FROST & SULLIVAN

60 Years of Growth, Innovation and Leadership

中国智算中心操作系统
市场研究报告（2024）

A Frost & Sullivan
White Paper

目录

执行摘要.....2

关键发现点.....3

章节一：大模型发展开启计算密集型新时代.....4-7

AI大模型发展推动计算需求及模式变革.....4

以处理器为代表的硬件形态进步驱动软件形态更新.....5-6

立足国家战略，智算中心建设赋能新时代.....7

章节二：智算中心市场概览.....8-14

智算中心定义、特征及演进.....8

智算中心产业链分析.....9

智算中心价值分析.....10

国内智算中心建设现状.....11

全球智算中心建设现状.....12

智算中心效益分析.....13

智算中心面临挑战.....14

章节三：智算中心操作系统价值.....15-20

智算中心操作系统是智算中心的“中枢神经”及能力底座.....15

智算中心操作系统赋能智算中心高质量发展.....16-17

智算中心操作系统赋能算力租赁商业模式变革.....18-20

章节四：智算中心操作系统领先实践.....21-25

智算中心操作系统领先厂商实力评估.....21

领先厂商介绍.....22-25

附录.....26

执行摘要

简介

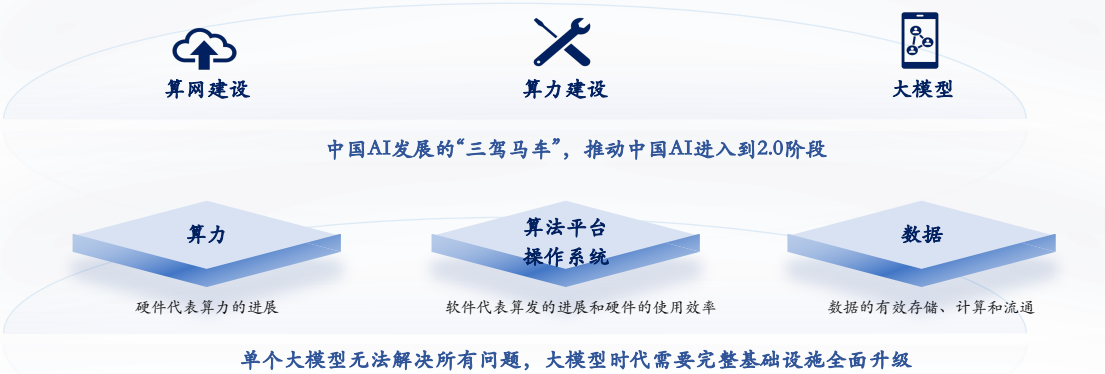
近年来，随着人工智能（AI）技术的迅猛发展，一种新型基础设施——智算中心（Intelligent Computing Center, ICC）应运而生。**智算中心专为处理大规模数据和复杂算法而设计，旨在支持AI、机器学习（ML）、深度学习等先进技术的应用。**AI和ML技术的快速发展，尤其是深度学习模型的兴起，对计算能力提出了前所未有的需求。现代社会产生的海量数据需要强大的计算资源进行处理和分析，金融、医疗、制造、智慧城市等行业对智算中心的算力支持需求日益增长。

AI大模型的发展预示着计算密集型新时代的到来，这一转变加速了硬件架构从CPU向GPU的转变，并推动了软件形态的更新，以满足AI时代对高并发计算的需求。智算中心采用最新的计算架构和技术，例如GPU集群，以提供高效的AI计算能力和数据算法服务，促进数字化转型和智能化升级，助力智慧城市建设和数字经济的发展。政府也出台了相关政策以支持智算中心的建设。展望未来，智算中心将成为科技创新和产业发展的关键平台，进一步推动AI技术发展和广泛应用。

智算中心操作系统（智算中心OS）在智算中心价值链中发挥着至关重要的作用。智算中心操作系统不仅是智算中心的"中枢神经"，也是实现资源统一管理和智能调度的基石。通过高效的资源管理和智能调度，确保了智算中心的高效运行和发展。这一操作系统不仅提高了算力的效能和降低了成本，而且通过提供全栈全流程的软件支持，降低了大模型算力系统的使用门槛。

智算中心产业链展现了从上游硬件制造商到下游智能服务应用的协同发展，为AI应用和创新提供了坚实的基础。智算中心操作系统产业生态中的领先厂商，如九章云极DataCanvas、百度智能云、浪潮云、商汤科技等，进一步证明了智算中心操作系统在金融、制造、能源、城市、医疗、媒体等多个行业中的应用价值，加速了这些行业的智能化进程。

本报告致力于深入探讨智算中心及智算中心操作系统的全貌，明确了智算中心操作系统的基本概念、关键技术、市场定位和核心价值，强调了智算中心操作系统在智能化时代所发挥的资源高效管理、智能调度等关键性支撑作用。此外，报告进一步评估了智算中心操作系统产业链中各典型厂商的综合竞争力，洞察行业引领者，为行业未来发展提供了方向与建议。



来源：弗若斯特沙利文

关键发现点

智算中心是融合人工智能与高性能计算技术的先进计算基础设施，旨在提供强大的计算能力以支持复杂的数据处理和分析任务。**智算中心的操作系统是其“中枢神经”和能力基础，它赋能智算中心的算力运行与业务运营，实现“用软件定义算力”的理念。**本报告将从行业宏观环境分析市场需求，从技术发展规律预测未来趋势，并识别在智算中心操作系统领域领先的厂商，对其综合竞争力进行评价。

驱动因素分析：智算中心操作系统行业的发展受到以下三个层面的驱动：**首先，随着大模型对算力需求的增加，以及5G、工业互联网、物联网等信息技术的加速发展，数据量呈现爆炸式增长。**《2022-2023中国人工智能计算力发展评估报告》显示，2021至2026年间，中国智能算力规模的年复合增长率预计为52.3%，而通用算力规模的年复合增长率为18.5%。**其次，随着下游应用场景的不断扩展，智算中心操作系统的需求也在迅速增长，它在大模型训练和推理中发挥着关键作用，并在智能制造、智慧城市、医疗健康等多个领域得到广泛应用。**预计未来，随着更多应用场景的实现和功能的优化，智算中心操作系统市场将加速增长。**最后，在全球对人工智能扶持的政策环境持续优化之下，中国对大模型技术的宏观战略规划为行业发展提供了有利条件。**目前，国家已经出台了包括“十四五”国家信息化规划、“十四五”数字经济发展规划在内的多项政策，以推动智算中心的发展和智算中心的建设，支持多模态大模型、科学计算大模型等的研发。

市场趋势分析：在核心驱动因素的推动下，中国智算中心操作系统市场正呈现出三大显著趋势，这些趋势在塑造市场竞争格局中逐渐清晰。**首先，人工智能技术的快速发展预示着未来大部分经济活动将依赖于AI应用。**据预测，约80%的经济场景将整合人工智能技术，全球范围内对智算中心的重视程度不断上升。在中国，目前已有超过30个城市正在积极建设或规划智算中心，这些中心往往与当地的经济开发区或高新技术开发区相结合，通过与产业需求紧密结合，旨在推动区域经济和技术创新的发展。

其次，智算中心操作系统的广泛应用正在推动构建一个普及和普惠的服务生态系统。这包括算法应用的普及和智能算力服务的普惠化，能够迅速形成高精度、高能力的行业解决方案，促进各行各业的智能化升级。智算中心为用户提供按需获取的算力、存储、带宽等专业服务，确保计算资源的广泛可用性和便捷获取。

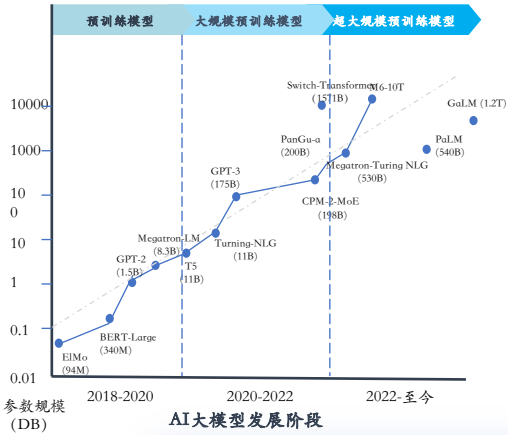
最后，智算中心操作系统作为智慧时代算力生产、调度的核心，扮演着至关重要的角色。它通过建立多层次、全方位的算力需求满足机制，专注于人工智能的核心任务，并提供集成了“算力、数据、算法、调度”的综合服务。这实现了全局加速优化和异构算力的管理和调度，为智算中心提供了高效的操作系统和有价值的服务，在智算生态系统中发挥着核心作用。

竞争格局：在智算中心操作系统的竞争激烈市场中，**关键竞争要素包括产品力、运营力、协同力、生态力**。通过性能优化、生态升级和差异化服务，各企业方能在长期市场竞争中实现持续创新增长。

章节一：大模型发展开启计算密集型新时代

AI大模型发展推动计算需求及模式变革

- 智能化时代已经到来，AI模型的发展经历了从基础预训练模型到超大规模预训练模型的多个阶段。随着模型参数数量的指数级增长，我们已经见证了千亿甚至万亿参数级别的大型模型的出现。
- 在人工智能技术的推动下，计算范式如并行计算、智能边缘计算、微服务等正在经历一场变革，这导致了对算力的需求呈爆发式增长。



AI大模型的发展从参数规模角度可以划分为三个阶段：预训练模型、大规模预训练模型和超大规模预训练模型。参数规模每年至少增长10倍，实现了从亿级到百万亿级的飞跃，目前千亿级参数规模的大模型已成为主流。

- **预训练模型：**这些模型依赖大量人工标注的数据进行训练，由于规模相对较小，它们在处理复杂任务时的能力受到限制；
- **大规模预训练模型：**起源于自然语言处理领域，以Google的BERT、OpenAI的GPT和百度的ERNIE（文心大模型）为代表，这些模型的参数规模逐渐增长至万亿级别；
- **超大规模预训练模型：**自2021年起，这一领域进入了快速发展阶段，以Vaswani提出的Transformer架构和OpenAI的GPT-3为代表，多模态大模型如Sora在视频生成领域引起了行业广泛关注，推动了业界对大模型在技术深度和规模广度上的持续探索。

AI大模型崛起为计算范式变革提供催化剂

并行计算

- 随着数据规模的指数级增长，传统的串行计算已经无法满足海量数据的处理需求，**并行计算技术的发展使得处理大规模数据和复杂问题成为可能。**
- 并行计算通过将任务分解为多个子任务，同时在多个处理单元上执行，极大提高计算效率，对于实现AI大模型的快速迭代和应用部署至关重要。

边缘AI

- **边缘AI将人工智能技术引入到边缘计算场景中。**随着AI大模型对实时性和低延迟的需求不断增长，边缘AI应需而生。
- **边缘AI能够在本地处理和存储数据，无需依赖互联网连接。这种本地化处理能力使得设备能够在几毫秒内做出决策。**随着边缘计算能力的增强，将大型AI模型部署在边缘侧和移动设备上已成为一种必然趋势。这一技术也推动了智能驾驶、智能制造、智能家居、虚拟现实等领域的应用迅速落地。

