IP核出货量达130亿颗 业界期待AI成就RISC-V生态

本报记者 李玉洋 上海报道

近期，两场RISC-V（开源指令集架构）会议在上海、杭州两地接连召开，一场是滴水湖中国RISC-V产业论坛（以下简称“滴水湖论坛”），另一场是2024 RISC-V中国峰会。

“这两个会其实有点区别，滴水湖重在产业落地，RISC-V中国峰会则重在产业的未来，但谈的都是高性能计算。”电子创新网创始人兼CEO张国斌告诉《中国经营报》记者，AI需要大算力，当然会成为RISC-V现在最重要的发展方向。

其中，中国工程院院士倪光南在“2024 RISC-V中国峰会开幕会”表示：“根据2023年年底的数据，在芯片领域RISC-V IP核出货量达到130亿颗，完成了ARM经过30年才走过的历程。”该消息让业界振奋。随着RISC-V在物联网、嵌入式系统等领域批量应用，并在桌面计算、服务器、人工智能等领域迅速发展，未来RISC-V有望成为继X86和ARM之后的第三大主流芯片架构。

而RISC-V国际基金会人工智能与机器学习专委会主席、北京大学讲席教授谢涛则期待AI能成就RISC-V生态，就“像当年PC成就X86生态，手机成就ARM生态”一样。

在半导体行业资深产业分析师黄烨锋看来，在今年滴水湖论坛产品推介中就能看到包含AI SoC、AI CPU在内的RISC-V芯片。“这在任何CPU指令集的发展历程中都是相当罕见的：短短四届滴水湖论坛，人们见证了RISC-V的一路狂飙，现在AI又为RISC-V这辆跑车提供了一次‘氮气加速’（指加速度很快）的机会。”他表示。

RISC-V优势在开放性、灵活性

“像当年PC成就X86生态，手机成就ARM生态，我们期待AI成就RISC-V这样的生态。”

“RISC-V是CPU指令集，谈AI芯片，CPU指令集和它有什么关系？”谢涛表示，今年4月11日，RISC-V国际基金会理事会官宣，人工智能/机器学习是2024年RISC-V国际基金会顶级关键战略最优先的战略。

在今年的世界人工智能大会上，RISC-V国际基金会理事长戴路也表示，RISC-V是最适合AI的指令集架构。此外，加拿大AI芯片独角兽Tenstorrent首席CPU架构师练维汉指出，RISC-V非常适合做AI计算，比如大语言模型的存取非常破碎、混乱，这是CPU最擅长做的。

不仅如此，嵌入式处理器开发商MIPS CEO Sameer Wasson也在滴水湖论坛上表示，作为技术奇点的生成式AI，及数据驱动的资本投入，推动着RISC-V时代的到来。

中国工程院院士倪光南认为，发展RISC-V基础软件是我国软件业新机遇。 视觉中国/图

谢涛认为，基于RISC-V构建AI算力的优势在于其开放性、灵活性，高度可扩展性、功耗和效率优势，以及生态系统和社区的强有力支持。“像当年PC成就X86生态，手机成就ARM生态，我们期待AI成就RISC-V这样的生态。”谢涛说。

“目前，RISC-V AI芯片有两种主要模式：一种是紧耦合模式（integrated），适合低功耗领域（RISC-V+AI），一种是松耦合模式（attached），适合大算力领域（AI+RISC-V）。”谢涛表示，前者以CPU主干为骨架，集成在CPU内部，共享程序计数器、寄存器等流水线单元，只是在执行单元部分增加矩阵或向量单元；后者则外挂在CPU上，会有独立的流水线、寄存器堆、缓存等，是协处理器，可以接收来自及一个或多个CPU的指令，异步执行不同CPU提交的任务。

黄烨锋进一步指出，RISC-V+AI的紧耦合模式就是通过指令集扩展实现AI加速，理论上，ARM公司的Neon、Helium（前者是适用于ARM Cortex-A系列处理器的一种128位SIMD扩展结构，后者是ARM Cortex-M系列产品的MVE一种新的矢量指令集扩展）都属于此类；AI+RISC-V的松耦合模式则在RISC-V CPU的基础上，增加协处理器或加速器——ARM家族的代表是Ethos NPU（该系列是ARM推出的AI微加速器）。

