

RISC-V 突破不断 正强势崛起为芯片架构第三极

本报记者 李玉洋 上海报道

全球首款 RISC-V 大小核处理器面市、全球首款 RISC-V 笔记本正式交付、全球首款开源万兆 RISC-V 网络交换机亮相、RISC-V 融合服务器全球首发、平头哥推出首个 RISC-V AI 平台……近段时间，RISC-V 产业链不断取得新突破。作为 x86、ARM 之外的芯片架构第三极，RISC-V 正在全球尤其是中国强势崛起。

RISC-V 是一个开发、免费的

RISC-V 生态高速发展

在过去的 2~3 年里，RISC-V 从支持小生态系统的芯片架构，已逐渐进入复杂生态系统，全栈能力不断提高。

在李春强看来，随着 RISC-V 的技术、生态、产业的高速发展，近年来还涌现出一些里程碑式的积极事件。

比如 RISC-V 服务器芯片的发布。基于玄铁 C910 的算力 SG2042 服务器芯片和融合服务器，展现了 RISC-V 在服务器市场广阔的应用前景。RISC-V 开源开放的特性，在高性能方向上得到了越来越多的关注和应

再比如，今年 6 月，由谷歌、英特尔、英伟达、高通以及平头哥等 13 家企业发起的全球 RISC-V 软件生态计划“RISE”，其主要的目的是加速 RISC-V 的软件生态建设及应用商业化进程，将进一步加速

高性能应用落地加速

目前，RISC-V 高性能全栈技术在多领域展开规模化落地。

需要注意的是，在数据中心、服务器等高性能应用领域将会产生新的 RISC-V 需求爆发点。

根据投资机构 ARK Invest 的预测，到 2030 年，ARM 和 RISC-V 可能成为新的处理器标准，在云业务领域取代英特尔 x86 架构，ARM + RISC-V 的组合所占的市场份额，将从 2020 年的零，增加至 2030 年的 71%。

黄焯锋表示，虽然 71% 这个数字有待商榷，但 ARM 和 RISC-V

中国企业和机构是主要贡献者

“很多本土 IC 公司、工具公司都在支持 RISC-V 产业发展，生态日益完善，做了很多好的探索。”

聚焦国内 RISC-V 产业，业界形成了哪些业务发展路线？哪些场景的产品已经落地？

张国斌表示：“很多本土 IC 公司、工具公司都在支持 RISC-V 产业发展，生态日益完善，做了很多好的探索，比如做 RISC-V 架构的 MCU、蓝牙芯片等。”

深度科技研究院院长张孝荣则表示相较于 x86 和 ARM，RISC-V

指令集架构，是由加州大学伯克利分校图灵奖得主 David Patterson 教授及其课题组基于 RISC 的 CPU 指令集架构，历经 30 多年研发、迭代五次而成，2015 年加州伯克利大学将 RISC-V 指令集架构开源。从此之后，芯片架构在 x86 和 ARM 之外又多了一个新的选择。

尽管略显年轻，但 RISC-V 架构在 2022 年年底就实现了 100 亿颗芯片的出货量。“ARM 架构用了 17 年完成了这一里程碑，而 RISC-V 只用了 12 年。”电子创新网 CEO 张

RISC-V 在移动通信、数据中心、边缘计算及自动驾驶等领域的技术和商业化进程。

对于当前 RISC-V 架构的发展历程及特征，李春强表示可大致分为三个阶段，“在 RISC-V 发展早期，大家把 RISC-V 处理器多应用于专用芯片，如 RF 通信、电源管理芯片等；随着 RISC-V 指令集的逐渐完善，越来越多的 IoT（物联网）、MCU（微控制单元）类的 SoC 芯片（系统级芯片）采用 RISC-V，包括蓝牙、Wi-Fi、智能语音芯片等”。

“最近两年，RISC-V 在高性能 AI、AP 类的芯片也得到很好的推进。”李春强指出，随着 RISC-V 从

在数据中心拿下更多市场是板上钉钉的。“数据中心有不同类别的应用，如果再泛化到网络设备，包括 networking、存储、HPC（高性能计算机群）、AI 等，虽然软件栈依赖仍然不见得变少，但对初创企业而言，解决好其中一两个应用市场与生态系统，就有机会将 CPU 产品做起来，那么企业就有机会存活。”他说。

