白皮书

在英飞凌 ModusToolbox™ 环境中使用 Arm® Ethos™-U55 NPU 实现机器学习应用

摘要

机器学习 (ML) 模型需要强大的计算资源来进行训练和推理，因此，它们通常在可以进行大算力数据处理的 PC 或云服务器上运行。然而，在计算机架构的革命性发展和软件工具的突破性进步的引领下，嵌入式系统 AI 和 ML 应用正在经历转型期。ML 应用和场景正迅速扩展到物联网和嵌入式系统领域。当视频和图像等数据利用深度学习 ML 模型时，这些应用需要高处理能力的运算单元和大量的内存。为了支持这些场景，有必要使用诸如 Arm® Ethos™-U55 等神经网络协处理器 (NPU) 来增强处理器的性能。

能耗效率和低成本是嵌入式 ML 应用的关键标准。除增强处理能力之外，还需要降低系统功耗和提供高效的软件开发环境。开发环境涵盖了用于模型开发、训练和部署的工具和优化器。

英飞凌的 PSoC™ Edge 平台及其 ModusToolbox™ 软件开发环境可以更好地利用平台硬件资源进行 CPU 密集型嵌入式 ML 推理。

目录

1 PSoC™ Edge MCU

PSoC™ Edge MCU 是一个高性能、低功耗 MCU 产品系列，这些产品专为提升计算性能、人机界面 (HMI)、机器学习 (ML)、增强传感、实时控制和低功耗应用而设计。

这个产品系列中的双 CPU 微控制器都具备神经网络协处理器、DSP 功能，并支持高性能内存扩展能力 (OSPI)、可进行高性能模数转换的低功耗模拟子系统和低功耗比较器等。这些产品还提供了物联网连接、通信接口、可编程模拟和数字模块，以及音频和图形模块等。

ModusToolbox™ 开发环境包括可安装的 SDK 和库、Arm®行业标准工具、RTOS 支持以及性能稳健且易于使用的 ML 和 HMI 软件和工具。支持的功能包括安全、通信和控制以及 DSP 功能。多域架构实现了细粒度功耗优化以及动态频率和电压按比例调节。

PSoC™ Edge 的常开域模块支持语音识别、触摸唤醒、电池监测和其他传感应用。提供这些功能仅需极低功耗。

PSoC™ Edge 高性能域系统级架构如图 1 所示。

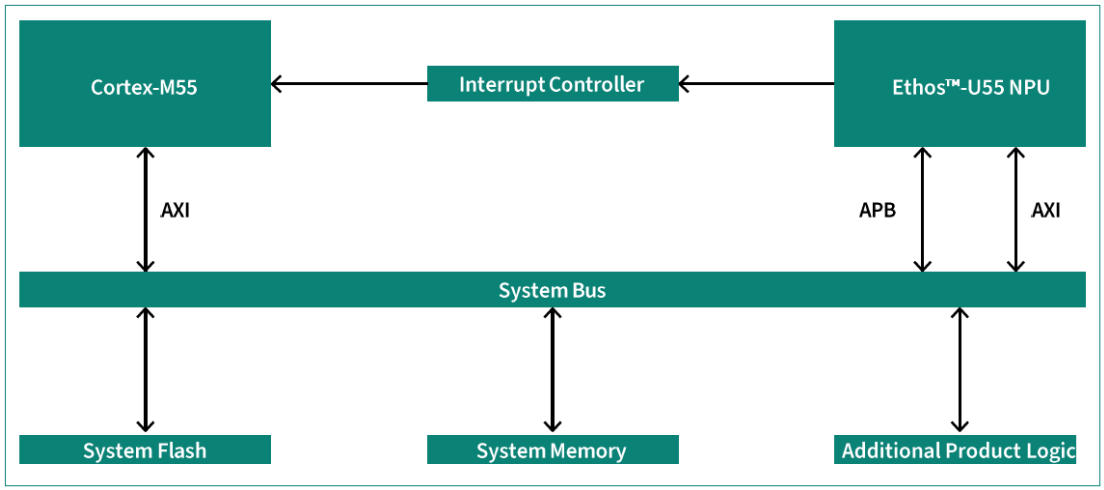

图 1 PSoC™ Edge 高性能域系统架构

2 Ethos™ -U55 NPU

Ethos™-U55 是第一代与 Cortex®-M 处理器兼容的 microNPU。Ethos™-U55 可支持以前不可能实现的新一代 TinyML 应用。使用它，仅需很小芯片面积和很低功耗即可实现神经网络加速推理计算。相比于单独使用 Cortex®-M 处理器，Cortex®-M55 和 Ethos™-U55 组合起来可将 ML 性能提升至最高 400 倍。

