腾讯云 | CAICT 中国信通院

智算赋能算网新应用白皮书

腾讯云计算（北京）有限责任公司
中国信息通信研究院云计算与大数据研究所

智算赋能算网新应用白皮书

腾讯云计算（北京）有限责任公司
中国信息通信研究院云计算与大数据研究所
2023年9月

编委会

编委成员

主编：

张晋、栗蔚

编委（排名不分先后）：

秦若毅、周锐、吴炳文、马飞、苏越、赵伟博、桑柳

参编单位：

腾讯云计算（北京）有限责任公司

中国信息通信研究院云计算与大数据研究所

版权声明

目录 / CONTENTS

前言 / FOREWORD

随着国家“东数西算”工程的启动，算力产业发展进入快车道，推动构建于算力网络之上的算网应用快速发展。伴随大模型训练、全真互联等人工智能浪潮的兴起，将全社会带入智算时代，智算服务成为激发数字经济发展的新动能、新引擎，一方面新场景激发算网新应用诞生，另一方面技术演进促进传统算网应用焕发新活力。

对此，国内外已形成建设智算服务共识，通过政策支撑、资金扶持等方式推动智算服务发展，助力其“内修”——从感知、部署技术到调度技术优化，提升智能算力利用率、生产率，“外治”——推陈出新、由浅入深，扩展算网应用场景支持广度与深度。

在关键技术演进上

本报告系统梳理智算服务关键技术，指出为支撑算网应用建设，当前产业在提升智能算力利用率、生产率上的发展重点与现状：一方面建设灵活感知、融合编排、泛在调度的智算技术矩阵提升智能算力利用率，另一方面打造大规模、高性能智算集群提升智能算力生产率。

在算网新应用发展上

本报告给出算网应用产业发展观察，按照算网应用特性从传统应用与创新应用两方面展开讨论：传统应用依托智算服务实现智能化升级，焕发新活力；创新应用借助智算服务快速发展，强化产业渗透。

在未来发展趋势上

本报告提出三个关键方向：一是在应用发展上，模型即服务具备强大发展潜力，未来将有效助力算力网络发展；二是在服务模式上，类比公有云与私有云，未来将形成通用与专用算网应用协同支撑产业发展的服务模式；三是在发展格局上，跨架构、跨地域提供服务的算网应用将成为全国算网一体化服务的关键支撑。

为进一步梳理智算服务、算网新应用等发展态势，腾讯云计算（北京）有限责任公司与中国信息通信研究院云计算与大数据研究所结合产业发展现状，立足双方智算服务与算力网络研究成果，深度分析产业需求与电信运营商等行业建设算网应用的诉求，输出《智算赋能算网新应用白皮书》。本报告内容仍有诸多不足，恳请各界批评指正。

01

智算服务赋能算网应用创新发展态势

随着新一轮科技革命和产业变革深入推进，以及元宇宙、大模型等新兴应用场景的发展，全球对智能计算的需求激增，智算服务正在成为数字经济发展的新引擎，推动算网应用在产业智慧化的浪潮下展现出全新生命力。算网应用以算力网络为构建基础、以算力任务及相关资源统一编排调度为目标、以算网协同为依托直接服务用户或者相关场景。

在智能计算的持续演进下，算网应用出现两点新变化：一是传统算网应用焕发全新活力，例如功能性能升级、场景支持深化等；二是顺应产业需求衍生创新算网应用，扩展算力网络在产业的支持广度与深度。

1.1 智算成科技发展新驱动，各国抢抓智算服务发展机遇

全球各国布局智算服务，拉开新一轮科技竞赛序幕

伴随智慧出行、智能制造等产业智能化的程度的提升，以及元宇宙、大模型等新兴应用场景的发展，全球对智能算力的需求激增，进入了智算服务的新一轮增长期。**政策上**，美国白宫科技政策办公室发布《国家人工智能战略研发计划》，此政策对AI研发关键领域、投资重点领域等内容进行规范，以确保美国在AI领域的领先地位；2023年，欧盟议会成员就《人工智能法》达成政治协议，该法案将管辖所有人工智能产品或服务的提供方，涵盖可以生成内容、预测、建议或影响环境的决策的系统。**算力规模上**，根据中国信息通信研究院《中国算力发展指数白皮书（2022年）》统计，2021年全球智能算力规模达232EFLOPS，2030年预计达到52.5ZFLOPS，平均年增速超过80%，占全球算力总规模的93%以上，智算算力将成为全球算力规模增长的主要驱动力。**研发投入上**，2020年美国《无尽前沿法案》中提出拟在未来5年投入1000亿美元研发包括芯片、人工智能在内的10大关键技术；2021年4月，欧盟以条例的形式通过“数字欧洲计划”，对包括人工智能在内的项目进行投资，总额达75.9亿欧元。

我国大力发展智算服务，产业布局提速

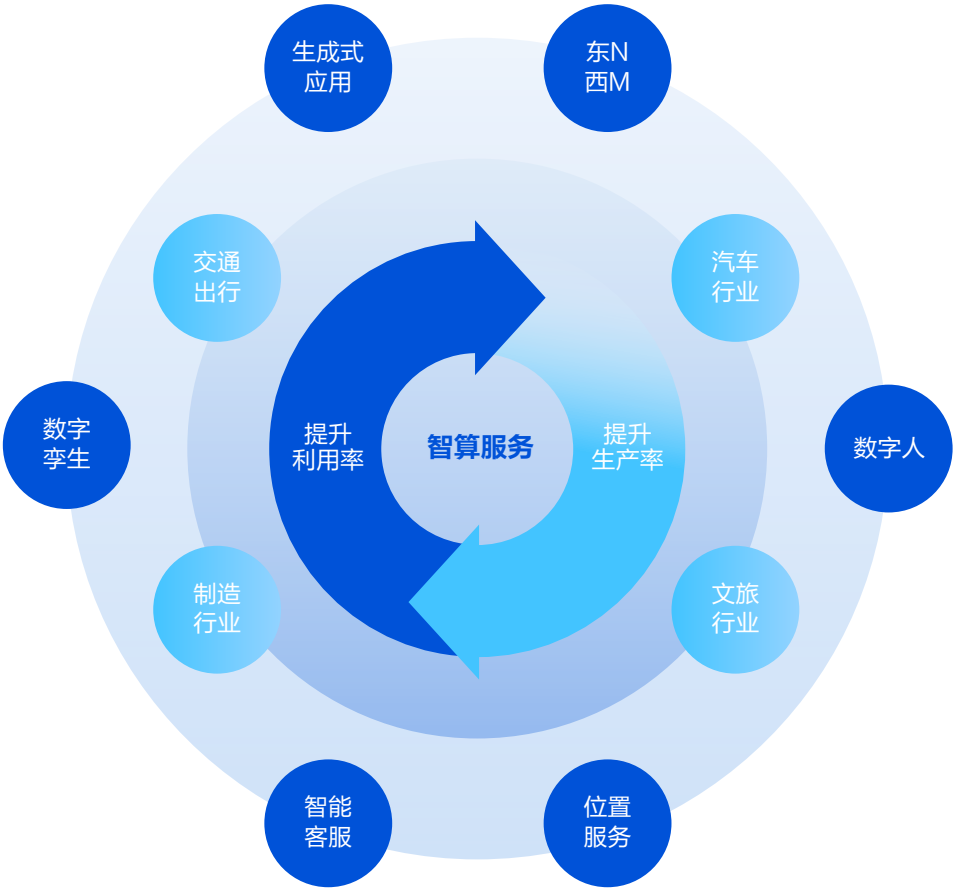
政策上，《新型数据中心发展三年行动计划（2021-2023年）》指出，引导新型数据中心智能化建设，加快高性能智能计算中心部署，支撑各类智能应用。《“十四五”数字经济发展规划》指出要推动智能计算中心有序发展，打造智能算力、通用算法和开发平台一体化的新型智能基础设施，提供体系化的人工智能服务。**算力规模上**，2021年我国智能算力规模达到104EFLOPS，在我国算力总规模中占比超过50%，增速为85%，成为算力规模增长的主要驱动。2022年中国人工智能核心产业规模已达5080亿元人民币。**研发投入上**，北京、上海、广东、山东等地设立专项基金用于人工智能相关技术、标准的研发和应用，打造泛在、标准的智算服务。

1.2 算网应用连接技术与用户，多样产业角色入局共建

算网应用构建于算力网络之上，以服务形式将算力网络技术能力统一输出给用户及应用场景。运营商、云服务商等不同产业角色均投入到算网应用的建设中来，运营商依托其强大的网络能力，打造连接云、边、端资源、服务一体化的算力网络，如中国移动《算力网络白皮书》中提出建设“网络无所不达、算力无所不在、智能无所不及”的算力网络；中国电信规划“核心+省+边缘+端”四级架构AI算力网络，提供算网数智等多要素融合的AI算力服务；中国联通将打造基于算网融合设计的服务型算力网络，构建云网边一体化智能调度和能力开放体系。云服务商依托其成熟的虚拟化技术与算力编排调度技术，建设统一资源管理平台，如“星辰算力调度平台”可实现异构算力资源灵活调度、弹性伸缩。

1.3 智算服务“内修外治”，助力算网应用赋能千行百业

智算服务向内聚焦智能算力利用率、生产率的提升，向外打造智能算力一体化供给服务，支持多样算网应用蓬勃发展。



▲ 智算服务框架

1.3.1 智算服务牵引智能算力利用率、生产率双提升

技术推陈出新，提升智能算力利用率

智算服务通过不断提升网络传输速度、优化算力调度技术等方式实现智能算力利用率的提升。**网络传输方面**，路由协议与芯片间高速互联技术高速发展。网络云化过程发展出了以IPv6+、SD-WAN（Software Defined Wide Area Network）、SRv6（Segment Routing over IPv6）、确定性网络为代表的路由技术，支持将业务需求与算力信息随数据包进入网络，打破网络与算力应用的边界，支撑算力服务下算与网的深度融合，打造坚实算力网络。NVIDIA推出NVLink技术，支持GPU之间业务数据高速互通，良好支撑大模型训练场景。**融合调度方面**，确定性提供高质量调度保障。算力调度领域发展出了同时考虑算力节点与网络传输性能的算网融合技术，提供兼具低时延与高可靠特性的算力服务。例如在智能制造场景下，由于工业制造环境复杂、协议多样，所以需要对算力、网络等支撑资源进行集中化的统一调度和编排。麦肯锡公司发布的《2021年离散制造业上云调查》报告显示：云的IT价值在敏捷性、弹性和经济性几个方面的充分呈现加上同5G技术和应用的结合，在制造、供应链和采购等价值链关键环节赋能作用明显，也催生出如车联网/车路协同、超高清视频流媒体、远程医疗等多行业应用场景。

资源化零为整，提升智能算力生产率

中国信息通信研究院《中国算力服务研究报告（2023年）》指出，算力应用依托有效算力进行计算并输出结果从而实现应用价值，有效算力则是真正完成计算任务的计算能力。提升算力生产率，是提升有效算力规模的关键手段之一。智算服务将社会闲散智能算力整合起来，通过服务化的方式完成智能算力交付，具体体现在以下两个方面：**一是平台化建设，实现资源集约与统一供给**。智算平台围绕人工智能及其衍生技术建设，向下深度适配CPU、GPU、FPGA、MLU、NPU、TPU等算力资源，屏蔽异构算力软硬件差异，构建无需用户理解、感知的资源池；向上提供标准化编程范式及智能计算工具链，提供诸如模型训练、推理、验证等能力，提供自然语言处理、语音处理、图像视频处理等应用，助力产业生态融通。**二是云边端协同，将资源供范围扩展到边、端零散算力**。智算服务结合计算任务特征判断所需计算设备规格及位置，使得边缘、终端智算资源也可运行较小规模、时延不敏感的计算任务，进一步扩大智算资源供给范围，打造泛在化智算服务。以移动云“中训边推”场景为例，人工智能计算任务通过中心云进行大规模模型训练，通过边缘云完成就近推理。该技术实现思路支持资源秒级自动优化、天然跨域容灾，可有效应对计算需求突增的场景。

