 容 以 太 网 和 Infiniband,除了传 统的 TCP/IP 协议,智能网卡大多支持 RDMA 协议 或者其他加速数据通路的协议,如 Portals 4, RoCE v1/v2, iWARP, 在数据中心或者高性能计算机集群中提供低延时、高带宽的网络服务; 智 能 网 卡 甚 至 支 持 存 储 方 面 的 协 议 , 如 NVMe-oF[53]. 此外,如 Mellanox 产品还支持MPLS (multi-protocol label switching) 协议、GPU-Direct,迎合当下虚拟化、人工智能的应用场景。

在网络协议处理方面,以微软为例,早期的工 作中已经将 TCP/IP 协议中模式固定、计算简单的 处理卸载到网卡,如 TCP 校验(chueksum)、IPsec (internet protocol security) 的卸载;之后, 逐渐有更多的网络协议处理卸载到往卡上,如 RSS 卸载、VMQ(virtual machine queues)卸载。近 些年, 为加速虚拟化的云场景的网络处理, 如 NVGRE, VxLAN 的处理也在网卡中得到卸载处理, 进一步释放主机端 CPU. 此外,智能网卡对多种加 解密方法提供加速,如 Stringrav 智能网卡,可 以卸载 Hash 计算、SHA、MD5、PKA (public key accelerator)等. 在学术界,不断有智能网卡卸 载网络协议的新方式提出,如 TriEC 对现有智能网 卡卸载纠删码 (erasure coding, EC) 的方式进行 了改善, 提出 3 分图式纠删码卸载模式. 1RMA 则 对现有 RDMA 网络可靠连接高开销的通信模式讲 行改善,提出 Connection-free 的网络连接模式, 同时通过软件保证连接的可靠性,并且对安全方面 提供了网卡卸载支持。

3.2 网络功能卸载

网络功能(NF)的作用是通过一系列方式对数据包进行检测和修改,典型的网络功能有防火墙(firewall)、网关(gateway)、入侵检测(instruction detection system, IDS)、负载均衡(load balancer)、域名服务(domain name service, DNS)等。微软基于 Catapult 硬件架构的 ClickNP 编程模型一文中主要对卸载网络功能进行了实现和评估,iPipe 中也对卸载网络功能进行了实现和评估,文献则对 DNS 进行了卸载,从网络功能的性能上看基于 FPGA 的 ClickNP 优于基于多核处理器的 iPipe,可见,相比于数据中心中的其他复杂应用,逻辑相对简单的网络功能更适合流式处理的实现方法。

此外,智能网卡在卸载 SDN 协议栈、加速 SR-IOV、卸载 NFV、卸载 OVS 方面也有很好的应用场景,在

微软的 AccelNet 工作中得到了充分的体现, NetBricks 也对 NFV 进行了卸载。在产业界, Mellanox 则将加速 OVS 的 ASAP2 (accelerated switching and packet processing) 技术应用到 新一代的智能网卡产品中。

3.3 SmartNIC在数据中心的应用

智能网卡在数据中心应用的十分广泛,在此总结为 5 类:

3.3.1 卸载一致性协议

如对 Paxos 一致性协议进行卸载,其中有在交换机端的卸载工作, 其中使用 P4 交换机完成了 Paxos 一致性协议的卸载,也有在网卡端完成一致性协议的卸载,如 P4xos 在网卡上完成了一致性协议的的卸载工作。

3.3.2 卸载 KVS 相关的应用

如 KV-Direct, Lake ,加速分布式共享内存 (distributed shared memory, DSM),如 FaRM, Grappa. 在智能网卡软硬件设计的相关研究中,如 iPipe,Floem,FlexNIC,sPIN,NICA 等,皆以 KVS 作为性能评测的重要指标,在商业智能网卡产品中,KV 加速部件也已成为重要的组件。

