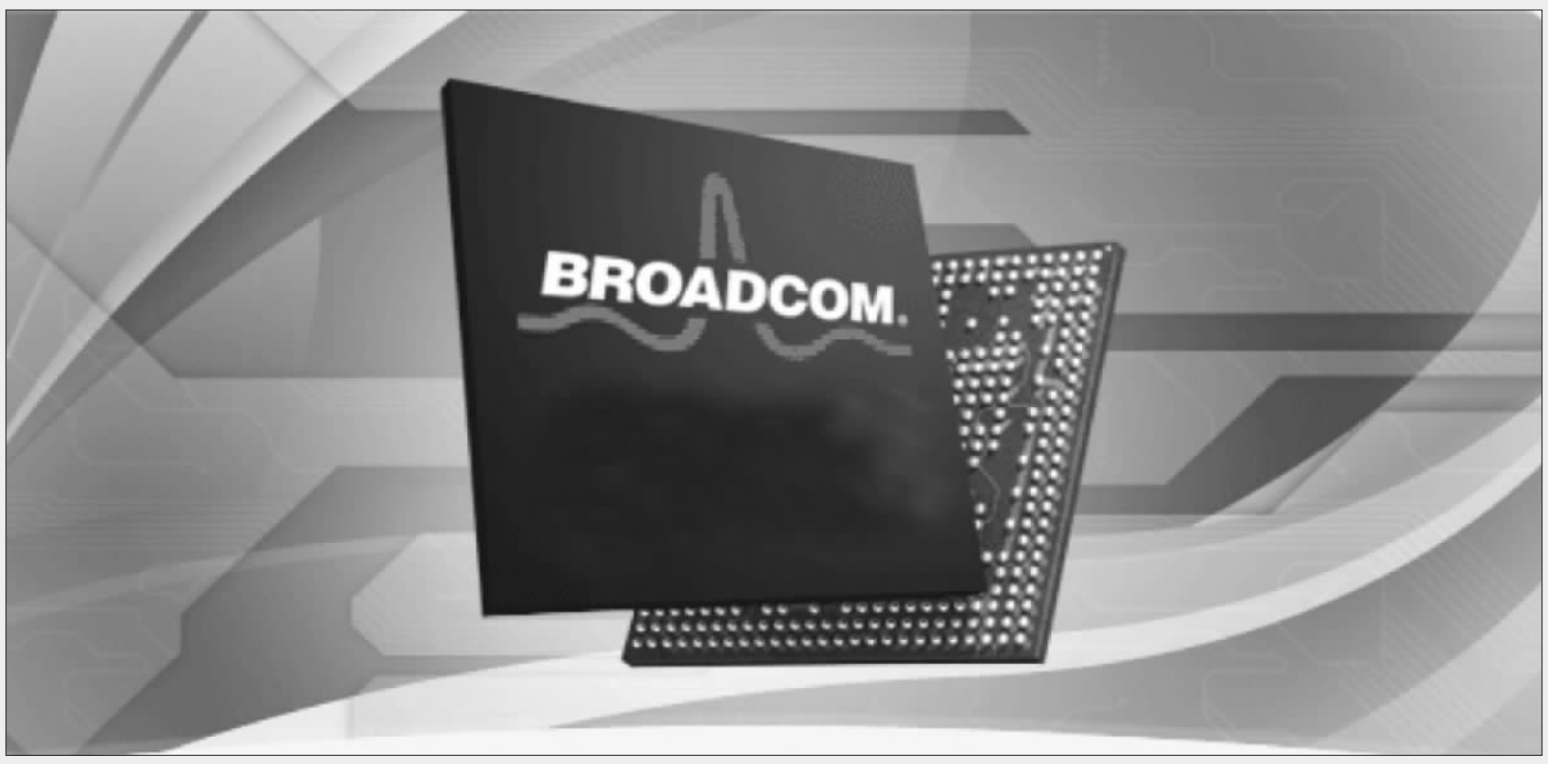


AI浪潮推高交换芯片需求

本报记者 姬晓婷 实习记者 夏冬阳

当前，半导体市场对生成式人工智能的关注大多集中于计算芯片，但在数据中心和网络通信基础设施加快建设的推动下，一颗以太网交换芯片的价格暴涨正引发关注：2024年第一季度以来，博通Tomahawk4系列的多款交换芯片价格异常走高，在其官网和其他交易平台上大多显示无库存，且交货期高达50周，其中BCM56990B0KFLGG市场报价已达4100美元左右，成为数据中心网络芯片市场的“黑马”。交换芯片价格暴涨、供应短缺背后，透露出怎样的信号？以太网交换芯片是否真的成为半导体市场在AI浪潮下的又一受益者？

长期备货量少，难以应付市场需求激增

交换芯片的价格变动趋势受到多种因素的影响，包括市场需求、供应链状况以及全球半导体市场的整体环境等。

网络交换机是一个扩大网络的装置，能为子网络提供更多的连接端口，以便连接更多的计算机。交换芯片是网络交换机的核心部件，主要负责数据的转发功能。当一个数据包到达交换机时，交换芯片会读取数据包的目标地址，并迅速将其转发到目的端口。以太网中的数据交换，特别是在大型网络中，依赖于高效的交换芯片来确保数据的快速、准确传输。可以说，交换芯片的性能直接决定了交换机的性能，并直接影响以太网的传输效率和响应速度。高性能的交换芯片可以处理更多的数据交换请求，从而确保网络的高吞吐量和低延迟。例如，在数据中心或云计算平台中，高性能的以太网交换芯片是支持大规模数据传输和处理的关键。

长期以来，网络交换机更新换代速度相对较慢，交换芯片市场规模相较于计算芯片、存储芯片等其他类型半导体产品来说，在整个半导体市场中的占比较低，未曾受到市场的广泛关注。

