ZTE中兴

数字基础设施
技术趋势白皮书

数字基础设施技术趋势白皮书

版本	日期	作者	备注
V1.0	2023.05	中兴通讯	

目录

1. 前言

技术创新是生产力进步、产业发展的核心驱动力。世界经济论坛创始人克劳斯·施瓦布（Klaus Schwab）在其所著的“第四次工业革命”中指出，近代以来，人类经历了四次由技术创新引领的工业革命。第一次起始于1760年左右，以蒸气机、铁路的发明和广泛应用为主要标志，使人类从手工工艺时期跃进到机械化生产阶段；第二次始于19世纪末，以电的发明和广泛应用为标志，电力工业、化学工业以及电报、电话等的迅速发展，使人类进入大规模工业化生产时代；第三次工业革命始于20世纪中期，以通信技术、计算机技术、互联网技术（简称信息通信技术，或ICT技术）为主要标志，人类进入自动化生产阶段；第四次工业革命目前正在发生，是第三次工业革命的延续，但技术创新的速度、广度和影响力以指数级别上升。第四次工业革命以数字化和智能化为主要特征，标志性技术有物联网、大数据、人工智能等，尤其是以深度学习为代表的人工智能技术的突破，使人类逐渐从数字社会迈向智能社会。

高效的数字基础设施是数字社会、智能社会的关键支撑。面向2030，产业互联、全息通信、元宇宙、自动驾驶等新型应用对信息通信技术提出了更高的需求。但应该看到，信息通信技术的发展是以19世纪末到20世纪中叶人类在电磁学、量子力学、信息论等数学、物理学的突破为基础的，而近几十年来，基础科学的突破有放缓的迹象，这使得信息通信领域的未来技术发展面临越来越严峻的挑战。传统的技术演进路线面临摩尔定律、香农定理的限制以及节能减排的约束，亟需在基础理论、核心算法和系统架构方面有根本性的创新。国际形势的风云变幻，也对技术自主创新提出了很高的要求。

本白皮书是对数字基础设施的未来技术发展趋势的解读，由中兴通讯技术专家委员会集体编写完成。与业界常见的以商业模式、应用愿景、技术需求为主的白皮书不同，本技术白皮书更多地把重点放在技术发展面临的挑战，以及解决这些挑战的技术实现路径上。

白皮书第二章介绍了未来的业务场景对于数字基础技术的需求，提出了数字基础设施最关键的三个技术要素是连接、算力和智能。但是这三个技术要素的发展，面临着香农定理极限、摩尔定律放缓和智能本质认知不足的问题，使得未来的技术进步面临很大的挑战。

第三~五章分别针对更宽的连接、更强的算力、更高的智能三个方向进行了具体技术趋

势的描述。每个技术方向均描述了未来的技术需求和面临的技术挑战，以及解决问题和挑战的技术发展路径。既有对业界发展现状和主流观点的描述，也有中兴通讯的技术创新和对未来发展的预判。

第六章是整个白皮书的总结，同时也针对数字基础设施能力如何更好服务于各行各业提出了我们的一些思考。

中兴通讯的技术创新瞄准技术发展趋势、行业发展方向和国家重大需求，依托于政、产、学、研、用各方的协同与支持，目前已经取得了丰硕的成果。我们将继续与产业伙伴一起，共同推动信息通信领域的技术创新工作，为人类迈向数字社会和智能社会作出贡献。

2. 数字基础设施技术发展的需求和挑战

2.1.未来业务发展对于数字基础设施技术的需求和约束

自从 1837 年美国人摩尔斯发明摩尔斯电码和电报以来，信息通信技术迅速发展，极大地改变了人类生活、生产的方式。通信业务从最初单一的电报、电话到现在涉及人类经济、社会、生活的方方面面。全球数字经济规模持续上涨。2021 年，全球主要的 47 个国家数字经济增加值规模为 38.1 万亿美元，同比名义增长 15.6%，占 GDP 比重为 45.0%[01]。2022 年，我国数字经济规模达到 50.2 万亿元，同比名义增长 10.3%，已连续 11 年显著高于同期 GDP 名义增速[02]。

九层之台，起于累土。高效的数字基础设施是数字经济的核心基础能力。在 ToC 和 ToH 领域，新冠疫情带来的工作和生活方式的变化、短视频及直播等应用的爆发、在线教育和远程办公的普及，对于网络带宽和覆盖提出了更高的要求；在 ToB 领域，从 ICT 向 OT（生产域）的纵深拓展和贯通融合，也对网络性能、经济便捷和安全可靠等提出更高的期望。

2023 年 2 月，中共中央、国务院印发了《数字中国建设整体布局规划》[03]（以下简称《规划》）。《规划》明确指出，数字中国建设按照“2522”的整体框架进行布局，即夯实数字基础设施和数据资源体系“两大基础”，推进数字技术与经济、政治、文化、社会、生态文明建设“五位一体”深度融合，强化数字技术创新体系和数字安全屏障“两大能力”，优化数字化发展国内国际“两个环境”。

《规划》指出，打通数字基础设施大动脉，有如下具体要求：加快 5G 网络与千兆光网协同建设，深入推进 IPv6 规模部署和应用，推进移动物联网全面发展，大力推进北斗规模应用；系统优化算力基础设施布局，促进东西部算力高效互补和协同联动，引导通用数据中心、超算中心、智能计算中心、边缘数据中心等合理梯次布局；整体提升应用基础设施水平，加强传统基础设施数字化、智能化改造。

根据《规划》的要求和中兴通讯对于行业的理解，数字基础设施有三个基本要素：连接、算力、智能。数字基础设施的目标就是，网络无所不达、算力无所不在、智能无所不及。

“连接”是互联网最核心的特征，连接速率从最初电报的每秒约 1 个字符，到现在的“双千兆”接入（即无线接入和光纤接入均达到千兆）和骨干网单光纤几十 Tbps 的速率。无线通信网络基本每十年进行一次更新迭代，速率提升约 10 倍。面向 2030（6G），随着全息通信、元宇宙等新业务的发展，预计业务对于连接的需求相比目前（5G）仍将增长 1~2 个数量级[04]。其中，带宽峰值速率将达到 1 Tbit/s（50 倍），用户体验速率达 20 Gbit/s（200 倍），时延可低至 0.5ms（8 倍），连接密度达到 100 个/m2（100 倍）。

“算力”在数字社会已成为像水电煤一样的基本设施。据 IDC&浪潮信息&清华全球产业院的评估，计算力指数平均每提高 1 个点，数字经济和 GDP 将分别增长 3.5‰和 1.8‰[05]；据中国信通院统计，2021 年全球计算设备算力总规模达到 615 EFlops，预计 2030 年达到 56 ZFlops，平均年增速达到 65%[06]。算力是实现其他技术需求的关键要素。比如通信容量的提升需要有各种编解码计算；视频领域的 AR/VR、全息等业务，需要视频编解码、图像渲染、动画生成等高计算量的技术；近十多年来广泛应用的人工智能技术对于算力的需求是前所未有的。

随着近十多年深度神经网络算法的突破，人工智能技术不断拓展应用的深度和广度，已经成为人类社会从数字化向智能化迈进的强大引擎。数字化的前提是用数学模型表示物理世界，数学建模是算法和软件的基础。而 AI 技术突破之前，现实世界有大量的复杂系统无法用数学模型表示。深度神经网络技术的本质，是用简单神经元节点的大规模互联来逼近各类复杂系统的数学模型（比如人类认知系统或者高度非线性的物理系统），极大拓展了数字化的应用广度和深度。从物理层链路的非线性补偿，到网络层资源的智能化调度，再到应用层的视频处理、人机交互、安全态势感知、自动驾驶等等。智能成为最为关键的数字化基础技术之一。

由此可见，连接、算力、智能是未来数字化应用的基本技术需求。未来数字社会的根基是融合的算网基础设施及智能化服务体系。在数据洪流对端、边、云的冲击之下，连接、算力、智能这三者相辅相成，体现出更加紧密的关系和更加模糊的边界，以实现海量数据的存储、交换和处理的全局效益最优。

连接、算力、智能也是实现其他技术需求的关键要素。比如，安全技术需求，最基本的技术就是各种加密解密算法和计算部件；可靠性需求的技术基础是器件、组件、系统的失效期计算，以及各种系统冗余、网络冗余的算法。目前 5G 的 URLLC 场景已经可以提供 4 个 9（99.99%）的可用性，未来要满足工业场景 5 个 9（99.999%）的可用性需求，可能的办法是引入 AI 算法实现信道预测、故障预测、干扰跟踪等技术手段。

图 2.1 给出了未来各种应用场景与三大技术要素的对应关系。这些场景取自于国际电信联盟（ITU）网络 2030 焦点组于 2020 年 6 月提出的未来网络 12 个应用场景[07][08]。

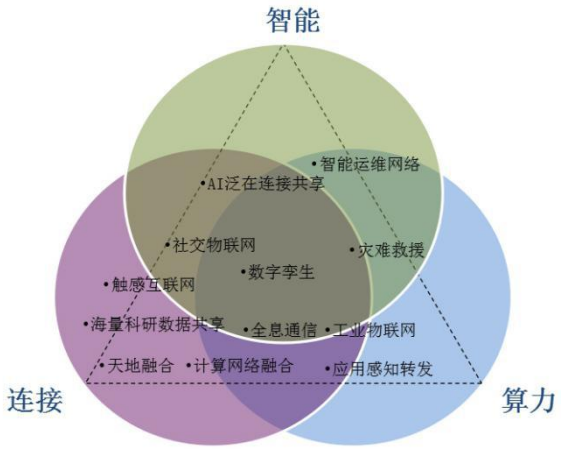

图 2.1：未来业务场景与三大技术要素的对应关系

与此同时，人类可持续发展的目标对于节能环保提出了越来越高的要求。各个国家都提出了双碳目标（碳达峰、碳中和）。尽管信息高效交互和处理可以提升物流和能量流的效率，从而降低整个人类活动的碳排放水平。例如 GeSI（Global Enabling Sustainability Initiative）发布的报告《SMARTer 2030》指出，ICT 技术有可能在 2030 年前帮助全球减少碳排放 20%[09]。但信息通信行业自身的碳排放仍然是不可忽视的一个方面。据上述 GeSI 报告的估算，预计到 2030 年信息通信产业的碳排放占全球总碳排的 1.97%。未来的技术发展，数字基础设施的建设，除了业务驱动外， 还必须把节能减排作为一个主要的约束条件。

另一个重要的约束条件是国际政治形势对于全球技术融通和共享的影响。未来的技术供应链、全球的技术协作都很可能面临越来越多的困难。

2.2.传统的技术发展路径面临的挑战

从 19 世纪后期到 20 世纪中叶，人类在电磁学、量子力学、信息论等数理理论的突破，是现代信息通信技术的基础。信息通信技术三大要素，连接、算力、智能，既有各自的发展路径，又呈现相互支撑、协同并进的态势。

香农信息论从理论上证明了，带宽恒定的信道在有噪声情况下存在传输容量上界，即香农极限。通信带宽的提升，一方面是研发更优的传输信道（无线、电缆、光纤等），另一方面是用不断演进的算法（调制解调、整形补偿、前向纠错等）逼近香农极限。先进的算法带来了计算复杂度的增加，必须依赖微电子技术的进步，以更强的数字信号处理能力满足通信算法的需求。

计算机出现之后，对算力的需求成为微电子技术进步的最大驱动力。著名的摩尔定律最初就是对微处理器发展的预测，即，每隔 18 或 24 个月，单位面积的芯片上所集成的晶体管数量翻一番，微处理器性能提升一倍，价格减半。

按摩尔定律发展的微处理器在过去 50 年来性能提升了 10 亿倍以上，其带来的半导体工艺技术的进步也同时带动了其他芯片的发展，包括用于通信的数字信号处理器（DSP）、网络处理器、交换芯片等。芯片技术的进步使得更加复杂的通信算法得以实现。

人工智能算法对于算力的需求前所未有，除了研发性能越来越强大的 AI 算力芯片之外，分布式计算也是解决大算力需求的必由之路。分布式 AI 计算对于连接带宽、时延提出了很高的要求，连接技术的进步使得跨地域的云-边高效协同的分布式 AI 计算成为可能。

反过来，连接带宽的提升也需要人工智能技术的支撑。目前 AI 算法已经在网络物理层优化、无损网络参数优化、神经网络视频编码等各层面发挥作用。

由此看出，连接、算力、智能这三大技术要素，每个要素的发展都依赖其他要素的支撑，反之，任一个技术方向的受阻，都会影响其他技术方向的演进。

目前看，这三个技术要素的未来发展路径均面临各自的困难。

（一）通信算法已经逼近香农极限

香农定理，*即 $C/W=log_2(1+S/N)$*，揭示了谱效（单位频谱宽度能传送的最大信息速率）与信噪比的关系。在通信实践中，经常根据信息速率（不含冗余信息）、通道宽度，来

