2025年算力调度平台行业

算力规模四年翻倍

释放万亿级基础设施机遇

概览标签：算力、算力调度

Computing Power Scheduling Platform Industry

計算力スケジューリングプラットフォーム業界

报告提供的任何内容（包括但不限于数据、文本、图表、图像等）均系头豹研究院独有的高度机密性文件（在报告中另行标明出处者除外）。未经头豹研究院事先书面许可，任何人不得以任何方式擅自复制、再造、传播、出版、引用、改编、汇编本报告内容，若有违反上述约定的行为发生，头豹研究院保留采取法律措施、追究相关人员责任的权利。头豹研究院开展的所有商业活动均使用"头豹研究院"或"头豹"的商号、商标，头豹研究院无任何前述名称之外的其他分支机构，也未授权或聘用其他任何第三方代表头豹研究院开展商业活动。

研究目标

研究背景

随着人工智能技术的迅猛发展，全球范围内对算力的需求呈现指数级增长，因此需要算力调度来实现跨地域、跨平台的算力资源整合与优化调度。

研究目标

- 了解人形机器人硬件端及软件端的发展情况
- 了解人形机器人不同类型厂商的发展情况

本报告的关键问题

- 梳理算力调度及异构算力调度的相关技术与联系
- 梳理国内主要算力调度平台
- 梳理全球主要的开源算力调度平台

观点摘要

01 异构算力调度面临的挑战：

◆ 异构算力调度面临多重核心挑战：资源异构性与软件环境碎片化显著增加调度复杂性；跨架构任务迁移成本高导致效率低下；缺乏统一调度标准引发资源错配与利用率低等。

02 国内主要的算力调度平台：

◆ 国家级算力调度平台多由政府主导或运营商/头部企业建设，强调跨区域协同与市场化交易；省级平台覆盖长三角、成渝、京津冀等重点区域，结合本地产业需求；市级平台则聚焦本地AI、智能制造等场景。

03 主流的开源算力调度技术平台：

◆ 国内算力调度平台或基于开源算力调度技术平台打造，openFuyao作为新兴的多样性算力调度平台，在国产化适配支持上具有优势，而Kubernetes、Slurm等成熟项目则在云原生和HPC领域有深厚积累。

目录

名词解释

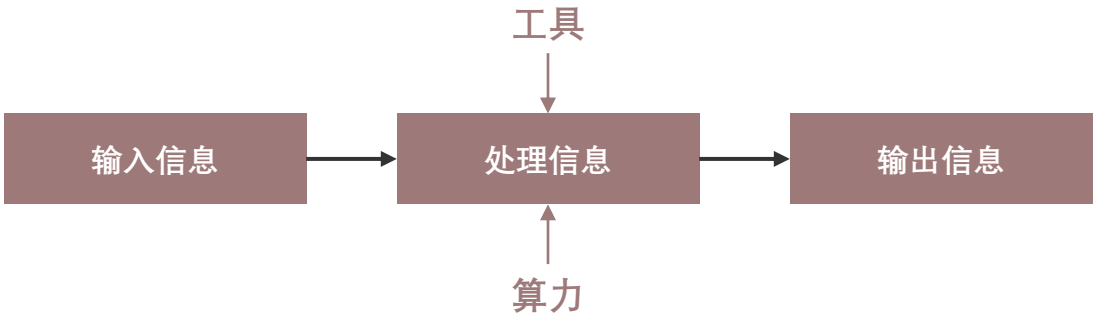
- **ASIC (Application-Specific Integrated Circuit)：**专用集成电路，是为特定目的设计制造的集成电路。相比通用处理器，ASIC在执行特定任务时效率更高、速度更快。
- **CUDA (Compute Unified Device Architecture)：**由NVIDIA开发的一种并行计算平台和应用程序接口模型，它允许开发者使用NVIDIA的GPU进行通用计算。
- **DPU (Data Processing Unit)：**数据处理单元，是一种新型的可编程处理器，主要用于数据中心内的网络、存储和安全等数据处理任务。
- **DSA (Domain Specific Architecture)：**领域特定架构，指专门为某一类应用或算法设计的计算机架构，以提高这些特定应用的性能和效率。
- **EFLOPS (ExaFLOPS)：**每秒百亿亿次浮点运算，是衡量计算机性能的一个单位，表示计算机每秒钟可以完成多少次浮点运算。
- **FPGA (Field-Programmable Gate Array)：**现场可编程门阵列，是一种硬件可重构的体系结构，用户可以根据需要通过软件定义来配置其逻辑功能。
- **GPU (Graphics Processing Unit)：**图形处理单元，最初设计用于加速图形渲染，现在也广泛应用于科学计算、机器学习等领域。
- **Hadoop：**一个开源框架，能够对大量数据进行分布式处理。它主要由HDFS（Hadoop Distributed File System）和MapReduce两部分组成。
- **Kubernetes：**一种自动化部署、扩展和管理容器化应用程序的开源系统，它使得用户能够更方便地管理和调度容器化的应用。
- **OpenCL (Open Computing Language)：**一种开放标准，用于异构计算，支持CPU、GPU及其他类型的处理器在同一平台上协同工作。
- **PyTorch：**由Facebook开发的深度学习框架，以其灵活性和易用性著称，特别适合研究和快速原型开发。
- **TensorFlow：**谷歌开源的第二代人工智能学习系统，被广泛用于机器学习和深度学习的应用开发。
- **TPU (Tensor Processing Unit)：**张量处理单元，是Google为加速其TensorFlow机器学习框架而设计的定制ASIC芯片。

