



专题:信息中心网络(ICN)

基于 NetFPGA10G 的网络组件存储功能设计与实现 *

张硕琳, 罗洪斌

(北京交通大学下一代互联网互联设备国家工程实验室 北京 100044)

摘 要:以内容为中心的下一代互联网体系结构是目前的研究热点。其网络组件需要同时具备基本转发功能和内容存储功能。在 NetFPGA10G 硬件平台上,设计并实现了具有存储功能的网络组件,可以同时进行数据转发和内容存储,提高了网络资源的利用率。

关键词:下一代互联网;网络组件;存储功能;NetFPGA10G

doi: 10.3969/j.issn.1000-0801.2014.09.003

Design and Implementation of Storage of Network Component Based on NetFPGA10G

Zhang Shuolin, Luo Hongbin

(National Engineering Laboratory for Next Generation Internet Interconnection Infrastructure,
Beijing Jiaotong University, Beijing 100044, China)

Abstract: Presently, it is hot to do research about the next generation network architecture based on the content centric network. The network component, as an important part of it, will be able to forward as well as save the content in the meantime inevitably. New type of network component was designed and implemented, based on NetFPGA10G hardware platform, to forward the data and store the content at the same time, thus to improve the utilization rate of network resources.

Key words: next generation network, network component, storage function, NetFPGA10G

1 引言

随着用户规模和应用数量的不断增加,现有互联网逐渐暴露出许多原始设计的不足。例如,现有互联网的体系结构采用的 host to host 通信模式,难以适应当前用户对内容获取和分发的多样化需求。当同一个子网的两个用户要获取同一个内容时,都要向远端的服务器获取,从而导致资源浪费。为了解决传统互联网的上述不足,以内容为中心的下一代互联网体系结构应运而生。

在以内容为中心的下一代互联网体系结构中,存储数

据分组将作为网络节点的功能之一。当数据分组流经网络节点时,网络节点应该能够根据控制信息的要求,存储相应的服务内容。当相同的服务再次被请求时,网络节点可以从自己的存储区域中提取相应的数据,就近为用户提供服务,从而提高网络资源利用率。

目前对于网络节点存储功能的研究大多停留在仿真阶段,并没有实际的硬件实现。本文基于 NetFPGA10G 硬件平台,设计具有存储功能的网络组件。之所以选择 NetFPGA10G 作为开发平台,是因为 NetFPGA 平台能够很好地支持模块化设计,使得复杂的硬件设计可以通过各个

* 高等学校“基本科研业务费”资助项目(No.2014JBM011)

模块的组合来完成。而且,NetFPGA 是一个基于 Linux 的开放性平台,所有对它感兴趣的人都可以利用平台上现有的资源,在前人开发的基础上搭建自己的系统;而自己开发的系统也可以被其他人所用,从而不需要重复地搭建外围模块、开发驱动和 GUI 等,只要添加自己的模块和修改现有的系统即可。NetFPGA 平台的模块化和开放性,大大简化了网络研究的任务量。

2 平台架构

传统网络中的路由器用于连接多个逻辑上分开,同时具有控制和转发功能的网络节点。而在下一代互联网体系结构的研究中,大多数研究人员认为,网络组件需要引入控制层面与数据层面的分离机制,将控制功能从网络组件中剥离出去,交给控制器完成,而网络组件只负责根据控制器下发的控制信息,转发相应的数据分组,从而实现网络组件最基本的功能。而在以内容为中心的网络体系结构下,网络组件必须具有存储功能。首先,本文基于 NetFPGA10G 的网络组件,在 NetFPGA10G 板卡上搭建 OpenFlow 交换机,上层控制器可以通过数据通路,根据 OpenFlow 协议,向 NetFPGA10G 板卡下发流表信息,当数据分组到来时,NetFPGA10G 查询流表,根据查询结果完成相应的转发动作,从而实现网络组件的转发功能。然后,本文重构了 NetFPGA10G 万兆光纤网卡驱动,使之在与 NetFPGA10G 通过 PCIE 连接的主机的内存空间中利用内存分配技术,分配 2 GB 的存储空间,并用 STL 提供的 list_head 对存储空间进行管理。当接到存储分组请求时,进行查询匹配,然后根据请求分组中的路由信息进行分组转发。基于 NetFPGA10G 的具有存储功能的网络组件如图 1 所示。

