SYNOPSYS®

WHITE PAPER

实现设备智能

从数据中心到边缘的AI芯片设计

作者

Stelios Diamantidis
Director, AI Products,
Design Group

David Hsu
Product Marketing Director,
Verification Group

Ron Lowman
Product Marketing Manager,
Solutions Group

人工智能的兴起

人工智能的芯片和软件的爆炸性增长正在变革关于连接性、能效、移动性和安全性的方方面面。机器学习（ML）技术已经用于计算机视觉、物体检测、语音识别和大数据分析。深度学习（DL）算法和神经网络正在推动芯片和软件快速发展，以满足在处理能力、每瓦性能、内存延迟和实时连接性等方面的新要求。

对于人工智能加速的需求是由以下两项深度学习任务所推动的：训练（training）和推理（inference）。高度专用的处理器（或者称 AI 加速器）孕育而生，以管理这些任务所需要的规模巨大且不断变化的计算强度。在数据中心中，具有高度并行、大量复制的计算结构的 AI 加速器正被用于训练数千万到数亿个神经元，而其功耗仅为通用 CPU 和 GPU 的一小部分。在推理方面，借助于到2021 年高达85亿部的智能手机出货量（参见：Gartner，2017 年3 月），移动边缘的 AI 加速器 —— 有些面积仅为1mm2 ——正在成为世界上规模最大的计算环境，通过他们经过训练的神经元，加速器在几毫秒内完成数万亿次计算，从而满足交互式应用的需求（图1）。

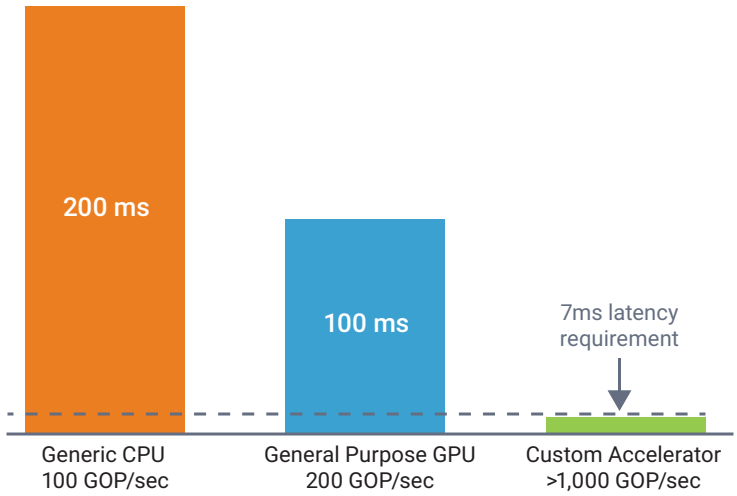

图1：满足语音识别中使用的55层、3400万权重递归神经网络（RNN）的推理延迟要求

AI加速器：高度动态和不断演进

当下已经出现了多种多样的硬件平台来满足从数据中心到边缘的 AI 计算需求。然而，为当前及未来的 AI 应用创建芯片并非易事。设计人员必须解决许多技术挑战，涉及到广泛的AI算法以及相应的硬件架构异构。设计人员还需要克服高性能、低功耗物理设计所带来的复杂性和成本问题。

算法创新

每天都有研究成果从大学及行业实验室中涌现出来，它们不断产生新的神经网络模型、增强现有模型，并生成大量数据集对各个模型进行训练和测试。一旦这些模型经过了训练，就会出现进一步的创新，以将这些模型压缩并映射到不同的硬件架构中。创新者需要平衡多项相互竞争的要求，比方说，带宽（例如，每个卷积每秒的乘法累加（MAC）运算次数）、量化或数据类型选择的影响、面积效率（每平方毫米每秒帧数）、内存带宽效率（MB/帧），等等（图2）。

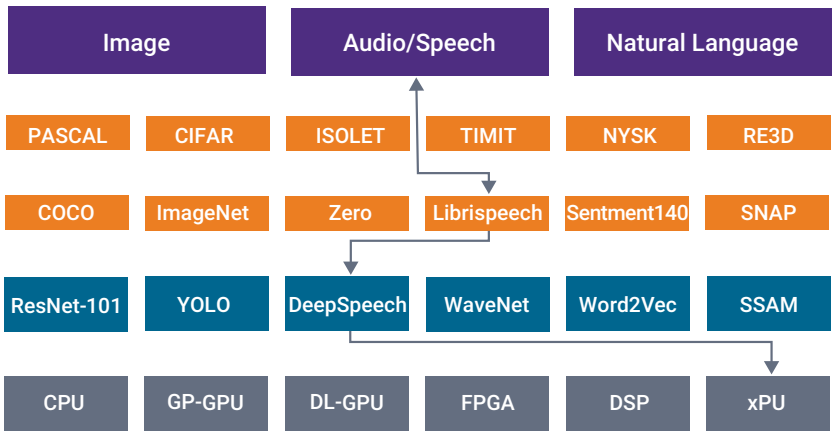

图2：将 AI 应用映射到硬件加速平台

多样化的架构

业界正在提出广泛的异构计算架构来加速计算，同时降低每次运算的总功耗。每个 AI 应用都有专门的计算、内存和互连需求。除了 AI 加速器功能本身之外，AI 芯片还包含各种其他组件。例如，数据中心设备与 AI 数据中心之间必须具有可靠且可配置的连接，而边缘设备则应当包含与传感器、图像、音频等之间的实时接口（图2）。内存选择则对于在低功耗条件下满足低延迟访问要求尤为重要。

