SAPER-AI加速器：一种基于脉动阵列的低能耗可重构人工智能加速器

Fahad Bin MUSLIM[1], Kashif INAYAT[2,3], Muhammad Zain SIDDIQI[1],
Safiullah KHAN[4], Tayyeb MAHMOOD[5], Ihtesham ul ISLAM[6]
[1]GIK工程科技学院计算机科学与工程系，巴基斯坦托皮，23460
[2]巴塞罗那超算中心，西班牙巴塞罗那，1-3 08034
[3]仁川国立大学电子工程系，韩国仁川市，22006
[4]曼彻斯特城市大学计算与数学系，英国曼彻斯特，M15 6BX
[5]Nextwave公司，韩国大田市，34134
[6]国立科技大学计算机软件工程系，巴基斯坦伊斯兰堡，H-12

摘要：深度学习加速器对于满足现代神经网络日益增长的计算需求至关重要。基于脉动阵列的加速器由二维网格状的处理元件组成，这些处理元件协同工作以加速矩阵乘法运算。此类加速器的能效至关重要，尤其是在边缘人工智能场景下。本文提出的SAPER-AI加速器是一种脉动阵列加速器，采用统一功耗格式来定义其功耗意图，几乎无需优化其微架构。该加速器以粗粒度方式关闭处理元件的行和列，从而使脉动阵列微架构能适应现代深度学习工作负载不断变化的计算需求。分析表明，在最佳情况下，32×32和64×64脉动阵列设计的能效分别提升10%至25%。此外，更大规模的脉动阵列的功率延迟积改善约6%。进一步，基于MobileNet和ResNet50模型的性能比较表明，脉动阵列在处理ResNet50工作负载时通常表现更佳。这是因为ResNet50所呈现的更为规整的卷积计算更受脉动阵列青睐，且随着脉动阵列规模增大，这一性能差距会进一步扩大。

关键词：人工智能加速器；专用集成电路设计；脉动阵列；低能耗设计
https://doi.org/10.1631/FITEE.2400867