Qualcomm 高通

让AI触手可及

高通AI白皮书

Qualcomm AI White Paper

高通AI白皮书

Qualcomm AI White Paper

序言
PREAMBLE

携手合作 拥抱AI终端创新的黄金时代

高通公司中国区董事长　孟樸

一年前，高通公司发布了《混合AI是AI的未来》白皮书，率先向业界分享了对人工智能（AI）技术发展趋势的洞察。那时，ChatGPT等生成式AI初露锋芒，这一现象级的应用引发了产业界对这场AI技术革命的广泛探讨和巨大期待。人们开始意识到，生成式AI将为各行各业生产力的提升带来质变。从那时起，大模型技术日新月异，商业化应用的步伐不断加快。当每个人都希望无时无刻地拥有“个人大模型”时，生成式AI走向终端，成为了一个不可逆转的趋势。智能终端的新应用、新形态、新场景，正在为AI技术的普及提供广阔的空间，AI终端创新的黄金时代已经到来。

• 从云到端：智能终端迎来新增长周期，让AI真正触手可及

当生成式AI展现出强大的能力和前景，我们也认识到，AI技术的真正价值在于其普惠性 —— 要实现AI人人可享、人人可用，需要让AI技术更加贴近用户，在人们触手可及的终端上运行。

由此，AI的计算重心正在从云端向终端迁移。这是由市场需求、技术趋势和用户体验共同驱动的结果。从主机到智能手机、个人电脑（PC）等终端，计算能力的下沉使得这些终端也能够进行AI加速计算。这种分布式计算平台的运行，不仅提高了计算效率，也加速了AI在终端侧的演进。与此同时，AI能够本地运行，并根据用户需求与云端交互，人机交互将变得更自然、更即时、更加个性化，隐私性也更有保障。在这个过程中，5G作为关键的连接“底座”，为AI在云端、边缘云和终端侧协同奠定了坚实的基础。预计到2025年底，全球5G连接规模将达到25亿[1]。这正是“5G+AI”协同发展所带来的令人兴奋的变革 —— 它改变了用户体验的定义，丰富了千行百业的智能连接用例，也推动了新一轮终端创新的浪潮。

在高通看来，这也正是生成式AI的革新意义 —— 智能终端让AI成为无处不在的个人助理，推动终端与云端的融合，为智能手机带来新的互动方式，让汽车成为全新的运算空间，为下一代PC带来强大的AI能力，智能终端市场迎来了新的增长动力。

[1] GSMI、GTI、中移智库：5G新技术创造新价值

智能手机、PC、智能网联汽车位于AI终端创新的最前沿。其中，智能手机市场规模庞大，年出货量高达十几亿台。目前，众多手机厂商积极推广生成式AI应用，使得智能手机有望成为生成式AI发展最快的领域之一。据预测[2]，生成式AI智能手机出货量将在2023到2027年迅速增长，预计2024年出货量占比达到11%，到2027年将达到5.5亿部，占比43%，年均复合增长率为49%。

• 从“百模”到“百端”：让高性能的AI处理成为可能，赋能终端侧AI规模化扩展

AI应用场景不断拓展，各类算法模型日趋多样化和复杂，对底层算力的需求也与日俱增。如何将“大模型”高效装载到“小设备”，满足多样化的生成式AI用例？——这有赖于终端算力的革新升级。

你的智能手机将成为个人AI助理的载体，帮你完成信息查找、场景识别、图像处理等各种任务。然而，这些任务对计算资源和处理能力的要求不尽相同。这就需要从以通用计算为核心的计算架构，向更加高性能的异构AI计算架构升级，让CPU、GPU和NPU等不同的计算单元“各司其职”。只有协同使用这些计算单元，异构计算才能在应用性能、能效和电池续航上实现最优化，让AI助理如虎添翼，赋能增强的生成式AI体验。

作为AI前沿科技的开拓者和探索者，我们看到，终端侧AI规模化扩展正在点燃产业界的热情和信心，推动智能终端软硬件和生态层面的创新。我们也倍感自豪，高通能够成为推动这一进程的重要力量。今年3月，我们发布了《通过NPU和异构计算开启终端侧生成式AI》白皮书，分享了高通在异构计算架构和NPU研究方面的创新成果。事实上，早在2007年，也就是生成式AI进入大众视野的15年前，高通就开始了对NPU的研究。多年来，高通致力于将高性能低功耗的AI计算能力带入终端设备，打造了专为AI定制设计的全新计算架构。通过异构计算AI引擎，我们将性能卓越的CPU、NPU和GPU进行组合，为行业提供了可行的解决方案，支持生态系统在跨多品类终端上开发并实现生成式AI用例、体验和领先产品，让智能计算无处不在。

• 从共享机遇到共建生态：共创AI终端创新的黄金时代

终端侧AI规模化扩展的发展浪潮，为大模型服务商、终端厂商、算力提供商、应用开发者等产业链各方，带来了前所未有的发展机遇。据预测[3]，对端侧AI能力的需求可能会引发新一轮的换机热潮，并有助于提高设备的平均销售价格（ASP），AI能力将成为手机厂商推进高端化的有效发力点。小米、荣耀、OPPO、三星等品牌均已推出支持丰富生成式AI应用的旗舰机型。在PC领域，预计到2027年[4]，超过60%出货的PC将是AI PC。

---

[2] Counterpoint：生成式AI智能手机出货量将大涨，2027年占比达43%
[3] Canalys：洞悉中国手机市场的AI趋势与潜力
[4] Canalys：Canalys报告摘要：AI PC的现在和未来

面对AI终端产业机遇，我们始终相信，要实现让智能计算无处不在、AI触手可及，需要产业链上下游的通力合作，需要包括中国在内的全球生态系统的创新与协作。这将加速AI技术在各领域的普及与应用，为形成新质生产力蓄势赋能。高通的AI领先优势得益于与业界的深度合作。无论是高通的异构计算能力，还是可扩展的AI软件工具等，都需要与客户的终端深度结合才能实现。我们也很高兴地看到，高通的AI解决方案和骁龙平台正在成为推动终端侧AI体验的关键引擎——手机厂商基于第三代骁龙8移动平台，为消费者打造突破性的AI体验；PC厂商通过骁龙X系列平台产品组合，为企业用户和消费者带来强大生产力、丰富创造力和沉浸式娱乐体验；汽车厂商也基于骁龙数字底盘，将智能网联汽车上的生成式AI应用与云端AI相结合，为用户创造更好的驾乘体验。目前，高通AI引擎赋能的终端产品出货量已经超过了20亿。

与此同时，为了与生态伙伴共建开放生态，高通推出了AI Hub，让开发者充分发挥前沿技术的潜力，共同推进终端侧AI的规模化商用进程。我们希望能够打造一个横向生态系统，让所有模型在终端上可以和谐共生，带来跨多个生态系统的全新AI体验。

在终端侧AI规模化扩展的机遇面前，我们倍感振奋，将一如既往地通过技术创新与合作共赢，担当推动终端侧AI发展的重要力量。期望各界能够从我们最新结集发布的《让AI触手可及——高通AI白皮书》中，更加系统性地了解高通在AI技术演进和应用落地方面的见解和洞察。这不仅是高通在AI领域持续探索、不断突破的有力见证，也凝聚了高通与行业伙伴共同智慧的结晶。

让我们携手共同迈向激动人心的AI新时代，一同探索AI终端创新的无限可能，见证AI科技变革千行百业、成就人类美好生活的壮阔进程。

2024年世界移动通信大会（MWC）期间，高通凭借领先的AI技术创新，荣获全球移动大奖（GLOMO奖）的“最佳人工智能创新奖”[5]，专为生成式AI而生的移动平台第三代骁龙8荣获“设备创新突破奖”[6]，赋能智能手机体验的全面突破，让智能计算无处不在。

全球移动大奖（GLOMO奖）是全球数字智能领域的最高奖项，表彰推动移动行业进步的巨擘级创新[7]。

[5] 奖项名称 Best AI Innovation，请以英文为准
[6] 奖项名称 Breakthrough device innovation，请以英文为准
[7] 奖项信息源自官方介绍，https://www.mwcbarcelona.com/mobile-awards

Table of contents

第一部分 PART ONE

通过NPU和异构计算开启终端侧生成式AI

Table of contents

第二部分 PART TWO

终端侧AI和混合AI开启生成式AI的未来

第三部分 PART THREE

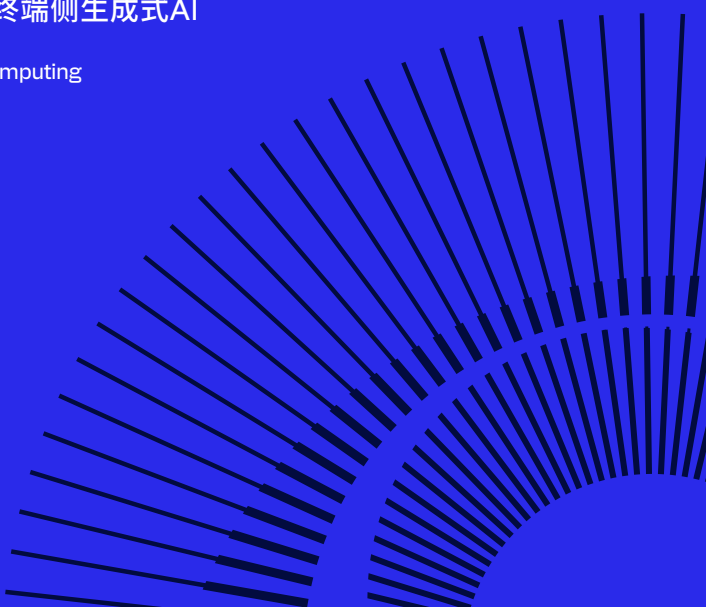
高通在推动混合AI规模化扩展方面独具优势

生成式AI时代需要何种算力？

高通AI白皮书 第一部分

通过NPU和异构计算开启终端侧生成式AI

Unlocking on-device generative AI
with an NPU and heterogeneous computing



第一部分 PART ONE

通过NPU和异构计算开启终端侧生成式AI

Unlocking on-device generative AI
with an NPU and heterogeneous computing

• 1. 摘要

生成式AI变革已经到来。 随着生成式AI用例需求在有着多样化要求和计算需求的垂直领域不断增加，我们显然需要专为AI定制设计的全新计算架构。这首先需要一个面向生成式AI全新设计的神经网络处理器（NPU），同时要利用异构处理器组合，比如中央处理器（CPU）和图形处理器（GPU）。通过结合NPU使用合适的处理器，异构计算能够实现最佳应用性能、能效和电池续航，赋能全新增强的生成式AI体验。

NPU专为实现低功耗加速AI推理而全新打造，并随着新AI用例、模型和需求的发展不断演进。优秀的NPU设计能够提供正确的设计选择，与AI行业方向保持高度一致。

高通正在助力让智能计算无处不在。业界领先的高通 Hexagon™ NPU面向以低功耗实现持续稳定的高性能AI推理而设计。高通NPU的差异化优势在于系统级解决方案、定制设计和快速创新。通过定制设计NPU以及控制指令集架构（ISA），高通能够快速进行设计演进和扩展，以解决瓶颈问题并优化性能。Hexagon NPU是高通业界领先的异构计算架构——高通AI引擎中的关键处理器，高通AI引擎还包括高通 Adreno™ GPU、高通 Kryo™ 或高通 Oryon™ CPU、高通传感器中枢和内存子系统。这些处理器为实现协同工作而设计，能够在终端侧快速且高效地运行AI应用。我们在AI基准测试和实际生成式AI应用方面的行业领先性能就是例证。

我们还专注于在全球搭载高通和骁龙®平台的数十亿终端设备上实现便捷开发和部署，赋能开发者。利用高通AI软件栈（Qualcomm AI Stack），开发者可在高通硬件上创建、优化和部署AI应用，一次编写即可实现在不同产品和细分领域采用高通芯片组解决方案进行部署。高通技术公司正在赋能终端侧生成式AI的规模化扩展。

Qualcomm

• 2. 处理器集成于SoC中的诸多优势

在不断增长的用户需求、全新应用和终端品类以及技术进步的驱动下，计算架构正在不断演进。最初，中央处理器（CPU）就能够完成大部分处理，但随着计算需求增长，对全新处理器和加速器的需求出现。例如，早期智能手机系统由CPU和环绕CPU分布的分立芯片组成，用于2D图形、音频、图像信号处理、蜂窝调制解调器和GPS等处理。随着时间推移，这些芯片的功能已经集成到称为系统级芯片（SoC）的单个芯片体（DIE）中。

例如，现代智能手机、PC和汽车SoC已集成多种处理器，如中央处理器（CPU）、图形处理器（GPU）和神经网络处理器（NPU）。芯片设计上的这种集成具有诸多优势，包括改善峰值性能、能效、单位面积性能、芯片尺寸和成本。

例如，在智能手机或笔记本电脑内安装分立的GPU或NPU会占用更多电路板空间，需要使用更多能源，从而影响工业设计和电池尺寸。此外，输入/输出引脚间的数据传输也将增多，将导致性能降低、能耗增加，以及采用更大电路板带来的额外成本和更低的共享内存效率。对于智能手机、笔记本电脑和其他需要轻巧工业设计，具有严格功率和散热限制的便携式终端，集成更为必要。

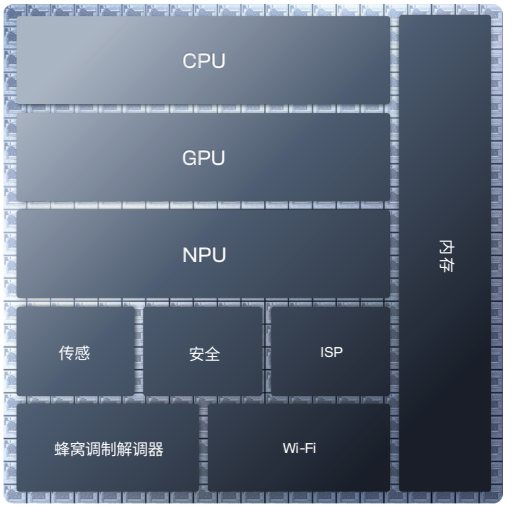


图1：现代SoC在单个DIE中集成多个处理器以改善峰值性能、能效、单位面积性能、工业设计和成本。

• 3. 生成式AI需要多样化的处理器

谈到AI，集成专用处理器并不新鲜。智能手机SoC自多年前就开始利用NPU改善日常用户体验，赋能出色影像和音频，以及增强的连接和安全。不同之处在于，生成式AI用例需求在有着多样化要求和计算需求的垂直领域不断增加。这些用例可分为三类：

1. 按需用例由用户触发，需要立即响应，包括照片/视频拍摄、图像生成/编辑、代码生成、录音转录/摘要和文本（电子邮件、文档等）创作/摘要。这包括用户用手机输入文字创作自定义图像、在PC上生成会议摘要，或在开车时用语音查询最近的加油站。

2. 持续型用例运行时间较长，包括语音识别、游戏和视频的超级分辨率、视频通话的音频/视频处理以及实时翻译。 这包括用户在海外出差时使用手机作为实时对话翻译器，以及在PC上玩游戏时逐帧运行超级分辨率。

3. 泛在用例在后台持续运行，包括始终开启的预测性AI助手、基于情境感知的AI个性化和高级文本自动填充。例如手机可以根据用户的对话内容自动建议与同事的会议、PC端的学习辅导助手则能够根据用户的答题情况实时调整学习资料。

