数据中心光互联模块发展趋势及新技术研究

Research on Development Trends and New Technologies of Optical Interconnection Modules in Data Centers

宋梦洋,朱 虎,江 毅(武汉光迅科技股份有限公司,湖北 武汉 430205)
Song Mengyang,Zhu Hu,Jiang Yi(Accelink Technologies Co.,Ltd.,Wuhan 430205,China)

摘 要:
介绍了算力网络数据中心光互联技术的发展背景,分析了当前800 Gbit/s光互联模块的标准化、技术方案和产业化发展现状,并结合未来光互联模块发展需求,对1.6 Tbit/s光互联模块的标准化和行业新型实现技术等进行介绍,最后对未来光互联模块的发展趋势进行展望。

关键词:
算力网络;800 Gbit/s;1.6 Tbit/s;光电合封;线性直驱
doi:10.12045/j.issn.1007-3043.2024.02.007
文章编号:1007-3043(2024)02-0036-05
中图分类号:TN913
文献标识码:A
开放科学(资源服务)标识码(OSID):

Abstract:
It introduces the development background of optical interconnection technology in computing power network data centers, analyzes the standardization, technical solutions, and industrial development status of 800 Gbit/s optical interconnection modules, and combines the future development needs of optical interconnection modules to introduce the standardization of 1.6 Tbit/s optical interconnection modules and new industry implementation technologies. Finally, the development trend of future optical interconnection modules is discussed.

Keywords:
Computing power network;800 Gbit/s;1.6 Tbit/s;CPO;LPO

引用格式:宋梦洋,朱虎,江毅. 数据中心光互联模块发展趋势及新技术研究[J]. 邮电设计技术,2024(2):36-40.

0 引言

随着云计算、大数据、超高清视频、人工智能、5G行业应用等快速发展,网络访问频率、接入手段、数据处理和计算需求不断增加。特别是随着AI大模型应用的快速发展,国内外公司纷纷推出相关模型,各地开始启动智算数据中心建设。国际知名咨询公司LightCounting数次上调光互联模块市场规模预测,主要驱动力均来源于AI大模型智算数据中心旺盛的算力需求。

根据中国信息通信研究院发布的《中国算力发展指数白皮书(2022)》(见图1),2021年美国算力规模占全球份额的34%,中国以33%的占比位居全球第二。美国、中国、日本的GDP依次位居全球前三,而三者的算力能力也为全球前三,算力规模与国家GDP呈现正相关关系[1]。2023年10月,工业和信息化部等六部门联合印发《算力基础设施高质量发展行动计划》,完善算力综合供给体系,提升算力高效运载能力。2023年12月,国家发展改革委等五部门联合印发《深入实施"东数西算"工程 加快构建全国一体化算力网的实施意见》,大力推进算力网络建设,预计到2025年底,综合算力基础设施体系将初步成型。

算力网络的发展需要高速率、高带宽与高能效的互联技术来支撑,将高性能计算数据中心、AI数据中

收稿日期:2024-01-26

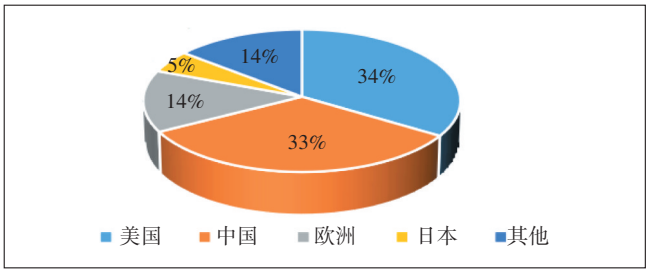

图1　2021年全球算力分布情况[2]

