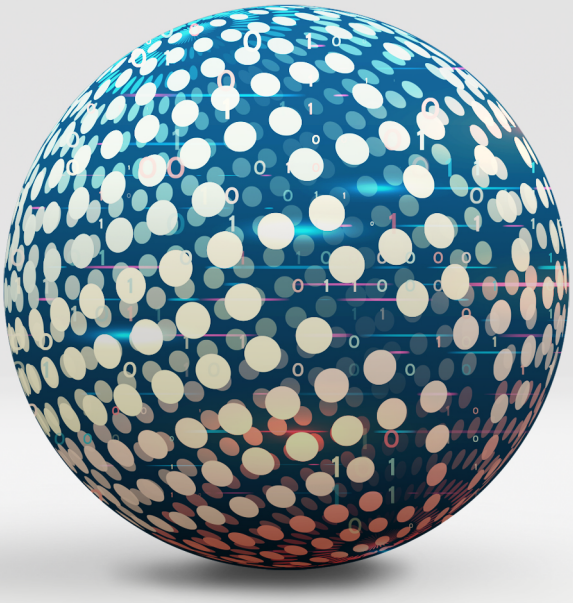
HUAWEI

计算 2030



构建万物互联的智能世界

前言

十年前，人类进入ZB[1]数据时代，移动互联网、云计算、大数据刚刚起步；今天，这些技术已经深刻地改变人类社会，计算发挥了前所未有的作用。

2030年，人类将迎来YB[1]数据时代，对比2020年，通用算力增长10倍、人工智能算力增长500倍[2]。数字世界和物理世界无缝融合，人与机器实现感知、情感的双向交互；人工智能无所不及，帮助人类获得超越自我的能力，成为科学家的显微镜与望远镜，让我们的认知跨越微小的夸克到广袤的宇宙，千行万业从数字化走向智能化；计算能效将持续提升，走向低碳计算，帮助人类利用数字手段加速实现碳中和目标。

未来十年，计算将帮助人类跨入智能世界，这是一个波澜壮阔的史诗进程，将开启一个与大航海时代、工业革命时代、宇航时代等具有同样历史地位的新时代。

目录

宏观趋势

计算经过半个世纪的发展，已经深深地融入了人类的生活和工作。未来10年，计算作为智能世界的基石，将持续推动社会经济发展和科学进步。

面向2030年，中国、欧盟、美国等均将计算作为战略方向重点布局。在中国十四五规划和2035年远景目标纲要中，将高端芯片、人工智能、量子计算、DNA存储等作为强化中国的战略科技力量；在欧盟《2030数字指南针：欧洲数字十年之路》中，计划到2030年，75%的欧盟企业将充分运用云计算、大数据或人工智能，打造欧盟首台量子计算机；而美国，则再次提出“无尽前沿”，借助法案和拨款，推动美国在人工智能、高性能计算&半导体、量子计算、数据存储和数据管理技术等领域的领先性研究。

2030年，数字世界和物理世界无缝融合，人与机器实现感知、情感的双向交互，计算具备模拟、还原、增强物理世界的能力，超现实体验将驱动计算走向边缘，云与边缘、边缘与边缘、虚拟与现实多维协同计算；人工智能将从感知走向认知，具备创造的能力，更加普惠并赋予万物智能；科学探索的边界将不断扩展，带来算力需求的快速增加，未来将出现100EFLOPS[2]级的超级算力和智能的科学研究新范式；碳中和目标驱动计算走向绿色，未来将更好地匹配绿色能源和业务体验。

计算所依赖的半导体技术逐步接近物理极限，计算将迎来创新的黄金10年，软件、算法、架构、材料的创新和突破将开启智能、绿色、安全的计算新时代。预计2030年，全球数据年新增1YB；通用算力增长10倍到3.3ZFLOPS，人工智能算力增长500倍超过100ZFLOPS[2]。

未来计算场景


更纵深地感知
食 智慧农业
企业 智能控制设备
生产机器人
更准确地模拟现实
企业 生产仿真100倍精度
风洞仿真
医 AI新药探索
更精确地探索未知
自然 生态监测
海洋预测
气象&地震预测
更聪明的AI
行 AI无人驾驶
AI智慧交通
城 智慧的城市
永无止境的探索
海量
数据
无处不在的感知
计算2030
无所不及的AI
普惠
算力
源源不断的创新
超越现实的体验
住 智慧交互
虚拟世界/元宇宙
AR/VR
数据驱动的业务创新
企业 算力挖掘数据价值
10倍的新业务开发
更高的运营效率
企业 精细化的资源使用
软件定义的运营
低碳DC
更普惠的AI
医 AI精准医疗
AI药物筛选
教育 个性化教育

更聪明的AI

行：AI智慧交通

智能交通领域，通过摄像头、雷达、气象传感器等采集各种数据，边缘完成车辆识别、交通事件识别、路面状况识别，生成局部路段的全息数据，并在云端形成城市级道路数字孪生，实现车道级实时路况、历史路况的全息呈现。通过云端策略计算，可以对每辆车、每条道路生成不同的交通指令，指挥车辆、调节交通信号，从而更高效、低碳的完成出行。预计2030年，全球道路上的电动汽车、面包车、重型卡车和公共汽车数量将达到1.45亿辆。每辆汽车行驶中产生的数据（一辆车平均每天行驶2小时，行驶中每秒上传的压缩数据将从现在的10KB升至1MB，10万辆车智能网联汽车每天需要传输的数据量大约为720TB）需要在汽车与城市之间频繁进行数据交换，借助智慧交通基础设施的海量数据存储和分析能力，城市通勤时间将得到大幅提升（日均通勤缩短15-30分钟），交通事故和汽车对城市碳排放量也随之大幅降低。以计算为核心，持续支撑交通的数字化升级和智慧化管理。大交通从“运力”时代进入“算力”时代是历史的必然选择，“算力”带来的交通安全、效率、体验的提升，必将释放出新的生产力，推动社会经济的发展。

行：AI无人驾驶

随着L4级的自动驾驶规模商用，数据被源源不断地送往数字孪生，AI在数字世界中不断学习训练，自动驾驶AI将变得越来越聪明，最终将在应对复杂路况、极端天气超越人类，实现更高级L5级的完全自动驾驶。智能驾驶算力需求增长会远超摩尔定律，随着边界案例（Corner Case）的不断积累，算力需求不断增长，到2030年L4+自动驾驶汽车的单车算力将达到5000T。智能驾驶将驱动将无监督学习或弱监督学习带入数据闭环中，利用车端快照获取的图片和视觉信息，实现自动化无监督的视频级AI机器学习训练。自动驾驶催生端云协同的计算需求，未来单个车厂的云端至少需要10EFLOPS以上的算力。

城：智慧的城市

城市占据全球2%的面积，居住着超过全球50%以上的人口，消耗了全球2/3的能源、排放了70%的温室气体（250多亿吨二氧化碳）。城市的智慧化治理成为实现城市可持续发展的必然选择。 智慧城市中的物联网传感器则持续生成城市运行的环境数据，未来，每一个物理实体都将有一个数字孪生，如城市楼宇、水资源、基础设施等将组成城市数字孪生，实现更加智能的城市管理。城市智慧治理将带来100倍的社会数据聚集，实现高效城市治理。

智慧能源基础设施借助数据的保存和分析能力将城市能源消耗中供需二者协调到一个系统中，以实时数据处理来实现城市能源的高效调度。例如：通过城市能源的消耗数据绘制出城市实时能效地图，动态监控能源

的使用情况，再针对性的进行能源调度，将居民高峰用电平均需求减少15%以上。

城市中每个居民息息相关的气象、海洋、地震等公共服务，背后也是依赖大量的数据计算处理。通过更多元、更大量的城市及自然环境数据，智慧公共服务将可以更好地预测天气、海洋和地震对城市生活的影响，从而使城市在面对极端事件时更具韧性。每个居民还可以通过这些智慧化的公共服务，结合自身地理位置等信息，以定制化的信息判断气候或突发事件对自身的影响。

数据是智慧城市高效运作的核心要素，如何对生成的海量数据进行有效管理和使用是智慧城市发展绕不开的主题。

更普惠的AI

医：AI精准医疗

在医疗领域，人工智能已经可以自动识别出微小的肺结节，与以往肉眼识别、手工标识相比可以节省医生大量的时间。未来人工智能将在更加复杂的问诊中，深入参与医生的病情推理过程，与医生“讨论病情”，为医生提供可解释的诊断依据和预期疗效分析。这将使得人工智能出方案、医生审核成为普遍的诊疗模式。世界卫生组织估计到2030年，将出现1800万卫生工作者的短缺，人工智能将为人类应对这一挑战提供有力的帮助。

医：AI药物筛选

AI将更加透明，不再是一个黑盒，不仅会告知结论，同时也会告知如何得出结论的，让人类明白AI的思考过程，和人类建立彼此的互信。有了这样的基础，AI就可以在更广的范围内发挥更大的作用，帮助人类完成复杂的任务，比如：进行抗病毒药物筛选，AI会告诉我们选出药物的原因，而不是只给出一个药物列表（通常情况下，如果我们只是看到一个结果，将很难做出决策）。

教育：个性化教育

人类训练人工智能的过程，同时也是认识自己的过程，人工智能使得认识人类的智能、人脑的规律变得更加重要，进而重新认识教育、改革教育[3]。未来人工智能将改变人类自己的学习、认知的过程。如人工智能教员通过精细化地分析学生的行为、习惯、能力等，制定个性化的教学内容、计划和教学方式，学生

的学习潜力将得到极大的挖掘，接纳新知识更多、更容易。

人工智能进入人类生活的方方面面，让我们更高效的思考、创作、学习，让优质稀缺的资源变得更加容易获得，将在精准医疗、创作设计、文化教育、老人护理、社区服务、自动驾驶等领域普惠每一个人。

更纵深地感知

预计到2030年，全球联接总数将达到两千亿，传感器的数量达到百万亿级，传感器持续不断地从物理世界采集数据，温度、压力、速度、光强、湿度、浓度等，为了让机器人具备“视觉、触觉、听觉、味觉、嗅觉”，需要更加多维的感知能力。数据量、时延等原因决定了产生感知的计算在边缘完成，边缘将具备智能的数据处理能力，例如模仿人类大脑工作的模拟信息处理技术等。未来，大量感知计算将在边缘完成，处理大约80%的数据。

感知智能让海量数据的采集、分析成为可能，让更多的行业获得感知自我的能力，并通过云端的数字孪生与物理世界形成协同，驱动行业的数字化创新。

食：智慧农业

未来将建立和完善天空地一体化的智能农业信息遥感监测网络，互联网、物联网、大数据、云计算、人工智能等现代信息技术与农业深度融合，具备农业信息感知、定量决策、智能控制、精准投入、个性化服务的全新农业生产方式将逐步实现。智慧大田、智慧大棚、智慧养殖、智慧种植、喷药无人机等对边缘AI计算有广泛的需求。农业智能传感与控制系统、智能化农业装备和农机田间作业自主系统将加快发展农业电子商务、食品溯源防伪、农业休闲旅游、农业信息服务水平，农业将迎来全方位全过程的数字化、网络化、智能化改造。

企业：智能控制设备

人工智能技术将在生产系统中高度普及，融入企业作业各个环节，这将带来工厂作业模式、人员配置、部门区域协同等一系列的升级。未来10年，人工智能技术将给关键生产环节带来大幅的质量提升与成本收益。AI可以帮助制造企业实现控制层智慧化运营管理、贯通层海量数据分析挖掘以及感知层更低时延诊断预警。中国制造2025提出，制造业重点领域全面实现智能化，试点示范项目运营成本将降低50%，产品生产周期缩短50%，不良品率降低50%。比如，工厂的轴承故障诊断、钢炉热异常检测、电力设备的检修等深度学习场景，制造工厂可以通过AI技术进行更低时延的诊断预警，提高生产检测效率，缩短订单交付周期。

企业：生产机器人

未来，从操作机械到指挥机器，人类告别恶劣极端的工作环境。人工智能将参与企业更多的非操作性任务，人与机器形成无缝的协作关系。从产品设计、生产、销售，到企业架构、员工的雇用和培训等各个环节，人工智能将驱动企业业务进行彻底的重

塑。如企业采用人工智能对经济发展、社会热点事件等进行分析，判断行业外部及企业的发展趋势，或者根据分析结果优化生产计划、形成方案，为产品概念的开发提供决策建议；特别是在满足个性化需求的柔性生产中，人工智能的创造能力不仅能够按照定制要求设计，更能综合需求变化和产品使用数据生成新的产品设计。预计到2030年，每万名制造业员工将与390个机器人共同工作，机器准确理解人的指令、准确感知环境、做出决策建议与行动。

今天，无人值守的黑灯工厂已开始规模部署，人工智能驱动机器人忙碌于生产线和物流系统，在重复性高的场景中，机器让人类告别重复枯燥的工作。未来，机器将帮助人类处理非确定场景下危险、恶劣工作，人将从现场操作转变为远程指挥，在更加舒适的环境中工作，远离危险。

如在采矿业，中国提出了煤矿智能化发展的目标，到2025年大型煤矿和灾害严重煤矿基本实现智能化决策和自动化协同运行，井下重点岗位机器人作业，实现井下少人到井下无人，2035年建成智能感知、智能决策、自动执行的煤矿智能化体系。[4]