这些AI用例面临两大共同的关键挑战。第一，在功耗和散热受限的终端上使用通用CPU和GPU服务平台的不同需求，难以满足这些AI用例严苛且多样化的计算需求。第二，这些AI用例在不断演进，在功能完全固定的硬件上部署这些用例不切实际。因此，支持处理多样性的异构计算架构能够发挥每个处理器的优势，例如以AI为中心定制设计的NPU，以及CPU和GPU。每个处理器擅长不同的任务：CPU擅长顺序控制和即时性，GPU适合并行数据流处理，NPU擅长标量、向量和张量数学运算，可用于核心AI工作负载。

CPU和GPU是通用处理器。它们为灵活性而设计，非常易于编程，“本职工作”是负责运行操作系统、游戏和其他应用等。而这些“本职工作”同时也会随时限制他们运行AI工作负载的可用容量。NPU专为AI打造，AI就是它的“本职工作”。NPU降低部分易编程性以实现更高的峰值性能、能效和面积效率，从而运行机器学习所需的大量乘法、加法和其他运算。

通过使用合适的处理器，异构计算能够实现最佳应用性能、能效和电池续航，赋能全新增强的生成式AI体验。

Qualcomm 高通

你的设备
可以像你一样思考吗？

#让AI触手可及

扫码下载
高通AI白皮书电子版

• 4. NPU入门

NPU专为实现以低功耗加速AI推理而全新打造，并随着新AI用例、模型和需求的发展不断演进。对整体SoC系统设计、内存访问模式和其他处理器架构运行AI工作负载时的瓶颈进行的分析会深刻影响NPU设计。这些AI工作负载主要包括由标量、向量和张量数学组成的神经网络层计算，以及随后的非线性激活函数。

在2015年，早期的NPU面向音频和语音AI用例而设计，这些用例基于简单卷积神经网络（CNN）并且主要需要标量和向量数学运算。从2016年开始，拍照和视频AI用例大受欢迎，出现了基于Transformer、循环神经网络（RNN）、长短期记忆网络（LSTM）和更高维度的卷积神经网络（CNN）等更复杂的全新模型。这些工作负载需要大量张量数学运算，因此NPU增加了张量加速器和卷积加速，让处理效率大幅提升。有了面向张量乘法的大共享内存配置和专用硬件，不仅能够显著提高性能，而且可以降低内存带宽占用和能耗。例如，一个NxN矩阵和另一个NxN矩阵相乘，需要读取$2N^2$个值并进行$2N^3$次运算（单个乘法和加法）。在张量加速器中，每次内存访问的计算操作比率为N:1，而对于标量和向量加速器，这一比率要小得多。

在2023年，大语言模型（LLM）—— 比如Llama 2-7B，和大视觉模型（LVM）—— 比如Stable Diffusion赋能的生成式AI使得典型模型的大小提升超过了一个数量级。除计算需求之外，还需要重点考虑内存和系统设计，通过减少内存数据传输以提高性能和能效。未来预计将会出现对更大规模模型和多模态模型的需求。

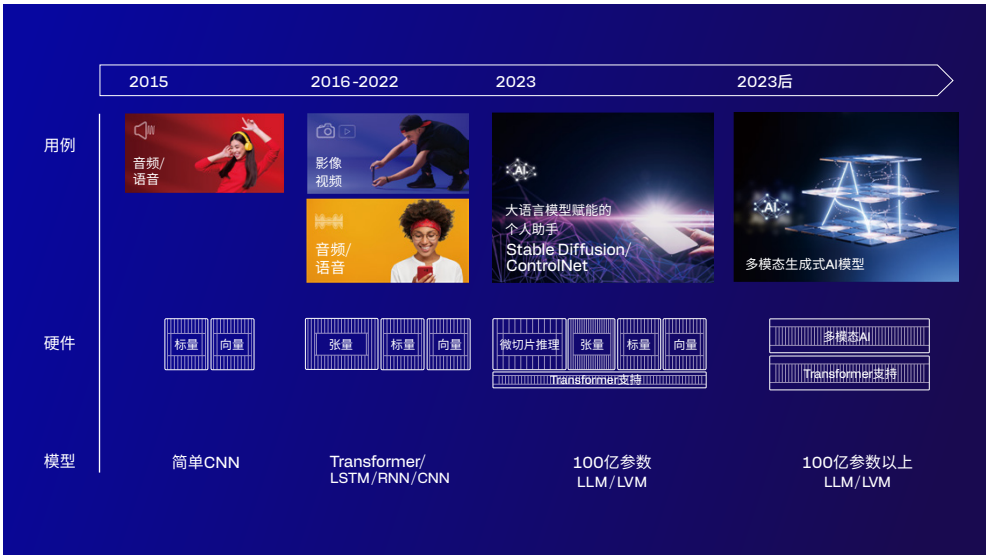


图2：NPU随着不断变化的AI用例和模型持续演进，实现高性能低功耗。

随着AI持续快速演进，必须在性能、功耗、效率、可编程性和面积之间进行权衡取舍。一个专用的定制化设计NPU能够做出正确的选择，与AI行业方向保持高度一致。

• 5. 高通NPU：以低功耗实现持久稳定的高性能AI

经过多年研发，高通 Hexagon NPU不断演进，能够满足快速变化的AI需求。2007年，首款Hexagon DSP在骁龙® 平台上正式亮相——DSP控制和标量架构是高通未来多代NPU的基础。

2015年，骁龙820处理器正式推出，集成首个高通AI引擎，支持成像、音频和传感器运算。2018年，高通在骁龙 855中为Hexagon NPU增加了Hexagon张量加速器。2019年，高通在骁龙865上扩展了终端侧AI用例，包括AI成像、AI视频、AI语音和始终在线的感知功能。

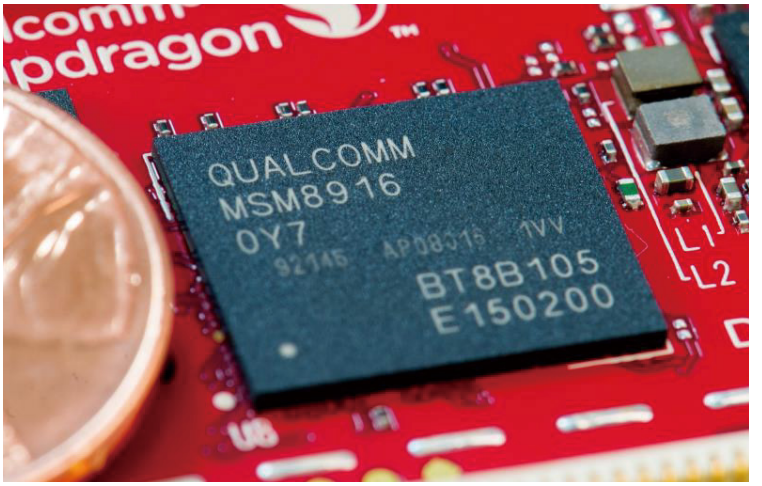


图3：2015年发布的骁龙820首次集成高通AI引擎。

2020年，高通凭借Hexagon NPU变革性的架构更新，实现了重要里程碑。我们融合标量、向量和张量加速器，带来了更佳性能和能效，同时还为加速器打造了专用大共享内存，让共享和迁移数据更加高效。融合AI加速器架构为高通未来的NPU架构奠定了坚实基础。

2022年，第二代骁龙8中的Hexagon NPU引入了众多重要技术提升。专用电源传输轨道能够根据工作负载动态适配电源供应。微切片推理利用Hexagon NPU的标量加速能力，

将神经网络分割成多个能够独立执行的微切片，消除了高达10余层的内存占用，能够最大化利用Hexagon NPU中的标量、向量和张量加速器并降低功耗。本地4位整数（INT4）运算支持能够提升能效和内存带宽效率，同时将INT4层和神经网络的张量加速吞吐量提高一倍。Transformer网络加速大幅加快了应用于生成式AI的多头注意力机制的推理速度，在使用MobileBERT模型的特定用例中能带来高达4.35倍的惊人AI性能提升。其他特殊硬件包括改进的分组卷积、激活函数加速和张量加速器性能。

第三代骁龙8中的Hexagon NPU是高通面向生成式AI最新、也是目前最好的设计，为持续AI推理带来98%性能提升和40%能效提升[1]。它包括了跨整个NPU的微架构升级。微切片推理进一步升级，以支持更高效的生成式AI处理，并降低内存带宽占用。此外，Hexagon张量加速器增加了独立的电源传输轨道，让需要不同标量、向量和张量处理规模的AI模型能够实现最高性能和效率。大共享内存的带宽也增加了一倍。基于以上提升和INT4硬件加速，Hexagon NPU成为面向终端侧生成式AI大模型推理的领先处理器。

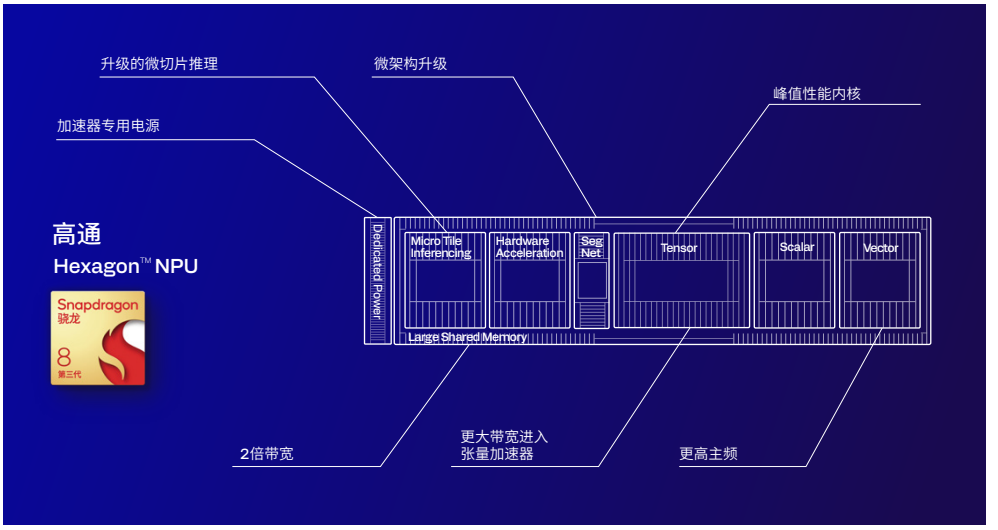


图4：第三代骁龙8的Hexagon NPU升级以低功耗实现领先的生成式AI性能。

[1] 与前代平台相比。

高通NPU的差异化优势在于系统级解决方案、定制设计和快速创新。高通的系统级解决方案考量每个处理器的架构、SoC系统架构和软件基础设施，以打造最佳AI解决方案。要在增加或修改硬件方面做出恰当的权衡和决策，需要发现当前和潜在的瓶颈。通过跨应用、神经网络模型、算法、软件和硬件的全栈AI研究与优化，高通能够做到这一点。由于能够定制设计NPU并控制指令集架构(ISA)，高通架构师能够快速进行设计演进和扩展以解决瓶颈问题。

这一迭代改进和反馈循环，使我们能够基于最新神经网络架构持续快速增强高通NPU和高通AI软件栈。基于高通的自主AI研究以及与广大AI社区的合作，我们与AI模型的发展保持同步。高通具有开展基础性AI研究以支持全栈终端侧AI开发的独特能力，可赋能产品快速上市，并围绕终端侧生成式AI等关键应用优化NPU部署。

相应地，高通NPU历经多代演进，利用大量技术成果消除瓶颈。例如，第三代骁龙8的诸多NPU架构升级能够帮助加速生成式AI大模型。内存带宽是大语言模型token生成的瓶颈，这意味着其性能表现更受限于内存带宽而非处理能力。因此，我们专注于提高内存带宽效率。第三代骁龙8还支持业界最快的内存配置之一：4.8GHz LPDDR5x，支持77GB/s带宽，能够满足生成式AI用例日益增长的内存需求。

从DSP架构入手打造NPU是正确的选择，可以改善可编程性，并能够紧密控制用于AI处理的标量、向量和张量运算。高通优化标量、向量和张量加速的设计方案结合本地共享大内存、专用供电系统和其他硬件加速，让我们的解决方案独树一帜。高通NPU能够模仿最主流模型的神经网络层和运算，比如卷积、全连接层、Transformer以及主流激活函数，以低功耗实现持续稳定的高性能表现。

Qualcomm

• 6. 异构计算：利用全部处理器支持生成式AI

适合终端侧执行的生成式AI模型日益复杂，参数规模也在不断提升，从10亿参数到100亿，甚至700亿参数。其多模态趋势日益增强，这意味着模型能够接受多种输入形式——比如文本、语音或图像，并生成多种输出结果。

此外，许多用例需要同时运行多个模型。例如，个人助手应用采用语音输入输出，这需要运行一个支持语音生成文本的自动语音识别（ASR）模型、一个支持文本生成文本的大语言模型、和一个作为语音输出的文本生成语音（TTS）模型。生成式AI工作负载的复杂性、并发性和多样性需要利用SoC中所有处理器的能力。最佳的解决方案要求：

1. 跨处理器和处理器内核扩展生成式AI处理

2. 将生成式AI模型和用例映射至一个或多个处理器及内核

选择合适的处理器取决于众多因素，包括用例、终端类型、终端层级、开发时间、关键性能指标（KPI）和开发者的技术专长。制定决策需要在众多因素之间进行权衡，针对不同用例的KPI目标可能是功耗、性能、时延或可获取性。例如，原始设备制造商（OEM）在面向跨品类和层级的多种终端开发应用时，需要根据SoC规格、最终产品功能、开发难易度、成本和应用跨终端层级的适度降级等因素，选择运行AI模型的最佳处理器。

正如前述，大多数生成式AI用例可分类为按需型、持续型或泛在型用例。按需型应用的关键性能指标是时延，因为用户不想等待。这些应用使用小模型时，CPU通常是正确的选择。当模型变大（比如数十亿参数）时，GPU和NPU往往更合适。电池续航和能效对于持续和泛在型用例至关重要，因此NPU是最佳选择。