心以及基础算力数据中心统一连接起来，形成集成化的数据中心，从而实现算力的协同调度。

在这些需求带动下，用于数据中心互联的光模块处于高速发展阶段：400 Gbit/s光互联模块发货量快速增长，800 Gbit/s进入批量化进程，更高速率的1.6 Tbit/s光互联模块研发工作已经开展，全球主要标准化机构和多元协议组织（MSA）纷纷启动了基于单通道200 Gbit/s的1.6 Tbit/s光模块标准的研究和制定，IEEE在2021年底给出了相关标准的预计完成时间（见图2）。

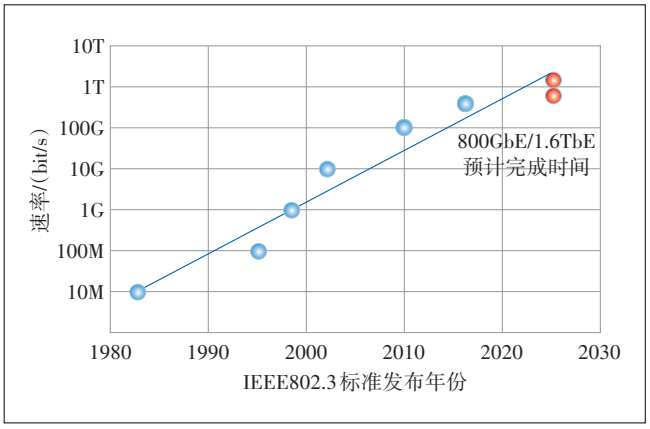

图2　基于单通道200 Gbit/s的标准化时间节点[3]

1　数据中心光互联模块发展现状

数据中心光互联应用主要分为两大类（见表1）：一类是数据中心内部的互联，典型光纤传输距离为2 km及以内；一类是数据中心之间的互联，典型光纤传输距离为80 km及以上。

典型光互联方式包含以下几种。

a）直连电缆（DAC），该方案采用铜缆，传输距离随带宽的增加而减少，但成本相对较低。

b）有源光缆（AOC），将光缆和光模块进行集成，光缆可根据传输距离进行配置，传输距离通常为100 m及以内。

c）光模块，根据传输距离需求采用不同规格的光模块，用于连接服务器、交换机等网络设备，承载高速数据的收发。

表1　数据中心互联场景[4]

互联场景		应用场所		典型距离	互联光模块类型
数据中心内部互联	场景1	服务器到TOR	机房内部	2 m（机架内） 30 m/50 m（机架间）	直调直检方案
	场景2	TOR到Leaf	楼栋内部	70 m/100 m	直调直检方案
	场景3	Leaf到Spine	楼栋之间	500 m/2 km	直调直检方案
数据中心之间互联	场景4	DC到DC	园区之间	80 km~120 km	相干方案

在数据中心内部互联场景中，以上几种互联方式均有采用，随着数据中心不断向高带宽、高速率演进，并且由于供电、GPU应用数量等原因，数据中心内部互联以基于直调直检方案的光模块和AOC为主。在数据中心之间的互联场景中，主要采用相干光模块进行连接。

数据中心内部光互联模块的发展与交换机交换芯片串行-解串行器（Serdes）的发展进度密切相关，交换机及光模块发展趋势见表2。2023年交换芯片Serdes的速率达到112 Gbit/s，交换芯片吞吐量相应达到51.2 Tbit/s，根据交换芯片演进趋势、市场需求及技术成熟度，预计2025年交换芯片吞吐量将达到102.4 Tbit/s，2027年将达到204.8 Tbit/s，光互联模块也需要相应演进到1.6 Tbit/s和3.2 Tbit/s对其实现有效支撑。

相干技术已经成为数据中心之间互联的主流方案。在多个标准化组织的大力推进下，400 Gbit/s光模块已发布多项标准，如400ZR、400G OpenROADM、Open ZR+等均采用DWDM技术，在C波段进行传输，结合DP-16QAM调制格式，可实现80~120 km（纯裸纤传输距离为40 km，增加光放可达到120 km）的高速传输。

随着400 Gbit/s光互联模块的批量化应用，800 Gbit/s光互联模块开始进入样品或小批量发货阶段，标准研究接近尾声（见表3和表4），后续随着标准的正式发布，将逐步走向批量化应用。

