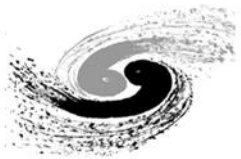


中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

高能所計算中心
IHEP Computing Center

Implement ParticleNet on FPGA

报告人：张玉涛

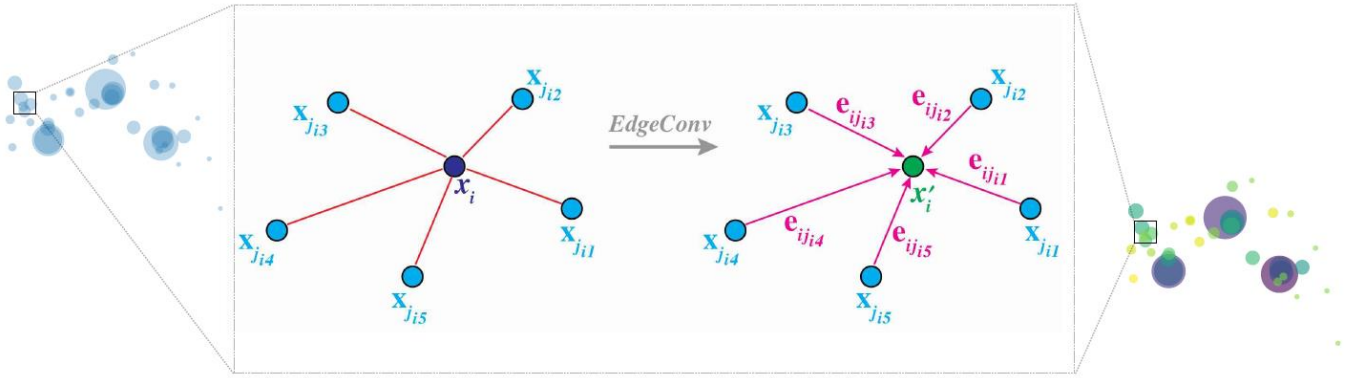
中国科学院高能物理研究所，计算中心

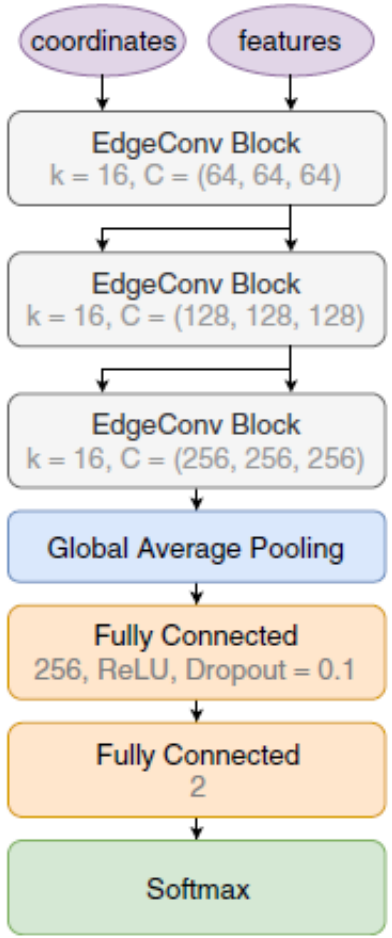
量子计算和机器学习青岛研讨会
2023年8月

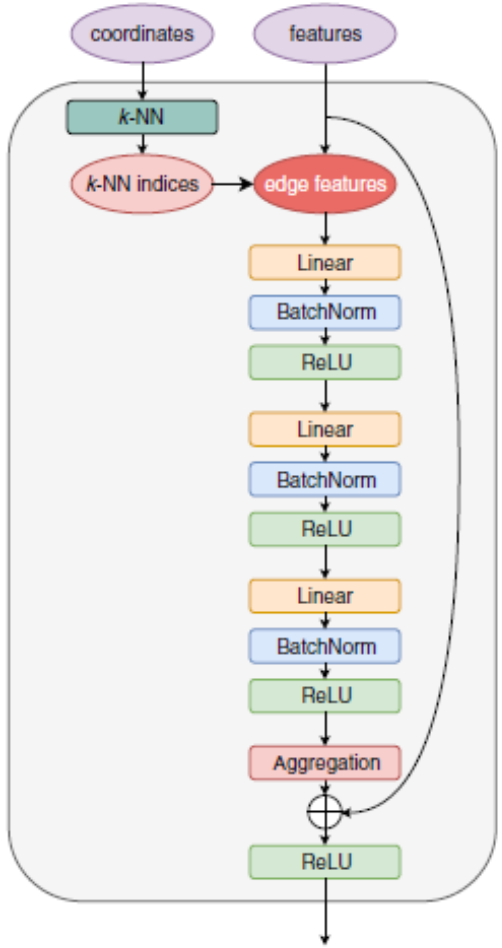
提纲

- ParticleNet
- 模型部署方案
  - hls4ml
- 模型推理加速技术
  - 剪枝、量化和图融合
  - 浮点到定点的转换
- FPGA实现结构
- 总结

