「人工智能+」时代
公共云发展模式与路径研究

国家信息中心
State Information Center
信息化和产业发展部
2024年3月

编委会（此排名不分前后）

主任：

单志广

副主任：

涂菲菲 于凤霞 王丹丹

编委：

张延强　房毓菲　吴洁倩　宋心荣

李春香　陈　栩　蔡丹旦　张雅琪

胡沐华　关乐宁　戴　彧　张　岳

主编单位：

国家信息中心信息化和产业发展部

目录

CONTENTS

前言

新一代人工智能技术正步入跨越式发展新阶段，成为引领现代产业变革的核心力量，大力发展通用人工智能已经成为全球潮流。2024 年我国政府工作报告中提出“开展‘人工智能+’行动，打造具有国际竞争力的数字产业集群”，既顺应全球人工智能发展趋势，也反映了未来中国产业升级的现实需求，开启了人工智能技术在各行各业广泛应用的新篇章。

随着人工智能技术和应用创新不断加速，模型复杂度和数据量急剧增加，对包括算力在内的新型基础设施建设提出了新需求和新要求。实践层面看，目前我国算力资源在规模和使用成本等方面仍然不能满足人工智能的规模化应用和快速迭代创新的需要，建立适应“人工智能+”时代的高质量算力服务体系迫在眉睫。

公共云是破解我国“人工智能+”时代算力“供不上、用不起”瓶颈问题的重要抓手。公共云具有大规模集群管理能力，能以资源利用效率最大化的方式，帮助提升我国算力供给能力，并通过规模经济效应推动算力门槛降低，让更多的用户享受普适普惠的算力服务。公共云和人工智能相结合还将推动“云计算”走向“云智算”，云计算不再局限于 IT 计算本身，而是提供包括算力、模型、数据、生态等与智能化发展相关的全方位创新服务，从而促进“人工智能+”切实落地，引领新质生产力不断发展。

本报告阐述了公共云的概念内涵及其对未来经济社会发展的重要意义，在分析我国公共云当前发展态势与面临问题的基础上，提出了“人工智能+”时代公共云发展的典型模式、模式要求、技术服务架构和健康长效发展路径，以期为相关从业者提供指导和参考。

01

公共云概述

（一）公共云的概念与内涵

公共云是指面向公众提供的云计算资源，其本质是将云计算资源规模化、大范围进行共享，聚合各类算力并通过在线的模式为各类用户提供简单易用、且近乎无限扩展的计算服务。

在传统的用户单位自建数据中心模式下，需要大量人力物力投入来建设基础设施、系统、中间件服务等，支出庞大且维护复杂。区别于私有云、公有云的“有”，公共云强调的则是“用”，重视使用权，是以需求为导向以应用为目的的一种公共服务模式。公共云模式下，用户单位无需过多关注上述技术细节，通过按需使用公共云服务商所提供的计算、存储、网络等开放资源，能够大大减轻 IT 成本，并转而投入更多的精力聚焦业务持续创新。

公共云是云计算发展的初心和核心价值所在。随着互联网技术的发展和虚拟化技术的成熟，云计算应运而生。云计算的核心理念是将计算资源集中起来，并通过互联网按需提供给用户；其核心价值和优势体现在资源共享、按需使用、灵活性和可扩展性。从实践层面看，全球公共云巨头亚马逊最早推出云计算服务 AWS（Amazon Web Services），企业和个人只需通过互联网就能按需获取计算和存储服务。AWS 自成立的第一天起，就肩负着以公共云对外提供云计算服务的使命。贝索斯曾这样描述 AWS 的愿景——“学生在宿舍里就能使用与世界上最大的公司一样的基础设施”。

面向未来看，人工智能特别是深度学习等技术的发展，对算力提出了更高要求，推动“云计算”走向“云智算”。人工智能的发展与公共云密不可分，公共云不仅支撑了人工智能大模型的突破，未来大模型也将以云的方式提供服务。“人工智能+”时代，从模型预训练到模型部署和推理应用，算力的需求呈指数级增长，公共云将成为破解算力资源紧张、算力成本昂贵的关键抓手。

（二）公共云有效支撑经济社会数智化发展

“人工智能+”上升为国家战略行动，人工智能从推动各行业提质增效的技术手段，升级为支撑经济社会数智化转型升级不可或缺的基础设施和核心能力。我国持续加强顶层设计，加快形成以人工智能为引擎的新质生产力。

随着人工智能应用和产业发展进入加速期，公共云成为推动全球经济增长和提升社会发展质量的关键驱动力。通过提供可扩展、灵活且经济实惠的计算资源，公共云成为连接不同行业、不同规模企业协同创新的桥梁，为各行各业提供了转型升级的新机遇，能够助推创新能力的整体提升，促进经济多元化、可持续发展。

1、推进技术、应用和模式创新

公共云降低了使用和试验新技术的门槛。公共云通过提供按需访问的计算资源、存储和广泛的服务，显著降低了开发者和企业创新研发的门槛，使个人开发者和中小型创业公司能够利用先进的技术栈进行产品实验和原型设计。如 SaaS（Software as a Service，软件即服务）允许个人和企业以订阅的方式访问最新的软件和工具，以较低的试错成本快速验证新设计的可行性，加速了技术创新和业务模式的演进。随着近年来人工智能等新技术的迅速发展，公共云的定位也在不断演变，不再仅仅是一种普惠、灵活的基础资源，还是个人和企业获取新技术新能力的重要渠道。

图 1 公共云助力构建创新生态

公共云提供了协同、开放发展的创新生态服务体系。公共云支持数据和应用程序的集中存储，通过提供丰富的 API、开发工具和集成服务等，为开发者构建和部署新的应用程序提供了创新平台，促进了企业、开发者和各类研究机构间协作，推动了商业模式和应用场景创新。例如，亚马逊 AWS、微软 Azure、谷歌等公共云厂商通过提供虚拟化的计算资源、存储空间和各种服务，支撑了医疗健康、能源、金融科技等各领域的数据挖掘和应用创新。

2、强力服务全行业数字化转型

公共云降低了企业数字化转型的成本。公共云服务允许企业根据计算能力、存储空间和网络带宽等实际需求快速调整多元的算力组合与服务模式，能够帮助企业、尤其是中小企业应对突发的业务量变化，减少了因资源不足而错失数字化转型机会的风险，对降低中小企业技术成本有至关重要的作用。

