

引用格式:杨钞翔,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. [YANG C X, KONG X W, LUAN W H, et al. Design of the physical coding sublayer based on the ethernet 10Gbase-R protocol [J]. Microelectronics & Computer,2019,36(2):16-20.]

基于以太网 10Gbase-R 协议的 PCS 层设计

杨钞翔¹,孔宪伟²,栾文焕¹,张远航¹,王志军¹,张 春¹,王自强¹

(1 清华大学 微电子学研究所,北京 100084;2 中国电子技术标准化研究院
集成电路测试与评价工业和信息化部重点实验室,北京 100176)

摘 要: 随着网络技术的飞速发展,以太网技术凭借其优良特性已经成为当今使用最广泛的局域网技术,以太网标准包含多种协议构成完整体系,10Gbase-R 协议作为其中重要的基础协议得到重视. 本文基于 10Gbase-R 协议的 PCS 层,深入探究 PCS 层内部关键模块:弹性缓冲器、64B/66B 编解码器、加解扰码器、位宽转换器等,提出完整设计架构,使用 verilog 语言完成其 RTL 级模块设计,搭建仿真验证平台,使用 Modelsim 软件进行功能验证,完成了相关设计.

关键词: 以太网;10Gbase-R 协议;PCS 层;设计

中图分类号: TN492

文献标识码: A

文章编号: 1000-7180(2019)02-0016-05

Design of the Physical Coding Sublayer Based on the Ethernet 10Gbase-R Protocol

YANG Chao-xiang, KONG Xian-wei, LUAN Wen-huan, ZHANG Yuan-hang,
WANG Zhi-jun, ZHANG Chun, WANG Zi-qiang

(1 Institute of Microelectronics, Tsinghua University, Beijing 100084, China; 2 Key Laboratory of Integrated Circuits Testing and Evaluation, Ministry of Industry and Information Technology, China Electronics Standardization Institute, Beijing 100176, China)

Abstract: With the rapid development of network technology, Ethernet technology has become the most widely used local area network technology with its excellent features. The Ethernet standard consists of multiple protocols to form a complete system. The 10Gbase-R protocol is one of the important basic protocol. This article is based on the PCS layer of the 10Gbase-R protocol, and deeply explores the key modules within the PCS layer: elastic buffers, 64B/66B codecs, scrambler/descrambler, bit-width converters, etc. It proposes a complete design architecture, and uses the verilog language to complete the RTL-level module design, then builds simulation verification platform and uses modelsim software to complete the functional verification, and completes related design.

Key words: ethernet; 10Gbase-R; PCS layer; design

1 引言

以太网技术是目前使用最广泛的局域网技术之一,发展应用 30 多年,从最初的半双工发展成全双工的工作方式,经历了 10 Mbps 初期以太网,100

Mbps 快速以太网,1000 Mbps 千兆以太网,10 Gbps 万兆以太网的发展历程,现在还提出了下一代的 40 Gbps、100 Gbps 高速以太网. 以太网标准包含多种协议,构成一个完整体系的协议族,现如今在局域网领域,以太网技术早已取代诸如令牌环、令牌总

线、FDDI、ARCNET 等技术,成为了局域网领域最主流的技术^[1]。

以太网技术具有多种优良特性;现如今不仅在局域网领域占据了重要地位,已经扩展到城域网、广域网等其他领域中。

2 10Gbase-R 协议的 PCS 层

2.1 以太网体系

以太网技术最重要的一个领域是 10G 万兆以太网领域,和传统以太网相比具有多种特性:支持星型结构网络,采用点对点方式,发展出了新的 PMD 物理媒介相关子层,数据速率高达 10G,保留了最大/最小帧长,全双工的工作模式。到现在已经发展成为一个庞大的协议族^[2],按照传输介质可以分为:

(1)光纤上传输的 IEEE 802. 3ae,包含 7 个规范,10GBase-LX4, 10GBase-W (10GBase-SW; 10GBase-LW;10GBase-EW),10GBase-R(10GBase-SR;10GBase-LR;10GBase-ER)。