从操作性工作到创造性工作，企业智能化重塑。未来人工智能深度参与人类的思考，与人形成互动，并呈现出推理的过程，成为可信任的智能，将在金融、医疗、司法等需要高质量决策的复杂场景中发挥巨大作用。未来10年，通过对物理世界的不断学习，人工智能将更加聪明，从确定性场景到非确定性场景，在越来越多的任务领域中增强人类，帮助人类获得超越自我的能力。

超越现实的体验

住：智慧交互

今天，人工智能已经在帮助人类完成一些过去难以完成的任务。例如，通过手机摄像头可以识别出我们所不认识的植物，并能获取它的生活习性、栽培方法；机器人帮助增强人类的行动力，如外骨骼机器人辅助病人进行康复；家用机器人则能帮助老人陪伴、家务劳动等智能化工作。预计2030年，家用智能机器人使用率将超过18%。

人工智能参与人类的思考和创造过程，需要结果具备可解释性，并符合人类思考问题的逻辑，具备与人类使用自然语言无缝交流的能力，未来人工智能将实现从感知到认知、从弱人工智能到强人工智能的跨越。

当前人工智能在写诗、作画上进行了初级尝试，未来人工智能将完成更加复杂的创造性工作，如电影制作、艺术创作和工业设计等。人工智能能够提供高度定制化的内容服务，人们可以随时获得一幅定制的画作，一部定制的电影。比如在互动电影的观看过程中，观众可以在观影中通过不同的选择来影响剧情走向，人工智能将完成每一条故事线的演绎和视频生成，因此相同的电影将产生不同的结局，整体内容也更加丰富。未来这种人类提出主题、人工智能实现细节的创作方式将极大地提升人类的创造力、丰富人

们的生活。

住：AR/VR

数据将构建出众多的数字空间，旅游景点、全息会议、虚拟展会……这些数字空间与物理世界共同组成了一个虚实融合的世界。在虚拟旅行中游览“真实的”山川、流水；与千里之外的朋友促膝交谈；对话先哲，与王阳明一起悟道，与普鲁塔克探讨特修斯之舟；人与人、人与社会、人与自然、人与机器的交流方式将发生革命性的改变，未来人类的生活、工作和学习方式将重新定义。预计2030年，超过30%的企业在数字世界中运营与创新，各种虚实结合的AR（Augmented Reality，增强现实）/VR（Virtual Reality，虚拟现实）用户数达到10亿。

住：虚拟世界/元宇宙

数字世界与物理世界的无缝融合，能够准确感知和还原物理世界，在虚实结合的世界中理解用户的意图，体验将驱动计算走向边缘，云与设备、设备与设备、虚拟与现实多维协同计算。云端将实现物理世界的建模、镜像，经过计算、加入虚拟的元素，形成一个数字的世界；边缘设备将具备听觉、视觉、触觉、嗅觉和味觉能力，人与设备之间实现实时交互；多维协同的计算将用户所处的环境整体变成一台超级计算机，计算环境信息、识别用户意图，并通过全息、AR/VR、数字嗅觉和数字触觉等技术进行用户呈现。

更精确地探索未知

今天，“高性能计算+物理模型”的方法已被广泛应用到众多的科学问题。未来，随着人类认知边界的不断扩展，量子力学、生命科学、地球大气、宇宙起源的研究，尺度从10^{-21}到10^{28}米，跨越微观世界到无垠宇宙，科学家需要处理的数据与计算量将爆炸性增长，数字世界算力的规模决定了物理世界探索的广度和深度。例如，2012年欧洲核子研究组织（CERN）大型强子对撞机（LHC）实验项目，全球超级计算机组成算力池，帮助科学家从近100PB数据中证明希格斯玻色子的存在；2027年底CERN将投入使用高光度大型强子对撞机(HL-LHC)，每秒发生约10亿次粒子碰撞，数据计算量将增加50-100倍[5]，存储需求达到ZB级。2030年，计算将在更多的领域帮助科学家解决基础性问题。

自然：生态监测

未来人类将环境保护作为重点，将新型科学技术与设备结合人工智能，可有效解决环境恶化带来的温室效应，土壤沙化和盐碱化等各种环境问题的挑战。以大数据为基础，利用模型，可以较好地预测出不同管理措施下的结果，并不断反馈给算法模型，得出更好的治理模式。如精确定位污染源，预测污染扩散等。

自然：气象

未来天气预报不断发展为更加复杂的动力数值模式，以求更准确和提前预报天气。如气象雷达质量控制、卫星数据反演及同化等气象数据处理；短时临近预报、概率预报、台风海洋天气预报、极端或灾害性天气预警、风暴环境特征分类、环境预报等天气气候分析； 以局部短时天气预报为例 短时强降雨具有极大的破坏性，但受限于海量数据和巨大算力需求，很难实现准确预测。天气预报从当前的10公里的精度，提升到公里、次公里，数据规模和算力需求提升100~1000倍。预计2030年随着100EFLOPS级超级计算机的出现，更高精度气候模拟和天气预报将成为可能，人类能够更加从容的应对极端天气。

自然：地震预测/海洋预测

未来应用人工智能监测地震、实时估算地震震源等将极大提高预报的准确性。从地震记录推算地震震源机制是个计算耗时的过程，自1938年地震学家第一次开始推算地震断层面解，震源机制参数一直是个难题。采用人工智能方法有效地解决了这个复杂计算问题，应用地震大数据训练人工智能神经网络，可完善预报系统的准确性和可靠性，实现地震预报领域的突破。

自然：宇宙结构探测

宇宙大规模结构是重要的科学前沿领域，研究宇宙结构形成和时间演化，从而揭示宇宙的物质组成以及宇宙演化过程、暗物质、暗能量等宇宙学问题。传统的办法是根据物理理论，使用超级计算机计算宇宙中各种大规模结构的演化，将其与观测数据进行对比，但是这需要对数十万到百万个宇宙论模型进行精确的计算，目前可观测的宇宙有2万亿个星系，万亿亿个星球，即使全球所有计算资源一起也难以完成。

更准确地模拟现实

企业：生产仿真100倍精度/风洞仿真

计算机风洞仿真已经成为飞机、高铁和汽车等高速运动产品的重要测试手段。但由于整机仿真计算量巨大，为了得到高精度的仿真结果，需要将测试系统分解成滑行轮胎、发动机等多个子系统，甚至发动机也要拆解成更小的系统，这对验证整机设计是否满足要求带来新的挑战。未来计算能力将提升2~3个数量级，风洞仿真有望实现更大级别子系统，甚至整机的高精度仿真测试。

医：AI新药探索

2013年诺贝尔化学奖授予了在开发多尺度复杂化学系统模型中做出突出贡献的科学家，评选委员会在声明中阐述道：对于化学家来说，计算机是与试管同等重要的工具，计算机对真实生命的模拟已成为当今化学领域中大部分新研究成果成功的关键因素。

组合量子力学/分子力学方法(QM/MM[6])建模是当前研究酶催化机理最可靠的计算模拟方法之一，核心区域采用高精度QM模型、外围采用低精度MM模型，兼顾量子力学的精确性和分子力学的高效性。用该模型模拟0.2微米生殖支原体细胞2小时的生长繁殖过程，超级计算机Summit[7]需要

耗费10亿年。对于更复杂的人脑思维、记忆和行为研究，如模拟人脑在特定刺激下的反应，每一小时模拟Summit需要计算10^{24}年[8]。

图灵奖得主吉姆·格雷(Jim Gray)将科学研究分为四类范式，即实验、理论、计算机仿真和数据密集型科学发现。今天，在生物、材料、化学、宇宙等演化复杂度极高的研究方向上，传统的计算方法面临变量数量、自由度增多带来的“维度灾难”挑战，算力需求呈指数级增长。

人工智能将为解决“维度灾难”开辟新的解决办法，为科学研究打开新的探索之道。例如，采用传统方法分析单个蛋白质的折叠结构，需要耗费科学家数年时间；通过人工智能学习已知的1.8万种蛋白质折叠结构，可以在几天内获得对未知蛋白质折叠的原子精度模拟结果。这一成果使得癌症、老年痴呆等细胞内蛋白质结构变化引起的世纪难题的预防、治疗成为可能。2020年戈登贝尔奖[9]的研究工作，利用人工智能实现了1亿原子规模模拟，比过去的同类工作计算空间尺度至少增大100倍，同时计算速度提高至少1000倍，实现了传统方法无法模拟的大尺度计算，将精确的物理建模带入了更大尺度的材料模拟中[10]。

未来科学计算将向着数据、计算、智能融合的方向发展，催生新的科学研究范式。通过人工智能学习已有知识、分析总结理论，在线迭代结合传统建模的方法将极大的提高科学探索效率，加速人类认知的不断扩展。

数据驱动的业务创新

企业：算力挖掘数据价值

云计算和大数据已经成为行业数字化的基础，驱动以管理效率提升为目标的数字化，其特点是优化生产关系，更好的匹配生产力和客户需求，如O2O（Online to Offline，线上到线下）服务、电商平台等。

企业：10倍的新业务开发需求

端边云全栈Serverless化成为支撑企业数字化、智能化转型中应用现代化改造的主流技术，基于云原生计算模式的编程语言、语言runtime、应用调度、运行、运维，成为构建全栈Serverless化、现代化软件的基础，实现全面应用上云，构筑10X的性能、效率、成本优势。

更高效的运营效率

企业：精细化的资源使用

云技术的广泛运用将使企业更加便捷地使用计算资源，新的计算技术可以让企业消费资源的粒度更小、调度的时间更短，这将大量减少企业计算资源的浪费。例如，在非云化时代，处理器仅有10%的利用率，容器技术则将这一比例提升到了40%以上，未来新的资源管理技术的广泛采用将进一步减少50%以上的资源浪费。

企业：软件定义运营

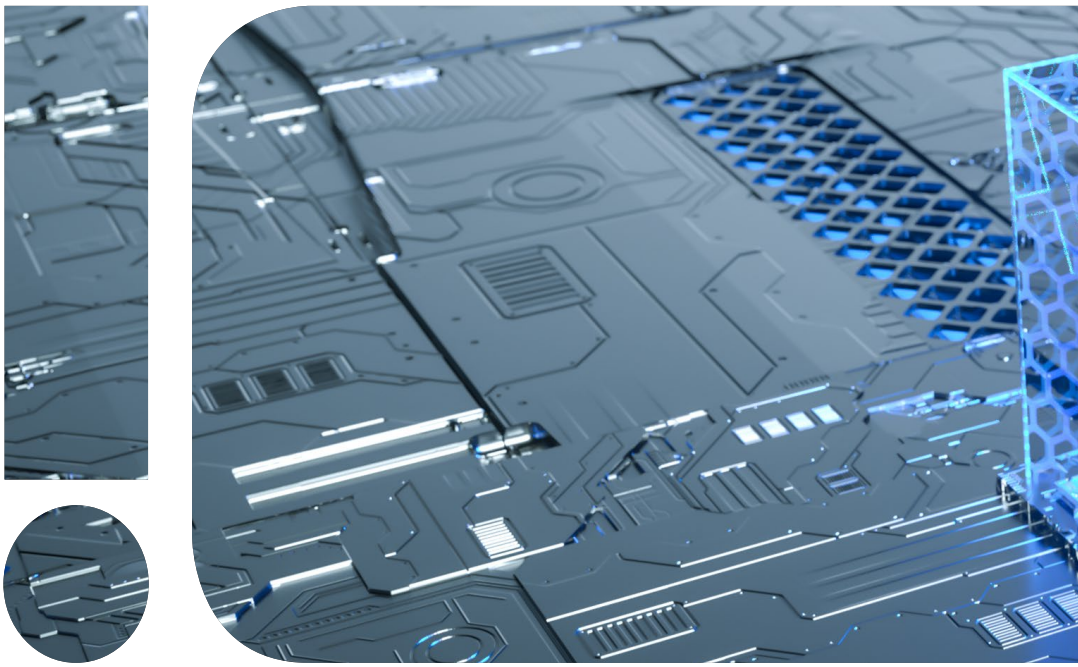
IT越来越成为企业生产系统的重要组成部分。互联网企业因为采用DevOps[11]（敏捷开发和开发运维一体化）而变得敏捷高效。2030年，工业制造等传统企业将在更加复杂的产业链上下游环境中实现由软件定义的高效企业运营。

工业物联网将驱动全球的供应、制造、维护、交付和客户服务等业务流程实现广泛联接，各类公司将集体组成一个跨越全球的价值网络，企业的数字化转型将从内到外转移至整个产业链的优化与协同，对数据的依赖从企业自身的数据扩展到产业链的上下游甚至是整个社会。未来企业将通过软件处理跨组织复杂协同，通过软件快速定义业务的运营，比如，流程自动化机器人、无代码/低代码等开发技术，通过人工智能支撑的自然语言编写程序，调用机器人自动化软件的能力，申请各类服务资源，编排各种业务流程，普通员工即可完成工作流程的优化和问题的解决。

