CORNING

光模块的发展以及在 Hyperscale 用户中的演进

康宁光通信中国 产品管理部

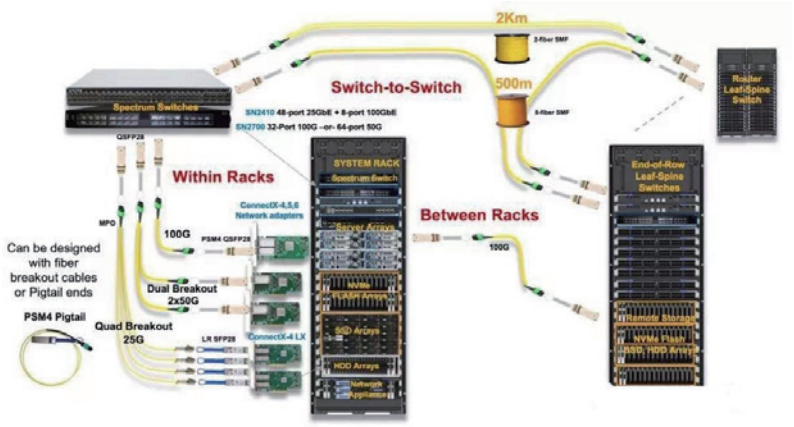
如果看过康宁公众号之前的一些康宁光连接方案和产品，那么相信读者对于康宁 MTP 和 LC 等数据中心中常用的连接产品标准，性能和优势略知一二。小编今天来介绍下用于连接 MTP 和 LC 接口的光模块（optical transceiver）的历史以及演进。

光模块发展历史

在正式展开光模块发展历史之前，我们先来聊聊促进光模块和数据中心发展的原始驱动力。

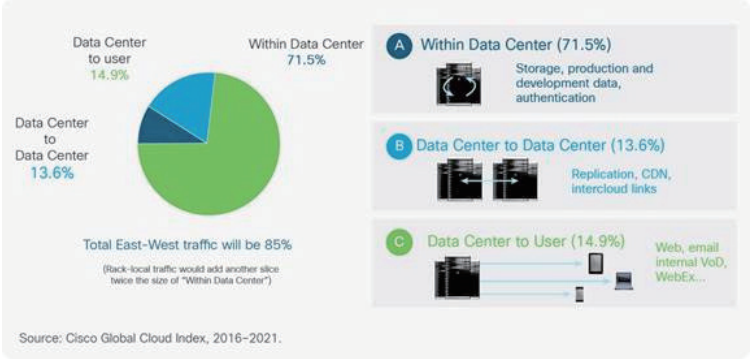
随着通信技术，基于互联网应用的不断发展，人们对于计算能力和数据存储的要求，渐渐地从个人主机往“云”上迁移，而企业原有的一些内部计算存储需求，也随着云计算壮大带来的成本和管理优势，迁移上了“云”。个人和企业对于计算和存储能力的需求，正是促进数据中心在最近数十年大发展的驱动力。而所谓的“云”对应的基础设施正是数据中心。

数据中心里有什么？数据中心里面其实就是码放整齐的服务器和各类交换机 / 路由器。而服务器和交换机上插满了各种光模块，以用于数据的传输和交换。



Source: mellanox.com

如下图所示，数据中心内部的数据交换量占了 70% 以上的比例，而不同的数据中心之间的 DCI（Data Center Interconnection）数据通信仅为 13% 左右，这也就可以理解为什么数据中心业务大发展阶段，与之对应的光模块发展如此的迅速。



讲到这儿，我们开始介绍下光模块（optical transceiver）的历史。

从原始社会通信基本靠吼，到飞鸽传书，再到电话电报，直到当今的光网络，通信技术一直不断往前发展。但是完成信息传递的三个基本要素，即信源、信通道和信宿，也就是信息的发送、传递和接受，这三点缺一不可；所有技术的发展都是围绕着这三点来实现的。

当通信进入到应用现代科技阶段时，先是以电为研究对象，从电的特性出发，改善通信的质量。从早期的固定电话，到 2G、3G 无线通信基本都是基于电的通信方式。大家最近这些年常常听到的“光进铜退”，指的就是由于电缆这种介质本身特性无法实现高速率信号的长距离传输，从而限制了它的进一步发展。用电传输信号，随着传输距离增加频率越高损耗越大，信号变形越厉害，从而引起了接收机的判断错误，导致通信失败。为了克服这个限制，光模块就是把电信号在发射端转成光信号，就是我们说的 Transmitter, 即发送器，它负责将设备产生的电信号转换成光信号发出；而在接收端再把收到的光信号转换成电信号，就是 Receiver，也叫做接收器。如果把 Transmitter 和 Receiver 做在一个封装模块里，就成了 Transceiver，既可以发送也可以接收，光模块（optical transceiver）就是这样形成。