1.3.2 智算服务助力算网应用推陈出新、由浅入深

推陈出新，智算服务驱动创新算网应用新发展

如元宇宙，大模型等应用场景，通常具有发展年限较短、智能算力规模需求大、性能要求高的特性。如GPT-3.5在微软Azure AI超算基础设施（由V100GPU组成的高带宽集群）上进行训练，总算力消耗约3640PF-days（即每秒一千万亿次计算，运行3640天）。智算服务提供大规模、高性能智算集群，支撑创新算网应用快速落地。

由浅入深，智算服务助力传统算网应用释放新价值

如出行行业应用、制造行业应用、交通行业应用，在大规模算力需求与一体化数据流动需求的同时，通常还具备较长的发展年限、复杂的使用场景、专业的应用软件以及超大规模用户群体等特点，智能化升级风险高、难度大。智算服务通过数据入云、专用服务统一输出等方式支撑传统算网应用平滑、稳定地升级。以智慧交通为例，2020年，高通推出可扩展的自动驾驶平台 Snapdragon Ride，包括安全系统级芯片、安全加速器和自动驾驶软件栈，满足从L1-L3级驾驶辅助以及L4-L5级自动驾驶运算需求；华为MDC智能驾驶计算平台集成智能驾驶操作系统、配套工具链及车路云协同系统，可支持400TOPS算力及200ms时延，助力传统算网应用场景焕发新活力。

02

智算服务关键技术

智算服务基于云计算、大数据、人工智能等技术，提供计算资源、算法模型以及一系列场景应用。通过智算服务，用户可以快速获取高性能计算资源、优化算法模型、提高计算效率，满足不同场景下的算力应用需求。智算服务通常包括算力资源、算法模型、算力应用等多种服务，用户可以根据需要灵活选择和组合使用。智算服务广泛应用于机器学习、大数据分析、科学计算、图像处理等领域，在朝着绿色、多模态、泛在演进。

2.1 智算服务发展聚焦绿色、多模态与泛在

随着技术的不断进步和应用场景的不断扩展，算力及算力应用的发展趋势，逐渐向着更加绿色、智能、泛在的方向发展。

2.1.1 绿色: 用“连接”引领低碳生活，助力产业低碳转型

在数字经济时代，计算力即生产力。但随着算力的增长，数据中心的能耗也在增加。在碳达峰和碳中和的背景下，提高效率、降低能耗是未来产业发展的一个重要课题。加快实现自身运营的碳中和，是企业碳中和行动的首要目标。

大型数据中心、边缘数据中心和5G基站等更多信息化基础设施建设进一步使得电信运营商的能耗成本支出和碳排放快速增长，给电信运营商的可持续发展带来巨大挑战。在“双碳”背景下，如何实现自身降碳目标，并保持可持续增长成为电信运营商在数字经济时代亟待解决的问题。三大电信运营商均已发布碳达峰碳中和绿色行动计划，全面启动双碳绿色行动，用创新催生“新绿色”。中国移动发布的《碳达峰碳中和白皮书》明确了“十四五”节能降碳工作目标：到“十四五”期末，公司单位电信业务总量综合能耗、单位电信业务总量碳排放下降率均不低于20%，企业自身节电量较“十三五”翻两番、超过400亿度，企业2025年自身碳排放控制在5600万吨以内，助力经济社会减排量较“十三五”翻一番、超过16亿吨。中国联通发布《“碳达峰、碳中和”十四五行动规划》，聚焦5大绿色发展方向。一是推动移动基站低碳运营，推广极简建站、潮汐节能等技术，有序提高清洁能源占比；二是建设绿色低碳数据中心，通过供电降损简配、空调利用自然冷源等，提高系统能效；三是深入推进各类通信机房绿色低碳化重构；四是加快推进网络精简优化，老旧设备退网；五是提高智慧能源管理水平。中国电信发布的碳达峰、碳中和行动计划是在“十四五”期末，实现单位电信业务总量综合能耗和单位电信业务总量碳排放下降23%。在“十四五”期间，实现4/5G网络共建共享节电量超过450亿度，新建5G基站节电比例不低于20%；大型、超大型数据中心占比超过80%，新建数据中心PUE低于1.3。另外，下一步中国电信将重点从三个方面推进“双碳”工作：一是建设绿色新云网，打造绿色新运营;二是构建绿色新生态，赋能绿色新发展;三是催生绿色新科技，筑牢绿色新支撑。腾讯于2022年2月发布《腾讯碳中和目标及行动路线报告》，以“减排和绿色电力优先、抵消为辅”的原则，包括节能提效、可再生能源替代、碳抵消等，提出“不晚于2030年,实现自身运营及供应链的全面碳中和。”从节能提效、可再生能源、碳抵消等三个方面开展重点行动，用科技助力实现零碳排放。通过引领绿色低碳生活、助力产业低碳转型，推动社会经济可持续发展。

2.1.2 多模态：AIGC技术大爆发，成为数智发展新引擎

AIGC作为新的生产力引擎，让我们从过去的PGC、UGC，已经不可避免地进入AIGC时代。AIGC代表着AI技术从感知、理解世界到生成、创造世界的跃迁，正推动人工智能迎来下一个时代。

01 基础的生成算法模型不断突破创新。

比如为人熟知的GAN、Transformer、扩散模型等，这些模型的性能、稳定性、生成内容质量等不断提升。得益于生成算法的进步，AIGC现在已经能够生成文字、代码、图像、语音、视频、3D物体等各种类型的内容和数据。

02 预训练模型，也即基础模型、大模型，引发了AIGC技术能力的质变。

虽然过去各类生成模型层出不穷，但是使用门槛高、训练成本高、内容生成简单和质量偏低，远远不能满足真实内容消费场景中的灵活多变、高精度、高质量等需求。而预训练模型能够适用于多任务、多场景、多功能需求，能够解决以上诸多痛点。预训练模型技术也显著提升了AIGC模型的通用化能力和工业化水平，同一个AIGC模型可以高质量地完成多种多样的内容输出任务，让AIGC模型成为自动化内容生产的“工厂”和“流水线”。

03 多模态技术推动了AIGC的内容多样性，进一步增强了AIGC模型的通用化能力。

多模态技术使得语言文字、图像、音视频等多种类型数据可以互相转化和生成。比如CLIP模型，它能够将文字和图像进行关联，并且关联的特征非常丰富。这为后续文生图、文生视频类的AIGC应用的爆发奠定了基础。

随着算力服务的智能化，算法的进步将带来更多激动人心的应用，语言模型会得到进一步发展，可以自我持续学习的多模态AI将日益成为主流，这些因素会进一步推动AIGC领域的蓬勃发展。

传统方式

- 反复查阅文献，人工临摹上色
- 部分破损严重且参考文献不足难以修复
- 耗时长（2-3月/幅）

AI辅助修复

- 基于GAN的多尺度生成模型和损失函数自动修复、上色
- 数据增强技术产生更多训练数据替代、补充文献
- 高效（<1秒/幅）

▲ 利用AIGC技术手段修复古壁画

2.1.3 泛在：让智能算力像水一样流动，随时随地按需取用

智算服务的泛在化意味着更多的人可以获得高效的计算资源和先进的机器学习算法，不再需要拥有昂贵的硬件设备或深厚的技术背景。更多的企业和个人可以基于业务，利用先进的算力和模型，开发出更加创新和高效的算力应用和服务，从而推动整个产业的快速发展。

01 随时，在各种算网应用中，都离不开连接的能力。

从文件传输、即时通讯、视频会议等，不同的应用场景对连接能力有不同的需求。视频会议场景中，在偶发的高丢包高抖动的环境下，如何保持通信的流畅度？智能驾驶场景中，如何将音视频的传输时延从300ms优化到100ms以内？不同场景下，应用的连接需求会给服务体验带来不确定性的挑战，安全广泛的通讯网络连接助推更广泛的生态互通。

02 随地，算网应用中突破了空间的限制，实现了人、物、机器、空间环境之间的连接与交互模式的重塑。

在音视频、数字孪生、3D引擎及空间计算的细节追踪支持下，能够让数字世界全细节化还原或超写实呈现。可为人、物、环境创建1:1还原的全面信息孪生体，让数字世界和真实世界相互连接、映射与耦合，实现数实世界之间的实时同步，是全真互联实现数实融合的呈现形态。在油气、电力、航空、轨道交通等行业中，一线员工经常会面对高风险的作业环境，同时作业设备设施比较复杂，一线员工在遇到自己无法准确判断的设备问题时，很难与他人清晰地交流和协作，而现在一线员工在碰到复杂问题时可以通过轻便的AR眼镜连线专家，专家随即可通过第一视角做出判断，并通过远程实时的AR标注、屏幕共享等功能指导现场员工进行作业，大幅提升现场维修维护时效。

03 按需，按需取用必然要求算力服务的丰富度和智能化水平。

以即时全息通信为例，必然离不开实时音视频、人机交互、XR、数字人、多模态感知、数字孪生等多种智算服务的场景化能力。当原子算力服务越丰富，适配场景的能力越强，其实际满足“按需取用”的契合度就会越高。

以上分析可以得出，实现绿色、多模态、泛在的智算服务，必然还要求算力能高效的流动起来，实现异构算力的一体化调度和高效流通；通过算力虚拟化、算力的分级SLA、混部等关键技术来充分提升算力的利用率；依靠高性能算力集群、框架层、模型层等优化进一步提升智能算力在模型训练和推理的生产效率。

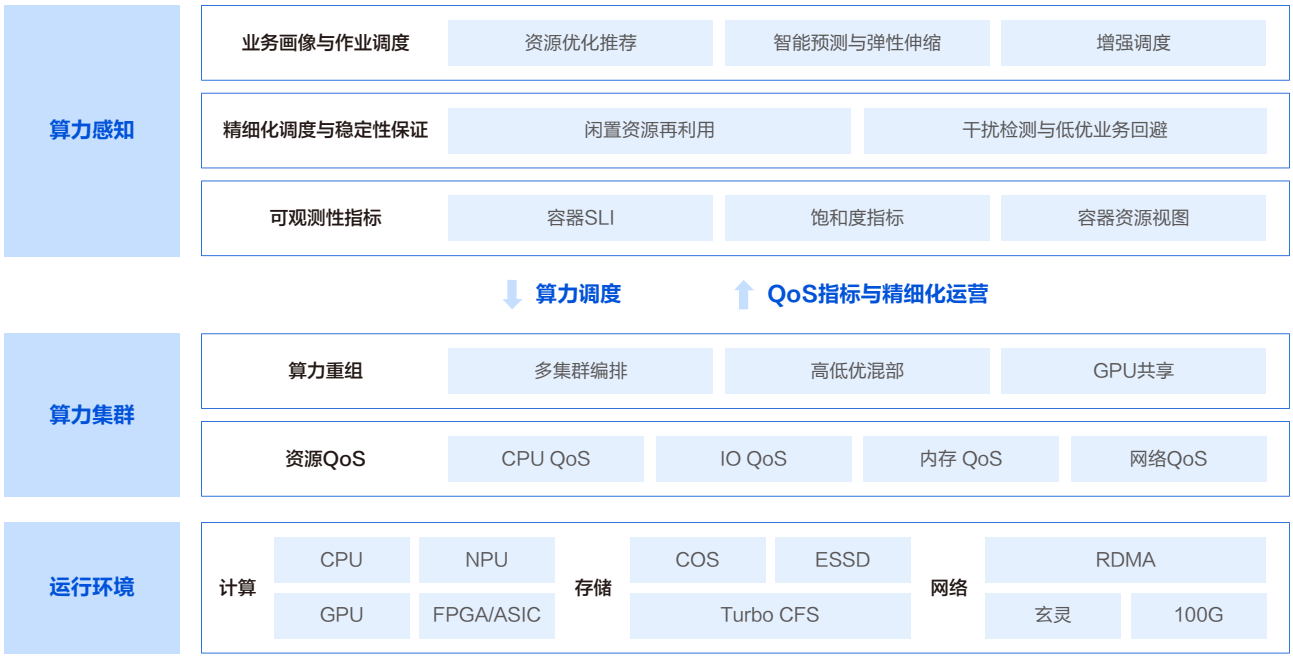
2.2 资源全面感知、精准调度，提升智能算力利用率

当前算力资源的使用还处于粗放式的发展，从目前的统计数据来看，算力的使用率低于30%，造成了大量的计算资源和能源成本的浪费。

算力利用率体现了计算机系统中正在使用的计算能力与该系统的总计算能力之间的比例。算力利用率越高，系统或网络的计算能力就越充分地被利用。

深化全栈技术体系应用，构建一体化智能算力调度服务，可有效提升智能算力的利用率，大幅提高生产效率，通过更少的硬件设施和能源成本，实现更高的经济效益，是实现算力绿色高效、可持续发展的有效手段。