凭借其内置 Helium 矢量处理扩展，Cortex®-M55 能够在嵌入式微控制器平台上运行 ML 模型，而通过与 Ethos™-U55 microNPU 集成，它还可以提供比前几代 Cortex®-M 处理器系列更出色的 ML 性能。采用相同软件堆栈的 Ethos™-U55 可将 Cortex®-M55 系统的 ML 性能提高至最多 30 倍。

图 2 所示为 Ethos™-U55 系统架构图

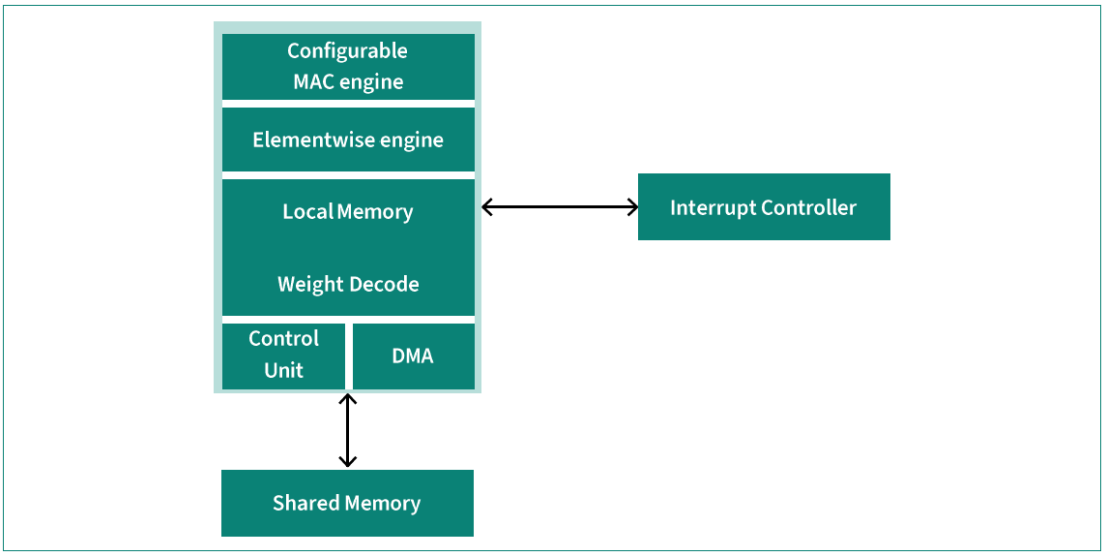

图 2 Ethos™-U55 系统架构

Ethos™-U55 NPU 亮点概览：

- 内存占用空间小且效率最高
- 集成可配置 MAC 引擎（32 位、64 位、128 位、256 位），可满足卷积、深度池

化、矢量乘积以及最大池化所需的操作等要求。PSoC™ Edge 为 Ethos™-U55 配置了 128 位 MAC
- 权重解码器和 DMA 用于动态权重解压缩
- 支持用于模型优化的离线工具
- 支持最常见的 ML 网络操作，包括 CNN 和 RNN，并可灵活支持未来的 ML 创新
- 中央控制 (CC) 执行工作单元队列并调度至 DMA 控制器、权重解码器、MAC 单元和输出单元
- 输出单元支持激活函数，包括：ReLU、ReLU1、ReLU6、Leaky ReLU、tanh、sigmoid 和可配置查找表 (LUT)
- 通过开放的 Q-Channel 从端口支持高级时钟和功率门控

2.1 使用 Ethos™-U55 NPU 的优点

在 PSoC™ Edge 平台中使用 Ethos™-U55 NPU 结合 Cortex®-M55 可实现：
- 更出色的 ML 性能
- 更低功耗
- 节省 Cortex®-M55 CPU 时钟周期用于其他任务，以便实现更复杂的 ML 应用