另一个关键区别在于AI模型为内存限制型（即性能表现受限于内存带宽），还是计算限制型（即性能表现受限于处理器性能）。当前的大语言模型在生成文本时受内存限制，

因此需要关注CPU、GPU或NPU的内存效率。对于可能受计算或内存限制的大视觉模型，可使用GPU或NPU，但NPU可提供最佳的能效。

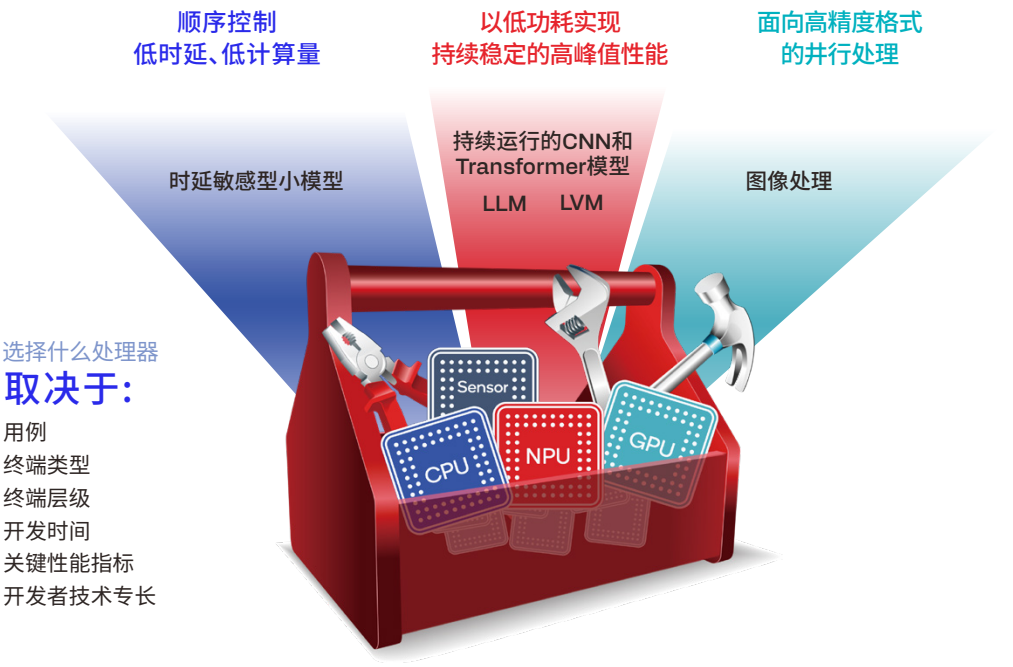


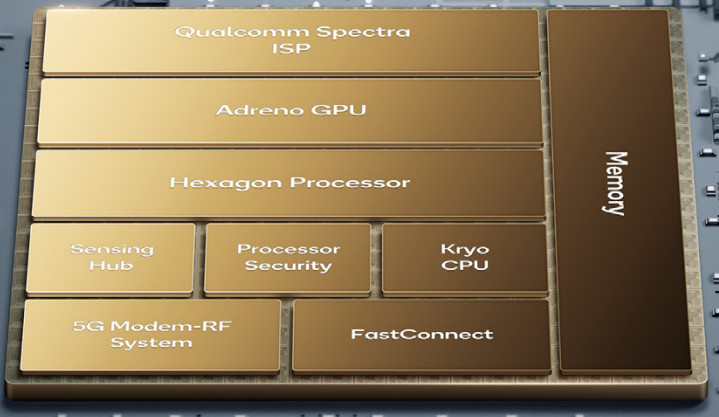
图5：正如在工具箱中选择合适的工具一样，选择合适的处理器取决于诸多因素。

提供自然语音用户界面（UI）以提高生产力并增强用户体验的个人助手预计将成为一类流行的生成式AI应用。语音识别、大语言模型和语音模型必将以某种并行方式运行，因此理想的情况是在NPU、GPU、CPU和传感处理器之间分布处理模型。对于PC来说，个人助手预计将始终开启且无处不在地运行，考虑到性能和能效，应当尽可能在NPU上运行。

Qualcomm 高通

如何让处理器各显神通获得更高效率？

#让AI触手可及



扫码下载
高通AI白皮书电子版

• 7. 高通AI引擎：面向生成式AI的业界领先异构计算

高通AI引擎包含多个硬件和软件组件，以加速骁龙和高通平台上的终端侧AI。在集成硬件方面，高通AI引擎具有业界最领先的异构计算架构，包括Hexagon NPU、Adreno GPU、高通Kryo或高通Oryon CPU、高通传感器中枢和内存子系统，所有硬件都经过精心设计以实现协同工作，在终端侧快速高效地运行AI应用。

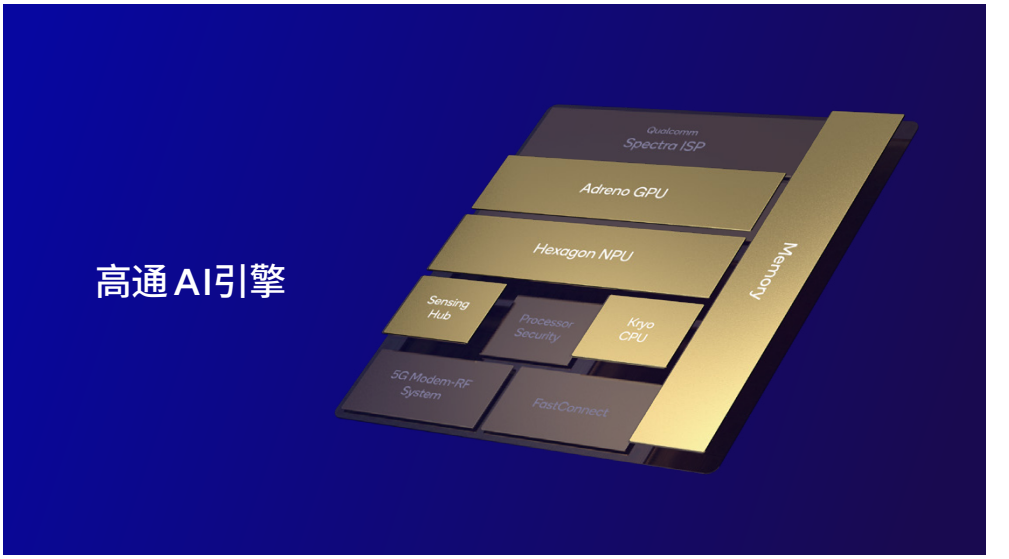


图6：高通AI引擎包括Hexagon NPU、Adreno GPU、高通Kryo或高通Oryon CPU、高通传感器中枢和内存子系统。

7.1 高通AI引擎中的处理器

高通最新的Hexagon NPU面向生成式AI带来了显著提升，性能提升98%、能效提升40%，包括微架构升级、增强的微切片推理、更低的内存带宽占用，以及专用电源传输轨道，以实现最优性能和能效。这些增强特性结合INT4硬件加速，使Hexagon NPU成为面向终端侧AI推理的领先处理器。

Qualcomm

Adreno GPU不仅是能够以低功耗进行高性能图形处理、赋能丰富用户体验的强大引擎，还可用于以高精度格式进行AI并行处理，支持32位浮点（FP32）、16位浮点（FP16）和8位整数（INT8）运算。第三代骁龙8中全新升级的 Adreno GPU实现了25%的能效提升，增强了AI、游戏和流媒体能力。基于Adreno GPU，Llama 2-7B每秒可生成超过13个tokens。

正如上一章节所述，CPU擅长时延敏感型的低计算量AI工作负载。在骁龙®X Elite计算平台中，高通 Oryon CPU作为PC领域的全新CPU领军者，可提供高达竞品两倍的CPU性能，达到竞品峰值性能时功耗仅为竞品的三分之一。

始终在线的处理器对于处理面向泛在型生成式AI应用的情境化信息至关重要。高通AI引擎集成的高通传感器中枢是一款极其高效、始终在线的AI处理器，适用于需要全天候运行的小型神经网络和泛在型应用，比如情境感知和传感器处理，所需电流通常不超过1毫安（mA）。第三代骁龙8中全新升级的高通传感器中枢相比前代性能提升3.5倍，内存增加30%，并配备两个下一代微型NPU，能够实现增强的AI性能。高通传感器中枢具备专用电源传输轨道，可在SoC其余部分关闭时运行，从而大幅节省电量。

高通AI引擎中的所有处理器相辅相成，能够实现AI处理效率的大幅度提升。

7.2 高通AI异构计算的系统级解决方案

异构计算涵盖整个SoC，包括多样化处理器、系统架构和软件三个层级，因此在异构计算解决方案中应用系统级方法至关重要。全局视角让高通架构师可以评估每个层级之间的关键约束条件、需求和依赖关系，从而针对SoC和最终产品用途做出恰当的选择，比如如何设计共享内存子系统或决定不同处理器应支持的数据类型。高通定制设计了整个系统，因此我们能够做出恰当的设计权衡，并利用这些洞察打造更具协同性的解决方案。

定制设计方法为高通解决方案带来了差异化优势，我们可以为每类处理器插入全新的AI指令或硬件加速器。高通致力于推动面向异构计算特性的架构演进，同时保持处理器多样性这一优势。如果所有处理器都采用相近的架构，那么SoC将变成同构系统。

相比之下，许多芯片组厂商通常选择授权多个第三方处理器，然后拼装在一起。这些处理器不一定能够紧密配合，也不一定是针对相同约束条件或细分市场而设计的。

高通AI引擎是我们终端侧AI优势的核心，它在骁龙平台和众多高通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，支持当前和未来的用例。搭载高通AI引擎的产品出货量已超过20亿，赋能了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。[2]

7.3 案例研究：使用异构计算的虚拟化身AI个人助手

在2023骁龙峰会上，高通在搭载第三代骁龙8移动平台的智能手机上演示了语音控制的AI个人助手，支持手机屏幕上的虚拟化身实现实时动画效果。该应用需要同时基于不同计算需求，运行众多复杂工作负载。实现优秀用户体验的关键在于充分利用SoC内的处理器多样性，在最匹配的处理器上运行合适的工作负载。

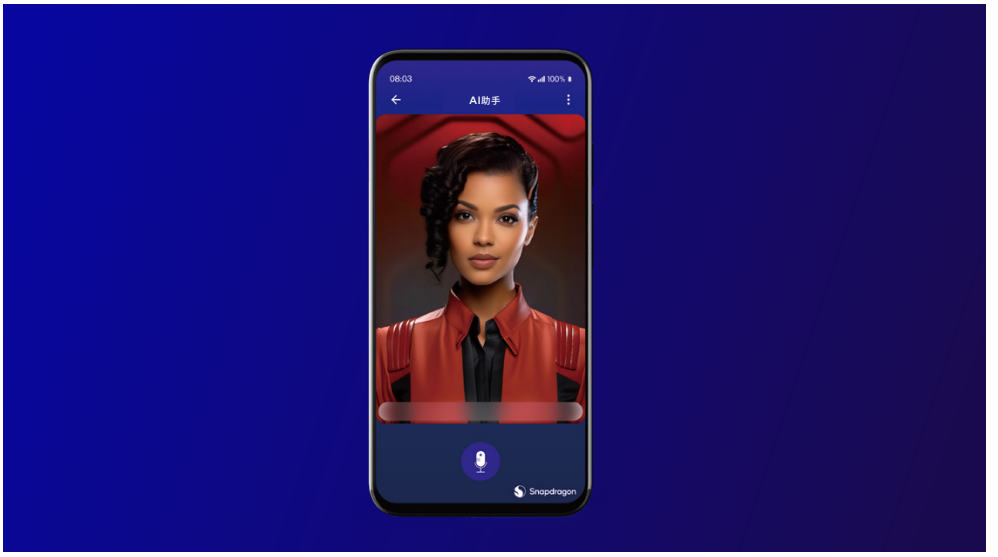


图7：虚拟化身AI助手包括众多复杂工作负载。

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

让我们看看该如何分配这一用例的工作负载：

1. 当用户与AI助手交谈时，语音通过OpenAI的自动语音识别（ASR）生成式AI模型Whisper转化为文本。该模型在高通传感器中枢上运行。
2. AI助手再使用大语言模型Llama 2-7B生成文本回复。该模型在NPU上运行。
3. 然后利用在CPU上运行的开源TTS模型将文本转化为语音。
4. 与此同时，虚拟化身渲染必须与语音输出同步，才能实现足够真实的用户交互界面。借助音频创建融合变形动画（blendshape）能够给嘴形和面部表情带来合适的动画效果。这一传统AI工作负载在NPU上运行。
5. 最终的虚拟化身渲染在GPU上进行。以上步骤需要在整个内存子系统中高效传输数据，尽可能在芯片上保存数据。

这一个人助手演示利用了高通AI引擎上的所有多样化处理器，以高效处理生成式和传统AI工作负载。

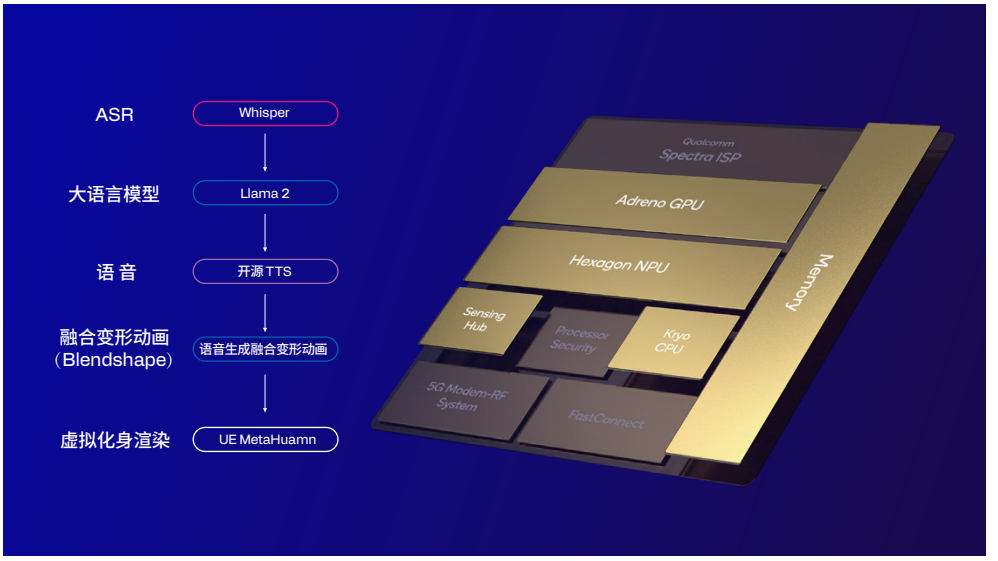


图8：支持虚拟化身的个人助手充分利用高通AI引擎的所有多样化处理器。

• 8. 骁龙平台领先的AI性能

实现领先性能需要卓越的硬件和软件。尽管每秒万亿次运算(TOPS)数值能够反映硬件性能潜力，但决定硬件可访问性和总体利用率的是软件。AI基准测试可以更好的展示性能，但最终的评估方式还是在实际应用中，测试峰值性能、持续稳定性能和能效。由于生成式AI基准测试和应用仍处于起步阶段，以下对当前领先AI指标的分析展示了骁龙平台的领先性能。

8.1 第三代骁龙8的领先智能手机上AI性能

在MLCommon MLPerf推理：Mobile V3.1基准测试中，与其他智能手机竞品相比，第三代骁龙8具有领先性能。例如，在生成式AI语言理解模型MobileBERT上，第三代骁龙8的表现比竞品A高17%，比竞品B高321%[3]。在鲁大师AIMark V4.3基准测试中，第三代骁龙8的总分分别为竞品B的5.7倍和竞品C的7.9倍。在安兔兔AITuTu基准测试中，第三代骁龙8的总分是竞品B的6.3倍。

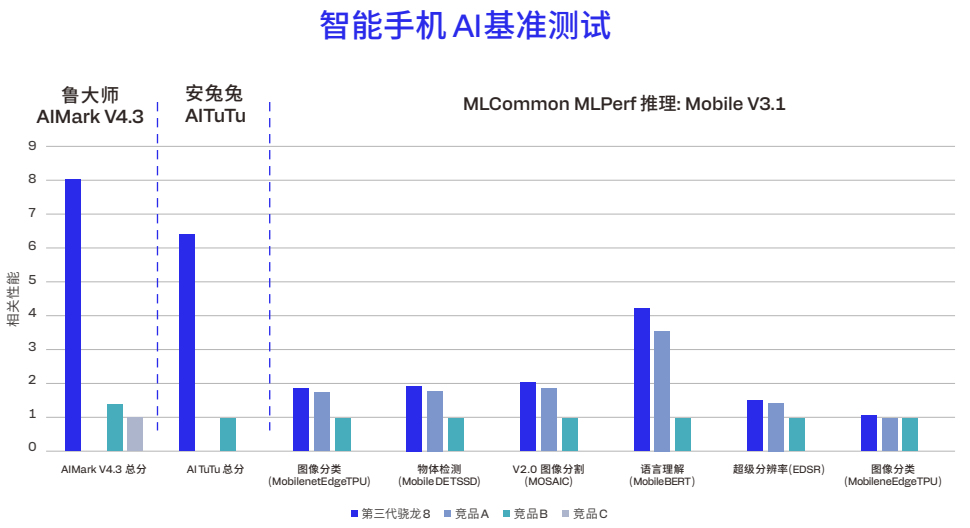


图9：第三代骁龙8在AIMark、AITuTu和MLPerf中具有领先的智能手机AI性能。

[3] 高通技术公司在搭载骁龙和竞品B平台的手机上运行和收集数据。竞品A数据为其自身披露。

在2023年骁龙峰会上，高通演示过两个生成式AI应用，展示了面向大语言模型和大视觉模型通用架构的真实应用性能。在第三代骁龙8上，个人助手演示能够以高达每秒20个tokens的速度运行Llama 2-7B。在不损失太多精度的情况下，Fast Stable Diffusion能够在0.6秒内生成一张512x512分辨率的图像[4]。高通有着智能手机领域领先的Llama和Stable Diffusion模型指标。