确定最小信噪比容限。这个信噪比容限代表可实现无误码传输时所需信号质量的极限值。

早在 2001 年就有研究论文指出[10]，目前无线通信中使用的 LDPC 编码算法，在白噪声信道中达到 10^{-6} 误码率的信噪比容限，是 0.31dB，而此场景下的香农理论极限是 0.18dB，两者只差 0.13dB，也就是实际信噪比容限比香农限只多了 3%（$10^{0.013}$≈1.03）。

另外，据中兴通讯实测数据，目前光传输算法在 4~16QAM 调制达到的信噪比容限，距离香农限约为 1dB 左右，只能等效提高 25%的传输距离，或提高 0.33bps/Hz 的谱效。

越接近极限，提升算法复杂度所带来的性能收益越低，往往用数倍的计算量，才能带来几个百分点的性能提升。因此未来的算法即使能够继续逼近香农极限，其对于算力效能的需求（单位功耗的算力）也大大超过了摩尔定律能达到的水平。

（二）微电子技术逐渐逼近物理学设定的边界

代表微电子技术进步的摩尔定律也逐渐遇到越来越大的困难。

在 28nm 工艺之前，业界通过缩小晶体管尺寸（如栅长）来增加单位面积的晶体管数量。但晶体管尺寸小到一定程度就会因为量子隧穿效应、寄生电容等问题而难以为继(一个硅原子的直径为 0.2 nm， 20 nm 的栅极长度大约只有 100 个硅原子)。因此从 22nm 开始，主要靠晶体管结构创新来增加单位面积的晶体管数量，从鳍式结构 FinFET（预计可以支撑到 3nm）到全环结构 GAA（预计可再支撑 2~3 代）。但是复杂的结构创新所需要的成本和功耗很高，因此继续提升工艺节点的经济性越来越成为限制因素。

目前看来，人类在微观世界的基础理论进展（量子力学）所带来的技术红利已经接近用尽。从二十世纪五十年代开始，以量子力学为基础的现代光学、电子学和凝聚态物理迅速发展，诞生了激光器、半导体和原子能等重大科技突破。但这次技术革命还只是从宏观统计的角度认识和利用量子现象，例如能级跃迁、受激辐射和链式反应，对于物理介质的观测和操控依然是宏观的手段和方法，例如电流、电压、光强等。

微电子技术的物理学基础是基于电荷数量的电平检测，随着电荷团持续缩微，带来了静态泄漏电流增大等难以解决的问题。业内在研究将“电荷团迁移”改为“材料的阻变”、或“量子自旋”等新物理机制来降低功耗，但操控的仍然是宏观物理量，仍将面临尺寸缩微的极限。

要全面展示量子态的特性，充分释放量子的潜力，就必须通过对光子、电子和冷原子等微观粒子系统及其量子态进行精确的人工调控和观测。科学界在这方面的研究尚处于起步阶段，未来的发展路线、方式、目标都存在很大的不确定性。

（三）智能的发展，缺少认知科学理论的指导

人工智能是对人类思维过程和智能行为的模拟。对人工智能的研究从计算机诞生不久的20世纪50年代就已经开始，真正在应用上取得突破是在2006年深度神经网络取得成功之后。但基于神经网络的人工智能算法是对人脑生理结构的浅层模拟，对于其深层工作机理的研究属于认知科学的范畴，目前尚未取得突破。

目前人工智能算法已经在多个难以用数学模型描述的复杂问题域取得应用，包括人类智能相关领域，和非人类智能相关领域。在人类智能相关领域，比如机器视觉、自然语言处理、自动驾驶等，虽然在某些场景下已经非常逼近甚至超过人类的水平，但在通用性方面离人类智能仍然有差距。业界逐渐认识到，人工智能算法要真正达到人类智能的水平，需要在认知科学基础理论有突破性的进展。在非人类智能相关领域，比如通信物理层的神经网络建模和补偿算法，虽然取得了一些学术研究成果，但由于缺少对其内在物理机理的了解，使得业界对于这种模型的适应性和可靠性存疑，阻碍了其真正在工程上的大规模应用。

目前的深度学习技术路线是靠算力和数据的大规模堆积来实现的，但在摩尔定律放缓，节能减排约束的背景下，这种技术路线从长期看是难以持续的。目前人工智能算法对算力需求增速远大于摩尔定律增速，特别是进入Transformer大模型时代后，训练模型所需的算力增速提升到平均每两年增长275倍，大幅超越摩尔定律每两年增长2倍的增速[11]。这种增速背离带来的直接结果，就是AI计算成本和对环境的压力快速上升。目前，全世界1%的发电量被用于AI计算。全球AI计算能耗年增长率为37%。据此估算，下一个10年，AI计算将消耗全世界发电量的15%左右，将为环境带来沉重的负担。

总的来说，信息通信技术的发展已经逼近数学（香农定理）、物理（量子力学）、认知科学三大基础理论所设定的边界。往前每走一步都要付出比以往更大的代价。而节能降耗的要求、技术效用的边际递减又给技术演进的经济可行性设置了更高的障碍。

浅层的矿藏已经挖完，深层的矿藏是否具有技术经济性尚属疑问。这是目前的技术演进路线面临的主要问题。

2.3. 未来技术发展路径概述

上一节提到，连接、算力、智能这三大技术要素的演进都碰到了瓶颈，技术提升所需要的成本、功耗与带来的收益越来越不成比例。而未来业务需求对于带宽、时延、算力的需求仍然是数量级的上升。如何突破技术瓶颈，打造“连接+计算+数智能力”的数字底座，是当前面临的重大课题。

本白皮书第 3 章~第 5 章分别从“更宽的连接”、“更强的算力”、“更高的智能”三个方面描述未来技术发展的可能路径。

美国思想家布莱恩·阿瑟在《技术的本质》一书中提出，技术的本质是被捕获并加以利用的现象的集合；技术的进化类似于生物进化，是一种组合进化的过程，新技术是已有技术的新组合。我们认为，未来的技术发展，除了继续单点突破，挖掘现有技术路径的潜力外，另一方面也将更多聚焦到多种技术的协同、系统架构的优化。

信息通信系统的架构，不管是计算架构还是网络架构，都是以模块化、分层、解耦为特征，比如冯·诺依曼计算架构的特征是计算和存储分离；网络架构采用协议分层和层间解耦的设计。分离和解耦的好处是各个模块独立发展，便于创新和维护。但模块间简单化、通用化的协作与接口，对于某些特定业务来说并不能达到性能最优。在单个模块的性能提升遇到瓶颈的时候，往往需要模块间的协同和融合去带来性能和功耗的收益。

在后面章节描述的技术路径中，既有现有技术路径的挖潜，比如无线、有线通信中开发新的频谱和信道；也有多种技术的协同和耦合，比如光电集成、存算一体、算网融合等等。

表格 2-1 是对于连接、算力、智能三个方向的技术发展路径的概述。

表 2-1 未来技术发展路径的概述

	单点纵深突破	立体协同耦合
更宽的连接	提高频谱效率继续逼近香农极限；扩展频谱带宽；空分复用	光电集成； 分组转发芯片架构创新
更强的算力	More Moore：半导体工艺继续缩微路线	算、存、网从分离到融合：存算一体、对等系统、算网融合
更高的智能	AI 芯片架构创新，实现更高算力/能耗比；AI 算法从多样化的分离小模型向通用大模型演进；	智能化能力向数字基础设施自身、行业、企业赋能；

3. 更宽的连接

3.1.综述

万物互联是数字化时代的主要特征。提升连接带宽是信息通信技术的主要追求目标。目前，无线接入（5G）和有线接入（10GPON）已经具备向用户提供“双千兆”接入带宽的能力。骨干光纤网开始部署长途 400G 单波传输技术。如第二章所述，未来 5~10 年，业务对于带宽的需求仍将有 1~2 个数量级的增长。

网络带宽提升不仅是物理层传输带宽的提升，还涉及分组层、应用层等数据处理能力的提高。同时，机架及设备内部互联的带宽也需要相应提升。

（一）物理层

通信物理层以电磁波理论为基础，电磁场表达式如以下公式所示：

$$\overrightarrow{E(x,y,z,t)} = \underbrace{\overrightarrow{e_p}}_{\text{偏振}} \cdot \underbrace{F_d(x,y)}_{\text{空间分布}} \cdot \underbrace{|A_m(T_S)|}_{\text{幅度}} \cdot \underbrace{exp[j\varphi_n(T_S)]}_{\text{符号周期 相位}} \cdot exp[\underbrace{j(\omega_k t - kz)}_{\text{波长}} + \varphi_0]$$

根据电磁场表达式，通信可复用的维度有偏振、空间分布、幅度、相位、波长和符号周期（波特率），一般把幅度和相位合称为 QAM 调制（Quadrature Amplitude Modulation，正交幅度调制），因此一共是 5 个维度。其中偏振、QAM、波特率与单波速率有关。因此总传输速率可以写成：

$$\text{总传输速率} = \underbrace{D}_{\text{空分}} \bullet \underbrace{K}_{\text{波道数}} \bullet \underbrace{R_s}_{\text{单波道速率}} \le D \bullet K \bullet \underbrace{B}_{\text{单波道带宽}} \bullet \underbrace{\log_2(1+SNR)}_{\text{信道香农定理}} = \underbrace{C}_{\text{信道容量}}$$

注：此公式是一个简化表示，无线领域涉及多个空分通道的计算公式较复杂

由公式可知，提升传输容量可通过提升单波速率、波段扩展和空分复用，合计 3 个技术途径。其中单波速率受限于香农定理，提升单波速率一方面通过高阶调制、偏振复用等技术提升谱效，逼近香农限，另一方面通过提高波特率，从而提升单波带宽；波段扩展是指增加通信可用的频段；空分复用通过多天线或者光纤的多个纤芯、模式，增加通道数量，实现容量数倍提升。

表 3-1 是无线通信、光通信对于上述 5 个维度、3 个技术途径的发展现状和未来技术路

径的简要总结。详细的描述分别见第 3.2 节和 3.3 节。

表 3-1：无线通信、光通信的技术发展现状和未来技术路径

5 个维度	三个技术途径	无线	光传输（长距）
幅度和相位（QAM）	单波提速（单自由度提速）	● 现状：1024QAM 已经标准化，但还没有商用；调制阶数越高，距离香农限距离越远 ● 趋势：更高调制阶数，提高星座成形增益，编码调制联合优化	● 现状：相干 4~16QAM 调制编码已逼近香农极限；波特率 64~128GBd ● 趋势：继续提升单波带宽（波特率）、新型光纤/放大器与 DSP 均衡算法提升 SNR（支持更高阶 QAM）
偏振/极化			
波特率			
波长	波段扩展/频谱使用效率提升	● 现状：已有 200MHz 载波聚合；子带全双工正在标准化 ● 趋势：载波聚合，全双工技术、毫米波/太赫兹、多址/复用	● 现状：C+L 波段 12THz 谱宽支持 80 波*400G/800G ● 趋势：向 S+C+L 波段扩展
空间	空分复用	● 现状：业界公认提升容量的最主要方法之一；64TR/16 流已商用；NCR 正在标准化 ● 趋势：eMIMO/Beam 、分布式 MIMO，超大孔径 ELAA，Cell-free、RIS、NCR 等	● 现状：尚未商用 ● 趋势：多芯少模；多芯弱耦合或率先商用

（二）分组层

互联网诞生以来，以 IP、以太网为代表的分组技术就是网络技术的核心。网络设备的分组处理能力往往成为提升网络容量和性能的瓶颈。分组处理需要兼顾容量和灵活性，其整体性能的提升，不仅依赖芯片工艺的进步，也跟分组处理芯片架构的改进密切相关。随着带宽增长，分组芯片容量每 2~3 年增长一倍，仅靠工艺进步难以满足芯片容量、成本、功耗的要求，需要芯片架构优化和算法优化，以兼顾容量、灵活性和低时延需求。本白皮书第 3.4 节描述了未来分组转发芯片架构的演进方向。

（三）应用层

在应用层提高信源压缩率的努力也与通信容量紧密相关。香农定理除了证明了信道编码的上限，也证明了无损信源编码的压缩率上限（与信息熵有关），但对于有损的信源编码，并没有压缩率的极限。随着 XR、全息等应用的发展，到 2030 年，视频流量预计将占到互联网总流量的 90%以上。因此如何通过视频编码技术进一步提高压缩率，是降低整个网络带宽压力的重要研究方向。本白皮书第 3.5 节描述了基于深度学习的视频编码提升视频压缩率。

（四）设备互连

随着外部链路带宽和端口密度的提升，通信和计算设备内外部互连总线也将成为瓶颈。光互联相对电互联在性能和功耗的优势凸显，随着光电共封装（CPO，Co-Packaged Optics）技术的日渐成熟，设备内部也将出现“光进铜退”，同时，光互连在空间距离上的损耗非常小，将推进信息通信设备的形态向分布式和大容量的方向演进。具体见 3.6 节。

