南京大学 NANJING UNIVERSITY

RINC Robotic Intelligence & Neural Computing Group

面向边缘计算平台的高能效神经网络研究

答辩人：俞诗航 502022330064

导师：申富饶 教授

日期：2025年5月16日

誠樸雄偉 勵學敦行

RINC
Robotic Intelligence & Neural Computing Group

目录

誠樸雄偉 勵學敦行

RINC
Robotic Intelligence & Neural Computing Group

第一部分

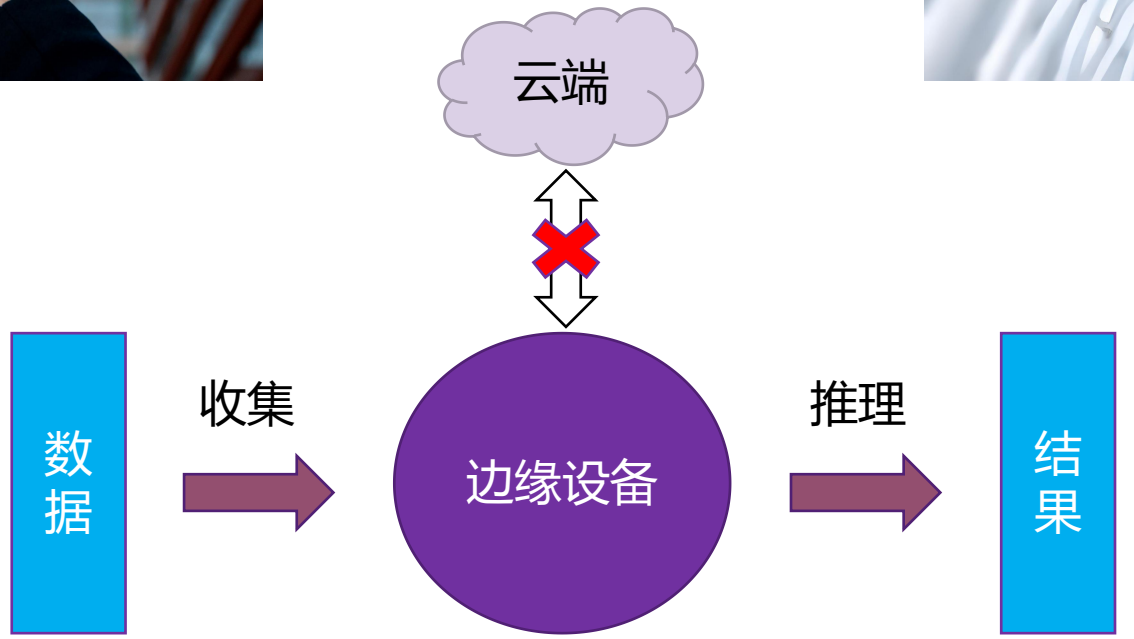
研究背景

Research Background

边缘智能 | 相关工作 | 困难与挑战

1.1 边缘智能

为边缘设备赋予智能



南京大学 NANJING UNIVERSITY

1.2 相关工作

推理能力

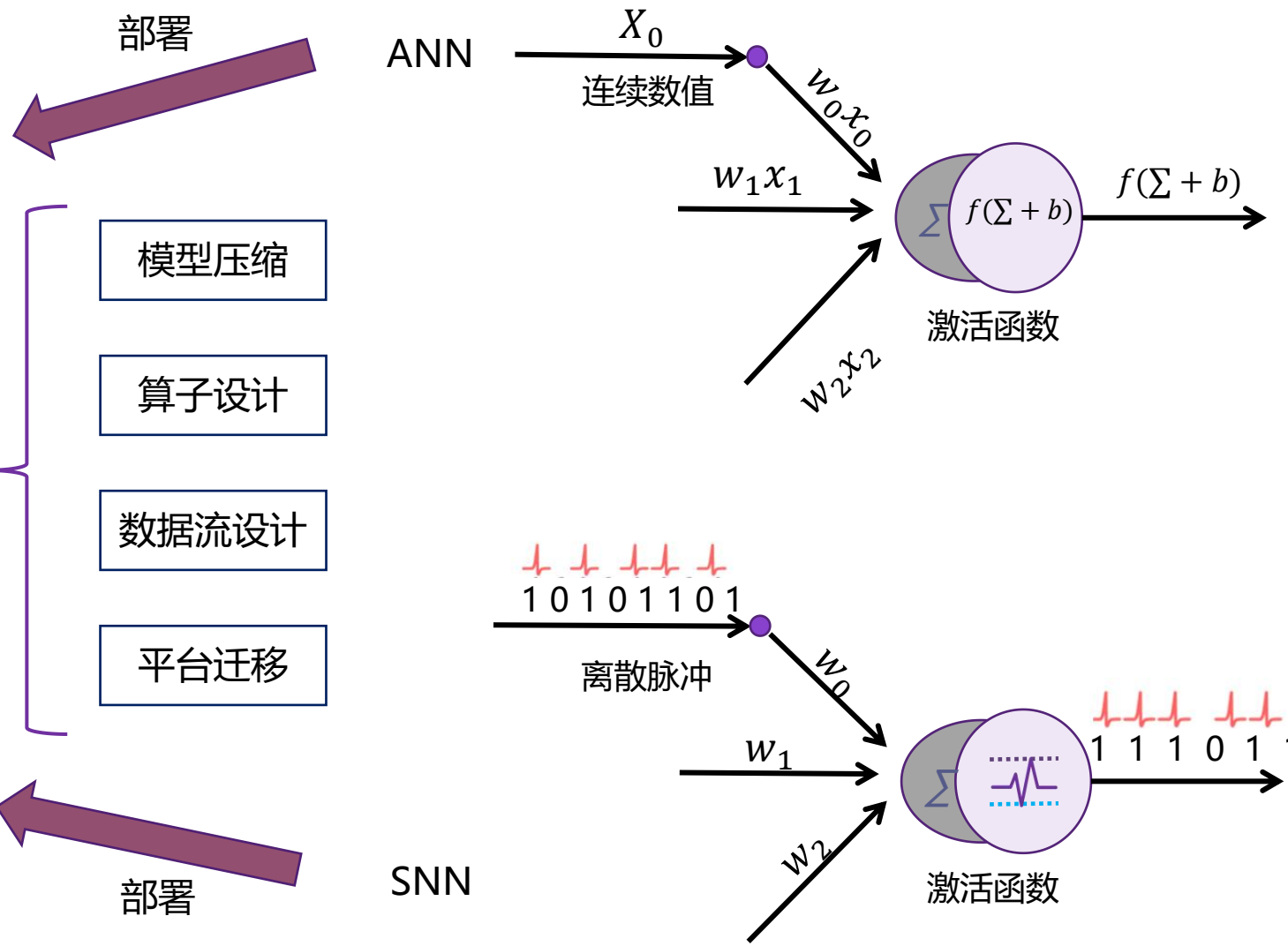
边缘设备

加速器设计

- 模型压缩
- 算子设计
- 数据流设计
- 平台迁移

部署

部署



RINC Robotic Intelligence & Neural Computing Group

1.3 研究难点 & 研究意义

- 软件模型的优化目标：
  - 速度快
  - 精度高

- 硬件资源的优化目标：
  - 运行功耗低
  - 存储占用少

◆ **研究难点：**

- 边缘设备**内存**有限，神经网络模型**内存**占用多
- 边缘设备**算力**有限，神经网络模型**算力**需求高
- 边缘设备**续航**有限，神经网络模型**能量**消耗大

