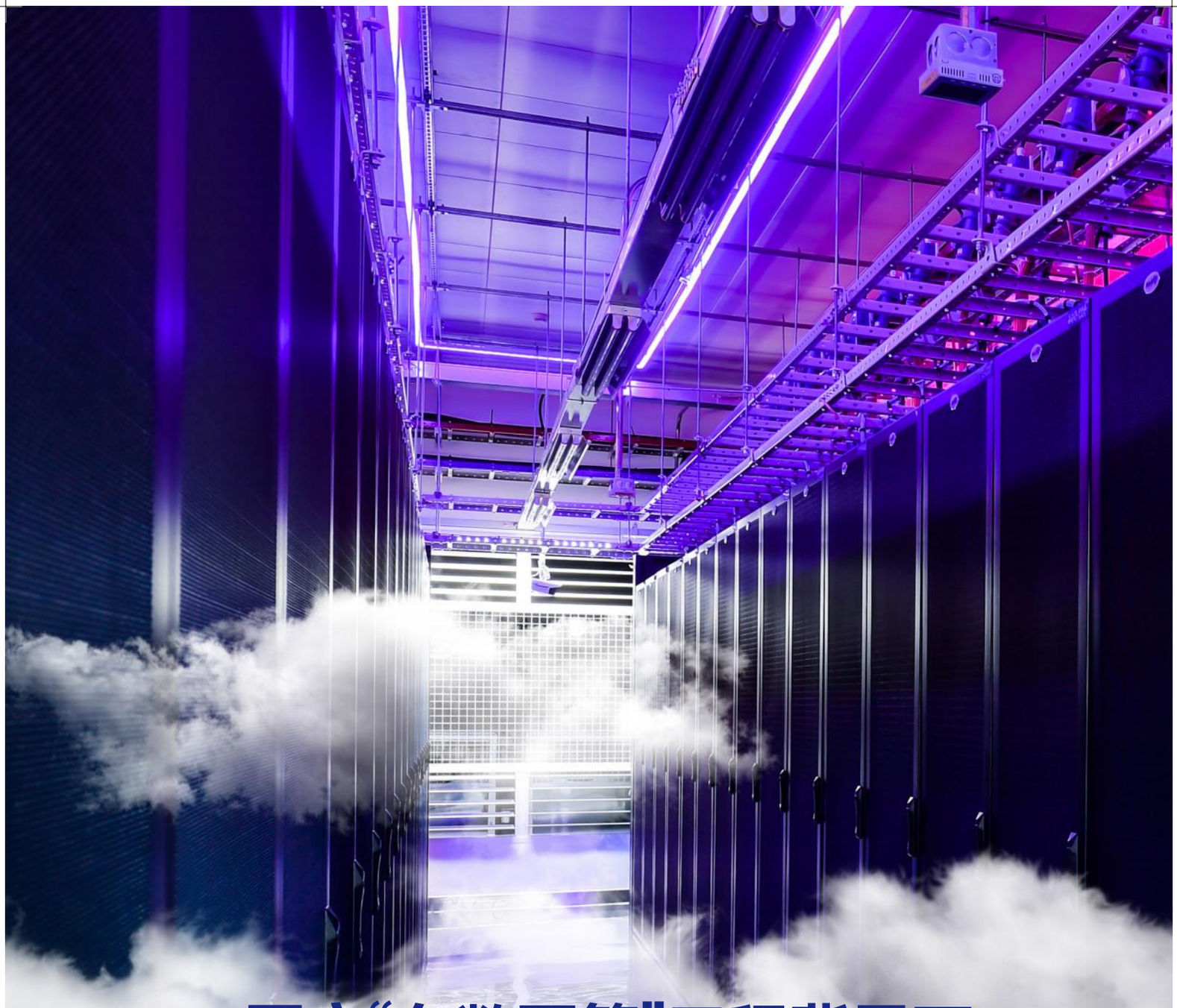
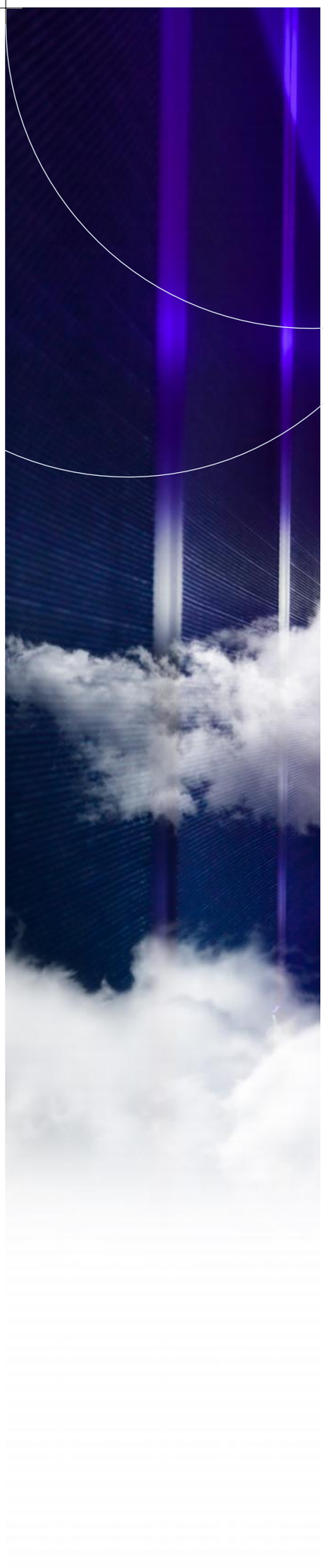


# 国家“东数西算”工程背景下
# 新型算力基础设施发展研究报告

单志广 何宝宏 张云泉 著

出品单位：智能计算 中国智能计算产业联盟

支持单位：首企研究院 E7 · RESEARCH



# 国家“东数西算”工程背景下
# 新型算力基础设施发展研究报告
# 编写委员会


顾问：　陈润生　陈国良　郑纬民　袁国兴

主编：　单志广　何宝宏　张云泉

执行主编：　安　静　王海峰　张广彬

编委：　沈文海　陈学斌　方　娟　贾海鹏　赖能和　袁　良
张延强　王丹丹　陈　栩　涂菲菲　刘　殷　戴　彧
宋心荣　舍日古楞　徐凌验　张　翼　李英浩



特别鸣谢：　清华大学　益企研究院



参编单位：　国家信息中心
信通院云大所
中科院计算所
清华大学
国家气象中心

# 目录 CONTENTS

P04 **“东数西算”定义和解读**

P08 **前言**

P09 **第一章 东数西算对算力新基建的影响**

P10 数据中心布局向供需协调有序、综合能效优化演进
P13 东数西算向时延要求低、存算要求高类型应用场景拓展
P15 算力结构呈现多元算力协同、算网一体化发展态势
P16 产业链上下游集聚发展，生态体系逐步壮大完善
P16 绿色低碳技术推广应用，清洁能源供给不断加大
P18 新老节点加快有序衔接，强化算力网络智能调度
P19 数字技能水平要求提升，技能人才需求不断凸显

P21 **第二章 算力新基建呈现的 10 大挑战和实践**

P23 算力基础设施化 保障资源多元供给
P32 关键信息基础设施的安全性要求
P36 信创产业化：国产化、自主化
P38 算力设施整体能耗偏高，绿色低碳应用仍需持续推广
P43 高密度 机柜功率密度提升
P46 算力智能调度：跨区域、跨云、云边调度
P50 多元算力 多样计算
P52 算力服务成为新业态
P54 原生应用：云原生、AI 原生
P57 规模化和算网融合

P63 **第三章 展望·面向 2030 年的算力基础设施**

P64 数字文明时代加速到来，要求算力基础设施资源充沛、泛在普惠
P65 隐私计算为代表的技术为组织间数据流通提供解决方案
P67 可信隐私计算是未来数据要素化的理想技术方案之一
P69 Web3.0 驱动规模化、泛在化的智能算力构建
P71 第五范式 AI for Science 对算力的需求
P73 大模型成为人工智能工程化重要方向，智能算力需求几何级增长
P75 边缘创新与新兴应用
P76 自动驾驶进入无人化新阶段，云边端高效协同

# “东数西算”定义和解读

## 一、东数西算工程

今年 2 月，国家发展改革委会同中央网信办、工业和信息化部、国家能源局等有关部门，同意在京津冀、长三角、粤港澳大湾区、成渝、内蒙古、贵州、甘肃、宁夏 8 地启动建设国家算力枢纽，并设立 10 个国家数据中心集群，正式启动“东数西算”工程，构建全国一体化大数据中心协同创新体系。

与“西气东输”“西电东送”“南水北调”等工程相似，“东数西算”是一个国家级算力资源跨域调配战略工程，针对我国东西部算力资源分布总体呈现出“东部不足、西部过剩”的不平衡局面，引导中西部利用能源优势建设算力基础设施，“数据向西，算力向东”，服务东部沿海等算力紧缺区域，解决我国东西部算力资源供需不均衡的现状。

2022年2月—至今，“东数西算”工程正式全面启动，各项政策引导持续跟进


| 东西部 | 枢纽节点 | 集群 | 起步区边界 |
|---|---|---|---|
| 西部 | 贵州枢纽 | 贵安数据中心集群 | 贵安新区贵安电子信息产业园 |
| | 内蒙古枢纽 | 和林格尔数据中心集群 | 和林格尔新区、集宁大数据中心产业园 |
| | 甘肃枢纽 | 庆阳数据中心集群 | 庆阳西峰数据信息产业聚集区 |
| | 宁夏枢纽 | 中卫数据中心集群 | 中卫工业园西部云基地 |

| 东西部 | 枢纽节点 | 集群 | 起步区边界 |
|---|---|---|---|
| 东部 | 京津冀枢纽 | 张家口数据中心集群 | 张家口市怀来县、张北县、宣化区 |
| | 长三角枢纽 | 长三角生态绿色一体化发展示范区数据中心集群 | 上海市青浦区、江苏省苏州市吴江区、浙江省嘉兴市嘉善县 |
| | | 芜湖数据中心集群 | 芜湖市鸠江区、弋江区、无为市 |
| | 粤港澳大湾区 | 韶关数据中心集群 | 韶关高新区 |
| | 成渝枢纽 | 天府数据中心集群 | 成都市双流区、郫都区、简阳市 |
| | | 重庆数据中心集群 | 重庆市两江新区水土新城、西部（重庆）科学城璧山片区、重庆经济技术开发区 |

图片来源：https://www.ndrc.gov.cn/fzggw/jgsj/gjss/sjdt/202209/t20220923_1336061.html?code=&state=123

## 二、东数西算内涵

“东数西算”是“全国一体化大数据中心协同创新体系”的一个下辖概念，而后者旨在推进技术、业务、数据融合，实现跨层级、跨地域、跨系统、跨部门、跨业务的数据协同管理和服务，其实现方式不是固定不变的。因此，不一定过度强调“东数西算”，面对不同应用场景，还可能有东数东算、南数北算等模式，应因地制宜。但无论哪种模式，都有着共同的目标：一是促进数据中心资源最大化共享、流通和利用，二是通过数据中心的系统化布局，促进国家碳达峰、碳中和战略实现。

## 三、东数西算与算力基础设施化

算力代表了对数据的处理能力，是数字化技术持续发展的衡量标准，也是数字经济时代的核心生产力。东数西算项目是促进算力、数据流通，激活数字经济活力的重要手段。

东数西算首次将算力资源提升到水、电、燃气等基础资源的高度，统筹布局建设全国一体化算力网络国家枢纽节点，助力我国全面推进算力基础设施化。

算力基础设施化并不简单等同于算力总量的拉升。算力的基础设施化并不是简单的算力堆砌，当前各类机构的算力总量测算方式都是将各行业、各公司的私有算力进行累加，甚至还会加上手机终端等移动端的算力，这些算力确实能够服务一定的群体，但算力资源并不能面向全社会提供统一一致的服务。

## 四、东数西算与绿色节能

东数西算是促进绿色节能，助力实现碳达峰、碳中和目标的重要手段。目前东部算力需求旺盛，但东部地区在气候、资源、环境等方面不利于低碳、绿色数据中心的建设。通过算力基础设施的西部迁移，可以充分发挥西部区域气候、能源、环境等方面的优势，引导数据中心向西部资源丰富地区聚集，扩大可再生能源的供给，促进可再生能源就近消纳，加强数据、算力和能源之间的协同联动，助力我国数据中心实现低碳、绿色、可持续发展，完成碳达峰、碳中和目标。

“东数西算”工程聚焦创新节能，在集约化、规模化、绿色化方面着重发力，支持高效供配电技术、制冷技术、节能协同技术研发和应用，鼓励自发自用、微网直供、本地储能等手段提高可再生能源使用率，降低数据中心电能利用率（PUE），引导其向清洁低碳、循环利用方向发展，推动数据中心与绿色低碳产业深度融合，建设绿色制造体系和服务体系，力争将绿色生产方式贯彻数据中心全行业全链条，助力我国在 2060 年前实现碳中和目标。

## 五、“东数西算”工程“五个一体化”的目标建设

东数西算把东部地区的非实时算力需求以及大量生产生活数据输送到西部地区的数据中心进行存储、计算并反馈。在其上则是希望构建更绿色、更平衡和更高效的国家算力网络体系，以满足新时代各行各业数字化转型、数字技术与生活场景加速融合所带来的海量计算、传输、存储需求，最大化实现数据中心产业绿色集约发展，推动资源统筹利用和西部数字经济建设。

- **网络一体化** 围绕集群建设数据中心直连网，建立合理网络结算机制，增大网络带宽，提高传输速度，降低传输费用。围绕集群稳妥有序推进新型互联网交换中心、互联网骨干直连点建设。

- **能源一体化** 从国家双碳战略整体规划出发，充分发掘西部丰富的风光等可再生资源，应对好可再生能源波动性问题，扩大清洁能源市场化交易范围，促进建立清洁能源消纳的市场化机制。从整体规划层面对数据中心集群进行统一能耗指标调配。

- **算力一体化** 在集群和城区内部的两级算力布局下，推动各行业数据中心加强一体化联通调度，促进多云之间、云和数据中心之间、云和网络之间的资源联动，构建算力服务资源池。

- **数据一体化** 建设数据共享开放、政企数据融合应用等数据流通共性设施平台。试验多方安全计算、区块链、隐私计算、数据沙箱等技术模式，构建数据可信流通环境。

- **应用一体化** 开展一体化城市数据大脑建设，选择公共卫生、自然灾害、市场监管等突发应急场景，试验开展“数据靶场”建设，探索不同应急状态下的数据利用规则和协同机制。

---

关于“东数西算”工程“五个一体化目标”阐述来源：
中国工程院院士、清华大学计算机科学与技术系教授郑纬民
https://www.ndrc.gov.cn/xwdt/ztzl/dsxs/zjjd1/202203/t20220321_1319866.html?code=&state=123

# 前言

2022 年初，国家发展改革委、中央网信办、工业和信息化部、国家能源 局联合印发通知，同意在京津冀、长三角、粤港澳大湾区、成渝、内蒙古、贵 州、甘肃、宁夏等 8 地启动建设国家算力枢纽节点，并规划了 10 个国家 数据中心集群，标志着东数西算”工程正式全面启动。

数据中心不仅是算力的聚集地，也是数据应用的发祥地，更是企业数字化 转型的根据地。2022 年发布的《“十四五”数字经济发展规划》第一条支线即为底层信息网络基础设施建设领域， 包括 5G、数据中心、光纤宽带等， 可以理解为此前反复强调的“新基建”的延续。

东数西算工程从国家战略、技术发展、能源政策等多方面综合考虑，将算力资源提升到水、电、燃气等基础资源的高度，统筹布局建设全国一体化算 力网络国家枢纽节点，在实现数据中心一体化协同创新的要求方面，给出了高质量的解决方案，助力我国全面推进算力基础设施化，其战略价值已经被大家认同。但在具体实施落地过程中，需要解决诸多问题，诸如实现数据中心有效整合、优化算力布局，降低算力成本、完成算力调度、实现算网融合、政府作用与市场力量有机结合等关键问题。

为此，中国智能计算产业联盟、益企研究院基于实践调查、探索研究后提 出了几个维度的洞察，并分析东数西算对新型算力基础设施发展的影响、新型算力基础设施的技术架构的迭代和演进，以及如何通过技术驱动提升算力新基建的竞争力。

# CHAPTER1

# 东数西算
# 对算力新基建的影响

## 数据中心布局向供需协调有序、综合能效优化演进

数据中心按照规模，分为超大型数据中心、大型数据中心和中小型数据中心。按照主要处理的业务类型，又可分为边缘计算类、低时延类、中时延类和高时延类。数据中心建设作为资本密集、技术密集型投资，数据中心地理位置的选择与其投资规模、投资成本、数据中心类型、服务质量、经济效益等因素紧密相关。随着全国一体化大数据中心体系总体布局的实施，在京津冀、长三角、粤港澳大湾区、成渝、内蒙古、贵州、甘肃、宁夏启动建设 8 个国家算力枢纽节点，并规划了张家口集群、长三角生态绿色一体化发展示范区集群、芜湖集群、韶关集群、天府集群、重庆集群、贵安集群、和林格尔集群、庆阳集群、中卫集群等 10 个国家数据中心集群。重点从顶层设计层面加强数据中心布局、算力、数据、网络、电力、能耗等方面的全国性统筹规划、一体化发展，数据中心的布局也将更加规范和优化。数据中心在选址布局时也将呈现如下变化。

一是数据中心选址向算力需求中心聚集。当前我国数据中心分布以大湾区、长三角、京津冀经济人口等较发达地域为主，在用数据中心中约 80% 集中在用户规模较大、应用需求强烈的互联网骨干节点所在省市及周边人口稠密、经济发达、总部企业密布一线城市。截至 2021 年底，北京及周边、上海及周边的数据中心机架数量排列分列一二。

图片来源：《2021 中国云数据考察报告》


但是随着北京、上海、广州等一线城市土地、能耗指标日益缩紧，对数据中心的政策约束愈加严格，已建数据中心已远远不能满足城市经济发展对算力的需求，一线城市周边地区逐渐成为互联网数据中心的首要选择。例如，紧邻北京的张家口、廊坊，靠近上海的南通、昆山，距离广深不远的韶关、云浮、清远等都是数据中心密集落户的地区。阿里云五大超级数据中心选址乌兰察布、张北、南通、杭州、河源，均位于在中心城市周边。随着 10 个数据中心集群的规划建设，数据中心供给结构优化，扩展算力增长空间，政策方面也明确要求对于符合条件且纳入国家枢纽节点数据中心集群范围的建设项目，积极协调安排能耗指标予以适当支持，实现大规模算力部署与土地、用能、水、电等资源的协调可持续。

二是数据中心选址向综合能效最优聚集。数据中心的主要运营成本包括土地、水、电、运维等要素，运营成本因为区位的不同有显著的差异，其中，高能耗导致的高电力成本是制约数据中心发展的主要因素之一。在土地、气候、政策、能源供给等优势加持下，数据中心建设选址向可再生能源丰富、气候适宜、数据中心绿色发展潜力较大、综合能 效最优的节点城市聚集，着重提升算力服务品质和利用效率，充分发挥资源优势，夯实网络等基础保障，积极承接全国范围需后台加工、离线分析、存储备份等非实时算


