Qualcomm

2023 年 5 月

混合 AI 是 AI 的未来

第二部分：

高通在推动混合 AI 规模化扩展方面独具优势

目录

1 摘要

正如白皮书第一部分所言，在云端和终端进行分布式处理的混合 AI 才是 AI 的未来。混合 AI 架构，或仅在终端侧运行 AI，能够在全球范围带来成本、能耗、性能、隐私、安全和个性化优势。

高通正在助力实现随时随地的智能计算。高通技术公司作为终端侧 AI 领导者，面向数十亿手机、汽车、XR 头显与眼镜、PC 和物联网等边缘终端提供行业领先的硬件和软件解决方案，对推动混合 AI 规模化扩展独具优势。高通的硬件解决方案具有行业领先的能效，智能手机解决方案的能效与竞品对比，大约有两倍的优势。凭借一系列基础研究，以及跨 AI 应用、模型、硬件与软件的全栈终端侧 AI 优化，我们的持续创新让公司始终处于终端侧 AI 解决方案的最前沿。

高通技术公司还专注于为全球数十亿、由高通和骁龙®平台支持的终端提供开发和部署的简便性，从而赋能开发者。利用高通 AI 软件栈，开发者可以在我们的硬件上创建、优化和部署 AI 应用，一次编写即能实现跨我们芯片组解决方案的不同产品和细分领域进行部署。凭借技术领导力、全球化规模和生态系统赋能，高通技术公司正在让混合 AI 成为现实。

2 高通技术公司是终端侧 AI 的领导者

凭借赋能数十亿边缘终端的终端侧 AI 领导力，高通技术公司正在助力打造混合 AI 新时代。可扩展的技术架构让我们能够采用一个高度优化的 AI 软件栈即可在不同终端和模型上进行工作。我们的 AI 解决方案旨在提供最佳能效，让 AI 无处不在。

高通 AI 引擎是我们终端侧 AI 优势的核心，它在骁龙平台和我们其他众多产品中发挥了重要作用。高通 AI 引擎作为我们多年全栈 AI 优化的结晶，能够以极低功耗提供业界领先的终端侧 AI 性能，赋能当前和未来的用例。搭载高通 AI 引擎的产品出货量已超过 20 亿，赋能极为广泛的终端品类，包括智能手机、XR、平板电脑、PC、安防摄像头、机器人和汽车等。[1]

高通 AI 软件栈将我们所有相关的 AI 软件产品集成在统一的解决方案中。OEM 厂商和开发者可在我们的产品上创建、优化和部署 AI 应用，充分利用高通 AI 引擎性能，让 AI 开发者创建一次 AI 模型，即可跨不同产品部署。

[1] https://www.qualcomm.com/products/mobile/snapdragon/smartphones/mobile-ai

骁龙和高通品牌产品是高通技术公司和/或其子公司的产品。

2.1 持续创新

我们开发的低功耗、高性能 AI，已经形成了一个跨智能手机、汽车、XR、PC、笔记本电脑以及企业级 AI 等现有市场和新兴领域的庞大终端 AI 生态系统。多年来，我们在照片与视频拍摄、先进连接、语音指令、安全和隐私等关键用例领域，持续利用 AI 赋能芯片组产品、打造差异化优势，以获得市场领先地位。

2.1.1 我们 AI 技术的发展历程

高通深耕 AI 研发已超过 15 年。在高通 AI 研究[2]，我们的使命是实现 AI 基础研究突破，并实现跨行业和用例的规模化扩展。高通正在推动 AI 进步，让感知、推理和行为等核心能力在终端上无处不在。我们的重要 AI 研究论文正在影响整个行业，推动高能效 AI 发展。通过汇聚领域内的杰出人才，高通正在不断突破 AI 可能性，塑造 AI 的未来。