3.2. 物理层（无线）：5G-A & 6G 需要更大力度的空分复用及扩展频段

自上世纪 80 年代以来，移动通信基本上以十年为周期出现新一代变革性技术，从 1G 逐步发展至现在的 5G， 目前 5G 已经在全球范围内开始大规模部署，各国更是已将 6G 列入未来几年的国家计划。

如 3.1 节所述，无线提升通信容量的方法有：提升单个自由度的频谱效率、增加带宽或带宽利用提效、增加空域复用阶数等。4G 采用高阶调制提升编码效率，引入 MIMO 多天线技术提高信道容量，引入载波聚合获得更大的频谱带宽。

5G 为了实现更高网络容量，主要采用两种方法，其一是继续增加天线数量，采用大规模天线阵列（Massive-MIMO）和超密集组网（UDN）。其二是拓展 5G 使用频谱的范围，从 4G 的 Sub-3GHz 频谱，扩展到 5G 的 Sub-6GHz 频谱。

如第二章所述，面对未来的业务需求，6G 在几个核心的指标，比如带宽、时延、可靠

性等，相比 5G 都有 1~2 个数量级的提升。3GPP 第一个 6G 版本预计会在 2030 年左右出现，在这之前还有 3 到 4 个版本的演进聚焦于 5G 增强技术“5G-Advanced”。

从频谱效率看，虽然 5G 技术低谱效到中等谱效下已逼近单链路的香农限，但在高谱效区距离香农限还有一定距离。另外，6G 需要在增加带宽或提升带宽利用率、增加空域复用阶数上更大力度地下功夫。包括载波聚合、全双工、更高频谱（Beyond 6G 和太赫兹）、非正交/OFDM 及其变形/高频波形/感知波形、超大规模天线和极致 MIMO、RIS（智能超表面）技术、网络控制 Relay（NCR）等。在上述技术中包含基础性的技术趋势，即：用越来越强的计算能力，尤其是底层的计算能力，去换取资源利用效率的提升。

下面针对几项典型技术分别进行阐述。

（一）高阶调制/星座整形/编码调制联合方案

当前调制方式可以到 1024QAM，每个符号可以携带 10 比特。6G 为了进一步提升谱效，可能会将调制阶数提升到 4096QAM 甚至更高。在高阶调制方式下，传统的方形 QAM 星座图的效率不是最优的，可能会导致谱效越高，距离香农限越远的情况。基于几何整形/概率整形的高阶调制、编码调制联合方案有望更加逼近香农限，尤其在高 SNR 区域。

（二）提升频谱使用效率：全双工及子带全双工

全双工是提高网络数据速率以及频谱使用效率的新技术。对于未来大带宽低时延的业务，全双工使用非成对频谱，通过释放 DL/UL 资源使用上互斥的限制，能够增强频谱使用效率和减少传输延时。但实现全双工需要基站或终端能够处理自干扰（SI）以支持同时进行的收发信功能，实现复杂度以及硬件代价还是相当大的，尤其是对于多天线 Massive MIMO 机型。因此实际上多天线技术与全双工有一定互斥性。

目前的研究大多是从天线数较少的机型和子带全双工开始，即在带内不同频率上分别配置上行和下行资源。这种方法能够灵活地配置更多上行资源，有助于降低上下行时延、提升上行覆盖和容量。虽然子带全双工降低了基站内部干扰消除能力要求，但是 UE 间的互干扰还是很严重，需要业界共同努力。

（三）扩展更多频谱：太赫兹技术

作为 6G 潜在的基础技术，太赫兹是指 100GHz～10THz 的频段资源，具有连续可用的大带宽，将有助于构建 6G 短距离、高速率的传输系统，支持超高速率的数据传输，满足超密集设备的连接需求，增强网络连接的可靠性，并支撑高能效的终端网络。

但太赫兹的缺点也比较明显。相比于毫米波，太赫兹频率的提高使传播路径损耗明显增大，室外通信在受到雨雾天气影响时也会带来额外损耗。此外，发射机功放功率低、低噪声放大器噪声系数高、高增益天线设计加工难度大等都极大地限制了太赫兹波的传输范围。

通过与多天线技术结合，太赫兹可借助极窄波束来克服路径衰落问题和扩展传播距离。此外，将 RIS（智能超表面，见第五条） 应用于太赫兹频段是未来的技术发展趋势。将 RIS 密集地分布在室内和室外空间中会对太赫兹覆盖空洞产生积极作用。

（四）更大力度的空分复用：超大规模天线和分布式 MIMO

超大规模天线能有效增强上行容量和提高新频段的覆盖性能。

一些新兴的工业互联网应用，例如现代工厂中的机器视觉类应用，对上行带宽的要求远高于下行带宽，极端情况下需要满足 Gbps 或 10Gbps 数量级的吞吐量。解决方案之一是 NR（5G 空口）的上行支持更多的天线或 MIMO 层数、支持更多用户的 MU-MIMO，以及更灵活的载波分配和聚合等。5G-Advanced 能支持最多 24 正交的解调参考信号 DMRS（DeModulation Reference Signal）端口，如果每个用户支持单流上行传输，在共同的时频资源可以支持最多 24 个用户的上行传输；此外，5G-Advanced 支持上行能力更强大的终端，单用户可以支持到 8 流，能大大提高峰值速率，在密集网络部署的情况下，有效提高上行吞吐量。

新增频段也对天线数目有要求。2023 年世界无线电通信大会（WRC-23） 将对 6Ghz-10Ghz 做分配，这些频段具有波长短，传播损耗大特点。为了提升这些频段的覆盖性能，系统设计时需要考虑增加天面，以增强天线增益，降低功耗和网络成本。我们认为由于芯片集成度提升，设备成本将快速下降，未来趋势是通过利用更多数字通道，改善赋形或者让接收波束更窄，以大幅提升无线性能。

另一方面，更大力度的空分复用未来发展趋势是分布式程度越来越高，等效孔径越来越大。从包含少量接入点（AP）的 MTP/ eCoMP，发展成为更大规模的异质分布式 MIMO，进而发展为超大 AP 规模的 Cell-free 网络。大规模的分布式 MIMO 需要解决时频同步、前传带宽和 AP 供电等问题。

（五）提升信道覆盖质量：智能超表面技术 RIS

智能超表面（Reconfigurable Intelligent Surfaces，RIS）是一种无线环境优化技术，具有低成本、低能耗、高可靠、大容量的特点。RIS 主要通过以下几种方式提升小区边缘用户的覆盖、吞吐量和能量效率：

(1) 在直射传播路径受阻时提供有效的反射传播路径，避免覆盖空洞；

(2) 为目标用户进行波束赋型，充分利用空间分集和复用增益；

(3) 对干扰用户进行零点波束赋型，实现小区间干扰抑制。

RIS 本质上是一种分布式空分复用技术，通过多天线技术更好地、更精细地利用空间电磁场特征，从而提升无线性能。相比之下，Massive MIMO 是一种集中式空分复用技术；RIS 由于成本低，容易做到更大规模，容易进行多点部署。

目前业界对 RIS 的优点已认识比较充分，对其所带来的问题，尤其是规模组网时的部署和性能问题认识得并不充分。最近随着研究的深入，越来越多的问题被认识，例如目前 RIS 器件调控精度低，导致 RIS 口径效率低，存在栅瓣，会引起用户间和网络间的干扰；另外，目前还缺乏基站与 RIS 间的标准控制接口，仍以静态/半静态调控居多，只能改善固定用户的覆盖性能，后续需要实现基站与 RIS 动态协同技术，以提升更广域场景用户的覆盖性能和移动性能，等等。这些问题有的造成应用上的限制，有的导致成本上升或性能下降，上述问题需要产业界和学界共同攻克，需要客观估计 RIS 的适用场景和产业化进程。

3.3.物理层（光传输）：单波提速、波段扩展和空分复用

高速、大容量和超长距是光传输的最重要需求。随着业务流量的增长，导致光传输系统单纤容量增长的压力越来越大。同时，在干线应用时，传输能力要求超过 1000km 以上。目前采用 super C（超宽 C 波段）的 200G PM-QPSK（偏振复用四相相移键控）系统已经广泛商用，400G PM-QPSK 预计 2023 年开始商用。

如 3.1 节所述，提升通信容量可通过提升单波速率、波段扩展和空分复用 3 个技术途径。单波提速同时降低单比特成本，是容量扩展最直接/经济方式，而新波段扩展(比如 L 波段与 S 波段)使得频谱成倍提升，可适配单波带宽与波道数的同步增长。随单纤波段继续扩展的性能/成本优势降低，空分复用或成为单纤容量持续提升的可选路径。

接下来就单波提速、波段扩展、空分复用就未来的演进趋势分别展开介绍：

（一）单波提速

根据香农定理，单波提速的途径一方面是谱效提升，另一方面是单波带宽/波特率提升。谱效的提升意味着对接收信号处的 SNR（信噪比）需求提升，而光路依托新型光纤(如 G.654 光纤、空芯光纤)降损耗/非线性，结合放大器降低噪声系数可显著提升信道 SNR，支持单信道容量成倍提升；先进相干 DSP 芯片采用高性能调制解调结合高编码增益 FEC（前向纠错编码），使得 SNR 容限继续趋近理论值，传输速率逼近信道容量上限；单波带宽提升方面，芯片与光器件带宽提升是单波提速与降比特成本主要方式，波特率从 64GBd 到 96GBd/128GBd，并将持续往 180GBd+演进。

（二） 波段扩展

随单波带宽提升与系统波道数要求，波段扩展是提升单模光纤容量主流方向。本着单波提速不减波数，容量翻倍的原则，长距模式 100G、200G、400G 分别应用了 C4T、C6T、C6T+L6T 的工作波段（如表 3-2 所述）。目前长距 400G 依托 C+L 波段逐步商用，长距 800G 结合 S+C+L 波段扩展将成为下一步容量提升方向。同时还需要推动规范 G.654E 光纤的截

止波长到 1460nm 以下，以及规范 U 波段的宏弯损耗和 E 波段低水峰光纤规格。

波段扩展依赖新波段光器件的材料工艺发展，以支持更宽波长范围。如放大器的 Tm/Bi 离子或基质掺杂工艺、光模块的 128GBd 以上 TFLN（薄膜铌酸锂）相干调制器、ITLA（可调谐激光器）多波段外腔技术、WSS（光波长交换）器件多波段增透镀膜设计等。

表 3-2 光纤通信的波段扩展

年度	2010	2017	2023★	2026
干线单波速率 bps	100G	200G	400G	**800G**
工作波段	C	Super C	Super C+L	**U/S+C+L**
工作波段带宽	4THz	6THz	12THz	**24THz/18THz**
光纤	G.652D	G.652D	G.654E	**G.654E**
干线单纤容量	8Tbps	16Tbps	32Tbps	**64Tbps**
波特率	32GBd	64GBd	128GBd	**256GBd/192GBd**
通道间隔	50GHz	75GHz	150GHz	**300GHz/225GHz**

（三）空分复用

通过空分复用技术，即光纤芯数和传输模式的增加，可以实现单纤容量的大幅提升。从技术路线上可分为多芯弱耦、多芯强耦、少模弱耦、少模强耦。其中多芯弱耦合光纤/器件相对成熟，具备长距传输能力。受限光纤维护问题与多芯光放性能，多芯弱耦相比多纤芯光缆的陆缆商用价值或不显著，但因多芯放大器的能耗与多芯光缆的尺寸/密度优势，更受海缆应用关注。少模弱耦合光纤/器件能力具备，受限光纤与连接器的模式间串扰，传输距离不足，在 DCI 短距互联或具有应用潜力；多芯强耦合光纤/器件能力具备，主要受限于多路复用 OTU（MiMo DSP 与相干光器件集成），短期无法实用；少模强耦合器件能力具备，主要受限光纤耦合度低与多路复用 OTU，尚不具备实用性。

3.4.分组层：兼顾容量和灵活性的分组转发芯片架构

在分组层（IP 层和以太网层），对于带宽影响最大的技术，是分组转发设备的处理能力，也就是通常说的吞吐量。分组芯片转发能力是提升网络带宽的关键。目前，业界已经发布 51.2Tbps 处理能力的芯片，按照 2~3 年芯片能力增长一倍的趋势，预计 2025~2026 年，单芯片处理能力将达到 102.4Tbps。2030 年，单芯片最大处理能力将有可能达到 204.8Tbps。

同时，未来十年，传统网络向算网融合的新型数字基础设施演进，分组芯片还需要提升业务灵活处理能力。一方面加强芯片可编程能力满足新业务创新；另一方面降低芯片转发时延，满足数字孪生、元宇宙等新场景对于低时延的需求。

