基于 KA200-S 类脑芯片的边缘计算终端设计与实现

王俊文，程超，狄广义，王耀坤，郭建民
国能数智科技开发（北京）有限公司，北京 100011

DOI: 10.61369/TACS.2026090006

摘　　要： 针对工业边缘计算场景对低功耗、高能效 AI 算力的需求，设计并实现了一款基于 KA200-S 类脑芯片的边缘计算终端。该终端采用 ARM 主控与类脑协处理器异构架构，通过 PCIe 3.0 总线实现数据交互。测试结果表明：系统在 YOLOv5s 任务中达 145 fps，功耗 6.73 W，能效比 21.55 fps/W，较 NVIDIA Orin NX 提升约 2.8 倍，较华为 Atlas 200 提升约 3.2 倍；系统在 −20℃ ~60℃环境下稳定运行，72 h 连续推理帧率波动小于 3%。

关 键 词： 类脑计算；边缘计算；异构架构；能效优化

Design and Implementation of Edge Computing Terminal Based on KA200-S Neuromorphic Chip

Wang Junwen, Cheng Chao, Di Guangyi, Wang Yaokun, Guo Jianmin
National Energy Information Technology Co., Ltd., Beijing 100011

Abstract: To meet the demands of low power consumption and high energy-efficient AI computing in industrial edge scenarios, an edge computing terminal based on the KA200-S neuromorphic chip is designed and implemented. The terminal adopts a heterogeneous architecture with an ARM host processor and a neuromorphic co-processor, with data exchange via PCIe 3.0 bus. Experimental results show that the system achieves 145 fps on YOLOv5s at 6.73 W with an energy efficiency of 21.55 fps/W, approximately 2.8 times higher than NVIDIA Orin NX and 3.2 times higher than Huawei Atlas 200. The system operates stably from −20 ℃ to 60 ℃ with frame rate fluctuation below 3% over 72-hour continuous inference.

Keywords: neuromorphic computing; edge computing; heterogeneous architecture; energy efficiency optimization

引言

随着 AI 技术在工业场景的深入应用，边缘计算设备面临双重挑战：一方面需足够的 AI 算力支撑目标检测等深度学习任务；另一方面受限于部署环境的供电和散热，对功耗有严格控制。传统 GPU 方案虽算力充足，但功耗较高，限制了其在矿山井下等场景的部署[1-2]。

类脑计算借鉴生物神经系统机制，采用存算一体架构和事件驱动模式[3]。灵汐科技 KA200-S 芯片基于存算一体、众核并行、异构融合架构，单芯片集成 25 万神经元和 2500 万突触，支持 DNN 与 SNN 异构融合模型。其片上 SRAM 与计算单元紧耦合设计，消除了传统架构的数据搬运瓶颈。

现有研究多聚焦芯片架构层面，针对类脑芯片在工业边缘场景的终端级系统集成报道较少[4]。本文以 KA200-S 为核心，设计并实现了一款面向工业边缘计算的完整终端系统，从硬件架构、软件平台到模型部署进行了全流程验证。

一、芯片特性与选型分析

（一）KA200-S 核心架构

KA200-S 采用存算一体架构，内部含 30 个类脑计算核心，最高 400 MHz，片上集成 30 MB SRAM。主要参数如表 1 所示。

表 1 KA200-S 类脑芯片主要参数

参数	指标
架构	存算一体、众核并行、异构融合
神经元数量	25 万（稠密）/200 万（稀疏）
突触数量	2500 万（稠密）/20 亿（稀疏）
INT8 算力	48 TOPS
FP16 算力	24 TFLOPS
类脑计算单元	30 核心，最高 400 MHz

作者简介：
王俊文，女，高级工程师，主要研究方向：边缘计算、类脑芯片应用。
程超，男，高级工程师，主要研究方向：嵌入式系统。
狄广义，男，高级工程师，主要研究方向：人工智能。
王耀坤，男，高级工程师，主要研究方向：项目管理、信息化工程。
郭建民，男，高级工程师，主要研究方向：边缘计算。