以紧耦合模式的RISC-V AI芯片为例，本次滴水湖论坛展示了来自进迭时空（杭州）科技有限公司的SpacemiT Key Stone K1，号称是“全球首款8核RISC-V AI CPU”，这颗芯片采用了进迭时空自研的RISC-V智算核X60，它拥有8个核心，频率最高2.0GHz，核心单核算力比ARM Cortex-A55高30%。

不过，谢涛也指出，我国乃至全球RISC-V+AI生态仍存在生态碎片化、资源投入严重不足、缺少组织统筹以及产学研协同不够的挑战。

如何撼动CUDA

英伟达的CUDA生态是相对封闭的。对此，谢涛指出历史上能够击败闭源霸主生态的往往是一个开源的生态。

在AI芯片领域，英伟达是绝对的市场霸主，其全球市场占有率高达90%。作为最有希望挑战英伟达霸主地位的公司，AMD仍与英伟达有一个数量级的差距。

从近期的财务数据来看，数据中心业务是这两大GPU巨头增长的主要驱动力，并在2024年保持高速增长。

财报显示，英伟达在2025财年第一财季（自然年2024年2月至4月）创下了260亿美元的季度收入纪录，其中数据中心业务贡献了226亿美元，同比增长427%。作为对比，AMD在2024年第一季度和第二季度的收入分别为55亿美元和58亿美元，数据中心销售额分别达到23亿美元和28亿美元，分别同比增长80%和115%。

一个越来越为人熟知的事实是，之所以英伟达能成为AI芯片市场霸主，除其硬件产品性能优秀外，更在于它构建起了以CUDA（英伟达推出的运算平台）为基础的软件栈。

“相比于英伟达，国产AI芯片除性能差距外，软件生态差距更大。英伟达的成功不仅仅在于其芯片，更在于其软件栈CUDA的成功。”谢涛表示，CUDA是2006年英伟达推向市场的，经过这么多年的发展，英伟达为CUDA生态投入120亿美元，目前CUDA开发者已有450万。

谢涛指出，如今国内高端AI芯片企业达40多家，但软件栈层面各自为战，整体市场份额不足10%。

“一些国产和国际AI芯片公司也会采用所谓的‘打不过就加入’的思路，兼容CUDA软件生态，特别是走GPGPU（通用图形处理器）的路线。这样的道路能解燃眉之急，但长远来看还是受制于人。”谢涛表示，当然还有一些AI芯片公司走的是非CUDA路线，但整体上来说AI算力软件生态呈现小、散、弱的局面。

他还指出，指令集不统一，硬件架构分散；软件栈不统一，用户学习成本高；算子覆盖度低，用户迁移成本高，以及企业各自为战，没有足够的生态竞争力，这些都导致国产AI芯片竞争力的缺乏。

但英伟达的CUDA生态是相对封闭的。对此，谢涛指出历史上能够击败闭源霸主生态的往往是一个开源的生态。

“在IT历史上，当一个闭源生态占据主导地位的时候，基本上没有看到一个成功的例子是说第二个后来居上的闭源生态撼动（原先）霸主生态。但有两大案例，是开源的生态去震撼闭源霸主的生态，一个Linus VS Windows，一个是Android VS iOS。”谢涛说，RISC-V指令集本身是开源的，且已有了相当的芯片出货量及开发生态基础。

言下之意，选择RISC-V做AI芯片的理由又多了一个，即用开源的RISC-V生态来撼动英伟达的CUDA生态。

针对以上当前构建我国RISC-V+AI生态存在的挑战，谢涛认为，可以采用自下而上的思路，以RISC-V指令集扩展+开源系统软件栈（并推成标准）为“公共开源根”，利用国际开放/开源社区“长叶”（基于开源根的商业软件/芯片），形成“根技术开源”与“叶技术竞争”的技术生态优势。