(2)同轴电缆上传输的 IEEE 802. 3ak,包含 1 个规范,10GBase-CX4。

(3)双绞线上传输的 IEEE 802. 3an,包含 1 个规范,10GBase-T。

(4)多模光纤上传输的 IEEE 802. 3aq,包含 1 个规范,10GBase-LRM。

(5)PCB 板上传输的 IEEE 802. 3ap,包含 2 个规范,10GBase-KX4 和 10GBase-KR。

以上规范按用途又可分为 3 类:

(1)局域网:10GBase-SR; 10GBase-LR; 10GBase-ER; 10GBase-LX4; 10GBase-CX4; 10GBase-T;10GBase-LRM。

(2)广域网:10GBase-EW; 10GBase-LW; 10GBase-SW。

(3)背板:10GBase-KX4;10GBase-KR。

其中,广域网物理层主要是在局域网物理层原有基础上增加了 WIS 子层,用来完成兼容性和速率适配。可以看出,10Gbase-R 协议既代表局域网物理层,也是广域网物理层的基础,属于以太网的关键技术,本文基于 10Gbase-R 协议研究其 PCS 层,完成相关设计。

2.2 PCS 层

如图 1 所示的 OSI 系统结构中,10Gbase-R 协议的 PCS 层属于物理层,位于 MAC 层和 PMA 层之间。在发送方向上,上层数据通过 MAC 层传给 PCS 层,再传给 PMA 层;在接收方向上,底层数据

通过 PMA 层传给 PCS 层,再传给 MAC 层。MAC 层被称为介质访问控制层,作为协议的控制引擎,与上层进行交互。PMA 层被称为物理介质接入层,实际为串化/解串模块,主要负责数据的串行和并行转换。PMA 层在发送路径上,提供发送时钟,将 PCS 层传来的并行数据,转换成串行数据,再传递给 PMD 层;PMA 层在接收路径上,提供接收时钟,传递状态链路信息,将 PMD 层传来的串行数据,进行数据解串,再转换成并行数据传给 PMA 层。

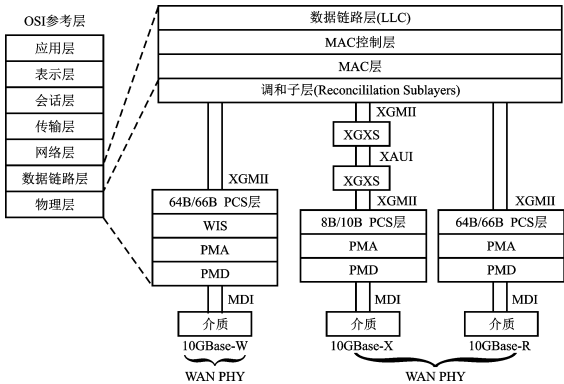


图 1 OSI 系统结构图

深入到 PCS 层内部,具体结构如图 2 所示,内部结构可以分为发送通路和接收通路,分别记为 TX 通路和 RX 通路。PCS 层和 MAC 层通过 XGMII 接口连接,接口的数据位宽为 64 bit,PCS 层和 PMA 层接口的数据位宽为 32 bit。TX 通路从上到下依次经过 TX 弹性缓冲器、64B/66B 编码器、扰码器、变速箱等模块,RX 通路从下到上依次经过数据同步模块、解扰器、64B/66B 解码器、RX 弹性缓冲器等模块。下面将对发送通路和接收通路包含的各个关键模块进行解析和设计。

