新一代以太网的结构模型和并行传输设计

敖志刚，解文彬，胡 琨，唐长春，张康益

(解放军理工大学工程兵工程学院 南京 210007)

【摘要】新一代以太网通常是指速度超过10 Gb/s的以太网，为开展和推动新一代以太网更深入的研究，介绍了国内外的研究进展，利用结构化功能分层方法构建了40 GbE和100 GbE在不同传输模式下的层次结构模型和功能模型，并对各参量进行了分析；研究了物理层各子层与接口之间的关系，以及传输模式与结构模型中各要素之间的关系；提出了并行传输的设计思路、技术方案和典型的实现。为研发新一代以太网产品提供了技术支持。

关 键 词 通信与网络; 新一代以太网; 并行传输; 结构模型

中图分类号 TP39 文献标志码 A doi:10.3969/j.issn.1001-0548.2013.05.025

Structure Model and Parallel Transfers Design for New generation Ethernet

AO Zhi-gang, XIE Wen-bin, HU Kun, TANG Chang-chun, and ZHANG Kang-yi

(Engineering Institute of Corps of Engineers，PLA University of Science and Technology Nanjing 210007)

Abstract The new generation Ethernet commonly refers to the Ethernet for speed to exceed 10 Gb/s. This paper first introduces research progresses in the field at home and abroad, and then constructs the layer structures model and function model of 40 and 100 Gigabit Ethernet in different transfer patterns. The parameters of the model are analyzed, including the relation between each physics sublayer and interface and the relation between different transfer patterns and every factor in the structure model. The design thought, technical schemes, and typical realization about parallel transfers are put forward, which provide technical support to the research and development of new generation Ethernet.

Key words communication and network; new generation Ethernet; parallel transfers; structure model

目前市场上的网络产品(如网卡、网桥、集线器、交换机、路由器和网关等)几乎都是以太网的序列产品。过去10年，以太网经历了从百兆向千兆以太网(gigabit Ethernet，GbE)及万兆以太网(10 GbE)的过渡，所谓新一代以太网是指速度超过10 GbE的以太网。IEEE已经明确新一代以太网速度为40 Gb/s和100 Gb/s，并有相关标准出台[1]，许多学者将这两种速度的以太网综合起来一块研究[2-4]。开发新一代以太网是面对不断增长的带宽密集应用为网络带来的带宽压力做出的激烈响应。40/100 GbE将为新一波更高速的以太网服务器连通性和核心交换产品铺平发展之路。它解决了数据中心、运营商网络和其他流量密集高性能计算环境中数量越来越多的应用的宽带需求，而数据中心内部虚拟化和虚拟机数量的繁衍，以及融合网络业务、视频点播和社交网络等的需求也是推动制订该标准的幕后力量。通过更快速的40/100 Gb/s管道，它还有望推动10 GbE的普及，可以提供更多的10 Gb/s链路汇聚；还有望降低运营支出，通过减少多个10 Gb/s链路汇聚以实现40 Gb/s和100 Gb/s速率的需求，从而改善能源效率。为应对服务器速率急剧增长而引发的汇聚链路与回程链路资源的过度占用，国际电联电信标准化部门(ITU-T)、IEEE和光因特网论坛(OIF)3个组织开展了40/100 GbE的相关标准化工作。ITU-T主要从光传送网的角度对40/100 GbE技术进行了规范；而IEEE主要从业务接口的角度规范了40/100 GbE的接口参数；OIF则主要关注100 GbE长途传输线路接口以及相关电接口的规范。

1 新一代以太网国际标准的制定与研究进展

各主要国际标准化组织在40/100 GbE的标准化

收稿日期：2011－11－15; 修回日期：2012－08－31

基金项目：江苏省基础研究项目(BK2009061)

作者简介：敖志刚(1959－)，男，教授，主要从事网络前沿技术方面的研究.

