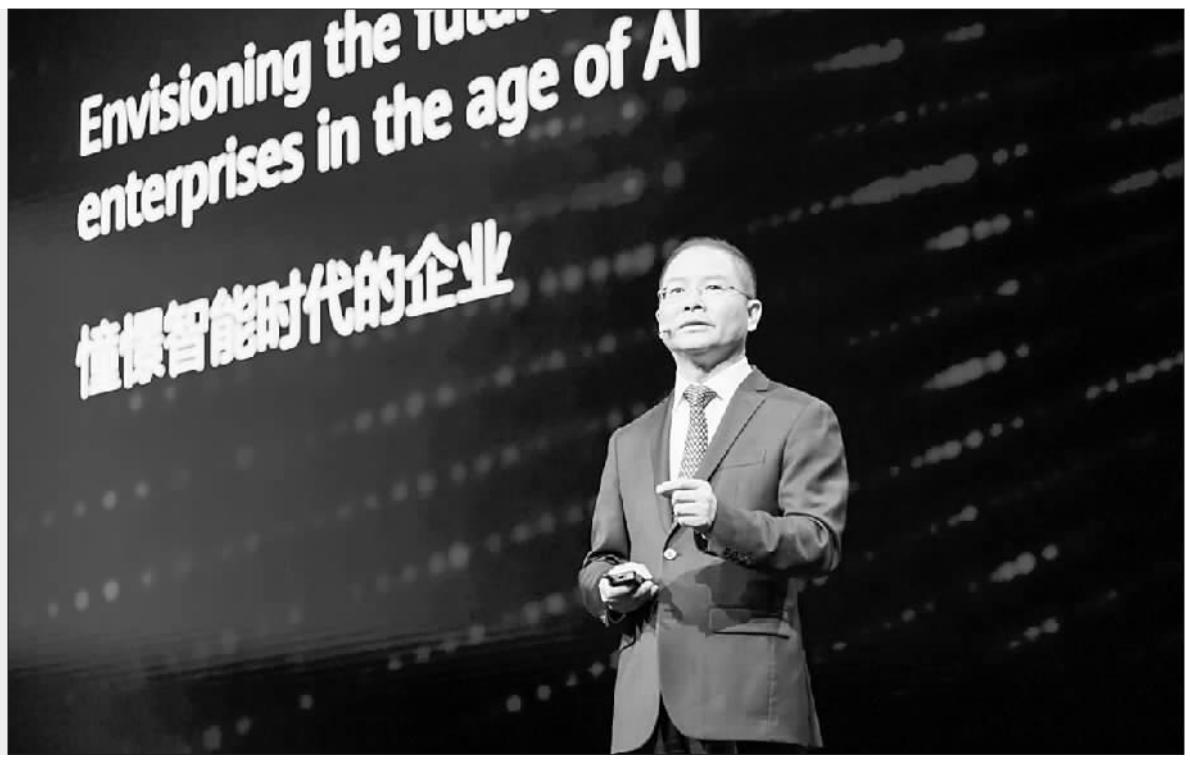
华为轮值董事长徐直军：

全面智能化时代已然来临



本报记者　张琪玮

近日，华为全联接大会2024在上海举办。会上，华为副董事长、轮值董事长徐直军表示：“AI技术的持续进步正在推动各行各业智能化的不断深化，全面智能化时代已然来临。”

算力供给既是机遇又是挑战

计算系统正不断发生结构性变化，系统算力的重要性已经逐渐超越单处理器算力，算力供应商的架构性创新迫在眉睫。

AI技术要落地应用，企业要实现智能化转型，算力是必不可少的“刚需”。徐直军指出：“智能化必将是一个长期过程，而算力将始终是智能化的关键基础。因此，要实现智能化的可持续，首先要实现算力的可持续。”随着人工智能成为主导性算力需求，计算系统也正不断发生结构性变化，系统算力的重要性已经逐渐超越单处理器算力，算力供应商的架构性创新迫在眉睫。