而在当前，全球以太网交换芯片市场正在经历前所未有的变革。在AI和大数据技术的推动下，传统的数据处理和传输方式已经无法满足日益增长的数据需求。特别是在云计算、大数据等技术的推动下，数据中心正在向大规模、高密度的方向发展。在这样的背景下，要求交换芯片不仅要有更高的性能，还需要有更强的可扩展性和灵活性。因此，交换芯片的设计和制造难度也在不断增加。

低性能交换机芯片已经无法满足用于大模型训练等对数据传输效率有着更高要求的场景需求，这才在短期内推高了博通高端交换芯片的市场需求。这也是BCM56/58系列芯片价格在短期内暴涨的原因。

据悉，BCM56990能够在单个芯片上提供高达25.6Tb/s的高带宽，无缝网络连接，可应用于超大规模云网络，存储网络及HPC（高性能计算）等场景。

交换芯片的价格变动趋势受到多种因素的影响，包括市场需求、供应链状况，以及全球半导体市场的整体环境等。芯谋研究企业服务部总监王笑龙在接受《中国电子报》记者采访时表示，博通的交换机芯片，由于单价高、用量低，一般采用订货、生产的方式，产品余量少，这才导致了在市场需求量突然增加的情况下，由于产品本身供给量有限而出现价格猛涨的情况。

数据中心规模扩张是主要增长动能

网络交换机和交换芯片的需求量、增长预期与数据中心数量及单个数据中心内建设的计算节点数量强相关。

超聚变数字技术有限公司算力基础设施领域CTO丁煜表示，2018—2023年，大模型训练算力需求每年以10倍的速度激增，而GPU算力每年以1倍的速度增长，算力需求与供给之间存在巨大的矛盾还没有得到解决，还需要大规模的集群计算。

由此判断，未来几年，大规模计算集群仍将持续建设，数据中心之间及其内部的数据传输量不断增长，需要更高带宽的交换芯片来满足数据传输的需求，进而给网络交换机及交换机芯片带来增长空间。