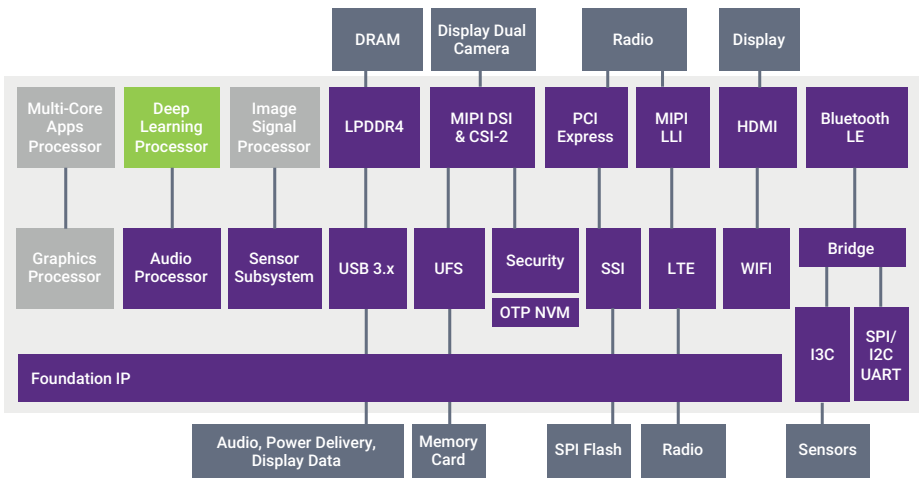

图2：示例：在边缘处用于增强现实/虚拟现实的 SoC 架构，包含一个深度学习加速器。

高性能设计

最近出现的面向数据中心的架构，往往包含超过200亿个晶体管以及数百个或数千个处理模块，其速度可以超过5GHz。以低延迟支持数据局部性的新型嵌入式内存拓扑，正在推动晶体管主导的布局，并带来布线拥塞（routing congestion）和宏时序路径闭合的挑战。在边缘处，设计人员正在运行1GHz 以上的推理引擎，它们需要以极端温度和电压角（voltage corner）来运行。要想满足因算法和数据集而异的功率和热量预算，就必须采用新的功率估算和分析方法。

因此，一套综合性的AI芯片设计方法必须解决 AI 设计的所有方面的问题——从加速算法创新周期（phase），到可靠地拼接各种架构，再到最终提供最佳的物理实现，直至完成制造 signoff。

Synopsys：支持从数据中心到边缘的AI设计

作为面向全球众多最具创新性的企业的 Silicon to Software™（“芯片到软件”）合作伙伴，Synopsys 公司帮助开发了我们每天所依赖许多电子产品和软件应用。Synopsys 与数据中心和云服务提供商合作开展训练和推理，与汽车半导体领导者合作开展自动驾驶汽车研究、视觉处理和决策，并在移动、物联网领域推进深度学习加速器，以优化性能、强化隐私保护。因此，世界各地数据中心的 AI 加速器几乎都采用 Synopsys 软件进行设计和验证。Synopsys 还与许多AI创业公司密切合作，使他们得益于我们的领域知识以及针对 AI 优化的解决方案。

AI架构探索

Synopsys 的 Verification Continuum 为加速和优化AI架构探索提供了独特的解决方案。Platform Architect 虚拟原型设计能够实现架构级的性能和功耗分析。基于 HAPS FPGA 的原型设计以及ZeBu仿真使得对极其庞大和复杂的RTL实现进行探索和验证变得切实、可行。

AI加速器和数据框架的协同验证

在卷积神经网络（CNN）上，进行三个16x16像素图像的 RTL 仿真对任何软件模拟器而言，都超出了当前业界最先进水平。Synopsys ZeBu 是业界速度最快的仿真系统，也是唯一经过验证的解决方案，能够满足全AI芯片仿真的容量和速度需求。与其他解决方案相比较，它提供了最高的容量（190亿个以上门限）和最低的拥有成本（功耗降低5倍，数据中心占用空间减少一半）。ZeBu 具有 AI 性能可视化功能，其中包括图形追溯、张量图吞吐量分析、内存性能分析等等。

针对硬件加速对AI模型进行优化

Synopsys 的 ASIP Designer 是一套业界领先的工具，用于设计完全可编程处理器以及AI加速器。ASIP Designer 通过自定义数据路径，自动化高度并行的、完全软件可编程的硬件的实现，并针对硬件处理和软件算法迭代优化了 AI 模型。（图4）

