·学术研究·

多核网络协议栈可扩展性解耦设计*

尚秋里[1,2] 王劲林[1] 陈 晓[1] 叶晓舟[1]

([1]中国科学院声学研究所,国家网络新媒体工程技术研究中心 北京 100190
[2]中国科学院大学 北京 100190)

摘要:高速网络环境下,多核网络协议栈的性能可扩展性尤为重要。针对多核可扩展性问题,本文提出了一种多核网络协议栈可扩展性解耦设计方案,包括数据包和网卡队列两个层面的解耦。其中,数据包层面解耦通过数据流分流映射的方法来实现多核全并行无锁处理;网卡队列层面通过多虚拟队列来解决多核的网卡队列竞争。实验表明,本文提出的解耦方案在 20GE 实验平台上实现了多达 14 个处理核心的线性网络性能扩展,比传统多核并发方案具有更优的可扩展性。

关键词:多核处理器,网络协议栈,可扩展性

A Scalability Decoupling Design for Multi – Core Network Protocol Stack

SHANG Qiuli[1,2], WANG Jinlin[1], CHEN Xiao[1], YE Xiaozhou[1]

([1]National Network New Media Engineering Research Center, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China,
[2]University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In high – speed network environment, the performance scalability of multi – core network protocol stack is crucial. For multi – core scalability, this paper proposes a scalability decoupling design for multi – core network protocol stack, including decoupling of both packets and NIC (Network Interface Card) queue. The packet decoupling obtains parallel lock – free processing by mapping the packet flows to cores. The NIC decoupling relieves queue competition via multiple virtual queues. Experiment shows that, by using the proposed decoupling scheme on 20GE prototype platform, the network performance is scalable linearly for up to 14 processing cores, which is much better than the traditional multi – core concurrent scheme.

Keywords: Multi – core processor, Network protocol stack, Scalability

高速增长的互联网流量和网络带宽对网络协议栈性能提出了挑战。Thumb 定律指出,对于传统基于软件方式的网络协议处理,CPU 每处理 1bit 网络数据,将消耗 1Hz 的处理能力。因此,对 10Gbit/s 以太网数据流的全双工处理,将消耗 20GHz 的 CPU 处理资源[1],这是当前主流企业级服务器都很难具备的硬件配置。另一方面,Gilder 定律预测人们对带宽的需求将以计算能力 4 倍的速率不断增长[2]。因此,多核处理器技术被广泛的应用来提高网络系统的处理性能。由于处理器设计和工艺等因素的限制,单纯提升处理器主频已经难以满足日益增长的性能需求,还会导致功耗和制造成本的增加[3,4]。因此,多核处理器(Multi – core Processors),即片上多处理器(Chip Multi – Processor,CMP)应运而生。多核处理器将多个微处理器内核,即处理核心集成在一起,通过多核多硬件线程并行的方式提高处理性能,具有结构简单、功耗较低等特点[5]。

本文于 2016 – 05 – 12 收到。
* 中国科学院战略性先导科技专项课题,未来网络架构研究与边缘设备研制(XDA06010302)。

然而,基于多核处理器的网络协议栈性能,随着运行处理核心数目的增长,会逐渐接近瓶颈,而无法实现期望的线性增长,即出现可扩展性不足的问题。网络存储系统向通用化和融合系统的发展趋势,要求系统具备弹性可伸缩的应用部署特点,所以提高系统的资源利用率,即性能密度,解决可扩展性问题迫在眉睫[6]。

以下两方面原因制约了多核网络协议栈的可扩展性:

第一,多核同步开销。多个处理核心在处理同一条数据流时,为了保证数据包级别的保序性,需要通过锁实现同步。在高速网络环境下,随着系统负载的升高,数据流在多核之间的锁同步开销急剧上升,极大限制了系统可扩展性。

第二,网卡队列竞争。在高速网络环境下,随着投入处理工作的处理核心数目增长,网卡队列成为了多核协议栈竞争的焦点,进而成为处理性能扩展的瓶颈。例如,关于 Fastsocket 的研究表明,Linux 内核协议栈在 12 核以上会出现性能下降[7]。

