工程与管理科学
Engineering and Management Science
第 7 卷◆第 10 期◆版本 1.0◆2025 年
文章类型：论文|刊号（ISSN）：2705-0637(P) / 2705-0645(O)

面向高级别自动驾驶的存算一体芯片架构设计与能效优化

黄小如
广州华立科技职业学院　511325

DOI：10.12238/ems.v7i10.15751

[摘　要] 随着高级别自动驾驶技术的快速发展，传统冯·诺依曼架构芯片在算力和能效方面面临严峻挑战。存算一体技术作为突破“存储墙”和“功耗墙”的关键路径，为自动驾驶芯片设计提供了新思路。本文深入分析了高级别自动驾驶对芯片的算力与能效需求，系统阐述了存算一体技术的原理与优势，并从架构设计、能效优化策略、关键技术实现三个维度展开研究。通过引入混合精度计算、新型存储介质、高效片上网络等创新技术，提出了一种面向自动驾驶场景的存算一体芯片架构方案。实验结果表明，该架构在典型自动驾驶任务中实现了算力密度与能效比的显著提升，为高级别自动驾驶的商业化落地提供了技术支撑。

[关键词] 高级别自动驾驶；存算一体；芯片架构设计；能效优化；混合精度计算

一、引言

高级别自动驾驶（L3 - L5 级）的商业化进程正加速推进，其核心挑战在于如何实时处理海量传感器数据并做出安全决策。以城市道路场景为例，自动驾驶车辆需同时处理 8 个以上摄像头、1 个激光雷达及多个毫米波雷达的数据，每秒数据量可达数 GB，且决策延迟需控制在毫秒级。传统冯·诺依曼架构芯片因“存储墙”和“功耗墙”问题，难以满足这一需求：一方面，处理器与存储器间的数据搬运能耗占比高达 63.7%（7nm 工艺下），导致整体能效低下；另一方面，存储带宽限制使算力利用率不足 40%，制约了复杂算法的部署。存算一体技术通过将存储与计算功能融合，直接在存储单元内完成数据运算，可消除数据搬运开销，理论上能将单位算力能耗降低一个数量级。本文聚焦高级别自动驾驶场景，系统研究存算一体芯片的架构设计与能效优化方法，旨在为下一代自动驾驶计算平台提供理论支撑与技术方案。

二、高级别自动驾驶的芯片需求分析

2.1 算力需求增长

高级别自动驾驶等级与算力需求呈指数级增长。L2 级辅助驾驶需 20－100TOPS 算力，主要处理车道保持、自适应巡航等简单任务，其算法相对简单，对实时性要求虽高，但计算量相对较小。L3 级条件自动驾驶需 100 - 500TOPS，需支持高速领航、自动变道等功能，此时算法复杂度增加，不仅要处理感知信息，还需进行一定程度的决策规划。L4 级高度自动驾驶需 500 - 1000TOPS，要应对城市复杂路况，如交通信号灯识别、行人避让、与其他车辆的交互等，对环境感知的精度和决策的准确性要求极高。L5 级完全自动驾驶则需数千 TOPS 算力，以覆盖所有极端场景，包括恶劣天气、突发事故等，算法需具备高度的自适应和学习能力。以特斯拉 FSD 芯片为例，其算力达 144TOPS，可支持 L2＋ 级功能，但面对 L4 级场景时仍需进一步提升。随着自动驾驶等级的提升，传感器数量和种类不断增加，数据量呈几何级数增长，对芯片的算力提出了巨大挑战。

2.2 能效比要求

自动驾驶芯片的能效比直接影响车辆续航与散热设计。传统架构芯片在处理 BEV（鸟瞰图）网络时，数据搬运能耗占比超 70%，导致整体能效低于 5TOPS/W。例如，某款基于 GPU 的自动驾驶计算平台，在 500TOPS 算力下功耗高达 500W，需配备复杂液冷系统，显著增加成本。存算一体架构通过消除数据搬运，可将能效比提升至 15TOPS/W 以上，为车载边缘计算提供可行方案。

2.3 实时性、鲁棒性与灵活性需求

自动驾驶场景具有高度动态性与不确定性，要求芯片具备实时性、鲁棒性与灵活性。实时性要求决策延迟需<100ms，以应对突发状况，如前方车辆突然刹车、行人突然闯入道路等。若芯片处理延迟过高，可能导致自动驾驶系统无法及时做出反应，引发安全事故。鲁棒性要求需在 - 40℃至 85℃