这是亚马逊自研 CPU 能够快速部署到自家服务器上的重要原

的应用领域仍较少，但在国内企业不断探索中，逐渐形成了一些发展路线和落地的产品场景。

“例如，国内一些芯片设计公司已经开始设计和生产基于 RISC-V 架构的处理器芯片，用于嵌入式设备、物联网设备等领域；一些大型互联网企业也开始在自己的服务器和数据中心中采用 RISC-V 架构的处理器，以提

国斌告诉《中国经营报》记者。

“作为开源开放的指令集架构，RISC-V 对 CPU、芯片行业有着深远的意义。”平头哥玄铁 RISC-V 软件研发负责人李春强表示，在短短 10 余年间，RISC-V 处理器的产业应用达到 100 亿颗出货量，显示出这个新出现的处理器架构旺盛生命力。

平头哥是国内 RISC-V 产业链的重要玩家之一。在近日推出首个自研 RISC-V AI 平台时，平头哥生态副总裁杨静甚至说“随着软硬件生态的逐步成熟，创新的形态不断涌现，所有



RISC-V 应用范围正在不断拓展，产业生态不断完善。

本报资料室图

低成本到高性能的不断拓展，RISC-V 能覆盖到对算力要求更高的领域，比如 AI，而参与到 RISC-V 生态中来的企业也越来越多。

“首先是操作系统厂商，包括 Google 开展 Android 系统拥抱 RISC-V 架构等。其次是 AI、车载和通用计算等领域的芯片公司也逐

渐开始关注或使用 RISC-V 指令集架构。最后是更多高性能的应用终端厂家往 RISC-V 转。”李春强说。

可以看到，在过去的 2~3 年里，RISC-V 从支持小生态系统的芯片架构，已逐渐进入复杂生态系统，全栈能力不断提高。“随着软硬件生态的逐步成熟，创新的形态不断涌现，所有芯片都值得用 RISC-V 做一

次。”杨静表示。

多位采访对象向记者表示，在经历多年的快速成长后，RISC-V 架构向高性能 AI、AP 类等芯片（比如以上提及的服务器、网络交换机）推进的趋势愈发明显。“RISC-V 的兴起，在很多人的概念里，应该是自嵌入市场而起。但实际上，在高性能计算市场，加速计算的重要性增强，CPU 与其配合的灵活性也变得重要。”半导体行业资深产业分析师黄焯锋表示，RISC-V 指令集通往高性能市场是必然会发生。

件技术全面升级，并实现 RISC-V 与 Debian、Ubuntu、安卓、OpenKylin、OpenHarmony、龙蜥、酷开 WebOS 等主流操作系统的深度适配。

据悉，目前 RISC-V 高性能全栈技术在多领域展开规模化落地。平头哥携手合作伙伴，实现首个基于玄铁高性能芯片的安卓商业化项目落地，量产多款视频视觉类产品，推出云计算、智能电视等多场景应用。

速引擎的。”李春强表示，RISC-V 在 AI 方向的技术将是众多芯片公司发力的方向。

此外，中国企业和机构已是 RISC-V 国际社区的主要贡献者之一。在 RISC-V 国际基金会中，平头哥投入很大力量，参与了 30 余个技术方向的标准制定，主导负责了安卓、数据中心等 12 个关键技术小组。

业面临的重大机遇和挑战，实现科技自立自强，高质量发展，为科技强国建设和掌握数字经济竞争主动权提供坚实支撑。”

海外相比也有一定差距，IDC 公布的《2023 年第一季度中国企业级存储市场跟踪报告》显示，中国企业级数据存储市场销售额同比增长 3.45% 至 70.14 亿元，全闪存存储销售额 15 亿元，市场占比 25%，混闪存存储销售额 38 亿元，市场占比 54%，相比全球全闪存存储市场份额 41.3% 的局面，中国全闪存存储市场还有很大的发展空间。

配置、均衡发展，才能充分发挥算力的作用。

倪光南表示：“经过存算比的测算，我国存力相对不足，存在重算力、轻存力的倾向。”在他看来，以数据存储能力、信息计算能力、网络运载能力为代表的存力、算力、运力都是我国信息产业发展的核心和基础，是建设科技强国的战略支撑。不仅如此，国内的存力水平与