2　数据中心光互联发展趋势及新技术

AI算力网络与常规数据中心相比，对计算的需求量每18个月增长10倍，对高带宽、低时延的光互联需求更加迫切。目前国内外标准化组织纷纷启动1.6 Tbit/s光互联模块的研究工作，主流光模块厂家均已完

表2 交换机及光模块发展趋势预测

类型		2010	2012	2014	2016	2018	2020	2023		2025(预)		2027(预)
交换机	带宽/(Tbit/s)	0.64	1.28	3.2	6.4	12.8	25.6	51.2		102.4		204.8
	Serdes 通道数	64	128	128	256	256	512	512		1 024/512		1 024
	Serdes 速率/(Gbit/s)	10	10	25	25	50	50	112		112/224		224
光互联模块	速率/(Gbit/s)	200	40	100	100	400	400	400	800	800	1 600	3 200
	通道数	2	4	4	4	8	8	4	8	4	8	16
	单通道速率/(Gbit/s)	100	10	25	25	50	50	100	100	200	200	200

表3 800 Gbit/s光模块标准化进展

标准化组织	800 Gbit/s
IEEE	IEEE 802.3df已发布3.2草案 IEEE 802.3dj 预计24Q1发布1.0草案
800G Pluggable MSA	已发布800 Gbit/s PSM8和800G FR4规范
QSFP-DD800	已发布基于800 Gbit/s QSFP-DD800的6.01版本
OSFP	已发布基于800 Gbit/s OSFP的4.0版本
IPEC	800 Gbit/s 500 m/2 km已发布1.0版本 800 Gbit/s 10 km 文稿正在讨论过程中
OIF	800LR/800ZR 即将发布
中国通信标准化协会(CCSA)	800 Gbit/s AOC 已发布 8×100 Gbit/s 正在编制过程中 4×200 Gbit/s 正在编制过程中 1×800 ZR 正在编制过程中

表4 800 Gbit/s光模块标准部分技术方案[4]

标准化组织/名称	技术方案		距离
IEEE 802.3df	8×100 Gbit/s	8对MMF	50 m
		8对MMF	100 m
		8对SMF	500 m
		8对SMF	2 km
IEEE 802.3dj	4×200 Gbit/s	4对SMF	500 m
		4对SMF	2 km
		4波长复用SMF	500 m
		4波长复用SMF	2 km
		4波长复用SMF	10 km
	1×800 Gbit/s	相干	10 km
		相干	40 km
IPEC	8×100 Gbit/s	8对SMF	500 m
	2×400 Gbit/s FR4	4波长复用SMF	2 km
800G Plug-gabble	8×100 Gbit/s	8对SMF	100 m
	4×200 Gbit/s	4波长复用SMF	2 km
OIF	1×800 Gbit/s	相干	10 km
		相干	80 km

成1.6 Tbit/s光模块的样机研制工作(见表5)。

随着速率和带宽的增长,功耗也随之成倍增长。交换芯片、SerDes和光模块是功耗增加的主要因素。

表5 1.6 Tbit/s光模块标准化进展

标准化组织	1.6 Tbit/s
IEEE	IEEE 802.3dj 预计24Q1发布1.0草案
QSFP-DD800	已发布基于1.6T QSFP-DD1600的7.0版本
OSFP	已发布OSFP-XD 1.0,支撑1.6 Tbit/s可热插拔封装应用
4×400G MSA	启动4×400 Gbit/s研究
OIF	启动1.6 Tbit/s相干研究
CCSA	8×200 Gbit/s立项申请中

