白皮书

云，数据中心

异构 FPGA 加速器

利用基于可扩展计算的 FPGA 突破 AI 边界

作者

Markus Adhiwiyogo
产品线经理
英特尔公司

Rohit D'Souza
AI 与加速
产品营销经理
英特尔公司

Steve Leibson
高级产品营销工程经理
英特尔公司

Ronak Shah
软件产品营销和产品线管理总监
英特尔公司

目录

简介

FPGA 可帮助客户通过系统内硬件定制快速开展创新，适应瞬息万变的市场趋势。当前的一个重要趋势是，企业对智能处理和人工智能（AI）的需求不断增长且越来越普遍，同时对定制的需求也在持续上升。创新的英特尔® FPGA 在结构和 I/O 具有灵活性以及低确定性延迟方面的内在优势，可满足企业的硬件定制和 AI 需求。

支持集成 AI 的硬件定制

通过添加 AI 功能，支持集成 AI 的硬件定制可满足传统市场中不断增长的创新需求。Microsoft Bing 搜索引擎应用就是一个很好的例子。

Microsoft Bing 搜索引擎

Microsoft Bing 搜索引擎将英特尔® FPGA 与实时 AI 相结合，可为用户更智能的提供搜索结果 [2]。Microsoft 部署了机器阅读理解功能，以理解用户查询背后的含义，从而在结果中提供 “基于网络中所有观点” [2] 的答案。机器阅读理解模型在许多网页上运行，并为用户提供汇总结果。这项功能要求很高的计算能力，同时不降低搜索速度。它是通过一个深度学习加速平台实现的，Microsoft 将该平台称为 ‘Project Brainwave’ [2]。

不过，随着创新需求的快速变化，人工智能领域也在不断演变。AI 模型的大小和复杂性都在呈指数级增长，这是一个重要的颠覆性趋势。

日趋复杂的 AI 模型

AI 模型的复杂性不断提高，其规模呈爆炸性增长，计算资源方面的创新无法与之匹配，单个设备上的内存容量已无法满足需求。如今，AI 模型的复杂度每 3.5 个月提高一倍或每年提高约 10 倍 [1]（见图 1），导致 AI 计算能力需求快速上升。由于模型中参数或权重的数量不断增加，因此对 AI 模型的内存要求也在增加。随着参数的增多，系统需要更多的片上存储来维持模型的持久性 [5]。内存需求的增加还会导致用户需要更高的 I/O 带宽，因为需要传输更多数据。这种增长趋势很可能会持续下去。

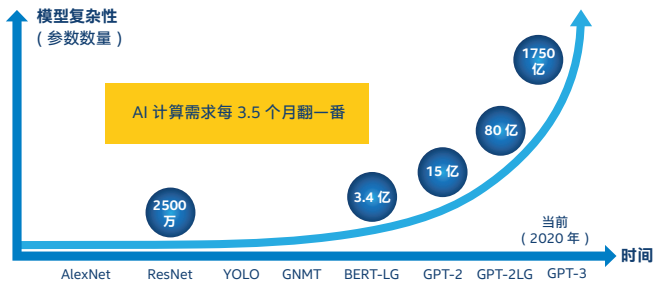

图 1. 日趋复杂的 AI 模型 [1, 6]

英特尔® Stratix® 10 NX FPGA 简介

英特尔® Stratix® 10 NX FPGA 是英特尔首款 AI 优化型 FPGA。英特尔开发了英特尔® Stratix 10 NX FPGA，以使客户能够随着 AI 复杂性的增加扩展其设计，同时继续提供实时结果。这个新的英特尔 FPGA 产品家族具有以下功能，可帮助您解决 AI 模型日趋复杂的挑战：

- 一种新型的高密度、低精度且经过 AI 优化的张量算术模块，专为满足 AI 模型需求而设计
- 用于高速网络和低延迟主机连接的高性能收发器块
- 集成的第二代高带宽内存（HBM）堆栈可满足大型 AI 模型的极端内存要求