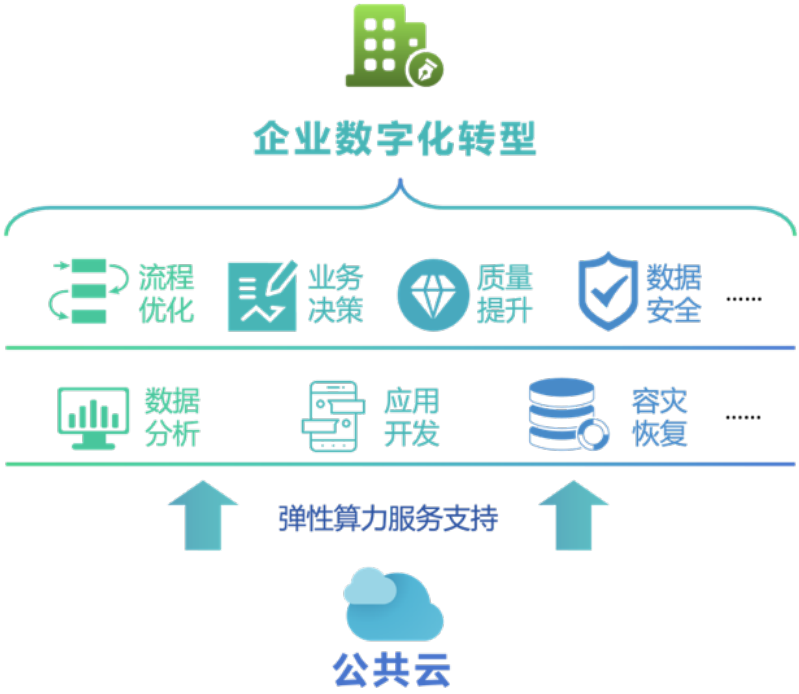


图 2 公共云助力数字化转型

公共云为企业提供广泛的数字化服务生态。公共云服务模式下，通过共享基础设施和资源，为广大使用者提供灵活的数据分析、应用快速开发和部署、容灾恢复等云服务，能够促进企业在业务流程优化、业务精准决策、产品质量提升、业务数据安全等方面实现数字化智能化转变，是初创数字企业贯彻数字化思维、推动中小企业跨越数字鸿沟、赋能企业节本增效、助力产业升级的战略性工具。

3、提升云服务用能效率促进绿色发展

公共云能够有效提升能源利用效率。与分散部署的传统数据中心相比，大型公共云平台的能源需求和消耗更加集中。公共云通过集中管理和运营，实现能源监控和负载平衡，确保能源消耗与实际需求紧密匹配，减少能源浪费。公共云平台通过计算资源集约化、转移工作负载，实现了更高的能源利用率（公共云的资源利用率是企业自建机房的 5-10 倍[1]），有助于减少能源浪费，从而降低碳排放。

图 3 公共云助力绿色发展

公共云有效推动清洁能源的利用。公共云对能耗的管控水平更高，为集中利用风能、太阳能等清洁能源提供了可行载体。例如，2023 年，亚马逊宣布当年已投资 78 个全新的太阳能和风能项目，预计到 2025 年，亚马逊 AWS、亚马逊物流中心、实体商店和公司办公室在内的所有业务运营 100%使用可再生能源[2]；谷歌、微软等公共云服务商也在积极探索通过使用可再生能源减少环境足迹。

[1] 来源：为绿色低碳注入科技动能——探访阿里云张北数据中心，http://www.xinhuanet.com/techpro/20210722/21ba34f4b01a410db5c1dc94e2378a62/c.html，2021 年 7 月 22 日。

[2] 来源：亚马逊宣布全球新增 78 个可再生能源项目，包括中国大庆和博白风能项目，https://www.amazonaws.cn/newsroom/2023/1115-sustainability/，2023 年 11 月 15 日。

02

公共云已成为“人工智能+”时代数智化发展的战略抉择

（一）公共云已成为支撑国际领先大模型的云服务首选

公共云为大模型训练提供万卡集群的技术能力。自 OpenAI 于 2022 年 11 月发布 ChatGPT 以来，全球大型科技企业掀起了基础大模型之战，不断升级大模型版本。当前，全球领先基础大模型的能力遵循“规模化定律”（Scaling Law），即模型参数、训练数据和算力规模越大，则模型效果越好。尽管公共云和专属云都可以用于训练基础大模型，但训练全球领先的基础大模型需要调度至少万张 GPU 高效协同工作。在此背景下，只有极少数公共云具备相应能力。例如，OpenAI 的 ChatGPT 部署在微软 Azure 云上，Anthropic 的 Claude 和 Meta 的 LLaMA 部署在亚马逊 AWS 云上，Midjourney 和 Google 的 Gemini 部署在谷歌云上。具体如表 1 所示。

表 1 全球领先基础大模型训练芯片规模及部署方式

厂商	基础大模型	GPU型号及数量	部署方式
OpenAI	GPT-4	2.2万张A100	微软公共云
	GPT-4.5	预计几万张卡	
	GPT-5	预计几万H100 甚至10万卡	
Anthropic	Claude	1.6万张H200	亚马逊公共云
Meta	LLaMA	2048张A100	亚马逊公共云
谷歌	Gemini	TPUv5（是GPT-4的4-5倍）	谷歌公共云
	Gemini Ultra		跨多个数据中心的大量 TPUv4
Midjourney			谷歌公共云

公共云为大模型应用提供“AI+云”的服务方式。智能时代云计算技术体系加速演进，从单点技术竞争升级为芯片、网络、计算、模型全体系技术综合能力的竞争。公共云巨头为新一代人工智能技术进步提供了强大支撑。以 OpenAI 和微软 Azure 的合作为例，微软提供装备了上万块英伟达 H100 GPU 和超过 20 万核 CPU 的超级计算系统，用于支持 ChatGPT 大模型训练和在线服务。在公共云上部署大模型，不仅能让用户更加及时地获取到大模型的最新功能和应用，还能通过其 PaaS 层和 SaaS 层为用户提供极为丰富的工具、组件和应用，从而支撑千行百业智能化应用。截至 2023 年 11 月，已有超过 18000 家组织通过 Azure 接入 OpenAI 大模型服务[3]。

[3] 来源：Microsoft 365 Copilot Set to Provide A Generative AI Revenue Boost，https://www.forbes.com/sites/robertdefrancesco/2023/12/19/microsoft-365-copilot-set-to-provide-a-generative-ai-revenue-boost/。