要实现架构性创新、打造可持续的算力解决方案，可谓任重道远。徐直军坦言，实现算力增长所需的芯片先进性制约是华为打造算力解决方案必须面对的一大挑战，“立足中国，只有基于实际可获得的芯片制造工艺打造的算力才是长期可持续的。为此，华为将基于实际可获得的芯片制造工艺，计算、存储和网络技术协同创新，开创计算架构，打造‘超节点+集群’系统算力解决方案，长期持续满足算力需求。”

同时，近年来大模型市场需求出现爆炸式增长，建设AI算力、训练大模型风靡千行百业，无疑也为算力供应商带来了重大利好。然而，徐直军指出，立足企业长远发展，如火如荼的“大模型热”也需要一些“冷思考”，“不是每个企业都要建设大规模AI算力，不是每个企业都要训练自己的基础大模型，不是所有的应用都要追求‘大’模型。”

对于当前大量企业纷纷建设大规模AI算力的现状，徐直军泼了几盆冷水：“一是AI算力集群建设成本高、技术迭代快，很容易面临要么浪费、要么无法满足需求的困境；二是AI硬件产品更迭速度快，企业受限于成本与算力规模，大多需要多个代际产品混合使用，导致资源调度复杂度高，无法充分发挥出新一代产品的性能；三是多代际产品混用的形式，对运营维护人员有较高的技能要求，这对于很多只具备传统IT维护能力的企业而言是重大挑战。”基于此，徐直军表示：“每个企业都要思考适合自己的获取AI算力的方式，而不仅仅是建设自己的AI算力。”

而要分别打造自己的基础大模型，也对企业提出了许多挑战。徐直军指出，一方面，当前训练基础大模型的预训练数据量进入10万亿tokens量级，所需的高成本与海量数据成为了企业的第一道“门槛”；同时，随着大模型技术发展不断深入，训练模型的难度也与日俱增，通常需要数月到数年时间完成模型迭代训练，这样长时间的投入，无疑会对AI尽快赋能核心业务产生负面影响；此外，当前大模型技术更新快、但具备实战经验的技术专家少，人才获取也是一大挑战。

立足盘古大模型的实践经验，徐直军指出，只要因地制宜，“小模型”也能发挥“大作用”：十亿参数模型可以满足科学计算、预测决策等业务场景的需求，并在PC、手机等端侧设备上广泛应用；百亿参数模型可以满足面向NLP、CV、多模态等大量特定领域场景的需求，如知识问答、代码生成、坐席助手、安全检测等；面向NLP、多模态的复杂任务，则可以用千亿参数模型来完成。“总而言之，企业需要的是根据自身不同业务场景需求，选择最合适的模型，通过多模型组合，解决问题，创造价值。”徐直军总结道。

智能化时代企业应做到“6个A”

“6个A”分别是自适应体验，自演进产品，自治的运营，增强的员工，全量全要素全联接，以及智能原生基础设施。

“AI已经成为对行业影响最大的技术，从没有一项技术进步像AI一样，在如此短的时间内产生如此大的影响。”徐直军直言。

一项麦肯锡和斯坦福大学的研究数据表明，当前，AI行业应用主要集中在产品开发、营销与业务运营三个环节，并在企业的积极推动下不断深化拓展到更多环节中。对此，徐直军认为，AI的行业应用不仅是为企业创造出“今天”的价值，更是为企业巩固“明天”的智能化竞争力。因此，立足全面智能化的时代浪潮，企业应及时调整发展方针，立足“6个A”发展智能时代的新特征。

徐直军告诉记者，这“6个A”分别是自适应体验（Adaptive User Experience），自演进产品（Auto-Evolving Products），自治的运营（Autonomous Operation），增强的员工（Augmented Workforce），全量全要素全联接（All-Connected Resources），以及智能原生基础设施（AI-Native Infrastructure）。