以关键字识别 (KWS) 场景的 ML 应用为例，展示在系统中使用 Ethos™-U55 NPU 带来的性能提升。

2.1.1 ML 性能提升

图 3 所示为 Ethos™-U55 NPU 结合 Cortex®-M55 CPU 带来的性能提升。数据来自关键字识别 (KWS) 应用测试。

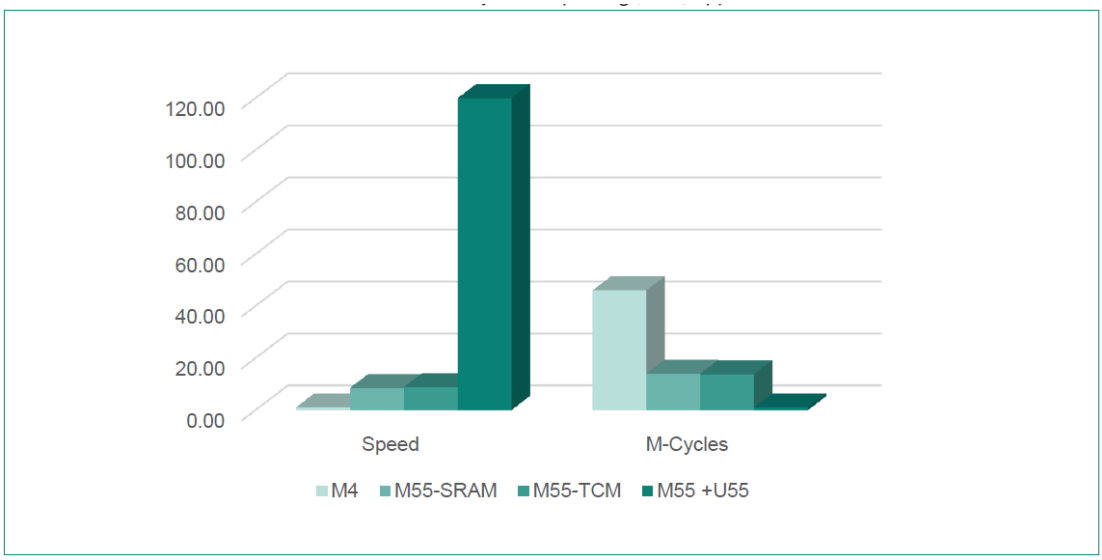

图 3 N55 NPU 的 ML 性能比较

图中比较了单独的 Cortex®-M4、使能 TCM 前后 的 Cortex®-M55 以及 Cortex®-M55 结合 Ethos™-U55 NPU 实现的性能。如图所示，相比于 Cortex®-M4 MCU，具备 Helium 扩展指令集的 Cortex®-M55 可以为 ML 应用提供稍好性能。然而，将 Cortex®-M55 与 Ethos™-U55 NPU 相结合，则可以大幅增强系统的 ML 性能，为开发基于嵌入式平台的复杂的 ML 应用创造了条件。

2.1.2 CPU 负荷率

为系统配置一颗 Ethos™-U55 NPU，可以大大降低 Cortex®-M55 CPU 的负荷率并提升其性能。由此节省的 CPU 资源可以用于其他任务，从而支持在 PSoC™ Edge 平台上实现更复杂的 ML 应用。

