FROST & SULLIVAN
沙利文

头豹
LeadLeo

2025年中国AI基础设施市场报告

AI基础设施、网络安全、综合性计算环境

2025年8月

头豹研究院
弗若斯特沙利文咨询（中国）

报告说明

沙利文联合头豹研究院谨此发布中国生成式式AI系列报告之《2025年中国AI基础设施市场报告》。本报告旨在梳理中国AI基础设施的市场发展现状、用户核心需求以及相关技术洞察，明晰市场需求，并结合市场发展前景判断AI基础设施领域内各类竞争者所处地位。

沙利文联合头豹研究院对AI基础设施参与厂商进行调研。

本市场报告提供的AI基础设施发展趋势分析亦反映出AI基础设施行业整体的动向。报告最终对市场排名、领袖梯队的判断仅适用于本年度中国AI基础设施领域发展周期。

本报告所有图、表、文字中的数据均源自弗若斯特沙利文咨询（中国）及头豹研究院调查，数据均采用四舍五入，小数计一位。

头豹 LeadLeo

FROST & SULLIVAN 沙利文

研究框架

章节一
AI基础设施行业概况

- AI基础设施指为人工智能应用提供支持的硬件、软件和网络资源的集合，是以数据、算法和算力作为核心要素，支撑AI应用的研发、部署和运维的基础架构，确保数据处理、模型训练和智能决策等AI功能能够高安全地运行。
- 算力、算法与数据为支撑AI产业发展的三大核心因素，数据为模型训练与优化提供了海量输入资源；算法则依据海量数据中的有效信息进行分析预测，直接决定了AI系统的性能；算力的发展则为整体运行提供更高性能的计算能力以满足更复杂的模型优化与设计需求。
- 伴随着移动互联网和大数据技术的普及以及近年来AI能力的爆发式提升，AI技术在各行各业得到广泛应用，对高性能计算、大规模存储、高效算法库等基础设施需求不断增长。AI基础设施正式步入稳定与革新并存的应用期。
- 算法作为AI基础设施的关键要素之一，算法本身的发展和优化是推动AI基础设施发展的核心动力，同时也直接推动了AI芯片和服务器的需求增长。算法的发展离不开政府、学术机构、企业等各个层面对于学术研究、技术创新、人才教育等的高度重视。
- 数据资产拥有方作为AI基础设施的数据提供者，对于初始元数据的开发治理与资产化为AI基础建设的发展提供了燃料，而其余通信设备的制造生产以及能源配套则为AI基础设施的建设提供了保障。

1.1 行业发展背景

AI基础设施在产业AI化等需求升级的背景下，逐步迈向“绿色化、普惠化”新阶段。

《2025年中国AI基础设施报告》发布，为各行业企业在如何构建高性能、高资源利用率的AI基础设施提供指引，推动了行业内更高质量的市场竞争。

❑ AI应用开发需求的爆发式增长，推动算力需求的飙升

2024年后半年开始，AI应用开发需求呈现爆发式增长，应用端生成能力的提升与Agent的发展，驱动更为复杂和精细的新一代AI基础设施应运而生。受到AI应用与各领域业务结合、Agent的推出、多模态渗透等因素，AI基础设施中算力需求飙升，其中AI算力需求从以往的训练端转向推理端。

❑ 头部厂商领先加码基础设施，国内算力行业再次“狂飙突进”

2025年阿里云宣布未来三年将投入超过3800亿元，用于建设云和AI硬件基础设施，创下中国民营企业在云和AI硬件基础设施领域有史以来最大规模的投资记录。

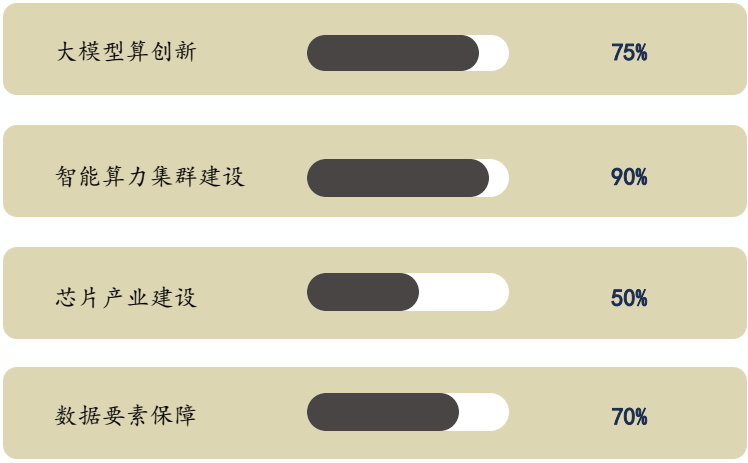
腾讯云于今年3月上海峰会中宣布将持续加大海外基础设施投入，计划在沙特建设首个中东数据中心以及印尼的第三个数据中心。

商汤科技在AI算力上采用了差异化策略，在AI通用算力外，为具身智能、AIGC及传统企业智能化升级等重点方向提供最优算力方案。截至2025年3月，商汤科技运营总算力突破2.3万petaFLOPS。

❑ 顶层设计强化+地方政策落地，共同助推新基建布局

国家政府层面以“东数西算”战略为基调，以《数字中国建设计划》等政策为顶层设计奠定基础，重点推动算力资源全国协同调度。地方政府积极把握发展机遇，通过营造良好的营商环境，大力支持人工智能产业的发展。北京、上海、广州等地政府通过建立产业集聚区、提供人才优惠等措施，积极打造人工智能产业生态。

各省市关于AI基础设施政策推动重点

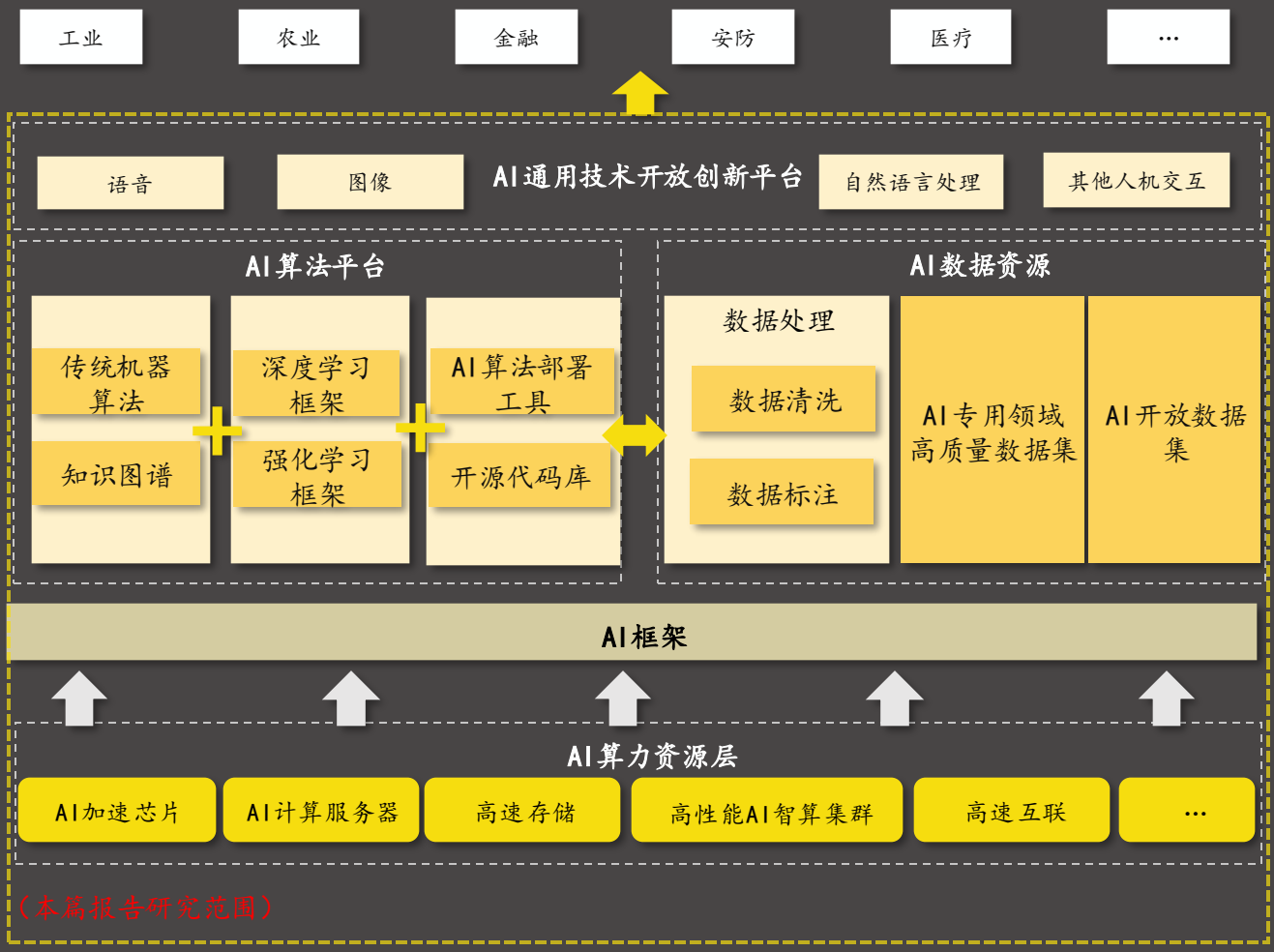


1.1 AI基础设施定义与研究范围

关键发现

AI基础设施指以高质量网络为重要支持，以数据资源、算法框架和算力资源为核心要素，用于支撑AI应用的设计、部署和运行的基础架构，也是确保系统可以正常运行并处理庞大数据和复杂计算任务的基石。

AI基础设施整体视图



❑ **AI基础设施指以高质量网络为重要支持，以数据资源、算法框架和算力资源为核心要素，用于支撑AI应用的设计、部署和运行的基础架构。**

- AI基础设施是整个AI系统的底层部分，是确保系统可以正常运行并处理庞大数据和复杂计算任务的基石。AI基础设施集成了算力、算法和数据三大资源要素，以AI硬件设施、算法平台、数据平台以及开放创新平台等为主要载体。随着人工智能的融合发展，AI基础设施将推动人工智能与5G、云计算等领域的相互耦合，加速人工智能与实体经济的深度融合，形成新一代信息基础设施赋能产业及各行业应用的核心能力。

1.3 AI基础设施核心要素

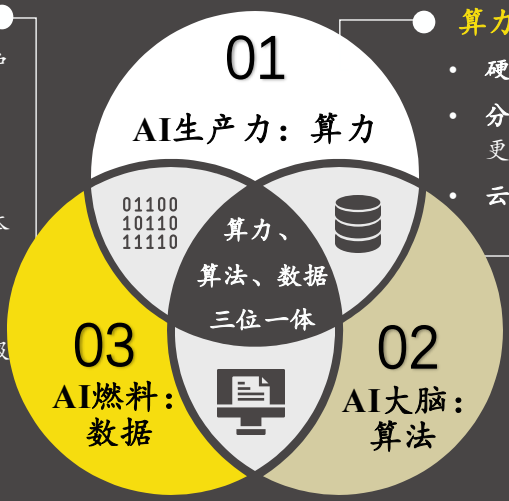
关键发现

算力、算法与数据为支撑AI产业发展的三大核心因素，数据为模型训练与优化提供了海量输入资源；算法则依据海量数据中的有效信息进行分析预测，直接决定了AI系统的性能；算力的发展则为整体运行提供更高性能的计算能力以满足更复杂的模型优化与设计需求。

AI基础设施三要素

资源突破：大数据

- **高质量数据收集**：传感器、用户输入、在线行为
- **数据预处理**：数据清洗、去噪、特征工程
- 数据标注：图像目标框选、文本分类
- **数据存储**：海量并发数据存储
- 数据安全与隐私：数据分类分级



算力突破：AI芯片

- **硬件加速**：GPU、CPU、FPGA等
- **分布式计算**：处理更大的数据集与更复杂的模型
- **云服务**：更灵活的算力资源

算法突破：深度学习

- **深度学习**：卷积神经网络、循环神经网络等
- **机器学习**：监督学习、无监督学习等
- **算法优化**：用于调整模型参数以最小化损失函数

❑ **算力、算法与数据为支撑人工智能产业发展的三大核心要素，被称为AI燃料的数据为模型训练与应用落地奠定基础；作为AI大脑的算法则是指引数据处理和决策制定的核心逻辑，直接决定了AI系统的性能与智能化水平；最后，高性能的计算能力为AI计算提供了强有力的支持。**

- **推动AI落地发展的核心基础-数据**：人工智能在各行业的大规模应用需要利用到海量数据对模型进行训练，数据在整个AI系统中扮演着至关重要的角色，它不仅是模型训练和学习的基础，也是AI系统不断优化的源泉。为提高数据的质量与规模，部分企业、高校与开源社区等主体选择开放数据集，加之全球数据量的爆炸式增长，为AI算法提供了广阔的学习与优化空间。

- **驱动AI创新发展的重要引擎-算法**：作为人工智能的核心逻辑，算法是产生人工智能的直接工具，优秀的算法可以高效地从海量数据中提取有价值的信息，并据此进行智能分析和预测。可以说算法的优化与创新直接决定了AI系统的性能与智能化水平。一方面，AI算法的发展推动计算机视觉、智能语音、自然语言处理等技术突破商用门槛，实现大规模应用；另一方面，以开源框架为核心的技术研发生态和以开放平台为核心的行业应用生态已经成为人工智能发展与应用的重要基础，并逐渐成为AI新基建的重要发力方向。