要同时满足芯片容量、灵活性和低时延要求，未来芯片不仅依赖工艺技术的进步，更需要在架构设计和算法上进行创新。

当前主流可编程转发架构有两种：（1）并行 RTC（Run To Complete）架构；（2）串行流水线架构。并行 RTC 架构采用多个包处理引擎并行进行报文处理，每个微引擎完成完整的报文处理；串行流水线架构将报文处理分成多个阶段，由多个串行的包处理引擎接力完成完整的报文处理。

并行 RTC 架构具有大容量表项，超大指令空间，能够处理复杂业务，但该架构转发延时大，且在重载时，引擎间调度冲突加剧，网络抖动较大，无法满足低时延业务需求。串行流水线架构具有较小的延时和确定的抖动，但转发表容量较小，编程能力有限，无法处理复杂业务。当业务复杂度超过流水线容量时，只能通过环回处理，导致性能折半下降。

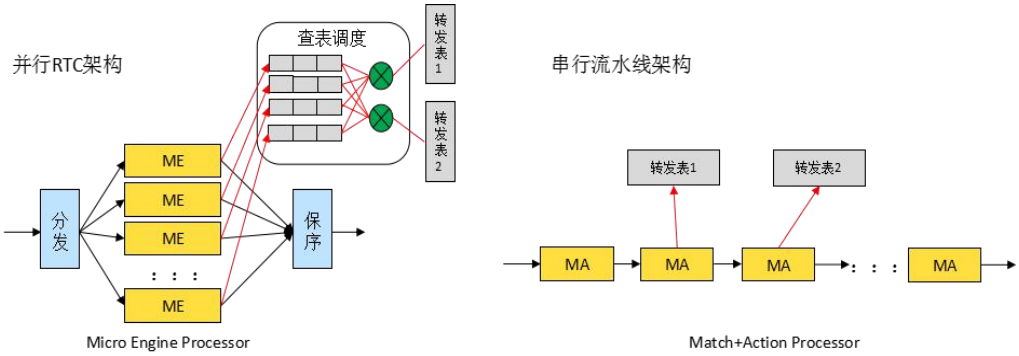

图 3.1 分组转发芯片的并行 RTC 架构和串行流水线架构

我们提出一种新型转发芯片架构：并行+串行的混合架构。该架构最主要的特点是：通

过编排将不同特点的业务分配到适合的转发架构上，同时满足未来网络性能、延时和业务扩展性的要求。低延时场景中，低延时业务全部由串行流水线处理，少量复杂业务（比如大容量转发表查找等），交由 RTC 并行架构处理。该场景下，由于 RTC 处理的业务较为简单，占用指令少，该架构依然能够保证较低的处理时延。对于其他场景（比如通用路由器），芯片要处理的业务非常复杂，涉及多级大容量表项的查找，并且不同表项之间有明确的前后级依赖关系。该场景下，需要将路由器业务进行适度分解，通过串行流水线和并行 RTC 混合路径处理，将复杂业务处理集中编排到并行 RTC 架构执行，由 RTC 集中完成大容量转发表的查找和复杂逻辑处理，从而有效解决单纯串行流水线编程能力不足的问题。

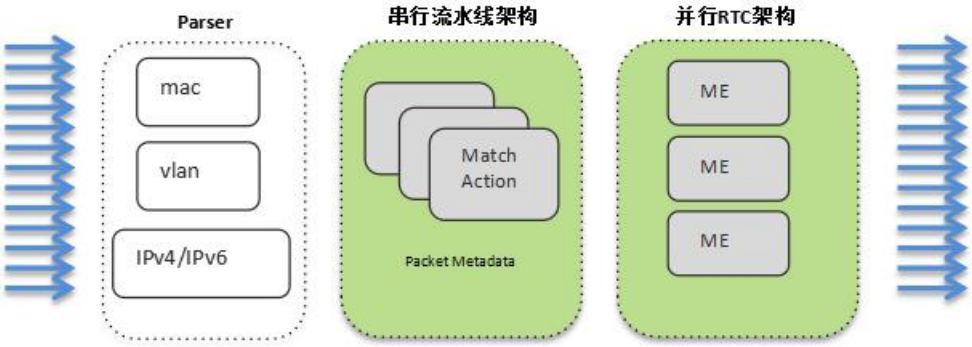

图 3.2 分组转发芯片的并行+串行混合架构

基于我们的技术评估，在选择合适业务排布模型的前提下，混合架构的转发时延与串行流水线架构基本相当，比 RTC 并行架构减少约 40%。混合架构的芯片面积比 RTC 并行架构和串行流水线架构少 15%-20%左右，功耗少 12%-20%左右。我们认为：未来的分组芯片引入串并混合架构，使芯片成为不同产品的通用平台，实现“一款芯片、多种产品”，满足多种业务场景需求必然逐渐被业界接受并推广。这不仅有利于降低芯片研制成本，而且在后继产品开发中，能显著缩短新功能开发和部署周期，快速适配用户持续变化的业务需求。

3.5.应用层：基于深度学习视频编码进一步提升视频压缩效率

视频是信息重要载体，据统计，2020 年视频流量已经占到整个互联网流量的 70%以上。对于视频内容编码质量以及相同视觉质量下压缩效率的追求，是视频技术发展的最主要驱动因素。视频编码目标是在可以接受的信息损失范围内尽量提高视频压缩效率，从而降低视频

传输带宽的需求，这是应对香农极限的另一个努力方向。

ISO/IEC JTC1 SC29 与 ITU-T SG16 VCEG 联合成立的视频专家组(Joint Video Experts Team, JVET) 于 2020 年 8 月发布新一代视频编码标准 H.266/VVC(Versatile Video Coding)[14]。H.266/VVC 沿用传统混合编码框架中的预测编码、变换编码和熵编码技术来降低视频空域、时域、频域、分量间以及视觉冗余，并引入了大量新的编码参数，更加准确地描述视频内容。相比上一代视频编码标准 H.265/HEVC，在相同视觉质量下可实现约 50%的编码码率节省。

然而随着更多样块划分方法与编码模式不断出现，更复杂预测与变换技术的不断引入，传统视频编码算法复杂度日益增长，完全基于传统编码框架技术来提升视频编码压缩效率也愈发困难。深度学习技术在图像分类、目标检测等计算机视觉任务上已取得了巨大的成功，近几年，深度学习技术为图像/视频编码框架定义了新的结构范式，实现了图像和视频编码器性能的显著提升，这赋予了图像/视频编码领域新的研究契机。

基于深度学习的视频编码技术主要包括：传统视频编码与基于神经网络编码相结合的混合视频编码技术，以及完全端到端的神经网络视频编码两大技术方向。

（一）传统编码与神经网络编码相结合的混合视频编码

混合视频编码旨在传统编码框架中引入深度神经网络来进一步提升压缩性能。其中一类技术与传统编码标准相兼容，利用深度学习策略实现对 H.266/VVC 和 AV1 等传统编码标准中的块划分、预测模式等大量待搜索对象快速判决，从而缓解编码端搜索压力、降低计算复杂度。另一类技术则属于仅实现压缩效率提升的非标准方案，其典型工具包括：超分辨率和后处理滤波。前者是在解码端对解码图像执行超分辨率操作，实现编码端低分辨率图像/视频输入，仍可获得高分辨率、高质量重建值，从而有效提升编码效率。后者则试图直接建立重建像素到原始像素之间的映射关系，通过滤波器策略实现对重建编码图像质量的提升。

（二）完全端到端的神经网络视频编码

端到端神经网络视频编码技术打破传统编码框架，完全使用深度学习方法实现编解码流程。基于神经网络视频编码利用海量数据集进行神经网络训练，学习去除视频压缩失真任务中的先验知识。强大的非线性变换和映射能力是端到端神经网络编码可获得更好压缩性能的一个主要原因。此外，端到端神经网络编码器针对整个编码环路进行端到端优化，可以避免

传统编码器中手工设计或独立优化所存在的局部最优问题，实现编码系统整体编码性能的提升。

基于深度学习的视频编码虽然相比传统视频编码可获得较大压缩效率提升，但其所带来的解码端复杂度大幅提升使得此类技术在短期内落地面临一定挑战。目前，业界厂商正在积极研究传统视频编码技术与基于深度学习视频编码技术联合优化。例如，JVET 开展的神经网络视频编码（NNVC）探索实验[15]和增强压缩视频编码（ECM）探索实验[16]，兼顾传统预测变换编码工具的压缩优势，以及深度神经网络智能编码工具的质量提升优势，测试结果表明，在 RA 和 AI 配置下，Y、Cb、Cr 三个通道 BD-rate 分别节省：{-21.17%, -32.29%, -33.05%}和{-11.06%, -22.62%, -24.13%}，具备演进成为下一代视频编码标准的技术潜力。

3.6. 设备互联：光互联将逐步渗透到设备间及设备内互联

更宽的连接，对于通信设备来说，意味着更大的单板容量、更高的通道速率和更高的带宽密度。同时，需要有更低的 bit 功耗和 bit 成本。目前，信息通信设备内部以电互连为主。比如，框式设备内部通过 PCB 电背板进行互连；盒式设备的主芯片和光模块之间也是电互连。随着速率的上升，电互连的高频信号插损（IL）变得很大。同时，由于回损（RL）、连接器的阻抗不连续和串扰（Crosstalk）等因素的影响，需要使用更加复杂的均衡算法和 FEC（前向纠错编码）来补偿信号损失，这就使得电互连的 Serdes（串行/解串行）部分功耗增加。光互连在容量和距离方面具有电互连无可比拟的优势，因此，随着数据速率的增加和连接密度的上升，设备内部的互连也出现光进铜退的趋势。

相比于电互连，光互连还有一个优点就是极大的拓展了设备的空间互连距离，在设备结构层面的表现就是解耦目前电互连机架中板卡之间的空间紧耦合关系，降低了散热和 SI（Signal integrity，信号完整性）的设计难度；在设备内部网络连接拓扑的层面上，表现为三级 CLOS 网络的基数（radix）可以增加很多，即，可以在一个三级 CLOS 架构下，让更多的交换板和更多的线卡进行互连，从而实现了超大容量（提高一个数量级）信息通信设备的

低成本、低时延和低功耗的连接方案[17]。

CPO（Co-packaged Optics）技术是减小光引擎（实现光收发功能的单元）的体积，将光引擎和主芯片共封装的技术，也是将光互连下沉到单板间互连及芯片间互连的关键技术。CPO 将带来功耗的减少、信号完整性的优化、成本的降低以及其它方面的收益。与面板可插拔光模块（FPP）相比，CPO 大幅缩短了主芯片和光器件（optics）之间的距离，显著地降低了成本和功耗。以 112G Serdes 为例，当 Serdes 的 PCB 长度从 1000mm（CEI-112G-LR），缩短到 50mm（CEI-112G-XSR）的时候，奈奎斯特频率处的插损从 28dB 变成了 10dB。相对应的 Serdes 功耗从 650mW 下降到 150mW，功耗大约节约了 75%[18]。对于 linear 链路的 CPO 来说，由于取消了内部的 DSP，可以更大幅度地降低整体成本和功耗[19]。

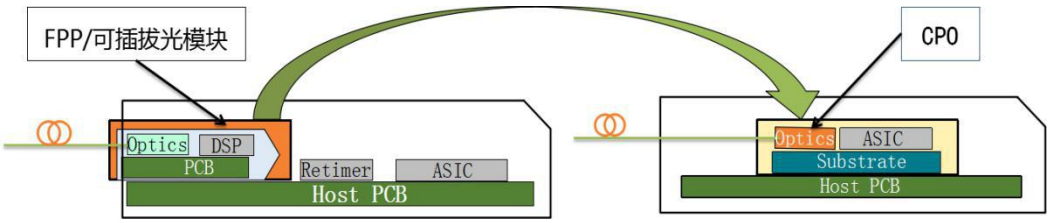

图 3.3 可插拔光模块向 CPO 演进示意图

共装低功耗、高密度和大容量的 CPO 是交换芯片的未来发展趋势。从 2010~2022 年，数据中心交换机带宽提升了 80 倍，功耗提升了 22 倍，其中光模块的功耗增长了 26 倍，交换芯片 Serdes 的功耗增长了 25 倍，两者的总功耗占到了交换机功耗的 70%[20]。面对功耗、SI 和成本方面的压力，行业各方都在大力推进 CPO 的标准化和产业化，预计 102.4T 容量阶段的交换机将是 CPO 规模部署的切入点。当然，面板可插拔光模块（FPP）也在不断随着各种新技术的发展而改进优化，尤其是近期 LPO（Linear-drive Pluggable Optics）受到比较多的关注，其在功耗和成本上，相比于非线性直驱的的可插拔光模块有比较大的优势，但是，要完全覆盖目前的非相干光模块场景还是比较困难的。由于技术之间的融合，LPO 也可以视为 CPO 技术的铺垫[21]。总之，在相当长的一段时间内，CPO 和可插拔光模块将会共存。