具体而言，“自适应体验”是华为对“未来企业应如何服务客户”这一议题给出的答案。徐直军指出，自适应体验指智能化企业应该能够感知并理解用户的行为、需求、兴趣、品味和环境变化，主动调整提供最符合用户需求的服务、与能够适时和同时满足海量个性化独特需求的产品。例如，智能时代的AI学习机应当能够根据学生年龄、学习进度、理解能力以及测试反馈等自动调整教学内容和难度，让每个学生在不同阶段都能获得适合自己的学习体验。“为客户提供预设的确定体验到自适应体验是一次跃迁，每个企业都需要提供适应智能化时代的客户体验。”徐直军表示。

“自演进产品”则是对智能时代企业需要打造的产品的预测。徐直军认为，智能化时代的产品将具备自主学习、持续迭代、适应变化的能力，能够进行自优化和自演进。他强调：“产品从‘数字化’到‘智能化’是一次跃迁，每个企业都需要思考如何把智能化能力融入自己的产品。”

智能化企业日常运营的未来，则是“自治的运营”，即要实现业务流高度自治运营，从感知、规划、决策到执行，形成端到端的自主闭环，通过AI技术完成运营效率提升的飞跃。例如，港口通过智能计划平台，能够自动生成作业计划，通过自动驾驶集卡自动完成集装箱水平运输。“企业运营自动化是多年以来很多企业一直在追求的，随着AI工具越发成熟，每个企业都需要思考如何在更广、更深的范围用AI赋能和改变企业运营。”徐直军坦言道。

“增强的员工”，是企业对未来员工工作体验和工作方式的愿景。“要让每个员工都有一个‘懂我’的智能助手，从而高效、高质量的完成每一件工作。”徐直军表示，让AI造福于人类是AI存在的意义，让员工有更好的工作体验是每个企业在智能化时代竞争力的关键基础。“可以预想，在未来，运营商基站现场维护人员通过维护助手APP快速获取故障位置，故障根因以及处理建议等信息，从而大幅提高工作效率。”

如果说前四点表征了智能化的效果，最后两点则表征了智能化的基础。徐直军表示，全量全要素全联接是指，要实现企业的资产、员工、客户、伙伴、生态等全互联，所有业务对象、过程、规则实现深度、全面的数字化，使企业具备智能化必须的数据和信息基础；智能原生基础设施则指ICT基础设施要同时兼顾两方面智能化——“ICT for Intelligence”和“Intelligence for ICT”：既要系统化构建，适应智能化应用的需要；又要保证基础设施本身的运维管理和体验保障的充分智能化。

“全面智能化时代已然来临，给每个人、每个企业带来新的机遇和新的挑战。”临近尾声，徐直军提出了其对全面智能化时代的期待，“未来，要让每个人都有自己专属的智慧助手，让每个企业成为智能化企业，让每辆车都能无人驾驶。”

逾400项新产品新技术将在第三届数贸会首次亮相

本报讯　记者**齐旭**报道：近日，国新办就第三届全球数字贸易博览会（以下简称“数贸会”）有关情况举行发布会。记者从发布会上获悉，9月25日至29日，第三届数贸会将在浙江杭州举办，目前各项筹备工作已基本就绪。在介绍大会基本情况时，商务部部长助理唐文弘分享了三个“聚焦”。

一是聚焦开放。在扩大国际合作中提升开放能力，把数贸会作为展现高水平开放的重要平台。在展览展示方面，32个国家和地区的龙头企业参展，国际企业参展数量和面积占比均超20%。在首发首秀方面，400余项新产品、新技术中约1/4来自国外，比例远超上届。在洽谈对接方面，首次同期举办投资中国—开发区对话500强等活动，国际客商数量约为上届的3倍。

