WP486 (v1.0) 2016 年 11 月 11 日

利用 Xilinx 器件的 INT8 优化开展深度学习

作者：Yao Fu、Ephrem Wu、Ashish Sirasao、Sedny Attia、Kamran Khan 和 Ralph Wittig

赛灵思 INT8 优化为深度学习推断提供了性能最佳、能效最高的计算技术。赛灵思的集成式 DSP 架构与其他 FPGA DSP 架构相比，在 INT8 深度学习运算上能实现 1.75 倍的解决方案级性能。

概要

本白皮书旨在探索实现在赛灵思 DSP48E2 Slice 上的 INT8 深度学习运算，以及与其他 FPGA 的对比情况。在相同资源数量情况下，赛灵思的 DSP 架构凭借 INT8 在 INT8 深度学习每秒运算次数 (OPS) 上相比其它 FPGA，能实现 1.75 倍的峰值解决方案级性能。由于深度学习推断可以在不牺牲准确性的情况下使用较低位精度，因此需要高效的 INT8 实现方案。

赛灵思的 DSP 架构和库专门针对 INT8 深度学习推断进行了优化。本白皮书介绍如何使用赛灵思 UltraScale 和 UltraScale+ FPGA 中的 DSP48E2，在共享相同内核权重的同时处理两个并行的 INT8 乘法累加 (MACC) 运算。本白皮书还阐述了要运用赛灵思这一独特技术，为何输入的最小位宽为 24 位。本白皮书还以 INT8 优化技术为例，展示了该技术与神经网络基本运算的相关性。

用于深度学习的 INT8

深度神经网络（DNN）已掀起机器学习领域的变革，同时运用新的达到人类水平的 AI 功能重新定义众多现有的应用。

随着更精确的深度学习模型被开发出来，它们的复杂性也带来了高计算强度和高内存带宽方面的难题。能效正在推动着深度学习推断新模式开发方面的创新，这些模式需要的计算强度和内存带宽较低，但绝不能以牺牲准确性和吞吐量为代价。降低这一开销将最终提升能效，降低所需的总功耗。

除了节省计算过程中的功耗，较低位宽的计算还能降低内存带宽所需的功耗，因为在内存事务数量不变的情况下传输的位数减少了。

研究显示要保持同样的准确性，深度学习推断中无需浮点计算 [参考资料 1][参考资料 2][参考资料 3]，而且图像分类等许多应用只需要 INT8 或更低定点计算精度来保持可接受的推断准确性 [参考资料 2][参考资料 3]。表 1 列出了精调网络以及卷积层和完全相连层的动态定点参数及输出。括号内的数字代表未精调的准确性。

表 1： **带定点精度的 CNN 模型**

	层输出	卷积参数	完全相连 (FC) 参数	32 位浮点基线	定点精度
LeNet（示例 1）	4 位	4 位	4 位	99.1%	99.0% (98.7%)
LeNet（示例 2）	4 位	2 位	2 位	99.1%	98.8% (98.0%)
Full CIFAR-10	8 位	8 位	8 位	81.7%	81.4% (80.6%)
SqueezeNet top-1	8 位	8 位	8 位	57.7%	57.1% (55.2%)
CaffeNet top-1	8 位	8 位	8 位	56.9%	56.0% (55.8%)
GoogLeNet top-1	8 位	8 位	8 位	68.9%	66.6% (66.1%)

注意：
1. 来源：Gysel 等，卷积神经网络的面向硬件的近似法，2016 年深度学习国际会议 (ICLR) [参考资料 2]

赛灵思 DSP Slice 片上的 INT8 深度学习

赛灵思的 DSP48E2 设计用于在一个时钟周期内高效地完成一个乘法累加算法，多达 18x27 位的乘法和多达 48 位的累加，如图 1 所示。除了采用回送或链接多个 DSP Slice，乘法累加 (MACC) 也能使用赛灵思器件高效完成。

