深度学习加速芯片架构设计中数据通路存储层次协同优化研究

王颖 韩鋈 徐新宇

中国电子科技集团第五十八研究所，江苏省无锡市，214000

摘要：本文聚焦深度学习加速芯片架构设计，针对传统芯片架构在处理深度学习任务时，数据通路存储层次协同效率低下的问题展开深入探索。深度学习算法计算量大、数据访问模式复杂，对芯片计算性能和存储能力要求严苛。通过构建基于深度学习算法特性的协同优化模型，在数据通路设计上，采用脉动阵列结构等优化矩阵运算路径，提升计算单元效率；存储层次方面，运用优化缓存策略、引入高带宽内存等技术，提高数据存储与访问性能。实验表明，协同优化方案显著提升了芯片性能，以AlexNet模型为例，训练时间缩短66.7%，推理时间减少80%，能耗降低40%，存储利用率大幅提高。

关键词：深度学习加速芯片；数据通路；存储层次；协同优化

作者简介：王颖(1991-),男,汉族,江苏省无锡市,硕士研究生,工程师,研究方向:芯片设计。

1 引言

随着深度学习在计算机视觉、自然语言处理、语音识别等众多领域取得显著成果，其对计算性能的需求呈指数级增长。深度学习模型，如卷积神经网络

(CNN)、循环神经网络(RNN)及其变体，包含海量参数和复杂的计算逻辑，传统的中央处理器(CPU)和图形处理器(GPU)在处理这些任务时，逐渐暴露出计算效率低下、能耗过高的问题，不仅限制了深度学习技术的快速发展，也阻碍了其在资源受限设备中的广泛应用。

数据通路作为芯片中数据传输和处理的通道，其效率直接影响芯片的整体性能；存储层次则负责数据的存储与管理，不同层次的存储在容量、速度和成本上存在差异。如何构建高效的存储层次结构，实现数据的快速访问和存储，是提高芯片性能的关键。然而，在传统的深度学习加速芯片架构设计中，数据通路存储层次往往独立设计，缺乏有效的协同优化，导致数据在传输与存储过程中产生大量延迟和能耗，严重制约了芯片的计算性能和能效比。

通过优化深度学习加速芯片架构设计中的数据通路存储层次协同工作机制，可以显著提高芯片的计算效率，降低能耗，从而加速深度学习模型的训练与推理过程，推动深度学习技术在更多领域的应用与发展，具有广泛的应用前景和经济价值。

2 深度学习加速芯片架构基础

深度学习算法中，卷积神经网络(CNN)与循环神经网络(RNN)及其变体应用广泛。CNN通过卷积层、池化层和全连接层处理图像，卷积核滑动提取局部特征，实现权值共享；池化层降维增强鲁棒性，如AlexNet可精准提取图像特征。RNN用于处理序列数据，但其处理长序列存在梯度问题，LSTM通过门控机制、GRU通过简化结构解决该问题，在机器翻译等任务中表现出色。这些算法计算复杂，要求芯片具备强并行计算、高效数据处理与快速存储访问能力。

传统CPU和GPU架构处理深度学习任务存在明显缺陷。CPU核心少、并行计算能力弱，内存带宽低，面对深度学习大量矩阵运算和数据读写时效率低下。GPU虽适合并行计算，但资源分配不灵活、能耗高，固定的存储层次难以适配深度学习动态数据访问需求，导致存储访问延迟大，影响任务执行速度。随着深度学习模型发展，这些局限日益凸显。

深度学习加速芯片架构专为高效处理深度学习任务设计，一般包含计算单元、存储单元、数据通路和控制单元。计算单元含大量并行核心执行运算，如TPU的矩阵乘法单元提升计算效率；存储单元分不同层次满足速度和容量需求；数据通路优化结构与带宽减少传输延迟；控制单元协调各单元工作，按需分配资源[1]。其工作原理基于对深度学习算法的优化，如采用脉动阵列实现高效矩阵乘法，优化存储层次、利用数据压缩技术等提升性能。相比传统架构，它在计算效率、能耗和存

储带宽利用上优势显著，如谷歌TPU在推理任务中计算效率远超传统GPU，成为推动深度学习发展的关键[2]。

3数据通路设计与优化

数据通路在深度学习加速芯片中负责数据在存储和计算单元间的传输，由算术逻辑单元（ALU）、寄存器组、数据总线构成。ALU执行深度学习中的矩阵乘法等基本运算；寄存器组临时存储数据，减少访问延迟，支持模型计算；数据总线决定传输速度，深度学习加速芯片常采用高速总线满足数据传输需求。各部件协作完成数据处理，如模型前向传播时，数据经存储单元到寄存器组，再由总线传至ALU计算，结果返回寄存器组并存储。

