

云计算白皮书（2024年） 解读

中国信息通信研究院云计算与大数据研究所副所长 栗蔚

2024.7.23

1.1 各国持续深化云计算战略，安全和智能化成发展重点

美国：持续加强云计算行业应用，关注云安全发展

《国家网络安全战略》

- 提出要提高包括云基础设施在内的关键设施的网络安全和弹性，……，提升整个联邦政府的网络安全态势

2023

3月

《安全策略文件》

- 提出优化政府云安全性的措施，建议国会通过立法进一步规范云服务安全

2月

《采取额外措施应对重大恶意网络活动方面的国家紧急情况拟议规则的通知》

- 加强美国云计算服务商对外国客户的身份识别和报告义务

1月

2024

欧盟：将云计算作为国家科技创新和数字经济时代的重要支撑

2023年5月，《欧盟云服务网络安全认证计划草案》

- 从运维、数据、法律和控制权等角度全方位强调对云主权的重点关注

2023年11月，《欧洲2030：数字强国—提升欧洲科技领导力和复原力的20种解决方案》

- 将公司用云率设为关键绩效指标，以衡量建设有竞争力的欧洲方面取得的进展

2023年12月，欧盟加大对云计算投资

欧盟为云计算和边缘计算技术投资12亿欧元，以支持欧洲先进云计算和边缘计算技术的研究、开发和首次工业部署

东南亚和中东地区国家：发展云计算作为国家长期战略

泰国：以云计算提升自身竞争力

- 2024年2月《2024年数字路线图》：致力于通过专注云计算、人工智能、数字劳动力以及网络安全等领域来提高在全球数字经济中的竞争力

越南：积极拥抱云计算

- 2023年11月《新电信法》：对外国企业在越南直接开展数据中心和云计算服务几乎没有施加限制，表明了其积极拥抱云计算的态度

新加坡：通过云平台推广人工智能

- 2023年5月，新加坡智慧国家和数字政府办公室与谷歌云联合宣布推出**人工智能政府云集群**，通过云平台加速新加坡公共部门对人工智能技术的采用

沙特：设立“云计算经济特区”

- 2023年4月，沙特宣布在新设立的四个经济特区中，单独设立了“**云计算经济特区**”，专项鼓励云计算服务企业入驻，力求打造沙特科技创新新高地

韩国和日本：推动本土云服务扩张和智能化发展

2023年9月“第五个科学技术基本规划的方向（2023-2027年）” & “第三个云计算基本计划”：……，加强云产业竞争力，打造云产业永续发展的生态圈。

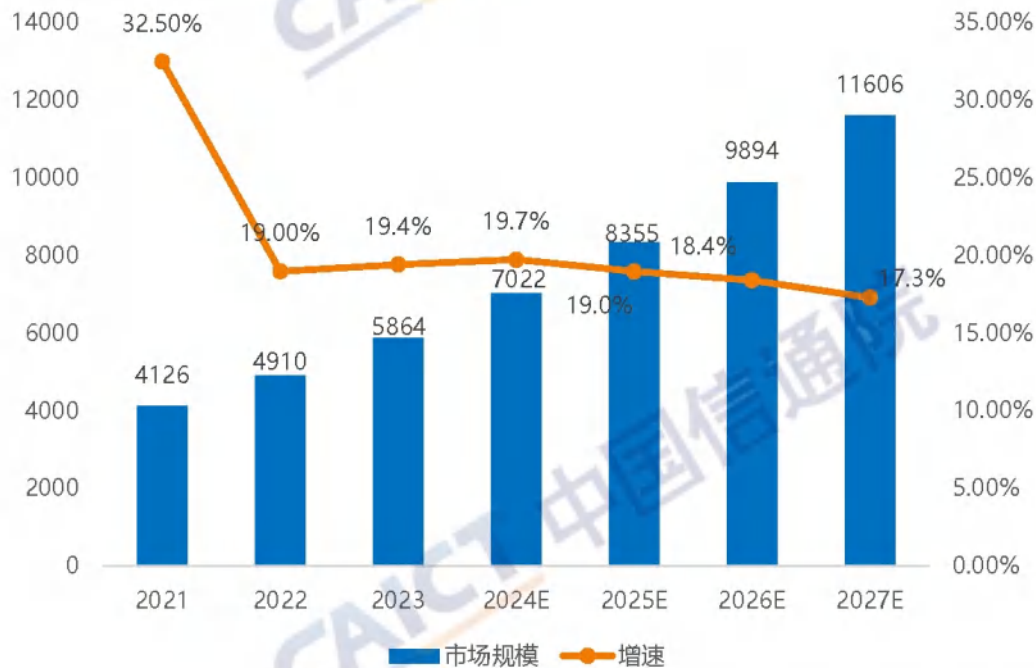
2024年2月，韩国科技信息通信部宣布，将投资9150万美元来推动本国云计算产业智能化发展。

2024年1月，日本政府接受亚马逊云在其本土投资的150亿美元，……，促进日本云计算产业的智能化发展。

1.2全球云计算市场持续稳定增长，行业巨头保持领先地位

云计算市场进入稳定增长阶段

2023年以IaaS、PaaS、SaaS为代表的全球云计算市场规模达到**5864亿美元**，**增速19.4%**。未来几年，随着云计算与生成式AI、大模型、算力的深度融合，市场将以18.6%的年复合率增长，预计到2027年全球云计算市场将突破万亿美元。



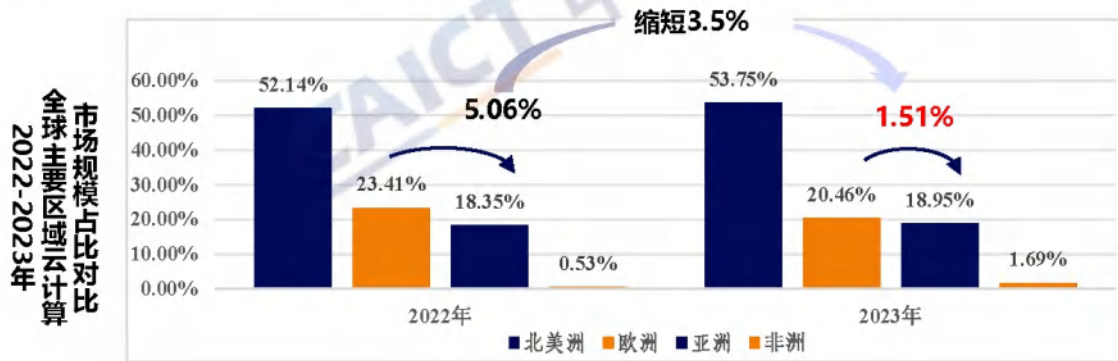
来源: Gartner, 2024年4月
2023年全球云计算市场规模及增速 (亿美元)