- **支撑AI高速发展的关键因素-算力**：AI算力包括AI芯片、AI计算服务器、高性能计算集群等，为人工智能技术和产业发展提供了强有力的算力支撑。当前，以深度学习为代表的的人工智能技术需对海量数据进行处理和训练，对算力提出了较高的要求。近年来，算力相关技术发展迅速，云计算平台通过弹性可扩展的计算资源降低了AI应用的门槛和成本；边缘计算则将计算任务推向数据源头，减少了数据传输延迟；GPU与TPU等专用加速芯片的出现也极大提升了AI计算的效率与速度。

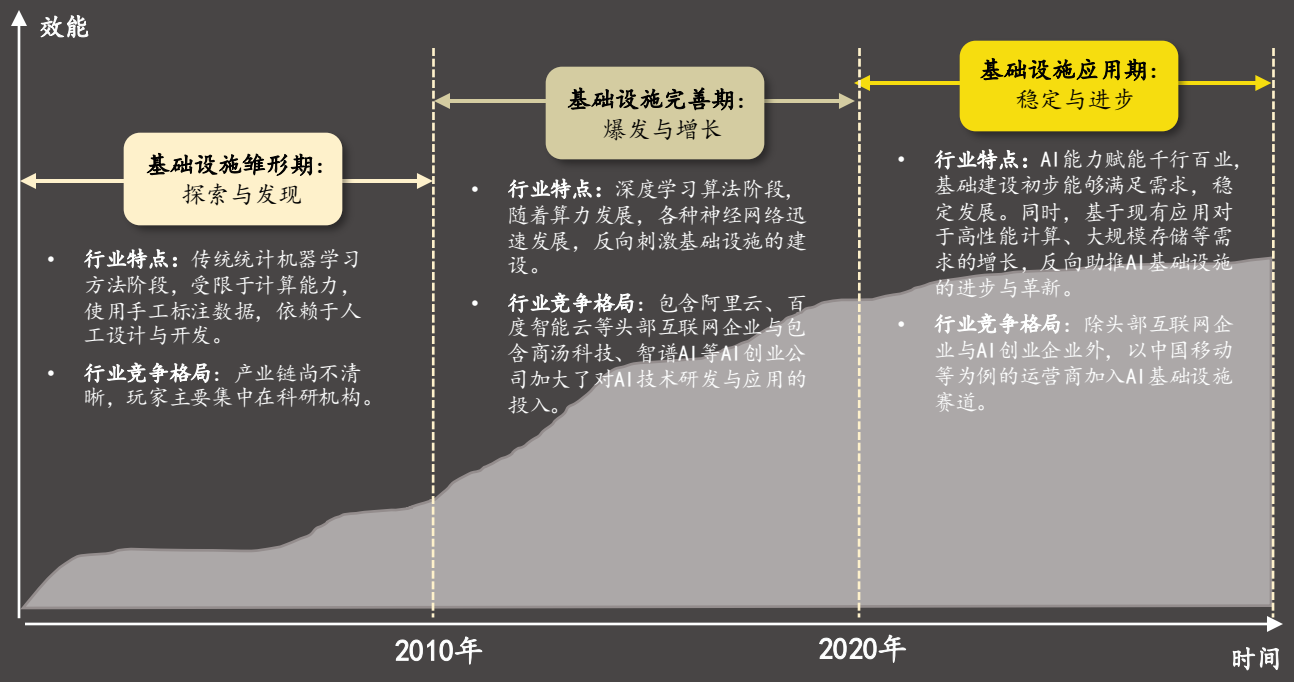
来源：阿里云、沙利文

1.4 AI基础设施发展历程

关键发现

伴随着移动互联网和大数据技术的普及以及近年来AI能力的爆发式提升，AI技术在各行各业得到广泛应用，对高性能计算、大规模存储、高效算法库等基础设施需求不断增长。AI基础设施正式步入稳定与革新并存的应用期。

AI基础设施发展历程及效能曲线



❑ **智能化转型趋势下，AI与各行业场景深度融合，企业用户在AI部署过程中对数据质量、计算能力等提出了更高的要求，AI基础设施行业进入稳定发展与进步革新并存的应用期。**

- **雏形期（2010年以前）**：该时期，AI基础设施处于探索与发现阶段，人工智能领域处于低谷时期，AI项目多数处于实验性阶段。因此，AI基础设施生态系统在该阶段尚不清晰，专业的人才与资源缺失，主导AI项目的玩家主要集中于科研机构。

- **完善期（2010-2019年）**：随着机器学习和深度学习等一系列AI技术的关键突破，AI基础设施行业开始迈入爆发与增长阶段。一方面，AI技术的进步，互联网发展带来的大数据时代以及AI开源社区的建立，均为AI技术的发展与传播提供了有力的支撑条件，进而对底层基础设施的技术发展产生积极性影响。另一方面，AI基础设施的商业模式逐渐清晰，头部互联网企业与部分AI创业企业逐步认知到AI的潜力，纷纷加入该赛道，加大了对AI基础设施的资源投入。

- **应用期（2020年-至今）**：2020年后，各行各业开始大规模应用深度学习技术实施创新应用，加快产业转型和升级。随着AI与各场景的深度融合与发展，计算能力、分布式系统和大规模数据处理能力等新能力反向助推了AI基础设施的创新与优化 。此外，随着5G与AI的融合发展趋势，以中国移动等为例的运营商也在该阶段进入赛道，开始积极布局包含AI基础设施新基建、平台能力新基建与云网新基建等。

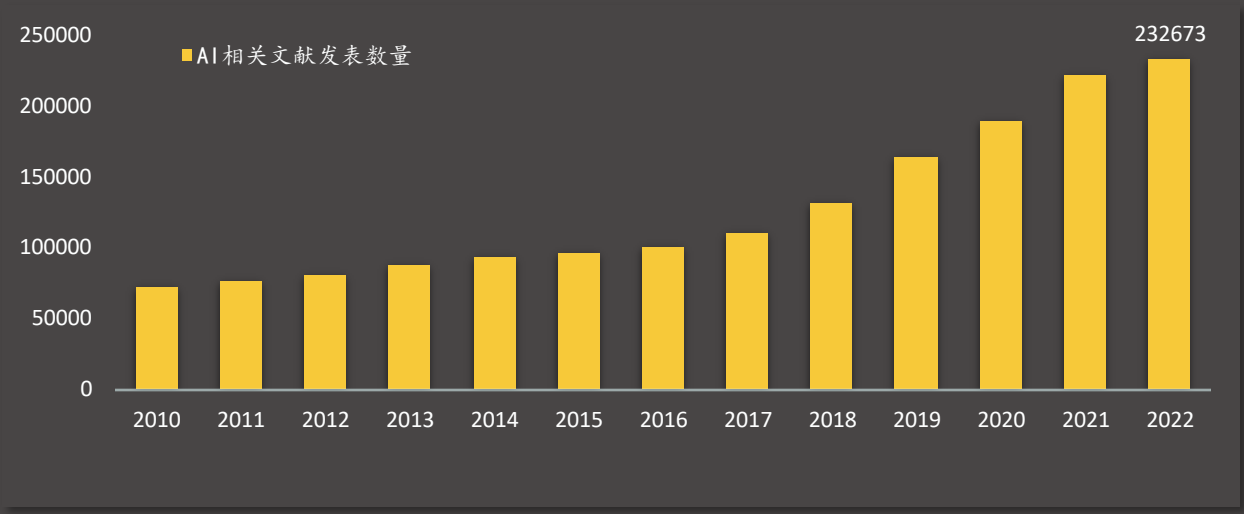
来源：沙利文

1.5 AI基础设施驱动因素 - 技术创新

关键发现

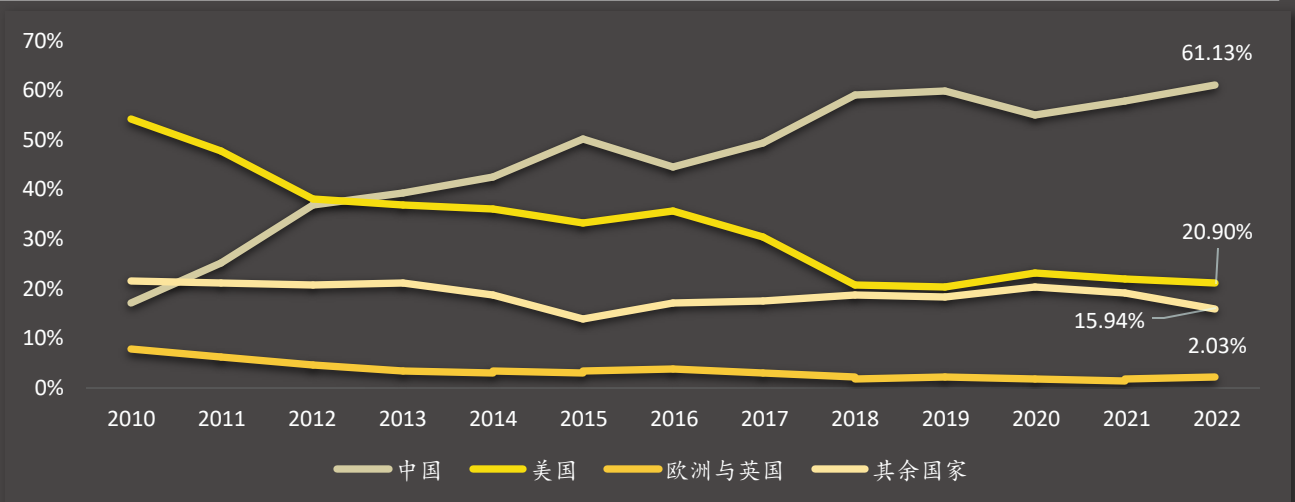
算法作为AI基础设施的关键要素之一，算法本身的发展和优化是推动AI基础设施发展的核心动力，同时也直接推动了AI芯片和服务器的需求增长。算法的发展离不开政府、学术机构、企业等各个层面对于学术研究、技术创新、人才教育等的高度重视。

全球AI相关文献发表，2010-2022年



- 2010至2022年期间，全球共发表了***165.8万***篇关于人工智能的学术论文，增长率高达***129%***。其中，中国作者发表的相关论文数量位居全球第一，且每篇论文的年均引用次数也位居全球前列。中国科学院的AI研究总量以及高引论文数量，均位居世界榜首。
- 此外，中国AI相关的专利技术占世界百分比稳步增长，于2022年达到***61.13%***

全球主要国家的人工智能专利（占世界总数的百分比）



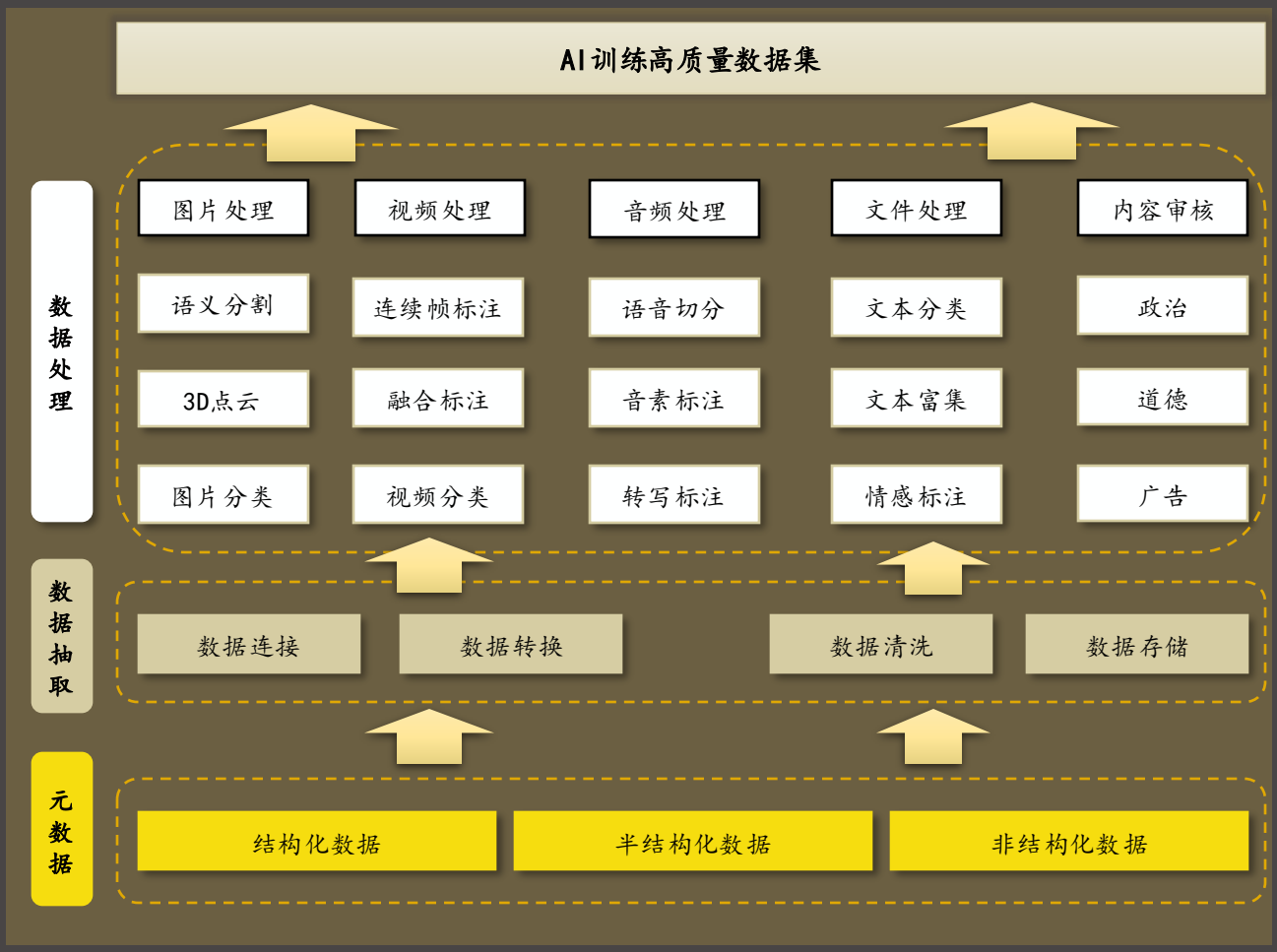
来源：斯坦福大学、沙利文

1.5 AI基础设施驱动因素 - 高质量数据资产

关键发现

数据资产拥有方作为AI基础设施的数据提供者，对于初始元数据的开发治理与资产化为AI基础建设的发展提供了燃料，而其余通信设备的制造生产以及能源配套则为AI基础设施的建设提供了保障。