基于 NetFPGA10G 的具有存储功能的网络组件主要

由两个重要部分组成:基于 NetFPGA10G 的 OpenFlow 交换机以及万兆光纤网卡驱动。OpenFlow 交换机用于接收上层控制器下发的控制信息,配置流表,查询流表,完成相应的转发、丢弃等动作。当不需要存储的分组到达 NetFPGA10G 板卡时,首先查询精确匹配表和通配符匹配表,如果有对应的流表项,则执行 action 域所规定的动作;如果没有对应的流表项,则向控制器发出请求信息,等待控制器下发控制信息之后再转发。当需要存储的分组到达板卡时,NetFPGA10G 会将分组通过 AXI-lite 总线发送给 PCI/DMA 模块,之后通过万兆光纤网卡驱动进行存储。当存储空间有空闲时,数据分组将直接存储下来;当存储空间为满时,执行替换算法,对存储的内容进行更新替换。当查询分组到达时,驱动将根据内容名字,查找存储空间是否有相应的请求内容,如果有就返回内容到查询分组进入的端口,如果没有就返回相应的提示信息。这就构成了基于 NetFPGA10G 的具有存储功能的网络组件。

3 主要技术实现

3.1 OpenFlow 交换机在 NetFPGA10G 上的设计与实现

在 NetFPGA10G 平台上搭建 OpenFlow 交换机,具体设计框架如图 2 所示。

数据分组可以通过 4 个光纤网口进入交换机进行处理,其主要模块分为以下 4 个部分:分组预处理模块、流表控制模块、行为处理模块以及主机接口模块。

(1) 分组预处理模块

数据分组进入该模块之后,该模块负责解析提取分组头部分的信息,将信息组装成 OpenFlow 协议规定的 12 元组,该信息称为流表项。之后,分组预处理模块会将流表项发送给流表控制模块,并且将原始数据分组发送到行为处