二是聚焦创新。举办高水平、专业化的数贸会，将数字贸易打造成为共同发展的新引擎。在展示内容上，首发首秀展数量是去年的4倍，首次设立未来产业专区，集中展示智能机器人、低空经济等领域前沿技术。在呈现方式上，将采用数字人、裸眼3D等技术增强现场互动。

三是聚焦共赢。努力把合作的清单拉长，以中国新发展为世界提供新机遇。在打造品牌方面，今年将围绕电商产业生态，进一步丰富活动内涵。同时，率先落实中非合作论坛峰会成果，策划“数贸非洲日”活动，打造数字贸易合作新典范。

记者还了解到，今年的数贸会综合展区邀请到了正在风靡全球的《黑神话：悟空》联动全球知名硬件厂商演绎“传奇故事”；人工智能展区还邀请了60多个智能机器人同台竞技。

此外，作为第三届数贸会的重要活动之一，由杭州市商务局（杭州自贸片区管委会）和北京赛迪出版传媒有限公司共同承办的2024全球数字贸易创新大赛将于同期举行决赛。

郑纬民院士详解“八卦炉”软件系统最新进展

本报讯　随着大模型的应用日益广泛，我国对算力的需求也呈现出爆发性的增长，构建国产智能算力系统成为推动我国人工智能产业发展的关键。近日，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在2024服贸会通用人工智能算力论坛上指出，发展国产智能算力系统需要“软硬兼备”，优秀的系统软件能够充分释放底层硬件算力的潜力，因此，我国人工智能产业不仅需要关注硬件性能，更需要重视软件生态的建设。

郑纬民指出，构建人工智能大模型主要分数据获取、数据预处理、模型训练、模型微调、模型推理等五个环节，而这每一个环节都对算力提出了极高的要求。“为了应对人工智能大模型带来的算力需求，我国对AI芯片的需求也在加大。”郑纬民表示，近年来，我国自主研发的AI芯片取得了显著进展，但产业依然面临智能算力的软件生态建设和软件支持不足等问题。

“智能算力的软件生态建设是当前制约我国AI芯片发展的关键因素。”郑纬民指出，硬件性能固然重要，但系统软件的完善同样不可或缺，只有两者兼备，才能真正推动我国智能算力的发展。优秀的系统软件能够充分释放底层硬件算力的潜力。通过优化软件生态，不仅可以提高算力效率，还能降低应用成本，为智能算力系统的广泛应用奠定坚实基础。

为了解决AI芯片生态建设的问题，清华大学开发了一套名为“八卦炉”的智算系统核心基础软件。这套软件包括并行系统、编程框架、AI编译器、算子库等多个组件，旨在优化AI芯片的性能并提高其易用性。

据郑纬民介绍，该“八卦炉”已在多个场景中得到了验证，如在神威平台上实现大模型训练，结果显示训练结果准确，而且成本较低（预计成本仅为英伟达GPU的六分之一）。此外，在与沐曦、燧原科技、摩尔线程等企业的合作中，“八卦炉”也显著提升了算力效率，降低了推理成本。“随着智能算力系统的不断完善，相信不久的将来，我们将看到更多基于国产技术的人工智能应用不断涌现，为我国的科技创新和产业升级注入新的活力。”郑纬民表示。　（谷　月）

阿里巴巴集团CEO吴泳铭：

生成式AI终将接管整个数字世界

本报讯　近日，一年一度的云栖大会在杭州拉开了帷幕。阿里巴巴集团CEO、阿里云智能集团董事长兼CEO吴泳铭在会上表示，AI最大的想象力绝不是在手机屏幕上打造一两个超级应用，而是会接管数字世界，改变物理世界。

过去20年，互联网浪潮的本质是“连接”，连接人、信息、商业、服务和工厂，通过连接提高了整个社会的协作效率，创造了更大的价值，改变了人们的生活方式。而生成式AI的出现，通过生产力的供给创造了新的价值，从而为世界创造了更大的内在价值，提高了整个世界的生产力水平。“这种创造价值，我们觉得可能是移动互联网通过连接创造价值的十倍，甚至是几十倍。”吴泳铭说道。