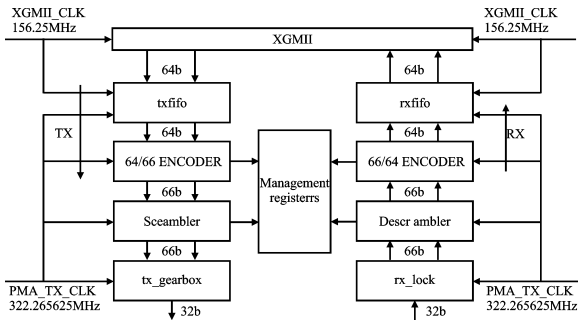


图 2 10Gbase-R 协议 PCS 层内部结构图

3 设计实现

PCS 层内部发送和接收两条通路的各个模块之间实际为互逆的模块组,如图 2 所示,txfifo 和

rxfifo 分别是发送和接收通路上互逆的弹性缓冲器, 64B/66B Encoder 和 64B/66B Decoder 分别是发送和接收通路上互逆的 64B/66B 编码器和解码器, Scrambler 和 Descrambler 分别是发送和接收通路上互逆的加扰器和解扰器, tx_gearbox 和 rx_lock 分别是发送和接收通路上互逆的位宽转换器。

3.1 弹性缓冲器

PCS 子层内涉及两个时钟域, MAC 子层提供的 XGMII_CLK 时钟域和 PMA 子层提供的 PMA_CLK 时钟域, 由于数据率为 10 Gbps, MAC 层和 PCS 层接口数据位宽为 64 bit, 所以 XGMII_CLK 时钟的频率设计为 $10 \text{ Gbps}/64 \text{ bit} = 156.25 \text{ MHz}$ 。PCS 层采用了高效率的 64B/66B 编码方式, 10 Gbps 的数据率变为 $10 \text{ Gbps} \times (66/64) = 10.3125 \text{ Gbps}$, 又由于 PCS 子层和 PMA 子层的数据位宽为 32 bit, 则 PMA_CLK 时钟的频率设计为 $10.3125 \text{ Gbps}/32 \text{ bit} = 322.265625 \text{ MHz}$ 。因此数据传输时需要进行跨时钟域操作, 通常采用异步 FIFO 完成跨时钟域, 但是考虑到数据率较高即使两时钟频率偏差很小, 随着时间的积累 FIFO 迟早会被写满或读空, 所以本设计将传统 FIFO 进行改进设计出一种具有弹性的缓冲器, 借助数据中的 IDLE 字符, 通过比较缓冲器两端的读写速率, 正确地插入/删除 IDLE 字符, 调节缓冲器读写速率, 进行了时钟频率补偿, 成功地实现了跨时钟域操作。

弹性缓冲器对应图 2 中 txfifo 和 rxfifo 模块, 设计为深度 16, 位宽 72 位, 以 TX 通路为例, 具体设计结构如图 3 所示, 包含双口存储器模块、读/写地址产生逻辑模块、二进制码转格雷码模块、跨时钟域模块、格雷码转二进制码模块、读写地址差值计算模块^[3]。图中读写数据都为 72 bit, 其中包括 64 bit 纯数据和 8 bit 控制码。双口存储器采用两写一读方式来实现, 两个写地址为 w1 和 w2, 一个读地址为 r1。写入数据受 XGMII_CLK 时钟控制, 读出数据受 PMA_CLK 时钟控制, 读地址 r1 先经过二进制转格雷码变换, 再从 PMA_CLK 时钟域跨到 XGMII_CLK 时钟域, 再将格雷码转回二进制码, 再与写地址作差值计算。比较差值判断读写速率快慢: 当写速率和读速率正常匹配时, 不进行附加操作; 当写速率快于读速率时, 写入端删除数据中的 IDLE 字符, 降低写速率; 当写速率慢于读速率时, 写入端往数据中插入 IDLE 字符, 提高写速率。当读写地址指针差值接近 0 或者 16 时, 指示信号 tx_overflow 和 tx_underflow 置 1, 分别表示写满和读空, 并进行复位操

作。RX 通路采用相同的方法, 但读写方式改为一写两读。如此调节读写速率偏差, 使弹性缓冲器维持在半满最佳状态, 实现频率补偿, 完成了高速数据的跨时钟域操作^[4]。