温度范围内稳定工作，同时要能抵抗电磁干扰、振动等外界因素的影响。在不同的地理环境和气候条件下，芯片都要保持可靠的性能。灵活性要求支持算法快速迭代与 OTA 升级，随着自动驾驶技术的不断发展，算法需要不断优化和改进，芯片应能够快速适应这些变化，通过软件更新实现功能的升级和扩展。传统 ASIC 芯片因功能固化，难以适应算法演进；而存算一体架构通过可重构计算单元设计，可兼顾性能与灵活性。

三、存算一体技术原理与优势

3.1 存算一体架构原理

存算一体架构通过修改存储单元的“读”电路，在数据读取过程中完成计算操作。以 SRAM 存内计算为例，其核心机制包括：在 6T SRAM 单元中增加 2 个晶体管，构建 8T 计算单元，支持 MAC（乘加）运算；将权重数据预存于存储阵列，输入数据通过字线（Word Line）广播，实现并行计算；采用模拟域累加与数字域量化，平衡精度与能效。在传统的冯 •诺依曼架构中，数据需要在存储器和处理器之间频繁搬运，而存算一体架构将计算逻辑嵌入到存储单元中，使得数据可以在存储的同时进行计算，大大减少了数据搬运的次数和时间，从而提高了计算效率和能效。

3.2 存算一体架构分类

存算一体架构可分为近内存计算(NMC)与存内计算(IMC)两类。NMC 通过 2.5D/3D 封装技术拉近存储与计算距离，如 HBM + GPU 方案，可提升带宽至 1TB/s，但数据搬运能耗仍占 30%。这种架构在一定程度上缓解了存储墙问题，但并没有完全消除数据搬运带来的能耗和延迟。IMC 完全融合存储与计算功能，如后摩智能鸿途 H30 芯片，采用数字存算一体架构，在 12nm 工艺下实现 256TOPS 算力，功耗仅 35W。IMC 架构将计算单元直接集成在存储阵列中，实现了真正的存算融合，具有更高的能效和更低的延迟。

3.3 存算一体在自动驾驶中的优势

存算一体架构在自动驾驶场景中具有显著优势。能效比提升方面，消除数据搬运后，访存功耗降低 90%，整体能效比提升 5－10 倍。在自动驾驶中，大量的数据需要不断进行处理和计算，减少访存功耗可以显著降低芯片的整体能耗。算力密度提高方面，单位面积算力可达传统架构的 20 倍，适合车载空间受限场景。车载芯片需要在有限的空间内实现高性能的计算，存算一体架构的高算力密度可以满足这一需求。延迟降低方面，计算与存储紧密耦合，端到端延迟<10μs，满足实时性要求。

四、面向自动驾驶的存算一体芯片架构设计

4.1 整体架构概述

本文提出一种分层异构存算一体架构，包含感知层、计算层、存储层与通信层。感知层集成 8 个摄像头接口与 1 个激光雷达接口，支持多模态数据融合。不同传感器的数据具有不同的特点和格式，通过感知层的设计可以实现数据的有效整合和预处理。计算层采用存算一体核心（IPU）与通用 CPU 协同设计，IPU 负责 AI 加速，CPU 处理逻辑控制。IPU 可以高效地处理自动驾驶中的深度学习算法，而 CPU 则可以完成一些控制和管理任务。存储层基于 SRAM 的存内计算阵列（128MB）与 DDR5 外存（8GB）结合，平衡容量与速度。SRAM 存内计算阵列用于存储频繁访问的数据和中间结果，提高数据访问速度；DDR5 外存则用于存储大量的模型参数和历史数据。通信层采用 NoC（片上网络）实现低延迟数据交换，带宽达 512GB/s，确保各层之间数据的高效传输。

4.2 存算一体核心（IPU）设计

存算一体核心（IPU）采用混合精度设计，支持 INT8/FP16/FP32 多精度计算。其计算单元包含 1024 个 8T SRAM 计算单元，每个单元支持 16 位 MAC 运算；通过权重驻留与输入广播机制减少数据搬运次数；根据任务需求自动切换计算精度，例如感知任务采用 INT8，规划任务采用 FP16。在感知任务中，如目标检测和图像分类，对精度的要求相对较低，采用 INT8 量化可以在保证一定精度的前提下，显著降低计算量和能耗。而在规划任务中，如路径规划和决策制定，需要更高的精度来确保安全性和准确性，因此采用 FP16 精度。

4.3 片上网络（NoC）设计

片上网络(NoC)针对存算一体架构的数据流特性设计层次化结构。全局层采用 2D Mesh 拓扑，连接 IPU、CPU 与存储模块，带宽 256GB/s；局部层在 IPU 内部采用 H－Tree 结构，实现计算单元间的低延迟通信（<5ns）。2D Mesh 拓扑结构具有良好的扩展性和可预测性，能够满足全局数据传输的需求。而 H－Tree 结构在 IPU 内部可以有效地减少信号传输的延迟，提高计算单元之间的通信效率。

