达摩院首席科学家孟建熠：

模型创新为算力架构带来新机会

本报记者 杨鹏岳

AI时代架构创新不断涌现，而DeepSeek的出现为整个AI市场带来了全新活力。近日，达摩院首席科学家、知合计算CEO孟建熠在2025玄铁RISC-V生态大会上表示，DeepSeek在一定程度上改变了行业对AI硬件架构的诉求，模型持续创新是所有算力架构的机会。对RISC-V发展而言，打造标杆产品是关键。

DeepSeek让大模型从云走向端

DeepSeek在一定程度上改变了大家对AI硬件架构的诉求，推动行业更加关注底层硬件能力的适配。

孟建熠表示，DeepSeek问世之后，业界产生了三种不同观点的争论：一是开源架构与闭源架构谁更好。DeepSeek证明了开源架构也有很好的表现。二是该用Dense模型还是MoE模型。前者是通用全能的模型，后者是更好的专家模型，二者各有所长。三是算力敏感与内存容量敏感之争。之前模型对算力的需求非常大，现在是容量很大，算力下降了，所以内存容量成为一项关键指标。

“DeepSeek在一定程度上改变了大家对AI硬件架构的诉求。”孟建熠认为。

模型深度优化为算力架构带来了全新可能。具体而言，一是MoE模型，以更低激活比达成更低的计算成本，并使模型的单机部署成为可能。二是稀疏计算与模型压缩技术，识别并跳过模型中不重要的计算节点（如权重接近零的部分节点），同时结合模型压信技术减少参数量。三是混合精度计算与量化技术，浮点计算转化为低精度计算（如INT8、FP8、FP16），同时保持模型精度。四是动态计算图优代技术，实时调整计算结构减少冗余计算。五是内存优化与数据流重构技术，减少内存访问延迟以及数据传输开销。六是分布式计算与负载均衡技术，将大规模模型推理任务拆分到多个计算节点，并通过负载均衡技术优化任务分配。

DeepSeek的出现，推动行业更加关注底层硬件能力的适配。“在很长一段时间里，大家都喜欢‘大炮打蚊子’，当然这样是效率很高，但是今天我们有了一个新思路——可以用软硬件融合的视角来看待整个AI的发展。”孟建熠强调，算力、内存、互联之间原有的平衡发生了剧变，对于新的算力架构机会而言，大家再次进入同一起跑线。同时，开源大模型单机部署成为可能，进一步推动实际应用落地。

另一个趋势是大模型走向趋同，帮助算子收敛。值得关注的是，大模型的参数量和计算量巨大，但如今算子的个数在逐步收敛，主要以矩阵计算为中心，而且通过开源相互学习正在走向趋同。

从云端协同的层面，DeepSeek帮助大模型从云走向端，由此也带来了几个变化：一是从算力瓶颈变为存储的带宽和容量瓶颈，容量瓶颈成为全量大模型最关键的因素，比如671B大模型。二是降低计算资源需求，让几T、几十T和几百T的算力成为可能。三是实现单机部署的可能，能够让开源模型被更多私有数据优化，形成私有解决方案。四是支持边缘设备，使得高性能AI应用能够在边缘设备上顺利运行。“大模型在云端的话，实施成本比较高，只有有限的企业可能在部分领域应用，而一旦到了端侧，就有大量的应用都会发展起来。”孟建熠表示。

模型创新是算力架构的机会

当前算力基础是以CUDA为代表的传统闭源硬件与生态，而开源大模型的不断涌现，给算力架构带来了新机会。

当前算力基础是以GPGPU（CUDA）为代表的传统闭源硬件与生态，而DeepSeek、Llama、Grok等开源大模型不断涌现，给算力架构带来了新机会。当然，这个机会对所有架构都有效，并非只对RISC-V而言。如今，这些开源模型正在吸引更多算力架构，包括x86、Arm这样的传统CPU架构，DSA、ASIC这样的自研架构，以及以RISC-V为代表的开源架构。“我们都在一个新的起点上，现在就看谁能够跑得快。”孟建熠说道。

作为三大主流指令集架构中最灵活、最开放的一个，RISC-V适应了AI时代的技术创新节奏。它在原生AI支持上，拥有开源与开放架构、更易实现的软硬件协同设计、更优的能效比，以及覆盖全球、不断成熟的生态。在孟建熠看来，“RISC-V架构+AI”是以AI原生成为架构创新的最终答案。随着开源RISG-V架构的快速发展，重新自研架构已意义不大，以RISC-V为基础构建处理芯片是未来的主流。