大量算力应用场景对算力资源的某些方面的可用性存在特殊要求，不同在线或离线业务对算力服务的质量要求千差万别。从供给侧来看，传统无差别算力服务提供模式无法为差异化应用需求提供个性化的可靠保障。按同样的要求进行规划保障，容易造成算力资源大量浪费或无法满足业务需求的两种极端情况。因此在提升算力利用率的同时，需要保障算力服务的可用性。通过算力虚拟化、算力隔离、算力感知、混合部署和调度等技术，来实现不同SLA要求的智算服务的可靠性保障。



▲ 算力感知技术框架

2.2.1 智能算力感知：构建智算感知能力体系，为资源细粒度优化提供依据

算力感知是算力调度的基础，通过建立算力感知的技术指标体系，一方面可定义业务应用的算力参数需求，如计算性能、网络时延等，另一方面定义算力运行的可观测性指标，包括全维度硬件QoS指标，如CPU、IO、内存、网络等。

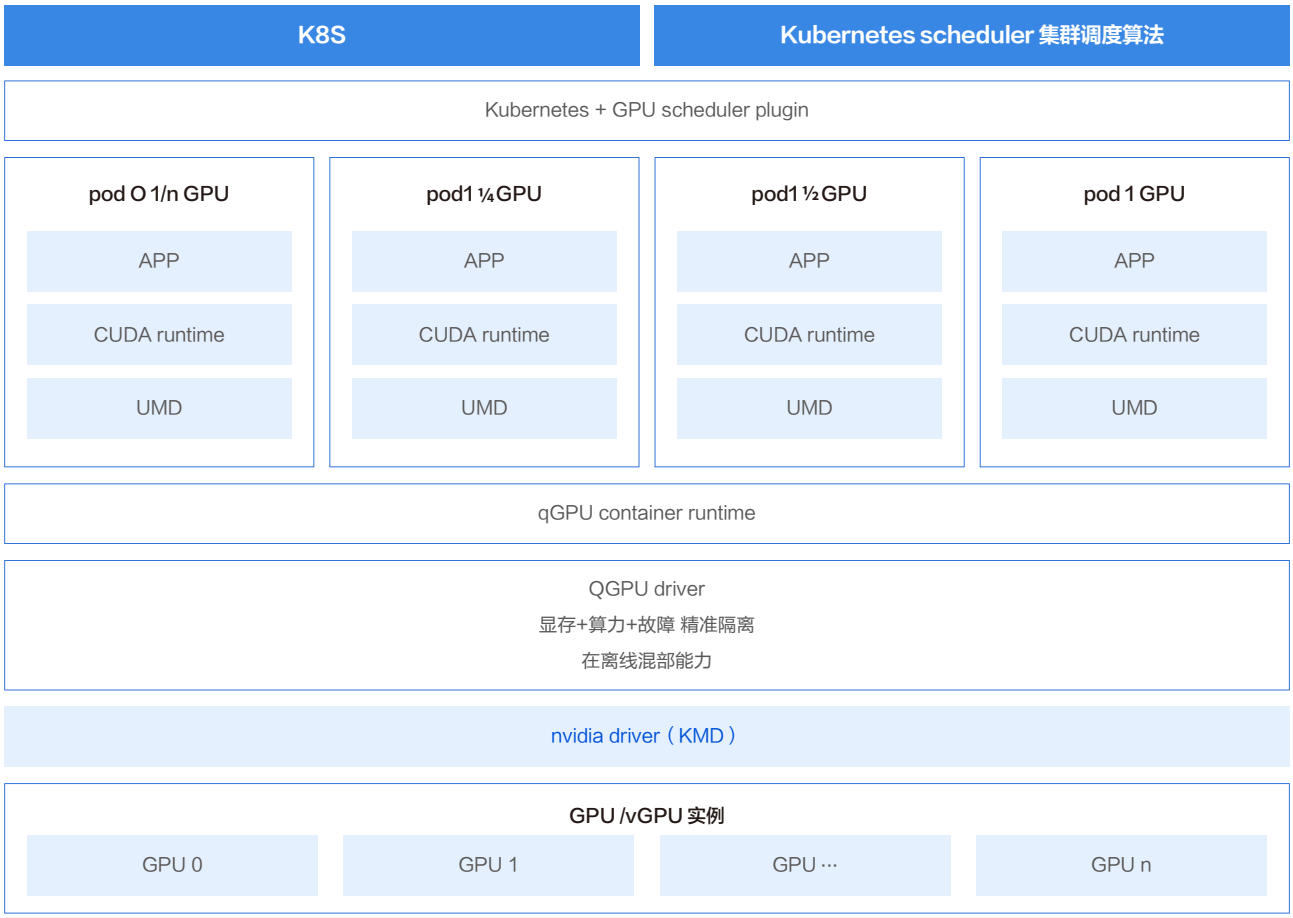
在算力服务实际的运行过程中，实时感知算力集群内各个算力资源的实际负载情况，根据算力资源需求重调度到更合适的算力节点，保障算力服务在各个算力节点的合理负载。并对其进行可视化展示和分析，了解资源的实际利用率以及周期性规律，在此基础上，针对业务应用的算力需求进行细粒度的资源优化。

通过算力应用的运行情况，可对算力应用进行画像，感知业务实际的资源用量，为业务智能推荐资源需求，智能预测峰值算力资源，做到按需弹性扩缩容，随取随用。

2.2.2 智能算力共享：精准隔离，有效提升智算应用部署密度

随着机器学习的不断发展，GPU的性能越来越强，提供并行算力已非常普遍。在实际的使用过程中，通常将完整的GPU卡分配给一个容器，对于模型开发和模型推理等场景资源浪费严重。因此通过GPU共享技术，可有效的提升算力应用的部署密度，提升GPU的利用率。

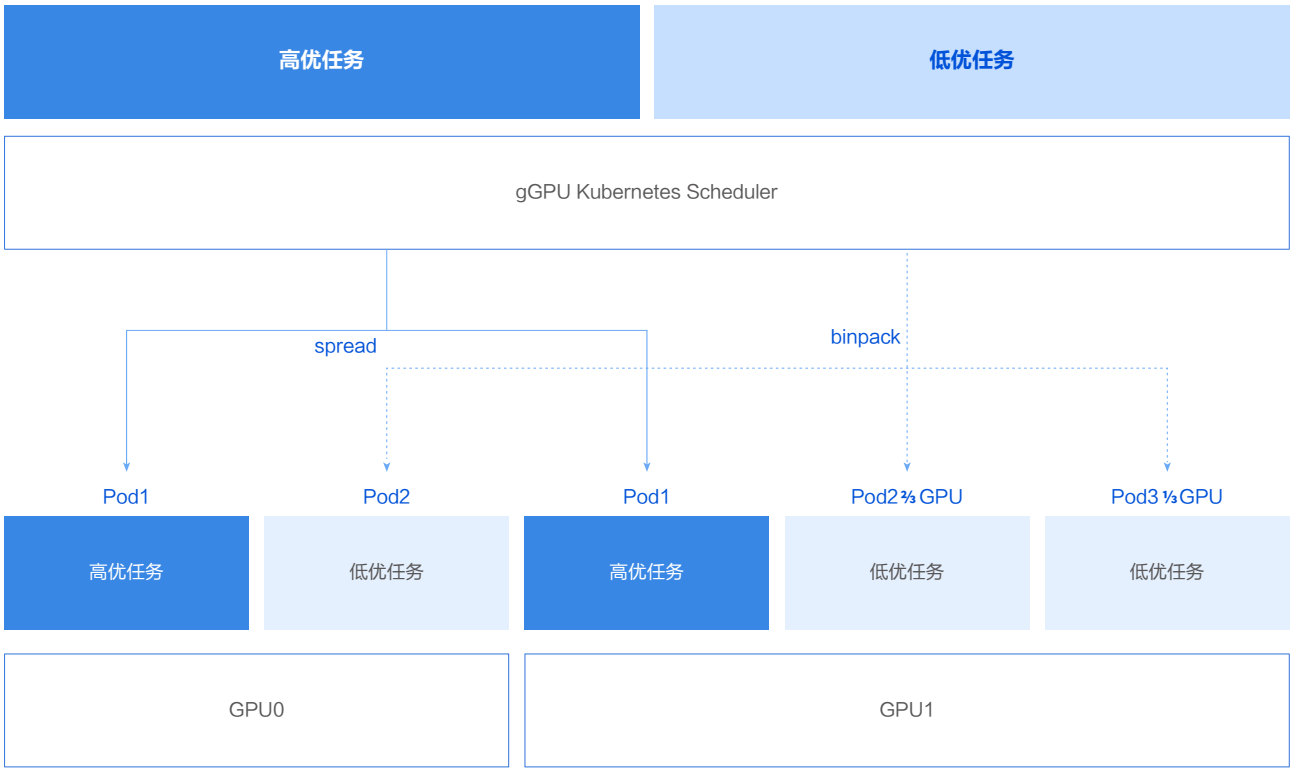
GPU共享需要解决容器间算力和显存精细隔离的问题，支持算力和显存的灵活配置，从而在精细切分GPU资源，最大程度保证业务稳定的前提下，大幅提升GPU利用率，以达到节约GPU资源成本的目的。同时需具备良好的兼容性和云原生的支持，实现业务无感接入。



▲ QoS GPU容器共享技术框架

2.2.3 混合部署：智算应用分级 QoS，削峰填谷，充分利用空闲算力

提高算力集群资源利用率，可对不同优先级的业务应用进行混合部署，通过不同的组合方式，如错峰业务组合、计算型和内存型任务的组合等，运行更多的算力任务。混合部署对算力隔离和精细调度要求高，只有对计算和显存提供强有力的 QoS 保障和完全的隔离能力，才能使得算力共享带来的利用率提升的同时，满足不同算力服务的可用性要求。



▲ QoS GPU容器混部技术框架

2.2.4 智能算力调度：一体化精准调度，最大化算力价值

通过算力感知，可分析算力的整体效率，提供可靠、便利、智能的算力调度优化技术方案，以满足算力应用的分级QoS和SLA要求，实现算力的调度优化。

算力调度的优化包括节点亲和性调度、基于负载的动态调度、基于SLA保障的重调度等。利用亲和性调度找到适配业务任务的资源，通过动态调度实现资源的总体负载优化，同时通过重调度保障业务的可用性。

基于动态调度策略，可解决资源碎片的问题，提高装箱率回收业务波谷时的冗余，通过算力应用弹性和混合部署，做到按需使用。对于固定资源池，对负载峰值在不同时段的在线应用、离线应用进行混部，做到分时复用，实现资源的池化、共享以及隔离。

