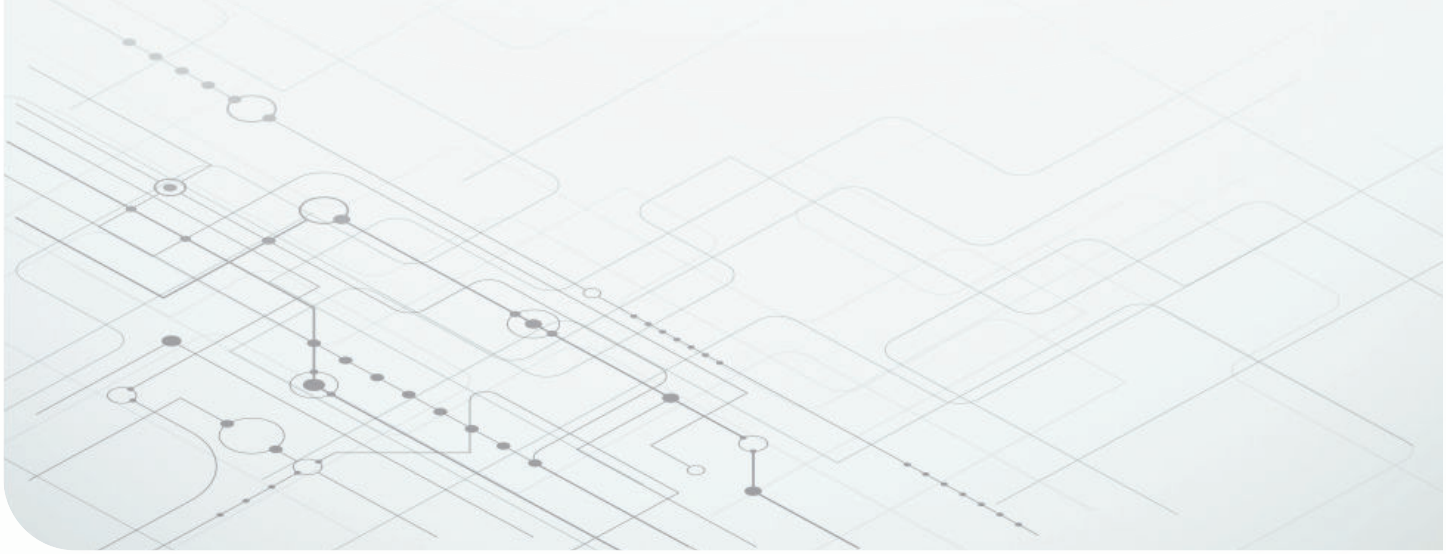
使用高度集成的处理器设计高效的边缘 AI 系统

Texas Instruments

Manisha Agrawal
Product Marketing
Jacinto™ Processors



自动化正在日益普及，现已几乎无处不在。

内容概览

概览本白皮书介绍了构建高效边缘人工智能 (AI) 系统的要求，以及视觉 AI 处理器如何通过异构架构和可扩展 AI 性能优化性能。

引言

当消费者在线订购产品时，自动化技术能够提高整个流程（从制造原材料、提高仓库生产力到快速送货上门）的效率，有时产品只需数小时即可送到消费者手上。要继续推动自动化技术取得这些显著进步，需要更好的机器感知和智能，并减少错误率，而这可以通过将人工智能 (AI) 引入边缘设备来实现。

要创造出更快、更智能和更精确的系统，我们需要从更多传感器收集更多的数据，并提高处理能力。但是，获取更多的数据和进行更多的计算会对系统的性能、功耗和成本要求方面产生挑战。因此，为了优化系统并缩短开发周期，我们需要采用实际可行的方法来设计边缘 AI 系统。

定义边缘 AI

边缘 AI 是指将 AI 算法放在本地设备而非云端处理，这项技术正在为深度神经网络 (DNN) 作为主要算法组件的工业和汽车应用带来新的可能性。为了在尺寸、功耗、散热和成本受限的环境下高效运行，边缘 AI 应用需要高速和低功耗处理，以及特定于应用和其任务的高级集成。**图 1** 展示了一些可使用边缘 AI 处理来提高性能和效率的应用。例如，使用视觉输入的边缘 AI 系统可以通过单个摄像头在生产线上实现质量控制，或通过多个摄像头帮助在汽车或移动机器人中实现功能安全。