神经网络加速器研究意义：
为**边缘**计算平台部署**高能效**的神经网络模型，兼具**高性能**与**低功耗**

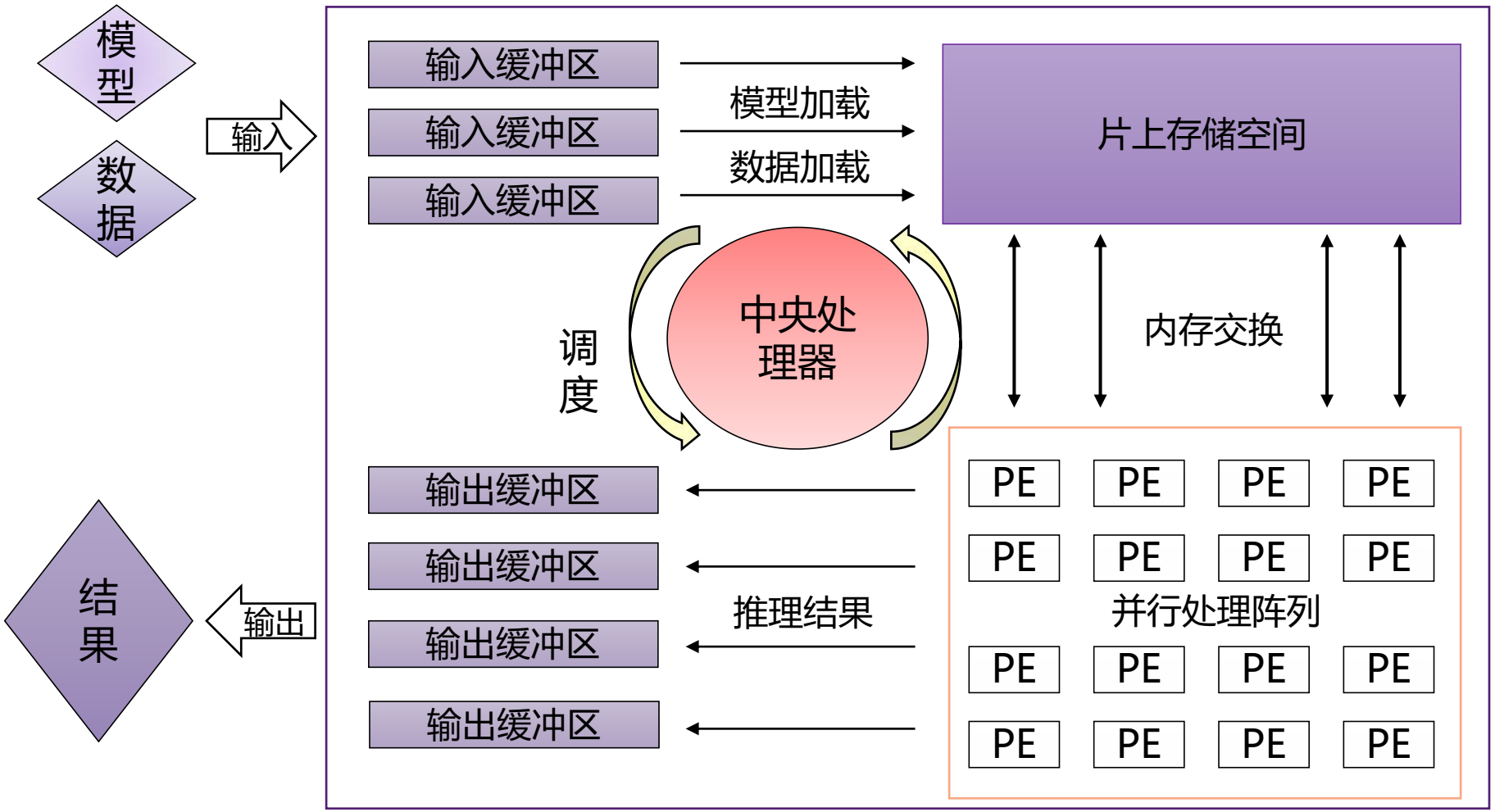
第二部分

研究内容

Research Content

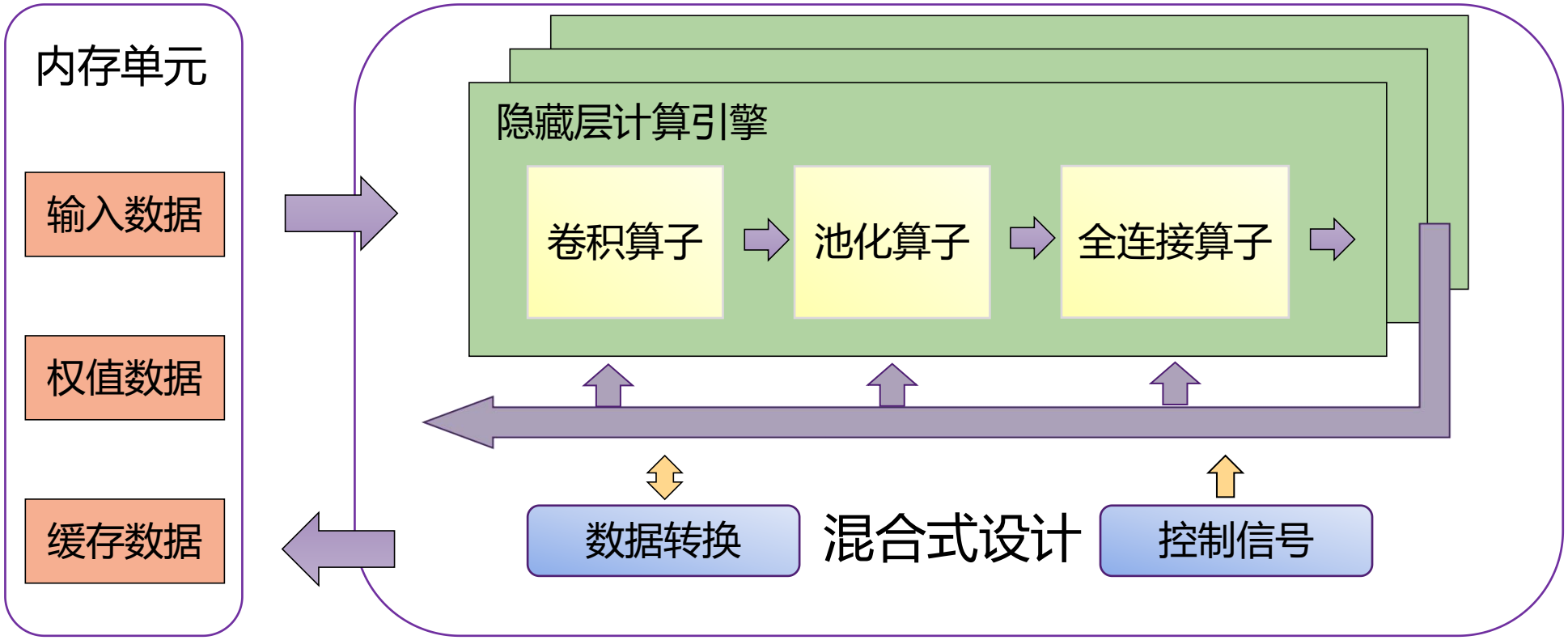
- 基于FPGA的低功耗卷积神经网络加速器设计
- 基于NPU的高能效脉冲神经网络加速器设计

基本架构


模型
数据
输入
输入缓冲区
输入缓冲区
输入缓冲区
模型加载
数据加载
片上存储空间
调度
中央处理器
内存交换
PE
并行处理阵列
输出缓冲区
输出缓冲区
输出缓冲区
输出缓冲区
推理结果
结果
输出

