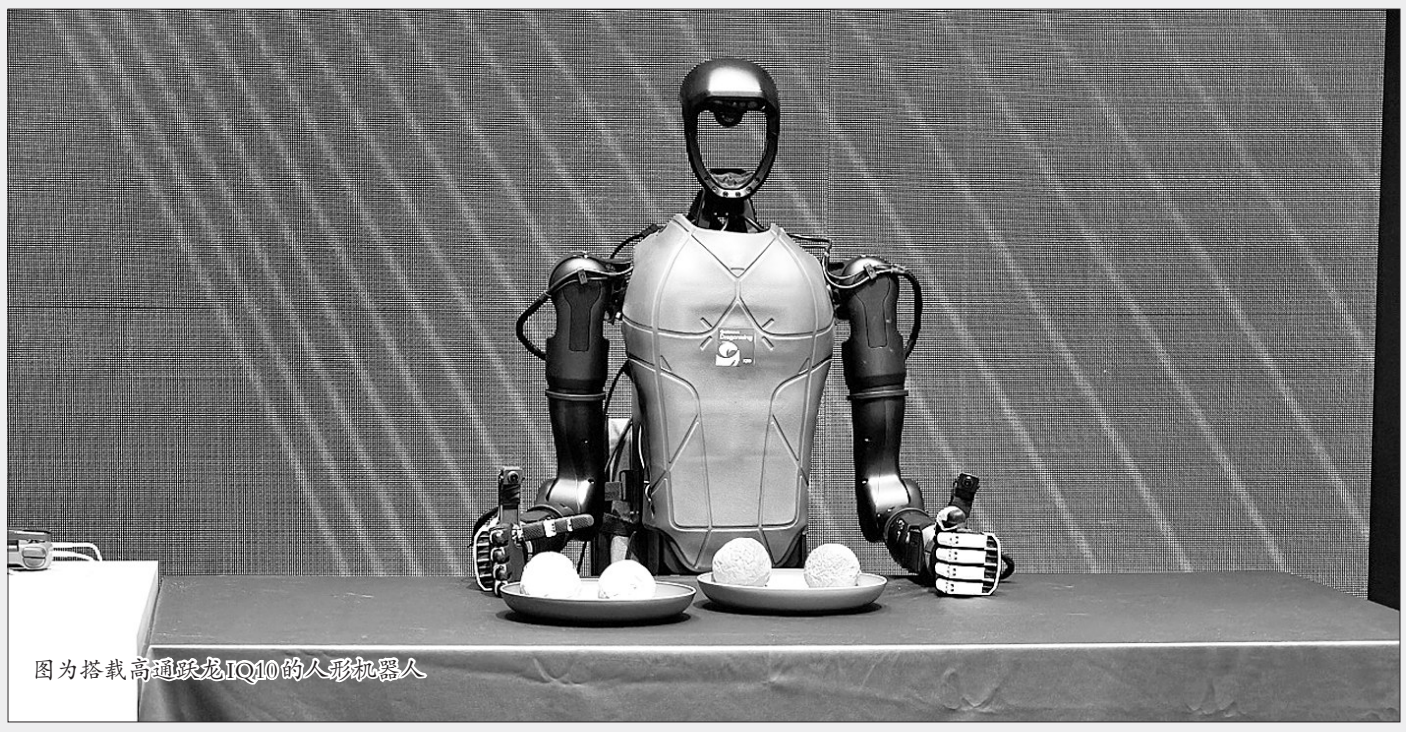
芯片公司的“上天”“落地”探索

本报记者 张心怡

一边“拔高”上限，一边务实“落地”——在CES 2026上，芯片继续成为最大的看点之一。AMD和英伟达等企业面向AI“深度思考”和智能体趋势，向YottaFlops（10的24次方浮点计算）级算力需求挺进，推动平台级别和机架级别的算力跃迁。与此同时，高通、英特尔、Arm、AMD、英伟达等企业，都在强调物理AI，打通AI奔赴现实世界的“最后一公里”。


图为搭载高通跃龙IQ10的人形机器人

算力“上天”
应对YottaFlops级计算需求

AI算力似乎永远不够用。表面上看，是AI用户数量急剧膨胀和推理任务日益复杂导致算力需求井喷，但更深层次的原因在于：AI的思考过程越来越详细。许多大模型不再像初代ChatGPT那样直接输出答案，而是有了思考路径，通过持续地自我验证乃至于自我否定来寻求最优解，每多出一个思考步骤，都意味着计算量的增长。