***图 1.** 边缘智能存在于许多不同的应用中*

边缘 AI 系统可以帮助提高仓库和工厂的效率，增加城市、建筑和农业的安全性和效率，并让家庭和零售环境智能化。以下是几个需要高效边缘 AI 处理的系统示例：

1. **高级驾驶辅助系统 (ADAS)：**ADAS 技术可以提供车辆周围环境的信息，使驾驶更加便利、轻松和安全。大多数 ADAS 功能都基于视觉系统，从多个摄像头传感器获取高分辨率输入，并使用深度学习和计算机视觉算法来解读这些图像。
2. **自主移动机器人和无人机：**要打造商业机器人，片上系统 (SoC) 必须能够以高速、低功耗的方式处理复杂的感知和导航栈，并具有优化的系统成本。为了尽可能低提高系统效率，SoC 还必须卸载各种计算密集型任务，比如图像去扭曲、立体深度估算、缩放、图像锥体生成和深度学习等。
3. **智能购物车：**大多数智能购物车都配有多个视觉传感器，可以通过摄像头和计算机视觉自动检测商品。智能购物车可以在商品放入购物车时计算订单总价，向消费者推荐购物清单项目，并允许消费者在购物车上为商品买单，从而为客户提供更加个性化的购物体验，并避免排队结算的麻烦。

4. **边缘 AI 盒：**边缘 AI 盒是零售自动化、工厂监控和楼宇监控系统中所用摄像头系统的智能扩展。尽管存在尺寸限制以及功耗和散热方面的挑战，但高吞吐量的 AI 技术使得该盒子能够面向更多摄像头执行智能处理。
5. **机器视觉摄像机：**机器视觉摄像机用于光学字符识别、物体识别、缺陷检测和机械臂引导，利用嵌入式 AI 技术进一步简化产品开发并提高系统准确性。

表 1 列出了各种应用的系统要求。

	ADAS	机器人	智能零售	机器视觉	边缘 AI 盒
深度学习加速器	x	x	x	x	x
多摄像头图像信号处理 (ISP)	x	x	x	x	x
视觉加速器	x	x	x	x	x
深度和运动加速器	x	x	x	x	x
以太网交换机	x	x			x
外围组件快速互连 (PCIe) 开关	x	x			
功能安全	x	x			

***表 1.** 边缘 AI 系统的关键处理和组件要求*

什么是高效的边缘 AI 系统？

在高效的边缘 AI 系统中，DNN 无法自行运行。高效的 AI 系统需要复杂的视觉流水线，通常包括单摄像头或多摄像头图像处理、传统计算机视觉，甚至可能包括多个 DNN。一些应用还可能需要视频编码器和解码器。为了处理所有这些输入，系统需要高性能计算。此外，系统可能需要增强的安全性和功能安全，因此系统复杂性和成本会有所增加。

高效的边缘 AI 系统应针对以下方面进行优化：

- **性能：**嵌入式处理器必须能够提供系统所需的速度、延迟和精度，同时即使在恶劣的环境中，也能可靠地运行。
- **设计限制：**嵌入式处理器必须能够在具有功率和散热限制的设计中运行，包括无风扇设计、被动冷却设计，或者需要依靠电池供电运行更长时间的设计。另外，为符合物理限制条件，处理器还必须满足尺寸和重量规格要求。
- **成本：**实现高性能且具有成本效益的处理，从而尽可能地降低物料清单 (BOM) 成本。

要构建高效的边缘 AI 系统，设计人员应考虑哪种架构和内核最适合完成系统所需的任务。

选择 SoC 架构

嵌入式处理器设计有两种选择，即同构架构和异构架构，通常会采用专门的处理功能来处理某些任务。您应该根据所需的内核类型评估哪种架构最符合您的边缘 AI 系统的需求。

边缘 AI 系统的目标是在最合适的内核上运行 AI、视觉、视频和其他任务，从而优化系统的性能功耗比和每秒 TOPS 的性能，以及成本、尺寸和重量。对于边缘 AI 系统，异构架构具有适合特定任务的正确内核至关重要。

并非所有具有异构架构的处理器都采用一样的设计。器件供应商必须选择合适的处理功能或工艺，并决定是在硬件中加速这些功能，还是将其设计为可配置或可编程。同时，他们还必须注意将内核集成到系统中的方式。总线架构和存储器子系统必须能够在内核之间高效地传输数据。

