AI芯片陷入“缺口”之争

本报记者 张心怡

“第一个问题是，7万亿美元到底是关于什么的?”在当地时间2月21日举办的Intel Foundry Direct Connect 2024上，英特尔CEO帕特·基辛格将这个问题抛给了OpenAI CEO山姆·奥特曼。

“首先，不要对媒体的报道照单全收……事实的核心是，我们相信对于AI计算、能源、数据中心的大量投资是非常重要的……这需要全球性的投资，会有助于开展许多不同的工作，帮助许多人员。除了芯片，还有整个AI基础设施。”奥特曼说道。

“7万亿美元能购买多少GPU?”在当地时间2月12日开幕的2024年世界政府峰会期间，阿联酋人工智能部部长奥马尔·阿尔·奥拉马向英伟达CEO黄仁勋提问。

“所有GPU。”黄仁勋说道。

上述两段发生在今年2月的对话，都指向了近日围绕奥特曼的媒体报道：他正在筹集巨额资金建设AI芯片工厂，重塑芯片产业链。对此，奥特曼表示，AI基础设施建设需要大量的全球性投资，但不要过于在意“7万亿美元”这个数字，他无法逐一纠正媒体的报道。

AI芯片出现巨大缺口

AI芯片作为生成式人工智能的核心算力单元，也出现了巨大的缺口。

此前，奥特曼在社交平台上表示，世界需要的人工智能基础设施超出人们的建设计划，包括晶圆厂产能、能源、数据中心等。建设大规模的人工智能基础设施和具有弹性的供应链，对于保持经济竞争力至关重要。OpenAI将尽力提供帮助。

自从ChatGPT引爆生成式人工智能的热潮后，产业界就陷入了深深的算力和能源焦虑。AI芯片作为生成式人工智能的核心算力单元，也出现了巨大的缺口。

至于这个缺口会进一步扩大还是缩小，取决于芯片技术进步与AI需求增长的竞速，前者快则缺口有望缩小，后者快则缺口或将增大。

从目前来看，黄仁勋更相信前者，而奥特曼更忧虑后者。

黄仁勋对于AI芯片缺口的计算方式，有着自己的理解。他认为，计算架构的性能正在持续提升。如果抛开计算速度的进步，只看计算单元的总量，就很可能得出全球需要14个行星、3个星系和4个太阳才能覆盖计算所需能源的结论。过去10年，英伟达等从业者作出的最大贡献之一就是将计算和人工智能的效能提升了100万倍。因此，无论面对什么样的计算需求，都应当把计算架构仍然具有百万倍提升空间这一点纳入考量。

AMD K8微架构的首席架构师吉姆·凯勒(Jim Keller)也在社交平台上表示，他可以用不到1万亿美元实现所谓的“7万亿美元目标”。首先要消除2~3层的利润堆叠(供应链中每个参与者为将产品提供给最终用户所付出的成本)，接下来是提升芯片的运行速度，使硬件与软件相匹配。

黄仁勋和凯勒的观点都趋向于一点：产业的算力和能源缺口，要通过芯片的性能提升来解决，核心在于技术的进步和创新。

亟须发力AI基础设施和供应链

生成式人工智能的训练往往需要几百张甚至上万张GPU，进一步加剧了GPU的紧俏程度。

而奥特曼的顾虑，多少与积极自研AI芯片的大型云服务厂商们有共通之处：当前的AI基础设施难以支撑生成式人工智能的爆发。

早在2023年上半年，OpenAI就曾表示GPU供应不足限制了ChatGPT能够处理的信息量和功能优化，部分初创AI企业表示不得不采用价格更加昂贵的GPU来填补算力缺口，云服务厂商面对客户的算力池扩张需求也倍感压力。尤其是生成式人工智能的训练往往需要几百张甚至上万张GPU组成的集群才能发挥最佳效能，这进一步加剧了GPU的紧俏程度和产品溢价。

