

引用格式:孔德伟,袁国顺,刘小强. 基于 FPGA 的万兆以太网链路的设计与实现 [J]. 微电子学与计算机, 2019, 36(12): 21-25. [KONG D W, YUAN G S, LIU X Q. The design and implementation of 10 gigabit ethernet link based on FPGA [J]. Microelectronics & Computer, 2019, 36(12): 21-25.]

基于 FPGA 的万兆以太网链路的设计与实现

孔德伟^{1,2}, 袁国顺¹, 刘小强^{1,2}

(1 中国科学院 微电子研究所, 北京 100029; 2 中国科学院大学, 北京 100049)

摘 要: 本文设计了基于 FPGA 的万兆以太网链路传输系统. 对现有七层 OSI 模型进行简化, 设计了五层网络传输模型, 并深入研究了传输层、网络层和数据链路层. 实现了完备的 UDP/IP 协议、ARP 协议和万兆数据链路层 MAC 控制器, UDP/IP 协议实现了数据的封装与解封, ARP 协议解决了 IP 地址与 MAC 地址映射的问题, 方便大规模的系统级联. 使用重复的 CRC 校验单元, 实现了 CRC 校验零延迟, 加快了数据的传输过程. 基于 Xilinx 的 Kintex UltraScale 器件进行了系统设计实现, 并使用 MC02 开发板进行了系统功能测试.

关键词: 万兆以太网; FPGA; UDP/IP; ARP; CRC

中图分类号: TP312

文献标识码: A

文章编号: 1000-7180(2019)12-0021-05

The design and implementation of 10 gigabit ethernet link based on FPGA

KONG De-wei^{1,2}, YUAN Guo-shun¹, LIU Xiao-qiang^{1,2}

(1 Institute of Microelectronics of China Academy of Sciences, Beijing 100029, China;

2 University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Based on the FPGA hardware, this paper designs a 10 Gigabit Ethernet link transmission system. Simplifying the seven-layer OSI model, the paper designs a five-layer network transmission model and deeply studies the transport layer, network layer and data link layer. A complete UDP/IP protocol, ARP protocol and 10 Gigabit data link layer MAC controller are implemented. The UDP/IP protocol implements data encapsulation and decapsulation. The ARP protocol solves the mapping problem between IP address and MAC address, which is conducive to build a large-scale system. Using repeated CRC units, the paper implements CRC with zero delay, which speeds up the data transmission process. The system was designed and implemented on the Xilinx-based Kintex UltraScale device and the system function test was performed on the MC02 development board.

Key words: 10 Gigabit Ethernet; FPGA; UDP/IP; ARP; CRC

1 引言

随着科技的发展和进步, 短视频业务的兴起, IoT 的布局以及人工智能 (Artificial Intelligence, AI) 在视频监控行业中的应用, 数据量将会陡增, 这对网络的带宽和实时性有了更高的需求, 千兆以太

网已经不能满足现有的要求. 虽然 5G 的兴起带来的高速、高带宽能满足其条件, 但因其铺设成本高昂, 损耗大等缺点, 目前无法应用于专用局域网络内, 10G 以太网通用化仍是未来发展趋势之一. 相比于有握手机制的 TCP (Transmission Control Protocol) 协议, UDP (User Datagram Protocol) 协议控制

选项少,在数据传输过程中延迟小,数据传输效率高,适合对数据传输实时性要求高的场合.

目前,基于不同的嵌入式平台主要有以下三种常见的网络数据传输方案:

(1)ARM+Linux+网卡,其主频芯片频率高,面对不同领域方便代码移植,但需要自行编写外设和驱动程序,前期设计成本高,后期维护难度大.

(2)DSP+PHY 芯片,相比于(1),虽开发难度有所降低,但是由于 DSP 内部的 CPU 仍是串行调用方式,因此,工作效率不高.

(3)FPGA+PHY 芯片,相比前两种解决方案,其使用纯硬件语言实现功能,不受 CPU 节制,可以实现高效率传输,又因目前成熟的开发环境,大量 IP 核的使用使开发难度大大降低.

文献^[1]采用了第三种方案,利用成熟的 IP 核,设计了千兆以太网传输链路.基于此,本文提出了一种基于 FPGA 的万兆以太网链路的传输方案,设计完成了 UDP/IP 协议,ARP 协议,万兆以太网 MAC 控制器,设计中采用 AXI4_stream 流协议,实现了数据实时高效的传输

2 整体架构

开放式系统互联通信参考模型(Open System Interconnection Reference Model,OSI),也就是著名的 OSI 模型^[2],它定义了网络互联的七层框架,包括应用层、表示层、会话层、传输层、网络层、数据链路层和物理层.