力需求，打造面向全国的非实时性算力保障基地。通信运营商、互联网企业等也纷纷将数据备份存储、大数据处理等对网络时延要求较低的业务向综合能效优势地区转移。例如，内蒙古乌兰察布“草原硅谷”，吸引了华为、阿里、快手等知名企业来此设立数据中心。西南地区的“云上贵州”吸引了苹果、腾讯、华为等企业，2021 年贵阳贵安成为全球集聚超大型数据中心最多的地区之一，数字经济占比达 34%。甘肃的“云天中卫”建成了亚马逊、美利云、中国移动、中国联通、天云网络、创客超算 6 个大型、超大型数据 中心，中国电信、炫我科技、爱特云翔、中国广电 4 个数据中心也在加快建设。

三是数据中心选址向用户终端靠近。随着超高清视频、虚拟现实 / 增强现实（VR/AR）、金融支付、金融证券、自动驾驶、工业制造、远程医疗等对网络时延要求较高的业务的广泛应用，靠近用户侧，作为算力“边缘”端的边缘计算型、中小型数据中心建设逐渐成为趋势。《广东省 5G 基站和数据中心总体布局（2021-2025）》提出，原则上只可建中型及以下的数据中心，承载边缘计算和低时延业务，中时延业务逐步迁移至粤东粤西粤北地区，高时延业务更要求迁移至省外。《北京市数据中心统筹发展实施方案（2021-2023 年）》提出，适度利用腾退后资源和空间改造建设边缘计算中心，支撑低时延业务应用，服务智慧城市、车联网等重点应用场景落地。除边缘计算中心外，东、西城区禁止新建或扩建数据中心。因此，数据中心选址时还会考虑数据

中心功能定位、数据处理要求、所承载业务的时延敏感性等因素合理选择新建数据中心的地理位置，例如自动驾驶（车联网）、工业制造（机器人）、远程医疗、金融证券等对时延非常敏感的网络应用的数据中心，可以选择在节点城市内部发展，服务后台加工、离线分析、冷数据存储备份等对时延不敏感的网络应用的数据中心，就可以优先向贵州、内蒙古、甘肃、宁夏节点转移，实现资源优化配置，提升资源使用效率。

## 东数西算向时延要求低、存算要求高类型应用场景拓展

带宽和时延是信息传输的两个关键指标，受限于物理规律，无论网络带宽多大，传输速度多快，传输时延都是客观存在的。因此，在“东数西算”中，工业互联网、灾害预警、远程医疗、自动驾驶等需要被计算节点频繁访问、网络时延要求高的实时在线类“热数据”不适合“西算”，而离线分析、后台加工、存储备份等离线类访问频率低、网络时延要求不高的“冷数据”以及介于两者之间的“温数据”，则更适合“西算”。虽然“东数西算”在网络时延上的限制使其不适用于时效紧迫型的数据应用，但是“东数西存”“东数西渲”“东数西训”，以及未来的“东云西库”等对存力、算力要求高，但对数据实效性要求不高的应用场景将成为“东数西算”未来应用的重要支点。

一是“东数西存”应用空间极其广阔。数字经济的发展推动海量数据的产生，这些数据的存储需要有强大的存力支持。从一般的统计来看，社会运行所产生的数据中，冷、温、热数据的占比分别为 80%、15%、5%，其中冷数据是存量最多的数据[1]。对于冷数据来讲，计算不是常态，其最主要的需求还

> “东数西算”只是“全国一体化大数据中心协同创新体系”的一个下辖概念，而后者旨在推进技术、业务、数据融合，实现跨层级、跨地域、跨系统、跨部门、跨业务的协同管理和服务，其实现方式不是固定不变的。

1. 邬贺铨：东数西算实为“东数西存”如何处理冷热数据值得研究
https://www.163.com/dy/article/HDHNHS2G0512D3VJ.html


是存储。随着“东数西算”工程的逐渐深入，内蒙古自治区、甘肃省、宁夏回族自治区、贵州省等枢纽节点省份存力规模将持续扩大，为西部数据中心承接东部“冷数据”“温数据”夯实了基础，其发展空间和发展潜力巨大。

二是“东数西渲”应用价值逐渐凸显。随着网络游戏、影视媒体的快速发展，云游戏XR、视频制作等渲染视频数据需求凸显，完成大规模的视频渲染离不开算力的支持，“东数西算”工程的实施为这些渲染业务提供了良好的基础能力支撑。随着云计算技术的逐渐成熟，渲染业务云化发展是大势所趋，通过建立渲染云应用平台，业务需求方可以将渲染任务快速提交到平台，从而获取算力、网络、存储一体化资源，实现资源编排、调度等的最优匹配。

三是“东数西训”应用潜力逐步释放。随着人工智能技术的快速发展，人工智能产业与经济社会发展的结合日益紧密。面向指数级增长的海量数据，想要有效激发数据资源的价值，离不开高级人工智能算法和强大算力的支持。特别是针对大规模人工智能任务应用场景下的复杂计算，东部算力资源的成本过高，为降低算力资源的使用成本，可以将训练数据和训练任务调度至西部枢纽数据中心集群进行上亿级参数的大模型深度学习，实现算网资源综合成本最优。

四是“东数西算”应用前景非常可期。随着一批中西部地区数据中心建成投运，国家高性能计算环境进一步完善，从试点走向规模应用，实现科学计算“东数西算”的基础条件日益成熟。一些算力需求巨大的科学计算应用，如格点量子色动力学、分子动力学

模拟、材料计算、生物信息等开始探索适用于“东数西算”场景。如格点量子色动力学的大规模数值模拟，涉及场景多、应用广，是最耗费计算资源的科研领域之一，其运算过程以及运算后海量组态数据分析，就适合于东数西算场景。又如材料基因组工程的各类高通量计算，每个计算任务之间无耦合，可以分配到任何可用节点，因此可以充分利用超算互联网连接的各超算中心的闲置资源。目前，中西部地区新建的数据中心已经开始面向全国科研用户提供绿色普惠的高性能一体化算力服务。

总的来说，“东数西存”“东数西渲”“东数西训”等应用场景是推动“东数西算”均衡发展的有效途径，通过对东西部算力、存力等资源供需关系的合理匹配，让“东数西算”融入到各个实际业务场景中，赋能千行百业的高质量发展。

## 算力结构呈现多元算力协同、算网一体化发展态势

一是算力设施多元化布局加快推进。一方面，“东数西算”工程布局空间跨度大，增加了数据传输时延，为有效解决这一问题，不仅需要高效灵活的东西部算力资源调度机制，也需要对东西部各类算力资源，包括通用算力、智算算力、超算算力、边缘算力等的配比进行优化，算力设施的异构化、多样化供给需求将明显增强。另一方面，智能化正以前所未有的速度在我国重塑各行各业，人工智能应用场景呈现出多元化、规模化发展趋势，除了通用算力，对智能算力的需求与日俱增。工信部《新型数据中心发展三年行动计划（2021-2023 年）》提出，到 2023 年底，全国数据中心总算力超过 200EFLOPS，高性能算力占比达到 10%。综上，我国算力基础设施迎来了大规模需求的繁荣期，结合不同应用场景需求的多元化布局将加快推进。

二是算力网络一体化格局加速形成。《全国一体化大数据中心协同创新体系算力枢纽实施方案》提出要构建一体化的新型算力网络体系，在通用算力方面，工信部数据显示，截至 2021 年底，我国在用数据中心机架总规模超过 520 万标准机架，平均上架率超过 55%。在智算算力方面，公开资料显示，当前全国智算中心已超过 20 个，主要分布在东部和中部地区。在边缘算力方面，我国还处于起步阶段，但在工信部公布的 2021 年国家新型数据中心典型案例名单中，已有 12 个边缘数据中心入选。随着“东数西算”工程的推进，以及多元算力适配与调度、算网融合等关键技术的突破，全国算力网络一体化格局将加速形成。

## 产业链上下游集聚发展，生态体系逐步壮大完善

算力新基建产业链条长、覆盖门类广、投资规模大、带动效应强，“东数西算”工程将推动新建数据中心尤其是大型、超大型数据中心向中西部地区以及北上广深等一线城市周边布局，同时带动相关产业有效转移集聚，促进东西部数据流动、价值传递。如中科曙光近年来在重庆、太原等地积极布局，探索实现了“以算促用”、“以算带动创新”。

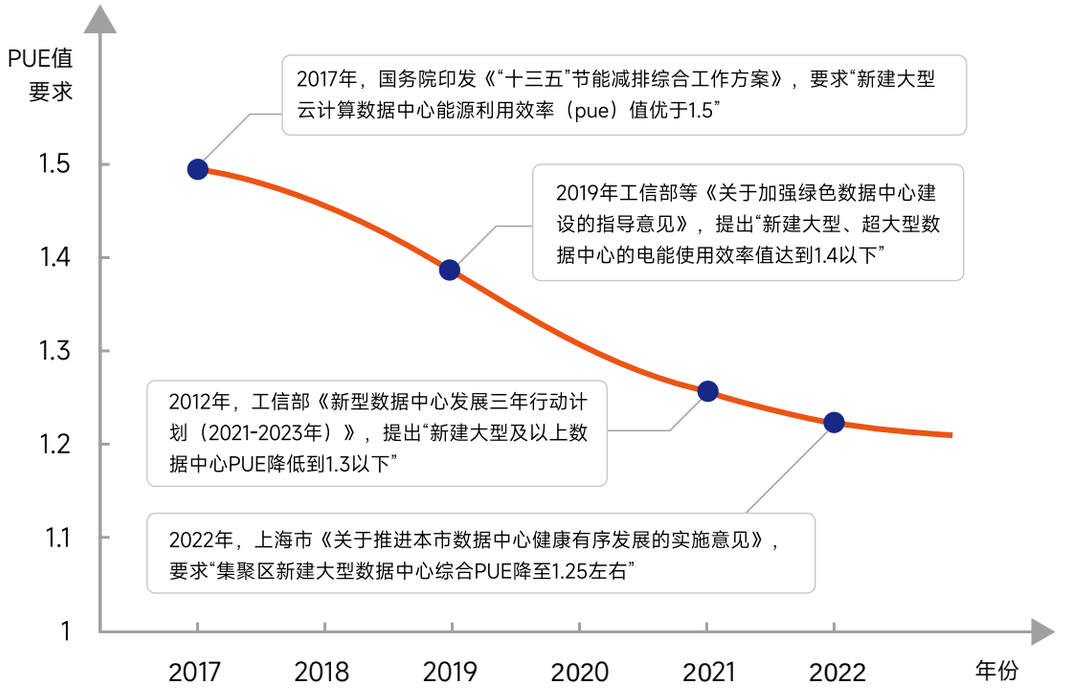
一方面，从中短期看将直接拉动产业链上游和中游发展。“东数西算”工程的启动实施，将直接拉动新一轮数据中心建设投资，IDC 基建率先受益。据统计，自 2022 年以来，全国 10 个数据中心集群中，新开工项目达到 25 个，数据中心规模达 54 万标准机架，带动各方面投资超过 1900 亿元 [2]。网络建设是算力均衡布局的基础，符合“东数西算”需求的网络特别是西部网络建设将全面提速，加速光纤通信向全光网演进。此外，在全球数据激增和“双碳”战略的大背景下，“东数西算”工程对数据中心建设标准更为严格，平均上架率至少要达到 65% 以上，对 PUE 也提出了更高的要求，绿色、低能耗的数据中心基建设备、边缘计算等环节将迎来持续发展机遇，温控散热技术有望实现升级。

另一方面，从长期看将逐步壮大完善产业生态体系。“东数西算”工程实施后将加速推动数据中心上游设备制造业和下游数据要素流通、数据创新型应用和新型消费产业等集聚发展，西部地区有望吸引数据加工、数据清洗、数据内容服务等偏劳动密集型产业落地。此外，随着数据规模和算力水平的提升，将有效激发数据要素创新活力，培育涌现出一批数据交易所、算力运营商、余热利用经销商、绿色数据中心评价机构等新模式新业态。

## 绿色低碳技术推广应用，清洁能源供给不断加大

因此，不一定过度强调“东数西算”，面对不同应用场景，还可能有东数东算、南数北算等模式，应因地制宜。但无论哪种模式，都有着共同的目标，一是促进数据中心资源最大化共享、流通和利用，二是通过数据中心的系统化布局，促进国家碳达峰、碳中和战略实现。

2. 国家发展改革委高技术司负责同志就“东数西算”投资建设进展相关问题答记者问
https://www.ndrc.gov.cn/fggz/fgzy/shgqhy/202204/t20220425_1323056.html?code=&state=123




国家和地方政策对数据中心 PUE 值要求演变图

一是数据中心绿色低碳发展。数据中心作为“东数西算”工程重要算力基础设施，高能耗是其显著特征。面向“碳达峰”“碳中和”发展目标，国家和地方持续出台一系列政策，进一步规范了数据中心的能耗管理和 PUE 值，如图所示，明确要求全国新建大型、超大型数据中心平均电能利用效率 (PUE) 降到 1.3 以下，国家枢纽节点进一步降到 1.25 以下，绿色低碳等级达到 4A 级以上。北京根据数据中心建设规模，将 PUE 准入水平分别设定为 1.3、1.25 和 1.15。上海数据中心新建项目 PUE 控制在 1.3 以下，改建项目控制在 1.4 以下，集聚区新建大型数据中心综合 PUE 降至 1.25 以下。旨在有序推动数据中心绿色高质量发展。

二是推动清洁能源有效利用。据统计，2021 年，我国数据中心年耗电量 2161 亿千瓦时，约占全国总用电量的 2.6%。且我国当前在用数据中心机架主要分布在北上广及其周边地区，能源使用压力巨大。从一体化大数据中心算力枢纽节点来看，内蒙古、甘肃、宁夏、贵州等省份是我国清洁能源大省，除贵州拥有丰富的水电资源外，其他三地都是风光资源的“富集区”，“东数西算”工程实施，承接东部算力需求的潜力，将大幅提升绿色能源

的需求，提高西部地区的绿色能源消纳水平。

三是节能减排实践加速涌现。液冷、蓄冷、高压直流、余热利用、蓄能电站等技术应用，以及太阳能，风能等可再生能源利用，进一步降低数据中心能耗及碳排放。中科曙光经过十年研制成功的浸没式相变液体冷却技术可以将计算系统的 PUE 值降到 1.04，达到全球领先水平。建设运营绿色低碳数据中心实践不断涌现，百度云计算（阳泉）中心应用市直供 +HVDC、自研“零功耗”置顶冷却单元及 AI 调优技术，年均达到 1.08。西部（重庆）科学城先进数据中心通过采用浸没液冷、光伏发电、微模块等技术，整体综合 PUE 低至 1.14。

## 新老节点加快有序衔接，强化算力网络智能调度

一方面，将加快新建算力设施和已有算力设施衔接配合。“东数西算”工程是一个让算力资源从过去的分散到相对集中，从个体运营到国家统筹的过程。“东数西算”不是单

图片来源：《2021 中国云数据考察报告》

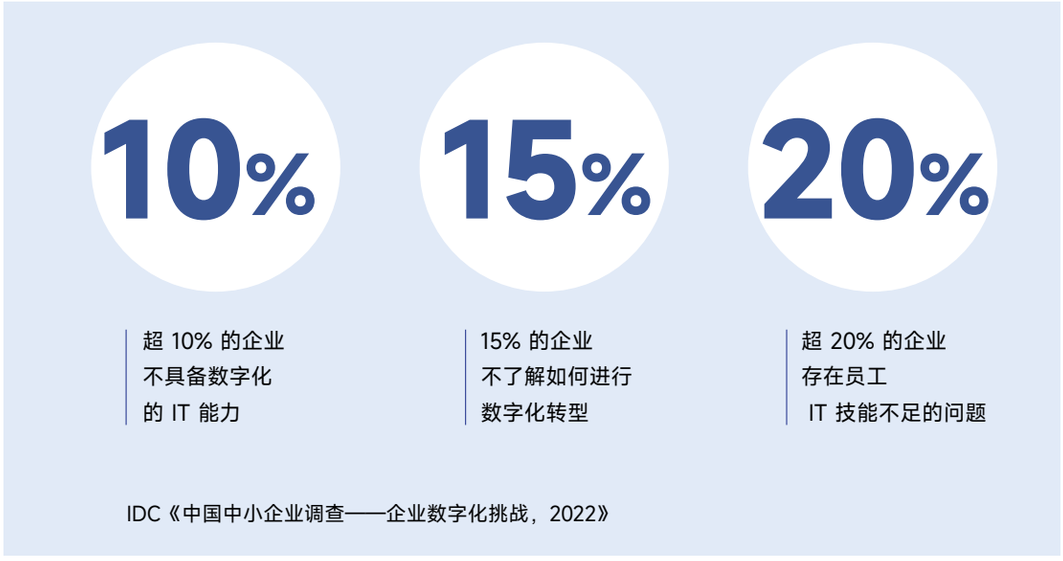
> “东数西算”工程是一个让算力资源从过去的分散到相对集中，从个体运营到国家统筹的过程。

纯的覆盖原有的算力设施，按照算力设施目前布局情况，除了“东数西算”要建设的节点集群外，原来各个地方已经有大量的数据中心、超算中心和智算中心，将促进新建和已有算力设施的有机融合。当前，京津冀、长三角、粤港澳大湾区、成渝、内蒙古、贵州、甘肃、宁夏、山东等地出台了一体化算力网络建设方案，通过科学规划布局，使国家数据中心的集群和现有的各地各类型的算力设施统筹衔接和配合，从而形成合理分工、科学联动、高效协作的机制。

另一方面，将推动东西部算力设施供需匹配和智能调度。“东数西算”要面临很多类型的数据处理需求，以及多样化的业务需求，如何保证东数和西算形成有效的供需匹配成为一个重要难题。如果没有业务保障，算力设施就无法发挥应有作用，从而造成设施空置和能源空耗。算力设施供需匹配将会在国家层面、省市级层面、企业层面、业务层面等多个层级开展面向性能、面向价格、面向效益的多方面测算，从而形成真正的应用需求供给和可持续的劳动力机制。进一步地，作为“东数西算”未来的神经中枢，算力网络的集中化调度是重中之重，有利于将所有的网络资源包括带宽资源和云资源进行统一调度，实现算力网络的云网协同。

## 数字技能水平要求提升，技能人才需求不断凸显

一方面，算力基础设施规模化绿色化智能化趋势明显，数据中心运营人才短缺。数据中心是数字技术创新的高地，随着数据中心绿色低碳、智能运营、算网调度等要求的不断升级，将加大对大数据、云计算和人工智能等相关高技术领域的人才需求。人社部中国就业培训技术指导中心的《新职业在线学习平台发展报告》指出，未来 5 年，大数据、云计算产业人才缺口将高达 150 万。掌握云架构、云配置管理、IT 基础设备管理、信息安全、数据中心综合管理以及节能减排等技能人员成为未来数据中心最为抢手的技术



