THE ENTERPRISE OBSERVER
企业观察报

责编 | 张郁
E-mail: yuzhang@cneo.com.cn
美编 | 牛大伟 校对 | 傅克伟

华为穿越 AI 生态周期的两张“底牌”

■ 毛烁

“AI 一天，人间一年。”尽管 AI 在持续狂飙突进，但底层的“算力焦虑”却从未消散。

参数规模的指数级增长、能力的边界不断拓展，让“通用人工智能”走向现实的步伐似乎一夜之间提速。

然而，在算力“狂飙突进”的表象之下，华为正试图用“鲲鹏”与“昇腾”两张“底牌”，给出新的解题思路。

鲲鹏：通用算力从“可用”走向“普惠”

在过去几年，鲲鹏“从有到优”的持续迭代，让其在通用计算领域扮演了重要角色，但智能时代的到来，对其提出了更高要求：如何不仅提供稳定的计算能力，更能为快速变化的智能应用提供高效支撑？

在近日举行的鲲鹏昇腾开发者大会 2025 上，华为 ICT Marketing 部部长周军在鲲鹏开发者峰会上用一组数据描绘了鲲鹏的“基本盘”。他指出，“过去一年，在开发者的共同努力下，鲲鹏产业生态持续繁荣发展，基于鲲鹏技术的应用创新不断，加速行业智能化升级。截至 2025 年 5 月，鲲鹏携手 6300 余家伙伴，孵化了超过 18000 多个解决方案，并广泛应用于千行万业，覆盖国计民生核心场景。”

数字的背后，是鲲鹏在银行、证券、互联网电商、能源等关键领域被逐步接受和应用的真实写照，进一步证明了鲲鹏在“能用”层面的成功。然而，当 AI 庞大的“算力胃口”和异构计算需求扑面而来时，通用算力必须重新思考定位。

周军也直言：“AI 时代，算力的需求无处不在，AI 应用的算力底座架构逐步走向异构融合。开发者需要简单易用的工具调用算力能力，实现算力的高效协同。”这也反映了通用计算在智能时代临的根本性策略：从单一向多元、从固定向融合的转变。

鲲鹏的“新解法”之一，是“鲲鹏 AI+ 解决方案”。其核心逻辑是——通用计算不再是智能计算的“旁观者”，而是其深度融合的支撑者。

这种融合并非简单的硬件“捆绑”，而是在 AI Core、AI Infra、AI 应用三个层面进行系统性优化。

在 AI Core 层面，鲲鹏推出了“鲲鹏 +xPU 推理方案”，强调其对昇腾及其他通用处理器的兼容性，覆盖数据中心到边缘的广泛场景。这意味着，企业无需在通用和智能算力之间进行艰难取舍，而是可以在一个统一的框架下实现算力的弹性调度和高效利用。趋境科技依托鲲鹏基础软硬件平台，在 Ktransformer 架构优化中取得突破性进展，并正式发布了鲲鹏 +xPU 解决方案。这展现了鲲鹏作为“多元算力调度者”的潜力。

在 AI Infra 层面，鲲鹏的重心落在“效率”与“安全”上。其提供的 AI 数据工程组件和 AI 安全组件，旨在全面提升数据处理效率，并构建多重安全防线，保护智能模型和私域数据安全。在大模型对数据质量和安全提出更高要求的背景下，这无疑是抓住了企业智能落地的“七寸”。

而在 AI 应用层面，鲲鹏则致力于打造“开箱即用、开箱即优”的解决方案。鲲鹏全新发布的“搜广推”解决方案，正瞄准了互联网核心的推荐场景。

更值得关注的是，此次“鲲鹏 RAG 解决方案 1.0”正式发布。

RAG（Retrieval Augmented Generation）作为当前大模型落地应用的关键技术，可通过外部知识增强（外挂知识库），解决大模型的“幻觉”和知识更新问题。

鲲鹏 RAG 解决方案基于鲲鹏处理器、昇腾处理器及第三方通用处理器，打造了包含业务编排、AI 微服务、部署调度、安全和存储的开源开放体系，意在定义企业级 RAG 的技术标准。

