AIGC算力全景与趋势报告

分析师：丁乔
dingqiao@qbitai.com
量子位智库 QbitAI Insights

序 言

2023年无疑是AIGC元年，ChatGPT引发的各类大模型竞赛中，行业内绕不开的一个话题便是算力从何而来。

算力目前已经在AIGC产业内形成新共识——算力成为AIGC发展的关键基石。随着英伟达今年一系列不断推陈出新的产品动作，可以看到国际上最先进的算力厂商如今已迈向由超级芯片组成的算力集群阶段。

此外，算力厂商也无疑成为AIGC产业下的率先受益方。然而，随着大模型参数的不断增长，OpenAI近期表明算力成为其发展的挑战之一。在AIGC产业繁荣的当下，可以预见的是未来对算力的需求会越来越大。那么，在这场AIGC盛宴中，应该如何应对当下面临的「算力危机」呢?

在《AIGC算力全景与趋势报告》中，量子位智库将从我国算力产业现状、算力产业变革、趋势预判等角度出发，通过广泛调研与深度分析，全面立体描绘我国当前AIGC算力产业全景与趋势。

我们期待，能够与众多投入、关注、期待中国AIGC算力产业的伙伴一起，共同见证并打造中国AIGC算力产业的蓬勃未来。

目 录

01

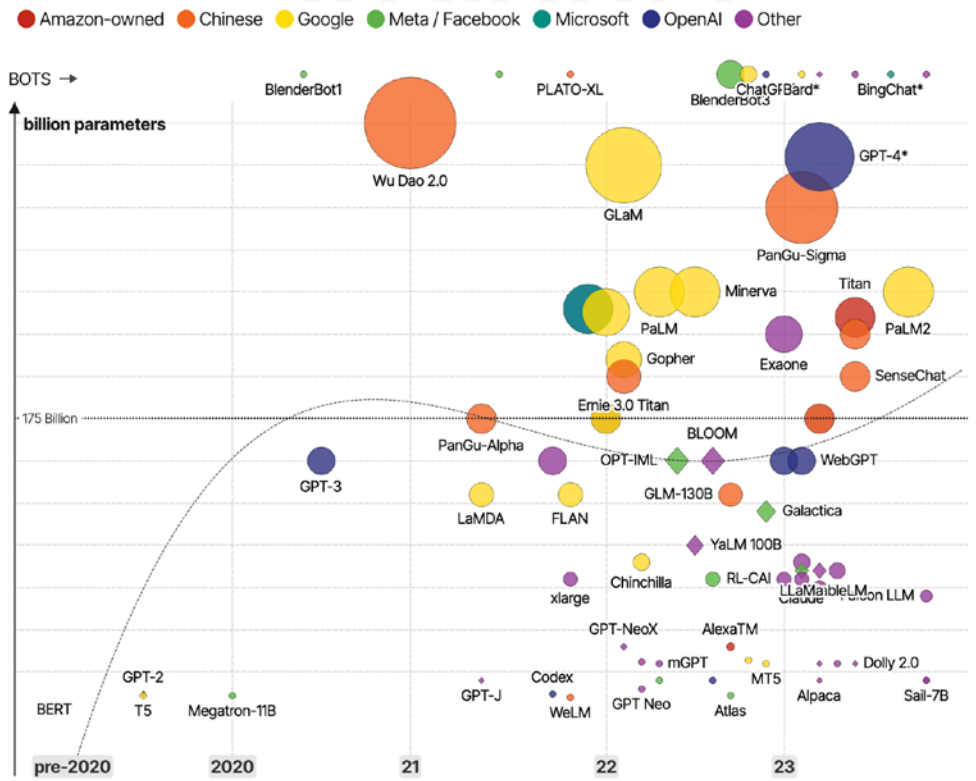
AIGC驱动，算力产业机遇空前

AIGC潮起，算力产业挑战巨大，机遇空前

OpenAI发布ChatGPT属于GPT系列中的聊天机器人模型。GPT系列中，GPT3是由1750亿参数组成的语言模型，而GPT4的参数更是达万亿级别。国内目前公布的大模型参数规模也普遍在百亿至千亿级别。如此庞大的参数规模，对于芯片提供商、云服务厂商以及服务器厂商都产生了新需求。

全球范围内，GPT具备从底层改变各行业规则的能力，作为AIGC产业的基建，算力产业在未来有望成为一项公共服务渗透入各行各业。基于此，智算中心作为公共算力基础设施，成为AIGC基建中的关键环节。

大模型参数量变化



来源：Information is Beautiful

云计算厂商	• 游戏规则被改写，MaaS能力成为竞争的关键变量
智算中心	• 在算力需求暴涨、数据和模型资源稀缺、AI技术广泛落地背景下，智算中心成为地区AI新基建
服务器厂商	• 大模型训练驱动AI服务器需求暴涨，并且正在催生新物种：AI模型一体机
芯片	• GPU为核心的AI训练芯片供不应求，是AIGC算力产业最大挑战和最大机遇

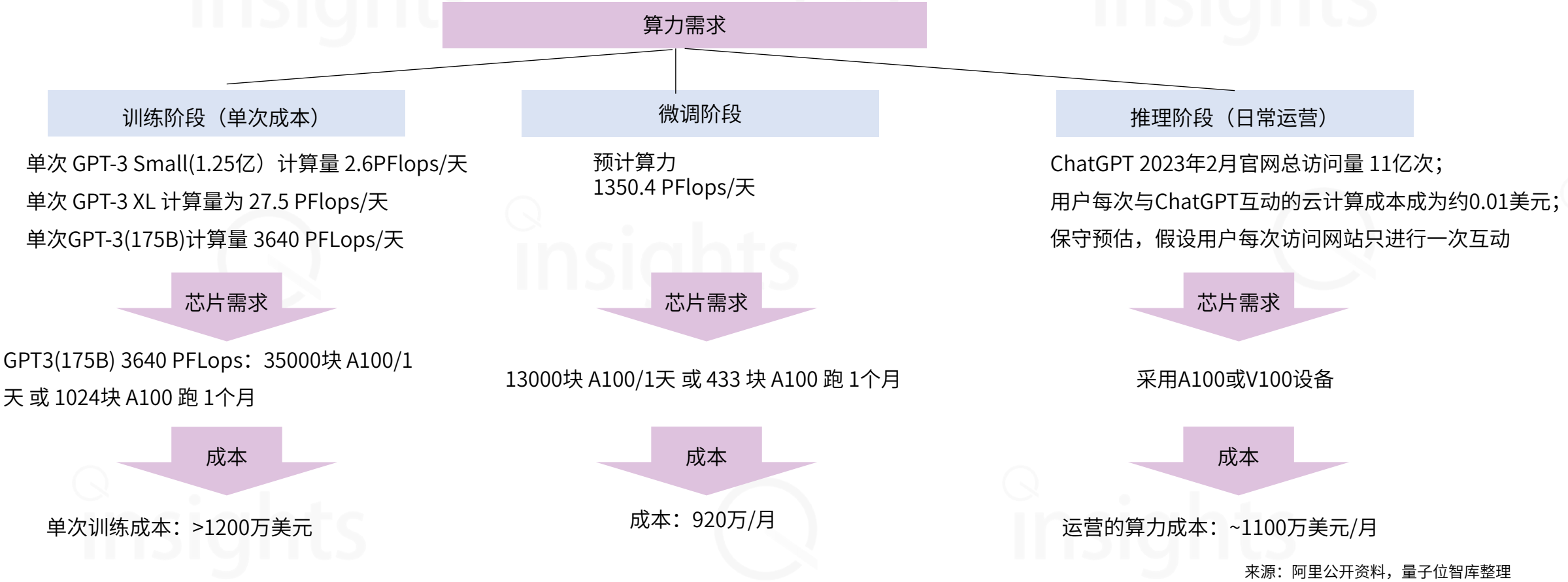
芯片：大模型训练需求暴涨，GPU供不应求

- 需求

当前大模型参数量在百亿至千亿参数规模，在训练阶段，对芯片的需求从CPU+加速器转变为以GPU主导的大规模并行计算。未来，当多数大模型参数规模到达万亿级别，将产生更大的算力需求。在单芯片性能之上，智算中心能够通过算力的生产-调度-聚合-释放，支持AI产业化发展。

- 缺口

目前市场对于英伟达芯片的需求远大于供给。经测算，一万枚英伟达A100芯片是做好AI大模型的算力门槛。国内具备此量级的公司最多只有1家，而GPU芯片持有量超过一万枚的企业不超过5家。