工作的进展如图1所示。

国外学者在该领域开展的实验和理论研究[2-10]，取得了一定的研究进展和技术突破。“100 GbE光以太网关键技术与传输实验系统”于2008年列入我国863研究计划。但是，国内40 GbE和100 GbE的研究才刚刚起步[11-12]，与国外存在较大差距。主要体现在：1) 缺乏新一代以太网系统结构研究与设计；2) 缺乏网络核心技术和创新活力；3) 缺乏新一代以太网的理论探索和技术攻关。

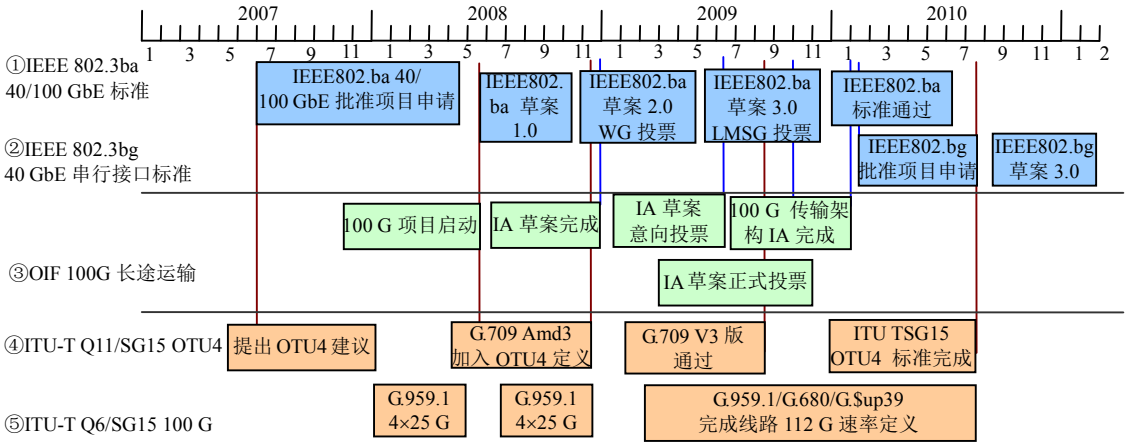

图1 国际标准化组织40/100 GbE标准工作进展

2 新一代以太网传输模式的命名

40/100 GbE有多种传输模式，其命名的表达式为40/100 GBase-*abc*，其中字母*a*、*b*、*c*分别表示40/100 GbE的媒介类型(传输距离)、物理层编码方案和波长(通路)复用数，如图2所示。

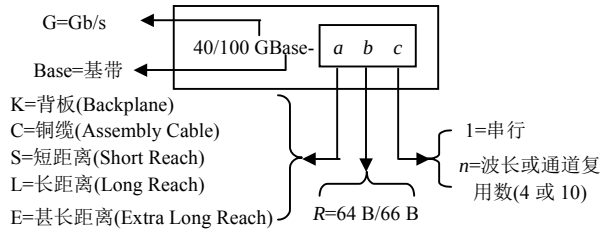

图2 40/100 GbE以太网传输模式的命名

图中当字母“*a*”为K、C、S、L和E时，对应物理媒介分别为背板、铜缆、短距离光纤，长距离光纤、甚长距离光纤。当字母“*b*”为R时，表示64 B/66 B的编码方案，40 GbE和100 GbE只有64 B/66 B这种编码方案。当字母“*c*”为1和*n*时，分别表示单个波长/通路(串行发送方案)和*n*个波长/通路复用方案；单个波长/通路时，通常在命名中省略最后的“1”，一般情况下*n*=4或10。

3 新一代以太网的结构模型

40/100 GbE的体系构架中都保留了以前各种速率以太网的物理编码子层(physical coding sublayer，PCS)、物理介质附属(physical media attachment，PMA)子层、物理介质相关(physical media dependent，PMD)子层、调和子层(relation sublayer，RS)、逻辑链路控制(logical link control，LLC)、介质访问控制(medium access control，MAC)和介质相关接口(media dependent interface，MDI)。为了满足新一代以太网的要求，其体系结构比以前的以太网增加了新的内容，这其中包括了针对40 GBase- KR4、40 GBase-CR4和100 GBase-CR10规范，在物理层设备(physical layer device，PHY)中增加了可选的前向差错校正(forward error correction，FEC)子层、自协商(auto negotiation，AN)子层；MAC、PHY间的片内总线使用4万兆位介质无关接口(40 gigabit media independent interface，XLGMII)和10万兆位介质无关接口(100 gigabit media independent interface，CGMII)，PHY间总线使用4万兆位附属单元接口(40 gigabit attachment unit interface，XLAUI)和10万兆位附属单元接口(100 gigabit attachment unit interface，CAUI)。新一代以太网支持全双工操作，保留使用IEEE 802.3 MAC帧格式和长度规范，所以标准在帧格式、服务、管理属性方面进行扩展已与先前的速率保持一贯性。此外，其他各子层都有相应的改动。