目前，国际交换机芯片整体市场主要由几家巨头主导，如博通、思科和美满等。这些公司凭借强大的技术实力和丰富的产品线，占据了市场的较大份额。特别是在超大规模的云数据中心和HPC集群领域，他们的产品具有显著的优势。最近在交换芯片市场表现最好的BCM56990B0KFLGG便是博通公司旗下高速率交换芯片的代表。

长期以来，博通的Tomahawk 4芯片一直被看成400Gbps光模块起量的标志，并做到约两年就将带宽增加一倍，一直在数据中心网络芯片市场保持领先。其在2022年8月推出的Tomahawk 5系列网络芯片，已将带宽提升到51.2T，可支撑64个800G的端口或128个400G端口，数据交换性能是Tomahawk 4的两倍。

而随着大模型训练带来的数据交换量增长，网络交换芯片将向着更高端口速率和转发速度的方向演进。

网络交换机和交换芯片的需求量、增长预期与数据中心数量及单个数据中心内建设的计算节点数量强相关。

本土企业增量空间大

到2025年，中国商用以太网交换芯片市场中，100G及以上带宽的高速交换芯片市场需求和规模将大幅增长，市场规模占比有望达到44%。

交换芯片主要分为两大类：自研型和商用型。自研芯片由制造商自行设计并专用于其交换机产品，构成了公司数据通信业务的核心基础，通常不对外销售，主要的生产厂商包括华为、思科和中兴等。而商业销售型芯片则是由芯片制造商生产后，直接面向市场销售给其他厂商，其中包括国际龙头企业博通、美满以及国内的盛科通信。

记者了解到，目前国内交换芯片品类最高的交换容量与国际品牌相比仍有2~3代差距，但随着国内厂商的技术进步和自主创新能力的提升，国内企业如盛科通信等，已经在交换芯片领域取得了显著进展，推出了具有竞争力的产品。例如其TsingMa.MX（交换容量2.4Tbps，支持400G端口速率）、GoldenGate（交换容量1.2Tbps，支持100G端口速率）等系列已导入国内主流网络设备商并实现规模量产。根据盛科通信招股书显示，其计划推出的Arctic系列目前正处于研发的后端设计阶段，交换容量最高达到25.6Tbps，支持最大端口速率800G，面向超大规模数据中心，有望对标行业一线龙头。“我们认为，随着国产交换芯片设计商技术不断迭代升级，国产芯片有望进一步导入中端交换产品，并逐步向中高端市场渗透。”中金公司研究部表示。

随着市场对高速网络解决方案的需求不断增长，交换芯片的速率也在同步提升。根据市场分析机构数据，到2024年，全球和中国市场中速率低于100M的交换机端口将逐渐被淘汰，千兆端口将继续占据市场的主导地位。Dell'Oro Group数据预测，2020—2025年期间，数据中心交换机市场的年复合增长率为9%。

商汤科技大装置事业群解决方案总监代继在接受《中国电子报》记者采访时表示，在搭建大模型时代的AI基础设施时，会采用国产交换机。“我们不仅要看芯片，还要看交换机产品本身，华为自研交换芯片会搭载自己的交换机做研发，我们在应用过程中会注重交换机的整体性能和软件能力。”

随着云计算服务的不断扩展和各行各业数字化转型步伐的加快，数据中心的规模正迅速扩大，对可扩展的高性能网络解决方案的需求日益增长。因此，为满足边缘计算应用而设计的以太网交换机芯片变得尤为关键，灼识咨询的预测指出，到2025年，中国商用以太网交换芯片市场中，100G及以上带宽的高速交换芯片市场需求和规模预计大幅增长，市场规模占比有望达到44%。

芯谋市场分析师顾文军在接受中国电子报记者采访时表示：“中国交换芯片的未来市场空间较大，关键在于我们自己的产品都不够强。”

公开数据显示，2020年全球以太网交换芯片市场规模为368亿元，2025年全球以太网交换芯片市场规模预计达到434亿元，2020—2025年年复合增长率为3.4%。灼识咨询数据预测，中国交换芯片市场预计2025年达到225亿元，年复合增长率约13%。能够看出，中国交换芯片市场规模增长率明显高于全球。

