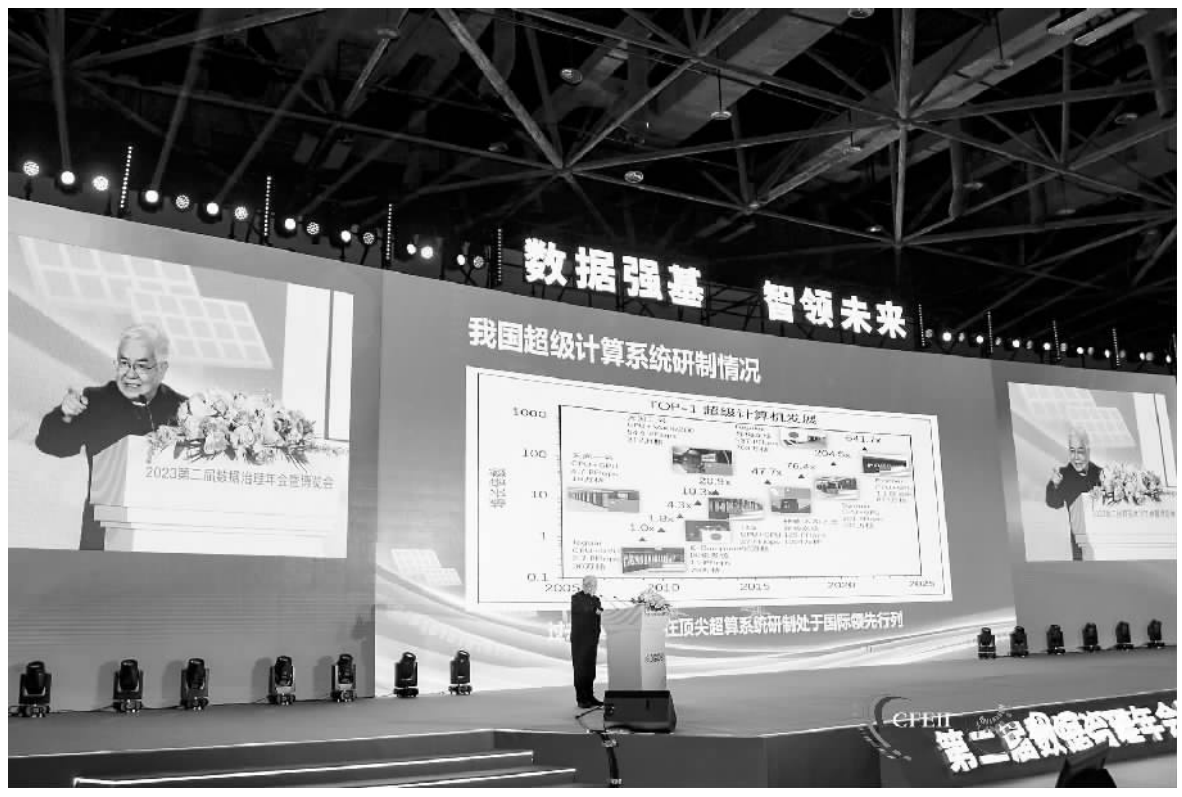
中国工程院院士郑纬民：

AI大模型基础设施亟待优化

本报记者 张琪玮

在12月13日召开的第二届数据安全治理年会上，中国工程院院士郑纬民表示，我国人工智能基础设施亟待优化，应从软硬件两方面突破瓶颈。



我国人工智能产业面临软硬件两方面瓶颈

在会上，郑纬民提出了算力“三大定律”：人类已经进入以算力为核心生产力的数字经济时代，算力就是生产力，这是“时代定律”；当下，算力每12个月就增长一倍，算力资源增速显著，已经打破摩尔定律，这是“增长定律”；算力每投入1元，就带动3~4元GDP经济增长，这是“经济定律”。

近年来，我国人工智能产业发展呈指数级增长。郑纬民表示，预计到2025年，中国人工智能产业规模将超过4500亿元，带动产生间接经济效益超1.6万亿元。

郑纬民直言，我国人工智能产业正面临着软件、硬件两方面的瓶颈。

从硬件角度看，一方面，我国国产芯片产量不足。郑纬民表示，2021年，我国人工智能服务器芯片总用量100万片，其中，美国英伟达市场份额高达95%左右。

从软件角度看，我国在算法等技术方面发展仍然有所欠缺，当前，谷歌和Meta的人工智能算法开发框架占中国人工智能市场份额的90%以上。

郑纬民坦言，要解决当下面临的问题，一是营造完善的人工智能服务器硬件生态，二是优化人工智能大模型基础设施架构。

“4个平衡”优化大模型基础设施

郑纬民强调，在设计大模型基础设施时，要思考“4个平衡”的优化问题。

一是半精度运算性能与双精度运算性能的平衡设计。在计算机系统的内存中，半精度、单精度和双精度是决定数据计算精确度的度量标准，双精度比半精度更精密，但同时要占据更多存储空间。郑纬民提出，大模型设计中不仅要考虑16位的半精度运算性能，还要考虑支持64位的双精度运算。他表示，最优的双精度与半精度运算性能比为1:100。