其中需要注意的是 IDLE 字符, 表示空状态, 存在于数据流中, 上层协议在解析数据时可以自动识别 IDLE 字符并将其滤掉, 因此可通过插入/删除 IDLE 字符, 进行时钟补偿, 保证数据的正确传输。

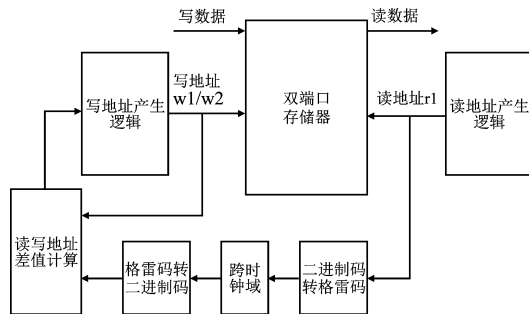


图 3 TX 弹性缓冲器结构图

3.2 64B/66B 编解码器

本设计采用 64B/66B 的编解码方式, 常用的实现方法有两种: 逻辑运算法和查找表法。逻辑运算法主要用于编码规则具有一定规律, 编码种类数目较少的情况。对于 64B/66B 编码, 由于编码较长、情况种类较多, 也不存在特定的编码映射规律, 所以此处采用查找表法来实现, 操作简单且效率较高, 编码器和解码器分别对应图 2 中 64/66 Encoder 和 66/64 Decoder。

64B/66B 编码属于 TX 通路, 对 64 bit 数据和 8 bit 控制码进行编码。控制码每 1 bit 对应数据 8 bit, 当控制码对应位为 0, 表示纯数据, 当控制码对应位为 1, 表示控制字。所以 64B/66B 编码根据这两种不同的情况分开进行, 具体参照图 4 编码规则进行。首先对控制码进行检测, 如果 8 bit 控制码全为 0, 表示 64 bit 数据是纯数据, 用 D0~D7 表示, 则加上 2'b01 同步头, 拼凑即可得到 66 bit 编码后的数据。如果 8 bit 控制码不全为 0, 表示 64 bit 数据是数据和控制字符的组合或者全为控制字符, 数据字符用 D0~D6 表示, 控制字符用 C0~C7、O、S、T 表示, 此时先加上 2'b10 同步头, 再根据数据类型添加区域块类型码, 再按照如图 4 对应情况完成编码。

64B/66B 解码属于 RX 通路, 实际上为 64B/66B 编码器的逆运算, 实现方式和基本原理相同, 也采用查找表法, 将 66 bit 数据解码成 64 bit 原始数据, 并译出 8 bit 控制码。关键是利用好 66 bit 数据

中的同步头字段和区域块类型码,首先检测 2 bit 同步头字段,如果同步头字段为 2b'01,则说明 66 bit 数据为纯数据,直接去掉同步头剩下的 64 bit 数据即为解码输出的数据,并且将 8 bit 控制码全部译为 0,表示纯数据;如果同步头为 2b'10,说明数据含有控制字符,则根据图 4 确定 66 bit 数据中包含的区域块类型码的种类,按照编码的逆运算进行解码,最终得到 64 bit 的解码输出数据,并且将 8 bit 控制码对应的位译为 1.