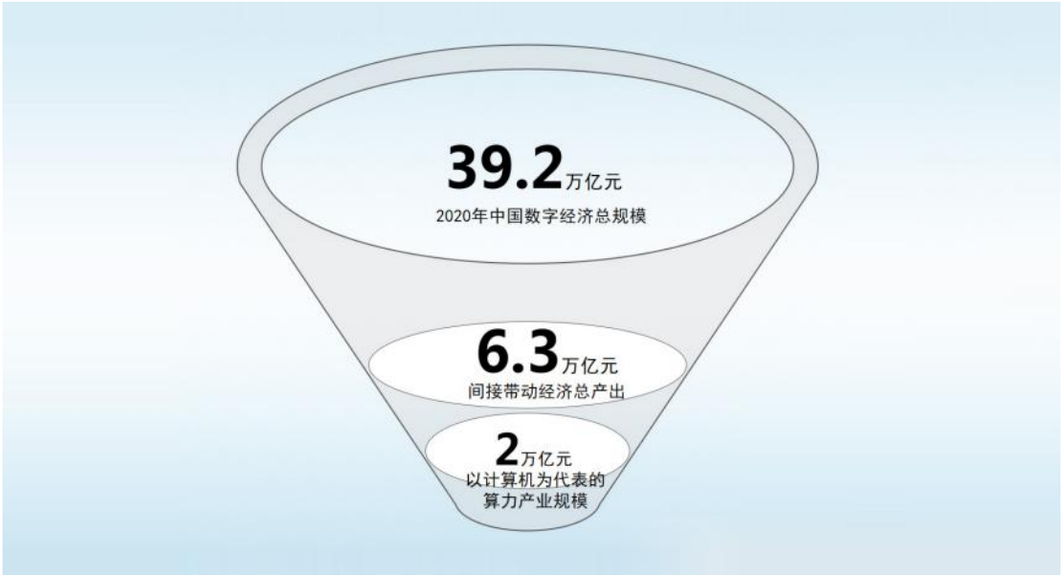

领域人才。目前，就业与招工难在数据中心行业同时并存，数据中心权威机构 Uptime Institute 的报告显示，调查的受访人当中，有一半表示目前很难找到空缺职位，远远高于 2018 年的 38%；从企业端看，数据中心对复合人才需求巨大，很多求职者并不符合招聘职位的要求，高等院校目前难以培养出足够的技术人才。

另一方面，算力设施普及将加速各领域数字化转型步伐，数字技能人才需求加大。“东数西算”不仅有助于改善数字基础设施不平衡的布局，而且有助于企业更好地提供云存储、云计算、数据工具、研发平台、AI 技术等服务，进一步降低上云用数成本，加快更多传统企业及中小企业实现数字化转型。数字化人才储备是数字化转型的关键，当前中国劳动力市场的数字化人才短缺。根据 IDC《中国中小企业调查——企业数字化挑战，2022》报告显示，超 10% 的企业不具备数字化的 IT 能力，15% 的企业不了解如何进行数字化转型，超 20% 的企业存在员工 IT 技能不足的问题。《数字经济就业影响研究报告》指出，2020 年中国数字化人才缺口接近 1100 万。根据人社部发布的相关报告测算，我国人工智能人才目前存在较大缺口，国内供求比例为 1：10，供需比例严重失衡。


# CHAPTER2
# 算力新基建
# 呈现的 10 大挑战和实践

2021 年 12 月，国务院印发《“十四五”数字经济发展规划》提出，到 2025 年，数字经济迈向全面扩展期， 数字经济核心产业增加值占 GDP 比重达到 10%。发展数字经济，已经成为中国经济战略的重中之重。

发展数字经济，算力又是重要支撑，算力作为核心生产力成为共识。2018 年，中国科学院计算技术研究所研究员张云泉提出算力经济概念，指出以计算为核心的算力经济将成为衡量一个地方数字经济发展程度的代表性指标和新旧动能转换的主要手段，算力经济涵盖算力基础设施、算力资源、算力服务和算力应用等产业。从中国信通院发布的数据来看，在算力中每投入 1 元，带动 3-4 元经济产出；算力发展指数每提高 1 点，GDP 增长约 1293 亿元。




数据来源：《中国算力发展指数研究报告》

同样，将罗兰贝格算力估算结果同 IMD 智能化水平评估结果进行比对，发现国家分类结果基本吻合，从而证明人均算力与国家智能化水平正相关。

算力在生产生活中的应用越来越多，数据潜力才会不断被挖掘，因此加快算力基础设施建设，优化算力资源布局、支持跨区域算力网络实时、灵活调度运营，提升算力应用强度至关重要。


资料来源：IMD，案头研究；罗兰贝格

## 一、算力基础设施化 保障资源多元供给

综合来看，算力是硬件和软件配合共同执行某种计算需求的能力。算力服务是提供算力的一种商业模式，是包括算力生产者、算力调度者、算力服务商以及算力消费者在内的算力产业链上算力经济模式的统称。

在加速算力服务核心技术创新发展方面，未来需要加强算力网络、基础设施化、开放应用模型、云边协同、云原生等算力服务核心技术布局，打造开放灵活的算力服务用户平台，推动算力经济供给侧改革，激发算力服务的范式创新。

让算力像水、电资源一样随取随用，使算力服务成为一种公共服务，是 61 年前“人工智能之父”约翰·麦卡锡的预测。但不同于标准化的电力，因为数据来源、结构存在多样性和复杂性，一些特定场景对计算能力的要求或者对计算特性的要求会越来越多，如

AI 服务、音视频服务等场景有足够的市场，显然通用计算无法满足其效率需求。与此同时，企业为寻求更加敏捷、灵活和高效的应用开发模式，以加速应用的创新和快速上市，如容器、微服务和 DevOps，这些应用开发模式拉近了业务和计算平台之间的联系，应用开发团队将定义基础设施的性能、可用性和规模，直接推动计算平台架构的变革和创新。

算力基础设施化并非易事，随着多样性技术路线的引入和发展，以 GPU、FPGA 为代表的异构计算与以 ARM 为代表 CPU 架构的兼容性问题更加突出，多样性算力的标准化度量与输出成为挑战。不同算力平台（超级计算中心、云数据中心、智能计算中心）的技术方案、系统架构、软件平台、硬件设备、服务保障存在很大差异。

要加快算力基础设施化进程，需要多类算力基础设施并行发展，保障算力资源的多元供给，围绕强化数字转型、智能升级、融合创新支撑来统筹布局云数据中心、智能计算中心，超级计算中心等算力基础设施建设。

### 1）算力服务能力是云数据中心的基石

云计算的推广，使得算力得以普惠化。用户按需采购算力、存储、带宽即可开展业务，可以将精力集中在拓展、开发新的应用，专注于本行业的知识创新，而不必在基础硬件、系统、网络、安全等需求上重复建设，也不用担心业务快速发展时受困于系统瓶颈。自此，不论是大中小型企业，亦或是个人，都可以通过不同形式的云 (IaaS、PaaS、SaaS) 获得需要的服务。

我国的云数据中心作为数字化基础设施的核心节点，这几年飞速发展。云数据中心不仅是算力的聚集地，还是数据应用的发祥地，更是企业数字化

“云计算的推广，使得算力得以普惠化。用户按需采购算力、存储、带宽即可开展业务，可以将精力集中在拓展、开发新的应用，专注于本行业的知识创新


转型的根据地。益企研究院在实地考察 8 个国家算力枢纽节点、7 个数据中心集群后发现，算力基础设施的使用效率，会直接影响到云服务商的创新能力和盈利能力。

全方位的计算力服务能力是云服务商竞争力的基石，云服务商不断优化硬件基础设施提升算力效率，尤其在服务器产品层面，通过高计算密度提供高算力和能效比，通过高速互联技术提升集群的扩展性，通过高度集成化设计、模块化和冗余设计简化交付部署流程，通过高效率散热系统打造绿色节能的集群系统。

在数据中心内部，基于云数据中心规模化优势，云服务商通过规模化、定制化能力支撑各项新型业务，将新的技术应用于云服务器来适配云端业务场景，通过完善从底层到应用层的自研技术体系，不断优化硬件基础设施提升算力效率，快速灵活对市场做出反应。

为通用算力输出单元的 CPU 也走向多元化：ARM 阵营百花齐放。对云服务商而言，一方面需要 CPU 有更强的核心和更多的核心数，另一方面需要不同的 CPU 满足客户多元化细分场景的需求，都与效率有着密不可分的关系。最大的变化是，在自主可控的大潮推动下，中国“芯”力量正在崛起，国产 CPU 龙头海光信息成功登陆科创板，成为 2022 年半导体领域知名的 IPO 事件，海光、龙芯、飞腾等产品的技术成熟度和应用范围正在追赶主流。

从数字中国万里行的洞察中发现，云数据中心完成多元算力的布局，但多元算力的多元的开发生态体系相对独立，应用的跨架构开发和迁移困难，亟需通过开源、开放的方式建立可屏蔽底层硬件差异的统一异构开发平台。

在数据中心基础设施层面， 新基建、双碳、东数西算，每年都有新热点，背后的指导思想则是一以贯之，兼顾效率与均衡、可持续的发展。从国家战略层面来说，希望通过建设高效集约、普适普惠的新型基础设施，推动计算力向绿色化和集约化方向加速演进。

### 2）智算中心成为新热点

人工智能需要海量的计算资源和存储空间，再加上非结构化数据的大爆发以及 AI 算法的快速演进，对传统计算范式造成了巨大的冲击，很多特定计算任务开始涌现，且需要在并行度、吞吐量和时延上做到极致。

无论是智慧城市还是智能制造、无人驾驶、数字孪生等场景，除了要有数据支撑以外，还要和各领域、各场景的知识模型、机理模型甚至物理模型相叠加，形成基于人工智能的新应用和场景实现。复杂模型、复杂场景势必需要面向 AI 的算力基础设施，即智算中心，智能计算中心。

集约化成为“智算中心”算力基础设施的趋势，通过 AI 服务器把算力高密度地集中在一起，解决了调度和有效利用计算资源、数据、算法等问题，同时减少闲置浪费，通过算力共享模式，大幅降低 AI 算力成本，支持更广泛的 AI 创新研究和应用。作为城市级公共算力平台，智算中心支撑类似大模型训练等大算力需求，满足区域内政府、企

业、高校等各类用户的算力需求，以 AI 专用芯片为计算算力底座，使用算力单位略有不同，集约化成为“智算中心”算力基础设施的趋势，通过 AI 服务器把算力高密度地集中在一起，解决了调度和有效利用计算资源、数据、算法等问题，同时减少闲置浪费，通过算力共享模式，大幅降低 AI 算力成本，支持更广泛的 AI 创新研究和应用。作为城市级公共算力平台，智算中心支撑类似大模型训练等大算力需求，满足区域内政府、企业、高校等各类用户的算力需求，需要配备多元融合算力。以曙光 5A 级智算中心为例，其通过分布式异构并行体系结构，搭载多类型芯片，实现全精度、多样性算力供应，满足包含数值模拟、AI 训练、AI 推理在内的不同应用场景需求。

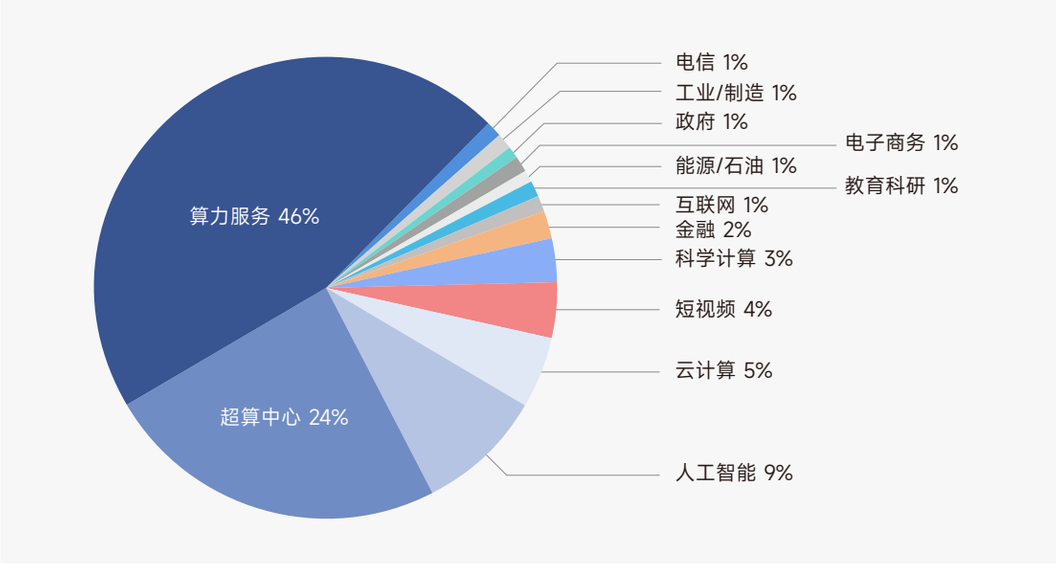
目前智算中心发展尚处于初期阶段却发展迅速，智算中心围绕模型算法来提供更好的数据和算力支持，需要将算法、模型、算力三者有机融合起来，输出 AI 的数据库、AI 的模型、AI 的开放平台等多种 AI 产品，让人工智能应用透明化，为政府、企业和科研院所提供普惠 AI 算力服务，真正支持数字经济的技术创新、平台创新、应用创新、生态创新和监管创新。

### 3）超算 2.0：赋能产业

超算算力是基于超级计算机等计算集群所提供的高性能计算能力，可进行普通计算机无法完成的工作，芯片以 CPU 为主，可含部分 GPU 加速器，以提供双精度浮点数（64 位）计算能力为主。

截至目前，已批准建立的国家超级计算中心共有十所，分别是国家超级计算天津中心、广州中心、深圳中心、长沙中心、济南中心、无锡中心、郑州中心、昆山中心、成都中心。

而从 2021 年 11 月发布的中国 TOP100 的行业应用领域趋势图和 Linpack 性能份额图来看，除了超算中心、人工智能、科学计算领域之外，高性能计算逐渐在生物制药、基因测序、动漫渲染、数字电影、数据挖掘、金融分析以及互联网服务等领域中扩展。




中国 TOP 100 行业应用领域机器 Linpack 市场份额图（2021.11）

数据来源：《2021 年中国高性能计算机发展现状分析与展望》

在应用领域新增算力服务，充分反映了在大数据、人工智能算法和算力三驾马车协同配合时代中算力经济的发展，算力的多样化正成为高性能计算领域的发展趋势。

目前，国家也重视超算互联网工程，整合多个超算中心包括云计算中心的软硬件资源，平衡算力的需求与供给，通过建设超算资源共享与交易平台，支持算力、数据、软件、应用等资源的共享与交易，同时向用户提供多样化的算力服务。

> 国家超级计算中心共有十所，分别是国家超级计算天津中心、广州中心、深圳中心、长沙中心、济南中心、无锡中心、郑州中心、昆山中心、成都中心

#### 4）云边融合 边缘暨核心

边缘计算可代表一类场景,基础需求是算力尽量地靠近用户。通常情况下,这些需求可以通过固网、光纤等技术来满足，但很多场景无法用“有线”来解决，比如在工业互联网、物联网、车联网等领域。

在不同场景下，人们对边缘的理解不同， 就运营商而言，一是从行政区划角度，从全国到省一级，再到地市级、县乡级，越远就相对越边缘；另一个角度是从运营商组网层面，从接入网到核心网，再到数据中心内部，靠近接入网的站点，就可定义为边缘站点。

而从应用的角度，除了大型数据中心以及云计算中心节点之外，都可以称作边缘。比如从最接近用户侧的家用路由器或者工厂里的工控机，到园区的计算以及数据处理设备，以及一些区域数据中心，都可以称作边缘的数据中心。

无论是 5G 还是边缘计算，主要目标就是为行业用户提供服务，尤其是算力服务。

但中国的行业种类繁多，不同的应用场景必然导致不同的算力需求，边缘计算的业务差异性大，这些差异性的业务在一个相对规模不是很大的汇聚节点要去呈现，在一个资源相对有限的边缘数据中心去满足不同业务提出的算力需求，需要边缘数据中心的设备能够尽量通用、开放。

这就需要积极构建城市内的边缘算力供给体系，支撑边缘数据的计算、存储和转发，满足极低时延的新型业务应用需求。引导城市边缘数据中心与变电站、基站、通信机房等城市基础设施协同部署，保障其所需的空间、电

力等资源，需求牵引，在工信部发布的《新型数据中心发展三年行动计划（2021-2023年）》中提到，深化协同，推动新型数据中心与网络协同建设，推进新型数据中心集群与边缘数据中心协同联动，促进算力资源协同利用，加强国际国内数据中心协同发展。

### 5）算网协同 算网融合

无论是边缘数据中心间，边缘数据中心与核心数据中心集群间的组网、还是不同算力集群之间组网，提升算力网络支撑能力，构筑新型的算力网络基础设施是推动算力基础设施化的重要前提和举措。强化算网的协同融合发展，优化东西部数据资源的结构，需要一体化、高质量的数据中心集群和互联网络协同支撑。

从工信部在《新型数据中心发展三年行动计划（2021-2023 年）》中提到的目标来看，到 2023 年底，全国数据中心机架规模年均增速保持在 20% 左右，平均利用率力争提升到 60% 以上，总算力超过 200 EFLOPS，高性能算力占比达到 10%。国家枢纽节点算力规模占比超过 70%。新建大型及以上数据中心 PUE 降低到 1.3 以下，严寒和寒冷地区力争降低到 1.25 以下。国家枢纽节点内数据中心端到端网络单向时延原则上小于 20 毫秒。

这其中，算力网络是一个系统工程，包括算力供给、算力管理、算力服务等多个方面。一方面要尽快补足算力枢纽节点间网络薄弱环节，另一方面，逐步建立算网协同联动

> 国家算力网络是一个系统工程，包括算力供给、算力管理、算力服务等多个方面。一方面要尽快补足算力枢纽节点间网络薄弱环节，另一方面，逐步建立算网协同联动机制，推动算力网络需求和供给有效对接。

机制，推动算力网络需求和供给有效对接。加快实现算力网络高效、智能、敏捷的调度与应用。

算力网络使能各行各业数字化转型，需要各方共同打造新架构和底层技术，构建灵活敏捷的算力底座，打造泛在多维立体的算力网络，来满足算力资源高效连接，按需分配，灵活调动。

为此，可充分发挥我国的体制优势和市场优势，提升自主创新能力。加快算力网络智能运维和融合架构等领域的创新突破，通过算力设施由东向西布局，带动相关产业有效转移，促进中西部地区数据流通、价值传递，推动算力设施能效水平和绿色用能水平的提升。

## 二、关键信息基础设施的安全性要求

数字世界有多高效，可能也就有多脆弱。安全问题近些年变得愈演愈烈，勒索病毒与黑客攻击无时无刻不在进行。受新冠疫情流行和全球数字化进程加快的驱动，一定程度上因网络开放度的提升和接口的增多，给勒索病毒提供了新的攻击面。

2021 年美国最大的成品油管道运营商科洛尼尔管道运输公司（Colonial Pipeline）就遭受病毒勒索，随后，科洛尼尔主动切断了某些系统的网络连接，造成油料运输不正常，导致东海岸 45% 的汽油、柴油等燃料供应受到影响，最后公司无奈支付了 500 万美元的赎金得以释放。就在2022年3月，丰田汽车供应商小岛工业（KojimaIndustries Corp）公司同样受到“勒索软件”攻击，导致丰田在日本所有的 14 家工厂 28 条产线全面停产，导致丰田当月产能下降 5%，损失上亿美元。

