Qualcomm

2024 年 3 月

通过 NPU 和异构计算
开启终端侧生成式 AI

目录

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。

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。

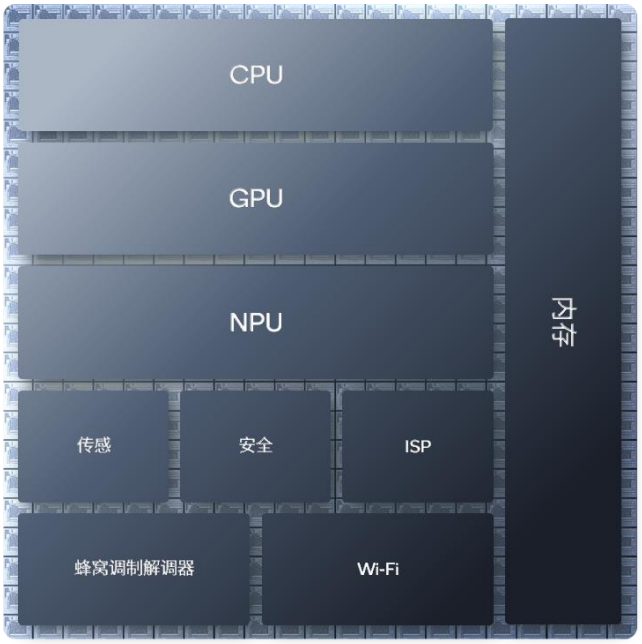

图 1：现代 SoC 在单个 DIE 中集成多个处理器以改善峰值性能、能效、单位面积性能、工业设计和成本。

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. 按需型用例由用户触发，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. 泛在型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。

