甲子光年
JAZZYEAR

2023

中国AIGC产业算力发展报告

AIGC爆发，算力服务机会or变革?

出品机构：甲子光年智库
报告指导：宋涛
报告撰写：刘瑶
发布时间：2023.08

前言

- 伴随着ChatGPT的爆红，AIGC产业链受到广泛关注，大模型的发展正推动AI算力市场进入新的发展阶段，强大的AIGC算力基础设施正在构建当中。大模型训练是复杂系统工程，AIGC产业的算力也对应是系统化的建设，基于此，甲子光年智库特此展开AIGC的算力研究，输出《中国AIGC产业算力发展报告》，在瞬息万变的AIGC产业发展过程中，试图捕捉到AIGC产业算力的发展变化。

AIGC时代已来，算力作为推动产业发展的关键资源，决定了产业的发展速度及

AI算力不止于训练，同时支持AI多场景应用，在多细分场景上具有潜力

AI技术（算法、模型）与算力的融合催生更多产品及服务模式

AIGC的爆发会重新改变负载AI技术的算力发展

本次报告探讨的问题

- *明晰概念：从需求出发，探究决定AIGC产业基础设施（infrastructure）——算力，包括哪些关键决定因素？*
- *产业分析：AIGC的算力产业链剖析，从芯片发展到云服务方案，”云边端“算力供应商的角色作用是什么？*
- *需求探讨：中国AIGC产业发展是否缺乏算力，还是缺乏针对企业的“高性价比”及“可落地”的AIGC算力解决方案？*
- *实践指南：针对当下国内的“百模大战”与企业对于AIGC应用落地的需求，目前AI算力领域有哪些解决路径及方案？*
- *趋势可能：AIGC的算力爆发是否可持续？未来对AIGC的算力提出哪些层面的要求？*

甲子光年重点关注企业——AIGC产业算力领域的领航者

浪潮信息

“基于大模型自身实践与服务客户的专业经验，**浪潮信息**发布大模型智算软件栈OGAI（Open GenAI Infra）‘元脑生智’，OGAI以大模型为核心技术的生成式AI开发与应用场景，提供从集群系统环境部署到算力调度保障和大模型开发管理的全栈全流程的软件，从而降低大模型算力系统的使用门槛、优化大模型的研发效率，保障大模型的生产与应用。”

intellifusion 云天励飞

“应用产生数据、数据训练算法、算法定义芯片、芯片赋能应用”是**云天励飞**一直坚持的人工智能发展之路。基于此，云天励飞构建了业界领先的算法、芯片、大数据全栈式AI能力，同时拥有大量创新应用和落地场景，横跨人工智能基础层、技术层和应用层。”

UCLOUD优刻得

“UCloud**优刻得**是中立、安全的云厂商，拥有超10年的公有云技术沉淀并积累了全面的系统工程能力，具备从数据中心、计算平台，到管理平台、网络服务、应用服务、生态接口等一站式产品和解决方案。凭借技术及工程能力沉淀，UCloud优刻得可在AIGC领域构筑强大的算力底座，通过优化网络和存储带宽提升大模型训练效率，并持续提供快速、高效、可控及安全的AI推理环境。”

Witmem 知存科技

“凭借存储与计算物理融合的架构优势，存内计算能够为神经网络模型指数级增长的算力需求提供强大基石。**知存科技**凭借率先量产商用存内计算芯片的产业积累，将继续推进存内计算架构创新，由3D存内计算架构向高速互联存内计算架构演进，实现产品“端-边-云”侧算力全面覆盖。”

博云

“BoCloud**博云**形成了系列产品以创新云技术支撑企业核心业务，构建数字化高效IT系统。公司自主研发的多项软件产品，包括边缘计算平台、企业级容器平台、统一云管平台、虚拟化产品等，已在金融、电力、石油、政务、IDC、航空等行业领域的生产系统中落地实施，为国有电力公司、股份制银行、大型支付机构等标杆行业客户的重要生产系统提供服务。”

亿铸科技

“亿铸科技在全球率先利用ReRAM（RRAM）的特性着手使用先进异构封装的方式来实现系统级的芯片优化方案，并且在2023年3月，亿铸科技正式公布了存算一体超异构芯片的创新理念——以存算一体(CIM)AI加速计算单元为核心，同时将不同的计算单元进行异构集成，以实现更大的AI算力以及更高的能效比，同时提供更为通用的软件生态，开创大模型时代的AI算力发展新方向。

目录

CONTENTS

甲子光年
JAZZYEAR

Part 01 产业基石，算力是AIGC产业的催化剂

Part 02 软硬兼得，AI新世代呼唤工程化导向的算力支撑

Part 03 层见叠出，商业浪潮下的算力选择思考

Part 04 实践真知，AIGC产业算力实践的新范式

Part 05 来日正长，AI技术的翻涌带来无限可能

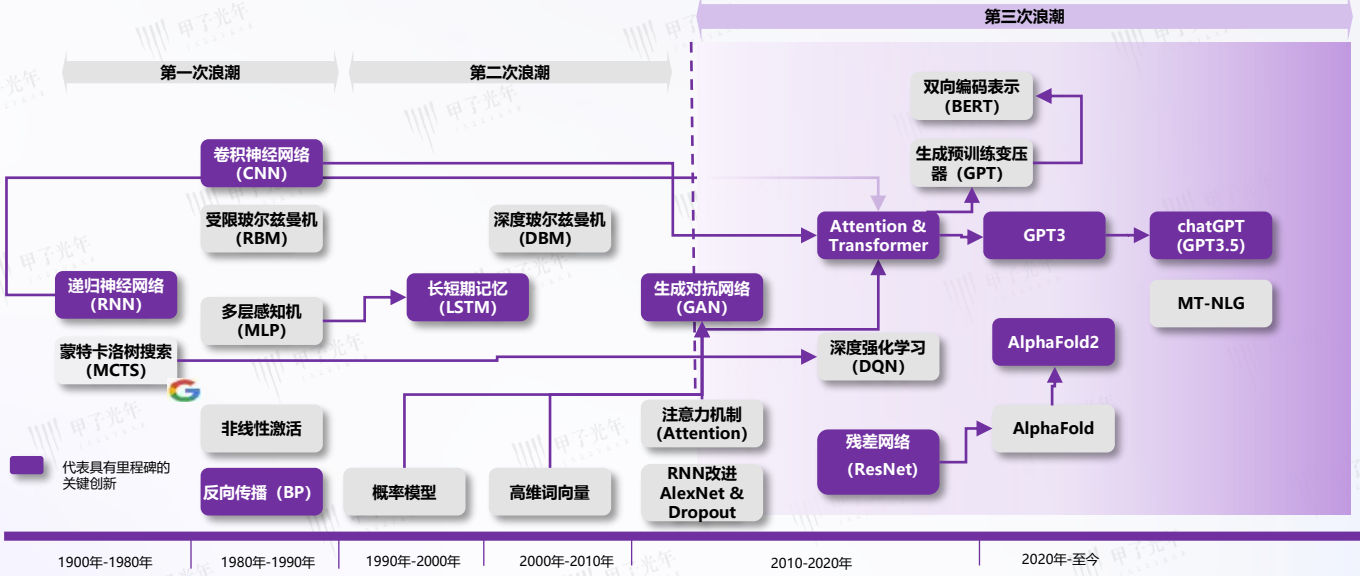
AI的新时代，生成式AI技术重新塑造AI技术的开发及应用

- 随着2023年大模型的热潮，AIGC早已超越内容生产的概念，而突出**生成式AI（Generative AI）**的概念，即如何通过生成式AI的技术思路解决以往决策式AI难以完成的问题，尤其在数据或者内容生成上实现 “质的突破”。
- 新的AI时代则是AIGC产业全面进击的时代，随着生成式对抗网络（Generative Adversarial Network，GAN）等的演进及迭代，生成式AI可以延展到流程、策略、代码、蛋白质结构等多种形式，即意味着凡是可以使用数字内容形式的产业，生成式AI均可以涉及。

AIGC

AI的新时代：更关注如何利用生成式AI技术在涉及数字内容的诸多领域实现改变及突破，生成式AI实际上扩大了“内容”的含义，凡是可以数字化的内容形式均为生成对象，而非传统意义下媒体环境的内容。

AIGC产业：新一代AI技术和理念，以“生成式AI”为代表技术的开发及应用产业，即如何利用资源发挥新的AI技术的应用，通过商业价值推动AI第三次浪潮的发展。



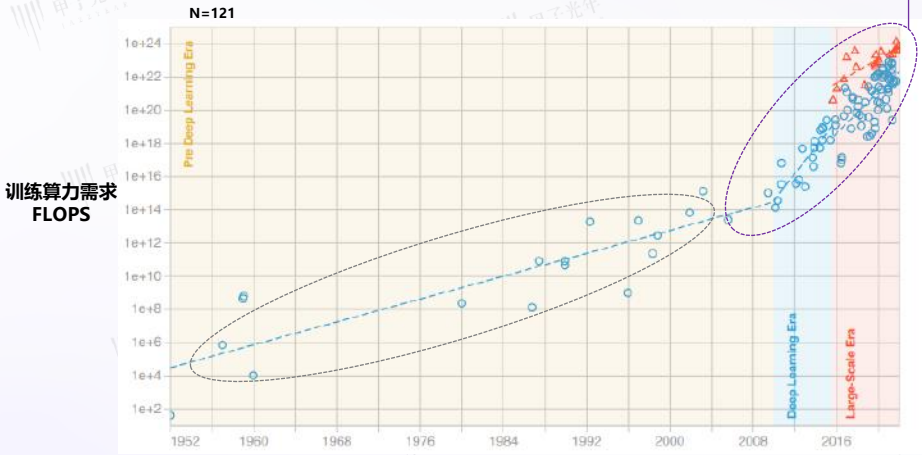
数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

纵观AI发展，算法的发展及迭代极大地拉动了算力的需求

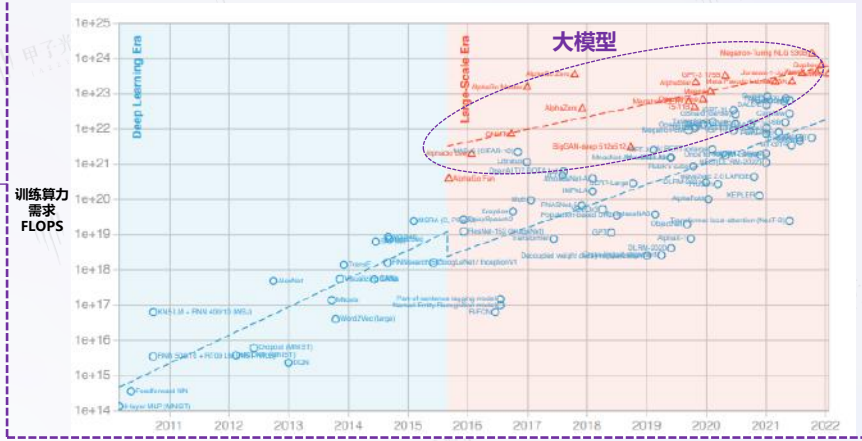
- 机器学习的训练计算大概可以分为三个时期，2015-2016 年左右开启了大模型时代，整体的训练计算量较之前的时期大2到3个数量级。
- 从2022年底，随着ChatGPT成功带来大规模参数通用大模型相继发布。这些大模型的训练需要千亿、甚至万亿级参数，以及上千GB的高质量数据，大模型的训练迭代将极大地拉动了智能算力的需求。

2010前	**深度学习时期之前，训练计算算力需求缓慢增长，算力翻倍需要21.3个月**
2010-2022	**深度学习不断取得进展，算力翻倍仅需要5.7个月，所需算力量级由 TFLOPs增至EFLOPs**

训练算力（FLOPS）需求与人工智能发展关系图（1952-2022年）



训练算力（FLOPS）需求与深度学习发展关系图（2000-2022年）



2016-2022	2015年开始逐渐出现大规模（更大参数量）模型，算力需求直接提升约两个数量级。

数据来源：《COMPUTE TRENDS ACROSS THREE ERAS OF MACHINE LEARNING》Jaime Sevilla等著、甲子光年智库总结整理, 2023.08

深究AI开发，“量变”的算法、数据可以带来“质变”

甲子光年 JAZZYEAR

“量变”

模型训练涉及的基础资源提升在方向上（理论上）能够决定模型训练的效率和结果

- 参数规模
- 充足的能耗
- 数据质量
- 数学理论
- 分布式计算效率
- ……

实际上可以看作“必要不充分条件”：难以明确的直接因果关系

调参过程实际上类似于“实验”：“调参”的结果与以往人工智能方式相比，具有更多的不确定性，需要进行多次的反复训练，模型训练中，模型即是训练结果，中间的过程则无法完全复制。

“质变”

训练的过程呈现“黑盒”性质

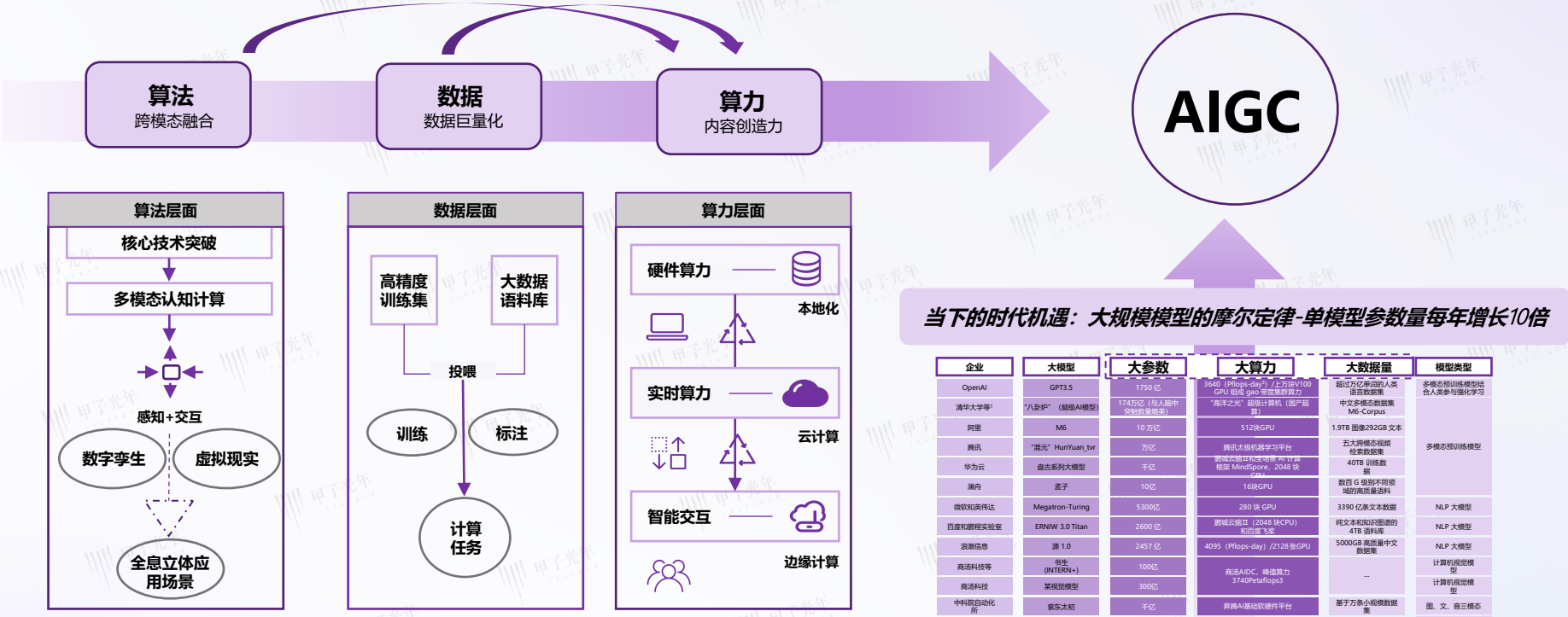
模型调参

收集数据及数据准备	特征工程	模型选择	模型训练	模型评估	模型调优	模型部署和应用
包括数据收集、预处理、存储；数据的质量和梳理对算法效果至关重要	从原始数据提取有代表性、可解释的特征	自开发实现模型使用现成的框架	无监督学习 监督学习 强化学习 ……	根据分类、回归、聚类等各种问题有相应评估指标	网络搜索、随机搜索、贝叶斯优化、梯度优化、集成方法等选取最优参数组合，提高模型性能	模型封装后，导入环境进行推__ API部署 边缘部署 集成部署 容器化部署 批处理部署

算力：提供底层动力源泉

资源“三剑客”中，算力承接算法及数据，成为AIGC产业发展基石

- 在现代人工智能领域，算力扮演着推动创新、实现突破的核心驱动力。算力、算法、数据和系统架构等多个方面的综合优化对于大规模模型训练的成功至关重要。从技术层面看，在大模型的研发过程中，预训练、微调和模型推理等环节是核心关键因素和主要计算特征。



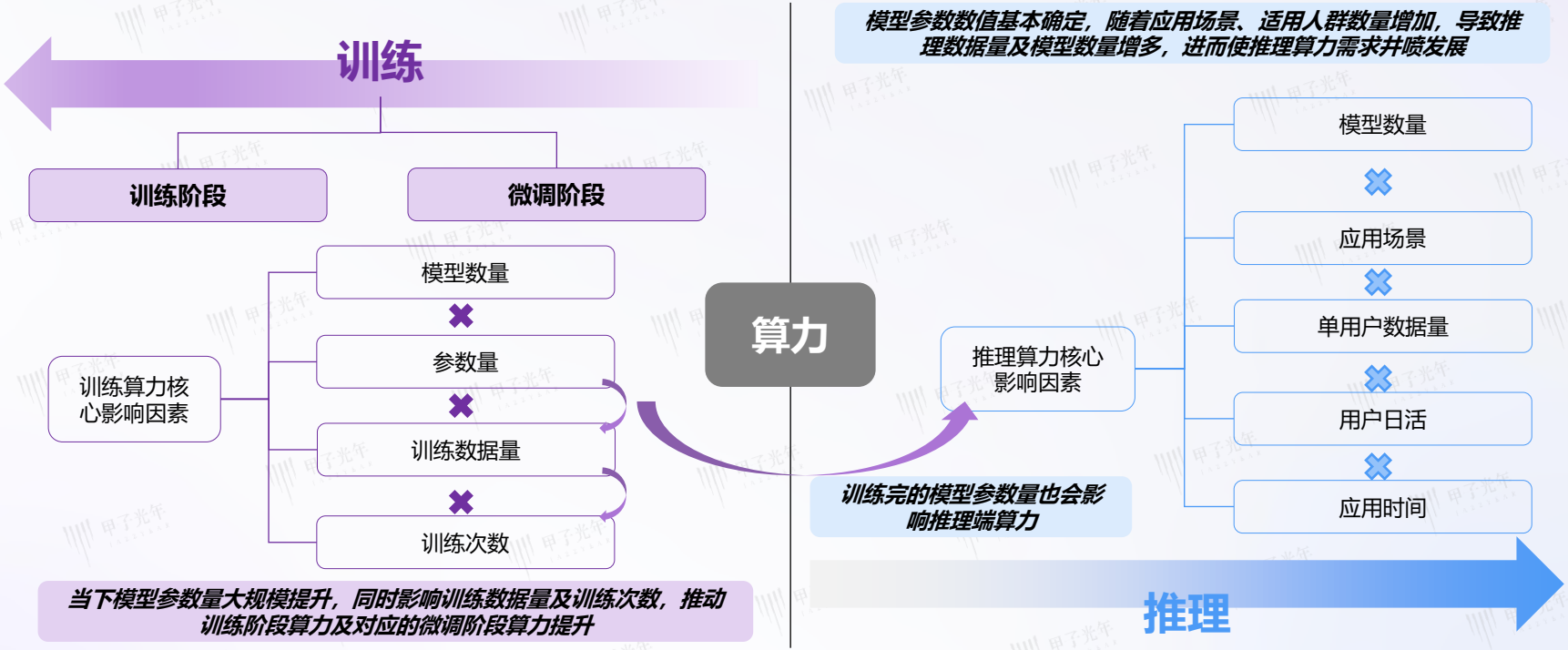
当下的时代机遇：大规模模型的摩尔定律-单模型参数量每年增长10倍

企业	大模型	大参数	大算力	大数据量	模型类型
OpenAI	GPT3.5	1750 亿	3640（Pflops-day*）/上万块V100 GPU 组成 gpu 可交互计算力	超过万亿单词的人类语言数据集	多模态预训练模型结合人类参与强化学习
清华大学等[1]	“八卦炉”（超级AI模型）	174万亿（与人脑中突触数量相等）	“神洋之光”超级计算机（国产超算）	中文多模态数据集 M6-Corpus	
阿里	M6	10 万亿	512块GPU	1.9TB 图像292GB 文本	
腾讯	“混元” HunYuan_tvr	万亿	腾讯太极机器学习平台	五大跨模态视频检索数据集	多模态预训练模型
华为云	盘古系列大模型	千亿	鹏城云脑II和全场景 AI 计算框架 MindSpore，2048 块	40TB 训练数据	
澜舟	孟子	10亿	16块GPU	数百 G 级别不同领域的高质量语料	
微软和英伟达	Megatron-Turing	5300亿	280 块 GPU	3390 亿条文本数据	NLP 大模型
百度和鹏程实验室	ERNIW 3.0 Titan	2600 亿	鹏城云脑II（2048 块CPU）和百度飞桨	纯文本和知识图谱的 4TB 语料库	NLP 大模型
浪潮信息	源 1.0	2457 亿	4095（Pflops-day）/2128张GPU	5000GB 高质量中文数据集	NLP 大模型
商汤科技等	书生（INTERN+）	100亿	商汤AIDC，峰值算力 3740Petaflops3	--	计算机视觉模型
商汤科技	某视觉模型	300亿			计算机视觉模型
中科院自动化所	紫东太初	千亿	昇腾AI基础软硬件平台	基于万条小规模数据集	图、文、音三模态
复旦大学	MOSS	百亿	复旦大学超算中心	--	对话式大型语言模型