在本设计中,采用精简的五层协议,如图 1 所示,将上三层(应用层,表示层,会话层)合并成一个应用层,保留了下四层,传输层为 UDP 协议,实现了数据的 UDP 封装、解封和 Checksum 校验,网络层包括 IPv4 协议和 ARP 协议,实现目的 IP 地址的灵活配置以及 MAC 地址的寻址,数据链路层则包括 MAC 地址的封装,流量控制和错误检测等功能,物理层则是物理通路的定义与建立.

设计的 UDP/IP 协议可以简单的分成两个工作流程,发送过程和接收过程.

发送过程:应用层来的数据首先会通过传输层,进行 UDP 报头封装,经过网络层时,会进行 IPv4 报头的封装,以及 MAC 地址的获取,数据链路层会形成完整的以太网帧,随后经过 PCS/PMA^[3]子层,实现 64B/66B 编码和加扰操作,以及数据的并/串转换最后交由发送器,通过 SFP+接口将数据发送出去.

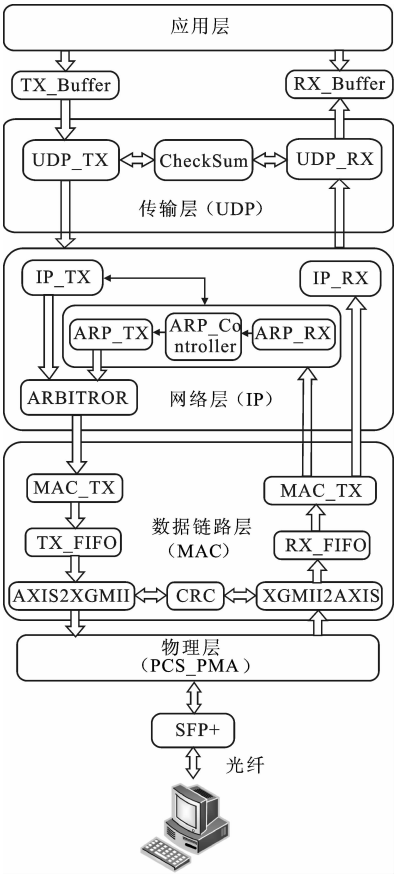


图 1 整体架构图

接收过程:与发送过程恰恰相反,主机通过光纤向从机发送数据,从机通过 SFP+接口接收到来自主机的数据,通过物理层实现数据的串并转换以及解码,而后 MAC 层对其进行以太网包解包以及错误检测,随后经过 IPv4 协议的解包,最后通过 UDP 层的解包到达指定上层应用.

3 各模块功能

3.1 UDP 模块

UDP 报头包括 4 个字段,分别为来源链接端口、目的链接端口、报文长度、校验和,每个字段占用两个字节,其中报文长度为 UDP 报头和数据总共占用的长度,最小长度 8 字节,来源链接端口和校验和是可选的,默认填充 0,来源链接端口和目的链接端口指定的是要发送和接收数据的应用进程.本文的 UDP 模块主要包括校验和产生模块、发送模块和接收模块.

本节主要关注发送过程,UDP 模块在接收应用进程的数据之后,首先会经过 Checksum 校验模块生成校验和,而后会送到发送模块,打包成 UDP 包,发送给 IP 模块.发送模块的状态转移图可以简