英特尔® Stratix 10 NX FPGA 采用基于 chiplet 的英特尔架构，其中 FPGA 内核芯片连接到同一封装中的定制芯片或 chiplet，以实现最大的灵活性和敏捷性。这种基于 chiplet 的架构可以让英特尔使用正确的工艺节点实施正确的功能组合，在单个封装中提供客户所需的系统功能。例如，收发器 chiplet 用于为系统开发人员提供网络和处理器连接功能的准确组合，让他们只需很短的开发时间和很少的资源即可开发出带有集成收发器的全新芯片（参见图 2）。这种设计和制造技术帮助系统开发人员缩短了产品上市时间。对于英特尔 Stratix 10 NX FPGA，FPGA 内核芯片的数字信号处理（DSP）模块面向 AI 优化的张量算术模块进行了修改。

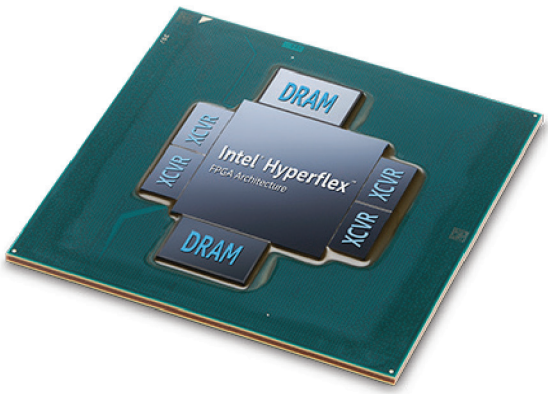

图 2. 英特尔® Stratix® 10 NX FPGA 中的 Chiplet

面向低精度 AI 推理工作负载的高张量吞吐量

英特尔 Stratix 10 NX FPGA 结构包含一种经过 AI 优化的新型张量算术块，称为 AI 张量模块。这些 AI 张量模块包含通常用于 AI 模型算术的低精度乘数的密集数组。这些 AI 张量模块中的较小乘数也可以进行聚合，构成更高精度的乘数。

AI 张量模块的架构（参见图 3）包含三个点积单元，每个点积单元都有 10 个乘数和 10 个累加器，每个模块中总共有 30 个乘法器和 30 个累加器。AI 张量模块针对用于 AI 计算的通用矩阵乘法或矢量矩阵乘法进行了调整，无论矩阵大小，均能高效运行。AI 张量模块乘数的基本精度为 INT8 和 INT4，使用共享指数以支持模块浮点 16（Block FP16）和模块浮点 12（Block FP12）数字格式。多个 AI 张量块可串联在一起以支持更大的矢量计算。总体计算性能和效率如下所示表 1。

精度	性能	效率
INT4	286 TOPS	2 TOPS/W
INT8	143 TOPS	1 TOPS/W
模块 FP12	286 TFLOPS	2 TFLOPS/W
模块 FP16	143 TFLOPS	1 TFLOPS/W
@600 MHz Max Frequency		

表 1. AI 张量模块的性能和效率 [3]