如果 SoC 存在内核类型不正确而无法加速任务，内核数量过多而无法高效地管理，或者总线基础结构和存储器子系统效率低下，基于视觉的边缘 AI 系统可能会效率低下。

可编程内核类型和加速器

下面我们来了解一下边缘 AI 系统中可能的内核类型：

CPU

中央处理单元 (CPU) 是可处理连续工作负载的通用处理单元。它们具有很高的编程灵活性，并可从庞大的现有代码库中受益。通常，大多数边缘 AI 系统具有 2 到 8 个 CPU 内核，用于管理平台和功能丰富的应用。但是，仅含 CPU 的系统不适合像素级成像、计算机视觉和卷积神经网络 (CNN) 处理等高度专业化的任务。CPU 还具有高功耗，但吞吐量却是不同内核类型中最低的。单核 CPU 系统与 AI 加速、图像处理等专用硬件模块配合使用，可以满足低成本应用的功率预算要求。

GPU

图形处理单元 (GPU) 具有数百到数千个小型内核，非常适合并行处理任务。GPU 最初设计用于实现一系列图形操作，但现在已经广泛应用于深度学习应用中，尤其是在训练 DNN 时特别有用。然而，由于内核数量众多，GPU 功耗很大并具有更高的片上存储器要求，这是其主要缺点之一。

DSP

数字信号处理器 (DSP) 是高能效的专用内核，通常用于解决多个复杂的数学问题。DSP 能够以低功耗处理来自现实世界中视觉、音频、语音、雷达和声纳传感器的实时数据，并有助于更大程度地提高每个时钟周期的处理能力。然而，由于其编程难度较大，需要熟悉 DSP 硬件的特性、编程环境和 DSP 软件优化，才能实现最佳性能。

ASIC

专用集成电路 (ASIC) 和加速器能够为系统应用提供最高的性能和最低的功耗。当您确定要加速的功能所属的核心内核时，通常会使用它们。例如，CNN 的核心计算通常涉及矩阵乘法。对于传统的计算机视觉任务，专用硬件加速器能够计算图像缩放、镜头失真校正和噪声滤波等操作。

FPGA

现场可编程门阵列 (FPGA) 是一类集成电路，可以对硬件模块进行重新编程并将其用于特定应用。它们的功耗低于 GPU 和 CPU，但高于 ASIC。不过，硬件编程比较难，并且需要掌握硬件描述符语言方面的专业知识，比如 Verilog 或超高速 IC 硬件描述语言等。

使用 TI 视觉处理器设计边缘 AI 系统

TI 的视觉处理器产品系列旨在为尺寸和功率限制是关键设计挑战的应用提供高效、可扩展的 AI 处理。

这些处理器包括 AM6xA 和 TDA4 处理器系列，其采用的 SoC 架构为视觉系统实现了广泛集成，其中包括 Arm® Cortex®-A72 或 Cortex-A53 CPU、内部存储器、多种接口和硬件加速器，可提供每秒 2 万亿次至 32 万亿次运算 (TOPS) 的深度学习 AI 处理能力。

AM6xA 系列使用 Arm Cortex-A MPU 将深度学习推理、成像、视觉、视频和图形处理等计算密集型任务卸载到专用硬件加速器和可编程内核，如**图 2** 所示。通过将高级系统组件集成到这些处理器中，有助于边缘 AI 设计人员简化系统物料清单。该处理器产品系列包括可扩展的处理选项，如适用于具有 1 至 2 个摄像头的低功耗应用的 AM62A 处理器和 AM68A（多达 8 个摄像头）、AM69A（多达 12 个摄像头）处理器。

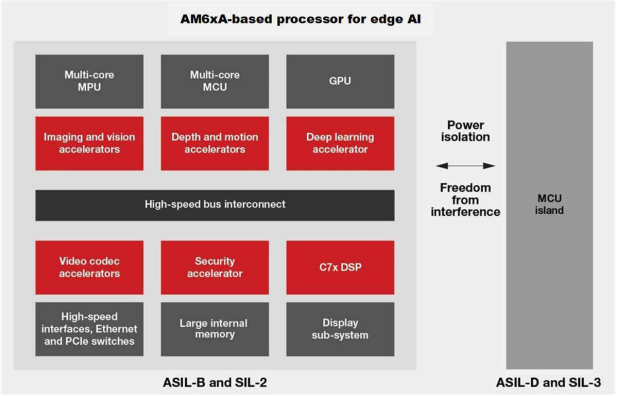


***图 2.** TI 视觉处理器边缘 AI 系统分区*

深度学习加速器

虽然 CPU 和 GPU 适用于其他任务，但它们并不是最适合加速深度学习任务的内核。CPU 的吞吐量有限且功耗高，而 GPU 则是所有内核中功耗最大的，并且内存占用量大。

