3D 混合键合集成的软件定义近内存架构

1 引言

1.1 背景需求

随着人工智能、物联网、大数据等新兴应用的快速发展，全球数据量呈爆炸式增长，推动着对高性能计算芯片的持续需求。然而，传统依赖于工艺节点演进的计算能力提升方式已逐渐失去动力。摩尔定律和 Dennard 缩放定律的失效，使得芯片性能提升面临“功耗墙”与“存储墙”的双重限制，能效问题日益成为限制系统性能进一步提升的关键瓶颈。与此同时，计算需求的多样化也提出了对芯片架构更高的灵活性要求。尤其是在人工智能与数字信号处理等领域，算法快速演化，传统固定功能硬件难以适应，导致硬件设计周期长、非重复性工程成本高，严重影响了应用部署的效率与经济性。因此，兼具高能效与高灵活性的计算架构已成为当前芯片设计的重要研究方向。

1.2 问题描述

目前主流的计算架构在能效与灵活性之间存在难以兼顾的矛盾。应用专用集成电路具有极高的能效比，但其硬件结构在设计完成后即固定，缺乏灵活性，生命周期短；通用处理器与图形处理器则通过软件进行指令调度，具备较强灵活性，但受限于冯·诺依曼架构下的数据搬移瓶颈，能效表现较差；现场可编程门阵列

提供了一定程度的硬件可编程能力，但其编程粒度细、资源抽象复杂，导致功耗高、使用门槛高，不适用于大规模商业部署。此外，传统计算架构中处理器与存储器的分离设计导致频繁的数据搬移，不仅带来显著的功耗开销，还限制了系统带宽。研究表明，在 40 纳米工艺下，计算操作的能耗仅占总体能耗的不足 10%，其余 90% 以上来自于数据的搬移与存储访问。由此可见，在优化计算能效的同时，若不能解决“存储墙”问题，系统整体性能仍将受限。

1.3 研究动机

为有效应对上述挑战，业界提出了软件定义芯片与近存计算等新型架构理念。SDC 通过引入可编程硬件与高层软件协同控制机制，在运行时实现计算结构的动态重构，兼具高灵活性与高能效，适用于快速变化的应用场景。近存计算则通过将部分计算功能前移至内存附近，显著降低数据搬移的能耗，提高带宽利用率。尽管 SDC 和近存计算各自具有明显优势，但目前尚缺乏一种有效融合两者优点、同时具备可制造性与实用性的架构实现。为此，本文提出一种基于三维混合键合集成的软件定义近存处理架构。该架构通过上下协同的方式，在硬件层面实现逻辑与存储的高效耦合，在软件层面引入动态可重构机制，实现数据密集型与计算密集型任务的统一加速，具有广阔的应用前景与重要的研究价值。

2 研究目标与内容

为应对当前高性能计算领域在能效、灵活性与可制造性方面所面临的挑战，本文围绕软件定义近存处理架构展开研究，提出一种融合软件可编程、硬件可重构及 3D 集成互联的新型芯片体系结构。该研究旨在突破传统芯片架构在计算与存储解耦、硬件固定化、数据搬移成本高等方面的瓶颈，提升系统在数据密集与计算密集场景下的处理效率与资源利用率。围绕这一核心目标，本文主要开展以下三个方面的研究工作：其一，设计并实现异构可重构的处理单元阵列，构建面向任务特性优化的空间并行计算平台；其二，采用成熟的三维混合键合工艺将逻辑与 DRAM 芯片垂直集成，构建低能耗、高带宽、强耦合的逻辑-存储协同体系；其三，开发高效、自动化的编译工具链，支持从高级语言至底层配置的自动映射与调度，实现用户对硬件资源的“无感”使用。上述研究内容相互支撑，共同构成 SDPNM 架构的软硬一体化解决方案。

2.1 异构处理阵列设计

本研究提出的 SDPNM 芯片采用多类型的处理阵列架构，由五个处理阵列构成，包括针对低复杂度控制类任务的 NPEA、面向乘法密集型信号处理的 FPEA，以及优化用于深度学习与高并行计算的 SPEA。每个处理阵列内包含 64 个可重构处理单元，支持不同粒度的指令集与数据通路结构，能够在运行时根据任务特征动态组合与重构，充分实现空间并行与任务异构处理。该设计