元数据开发与治理图谱



❑ 人工智能的大规模应用需要高质量的数据集

- 从信息技术革命爆发开始，到近年来移动互联网的蓬勃发展，同时伴随着互联网+概念的不断深入演进和落地，各行各业的企业积累了越来越多的数据，相应的人工智能也需要大量的数据来教授和培训，然而并非任何数据都可以作为AI基础设施并投喂给AI模型，高质量大规模的训练数据集才是深度学习进行模型训练的关键，因此大量的元数据在产生之后需要进行开发治理与存储。

章节二
AI基础设施产业链分析

- 产业链图谱
- 产业链上游分析
- 产业链中游分析
- 产业链下游分析

- AI基础设施行业产业链上游为各类硬件原材料供应商、数据集供应商以及网络服务；中游包括服务于各行业场景企业用户与消费级终端用户的云计算厂商、AI原生厂商、硬件系统厂商以及运营商；下游各行业应用中，互联网等数字原生行业为应用进展先行者。
- AI芯片作为人工智能基础设施的算力核心支撑，其重要性日益凸显。随着近期美国对中国AI芯片的进一步封锁措施，给中国获取先进AI芯片带来了短期困难。面对这一挑战，中国AI芯片的后续发展将依赖于国产替代进程的推进。
- 云计算、AI原生、硬件系统与运营商四类主要厂商塑造AI基础设施市场格局，提供全栈服务能力的云计算、提供垂直一体化解决方案的AI原生、提供高效硬件支持的硬件系统厂商与整合网络资源的运营商共同推动AI基础设施行业生态完善与技术演进。
- AI基础设施商业模式根据资产控制权等因素可分为重资产、轻资产以及混合模式三种主要形式。其中混合模式核心为“核心自控+外部弹性扩展”与“软硬结合”，凭借其灵活性以及复杂场景的高适配性成为多数企业的现实选择。
- 中国AI基础设施渗透率较高的行业包含互联网、教育、金融与医疗等具备丰富数据资源的行业，其中安全隐私性为下游用户的首要考量因素，尤其在金融与医疗监管严格的行业。此外，高性能与可扩展性也为互联网与制造业等应用场景复杂的主要考量。

2.1 中国AI基础设施产业链图谱

关键发现

AI基础设施行业产业链上游为各类硬件原材料供应商、数据集供应商以及网络服务；中游包括服务于各行业场景企业用户与消费级终端用户的云计算厂商、AI原生厂商、硬件系统厂商以及运营商；下游各行业应用中，互联网等数字原生行业为应用进展先行者。

AI基础设施产业链图谱

来源：智算联盟、商汤科技、沙利文

2.2 产业链上游-AI芯片国产替代进程加速

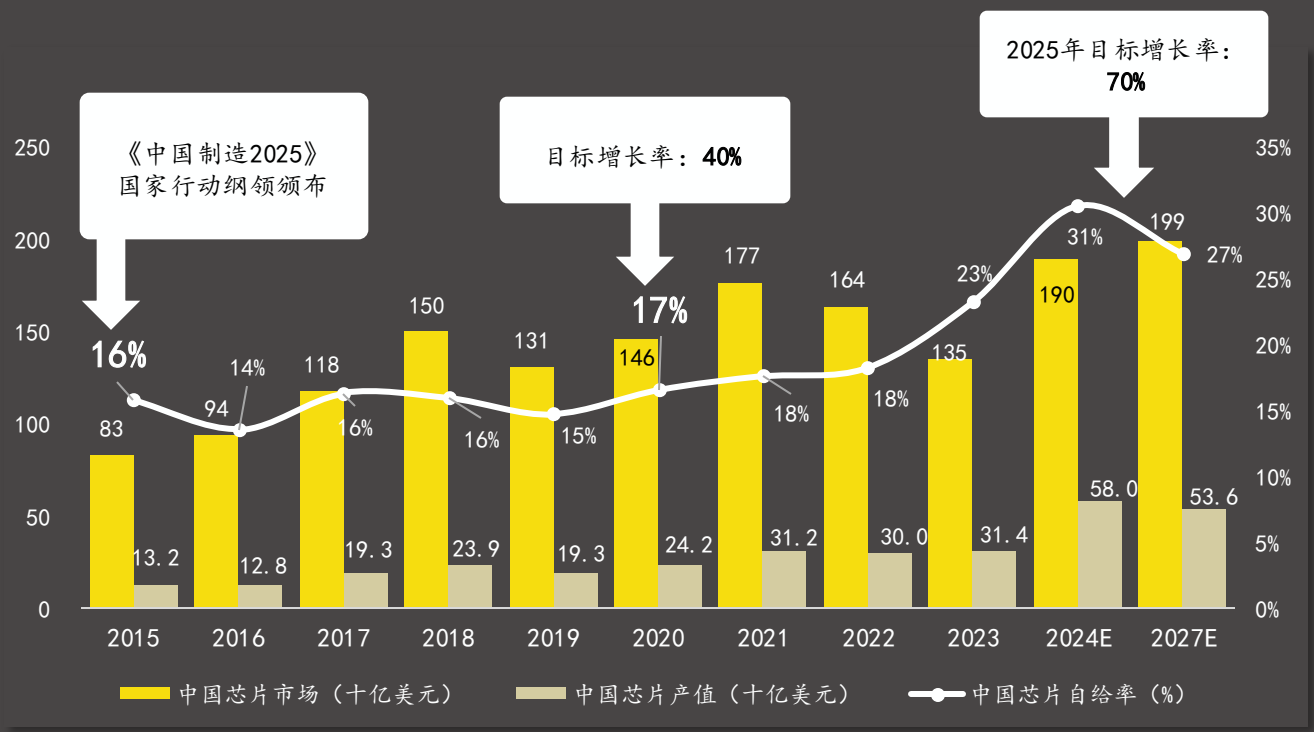
关键发现

AI芯片作为人工智能基础设施的算力核心支撑，其重要性日益凸显。随着近期美国对中国AI芯片的进一步封锁措施，给中国获取先进AI芯片带来了短期困难。面对这一挑战，中国AI芯片的后续发展将依赖于国产替代进程的推进。

中国芯片自给率，2015年 - 2027E

中国芯片自给率：

2015年：16%；**2020年**：17%；**2023年**：23%



❑ **AI芯片作为AI产业链的基础层，直接决定了AI系统的计算能力与效率，是上游的核心领域。随着2025年美国对于AI芯片限制的进一步加强与一系列国产芯片支持的政策推出，AI芯片国产替代进程加速，中国芯片自给率稳步提升，但仍亟待提高。**

- 中美科技博弈持续演绎的背景下，作为核心基础层的AI芯片政策战略地位高，积极导向的政策颁布与有力的资金支持，助推芯片国产化率提升。长期看，一方面，中国企业将加大对芯片制造技术的研发投入，努力提升自身的制造能力和技术水平，减少对海外代工厂的依赖；另一方面，国内的芯片制造设备和材料供应商也将获得更多的发展机会，推动整个芯片制造产业链的国产化进程。根据Techsights数据，中国芯片自给率呈现稳定增长态势，从2015年的***16%***增长至2023年的***23%***。此外，在2020年国务院颁布的《新时期促进集成电路产业和软件产业高质量发展的若干政策》中，明确提出2025年芯片自给率目标为***70%***，国产芯片未来增长空间可观。
- 然而，高端芯片的技术壁垒急需攻破，英伟达等国外厂商在高端芯片技术上仍保持领先地位。将国产主流AI芯片与英伟达产品对比，可发现部分运算性能仍存在较大差距，例如H100芯片采用的4nm工艺国内尚未实现。国产芯片在运算性能、市场份额与生态建设仍需逐步提升。

来源：Techsights、沙利文

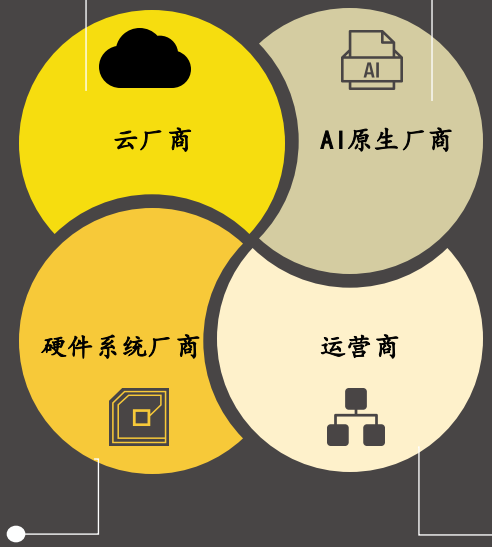
2.3 产业链中游 – 厂商类型（1/2）

关键发现

云计算、AI原生、硬件系统与运营商四类主要厂商塑造AI基础设施市场格局，提供全栈服务能力的云计算、提供垂直一体化解决方案的AI原生、提供高效硬件支持的硬件系统厂商与整合网络资源的运营商共同推动AI基础设施行业生态完善与技术演进。

AI基础设施厂商类型及特点梳理

- **厂商特点：**加强MaaS服务，为开发者和企业提供了构建、运行和扩展AI应用所需的基础设施和服务，尤其是在算法框架研发等方面。
- **优势：**成熟的云计算相关产品，提供全栈服务能力
- **劣势：**趋于平台化与标准化，对于定制化需求响应灵活度较低
- **代表厂商：**阿里云、腾讯云、百度智能云等



- **厂商特点：**着重算力层面的建设。
- **优势：**较早开始布局AI基础设施，具备一定的先发优势，且拥有丰富的AI落地经验与深度AI技术积累。
- **劣势：**需要加强自有算力的投资与建设。
- **代表厂商：**商汤科技、第四范式、科大讯飞等

- **厂商特点：**重点布局AI相关的产品技术与应用服务
- **优势：**领先的算力硬件设施产品，在算力层拥有深厚的技术经验
- **劣势：**产品矩阵有限
- **代表厂商：**华为、寒武纪、浪潮等

- **厂商特点：**重点与各类AI算法企业、集成商等合作共同服务客户
- **优势：**国家网络资源的整合度高，符合高标准的安全合规性，政企行业覆盖广泛
- **劣势：**AI相关的技术产品建设滞后
- **代表厂商：**联通数智、天翼云、移动云等

❑ **AI基础设施厂商指为AI技术开发、训练和部署等提供底层硬件、软件以及平台的企业，目前主流AI基础设施厂商可分为云厂商、AI原生、硬件系统厂商以及运营商。**

- **云计算厂商：**头部云厂商具备完善的全栈服务能力、庞大的开发者生态与客户基础以及丰富的C端产品经验与技术。以阿里云、腾讯云为例的该类厂商在技术创新、算法研发与平台支持层面起到关键作用。
- **AI原生厂商：**较早开始布局AI技术与产品，具备深厚的AI算法沉淀。以商汤科技为例的该类厂商在端到端的垂直一体化解决方案以及长尾场景下的定制化能力具备优势。
- **硬件系统厂商：**专注于开发和提供以GPU、服务器等处理AI任务的硬件设备以提升AI任务处理效率。以华为等为例的该类厂商具备领先的硬件设计能力，在为各类AI任务提供高效计算能力层面至关重要。
- **运营商：**覆盖广泛的网络资源与政企相关渠道，提供符合最高标准的安全合规性产品。以中国移动等为例的该类厂商通过与AI产业链其他玩家合作共同服务企业客户，提供“AI+DICT”解决方案。

来源：商汤科技、沙利文

2.3 产业链中游 – 商业模式（2/2）

关键发现

AI基础设施商业模式根据资产控制权等因素可分为重资产、轻资产以及混合模式三种主要形式。其中混合模式核心为“核心自控+外部弹性扩展”与“软硬结合”，凭借其灵活性以及复杂场景的高适配性成为多数企业的现实选择。

AI基础设施厂商商业模式

商业模式	主要覆盖厂商	主要覆盖客群	优势	挑战
重资产：企业用户从硬件到软件系统均由企业用户自建	• 头部云计算厂商 • 运营商 • 算力硬件厂商	• 互联网大厂 • 科研机构与高校 • …	• 用户自主掌控 • 深度定制化 • 安全隐私防护能力 • 高性能、大规模的算力输出	• 初始投资成本高 • 技术迭代风险 • 能耗与运维成本高
轻资产：企业用户通过与第三方合作建设AI基础设施，或租赁、购买第三方提供的AI算力服务	• Maas供应商 • …	• 中小型创业企业 • 数字化转型传统企业 • …	• 更快的部署速度 • 灵活性高 • 运维负担小	• 不适于用于算力需求庞大的企业用户 • 定制化程度受限 • 迁移与兼容挑战 • 底层硬件依赖性
混合模式：企业用户灵活利用云服务或合作方资源进行补充和扩展	• 云计算厂商 • AI原生厂商 • …	• 中小微商业体 • …	• 可在安全合规、成本控制与定制化需求中取得最佳平衡点	• 成本控制的精确计算和预测 • 不同环境下的数据安全 • 对企业用户的技术能力与资源正能力要求较高