流水线技术将复杂计算分解为多阶段，各阶段由专门部件执行，中间插入寄存器暂存结果，提升数据通路吞吐量和处理速度，在深度学习卷积和矩阵乘法运算中广泛应用，如将卷积运算分阶段，使多个计算任务可同时处理。并行处理技术通过增加计算资源，如多个ALU等，让多个任务并行处理，契合深度学习大量矩阵运算和并行操作的特点，还可采用多线程技术提升效率。综合运用这些方法技术，能设计出满足深度学习高性能需求的数据通路[3]。

深度学习任务计算特殊、数据流动复杂，对数据通路要求高。模型中矩阵和卷积运算多，数据读写量大，访问模式不规则。为满足需求，可优化矩阵运算路径，采用脉动阵列结构减少数据传输存储开销。还能优化数据通路结构，用高速缓存减少片外内存访问，设计高速总线或多总线结构提升带宽，利用数据压缩技术降低传输量，以此提高数据通路性能，满足深度学习对计算效率和数据处理的要求[4]。

4存储层次设计与优化

在深度学习加速芯片里，存储层次结构负责数据存储与管理，不同层级存储部件协同，满足芯片对数据存储容量、速度、成本上的多样需求[5]。寄存器紧邻计算单元，访问速度极快但容量小，用于暂存中间结果、模型参数，支撑快速运算。缓存（Cache）处于CPU和主存间，基于SRAM技术，通过预存数据减少CPU访主存次数，提升访问速度。主存采用DRAM技术，容量大但访问速度慢，存储模型完整参数、训练及推理数据。辅助存储器（如硬盘、SSD）容量巨大，用于长期存储海量数据和程序，但访问速度极慢，用时先加载到主存。

存储层次的访问速度、存储容量和带宽等性能指标对深度学习加速芯片至关重要。访问速度方面，寄存器最快，缓存次之，主存较慢，辅助存储器访问最慢，存储层次访问速度慢会延长深度学习模型训练时间。存储容量方面，寄存器和缓存容量有限，主存满足基本存储需求，辅助存储器容量巨大，用于长期存储大规模数据，存储层次容量不足会影响芯片性能。带宽反映数据传输能力，带宽不足会增加数据传输延迟，拖慢计算速度。高效的存储层次结构能快速提供数据、合理存储，并保证数据快速传输。

深度学习对存储层次要求高，需采用多种优化方法。优化缓存策略，改进传统LRU算法，根据深度学习任务特征动态调整缓存数据，优化缓存关联性和大小；采用高带宽内存，如DDR4、DDR5或多通道内存技术，提升数据传输速率；利用新型存储技术，如NVM（PCM、RRAM），作为主存或缓存补充，采用分布式存储架构，提高存储系统可靠性和扩展性。

5数据通路与存储层次协同优化策略

传统设计中数据通路和存储层次独立，导致数据传输延迟和访问冲突。数据通路与存储层次协同优化旨在提高计算效率和降低能耗。通过优化协同机制，减少传输延迟和访问冲突，提升计算单元利用率，减少数据冗余传输和存储。

数据预取技术依据数据访问模式和计算需求，预测下一次计算所需数据并提前预取到缓存。缓存一致性维护（如MESI协议）保证数据在不同缓存层次和计算单元间一致，提升系统可靠性和性能。优化存储与计算单元接口和通信机制，根据深度学习任务特点，合理分配存储和计算资源，实现高效协同。

结合深度学习算法特点构建协同优化模型，包含任务分析模块（负责分析深度学习任务类型、结构和计算需求，提取关键特征）、资源分配模块（依据任务分析结果，动态分配数据通路和存储层次资源）、调度模块（协调数据通路和存储层次工作，按任务执行顺序和资源分配情况，合理调度数据传输和计算操作）和反馈模块（根据任务执行结果调整优化模型），实现高效协同，提升深度学习加速芯片性能。

6实验与验证

实验选用CNN中的AlexNet和RNN中的长短时记忆网络（LSTM）模型，对比优化前后芯片运行性能。实验环境基于FPGA开发板和高性能服务器，用Verilog设计实