数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

AIGC的突破依赖于算力的“暴力美学”，应用依赖于算力在场景中的释放

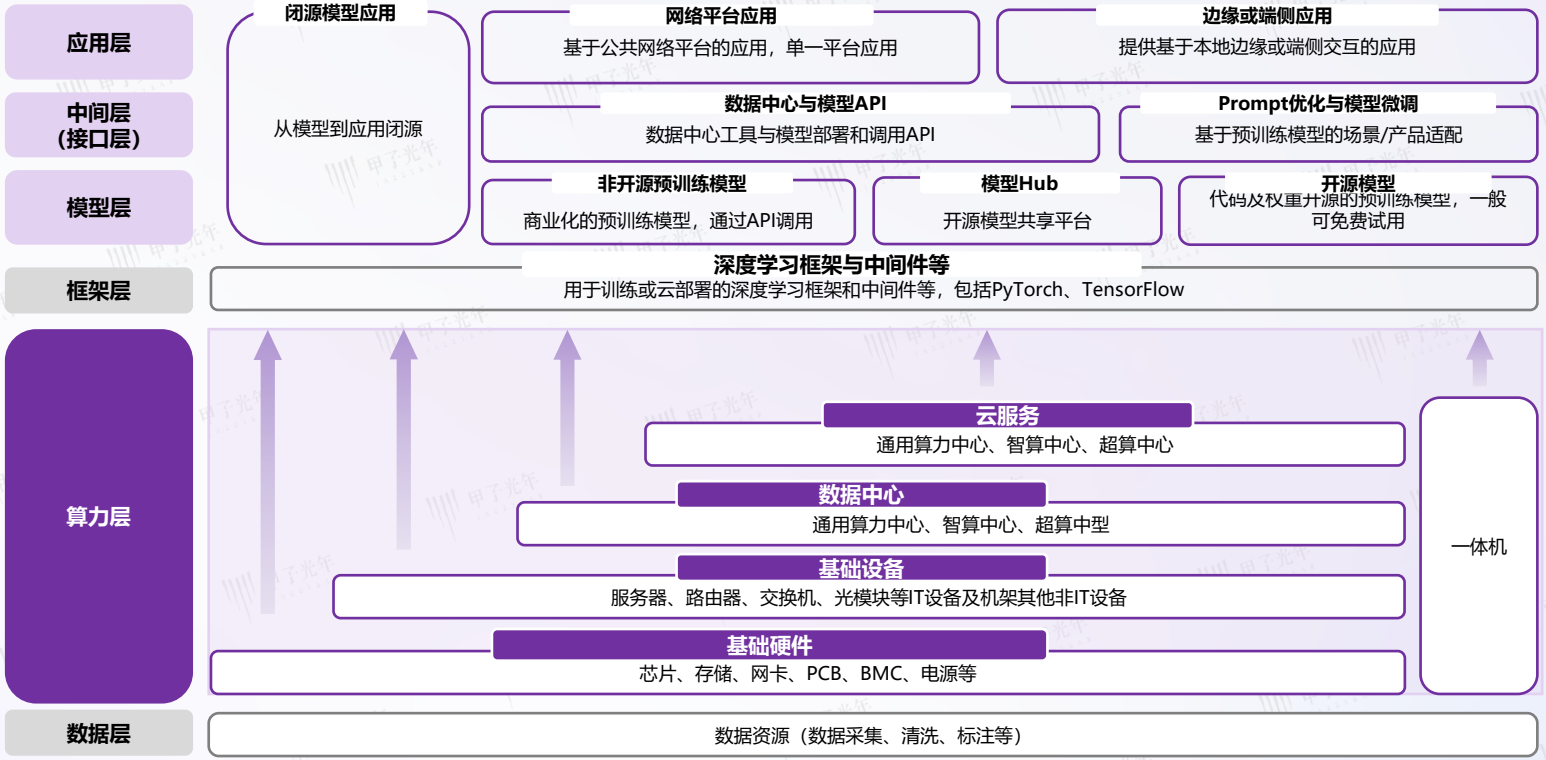
- AI技术在实际应用中包括两个环节：训练（Training）和推理（Inference），AIGC的算力需要考虑训练及推理两个方面。
- 训练是指通过数据开发出AI模型，使其能够满足相应的需求，一般为AI技术的研发。因此参数量的升级对算力的需求影响大。
- 推理是指利用训练好的模型进行计算，利用输入的数据获得正确结论的过程，一般为AI技术的应用。推理部署的算力主要在于每个应用场景日数据的吞吐量。



数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

AIGC算力具备软硬件的复杂性，并且以不同产品/服务/方案为应用赋能

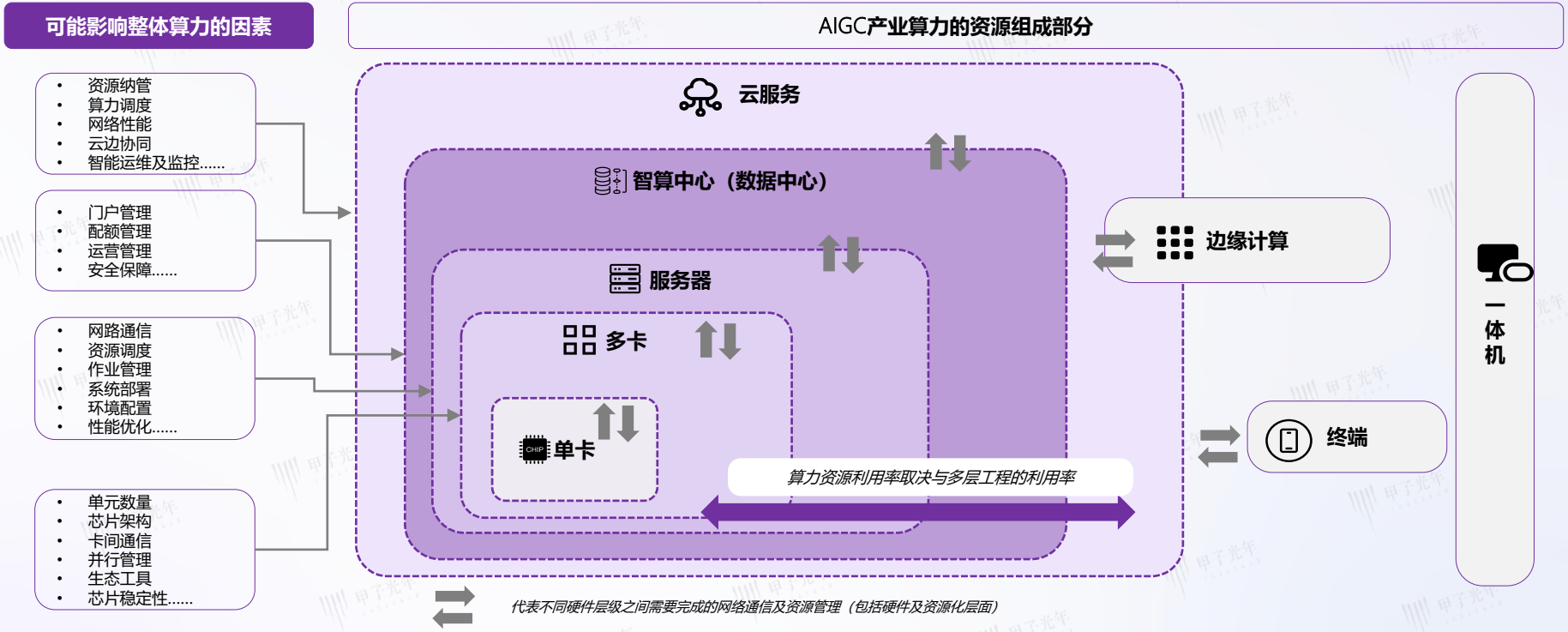
基于AIGC的技术栈，算力层作为上层模型集应用的重要支撑



数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

AIGC产业的算力是工程化结果，是从芯片到资源服务的多层次构造

- 未来大模型的产业化发展是一套复杂的系统工程，构建高效稳定的算力平台是核心要义，成熟的算法、数据产业链，配套工具链及丰富的生态链是关键因素，亟需以系统的方式寻找最优解。
- 算力设备软硬件兼容性和性能调教上的Know-How，可以保证AI算力的适配性和稳定性，并非单一因素的参数能简单决定。

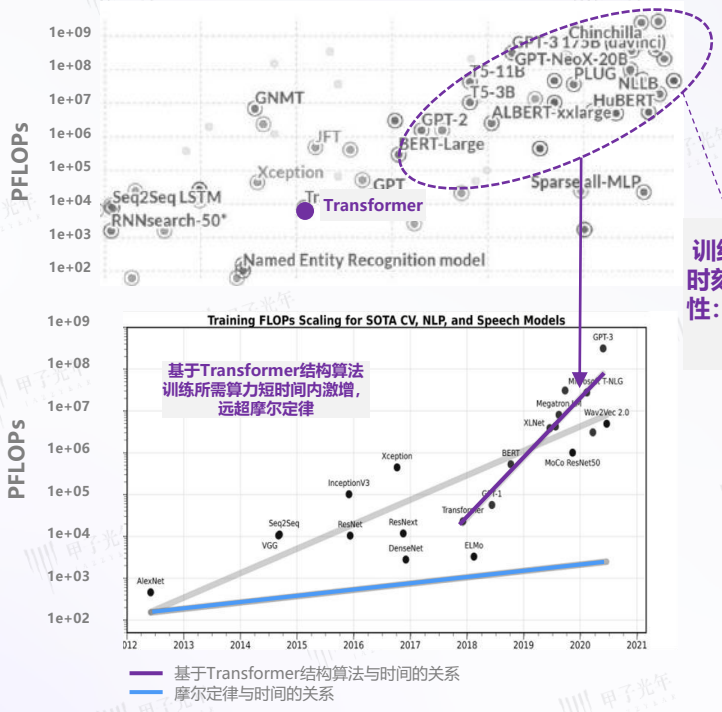


数据来源：公开资料、浪潮信息，专家访谈、甲子光年智库总结整理，2023.08

Transformer结构解锁新AI时代的iPhone时刻，实现资源及应用的飞轮增长

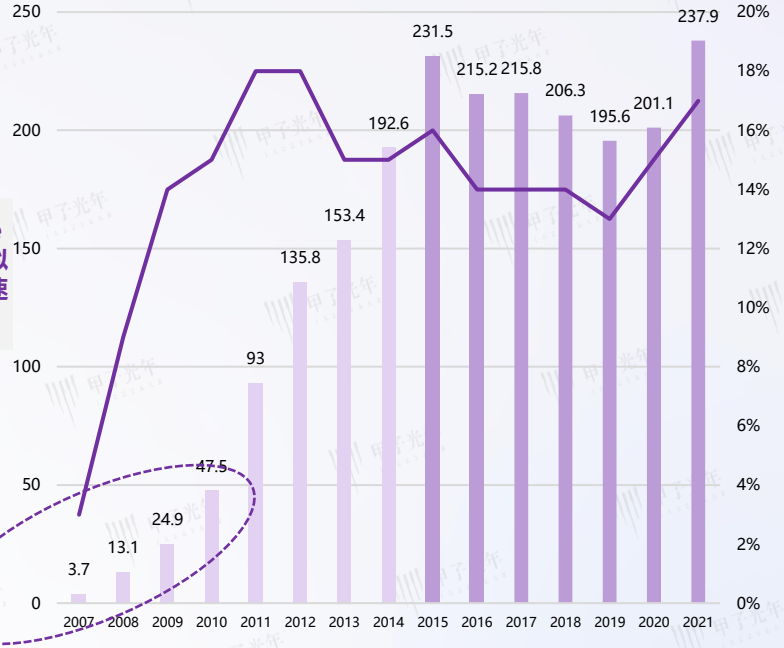
- Transformer的应用标志着基础模型时代的开始（基础模型的庞大规模和应用范围突飞猛进），模型参数量指数级增长，带动算力超过摩尔定律，可以称为AI技术的iPhone时刻。
- 技术层面上，基础模型通过迁移学习（Transfer Learning）和规模（scale）得以实现；深度学习中，预训练又是迁移学习的主要方法：在替代任务上训练模型（通常只是达到目的的一种手段），然后通过微调来适应感兴趣的下游任务。迁移学习（Transfer Learning）使基础模型成为可能。

Transformer结构对于基础模型训练算力需求的推动作用



训练算力与iPhone时刻的阶段具有类似性：实现革命，快速增长

2007-2021年苹果iPhone出货量（百万台）及市占率

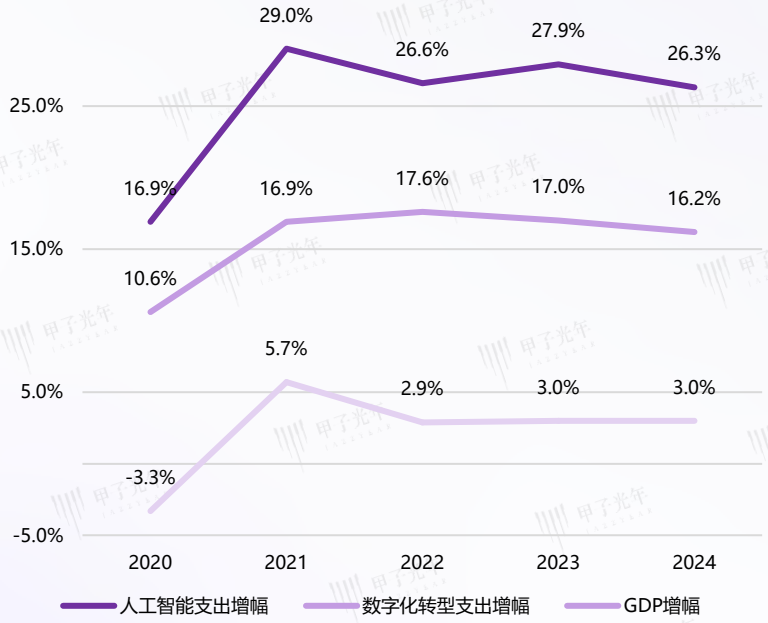


数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

数字经济时代，人工智能相关产业必然迎来广阔发展，AI算力正当时

- 企业数字化转型推动人工智能领域支出，人工智能支出已经成为支持企业数字化转型支出的主力之一。IDC数据统计，全球范围内，企业在包括硬件、软件和服务在内的人工智能（AI）市场的技术投资从2019年的612亿美元增长至2021年的924亿美元，预计将在2022年（同比）增长26.6%至1170亿美元，有望到2025年突破2000亿美元，增幅高于企业数字化转型支出整体增幅。
- 从国家到地方，关注通用人工智能的系统建设，探索通用人工智能新路径，推动创新场景应用成为思想共识、政策共识、发展共识。

全球人工智能支出、数字化转型支出及GDP增长趋势预测，2020-2024



从中央到地方，出台相应政策推动通用人工智能发展

2023年4月

中央重视通用人工智能发展

中央政治局召开会议，指出要重视**通用人工智能发展**，营造创新生态，重视防范风险。为贯彻落实国家相关决策部署。

2023年5月

发改委：加快发展数字经济，重视通用人工智能发展

把握数字化、网络化、智能化方向，大力推进数字产业化和产业数字化，重视**通用人工智能发展**，支持平台企业在引领发展、创造就业、国际竞争中大显身手。

2023年5月

北京市：《北京市促进通用人工智能创新发展的若干措施》

提出系统构建大模型**等通用人工智能**技术体系，开展大模型创新算法及关键技术研究，加强大模型训练数据采集及治理工具研发，建设大模型评测开放服务平台，构建大模型基础软硬件体系，发展面向通用人工智能的基础理论体系。

同时北京正在加快推进国家新一代人工智能创新发展试验区和国家人工智能创新应用先导区建设，打造具有全球影响力的人工智能创新策源地

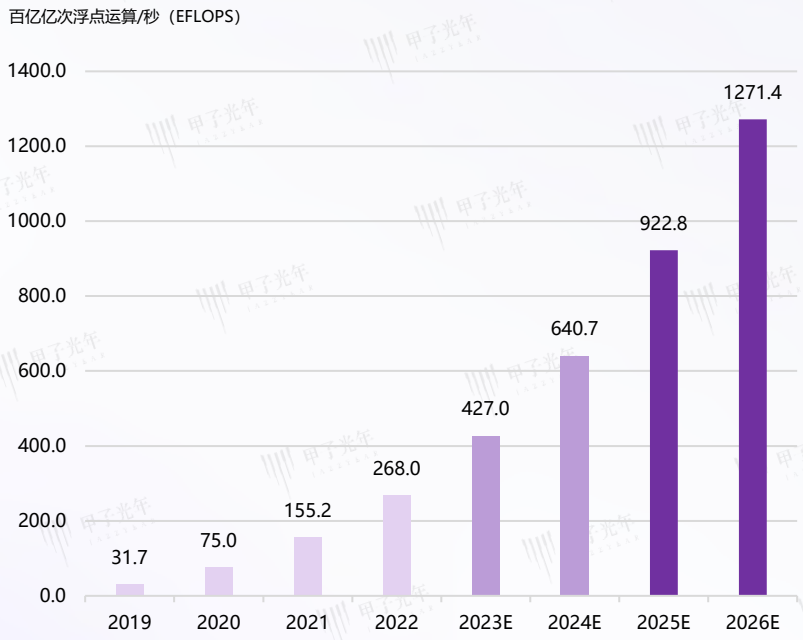
- 北京市经信局联合市科委中关村管委会、市发改委等共同发布「北京市通用人工智能产业创新伙伴计划」，旨在搭建人工智能大模型的开放合作平台，建立协同合作机制，通过持续优化产业链布局，大幅提升优质算力、高质量数据供给支撑能力，培养一批应用大模型技术实现突破性成长的标杆企业。