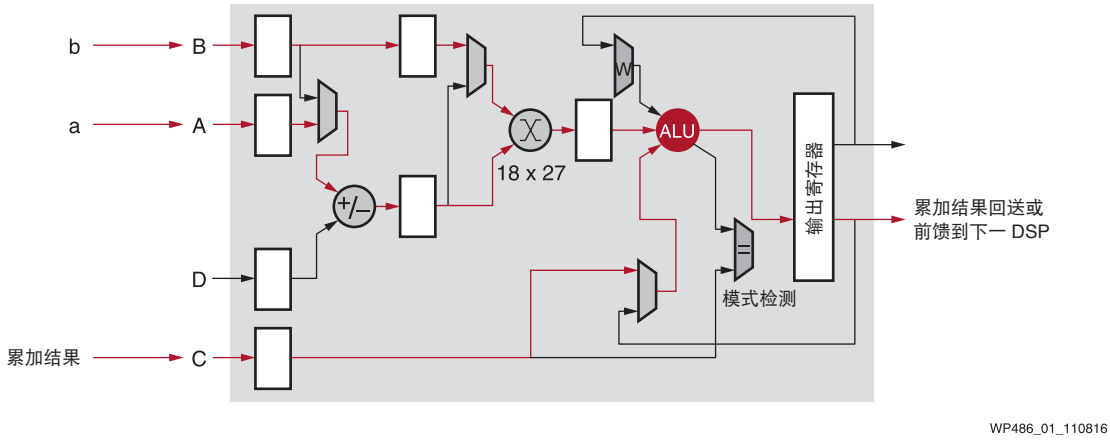

图 1： **使用 MACC 模式的 DSP Slice**

在运行 INT8 计算时，较宽的 27 位宽自然占有优势。在传统应用中，预加法器一般用于高效实现 (A+B) x C 计算，但这类计算在深度学习应用中很少见。将 (A+B) x C 的结果拆分为 A x C 和 B x C，然后在独立的数据流中进行累加，使之适用于典型深度学习计算的要求。

对 INT8 深度学习运算来说，拥有 18x27 位乘法器很占优势。乘法器的输入中至少有一个必须为最小 24 位，同时进位累加器必须为 32 位宽，才能在一个 DSP Slice 上同时进行两个 INT8 MACC 运算。27 位输入能与 48 位累加器结合，从而将深度学习求解性能提升 1.75 倍（1.75:1 即为 DSP 乘法器与 INT8 深度学习 MACC 的比率）。其他厂商提供的 FPGA 在单个 DSP 模块中只提供 18x19 乘法器，DSP 乘法器与 INT8 MACC 之比仅为 1:1。

可扩展的 INT8 优化

目标是找到一种能够对输入 *a*、*b* 和 *c* 进行高效编码的方法，这样 *a*、*b* 和 *c* 之间的相乘结果可以容易地分解为 *a* x *c* 和 *b* x *c*。

在更低精度计算中，例如 INT8 乘法中，高位 10 位或 19 位输入用 0 或 1 填充，仅携带 1 位信息。对最终的 45 位乘积的高位 29 位来说，情况一样。因此可以使用高位 19 位开展另一计算，不会影响低位 8 位或 16 位输入结果。

总的来说，要把未使用的高位用于另一计算必须遵循两条规则：

1. 高位不应影响低位的计算。
2. 低位计算对高位的任何影响必须可检测、可能恢复。

为满足上述规则，高位乘积结果的最低有效位不得进入低位 16 位。因此高位的输入应至少从第 17 位开始。对一个 8 位的高位输入，总输入位宽至少为 16+8=24 位。这样的最小 24 位输入宽度只能保证同时用

一个乘法器完成两次相乘，但仍足以实现 1.75 倍的 MACC 的总吞吐量。

接下来的步骤是在一个 DSP48E2 Slice 中并行计算 *ac* 和 *bc*。DSP48E2 Slice 被用作一个带有一个 27 位预加法器（输入和输出均为 27 位宽）和一个 27 x 18 乘法器的算术单元。见图 2。

1. 通过预加法器在 DSP48E2 乘法器的 27 位端口 p 打包 8 位输入 *a* 和 *b*，这样 2 位向量能尽量分隔开。输入 *a* 左移位仅 18 位，这样从第一项得到的 27 位结果中的两个符号位 *a* 以避免在 *b*<0 和 *a*=–128 时预加法器中发生溢值。*a* 的移位量为 18，恰好与 DSP48E2 乘法器端口 B 的宽度一样。