应用上，会议期间多家企业介绍了基于鲲鹏 RAG 解决方案 1.0 打造的各自行业的差异化 RAG 解决方案。这表明鲲鹏正在从底层能力提供商，向行业解决方案的“联合设计者”和“使能者”转变，直接深入到客户的业务场景中，解决实际问题。

此外，为了解决大规模集群的管理复杂性，openFuyao 社区正式开源，这将极大地降低企业部署和管理鲲鹏算力集群的门槛，从而加速鲲鹏算力的普及。这也体现了华为从“硬件堆叠”到“软件赋能”的思维转变。

鲲鹏的变革，核心在于如何从满足通用计算需求，跃迁到为智能应用提供普惠、高效、安全的混合算力支撑。这是技术的深度革新，更是生态战略的再一次定位。

昇腾：智算“生态奇点”下的“自生长”与“广阔触达”

如果说鲲鹏是智算“骨架”，那么昇腾就是其跳动的“心脏”。

作为华为智能战略的核心，昇腾承担着为 AI 应用提供强大算力支撑的重任。在昇腾 AI 开发者峰会上，周军再次强调了华为在智能领域的战略：“华为始终坚持‘硬件开放、软件开源、使能伙伴、发展人才’的生态策略，持续投入根技术创新和系统架构创新，携手伙伴和开发者，共同打造开放繁荣的计算产业生态。他指出，截至 2025 年 5 月，鲲鹏、昇腾已发展超过 665 万开发者，8800 多家合作伙伴，完成 23900 多个解决方案认证。”

其中，昇腾的 300 万开发者，亦是其生态从“拓荒期”迈入“加速期”的有力证明。

然而，智算的挑战在于其快速迭代和高门槛。智能平台和工具的易用性、灵活性和可扩展性，成为了开发、部署和运营的重要基础。昇腾正是在这一背景下，持续增强基础软硬件能力，打造好用易用的开发平台和工具，加速行业应用创新。

在核心技术层面，昇腾从底层架构到行业落地进行了全链条革新。通过 CANN 分层开放、超节点架构的极致效能，以及 MindSpeed RL、MoE 并行推理等前沿工具的发布，昇腾正为全球开发者构建“所想即所得”的创新土壤。

在应用部署层面，MindIE Motor 推理服务加速库和推理微服务 MIS 的发布，致力于让应用部署更加流畅高效。行业挑战上，大模型的部署和推理，往往伴随着巨大的资源消耗和复杂的调度管理。MindIE Motor 和 MIS 将帮助开发者更高效地将大模型部署到昇腾平台上，并实现高性能的推理服务。

这直指大模型落地过程中最常见的“最后一公里”问题，旨在让开发者从繁琐的部署工作中解放出来，更专注于模型本身。

在模型训练和推理层面，昇腾持续升级分层开放 CANN 的能力、MindSpeed RL 强化学习套件，以及大规模专家并行推理解决方案，旨在让模型训练和推理更为高效。尤其是针对大模型训练和推理的优化，这关系到智能应用的性能瓶颈突破，也是在核心技术层面构建竞争壁垒的关键。

昇腾超节点架构以突破性创新打破集群互联瓶颈，通过技术革新实现节点间高效协同，让集群运行如同一台强大计算机，大幅提升整体计算效率。其构建业界最大规模 384 卡高速总线互联体系，相比传统节点，训练性能实现 3 倍飞跃，以强劲算力支撑大规模 AI 任务快速推进。同时，超节点架构深度适配 MoE，充分释放 MoE 模型潜力，为模型训练与推理提供高效支持，使昇腾成为 MoE 模型开发与应用的最优选择，在 AI 计算领域树立新标杆。

科大讯飞星火大模型训练工程资深技术专家张海俊分享了其实践经验。他表示：“科大讯飞与昇腾的合作已深入到大模型训练的方方面面。双方不仅在算子开发上与昇腾团队紧密协作，更在模型部署和推理优化上取得了显著进展。”

清华大学博士生、vLLM 社区 Maintainer 游凯超也表达了对昇腾地期许，他指出：“vLLM 在推理加速方面做得非常好，我们看到昇腾也正在加速库的适配。我们现在看到了与 vLLM 的深度合作，第三方的发展非常快，希望昇腾持续的保持关注。”