图 4 所示为 Cortex®-M55 CPU 在配置 Ethos™-U55 前后的情况下用于关键字识别 (KWS) 应用时的占用率。

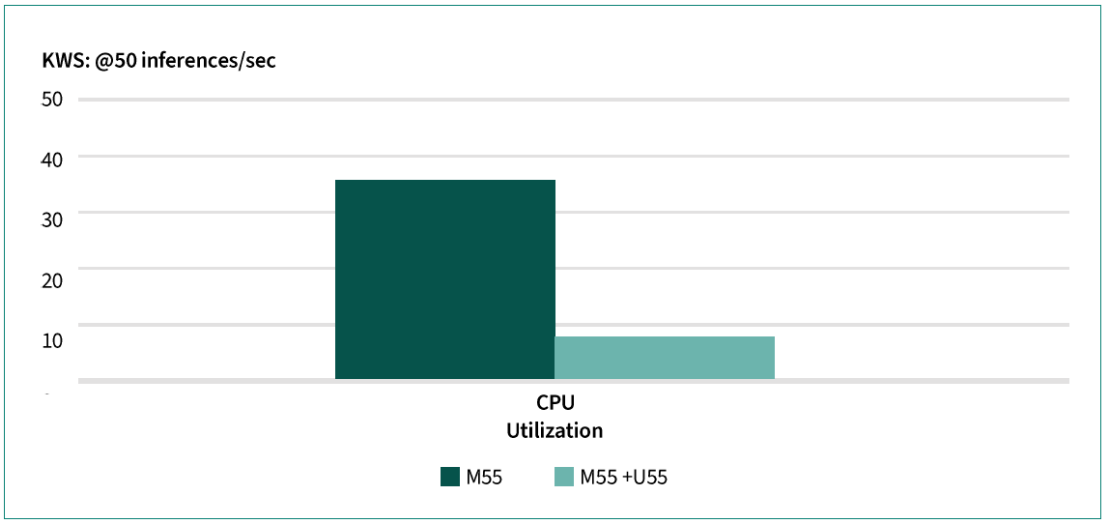

图 4 配置 Ethos™-U55 NPU 前后时的 CPU 负荷率对比

2.1.3 功耗

除提升性能之外，通过配置 Ethos™-U55 NPU 来实现 NN 加速的同时，还可以降低 Cortex®-M55 CPU 负荷率，从而大幅优化功耗。

2.2 Ethos™-U55 NPU 加速流程

这个流程从将要训练或获取待加速的 TensorFlow 模型开始。将模型量化为 8 位整数格式，并转换为标准 TensorFlow Lite 格式。

图 5 所示为通用 ML 模型优化流程：

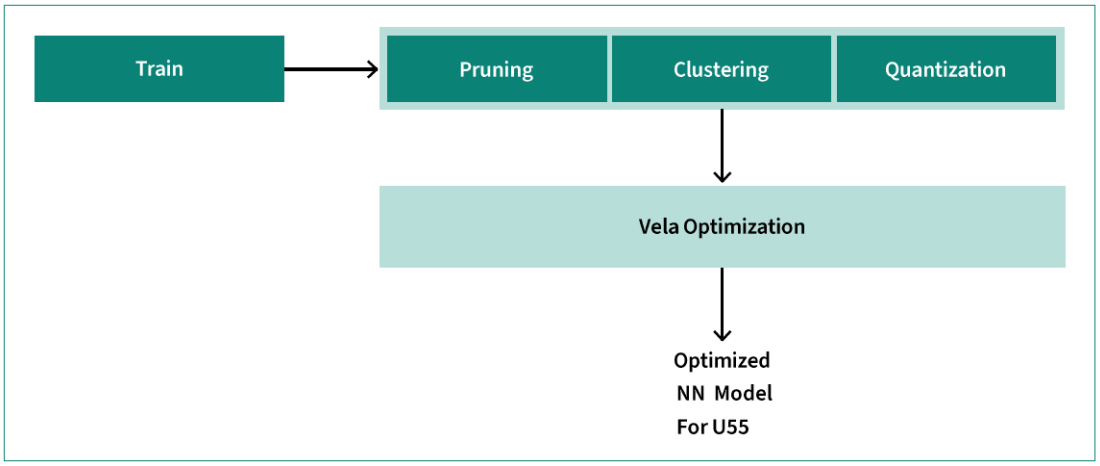

图 5 适用于 Ethos™-U55 的通用 ML 模型优化流程

Arm® 提供的 NN 优化工具（Vela 编译器）读取 TensorFlow Lite 文件作为输入并进行格式转换，以使之准备就绪进行部署。 TensorFlow Lite Micro (TFLM) 运行环境被创建，以便在嵌入式设备的局限范围内执行。在主机上离线创建的 TensorFlow Lite 平面文件（平面缓冲区格式）将被部署到目标设备上。这个平面文件载明了哪些神经网络层在 Ethos™-U55 上执行，哪些神经网络层仍在 Cortex®-M 处理器上执行。Ethos™-U55 支持的神经网络层在它上面实现加速运算，而其余的神经网络层则保留在 Cortex®-M 处理器上执行。如果有相应的内核可供使用，可通过 CMSIS-NN 软件库加速在 Cortex®-M 处理器上执行的神经网络层，否则，就使用 TensorFlow Lite Micro 参考内核。

3 ModusToolbox™

ModusToolbox™ 软件提供了一个现代的可扩展开发环境，可支持各类型英飞凌微控制器，包括 PSoC™ Arm® Cortex® 微控制器以及多种不同的英飞凌连接方案。

ModusToolbox™ 软件提供了一系列开发工具、库和嵌入式运行环境资源，其架构化设计可提供灵活、全面的开发体验。

运行环境软件包括中间件、设备驱动程序和代码示例，通过大量 GitHub 存储库提供。

英飞凌开发者中心提供的安装包，包含了支持 Windows、Linux 和 macOS 的开发工

具。这些桌面应用可用于创建新的嵌入式应用、管理软件组件、配置设备外设和中间件，还包括用于编译、编程和调试的嵌入式开发工具。ModusToolbox™ 开发工具直接与可用的运行时软件库对接，便于轻松获得最新开发资源。

从英飞凌开发者社区，可以轻松访问社区论坛、知识型文章和技术类博客文章。可增强 ModusToolbox™ 开发体验的其他资源包括，开发工具和运行时软件的全面的技术文档、详细的培训和教程视频等。