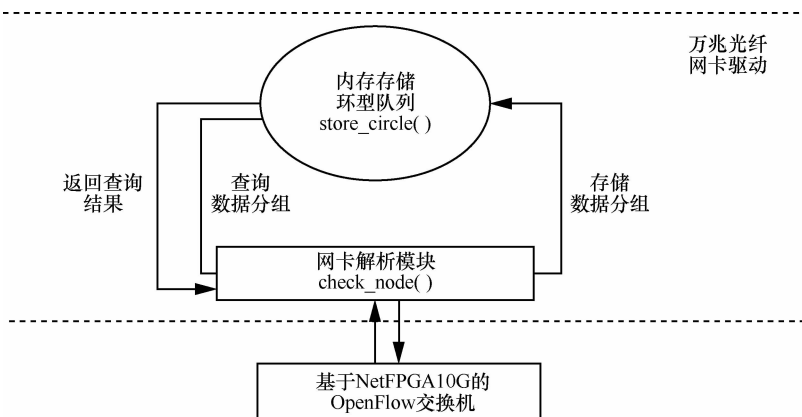


图 1 具有存储功能的网络组件

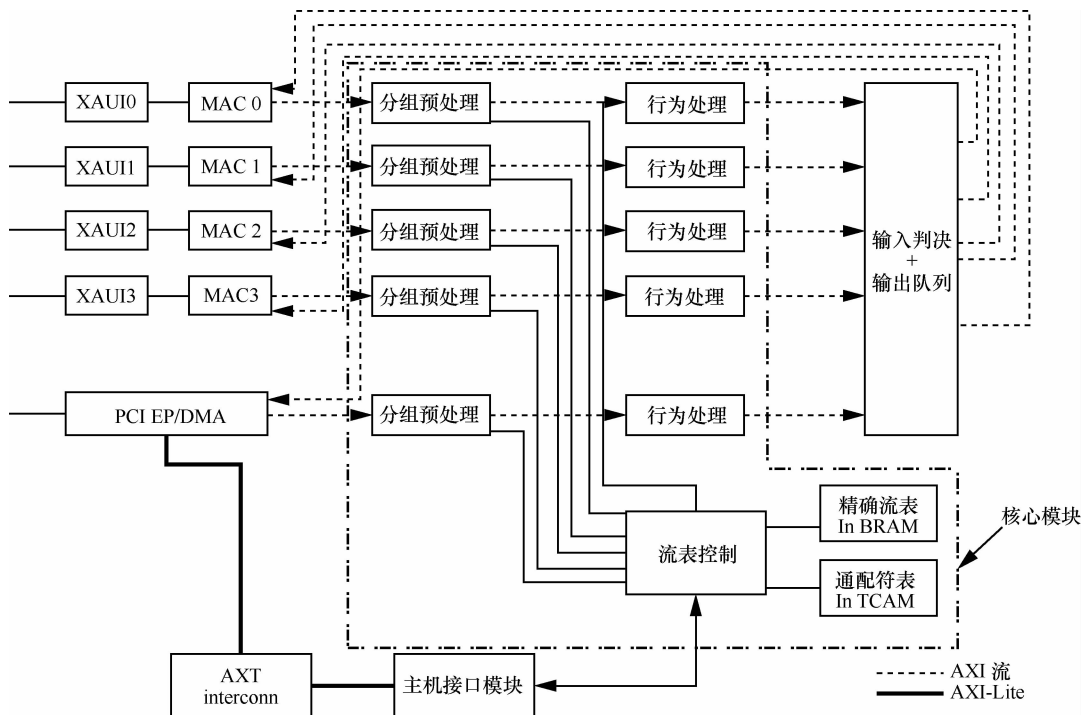


图2 OpenFlow 交换机硬件设计

理模块等待。

(2) 流表控制模块

流表控制模块通过一个通往通配符流表和精确流表的单一管道来访问交换机的流表,该管道可以被所有的端口共享,所以该模块响应各端口对其发起的所有请求。在接收查找流表项请求后,流表控制器将同时查找精确匹配流表与通配符流表,若在流表中找到了匹配表项,流表控制器就从内存中抓取相应的行为列表并把它发给对应端口的行为处理器,同时更新其统计数据。若流表控制器在精确匹配流表和通配符流表中都匹配成功,则使用精确匹配流表中的转发行为。

(3) 主机接口模块

该模块是硬件交换机与上位控制器之间沟通的桥梁,控制器可以通过它来访问交换机中的寄存器,并对流表进行修改、添加和删除等操作。因为主机向交换机写入流表是以字为单位的,首先在主机接口模块中缓存所有信息,一旦主机的一个写操作完成,主机接口触发通信,通过使用分组预处理模块与流表控制模块通信时一样的握手协议和流表控制器进行通信。

(4) 行为处理模块

行为处理模块根据流表控制模块发送的转发行为列

表转发数据,更新其分组头域与数据分组长度。之后将重新封装好的数据分组发送到输出队列排队等待处理。

以上4个模块构成了基于NetFPGA10G的OpenFlow交换机核心功能,完成了解析分组、组装流表项、查询流表项、重构流表、封装分组、执行转发行为等一系列分组处理过程,实现了实际硬件网络组件的基本转发功能。

3.2 存储功能设计与实现

网络组件存储功能的实现,是在网络组件交换功能的基础上,增加一个行为(action),当需要查询数据分组或存储数据分组到达时,硬件将该数据分组的行为定义为cache,转发端口定义为PCI/DMA端口(5号端口),所以该数据分组将通过PCI通道进入与NetFPGA10G相连的主机中进行查询和存储。具体分组处理流程如图3所示。

在上述查询分组和存储分组的处理流程中,关键技术部分是:存储空间的管理。其主要分为以下几个步骤实现。

(1) 创建存储环型空间

物理地址被分成离散的单元,称为页。系统内部许多对内存的操作都是基于单个页的,目前大多数系统都使用每页4 096 byte。在本设计中,存储的单个数据分组最大为1 533 byte,本文选择每页存储两个数据分组,即存储空间