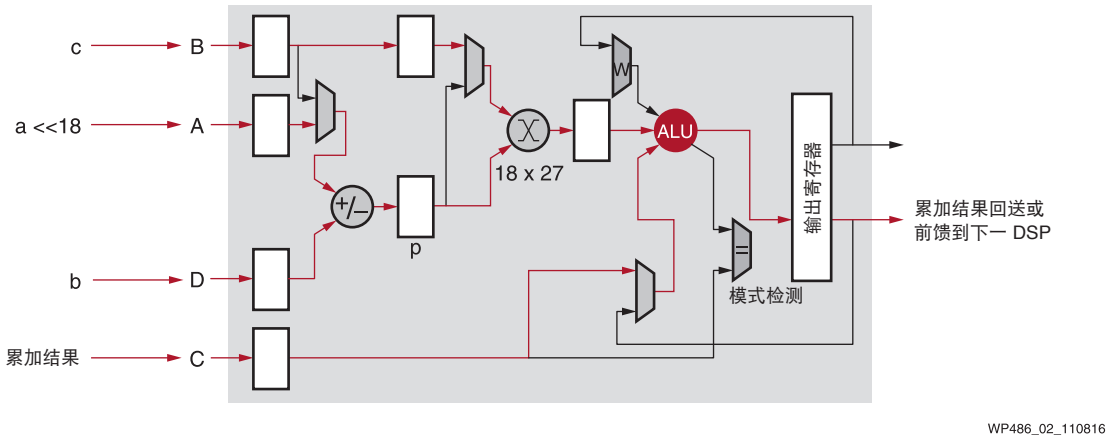

图 2： **8 位优化**

2. DSP48E2 27x18 乘法器用于计算打包的 27 位端口 *p* 和以二进制补码格式表达在 18 位 *c* 中的 8 位系数的积。现在该 45 位乘积是二进制补码格式的两个 44 位项的和：左移位 18 位的 *ac* 和 *bc*。

后加法器可用于累加上述包含单独的高位乘积项和低位乘积项的 45 位乘积。在累加单个 45 位积时，对高位项和低位项进行了校正累加。最终的累加结果如果没有溢值，可以用简单运算分开。

这种方法的局限在于每个 DSP Slice 能累加的乘积项的数量。由于高位项和低位项间始终保持两位（图 3），可以保证在低位不溢值的情况下累加多达 7 个项。在 7 个乘积项之后，需要使用额外的 DSP Slice 来克服这一局限。因此这里 8 个 DSP Slice 执行 7 x 2 INT8 乘法 - 加法运算，与拥有相同数量乘法器的竞争型器件相比 INT8 深度学习运算的效率提升 1.75 倍。