数据来源：IDC&浪潮信息、公开资料、甲子光年智库总结整理，2023.08

模型的训练及推理离不开智能算力，AI应用催生智能算力需求

- 通用大模型、垂直行业大模型的训练及微调，及基于大模型推理的行业应用需要大量的AI算力实现支撑。AI渗透千行百业，拉动智能算力规模高速增长。2022年，各行各业的AI应用渗透度都呈不断加深的态势，尤其是在金融、电信、制造以及医疗领域。

中国智能算力规模及预测，2019-2026



国内重要的算力政策文件内容，2021-2023

时间	发文部门	文件名称	主要内容
2023.2	十三届全国人大常委会第三十七次会议	关于数字经济发展情况的报告	应统筹**通信和算力基础设施建设**，适度超前部署5G基站.推进“东数西算”工程。加快建设空天地海一体化网络。
2023.2	中共中央国务院	数字中国建设整体布局规划	系统优化算力基础设施布局，促进东西部算力高效互补协同联动，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。
2022.8	科技部等六部门	关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见	鼓励**算力平台**、共性技术平台、行业训练数据集、仿真训练平台等人工智能基础设施资源开放共享，为人工智能企业开展场景创新提供算力、算法资源。鼓励地方通过共享开放、服务购买，创新券等方式，降低人工智能企业基础设施使用成本，提升人工智能场景创新的算力支撑。
2022.8	科技部财政部	企业技术创新能力提升行动方案（2022-2023年）	推动**国家超算中心、智能计算中心**等面向企业提供低成本算力服务。
2022.1	国务院	关于印发“十四五”数字经济发展规划的通知	推进**云网协同和算网融合发展**。加快构建算力、算法、数据、应用资源协同的全国一体化大数据中心体系。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，面向政务服务、智慧城市、智能制造。自动驾驶、语言智能等重点新兴领域，提供体系化的人工智能服务。
2022.1	国家知识产权局	关于印发知识产权公共服务“十四五”规划的通知	加强国家知识产权大数据中心建设。依托全国一体化大数据中心体系，建设国家知识产权大数据中心，**强化算力统筹和智能调度**。
2021.8	工业和信息化部	新型数据中心发展三年行动计划（2021-2023年）	需求牵引，深化协同。坚持市场需求导向，建用并举。推动新型数据中心与网络协同建设，推进新型数据中心集群与边缘数据中心协同联动。促进算力资源协同利用，加强国际国内数据中心协同发展。
2021.5	国家发改委	《全国一体化大数据中心协同创新体系算力枢纽实施方案》	方案明确**国家算力枢纽建设方案，加快建设全国一体化算力枢纽体系**，提出布局全国算力网络国家枢纽节点，启动实施“东数西算”工程构建国家算力网络体系。进动微提中贪理局，供给平衡，绿色集约及互联互场

数据来源：IDC&浪潮信息、公开资料、甲子光年智库总结整理，2023.08

多地重视人工智能及算力产业发展，相继出台相关政策

甲子光年 JAZZYEAR

国内典型城市大力发展算力的相关政策汇总

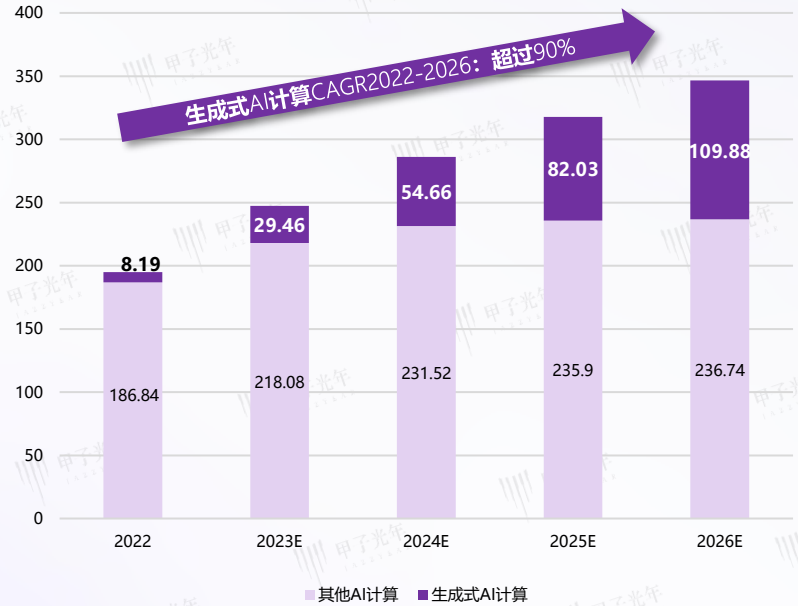
省市	时间	文件名称	主要内容
上海	2023.5	《上海市加大力度支持民间投资发展若干政策措施》	充分发挥人工智能创新发展专项等引导作用，支持民营企业广泛参与数据、算力等人工智能基础设施建设
北京	2023.5	《北京市促进通用人工智能创新发展的若干措施》	将新增算力建设项目纳入算力伙伴计划，加快推动海淀区、朝阳区建设北京人工智能公共算力中心、北京数字经济算力中心，形成规模化先进算力供给能力，支撑千亿级参数量的大型语言模型、大型视觉模型、多模态大模型、科学计算大模型、大规模精细神经网络模拟仿真模型、脑启发神经网络等研发。
北京	2023.2	《关于北京市推动先进制造业和现代服务业深度融合发展的实施意见》	支持园区加快计算中心、算力中心工业互联网、物联网等基础设施建设,建设园区大脑、数字孪生园区。
成都	2023.1	《成都市围绕超穿智算加快算力产业发展的政策措施》	建立以“算力券”为核心的算力中心运营统筹结算分担机制，结合区块链等新技术实现“算力券”有效监管。每年发放总额不超过1000万元的“算力券”，用于支持算力中介服务机构、科技型中小微企业和创客、科研机构、高校等使用国家超算成都中心、成都智算中心算力资源。
黑龙江	2022.10	《黑龙江省现代信息服务业振兴行动方案(2022-2026年)的通知》	完善新型基础设施布局,增强数字化转型支撑低能能力。不断增强骨干网承载能力,构建算力产业体系，建设区块链等新技术基础设施，助力构筑哈大齐协同一体科创走廊和工业走廊，促进网络基础设施广泛融入生产生活，有力支撑政务服务、公共服务、民生保障和社会治理。
河南	2022.9	《河南省元宇宙产业发展行动计划(2022-2025年)的通知》	构建多层次算力设施体系。统筹布局算力基础设施构建“超算+智算+边缘计算+存储“多元协同、数智胜合多层次算力体系。
上海	2022.7	《上海市数字经济发展“十四五”规划》	推动建设绿色数据中心,强化算力统筹和智能调度,提升数据中心跨网络、跨地域数据交互能力,推动数据中心供电、冷却、网络、服务器等智能协同,实现数据中心自动化能效调优,提升数据中心能效密度
河北	2021.11	《河北省建设全国级产业转型升级试验区“十四五“规划的通知》	建设全国一体化算力网络京津冀国家枢纽节点,加快向建工业互联网网络体系,改造升级省级北斗导航系统,规划建设低轨卫星互联网地面信关站。
天津	2021.3	《天津市新型基础设施建设三年行动方案(2021-2023年)的通知》	打造超算资源算力供给体系。依托国家超级计算天津中心,推动超算与人工智能深度融合。加快与量子计算、区块链技术融合发展,提供多层次智能算力服务打造各类创新平台协同创新算力载体。
北京	2021.2	《数字经济领域“两区"建设工作方案》	以支持数字经济发展的新基建为契机,推动形成5G网络、卫星网络、新型算力、新型数据中心、车联网等集聚、协同联动的数字经济基础建设体系

数据来源：公开资料、甲子光年智库总结整理，2023.08

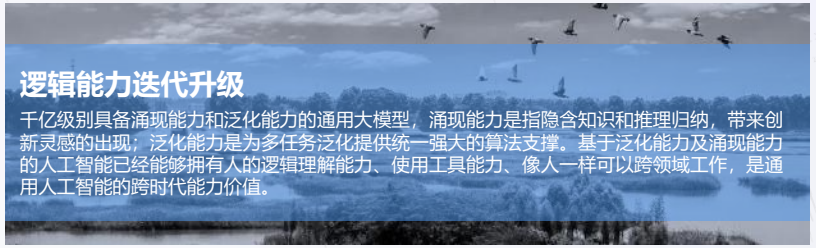
产业趋势叠加时代趋势，专用智能走向通用智能

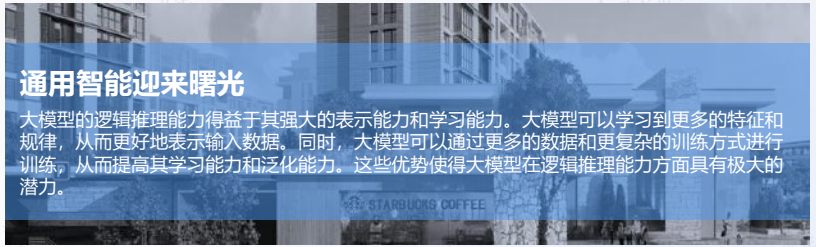
- 产业角度：大模型的产业发展仍处于起跑阶段，相关技术的行业应用和场景化落地存在无限可能。IDC预测，到2026年，全球AI计算市场规模将增长到346.6亿美元，生成式AI计算占比从2022年4.2%增长到31.7%。
- 时代趋势：随着LLM大规模语言模型技术的不断突破，以ChatGPT为代表的生成式人工智能引发广泛关注，人工智能正在从专用智能迈向以大模型为基座的通用智能，逐渐走向“开领域，走向通用”。

全球生成式AI计算的规模及预测（亿美元），2022-2026



发展通用人工智能，关注通用大模型的逻辑推理能力提升





数据来源：IDC&浪潮信息、公开资料、甲子光年智库总结整理，2023.08

目录

CONTENTS

甲子光年
JAZZYEAR

Part 01 产业基石，算力是AIGC产业的催化剂

Part 02 软硬兼得，AI新世代呼唤工程化导向的算力支撑

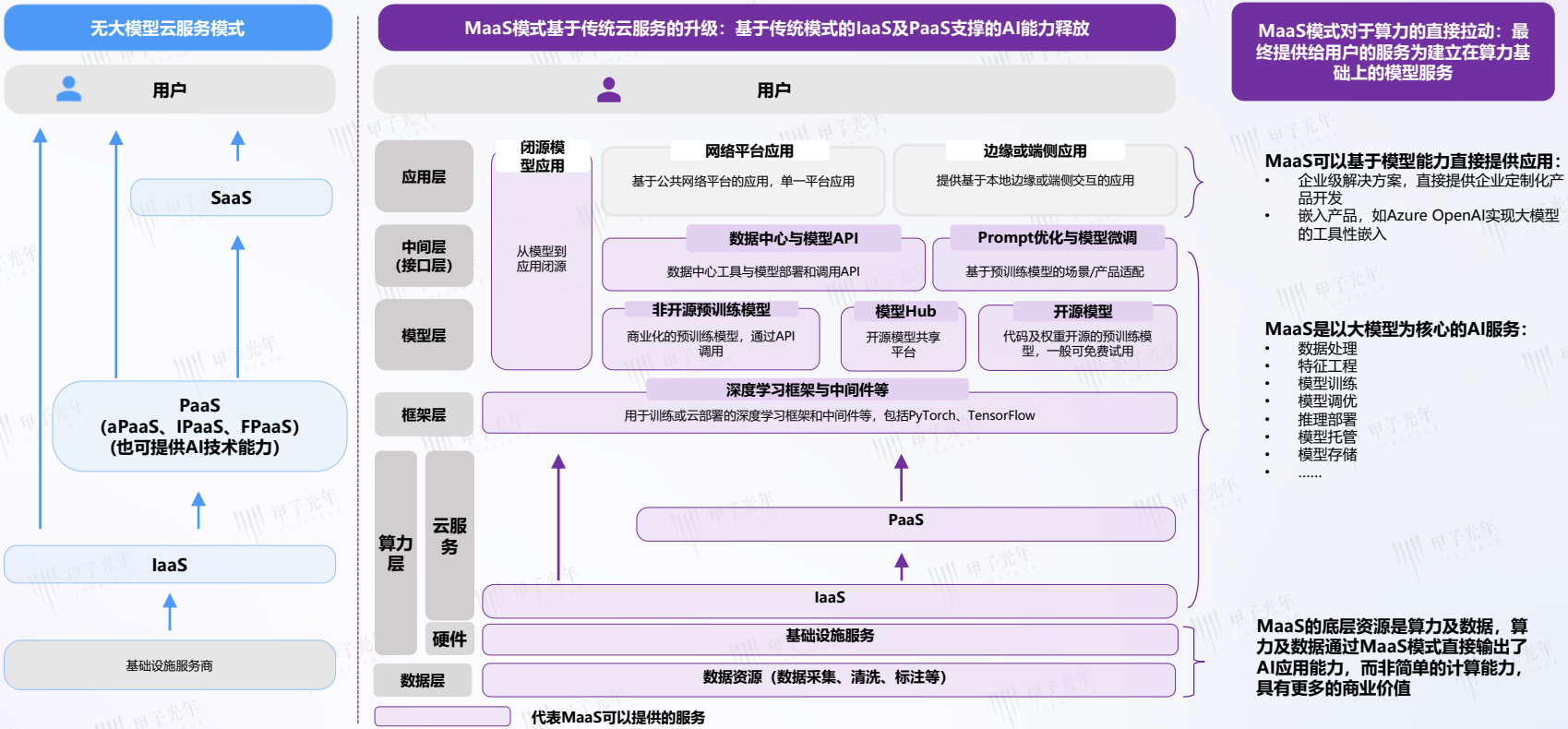
Part 03 层见叠出，商业浪潮下的算力选择思考

Part 04 实践真知，AIGC产业算力实践的新范式

Part 05 来日正长，AI技术的翻涌带来无限可能

MaaS是AI新时代云服务模式的破与立，构建新的“算力+算法”服务模式

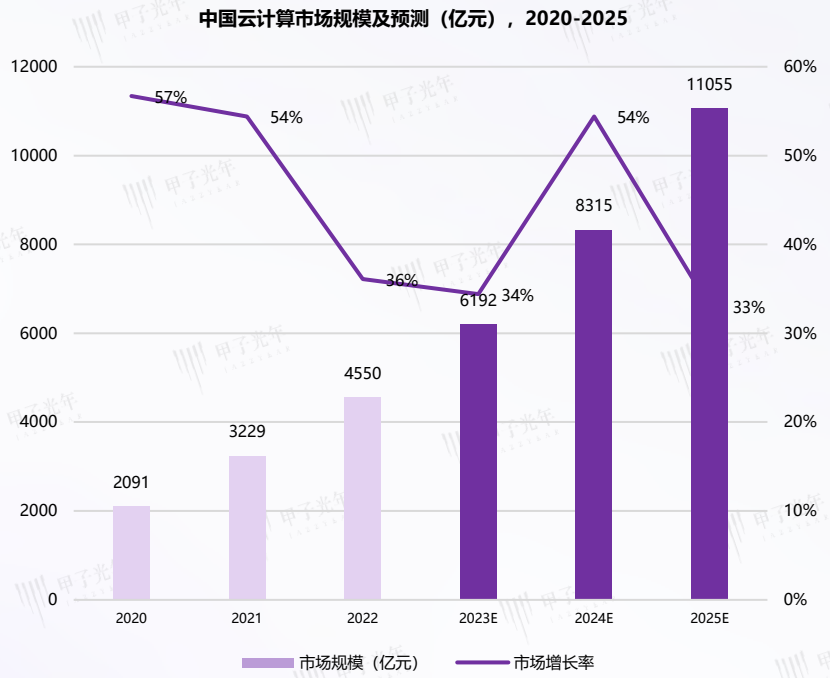
- MaaS（Model as a service）模型即服务，是指将大模型作为一项服务提供给用户使用的新业态，MaaS中模型训练（主要指微调）及推理的技术路线成立必须依赖云计算的算力支撑，同时算力及其他资源通过MaaS模型实现AI层面上更好的价值释放。

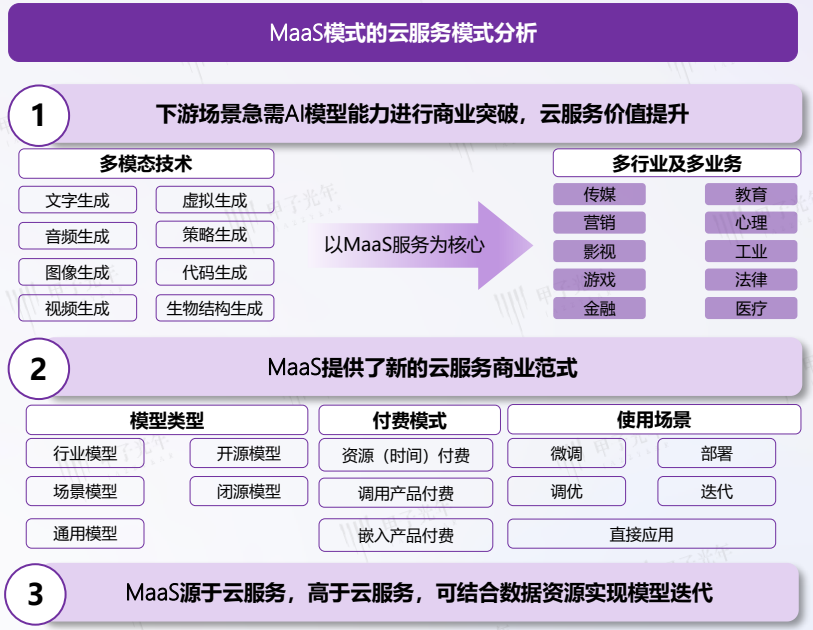


数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

基于AI时代的训练及推理需求，MaaS模式是当下云服务厂商新的机会点

- 2025年，根据中国信通院数据，我国云计算规模将超万亿云，其中重要的增长点是AIGC行业发展，MaaS服务契合当下AIGC产业发展，提供云服务商业应用价值，带动整体云计算增长。



