阿里云 | Deloitte.

公共云优先，打造智能时代的

开放创新基础设施

核心观点

智能时代来临，第二次“大航海时代”的号角已经吹响

1.历次产业革命，被载入史册的都是技术跨过普惠拐点的时刻。

2.唯有实现算力普惠和模型普及，才能更早迎来AI技术的普惠拐点。

全球公共云千帆竞渡，体系化发展是关键

1.从芯片层面看，2023年，美国和中国在全球算力规模中的份额分别为43%和26%，领跑全球，但同2021年相比，差距进一步拉大。

2.从芯片和云计算层面看，美国公共云占算力规模的65%，中国为28%。公共云的计算资源利用率约40%，私有云不超过5%，导致美国和中国在全球算力规模中的实际份额分别为9.54%和5.12%，差距拉大。

3.从芯片、云计算、模型服务和开发者生态共同组成的创新系统看，2023年全球AI初创企业投资份额中，美国占比为50%，中国份额为30%，差距进一步扩大。

公共云优先是大潮所指，私有云是过渡形态

1.过去十年，美国政府实施“公共云优先”战略，通过联邦数据中心整合计划（FDCCI）、联邦政府信息技术采购改革法案（FITARA）、数据中心优化倡议（DCOI）等政策措施，数据中心数量减少7000个，减少约50%；部分服务器利用率从5%提升到65%以上。

2.以公共云为主要基础设施，美国SaaS企业市值超8万亿美元；2022年，超过50%的生成式AI初创公司位于美国，吸引全球七成私营投资。

3.中国以私有云为主要基础设施，SaaS企业市值约1000亿美元。

4.OpenStack私有云的发起者Rackspace从“ALL in OpenStack”转为“多云”战略。Mirantis裁员。惠普与思科相继关停OpenStack私有云。英特尔提前退出了与Rackspace合作的一个OpenStack创新项目。赛门铁克、Juniper退出黄金会员席位。OpenStack历经十年后走向衰落。

公共云是智能时代必备的创新基础设施

1.大语言模型在过去两年对算力的需求增长了750倍。单机八卡H800整机功耗10KW，已接近IDC机柜供电上限；同时在算力、显存容量和带宽等诸多方面无法满足训练需求，分布式AI训练系统已成刚需。

目录

Contents

1.号角：第二次“大航海时代”的历史召唤

当前，通用人工智能等前沿技术领域正在发生革命性突破，推动数字经济发展迈向新阶段。人工智能的规模化应用，是推进产业体系智能化的关键变量，将对全球产业体系变革带来历史性机遇，人类的第二次“大航海时代”已经开启。回顾历史，从技术突破演进成产业革命，能否快速实现技术的普惠应用，是促成技术扩散和产业体系升级的胜负手。我们必须抓住智能时代的历史机遇，以公共云构建普惠计算服务，坚实支撑起人工智能产业的爆发。

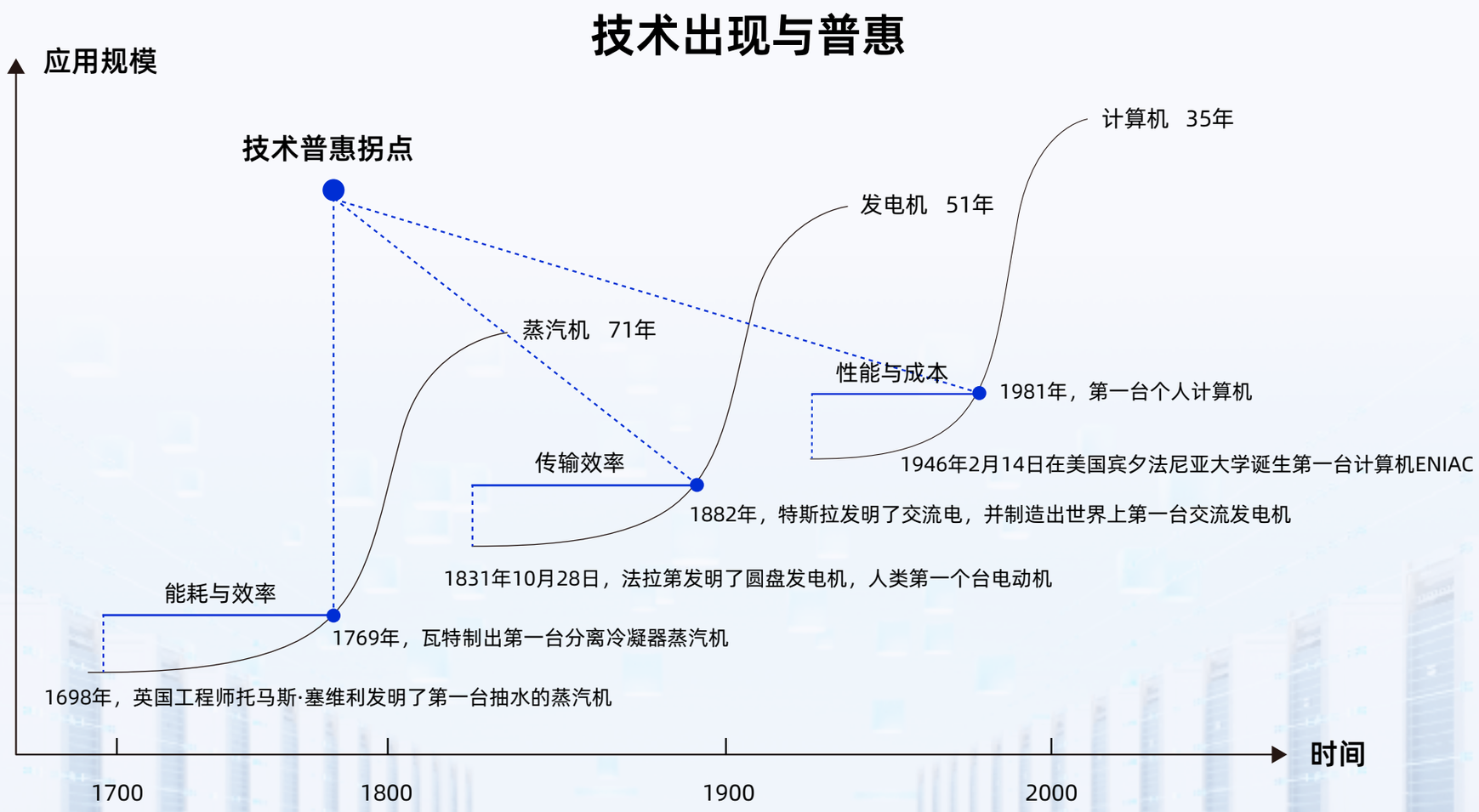
1.1 产业革命，技术普惠是胜负手

回顾历次产业革命过程，被载入史册的关键技术标志，都是技术跨过普惠拐点的历史时刻，而并非技术的原始发明时刻。技术发展只有跨越规模化应用的门槛，才有可能与行业普遍结合，进而有机会推动产业体系的重塑。

1 关键技术只有跨过普惠拐点，才能促成产业革命

瓦特是蒸汽机的改良者而非发明者。蒸汽机是第一次工业革命的标志，它让大工业的发展有了统一性，为手工生产走向模块化、标准化的工业生产提供了通道。瓦特并非蒸汽机的发明者，作为改良者的他为何被世人铭记？就是因为瓦特真正解决了蒸汽机走向应用的问题。他从技术、产品、商业模式、使用成本等环节上让蒸汽机具有了与行业结合的可行性。瓦特的改良，是蒸汽机实现普惠应用的拐点，也是第一次工业革命的真正发端 。

图1-1 历次产业革命都发端于关键技术跨过普惠拐点



电气化革命和计算机革命也起源于相应技术的普惠突破。爱迪生发明了直流电，但直至特斯拉发明交流电，解决了电力远距离低损耗传送问题，才让电力得以大规模应用，第二次产业革命由此开启。第一台通用计算机ENIAC发明于美国宾夕法尼亚州立大学，但直至IBM研发成功商用计算机并进而推出个人电脑，计算机才得以走进工厂企业和千家万户，第三次产业革命随之席卷而来。

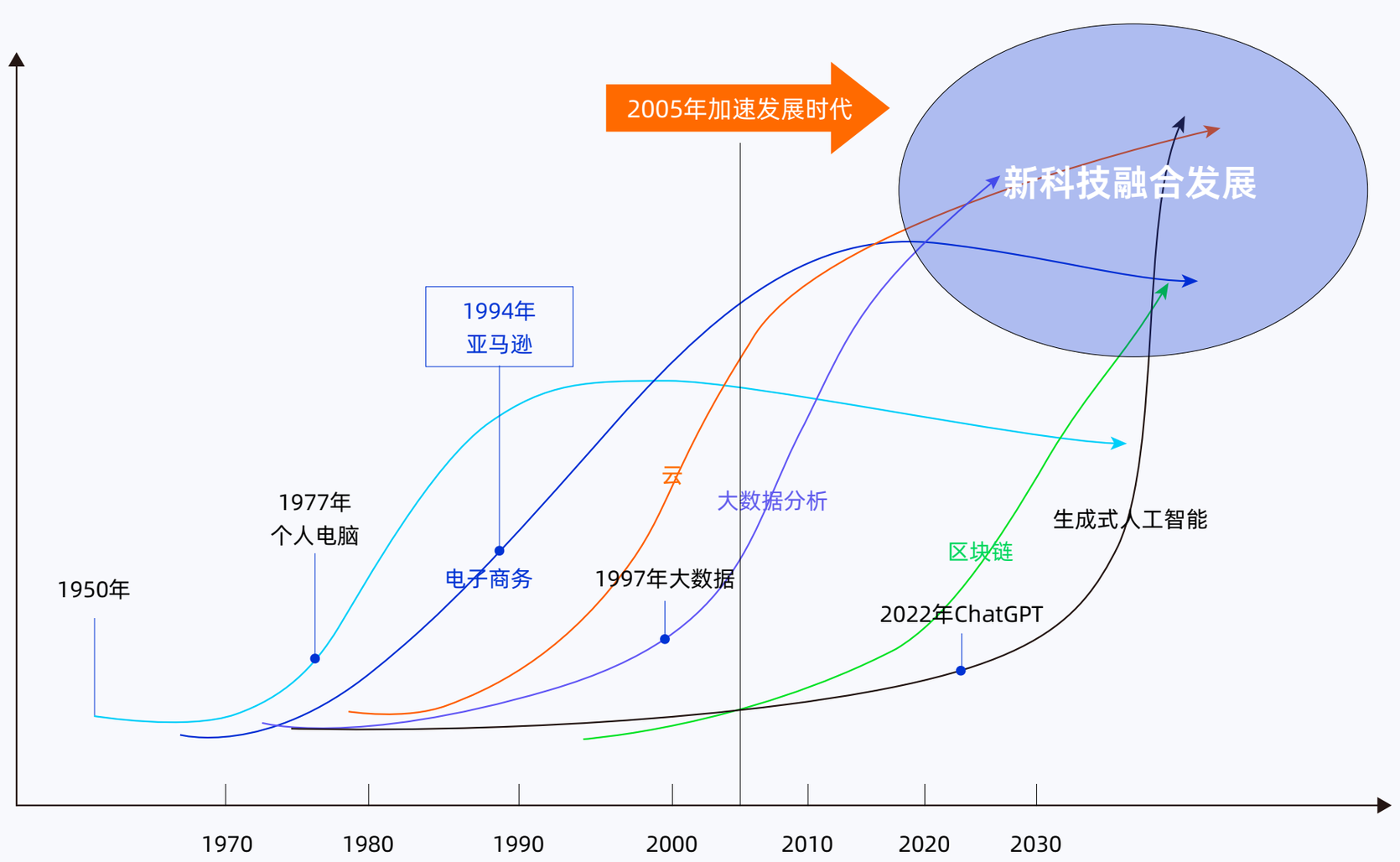
今天，面对新一轮智能革命，全世界都在寻找推动人工智能技术普惠的那部“蒸汽机”。当前全球大模型竞赛，虽然模型参数规模越来越大，数据集记录不断刷新，但真正能在产业中规模化、标准化应用的大模型还没有。我们还处在制造一台不能走进工厂的“蒸汽原型机”阶段。面向未来，唯有实现算力普惠和模型普及，才能为AI产业规模化铺平道路。人工智能模型训练对算力的需求只是起点，未来产业应用对算力的需求必将指数级爆发，公共云是实现算力普惠和模型普及的唯一路径。

2 技术普惠有三大内生要求，高可用、低门槛、可持续

高可用，让技术产品和服务无处不在，随时可用。并非所有技术都能发展成为产业，而一切真正成为产业的技术必须跨越工程化的门槛，解决高可用问题， “高可用”的本质是“可信任”。一项技术要被产业广泛接纳，并成为产业基础技术之一，除了其本身先进性之外，它还必须足够易得、可靠和安全。蒸汽机、电力、计算机、互联网全都如此，云计算、人工智能亦无例外。弹性计算服务等级协议SLA是国际通用的云计算可用性评价标准，只有那些能拿出SLA三个9、五个9（分别指SLA可用性99.9%和99.999%）的厂商才是真正的云计算厂商。

低门槛，让技术产品和服务经济易用，谁都能用。“低门槛”是一种关键能力。不够经济，人们用不起，是技术扩散的门槛；不够易用，能用起来的人不多，也是技术扩散的门槛。经济性和易用性通常伴随着规模化程度的提高而逐步达成，符合学习曲线规律。降低应用门槛是技术提供方的责任，它体现了技术提供方的持续投入水平和工程能力领先程度。正如IBM推广了PC、淘宝推广了电子商务、亚马逊和阿里云推广了公共云计算一样，这些公司之所以广为人知，不仅在于他们发展了关键技术，更在于他们围绕技术应用持续建设了丰富的服务体系，让中小企业和个人也能使用与大公司一样的技术，践行了低门槛的技术普惠理念。也正是这些低门槛的技术不断汇聚，为今天全社会的智能革命拉开序幕 。

图1-2 低门槛的技术产品和服务汇聚开启智能革命序幕



可持续，让技术产品和服务资源集约，绿色高效。工业革命以来，人类社会的发展不断提速，但对赖以生存的自然资源消耗也持续加速。在绿色发展的国际共识之下，面向未来的产业革命，资源集约、绿色高效必须成为关键技术的普惠要求之一。任何资源浪费的技术及其应用模式，都不可能成为下一代产业革命的基础。为智能时代的提供有效算力，以公共云并池调度，为社会提供“弹性和多租复用”的公共计算服务是实现计算资源集约利用的必由之路 。

3 通用技术跨越普惠拐点越来越快，降低人工智能技术成本十分迫切

几千年来，人类社会的技术可以分为两类：专用目的技术（SPT, Special Purpose Technology）和通用目的技术（GPT, General Purpose Technology） 。 通用目的技术是对人类经济社会产生巨大、深远而广泛影响的革命性技术。所有促成产业革命的关键技术，都是通用目的技术。

有史以来共有24种技术属于通用目的技术，可以按产品（P）、流程（Pr）和组织（O）分为三类：

产品类有14项：轮子、青铜、铁、水车、三桅帆船、铁路、铁轮船、内燃机、电力、机动车、飞机、计算机、互联网。

流程类有7项：植物驯化、动物驯养、矿石冶炼、写作、印刷、生物技术、纳米技术。

组织类3项：工厂体系、批量生产/连续过程/工厂、精益生产。

通用目的技术有四个特征：一是可以广泛地应用在各个领域；二是通过促进生产率提高、降低使用成本，颠覆性带来生产力跃升；三是与各类技术间存在着强烈互补性（complementarity），强烈外部性促进新技术创新；四是促进生产流通和组织管理方式变革，推动组织形态与生产关系重塑。

通用人工智能（AGI）将成为人类社会第25种通用目的技术。在工业时代，以蒸汽机、内燃机为代表的通用目的技术，替代、赋能体力劳动者；在数字时代，通用人工智能改变了人类知识检索、创造、运用的基本方式，是支持、赋能脑力劳动者的新生产力。

回顾历史，通用目的技术跨越普惠拐点所需要的时间越来越短。印刷术在德国古登堡1448年发明300年后，出版大众化廉价化才开始兴起；蒸汽机在瓦特1796年改良之后80年才具有普及的商业意义；电力发明50年之后一般民众才能普遍受惠；IBM大型机30年后才对一般人的日常生活有深刻影响。

而通用人工智能技术的扩散速度再次刷新记录。以ChatGPT为例，两个月用户数达到1亿，四个月用户数达到10亿。2023年1月，微软董事长兼CEO萨蒂亚评论说：“我一生中从未见过，至少在我从事科技行业的30年中，美国西海岸的先进科技可以在几个月内，以非常真实的方式出现在印度农村。”[2]

[1] Andrew Reamer，“The Impacts of Technological Invention on Economic Growth”, 2014.2
[2] 微软公司CEO、董事长萨蒂亚·纳德拉，在瑞士达沃斯举行的世界经济论坛公开发言，2023年1月。

1.2 计算革命，云计算是新纪元

数据只有通过计算才能释放价值。为了从数据中挖掘更多价值，计算在其漫长的发展进程中始终朝向普惠的方向。进入数字时代，数据要成为生产要素，必须使用公共云作为与之相适应的计算设施。唯有以弹性和多租，让共享计算资源池的规模效益充分发挥，计算才能人人能用、人人可用。这是云计算为全社会而非少数人所创造的技术红利。私有云是云计算发展的过渡形态。另外，一些倒卖服务器和GPU卡的商业行为，是硬件的简单堆砌，没有通过技术创造客户价值和社会红利，不是云计算 。