行业巨头持续保持领先地位

区域层面

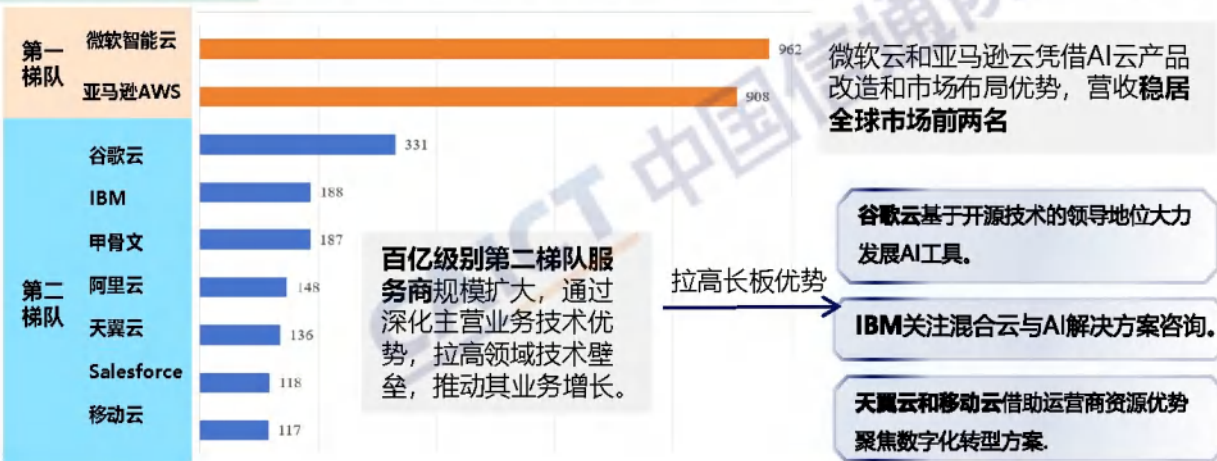
亚洲市场潜力巨大，中东地区发展迎来热潮

2023年，北美洲云计算市场规模占比53.75%，保持领先地位。**得益于部分国家的利好政策等因素，亚洲云计算市场实现快速发展，占比与欧洲的差距缩小到两个百分点以内。**大洋洲和南美洲与去年持平，整体占比变化不大，非洲则实现了3倍增长至1.69%。



服务供应商层面

云计算巨头持续领跑全球，第二梯队深耕细分技术领域保持追赶



微软云和亚马逊云凭借AI云产品改造和市场布局优势，营收稳居全球市场前两名

百亿级别第二梯队服务商规模扩大，通过深化主营业务技术优势，拉高领域技术壁垒，推动其业务增长。

拉高长板优势

谷歌云基于开源技术的领导地位大力发展AI工具。

IBM关注混合云与AI解决方案咨询。

天翼云和移动云借助运营商资源优势聚焦数字化转型方案。

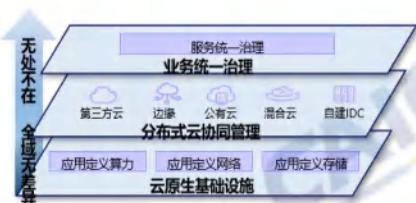
1.3 云计算技术融合价值凸显，成为新技术创新发展的土壤1

应用现代化

促进技术革新和产业需求相互融合，推动企业数字化升级转型

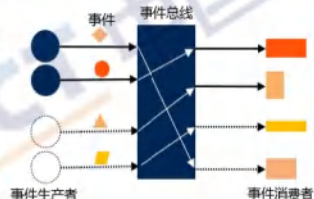
技术架构维度

分布式云原生有力支撑多云应用的一体化管理诉求。



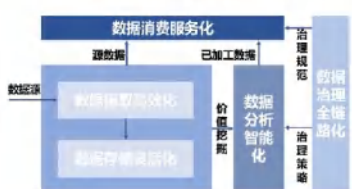
应用架构维度

事件驱动为应用提供高效灵活的架构模式。



数据架构维度

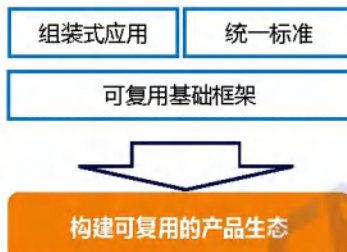
云原生融合大数据技术加速释放应用的数据价值。



行业云平台

具有共享化、场景化和市场化特征，成为大型企业建云新模式

架构方面



应用方面



运营方面



云成本治理

与云原生、大数据、SaaS等技术结合，引领企业迈向精细化运营管理阶段



云上系统稳定性

融合多种技术推动“稳保”进入主动化与透明化新阶段



云原生安全

企业云原生转型进程加快，云原生安全呈原生化、一体化、智能化发展



国家层面：政策支持力度不断加大，持续加强云计算基础设施建设和行业应用

基础设施

- 2024年，**政府工作报告**指出发挥好政府投资的带动放大效应，重点支持以云计算等为代表的新技术基础设施。
- 2023年10月，工信部等六部门联合印发《**算力基础设施高质量发展行动计划**》，提出“推动以云服务方式整合算力资源，充分发挥云计算资源弹性调度优势”，促进多方算力互联互通。
- 2024年1月，工信部在《**云计算综合标准化体系建设指南**》中，明确优先制定云计算创新技术产品、新型服务应用和重要缺失领域的关键标准。

行业应用

- 2023年3月，中共中央办公厅、国务院办公厅印发了《**关于进一步完善医疗卫生服务体系的意见**》，提出要加快推进云计算等在医疗卫生领域中的应用。
- 2023年11月，《**工业和信息化部关于健全中小企业公共服务体系的指导意见**》提出，要充分运用云计算等技术，建立贯通国家、省、市、县各级中小企业公共服务机构的全国中小企业服务“一张网”。
- 2023年12月，工信部等八部门联合发布《**关于加快传统制造业转型升级的指导意见**》，提出加快云计算等信息技术与制造全过程、全要素深度融合。

地方层面：各地积极贯彻落实国家战略部署，结合区域特点和发展需求，推动云计算技术创新和产业智能化应用

• 各省市通过政策支持和资金投入，大力推动云计算关键技术的研发与创新



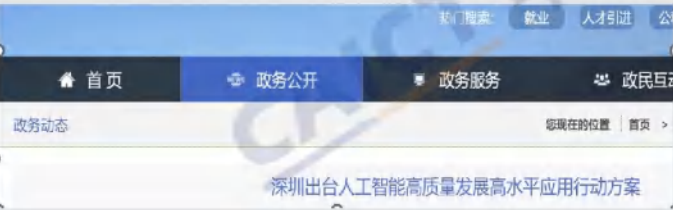
北京、山东、浙江等地聚焦云计算领域，通过相关制度激励，推动云计算重大研究项目和关键技术攻关。

• 各地政府结合本地产业特点和发展需求，积极推动产业数字化转型和智能化升级，促进云计算技术的实际应用

北京、天津等地通过实施数字化转型和智能制造工程，推动中小企业“上云用数赋智”，加快产业数字化转型。



• 各地积极探索大模型和人工智能技术在云计算中的应用



深圳、湖北等地通过建设智能算力平台，推动智能算力云服务，提升人工智能在云计算中的应用水平。

2.2我国云计算市场迎来新一轮增长，AI加速产业格局重新洗牌

市场层面

云计算市场保持较高活力，产业发展将迎来新一轮增长曲线

■ **整体市场来看**，2023年云计算市场规模总计**6,165亿**，较2021增长**35.5%**，随着AI原生带来的云计算技术革新以及大模型规模化应用落地，我国云计算产业发展将迎来新一轮增长曲线，**预计到2027年我国云计算市场规模将超过2.1万亿元。**



来源：中国信息通信研究院，2024年5月
中国云计算市场规模及增速（亿元）

■ **从细分领域来看**，AI推动市场增长点向PaaS、SaaS上移。



- **IaaS**: 电信运营商在IaaS领域的市场份额稳步上升;
- **PaaS**: PaaS市场同比增长74.9%，在公有云出海业务及AI发展的需求驱动下将持续增加;
- **SaaS**: SaaS市场增长率为23.1%，未来随着AI大模型进入商业落地阶段有望成为中小创新企业竞争的主要方式。