❑ **AI基础设商业模式根据资产所有权与控制权可分为重资产、轻资产以及混合模式三种主要模式。其中混合模式凭借“核心自控+外部弹性扩展”的灵活性优势，符合大多数企业和场景的现实选择，愈发成为明显的未来趋势。**

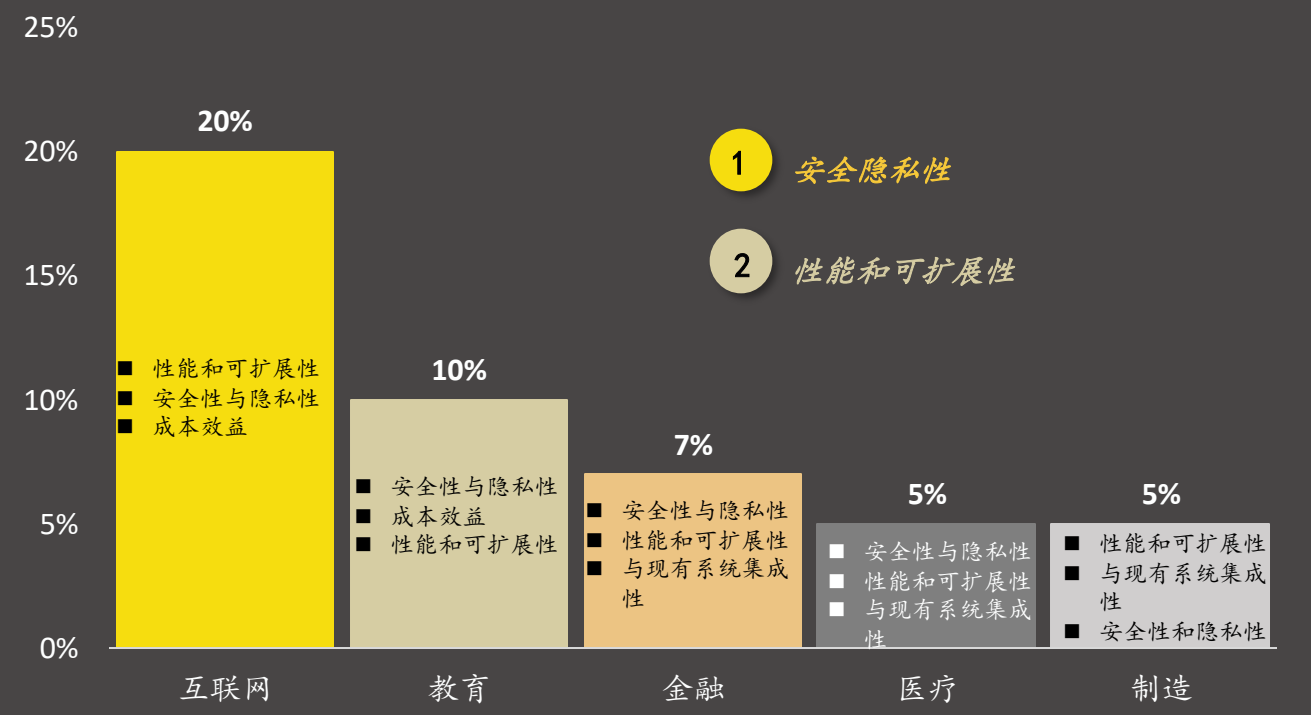
- **重资产**：该模式的核心为规模化的硬件投入和基础设施建设，为企业客户提供稳定且高性能的算力资源。然后完全自建算力中心对多数企业而言经济负担大且具备较高的技术迭代风险和运维难度。
- **轻资产**：该模式的核心在于通过软件、平台等自身不持有大量物理资产，更聚焦于软件的敏捷开发和生态构建。然而对于企业用户的数据控制和管理能力相对较弱且对厂商具备较高的依赖性。
- **混合模式**：核心在于结合前两种模式的优势，既拥有部分核心资产以保证关键服务的可控性，又通过合作、租赁等轻量合作模式保持其灵活性，加强了企业用户对复杂场景需求的适应性。

2.4 产业链下游 – 应用情况

关键发现

中国AI基础设施渗透率较高的行业包含互联网、教育、金融与医疗等具备丰富数据资源的行业，其中安全隐私性为下游用户的首要考量因素，尤其在金融与医疗监管严格的行业。此外，高性能与可扩展性也为互联网与制造业等应用场景复杂的主要考量。

中国AI基础设施主要渗透行业及考量因素



❑ **按照企业用户垂直行业划分，AI基础设施渗透率较高的行业包括互联网、教育、金融、医疗与制造。根据主要渗透行业用户的考量因素排序，安全与隐私性为多数行业的首要条件，尤其在金融与医疗等监管环境严苛的行业中至关重要。**

- **应用覆盖行业：**AI基础设施在互联网行业渗透率最高，达到近20%，其主要原因为互联网丰富的数据资源属性以及对于业务创新、运营优化等需求驱动，此外互联网先天的数字属性将其接入AI基础设施的摩擦成本大幅降低。其中数据资源为影响渗透率的关键因素，除互联网外，在教育、金融与医疗等行业，通常累积了大量的数据资源为AI基础设施的应用建设提供了充分的训练数据，使得模型可生成更精准且具备针对性的解决方案。

- **应用行业考量因素：**在渗透率较高的行业中，安全隐私性成为了诸多用户的首要考虑条件，尤其在金融与医疗等监管严格的行业，大量的数据（包括交易数据、病患信息等）需得到妥善的保护，因此该类行业用户对于AI基础设施的风险实时监测等能力有较大需求。而互联网与制造业由于场景复杂，对于高性能与可扩展性有着迫切需求。此外，不同的解决方案部署方式以及AI应用成熟度也对各行业优先考量条件产生一定影响。

来源：微软、沙利文

章节三
AI基础设施关键技术突破点

- 高通量网络技术
- 高性能存储能力
- 资源弹性调度能力

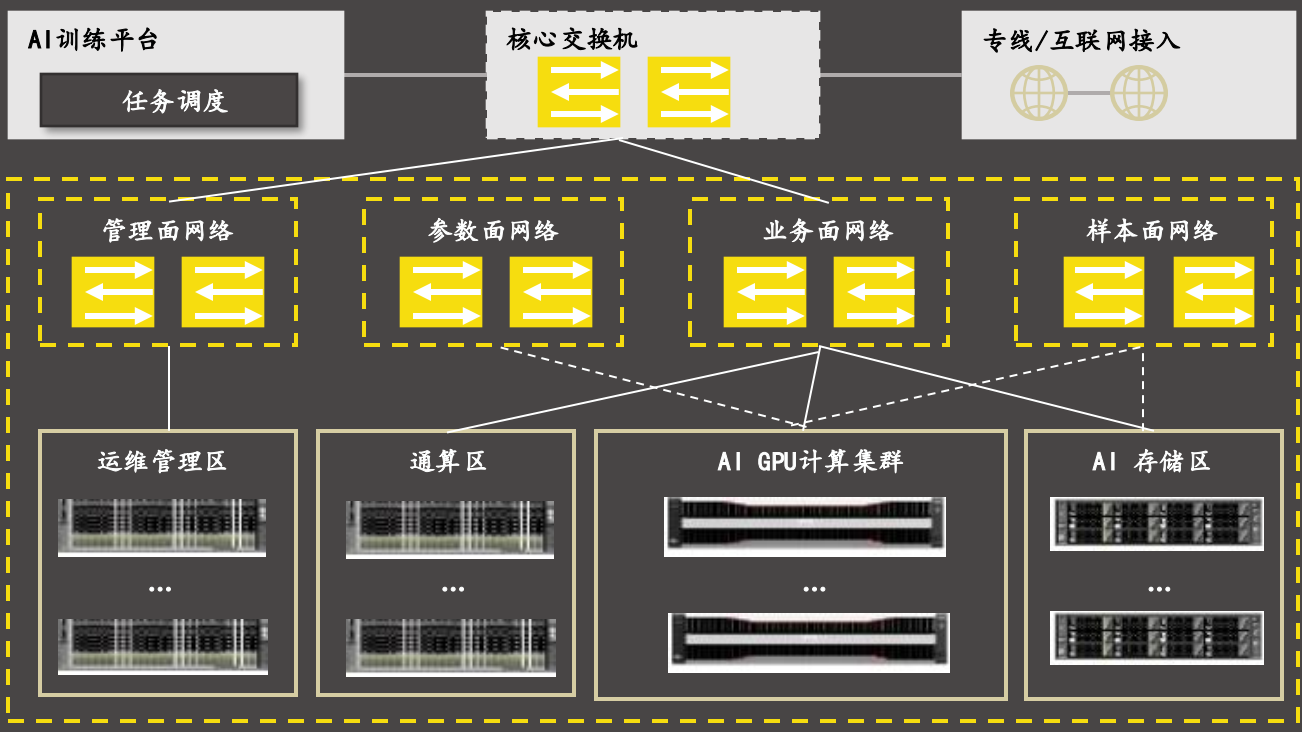
- 高通量网络技术是应对目前AI计算规模急剧扩张的关键突破口，其直接影响了大规模集群的协同效率和整体算力输出。随着国内高通量以太网联盟协议标准的发布，标志着以高通量以太网为代表的新一代开放智算网络，正加速迈向产业化落地。但协议的标准化与运维成本仍是主要面临难点。
- 高性能存储是解决数据归集时间长、数据处理效率低等问题的核心环节。在训练侧，高性能存储可显著降低GPU空耗；在推理侧，“以存代算”通过SSD作为分级缓存承载KV-cache，有效缓解显存瓶颈。目前，全球各大厂商均在积极研发适用于AI工作负载的高性能存储解决方案，加速布局AI SSD产品线。但性能、能耗与成本问题仍有待解决。
- 高效的算力调度能力一方面有助于提升集群资源利用率，直接降低企业用户的运用成本；另一方面，AI应用存在显著的波峰波谷和突发请求，弹性调度可基于业务需求，实现资源的动态分配，保障AI服务的稳定性与连续性。但现行的调度机制往往无法高效利用计算资源，无法根据实时需求动态调整配置。

3.1 高通量网络技术

关键发现

高通量网络技术是应对目前AI计算规模急剧扩张的关键突破口，其直接影响了大规模集群的协同效率和整体算力输出。随着国内高通量以太网联盟协议标准的发布，标志着以高通量以太网为代表的新一代开放智算网络，正加速迈向产业化落地。但协议的标准化与运维成本仍是主要面临难点。

AI基础设施网络互联



- **高通量网络技术是应对目前AI计算规模急剧扩张的关键突破口，其直接影响了大规模集群的协同效率和整体算力输出。**
  - **高通量网络技术重要性：**随着AI计算集群规模不断扩大，如何保持算力的线性扩展成为业界难题。AI模型的训练与推理具备规模大、周期长的特点，且图所示，AI大模型的训练涉及计算、网络、存储等多系统协调，网络间的高效互联可助力大模型的高效训练。AI训练过程依赖的庞大算力支撑与推理的实时性要求均对网络的高带宽、低延时与高稳定性提出了苛刻的要求，构建高通量规模网络成为保障AI工作负载的核心条件之一。
  - **高通量网络技术进展：**2024年9月在CCF全国高性能计算学术年会上，阿里云、中国科学院计算技术研究所等40余家机构举办发布会，联合发布国内首个高通量以太网ETH+协议标准，可实现集合通信性能***30%***的提升。同时，基于该协议的基础网络协议IP、开源网卡等硬件及系统也正式亮相，标志着以高通量以太网为代表的新一代开放智算网络，正加速迈向产业化落地。
  - **高通量网络技术面临挑战：**目前多组织协议标准协同、网络运维成本等为主要面临挑战。
    - ✓ **多组织协议标准协同：**高通量网络技术的发展受到业内各地协议标准化的影响，然而目前国内外各组织对于该技术发展的侧重点难以统一，为避免技术碎片化，推动统一的协议标准仍是个复杂且艰难的过程。
    - ✓ **网络运维成本：**在高速率、大规模的网络环境中，链路拥塞、设备故障或其他性能短板都可能成为影响整个AI计算任务进度和稳定性的瓶颈。传统的运维工具和方法可能难以快速、精准地定位故障，运维的复杂性和成本呈指数级增长。

来源：昇腾社区、阿里云、沙利文

3.2 高性能存储能力

关键发现

高性能存储是解决数据归集时间长、数据处理效率低等问题的核心环节。在训练侧，高性能存储可显著降低GPU空耗；在推理侧，“以存代算”通过SSD作为分级缓存承载KV-cache，有效缓解显存瓶颈。目前，全球各大厂商均在积极研发适用于AI工作负载的高性能存储解决方案，加速布局AI SSD产品线。但性能、能耗与成本问题仍有待解决。