中国工商银行软件开发中心大数据和人工智能实验室资深经理夏知渊则从行业应用的角度提出了期待：“我们希望昇腾的生态能够进一步完善，让更多的行业应用能够快速地迁移和部署到昇腾平台上，从而加速金融行业的智能化转型。同时，我们也希望昇腾能够提供更完善的开发者工具和社区支持，帮助我们解决在实际应用中遇到的各种问题。”

来自金融行业用户的声音，代表了对昇腾生态成熟度和易用性的更高要求，也指明了未来昇腾生态建设的重点方向：从“能用”到“好用”，再到“省心用”，最终实现“自生长”。

面对这些期待，华为昇腾计算业务总裁张迪煊明确表示：“昇腾 AI 的发展需要‘根’扎得更深，‘枝叶’更繁茂。昇腾致力于成为智能基础设施的坚实底座，打造繁荣的智能应用生态。我们坚持‘硬件开放、软件开源’，持续深耕 CANN、MindSpore 等基础软件平台，并与伙伴和开发者一起，共同推动智能产业的发展。”

诚然，昇腾在技术“深耕”与生态“广拓”上的平衡策略，透露出其对构建长期价值的追求。

而昇腾的“生态裂变”也不仅在于技术上的持续突破，更在于生态的深度开放与协同。上海交通大学副校长、鲲鹏昇腾卓越中心管委会主任管海兵指出：“上海交通大学作为鲲鹏昇腾卓越中心的重要合作伙伴，一直致力于推动产学研深度融合。”他进一步指出，我们与华为共同开展了多项前沿研究，并在人才培养方面取得了显著成效。未来，我们将继续深化合作，为中国 AI 产业发展贡献力量。”

硅基流动创始人兼 CEO 袁进辉则更为直接地指出了选择昇腾的考量。他坦言道：“从去年上半年创业开始，我们认为在中国长期做基础设施服务，必须选择没有‘后顾之忧’的通用处理器和算力。所以，我们当时就决定全力拥抱昇腾。”

经过不到一年的实战，效果显著。袁进辉表示：“如今，我们团队已经完全掌握了在昇腾上编程、优化和适配的技巧，快速推出各种模型更是得心应手。实际上，我们使用的昇腾卡甚至超过了其他类型的卡。”这不只是技术上的认可，更是基于对长期发展和创新的深思熟虑。

华为首席战略架构师党文栓则勾勒了更宏大的图景：“华为的目标是打造一个开放、协作、共赢的 AI 生态。”他坦言，任何一家公司都无法单枪匹马完成所有创新，只有与全球开发者和伙伴紧密合作，才能真正推动 AI 技术和产业的进步。

诚然，昇腾要的不是一枝独秀，而是百花齐放。这也是为什么越来越多企业，选择与昇腾“同频共振”的原因。

鲲鹏与昇腾，从“两条腿”到“一体两翼”

鲲鹏昇腾开发者大会已经举办了三个年头，之所以将鲲鹏昇腾放在一起，背后蕴含着华为对未来计算架构的深刻洞察：通用计算与智能计算的深度融合，将是实现智能普惠化的关键路径。鲲鹏与昇腾，正持续走向协同发展，构建一套更加立体、更有韧性的计算体系。

事实上，鲲鹏与昇腾，正在从底层技术到上层应用，实现全方位的深度融合，共同打造一个面向智能时代的全栈计算基础设施，让开发者能够更简单、更高效地使用多元算力，加速创新。”

“全栈”二字，或许明确了华为在计算领域的布局，意在打通底层硬件到上层应用的全部链条，构建一个无缝的衔接体系。

这种融合是软硬协同、生态互通的系统性工程。其底层便是基于鲲鹏处理器、昇腾处理器以及第三方通用处理器的混合架构。

这意味着，无论是通用计算任务还是智能推理任务，都可以在统一的框架下得到优化和调度，从而实现资源利用率的最大化和应用性能的最优。这为解决大模型时代多元化、复杂化的算力需求提供了有效方案，避免了“各自为政”带来的效率损耗和资源浪费。