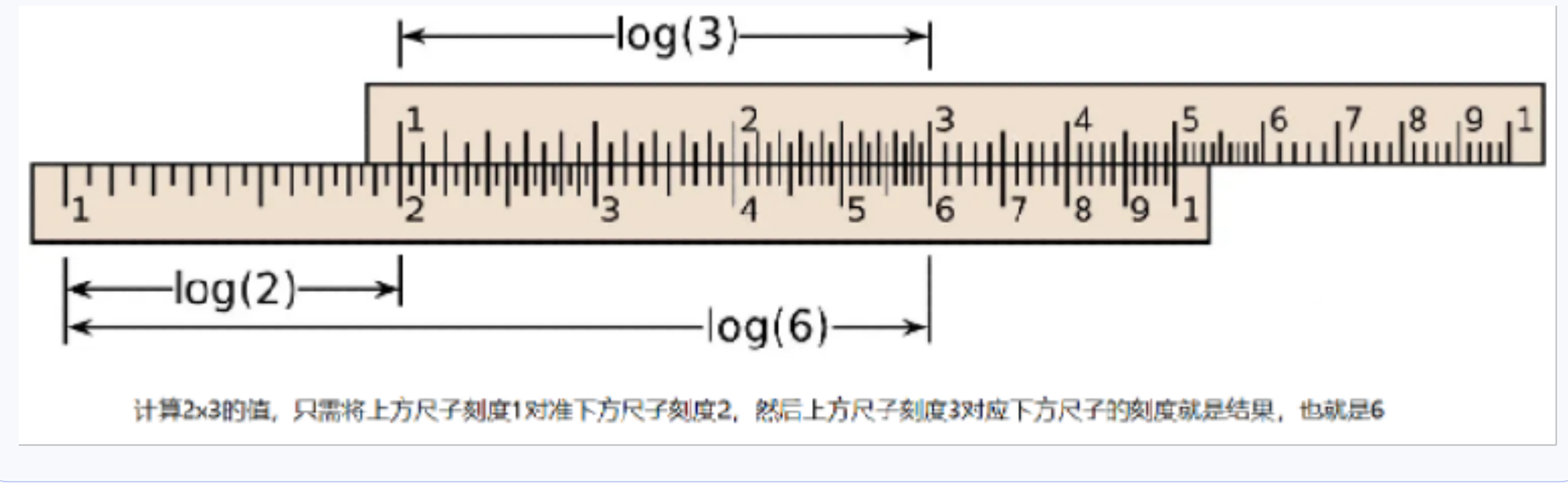
1 计算，让数据释放价值

计算，是从数据中获得信息和知识的过程。单独存在的数据没有意义，只有它们存在于一个计算式中，存在了计算关系，数据才具有了意义。“一个人每天走过的路，会有脚印落在地上。不去关注它，脚印就没有意义。而计算可以把他无数的脚印、走过的路、每天的心跳都变成数据，把这些普通的点点滴滴变成有意义的数据，从而为生活带来更美好的改变。这就是数据的价值。”[1]

计算在其漫长的发展历史中，普惠始终是方向。从手指算、垒石算到算盘，计算工具的发展，都是为了更方便地满足人们的计算需求。算盘在中国大约被使用2000年之久，一个很重要的原因是其成本低廉、易于制作且容易上手，人们通过背诵和使用珠算口诀就可以完成计算任务。

后来，计算尺的发明，虽然其背后的对数原理非常复杂，但却让使用者计算乘法变得非常简单。

图1-3 原理复杂但使用特别简单的计算尺



计算2x3的值，只需将上方尺子刻度1对准下方尺子刻度2，然后上方尺子刻度3对应下方尺子的刻度就是结果，也就是6

再后来，帕斯卡加法器、莱布尼茨步进式计算器、巴贝奇差分机先后出现，发明家们用精巧的机械结构把制作计算用表的工作不断简化，而计算用表又让航海、建筑等行业一线的工人完全不需计算就能查表获得结果，相当于享用了计算的服务。

[1] 王坚，中国工程院院士、阿里云创始人，《因为计算，创造了更多无法计算的价值》，2015.10

电气化时代来临后，赫尔曼为美国人口普查任务发明了电动打孔卡片制表机，后来发展成为国际商业机器公司（IBM）。在计算机时代，IBM等公司把通用计算机带给千家万户，Intel等公司让拇指指甲大小的芯片拥有超强的计算性能，苹果等公司用一部手机让大众享受到各种智能服务，他们的背后是谷歌、亚马逊、阿里巴巴等公司用互联网和公共云，让人类第一次可以随时随地享用计算服务。

只有计算成为公共服务，数据才能成为生产要素。中国工程院王坚院士曾经这样解释“计算、数据与互联网”之间的关系。“当互联网成为信息基础设施，像公路、港口、水、电、煤等一样，越来越成为国民经济各项事业发展的基础，越来越成为国民经济发展新的引擎之时，数据就会以更低的成本被自然沉淀，数据便成了生产资料。”不过，数据本身还不能释放价值，“要让数据产生价值，就需要计算，只有被计算处理过的数据才有用。”数据这时才从生产资料成为了生产要素。“计算一定要成为公共服务，才能让每个老百姓都能享受到计算为社会带来的价值。”把计算做成公共服务，正是王坚创立阿里云的初心。[1]

2 数字时代对计算普惠提出了更高要求，公共云应运而生

公共云的诞生有三大驱动力。

一是成本压力。1999-2001年，亚马逊公司的技术成本（研发工资、系统及通信基础设施开支）占销售收入的8-10%[2]。华尔街分析师认为贝索斯面临现金流枯竭，公司年度债务利息高达1.3亿美元[3]，公司股价最高下跌80%。为了解决经营困境，贝索斯提出“API六条”作为降本增效和创新增收的手段。亚马逊2006年推出云服务时，其公司年度经营利润率只有3.63%[4]。

阿里云诞生时的情景与亚马逊如出一辙。2008年时，阿里巴巴已是全亚洲最大的数据库用户，昂贵的服务器、数据库等投入和运维开支“躺着拿走淘宝的利润”。2009年底，淘宝技术预算负责人在给董事会的报告中提出“淘宝2010年起不再购买小型机”[5]

【案例1-1】AWS的创业始于无奈[6]

亚马逊是一家低利润零售商。贝索斯秉持“选品与便利-客户体验-更多流量-更多供货商-低成本结构-更低价格-更好客户体验”的“飞轮理念”，实现了线上零售快速扩张。

遭遇瓶颈。随着用户体验的快速变化，软件开发和迭代速度跟不上用户体验变化的速度。安迪·加西，即后来AWS的CEO，与开发团队沟通后发现：一个10-20多人的微小服务团队，即使发现了用户痛点，也一样要做服务器安装、搭存储、安装操作系统、开发环境、加载数据库等重复性工作，后期维护更是占去了一半的工程师资源，没有时间处理差异化的业务逻辑。另一方面，电商是一个薄利润产业，急速扩张下的运营带来季节性峰值挑战。亚马逊必须要为电商发明一个具有弹性的基础设施。

[1] 王坚，《在线》，中信出版社，2016.9
[2] 王坚，《在线》，中信出版社，2016.9
[1] 2001 Amazon财报
[3] Lehman：Revisiting Amazon’s Liquidity Issues
[4] 2006 Amazon财报
[5]《“云”和大数据，阿里巴巴去IOE之路》，2020.7
[6] 顾凡，AWS大中华区云服务产品管理总经理，《亚马逊技术背后的创新故事》，2020.8。

长期投资。解决完自己的问题后，亚马逊发现这也是整个市场共同的诉求，AWS由此创立并推出了第一个云服务S3。随后亚马逊长期投资把AWS做成云平台，陆续推出了更多云服务，成为今天全球规模最大的云服务厂商。

二是性能困境。2008年时，阿里巴巴所使用的庞大数据库已经难以维护，也无法继续扩展。“请来全国最顶尖的数据库管理员，一个晚上一张报表都做不出来。全球也没有任何一家公司能为阿里提供完整的技术服务”。同时，计算和存储的扩容周期也已经无法满足电商业务的需要，“为准备一次秒杀营销而做的IT系统扩容，从采购到部署至少需要半年。等系统准备好时，业务时机早过去了。”

【案例1-2】阿里云诞生于背水一战[1]

阿里巴巴“计算力”告急。几亿用户在淘宝购物要靠巨大的计算力支持。2008年，每天早上八点到九点半之间，阿里服务器的使用率都会飙升到 98%，离爆棚就差两个点。对业务增长最迫切的阻力来自IT基础设施瓶颈。按照业务增速估算，扩容开支足够让阿里破产。2008年中旬，马云召开内部会议决定，研发新的技术架构换掉旧引擎。新的计算引擎必须便宜、好用。这个想象中的云计算系统被定名为“飞天”。

云梯计划。2009年淘宝网只有一个季度勉强盈利，系统扩容入不敷出。阿里云启动云梯计划，解决淘宝网“大规模数据计算”需求。“云梯计划”关乎阿里生死，公司做了两手准备：“云梯1”用开源软件为基础研发数据计算系统；“云梯2”以“飞天”为基础纯自研一套数据计算系统。两座云梯以实现独立调度5000台服务器为线，谁先跑到为赢，这个目标写为5K。2012年底，以Hadoop为基础的“云梯1”实现了4000台集群调度，纯自研的飞天“云梯2”还在1500台集群的数量徘徊。三年没进展，让飞天成为笑话、王坚成为“骗子”，更让淘宝心急如焚。淘宝网预警两套云梯系统的存储和计算能力都将在2013年6月左右到达瓶颈，届时电商数据业务将会停滞。阿里云将近80%的同学转岗或离职。

All in飞天。王坚在重压之下，在2012年阿里云年会上泣不成声。马云宣布坚定支持后，阿里集合技术精锐背水一战，终于在2013年6月底让飞天5K通过“拔电源”稳定性测试，随即以飞天全盘替代IOE和云梯1。2013年5月，阿里最后一台小型机下线；7月，淘宝最后一个Oracle数据库下线。

持续投入。飞天随后又实现了单集群10K调度，进而实现了无限制扩展。2018年，飞天云计算操作系统获得中国电子学会颁发的特等奖。阿里云陆续推出神龙云服务器、POLARDB商业数据库、飞天2.0、机器学习平台PAI等系列云服务，成为围绕AI具备体系化能力的新一代云计算公司。

三是技术创新。2003~2006年，谷歌公司连续发表了四篇重磅论文，分别关于分布式文件系统（GFS）、并行计算（MapReduce）、数据管理（Big Table）和分布式资源管理（Chubby）。这些技术思想展现了谷歌公司如何应对数据激增所带来的计算挑战。虽然这些论文中所公开的技术未必是当时最先进的（谷歌云真正使用的云计算技术从未开源），但启发了更多的云计算从业者，为全球公共云的发展指示了方向。

[1] 史中，《阿里云的这群疯子》，2018.10

站在今天回看，无论是成本压力、技术困境还是技术创新，公共云诞生的三大驱动力有一个共同的本质，就是全球第一批互联网公司率先感受到，传统IT架构无法有效计算由互联网自然沉淀下来的海量、非结构化数据，所以不得不另起炉灶，从头自研。面向全新的智能时代，数据的规模仍在加速暴涨，模型训练和推理对算力的需求起伏更加跳跃，如张量数据等新型数据结构的演进还在加速，公共云正是专为应对这些挑战而生。

3、什么是云计算，理清云计算和计算的关系

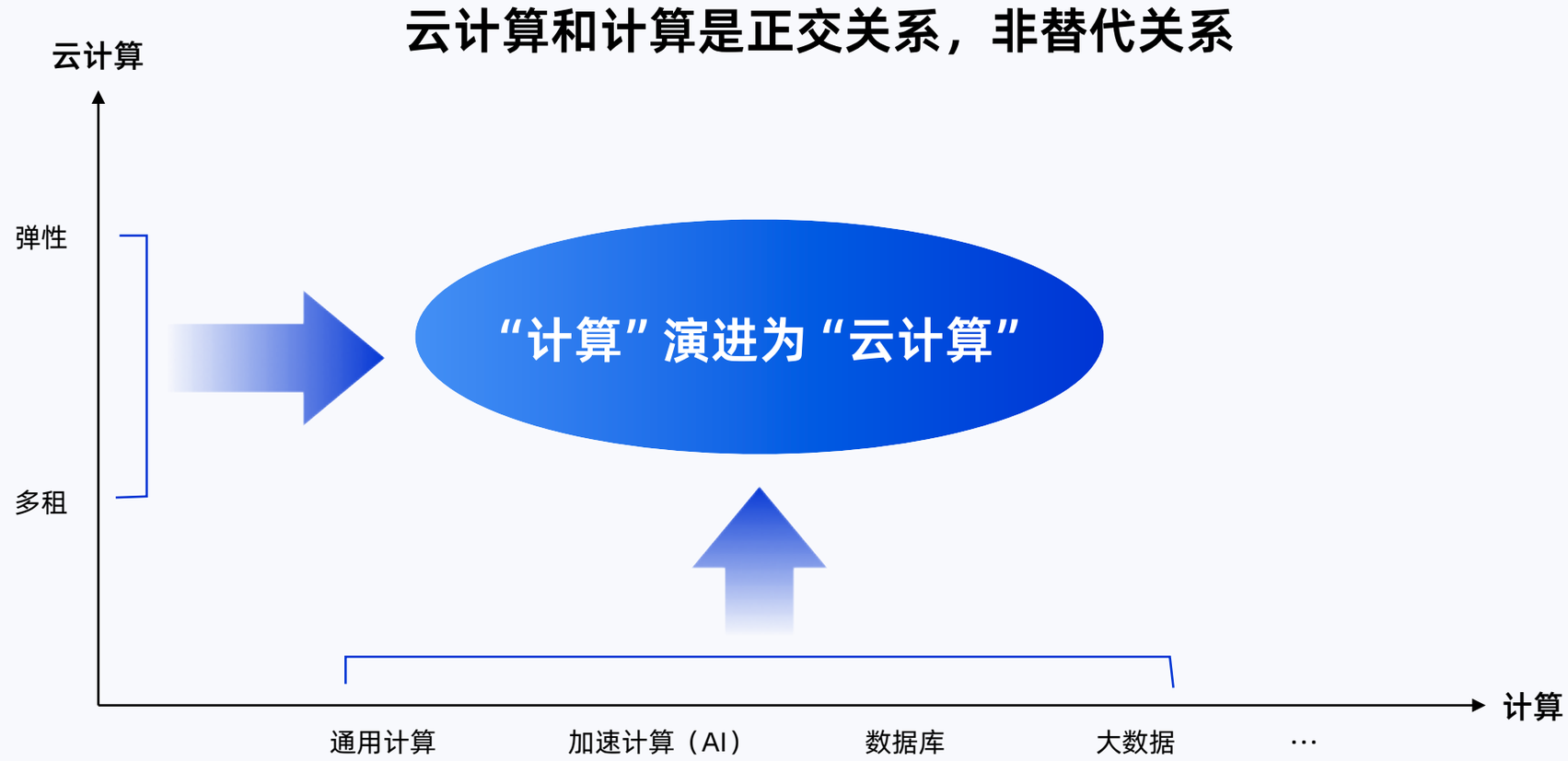
“弹性和多租”是云计算相对于计算的主要增量业务价值。2009 年美国伯克利大学对云计算做出了精确定义[1]，只有符合这六条定义的才是云计算。今天，有不少实质上属于倒卖服务器、倒卖GPU 卡的商业行为，没有通过技术为社会创造红利，没有创造客户价值，他们不是云计算，从长期来看他们是不可持续的。

实现云计算的核心关键是统一计算资源池。只有实现计算资源池的统一调度，同时技术层面实现多租安全隔离，才能够真正意义上实现全Region（地域）和全AZ（可用区）的计算资源弹性。

表1-1 云计算的精确定义

云计算的精确定义	解释
云计算是按需使用 无限量计算资源	线下IDC是基于本地部署（On Premise）供应，需在使用前预测需求，并提前完成服务器/交换机/存储的构建。而云计算服务是按需供应，根据实际业务需求做到计算、存储、网络等资源的极致伸缩。
消除云用户的预先承诺	公共云用户在任何一个区域（Region）或可用区（AZ），只需要付费成功就可以按量使用计算存储网络的资源，并不需要向云提供商承诺要在哪一个地方消费多少资源。因此，云服务商必须在全AZ可用区和全region地域供应各种算力需求 。
根据实际需要支付短期 使用计算资源的费用	用户不需要承诺购买的计算资源的预期使用时间。对于云计算所使用的容器或者函数，计时单位可以精确到毫秒级。用户购买后即使只用几毫秒也可以。当一个计算资源必须让用户使用四年、五年或者整个生命周期，即说明它不是云计算，而是 IDC或者私有云。
通过规模经济 显著降低成本	云服务商希望通过使用超大的数据中心，以规模经济放大供应链效益，摊薄研发成本。
通过资源虚拟化技术简化操作并提高资源使用率	云服务商希望通过资源虚拟化，把计算、存储、网络抽象后，达到简化操作，提升资源利用率。
通过多路复用的方式运行来自不同组织的负载 提高硬件资源利用率	核心是多路复用。通过多租户使用同一套计算资源大池，削峰填谷，保证公共云的硬件资源利用率显著提升。今天智算中心就面临大任务算力不够用（比如模型预训练生成），小任务跑不满（比如模型推理应用）的情况，本质上是因为它们没有统一的云架构 。

图1-4 云计算和计算的关系



云计算与计算是正交关系，而非替代关系。以AI 分布式训练系统为例，微软公司的开源DeepSpeed可以部署在线下IDC算力集群上，也可以部署在云端弹性算力集群上。DeepSpeed让分布式训练变得容易、高效和有效，它解决的是“计算”问题。而云计算的任务是当DeepSpeed系统部署在云端弹性算力集群上时，让AI分布式训练进一步做到极致的弹性多租、普惠和高效。