TI 视觉 AI 处理器集成了一个深度学习加速器，其中包含 ASIC 中的矩阵乘法加速器 (MMA)，并固定在可编程 C71 DSP 上。MMA 支持高性能（每个周期可以进行 4K 8 位固定乘法累加）和低功耗传感器加速，而 C71 DSP 负责加速矢量和标量运算并管理 MMA。

由于将 MMA 和 C71 DSP 结合起来，该加速器能够提供非常出色的性能（每秒推理次数）和能效（每瓦推理次数）。C71 内核的编程灵活性可以满足边缘 AI 创新的需求。当不用于深度学习时，该内核能以低功耗处理其他计算密集型任务。

智能存储器架构实现了加速器的高效利用。该加速器配置了内置的存储器子系统，用于数据传输的专用 4D 可编程直接存储器存取 (DMA) 引擎，以及专用的流硬件。这些流硬件可以将数据直接从外部存储器传输到 C71 内核和

MMA 的功能单元，而绕过高速缓存。平铺和超平铺功能可更大限度地减少与外部存储器之间的数据传输。

表 2 显示了 AM68A 和集成 8TOPS 加速器的 TDA4VM 上的 8 位固定推理性能。报告的性能采用批次 1 和单个 32 位 LPDD4。

网络	图像分辨率	每秒帧数 (fps)
MobileNet_v1	224 × 224	741
Resnet-50 V1.5	224 × 224	162
SSD-MobileNets-V1	300 × 300	385

***表 2.** MLPerf 推荐模型的推理基准测试*

免责声明： TI 使用 MLPerf 建议的模型和准则进行了边缘 AI 推理基准测试。TI 尚未将结果提交给 MLcommons 组织。

成像和计算机视觉硬件加速器

基于视觉的边缘 AI 系统通常包括单摄像头或多摄像头图像处理和传统的计算机视觉任务。在 CPU 或 GPU 中执行这些任务会产生大量功耗，并且存在吞吐量限制。

此类边缘 AI 处理器 SoC 可以在硬件上加速计算密集型低级别强力像素处理视觉任务，例如视觉处理加速器内核中的 ISP、镜头失真校正、多尺度缩放和双边噪声滤波。深度和运动感知加速器内核可以加速立体深度估算和密集光流，有助于增强对环境的感知，如**图 3** 所示。

***图 3.** 视觉加速器功能*

通过在硬件上加速这些任务，可以实现低功耗和小尺寸。尽管这些任务是在硬件中加速的，但其可配置性仍然提供了灵活性，从而可以使用加速器功能来更好地满足您的系统需求。

这种集成和加速可以省去对定制 ISP 或 FPGA 的需求，同时还能释放 CPU 的性能，用于在硬件中处理计算密集型成像和视觉任务。例如，单个视觉处理加速器内核能够以 30fps 的速度处理多达 8 个 200 万像素或 2 个 800 万像素摄像头的数据。深度和运动处理加速内核能够以每秒 8000 万像素的速度进行立体深度估算，并以每秒 15000 万像素的速度处理运动矢量。

智能内部总线和存储器架构

通过监控数据移动和处理器的存储器架构，防止在同时运行多个内核时出现各种内核阻塞和延迟，有助于更大限度地提高整体系统性能。

TI 视觉 AI 处理器采用高带宽总线互连，具有非阻塞基础结构和大型内部存储器。多个专用可编程 DMA 引擎以极高的速度自动执行数据移动。此设计实现了硬件加速器的高利用率，并显著节省了双倍数据速率 (DDR) 带宽。通过减少 DDR 实例的数量，可以降低 DDR 存取消耗的功率，从而降低总体系统功耗。

