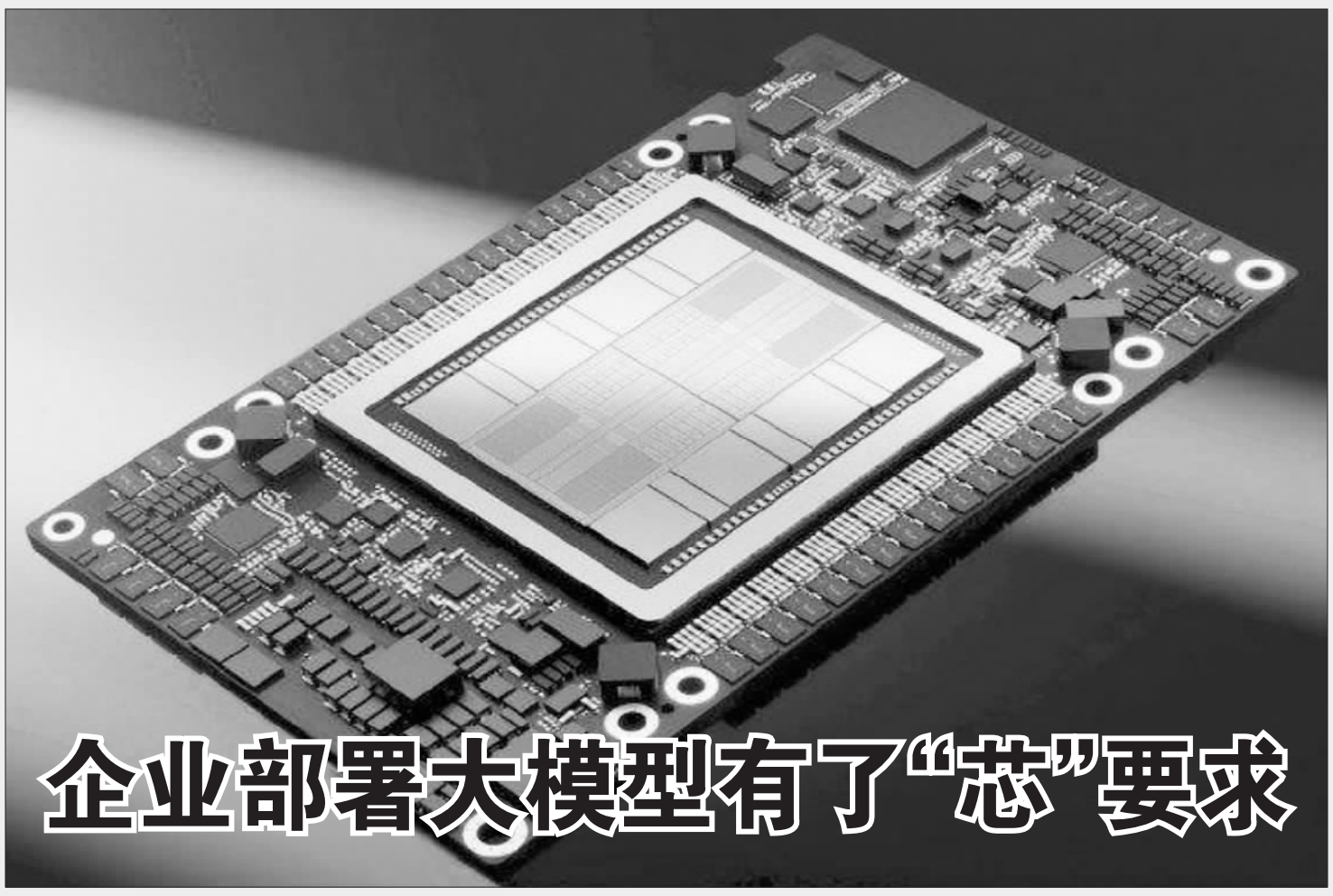

企业部署大模型有了“芯”要求

本报记者 张心怡

在消费者对ChatGPT等消费级AI应用进行尝鲜和玩票之后，企业也迎来了AI发展的转折点。相比2023年企业纷纷训练自己的大模型，2024年企业对大模型的关注点转向推理，以实现差异化和变现。一方面，企业部署大模型给计算架构带来了新的挑战；另一方面，企业对推理的重视，也使其对算力架构的选择走向多元化。