根据实际用例的要求，这种方法有多种变化形式。带有校正线性单元 (ReLU) 的卷积神经网络 (CNN) 产

生非负激活，同时*无符号* INT8 格式将精度增加一位以上且峰值吞吐量提升 1.78 倍。

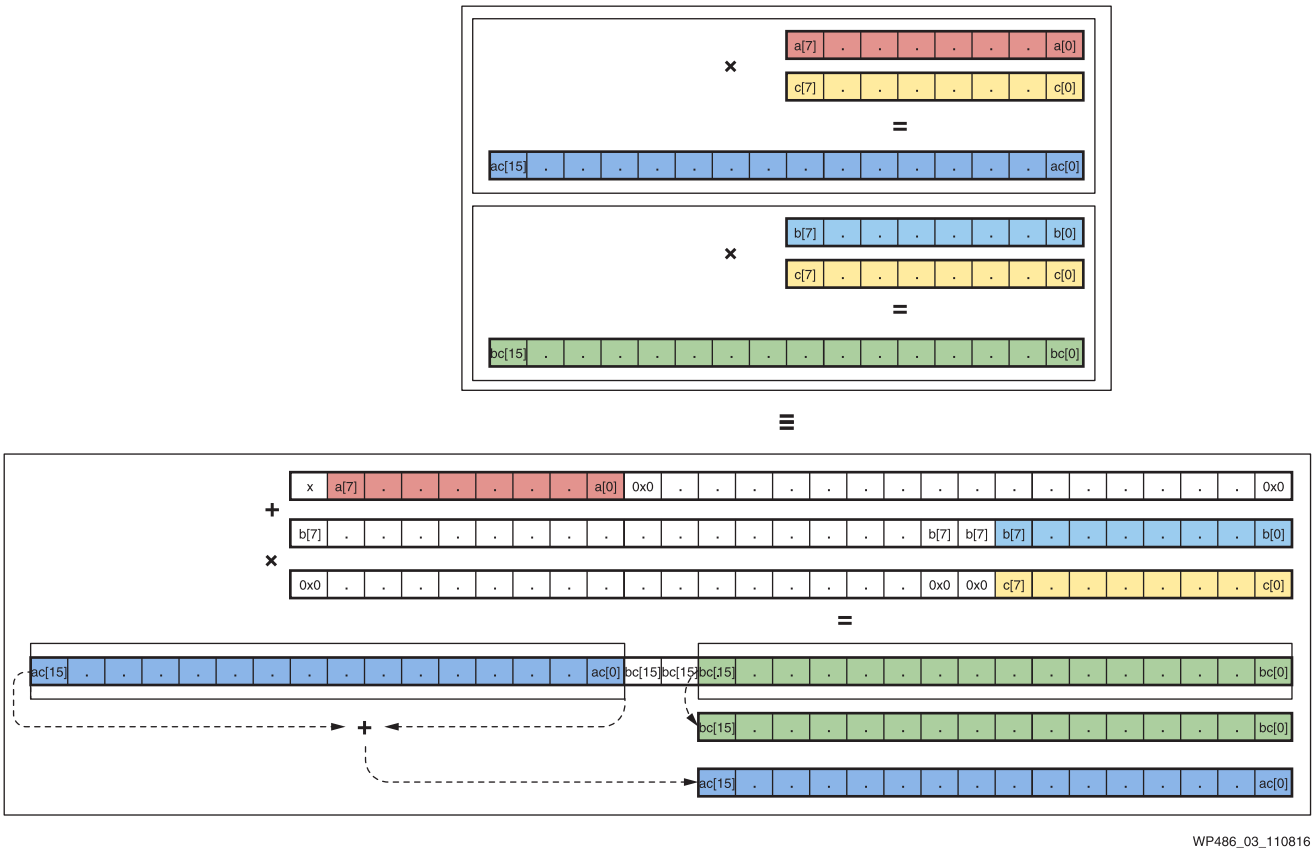

图 3： **用单个 DSP48E2 Slice 打包两个 INT8 乘运算**

CNN 的计算要求

现代神经网络大部分是从这个原始概念模型 [参考资料 4] 衍生而来。见图 4。

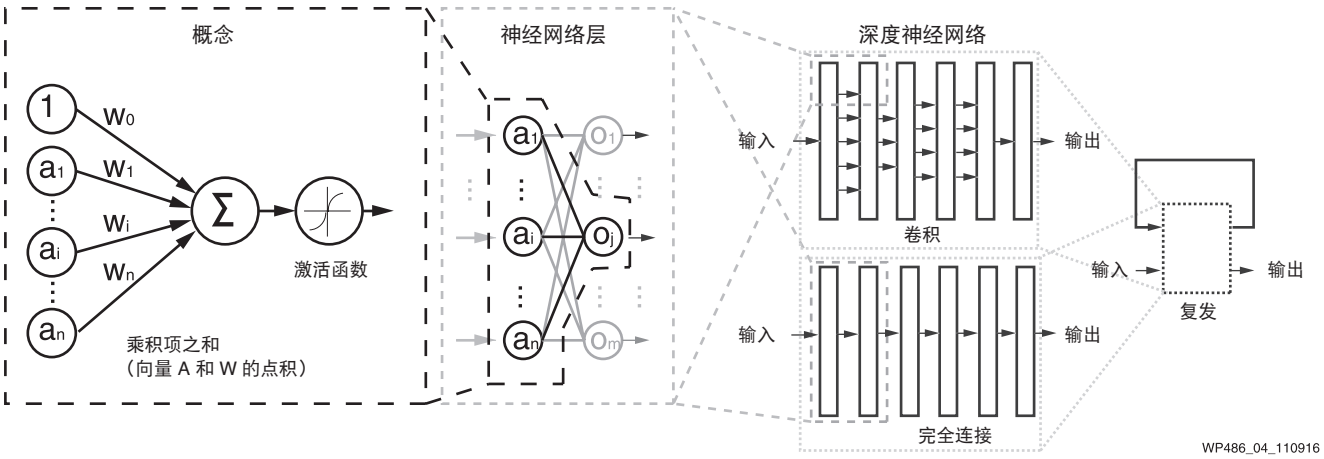

图 4： **概念和深度神经网络**

虽然从标准感知器结构开始已有相当程度的演进，现代深度学习（也称为深度神经网络 (DNN)）的基本运算仍然是类感知器的运算，只是有更广大的总体和更深入的堆叠感知器结构。图 4 所示的是一个感知器的基本运算。在每个典型的深度学习推断中它穿过多个层，最终重复数百万至数十亿次。如图 5 所示，在一层神经网络中计算 m 个感知器 / 神经元输出中的每一个的主要计算运算为：

$$o_j \ (j \in [1, m])$$

将全部的 n 个输入样本

$$a_i (i \in [1, n])$$

乘以对应的内核权重

$$w_{i,j} \ (i \in [1, n], j \ \in [1, m])$$

并累加结果

$$o_j = f(\sum_i a_i \, w_{i,j}), \qquad (i \in [1, n])$$