Input Data	Syn C	Block Payload
Bit Position: 0 1 2		65
Data Block Format:		D ₀ D ₁ D ₂ D ₃ D ₄ D ₅ D ₆ D ₇
Control Block Formats:		Block Type
C ₀ C ₁ C ₂ C ₃ C ₄ C ₅ C ₆ C ₇	0x10	C ₀ C ₁ C ₂ C ₃ C ₄ C ₅ C ₆ C ₇
C ₀ C ₁ C ₂ C ₃ D ₀ D ₁ D ₂ D ₃	0x2d	C ₀ C ₁ C ₂ C ₃ D ₀ D ₁ D ₂ D ₃
C ₀ C ₁ C ₂ C ₃ D ₀ D ₁ D ₂ D ₃	0x33	C ₀ C ₁ C ₂ C ₃ D ₀ D ₁ D ₂ D ₃
D ₀ D ₁ D ₂ D ₃ D ₄ D ₅ D ₆ D ₇	0x66	D ₀ D ₁ D ₂ D ₃ D ₄ D ₅ D ₆ D ₇
D ₀ D ₁ D ₂ D ₃ D ₄ D ₅ D ₆ D ₇	0x55	D ₀ D ₁ D ₂ D ₃ D ₄ D ₅ D ₆ D ₇
S ₀ D ₁ D ₂ D ₃ D ₄ D ₅ D ₆ D ₇	0x78	D ₀ D ₁ D ₂ D ₃ D ₄ D ₅ D ₆ D ₇
D ₀ D ₁ D ₂ D ₃ C ₄ C ₅ C ₆ C ₇	0x4b	D ₀ D ₁ D ₂ D ₃ C ₄ C ₅ C ₆ C ₇
T ₀ C ₁ C ₂ C ₃ C ₄ C ₅ C ₆ C ₇	0x87	C ₀ C ₁ C ₂ C ₃ C ₄ C ₅ C ₆ C ₇
D ₀ T ₁ C ₂ C ₃ C ₄ C ₅ C ₆ C ₇	0x99	D ₀ C ₁ C ₂ C ₃ C ₄ C ₅ C ₆ C ₇
D ₀ D ₁ T ₂ C ₃ C ₄ C ₅ C ₆ C ₇	0xaa	D ₀ D ₁ C ₂ C ₃ C ₄ C ₅ C ₆ C ₇
D ₀ D ₁ D ₂ T ₃ C ₄ C ₅ C ₆ C ₇	0xb4	D ₀ D ₁ D ₂ C ₃ C ₄ C ₅ C ₆ C ₇
D ₀ D ₁ D ₂ D ₃ T ₄ C ₅ C ₆ C ₇	0xcc	D ₀ D ₁ D ₂ D ₃ C ₄ C ₅ C ₆ C ₇
D ₀ D ₁ D ₂ D ₃ D ₄ T ₅ C ₆ C ₇	0xd2	D ₀ D ₁ D ₂ D ₃ D ₄ C ₅ C ₆ C ₇
D ₀ D ₁ D ₂ D ₃ D ₄ D ₅ T ₆ C ₇	0xe1	D ₀ D ₁ D ₂ D ₃ D ₄ D ₅ C ₆ C ₇
D ₀ D ₁ D ₂ D ₃ D ₄ D ₅ D ₆ T ₇	0xf7	D ₀ D ₁ D ₂ D ₃ D ₄ D ₅ D ₆ C ₇

图 4 64B/66B 编码规则

3.3 扰码器/解扰器

64B/66B 编码器编码效率高,但是存在长连 0、长连 1 的问题,不能进行直流平衡,不利于采样时钟恢复. 所以还要对编码后的数据增设扰码操作,扰码可以降低数据直流分量,有效平衡数据流中 0 和 1 的数量,使数据尽量接近伪随机序列,具体对应于图 2 的 Scrambler 和 Descrambler 模块.

根据协议要求,编码器输出的 66 bit 数据,其中 2 bit 同步头不进行扰码,只将后 64 bit 数据输入扰码器进行操作,扰码多项式为:

$$G(x) = 1 + x^{39} + x^{58} \tag{1}$$

扰码规则如图 5 所示,扰码电路由 58 个 D 触发器和 2 个异或门搭建而成,D 触发器序列编号为 S0~S57,每个 D 触发器输出的结果用 X1~X58 表示,先将 S38 和 S57 寄存器当前输出的两个结果 X38 和 X58 进行异或得到 X0,再将 X0 和输入的待扰码数据进行第二次异或,一方面得到扰码输出结果,另一方面将此结果送回 S0 寄存器,重新回到寄存器序列中不断补充新数据. 因此扰码输出不仅由当前输入决定,还和上一次寄存器序列的状态有关. 数据位宽 64 位,数据率要达到 10 Gbps,串行扰码方式难以满足要求,所以本文采用并行扰码方式来

等效实现,满足了高数据率的要求. 解扰器为扰码器的逆向操作,例如扰码操作为: $A \oplus B = A'$,则解扰操作为: $A' \oplus B = A$. 解扰器也是根据公式(1)来工作,采用并行解扰的方式来等效实现,满足高数据率的解扰需求^[5].