再如，NVIDIA公司在GPU加速计算方面的创新使其市值高涨，这是资本市场对其在”计算”维度创新的回报。而云计算公司的使命，不仅是持续地增强计算层面的竞争力，更需要通过弹性和多租的技术，让计算更普惠和高效，能让“妙鸭相机”这种创新应用的训练和部署成本变得足够低。

1.3 智能革命，公共云+AI是主引擎

云计算的发展自始至终只有一条技术路线，就是基于统一计算资源池，实现弹性和多租服务，即公共云。云计算的起点是公共云，亚马逊、阿里巴巴等互联网公司必须创造新的计算架构，解决内部降本增效，对外创新增收的问题。云计算的未来还是公共云，智能时代的算力需求比互联网时代有指数级增长，唯有对算力资源做好统筹和复用，才能让人工智能真正形成产业、进入产业，为全社会而非少数人创造更大价值。迎接智能革命，不是简单的单点技术演进，而必须发挥好公共云和AI技术体系的主引擎作用。

1 云计算的起点是公共云，未来还是公共云

公共云是云计算的起点。亚马逊首创云计算，就是为了构建统一的计算资源池，对内共享复用计算资源得以降本增效，对外还能创新增收。千万不要低估了亚马逊对创新增收的考虑。根据亚马逊2009-2012年的财报数据，其当期互联网基础设施的成本压力依然巨大，服务器、存储、交换机、商用软件等增长速度是收入增长的1.5倍，带宽、电费等占年收入比例达9-20%。亚马逊电商业务忙闲不均特性显著，基础设施日常利用率不到15%[1]。因此，AWS自成立的第一天起，就肩负着以公共云对外提供云计算公共服务的使命。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”[2]。

当前，计算的性价比提升面临着前所未有的挑战。摩尔定律曾经是芯片性能提升的主要驱动力，但由于物理空间的限制，芯片制造商正逐渐难以继续增加晶体管的数量。尽管摩尔定律尚未完全失效，但其有效性已受到质疑，且芯片的耗电量和冷却成本也构成了新的瓶颈。与此同时，企业的基础设施正在快速向计算密集型转型。工业自动化和数字孪生技术的发展，加之物联网设备的广泛部署，均产生了大量的数据和计算需求。特别是机器学习和生成型AI，需要处理海量数据的复杂算法，这些都对计算资源提出了更高的要求。

私有云服务仍然适用于标准化后台流程系统，常规的云服务虽然仍然能够为大多数的日常运营提供足够的支撑，但企业还需要通过高度优化和专用化的计算环境来驱动新数字化场景，从而产生竞争优势。然而，对于推动业务增长的数字化场景，其成本可能会吞噬整个企业的IT预算。企业必须在提升计算能力和控制成本之间找到平衡，领先的企业正在寻找新的方法来充分利用现有的基础设施，通过公共云专用化的高性能计算解决方案可能成为应对这一挑战的有效途径，以确保在数字化转型过程中保持竞争优势。

因此，公共云才是云计算的未来。面向智能时代，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。与此同时，不同类型的计算任务对算力类型的需求也不同，计算芯片价格昂贵且折旧迅速（如GPU卡折旧一般三年）。这些都对计算资源的利用率提出更高要求。如果不设法做好多元算力的调度统筹，必将因为算力设施空闲造成巨大浪费。这些浪费不仅带来投资压力，更将转嫁为用户使用算力服务的成本，拖累产业发展的进程。AI和公共云密不可分，就像OpenAI和Azure、DeepMind和谷歌云的关系一样。

[1] CSLA, 2009-2012 根据Amazon年报数据整理。
[2]《The Everything Store：Jeff Bezos and the age of Amazon》。

私有云只是云计算发展的过渡阶段。企业自建私有云，自行购置设备，并进行运维，扩容时周期长，难以快速匹配业务需求，导致企业业务用户，通常会留足算力冗余，造成算力闲置，且业务收缩时难以缩容进一步加剧闲置。此外，复杂的数字基础设施需要专业团队维护，增加了企业的人力成本，并不适合未来的高速发展。AWS EC2推出时定价10美分/小时，迅速受到中小企业和初创企业欢迎。其最常见的客户企业规模是拥有10-50位员工、年营收在100万-1000万美元之间，这些企业使用云服务的每月账单通常少于1000美元[1]。

这种新型的服务方式对整个美国IT市场带来不同程度的影响。斯坦福大学发布的研究报告显示，数据库、计算、存储领域的公司营收尤其受到冲击，包括IBM、Oracle、EMC、VMWare、HP、RedHat等[2]。2010年7月，美国宇航局NASA和美国IT公司Rackspace联合推出开源软件项目OpenStack私有云，意在与公共云抗衡。两年内，IBM、惠普、Intel、RedHat等传统IT巨头纷纷加入。

但是，十年后的今天，Rackspace自己已从“ALL in OpenStack”转变为“多云”战略，即同时支持AWS、Azure和OpenStack多种云平台；OpenStack基金会的名字更换为“OpenInfra基金会”；Mirantis（OpenStack圈子里的明星企业）宣布裁员；惠普与思科相继关停各自的OpenStack云；英特尔提前退出了与Rackspace合作的一个OpenStack创新项目；赛门铁克、Juniper退出了黄金会员席位。越来越多的公司已经意识到，公共云是提供智能时代算力保障的唯一出路。企业在拥抱开源寻求发展的同时，更要看清技术大势：私有云是云计算发展进程中的短期过渡形态。

在此背景下，企业应当考虑以下几点策略：首先，评估现有IT基础设施的利用率，识别并优化低效部分；其次，探索混合云策略，将非核心计算任务迁移至公共云，以释放私有云资源；最后，关注新兴技术，如量子计算和边缘计算，这些技术可能在未来提供显著的性能提升和成本节约。通过采取这些措施，企业可以在快速变化的数字化环境中保持竞争力，同时有效管理IT预算。在电力普及的过程中，也发生过同样的情况。电灯泡曾是家庭中唯一的电器，以至于当通用电气公司推出家用电烤面包机时，插头还要采用灯泡的接口形式。后来，为了方便更多家用电器的使用，出现了“集成在灯座上的两相插座”。今天，我们已经知道接入电网的标准插座。“灯座上的插座”是电力成为社会基础设施过程中的过渡形态。私有云就是云计算成为基础设施过程中“灯座上的插座”，是短暂的过渡形态。

[1] Morgan Stanley 2015.11。

[2] Stanford，《Amazon Enters the Cloud Computing Business》。

图1-5 “灯座上的插座”：从“灯座”进化到“插座”

2 理解人工智能对计算和云计算带来的挑战

计算视角出发，AI对计算体系结构的挑战有三个方面。

一是功耗墙。分布式大语言模型在过去两年对算力的需求增长了750倍。算力需求提升导致GPU功耗持续提升。2017年V100单卡功耗峰值250W，2020年A100单卡功耗峰值400W，2023年H800单卡功耗峰值750W，芯片制程和散热的挑战愈来愈艰巨。单机八卡H800当前整机功耗10KW，接近传统IDC机柜供电上限。功耗约束，也是促成Nvidia下一步超级芯片grace hopper技术的重要原因。

二是内存墙。大语言模型在显存受限的情况下，对内存带宽提出极高要求，因此GPU普遍采用HBM大带宽内存技术。当前GPU显存与系统内存的交互带宽受限于PCIe Gen5标准，已经成为日益严重的带宽瓶颈。因此各个GPU/CPU的硬件设计厂商都在尝试通过chip2chip的方式构建高速互联通道，绕开PCIe Gen5瓶颈限制。

三是通讯墙。由于单机八卡在算力和显存容量和带宽等诸多方面，均无法满足大语言模型的需求。通过数据中心网络互联技术，完成多机scale out分布式AI训练系统的需求就成为了刚需，大语言模型上下文，数据中心网络的大带宽、GPU direct RDMA等技术已经成为必须。

图1-6 计算视角下AI带来的挑战

在智能时代，计算的范式正在发生根本性的变化，AI计算的重要性正在超越传统计算。大模型驱动的AI计算最终将成为整个数字世界的基石，而且会成为整个数字世界计算资源的调度者和中枢神经，并最终接管以CPU为主的传统计算资源。AI对云计算基础技术体系提出了更高要求。

一是要做到AI训练、AI推理以及HPC超算的资源并池。只有多业务并池，才能提升资源弹性和资源利用率。AI训练业务属于离线业务，对于Region和Az的要求不高。但是如果AI训练业务和AI推理业务不并池统一调度，离线AI训练集群的资源闲置风险会放大。

二是新增的AI推理业务要能和已有的其他在线业务整体统一部署。如果部署AI推理业务要求整体迁移全部在线业务，不符合降低AI应用门槛的要求。

图1-7 云计算视角下AI带来的挑战

3 公共云是AI时代的必备基础设施

只有公共云才能成就人工智能的产业化。大模型是一场“AI+云计算”的全方位竞争，超千亿参数的大模型研发，并不仅仅是算法问题，而是囊括了底层庞大算力、网络、大数据、机器学习等诸多领域的复杂系统性工程，需要有超大规模AI基础设施的支撑。目前，AI与云计算相互依赖，技术创新和产业化与云计算紧密关联。算力成本、数据成本、商业闭环已成为制约AI发展的重大挑战，公共云是AI大规模普及的最佳方式。

首先，基础大模型的性能决定AI应用的水平，模型训练要求高性能、稳定和普惠的并行计算支持，全球领先的基础大模型优先基于公共云完成训练。

支持基础大模型训练，必须让万张GPU像一台计算机一样高效运行，具体有三方面要求：

一是要高效率实现高性能并行计算，对计算、网络和存储有特殊要求。在计算领域，借助虚拟化和容器化技术，可以构建大规模的CPU+GPU异构算力。形成标准算力单元，并通过适应性策略和敏捷框架进行精准匹配，从而实现高性能并行计算。这种方法不仅满足敏捷开发的需求，还能实现弹性灵活调度及快速部署；在存储领域，要针对大规模训练任务数据的高频、快速读取特点进行软件策略和硬件性能层面的优化；在集群计算网络领域，需要高带宽、低时延、性能稳定可预期网络技术支持。

二是要实现长时间稳定计算，对系统的故障处理和计算任务断点恢复功能有更高要求。在系统故障处理方面，要能迅速发现故障，迅速迁移计算任务到无故障的计算资源上。在计算任务断点恢复方面，要能在不损失训练性能的前提下，尽量对每一步训练迭代都设置可恢复的还原点，把故障造成的训练损失降到最低。

三是要实现普惠的计算服务，对算力的可获得性和成本具有更高要求。

在算力的获取方面，模型训练需要随时随地调取算力，实现秒级扩展。用户无需理会底层技术的复杂性和庞大架构，也不需要购买服务器，就能随时根据需求享受便捷的综合性服务。这种高效的算力使用方式，不仅提升了工作效率，还显著降低了技术的门槛，使更多用户能够集中资源于核心业务和创新。

在算力成本方面，本地构建和维护AI基础设施的成本非常高昂。大模型训练所需的算力资源消耗巨大，费用也非常高，短期内难以回收投入成本，导致投入产出比低。因此，需要寻求低成本的训练算力解决方案。公有云提供即用即付模式，使组织能够访问强大的资源，而无需进行大量前期投资。这种模式推动了人工智能的普及，使各种规模的企业都能够使用它。

理论上公共云和私有云都能提供大模型训练的解决方案。但由于算力基础设施的技术挑战、建设周期、综合成本等原因，全球领先的基础大模型事实上都选择了基于公共云完成训练。比如，OpenAI基于微软公共云训练出了GPT-4o大模型；Anthropic基于亚马逊公共云训练出了Claude大模型；谷歌基于谷歌公共云训练出了Gemini1.5 Pro大模型；xAI公司基于Oracle公共云训练出了Grok大模型，如表1-2所示。

表1-2:全球领先基础大模型训练芯片规模及部署方式

云	基础大模型	GPU型号及数量	部署方式
OpenAI	GPT-4、GPT-4.5、GPT-5	2.2 万张 A100、预计几万张卡 预计几万H100卡、甚至10万卡	微软公共云计算集群
亚马逊	Anthropic的Claude	1.6万张H200	亚马逊公共云集群
谷歌	Gemini、Gemini Ultra	TPUv5（是GPT-4的5倍）	公共云，支持客户模型训练 跨多个数据中心的大量 TPUv4 利用AWS的云
Meta	LLaMA、LLaMA3	2048张A100 未公布，估计至少2万张A100	拥有1.6万AI超级计算机集群，用于训练和推理，服务全球数十亿用户 预计到2024年底，将拥有35万张H100，公司算力总水平将相当于近60万块H100。
xAI	Grok 1	数万个 GPU 集群	外媒 Techcrunch 透露，可能是由 Oracle 提供(财报显示，Oracle是xAI的云服务商。)

数据来源：公开材料整理

其次，模型应用需要高性价比、高弹性、高可用和就近合理分布的算力支持，公共云是AI规模化应用的必然选择。

模型应用与训练所需算力的要求完全不同，更强调计算的性价比和可得性。

一是算力的性价比直接影响AI应用的规模，公共云更能发挥规模效应优势，能持续降低推理算力价格。

过去五年，模型应用所需的推理算力，随模型参数规模的增加而快速攀升。GPT4 的推理算力消耗，比第一代GPT扩大了1.5万倍。与此同时，为了促进模型被更广泛应用，OpenAI在过去一年内完成了三次降价。以GPT 3.5 turbo为例，当前版本每生成1000 个token （约750个英文单词或400-500个汉字）收费0.002 美元，半年下降了70%，如表2所示。

表1-3:GPT 3.5 turbo 不同版本价格（美元/1000token）

不同版本	输入价格	输出价格	部署方式
2023年6月14日更新	0.003	0.004	0.007
2023年11月7日更新	0.001	0.002	0.003
2024年1月29日更新	0.0005	0.0015	0.002

数据来源：公开材料整理

二是推理算力要具有高弹性、高可用的能力，公共云是支撑大范围、高并发AI应用访问的必然选择。

2023年1月，OpenAI 的网站访问量高达6.72亿次，3月最高达16亿，较2022年11月份的访问量增长了850倍。OpenAI发布新模型版本后，由于短时间内用户访问负载过重，导致服务中断近两小时。凸显了模型应用对极致弹性、高可用算力的迫切需求。

公共云的弹性伸缩优势，能够有效地应对AI大模型推理应用过程中产生的大规模计算需求及流量波动的变化。云计算服务是按需供应，根据实际业务需求做到计算、存储、网络等资源的秒级扩容或资源收缩。在面对诸如GPT 4等服务出现的用户访问高峰时，公共云能够通过快速扩容来应对流量洪峰，确保服务稳定性和响应速度。当用户访问回归平稳时，公共云能迅速收缩、释放闲置算力资源，控制算力成本。这种即时、灵活的资源调度能力，使得无论是在常态下的高效服务提供，还是在突发流量下的快速响应，公共云都能为AI大模型应用与服务提供最佳的支持。

三是部分模型应用对延时较为敏感，依赖于就近可得的推理算力服务，公共云在全球范围内具有更强的分布式覆盖能力。

部分在线类的模型应用对响应的时间要求高，比如生产控制、金融交易、医疗诊断等AI应用，要求推理算力能就近位置可获得。公有云将在实现边缘AI方面发挥关键作用，提供在更靠近数据源的边缘部署和管理AI模型所需的基础设施和服务。公共云在全球范围内大规模部署地域分散的服务节点，与边缘云和物联网终端共同组成“云边端一体化协同”的混合云服务形态，满足不同行业差异化的算力服务需求。根据应用调用发起的位置，就近提供计算和网络覆盖，优化模型响应的网络延迟，通过稳定、高速、安全的网络，实现算力和连接的全局优化与高效协同。

最后，随着人工智能的普及，安全和隐私问题将愈加突出。公有云提供商必须优先考虑强大的安全措施和数据治理实践，以确保人工智能的负责任使用。

综上所述，公有云完全有能力满足日益增长的人工智能需求。其可扩展性、成本效益以及对尖端技术的访问，使其成为开发和部署AI解决方案的理想平台。随着人工智能的不断进步，我们可以预见公有云提供商将在塑造这一变革性技术的未来中发挥更为重要的作用。

4 大模型加速云计算技术体系升级

公共云的快速演进正在推动以模型为中心的新一代云计算技术体系的形成。AI大模型所需的计算架构在现有的IaaS、PaaS、SaaS三个层面上，进一步发展出了MaaS层（Model as a Service），从而推动了整体云计算架构的升级。

AI大模型的工作负载资源密集，需要大量的计算能力和存储资源。公有云具备无与伦比的可扩展性，使组织能够根据需求快速扩展或缩减资源，这种弹性对于处理AI大模型工作负载的动态性质至关重要。公有云提供商一直处于AI创新的前沿，不断开发和提供新的AI服务、框架和工具，使企业无需具备内部专业知识即可获得最新和最好的AI技术。

此外，公有云提供预构建的AI平台、API和工具，简化并加速了AI开发过程，使开发人员能够专注于构建和部署AI解决方案，而无需管理复杂的基础设施。公有云还为人工智能开发营造了一个协作环境，提供对海量数据集、预训练模型以及充满活力的开发人员和研究人员社区的访问，从而促进知识共享和创新。

图1-8 以模型为中心的新一代云计算技术体系

表1-4 新一代云计算技术体系分层功能

SaaS层	应用效果“类人化”驱动场景创新。AI将以更“类人化”的应用效果，广泛应用于聊天机器人、虚拟助手、文本生产、文本摘要、语音识别等多领域。
MaaS层	MaaS（模型即服务）作为AI大模型服务层应运而生，将演变成一种新型的AI基础设施。
PaaS层	AI算法库与工程框架是核心。AI算法库是人工智能知识体系的“树根”，AI工程框架显著降低模型开发的门槛。
IaaS层	异构芯片成为AI计算的关键计算资产，集成GPU、TPU、NPU等不同架构的芯片，能够更高效的支撑AI大模型进行各类复杂任务的训练及应用。