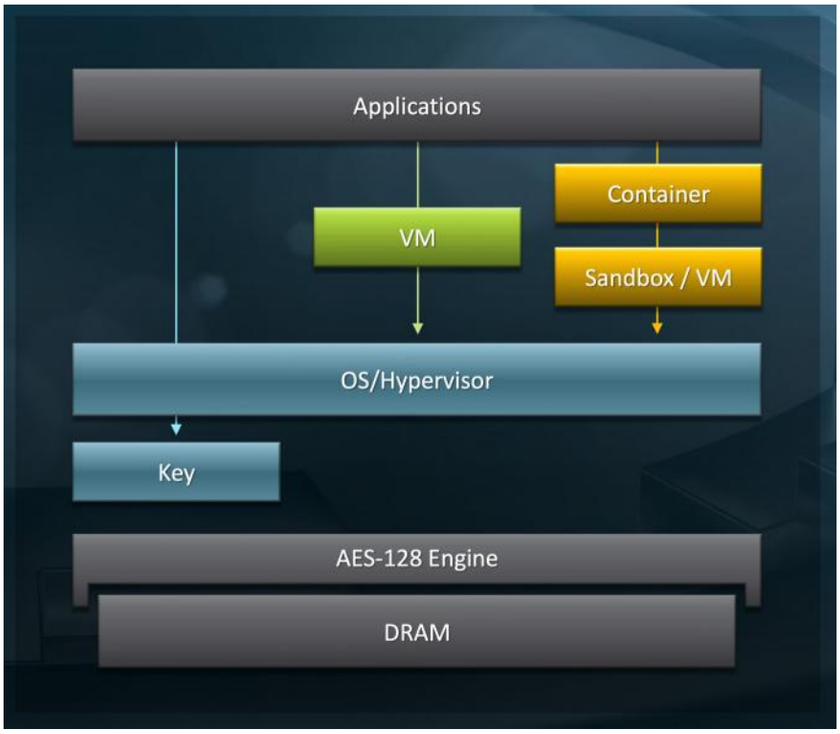
当恶意攻击行为的攻击面扩大，客户希望得到从芯片到安全 OS、软件接口、安全应用等业务的全流程、全生命周期的数据安全服务，基于硬件的可信执行环境 TEE（Trusted execution environment）在云计算环境中成为趋势，算力提供商正努力在源头上封堵可能存在的漏洞。

### 1、独立的安全子系统

从算力基础单元 CPU 来看，当 CPU 的核心越来越强大， L3 Cache 的容量也一直在成倍增长。这些核心功能的持续进步，容易让人忽略，其安全特性也在不断完善。

以 AMD 第三代 EPYC 处理器为例，其安全性建立在一个独立的安全子系统之上，其核心是 CPU 集成的安全协处理器，这是一个基于 Arm Cortex-A5 的 32 位微控制器。安全协处理器运行一个安全的操作系统 / 内核，安全的片外非易失性存储 ( 如 SPI ROM) 保存固件和数据，提供安全密钥生成和密钥管理等加密功能，启用经过硬件验证的引导。

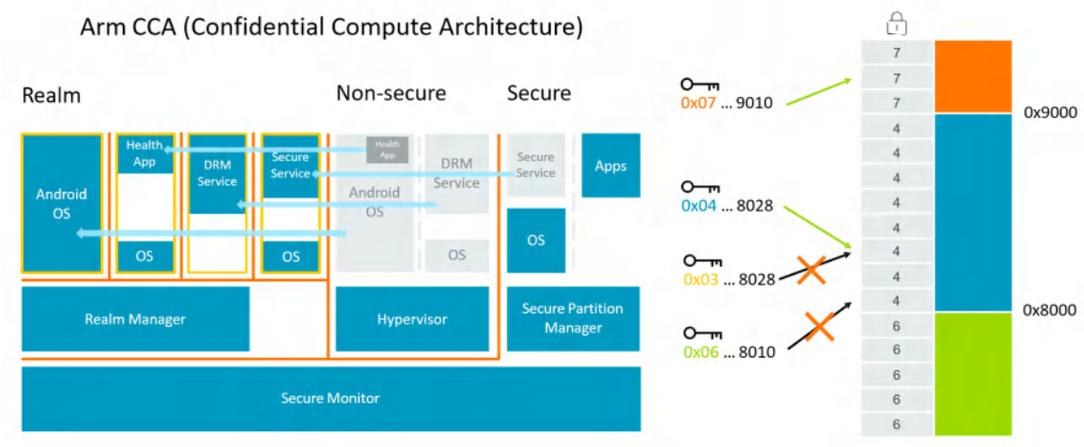
在硬件验证的引导过程中，安全协处理器加载片上引导 ROM，该 ROM 加载并验证片外引导加载程序。引导加载程序在 x86 核心开始 执行 BIOS 代码前验证 BIOS，也验证和加载安全协处理器使用的代码以提供密钥管理。




EPYC 处理器的 SME 设计　　图片来源：《2021 中国云数据考察报告》

安全加密虚拟化 (Secure Encrypted Virtualization，SEV) 在云计算时代的重要性不言而喻，它在虚拟机以及 hypervisor 之间提供强加密隔离，根据虚拟机 ID 选择活动的加密密钥。

在安全性方面，Confidential Compute Architecture(CCA)，中文名称为机密计算架构，这是一种基于架构层面的安全防护能力， 通过打造基于硬件的安全运行环境来执行计算，保护部分代码和数据，免于被存取或修改，乃至不受特权软件的影响。




Arm 机密计算架构（左），Android 11 和 OpenSUSE 引入的内存标签扩展技术（右）

图片来源：《2021 中国云数据中心考察报告》

为此 CCA 引入了动态创建机密领域 (Realms) 的概念：一个安全的容器化执行环境，支持安全的数据操作，可将数据与 hypervisor 或操作系统隔离。Hypervisor 的管理功能由“领域管理器”(realms manager) 承担，而 hypervisor 本身只负责调度和资源分配。使用领域的优势在于极大地减少了在设备上运行给定应用程序的信任链，操作系统在很大程度上对安全问题变得透明，也允许需要监督控制的关键任务应用程序能够在任何设备上运行。

在实际应用中，内存是非常容易被攻击的一环，内存安全也一直成 为行业的关注点，如何在内存安全漏洞被利用之前就能发现问题，是提高全球软件安全的重要一步。

## 2、从硬件直达云上的内生安全能力

从数据中心的角度，需要实现从数据中心的防护到数据防护的一体化解决方案，这意味着从数据中心的设计规划到建造到使用和运维，需要从生命周期安全、技术安全、管理安全、安全运维等多维度层面保证安全。

而从云平台的角度，通过对计算、存储、网络、安全、基础软件等软硬件资源的统筹管理，借助 IaaS 和 PaaS 的资源实现安全能力，发挥云计算的优势，实现物理安全、网络安全、应用安全和数据安全。

尤其在 2021 年 9 月《国家数据安全法》颁布后，在如何实现数据流通和数据安全的平衡方面，开始探索数据治理的新理念和新方法，从法律、技术、安全、制度等多个角度，为部委、省市政府、央企等客户系统性地推进数据治理工程提供有效支撑与服务。

以数字中国万里行考察的数字大理苍洱云平台为例，该平台通过物理安全、硬件安全、系统安全、虚拟化安全建设，聚焦云平台安全监控和安全运营、身份访问控制和全链路数据保护，采用架构服务器构建可信云底座，构建以“本质 + 过程”的全栈云原生安全防护体系，满足大理州云上应用系统安全防护需求。

图片来源：《2021 中国云数据考察报告》

## 三、信创产业化：国产化、自主化

从 2020 年开始，信创趋势越发明确，站在用户角度，信创落地最佳的方案是通过云计算的方式，建立一云多芯的混合 IT 体系，同时基于信创平台进行可进化云原生研发。2021 年，一个新的现象是全栈国产化云在政府、物流、金融、交通、电力等行业应用落地速度加快，为网络信息体系建设提供完全自主可控的基础支撑环境。国产化云采用自主和安全的架构体系搭建，从底层芯片、服务器的底层硬件与操作系统、数据库到云服务的全栈打通，整合集成、监理、运维、安全、项目管理等服务，既满足行业的数字化升级需求，还有效保证国家和企业云服务的安全性和可控性。

在自主可控的大潮推动下，算力成为了企业发展的核心助力之一。 基础设施的技术架构迭代中，中国“芯”力量开始登上舞台，国产技术的成熟度和应用程度正在提升，从传统的电脑到服务器，从芯到云，信创产业化加速进程中，“承上启下“的适配作用不可忽视，如果数字底座全为国产化，需要全流程的适配服务，包括为国产信创云的迁移适配提供组织规划、适配认证服务、资源保障、人才保障以此来支撑云平台上的不同智慧场景。这样国产化从“可用”迈向“好用”。

中国电子信创云基地（顺义）　　图片来源：《2021 中国云数据考察报告》

在顺义，中国电子按照国家关键信息基础设施的标准打造了中国电子信创云基地，支撑异构多节点云的管理，整体架构基于飞腾 ARM 架构和 x86 架构构建云平台资源池，其中国产化飞腾 ARM 体系满足国家安全规定，自主安全要求的信创基础设施资源池，x86 体系的资源，作为现有部分适配难度较大的业务运行的非信创过渡资源池，服务诸多央企和政府用户。

长沙人工智能计算平台核心模块，采用曙光 5A 级智算中心建设方案。一方面基于自主硬件构建异构计算平台，确保实现混合多元算力覆盖，满足不同需求的计算模拟仿真、人工智能模型训练推理、大数据分析与可视化等多类应用场景；另一方面，基于全栈软硬件技术，包括操作系统、深度学习框架、管理平台、开发平台、大模型等，促进自主硬件与学习框架的深度适配与优化，支撑企业用户进一步开发、移植和优化算法模型和应用软件。

## 四、算力设施整体能耗偏高，绿色低碳应用仍需持续推广

新型节能新技术的应用程度有待提高。我国数据中心总体上还处于小而散的粗放建设阶段，大型、超大型数据中心占比仅为 12%。大部分中小数据中心多依赖空调、冷水机等设备来降温，受自然冷源、气候等环境因素影响，解热极限相对较低。据数据中心绿色能源技术联盟统计，2021 年度全国数据中心平均 PUE 为 1.49，并且有相当数量的数据中心 PUE 超过 1.8 甚至 2.0。随着 ICT 设备器件性能提高和单机柜功耗的增加，发热量随之上升，数据中心制冷系统的电能消耗还在不断持续增高。由于早期政策相对宽松，精确监管存在困难，市场应用规模有限，产业链成熟度不足，可靠性不足和不合理等原因，相关创新节能技术并未大规模应用。

新型液冷技术有待加大推广应用。液冷是指借助高比热容的液体作为热量传输介质满足服务器等 IT 设备散热需求的一种冷却方式。有数据显示，液冷比传统风冷具备更强的冷却能力，其冷却力是空气的 1000-3000 倍，热传导能力是空气的 25 倍。同等散热水平时，液冷系统噪音比风冷低 25-35 分贝，相比传统风冷系统约节电 30%-50%，数据中心 PUE 值可降至 1.2 以下，甚至接近于 1。例如在北京冬奥云数据中心部署了浸没式液冷集群，对数据设备采用了环保节能的自然冷却技术，年平均 PUE 低于 1.2，大幅度降低了碳排放量。受限于我国数据中心建设规模和政策要求，液冷技术尚未得到广泛应用。

> 在东数西算的布局中，从双碳角度看，西部的数据中心，从双碳角度应该有两大优势：一是就地消纳丰富的能源供给，特别是可再生能源（如风能和太阳能）；二是气候条件好，可以充分利用自然冷源，降低对电能的消耗。

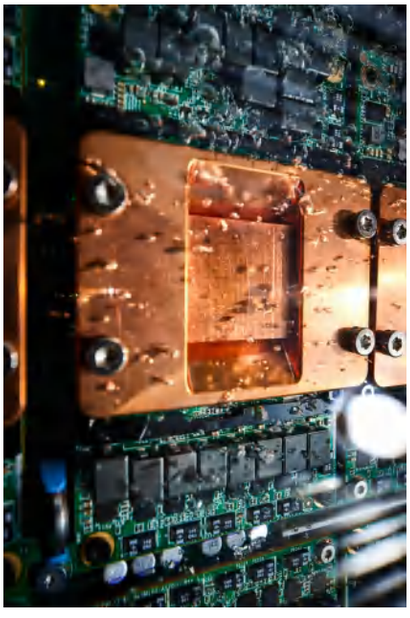
图片来源：《2021 中国云数据考察报告》

从我国数据中心的实践来看，这个问题可能更加棘手。统计数据显示，目前我国仅有 41% 的数据中心 PUE 在 1.4 以下。而在“东数西算”工程发布的文件中可以看到，此次各地区 PUE 目标东部地区不超过 1.25，西部地区不超过 1.2，能效指标更严格。

如何通过技术方案有效降低 PUE 是温控系统发展的重点。为了给数据中心计算节点的核心部件降温，技术专家们尝试了风冷、水冷、温水冷却、冷板式液冷，最后将目光投向了浸没式液冷。

这些特殊液体沸点较低，通过遇热气化将设备内部的 CPU、内存、电源系统等发热部件产生的热量转移出 IT 设备，再与水做热交换，最终将热量排出，以达到良好的降温节能效果。

图中液冷计算节点能够将数据中心能效比 PUE 降至 1.1 以下，比传统风冷技术节电 20%。这是曙光自主研发的浸没式相变液冷技术，曙光浸没



《求是》杂志 2022 年第 2 期稿件配图，摄于中科曙光北京一数据中心

式相变液冷技术可助数据中心实现全地域全年自然冷却，PUE 值最低可降至 1.04。液冷方案的优势主要是靠近热源、温度均匀、能耗低，其方案占比正在快速提升。

当然任何技术的发展不是一蹴而就，十年前曙光开始探索液冷技术，并率先在全国开始浸没式液冷服务器大规模应用的研发。截至目前，曙光拥有液冷核心专利超 60 项，部署的液冷服务器节点已达数万台，居国内市场份额之首。

另外，数字中国万里行团队考察阿里巴巴浙江云计算仁和数据中心发现，采用了服务器全浸没液冷等多项节能技术进行规划设计与建造，运算产生热量可被直接吸收进入外循环冷却，全程用于散热的能耗几乎为零，PUE 低至 1.09。

在东数西算的布局中，从双碳角度看，西部的数据中心，从双碳角度应该有两大优势：一是就地消纳丰富的能源供给，特别是可再生能源（如风能和太阳能）；二是气候条件好，可以充分利用自然冷源，降低对电能的消耗。其中，很多符合“西算”标准的数据中心，广泛应用了以间接蒸发制冷为代表的节能方案，在张家口数据中心集群、和林格尔数据中心集群的数据中心，一年有 10 个月以上的时间可以使用自然冷源，年均 PUE 可达 1.2 。

大型互联网和云计算公司主导的超大规模数据中心，将对液冷服务器的大规模应用产生决定性影响。因为他们既有足够的体量和应用需求，对数据中心建设也有足够的掌控能力。

以“东数西算”成渝枢纽节点内的曙光承建的西部（重庆）科学城先进数据中心为例，该数据中心采用了浸没液冷技术、余热回收、绿色建筑、清洁能源（光伏）等多种相关技术，项目年均 PUE 可达到 1.144，年节省用电约为 14624.8 MWh，年节省标准煤 4870 吨，年减少二氧化碳排放 13149 吨。 真正做到了从能源的使用、机架的合理选用、到散热的合理规划、机房设计、布局和使用等多方面的合理布局，全面提高机房散热效率，降低机房的整体能耗，最终达到节能减排的目标。

对清洁能源的开发利用还有较大提升空间。数据中心面临着区域性发展不均衡的问题，东部地区供给不足和西部地区供给过剩的结构性矛盾较为突出。据测算，由于光伏和风力等可再生能源的不稳定特点，我国西北部地区每年弃风弃光电量约 125 亿度，如果在这些地方依托电厂和电网布局就近建设大型以上数据中心，并利用储能系统和调度系统创新解决稳定负载的柔性供能问题，可以促进可再生能源开发利用，有效降低中西部地区弃风和弃光电量，进一步减少碳排放。

绿色低碳循环发展需要持续推进。目前，数据中心节能减排主要集中在前端绿电应用、制冷系统节能减排、IT 系统降耗等方面，余热回收利用方面因大多数数据中心采用风冷降温，携带热量介质为空气，存在余热流动缓慢、不适合长距离运输等缺陷，余热收

图片来源：《2021 中国云数据考察报告》


集及运输难度较大，成本较高、回收利用率低，所以，绝大部分余热直接排向空气。2021 年 7 月，国家发改委印发《“十四五”循环经济发展规划》，提出将“推进工业余压余热、废水废气废液的资源化利用，实现绿色低碳循环发展，积极推广集中供气供热”作为重点任务。数据中心余热回收利用也是通过梯次综合能源利用，是促进全行业节能降碳重要探索方向。通过对来自数据中心的热量进行回收再利用，为附近住宅、医院、办公、酒店等用热单位持续供暖，替代其他用于供暖的能源。

据测算，从数据中心总耗电量中，可大约提取回收 11.2% 电力消耗产生的余热。以我国 2020 年数据中心耗电量 2000 亿千瓦时估算，如果这些余热被完全利用将减少约 2230 万吨二氧化碳排放。

在数据中心节能方面，目前业界对清洁能源利用、机房建筑节能设计、余热回收、服务器硬件节能等方面进行了较多的探讨。除此之外，对于软件复合节能优化的研究开始在起步阶段，如数据库作为云计算的基础服务之一，其性能的提升将会直接影响硬件设备的使用效率。今年 8 月，腾讯云联合多家产业机构与中国电子节能技术协会发

> 2021 年 7 月，国家发改委印发《“十四五”循环经济发展规划》，提出将“推进工业余压余热、废水废气废液的资源化利用，实现绿色低碳循环发展，积极推广集中供气供热”作为重点任务。数据中心余热回收利用也是通过梯次综合能源利用，是促进全行业节能降碳重要探索方向。

布国内首个数据库节能减排报告《键值型数据库技术及节能要求》，这个标准的应用，让腾讯数据中心的节能能够达到 30% 以上，为行业内软件的节能减排提供了解题新思路。

腾讯云全网超过 100 万台服务器，1.5 亿核 CPU，相当于 2500 万台主流配置的个人 PC，超过 2021 年中国台式 PC 全年出货量，通过腾讯云遨驰分布式云操作系统的高效调度，可以提升 30% 以上利用率，相当于节省了 30 万台服务器，一年可以节电约 2.5 亿度，减排二氧化碳量达 5.24 万吨，碳排放当量约为种植 286 万棵大树，约合 2500 个足球场面积的森林。

## 五、高密度 机柜功率密度提升

数据中心的生命周期包括规划、设计、建设、运行和评估等阶段，功率密度是数据中心在规划和设计阶段需要明确的一个重要参数，当前业界普遍接受用“单机架用电”参数来表示数据中心功率密度。

通过梳理全球数据中心产业的发展情况，发现近年来数据中心作为信息基础设施，其功率密度在逐年上升。据 Uptime Institute 发布的《2020 全球数据中心调查报告》显示，2020 年全球 71%的数据中心平均功率密度低于 10 kW / 机架，最常见的密度是 5~9 kW / 机架，平均功率密度高于 20 kW/ 机架的数据中心约占 16%。虽然整体功率密度相较高性能计算（HPC）等领域还不高，但总体上升趋势明显。2020 年数据中心平均单机架功率

为 8.4kW/ 机架，相比于 2017 年的 5.6kW/ 机架、2011 年的 2.4 kW/ 机架有明显提高，年复合增长率达到 15%，预计未来数据中心的功率密度还将继续上升。