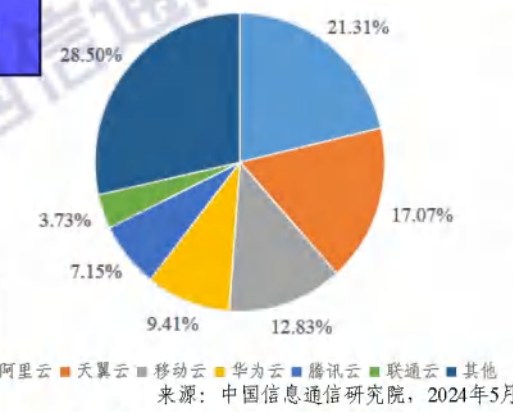
厂商格局层面

产业升级带来新机遇，市场格局或面临重新洗牌。

头部厂商：加快云智能市场规模化布局

- **头部互联网厂商保持IaaS领先优势的基础上**，聚焦软件层性能优化、PaaS层模型训练和SaaS层智能化应用。

阿里云	通过 开源策略 来构建生态，包括通义大模型家族，通过开源社区魔搭提供服务，促进了大模型的广泛应用和二次开发。
华为云	推动大模型在混合多云环境中的应用，针对不同行业提供例如矿山、药物分子 定制化的大模型解决方案 ，加速了大模型在具体行业的落地和应用。



- **运营商借助原行业渠道优势扩大IaaS层的优势**，进一步升级智算供给能力。

天翼云/移动云/联通云	天翼云提出“五位一体”的智算云能力体系在算力、模型、数据、应用等方面加强生态合作；移动云推动“五融发展”云网、云边、云数、云智、云安多维度的开放合作生态；联通云继续构建人工智能基础设施体系，推动“云网智”融合发展，全面向“AI+”战略转变。
-------------	--

中腰部厂商：

借助科技平台优势强势发力，快速推动原有云业务智能化转型从而抢占市场，厂商格局或借此机遇进行洗牌

京东云：京东推出了言犀大模型，面向实际商业场景实现在供应链优化、物流、营销、内容生成、数字人等多个场景中落地实践

火山引擎：火山引擎发布了9款豆包大模型，在字节跳动内部50多个业务中得到应用，在App Store的AIGC类应用中名列前茅。

金山云：金山云主要通过自身办公软件优势融合智能化业务，与微软相类似“产品+智能”的路线。

2.3各行业上云积极性持续攀升，梯队分层效应依旧明显

政务方面，政务云正步入“融合升级”与“智能跃进”的新时期

■ 政策环境优化升级，一体化建设提速
面向新阶段，在国家顶层指导下，各省市如河南、江苏、广东等地陆续发布相关政策，围绕构建履职能力、安全保障、制度规则、数据资源、公共支撑等内容，**推动“一朵云”建设，完善一体协同的数字政府发展体系。**



■ 新技术融合创新，核心价值释放
随着5G、大数据、人工智能等技术的深入应用，政务云的基础设施将更加完善，支撑起更复杂、更智能的政务应用。推动**政务服务向一体化、智能化方向**发展。



■ 场景应用多元化，生态产业新格局
“云+AI”将赋能政务应用场景，推动云服务商与AI大模型企业协同创新，**服务模式**不断创新，助力产业生态新格局。

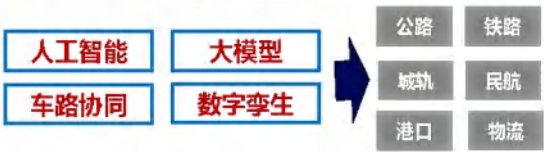


交通方面，云重心从基础设施向应用层转移，水陆空部分细分行业进入发展快车道

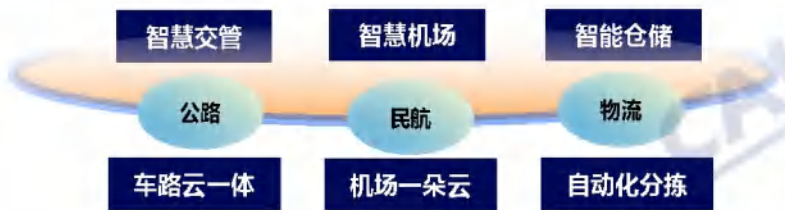
■ 资源上云向应用上云过渡，PaaS、SaaS等高阶服务渗透率逐步提高。
机场云、城轨云领域由传统虚拟机、一体机逐步迈向**容器化、微服务等云原生架构**。物流行业企业，将云原生架构应用于仓储物流等核心环节及系统构建。



■ AI、车路云一体化、大模型等新兴技术加速升级，驱动交通云底座建设再上新台阶
腾讯、百度等厂商，积极探索并推进**AI、车路协同、大模型、实时数字孪生**等新兴技术加速升级，有效驱动云计算在交通行业的深度使用。

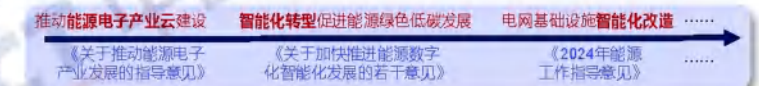


■ 水陆空云上应用多点开花，公路、民航、物流领域进入发展快车道
公路、民航、物流等细分领域加速云上深度实践，打造精品工程，不断提升服务质量及用户满意度。



能源方面，能源行业处于夯实基础阶段，聚焦云底座规划和建设能力

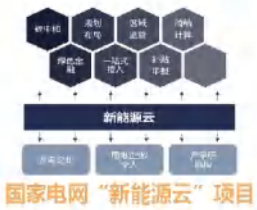
■ 能源行业转型升级政策持续加码，提升云平台建设水平成共识
✓ 多政策出台加强能源领域自主创新、新能源基础设施、电网智能化改造等方向，随着能源行业相关政策落实，能源行业用云水平将进一步提升。



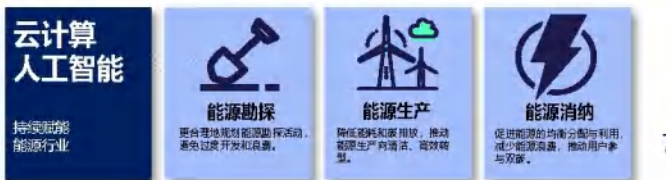
■ 能源云建设从基础设施云化改造，向综合能源服务云平台转变
✓ 围绕能源设施云化改造，能源企业**充分挖掘能源数据价值**
✓ 加强能源产业链协同，形成人工智能等云上应用，对外提供服务。



■ 能源云“龙头效应”明显，电力、油气领域应用落地进程较快
✓ 能源产业重点领域大型央企牵头上云，其他中小企业积极跟进
✓ 头部企业关注**能源云的二次开发与应用**
✓ 中尾部企业关注能源深度用云场景



■ “云计算+人工智能”模式应用于能源业务，推动“双碳”落实



- ❑ **整体来看**，各地区的供云用云水平与其数字经济规模呈明显正相关关系。以京津冀、长三角、珠三角为首的中东部地区供云用云较为成熟，供云规模与用云数领跑全国。西部地区的数字经济建设起步较晚，目前仍处于发展储能阶段。
- ❑ **供需对比来看**，我国大部分省份供需量保持一致，但部分省份存在云资源供需交叉不平衡的情况，例如贵州、宁夏、内蒙古等省份的供云量远高于其用云量，川渝地区的用云量则高于其自身供云量，亟需优化资源配置，促进各地数字经济均衡发展。