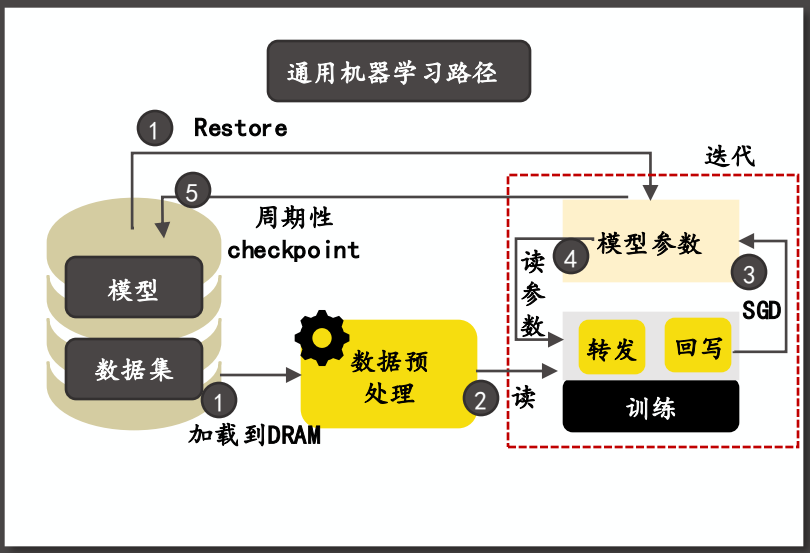
AI训练流程中的存储与需求

典型AI训练流程中的存储

- 通过Data Loader对训练前的数据加载样本集并进行预处理的过程涉及大量的IO读和数据请求
- 训练过程中周期性的Checkpoint

对存储的要求

- 高吞吐带宽
- 低延时
- PB级乃至EB级存储容量
- 线性扩容
- 高可靠性与可用性
- 数据安全与合规



❑ **数据总量与质量决定AI模型上限，海量数据的准备效率与流转效率影响着大模型端到端生产成本，高性能存储则是解决数据归集时间长、数据处理效率低等问题的核心环节。**

- **高性能存储能力重要性：**AI的发展离不开海量的数据，高性能存储能力能够满足AI系统对数据存储容量的需求，确保数据安全可靠性。此外，高性能存储可支持对大规模数据的快速处理，提高整个AI系统的运行效率。在训练侧，高性能存储可显著缩短数据加载与CKPT读写时间，降低GPU空耗；在推理侧，“以存代算”通过SSD作为分级缓存承载KV-cache，有效缓解显存瓶颈，降低时延。

- **高性能存储能力技术进展：**全球各大厂商均在积极研发适用于AI工作负载的高性能存储解决方案，加速布局AI SSD产品线。国内以华为的“以存代算”存储策略与存算一体AI SSD为代表的技术取得显著进展，其中，华为发布的全球首款AI SSD，首次在存储介质内直接进行AI推理计算，将“存储-计算-返回”流程的时延大幅降低***78%***。

- **高性能存储能力面临挑战：**随着AI模型越发复杂，高性能存储面临着性能、能耗以及成本等多方面挑战与压力。性能层面，目前存储带宽增速远远落后于算力增速。成本层面，模型及数据集容量激增，企业级SSD单位容量成本显著高于HDD，存储投资压力较大。能耗层面，数据中心电力消耗巨大，作为重要组成部分的存储设备，高密度SSD功率密度提升对可持续发展产生影响。

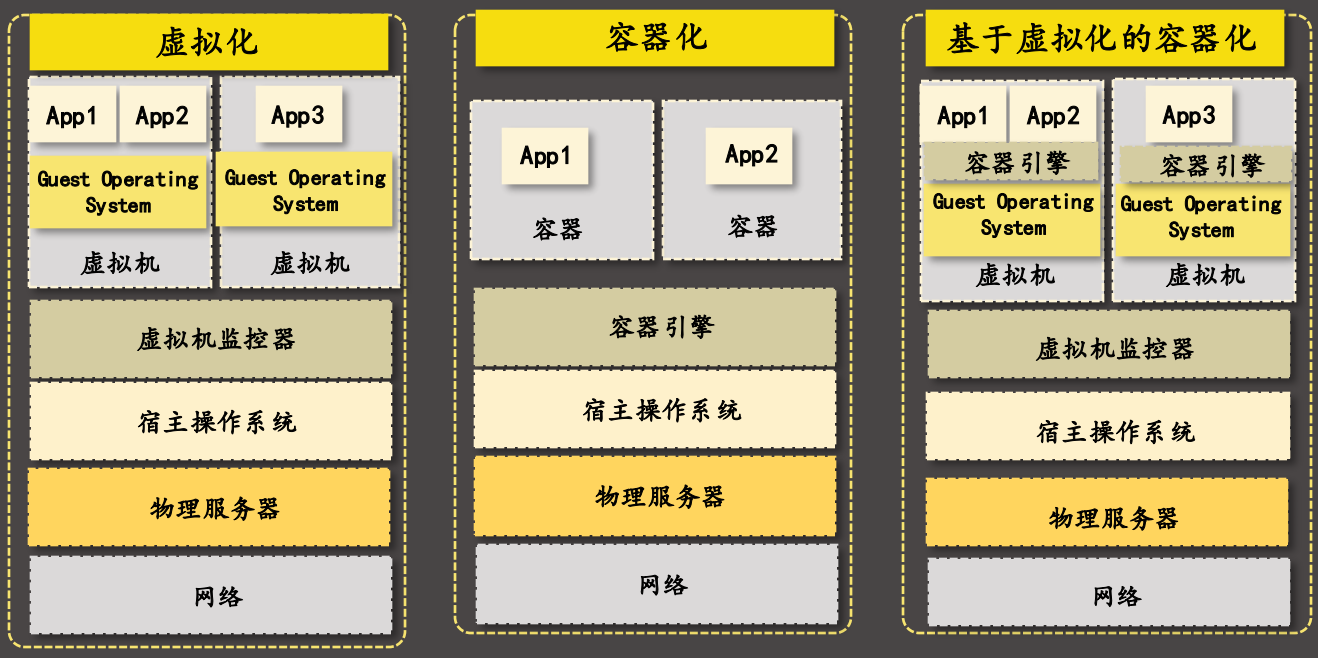
来源：腾讯云、百度智能云、沙利文

3.3 算力弹性调度能力

关键发现

高效的算力调度能力一方面有助于提升集群资源利用率，直接降低企业用户的运用成本；另一方面，AI应用存在显著的波峰波谷和突发请求，弹性调度可基于业务需求，实现资源的动态分配，保障AI服务的稳定性与连续性。但现行的调度机制往往无法高效利用计算资源，无法根据实时需求动态调整配置。

算力弹性调度基础 – 虚拟化 VS 容器化



- **高效的算力调度能力不仅可以优化资源配置，还可提升整体的资源利用率进而降低运营成本，是释放AI巨大潜力的重要保障。**

  - **算力弹性调度能力重要性：**一方面高效的弹性调度有助于提升集群资源利用率，直接降低企业用户的运用成本；另一方面，AI应用存在显著的波峰波谷和突发请求，弹性调度可基于业务需求，实现资源的动态分配，保障AI服务的稳定性与连续性。

  - **算力弹性调度能力相关技术：**虚拟化、容器化与池化为算力弹性调度基础。虚拟化技术是将物理芯片分割为多个独立虚拟逻辑单元进而合理匹配不同的任务；容器化则是将软件代码与其所有必要组件打包隔离在各自的容器中；池化则是统一管理计算资源，将分散的碎片化资源进行再整合。此外，目前异构兼容技术与预测性弹性伸缩技术逐渐成为基础技术外的算力调度新选择。

  - **算力弹性调度能力面临挑战：**随着各行业对算力调度的需求日益增加，当前的算力调度仍面临众多挑战。首先，算力资源的分布日益复杂，区域之间存在供需不平衡现状。这不仅增加了算力调度的难度，还导致了资源浪费和服务延迟。其次，现行的调度机制往往无法高效利用计算资源，无法根据实时需求动态调整配置。

章节四
AI基础设施市场空间梳理

- 中国AI基础设施市场规模分析
- 中国AI基础设施核心要素市场规模分析

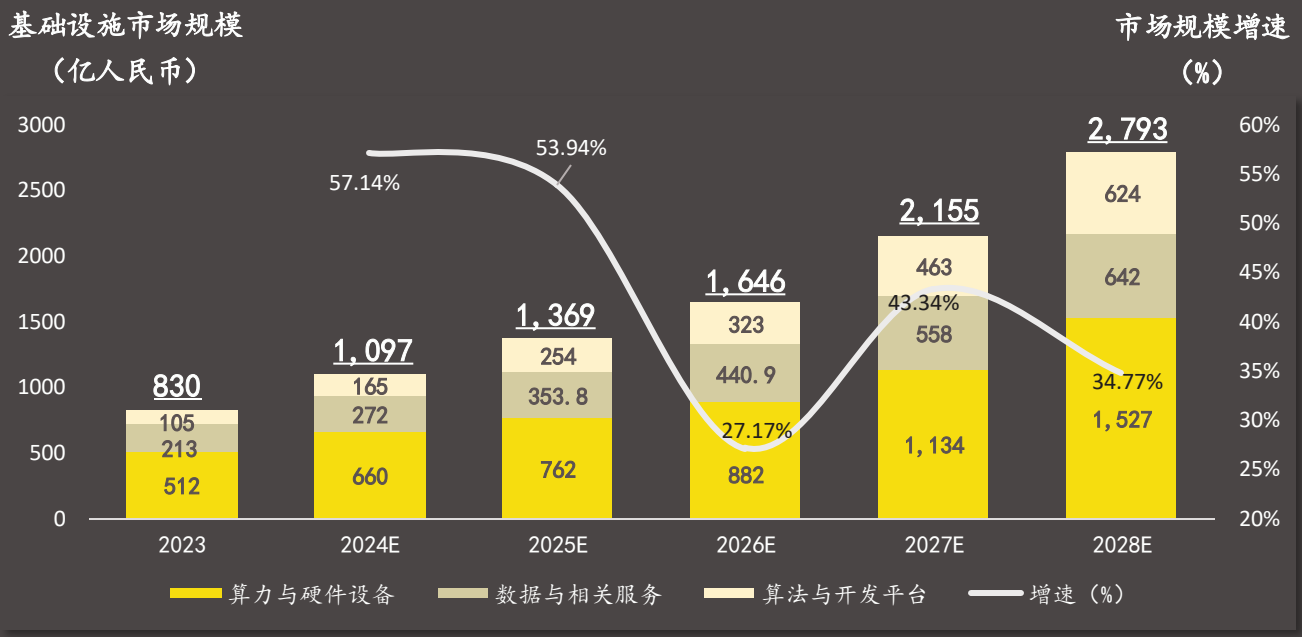
- 中国AI基础设施行业划分为算力、数据和算法核心要素，2023年市场规模达到830亿元，预计2028年AI基础设施市场将达到2,793亿元。随着AI基础设施逐渐落地各类应用场景，加之政策红利与国产技术创新，合力助推中国AI基础设施市场呈现持续增长态势。
- 当前，我国算力市场主要参与主体为三大基础电信运营商、算力中心服务商和云计算厂商。我国算力建设规模持续攀升，截至2024年第三季度算力总规模达到246EFLOPS，国家枢纽间20ms时延保障能力全面实现。随着AI应用场景的不断拓展，算力规模有望进一步增长。

4.1 中国AI基础设施市场规模分析

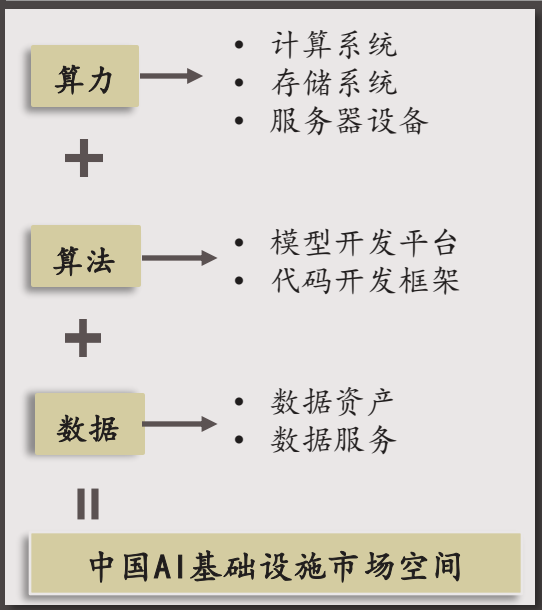
关键发现

中国AI基础设施行业划分为算力、数据和算法核心要素，2023年市场规模达到830亿元，预计2028年AI基础设施市场将达到2,793亿元。随着AI基础设施逐渐落地各类应用场景，加之政策红利与国产技术创新，合力助推中国AI基础设施市场呈现持续增长态势。

中国AI基础设施市场规模，2023-2028E



中国AI基础设施市场规模测算逻辑



- **中国AI基础设施行业按要素划分为算力、数据和算法。2023年中国AI基础设施行业市场规模为*830亿元*，预计2028年AI基础设施市场将达到*2,793亿元*，年增长率约*35%*。**
- 2023年中国AI基础设施市场规模为830亿元，预计2028年，市场规模将达到2,793亿元，2023-2028年复合增长率达到27.47%.
- 需求端，随着金融、医疗、制造等行业的智能化转型加速，推动了各行业对AI基础设施的迫切需求。例如金融领域的反诈等数字化系统、制造业的自动化质控等应用，均需依托高性能算力和数据处理能力支撑。此外，各行业数据量面临爆炸式增长与标注需求，进一步推动了底层存储和计算资源的扩张。
- 供给端，一方面，在国家支持AI基础设施建设的明确导向下，企业可通过税收优惠、资金补贴等政策积极投入相关资源；另一方面，国产技术实现持续突破，降低了对外部技术的依赖。需求端的应用拓展叠加更极端政策红利与技术创新，共同助推AI中国基础设施市场规模进一步增长。