微服务

- 高效的推理微服务正加速人工智能的大规模部署。随着AI模型规模和复杂性的增长，传统架构越来越难以满足市场对快速迭代和灵活部署的需求。**微服务架构通过将大型复杂的应用程序分解为一系列小型、独立且功能专一的服务，有效解决了这一挑战。每个微服务都负责实现特定的业务功能，这样的设计简化了AI部署流程，提高了部署效率。**
- 例如，英伟达在2024年推出的生成式AI微服务，极大地助力了开发者构建生成式AI应用，将开发所需时间从数周大幅缩短至几分钟。

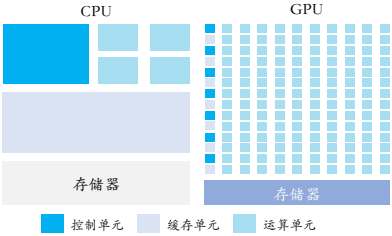
来源：弗若斯特沙利文

以处理器为代表的硬件形态进步驱动软件形态更新（1/2）

- 随着人工智能（AI）技术的不断进步，GPU的高并发计算能力已成为满足AI时代算力需求的核心要素，并且GPU在AI模型构建中的适配性日益增强。
- 随着AI模型参数的指数级增长，对底层算力的需求也在迅速提升，这推动了AI超级计算机和大规模GPU集群的发展。这种硬件层面的转变要求软件层面必须同步发展，以确保与硬件的最佳适配和性能优化。

□ AI大模型发展背景下GPU更适合AI计算

- **以CPU为代表的传统通用计算能力越来越难以满足AI时代的算力需求。**随着人工智能技术的持续进步，AI模型训练所需的算力正以爆炸性的速度增长。例如，从2020年6月GPT-3的发布到2023年3月GPT-4的发布，大型模型的计算需求增长了约7倍。相较于摩尔定律预测的芯片算力增长速度，大模型训练所需的算力增长速度预计远远超出了算力性能的提升速度。
- **相比之下，GPU更适合处理高强度并行计算场景。**AI计算中包含大量的并行计算任务，GPU拥有大量的运算单元，每个运算核心计算能力也更强，这使得GPU不仅适合处理复杂的并行运算，还具备强大的内存访问带宽，适合高并发的计算需求。

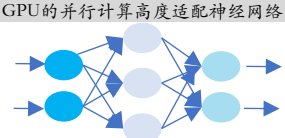


CPU和GPU架构对比

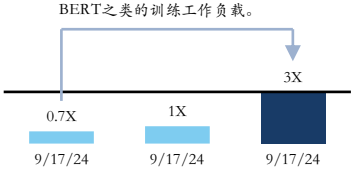
GPU高度适配人工智能计算，原因如下：1）传统的通过增加处理器内核数来提升算力的方法正面临物理极限的挑战，CPU性能的增长速度放缓；2）GPU拥有大量运算单元和简化的流水线设计，这使得它们非常适合执行计算密集型和易于并行化的程序。GPU在设计上注重吞吐量的优化，并具备强大的内存访问带宽，同时在设计上也注重延迟的优化，包括复杂的多级缓存系统（如L1、L2、L3）和逻辑控制单元；3）AI模型训练阶段需要处理大量数据和复杂的神经网络结构，而推理阶段则要求低延迟和高效的重复计算。深度学习中的常用算法，如矩阵乘法和卷积，可以分解为大量并行任务以提高计算效率。GPU凭借其强大的并行处理能力和高速的内存带宽，在处理训练和推理任务时表现出色。

AI模型构建（以英伟达A100为例）

GPU的并行计算高度适配神经网络

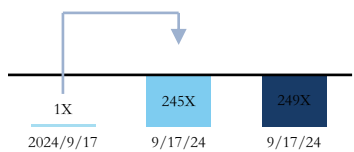


- **GPU可以帮助高速解决问题：**2048个A100 GPU可在一分钟内成规模地处理BERT之类的训练工作负载。



推理过程

- 多实例GPU（MIG）技术允许多个网络同时在单个A100 GPU上运行，从而提升计算资源的利用率。
- 在A100已有的推理性能增益基础上，仅结构稀疏支持一项就能带来高达两倍的性能提升。
- 在BERT等先进的对话式AI模型上，A100可将推理吞吐量提升至CPU的249倍。



ChatGPT引发GPU热潮

百度：推出文心一言（ERNIE Bot）

苹果：引入AI加速器设计的M2系类芯片（M2 Pro&M2 Max）将被搭载于新电脑

OpenAI：随着ChatGPT的使用量激增，OpenAI需要更强的计算能力来响应百万级别的用户需求，因此增加对英伟达GPU需求

AMD：计划推出于苹果M2系类芯片竞争的台积电4mm工艺"Phoenix"系列芯片，以及使用Chiplet工艺设计的"Alveo V70"AI芯片，这两款芯片均计划在今年推向市场，分别面向消费电子市场以及AI推理领域

□ 超万卡集群建设步伐加快

✓ **大模型驱动智能算力爆发增长，超万卡集群建设成为显著趋势：**以GPT-4为例，该模型包含16个专家模型，共计1.8万亿参数，其单次训练任务需在约25,000个A100 GPU上进行90至100天的训练。领先的AI公司正积极部署千卡至超万卡规模的计算集群，例如：谷歌推出了配备26,000块英伟达H100 GPU的超级计算机A3 Virtual Machines；Meta在2022年推出了配备16,000块英伟达A100 GPU的AI研究超级集群AI Research Super Cluster。

✓ **硬件形态转变呼吁软件同步迭代以实现高效的深度学习模型训练：**在超万卡集群环境下，数据规模和复杂度的提升要求软件和框架进行持续创新，具体包括：1）针对以裸金属和容器为主的集群资源，需要进行有效的管理和优化，以支持应用跨架构迁移和异构混合训练的能力；2）在应用场景中，基于现有框架的能力，不断研发自动化的分布式训练框架，并积累数据服务、模型仓库等工具集；3）在运维领域，通过专用软件支持计量计费、任务调度和租户隔离等功能。

来源：弗若斯特沙利文

以处理器为代表的硬件形态进步驱动软件形态更新（2/2）

- 面对大型模型训练过程中的硬件集成和软件兼容性挑战，以及训练周期长和成本高昂的问题，我们需要提升分布式训练的效率、优化算力资源的调度，**发展超万卡规模的集群及构建高效的算力平台。**
- 算力平台的建设不仅仅是将服务器、存储和网络等硬件设备进行简单的集成，它还涉及到这些设备之间软硬件的兼容性和性能优化。**采用全栈全流程的软件架构有助于实现资源的统一管理和弹性扩展。**

GPU

- GPU相对于CPU而言更满足AI时代的算力要求——AI模型训练所需算力呈现爆发式增长

超万卡集群

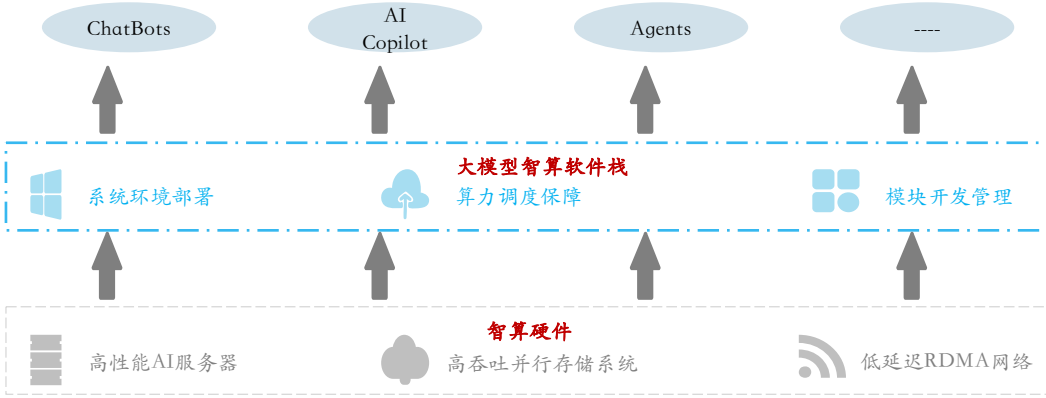
- 大模型驱动智能算力规模需求爆发增长，超万卡集群建设成为显著趋势

万卡集群训练所面临的挑战

- **训练周期长**：训练过程比传统的分布式训练复杂，训练周期长达数月。
- **成功概率低**：集群计算效率低、故障频发且处理复杂，会导致训练中断后不能及时恢复，降低成功概率。
- **成本高**：超万卡集群建设成本高，导致AI模型训练成本居高不下。

> **AI集群就像是一台动力强劲的法拉利，但目前为止我们还没有很好的掌握如何驾驭它**
>
> -----某科技公司IT架构总监

软件的优化和升级以应对大模型的挑战



- 超万卡集群下，模型训练规模、数据集复杂度指数级提升，**需要在软件方面持续创新，聚焦于自动化、跨平台支持、大规模模型训练、跨集群训练、边缘训推等方面不断优化完善，实现高效、可靠和快速的深度学习模型训练**，提高模型的准确性和训练效率，降低用户开发大模型的使用门槛和资源消耗，提供更加高效、易用的模型训练工具。

- **软件依赖于物理计算资源（如CPU、GPU、内存和存储等），因此需要对这些资源进行有效的统一管理和监控。**我们建议智算中心操作系统采用标准化的扩展接口，以实现数据安全保护和网络安全防护的深度整合。智算中心操作系统提供适合生成式AI应用的无损裸机算力环境，确保计算过程中的零损耗，并最大化地发挥CPU和GPU的算力，以及提供高性能的裸金属网络支持。

来源：弗若斯特沙利文

立足国家战略，智算中心建设赋能新时代

截至2024年5月中旬，国内大模型项目中标数量远超2023年全年水平，已发布超

230个

国家积极加强智算中心建设

在当前迅速发展的数字化时代，数字基础设施已经成为推动经济增长、改善社会治理和提升民众福祉的关键因素。**人工智能技术的快速进步，使得建设智能时代的新型基础设施——智算中心，成为战略布局的重中之重。**国家正全力推进智算中心的建设，以加速数字经济和数字社会的发展。

智算中心，作为数字中国的核心基础设施和未来发展的基石，对于促进数字化转型、智能化升级以及实现可持续发展目标扮演着至关重要的角色。这些中心的建设不仅满足了日益增长的计算需求，提供了强大的算力支持，而且还能应对复杂的数据分析和机器学习挑战。智算中心已成为技术创新的催化剂和产业升级的引领者，为社会经济的全面提升提供了强大的动力。

目前，国家已在多个关键领域和地区布局建设了一批智算中心。**据最新统计，全国已有超过30个城市启动或规划了智算中心的建设，项目总数超过100个，总投资规模达到百亿级别。**建设主体包括各级政府机构、国内主要电信运营商以及多家领先的互联网企业。