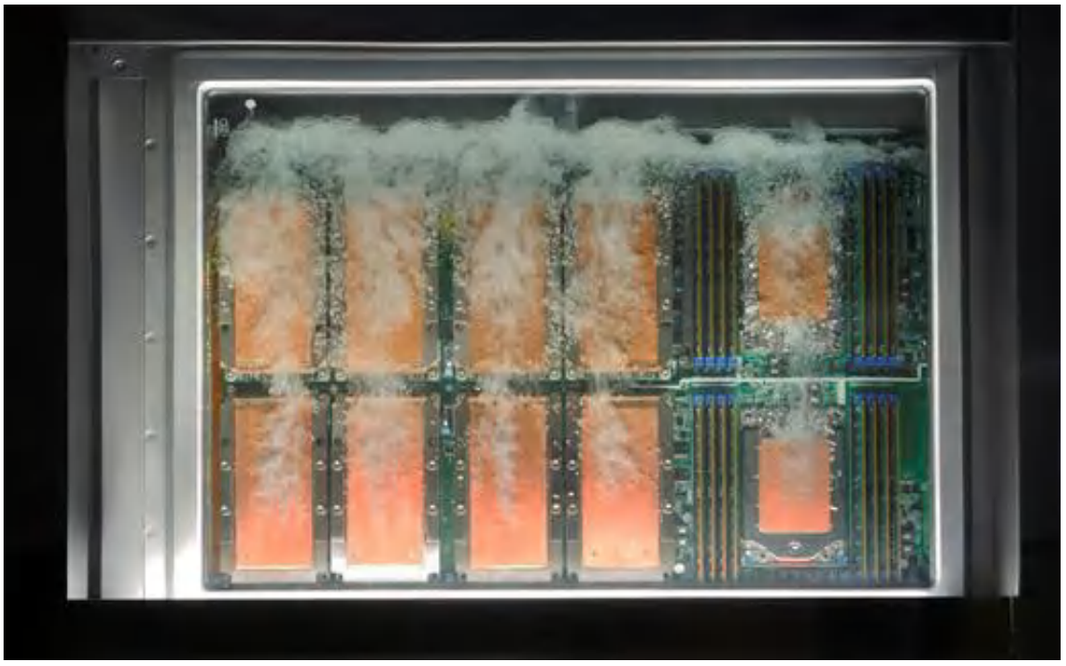
基础电信运营商、第三方数据中心服务商及大型互联网企业是我国数据中心的主要参与者。目前在我国数据中心市场中基础电信运营商仍占据主要市场，第三方数据中心服务商是除基础电信运营商外的重要组成部分，BAT 等大型互联网企业也成为重要的数据中心持有和运营主体。经过统计分析，基础电信运营商、第三方数据中心服务商和大型互联网企业的数据中心功率密度情况为：

1）我国三大基础电信运营商建设数据中心主要满足通信建设需要及带宽租用、云计算服务等业务。根据中国信通院对截至 2020 年已用情况进行调研统计，当前三大基础电信运营商在用数据中心功率密度平均约为 4.46kW/ 机架。同时发现基础电信运营商数据中心功率密度与数据中心规模呈正相关关系，数据中心规模越大，部署功率密度相对也越高。

2）除基础电信运营商外，数据中心租赁和服务市场最大的参与群体是第三方数据中心服务商，目前我国第三方数据中心服务商数量多、分布零散。通过分析，第三方数据中心服务商数据中心功率密度与服务的客户和承载的业务紧密相关，当数据中心的用户群体集中为互联网企业、云服务商，主要提供批发服务时，功率密度受上层密集的计算业务影响会相对较高。当数据中心的用户群体较分散，或主要面向中小企业提供零售型服务时，功率密度则会相对较低。

3）随着互联网业务复杂度不断提高和需求量的快速扩张，互联网企业开始自建自运营数据中心，并自研适合业务定位的关键设备和系统，积累了众多技术创新成果，如整机柜服务器、微模块、HVDC、间接蒸发冷却、液冷等。

4）在高端计算领域，中科曙光基于浸没液冷技术和高密度刀片系统高密集成设计，已经将单机柜功率做到了惊人的 160Kw，同时又创新性地将浸没液冷计算系统与立体模块化组装设计相结合，将单位机房计算密度提高了 30 倍，形成了独具特色的算力中心方案，在合肥、兰州等十多个地方都实现了落地部署。



曙光相变液冷技术

经过分析，造成数据中心高密度发展趋势的原因主要有以下几个方面：

1）IT 硬件产品迭代。芯片是数据中心 IT 设备的重要基础组件，芯片的性能与功耗极大影响了数据中心的功率密度与运行处理效率。当前人工智能、物联网、超级计算及其相关应用对芯片提出了更高的性能要求，高算力已成为芯片的主要突破方向，为了满足高算力负载的需求，需要叠加多核处理器，或者提高单核心的主频，无论哪种方式，都会显著增加 IT 硬件的处理器功耗，从而使得数据中心功率密度越来越高。

2）承载业务的计算需求变化。近年来科学技术发展日新月异，计算密集型应用场景（例如 AI、IoT、区块链以及 AR/VR 等）的激增导致承载这些应用负载的服务器设备（虚拟机、刀片机、多节点服务器等）功耗也大幅增加，从而导致数据中心功率密度呈现逐年增大的趋势。

3）投资回报的统筹考量。部分城市和地区土地资源紧张、费用高昂，如何利用更小的空间、尽可能低的成本满足更多的业务需求是数据中心建设主体必须考虑的问题，在此情形下，数据中心的密度不得不被设计得越来越高。另外，数据中心运营成本中电费是最大的开支，提高功率密度可一定程度上提升配套设施的利用效率，降低PUE，节省电费开支。

当前“新基建”“东数西算”等政策正在持续促进数据中心市场规模扩大，并逐步引导数据中心产业发展格局完善。无论是从发展现状还是驱动因素来看，功率密度提升将成为未来数据中心一个重要演进趋势，并将引起建筑、供电、制冷等多个系统的变革。

## 六、算力智能调度：跨区域、跨云、云边调度

纵观我国算力经济发展情况，当前算力基础设施规模已位居世界前列，但人均算力尚低，亟需算力服务灵活地为算力资源供需者协调、匹配、调度算力资源，实现算力资源最大化的利用，其中算力智能调度是衡量算力服务水平的关键。

从国家层面来看，“十四五”规划明确指出，要加快构建全国一体化大数据中心体系，强化统筹算力智能调度，建设若干国家枢纽节点和大数据中心集群。工信部印发《新型

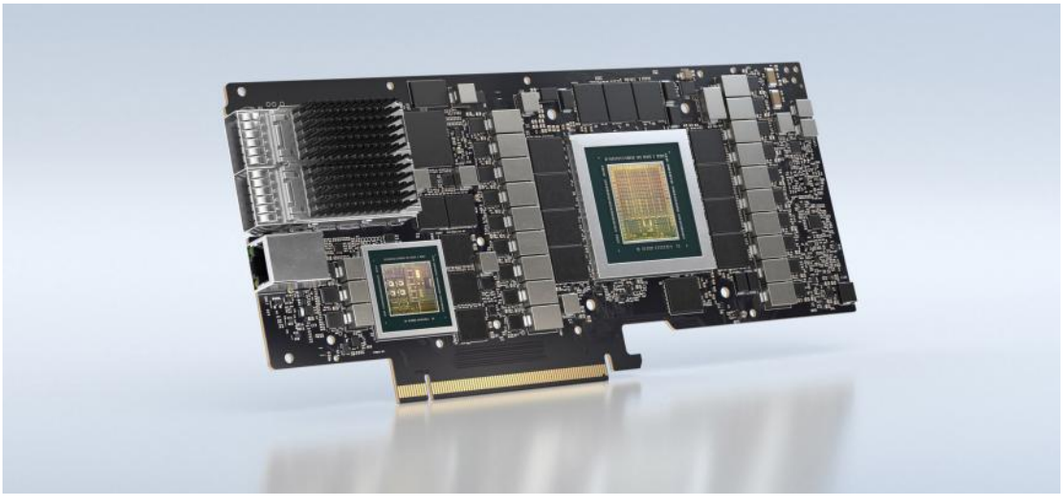
图片来源：《2021 中国云数据考察报告》

> 由于边缘计算节点所处位置一般较为分散且偏远，而运维中心一般集中在地市、云端，这就需要边缘云平台具备自动维护、自愈、修复等能力，保证在无人力值守的情况之下仍然能够具备智能化的健康监测、边缘自治能力。

数据中心发展三年行动计划（2021-2023 年）》，明确用 3 年时间，基本形成布局合理、技术先进、绿色低碳、算力规模与数字经济增长相适应的新型数据中心发展格局。在“东数西算”工程正式启动的背景下，算力智能调度的重要性凸显，云厂商、运营商等在该方向上进行了初步探索，未来在跨区域、跨云、云边等方面还存在诸多挑战。

一是算力跨区域调度与网络协同难度大，智能化程度不足。当前的算力资源提供商大多以地理区域为单位，部署服务管理平台，主要为特定地理区域范围内的用户提供算力服务。当涉及到跨区域的算力调度时，首先，需要保障各区域之间算力枢纽的协同联动，推动跨区域的算力资源与网络的供需对接，实现算力资源的敏捷、智能化的调度；其次，目前算力资源智能化调度模型整体调试、测试周期较长，对于资源请求响应较慢，业务调度效率较为低下；最后，由于我国不同地区之间网络基础建设程度层次不齐，存在算力枢纽节点之间网络薄弱的问题，这将有可能导致算力资源传输时间、响应时间过长等问题。

针对跨区域的算力资源调度面临的协调、管理难度较大的问题，需要建立起算力、算网的跨区域协同联动机制，在资源方面，算力资源提供方可以通过在调度的各区域之间建立起统一的算力资源管理平台的方式解决，平台需要建立算力资源与网络地址的映射机制，当算力资源的需求方需要跨区域的算力资源时，算力资源管理平台将解析出符合算力需求方要求的算力资源所在的地址，通过建立需求方与提供方的网络联接实现资源的智能化调度。在智能化调度模型方面，通过弹性可伸缩架构、低延迟轻量化设计、A/B 测试滚动发布、多模型加权评估等技术创新，优化智能调度模型，实



图片来源：《2021 中国云数据考察报告》

现计算资源的高效利用和快速部署。面对跨区域调度的网络传输问题，需要建立算网一体化协同调度能力，例如通过引入 AI、SRv6 等技术构建新一代承载网络，实现通过网络智能化感知业务需求、网络资源和算力资源；另外，运营商应当结合“东数西算”的背景，在算力枢纽节点之间强化网络建设，保障资源在算力枢纽之间的快速调用。

二是算力跨云调度面临不同云厂商和云形态两方面异构的问题，难以统一管理。随着业务发展带来的数据量的增加，用户对于计算资源的需求开始呈现多样化的趋势，单一的云环境逐渐难以满足多样的计算需求，跨云环境下的计算资源调度开始被广泛应用。当涉及到的算力资源属于不同提供商时，一方面，多个服务管理平台需要进行接口的打通对接，另一方面，也涉及到不同算力资源的安全性的认证保障的问题；当涉及到不同云形态的资源调度时，一方面，算力资源存在异构化、差异化的特点，导致资源的统一分配、调度、部署较为困难，另一方面，由于不同的云环境之间存在网络隔离，如何实现跨云组网，在不同的云服务商之间部署工作流，避免网络结构过于臃肿，请求无法敏捷快速响应，将是面临的又一挑战。

针对跨云调度面临的问题，目前产业内企业、第三方服务商等，开始建设大型多云管理平台，用以屏蔽底层异构资源的差异性，实现跨云资源的无差异调度。目前较为通用的多云管理技术架构能够支持多种云资源池的接入，实现对多云资源的统一纳管、认证和监控。多云管理技术能够实现对于虚机和容器的统一编排调度，提供无服务器模

式的业务访问能力，使用户不需要关注底层资源的调度、分配，主要关注业务流程的开发上。

三是算力云边调度面临节点统一管控难度大、边缘节点自治能力待提高的问题。在云边协同的背景之下，边缘计算节点能够将云计算中心的计算和存储能力下沉，屏蔽掉资源的异构化和地理位置差异，提供资源一致化的服务。但由于边缘计算节点较为分散，所处环境、网络、稳定性等存在不一致的情况，因此，如何将单个节点的能力与其他节点共同整合并与中心云联动，进行统一的管控调度是一个难题。另外，由于边缘计算节点所处位置一般较为分散且偏远，而运维中心一般集中在地市、云端，这就需要边缘云平台具备自动维护、自愈、修复等能力，保证在无人力值守的情况之下仍然能够具备智能化的健康监测、边缘自治能力。但由于当前边缘侧面对的场景大多呈现碎片化的特点，因此在网络问题、攻击问题等方面仍存在较大的治理压力，实现全方位智能化的平台保护、自治存在较大难度，边缘节点自治能力仍有待提高。

针对边缘计算资源分散，难以统一管理的问题，目前边缘计算节点多数采用 Kubernetes 多集群的方式，来实现多个边缘计算集群的协同管理和计算资源的管理。在分布式计算节点的健康监测方面，应建立健康能力检测的可视化平台，分布式健康监测节点，在边缘侧持续收集节点的故障信息，快速定位并及时报告。在边缘节点自治能力方面，面对碎片化的场景，需要进行分层的应对方案设计，包括设备层、网络层、数据层和应用层，提供针对性的解决方案。例如，以 SuperEdge 为代表的边缘容器方案提供的边缘自治能力，能够保障当边缘节点与云端网络连接不稳定或处于离线状态时，边缘节点仍可以自主工作，化解由于网络波动带来的不利影响。

> 针对多元算力领域当前的挑战，需共建多样性算力产业体系，打造多元产业生态、推动产业协同，才能为东数西算提供强有力的新型算力基础设施。

## 七、多元算力 多样计算

多元算力技术和服务不断成熟，产业生态有待完善。随着 5G、人工智能、云计算、大数据等新一代信息技术的广泛应用，行业应用的多样性日益丰富，现有主流算力难以满足多样化的场景需求。计算密集型应用需要计算平台执行逻辑复杂的调度任务，而数据密集型应用则需要高效的海量数据并发处理，算力需求和供给结构之间的矛盾逐步显露。单一计算架构和平台难以适应所有计算诉求，面向未来多样化的业务需求，多样性算力成为必然。

而多样性算力在形势、产业、技术、市场等方面，面临着如下挑战。

一是国产化计算芯片的性能及工艺仍需提高，自主研发刻不容缓。在 CPU 领域，部分国产 CPU（例如龙芯）已具备自主指令架构系统，同时可兼容已有的基础软件或平台，但仍面临应用生态的挑战，在国产化开源操作系统、自主编程语言和编程框架的推广普及等方面任重道远。

同样是国产 CPU 的海光处理器，因兼容 x86 指令集，具备成熟而丰富的应用生态环境，可支持云计算数据中心、大数据分析、边缘计算等多领域应用，满足互联网、电信、金融、交通、能源等行业广泛需求。

在 FPGA 领域，AMD 作为行业头部企业完成对赛灵思的合并后，将其 CPU 与赛灵思的 FPGA 结合为了 CPU+FPGA 的异构计算模式。而国产 FPGA 目前主要集中在中低端市场，仍需约三到五年达到 16nm/28nm 工艺水平，与国际先进工艺有 2-3 代的差距。

在 GPU 领域，部分国内 GPU 企业通过购买国外公司的 IP 授权，成功流片或量产，迈出了 GPU 国产化第一步。GPU IP 自研需要大量的时间与人力，芯动科技等企业采用外购 IP 加上自研设计的方式实现商业变现，极大降低研发周期和风险。

二是亟需构建高效、系统化的协调统一的异构算力系统。异构计算一般指在完成一个计算任务时，采用一种以上的硬件计算单元、互联协议、差异化架构、软件接口等。异构算力包括 CPU、GPU、DPU、FPGA 等，可提高算力和性能，同时降低功耗和成本，又具备多类型任务处理能力。但不同的硬件设备、协议、软件应用层接口等差异较大，异构算力在流程协同、芯片互连和软件适配方面均面临挑战，亟需构建高效、协调统一的异构算力系统，推动算力经济供给侧改革。

在技术设计流程的协同上，需保证不同厂商芯片的互联互通，可正常协同工作；在互联标准上，需统一各厂商芯片之间的互连标准。在软件层面上，不同厂商的计算芯片之间需搭建适配的 I/O、内存通道。

三是多元算力的供给形态不断丰富，需适配高性能计算、超算、智算等算力服务平台。多样性算力服务如何输出是算力服务的核心问题之一，当前算力服务的主要供给形态包括虚拟机、容器、API 等。

虚拟机方面，虚拟机是通用计算云服务的主要服务形态，但针对超算平台的虚拟化难度较大，难以使用 x86 虚拟机架构适配。虚拟化提供算力资源 是未来算力发展的固有趋势，但虚拟机的表现形态可能发生改变，适配异 构硬件。

容器方面，容器技术屏蔽了底层不同的硬件，实现微服务化调用，有益于 算力合理分配。容器技术生于云长于云，对于超算、智算等暂未完成适配，可能会出现算 力异构方面的问题，特别是对于超算平台。

API 方面，当前较少服务商使用 API 供给的方式提供服务， API 的方式提 供服务对算力服务使用者来说门槛较高，但也会获得更好的灵活性。

针对多元算力领域当前的挑战，需共建多样性算力产业体系，打造多元产 业生态、推动产业协同，才能为东数西算提供强有力的新型算力基础设施。

一是标准引领。建立多元算力标准与测评体系，推动通用算力、异构算力、智能算力等规范化建设与落地实践。

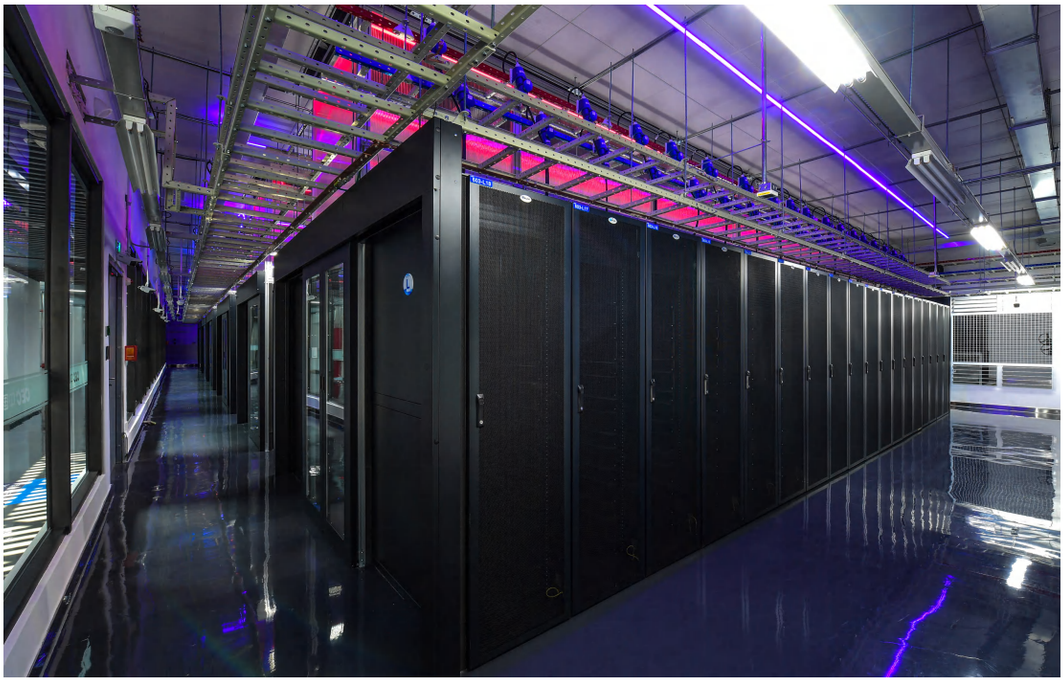
二是服务升级。提升各类社会算力的资源利用率，提升多样化算力效率，做到“物尽其用”，利用算力网络形成算力、网络、人工智能、区块链等多要素融合的一体化服务，推动算力服务全面升级和产业数字化转型，激发算力服务的范式创新。