▲ 一体化算力调度架构

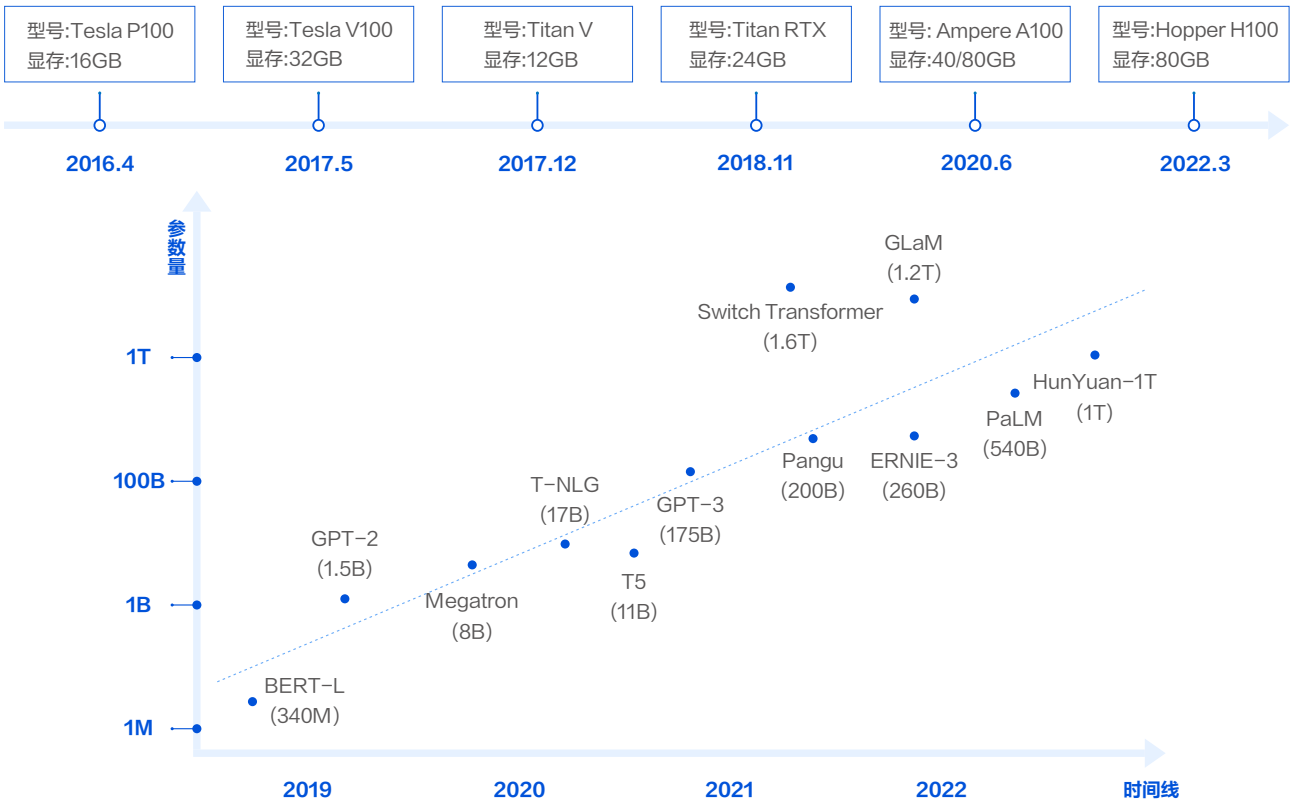
2.3 提升智算生产率，推动算力泛在化发展

随着信息技术的飞速发展，人工智能、大数据分析、自动化等智能化技术正在快速渗透到各行各业。智算生产率提升指的是通过这些技术手段实现生产效率和质量的提升，降低生产成本，提高竞争力。智算生产率提升还有助于节约资源，实现经济可持续发展。因此，智算生产力提升已经成为算力经济低碳化、普惠化发展的重要手段，是推动社会进步和经济发展的重要动力之一。

智算生产率体现了智能计算技术在生产过程中所创造的产出价值与所投入资源之比，反映了智能计算技术在生产中创造价值的效率和贡献。

以大模型生产为例，近几年NLP预训练模型规模的发展，模型已经从亿级发展到了万亿级参数规模。2018年BERT模型最大参数量为3.4亿，2019年GPT-2为十亿级参数的模型。2020年发布的百亿级规模有T5和T-NLG，以及千亿参数规模的GPT-3。在2021年末，Google发布了Switch Transformer，首次将模型规模提升至万亿。

然而硬件发展的速度难以满足Transformer模型规模发展的需求。近四年中，模型参数量增长了十万倍，随着模型训练的要求越来越高，动辄需要数千卡的资源投入，需要更多的算力和时间，这导致了更高的资源成本。



▲ 大模型训练参数增长趋势

因此，提高智能算力的生产率，能有效的减少算力投入的门槛，缩短生产时间，是算力普惠的关键路径。智算生产率提升的关键一方面在于构建高性能智算集群平台能力，包括提供高性能的计算、网络、存储能力，另一方面在于提供智算的加速框架层优化及模型优化等。高性能智算集群是基于高性能计算和人工智能技术的计算机集群，旨在提供更高的计算性能和更快的人工智能应用速度。高性能智算集群可用于执行大规模、高计算密度的人工智能任务，如NLP预训练模型，它可以提供更快的计算速度和更高的精度，以便处理大型数据集和复杂的计算任务。通过对算力、网络架构和存储性能进行协同优化，能够为大模型训练提供高性能、高带宽、低延迟的智算能力支撑。高性能智算集群具备计算性能强、通信能力强、存储读取快等特点。

▲ 高性能智算集群生产效率优化

2.3.1 高性能计算：提升单节点计算能力，并向分布式、混合并行模式演进

大模型进入万亿参数时代，训练数据量和模型参数量发生了两个关键层次的变化，一是随着数据量的扩大，从单卡训练转变为分布式训练，二是数据并行训练升级到多维混合并行训练。

在数据并行方案中，数据集被切分成后分配给各卡并行处理。每张卡上运行完整的模型，保证了各卡之间模型的一致性。在模型参数特别大的情况下，如千亿级别，单卡已无法容纳完整模型，因此除数据并行外，需要同时采用模型并行的方案，实行多维的混合并行训练。

因此在构建高性能算力集群，需要对处理器、网络架构和存储性能进行全面优化，一方面优化单计算节点运行时的 I/O、CPU预处理、CPU/GPU数据通信、GPU计算等方面的性能开销，另一方面需要解决大模型场景下多节点协作的性能损耗问题，为大模型训练提供高性能、高带宽、低延迟的高性能计算支撑。

2.3.2 高性能网络：建设高性能通信网络，有效提升智能算力集群性能

当模型达到一定规模时，需要实现分布式的多维混合并行训练，计算节点间存在海量的数据交互需求。随着集群规模扩大，通信性能会直接影响训练效率，通过高性能网络架构保障算力性能的线性增长是有效发挥算力集群性能的关键因素。如“东数西算”宁夏枢纽搭建的智算无损网络，实现单GPU服务器之间800G的大带宽；“星脉”网络搭载了3.2T的超高通信带宽，在同样的GPU卡上星脉网络相较前一代网络，将集群整体算力提升20%。高性能通信网络使得超大算力集群能保持优秀的通信开销比和吞吐性能，并支持单集群高达十万卡级别的组网规模，满足更大规模的大模型训练及推理。

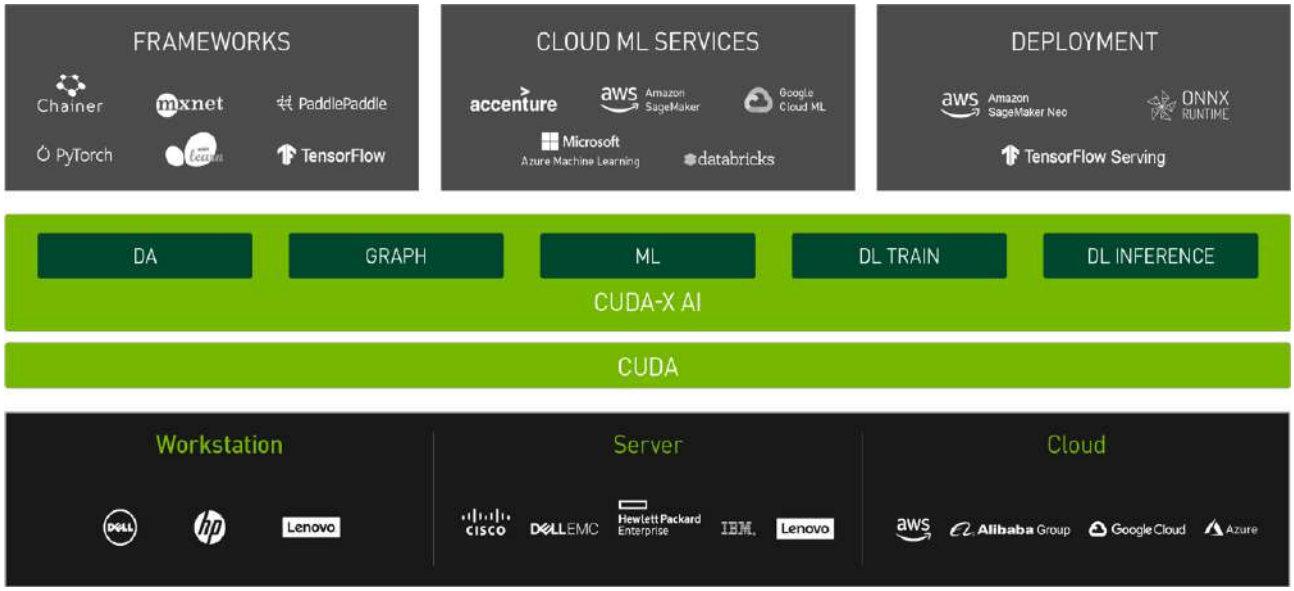
另外面对定制设计的高性能组网架构，开源的集合通信库（比如NCCL）并不能将网络的通信性能发挥到极致，从而影响大模型训练的集群效率。为此需要开发高性能通信加速库，在网卡设备管理、全局网络路由、拓扑感知亲和性调度、网络故障自动告警等方面融入了高性能定制设计的解决方案，以此提升大模型训练的集群效率，优化大模型训练的负载性能，减少网络原因导致的训练中断问题。

2.3.3 高性能存储：提升缓存命中率，降低数据读取耗时

大量计算节点同时读取一批数据集，需要尽可能缩短加载时长。对于文件存储、对象存储架构，需要具备TB级吞吐能力和千万级IOPS，充分满足大模型训练的大数据量存储要求。超大带宽：可以提供超大的内网带宽，满足机器学习场景大带宽需求。多数据源支持：可对接多种数据源，允许存储任意规模的结构化、半结构化、非结构化数据。性能加速：通过数据多级加速服务，实现超越本地HDFS的性能。可以利用数据加速器结合对象存储作为数据存储底座的成本优势，为数据生态中的计算应用提供统一的数据入口，加速海量数据分析、机器学习、人工智能等业务访问存储的性能。相比直接读写对象存储上的数据，数据加速器能够为上层计算应用带来十倍以上的性能提升，极大地提高生产效率。此外，数据加速器需要具备分布式集群架构，具备弹性、高可靠、高可用等特性；为上层计算应用提供统一的命名空间和访问协议，方便用户在不同的存储系统管理和流转数据。

2.3.4 计算加速框架：集成模型工具箱，大幅提升大模型生产效率

数据科学是推动AI发展的关键力量之一，而AI能够改变各行各业。 但是，驾驭AI的力量是一个复杂挑战。 开发基于AI的应用程序涉及许多个步骤（包括数据处理、特征工程、机器学习、验证和部署），而且每个步骤都要处理大量数据和执行大规模的计算操作。需要使用加速计算技术加速数据科学工作流。