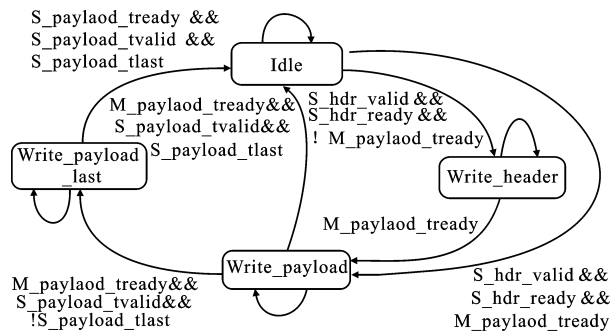


图2 UDP发送状态图

化为图2,整个状态流程可以分为四个状态:idle状态,Wirte_header状态,Write_payload状态,Write_payload_last状态. idle为初始状态,当UDP接收模块准备好接收数据时(S_hdr_ready有效),且从上层接收到的数据(S_hdr_valid)有效时,如果此时下层模块没有准备好接收数据(M_payload_tready信号无效)时进入Wirte_header状态.反之如果已准备好,直接发送报头数据,随后进入Write_payload状态.在Wirte_header状态时,当下层模块准备好接收数据时,发送报头数据,随后进入Write_payload状态.在Write_payload状态中,发送来自应用层的数据和相应的标志信号,当发送计数器达到指定数值时,即数据的长度值时,如果此时标志数据已发送完的结束信号(S_paylaod_tlast)到来时,则发送结束,进入Idle状态,整个发送周期结束,等待下一次发送周期,否则进入Write_payload_last状态,等待结束信号的到来,一旦到来,则发送结束,进入Idle状态.UDP接收模块是发送模块的反过程,与此类似,不过其对应的是解包过程.为提高整个模块的收发速度,模块中插入一级缓存寄存器.

3.2 IP和ARP模块

整个网络层主要三大模块:IP模块,ARP模块,仲裁选择模块.IP模块主要包括两个部分,发送模块和接受模块,实现了对来自UDP层数据的IPv4协议的封装,和对来此MAC层数据的IPv4协议的解封,其封装与解封流程与UDP类似.ARP模块主要包括三大部分,ARP发送模块、ARP接收模块、ARP缓存模块.整个ARP模块实现了目的主机的MAC地址的获取和响应,其它主机的ARP请求.仲裁选择模块实现了IP数据包和ARP数据包的选通发送,使用优先级编码,优先发送ARP数据包.

ARP协议消息格式包含单一的地址请求或响

应.其组成主要包含两个部分,14字节的以太网首部和28字节的ARP请求/应答部分.以太网首部包括目的以太网地址、源以太网地址和以太网类型.当需要请求某主机的MAC地址时,目标以太网地址需要填写为路由地址或广播地址(FF:FF:FF:FF:FF:FF),路由器查询目的主机的MAC地址,或者向网络中所有主机发送请求,等待目的主机的回应.

当数据流从UDP层流入IP层时,整个IP层发送流程可简化为图3形式.

IP层等待来自UDP的数据流,当有数据流到来时,先请求ARP模块查询目的IP的MAC地址,如果ARP缓存中存在该地址,则返回该MAC地址,随后交由IP模块对数据进行IPv4协议封装,如果ARP缓存中不存在指定IP的MAC地址,则ARP模块产生ARP请求,获取指定IP的MAC地址.

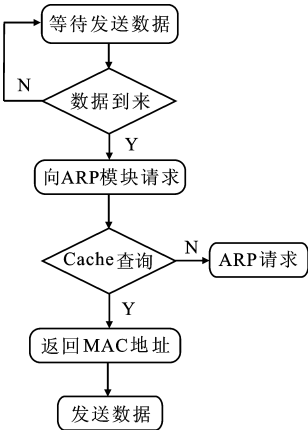


图3 网络层发送流程图

在接收过程中,对来自MAC层的数据解析要比发送模块更加复杂,需要区分协议类型和ARP操作类型,我们将其简化为图4.

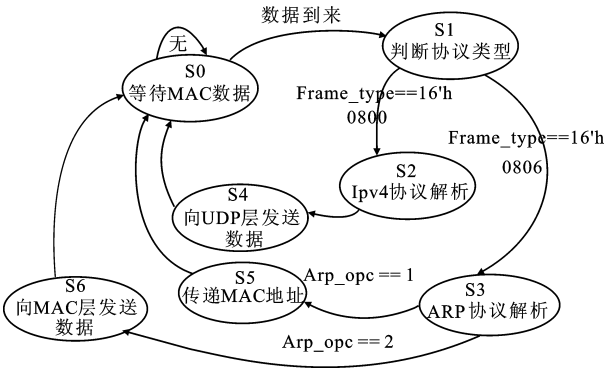


图4 协议解析流程图

当有数据传入时,通过提取数据帧类型进行协议解析,如果 frame_type == 16'h0800 时,则为

IPv4 包,随后进行 IPv4 协议的解析,并将解好包的数据发送到 UDP 层,发送结束之后,随后进入到 S0 状态,等待下一个数据的到来. 如果 frame_type == 16h0806,则为 ARP 包,进入 S3 状态,对 ARP 进行协议解析,首先判断 ARP 操作码类型,如果 Arp_opc==1,则为回复类型,随后将提取出来的 MAC 地址发送给 IPv4 协议发送模块,并将此 MAC 地址存入 ARP 缓存中,方便下次调用. 如果为请求类型,即 Arp_opc == 2 时,则应答请求方,将本机 MAC 回复给请求方,并将请求方的 MAC 地址存入缓存中,回复结束后,进入 S0 状态,等待下一个包的到来.