针对上述问题,本文提出了一种多核网络协议栈可扩展性解耦设计方案,包括数据包和网卡队列两个层面的解耦。其中,数据包层面解耦通过数据流分流映射的方法来实现多核全并行无锁处理;网卡队列层面通过多虚拟队列来解决多核的网卡队列竞争。

1 多核拓扑模型

基于多核处理器的网络系统,一般由管理平面、控制平面、数据平面等主要模块构成。其中,数据平面负责网络协议栈处理和网络数据包收发,主要包括上层应用接口、应用层网络协议栈、TCP/IP 协议栈、网络适配层等。数据平面对整个系统的网络数据处理性能起关键作用,通常需要具备高吞吐和高可扩展性的特点。

数据平面的多核拓扑模型,即处理核心上任务部署和映射关系,主要有 RTC(Run to Complete)和 SPL(Software PipeLine)[8]两类模型。其中,RTC 模型为全并行结构,各处理核心并行执行任务,每个核独立完成完整的任务处理流程,而互相之间不进行切换且不构成流水线关系。SPL 模型的各处理核心之间逐级构成流水线结构,每一级处理核心依次完成任务处理流程中的不同阶段。SPL 模型又分为两类:S - SPL(Single Software PipeLine)为严格的单流水线结构,通常处理核心数目等于处理阶段数目;P - SPL(Parallel Software PipeLine)结合 RTC 和 S - SPL,采用并行结构扩展 S - SPL 模型中各个阶段的处理核心数量,以提供对关键处理阶段的优化。

由于 P - SPL 模型能兼顾处理性能和扩展性,适用于网络协议栈处理,因此网络系统的数据平面一般按照 P - SPL 来组织各逻辑模块[9,10]。

图 1 中的数据平面以 P - SPL 拓扑模型部署在 7 个处理核心上。其中,网络协议栈以 RTC 方式部署在 4 个处理核心上,可以独立完成各自网络协议栈(以 SCSI/iSCSI + TCP/IP 协议为例)处理任务。同理,网络适配层也以 RTC 的方式部署在 2 个处理核心上,负责网络数据包输入/输出和数据流管理调度。在宏观处理流程上,数据平面中应用适配层、网络协议栈、网络适配层又构成了逐级串联的 S - SPL 流水线关系。由于部署在 Core 1 - Core 4 上的网络协议栈为全并行方式,因此各处理核心进行数据包保序操作所引发的核间同步,以及各处理核心进行网卡轮询所引发的队列竞争,导致了不可忽视的开销,制约了系统的性能可扩展性。

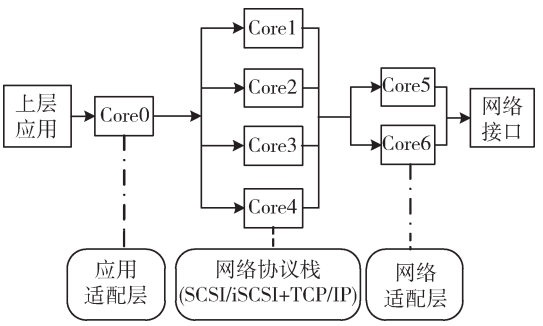

图 1 数据平面的多核拓扑模型

2　可扩展性解耦设计

本文以 Cavium OCTEON 系列多核网络处理器[11]为原型平台，研究多核网络协议栈的可扩展性解耦设计，主要包括数据包和网卡队列两个层面的解耦。

2.1　数据包层面解耦

数据包层面解耦，通过数据流映射分流的方法实现。本文以数据流（flow）为单位来调度数据包，将数据流与处理核心进行映射绑定，避免数据流在处理核心之间的切换。这种方法利用处理核心对数据流的亲和性，一方面缓解多核同步开销，一方面提高 Cache 利用率。

基于原型平台提供的基本机制，本文采用工作组（work group）机制和标签（tag）机制来实现数据流映射。

对于处理核心，使用工作组来实现分组。假设多核处理器平台有 N 个处理核心，序号为 $core_0 - core_{N-1}$，则可以将其划分为 N 个工作组，工作组序号为 $grp_0 - grp_{N-1}$。每个处理核心只能存在于唯一一个工作组中，每个工作组中可以有 0 – N 个处理核心。