供云量：中东部地区供云量领跑全国，西部地区发展潜力巨大

- 当前，我国31个省（自治区、直辖市）云计算对用户可供给**vCPU核数超过10亿个，云存储规模超50EB**。
- 各大云服务商在中东部地区的云计算供给量占整体公有云资源的**60%左右**
- 西部地区整体处于**发展和蓄能阶段**。

2023年31个省（自治区、直辖市）供云量指数



供云量指数

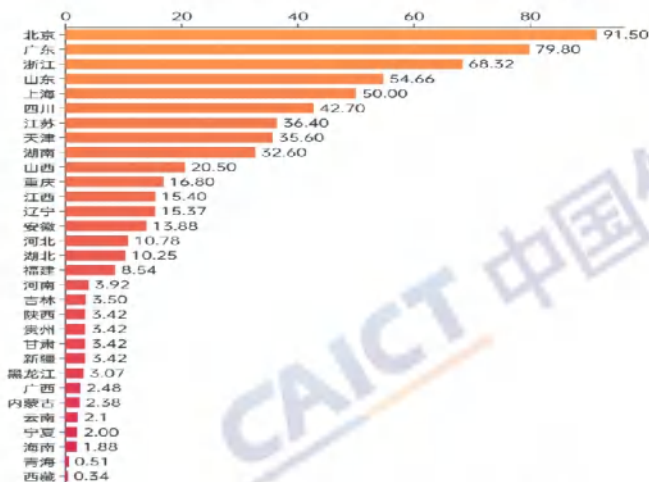
云服务商在各地区能够提供的云计算供给规模，反映各地区云计算供给水平。

数据来源：中国信通院，2024年5月

用云量：东南地区高于中西部地区，第一梯队集中在京津冀、长三角、珠三角

- **各省用云水平深度与数字经济发展成正相关**。随着省市用云量提高，云计算所拉动的产业数字化增长随之升高。
- **从地域分布看**，全国大部分省份已完成初步上云，用云量第一梯队主要集中在京津冀、长三角、珠三角等地，东南地区用云量明显高于中西部地区

2023年31个省（自治区、直辖市）用云量指数



用云量指数

指各地区使用云计算相关产品或服务的情况，反映各地区上云用云的程度。

数据来源：中国信通院，2024年5月

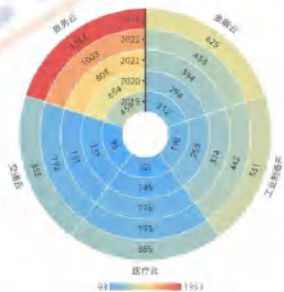
3.2 云服务应用从资源上云转向深度用云，服务质量还有提升空间

评价体系显示，中国云服务应用水平展现出了从行业龙头到中小企业全面铺开、从“资源上云”到“深度用云”，广泛渗透的发展格局：

- 从行业看，泛互联网行业云化程度领先，上云理念不断向传统行业渗透；
- 从规模看，大型企业未来几年将实现全面上云，中小企业有机会成为上云“生力军”；
- 从质量看，云的高可用性已成为服务商发展必备条件，开始着力于提升服务质量。

泛互联网行业用云量大幅领先，政务上云热度高

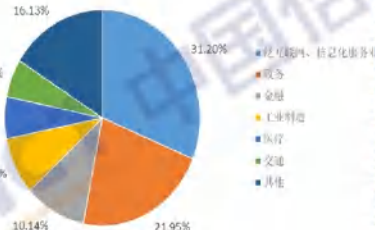
- 第一梯队：**政务用云热度保持领先；
- 第二梯队：**金融、制造行业在行业政策推动下，积极性将大幅提高；
- 第三梯队：**医疗、交通行业随医疗云、汽车云等相关概念兴起，未来发展空间巨大。



2019-2023年典型行业用云热力图

数据来源：中国信通院

- 数字行业：**泛互联网、电商行业信息化程度领先，占总用云量30%以上；
- 传统行业：**政务、金融、制造等传统行业上云用云规模将随“上云用数赋智”理念深入不断扩大。

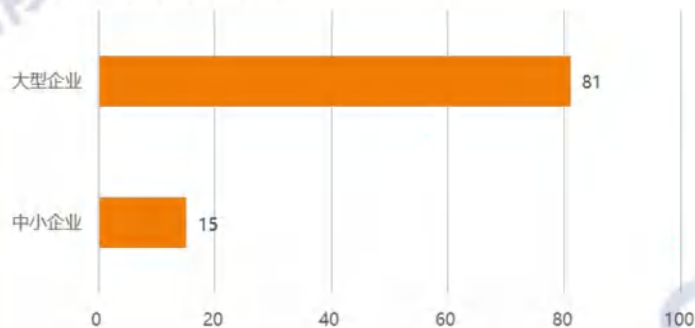


2023年各行业用云量占比

数据来源：中国信通院

大型企业发展深度用云，中小企业快速上云

- 大型企业：**企业结构复杂，需加速推动核心业务系统上云实现降本增效；同时自身盈利能力强，企业具备足够资本推动上云用云。目前大型企业正式向“深度用云”阶段演进，各系统上云率超过80%；
- 中小企业：**具备资金流转快、技术人才需求高、用云速度快等特点，SaaS模式上云契合中小企业特点。随我国SaaS行业持续发展，将持续改善中小企业上云情况。



2023年不同规模企业上云率

数据来源：中国信通院

东部IaaS服务质量稳定，SaaS服务质量波动明显

- 东部服务质量稳定：**整体运行情况较为稳定，2023年平均服务可用性达98.43%；
- 中西部服务稳定性存优化空间：**在可用性、访问时延等方面仍有优化空间，服务质量相比于京津冀、长三角、粤港澳三大经济区低20%左右；
- 企业级SaaS质量波动较大：**年平均可用性高于99.5%，但服务平均响应时间波动较大。其中，企业协作、财务管理、点播直播等企业级SaaS服务的整体运行情况表现较好。



2023年31个省（自治区、直辖市）云服务体验指数

数据来源：中国信通院

3.3算云融合呈阶梯状分布，数智云化水平仍有较大提升空间

云技术融合发展水平反映算力、云计算等技术融合情况，通过算云融合水平与数智云化水平进行展现，具体从算力资源云化、数据服务云化、智算服务云化三个视角展开分析。

算云融合水平 算云融合水平指通用算力基础设施融合虚拟化、云原生等云计算技术实现算力及其关联资源的云服务化输出，反映各区域的算力云化程度。

第一梯队：以北京、广东、上海、浙江为代表，对算力规划布局早，云化比例均超过65%，算力建设经验丰富且已经形成了规模化的算力基础，同时大力开展云计算、算力服务相关技术与落地，形成了体系化的算力云化输出，算云融合指数高。