虽然英伟达等厂商持续缩短GPU的交付周期，但大模型性能进步、数量提升、应用扩展三重因素的叠加，还是令下游市场忧心忡忡。以OpenAI为例，从2022年11月ChatGPT上线，到2023年3月多模态大模型GPT-4上线，再到2024年2月以文本描述生成视频的人工智能模型Sora的问世，在不到两年的时间里，大模型就实现了飞跃式发展。再比如百度研发的文心大模型，其3.5版本相比上一个版本，训练速度提升了2倍，推理速度提升了30倍。与此同时，大模型的数量也在飞速增长。2023年1—11月，国内发布了238个大模型。这意味着，在这200多天的时间里，平均每天都有一个大模型在中国问世。此外，大模型的应用范围也在不断拓展。

这或许是为什么以模型和工具开发为主营业务的OpenAI团队，打算发力AI基础设施和供应链，协助产业筹建更多晶圆厂、能源设施和算力设施的原因。

当然，奥特曼本人是一位富有创新精神的企业家。OpenAI协助“建设大规模AI基础设施和弹性供应链”的计划也暂未披露具体方式、有哪些机构参与、各自有着怎样的利益诉求。如果这项计划只是聚焦购买设备、筹建晶圆厂等产业已有的重复性投入，实际的效益或许仅仅是增加AI芯片的产量和算力设施所需的能源供给。

但若这项计划聚焦计算架构和能源科技的创新，再围绕创新成果进行产业化部署，或许能给计算产业带来不同的图景。毕竟，计算产业已经来到了后摩尔时代的十字路口，量子、硅光等新的技术路径蓄势待发。

信息显示，奥特曼曾参与Riget-ti Computing、PsiQuantum、Quan-tinuum等量子计算公司的融资，并投资了该融合公司Helion Energy。

技术创新的跃进，来自创新和实干。在计算产业的发展过程中，许多关键的技术节点，都少不了技术狂人的灵光一闪和实干团队的商业化推进。我们期许如英伟达等企业研发团队、凯勒等技术专家的创新灵感，也需要马斯克、奥特曼等愿意整合资源改进供应链的推动者——或许两者相向而行，才能在计算产业触发一场呈燎原之势的变革。

AI推动数字人加速“走进生活”

本报记者 姬晓婷

拜年，这一延续了千年的传统习俗，在这个春节迎来了新的表达方式——数字人拜年。

在AI技术突飞猛进的当下，数字人急速走进我们的生活，它们的“生产者”，怀抱着不同寻常的“速度与激情”，热辣滚烫，飞扬青春。

北京聚力维度科技有限公司(以下简称“聚力维度”)是一家主营3D数字人内容制作的公司。该公司的品牌负责人谢京华告诉《中国电子报》记者：“今年春节前，公司接到最多的就是年会和拜年视频的订单。”谢京华向记者展示了他们近期制作的数字人年会视频：一位身穿旗袍的小姑娘，与两位身披铠甲的将军在红灯笼装扮的舞台上演唱《我的未来不是梦》。谢京华告诉记者，这条视频就是在公司的演播厅摄制完成的，录制团队最少仅需要两个人。

记者走进数字人视频内容摄制厅，才发现这个演绎了无数数字人喜怒哀乐的地方，竟然只有一间卧室面积的大小。拍摄导演孟瑶告诉《中国电子报》记者，数字人内容制作所需的设备非常简单：一个单目摄像头用于捕捉演员的动作和表情，一台配置较好显卡的计算机用于实时生成数字人，一台显示器帮助演员看到自己实时生成的数字人影像，再加上两盏补光灯，便能够支持数字人内容录制。

一位叫周瑶的演员走进拍摄现场。这天，她要以三种不同的虚拟人形象送上新春祝福。这三位女性形象，一位古灵精怪，一位沉稳大方，还有一位温柔可爱，她们是不同企业设定的数字人形象。周瑶，就在摄像头前，通过不同的虚拟形象，传达着对新春最美好的祝福。