国家指导政策推动智算中心建设布局

智算中心

大模型训练

超大规模新型智算中心。加快突破GPU芯片、集群低时延互连网络、异构资源管理等技术，建设超大规模智算中心，满足**大模型迭代训练和应用推理需求**。

——《关于推动未来产业创新发展的实施意见》

来源：工业和信息化部等七部门 2024.01

数字经济

基础设施建设

系统算力是集信息计算力、网络运载力、数据存储力于一体的新型生产力，主要通过算力基础设施向社会提供服务。**算力基础设施**是支撑**数字经济发展**的重要资源和基础设施，呈现多元泛在、智能敏捷、安全可靠、绿色低碳等特征，对于实现数字化转型、培育未来产业，以及形成经济发展新动能等方面具有重要作用。

——《算力基础设施高质量发展行动计划》

来源：工业和信息化部等六部门 2023.10

算力优化

东数西算

系统优化**算力基础设施**布局，促进东西部算力高效互补和协同联动，引导**通用数据中心、超算中心、智能计算中心、边缘数据中心**等合理梯次布局。

——《数字中国建设整体布局规划》

来源：中共中央 国务院 2023.02

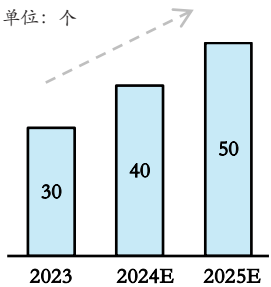
2023年底，全国带有“智算中心”相关项目有共有

129个

中国智能计算中心建设指标

2023-2025E

单位：个



各省份激励智算中心产业发展

各省份相关政策（2023-2024）	来源单位名
《关于加快人工智能产业发展的指导意见》	浙江省
《河南省重大新型基础设施建设提速行动方案（2023—2025年）》	河南省
《2024年项目投资攻坚年行动方案》	宁夏回族自治区
《2024年〈政府工作报告〉重点工作责任分工方案》	贵州省
《关于加快推进数字经济创新发展的若干措施》	山东省
《关于加快建设通用人工智能产业创新引领地的实施意见》	广东省

来源：弗若斯特沙利文

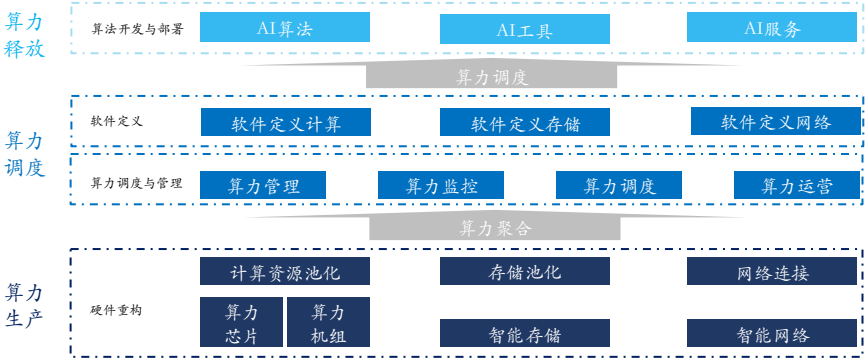
章节二：智算中心市场概览

智算中心定义、特征及演进

- 智算中心是基于前沿人工智能理论，采用尖端的人工智能计算架构，为人工智能应用提供所需的算力服务、数据服务和算法服务的新型算力基础设施。
- 智算中心以其技术领先和算法融合等特性脱颖而出，其发展经历了从传统的物理数据中心，到互联网数据中心，再到云化数据中心，最终演进至当前的智算中心的连续升级过程。

智算中心定义

- 智算中心是基于前沿人工智能理论，采用尖端的人工智能计算架构，为人工智能应用提供所需的算力服务、数据服务和算法服务的新型算力基础设施。
- 智算中心通过算力的生产、聚合、调度和优化配置，有效地支持数据的开放共享、智能生态的构建、产业创新的集聚，从而显著推动人工智能产业化、产业智能化以及政府治理的智能化进程。



- **算力释放的价值最大化**
  企业应能够为AI类产品的应用开发提供优越的开发环境，包括模型赋能、开发工具以及开发支持、安全保障等服务，以最大化算力释放价值。
- **算力调度的效率和成本**
  企业应追求算力调度的效率最优化和成本最优化，实现资源的最佳配置。
- **算力资源的规模范围**
  企业在生产资料的投资建设中应有规模化的布局，确保算力聚合和获取的稳定性、可靠性和安全性。

智算中心特征

- **技术先进**：智算中心基于最新的人工智能理论，采用前沿的计算架构，确保技术领先性和创新性。
- **服务公共性**：作为公共算力基础设施，智算中心提供算力、数据和算法服务，服务于社会和行业需求。
- **资源高效聚合**：智算中心高效地管理和调度算力资源，支持数据开放共享和智能生态建设。
- **算力与数据融合**：智算中心结合算力、数据和算法，驱动AI模型对数据进行深度加工，支撑AI技术的应用和演进。

智算中心发展历程

主要指电信企业面向大型企业提供的机房，包括场地、网络、通信设备等基础电信资源和设施的托管和线路维护服务。

围绕主机托管提供数据存储管理、安全管理、网络互连、出口带宽网络服务等等，这一阶段的数据中心由互联网企业自行搭建。

数据中心从分散在各地走向集中式，一般都是科技巨头搭建的大型化、虚拟化、综合化数据中心，如今云化数据中心已被普遍应用。

智算中心是对数据中心的虚拟化，它可能是多个云数据中心协作起来提供算力，也有可能只是云数据中心中的部分机组构成。

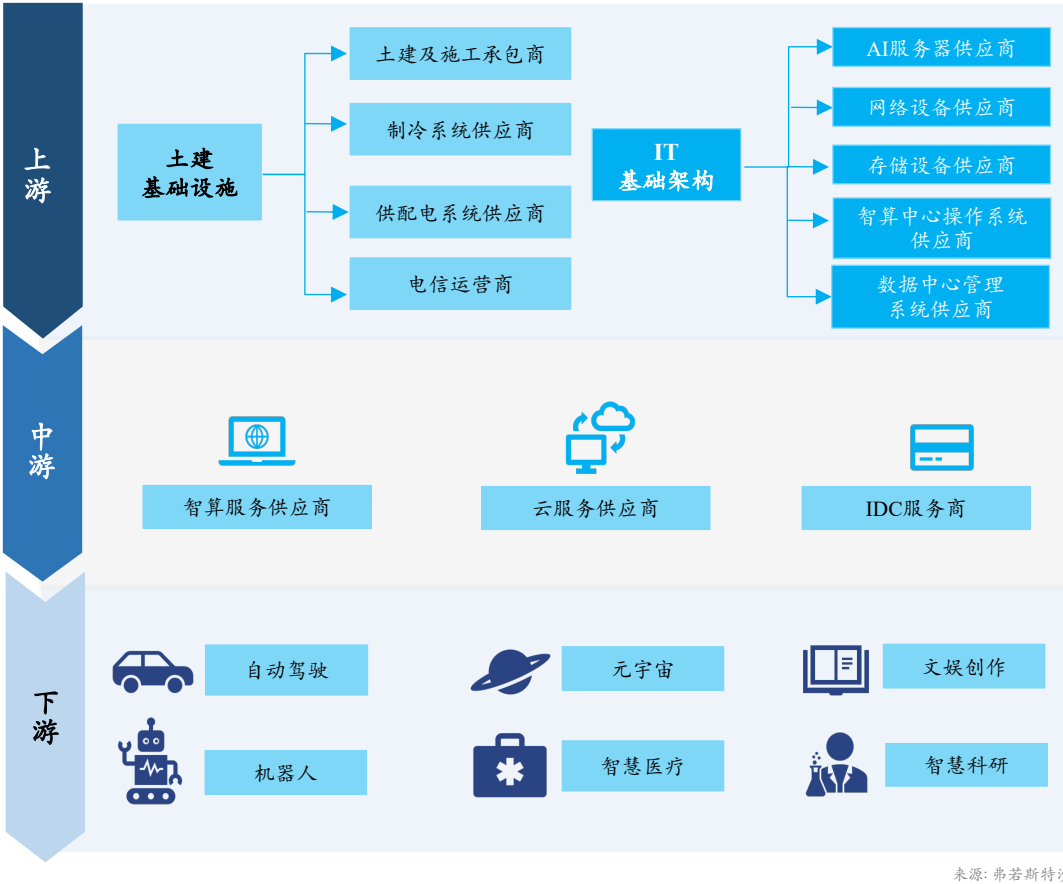
来源：弗若斯特沙利文

智算中心产业链介绍

关键发现

- **智算中心拥有专业人才资源、创新资源，为形成产业链协同发展提供了基础条件。**智算中心以建设配套产业园区等形式，打造联动上中下游产业链，提供优质的环境，促进了上游供应商、中游服务商以及下游相关产业的协同发展。
- **协同发展的产业链有效加强了IT基础架构和土建基础设施供应商等上游参与者，以及智算服务供应商、云服务供应商等中游相关产业的协同，并服务于互联网、金融、交通等行业用户的人工智能应用需求。**同时，实现了对自动驾驶、机器人、元宇宙、智慧医疗、文娱创作、智慧科研等下游相关产业带动。

智算中心产业链分析



来源：弗若斯特沙利文

智算中心价值分析

- **智算中心已经成为提升国家国际竞争力的关键基础设施，并且是数字经济高质量发展的重要支撑。**它们通过提供强大的计算能力和数据处理能力，支持了人工智能技术的研究和应用，从而加速了AI产业化的进程。
- **智算中心通过推动核心技术的发展和创新融合，为人工智能技术的高速发展提供了动力。**这些中心不仅促进了AI技术在多个行业的应用，还助力政府治理的智能化，提高了公共服务的效率和质量。

智算中心的具体作用

推进AI产业化

- **智算中心已成为推进AI产业化的重要引擎。**它提供人工智能专用的算力服务、数据服务和算法服务，通过集中生产、聚合和调度算力资源，为人工智能创新与关键技术研发提供强有力的支撑。
- **智算中心为企业、高校等各类研发机构创造了低成本、高可靠性、便捷简易的人工智能应用开发与场景试验条件**，从而促进了人工智能技术的成果转化和应用推广。

赋能产业AI化

- **智算中心作为数智化基础设施**，赋能产业数字化和智能化发展，带动一二三产业提质增效。
- **智算中心的孵化功能为新兴业态提供了肥沃的土壤。**它们通过提供必要的基础设施支持，促进了智能网联汽车、数字交易、元宇宙等新兴业态的快速发展。智算中心通过为日益丰富和个性化的数字产品、服务的研发提供关键资源，加速了创新技术的商业化进程和市场应用。

助力治理智能化

- 利用智算中心提供的计算平台、框架和底层算法，**可以加速业务与人工智能技术的融合**，包括计算机视觉、自然语言处理和机器学习等领域。
- 这种融合在政务服务、健康医疗、教育科研等公共服务领域，以及交通治理、灾害预警、生态环境监测等城市治理领域，**极大地推动了公共服务的便捷化，提升了公共治理的效率**。