据推算,交换机从640 Gbit/s发展至51.2 Tbit/s,带宽增长80倍,功耗同时增长22倍。其中,专用集成电路核心(ASCI Core)的功耗增长8倍,系统风扇的功耗增长11倍,交换芯片SerDes的功耗增长25倍,光模块功耗增长26倍。将51.2 Tbit/s交换机的整机功耗按照上述4个维度进行分解,光模块的功耗约占交换机整机功耗的一半。因此在进行更高速率光互联模块开发设计时,功耗是无法绕开的瓶颈,目前行业中正在蓬勃发展的各项新技术,其主要需求驱动均为功耗控制。

2.1 光电集成——硅光技术

硅基光电子(简称硅光子)基于微米/纳米级光子、电子及光电子器件的新颖工作原理,可使用与硅基集成电路技术兼容的技术和方法,在同一硅衬底上实现单片或混合集成[5]。以此为基础的硅基光电子集成平台,可以利用现有的微电子工艺和成果,在硅衬底上同时集成微纳米尺寸光学回路与各类CMOS集成电路如调制器、探测器、互阻放大器、数字信号处理模块以及各类无源器件等,形成具有若干种功能的大规模集成芯片。硅光子技术结合了CMOS技术的超大规模逻辑、超高精度制造特性和光子技术超高速率、超低功耗优势,是一种可解决技术演进与成本矛盾的创新性热点技术,并已在通信光模块应用中发挥了积极作用,目前其技术发展势头强劲,产业规模不断扩大,产品与应用进展也在不断推进[5]。

硅光子集成技术的发展受到多方力量驱动。首

先，从集成角度来看，集成光器件相比分立光器件具有体积小、稳定性高等优势，可以大大减少分立光器件的数量和封装界面，减少传输路径，从而降低产品功耗；其次，从材料角度来看，硅相对于InP和GaAs等半导体材料价格更为低廉，且有望基于现有成熟、发达的微电子工艺，发挥规模优势提高工业化水平，从而进一步降低成本。

硅光子产品主要包括硅光子集成芯片和硅光子光模块（见图3）。目前硅光子技术凭借高集成度、低功耗、小型封装、大规模可生产性等优势，与共封装技术和薄膜铌酸锂调制技术联合应用，有望在800 Gbit/s、1.6 Tbit/s甚至更高速率的短距和相干应用中成为主力方案。

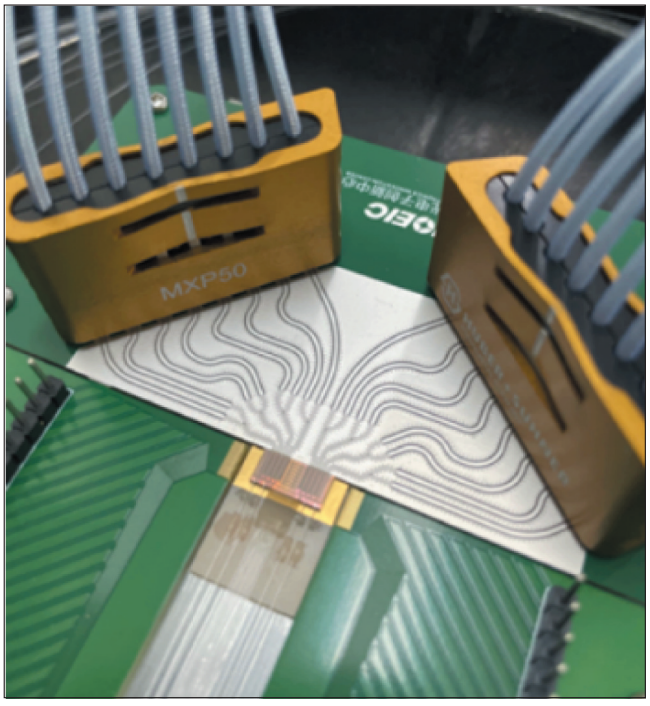

图3　1.6 Tbit/s光收发模块的COB封装器件

根据LightCounting预测，到2028年，硅光产品的市场份额将从2022年的25%增长至43%，领先的供应商都在布局硅光技术。但硅光技术的耦合效率仍相对较低，产业链完整性相比于Ⅲ－Ⅴ族仍有缺失，垂直整合能力有限。目前，国内硅光厂家已经开始硅基光电芯片晶圆级测试方法、硅光集成芯片技术规范等方面的行业标准编制工作。