3.3.3 加速搜索引擎

如微软加速了 Bing 搜索引擎业务,将吞吐提高了 95%。

3.3.4 加速人工智能应用

如 Lynx,搭建了以智能网卡为中心来调度管理异构 AI 加速器的神经网络训练、推理加速平台,将 CPU 从任务中释放出来做其他事务的处理;也有的则把网络设备作为一种神经网路加速器来使用,卸载神经网络模型中的某些层甚至整个模型,数据在网络传输中被计算,降低延时的同时减轻终端加速器的负载。

3.3.5 提供虚拟化、云环境支持

如微软在 AccelNet 中加速 SR-IOV、卸载 OVS, Freeflow, MasQ则对 RDMA 网络进行了虚拟化, 向多租户提供接近物理网卡性能的虚拟 RDMA 接口. 而 FairNIC, PIEO, Loom, 1RMA[95]则对云环境下的包调度、性能隔离和数据加密方面进行了研究. Pythia 则对 RDMA 数据安全性方面进行了侧信

道攻击的尝试,对网卡安全提出了更高的要求。

3.4 科学计算应用

智能网卡在科学计算中的应用首先表现在通信加速上,如利用 RDMA 的特性进行非连续数据通信的加速、集合通信加速、MPI 加速。其次,智能网卡在科学计算的应用中也可以起到卸载计算的作用,如集合通信加速时

Allreduce 操作中的计算、MPI Tag-Matching均可卸载到网卡处理. 再者,针对科学计算应用中存在大量的矩阵计算的特点,在 INCA中,作者使用网卡进行了矩阵转置、卷积、矩阵乘等与应用紧耦合的计算任务的卸载,可将高性能计算应用加速 11%。

智能网卡在分布式应用中几乎无处不在. 在传统 的 网络 通信 方面,智能 网卡可以满足RDMA, TCP/IP等协议下,基本的网络数据传输甚至部分网络协议的硬件卸载,在如今数据中心网络虚拟化的大趋势下,智能网卡可以提供网络功能的卸载、SR-IOV 的支持、虚拟交换机的卸载等,将部分 CPU 的资源从网络处理中释放出来,E3 研究表明,高效利用智能网卡中的低功耗处理器处理合适的数据中心任务可以将能效比提高 3 倍;在用户应用方面,智能网卡在加速一致性协议、KVS 相关应用、搜索引擎、分布式共享存储、神经网络等数据中心应用方面皆有优秀表现,在科学计算领域的集合通信加速、MPI 加速、矩阵计算中也表现出重要价值。

上述从四份方面来说明了智能网卡(SmartNIC)应用的研究进展,都是从大概的方面谈的,虽然其中也举了一些最新研究的例子,但是并未详细说明,所以说接下来,我将使用自己学习、研读过的论文来具体的说明一下 SmartNIC 的各种应用的研究进展。

3.5 LineFS

第一篇,就已第一节中说过的另一篇论文为例,LineFS: Efficient SmartNIC Offload of a Distributed File System with Pipeline Parallelism,即 LineFS: 高效的具有管道并行性的分布式文件系统的智能网卡卸载。

为了缓解远端写操作的高延迟问题, 并且充分

发挥持久内存的优势,有人提出了Client-local PM的模型,其主要的设计理念是将写操作放在本地机器上,数据写入本地 PM,这样写延迟大大降低。但是 PM 又需要客户机本地的文件系统进行管理,占用本地的 PM 空间,导致了文件管理与应用进程争夺内存和计算资源,导致系统性能下降。为了缓解DFS 占用大量 cpu 开销,现有方案主要利用SmartNIC实现: 1) 让 smartNIC 分担部分数据分类和处理操作; 2) RDMA 时,将远程文件访问、复制中某些操作就近交给 smartNIC 执行。但是现有的方案没有完整地考虑将 DFS 的功能卸载到智能网卡上。本篇论文就是将 DFS 完整的卸载到智能网卡上。

当然把 DFS 功能卸载到 smartNIC 上,会有很多的挑战: 1)、smartNIC 是插在 PCIe 总线上,相比于 cpu 访问 PM,延时也有数量级的差别,因此访问有很大的时间开销,那要如何隐藏这部分开销呢? 2)、smartNIC 上的处理单元的性能相比与主机cpu 性能差很多,因此和 CPU 相比是一个弱的执行环境。我们要如何在弱的执行环境上实现复杂的DFS 呢?