促进产业集群化

- 智算中心通过与高校、科研院所以及各行业领先企业的紧密合作，提供创新孵化、技术转移和人员培训等服务，**致力于培养数据处理、数据分析和运营服务等领域的人才和企业。**
- 利用人才优势、产品优势和产业链资源的整合，智算中心不仅能够推动人工智能产业的发展，**还能构建技术研发与商业应用的双向闭环，进而促进地区人工智能产业的集群化发展**。

智算中心建设的宏观意义

➢ 智算中心已成为提升国际竞争力的关键基础设施

算力在新科技革命和产业变革中发挥着基础支撑作用，是提升国际竞争力的关键因素，也是衡量综合国力的重要指标之一。智能算力水平不仅体现了国家在智能化和数字化方面的发展水平，而且构成了数字化应用建设和发展的坚实基础。

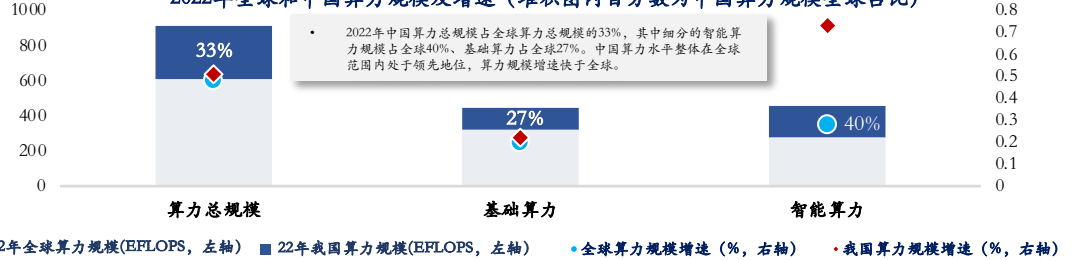
➢ 智算中心建设已成为数字经济高质量发展的重要支撑

算力作为数字经济浪潮中的关键技术，将带来深远的变革和显著的赋能效应。在数字化和智能化转型加速的背景下，数据总量呈现出爆发式增长。智算中心的建设不仅将为AI产业化发展提供强大动力，而且将推动整个产业向AI化转型。

➢ 智算中心已成为构建智慧城市的必然选择

泛在算力是智慧城市建设的基石。智算中心作为公共算力基础设施，为政府、行业、企业和公众等广泛的用户群体提供智慧城市管理和运行所需的算力、数据和算法服务。

2022年全球和中国算力规模及增速（堆积图内百分数为中国算力规模全球占比）



- 2022年中国算力总规模占全球算力总规模的33%，其中细分的智能算力规模占全球40%、基础算力占全球27%。中国算力水平整体在全球范围内处于领先地位，算力规模增速快于全球。

来源：弗若斯特沙利文

国内智算中心建设现状

- **从算力部署的规划角度来看，在各地政府的统筹规划和引领下，智算中心的建设有望迎来新一轮的建设高潮。**与以往的数据中心建设相比，本轮新基建的建设将更加重视建设与运营的一体化、能耗的合规性、成本的有效控制以及安全性和可信度。此外，本轮建设还将注重算力部署的运载和存储需求，以及与应用需求的协同发展，以建设更加高效和可持续的智算生态系统。

□ 统筹规划以四大主要目标为导向，引领新一轮新基建建设高潮

高速扩张阶段（2017-2021年）

区域政府和企业"各自为战"，缺少统筹规划

行业引领阶段（2022年至今）

发布《算力基础设施高质量发展行动计划》，统筹规划

2017

产业萌芽期
智算中心首次在《新一代人工智能发展规划》被提及

2019

发展探索期
多地积极布局智算中心建设，抢抓智能发展高地

2020

快速扩张期
智算中心纳入新基建，激活各地投建浪潮

2022

政策引导期
国家政策引导力度增强，基于"东数西算"强化算力资源统筹和智能调度；建设"全国统一大市场"，强调资本、数据、技术流通开放

2023

规划落地期
提出4方面建设目标；明确多元泛在、智能敏捷、安全可靠、绿色低碳化要求

《算力基础设施高质量发展行动计划》2025年4大主要目标

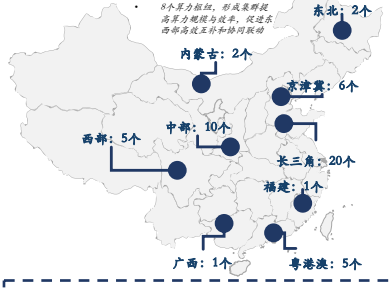
01 **计算力** 算力规模超过300EFLOPS

02 **运载力** 国家枢纽节点数据中心集群间基本实现不高于理论时延1.5倍的直连网络传输

03 **存储力** 存储总量超过1800EB，先进存储容量占比达到30%以上

04 **应用赋能** 围绕工业、金融、医疗、交通、能源、教育等重点领域，各打造30个以上应用标杆

全国算力枢纽构建一体化算力网络



行业引导2.0阶段关键发展特征

- **算力融合** 加速芯片异构组合，实现超异构计算
- **软硬协同** 实现产业链要素与基础层、框架层及应用层的"横纵向"融合
- **投建运一体化** 投建运一体化布局
- **能耗低碳** 利用液冷、储能、蓄冷、算力调度及智能运维加速绿色节能建设
- **成本优化** 充分释放每P算力的价值成效，促进普惠化发展
- **需求牵引** 更多以市场化、平台化模式运营
- **安全可信** 考虑信息安全和产业安全，重视数据存储安全以及安全运维能力

□ 智算中心建设以一线和新一线城市为主，逐步向地级及以下城市渗透

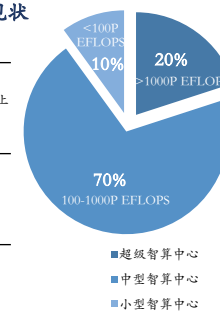
主要合作方

目前建设运营模式主要有政府投资建设、企业建设运营、政府购买服务、政府和社会资本合作等模式，主要产业合作方如下：

- **云厂商及AI科技企业**
  具备智算中心操作系统和AI算法的开发部署能力
- **三大通信运营商**
  拥有全国布局的网络基础资源、数据中心资源和云资源池
- **第三方数据中心厂商**
  具备高标准等级的算力硬件和运维能力，为高性能算力和节能方案提供支撑

中国智算中心格局分部现状

	现有项目占比及概况	主要应用领域	投资规模
超级智算中心	项目占比约20% 主要分布在京津冀区域、长三角和珠三角区域	面向科研和科学计算场景提供支撑服务：为AI大模型、自动驾驶、空间地理等人工智能探索应用提供强大的智能算力服务	百亿元及以上
中型智算中心	项目占比约70% 分布在一线、新一线及二线城市	主要服务于人工智能产业链或细分行业智能化集群	10亿元以上
小型智算中心	项目占比约10% 主要分布在二线及以下城市	主要服务于企业的智能化建设或小型人工智能产业集群建设	10亿元以下



来源：弗若斯特沙利文

全球智算中心建设现状

- **全球的智算中心正在以迅猛的速度发展，**美国、欧盟、新加坡等国家和地区均高度重视智算中心的发展。这些中心通常是由政府、企业和学术机构共同建立的，以促进人工智能（AI）的研究和应用。它们提供了强大的计算资源和专业知识，为AI领域的创新和发展奠定了坚实的基础。此外，全球智算中心的建设正日益呈现出低碳化和可持续化的特征。

□ 美国、欧盟及新加坡智算中心建设情况分析

➢ 建设情况　　➢ 发展模式

美国

- **三大智能超算中心体系** 1. 美国能源部的六大国家实验室；2. 美国国家科学基金会支持的高校智能超算中心；3. NASA Ames研究中心超级计算中心
- **全球领先的智能计算中心** 美国在智能计算领域处于国际前列。美国能源部的智能计算中心在构思、设计、建造及运维方面全球领先。能源部下属的橡树岭国家实验室、劳伦斯·利弗莫尔国家实验室与IBM、NVIDIA公司合作成立了超级计算机卓越实验中心。
- **新一代HPC计算机的开发** 这些中心联合开发新一代HPC计算机，使用IBM Power处理器和NVIDIA Tesla加速卡，浮点性能达到至少10亿次，最高可达30亿次。
- **广泛的研究领域** 这些计算资源主要用于核武器、核安全、天文、能源、气候、宇宙、新能源等领域的研究。

政府主导+多维度机构配合共同发展智算并注重绿色环保

□ **高校及研究机构** 1.国家科学基金会支持的研究机构，如伊利诺伊大学，提供顶级超级计算资源。2. 国家超级计算应用中心（如主机Abe）和德克萨斯高级计算中心（如主机Stampede）构成了主要的计算资源布局。

□ **NASA Ames研究中心** 应用于天文、航天、气候、军事等领域，支持太空任务模拟和气候模型预测。

□ **企业智能计算中心** 谷歌实现24/7无碳能源目标，利用碳智能计算平台匹配风能和太阳能。

□ **政府主导模式** 1.美国能源部投资18亿美元建设3个E级超算中心（Aurora、El Capitan、Frontier），以支持人工智能研究。2. 白宫发布《人工智能倡议首年年度报告》，指导联邦机构优先分配资源以推进AI研究。

欧盟

- **GAIA-X项目** 2020年10月发布，2021年上线。目标是减少对国外云厂商的依赖，提供强大、安全的数据存储基础设施。
- **数字欧洲计划** 2019年启动，预算75亿欧元，2020年增加至92亿欧元。**主要投入：** 22亿欧元用于超级计算，21亿欧元用于人工智能。**具体措施：** 2021年收购百万兆级超级计算机，建立AI数据空间，部署泛欧量子通信基础设施，支持网络安全认证和AI、高级计算、网络安全计划。
- **欧洲高性能计算联合执行体** 2020年9月发布新章程，投资80亿欧元。**重点：** 支持百亿次计算和量子计算的研究与创新，打造世界级超算生态。**聚焦领域：** 基础设施、超算服务、技术、应用和技能培训。

政策高度关注打造智能计算生态环境

□ **《欧洲高性能计算共同计划（EuroHPC）》** 2019年启动，预算8.4亿欧元，建设8个超级计算中心。这些中心将服务于医疗、药物设计、生物工程、天气预报和气候变化等多个领域。

□ **依赖全球科技巨头**
合作伙伴包括亚马逊、微软、谷歌、IBM等，为智算中心提供了强大的技术支持。此外，阿里云和腾讯云等亚洲科技巨头也参与其中。西门子、雷诺、德意志银行、汉莎航空、大众汽车等欧洲知名企业也利用这些云服务来提升其业务效率和创新能力。

□ **绿色环保**
2020年启动GREENDC项目，降低数据中心碳排放。2021年25个数据中心服务商和17个协会签署协议，目标2030年实现能源中立。

新加坡

- **建立国家超级计算中心** 新加坡于2016年建立了国家超级计算中心，提供国内最先进的计算、数据和资源设施，以支持用户解决各类科学技术问题。该中心通过高速宽带网络连接各种资源，确保用户无论身处何地都能享受到高速的使用体验。此外，中心还积极促进各行业利用超级计算资源来解决复杂问题、进行试验设计和优化技术。
- **联合德国在新建立英飞凌人工智能中心** 在2020至2023年间，德国英飞凌半导体公司投资2亿美元在新加坡建立了全球首个人工智能中心。该公司计划在2023年前启动25个人工智能项目，并致力于培养掌握人工智能技能的高端人才。