存储设备，应加强政策引导。

不仅如此，倪光南指出，国产数据存储产业正迎来发展窗口期。存储领域企业应以此为契机，重视对存储产品的安全审查，包括供应链安全、信息安全、数据安全等。尤其是对于存储整机、主控芯片、存储文件系统等相关核心技术，建议参照信创工作经验，实施“自主可控测评”。

快手入局大模型：站内检索先落地 AIGC 提升中

本报记者 李昆昆 李正豪 北京报道

近日，快手公布基于自研大语言模型应用的最新进展：“快手 AI 对话”功能已经在安卓版本开放内测。用户在快手搜索首页右上角 AI 图标即可进入产品首页，选择输入内容或对话或点击引导问题便可开启对话，或者触发“搜索智能问答”卡片，通过底部入口可以启用该功能。

快手 AI 对话

记者在快手 App 里搜索 AI 对话，并问到快手为什么要做 AI 对话，而不是像文心一言那样做独立的 App。快手 AI 对话回复称，“快手选择在快手 App 里做 AI 对话，是因为它更符合用户的使用习惯和需求。快手 App 拥有庞大的用户群体和丰富的内容生态，而 AI 对话能够与这些内容相结合，提供更加智能、便捷、个性化的服务。此外，在快手 App 里做 AI 对话也有利于快手更好地了解用户需求，收集用户反馈，进一步优化和改进产品功能。”

盘古智库高级研究员江瀚告诉本报记者，快手选择将 AI 对话功能集成到其主应用程序中，这可能意味着他们希望为用户提供更综合和无缝的体验，使用户在同一个应用程序中能够完成更多的任务。如果像文心一言那样做独立的 App 可能需要更多的投入和更高的成本。这两种方法各有优缺点，具体取决于快手的策略和用户需求。

“现在在做独立 App 的，也有不做独立 App 的，对于公司来讲，他可能考虑的一个核心原因是成本问题。另外，新的 App 出来以后，还要去吸引用户，做一些推广，这也要花一些成本。”财经专栏作者马继鹏向本报记者说。

快手自研的大语言模型叫

持续投入 AIGC

抖音也在做 AI。近日，字节跳动旗下 LLM 人工智能机器人“豆包”现已开始小范围邀请测试，用户可通过手机号、抖音或者 Apple ID 登录。据称，“豆包”是字节跳动公司基于朱雀模型开发的 AI 工具，提供聊天机器人、写作助手以及英语学习助手等功能，它可以回答各种问题并进行对话，帮助人们获取信息，支持网页 Web 平台、iOS 以及安卓平台，但 iOS 需要使用 TestFlight 安装。

除 AI 对话以外，在短视频平台中，当前整个内容领域或多或少地受到了 AI 的冲击，AI 写作、AI 绘画的技术已经相当普及，AI 短视频、AI 直播的技术离成熟也不远了。比如在抖音平台的视频中，也有不少 AI 制作的虚拟场景，不经提示很难分辨出来。

“大模型做得好不好，需要几个硬实力。首先是芯片层，芯片层在国内做的其实不多，但大部分公司没有能力做芯片。”马继鹏说，在芯片层之外，框架层其实没什么门槛，都是一些开源框架，包括国内大模型，都是在开源框架上训练出来的大模型。

上述人士称，“所以大模型到底好不好，其实除了技术实力之外，最重要的就是训练次数有多少，多少人能参与。ChatGPT 之所以那么火，是因为参与的人数特别多，全球那么多人参与，一上线就有几百万人参与，所以

《中国经营报》记者了解到，“快手 AI 对话”依托于快手站内社区内容生态，通过互动对话的形式，可以帮助用户快速查找短视频、达人、百科等内容，体验全新的信息获取形态。同时，利用大模型背后海量的数据知识，“快手 AI 对话”为快手搜索打破了站内内容的生态壁垒，有望提供全网检索服务，为用户提供全新的信息获取形态。

“快意”（KwaiYi），为业务团队提供标准 API 和定制化项目合作方案。快手内部人士称，“快意”背后的 AI 团队自 2022 年至今已经在国际顶级会议和期刊上发表了 130 多篇论文，覆盖机器学习、自然语言处理、图像识别、语音识别、搜索推广等多项研究领域。