其中：f(x) 可以是任何选择的激活函数。

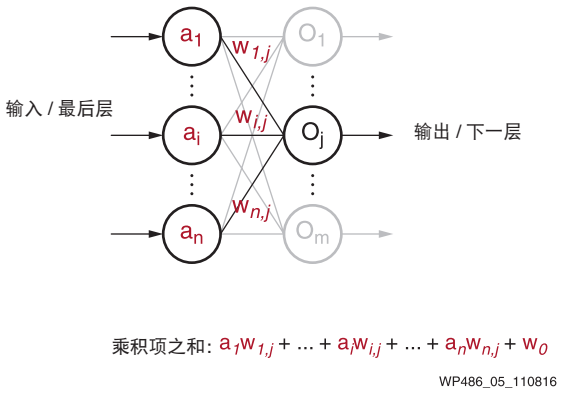

图 5： **深度学习中的感知器**

如果 a_i 和 $w_{i,j}$ 的精度限定为 INT8，该乘积之和是 INT8 优化方法中介绍的并行 MACC 中的第一个。

第二个乘积和使用相同输入 $a_i \ (i \in [1, n])$，但内核权重不同 $w_{i,k} \ (i \in [1, n], k \ \in [1, m],$ 且 $k \ \neq j)$。

第二个感知器 / 神经元输出的结果是

$$o_k = f(\sum_i a_i \, w_{i,k}), \qquad (i \in [1, n], k \ \neq j)$$

见图 6。

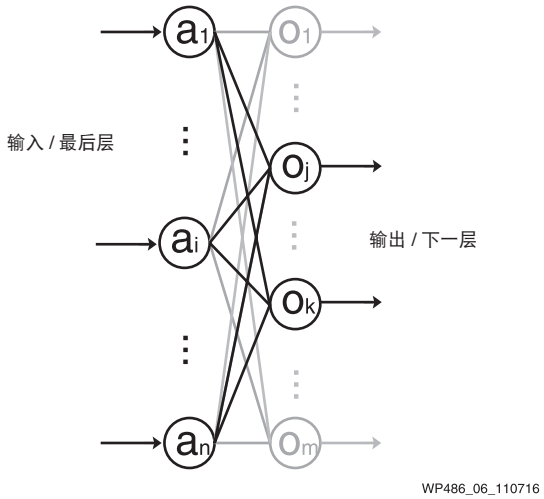

图 6： **使用共享输入并行得到两个乘积项和**

使用 INT8 优化方法将 $w_{i,k}$ 值向左移位 18 位，每个 DSP Slice 就得出最终输出值的部分且独立的一部分。用于每个 DSP Slice 的累加器有 48 位宽并链接到下一个 Slice。为避免移位 $w_{i,k}$ 饱和影响到计算，链接的模块数量被限制为 7 个，即对总共 *n* 个输入样本使用 *2n* 个 MACC 和 *n* 个 DSP Slice。