3.3 MAC 层

MAC 层主要负责数据帧的封装、解封、地址匹配、CRC 校验及实现流量控制等功能. 其帧格式如表 1 所示. MAC 的主要功能可以简单描述为:发送时,接收来自上层的帧时,在数据帧头部加入 7 字节前导码和 1 个字节起始符,6 字节的地址和 6 字节的源地址,以及 2 字节的网络类型,在数据帧尾部插入 4 字节的 CRC 校验码和 12 字节的包间距. 接收时,去掉添加的前导码和起始符,判断目的地址是否是本机 MAC 地址或广播地址,并数据帧尾的 CRC 进行校验,校验正确,则向上层发送 good_frame 信号,否则发送 bad_frame 信号.

表 1 以太帧格式

前导码	7 octets
起始符	1 octet
目的地址	6 octets
源地址	6 octets
以太类型	2 octets
负载	46—1500 octets
CRC	4 octets
帧间距	12 octets

与千兆以太网所不同的是,MAC 层使用 64 bits 的数据位宽和 8 bits 的数据使能信号^[4],另外,万兆以太网使用 XGMII 接口.

分析了 MAC 层帧格式,本文将其封装过程分为两个部分,中间插入异步 FIFO,加快数据传输. 第一部分,实现了对目的地址、源地址和以太类型封装与解封. 第二部分则实现前导码、起始符、CRC 校验码插入.

因为在以太帧的起始或结尾处需要校验的字节数不确定,这 CRC 校验^[5]带来了困难,因此单一模式不再合适. 为了提高速度,本文选择了无延时的

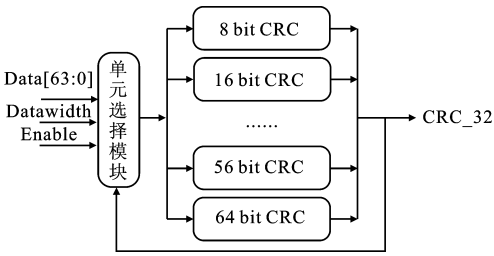


图 5 CRC 校验模块

CRC 校验模块,如图 5 所示.

该方案用较多的逻辑资源去换取高效的速度,通过重复 CRC 逻辑,例化了不同位数的 CRC,并使用选择单元选择相应的 CRC 模块,加快 CRC 校验过程.

3.4 PHY 层

PHY 层,本文采用 Xilinx 的 10GBASE-R IP 软核,该 IP 核实现了 Physical Coding Sublayer (PCS) 和 Physical Medium Attachment (PMA) 功能,符合 IEEE802.3—2012 条款 45、49、72 和 74 标准,其连接本文所设计的 MAC 层和输出的 10G PHY 接口. 10G BASE-R 支持两种数宽,64 bits XGMII 和 32 bits XGMII,时钟频率分别为 156.25 MHz 和 312.5 MHz. 本文中选用的是 64 bits 数据位宽,采用 SFP+ 光纤接口发送数据^[6]. 其发送部分主要包括 64B/66B 编码器,扰码器和变速器. 其接收主要包括 64B/66B 解码器,解扰码器,同步块,和误码率检测器. 数据的发送和接收通过一个 SERDES 模块实现.

4 系统测试

本文采用 Xilinx 的 Kintex UltraScale 系列的 xcku040-ffva1156-2-e 器件进行设计实现,使用 vivado 2018.3 工具进行综合、映射和布局布线. 使用杭州某公司的 MC02 开发板进行下载测试. 综合后的资源报告和时序结果如图 6 所示,结果表明,本文的设计远远满足 156.25 MHz(6.4 ns)时序约束的要求.

Setup	Hold
Worst Negative Slack (WNS): 2.420 ns	Worst Hold Slack (WHS): 0.031 ns
Total Negative Slack (TNS): 0.000 ns	Total Hold Slack (THS): 0.000 ns
Number of Failing Endpoints: 0	Number of Failing Endpoints: 0
Total Number of Endpoints: 17032	Total Number of Endpoints: 17014

(a) 时序报告

Resource	Utilization	Available	Utilization %
LUT	6332	242400	2.61
LUTRAM	194	112800	0.17
FF	7440	484800	1.53
BRAM	11.50	600	1.92
IO	4	520	0.77
GT	1	20	5.00
BUFG	6	480	1.25
MMCM	1	10	10.00