IDE 选项：
- 适用于 ModusToolbox™ 的 Eclipse IDE
- 微软 Visual Studio Code
- IAR Embedded Workbench
- Arm® µVision®

编译器选项：
- GNU
- IAR
- Arm®

调试选项：
- Segger J-Link
- 英飞凌 MiniProg4
- IAR I-jet
- Arm® ULINK™

3.1 ModusToolbox™ ML

适用于机器学习的 ModusToolbox™ 可满足三类主要的机器学习产品制造商的需求：
- 使用自有模型
- 训练自有模型
- 购买自有模型

借助 ModusToolbox™ 机器学习 (ML)，用户可以在英飞凌 MCU 上快速评估和部署 ML 模型。ModusToolbox™ ML 经专门设计，可与 ModusToolbox™ 生态系统无缝协作，并且可以添加到现有项目中，以便在低功耗边缘设备上执行推理。

适用于机器学习的 ModusToolbox™ 还集成了合作伙伴解决方案，通过

ModusToolbox™ 和 Friends Ecosystem 将英飞凌软件和合作伙伴软件的优势结合起来。合作伙伴提供了易于使用的 AutoML 训练平台，以便开发人员专注于数据和用例，而不是传统 ML 训练过程中繁重的基础设施要求和学习曲线。

合作伙伴解决方案可以实现：

- 采集数据和打标签
- 云端训练基础设施
- 轻松集成
- 针对嵌入式性能进行了优化

3.2 ModusToolbox™ ML 软件架构

图 6 所示为 ModusToolbox ML 环境的各种组件和软件库。

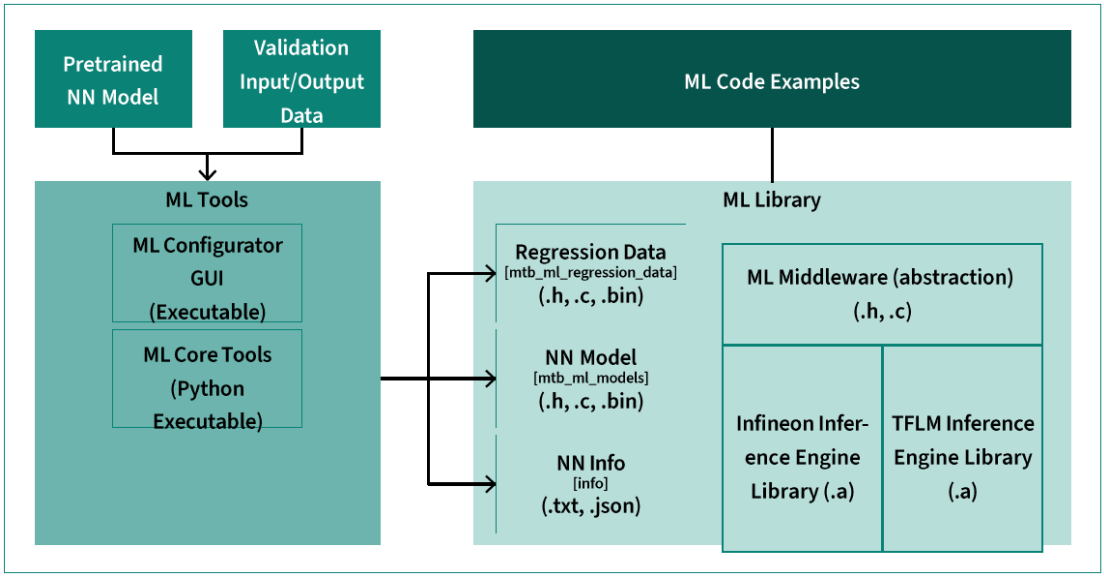

图 6 ModusToolbox™ ML 软件组件

ModusToolbox™ ML 提供的一系列功能，有助于在 PSoC™ 平台无缝进行深度学习模型部署和验证。这些功能包括：

- 适用于微控制器的 TensorFlow Lite，包括解释器推理和无解释器推理
- 支持 .tflite 和 .H5 模型格式
- 常用 NN 内核：MLP、GRU、Conv1d、Conv2d、LSTM
- 辅助 NN 内核：压平层、丢弃层、重塑层、输入层
- 激活函数：relu、softmax、sigmoid、线性、tanh
- 输入数据量化级：