在 HPC（High Performance Computing 高性能计算）/AI 的网络和设备中，同样面临着功耗、成本和时延方面的巨大压力。Optical I/O 作为 CPO 的特定形态，在计算芯片 CPU、GPU 以及 XPU 等之间的互连(chip to chip interconnect)方面，具有低功耗、高带宽、低延迟的优势。从 Nvidia 评估的数据来看，GPU 在板内的电互连（PCIe 总线），功耗约为 6pJ/bit@0.3m，升级为全光互连后，功耗降低为 4pJ/bit，连接距离可以增加到 100m[22]。预计未来在 200G 的通道带宽下可以实现 0.1pJ/bit 的更低功耗[23]。由于近期 ChatGPT 的热度持续，预计 Optical I/O 形态的 CPO，将会首先进入规模商用，在 HPC/AI 的网络和设备中进行部署。

从目前产业链发展情况来看，Ⅲ-Ⅴ材料和硅基的异质集成和以 Chiplet 为重点的异构集成，将是 CPO 发展的有效途径。同时，CPO 和主芯片的 2.5D 集成甚至 3D 集成也将是重要的研究方向。CPO 发展的最终目标将是光电的单片集成，即在 wafer 级上，实现光功能模块和电功能模块的集成，这是光电集成的圣杯，也意味着巨大的挑战。

4. 更强的算力

4.1. 综述

随着人工智能、隐私计算、AR/VR 以及基因测试/生物制药等新型高性能计算应用的不断普及，对算力的需求也不断持续增加。比如，以 ChatGPT 为代表的大模型需要巨大算力支撑。大模型对算力的需求增速远大于摩尔定律增速。

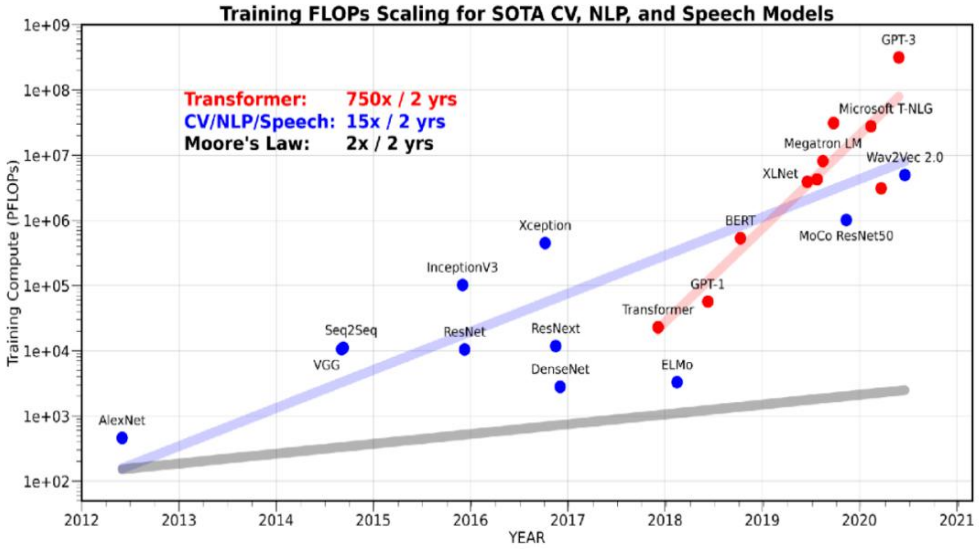

图 4.1 AI 大模型对于算力需求的增速远大于摩尔定律增速[11]

自微处理器诞生以来，算力的增长按摩尔定律发展，即通过增加单位芯片面积的门电路数量来增加处理器算力，降低处理器成本和功耗。但近年来这条路已经遇到越来越大的困难，通过持续缩微来提升性能已经无法满足应用的需求。

在后摩尔定律时代，一方面可以通过持续在工艺和材料上的创新来提升芯片算力：

- More Moore：继续追求更高的晶体管单位密度。比如晶体管工艺结构从鳍式结构 FinFET 到环形结构 GAA，以及纳米片、纳米线等技术手段有望将晶体管密度继续提升 5 倍以上。但这条路在成本、功耗方面的挑战非常大。
- Beyond CMOS：放弃 CMOS 工艺，寻求新材料和新工艺。比如使用碳纳米管、二硫化钼等二维材料的新型制备工艺，和利用量子隧穿效应的新型机制晶体管。但这条路径的不确定性较大，离成熟还需要很长时间。

另一方面利用架构的创新来提升算力密度，优化算力资源，从而延续摩尔定律。这也是本章要重点阐述的内容：

- 在芯片层面坚持领域定制的技术路线，进行软硬件协同协同设计，同时利用 3D 堆叠和 Chiplet 技术来降低芯片的设计和制造成本。（见 4.2 节）
- 引入新的计算架构和计算范式，如存算一体设计，在系统、体系和微架构层面进行计算、存储协同设计，从而实现高能效计算。（见 4.3 节）
- 采用“对等系统”等体系结构创新，优化计算、控制和数据路径，用全局最优替代局部最强，减少计算性能提升对先进工艺的依赖。（见 4.4 节）
- 在网络架构层面进行创新，通过算网融合提升算力资源调度效率。（见 4.5 节）

4.2. 芯片架构：DSA & 3D 堆叠 & Chiplet

图灵奖获得者 John Hennessy 和 David Patterson 在 2019 年共同发表的《计算机架构的新黄金时代》中提出：当摩尔定律不再适用，一种软硬件协同设计的 DSA（领域定制架构 Domain Specific Architecture）架构会成为主导，这种设计的核心在于针对特定问题或特定领域来定义计算架构。近几年火热的人工智能 AI 芯片和方兴未艾的 DPU 都是 DSA 领域的典型代表。

DSA 针对特定领域的应用采用高效的架构，比如使用专用内存最小化数据搬移、根据应用特点把芯片资源更多侧重于计算或存储、简化数据类型、使用特定编程语言和指令等等。与 ASIC 芯片（Application Specific Integrated Circuit，专用集成电路）相比，DSA 芯片在同等晶体管资源下具有相近的性能和能效，并且最大程度的保留了灵活性和领域的通用性。例如中兴通讯提出的计算和控制分离的人工智能领域定制芯片架构“夸克”，针对深度神经网络的计算特点，将算力抽象成张量、向量和标量引擎，通过独立的控制引擎（CE）对各种 PE 引擎进行灵活编排和调度，从而可以高效实现各种深度学习神经网络计算，完成自然

语言处理、AI 检测、识别和分类等各种人工智能应用。由于采用软硬件协同设计的定制化方案，DSA 芯片在相同功耗下可以取得比传统 CPU 高数十倍甚至几百倍的性能。

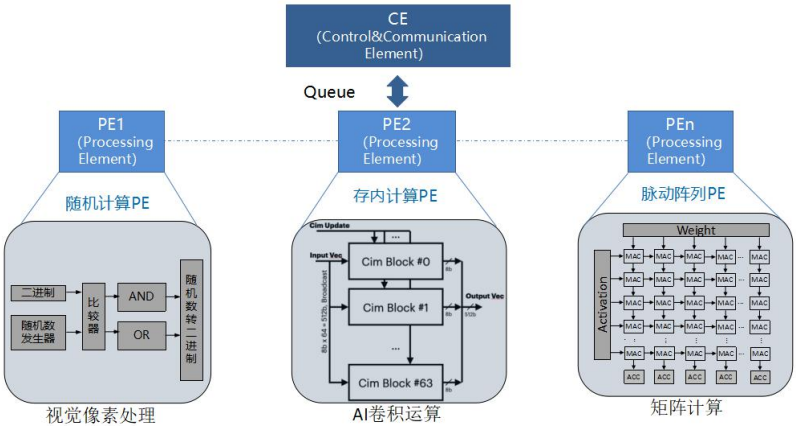

图 4.2 中兴“夸克”领域定制架构

摩尔定律本身是在 2D 空间进行评估的，随着芯片微缩愈加困难，3D 堆叠技术被认为是提升集成度的一个重要技术手段。3D 堆叠就是不改变原本封装面积情况下，在垂直方向进行的芯片叠放。这种芯片设计架构有助于解决密集计算的内存墙问题，具有更好的扩展性和能效比。

Chiplet 技术被认为是延续摩尔定律的关键技术。首先 Chiplet 技术将芯片设计模块化，将大型芯片小型化，可以有效提升芯片良率，降低芯片设计的复杂程度。其次，Chiplet 技术可以把不同芯粒根据需要来选择合适的工艺制程分开制造（比如核心算力逻辑使用新工艺提升性能，外围接口仍采用成熟工艺降低成本），再通过先进封装技术进行组装，可以有效降低制造成本。

与传统芯片方案相比，Chiplet 模式具有设计灵活性、成本低、上市周期短三方面优势。Chiplet 技术面临的最大挑战是互联技术，2022 年 3 月 2 日，“UCIe 产业联盟”成立，致力于满足客户对可定制封装互联互通要求。Chiplet 产业会逐渐成熟，并将形成包括互联接口、架构设计、制造和先进封装的完整产业链。

4.3.计算架构：存算一体使得计算和存储从分离走向联合优化

经典冯·诺依曼计算架构采用计算和存储分离架构，如果访存的速度跟不上 CPU 的性能，就会导致计算能力受到限制，即“内存墙”出现。谷歌针对自家产品的耗能情况做了一项研究，发现整个系统耗能的 60%以上花费在 CPU 和内存的读写传输上[24]。而读写一次内存耗费的能量比计算一次数据耗费的能量多几百倍。由于“功耗墙”的存在，大量的数据访问也会严重限制计算性能。随着大数据和人工智能应用的发展，传统计算架构在内存墙和功耗墙的双重限定下，对新兴数据密集型应用的影响变得越来越突出，亟需新的计算架构解决这一问题。

存算一体技术就是从应用需求出发，进行计算和存储的最优化联合设计，减少数据的无效搬移、增加数据的读写带宽、提升计算的能效比，从而突破现有内存墙和功耗墙的限制。

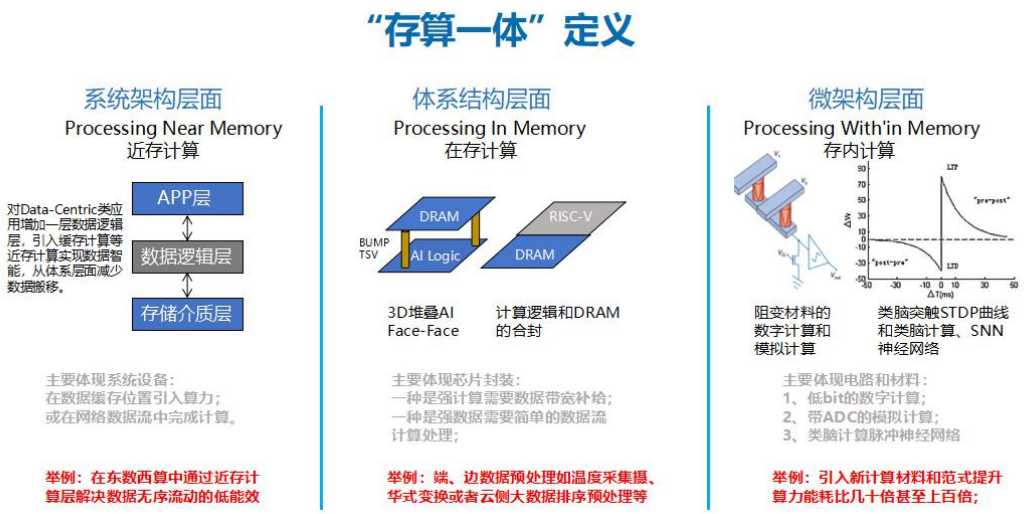

图 4.3 存算一体的三种架构

存算一体包含系统架构、体系结构和微架构多个层面。系统架构层面，在传统计算和存储单元中间增加数据逻辑层，实现近存计算，减少数据中心内、外数据低效率搬移，从系统层面提升计算能效比；体系架构层面，利用 3D 堆叠、异构集成等先进技术，将计算逻辑和存储单元合封，实现在存计算，从而增加数据带宽、优化数据搬移路径、降低系统延时；微架构层面，进行存储和计算的一体化设计，实现存内计算，基于传统存储材料和新型非易失存储材料，在存储功能的电路内同时实现计算功能，取得最佳的能效比。

（一）系统架构层面的近存计算（Processing Near Memory）

近存计算在数据缓存位置引入算力，在本地产生处理结果并直接返回，可以减少数据移动，加快处理速度，并提升安全性。如图 4.3 所示，通过对 Data-Centric 类应用增加一层数据逻辑层，整合原系统架构中的数据逻辑布局功能和应用服务数据智能功能，并引入缓存计算，从而减少数据搬移。在“东数西算”工程中，可以通过设置近存计算层，解决数据无序流动的低能效问题。