（二）公共云已成为破解全球算力瓶颈的核心手段

公共云能够缓解人工智能发展算力紧缺问题。大模型发展带来了 AI 算力需求的快速上升，所消耗的计算资源每 3-4 个月翻一倍[4]，算力需求的增长速度已经远超芯片性能提升和产能扩张速度的上限。随着人工智能大模型规模化应用，支撑海量用户频繁使用所需要的推理算力成本也将急剧上升，尤其是多模态大模型对于算力的消耗将远高于文本类大语言模型。算力资源紧缺已成为制约大模型规模化应用的主要瓶颈。据估计，截至 2023 年 8 月，全球 H100（英伟达主流高端 GPU）的供给缺口超过 43 万张[5]。受限于封装技术及产能不足，H100 订单交货周期长达 36 周到 52 周不等，即使是美国订单也无法得到保障。公共云可以多路复用，通过多租户使用同一套计算资源大池，削峰填谷，显著提升硬件资源利用率。随着公共云技术体系加速升级，AI 训练、AI 推理以及 HPC 超算等计算资源将并池管理，实现算力普惠和模型普及。

公共云能够实现我国算力供给能力的边界突破。国内大量企业自建数据中心的平均资源使用效率不到 5%，而亚马逊 AWS、谷歌等公共云厂商的数据中心资源使用效率一般可达 25%-40%[6]。在我国高端算力芯片进口受限的背景下，破解我国算力瓶颈的关键路径之一在于提高既有芯片和产能的利用率，而非盲目新建投入。公共云通过集群的计算、网络、存储平衡设计

[4] 来源：吴泳铭《拥抱人工智能驱动的产业智能革命》

[5] 来源：Nvidia H100 GPUs: Supply and Demand，https://gpus.llm-utils.org/nvidia-h100-gpus-supply-and-demand，2023 年 11 月。

[6] 来源：阿里云 2024 新战略：全面降价 20%，https://mp.weixin.qq.com/s/smviQkHlKUTKFvvY1OLI9A，2024 年 2 月 29 日。

和软硬一体化加速技术，调度“盘活”已有芯片，可以形成超大规模算力资源池，实现芯片复用、弹性可扩展。发挥公共云大规模机器调度、异构芯片兼容能力，不仅能将已有先进芯片集约化利用，还能充分利用已有的通用 CPU 资源，为 AI 大模型训练和推理应用提供必要的算力支持。

（三）公共云已成为智能产业降本增效的必由之路

公共云规模效应有利于算力普惠。算力成本是限制大模型发展的关键因素之一，中小型人工智能企业往往难以通过自建算力设施解决训练和推理算力需求，因此，算力租赁需求旺盛。而由于美国芯片出口禁令等外部因素，当前市场上算力租赁业务价格不菲。有观点认为，一旦算力成本降到目前的1%，大模型就能实现真正的普及[7]。公共云以其网络效应和规模效应，具备降低成本的潜力。随着用户规模增加、技术优化和运营效率提升，公共云厂商持续降低云计服务价格。例如，亚马逊 AWS 曾连续三年每年降价 12 次；过去十年阿里云将计算成本降低了 80%，存储成本降低了近 90%。

[7] 来源：创新工场汪华：两年后，模型基础设施成本可降至当前 1%，https://chuangke.aliyun.com/info/1067538.html，2023 年 7 月 25 日。

公共云加速人工智能产业化进程。人工智能原生企业提供新技术和新工具，是人工智能产业化的生力军和未来的主力军，也是人工智能产业成熟度的重要标志。美国涌现出了 Stability AI、Midjourney、Pika 等一批 AI 原生企业。其中，Midjourney 成立不到一年，用户超过 1300 万，市值超过 10 亿美元；Pika 成立不到半年即获得 5500 万美元融资，估值也超过 2 亿美元。公共云厂商为这些 AI 原生企业注入了关键的算力资源，比如，微软累计向 OpenAI 投资 130 亿美元，大部分是算力资源，CoreWeave 通过质押万卡 H100 获得英伟达 23 亿美元融资，如表 2 所示。此外，公共云平台通过“平台+低代码+生态”的方式，构建普惠化、个性化、低成本的技术架构和解决方案，帮助企业专注于业务创新，加速技术的迭代与升级。例如，Discord 依托谷歌云平台，从一个聊天沟通平台变成 AI 创新应用的“首发”验证平台，支持了 Midjourney 和 Pika 的成长。

表 2 AI 独角兽企业主要投资方及投资方式

企业	估值（亿美元）	主要投资方/投资额（亿美元）	投资方式
OpenAI	1000	微软/130	算力+现金
Anthropic	184	谷歌/20 亚马逊/40	算力+现金
Midjourney	100	基于Discord平台	
CoreWeave	70	英伟达	抵押万卡H100获23亿美元融资
Character.AI	50	谷歌/未公布	云服务+现金
Hugging Face	45	谷歌、亚马逊、英伟达、英特尔/2.35 Salesforce/2	
Inflection AI	40	微软、英伟达/13	2.2万H100
Stability AI	40	英特尔	Xeon+Gaudi2芯片+5000万现金
Cohere	30	英伟达、甲骨文 Salesforce/已融2.7，拟融10	算力+现金
Mistral AI	20	英伟达、Salesforce/5	算力+云服务+现金