引言 神经网络加速器设计

数据流



层间算子重用 | 层内流式传输

2.1 基于FPGA的低功耗卷积神经网络加速器设计：研究动机

研究目标：

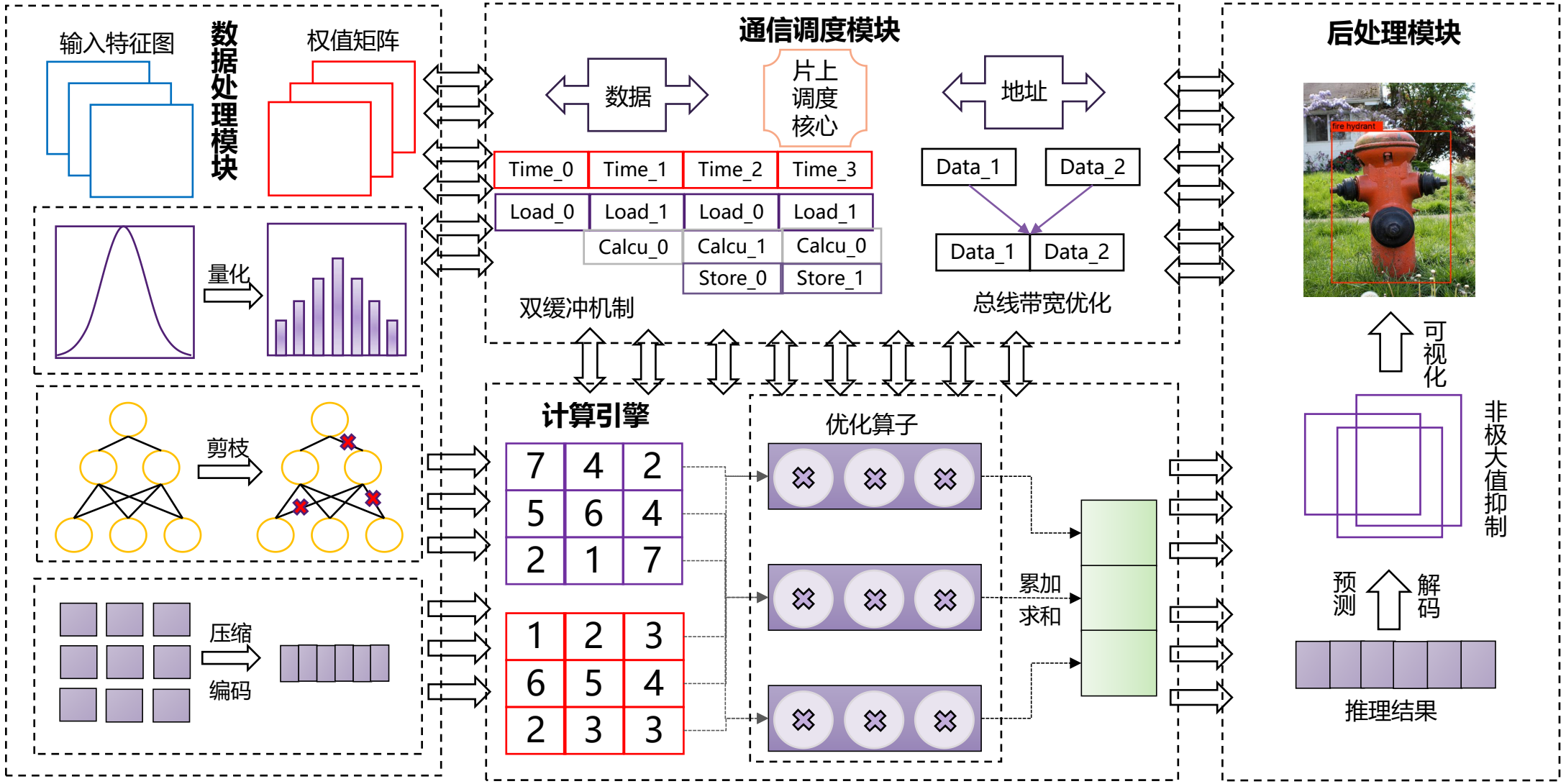
- 将深度卷积神经网络模型部署在**资源功耗受限**的FPGA平台上
- 发挥FPGA平台**低功耗**的优势，设计**高能效**的硬件加速器
- 在不降低或略微降低精度的情况下，实现模型**轻量化部署**

核心思路：

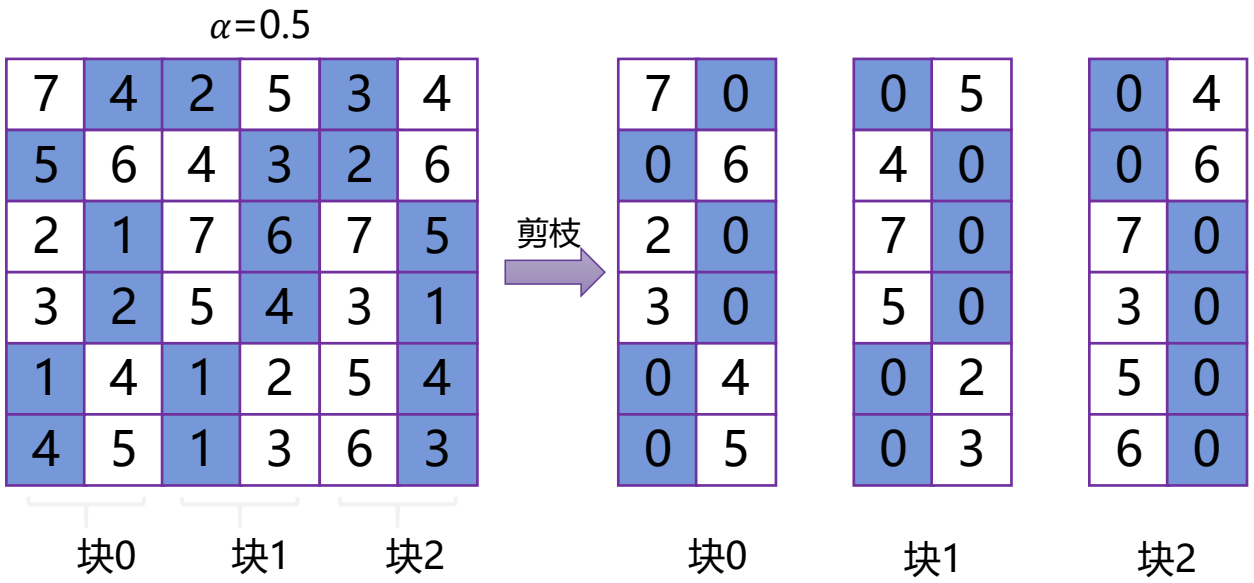
- 通过**剪枝**，**量化**等轻量化方法**降低深度CNN模型复杂度**
- 在**软件**模型层面对神经网络算子进行**硬件适配优化**
- 在**硬件**设计层面对加速器进行内存带宽等**资源优化**

南京大学 NANJING UNIVERSITY

2.1 基于FPGA的低功耗卷积神经网络加速器设计：整体设计架构



2.1 基于FPGA的低功耗卷积神经网络加速器设计：模型轻量化处理



块均衡结构化剪枝

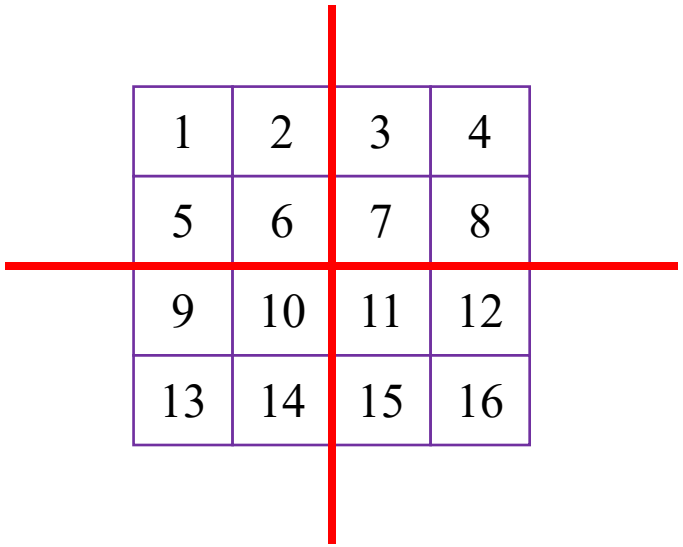
$$V_{float} = (-1)^{Sign} \times 2^{Exp-127} \times (1 + \sum_{i=1}^{23} b_{23-i} \cdot 2^{-i})$$