根据新一代以太网的功能、特点和要求，利用结构化功能分层方法，综合参考文献[1]的分层架构，可得如图3所示新一代以太网的结构模型。

图3中的调和子层RS将串行的比特流转换为可用于并行分发的串行码块即64B。PCS主要完成64B/66B编解码、码块的分发重组、控制信令的传送、速率变换，以及提供传送时钟，将多个通道绑定在一起。PMA完成传输、接收、碰撞检测、时钟恢复、队列扶正、复用/解复用，将L路虚通道数据按位复用形成M路的CAUI(XLAUI)接口通道数据，再将M路的CAUI(XLAUI)接口数据转换为N路通道接到PMD层。PMD子层用作对编码数据的转换，即把PCS层的64B/66B数据转换为适应具体介质的传输信号。FEC的主要功能是提供编码增益以提高链路预算和误码率性能。AN的主要功能是提供一种在链路两端设备交互信息的方法，可以自动地配置两端设备，使它们工作在共有的最高性能模式下。

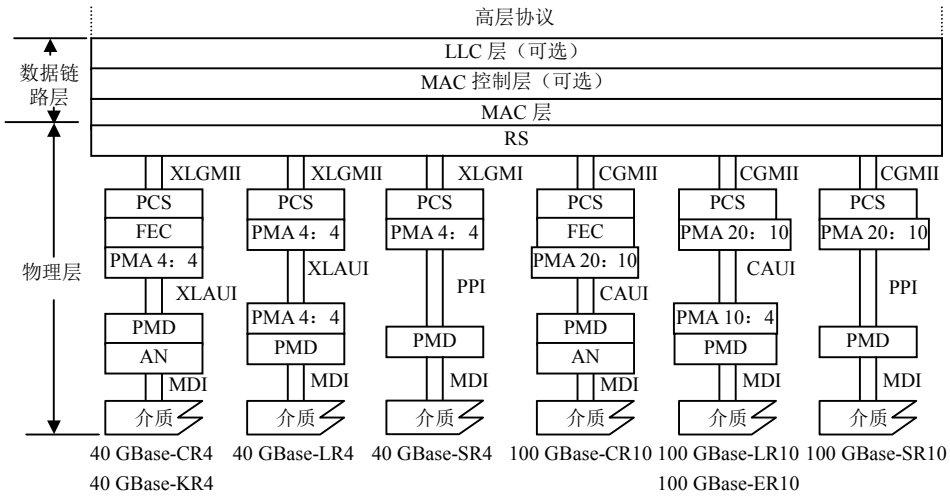

图3　下一代以太网的结构模型

图3中的介质无关接口(MII)是将MAC与PHY连接的逻辑接口；附属单元接口(AUI)是扩展PCS和PMA之间的连接，主要应用于芯片与芯片之间或者芯片与模块之间的连接。这些接口的名字跟踪10 GbE所建立的协议，10 GbE中XAUI和XGMII的“X”表示罗马数字10。由于40的罗马数字是“XL”，100的罗马数字是“C”，因而得出XLAUI和XLGMII用于40 Gb/s，CAUI和CGMII用于100 Gb/s。并行物理接口(parallel physical interface，PPI)用于40 GBase-SR4和100 GBase-SR10连接PMA与PMD的物理接口。管理数据输入/输出(management data input/output，MDIO)接口，目的是访问设备的寄存器，控制链路状态、速度、节能情况、故障、反馈、PCS的错误计数、测试模式等。介质相关接口MDI是PMD与介质之间的接口，包括连接的缆线和PMD插座，并要满足特定的接口性能规范。