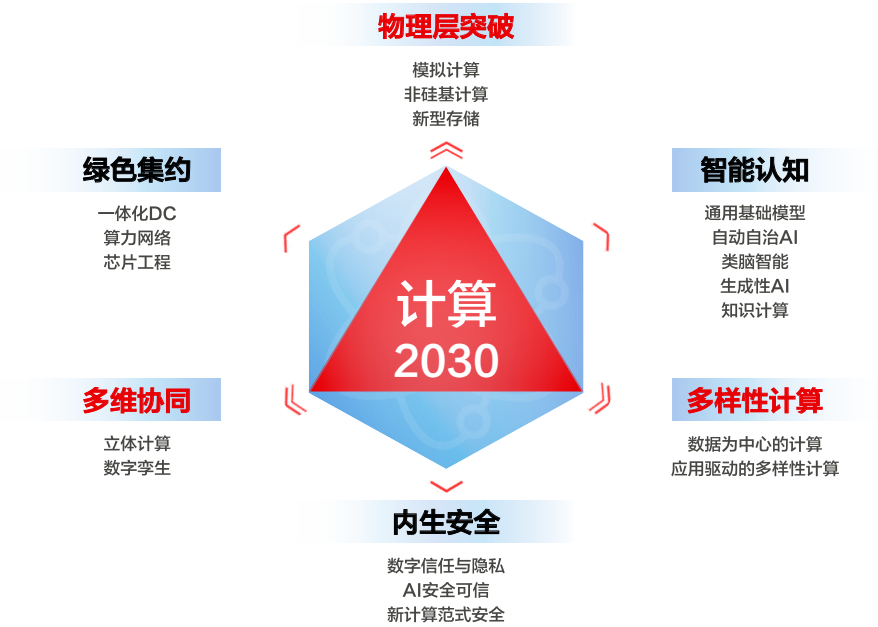
企业：低碳DC

2030年，数据中心将在算力提升百倍的同时实现低碳的目标，企业将获得更加绿色的计算资源。创新计算架构的引入，计算能效将极大的提升。例如，传统计算过程中超过60%能耗集中在数据迁移，而未来以数据为中心的计算将使得能效提升数十倍。模拟计算如量子计算、模拟光计算将逐步成为重要的算力来源，能源效率更能得到指数级的提升。

碳中和目标的驱动下，未来数据中心将受能源分布、算力需求分布的双重影响，计算架构将在更大的空间维度上发生变化，通过算力网络可以更好的匹配绿电、时延、成本的差异，实现全局最优的PUE（Power Usage Effectiveness，能源利用效率）与碳排放。可将人工智能训练、基因测序任务放到绿色能源丰富和气温较低的地区，工业控制应用、AR/VR放到靠近客户生产环境的地区。

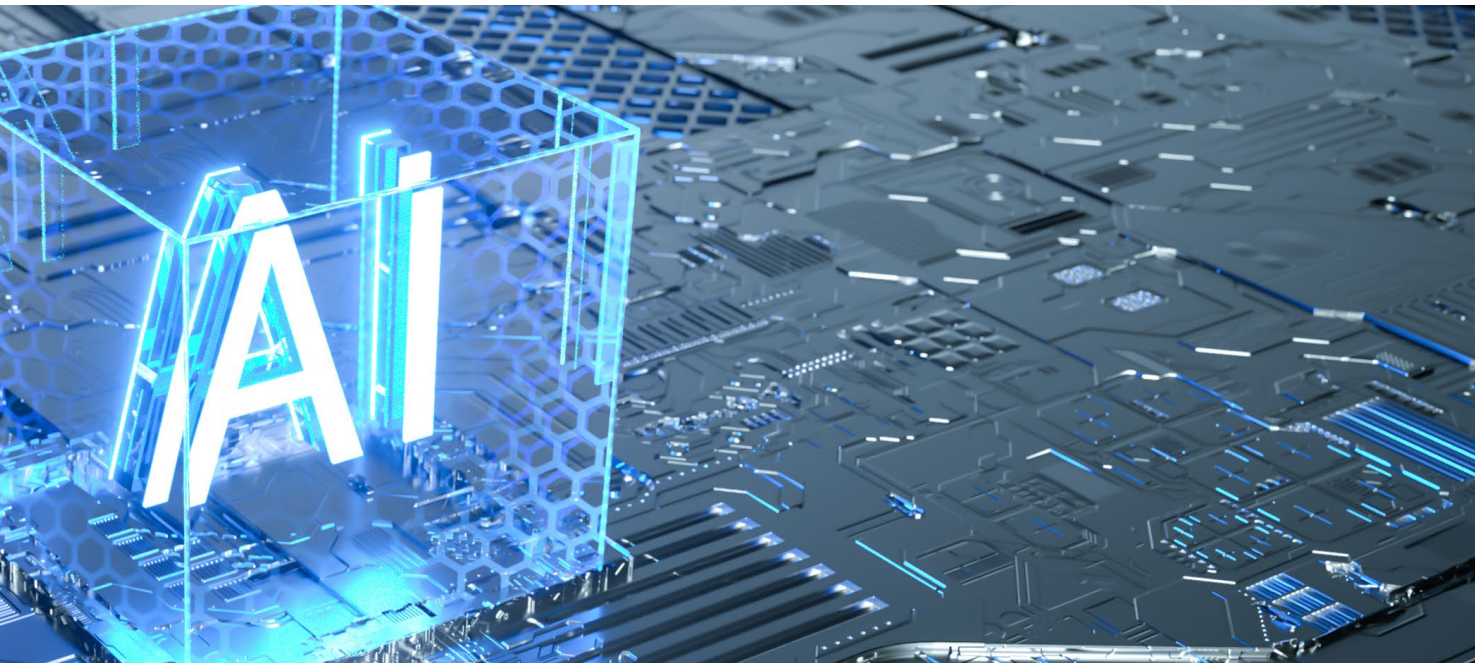


计算 2030 愿景及关键特征



智能认知

AI从感知走向认知是必然趋势，认知智能是人工智能技术发展的高级阶段，旨在赋予机器数据理解、知识表达、逻辑推理、自主学习等的能力，使机器成为人类改造世界、提升能力的得力助手。在从感知智能到认知智能的发展过程中，语义及知识的表达和逻辑推理，是进行认知的重要手段，而多模态学习则是获得信息融合和协同的重要手



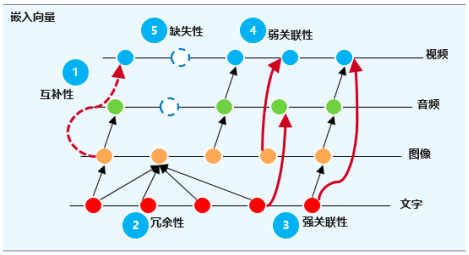
段。通过构建多模态的大规模基础模型，可以学习多种信息的融合表征，建立模态转换和协同关联，从而提高AI系统对于复杂环境的认知和理解能力，进而获得多场景多任务的AI应用能力。

通用基础模型

AI从感知走向认知是发展趋势：人工智能从早期的计算智能，升级到现在的感知智能，并将逐步走向认知智能。机器在运算速度和存储方面具有一定的优势；感知智能方面，利用深度学习和大数据分析，机器在视觉、听觉、触觉等方面执行确定任务的能力上接近人类；认知智能使得机器具有类人的理解和推理等能力，成为人类认知世界，改造世界的有利工具。

从弱人工智能到强人工智能发展的路线上，提高机器解决问题的“泛化”能力，是重要的手段。在场景泛化、模态泛化、任务泛化等方面，通过大规模基础模型的领域通用方法，赋予AI系统解决多问题的能力。

多模态学习是构建基础模型的重要手段：多模态首先要解决数据异构性问题，由此带来的一系列挑战包括：如何利用跨模态数据的互补性及冗余性做好表征学习；如何处理跨模态数据的强弱关联性做好表征学习后的向量关系映射；如何处理训练场景下某个或某类模态数据缺失后模型自适应的学习及迁移能力保障模型精度维持在可接受范围内；如何处理推理场景下某个或某些模态数据缺失后的模型拓扑路由提高推理增益等。从发展趋势看，多模态模型应具备跨模态自监督学习与通用知识迁移能力，可以使不同领域任务在多模态框架下实现统一。



- 高效表征学习
- 精准关系映射
- 自适应学习（训练）
- 模型拓扑路由（推理）

多模态学习将在以下关键领域实现突破：一是预训练数据标注技术，关联解译文本、音频和视频帧等；二是多流编解码技术，从单模

态预训练模型到多模态关联编码，可实现多模态信息弱关联学习，解码器支持跨模态转换与生成任务；三是自监督学习技术，实现文本、语音、视觉等各模态信息的语义对齐及相互预测；四是下游任务微调技术，实现多模态语义理解、多模态生成等任务；五是多模态模型小型化技术。

自动自治AI

目前，深度学习的开发及应用并未突破主流监督学习的模式，数据清洗、数据标注，模型的设计、开发、训练和部署等都需要大量人力投入。迁移学习、小样本、零样本、自监督、弱监督、半监督、无监督及主动学习等新方法，将推动人工智最终实现“自治”，解决模型训练、迭代、设计对人工的依赖。未来AI自治使得模型更加归一，多种任务共享相同的模型结构，数据规模进一步扩大，不再需要人工干预，模型可以在线学习吸收新的数据知识，实现自身能力的迭代提升。数据规模扩大及在线学习将使模型的生产更加集约化，各行业的业务模型会汇聚成几个甚至一个超大模型。自治AI仍面临如下挑战：

1）从依赖人工显式标注转向自监督，由训练过程中转向执行过程中同步自反馈。

2）目前模型学习到的表征都是自然产生的，多次训练的模型内在表征大相径庭，需要克服灾难性遗忘，在线持续学习，形成流式训练、训推一体。

3）从人工设计多个模型匹配不同任务，到单模型学习多任务编码，在线按需切换。

类脑智能

当前的深度学习技术主要以数据驱动，严重依赖于大量的标签数据和超强算力，基于反向传播的训练算法，在模型鲁棒性、泛化能力和可解释性上都需要持续增强。类脑智能期望借鉴和模仿生物神经元的工作模式，通过构建功能更加丰富的神经元，具有事件触发、脉冲编码、时间和空间信息协同处理的能力；利用神经动力学原理，可实现短时可塑性和长期记忆，在开放环境中具备自适应调整和学习能力；借鉴生物脑的稀疏连接和递归特性，没有脉冲发放就不会产生计算，可大大减少能耗。如果能够突破相关技术，未来五到十年类脑计算可能会在很多计算任务中展现出性能和功耗方面的优势，并在智能终端、穿戴式设备、自动驾驶等领域得到应用。

由于当前对人脑学习机理的研究还不够透彻，其学习效率、运算精度相对深度学习还有差距。未来类脑智能需要从两个方向突破：一是自下而上，模拟生物脑中的脉冲神经网络，借助神经形态芯片实现一定规模神经元和突触，并在时序相关的应用中实现低功耗低时延；二是自上而下，从功能角度构建神经动力学理论和认知理论，并将其与人工智能结合，实现更鲁棒、更通用的智能。

生成性AI

生成性AI（Generative AI）技术作为最佳的自动化内容生产力要素，允许计算机抽象与输入（例如文本，音频文件或图像）有关的基础模式，使用它来生成期望的内容，可以用于身份保护、图像修复、音频合成、抗菌肽（AMP）药物研究等领域。

生成性AI与训练数据保持相似，而不是简单的复制，可将人类创意融入设计和创作过程。如结合3D游戏生成引擎，测试挑战智能体的视觉、控制、路线规划和整体游戏能力，加速智能体的训练。在生成性AI应用开发中，具有随时间动态改进、自我进化能力的生成模型是关键。

生成性AI具有如下挑战：

1）某些生成模型（例如GAN，Generative Adversarial Network，生成式对抗网络）不稳定且难以控制其行为，如生成图片的精确度不足，无法产生预期的输出，并很难判断原因。

2）当前生成性AI算法仍需要大量的训练数据，不能创造全新的事物，这要依赖自我更新、自我进化的算法突破。

3）恶意行为者可以将生成性AI用于欺诈目的，利用人工智能工具的本身漏洞进行远程攻击，导致数据泄露、模型篡改、虚假垃圾邮件等事件，对网络信息安全形成极大威胁。

知识计算

人工智能在行业中的应用，要能够通过跨学科的领域专家知识进行综合决策，形成完善的知识抽取、知识建模、知识管理、知识应用的技术体系。未来十年知识计算将实现知识抽取从文本、结构化特征，到多模态知识对齐、抽取与融合，复杂任务知识抽取，跨领域综合知识抽取等复杂、多层次知识发展的跨越；知识建模则向垂直场景化、原子化、自动化、规模化的知识图谱，进一步向垂直场景知识图谱与通用知识图谱的融合发展；知识的应用从简单的查询、预测，向因果推理、长距离推理、知识迁移等高阶认知方向发展。

知识计算的应用需要在算法上突破海量稀疏信息检索、不定长的知识引入、知识注意力（Knowledge Attention），大规模图式计算；在认知智能的训练模式上，需要突破训练推理时高频度知识检索、知识结合的训练特征提升等；在计算上，需要解决高频度的随机检索训练与推理，高速数据通路，诸如随机漫步(RandomWalk)、结构采样的图式计算等问题。

内生安全

计算云化打破了传统安全边界，传统基于信任域划分的外挂式安全防护方案已经无法应对各种新型攻击方法的挑战。安全应该具备内生的特点；1）安全是系统的内生能力，是芯片、固件、软件必备的基本特性；2）安全贯