    - 32 位浮点
    - 16/8 位整数
- NN 权重量化级：
    - 32 位浮点
    - 16/8 位整数
- 回归数据评估
- 周期和内存估算
- 基于 PC 的推理引擎
- 基于目标设备的推理引擎（经优化）

3.3 运行时软件流程

模型准备、优化、下载到设备以及在设备上执行的软件流程如图 7 所示。

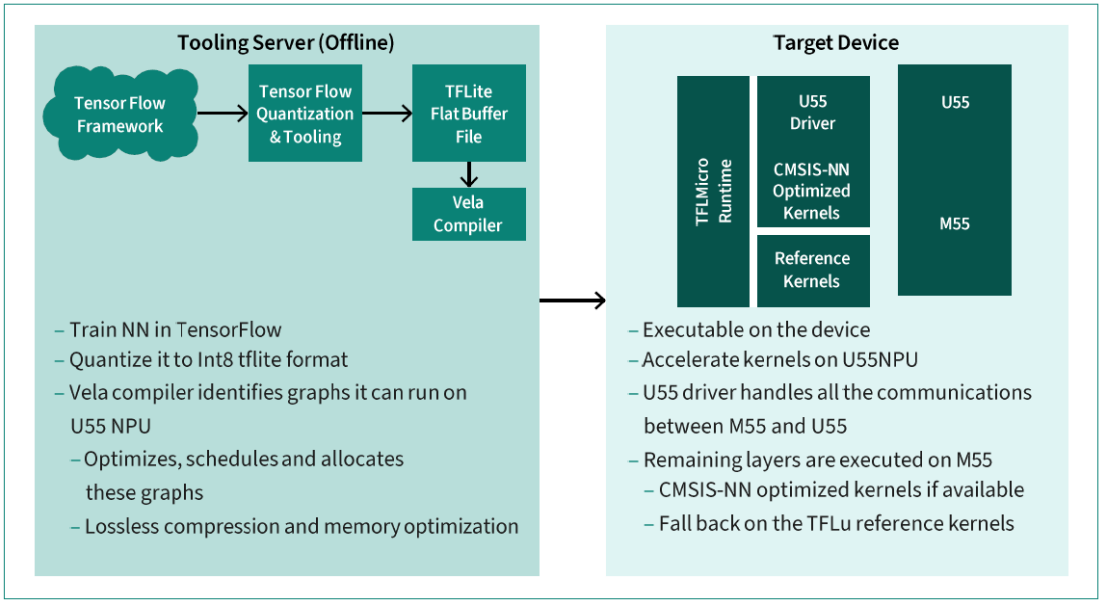

图 7 软件堆栈及运行流程图

运行时软件堆栈由以特定方式相互作用的组件构成。

- **用户应用**

用户应用在执行模型推理时运行必要的功能并调用适用于微控制器的 TensorFlow Lite（TFLM 库）。

- **适用于微控制器的 TensorFlow Lite**

TFLM 框架被编译成 C++ 库，其中包含优化模型副本以及各种参考内核和 CMSIS-

NN 内核版本。然后，用户应用使用这个库来执行推理。在推理过程中，按一次一个运算符对模型进行解析并执行相应的内核。Vela 编译的子图表示为特殊自定义运算符的实例，其关联“内核”直接将关联的 U55 指令序列和关联的 Tensor 数据传送给 Ethos™-U NPU 驱动程序。

- **参考内核**

包含了一组适用于 TFLM 框架中的所有运算符的内核。

- **CMSIS-NN 内核**

CMSIS-NN 包含高度优化的高性能内核，可以加速 TFLM 框架中的一个运算符子集。

- **NPU 驱动程序**

Ethos™-U NPU 驱动程序控制 TFLM 框架和 Ethos™-U NPU 之间的通信，以处理自定义运算符。当 Ethos™-U NPU 完成处理时，它会向驱动程序发送信号，驱动程序再通知 TFLM 库。