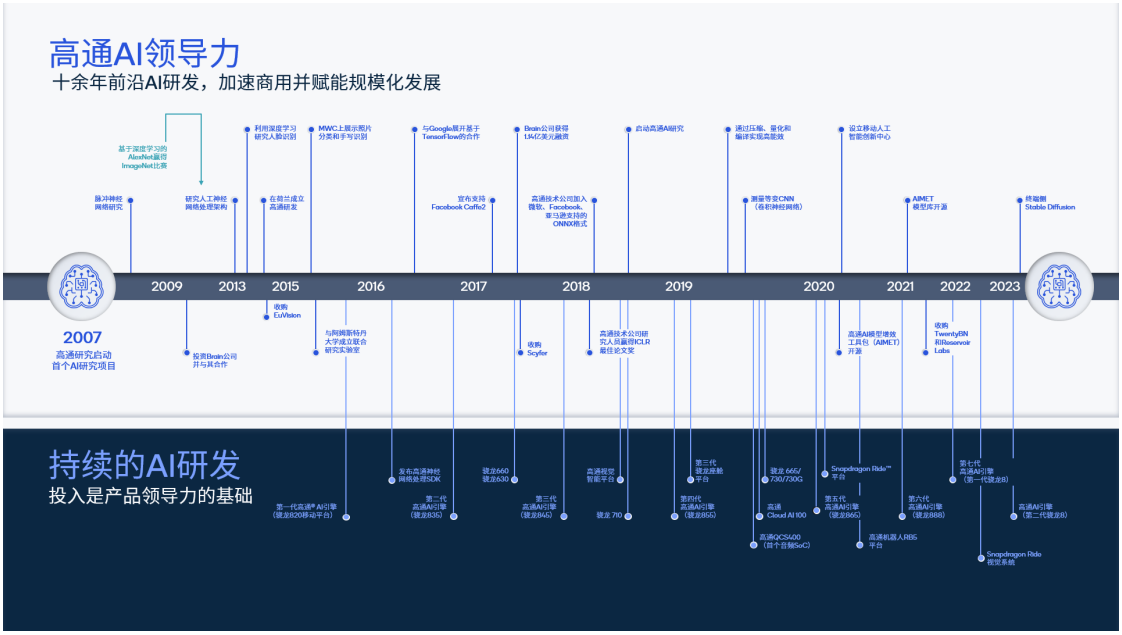

图1：高通持续的 AI 研发投入是产品领导力的基础。

3 我们在终端侧生成式 AI 领域的领导力

多年来，高通 AI 研究团队一直在探索生成式 AI。生成式 AI 可追溯到生成式对抗网络（GAN）和变分自编码器（VAE）。最初，我们探索了生成式模型是否能够很好地压缩，并进一步提升生成痕迹（Artifact）的感知效果。我们利用 VAE 技术创建更好的视频和语音编解码器，将模型规模

[2] 高通 AI 研究是高通技术公司的机构。

控制在 1 亿参数以下。我们还将生成式 AI 理念延伸到无线领域来替代信道模型，让通信系统更加高效。

近期，我们已在终端侧实现支持超过 10 亿参数的生成式 AI 模型，比如 Stable Diffusion，并计划未来在终端侧支持参数高达数百亿的模型。我们不仅在研究如何将生成式 AI 模型用作通用代理来构建计算架构并使用语言来描述相关任务和行为，同时也正在研究如何能够通过增加感知输入（比如视觉和音频），进一步开拓这一能力以及环境交互能力，比如对机器人生成指令或运行软件。

3.1 突破终端侧和混合 AI 边界

高通技术公司具有独特专长，我们能够提供在边缘侧终端上低功耗运行生成式 AI 所需的处理性能，例如大语言模型（LLM）等。若要让生成式 AI 得到广泛采用，就不能像目前这样仅在云端进行推理，还必须在终端侧进行大量 AI 处理。为了让生成式 AI 融入日常生活，AI 处理需要同时使用云端和终端。最终，AI 能力将成为用户选购下一款手机、PC 或汽车的主要影响因素。

通过 AI 硬件加速和简化开发的软件解决方案（比如高通 AI 软件栈），高通已经在引领终端侧 AI 推理。目前，我们能够支持在终端侧运行参数超过 10 亿的模型，预计在未来几个月，终端侧将可以支持超过 100 亿参数的模型。

我们的 AI 加速架构具备灵活性和稳健性的特点，能够应对生成式 AI 模型架构的潜在变化。随着大语言模型和其他生成式 AI 模型持续演进，高通 AI 软件栈和技术将随之不断发展。能够轻松开发混合 AI 应用是关键所在，而我们跨产品组合的通用 AI 架构以及 AI 工具正是面向这一未来而设计。

3.2 负责任的 AI

高通力求创造能为社会带来积极影响的 AI 技术。高通的终端侧 AI 愿景基于透明、负责、公平、管理环境影响和以人为本等原则，我们的工作将产生广泛深远的影响，因此我们致力于负责任地管理 AI，并采取措施以规避潜在危害。高通终端侧 AI 解决方案旨在赋能增强的隐私性和安全性，这对打造稳健可信的 AI 生态系统至关重要。