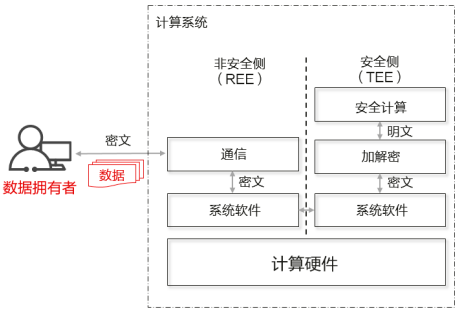
穿存储、计算、传输等数据处理的全过程，以抵御各环节安全攻击；3）硬件构建安全信任根，由于系统权限分级的原因，安全功能需要基于硬件的最高特权来实现，才能在操作系统及应用上提供可靠的安全服务，并且通过硬件加速的方式来提升安全服务的性能。4）安全开源开放，为了使安全服务能充分融入到各个业务软件中，安全服务应以开源开放的形式提供，让业务软件结合自身软件架构特点，将安全特性融入到业务中，从而保证业务安全。

数字信任与隐私

数据处理的本质是算法施加算力于数据。如果这3个要素全部由数据所有者控制，则不涉及数字信任与隐私问题；但云计算导致要素分离，算法与算力都是由算力提供商提供，用户（数据拥有者）需要上传数据到云端处理，即使用户信任算力提供商，也无法信任算力提供商拥有特权的管理员。因此云计算场景下安全的主要挑战在于如何保护用户数据与隐私，需要重建数字信任体系。

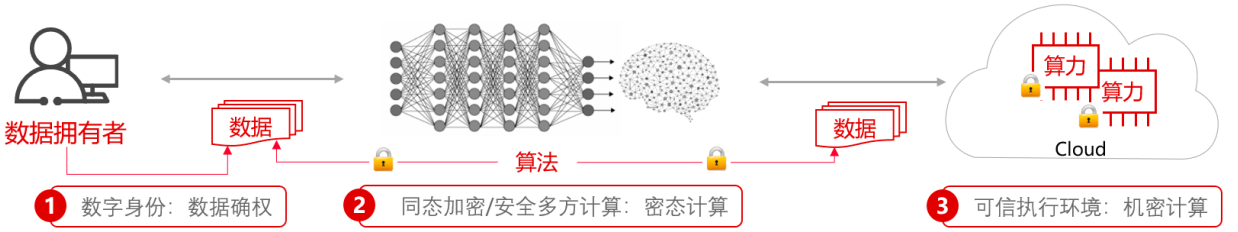
为重建数字信任体系各国政府相继出台数据保护法，为数字信任体系的建立确立了法律依据。同时，数字身份与隐私计算成为重建整个数字信任体系的关键技术，其中数字身份是数据确权的基础，隐私计算可以在保护数据本身不对外泄露的前提下实现数据分析处理：

1）基于TEE（Trusted Execution Environment，可信执行环境）实现敏感数据处理的硬件隔离技术，主要挑战在于硬件安全隔离机制实现的完备性无法用数学证明，难以自证清白，存在安全漏洞风险。但和密码学技术相比，TEE对性能影响小，未来基于TEE的隐私计算将成为公有云、互联网以及企业重要业务的普遍需求，预计2030年50%以上的计算场景将使用该技术。



2.）基于密码学的同态加密、安全多方计算技术因其安全性在数学上可证明，从而成为业界公认最理想的隐私计算技术。但主要挑战在于其性能比常规计算降低一万倍以上，需要大幅度提升才能满足应用需求。随着近似算法的成熟，同态加密、安全多方计算技术在人脸验证、健康数据分享等特定领域已获得应用。未来，突破基于硬件加速的同态加密、安全多方计算技术，将在金融、医疗等行业的高安全应用场景获得广泛商用。

3）多方计算的基础是多方之间共享秘密，如果通过零知识证明等密码学方法实现，性能开销非常大，利用TEE来实现多方之间的秘密共享，不但可以大幅提升多方计算性能，而且在信任TEE的基础上安全性可数学



证明，未来有广泛的应用前景。

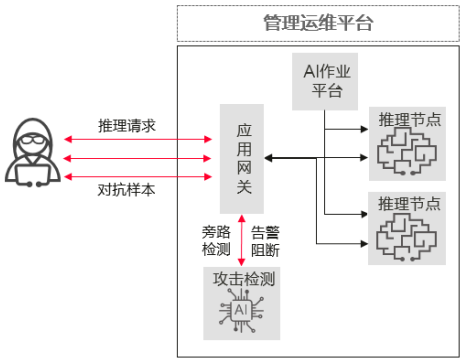
AI安全可信

随着AI应用的普及，特别是在医疗、自动驾驶等关键领域的应用，AI面临日趋严峻的安全挑战：1）AI模型和训练数据是AI应用厂商的核心资产，如果保护不善可能被恶意逆向恢复。2）AI模型本身存在脆弱性，导致针对AI模型的闪避和药饵等攻击越来越多，在关键领域中使用的AI模型被攻击导致误判将带来严重后果。3）因为人类对AI顾虑越来越大，AI伦理、取证成为新的安全挑战。

为应对AI日益严峻的安全挑战，AI监管合规与治理成为AI生态中各参与方的必选项，同时也需要创新的技术手段支持对多参与方的责任追溯，从而实现负责任的AI（responsible AI）：

1）AI模型与训练数据保护：AI模型与训练数据需要通过加密、强制访问控制、安全隔离等手段保证AI模型与训练数据在收集、训练及使用阶段的全生命周期安全。核心挑战在于如何对NPU（Neural network Processing Unit，神经网络处理器）芯片的高带宽的内存数据进行实时的密态处理，并确保性能无损。未来需要突破高性能、低时延的内存加密算法，以及突破NPU片上的内存硬件加密引擎的架构设计，提供全生命周期的保护能力。

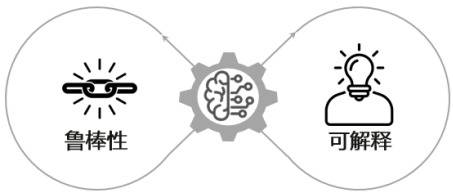
2）AI攻击检测与防护：通过增加外部对抗样本检测模型实现对数字闪避和物理闪避等AI攻击的识别，阻断攻击路径，防止AI模型受到攻击后产生误判。主要挑战在于持续的进行对抗训练以适应新的攻击类型，未来会出现针对AI攻击的独立安全产品与服务。



3）除上述针对已知攻击手段所做的防御之外，也应增强AI模型本身的安全性，避免未知攻击造成的危害。包括增强模型鲁棒性、模型可验证性以及模型可解释性。

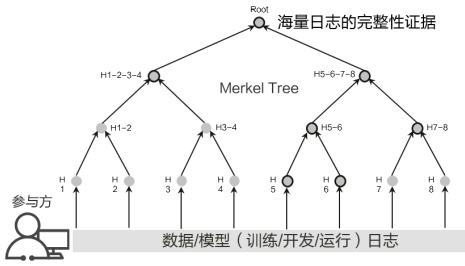
通过对抗训练，提高抗攻击能力是AI模型安全能力提升的主要技术路径；对抗样本的泛化能力，模型正则化将是需要突破的关键技术；未来对抗鲁棒性有望从当前较低的水平提升到80%。

未来针对小模型存在有效的形式化验证方法，可证明模型的安全性；面对大模型的形式化验证还面临巨大的挑战。



为了防止AI带来业务法律风险或者逻辑风险，需要了解AI模型做出判断的依据。未来通过建模前的“数据可解释”，可以构建事前“可解释模型”。目前线性模型基本都具备可解释性。针对非线性模型，还将面临巨大的挑战，目前还无法做到AI模型的全局可解释，但是，对网络模型的分层可视化和局部可解释，将会是未来很长一段时间的可能实现的技术路径。

4）为了满足AI监管要求，未来在AI模型运行过程中必须持续监控与审计，并通过区块链等技术保证审计结果可信，实现AI问题实时可追溯。



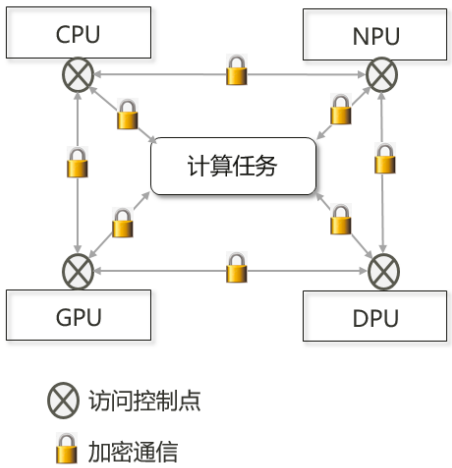
新计算范式安全

在以数据为中心计算场景下，算力下移，特别是内存计算PIM（Processing-In-Memory，内存内处理）将算力下移到内存，导致传统内存加密机制失效，无法部署基于硬件的隐私计算技术。即使在应用层加密数据、数据处理过程中，也将是明文状态，从而导致无法防止特权用户、进程窃取数据。针对这种场景唯一的选择是部署同态、多方计算等基于密码学的隐私计算技术，从而建立用户对于算力提供商的信任。

在多样性算力数据中心场景下，云化导致网络安全边界模糊，传统的基于边界的安全防护模式逐渐失去价值。针对这样的趋势，零信任安全架构[12]通过强化访问策略、主动监测、加密等技术以应对环境不可信的安全挑战。零信任安全架构与多样性算力发展趋势相结合确定了未来多样性算力安全技术走向：

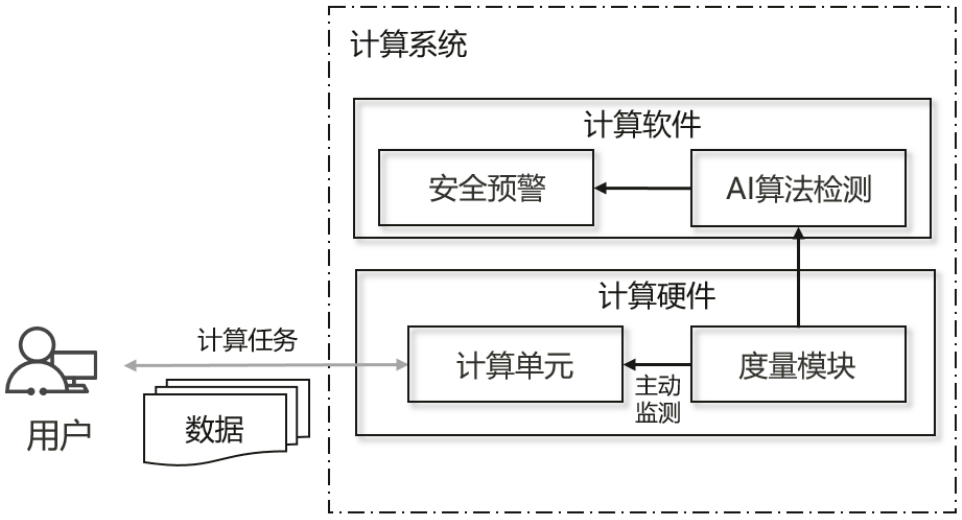
1）安全与在网计算架构相结合：零信任架构打破安全边界后需要更细粒度的权限与访问控制，实现动态的身份验证和资源访问策略，软件实现将占用大量CPU资源；在网计算架构中融入正则表达式硬件加速机制，可以有效提升策略执行效率10~15倍。

2）安全与多样性计算架构相结合：零信任架构假设网络环境不可信，无论在网络的任何位置，通信都应该加密，包括计算节点间、数据中心间。因此需要在多样性计算架构的每个xPU（x Processing Unit，泛指各种处理器）中融入加密通信的高性能硬件卸载能力，并支持后量子加密算法，以应对未来量子攻击风险。



3）安全与数据为中心的对等计算架构相结合：未来，在数据为中心的对等计算架构中，非易失性高性能的内存介质将会接入到系统的内存总线上，掉电后内存中残余的数据目前尚无加密机制，数据与隐私泄露风险将大幅度提升；在数据为中心的对等计算架构中如何实现数据安全将成为新的挑战。例如：在分布式集群系统中，面对跨数百计算节点共享的大内存，如何进行数据保护，实现内存访问的带宽性能下降逼近理论极限，<3%。

4）DC（Data Center，数据中心）级的



动态度量和主动监测：当前的算力平台对系统中运行的计算任务通常并不感知，即使系统被攻击也无法有效区分恶意行为与正常计算。在DC场景下，如何实现计算设备对系统中的计算任务的深层感知，主动度量系统状态并监测计算任务，自适应地感知和防护潜在的恶意行为，保证算力安全、防止算力被盗用等，还面临诸多挑战。

绿色集约

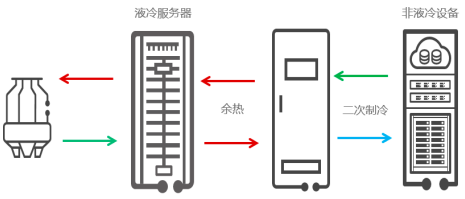
全球数据中心能耗约占电力需求的1%，通用计算的总能耗每3年增长1倍，碳中和目标将驱动算力提升百倍的同时提升能源效率。在芯片上，新的封装和架构持续优化，不断提升算力密度和能源效率，芯片出光减少高频数据交换损耗。一体化数据中心利用人工智能实现供电、服务器、负荷的协同，形成更优的PUE，并不断挑战PUE极限，甚至向小于1发起挑战。通过算力网络将提供对等服务的分布部署的数据中心资源统一起来，更好匹配时延、绿电、成本等差异，达到全局最优的PUE和碳排放。