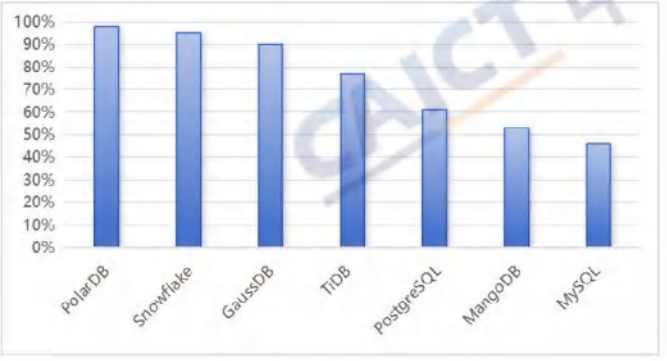
第二梯队：以宁夏、甘肃为代表，算力规划和建设虽然较北上广等起步晚，但及时拥抱算力等新技术，在建设过程中即与云计算深度融合，算云融合程度较高。

第三梯队：以贵州、内蒙古为代表，这些省份基于地理、环境等优势完成了较大规模的算力资源建设，但整体云化比例偏低，在算云融合方面仍需要持续投入，以实现算力资源的高效利用与价值转化。



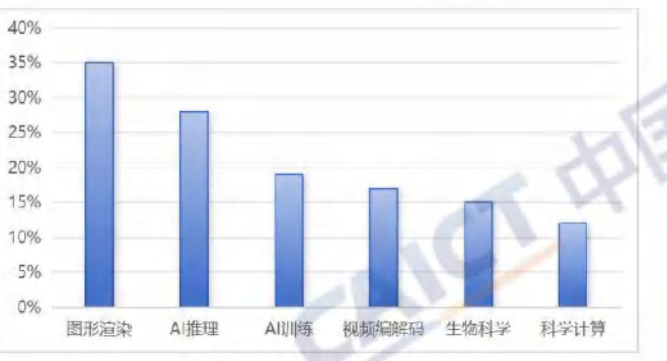
2023年31个省（自治区、直辖市）算云融合指数
数据来源：中国信通院

数智云化水平 数智云化水平指数据服务与智能计算服务融合虚拟化、云原生等云计算技术实现在不同场景的应用和服务部署的云化程度。



数据服务云化部署比例
数据来源：中国信通院

在数据服务云化部署方面，云原生数据库由于其高弹性、高可用等优势逐渐成为用户数据库服务的主流选择，云化部署比例极高，如PolarDB、Snowflake、GaussDB、TiDB等，均超过70%。



智能计算服务云化部署比例
数据来源：中国信通院

在智能计算服务云化部署方面，图形渲染、AI推理等场景云化部署比例较高，均超过25%，生物科学、科学计算等场景因受其计算范式所限，多使用本地部署的智能算力，云化部署比例在10%-20%之间。

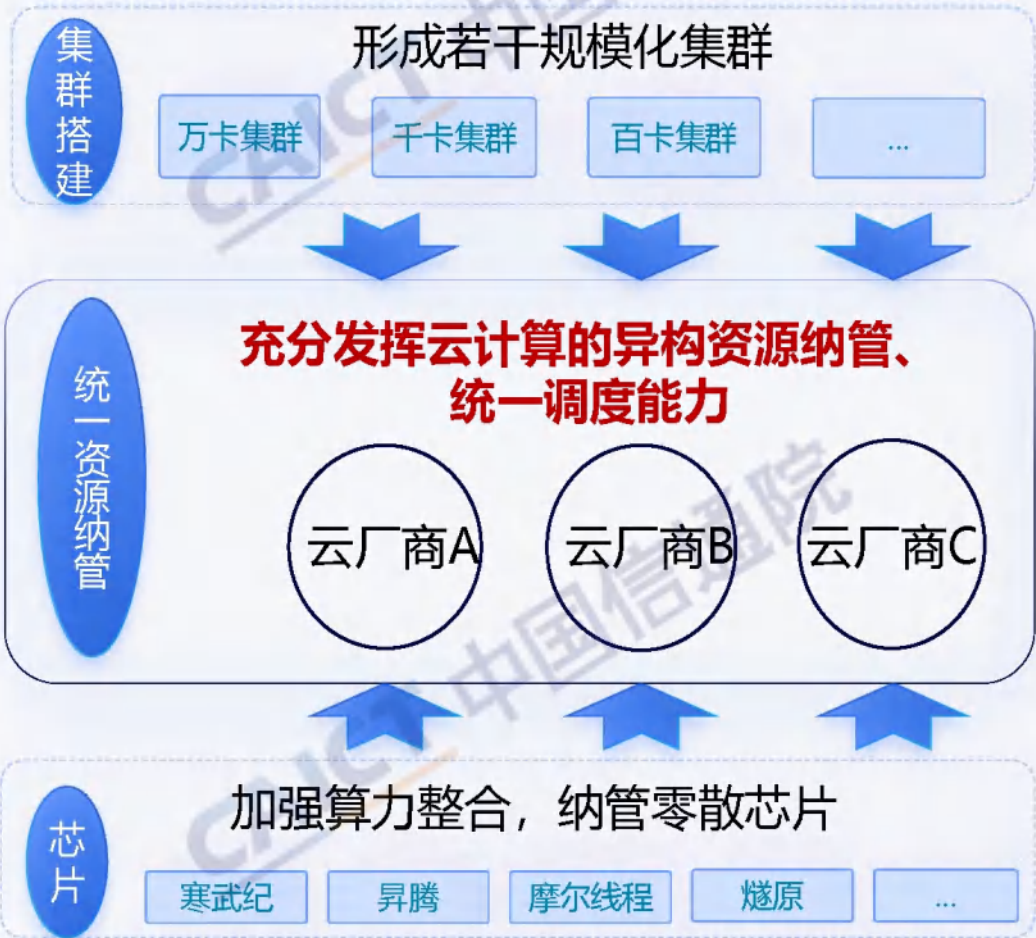
智能云体系包括智能云基础设施服务 **AllaaS**、智能云平台服务 **AI PaaS**、模型云服务 **MaaS**、智能云应用服务 **AI SaaS**。智能云作为新一代人工智能发展的驱动力量，正在 **重构云服务产业格局**，加速形成新质生产力。