据了解，大语言模型需要四个层：芯片层、框架层、模型层、应用层。快意数据大语言模型属于模型层，快手 AI 对话是在模型层基础上延伸出来的一个应用，属于应用层。

江瀚表示，“快手自研的大语言模型快意与快手 AI 对话之间是密不可分的。大语言模型是实现 AI 对话的关键技术之一，而快手自研的快意大语言模型可能为快手 AI 对话提供了强大的技术支持。因此，我认为快手自研的大语言模型快意与快手 AI 对话之间是一种支持和互补的关系。”

和传统大模型相比，快手 AI 对话是对于搜索新场景的探索。一方面将快手站内大量的内容资源作为索引，解决大语言模型 AI 幻觉的问题，提升回答准确性；另一方面也更加有效的资源组织形式满足用户多元化需求，不仅覆盖生活常识、服务查询等内容，用户还可以进行追问，在个性化的场景中寻找到更适合自己的搜索结果。

会越来越聪明。但是国内其实现在不太一样，监管方面也出台了相关政策。对每一个公司的大模型，根据监管要求，这些大模型在公测时也要符合相关监管规定。”

快手科技创始人兼首席执行官程一笑介绍，基于目前在大模型的技术积累，快手已经实现了多个应用场景的落地。首先是在搜索方面，于 7 月 8 日启动了智能问答产品的内测，8 月 8 日启动“AI 对话”内测，并于 8 月 18 日在快手 App 安卓版开放内测“快手 AI 对话”功能，这是短视频和直播行业首个基于大语言模型落地的智能问答产品，在搜索场景为用户带来智能问答和文本创作等新功能。

其次是在 AIGC 方面，快手已经打造“全模态大模型 AIGC 解决方案”。基于自研的基座大模型，为用户提供包括文本生成、图像生成、3D 素材生成、音频生成、视频生成等在内的多种技术能力。

江瀚称，他对快手在 AI 方面的发展持乐观态度。一方面快手在短视频领域拥有庞大的用户基础和强大的社交属性，这为快手提供了丰富的数据资源，可以为 AI 技术的发展提供有力的支持；另一方面，随着人工智能技术的不断发展，AI 在各个领域的应用越来越广泛，这也为快手在 AI 方面的发展提供了广阔的市场前景。

倪光南：应建立合理的“算存比”

本报记者 秦晁 哈尔滨报道

ChatGPT 发布至今，AI 大模型正在进入全新的生态模式，大模型

重算力，轻存力

随着大模型产业的快速发展，AIGC 模型训练数据量呈现指数级增长，带动算力需求爆发。《中国经营报》记者了解到，模型参数量从 GPT-1 的 1.17 亿增加到 GPT-3 的 1750 亿；训练数据量也由 GPT-1 的 5GB，增加到 GPT-3 的 45TB。这就导致面向 AI 大模型的数据准备时间长、数据来源分散、归集慢。

数据统计显示，目前大模型

掌握先进数据存储主动权

IDC 预计全球数据量到 2025 年将达到 175ZB，其中我国的数据量也将由 2018 年的 7.6ZB 增至 48.6ZB，跃居全球第一，而拥有强大、先进的数据存储产业作为支撑，才能有发展的主动权。

时代，数据决定 AI 智能的高度。作为数据的载体，数据存储成为 AI 大模型的关键基础设施。

中国工程院院士倪光南表示：

算力成本约占整个成本的 25%，而数据清洗、预处理等工作，在不算数据存储硬件的情况下，占到成本的 22%。从这个角度看，数据机器存储过程，在大模型时代越来越重要。

“大家对算力的理解存在一定的片面性。”倪光南认为，真正的人工智能不仅需要算力，还需要存力、运力，三者缺一不可，只有三者平衡

提出“算力”与“存力”的适当比率范围。

我国算力总规模已位居全球第二。工信部发布的数据显示，截至 2022 年年底，我国数据中心标准机架总规模超过 7000 万架，服务器规模约 2400 万台，算力总规模达到 180EFLOPS（每秒 18000 京次浮点运算）。

倪光南指出，要避免大力发展