（二）体系架构层面的在存计算（Processing In Memory）

在存计算主要在存储器内部集成计算引擎，这个存储器通常是 DRAM。其目标是直接在数据读写的同时完成简单处理，而无需将数据拷贝到处理器中进行计算。例如摄氏和华氏温度的转换。在存计算本质上还是计算、存储分离架构，只是将存储和计算靠近设计，从而减少数据搬移带来的开销。目前主要是存储器厂商在推动其产业化。

（三）微架构层面的存内计算（Processing Within Memory）

存内计算是把计算单位嵌入到存储器中，特别适合执行高度并行的矩阵向量乘积，在机器学习、密码学、微分方程求解等方面有较好的应用前景。

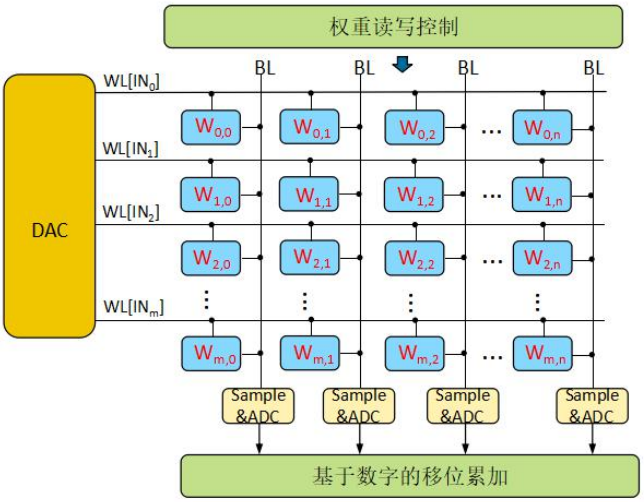

图 4.4 存内计算架构示意

存内计算采用计算、存储统一设计的架构。以深度神经网络的矩阵向量乘加操作为例，一般采用图 4.4 所示架构，由输入端的 DAC、单元阵列、输出端的 ADC 以及其他辅助电路组成。存储单元中存放权重数据，输入经过 DAC 转换后变成对存储数据的读写操作，利用欧姆定律和基尔霍夫定律，不同的存储单元输出电流自动累加后输出到 ADC 单元进行采样，

转换成输出的数字信号，这样就完成了矩阵向量乘加操作。

存内计算按照存储介质不同可以分为易失材料（DRAM/SRAM）和非易失材料(RRAM/PRAM/MRAM/FLASH 等）2 类。在计算电路的具体实现上包括模拟计算、数模混合计算和全数字计算。基于 SRAM 的全数字存内计算芯片，由于工艺成熟，同时能够实现比较高的计算精度（≧8bit），目前已经逐步进入商用阶段。而基于非易失材料的模拟存内计算，还需要解决器件稳定性和功耗问题。

4.4.计算架构：基于对等系统的分布式计算架构

传统的计算系统以 CPU 为中心进行搭建，业务的激增对于系统处理能力要求越来越高，摩尔定律放缓，CPU 的处理能力增长越来越困难，出现了算力墙。通过领域定制（DSA）和异构计算架构可以提升系统的性能，但是改变不了以 CPU 为中心的架构体系，加速器之间的数据交互通常还是需要通过 CPU 来进行中转，CPU 容易成为瓶颈，效率不高。

基于 xPU（以数据为中心的处理单元）为中心的对等系统可以构建一个新型的分布式计算架构。如图 4.5 所示，对等系统由多个结构相似的节点互联而成，每个节点以 xPU 为核心，包含多种异构的算力资源，如 CPU、GPU 及其它算力芯片。xPU 主要功能是完成节点内异构算力的接入、互联以及节点间的互联，xPU 内部的通用处理器核可以对节点内的算力资源进行管理和二级调度。节点内不再以 CPU 为中心，CPU、GPU 及其它算力芯片作为节点内的算力资源处于完全对等的地位，xPU 根据各算力芯片的特点及能力进行任务分配。

对等系统的节点内部和节点之间采用基于内存语义的新型传输协议，即，采用 read/write 等对内存操作的语义，实现对等、无连接、授权空间访问的通信模式，通过多路径传输、选择性重传、集合通信等技术提高通信效率。与 TCP、RoCE 等现有传输协议相比，基于内存语义的传输协议基于低延时、高扩展性的优势。节点内 xPU、CPU、GPU 及其他

算力芯片之间通过基于内存语义的低延时总线直接进行数据交互。节点间通过 xPU 内部的高性能转发面实现基于内存语义的低延时 Fabric，从而构建以节点为单位的分布式算力系统。同时 xPU 内置安全、网络、存储加速模块，降低了算力资源的消耗，提高了节点的性能。

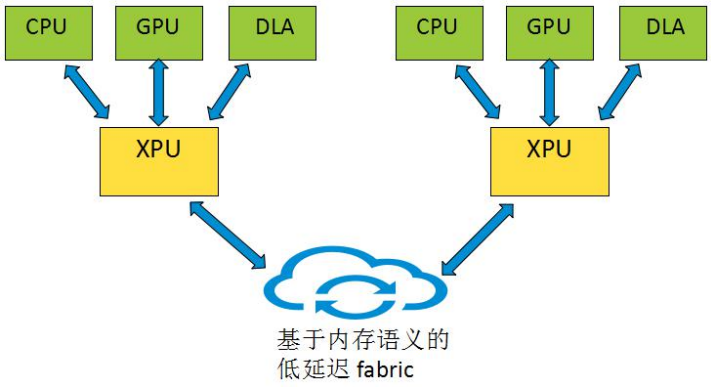

图 4.5 对等系统示意图

基于对等系统架构的服务器可以看成一个“分布式计算系统”，有利于产业链上各节点独立规划开发，发挥各自优势。比如 xPU 卸载 + 库/外 OS 演进 + APP direct 模式解决公共能力（存储、网络），整体性能的提升不再依赖于先进工艺；基于对等内存语义互联实现系统平滑扩展，将庞大分布式算力视为一台单一的“计算机”。

4.5. 网络架构：支撑算网融合的 IP 网络技术实现算力资源高效调度

随着边缘计算的发展、东数西算的推进，算力资源呈现分布式部署的模式。如前所述，网络带宽的提速受到香农定理的限制，算力能力的提升受到摩尔定律的约束，两者又都受到节能降耗的影响，预计网络和算力资源的高效调度、精细化运营成为必须选择的方向。算网络融合的目标借助高速、灵活、智能的网络，将跨地域的算力节点组织起来，协同提供开放的算力服务，并提高算网资源的有效利用率。

算网深度融合有两大驱动力，一是需求侧，实现算力和网络的协同调度，满足业务对算力资源和网络连接的一体化需求。比如，高分辨率的 VR 云游戏，既需要专用图形处理器（GPU）计算资源完成渲染，又需要确定性的网络连接来满足 10 ms 以内的端到端时延要

求。二是供给侧，借助于网络设施天生的无处不在的分布式特点，算网深度融合可以助力算力资源也实现分布化部署，满足各类应用对于时延、能耗、安全的多样化需求。

算网融合给 IP 网络技术提出了挑战。在互联网整个技术架构中，通常来说算对应着上层的应用，网对应着底层的连接，IP 技术作为中间层，起到承上启下的枢纽作用。传统的 IP 网络遵循的端到端和分层解耦的架构设计，使得业务可以脱离网络而独立发展，极大降低了互联网业务的创新门槛，增加了业务部署的便利。但是在这样的设计架构之下，业务和网络处于“去耦合“的状态，最终绝大多数业务只能按照“尽力而为”的模式运行。

如何建立业务和网络之间的桥梁，实现算力资源、网络资源的协同和精细化管理，是未来 IP 网络面临的一大挑战。中兴通讯提出的“服务感知网络（SAN，Service Awareness Network）”是在这个方面的创新尝试[25]。

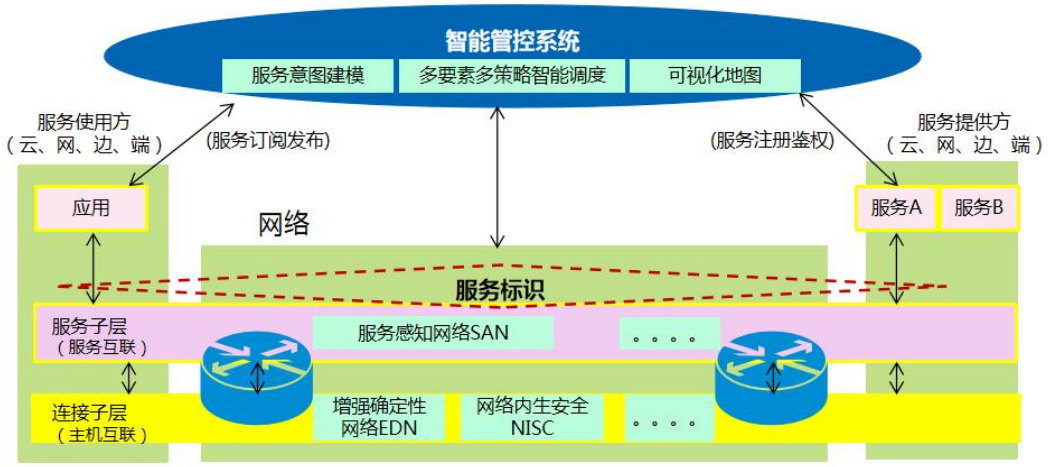

图 4.6 服务感知网络（SAN）架构示意

服务感知网络的架构如图 4.6 所示。其核心思想是把服务提供商对外提供的算力资源和网络资源封装为“服务”，以服务标识来表示；服务按需在网络各处部署；在 IP 网络层引入“服务子层”，实现对服务的感知、路由和调度。因此服务感知网络具有三个核心设计要素：

（一）横向贯穿端网云、纵向打通应用和网络的服务标识

服务标识是端、网、云统一的服务治理对象，包括连接类服务和算力服务。应用层直接

用服务标识发起位置无关传输层连接，无需 DNS 域名解析过程，大大缩短了服务响应时间，并且内含了对移动性的支持。

（二）在 IP 网络层引入服务子层（3.5 层），以网络为中心实现服务互联

在保留传统 IP 主机路由的基础上引入以服务标识为中心的服务子层，使能网络对服务使用方的算力需求的感知，和服务提供方的算力资源状态的感知，从而通过服务路由，实现服务需求到服务资源的高效连接，实现了网络从传统模式下的主机互联到服务互联的演进，

（三）能力增强的连接子层

连接子层对网络的基础能力进行增强，比如确定性服务能力和内生安全能力。服务的网络连接级服务质量需求由连接子层满足。

服务感知网络实现了算力服务和网络服务的一体化供给，实现算网资源的高效调度，既保障了服务质量，又能将节能减排的要求落到实处。

5. 更高的智能

5.1. 综述

智能技术是推动人类进入智能时代的决定性力量。全球发达经济体充分认识到人工智能技术引领新一轮产业变革的重大意义，纷纷推进智能基础设施建设和各领域应用研究。

人工智能技术的基本要素是“三算”：算力、算法和数据（算料）。其中数据跟具体业务领域相关，开放、共享、可流通的数据资源体系建设是《数字中国建设整体布局规划》[03]关注的两大基础之一；而算力和算法是数字基础设施应具备的基本能力。

从 2016 年以来的新一轮人工智能技术在“三算”方面已取得重要突破，正从之前“不能用”到现在“可以用”，但距离“很好用”仍有诸多瓶颈。如强大的智能能力依赖于复杂的算法叠加庞大的算力而获得，成本功耗均很高，对于环境的压力很大；面向特定领域的专用人工智能因建模相对简单、算料标注充沛，已经取得不少优于人类能力的成果，但是通用智能仍处于起步阶段，处于“有智能无智慧、有智商无情商、有专才无通才”状态。正如第二章所述，人类尚未在人工智能的基础理论（认知科学）取得突破，缺乏理论指导。

在 AI 算力需求的增长远大于摩尔定律增长的情况下，如何实现更有效率的 AI 芯片是业界面临的重要课题。本章第 5.2 节论述了 AI 芯片架构的创新方向，以实现更高算力/能耗比。

以 ChatGPT 为代表的生成式 AI（AIGC）的成功，使多模态大模型成为通用 AI 算法最有潜力的拓展方向。本章第 5.3 节论述了大模型技术的发展趋势及其不断拓展的应用领域，进而有可能演进为新的平台层，模型即服务（MaaS）成为可能商业模式，为多种行业提供通用化的 AI 能力。

本章第 5.4 节把“网络智能化”作为智能基础设施的一个应用案例。人工智能如何赋能网络的运营运维，实现网络自身的数字化转型，是电信行业始终关注的话题。在更高效的 AI 算力，以及大模型等新型算法的支撑下，网络智能化有望从目前的 L2~L3 级别向 L4~L5 更高级别迈进。