高通密切关注并配合参与全球各地政府的监管框架、指导方针和最佳实践，包括政府间政策指导（比如，世界经济合作与发展组织推出的《人工智能发展建议》）和区域与国家框架（比如欧盟制定的《人工智能法》和美国国家标准与技术研究所发布的《人工智能风险管理框架》）。这些

法规和政策指导方针为负责任地开发和部署 AI 技术提供了重要的法律和道德考量标准。遵守 AI 法规和最佳实践是高通致力打造道德、负责的 AI 创新的基础，我们的工作实践将持续看齐不断演进的 AI 治理格局。

最后，作为我们参与和领导行业协作、标准机构组织和联盟的一部分，高通支持并倡导 AI 标准、数据与隐私保护和稳健的网络安全。一直以来，高通深知拥有稳健的综合性标准，对于指导负责任的新技术开发部署具有重要意义。

携手合作开发稳健有效的 AI 标准，是迈向打造可持续且可信赖的 AI 生态系统的关键一步。

4 卓越的终端侧 AI 技术和全栈优化

高通为应用、神经网络模型、算法、软件和硬件进行全栈 AI 研究和优化。异构计算方法利用硬件（比如 CPU、GPU 和 AI 加速器）和软件（比如高通 AI 软件栈）来加速终端侧 AI。我们的团队跨上述全部领域联合工作，共同开发最为优化的解决方案。

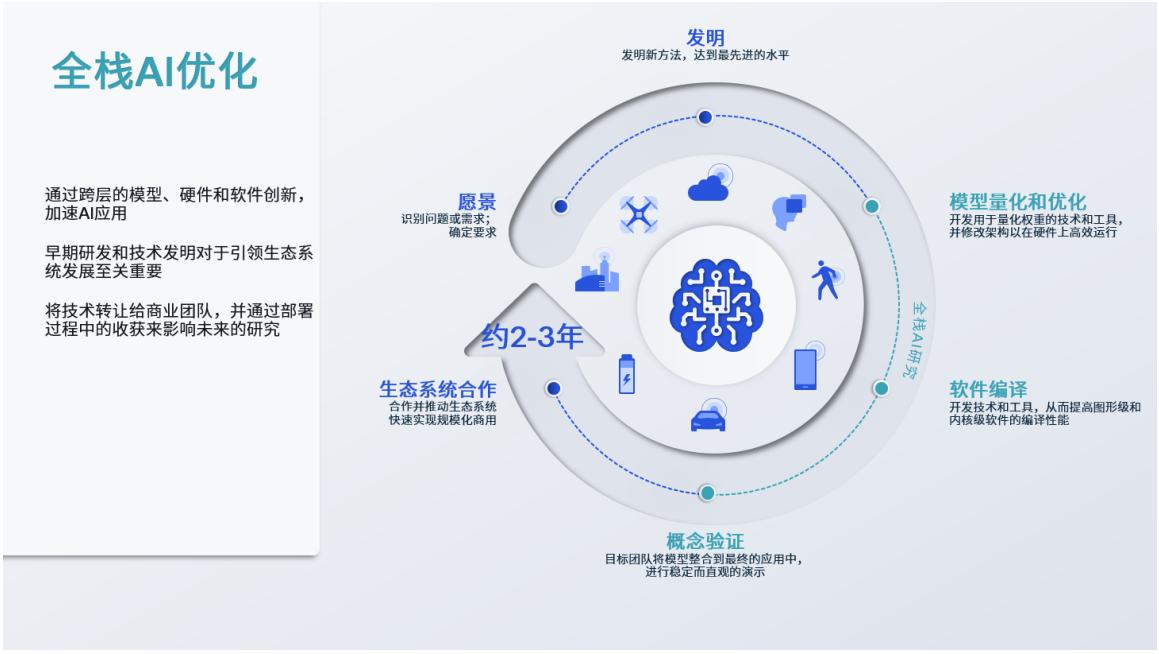

图2：高通全栈 AI 研究和优化赋能技术持续改进并引领高能效解决方案发展。

上图展示的循环创新方式让我们能够基于最新神经网络架构，针对硬件、软件和算法持续改进高通 AI 软件栈。高通在 AI 基础研究方面具备独特能力，能够支持全栈终端侧 AI 研发，赋能产品快速上市并围绕终端侧生成式 AI 等关键应用实现优化部署。