公共云是智能化时代的创新引擎，也是AI发展的基石。唯有 “云智一体”才能让智能创新触手可及，让产业全面迈向智能。全球产业界关于AI大模型的这一轮竞争，核心是“大模型训练能力+公共AI服务能力+公共计算服务能力”的竞争，最终比拼的是谁能用最低的成本、最高的效率、最专业化的技术、最大化地挖掘数据要素中的无限价值。

2. 激浆：全球云计算千帆竞渡

公共云已经成为全球主要IT基础设施，正从技术、产业、生态等方面构建起完整的产业生态体系。领先者与追赶者的差距，从核心技术到产业体系呈现层层放大的“喇叭形”结构。美国领跑全球，中国加速追赶，但在云计算理念认知、发展战略及行动落地方面还存在差距。

2.1 技术：单点突破和系统创新

智能时代的公共云的竞争，不仅仅是芯片、操作系统、数据库等单点技术的突破，更需要的公共云、AI大模型、开源开放等体系化能力提升。

过去十年，云计算单点技术各有特色，中国企业有赶超之势。亚马逊拥有Nitro、Graviton等自研芯片、AWS云计算操作系统、RDS数据库；微软拥有自研AI芯片Athena、Azure云计算操作系统、Azure SQL数据库；阿里云拥有自研服务器芯片倚天、AI芯片含光、飞天云计算操作系统、龙蜥服务器操作系统，PolarDB和AnalyticDB等数据库。这些核心技术共同构成了传统意义上的云计算系统能力：计算、存储、网络、安全。

中美头部云计算厂商领跑全球。Gartner在2021年对全球主要云厂商的能力评分显示，阿里云在计算，存储，网络，安全的单项得分均为全球第一；且与2020年相比，阿里云产品能力得分项从168项提升至203项（共270项），补齐35项产品能力，说明中美企业之间在传统云计算领域的技术差距逐渐缩小。

表2-1 全球主流云厂商核心技术

公司	芯片	操作系统	数据库
亚马逊 AWS	Nitro、Graviton、Inferentia、Trainium	AWS	RDS
微软 AZURE	Azure Cobalt、Azure Maia	Azure	Azure SQL
阿里云	倚天、含光	飞天、龙蜥	PolarDB、AnalyticDB

评比类别	阿里云名次	AWS名次	Azure名次	Oracle名次	GCP名次
存储	第一	第二	第三	第四	第五
云服务商管理&审计	第五	第二	第一	第三	第四
安全	第一	第一	第三	第四	第二
网络	第一	第二	第四	第三	第四
计算	第一	第二	第四	第五	第三
运营&治理	第三	第一	第二	第四	第四
韧性	第四	第二	第三	第一	第五
数字业务基础设施	第三	第一	第二	第五	第四
软件基础设施	第二	第一	第三	第五	第四
IaaS基础能力（计算+存储+网络）	第一	第二	第三	第四	第五
IaaS+PaaS基础能力 （除云服务商管理&审计）	第二	第一	第三	第四	第五
总排名	第三	第一	第二	第四	第五

表2-2 全球云计算厂商基础技术比较

数据来源：Gartner 2021对五家主要云服务厂商的Solution Scorecard评估报告

智能时代云计算技术体系加速演进，竞争升级。AI大模型崛起，推动云计算进入了“AI+云”的新一轮竞争。OpenAI+微软云，做出了ChatGPT；DeepMind+谷歌云，做出了颠覆了生命科学传统研究的AlphaFold。AWS有云但缺AI，Meta有AI但缺云；微软同时涉足云计算和AI技术领域，发展大大提速。智能时代技术比拼的大幕已经拉开，这场竞赛不是单点技术的竞赛，而是芯片、网络、计算、模型全体系技术综合能力的竞争。阿里云也已经不是传统意义上的云计算公司，正在围绕AI全面提升云计算的体系化基础能力。

AI架构格局	模型的服务 商业/开源社区	无	OpenAI （商业）	Hugging face 开源			无		魔搭 Modelscope 开源	千帆 （商业）
	AI计算架构 及市场分割		Pytorch +TensorFlow	无	推出Pytorch （16.7%国际） （中国近3成）	推出 TensorFlow （78%国际） （中国近3成）		Mindspore 不支持Tf和Pt （0.05%国际） （中国近10%）	Pytorch +TensorFlow	飞桨 Paddlepaddle （4.6%国际） （中国近3成）
算力分配能力	可实现的算力 存储 颗粒度	容器 或依赖Azure	1、裸金属（最初级：传统分配物理服务器） 2、容器级（中级：可以将物理机虚拟化成为若干虚机后再分配） 3、任务级（最优级：实现算力资源按照任务分配）					1、裸金属 2、容器级	1、裸金属 2、容器-ACK 3、任务-PAI	1、裸金属 2、容器
	算力调度 （公共云或集群）	DGX cloud 托管给Azure	Azure 公共云	AWS公共云 （Meta使用AWS）		GCP公共云	无	Atlas （各地智算集群）	飞天 （公共云）	百度智算 （自用的集群）
传输加速架构	数据中心的网络架构RDMA网络服务器间加速	InfiniBand	InfiniBand	Scalable Reliable Datagram （SRD）	RoCE	Aquila	InfiniBand	RoCE	RoCE	RoCE
	芯片核间和 多个芯片之间 带宽情况	1、Nvidia GPU卡，内部核间双向通道的传输速率是50GB 2、NV link 对的CPU-GPU、GPU-GPU间的速率，新一代的NV switch3 可以达到900GB					OpenVINO 框架	国内情况不明 之前买的应该可以达到早期的300GB/s水准 而正常的PCIE一般在64-128GB/s		
芯片硬件架构	目前是否主要用 或支持CUDA芯 片计算加速架构	均主要使用Nvidia的GPU，获取CUDA生态资源						达芬奇DaVinci （<10%国内）	均主要使用Nvidia的GPU 获取CUDA生态资源	
	独立研制 AI芯片情况 （面向未来）	GPU A100 （>85%国内、 国际）	Athena计划 Azuce Cloud AI 2023年11月发	训练：Trainium 推理：Inferentia	MTIA芯片 ASIC架构 功耗仅25W	TPU v4 A100 的1.7倍 国际市场	Gaudi 2 供给中国	昇腾 NPU 算力A100相当	寒光 倚天	昆仑芯片
	部分主要 软硬件厂商	Nvidia （硬件）	Microsoft （互联网）	Amazon （互联网）	Meta （互联网）	Google （互联网）	Intel （硬件）	华为 （硬件）	阿里 （互联网）	百度 （互联网）

图2-1 智能时代的技术体系竞争

首先，阿里云通过自研软件的能力突破，将成千上万颗芯片高效连接并且调度起来，形成规模化的高效算力供给和调度分配，不再依赖单一芯片处理能力。以阿里云适配的某国产芯片为例，单芯片性能只有主流芯片性能的15%，但通过阿里云计算系统调度后，100颗国产芯片的整体处理能力与75颗主流芯片的整体处理能力相当。阿里云已建成88个云计算数据中心覆盖全球29个地区，为AI发展提供算力支撑。

其次，在自研云计算基础上，阿里云也是国内最早开始研发这一代AI技术的企业。2014年推出了国内首个人工智能平台PAI（Platform of AI）；2017年成立达摩院继续把AI作为重点研究方向；2021年做出了国内第一个百亿级参数规模的大模型；2022年在国内第一个发布了自研通义大模型，并率先提出Model as a Service的MaaS理念；2023年在国内第一个提供开源的百亿参数规模大模型，在云栖大会上发布千亿级参数规模的大模型通义千问2.0。

第三，阿里云自研并兼容各种国产化硬件。通过自研AI芯片、高性能的云计算网络设备、存储系统，阿里云围绕AI计算全面升级了软硬件体系架构并可以兼容其他国产硬件。

2.2 产业：产业规模和创新能力

公共云产业层面，IaaS、PaaS、SaaS等产业结构比例体现了发展的成熟度，企业利润及研发投入决定了产业发展的创新动能和可持续能力。

1 从产业规模看，中国公共云市场增速放缓

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2763.6亿元[1]。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年同比增长11.4%[2]，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%[3]。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

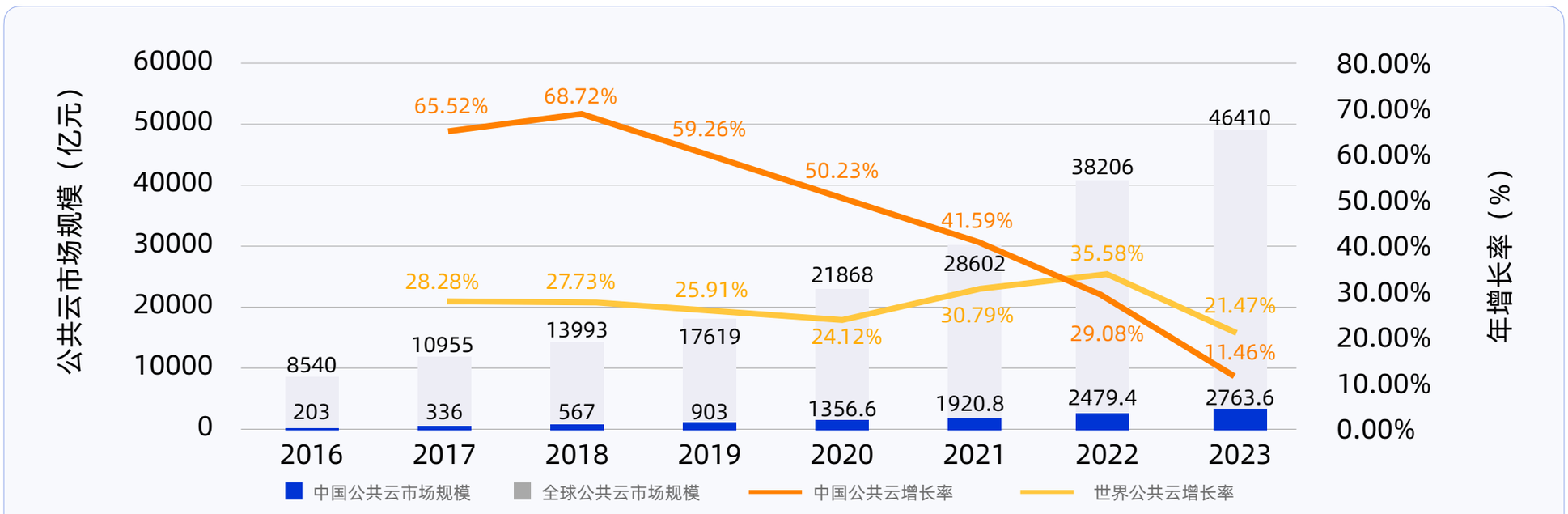


图2-2 2022年全球与中国公共云市场结构　数据来源：Gartner

[1] 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。

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

[3] 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》

全球公共云IaaS市场，按市场份额排名的主要厂商依次为亚马逊（39.03%）、微软（23 %）、谷歌（8.18%）和阿里云（7.94%）。其中，阿里云是我国唯一排入全球前四的国产厂商，但仅为领跑者亚马逊的五分之一左右。

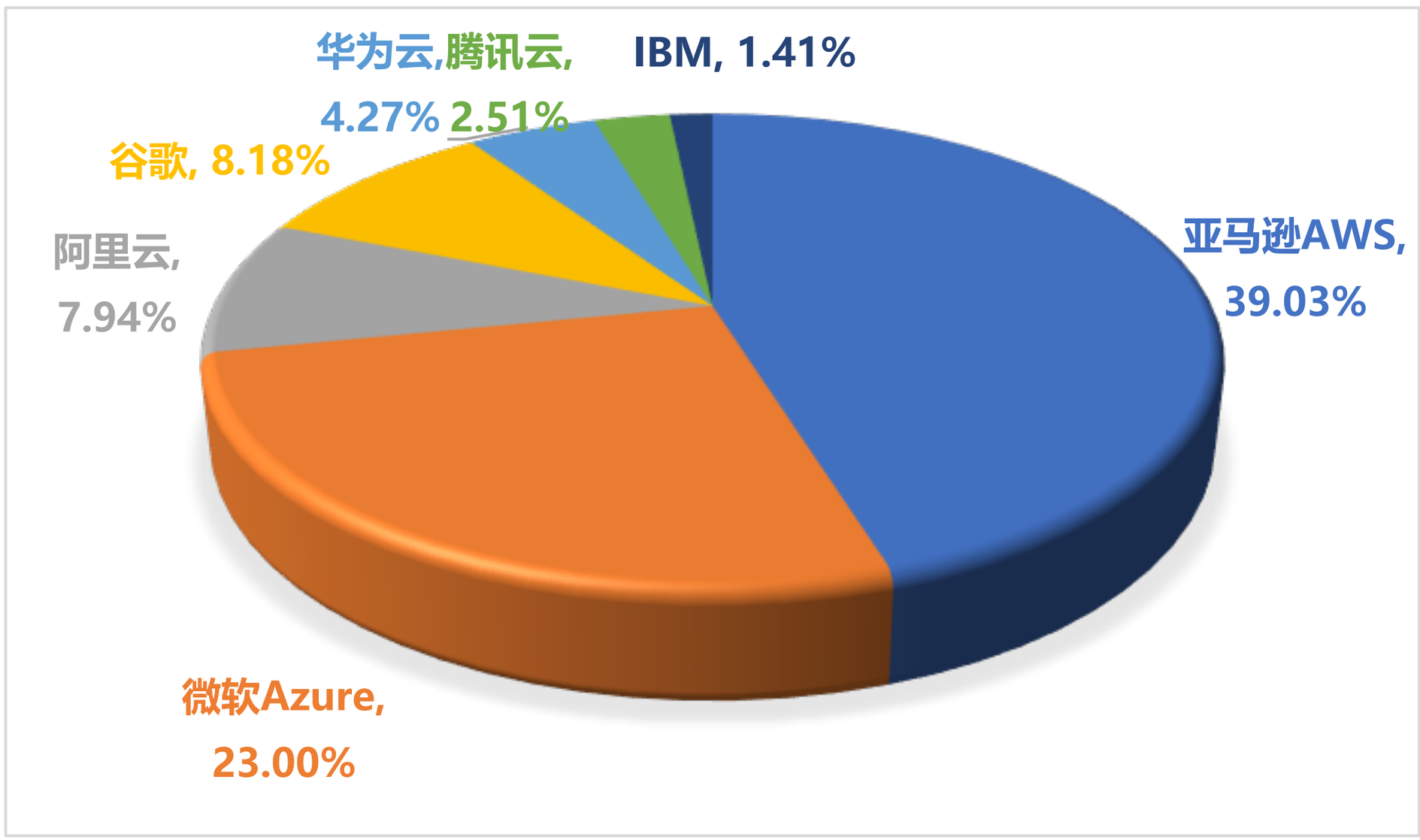


图2-3 2022年全球与中国公共云市场结构　数据来源：Gartner

2 从利润水平看，中美云服务商差距拉大

2022年-2023年，中国、美国头部云厂商差距在持续拉大。2021年是中美头部云厂商差距最小的时候。当时，美国三大云厂商（微软云、亚马逊AWS、谷歌云）的营收规模分别是阿里云的6.6倍、5.5倍、1.7倍。然而到2023年上半年，美国三大云厂商（微软云、亚马逊AWS、谷歌云）的营收规模和阿里云的差距扩大到了11.1倍、8.2倍、2.9倍。

阿里云的营收剔除了阿里体系内收入；图中的倍数为微软云、亚马逊AWS、谷歌云当期收入。

从利润看，亚马逊AWS、微软云、谷歌云在2022年掌握了全球66%市场份额，营业利润总和超3700亿元。相比之下，中国七朵云全球份额低于20%，七朵云大部分在战略亏损，2022年营业亏损总和超百亿元。[1]

营收和利润差距，也从一个侧面反映出中国云计算产业的成熟度低，还没有进入市场激励创新的正向循环。

[1] https://yunzhidao.alibabainc.com/yunzhizhen/market/globalPublicCloudMarket）
数据来源：《财经十一人》5月4日文章，《中国算力，雄心软肋》