8.2 骁龙X Elite的领先PC上AI性能

骁龙X Elite上集成的Hexagon NPU算力达到45 TOPS，大幅领先于友商最新X86架构芯片NPU的算力数值。在面向Windows的UL Procyon AI基准测试中，与其他PC竞品相比，骁龙X Elite具有领先的性能。例如，骁龙X Elite的基准测试总分分别为X86架构竞品A的3.4倍和竞品B的8.6倍。

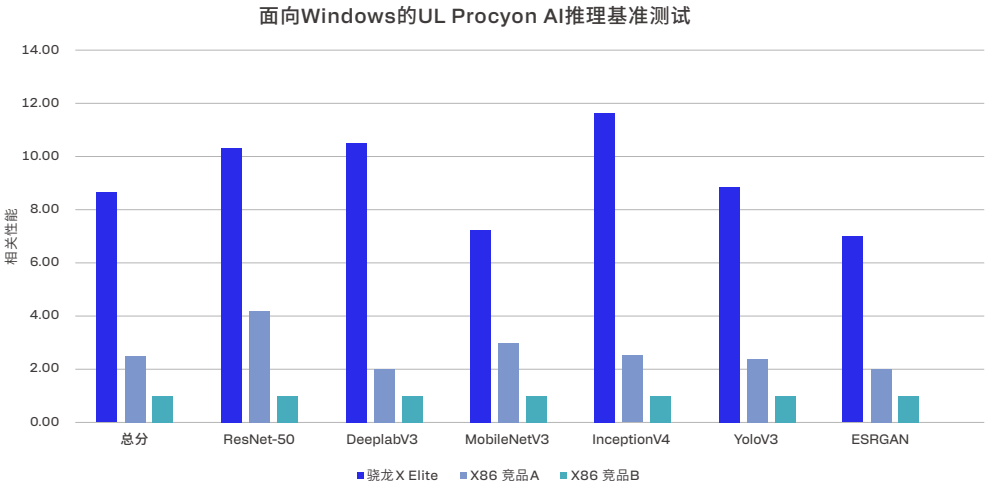


图10：骁龙X Elite在Procyon基准测试中具有领先的笔记本电脑AI性能。

在骁龙X Elite上，Llama 2-7B模型能够在高通Oryon CPU上以高达每秒30个tokens的速度运行。在不损失太多精度的情况下，Fast Stable Diffusion能够在0.9秒内生成一张512x512分辨率的图像。高通有着笔记本电脑领域领先的Llama和Stable Diffusion模型指标。

[4] 基于对比性语言-图像预训练（CLIP）模型分数，用于评估准确性，接近基线模型。

• 9. 通过高通软件栈访问AI处理器

仅有优秀的AI硬件还不够。让开发者能够获取基于异构计算的AI加速，对于终端侧AI的规模化扩展至关重要。高通AI软件栈将我们的互补性AI软件产品整合在统一的解决方案中。OEM厂商和开发者可在高通的产品上创建、优化和部署AI应用，充分利用高通AI引擎的性能，让开发者创建一次AI模型，即可跨不同产品随时随地进行部署。

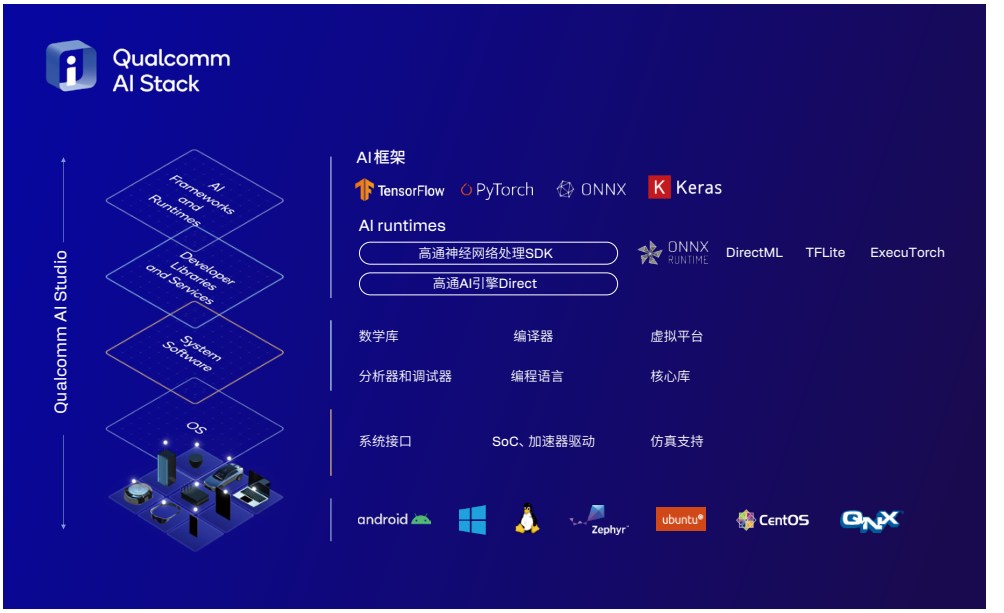


图11：高通AI软件栈旨在帮助开发者一次编写，即可实现随时随地运行和规模化扩展。

高通AI软件栈全面支持主流AI框架（如TensorFlow、PyTorch、ONNX和Keras）和runtime（如TensorFlow Lite、TensorFlow Lite Micro、ExecuTorch和ONNX runtime），面向以上runtime的代理对象可通过高通AI引擎Direct软件开发包（SDK）直接进行耦合，加快开发进程。

Qualcomm 高通

AI应用如何一次开发多端运行？

#让AI触手可及

扫码下载
高通AI白皮书电子版

此外，高通AI软件栈集成用于推理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。

在软件栈更底层，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。我们还跨不同产品线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）。

对于GPU的直接跨平台访问，我们支持OpenCL和DirectML。由于易于编程且应用于所有平台，CPU通常是AI编程的首选，我们的LLVM编译器基础设施优化可实现加速的高效AI推理。

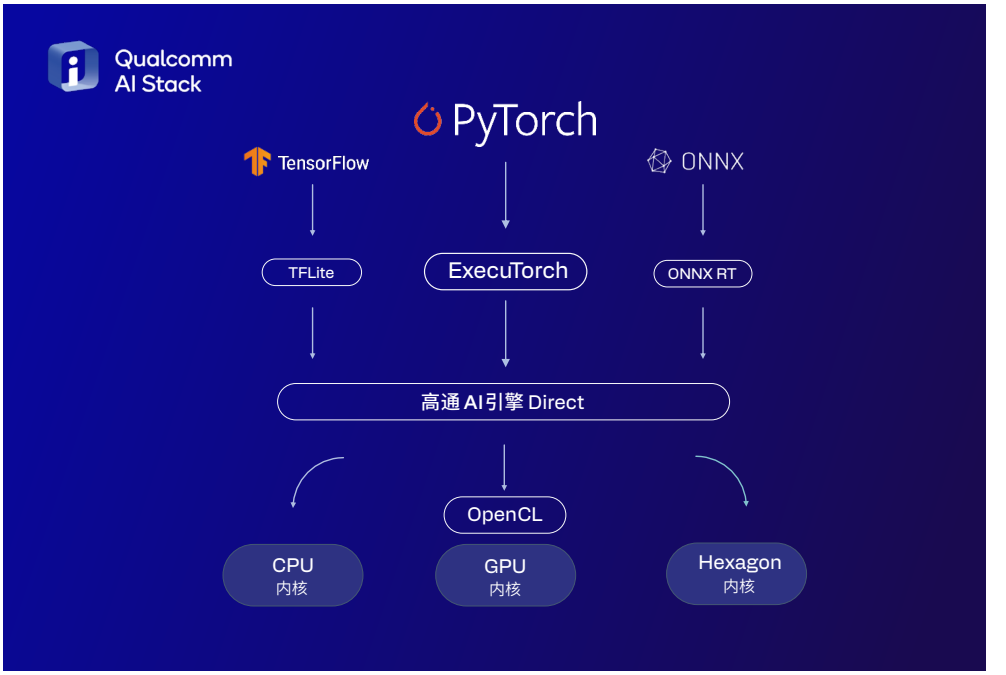


图12：高通AI软件栈支持关键框架和runtime。

高通专注于AI模型优化以实现能效和性能提升。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，在不牺牲太多准确度的前提下缩减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。

例如，量化有益于提升性能、能效、内存带宽和存储空间。Hexagon NPU原生支持INT4，高通AI模型增效工具包（AIMET）[5]提供基于高通AI研究技术成果开发的量化工具，能够在降低位数精度的同时限制准确度的损失。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。

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。事实上，INT4已成为大语言模型的趋势，并逐渐成为范式，尤其是面向开源社区和希望在边缘终端上运行大型参数规模模型的情况下。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。使用低位整数型精度对高能效推理至关重要。

• 10. 总结

利用多种处理器进行异构计算，对于实现生成式AI应用最佳性能和能效至关重要。与竞品相比，专为持久稳定的高性能AI推理而打造的Hexagon NPU具有卓越性能、能效和面积效率。高通AI引擎包括Hexagon NPU、Adreno GPU、高通Kryo或高通Oryon CPU、高通传感器中枢和内存子系统，能够支持按需型用例、持续型用例和泛在型用例，为生成式AI提供业界领先的异构计算解决方案。

通过定制设计整个系统，高通能够做出恰当的设计权衡，并利用这些洞察打造更具协同性的解决方案。我们的迭代改进和反馈循环，使高通能够基于最新神经网

[5] 高通AI模型增效工具包（AIMET）是高通创新中心公司（Qualcomm Innovation Center, Inc.）的产品。

络架构，持续快速增强高通NPU和高通AI软件栈。我们在面向智能手机和PC的AI基准测试与生成式AI应用中领先的性能表现，是高通差异化解决方案和全栈AI优化的结晶。

高通AI软件栈赋能开发者跨不同产品创建、优化和部署AI应用，使得高通AI引擎上的AI加速具备可获取性和可扩展性。通过将技术领导力、定制芯片设计、全栈AI优化和生态系统赋能充分结合，高通技术公司在推动终端侧生成式AI开发和应用方面独树一帜。

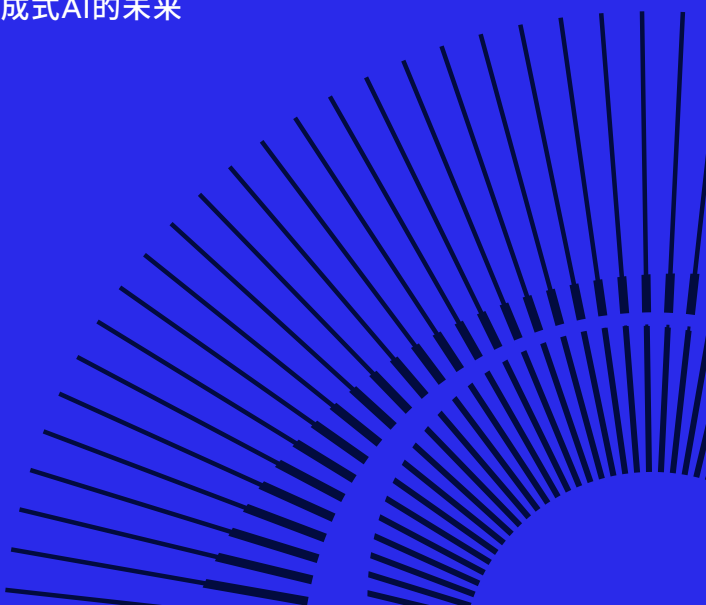
该部分信息发布于：2024年3月

生成式AI普及的关键是什么?

高通AI白皮书 第二部分

终端侧AI和混合AI开启生成式AI的未来

Unlocking the generative AI future
with on-device and hybrid AI



第二部分 PART TWO

终端侧AI和混合AI开启生成式AI的未来

Unlocking the generative AI future
with on-device and hybrid AI

• 1. 摘要

混合AI是AI的未来。随着生成式AI正以前所未有的速度发展[1]以及计算需求的日益增长[2]，AI处理必须分布在云端和终端进行，才能实现AI的规模化扩展并发挥其最大潜能——正如传统计算从大型主机和瘦客户端演变为当前云端和边缘终端相结合的模式。与仅在云端进行处理不同，混合AI架构在云端和边缘终端之间分配并协调AI工作负载。云端和边缘终端如智能手机、汽车、个人电脑和物联网终端协同工作，能够实现更强大、更高效且高度优化的AI。

节省成本是主要推动因素。举例来说，据估计，每一次基于生成式AI的网络搜索查询（query），其成本是传统搜索的10倍[3]，而这只是众多生成式AI的应用之一。混合AI将支持生成式AI开发者和提供商利用边缘终端的计算能力降低成本。混合AI架构或终端侧AI能够在全球范围带来高性能、个性化、隐私和安全等优势。

混合AI架构可以根据模型和查询需求的复杂度等因素，选择不同方式在云端和终端侧之间分配处理负载。例如，如果模型大小、提示（prompt）和生成长度小于某个限定值，并且能够提供可接受的精确度，推理即可完全在终端侧进行。如果是更复杂的任务，模型则可以跨云端和终端运行。混合AI还能支持模型在终端侧和云端同时运行，也就是在终端侧运行轻量版模型时，在云端并行处理完整模型的多个标记（token），并在需要时更正终端侧的处理结果。

随着强大的生成式AI模型不断缩小，以及终端侧处理能力的持续提升，混合AI的潜力将会进一步增长。参数超过10亿的AI模型已经能够在手机上运行，且性能和精确度水平达到与云端相似的水平。不久的将来，拥有100亿或更高参数的模型将能够在终端上运行。

混合AI方式适用于几乎所有生成式AI应用和终端领域，包括手机、笔记本电脑、XR头显、汽车和物联网。这一方式对推动生成式AI规模化扩展，满足全球企业与消费者需求至关重要。

[1] https://www.statista.com/chart/29174/time-to-one-million-users/
[2] https://siliconangle.com/2023/02/05/generative-ai-drives-explosion-compute-looming-need-sustainable-ai/
[3] https://www.reuters.com/technology/tech-giants-ai-like-bing-bard-poses-billion-dollar-search-problem-2023-02-22/

Qualcomm

• 2. 生成式AI简介和当前趋势

ChatGPT激发了人们的想象力和好奇心。自2022年11月推出后，短短两个月内其月活用户便达到1亿，成为有史以来增长速度最快的消费类应用和第一个杀手级的生成式AI应用。随着创新节奏的加快，想要紧跟生成式AI的发展速度，难度越来越大。大型聚合网站的数据显示，目前已有超过3,000个可用的生成式AI应用和特性[4]。AI正迎来大爆发时期，就像此前电视、互联网和智能手机的问世，而这仅仅是一个开始。

ChatGPT和Stable Diffusion等生成式AI模型能够基于简单的提示创作出全新的原创内容，如文本、图像、视频、音频或其他数据。这类模型正在颠覆传统的搜索、内容创作和推荐系统的方法——通过从普通产业到创意产业的跨行业用例，在实用性、生产力和娱乐性方面带来显著增强。建筑师和艺术家可以探索新思路，工程师可以更高效地编写程序。几乎所有与文字、图像、视频和自动化相关的工作领域都将受益。