4　传输模式与各子层和接口之间的关系

表1描述了40 GbE和100 GbE传输模式与各物理子层和接口之间存在的关系，其中M(mandatory)表示必选项，O(optional)表示可选项[1]。

表1　40 GbE和100 GbE传输模式与各物理子层和接口之间的关系

	对应关系	传输模式							
		40 GBase-				100 GBase-			
		KR4	CR4	SR4	LR4	CR10	SR10	LR4	ER4
物理子层和接口	AN	M	M			M			
	100 GBase-R FEC	O	O			O			
	RS	M	M	M	M	M	M	M	M
	XLGMII	O	O	O	O				
	CGMII					O	O	O	O
	40 GBase-R PCS	M	M	M	M				
	100 GBase-R PCS					M	M	M	M
	40 GBase-R PMA	M	M	M	M				
	100 GBase-R PMA					M	M	M	M
	XLAUI	O	O	O	O				
	CAUI					O	O	O	O
	40 GBase-KR4 PMD	M							
	40 GBase-CR4 PMD		M						
	100 GBase-CR10 PMD					M			
	40 GBase-SR4 PMD			M					
	100 GBase-SR10 PMD						M		
	PPI			O			O		
	40 GBase-LR4 PMD				M				
	100 GBase-LR4 PMD							M	
	100 GBase-ER4 PMD								M

表2对各种不同的物理层PMD接口，从传输距离、所用线缆、信号方式和实现方式共4方面进行了

总结[1]。

由表2可以看出，有4类不同的传输模式：背板、铜缆、多路并行传输的多模光纤(multi mode fiber，MMF)(纤芯直径为50 μm或62.5 μm)和单模光纤(single mode fiber，SMF)(纤芯直径为8.3 μm)。它们的传输距离、采用的线缆、信号方式和实现方式有很大的不同。表中缩写分别是：稀疏波分复用(coarse WDM，CWDM)；密集波分复用(dense WDM，DWDM)；光多模3(optical multimode 3，OM3)；半导体光放大器(semiconductor optical amplifier，SOA)是一个没有或有很少光反馈的激光二极管，功率消耗低，便于光学集成。

表2 新一代以太网传输模式的部分参量

PMD	40 GBase-KR4	40 GBase-CR4	40 GBase-SR4	40 GBase-LR4	100 GBase-CR10	100 GBase-SR10	100 GBase-LR4	100 GBase-ER4
传输距离/m	1	10	100	10 000	10	100	10 000	40 000
线缆	背板传输	铜缆	OM3并行MMF	SMF	铜缆	OM3并行MMF	SMF	SMF
信号方式	—	—	CWDM	CWDM	—	DWDM	DWDM	DWDM +SOA
实现方式	4×10 Gb/s	4×10 Gb/s	4×10 Gb/s	4×10 Gb/s	10×10 Gb/s	10×10 Gb/s	4×25 Gb/s	4×25 Gb/s

5 新一代以太网的设计实例——100 GbE的数据并行传输机制

100 GbE物理层对于MAC层来的数据处理的出发点是尽量提供最大的数据率和有足够的数据吞吐量，因此，在物理层的串行通道接口中采用64B/66B编码，经此编码后，速率提高到103.125 Gb/s (=100 Gb/s× 66/64)。对于如此高的数据流量，其时钟频率将高达1.6 GHz，这样的时钟频率在逻辑设计中很难达到。如果采用多通道(虚通道)并行处理，如10路并行处理的通道，每个通道数据位宽64 bit，这样161 MHz的时钟便可以满足设计需求(64 bit×10×161.1 328 125 MHz=103.125 Gb/s)。