来源：阿里公开资料，量子位智库整理

服务器：业务增长显著，高端芯片AI服务器火爆

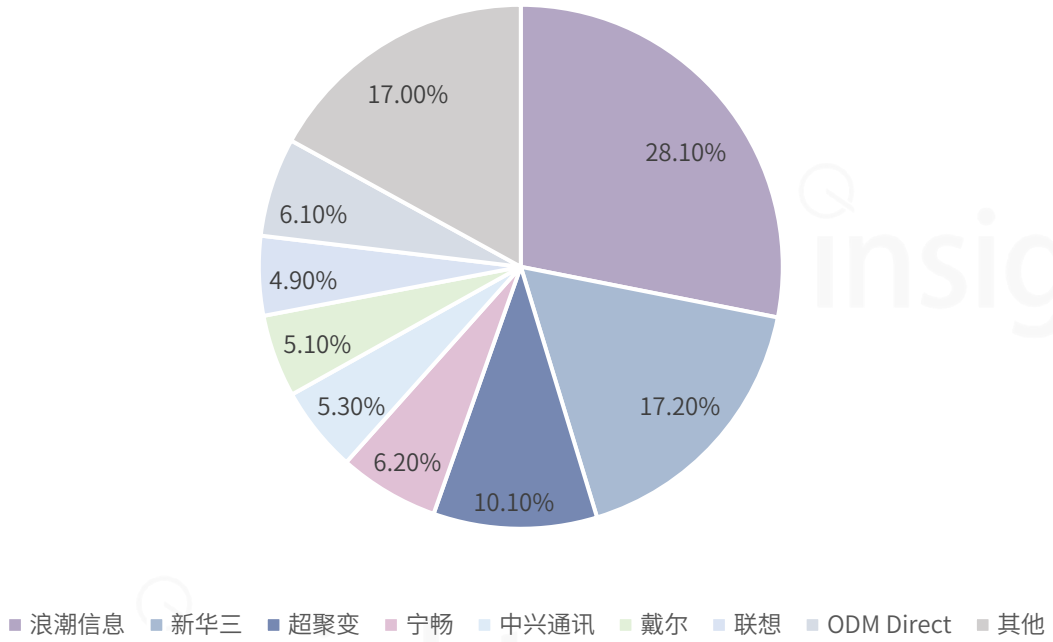
- 现状

AIGC产业的发展将加剧AI服务器行业的增长速度，国产服务器厂商普遍业务增量在30%以上；国内市场中，服务器重新进入洗牌期。

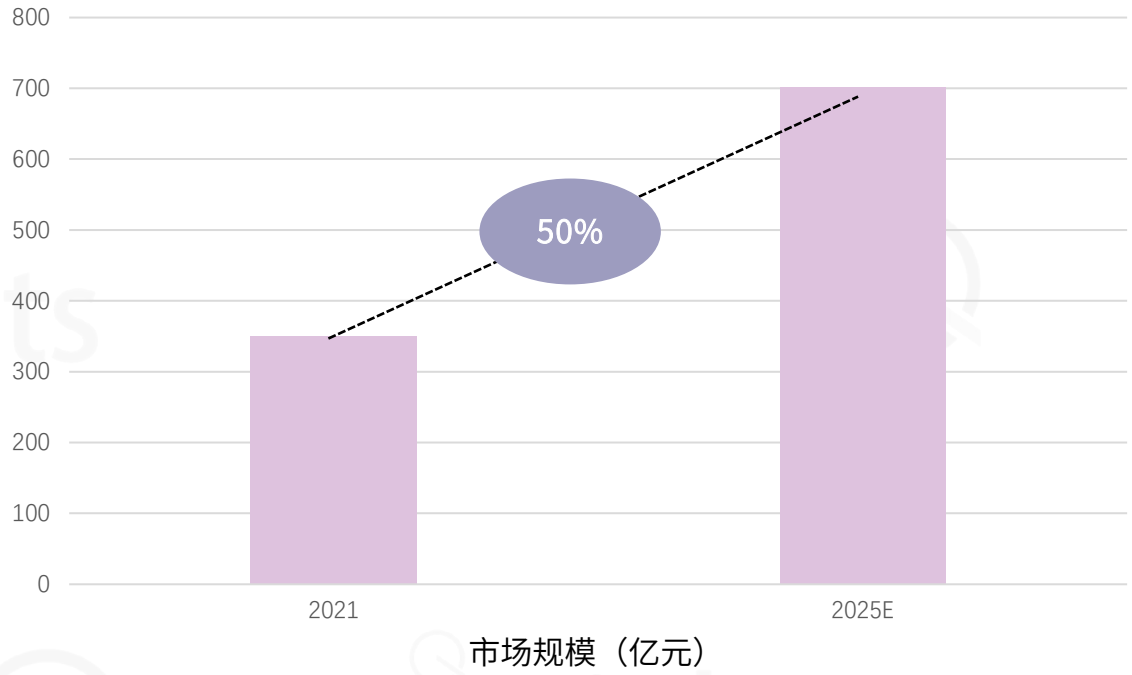
- 需求趋势

由于AIGC对于高性能计算的需求，云厂商在服务器的选择上以AI服务器为主。据IDC数据，2025年全球AI服务器市场规模将达317.9亿美元，年复合增长率为19%。英伟达GPU短期内面临产能不足问题，或将一定程度上限制AI服务器生产，从而影响出货量。

2022年中国服务器市场份额占比



2021-2025中国AI服务器市场规模预测



数据来源：IDC，量子位智库整理

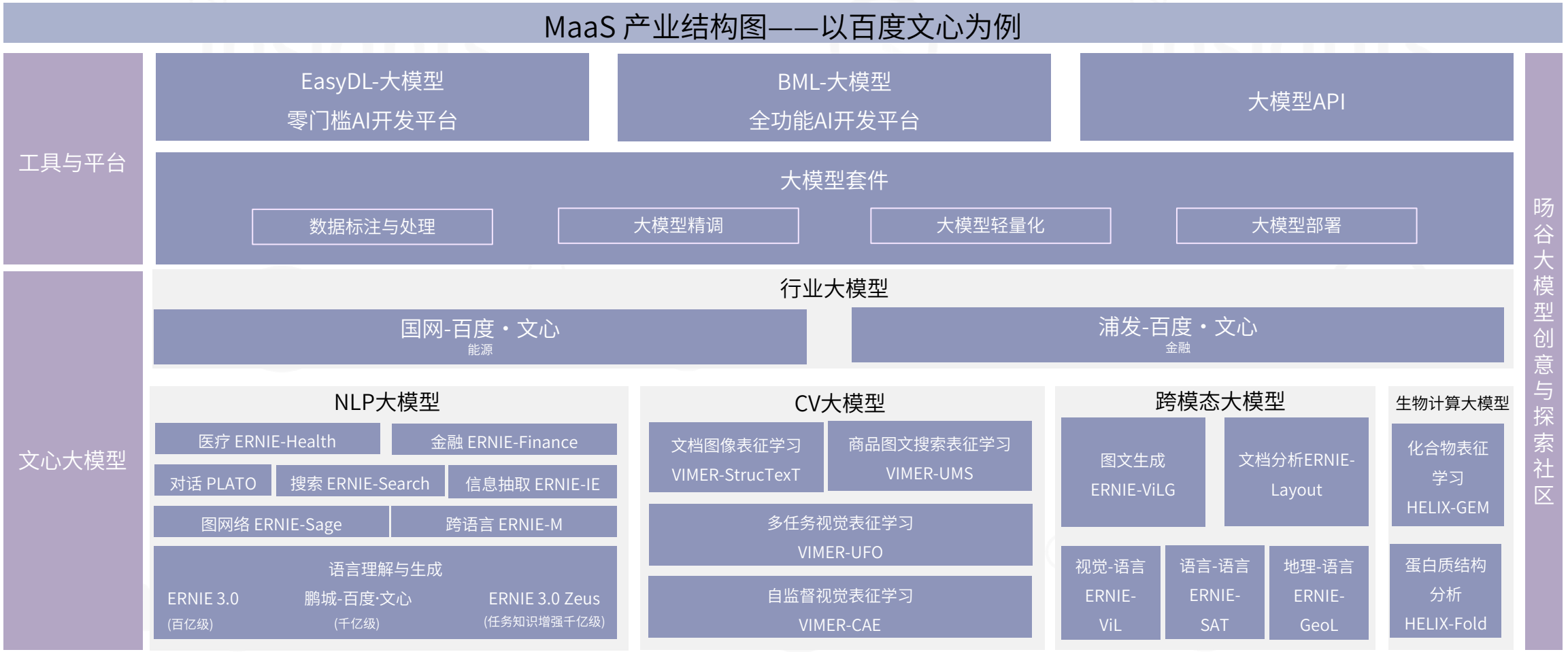
云计算厂商：服务范式变革，MaaS带来新商业路径

- 范式转变

MaaS成为云计算服务的新范式，云计算判别标准从算力能力转向「云智一体」的AI产品能力。

- 成本

自研芯片：根据 IDC 2018年服务器成本结构数据显示，高性能服务器中，芯片成本占比高达 50%～83%；全球头部云厂商（谷歌、微软、腾讯等）为摆脱过于依赖芯片厂商的局面，均加大芯片自研力度。



智算中心：基建级AI算力供应，打造地区经济增长新引擎

《智能计算中心创新发展指南》指出，在智算中心实现80%应用水平的情况下，城市/地区对智算中心的投资可带动人工智能核心产业增长约2.9-3.4倍，带动相关产业增长约36-42倍；

未来80%的场景都将基于人工智能，所占据的算力资源主要由智算中心提供，智算中心将成为经济增长的新动力引擎。

公共基建

全国超30座城市落地智算中心：

北京、天津、河北、南京、无锡、宁波、杭州、武汉、沈阳、成都、哈尔滨、许昌、广州、宿州、乌镇、昆山、甘肃、长沙

……

企业方

阿里云张北超级智算中心、乌兰察布智算中心

商汤科技人工智能计算中心

百度智能云-昆仑芯（盐城）智算中心

百度智能云（济南）智算中心

腾讯长三角（上海）人工智能先进计算中心

腾讯智慧产业长三角（合肥）智算中心

曙光5A级智算中心

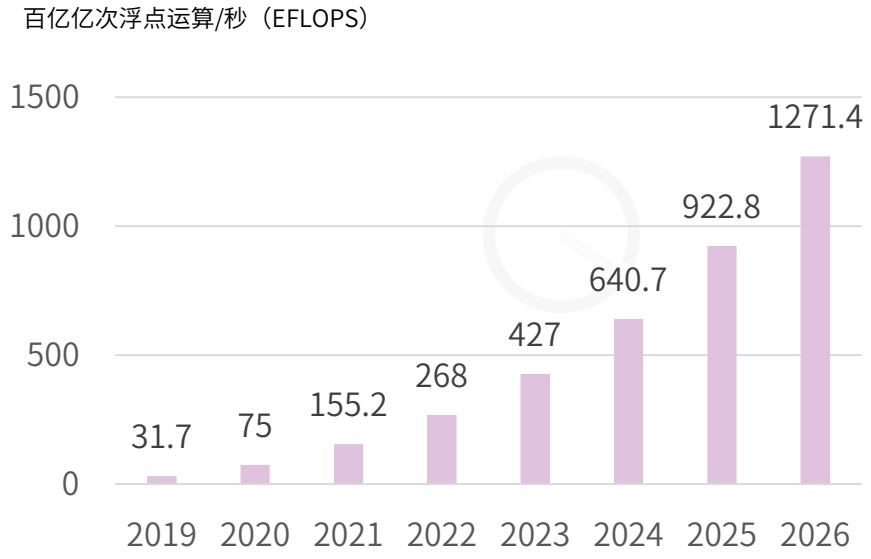
克拉玛依浪潮智算中心

中国电信京津冀大数据智能算力中心

中国联通广东 AI 智算中心

……

中国智能算力发展情况及预测



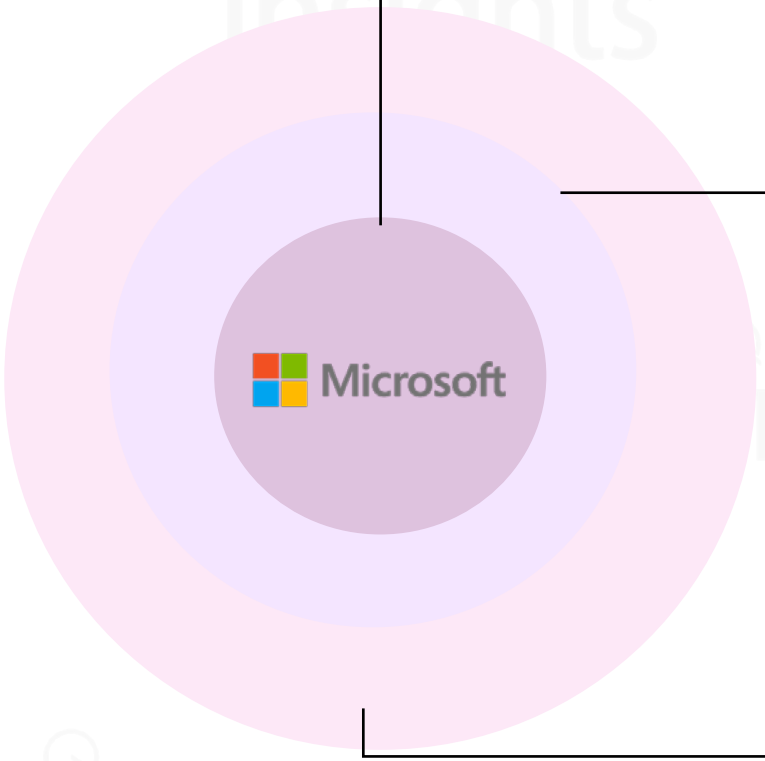
来源：国家信息中心《智能计算中心创新发展指南》，IDC，量子位智库整理

02

AIGC算力产业全景

AIGC算力现状链路：芯片—服务器—云平台—模型应用

以微软为例



芯片资源

- 外部：Azure云服务为ChatGPT构建了超过1万枚英伟达A100 GPU芯片的AI计算集群
- 内部：微软正在自研AI芯片——雅典娜（Athena），将由台积电代工，采用5nm制程