网络搜索是生成式 AI正在变革的诸多应用之一。另一个例子则是Microsoft 365 Copilot，作为一项全新的生产力特性，它能够利用生成式AI帮助编写和总结文档、分析数据，或将简单的书面想法转化为演示文稿，嵌入于Word、Excel、PowerPoint、Outlook和Teams等微软应用中。

生成式AI的出现也标志着用户开始向探索更加多样化、个性化的数字世界迈出了第一步。由于3D设计师可以借助生成式AI工具更加快速高效地进行内容开发，3D内容创作有望得到普及。这不仅将加速沉浸式虚拟体验的创建，而且能够降低个人创作者自主内容制作的门槛。

我们即将看到从生成式AI中涌现出各种各样的全新企业级和消费级用例，带来超越想象的功能。GPT-4和LaMDA等通用大语言模型（LLM）作为基础模型，所具备的语言理解、生成能力和知识范畴已达到了前所未有的水平。这些模型大多数都非常庞大，参数超过1千亿，并通过API向客户提供免费或付费服务。

基础模型的使用推动大量初创公司和大型组织利用文本、图像、视频、3D、语言和音频创建应用。例如，代码生成（GitHub Copilot）、文本生成（Jasper）、面向艺术家和设计师的图像生成（Midjourney），以及对话式聊天机器人（Character.ai）。

[4] 截至2023年4月，生成式AI应用和特性：https://theresanaiforthat.com/

据初步估计显示，生成式AI市场规模将达到1万亿美元[5]，广泛覆盖生态链的各个参与方。为把握这一巨大机遇，并推动AI成为主流，计算架构需要不断演进并满足大规模生成式AI日益增长的处理和性能需求。

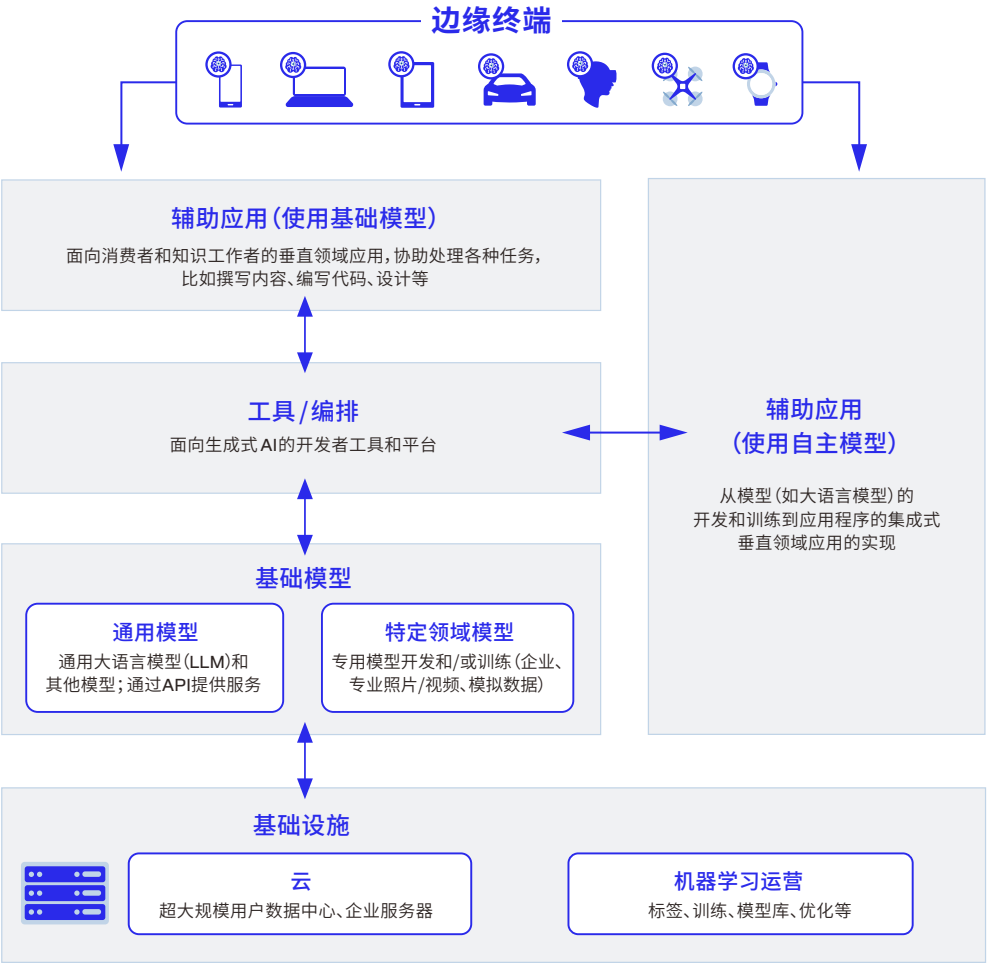


图1：生成式AI生态链使应用数量激增

[5] 瑞银，2023年2月

Qualcomm 高通

生成式AI搜索成本是传统搜索的10倍*

终端侧AI可以
花更少,算更多

#让AI触手可及

扫码下载
高通AI白皮书电子版

*来源：摩根士丹利，《How Large are the Incremental AI Costs...and 4 Factors to Watch Next》，2023年2月

• 3. 混合AI对生成式AI规模化扩展至关重要

拥有数十亿参数的众多生成式AI模型对计算基础设施提出了极高的需求。因此，无论是为AI模型优化参数的AI训练，还是执行该模型的AI推理，至今都一直受限于大型复杂模型而在云端部署。

AI推理的规模远高于AI训练。尽管训练单个模型会消耗大量资源，但大型生成式AI模型预计每年仅需训练几次。然而，这些模型的推理成本将随着日活用户数量及其使用频率的增加而增加。在云端进行推理的成本极高，这将导致规模化扩展难以持续。

混合AI能够解决上述问题，正如传统计算从大型主机和瘦客户端演变为当前云端和PC、智能手机等边缘终端相结合的模式。

3.1 什么是混合AI?

混合AI指终端和云端协同工作，在适当的场景和时间下分配AI计算的工作负载，以提供更好的体验，并高效利用资源。在一些场景下，计算将主要以终端为中心，在必要时向云端分流任务。而在以云为中心的场景下，终端将根据自身能力，在可能的情况下从云端分担一些AI工作负载。

3.2 混合AI的优势

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。

3.2.1 成本

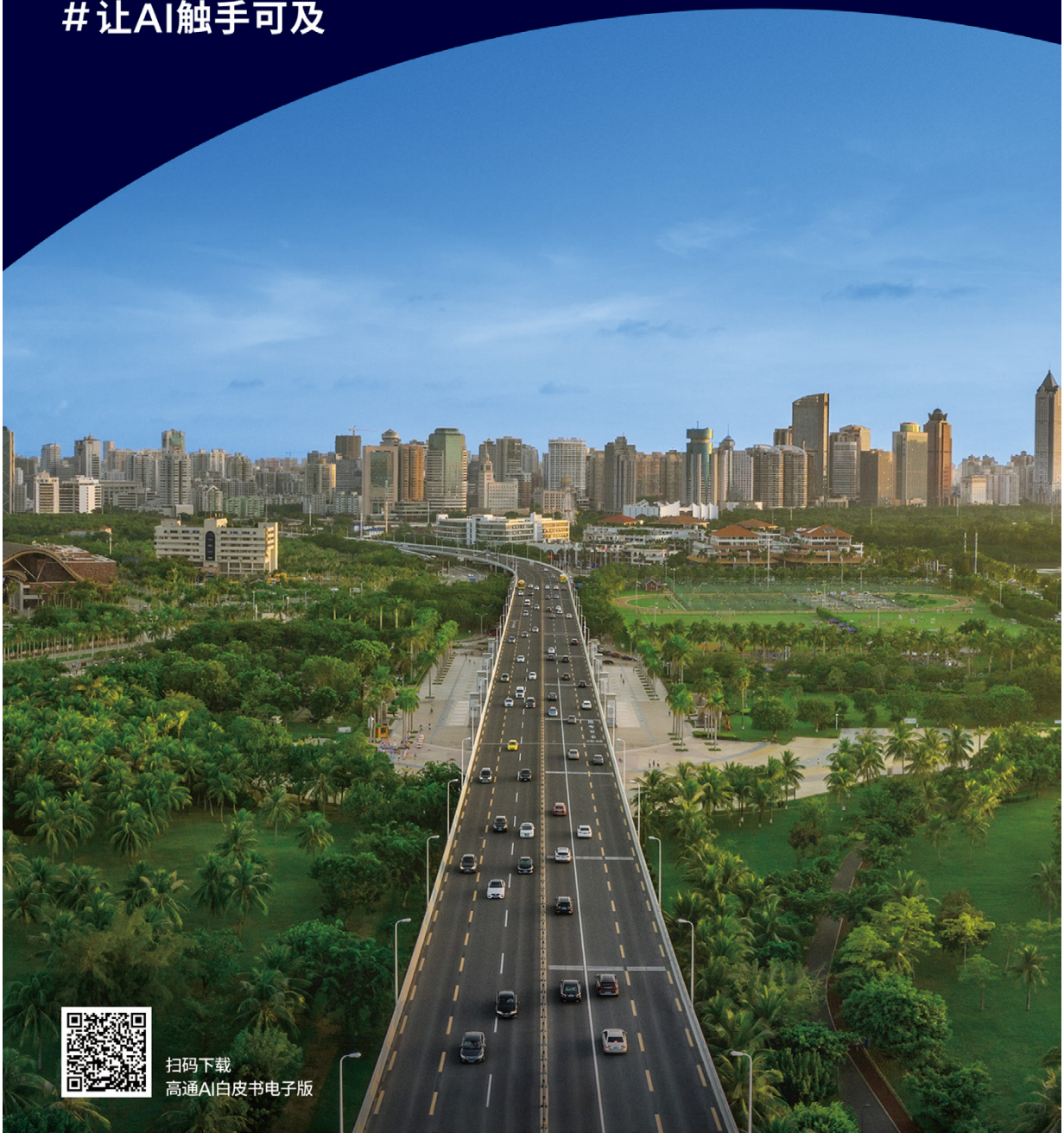
随着生成式AI模型使用量和复杂性的不断增长，仅在云端进行推理并不划算。因为数据中心基础设施成本，包括硬件、场地、能耗、运营、额外带宽和网络传输的成本将持续增加。

例如，当前面向大语言模型推理的云计算架构，将导致无论规模大小的搜索引擎企业负担更高运营成本。试想一下，未来通过生成式AI大语言模型增强的互联网搜索，比如GPT，其运行参数远超1750亿。生成式AI搜索可以提供更加出色的用户体验

Qualcomm 高通

用更节能的AI
开启更有生机的未来

#让AI触手可及



扫码下载
高通AI白皮书电子版

和搜索结果，但每一次搜索查询(query)其成本是传统搜索方法的10倍。目前每天有超过100亿次的搜索查询产生，即便基于大语言模型的搜索仅占其中一小部分，每年增量成本也可能达到数十亿美元。[6]

将一些处理从云端转移到边缘终端，可以减轻云基础设施的压力并减少开支。这使混合AI对生成式AI的持续规模化扩展变得至关重要。混合AI能够利用现已部署的、具备AI能力的数十亿边缘终端，以及未来还将具备更高处理能力的数十亿终端。

节省成本也是生成式AI生态系统发展的重要一环，可以支持OEM厂商、独立软件开发商(ISV)和应用开发者更经济实惠地探索和打造应用。例如，开发者可以基于完全在终端上运行的Stable Diffusion创建应用程序，对于生成的每个图像承担更低的查询成本，或完全没有成本。

3.2.2 能耗

支持高效AI处理的边缘终端能够提供领先的能效，尤其是与云端相比。边缘终端能够以很低的能耗运行生成式AI模型，尤其是将处理和数据传输相结合时。这一能耗成本差异非常明显，同时能帮助云服务提供商降低数据中心的能耗，实现环境和可持续发展目标。

3.2.3 可靠性、性能和时延

在混合AI架构中，终端侧AI处理十分可靠，能够在云服务器和网络连接拥堵时，提供媲美云端甚至更佳的性能[7]。当生成式AI查询对于云的需求达到高峰期时，会产生大量排队等待和高时延，甚至可能出现拒绝服务的情况[8]。向边缘终端转移计算负载可防止这一现象发生。此外，混合AI架构中终端侧处理的可用性优势，让用户无论身处何地，甚至在无连接的情况下，依然能够正常运行生成式AI应用。

3.2.4 隐私和安全

终端侧AI从本质上有助于保护用户隐私，因为查询和个人信息完全保留在终端上。对于企业和工作场所等场景中使用的生成式AI，这有助于解决保护公司保密信息的难题。例如，用于代码生成的编程助手应用可以在终端上运行，不向云端暴露保密信息，

[6] 摩根士丹利，《How Large are the Incremental AI Costs...and 4 Factors to Watch Next》，2023年2月
[7] https://www.qualcomm.com/news/onq/2023/02/worlds-first-on-device-demonstration-of-stable-diffusion-on-android
[8] https://www.digitaltrends.com/computing/chatgpt-is-at-capacity-and-is-frustrating-new-people-everywhere/

从而消除如今众多企业面临的顾虑[9]。对于消费者使用而言，混合AI架构中的“隐私模式”让用户能够充分利用终端侧AI向聊天机器人输入敏感提示，比如健康问题或创业想法。此外，终端侧安全能力已经十分强大，并且将不断演进，确保个人数据和模型参数在边缘终端上的安全。

3.2.5 个性化

混合AI让更加个性化的体验成为可能。数字助手将能够在不牺牲隐私的情况下，根据用户的表情、喜好和个性进行定制。所形成的用户画像能够从实际行为、价值观、痛点、需求、顾虑和问题等方面来体现一个用户，并且可以随着时间推移进行学习和演进。它可以用于增强和打造定制化的生成式AI提示，然后在终端侧或云端进行处理。用户画像保留在终端内，因此可以通过终端侧学习不断优化和更新。

个性化不仅仅适用于消费者，企业或机构可以借助它标准化代码的编写方式，或者制作具有特殊语气和声音的公共内容。

3.3 AI工作负载的分布式处理机制

我们期望打造能够支持不同工作负载分流方式的混合AI架构，可以根据模型和查询复杂度进行分布式处理，并能持续演进。例如，如果模型大小、提示和生成长度小于某个限定值，并且能够提供可接受的精确度，推理即可完全在终端侧进行。如果是更复杂的任务，模型则可以跨云端和终端运行；如果需要更多最新信息，那么也可以连接至互联网获取。

3.3.1 以终端为中心的混合AI

在以终端为中心的混合AI架构中，终端将充当锚点，云端仅用于分流处理终端无法充分执行的任务。许多生成式AI模型可以在终端上充分运行（参阅图2），也就是说终端可通过运行不太复杂的推理完成大部分处理工作。

例如，用户在笔记本电脑上运行Microsoft 365 Copilot或必应Chat时，包含高达数百亿参数的模型将在终端上运行，而更复杂的模型将根据需求在云端进行处理。对用户来说，这种体验是无缝的，因为终端侧神经网络或基于规则而运行的判决器（arbiter）将决定是否需要使用云端，无论是为了有机会使用更好的模型还是

[9] https://www.pcmag.com/news/samsung-software-engineers-busted-for-pasting-proprietary-code-into-chatgpt

检索互联网信息。如果用户对请求处理结果的质量不满意，那么再次尝试发起请求时可能就会引入一个更好的模型。由于终端侧AI处理能力随着终端升级和芯片迭代不断提升，它可以分流更多云端的负载。