ParticleNet： Jet Tagging的动态图卷积网络







- 将*jet*表示为Particle cloud

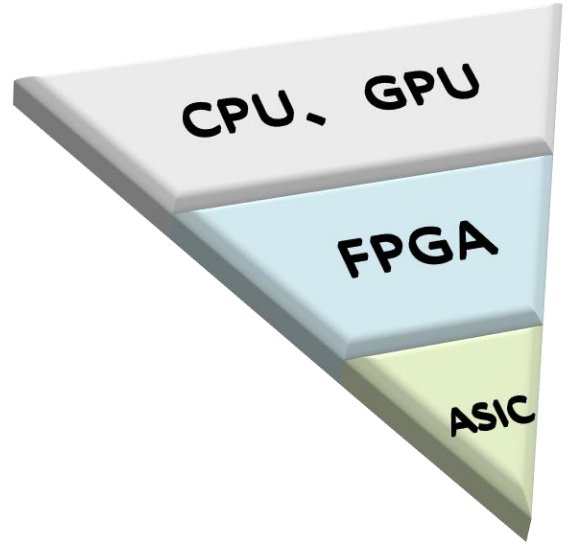
模型部署

深度学习领域中计算需求量大、推理时延大

高并行、高带宽、适合训练
在某些应用场景下功耗高、散热成本高

可编程性、低功耗
可针对特定应用进行半定制化开发，实现加速

根据用户需求进行定制化设计、达到更快计算速度
开发成本高、周期长、门槛高

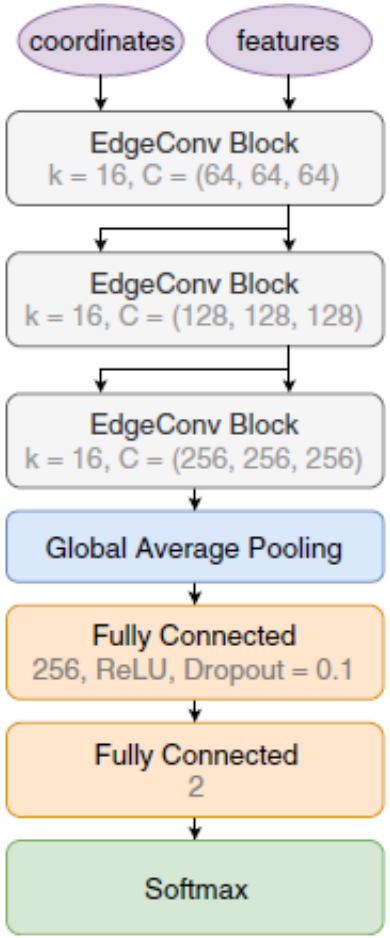


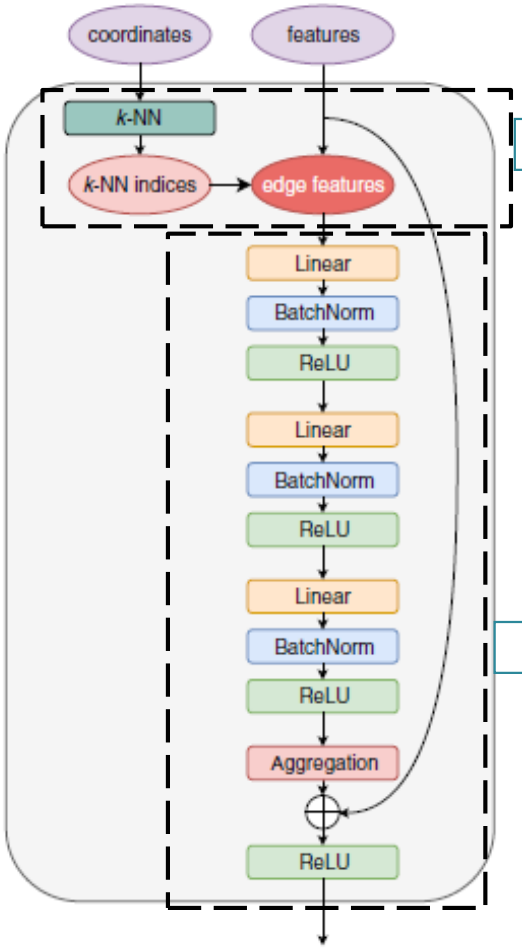
通用型芯片

半定制芯片

定制化芯片

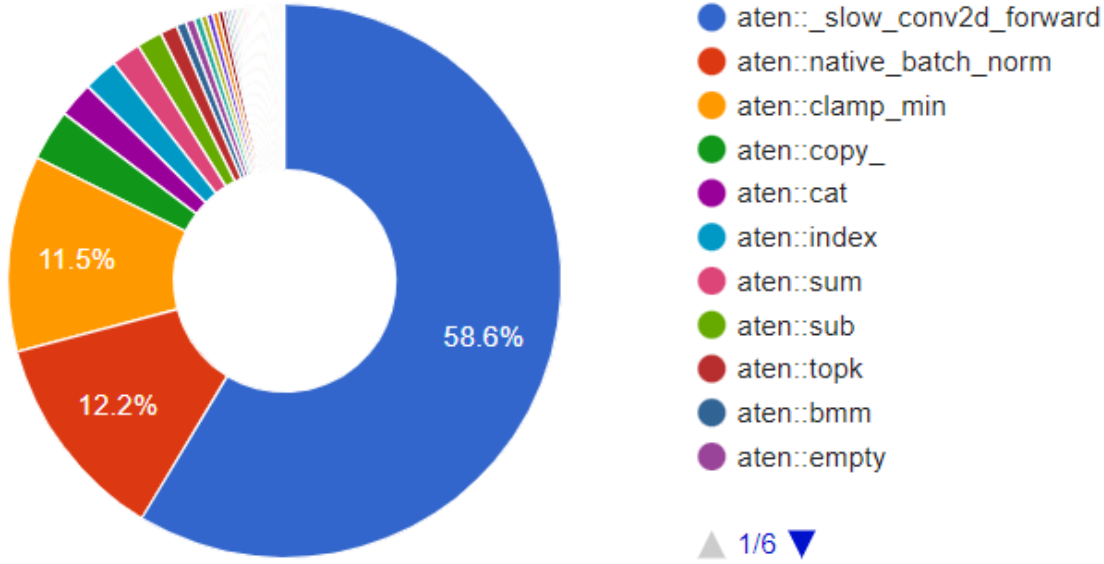
模型部署





CPU

FPGA



占总耗时的~70%

形成软硬件融合，CPU+FPGA的异构计算模式