首个目标：为OpenAI提供算力引擎，以替代英伟达A100/H100

云基础设施平台Azure

微软是OpenAI唯一云服务提供商，为GPT训练提供计算资源、存储资源、自动化部署和管理等支持

模型即应用（MaaS）

1）Azure OpenAI 服务：

企业级解决方案：借助 Azure OpenAI，用户可以汇总文本、获取代码建议、为网站生成图像等

2）Microsoft 365 Copilot：

使用了GPT-4作为其核心的LLM，将用户的自然语言输入转化为高效的生产力工具，集成在Word、Excel、PowerPoint、Outlook、Teams等多个应用中

芯片层现状：AIGC算力2大路线，GPU通用路线和AISC专用路线

AI芯片目前有两大路线，一种是英伟达代表的GPU路线，更适合当前AIGC产业对大算力的需求，与AIGC大模型的训练及推理适配度极高。另一种路线则是以国内华为（主力产品）、寒武纪厂商为代表的专用AI芯片路线，此路线下的芯片更适用于垂类小模型，为其提供能效比更高的芯片。此外互联网云厂商的自研芯片也是专用路线，芯片主要服务于自家产品，为自身产品打造性能更优的算力底座。

通用芯片路线 (Graphics processing unit)

能够完成多样化算力任务

优势

- 擅长大规模并行计算
- 兼容英伟达生态，是最快也是最适用于当下的解决方案

局限

- 在厂商被迫「重复造轮子」的前提下，追赶上英伟达的难度极高
- 芯片总体功耗高

专用芯片路线 (Application-specific integrated circuit)

用来执行专门/定制化任务

优势

- 专用场景中能够做到更优的能效比
- 跳出当前的已有生态，长期来看有可能实现真正超越

局限

- 研发周期长、商业风险较大，产品易受市场变化影响
- 不易扩展，难以满足后续增加功能的需求

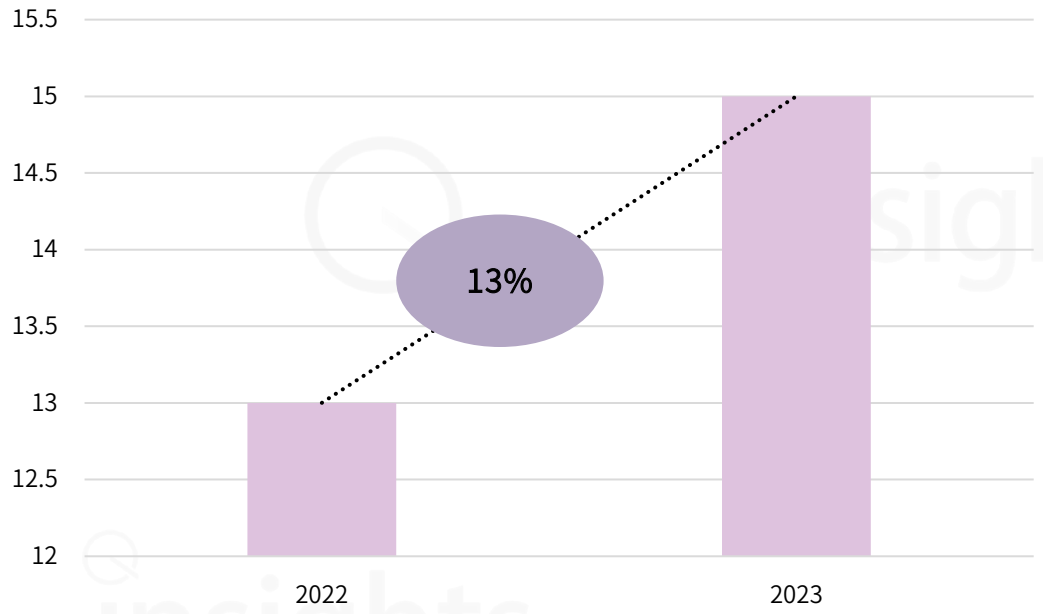
服务器层现状：AI服务器成主要增长点，采购占比互联网客户为主

服务器作为算力的载体，是AIGC基础设施的核心硬件。由于AIGC对于高性能计算的需求，云厂商在服务器的选择上以AI服务器为主。据IDC数据，2025年全球AI服务器市场规模将达317.9亿美元，年复合增长率为19%。AIGC产业的发展将加剧AI服务器行业的增长速度，国产服务器厂商普遍业务增量在30%以上；

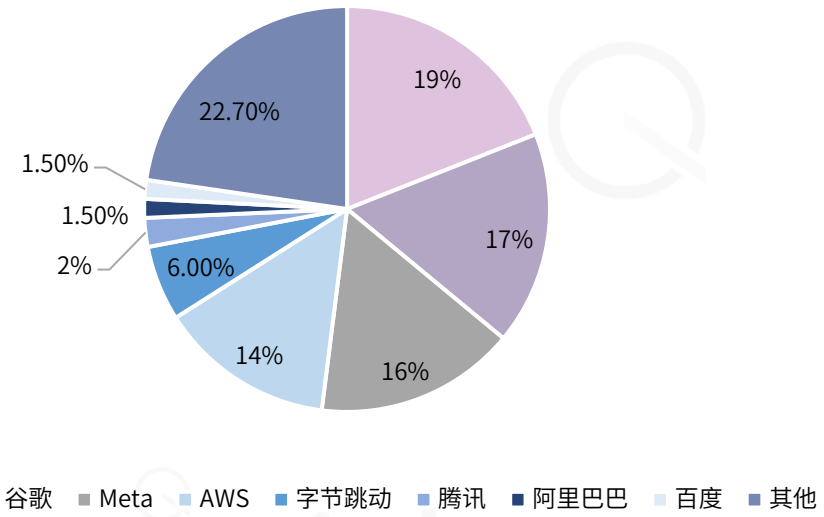
TrendForce日前发布预测，指出随着AI服务器与AI芯片需求同步看涨，预计2023年AI服务器（包含搭载GPU、FPGA、ASIC等主芯片）出货量将接近120万台，年增38.4%，并将2022-2026年AI服务器出货量年复合增长率上调至22%。