凭借 3960 个 AI 张量模块，最大目标频率为 600 MHz 的 INT8 精度的峰值性能数据可以按照如下方式进行计算 [3]：

（3,960 个 AI 张量模块）（10 输入 * 3 列）*（每个乘法 2 次运算）*

=（3,960 个 AI 张量模块）（每个 AI 张量模块 30 个乘法）*（每个乘法 2 次运算）*

= 237,600 次运算

假设最大频率为 600 MHz

（237,600 次运算）（600 MHz）*

= 142.56 TOPS

~143 TOPS

同样，INT4 精度的峰值性能是 INT8 的两倍，或大约 286 TOPS

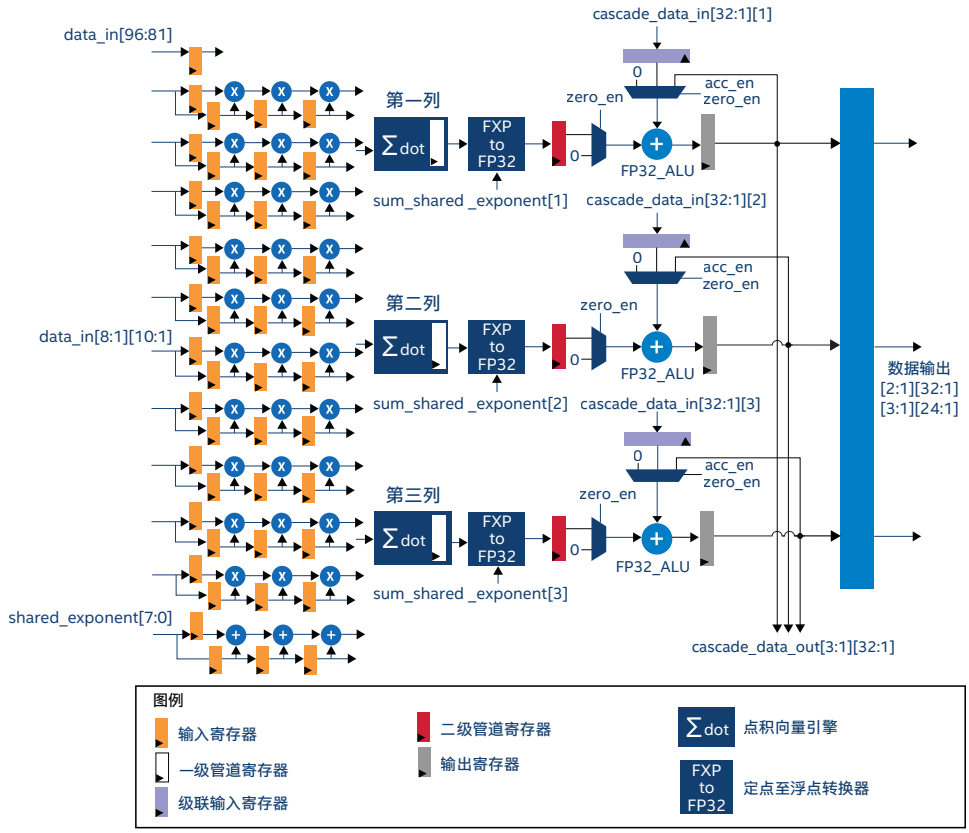

图 3. AI 张量模块架构

简而言之，英特尔预计 AI 张量模块的特性和功能支持单芯片实施 ResNet50，以大约每秒 7,000 帧（FPS）的速度处理图像（参见图 4）。该预计基于 500 MHz 时钟速率并假设 AI 张量模块的利用率为 75% [3]。

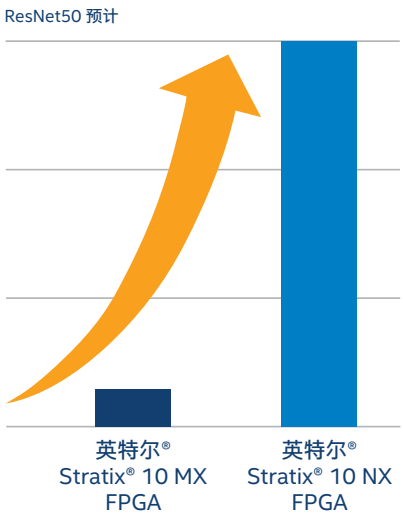

图 4. ResNet50 拓扑的理论性能 [3]