Chapter 1
算力调度行业综述

算力调度行业综述——算力的定义与分类

- 算力（Computing Power）是指计算机设备或数据中心处理信息的能力。根据计算方式、算力核心、应用领域分类，算力可分为通用算力、智能算力和超算算力

算力的定义



算力（Computing Power）是指计算机设备或数据中心处理信息的能力，是硬件和软件协同执行计算任务的综合性能。狭义上，算力通常用每秒浮点运算次数（FLOPS）衡量，例如 TFLOPS（万亿次/秒）、PFLOPS（千万亿次/秒）。广义上，算力还包括数据存储、网络传输、任务调度等能力，是支撑数字经济发展和科技创新的核心生产力。

算力的分类

	通用算力	智能算力	超算算力
计算方式	基础通用计算	人工智能计算	尖端科学领域计算
算力核心	基于CPU的服务器	基于GPU、FPGA、ASIC的AI加速芯片	基于超级计算机的高性能计算机群
应用领域	云计算、边缘计算、企业办公系统、互联网服务	人工智能训练与推理	科学研究、工程仿真

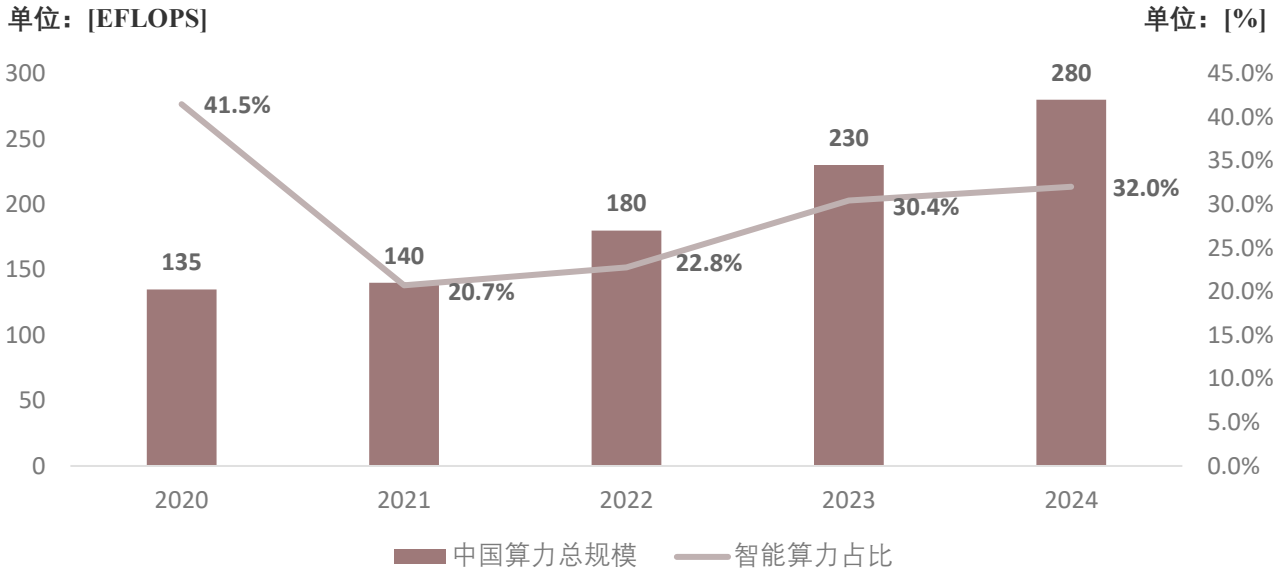
- 根据计算方式、算力核心、应用领域，算力可分为：**1）基础通用算力：**由基于CPU的服务器提供，适用于日常通用计算任务；**2）智能算力：**基于GPU、FPGA、ASIC等芯片的加速计算平台，专为人工智能任务设计；**3）超算算力：**由超级计算机或高性能计算集群提供，解决复杂工程问题。
- 算力作为数字时代的“新型电力”，是推动经济高质量发展和科技创新的核心驱动力，其重要性体现在：1）数字经济基石：算力是数据处理的核心能力，直接影响数字经济效率；2）科技创新驱动力：支撑AI、区块链、基因组学等前沿领域的突破；3）社会基础设施：算力如同电力一样，已成为现代社会运行的底层资源。

来源：头豹研究院

算力调度行业综述——算力规模与数据生产总量

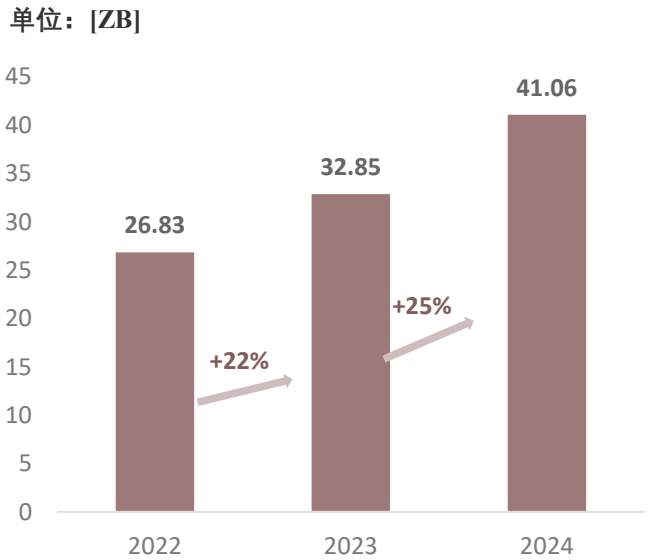
- 近年来，中国算力规模快速增长，由2020年的135EFLOPS增长至2024年的280EFLOPS，2024年智能算力占比达32%；2024年中国数据生产总量达41.06ZB，同比增长25%，增长主要受到智能应用的推动