二是网络平衡设计。郑纬民指出，在网络设计方面，高带宽、低延迟的网络是极大规模预训练模型运行的必要条件。“在训练过程中，我们采用数据并行、模型并行和专家并行三种不同的并行方式，但这三种方式对互联有不同的要求。”郑纬民表示，“只有把通信做好，大模型才能顺畅跑通。”

三是体系结构感知的内存平衡设计。通俗而言，大模型在训练过程中使用的大量数据会产生大量的内存访问请求；对内存平衡的优化，目的是提升模型访存性能，从而提高模型训练效率。

四是输入输出子系统平衡设计。郑纬民指出，机器在执行大规模训练任务时，发生硬件、软件错误在所难免。针对这样的情况，容错检查点成为了大模型训练中的一道“保险闸”。容错检查点设置不足，会导致模型训练效率降低；检查设置过于频繁，则会浪费大量时间和存储空间。因此，优化检查点存储在大模型训练中的重要性不言而喻。

“以上四点平衡的问题得到解决，AI大模型将实现快速发展。”郑纬民总结道。

亚马逊云科技发布多项创新成果

本报记者 宋婧

作为全球云计算行业创新的风向标，一年一度的亚马逊云科技re:Invent大会备受业界瞩目。12月12日，2023亚马逊云科技re:Invent中国行首站在北京召开。会上，亚马逊云科技宣布了多项新服务，以及人工智能、机器学习和芯片技术方面的创新成果。

亚马逊云科技大中华区产品部总经理陈晓建表示：“我们在基础设施、计算、存储、数据等领域持续重塑云计算，并围绕最具变革性的生成式AI技术推出重磅新服务及功能，希望通过这些技术帮助更多企业加快创新速度，利用生成式AI全面重塑未来。”

低轨卫星

云服务正在从中心拓展至边缘，甚至走向太空。“我们正在建设的一个通过数千颗近地轨道卫星组成的卫星网络，希望未来人们在地球上的任何一个角落都能够享受快捷的互联网接入。”陈晓建说道。

记者了解到，亚马逊云科技启动了一项名为“Project Kuiper”的太空互联网项目，旨在通过发射和运营由3236颗低地球轨道（LEO）卫星组成的星座，为全球数千万人提供低延迟、高速宽带互联网服务。2022年第四季度，亚马逊成功发射了其第一颗“Project Kuiper”互联网卫星；2023年10月6日，载有两颗亚马逊原型卫星的Atlas V火箭从美国佛罗里达州发射升空，这两颗测试卫星将用于展示、发送并接收宽带信号。据陈晓建透露，早期客户测试将于2024年下半年开始。

生成式AI

生成式AI重塑各行各业，带来了巨大的市场潜力。陈晓建介绍说：“亚马逊云科技为生成式AI提供三层架构，包括利用基础模型构建的应用程序、使用基础模型进行构建的工具和用于基础模型训练和推理的基础设施。亚马逊云科技在每一层架构都持续创新，帮助客户更轻松、安全地构建和应用生成式AI，进一步降低利用生成式AI的门槛。”

在自研大模型方面，亚马逊云科技的Claude发布2.1版本的重大更新，带来了更多模型选择和全新功能。相比之前版本，虚假或幻觉类陈述减少了200%。新的测试版功能允许Claude与用户现有的流程、产品和API集成。

此外，生成式AI应用搭建平台Bedrock也值得关注。亚马逊云科技将其称作“最简单的利用大模型搭建和扩展生成式AI的方式”。用户可基于此平台对Amazon Titan、Claude、Llama2等多款主流大模型在关键标准上的表现进行评估，从而做出最佳权衡。

针对微软推出的由GPT驱动的Copilot生成式人工智能助手，亚马逊云科技推出了一款名为“Amazon Q”的生成式人工智能助手。这意味着亚马逊云科技在生产力软件领域开始挑战微软和谷歌的权威。