▲ NVIDIA CUDA-X AI 架构

以NVIDIA为例，推出软件加速库的集合CUDA-XAI来加速计算。这些库建立在CUDA（NVIDIA 的开创性并行编程模型）之上，提供对于深度学习、机器学习和高性能计算必不可少的优化功能。这些库包括 cuDNN（用于加速深度学习基元）、cuML（用于加速数据科学工作流程和机器学习算法）、NVIDIA TensorRT（用于优化受训模型的推理性能）、cuDF（用于访问pandas之类的数据科学API）、cuGraph（用于在图形上执行高性能分析）等。这些库加快了基于 AI 的应用程序的开发和部署速度。

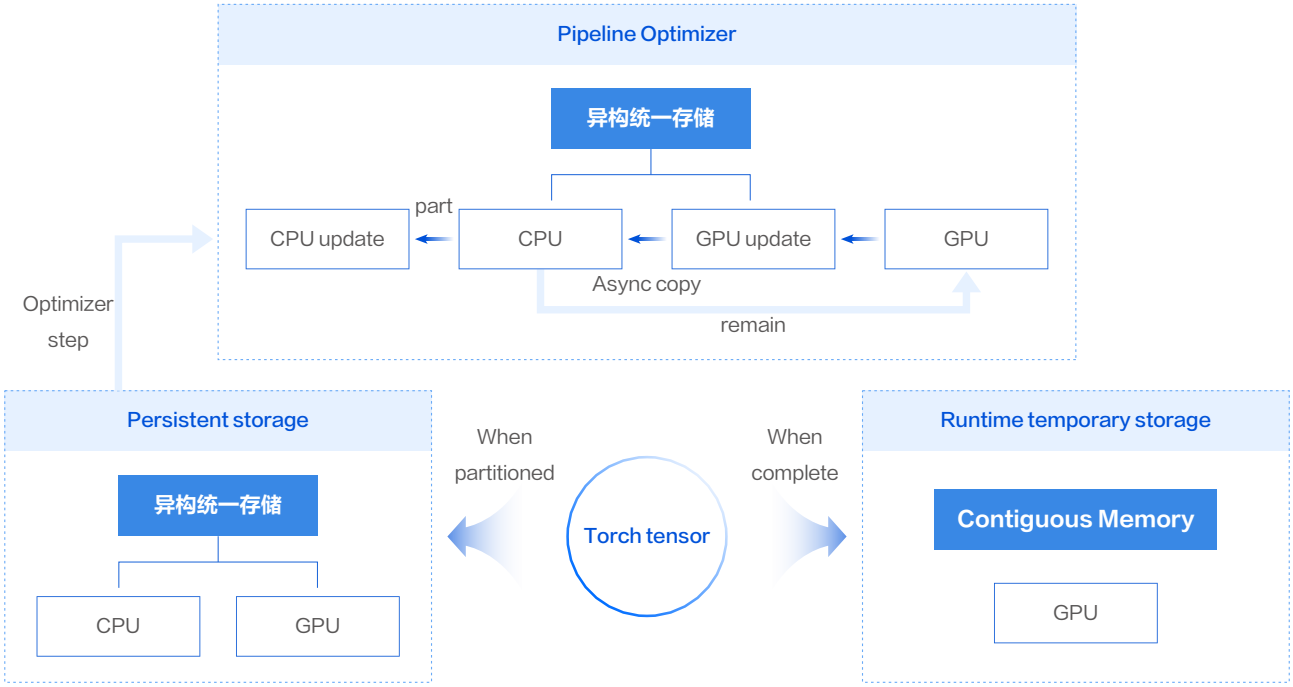
但随着模型参数的快速增长，万亿参数的模型训练仅参数和优化器状态便需要1.7TB以上的存储空间，至少需要数百张高端训练卡，这还不包括训练过程中产生的激活值所需的存储。在这样的背景下，大模型训练受限于巨大的准入门槛。

在大模型训练中，多级存储访问带宽的不一致很容易导致硬件资源闲置，如何减少硬件资源的闲置时间是大模型训练优化的一大挑战。

模型训练时的模型状态存储于CPU中，在模型训练过程中会不断拷贝到GPU，这就导致模型状态同时存储于CPU和GPU中，这种冗余存储是对本就捉肘见襟的存储空间的严重浪费，如何彻底的去处这种冗余，对低成本训练大模型至关重要。

因此在模型生产层面，需要对准入门槛高、多级存储访存带宽不一致、模型状态冗余存储、内存碎片过多等进行优化，提升模型的生产效率。

如在存储优化方面，可采用显存、内存统一存储视角，来扩充存储容量的上限。如太极AngelPTM，基于ZERO策略，将模型的参数、梯度、优化器状态以模型并行的方式切分到所有GPU，自研ZeRO-Cache框架把内存作为二级存储offload参数、梯度、优化器状态到CPU内存，同时也支持把SSD作为第三级存储。



▲ 统一存储视角管理

太极AngelPTM同时将多流异步化做到了极致，在GPU计算的同时进行数据IO和NCCL通信，使用异构流水线均衡设备间的负载，最大化提升整个系统的吞吐。通过将GPU显存、CPU内存统一视角管理，减少了冗余存储和内存碎片，增加了内存的利用率，将机器的存储空间压榨到了极致。

太极AngelPTM可采用大模型训练框架ZERO-Cache，高性能MOE组件，以及数据并行、流水并行、张量并行、专家并行的4D策略，方便用户结合多种并行策略进行大模型训练。其通用加速组件包含可减少显存并提高精度的异构Adafactor优化器，可稳定MOE半精度训练loss的Z_loss组件，选择性重计算组件和降低通信代价的PowerSGD组件。通过PowerSGD梯度压缩技术，对梯度进行低秩矩阵分解后进行通信，降低通信量，提升通信效率。通过提高算力的生产率，可大幅降低大模型训练的算力门槛，缩短训练的时间。算力生产率的提升将大大降低业务应用的算力生产成本，是推进算力绿色泛在的关键因素。

综合以上分析，算力作为数字经济时代的重要生产力，其产出不仅和算力的投入有关，同时和算力的利用率、生产率、以及算力服务化的水平相关。通过提升算力的利用率和生产率可大幅优化算力的投入产出比。

电信运营商作为算力网络建设的主要参与方，在碳达峰和碳中和的背景下，通过技术演进，提高算力效率、降低能耗，是其实现产业碳中和的关键路径；而电信运营商通过算力赋能千行百业的高质量发展，必然要求实现算力随取随用的泛在化和无限可能的智能化。

云服务商在智算中心建设、智算云效能增强、视频云算力平台建设、算力资源融合等应用场景与电信运营商能形成较好的能力互补，在算力优化方面，采用驱动层的GPU共享技术、基于内核层的算力感知和隔离技术、基于调度层的成本优化组件，来提升整体算力利用率；在智算生产方面，采用基于网络层的高性能RDMA网络通信加速库，基于框架层输出统一视角存储管理、高性能MoE、自动流水并行等框架加速能力，基于模型层的算子、编译、计算图等模型优化能力，全面提升智算生产率。在算力服务应用方面，采用音视频编解码、传输、识别、质检、增强等解决方案，提升场景化连接能力。

03

智算服务
赋能算网应用创新升级

随着 5G、大数据、人工智能、区块链等信息通信技术的推广应用，经济社会向数字化转型升级的趋势愈发明显。2020 年以来，国家发布了以“新基建”为导向的一系列政策，旨在通过加快建设数字化基础设施，引领重大科技创新、重塑产业升级模式，为社会发展注入更强动力。

算力服务作为“新基建”的重要组成部分，已经成为整个社会发展的基础，正推动各行业向数字化转型、再造的深水区进军，为各行业带来了红利。算网应用面向垂直行业和具体客户提供适配多样场景的服务，从而加快数字化转型，提升数字化水平。

云服务与电信运营商行业有较大合作空间。一方面是利用云服务公司的技术能力、产品能力，帮助电信运营商补齐在 AI、视频处理、内容生产方面的能力。例如 ASR、数字人等能力，帮助电信运营商完善“服务质量分析”、“远程业务办理”等场景。另外一方面是利用云服务公司在解决方案、生态能力方面的优势，协同电信运营商推动各个行业的数字化转型。例如同电信运营商共同承建文旅、政务等数字化能力提升项目。

3.1 算网应用呈现场景化、多样化、个性化特点

随着算力服务的建设规模持续扩大，算力服务的结构不断演化，算力服务的建设、运营环境也越来越好，主要体现在以下三方面。一是持续优化的算力基础资源和网络环境，为算网应用的孵化提供了基础能力。二是消费和行业数字化需要更多应用来支撑。三是政府的政策利好，企业的大力投资，推动了算力服务生态的发展。正是在这种趋势下，算网应用也在不断的演化、进化。

算网应用不断推陈出新，同当前的主流技术不断融合，应用在千行百业的新场景中，满足生产、运营、管理的需要。算网应用具备新的特点：场景化、多模态、个性化。特点一：场景化，从面向能力到面向场景。特点二：多样化，从较单一种类到多种类百花齐放。特点三：个性化，从集中式统一发展到分布式个性发展。算网新应用的外延，包含两个方面，一方面是技术演进，驱动传统算网应用萌生新活力，例如：交通出行应用。另外一方面是场景创新，激发创新算网应用打开新局面，例如：东N西M应用。通过归纳，梳理出七大典型的算网新应用，如下表。

#	分类	应用	适用场景	应用行业	相关技术
1	传统应用	交通出行应用	路况预测、道路监控、行程规划	交通、出行等	数字孪生、交通 0S、渲染引擎、数据处理、人工智能等
2	传统应用	汽车产业应用	汽车制造辅助、辅助设计和测试	汽车制造、汽车检验等	仿真 HPC 能力、自动云驾驶能力等
3	传统应用	制造产业应用	全场景链接、全流程数字能力	机械制造等	模块化数字能力等
4	创新应用	东N 西M 应用	动画制作、交互式数字人等	政务、传媒、影视、互联网等	专用高速网络、渲染引擎等
5	创新应用	生成式应用	内容创作、自动编程、知识问答等	政务、电信运营商、金融、互联网、信息化等	数据标注、模型调优、内容生成、大语言模型、内容安全等
6	创新应用	数字孪生应用	工厂生产控制、工厂运营监控	机械制造、矿山开采等	数字孪生、5G 远控、渲染引擎、物联网控制等
7	创新应用	数字人应用	内容创作、交互式客服、电商直播等	泛互、文旅等	小样本数字人快速制作、虚拟直播数字人能力等

▲ 典型的算网新应用

3.2 技术演进，驱动传统算网应用萌生新活力

3.2.1 交通出行应用

智慧交通的信息化系统建设大多以实现特定功能为目标，存在整体设计架构固化、开发成本高、复用率低、升级扩展困难等一系列问题，不具备从感知、反应、学习到进化的智能化属性；主流的行业应用大多是传统烟囱式建设，单系统自我生长，不支持多专业业务协同和能力演进，未形成联动多业务协调与能力进阶的体系架构，存在资源浪费、运营效率低、运维繁杂等问题。智慧交通缺乏整体规划和体系架构创新，尚未出现深入的多专业整合、体系架构整体创新的行业解决方案。在智慧交通领域，需要提供覆盖多种交通方式的运营、管理、服务集成整体解决方案。交通OS是面向行业提供的数字交通基础设施之一，兼备工业级稳定性与互联网敏捷性，为企业与合作伙伴输送能力与竞争力。