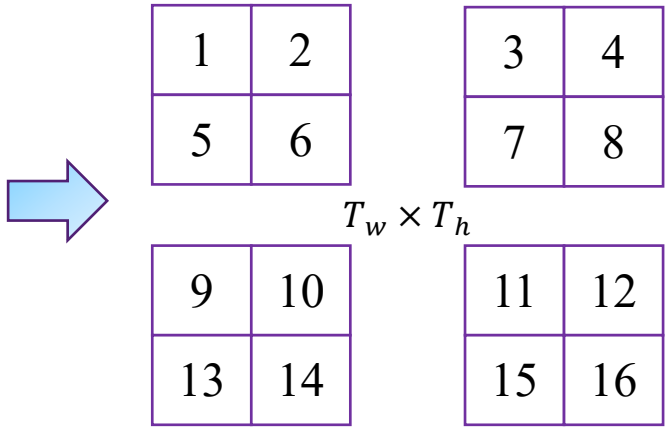
$$V_{fixed} = \sum_{i=0}^{Length-1} b_i \times 2^{-Exp} \times 2^i$$

定点数量化

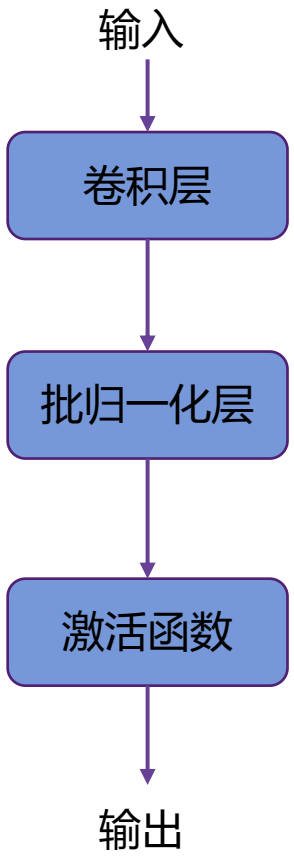
南京大学 NANJING UNIVERSITY

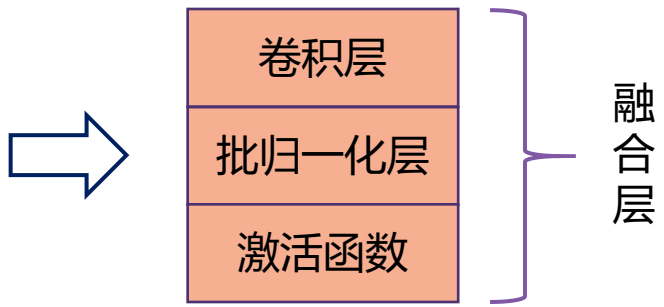
2.1 基于FPGA的低功耗卷积神经网络加速器设计：算子重构





循环平铺





层融合

2.1 基于FPGA的低功耗卷积神经网络加速器设计：数据流优化

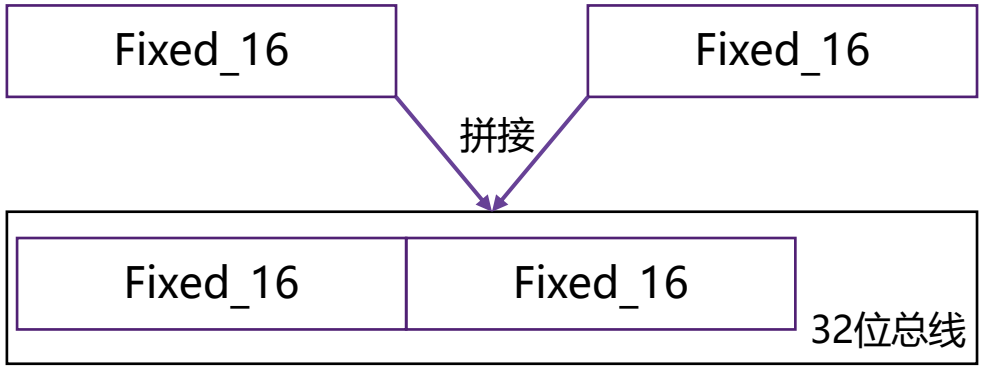
单缓冲结构

Time_1	Time_2	Time_3	Time_4	Time_5
Load	Calcu	Load	Calcu	Load
		Store		Store

双缓冲结构

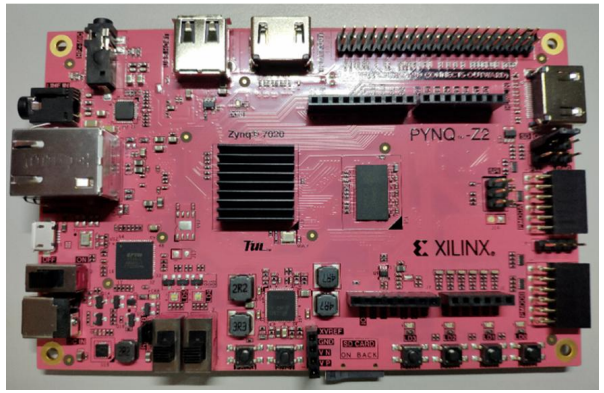
Time_1	Time_2	Time_3	Time_4	Time_5
Load_0	Load_1	Load_0	Load_1	Load_0
	Calcu_0	Calcu_1	Calcu_0	Calcu_1
		Store_0	Store_1	Store_0

双缓冲机制



带宽优化

2.1 基于FPGA的低功耗卷积神经网络加速器设计：实验结果



硬件平台：PYNQ-z2

表 3-1 原始模型参数分析

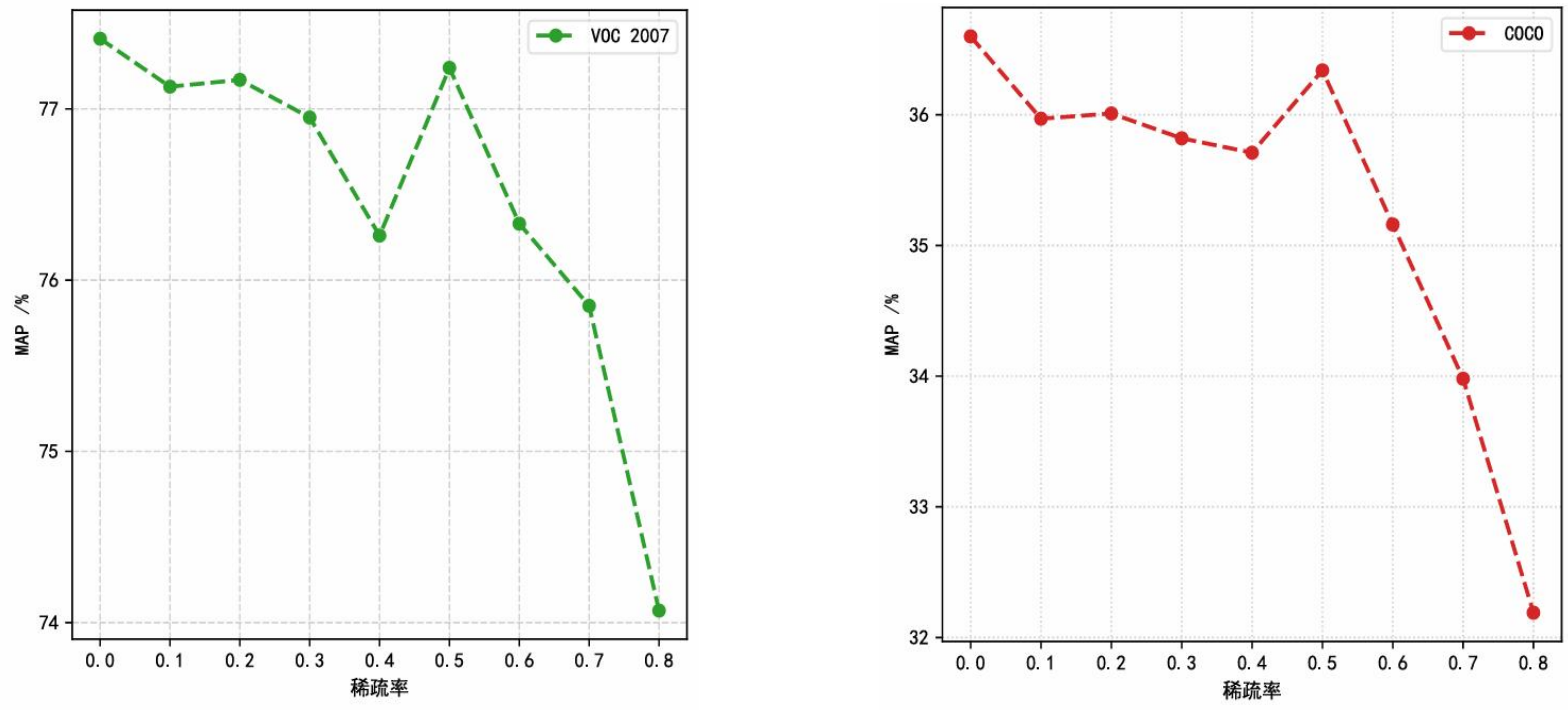
参数个数	模型大小	点积运算	浮点运算	运行内存
6056606	73.52MB	6.93GMAdd	3.47GFlops	176.34MB