台积电第一季度营收环比下降5.3%
地震将影响第二季度毛利率

本报讯 记者**王信豪**报道：4月18日，台积电公布2024年第一季度财务报告，总营收为5926.4亿新台币（约合人民币1324.1亿元），环比下降5.3%，同比增长16.5%；净利润为2259.9亿新台币（约合人民币504.9亿元），环比下降5.5%，同比增长8.9%。

台积电表示，受到智能手机周期性波动的影响，第一季度营收环比小幅下滑，不过下降程度被HPC（高性能计算）抵消了一部分。从业务营收构成来看，台积电HPC业务营收占比达46%，环比增长率为3%，而智能手机业务占比为38%。

“在未来几年中，几种AI处理器将成为我们HPC业务增长的最强推动力，并成为增量收入的最大贡献者。”台积电总裁魏哲家表示。

台积电预测，2024年AI服务器的收入贡献将增加一倍以上，占2024年总收入的10%，并于2028年持续增长至20%以上。台积电表示，越复杂的AI大模型越需要更好的半导体硬件支持，即更先进的半导体工艺以及封装技术。

在先进制程方面，台积电7nm、5nm及3nm制程于本季度分别带来19%、37%和9%的营收，总占比达65%，环比增长7%。关于2nm，台积电透露，预计于2025年年底投产，并于2026年上半年开始交付并赢利。“我们观察到客户对2nm的兴趣和参与度很高。”魏哲家直言道。

目前，台积电3nm制程主要应用于智能手机中，而AI加速器普遍使用5nm或4nm制程。魏哲家表示，能源效率是HPC客户必须考虑的问题，更先进的制程具有更良好的功耗表现，因此，未来AI芯片也将加速实现从4nm到2nm的过渡。

在先进封装方面，台积电表示，CoWoS封装的需求在近两年内都非常强劲，台积电2024年的CoWoS产能相较去年已经提升了一倍以上，但仍供不应求。“我们正在尽最大努力增加产能以缓解短缺。”魏哲家说，“当然，面对GPU之外的AI芯片（比如ASIC等）在先进封装上的需求，台积电也乐意服务这类型的客户。”

4月3日，中国台湾地区花莲县东部海域发生7.3级地震。受此影响，台积电等晶圆制造厂均出现生产线停机等状况。

在此次电话会议中，台积电首席财务官黄仁昭详细阐述了员工及工厂情况。“凭借台积电在地震响应和灾害预防方面的经验，所有台积电员工都很安全，且工厂的整体工具回收率在震后10小时内达到了70%以上，并在第三天结束时完全恢复。”黄仁昭表示，工厂没有受到结构性破坏，关键工具（包括EUV光刻机）保存完好。

但是，受地震影响，台积电生产的大量晶圆片不得不作报废处理。黄仁昭表示，由于晶圆废片和相关材料的损失，地震将稀释台积电第二季度约0.5%的毛利率，不过损失的大部分产能也将在第二季度恢复。

此外，由于中国台湾地区电价上涨，用电成本的提升间接导致其他化学材料及气体成本波动。台积电预计，这部分因素将影响公司第二季度0.7%至0.8%的毛利率。相比于第一季度53.1%的毛利率，台积电将第二季度的毛利率指引调整为51%至53%。

英特尔发布大型神经拟态系统Hala Point
或大幅降低大模型训练能耗

本报讯 记者**姬晓婷**报道：北京时间4月18日凌晨，英特尔发布了代号为Hala Point的大型神经拟态系统。Hala Point基于英特尔Loihi 2神经拟态处理器打造，旨在支持类脑AI领域的前沿研究，解决AI目前在效率和节能等方面的挑战。Hala Point在英特尔第一代大规模研究系统Pohoiki Springs的基础上改进了架构，将神经元容量提高了10倍以上，性能提高了12倍。