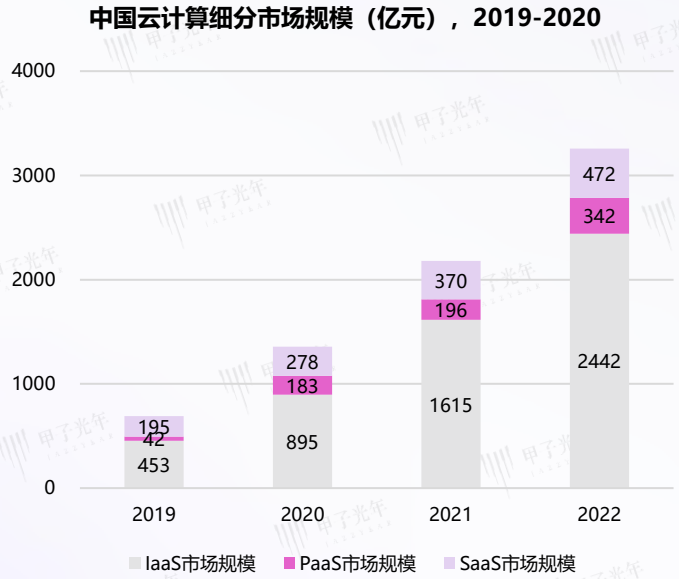


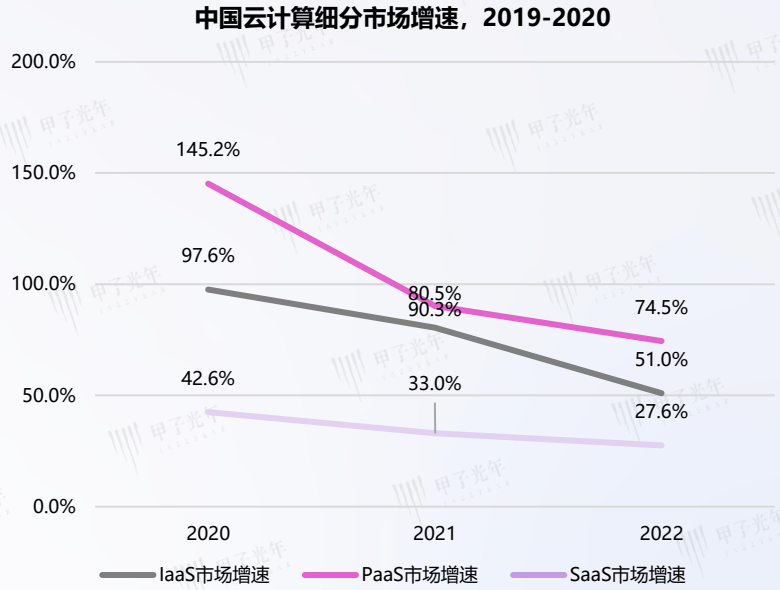
数据维度的治理、标注、数据库资源及模型维度配合用户的业务积累、数据回流等，可形成模型迭代，实现长期稳定服务及增量服务。

数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

MaaS服务未来将带动云服务市场的第二增长曲线

- 2022年，IaaS市场收入稳定，规模依然保持在2442亿元，并且是 PaaS+SaaS 的三倍，增速依然可以达到50%以上；引人注意的PaaS市场在容器、微服务等云原生应用带来的刺激增长，总收入已达到342亿元，并且增长率达到74%。
- PaaS模式在技术架构上，易结合AI技术应用，并且可以与MaaS服务作为增值服务提供，目前多家云服务厂商已经推出自研大模型、接入开源大模型，基于模型提供新的云服务增长点。根据微软财报数据，微软23Q3（相当于2023年第一季度）Azure Open AI服务客户数目达到2500+个，微软23Q4（相当于2023年第二季度）Azure Open AI服务客户数目继续增长至11000+个，环比增加340%，且本季度每天新增近100名新客户。



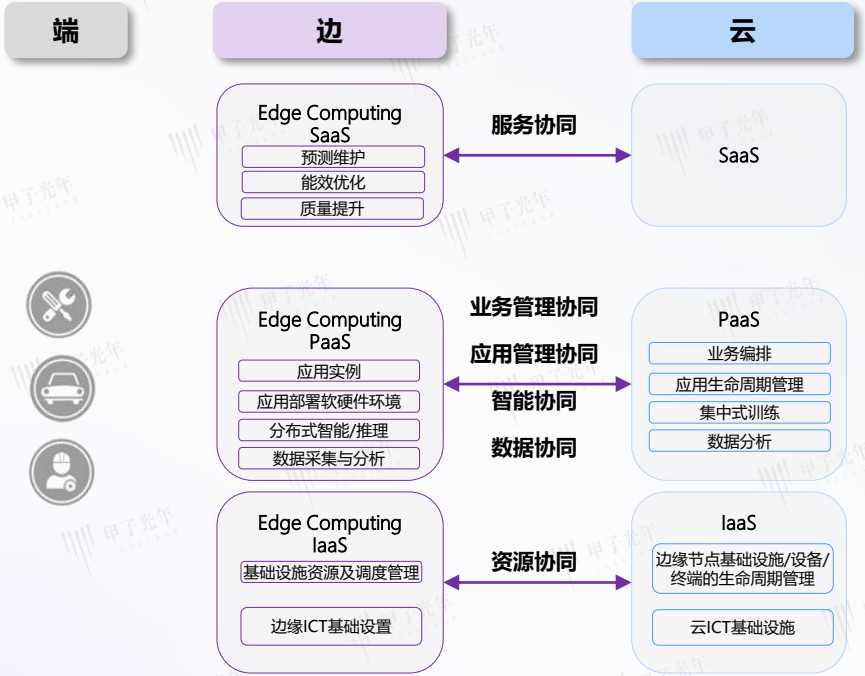


数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

云边协同，从计算、通信、安全、时效等方面提升AI应用落地可能性

- 边缘计算可负责范围内的数据计算和存储工作。同时，负责将连续数据汇总至计算层，最终由云计算层完成分析挖掘、数据共享工作，下发结果或模型至边缘和终端层，形成云-边-端协同。
- 边缘计算的核心价值：边缘侧完成数据的计算，并且实现云、端间的数据及计算结果的协同。边缘云及边缘芯片的发展将推动AIGC的更快落地。

云边系统实现资源协同，可最大程度实现的算力资源的合理分配



边缘计算的核心在于避免数据多次传输，从而打破完全中心的计算困境

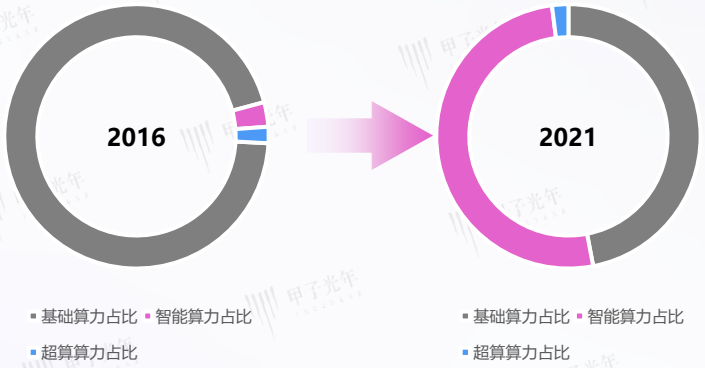
时效	在实体场景（例如物联网，工厂互联下），可以节省数据向中心端的传输时间，从而提高数据的实时处理能力
传输	面对物联网中长距离数据传输，边缘计算解决大量数据回传云端时网络带宽压力、性能瓶颈以及网络吞吐量
协同	解决云、端割裂，边缘计算为末端设备提供服务协同、资源协同、应用协同能力
安全	直接在边缘计算，实现在断网、断点情况保持稳定性、安全性
灵活	减轻终端、云端的数据压力，边缘计算仅对需要的、必要的数据提取计算，解决云端承受海量数据压力，实现终端的算力

数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

智能算力持续增长，未来需求增加，进一步加快智算中心建设及相关设备增长

- 据中国信通院，2016年，智能算力在我国算力中的占比仅为3%，而2021年，我国智能算力占比已经超过基础算力，达到51%，成为算力快速增长的驱动力。
- IDC预测2021年到2026年期间中国智能算力规模年复合增长率为52.3%，远高于同期预测的基础算力的增长率。

中国各类算力类型占比，2016VS2020



主要指标	基础算力中心	智算中心	超算中心
来源	一般为基于CPU芯片的服务器	一般为基于AI芯片的加速计算平台	超级计算机等高性能计算集群
建设目的	帮助用户降本增效或提升盈利水平	促进AI产业化、产业AI化、政府治理智能化	面向科研人员和科学计算场景提供支撑服务
技术标准	标准不一、重复建设CSP内部互联、跨CSP隔离安全水平不一致	统一标准、统筹规划、开放建设、互联互通互操作、高安全标准	采用并行架构，标准不一，存在多个技术路线，互联互通难度较大
具体功能	能以更低成本承载企业、政府等用户个性化、规模化业务应用需求	算力生产供应平台、数据开放共享平台、智能生态建设平台、产业创新聚集平台	以提升国家及地方自主科研创新能力为目的，重点支持各种大规模科学计算和工程计算任务
应用领域	面向众多应用场景应用领域和应用层级不断扩张，支撑构造不同类型的应用	面向AI典型应用场景，如知识图谱、自然语言处理、智能制造、自动驾驶、智慧农业、防洪减灾等	基础学科研究、工业制造、生命医疗，模拟仿真、气象环境、天文地理等

数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

智算中心的建设蓬勃发展，将成为AIGC算力的坚实基础

甲子光年 JAZZYEAR

- 2021年至2023年，国内各地实现多家智算中心的完工、揭牌、上线，支撑AIGC产业的研发及多行业应用。

序号	智算中心名称	运营状态	算力
1	北京昇腾人工智能计算中心	2023年2月13日上线	一期100P;短期500P;远期1000P
2	天津人工智能计算中心	2022年12月30日一期完工	300P
3	河北人工智能计算中心	2022年2月14日揭牌	计划100P
4	济南人工智能计算中心	已接入中国算力网	-
5	青岛人工智能计算中心	已接入中国算力网	100P
6	南京鲲鹏·昇腾人工智能计算中心	2021年7月6日上线	800P
7	南京智能计算中心	2021年7月16日投入运营	800P
8	太湖量子智算中心	2023年1月1日揭牌	-
9	腾讯长三角人工智能超算中心	在建	预计1400P
10	商汤人工智能计算中心	2022年1月24日投产	同时接入850万路视频; 单日处理时长23600年的视频
11	杭州人工智能计算中心	2022年5月20日	40P
12	淮海智算中心	在建	300P
13	中国-东盟人工智能计算中心	2022年9月23日揭牌	一期40P训练/1.4P推理
14	福建人工智能计算中心	2023年4月26日揭牌	一期规划105P;总体400P
15	深圳人工智能融合赋能中心	2019年打造人工智能融合赋能平台	-
16	广州人工智能公共算力中心	2022年9月15日上线运营	一期100P，五年内1000P
17	浙江“乌镇之光”超算中心	2021年9月25日正式启用	181.9P
18	宁波人工智能超算中心	2023年1月10日上线	一期100P(FP16)/5P(FP64) 二期300P(FP16)/15P(FP64)
19	昆山智算中心	2021年12月1日寒武纪中标	峰值500P(FP16)
20	阿里云张北超级智算中心	2022年8月30日上线	12000P

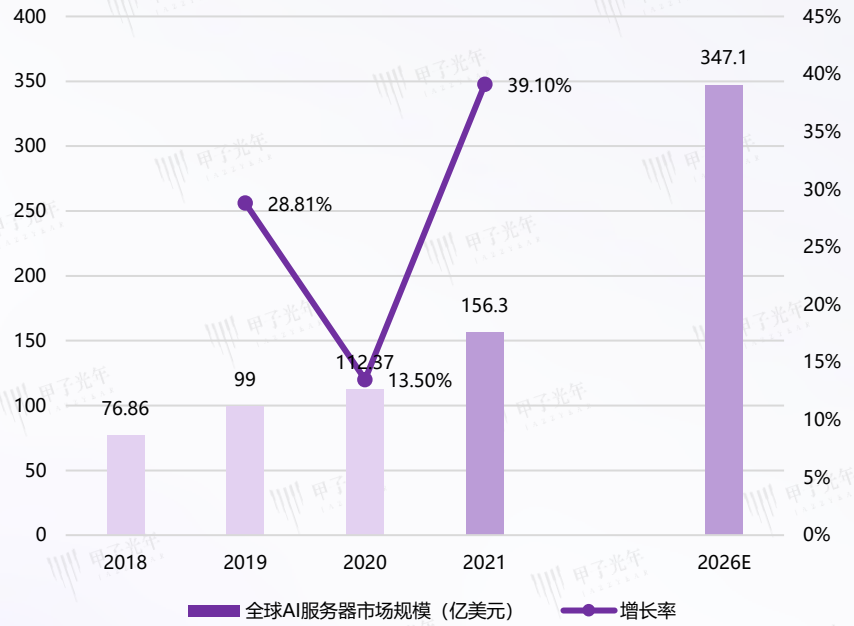
序号	智算中心名称	运营状态	算力
21	浙江省青田县元宇宙智算中心	2022年11月17日投产	100P
22	上海有孚临港云计算数据中心	-	-
23	中国电信京津冀大数据智算中心	2021年底投入运营	1-10P
24	北京数字经济算力中心(规划)	2022年4月落户	规划超过1000P
25	阿里云华东智算中心	2020年开工，2025年达产	-
26	上海市人工智能公共服务算力平台	2023年2月20日揭牌	-
27	山西先进计算中心	2018年10月运行	2.05
28	百度阳泉智算中心	2022年12月27日开机上线	计划100P
29	中原人工智能计算中心	2021年10月21日	计划100P
30	长沙人工智能计算中心	2022年11月4日	200P;2025年1000P
31	武汉人工智能计算中心	2021年5月31日	100P
32	横琴人工智能超算中心	2019年12月成立	1.16E(2019年);4E(完全建成)
33	合肥人工智能计算中心	在建	100P
34	成都人工智能计算中心	2022年5月10日	300P
35	未来人工智能计算中心	2021年9月9日	一期300P
36	重庆人工智能计算中心	在建	一期400P
37	甘肃庆阳智算中心	预计2023年8月建成使用	-
38	大连人工智能计算中心	在建	计划100P
39	哈尔滨人工智能先进计算中心	2020年底运营	-
40	沈阳人工智能计算中心	2022年8月9日上线	100P; 后期300P

数据来源：公开资料、甲子光年智库总结整理，数据截至2023年3月

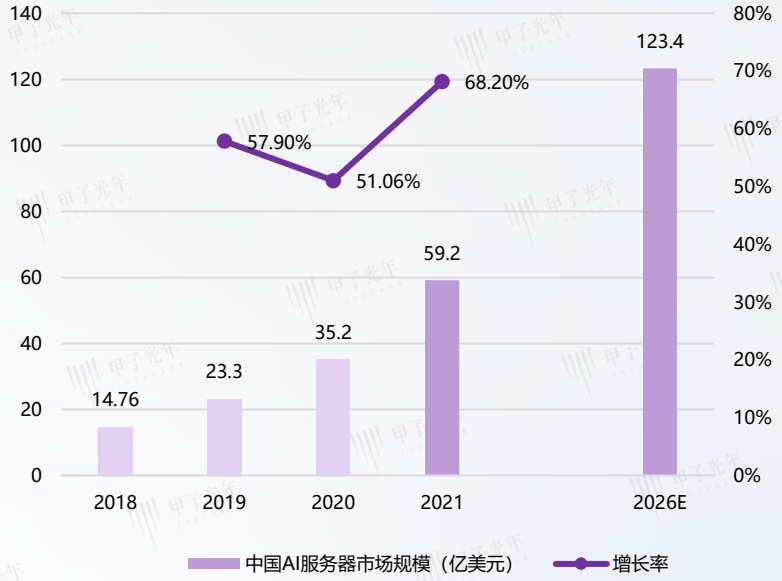
作为核心的智算硬件，AI服务器全球及中国的市场规模持续提升

- AI服务器一般是异构服务器，可以根据应用范围采用不同的组合方式，提供AI技术所需的算力。可以针对不同的需求进行硬件的选型及组合。
- AI服务器可用于智算中心、私有化部署、云服务等方面，AIGC的产业发展要求算力层各供应商产业升级，加快AI服务器的产品迭代及扩大规模。

全球AI服务器市场规模，2018-2026



中国AI服务器市场规模，2018-2026

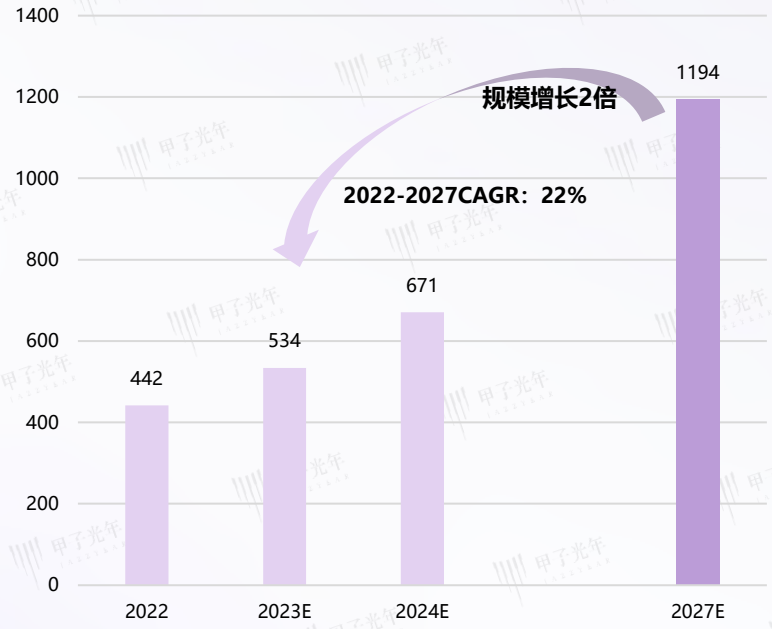


数据来源：公开资料、甲子光年智库总结整理，2023.08

AIGC的产业发展极大地推动了AI芯片市场的未来增长速度及产品丰富性

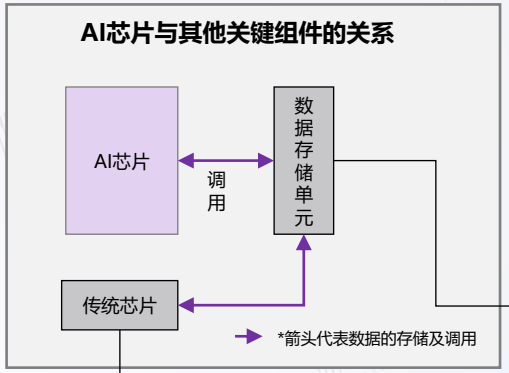
- 生成式AI的发展和各种基于AI的应用在数据中心、边缘基础设施和端点设备中的广泛使用，将推动AI芯片的生产和部署。到2027年，AI芯片规模预计将比2023年的市场规模增长一倍以上，达到1194亿美元。

全球AI芯片市场规模及预测（亿美元），2022-2027，Gartner



广义的AI芯片，可指运算AI算法的芯片，可包括深度学习，也可以包括其他机器算法，AIGC对AI芯片的推动不仅在于对单卡芯片算力的需求，同时也包括对多卡芯片的管理、AI芯片的架构升级、算法与芯片在设计层面的融合，通过算法调整解决解决芯片硬件瓶颈等

AI芯片与其他关键组件的关系



AI训练对内存需求较高

数据存储单元：进行数据实时的存储

传统芯片：主要运行非人工智能算法，
- CPU（Center processing unit，中央处理器）
- DSP（Digital Signal Processing，数字信号处理）
- ISP（Image Signal Processing Unit，图像信号处理）
- ……

数据来源：Gartner，公开资料、甲子光年智库总结整理，2023.08

大模型参数量呈爆炸式，短期内训练侧GPU集群化运算成为必须

甲子光年 JAZZYEAR

- 狭义的AI芯片指针对人工智能算法做了特殊加速设计的芯片，目前来讲，由于深度学习算法在人工智能领域认可度及应用程度不断上升，AI芯片一般指针对大量数据进行数据训练（training）与推断（inference）设计的芯片。