3 从研发投入和创新能力看，中美企业差距更大

在研发投入方面，美国云计算企业研发/资本的支出差距较大，如图所示。研发投入的规模决定产业创新的速度。

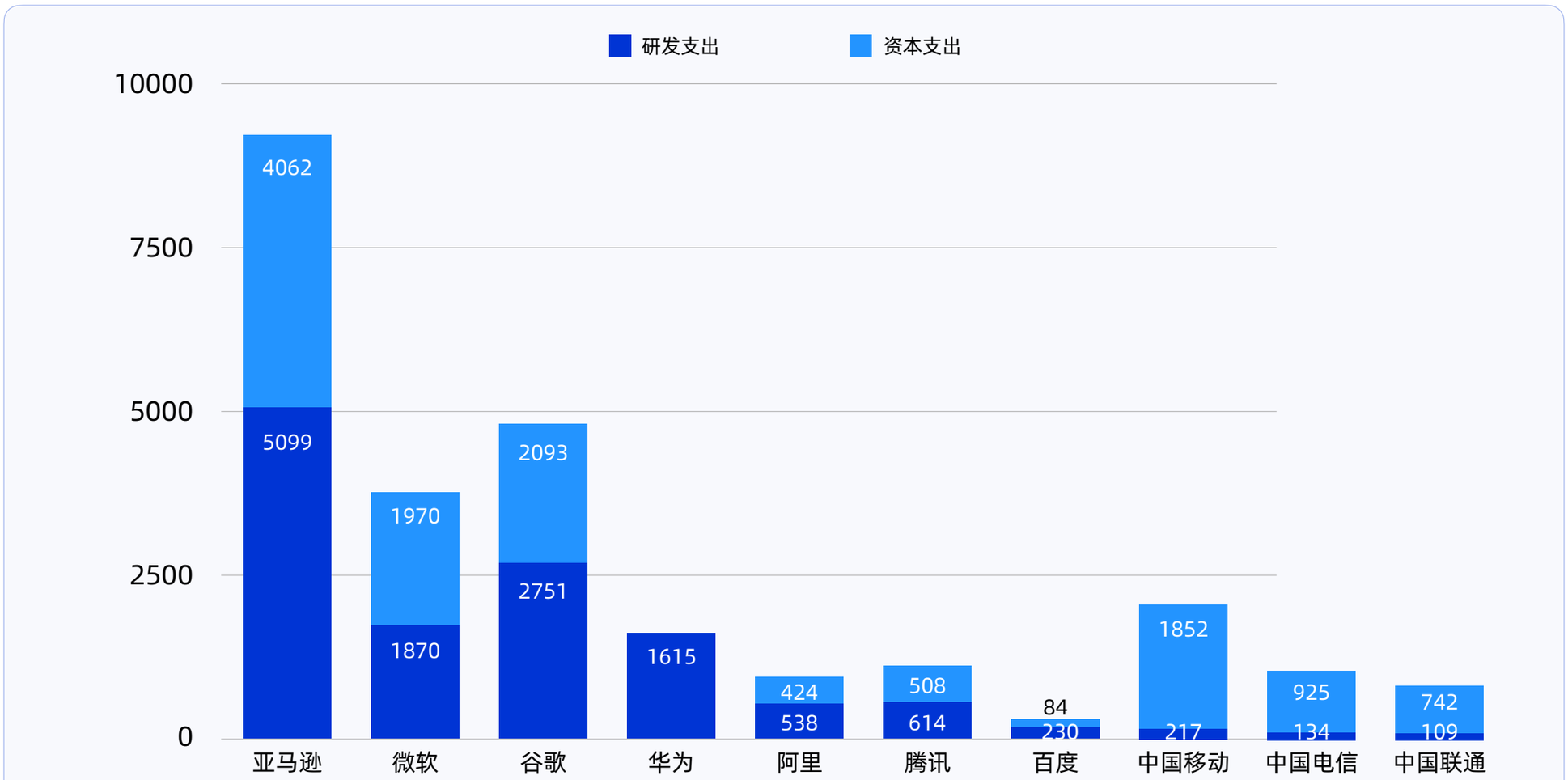


图2-4 主要云计算企业研发/资本支出占比（2022）[1]

根据欧盟委员会发布的《2023年工业研发投资记分牌》报告，2022年全球研发投入排名前2500名的企业中，美国企业占据了802席，总研发投入高达5010亿美元。而中国企业仅有768家上榜，总研发投入为3590亿美元。

在研发投入占营收比例方面，美国企业通常将更高的比例的收入投入研发。以科技巨头为例，亚马逊、谷歌、微软等公司的研发投入占营收比例普遍在10%以上，部分企业甚至超过20%。而中国企业研发投入占营收比例普遍较低，多数企业在5%以下。

在研发投入方向上，美国企业在基础研究、应用研究和产品开发等各个层面具有明显优势，并在多个前沿科技领域保持领先地位。美国企业在基础研究领域的投入更大，更注重原始创新和突破性技术。而中国企业则更注重应用研究和技术商业化，在基础研究领域的投入相对较少，影响了整体创新能力的提升。

在研发人才方面，美国在吸引和培养顶尖科技人才方面具有优势，拥有更完善的科研体系和人才培养机制。中国近年来在人才引进方面取得了一定进展，但高端人才仍然相对缺乏。

此外，美国拥有成熟的创新生态系统，包括风险投资、孵化器、科技园区等，能够有效支持企业的研发活动，并促进科研成果的商业化。中国的创新生态系统正在快速发展，但在风险投资和科研成果转化方面仍有待完善。

[1] 数据来源：公司财报。《财经》记者整理。

2.3 生态：生长环境和开放开源

产业生态体现产业实力。公共云是基础设施，产业生态是公共云价值的放大器，是云计算产业发展的果实。产业生态是否繁荣，一方面取决于云计算的开放开源程度和服务水平，另一方面，也取决于政府、企业对公共云的认知和政策等生长环境。

过去十年，美国是云计算产业生态最繁荣的国家。美国在全球SaaS市场中的占比超过七成，是全球最大的SaaS市场。美国拥有SaaS企业约1.2万家，其中SaaS上市企业约300家，总估值约8万亿美元。市值TOP10的美国本土SaaS企业，其总市值超过1.1万亿美元，其中有4（Adobe、Salesforce、Intuit、ServiceNow）家的市值均超过了1000亿美元。

对比来看，中国SaaS市场发展不足。截至11月23日，十家中国软件公司总市值为5238亿元（约合733亿美元），仅为美国前10强SaaS企业的6.5%。其中仅有2家软件公司市值超过1000亿元。中国市值前10的软件公司上半年营业利润仅为19.7亿元（约合2.8亿美元），仅为美国前10强SaaS企业的3.3%。研发支出为65.4亿元（约合9.2亿美元），仅为美国前10强SaaS企业的7.3%。

表2-3：中国SaaS 企业市值前10强

企业	领域	半年营收	营业利润	毛利率	营业利润率	研发支出	成立时间	市值
金山办公	办公软件	21.7	6.4	86.1%	29.5%	7.2	1988	1466
宝信软件	钢铁工业	56.8	13.5	39.5%	23.8%	6.4	2000	1024
恒生电子	金融证券	28.3	4.6	72.0%	16.3%	11.7	2000	605
用友	企业管理	33.7	-4.7	48.9%	-13.9%	10.1	1995	604
金蝶	企业管理	26.2	-4.9	61.9%	-18.7%	7.4	2001	386
广联达	建筑施工	30.7	2.9	85.5%	9.4%	9.1	1998	344
石基信息	酒店餐饮	12.0	0.4	49.5%	3.3%	3.5	1998	305
中科软	保险政企	26.3	2.1	33.4%	8.0%	4.8	1996	196
卫宁健康	医疗健康	11.9	-0.5	40.7%	-4.2%	3.5	2004	172
中望软件	研发设计	2.8	-0.1	97.5%	-3.6%	1.7	1998	136
合计		250.4	19.7			65.4		5238

资料来源：公司财报，《财经十一人》整理

备注：市值选取11月23日收盘价；营收、营业利润、毛利率、营业利润均为2023年最新半年报财报。单位：亿元。

2022年中国SaaS市场规模只是美国市场规模的8.3%，两国排名前10的SaaS上市公司市值相差15倍。[1]

智能时代来临，开源共享激发了大模型的持续创新。从大模型技术发展路线看，大模型主要是基于开源的技术，即Google 的 transform模型。当前全球主流大模型，基本上都是从transform 中的解码器和编码器等技术演化而来，形成了全球大模型的繁荣图景。

美国积极发展AI开源开放社区，推动AI大模型创新发展。hugging face是全球最具影响力的大模型开源社区，截止到2024年4月，有超过46万开源模型，超过200种语言，企业用户超过1.5万家，包括微软、谷歌、英特尔这些大企业都是它的用户。此外，Open AI也正在打造开放的应用市场，自11月7日发布GPTs 应用市场以来，两个月时间到2024年1月 应用超过300万个。

阿里云搭建和运营的魔搭社区(ModelScope)，从发展一年多以来，魔搭社区成为国内最具影响力的AI模型社区之一的AI模型社区，汇聚5500多款优质模型和上千数据集，为超过560万开发者提供了模型及免费算力服务，模型的累计下载量超过亿次。

为了推动AI普及，一方面，魔搭社区为每位AI新开发者提供100小时的免费GPU算力，已累计提供超过3000万小时的免费GPU算力。这些对AI开发者的巨大投入与支持，体现了阿里云为降低AI创新者成本，培育AI创新生态所作的努力。另一方面，阿里云推出了一项面向全国高校的重磅计划——“云工开物”计划，为中国4000多万高校学生每人送一台云服务器，希望帮助广大青年运用云和AI探索科技创新，中国高校在读大学生均可在官网上免费领取university.aliyun.com。

开源开放对产生生态具有重大的促进作用，中美对比来看，以公共云为主要基础设施的美国收获了更多AI初创企业。全球AI初创企业的投资份额中，美国2023年占比达50%，中国份额为30%。

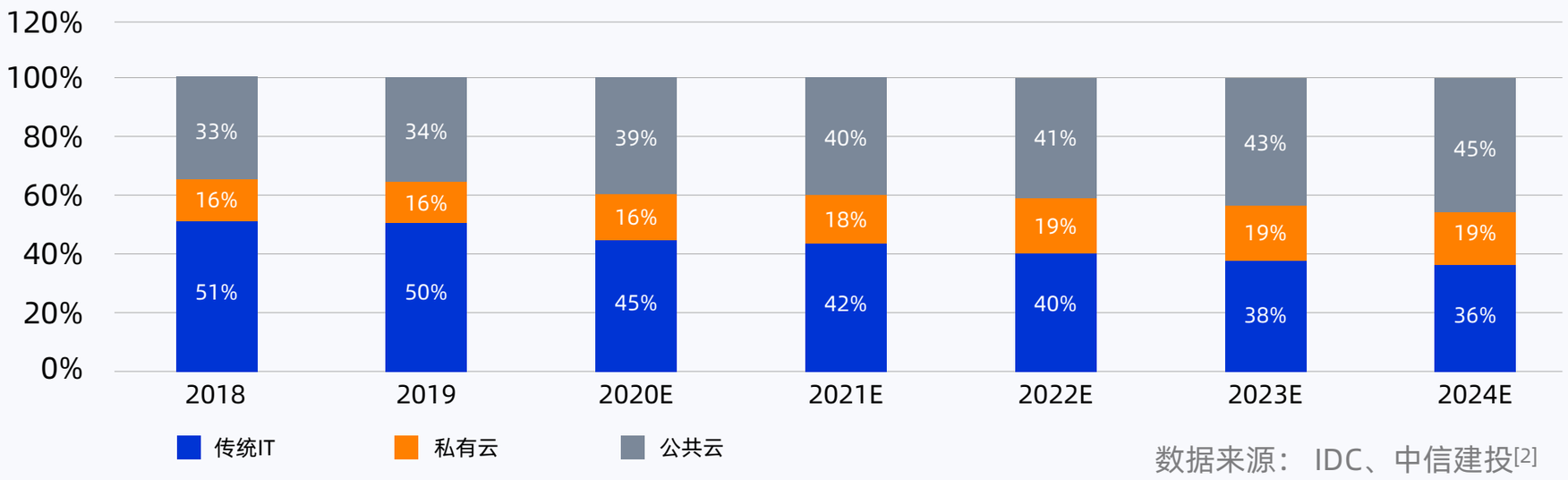
[1]《财经十一人》整理。

2.4 市场：应用场景和用户类型

公共云市场规模反映了用户认知的水平。从全球范围看，公共云被广泛接纳都经历了认知转变的过程。从全球范围看，美国、欧洲、日本等国家，公共云已经成为经济社会运营的底座，公共云市场规模不断扩大。中国公共云渗透率也在缓慢提升。

公共云占全球IT投入比重不断上升。从全球IT投入的结构看，从2019年开始，云计算的投资占全球IT投资的比重占50%，标志着云计算真正成为IT基础设施的主流。2022年，美国IT支出占总GDP的4.9%，其中15.1%用于公共云服务[1]。我国IT支出占总GDP的3%，其中用于公共云服务的比例6.6%。公共云服务支出总额，美国是中国的五倍。

图 2-5 2018-2024年全球云基础设施占IT支出比例



数据来源： IDC、中信建投[2]

企业用户是公共云市场发展的主体。中美公共云市场表现出来的规模和结构差异，核心还是看企业用户规模。目前，我国企业整体上云率30%左右，低于美国85%，欧盟70%等欧美上云水平，未来市场空间潜力巨大。[3]

[1] 数据来源：Gartner。

[2] http://www.dyhjw.com/gold/20220202-50098.html

[3] 数据来源：中国信息化百人会中国信通院院长余晓辉公开演讲材料，2023.7.

2.5 全球化：基础设施和服务运营

公共云是数字经济的基础设施，企业主体在技术、产业、生态、市场等方面的能力组合起来，最终外化体现为各云计算厂商在全球版图上的服务覆盖。

美国云计算企业全球布局数据中心，运营着丰富的生态合作伙伴。亚马逊AWS在全球 32 个地理区域、102 个可用区，115 个站点，服务覆盖245个国家和区域，合作伙伴超过12万家。微软Azure凭借多年积累的商业应用场景，与超过40万家生态伙伴，为全球60个区域的客户提供服务。

表2-4 全球头部云计算厂商全球基础设施布局情况

公司	全球基础设施	全球生态合作伙伴
亚马逊AWS	全球 32 个地理区域、102 个可用区，115 个站点，服务245个国家和区域	超12万家
微软Azure	全球 60个地理区域	超40万家
阿里云	全球30个区域建设了89个数据中心，3200余个网络节点，服务覆盖全球 200 多个国家和地区	全球超1.2万家

数据来源：公开数据整理

阿里云海外市场快速增长。阿里云是唯一在全球自建大量数据中心的中国云计算公司，在30个地区建设部署了89个数据中心、超过一百万台服务器，为全球超过500万用户提供服务，其中包括超三分之一的全球500强企业，以及超过22万家中国出海企业。支撑2021年东京奥运会、2022年北京冬奥会和2024年巴黎奥运会全面上云。

对比来看，美国云计算头部企业在全球范围内的服务器规模、可用区规模、生态伙伴规模等方面全面领先，阿里云处于追赶状态。

2.6 中国公共云发展减速的原因

中国云计算起步不晚，但时至今日与全球领先者之间仍有明显差距，主要问题是，近几年国公共云的发展速度相对国外慢了许多。问题产生的原因是多方面的，既包括全社会对云计算发展战略认识不足，也体现在政策指引不够坚决，建设管理不够完善。

1 对云计算的战略作用和技术复杂度认识不足

与美欧等国相比，今天中国企业和政府对公共云的接受程度比较低，而偏爱过渡形态的私有云。这既有客观上数字化基础薄弱、行业云解决方案不足、人才和技术就绪度不够等问题，也有各界对云计算安全、技术切换成本、风险收益顾虑等认知偏差有关。云计算发展存在过渡阶段是正常的，美欧也都经过了这个阶段。但是今天，全球领先者已明确“公共云优先战略”，中国需要加快追赶，在政府部门率先垂范、强化政策指引等方面，有更强的紧迫性和工作力度。

战略上，政府部门对公共云的基础作用和技术复杂性认识不清。公共云既是数字时代整个经济社会高效运行的技术底座，也是孕育和孵化新一代数字技术、新一代工业控制技术和企业的摇篮，是全球产业竞争的制高点。今天，全球所有的IT（信息技术）、CT（通信技术）及OT（工业控制技术）领导企业都在围绕公共云重构技术、产品和商业模式，失去公共云会失去未来。欧盟举全欧洲之力自主开发的盖亚云（Gaia-X）失败，侧面说明了构建公共云技术的复杂性和艰巨性。公共云是数字时代最重要的核心技术体系之一。

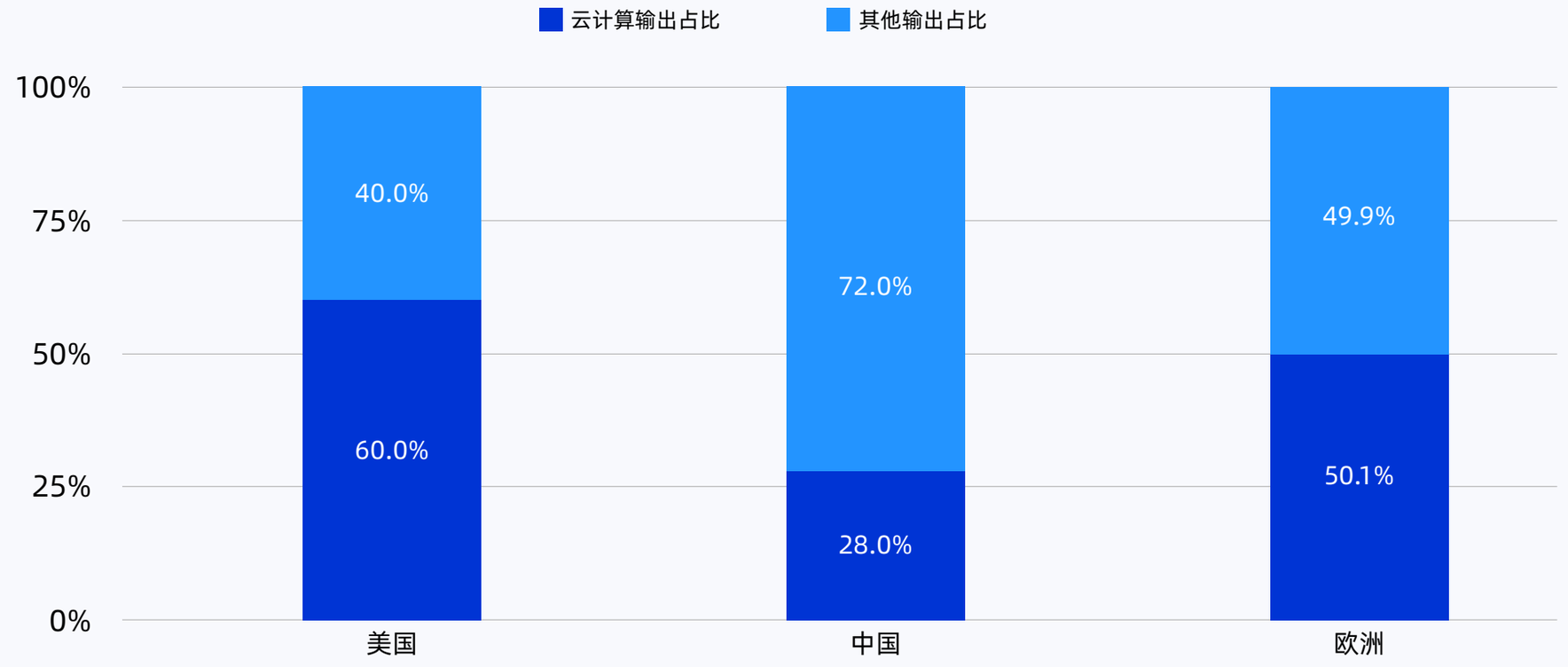
应用上，企业对云计算的安全性、经济性认识不足。中国消费者对新生事物的接受程度超过了欧美，但中国企业对云计算的接受程度，却落后于欧美。许多人认为公共云就是不安全，许多企业上云是基于云技术先进性而非经济性，会格外重视云上数据的管控能力，通常选择私有化部署，而只把少数公众服务业务或创新型业务放在公共云上。这些认识都是片面的。