中国算力总规模，2020-2024



- 近年来，中国算力规模快速增长，算力已发展成为集信息计算力、网络运载力、数据存储力于一体的大规模计算资源。根据工信部数据，截至2024年底，中国算力总规模达到280EFLOPS，其中智能算力占比32%（90EFLOPS）。稳居全球第一梯队。

中国数据生产总量，2022-2024



- 数据是算力的需求源，算力是数据的价值放大器。二者在政策引导、技术突破和市场需求的共同作用下，正加速构建数字时代的“新生产力”，为人工智能、工业互联网、智慧城市等场景提供核心支撑，推动中国经济向高质量发展转型。2024年，中国数据生产总量达41.06泽字节（ZB），同比增长25%，增速较去年提高2.56个百分点。数据生产总量的快速增长主要受到智能应用的推动，随着人工智能快速发展，用于开发、训练和推理的数据量同比增长40.95%。智能家居、智能网联汽车等智能设备数据增速位居前列，分别为51.43%、29.28%。

来源：工信部，中国信通院，《全国数据资源调查报告（2024年）》，头豹研究院

算力调度行业综述——算力网络与算网融合综述

- 算力网络是一种基于现有互联网基础设施的新型信息网络，其核心目标是通过标准化标识、智能调度和弹性网络能力，将不同地域、不同主题的算力资源互联互通，实现按需分配、实时感知、灵活调度

算力网络与算网融合对比

	算力网络	算网融合
定义	算力网络是以“算为中心、网为根基”的新型信息基础设施，通过将计算资源（算力）和网络资源深度融合，实现对云、边、端算力的**按需分配和灵活调度**。	算网融合是计算资源与网络资源在多个层面（硬件、软件、平台、应用等）的**深度整合**，目标是实现算力的“即插即用”和网络的“按需适配”，最终达成“网络无所不达、算力无所不在、智能无所不及”的愿景。
核心目标	实现算力的泛在化与灵活调度	实现算力与网络的深度协同与一体化服务
技术侧重点	网络主导的资源调度与跨域组网	计算与网络的双向适配与能力融合
典型特征	集中式/分布式调度、算力标识、算力路由	意图驱动、SDN/NFV、边缘与云协同
应用场景	大规模算力调度、跨域资源共享	智能业务优化、弹性资源扩展

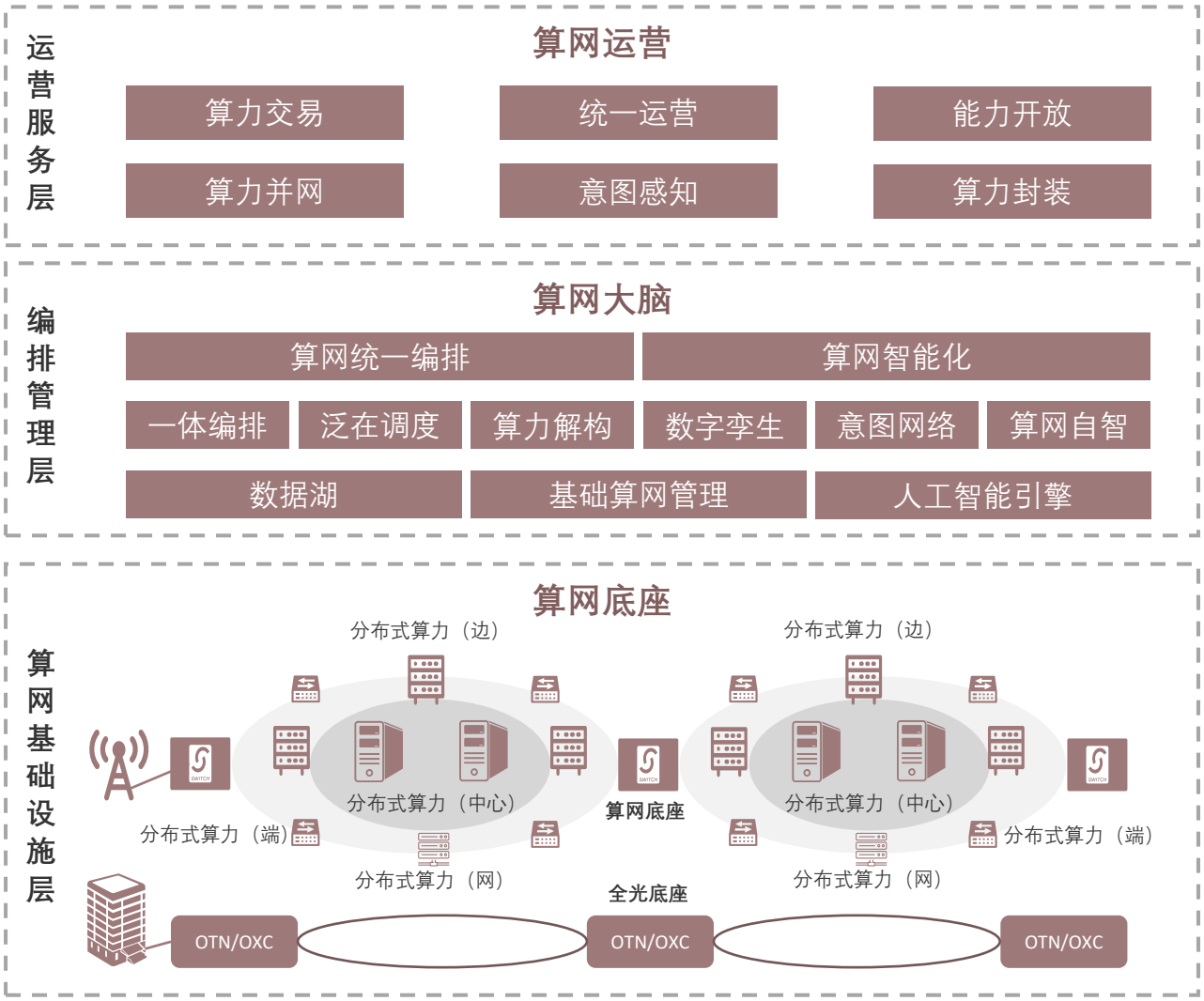
- **算力网络**（Computing Power Network）是一种基于现有互联网基础设施的新型信息网络，其核心目标是通过**标准化标识、智能调度和弹性网络能力**，将分散在不同地域、不同主体（如企业、数据中心、边缘节点等）的算力资源（如CPU、GPU、FPGA、ASIC等）**互联互通**，实现算力资源的**按需分配、实时感知、灵活调度**，最终让用户像使用水电一样便捷地获取算力服务。
- **算网融合**则是计算资源与网络资源在多个层面（硬件、软件、平台、应用等）的深度整合，目标是实现算力的“即插即用”和网络的“按需适配”，最终达成“网络无所不达、算力无所不在、智能无所不及”的愿景。它更强调技术层面的协同优化，而非单纯的基础设施建设。
- **算力网络是算网融合的基石，算力网络为算网融合提供了底层基础设施（如跨域组网、算力标识），而算网融合是算力网络的高阶目标。此外，算力网络侧重“资源调度”，算网融合则进一步追求“能力共生”，例如通过网络感知算力状态、通过计算优化网络路径。**