关键能力：交通 OS

通常情况下，交通 OS 需要具备开放式、组件化的设计理念，融合工业控制、互联网、物联网以及云计算等关键技术，向下连接交通场景中的人、工具、设施、环境、服务，向上支持行业应用的快速构建，汇聚数字技术，融合集成物理空间与数字空间，为企业客户与合作伙伴提供标准化能力的支撑，助力实现数字化转型升级。交通 OS 具备四大关键价值：

01 实现应用解耦。

借助交通OS，智慧交通各系统将业务应用与关键功能模块进行解耦，可复用的功能以组件形式沉淀至平台，帮助应用实现解耦，有利于围绕平台构建应用，随技术发展不断迭代更新底层技术支撑，确保系统随技术迭代发展，不再是推倒式重建。

02 提高开发效率。

相较于传统建设方案，在业务场景应用开发中，利用交通 OS 系统轻松访问所需的设备及数据，复用平台沉淀的基础服务、业务流程、大数据、算法等关键组件和能力，可根据业务需求快速进行业务开发部署，开发效率相较于传统模式大幅提升，并且无需投入大量专业研发人员及额外建设资金，一次建设，可被既有项目和新建项目快速复用，降低系统应用开发、重复建设、接口调试、建设对接成本。

03 打破应用孤岛。

设备、数据等资源由交通 OS 实现统一接入和管理，建设行业统一的数字化资产库，打破智慧交通操作系统间的壁垒，实现数据有效流通。同时，各类算法服务、数据服务及其他通用服务沉淀在平台上，基于平台级接入与开发标准，确保各应用系统实现融合共享，从根本上改变了孤岛式的建设方式。

04 促进场景创新。

基于便捷的“拖拉拽”式的轻开发工具集，进一步降低平台应用开发门槛，使得没有软件开发基础的一线业务人员也可以很容易地参与到交通 OS 的应用生态建设中来，灵活组装所需业务场景，共享并复用开发成果，大幅提高应用开发的生产力。各级应用系统也可以借此灵活组织业务，实现业务的随需动态调整，满足业务快速变化的场景化需求，快速进行场景创新。

▲ 智慧交通操作系统交通OS功能架构

交通 OS 由边缘接入层、平台服务层、应用使能层构成：边缘接入层面向企业的 IT 系统、OT 系统与资源提供强大的连接工具。面向企业应用、系统、服务、API、设备等资源，提供安全、高可用、轻量化的连接器能力，为企业及生态合作伙伴打破应用和系统间的数据壁垒。

平台服务层集成多种关键能力，实现企业各类资源的统筹管理，包含交通物联服务、数字资产管理、流程自动化、工业时序数据库、事件网格、实时数据管道与分析、API 集成、管理与安全、云边协同等能力，实现对企业数字资源的沉淀与协同应用。应用使能层面向开发者与业务人员，提供开放协作平台与轻开发工具集，为用户提供低门槛、开箱即用的数字化转型工具，面向应用提供准化、模型化的数字资产共享与协作平台，形成资源和服务的沉淀与共享能力。

3.2.2 汽车产业应用

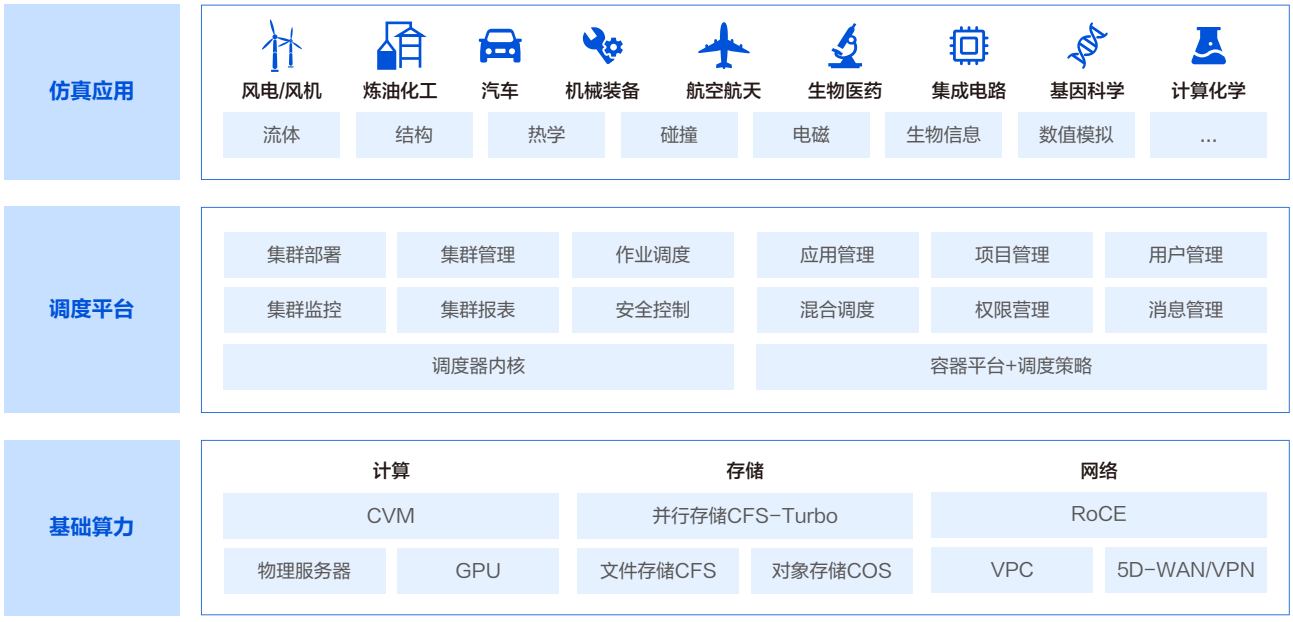
我国大力投入发展智能网联汽车以来，汽车产业已经在数字化的道路上进行了深入的探索：稳扎稳打的国产品牌通过转型实现弯道超车，积淀厚重的跨国巨头持续深入开展本土化变革，一大批造车新势力则依托原生的数字化基因，在国内市场异军突起，有些甚至在国际市场上打开一片新天地。产业变革的动力是技术进步，而技术进步的源泉是人类对美好生活的向往。让用户的交通出行更加安全舒适、让汽车产品和服务的效率更优、体验更好，环保可持续性更高，理应是汽车产业数字化转型的初心所在。云服务商积极参与产业生态共建，持续以数字化技术赋能汽车产业转型升级，旨在通过加快云计算、大数据、AI等创新技术的应用，构建智慧出行的基础设施，助推汽车产业全链路的进化,努力提升全生命周期服务体验。

汽车产业正在从传统工业时代向数字时代迈进，从机械化向以电动化、网联化、智能化和共享化为代表的“新四化”演进。紧抓机遇，勇立潮头，数字化转型不再是汽车行业的“可选项”，而是“必选项”,更是“最优解”。 当下,汽车产业正经历大变革，具体表现在产品形态之变、用户需求之变和产业价值之变这三重转变。转变一产品形态之变，汽车正成为绿色化和智能化的新物种。转变二用户需求之变，Z世代和“她”经济崛起为代表的新消费。转变三产业价值之变，汽车是移动数字生活的新空间。

关键能力：仿真 HPC

近几年汽车行业发生了翻天覆地的变化，传统车企纷纷转型，新势力异军突起，越来越多的车型应运而生。优质车型通常需要进行大量的 CAE 仿真模拟测试，根据行业趋势观察，近 5 年来，随着中国 CAE 行业市场规模持续稳定增长，CAE 软件辅助车企研发生产的重要性日益凸显。

随之而来的 HPC 集群资源需求每年都在成倍增加，车企选择自建 HPC 集群已经无法满足今天市场对研发制造的需求，大规模和突发性成为经常和必须要面对的情况。随着云计算的普及以及在市场需求的驱动下，逐渐出现与汽车产业深度融合的 HPC 仿真云，其架构自下而上由基础算力、调度平台、仿真应用构成。



▲ HPC仿真云架构图

近一两年 CAE 借助云计算服务快速上云，利用公有云每年最新算力机型，弹性的利用模式，用户可以实现上传计算文件、选择求解器、确定配置参数等操作。目前，利用云计算平台进行数值仿真以辅助产品设计分析，已是越来越常见，仿真 HPC 已经能满足较高性价比、适配不同类型 CAE 软件、支持多类 GPU 资源池的需求。

关键能力：自动驾驶云平台

自动驾驶是汽车产业与人工智能、物联网、高性能计算等新一代信息技术深度融合的产物，是当前全球汽车与交通出行领域智能化和网联化发展的主要方向。

许多企业在自动驾驶研发这一领域面临以下难题。难题一：短期内对自动驾驶流程的认知有限，自动驾驶的研发链路和架构规划需要咨询服务。难题二：现有的自动驾驶研发流程比较割裂，不够集中，自动化程度不高，效率低。难题三：现有的仿真效率不够，迫切需要云仿真提高效率，加快研发节奏。难题四：数据存储成本高，仿真软件采购成本高。难题五：对数据合规的理解与风险把控不足。

工欲善其事，必先利其器。自动驾驶平台以数据为核心，专注于为自动驾驶技术研发提供全链路服务。2020 年 5 月 15 日，中国联通正式发布《中国联通自动驾驶网络白皮书》，围绕网络的规划，建设，维护，优化等业务场景，定义了自动驾驶网络 L1~L5 的分级标准和描述，制定了自动驾驶网络的终极目标：实现网络“规、建、维、优”全生命周期的闭环自治。平台广泛集成、聚合行业内的优秀解决方案，有效串联起从数据采集、存储、标注，到感知算法训练、仿真与评测，再到数据回传、数据运营等自动驾驶研发的全链路、全生命周期的方方面面。

▲ 自动驾驶云平台业务架构

3.2.3 制造行业应用

当前，越来越多从工业时代成长起来的企业，开始不满足于旧有模式与规则下的改良，而是寄希望于借助数字化手段实现倍增创新。通过创新，在更短时间内，以更少的人力与资源投入，实现业务的倍速增长、极致的运营效率以及卓越的用户体验。

尽管不同企业在数字化转型中的侧重点各不相同，但本质上都是在做同一件事，即“破圈”:开辟一条新的思维途径、引入更多创新技术、重新定义运营与商业模式。宁德时代制造专家曾提出，“智能制造的未来方向是极限制造”，并成功将动力电池缺陷率提升到十亿分之一。这样的极限效率，仅依赖传统工具与固有经验绝非可能。因此，企业唯有将原有的生产方式、运营流程、供应链体系用数字化的方式重构，方能响应复杂市场环境所催生的极致需求。

关键能力：模块化数字能力

制造业的数字化转型无论从哪里切入，或是朝着什么方向发展，终局都会形成一套复杂系统。转型越复杂，就越需要复杂系统来支撑。但复杂系统仅是转型的结果，不是目的。因此在数字化能力的构建上，采用的是“模块化”的设计理念。

模块化能力架构具体体现四方面的优势。优势一，可集成，更加容易集成来自合作伙伴的新技术、新产品，将其变成“子模块”，不断丰富解决方案内容。优势二，可扩展，当新技术出现时，无需进行整体变更，只需要升级相应模块便可以完成升级。优势三，可复用，模块设计之初会考虑用户的共性需求。通过将共性能力抽象成公共模块，增强能力的可复用性。优势四，可组合，根据企业需求对预设好的模块进行积木式组合，可以快速、低成本地满足不同业务场景的需求。

▲ 面向研、产、供、销、服的整体数字化转型方案

3.3 场景创新，激发创新算网应用打开新局面

3.3.1 东N西M应用

我国东部地区算力需求旺盛，但受土地、电力、能耗指标等限制，算力供应不足。西部地区资源丰富，可利用空间大。在西部部署算力资源，通过灵活调度，承接东部地区算力外溢需求，推动东西部资源和需求的再配置。由此，在西部开展大规模的算力构建，通过高速网络连通，满足东部地区对数据计算、数据存储、内容生成方面的需求，诞生了东数西算、东数西存、东数西训、东视西渲等应用。建立全国算力一体化协同体系，推进数据要素跨域流通，实现资源合理利用和区域协调发展。