一体化DC

1）DC级全栈融合架构

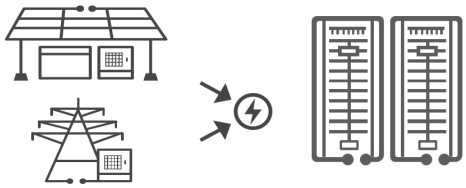
随着人工智能、超算、云等计算场景的快速发展，未来将会出现百万级的数据中心。重点要解决端到端散热问题，灵活的硬件配置与资源使用效率问题，百万级的中心节点和海量的边缘设备的统一运维的问题。

一体化数据中心的能耗将超过百万瓦。需要持续优化数据中心的能效，才能满足各国的建设要求：包括去空调，去冷机，液冷技术普遍应用，围绕液冷的余热回收，如供热、二次制冷、发电等产业成为热点，新技术逐步完善并开始商用，PUE不断逼近1.0，甚至有望挑战小于1.0。随着芯片制程、封装技术的发展，AI、HPC（High-Performance Computing，高性能计算）等重算力芯片，芯片的热流密度超过150~200w/cm^2，开始出现原生液冷芯片。AI技术普遍应用，DC级从供电、制冷、到芯片工作模式，结合业务调度和业务负载特点的全栈自动化协同优化。



数据中心供电路径需要更短，更高效。2.5D、3D、WLC（Wafer Level Chip，晶圆级芯片）等新的封装技术使能的KA（Kiloampere，千安培）级芯片供电，需要有新工艺新器件新拓扑的创新。芯片超频大动态负载的功率波动，对服务器的供电设计带来挑战。液冷相对传统风冷在机房建设、服务器生产、交付、运维等流程和人员技能有更高的要

求。冷板、液冷工质等核心部件需要在加工工艺、可靠性等方面持续提升适应海量部署要求。



芯片3D封装的普遍应用，芯片封装内部温升增加，占散热全路径温升接近50%，对散热提出了更高的要求。TIM（Thermal Interface Material，热界面材料）材料、冷板热阻需要降低50%，依赖材料、工艺的创新。WLC等大尺寸芯片封装，在冷板装配强度，共面度，可靠性也提出了挑战。进一步的散热解决方案是芯片封装技术和液冷技术的融合，去掉TIM层，液冷工质进芯片封装和DIE（裸片）直接接触，带来DIE表面的强化散热处理，射流均流，长期冲刷腐蚀，封装密封等可靠性的挑战。

余热回收效率和水温强相关，而芯片散热、性能的考虑，要求水温又不能很高，如不超过65℃，低水温对余热回收系统在数据中心场景应用提出了更高的要求。余热二次制冷有望在2025年内规模应用，而余热发电当前效率小于5%，规模应用依赖关键技术的突破，如高ZT值新型发电材料等。同时余热回收需要稳定的热源，而液冷回水温度和芯片负载相关，需要结合业务调度，负载控制，液冷流量控制为余热回收系统提供稳定水温的热源。

DC级全栈的能效优化，需要DC内冷塔、水泵、CDU（Coolant Distribution Unit，冷液分配装置）、UPS（Uninterruptible Power Supply，不间断电源）、电表、服务器等需要开放状态监控和集中控制的接口，制定相应的接口规范。

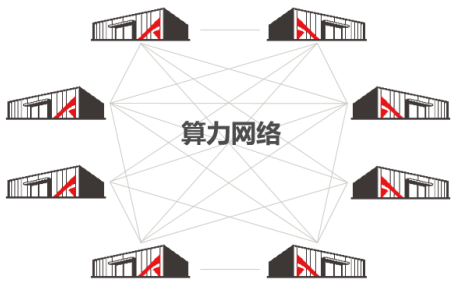
灵活可变的硬件配置：业务种类多样化，处理器平台多样化，未来云计算/100E级超算数据中心IT资源的规模和复杂度都将大幅增加；从服务器为粒度的交付演变到以资源部件为粒度的交付方式，资源有效使用率从当前30%提升到70%；为了配合自动化运维和部件化供应，需要对硬件形态和软硬件接口制定规范。

自动智能的设备运维：中心机房百万级服务器规模，自动化可以数量级提升交付/运维的效率和准确度；庞大数量的边缘节点，集成自动化大幅降低人力和运营成本，提升故障处理能力；基于AI与大数据的复杂系统优化决策，自学习+自动化高效动态调整软硬件资源的配置和部署，提升IT资源和能源效率；疫情等突发事件都要求未来机房具备非接触式的交付和运维；随着工业4.0和AI的发展，自动化技术在加速成熟；智能无人的自适应数据中心（Adaptive DC）将开始逐步推广，实现DC与业务的无人、动态匹配。

算力网络

1）跨地域的超级分布式数据中心

算力网络的核心思想是通过新型网络技术将地理分布的算力中心节点连接起来，动态实时感知算力资源状态，进而统筹分配和调度计算任务，传输数据，构成全局范围内感知、分配、调度算力的网络，在此基础上汇聚和共享算力、数据、应用资源。



算力中心呈现多层次，多管理域的布局。不同的算力中心间存在巨大的差异性，从资源的角度看，部署的应用类型、保存的数据集、算力的体系结构可能不同；从管理的角度看，管理策略、计费标准、碳排放标准可能不同。因此，算力网络的建设须面对不同算力中心间的高效协同，算力、数据、应用可信交易与管理机制设计，缺乏一体化标准等挑战，最终构建成为开放的、高资源利用率、高能效的计算基础设施。

2）融合应用形成数字连续体

人工智能模型规模的不断提升，数据规模的激增以及科学计算对模拟精度与时效性需求的不断提升，一方面带来算力需求的激增，另一方面也在推动应用的变革。未来的分布式应用，将融合实时与非实时数据处理，模型的训练与推理、仿真与建模、物联网、信息物理系统等一起形成了“数字连续体”，解决的单算力中心无法解决的问题，例如：结合了神经网络与实时数据的数字气象模型，可以提供高频率、高分辨率的短临天气预报，为国民生产生活提供保障；分布式的大模型利用多个算力中心的资源加速模型的训练过程。新应用程序的出现，将促进算力中心之间，以及算力中心与边缘计算的连接；算力中心将不再是独立的系统，而是形成相互联结的算力网络，多个组织的用户在多个算力中心共享算力和数据，完成复杂应用对计算和数据处理的需求。

3）跨域算力中心协同调度

地理分布的多个算力中心将联结在一起，为新型分布式融合应用提供支撑。超大模型的训练可能需要协同多个算力中心的资源完成，复杂的融合应用可能利用不同算力中心的多种算力与数据集协同完成。应用的差异性、算力中心资源的异构性以及不同管理域的管理策略将给调度系统带来新的挑战。调度系统需要感知应用所需算力与存储资源，感知应用所需数据的所在位置以减少数据移动开销，感知应用的通信模式以减少通信开销；调度系统还需要实时地感知不同算力中心资源的可用性与异构性，算力中心间的网络状态；此外，由于不同算力中心的资源定价、碳排放等标准的差异，调度系统还需要在性价比与能效比的约束下作出最优决策。需要调度系统具备全局的资源的发现能力、感知应用特征、感知算力中心的软硬件异构性，具备感知局部管理策略的能力，从全局视角，获得计算效率、数据移动效率与能耗效率的最优。

芯片工程

1）2.5D Chiplet芯片封装集成技术持续提升芯片算力和产品竞争力

传统芯片受wafer（硅片）曝光尺寸限制（1 Reticle：25mm*32mm），芯片Die的尺寸及Die良率提升受到严重技术瓶颈，直接制约芯片整体性能提升及芯片成本降低。

2.5D Silicon/FO Interposer+Chiplet技术可以有效提升Die良率、降低芯片成本，堆叠集成实现更大规模芯片性能，且对于不同产品规格应用更加灵活。同时2.5D封装性对于传统封装板级互连方案单bit能耗降低至约1/2。

基于行业发展与超大规模芯片需求，预计2025年2.5D silicon/FO interposer 尺寸将超过4xReticle，未来封装substrate（基板）预计会超过110mm*110mm。更大尺寸的2.5D与substrate应用直接面临良率、交期、可靠性等一些列工程难题，融合创新基板架构需求迫切。

2）3D芯片技术在芯片性能方面的综合表现远高于传统架构，预计提升数十倍

与传统2D/2.5D先进封装及异质集成芯片

技术相比，3D芯片技术在互连密度及带宽、芯片尺寸、功耗性能、芯片综合性能方面优势显著，是解决未来高性能计算、AI等关键场景芯片与系统集成的核心技术。

3D芯片技术未来会从D2W（Die-to-Wafer，芯片到晶圆）->W2W（Wafer to Wafer，晶圆片对晶圆片），uBump->Hybrid Bonding->Monolithic 3D技术逐渐演进，应用场景将会广泛覆盖3D Memory on Logic、Logic on Logic及Optical on logic等，并且未来会逐步走向更多层异质堆叠。

3D芯片在堆叠工艺方面需要采用小于10μm甚至更小pitch超高密Bonding技术，3D芯片相对于传统2.5D封装在带宽及功耗性能优势显著，单bit功耗降低有望降低至1/10。更小尺寸TSV（Through Silicon Via，硅通孔）技术需要从材料、工艺基础技术深入持续探索；同时3D堆叠带来局部功耗密度和电流密度倍增，直接影响系统整体供电与散热路径。

3）芯片出光，实现T级高带宽端口

高算力芯片(如xPU、Switch、FPGA等)的IO带宽越来越高，预计2030年，端口速率将达T级以上。随着单路速度提升，100/200G Gbps以上的高速串行通信带来功耗、串扰和散热挑战，传统光电转换接口将无法满足算力增长需要，芯片出光相比传统方案端到端能效有望降低至1/3。光电转换芯片和主芯片共封装(Co-packaged Optics)，替代可拔插光模块(Pluggable Optics）和板载光模块(On-board Optics)，芯片出光成为未来突破带宽瓶颈的关键技术。同时芯片出光面临PIC（Photonic Integrated Circuit，光子集成电路）与EIC（Electronic Integrated Circuit，电子集成电路） 3D封装，超大封装基板及OE（Optical Engine，光引擎）集成，单芯片功耗密度提升等一系列工程技术挑战。

4）大功耗芯片供电技术探索

算力需求与Chiplet技术持续推动芯片功耗提升，千瓦级芯片供电已经在望，万瓦级Wafer level芯片需要更加创新及高效的供电策略。高压单级变换、开关电容混合变换等新型供电架构配合低压氮化镓（GaN）功率器件和高频集成磁等工程技术的应用，可以进一步提升单板供电的端到端能效和功率密度。

芯片48V高压直供是解决芯片供电瓶颈的关键技术路径。基板、封装承受高压的材料研究与工艺改进是芯片高压直供的前提，同时高效的片上电压转换技术与分核供电技术也是关键研究方向。

5）未来芯片层面散热技术探索

随着计算芯片功耗的急速上升，散热已成为制约芯片性能提升的主要瓶颈之一，新型散热技术及材料亟待开发。通过开发高导热TIM1材料降低路径热阻，Lidless（无顶盖封装）芯片散热、封装与芯片级先进液冷技术，有望为未来芯片提供千瓦级与万瓦级散热能力，为芯片性能的跨越式提升提供散热基础。芯片动态热管理技术与整机系统散热协同设计也是未来超大功耗芯片散热关键设计技术。

多样性计算

未来的计算，数据将在最合适的地方，以最合适的算力来处理，例如网络数据在DPU（Data Processing Unit，数据处理单元）上就近被处理，神经网络模型在NPU上训练；算力无处不在，硬盘、网卡、内存等外设开始逐渐具备数据分析和处理能力。融合应用呼唤多样性计算的统一架构出现；当前各厂商工具的烟囱化，严重制约了多样性计算的发展。

数据为中心的计算

1）对称计算架构（数据全内存处理）

冯诺依曼的经典架构，需要把数据搬移到CPU进行处理，这种数据搬移消耗了大量的系统算力和能量，而且数据在处理和交换过程中，存在着大量的反复的内存格式，存储格式，传输格式的各种转换，这种格式转换消耗大量CPU时间，而且能率很低；同时受到硬件发展的制约，而数据爆发凸显了IO（Input/Output，输入输出），算力，网络等瓶颈，这些瓶颈都影响数据搬移的速度和处理效率，影响整体的系统能效。

对称计算架构通过内存池化，在数据全生命周期使用统一的内存语义进行数据处理和交换，甚至数据全在内存中进行处理。该架构可避免数据格式的转换，提高数据的移动速度，扩大应用的可使用内存，从而极大的提升整体

系统数据处理能力，是未来提升计算效率的重要路径。实现该架构需要在多层级内存架构、大容量非易失性内存等关键技术上突破创新。

2）泛在计算（外设智能化）

未来除了xPU各类算力之外，我们认为计算架构将走向泛在的近数据计算，数据在最合适的地方，以最合适的算力来计算，减少数据搬移，提高整体系统的性能。泛在近数据计算包括以下几个方向：

近内存计算，当前的瓶颈在于数据搬移的有效带宽受外部总线带宽约束，未来通过在DRAM（Dynamic Random Access Memory，即动态随机存取存储器）的控制电路上增加多并发的可编程计算单元，同时优化DRAM阵列结构提升内部访问数据的并发度，实现DRAM内数据运算有效带宽的倍数级提升，打破内存墙造成的数据带宽瓶颈；