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

芯片内置硬件图像预处理引擎，支持 Resize、Mirror、Flip 等操作，可减轻主控负担。

（二）选型对比

与主流边缘 AI 芯片的对比如表2所示。

表2 边缘 AI 芯片选型对比

对比项	KA200-S	NVIDIA Orin NX	华为 Atlas 200
INT8算力	48 TOPS	70 TOPS	22 TOPS
存算一体	支持	不支持	不支持
SNN 支持	支持	不支持	不支持
DNN/SNN 异构	支持	不支持	不支持

KA200-S 在绝对算力上低于 Orin NX，但凭借存算一体架构在能效比方面具有优势，且是唯一支持 DNN 与 SNN 异构融合模型的方案。

二、 系统硬件设计

（一）总体架构

终端采用 ARM 主控与 KA200-S 协处理器异构架构，如图1所示。主控选用瑞芯微 RK3568（4×Cortex-A55，2.0 GHz），负责系统管理与外设控制。KA200-S 通过 PCIe 3.0与主控连接，承担 AI 推理加速。

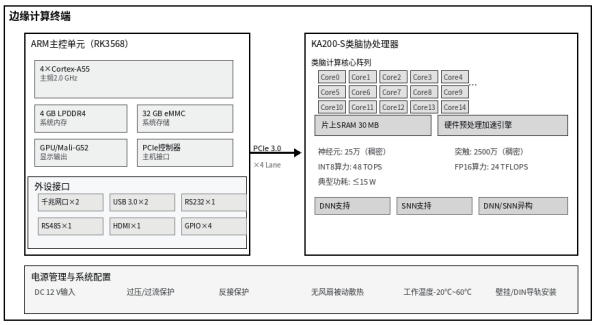


图1 系统总体架构图

（二）接口与电源

系统配置4 GB LPDDR4和32 GB eMMC，支持 TF 卡扩展。外设包括：千兆网口 ×2、USB 3.0×2、RS232×1、RS485×1、HDMI×1、GPIO×4。电源支持 DC 12 V 输入，具备过压、过流、反接保护。

（三）环境适应性

终端采用无风扇被动散热，工作温度 -20℃ ~60℃，支持壁挂和 DIN 导轨安装，适用于矿山、电力等复杂环境[5-7]。

三、软件平台设计

（一）操作系统与驱动

系统基于 Ubuntu 20.04 LTS，完成 KA200-S 驱动移植。驱动提供芯片初始化、模型加载、推理执行、状态查询等接口，芯片注册为 /dev/lynxi 设备节点，应用程序通过标准文件操作调度推理[8]。

（二）部署工具链

工具链支持 TensorFlow、PyTorch 模型导入、量化、编译与部署，流程如图2所示。支持 INT8/FP16混合精度量化，在保持精度前提下提升速度并降低功耗。

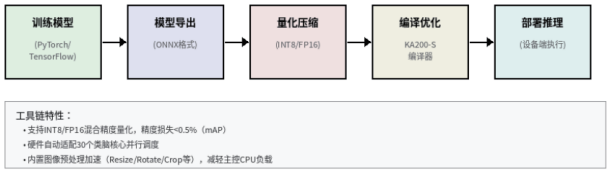


图2 模型部署工具链流程图

（三）算法部署

完成了 YOLOv5s 目标检测、ResNet50分类、PP-OCR 文字识别等算法的边缘端部署。部署流程：PyTorch 导出→ ONNX 转换→ INT8量化→ KA200-S 编译→推理执行。

四、 性能测试与分析

（一）测试环境

平台：RK3568主控 +KA200-S 协处理器，4 GB LPDDR4，Ubuntu 20.04 LTS。功耗通过功率计在稳态运行时测量（60 s 均值）。测试模型包括 YOLOv5s/m/l、ResNet50/18、MobileNet-v1等。