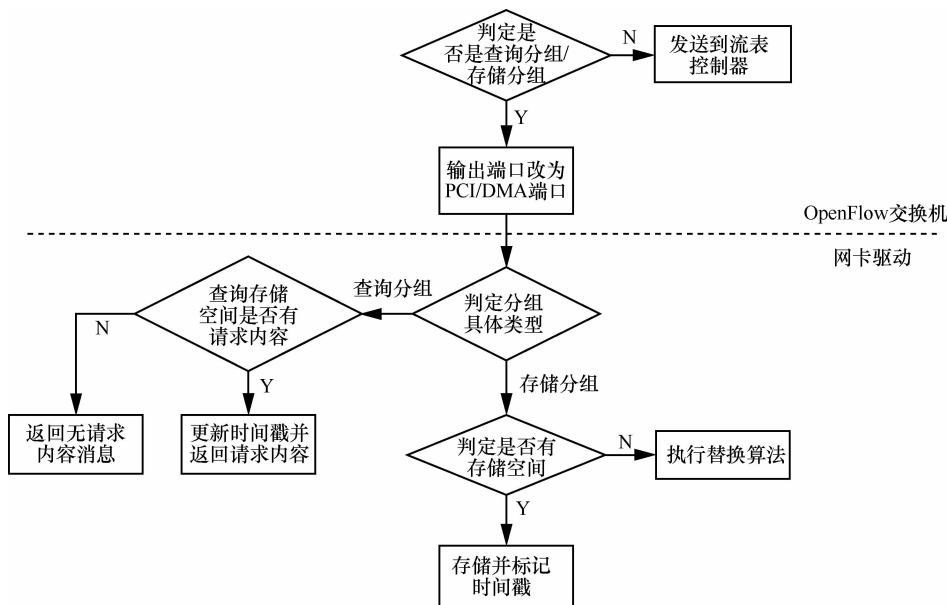


图 3 查询分组/存储分组处理流程

单元是 2 KB。同时,本设计总共分配了 2 GB 的内存作为网络组件存储空间,并用 `get_free_pages` 进行内存申请。其中存储单元类型为 `store_node_list`,并定义其结构体如下:

```

typedef struct{
    struct sk_buff *s_skb;
    int port;
    ktime_t tstamp;
    struct list_head list;
}store_node_list;
    
```

其中, `port` 为存储输入端口号; `tstamp` 为每个分组更新的最新时间戳,每一次数据分组被请求时,时间戳都会更新; `list` 是查询和维护存储空间的数据结构,后面将有具体说明。

(2) 存储空间的维护

由 `get_free_pages` 获得的内存是不连续的,需要数据结构进行管理和维护。利用 Linux 内核中的双向链表 `list_head` 将多个存储单元连接起来。 `list_head` 结构具备双向链表功能,通常它都组织成双循环链表。与普通链表不同的是, `list_head` 没有数据域, `list_head` 具体定义如下:

```

struct list_head {
    struct list_head *next, *prev; };
    
```

在 Linux 内核链表中,不是在链表结构中包含数据,而是在数据结构中包含链表节点,正如 `store_node_list` 结构体中的定义,包含 `list_head` 结构成员。在遍历链表时,由于

存储空间链表中仅保存了数据项结构中 `list_head` 成员变量的地址, Linux 为此提供了一个 `list_entry(ptr, type, member)` 宏,可以根据 `list_head` 的地址获得数据存储地址,从而对相应的存储单元进行操作。

(3) 存储数据分组

当需要存储的数据分组到达时,如果有空闲的存储空间,数据分组就顺序存储在空闲的单元中,并且为该数据分组打时间戳;如果没有空闲的存储单元,将遍历存储空间,选择时间戳最大的单元(即热度最小的存储分组)进行替换。

4 测试

对本文所设计的基于 NetFPGA10G 的具有存储功能的网络组件进行功能测试。测试的硬件环境为: CPU 为 Intel Core i5-2400S CPU @ 2.50 GHz, 内存为 4 GB, 数据分组存储空间为 2 GB, 网卡为万兆光纤网卡; 软件配置为: Linux 操作系统, 内核版本为 2.6.39, 利用 TCPdump 查看本设计模块功能测试结果, 如图 4 所示。

图 4 中标号为 0 的是查询分组, 标号为 4 的是数据分组, 本文设计模块成功实现了数据分组的存储和查询功能。

08:32:16	IP10 bad-hlen 0
08:32:16	IP10 bad-hlen 0
08:32:16	IP10 bad-hlen 4
08:32:16	IP10 bad-hlen 4