既保障了运算能效，也显著提升了任务适配能力。

2.2 三维混合健合集成结构

为解决传统“处理器-内存”分离架构所造成的数据搬移瓶颈，本文采用先进的三维混合键合技术将逻辑层与 DRAM 层进行垂直集成。逻辑芯片与内存芯片分别使用 40nm 和 25nm 工艺制备，并通过 Cu-Cu 互联在低于 400℃ 温度下完成高密度、低电阻的金属键合。该结构不仅缩短了逻辑与内存之间的互连路径，提升带宽密度，同时具备良好的兼容性与可制造性，为低成本实现高能效逻辑-存储协同提供技术保障。

2.3 自动化编译系统研发

为了释放芯片潜力并降低开发复杂度，本文设计了一套基于 LLVM 的自动化编译系统。该系统支持将高层算法程序自动转化为适配底层 PE 阵列结构的配置上下文，编译流程涵盖任务划分、数据依赖建模、数据流图优化、图映射调度与机器码生成等环节。通过引入整数线性规划（ILP）与模块化调度机制，编译器能够实现高质量的指令调度与资源分配，避免冗余通信与内存冲突，最终输出高性能、低能耗的运行配置。该系统支持秒级编译时间，并具有良好的可移植性与扩展性。

3 技术方案

为实现软件定义的近存计算目标，本文设计并实现了一套完整的软硬协同、上下贯通的技术体系，涵盖了硬件系统结构、软件编译工具链以及逻辑-存储集成方式等多个层面。在硬件方面，提出一种采用三维混合键合集成的异构计算架构，通过构建多类型处理阵列与高带宽内存接口，实现计算与存储的协同优化；在软件方面，设计了基于 LLVM 的自动化编译系统，通过多层优化、数据流分析与调度映射，将高层算法高效映射至底层硬件资源，提升编程效率与系统性能。以下对各个关键技术点进行具体阐述：

3.1 异构处理阵列架构设计

SDPNM 架构采用空间可重构的处理阵列结构，整体由五个处理阵列（PEA）组成，划分为三种类型：NPEA（用于低复杂度定点计算）、FPEA（用于浮点或乘法密集型运算）、SPEA（负责深度学习类并行计算）。每个 PEA 内包含 64 个处理单元，支持包括加法、乘法、FFT、卷积、复数乘加等操作，通过不同的指令子集实现对不同计算类型的加速。异构 PEA 的设置实现了“按需调度、精细分工”的架构优化策略，提升了面积与能耗利用率。阵列之间采用可配置数据路径互连，便于灵活构建适配算法的数据流结构，异构加速效率增益公式如下所示：

$$\eta_{\text{hetero}} = \frac{\sum_{i \in \{\text{N,F,S}\}} n_i \cdot \alpha_i \cdot \beta_i}{\text{Area}_{\text{total}}} \cdot \frac{1}{\epsilon_{\text{power}}} \tag{1}$$

3.2 逻辑-存储三维混合键合集成

为突破“存储墙”限制，本文采用 3D 混合键合技术实现逻辑层与 DRAM 层的紧密集成。逻辑芯片采用 SMIC 40nm 工艺制造，DRAM 芯片采用 PSMC 25nm 工艺制造，二者面积一致，并通过 XMC 提供的 Cu-Cu 键合工艺进行垂直集成。该集成方式具备高密度、低电阻、低温加工等优势，相比传统微凸点方式具有更短的互连距离和更高的带宽密度。实验显示，该芯片的逻辑-存储接口带宽可达 287 GB/s，能耗仅为 0.88 pJ/bit，显著优于主流 HBM 方案，互连带宽密度公式如下所示：

$$B_{\mathrm{3D}} = \frac{N_{\mathrm{bonds}} \cdot f_{\mathrm{signal}} \cdot W_{\mathrm{bit}}}{A_{\mathrm{chip}}} \tag{2}$$

此技术路径同时具备成熟度高、制造成本低的工程优势，有助于推动 SDPNM 架构的大规模实际应用。

3.3 灵活可重构互连机制

SDPNM 中的 PE 阵列之间通过可配置互连单元连接，支持多种拓扑模式（如一维、二维 Mesh），并可动态配置数据传输路径与运算逻辑。每个 PE 具备独立的配置上下文，支持输入、输出来源的灵活切换（来自共享内存、本地寄存器、其他 PE 等）。互连机制不仅支持高效的点对点通信，还具备良好的局部数据复用能力，适配多种计算场景如卷积、矩阵乘、FFT 等。系统整体呈现出高度的空间并行性和数据流导向特性，为高吞吐量、低延迟的任务调度提供支撑。