Hala Point系统由封装在一个六机架的数据中心机箱中的1152个Loihi 2处理器（采用Intel 4制程节点）组成，大小相当于一个微波炉。该系统支持分布在140544个神经形态处理内核上的多达11.5亿个神经元和1280亿个突触，最大功耗仅为2600瓦。Hala Point还包括2300多个嵌入式x86处理器，用于辅助计算。

在大规模的并行结构中，Hala Point集成了处理器、内存和通信通道，内存带宽达每秒16PB，内核间的通信带宽达每秒3.5PB，芯片间的通信带宽达每秒5TB。该系统每秒可处理超过380万亿次8位突触运算和超过240万亿次神经元运算。

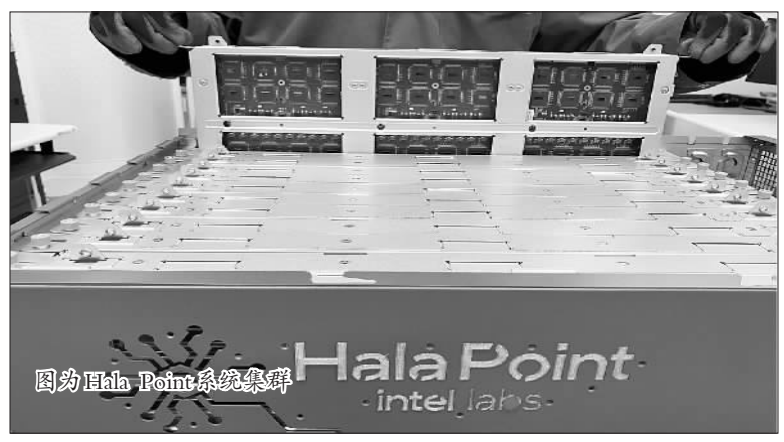
在用于仿生脉冲神经网络模型时，Hala Point能够以比人脑快20倍的实时速度运行其全部11.5亿个神经元，在运行神经元数量较低的情况下，速度可比人脑快200倍。虽然Hala Point并非用于神经科学建模，但其神经元容量大致相当于猫头鹰的大脑或卷尾猴的大脑皮层。

Loihi 2应用了众多类脑计算原理，如异步（asynchronous）、基于事件的脉冲神经网络（SNNs）、存算一体，以及不断变化的稀疏连接，以实现能效比和性能的数量级提升。神经元之间能够直接通信，而非通过内存通信，因此能降低整体功耗。

在执行AI推理负载和处理优化问题时，Loihi 2神经拟态芯片系统的速度比常规CPU和GPU架构快50倍，同时能耗降低至原来的1%。早期研究结果表明，通过利用稀疏性高达10比1的稀疏连接和事件驱动的活动，Hala Point运行深度神经网络的能效比高达15TOPS/W，同时无须对输入数据进行批处理。批处理是一种常用于GPU的优化方法，会大幅增加实时数据（如来自摄像头的视频）处理的延迟。尽管仍处于研究阶段，但未来的神经拟态大语言模型将不再需要定期在不断增长的数据集上再训练，从而节约数千兆瓦时的能源。

神经拟态计算是一种借鉴神经科学研究的全新计算方法，通过存算一体和高细粒度的并行计算，大幅减少了数据传输。研究显示，在运行传统深度神经网络时，该系统能够每秒完成2万万亿次（20pet-aops）运算，8位运算能效比达到了15TOPS/W，相当于超过了基于GPU和CPU的架构。

目前，Hala Point是一个旨在改进未来商用系统的研究原型。英特尔预计其研究将带来实际技术突破，如让大语言模型拥有从新数据中持续学习的能力，从而有望在AI广泛部署的过程中，大幅降低训练能耗，提高可持续性。



图为Hala Point系统集群

社长：张立 社址：北京市海淀区紫竹院路66号赛迪大厦18层 邮编：100048 每周二、五出版 周二8版 周五8版 零售4.50元 全年定价420元 广告部：010-88558848/8808 发行部：010-88558777 广告许可证：京石工商广登字20170003号 发行单位：中国电子报社 印刷单位及地址：经济日报印刷厂 北京市西城区白纸坊东街2号