表 3-2 16 位定点数量化分析

	LUT	Register	BRAM	DSP
总资源	53200	106400	140	220
量化前 (float32)	98508(185%)	94330(88.6%)	140(100%)	815(370%)
量化后 (AP16)	36325(68.2%)	34064(32.0%)	93(66.4%)	217(98.6%)

定点数量化大幅减少资源占用

2.1 基于FPGA的低功耗卷积神经网络加速器设计：实验结果



不同剪枝率下的精度损失

选择稀疏率α=0.5可在精度损失极小的情况下压缩模型

2.1 基于FPGA的低功耗卷积神经网络加速器设计：实验结果

表 3-3 不同硬件平台下的实验结果

硬件平台	CPU	GPU	FPGA
硬件型号	Intel(R)core(TM)i5-9400	NVIDIA RTX 2080Ti	Zynq XC7Z020
工作频率 (Hz)	2.9G	1.35G	100M
数据精度	float32	float32	AP16
推理速度 (s/张)	0.3721	0.0025	0.0427
平均功耗 (W)	65	260	2.33
能效 (张/J)	0.041	1.53	10.05

FPGA平台拥有最高的能效

表 3-4 与其他基于 FPGA 的相关工作对比

	Preußer 等人[111]	Nakahara 等人[112]	Montgomerie 等人[113]	Heller 等人[114]	本设计
网络模型	Tincy-YOLO	Lightweight-YOLOv2	YOLOv3-tiny	YOLOv4-tiny	YOLOv4-tiny
FPGA 平台	Zynq Ultrascale+	Zynq Ultrascale+	VCU110	Kria KV 260	PYNQ-z2
工作频率 (Hz)	N/A	300M	220M	N/A	100M
准确率 (mAP)(%)	48.5	67.6	33.9^{coco}	73.8	36.34^{coco} 77.24
DSP 消耗	N/A	377	1780	N/A	217
功耗 (W)	6	4.5	15.4	8	2.33
吞吐量 (FPS)	16	40.81	69	15	23.42
能效 (FPS/W)	2.67	9.06	4.48	1.88	10.05