模型部署

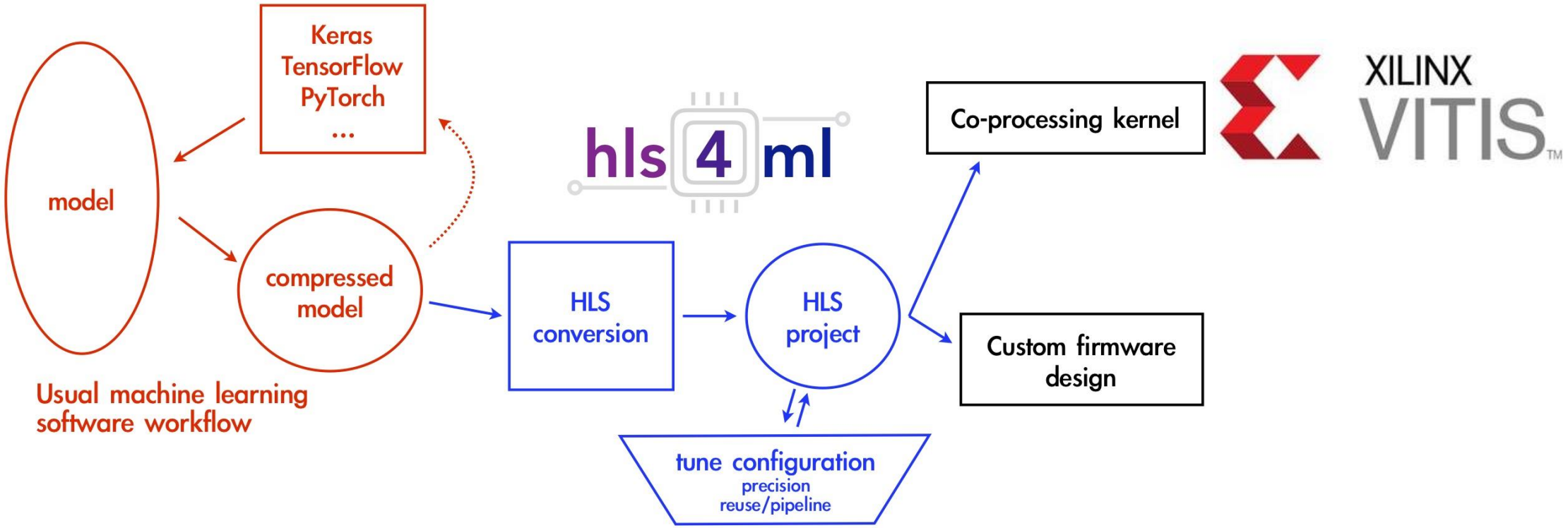
hls4ml

Applied Physicist, CERN

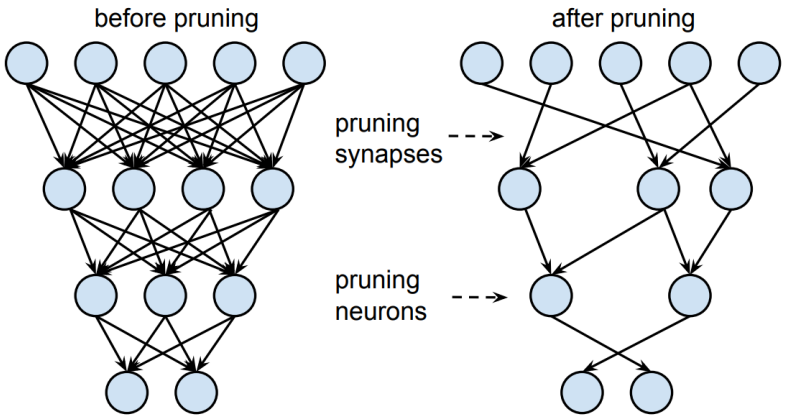
I work on the Level 1 Trigger of the CMS Experiment at the LHC. For the Phase 2 Upgrade this involves developing fast and efficient physics reconstruction algorithms for FPGAs. I've contributed to track finding, vertex reconstruction, particle flow, pileup substraction (PUPPI), jet reconstruction, and providing a platform