2022年，国内互联网大厂成为AI服务器的最大买家；2023年，随着AIGC的爆发，根据业内消息，互联网厂商依旧是AI服务器的最大买方。

2023AI服务器出货量预测



2022年AI服务器采购量占比



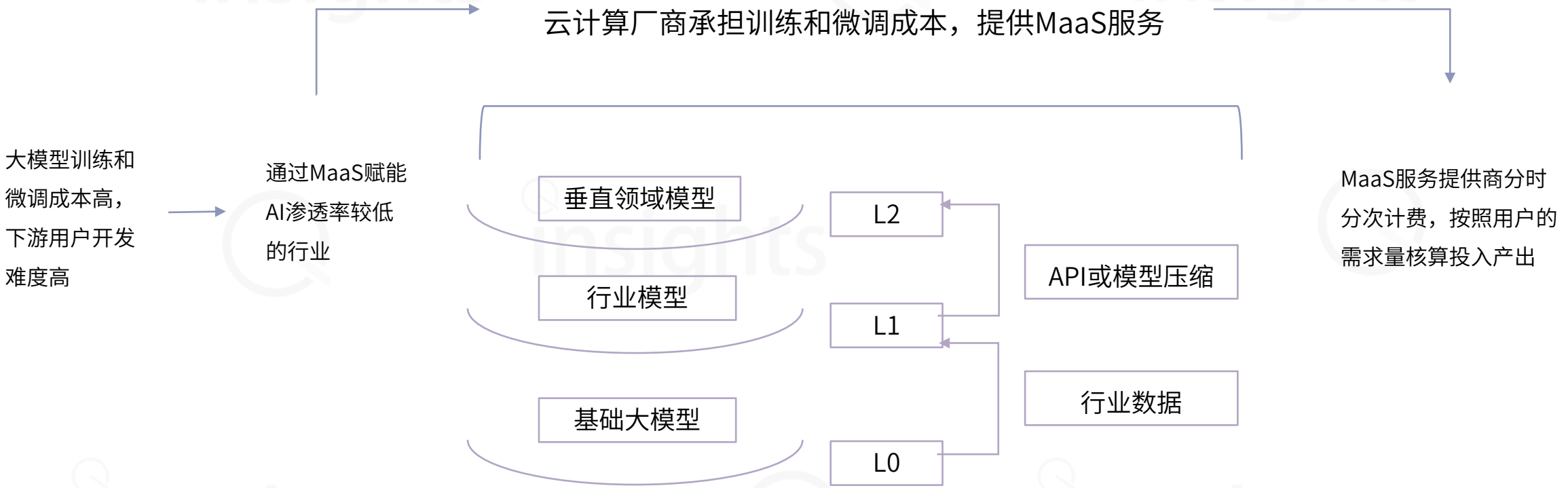
数据来源：TrendForce，量子位智库整理

云计算现状：MaaS重塑服务模式，新老玩家重构竞争力

大模型成为MaaS的基座，MaaS所打造的商业模式也是大模型厂商的主要变现模式——基于大模型产生有实际应用价值的产品。

MaaS模式最早由阿里提出，随后互联网大厂、人工智能企业（如商汤）均已引入MaaS模式。此外，互联网大厂、华为等企业已经将自研芯片用于MaaS底座构建中。

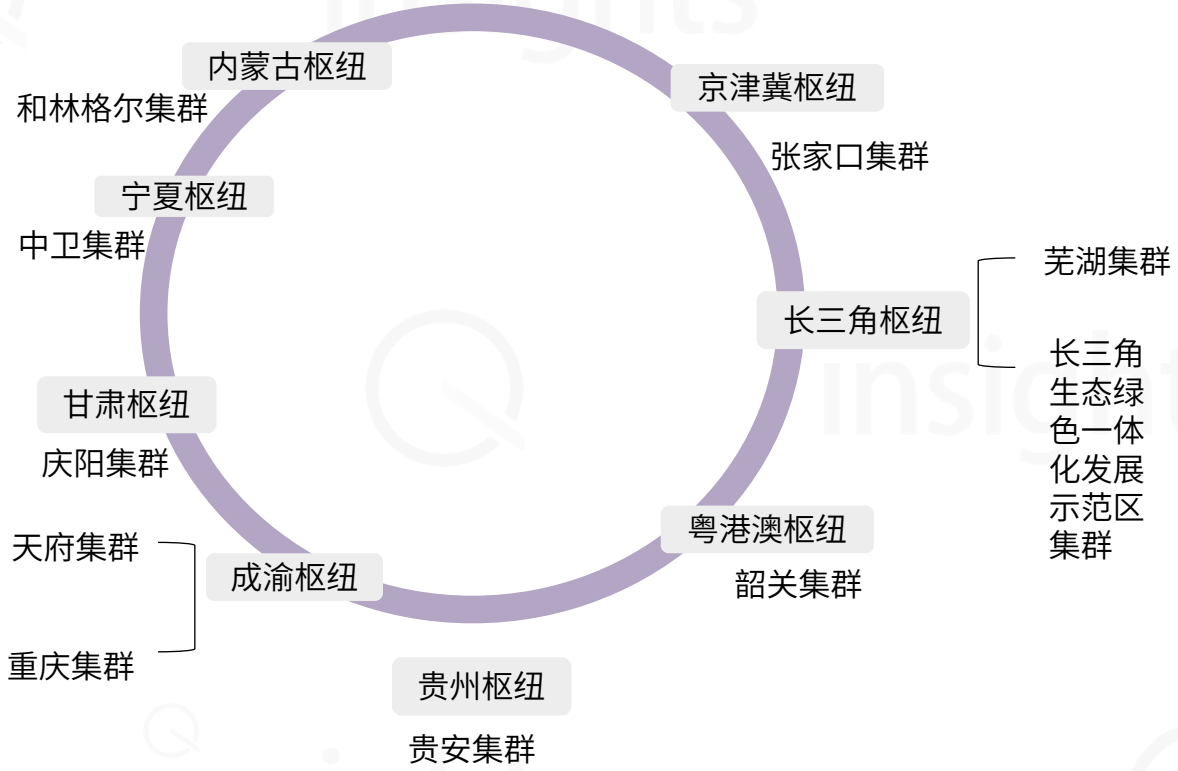
云厂商是MaaS的提出者，也是主要参与方。MaaS模式基于大模型，能够最大限度消除大型企业数字化过程中规模化、标准化壁垒，降低企业部署难度；对于C端用户来讲，MaaS可在不同层级里产生价值，有望为云计算厂商带来新增长曲线。



智算中心现状：地域发展差异明显，「东数西算」成算力调度关键

智算中心是对原有数据中心的升级，其提供的首要是AI算力。具体来讲，智算中心提供包括算力、框架、模型，以及支持应用场景具体的基础设施，将不同层级进行打包，通过本地化部署完成智算中心建设。相比于数据中心，智算中心更贴近应用和产业方。

东数西算整体规划



从计算设备分布来看

北京、广东、浙江、上海、江苏在服务器和AI服务器市场中居前五，市场份额总计分别达到75%和90%（2021年数据）。

从需求角度看

AIGC算力需求主要来源为京津冀地区、长三角及大湾区。

从供给角度来看

目前智算中心多分布在东部和中部省份，而AIGC业务需要处理海量数据导致东部算力资源成本过高。将大模型训练等对计算要求高的任务移至西部地区，形成“东数西训”，能够有效降低成本，实现算网资源综合成本最优。

具体来讲，针对算力需求供需不平衡等问题，需要通过算力调度将东部的算力和数据处理需求转移至成本较低的西部地区。其中，优化东西部之间互联网络和枢纽节点间直连网络是提升算力调度水平的关键。

AIGC算力产业全景图

MaaS层

阿里云 通义千问 | 百度智能云 文心 | 腾讯云 混元 | 华为云 盘古 | 商汤 日日新 | 云从科技 从容 | 火山引擎 火山方舟 | 京东云 言犀

云计算平台

阿里云 | 百度智能云 | 腾讯云 | 华为云 | 火山引擎 | 天翼云 | 移动云 | 浪潮云

HUAYUN 华云 | UCLOUD优刻得 | 紫光云 | 金山云 | 首云 CAPITALONLINE | 九州云 Cloud | 易捷行云 EasyStack

服务器厂商

中科曙光 Sugon | inspur 浪潮 | H3C | HUAWEI | Lenovo 联想 | ZTE中兴 | Nettrix宁畅 | 安擎 ENGINETECH

计算类芯片

通用芯片

CPU

HYGON 中科海光 | HUAWEI | Phytium飞腾 | 阿里巴巴 | 龙芯中科 LOONGSON TECHNOLOGY | 兆芯 | 申威

GPU

HYGON 中科海光 | 天数智芯 Iluvatar CoreX | 壁仞科技 BIREN TECHNOLOGY | 摩尔线程 MOORE THREADS | AZURENGINE | 登临科技 | 景嘉微 JINGJIA MICRO

专用芯片

ASIC

HUAWEI | 阿里巴巴 | Cambricon 寒武纪

FPGA

京微齐力 | FUDAN MICRO | 紫光同创 PANGOMICRO | GOWIN高云 | ALOGIC

DSA

MOFFETT AI 墨芯人工智能

存储类芯片

DRAM

GXMT | GigaDevice 兆易创新 | 紫光国微 GUOXIN MICRO | dosilicon 东芯半导体股份有限公司 | 君正 Ingenic

NAND

君正 Ingenic | Netac 朗科 | GigaDevice 兆易创新 | TWSC | longsys | MaXio | dosilicon 东芯半导体股份有限公司

Nor Flash

GigaDevice 兆易创新 | Puya | 恒烁半导体 Zbit Semi, Inc. | dosilicon 东芯半导体股份有限公司 | XTX 芯天下

EEPROM

Puya | 聚辰半导体 GIANTEC SEMICONDUCTOR | FUDAN MICRO

3D NAND

长江存储 YANGTZE MEMORY

03

AIGC算力产业「五新」趋势

背景：算力供给趋于复杂，大规模运算需要系统级工程支撑

芯片在AIGC算力产业中是最底层也是最关键的硬件产品。AIGC爆发，既是芯片厂商的一个重要分水岭，也将芯片厂商的目标重新聚焦于大算力方向。

芯片作为算力直接来源，其发展逻辑是从应用端的需求出发，根据应用端所需要的算力特点提供相应的算力服务。在ChatGPT相关大模型爆发之前，国内芯片厂商一方面在做GPU布局，另一方面更多在满足垂直行业中的特定需求，且后者在国内市场更常见。此外，国产GPU厂商的设计初衷也多是按照推理芯片设计。

在AIGC爆发后，对芯片的需求集中在训练侧，并且对于训练芯片的算力要求极高，目前只有英伟达能够满足。然而，OpenAI 表示目前英伟达的产能已无法满足其更高的算力需求。未来，随着大模型参数量不断攀升，以及芯片制程走到尽头等问题，对于算力的定义将从单芯片性能逐渐转向超算/智算集群的计算能力。

国产处理器厂商的挑战与机遇

挑战

硬件

- 目前在高端AI芯片中，英伟达占据绝对优势，而英伟达的高端系列在中国只有存量没有增量。
- 在芯片代工层面，目前优于7nm制程工艺没有对应的国产代工厂可以承接。

软件

- 业内普遍认为国产芯片在10年内很难突破英伟达的CUDA生态。

机遇

- 市场将给予国产GPU厂商更多机会。
- 国产GPU厂商可选择成熟制程+先进封装的方案来达到与英伟达近似的性能指标。在服务器集群层面，通过高速互联技术实现高性能计算。
- 目前国产芯片厂商采用两种路径：
  1）兼容CUDA生态；2）构建自身生态
- 短期来看，兼容CUDA生态的厂商更适合为通用大模型提供算力。对于构建自身生态的厂商来说，其产品更适用于垂类小模型。

趋势01——新机遇：芯片竞逐高性能大算力，引入新计算架构

需求方变化

对芯片回归到最原始的需求 → 模型需要大算力支持

大模型不同阶段对应不同的芯片需求

大模型训练阶段

- 芯片类型：GPU为主
- 芯片需求：执行大量矩阵运算和计算密集任务
- GPU优势：高并发和浮点计算能力，可大幅提升训练速度
- GPU劣势：功耗高、成本高

大模型推理阶段

- 芯片类型：ASIC/FPGA/NPU与GPU均可
- 芯片需求：低延迟、低功耗（专用芯片更符合）
- 专用芯片优势：更高的能源效率和计算密度
- 专用芯片劣势：缺乏通用性

供给方变化

技术趋势

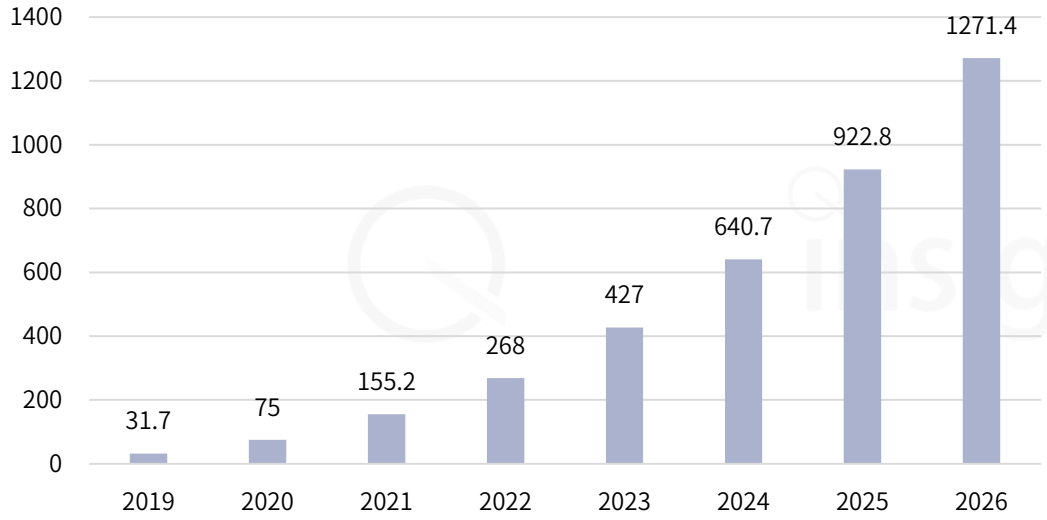
突破芯片算力性能	硬件层面	AI芯片厂商	存算一体架构	通过减少/消除数据搬运，降低功耗	亿铸科技、后摩智能、知存科技等
		GPU厂商	GPU+片间互联	实现大规模并行计算	海光信息、燧原科技等
	软件/算法层面	AI芯片厂商	稀疏计算	通过减少计算量，降低成本	墨芯人工智能
		高性能网络	RDMA互联	高带宽接入，提升通信性能，缩短训练时长	腾讯云星脉网络

趋势02——新增长曲线：AI服务器异军突起，红利曲线先训练后推理

AI大模型对算力需求呈指数级增长，使得具有更高配置的AI服务器成为AIGC算力的主要载体。相比于传统服务器，AI服务器的计算、存储以及网络传输能力能达到更高的水平。例如，NVIDIA DGX A100服务器 8 个 GPU+2 个 CPU 的配置远高于传统服务器 1~2 个 CPU 的配置。