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。

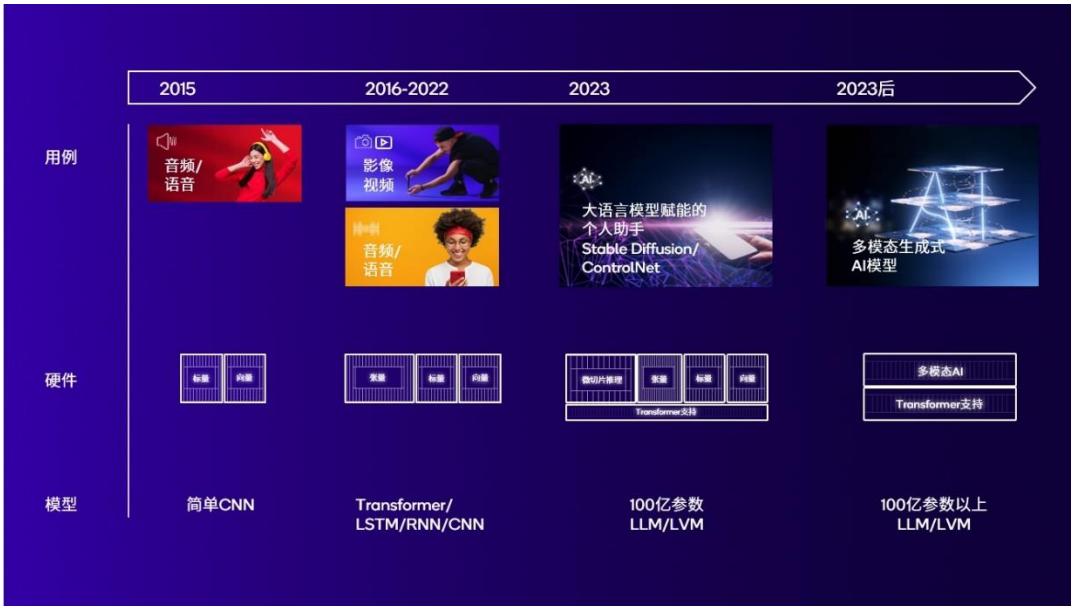

图 2：NPU 随着不断变化的 AI 用例和模型持续演进，实现高性能低功耗。

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。

[1] 与前代平台相比。

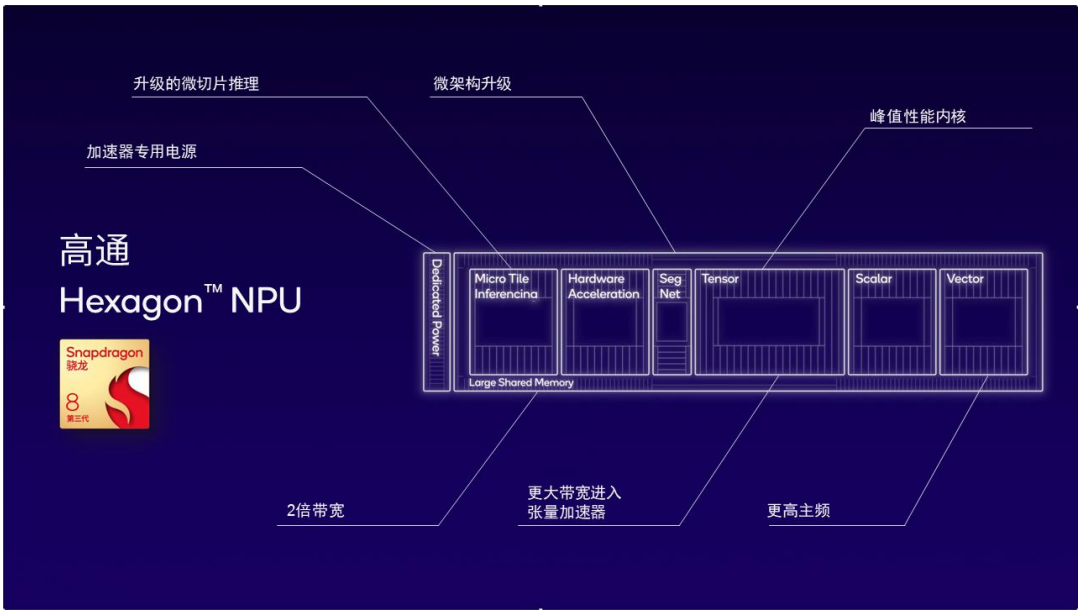

图 4：第三代骁龙 8 的 Hexagon NPU 升级以低功耗实现领先的生成式 AI 性能。

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。

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。

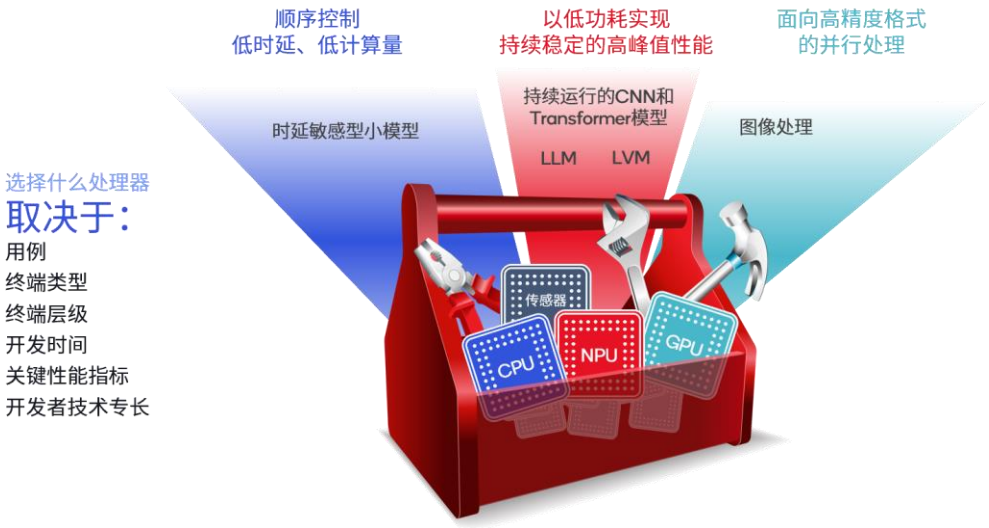

图 5：正如在工具箱中选择合适的工具一样，选择合适的处理器取决于诸多因素。

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。