for particle-based algorithms. I also develop algorithms and tools for machine learning in the trigger such as hls4ml and conifer.

来源：https://sioni.web.cern.ch/

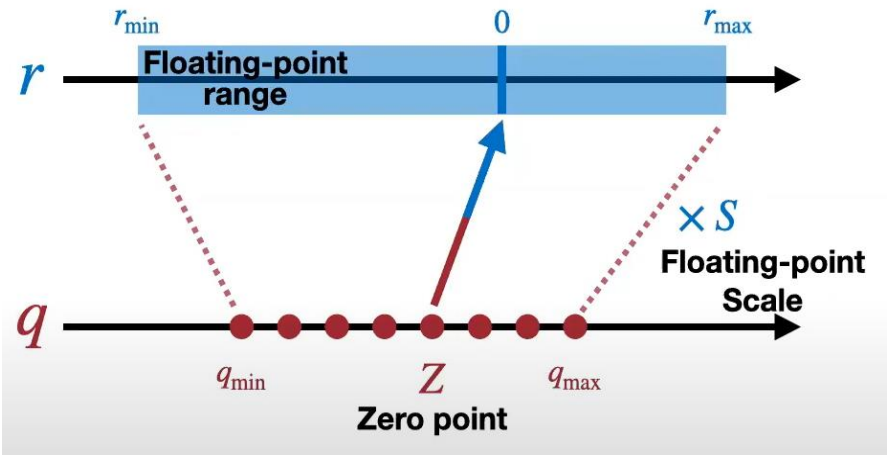
模型部署


Keras
TensorFlow
PyTorch
...
model
compressed model
Usual machine learning software workflow
hls4ml
HLS conversion
HLS project
tune configuration
precision
reuse/pipeline
Co-processing kernel
Custom firmware design
XILINX
VITIS