典型的 DNN 每层有数百到数千个输入样本。但是在完成 7 个项的累加后，48 位累加器的低位项可能饱和，因此每 7 个项之和就需要一个额外的 DSP48E2 Slice。这相当于每 7 个 DSP Slice 和 14 个 MACC，另加一个 DSP Slice 用于防止过饱和，从而带来 7/4 或 1.75 倍的吞吐量提升。

在卷积神经网络（CNN）中，卷积层一般主要使用同一组权重，从而形成 *a x w* 和 *b x w* 类型的并行 MACC 运算。因此除输入共享外，还可以使用权重共享（见图 7）。

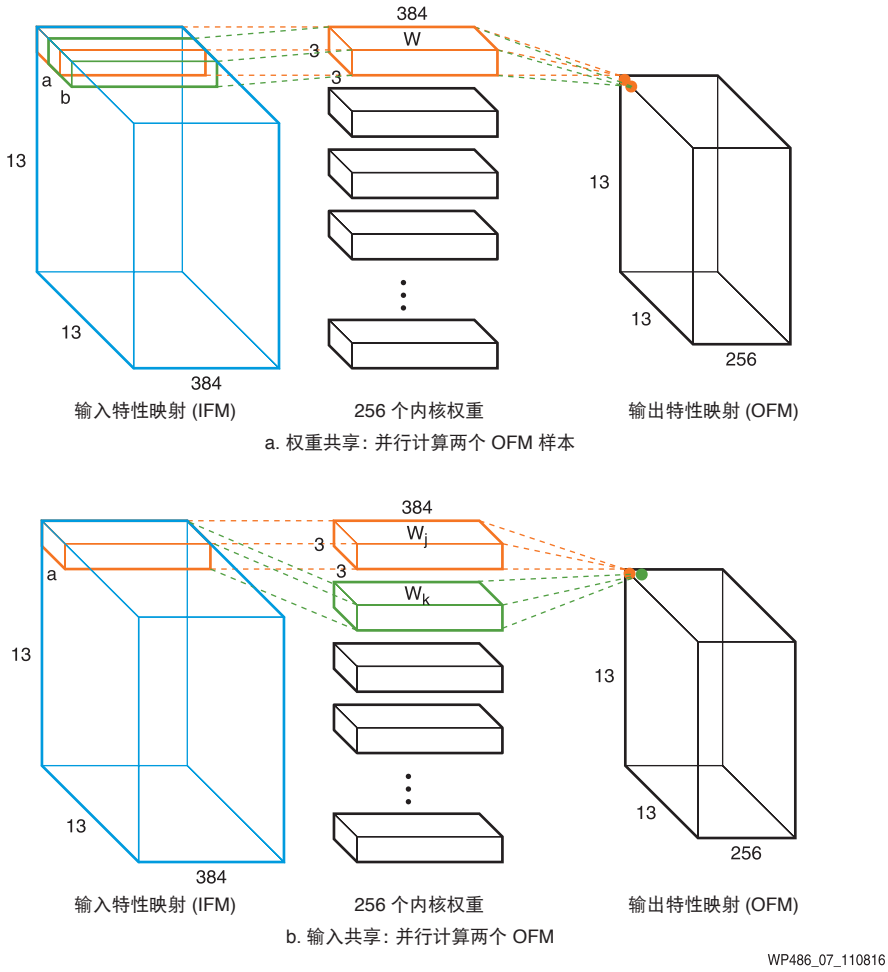

图 7： **权重共享和输入共享比较**

创建 INT8 链接 MACC 的其他方法

INT8 MACC 还能用 FPGA 架构内与 DSP Slice 频率近似的 LUT 来构建。根据 FPGA 的使用情况，这可以显著提升深度学习性能，在某些情况下性能可提升三倍之多。许多情况下相对于其他非 FPGA 架构而言，在计算可用深度学习运算时这些可用的计算资源并未考虑在内。