早期的光模块从一开始 155Mb/s(一秒钟传输 155 万个比特)，到 622Mb/s,1.25Gb/s,2.5Gb/s 一直到 10Gb/s，利用的是时分复用的技术，也就是 TDM(Time Division Multiplexing)，在单位时间内传输更多的比特数。但一个光模块的传输速率再快，也不如几个同时一起传输，那么就慢慢有了并行传输，称之为 parallel，4 个并行的叫 QSFP, 12 个并行的叫 CXP。

对于短距离传输来说，并行传输多用的一些光纤对于材料和建设成本并无多大区别。但对于长距离通信来说，铺设一根光缆的建设成本远远大于材料本身，长距离传输时，TDM 时分复用会受限于电子器件开关频率，那么用一根光纤来传输多个波长的技术就随之诞生，我们称之为 WDM(Wavelength Division Multiplexing)， WDM 又分为两种，20nm 间隔的 CWDM（Corse DWM, 粗波分复用）和 0.8nm 间隔的 DWDM（Dense WDM 密集波分复用）。

光模块的封装形式

接下来，我们再来了解下光模块的封装形式 (Package Form)，这个是最首要的 Transceiver 的分类方式，也是产品线划分的依据。

在光模块行业成型之前，早期由各大电信设备制造商各自开发，接口五花八门，互不通用。这样导致光收发模块大家用起来都不方便，于是大家一起制定规则，就有了 MSA（Multi Source Agreement）, 多源协议。有了 MSA 标准之后，独立专注于开发 Transceiver 的公司开始崭露头角，随之行业兴起。

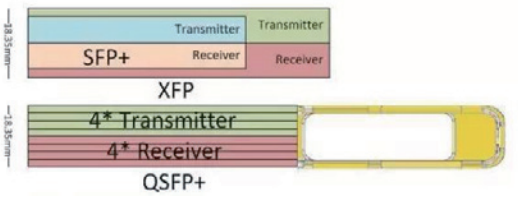
篇幅原因，无法详细解释所有光模块。那么我就来看看常见的几种，例如 SFP/XFP/SFP。

随着工艺水平的不断提高，Transceiver 尺寸越做越小，有了 SFP (Small Form-Factor Pluggable)

的 Transceiver 模块，也称为小封装可插拔模块，支持热插拔，即插即用。SFP 的速率越做越高，从 1.25G， 2.5G， 4G， 6G，到了 10Gb/s 以后，原先的封装大小就放不下那么多的元器件了，就定义了新的标准 XFP。XFP 指的是 10Gb/s 速率的可插拔光模块。技术一直在进步，到了 2009 年集成工艺的提升，终于可以把 XFP 塞进 SFP，这种新的 SFP 的 Transceiver 称作 SPF+，即增强型 SFP 模块。SFP 和 SFP+ 尺寸大小，连接器定义，功能完全相同；为了区分，把支持 8Gb/s 以上的 SFP 称为 SFP+。

Parallel并行光模块的演进

人们对于带宽的需求不断增加，那么同样对于数据中心内的 Transceiver 速率提出了更高的要求，TDM 光模块的限制显现出来了。到了 10Gb/s 的时候，新的设计思路是把 Transceiver 做的更小，同时把几个 Transceiver 装在与原来大小一样的封装里，Parallel 的光模块由此诞生。如下图所示，QSFP+ 和 SFP+ 封装的尺寸对比。



Parallel QSFP+ 光模块的兴起，并不意味着 SFP+ 的退出。用户会根据自己的需求以及产品的性价比来选择不同的光模块。因此，SFP+ 仍有其存在的理由和市场；随着技术的不断提升，SFP+ 也推出了 28Gb/s 的产品。

谈到这儿，顺便提一下光模块的光接口类型。对于传统的 XFP/SFP+，有两个光接口，一收一发，使用的是 LC 类型的接口。对于 Parallel 模块，QSFP(4 收 4 发) 甚至 CXP(12 收 12 发)，那么就需要用到多发多收，即 MPO（MPO 是 Multiple fiber push-on/push off）接口。康宁的数据中心解决方案里使用了更高工艺标准的 USCONEC MTP 接口，根据光模块接口类型，我们提供 MTP8，MTP12， MTP 16 和 MTP24 接头类型。

光模块作为信号的信源和信宿，担当了通信三要素中的亮点。与此同时，康宁提供的数据中心产品方案则是信道，信道的可靠性对通信起着至关重要的作用。关于如何选择 LC 和 MTP 接头的连接产品，可以参考公众号之前发布的文章。

对于光模块而言，应用在什么交换机上不重要，重要的是根据距离选择合适的模块类型，以 100G 为例：

传输距离	模块类型
<20m	25G AOC, 100G QSFP AOC
<100m	100G QSFP SR4
<500m	100G QSFP PSM4
<2km	100G QSFP CWDM4, CLR4
<10km	100G QSFP LR4
<40km	100G QSFP ER4
>80km	Coherent CFP/CFP2 ACO, DCO

当然目前 100G 主流的光模块应用仍然是小于 100m QSFP SR4。大型的 Hyperscale 云服务供应商在光模块选型上，更多的会从产业成熟度、稳定供应、平滑升级的几个方面去考虑并设计对应的架构。