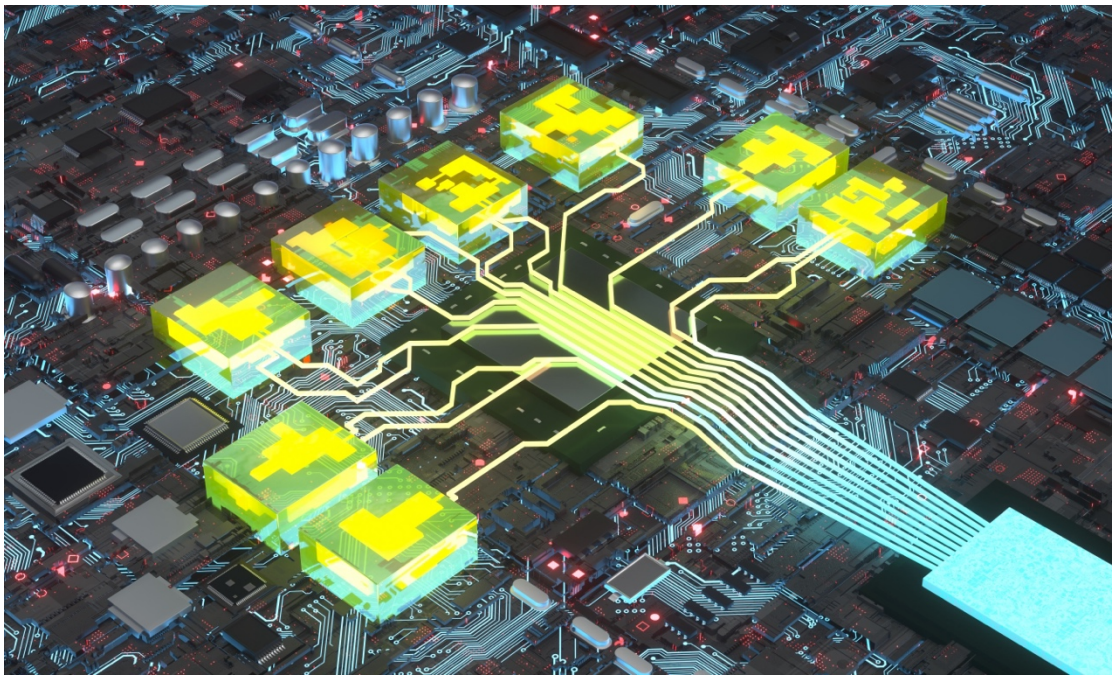
数据来源：课题组根据公开资料整理（截止到 2024 年 1 月 31 日）

03

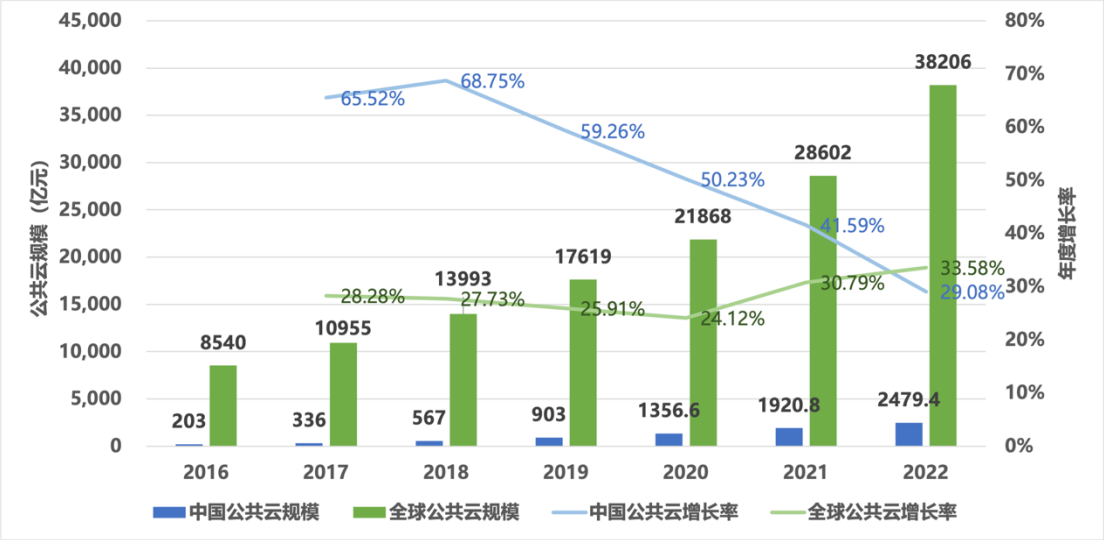
我国公共云
发展态势和面临的问题

（一）市场增长乏力，需求引导不足

无论是从理论上还是国际市场发展趋势看，算力服务最经济的方式是将算力集中在云厂商手中，以公共云模式提供算力服务。特别是智能计算时代，算力采购成本更高、集群管理更难、研发投入更大，公共云可以利用规模效应降低算力成本。近年来，国际市场算力正变得更集中，公共云保持快速增长态势。然而，我国公共云市场却呈现出增速放缓、市场内生算力需求不足、算力供给分散等特征。



2016-2022 年，全球公共云服务市场实现了蓬勃发展，市场规模从 2016 年的 8540 亿元增长到 2022 年的 38206 亿元。其中，我国公共云服务规模从 2016 年的 203 亿元增长到 2022 年的 2479.4 亿元[8]。在经历了快速增长期之后，近年来我国公共云市场增速逐步放缓。2023 年上半年，我国公共云服务整体市场规模为 190.1 亿美元，从 IaaS（Infrastructure as a Service，基础设施即服务）+PaaS（Platform as a Service，平台即服务）市场来看，2023 上半年同比增长 15.9%[9]，为近三年来同比增速新低。据 IDC 预测，到 2027 年全球收入将达到 1.34 万亿美元(约 93800 亿元)，复合增长率达 19.4%[10]。与全球趋势对比，我国公共云市场规模不仅份额小，发展动力也不足。



数据来源：IDC 公开资料整理

图 4 全球公共云服务规模

[8] 来源：国际数据公司（IDC），以美元为单位的数据按照 7：1 的汇率折算成人民币。
[9] 来源：国际数据公司（IDC）《中国公有云服务市场（2023 上半年）跟踪》
[10] 来源：国际数据公司（IDC）《全球半年度公共云服务（2023 上半年）追踪》

我国公共云市场增速持续走低的主要原因，一是需求不足，二是供给分散。

从需求端看，最积极使用公共云的互联网行业近几年发展持续低迷，占据 IT 支出大头的政企行业更加青睐私有云、混合云和专属云。据国际咨询机构 Gartner 数据显示，2021 年中国混合云占比达到了 42%，2024 年中国混合云占比将达到 70%，远高于全球平均水平（50%）[11]。可见我国公共云市场增速在持续下滑，而本地部署的云（混合云、私有云、专属云）保持稳定增长。**从供给端看，**近年来我国云市场出现了明显的分野，腾讯云、阿里云等更聚焦发展公共云，而有大量云厂商更多在混合云、私有云、专属云等本地部署的云上发力，导致算力建设日趋分散化。

与国外发达国家相比，我国公共云市场在总体规模、发展速度、全球市场份额等方面仍然存在较大差距。美国亚马逊、微软和谷歌三大科技巨头云计算业务牢牢占据了近 70%的市场份额，算力资源集中。这三大云厂商普遍基于公共云提供服务，其优势是规模大、效率高，天然能向全球市场扩张，不断摊薄算力、研发成本，推动美国成为全球云计算市场的领导者，其市场规模全球占比超过四成。私有云和专属云过多，会导致我国算力产业和软件服务业碎片化，对未来人工智能产业、数字经济核心产业以及新质生产力整体提升都将造成一定的阻碍。