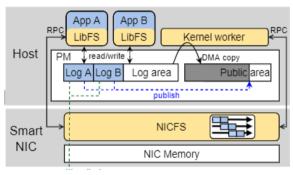


图 2 LineFS 的总体设计图

LineFS 的设计目标是最小化对主机上应用的干扰和降低DFS的延迟。LineFS分为两个部分,host的 LibFS 和智能网卡的 NICFS。将智能网卡上的 DFS 任务分为两部分: persist 和 publish。在 host中,LineFS 根据客户机的 PM 分配一部分 PM 作为日志。LibFS 将数据和元数据更新到私有日志,日志块满足一定大小后,异步发布到本地公共 PM 中,然后由 NICFS 复制到远程 PM中。LibFS 使用主机内核的DMA 将数据和元数据持久化。NICFS 在智能网卡内核后台发布和复制更新,避免主机内核的管理,从而隐藏了时间。[7]

管道并行性: LineFS 利用管道并行来发布和复制日志块。LineFS 将 dfs 操作分为不同的执行阶段,

在同一个执行阶段间构建执行管道。通过管道并行处理每个日志块。LineFS协调管道的沟通,线性化的共享更新。为了便于管理这些管道,NICFS为每个阶段分配一个线程,当LibFS中的日志块满后,发送rpc启动管道。通过监控每个阶段的使用时间,可以判断瓶颈,然后调整分配的线程量来加速处理。

对于发布流程:首先,管道将日志数据提取到NICFS,对其进行验证,然后进行发布和等待验证。在发布之后,LibFS 会回收相应空间,腾出空间。主要的挑战有两方面,PCIe 总线的延迟,验证许可需要消耗一定的计算力。因此为了分摊 PCIe 的传输开销,执行这样的并行线性的管道操作。并且通过主机 DMA 复制来完成发布,避免内核 CPU 参与。另外,在这些周期里面有一些是空闲时间段,可以添加额外的管道,利用 smartNIC 处理性能,如压缩阶段来合并跳过不必要的条目,减少写放大,提高 PM 的使用寿命。这个需要根据 smartNIC 的负载决定。

复制过程也包括四个阶段,前两个阶段和发布过程相同,因此这两个管道是通用的,避免了复制造成的浪费。LineFS 使用 RDMA 线性地将日志复制到多个副本中,以此保持一致性。复制延迟会降低DFS 写入性能。为了降低复制的延迟,将网络操作预发布到 RDMA 的发送接收队列,并在轮询中完成,预发布和轮询不能延迟,并且要在独立的 cpu 下运行。所以将这项工作卸载到 smartnic 上。所以smartnic 将 DFS 请求分为: 低延迟和高吞吐的请求,使用不同的接口分别进行处理。

通过这些操作,我们卸载了一个完整的 DFS (分布式文件系统)到智能网卡上,并且实现了高可用性。但实际上我们可以将这篇论文与前面讲过的第二点联系起来,即网络功能卸载,虽然 DFS 不是一个NF (网络功能),但它们两者的核心本质是相同的,即都是将主机操作系统中的某些功能卸载到网卡上,智能网卡都是当做一个卸载平台来使用的。因为智能网卡上有 cpu core 和内存资源等,足以完成任务,当然关于速度慢的问题,LineFS 使用了管道并行性和后台执行等多种方式来弥补。

3.6 KV-Direct

接下来,看第二篇论文,KV-Direct: High-Performance In-Memory Key-Value Store with Programmable NIC,即KV-Direct:具有可编 程 NIC 的高性能内存键值存储。这篇文章主要是 针对数据中心中非常重要的 KVS (内存键值存储), 也即是与我们上面说的第三点中的某一小点相关。