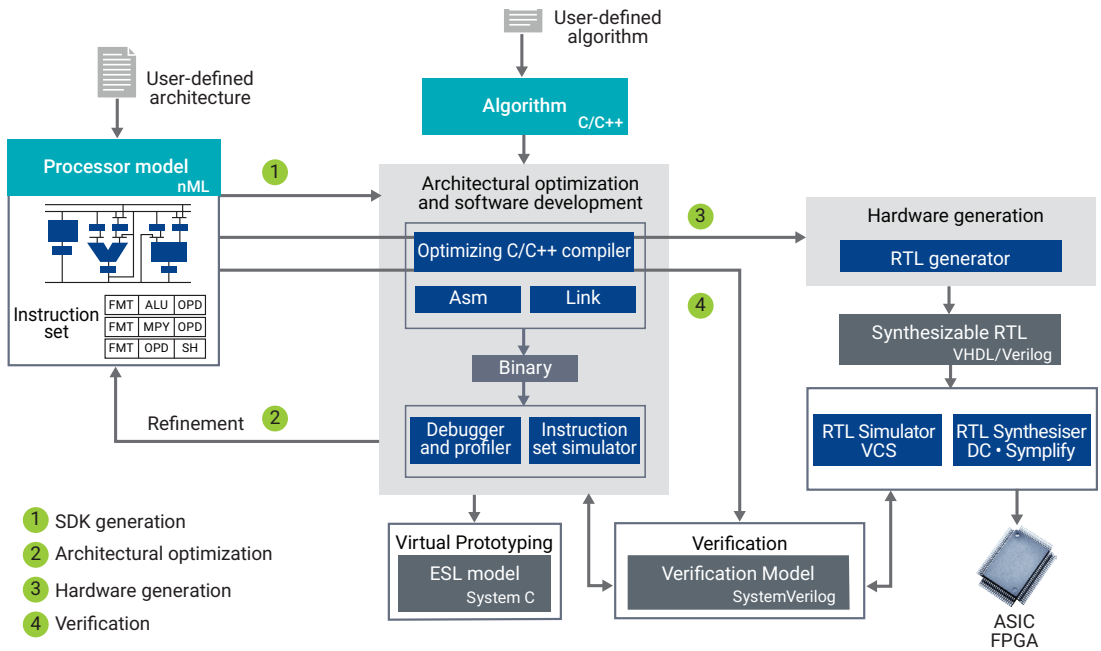

图4：Synopsys 的 ASIP Designer

业界最广泛、经过芯片验证的、面向AI的IP

Synopsys 经过芯片验证的 DesignWare IP® 产品组合能够满足 AI 市场对各种各样处理能力、内存和连接性的要求，其包括移动、物联网、数据中心、汽车和数字家居。处理器能够管理面向机器学习和深度学习任务的大量且不断变化的计算需求；内存 IP 解决方案能够针对不同内存约束条件支持高效的架构，这些约束条件包括带宽、容量和缓存一致性等；接口IP解决方案针对各种 AI 应用（包括视觉、自然语言理解以及上下文感知）提供与 CMOS 图像传感器、麦克风以及运动传感器之间的可靠连接。

突破性能、功率、面积（PPA）方面的限制

采用 Fusion 技术™的 Synopsys 设计平台是开发具有最佳功率、性能、面积和良品率的高级数字、定制及模拟/混合信号设计的首选。作为全球90%的 FinFET 设计背后的技术，Synopsys 的数字设计实施解决方案已经借助于若干项关键的、以人工智能为重点的优化技术进行了增强，这些技术包括针对数百个复制模块的 AI 芯片互连规划、MAC 拓扑优化、完整的 AI IP 参考流程等等。通过与世界领先的代工厂密切合作，Synopsys 的设计平台能够提供领先的工艺技术所具有的最大优势，同时满足 AI 芯片时间上的严格要求。

使用AI增强的设计工具进行创新

诸如机器学习等人工智能技术可以帮助解决高度复杂、高成本的挑战，不仅适用于 AI 设计，也适用于其他各种设计工作。Synopsys 推出的 AI 增强型设计平台以及 VC Formal 回归模式加速器充分体现了人工智能的颠覆性能力，它们能够在客户环境中不断地学习并改善性能 —— 一个有别于传统系统的明显特征。AI 增强工具能够加速计算密集型分析、预测可以促进更好决策的结果，并充分利用过去学习的内容来智能地指导调试，从而提高设计人员的工作效率。

总结

许多创新性应用正在推动具有 AI 功能的芯片的增长。深度神经网络需要专门的加速器，这反过来又会为计算、内存、电源和连接性带来新的架构要求。Synopsys 提供了一套全面的解决方案，能够解决 AI 设计各个方面的问题 —— 从加速算法创新周期（phase），到探索和验证各种不同的架构，到最终提供最佳的物理实现，并同时最大限度体现领先的代工厂节点的优势。

行业领导者以及世界上大多数最具创新性的AI公司都依靠 Synopsys 的解决方案来实施从数据中心到边缘的 AI 芯片。

有关SynopsysAI设计解决方案的更多信息，请访问 www.synopsys.com/ai.

03/18/19.CS328362694_AI-Whitepaper-CN.

Pub: Sept. 2018