高通演示的全球首个在 Android 智能手机上运行的 Stable Diffusion，突显了我们全栈策略的优势。所有让 Stable Diffusion 实现 15 秒内完成终端侧运行的全栈研究和优化，现已集成进高通 AI 软件栈，并将助力提升未来硬件设计。此外，让 Stable Diffusion 能够在手机上高效运行的优化方式也可以用于其他平台，比如高通技术赋能的笔记本电脑、XR 终端和几乎任何其他终端。

4.1 算法和模型开发

高通研究团队从事神经网络架构开发和调整工作，以在不牺牲准确度的前提下提高效率，例如动作识别和超级分辨率。

面向动作识别设计的传统深度学习模型会逐帧、逐层地处理视频序列，虽然这会带来准确的处理结果，但它是计算密集型的、时延高，并且能效低。高通现已推出的 FrameExit 模型能够自主学习，针对较简单视频处理更少帧，针对较复杂视频处理更多帧，以减少能耗并提高性能。除模型结构创新之外，高通全栈 AI 优化还包括最先进的量化技术和创新的编译器（compiler）栈。我们在移动终端上演示了这一技术，在常用动作识别基准测试平台上相较于其他方法计算量和时延（平均）可减少五倍。

面向高清屏幕上的游戏和视频播放等应用，超级分辨率能够让图像更清晰、锐利，实现分辨率升格。尽管基于 AI 的超级分辨率相比传统解决方案能够实现出色的视觉质量，但在移动终端上实时运行颇具挑战性。高通对 AI 全栈进行了优化，包括基于我们 Q-SRNet 模型的算法、采用 INT4 量化的软件，以及支持 INT4 加速的第二代骁龙 8 硬件。我们利用 INT4 模型实现全球首个实时超级分辨率终端侧演示，大幅改善了时延和功耗。实际上，与 INT8 相比，INT4 性能和能效提高了 1.5 倍至 2 倍。

4.2 软件和模型效率

高通 AI 软件栈旨在帮助开发者实现一次开发，即可跨高通所有硬件运行 AI 负载。高通 AI 软件栈全面支持主流 AI 框架，比如 TensorFlow、PyTorch、ONNX 和 Keras，以及包括 TensorFlow Lite、TensorFlow Lite Micro 和 ONNX Runtime 等在内的 runtime。此外，它还集成了推理软件开发包（SDK），比如我们广受欢迎的高通神经网络处理 SDK，包括面向 Android、Linux 和 Windows 的不同版本。高通开发者库和服务支持最新编程语言、虚拟平台和编译器。在更底层，我们的系统软件集成了基础的实时操作系统（RTOS）、系统接口和驱动程序。我们还支持广泛的操作系统（包括 Android、Windows、Linux 和 QNX），以及用于部署和监控的基础设施（比如 Prometheus、Kubernetes 和 Docker）。

高通 AI 软件栈还集成了 Qualcomm® AI Studio，支持从模型设计到优化、部署和分析的完整工作流。它将高通提供的全部工具集成到一个图形用户界面，并利用可视化工具以简化开发者体验，支持开发者实时查看模型开发进度，这其中包括高通 AI 模型增效工具包（AIMET）、AI 模型增效工具包模型库、模型分析器和神经网络架构搜索（NAS）。[3]