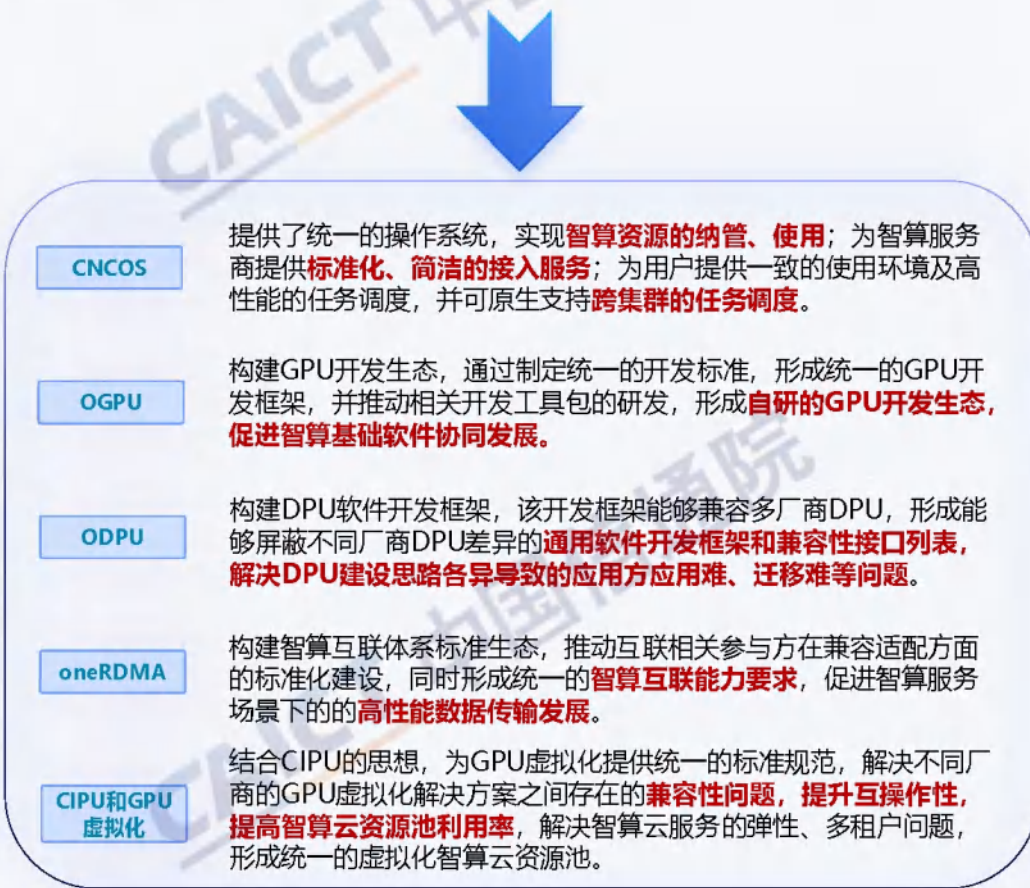
智能云体系架构图

- AI SaaS推动AI应用落地实践，加速企业数字化转型进程
- MaaS打造AI模型管理与应用开发工具平台，敏捷赋能AI应用创新。
- AI PaaS聚焦智算服务开发能力，成为AI工程化实践最优解
- AllaaS夯实异构智算资源底座，为AI产业发展提供灵活高效的云上算力支持

AllaaS以智算资源为核心，兼容纳管异构智算资源提供**智能化IaaS服务**，并使用分布式编排调度工具充分发挥智能云资源池能力，**统筹异构智算资源，缓解芯片短缺**。



通过**智算基础软件创新**，充分发挥智算集群的工程化能力，实现**千卡万卡芯片资源高效协同工作**，最大程度发挥芯片的性能，**突破算力瓶颈**，显著提升了AI训练与推理的效率。



4.3 AI PaaS 聚焦智算服务开发能力，成为AI工程化实践最优解

智算云服务能力的关键在于对智能算力进行高质量管理和应用。AI PaaS提供丰富的AI工具组件和模型及算法仓库以及稳定的运行环境让用户模块化、定制化进行人工智能开发，成为大模型时代应用生成的加速器。实现**资源高效调度、设施绿色低碳、算力灵活供给、服务智能按需**。



增强传统云服务**平台架构设计、平台能力、服务构建能力**
加快通用云服务向智能云服务转型

以云计算底座建设高质量**智算云服务平台**
全流程打通**数据服务与AI应用开发**

4.4MaaS打造AI模型管理与应用开发工具平台，敏捷赋能AI应用创新 CAICT 中国信通院

MaaS面向基于模型做AI应用开发的用户，将复杂的底层能力进行了**标准化封装**，为用户提供了一个**便捷化的模型服务平台**，使得用户只需专注于垂域模型训练和AI应用开发本身。



MaaS提供**通用AI模型管理**和**垂域AI模型训练**服务，降低AI模型应用门槛，推动AI模型普惠化。



MaaS提供基于模型的**全流程AI应用开发工具链**，实现AI应用敏捷开发，加速AI创新与价值落地



国内外厂商推出全流程应用开发的MaaS平台，帮助下游用户更加高效、便捷地实现大模型的应用开发



MaaS提供丰富的**基础AI云服务**，赋予了企业**定制开发AI应用**的能力，支撑各行业生产力变革



AI结合SaaS轻资产、快速应用的优势，经过封装形成开箱即用的AISaaS，极大拓展AI技术的应用边界：

- **理解业务细节**：结合行业SaaS落地行业应用经验，深入按行业用户需求，解决实际问题；
- **简化用户应用**：SaaS即开即用、按量付费特性快速落地生产级AI应用，降低构建以及使用成本；
- **提升训练质量**：SaaS实践多年已积累大量真实数据，可用于训练完善AI能力，为数据不足用户提供完善模型。

通用行业

AISaaS+业务场景

垂直行业

市场业务

- 客户行为分析，预测客户需求，形成个性化营销方案
- 分析市场趋势，调节团队人员配比，提高销售转化率
- 重塑消费者画像，定制化营销方案，提高营销行为ROI

公司管理

- **人力资源SaaS**：简历智能筛选，职位智能匹配
人员风险预估，个性员工培训
- **财税管理SaaS**：自动账务处理，财务风险评估
智能发票识别，减少人工失误

协同办公

- 企业员工：AIGC+OA智能化办公，提升办公效率
- 管理人员：智能化数据/资源分析，辅助项目决策

农业

SaaS化物联网设备管理与AI算法分析，智能化提供作物管理建议，提升种植效率，推动农业管理现代化与精细化。

工业

生产、运维、质量监测SaaS接入大模型，实现精细化参数调整。计算机视觉替代传统人工质检，提高工业产品质量控制的准确性与效率。

医疗

医疗大模型辅助诊断诊疗，医学影像SaaS实现影像智能化分析，提高问诊效率与患者的就医体验。

零售

智能化供应链管理，预测需求变动，优化采购计划，降低物流与仓储成本，提高货物周转率。

我国算力资源相对分散

算力资源提供主体多，地域分布广，规模不一



中国部分省份算力基础设施规模分布 (2022年)

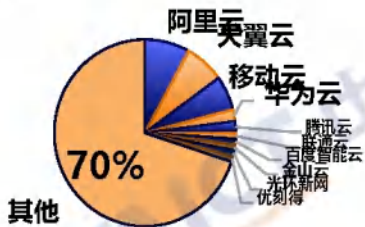
数据来源：中国信息通信研究院《中国算力发展指数白皮书（2023年）》

截止2023年6月，

- 通用算力：数据中心服务商7387个，其中云服务商5183家
- 智能算力：智算服务商114家
- 超级算力：国家级超算中心15所

找算力难，供需匹配失衡

- 我国算力集中度低，前10的服务商只运营30%的算力，也很难统筹其余5000多家分散的算力。



中国算力集中度

调算力难，应用调度低效

- 调度系统能力弱，我国大部分算力运营主体普遍不具备调度能力，服务化程度低；

我国服务商调度能力

前10家服务商	调度系统能力 较强
5000多家服务商	调度系统能力 无或弱

- 网络：数据传输带宽低、成本高

用算力难，应用迁移困难

- 我国计算芯片多达十余种，计算框架各不相同。

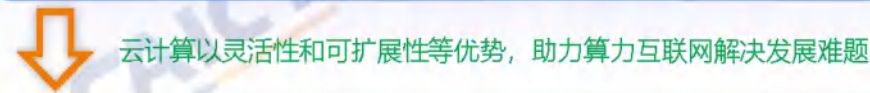
我国智算芯片和计算框架分布

	我国智算芯片和计算框架分布			
计算框架	CANN	Tapside	MXMACA	
	天数软件栈	ROCm	MUSA	BIREN SUPA
智算芯片	华为	燧原科技	天数智芯	
	海光	摩尔线程	沐曦	壁仞