“随着越来越多的通用大模型被训练出来，今年企业的关注点转向了推理。”英特尔公司市场营销集团副总裁、中国区数据中心销售总经理庄秉翰向《中国电子报》记者表示，“我们看到一些客户愿意尝试用CPU做大模型推理，还有一些互联网公司，之前更多提供的是基于GPU的大模型服务，现在也提供基于CPU的大模型，尤其是在推理上。”

企业迎来大模型部署浪潮

市调机构数据显示，预计2026年80%的企业会使用生成式人工智能，至少50%的企业会在边缘计算部署机器学习或者深度学习。

企业级场景正在成为AI大模型的蓝海市场。市调机构数据显示，预计2026年80%的企业会使用生成式人工智能，至少50%的企业会在边缘计算部署机器学习或者深度学习，从而提升企业竞争力。在企业支出方面，预计企业在生成式人工智能的投资将在今年达到400亿美元的规模，到2027年达到1510亿美元的规模。

庄秉翰指出，企业AI的发展可以分为三个阶段。一是AI辅助阶段，AI作为企业的辅助工具，提供会议纪要总结、大纲提炼、文生图等辅助功能。二是AI助手阶段，AI赋能各领域的工作流程自动化，提供行程管理等助手型功能，以及面向客户的智能客服功能。三是全功能AI阶段，AI提供全方位、更精准的自动化服务，为企业经营带来更大价值。

无论是在AI助手阶段还是全功能AI阶段，最大的挑战都在于企业数据与通用模型的结合。

“企业对自己的数据，比如传统的数据如何保存使用，哪些数据应该在公有云或者私有云使用，有很明确的规定。但是，现有AI模型大多是通用模型，一旦企业把数据上传到这些AI模型再做提炼升华，就存在数据泄露的风险，也会对企业的竞争力带来影响。所以我们提出企业AI的概念，其核心在于以更加开放、更具规模性、更加可靠的方式，帮助企业释放AI潜力。”庄秉翰向《中国电子报》记者表示。

按照技术架构，企业AI能力的构建可以分为四个层次。一是兼具可获取性和隐私性的底层算力。二是具备可扩展性和标准化的基础设施。比如企业在私有云部署AI，可能采用单节点、多节点、平均式的部署，或者根据企业的发展规划从单节点小集群逐步走向更大的集群，这就需要算力基础设施具备可扩展性和标准化。三是安全可靠的软件生态。四是便捷开放的应用生态。

面向企业AI的部署需求，英特尔搭建了算力、基础设施、软件、应用四层生态的开放生态堆栈。其中算力生态包含基于酷睿处理器、vPRO商用PC芯片平台的AI PC，基于至强处理器、Gaudi AI加速芯片、ARC显卡的边缘AI与数据中心AI。基础设施生态包含OEM、ODM、CSP（云端服务供应商）、OSV（操作系统集成商）等。

CPU成为AI推理选项

当大模型和生成式人工智能热度退去，需要变现并产生价值的时候，就需要思考落地的经济适用性和最适合的方案。

2023年，AI大模型迎来“百模大战”的盛况。但在大模型走向落地期的过程中，企业越来越注重大模型的投入产出比和后续的盈利能力，这一点也反映在企业对底层算力架构的选择上。

庄秉翰在接受《中国电子报》专访时表示，2023年，企业对大模型的关注聚焦在训练上，更注重性能，对成本和功耗没有那么重视。由于企业都希望训练自己的通用大模型，就出现了“百模大战”的现象。