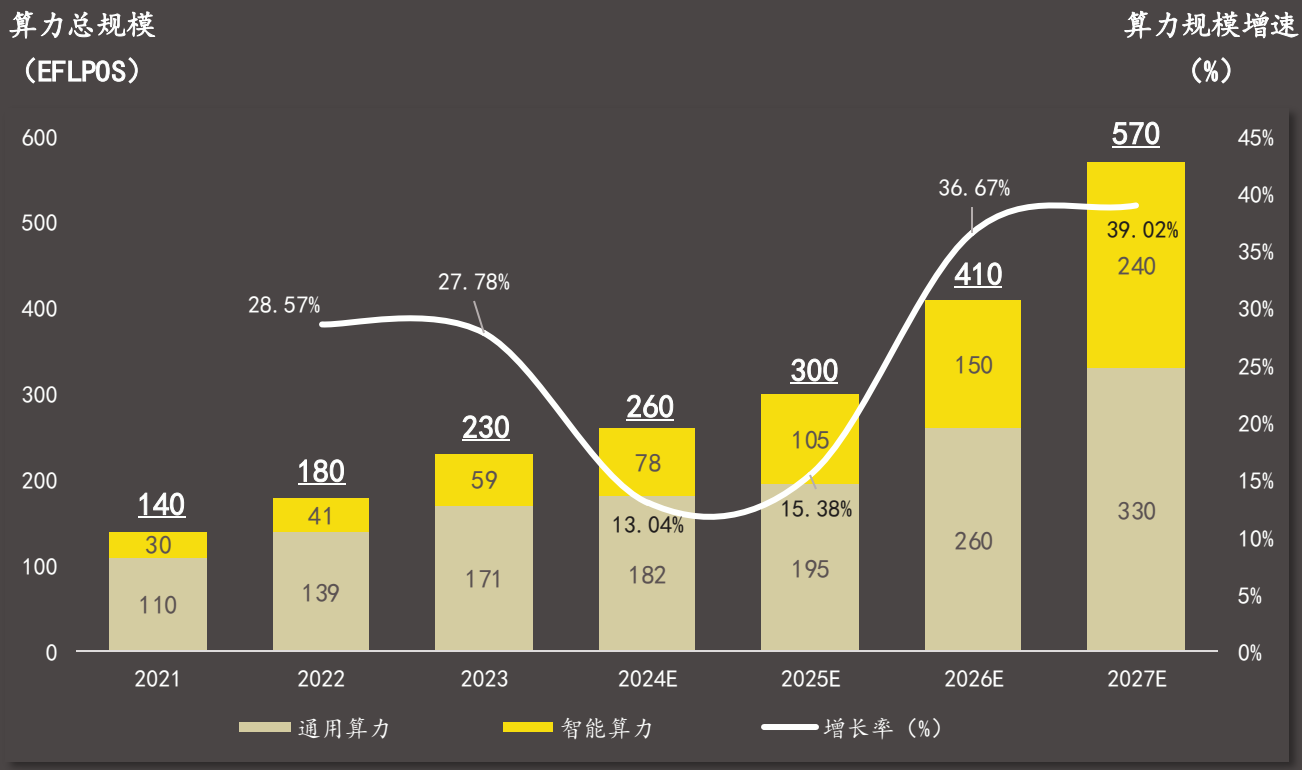
来源：沙利文

4.2 中国AI基础设施核心要素市场规模分析

关键发现

当前，我国算力市场主要参与主体为三大基础电信运营商、算力中心服务商和云计算厂商。我国算力建设规模持续攀升，截至2024年第三季度算力总规模达到246EFLOPS，国家枢纽间20ms时延保障能力全面实现。随着AI应用场景的不断拓展，算力规模有望进一步增长。

中国通用算力与智能算力规模，2021-2025E



- **在AI基础设施中，算力是推动创新与实现突破的核心驱动力。随着AI与云业务的不断发展，算力规模呈现稳定增长态势，截止2023年，通用算力与智能算力分别达到171与59EFLPOS，预计2027年通用与智能算力将分别达到330与240EFLPOS，整体增速达到39%。**

  - 供给端：我国算力市场主要参与主体为三大基础电信运营商、算力中心服务商和云计算厂商。近年来，算力基础设施投入不断加大，数据中心、超算中心等建设规模持续扩大，为通用算力与智能算力提供了坚实的支撑。其次，AI芯片与服务器等技术的快速发展，提升了算力的供给效率和性能。

  - 需求端：随着AI大模型应用场景的不断拓展，国内厂商对于算力资源的需求呈井喷式增长，其次，作为算力的重要输入-数据，在数字化推进加速背景下，数据量呈指数级增长，对算力需求也相应增加。未来，随着以农业、房地产等AI阶段相较缓慢的产业逐步迈入AI应用成熟落地阶段，对高算力需求将不断涌现，进一步推动了智能算力的智能需求。

  - 未来，通用算力与智能算力有望进一步增长，其中智能算力在AI Agent等大模型应用的发展推动下增速更可观，为算力规模增长的主要驱动力，预计2027年智能算力增速达到***60%***。

来源：信通院、沙利文

章节五
AI基础设施厂商竞争力分析

- 厂商竞争力因素
- 厂商竞争力评价指标
- 2025年中国AI基础设施综合竞争表现雷达图
- 领导者介绍

- 本报告设立创新指数评估体系对AI基础设施进行评价及分析，下设计算资源的高效分配利用、高性能网络架构设计、高性能存储系统构建、灵活可扩展的基础设施架构四大指标。
- 本报告设立增长指数评估体系对AI基础设施进行评价及分析，下设安全合格与风险管理、成本效益优化策略、专业交付与服务支持、生态系统与集成能力、行业客户与场景应用方案五大指标。

创新指数评价指标

关键发现

本报告设立创新指数评估体系对AI基础设施进行评价及分析，下设计算资源的高效分配利用、高性能网络架构设计、高性能存储系统构建、灵活可扩展的基础设施架构四大指标。

一级指标	二级指标	指标要点
计算资源的高效分配利用	算力供给能力	高效算力供应能力、AI算力分配与管理、按需推理服务、平台弹性伸缩能力等
	资源利用率优化管理	高性能工作负载的开发部署支持、现代推理加速工具集成、推理成本优化创新、高效调度策略等
高性能网络架构设计	网络性能优化	高性能网络架构、网络优化技术等
	网络拓扑和可扩展性	网络拓扑结构种类、网络架构可扩展性、通信优化策略、数据传输效率优化策略等
	网络监控与故障恢复	网络性能监控和分析工具、故障快速恢复机制、网络安全保护措施措施等
高性能存储系统构建	存储架构设计	训推存储优化策略、存储系统的可扩展性设计、数据存储解决方案的韧性等
	AI特化存储优化	小文件存储优化策略、大规模模型检查点的存储和快速恢复机制等
	智能数据管理	支持AI工作流的数据管理、训推数据流水线优化、数据层高实时性事务处理能力等
灵活可扩展的基础设施架构	架构可扩展性及多样化部署选项	平台弹性伸缩能力、大规模GPU集群的稳定性和安全性、支持的部署模式、基础设施的平滑扩展和技术升级等
	集群管理与资源配置	AI计算集群的快速创建和扩缩容、队列管理的灵活性和易用性等
	资源动态调度	计算资源的动态分配和回收、多租户支持和资源隔离机制等
	可观测性和管理	大规模AI基础设施的日常运维工作简化方案、系统异常检测和自动修复、AI工作负载的性能分析工具、全栈监控和可观测性能力等

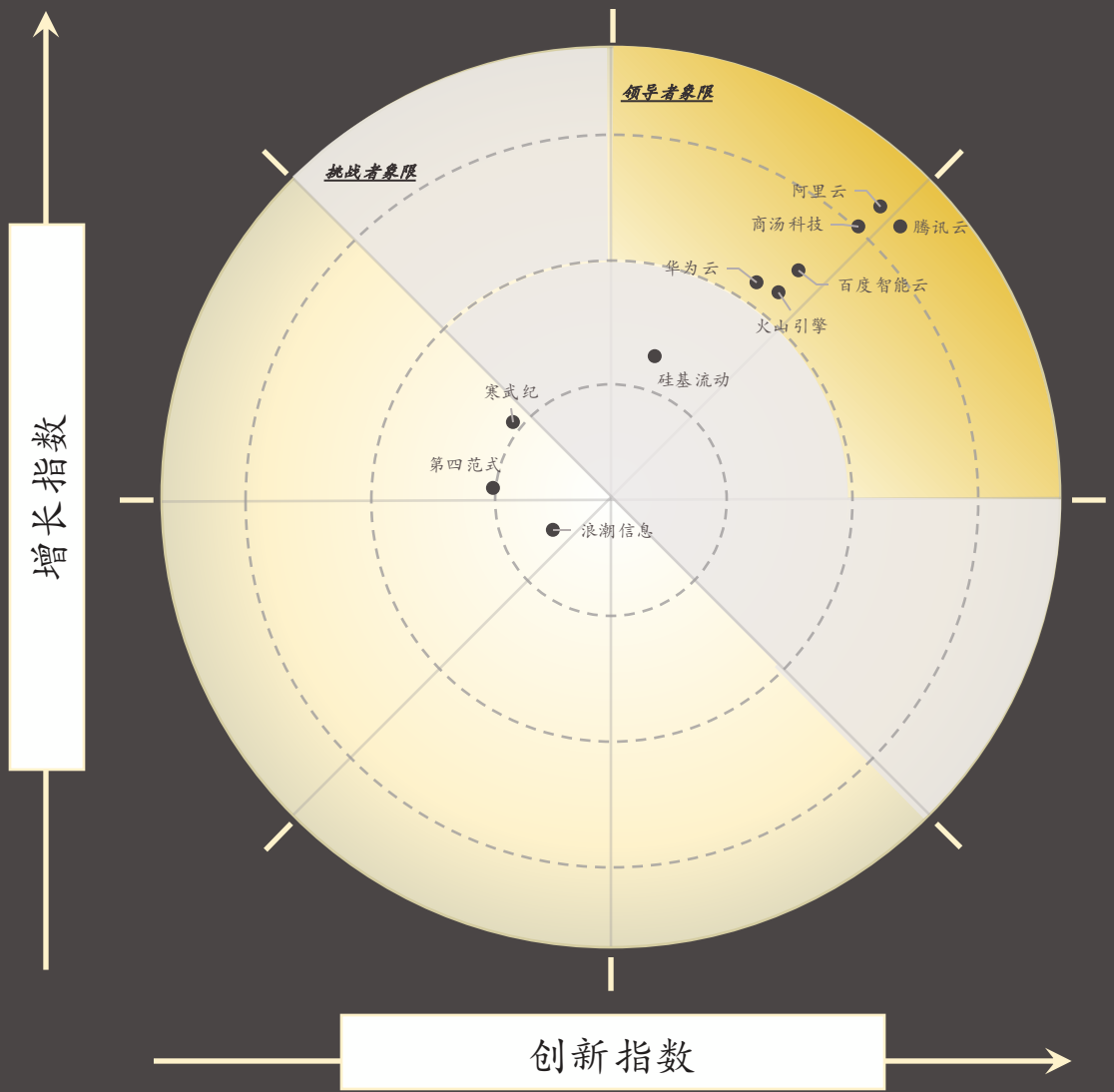
增长指数评价指标

关键发现

本报告设立增长指数评估体系对AI基础设施进行评价及分析，下设安全合格与风险管理、成本效益优化策略、专业交付与服务支持、生态系统与集成能力、行业客户与场景应用方案五大指标。

一级指标	二级指标	指标要点
安全合规与风险管理	数据安全与隐私保护	训推过程中的数据安全保护措施、多租户环境下的数据隔离与隐私保护等
	模型安全与完整性	模型知识产权保护、AI伦理与负责任的AI、恶意操纵与攻击防御等
	基础设施安全	网络安全、物理基础设施安全等
	安全运营与响应	平台安全检测、应急措施与恢复机制、安全隔离识别与评估等
成本效益优化策略	资源配置与成本模型	资源配置商业模式支持、成本预估与优化、长期运维成本方案等
	性能与成本平衡	差异化成本优化策略、性能与成本优化平衡等
	能源效率与可持续性	能源效率提高方式、可持续性长期规划、用户长期TCO优化等
专业交付与服务支持	项目实施和迁移服务	AI基础设施部署与迁移支持服务等
	客户成功和关系管理	长期合作管理策略、客户反馈收集与响应等
	技术支持与咨询服务	技术支持类型、紧急故障响应机制、AI基础设施设计和优化咨询服务、规范化产品文档等
	培训和增值服务建设	培训项目类型、增值服务等
生态系统与集成能力	兼容性	训推加速工具、数据源与数据管理系统集成、针对行业的预集成解决方案、技术合作伙伴生态系统等
	开放性	第三方开发支持、生态系统发展战略、AI模型部署与服务集成简化能力等
行业客户与场景应用方案	整体用户规模	用户规模、注册用户增长速度等
	各应用行业（金融、医疗等）	各行业的覆盖场景、细分场景的服务功能、采用的优势技术与服务等

2025年中国AI基础设施市场综合竞争表现——Frost Radar（弗若斯特雷达）™



注：圆环按由内向外递增的逻辑对应由低至高的综合评分，竞争力由“创新指数”以及“增长指数” 综合得出。结论仅适用于该阶段AI基础设施市场发展情况。

- **纵坐标代表“增长指数”：**
  - 衡量竞争主体在AI基础设施增长维度的竞争力，位置越靠上方，AI基础设施的安全合规与风险管理、成本效益优化、专业交付与服务支持、生态系统与集成等水平越高。
- **横坐标代表“创新指数”：**
  - 衡量竞争主体在AI基础设施创新维度的竞争力，位置越靠右侧， AI基础设施在计算资源高效分配利用、高性能网络架构设计、高性能存储系统构建等方面的能力越强。