近存储计算，当前的方式是在SSD（Solid State Drive，固态硬盘）控制器上增加固化的数据加速单元（如压缩引擎）实现单一的数据处理功能，未来将演进到通过API（Application Programming Interface，应用程序接口）编程接口灵活调用SSD控制器内多种算子引擎，配合编译器实现更为灵活的算力卸载，在通用场景下大幅提升数据运算的能效比；

从基于SmartNIC（智能网卡）的在网计算演进到基于DPU的以数据为中心的计算架构，未来将实现灵活的可编程在网算力、开放的异构编程框架、业务驱动的在网计算范式。支持对存储、安全、虚拟化等的全面加速，支撑HPC+AI融合、大数据、数据库等分布式应用性能倍增。未来将进一步实现对整个DC计算资源的细粒度动态调度、高效交互。

3）存算一体

存算一体是计算单元和存储单元紧耦合的一种方式，即存储介质既能做存储单元又能做计算单元，打破算力和存储的边界，有效改善功耗墙和内存墙，相比传统冯诺依曼架构有着预计十倍以上的能效提升。

基于SRAM（Static Random-Access Memory，静态随机存取存储器）、NOR Flash（非易失闪存）等成熟存储器实现的存算一体，将有望在2-3年内规模商用，在端侧、边缘侧的人工智能推理运算中展现出10倍能效优势。基于ReRAM（Resistive random-access memory，可变电阻式内存）、PCM（Phase Change Memory，相变存储器）、MRAM（Magnetoresistive Random-Access Memory，磁性随机存储器）等新型非易失存储器的存算一体还在探索中，因其具有高性能、低功耗的特点，未来十年有望在数据中心侧实现突破。

存算一体大规模应用还需要在以下方向突破：

计算精度：由于器件的一致性、稳定性导致的误差，以及计算过程中存在的噪声，使得存算一体的精度相比数字计算有一定下降，需要结合电路特征优化算法，使得计算结果满足应用需求。

软件生态：存算一体是一种数据驱动的

计算，需要将神经网络模型部署在合适的存储单元上，并通过数据流调度来控制整个运算过程，需要更加智能、高效、便捷的数据映射工具将不可或缺。

体系架构：当前新型非易失存储器的存算一体主要是基于向量乘矩阵的计算方式，常用于特定的机器学习应用（如神经网络推理、训练），难以扩展到其他的应用场景，且无法与现有的存储系统配合，进行数据的高效处理。未来需要突破从存储器件到编程模型，再到系统架构和应用的“全栈协同设计”，使得存算一体架构走向通用。

4）计算总线从板级走向DC级

随着算力和数据的成倍增长，以AI、HPC和大数据为主要业务的大型集中数据中心成为发展趋势。而连接整个数据中心的网络，相比节点内总线有巨大的时延、带宽“鸿沟”和厚重的网络软件栈开销，制约了算力的发挥。“内存语义”总线将高带宽、低时延和内存语义的轻量软件栈，从板级平滑扩展到全数据中心，实现全数据中心性能和能效比最优。

“内存语义”总线的最大挑战在于构建开放、平等、互通互操作性的总线、接口和协议标准，避免计算系统总线走向7国8制，限制计算性能发挥和规模构建。

应用驱动的多样性计算

以特定领域专用硬件、特定领域编程语言、开放式架构、原生安全架构为代表的新计算范式将会成为下一代计算系统的主流。

1）智能科学计算（HPC+AI）

AI计算方法和AI算力架构持续突破，将机器学习与基于第一性原理的物理建模相结合的智能科学计算方法，正成为科学研究的一个新范式。未来十年，智能科学计算将深入到科学研究和技术创新的各个方面。如何将AI算法与科学计算高效融合，面临前所未有的挑战和机会。

在基本层面，面临新计算模式的计算框架和数学方法挑战。新的计算框架和方法，首先需要明确给定的问题是可通过AI方法被有效地解决。即，计算数学方法及框架需要满足可计算性、可学习性、可解释性。与此相应，在未来十年，软硬件基础设施也必须以数学和AI研究为基础，提供合适的实施、评估和测试体系。

在数据层面，通过AI方法加速科学、工程和制造需要大量不同的数据源。一，当前，不同领域科学问题来自仪器、模拟、传感器网络、卫星、科学文献和研究成果的数据源，在数据可获得和可共享性具备较大挑战。二、利用AI方法产生有效的从物理原理出发和符合物理基本定律（比如对称性和守恒定律）的数据。这个挑战，需要领域科学家、AI专家、数学家、计算机科学家广泛的协同设计工作来跨越。

2）AI使能存储智能化

存储系统需要承载的业务诉求也越来越多样和复杂，既需要应对不断变化的多样化业务负载，又需要简化系统管理运维。

未来的存储系统需要基于AI实现主动管理和响应其内外部环境、持续学习、感知负载自适应响应、自动优化系统等智能化功能，获得资源分配、成本、性能、可靠性、易用性、功耗等综合收益，同时运维方式也需要基于AI从传统的人工运维向免人工智能运维逐步演进。

目前AI技术应用在存储系统的索引管理、自动调优、资源分配等方向已经取得一些进展，但仍需在以下四个方面进行突破：

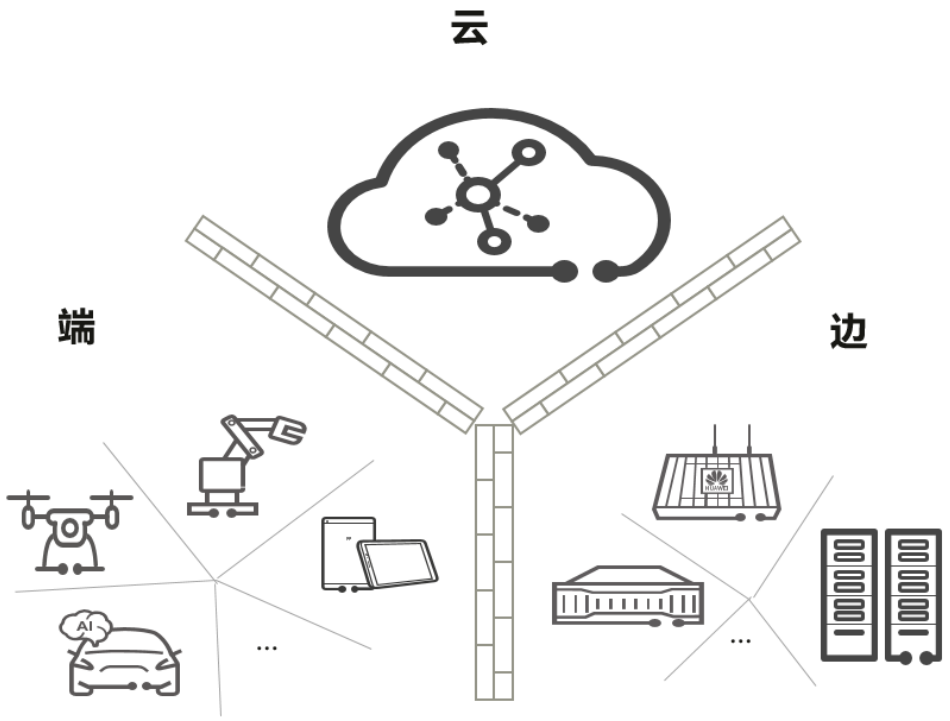
负载域：从系统性能维度对IO 负载进行建模，分解出影响系统模块性能的关键指标及因素，精确评估系统性能和模拟用户真实业务场景；

数据域：感知数据布局、感知数据生命周期和感知数据内容上下文等信息，提升数据访问性能、降低系统后台垃圾回收资源消耗和提升数据缩减率；

系统域：捕捉历史规律和模式、高效安置和调度计算任务、进行运行时优化，优化系统参数和资源配置、降低系统能耗、保证系统性能波动可控且不影响可靠性；

运行域：实现免人工运维、自动故障根因分析、系统亚健康检测自动预防与修复。

综合自顶向下的负载建模和自底向上的系统自学习技术使能存储智能化，实现具备性能自动调优、服务质量自动化控制、数据智能感知、规则与策略自学习、智能调度、低功耗控制、极简规划和配置、系统问题提前预测、故



障根因自动分析的智能存储系统将成为重要的研究方向。

多维协同

多种计算、存储等设备分布在云、边、端不同的位置，将这些设备横向及纵向进行协同与协作，实现优势互补，形成立体计算。解决业务体验不好、算力分布不均、算力利用率低、信息孤岛等一系列的问题与挑战。

通过多维感知与数据建模技术，物理世界被镜像、计算、增强，形成孪生的数字世界；利用光场全息渲染、AI内容生成等技术，实现数字世界到物理世界的精确映射。结合时间与空间、虚拟与现实的多维协同，实现物理世界与数字世界的无缝融合。

立体计算

1）边缘计算

未来是万物互联的智能世界，随着5G技术的成熟与应用，边缘计算开始在ICT行业广泛部署，预期2030年全球市场规模将从100亿美元增长到数千亿美元，市场潜力巨大，影响边缘计算大规模应用的主要问题与挑战包括：边缘智能、边缘算力网络、边缘安全、边缘标准与开放生态等。

边缘智能：制造、电力、城市、交通、金融等垂直行业的智能化升级与改造，是边缘计算在这些行业规模应用的重要驱动因素，将带来爆发式的增长。需提供增量学习、迁移学习、硬件亲和的模型压缩、推理调度部署等AI基础能力开发套件，来解决跨行业的智能化共性问题；以及面向制造行业的复杂背景、弱对比、小样本、弱监督等应用特征提供开发套件，来解决智能制造的共性问题，其它行业依次类推。进而形成一整套功能完备的应用使能SDK（Software Development Kit，软件开发工具包）。

边缘算力网络：边缘设备因未来业务发展多样化的诉求，逐渐向小型化、移动化、低功耗的方向发展，算力、存储、带宽、时延等越来越成为瓶颈。全息及多维感知类业务对算力提出至少100倍于当前能力的要求，对存储提出100倍乃至1000倍于当前能力的要求，对网络带宽的诉求高达到10Tbit/s级别；智能制造、智慧电力、智能交通等行业基于自身的业务特点提出了毫秒级时延及确定性时延的要求。为了满足边缘加速、卸载和突破性能瓶颈的诉求，要求进行计算、存储、网络的协同与超融合，以及多样算力的有效利用，对边缘软、硬件架构带来新的挑战。

边缘安全：边缘设备在物理位置上通常离

攻击者比较近，所处环境复杂，更容易遭到来自物理硬件接口、南北向业务接口、北向管理接口等的攻击。数据往往是用户的核心资产，丢失或被窃取可能使用户遭受重大损失。预计2030年将有80%的数据在边缘进行处理，需加强在边缘进行数据采集、存储、处理、传输过程中的安全与隐私保护；严格保护边缘应用、模型等核心资产的安全与隐私；避免因数据隐私保护形成数据孤岛，导致数据与AI算法在医疗、金融、工业等领域的潜在价值无法充分发挥。

边缘标准与开放生态：面向不同行业应用的边缘设备在软硬件形态、算力、功能、接口等方面差异巨大，各厂商提供的私有软、硬件方案及接口协议，相互之间难以兼容互通，很大程度上影响了边缘计算的推广与普及。需要将边缘计算系统与软硬件框架，及相关的接口与协议标准化，并建立对应的测试验收标准，以促进边缘设备、软件与协议的兼容互通。同时面向各个行业建设开放生态，吸引更多厂商与合作伙伴的投入来共融共建。

2）多设备协作

蚂蚁、蜜蜂等生物群体通过个体协作产生集体智能，多设备协作技术的目标正是寻求类似突破以提升多设备所形成系统解决问题的能力、整体性能、鲁棒性等。

多设备协作技术存在任务分担、结果共享、智能体等多种模式。任务分担模式是设备之间通过分担执行整个任务的子任务而相互协作；结果共享模式是设备通过共享部分结果相互协作，各设备在任何时刻进行的处理取决于当时该设备自身拥有或从其他设备收到的数据和知识；智能体模式是每个设备在独立性和自主性基础上的相互协作。

多设备协作技术面临多设备之间的合作与冲突消解、全局最优化、交互协作一致性等挑战。

合作与冲突消解：多设备协作过程中可能导致死锁或活锁，死锁使得多个设备无法进行各自的下一步工作，活锁使得多个设备不断工作却无任何进展，如何在交互过程中避免死锁或活锁，协调的机制和算法是系统的核心挑战。

全局最优化：多设备间根据局部信息的协作难以达到协作的全局最优，采用全局视野的协作往往意味着通信量大，会给系统带来沉重的负担。如何高质量、高效率、高可靠、高安全取得全局环境的态势估计以此进行多设备的协同规划和协调，决定了多设备协作的效率与效果。

交互协作一致性：各设备通过网络通信获取其他设备信息，并以此调整自身的状态，实际系统中由于多设备间的通信连接不可靠或通信存在限制，如何解决由通信不确定性带来的协作一致性问题，决定了多设备协作系统的鲁棒性。