随着越来越多的通用大模型被训练出来，今年企业的关注点转向了推理。对于企业来说，大模型是需要变现且能够赢利的，但目前市场上的大模型大多基于开源，用作训练的数据也差不多，很难通过差异化实现赢利。而企业AI能够让企业将自身数据融合在大模型的训练过程中，使大模型真正帮助企业解决业务上的问题，增强产品竞争力。

而企业对于大模型赢利能力的重视，也体现在对底层算力架构的选择上。庄秉翰表示，推理基于大规模的算力部署，需要对智算中心的成本、功耗以及整体的运营运维进行考量。在这种趋势下，一些企业正在尝试用CPU做大模型推理。从许多案例可以看到，CPU可以支持130亿参数规模以下大模型的推理。

“对一些企业来说，大模型部署处在初始阶段。如果立即部署一个很大的GPU集群，对于运维和开发来说都是很大的挑战。如果采用逐步部署生成式人工智能的节奏，就可以通过CPU先来做一些大模型的应用部署。当不需要生成式大模型时，还可以转换到通用的应用，这也是一种可以实现赢利的方式。我们看到一些客户愿意尝试用CPU做大模型推理，还有一些互联网公司，之前更多提供的是基于GPU的大模型服务，现在也提供基于CPU的大模型，尤其是在推理上。”庄秉翰向《中国电子报》记者表示。

而算力架构的选择，也与企业类型和所处阶段息息相关。英特尔公司市场营销集团副总裁、中国区云与行业解决方案部总经理梁雅莉表示，企业要因地制宜，选择最适合企业的人工智能策略，并基于该策略选择最适合的基础设施和架构。

“对于头部互联网和大模型公司来讲，今年面临的挑战是大模型的落地和变现。对于其他企业来说，如何挑选合适的大模型融入生产或业务流程以创造价值，是更重要的命题。”梁雅莉向《中国电子报》记者表示，“当大模型和生成式人工智能热度退去，需要变现并产生价值的时候，就需要思考落地的经济适用性和最适合的方案。”

2023年全球Top25半导体供应商名单发布

本报讯 半导体市场研究机构TechInsights近日发布的2023年全球Top25半导体供应商名单显示，上榜企业没有变化，但排名发生明显变化，台积电超越三星位居榜首，总销售额达到692.76亿美元；英伟达成为增长最快的厂商，由第八名跃居至第四名，增幅102%，销售额达到496亿美元；中芯国际上榜。

排名第二到第五的厂商分别是英特尔、三星、英伟达和高通，销售额需要达到59亿美元左右才能跻身榜单。具体来说，2023年台积电营收虽然同比下滑9%，至692.76亿美元，但是英特尔的营收同比更是大幅下滑了14%，至515.05亿美元，这也使得台积电超越英特尔排名第一。

排名第三的三星由于存储芯片业务的影响，其营收同比更是大跌了34%至483.63亿美元。

相比之下，排名第四的英伟达则是2023年营收增长最快的半导体厂商，其营收同比飙升了102%，达到496亿美元，也带动了其排名由去年的第八升至了第四。英伟达营收的暴涨主要得益于其面向数据中心服务器AI GPU需求的巨大增长。

排名第五的高通，其2023年营收同比下滑了16%，至309.13亿美元，这主要是受到了2023年全球智能手机市场下滑的影响。

全球Top25的半导体公司分别是：台积电、英特尔、三星、英伟达、高通、博通、SK海力士、AMD、英飞凌、意法半导体、美光、德州仪器、苹果公司、联发科、恩智浦半导体、ADI、索尼、瑞萨电子、Microchip、安森美、Global-Foundries、联华电子、铠侠、中芯国际、西部数据。可以看到，受2023年全球半导体市场需求下滑的影响，在Top25厂商当中，仅有英伟达、英飞凌、意法半导体、恩智浦半导体、索尼、Microchip等少数厂商实现了营收的同比增长。这些厂商主要受益于来自汽车芯片市场的旺盛需求，索尼则部分受益于高端智能手机市场需求的增长。