(b)资源报告

图 6 综合报告

在进行下板测试时,为了更好的测试链路功能,本文设计了如下测试方案,如图 7 所示。

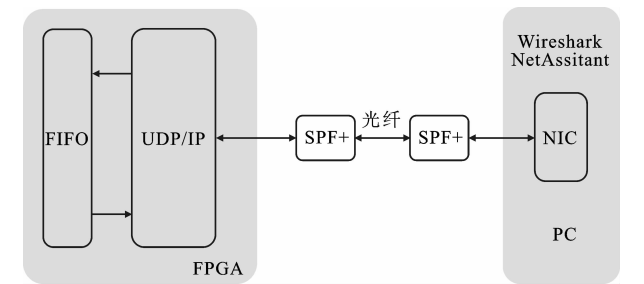


图 7 测试环路

在该测试环路中,PC 端使用 NetAssistant 软件对测试模块发送数据,测试模块接收到数据之后,将数据存入 FIFO 中,随后从 FIFO 中取出数据发回 PC。此时,我们通过 WireShark 软件就可以查看完整的通信过程。结果如图 8 所示。

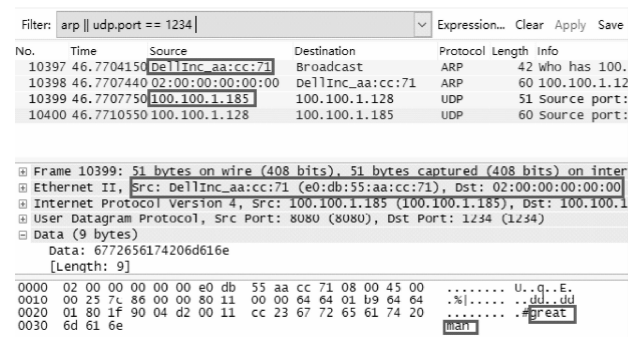


图 8 系统功能测试

从图 8 可以看出,DELL_aa:cc:71 先发送了广播地址,查询 IP 为 100. 100. 1. 128 的 MAC 地址,随后测试模块,其地址为 02:00:00:00:00:00,发送 ARP 回复,响应主机,发送本机 MAC 地址。紧接着,主机向测试模块发送了一串字符,即为图中的“great man”,随后测试模块将这串字符回送回来,

完成了整个链路的测试。通过 LOOPBACK 环路测试测试本模块性能,即自发自收测试,通过计数器统计所收到的帧数,测试结果表明,传输速度可高达 10 Gb/s。

5 结束语

针对日益增长的数据量,本文设计了一种完备的万兆以太网传输链路。使用硬件实现了 UDP/IP 协议和 MAC 层的设计,并加入了 ARP 模块,解决了 IP 地址与 MAC 地址的映射问题,使得网络数据传输更加便利和灵活,达到了大规模系统级联的目的。通过使用重复了 CRC 校验模块,加速了 CRC 校验过程,实现了校验过程零延时,加快了数据的传输速度。最后通过 FPGA 验证,实现了完整功能。因为 FPGA 的灵活性,使的本设计灵活通用,能够应用于多种嵌入式场景,尤其是对于实时性要求很高的场景具有很高的应用价值。

参考文献:

[1] 崔鹤,刘云清,盛家进. 基于FPGA的UDP/IP协议栈的研究与实现[J]. 长春理工大学学报(自然科学版), 2014(2):133-137.

[2] ABHISHEK SAHU, PRAJYANT PATHAK. A Novel Approach for Implementing of UDP/IP Stack in FPGA[J], ijmern. 2017 ,5(1):24-27.

[3] 杨钊翔,孔宪伟,栾文焕,等. 基于以太网 10Gbase-R 协议的 PCS 层设计[J]. 微电子学与计算机. 2019,36 (2):16-20.

[4] 赵柏山,王禹衡,刘佳琪. 用于视频传输的 10G 网络接口设计[J]. 微处理机,2018,39(3):28-32

[5] 曹政,李磊,陈明宇. 万兆以太网媒体访问控制层研究 [J]. 计算机工程,2007(17):31-33.

[6] 徐光辉,聂永军,冯秀妍. 基于FPGA的万兆以太网传输设备设计[J]. 光通信技术,2015,39(11):4-6.

作者简介:

孔德伟 男,(1995-),硕士研究生,研究方向为 SOC 设计。
E-mail:kongdewei@ime. ac. cn.

袁国顺 男,(1966-),研究员,博士生导师,研究方向为数模混合集成电路设计。

刘小强 男,(1990-),博士研究生,研究方向为图像处理和 SOC 设计