算力互联网定义：算力互联网是互联网面向算力应用与调度需求进行能力增强和系统升级形成的新型基础设施和技术产业体系，其本质是在互联网体系架构上构建统一算力标识符，以算网云调度操作系统和高性能传输协议为基础增强计算和网络能力形成弹性网络，具备智能感知、实时发现、随需获取能力，形成算力标准化、服务化的大市场 and 算力相互连接、灵活调用的一张逻辑上的网，使应用和数据在算力资源间高效供需匹配、流动互通、迁移计算。

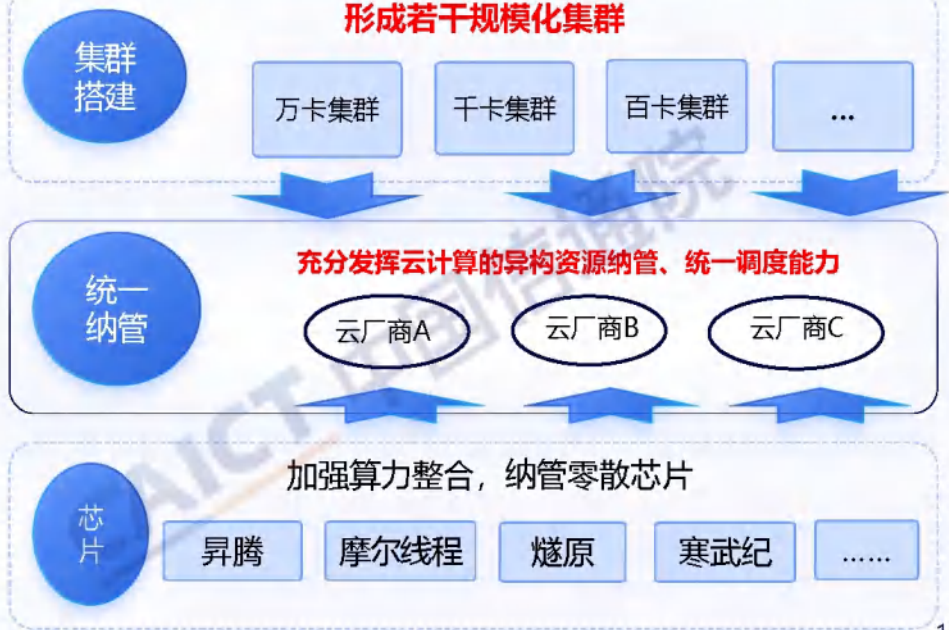
机遇：大模型计算等场景催生算力互联互通需求，算力互联网应运而生。将不同主体、不同架构、不同地域的算力资源标准化互联，实现任务和数据在算力资源间高效流动互通，成为推动算力发展的关键。

挑战：一是算力市场分散，“找算力”资源成本待降低；二是调度能力不足，“调算力”应用传输待提升；三是计算框架不同，“用算力”架构适配待优化。



云计算成算力互联网操作系统：
以算网云调度操作系统和高性能传输协议为基础

形成若干规模化集群



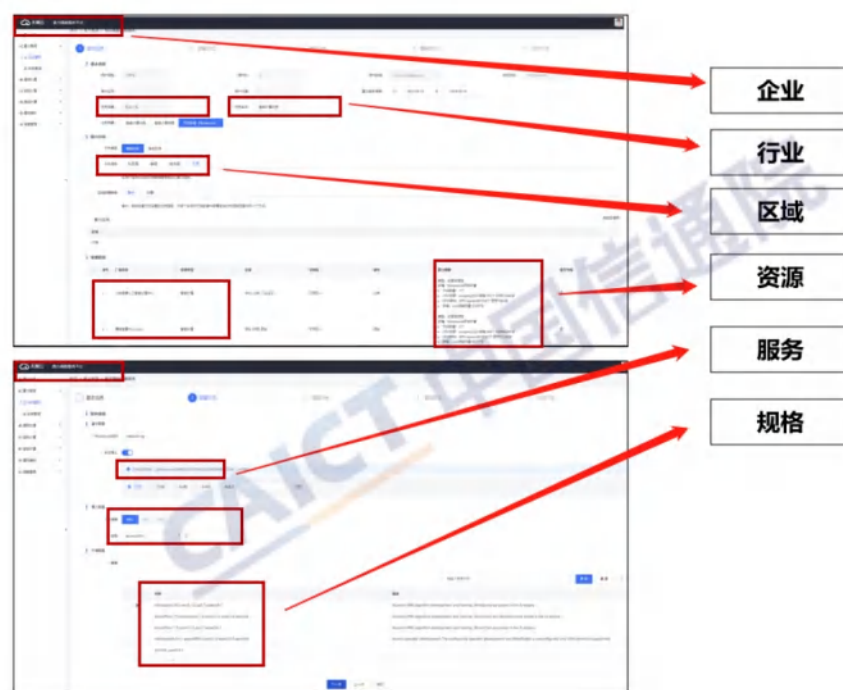
资源感知方面，云计算基于算力标识推动算力互联网资源与服务的描述、定位及寻址

- **算力互联网标识符（简称：算力标识）C-URL**：类似互联网使用资源标识符URI（URL）对访问的文件进行定位，包括了域名和文件所在的路径。C-URL包括**唯一标识和算力路径**。计算任务在应用层基于算力标识匹配合适的资源 and 路径。
- **算力标识到芯片**：人工智能等应用需编排调度智算到GPU芯片颗粒度，因此标识需细化到芯片层以支撑更灵活的算力调度策略：



资源互联方面，云计算基于算力标识推动算力互联网资源与服务间的互联互通。

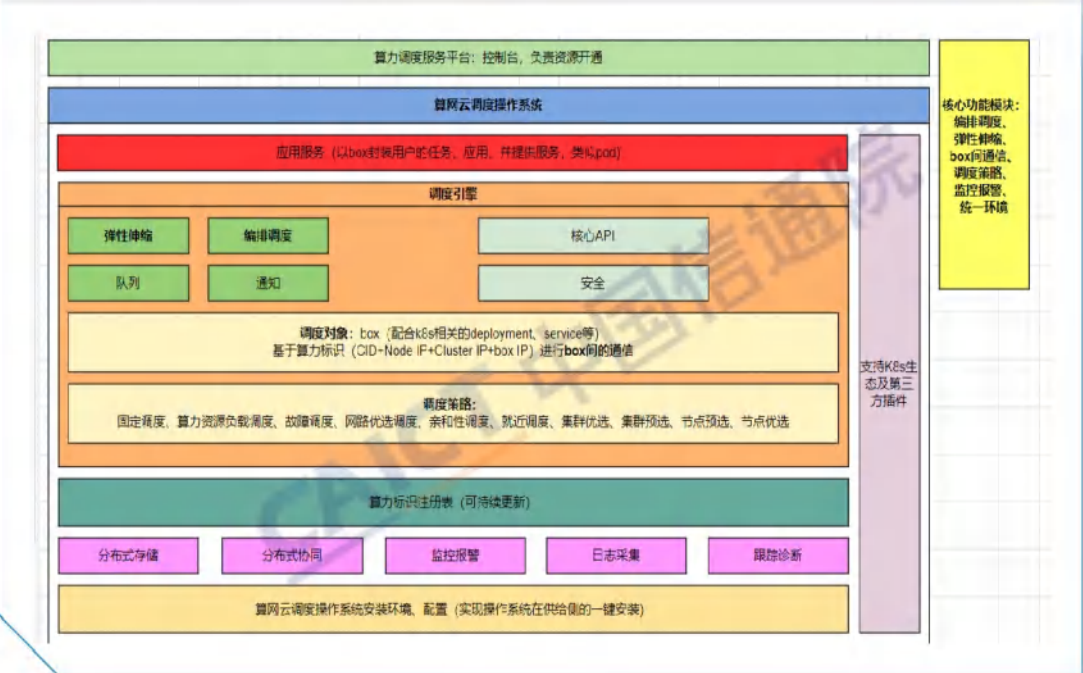
- **动态标识：**任务型应用算力需求高弹性，C-URL通过**实时感知资源状态**支撑算力资源的按需调度。
- **资源互联开通：**各算力资源提供商按照云计算的资源及服务开通流程，提供**基于算力标识描述的各类资源及统一接口**；算力调度服务商通过统一算力标识及接口规范，将算力提供商资源在自身调度平台，用户通过任一调度平台均可实现全局资源的检索及开通使用。