阿里云

2025年AI基础设施Frost Radar排名说明

✓阿里云在2025年AI基础设施Frost Radar中增长指数排名第一
✓阿里云在2025年AI基础设施Frost Radar中创新指数排名第二

阿里云在增长指数排名第一，在以下指标项得分最高：

■ **阿里云在安全合规与风险管理模块得分最高：**

- 阿里云在AI训推过程中，从数据存储安全、数据传输安全与数据处理安全方面均提供完善的保护机制。此外，在多租户环境下，阿里云在数据隔离、数据加密等技术基础上，通过全链路可信身份传递、合规性与审计、审计日志与监控等功能增强数据全流程安全性。模型安全与完整性方面，阿里云提供了全面的技术功能以保护AI模型的知识产权，防止模型被恶意操纵或攻击。且阿里云通过AI伦理委员会的设立、AI偏见检测和解析工具的提供以及AI决策透明度增强技术，在AI伦理和负责任的AI方面做出长期承诺计划。基础设施安全层面，阿里云采取了DDoS高防、Web应用防火墙、云防火墙与数据安全中心等措施全面保护AI基础设施免受网络攻击。此外，针对物理基础设施，阿里云采取动环监控实现全环境的动态监控与警报，物理接口加固以防止未经授权的人员通过接口访问设备。

■ **阿里云在生态系统与集成能力模块得分最高：**

- PAI平台为用户提供了丰富的自研分布式训练加速框架与第三方分布式训练框架支持、数据处理加速和推理优化工具等协助用户提高训推效率。此外，阿里云拥有领先的合作伙伴生态系统，拥有超过一万两千家生态合作伙伴，伙伴类型覆盖伙伴类型覆盖渠道、ISV、SI、MSP所有类型，在AI基础设施解决方案领域设有特定伙伴。合作伙伴中，阿里云面向高校和科研机构推出了“云工开物”计划，提供支持高校教育和科研的权益支持，帮助高校和科研机构更好地使用AI相关服务。目前已支持超过国内外***150所***高校，超过***1,000项***合作项目。最后，阿里云积极参与开源社区，贡献了大量的开源项目， 是Apache 基金会的顶级赞助商之一，贡献了多个开源项目。阿里云还发起成立了中国最大最活跃的开源模型社区ModelScope，开源了一系列自研基础模型，千问系列开源模型已超过Llama成为衍生模型数量最大的模型家族。

阿里云在创新指数排名第二，在以下指标项得分最高：

■ **阿里云在高性能存储系统构建模块得分最高：**

- 存储架构设计方面，阿里云提供大规模、低成本、高可靠的云存储服务-对象存储OSS，专为AI和HPC场景设计的高性能文件存储CPFS（提供高达***2TB/s***的吞吐量和***3,000万IOPS***，具备去中心化的分布式架构，支持全并行I/O访问），以及面向ECS实例、E-HPC和容器服务等计算节点的文件存储服务-NAS等丰富的存储系统以支持AI工作负载。AI特化存储优化方面，阿里云PAI-DLC平台提供EasyCkpt高性能状态保存恢复、AIMaster的容错监控、SanityCheck健康检查等多种方式提升容错，以支持AI模型检查点的高效存储和恢复。智能数据管理方面，阿里云完整覆盖数据版本控制、元数据管理、生命周期分层冷热数据分层存储等数据管理功能以支持AI工作流。

■ **阿里云在架构可扩展性及多样化部署选项模块得分最高：**

- 架构可扩展性及多样化部署选项层面，阿里云的PAI灵骏智算服务提供了包含AI Master自动容错弹性训练、训练任务作业弹性伸缩等多种弹性伸缩和资源管理的能力，确保在扩展过程中的性能一致性。阿里云专有云和公共云技术同源，因此对于客户本地部署的需求，公共云产品的升级，均可以在专有云平台（本地部署）平滑同步升级，满足用户扩展升级灵活性。

来源：阿里云、沙利文

领导者：阿里云

关键发现

阿里云是AI基础设施领域的领导者。阿里云为企业级用户提供了全栈自研AI基础设施，凭借强大的自有生态网络、大规模商用实践经验以及领先的网络架构，为用户提供了稳定安全的计算环境。

阿里云AI基础设施架构

❑ **阿里云为企业级用户提供了全栈自研AI基础设施，具备完善的自有生态网络、大规模商用实践经验以及稳定安全的AI基础设施架构。**

- **全栈自研、积极创新：**阿里云通过全栈自研构筑壁垒，从底层的算力到AI平台，始终坚持全栈技术创新。其自研产品及功能包含但不限于倚天等自研芯片、自研训推服务框架、自研HPN7.0网络架构等。

- **完善的自有生态网络：**阿里云拥有强大的生态系统，一方面阿里云拥有完善的产品服务生态；另一方面阿里云拥有***12,000+家***生态合作伙伴，与高校和科研机构保持紧密的合作关系，此外，阿里云还发起了全国最活跃的开源模型社区-魔搭社区。强大的生态优势进一步增强了阿里云的用户粘性与应用场景的深入。

- **安全稳定的基础设施架构：**阿里云为企业级用户提供了强大的安全防护体系与领先的网络架构。阿里云采取了DDoS高防、Web应用防火墙、云防火墙与数据安全中心等措施全面协助用户保护AI基础设施免受攻击。此外，阿里云提供了多种方式来监控和评估节点的健康状态，帮助用户快速定位问题并进行故障排查，满足用户对于产品安全稳定的要求。

来源：阿里云、沙利文

腾讯云

2025年AI基础设施Frost Radar排名说明

✓腾讯云在2025年AI基础设施Frost Radar中创新指数排名第一
✓腾讯云在2025年AI基础设施Frost Radar中增长指数排名第二

腾讯云在增长指数排名第二，在以下指标项得分最高：

■ **腾讯云在资源配置与成本模型模块得分最高：**

- 资源配置与成本模型方面，腾讯云提供了丰富的工具协助用户进行成本预估与优化。例如腾讯云费用中心的成本分析模块可协助用户依据费用类型、产品、地域以及项目进行细化维度的筛选和统计，实时监控资源使用情况以识别资源浪费。此外，腾讯云开源的Crane项目内置预测算法，相较传统弹性工具，Crane项目可根据预测分析提前触发横向与纵向弹性伸缩，避免滞后性导致的资源不足问题。

■ **腾讯云在能源效率与可持续性模块得分最高：**

- 能源效率与可持续性方面，腾讯云拥有先进的数据中心技术架构设计，其中腾讯第四代数据中心通过模块化、标准化、产品化以及预制化的设计理念，有效缩短了建设周期，大幅降低了建设过程的碳排放量。该数据中心在运营阶段可实现年均PUE达到*1.2*，节能效果显著。

■ **腾讯云在安全运营与响应模块得分最高：**

- 安全运营与响应方面，腾讯云构建了“感知-决策-响应”一体化安全运营体系进行风险智能监控和自动化响应。在风险事件发生时，腾讯云通过“三步应急”机制，即快速抑制、数据恢复与溯源复盘最小化安全事件的影响，并通过生成的事件报告持续优化防护策略。

腾讯云在创新指数排名第一，在以下指标项得分最高：

■ **腾讯云在资源利用率优化管理模块得分最高：**

- 资源利用率优化管理方面，腾讯云集成了领先的现代推理加速工具，拥有自研的TACO-DiT和TACO-LLM框架，其中TACO - LLM 的吞吐性能相对于社区 SOTA 提升 ***1.8 - 2.5倍***；营运成本降低 ***44% - 64%***。**此外，腾讯云提供了多样化的调度策略与资源优化工具，包含但不限于TCCL通讯库优化、分布式训练框架优化、存储性能优化与星脉网络优化等。其中，TCCL通讯库优化可实现网络负载性能提升40%；分布式训练框架优化可减少*33%*的模型状态存储空间；星脉网络优化可提升通信效率*30%*，并使得大模型训练效率提升*10%*。**

■ **腾讯云在网络性能优化模块得分最高：**

- 网络性能优化方面，腾讯云提供智能高性能网络IHN（星脉）-基于RoCEv2技术栈提供软硬一体的网络解决方案，为用户提供高带宽、低延时的网络解决方案。此外，针对AI工作负载中的突发性高带宽需求，腾讯IHN网络流量调度系统可通过构建全局拓扑结构，绘制全网流量和业务矩阵，计算最佳流量分布，进而自动进行流量规划和动态感知调度以实现全网吞吐性能最优。相比传统方式，腾讯INH网络流量调度系统可实现通信性能提升***30%-50%***。

来源：腾讯云、沙利文

领导者：腾讯云

关键发现

腾讯云集结软硬自研产品能力，打造了算、存、网、数一体的高性能智算底座，在稳定性和性能上业内领先，为AI创新输出性能领先、多芯兼容、部署灵活的智算产品能力。企业凭借产品的性能与稳定性、积极创新的研发态度与各行业丰富的经验领先行业。

腾讯云AI基础设施全景能力图

❑ **腾讯云集结软硬自研产品能力，打造了算、存、网、数一体的高性能智算底座。凭借AI基础设施产品的智能性能与稳定性、积极创新的研发态度与各行业丰富的经验领先行业。**

- **智算性能和稳定性：**腾讯云可实现集群千卡单日故障数低至*0.16*，为行业水平的*1/3*。此外，腾讯云的集群千卡*1分钟*能完成万卡checkpoint写入，数据读写效率为业界的*10倍*，通信时间更为业界的一半。

- **积极创新：**腾讯云集结软硬自研产品能力，打造了算、存、网、数一体的高性能智算底座，在AI高性能计算、AI高性能存储、AI高性能网络、向量数据库等方面均进行了系统性的优化与创新，打破了算存网的“木桶”效应。

- **应用场景丰富：**腾讯云产品应用场景丰富，其覆盖B端大客户、中小型企业与开发者等不同类型用户。覆盖包含互联网应用、出行与智能驾驶、大模型训练、广告搜索推荐、智能制造、医疗与教育等*15+*个行业，积累了丰富的行业实践经验。

来源：腾讯云、沙利文

商汤科技

2025年AI基础设施Frost Radar排名说明

✓商汤科技在2025年AI基础设施Frost Radar中增长指数排名第二
✓商汤科技在2025年AI基础设施Frost Radar中创新指数排名第三

商汤科技在增长指数排名第二，在以下指标项得分最高：

■ **商汤科技在项目实施和迁移服务模块得分最高：**

- 项目实施和迁移服务方面，商汤科技SenseCore 2.0 提供一整套端到端的实施服务和工具，确保客户能平稳过渡，快速上线并实现长期成本和性能优势。最关键的，商汤大装置有专门团队提供“AI专家服务”，为不同客户群体提供针对性的服务，包括模型研发客户、应用开发客户、私有化客户、本地应用开发客户等。以协助客户完成AI大模型训练与推理基础设施的部署和迁移，包含从需求评估、TCO分析、架构设计、自动化工具应用、数据迁移以及后续运维优化等流程，为用户提供覆盖全流程的专业迁移服务。

■ **商汤科技在培训和增值服务建设模块得分最高：**

- 培训和增值服务建设方面，商汤科技除了为用户提供基础的培训课程外，还提供周期性的深度共创培训，即针对潜在客户和现客户，提供主题课程类的深度培训，从技术发展趋势到具体产品应用。此外，商汤科技还为用户提供高级数据分析服务、行业特定AI解决方案、实时性能监控工具与专业的模型评估服务等增值服务，一方面协助用户基于需求选择更适合的AI基础设施产品；另一方面协助用户对未来的业务趋势、市场变化、用户行为等进行预测，帮助客户提前做好规划和准备，抢占市场先机。

商汤科技在创新指数排名第三，在以下指标项得分最高：

■ **商汤科技在算力供给能力模块得分最高：**

- 算力供给能力方面，商汤科技具有领先的AI专用算力集群规模。目前，商汤大装置已经纳管了***20,000P***的算力资源，已上线的GPU数量达到***5.4万***张。此外，商汤还拥有建设中的智算中心数十个，已计划***20,000P***新增算力。同时，适配了20多款国产AI芯片，能够保障未来国产化浪潮来袭。最后，商汤大装置通过虚拟化、容器等技术，以及全国接近***10个***智算中心网络节点的方式，让客户可以随时根据需求进行资源的调配和扩容，以保证在硬件层面为用户持续供应高效的GPU算力。

■ **商汤科技在集群管理与资源配置模块得分最高：**

- 集群管理与资源配置层面，商汤科技SenseCore 2.0 控制台提供了丰富的功能与工具以支持AI计算集群的快速创建和扩缩容。控制台提供针对不同规模的生成式 AI 和大模型任务，预设了计算资源、存储资源、网络配置等参数组合的多类型预置模版，用户只需根据自身需求选择相应模板。同时，控制台以图形化方式展示集群创建的各个步骤，用户无需记忆复杂的命令和参数，通过简单的鼠标点击和文本输入即可完成诸如选择节点类型、设置节点数量等操作，大幅降低了用户的操作门槛，提高创建效率。此外，控制台集成了实时资源监控功能，可直观展示GPU、内存、存储等资源的使用情况。用户可设置资源阈值，当资源利用率超过设定的上限或低于设置下限时，系统自动触发扩缩容操作，以实现资源的高效利用。

来源：商汤科技、沙利文

领导者：商汤科技

关键发现

商汤科技大装置以高度集成的端到端架构，致力于为企业提供敏捷、灵活、可靠的全栈的原生AI基础设施服务，以极致性价比推动大模型技术的高效落地与规模化应用，为不同行业、不同场景的客户提供贴身的解决方案，切实帮助客户解决技术问题、产品问题和商业化问题，从而推动人工智能产业的长远发展。