谢涛提出，应聚焦边缘计算和智能终端等多样化应用场景，推动软件生态的发展，进而带动云上软件生态，这种“农村包围城市”的策略来与现有巨擘抗衡，逐步建立RISC-V在AI领域的市场地位。再依托日益强大的RISC-V软硬件生态，聚焦全球开源工具创新，最终达成类似Android VS iOS或Linux VS Windows的竞争格局。

关于具体破局思路，谢涛提出了国际标准+开源社区两抓手。“一是以推动RISC-V国际标准为抓手到国际借力，把握‘根技术’，快速布局新市场（如智能终端、AI PC等），以推动国际基金会标准来依托上游国际开源社区贡献系统软件栈。二是以共建国际开源软件生态为抓手到国际借力，到国际开源软件生态（如Triton、SYCL）中发出中国强声音。”谢涛说。

在谢涛看来，Triton（开源的GPU编程语言）与SYCL（由英特尔主推，和CUDA同层级的跨平台抽象层）是RISC-V AI生态发展的关键，SYCL被他类比为“编程模型中的RISC-V”，相对的CUDA是编程模型中的X86。

黄烨锋指出，Triton实现了硬件无关的中间层表示，生态兼容负担小，编程难度相较CUDA更低，但仍能实现接近于CUDA极限生态的性能。

大模型融资热背后：科技大厂与投资机构的排兵布阵

本报记者 曲忠芳 北京报道

历史宛如一个循环的圆，不断上演着相似的故事，唯一变化的只是故事中的角色。当把目光投向科技领域的投资时，相似的故事与变化的主角同样在交织上演。

近几个月，AI大模型企业的融资一个接一个：月之暗面在8月进行了3亿美元的融资，目前估值达到33亿美元，这是最近6个月里月之暗面的第三次融资，三次融资总额达13亿美元；同样在8月，零一万物完成新一轮数亿美元的融资，最新估值为104亿元人民币；百川智能于7月已完成50亿元人民币的A轮融资，并且以200亿元估值开启B轮融资；智谱AI在今年1月、7月分别完成一轮股权融资……初创型企业借助股权融资进行“输血”和“赛跑”的戏码已经屡见不鲜，因此，《中国经营报》记者综合IT桔子、上奇产业通、天眼查等平台数据以及第三方专家，希望将目光聚焦于背后为“选手”提供“输血”的科技大厂与投资机构，以期在“旧故事”中观察其全新的站位与布局。

大厂主导、押注多方

“不要把所有鸡蛋放在同一个篮子里”是一种常见的投资和风险管理原则，在对大模型独角兽的投资中，科技大厂们也普遍遵循了这一原则。记者仅选择6家“明星”级的AI大模型公司作为考察样本——这些明星公司的特征是创始团队背景靓丽，估值均超过10亿元、部分向200亿元大关挺进，以及均已完成多轮股权融资，它们分别是月之暗面、百川智能、智谱AI、MiniMax、面壁智能、零一万物，记者基于公开信息梳理了它们的融资历程。结果显示，阿里巴巴、腾讯、美团等科技大厂均多次“出手”：阿里巴巴分别投资了百川智能、零一万物、智谱AI、月之暗面、MiniMax；腾讯则是百川智能、MiniMax、智谱AI、月之暗面的投资方；月之暗面、智谱AI的投资队伍中亦有美团的身影。

如果将AI大模型“明星”企业的融资历程表，与上一轮AI领域中计算机视觉识别技术起家的“AI四小龙”融资历程做对比，会很容易发现，AI大模型的投资更多地由科技大厂在“主导”，而非由机构资本所主导。谦询智库创始合伙人龚斌指出，产业发展的规模或曲线其实是类似的，但时间周期在缩短。投资领域一直有一种普遍的情绪叫FOMO：Fear of Missing Out，意为“错失恐惧症”——总是怀疑和担心自己错过什么，更为重要的是，害怕因为错过什么而错失更多的机遇。与前一轮融资相比，无论是对于大厂，还是对机构资本而言，这一轮FOMO的时间窗口更短，在快速的脉冲式“热炒”后迅速回落。当然需要指出的是，受经济金融周期、整体大环境的影响，这一轮的投资实际上更为理智。