多设备协作技术强调多个设备之间的紧密群体协作，协作系统将从简单的合作与连接逐渐发展成独立自主的群体智能系统。

3）端边云协同

智能制造、智能城市、智能巡检、智能交通等AI和新兴数据密集型应用在快速发展，低时延响应、节约带宽成本、保护数据隐私安全等应用体验驱动计算向端边云协同发展，要实现一体化的计算架构，面临如下挑战。

任务协同：如何进行合理的计算任务划分将应用分割为多个子任务，并且进行子任务在端边云的部署与调度，比如子任务在端、边、云何处执行，何时执行，以及计算子任务跨云、跨集群、跨节点如何迁移均充满挑战。

智能协同：“云端训练、边缘推理”的模式正在走向端边云“合作式”的训练和推理，如何解决协同训练的精度和收敛速度问题，如何解决协同推理时延和准确率问题，如何解决端边云协同智能中存在的数据孤岛问题、小样本问题、数据异构问题、安全隐私问题、通信成本问题、端/边设备的资源受限问题等。

数据协同：数据是智能的基础，数据的接入、聚合、交互、处理面临着多样化和异构的挑战。

网络协同：随着端边云计算网络的规模越来越大，大量设备及子网的接入带来设备、网络、业务管理的巨大挑战，如何确保联接的实时性可靠性是必须要解决的问题。

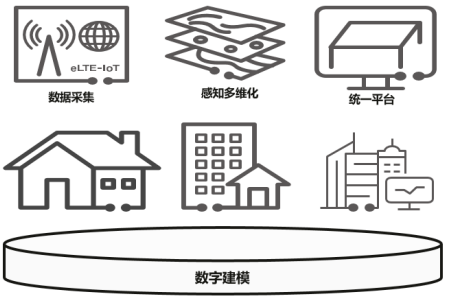
安全可信挑战：边缘侧设备和产生的数据接入云端的安全和隐私如何保证，云端如何抵御来自边缘侧的攻击，云端下发到边缘侧的数据如何保证安全等。

数字孪生

1）统一数字孪生平台成趋势

在智慧工厂、智慧城市、虚拟社交媒体等各行业数字化浪潮之下，缺乏一套能够创建富有个性化数字孪生系统的统一平台。该平台需重点关注三维模型的数据格式、开发工具等的统一，能够提供多样性算力及存储空间以满足大量数据建模的需要。

2）多维感知与数字建模技术



未来的物理世界将会有一个孪生的数字世界，数字世界和物理世界无缝的衔接、协同，以提升产品设计、产品制造、医学分析、工程

建设等领域的效率。物理世界到数字孪生的映射过程将面临感知多维化、三维建模、光场采集数据存储等多方面的挑战。

感知多维化：物理世界里影像、视频、声音、温度、湿度、力学等各种数据经采集、存储后数据量非常庞大。更多维数据的获取、处理与融合，需要高分辨率的感知、定位、成像和环境重构能力，形成的数据量更加庞大。这些海量数据的筛选、预处理、建模、仿真等过程都依赖于强大的算力，以及人工智能、认知科学、控制科学、材料科学等多学科的深度融合。

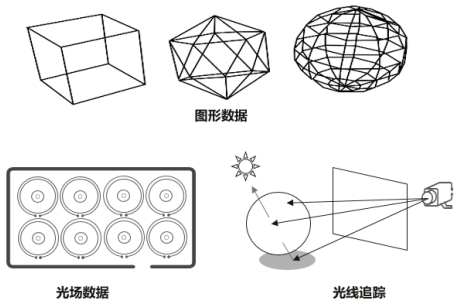
三维建模算力需求增加100倍:根据不同角度根据不同角度的图片与视频流，以及阵列相机、深度相机等采集的海量数据进行三维建模需要强大的算力。使用100+路摄像机阵列采集的高精度数据的数据量，比传统2D图像数据量增加100倍以上，分辨率提升到8K，单路算力提高4倍，所需建模算力也增加100倍以上。管理多维海量数据，并将之转化为三维模型面临巨大挑战。同时消费市场可通过手机3D相机获取影像的深度信息，并根据深度信息在手机端完成中低精度的建模，3D相机通常是双目、结构光或者ToF（Time of Flight，飞行时间）相机。需要提供一套统一、高效、经济的三维建模软硬件系统，来同时满足高阶与消费级建模的诉求，和促进各行业的数字化转型及数字孪生产业的繁荣。

基于AI技术的数字建模材质生成：未来，基于AI影像辨识技术、智能生成算法及强大的AI算力，自动辨识图片中材质的金属性、粗糙度、反射率、折射率、表面法向量等物理特性，并协同三维模型生成现实生活中的材质。面向未来，需要建立一套统一、开放的材质描述语言，从而实现不同行业3D图形数据的交换。

光场数据增长百倍压缩技术成关键：光场相机阵列采集的图片与视频流数据增长100倍，基于光场数据合成三维视频流，以及渲染的光线着色等，数据的存储与处理都存在巨大瓶颈，光场数据的快速压缩与存储相关的技术突破将成为后续渲染与成像的关键。

3）光场全息渲染技术

具有真实世界感官体验的数字孪生显示系统，需要在视觉、互动技术上进行突破。目前多数产品在渲染质量、逼真度、渲染时延上还不能满足要求。实时光线追踪、零时延传输是达成现实级逼真渲染效果的关键技术，直接影响用户体验。高阶渲染光线追踪相比传统渲染算力需求增加10x以上，以存代算技术可有效缓解算力需求的矛盾，同时可降低时延，但需要更大的存储空间。基于云的光场全息渲染技术将成为未来的重要技术方向。



高阶渲染技术，分辨率提升64X:光场全息渲染的主流技术从光栅化渲染，逐渐向光线追踪等高阶渲染技术发展。在游戏、XR（Extended Reality，扩展现实）等场景，要实现逼近现实的体验，达到双目16K分辨率、120FPS帧率、8ms时延，强交互场景对时延要求标准更高达5ms，算力需求提升64倍以上，需要突破三维建模、材质生成、光线存储等关键技术。同时依赖跨端边云计算集群的渲染、AI、视频流化的融合算力，以及面向高阶渲染的内容制作软件的突破，从而实现近实时、高性能的整体渲染解决方案。

基于AI技术的内容生成：基于AI技术实现3D模型构建、材质自动生成、超分、降噪等。基于GAN、NLP（Natural Language Processing，自然语言处理）与NLG（Natural Language Generation，自然语言生成）等AI技术，实现逼真的数字人3D成像、表情与真实的语言对话，让世界各国的人们可以完成高效的沟通和交流。AI内容生成还可应用于工业设计、XR内容创作、影视特效制作等。

4）亿级用户虚实协同与交互

亿级用户在数字世界与物理世界的协同、

联动和同步，对算力、存储、网络挑战极大，大量的状态查询与消息传送，如何满足人和物两两之间交互时延小于5~10ms，单用户数百Mbps带宽、单用户数十Tflops算力，网络与端云协同，亿级用户数据实时处理与传输，将面临巨大挑战。

物理层突破

学术界、工业界都在寻求物理层突破，通过探索模拟计算、非硅基计算、新型存储器以及优化芯片工程技术，在未来继续提升计算能效和存储密度。例如：量子计算在数据表达和并行计算能力上具有指数级优势，模拟光计算在特定计算中展现出低能耗和高性能；二维材料和碳纳米管具有载流子迁移率高、沟道短的特点，有望成为替代硅基的新材料；铁电、相变材料和器件结构取得较大突破，存储密度和读写性能大幅提升，多层多维的光存储在冷数据长期保存上有较大潜力；未来还有DNA存储等有待突破。这些物理层关键技术的不断突破，将对计算和存储领域带来革命性改变。

模拟计算

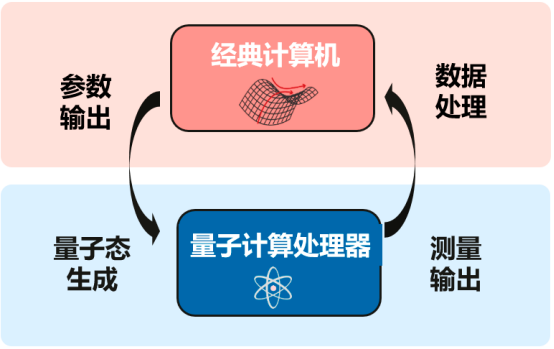
量子计算：量子计算是未来高性能计算的必争之地

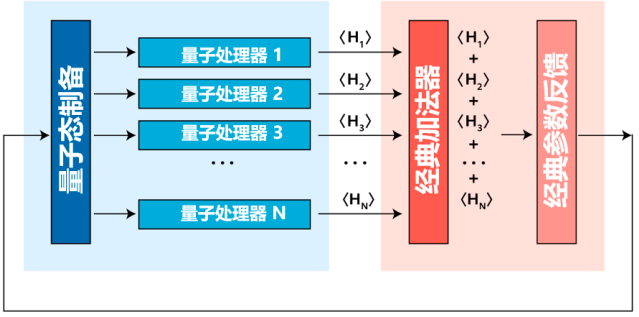
量子计算目前处于高速工程化的阶段，预计未来五年将出现超过1000比特的量子芯片。目前量子计算处于含噪声的中等尺度量子（NISQ）时代，基于精确计算的经典计算机与高性能量子计算机，构建混合计算架构是最具可行性的技术方向。其中量子化学模拟、量子组合优化算法及量子机器学习三大方向是最具商业价值的落地场景。量子化学模拟能为药物研发与新型材料研发提供新算力；量子组合优化算法把组合优化问题编码为量子计算过程，能更快更好的解决物流调度、行程规划及网络流量分配问题；量子机器学习将作为人工智能计算加速的新路线。

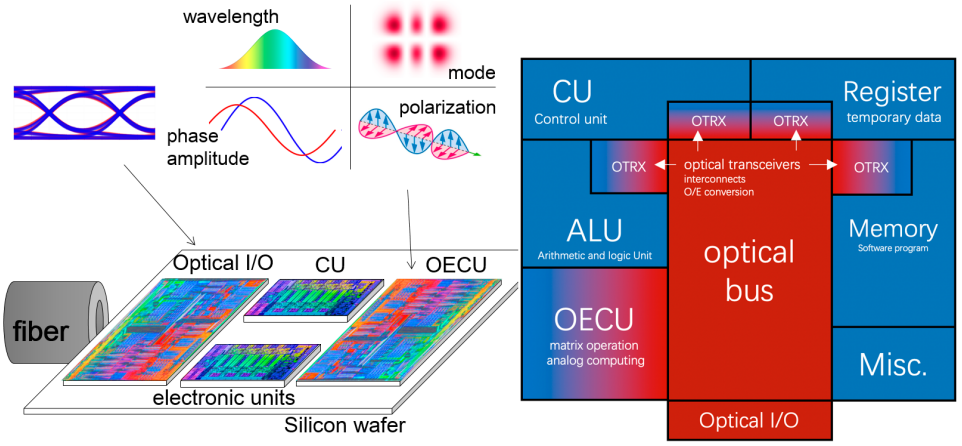
未来十年重点是实现基于NISQ的专用量子计算机，需要不断提升单量子芯片的物理比特规模，增强相干时间和保真度，并通过量子芯片的互联提升系统的扩展能力，获得解决复杂问题的算力；同时增强量子计算的容错设计，提升系统可靠性，结合应用场景不断优化量子算法，降低线路深度和复杂度，完善量子软件栈，逐步推动NISQ量子计算走向商用。但要实现一台通用量子计算机，道路更加漫长、更加充满挑战。

模拟光计算：模拟光计算将在部分复杂计算中展现优势

光的传播速度快、能耗低，其干涉、散射、反射等物理现象背后，都有对应的数学模型，通过对光信号的调制、控制、探测，可完成某些特定的计算任务。同时光作为玻色子天然具有波分复用、模分复用、OAM（Orbital







Angular Momentum，轨道角动量）复用等特性，通过模拟光计算实现多维度并行，是未来光计算发展的重要方向，有望在卷积计算、伊辛模型求解、蓄水池计算等领域率先突破，并成为光信号处理、组合优化、序列比对、AI加速等场景的利器。

光计算要实现规模应用，首先需要解决有源器件、无源器件在芯片上的异质集成问题，提升光信号耦合效率、控制插损和噪声，满足特定应用场景的计算精度要求。另外，光计算的驱动电路也需要进一步与光芯片集成，降低功耗和面积。光计算和电计算各有优势，光电混合的计算架构是未来发展的重要方向。

非硅基计算

二维材料：二维材料有望成为延续摩尔定律的终极材料

二维材料晶体管具备沟道短、迁移率高、可2D/3D异质集成的优势，有望作为晶体管沟道材料延续摩尔定律至1nm节点。此外具有超低介电常数的二维材料，也可以用作集成电路的互连隔离材料。二维材料有望首先在光电、传感等领域应用，最终在大规模集成电路和系统中实现应用。

当前二维材料及其器件仍处于基础研究阶段，需要从材料、器件、工艺等层面突破。未来五年，首先需要解决工业级二维材料晶圆制备的产业化良率问题；其次要不断改善电极和器件结构，提升二维晶体管器件综合性能；在此基础上，未来十年有望大规模集成电路产业实现应用。

碳材料晶体管：碳基电子学可能是未来最有希望延续摩尔定律的技术

碳纳米管具有超高的载流子迁移率、原子级的厚度，具有高性能、低功耗的巨大优势。在尺寸极端缩减的情况下，碳管晶体管能效比硅基晶体管提升约10倍，3~5年内有望在生物传感、射频电路实现商用。