分类	典型特征
GPU	图形处理器（Graphics processing unit），在计算方面具有高效的并行性。用于**图像处理的GPU芯片因海量数据并行运算能力**，被最先引入深度学习。
FPGA	现场可编程门阵列（Field programmable gate array），是一种集成大量基本门电路及存储器的芯片，**最大特点为可编程**。具有能耗优势明显、低延时和高吞吐的特性。
ASIC	专用集成电路（Application specific integrated circuit，特定应用集成电路），**是专用定制芯片，为实现特定要求而定制的芯片**。除不能扩展应用以外，在功耗、可靠性、体积方面都有优势。
类脑芯片	“类脑芯片”是指参考人脑神经元结构和人脑感知认知方式来设计的芯片。**目前仍然处于探索阶段。**
量子计算（量子芯片）	基于量子力学的新型半导体芯片，它利用量子力学的特性来实现信息的存储、处理和传输，实现多量子比特的耦合，实现更高的计算能力和更复杂的逻辑运算。具有高迁移率，即量子芯片可以同时处理多个任务;同时具备强稳定性。制造和维护成本高，研发和生产成本也非常高。**目前仍然处于探索阶段。**
光子芯片	利用量子力学原理来制造的特殊芯片，它可以实现对光子信息的操作和处理，光子芯片具有高计算速度、低功耗、低时延等特点，且不易受到温度、电磁场和噪声变化的影响，光子器件很难做成芯片，后面需要很长的基础物理学研究，解决大量工程学问题。**目前仍然处于探索阶段。**

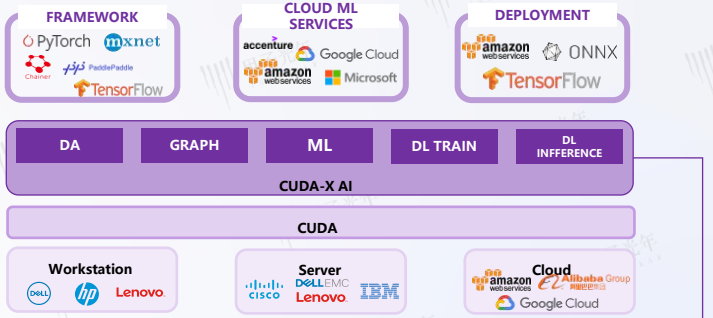
短期内，大模型训练端需要GPU多卡集群完成，时间压力下用户倾向选择成熟方案

长期看，FPGA、ASIC依然在推理侧具有一席之地，但需要明确场景需求

英伟达通过对硬件单元的改进与显存升级增强了单张GPU算力的释放，及形成护城河的网络通信、AI生态工具，并且全球内具有成功案例，短期内是大模型训练最适用的AI芯片。

随着 Transformer 模型的大规模发展和应用，模型参数量呈爆炸式增长，单卡无法完成相应训练，大模型参数量的指数级增长带来的诸多问题使GPU集群化运算成为必须。

英伟达的CUDA生态在AIGC领域已经逐步养成客户习惯，短期内具有竞争力



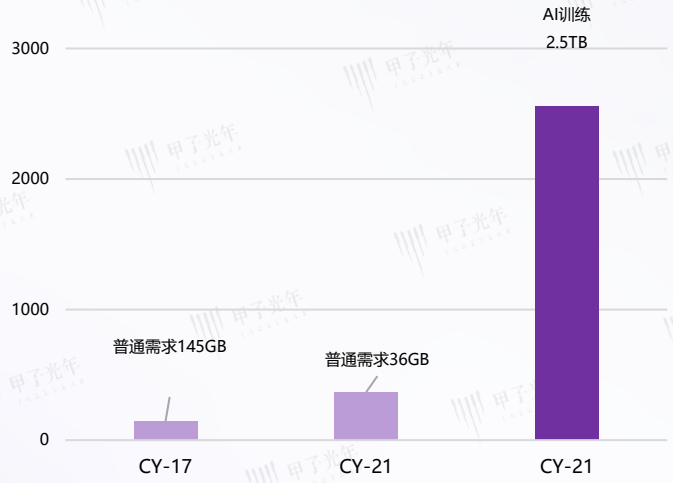
- CUDA-X AI 是软件加速库的集合，建立在CUDA之上，提供对于深度学习、机器学习和高性能计算的优化功能。
- 库与NVIDIA Tensor Core GPU 配合工作，能够将机器学习和数据科学工作负载加速至高达50倍。
- CUDA-X AI的软件加速库集成到所有深度学习框架和常用的数据科学软件中，且可以部署到多种设备内的NVIDIA GPU上

数据来源：公开资料、甲子光年智库总结整理，2023.08

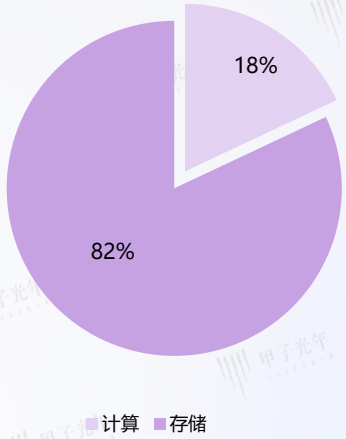
AIGC时代，内存对算力的整体利用率影响提升，需要新的内存解决思路

- Transformer模型中的参数数量呈现出2年240倍的超指数增长，而单个GPU内存仅以每2年2倍的速度扩大。而训练AI模型的内存需求，通常是参数数量的几倍，AI训练不可避免地撞上了“内存上限”，“内存上限”不仅是指内存容量，也包括内存传输带宽。
- 同时通信成为算力的瓶颈。无论是芯片内部、芯片间，还是AI加速器之间的通信，都已成为AI训练的瓶颈。过去20年间，运算设备的算力提高了9万倍，虽然存储器从DDR发展到GDDR6x，接口标准从PCIe1.0a升级到NVLink3.0，但是通讯带宽的增长只有30倍。长期看，无法实现堆积显存解决问题。

以AI训练服务器为例，AI训练需要更高的内存容量



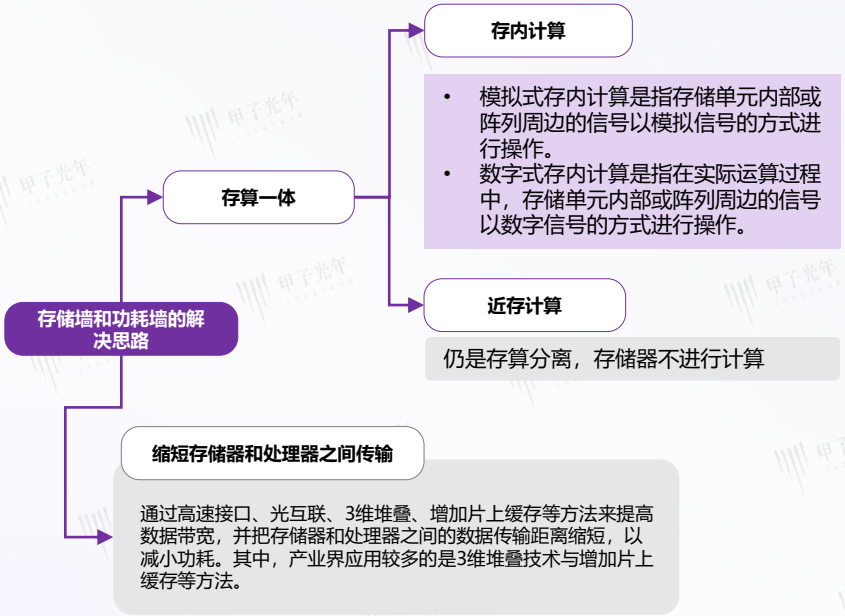
Cadence数据显示，在自然语言类AI负载中，存储消耗的能量占比达到82%



数据来源：公开资料、甲子光年智库总结整理，2023.08

存算一体随存储器介质的多样性逐步走向应用成熟，解决AGI时代的存储墙问题

- 目前，主流芯片如CPU、GPU以及DPU均按照冯·诺依曼架构设计，由于冯·诺依曼架构的局限性，数据的处理遇到了存储墙和功耗墙两大问题。
- 存储器的访问速度远远小于处理器的运算速度，系统整体会受到传输带宽的限制，导致处理器的实际算力远低于理论算力，难以满足大数据应用的快、准响应需求，数据在存储器与处理器之间的频繁迁移带来巨大的传输功耗。
- 存算一体有Flash、SRAM、DRAM等成熟存储介质，同时ReRAM、MRAM等新型存储介质也在快速发展，ReRAM存内计算技术未来具有非常大的应用潜力，尤其是实现大算力的方面，虽然目前工艺成熟度相对不足，但有待突破。



基于不同存储器介质的存算一体芯片之间的性能

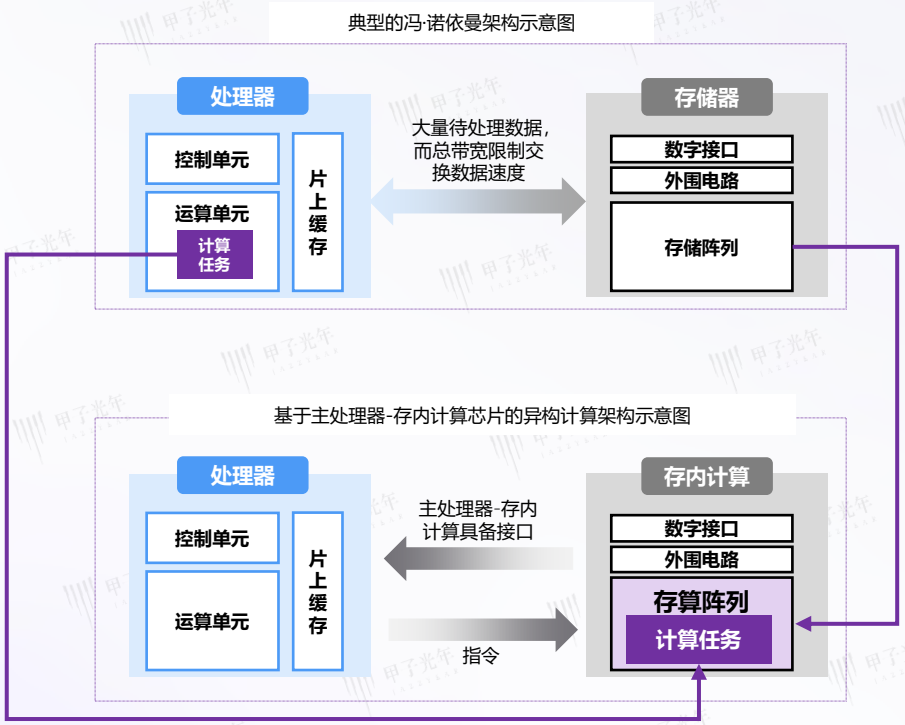
标准	SRAM	DRAM	Flash	ReRAM	PCM	FeFET	MRAM
非易失性	否	否	是	是	是	是	是
多比特存储能力	否	否	是	是	是	是	否
面积效率	低	一般	高	高	高	高	高
功耗效率	低	低	高	高	高	高	高
工艺微缩性	好	好	较差	好	较好	好	好
成本	高	较高	低	低	较低	低	低

数据来源：公开资料、甲子光年智库总结整理，2023.08

存内计算技术革命性解决存储墙问题，高效适配Transformer结构算法

常用的向量矩阵乘法在深度学习计算中，如果转化到存算一体中，只需要1次存储器的读取操作，就可以完成百万级参数的乘法和加法计算。如果用传统的GPU架构，百万级的乘法加法计算，光是存储器的读取次数就要超过5万次。……我认为未来几年都是存算一体飞速发展的黄金时代，这就像以前90年代摩尔定律一样，每年都有几倍的算力提升。存算一体在未来3-5年内可能提升速度更快，每年可能都超过8倍的算力提升。

——知存科技创始人兼CEO 王绍迪



在冯·诺依曼架构中，数据从存储单元外的存储器获取，处理完毕后再写回存储器，计算核心与存储器之间有限的总带宽直接限制了交换数据的速度，**带来存储墙和功耗墙问题。由于带宽导致的实际算力往往低于处理器的理论算力。**

解决处理器与内存间数据传输即可直接提升实际算力

存内计算不仅可打破数据搬运产生的“存储墙”，并且适用于AI算法

存内计算，作为一种新型计算架构，直接利用存储器本身进行数据处理，从根本上消除数据搬运，实现存储与计算融合一体化，可突破冯·诺依曼架构带来的**带宽墙和功耗墙问题。**

此外，存内计算及存内逻辑非常适合人工智能/深度学习的算法计算，人工智能/深度学习的算法中有大量的矩阵乘法计算，其本质是乘累加（Multiply-Accumulate，MAC）运算，存算架构可以将计算直接映射到存储结构中，具有高能效比和低延迟性。生成式AI中大量算法所依赖的Transformer结构同样可以适合存内计算完成。

例如，Mythic已于2021年推出基于NOR Flash 的存内计算量产芯片M1076，可支持80MB神经网络权重，单个芯片算力达到25TOPS。

数据来源：《存内计算芯片研究进展及应用》，专家访谈，甲子光年智库整理，2023.08

目录

CONTENTS

甲子光年
JAZZYEAR

Part 01 产业基石，算力是AIGC产业的催化剂

Part 02 软硬兼得，AI新世代呼唤工程化导向的算力支撑

Part 03 层见叠出，商业浪潮下的算力选择思考

Part 04 实践真知，AIGC产业算力实践的新范式

Part 05 来日正长，AI技术的翻涌带来无限可能

AIGC产业算力理念：需要基于目标与资源的分配去达成工程学平衡

甲子光年 JAZZYEAR

- AIGC产业落地的算力选择，更应该强调最优解，而非最大解。在实现AIGC的技术落地过程中，模型的参数量及涌现结果固然重要，但模型在运行过程中所需的算力成本、能耗成本、运营成本等是否能匹配AIGC技术提供的效果及价值突破更为重要。

以终为始，贴合行业需求，实现目标与资源平衡，是AI新世代下的算力选择依据

资源分配：通过选择合适的技术路径实现算力的成本优化

核心目标：基于行业Know-How需要实现的AIGC技术功能拆分，实现精准的需求分析

功能需求决定推理能力，推理能力取决训练水平，有限算力资源要进行主次的优先选择

训练需求（一次开发）

考虑到模型训练“黑盒”机制与多次调优，所需算力与开发过程强相关

- 技术实现路径
- 算法结构
- 训练数据量
- 参数量规模
- 预训练
- 基于需求进行fine-tune

推理需求（长期运营）

模型推理阶段的算力主要为运行模型和数据处理，并且需要考虑产品的使用体验

- 模型规模（参数稀疏程度）
- 模型种类（算力需求系数相关
- 数据吞吐量
- 时延
- 网络通信
- 安全性与稳定性

其他成本制约因素

- 时间成本（是否尽快抢到实践化的落地）
- 能耗成本（云服务或者算力的使用成本）
- 人员成本（工程化协作的团队）

行业Know-How不仅仅表现在丰富的行业实践经历，而是深入理解客户的业务需求，并且通过管理项目开发的流程完成，在细化需求中寻找到主要矛盾并解决。

Why（用户分析）

- 基于用户的细分行业属性，熟悉细分行业的需求价值
- 基于用户的业务流程细节，对用户的需求矛盾分析
- 基于用户的资源能力，明确用户的负担上限
- ……

✖

How（项目执行）

- 在不同阶段和层面对项目的工作内容从主项、分项、子项甚至单体的各个部分进行拆分（例如采用WBS），实现项目关键节点的管理，
- 完成项目人员的协同、管理、分工及时间资源调配
- 对风险的预知、判断及合理控制
- ……

数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

AIGC的算力资源选择，需要结合自身部署能力及应用需求综合考量

- 算力资源的维度不仅包括算力规模大小，要考虑算力部署及运营过程中可以利用的程度。算力是工程化结果，是从芯片到资源服务的多层次构造，需要算力服务方自身在自身专业能力及经验案例上的实际Know-How作为基础。
- 不同需求程度的用户不能唯算力的参数而论，而是要结合自身对于算力部署的能力进行进一步探究。

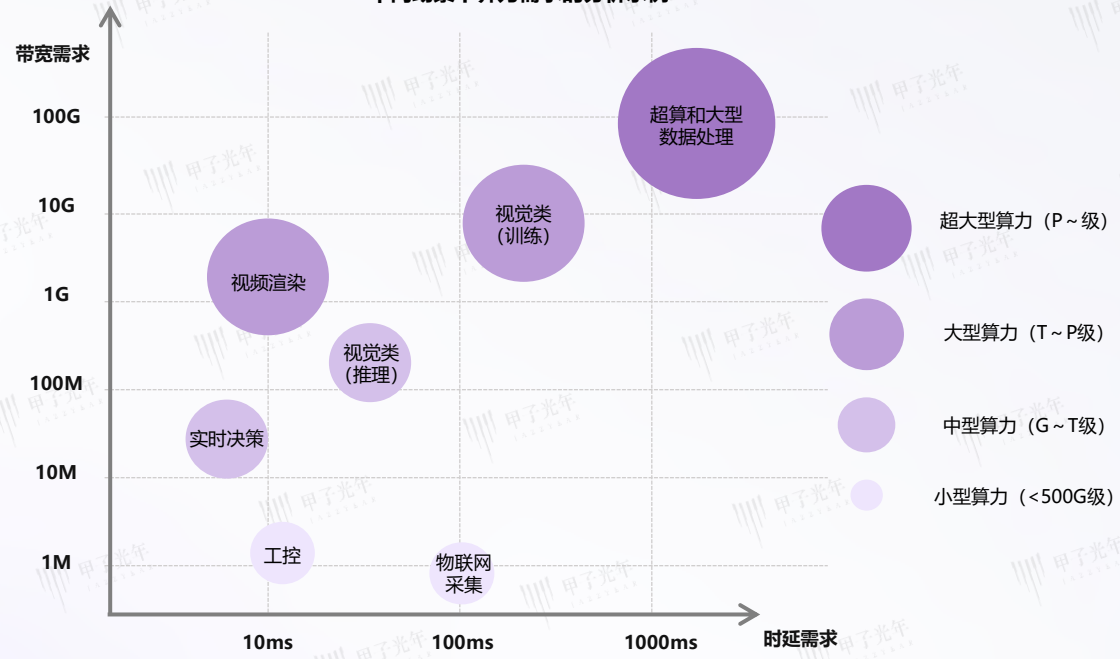
	影响算力资源利用的维度 （算力提供方在AI算力领域的Know-How及经验）		算力直接使用者所需技术要求
云服务	• 核心计算单元的算力参数 • 对应的运算精度 • 单元数量	• 芯片的选择及适配 • 智算硬件的选择及适配 • 智算中心的选择及适配 • 接入方式、算力调度、需求分配、弹性扩展、高效稳定、算法优化、通讯传输、第三方生态、故障排查、大模型相关数据及训练工具包（生态）、模型的纳管及生态合作、云边端协同	★★ 可按需适配资源及弹性适配，部署时间更快，可以选择适配AIGC产品/服务的算力资源，减少对于AI算力环境优化的时间及人力成本
智算中心		• 芯片的选择及适配 • 智算硬件的选择及适配 • 租户管理、配额管理、运维管理、资源及作业调度管理、系统监控、安全及稳定	★★★ 按需取用、灵活扩展、无需各IT系统的复杂运维，直接在完成优化的环境下进行开发
智算硬件		• 芯片的选择及适配 • 硬件选型及适配（如内存）、异构算力的调度及配合、网络传输、软件优化、集群架构、环境优化	★★★★ 通过服务器等硬件完成自有算力的部署，环境调试，完成大量不同硬件设备的选型、优化及稳定性保障，需要具备成熟的项目案例经验
芯片		• 内存/显存、片内互联及片间互联、AI适配生态工具（包括适配算法及其他硬件）、物理环境支撑、折旧速率	★★★★★ 需要从芯片层面解决工程问题，包括芯片互联、构建网络、适配从应用到硬件的环境，工作量大且繁杂，需要具备从0到1的经验丰富的技术团队支持

数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

算力作为逻辑资源，场景的复杂性导致算力的评价指标不唯一

- 算力作为逻辑资源，与水电等标准化资源相比就更加复杂、具备更多维度，而技术的发展催生了丰富的计算场景，不同的行业、应用场景对算力更提出了不同的需求。

不同场景下算力需求的分析示例



以上数据为估算示意，具体情况下的数据需具体分析

根据需求考量的算力维度示例

算力精度：双精度算力、单精度、半精度、 整型计算的选择
如天体物理、气象研究、航空航天等高精尖科研领域需要能够支持复杂运算、性能高的双精度算力；对于AI模型训练及推理来说，处理文字、 语音、图片或视频等需求较大，单精度、半精度、甚至整型的计算即可满足应用需要。