对于数据包，采用标签（tag）机制来区分数据流。每个数据包的 32bit 标签值通过对五元组（源 IP 地址、目的 IP 地址、协议号、源端口、目的端口）的 Hash 运算得到。属于同一个数据流的数据包具有相同值的标签。因此每个数据包可以通过 Hash 查找其 tag 值，唯一映射到其对应的数据流。处理核心与数据流，即标签与工作组的映射关系，通过自定义的 Hash 函数实现。例如，一种实现数据流到处理核心唯一映射的 Hash 函数为式（1）。

$$grp = Hash(tag) = tag \bmod N \tag{1}$$

另外，在本文原型平台 OCTEON 多核网络处理器中，可以使用 PIP/IPD 引擎提供的 Hash 硬件协处理器来卸载 Hash 映射的计算和查找，并处理 Hash 哈希碰撞，从而减小计算开销，提升处理效率。

上述方法确保在某一并行处理阶段，某一数据流只映射到同一组（个）处理核心上，避免多核之间的切换，实现了数据包的全并行无锁处理，即数据包层面的解耦。图 2 为采用上述方法对数据包分流解耦的一个例子。

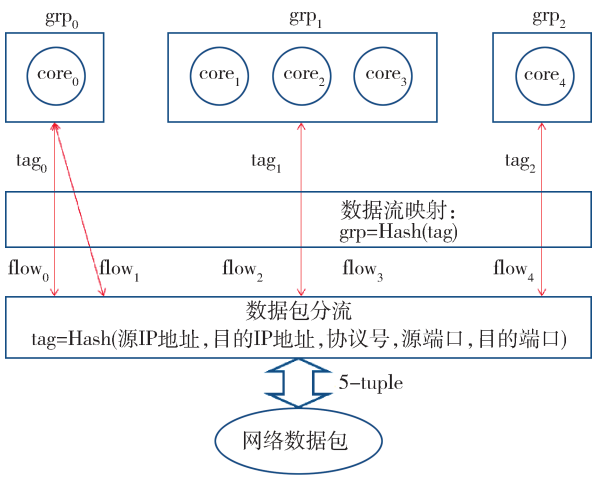

图 2　数据包分流解耦示例

2.2　网卡队列层面解耦

在现有软件框架中，运行网络协议栈的多处理核心并发竞争网卡队列控制权。获得网卡队列控制权的处理核心对队列进行加锁，如果此时其他核心需要访问该网卡队列，将不断尝试获得控制权并处于忙等待状态，这将引入极大的竞争开销。而在高速网络环境中，随着业务负载的加重，对于网卡队列的竞争将成为影响系统性能扩展的瓶颈，甚至整个系统的网络性能将完全取决于该网卡队列吞吐量。

网卡队列层面解耦，通过网卡多队列的方式实现，通过多队列与多处理核心/处理线程的映射和绑定，将网卡上数据分流到多核并行协议栈上，从而实现网卡队列竞争的解耦合。

RSS（Receive – Side Scaling）是目前应用很广泛的一种网卡多队列技术，但是，RSS 存在的问题包括：第一，需要依赖专用网卡提供的多队列功能；第二，网卡支持的队列数目有限，无法扩展，在处理核心较多的多核平台上（CN58XX 有 16 个处理核心），无法做到队列到处理核心的一一映射，例如 Intel 82576 网卡最多支持 8 条队列；第三，RSS 基于中断调度队列，都很难在网卡层面，根据特定业务灵活地部署多队列调度策略。

因此，针对上述问题，本文提出的网卡队列解耦方法，不依赖于特定网卡功能，在数据平面的网卡资源适配层中，为每个网卡实现一组逻辑通道（虚拟队列）来作为网卡的虚拟队列，并由数据平面中负责数据包

收发的处理核心，来负责维护、管理和调度逻辑通道。在网卡适配层中，为每一网卡分配 N 个逻辑通道（N 为工作组数），并将逻辑通道映射到处理核心工作组。这样，不同工作组就可以通过访问对应的逻辑通道，并发、有序地访问同一网卡，从而缓解队列竞争。