东部人多、西部人少	东部发达、西部欠发达	东部耗能高、西部产能多
东部人口密度大、资源紧张，西部地广人稀、年均气温低，有先天资源优势，利于算力设施的建设和运营。	东部GDP占比高，是国家经济发展的发动机，西部需要东部的带动和需求转移，实现协同发展。	东部能耗难以自给，西部产能多，满足自用同时，可在本地承接东部算力需求，实现资源、需求的再配置。
		

▲ 东西部资源分布和经济发展特点比较

东N西M应用不仅需要算力等基础设施，也需要AI训练加速、多媒体云渲染、视频传输编解码等能力，更好的实现低成本AI训练、实时云渲染、低带宽视频传输和低容量存储。在这方面，云服务商已经同电信运营商形成了较为成熟的解决方案，如同电信运营商合作的视频编解码产品，应用在大视频的存储场景和降低视频传输带宽场景，还向电信运营商提供异构算力优化能力，支持多个容器共享一张 GPU 卡的容器共享技术。以容器插件的方式接入业务，提供 GPU 算力与显存灵活切分与隔离能力，助力业务提高 GPU 硬件资源利用率降低使用成本。

关键能力：多媒体云渲染

云渲染是指渲染应用客户端（UE、Unity 等应用）运行在云端 GPU 机器上，用户通过视频流的方式访问云上应用。云渲染并发代表着一系列虚拟计算资源的集合，包含 CPU、带宽、磁盘、GPU 等，一路并发支持一个用户同时访问。多媒体云渲染需要具备三大能力：一是低延时，随着应用场景的丰富以及多媒体终端的普及，对云渲染的实时性要求日益提升；二是高画质，保障用户在使用时可以获得优质的视觉体验以及较低的流量消耗；三是弱网保障，结合 RTC 带宽评估、丢包重传以及智能码控等技术，确保用户在弱网情况下也可得到清晰流畅的使用体验。

关键能力：视频编解码

音视频编解码可将原视频码流转换成另一个视频码流，可调整原始码流的编码格式、分辨率和码率等参数，从而使原视频可以在不同的终端和网络环境下播放，满足不同场景的应用需求。随着人工智能相关技术的发展，音视频编解码可通过视频AI 算法，根据视频场景分类实时识别结果，并结合视频源码率、帧率、分辨率、纹理、运动变化幅度以及机器负载和 ROI 检测等维度，选择最优编码参数，有效提升肉眼画质，减少带宽损耗，通过智能场景识别、动态编码匹配和画质增强修复等功能相结合，实现智能动态编码，可为直播、点播以及媒体等行业，实现以更低码率提供更高清的流媒体服务。

电信运营商的互联网业务和云服务业务市场拓展越来越广，运营商依托云服务商提供的渲染引擎建设“渲染云”，一方面提升云渲染能力，支撑数字人智能服务、全真营业厅等场景，另一方面一同助力传统行业的数字化转型，应用在影视动画制作、数字孪生工厂等场景。

3.3.2 生成式应用

纵览生成式人工智能进化史，从AI诞生之始，人们就试图让机器生成内容，与其对话，并诞生了最早的图灵测试标准。多年来，生成式AI的发展一直不温不火。2022年11月ChatGPT的横空出世，引发了现象级热潮，让生成式人工智能走入了亿万用户的视野。生成式人工智能无论是服务C端还是B端场景，都需要有多种技术为其赋能，例如：数据标注、模型调优、内容生成、大语言模型、内容安全等。生成式应用需要行业大模型和AI能力结合，可以快速提升“对话理解”和“智能问答”能力。比如，在学习了汽车场景的数据后，车载语音助手可根据车辆状态、用户状态、历史数据等信息，做主动触达和场景运营，提供更人性化的场景服务。当前，生成式应用已经在智能客服场景中得到应用，其背景是对智能客服的需求已经从原来的追求效率和分流，到现在更重视的往智能客服的运营价值和智能客服的交互体验上发力。通过数据来驱动智能客服的业务升级，平衡降本增效与用户体验。通过AI的自主学习能力，让智能客服加速自我进化，并通过轻量化方式实现自主运营。运营商与云服务商已经在多个能力方面达成合作，如支撑运营商构建内容标签、智能影集、智能搜索、AI头像、智能超分和插帧、美颜特效等的应用能力，以及文生图、图生图、图片和视频人脸融合等能力。

另外，从用户侧类型划分，生成式应用在C端和BG端市场呈现出两类路径，其中C端已经达到可用、甚至好用的临界点，BG端将从高价值先导领域向模型即服务（MaaS）生态扩展。生成式应用中，模型即服务（MaaS）、内容安全等能力是其中的关键。

关键能力：模型即服务（MaaS）

大模型驱动“智慧涌现”，产业场景已成为最佳练兵场，在智能问答、内容创作、智能决策、智能风控等很多业务场景，具有非常广泛的应用价值。大模型的良好应用还存在诸多挑战。第一，是计算资源紧张。大模型的训练和推理对计算资源和存储资源有很高的需求，对很多客户来说门槛太高。第二，数据质量差。构建大模型是成本极高的系统工程，大模型需要大量的高质量数据进行训练，数据还必须经过清洗和预处理。数据质量差，会导致模型的效果和效率无法得到保障。第三，投入成本高。为确保业务使用的效果需要投入大量的数据、计算资源来训练，还需要持续地调试和优化。第四，专业经验少。大模型的部署需要考虑到计算资源、网络带宽等多个方面的问题，大模型的开发和落地需要很多的技术和人力资源。此外，安全、合规，也是企业需要考虑的关键因素。

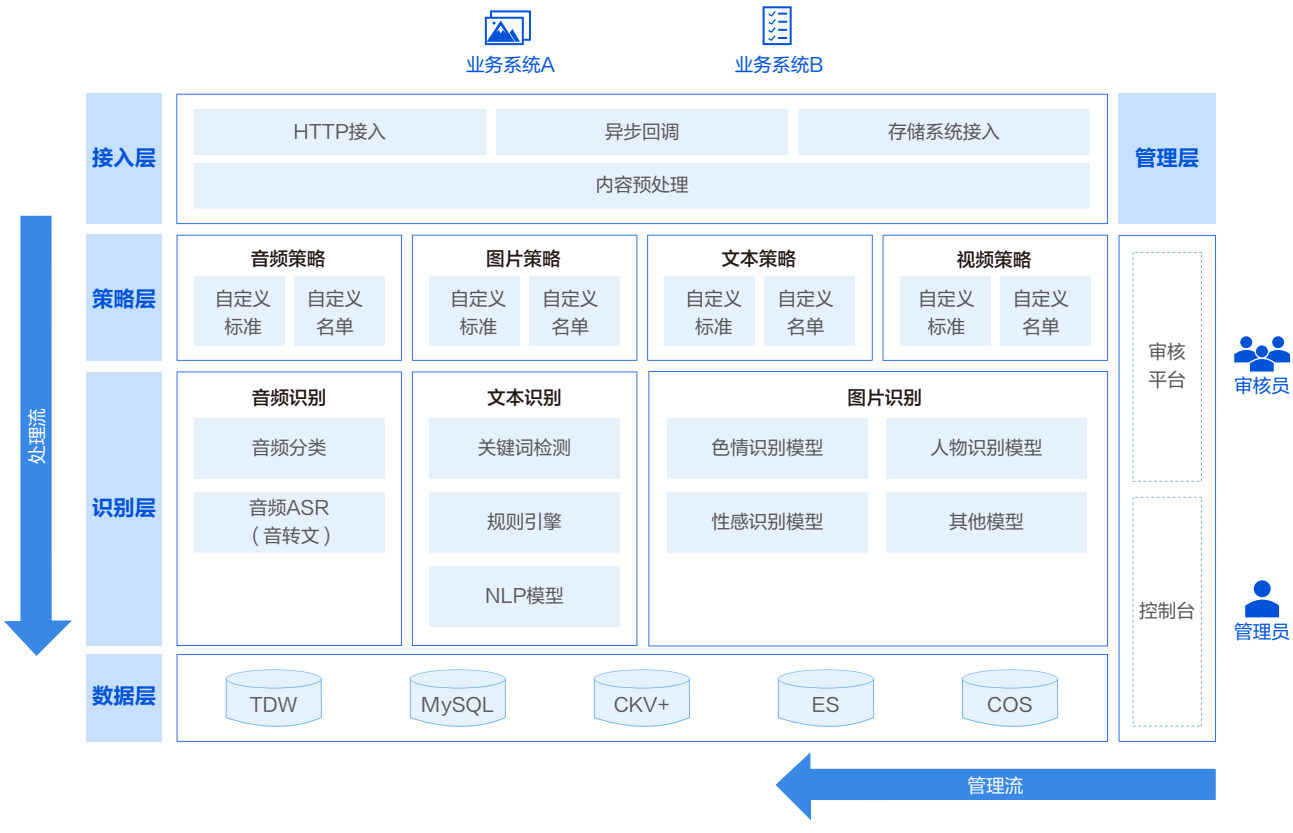
产业各界正在探索行业大模型的应用，2023年7月，中国信息通信研究院联合金融企业、云服务商共同启动行业大模型标准联合推进计划，并率先在金融行业展开研究。云服务商加大在大模型领域的投入，2023年6月亚马逊云宣布投资1亿美元，用于建立生成式AI中心，打造大模型精调工具链，支持客户加入自己独有的场景数据，进行精调训练，客户可根据自身业务场景需求，定制不同参数、不同规格的专属模型。电信运营商加速推进产业智慧化转型，借助云服务商提供的相关能力，升级智能客服、智能质检、网络运维等场景，并进一步扩展到电信运营商的网上营业厅、客服热线、内部知识库等IT系统。

关键能力：内容安全

大模型的应用，安全、合规是前提，在问题侧、模型侧、答案侧三个层面进行敏感信息的过滤和规避，让最终的答案符合安全、规范的要求。

内容安全依托基础 AI 识别引擎（视觉引擎、ASR 和文本引擎等），通过统一的管理调度，提供针对全内容（文本、音频、视频、PDF、WORD 等）的安全检测。防止生成式内容中包含色情图片、谩骂文本、违规音视频。

根据生成式内容安全的需求，构建基于分布式架构的内容安全系统。内容安全系统整体技术框架设计如下：



▲ 内容安全系统架构设计

01 接入层

为企业现有的业务系统提供统一的接入服务，支持两种方式接入，直连方式（业务调用同步接口或异步回调的方式将内容传入内容安全系统），间接方式（业务将数据图片、视频、音频等内容存储到COS里面，内容安全系统自动拉取内容进行识别）；内容预处理针对WORD文件、PDF文件等非文本、图片、音频的文件进行预处理，将里面的内容拆解为文本、图片和音频分别进行识别再合并结果。

02 策略层

策略层针对客户不同业务系统接入支持不同的策略配置，策略配置包括黑白名单配置（支持针对文本、图片、视频的配置），支持不同识别标准。

03 识别层

设计为整个内容识别的关键部件，包括了针对文本、音频、图片的识别引擎，分别包括涉黄、涉恐等类型的，该层的每个模块都支持平行扩容，热插拔。

04 数据层

设计为保存内容识别系统中的产生的数据，包括识别结果数据、访问日志、策略配置等数据。

05 管理层

为企业系统统一的管理系统，支持对每个业务系统接入数据的查询和处置功能，同时为了满足某些高准确率的系统，提供审核台的支持，业务系统安排人工即可对识别为不良的内容进行人工二次审核。

互联网企业每天都面临复杂多样的攻击方式，需要长期与黑客进行攻击与防守的对抗，生成式内容安全方案为企业业务系统安全运行保驾护航，已经在电信运营商的内容审核等场景中得到应用。