RISC-V在AI领域具备很高的包容性，可以支持做CPU/DSA，也支持做GPU、多核产品或者近内存计算。“大家可以在硬件上不断创新，而生态上统一在RISC-V。尽管目前有不同的实践方案，但最终一定会走向生态统一。”孟建熠表示。

RISC-V如何真正走出来，是备受业界关注的一个问题。发展至今，RISC-V生态已经慢慢建立起来，从IoT等功耗敏感型场景向服务器等高算力场景成长，从纯通用计算向AI计算与通用计算融合成长，其中已经有了一些标杆性产品。

打造标杆产品是RISC-V成长路径的关键环节。孟建熠指出，RISC-V发展中的一个新趋势是从“小标杆产品”向“大标杆产品”成长，引领软件生态加速配适，吸引产业资源加大倾斜。

实际上，国内外企业都在尝试打造RISC-V的标杆产品。比如，国内的达摩院2022年发布了基于玄铁C910 RISC-V核的4核产品，推动了包括安卓在内的软件厂商加入RISC-V生态。国际上，Tenstorrent、Vantana和SiFive等企业也推出一些标杆产品。其中，Tenstorrent最新的RISC-V核“Ascalon”采用了CPU中少见的8-Wide指令集解码器设计。孟建熠认为，下一代RISC-V标杆产品在服务器场景、AI PC场景、AI场景有着一些关键指标。要真正从产业中走出来，性价比很重要。

“标准建设是RISC-V下阶段发展的重中之重。”孟建熠表示，国内产业需要在标准建设中尽快形成合力。目前，国际上在指令架构上的贡献明显高于国内，国内力量的参与度还不够。国内已经建立多个组织，都在进行相关的指令集的制定工作，需要联合起来统一到一个平台工作。另外，技术路线上要考虑相对集中，以AI为目标先做一轮国内标准制定的尝试。此外，计算原语是相似的，所以CPU、GPGPU、TPU在扩展上要形成一定的梯度，不能把指令集做成很多套并行大而全的扩展，这样生态就无法形成。

我国成功开发出世界首款光子时钟芯片

本报讯 芯片的信息处理需要做好时间调控，而调控的速度与精准度，直接决定了芯片的性能。北京大学常林研究团队与中国科学院空天信息创新研究院合作，成功开发出世界首款光子时钟芯片，可将芯片上的时间调控速度提升100倍，从而极大提升未来智能计算、6G通信、空天遥感等一系列现实应用的性能。相关成果日前发表于《自然·电子学》。

“传统芯片要想产生高速的信息处理能力，通常需基于电子的振荡器来产生时钟信号。但从目前来看，该方案的速度并不理想，且会消耗大量功率，产生较高热量。同时，一个芯片往往只能产生一定频率范围内的时钟，导致不同应用，比如6G、车载毫米波雷达、GPU等，需要完全不同的芯片制造技术，从而极大提升了芯片成本。”常林告诉记者，“不同于传统方案，我们开发的光子芯片技术‘以光为媒’，通过光子产生时钟信号。我们都知道，就速度而言，光比电快很多，因此用光子时钟来处理信息，速度比电子时钟快得多。”

常林介绍，该芯片之所以能研制成功，关键在于对“光频梳”技术的“改造”。在过去，这一技术只能依靠昂贵的设备来实现，一台售价几百万元，且只能依赖进口。“我们成功实现了‘光频梳’技术的芯片化，通过在芯片上构建了类似于跑道形状的环，让光在其中以光速不断‘奔跑’，而每跑一圈的时间，就可以作为片上时钟的标准。由于这一时间非常短，通常为1秒的几十亿分之一，因此光子时钟能以超高速进行时间调控。”

“通过这种方案，我们可以基于光来进行芯片上的信息传输与处理，从而极大提升了传统芯片的性能。”常林团队在实验中发现，他们可以只用一个芯片就能覆盖目前所有微波频段的时钟。“这也意味着，这一芯片可以支持从5G到6G，甚至更高速度的手机通信，从而避免了过去每升级一次通信方案，就需要更新一次手机硬件的问题。”

常林还透露：“这一技术的另一个重要应用，是提升计算的主频。目前，无论是GPU还是CPU，主频一般都在2GHz~3GHz，而目前我们团队实现的时钟频率已超过100GHz。这相当于在更短的时间内，我们可以计算更多次数，从而为人工智能发展提供更强算力。”（文 编）