“OpenAI o1模型的引入是AI的转折点，推理不再是一次性给出答案，而是一个思考过程。为了教会AI思考，强化学习和大量计算被引入后训练阶段，让计算机通过自我尝试来学习如何执行任务，导致用于预训练、后训练、测试时缩放的计算量呈爆炸式增长。现在，我们每进行1次推理，都可能生成2个token而不是1个，测试时缩放导致模型生成的token数量每年增加5倍。”英伟达创始人兼首席执行官黄仁勋在CES 2026演讲中指出。

而智能体（Agent）的盛行，正在推动AI从“被动响应”向“主动决策”的根本性转变，并进一步推高算力需求。

“当我们将AI扩展到更广泛的智能体时，全球计算基础设施需求激增的趋势将更加深远。我们需要将计算能力再提高100倍，在未来5年内达到超过10 Yotta-Flops，也就是我们在2022年所拥有算力的1万倍。”AMD首席执行官苏姿丰（Lisa Su）在CES 2026开幕演讲中表示。

面向AI“深度思考”及智能体时代的算力需求，AMD、英伟达等企业基于新一代计算单元和不同组件之间的协同设计，以更快的迭代频率、更系统化的计算平台，应对全球AI开发者的算力焦虑。

Helios，是AMD面向Yotta级AI算力需求的新一代机架级平台。据介绍，每个Helios机架拥有超过1.8万个GPU计算单元和超过4600个Zen6 CPU核心，提供2.9 Exa-flops性能。每个机架还具备31TB的HBM4内存，260TB/s的纵向扩展带宽，以及43TB/s的横向扩展带宽。

Helios的核心是计算托盘，每个托盘包含4个Instinct MI455X GPU，搭配新一代EPYC“Venice”CPU和Pensando Vulca-no NIC（网卡），通过开放的ROCm软件栈整合。

其中，AMD新一代Instinct MI455X加速器被苏姿丰形容为“我们制造过的最先进的芯片”，拥有3200亿个晶体管（比MI355多70%），包含12个2纳米和3纳米制程的计算及I/O Chiplet，以及432GB的HBM4显存，所有单元通过AMD下一代3D芯片堆叠技术连接。驱动MI455X GPU的是代号“Venice”的下一代EPYC CPU，采用2纳米工艺，拥有256个最新高性能Zen 6核心；在机架规模下，Venice能全速为MI455X供应数据，实现协同工程。以上组件与800G带宽的Pen-sando Volcano网络芯片、Salina DPU集成，提供超高带宽和超低延迟，使成千上万的Helios机架能够在数据中心内扩展。

另据苏姿丰透露，下一代MI500系列的开发已经在进行中，该系列基于AMD下一代CDNA 6架构，采用2纳米工艺，并使用更高速的HBM4E内存。

“随着2027年MI500系列的推出，我们有望在4年内实现1000倍的AI性能提升，让更强大的AI惠及所有人。”苏姿丰说道。

而英伟达也一改“每一代新平台最多迭代一两颗芯片”的原则，一口气推出集成Vera CPU、Rubin GPU、NVLink 6交换机、ConnectX-9 SuperNIC、BlueField-4 DPU、Spectrum-6以太网交换机6款全新芯片的新一代AI计算平台NVIDIA Rubin。据悉，Rubin平台在MoE模型训练中使用的GPU数量仅为Blackwell平台的1/4，生成每个token的成本低至1/10。

有意思的是，Rubin GPU在NVFP4（英伟达提出的4位浮点格式）精度下的推理性能达到50 PFLOPS，是Blackwell的5倍，但晶体管数量只有Blackwell的1.6倍，实现这一目标的关键是英伟达的创新技术：NVFP4 Tensor Core。这种采用新格式的运算引擎不是简单地在数据路径中嵌入某种4位浮点数，而是一个完整的处理器单元，懂得如何动态、自适应地调整精度和结构，以应对Transformer模型的不同计算阶段，从而在允许损失精度时实现更高的吞吐量，在需要的时候再恢复到最高精度。该技术使Rubin GPU能够以相对更少的晶体管增长来实现更大幅度的性能提升，也有助于进一步降低AI算力成本。