时延：需要实时渲染的游戏、自动驾驶决策、远程手术、工业控制等领域对延迟的要求非常高，而模型训练等场景则对延迟没有很高要求。

带宽：基于AR、VR等渲染场景，模型训练、超算类等场景对大带宽的需求较高，工控、物联网采集等则对带宽要求不高。

数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

当下的AIGC算力关注热点在训练端，但商业突破及应用需要推理侧支持

- 整体市场的算力核心判断指标取决于市场发展阶段对应的训练及推理需求，当下是AIGC产业技术与商业结合的重要拐点，一方面，国内外均有商业落地的场景及对应模型出现，技术路线实现大方向确认；另一方面，具体场景的商业模式及盈利模式仍待寻找，需要大量算力支持各行各业企业持续探索。
- 市场算力的核心指标变化：重训练阶段——算力支持模型迭代，重推理阶段——关注应用的用户数量。

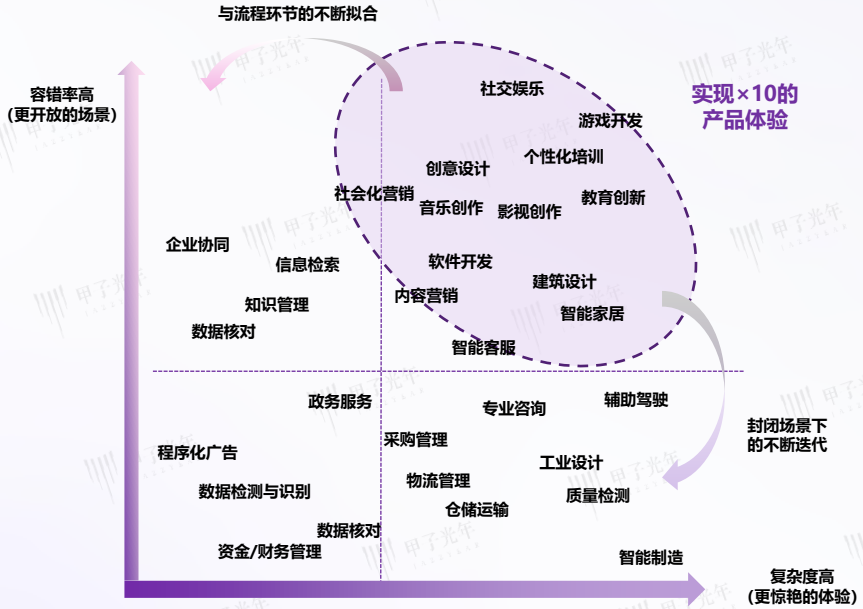
- *总体技术路线已确认：国内外大量通用模型出现，并且有部分标杆性案例，可在部分场景下寻找稳定商业模式*
- *细分赛道的技术路线待定：当下的模型能力如何赋能细分赛道依然值得探索*

数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

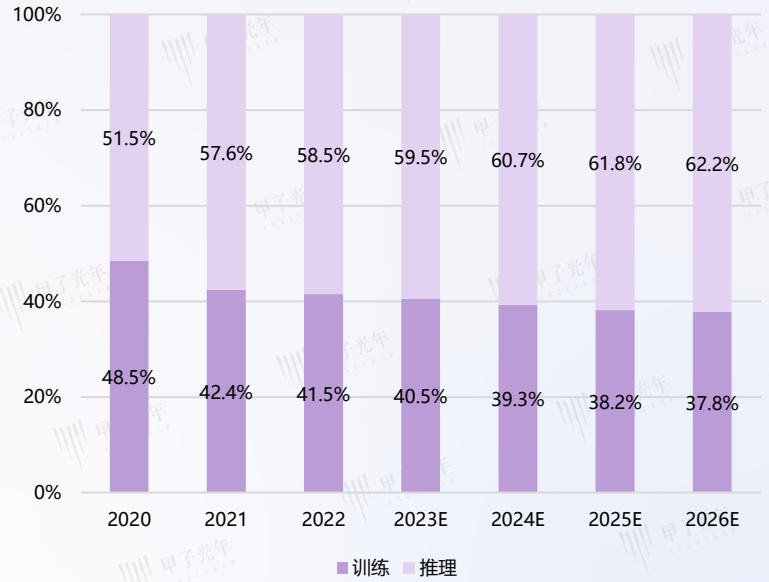
多模态*多行业，AIGC的C端垂类应用体验将处于快速探索阶段

- 考虑到当下AIGC的可信性及成本的局限性，AIGC的应用在复杂度高（智能化水平），及更开放的场景实现，推动AI落地侧应用以及模型的迭代。
- 根据IDC数据，2022年中国数据中心用于推理的服务器的市场份额占比已经过半，达到58.5%，预计到2026年，用于推理的工作负载将达到62.2%。未来随着AIGC产业发展，训练侧及推理侧均具有发展潜力。

AIGC应用场景的分析示例，产品体验将催生大量C端垂类产品



中国人工智能服务器工作负载占比预测，2020-2026



数据来源：IDC，公开资料、专家访谈、甲子光年智库总结整理，2023.08

AIGC给予了企业流程新的管理思路，催动算力支撑AI技术融入数字化管理

- 既定业务流程的建模、标准化、自动化、执行、控制、度量和优化（例如BPM）是基于现有数据交互协同的思路管理员工。
- AIGC通过解决部分交互及协同标准，可以更好地实现流程上的协同管理，但仍需企业对场景的不断探索后才能明确需求。

数据在企业的人、财、物、资中无处不在，数字化的重要作用是通过数字世界的信息流完成企业数字化管理

人“找”数字化流程 ***AIGC*** ***数字化流程“找”人***

企业服务领域的AIGC应用，与企业的数字化流程融合

企业现有算力需要升级及迭代

要求企业数字化管理的算力需要支撑AIGC的应用，满足模型在推理端的使用及探索

基于自有流程进行专属大模型的训练/微调

企业的专属数据可训练适配自身的模型，而企业自身数据具有私密性，部分行业严禁外流（例如金融），因此需要建设相关算力支撑

算力需保证企业流程的稳定性及安全性

企业服务与个人开发的区别在于，模型的训练及使用，以及对于流程的改造，绝不能中断企业现有数字化节奏，需要更好的算力稳定保证活迁移方案

目录

CONTENTS

甲子光年
JAZZYEAR

Part 01 产业基石，算力是AIGC产业的催化剂

Part 02 软硬兼得，AI新世代呼唤工程化导向的算力支撑

Part 03 层见叠出，商业浪潮下的算力选择思考

Part 04 实践真知，AIGC产业算力实践的新范式

Part 05 来日正长，AI技术的翻涌带来无限可能

云服务、大模型一体机、智算中心、服务器及计算芯片服务商为当下AIGC算力核心提供企业

*随着AIGC在多场景、多领域的不断应用，将不断推动更多企业进入AIGC产业算力服务领域，图谱1.0版数据截至2023年8月，顺序不分先后

重点厂商产品及服务能力分析——浪潮信息

打造大模型智算软件栈OGAI，保障大模型开发和应用的算力基础设施需求

➢ **公司简介：**浪潮信息（股票代码SZ000977），全球领先的IT基础架构产品、方案及服务提供商，业务覆盖计算、存储、网络三大关键领域，提供云计算、大数据、人工智能、边缘计算等在内的全方位数字化解决方案。公司秉持“计算力就是生产力”，致力于通过计算技术的不断创新推动社会文明的持续进步。

浪潮信息

大模型智算软件栈OGAI

OGAI（Open GenAI Infra）“元脑生智”，为大模型提供算力系统环境部署、算力调度及开发管理能力的全栈全流程智算软件栈
OGAI由浪潮信息基于大模型自身实践与服务客户的专业经验而开发，旨在为大模型与应用打造高效生产力，加速生成式AI产业创新步伐。

大模型智算软件栈OGAI整体架构图

L4	多模纳管	MModel	提供多模型接入、服务、评测等功能的纳管平台，加速模型的部署和应用
L3	模型工具	YLink	面向大模型数据治理、预训练、微调的高效工具链，降低大模型开发和落地的门槛
L2	调度平台	AIStation	面向大模型开发的商业化人工智能算力调度平台。针对大模型训练中常见的训练中断难题，可实现训练异常快速定位，断点自动续训
L1	系统环境	PODsys	开源、高效、兼容、易用的智算集群系统环境部署方案，实现自动化部署和弹性扩展，提高系统可用性和扩展性
L0	基础设施	智算中心OS	面向大模型算力服务的智能算力运管平台，满足多租户以裸金属为主的弹性AI算力运管需求。

智算硬件

高性能AI服务器 | **高吞吐并行存储系统** | **低延迟RDMA网络**

OGAI在不同的层次强调了性能优化，从服务器BIOS的调教到大规模集群组网性能和算力调度策略的多尺度、多层次的性能优化，来充分释放智算集群性能

致力于使大模型的安全管理、便捷使用与能力测评变得更加高效和易用，核心组件包括：数据集管理、模型纳管和评测

聚焦于大模型的数据治理、预训练、微调等开发过程，集成浪潮信息在大模型研发中的自研工具和开源工具

聚焦于AI开发场景,通过云原生技术对集群系统中的计算资源、存储资源和网络资源进行统一的接入和纳管

聚焦智算集群部署场景，从OS、驱动到系统监控可视化、资源调度等集群系统环境要素全面覆盖

高效的裸金属服务支持分钟级部署上千规模裸金属节点并按需进行弹性扩展，实现异构计算芯片、IB、RoCE高速网络、高性能存储等环境一键获取

OGAI融合了浪潮信息在MLPerf性能评测、服务客户实际需求、源大模型开发中的最佳实践

数据来源：浪潮信息，甲子光年智库整理

千亿参数规模的源大模型创新实践，助力AIGC落地千行百业

甲子光年 JAZZYEAR

浪潮信息 源 YUAN

➢ **项目概况：**作为最早布局千亿参数规模大模型的企业之一，浪潮信息在业界率先推出了中文AI巨量模型“源1.0”，参数规模高达**2457亿。**“源1.0”在语言智能方面表现优异，获得中文语言理解评测基准CLUE榜单的**零样本学习（zero-shot）和小样本学习（few-shot）两类总榜冠军。**

“源1.0”，千亿巨量模型的工程实践

大模型训练全栈能力

- **训练数据，**通过自研海量数据过滤系统（MDFS），建立从数据采集、粗滤、质量分类、精滤的全自动化的端到端数据工作流程，通过清洗866TB海量数据，获得5TB高质量中文数据集。
- **算法创新，“源1.0”针对大模型的**Attention**层和前馈层的模型空间进行结构优化，**改进注意力机制聚焦文章内部联系的学习，面向中文的语言理解和生成能力业界领先。
- **计算优化，首次提出面向效率和精度优化的大模型结构协同设计方法，**围绕深度学习框架、训练集群IO、通信开展了深入优化，在仅采用2x200G互联的情况下，**源1.0的算力效率达到45%，**算力效率世界领先。

大模型智算软件栈OGAI+源大模型 = AIGC产业基石

应用于智算中心，
让智算力可以像水和电一样便捷地提供给千行百业

- 浪潮信息作为智算新基建的倡导者和推动者，联合伙伴相继推动南京智算中心、淮海智算中心、青田元宇宙智算中心、“钱塘江”液冷智算中心解决方案等项目与方案的落地。
- 将OGAI智算软件栈、“源1.0”大模型的系统工程经验应用于智算中心，**对集群架构、高速互联、算力调度等进行全面优化，对大模型三大并行训练策略优化，**精准调整模型结构和训练过程的超参数，**最终实现千亿参数规模的大模型训练算力效率达至53.5%。**

应用于企业，助力大模型高效率研发

助力网易伏羲中文预训练大模型“玉言”登顶中文语言理解权威测评基准CLUE分类任务榜单，并在多项任务上超过人类水平。

- “玉言”大模型参数**达到110亿，**结构由深层Encoder和浅层Decoder组成。这种结构使得大模型具有优秀的理解能力和生成能力，同时方便训练任务的设计，不需要复杂的掩码策略，上线即登顶CLUE分类任务榜单。
- “玉言”具有良好的泛化性，在各类任务上都有着出色的性能。目前，大模型相关技术和成果已应用在网易集团内的文字游戏、智能NPC、文本辅助创作、音乐辅助创作、美术设计、互联网搜索推荐等业务场景，取得了显著的业务效果。

强大稳健的基础大模型，让大模型即服务（MaaS）垂手可得，助力AIGC落地

智能客服

通过将“源”大模型的智能力与复杂的服务场景进行深度融合，打造专家级数据中心智能客服大脑，凭借强大的学习能力，“源晓服”能够对知识库进行自主化学习，可覆盖终端用户**92%的咨询问题，**将复杂技术咨询的业务**处理时长降低65%，整体服务效率达160%，获评哈佛商业评论鼎革奖。**

智慧政务

在智慧政务领域，基于源大模型打造的AI社区助理“临小助”，可为基层社区工作者提供沉浸式、针对性的一对一培训，社区工作者通过手机终端与“临小助”进行互动对话，可快速提高自身服务群众的能力。目前，“临小助”已投入使用，在某高频场景中，社区工作人员培训**学习效率提升5倍,有效辅助问答建议达到75%。**

智慧文创

在智慧文创领域，基于“源1.0”，开发者开发出首个AI剧本杀；开发并上线会“闹情绪”的AI陪练，助力心理咨询师在模拟演练中提高自身技能。除此之外，**“AI数字人鲁迅”**、数字演员、陪伴机器人、游戏NPC对话等极具创新的应用也在不断地孵化落地。

数据来源：浪潮信息，甲子光年智库整理

重点厂商产品及服务能力分析——云天励飞

具备“芯片+算法+大数据”全栈式能力，自主设计开发的新一代边缘计算芯片

➢ **公司简介：**云天励飞（股票代码688343）成立于2014年，是拥有算法、芯片和大数据全栈式能力的人工智能企业，可提供融合全栈能力的系统解决方案，业务领域覆盖智慧警务、城市治理、智慧交通、智慧园区等多个领域。具备自研“造芯”能力，并且独特地选择了算法与芯片融合发展的路径，打造了“算法芯片化”等核心能力，让算法和芯片两大技术相互配合，高效赋能各类边缘计算、大模型系统解决方案和应用场景。

intellifusion 云天励飞

自主设计开发的新一代边缘计算芯片DeepEdge10，采用D2D/C2C高速互联技术，满足边缘计算场景的算力多样化的要求

纯国产AI芯片，完全自主可控

国产设计；国产工艺；国产基板；国产封测

D2D Chiplet，国产工艺首创

满足国产Chiplet UCIE标准；创新D2D高速互联chiplet技术；快速实现通用、专用算力扩展

DeepEdge10芯片产品平台

可满足不同算力、不同性价比的边缘计算场景的需求，包括边缘CV大模型并且可以解决大模型应用和部署过程中的各类挑战

Edge10的产品优势：

- AP级通用SoC（CPU/GPU/MM/双显）
- 异构多核，大算力
- 接口丰富、灵活、扩展强

Edge10C

8核CPU
8T算力

Edge10标准版

10核CPU
12T算力

Edge10Max

40核CPU
最大**64T算力**

边缘侧产品的布局

边缘网关： X200 miniPCIe卡（8路/6Tops）；X2000工控机加速卡（40路/25Tops）

单芯片AI盒子： Edge10盒子（单芯片，**16路/12Tops**）；Edge10max盒子（单芯片，**64路/50Tops**）

大模型的推理加速卡布局

X5000加速卡：64T算力，支撑CV大模型落地
X6000加速卡：256T算力，支撑100亿大模型落地
X6000一体机：2048T算力，支撑1000亿大模型落地

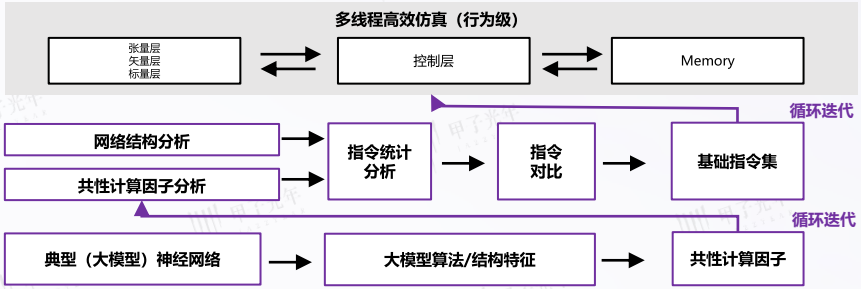
数据来源：云天励飞，甲子光年智库整理

“算法芯片化”结合“云边协同”，满足新AI时代的计算需求

- **“算法芯片化”**：云天励飞凭借对行业的战略前瞻和持续多年的深耕布局，让算法和芯片两大技术相互配合，**基于对场景的深刻理解以及对算法关键计算任务在应用场景中的量化分析，将芯片设计者的理念、思想与算法相融合的AI芯片设计流程**，为解决方案和应用场景更高效地赋能。
- 基于该理念，DeepEdge10边缘神经网络推理芯片在技术架构上采用架构统一的神经网络处理器芯片平台和创新的指令集架构，**满足大模型基于Transformer结构所带来的新的神经网络计算范式以及高带宽传输、分布式并行计算、低精度混合计算需求。**

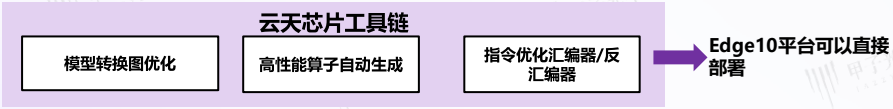
DeepEdge10边缘神经网络处理器芯片的典型技术优势

DeepEdge10边缘神经网络处理器芯片**根据新的神经网络计算范式的转变，设计更高效的、更灵活的处理器指令集**，以达到大模型的新型计算、访问模式以及混合精度计算等方面的要求



DeepEdge10边缘神经网络推理芯片采用D2D高速互联的chiplet芯片技术路线，通过多颗Edge10 Die片上MESH互联设计，可**实现单芯片算力的灵活可扩，满足不同场景应用对AI芯片算力的差异化需求**

DeepEdge10边缘神经网络推理芯片采用AP+NPU+GPU+CV多核异构计算架构技术路线，**通过统一的工具链技术平台，实现对人工智能通用大模型在边缘高效推理部署对不同计算资源的需求。**



“云边协同”技术路线，赋能AI应用场景加速落地

可应用场景丰富

云天AI芯片赋能硬件ODM厂商，打造边缘计算网关产品，为世界500强云服务提供商**提供云边协同的AI推理算力，云端训练+边缘推理部署，广泛用于数字城市的安防和泛安防领域的各类视频物联边缘计算场景**，如智能社区、智慧园区、智能楼宇、智慧工厂的视频分析应用

场景核心痛点

- **规模化视觉端云协同，云端推理需要大量的数据传输：**
  - ✓ 网络带宽占用大
  - ✓ 传输响应延时大
  - ✓ 影响业务闭环的响应速度

产品及技术优势

- **边缘计算网关产品通过解决数据采集、延迟、带宽、安全、可扩展性等方面的痛点问题，为企业提供了更高效、安全和可靠的数字化转型解决方案**
- **减轻了云端负载：**
  - ✓ 降低了网络带宽的投入
  - ✓ 缓解了网络拥堵问题
  - ✓ 在部分无网环境实现AI业务落地
- **高效的AI业务赋能：**
  - ✓ 低成本的本地化部署
  - ✓ 灵活的算法调度
  - ✓ 实时的业务闭环响应
  - ✓ 断网不断业务