华南理工大学计算机科学与工程学院教授、博士生导师陆璐，从学术角度肯定了这种融合的必要性。他指出，大模型时代，单一的算力很难满足复杂计算需求。鲲鹏和昇腾的融合，提供了一种更高效、更灵活的异构计算平台，为科研和产业应用提供了坚实的基础。我们期待看到更多基于这种融合架构的创新应用出现。

从更宏观的战略层面来看，鲲鹏与昇腾的融合，是华为对未来算力基础设施的预判和布局。当智能模型变得越来越大、越来越复杂，对算力的需求也水涨船高。单一的通用处理器算力在面对大规模智能训练和推理时会显得力不从心，而单一的智能专用算力在处理通用业务逻辑时则可能效率低下。

因此，将二者有机结合，形成一个既能处理通用业务，又能高效支撑智能任务的融合计算基础，才是应对未来挑战的根本之道。这种“一体两翼”的协同，正是华为在计算领域构建核心竞争力的关键底牌。

这种融合也体现在对开发者工具链的持续优化上。

过去，开发者在鲲鹏和昇腾平台上进行开发时，可能需要面对两套相对独立的工具和开发流程。而现在，随着鲲鹏 DevKit 的智能能力支持台账自动分析、自动配置和语言自动改造，以及昇腾 CANN 的持续升级和各类加速库的发布，开发者将能够在一个更加统一、自动化程度更高的环境中进行开发，极大地降低了开发门槛和工作量。

这种集成化的开发体验，是吸引更多开发者投入生态，并高效进行创新的关键一步，也是从“各自为营”走向“统一调度”的必经之路。

生态张力：开放共创，构建“信任护城河”

无论是鲲鹏还是昇腾，其成功的关键都离不开健康、繁荣的生态系统支撑。华为深知，仅凭一己之力难以构建起一个庞大的计算产业。因此，“硬件开放、软件开源、使能伙伴、发展人才”一直是其长期坚持的生态策略，这不仅是口号，更是其行动准则，贯穿始终。

本次大会上，华为再次强调了对生态伙伴和开发者的承诺。除了“三年 30 亿”“同辕开发激励计划”，还有更多具体措施落地。例如，鲲鹏与麒麟操作系统、TDSQL、长亮科技等合作伙伴在金融核心交易领域的联合解决方案，已帮助客户在关键场景和关键业务性能上提升 20%。

这为更多合作伙伴提供了可复制的成功路径，形成了正向的示范效应。这些“样板”的成功，是生态复制和扩散的基础，也是构建“信任护城河”的基石。

科大讯飞副总裁、AI 工程院院长表示，科大讯飞与华为在智能领域有着长期而深入的合作。他指出，昇腾作为星火大模型的重要算力基础，提供了强大的支持。

这种紧密的合作关系，不仅体现在大型企业和机构之间，也深入到个人开发者和高校。鲲鹏日、鲲鹏创想日等活动，以及各类线上线下学习资源，都在为开发者提供学习和成长的机会。

不难看出，华为的目标，是打造一个充满活力的开发者社区，让每一个有创新想法的开发者，都能在鲲鹏昇腾的平台上找到舞台。这是一种通过培养人才来壮大生态，进而实现“涓涓细流汇成江海”的长期策略，也是为未来储备的创新力量。

生态的开放，不光意味着技术的开放性，更意味着商业模式的开放和合作机制的健全。华为通过与 ISV（独立软件开发商）、SI（系统集成商）等各类伙伴的深度合作，共同打造面向行业的定制化解决方案。

这种模式，使得鲲鹏昇腾的技术能力能够快速渗透到各个行业，解决实际业务痛点。这是从底层技术到上层应用，形成价值闭环的关键，也是生态张力的直接体现。

然而，生态建设并非一蹴而就，它是一个需要长期投入和持续耕耘的过程。正如袁进辉所言，要要让大多数开发者“不经常麻烦华为的工程师”，这需要海量的开源项目、完善的文档以及真正开放的接口。

这代表着，华为在生态建设上，还需要持续投入和打磨，才能真正实现生态的“自生长”和“自我驱动”，让更多创新在开放环境中自然涌现，而不仅仅是依靠华为的直接推动。

无疑，“放权”与“赋能”是生态成熟的更高境界，也是构建长期信任的关键。E

（文章来源：科技行者 techwalker.com）