[11] 来源：Gartner《中国混合云运营的三个重要经验》

（二）重硬轻软失衡，服务效能不佳

云计算通常以 IaaS、PaaS、SaaS 等方式向外提供服务。早期以 IaaS 服务为主，随着产业发展的深入，各行业领域对于上云用云服务的需求愈加多样化，应更加重视服务的创新和对应的生态支持。纵观全球云计算市场，SaaS 服务的占比不断提升，而我国呈现出 IaaS 服务占比提升、SaaS 发展迟缓的趋势，灵活便捷的软件应用支撑服务供应不足，使得用户用云服务成本较高。从长远来看，难以形成可持续发展的优质云服务生态。

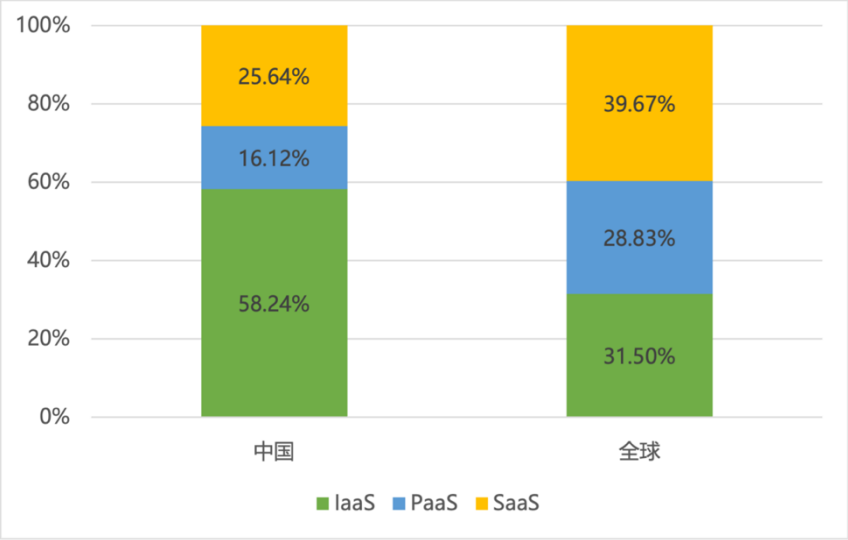


图 5 2022 年中国与全球公共云服务结构

从公共云服务结构来看，我国与全球发展情况存在差异，我国公共云服务形式主要以提供 IaaS 为主，PaaS、SaaS 的市场占比较低，存在平台和软件方面的服务短板。Gartner 数据显示，2022 年，我国公共云市场 IaaS、PaaS、SaaS 市场规模占比分别为 58.24%、16.12%、25.64%，以 IaaS 为主。同年，全球公共云市场 IaaS、PaaS、SaaS 市场规模占比分别为

31.50%、28.84%、39.67%，SaaS 是主体。此外，IDC 数据显示，2023 年我国 IaaS 市场占比进一步提升，SaaS 市场占比进一步降低，与全球公共云服务结构差异愈加凸显[12]。

[12] 来源：国际数据公司（IDC）《中国公有云服务市场（2023 上半年）跟踪》

（三）算力资源分散，利用效率不高

近年来，我国算力总规模高速增长，但算力资源未能实现充分利用。2018-2022 年，我国在用数据中心机架总规模年均增速超过 30%，但 2022 年上架率仅为 58%[13]，低于全球平均水平（60%）和欧美发达国家平均水平（65%）。其中一个重要原因是，专属云建设比例高，开放不充分的算力服务市场导致大量算力供给浪费。

从使用效率看，公共云 CPU 利用效率可以达到 40%，而专属云部署的 CPU 使用效率通常为 5%-10%[14]。2022 年，我国以公共云形式提供服务的算力占比仅为 28%，大部分服务器以私有化部署的形式存在[15]，部分城市通过自建数据中心来承载政务信息系统，金融机构和大型央国企出于数据安全和资产保值的目的大多也自建数据中心。算力资源分散，小规模、分散运营的传统机房普遍存在，难以形成统一的公共云服务市场，缺乏可持续发展的竞争力。

[13] 来源：中国信息通信研究院《中国算力发展指数白皮书（2023 年）》

[14] 来源：王轶辰《不可轻视数据中心高能耗问题》，https://mp.weixin.qq.com/s/5mP822sF7BW0zATgPqFiqQ?poc_token=HGMUBWajc4YjrG2TkgoCTJLrw0QgH6pRG9WA477F，2023 年 3 月 16 日。

[15] 来源：中国信通院，2022 年 7 月中国信息化百人会公开演讲《推动我国算力设施发展的思考》。

04

“人工智能+”时代
我国公共云发展模式

（一）典型模式

1、市场主导的建运投一体化模式

市场主导的建运投一体化模式，一般由单一市场主体同时作为投资方和运营方，采用自建或代建方式形成自有资产的算力资源，通过自营方式整合生态运营伙伴资源，共同面向用户市场化提供多层次、多种类的公共云服务资源，这也是目前腾讯、阿里、百度等云服务商和电信运营商提供的公共云服务的主要模式。根据主导企业的不同，又可以进一步细分为科技企业主导的平台服务模式、运营商主导的垂直整合模式和 IT 厂商主导的建设运营模式等。

这一模式的优势是运营主体拥有对服务资源的完全配置权，能够自主面向用户需求快速调整产品开发和服务供给策略，从而更好地满足用户需求。但这一模式也要求市场主体同时具备雄厚的资金实力和较高技术和管理能力。此外，不同细分模式具有不同的相对优势，如 IT 厂商主导的模式服务器采购成本较低，运营商主导的模式网络和终端客户资源更丰富，科技企业主导的模式平台化运营能力更强。

2、政企合作的多元生态运营模式

政企业合作的多元生态运营模式，投资方、建设方、运营方可以是不同主体，例如投资方可以是政府、市场主体或其他社会机构，或者三类主体的多元组合；建设方可以是具备技术实力的云服务商、电信运营商；运营方一般作为投资方的授权代表，通过有效的组织管理，基于自身技术能力或采购相应的技术服务、资源服务、运营服务甚至集成服务，例如地方政府主导的公共算力普惠供给模式，多采用由政府指定机构进行投资建设和运营的方式，由政府提供资金、土地、政策等支持，并对数据中心的规划、建设、运营进行监管。