商汤大装置-原生AI基础设施

❑ **商汤科技凭借超大规模算力支持、卓越的训推协同优化能力、深度布局的场景化解决方案，以及全链条的专家服务，成为“最懂大模型的AI基础设施”，帮助用户能以更高的投入产出比拥抱AI变革。**

- 超大算力规模：商汤大装置具备23,000 PetaFLOPS的算力规模，智算节点覆盖长三角、粤港澳、京津冀、中西部等重点区域，并实现全国联网统一调度。针对国产芯片的结构与通讯机制完成深度调度优化，实现了规模化商用，支持多种异构芯片5,000卡集群上单一大模型训练任务调度与运行，异构训练效率达同构训练的95%。
- 强大的训推优化能力：通过算力与模型的深度协同优化，在训练与推理效率上构建显著技术壁垒，拥有自研训练框架，开发并优化多种并行策略，提升训练性能和显存管理，另外，支持开源vLLM及自研LightLLM双推理引擎，极大提升推理效率并压低推理成本。
- 开箱即用的场景化解决方案：从场景定义到业务落地全链路赋能：基于对行业共性需求与具体业务痛点的深度理解，结合行业知识与业务流程，构建服务端壁垒，为客户提供场景化解决方案，具体包括具身智能、AIGC、AI4S和产业智能化等场景化解决方案，既能“开箱即用”，也可“快速集成”。
- 积极拥抱开源生态：通过OpenAPI兼容、K8S原生平台适配、开源大模型托管服务、全栈开源工具链及丰富的开源组件等，实现了技术栈的无缝整合与敏捷迭代，满足快速发展的技术栈应用诉求。
- **全链条的大模型专家服务：**面向大模型落地，提供全面的AI专家服务，包括大模型咨询、定制、场景部署与优化和维护等环节，形成从需求分析、模型开发到产业化落地的完整闭环。

来源：商汤科技、沙利文

百度智能云

2025年AI基础设施Frost Radar排名说明

✓百度智能云在2025年AI基础设施Frost Radar中增长指数排名第三
✓百度智能云在2025年AI基础设施Frost Radar中创新指数排名第四

百度智能云在增长指数排名第三，在以下指标项得分最高：

■ **百度智能云在性能与成本平衡模块得分最高：**

- 性能与成本平衡方面，在计算资源不变的情况下，百度智能云通过提升模型训练性能（算子融合、显存优化），提升单位算力效率；在计算性能不变的情况下，则通过并行策略、自适应调度策略，使单位计算资源能够并行更多AI任务以提升算力利用率；在计算量不变的情况下，通过多芯策略，调整最佳算力结构，以降低整体算力成本。此外，针对不同规模的AI任务，平台提供具有差异化的方案。在大规模训练场景，平台通过并行计算、显存优化、算子融合、多芯策略、智能调度、过程容错等机制提升训练效率，降低算力成本；针对于小模型的场景，平台提供GPU切分能力，支持多个任务共享GPU，进一步降低生产成本。

■ **百度智能云在行业客户与场景应用方案模块得分最高：**

- 行业客户与场景应用方案方面，百度智能云凭借领先的创新技术（包括但不限于自研的AIAK训推加速框架、自研昆仑芯片等）、成熟的客户管理体系与完善的客户服务支持等优势，在行业内沉淀了可观的客户规模。其行业客户与应用场景全面覆盖了金融、互联网、汽车、能源等AI应用阶段领先的行业，其中互联网与汽车行业中诸多头部厂商与百度智能云展开密切合作，为百度智能云具备核心优势的应用行业。

百度智能云在创新指数排名第四，在以下指标项得分最高：

■ **百度智能云在AI特化存储优化模块得分最高：**

- AI特化存储优化方面， 百度智能云的PFS提供了基于内存缓存的元数据管理技术，通过客户端缓存机制，在保证语义的前提下，能安全的命中缓存，减少跨网络和磁盘访问开销。针对数据读写逻辑较为简单的AI训练场景，百度智能云利用Page Cache 将一些比较小的文件缓存到内存空间，通过缓存加速了对小文件的访问性能。此外，百度智能云拥有完善的数据存储方案，包含但不仅限于高性能的NVMe SSD存储设备、元数据服务器扩展、高效的网络协议与数据预加载等，协助用户大幅提高GPU与存储设备之间的吞吐量和降低数据访问延迟，减少算力空置损失。最后，百度智能云提供高性能的checkpoint保存框架，相较开源的方案，单次checkpoint保存耗时最高减少***93到99%***。

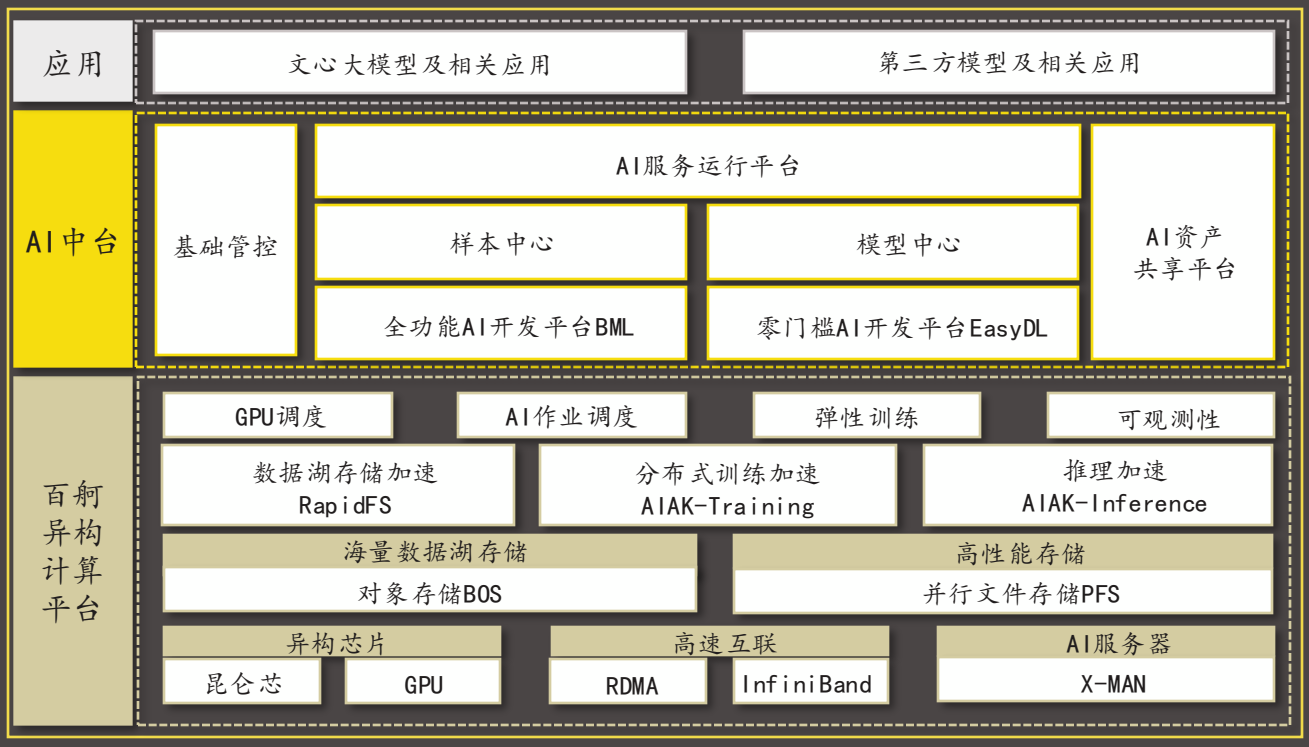
来源：百度智能云、沙利文

领导者：百度智能云

关键发现

百度智能云AI大底座作为中国AI领域首个基于产业深度实践的基础设施，为广泛行业用户提供了一站式自主创新的解决方案，支撑企业在智能算力基础设施建设上实现集约化建设、高性能应用与持续性增强。

百度智能云AI大底座全景图



❑ **百度智能云坚持将自研技术贯穿完整架构，以构筑自主可控的核心产品与技术。其自研的AI大底座为各行业客户提供了从IaaS到PaaS的自主可控、自我进化的解决方案。**

- **广泛的行业覆盖：**百度智能云的AI基础设施解决方案已覆盖互联网、制造、能源、金融、交通物流等多个领域。此外，百度智能云拥有庞大的开发者社区和合作伙伴生态，为企业提供全方位支持，以确保客户建立长期合作关系。

- **领先的自主创新能力：**百度智能云自研的“AI大底座”由百度百舸计算平台与百度AI中台解决方案两大平台构成，为企业用户提供了从IaaS到PaaS的全栈自研、自主可控、自我进化的解决方案，支撑企业在智能算力基础设施建设上实现集约化建设、高性能应用与持续性增强。

- **高性能“云智一体”架构：**百度智能云坚持“云智一体”架构，以各行各业的核心场景为切入点，构建更低成本的异构算力和更高效的开发运营能力，为各行业用户提供极致能效。

来源：百度智能云、沙利文

名词解释页（1/2）

- **智能体:**智能体（Agent），作为人工智能领域的一个重要概念，是指能够自主感知环境、做出决策并执行行动的系统。它具备自主性、交互性、反应性和适应性等基本特征，能够在复杂多变的环境中独立完成任务。智能体的出现，标志着人工智能从简单的规则匹配和计算模拟向更高级别的自主智能迈进。
- **向量数据库:**向量数据库是专门用来存储和查询向量的数据库，其存储的向量来自于对文本、语音、图像、视频等的向量化。与传统数据库相比，向量数据库可以处理更多非结构化数据（比如图像和音频）。在机器学习和深度学习中，数据通常以向量形式表示。
- **PaaS层**：PaaS层（Platform as a Service）是云平台中的一个关键架构层次，位于底层的IAAS（Infrastructure as a Service）和顶层的SAAS（Software as a Service）之间。PaaS通过提供一系列服务，使得开发人员能够更高效地开发和部署应用程序，而无需关心底层基础设施的细节。PaaS将软件研发的平台作为一种服务，允许用户在平台上完成应用程序的开发、部署、运行和管理。
- **MAAS**:MaaS是一种将机器学习模型作为服务提供给用户的云计算模式。它允许用户在不需要拥有自己的硬件设备或专业技能的情况下，通过API接口或在线平台调用预训练的机器学习模型，实现数据的实时预测、分析等功能。MaaS简化了AI模型的开发和部署流程，降低了技术门槛和成本，使得更多企业能够轻松拥抱AI技术。
- **SAAS**：Saas（软件即服务）是托管在云端的应用软件，由 Web 浏览器、移动应用程序或瘦客户端通过互联网连接使用。SaaS（软件即服务）提供商负责运营、管理和维护软件以及作为软件运行平台的基础架构。 客户只需创建一个帐户并支付费用，即可开始工作。
- **API**：应用程序编程接口（Application Programming Interface，简称：API），是一些预先定义的函数，目的是提供应用程序与开发人员基于某软件或硬件得以访问一组例程的能力，而又无需访问源码，或理解内部工作机制的细节。

方法论

- 头豹研究院布局中国市场，深入研究19大行业，持续跟踪532个垂直行业的市场变化，已沉淀超过100万行业研究价值数据元素，完成超过1万个独立的研究咨询项目。
- 头豹研究院依托中国活跃的经济环境，研究内容覆盖整个行业发展周期，伴随着行业内企业的创立，发展，扩张，到企业上市及上市后的成熟期，头豹各行业研究员积极探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业视野解读行业的沿革。
- 头豹研究院融合传统与新型的研究方法论，采用自主研发算法，结合行业交叉大数据，通过多元化调研方法，挖掘定量数据背后根因，剖析定性内容背后的逻辑，客观真实地阐述行业现状，前瞻性地预测行业未来发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。
- 头豹研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。
- 头豹研究院秉承匠心研究，砥砺前行的宗旨，以战略发展的视角分析行业，从执行落地的层面阐述观点，为每一位读者提供有深度有价值的研究报告。

法律声明

- 本报告著作权归头豹所有，未经书面许可，任何机构或个人不得以任何形式翻版、复刻、发表或引用。若征得头豹同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“头豹研究院”，且不得对本报告进行任何有悖原意的引用、删节或修改。
- 本报告分析师具有专业研究能力，保证报告数据均来自合法合规渠道，观点产出及数据分析基于分析师对行业的客观理解，本报告不受任何第三方授意或影响。
- 本报告所涉及的观点或信息仅供参考，不构成任何证券或基金投资建议。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告或证券研究报告。在法律许可的情况下，头豹可能会为报告中提及的企业提供或争取提供投融资或咨询等相关服务。
- 本报告的部分信息来源于公开资料，头豹对该等信息的准确性、完整性或可靠性不做任何保证。本报告所载的资料、意见及推测仅反映头豹于发布本报告当日的判断，过往报告中的描述不应作为日后的表现依据。在不同时期，头豹可发出与本报告所载资料、意见及推测不一致的报告或文章。头豹均不保证本报告所含信息保持在最新状态。同时，头豹对本报告所含信息可在不发出通知的情形下做出修改，读者应当自行关注相应的更新或修改。任何机构或个人应对其利用本报告的数据、分析、研究、部分或者全部内容所进行的一切活动负责并承担该等活动所导致的任何损失或伤害。

主笔分析师
江林烨
17721418134
chloe.jiang@frostchina.com

深度研究小组负责人
李庆
13149946576
livia.li@frostchina.com

www.frostchina.com ; www.leadleo.com
https://space.bilibili.com/647223552
https://weibo.com/u/7303360042