5.3云计算提供跨域跨架构跨主体的任务调度，实现业务高效“调算力”CAICT 中国信通院

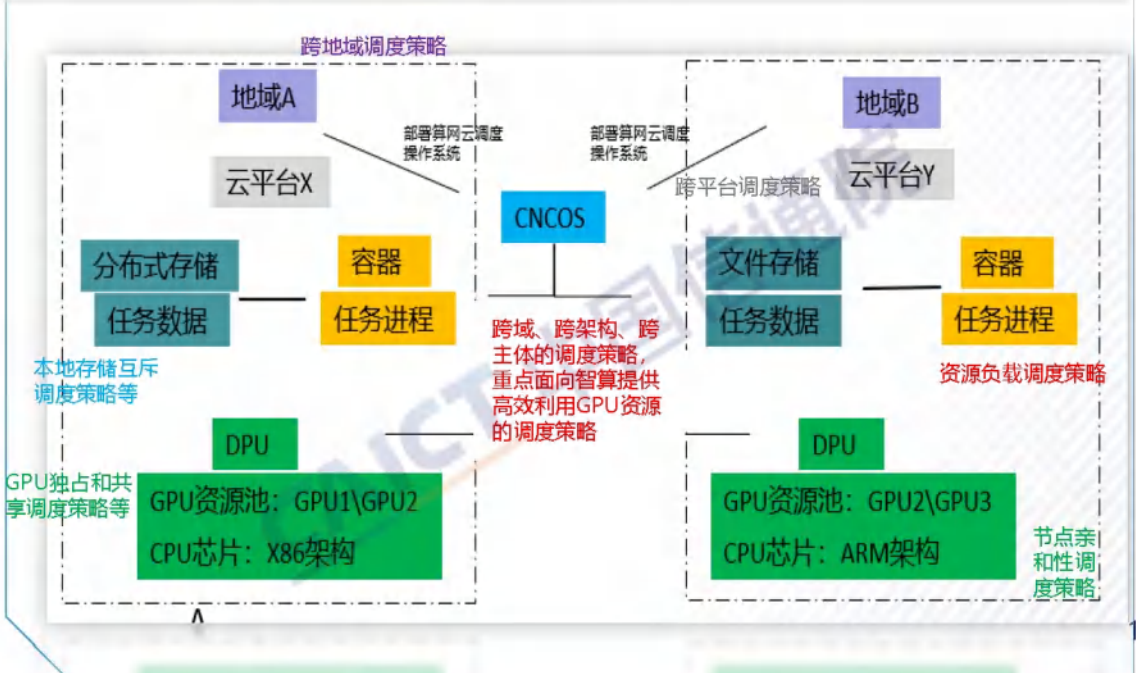
调度能力方面，云计算提供智算、通算等多类场景资源及任务调度能力，助力算力互联网协同调度。

- 调度能力：**算力互联网的应用在异构算力、异构网络和多云环境中部署运行，需要基于算力标识、算网参数进行统一编排调度。实现基于异构算力与开发架构的统一操作系统环境部署、网络间高性能传输、用户应用、任务封装及跨地域、跨架构、跨主体调度部署。



调度策略方面，云计算围绕调度场景及应用需求，提供丰富的调度管理，推动算力互联网资源高效使用。

- 调度策略：**云计算基于丰富的**调度策略匹配任务需求及资源现状**，尤其是面对智算场景调度层面，可以提供包含GPU独占和共享调度、卡间传输监测调度、节点和应用资源感知调度、资源预留调度等诸多调度策略，充分利用闲散资源支持多云场景下的算力资源统一管理

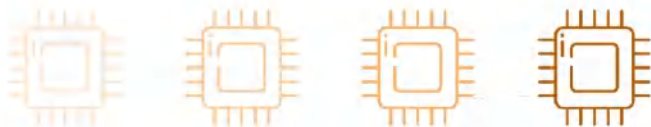


5.4 云计算推动基础设施软硬一体协同发展，实现业务高效“用算力” CAICT 中国信通院



计算能力方面，云计算通过**GPU资源虚拟化技术**提升计算利用率，整合碎片化算力资源。

GPU



vGPU1 vGPU2 vGPU3 vGPU4

弹性的GPU算力服务

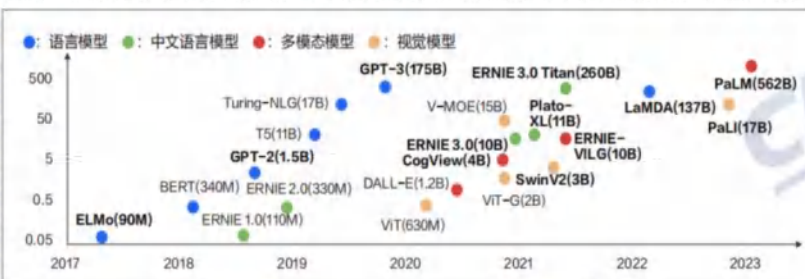
良好的用户资源隔离

算力互联网的应用场景较为广泛，执行的计算任务更为复杂多样，通过GPU虚拟化能够加速资源的高效分配，将算力市场中零散闲置的资源提供给用户使用。



网络能力方面，云计算通过部署**高性能互联网络**，大幅提升算力互联网计算效能。

传统的云计算采用以**TCP/IP**为核心的高速以太网作为高性能集群间的组网方式



随着模型规模和训练数据规模的不断增大，面对日益增长的数据量，以太网仍然存在性能瓶颈，因此，需要在云环境下构建**高性能数据传输网络**，提升云计算场景的数据传输能力，消除网络性能瓶颈。



IB网络的设计区别于已有的TCP/IP网络，具有**高带宽、低延时和无损**的特点



RoCE网络保留了IB网络的传输层，在其他层使用了传统以太网的设计理念，具备**兼容传统以太网设备**的特点

ECO

生态构建方面，云计算通过部署**GPU统一开发生态**，平衡AI芯片的通用性与性能，促进多种架构的互通。

CUDA生态较为完善，开发者的数量超过400万；国内GPU生态尚未统一，不同类型、不同主体间应用接口、协议架构等标准不一，适配困难，跨架构迁移部署耗时较长。

在云环境下部署GPU统一开发生态，可以整合不同硬件厂商的开发架构。

MUSA架构

SUPA架构

CANN架构

Topsrider架构

通过构建GPU开发生态，从计算基础层、计算执行层、兼容适配层和应用层实现指令集、算子库、编译器等层次的自研，加速实现软硬一体协同式发展。

感谢观看
THANKS