未来五年还要继续改进碳管材料的制备工艺，降低表面污染和杂质，提升材料纯度和碳管排列的一致性；优化器件接触电阻和界面态，提升注入效率；配套EDA（Electronic Design Automation，电子设计自动化）工具的开发；通过小规模的集成电路验证碳基半导体端到端的成熟度，有望在柔性电路领域初步得到应用。展望未来十年，当碳基半导体器件的尺寸能够微缩到与硅基先进工艺相当水平时，在高性能、高集成度的应用场景中，将迎来规模应用的机会。

新型存储

传统存储以磁介质为主，新型存储全闪存将成为主流，预计未来将有72%的企业存储基于全闪存。全闪存不仅用于主存储（primary）存储，还将延伸到辅助存储（secondary）存储，预计，企业将会有82%的业务数据存在备份需求。围绕着数据全生命周期的热温冷差异，未来介质也将向高速高性能和海量低成本两个方向演进。

1）新型内存型介质技术

当前热数据存储在SSD中，搬移到DRAM中处理，SSD时延与DRAM相差1000倍，而DRAM受物理特性的限制，密度和电压都已无法继续扩展，所以SSD和DRAM都无法完全满足热数据存储的需求。目前业界已经涌现了许多新型内存型介质技术，如PCM、MRAM、ReRAM、 FeRAM（Ferroelectric Random-Access Memory，铁电式随机存取内存）等。这些介质在性能、容量、成本、寿命、能耗、可扩展性等各方面都将优于DRAM，支持字节级访问和持久化，不需要再进行数据搬移，将成为热数据存储的主流介质，但面临如下技术挑战：

容量的挑战：到2030年，热数据总量将相当于当前SSD存储数据的总量，热数据介质的容量密度至少需要扩大十倍左右达到当前SSD的1Tb/die，还要支持按需扩展，不受处理器、内存接口、网络时延和带宽的限制。而FeRAM、ReRAM和MRAM等介质则面临着结构和材料等方面的挑战。

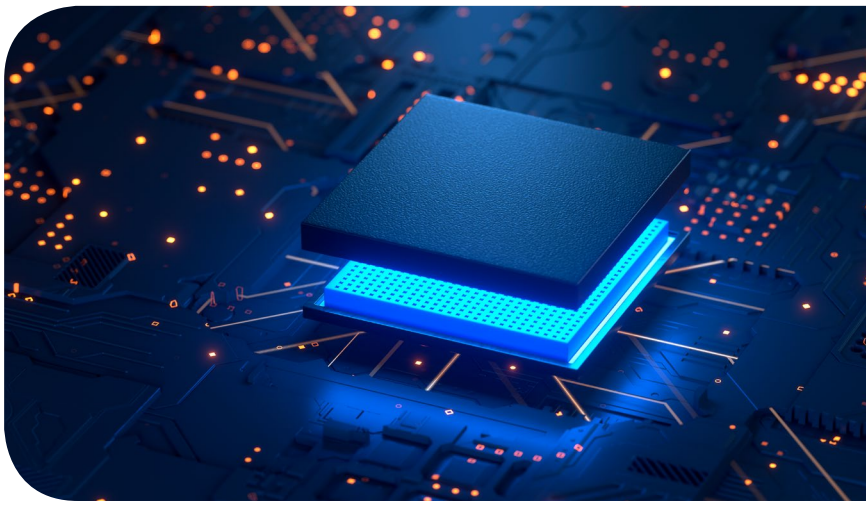
能耗的挑战：在“碳中和”的背景下，作为海量热数据的存储介质，面临功耗的巨大挑战。PCM、ReRAM等基于电阻的数据存储技术，数据写入电压更高，功耗更大。ReRAM和MRAM的单位bit功耗是FeRAM的10倍，而PCM更是高达100倍，FeRAM类低工作电压介质潜力更大。

2）高密NAND Flash介质技术

未来大部分热数据需要从温数据中产生，温数据成为热数据最大的“蓄水池”，所以温数据介质需要兼顾性能、容量和低成本。NAND作为温数据的主存储介质取代HDD（Hard Disk Drive，硬盘驱动器），向Cell多值（1个存储单元存储多个bit）和3D堆叠方向演进；在保持性能和寿命与当前QLC（Quad-Level Cell，四层式存储单元）相当的前提下，实现容量扩展和成本下降是最大的挑战：

Cell多值的性能和寿命挑战：Cell每多存储一个比特，表示数据的电压级数将增加一倍，读写性能和寿命下降数倍。

3D堆叠的工艺挑战：预计2030年堆叠层数将从当前的百层量级达到千层量级，介质硅通孔宽深比将达到120比1（或提升1倍），带来巨大的加工难度。

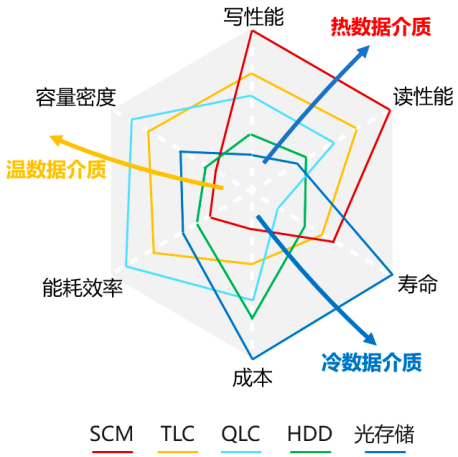


3）光存储技术

未来冷数据长期存储规模将从1.2ZB增至26.5ZB，同时存储寿命需要提升5~10倍。以中国国家档案馆为例，关键档案数据的存储寿命要从100年提升到500年，数据规模将从100PB增长到450PB。传统的硬盘和磁带将无法满足需求，随着对石英玻璃、有机玻璃等透明体材料读写原理及编解码算法的研究，光存储将成为海量冷数据的主流存储介质。挑战如下：

1、介质寿命要提升十倍，且在寿命周期内能应对各种复杂恶劣环境。

2、与蓝光相比，容量要达到10倍，成本下降5倍，性能提升10倍。



计算 2030 倡议

计算在过去的半个多世纪中加速了科学进步和经济发展，已经深深融入了人类社会的方方面面，是全人类的共同财富，也是未来智能世界的基石。

面向2030年，计算将更加开放和安全，每一个人、每一个组织都能够平等的参与未来计算产业的构建和创新，共享计算技术创造的价值。

让我们共同努力，开创计算新时代！

附录

参考

[1]Zettabyte (ZB)，Yottabyte (YB)：数据存储容量单位，1ZB=10^{21}Byte，1YB=10^{24}Byte

[2]华为预测，2030年通用算力（FP32）3.3ZFLOPS，对比2020年增长10倍，AI算力（FP16） 105ZFLOPS，对比2020年增长 500倍；FLOPS：每秒浮点运算次数；EFLOPS：一个EFLOPS（exaFLOPS）等于每秒一百亿亿（10^{18}）次的浮点运算；ZFLOPS：一个ZFLOPS（zettaFLOPS）等于每秒十万亿亿（10^{21}）次的浮点运算

[3]参考中国工程院院士李德毅在首届中国智能教育大会上的讲话，2018

[4]中国《关于加快煤矿智能化发展的指导意见》2020.03

[5]欧洲核子研究中心CERN, https://home.cern/science/computing

[6]QM/MM: 组合量子力学/分子力学方法，在QM/MM方法中，一部分体系使用量子力学(QM, quantum mechanics)方法进行处理(非常耗时)，另一部分体系使用基于力场的标准分子力学(MM, molecular mechanics)方法进行处理

[7]Summit，美国橡树岭国家实验室超级计算机，算力148.6P FLOPS，2021世界排名第二

[8]Roland R. Netz, William A. Eaton, Estimating computational limits on theoretical descriptions of biological cells，PNAS 2021

[9]戈登贝尔奖，由国际计算机协会（ACM）颁发，旨在奖励时代前沿的并行计算研究成果，特别是高性能计算创新应用的杰出成就

[10]Weile Jia, Han Wang, Mohan Chen, Denghui Lu, Lin Lin, Roberto Car, Weinan E, Linfeng Zhang,Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning，2020

[11]DevOps，敏捷开发和开发运维一体化

[12]Forrester分析师约翰•金德维格在2010年提出零信任安全架构

缩略语

缩略语	英文全称	中文全称
3D	3 Dimensions	三维
AI	Artificial Intelligence	人工智能
API	Application Programming Interface	应用程序接口
AR	Augmented Reality	增强现实
BP	Back Propagation	反向传播
CDU	Coolant Distribution Unit	冷量分配器
CERN	European Organization for Nuclear Research	欧洲核子研究组织
CPU	Central Processing Unit	中央处理单元
CSP	Cloud computing Service Provider	云算力提供商
D2W	Die-to-Wafer	芯片到晶圆
DC	Data Center	数据中心
DNA	Deoxyribonucleic Acid	脱氧核糖核酸
DPU	Data Processing Unit	数据处理单元
DRAM	Dynamic Random Access Memory	动态随机存取存储器
EDA	Electronic Design Automation	电子设计自动化
EFLOPS	exa Floating-Point Operations Per Second	每秒浮点运算百亿亿次
EIC	Electronic Integrated Circuit	电子集成电路
FeRAM	Ferroelectric Random-Access Memory	铁电式随机存取内存
FPGA	Field Programmable Gate Array	现场可编程门阵列
GAN	Generative Adversarial Network	生成式对抗网络
HDD	Hard Disk Drive	硬式磁盘驱动器
HL-LHC	High Luminosity - Large Hadron Collider	高光度大型强子对撞机
HPC	High-Performance Computing	高性能计算
ICT	Information and Communications Technology	信息和通信技术
IO	Input/Output	输入输出
KA	Kiloampere	千安培
MM	Molecular Mechanics	分子力学
MR	Mixed Reality	混合现实
MRAM	Magnetoresistive Random-Access Memory	磁性随机存储器
NISQ	Noisy Intermediate-Scale Quantum	嘈杂中型量子

NLG	Natural Language Generation	自然语言生成
NLP	Natural Language Processing	自然语言处理
O2O	Online to Offline	线上到线下
OAM	Orbital Angular Momentum	轨道角动量
OE	Optical Engine	光引擎
PCM	Phase Change Memory	相变存储器
PB	Petabyte	拍字节，千万亿字节
PIC	Photonic Integrated Circuit	光子集成电路
PIM	Processing-In-Memory	内存内处理
PUE	Power Usage Effectiveness	能源利用效率
QLC	Quad-Level Cell	四层式存储单元
QM	Quantum Mechanic	量子力学
REE	Rich Execution Environment	富执行环境
ReRAM	Resistive Random-Access Memory	可变电阻式内存
SDK	Software Development Kit	软件开发工具包
SRAM	Static Random-Access Memory	静态随机存取存储器
SSD	Solid State Drives	固态硬盘
TEE	Trusted Execution Environment	可信执行环境
TIM	Thermal Interface Material	热界面材料
ToF	Time of Flight	飞行时间
TSV	Through Silicon Via	硅通孔
UPS	Uninterruptible Power Supply	不间断电源
VR	Virtual Reality	虚拟现实
W2W	Wafer to Wafer	晶圆片对晶圆片
Wafer Level	Wafer Level	晶圆级
WLC	Wafer Level Chip	晶圆级芯片
xPU	x Processing Unit	泛指各种处理器
XR	Extended Reality	扩展现实
YB	Yottabyte	尧字节，一亿亿亿字节
ZB	Zettabyte	泽字节，十万亿亿字节
ZT	Thermoelectric Figure of Merit	热电优值

致谢

计算2030编写过程中得到了来自华为内外部多方的大力支持，300多位来自华为的专家和社会各界知名学者参与了材料的讨论、交流，贡献思想、共同畅想了2030年计算产业的发展方向和技术特征，在此对所有参与技术交流和讨论的学者们致以诚挚谢意！
（学者名单按照姓名字母排序，不分前后）

André Brinkmann（美因茨大学，教授）

Bill McColl（前英国牛津大学教授）

陈文光（清华大学，教授）

冯丹（华中科技大学，长江学者特聘教授）

冯晓兵（中科院计算所，研究员）

甘霖（清华大学，副研究员）

管海兵（上海交通大学，长江学者特聘教授）

过敏意（上海交通大学，教授，IEEE Fellow，欧洲科学院院士）

Jaroslaw Duda（雅盖隆大学，助理教授，ANS压缩算法发明人）

贾伟乐（中科院计算所，副研究员）

金海（华中科技大学，长江学者特聘教授，IEEE Fellow）

金钟（中科院计算机网络信息中心，研究员）

缪向水（华中科技大学，长江学者特聘教授）

Onur Mutlu（苏黎世理工大学，教授，ACM&IEEE Fellow）

潘毅（中科院深圳理工大学，教授，美国医学与生物工程院院士，乌克兰国家工程院外籍院士，英国皇家公共卫生院院士）

舒继武（清华大学，长江学者特聘教授，IEEE Fellow）

孙家昶（中科院软件所，研究员）

田臣（南京大学，副教授）

田永鸿（北京大学，教授）

王金桥（中科院自动化所，研究员）

吴飞（浙江大学，教授）

谢长生（华中科技大学，教授）

薛巍（清华大学，副教授）

杨广文（清华大学，教授）

郑纬民（清华大学，教授，中国工程院院士）