来源：中国移动，头豹研究院

（接上页——算力网络与算网融合综述）

- 算网融合则是计算资源与网络资源在多个层面（硬件、软件、平台、应用等）的深度整合，目标是实现算力的“即插即用”和网络的“按需适配”，最终达成“网络无所不达、算力无所不在、智能无所不及”的愿景

算网融合应用架构



- 中国移动的算网融合架构从逻辑上分为算网基础设施层、编排管理层和运营服务层。其中，1）算网基础设施层：作为算力网络的物理底座，依托全光网络与IP承载技术，构建云-边-端算力的高速互联通道，保障数据的高效、无损流动；2）编排管理层：作为算力调度的核心中枢，融合人工智能、大数据等技术手段，实现对底层算力资源的统一纳管、动态编排与智能优化，从而提升算力网络的整体效能；3）运营服务层：作为面向用户的服务能力平台，通过整合算网资源与场景需求，提供一体化的算力服务供给，使客户可便捷获取一站式服务并享受智能化、无感化的使用体验。

来源：中国移动，头豹研究院

算力调度行业综述——异构算力的定义与分类

- 异构算力是指通过结合两种或多种不同类型处理器或控制器架构（如CPU、GPU、FPGA、ASIC等）的计算单元，构建一个协同工作的系统，以发挥不同硬件架构优势，提升整体计算性能、能效比和灵活性

不同处理器的灵活性与性能对比

	CPU	GPU	FPGA	DSA	ASIC
架构特点	通用处理器，擅长复杂指令和串行任务	大规模并行计算单元，适合高吞吐量任务	可编程硬件，支持自定义逻辑电路	针对特定领域优化的专用硬件	针对单一任务完全定制的专用芯片
性能	单线程性能强，但并行能力有限	并行能力强，适合大规模数据处理	性能介于CPU和ASIC之间，延迟低	在特定领域性能极高	专为特定任务设计，性能最优
灵活性	高度灵活，可运行各种软件	灵活性中等，需依赖并行算法	灵活性高，可通过重新配置适应不同任务	灵活性低，仅适用于特定领域	完全无灵活性，只能执行固定功能
功耗	功耗较高，尤其在高性能计算时	功耗较高，但单位计算能耗优于CPU	功耗较低，但比ASIC高	功耗较低，针对特定任务优化	功耗最低，专为效率优化
开发成本	开发成本低，成熟工具链	开发成本中等，需掌握并行编程技术	开发成本较高，需硬件设计知识	开发成本较高，需深度领域知识	开发成本最高，需要大量前期投入
开发周期	短	中等	较长	较长	较长
上市时间	快	较快	较慢	较慢	最慢
升级与迭代	易于升级，只需更换软件	需要重新设计算法，但硬件兼容性较好	可通过重新编程升级，硬件无需更换	升级困难，需重新设计	无法升级，需重新制造

更强的灵活性 → 更强的性能

- 异构算力是指通过**结合两种或多种不同类型处理器或控制器架构**（如CPU、GPU、FPGA、ASIC等）的计算单元，构建一个协同工作的系统，以充分发挥不同硬件架构的优势，提升整体计算性能、能效比和灵活性。其核心目标是通过**异构资源的互补性**，解决单一架构在性能、灵活性或成本上的局限性。
- 依据指令的复杂度，处理器引擎分为CPU、GPU、FPGA、DSA和ASIC等，如图，从左向右，单位计算依次复杂，性能逐渐提升，但灵活性不断降低。
- **狭义的异构计算**聚焦于单一计算系统内部（如服务器或终端设备）通过集成多种类型处理器（如CPU、GPU、FPGA等）实现硬件协同和任务分工，以提升性能与能效；而**广义的异构计算**则突破物理边界，通过跨设备、跨网络的多层级技术整合（包括硬件、软件、编程框架、通信协议等），实现异构资源的高效调度与统一管理，强调对复杂场景下多样化计算需求的全面适配。狭义侧重硬件架构的协同优化，广义则追求跨领域、跨层级的系统级融合与生态化发展。