物理AI“落地”
打通“奔赴现实世界”最后一关

走进CES高通展台，搭载高通跃龙（Dragonwing）IQ10系列的人形机器人正在捡拾水果。在展台上，摆放着红色、绿色的盘子各一个，红色、绿色的塑料水果各两个，机器人会将水果抓起来，放到与水果颜色一致的盘子里。无论工作人员或参观者如何变动水果的位置，还是交换两个盘子的位置，机器人依然会稳稳地拿起水果，放到同色的盘子中。

在本次CES上，高通、英特尔、英伟达、AMD、Arm等芯片公司，都在强调“面向真实世界部署”的物理AI。高通公司AI产品技术中国区负责人万卫星曾在公开演讲中，将AI应用的演进分为四个阶段。一是感知AI，比如传统的自然语言处理、语音降噪、图片识别；二是生成式AI，基于训练数据创作内容从而响应用户提示；三是智能体AI，能够自主行动和决策；四是物理AI，可以理解真实的物理世界，并根据物理定律做出反馈和响应。

要实现“AI无处不在”的愿景，物理AI显然是必不可少的“最后一公里”。

汽车、机器人是高通构建物理AI的主要抓手。汽车方面，高通的骁龙数字底盘解决方案已经被全球超过4亿辆汽车采用。机器人方面，高通推出下一代完整的机器人技术栈架构，集成硬件、软件和复合AI，并发布了最新高性能机器人处理器高通跃龙IQ10系列。此外，在物联网方面，高通推出全新高通跃龙Q-8750和Q-7790处理器，聚焦终端侧AI、多媒体能力、安全特性以及其他先进功能，从而更好地支持无人机、视觉系统、智能摄像头和AI电视等广泛的物联网产品形态。

在最新处理器的基础上，高通强调技术栈构建和完整的产品组合。比如，在机器人领域，高通提供了通用型机器人架构，结合视觉语言动作模型（VLA）和视觉语言模型（VLM）等端到端AI模型，支持先进感知和运动规划，从而赋能泛化操作能力以及人与机器人的交互能力。搭载高通跃龙IQ10的通用型机器人架构提供了完整的技术栈，包括异构边缘计算、边缘AI、混合关键级系统、软件、机器学习运维和AI数据飞轮，并依托合作伙伴生态系统与开发者工具套件，使机器人能更高效地进行推理并智能地适应时空环境，经优化后能够在多种形态下实现工业级可靠性的规模化部署。

英特尔在CES 2026正式发布第三代英特尔酷睿（Core）Ultra处理器，覆盖从PC到边缘领域的应用。该处理器是首款基于In-tel 18A制程打造的计算平台，旗舰型号酷睿Ultra X9 388H配备16个CPU核心（4个性能核、8个能效核、4个低功耗能效核）、12个Xe核心（核显）和50 TOPS NPU算力，将赋能超过200多款PC产品设计。

值得注意的是，英特尔在3系列处理器上，首次实现了边缘处理器与PC版本同步发布，并首次获得了针对嵌入式和工业边缘场景的测试与认证，包括宽温范围支持、确定性以及7×24小时全天候可靠性，加速AI在具身智能、智慧城市、自动化与医疗领域的部署。据悉，搭载第三代英特尔酷睿Ul-tra处理器的边缘系统预计将于2026年第二季度开始面市。

Arm将物理AI视为AI领域发展的核心动能，基于算力支撑，赋能汽车、机器人及各类设备感知、理解现实环境，并在实际场景中安全可靠地运行。特斯拉新一代AI5芯片基于Arm计算平台打造，其AI性能相较上一代提升40倍。基于Arm架构的NVID-IA DRIVE Thor平台为文远知行L4级自动驾驶出租车GXR所搭载的联想HPC 3.0高性能计算平台提供算力支撑。HERE Technologies借助基于Arm架构的Amazon Graviton基础设施，更高效地将工作负载从云端迁移到生产环境。

在演讲中，苏姿丰将物理AI视为AI技术领域最严峻的挑战之一，需要构建能够无缝集成多种类型处理器的机器，以理解环境、做出实时决策，并在无须任何人工输入的情况下采取精确行动，且整个过程对误差零容忍。目前，AMD技术已经用于训练模拟物理因果关系的大模型系统，支持World Labs（世界实验室）构建遵守物理定律和动力学的空间智能，并应用于机器人和太空探索。

“交付物理AI需要全栈式的方法，包括用于运动控制和协调的高性能CPU，用于处理实时视觉和环境数据的专用加速器，以及开放的软件生态系统，使开发者能够快速行动，并在平台和应用程序之间无缝迁移。”苏姿丰表示。