政府主导国家型智算建设
联合外国科技企业创新发展智算

□ **政府主导建设智算中心**
政府主导建设的智算中心为新加坡的科学研究和技术创新提供了坚实的基础，支持了多个行业和学术领域的发展，并推动了本地创新生态系统的繁荣。

□ **合作推动智算创新发展**
在合作推动智算创新发展方面，新加坡的创新机构、初创企业、高等学府及研究机构正利用英飞凌提供的丰富数据资源来开发创新解决方案。

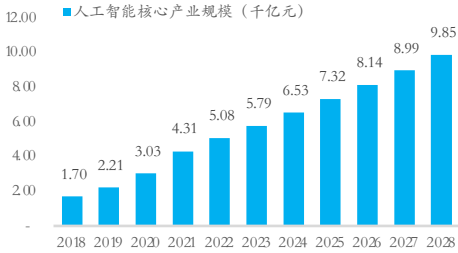
英飞凌与新加坡国立大学系统科学院的合作，致力于培养新一代的科技人才和创新者。通过培训和合作，这些努力不仅推动了本地创新生态系统的发展，也增强了新加坡在全球科技领域的竞争力。

来源：弗若斯特沙利文

智算中心效益分析

- **智算中心经济效益显著：**作为数字经济的核心驱动力，预计到2025年，人工智能核心产业的规模将超过7000亿元人民币，到2028年将接近1万亿元人民币。智算中心作为重要的智算基础设施，不仅加速了科技研发和技术迭代，还优化了数字经济的生态系统，推动了产业规模化发展。

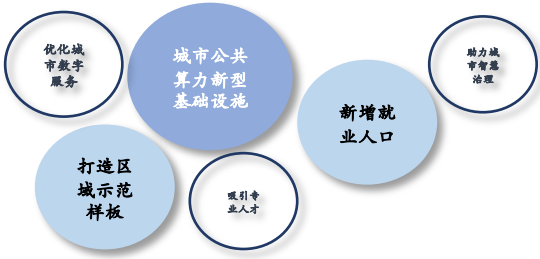
智算中心经济效益分析



- **推动人工智能核心产业发展：**根据沙利文数据的统计，2023年中国人工智能核心产业的规模为5787亿元人民币，预计到2025年将达到7316亿元人民币，期间的复合年增长率（CAGR）为14.2%。这一增长预计将带动相关产业规模达到约5万亿元人民币。展望2028年，人工智能核心产业的规模预计将达到9854.4亿元人民币，十年间的复合增长率达到19.2%。
- **推动智算产业集群化及相关技术创新发展：**智算中心作为信息技术的集成载体，推动通信、大数据、人工智能等技术的快速迭代，促进技术创新。研究表明，在“十四五”期间，当智算中心的应用水平达到80%时，城市或地区在智算中心建设上的投入增长量可对创新产出贡献约14%至17%。
- **带动相关产业快速发展：**智算中心支撑人工智能产业的快速发展，预计到2025年，人工智能核心产业的规模将达到7316亿元人民币，带动相关产业规模达到5万亿元人民币。在“十四五”期间，智算中心若达到80%的应用水平，其投资有望使人工智能核心产业增长2.9至3.4倍，相关产业增长36至42倍。
- **支撑优化数字经济发展生态：**智算中心不仅提供公共算力基础设施，还充当产业的“聚合器”和“孵化器”，有效促进人才、资金、技术、数据等要素的联动。通过构建跨行业领域的沟通平台，整合政府、企业、研究院所等多方资源，推动更多人工智能应用场景的实现。这种综合作用有助于避免各自为政和产业链分割的问题，对于吸引数字企业、培育市场主体、推动数字经济的规模化发展具有关键性作用。

智算中心社会效益分析

智算中心在支持城市公共服务和治理的智能化应用方面发挥着关键作用。它通过提供计算平台、框架和算法，促进了政务、健康、教育等公共服务领域的智能化发展，同时也推动了交通治理、灾害预警、环境监测等城市治理领域的智能化进程。



- **提供创新就业机会：**智算中心的建设和运营为数字经济和智能城市建设提供了新的就业机会。智算中心融合了计算机视觉、自然语言处理和机器学习等人工智能技术，实现了公共服务的便捷化和城市治理的精准化，从而提升了公众的幸福感和获得感。
- **吸引专业人才和技术：**智算中心作为高科技研发和应用的核心，吸引了大量高端技术人才和专业人士。中心需要大量具备人工智能、数据科学、机器学习等专业技能的工程师和研究人员，这促进了相关领域的就业增长。培养和吸引这些专业人才不仅增强了当地的技术创新能力，还提升了整体产业的竞争力。
- **推动构建科技创新生态环境：**智算中心在地方和国家层面都是科技创新的重要驱动力。它为创业者、研究人员和企业家提供了平台和资源，促进了科技创新生态系统的形成和发展。智算中心的技术创新和服务不仅促进了当地产业的转型升级，还带动了相关产业链的发展和壮大。

来源：弗若斯特沙利文

智算中心面临挑战

- 当前，算力标准化进程缓慢，尚未形成统一的标准化供应体系，导致智算中心的算力使用和流通未能形成一个有效的闭环，算力的互联互通和全面商业化尚未完全实现。
- 此外，算力资源的抽象功能尚未完全实现，缺乏完善的基于抽象的算力生成、传输和消费工具，这使得下游用户难以便捷地使用算力。

标准化挑战：算力基础设施化及算力互联互通、完全商业化未真正实现

标准不一

- **当前算力标准化进程缓慢，导致市场上缺乏统一的算力供给标准。这种情况使得智算中心在提供服务时难以实现标准化的算力计价和供给模式。**不同智算中心可能采用各自的计价方式和标准，这给下游客户在购买和使用算力时造成了困扰和不便。

流通不畅

- 由于缺乏标准化，算力在应用和流通过程中未能形成有效的闭环体系。这种情况导致算力资源的利用效率降低，并且增加了算力服务的成本与复杂性。

基建落后

- **中国智算基础设施的建设也受到了地方竞争和数据中心建设与应用不足的矛盾影响，导致一些新建的算力中心可能面临收入难以覆盖成本的问题。**跨域调度和跨中心调度也面临商业上的利益冲突和技术上的异质互联问题，这限制了算力资源的有效利用。

抽象化难题：算力资源抽象的功能尚未实现

- **算力资源抽象技术理念**：该理念旨在通过新技术抽象化算力资源，使用户在开发流程中无需感知底层算力资源的复杂性。通过建立一个支持多种异构算力资源的统一开发框架，实现对这些资源的开发和编译工作，从而解决异构算力资源开发生态的隔离问题、应用迁移难题、算力资源管理的复杂性以及资源利用率低下等问题。
- **算力资源抽象功能尚未实现**：目前市场上缺乏能够有效将底层硬件资源和网络架构抽象化，使之易于理解和操作的工具和服务。这种抽象化的不足导致下游用户难以根据自身需求直接和便捷地调配及使用算力资源。从算力资源的生成（资源的准备与配置）、传输（数据和计算任务在不同节点间的有效调度）到消费（用户端的实际应用），整个流程中的工具支持尚不充分，这限制了算力资源价值的充分发挥。

大模型构建瓶颈：缺乏有力的工具和平台进行支撑

技术与资源门槛高

- **技术挑战：**随着模型规模的扩大，设计高效且可扩展的算法与模型架构变得至关重要。这涉及到减少过度拟合、增强模型的泛化能力、以及提升模型训练的效率等多个方面。对于大多数企业（尤其是中小企业）而言，这些都是难以克服的难题。
- **资源需求**：开发大规模机器学习或深度学习模型不仅需要处理大量数据，还需要依赖高性能的计算资源和深厚的算法专业知识。

工具与平台支持匮乏

- **缺乏集成开发环境**：目前市场上缺乏集成数据处理、模型训练、优化与部署等全链条支持的综合性开发平台，这限制了大模型构建和应用的普及，使其成为少数技术巨头的专属领域。
- **模型可解释性与安全性**：大模型的“黑箱”特性增加了解释其决策过程的复杂性。此外，确保模型安全，包括防止数据泄露和抵御模型攻击，也是技术难题。

来源：弗若斯特沙利文

章节三：智算中心操作系统价值

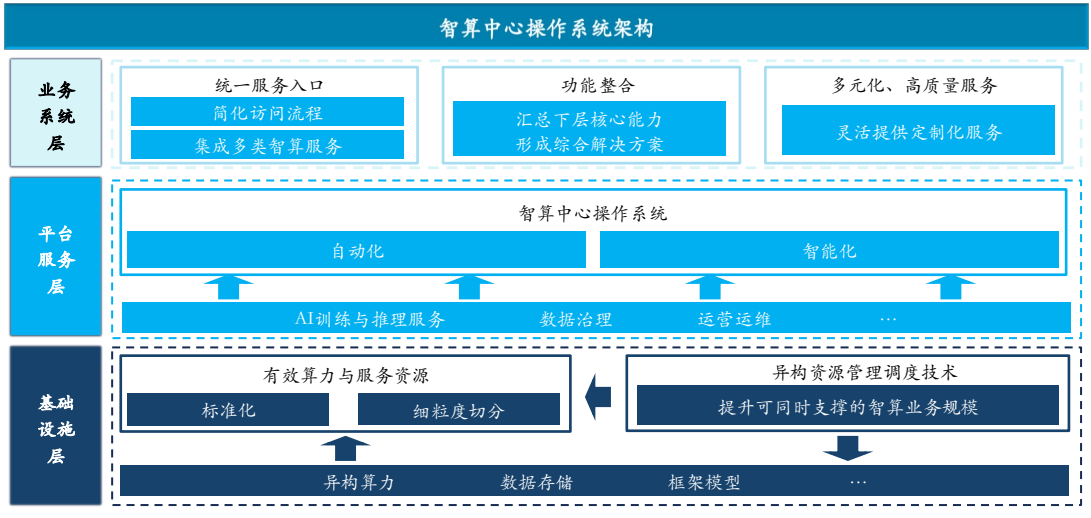
智算中心操作系统是智算中心的“中枢神经”及能力底座

- 智算中心操作系统是专门针对智算服务需求而设计的，它对智算中心的基础设施资源池进行高效管理和智能调度。智算中心操作系统紧跟GPU硬件的发展和形态升级趋势，优化硬件资源的利用和管理，从而提升算力调度的能力和效率；并进一步实现了智能化的管理和运维，为智算中心的长期可持续发展提供了支持。

智算中心操作系统定义

智算中心操作系统紧跟GPU硬件形态的升级趋势，实现了对硬件资源的有效利用和管理，从而提高了算力调度的能力和效率。它通过智能化的管理和运维，为智算中心的可持续发展提供了支持。