准确率与能效均为最优

2.2 基于NPU的高能效脉冲神经网络加速器设计：研究动机

全球高端GPU市场由英伟达和AMD主导，他们占据了**80%**以上的市场，我国 AI 行业面临着关键设备断供的风险。

国产设备AI落地需求

RTX 3090

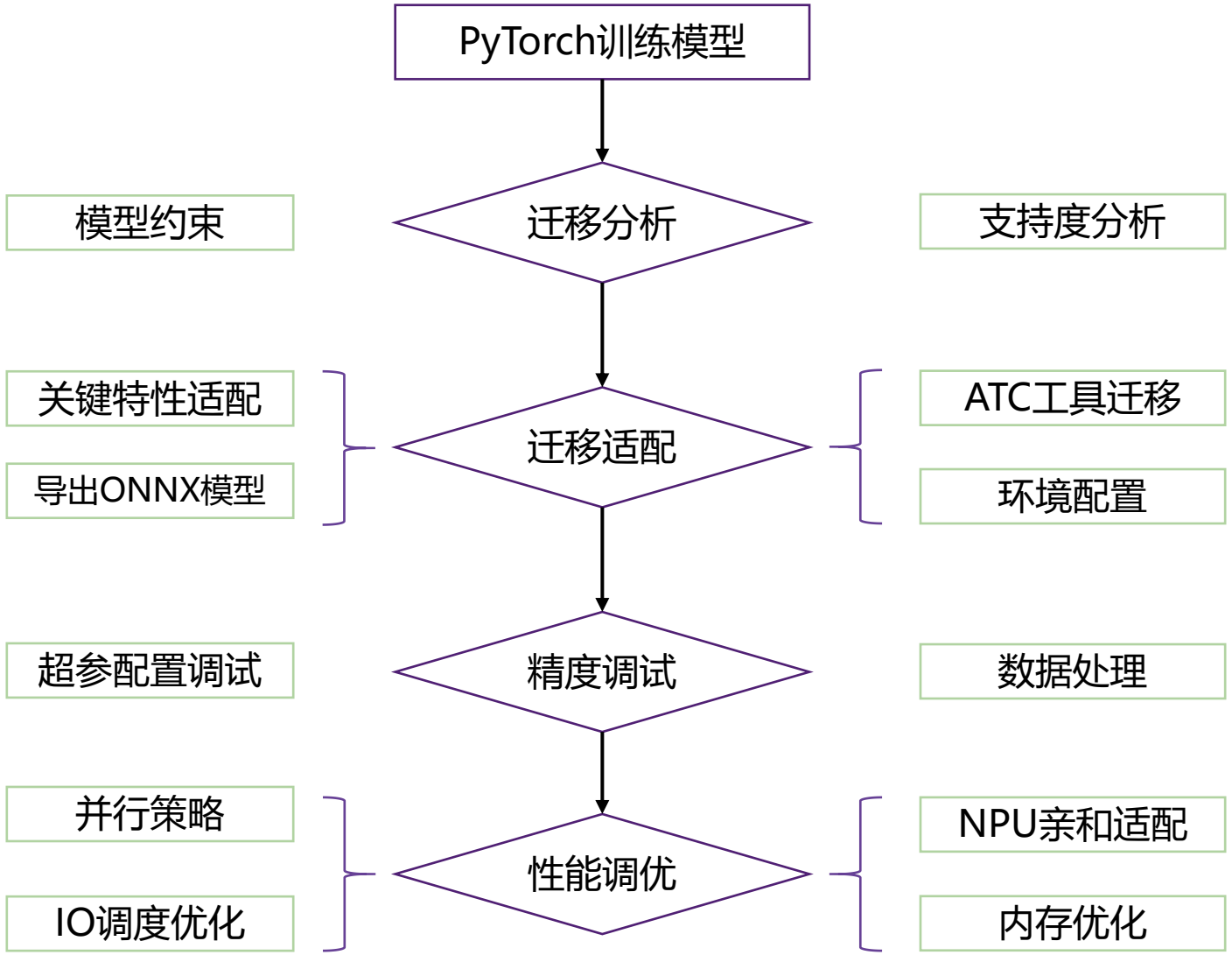
面临问题：

- **硬件架构**和**编程开发库**不同，无法直接部署
- 模型迁移时，**平台支持度**无法保证
- 迁移完成后，**模型性能**可能出现严重损失

研究思路：

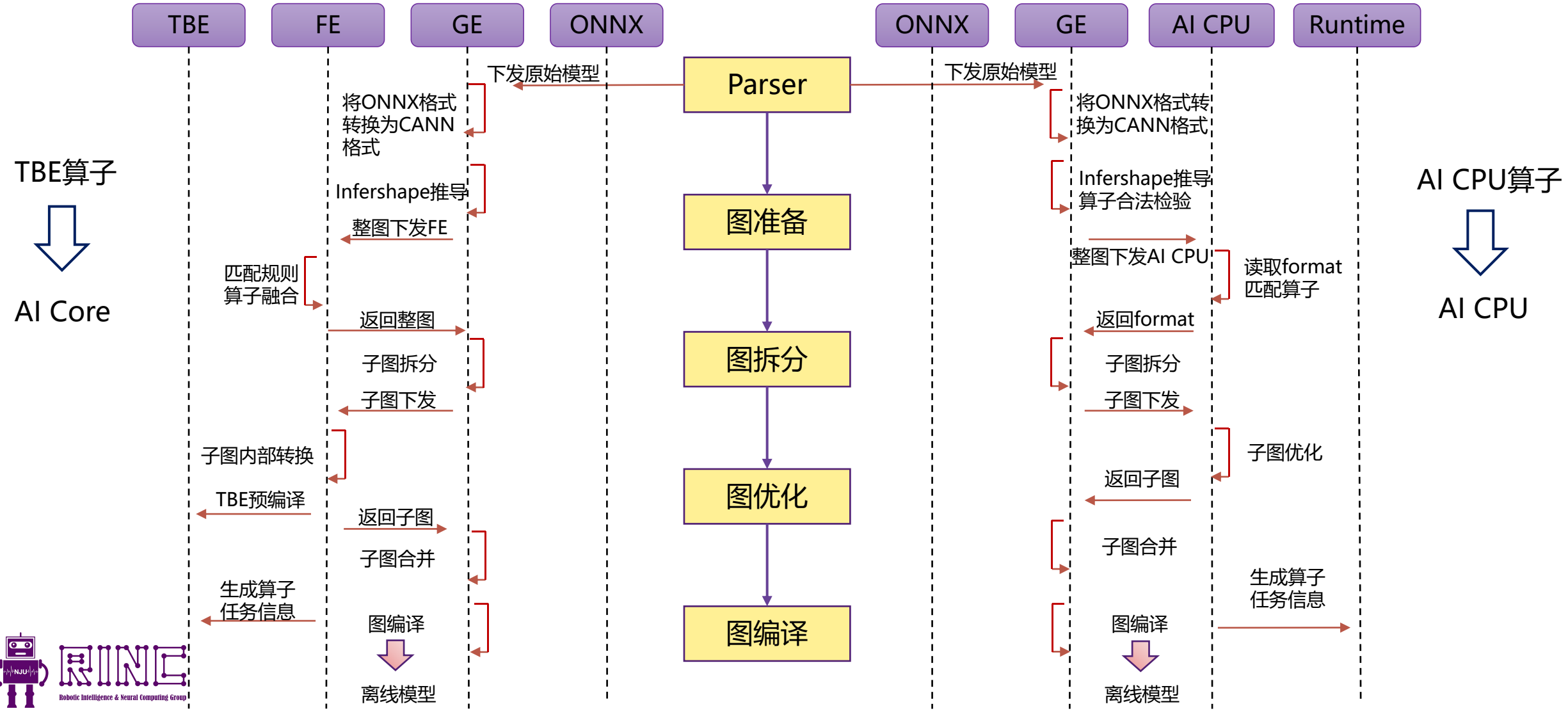
- 针对**国产**NPU平台进行模型迁移
- 选择第三代神经网络**脉冲神经网络**为迁移模型
- 通过**软硬件协同设计**对加速器进行优化