这一模式的优势是能够更好发挥参与各方的资源和能力优势，更多地从社会利益的角度考虑资源配置。如地方政府在公共算力普惠供给中引入技术和生态合作伙伴，可以有效弥补财政资金难以持续投入、技术和运营能力不足等问题。但因为涉及主体较多，这一模式也往往需要更多的协同配合，需要形成能够平衡各方利益的建设运营方案，组建分工合理、责权明确的专业化团队，对运营方的协调管理能力具有较高要求。

（二）模式要求

1、市场化供给

市场作为一种高效、灵活的配置资源方式，一是能够通过竞争机制激发市场主体活力，促进企业提高效率、降低成本，提高产品质量和服务水平，形成更加丰富、更高质量的市场供给；二是能够通过价格机制调节供需，使资源流向市场需求较高的领域，实现更加有效的资源配置。为有效破解算力产业发展过程中出现的资源易闲置难题，增加高质量、高效率的算力资源供给，形成更有韧性、更具竞争力的产业生态，应采用市场化方式实现算力资源供给。

2、规模化经营

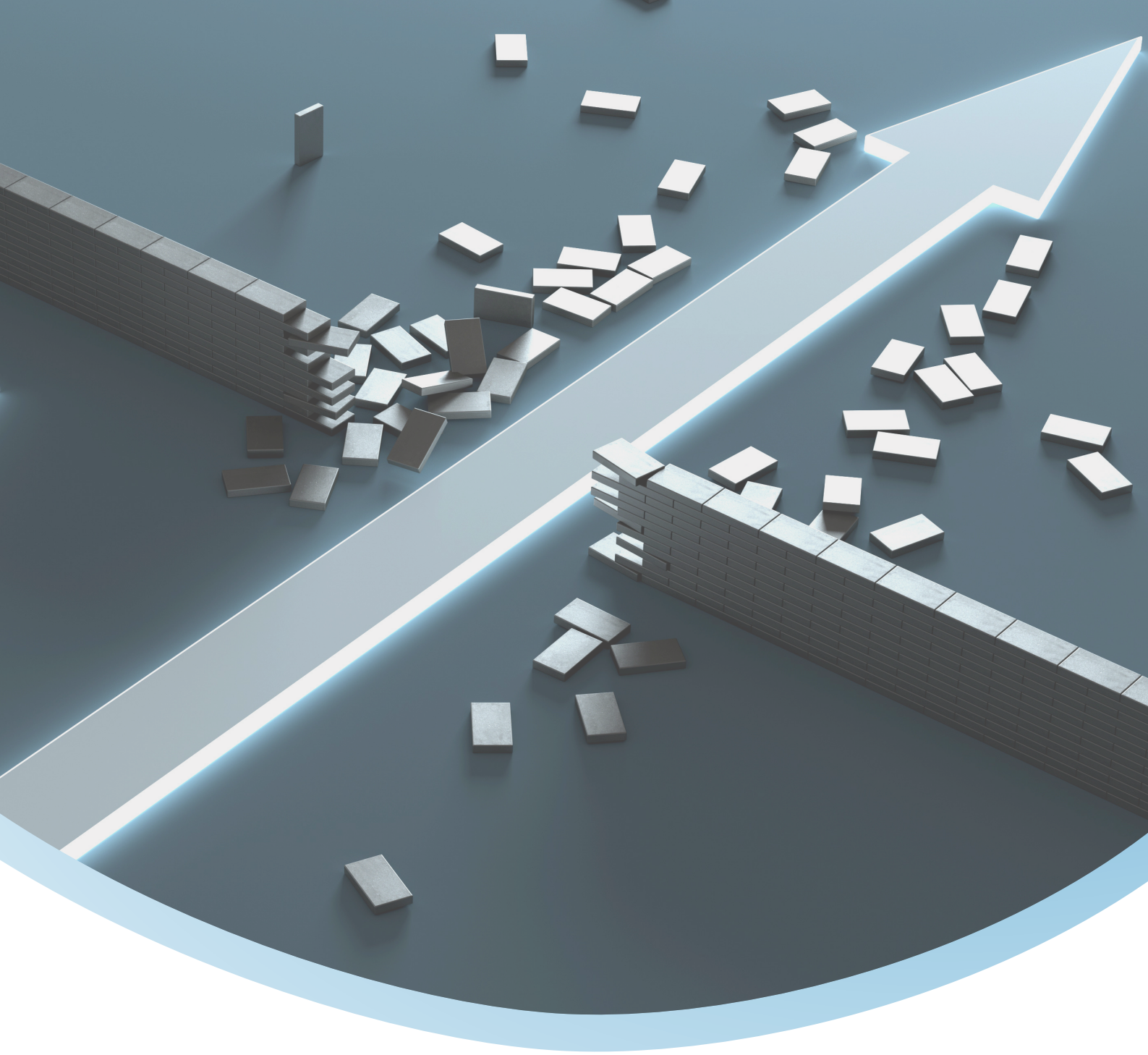
由于算力资源供给具有初始固定投入高、边际成本递减和边际收益递增的特点，通过规模化经营、多用户共享资源、按使用量计费，能够使公共云具备更加极致的资源弹性调度能力，在更好满足用户算力资源弹性使用需求的同时，提高资源的使用率、降低平均使用成本。为有效破解算力资源使用成本高的难题，提供更具性价比的算力资源服务，应采取规模化经营的路线，避免数据中心分散建设、小散化经营。

3、专业化运营

公共云可以有效屏蔽技术细节，为用户提供即时开通、按需使用、灵活可扩展的云服务，但相应的也需要服务团队具备极高的专业技术能力。例如为满足不同用户的不同需求，须掌握“一云多芯”“一云多算”等技术，支持对异构芯片进行统一纳管、池化和调度和对异构算力集群的统一管理，形成对多源异构算力资源的统一调度能力。为有效降低用户获取资源的技术门槛，提供可靠、稳定的技术服务，应采用专业化运营的方式。

4、生态化发展

为适应多场景、个性化的应用需求，一方面，需要更多懂业务、有数据的生态伙伴参与 SaaS/MaaS（Model as a Service，模型即服务）层应用产品和服务开发，另一方面，也需要公共云的技术框架更加开放兼容，能够适配多种技术和服务产品，从而在传统算力服务之上，提供更加丰富、便捷可选的应用解决方案。为使公共云服务更好贴近用户需求、满足业务应用需要，应注重打造开放协同的创新生态，这也是公共云产业链繁荣发展的有效途径。