4.4 存储系统设计

存储系统采用分级策略。一级存储为 SRAM 存内计算阵列，容量 128MB，带宽 1TB/s，用于权重与中间结果缓存；二级存储为 DDR5 外存，容量 8GB，带宽 51.2GB/s，用于存储模型参数与历史数据。一级存储的高速特性可以满足 IPU 对数据的快速访问需求，减少计算等待时间。二级存储的大容量则可以存储大量的模型参数和历史数据，为自动驾驶算法的训练和优化提供支持。

五、能效优化策略

5.1 动态精度调整

通过动态精度调整技术，在保证任务精度的前提下降低计算能耗。感知任务中，目标检测采用 INT8 量化，精度损失<1%，能耗降低 75%；规划任务中，路径规划采用 FP16，避免精度损失导致的安全风险。在实际应用中，可以根据不同的任务场景和需求，动态地调整计算精度。

5.2 多电压域电源管理

设计多电压域电源管理系统，高性能域 IPU 核心工作电压 1.2V，频率 1GHz；低功耗域 CPU 与外设工作电压 0.8V，频率 200MHz；通过动态调压根据负载需求实时调整电压，静态功耗降低 40%。在不同的工作状态下，芯片的各个部分对电压和频率的需求是不同的。通过多电压域设计，可以根据实际负载情况动态调整电压和频率，在保证性能的同时降低功耗。

5.3 近似计算技术

针对自动驾驶中的容错任务（如环境感知），引入近似计算技术。神经元剪枝移除冗余连接，模型规模缩小 60%，精度损失<2%；激活函数近似用分段线性函数替代 ReLU，计算能耗降低 50%。在环境感知任务中，一些细微的误差对最终的决策结果影响较小，因此可以采用近似计算技术来降低计算复杂度和能耗。神经元剪枝可以去除神经网络中不重要的连接，减少参数数量和计算量；激活函数近似则可以用简单的分段线性函数替代复杂的 ReLU 函数，降低计算能耗。

六、实验验证与结果分析

6.1 实验平台与测试任务

基于 12nm CMOS 工艺流片，测试芯片包含 1024 个存算一体单元，核心频率 1GHz，支持 INT8/FP16/FP32 多精度计算。测试任务包括 ResNet－50 图像分类（输入分辨率 224×224，Batch Size＝8）与 BEV 网络端到端测试（输入点云数据 100k 点/帧，输出 BEV 特征图 200×200）。选择这些任务是因为它们代表了自动驾驶中常见的感知和决策任务，能够全面评估芯片的性能和能效。

6.2 性能对比

与传统架构芯片对比，存算一体芯片在能效与算力密度方面优势显著。传统架构（GPU）能效比 3.2TOPS/W，存算一体架构 15.6TOPS/W，提升 4.88 倍；传统架构算力密度 1.2TOPS/mm^2，存算一体架构 24.5TOPS/mm^2，提升 20.4 倍；传统架构端到端延迟 120μs,存算一体架构 35μs,降低 3.43 倍。实验结果表明，存算一体架构在能效、算力密度和延迟方面都明显优于传统架构，能够更好地满足高级别自动驾驶的需求。

6.3 能效优化效果

混合精度计算与近似计算技术可显著降低能耗。混合精度在 ResNet － 50 任务中，INT8 模式下能耗降低 72%，精度损失仅 0.8%；近似计算在 BEV 网络中，神经元剪枝使能耗降低 58%，精度损失 1.2%。这些能效优化策略在保证任务精度的前提下，有效地降低了芯片的能耗，提高了能效比。

七、总结

本文针对高级别自动驾驶场景，提出了一种存算一体芯片架构设计与能效优化方案，通过混合精度计算、片上网络优化与近似计算技术,实现了能效比与算力密度的显著提升。实验结果表明，该架构在典型自动驾驶任务中能效比达 15.6TOPS/W，算力密度 24.5TOPS/mm^2，满足 L4 级自动驾驶需求。存算一体架构的创新设计有效地解决了传统架构在自动驾驶应用中面临的算力和能效瓶颈问题，为高级别自动驾驶的发展提供了有力的技术支撑。

[参考文献]

[1]彦明峰，智能驾驶芯片算力需求演变与架构优化方向,《汽车电子技术》，2022

[2]孙斌,混合精度高能效存算一体芯片架构及关键技术研究,《微电子学与计算机》，2023

[3]炜力,面向存算一体系统的设计空间探索和系统优化方法研究,《计算机研究与发展》，2021