“我们认为生成式AI将逐渐渗透数字世界，并接管数字世界。”吴泳铭进一步解释说，“物理世界的大部分事物都会具备AI能力，形成下一代的具备AI能力的新产品，并与云端AI驱动的数字世界产生协同效应。很长一段时间，AI的焦点主要集中在模拟人类的感知能力，比如自然语言理解、语音识别、视觉识别。但是深层次AI的崛起带来了质的飞跃，AI不再局限于感知，而是首次展现了思考、推理和创造的力量。”

生成式AI让世界有了一个统一的语言——“token”，它可以是任何文字、图像、视频、声音的代码。大模型可以通过物理世界数据的“token化”，理解真实世界的方方面面，比如人类行走、奔跑、驾驶车辆、使用工具、绘画、作曲、写作、编程等，甚至是开公司创业。有了这样的理解，AI就可以模仿人类去执行物理世界的任务，这将带来新的产业革命。

可以看到，汽车行业正在发生这样的变革：之前的自动驾驶都是靠人类算法去填写规则，通过非常多的代码仍然无法重现所有的驾驶场景。现在汽车行业采用端到端的大模型技术训练后，AI模型能够直接学习海量的人类驾驶视觉数据，让汽车具备超越大部分时期的驾驶能力。

机器人将会是下一个迎来巨变的行业。所有能够移动的物体都将变成智能机器人。它可以是工厂里的机械臂、工地里的起重机、仓库里的搬运工、救火现场的消防员，甚至家庭里的宠物狗以及保姆助理。未来工厂里会有很多的机器人在AI大模型的指挥下生产机器人。每个城市家庭可能都会有一个或两三个智能机器人来帮他们解决生活困扰，或者帮助他们提升生活中的效率。“可以预见，AI驱动的数字世界连接着具备AI能力的物理世界，未来整个世界的生产力水平将大幅提升，对物理世界的运行效率也将产生革命性的影响。”吴泳铭表示。

在吴泳铭看来，AI计算正在加速演进，成为整个计算体系的主导。无论是端侧的计算还是云端的计算，这都是一个非常明显的趋势。生成式AI对数字世界和物理世界的重构，将带来计算架构的根本性变化。过去几十年，CPU主导的计算体系正在加速向GPU为主导的AI计算体系转化。未来几乎所有的软硬件都会具备推理能力，它们的计算内核将会变成以CPU　AI算力为主，传统CPU计算算力为辅的计算模式。“在新增的算力市场和算力需求中，超过50%以上的需求都由AI驱动产生，AI算力的需求渗透已经超过50%，已经占据主导地位，未来这一趋势还会持续地扩大。”吴泳铭说道。

据他透露，过去一年，阿里云投资新建了大量的AI算力基础设施，但还是远远不能满足客户的旺盛需求。“今天我们接触到的所有客户、所有开发者、所有CTO，几乎都在用AI重构他们自己的产品”，大量的新增需求正在由GPU算力驱动，大量的存量应用也在用GPU重新改写。

“AI计算正在加速向汽车、生物医药、工业仿真、气象预测、教育企业软件、游戏等各个行业渗透。看不见的新兴产业革命正在悄然演进，所有行业都需要性能更强、规模更大、更适应AI需求的算力基础设施。”吴泳铭说道。

“从历史经验来看，人们对于新技术革命的期待会在短期内高估，又会对其长期发展低估。在新技术应用早期，渗透率比较低，人们往往会对经验中还没有发生过的事情产生怀疑。这很正常，新技术革命会在大多数人的怀疑中成长，也会在大多数人的怀疑中错过。站在AI时代新浪潮的开端，我感到无比的兴奋。”吴泳铭感慨道。

（宋　婧）