赛灵思 FPGA 中的编程架构是独有的，因为它能并行且高效地处理多样化工作负载。例如赛灵思 FPGA 能并行执行 CNN 图像分类、网络加密和数据压缩。我们的深度学习性能竞争分析并未将 MACC LUT 考虑在内，因为一般 LUT 用于执行 MACC 功能比用于执行其他并行功能时更有价值。

竞争分析

在本竞争分析中，将英特尔（前 Altera）的 Arria 10 和即将推出的 Stratix 10 器件与赛灵思的 Kintex® UltraScale ™ 和 Virtex® UltraScale+ ™ 进行了对比。对这种高计算强度的比较，选择的器件均为每个产品系列中 DSP 密度最高的器件：Arria 10 (AT115)、Stratix 10 (SX280)、Kintex UltraScale (KU115)、Virtex UltraScale+ (VU9P) 和 Virtex UltraScale+ (VU13P) 器件。比较的重点是能用于包括深度学习在内的众多应用的通用 MACC 性能。

英特尔的 MACC 性能基于运用预加法器的算子。但是这种实现方案产生的是乘积项和非唯一单独乘积项之和，因此英特尔的预加法器不适用于深度学习运算。

英特尔器件的功耗使用英特尔的 EPE 功耗估算工具估算，并假设在以下最坏情况下：

1. 在最大频率 (F_{MAX}) 下 DSP 利用率为 90%
2. 时钟速率为 DSP F_{MAX} 时逻辑利用率为 50%
3. 时钟速率为 DSP F_{MAX} 的一半时 ,block RAM 利用率为 90%
4. 4 个 DDR4 和 1 个 PCIe Gen3 x 8
5. DSP 触发率为 12.5%
6. 80° T_J

图 8 所示为深度学习运算的能效比较。凭借 INT8 优化，赛灵思 UltraScale 和 UltraScale+ 器件在 INT8 精度上相比 INT16 运算（KU115 INT16/KU115 INT8）能效提升 1.75 倍。与英特尔的 Arria 10 和 Stratix 10 器件相比，赛灵思器件在深度学习推断运算上能效高出 2-6 倍。