优化的系统 BOM

下面我们来了解一下 TI 视觉 SoC 中的高级集成系统组件和功能，这些组件和功能可以为多种类型的边缘 AI 应用降低系统 BOM 成本：

- **ISP：**集成的 ISP 内核消除了对外部 ISP 或 FPGA 设计的需求。机器视觉、智能购物车、机器人和 ADAS 等所有单摄像头和多摄像头 AI 应用都可以从这种集成中受益。
- **安全：**符合汽车安全完整性等级 (ASIL) D 和 SIL 3 的集成式安全微控制器 (MCU) 具有 Cortex-R5 内核，无需外部安全 MCU 即可帮助实现安全目标。由于其余的处理也符合 ASIL B/SIL 2 标准，这种架构支持 ADAS、机器人、建筑和农业电子控制单元应用。
- **以太网和 PCIe 开关：**集成的以太网和 PCIe 开关消除了对外部交换机组件的需求。
- **安全性：**集成的安全加速器提供先进的安全支持。
- **DDR 存储器：**内联纠错码保护以及与典型存储器架构相比更少的 DDR 存储器实例（通过智能存储器）可以节省成本。

易于使用的软件开发环境

TI 提供了综合的软件环境（如**图 4** 中所示），让您无需学习 TI 硬件或专有软件，就能够采用异构架构并充分发挥器件性能的潜力。通过生产品质驱动程序对硬件加速器进行抽象，同时还使用业界通用的应用程序编程接口 (API) 为

MPU 上的高级操作系统提供应用程序开发接口，从而加快软件开发。TI 的底层软件能够自动将成像、视觉、深度学习和多媒体任务分配到正确的硬件加速器上，从而简化高性能应用程序编程。

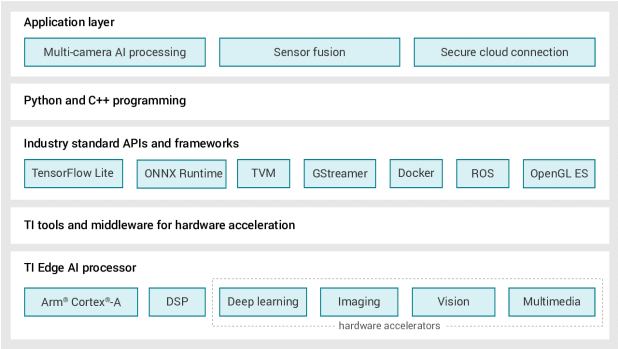


图 4. *适用于边缘 AI 应用的软件开发环境*

结论

越来越多的应用采用异构架构。TI 的视觉 AI 处理器具有加速的深度学习、视觉和视频处理，专用系统集成和高级组件集成功能，可使商业上可行的边缘 AI 系统针对性能、功耗、尺寸、重量和成本进行优化。TI 的边缘 AI 软件开发环境围绕业界通用的开源 API 构建，具有硬件加速器的自动加速功能，可加快边缘 AI 应用的开发。

AI 技术正在快速发展，为边缘 AI 应用的各个方面带来了创新。这项技术正在推动应用不断突破对更高计算能力的需求。当通过实施嵌入式处理器以更低的功耗和更低的系统成本实现时，边缘 AI 可以为嵌入式应用带来更广阔的前景。

重要声明和免责声明

TI“按原样”提供技术和可靠性数据（包括数据表）、设计资源（包括参考设计）、应用或其他设计建议、网络工具、安全信息和其他资源，不保证没有瑕疵且不做出任何明示或暗示的担保，包括但不限于对适销性、某特定用途方面的适用性或不侵犯任何第三方知识产权的暗示担保。

这些资源可供使用 TI 产品进行设计的熟练开发人员使用。您将自行承担以下全部责任：(1) 针对您的应用选择合适的 TI 产品，(2) 设计、验证并测试您的应用，(3) 确保您的应用满足相应标准以及任何其他功能安全、信息安全、监管或其他要求。

这些资源如有变更，恕不另行通知。TI 授权您仅可将这些资源用于研发本资源所述的 TI 产品的应用。严禁对这些资源进行其他复制或展示。您无权使用任何其他 TI 知识产权或任何第三方知识产权。您应全额赔偿因在这些资源的使用中对 TI 及其代表造成的任何索赔、损害、成本、损失和债务，TI 对此概不负责。

TI 提供的产品受 TI 的销售条款或 ti.com 上其他适用条款/TI 产品随附的其他适用条款的约束。TI 提供这些资源并不会扩展或以其他方式更改 TI 针对 TI 产品发布的适用的担保或担保免责声明。

TI 反对并拒绝您可能提出的任何其他或不同的条款。

邮寄地址：Texas Instruments, Post Office Box 655303, Dallas, Texas 75265