图 4 测试结果

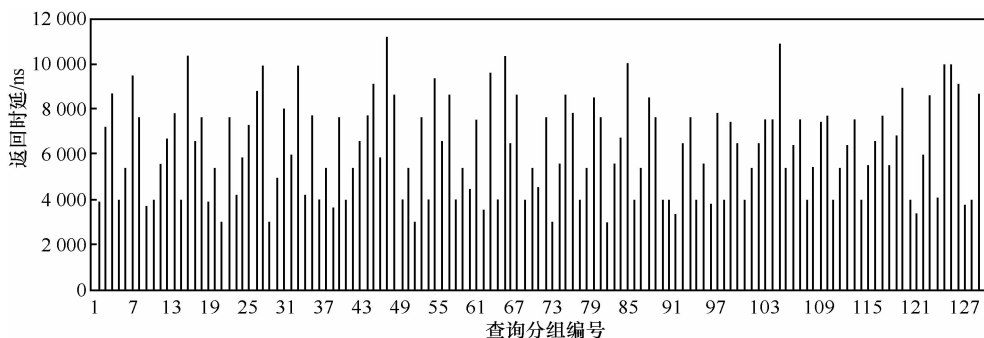


图5 1 Gbit/s 存储速率下查询分组返回时延

关于本文设计模块的性能测试见表1。

表1 性能测试结果

数据分组存储速率	分组丢失率	查询分组平均返回时延/ μ s	CPU 占用率
500 Mbit/s	3.6%	6.2	82%
1 Gbit/s	6%	6.9	85%
2 Gbit/s	9%	8.1	93.2%

可以看出,在不同的存储速率下,本文设计的模块功能实现相对稳定,由于本文设计中,对于存储空间的维护和管理使用的是链表方式,当所查询的数据分组在存储空间的不同位置时,所花费的查询返回时延是不定的,本文选择在1 Gbit/s 存储速率下,查询分组返回时延数据的详细说明,具体时延分布如图5所示。

由于查询分组查询的内容在存储空间中是随机分布的,查询方式根据索引值进行遍历,可以看出,查询返回时延分布在3~11 000 ns,平均返回时延为6.9 ns。

5 结束语

本文对下一代互联网体系结构中的热点问题进行了研究,创新性地对具有内容存储功能的网络组件设计给出了实现方案,提高了网络资源的利用率,对下一代互联网研究有着非常积极的作用。

参考文献

- 1 Arianfar S, Nikander P, Ott J. On content-centric router design and implications. Proceedings of the 6rd International Conference on Emerging Networking Experiments and Technologies (CoNEXT), Philadelphia, USA, 2010
- 2 Luo H B, Chen Z, Cui J B, *et al.* ColoR: an information-centric

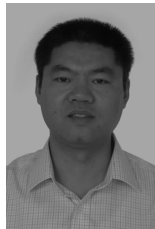
internet architecture for innovation. IEEE Network Magazine, 2014(2)

- 3 OpenFlow implementation on NetFPGA 10G design document. <http://www.OpenFlowswitch.org>
- 4 刘靖龙,刘颖,张思东. 基于 RTL8169 网卡的零拷贝技术研究与实现. 计算机技术与发展, 2011, 21(8)
- 5 Min S H, Choi Y C, Kim N, *et al.* Implementation of a programmable service composition network using NetFPGA-based OpenFlow switches. Proceedings of 1st ASIA NetFPGA Developer's Workshop, Korea, 2010
- 6 Corbet J, Rubini A. Linux 设备驱动程序. 魏永译. 北京: 中国电力出版社, 2006

[作者简介]



张硕琳,女,北京交通大学下一代互联网互联设备国家工程实验室硕士生,主要研究方向为下一代互联网体系结构,参与国家“973”计划基础科研项目。



罗洪斌,男,北京交通大学教授、博士生导师,下一代互联网互联设备国家工程实验室副主任,中国通信学会第四届青年工作委员会委员,北京通信学会第八届理事会理事,入选教育部新世纪优秀人才支持计划。目前主要从事未来互联网体系架构、理论与关键技术方面的研究工作,主持和参与多项国家“973”计划、“863”计划、国家自然科学基金项目。近年来在 IEEE/ACM Transactions on Networking、IEEE Journal on Selected Areas in Communications、IEEE Network 等本领域高水平期刊与国际会议发表论文 50 余篇。

(收稿日期:2014-07-20)