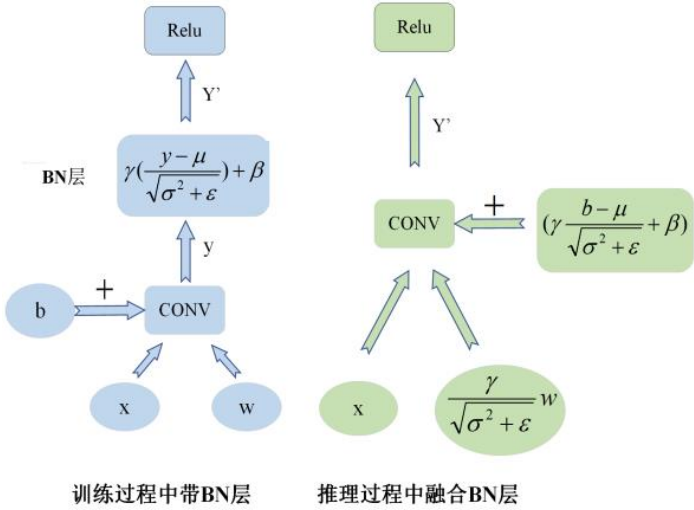
模型推理加速技术



剪枝



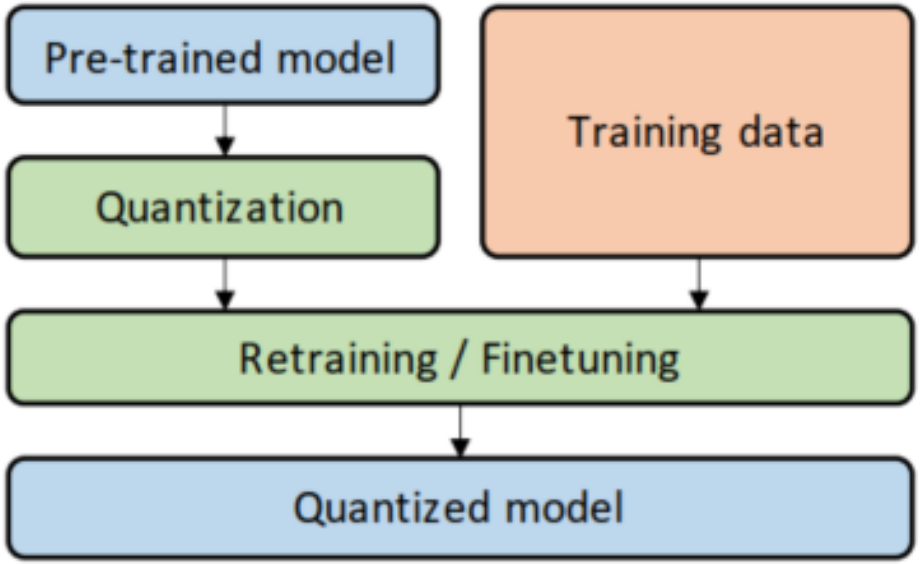
量化



融合

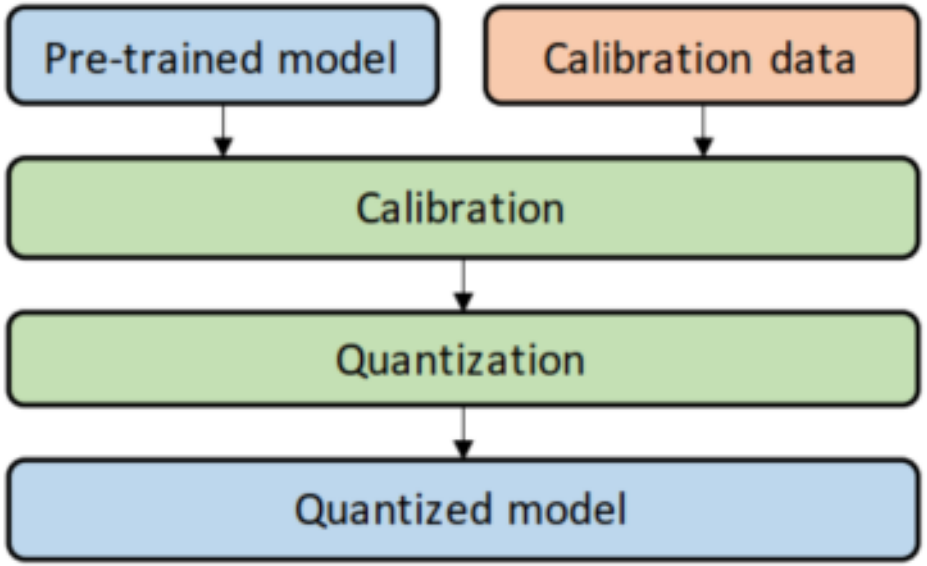
模型推理加速技术

量化类型：QAT和PTQ



量化感知训练： Quantization-aware training(QAT)