2 对算力认知有偏差，公共云渗透率不高，资源使用效率低

当讲到算力中心时，政府和社会各界关注的核心关键词首先是EFlops（浮点运算次数），指向以硬件能力建设（大量购买CPU、GPU、NPU芯片）为主导的建设模式。许多项目打着“数据中心”的名目，实际在建设传统“物理机房”。只重视服务器的“资产供给”，而不注重通过数字化转型实现“算力消费”；只重视建设，不重视运营；只看重资产规模，不看重资产服务能力和利用效率。

2021年，美国和中国在全球算力规模中的份额分别为34%和33%[1]。但其中，美国以公共云形式提供服务的算力占比为65%，而中国仅为28%[2]。公共云CPU利用效率高达25%-50%。私有云部署的CPU使用效率通常仅为1%-2%，一般不超过5%[3]，导致美国和中国在全球算力规模中的实际份额上分别为9.54%和5.12%，相差近一倍。

图2-6：2022年美国、中国、欧洲算力形态对比



数据来源：中国信息通信研究院《中国算力发展指数白皮书（2022年）》

当前对模型训练的算力需求只是起点，产业应用对算力的需求还将几何式爆发。每一次模型调用，都需要强大的AI云计算支撑。但国内算力中心的普遍现状是建得多、用得少、用不好，不能对AI发展提供有效支撑。

算力中心建设的核心问题是，并没有统一的云计算架构，导致算力中心的服务器资产，没有办法被充分调度起来。尤其在面临AI大模型的计算需求时，大任务算力不够用（比如：模型预训练生成），小任务跑不满（比如：AI模型的推理应用）。

算力网络应该像民航系统一样，机场形成全国的民航网络，航线上可以跑各种不同厂商、不同型号的飞机，这样的调度效率才是最高的。关键是要有统一的技术标准和全局资源调度能力。

[1] 数据来源：中国信息通信研究院《中国算力发展指数白皮书（2022年）》。
[2] 中国信通院数据，于2022年7月中国中国信息化百人会中国信通院公开演讲材料。
[3] 经济日报文章“不可轻视数据中心高能耗问题”，2023年3.16；中国大量数据中心服务器的CPU平均利用率仅为5%至10%（不可轻视数据中心高能耗问题）。

3 政策体系的持续引领性不够

中国是最早出台支持云计算发展及企业上云政策措施的国家之一，有力推动了云计算产业的发展。当前中美两国领跑云计算，优良的政策环境发挥了重要作用。但近几年，中国在云计算发展政策指引方面的持续引领性不够，欧美国家率先高举“公共云优先”旗帜，引领了新一轮技术产业的升级。

美国联邦政府对云计算的商业和战略价值的认知深刻，从2009年奥巴马政府提出的Cloud First（公共云优先）到2018年特朗普政府提出的Cloud Smart（用好云），体现了美国联邦政府对推动本国云计算发展的坚定方向。美国政府先行先试、关闭政府自建数据中心，率先要求CIA、DoD（国防部）、NSA（国土安全部）、NASA等国家关键的部门选择与云计算公司合作解决可用性问题。客观上说，美国政府率先指引美国企业基于云计算架构实现了去IOE的进程，并鼓励将本国的公共云计算厂商作为全球基础设施辐射全球。在云计算的应用场景方面，美国在超过10年的政府推动下，已经迈过大规模企业上云的阶段，开启了实质性的云上价值挖掘阶段，形成了数字经济的巨大增量。

美国政府通过把CIA、NSA、国防部等国家级关键系统授予亚马逊、微软等商业云计算公司，既减少了政府自建系统的投资预算和持续运维成本，又极大推动了公众和企业使用云计算服务，推广云计算应用，激活社会创新动力。并进一步通过创造巨大的社会需求推动了美国云计算企业完善产品和全球的互联网基础设施布局。

中国政策指引没有对公共云特别鼓励。从财政管理视角看，各级政府主管部门更重视建设，而不重视运营；更看重资产规模，不看重资产服务能力。在数字新基建投资中，考核“硬资产投资”，难考核或不考核“投资回报”，没有建立起引导云计算建设、运营、管理、考核的科学机制。智能时代来临，各地超算、智算中心建设热情高涨，又出现了只看算力供给，不看资产运营效率的惯性。一些智算中心在没有明确“数据生产资料”的供给方，缺乏需求场景的情况下上马，有算力却用不起来，不仅造成资源浪费，也造成了算力成本高企，没有能够有效支撑智能产业的创新。

3. 灯塔：创新领先者背后的三次公共云发展浪潮

一个时代的创新者，是重新定义这个行业的生存法则、快速适应复杂环境，形成持久的创新与竞争优势的榜样，这就是灯塔的意义。云计算最初被互联网和数字原生企业所用，后来被政府部门及更多传统行业所用，今天又承载了智能时代的创新赋能使命。一座一座的灯塔，标记了公共云发展的三次浪潮。

3.1 第一次浪潮：互联网企业是先锋队，数字原生企业是生力军

云计算的第一次浪潮改变了互联网的天下。云计算让计算资源像电一样方便普惠的提供给互联网创新团队，为数字原生企业提供了快速创新，获取市场的能力，快速抓住移动互联网的发展机遇，这背后都离不开云计算平台的支撑，这是云计算发展的第一次浪潮。

在云上已孵化出一批数字原生的创新型企业，核心在于以数据为驱动的战略与业务导向，并通过数据要素为生产全流程。美国市场，基于亚马逊、微软云平台上的丰富数字技术资源，快速成长了一批技术创新公司，如Snowflake、Palantir、Salesforce、Airbnb、Uber等。例如，Snowflake成立之初就长在亚马逊AWS，打造了云原生数据库，掀起了数据库革命。

【案例3-1】Snowflake：云原生引发了数据库新革命

美国云数据库公司Snowflake是生于云、长于云的云原生数字时代的新物种。Snowflake成立于2012年，2020年9月上市成为全球最大IPO，当年销售收入不到3.5亿美元，但市值超过750亿美元。Snowflake是作为软件即服务（SaaS）提供的分析数据仓库。与传统的数据仓库产品相比，Snowflake提供了一个更快，更易于使用且更加灵活的数据仓库，引发了数据库革命。Snowflake 的成功是源于充分利用了云原生架构，支持任意云平台和数据源，较好地满足了客户互联网数据处理与分析需求。目前，其技术、产品和服务全是基于云，涵盖了从数据生产到集成、传输到备份、交易到分析、智能化应用和挖掘等数据全生命周期。

大数据分析公司Palantir敏锐的抓住政府和企业上云数据分析的新趋势，基于公共云平台，快速构建了服务超大型政企客户的解决方案，短短几年跻身百亿俱乐部，成为全球估值排名第四的初创公司。

【案例3-2】Palantir：开创了云上大数据分析的新范式

Palantir成立于2004年，定位于为政府部门、超大型商业客户解决大规模的、极复杂的业务难题提供大数据分析，市值超过400亿美元。

首先，创新云上大数据分析新范式。Palantir的技术创新在于将分散、多源、异构、高关联性、动态性用户数据，转换为统一数据本体，形成了全新的数据处理与分析方式，开创了云上大数据分析新范式。Palantir主要提供两个数据融合平台，即面向政府机构的Gotham和面向商业领域的Foundry，服务

超过150个国家、36个行业，其收入主要来自软件订阅费用。

其次，利用云平台形成规模优势。PALANTIR与亚马逊AWS和微软AZURE合作，利用云平台拓展用户，扩大产品覆盖范围，减少单个软件开发成本。比如，与向AWS的客户提供其企业资源规划ERP系统，拓展云上用户。

第三，开发新的云原生软件平台产品。2016年，Palantir在金融大数据分析平台Metropolis基础上，构建了基于公共云的、具有微服务架构的、更通用的云原生软件平台Foundry，无缝连接公司内外部数据。

中国互联网也经历了高速的发展阶段，成长起来了一系列科技平台企业，他们天然具有互联网的基因，在云上数字化、产品快速迭代与用户洞察上不同于传统企业，这使得他们可以超越传统企业几十倍速的获取用户与营收规模，成为引领企业数智化的时代风向标。例如，米哈游就是一个从初创团队开始就用云，到现在成为全球第一梯队游戏厂商的成功案例，他们从创业到壮大，全球的核心系统始终在公共云上。

【案例3-4】米哈游：生于云、长于云，逐渐成长为全球领先的数字文化科技企业

上海米哈游公司是一家100%在云上成长起来的元宇宙公司，2022年被评为中国文化企业30强。2011年创立之初，在阿里云搭建全部业务，以“轻资产”方式在公共云上实现了全球化服务能力。米哈游充分利用阿里云全球稳定的云服务器、数据库、存储、网络、安全等弹性扩缩能力，实现了一套架构、多地部署，给用户提供了稳定、高质量的产品体验。目前全球超1亿用户，其中50%是海外客户，利润已超过了阿里云。（2022年米哈游利润22亿美元，阿里云利润13亿人民币，索尼play station 18亿美元）。

在汽车行业，特斯拉是数字原生的典型代表，通过数据实现了全流程端到端管理与服务的最佳案例，从一开始就是通过软件来重新定义了新的电动车动力系统、电池管理系统、用户交互系统、客户服务系统以及整个车机的控制。革命性的重新定义了汽车行业的数字原生企业的灯塔工程。

【案例3-5】特斯拉：智能汽车的开创者，真正把汽车从“功能机”升级为“智能机”

特斯拉目前牢牢占据全球车企市值第一的交椅，其价值被放大的核心原因在于1）相比于传统车企，特斯拉更像是 TO C的科技公司；2）在软件业务（FSD）的推动下，特斯拉的商业模式越来越像苹果了。

特斯拉基于软件+硬件（高集成的电子电气架构），以整车 OTA 为桥梁，可以实现汽车的持续升级和常用常新，客户可以通过与云端升级各项功能，特斯拉也可以通过软件系统持续与客户保持数据联动，通过数据与算法优化自动驾驶、休闲娱乐、维修售后等流程，真正把汽车从“功能机”升级为“智能机”。

特斯拉真正认识到并将“数据运营商“+”客户运营商“的基因刻在了每一个生产制造、服务运营的环节中，而不是传统车企的TOB观年的线上化。相对于传统车企的 To B 属性，特斯拉更是一个”进化型组织“，从企业文化、组织结构、技术水平等方面来看，更像是 TO C 的科技互联网公司，通过小团队的快速产品迭代实现了持续的创新。

中国的智能车已成为全球不可忽视的重要新势力。并且在数字化与智能化产业发展的大潮下，预测到2025年中国将有超过70%的汽车将装备有自动驾驶功能，这背后离不开强大的智能算力支撑基础设施。自动驾驶水平每提升一个级别，车载算力需求将提升一个数量级，而训练研发所需算力则要提升两个数量级。目前，阿里云“汽车云”在国内已服务超过70%的汽车企业，小鹏、一汽、吉利、长城、地平线等均已上云。

【案例3-6】小鹏汽车：数据驱动的智能汽车新势力，用智算加速车轮上的大脑

面对海量数据，上云成为提高自动驾驶研发效率的重要手段。目前，小鹏汽车与阿里云合建了中国汽车行业最大的自动驾驶智算中心，将自动驾驶模型训练提速近170倍。基于飞天智算平台，资源虚拟化利用率提高3倍，存储吞吐比业界20GB/s的普遍水准提升了40倍，模型训练部署、推理优化等AI工程化工具也让开源框架训练性能提升了30%以上。同时，智算中心的PUE（年均能耗电力电源使用效率）低于1.2。

此外，阿里云“自动驾驶云”可处理万级的仿真并发，将仿真提速2倍，并提供80多种复杂工况场景支持仿真训练。大大缩短了新车研发的周期。

3.2 第二次浪潮：传统企业是主力军，公共部门是宣传队

传统行业和政府部门上云是云计算发展的第二次浪潮，也就是互联网、云计算技术开始从信息服务业向更加广泛的，具有数字化转型战略眼光和能力的企业渗透。同时，伴随着消费者个性化、定制化、精细化的诉求日益明显，传统行业无法有效应对当需求端的不确定性。互联网和数字原生企业的实践启发了更多行业的数字化转型，公共云也被验证可以为更多行业提供转型赋能，越来越多的传统个行业和政府部门开始上云，通过云计算平台，实施精准的客户运营，这构成了云计算发展的第二次浪潮。

微软、亚马逊抓住欧美产业数字化和智能化革命浪潮，成功地与行业生态合作伙伴，构建了“平台+生态”的产业数字化转型的新模式。比如，亚马逊云科技已经在金融、制造、汽车、医疗、能源等十大重点领域打造了近百个行业方案。目前，亚马逊云科技在全球的合作伙伴数量已经超过12万，覆盖150个国家。

【案例3-7】大众：携手亚马逊云科技构建工业云，大幅提升供应链效率

大众集团拥有大众、奥迪和保时捷12个标志性汽车品牌，每年生产约 1100 万辆汽车，每天向其工厂输送 2 亿个零部件。大众汽车与亚马逊云科技合作，将其124 个工厂站点迁移到单一技术架构：大众汽车工业云（Volkswagen Industrial Cloud）。并且，在集团层面实施了一个数字化生产平台（DPP），该平台改变其制造和物流流程，将其在 122 家工厂的设备连接起来，并连接 1,500 多家供应商，将整体供应链生产率提升了 30%。

推动大模型更易在千行百业集成落地，阿里云基于通义发布了8个行业大模型，涉及到投资、创造、编程等多个行业。阿里云正与制造业等多个行业紧密合作，通过“云计算+AI+大模型”的完整技术体系，服务千行百业的智能化升级， 包括vivo、传音、美的等消费电子企业，吉利、一汽红旗等汽车行业，中国石油、中国石化、中国海油、国家管网等大型制造企业。截止到7月底，阿里云平台服务了1.6万家制造业数字化转型。

【案例3-8】中石化：云上构建全国一体系化零售与交易平台

中国石油化工集团有限公司是中国最大的成品油和石化产品供应商、第二大油气生产商，是世界第一大炼油公司、第二大化工公司，加油站总数位居世界第二，在2020年《财富》世界500强企业中排名第二位。全国拥有30多家化工公司、30多家成品油销售公司、3万座加油站、2.8万家便利店，2亿车主会员。

中石化在数字化发展的进程中，各类交易与业务系统激增，导致全国各省分散经营，发展不均，缺少统一的管理，导致创新业务发展受限。因此，中石油与阿里云合作在公共云上建立了包括工业品采购端-易派客、化工品销售-石化e贸、零售端-易捷加油等多个重要的链接供需生态体系的在线平台。

以易捷加油的加油卡业务系统为例，首先，基于阿里云的云原生技术构建业务中台能力，实现了石化钱包与一键加油业务，使得交易时间由分钟级提升到秒级，提升了支付体验。第二，建立了统一的移动端APP流量入口，全国进行统一会员经营，全国30多省市实现统一加油体验入口，使用体验逐步统一，客户满意度提升。第三，业务中台能力共享，逐步形成敏捷的应用创新平台，基于中台能力构建易捷便利店在线化业务，形成面向全国消费者的易捷电商，进行全面推广；基新应用的设计开发提升效率35%，业务推模式由地域性推广提升到在线全国线推广，成功率得到提升。第四，多活容灾构建同城双活和移动容灾能力，保障业务高可用到99.99%，基于多活容灾组件，在资源有效利用的条件下，实现应用的秒级切换，保障业务的4个9高可用。

中石化多年来通过云上能力建设，大幅提升了数字化转型的创新效率，成为了央国企中数字化转型成功的标杆。

在工业制造业领域，AI大模型为产业智能化带来了生产方式变革的重要工具，也成为了制造业的主战场。AI大模型为代表的新一代人工智能技术将以更低门槛、更高效率，打通制造业数据断流节点，推动数据高效畅通流动，弥合制造业数据流断点，加速了企业数据要素化的进程。

首先， AI驱动软件升级是大模型赋能制造业的主要途径。未来SaaS行业也将普遍集成AIGC能力，AI大模型将成为SaaS软件的“标配”。第二，进入控制环节是AI应用制造业的标志。目前，阿里云正在实验将千问大模型接入工业机器人，在钉钉对话框输入一句人类语言，即可远程指挥机器人工作。今年6月29日，阿里云与西门子建立战略合作，西门子Xcelerator与“通义大模型”共同探索人工智能在工业场景的应用与创新，提升西门子Xcelerator线上平台的用户体验。

在云计算发展的第二次浪潮中，公共部门成为云计算的宣传队。其中，一个标志性事件就是奥运上云，在2022北京冬奥会上，云计算第一次承载了核心系统，技术让更多参会选手被看见，大幅降低了转播的成本与难度，让科技照亮体育竞技比赛的初衷–突破人类自身极限，团结世界人民。