来源：头豹研究院

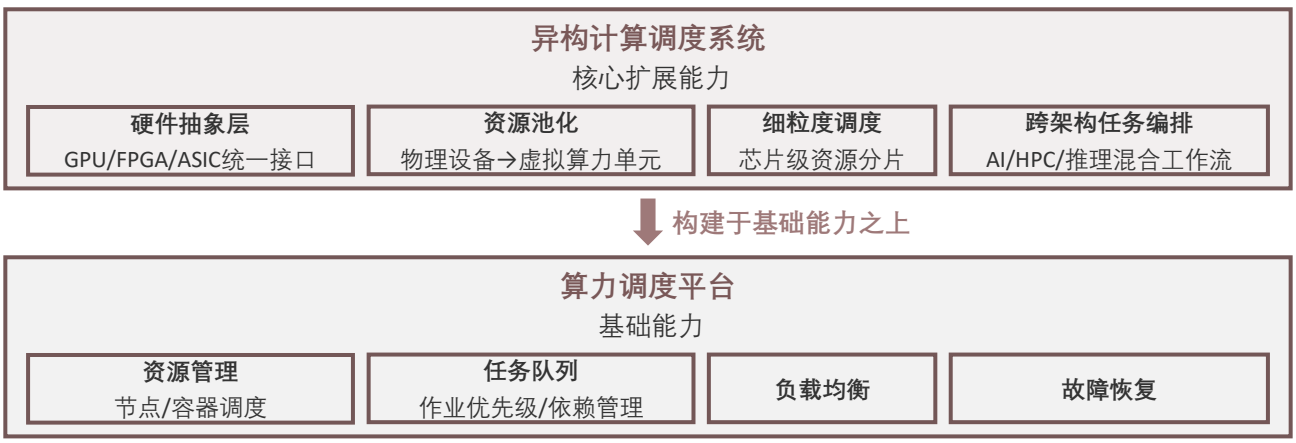
算力调度行业综述——算力调度平台与异构计算调度系统

- 算力调度平台是一种面向多类型计算资源（CPU、GPU、存储、网络等）的统一调度系统，而异构计算调度系统是专门针对异构硬件架构（CPU、GPU、FPGA、ASIC等）的任务调度系统

“算力调度平台”与“异构计算调度系统”的区别

	算力调度平台	异构计算调度系统
资源类型	多维度资源（计算、存储、网络、带宽等）	多架构硬件（CPU、GPU、FPGA、ASIC等）
调度颗粒	粗粒度（任务级、服务级）	细粒度（代码段级、指令级）
核心算法	资源感知、动态负载均衡、博弈优化算法	任务划分、异构资源匹配、通信优化算法
关键技术	算力路由、算网编排、资源池化	异构编程框架（CUDA/OpenCL）、RDMA、CCL

“算力调度平台”与“异构计算调度系统”的关系



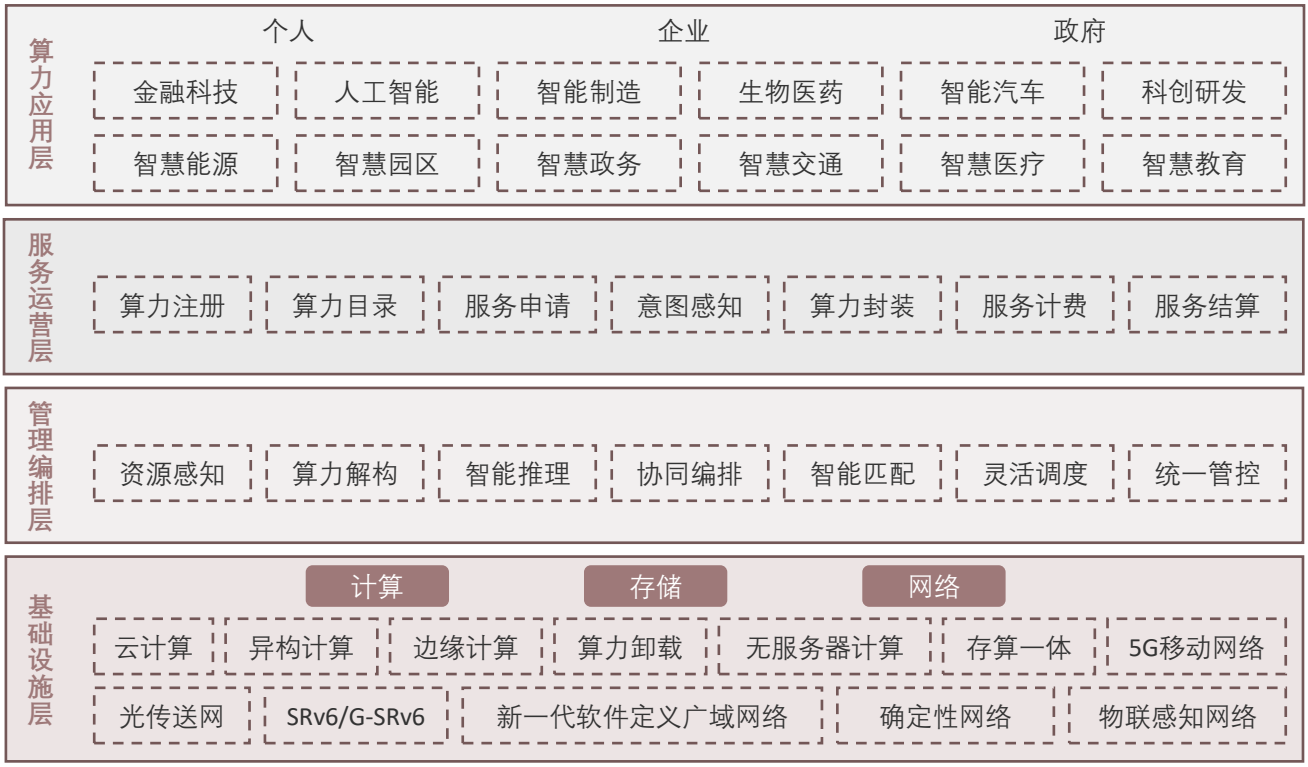
- **算力调度平台**是一种面向多类型计算资源（CPU、GPU、存储、网络等）的统一调度系统，旨在实现资源的**高效整合、动态分配和跨域协同**，满足业务需求（如AI训练、云计算、智慧城市等）。其目标是通过智能算法（如博弈论、强化学习）实现资源的**最优匹配**，解决算力供需矛盾，提升资源利用率和业务响应效率。
- **异构计算调度系统**是专门针对**异构硬件架构**（CPU、GPU、FPGA、ASIC等）的任务调度系统，核心是**将不同计算单元的性能优势与任务特性匹配**，以最小化执行时间。其目标是通过任务划分、负载均衡和通信优化，最大化**异构资源的并行性和利用率**，解决因硬件差异导致的性能瓶颈。
- 尽管两者有差异，但在实际应用中常结合使用。**算力调度平台负责宏观资源分配**（如将某个AI训练任务分配到某个数据中心），而**异构计算调度系统负责微观任务优化**（如在分配到的GPU集群中进一步拆分任务到不同卡）。总而言之，算力调度平台是“资源管家”，解决“哪里有资源、如何调用”的问题；异构计算调度系统是“性能调优师”，解决“如何让资源发挥最大效能”的问题。