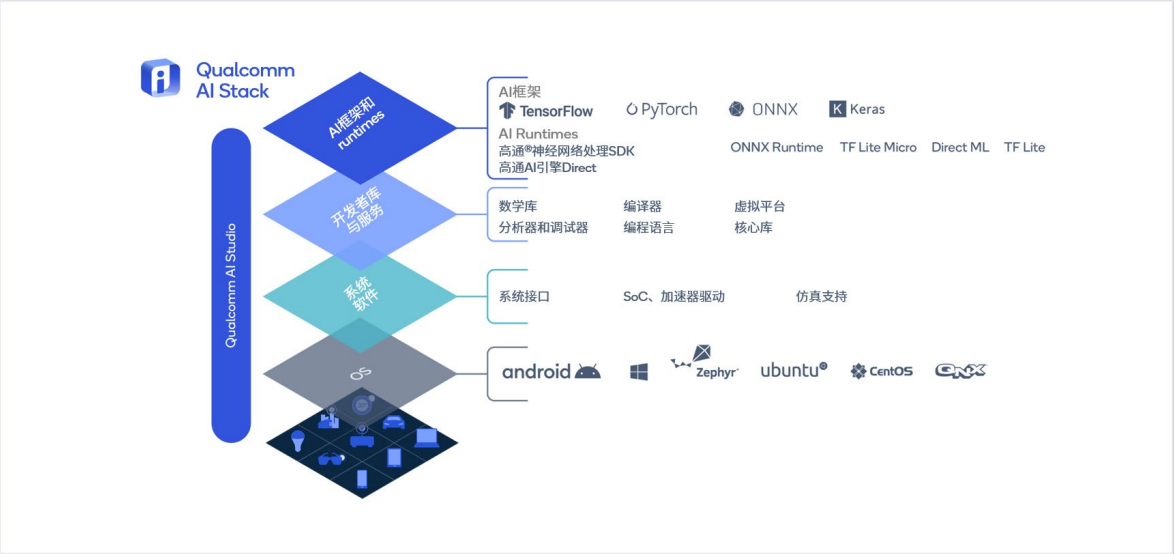

图 3：高通 AI 软件栈旨在帮助开发者一次编写、随处运行，实现规模化部署。

高通专注于 AI 模型效率研究以提高能效和性能。快速的小型 AI 模型如果只能提供低质量或不准确的结果，那么将失去实际用处。因此，我们采用全面而有针对性的策略，包括量化、压缩、条件计算、神经网络架构搜索（NAS）和编译，在不牺牲太多精度的前提下缩减 AI 模型，使其高效运行。即使是那些已经面向移动终端优化过的模型我们也会进行这一工作。

[3] 高通 AI 模型增效工具包（AIMET）和 AI 模型增效工具包模型库是高通创新中心公司的产品。

图 4：高通 AI 研究采用整体 AI 模型效率研究方法。

4.2.1 量化

面向高效整数推理的量化是我们的重点关注领域之一。过去几年，我们通过论文和演示分享了高通领先的 AI 量化研究，包括训练后量化（PTQ）技术，比如无数据量化和自适应舍入(AdaRound)，以及联合量化和剪枝技术，比如贝叶斯比特。量化不仅能够提高性能，降低内存要求，还能通过让模型在高通专用 AI 硬件上高效运行，降低内存带宽占用，以节省功耗。例如，将 FP32 模型量化压缩到 INT4 模型，可带来高达 64 倍的内存和计算能效提升。

对于生成式 AI 来说，由于基于 transformer 的大语言模型（比如 GPT、Bloom 和 LLaMA）受到内存的限制，在量化到 8 位或 4 位权重后往往能够获得大幅提升的效率优势。包括高通在内的多项研究工作显示，4 位权重量化不仅对大语言模型可行，在 PTQ 设置中同样可行，并能实现最优表现。这一效率的跃升已经超越了浮点模型。

高通 AI 模型增效工具包提供基于高通 AI 研究技术成果开发的量化工具，目前已纳入 Qualcomm AI Studio。借助量化感知训练和/或更加深入的量化研究，许多生成式 AI 模型可以量化至 INT4 模型。INT4 支持将在不影响准确性或性能表现的情况下节省更多功耗，与 INT8 相比实现高达 90%的性能提升和 60%的能效提升，能够运行更高效的神经网络。使用低位数整型精度对高能效推理至关重要。

4.2.2 编译

编译器作为高通 AI 软件栈中的关键组件，让 AI 模型能够以最高性能和最低功耗高效运行。AI 编译器将输入的神经网络转化为可以在目标硬件上运行的代码，同时针对时延、性能和功耗进行优

化。编译包括计算图的切分、映射、排序和调度等步骤。高通在传统编译器技术、多面体 AI 编译器和编译器组合优化 AI 研究方面的技术专长已经实现了诸多先进的技术成果。

例如，高通 AI 引擎 Direct 框架基于高通 Hexagon™处理器的硬件架构和内存层级进行运算排序，以提高性能并最大程度减少内存溢出。我们的优化有助于减少 DRAM 存取量，并显著降低 runtime 的时延和功耗。