5.2. 智能芯片：提高算力/能耗比的技术方向

如第二章所述，目前 AI 计算的能耗快速增长，将为环境带来沉重的负担。为了实现绿色可持续发展，必须不断研究更有效率的 AI 芯片。

实现 AI 芯片高 Tops/W （算力/能耗比）的两个可行的方向是空间计算和近似计算。

（一）空间计算

AI 芯片功耗与数据在芯片内搬运的距离正相关。借助创新的芯片架构设计，减少完成每次操作数据在芯片内需要移动的距离，可以大幅降低芯片的能耗。

将一个包含大计算、大存储单元的计算核心拆分为多个包含小计算、小存储单元的计算核心，可以有效降低每次计算数据移动的平均距离，从而降低芯片能耗。这也成为新一代 AI 芯片的设计趋势。然而，这种多核并行计算会引入额外的开销，导致计算效率降低。“空间计算”是通过软硬件架构协同设计，将一个计算任务拆分为多个子任务，然后将子任务指派到不同的计算核心上，并规划任务之间数据传输路径，最优匹配芯片的算力、存储、数据传输带宽、互联拓扑结构，减少数据移动距离，从而实现性能最优、功耗最低。

实现多核空间计算需要软硬件协同设计。在硬件方面，为提升并行计算效率，计算核心可以增加对 AI 并行计算常用通信模式的硬件支持，如 Scatter、Gather、Broadcast 等，对数据包进行封装、压缩等，在核间互联上优化片上网络拓扑结构和动态路由能力；在软件方面，由于空间计算的优化非常复杂，非开发人员所能负担，需要编译器自动实现任务的拆分、指派、路由规划，在运行时需要完成计算过程控制，特别是对空间计算过程中产生的各种异常（如丢包、乱序、拥塞）进行处理。

未来空间计算的一条演进路线是在存计算。在存计算可以把一个大的计算核心拆分为上万个微型计算核心，而不仅仅是上百个小核心。在这种架构下，每个计算数据平均移动距离将进一步降低至微米级，功效比可以超过 10 TOPS/W@INT8。例如 Untether AI 公司的 Boqueria 芯片拥有超过三十万个处理引擎 （Processing Elements），功效比高达 30 TFLOPS/W@FP8[26]。

空间计算技术的另一条演进路线是确定性设计。编译器优化能力对空间计算的性能至关

重要，但只能利用静态信息对计算进行调度。因此，重新设计系统的软件-硬件界面、静态-动态界面，使编译器能够利用更多的静态信息，成为一个新的技术演进方向。例如，Groq 公司的张量流处理器（TSP）采用确定性硬件设计[27]，编译器可以精确地调度每个核上的计算、内存访问和数据传输，避免共享资源的访问冲突。

（二）近似计算

深度学习模型的一个特征是对精度要求不高。计算过程中出现的误差并不会显著影响模型的最终判定结果。近似算法可以减少内存使用和计算复杂度，使计算更加高效。

低精度计算是深度学习近似计算一个重要的技术方向。使用低精度的数据类型，可以有效减少芯片面积和功耗。例如，INT8 的乘法和加法运算所消耗的能量仅为 32 位浮点数（FP32）的 1/30 和 1/15[28]。目前混合精度训练技术可以使用 FP16 位半精度浮点数和 FP32 单精度浮点数配合完成模型训练。

由于推理对精度的要求更低，因此在完成模型训练之后，可以将模型转化为更低精度的数据类型表示，这个技术称之为模型量化。目前，INT8 量化技术已经相当成熟，INT4 量化技术仍然面临一些困难。

近似计算的另一个演进路线是稀疏计算。研究发现，深度学习模型的权重存在一定的稀疏性，即部分权重值为零或者非常接近于零，特别是 Transformer 模型的稀疏度更大。利用模型的稀疏性可以省略不必要的计算，从而提升模型计算的效率。例如，Nvidia A100 GPGPU 中的 4 选 2 稀疏加速可以将芯片等效算力提升一倍[29]，同时功耗保持不变。Tenstorrent Wormhole 芯片更是可以在模型稀疏度 90%的情况下，将芯片等效算力提升 100 倍。未来软硬件协同下稀疏计算仍然会是一个非常有前景的技术方向。

未来 10 年，依靠制程提升能效比的难度越来越大，而空间计算、近似计算在提升芯片能效比方面存在巨大潜力。相对于目前的主流 AI 芯片，未来的芯片效能将有数十倍的提升，是 AI 产业实现双碳目标的有力保障。

5.3. 智能算法：多样化分离小模型向通用大模型演进

AI 算法的本质是提供一种现实世界与数字世界的映射，算法的好坏取决于用数学模型去表征现实问题的准确程度。AI 算法领域从初期的统计机器学习到 CNN、BERT、Transformer，直到最近出现的 GPT 大模型，构建的数字化模型规模越来越大，与现实世界的匹配度也越来越好。尤其是 ChatGPT 和 GPT-4 等 GPT 类大模型的出现，在 AI 领域掀起了一场革命。大模型已经成为人工智能算法的发展方向，拉开了通用人工智能的发展序幕。

（一）AIGC（生成式 AI）背后的基础模型：Transformer

2016 年 Google 发明了一种基于注意力机制的全新架构深度学习模型 Transformer，最初只是用来做机器翻译，但 2017 年 BERT[30]将面向单一任务训练分为两个阶段：任务无关的预训练和任务相关的微调，使 Transformer 成为能够处理多种语言任务的通用模型。同期，OpenAI 同样基于 Transformer 的 GPT 模型使用了和 BERT 不同的预训练思路，即仅使用 Transorformer 的 decoder 部分预训练语言模型，同样证明了其通用性，并且在扩大数据和模型规模后取得了更好的效果。2020 年，业界首个千亿级参数大模型 GPT-3 诞生，引发了一场训练大模型的算力军备竞赛。

（二）多模态大模型：CLIP

Transformer 在语言领域一统天下，成为一种通用的自然语言处理模型，那么，它的通用性是否能够延伸至语言之外的任务？2020 年，ViT[31]证明 Transformer 能够处理图像任务，而且比这个领域传统霸主卷积神经网络（CNN）处理的更好。接下来的进展则越来越快，CLIP 模型证明同一个 Transformer 模型就可以同时处理自然语言和图像两种模态的数据，随后，中国的研究人员也提出了三模态模型。根据文本创作图像这样的应用开始涌现出来，2022 年基于扩散模型的开源 StableDiffusion[32]更是能够生成高分辨率的清晰图像，使 AIGC 的应用场景进一步扩大。

（三）人工反馈的强化学习：ChatGPT

研发出 GPT-3 模型之后，OpenAI 就对其潜力展开深入研究。2021 年 CodeX 使用源代码代替自然语言做为训练语料，使 CodeX 模型（同样基于 GPT-3）具备代码生成能力。2022 年 GPT-3.5 则使用自然语言与源代码混合语料训练，使模型具有思维链能力。InstructGPT[33] 则通过使用人工反馈让模型生成的内容更加符合人类价值观。这一切最终导致了 ChatGPT 的诞生，其增强了对历史对话进行建模的能力，可以有效捕捉用户的意图，完成上下文理解实现连续性对话，能够从海量数据中归纳提炼有用知识，并有逻辑的应用。

（四）大模型激发大量行业应用需求

随着 GPT-4 大模型的发布以及性能飞跃，大模型在各领域有望迎来进一步的落地应用。大模型技术以其真实性、多样性、可控性、组合型的特征，有望帮助企业提高内容生产的效率，以及为其提供更加丰富多元、动态且可交互的内容。

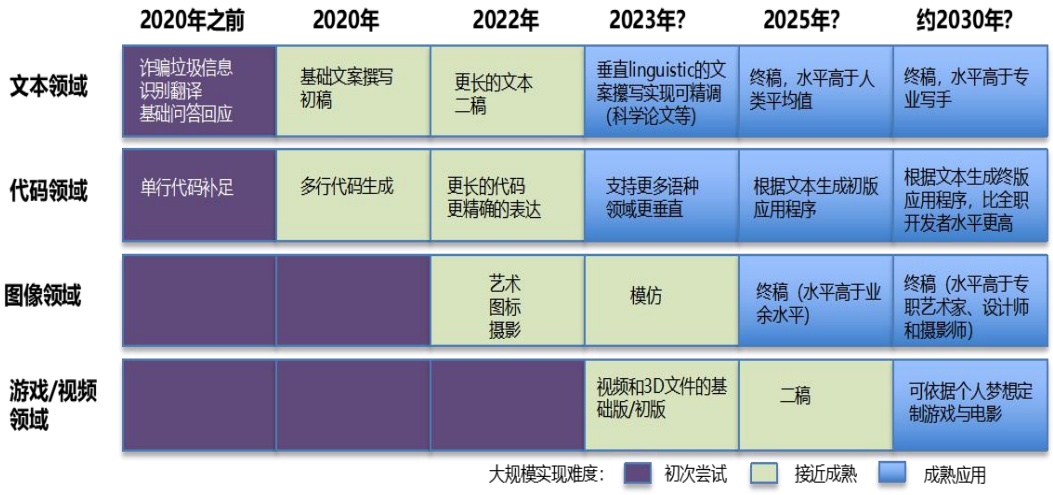

图 5.1 多模态通用大模型的创造能力趋势[34]

大模型在企业数字化的应用场景：

a. 自动对话与写作：自动完成文案写作、客服对话、邮件、会议纪要等。
b. 自动视频生成：通过文字描述，完成动画、视频自动生成。
c. AI作图：输入关键词即可生成图片。
d. 智能辅助编程：可以完成代码生成、补全、代码解释等多种任务，大幅提升软件开发效率。

大模型技术是深度学习跨时代的技术，其与传统深度学习算法最重要的区别是在相当程度上实现了通用性。传统深度学习模型只有处理单一任务的能力，因此过去几年人工智能应

用普遍存在碎片化、跨场景迁移成本高等问题，以至于落地进展较慢。但大模型技术所具备的通用性，使得训练一个模型就可以完成几十种甚至更多的任务，上下文学习能力使模型学习新任务也不需要重新对模型进行训练，这种通用性使得大模型可以成为一个新的平台层，为上层多种应用赋能，为垂直行业客户提供通用化的AI能力。

5.4.智能网络：网络自智向 L4/L5 等级迈进

自智网络包括网络自动化和运维智能化。网络自动化是网络自身实现自动配置、故障自愈、自动优化，具备灵活的业务发放和高可靠性、高性能。运维智能化是在自动化基础上借助 AI 能力实现跨域、跨厂商、跨专业的自动化闭环管理。

结合通信网络要求和业界需求，TM Forum 在 2019 年提出自智网络（Autonomous Networks）概念，并在 2022 年发布《自智网络赋能数字化转型-从战略到实现（Autonomous Networks :Empowering digital transformation – from strategy to implementation）》。TM Forum 提出了三零三自的愿景目标，即在网络运维层面三自（Self-serving，Self-fulfilling，Self-assuring），支撑上层客户三零（Zero Wait，Zero Touch，Zero Trouble），如图 5.2 所示。

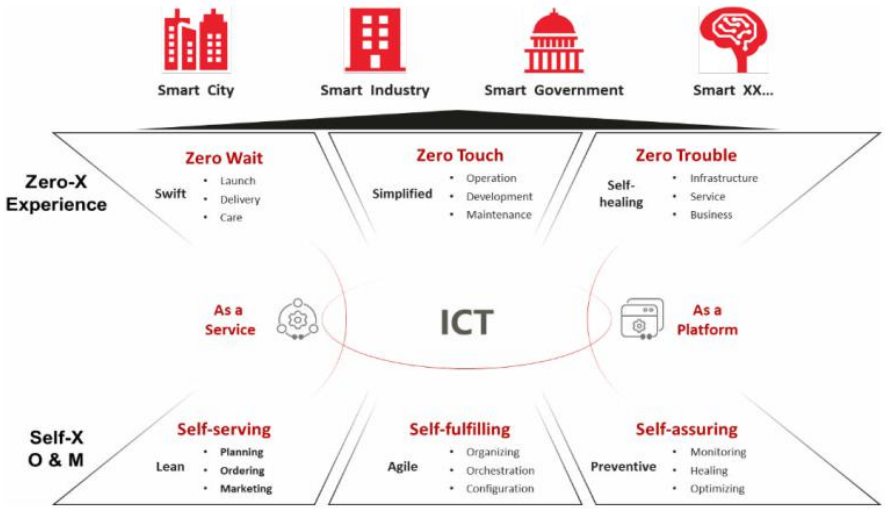

图 5.2 TMF 自智网络“三零三自”愿景目标

从自智网络能力演进升级的角度上，TM Forum 提出了能力分级标准，按程度分为 L0 至 L5 共 6 个级别，和执行、感知、分析、决策、意图/体验和应用六个维度，如图 5.3 所