来源：头豹研究院

算力调度行业综述——算力调度平台架构与技术

- 从技术架构角度，算力调度分为基础设施层、管理编排层、服务运营层和算力应用层；算力调度的关键技术包括算力感知、算力度量、算力路由、算网编排、算力交易等

算力调度平台体系架构



- 从技术架构角度，算力调度平台分为基础设施层、管理编排层、服务运营层和算力应用层。其中，**基础设施层**负责整合计算、存储、网络资源，提供池化算力与优质连接；**管理编排层**负责统一管控、智能调度，实现资源最优匹配与动态调整；**服务运营层**支撑算力交易与服务运营，保障安全、可信、高效流通；**算力应用层**则面向多行业场景，提供多样化算力服务，推动深度融合。

算力调度平台关键技术

算力感知

实时监测和采集计算节点的资源状态（利用率、负载、网络带宽等），为调度决策提供数据支撑。

算力路由

综合计算资源和网络状态，将任务流量路由至最优节点，支持跨域、跨架构的灵活调度。

算网编排

对计算资源和网络资源进行统一编排，实现任务路径的自动规划与资源协同。

算力度量

对异构算力资源（计算、存储、网络）进行量化评估，建立统一的多维资源模型，支持跨架构资源对比和标准化计费。

算力交易

根据用户的差异化需求，实现智能、公平、泛在、可溯、可信的算力交易。

来源：中国信通院，头豹研究院

算力调度行业综述——异构算力调度面临的挑战

- 异构算力调度面临多重核心挑战：资源异构性与软件环境碎片化显著增加调度复杂性；跨架构任务迁移成本高导致效率低下；缺乏统一调度标准引发资源错配与利用率低等

异构算力调度面临的主要挑战

➢ 资源异构性与软件环境差异导致调度复杂性显著增加

算力资源涵盖CPU、GPU、FPGA、ASIC等多种架构，每种硬件的性能特点（如并行计算能力、能效比）和指令集差异显著。例如，GPU擅长并行计算（如深度学习训练），而CPU更适合逻辑控制密集型任务。同时，不同任务依赖特定的软件框架（如TensorFlow、PyTorch）和库（如CUDA、OpenCL），调度系统需确保软件环境兼容性。若任务代码未适配目标硬件的编程接口（如CUDA代码无法直接在HIP生态中运行），则需进行代码迁移或重构，显著增加调度成本。

➢ 跨架构任务迁移成本高，适配技术复杂多样

异构算力调度中，任务从一种硬件迁移到另一种硬件（如CPU到GPU）往往需要重写代码或调整算法。例如，深度学习模型在CPU上训练时需依赖通用计算库，而迁移到GPU则需调用CUDA或ROCm接口，甚至需要重构数据并行策略。这种迁移不仅依赖开发者对多架构编程模型的掌握，还受限于不同硬件的内存管理机制（如GPU显存与CPU内存的差异）。此外，FPGA和ASIC等定制化硬件的适配需依赖专用开发工具链（如Vivado、Xilinx SDK），进一步提高了技术门槛。

➢ 缺乏统一调度标准与接口，生态割裂导致资源错配

当前算力调度缺乏统一的计量标准和接口规范，导致跨架构资源难以互通。例如，不同厂商的芯片（如NVIDIA GPU与国产GPU）在驱动、算子库和编程接口上存在差异，调度系统需针对每种硬件定制适配层。此外，跨厂商的作业调度生态支持能力弱，导致调度平台难以兼容所有硬件，资源利用率受限。

➢ 异构硬件性能特性非线性且动态变化，传统调度模型无法精准预测任务执行损耗

异构硬件（如GPU、NPU、FPGA）的性能表现往往受多重因素影响，呈现非线性动态特性。例如，GPU在高负载时可能因散热限制主动降频（如NVIDIA A100在持续高负载下性能衰减20%），而FPGA的能效比会随任务并行度动态变化。传统调度模型通常假设硬件性能为固定值，无法捕捉这些动态波动。

来源：头豹研究院

算力调度行业综述——国内主要算力调度平台

- 国家级算力调度平台多由政府主导或运营商/头部企业建设，强调跨区域协同与市场化交易；省级平台覆盖长三角、成渝、京津冀等重点区域，结合本地产业需求；市级平台则聚焦本地AI、智能制造等场景