4.3 硬件加速

高通硬件提供行业领先的能效，是移动领域竞品的近 2 倍。

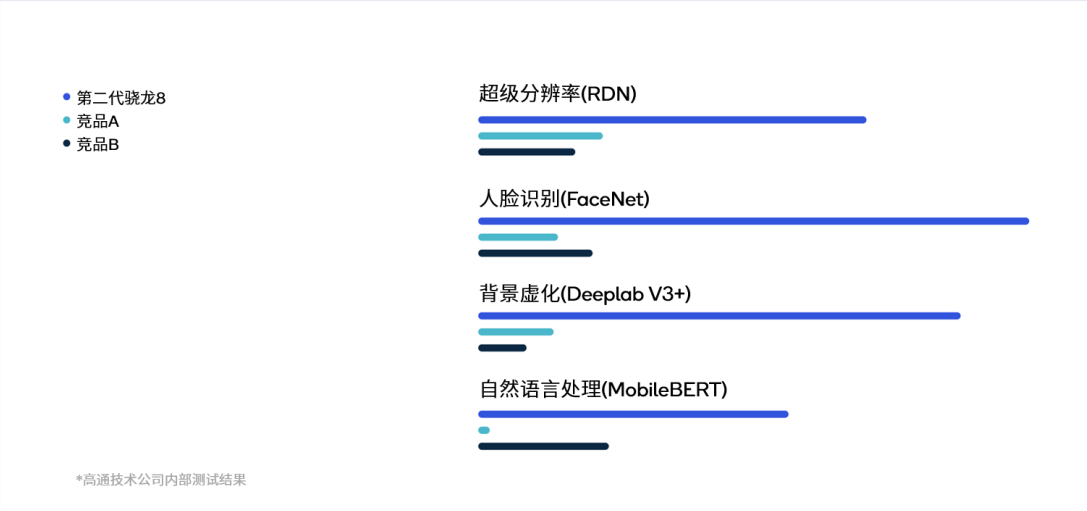

图5：与移动领域竞品相比，第二代骁龙 8 提供领先的 AI 能效。

高通 AI 引擎由多个软硬件组件构成，能在骁龙和高通平台上实现终端侧 AI 加速。在硬件方面，高通 AI 引擎采用异构计算架构，包括 Hexagon 处理器、高通 Adreno™ GPU 和高通 Kryo™ CPU，全部面向在终端侧快速高效地运行 AI 应用而打造。通过异构计算的方式，开发者和 OEM 厂商可以优化智能手机和其他边缘侧终端上的 AI 用户体验。

基于多年的专项研究投入，Hexagon 处理器不断演进，已经成为了高通 AI 引擎最关键的部分，并能够应对不断变化的 AI 需求。2007 年，我们在骁龙平台上推出了首个 Hexagon 处理器。2015 年，骁龙 820 处理器推出，集成了首个专门面向移动平台的高通 AI 引擎，以支持图像、音频和传感器的运算。2018 年，我们在骁龙 855 中为 Hexagon 处理器增加了张量加速器。2019 年，我们在骁龙 865 上扩展了终端侧 AI 用例，包含 AI 图片、AI 视频、AI 语音和始终在线的传感器中枢。

2022 年，第二代骁龙 8 为整个系统提供了开创性的 AI 技术，搭载了迄今为止最快、最先进的高通 AI 引擎。用户可以体验更快速的自然语言处理所带来的多语种翻译，或享受由 AI 赋能的电影模式视频拍摄所带来的乐趣。最新的 Hexagon 处理器采用专用供电系统，能够按照工作负载适配功率。特殊硬件提升了分组卷积、激活函数加速和 Hexagon 张量加速器的性能。支持微切片推理和 INT4 硬件加速能够在提供更高性能的同时，降低能耗和内存占用。Transformer 加速大幅提升了生成式 AI 中充分使用的多头注意力机制的推理速度，在使用 MobileBERT 的特定用例中能带来高达 4.35 倍的惊人 AI 性能提升。

5 无与伦比的全球边缘侧布局和规模

高通技术公司部署的边缘侧终端规模十分庞大，搭载骁龙和高通平台的已上市用户终端数量已达到数十亿台，而且每年有数亿台的新终端还在进入市场。[4]

我们的 AI 能力赋能一系列广泛的产品，包括手机、汽车、XR、PC 和物联网。我们开发 AI 加速解决方案（比如高通 AI 引擎）以及所有面向顶级产品的其他关键 IP 创新和技术，通常每年作为高通可扩展技术架构的一部分进行迭代，跨细分领域快速普及相关功能并下沉到主流和入门级产品。