（二）推理性能

各模型推理性能如表3所示，功耗维持在6~8 W 区间。

表3 KA200-S 典型模型推理性能

模型	输入尺寸	精度	帧率 /fps	功耗 /W	能效 /（fps · W^{-1}）
ResNet50	224×224×3	FP16	211	7.24	29.14
ResNet18	224×224×3	FP16	538	7.22	74.52
MobileNet-v1	224×224×3	FP16	1 035	6.01	172.21
YOLOv5s	640×640×3	FP16	145	6.73	21.55
YOLOv5m	640×640×3	FP16	69	7.83	8.81
YOLOv5l	640×640×3	FP16	46	8.32	5.53

（三）能效对比

以 YOLOv5s 为基准，能效对比如表4所示。

表4 能效对比（YOLOv5s，640×640×3，FP16）

平台	帧率 /fps	功耗 /W	能效 /(fps · W^{-1})
KA200-S	145	6.73	21.55
NVIDIA Orin NX	127	16.29	7.80
华为 Atlas 200	87	12.87	6.76

KA200-S 能效比为21.55 fps/W，约为 Orin NX（7.80）的2.8倍，约为 Atlas 200（6.76）的3.2倍。优势源于存算一体架构对数据搬运开销的消除——传统架构中数据搬运功耗占50%以上。

（四）稳定性测试

系统在 -20℃ ~60℃进行72 h 连续运行测试（含低温启动、常温运行、高温老化），持续运行 YOLOv5s 推理。60℃下芯片表面温度稳定在85℃以下，推理帧率波动 <3%。

 | Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

五、工程应用验证

基于本终端完成矿山人员入侵检测部署验证。系统接入4路1080P摄像头，实时运行YOLOv5s对人员、车辆识别跟踪，检测到目标后通过RS485联动报警器，同时经千兆网口上报监控中心[9-11]。

4路并发下终端整体功耗约25 W（含主控、类脑芯片及4路摄像头供电），单帧检测延迟<50 ms。连续运行30天，累计告警1 200余次，漏检率<2%，验证了系统在真实工业环境中的可靠性。

六、结论

本文基于KA200-S类脑芯片设计并实现了一款边缘计算终端。系统采用ARM主控与类脑协处理器异构架构，充分发挥了存算一体架构的能效优势。测试结果表明：YOLOv5s任务中能效比达21.55 fps/W，约为NVIDIA Orin NX的2.8倍；系统可在-20℃~60℃工业环境稳定运行。矿山人员入侵检测验证进一步证实了其在实际场景中的可行性。该设计为类脑计算在工业边缘场景的工程化应用提供了可行方案。

参考文献

[1] 灵汐科技 . 类脑芯片领启 KA200(-S) 产品手册 [Z]. 北京：北京灵汐科技有限公司，2024.

[2] 冯烁，路冬冬，尹飞，等 . 面向分布式计算的类脑智能处理器指令集设计 [J]. 计算机研究与发展，2026，63(1)：1-14.

[3] 存算一体类脑计算架构研究综述 [J]. 电子与封装，2024.

[4] 基于瑞芯微 RK3568+ 灵汐 KA200高能效边缘 AI 计算平台 [EB/OL]. (2024-06-15). https://www.lxsemi.com.

[5] 李源堃，王泽，张清天，等 .NAS4CIM：面向忆阻器存算一体芯片的神经网络结构搜索框架 [J]. 电子与信息学报，2025，47(12)：4948-4958.

[6] 杨玉超，张柏骏，袁锐，等 . An end-to-end memristive hardware system based on single-spike coding for human-machine interfaces[J]. Nature Electronics，2026.

[7] WANG M J，et al. Brain-inspired AI for Edge Intelligence：a systematic review[J/OL]. arXiv preprint，arXiv：2603.26722，2026.

[8] SynSense 时识科技 . 更稀疏的神经元编码转换机制显著降低类脑芯片功耗 [EB/OL]. (2023-09-20). https://www.synsense.ai.

[9] 类脑计算模型及硬件加速关键技术研究 [D]. 南京大学，2024.

[10] 华为技术有限公司 . Atlas 200 AI 加速模块技术白皮书 [Z]. 2023.

[11] 刘升阳 . 基于 FPGA 技术的物联网智能家居控制系统设计 [J]. 电子产品世界，2025，32(5).

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License