关于大厂主导的投资与VC（风险投资）的区别，一位投资界人士解释说，大厂的投资往往更注重战略性布局，会关注投资对象的长期发展潜力与自身业务的协同效应，同时获取或共同开发前沿技术，而不仅仅是短期的财务回报，大厂通常资金雄厚，对投资风险的容忍度相对更高。而VC主导的投资更倾向于财务回报，通常会寻找具有高成长潜力的初创企业进行投资，并在企业成熟后通过IPO或并购等方式退市以实现资本增值。

除阿里巴巴、腾讯、美团外，华为、商汤、小米、知乎、小红书、华策影视等企业或旗下基金也在“明星”大模型企业的投资队列中。值得一提的是，作为大模型明星投资标的的智谱AI，另一个身份则是投资方。截至8月22日，IT桔子收录了智谱AI自2024年4月至今的17次公开“出手”，仅2024年1—7月就投资了8家AI企业。从投资方向上不难看出，智谱AI的投资涉及大模型产业的上中下游，既有行云集成电路、无问芯穹等AI基础层公司，又有包括面壁智能、生数科技等模型层公司，同时还有百奥几何、幂律智能等面向垂直行业领域的公司。

投资机构出手更加谨慎

与大厂的积极、占据主导相比，机构资本的投资在一定程度上则更加谨慎，这或许从更广范畴的数据分析结果中得到一定印证。上奇产业通于8月21日向记者提供的《近三年股权融资情况分析》报告显示，从融资金额数量及增速来看，2021年融资金额54010.83亿元，为近三年最高，之后开始逐年下降；其中，2022年的增速下降至-52.69%，为近三年最低。从融资事件数量及增速来看，2021年融资事件为26035笔，同样是近三年最高，之后开始逐年下降；其中，2023年的增速下降至-44.63%，为近三年最低。从同比变化来看，2024年上半年发生融资事件4184笔、融资总额3647.58亿元，事件及金额均低于近三年同期水平。若按此趋势发展，2024年全年融资总额或将跌进万亿元，相较于2021年的5.4万亿元，预计下降超80%。从参投机构数量来看，2024年上半年有1767家投资机构参投过项目，低于近三年同期水平。投资机构“只看不投”，出手越来越谨慎。

当然需要指出的是，上奇产业通报告称，从融资金额产业分布及融资事件数量赛道排名来看，人工智能占据第三位，近三年里人工智能赛道融资总额达31959亿元，融资事件为18244笔。

值得一提的是，在AI大模型及生成式AI的热潮中，国资背景的资金也在积极出手。天眼查专业版数据显示，2023年年底创立的北京市人工智能产业基金，截至目前已投资9家企业，其中包括智谱AI、面壁智能、百川智能、生数科技、昆仑芯等。

按照北京市人工智能产业投资基金的规划，基金目标总规模为100亿元，围绕北京市在人工智能领域的总体布局开展直接股权投资，重点投向人工智能芯片、训练数据及相关软件等底层技术领域，大模型算法创新、可信AI等关键领域，以及大模型等人工智能技术产品开发和垂直行业创新应用等相关领域。除此之外，上海、深圳等其他地区的国资背景的基金也保持活跃。今年7月刚完成50亿元融资的百川智能也是同时获得北京市人工智能产业基金、上海人工智能产业投资基金、深创投等三地国资基金的一家大模型企业。

国际数据公司（IDC）最新报告显示，2023年中国大模型平台及相关应用市场规模达17.65亿元人民币。报告指出，在过去的一年中，行业对于大模型更多的是早期投入，甚至是观望而不重投入，因此2023年整体市场规模增长并不显著；并且市场格局也主要还是由早期投入者如百度、商汤、智谱、百川等公司构成。进入2024年，头部互联网公司加大对大模型的投入且发起价格战，为早期的大模型初创企业带来一定的竞争压力。预计未来2—3年，市场格局将发生多轮巨变。毫无疑问，无论是大模型公司，以及背后“输血”的科技大厂及机构资本，赛跑仍在持续激烈地进行，谁能笑到最后还有待时间的检验。