3.4 自动化编译系统设计

为了降低系统开发复杂度，本文构建了一套完整的自动化编译工具链。该系统基于 LLVM 编译框架开发，支持将高级语言编写的算法自动转译为可在 SDPNM 芯片上运行的配置上下文。编译过程包括：任务划分、数据依赖分析、数据流图生成、基本块合并、图映射与调度、指令生成与编码等多个阶段。其中，映射阶段使用整数线性规划模型对数据流图进行最优调度，并支持将数据依赖自动映射至互连结构或共享内存，显著提升运行效率。该编译器支持秒级编译时间，具备高度可移植性与开发友好性，远优于传统 FPGA 工具链。

4 实验设计

为验证本文提出的软件定义近存处理架构（SDPNM）的可实现性与性能优势，本文设计并完成了原型芯片的流片制造、系统集成与功能评估流程，并构建了完整的实验平台用于测试与分析。实验设计主要围绕以下几个目标展开：（1）验证 SDPNM 芯片在物理集成层面的工艺可行性与接口可靠性；（2）评估处理器阵列在典型计算任务中的性能表现与能效水平；（3）测试自动化编译系统在实际任务下的映射效率与可用性。为此，实验分为原型芯片实现、评估平台搭建、基准任务设计与测试流程组织等几个阶段，确保对芯片架构、编译系统和实际应用能力进行全面评估。

4.1 原型芯片设计与流片实现

本研究完成了 SDPNM 芯片的逻辑层与 DRAM 层的版图设计与流片制造。逻辑芯片基于 SMIC 40nm 工艺制备，内含五个异构处理阵列、配置控制模块及内存接口逻辑；DRAM 芯片采用 PSMC 25nm 工艺制造，提供总容量为 1 GB 的存储资源。二者通过 XMC 提供的混合键合技术实现三维集成，采用 Cu-Cu 互联方式构建高密度、低能耗的逻辑-存储接口，互连面积与位置在物理设计阶段严格对齐以保障键合精度。最终流片芯片尺寸为 16.96 mm × 11.87 mm，逻辑与存储层频率均为 200 MHz，键合接口带宽达到 287 GB/s，逻辑-内存访问能耗为 0.88 pJ/bit。

4.2 评估平台搭建与通信接口设计

为实现芯片调试与应用评估，本文搭建了一套完整的评估开发板。该开发板为多层 PCB 设计，支持 JTAG 调试、串口通信与 USB 数据传输功能，配套连接 PC 端用于编程下载与数据回传。主控电源由 12 V DC 输入，通过板载电压转换模块提供 1.1 V（核心）与 3.3 V（IO）工作电压，确保逻辑与 DRAM 工作稳定。芯片插座采用高精度封装支架，支持原型芯片直接更换。板上还搭配外部时钟源、配置寄存器与电平转换器，适用于不同工作模式下的信号输入输出需求。

4.3 基准测试任务设计与映射实现

为了全面评估 SDPNM 架构在不同任务类型下的性能表现，本文选取了代表性的计算密集型任务与数据密集型任务进行测试。这些任务均通过本文提出的自动化编译系统完成从高层程序到 PEA 配置上下文的转化，采用统一测试接口加载至芯片中运行。测试过程中记录任务执行时间、吞吐量与能耗，支持与 FPGA 及 DSP SoC 进行横向对比，量化系统在实际应用中的综合优势。

4.4 测试流程与数据采集方式

实验测试采用“输入—加载—运行—输出”四阶段流程。输入数据通过串口或 USB 接口发送至芯片内存中，随后通过 JTAG 发送配置上下文启动任务运行；运行过程由板载控制逻辑自动管理状态切换与时序控制；任务完成后，输出数据通过串口返回至 PC 进行比对验证。能耗测试通过片外功率监测模块测量实际运行过程的电流变化，结合已知工作电压进行计算。为确保数据准确性，每项任务均重复运行多次，取平均值作为最终结果。

5 结果分析

在完成 SDPNM 原型芯片的流片制造与评估平台搭建后，本文围绕芯片物理特性、编译系统性能以及应用任务执行效果进行了系统的实验测试与数据对比分析。测试结果充分验证了所提架构在能效、带宽利用率、可编程性及应用适配性等方面的优势。与主流 FPGA 与 DSP 芯片相比，SDPNM 在多个典型任务中表现出