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。

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。

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。

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

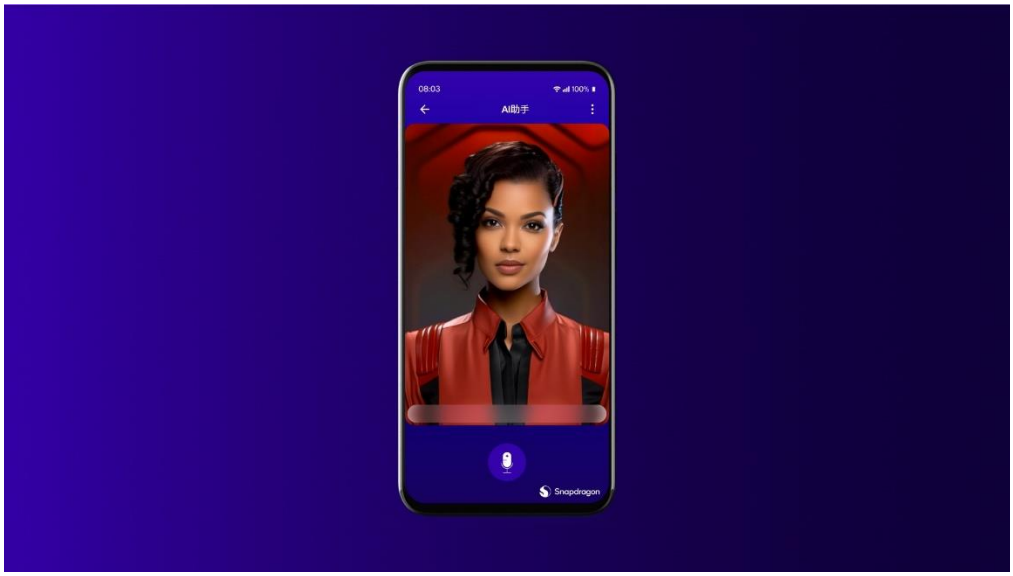

图 7：虚拟化身 AI 助手包括众多复杂工作负载。

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。

[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。

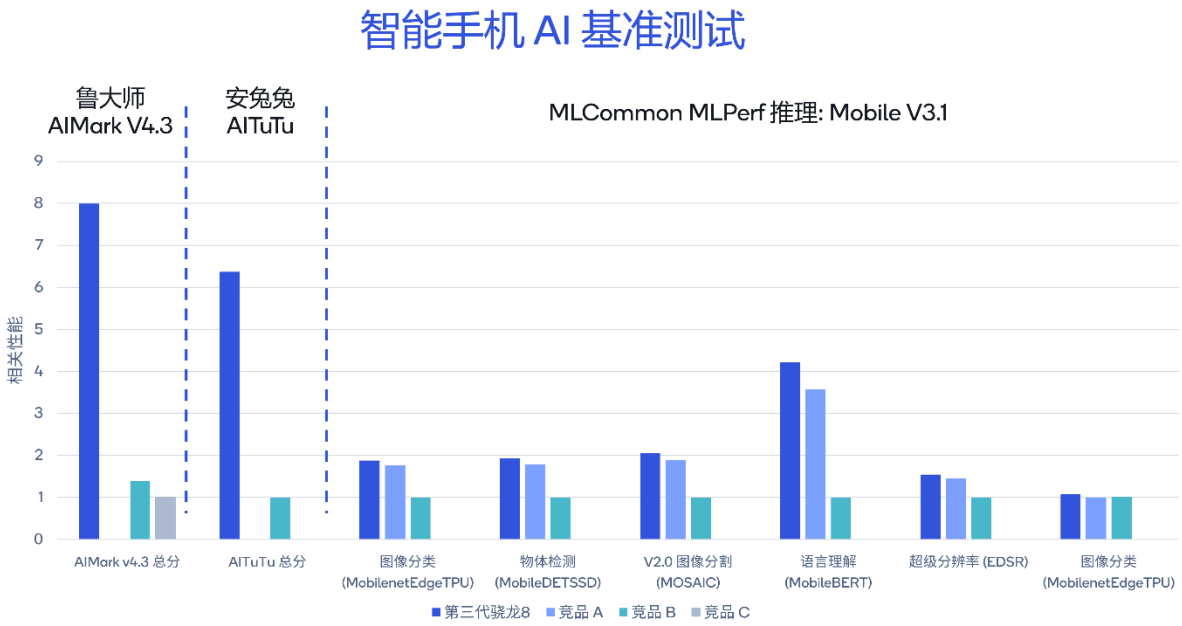

图 9：第三代骁龙 8 在 AIMark、AITuTu 和 MLPerf 中具有领先的智能手机 AI 性能。

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。

[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。