假设有 M 个网卡，则网卡 $k(0 \leqslant k \leqslant M-1)$ 的 ID 为 NIC_k，其逻辑通道 ID 为 $Q_{k.0} \sim Q_{k.N-1}$，则共有 $M * N$ 个逻辑通道。具体的映射的方法是，将网卡 NIC_k 上所有工作组为 grp_i 的数据流集合，映射到逻辑通道 $Q_{k.i}$，则由工作组 grp_i 来处理该数据流集合。数据流在不同逻辑通道上的映射（分流），同样可以借助 OCTEON 提供的 Hash 协处理器，来减小计算开销。这样，每个工作组就可以独立地管理和调度 M 个逻辑通道，同一网卡的 N 个逻辑通道也可以并发、有序地被访问。

如图 3 所示，是一个简单的逻辑通道映射例子。图 3 中将 4 个处理核心，分别划分成 4 个工作组，两个物理网卡分别虚拟出 4 的硬件队列，与工作组一一映射。

上述方法提供了一种针对多核平台网卡资源竞争的解耦机制，也为根据具体业务特征灵活部署网卡多队列调度算法提供了基础框架。例如，可以在图 3 中，根据 4 个工作组的负载强度变化（或业务流量特征），实现网卡 NIC_1 上 4 个逻辑通道的多队列优先级轮询算法、或负载均衡算法，从而优化业务服务质量。

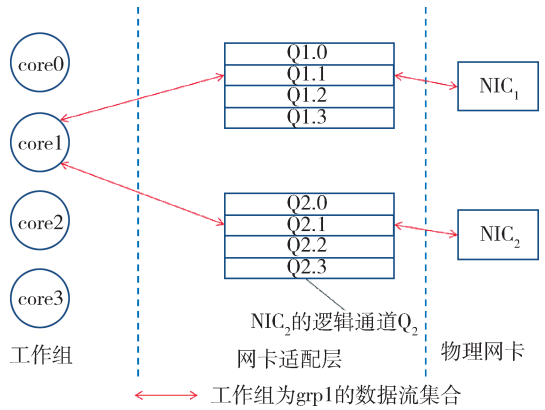

图 3 逻辑通道映射举例

3 实验与分析

3.1 实验环境

本文采用基于 Cavium OCTEON 多核网络处理器的原型平台作为实验环境，测试所提出的多核网络协议栈可扩展性解耦设计方案（简称“解耦方案”）的性能，本实验所选用的 OCTEON 处理器的型号和配置如表 1 所示。对比组采用朴素的全并行数据平面方案（简称“并发方案”），这种方案对全并行网络协议栈的多核并发竞争和锁同步不加以干预和调度。

在数据包层面，并发方案通过 OCTEON 处理器提供的 ordered 机制[11]来调度数据包，从而实现数据流加锁和数据包保序。在网卡队列层面，并发方案中各并行协议栈依次以轮询的方式访问网卡，隐含了对网卡资源的加锁和互斥。不失一般性，本实验中解耦方案的每个并行处理核心划分为一个独立的工作组，并在网卡资源解耦中为每个处理核心分配一个逻辑通道。本文实验中的网络协议栈采用 SCSI/iSCSI + TCP/IP 协议栈，以多核 RTC 的方式部署在 OCTEON 多核处理器的简单执行（Simple Executive，SE）环境中，作为网络系统数据平面的一部分。

表 1 实验平台参数

处理器名称	Cavium OCTEON 5860
处理器频率	MIPS64
处理器规格	800MHz
处理器核数	16
内存	8GB
网络接口	10GE * 2

3.2 实验结果

如图 4、图 5 所示，分别为通过 1 个、2 个万兆网络接口（10GE）时，两种方案的网络吞吐率实验结果，其中横坐标为运行的处理核心总数。本实验中将数据平面的应用适配层和网络适配层分别部署在 1 个处理核心上，网络协议栈部署的处理核心数目范围为 4 个 - 14 个。

如图 4、图 5 所示，平均意义上，解耦方案的吞吐率略优于并发方案，平均超过 3.6%（10GE）和 9.1%（20GE）。随着处理核心数目的增加，解耦方案的优势逐渐显现，当处理核心数大于 10 时，解耦方案的性能超过并发方案 9.7%（10GE）和 14.3%（20GE）。从测试结果可以看到，随着处理核心的增加，并发方案逐渐增长放缓，其吞吐率平均增长率分别从 0.53Gbps/core（10GE）和 1.1Gbps/core（20GE），即 0.593bit/Hz（10GE）和 1.375 bit/Hz（20GE）下滑到了 0.319 bit/Hz（10GE）和 0.481bit/Hz（20GE），即出现了可扩展性不