虚通道中通道数L为PMA层中电层片间接口通道数M和PMD层连接到光纤介质通道数N的最小公倍数，即L=LMC(M,N)。如在实际的逻辑设计中，期望100G的以太网的设备可以适配不同的光模块以实现网络对接。当前用于100 Gb/s传输的光模块通常有10×10 Gb/s、4×25 Gb/s和5×20 Gb/s，即N取(10,4,5)，而电层片间通道数M通常选取M=10，根据L=LMC(M,N)，可知在PCS层中应设计20路虚通道。100 GbE的CGMII被定义为具有64 bit宽数据信号、8 bit宽控制信号和1.5 625 GHz时钟的逻辑接口。为了在一个循环处理一个64 B/66 B码块，必须减小时钟频率和采用并行传输的思路。如8并行CGMII会有512 bit宽数据信号、64 bit宽控制信号和195 MHz时钟。根据100 GbE的结构模型和上述的设计思路，提出了如图4所示设计实例。

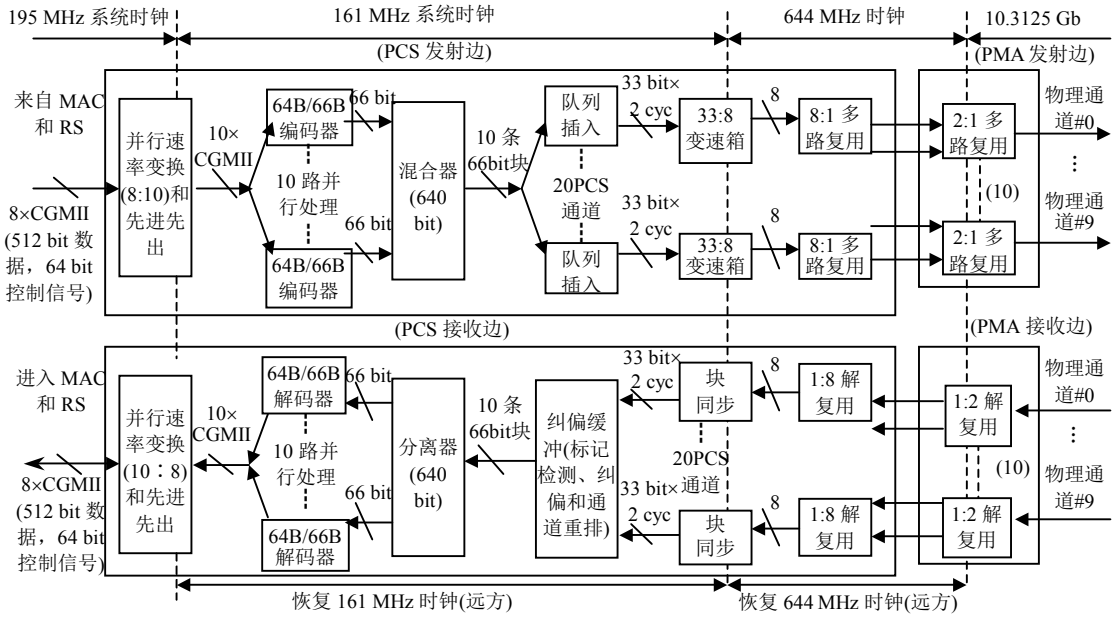

图4 100 GbE的PCS和PMA的并行实现框图

当并行处理时，需要把这种过程分成可并行的和串行的过程。如64B/66B编码过程分成代码变换过

程和数据混合过程。在代码变换处理过程中，不依赖于前后代码块，所以并行代码变换电路能够简单地并行。在数据混合处理过程中，一个给定的混合计算的块取决于串行比特流原块的处理结果，所以不能使用用于代码变换电路的相同并行处理方法。因此在一个单循环内，为了处理一个并行数据信号(如512 bit宽)，需要数据混合电路结构。

若选择161 MHz运行时钟，该时钟是103.125 Gb/s的1/640，把64 B/66 B编码解码并行成10路，同时把10个代码变换单元放在一路，每一路在一个循环处理一个块(64 bit)。对于数据混合电路来说，并行安放串行混合器以便在一路处理640 bit。顺便指出数据混合单元能够在一个单循环处理10个块。