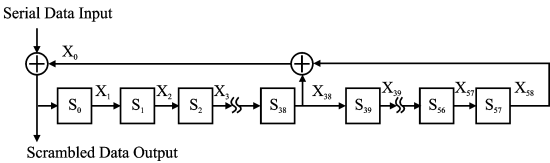


图 5 扰码规则

3.4 位宽转换器

PCS 子层和 PMA 子层进行数据传输时需要进行 66B/32B 的位宽变换,具体对应到图 2 的 tx_gearbox 和 rx_lock 模块.

TX 通路的 tx_gearbox 为变速箱,主要将扰码器传来的 66 bit 数据转换成 32 bit 数据,再传输给 PMA 子层. 具体采用的是最小公倍数法则和分数分频法,操作如图 6 所示, $66b \times 16 = 32b \times 33 = 1056b$,将 66 bit 数据进行缓存,每 16 次 66 bit 的写入正好和 33 次 32 bit 的读出匹配. 通过分数分频构造出写入频率为读出频率的 16/33,一个完整操作周期正好处理 1 056 bit 数据,然后以此周期循环,最终实现 66 bit 到 32 bit 的位宽转换.

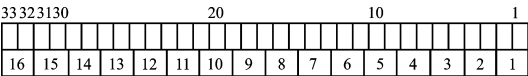


图 6 位宽转换示意图

RX 通路的 rx_lock 模块需要将 PMA 子层传来的 32 bit 数据进行位宽变换,转换成 66 bit 后传给解扰器,实现原理和 tx_gearbox 类似,将 32 bit 数据进行缓存,每 33 次 32 bit 的数据写入正好和 16 次 66 bit 的数据读出匹配. 通过分数分频构造出读出频率为写入频率的 16/33,一个完整操作周期正好可以处理 1 056 bit 数据,以此实现 32 bit 到 66 bit 数据的位宽转换. 但是除此之外,rx_lock 还需要实现同步对齐功能. 66 bit 数据含有一个 2 bit 同步头字段,只能是 2b'01 或 2b'10,2b'01 表示纯数据,2b'10 表示数据和控制字符的组合,此同步头每隔 64bit 便会出现在一次,利用此特性对数据进行偏移校正,检测 66 位数据的最高两位,如果不是 2b'01 或者 2b'10,说明数据没有对齐,则将数据移动一位继续检查,最多移动 65 次就能找到同步头字段,以此完成数据的同步对齐.

4 仿真验证及结果分析

以上对 10Gbase-R 协议 PCS 层各关键模块进行了解析,然后使用 verilog 语言完成 RTL 级代码设计,搭建数据发送通路和接收通路,设计测试激励,采用回环测试法,将发送通路处理完的数据回环到接收通路,构造完整测试路径,由于发送通路和接收通路实际上为两个操作互逆的通路,因此可以通过对比接收到的数据和发出的数据是否相同,如果完全相同则可以证明各模块功能正确,如果不同,说明数据在传输过程中出错,则分析数据路径并对出错地方进行修改,最终借助 Modelsim 仿真工具完成 PCS 层的功能仿真,对设计进行验证。