国内主要算力调度平台分类——按行政层级分类

维度	子类	代表性平台	说明
行政层级	国家级	全国一体化算力算网调度平台	中国信通院与中国电信联合发布的首个实现多元异构算力调度的全国性平台，汇聚通用、智能、高性能、边缘算力资源，支持跨厂商、跨架构的异构算力动态感知与作业智能分发调度，推动“东数西算”战略落地
		国家新一代人工智能公共算力开放创新平台	科技部于2022年启动申报工作，2023年批复9家正式建设平台和16家筹建平台。该平台体系通过整合高性能计算与智能计算资源，采用“一中心多平台”架构模式，为科研机构、高校及企业提供普惠算力服务
		东数西算一体化算力交易平台	国内首个一体化算力交易调度平台，覆盖智算、超算、通算资源，通过市场化撮合实现东西部算力高效互补，促进数字经济产业链向西部延伸
		北京算力互联互通和运行服务平台	服务京津冀蒙及西部、北部地区的综合性算力枢纽，推动算力服务标准化。
		国家超级计算中心体系	由国家主导建设的超算中心网络（如天津“天河”、广州“天河二号”、济南“神威”、无锡“曙光”等），提供每秒百亿亿次级算力，支撑科研、气象、工业仿真等国家战略需求
	省级/跨省级	长三角一体化示范区异构智算云网调度平台	长三角区域算力协同枢纽，由上海电信主导建设
		粤港澳大湾区算力调度平台	整合大湾区多元算力资源，推动区域供需高效匹配
		四川省算力调度服务平台	全省算网融合关键基础设施，由四川能投集团牵头建设
		山东省黄河工业算力调度服务平台	全国首个面向黄河战略的工业算力枢纽，由国家超级计算济南中心运营
		甘肃省算力资源统一调度服务平台	该平台由丝绸之路信息港公司研发，采用“1+N+X”架构体系，是国内首个实现省级算力统筹调度的基础设施平台
		河南省算力调度服务平台	平台依托“一核四极多点”总体，统筹管理全省通算、智算、超算等异构算力资源
	市级	深圳市智慧城市算力统筹调度平台	全市多元异构算力汇聚枢纽，由深智城集团建设
		杭州市算力资源调度服务平台	杭州首个多元异构算力调度平台，提供算力超市与算力券服务
		武汉市算力公共服务平台	推动武汉算力交易供给侧改革，由武汉数据集团与运营商联合运营
		青岛市算力调度服务平台	实现青岛市算力资源池的统一纳管和编排调度，由青岛国实集团牵头建设并运营
		天津市算力交易中心	北方首个 “通智超” 一体的算力交易枢纽，由天津联通主导建设

来源：头豹研究院

（接上页——国内主要算力调度平台）

- 由运营商主导的平台可依托其全国范围的网络覆盖和基础设施，整合跨区域、跨服务商、跨架构的异构算力资源；由科技企业、行业服务商主导的平台则聚焦垂直领域或特定技术，资源整合范围相对集中

国内主要算力调度平台分类——按运营主体分类

维度	子类		代表性平台	说明
运营主体	运营商主导	中国电信	全国一体化算力算网调度平台	国内首个实现跨地域、跨厂商、跨技术路线的国家级平台，整合天翼云、华为云、阿里云等主流资源池
			“息壤”算力分发网络平台	天翼云4.0算力分发网络平台“息壤”是中国电信自主研发的基于云原生和跨域大规模调度技术的平台
			长三角一体化示范区异构智算云网调度平台	由上海电信主导，通过作业智能分发引擎实现跨厂商算力调度，覆盖长三角区域，服务气象预测、工业互联网等场景
		中国移动	移动云智能算力调度平台	基于“4+N+31+X”梯次化算力布局，整合通用算力、智能算力、超算算力、边缘算力，实现跨主体资源互联互通
			长三角枢纽芜湖集群算力公共服务平台	全国首个“四算合一”（通算/智算/超算/量算）省级平台，接入移动云、华为云等27个数据中心
		中国联通	联通云星罗先进算力调度平台	提供通智超一体化算力供给，支持CPU/GPU/NPU等异构芯片统一纳管，千卡级集群管理能力
			天津市算力交易中心	北方首个“通智超”一体的省级算力交易枢纽，由天津联通主导建设
	企业/厂商主导		中科曙光一体化算力交易调度平台	由中科曙光联合政府及科研机构打造，是国内首个基于“算力可用、可控、可计量”的交易调度平台
			阿里云震旦异构计算平台	阿里云推出的异构算力调度解决方案，支持CPU、GPU、NPU等多架构芯片统一纳管，通过容器服务ACK实现资源弹性分配
			百度百舸AI异构计算平台	百度自主研发的AI异构计算基础设施，深度优化飞桨框架与昆仑芯片协同，支持大规模模型训练与推理，在自然语言处理、自动驾驶等领域实现高效算力调度
			华为公共多样性算力服务平台	华为发布的业界首个公共多样性算力平台，通过集群计算解决方案整合昇腾、鲲鹏等芯片资源，支持AI、HPC、大数据多场景混合调度
			浪潮AI计算系统及推理平台	浪潮推出的一体化AI算力解决方案，支持GPU/DPU/ASIC等异构芯片统一管理，通过多级缓存与并行优化技术实现千卡集群线性加速比96%，支撑20个千亿大模型同时训练
			博云·海纳算力调度运营平台	博云科技研发的多元异构算力调度平台，整合通算、智算、超算资源，提供跨中心 / 跨区域智能调度与运营工具
			京东云 vGPU AI 算力平台	京东云推出的GPU资源池化解决方案，通过内核级细粒度切分技术实现1%算力与MB级显存调度，支持混合异构算力调度，服务金融、医疗等场景
			软通智算“天元智算服务平台”	软通动力旗下软通智算自主研发的去中心化算力共享平台，支持跨厂商CPU/GPU/国产芯片统一纳管与性能调优，已接入粤港澳大湾区公共算力服务平台