“2024，华硕idol祝您：福起新年，万事顺遂。”周瑶在摄像头前，微笑着说完这句话，走到拍摄场地旁边的电脑旁坐下，为下一场拍摄的虚拟人物选择合适的数字装扮。

随着周瑶操作鼠标在屏幕上滑动，多彩的服装在人物身上来回替换。周瑶告诉《中国电子报》记者，这些服装，都是公司自己的“数字资产”。拍摄场地旁边的办公区，便是聚力维度工程师们工作的地方，从皮肤质地到服装、头发，数字人背后的各项技术都在这里完善。除了自制的“数字资产”，聚力维度还训练了“赛娲大模型”，专门用于生成不同类型的3D数字人模型。只需键入几个关键词，大模型便可自动生成符合要求的数字人模型。

在接受《中国电子报》记者采访时，聚力维度创始人赵天奇坦言，从产业发展周期来看，数字内容制作产业还处于起步阶段。而市场认可度，是制约数字人内容制作产业发展壮大最主要的因素。“数字人内容制作需要越来越多的人跳出传统，拥抱新技术流程才能实现普及。”赵天奇说道。

当被问及新的一年有什么计划时，聚力维度美术团队负责人曹泽恒向《中国电子报》记者说：“今年我们将把技术重点应用在AI短剧制作上。不断提升AI制作比重，不断提升数字人内容级别，争取让一个人就能每天创作10分钟影视级视频内容。”

中国移动5G轻量化技术已具备全面商用条件

本报讯 记者**张琪玮**报道：近日，中国移动携手10余家合作伙伴率先完成了全球最大规模、最全场景、最全产业的5G轻量化(RedCap)现网规模试验，推动首批芯片、终端具备商用条件，RedCap端到端产业已全面达到商用水平。截至目前，中国移动支持RedCap的5G基站总规模超10万，覆盖全国52个城市，实现城区连续覆盖，率先构建全国规模最大的RedCap商用网络。

记者了解到，当前最新一代终端设备在2.6GHz网络中上/下行峰值速率可达20Mbps/145Mbps，平均组网速率为12Mbps/71Mbps；在700MHz网络中上/下行峰值速率可达112Mbps/210Mbps，平均拉网速率为71Mbps/90Mbps；用户面时20~30ms，切换成功率达100%，基本功能与性能符合预期；网络侧支持多BWP等增强功能、切片资源预留等融合功能，可满足视频监控等场景大容量需求、可穿戴设备语音及移动性需求、电力等场景高隔离需求等多种应用场景需求。中国移动表示，RedCap网络、芯片、模组已具备全面商用条件。

工业和信息化部于2023年印发的《关于推进5G轻量化(RedCap)技术演进和应用创新发展的通知》(以下简称《通知》)为产业提出了发展目标，到2025年，5G RedCap产业综合能力显著提升，新产品、新模式不断涌现，融合应用规模上量，安全能力同步增强。在《通知》的指导下，中国移动打造了1个产业集群创新中心、5个技术创新之城和5个应用示范之城，全面加速RedCap产业成熟和应用创新。记者从中国移动获悉，其RedCap现网规模实验具有三个特点。

一是试验网规模最大。中国移动在浙江宁波、广东广州、上海、湖南岳阳、湖北十堰5个技术创新之城，构建“五城五网超千站”的最大规模试验网，组建“芯—模—网”的最全商用产业链，开展覆盖最全场景的RedCap端网试验。

二是测试场景最完善。开展网络性能测试和芯片终端测试，涵盖实验室和外场测试，重点验证网络基本功能与增强功能、首批芯片/模组功能与性能，以及与网络兼容性、RedCap与网络切片等5G优势技术融合功能，发现并解决部分场景速率低、不同系统厂家切换异常等近10项端网兼容问题，创新提出多BWP(部分带宽)灵活扩容、用户体验无缝的互操作增强等解决方案，满足业务的大容量和高移动等需求。

三是覆盖厂商最全面。参测厂商涵盖网络、芯片、模组等10家业界主流和先发厂商，有效加速芯片和模组成熟。