内存键值存储(KVS)的性能之所以这么重要,是因为现代 KVS 超越了传统的对象缓存工作负载,并成为支持数据中心分布式主内存计算的基础架构的关键。然而数据中心的网络带宽快速增长,大多数 KVS 的瓶颈从网络转移到 CPU。 支持RDMA 的 NIC 部分缓解了这个问题,但 RDMA 抽象提供的原语相当有限。同时,SmartNIC 可用于数据中心,从而实现网络内处理。

早期的键值系统建立在传统的操作系统(OS)抽象之上,例如操作系统锁和 TCP/IP 堆栈。 这给操作系统的性能带来了相当大的压力,尤其是网络堆栈。 由于数据中心应用程序对带宽的大量需求,物理网络传输速度在过去十年中取得了巨大进步,这一事实加剧了瓶颈。随着单核频率扩展和多核架构扩展速度放缓,分布式系统的一个新研究趋势是利用 NIC 上的远程直接内存访问(RDMA)技术来减少网络处理成本,有使用单边 RDMA,也有使用双边 RDMA 的,但效果都不太理想,究其原因,是因为由于缺乏事务支持,RDMA 提供的抽象并不适合构建高效的 KVS。^[8]

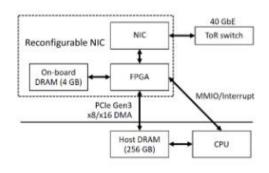


图 3 可编程网卡的架构

在这篇论文中中,介绍了 KV-Direct,这是一种高性能 KVS,它利用可编程 NIC 来扩展 RDMA 原语,并启用对主主机内存的远程直接键值访问。同时开发了几种新技术来最大化吞吐量并隐藏 NIC 和主机内存之间的 PCIe 连接的延迟。KV Direct,顾名思义,直接获取数据并在主机内存中应用更新来服务 KV 请求,绕过主机 CPU。KV-Direct 将 RDMA 原语从内存操作(READ 和 WRITE)扩展到键值操作(GET、PUT、DELETE 和

ATOMIC 操作),还提供了新的向量原语 UPDATE、REDUCE 和 FILTER,允许用户定义活动消息并将某些计算委托给可编程 NIC 以提高效率。

由于键值操作被卸载到可编程网卡,本文将设 计重点放在优化网卡和主机内存之间的 PCIe 流 量上。KV-Direct 采用了一系列优化来充分利用 PCIe 带宽并隐藏延迟。首先,论文作者设计了一 个新的哈希表和内存分配器,以利用 FPGA 中可用 的并行性并最大限度地减少 PCIe DMA 请求的数 量。其次,为了保证依赖的 KV 操作之间的一致性, KV-Direct 包含一个乱序执行引擎来跟踪操作依 赖关系,同时最大化独立请求的吞吐量。第三, KV-Direct 通过在 FPGA 中实现基于硬件的负载 分配器和缓存组件来利用可编程 NIC 上可用的板 载 DRAM 缓冲区,以充分利用板载 DRAM 带宽和容 量。KV-Direct 旨在绕过服务器 CPU, 仅使用一部 分主机内存进行 KV 存储。 因此, CPU 仍然可以 运行其他应用程序。 我们的测量发现, 当单个 NIC KV-Direct 处于峰值负载时,对服务器上其他工作 负载的影响最小。它充分利用了 NIC DRAM 和主机 DRAM, 使我们基于 FPGA 的键值存储系统具有通用 性和大规模部署能力。 此外,作者们精心的硬件 和软件协同设计,以及对 PCIe 和网络的优化,将 性能推向了物理极限,推进了最先进的解决方案。 这一篇论文是关于智能网卡卸载键值存储的研究 实践, 并取得了不错的结果, 当然做类似研究的人 很多,我只选取了其中的一篇来讲是完全不够的, 有兴趣的人应该自行去搜索。上面的两篇论文讲解 分别对应应用方向的第二、三大点,当然也有论文 对应其他的点,这里由于篇幅原因就不在赘述了。 我们只需要明白一件事, 那就是智能网卡的研究、 应用仍然很火热,并且在各位科研人员的研究下, 被广泛的应用在各个方面。