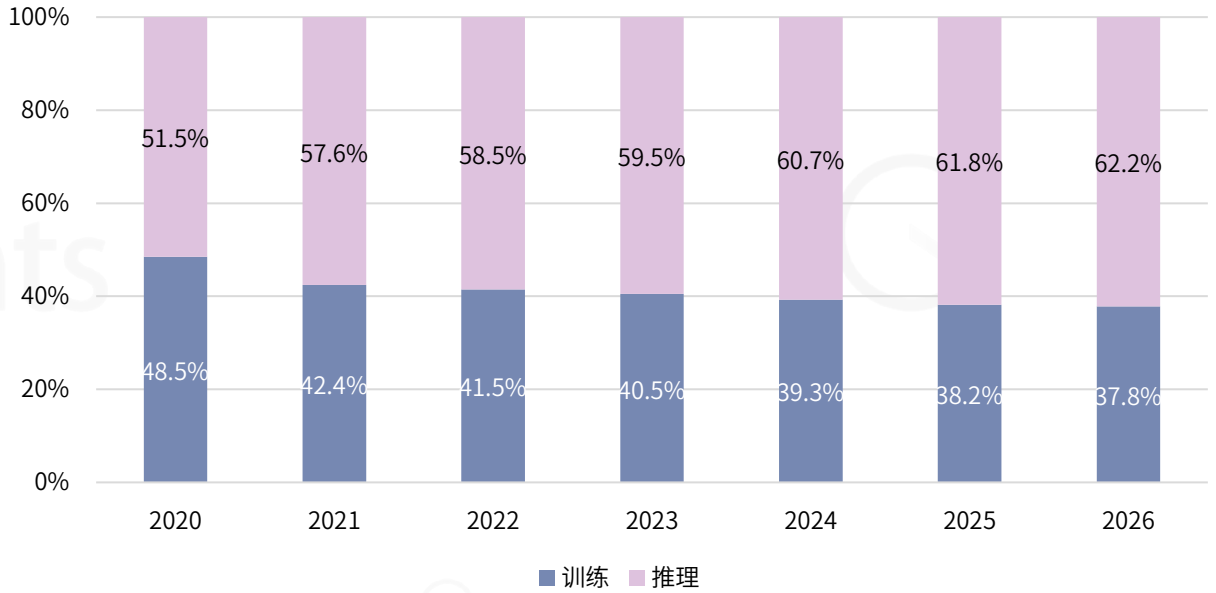
智算中心作为提供算力资源的公共基础设施平台，其算力机组以AI训练服务器和AI推理服务器为主。随着大模型训练阶段完成，未来AI服务器的主要需求将向推理侧转移。根据IDC的预测，到2026年，AIGC的算力62.2%将作用于模型推理。

中国智能算力规模及预测



单位：百亿亿次浮点运算/秒(EFLOPS)

中国AI服务器工作负载预测



数据来源：IDC，量子位智库整理

趋势03——新游戏规则：MaaS重塑云服务范式，AIGC商业模式闭环

MaaS（模型即服务）：在算力、算法和应用层中嵌入大模型，以智能底座集成应用并统一对外输出。MaaS的本质是将行业内通用的基础技术提炼整合成服务，满足各类应用场景需求；

云计算服务能力的判别式从算力水平转向「云智一体」能力，在算力基础设施之外，核心竞争力变为把算力、模型和场景应用打造成标准化产品的能力。

商业化路径

AI 开发者　AI 研究者　AI 使用者　AI 爱好者 ……

多样化应用开发，更多面向C端市场

模型体验　模型使用　模型定制　云端模型部署

付费使用接口，直接调用基础模型，基于不通过行业的数据进行fine-tune，形成垂直大模型，更多面向B端市场

数据仓库　模型仓库　算力平台

大模型基础能力

C端市场

商业模式：软件订阅

B端市场

盈利模式：按需计费，根据实际计算量收费

趋势04——新物种：AI模型一体机呼之欲出，传统产业「开箱即用」

传统行业构建AIGC产品的痛点

- 数据 → 内外部数据流通中难以确权
- 模型 → 构建自身模型难度大
- 算力 → 算力资源紧缺

确保算法模型部署到服务器上并能高效运行 → AI模型一体机

- 云服务厂商在硬件和软件层面完成系统级工程、调试测试环节，最后在用户侧可以直接使用的AI模型一体机。
- 对于数据安全性敏感的行业，通过AI模型一体机，完成模型+服务器的一体化部署，能够保证数据的安全。

→ 用户开箱即用

主要参与方

硬件类（服务器）云厂商

- 优势

硬件部署能力，如何让模型在服务器上运行效率达到最高

- 劣势

软件算法能力以及大模型研发能力

互联网云厂商

- 优势

具备通用大模型能力

- 劣势

硬件能力

百度智能云 飞桨一体机-产品架构

趋势05——新基建：智算中心护航AIGC运营，算力租赁模式成新解

算力租赁模式可以有效降低大模型研发门槛，对于研发垂类行业的小模型企业来说，没有购买足够AI服务器的实力，公共算力基础平台将帮助中小型企业搭建其自身所需模型。企业无需购买服务器，通过浏览器便可访问算力中心，并使用算力服务。对于中小企业来讲，无需依赖云厂商所构建的大模型底座进行二次开发，而是通过租用公有算力平台的算力资源，研发垂类行业小模型。

大模型训练推理过程消耗大量算力资源，成本高昂 → 中小企业有模型研发需求，但无法承担高昂的算力成本 → 算力平台向B端用户直接销售算力

- 国外：以英伟达为代表的超级计算机，目前已建成5座AI工厂
- 国内：
  1）在建及投入使用的智算中心
  2）云厂商单独租赁

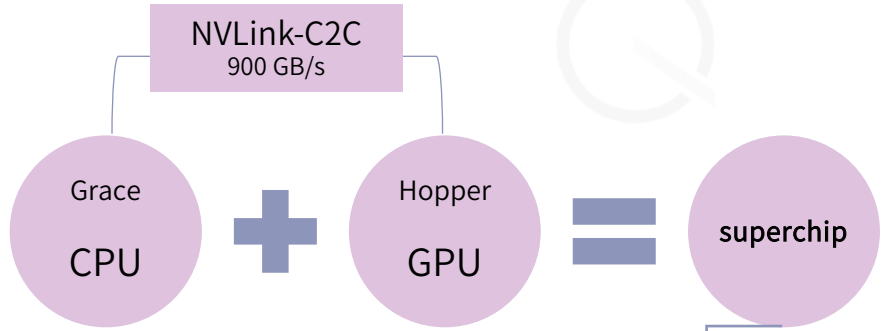
国内：智算中心完成系统级工程

AI算力一体化交付流程

生产算力 → 聚合算力 → 调度算力 → 释放算力

算力供需失衡的前提下，算力租赁有望成为行业内供给侧的最优解

国外：英伟达DGX Cloud模式



8个Ampere A100或Hopper H100旗舰HPC加速器，每个节点共有640GB显存

每月3.7万美元

算力租赁平台 DGX Cloud

04

AIGC算力产业周期预测

一个周期三个阶段：基建期、开发期，商业期

AIGC基建期

率先受益方：算力基础服务提供方

国内　国外

服务器厂商　GPU厂商

竞争要素

- 高性能芯片数量
- 计算集群建设能力
- 算力大小
- 带宽大小

AIGC开发期

大模型厂商「制胜点」

大模型持续迭代的能力

算法、算力、数据、知识

更具竞争力的企业具备两项能力

自研芯片能力　集成创新能力

AIGC商业期

技术创新型公司迎来红利

存算一体　光子芯片

类脑芯片

优势

功耗　成本　计算效率

0-2年　3-5年　5年以上

AIGC基建期：2023年~2025年

全球范围内，OpenAI的GPT初步实现智能涌现，其背后的算力支撑是英伟达高性能GPU。在追赶GPT的过程中，需要大模型企业在短时间内完成模型的训练过程。因此，在AIGC初期阶段，大模型研发企业不会考虑除英伟达之外的芯片作为训练阶段的芯片。

核心点

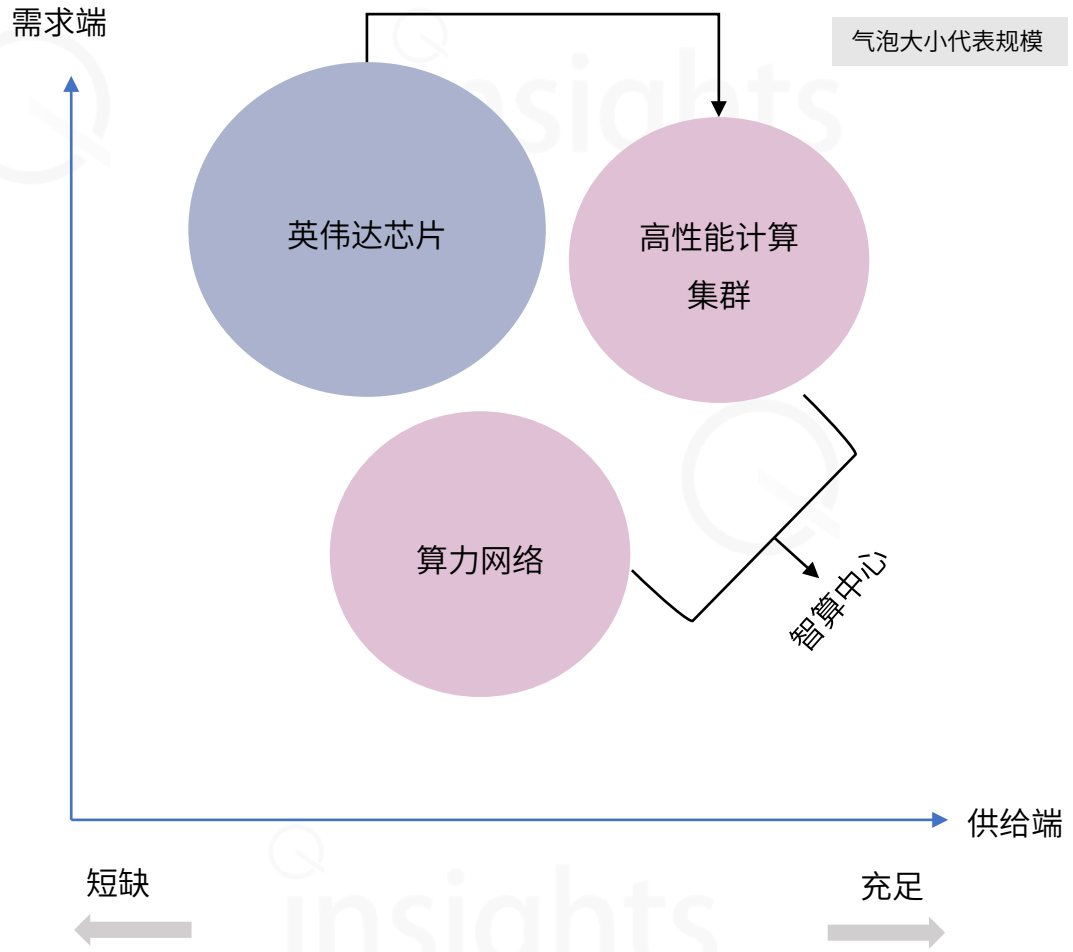
训练速度为王

- 英伟达掌握行业绝对话语权
- 英伟达拉动下游代工厂业务，二者成为AIGC基建期的关键奠基者

智算中心加快建设

- AI服务器需求保持持续、稳定增长，国内服务器厂商率先受益
- 算力网络建设是智算中心运营的关键，算网运营商在算力网络建设中为主导方

在初期阶段，高性能GPU厂商将成为最大收益方；同时，AI服务器厂商是此阶段的强势供给方。目前，AI服务器领域已经出现供不应求的局面



AIGC开发期：2026年~2028年

在中期阶段（5年内），推理芯片将成为主要需求方。相比于GPU的高算力高功耗以及对应的算力浪费，推理芯片更注重芯片的算效比，对于功耗和成本有更优的把控。此外，这个阶段也会是创新型芯片的机会。分析师预计存算一体芯片、类脑芯片、硅光芯片将有更多市场机会。