3.3.3 数字孪生应用

在数实融合的大背景下，数字孪生作为产业数字化的重要技术，是促进数字经济发展的重要抓手，已经进入一个产业爆发期。抢抓数字孪生发展机遇成为各界共识，数字孪生正处于从探索走向规模化应用的关键阶段。数字孪生综合运用感知、计算、建模等信息技术，将物理空间数字化为可计算的数字空间，对物理空间进行描述、诊断、预测、决策、控制，进而实现物理空间与数字空间的交互映射和闭环控制。数字孪生发展提速，成为产业数字化热点。现阶段，数字孪生被应用于交通、建筑、园区、能源、文旅、城市、制造、环境保护等行业，各行业的需求日益高涨，发展前景广阔。Precedence Research 市场研究机构公布的报告显示,全球数字孪生市场规模在2022年已达到了115.5亿美元，并预计在2022年至2030年期间以38.87%的复合年增长率增长，将达到约1597.7亿美元。

数字孪生工厂是数字孪生在制造业场景中的具体应用，是一种基于数字化技术和数据模型的制造业生产模式，通过对生产过程进行数字化建模和仿真，实现对生产过程全方位监测、优化和管理，以提高生产效率、降低成本、提高产品质量和灵活性。数字孪生工厂可以创建虚拟的生产环境，模拟各种生产场景，从而优化生产流程、预测故障和优化设备维护。

为保障工厂各种设备安全、有序运行，瑞泰马钢打造“数字孪生工厂”，以自动化装备系统、信息化系统（PLM、ERP、PMS等）为基础，以数字孪生底座为核心，利用数字孪生、动态3D大数据可视化、AI数据应用分析及5G远控等新技术，助力厂区“现场环境、产品质量”双提升。

通过“数字孪生工厂”的操作平台实现产品设计、工艺、制造、服务到退役全生命周期数据和商业智能分析（BI）的透明化，并形成装备、数据运行预警与处置的工业互联网应用。其中数据运行预警功能全覆盖厂房各区域，自动对问题与故障进行预测性分析告警与阈值告警，毫秒之间，直接捕捉问题，快速甄别根因告警，协助业务快速恢复运行。

关键能力：数字孪生底座

数字孪生底座提供一站式数字孪生应用的构建能力，支持合作伙伴及客户一站式的完成设备的接入与物联数据处理，空间数据的接入与治理，业务数据的接入与治理，业务数据、物联网数据与空间数据的融合处理，AI 算法的接入与调用，数据及业务的联动编排，数字孪生应用的可视化编排，实现数字孪生应用一站式快速搭建。

数字孪生底座通过无代码 / 低代码的方式为用户提供服务，以降低用户使用平台的技术门槛。物联方面，提供无代码的设备接入工具及标准化的物模型，用户通过简单的配置即可实现感知设备的快速接入、数据获取与消费、设备反控等；空间数据管理方面，提供多类 BIM 文件的导入、数模分离、轻量化、坐标转换与配准、模型渲染工具，可快速实现对三维数据的管理和应用；数据融合处理方面，提供可视化的数据编排画布及丰富的原子节点，可实现对数据的采集、处理、消费的无代码编排，满足各类大数据治理、数据融合处理、实时 / 离线数据运算、复杂数据联动场景编排等；应用编排方面，提供丰富的图表组件和模版，可快速上手，无代码编辑各类二三维一体的数据展示和分析场景，实现数字孪生应用的即时创建即时运行，所见即所得。

数字孪生底座是行业数字化转型的抓手之一，例如联通智网正在建设面向全集团的 IDC 智能化运营平台，重点建设数字可视、自动运维、智能运营、能耗调优、安全可信 5 大板块能力，实现 IDC 运营可视可管可信。

▲ IDC智能化运营平台

关键能力：5G 远控

5G 远程操控应用产品定位于为高危岗位操作、自动驾驶、高危 / 恶劣环境作业提供基于 5G 网络的低时延远程操控能力，可应用于工厂、露天矿区、港口、物流、网联无人机、L3/L4 乘用车等场景。5G 远控通常具备以下特点。

01 极低带宽，缓解无线网络压力，大幅降低网络建设成本。

通过网关硬件编码，能够将单路1920x1080分辨率的视频数据压缩至1-2M码率，常见的需要支持h264、h265硬件编码、av1编码，提供多种的视频压缩方式。

02 实时音视频通信，极低时延，满足流畅操控需求。

与传统的网络摄像头方案相比，5G远控采用P2P的传输架构，并基于音视频传输优化，将端到端的画面延时进行极致的优化。在同城网络环境下，可最低做到130ms的端到端画面传输延时。

03 弱网优化，解决网络覆盖不均匀导致的画面卡顿问题。

与传统的音视频传输场景不同，远控场景对画面的实时性、可靠性有着更加苛刻的要求。5G远控通过跨帧编码降低帧间耦合关系，在数据传输过程中即使出现部分帧丢失，也能还原出丢失部分的完整画面。动态码率调节能实时探测视频通道的丢包及带宽，并自适应调节传输码率，以保障画面传输的稳定性。

5G远控能力已经在能源行业、工业行业、运输行业得到广泛的应用，例如在武汉操控1500公里外鄂尔多斯矿区的卡车，延时低至 100 毫秒，支持电信运营商共同完成远程矿卡驾驶、码头AGV控制等5G远控解决方案的落地。

3.3.4 数字人应用

数字人是运用数字技术创造出来的、与人类形象接近的数字化人物形象。数字人由计算机图形学、图形渲染、动作捕捉、深度学习、语音合成等计算机手段创造及使用，并具有多重人类特征（外貌特征、动作表情能力、人类交互能力等）的综合产物。数字人主要通过动作捕捉、二维/三维建模、语音合成等技术高度还原真实人类。由人工智能所驱动的数字人，拥有近似真人的形象以及逼真的表情动作，唇形动作能与声音实时同步，且具备表达情感和沟通交流的能力。打造出的高度拟人化虚拟数字形象，能像真人般与人互动沟通，带来全新的感官体验。

目前，我国数字人商业化应用场景越来越丰富，数字人已经在运营商、金融、传媒、游戏、文旅等行业做出快速探索。这得益于数字人产业底层技术、应用平台的高速发展，从技术开发到落地应用的产业链也正趋于完善。数字人的发展也呈现出多种趋势，重点的趋势如下。趋势一：数字人制造和运营服务的B端市场不断扩大，将面向更广大的C端用户提供服务，各类数字人价值定位和商业模式有差异。趋势二：技术集综合迭代驱动数字人形似人，制作效能将继续提升。趋势三：AI技术驱动数字人多模态交互更神似人，并逐步覆盖数字人全流程。趋势四：数字人技术与SLAM、3D交互、体积视频、空间音频等技术深度融合，渲染将从本地到云端。

关键能力：小样本数字人快速制作能力

小样本数字人即通过少量的小样本素材（3 ~ 5 分钟），即可导入训练模型，生成与真人无异的数字人分身，五官、动作、表情完全模仿真人。仅需通过输入文本或音频，即可快速生成数字人分身视频，大幅节省每次拍摄的时间、空间、用人成本。适用于内容讲解、口播视频生产、直播带货等需要真人出镜的场景，节约成本，全年无休。

绿幕 （视频换背景效果顺序依次为绿幕蓝幕、纯色幕、其它。且幕布的颜色和衣服桌子要有差异）	提词器、三脚架、摄像机或手机（拍摄） 手机或pad（题词）	灯光
		

▲ 小样本数字人拍摄物料

坐姿示例	站姿示例
	

▲ 小样本数字人拍摄景别

小样本数字人制作，将数字人的应用场景从企业使从企业专属，扩展至更广泛的个人商用场景，如电信运营商的5G名片、泛互行业的电商直播场景等，助力运营商更好的提升内容制作效率及营销转化率。

关键能力：虚拟直播数字人能力

随着全真互联概念持续走热，作为人类未来在全真互联世界的虚拟分身，数字人直播也受到广泛关注。当下，将数字人嵌入真实场景或者虚拟空间进行直播，已经成为主播直播的新方式。从直播电商成交额来看，从无到万亿成交，市场仅花了不到4年时间。2017年中国直播电商成交额为268亿元，2020年上升为12881亿元，增长4700%，发展迅速。2021年上半年，我国直播电商成交额就达到了10941亿元由于直播电商行业发展迅猛，在线直播用户规模也在不断增加。数据显示，2021年我国在线直播用户规模达6.35亿人，同比增长8.2%。预计2022年我国在线直播用户规模将进一步增至6.6亿。在数字人、AI、5G、AR等技术的推动下，虚拟直播将迎来发展红利期，数字人将取代真人主播，真正实现24小时、更低风险的直播。数字人和虚拟空间的结合可解决直播中人、货、场的问题，支持用户六大需求。需求一：多样化的主播形象选择，支持平台预设数字人主播形象租赁，支持批量低成本生成和真人主播一样的数字人主播，降低主播人力投入。需求二：提供文字/语音/视频的多元化驱动模式，真人主播可随时接管，同时数字人NPC可协助主播增加直播体验和趣味性。需求三：数字人客服，支持数字人客服接入，预设问答知识库，降低真人客服成本投入。需求四：灵活的直播业务配置，客户可通过接口获取主播列表、主播可用资源列表、替换背景图、调整数字人的大小和位置。需求五：多平台直播推流，输出带绿幕视频，直播平台可将直播视频推流至视频号、抖音等。需求六：aPaaS交付，降低客户接入成本，提供aPaaS接口的低成本交付模式，客户按量付费。

04

算网应用
未来发展趋势

应用发展上，MaaS 将引领算网应用新一轮产业变革

模型即服务(Model as a Service)是指通过云服务将数据处理和机器学习模型的功能集成到现有业务中，为企业提供智能化、自动化的解决方案。通过 MaaS 的数据处理、数据分析、智能决策、模型训练等能力，帮助客户构建自有的行业大模型应用，将成为算网应用的新发展方向。MaaS 支持用户直接访问和使用典型模型，无需在模型开发和训练投入更多精力，极大地节省了时间和资源投入。MaaS 有效支撑算网新应用深化产业渗透，将成为提升企业和个人生产与生活效率的主要方式之一。

服务模式上，将形成通用应用与专用应用长期并存、高效协同的模式

“通用算力 + 专用算力”将成为人工智能算力基础设施的关键。算力基础设施应满足广泛应用场景的通用性，并支持高要求个性化应用场景的高效性。随着全球数据量的指数级增长，人工智能、区块链、数据中心和边缘计算等场景对算力的需求不断增强，为了应对多元化的算力需求和应用场景，未来基础计算架构将不断引入更多种类的基础资源来加速计算，除基础通用计算的 CPU 计算单元外，还包括如 GPU、DPU 以及 AI 加速芯片等异构资源以及专用硬件计算芯片等。现阶段芯片提供商多依靠自身硬件条件构建计算架构，彼此之间存在较大差异，难以实现应用跨架构的开发、迁移等。未来将通过开源框架、开源接口等方式建立统一、规范且支持屏蔽底层软硬差异的计算架构平台，支撑不同类型资源间实现联合协作，从底层优化算力服务性能。

发展格局上，跨架构、跨地域“双跨”应用将有力支撑全国算网一体化发展

算力服务依托相对成熟的云计算技术，综合考虑用户计算需求，算力、网络等多样资源状态，构建全域一体、算网融合的多要素融合编排体系，完成从调度单一资源到调度多样资源的跃迁。具备多要素融合编排调度能力的算网大脑产品已成为算力服务在融合调度领域的典型落地实践，将来，可以根据算力的性能、模态、单价等信息的综合判断，形成可支持跨架构、跨地域的算网编排方案，并完成相关资源部署，以支多场景运算需求。

腾讯云小程序