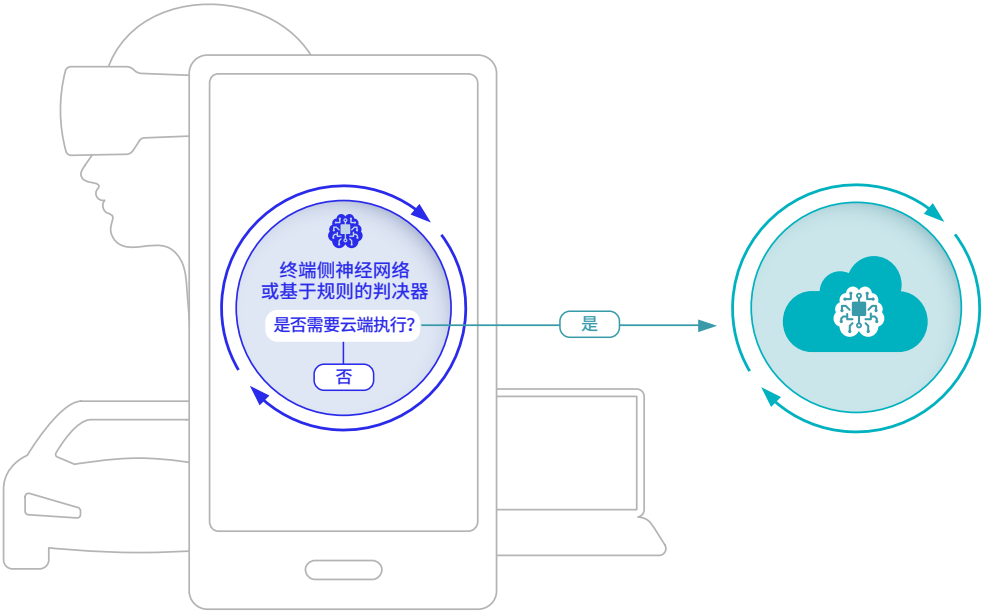


图2：在以终端为中心的混合AI架构中，云端仅用于分流处理终端无法充分运行的AI任务。

对于各种生成式AI应用，比如创作图像或起草邮件，快速响应式的推理更受青睐，即使它在准确度上会稍有损失。终端侧AI的快速反馈（即低时延）可以让用户使用改进的提示来快速迭代推理过程，直至获得满意的输出结果。

3.3.2 基于终端感知的混合AI

在基于终端感知的混合AI场景中，在边缘侧运行的模型将充当云端大语言模型（类似大脑）的传感器输入端（类似眼睛和耳朵）。例如，当用户对智能手机说话时，Whisper等自动语音识别（ASR）的AI模型将在终端侧运行，将语音转为文字，然后将其作为请求提示发送到云端。云端将运行大语言模型，再将生成的文本回复发回终端。之后，终端将运行文本生成语音（TTS）模型，提供自然免提回答。将自动语音识别和文本生成语音模型工作负载转移至终端侧能够节省计算和连接带宽。随着大语言模型变为多模态并支持图像输入，计算机视觉处理也可以在终端上运行，以进一步分流计算任务并减少连接带宽，从而节省成本。

在更先进的版本中，隐私将得到进一步保护，终端侧AI能够承担更多处理，并向云端提供经过改进且更加个性化的提示。借助终端侧学习和终端上的个人数据，比如社交媒体、电子邮件、消息、日历和位置等，终端将创建用户的个人画像，与编排器（orchestrator）程序协作，基于更多情境信息提供更完善的提示。例如，如果用户让手机来安排与好友会面的时间并在喜爱的餐厅预订座位，编排器程序了解上述个性化信息并能够向云端大语言模型提供更佳提示。编排器程序可在大语言模型缺乏信息时设置护栏并帮助防止产生“AI幻觉”。对于较简单的请求，较小的大语言模型可在终端侧运行，而无需与云端交互，这类似于以终端为中心的混合AI。

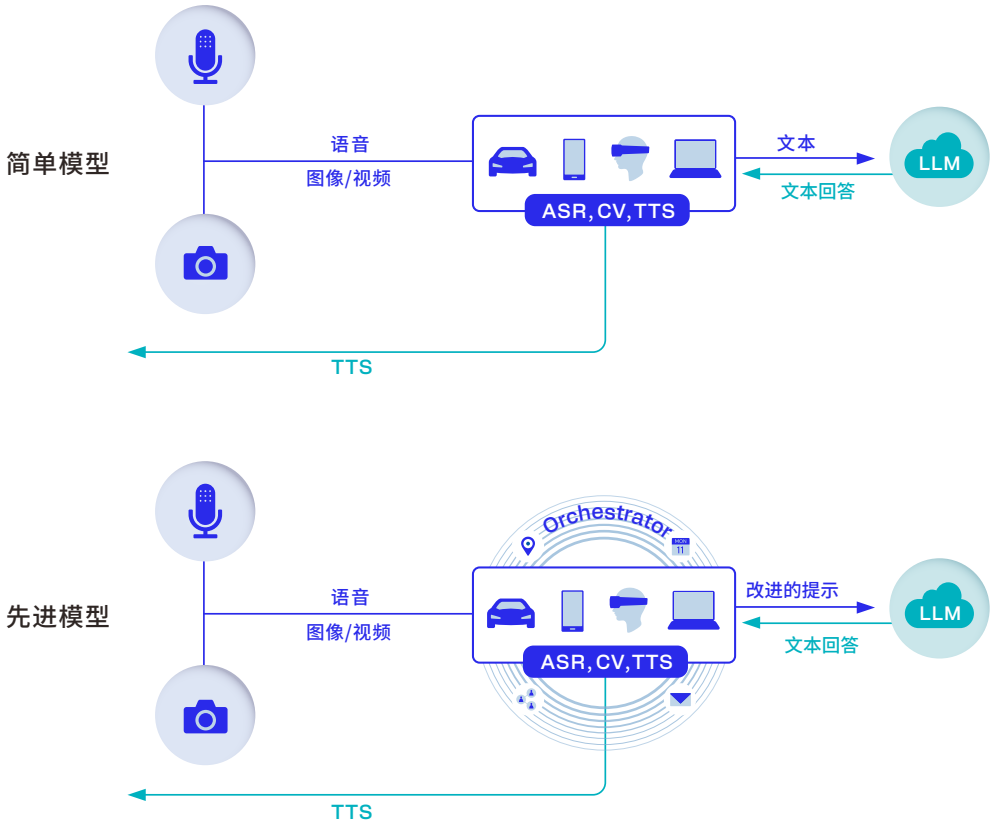


图3：对于基于终端感知的混合AI，自动语音识别、计算机视觉和文本转语音在终端侧进行。
在更先进的版本中，终端侧编排器程序能够向云端提供经过改进且更加个性化的提示。

3.3.3 终端与云端协同处理的混合AI

终端和云端的AI计算也可以协同工作来处理AI负载，生成大语言模型的多个token就是一个例子。大语言模型的运行都是内存受限的，这意味着计算硬件在等待来自DRAM的内存数据时经常处于闲置状态。大语言模型每次推理生成一个token，也就是基本等同于一个单词，这意味着GPT-3等模型必须读取全部1750亿参数才能生成一个单词，然后再次运行整个模型来生成下一个token，完整的推理过程可以以此类推。鉴于内存读取是造成推理性能的瓶颈因素，更高效的做法就是同时运行多个大语言模型以生成多个token，并且从DRAM一次性读取全部参数。每生成一个token就要读取全部参数会产生能耗和造成发热，因此使用闲置的算力通过共享参数来推测性并行运行大语言模型，可谓是在性能和能耗上实现双赢。

为了生成四个token，一个近似的大语言模型（比原始目标大语言模型小7至10倍，因此准确性更低）要在终端上按顺序连续运行四次才可以。终端向云端发送这四个token，云端高效运行四次目标模型来检查其准确度，而仅读取一次完整的模型参数。在云端token是被并行计算的，每个目标模型都有零个、一个、两个、三个或四个预测token作为输入。这些token在被云端确认或校正之前被认为是“近似的”。上述推测性解码过程将持续到完整的答案出现时为止。我们的早期实验和其他已发布结果[10]显示，通过四个token的推测性解码，平均两到三个token是正确可被接受的，这会带来单位时间内生成token数的增加，并节省能耗。

[10] Leviathan, Yaniv, Matan Kalman和Yossi Matias。《Fast Inference from Transformers via Speculative Decoding》。arXiv preprint arXiv:2211.17192 (2022)

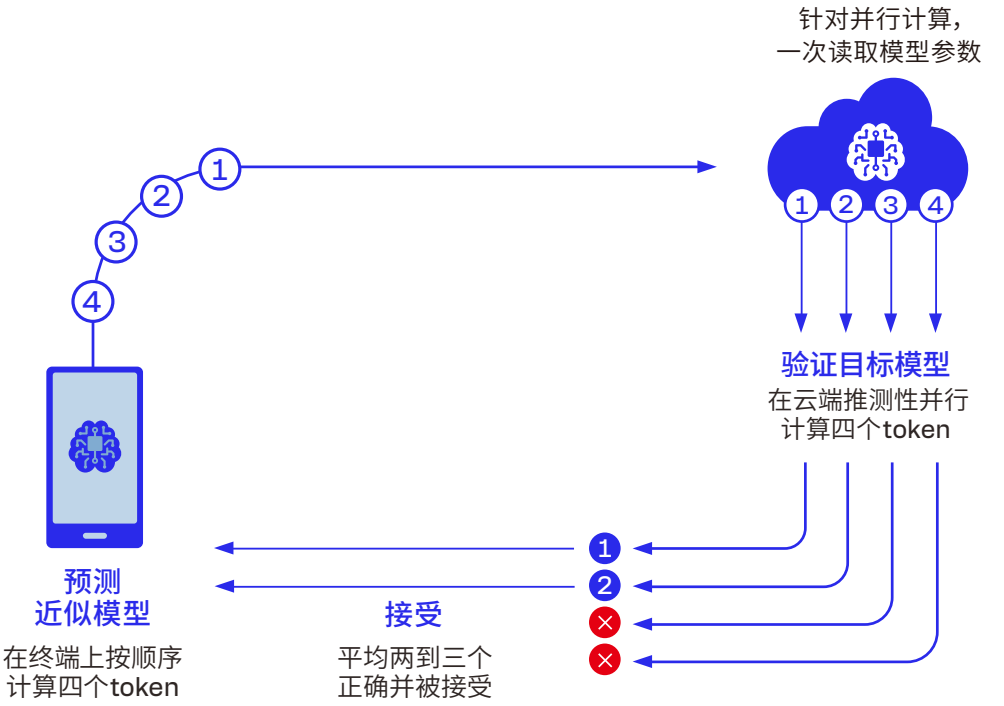


图4：协同处理混合AI的四个token推测性解码示例。

Qualcomm 高通

让个人隐私
留在个人身边

#让AI触手可及

扫码下载
高通AI白皮书电子版

• 4. 终端侧AI的演进与生成式AI的需求密切相关

终端侧AI能力是赋能混合AI并让生成式AI实现全球规模化扩展的关键。如何在云端和边缘终端之间分配处理任务将取决于终端能力、隐私和安全需求、性能需求以及商业模式等诸多因素(参阅第3.3章节)。

在生成式AI出现之前，AI处理便持续向边缘转移，越来越多的AI推理工作负载在手机、笔记本电脑、XR头显、汽车和其他边缘终端上运行。例如，手机利用终端侧AI支持许多日常功能，比如暗光拍摄、降噪和人脸解锁。

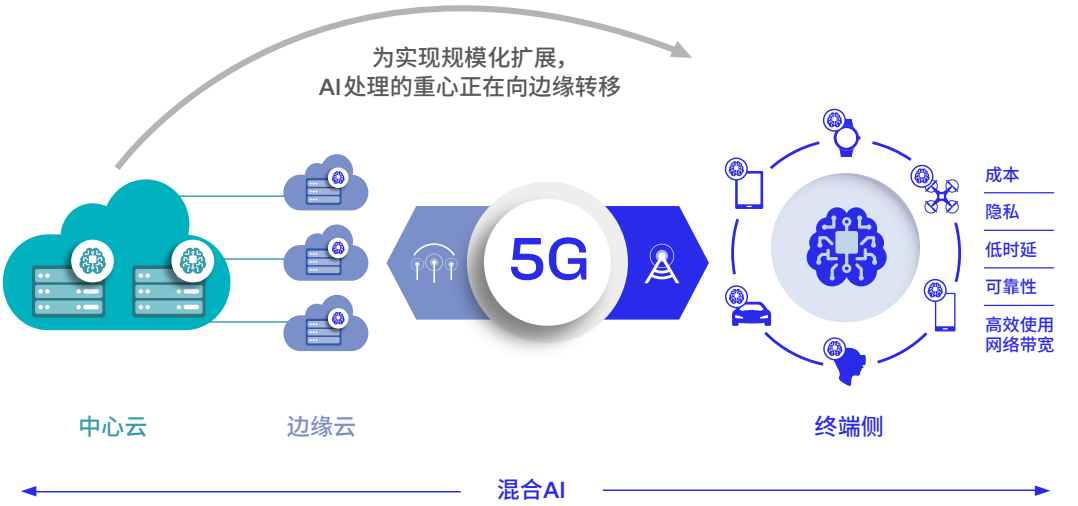


图5：AI处理的重心正在向边缘转移。

Qualcomm 高通

不上线
AI也始终在线

#让AI触手可及

扫码下载
高通AI白皮书电子版

4.1 终端侧处理能够支持多样化的生成式AI模型

如今，具备AI功能的手机、PC和其他品类的便携终端数量已达到数十亿台[11]，利用大规模终端侧AI处理支持生成式AI有着广阔前景，并且将在未来几年稳步增长。

关键问题在于，哪些生成式AI模型能够以合适的性能和准确度在终端侧运行。好消息是，性能十分强大的生成式AI模型正在变小，同时终端侧处理能力正在持续提升。图6展示了可以在终端侧运行的丰富的生成式AI功能，这些功能的模型参数在10亿至100亿之间[12]。如Stable Diffusion等参数超过10亿的模型已经能够在手机上运行，且性能和精确度达到与云端处理类似的水平。不久的将来，拥有100亿或更多参数的生成式AI模型将能够在终端上运行。

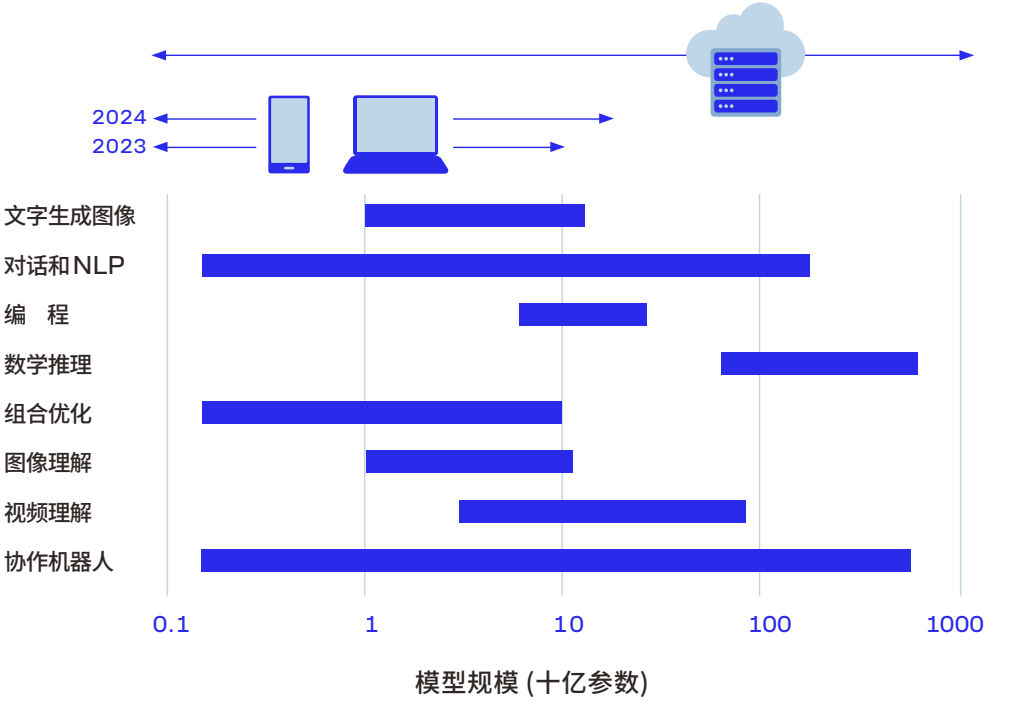


图6：数量可观的生成式AI模型可从云端分流到终端上运行。

[11] https://www.qualcomm.com/products/mobile/snapdragon/smartphones/mobile-ai
[12] 假设使用INT4型的参数