2.2　共封装技术

光电合封（CPO）是将交换芯片、专用集成电路（ASIC）和光/电引擎（光收发器）共同封装在同一基板上，使引擎尽量靠近ASIC，以最大程度地减少高速电通道损耗和阻抗不连续性，从而有效降低整个系统的功耗。

光电合封的技术方案和应用场景主要聚焦在以下2个方面。

a）基于垂直腔面发射激光器（VCSEL）和多模光纤的解决方案，以30 m以内的应用为主，主要面向超算及AI机群的短距离光互联。

b）基于硅光和单模光纤的解决方案，以2 km以内的应用为主，主要解决大型数据中心机架及机群之间的光互联。

相较于可插拔光模块以及板载光模块，光电合封技术有如下优势。

a）光模块中高速电信号在印制电路板（PCB）上传输越来越困难，目前的PCB技术将112 Gbit/s以上的电信号从交换芯片传送到位于交换机面板的光模块难度较大。光电合封技术将交换芯片与光电转换单元封装在一起，可降低高频线路以及信号完整性电器件的使用要求，突破电信号传输瓶颈，提高数据通信的交换容量。

b）光电合封技术将光引擎置于板载上，靠近ASIC芯片，可释放前面板的压力。

c）在热插拔光模块中，DSP是高功耗的主要来源。在光电合封场景中，考虑到交换芯片本身具有均衡能力，可直接由交换芯片的Serdes驱动光引擎，光模块中高功耗的DSP/CDR可被省略，从而降低功耗。此外，光电合封技术采用外置光源方案，将激光器置于光收发单元外部，可降低光收发单元的热量，同时便于维修，出故障时只更换激光器即可，进而降低成本[6]。

1.6 Tbit/s光模块虽然已基本确认仍将采用可插拔方式设计生产，但随着AI计算等大交换机吞吐量需求的出现，光电合封技术在提升整个链路性能方面具有较大潜力，或将是更高速率光互联的主流解决方案。CPO技术目前仍有许多亟待解决的关键技术问题，例如高密度光纤连接的管理、散热管理、封装测试的良率以及可靠性等问题，需要业界共同摸索和实践。标准化方面，OIF已发布外置光源和3.2 Tbit/s的CPO标准，CCSA也立项了关于外置光源的行业标准。

2.3　薄膜铌酸锂调制技术

伴随着光刻技术的不断进步和混合集成工艺的发展，铌酸锂薄膜脊型波导结构的制造为薄膜铌酸锂

调制器的开发奠定了基础。薄膜铌酸锂调制器继承了体材料铌酸锂良好的物理化学稳定性，具有光学窗口宽、电光系数大、高线性度等优点，并通过优化设计可有效避免调制效率低、尺寸大等缺点。薄膜铌酸锂调制器的技术核心是对光波导结构和电极结构进行匹配设计，使其光电响应匹配，提高工艺精度，损耗减少，从而实现高性能、低功耗、小尺寸、低驱动电压的新型调制器，降低光互联模块的功耗。

各大厂商正在积极开发800 Gbit/s光互联芯片，128 GBaud波特率相干光通信芯片以及60~70 GHz以上带宽的特种通信用调制器芯片，它将成为未来超高速光互联领域的主流方案之一。

2.4 线性直驱技术

在光模块内部，发送端信号需经过数模转换（DAC），将数字信号转换为模拟信号；在接收端，模拟信号经过模数转换（ADC）后，再转换为数字信号。数字信号处理（DSP）芯片的主要功能是进行ADC/DAC、变速管理芯片（gearbox）的信号变速、电信号劣化补偿以及时钟数据恢复（CDR）。DSP是高速光模块的关键部件之一，但是功耗较高，其功耗约占光模块总体功耗的50%~60%。