中期阶段AIGC市场将呈现收敛趋势，从百花齐放到逐步淘汰，此阶段主要是模型层公司之间的淘汰战。在此阶段，AI服务器厂商的红利期逐渐见顶，智算中心与超算中心走向融合；芯片也从GPU转向NPU/ASIC/FPGA/CPU等多种形式并存。创新型芯片路线中，看好存算一体架构的发展。

核心点

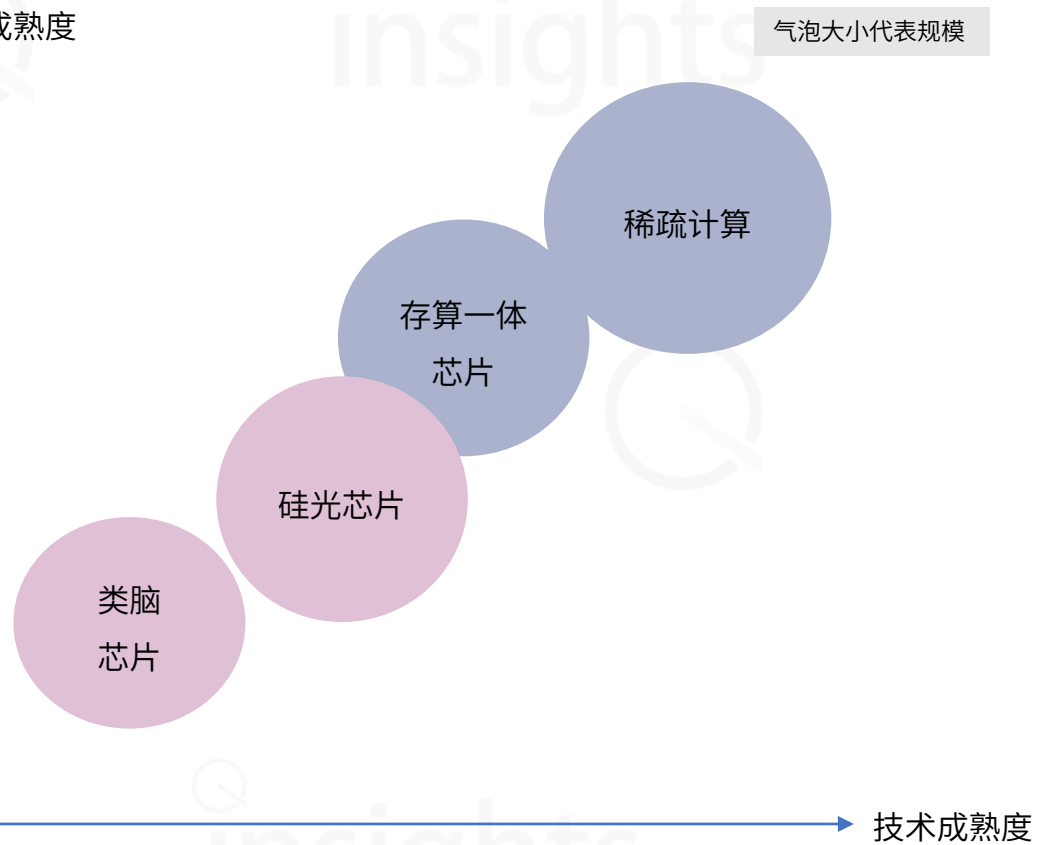
推理类芯片占比上升，芯片需求趋于多元化

- 大模型由训练阶段过渡到推理阶段，企业更加注重降低算力成本，对于功耗高的GPU集群，企业趋向寻求替代方案
- 能效比更高的芯片将迎来机会点

智算超算中心融合，满足多样需求

- 智算中心在满足人工智能业务的基础上，为了覆盖更多业务需求，将逐步与超算中心走向融合

中期阶段，具备底层创新能力的芯片厂商有望成为最大获益方



AIGC商业期：2028年后

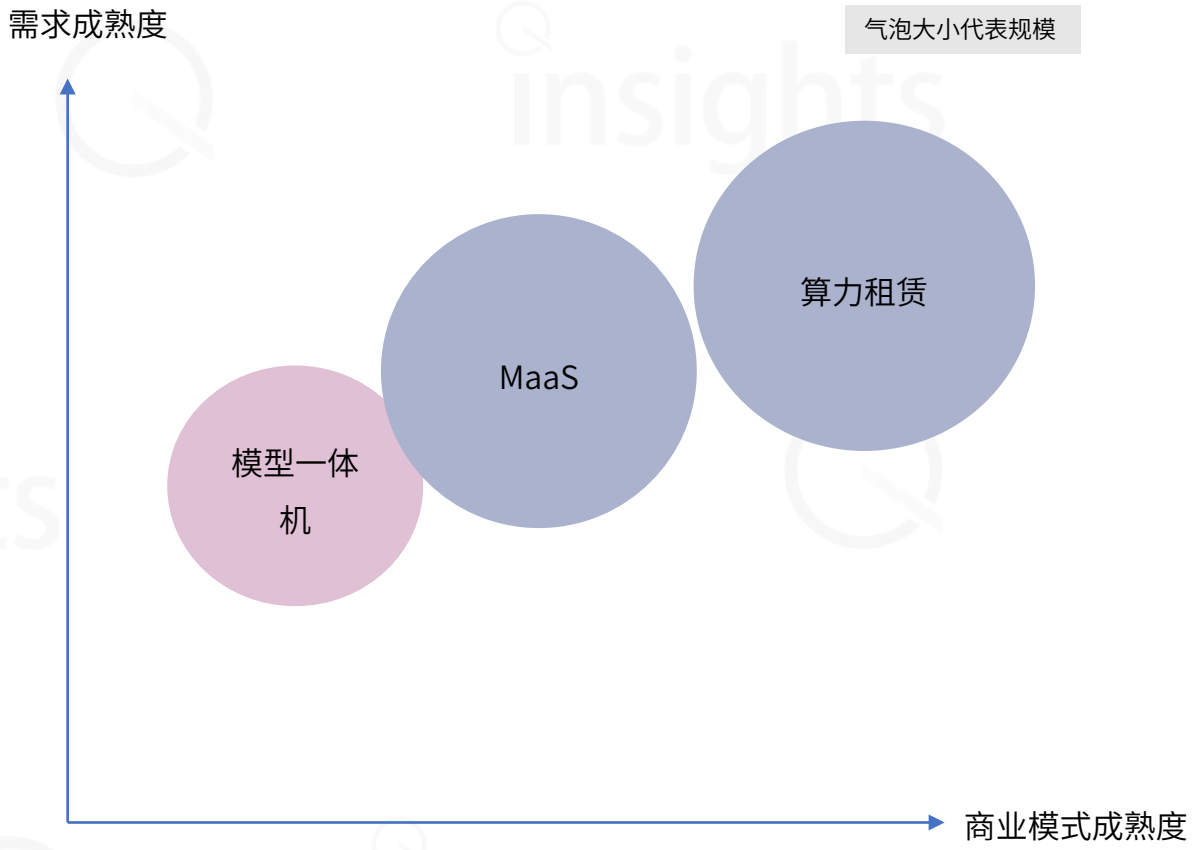
后期阶段（10年内）将在应用端呈现出百花齐放的趋势；

届时，AIGC应用将呈现出类app模式，出现各类细分领域的应用程序，通过模型层提供的API接口来发展各自的应用程序。此阶段，大众层面几乎不会感知到算力问题。

核心点

AIGC应用爆发，算力感知弱化，AIGC算力成为公共资源

- AIGC算力层的新商业模式趋于成熟，业务量与下游应用需求成正比
- 垂类模型使得算力租赁和MaaS具备更优的商业前景



后期阶段，算力租赁有望成为算力平台最具规模的盈利模式

05

AIGC算力行业案例集

阿里云 阿里巴巴

2022年，阿里云在国内首倡MaaS（Model as a Service，模型即服务）理念，提出以AI模型为核心的开发范式，并搭建了一套以AI模型为核心的云计算技术和服务架构，积累了丰富的大模型研发经验、工具和平台，这套能力将全部向大模型初创企业和开发者开放，提供包括模型训练、推理、部署、精调、测评、产品化落地等的全方位服务。

以模型为中心，打造MaaS平台服务

提供模型不同阶段需要的平台服务

模型社区

ModelScope

- 国内最活跃的模型社区，提供丰富的预训练SOTA模型、多元数据集和模型知识库
- 开源Python package，统一模型接入接口

模型开发平台

PAI 机器学习平台

- 交互式建模与可视化建模
- 支持万亿参数级模型训练
- 单任务集群规模可达万卡GPU

模型服务

DashScope 模型服务灵积

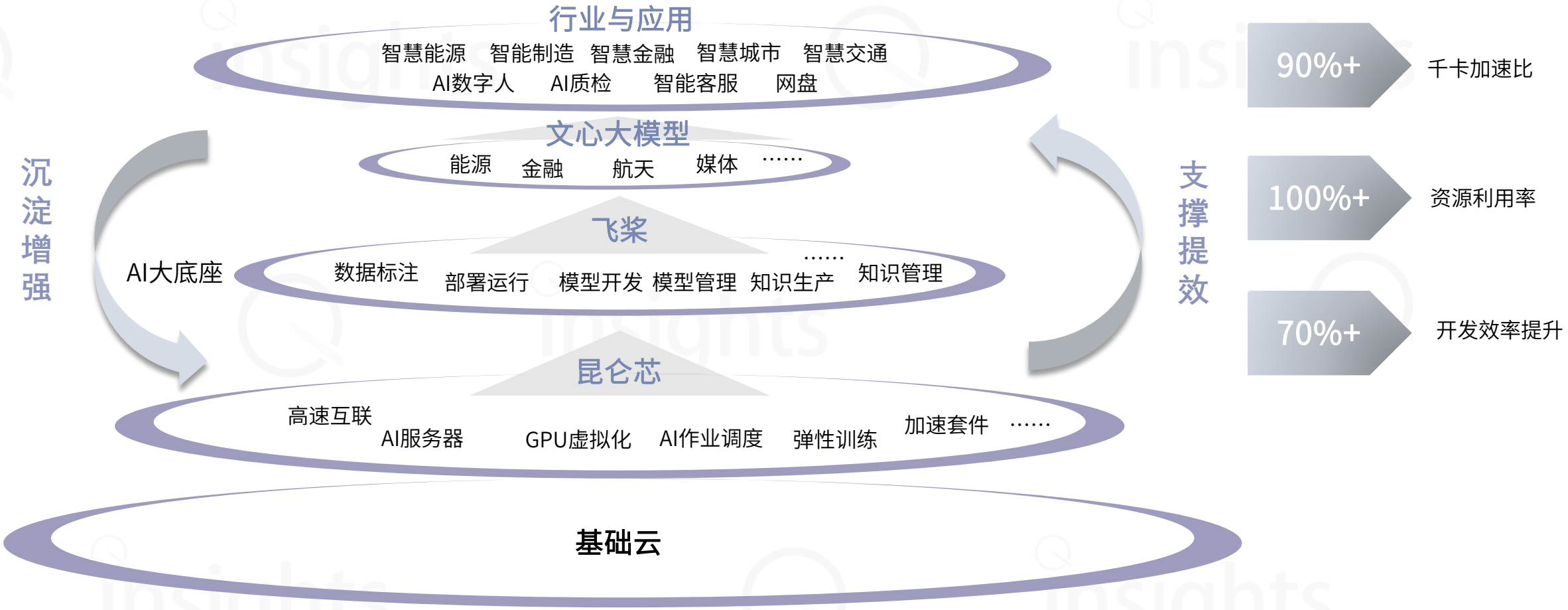
- 提供灵活、易用的模型API接口与SDK
- 自适应推理优化与高效微调训练
- 基于云底座的多区域弹性伸缩能力