在美国，公部门的数字化基本都在公共云平台上。美国政府在2013年到2021年间，推进了一系列大型公共云计算的项目，其中主要由亚马逊、微软、谷歌等商业云计算公司提供长期服务，服务部门包括中央情报局、美国联邦航空管理局、美国国防部、国家安全局等重要部门。

我国政府使用公共云与美国存在较大差距，主要采用政务专有云。目前各地政府自建的专有政务云主要集中在IaaS层，无法支撑AI时代快速发展下的能力需求。相比之下，公共云模式的云服务平台政务行业云基于跨平台、安全可控的飞天云操作系统，超大规模、一体化管理和水平可扩展的底层架构，同时具有领先人工智能、大数据技术与服务能力，拥有全球顶尖的立体安全防护体系与技术团队，以广泛的开放性和生态共创能力。

阿里云成为首个通过《面向公共云模式的政务云服务》测评的厂商，其政务行业云是面向政府客户打造的一朵政务属性加持的行业云，以公共云的服务模式对外提供服务，但同时具备更高级别的安全防护措施，完全满足政务系统的安全合规要求，平台能够承载非涉密的敏感信息和重要政务业务。

【案例3-9】河南省政府：基于公共云推进城市精细化治理

豫信电子科技集团与阿里云共建了省级政务公共云，目前稳定支撑200余个关键应用，搭建了多个数字化城市管理以及城市服务的场景。在政府治理领域，双方联合建设“互联网+统一指挥+综合监管”的省交通管理服务综合平台，建立以数据为驱动的行业监管新模式，实现交通行业治理“一网通管”；在民生服务领域，双方共同搭建省“互联网+医疗”政务服务一体化平台，通过“云、数、智”实现技术融合。

3.3 第三次浪潮：创新企业是弄潮儿，科研院所是攀登者

第三次浪潮是计算对科技创新本身带来的一次革命性的变化，就像AlphaFold的发生，是因为科研工具的变革，才使得科学发现有了新的突破，这也是云计算和GPT的关系，因为有了云计算，才使得各类人工智能等新的科技创新更加便捷、普惠、触手可得，大幅降低了创新的门槛。公共云是AI创新发展的基础，是AI融入千行百业的支撑，这是公共云发展的第三次浪潮。

以美国为例，美国强大的公共云基础设施，成就了这一轮的AI的全球领先地位，如生成式人工智能初创公司OpenAI、AI绘图工具Midjourney、AI视频生成公司Runway等等，都在短短几年时内成长为全球范围内的科技独角兽企业。据调查，超过50%的生成式AI初创公司（309 家公司中的167家）位于美国，并吸引了全球近70%的私营投资[1]。

【案例3-10】OpenAI：基于微软云成长起来的大模型企业引领者

OpenAI成立于2015年，专注于开发通用人工智能（AGI），2022年11底发布了GPT-3大模型，引爆了人工智能革命，目前周活跃用户达到1亿，拥有超过200万名开发者，超过 92% 的财富 500 强企业在使用该平台。一个创业公司，为何能在短短7年时间，引爆智能革命？有赖于微软对其在资金、云资源、场景、生态等全方位的支持。

一是利用大规模智算集群训练大模型。微软为OpenAI提供了专门定制的Azure超级计算系统，含英伟达上万块H100 GPU和超过20万核的CPU，用于支持ChatCPT大模型训练和在线提供服务。

二是利用微软产品推动商业化。在ChatGPT发布之初，微软将OpenAI的人工智能模型集成到了Azure、Office 365、Windows、Xbox等产品和服务中，以提供更智能、更便捷、更个性化的用户体验,这是OpenAI早期商业化最重要的应用场景。

三是为大模型开发和应用提供用户和开发者资源。微软还将OpenAI的人工智能模型提供给了Azure客户和开发者，让他们可以在云平台上轻松地构建各种人工智能应用程序。截止到2023年11月，已有超过 4500 家企业客户采用 Azure OpenAI 服务。

在AI大模型时代，各类创新企业迫切想快速构建自身的大模型能力。阿里巴巴集团董事会主席蔡崇信在2023年云栖息大会提出，“阿里巴巴要打造AI时代最开放的云”。 阿里云提供的“公共云+AI”为创新者提供了全球化的算力基础设施和丰富的AI服务，以“轻资产”的方式，帮助各类企业快速获取高端算力，加速成果转化。阿里云作为一家科技平台公司，已经服务了中国80%的科技企业和一半大模型公司。例如百川智能、昆仑万维、小红书等都基于公共云算力与大模型资源快速构建了自身的AI大模型，为企业创新发展快速赢得了战略机遇。以百川为例，不到6个月时间，基于阿里云公共云计算平台，先后训练了Baichuan、Baichuan2-192K两代基模，成为估值过10亿美元的独角兽科技企业。李开复创办的AI公司零一万物，仅用5个月时间，很快推出34B和6B Yi系列大模型，估值已超10亿美元，跻身AI 2.0 独角兽行列。

[1] 全球人工智能报告https://36kr.com/p/2322218377053828

【案例3-11】百川智能：专注创新，利用公共云上算力快速构建自身大模型

百川智能是2023年4月成立的一家以自研AI大模型为核心业务的创业公司，成立至今不到6个月，完全基于阿里云开发了百川大模型，在发布了百川大模型后，其新一轮公司估值已超过10亿美元。

通过阿里云的公共云服务，百川智能以“轻资产”方式迅速获得AI算力，专注于AI大模型核心技术的攻坚，迅速开发出多项指标都超越Meta公司的Llama 2开源大模型（Llama是全球最大开源模型社区Hugging face上评价和下载次数最多的开源大模型）。

通过使用阿里公共云提供的AI基础设施，百川将需要数亿人民币才能支持的基础建设投资，转变为弹性付费的“租用“云服务的算力消费方式，财务上幅降低了创业公司早期的基础设施建设成本，帮助企业能够在成立之初就聚焦核心技术能力的提升，并迅速开展研发攻坚，缩短了核心技术的研发周期。

阿里云不仅提供了AI发展所需的普惠高效的公共云算力资源，而且提供了阿里云自研的通义大模型能力，并通过API的模式开放给产业各界。各行业根据自身的业务场景、技术情况与资源，选择灵活的合作方式，快速将AI大模型能力嵌入自身的业务体系、产品与生产环节。如金山WPS与广东拓斯达机器人就是通过通义大模型，完成了自身的大模型应用场景与创新能力的落地。

【案例3-12】金山WPS：利用通义大模型快速转型成为AI内容创作公司

金山WPS通过阿里云自研的通义大模型，引入了AI绘图助手的功能，使得金山WPS从一家以国产化替代为核心的传统软件公司，转型成为一家能与用户实现内容互动创作的AI科技公司，并计划将通义万象的“文生图”API服务嵌入到10月底发布的新版WPS的全套产品中。

【案例3-13】广东拓斯达：借助通义大模型，完成机器人行业模型训练和应用场景落地

以前，在机器人现场部署过程中，需要针对不同场景由专业编程人员开发特定代码，再由机器人执行代码，完成任务，这对人员编程能力要求较高，耗时较长。

现在，拓斯达借助通义千问大模型，只需要一位略微懂代码的员工，就可以通过对话进行机器人的现场交付，在特定场景中实现更复杂和柔性的工艺。比如码垛这个场景，从A形状改为B形状，现在只需要一句话，机器人就能自动更改代码，精准完成码垛任务。未来，拓斯达还计划基于阿里云视觉和代码大模型，打造大模型+工业控制一体机，实现生产的精准控制。

拓斯达依托公共云的高性能、低成本弹性算力，完成机器人行业模型的训练和部署，率先实现了“大模型+机器人”的商业场景落地。并实现了不同场景的代码自动生成，大幅降低了技术门槛和开发成本，缩短了开发部署周期，提高了生成效率。同时，提高了产品一致性和品质，减少了人工干预带来的影响，提升了产品稳定性和可靠性

广东拓斯达机器人就是借助通义千问大模型，开发机器人行业模型训练，通过人机对话实现机器人代码自动生成，降低技术门槛，提高生成效率。

更为重要的是，公共云正在推动人工智能深度重塑科学研究的范式。AI for Science（AI4S）不是科研工具的更新，而是科研范式革命的工具。AI4S通过利用机器学习和其他AI技术可以解决科学研究中的问题，从预测天气和蛋白质结构，到模拟星系碰撞、设计优化核聚变反应堆，甚至像科学家一样进行科学发现，被称为科学发现的"第五范式"。这就推动科学研究从过去的假设驱动，到现在是数据驱动。根据Google学术数据表明，近3年使用AI的论文数量增长率超3倍，材料科学、生命科学、能源科学使用AI开展科研的比例超过34.5%。科技平台为科研机构的提供智算、AI等技术能力，是科学研究新范式的主力军。

2022年7月28日，DeepMind公布了从细菌到人类的几乎所有已知2亿多个蛋白质的可能结构。目前，通过AI和云计算来加速生物医学研究、药物研发、个性化的精准治疗已经成为趋势，越来越多的人类疑难杂症在科技创新的加速下解开谜团。

【案例3-14】谷歌Deep Mind：利用AI实现了生命科学的突破

理解蛋白质三维空间结构一直是生物学中的重大挑战，直到2020年，这一困扰生物学家50多年的问题被Deep Mind团队开发的Alpha Fold2模型彻底打破。在第十四届国际蛋白质结构预测竞赛（CASP14）上，Alpha Fold2的预测精准度几乎达到冷冻电子显微镜等实验技术的水平。2022年底， Alpha Fold2模型已经预测了2.2亿种蛋白质结构，几乎覆盖DNA数据库中所有已知生物体的蛋白质，其精确度已经到了原子水平，并将时间成本从数月缩短至几分钟，为新药研发等科技前沿发展做出了贡献。

AI4S的发展需要“算力+数据+算法”的支持，云计算企业是牵头产业创新生态链的最佳选择。阿里云服务了中科院国家天文台、FAST望远镜数据上云、DDE深时数字地球国际大科学计划、中山大学病毒发现、复旦大学科研智算平台等多个国家重大科研项目，在AI4S领域走在了全国前列。其中复旦大学CFFF是中国第一次让高效拥有了和科技巨头一样的研究计算平台，让科学研究真正的进入了计算驱动的时代 。

【案例3-15】复旦：打造国内首个异构算力集群， 让研究真正进入了计算驱动的时代

2023年6月27日，中国高校首个AI与大数据融合、智能计算与通用计算融合的千卡算力集群CFFF（Computing for the future at Fudan）正式上线。

CFFF基于高速数据传输网和阿里云全球领先的大规模异构算力融合调度技术、分级存储技术、AI与大数据一体化技术，部署在复旦校内数据中心集群，和托管在1500公里外阿里云乌兰察布数据中心集群，连成了一台真正意义上的“超级计算机”。复旦四校区的所有实验设备都能高速接入，做到异构算力统一管理，计算任务统一调度，满足不同应用场景下的科学智能研究与应用需求。

复旦大学校长、中国科学院院士金力表示：“以CFFF平台为代表的智算平台作为一种新兴的科研

超算架构，将成为科研的重要支撑力量，极大提升科研效率、降低科研成本，加速科学原理发现和技术突破，并有力推动科学大模型的落地。”目前，首个基于CFFF平台训练的科学大模型成果已正式发布，45亿参数大模型一天训完，气象预测速度从小时级缩短到了3秒内

在生命科学研究方面，在公共云大算力的加持下，AI推理与计算速度也得到了大幅提升。例如，中山大学加速了RNA病毒的研究，重新定义了人类对于病毒演化的历史和认知；深势科技也通过阿里云智算平台提升AI推理速度2-6 倍，算力成本降低为原来的30% 。

【案例3-16】中山大学：用云和大模型加速RNA病毒发现

中山大学医学院通过阿里公共云计算平台上提供的大规模并行计算服务（ODPS)以及与阿里云合作研发的RNA病毒发现AI模型，帮助RNA病毒发现技术突破了传统高性能计算架构的限制，将RNA病毒发现的效率从2-3月，提升为一周，将全球RNA病毒库的多样性进行了巨大的扩充——近30倍的新发现病毒种、约9倍的新发现病毒超群。云和大模型大幅提升了人类发现RNA新病毒的效率，重新定义了人类对病毒演化历史的认知 。

同时，开源开放的AI社区也是孵化创新企业的必备生态。短短一年时·间，魔搭汇聚了400万开发者、3000多个优质模型，模型下载量超过2亿，成为中国规模最大、最活跃的AI开发者社区，平台上也涌现了多个AI初创公司基于魔搭社区的各项能力与模型开启了自己的AI创想之路 。

【案例3-17】Hill Research：利用AI大模型加速医药研究

Hill Research是一家硅谷的生物医药研究机构，其客户遍布全球。通过使用了阿里云的魔搭模型社区上的“高质量”模型集、数据集和工具集，迅速搭建了AI模型驱动的工作流程，用于其先进的医药临床研究验证，具体体现在以下几个方面：

在开发前期，他们利用社区的AI模型（命名实体识别，文本分类，关系提取，分词等模型），快速搭建他们的AI Pipeline，来进行产品可行性验证。利用社区资源，得以在几周之内快速搭建出了整个AI框架（原本这需要数月时间）；

在产品前期，他们还利用到了魔搭社区的文本分类数据集（CBLUE医疗搜索查询词相关性）来训练AI模型。在前期，他们的临床数据还没有完全到位，而魔搭的文本分类数据集帮助他们加速了AI模型的训练;

在产品成熟后，他们依然利用了魔搭社区的分词模型来辅助搭建自己的data preprocessing pipeline，数据预处理的工作效率提升近10倍；利用了魔搭社区的StructBERT模型来搭建自己的AI模型完成一系列临床数据分析任务，如医学名词的标准化，医疗文本数据的疾病分类等。并就这几项技术申请了专利 。

4. 航迹：公共云优先是国际共识

公共云优先已经成为欧美等发达国家战略，从社会认知、产业政策、重大工程等方面已经形成了体系化的能力。美国、欧盟和日本等成为支持公共云发展的先行者，对中国等后来国家具有重要的启示和借鉴意义。

4.1 美国

美云计算产业发展较早，政府和企业也同时较早、较快地接受了云计算概念，并将“公共云”作为作为国家战略。尤其是美国政府率先将CIA、国防部、NSA等国家安全核心系统基于公共云平台建设，极大地推动了美国云计算产业快速发展，创造了巨大云计算市场空间。

图4-1 美国公共云国家战略演进

2009年，时任美国总统奥巴马推出“云优先（Cloud First）”战略，特别加大云计算投资，要求美政府部门在任何一个IT项目投资前，都必须先评估是否可利用安全和可靠的公共云技术实现。

2014年，美国防部公开的“获取和使用商业云计算服务的更新指导”中更新了对于国防部下属各部门采购云服务的流程，说明在保证满足端到端的安全要求情况下，国防部下属个部门可以从经过国防部首席信息官的批准的公共云服务商处直接获取云服务，免除了美国国防信息系统局（DISA）的批准流程，在保证政府数据和网络安全的情况下，简化了国防各部门使用公共云服务的流程，从而推广了云服务在政府部门中的应用。

2018年，特朗普政府进一步将“云优先（Cloud First）”升级为“云智能（Cloud Smart）” 战略，要求美各机构18个月内采取行动，进一步加强政府系统通过云技术实现系统重构和数字化转型。

2020年11月，美国国家科学技术委员会发布报告，建议联邦机构应启动和支持试点项目，以探索商业云在进行联邦资助的人工智能研究中的优势和挑战，并文档和广泛分享相关经验。

2021年5月，美国国防部公布美国本土以外（OCONUS）的战术边缘云战略，明确提出将通过云战略获取全球优势。

2022年7月发布“联邦政府利用云计算支持人工智能研发的经验”报告，总结空军研究实验室利用其云计算平台进行了大规模的人工智能训练实验，而NASA则举办了黑客马拉松等活动来教育员工如何使用云计算平台等经验。

2022年10月，美国陆军公布了《2022年陆军云计划》，该战略利用云技术来保持信息优势，并提供数字优势，使军队能够以最快的速度响应数字战场上的各种情况。

美国国立卫生研究院（NIH）表示将在2023年实施新的数据管理政策，促进更多的研究人员使用云计算。此外，美国在2022年9月发布了《国家竞争力面临的十年中期挑战》，其中提到通过发展云计算等高新科技，健全数字基础设施，以扩大其在经济、军事、科技等方面的竞争优势。

2023年，美国国务院发布《人工智能战略》（Enterprise Artificial Intelligence Strategy FY2024-FY2025）提出要集成人工智能技术到可持续且安全的智能化集成设施中，并构建和扩展各种人工智能应用。

4.2 欧洲

欧洲方面，欧盟委员会认为有必要在整个欧洲建立统一、标准化的云服务以环节目前欧洲地区数据中心利用率低、资源需求分散和重复建设等问题。多位欧洲国家领导人认为云计算与数据安全领域强相关，但目前欧盟还没有足够强大的云计算科技公司以支撑欧盟各国政府以及欧盟企业的上云用云需求，过度依赖于外国云服务将进一步加深欧洲对自身数据主权的担忧。数据显示，亚马逊、Azure、谷歌等美国云厂商占据了欧盟将近75%的云服务市场份额。

欧盟启动欧洲云计划Gaia-X，旨在建立一个欧洲主导的数据基础设施，成为欧盟的“母云端”，并通过创立通用云标准、参考云架构和互操作性要求等，提升自身话语权。2021年5月，欧盟通过了《欧盟云行为准则》，为云服务商如何遵守欧盟的隐私法规提供了详细指导。2021年5月，法国政府发布《国家云战略》，通过促进和支持对主权云服务的访问来帮助公共和私营部门进行数字化转型。该战略基于三大支柱：“可信云”认证、“云中心”政策和工业战略。2021年6月，意大利政府宣布了云计算的国家战略，创建存储所有公共部门应用程序和公民数据的国家级云计算系统，并将相关数据向“国家云”转移。