- **智算中心操作系统是专门针对智算中心基础设施资源池的高效管理和智能调度解决方案。**该系统使智算中心能够更有效地对外提供算力、数据、算法和智能组件等服务，显著降低了算力使用的门槛，并提升了资源调度的效率。它支持各种智慧应用场景的实施，是智算中心的“中枢神经”。



智算中心操作系统特征

- ✓ **开放性：**智算中心操作系统的开放性特征体现在采用开源技术构建，支持标准化和模块化设计，提供开放API接口，兼容多种硬件资源，融入云原生技术，并依托活跃社区和合作伙伴网络，共同推动资源高效利用和技术创新，为AI发展提供灵活、可扩展的基础设施。
- ✓ **融合性：**智算中心拥有多样化的算力、应用和场景。这些元素融合于一个数据中心内，需要智算中心操作系统进行有效的调度和管理。这包括实现可变粒度的资源池化管理、统一调度云数智应用，以及满足跨云、灾难恢复、多云管理和云边协同等需求。
- ✓ **敏捷性：**业务的快速发展和前端业务需求的快速变化对后端基础设施提出了更高要求。传统的瀑布式研发、生产和标准化交付模式已无法满足当前业务对IT的需求。现在，我们需要以更敏捷的方式创造计算力，实现快速迭代、快速变更、快速上线和持续交付，为前端应用提供强大的动力。
- ✓ **智能化：**智算中心涵盖了算力的生产、聚合、调度和释放。在算力的生产和聚合阶段，需要充分利用各种智能加速设备，并结合数据和智能技术进行算力的调度和释放。通过智能平台提供智能服务的同时，还需融合人工智能技术，提升平台的运维和管理智能化水平，增强运维和管理能力。

来源：弗若斯特沙利文

智算中心操作系统赋能智算中心高质量发展（1/2）

- 随着大模型技术的持续进步，利用自然语言进行编程逐渐变为现实，这标志着智算中心操作系统正在经历一场革命性的变化。**系统的角色从传统的管理进程和微服务，演变为对智能化的管理。**
- **通过高效调度和管理底层硬件，系统能够兼容和调度多个异构GPU，从而提升算力效能并降低算力成本。**同时，通过充分利用底层资源管理和资源驱动机制，系统支持上层服务，有效降低了模型开发的门槛。

智算中心操作系统对智算中心发展的意义

智算中心操作系统赋能智算中心算力运行与业务运营，实现“用软件定义算力”

- **发展背景**：随着大模型技术的持续进步，利用自然语言进行编程正逐渐变为现实。编程范式将不再仅仅面向过程或面向对象，而是更加面向需求本身。这一转变使得编程过程转变为开发者表达需求和愿望的过程，进而为操作系统带来革命性的变化。在操作系统的核心层面，传统的以CPU为中心的硬件架构正逐步演变为以GPU算力为核心的架构，并且融入了由大模型压缩的世界知识。操作系统管理的对象正在发生根本性的变化，其职能从管理进程和微服务，逐步进化到管理智能实体。
- **发展原因**：智算中心操作系统能够对中心内的大规模、复杂资源以及多样化的算力进行有效的统一调度和管理。智算中心不仅仅是资源的集合和调度，而是通过“上云、用数、赋智”的策略，更高效地利用中心内的数据资源。它还强调利用算力来驱动人工智能模型和算法，以催生更多智能化的应用。
- **发展意义**：与传统的云计算操作系统不同，后者通常从硬件层面出发，通过虚拟化技术将资源切分给上层应用使用。大模型训练则需要一种自上而下的管理策略。智算中心操作系统通过有效调度和管理底层硬件，能够兼容和调度多个异构GPU，使得算力资源得到更高效的利用。同时，它通过充分利用底层资源管理和资源驱动机制来支持上层服务，确保在AI平台上对模型训练算法和参数进行精细调优，以实现最佳的训练效果。

智算中心操作系统核心价值分析

- 智算中心操作系统是依托智算中心为核心载体、所构建的一个多元化和开放的智算服务平台。该平台整合了国内外的先进人工智能技术，形成了一系列标准化、模块化的模型、中间件和应用软件。
- 通过提供开放的接口、丰富的模型库和算法包，智算中心操作系统为用户提供包括行业大模型、自动驾驶、智慧科研等在内的多项人工智能服务。这不仅推动了人工智能技术成果的开放性和共享性，还推动构建一个开放和协作的智算生态系统。

智算中心建设两大重大挑战

成本高昂

以构建1000P算力的智算中心为例，包括硬件、能源等在内的总成本需要5.5个亿，其中硬件成本达到80%以上；同时根据不同电力计价水平，一年的能源运维成本大约在1000万到1500万之间。

稳定性难以保证

随着算力集群规模的扩大，算力集群的稳定性越来越难以做到，效率也难以明显提升。

呼唤

赋能

智算中心操作系统核心价值

算力效能提升

智算中心操作系统可以不断提升算力利用效能，显著提高千行百业落地行业大模型、垂类大模型所带来的经济收益，达成规模经济。

模型开发门槛降低

智算中心操作系统可以有效降低模型开发门槛，提供模型的全生命周期管理能力，让用户无缝、无感地将智能算力使用起来。

算力成本降低

智算中心操作系统通过优化资源调度和提高资源利用率，以及有效增强智算中心的运维稳定性，能够显著降低算力成本。

来源：弗若斯特沙利文

智算中心操作系统赋能智算中心高质量发展（2/2）

- 智算中心操作系统的设计目标是实现高效且全面的资源管理，它不仅要能够兼容并调度多种不同架构的GPU，而且关键在于能够在人工智能平台上对模型训练算法和参数进行精细调优。
- 智算中心操作系统的核心价值体现在六个方面：异构算力的集中管理、算力资源的智能调度、AI应用的加速优化，以及其他三个关键领域，共同构成了其全面的服务能力。

□ “智算”雄起，智算操作系统要“顶天立地”

顶天

技术层面

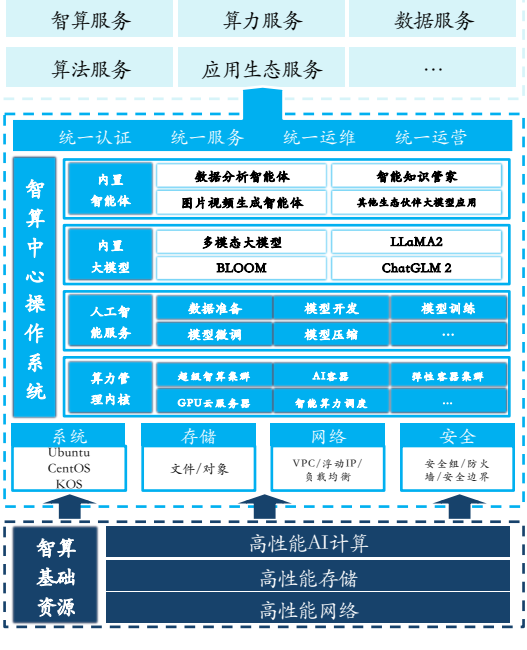
- **兼容调度：**高效管理异构计算资源，如GPU、TPU等，实现资源的智能调度和优化，最大化资源利用率和计算效率。
- **自动调优：**运用自动化技术优化算法参数，实现模型训练的高效与精准，探索出更加高效和准确的模型结构。
- **资源管理：**动态调整资源分配，确保计算效率，同时支持多租户的公平使用，保障各用户项目的独立性和优先级。

应用层面

- **自动建模：**提供低代码工具，简化模型构建流程，从而降低AI应用的门槛，以支持非专业人员的AI模型开发。
- **算法组合：**允许用户灵活组合算法，创新模型架构，提升训练效果，帮助用户快速获得高性能的定制化模型。
- **用户体验：**设计友好界面与文档，简化模型训练和部署，增强错误检测与调试能力，提高模型开发的效率和质量。

立地

□ 智算中心操作系统直面挑战，持续创造价值



异构算力纳管

- 提供对国内外多种算力资源的适配、纳管、调度和优化。通过实时监控和管理各个智算中心的设备，结合创新的调度算法，实现跨智算中心的大规模算力资源的智能调度，显著提高了算力资源的利用率。

智能算力调度

- 通过对不同应用场景下算力资源利用率的深入分析，可以更加合理地分配算力资源。基于此，不仅能够提高整体的算力利用率，减少资源浪费，还能显著提升数据处理速度，并有效降低延迟。

灵活扩展与成本效益

- 利用云原生架构的弹性扩展特性，可以动态调整计算资源，从而有效降低运营成本。同时，采用分布式训练与推理技术，能够高效处理大规模数据集，进一步提升整体工作效率。

加速AI优化

- 全面兼容主流人工智能开发工具及框架，并提供集成开发环境（IDE），显著提升开发效率。配备丰富的预训练模型和算法库，结合高性能计算资源，有效加速大规模模型的训练与推理过程。

算力配额管理

- 统一算力服务的计量单位，并通过这一标准实现算力的计量和计费，确保用户能够享受到“即买即用”的便捷算力服务。这为未来算力资源的互联互通奠定了坚实的商业基础，促进了算力市场的健康发展。

智能模型与数据管理

- 自动化模型管理涵盖了版本控制、部署和监控等关键环节。此外，系统具备强大的数据处理能力，包括数据清洗、特征工程、标注等处理步骤，能够适应结构化和非结构化数据的需求。

来源：弗若斯特沙利文

智算中心操作系统赋能算力租赁商业模式变革（1/3）

- 传统的算力商业模式主要包含算力租赁和算力购买。算力租赁包含整台服务器租赁、算力规模租赁和单张GPU租赁的模式；算力购买则主要包含裸金属采购。
- 在传统的算力商业模式下，可能会产生一系列问题，如：算力预算的局限性、资源缺乏弹性、资源浪费严重以及资源整合上的难题等。

传统算力商业模式

算力租赁模式

整台服务器租赁

- 租金以每台服务器为单位按月计费，这种模式适合于那些需要长期稳定使用大量CPU资源的客户。此外，**这种计费方式还具有较低的服务器运维成本，为客户提供了成本效益。**

算力规模租赁

- 租金按年度计量，允许客户根据其算力需求选择最合适的算力规模，以实现资源配置的灵活性。**这种模式有助于降低算力的总体成本，同时提升算力的使用效率。**

单张租赁GPU

- **租金是按照每张GPU每小时来计算的，这种计费模式非常适合需要短期或临时性算力支持的客户。**用户可以根据自己的实际需求灵活地调整租赁GPU的数量。

算力购买模式

裸金属采购

- 企业或个人可以选择直接购买服务器硬件，并负责将其安装在自有的数据中心或通过第三方进行托管。**这种模式通常适用于那些对长期稳定运行有严格要求的关键任务型应用。**

传统算力商业模式面临的问题

1 **算力预算有限**

- 随着大模型技术的快速发展，企业在投资智能计算服务器及其相关配套设备时，必须综合考量其高昂的购置成本、维护费用，以及如何有效平衡公司的收支状况。

2 **资源弹性不足**

- 在大模型的整个生命周期中，算力需求在不同阶段会出现较大的波动。在传统的算力商业模式下，用户难以根据每个阶段的具体需求，灵活地扩展或缩减算力资源，并通过相应的计费模式来使用这些资源。