可扩展且灵活的 I/O 连接带宽

英特尔 Stratix 10 NX FPGA 包含多达 96 个串行器/解串器（SERDES）收发器。这些收发器的最大数据速率为 58.8 Gbps PAM4 或 28.9 Gbps NRZ。每个型号还包括 PCI Express*（PCIe*）Gen3 x16 和 100G 以太网媒体访问控制（MAC）或物理编码子层（PCS）IP 硬核。这些收发器允许设计人员使用多个标准或自定义协议，将英特尔 Stratix 10 NX FPGA 连接到其他各种设备。借助所有这些功能，英特尔 Stratix 10 NX FPGA 可提供高达 668 GB/s 的连接带宽，并允许设计人员实施具有更高连接带宽要求的多节点 AI 推理解决方案。这有助于构建可扩展的连接解决方案，从而灵活地适应不断变化的带宽要求。

HBM2 解决内存带宽瓶颈，提供低延迟

英特尔 Stratix 10 NX 器件在一个高性能 FPGA 封装中集成了两个 HBM2 内存堆栈，从而使这些 FPGA 能够有效应对 AI 内存带宽瓶颈带来的挑战。每个 HBM2 堆栈可提供高达 256 Gb/s 的带宽，在一个封装中可提供高达 512 Gb/s 的总带宽 — 带宽远高于四个外部连接的 DDR4 内存。

该产品家族封装了 2 个 HBM2 内存堆栈，可提供 8 GB 或 16 GB 容量 2 个选项，这 2 个 HBM2 容量选项的总带宽相同。集成的 HBM2 内存堆栈允许大型 AI 模型保留在大型片上 HBM2 内存中，与片外内存相比降低了访问延迟。通过提供较高的内存带宽，HBM2 内存堆栈化解了内存密集型工作负载的内存瓶颈，相比没有 HBM 的器件最终降低了总体延迟。

图 5 展示了这种优势：与使用外部 DDR4 SDRAM 的 FPGA 相比，随着模型尺寸的增加，带 HBM2 DRAM 的英特尔 Stratix 10 NX FPGA 的延迟增加率显著降低。图 5 显示，在消耗所有内部 FPGA SRAM 资源并强制使用外部 DDR4 SDRAM 之后，不带 HBM 的 FPGA 的延迟（蓝线）急剧增加。消耗 FPGA 的所有片上 SRAM 资源后，带 HBM2 的英特尔 Stratix 10 NX FPGA 的延迟（橙线）增加速度会慢很多。此外，与单独使用外部 DDR4 SDRAM 相比，HBM2 内存堆栈的集成降低了系统功耗并减小了主板尺寸，从而降低了总拥有成本（TCO）。

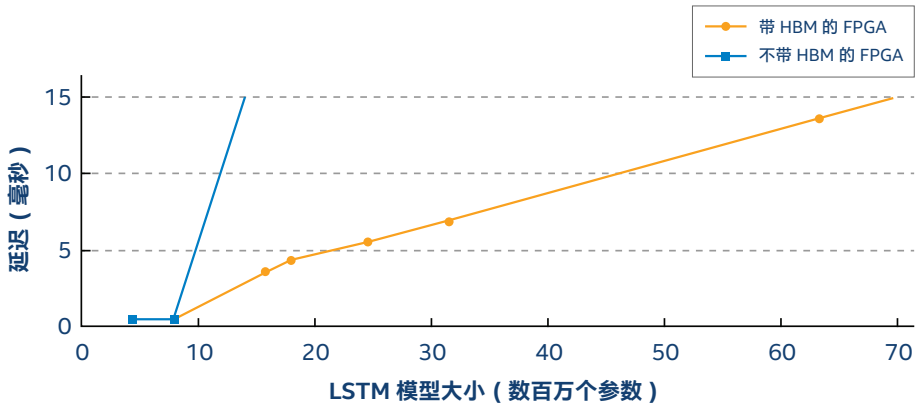

图 5. 带有 HBM 和不带 HBM 的 FPGA 之间的延迟比较 [4]

英特尔 Stratix 10 NX FPGA 应用

通常，FPGA 在需要实时、低批量和低延迟的 AI 应用中表现出色。以下示例应用演示了支持集成 AI 的硬件定制如何使用英特尔 Stratix 10 NX FPGA，在满足实时需求方面发挥关键作用。