Amazon Q是一项新型生成式人工智能辅助服务，其定位类似于一个专家助理。Amazon Q只专注于工作场景，而非面向普通消费市场。

现场，陈晓建演示了Amazon Q的多项功能，包括开发应用程序、转换代码、充当商业应用程序等。科技行业研究公司Futurum表示，亚马逊云科技的Amazon Q将改变客户的游戏规则，预计Amazon Q在未来几个月内会被开发人员和云管理员广泛采用。

模型芯片

自研芯片已成为云厂商共识。AWS作为在全球云市场占比最高的头部大厂，自去年以来就一直在增加基础设施功能和芯片的投入，以支持具有增强能效的高性能计算，并宣布了其Graviton和Trainium芯片的最新迭代。这两款芯片分别为大模型训练和推理而生。

据介绍，相比前一代Graviton3处理器，Graviton4处理器提升了30%的计算性能、50%的内核和75%的内存带宽。Trainium 2是用于生成式AI和机器学习训练的专用芯片，训练速度是第一代Trainium芯片的4倍，内存和能效则分别提升了3倍和2倍。

除了自研芯片，AWS还扩展了与英伟达的合作伙伴关系，二者正在合力打造全球首个云上GH200AI超级计算集群。基于合作，亚马逊云科技将成为第一家部署英伟达GH200芯片的云计算服务商。

量子计算

与微软、谷歌等科技大厂一样，AWS也认为，量子计算和量子网络将成为先进计算领域的重要参与者。“量子计算来了，它将改变世界。”陈晓建感叹道。

会上，亚马逊云科技展示了他们在量子计算领域的一些最新进展和目标，其中包括最新的量子纠错硬件设计。据了解，这种新的架构有望用更少的超导组件来生产可控的逻辑量子比特，从而为超大规模量子计算机的构建铺平道路。

陈晓建表示，新的超导量子芯片实现量子纠错的效率是标准纠错方法的6倍，商用量子计算机的到来，可能比之前预计的更早。这一步是开发高效硬件、扩展量子纠错的重要组成部分。“目前的技术水平是每1000次量子操作大约出现一次错误。下一步的目标是每1000亿次出现一次错误。”陈晓建说道。

裸眼3D等项目率先开启5G-A试点探索

本报讯 记者**张琪玮**报道：记者从2023世界5G大会上获悉，5G-A（5G Advanced）的关键技术和价值已得到运营商验证，浙江、广东、北京、上海等地已经启动了裸眼3D、物联、车联、低空、通感等多样化的5G-A试点项目。

据悉，5G-A将带来能力和体验的再升级，具备“万兆下行、千兆上行、确定性体验、通感一体、原生智能、千亿物联”等六大特征。

据记者了解，随着5G-A网络的带宽和时延等性能的进一步提升，裸眼3D技术的体验将更加流畅和逼真，为观众带来更加震撼的视觉效果。裸眼3D无须依赖任何辅助设备或空间场地，只需通过一部随身携带的手机或平板电脑，用户即可随时随地享受3D视觉盛宴。

据了解，2023年是中国5G商用第五年，5G已经进入应用规模化发展阶段，商用成绩斐然。统计数据显示，2022年5G直接带动经济总产出为1.45万亿元，间接带动经济总产出约为3.49万亿元，中国5G应用已广泛融入97个国民经济大类中的67个，应用场景向生产控制等核心环节稳步拓展，5G正在不断夯实数字经济发展的基础设施底座。

欧盟就全面监管人工智能法案达成初步协议

本报讯 当地时间12月8日，欧盟成员国及欧洲议会议员就全面监管人工智能法案达成初步协议。据悉，这是全球首次以全面的、基于伦理的方式，尝试对人工智能技术进行监管。

欧盟委员会内部市场专员蒂埃里·布雷顿（Thierry Breton）表示：“欧盟成为第一个为人工智能的使用制定明确规则的地区。‘人工智能法案’不仅是一本规则手册，更将成为欧盟企业和研究人员引领全球人工智能竞赛的启动平台。”

据悉，协议对包括ChatGPT在内的所有通用人工智能模型的透明度提出了严格要求，包括限制人脸识别技术的应用，以及禁止对人类安全造成“不可接受风险”的人工智能系统等。

此前，欧盟各成员国和欧洲议会议员已经就如何监管人工智能进行了多轮磋商，2021年欧盟委员会正式提议通过“人工智能法案”。上月底，德国、法国和意大利三国刚刚就人工智能监管达成一致。据了解，该法案的相关细节还需进行进一步探讨，具体落地时间仍未可知。

（赵 晨）