3 **资源浪费严重**

- 在大模型的全生命周期内，算力需求呈现周期性波动，而传统的裸金属服务器租赁模式，通常是按月或按年整租，这种模式在需求低谷期可能导致算力资源的显著浪费。

4 **资源整合问题**

- 算力资源通常分布在不同的服务提供商和地理位置的数据中心。这种分布特性可能导致资源的统一管理和优化面临挑战。

来源：弗若斯特沙利文

智算中心操作系统赋能算力租赁商业模式变革（2/3）

- 算力包是一种**灵活的云计算服务**，它允许用户根据自己的需求选择和调整计算资源，从而减少成本并简化管理。这种服务提供快速部署、高安全性和专业支持，适合不同规模的企业，帮助它们实现数字化转型。
- **算力包的商业模式是由九章云极DataCanvas率先提出的，**它弥补了算力市场传统商业模式的空缺，为行业带来了创新的解决方案。

“算力包”定义

定义

智算操作系统可实现对庞大的GPU算力资源池的精确切分、管理，将其细化为小块算力单元。这些单元以“度”作为基本计量单位，**实现按算量收费**。系统采用**实时扣费**的模式，**确保计费的准确性和透明性。**

按“度”计算的必要性

1. 为下游客户提供直观衡量标准，使模型训练和推理任务成本效益预估合理，从而促进了整个行业**资源优化**。 — 估算成本和收益
2. 不同类型智算芯片性能不同，统一算力计量单位后可以**忽略不同品牌差异**，使用统一性能基准计费。 — 统一性能标准
3. 灵活取用机制允许用户根据需求快速扩展或缩减资源，从而**提高资源利用率同时有效控制成本消耗**。 — 资源管理优化
4. 使不同企业公司智算中心间**无缝算力调度**，实现算力资源的即时获取和使用，从而**提高算力便捷性**。 — 构建算力一体化

算力包的一“度”服务内容

算力包的运用场景

来源：弗若斯特沙利文

智算中心操作系统赋能算力租赁商业模式变革（3/3）

- **传统算力**租赁通常以长期合同和固定费用为特征，适用于计算**需求稳定且长期场景**。
- **算力包**相比于传统算力租赁模式，提供了**更高灵活性和动态调整能力**，用户可以按需购买和管理计算资源，实现弹性扩展和缩减，从而提高资源利用率并控制成本。

算力包和裸金属优劣势对比

算力包	VS	裸金属
按**算量**计费，仅在计算任务运行时计费，算力资源较少浪费	计费模式	按**算时**计费，机器空转时仍需分摊租用成本，平均算力利用率≤20%，算力资源浪费明显
灵活调度GPU资源，按需弹性扩展/收缩计算资源	灵活性	固定计算资源，扩展门槛偏高，需求较低期间算力空置严重
提供完整的大模型+Agent开发工具链	配套工具	不提供相应工具，需自行搭建环境和工具，运维成本偏高

算力包服务优势

1
- 统一全国**智算中心建设和运营**。
- 可以结合多种算力资源组合，提供足够的算力储备。

2
- **按算力消耗计费**，针对算力任务计费。
- **灵活调度GPU资源**，节约任务提交过后仍剩余在智算中心的空余资源。

3
- 内置完整的大模型工具链，提供**高性能模型训练及推理能力**。
- 通过算法加速、编译优化、内存优化和通信加速，实现训练效率的提升、GPU利用率的提高以及推理速度的加快。

4
- **快速落地**基于大模型的垂类应用。
- 根据行业类型给予相应转变。

5
- 支持企业伙伴**大模型开发、智能体应用**。
- 支持整合行业生态环境。

来源：弗若斯特沙利文

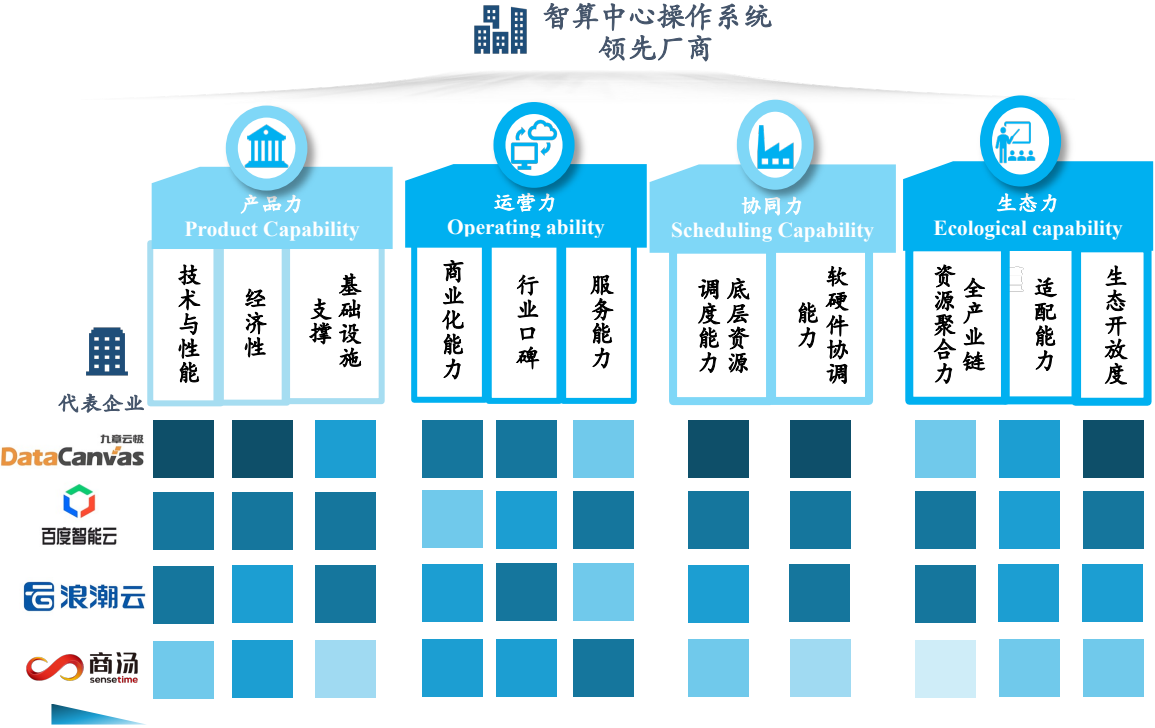
章节四：智算中心操作系统领先实践

智算中心操作系统厂商实力评估

关键发现

- **总体而言，智算中心操作系统市场目前处于高度景气状态。**它对于智算中心的建设和运营，以及大模型训练、推理等环节，尤其在生物制药、智能制造、自动驾驶等重点战略应用领域，都具有重大的意义，市场格局已初步显现。**然而，业界尚未有从智算中心操作系统服务提供商的价值视角出发，对参与厂商进行全面评价的先例。**
- **沙利文从技术、市场、资源、生态等视角入手，建立了综合服务价值分析模型。**该模型包含“产品力、运营力、协同力、生态力”四大维度，细分为11个一级指标和19个二级指标，旨在系统性地分析国内智算中心操作系统厂商的综合价值表现，赋能算力市场更自主化、成熟化发展。

智算中心操作系统厂商实力评估



备注：颜色越深能力越强

来源：弗若斯特沙利文

九章云极DataCanvas

关键发现

- 九章云极DataCanvas于2013年成立，**以“创造智能，探索未知”为使命**，以“助力全球企业智能升级”为愿景，**是中国领先的人工智能基础设施提供商**。沙利文认为，领先的技术开发优势、基于场景的技术服务能力是其核心优势之一。
- 公司拥有杰出的数据科学家和产品研发团队，深耕专业技术，**在智算中心操作系统、自动机器学习、数据建模、大数据分析等领域拥有前沿的行业实践经验**，凭借其在人工智能基础设施的技术优势、产品实力和优秀案例，**曾获沙利文评选“AI基础软件的领导者”**。

九章云极智算中心操作系统（DataCanvas Alaya NeW）

为AI而生：专注于人工智能核心任务，包括大模型的训练、微调、部署和推理等服务。**提供集“算力、数据、算法、调度”于一体的融合服务**，旨在推动人工智能产业化的创新发展。

算力包首创：提出了“算力包”产品，为业内首创，为算力服务提供了统一的**计量单位“度”**（DCU），为用户提供“买到即用到”的算力服务，为未来算力互联互通打下良好商业基础。

全局加速优化：**采用算法加速、数据三级缓存、内核优化、并行加速和模型压缩等方案，充分释放硬件效能。**集群训练效率提升100%，单卡利用率提升50%；模型推理速度提高4倍，Token吞吐量增加5倍。

异构算力纳管与调度：提供包含国内外**多种算力资源的适配、纳管、调度和优化服务**。通过实时监控和管理各智算中心的设备、创新的算法，实现海量算力智能调度。

智算中心操作系统产品优势

运用实例

- **中科曙光智能基础服务合作**

九章云极DataCanvas和中科曙光共同深化人工智能产业基础服务，推动AI新技术的深入应用和业务模式的创新突破。

- **马鞍山智算中心建设**

九章云极DataCanvas为马鞍山辐射地区提供对标国际一流水平的行业大模型能力和高质量的“算力+算法”一体化服务。

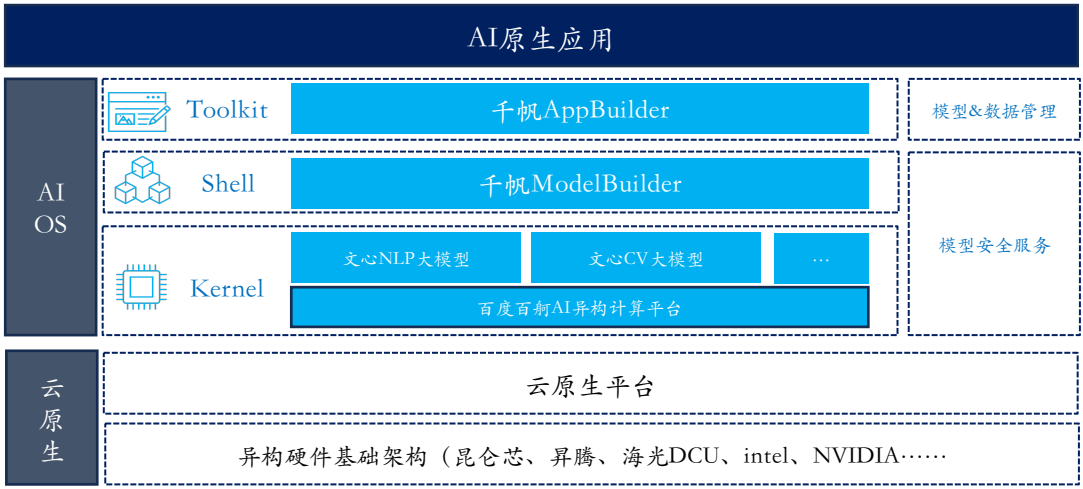
来源：弗若斯特沙利文、公司官网

百度智能云

关键发现

- 百度智能云**以“云智一体”为核心战略，**赋能各行各业。作为国内领先的综合性互联网服务提供商，其核心优势之一在于其先进的AI开发能力和丰富的数据积累。
- **百度智能云的智算中心操作系统“万源”是面向AI原生时代的智能计算操作系统，它通过提供极简的开发体验、智能的系统内核和高效的异构算力，为用户提供强大的支持。**万源操作系统的核心优势在于其先进的AI开发能力和数据积累，它能够管理大规模的集群，发挥GPU、CPU的性能，并拥有强大的大模型作为核心引擎，构建起全面的大模型服务能力。