来源：头豹研究院

算力调度行业综述——开源算力调度技术平台

- 国内算力调度平台或基于开源算力调度技术平台打造，openFuyao作为新兴的多样性算力调度平台，在国产化适配支持上具有优势，而Kubernetes、Slurm等成熟项目则在云原生和HPC领域有深厚积累

主流开源算力调度技术平台对比

平台/社区	性质	核心能力	目标场景
openFuyao	由华为、中国移动、联通等主导的开源社区，专注于**多样化算力集群调度**	• **异构硬件资源池化：**支持CPU、GPU、FPGA等多架构算力统一调度 • **大规模集群调度：**优化超大规模集群的资源利用率 • **在离线混部调度：**实现在线服务与离线任务的混合部署 • **NUMA亲和调度：**提升分布式计算性能	云原生+AI时代下的通算与智算集群，**强调多样性算力调度和本土化场景适配**（如2025年Q3正式开源，计划孵化全球算力互联网标准）
Kubernetes	云原生容器编排平台，核心功能为**容器化工作负载管理**	• **自动调度与扩缩容：**基于资源需求动态分配节点 • **服务发现与负载均衡：**支持微服务架构 • **多集群管理：**通过Karmada等工具实现跨集群调度	云原生环境下的容器化应用，如Web服务、微服务
Slurm	**HPC**（高性能计算）领域的**作业调度系统**	• **分区管理：**按计算资源类型划分作业队列 • **资源隔离：**保障作业间的资源隔离性 • **公平调度：**支持多用户/多任务的资源公平分配	超算中心、科研机构的批处理任务调度（如科学计算、仿真模拟）
Volcano	**Kubernetes**的**批处理调度插件**，专为AI/大数据优化	• **Gang调度：**确保AI训练任务的所有Pod同时调度 • **多集群调度：**通过Volcano Global支持跨集群任务分发 • **优先级与公平调度：**满足高优先级AI任务的需求	AI训练、推理及大规模数据处理（如TensorFlow、PyTorch作业）
YARN	**Hadoop**生态的**资源调度框架**	• **资源抽象：**将计算资源（CPU、内存）抽象为统一单位 • **多框架支持：**兼容MapReduce、Spark等计算框架 • **弹性资源分配：**动态调整资源分配策略	大数据批处理（如日志分析、ETL任务）

来源：头豹研究院

头豹
LeadLeo

头豹业务合作

全球视野 · 本土洞察 · 助力企业把握市场先机

核心业务

行业数据API

开放原创报告与研究数据接口，支持企业知识库、系统平台及AI应用高效接入和调用

KNIT解决方案

构建企业可信内容体系，提升品牌在AI搜索与问答中的可见度、准确性与转化效果

报告会员账号

可阅读全部原创报告和百万数据，提供PC及移动端，方便触达平台内容

定制报告/白皮书

对产业及细分行业进行现状梳理和趋势洞察，输出全局观深度研究报告

商业尽调

面向投资并购和商业决策，评估标的公司的商业前景、价值及风险

招股书引用

研究覆盖国民经济19+核心产业，内容可授权引用至上市文件、年报

业务咨询

客服电话：
400-072-5588

官方网站：
www.leadleo.com

商务咨询与深度合作

报告作者

袁栩聪
首席分析师

张俊雅
行业分析师

service@leadleo.com

办公地点

深圳办公室
广东省深圳市南山区粤海街道华润置地大厦E座4105室
邮编： 518057

上海办公室
上海市静安区南京西路1717号会德丰国际广场2701室
邮编： 200040

南京办公室
江苏省南京市栖霞区经济开发区兴智科技园B栋401
邮编： 210046

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。

- 研究院依托中国活跃的经济环境，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员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，以专业的视野解读行业的沿革。

- 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，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。

- 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。

- 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

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。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。本报告所指的公司或投资标的的价值、价格及投资收入可升可跌。

- 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。本文所载的资料、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，头豹可发出与本文所载资料、意见及推测不一致的报告和文章。头豹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。

FROST & SULLIVAN 沙利文 | CAA 中国广告协会 CHINA ADVERTISING ASSOCIATION

2026

中国GEO营销实践与合规发展报告

企业案例征集正式启动

共建合规、透明、可信、可持续的GEO营销生态

聚焦生成式引擎优化服务实践、品牌AI可见度提升、可信内容体系建设与行业合规发展，遴选年度代表性企业及优秀产品案例，推动GEO行业规范化与高质量发展

重点征集方向

品牌AI可见度诊断

围绕AI搜索、AI问答及生成式引擎中的品牌识别、理解、引用与推荐表现开展监测与分析

内容资产优化与可信信源建设

围绕品牌内容结构化、权威信源建设、多渠道传播与可信内容供给形成实践案例

GEO产品与服务能力

展示GEO诊断模型、监测工具、内容优化工具、报告交付体系及全流程服务能力

合规实践与行业示范

重点呈现符合法律法规、平台规则、广告合规与内容伦理要求的代表性实践案例

重要时间

项目启动
2026年6月1日-2026年7月31日

产品及案例申报评选
2026年8月1日-2026年8月31日

报告发布
预计2026年10月中国国际广告节期间正式发布

扫码参与征集

发布单位： 弗若斯特沙利文

指导单位： 中国广告协会

传播支持： 头豹研究院

官网： www.frostchina.com