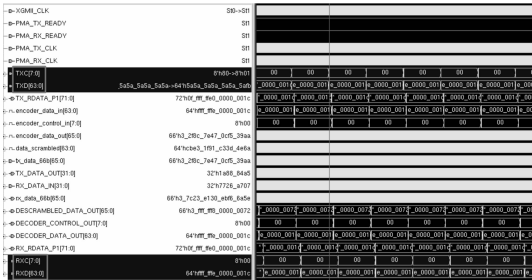


图 7 仿真结果图

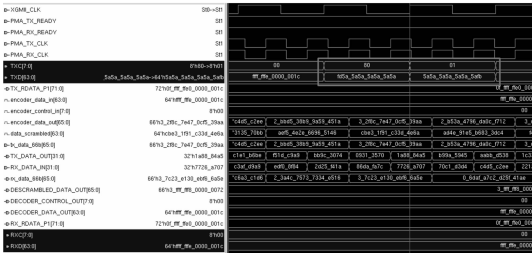


图 8 仿真结果局部放大图

功能仿真结果如图 7、图 8 所示,图 7 方框标注的信号 TXD[63:0]和 TXC[7:0]分别是 XGMII 接口输入到 PCS 层发送通路上的 64 bit 数据和对应的 8 bit 控制码,信号 RXD[63:0]和 RXC[7:0]分别是 PCS 层接收通路上输出到 XGMII 接口的 64 bit 数据和对应的 8 bit 控制码,回环测试的仿真结果表明,数据经过所有模块处理后,由局部放大图 8 中的方框标注可以得出,所有数据包的起始都为 TXC[7:0]=8b'0000_0001, TXD[63:0]=64h'5a5a5a5a_5a5a5afb,其中 8h'fb 是帧头 /S/ 控制字符;所有数据包的结尾都为 TXC[7:0]=8b'1000_0000, TXD[63:0]=64 h'fd5a5a5a_5a5a5a5a,其中 8h'fd 是帧尾/T/控制字符;中间数据对应的 TXC[7:0]=8b'0000_0000,表示纯数据. 因此发送和接

收到的数据包格式都符合协议要求,且接收到的数据和发出的数据完全相同,仿真结果正确,表明设计无误,完成了 10 Gbase-R 协议 PCS 层的功能仿真验证工作。

5 结束语

本文基于万兆以太网 10Gbase-R 协议的 PCS 子层,从发送和接收通路分别深入探究 PCS 子层内部关键模块:弹性缓冲器、64B/66B 编解码器、加解扰码器、位宽转换器等,提出完整设计架构,使用 verilog 语言完成其 RTL 级模块设计,搭建仿真验证平台,使用 Modelsim 软件进行功能验证,完成了相关设计。

参考文献:

[1] DAmbrosia J. Defining the next generation of ethernet [J]. Communications Magazine IEEE, 2008, 45(11): 38-38.

[2] Law D, Dove D, D'Ambrosia J, et al. Evolution of ethernet standards in the IEEE 802.3 working group [J]. Communications Magazine IEEE, 2013, 51(8): 88-96.

[3] Verilog E, Cummings C E, Simulation and Synthesis Techniques for Asynchronous FIFO Design[J]. Snug, 2002, 2(3):1-18.

[4] Michelogiannakis G, Balfour J, Dally W J. Elastic-buffer flow control for on-chip networks[C]// IEEE, International Symposium on High PERFORMANCE Computer Architecture. IEEE, 2009:151-162.

[5] Lin C H, Chen C N, Wang Y J, et al. Parallel scrambler for high-speed applications[J]. IEEE Transactions on Circuits & Systems II Express Briefs, 2006, 53(7):558-562.

作者简介:

杨鈔翔 男,(1991-),硕士研究生.研究方向为集成电路设计. E-mail:1159580239@qq.com.

孔宪伟 男,(1986-),硕士,工程师.研究方向集成电路测试.

王志军 男,(1982-),硕士,工程师.研究方向为数字集成电路设计.

张春 男,(1972-),博士,副研究员.研究方向为数模混合信号集成电路设计、嵌入式微处理器设计.

王自强 男,(1975-),博士,副研究员.研究方向为模拟集成电路设计.