百度智能云智算中心操作系统（万源）框架介绍



百度智能云（万源）核心功能及竞争优势

1. 万卡集群有效训练时长98.8%
2. 百卡混合训练效能97%
3. 文心大模型能力
4. 官方组件数55+

全面的工具链

中文理解能力

集成加速框架

节约成本

百度智能云（千帆）典型运用范式

电商　短视频　金融　教育　医疗　游戏

运用实例：沈阳皇姑智算中心

- **百度采用“万源”架构中的“千帆”功能，**助力沈阳文旅在大模型技术与产品方面的积累。**这一合作推动了全国首个集场景、算力、平台、模型于一体的端到端全栈AI智算中心的建设。**

来源：弗若斯特沙利文、公司官网

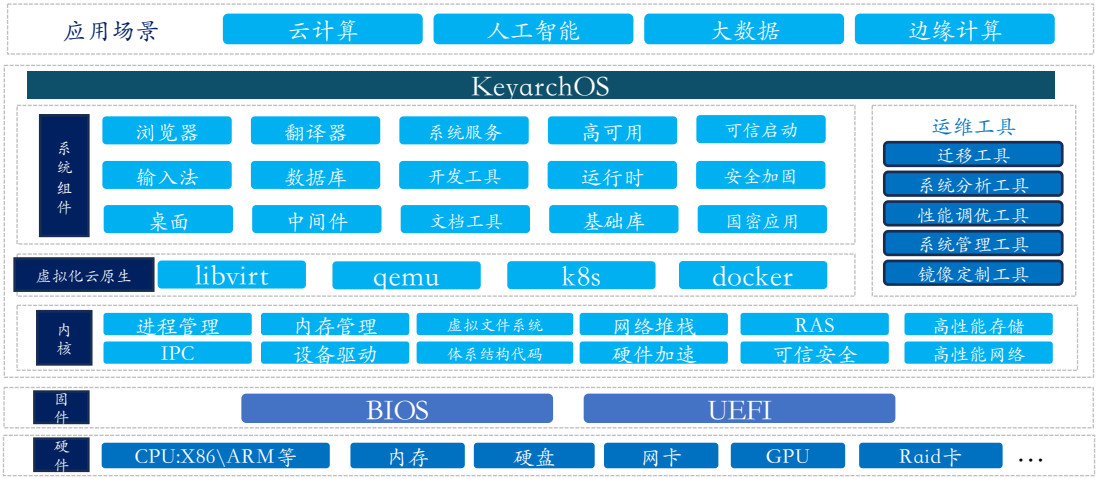
浪潮

浪潮云

关键发现

- 浪潮云成立于2015年3月9日，是中国领先的云计算和大数据服务提供商。在智能算力领域，浪潮云的核心优势之一体现在其软硬件业务的全面布局。
- **浪潮云智算中心操作系统（KOS）提供了全流程、全栈的智算软件栈，涵盖了AI算力系统环境的部署、算力调度及开发管理能力。**KOS集成了虚拟化云平台、容器云平台和云管理功能，能够统一调度和管理云数据中心的计算、存储、网络和安全等资源。这实现了业务的动态调整、智能资源管理和服务的自动化交付，从而优化了智算中心的运营效率。

浪潮云智算中心操作系统（KOS）框架介绍



智算操作系统（KOS）核心功能及竞争优势

稳定可靠	• KOS提供RAS增强、应用高可用等能力，在安全可信方面实现全栈可信链内置EDR安全加固套件，有效保障业务的连续性和可靠性。
高效协同	• KOS通过软硬件协同调度优化，基于浪潮信息KOS的人工智能、虚拟化、云原生等应用开发、运行效率、扩展性等方面都得到了增强。
全天候运维	• KOS提供深度监控、专家诊断规则、云端运维服务接入等能力，支持超过700余种深度监控数据，支持基于专家规则自动化运维诊断。
广泛兼容	• 依托上下游合作伙伴的广泛合作，KOS完成大量兼容认证覆盖服务器整机100余款、板卡350余款、数据库100余款、中间件150余款。

运用实例（KOS）

龙蜥CentOS迁移

- 浪潮云KOS助力推动并整合了龙蜥的ANCHK内核，实现了与市场主流芯片、板卡、数据库、中间件等软硬件的广泛兼容性。
- 针对关键行业对CentOS迁移替代的需求，浪潮云为龙蜥定制了一套全面的迁移方案，使其业务系统能安全可靠地过渡到浪潮信息的KOS平台。

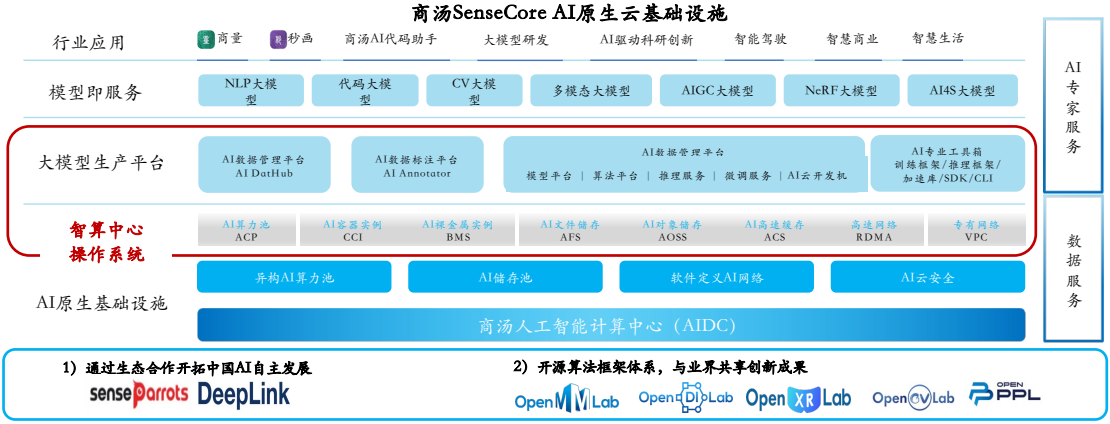
来源：弗若斯特沙利文、公司官网

商汤

关键发现

- 商汤科技自2014年成立以来，其智能计算中心操作系统相关产品作为SenseCore商汤AI大装置中计算基础设施层的关键组成部分和软件平台，与其AIDC共同构成了软硬协同的核心竞争力。
- SenseCore商汤AI大装置建立了一个内涵丰富的AI基础设施架构，涵盖了AI原生算力基础设施、大模型生产平台以及模型即服务（MaaS）。此外，还提供AI专家服务和数据服务，以支持相关技术措施的有效实施。

SenseCore商汤大装置全景图



1）通过生态合作开拓中国AI自主发展

sensetime Parrots DeepLink

2）开源算法框架体系，与业界共享创新成果

OpenMMLab OpenDILab OpenXRLab OpenCVLab OpenPPL

高效算力基础设施：**以大模型开发、生成和应用为核心的新一代人工智能基础设施。**其通过高稳定性、高性能和弹性扩展能力支持大规模并行训练，并确保长时间训练的高效和稳定运行。

AI数据管理平台：提供开放、高效的AI数据管理平台（AIDMP），涵盖从数据产生到使用的各个环节，并为大模型微调、自动驾驶等场景提供高质量、低成本的数据标注服务，显著提升**数据处理的效率和安全性。**

MaaS平台支持：通过MaaS（模型即服务）平台，整合自研大模型和开源大模型，提供包括模型微调、推理、内容安全等功能的全套生产工具，**帮助企业客户快速、低成本地接入并应用大模型。**

国产化生态适配：深耕国产化芯片适配，与多家国产芯片厂商合作，支持大模型的训练和生成式AI应用，构建了坚实的AI基础设施，并通过合作项目加速国产化落地，**推动AI生态的自主可控。**

智算中心操作系统产品优势

- 高性能AI计算，大规模并行处理
- 丰富的模型库，快速部署
- 提供全流程一站式AI服务
- 灵活的资源调度和管理
- 国产适配，兼容异构计算资源

运用实例

- **中公教育AI数字人**

中公教育与商汤科技合作，基于SenseCore大装置和“日日新”体系，开发了虚拟数字讲师“小鹿老师”，用于在线教育课程。

- **重庆智算中心**

商汤科技与重庆南岸区合作建设了人工智能计算中心（AIDC），其核心是国产化算力。

来源：弗若斯特沙利文、公司官网

附录

名词解释

ICC	**智算中心，全称智能计算中心（Intelligent Computing Center, ICC）：**是基于最新人工智能理论，采用领先的人工智能计算架构，提供人工智能应用所需算力服务、数据服务和算法服务的公共算力新型基础设施。
CPU	**中央处理器（Central Processing Unit，简称CPU）**：计算机系统的运算和控制核心，是信息处理、程序运行的最终执行单元，主要负责执行计算机程序中的指令，进行基本的算术、逻辑、控制和输入/输出操作。CPU自产生以来，在逻辑结构、运行效率以及功能外延上取得了巨大发展。
GPU	**图形处理器（Graphics Processing Unit，简称GPU）：**GPU是图形处理器，又称显示核心、视觉处理器、显示芯片，是一种专门在个人电脑、工作站、游戏机和一些移动设备（如平板电脑、智能手机等）上做图像和图形相关运算工作的微处理器。GPU使显卡减少了对CPU的依赖，能很好地解决电影级图像质量需要解决的透明性、高质量反走样、运动模糊、景深和微多边形染色等问题。
HPC	**高性能计算(High performance computing，缩写HPC)**：利用超级计算机实现并行计算的理论、方法、技术以及应用的一门技术科学，围绕利用不断发展的并行处理单元以及并行体系架构实现高性能并行计算这一核心问题，该领域研究范围包括并行计算模型、并行编程模型、并行执行模型、并行自适应框架、并行体系结构、并行网络通信以及并行算法设计等。
TPU	**TPU，张量处理单元芯片（Tensor Processing Unit），**是由谷歌开发设计的一种专门为了优化深度学习模型训练和推理过程的特殊芯片，被广泛应用于各种AI和ML任务中。与传统的CPU和GPU相比，TPU芯片专门优化了对大规模矩阵运算的性能，能够以适应更复杂的AI模型和更大规模的数据处理需求。
IDE	**集成开发环境（IDE，Integrated Development Environment）**:用于提供程序开发环境的应用程序，一般包括代码编辑器、编译器、调试器和图形用户界面等工具。集成了代码编写功能、分析功能、编译功能、调试功能等一体化的开发软件服务套。