三是共建生态。打造多样性算力应用与工具体系，在硬件架构、算法、软件等方面实现全链条自主可控，统一标准规范，共建共享丰富多彩的软硬件适配生态。

## 八、算力服务成为新业态

算力是信息时代的新生产力，是支撑数字经济发展的坚实基础，算力的发展推进了技术的升级换代、应用的创新发展、产业规模的不断壮大，而云计算作为算力的生产工具，推动了算力向服务化迈进。算力新基建热潮则进一步推动了算力服务商的发展，如中科曙光专门成立了曙光智算信息技术有限公司，以统筹各地算力中心的一体化运营和对外服务。 随着 5G、工业互联网、人工智能、云计算等新技术快速发展，GPU、FPGA、AI 芯片等异构算力迎来繁荣期，满足各场景计算需求的同时，从成本的角度大幅度降低计算门槛。算力的服务对象也将从企业和大客户扩展至小微企业及个人，真正推动算力成为触手可及的普惠资源。随着算力泛在性程度的不断加深，算力逐步从中心侧向边缘侧、端侧扩展，形成了云 - 边 - 端三级算力架构，以支撑智能安防、游戏、视频等多样化智能应用场景的不断落地。算力和网络的发展日益呈现一体共生之势，从网随算动到算网融合再到算网一体，网络从支持连接算力，演进为感知算力、承载算力，实现网在算中、算在网中。与此同时，算力度量、算力原生、算力路由等算网一体技术日益受到产学研各界重视，相关技术的成熟将持续推动算网融合向纵深发展。

一是计算服务化在架构方面具有灵活的扩展性。首先，算力服务将通用计算、智能计算、并行计算等多样性算力统一纳管和调度，屏蔽不同硬件架构差异，实现大规模异构计算资源的统一调度，适应不同需求，实现算力的普惠化。其次，计算服务化是将中心算力协同算力节点，构建算力管理多级架 构，提升算力的覆盖范围和调度能力，实



图片来源：《2021 中国云数据考察报告》

现算力的泛在化。再者，计算服务化通过整合算力资源，以统一的标准输出服务，避免算力应用被固定形式的算力需求所捆绑，实现算力的标准化。

二是计算服务技术上不断创新。网络控制与转发方面，算力网络等技术满足算力对网络感知和自智等需求；算力编排与度量方面，通过分布式操作系统、基础设施即代码（IaC）、算力建模形成标准可量化的算力单元，对泛在算力进行整体度量、编排与调度；算力运营与服务方面，云原生、开放应用模型（OAM）等技术对跨域和跨多样性算力进行整体封装，并以服务的形式对外提供，实现随需随用的算力应用“插座”。

三是计算服务化生态已成形。首先，以 IaC 为代表的新型算力管理技术快速发展，突破了超算、智算、云计算等异构算力难以标准化的技术瓶颈，新型算力管理服务商应运而生；其次，由企业数字化带来的云边协同、高性能等算力服务需求，云服务商开始提供高性能以及分布式算力等应用服务；再者，与人工智能、大数据、区块链等新一代信息技术融合的算力服务将展现雏形，实现数据分析、科学计算、工程计算等领域算力应用部署。

可以看到，算力服务作为一种新兴的业态，我国已取得初步进展，但目前算力服务发展仍存在几点不足：

第一，算力服务认知水平尚浅。我国算力基础设施规模已位居世界前列，但整体算力服务化水平尚低，在算力经济全面升级和产业数字化转型的背景下，对算力服务的定位、要求与实施路径重视程度不足，无法充分发挥算力服务对算力经济发展的支撑作用。

第二，算力服务化技术水平不足。一方面，算力网络、IaC、OAM、云边一体分布式操作系统等算力服务核心技术目前仍在初步发展阶段，尚不能满足产业上下游企业对算力服务的需求；另一方面，当前我国只实现了以云服务为代表的通用计算服务化，尚未建立对高性能计算、智能计算、并行计算等多样性算力服务的技术积累。

第三，算力服务化新业态发展缓慢。传统的一对一租赁的云服务提供模式不能满足算力消费方与算力提供方的需求，需要集中算力提供方的各类算力资源，统一出售给算力消费方，但算力供给 - 算力交易平台 - 算力需求的交易分配模式存在参与角色较多、供给关系复杂、缺乏产业组织引导等问题，导致算力服务新业态发展无法跟上快速增长的算力需求。

在算力服务的实践过程中，曙光公司积极参与东数西算工程与算力网络布局，目前实现了与 10 余家计算中心（先进计算中心、智能计算中心、一体化大数据中心）的极速互联与融合调度，并初步建设集算力、存储、数据和增值服务为一体的全国一体化算力服务平台，配备面向算力服务的应用支持、技术服务相关运营及运维团队，共助公共算力服务事业。

目前，曙光算力平台已服务国内外 10 万 + 用户，作业提交量突破 3000 万。

## 九、原生应用：云原生、AI 原生

数字宇宙时代，云智应用需求极剧攀升。中国信息通信研究院发布的《全球数字经济研究报告（2022 年）》显示，2021 年，全球 47 个主要国家数字经济增加值规模 38.1 万亿美元，同比名义增长 15.6%，占 GDP 比重为 45.0%。其中，中国数字经济规模达到 7.1

万亿美元，同比名义增长 16.2%，高于同期 GDP 名义增速 3.4 个百分点，占 GDP 比重达到 39.8%。“数字定义世界、软件定义未来”的时代正朝我们加速走来，工业互联网、车联网、智慧城市、智慧政府等以云计算为关键技术底座的应用场景在不断被挖掘，数据指数级爆发式增长，IDC 发布的《数据时代 2025》报告显示：未来一家数字化工厂一天可能产生超过 1PB 的数据；一辆联网的自动驾驶汽车每运行 8 小时将产生 4TB 的数据。在万物智联的数智时代，企业对数据挖掘和人工智能的需求极速提升，对应用的敏捷性要求持续攀升，以期在企业“竞存游戏”中，通过高效预测、高速分析、高频迭代，高质创新，提升企业的预见性和决策力，获取比较优势，以更快的速率感知市场、占领市场。而企业应用的云智化离不开强大算力的支撑，这对算力基础设施提出了新要求。

软件系统面临新挑战，应用呈现云原生化趋势。未来企业都将是软件企业，IDC 预测，到 2024 年数字经济的发展将孕育出超过 5 亿个新应用 / 服务，这与过去 40 年间出现的应用数量相当。到 2025 年 2/3 的企业将每天发布新版本软件产品。

这对应用软件提出了简化、标准化、自动化、智能化、敏捷性、稳定性、低成本、高效率等更为严苛的要求，而云原生以业务应用为中心，通过剥离软件中非业务逻辑的成分，实现聚焦价值、敏捷交付的目标，呈现出软件元素间关系的松耦合、结构的分布式、属性的高韧性的特征，赋予应用标准化封装部署、声明式描述 、持续集成持续交付、按需弹性的能力恰好符合应用进化的需求，故而应用呈现出云原生的技术倾向，Gartner 预测到 2025 年，云原生平台将成为 95% 以上新数字倡议的基础。

弹性按需是云原生的资源利用优势，但如果资源配置策略设置不合理可能会导致资源的浪费，同时如果云原生资源利用的计量方式不够灵活，会使得企业难以准确调控用云成本，造成能耗的浪费。

“在万物智联的数智时代，企业对数据挖掘和人工智能的需求极速提升，对应用的敏捷性要求持续攀升，以期在企业“竞存游戏”中，通过高效预测、高速分析、高频迭代，高质创新，提升企业的预见性和决策力，获取比较优势，以更快的速率感知市场、占领市场。

通过云原生技术的应用，对云资源规格、数量进行灵活调整，利用对业务的架构优化、以及通过弹性能力和资源混部等手段提升资源利用率。比如腾讯云推出碳排放优化器 Crane，就是首个开源云原生应用碳排放优化器。该优化器基于运行在 Kubernetes 平台上的应用的实际资源消耗，计算对应服务器功耗，进而计算出应用运行所产生的碳排放量。

为了最大化享受人工智能技术红利，AI 和云原生的联动也成为最佳选择，AI 开始步入面向业务应用的 AI 原生时代，运用松耦合、分布式的云原生特征，实现算法组件化，通过流水式编排开发降本增效，提供面向 AI 场景的弹性高性能异构算力，屏蔽底层资源异构性，提供低门槛的开放平台，最大化降低开放难度，加速 AI 能力的应用和落地。

应用进化带来新机遇，算力呈现算网原生趋势。应用的原生化演进对服务于应用数据的算力系统提出了新的挑战，算力新基建呈现出算网原生的发展趋势。 一方面，为全面适配应用的原生化技术倾向，需要屏蔽底层资源细节，将算力资源池化，全面整合底层基础设施的计算、网络存储、GPU 等资源，实现 GPU 的灵活调度，让应用用户可共享数据中心内所有服务器上的 GPU 算力，提升企业应用开发敏捷性。另一方面，为解决算力自身的发展需求。由于数据计算波峰波谷效应明显，所以传统算力架构资源浪费、弹性不足的问题凸显，而算网原生正是最佳解决方案。结合资源池化，将算力向水、电一样按需供给，即用即取。这需要算力系统满足两个要求，一是敏捷感应上层应用的工作负载，智能匹配最佳算力，二是现底层器件高效协同，充分释放算力潜能。算网原生可以通过 Kubernetes 的容器编排技术，实现各应用 GPU 需求的充分感应与最佳分配，通过提供多元异构算力服务，和大数据计算、深度学习计算、业务计算等场景深度融合，实现大规模的 GPU 集群的高效计算，有效降低应用开发、应用难度，缩短产品上线周期，加速敏捷迭代。

算力新基建的原生化，实现了算力资源与算力需求的最优匹配，同时微服务容器化松耦合实现了应用的安全隔离，以响应用户对流量的不同需求，将强大算力和云服务的安全性易用性相结合，其在智能制造、证券金融等领域均有良好应用。举例来说，证券公司提供的“智能投顾”，“智能投研”等人工智能预测服务，在传统架构下，业务并发量受限于集群内物理 GPU 的数量，业务伸缩能力受阻。而算网原生则通过资源池化实现了 GPU 资源的统一调度、纳管，支持不同代的算力卡混合池化，基于容器编

排技术实现跨域调度 CPU 和 GPU 资源的能力，充分满足业务高并发场景需求，系统弹性显著增强。

## 十、规模化和算网融合

国家发展改革委会同有关部门研究制定的《全国一体化大数据中心协同创新体系算力枢纽实施方案》指出：起步阶段，京津冀、长三角、粤港澳大湾区、成渝等跨区域的国家枢纽节点（“东数”区），原则上布局不超过 2 个集群；贵州、内蒙古、甘肃、宁夏等单一行政区域的国家枢纽节点（“西算”区），原则上布局 1 个集群。

以“充分发挥本区域的优势”为例，京津冀、长三角、粤港澳大湾区、 成渝列举的都是“市场、技术、人才、资金”，贵州、内蒙古、甘肃、 宁夏列举的都是“气候、能源、环境”；在发展数据中心集群的要求上，除了高能效、低碳、优化东西部间 互联网络和枢纽节点间直连网络这些共性，贵州、内蒙古、甘肃、 宁夏强调“高可靠”，京津冀、长三角、粤港澳大湾区、成渝则强 调“高密度”，还多一条“提升数据供给质量”。

结合相对的地理位置，在全国一体化大数据中心协同创新体系中：

- 京津冀、长三角、粤港澳大湾区、成渝 4 大枢纽是“东数”，自给而不自足，内部消化为主，对外转移部分需求（优化数据中心供给结构， 扩展算力增长空间，满足重大区域发展战略实施需要）；
- 贵州、内蒙古、甘肃、宁夏 4 大枢纽是“西算”，定位在供给方，主要承接转移过来的需求（积极承接全国范围需后台加工、离线 分析、存储备份等非实时算力需求，打造面向全国的非实时性算力保障基地）。

以粤港澳大湾区国家枢纽节点为例，根据《南方都市报》等媒体的 报道: 到 2025 年，广东省 70% 的数据中心在省内建设， 30% 的数 据中心通过“东数西算”向西部地区国家枢纽节点转移；韶关数据中心集群将建成 50 万标准机柜、500 万台服务器规模，投资超 500 亿元（不含服务器及软件）。

在“东数西算”工程的 8 个国家枢纽节点中，如果严格按照地理位 置来划分，总会由于成渝、贵州两个枢纽的特殊性，呈现“5+3”而非“东西各 4”的格局：

- 从东、西部的角度， 成渝和贵州都位于西部， 结果是“西 5 东 3”；
- 用“胡焕庸线”来切割，成渝和贵州都在东南， 就变成“西 3 东 5”……

1935 年提出的“胡焕庸线”( 见下图 ) 从黑龙江省的黑河 ( 瑷珲 ) 到云南省的腾冲划一条直线， 将中国地图一分为二： 右侧 ( 东南方向 ) 地势较低， 多平原和水网， 平均气温和人口密度较高；左侧 ( 西北方向 ) 地势较高，多草原和沙漠，平均气温和人口密度较低。


10 个国家数据中心集群 ( 起步区 ) 大致位置，“长三角”是长三角生态绿色一体化发展示范区的简写

图片来源：《2021 中国云数据考察报告》

这条斜线的划分也只是相对准确，譬如贵州位于胡焕庸线的东南侧， 反而是成都地区压在线上，成都市甚至还“越线”到了西北侧—— 在一些解读中，成渝枢纽也被划为既向甘肃枢纽和贵州枢纽转移数据，又承接长三角枢纽算力需求的“中间地带”。但在实际资源禀赋上，成都平原众所周知，贵州境内 90% 以上的面积是山地和丘陵；从人口分布和经济发展状况等方面来看， 相对偏西北的成渝枢纽属于“东 数”，位在其东南的贵州枢纽属于“西算”，确实是合理的。

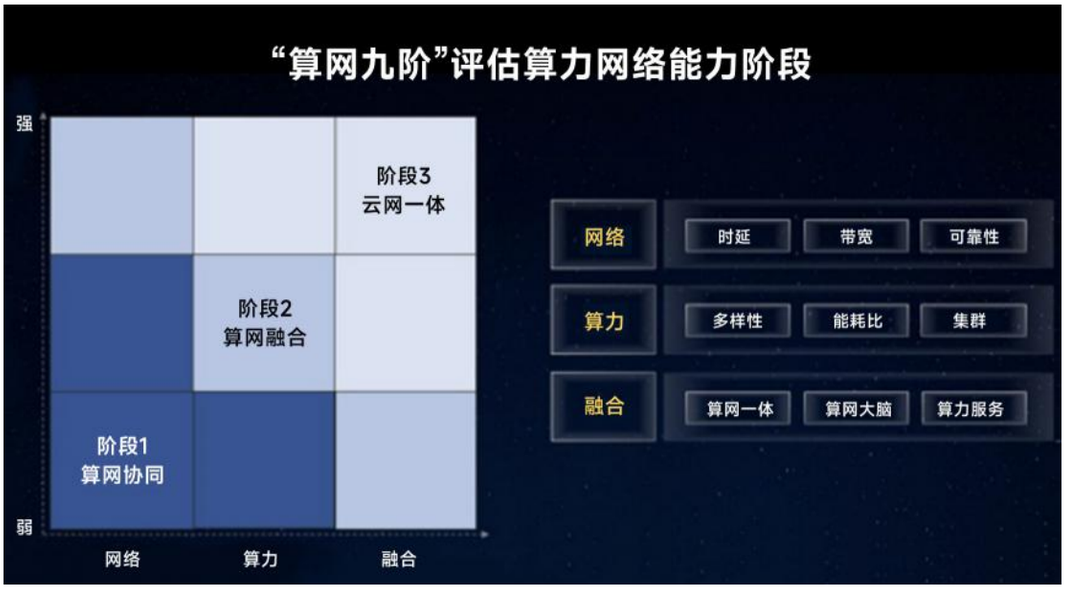
胡焕庸线的“相对”还在于，其划分方式在很多区域内部同样适用， 譬如京津冀、长三角、粤港澳大湾区的内部，也是东南部的经济发 展更好。像张家口数据中心集群之于

京津冀、韶关数据中心集群之 于粤港澳大湾区，基本都处在多山地的西北部。

“东数西算”的目的是将东部地区过于旺盛的算力和数据处理需求，转移一部分给更具成本效益、更可持续发展的西部地区承接，其中的关键是“国家枢纽节点之间进一步打通网络传输通道”“优化东西部间互联网络和枢纽节点间直连网络”，才能“提升跨区域算力调度水平”。

从国家层面来看，2020 年提出了算力基础设施这一概念，推动算力网络的发展，并根据技术演进和事件陆续推出了算力网络研究报告。电信运营商算力网络看成是 6G 的一部分，希望在云的连接上，加上计算一体化的服务场景实现业务的拓展，重视算力的感知，避免被管道化。

从 2020 年开始，中国电信和中国移动已经为“东数西算”工程调整了规划，分别推出了 “2+4+31”和“4+3+X”的全国数据中心布局， 其中“4”都对应京津冀、长三角、粤港澳（大湾区）、（陕） 成渝， “31”和“X”对应多个省级中心， “2”是内蒙信息园和贵州信息园， “3”是呼和浩特、哈尔滨、贵阳三大跨省中心。




算力网络的发展可以分为三个阶段：起步阶段（算网协同）、发展阶段（算网融合）和成熟阶段（云网一体），从目前来看，发展算力网络要解决几个矛盾：算力资源布局与需求之间的矛盾；算力效率水平与算力规模之间的矛盾；双碳目标与算力提升的矛盾；算力信息互通和调度与算力网络标准不完善的矛盾。

建设算力网络，需要加强技术投入和创新，在具体实践层面，超聚变建议用“算网九阶”模型来评估算力网络的能力阶段，设定算力、网络、融合三个维度，九大因子对算力网络发展的三个阶段能力进行综合评估，从而形成对算力网络的一致性标尺，助力各企业明确自身发展阶段，进行合理的规划与预测。

同时，东数西算作为前所未有的算网融合工程，东西横跨上千公里，接入全国各地的算力节点，对管理框架提出了新的挑战。各算力节点的建设周期不同步，算网架构需要逐步迭代、分级分区域演进；

- 东数西算集群间庞大的算网，对算力调度的计算和维护量巨大，计算和管理工作需要分层分解；
- 不同区域和集群间的云专网由不同厂商的设备组成，厂商之间接口不兼容，对算网统一调度的需求，需要大量的对接测试，需要兼顾不同厂商的设备能力。

要实现算力和网络的融合运营、智能编排、统一服务，也需要建立完善的评价体系和算网融合产业生态，目前看算力、网络厂商从芯片级、设备级、集群级、地域级 4 个维度寻找突破口，通过核心技术攻关解决从芯片到广域的 IO 不均衡问题，助推算力产业高质量发展。

在芯片级，目前“存算一体”的发展思路可以有效平衡计算和内存的配比，缩短数据搬运路径，降低搬运功耗，实现芯片级算力与 IO 的平衡，为算网融合构建算力基石。

在设备级，业界已经开始尝试多种总线互联和扩展技术，从内存、GPU、存储等多个角度入手，从互联设计角度出发，对资源进行分布式池化设计，从而平衡数据 IO 和计算密度。