示：

P：人工 AI：人工智能

等级定义	L0：人工运维	L1：辅助运维	L2：部分自智	L3：基本自智	L4：高度自智	L5：完全自智
执行	P	P+AI	AI	AI	AI	AI
感知	P	P+AI	P+AI	AI	AI	AI
分析	p	p	P+AI	P+AI	AI	AI
决策	P	P	P	P+AI	AI	AI
意图/体验	P	P	P	P	P+AI	AI
应用	N/A	特定场景				全场景

图 5.3 TMF 自智网络分级标准

其中 L0 为最低，全部需要人工操作维护，自动化程度最低；L1 是辅助操作维护，即在执行方面有系统可以实现人工完成的操作记录；L2 是部分自智网络，在执行方面由系统自动完成，而感知方面实现人工和系统的配合操作；L3 是有网络基本自治，在执行和感知方面由系统自动完成，在分析/决策方面由人工和系统配合完成；L4 是高度自智网络，即在意图驱动下，执行、感知和决策由系统自动完成，在体验方面由人工和系统配合完成；最高阶段 L5 完全自智。

国内三大运营商已经达成共识，在 2025 年实现 L4 自智网络目标。目前运营商自智网络大致处于 L2-L3 等级。

技术创新是自智网络演进升级的核心能力，要实现自智网络的 L4 等级目标，关键是要把 AI 算法融入到“网络自配置、故障自修复、质量自优化”等场景中：

（一）基于意图闭环，支撑网络自配置

开通配置等业务在现阶段主要以策略配置辅以算法优化调整参数，实施以人工结合自动化为主。随着意图在自智网络中不断成熟，通过客户意图感知和意图翻译，并基于 AI 算法结合人工反馈实现意图闭环验证，将能够自动配置业务参数和系统参数，实现业务自开通自配置。另外，根据业务使用情况，通过实时的意图洞察和趋势分析，自动实现参数调优以达到更好的意图体验。基于意图闭环的网络自配置，无论在 ToC 场景还是 ToB 场景，都将会有更好的零等待零接触体验。

（二）数据多维度分析，支撑网络故障自恢复

现阶段故障恢复的主要技术，是对告警、性能 KPI、日志和业务指标等多维度数据进行聚合分析，形成事件，完成基于智能事件管理的故障闭环处置。在具体实施中，通过知识图谱结合 AI 算法能够有效提升自智网络故障运维能力，如基于网络熵和图注意力网络的时空聚合完成感知和分析，基于因果推断辅助定位决策，基于工单向量化相似度学习推荐处理措施，以及基于 NLP 支撑智能质检完成故障闭环等。在可预见的未来，借助多模态大模型，如 NLP 大模型、网络大模型和视觉大模型等，结合更多维度的数据，通过端到端训练和知识蒸馏等技术，涌现更多的运维知识、运维能力，大幅提高运维精度和扩展运维场景。

（三）算法模型趋向可解释白盒化，支撑网络质量自优化

算法决策结果具备可解释性是自智网络应用场景的需要。算法模型的可解释性是指人能够理解算法模型在其决策过程中所做出的选择，包括做出决策的原因，方法，以及决策的内容。简单的说，可解释性就是把算法模型从黑盒变成了白盒。电信领域自智网络解决的场景直接关系到用户通信和上网的质量，算法推荐决策的结果一旦出现问题有可能会引起投诉事件，算法模型的白盒化有助于用户放心的将算法推荐结果实施到生产环境，另外运维人员可以通过可解释性理解模型做出的决策，找出偏差出现的原因，从而优化提升模型的性能。

中长期看来，自智网络在未来通信技术、大数据和算力网络的支撑下，通过深度学习、多模态大模型和数字孪生等前沿人工智能技术，有序逐步演进到全栈自动化、智能化的 L5 等级，最终实现自智网络中 Self-X 完全自治，实现零等待、零接触和零故障的 Zero-X 目标。

6. 结语

技术的发展理应为更高质量的经济发展、更美好的社会治理所用。自 18 世纪以来，技术创新就是核心生产要素之一。在 21 世纪已过往的 20 年中，随着更多新兴 ICT 技术，如云计算、大数据、人工智能、移动通信、光通信和基础芯片技术等快速发展，人类利用 ICT 技术从海量数据中挖掘信息、获取知识、产生辅助决策的能力越来越强。数据的核心价值被全世界广泛重视，并在我国被列入到五大生产要素之一。数据要素将与其他要素一起驱动数字经济的高质量增长，赋能数字社会、数字政府等方方面面的建设。

面向 2030，“连接+算力+智能”的数字基础设施是数字化时代的核心基础能力。“更宽的连接”将使能更丰富的新型高带宽应用，如元宇宙、3D 全息通话等，也便于海量数据快捷传递，为人工智能训练提供强劲的“燃料”；“更强的算力”将支撑巨量数据的存储与实时处理，使得人工智能产生类人智慧和更强的决策能力成为可能。“更高的智能”将为数字基础设施注入强大的智能要素，进一步驱动通信网络走向智能网络、数字经济升级为智能经济、数字社会迈向智能社会。

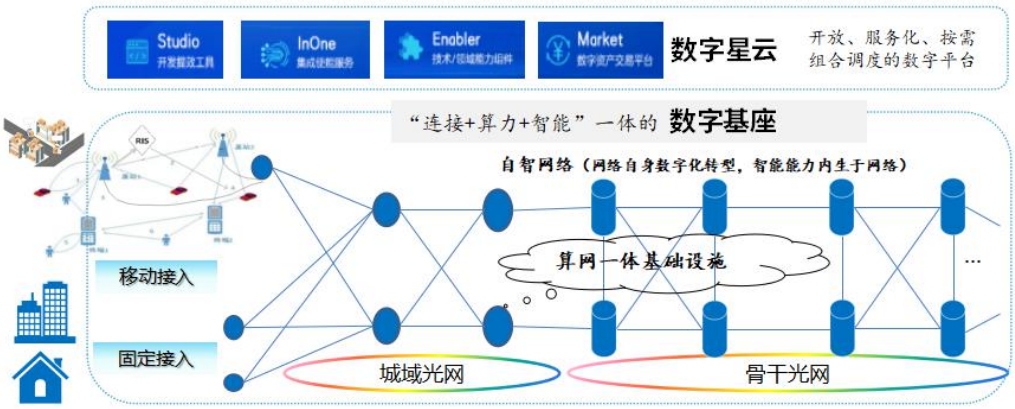

图 6.1 数字星云和数字基座赋能千行百业数字化转型

“更宽的连接”、“更强的算力”和“更高的智能”要更紧密的运作并互为支撑，离不开体系化系统能力的构建。能力构建的关键是形成可组装的复杂软件系统，按需组合、调度并开放各类 ICT 新兴技术能力向外赋能。千里之行、始于足下！从 2022 年起，中兴通讯在业界率先推出数字星云 Digital Nebula（简称 DN），并在 2023 年再次升级为数字星云 2.0[35]，

旨在打造的基于云原生、可服务化、数据驱动的数字化方案和数字化平台，充分组合并发挥“连接+算力+智能”一体的数字基座能力。行业客户可利用“数字星云”进一步构建自己的数字平台，破解多样应用、统一治理和降本增效之间的矛盾，实现业务有韧性、系统可生长、成本能降低。

中兴通讯始终坚持“数字经济筑路者”的定位，坚持“开放共赢”的理念。中兴通讯一方面定位为数字基础设施产品与技术提供商，以用户场景和体验为驱动，提供全球领先的云、网、边、端、软、业产品，并积极开放自身核心原子能力，助力数字运营体及大型企业。另一方面以自身能力带动“隐形冠军”类中小企业，坚持与生态伙伴共生、共赢、共智。

愿此白皮书的发布能带来业界同仁对信息通信领域技术发展的进一步深度交流与诚挚反馈！

7. 参考文献

[1] 中国信息通信研究院：全球数字经济白皮书（2022 年），2022 年 12 月

[2] 中国信息通信研究院：中国数字经济发展研究报告（2023 年），2023 年 4 月

[3] 中国政府网：中共中央 国务院印发《数字中国建设整体布局规划》http://www.gov.cn/xinwen/2023-02/27/content_5743484.htm

[4] 方敏，段向阳，胡留军：6G 技术挑战、创新与展望，中兴通讯技术 2020 年 6 月 第 3 期

[5] IDC&浪潮信息&清华全球产业院：《2021-2022 全球计算力指数评估报告》

[6] 中国信息通信研究院：《中国算力发展指数白皮书(2022 年)》

[7] ITU-T FG-NET2030：Representative Use Cases and Key Network Requirements for Network 2030，2020 年 1 月

[8] ITU-T FG-NET2030：Additional Representative Use Cases and Key Network Requirements for Network 2030，2020 年 6 月

[9] GeSI： SMARTer2030 - ICT solutions for the 21st Century ，2015

[10] Design of Capacity-Approaching Irregular Low-Density Parity-Check Codes， IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 47, NO. 2, FEBRUARY 2001

[11] Amir Gholami 等，https://github.com/amirgholami/ai_and_memory_wall/blob/main/imgs/pdfs/ai_and_compute.pdf

[12] Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient transformers: A survey. ACM Computing Surveys, 55(6), 1-28

[13] STRUBELL E, GANESH A, MCCALLUM A. Energy and policy considerations for deep

learning in NLP [EB/OL]. https://arxiv. org/abs/1906.02243

[14] ISO/IEC 23090-3 Information technology — Coded representation of immersive media —Part 3:Versatile video coding First edition 2021-02

[15] JVET-AB2023 EE1: Summary of Exploration Experiments on Neural Network-based Video Coding

[16] JVET-AB2024 Exploration Experiment on Enhanced Compression beyond VVC capability (EE2)

[17] Alexey Andreyev, Xu Wang, Alex Eckert，“Reinventing Facebook’s data center network”，MARCH 14, 2019

[18] M. LaCroix et al.，“A 116Gb/s DSP-Based Wireline Transceiver in 7nm CMOS Achieving 6pJ/b at 45dB Loss in PAM-4/Duo-PAM-4 and 52dB in PAM-2,” ISSCC, pp.132-133, Feb. 2021.

[19] ODCC-2022-0300A，“112G 线性光互联解决方案白皮书”，P7，2022-09

[20] Rakesh Chopra,"Looking Beyond 400G"P5，TEF2021，January 25，2020

[21] Janet Chen, Meta，Rob Stone, Meta，“Perspective on Linear Drive Pluggable optics”，OIF2023.123.01，

[22] William Dally，"Accelerating Intelligence"，P60,GTC China，December 14, 2020

[23] LightCounting comments on CPO panel discussion at Photonics West，“Our industry is at a crossroads”，February 2023

[24] A. Boroumand, et al.，“Google workloads for consumer devices: Mitigating data movement bottlenecks”， Proc. 23rd Int. Conf. Support Program. Lang. Operating Syst.，2018

[25] 中兴通讯股份有限公司：IP 网络未来演进技术白皮书 2.0，2022 年 8 月

[26] BEACHLER R, SNELGROVE M. Untether ai: boqueria [C]//Proceedings of 2022 IEEE Hot Chips 34 Symposium (HCS). IEEE, 2022: 1-19. DOI:10.1109/HCS55958.2022.9895618

[27] ABTS D, KIM J, KIMMELL G, et al. The Groq Software-defined Scale-out Tensor Streaming Multiprocessor: from chips-to-systems architectural overview [C]//Proceedings of 2022 IEEE Hot Chips 34 Symposium (HCS).IEEE, 2022: 1-69. DOI: 10.1109/HCS55958.2022.9895630

[28] HOROWITZ M. 1.1 Computing ’ s energy problem (and what we can do about it) [C]//Proceedings of 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC). IEEE, 2014: 10-14. DOI:10.1109/ISSCC.2014.6757323

[29] POOL J. Accelerating inference with sparsity using the Nvidia ampere architecture and NVIDIA TENSORRT [EB/OL]. [2022-10-12]. https://developer. nvidia. com/blog/accelerating-inference-with-sparsity-usingampere-and-tensorrt

[30] Lee, J. D. M. C. K., & Toutanova, K. (2018). Pre-training of deep bidirectional transform ers for language understanding. arXiv preprint arXiv:1810.04805.

[31] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale [EB/OL]. [2022-10-12].https://arxiv.org/abs/2010.11929

[32] Borji, A. (2022). Generated faces in the wild: Quantitative comparison of stable diffusio n, midjourney and dall-e 2. arXiv preprint arXiv:2210.00586.

[33] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (202 2). Training language models to follow instructions with human feedback. Advances in

Neural Information Processing Systems, 35, 27730-27744.

[34] 红杉资本：Generative AI: A Creative New World，

https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/

[35] 数字星云 2.0 https://www.zte.com.cn/china/about/news/20230419c9.html