足的问题；而解耦方案的性能以接近线性的方式增长，其吞吐率平均增长率（bit/Hz）始终保持在[0.625，0.688]（10GE）和[1.125，1.575]（20GE）区间。因此，上述实验表明本文提出的解耦设计方案显著改善了多核网络协议栈的性能可扩展性。

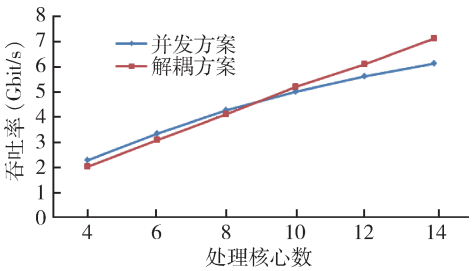

图 4　吞吐率性能（10GE 网络）

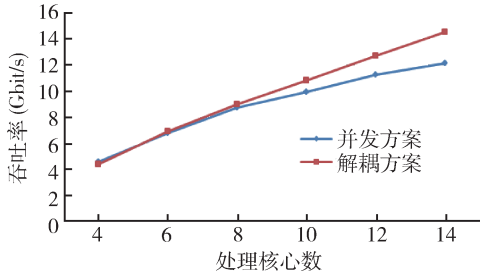

图 5　吞吐率性能（20GE 网络）

4　结束语

随着互联网流量和网络带宽的高速增长，多核处理器已经成为网络协议栈高速处理的首选平台。多核网络协议栈的可扩展性问题不容忽视，包括多核同步开销和网卡队列竞争两方面。针对上述问题，本文提出了一种多核网络协议栈可扩展性解耦设计方案，包括数据包和网卡队列两个层面的解耦。其中，数据包层面解耦通过数据流分流映射的方法来实现多核全并行无锁处理；网卡队列层面通过多虚拟队列来解决多核的网卡队列竞争。实验表明，本文提出的解耦方案在 20GE 实验平台上实现了多达 14 个处理核心的线性网络性能扩展，比传统多核并发方案具有更优的可扩展性。

参　考　文　献

[1] Wang W F, Wang J Y, Li J J. Study on Enhanced Strategies for TCP/IP Offload Engines[C]//International Conference on Parallel and Distributed Systems, 2005. Proceedings. IEEE Xplore, 2005.

[2] Gilder G. Telecosm: How infinite bandwidth will revolutionize our world[M]. Free Press, 2000.

[3] 查奇文，张武，曾学文，等. 基于多核处理器的 TCP/IP 协议栈加速技术[J]. 网络新媒体技术，2013，2(1)：58－64

[4] Guenter B, Jain N, Williams C. Managing cost, performance, and reliability tradeoffs for energy－aware server provisioning [C]//INFOCOM, 2011 Proceedings IEEE. IEEE, 2011.

[5] 周伟明. 多核计算与程序设计[M]. 华中科技大学出版社，2009.

[6] Gu Q, Wen L, Dai F, et al. StackPool: A High－Performance Scalable Network Architecture on Multi－core Servers[C]//IEEE International Conference on High PERFORMANCE Computing and Communications. 2013.

[7] Fastsocket[EB/OL]. http://www.oschina.net/p/linux－fastsocket

[8] 贺鹏程. 面向流的多核分组调度与传输技术研究[D]. 中国科学院研究生院，2010.

[9] 贺鹏程，王劲林，邓浩江，等. 多核分组处理系统软件结构研究[J]. 网络新媒体技术，2010，31(9)：12－20

[10] 郭秀岩，张武，王劲林，等. 用于视频点播系统中实时数据流发送的多核结构[J]. 小型微型计算机系统，2011，32(7)：1310－1316

[11] Networks C. OCTEON Plus CN58XX Multi－Core MIPS64 Based SoC Processors [EB/OL]. http://www.cavium.com/OCTEON－Plus_CN58XX.html.

作者简介

尚秋里，男，博士研究生，主研领域：网络系统、TCP/IP 协议栈。

王劲林，男，硕士，研究员。

陈晓，男，硕士，研究员。