2.2 基于NPU的高能效脉冲神经网络加速器设计：模型迁移总体流程



南京大学 NANJING UNIVERSITY

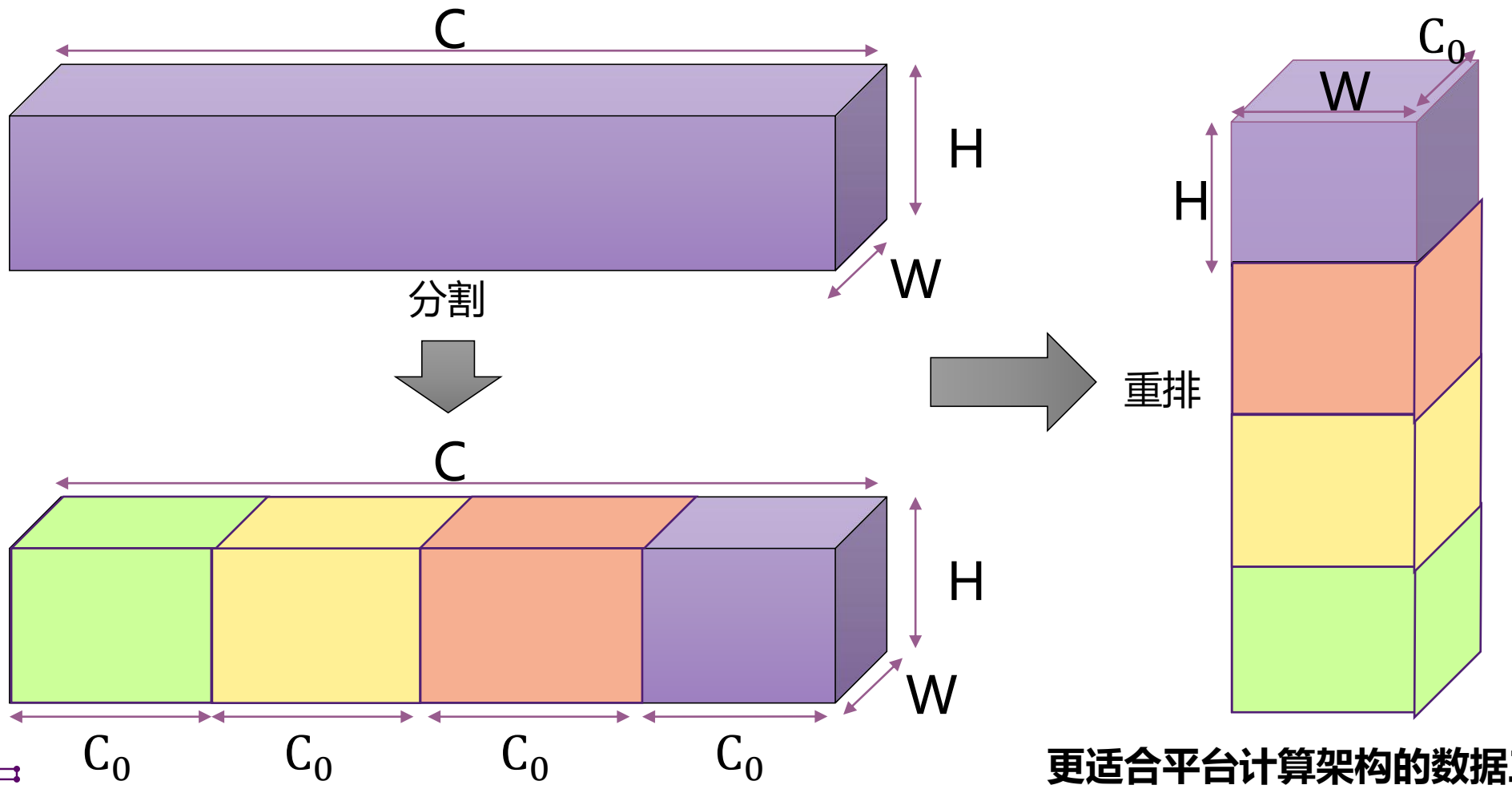
2.2 基于NPU的高能效脉冲神经网络加速器设计：算子调度优化



RINC Robotic Intelligence & Neural Computing Group

南京大学 NANJING UNIVERSITY

2.2 基于NPU的高能效脉冲神经网络加速器设计：权值数据重排



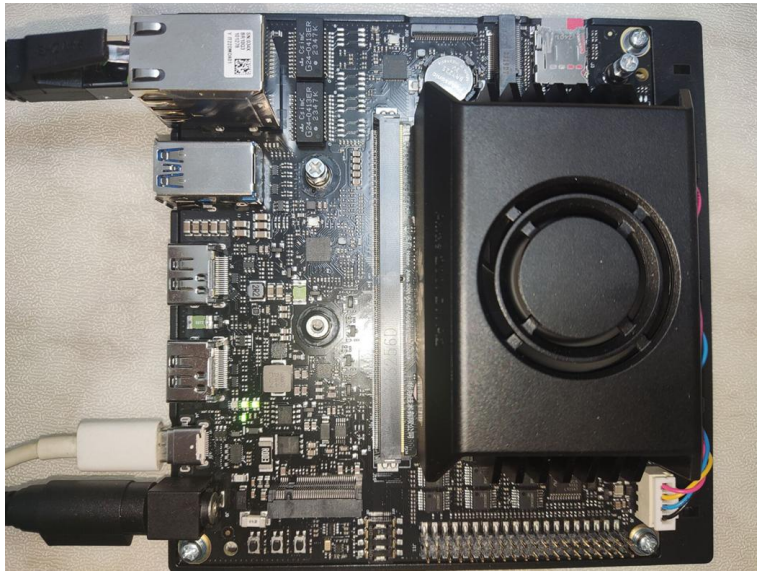
更适合平台计算架构的数据三维尺寸

2.2 基于NPU的高能效脉冲神经网络加速器设计：实验结果

```
{'alp': 0.65527344, 'ski': 0.3347168, 'tent': 0.006374359, 'shovel': 0.0022735596, 'dog sled': 0.0004878044}
Average inference time: 1.98 ms
*****run finish******
Releasing resources stage:
Resources released successfully.
```

图 4-5 SNN 模型在 NPU 上的推理效果

单张图片推理速度约为1.92ms



硬件平台：华为Atlas 200I DK A2

2.2 基于NPU的高能效脉冲神经网络加速器设计：实验结果

表 4-3 转换精度的对比: CIFAR-10

模型结构	ANN	SNN	迁移后的 SNN
VGG-16	95.49%	95.37%	94.75%
ResNet-18	96.02%	95.89%	95.02%
ResNet-50	96.76%	96.14%	94.88%

表 4-4 转换精度的对比: ImageNet

模型结构	ANN	SNN	迁移后的 SNN
VGG-16	74.27%	72.81%	71.24%
ResNet-34	74.30%	72.34%	71.03%

经过模型转换与平台迁移后，精度损失处于可接受范围

2.2 基于NPU的高能效脉冲神经网络加速器设计：实验结果

表 4-5 不同平台推理性能

硬件平台	CPU	GPU	NPU
硬件型号	Intel(R)core(TM)i5-9400	NVIDIA RTX 2080Ti	DaVinciV300 AI core
工作频率 (Hz)	2.9G	1.35G	500M
推理速度 (ms/张)	22.97	1.04	1.98
平均功耗 (w)	65	260	24
能效 (张/J)	0.67	3.70	21.04

NPU平台拥有最高的能效

表 4-6 与国外相关工作的对比

	Gerlinghoff 等人[40]	Rueckauer 等人[82]	本设计
硬件平台	Xilinx Virtex UltraScale+ XCVU13P	Loihi	华为 Atlas 200I DK A2
原始模型	AlexNet	MobileNet	ResNet-50
推理速度 (ms/张)	69.93	4.35	1.98
准确率 (%)	80.60	91.48	94.88
能效 (张/J)	3.04	9.80	21.04