Hyperscale客户对于未来光模块的需求

目前 Hyperscale 超大型数据中心的建设方主要是云服务提供商，例如亚马逊， 微软， 谷歌，阿里巴巴，腾讯等。对于以上云服务提供商的数据中心，单体数据中心的机柜数量往往在 1 万以上，规模非常庞大。这些云服务提供商对于各自数据中心物理和逻辑架构的设计各有千秋。上文中提到的数据中心流量的覆盖，绝大部分在数据中心内部，也就是服务器和交换机之间。大量的数据中心内部流量覆盖也对应了庞大的光模块需求数量。数据中心作为高投资和高能耗的设备集群，对于建设的要求始终是更高的投资回报率、更低的能源消耗和使用率。那么对应到光模块的要求，更多会考虑整体成本以及更低的功耗。

以 400G 为例，中国的阿里巴巴和腾讯均在公开场合公布了 400G 的建设计划，并预计在 2020 年会有试验的 400G 网络，逐步铺开，并将根据整体架构设计综合考虑光模块的选型。下图是目前 400G 光模块的潜在产品和标准。我们可以看出，多模并行和单模并行的光模块都是继续 8 芯的应用。康宁早在 2015 年就推出了 EDGE8TM 的数据中心连接方案，提供了灵活的产品配置以满足 400G 的组网和架构需求。

光纤类型应用	PMD	波长 / 单波长速率 / 接头	传输距离
MM WDM	No solution	/	/
SM WDM	400G-LR8 400G-FR8	8λ@50G, 2F LC 8λ@50G, 2F LC	10KM 2KM
MM Parallel + WDM	400G-SR4.2 400G-SR8	2λ@50G, 8F MTP 1λ@50G, 16F MTP	150M-550M
SM Parallel	400G-DR4	1λ@100G, 8F MTP	500M

再来展望一下 hyperscale 数据中心基于业务和应用的前景。近几年我们常听到 AI 人工智能，BLOCK CHAIN 区块链和大数据等前沿技术，那么这些技术对于数据中心有什么影响呢？就拿 AI 人工智能举例。早在 2014 年，谷歌公司就在其中一个数据中心设施中部署了 Deepmind AI（使用机器学习和人工智能的应用程序）。其结果是，能够将数据中心用于冷却的能源减少 40%，这相当于在考虑到电气损耗和其他非冷却效率之后，PUE（Power Usage Effectiveness 一种评估数据中心能源效率的指标）值减少了 15%，这也产生了该数据中心有史以来最低的 PUE。基于这些显著的成本节省，谷歌公司希望在其他数据中心中部署该技术，并建议其他公司也这样做。

Facebook 公司秉承的使命是“让人们有能力建立社区，让世界更紧密地联系在一起”，Facebook 公司的应用机器学习白皮书从数据中心基础设施视角进行概述，它描述了支持全球范围内机器学习的硬件和软件基础设施。

为了让人们了解人工智能和机器学习需要多少计算能力，百度公司硅谷实验室的首席科学家 Andrew Ng 表示，培训百度的中文语音识别模型不仅需要 4TB 的训练数据，还需要 20 个计算机的 exaflops（百亿亿次）计算量，也就是整个培训周期内需要 200 亿次数学运算。而这些计算能力需要靠数据中心具备更高的计算能力，以及更快的内部传输速率来实现。

如何搭建数据中心内部或数据中心之间的流动数据高速公路？其实从 hyperscale 100G 的数据中心架构开始，SPINE-LEAF 脊叶架构成为主流的架构。这种架构满足了数据中心东西流量需求的同时，也带来了更为复杂的服务器到交换机，以及不同层级交换机之间的交叉互联，这种交叉连接的部署，对于布线的挑战，甚至布线管理的难度，提出了更高的要求。我们将这种交叉连接称为 MESH 网络，hyperscale 数据中心在设计时除了考虑基础设施的 PUE，设备，光模块的选型之外，如何满足迅速扩展的云计算业务，加速数据中心 scale out（横向扩展） 和 scale up（纵向扩展）也是非常重要的考虑因素。康宁作为专业的光通信连接方案厂商，能提供优于行业标准

的，更可靠稳定的产品和方案，针对于 MESH 的应用，康宁也有 MESH 模块（ module）产品帮助客户快速实现数据中心的 scale out/up。

另外，对于 400G 超大规模数据中心的连接设计，单模 SM MPO 连接产品使用比例进一步攀升，对于 SM MPO 接头工艺的要求变得更高。康宁高品质的 MTP 单模和多模连接产品，能为数据中心的快速建设和安全运营提供更得力的保障。

更多康宁光连接产品方案，请访问康宁产品网站： http://www.corning.com/cn/zh/products/communication-networks/products.html

康宁光通信中国
上海市漕河泾高新科技开发区桂箐路 111 号立明大厦 3 楼
电话：86 21 5450 4888
传真：86 21 5427 7898
www.corning.com