线性直驱技术（LPO）在光模块中去除DSP/CDR芯片，模块内部只处理线性信号，由设备侧进行非线性信号的处理，从而降低光模块的功耗和成本。目前，LPO技术已在400 Gbit/s、800 Gbit/s速率上得到一定应用，但传输距离主要为500 m及以内，后续的规模化应用还需要技术的进一步发展、测试方法的建立以及标准的牵引。目前中国通信标准化协会已经启动LPO光模块研究课题，为后续标准化进行行业和技术分析。IPEC也启动了关于LPO方面的研究。

2.5 液冷光模块技术

为解决数据中心高密度设备散热和降低电源使用效率（PUE）的难题，液冷技术已获得广泛应用，从而产生对能够在液冷环境中配套使用的液冷光模块的需求。

液冷光模块需防止冷却液进入光模块内部光路，即光器件、光器件与光接口之间、光接口与尾纤之间存在的光路需整体采用密闭封装（液密封装），以实现同外部冷却液的完全隔离。液冷光模块的密封技术包括气密封装和液密封装或者2种封装方式的结合，这些技术保证光模块的密封性，防止气体或液体从内部泄漏到外部或从外部进入内部等。

液冷光模块能够很好配合系统进行散热，但相比常规光模块在成本方面有一定增加。主要体现在2个方面：一是物料成本，需采用绝缘、导热性能好、稳定性强的密封材料；二是加工成本，需通过较高的工艺和制造水平实现密封，且不能影响原性能参数、电磁兼容特性等要求[7]。

目前中国通信标准化协会已经完成用于液冷系统中的光模块的研究课题，正在进行用于液冷系统的光模块的标准立项申请。

3 结束语

随着算力网络概念的提出和推进，数据中心将进入高速发展快车道，光互联模块的机遇与挑战并存，国内外企业积极开展各项新技术的研究和实践。建立和完善涵盖光互联模块产业上下游的产业生态至关重要，为数据中心互联的健康发展提供有效支撑。

参考文献：

[1] 陈荣. 中国工程院院士郑纬民：算力互联汇聚超级计算力量[EB/OL]. [2023-10-22]. https://rmh.pdnews.cn/Pc/ArtInfoApi/article?id=37308579.

[2] 中国信息通信研究院. 中国算力发展指数白皮书[EB/OL]. [2023-10-22]. https://baijiahao.baidu.com/s?id=1777342370648807069&wfr=spider&for=pc.

[3] IEEE 802.3 Beyond 400 Gbit/s Ethernet Study Group. Project overview-IEEE P802.3df: 200 Gbit/s, 400 Gbit/s, 800 Gbit/s, and 1.6 Tbit/s ethernet[EB/OL]. [2023-10-22]. https://www.ieee802.org/3/B400G/public/21_1028/B400G_overview_c_211028.pdf.

[4] IMT-2020（5G）推进组. 5G承载与数据中心光模块白皮书[EB/OL]. [2023-10-22].https://max.book118.com/html/2021/1101/8011111037004030.shtm.

[5] 中国通信标准化协会. 硅光子技术在通信光模块中的应用研究[EB/OL]. [2023-10-22]. https://www.ccsa.org.cn/.

[6] 中国通信标准化协会. 光电合封技术研究[EB/OL]. [2023-10-22]. https://www.ccsa.org.cn/.

[7] 中国通信标准化协会. 用于液冷系统中的光模块研究[EB/OL]. [2023-10-22]. https://www.ccsa.org.cn/.

作者简介：
宋梦洋，毕业于华中科技大学，IEC注册专家，教授级高级工程师，硕士，主要从事光通信器件新技术可靠性研究以及标准化工作；朱虎，毕业于华中科技大学，高级工程师，硕士，主要从事光电子器件开发与市场推广工作；江毅，毕业于华中科技大学，高级工程师，硕士，主要从事通信光电子器件研发制造等工作。