在国产NPU平台上的表现优于国外FPGA平台与神经形态计算芯片

RINC
Robotic Intelligence & Neural Computing Group

第三部分

实际应用

Applications

系统简述 | 系统设计 | 运行效果

3.1 系统简述

■ 基于边缘计算平台的实时目标检测系统

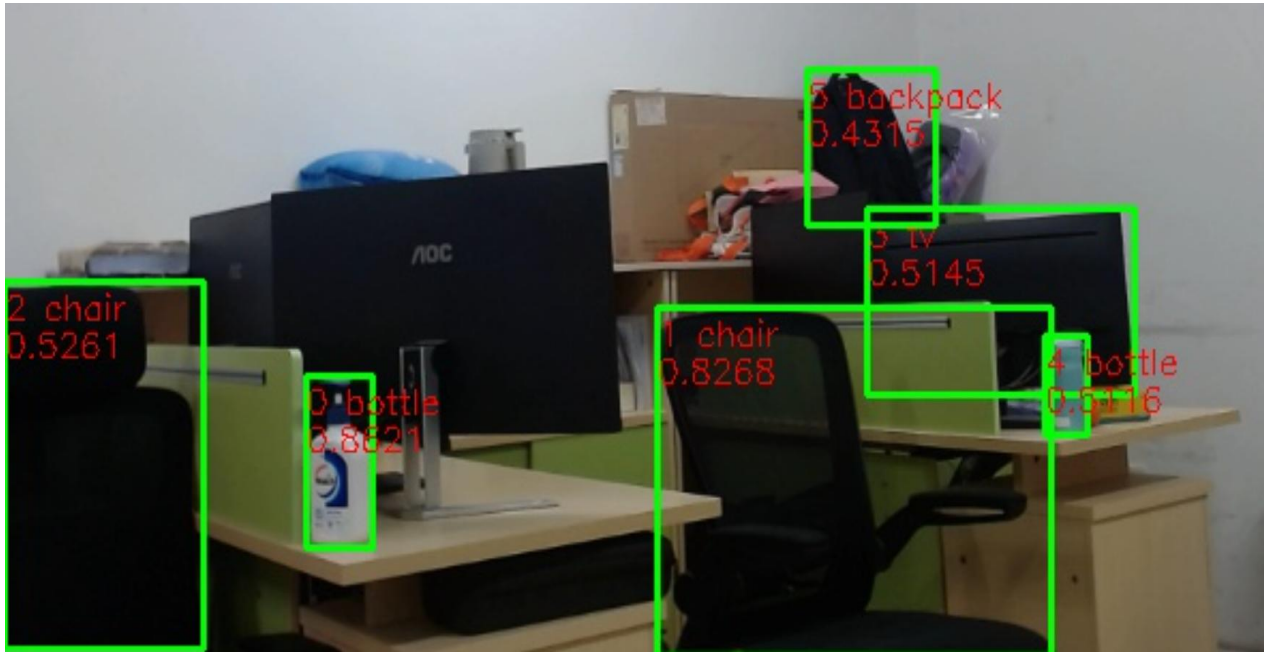
■ 系统需求：

◆ 用户需求：

- ✓ 体验友好的交互模块
- ✓ 安全隐私的认证控制
- ✓ 方便调整的检测视角
- ✓ 快速准确的可视结果

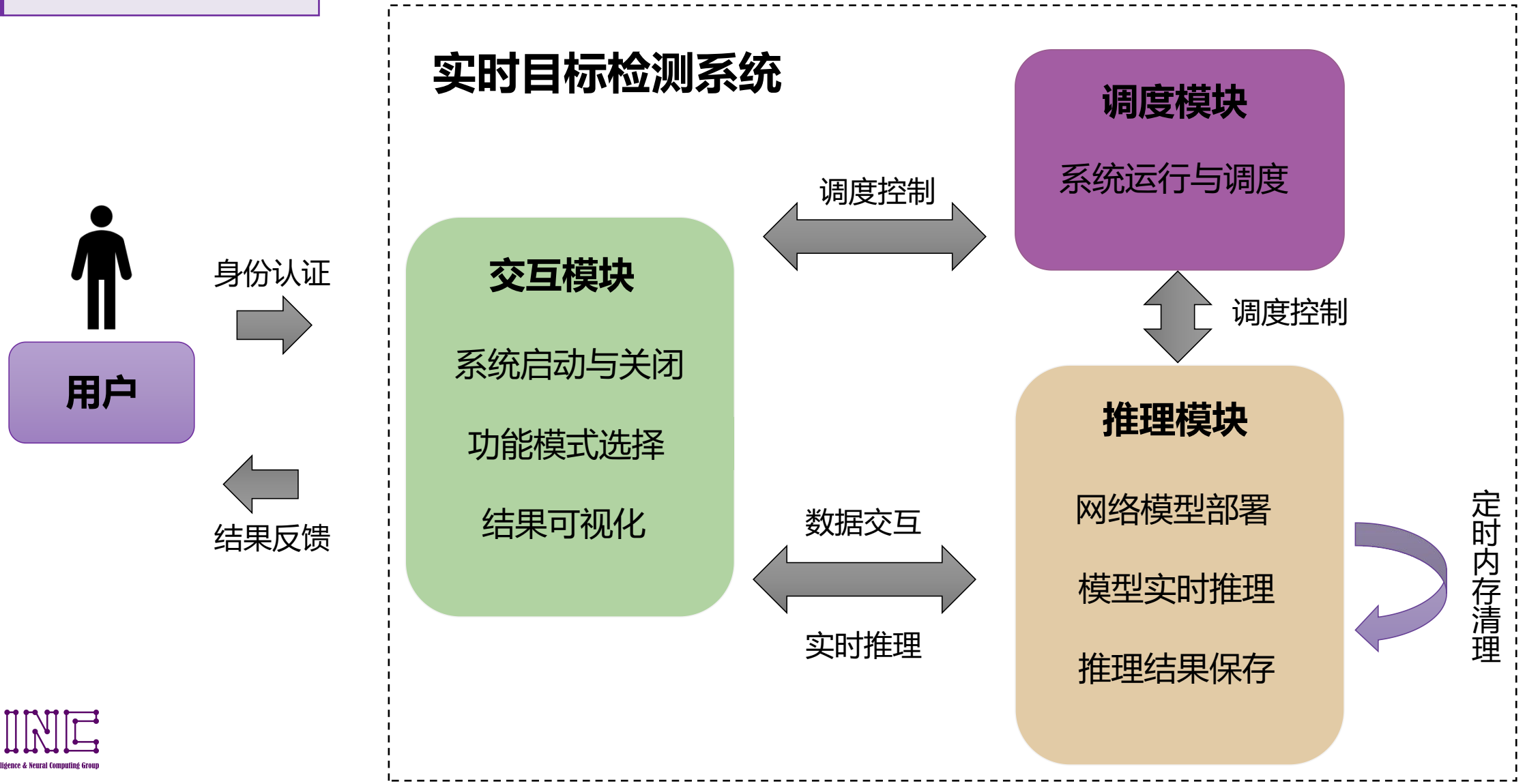
◆ 功能需求：

- ✓ 系统的启动与关闭
- ✓ 灵活化的功能定制
- ✓ 保证目标检测效果
- ✓ 持久化的待机部署



南京大学 NANJING UNIVERSITY

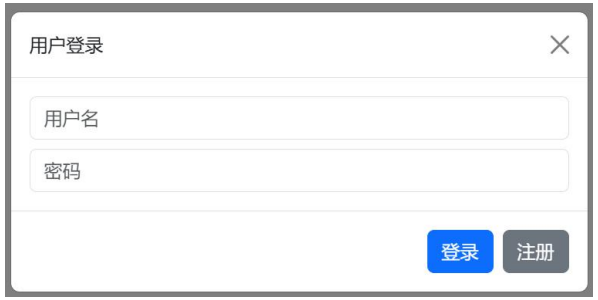
3.2 系统设计



南京大学 NANJING UNIVERSITY

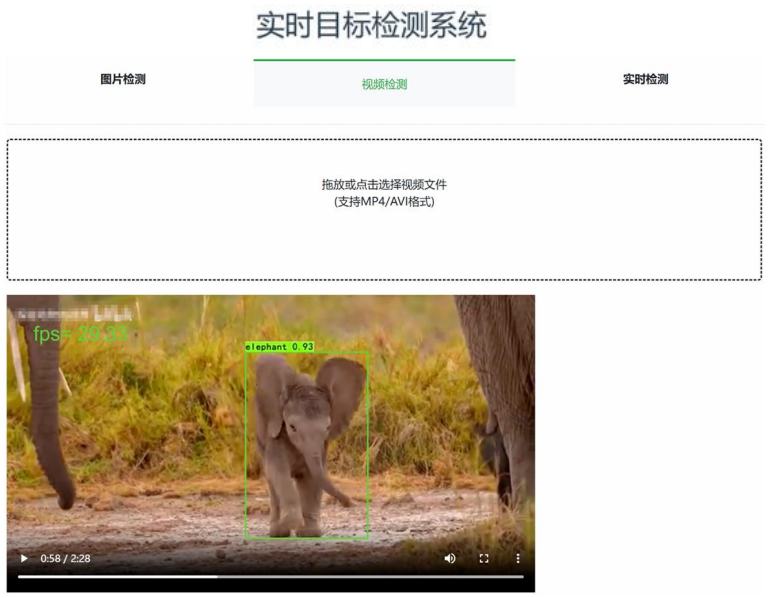
3.3 运行效果

登录系统



功能选择

开始检测



RINC
Robotic Intelligence & Neural Computing Group

第四部分

研究生期间工作成果

Work Product

➢ 论文

- 俞诗航,易梦军,吴洲,申富饶,赵健.神经形态计算:从脉冲神经网络到边缘部署[J]. 软件学报,2025,36(04):1758-1795.DOI:10.13328/j.cnki.jos.007298.

➢ 项目

- 数字化安全管控边缘计算装置自主可控关键技术研究及应用(项目编号5700-202319302A-1-1-ZN，负责该项目研究的边缘计算部分)
- 基于神经可塑性的脉冲网络高效学习机制与类脑智能系统(科技创新2030项目，负责脉冲神经网络研究部分)

➢ 竞赛

- 2024 华为软件精英挑战赛——江山赛区二等奖

➢ 荣誉

- 南京大学2024年度优秀研究生

第五部分

总结

Summary

面向边缘计算平台的高能效神经网络研究

基于FPGA的低功耗卷积神经网络加速器设计

- YOLOv4-tiny模型轻量化
- 算子重构与数据流优化
- 目标检测模型在FPGA平台高能效部署

基于NPU的高能效脉冲神经网络加速器设计

- SNN模型迁移适配
- 算子重构与数据流优化
- SNN模型首次部署在华为NPU平台

基于边缘计算平台的实时目标检测系统

- 详细的需求分析
- 目标检测模型在边缘计算平台落地
- 通过交互式网页，将检测结果实时显示

南京大学 NANJING UNIVERSITY

RINC Robotic Intelligence & Neural Computing Group

敬请各位老师批评指正

答辩人：俞诗航 502022330064

导师：申富饶 教授

日期：2025年5月16日

誠樸雄偉 勵學敦行