Qualcomm

• 5. 跨终端品类的生成式AI关键用例

基于基础模型的生成式AI迅速兴起，正在驱动新一轮内容生成、搜索和生产力相关用例的发展，覆盖包括智能手机、笔记本电脑和PC、汽车、XR以及物联网等终端品类。混合AI架构将赋能生成式AI在上述这些终端领域提供全新的增强用户体验。

5.1 智能手机：搜索和数字助手

面对每日超过100亿次的搜索量且移动端搜索占比超过60%的情况[13]，生成式AI的应用将推动所需算力的实质性增长，尤其是来自智能手机端的搜索请求。由于基于生成式AI的查询能够提供更令人满意的答案，用户的搜索方式已经开始发生转变。

对话式搜索的普及也将增加总体查询量。随着对话功能不断改进，变得更加强大，智能手机将成为真正的数字助手。精准的终端侧用户画像与能够理解文字、语音、图像、视频和任何其他输入模态的大语言模型相结合，让用户可以自然地沟通，获取准确、贴切的回答。进行自然语言处理、图像理解、视频理解、文本生成文本等任务的模型将面临高需求。

5.2 笔记本电脑和PC：生产力

生成式AI基于简单提示就能快速生成优质内容，它也正在凭借这项能力变革生产力。以笔记本电脑和PC上的Microsoft Office 365为例，全球有超过4亿Microsoft Office 365商业付费席位和个人订阅者，如果将生成式AI集成至用户日常工作流将带来重大影响[14]。此前需要数小时或数天的任务，现在仅需几分钟就能完成。Microsoft 365 Copilot同时利用大语言模型的功能和Microsoft Graph与Microsoft 365应用中的用户数据，能够将提示转化为强大的生产力工具[15]。

Office工作者可通过后台运行大语言模型，在Outlook中阅读或撰写电子邮件，在Word中编写文档，在PowerPoint中创建演示文稿，在Excel中分析数据，或在Teams会议中协作。生成式AI模型（比如自然语言处理、文本生成文本、图像生成、视频生成和编程）需要经过海量处理，才能支持这些被重度使用的生产力任务。在以终端为中心的混合AI架构中，大部分处理能够在PC上进行。

[13] https://www.statista.com/statistics/297137/mobile-share-of-us-organic-search-engine-visits/
[14] 微软财报
[15] https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/

5.3 汽车：数字助手和自动驾驶

得益于车内和车辆周围环境相关数据所提供的信息，如今AI驱动的座舱能够提供高度个性化的体验。类似于智能手机和PC，车载数字助手将能够让驾乘人员通过免提的友好用户界面保持无缝互联，同时为生态系统创造全新的创收机会。

数字助手可以访问用户个人数据，比如应用、服务和支付信息；以及来自车辆的传感器数据，包括摄像头、雷达、激光雷达和蜂窝车联网（C-V2X）等。企业API也支持第三方服务提供商集成他们的解决方案，将客户关系延伸到车上。例如，主动式驾驶辅助将大幅改善导航体验，比如会影响驾驶员常用出行路线的交通和天气信息更新，汽车充电或购买停车券提醒，此外，用户可以通过简单地请求即可用已绑定的信用卡预订自己喜欢的美食。如果汽车能够识别每位驾乘人员并提供定制化的音乐和播客等体验和内容，座舱的媒体娱乐体验也将会变革。随着车载AR应用变得更加普遍，数字助手可以按照驾乘人员的偏好提供定制化的显示。

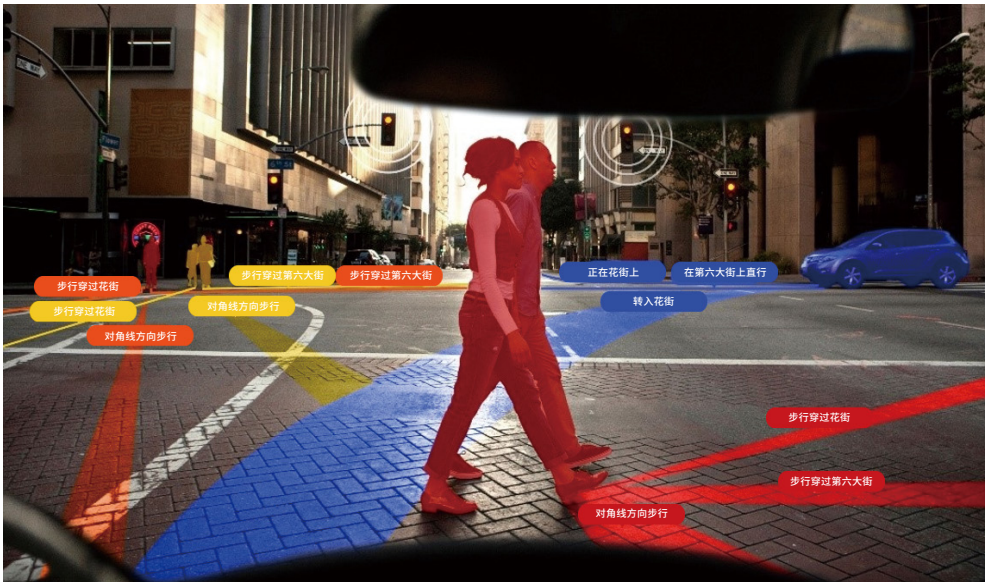


图7：生成式AI可用于先进驾驶辅助系统/自动驾驶（ADAS/AD），通过预测不同行为主体的轨迹和行为，帮助改进驾驶策略。

Qualcomm 高通

让数字助手
想你所想,行你所行

#让AI触手可及

扫码下载
高通AI白皮书电子版

汽车维修保养和服务也将变得更加自主和无缝。通过分析传感器输入、维修保养历史和驾驶行为等数据，数字助手可以预测何时需要进行保养。利用生成式AI，数字助手可针对汽车如何维修提供信息，或为用户提供咨询，找到合适的服务提供商，提高车辆可靠性，同时减少时间和成本。

感知软件栈从未遇到过的罕见或陌生物体，经常会对高级驾驶辅助系统和自动驾驶（ADAS/AD）解决方案产生干扰。这种情况通常由光线不佳或恶劣天气条件造成，会导致驾驶策略软件栈产生难以预测、有时甚至很危险的结果。为了在未来预防类似情况，必须妥善采集和标记这些极端场景的数据并重新训练模型。这个循环可能耗时费力，而生成式AI可以模拟极端场景，预测不同道路行为主体的轨迹和行为，比如车辆、行人、自行车骑行者和摩托车骑行者。规划者可以利用这些场景确定车辆驾驶策略。

驾驶策略软件栈以及感知软件栈始终在汽车的AI算力可支持的情况下本地运行。严苛的时延要求决定了云端无法针对这些AI工作负载在决策过程中发挥任何作用。随着ADAS/AD解决方案采用支持适当后处理的生成式AI模型，汽车必然需要具备显著高能效的AI计算能力。

5.4 XR：3D内容创作和沉浸式体验

生成式AI能为XR带来巨大前景。它有潜力普及3D内容创作，并真正实现虚拟化身。下一代AI渲染工具将赋能内容创作者使用如文本、语音、图像或视频等各种类型的提示，生成3D物体和场景，并最终创造出完整的虚拟世界。此外，内容创作者将能够利用文本生成文本的大语言模型，为能够发出声音并表达情绪的虚拟化身生成类人对话。总而言之，这些进步将变革用户在XR设备上创造和体验沉浸式内容的方式。

生成式AI为XR提供的前景无疑令人兴奋，但很难预测这些技术何时才能被广泛采用。不过，根据近几个月快速的创新步伐，可以肯定地说，我们可以期待在未来几年内取得重要进展。

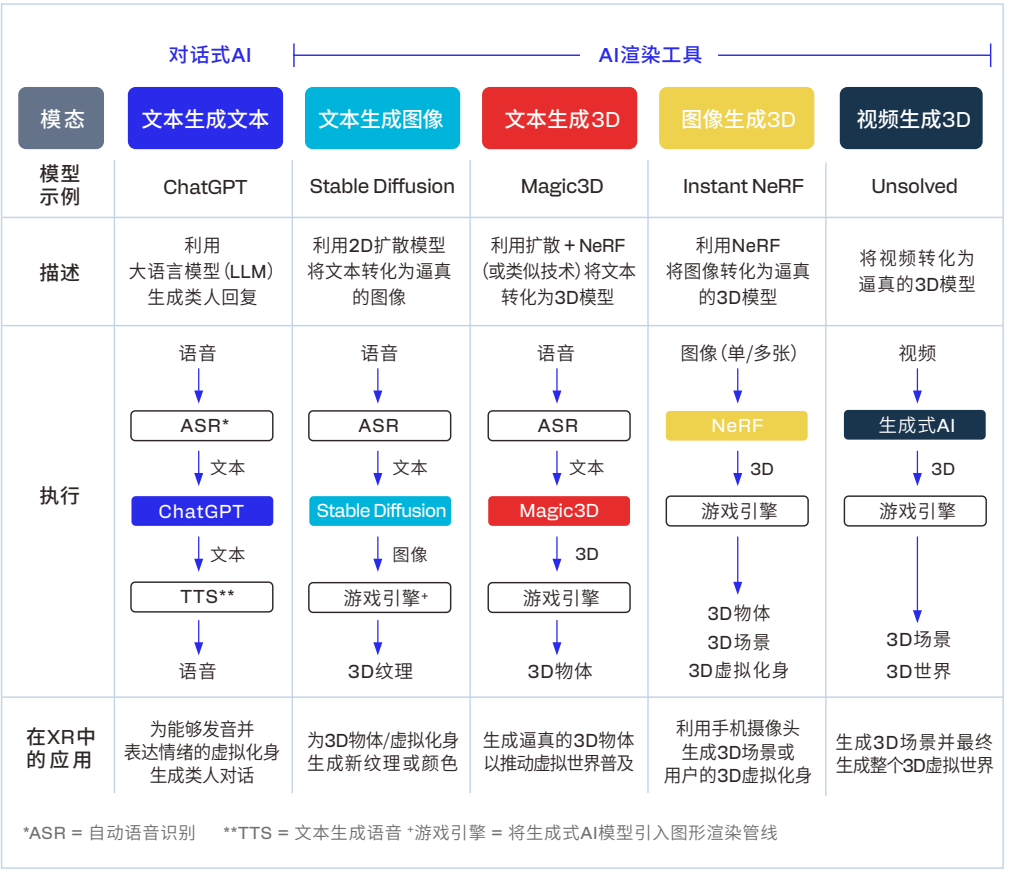


图8：生成式AI模型将面向XR赋能对话式AI和全新渲染工具。

对于沉浸式世界，Stable Diffusion等文本生成图像类的模型很快将赋能内容创作者在3D物体上生成逼真的纹理。我们预计，一年内这些功能将在智能手机上实现，并延伸到XR终端。XR中的部署需要“分布式处理”，即头显运行感知和渲染软件栈，与之配对的智能手机或云端运行生成式AI模型。未来几年，首批文本生成3D和图像生成3D类的模型将可能实现边缘侧部署，生成高质量的3D物体点云。几年后，

这些模型将通过提升，达到能够从零开始生成高质量3D纹理物体的水平。在大约十年内，模型将更进一步，支持由文本或图像生成的高保真完整3D空间和场景。未来，文本生成3D和视频生成3D类的模型最终或能让用户踏入从零开始生成的3D虚拟世界，例如自动构建满足用户任何想象的3D虚拟环境。

图9：生成式AI将有助于基于简单提示创造沉浸式3D虚拟世界的过程，
比如“超现实世界、水母四处游动、美丽的瀑布、神秘的湖泊、巍峨的高山”。

虚拟化身将遵循类似的发展过程。文本生成文本的模型，比如有130亿参数的LLaMA，将运行在边缘终端，为虚拟化身生成自然直观的对话。此外，文本生成图像的模型将为这些虚拟化身生成全新的纹理和服装。未来几年内，图像生成3D和编/解码器模型将能够为人类生成全身虚拟化身，支持远程通信。最终，人们将能够利用语音提示、图像或视频生成逼真、全动画、智能、可量产的类人虚拟化身。

5.5 物联网：运营效率和客户支持

目前，AI已广泛应用于各种物联网垂直领域，包括零售、安全、能源和公共设施、供应链和资产管理。AI依靠近乎实时的数据采集和分析改进决策质量，优化运营效率，并赋能创新以打造差异化竞争优势。通过生成式AI，物联网细分领域将进一步从AI的应用中受益。

以零售业为例，生成式AI可以改善顾客和员工体验。在售货亭或智能购物车旁的导购员可以基于每周特价商品、预算限制和家庭偏好帮助顾客定制带有菜谱的菜单。商店经理可以根据即将发生的事件预测非周期性的促销机会并进行相应准备。如果一个运动队来到其所在的城市，那么商店经理可以利用生成式AI查询粉丝喜爱的商品品牌，并相应地增加库存。另一个用途是参考来自相似社区的商店的优秀案例和成功经验，重新进行店面规划。生成式AI可以利用简单提示帮助商店经理重新排列货架商品，为利润高的产品腾出空间，或者利用附近连锁店的数据，尽可能降低产品缺货情况的发生。

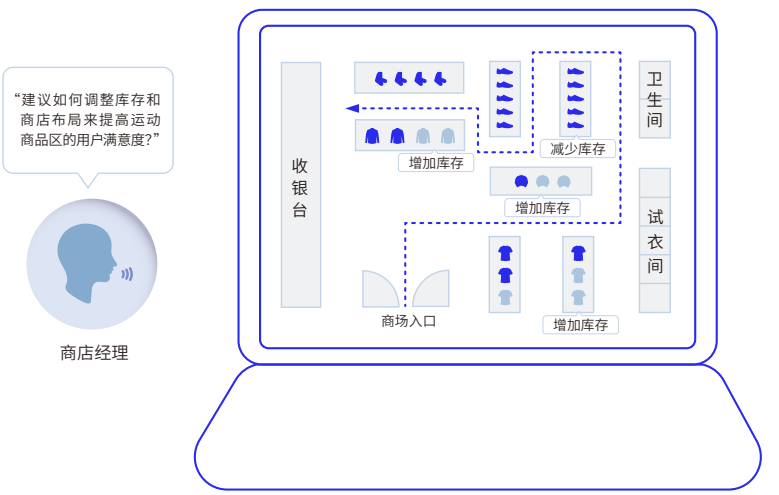


图10：以零售业为例，生成式AI有助于提升顾客和员工体验，比如提供库存和商店布局推荐。

能源和公共设施领域也将受益于生成式AI。运营团队可以创建极端负荷场景并预测电力需求，以及特殊情况下潜在的电网故障，比如农村地区在炎热的夏季出现强风和局部火灾的情况，从而更好地管理资源、避免电力中断。生成式AI也可以用于提供更好的客户服务，比如解答断电或账单计费问题。