自然语言处理

数百万用户与 Siri 或 Alexa 等基于云的设备进行交互，期望获得实时、交互式的智能响应（参见图 6）。为了实时响应数百万个同步执行的客户查询，系统必须小批量处理每个请求，同时保持低延迟。凭借 AI 张量模块以及可定制内存层级和流水线并行架构，英特尔 Stratix 10 NX FPGA 能够做到这一点。通过在多个节点中以流水线方式构建整个系统，系统设计人员可以创建实时交互式解决方案。

图 6. 语音转文本和文本转语音 — 自然语言处理应用

英特尔合作伙伴 Myrtle.ai 在英特尔 Stratix 10 NX FPGA 上演示了文本转语音实时合成应用。该应用部署了 WaveNet，后者是一个 AI 模型，可提供出色的音频质量。有关此演示的更多详细信息，请查看以下资源：

- 演示视频：
  www.intel.com/text-speech-fpga-demo
- 白皮书：
  www.intel.com/text-speech-fpga-paper

金融欺诈检测

金融欺诈检测是一种时限要求极高的应用，目的是在处理欺诈性信用卡交易之前检测并阻止该交易。英特尔 Stratix 10 NX FPGA 具有创建低延迟和确定性延迟的自定义硬件的能力以及通过高带宽网络和内存访问支持模型持久性的能力，可以实现此目标。

在一个真实的金融欺诈检测系统案例中（参见图 7），从在销售点（POS）终端刷信用卡到验证交易的总往返延迟要求必须小于 400 毫秒。然而，AI 算法只有 10 毫秒的时间来运行 AI 推理算法并确定交易是否具有欺诈性质。该应用运行的批次大小为 1，这是 FPGA 的优势所在。英特尔 Stratix 10 NX FPGA 拥有高性能 AI 张量模块和高带宽 HBM2 内存，可以轻松满足金融欺诈检测应用中严格的低延迟要求。

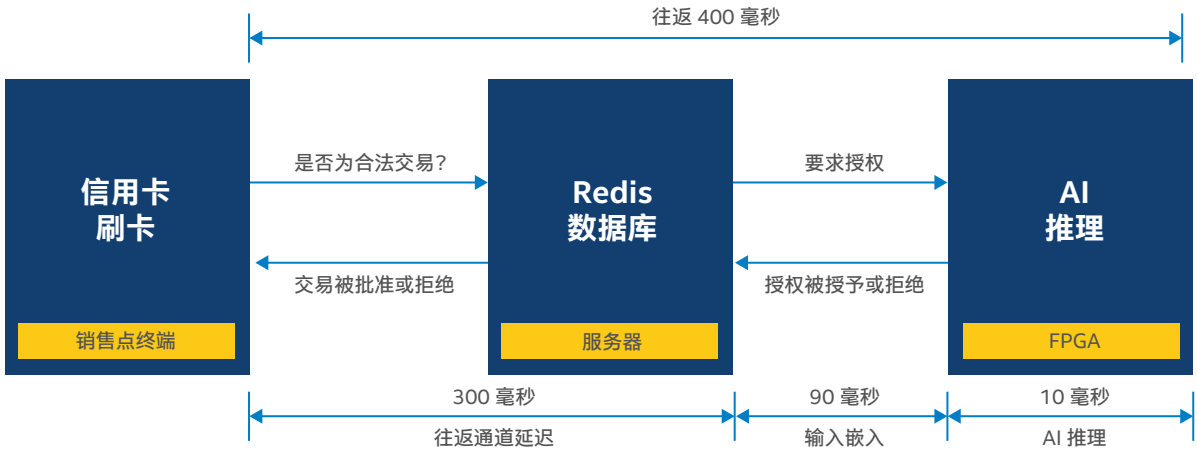

图 7. 金融欺诈检测系统和相关延迟的高级视图 [3]