英伟达第四财季营收393.31亿美元 净利润同比大增80%

本报讯 记者**许子皓**报道：北京时间2月27日，英伟达发布了截至1月26日的2025财年第四财季和全年业绩报告，数据亮眼。报告显示，英伟达第四财季营收同比激增78%至393.31亿美元，环比增长12%，净利润为220.91亿美元，同比增长80%，环比增长14%。从全年数据来看，2025财年英伟达营收高达1304.97亿美元，同比剧增114%，净利润为728.8亿美元，同比增长145%。

在英伟达各大业务板块中，数据中心业务无疑是最大亮点。第四财季，其数据中心业务营收达到创纪录的356亿美元，同比大幅增长93%，环比增长16%。2025财年全年，该业务营收更是高达1152亿美元，同比激增142%。英伟达表示，数据中心业务的强劲增长，主要得益于全球对AI计算需求的爆发式增长。无论是大型科技公司构建AI基础设施，还是各类企业加速数字化转型、开展AI相关业务，都对英伟达高性能计算芯片有着旺盛需求。例如，云服务提供商亚马逊网络服务、谷歌云平台、微软Azure等，都在积极引入英伟达的AI加速卡，以满足市场对人工智能激增的需求。

据了解，英伟达的旗舰AI芯片Blackwell系列产品在第四财季创造的营收达到110亿美元，这也是英伟达历史上最快的新产品营收拉升。英伟达CEO黄仁勋在财报电话会议上指出：“市场对Blackwell芯片的需求十分惊人，英伟达已经成功地大规模生产了Blackwell人工智能超级计算机，在其首个季度实现了数十亿美元的销售额。目前已经完全解决了Blackwell的供应链问题，Blackwell AI芯片已经大幅增产。下一代Blackwell Ultra芯片计划于2025年下半年发布。”并且他还提出，“推理”AI为行业带来了新的扩展法则，即增加用于训练的计算能力可使模型更智能，而增加用于深度思考的计算能力则能让答案更智能。这一法则的提出，既凸显了Blackwell芯片在满足市场对AI推理需求方面的关键作用，也从侧面再次回应了DeepSeek与英伟达之间是互需关系，而非对立。

黄仁勋表示：“DeepSeek-R1激发了全球的热情，是一个出色的创新。但更重要的是，它开源了一个世界级的推理AI模型。DeepSeek-R1这样的推理模型，应用了推理时间扩展，未来的推理模型可以消耗更多的计算量。”他认为，DeepSeek的成功恰恰证明了英伟达芯片的实用性，并强调“推理阶段需要大量英伟达GPU和高性能网络”，未来市场需求将随着AI应用扩展而增长。

亚马逊首款量子芯片Ocelot发布 量子纠错成本降低90%

本报讯 记者**许子皓**报道：继谷歌、微软发布量子计算芯片之后，亚马逊也在近日发布了自家的第一代量子计算芯片Ocelot，首次实现了可扩展的玻色子纠错架构，量子纠错成本可降低90%。

据了解，Ocelot芯片由亚马逊AWS量子计算中心与加州理工学院联合研发，核心技术在于采用“猫量子比特”（Cat qubit）架构。这一命名源自“薛定谔的猫”思想实验，其核心是利用超导微波振荡器的量子叠加态存储信息。与传统的transmon量子比特不同，猫量子比特通过增加振荡器中的光子数，使位翻转错误率呈指数级下降，同时通过重复编码抑制相位翻转错误。

亚马逊AWS量子硬件总监奥斯卡·佩恩特认为，目前最大的挑战不仅是构建更多的量子比特，还在于使之可靠工作。制造实用的量子计算机需要将量子纠错放在首位。所以，Ocelot从设计之初就将纠错融入系统架构，通过材料优化和电路创新，使纠错效率提升与硬件成本下降形成正向循环。

亚马逊表示，量子计算机对环境的微小变化或“噪声”极度敏感，像是振动、温度变化、手机的电磁干扰，甚至是来自外太空的宇宙射线等，都可能使量子比特脱离量子态，导致量子计算出现误差。因此，量子计算机十分依赖量子纠错技术，即通过将量子信息以“逻辑量子比特”形式编码至多个物理量子比特中，将量子信息与环境干扰隔离开来。

当前Ocelot芯片仍处于原型阶段，包含5个猫量子比特和4个辅助比特，主要验证量子存储与纠错功能。亚马逊计划下一步扩展至更多量子比特，实现逻辑门操作，并探索与表面码等纠错技术的结合。