在集群级，运用 AI 技术和网络设备的在网计算能力，可以实时收集并分析组网、设备、流量等综合信息，并通过强化学习对业务流量模型进行算力拓扑规划和动态调整，从而保障算力拓扑始终处于最优状态，满足大规模计算集群的部署需求。

最后在地域级，目前以新华三为代表的企业提出“确定性网络”确定性服务是广域算力互联及调度的关键点，在实际应用中具备诸多优势。在网络传输层面，传统的广域网传输是尽力而为的转发方式，通过引入确定性网络技术，可以保证网络层面全方位确定性传输，数据跨区域传输时延确定可控; 在算力调度层面，通过分布式算网大脑统筹考虑可用算力容量、成本、网络传输效率等多维属性，可以为客户提供确定性有保障的服务。

> 和交通和能源网络相比，算力网络要复杂多，网络体系结构的调整与演进，算网融合与算力调度体系的构建，需要做很多基础性的原始创新，还得做大量的技术攻关和事件，最终形成东西互补，南北贯通的一体化的算力网。

和交通和能源网络相比，算力网络要复杂多，网络体系结构的调整与演进，算网融合与算力调度体系的构建，需要做很多基础性的原始创新，还得做大量的技术攻关和事件，最终形成东西互补，南北贯通的一体化的算力网。

“东数西算”战略和“一体化大数据中心体系”把算力架构扩展至多数据中心的全局范围。这些为行业的信息系统体系结构的改进和发展提供了引领方向，比如气象业务是高度信息化和特性化的业务。在气象业务体系中存在一部分和其他行业类似的业务模式，但也具有非常鲜明的个性化的业务特性，与其他行业相比有明显的差异。例如处于核心地位的数值预报业务，需要极强的高性能算力支持，在常规行业中很难找到可复用的需求和解决方案；而气象数据在种类、使用方式的高度复杂性、应用时效等方面的个性化特征，以及其体积的巨大，使得气象大数据与气象高性能计算资源之间至今无法实现物理空间上的远距离分离。“东数西算”的资源布局趋势，对气象业务信息系统在设计实现和发展演进中提出了巨大的挑战。

“算力体系结构”在气象行业的应用中具备一定的特殊性。首先，数值预报对算力资源有着特殊的需求，主要体现在大规模高密度浮点运算能力和计算节点间高性能紧耦合通信能力等方面，因此高性能计算支撑能力必不可少。第二，气象数据资料类型复杂、种类多样、数据收集、处理、存储和应用等各个环节的数据量巨大、时效要求高，从而导致气象数据资源与高性能计算资源之间的高速、高效、高可靠性等的个性化需求。第三，各类气象业务应用，主体上具有高强度数据 IO 密集型的特性，对存储和通信资源及其支撑能力要求较为苛刻。

高性能计算、常规计算和数据分析处理在资源和应用方式等层面虽然存在较大差异，但业务应用的流程是需要总体贯通的，在控制调度上必须以“一体化”的视角将三者紧密衔接。

因此，“东数西算”背景下气象部门的“算力体系结构”，主要包含“超级计算能力”，“常规通用计算能力”，“超级数据处理能力”，“超级通信传输能力”这四个部分，可简称之为“超常算数通”。其核心思想是以“超级计算能力”支撑数值预报等核心气象业务，以“常规通用计算能力”支撑气象各单位常规型业务应用，以“超级数据处理能力”支撑大规模数据处理和存储以及数值预报周边的所有辅助型业务，以“超级通信传输能力”实现“东数西算”中数据在东西数据中心节点间以及数据中心内部稳定高速流动传输。

# CHAPTER3
# 展望
# 面向 2030 年的算力基础设施

## 数字文明时代加速到来，要求算力基础设施资源充沛、泛在普惠

当今世界正经历百年未有之大变局，互联网、大数据、云计算、人工智能、区块链等技术加速创新，日益融入经济社会发展各领域全过程，以信息技术和数据作为关键要素的数字经济成为全球新一轮科技革命和产业变革的重要引擎。习近平总书记在 2021 年在致世界互联网大会乌镇峰会的贺信中指出，“要激发数字经济活力，增强数字政府效能，优化数字社会环境，构建数字合作格局，筑牢数字安全屏障，让数字文明造福各国人民。”近几年，人类向数字文明的过渡大幅增进，特别是远程医疗、在线教育、共享平台、协同办公、网络直播、以 NFT 为代表的数字资产等数字化新事物、新业态、新模式，推动各个领域加快迈向数字文明时代。“万物互联”“万物智联”“人人享联”成为数字文明的基本特征。

一是“万物互联”对算力设施供给总量提出了更高要求。近年来，我国数字经济发展迅猛，带动数据量年均增速超过 50%，我国已成为全球数据资源规模最大、增长最快的数据圈，预计到 2025 年数据总量将跃居世界第一，全球占比有望达到 27% 以上。算力成为影响数字经济发展的核心要素。IDC 发布的《2021-2022 全球计算力指数评估报告》显示，计算力指数平均每提高 1 点，数字经济和 GDP 将分别增长 3.5‰和 1.8‰。信通院发布的《中国算力发展指数研究报告》指出，2016-2020 年期间，我国算力规模平均每年增长 42%，数字经济规模增长 16%，在算力中每投入 1 元，将带动 3-4 元经济产出。据测算，到 2025 年我国数字经济规模有望突破 80 万亿，2030 年破百万亿，数字经济快速增长、万物皆可“云”的时代要求算力供给资源充沛。

**3.5‰**
**1.8‰**

计算力指数平均每提高 1 点，数字经济和 GDP 将分别增长 3.5‰和 1.8‰。

IDC《2021-2022 全球计算力指数评估报告》

**42%**
**16%**

2016-2020 年期间，我国算力规模平均每年增长 42%，数字经济规模增长 16%

**80 万亿**
**100 万亿**

到 2025 年我国数字经济规模有望突破 80 万亿，2030 年破百万亿

信通院《中国算力发展指数研究报告》

二是“万物智联”让智算中心成为算力设施“主力军”。未来社会 80% 的应用场景都是基于人工智能，无论是智慧城市还是智能制造、无人驾驶、数字孪生等场景，除了要有数据支撑以外，还要和各领域、各场景的知识模型、机理模型甚至物理模型相叠加，形成基于人工智能的新应用和场景实现。复杂模型、复杂场景势必需要面向 AI 的算力基础设施，智算中心将算法、模型、算力三者有机融合起来，向外界、向园区、向企业、向政府输出 AI 的数据库、AI 的模型、AI 的开放平台等多种 AI 产品，让人工智能应用透明化，进一步让算力基建化、算法基建化。未来，人工智能计算需求将占据 80% 以上的算力资源，主要由智算中心承载。

三是“人人享联”要求算力设施更加泛在普惠。数字文明时代，随着人类更加便捷地进入虚拟空间，信息网络空间将从以“物”为核心的“赛博空间”向以“人”为中心的“智能化数字空间”转变。一方面，要求算力泛在化。根据第 49 次《中国互联网络发展状况统计报告》显示，截至 2021 年 12 月，中国网民人均每周上网时长达到 28.5 个小时，平均每天上网超过 4 个小时。随着数字时代的发展，人们在“智能化数字空间”中的工作学习生活时间进一步变长，随地接入、随时访问要求算力无处不在。另一方面，要求算力普惠化。数字文明时代要求把发展数字经济的出发点和落脚点聚焦到人民对美好生活的向往上，推进数字城乡区域融合发展，通过算力调度和补偿机制，让广大人民群众共享算力建设成果。

## 隐私计算为代表的技术为组织间数据流通提供解决方案

在提升组织内部数据管理能力的基础之上促进各主体间的数据的有序流通是释放数据价值的关键阶段，数据会在此阶段通过开放、共享和交易等方式实现真正意义上的价值释放。

以隐私计算为代表的价值流通技术体系为数据流通提供核心动能。隐私计算是指在保证数据提供方不泄露原始数据的前提下，对数据进行分析计算的一系列信息技术，保障数据在流通与融合过程中的“可用不可见”。从技术角度出发，目前主流的隐私计算技术可分为三大类：第一类是以多方安全计算为代表的基于密码学的隐私计算技术；第二类是以联邦学习为代表的人工智能与隐私保护技术融合衍生的技术；第三类是以可信执行环境为代表的基于可信硬件的隐私计算技术。隐私计算的应用主要集中在金


融、互联网和拥有大量数据源和数据流通需求的医疗、政务领域，应用场景则聚焦在联合风控、营销、智慧医疗和电子政务。联合风控方面，以银行机构为例，隐私计算技术可以实现跨机构间数据价值的联合挖掘，更好地进行客户检测，降低欺诈及合规风险，从而综合提升风控能力。2021 年，招商银行深圳分行发布行内首个运用联邦学习技术的贷款产品，产品对接了深圳市公共信用中心的系统和模型，在无需各方披露底层数据的前提下实现联合运算并进行风险控制。联合营销方面，隐私计算技术可以帮助机构在不输出原始数据的基础上共享各自的用户数据，提升建模结果准确度并根据模型制订营销策略，提高营销转化率。2020 年，建设银行联合子公司建信基金成功部署了行内首套联邦学习环境，通过“速盈客群价值提升场景”，实现跨双方模型训练，定位目标客群，助力客户价值提升，在实际的营销中针对目标客群的响应率相对提升 34%。智慧医疗方面，隐私计算技术可以实现医疗数据在“可用不可见”的前提下在机构间共享，推动各医疗机构的临床研究，解决医疗系统数据孤岛难题。如欧盟通过“机器学习分类账编排药物发现”项目构建了 Owkin Connect 平台，让各医疗机构利用制药企业的脱敏数据创建更有效的模型。在中国，多地医疗机构也开始运用隐私计算技术解决临床研究的问题。厦门市基于隐私计算的健康医疗大数据应用开放平台已完全落地，做到了数据“可用不可见”“可算不可识”；上海长宁区搭建了健康大数据开发平台；福建省也在筹备打造基于隐私计算的防疫平台和智能系统。电子政务方面，隐私计算技术典型应用实践就是电信反诈骗，如中国移动基于 CMBaaS 平台建立针对电信反欺诈识别的联邦模型，将运营商的用户静默等级、交际圈稳定程度、运营商业务

量变化情况等数据、公安的诈骗号码库进行联合建模，实现电信欺诈联合预测，降低财产损失。

此外，以区块链、数字签名、数字水印、数据指纹为代表的追溯审计技术体系为数据流通提供了重要保障；以数据脱敏、数据失真、态势感知等为代表的数据安全与信息保护技术体系为数据流通各环节保驾护航，与隐私计算技术体系共同为数据流通提供价值流通、追溯审计、数据安全与信息保护等方面的相关解决方案。

## 可信隐私计算是未来数据要素化的理想技术方案之一

可信隐私计算在应用过程中，其安全性、可用性和隐私保护能力等应符合设计声明预期，以满足数据需求方、数据提供方和监管方等各方的需求，一般包含安全可证、隐私保护、流程可控、高效稳定、开放普适等基本特征。图 2 给出了可信隐私计算的总体框架。


可信隐私计算总体框架

支撑技术层面，围绕着安全可证、隐私保护、流程可控、高效稳定、开放普适等可信的基本特征，以理论研究为抓手，弥补当前技术的不足，缩小应用的差距。例如研究能抵抗恶意攻击、合谋攻击的安全保护技术、研究保证精度损失可接受条件下性能有效提升的技术方法、研究保证计算全流程可审计的技术方法等，都需要学术界和工业界的积极探索。

企业实践层面，隐私计算从概念验证到应用落地依赖于企业将技术产品化。因此，企业在可信隐私计算的应用实践是可信方法中至关重要的环节。同时，应该注意到没有完美的技术，关键在于如何正确的使用技术，需要在产品研发使用的全生命周期过程中贯彻可信特征的要求，从产品源头保证“可信”。

行业组织层面，可信隐私计算需要整个行业的参与，包括可信隐私计算标准体系的建设、可信隐私计算评估测试等，通过可度量可验证的方式来减轻隐私计算技术和系统应用带来的风险。

### 第一要素：安全可证

安全可证是可信隐私计算的第一要素。隐私计算通过只输出中间参数、标签等信息，或通过在可信受控环境中对数据进行处理的方式，保障了数据的安全性，提高了数据流通的主动性。但隐私计算的安全性自证是技术应用过程中面临的难题，隐私计算产品安全边界的界定需要考虑不同行业、不同场景和不同技术的差别，也需要平衡计算准确性和计算效率的要求。因此，如何评价和验证系统的安全性亟需明确。

### 核心要素：隐私保护

隐私计算的核心目标是要保护隐私。个人隐私信息如个人身份标识、属性行为、位置轨迹等一旦泄露、非法提供或滥用将会危害个人或组织的相关权益。可信隐私计算通过技术手段对数据隐私进行保护，并将进一步保障数据使用可控，有效防止了数据的盗用、滥用和误用。可信隐私计算要对全周期隐私信息有保护。数据在不同参与方实体之间流转时，应采用隐私计算等技术措施，增强个人对处理者的信任度，应履行采取相应的隐私保护技术措施的义务，防止未经授权的个人信息泄露、篡改和丢失。

### 信任基础：流程可控

隐私计算虽有不同的技术路线，但是由于涉及多个参与方、普遍依赖密码学方法进行计算，所以数据使用的可控可计量、计算流程的可监控、全流程的可审计等至关重要，这些也是用户信赖隐私计算产品的基础。

### 落地抓手：高效稳定

除了具备安全可证、隐私保护、流程可控，想要实现隐私计算系统的真实可用和场景落地，高效稳定是可信隐私计算应用的重要抓手。

### 规模化前提：开放普适

日渐增加的隐私计算产品在丰富市场选择的同时也带来了新的需求，一是技术实现方法的多样化使得不同技术平台所托管的数据无法跨平台交互，可能造成“计算孤岛”现象，由此市场对平台的开放扩展兼容能力、互联互通能力提出了新的要求；二是系统操作简便、容易部署、容易运维也是实现各行业场景落地、规模化应用必不可少的前提条件。

## Web3.0 驱动规模化、泛在化的智能算力构建

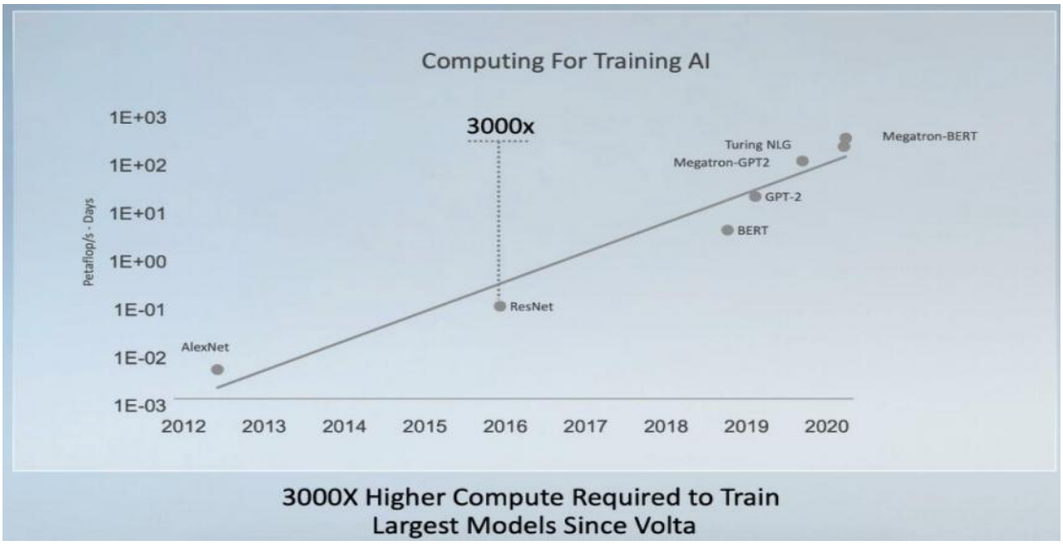
Web3.0 时代加速到来，元宇宙将成为主要形态和入口。一方面，随着非同质化通证、隐私增强技术等新一代关键技术的发展，一个以用户为中心，强调用户拥有身份、数据、算法、收益、协作等方面的自主权，打破了中心化模式下互联网平台对数据、交互天然垄断的下一代互联网发展形态 Web3.0 时代正加速到来，这将有利于释放数据要素价值，催生数字经济发展新模式。另一方面，元宇宙作为数字世界和现实世界融合的载体，既有现实世界的数字化复制物，也有虚拟世界的创造物，是 Web3.0 的特定应用形态。Web3.0 在沉浸式 AR/VR 终端、触觉手套等先进设备，以及动作捕捉、空间感知、数字孪生等相关技术的加持下，将为用户提供前所未有的交互性、高度的真实性以及深度的沉浸感和参与感。依托元宇宙这一关键入口，Web3.0 将“飞入寻常百姓家”。

规模化智能算力是 Web3.0 时代构建元宇宙不可或缺的基础。一是元宇宙将产生人类历史上前所未有的对规模化智能算力和计算资源的巨大需求，如为达到理想的沉浸式体验效果，VR/AR 需要实现更高的视网膜分辨率、更广的观察视角、更精准的位置感应（如 3D 音频动态跟踪、手势位置动态追踪、眼球追踪等），而这是传统数据中心和云服务商难以满足的。例如，即便是最初搭建在 AWS 上的元宇宙游戏 Roblox，也因为网络容量不足而不得不建立自己的数据中心[1]。二是算力不足带来的延迟、覆盖率低下等问题

1. 金色财经 . CCN 测试网 2.0 启动在即：CCN 将为 Web3 和元宇宙的建构提供算力基础 [EB/OL]. [2022.8.4]. https://mp.weixin.qq.com/s/5LdlP-gofPlsw7vLXDnTgA.