欧盟委员会启动欧洲数据战略，保障欧洲企业和公共机构能够基于公共云服务实现安全的数据存储和传输。欧盟计划以欧盟云规则手册和数据处理服务公共采购指南的形式编制一套规则。规则手册将为欧洲的云服务用户和提供商提供单一欧洲框架相关的约束性和非约束性规则。为了提高欧洲数据处理服务公共采购的效率和质量，该指南将提出实施一致的国家政策的建议，并辅以公共部门机构在招标过程中考虑的一套全面的数据处理服务基本标准。通过通用数据保护条例和云上数据保护准则保护云用户的数据安全，推进网络安全立法以提振企业、公共行政部门、公民上云的数据安全信心，推广上云用云新的基于云的服务必须响应数据保护、性能、弹性和能源效率方面的高标准要求，服务和基础设施必须满足工业和公共部门未来的数字化需求。

欧盟委员会发布《发挥欧洲云计算潜力》战略，提振云计算产业发展。战略旨在推动欧洲地区云计算发展进程，创造更多就业机会，带动经济进一步增长。欧盟委员会及成员国共投资450亿欧元用于云计算建设，到2020年在欧洲地区累计新增380万个就业岗位，创造9570亿欧元的产值。

4.3 日本

日本政府成立数字厅，推动政府业务上云。政府出台政策加速全国政务基础设施联通，便于中央和地方政府开展数据迁移，降低服务器的使用和运营成本，并计划于2025年以前构建所有中央机关和地方自治团体能共享行政数据的云服务，2026年3月份前实现全国各市町村的基础设施与云服务互联互通。日本经济产业省牵头成立由产业界、学术界专家组成的“云计算与日本竞争力研究会”，出台了培育创新、完善制度、营建平台等三位一体的推进政策。

2022年，日本国会参议院正式通过《经济安全保障推进法案》，其中将云服务列为需要“保障特定社会关键基础服务设施的稳定供应”，要求日本境内企业在引入对应行业的云服务时需要提前向主管部门提交计划审查。此举也是政府为保障日本境内云服务的本土化、加强运营商的科研开发、提升数据安全水平、防范信息泄露，确保政府数据安全。

5. 扬帆：拥抱公共云驶向智能时代

ChatGPT 引发的人工智能革命，正在加速人类进入通用人工智能时代，公共云已经成为人类驶向智能时代的“船票”。政府部门、科技企业、传统行业、中小企业都需要从理念到行动迎头赶上，重塑智能时代新的竞争力。

5.1 造大船：云服务商成为科技平台企业

AI与云计算的深度融合，也将成为云计算迭代的重要动力。我们相信，用户未来的需求定义、应用开发、运维管理、资源调度的整个运作方式，也将由大模型驱动的云计算来提供服务。云计算是智能社会的关键基础设施，云服务商要成为新航海时代的“大船”，成为支持社会创新的科技平台企业。全球的科技平台屈指可数，有美国的微软、亚马逊和中国的阿里云。科技平台的核心是要以公共云平台为载体，以激活数据潜能为目标，以构建创新生态为使命，通过基础设施即服务、数据即服务、模型即服务等新型服务模式，不断孕育、孵化新技术、新企业，赋能传统业务数字化升级，以数字创新生态驱动发展的新型平台。要成为全球性的科技平台企业，核心要打造自身核心技术体系，抓住AI大模型机遇，

成为合格的科技平台企业，云服务商应有三个“做到”。

一是要抓住云原生、智能原生技术发展趋势，做到核心技术持续领先，提升云计算的体系化基础能力。必须根据智能时代的新要求，不断提升计算、存储、网络、数据库等关键领域的研发工作，夯实技术能力，提供多元化服务。按照“云+AI”双轮驱动，从芯片、云、大数据、模型、生态等体系化提升核心能力，始终保持安全、稳定、可靠、高效、普惠和可持续，通过持续改进用户界面和用户体验，提供便捷、高效的云服务管理工具和支持服务，不断优化客户体验，，提高用户满意度，以扎实的技术能力和持续的技术创新兑现对客户的服务承诺。

二是要抓住AI大模型重构软件开发及应用范式的机会，主动开源开放AI能力，通过开放平台和API，公共云服务商吸引第三方开发者和合作伙伴，共同构建丰富的应用和服务生态，增强用户粘性和平台价值，做到新技术企业的持续孵化，让软件生态繁荣健康。生成式大模型可以完成向相当一部分的基础通用开发型工作，大幅度简化开发流程，提高开发的效率，缩短开发周期。这些优势意味着云服务商有责任推动模型的开源开放，在孵化新技术公司和培育面向具体应用的软件生态方面发挥关键作用。

三是要抓住AI大模型重构企业软件的机会，与生态伙伴一起，做到让普惠智能真正走进产业，为传统产业智能化升级赋能。生成式AI极大优化了软件界面和交互方式，不仅能够革命性提升用户体验，更能降低用户的使用门槛，简化用户的使用操作，让用户可以更加专注与用好软件，创造更多业务价值。因此，云服务商有责任积极推动智能模型与行业的结合，主动与行业伙伴一起从技术验证到模型落地做好全程服务，扎实推动普惠智能走进千行百业。

5.2 树桅杆：应用开发者丰富基于云的第三方应用生态

新的技术总是带来新的创业机会。在OpenAI这样的明星创业者感召下，更多创业者涌入AIGC领域，创投资金也非常活跃。随着越来约多优秀人才的涌入，AIGC行业具有非常广阔的发展前景。

要把握AIGC的浪潮，创业者很讲究借力而起，顺势而为，应有三个“谨慎”。

一是谨慎控制“模型野心”，把握好场景机会。当前AI大模型的技术演进与过去的计算机视觉（CV）技术演进存在很大差异。计算机视觉开发的算力需求较低，即使是小型公司也可以通过领先的算法在市场上取得一席之地；而在大模型领域，则需要同时考虑数据、算力和资金等多方面。对初创企业而言，“基础大模型”的开发是一个巨大的挑战。近几年来开源模型发展迅速，从经验看通常会形成一种商业模型领先、多种开源模型跟进的格局。创业公司如果用好这些模型，可以降低训练模型的巨大成本。搞“基础大模型”并不是必选项，创业者在基础模型之上进行开拓和创新倒是好机会。

二是谨慎管控过高期望，基于对AIGC技术的客观认知，深挖垂直领域的需求痛点。以ChatGPT为代表的AIGC表现出卓越的应用效果和发展潜力，但当前的人工智能领域对于AIGC的期望较高，可能已经超越了该技术所能达到的水平。创业者在大模型基础上寻找应用价值，更要建立起相对客观的认知，多加考虑现有技术路线所能实现的能力。生成式AI在多模态领域具备优势，如图像生成、视频生成等方面，将AIGC作为一种工具来为业务场景服务不失为好的选择。以AI小程序妙鸭相机为例，妙鸭相机从C端业务场景切入，所使用的AI技术并不是最先进的，但能够切中用户对照片的需求。据公开信息显示，自上线以来，妙鸭相机吸引超过6万用户访问。类似的垂直领域中，解决需求痛点都有很好的突破机会。

三是谨慎规避合规风险，抢抓出海机会。新的大航海时代，创业者们应该拥有全球视野，不仅要关注中国市场，还要关注全球市场。尽管出海的过程中会面临很多难题，比如合规安全、收款管理、人员管理等，但越来越多有实力提供全球服务覆盖的云平台，也在提供越来越好的“护航”服务。越来越多的中国企业出海是大势所趋。

5.3 换动能：传统行业成为客户运营商

智能化是现代化产业体系的应有之义。传统企业多年来持续推进信息化和数字化，技术水平不断提升，差异化竞争优势持续构建，经营效率不断提高。今天，智能时代加速到来，企业更要转型与原生并行，加速成为客户运营商。

成为客户运营商，是企业实现智能化的目的。企业需要努力实现智能化的根本原因在于，其内外部环境正在变得越来越复杂，越来越难以预测，越来越难以管控。这种不确定性，既体现在产品的复杂度越来越高，也体现在生产系统和供应链的复杂度越来越高，还体现在市场需求的变快越来越快，越来越多元。智能化旨在利用数据的自动流动，实现对需求的实时感知、实时满足，以机器智能增强人的智能，帮助企业应对日益增加的复杂性和不确定性。完成了智能化升级的企业，其外部形态会发生显著变化。最突出的一点，就是他们从原来以产品管理、以供应链管理为中心，全面转向以运营客户为中心。所谓客户运营商，就是将客户资产作为核心资产，能够实时洞察、筛选客户需求，实时满足客户需求，与客户持续交互、持续服务，在

为客户提供极致体验的每一个环节都创造价值的企业。客户运营商的本质，是他们认为市场的不确定性是常态，是难以预测和管控的，只能高频保持与客户的互动，围绕客户需求敏捷调动企业的全部资源，尽力跟上需求的变化，在新的服务水平上创造客户黏性和品牌忠诚度 。

成为客户运营商的前提，是成为数据运营商。数据运营商是指成功让数据成为了生产要素的企业，是企业数字化转型的结果。实现了全流程、全要素数字化的企业，其数据的采集和处理已经能够达到“实时、精准、端到端”的要求，并且拥有了符合业务逻辑的算法和模型。通过模型对数据的计算，企业可以不断挖掘数据的价值，得出可以辅助商业决策优化的建议。

打造进化型组织，保持敏捷创新。当企业真正基于数据驱动重构流程，并将经营理念切换为客户运营商时，企业的组织必然也将从“稳态”切换为“敏态”，成为根据环境变化而持续进化的强适应型组织。从稳定组织切换为强适应型组织需要在以下六个方面完成切换：从利润驱动切换为目标驱动，从以内部为中心的组织切换为以客户为中心的生态系统，从固定层级结构切换为灵活的自组织团队网络，从孤立的官僚互动切换为协作和敏捷治理，从放之四海而皆准的人才管理切换到个性化以及参与式的人才管理，从拒绝改变的文化切换到拥抱变化并持续学习的文化。[1]

[1] 阿里云新零售、阿里研究院、德勤，《数智化转型升级的企业组织变革白皮书》，2021

5.4 立潮头：中小企业勇当科技弄潮儿

中小企业是科技创新的主体。在智能时代，中小企业比以往任何时代都有条件创新，他们拥有与大型科技企业同样的创新基础设施。公共云平台正在打造一条普惠、低门槛的“平台+低代码”的中小企业数字化转型道路。我们看到，越来越多的创新型中小企业，他们借助云科技平台逐渐成为“独角兽”企业，成为智能时代创新引领者。通过公共云，可以有效破解中小企业数字化转型中遇到的“不想转、不会转、转不起”等“三座大山”，传统行业真正可以拥抱智能技术，释放智能化红利。中小科技企业面对市场更加敏捷，更能快速响应市场需求，将是智能时代创新的弄潮儿，具体有三个路径：

上云用数。 阿里云创始人王坚院士表示，“中小企业上云以后，才会觉得数据是非常重要的一个资产，才会有数据要素市场。”因此，上云是中小企业发挥数据要素价值的第一步，要借助科技平台提供的“平台+低代码+生态”体系化资源，快速上云，最短路径实现与客户连接，快速响应市场需求，迭代产品。

拥抱智能。在智能时代，随着数字技术的融合创新，云计算逐渐与大数据、人工智能、区块链等数字技术融合，IaaS、DaaS、MaaS等新的服务方式加速普及。尤其是MaaS和开源开放的社区，为中小企业拥抱智能时代提供了厚实的舞台，能低成本、高效地与使用最先进的大模型和开发自己的应用。 比如，Hugging Face和魔搭是目前全球最大的两家开源模型社区，聚集了全球最好的开源大模型、数据集和开发者和普惠的算力基础设施。中小企业借助开源社区，是拥抱智能时代的最短路径。

专注创新。在数字技术普惠的时代，中小企业要更专注于领域创新，利用公共云平台，快速开发、迭代新产品，响应市场需求。比如，阿里云目前服务了中国的65%“小巨人”企业，支持小巨人企业快速创新。

5.5 指航向：政府部门实施公共云优先战略

在战略层面，希望政府能将“公共云优先战略”作为行业规划的重要内容和数字政府建设的重要理念施。支持科技平台核心技术研发，顺应时代需求，升级智能时代的新基础设施，参与全球竞争，打造世界一流企业。建议我国积极开展公共云安全性、可控性、自主性、先进性验证，向成熟、自主的公共云计算厂商进一步扩大开放政务服务场景，进一步放开对公共云的采购准入，通过公平竞争，以市场手段推动云计算厂商技术能力持续升级，培育我国公共云产业核心能力。

在灯塔方面，可总结浙江政务“一朵云”经验，形成全省统建，通过购买服务的方式，为政府各部门提供弹性的云计算服务，打造专属的政务“公共云”。此外，对于面向公众的政务服务，需要借助公共云，提升资源的稳定性和服务能力。

在监管方面，坚持包容审慎的监管原则，平衡“发展和监管”，对AI应用风险较低的场景，以备案、行业自律等方式进行管理，为我国新一轮AI产业创新发展和国际竞争创造更好的营商环境。

在企业出海方面，鼓励中国出海企业使用国内云计算厂商在海外提供的云计算服务，确保经济安全和发展利益；支持建立离岸数据中心并提供独立的公共云服务，在跨境数据监管、财税等方面给予政策支持，帮助中国云技术更好开拓海外市场。政府或协会组织搭台、企业参与，以培训等方式切入云计算暂未发展的国家地区。

6. 彼岸：认知、战略

本报告记录了云计算从创立之初，一直到今天拥抱智能时代，整个发展历程中最鲜活的最佳实践，并从历史、技术、产业、政策等多个角度，对如何拥抱公共云驶向智能时代做了分析。在此，我们试图进一步拉高视角，从认知和战略的层面做进一步的思考 。

6.1 认知：把握第二次大航海时代的历史机遇

TeleGeography网站每年都会绘制一副全球互联网地图（Global Internet Map）[1]，展现通信如何让世界紧密相连。在2018年版的全球互联网地图中，展示了光缆把世界连接在一起，主要的网络节点是各大光缆干线枢纽。

而在2022年版的地图上，呈现方式已经悄然变成了绘制数据在全球数据中心之间的传输。从呈现光缆连接到呈现数据连接，反映了人们对互联网理解越来越清晰：光缆确实为世界的连接提供了载体，但真正让世界关联的，是光缆中传输的数据。

但数据必须通过计算才具有意义。在由数据连接的这张全球网络中，主要的枢纽节点是公共云的数据中心。公共云是互联网的基础设施，全球最主要的三朵公共云（AWS、Azure、AlibabaCloud，业界称为3A），他们的服务覆盖是世界地理在数字世界中的映射。

15世纪，哥伦布驾驶轮船开启了人类的第一次大航海时代，无数的新发现随之而来。21世纪，公共云就是人类在数字疆域中开启第二次大航海时代的轮船，我们期待产业中的创新先锋，驾驶这艘大船为人类带来更多创新。、

[1] https://www2.telegeography.com/

6.2 战略：打造世界级的科技创新平台

公共云是创新的基础设施。进入21世纪以来，全球数字经济发展格局演变的一个重要启示是，平台型创新体系已成为全球数字经济争先的制高点。平台型创新体系是基于公共云基础设施，面向海量创新需求进行精准感知和洞察，通过对全球创新资源的广泛连接、高效匹配和动态优化，构建起多主体协作、多资源汇集、多机制联动的创新生态，进而形成新技术、新产品、新业态快速孵化、规模扩散、持续迭代的新体系。互联网创新的背后是公共云平台为基础设施，电子商务创新的背后是公共云平台为基础设施，接下来通用人工智能崛起的背后仍是公共云平台为基础设施。因为从本质上，无论是互联网、电商还是人工智能的创新，它们都是源自对数据价值的无限挖掘。公共云专为处理大规模、异构、实时数据计算而生，是数字经济的基础设施。

平台型创新体系崛起的背后，是传统的产业创新体系正在解构与重组。数字化让数据成为了关键要素，将全球科技创新带入到平台竞争、生态竞争、体系竞争、多维竞争、高频竞争的新阶段。数据通过计算而释放丰富的信息和知识，从市场需求发现，到供需实施匹配；从高效验证研发创意，提高研发效率，到数据发现问题，重塑科研范式，核心是数据要素通过公共云基础设施，在供需之间让创新资源高效匹配，为科技创新与产业化之间的“断裂带”找到了新路。

云计算从头到尾只有一种–公共云。亚马逊创立AWS、阿里巴巴创立阿里云、微软发布Azure时，它们共同的名字都是“Cloud Computing，云计算”，而没有特别叫做“Public Cloud Computing，公共云”。云计算特指公共云，本不需要再加定语。面向特定需求场景，公共云可以划分高安全隔离和权限保障的专用区，但不需要重复堆砌建设独立的服务器集群。面向智能时代的机遇，在需求的驱动下，并池调度的资源规模越来越大，算力类型越来越多元，网络时延要求越来越高，公共云的技术体系正在加速升级。谷歌云和DeepMind、Azure和OpenAI、阿里云和通义系列大模型以及魔搭开源社区，面向新一代智能创新基础设施的竞赛早已打响。必须看清的是，智能时代的竞争是“公共云+AI”的体系赛跑，而绝非单点突破。

第二次大航海时代的号角已经吹响，数字经济发展掀开新篇章，智能时代的创新大戏已经拉开大幕。时不我待，行则降至，拥抱公共云，加速驶向智能创新的全新疆域！

阿里云智能集团：

安筱鹏　安　琳　张影强　曹珅珅　史大治　王中子
张献涛　陈　绪　杨　航
张　启　张晓茜

德勤：

孙晓臻　德勤中国咨询业务云服务全国主管合伙人
管　乐　德勤中国阿里云联盟领导合伙人
亓新国　德勤中国咨询业务云服务合伙人
张志钢　德勤中国咨询业务云服务总监