重点厂商产品及服务能力分析——UCloud优刻得

提供中立安全云计算，打造稳定可靠的大模型算力底座

➢ **公司简介：**UCloud优刻得（优刻得科技股份有限公司）成立于2012年，2020年正式登陆科创板，成为中国第一家云计算科创板上市公司。始终坚持关键核心技术攻关，推出公有云、私有云、混合云、边缘云等全线云产品；自主研发IaaS、PaaS、安全屋大数据流通平台、AI多种终端及服务平台。UCloud在全球建设和运营32个可用区，遍及国内、东南亚、欧洲、北美、南美、非洲等25个地域，构建了云网融合、安全稳定、智能敏捷、绿色低碳的数字信息基础设施。

以中立身份，通过“算力+平台+模型”的模式，提供开放、安全、定制的AIGC解决方案

应用服务	开源模型	合作模型	视觉绘图	模型微调	专家服务	安全方案
管理平台	资源纳管	租户管理	运维管理	实时监控	需求调研 方案设计 工程建设 交付实施 网络变更 设备测试 故障处理 迭代优化 7*24 NOC跟踪 服务报告	DDOS防护 WAF 堡垒机 数据库审计 日志审计 SSL VPN
网络服务	RoCE/IB交换机	BGP|单线带宽	专线|SD-WAN	云联网		
资源平台	混合云AI资源池	云主机AI资源池	裸金属AI资源池	存储资源池		
数据中心	机柜	围笼改造	定制化巡检	RFID资产管理		
	综合布线服务	7*24运维保障	机房搬迁服务	DCIM系统		

超10年云计算技术积累，拥有全面系统工程能力

模型可快速部署，高效应用

接入优质开源模型，建立国内模型生态合作，具备应用侧快速部署能力

通过内部孵化项目实践和优化模型微调、模型推理等流程，**实现高效能的资源利用**

强大技术底座及产品架构经验

管理：可使用UK8S进行任务调度，**从而实现完备的训推一体平台**

网络：联动RoCE、IB的高性能网络，**满足大规模算力集群对高速网络的性能需求**

资源：GPU云主机、裸金属等构建算力单元，US3、UFS构建存储池

物理层自主管理保证安全可靠

通过为用户提供了物理完全隔离的独享机柜、网络、服务器、存储资源，**结合完整的安全方案和专家服务，确保了用户业务平稳运行**

中立： **专注底层算力基础设施，不涉足大模型业务，不触碰用户隐私数据**

开放： **与模型厂商建立稳定生态合作，通过私有化大模型一体机的模式，帮助客户推广大模型**

安全： **纯内资背景，并且具备系列产品及服务保障数据安全**

定制： **提供公有云服务，也支持私有化部署，根据客户需求调优网络和存储带宽**

数据来源：UCloud优刻得，甲子光年智库整理

大模型算力集群解决方案，助力大模型服务商化解算力“燃眉之急”

甲子光年 JAZZYEAR

UCLOUD优刻得 ✖ 大模型服务商

➢ **背景信息：**随着大模型训练及应用的需求上升，大规模分布式算力集群的需求呈现井喷式涌现，相关专业性的产品及服务呈现极大缺口。UCloud优刻得针对国内某知名大模型公司的需求，在市场热潮初期与该公司进行了产品打磨，**满足其训练和推理算力需求的同时，以完整的基础架构方案对接上层的资源调度系统，推出了大模型训练集群+推理集群+存储+管理的完整云服务解决方案。**

项目难点

网络方案需迅速找到工程化解法

训练集群GPU服务器之间需要高速网络互联，成熟的RoCE网络方案虽然具有成本优势，但需要从工程化角度针对大模型场景进行专门的适配和优化

存储系统面对训练过程的高压

大模型训练过程依赖于存储系统，对存储系统的吞吐和使用方式都有新的要求

降低算力成本，保证资源最优

大模型集群建设成本高是算力紧缺的原因之一，因此针对资源的利用率优化，可以最大程度保证客户的建设成本

执行亮点

基于UCloud优刻得的工程化经验，针对客户的大模型训练需求进行快速调优，形成专属方案

- **自建大模型训练集群RoCE网络：**为满足算力集群对高速网络的性能需求，基于UCloud在公有云大规模使用RoCE网络的经验，自建大模型训练集群RoCE网络，并带来成本优势，供应链更加灵活
- **针对大模型提供冷热分层的存储方案：**基于分布式文件系统和US3对象存储集群，针对大模型训练集群提供冷热分层的存储方案，使用便捷，吞吐量高，提高训练效率
- **基于成熟UK8S产品优化资源调度：**资源调度上即可基于成熟的UK8S产品方案一键部署，也可为客户自行建设的调度集群提供完整适配解决方案
- **高效的故障处理能力：**完善的监控体系，快速定位硬件、网络故障

自有乌兰察布和青浦云计算中心，为大模型提供智算底座

- **深度定制：**专门为GPU集群建设的高电力机柜以及专门优化的走线设施，满足H800/A800等各种GPU相应的电力/网络需求
- **高性价比：**乌兰察布数据中心相较于北京、上海同等质量数据中心，价格降低40%
- **高速互联：**乌兰察布数据中心覆盖京津冀区域，可实现“训推一体”，效率提升；青浦数据中心覆盖长三角区域，更加适用于对延时敏感的推理任务

实践效果

联合打磨产品和方案，快速完成客户需求，助力客户实现在市场初期的机会把握

完成**千卡A800 GPU服务器+RoCE网络+存储系统的算力集群建设**，验证UCloud提供大模型完整解决方案的能力

完成网络、硬件、内核、K8S层的适配，对接客户自建**K8S**调度系统，为客户的大模型训练提供高性价比的训练算力

提供全栈大模型算力基础设施方案、技术支持以及运维服务，**客户专注于模型迭代，满足客户对时间和效率的极致要求**

数据来源：UCloud优刻得，甲子光年智库整理

重点厂商产品及服务能力分析——知存科技

专注存算一体芯片及技术，全球率先实现商业化量产存内计算芯片

- **公司简介：**知存科技是全球领先的存内计算芯片企业。凭借颠覆性的技术创新，知存科技突破传统计算架构局限，利用存储与计算的物理融合大幅减少数据搬运，在相同工艺条件下**将AI计算效率提升2个数量级**，充分满足快速发展的神经网络模型指数级增长的算力需求。公司针对AI应用场景，**在全球率先商业化量产基于存内计算技术的神经网络芯片。**
- 2023年1月完成B2轮融资，获得多家科技领军企业和顶级财务投资机构的持续支持，并且被正式纳入国家级专精特新小巨人培育企业名单。

Witmem 知存科技

长期专注存内计算技术研发，实现存内计算芯片从0到1的工艺突破，具备成功流片多款芯片，量产2款芯片的能力，产品矩阵覆盖端、边、云

企业优势：深耕技术，关注产业链生态建设

团队深耕存算一体技术：知存科技创始团队自2012年开始存算一体芯片开发，2016年研发出全球第一个支持多层神经网络的存内计算芯片，首次验证了存内计算在深度学习应用中的优势。

致力于推动存算一体产业化链条建设：率先搭建并不断强韧存内计算芯片产业化链条。不仅与国内外主流芯片生产厂家建立了长期稳定的合作关系，还共研实现了存内计算芯片从0到1的工艺突破，成功流片多款芯片，量产两款产品。

不断完善技术生态合作：与北京大学、清华大学、中科院、南京大学等高校联合承担了多个国家级和省部级科研项目，2023年与北京大学深圳研究生院成立“存算一体联合实验室”。

关注客户需求及技术结合：知存科技与智能终端集成与应用领军企业、智能语音应用领军企业等产业客户建立了密切的协同开发与合作关系。

产品路线：逐步提高算力，扩大场景

云

WTM-C系列云侧AI芯片
基于大容量存储的原位计算，实现数据通信需求10倍降低，大幅度提高计算效率。

边

WTM-8系列边缘侧AI芯片
新一代移动设备计算图像芯片，64M模型参数，**>24Tops算力，**12-bit运算精度，4核存算MPU，支持linux，支持AI超分、插帧、HDR、识别和检测。

*知存科技自主研发的**边缘侧算力芯片WTM-8系列即将量产**。该系列芯片能够提供至少**24Tops算力**，而功耗仅为市场同类方案的5%，将助力移动设备实现更高性能的图像处理和空间计算*

端

WTM-2系列端侧AI芯片
针对端侧AI计算市场推出的存内计算芯片，采用40nm制程，拥有高算力存内计算核，相对于NPU、DSP、MCU计算平台，**AI算力提高10-200倍。**

*2021年推出第一代芯片WTM1001，2022年则实现较**第一代芯片提升接近10倍算力**的WTM2101量产。该芯片已被多家国际知名企业用于智能语音、AI健康监测等场景，相比传统芯片**在算力和功耗上优势显著**，赋能行业用户实现端侧AI能力的提升和应用的推广*

数据来源：知存科技，甲子光年智库整理

存内计算芯片创新性融入智能穿戴产品功能革新，实现多场景商业落地

甲子光年 JAZZYEAR

Wltmem 知存科技

➢ **背景信息：可穿戴设备将会以更多元化的方式更深入到人们的生活，**根据市场调研机构IDC数据预测，2023年可穿戴设备的出货量将同比增长4.6%，达到5.39亿部，全球可穿戴设备市场将以5.1%的五年复合年增长率增长，到2026年底出货量将达到6.284亿部。

市场痛点

智能穿戴产品的革命需要算力升级

智能穿戴易出现产品形态、功能趋同等问题，为占据市场份额，企业需要从细节革新、功能深入，进一步提升产品市场竞争力，例如

- TWS耳机开始向多功能化、智能化方向发展
- 智能手表提供更加“精准”的健康管理功能
- VR/AR作为虚拟世界的入口，向成为下一代移动硬件终端演进

部分替代方案难以实现严苛的低功耗、低时延和低成本，需要新的算力方案提供

产品亮点

算力升级实现产品的功能智能化进阶

- WTM2101基于知存科技存内计算平台，可使用sub-mW级功耗完成大规模深度学习运算，提供智能语音、智能健康解决方案
- 可实现：AI降噪（环境降噪+极致人声保留）实时健康监测（心率/血压/血氧/身体状态）、离线语音识别、AI啸叫抑制、AI人声增强、脑电+肌电监测（情绪识别/指令识别/手势识别）等

WTM2101芯片产品特点

1.基于存算一体技术，实现NN VAD和上百条语音命令词识别

2.超低功耗实现NN环境降噪算法、健康监测与分析算法

3.典型应用场景下，工作功耗均在微瓦级别

4.采用极小封装尺寸

商业价值

魅蓝Blus K耳机

解决了混音、低延迟返听和K歌音效三大难题

作为魅蓝秋季发布的专业K歌耳机，Blus K搭载WTM2101芯片实现了耳返功能，通过神经网络AI算法真正**做到人声与伴奏完美融合、KTV混响和回声效果**

INMO Air2 AR智能眼镜

低功耗下实现更轻量化、更高性能的用户体验迭代

INMO Air2通过WTM2101在**低于1mA的功耗下实现了免唤醒命令词功能**，可以更快速打开应用/操控设备，反应更及时，实现更智能的语音控制、更轻松的操控体验

智能手表

低功耗下突破算力限制，实现实时健康监测

过往产品血压、运动心率、疾病风险预测等AI算法必须运行在移动端（需要装APP）或者云端，无法实时监控数据。通过WTM2101健康监测方案，客户实现了PPG实时健康监测（心率/血氧等），**运行功耗极低，测试数据更精准**

助听器

较传统产品实现革命性突破，并且产品成功通过Fonix、盐雾等医规测试，即将批量生产

- **防啸叫：完成助听领域首个NN啸叫抑制应用**，能够做到**40dB增益下无啸叫，**而过往助听器的增益超过30dB SPL就可能引起回声啸叫。
- 人声增强：同时实现**8ms NN人声增强**，而其他市场方案通常在数百毫秒到数秒之间。
- **功耗显著降低：**运行功耗低至1.8mA，而其他市场方案在数mA到数十mA之间。

数据来源：知存科技，甲子光年智库整理

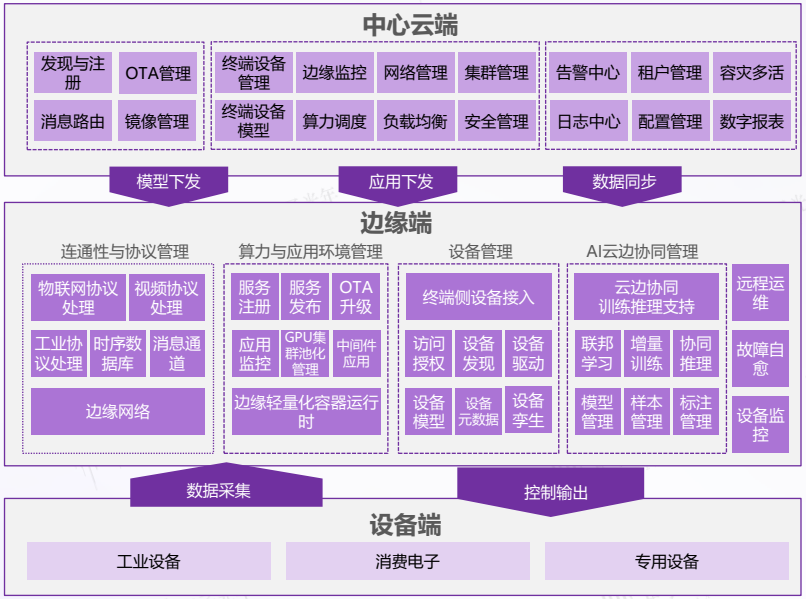
重点厂商产品及服务能力分析——博云

自主研发的一体化边缘计算平台，充分释放物联网场景下的AI算力

➢ **公司简介：**博云成立于2012年，坐落于苏州工业园区，是一家集科研创新于一体的国家高新技术企业公司。专注为企业级客户提供自主可控，以容器和云资源调度技术为基础的云操作系统相关的基础软件、解决方案与技术服务，包括容器云产品族、云资源管理系统、开发支撑软件DevOps三大系列产品与服务。公司为企业的数字化转型提供有力支撑，包含传统应用、新型云原生应用、边缘计算应用和高性能计算应用等广泛异构应用的运行承载与调度，同时通过开发支撑软件DevOps大幅提升应用开发效率。通过多云资源管理、应用承载调度和开发支撑三位一体的软件工具助力企业IT架构革新和业务敏捷创新。

博云

博云边缘计算平台架构



边缘计算技术优势

- **实时数据处理：**可以将物联网产生的大量数据的处理和分析推向设备的边缘，实现实时的数据处理和决策，可减少数据传输的延迟和带宽需求，提高系统的响应速度和效率。
- **数据过滤和筛选：**在设备的边缘对数据进行过滤和筛选，只将有用的数据传输到云端进行进一步处理和分析，可减少数据传输的量，降低网络带宽的需求。
- **边缘智能和决策：**将一些简单的数据处理和决策逻辑放在设备的边缘，减少对云端的依赖，可以提高系统的可靠性和稳定性，同时减少对云端计算资源的需求。
- **数据安全和隐私保护：**将一些敏感的数据处理和存储在设备的边缘，减少对云端的数据传输。这样可以提高数据的安全性和隐私保护，降低数据泄露的风险。

边缘计算解决传统物联网计算难题

- **解决延迟问题，提高数据处理实时性：**实现近距离的数据处理，从而大大降低延迟，可应用于实时性要求较高的应用场景。
- **减少网络依赖，解决带宽问题：**仅需重要价值的数据传输到云端，从而减少数据传输量，降低对网络带宽的需求。
- **注重敏感数据，解决数据安全隐患：**将敏感的数据处理和存储在设备的边缘，提高数据的安全性和隐私保护，避免数据泄露和隐私问题。
- **避免网络不可靠：**减少对云端的依赖，提高系统的可靠性和稳定性，即便网络出现故障或断连，也可保证系统部分正常工作。
- **利用边缘智能，提升数据分析效率：**减少对云端计算资源的需求，提高数据分析的效率。

数据来源：，甲子光年智库整理

云边协同解决方案，实现大规模核心的调度、运维及管理

甲子光年 JAZZYEAR

博云

➢ **项目简介：**某技术研究所仿真平台项目需要完成千核规模级别的调度满足算力需求，并且需要在windows环境下完成相应的算法迭代、繁琐运维，博云采用自身研发的云边协同解决方案完成客户需求。

项目难点

性能难以满足

原有高性能软件调度极限仅300核，无法满足现有业务所需的成千上万核调度能力

无法敏捷迭代

由于每天都会调整算法，原有模式部署需要一周完成，急需流水线提高部署效率

隔离差运维繁

原有平台经常遇到资源占用高、作业缺乏隔离、删除作业需后台手动操作等问题

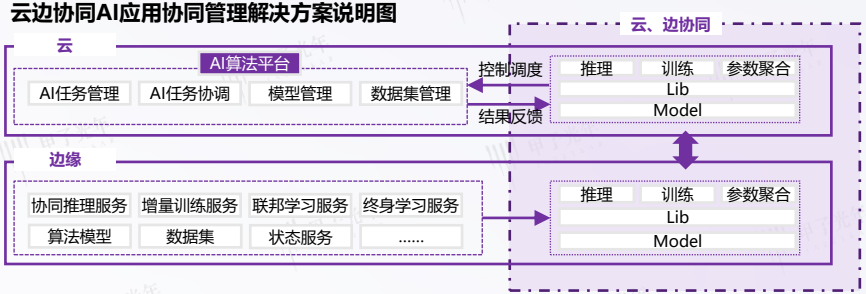
大量windows应用

客户领域内的高性能应用大量采用windows计算环境

执行亮点

项目初期完成对3套业务的7种作业类型支持，覆盖200+计算节点（每台cpu72C、men1TB），超240块GPU卡，验证核心数规模在5000+

云边协同AI应用协同管理解决方案说明图



边缘集群算力调度与应用运行支撑说明图

实践效果

- **自动调度：**实现自动化的任务排序和调度管理
- **简化运维：**自动对异常中止的任务和中间过程信息进行清理
- **简单管理：**实现了任务可视化状态管理，降低新用户使用成本
- **数据安全：**数据加密存储在后端，保障了任务结果的数据安全
- **信息可视：**实现对用户进行资源和任务执行过程限定和分析
- **系统兼容：**实现了兼容Linux和Windows计算节点

数据来源：博云，甲子光年智库整理

重点厂商产品及服务能力分析——亿铸科技

技术创新+学术实力，致力于打造新一代AI大算力芯片

➢ **公司简介：**亿铸科技成立于2020年6月，是全球首家基于存算一体这一创新架构，面向数据中心、云计算、中心侧服务器、自动驾驶及边缘计算等场景的**AI大算力芯片公司。**亿铸科技首次将新型存储器ReRAM及存算一体计算架构相结合，通过全数字化的芯片设计思路，致力于实现数倍性价比、更高能效比、更大算力发展空间的新一代AI大算力芯片。**初代产品基于传统工艺制程，可实现1000T（1P）以上的单卡算力。**

亿铸科技具备“四新一强”的优势，并且基于长远市场考虑及技术战略定力，专注国产存算一体大算力研发

“新”：技术全方面创新

存算一体架构创新

消除存储墙	减少能耗墙	降低编译墙

ReRAM新型忆阻器应用创新

“ReRAM是业内普遍认为最适合做存算一体大算力的存储介质，未来具有无限潜力”

非易失性	读写速度快	稳定性强
功耗低	CMOS工艺兼容	密度极大
高低阻值差异大	成本优势	微缩化发展
工艺成熟，可量产出货		

全数字化技术路径应用创新

“在满足大算力的同时支持高精度，使得存算一体架构真正在AI大算力方向落地”