给实时性、沉浸式的元宇宙体验造成困扰。不同于传输环节的丢帧，因算力不足等原因产生的弃帧将使元宇宙用户感受画面卡顿、跳跃与拖尾，极大破坏沉浸式体验。三是 Web3.0 通过聚合来自世界各地的计算资源可提供规模化智能算力服务，夯实元宇宙发展基础，以此确保了 Web3.0 时代处理海量数据处理需求的“游刃有余”“运转自如”。

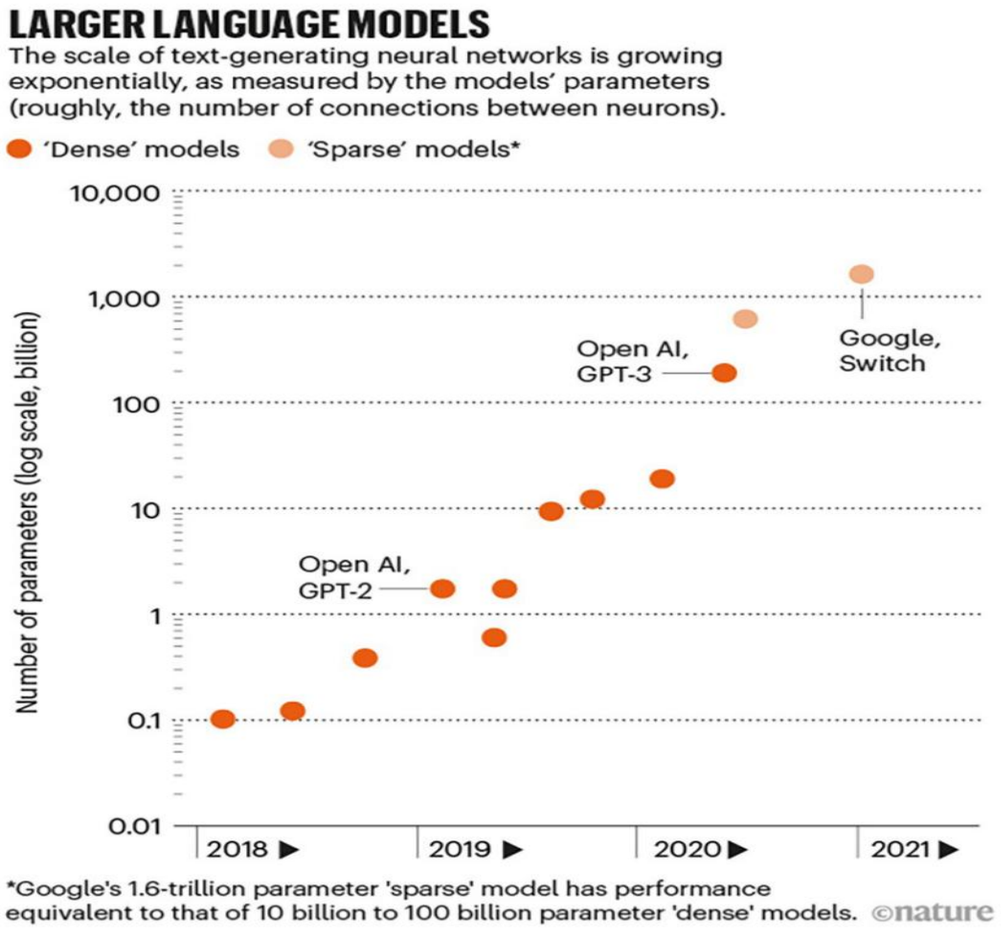
从“单点突破”迈向“泛在智能”，泛在智能算力是 Web3 应用场景落地的敲门砖。一是面向虚拟世界中更真实体验的机器学习、计算机视觉、自然语义处理等训练模型架构设计上趋向大规模并行，数据量已达千 G 量级，参数量迈向万亿级[2]。AR/VR 云游戏、元宇宙等场景对数据传输处理速度和快速分析、推理、决策能力也提出了更高要求。二是通过完善在云、边缘、现场终端不同层级的泛在智能算力体系，有助于实现更快、更低时延、更低成本的算力输出。2020 年中国总算力规模中智能算力占比达到 41%，预计到 2023 年智能算力的占比将提升至 70%[3]。据 Intel 测算，到 2030 年，每人拥有 1Petaflops 的算力和 1PB 的数据，时延不到 1 毫秒，元宇宙得到充分发展[4]。三是泛在智能算力牵引效率和创新能力加速突破。构建多元化、规模化、泛在化的智能算力，有效拓宽应用场景的边界，促使元宇宙从“单点突破”迈向“泛在智能”，助推随地、随需、随形的多元场景应用落地，引领 Web3.0 未来发展潮流。




2. 中国移动 . 算力网络技术研究报告（2022）

3. 中国算力发展指数研究报告》. 中国信通院

4. Web3.0 时代，得算力者得未来 . 雨晴，陆玖财经 .[EB/OL]. https://mp.weixin.qq.com/s/xZLxYEQ6_PlVa9uyD4ImLA.-2022.8.4.



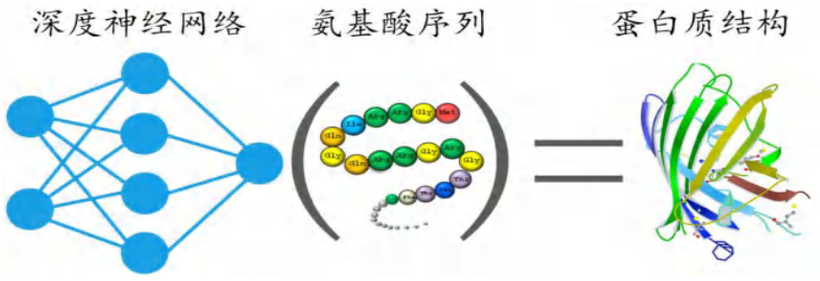

## 第五范式 AI for Science 对算力的需求

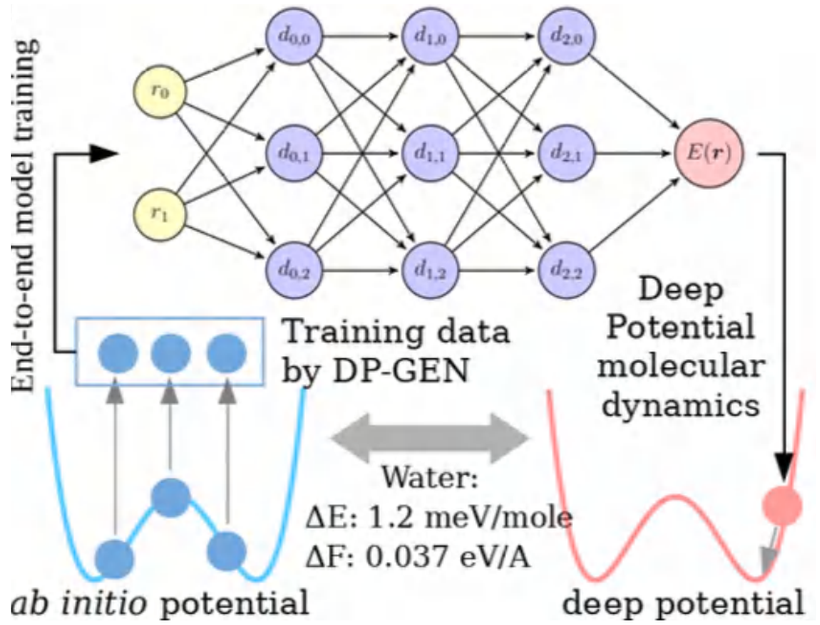
2020 年 5 月，OpenAI 发布了当时全球规模最大的预训练语言模型 GPT-3，具有 1750 亿参数，数据量达到 45TB，训练费用超过 1200 万美元。GPT3 的出现意味着 AI 对算力的需求进入新阶段！谷歌在 2021 年初推出超级语言模型 Switch Transformer，将参数量提升至万亿级别。据预测，GPT4 将至少有 2.5 万亿参数，比人类写得更好，当对答案不确定时，还可以进行研究。MIT 预测，训练 GPT-4 预计会花费 26 亿美元，2032 年才能降低到 5 百万美元。

当计算改变了科学，催生了数据密集型科学发现的第四范式。机器学习是第四范式中日益重要的组成部分，能对大规模实验科学数据进行建模和分析。

现在，深度学习的一个新用途是兼顾科学发现的速度与准确性的强大工具。这种使用机器学习的新方式与第四范式数据建模截然不同，因为用于训练神经网络的数据来自科学基本方程的数值解。

在精确程度和计算速度两难困局下，“AI+Science”被称之为第五范式，带来了两种范式结合的可能性：将 AI 的复杂数据处理能力和 Science 的第一性原理的结合，是两种传统研究范式的结合。以机器学习为代表的 AI 具备复杂数据处理的能力，可以使得训练出来的物理模型兼具微观尺度的精度和宏观尺度的效率，能够真正有效解决科学计算中的“维度灾难”问题：



- 处理海量数据所面对的维度灾难斩获 2020 年戈登贝尔奖的 Deep Potential 方法，展示了 AI 和分子动力学模型的有效结合；在保证精度的同时，指数级地提升了物理模型的效率；
- 复杂场景中求解物理模型所面对的维度灾难：系统性地解决药物设计、材料设计和化工设计等领域中的微观设计层面问题，实现「既快又准」的计算模拟；在宏观的飞机、汽车、火箭设计领域也将有丰富的应用。

采用深度学习等 AI 方法来处理数据，最成功的例子当属 AlphaFold2 。蛋白质结构预测问题是一个典型的高维问题，AlphaFold2 直接将蛋白质一级序列和三维结构通过一个精妙的深度神经网络关联了起来，这就像是 DeepMind 找到了一个优美的数学公式，可以将蛋白质的序列和结构用等号连接起来，AlphaFold2 彻底改变了蛋白质结构解析的技术路线。

AI for Science 的数据来自各个学科的数据积累，模型来自各领域科学家发现的科学原理和规律；算法源自机器学习算法和数值方法等方面的创新；需要多样算力融合的综合型智能计算平台，通过分布式异构并行体系结构，实现多样算力的融合、优势互补，为 AI 训练、AI 推理、数值模拟等不同应用提供不同算力，实现高精度到低精度算力的全覆盖、多种计算类型的全覆盖，以及 AI 训练 + 推理全覆盖。

## 大模型成为人工智能工程化重要方向，智能算力需求几何级增长

一方面，人工智能大模型成为世界性趋势。人工智能落地面临长尾场景应用的“碎片化”和应用开发的“高门槛”等挑战。为了增强 AI 通用性、加速 AI 工程化，“超大规模预训练模型”成为世界性趋势。自 2011 年以来，全球人工智能模型参数急剧增长，已突破千亿级。2019 年谷歌推出的 BERT 大模型拥有 3.4 亿个参数，使用了 64 个 TPU。2020 年，OpenAI 推出的 GPT-3 深度学习大模型拥有 1750 亿参数，是当时全球最大的 AI 巨量模型。2021 年浪潮发布的“源 1.0”参数升至 2457 亿，是当前全球最大规模的中文 AI 巨量模型。同年，微软和英伟达使用了 4480 个 GPU 训练出的拥有 5300 亿参数的 MT-NLG 大模型。通过构建大模型提升人工智能处理性能，已成为未来模型发展的重要趋势。


另一方面，模型超大规模化促进智能算力网络发展。据 OpenAI 统计，自 2012 年以来，业界最复杂的 AI 训练任务所需算力每 3.43 个月就会翻倍。AI 大模型对算力的需求远远超过了芯片产业长期存在的摩尔定律（每 18—24 个月芯片的性能会翻一倍）。当 AI 大模型成为推动 AI 能力提升的重要工具和手段，其非线性甚至几何式高速增长的参数数量，导致 AI 大模型、巨量模型的计算规模越来越大，需要的硬件资源（内存、GPU）越来越多，对算力的需求极其巨大，一般的算力基础设施很快将难以胜任。建立以 AI 芯片为主的高效率、低成本、大规模的智能算力基础设施将成为训练 AI 大模型的前提。为了提供相匹配的超大规模的算力支撑，亟需构建云化的智能算力网络，通过在区域内感知、分配、调度人工智能算力，根据各中心算力资源的情况和各地区的需求情况进行算力动态调配。

“当 AI 大模型成为推动 AI 能力提升的重要工具和手段，其非线性甚至几何式高速增长的参数数量，导致 AI 大模型、巨量模型的计算规模越来越大，需要的硬件资源（内存、GPU）越来越多，对算力的需求极其巨大，一般的算力基础设施很快将难以胜任。

## 边缘创新与新兴应用

云化 5GC 时代到来，边缘计算迎来爆发式增长。

在行业垂直领域，用户的信息数据使用需求，往往先通过边缘点、边缘云来收集；然后再通过大批量的复杂计算，传输到中心云进行处理，从而完成边缘云与中心云的联动协同。

而 5G 网络所具备的种种优势，能大幅提升传输数据、响应需求的速度；同时，也能进一步保障用户存储数据的安全性。因此，5GC 可以使得整个边缘云和中心云之间的联动过程比以往更加顺畅高效，也可以更好地满足不同用户的多样化边缘场景业务需求。

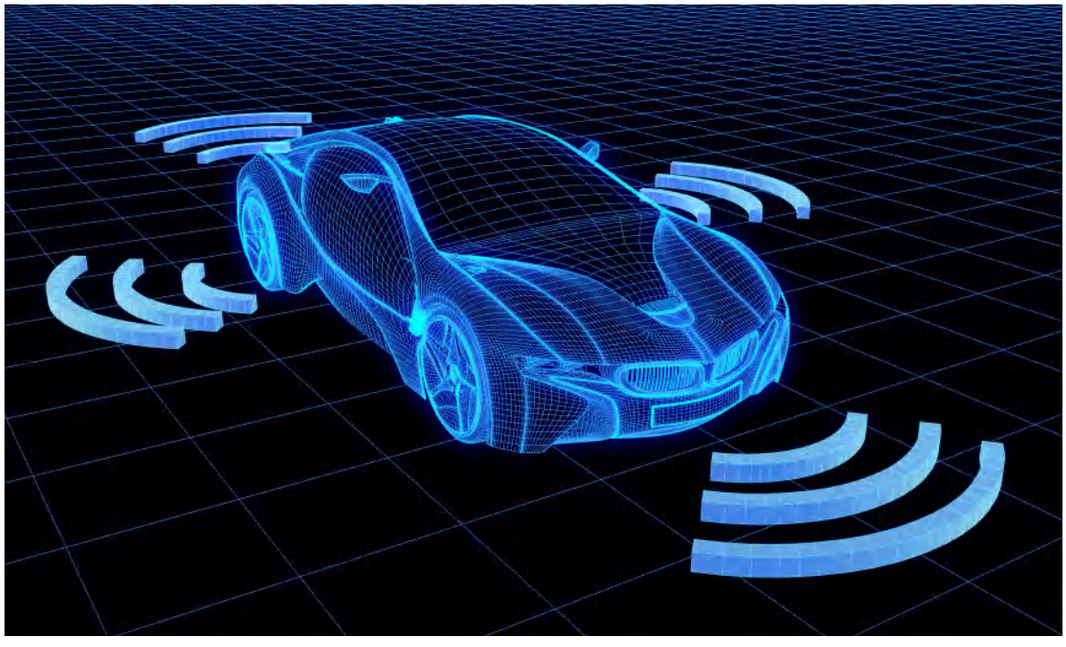
云边端部署模式将支撑未来新兴应用发展。在算力方面呈现海量计算需求和快速实时响应两个方面的特点。因此，传统计算技术及计算架构需要进行变革，云边端相结合的模式是算力体系支撑的新兴应用的重要方式，需要各类计算处理能力融于一体。在此基础上，需要大型数据中心承载集中算力，需要边缘数据中心承载实时算力。

在产业横向拓展中，云游戏、自动驾驶、VR/AR、物联网及工业互联网新型应用对网络时延、数据安全提出了更高要求。以云游戏为例，其背后所代表的是终端算力云化的大趋势，Arm 架构等移动终端等通过与云计算的结合，突破终端算力性能、功耗的瓶颈，进一步扩展高算力覆盖场景与需求，而在云计算的角度，端云的融合也产生了对通用计算、渲染、视频编解码、网络处理等算力的新组合与应用，引入新的架构，多元发展。技术层面，如何更好的调度、融合不同的计算资源，需要芯片、操

作系统、虚拟化等技术协同，组合新的参考架构助力行业发展。生态 / 应用层面，以 Android、鸿蒙、国产 Linux 为代表的操作系统也迎来了新的发展机会，移动 /PC/ 云操作系统产生融合。而业务层面，端云硬件组合的变化，也打破了不同厂商业务边界，终端厂商布局云端，而云厂商也可以触达终端用户。产业需要芯片、方案、云、业务等分属不同生态位的厂商共同配合、摸索新的技术组合。比如工业制造、医疗健康等注重精细和实时性要求的场景对 AR 应用需求的不断增加，需要平台部署更有力的软硬件，来提供高强度的音视频编解码和渲染能力，从而避免出现抖动、扭曲和画面丢失问题。 同时 AR 应用在成为业务效率倍增器的同时，也需要更强的实时 AI 算力，因此，图像识别、边缘检测等计算机视觉 AI 算法正被引入 AR 应用中，当然也对 AR 平台的算力提出了更高要求，需要实施 AI 加速优化，以避免使用时出现数据不同步、卡顿、超时等问题。因此，未来基于 MEC 技术的 AR 可视化方案将会在更多场景中得到应用。

## 自动驾驶进入无人化新阶段，云边端高效协同

展望 2030 年，按照我国自动驾驶产业的顶层设计和路线图规划，高度自动驾驶汽车



将从 2025 年的限定区域和特定场景商业化应用，向高度自动驾驶汽车实现规模化应用转变。部分和有条件自动驾驶级智能网联汽车市场份额超过 70%，高度自动驾驶级智能网联汽车市场份额将达到 20%，并在高速公路广泛应用、在部分城市道路规模化应用 [5]。自动驾驶产业是技术密集型产业，其发展离不开技术的创新与突破。这其中涉及到复杂环境融合感知、智能网联决策与控制、车载智能计算平台、高精度动态地图与定位、安全测试与验证等关键核心技术，而这些技术的研发与应用需要新型算力基础设施的有力支撑。我国应加快推动自动驾驶产业与“东数西算”战略布局融合，以新型算力基础设施支撑自动驾驶各环节技术加速创新应用，构建“车 - 路 - 云”一体化高效协同体系，助力我国智能网联汽车产业快速发展。

自动驾驶各环节技术创新应用需要新型算力的强力支撑。对于自动驾驶而言，算力不足是最为关键的发展瓶颈。从 L1 到 L5，每增加一个级别，计算的复杂性就会增加一个数量级，对算力的需求将会成 100 倍的增长。在自动驾驶智能模型构建方面，随着自动驾驶场景的日益丰富，自动驾驶训练模型也日益复杂化和规模化，需要足够的算力对算法与模型的开发应用提供支撑。在“东数西算”战略下，利用西部地区数据中心目前较高的能源使用效率，在西部建立更加先进的自动驾驶智算中心，实施自动驾驶的“东数西训”，在西部地区进行更大规模智能模型训练与学习，在自动驾驶的实际应用场地进行算法与模型应用。在自动驾驶精准感知与识别技术方面，感知与识别能力需要基于 AI 芯片的智能驾驶座舱系统和车载智能计算平台提供强大的海量数据计算、高精度、低延迟（毫秒级）的计算能力，对各种驾驶与交通环境类目标进行实时、动态、精准捕获与识别。同时，对车内驾驶员状态、语音和动作等进行实时监测与分析，为自动驾驶决策与控制提供判断依据。在自动驾驶本质安全方面，智能网联汽车预期功能安全需要经过测试、评估和验证等多个阶段，这需要具备具有强大算力的智能化虚拟仿真和测试验证平台的有力支撑，从而提升自动驾驶汽车性能评价与检测认证能力，从本质上确保自动驾驶的安全。

5.《智能网联汽车技术路线图 2.0》

“车 - 路 - 云”高效协同需要 AI 和云边协同新算力的支撑。自动驾驶的未来是智能网联，智能网联的核心是构建一体高效协同的“车 - 路 - 云”高效协同体系。而支撑“车 - 路 - 云”高效协同运行的关键是对多源异构和多层级数据资源的融合和计算能力。因此，建设海量异构数据融合与计算处理平台支撑自动驾驶上层应用场景的构建就至关重要。自动驾驶的高精动态地图就是“车 - 路 - 云”协同体系下的一种典型应用。高精地图融合应用 AI 和边缘计算等技术，利用道路侧智能设备与移动车提高数据采集和更新效率，提升自动驾驶车辆的定位、导航、控制、决策和安全行驶水平。高精地图的构建极具挑战。高精地图包含海量异构数据，涉及路面结构、道路标识、道路环境模型、交通信号和可行驶路面等数据。多源异构数据的融合计算对终端设备的计算和存储能力要求极高。为此，高精地图的构建需要利用更加高效的 AI 制图、AI 智能图像精准识别、三维道路场景智能语义分析与重建等技术，同时，支持云边协同计算与存储模式，利用云边协同新算力支撑海量地图数据的动态计算与实时更新。


智能计算
中国智能计算产业联盟


益企研究院
E7 · RESEARCH

智能计算
中国智能计算产业联盟

益企研究院
E7 · RESEARCH