从Top25厂商的总部所在地来看，其中有13家供应商的总部设在美国；欧洲、中国台湾地区和日本各有3家；韩国有2家；中国大陆地区有1家。 （微 文）

AMD推出最新自适应SoC 提升边缘计算芯片的灵活性和适应性

本报讯 记者**张心怡**报道：在大模型走向落地的过程中，边缘计算凭借靠近数据源的优势，成为大模型向智能终端、智能网关拓展从而触达更广大用户的重要载体。与此同时，大模型与边缘设备的融合，为边缘侧带来了更高的工作负载，这对本就在功耗和封装尺寸存在诸多限制的边缘计算芯片带来了更多挑战。近日，AMD推出了第二代Versal自适应SoC，包括面向AI驱动型嵌入式系统的Versal AI Edge系列，以应对边缘计算场景繁多，以及嵌入式系统开销受限的痛点。AMD自适应与嵌入式计算事业部Versal产品营销总监Manuel Uhm向记者表示，要适应快速变化的AI行业，需要提升边缘计算芯片的灵活性和适应性，并使芯片架构和封装形式适应嵌入式系统的需求。

“AI处于快速变化中。比如Transformer模型，5年前几乎没有人谈论，现在无论ChatGPT还是生成式人工智能都绕不开它。未来的（主流）模型很有可能是刚刚发端的甚至是全新的，要在这样快速变化的行业生存下来，就需要适应性、灵活性更强的计算平台。”Manuel向记者表示。

在智能物联网时代，FPGA厂商赛灵思（2022年被AMD收购）曾基于FPGA的硬件可编程能力，引入AI引擎，推出了Versal自适应计算加速平台。面向AI时代的边缘计算需求，AMD沿袭了子公司赛灵思“自适应”计算加速平台的理念，并进一步提升了AI计算性能和支持的数据类型。据悉，相较AMD第一代Versal AI Edge芯片，第二代Versal自适应SoC每瓦TOPS最多提升3倍，标量算力最多提升10倍。

“自适应”的核心在于灵活性。Manuel表示，自适应意味着这款SoC能够和多种传感器连接、多种接口连接。基于“可编程逻辑”的特性，自适应SoC能够对硬件进行实时编程。“处理器受限于指令集，只能处理指令集有的内容。而自适应SoC可以对硬件实现定制，适配不同的传感器、不同的性能，通过可编程实现真正的灵活性。”他说道。

嵌入式系统与边缘计算，向来相辅相成，同为实现智能物联网的关键技术。AMD认为，AI驱动的嵌入式系统，其计算加速可分为预处理、AI推理和后处理三个阶段。

其中，预处理是边缘计算的主要环节，对于嵌入式系统尤为关键。嵌入式系统往往搭载了雷达、激光雷达、摄像头等多种传感器，其采集的信息需要经过预处理阶段的传感器处理、传感器融合、数据调节，才能转化为能够被人工智能系统使用的数据。

AI推理使嵌入式系统能够处理更加复杂的工作负载。该阶段面临的挑战是庞大的数据吞吐量，需要在提升数据吞吐能力的同时保障推理精准度，并支持更多的数据类型。

而后处理是对AI推理得出的结果进行决策，需要高性能的嵌入式CPU来执行。

面向不同加速阶段的需求，AMD配置了三种处理器，用FPGA进行预处理，用AI引擎进行推理，再用CPU内核进行后处理。

但如果将三种芯片直接部署在边缘侧，会使本就在尺寸和功耗受到限制的边缘芯片产生庞大的系统开销，包括更高的功率需求和供电复杂性，更高的占板面积和终端系统尺寸，更高的内存需求等。因此，AMD第二代Versal自适应SoC采用了单芯片的封装形式，将预处理、AI推理与后处理集成于单器件中，以实现更小、更高效的AI嵌入式系统。