高精度	大算力	超高能效比
将存算一体架构在大算力真正落地		

存算一体超异构系统级创新

*“AI大算力芯片的系统级创新概念，从整体设计角度结合存算一体和异构计算优势，**为大模型时代AI大算力芯片换道发展提供了全新思路**”*

有效算力更大	放置参数更多	能效比更高
软件兼容性好	发展天花板更高	

“强”：团队阵容实力强劲

创始人具有深刻的行业战略及产品理解

创始人，董事长及CEO熊大鹏博士在中美有近30年的芯片行业经验，涉及从研发、产品定义、产品销售，到企业的整体管理层面的多方面经历，对中国市场的客户需求与产品有着深刻的理解。

团队具备丰富的科研级工程落地能力

其核心研发团队成员均为来自国内芯片大厂的资深专家，毕业于斯坦福大学、德克萨斯大学奥斯汀分校、哈佛大学、上海交通大学、复旦大学和中国科学技术大学等。研发能力覆盖从ReRAM器件、全数字存算一体计算架构、AI芯片设计、编译器、算子库、应用开发平台等全链条。

- **10+** 位世界知名院校博士及以上学位的专家教授
- **20+** 颗SoC芯片的设计、量产及销售经验
- **25+** 年高端集成电路设计和量产经验（工程团队成员平均）
- **40+** 篇顶会论文发表（研究团队合计）

数据来源：亿铸科技，甲子光年智库整理

具备云计算、边缘推理计算颠覆性的国产新型芯片的解决方案

亿铸科技

国内少数基于传统工艺制程落地AI大算力芯片的项目，可满足各类中心侧、边缘侧大算力、低能耗等需求。**原型验证芯片(POC)已成功回片点亮。**

产品在场景应用当中的优势

大算力

利用ReRAM新型忆阻器作为介质，充分发挥存算一体架构的优势，完全可以满足数据中心对于单位面积的算力产出、满足边缘侧（如自动驾驶）等升级换代所带来的算力提升需求

高算力密度

由于解决了存储墙的问题，无需数据搬运，同等算力下，能耗更小，面积更小，适配更多边缘侧场景的物理要求（温度、体积等），并且可以充分利用成熟的工艺制程完成先进制程芯片达到的算力，并且能效比达到10倍以上

高时延确定性、高精度

基于ReRAM的存算一体设计，外围物理环境改变不会降低精度，保证了芯片在应用场景下的高稳定性；另一方面，全数字化设计让芯片保证大算力的同时还能做到支持高精度，可实现AI大算力的多场景应用

易部署

软件调优简单，易于客户工程落地，可以打破对于国外GPU生态的强依赖

典型应用场景

中心侧

- 大模型
- 数据中心
- 金融
- 教育
- ……

边缘侧

- 自动驾驶
- 特种车辆
- 无人机
- 智转数改
- 工业检测
- 安防
- 超分辨率
- 智慧交通
- ……

数据来源：亿铸科技，甲子光年智库整理

目录

CONTENTS

甲子光年
JAZZYEAR

Part 01 产业基石，算力是AIGC产业的催化剂

Part 02 软硬兼得，AI新世代呼唤工程化导向的算力支撑

Part 03 层见叠出，商业浪潮下的算力选择思考

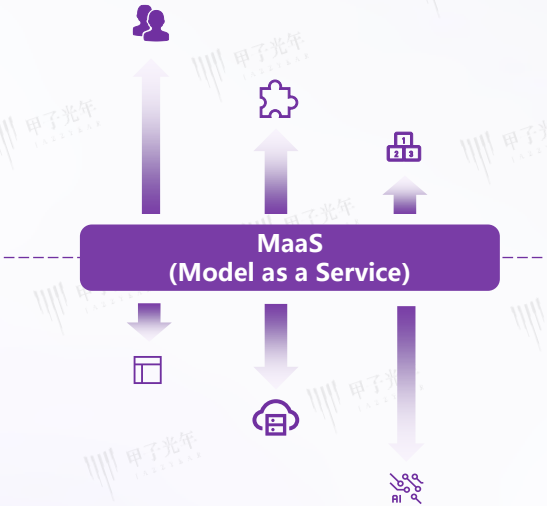
Part 04 实践真知，AIGC产业算力实践的新范式

Part 05 来日正长，AI技术的翻涌带来无限可能

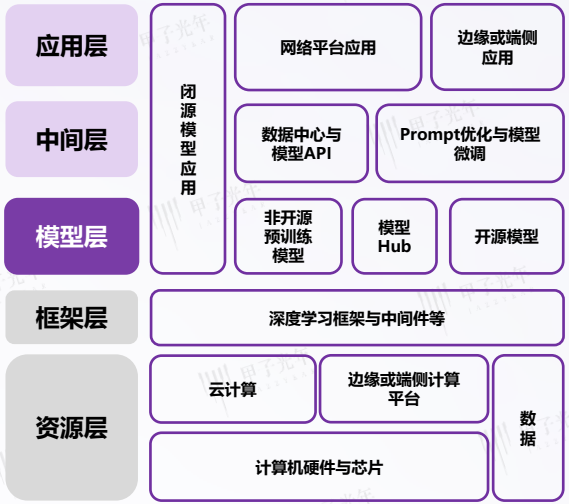
基于模型的复杂技术栈将形成产业更多生态组合，推动算力层能力下移

- AIGC的时代仍在快速发展，短期内算力服务商、AI技术服务商、AI应用企业将相互合作，探索可能场景及模式。
- 模型仍需要进化及迭代，并且训练模型的工程化暂无唯一确定的答案，模型的工程化能力还需要探索。

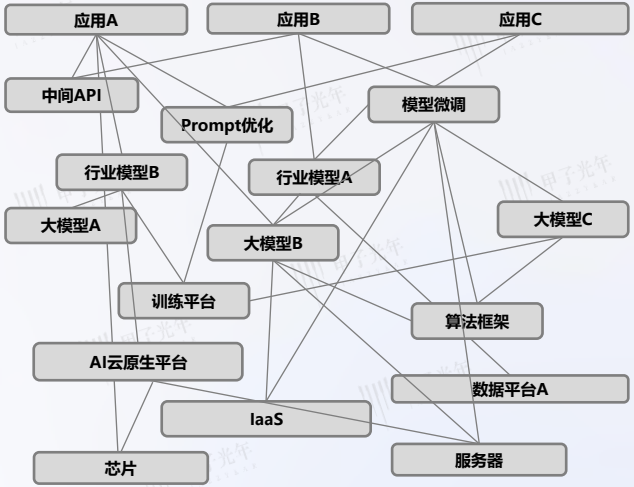
“MaaS”模式形成“算力结合模型”的服务理念



层次复杂的技术栈



更多的生态合作模式亟待解锁，尤其是算力层



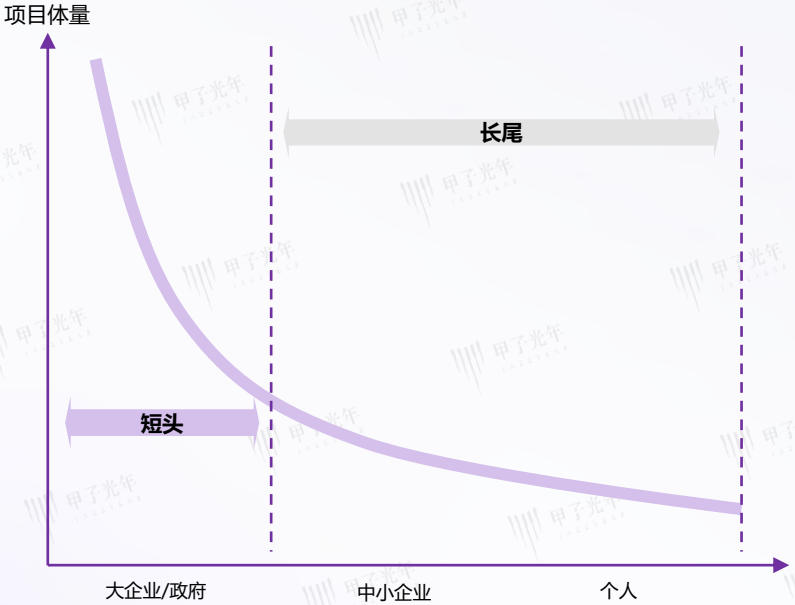
以上为示例

数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

AIGC会出现大量的长尾场景，算力层需要根据应用寻找细分赛道

- 随着大模型技术的出现，并且多个开源模型进入市场，AIGC的商业价值更多来自于“长尾”而非“头部”场景。在大模型出现之前，AI技术在长尾场景中的落地困难在于AI技术实现效果所需要的成本，大量专业细分的场景可以实现小成本实现，即来自于大模型基础上进行模型能力迁移而生成的中小模型。

大量长尾场景可以重新被AI技术实现



AI算力层并非去追求大算力，而是关注大量未发掘场景的可能性

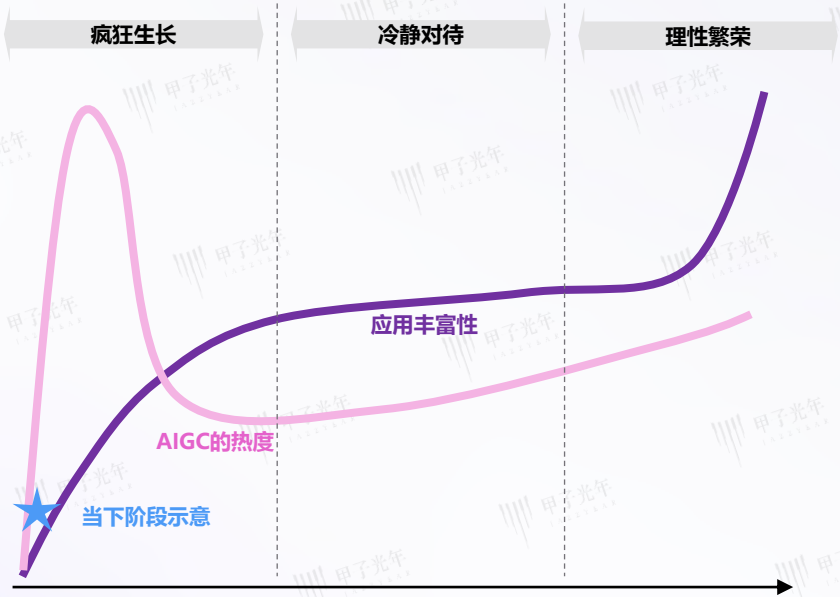
		产业各层商业逻辑
应用层		多点开花，在典型/行业场景内出现集中
中间层		多点开花，在典型行业/场景内出现集中
模型层		模型数量及参数量不断升级，支持上层应用
算力层	云服务	云服务厂商不断向上拓宽AIGC服务能力，实现尽量去适配多场景的能力
	智算硬件	实现芯片算力的最大适配及利用率，满足多行业的硬件选型与网络适配
	芯片	推理端的机会可能更多，针对成本有限的中小客户及细分赛道进行专业性优化

数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

AIGC的应用落地是一次演化过程，技术及垂直应用尚无定势

- AIGC的大量应用还未落地，商业模式尚未稳定，目前大量的资源及企业在进行路线探索。算力企业将面临一次场景丰富选择的机遇，国产算力服务厂商有望在国内应用探索期实现突破。

AIGC热度及应用丰富性分析，AIGC时代才刚刚开启



部分应用垂类应用已经引发用户长期关注

SimilarWeb数据：周（2023年7月9日-2023年）7月15日访问量同比增长

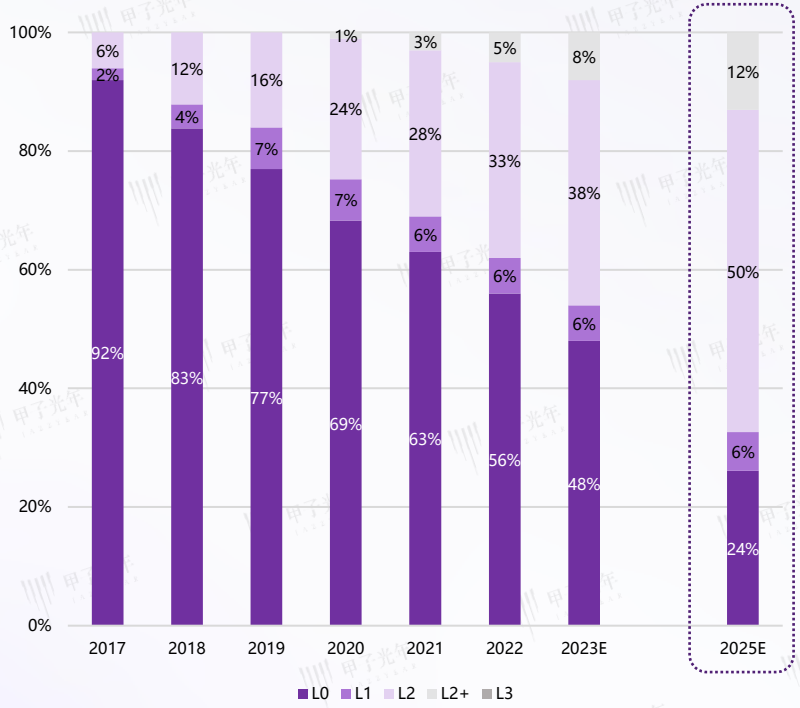
Tome（教育辅助	18.73%
Shopify（商家客服咨询）	5.58%
NovelAI（小说生成工具）	16.00%
CharacterAI（互动聊天）	5.58%
CodeWhisperer（编程辅助）	5.60%
Midjourney（图像生成）	10.24%

数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

终端算力的可能性，智能驾驶与AIGC的结合建立算力新赛道与格局

- 辅助驾驶是重要的AI应用场景，也是AI云端算力及终端算力的结合点，随着AI需求的增加，AI算力需求增加，如AI芯片厂商需要担起研发和整合的责任，其角色不断向整车厂靠近，将逐步整合Tier1和Tier2厂商的能力，和整车厂合作去打造整体的智能驾驶解决方案，重塑传统供应链的界限。

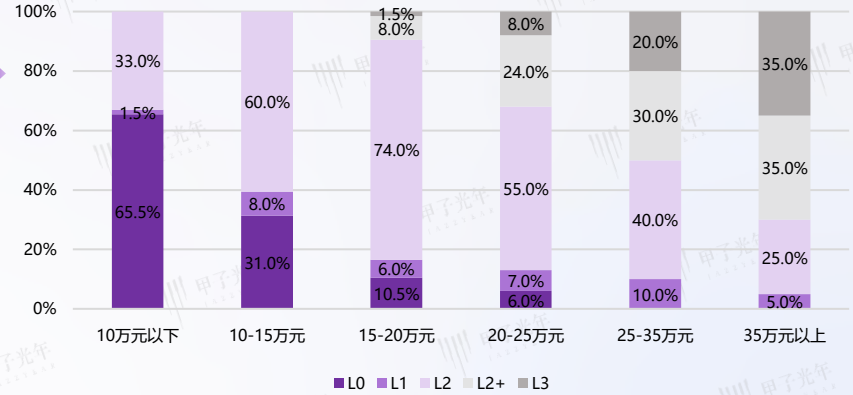
2017-2050E分智能辅助驾驶渗透率



AIGC在智能驾驶中的作用示例

自动驾驶场景的重建和数据生成	合并用于检测不同任务的小模型
数据自动标注	车道拓扑预测
测车端模型的性能上限	物体检测
数据挖掘	自动驾驶仿真

2025E分价格段智能驾驶分级别渗透率预测

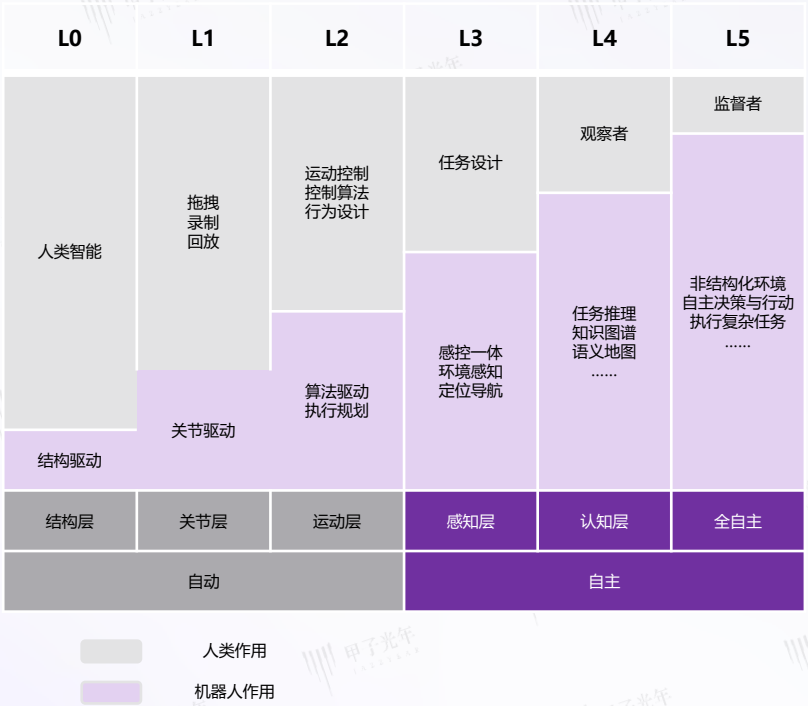


数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

大模型可推动现有机器人智能程度的突破，同时需要机器人端侧的算力支撑

- 机器人应用场景逐渐泛化，“自动”向真正地“智能”演进能力突破是关键。未来的智能机器人可拥有更丰富的传感器，不仅能获取并处理外部综合信息，甚至能据此自己制定行动目标，其智能主要体现在感知交互、独立决策、自我优化三个方面。其自主性的技术能力突破是当下关键，AIGC技术可成为当下机器人智能的突破关键点。

机器人的智能程度分级，L3-L4需要大模型的突破



机器人的应用场景

制造业	焊接、装配、喷涂、搬运、磨抛等机器人
农业	耕整地、育种育苗、播种、灌溉、植保、采摘、分选、巡检、挤奶等作业机器人
建筑	测量、材料配送、钢筋加工、混凝土浇筑、楼面墙面装饰装修、构部件安装和焊接、机电安装等机器人
能源	能源基础设施建设、巡检、操作、维护、应急处置等机器人
商贸物流	自动引导车、自主移动机器人、配送机器人、自动码垛机、智能分拣机、物流无人机等产品
商业社区	餐饮、配送、迎宾、导览、咨询、清洁、代步等商用机器人、以及烹饪、清洗、监护、陪伴等家用机器人
医疗健康	手术、辅助检查、辅助巡检、重症护理、急救、生命支持、康复、检验采样、消毒清洁等医疗机器人
养老服务	残障辅助、助浴、康复训练、家务、情感配合、娱乐休闲、安防监控等助老助残机器人
教育	交互、教学、竞赛等教育机器人产品机编程系统，分类建设机器人
安全应急	矿山、民爆、社会安全、应急救援、极限环境等机器人

数据来源：公开资料、专家访谈、甲子光年智库总结整理，2023.08

甲子光年
JAZZYEAR

THANKS

谢　谢

北京甲子光年科技服务有限公司是一家科技智库，包含智库、媒体、社群、企业服务版块，立足于中国科技创新前沿阵地，动态跟踪头部科技企业发展和传统产业技术升级案例，致力于推动人工智能、大数据、物联网、云计算、AR/VR交互技术、信息安全、金融科技、大健康等科技创新在产业之中的应用与落地

关注甲子光年公众号

扫码联系商务合作

分析师

刘瑶微信
18401669467

智库院长

宋涛微信
stgg_6406