佩恩特指出：“因为纠错所需的资源大幅减少，未来根据Ocelot架构构建的量子芯片的成本可能只有现有方案的五分之一，将最多提前五年制造出实用量子计算机，并在10年内实现有实用价值的量子计算。”研究团队估计，利用Ocelot开发成熟的量子计算机所需的资源仅为标准量子纠错方法的十分之一。

Arm发布新一代边缘AI计算平台 可运行超10亿参数模型

本报讯 记者**杨鹏岳**报道：2月27日，Arm控股有限公司发布了Armv9边缘人工智能（AI）计算平台。据悉，该平台可支持运行超10亿参数的端侧AI模型。

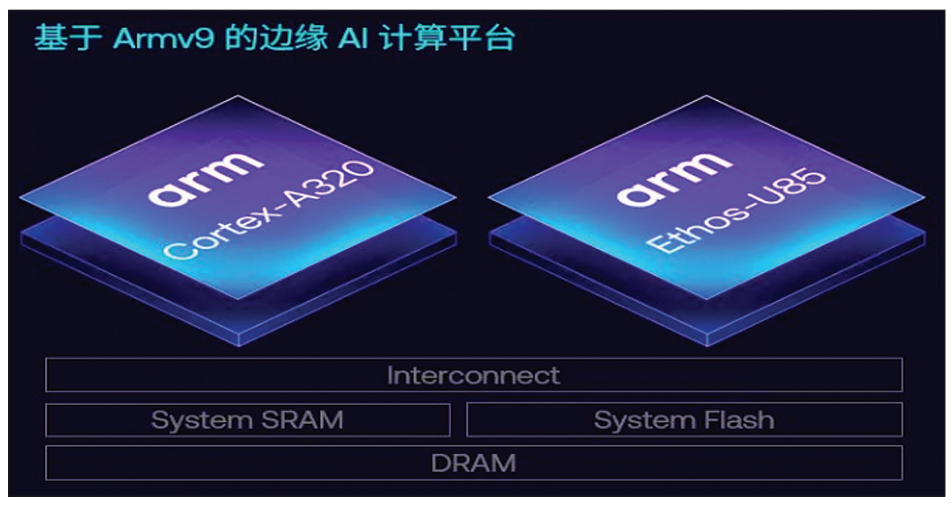
Arm高级副总裁兼物联网事业部总经理PaulWilliamson表示：“AI的革新已不再局限于云端。随着世界的互联和智能化水平的日益提升，从智慧城市到工业自动化，在边缘侧处理AI工作负载不仅带来显著的优势，其必要性更是不可或缺。专为物联网打造的Armv9边缘AI计算平台的推出，标志着这一发展趋势迈入了重要的里程碑。”

据介绍，Arm此次发布的计算平台集成了全新的超高能效Armv9CPU——Cortex-A320和支持Transformer算子网络的Ethos-U85NPU，打造出全球首个专为物联网优化的Armv9边缘AI计算平台。相较于去年推出的基于Cortex-M85的平台，新的边缘AI计算平台的机器学习（ML）性能提高了8倍。Arm表示，从授权该技术以构建SoC芯片的合作伙伴，到构建下一代设备的ODM和OEM厂商，该平台受到了包括亚马逊云科技（AWS）、西门子、瑞萨电子、研华科技和Eurotech在内的多家业界合作伙伴的支持。

记者在现场了解到，该平台能够支持基于智能体的AI应用上运行经过调优的大语言模型（LLM）和小语言模型（SLM），从而开辟全新类别的边缘应用场景。在未来的场景中，智能决策将更接近数据采集源头，这不仅能显著减少延迟，还能有效提升隐私保护水平。

Arm物联网事业部业务拓展副总裁马健表示：“在现阶段AI百模大战时代，焦点在云数据中心的集中式训练。但是训练本身不能产生价值，推理才是AI释放价值的关键。AI推理将从云端下沉到我们身边，无处不在。”

据了解，在工业自动化、智慧城市和智能家居等领域，OEM厂商、软件开发者们正在积极寻求与Arm联手营建边缘AI推理生态系统，释放AI的巨大价值。



社长：张立 社址：北京市海淀区紫竹院路66号赛迪大厦18层 邮编：100048 每周二、五出版 周二8版 周五8版 零售4.50元 全年定价420元 广告部：010-88558848/8808 发行部：010-88558777 广告许可证：京石工商广登字20170003号 发行单位：中国电子报社 印刷单位及地址：经济日报印刷厂 北京市西城区白纸坊东街2号