6. 总结

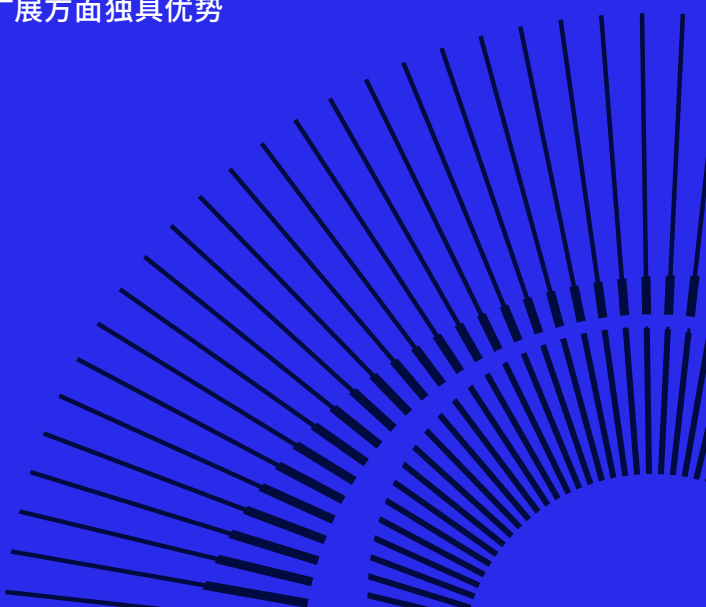
混合AI势不可挡。生成式AI用例将持续演进并成为主流体验，云端和其基础设施需求将不断增加。凭借终端侧AI的先进能力，混合AI架构将规模化扩展，以满足企业和消费者的需求，带来成本、能耗、性能、隐私、安全和个性化的优势。云端和终端将协同工作，依托强大、高效且高度优化的AI能力打造下一代用户体验。

该部分信息发布于：2023年5月

与高通一起
让AI
人人可享

高通AI白皮书 第三部分

高通在推动混合AI规模化扩展方面独具优势
Qualcomm is uniquely positioned
to scale hybrid AI



第三部分 PART THREE

高通在推动混合AI规模化扩展方面独具优势

Qualcomm is uniquely positioned
to scale hybrid AI

• 1. 摘要

正如白皮书第二部分所言，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。

Qualcomm

• 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。

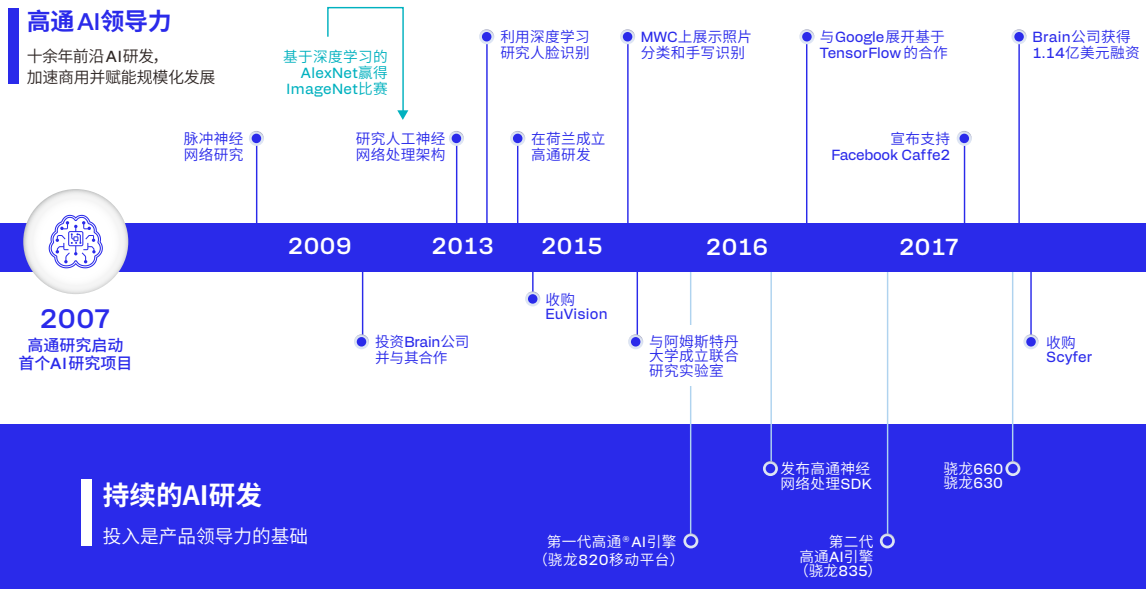


图1：高通持续的AI研发投入是产品领导力的基础。

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

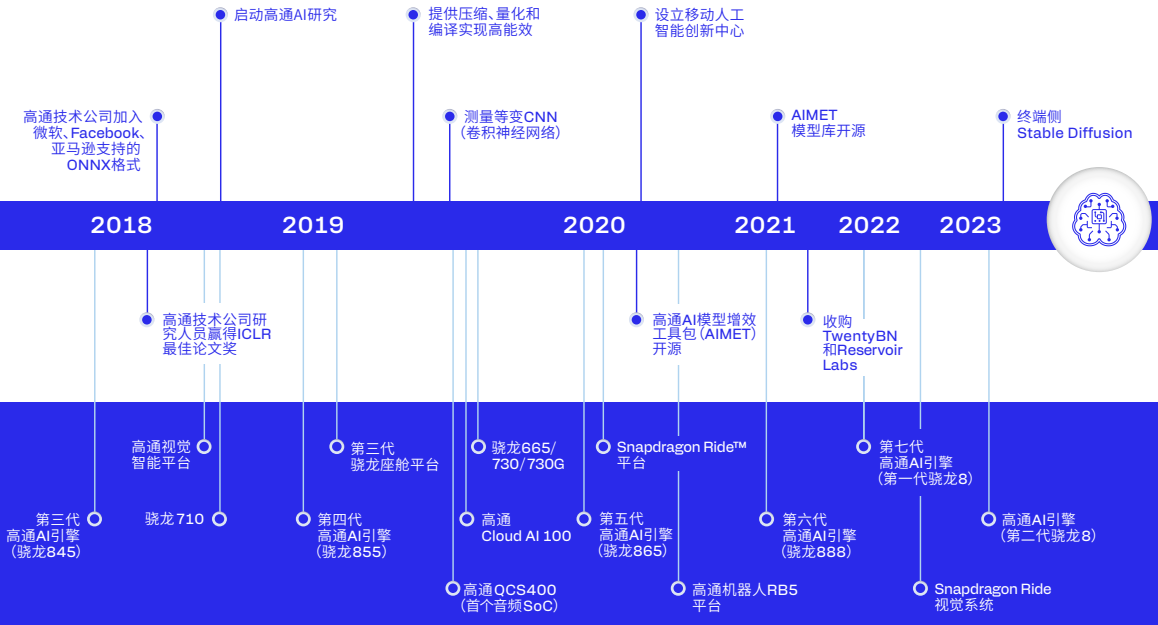
骁龙和高通品牌产品是高通技术公司和/或其子公司的产品。

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。



[2] 高通AI研究是高通技术公司的机构。

Qualcomm

• 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。

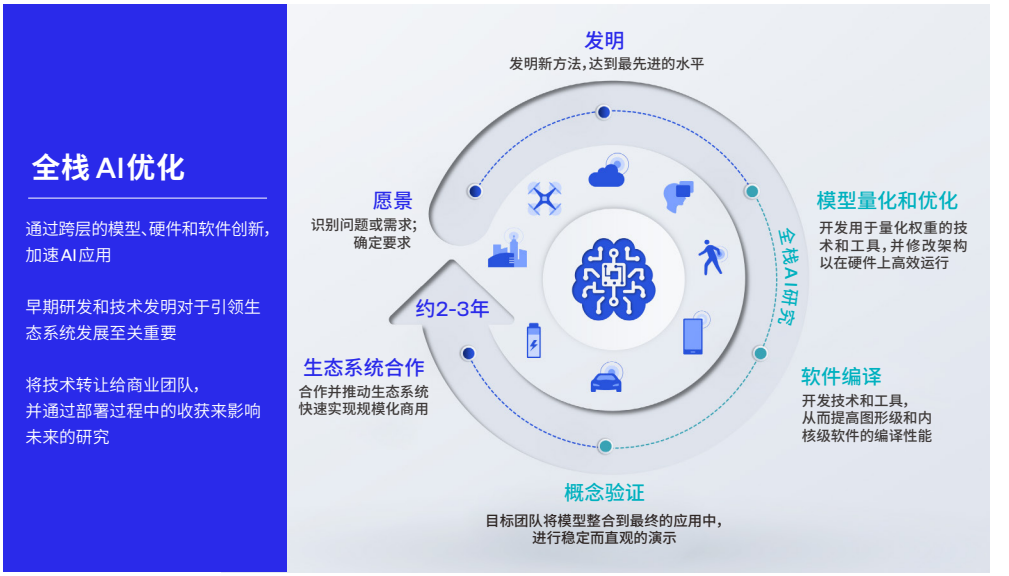


图2：高通全栈AI研究和优化赋能技术持续改进并引领高能效解决方案发展。

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]

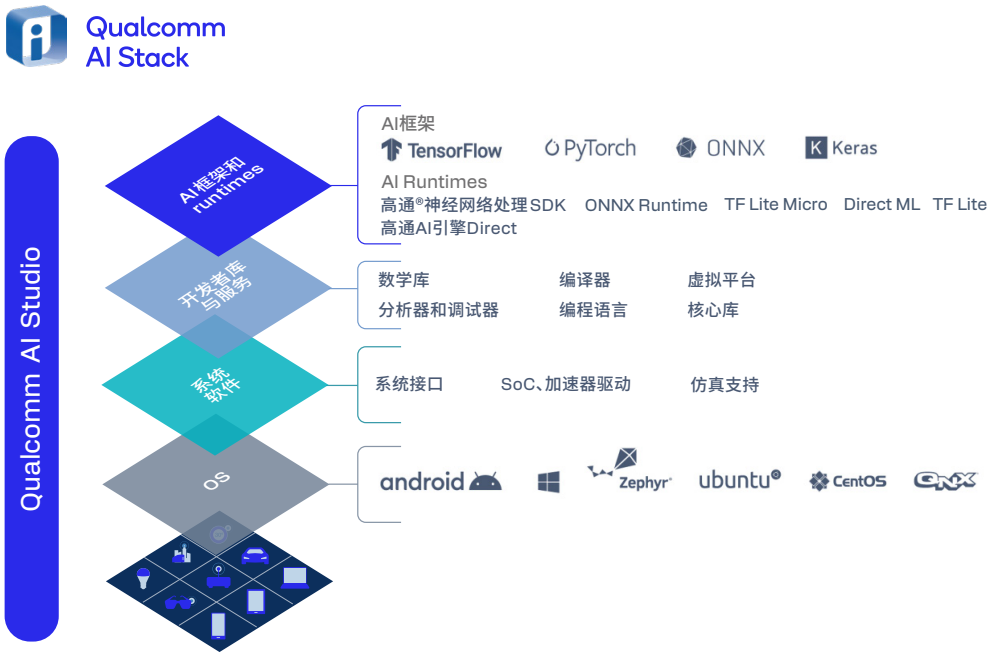


图3：高通AI软件栈旨在帮助开发者一次编写、随处运行，实现规模化部署。

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。

Qualcomm
AI Stack

统一的软件栈
赋能人工智能新时代

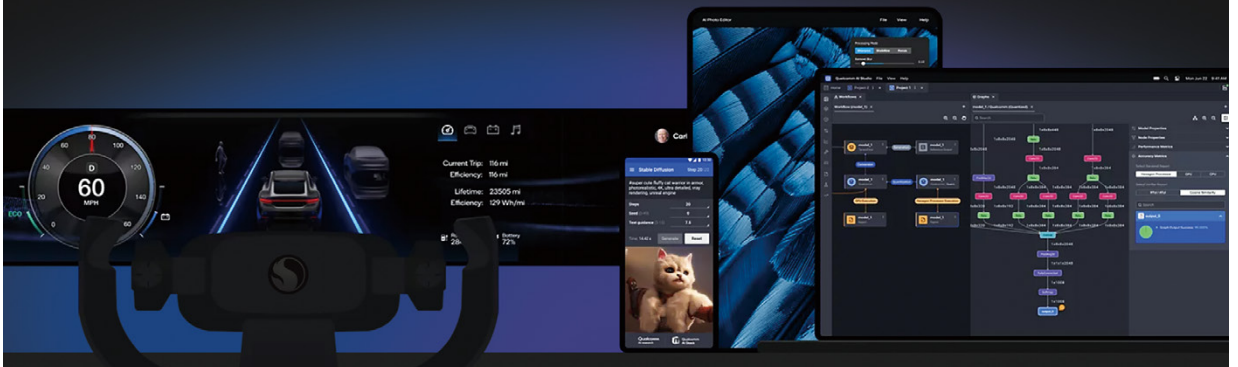


图4：高通AI研究采用整体AI模型效率研究方法。

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。

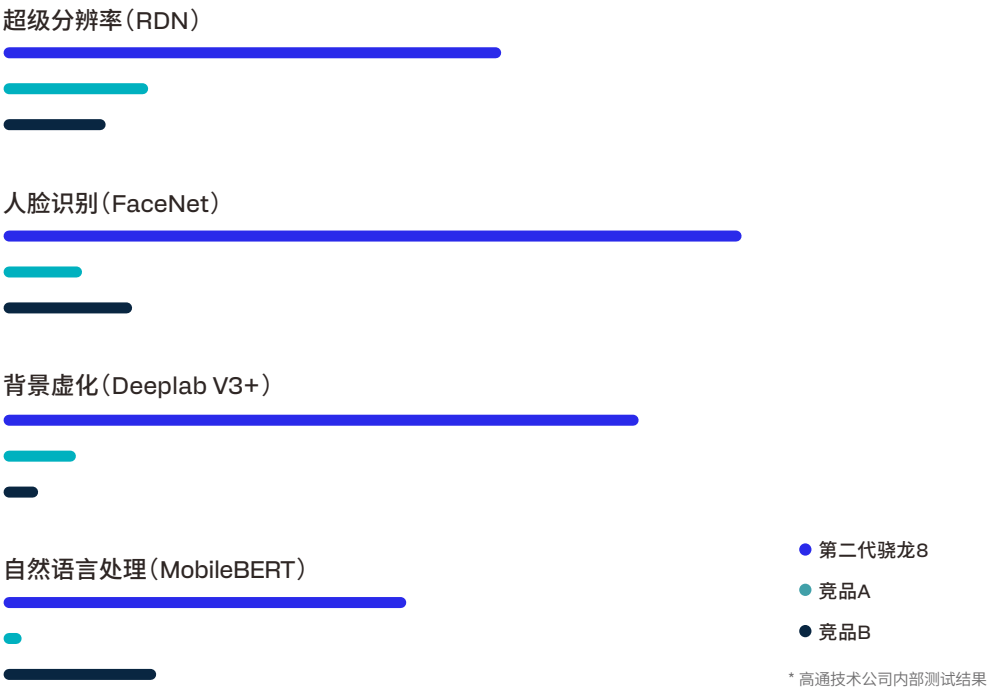


图 5：与移动领域竞品相比，第二代骁龙8提供领先的AI能效。

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。

Qualcomm AI Stack

• 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

Qualcomm

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。

该部分信息发布于：2023年5月

Qualcomm 高通

Qualcomm

Engineering Human Progress

请关注我们：

了解更多信息
请扫描二维码