05

“人工智能+”时代
我国公共云发展路径

（一）技术服务架构

“人工智能+”时代的公共云技术服务架构包括 IaaS、PaaS、MaaS、SaaS 四层，如图 6 所示。

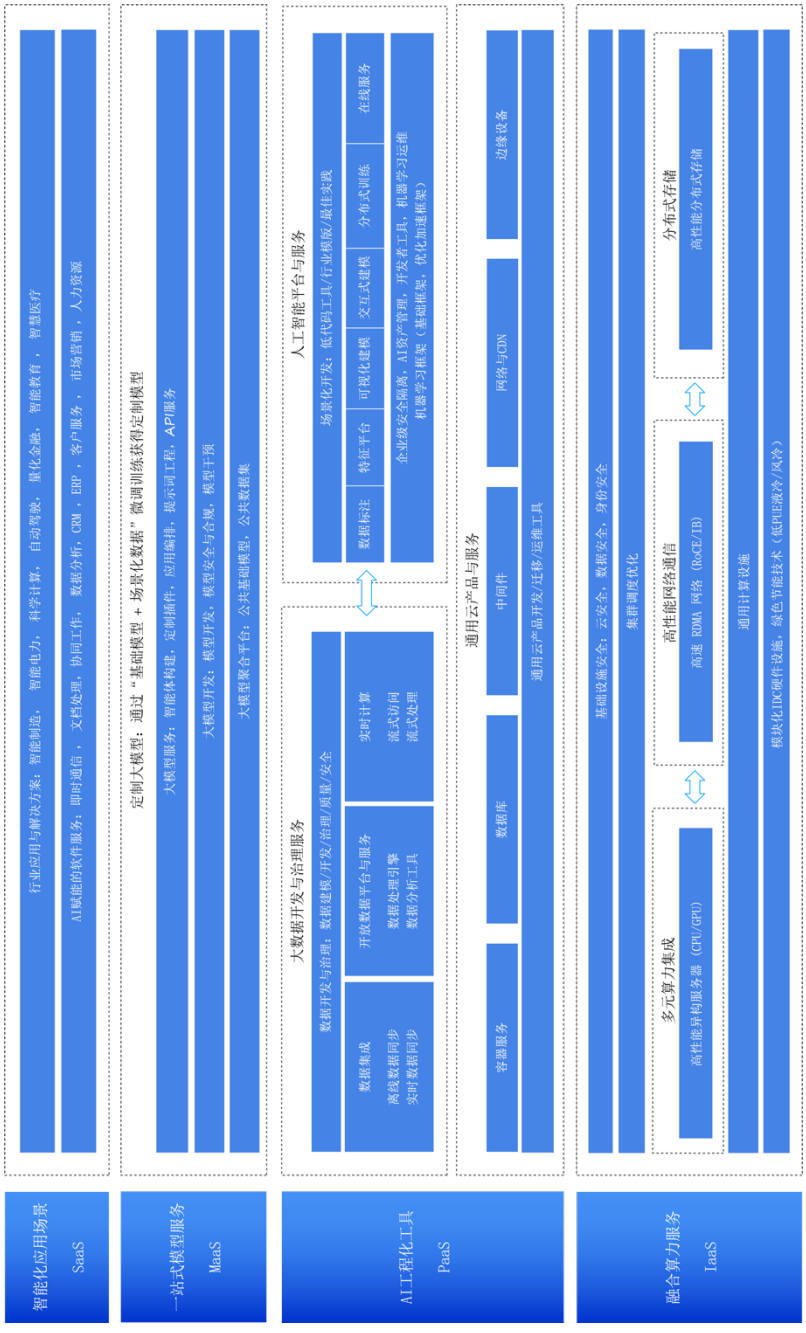


图 6 “人工智能+”时代的公共云技术服务架构

1、以融合算力设施为支点的 IaaS

人工智能的快速发展引爆算力需求，公共云通过虚拟化技术将计算、存储、网络等基础设施资源进行高效聚合，对外输出融合异构算力服务，能屏蔽底层复杂硬件、开源框架的技术细节，兼容 X86、GPU、ARM 等多种芯片类型服务器，是破解我国当前 AI 算力瓶颈的最优路径，也是实现 AI 大模型大规模商业化的必由之路。

公共云 IaaS 底层由模块化的 IDC 硬件设施和通用计算设施构成，通过云操作系统将多元算力资源、网络通信资源和分布式存储资源等进行调度优化，让用户能够根据业务需求动态调整计算资源的数量和配置，以满足不同的业务负载需求，并提供数据安全、身份安全等基础设施安全服务。在算力集成方面，通过聚合 CPU、DSP、GPU、DCU、PPU、ASIC、FPGA 等多种计算单元来提升计算性能，面向 AI 应用提供高效训练和精准推理能力，并支持算力资源的大规模部署和弹性扩容；在网络通信方面，支持高速 RDMA（Remote Direct Memory Access，远程直接内存访问）链路服务，实现零拷贝数据传输和智能网络监控管理，减少算力芯片负担的同时显著提高通信性能和效率；在分布式存储方面，针对不同应用场景提供块存储、对象存储等高性能分布式服务。