现芯片架构，借助Xilinx ISE开发工具进行综合、仿真，使用TensorFlow框架运行深度学习模型。

以AlexNet模型为例，优化前累计训练时长30小时，在执行推理任务时，单次推理所需时间为150毫秒。而在采用数据通路存储层次协同优化策略之后，训练时间锐减至10小时（缩短66.7%），推理时间仅需30毫秒（减少了80%）。性能提升的关键原因在于，协同优化有效地缓解了数据在存储层次与数据通路之间传输时产生的延迟问题，极大地降低了存储访问冲突的发生频率。

能耗指标的测试选取了LSTM模型。当LSTM模型运行在未优化的芯片架构上时，芯片的能耗稳定保持在200瓦。而在引入协同优化策略后，能耗值降至120瓦（降低了40%）。这主要是因为优化策略对数据传输和存储操作进行了深度优化，摒弃了大量不必要的数据搬运以及存储动作。通过减少这些冗余操作，从根本上降低了芯片在运行深度学习任务过程中的整体功耗，这对于提升芯片的能源利用效率、降低运行成本以及推动绿色计算具有重要意义。

在存储利用率的分析中，本文重点关注存储访问次数和缓存命中率这两个关键指标。优化前每完成一次深度学习任务，存储访问操作极为频繁，平均访问次数高达2000次，其中对主存的访问次数就多达1600次。而优化后，存储访问次数下降至1000次，主存访问次数减少到600次。主存访问次数占比从原本的80%下降至60%。

缓存命中率方面，未优化时缓存命中率仅为20%，但协同优化策略通过引入数据预取技术和缓存一致性维护技术，确保了缓存中数据的准确性与有效性，避免了数据不一致问题导致的计算错误，缓存命中率提升至50%，提升幅度高达150%，降低了存储访问延迟和带宽占用，有效提升存储资源的利用效率，优化了芯片的整体性能。

将本文协同优化方案与英伟达GPU架构和谷歌的TPU架构进行对比分析。计算性能上，处理卷积运算密集的图像识别任务时，本方案与英伟达GPU架构相当甚至更优，但GPU软件生态完善，通用性更强；与谷歌TPU架构比，本方案更具灵活性，但TPU在特定框架和任务优化程度上效率更高。能耗方面，本方案比GPU架构优势

明显，能耗更低；与TPU架构比，能耗表现相当，不同应用场景下能耗有差异。存储利用率上，本方案通过优化存储层次结构和数据访问策略，存储资源利用率优于现有方案，但在存储带宽利用效率上，不如专门优化存储带宽的方案。

7结论

本文针对深度学习加速芯片架构，实现数据通路存储层次协同优化，取得了一系列具有重要价值的研究成果。在数据通路优化上，基于深度学习算法数据特性，提出优化策略。通过采用新型架构与并行计算单元，提升数据处理与传输效率；优化传输总线，有效降低数据传输延迟，保障深度学习任务高效运行。存储层次优化方面，研究各层级性能差异，构建合理架构。通过优化缓存参数，引入高带宽内存、NVM等技术，提高数据存储访问效率，减少存储冲突，满足深度学习对数据存储的严苛要求。探索出数据通路存储层次的协同机制，提出基于任务动态特征的自适应协同优化方法。该方法可根据任务和运行状态，智能调配工作模式与资源，实现两者深度协同。构建的协同优化模型，验证了方法的有效性。实验采用AlexNet和LSTM模型验证，结果表明，协同优化后的芯片架构计算性能显著提升，训练与推理时间大幅缩短，能耗降低，存储利用率提高，具有良好的应用前景。

参考文献

- [1] Shirmohammadi Z, Taali M. JoBiS: Joint capacitance and inductance bit stuffing CAC for interposer based multi-chip Deep Learning Accelerator[J]. Integration, 2025, 102102347-102347
- [2] Taub L, Hampton H T, Sarkar S, et al. E.PathDash, pathway activation analysis of publicly available pathogen gene expression data. [J]. mSystems, 2024, e0103024.
- [3] 申云飞. 深度学习编译器关键技术研究与应用[D]. 安徽大学, 2023.
- [4] 龚烽. CIS片内神经网络加速器的设计与建模[D]. 上海交通大学, 2022.
- [5] 鲁蔚征, 张峰, 贺寅烜, 等. 华为昇腾神经网络加速器性能评测与优化[J]. 计算机学报, 2022, 45(08): 1618-1637.