依托于云，提供高效可靠的服务

阿里云底座

国内首个全栈自研的AI基础设施：百度智能云跨越芯片层、框架层、模型层、应用层四层，实现端到端的「云智一体」全栈AI设施，其中包含自研AI芯片昆仑，自研的深度学习框架和平台，自研的AI大底座，自研的大模型和深入千行百业的垂直行业应用。

端到端优化带来实际效果的显著提升：「云智一体」四层结构互相反馈和相互适配，全栈且深度融合带来的端到端优化，在大模型的训练和推理上均带来了更多的效果提升，具有显著优势。



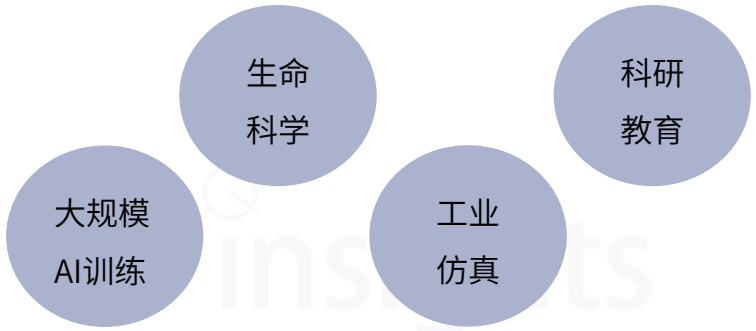
腾讯云新一代HCC（High-Performance Computing Cluster）高性能计算集群，采用腾讯云星海自研服务器，搭载英伟达最新代次H800 GPU，服务器之间采用业界最高的3.2T超高互联带宽，为大模型训练、自动驾驶、科学计算等提供高性能、高带宽和低延迟的集群算力。

腾讯高性能计算集群为MaaS提供充沛算力。

高性能计算集群

高性能计算集群（Tencent High-Performance Computing Cluster，THCC）以高性能云服务器为节点，通过RDMA（Remote Direct Memory Access）互联，提供了高带宽和极低延迟的网络服务，大幅提升网络性能，能满足大规模高性能计算、人工智能、大数据推荐等应用的并行计算需求。

应用场景



腾讯云MaaS全景图

MaaS			
客户专属大模型	细分领域模型训练平台	应用平台	智能应用
行业大模型精调解决方案	TI-OCR训练平台	媒体AI中台	数智人
	TI-AOI训练平台	智能视频分析平台	AI语音助手（车载/家居）
TI平台			
平台&工具			
TI-DataTruth 数据标平台	TI-ONE 训练平台		TI-Matrix 应用平台
太极Angel加速组件			
技术底座			
行业大模型			
金融大模型	政务大模型	文旅大模型	传媒大模型	教育大模型	……
基础设施			
高性能计算集群HCC	高性能网络：自研星脉计算网络架构		向量数据库

华为云 华为

算力是训练大模型的基础。华为在最底层构建了以鲲鹏和昇腾为基础的AI算力云平台，以及异构计算架构CANN、全场景AI框架昇思MindSpore，AI开发生产线ModelArts等，为大模型开发和运行提供分布式并行加速、算子和编译优化、集群级通信优化等关键能力。基于华为的AI根技术，大模型训练效能可以调优到业界主流GPU的1.1倍。昇腾AI云服务除了支持华为全场景AI框架昇思MindSpore外，还支持Pytorch、TensorFlow等主流AI框架。

昇腾聚焦AI基础软硬件，分层开放，促进行业智能升级

行业解决方案

昇腾训练解决方案　　昇腾推理解决方案

深度学习平台　讯飞火石平台　招行AI平台　星河AI平台　联通AI平台　……

MindX 昇腾应用使能　深度学习使能 | 智能边缘使能 | 优选模型库 | 行业SDK

AI框架　昇思MindSpore　PyTorch　TensorFlow　飞桨

CANN异构计算架构　Ascend C 编程语言 | 1400+高性能算子 | 6大算子库 | 基础加速库 | …

AI基础硬件　昇腾AI系列硬件

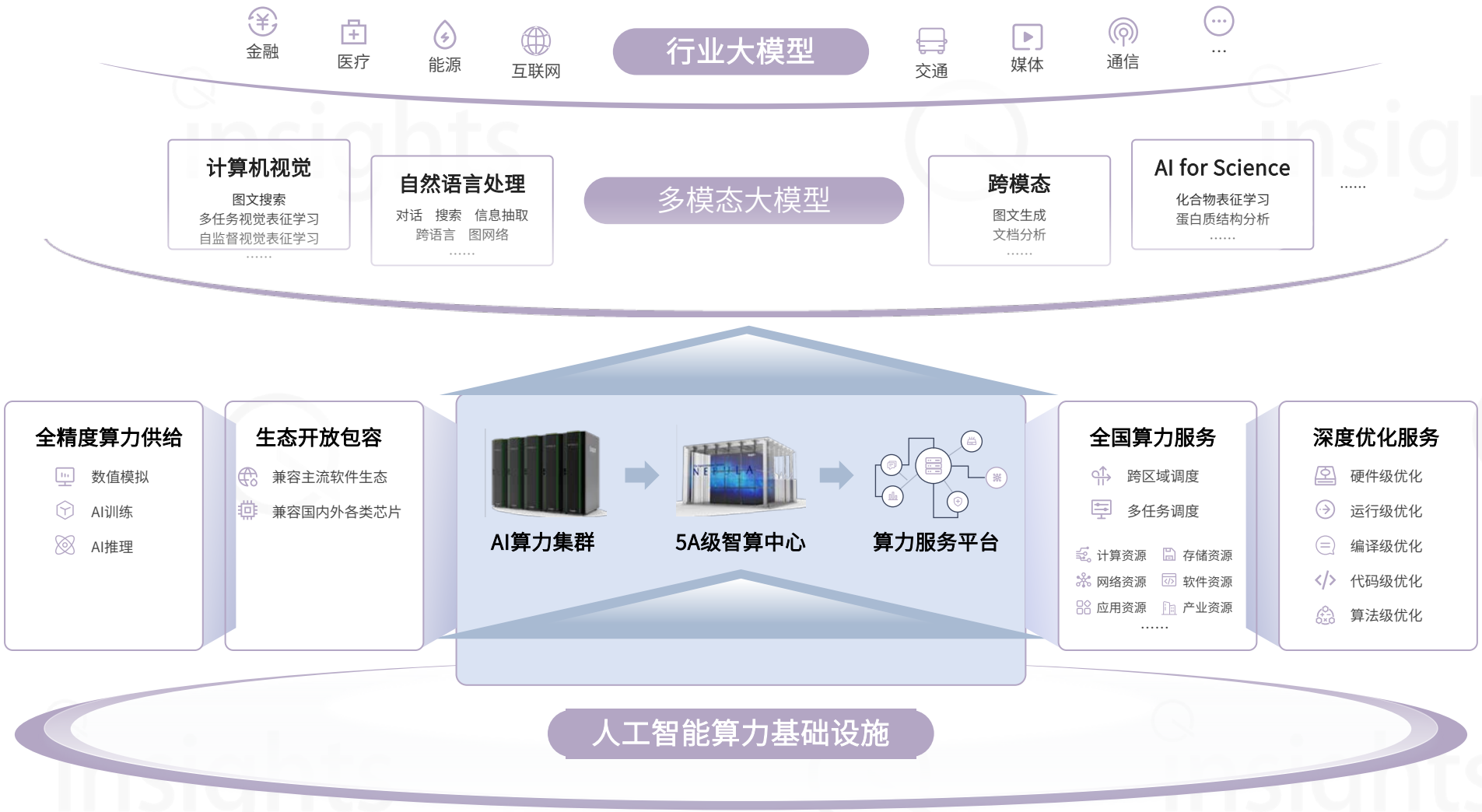
框架适配能力　插件化Adapter
支持PyTorch、TensorFlow、飞桨等业界框架
演进版本快速适配
3个月 → 1个月

动态Shape能力　二进制算子库
消除算子编译时间
性能满足场景需求
动态Shape算子满足度
70% → 95%
提升整网性能，并在CV、NLP等典型场景性能领先

算子开发能力　Ascend C编程语言
C/C++编程规范
结构化核函数编程方式
算子开发周期
2人月 → 2人周
已支持运营商、互联网等客户多个自定义大Kernel开发

中科曙光

中科曙光基于在智能计算领域的深耕，构建了完备的异构智能算力技术体系，实现了包含核心部件、整机、计算系统在内的诸多突破，打造了开放通用的算力底座。此外，还积极协同产学研用，孵化场景化解决方案，推动AI技术的实际应用和落地。



目前已完成

30+

国内外主流大模型适配及孵化

包含

GPT系列

LLaMA系列

GLM系列

ERNIE系列

……

同等条件下

大模型训练效率及训练稳定性

显著提升

墨芯人工智能

墨芯人工智能——稀疏计算引领者

随着AI大模型参数的日益攀升，稀疏计算已成为公认的AI发展必然趋势，从根本上解决大模型发展与算力的矛盾。

墨芯人工智能通过核心的稀疏计算技术建立起深厚优势，成为AIGC时代具有代表性的算力企业：率先基于原创的双稀疏算法，推出新一代AI计算平台，在算力、功耗、能效比等方面实现大幅优化，缓解大模型的高算力需求、高功耗、高费用等痛点，带来“多赢”的效果；并且在大模型算力的相关技术、产品、商业落地等方面，均已取得积极进展。

技术：独创双稀疏算法，并率先将稀疏化算法与硬件结合落地

- **推出全球首颗高倍率稀疏芯片Antoum®，支持高达32倍稀疏**:将此前的业界纪录提升16倍。

产品：屡获MLPerf冠军，性能位居行业领先

- 基于Antoum®芯片的墨芯AI计算卡产品，在国际权威基准测评MLPerf中**连续两届获得冠军**，并在MLPerf 3.0中**获得双料冠军**。

应用：支持千亿参数大模型，实现高吞吐、低延时，表现优异

- **在1300亿参数的GLM-130B大模型上，仅用8张墨芯S30计算卡，吞吐达432 token/s**，为AIGC大幅加速。
- **应用范围广**：支持 BLOOM、OPT、GPT-X、LLaMA、StableDiffusion等主流大模型。
- **高算力，低功耗，助力降本增效**：有效缓解AI企业的算力基础设施与运营成本高昂等难题，为企业拓展AIGC应用和业务提供强大算力支持。

商业落地：实现量产，多领域落地

- **产品已在互联网、交通、生命科学领域成单落地**：同时适用于运营商、金融、制造、医疗、能源、自动驾驶等众多行业与场景，获得市场认可。

全面赋能大模型行业落地与AIGC等应用

加速AIGC应用

文本生成 | 音频生成 | 图像生成 | 视频生成 | 跨模态生成 | 代码生成 ……

支撑大模型行业落地

互联网 | 运营商 | 生命科学 | 自动驾驶 | 金融 | 制造 | 物联网 | 智慧城市 ……

适用于各类型AI任务与模型

支持模型200+

视觉	语音	自然语言处理	多模态
ResNet-50, YOLO v3, YOLOv5...	FastSpeech, Conformer...	BLOOM, BERT,T5, GPT....	CLIP...