2、以 AI 工程化工具为重点的 PaaS

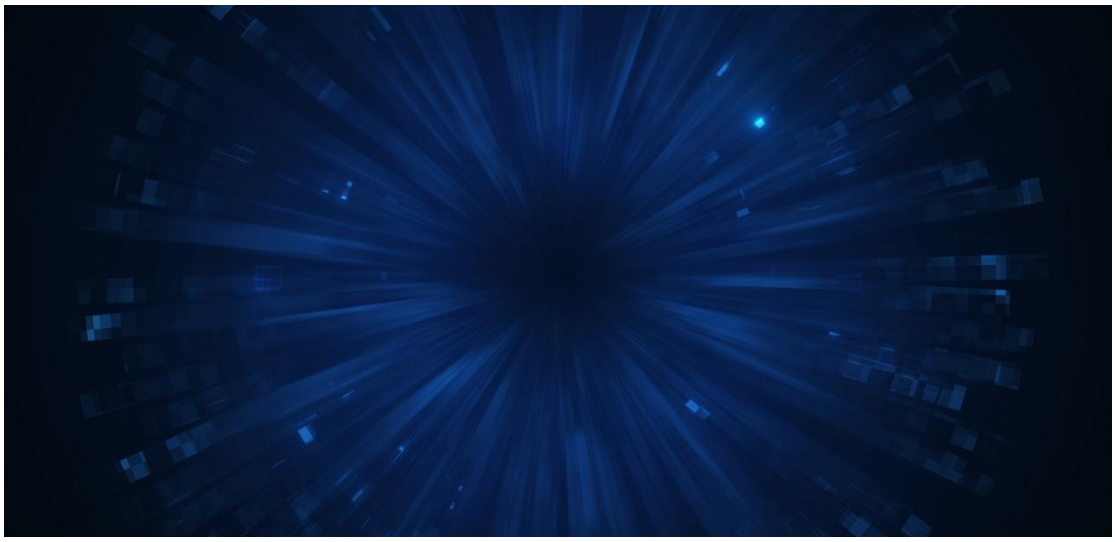
在 PaaS 层，公共云服务可以为用户提供统一的应用研发、测试、运维工具和服务，同时面向人工智能开发需求，提供高质量智算服务和机器学习框架，输出一站式 AI 开发能力。PaaS 层使用户无需在本地构建和维护平台就可以拥有面向 AI 的工程化能力，在提升灵活性和效率的同时极大降低了开发成本。

PaaS 层的主要包括通用云产品与服务、大数据开发与治理服务和人工智能平台与服务。面向通用的云服务需求，为用户提供容器服务、数据库、操作系统、网络与 CDN 和各种中间件，使用户无需关注底层硬件、系统软件和设备运维管理，专心开发应用程序，简化开发操作。面向人工智能算法模型训练对海量数据的需求，提供规模化数据采集、清洗、存储、分析、治理等功能，实现高效的大数据开发与治理。面向人工智能开发与训练需求，提供场景化低代码开发工具、数据标注、特征提取、可视化建模、分布式训练、机器学习框架、AI 运维管理等工具，利用海量弹性的云上算力资源，支撑人工智能模型训练任务的快速开展。

3、以一站式模型服务为核心的 MaaS

大模型是通用人工智能竞争的焦点，推动云计算进入“AI+云”的新一轮竞赛。公共云可以通过大模型和开放 API 打造实际落地的 AI 服务能力和产业生态，赋能经济社会的千行百业。因此，MaaS 层作为“人工智能+”时代特有的一层应运而生，通过为开发者提供集模型聚合、模型开发、模型服务为一体的一站式模型服务，解决传统 AI 应用中壁垒多、部署难等问题，降低 AI 技术应用到千行百业的门槛。

MaaS 层主要为泛 AI 开发者提供灵活、易用、低成本的一站式模型定制服务。通过大模型聚合平台，汇集行业领先的预训练模型，减少开发者的重复研发投入，面向具体应用场景，基于基础大模型和场景化数据，提供定制大模型微调和训练的整套工具，让用户能够更方便地打造符合个性化需求的专属模型，让大模型服务在专属行业领域能够应用落地。



4、以智能化应用场景为特色的 SaaS

在 SaaS 模式下，软件应用程序以云的方式通过网络提供给用户使用，使用户无需自己安装、维护和管理软件。SaaS 模式具有访问灵活性、多租户架构、高度可扩展、自动化升级和维护、按需付费、版本统一性等特点。“人工智能+”时代，SaaS 层可在 MaaS 层提供的定制模型的基础上，构建适用于智能化应用场景的软件服务，进而在软件服务的基础上结合各个行业的特点和需求形成完整的行业解决方案。

SaaS 层包括 AI 赋能的软件服务、行业应用与解决方案。基于 MaaS 层的场景化定制模型，SaaS 层实现了 AI 赋能的软件服务，将各种传统的场景智能化，例如，即时通信、文档处理、协同工作、数据分析、CRM、ERP、客户服务、市场营销、人力资源等领域。为了满足 AI 在不同行业落地的需求，通过集成 AI 赋能的场景化软件，实现行业化的智能解决方案，例如，智能制造、智能电力、智能教育、智慧医疗等。

（二）长效发展路径

1、需求导向，完善服务生态优化服务供给

为有效扩大公共云使用需求，需从完善产业发展生态和优化服务供给两端发力。一方面，充分发挥政府在优化营商环境的中的主导作用，营造更加公平的市场竞争环境，助力公共云产业生态培育和健康发展。一是在算力基础设施建设中给予网络直连、能耗、土地、税收等方面政策倾斜。二是鼓励财政资金购买公共云服务，将企业购买云服务纳入研发费用加计扣除税收优惠。三是将基于云计算构建的数字产品与服务纳入资产入表范围。另一方面，面向用户多元化、个性化的用云需求，提升多元算力的融合供给能力。一是加强对大规模异构算力资源的统一管理，提供更细粒度、更具弹性、更加灵活的算力资源供给。二是推广跨平台技术应用，允许和协助用户转移其使用公共云承载的系统和数据，并保证转移前后的功能一致与服务持续，提升云服务可迁移性，推动公共云更加开放。

2、应用为王，强化公共云服务效能升级

面向多元化的场景应用需求，丰富优化 SaaS、MaaS 服务方式，整合生态资源，提供更多结合行业领域业务场景的应用服务。一方面，充分发挥行业龙头企业在生态培育方面的引领作用，以开放兼容为原则，加快技术研发、产品开发、开源社区建设等。另一方面，在普及 IaaS 服务的基础上，推广 SaaS、MaaS 等服务方式，结合行业数字化转型需求，积极发展包括诊断咨询、方案设计、迁移服务、数据应用服务、知识技能培训、资源运维管理等上云应用服务，进一步建立用云安全理赔机制，打消用户使用公共云服务安全顾虑。

3、统筹衔接，推动集约化高效化发展

集约化发展是公共云实现规模化经营，提升产业竞争力的必然要求。要强化增量资源布局，推动存量资源整合，提高算力资源综合利用率。一是加强对算力资源布局和结构优化引导，鼓励适度超前、“质”“量”同步规划建设算力资源，支持以公共云服务方式提供算力服务，避免盲目上马、无序建设造成重复投资。二是鼓励市场主体通过集约化建设、规模化经营降低单位算力资源供给成本，提供更有竞争力的产品和服务。三是加强政企协同，通过多元化生态运营的模式，推动区域小散数据中心资源整合，提高存量资源的综合利用率。

国家信息中心信息化和产业发展部