6 结　论

新一代以太网技术是下一代互联网和物联网的基础。40 GbE和100 GbE是新一代以太网主要内容的两个方面，它们在技术上具有共同点，又有不同点。共同点是：只支持全双工通信，与以前的以太网MAC层的帧格式(包括最低和最高帧长度)相同，支持更好的不大于10^{-12}的误码率。不同点主要是速度和传输距离的问题，40 GbE的传输速度为40 Gb/s，在单模光纤、多模光纤、铜缆和背板上分别至少传输10 km、100 m和10 m；而100 GbE的传输速度为100 Gb/s，在单模光纤、多模光纤、铜缆上分别至少传输40 km(10 km)、100 m、10 m，它的背板传输模式目前还没有定义。另外它们在结构、信号方式和实现方式上也略有差距，40 GbE采用CWDM技术，实现方式只有唯一的4×10 Gb/s(4路并行传输)机制；而100 GbE采用DWDM技术，实现方式有10×10 Gb/s(10路并行传输)和4×25 Gb/s(4路并行传输)两种机制。

本文在综述新一代以太网的层次结构模型的基础上，介绍了各物理子层和接口的功能以及它们之间的关系，对40 GbE和100 GbE的传输模式与结构模型中各要素之间的关系进行了综合分析和比较。阐明了新一代以太网的基本原理和一些细节，解决了功能设计、接口设计、并行传输等的技术问题，这对于推动新一代以太网的研究、发展和应用具有重要参考价值。

参　考　文　献

[1] IEEE 802.3ba 40 Gb/s and 100 Gb/s Ethernet Task Force. LAN/MAN Standards Committee of the IEEE Computer Society[S]. New York, NY, USA: IEEE, 2009.

[2] AMBROSIA J D. 40 Gigabit Ethernet and 100 Gigabit Ethernet: the development of a flexible architecture[J]. IEEE Communications Magazine, 2009, 47(3): 8-14.

[3] MELLE S, INFINERA C，SUNNYVALE, et al. Bandwidth virtualization enables long-haul WDM transport of 40 Gb/s and 100 Gb/s services[J]. IEEE Communications Magazine, 2008, 46(2): S22-S29.

[4] UMBACH A. Advanced photoreceivers for 40 and 100 Gb/s optical communication networks[EB/OL]. [2011-10-15]. http://grouper.iee.org/groups/802/3/ba/index/html.

[5] FUJISAWA T, KANAZAWA S, TAKAHATA K, et al. 1.3 μm, 4×25 Gb/s, EADFB laser array module with large-output-power and low-driving-voltage for energy-efficient 100 GbE transmitter[J]. Optics Express, 2012, 20(1): 614-620.

[6] DROLET P, DUPLESSIS L. 100 G Ethernet and OTU4 testing challenges: From the lab to the field[J]. IEEE Communications Magazine, 2010, 48(7): 78-82.

[7] LYUBOMIRSKY I. Quadrature duobinary for high-spectral efficiency 100G transmission[J]. Journal of Lightwave Technology, 2010, 28(1): 91-96.

[8] CHACIŃSKI M, WESTERGREN U, STOLTZ B, et al. 100 Gb/s ETDM transmitter module[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2010, 16(5): 1321-1327.

[9] HERMSMEYER C, SONG H ,SCHLENK R, et al. Towards 100 G packet processing: challenges and technologies[J]. Bell Labs Technical Journal, 2009, 14(2): 57-80.

[10] DUELK M. GUTIERREZ-CASTREJON R. 4 ×25 Gb/s 40 km PHY at 1310 nm for 100 GbE using SOA-based preamplifier[J]. Journal of Lightwave Technology, 2008, 26(12): 1681-1689.

[11] 敖志刚, 赵水宁, 付成群, 等. 100吉比特以太网的系统架构与技术实现[J]. 解放军理工大学学报, 2011, 12(5): 445-448.
AO Zhi-gang, ZHAO Shui-ning, FU Cheng-qun, et al. System frame and technical realization of 100 gigabit Ethernet[J]. Journal of PLA University of Science and Technology, 2011, 12(5): 445-448.

[12] 敖志刚. 万兆位以太网及其实用技术[M]. 北京: 电子工业出版社, 2007.
AO Zhi-gang. 10 gigabit Ethernet and its practical technology[M]. Beijing: Publishing House of Electronics Industry, 2007.

编　辑　张　俊