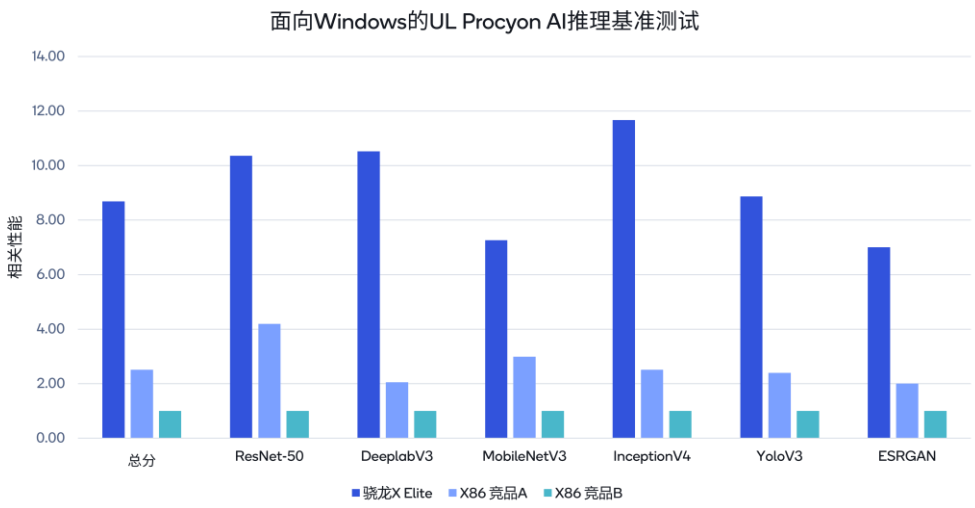

图 10：骁龙 X Elite 在 Procyon 基准测试中具有领先的笔记本电脑 AI 性能。

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。

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。

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。

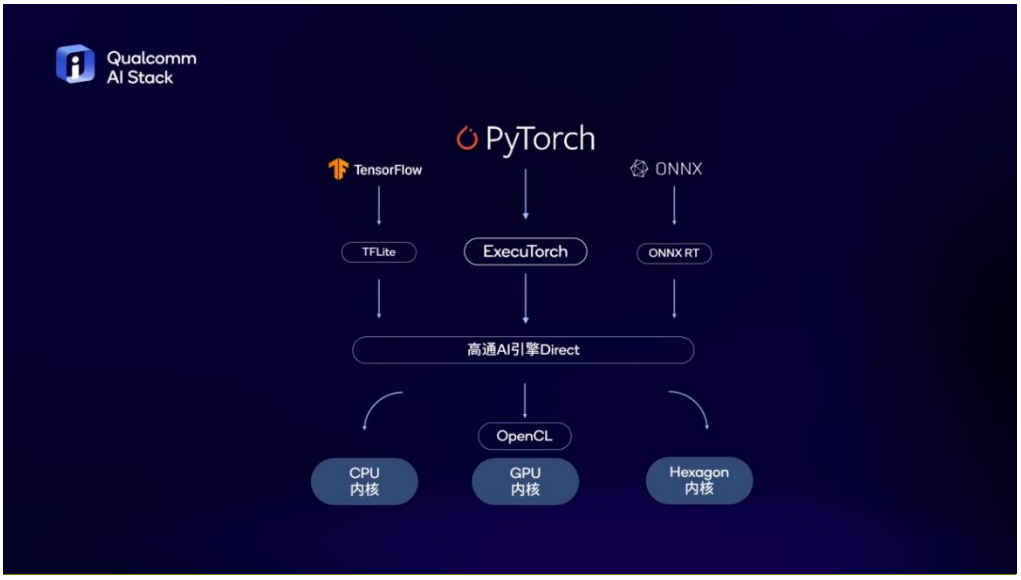

图 12：高通 AI 软件栈支持关键框架和 runtime。

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。**

[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。

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。

欲了解更多相关内容

欢迎订阅《未来 AI 和计算技术》简讯

欲了解更多信息，请访问
qualcomm.cn