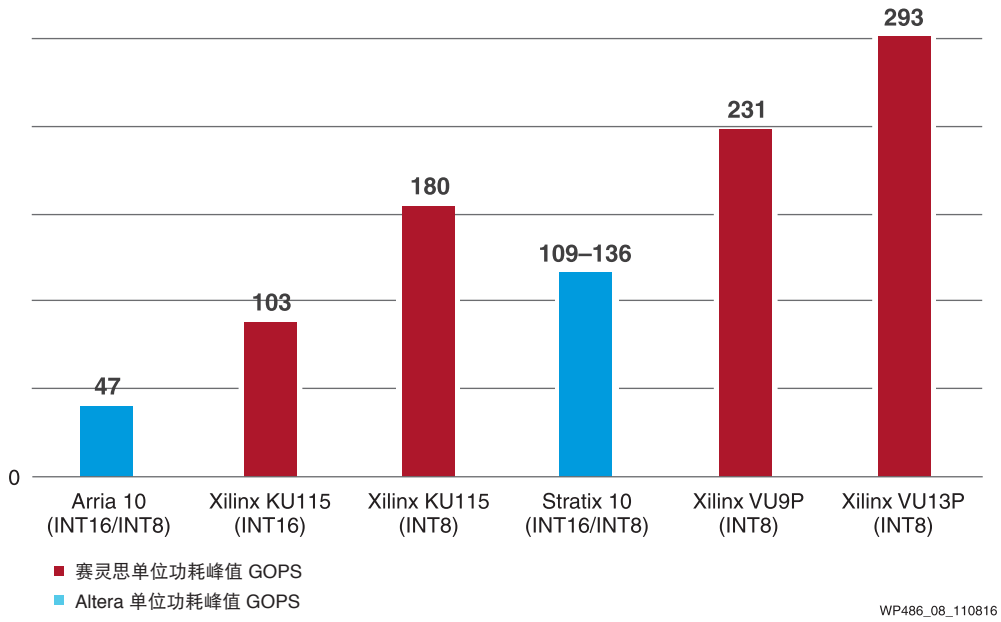

图 8： **INT8 深度学习能效对比：赛灵思对比英特尔**

结论

本白皮书探讨了如何在赛灵思 DSP48E2 Slice 上优化 INT8 深度学习运算，从而实现 1.75 倍的性能提升。赛灵思 DSP48E2 Slice 可用于在共享相同内核权重的同时实现并行 INT8 MACC。为高效地实现 INT, 需要采用 24 位输入宽度，这项优势只有赛灵思 UltraScale 和 UltraScale+ FPGA DSP Slice 能够提供支持。赛灵思非常适合用于深度学习应用中的 INT8 工作负载（例如图像分类）。赛灵思不断创新新的基于软 / 硬件的方法，以加快深度学习应用的发展。

如需了解有关数据中心深度学习的更多信息，敬请访问：

https://china.xilinx.com/accelerationstack

参考资料

1. 1.Dettmers, 8-Bit Approximations for Parallelism in Deep Learning, ICLR 2016
   https://arxiv.org/pdf/1511.04561.pdf
2. Gysel et al, Hardware-oriented Approximation of Convolutional Neural Networks, ICLR 2016
   https://arxiv.org/pdf/1604.03168v3.pdf
3. Han et al, Deep Compression:Compressing Deep Neural Networks With Pruning, Trained Quantization And Huffman Coding, ICLR 2016
   https://arxiv.org/pdf/1510.00149v5.pdf
4. Rosenblatt, F., The Perceptron:A Probabilistic Model for Information Storage and Organization in the Brain, Psychological Review, Vol. 65, No. 6, 1958
   http://www.ling.upenn.edu/courses/cogs501/Rosenblatt1958.pdf

修订历史

下表列出了本文档的修订历史：

日期	版本	修订描述
2016-11-11	1.0	赛灵思初始版本

免责声明

本文向贵司 / 您所提供的信息（下称“资料”）仅在选择和使用赛灵思产品时供参考。在适用法律允许的最大范围内：(1) 资料均按“现状”提供，且不保证不存在任何瑕疵，**赛灵思在此声明对资料及其状况不作任何保证或担保，无论是明示、暗示还是法定的保证，包括但不限于对适销性、非侵权性或任何特定用途的适用性的保证；**且 (2) 赛灵思对任何因资料发生的或与资料有关的（含对资料的使用）任何损失或赔偿（包括任何直接、间接、特殊、附带或连带损失或赔偿，如数据、利润、商誉的损失或任何因第三方行为造成的任何类型的损失或赔偿），均不承担责任，不论该等损失或者赔偿是何种类或性质，也不论是基于合同、侵权、过失或是其他责任认定原理，即便该损失或赔偿可以合理预见或赛灵思事前被告知有发生该损失或赔偿的可能。赛灵思无义务纠正资料中包含的任何错误，也无义务对资料或产品说明书发生的更新进行通知。未经赛灵思公司的事先书面许可，贵司 / 您不得复制、修改、分发或公开展示本资料。部分产品受赛灵思有限保证条款的约束，请参阅赛灵思销售条款：http://china.xilinx.com/legal.htm#tos；IP 核可能受赛灵思向贵 司 / 您签发的许可证中所包含的保证与支持条款的约束。安全保护功能，不能用于任何需要专门故障安全保护性能的用途。如果把赛灵思产品应用于此类特殊用途，贵司 / 您将自行承担风险和责任。请参阅赛灵思销售条款：http://china.xilinx.com/legal.htm#tos。

汽车应用免责申明

汽车产品（产品部件号中标识为“XA”）不保证用于安全气囊的开发或用于影响车辆控制的应用（“安全应用”），除非在该赛灵思产品中具备故障安全保护或者额外功能，符合 ISO 26262 汽车安全标准（“安全设计”）。为安全起见，客户应在使用或分销任何集成有该产品的系统之前，对这些系统进行全面测试。在没有安全设计的安全应用中使用产品的风险完全由客户承担，仅受有关产品责任的适用法律和法规限制。