显著的能效提升，尤其在数据搬移频繁的应用场景下，其混合键合带来的存储耦合优势得到充分体现。以下从芯片性能指标、编译效率表现及算法运行结果三个方面进行具体分析。

5.1 芯片物理性能指标

SDPNM 芯片在制造完成后进行了全面的功能验证与性能测试。逻辑芯片与 DRAM 芯片通过 Cu-Cu 混合键合实现紧密集成，实际测试显示逻辑—存储接口带宽达到 287 GB/s，远高于传统 DDR 或 HBM 接口。能耗方面，数据传输平均消耗为 0.88 pJ/bit，优于多种在研的近存或存内计算结构。在 200 MHz 工作频率下，系统功耗稳定维持在 1.5 W 以内，逻辑与存储分别约为 0.5 W 和 1 W，符合高能效计算芯片的预期要求，表明该架构具备良好的硬件可实现性与功耗控制能力。

5.2 编译系统效率与可用性

本文提出的自动化编译工具链在多个典型算法上进行测试，与主流 FPGA 工具链和已有软件定义芯片（SDC）编译器进行比较。结果表明，SDPNM 编译系统在不牺牲映射质量的前提下，将编译时间缩短至秒级，明显优于 FPGA 工具链数小时的综合与布局时长，也优于部分专用 CGRA 工具链的分钟级编译过程。编译系统支持 C 语言高层程序输入，自动完成任务划分、数据依赖分析、图调度映射与二进制生成，用户无需手动干预底层配置，大大降低了开发门槛，提升了芯片的可编程性与实用性。

5.3 算法运行性能与能效对比

在应用层面，SDPNM 对多个代表性计算任务进行了映射与运行测试，并与 FPGA 和 DSP SoC 进行了对比分析。结果如表所示，在数字下变频系统中的关键模块如 NCO、CIC、FIR，以及通用计算任务如 FFT、GEMM 和 YOLO V3 中，SDPNM 均展现出优异的能效性能。例如，在 FIR 滤波器任务中，SDPNM 达到 372 GOPS 的吞吐率，能效达 530 GOPS/W，为 FPGA 的 29.4 倍，为 DSP 的 14.3 倍；在 YOLO V3 任务中，SDPNM 能效为 148 GOPS/W，较 FPGA 提升 8.2 倍，较 DSP 提升 3.08 倍。整体而言，SDPNM 在多个应用中平均能效提升为 FPGA 的 33.1 倍，最高可达 104.1 倍，充分验证了其在能效敏感型任务中的应用潜力，性能和能效方面的比较如下图所示：

	Throughput (GOPS)			Energy efficiency (GOPS/W)		
	FPGA	DSP [27]	This work	FPGA	DSP [27]	This work
NCO	900	40	233	18	36	331
CIC	900	40	117	18	53	166
FIR	900	40	372	18	37	530
FFT	63	8	93	1.26	7	131
GEMM	900	40	372	18	36	530
YOLO V3	900	40	104	18	48	148

图 1: 性能和能效方面的比较

本文针对当前计算系统在能效、灵活性与存储访问效率方面面临的挑战，提出了一种基于三维混合键合集成的软件定义近存处理架构。该架构通过异构可重构处理阵列与存储资源的深度协同，构建了一个兼具高能效、高带宽与高灵活性的计算平台。通过引入空间可重构硬件设计、高密度逻辑-存储互联结构以及自动化编译工具链，本文实现了从体系结构、物理实现到软件支持的全栈创新，具备良好的可制造性与实际应用前景。在硬件层面，本

文完成了 SDPNM 芯片的流片制造与评估板搭建，实测结果显示其逻辑-存储接口带宽达 287 GB/s，能效为 0.88 pJ/bit，优于当前主流高带宽存储解决方案。在软件层面，基于 LLVM 框架构建的编译系统可实现秒级高效编译，并成功将多个代表性算法映射到 SDPNM 平台上，达到与手工配置近似的运行性能。在实际任务验证中，SDPNM 在多项计算与数据密集型应用中的能效表现均显著优于主流 FPGA 与 DSP 系统，平均提升超过 30 倍，展现出其作为下一代低功耗高性能计算平台的巨大潜力。尽管本研究已在架构设计、物理实现与系统评估等方面取得了较为系统的成果，但仍存在若干值得进一步研究的方向。首先，当前 SDPNM 架构主要面向典型的可预测数据流任务，未来可考虑引入分布式存储调度机制，以支持更复杂的数据访问模式与控制密集型应用。其次，编译系统可进一步引入机器学习辅助优化与跨层自适应调度机制，提升对大规模应用的映射质量与编译效率。最后，在系统扩展性方面，可研究多芯片互联与片间通信协议设计，推动 SDPNM 向多核异构集群方向演进，以满足更高性能计算场景的需求。