正因如此，高通技术公司对在全球范围赋能混合 AI 规模化扩展独具优势。

图 6：搭载骁龙平台的终端能够推动混合 AI 扩展至跨不同细分领域和层级的数十亿产品。

[4] Counterpoint Research，2023 年 5 月

5.1 手机

骁龙是提升顶级 Android 体验的领先移动平台，其中就包含已出货的 20 多亿个具备 AI 能力的处理器。骁龙平台在移动平台 AI 基准测试中也处于领先地位，比如在行业知名的 AI Benchmark 中占据前 20 位。[5]

2023 年第二季度，领先的市场调研公司 TechInsights 预测，高通技术公司将以超过 40%的市场份额保持 AI 智能手机处理器出货量的领导地位，远远超过苹果（25%）和联发科（24%）等其他公司。[6]

5.2 汽车

高通技术公司是座舱和车载信息娱乐解决方案的领导者，全球所有主要汽车制造商都选择骁龙座舱平台来赋能他们的数字座舱系统。其中许多汽车制造商已经启动量产项目，或目前正在设计采用高通解决方案的平台。这些汽车制造商包括本田、梅赛德斯、雷诺、沃尔沃、捷豹路虎、Stellantis、宝马、通用汽车/凯迪拉克、长城汽车、Mahindra、Togg、丰田、小鹏汽车、广汽集团、捷途汽车、蔚来和威马汽车。

随着最新一代骁龙座舱平台的推出，高通汽车解决方案旨在提供业界领先的车内用户体验，以及安全性、舒适性和可靠性，在网联汽车时代为数字座舱解决方案树立全新标杆。

Snapdragon Ride™平台能够提供扩展的产品路线图，包括基于 5 纳米工艺制程打造的首款可扩展自动驾驶 SoC 平台，拥有更广泛的软件生态系统，提供经行业验证的视觉感知、泊车和驾驶员监测软件栈。

5.3 PC 和平板电脑

骁龙计算平台集成高通 AI 引擎，支持强大的终端侧加速，能够为最新应用带来更佳质量、性能和效率。除文本、图像和视频创作等生成式 AI 应用外，高通 AI 引擎还支持一系列传统 AI 用例，从提升安全性的快速威胁检测，到增强视频会议体验的眼神接触和降噪。利用 Hexagon 处理器能够提升性能和效率，实现长时间电池续航，同时不占用 CPU 和 GPU 等其他系统资源，能够帮助用户提高生产力。

---

[5] 基于 ai-benchmark.com 分数，截至 2023 年 5 月

[6] TechInsights，2023 年 4 月

5.4 物联网

高通技术公司是物联网领域的主要技术提供商，拥有跨不同垂直领域超过 16,000 家的客户。嵌入高通物联网芯片组和平台的 AI 处理能力支持以高效可行的方式进行终端侧数据分析（比如视频），推动跨多个细分领域的创新和转型，包括机器人、智能摄像头、零售和城市基础设施。

5.5 XR

VR 头显和 AR 眼镜等 XR 终端也集成了高通终端侧 AI 和 Snapdragon Spaces™技术，以提供更具沉浸感的体验，更好地适应周围世界。

迄今为止，已有超过 65 款采用骁龙平台的 XR 终端发布，包括 Meta、PICO 和联想等品牌推出的众多广受欢迎的终端。

6 总结

混合 AI 势不可当。云端和终端将协同工作，依托强大、高效且高度优化的 AI 能力打造下一代用户体验。终端侧 AI 领导力赋予高通面向混合架构转型的独特优势。随着大量的工作负载正从云端转向边缘终端，因此需要边缘侧处理的高性能和出色能效。凭借具备前瞻性的早期研究和产品开发投入，目前骁龙平台能够支持参数超过 10 亿的生成式 AI 模型，并即将支持 100 亿或更多参数的模型。

高通拥有无与伦比的边缘侧布局，全球搭载骁龙和高通平台的终端装机量已达到数十亿台，有望推动生成式 AI 规模化扩展，为无数人的生活带来积极影响。高通技术公司将支持开发者、OEM 厂商和其他生态系统创新者快速且经济高效地构建全新生成式 AI 应用和解决方案。技术领导力、全球规模和生态系统赋能完美结合，让高通技术公司在推动混合 AI 开发和应用方面独树一帜。

欲了解更多相关内容

欢迎订阅《未来移动计算技术》简讯

Qualcomm

请关注我们：

欲了解更多信息，请访问

qualcomm.com