支持算子100+　算子涵盖大部分常见模型，并支持客户定制算子

全方位生态兼容

主流框架：ONNX | PyTorch | TensorFlow | 飞桨 ……

操作系统：支持Ubuntu CentOS等主流版本;欧拉、龙蜥、麒麟等国产操作系统

服务器平台：适配浪潮、新华三等主流服务器厂商X86 服务器平台，飞腾为代表的Arm服务器平台

软硬协同AI计算平台

Network Sparse Optimization	Sparse OPT		
SDK	SOLA Runtime	Sparse RT	Codec
Driver	Container Toolkit	Virtualization Toolbox	SOLA SEAL(KMD&UMD)

Antoum®芯片　AI计算卡系列　S4　S10　S30

天数智芯

天数智芯是中国领先的通用GPU高端芯片及超级算力系统提供商。

作为国内拥有云边协同、训推组合的完整通用算力系统全方案提供商，其系统架构、指令集、核心算子、软件栈均为自主研发，可独立发展演进。天数智芯已与国内重要行业合作伙伴携手，从源头对设计进行定义，率先实现大规模商业化量产，产品开发和商业应用进度领先国内同行。

通用GPU训练产品——天垓

支持200余种人工智能模型 —— 应用广泛 —— 开发易迁移 —— 适配主流CPU/服务器厂商

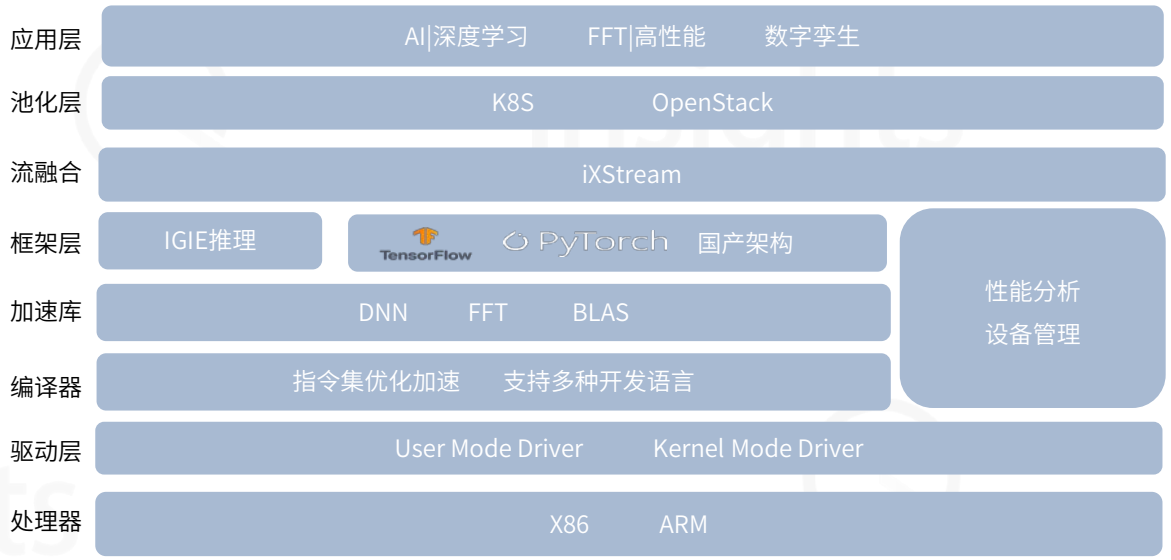
基于通用GPU架构 —— 性能可预期 —— 全栈可定制 —— 本地团队根据客户需求提供定制化服务

通用GPU推理产品——智铠

与市场主流产品相比，实现**2-3**倍实际使用性能提升

计算性能高 | 应用覆盖广 | 使用成本低

天数智芯软件栈



生态兼容

云平台	灵雀云、DAOClooud、联想云计算、iStation、天翼云、时速云、观测云
算法框架	PyTorch、TensorFlow、飞桨、计图、ONNX
OS	CentOS、ubuntu、银河麒麟、统信、OpenEuler、OpenAnolis
服务器	市场主流服务器
CPU芯片	市场主流CPU

摩尔线程

摩尔线程是一家以全功能GPU芯片设计为主的国家高新技术企业，能够为科技生态合作伙伴提供强大的计算加速能力，致力于打造为下一代互联网提供多元算力的元计算平台。

摩尔线程基于图形计算、人工智能计算，打造了第一款AIGC内容生成平台摩笔马良——支持中英双语，以及支持在简笔画、照片、真实画作、AI画作等多种模式输入下的图文生成、图文编辑，为用户提供零门槛的创作平台，促进用户自由创新表达。

基于AI+Graphics的智能内容创作（AIGC）平台

丰富功能	简单易用	软硬一体
中英文图文生成、图文编辑、图像超分、简笔画创作、视频生成等	零创作门槛	和MUSA架构深度集成优化

画作赏析

视拓云

视拓云团队的前身是中科院计算所山世光老师创建的中科视拓 SeeTaaS 部门，从2017年开始专注 C 端云计算市场和算法社区的研发和运营，面向“大 AI 圈”内的科研工作者和科技企业，运营 AI 算力服务平台 AutoDL.com 和算法社区 CodeWithGPU.com。

产品介绍

AutoDL.com是一个算力互联网平台，面向“大 AI 圈”内的科研工作者和科技企业，提供弹性、省钱、好用的普惠 AI 云算力服务。AutoDL 整合了来自全国各地的不同 IDC、运营商和云计算厂商的高性价比算力，共有超10个型号的7000多片 GPU 和国产 AI 加速芯片。自2021年10月公测至今，伴随着生成式 AI 和东数西算的红利，AutoDL 实现了快速增长，仅根据开票数据统计，AutoDL服务了130所985、211、双一流高校和300多所其他高校，超过400家企业，以及10万+个人开发者，是目前全国最大的 C 端 AI 算力入口之一。目前，AutoDL正在形成AI开发者社区CodeWithGPU.com 和以大模型训练/AIGC推理为特色的中立算力交易平台(整合国家东数西算算力节点)。

特点

1.**弹性**

弹性充分发挥容器相比虚拟机高效、性能损失小的优势，结合灵活弹性的按量计费方式、零成本升降配置等产品设定，使得AutoDL 在架构本身、计费方式功能上都极具弹性。

2.**省钱**

省钱基于 GPU 算力应用场景的特点，以高配置主机、独占算力提升核心体验，精简非必要组件、共享高成本资源降低服务成本使 AutoDL 成为好用不贵的GPU 云算力平台。

3.**好用**

站在科研工作者和科技企业的角度提供好用的功能，比如: 打通CodeWithGPU.com 使得便利的分享或使用镜像和模型; 团队子账号管理; 接口调用AIGC弹性部署等。在弹性、省钱的同时，还有很好的用户体验。

案例

曾服务某 AI 生成 LOGO 用户，在短短三天内提供了 800+ 卡3090+A5000 的混合 GPU 资源池，帮助用户成功应对了社区裂变带来的流量高峰，保证了终端客户的用户体验。

亿铸科技

亿铸科技致力于基于新型存储器 ReRAM 研发存算一体AI大算力芯片，是全球首家基于存算一体超异构创新架构，面向数据中心、云计算、自动驾驶、中心侧服务器等场景的 AI 大算力芯片公司。初代产品基于传统工艺制程，可实现500-1000T单卡算力。

亿铸科技

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

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

- 高精度
- 大算力
- 超高能效比
- 将存算一体在大算力真正落地

存算一体超异构系统级创新

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**

应用场景（部分）

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


智能数改


工业检测


安防

超分辨率


智慧交通

清微智能

清微智能是可重构计算（CGRA）领导企业，全球首家也是出货量最大的可重构计算芯片商用企业。核心团队来自于清华大学以及海思、英伟达、苹果、AMD等知名企业，专注于可重构计算芯片的创新研发和产业应用，面向云端训推一体，边端自动驾驶，安防监控等智能计算场景，提供高性能算力支持，致力于打造自主可控的可重构通用计算生态。

	TX5系列	TX8系列	TX2系列
产品简介	中算力CGRA 高性能端侧/边缘AI芯片	大算力CGRA 高性能云端训推一体芯片	小算力CGRA 高性能AIOT芯片
产品亮点	• CGRA可重构网络引擎，CGRA通用计算引擎 • 高能效比，图像处理性能超海思等同类芯片2-6倍，面积效率较国际顶级IP产品1.4-4倍 • 可重构ISP、CV、GPU、DSP	• 以国内相对成熟工艺实现国外先进工艺下顶级性能 • 同算力下功耗价格有数倍优势	CGRA可重构计算引擎，超低功耗可穿戴，能效比为传统音频DSP2-5倍以上
可重构技术验证	• 多核拓展技术 • 高能效通用处理	• 时空域数据流拓展技术 • 跨芯片边界互联	• 可重构编译技术 • 高能效AI处理

量子位 量子位智库 insights

关于量子位智库：

量子位旗下科技创新产业链接平台。致力于提供前沿科技和技术创新领域产学研体系化研究。

面向前沿AI&计算机，生物计算，量子技术及健康医疗等领域最新技术创新进展，提供系统化报告和认知。

通过媒体、社群和线下活动，基于专题技术报道及报告、专项交流会等形式，帮助决策者更早掌握创新风向。

关于量子位：

量子位（QbitAI），专注人工智能领域及前沿科技领域的产业服务平台。

全网订阅超过500万用户，在今日头条、知乎、百家号及各大科技信息平台量子位排名均为科技领域TOP10，内容每天可覆盖数百万人工智能、科技领域从业者。

微信号：Qbitbot020
量子位智库小助手