3.4 Cortex®-M55 CPU 与 Ethos™-U55 NPU 通信

图 8 所示为 Cortex®-M55 CPU 与 Ethos™-U55 NPU 的通信。

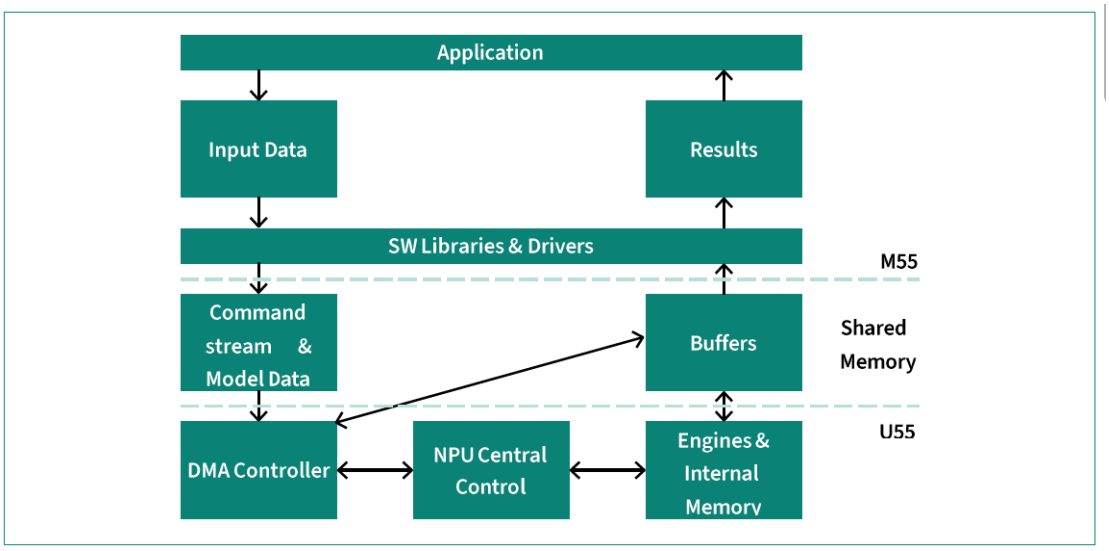

图 8 Cortex®-M55 与 Ethos™-U55 通信

共享内存、DMA 和中断提供了 Cortex®-M55 CPU 与 Ethos™-U55 NPU 之间实现高效通信所需的关键接口。

3.5 工具

3.5.1 ML 配置程序

在 ML 应用中，可使用 ModusToolbox™ 机器学习 (ML) 配置程序来对预训练学习模型进行调整，以使之适应英飞凌目标平台。这个工具接受预训练 ML 模型并生成嵌入式模型（作为库），这个模型可以和应用代码一起用于目标设备。
使用 ModusToolbox™ ML 配置程序，可将选定的预训练模型和一组优化参数发送至目标设备。这个工具是将 ML 工具生态系统中的其他资源汇聚到一起的中心资源，包括 CoreTools、推理引擎等等。

3.5.2 ML CoreTools

MTB-ML CoreTools 是后端实用程序，通过与它交互，ML 配置程序可以导入、转换和部署所支持的嵌入式 ML 模型。CoreTools 包括一个跨域验证工具，可用于验证转换后的 ML 模型相对于原始参考模型的性能，并确定转换过程是否符合准确度标准。这些后端实用程序表示为特定于平台的可执行文件。
ML CoreTools 可用于解决方案的配置、部署和跨域验证 (CDV)。CoreTools 的大部分功能均与 ML CoreTools 应用的“部署”运行模式有关。部署的目的是以训练权重来加载模型，对模型进行解析和优化，创建参考验证数据，并针对不同精度表示将模型量化。得到的输出是一组文件，其中包含参考模型的以多种不同精度表示的转换模型，以便在 MTB-ML 中进一步验证。
ML CoreTools 支持命令行界面和 JSON 配置文件选项。
在模型部署模式下，CoreTools 从文件加载 Keras H5、TFLite 预训练模型，对其进行量化和优化，然后使用随机或正常数据执行参考评估。

3.5.3 Vela 编译器

Vela 编译器是一个在工具机上执行的基于 Python 的优化器，具有以下特性和功能：
- 在嵌入式设备上实现以前不可行的 NN
- 开放源代码
- 读取 .tflite 文件并生成修改后的 .tflite 文件

- 输出包括 microNPU (Ethos™-U55) 指令
- 优化子图调度
- 无损压缩权重
- 减小 SRAM 和闪存占用空间

这个工具可用于将 TFLite 模型编译成可以在 Ethos™-U55 NPU 上运行的优化版本。模型中可以由 Ethos™-U55 NPU 加速的部分，被替换为调用特殊自定义运算符来调用 Ethos™-U55。模型中不能加速的部分保持不变，并在 Cortex® M 系列 CPU 上使用适当的内核运行。Vela 尝试不同的编译策略，并用代价函数来处理每种策略。然后，它为所支持的每个运算符或每组运算符选择最佳执行调度。