“（如果采用多芯片系统）需要在不同的芯片当中传输数据，（通信）时延就会延长，而且需要在不同的器件分享数据，需要使用外置的内存，也需要更多的占板面积和功耗。如果将多个芯片集成到一个紧凑的可编程逻辑（单芯片），就能够克服上面的挑战。”Manuel告诉记者。据悉，斯巴鲁公司已选择AMD第二代Versal AI Edge系列用于公司的下一代ADAS视觉系统开发，包括自适应巡航控制、车道保持辅助以及预碰撞制动等功能。AMD预计于2025年上半年提供第二代Versal系列芯片样片，并预计于2025年年末提供量产芯片。

Meta发布新一代AI芯片MTIA 专为AI工作负载设计

本报讯 近日，Meta宣布推出新一代训练和推理加速器（MTIA）。MTIA是Meta专为AI工作负载而设计的定制芯片系列。

去年5月，Meta推出了MTIA v1，为该公司的第一代人工智能推理加速器。MTIA v1旨在与Meta的高质量推荐模型完美配合。该系列芯片可以帮助提高训练效率，并使实际的推理任务变得更容易。

Meta表示，新一代MTIA的计算和内存带宽比以前的解决方案增加了一倍多，同时保持了与工作负载的紧密联系。

据Meta官方介绍，新一代MTIA由8×8网格的处理元件（PE）组成。这些PE大大提高了密集计算性能（是MTIA v1的3.5倍）和稀疏计算性能（提高了7倍）。新一代MTIA设计还采用改进的片上网络（NoC）架构，使带宽加倍，并允许以低延迟在不同PE之间进行协调。

为了支持新一代芯片，Meta还开发了一个大型机架式系统，最多可容纳72个加速器。它由3个机箱组成，每个机箱包含12块板，每块板包含2个加速器。

此外，Meta还将加速器之间以及主机与加速器之间的结构升级到PCIe Gen5，以提高系统的带宽和可扩展性。

从两代MTIA芯片的对比来看，新一代MTIA芯片采用的是台积电5nm工艺技术，拥有256MB的片上内存，频率为1.3GHz；而MTIA v1的片上内存为128MB，频率为800GHz，采用的是台积电7nm工艺技术。

新一代MTIA芯片的平均频率达到1.35GHz，比MTIA v1的800MHz要高出不少，但同时它消耗的功率（90W）也要比MTIA v1（25W）高出3倍多。

Meta相关负责人表示，公司内部正在设计定制芯片，以便与其现有的基础设施以及将来可能利用的新的、更先进的硬件（包括下一代GPU）配合工作。 （文 编）

瑞萨电子重启甲府工厂运营 以满足车用功率半导体需求

本报讯 车用芯片供应商瑞萨电子近日宣布重新启动其位于日本山梨县甲町市的甲府工厂运营。

瑞萨方面表示此举意在提升功率半导体产能，应对不断增长的电动汽车行业需求。

甲府工厂隶属瑞萨电子全资子公司瑞萨半导体制造有限公司，拥有6英寸和8英寸生产线，在2014年10月停止运营。

瑞萨电子方面于2022年宣布斥资900亿日元，将该工厂改建为12英寸晶圆厂，以应对功率半导体领域持续攀升的需求。

该工厂目前洁净室面积1.8万平方米，将于2025年开始量产IGBT、功率MOSFET等功率器件，翻倍瑞萨电子整体的功率半导体产能。

瑞萨CEO柴田英利表示：“从2014年开始停止运营的甲府工厂，在10年后的今天作为300mm晶圆功率半导体产线重新投入运营，对此我感到非常高兴……我们将通过甲府工厂生产的功率半导体，为电动汽车和人工智能普及，以及扩大所需的功率半导体的大规模应用做出贡献。” （瑞 译）

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