4 总结和展望

我们可以确认在未来,依然会有各色的新应用 驱动智能网卡向更强大的方向发展,如何适应新的 应用需求将是智能网卡设计中的一个重要的问题, 在此,我们从应用场景的专用性和通用性两个方面 做简单的总结: 1)专用性,智能网卡是应用驱动 的产物,在应用相对固定的情况下,如网络功能卸 载、机器学习模型训练、KV 处理、特定算法的加解密,可以针对特定的一类或者几类应用,做针对性的加速设计,如采用 FPGA 构、ASIC 的加速部件。2)通用性,在应用场景相对复杂的情况下,如云环境、多租户场景,需要通用性强的智能网卡,可以采用基于网络处理器、甚至通用处理器的多核架构,提供灵活性更友好的编程接口,满足新应用场景下的可用性。

从这两个角度来看并不是不可以,但是也可以 从另外四个方向来看智能硬件的未来研究领域,分 别是可编程性提升、人工智能增强、光学智能网卡 和开源硬件,这是伴随 5G/物联网/边缘计算的部 署应用和超大规模数据中心东西向流量激增,我们 可以归纳梳理提出的。

当然也还有其他方面的智能网卡提升领域,总的来说就是随着硬件水平的发展,智能网卡本身这个硬件的发展也值得期待。

硬件本身只是硬件,将之应用到各种场景之中才是它的最终归宿。我们可以期待的是,未来关于智能网卡的应用研究会百花齐放,而对于智能网卡开发难问题,我们希望后续会形成社区,供学习者相互交流、学习,解决智能网卡开发研究的闭门造车的问题

参考文献

- [1] 智能网卡(Smart NIC)简介
- ---https://zhuanlan.zhihu.com/p/166482296
- [2] 中国智能网卡(SmartNIC)行业概览
- ---https://jishuin.proginn.com/p/763bfbd6155e
- [3] 马潇潇,杨帆,王展,元国军,安学军.智能网卡综述[J/OL].计算机研究 与发展:1-20[2022-01-06].

http://kns.cnki.net/kcms/detail/11.1777.TP.20210225.1903.008.html.

[4] 张登科,王兴伟,贾杰,李婕.智能网卡研究新进展[J/OL].小型微型计算机系统:1-8[2022-01-06].

http://kns.cnki.net/kcms/detail/21.1106.tp.20210319.0906.004.html.

[5] 刘畅,陶云祥,戎胤.智能网卡应用于云网络加速方案研究[J].电信工程技术与标准化,2020,33(08):76-81.DOI:10.13992/j

.cnki.tetas.2020.08.017.

[6] 马潇潇,杨帆,王展,元国军,安学军.智能网卡综述[J/OL].计算机研究与发展:1-20[2022-01-06].

http://kns.cnki.net/kcms/detail/11.1777.TP.20210225.1903.008.html.

- [7] Kim J, Jang I, Reda W, et al. LineFS: Efficient SmartNIC Offload of a Distributed File System with Pipeline Parallelism[C]//Proceedings of the ACM SIGOPS 28th Symposium on Operating Systems Principles. 2021: 756-771.
- [8] Li B, Ruan Z, Xiao W, et al. Kv-direct: High-performance in-memory key-value store with programmable nic[C]//Proceedings of the 26th Symposium on Operating Systems Principles. 2017: 137-152.
- [9] Qiu Y, Xing J, Hsu K F, et al. Automated SmartNIC Offloading Insights for Network Functions[C]//Proceedings of the ACM SIGOPS 28th Symposium on Operating Systems Principles. 2021: 772-787.
- [10] Sidler D, Wang Z, Chiosa M, et al. StRoM: smart remote memory[C]//Proceedings of the Fifteenth European Conference on Computer Systems. 2020: 1-16.