智慧城市和零售

某些视频分析应用有严格的实时响应要求。这些应用通常需要将视频采集和处理与定制算法集成在一起。FPGA 凭借出色的硬件自定义功能，在这些独特的视频分析应用中表现出色，可以实现自定义处理和自定义 I/O 协议。智能零售应用的视频分析示例如图 8 所示。该应用可以识别正在购买的商品的实时视频图像，并将其与扫描的条形码进行匹配。此应用旨在减少客户扫描一件商品但将另一件更贵的商品放入购物袋时造成的财务损失。由于直接视频输入和视频采集、转换和 AI 推理阶段的流水线化全部由 FPGA 提供支持，因此整个流程只需不到 50 毫秒的时间。

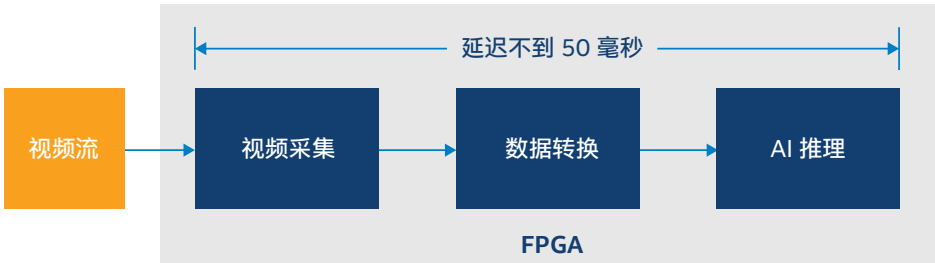

图 8. 基于 AI 的结账应用在智能零售应用中的流水线实施 [7][3]

结论

英特尔 Stratix 10 NX FPGA 提供了实施高性能 AI 系统所需的独特功能组合。它全面解决了 AI 模型日趋复杂的问题，具体包括以下四个方面：

- **高计算密度：**增强的 AI 张量模块可提供多达 286 个 INT4/Block FP12 或 143 个 INT8/Block FP16 TOPS/TFLOPS [3]，为支持客户应用保留了可用的软逻辑。
- **快速连接：**每个英特尔 Stratix 10 NX FPGA 均可提供高达 668 Gb/s 的高带宽网络，从而支持跨多个英特尔 Stratix 10 NX FPGA 的多节点应用。
- **丰富的近计算内存：**除片上 SRAM 外，还可使用 8 或 16 GB 的 HBM2 内存来存储大型 AI 模型和数据集。
- **支持集成 AI 的硬件定制：**在同一设备上支持自定义应用和 AI。

凭借这些独特的功能，企业正在采用英特尔 Stratix 10 NX FPGA 来应对 AI 模型规模不断增加的趋势，满足更高计算密度、更大内存带宽和跨多个节点扩展的需求。

更多信息

访问英特尔 Stratix 10 NX FPGA 网页 www.intel.com/stratix10nx

参考资料

[1] https://s21.q4cdn.com/600692695/files/doc_presentations/2019/11/intel-ai-summit-keynote-slides.pdf

[2] www.intel.com/content/www/us/en/programmable/b/bing-intelligence-search-with-intel-fpgas.html

[3] 基于英特尔内部预测

[4] E. Nurvitadhi 等人，“在带多个 FPGA 的 CPU 服务器上执行可扩展低延迟持久性神经机器翻译”，2019 年国际现场可编程技术会议（ICFPT），中国天津，2019 年，第 307-310 页。

[5] E. Nurvitadhi 等人，“要合作不要竞争：面向持久 RNN 的 FPGA-ASIC 集成”，2019 IEEE 第 27 届年度现场可编程自定义计算机（FCCM）国际研讨会，美国加利福尼亚州圣地亚哥，2019 年，第 199-207 页。

[6] https://medium.com/@Synced/openai-unveils-175-billion-parameter-gpt-3-language-model-3d3f453124cd

[7] https://megh.com/video-analytics-solution/

WP-01301-1.0