Vela 编译器还执行各种内存优化，通过压缩模型中的权重来降低闪存和运行时 SRAM 内存要求。

Vela 使用层叠来解决运行时内存使用问题。层叠可将一组连续支持的运算符的特征映射 (FM) 分割成条带，从而降低最大内存需求。条带既可以是 FM 的完整或部分宽度，也可以是 FM 的完整或部分高度。然后，用这一组的所有运算符依次处理每个条带。

将模型中可以优化和加速的部分分组并转换为 TensorFlow Lite 自定义运算符。然后，将运算符编译成 Ethos™-U55 NPU 可以执行的指令流。最后，将优化模型写成 TFLite 模型，并生成性能评估报告，提供诸如内存使用情况和推理时间等统计数据。

编译器包括许多配置选项，允许用户就嵌入式系统配置的各个方面进行指定，包括 Ethos™-U55 NPU 配置、内存类型和内存大小等。

4 ML 应用

下面列举了部分可受益于 PSoC™ Edge 硬件和软件功能的 ML 应用：

- 关键词识别
- 说话人辨认
- 图像分类
- 视频对象检测
- 手势/人员活动检测
- 可穿戴设备
- 智能设备/语音控制
- 预测性维护

5 结语

英飞凌 PSoC™ Edge MCU 由功能强大的 Cortex®-M55 MCU、Ethos™-U55 NPU、ModusToolbox™ 软件开发环境和适用于机器学习应用开发的工具等组成。它是一个极具吸引力的可选方案，可以在嵌入式/物联网环境中以低功耗实现复杂的计算密集型 ML 用例和应用。其多核架构具有独立的常开域和高性能按需唤醒功能，是适用于复杂的嵌入式/物联网 ML 应用的出色平台。

参考文献

[1] https://www.arm.com/products/silicon-ip-cpu/ethos/ethos-u55

[2] https://developer.arm.com/AI%20and%20ML#aq=%40navigationhierarchiescategories%3D%3D%22AI%2FML%22&numberOfResults=48

[3] https://www.arm.com/-/media/Files/pdf/ML%20on%20Arm/Arm_Ethos_U55_Product_Brief.pdf?revision=921d33c8-d166-4fb6-abf8-92ec306a0eeb&rev=921d33c8d1664fb6abf892ec306a0eeb&hash=1D3B17357D02451F7B4788F5BAF00F2F

[4] https://www.arm.com/-/media/Files/pdf/ethos/Arm_Accelerating_ML_Compute_for_Embedded_Market_white_paper1.pdf?revision=8772b5c9-89fa-420b-92a3-0d77e91c4597&rev=b5af90cebeb748b3ba54e5e71a988c7b&hash=673095A6389EE3DA9E7C1E05FA6C83A2

[5] https://www.infineon.com/cms/en/design-support/tools/sdk/modustoolbox-software/?gclid=CjwKCAjw-IWkBhBTEiwA2exyOzBwjtIFYIzxSJQrsXGqVRpzWACcslwmCTcZOMU9WhGV0T2IWgqpoxoCIUAQAvD_BwE&gclsrc=aw.ds

[6] https://www.infineon.com/cms/en/design-support/tools/sdk/modustoolbox-software/modustoolbox-machine-learning/

英飞凌科技股份公司印制

Am Campeon 1-15,85579 Neubiberg Germany

公开

日期：2023 年 10 月

关注我们

扫描二维码，探索我们提供的产品和解决方案

www.infineon.com

重要提示

本文档所提供的任何信息绝不应当被视为针对任何条件或者品质而做出的保证（质量保证）。英飞凌对于本文档中所提及的任何事例、提示或者任何特定数值及/或任何关于产品应用方面的信息均在此明确声明其不承担任何保证或者责任，包括但不限于其不侵犯任何第三方知识产权的保证均在此排除。

此外，本文档所提供的任何信息均取决于客户履行本文档所载明的义务和客户遵守适用于客户产品以及与客户对于英飞凌产品的应用所相关的任何法律要求、规范和标准。

本文档所含的数据仅供经过专业技术培训的人员使用。客户自身的技术部门有义务对于产品是否适宜于其预期的应用和针对该等应用而言本文档中所提供的信息是否充分自行予以评估。

如需产品、技术、交付条款和条件以及价格等进一步信息，请向离您最近的英飞凌科技办公室接洽（www.infineon.com）。

警告事项

由于技术所需产品可能含有危险物质。如需了解该等物质的类型，请向离您最近的英飞凌科技办公室接洽。

除非由经英飞凌科技授权代表签署的书面文件中做出另行明确批准的情况外，英飞凌科技的产品不应当被用于任何一项一旦产品失效或者产品使用的后果可被合理地预料到可能导致人身伤害的任何应用领域。