将训练过的模型量化后又再进行重训练。由于定点数值无法用于反向梯度计算，实际操作过程是在某些op前插入伪量化节点（fake quantization nodes）



训练后量化： Post-training quantization(PTQ)

使用一批校准数据对训练好的模型进行校准，将训练过的FP32网络直接转换为定点计算的网络，过程中无需对原始模型进行任何训练。

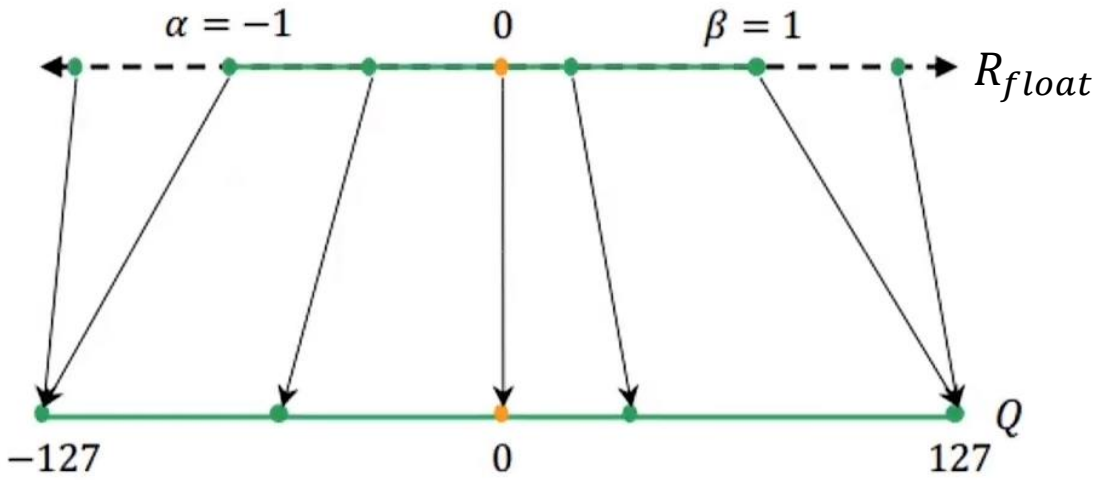
模型推理加速技术

量化的方法：

对称量化：

量化：$Q = clamp(round(\frac{R_{float}}{S}))$

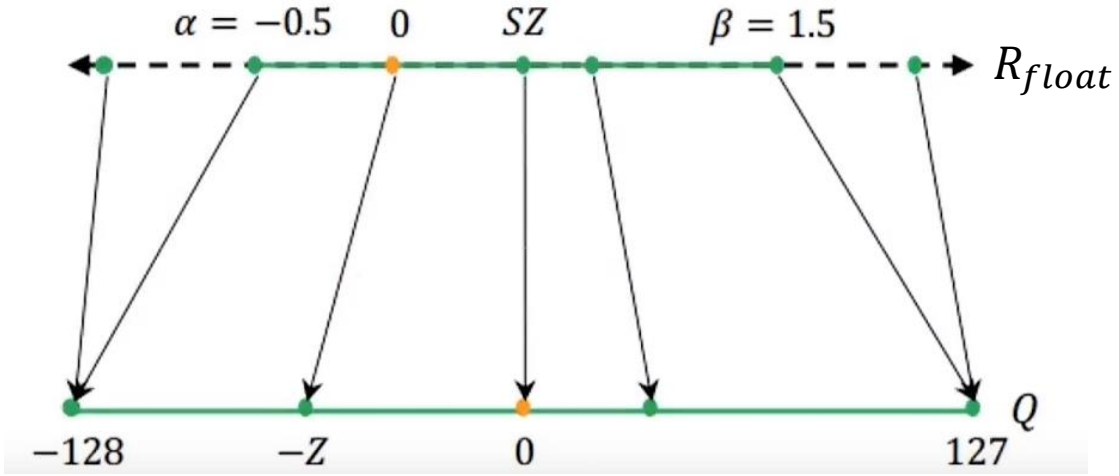
反量化：$R_{float} = S * Q$



非对称量化

量化：$Q = clamp(round(\frac{R_{float}}{S}) - Z)$

反量化：$R_{float} = S * (Q + Z)$



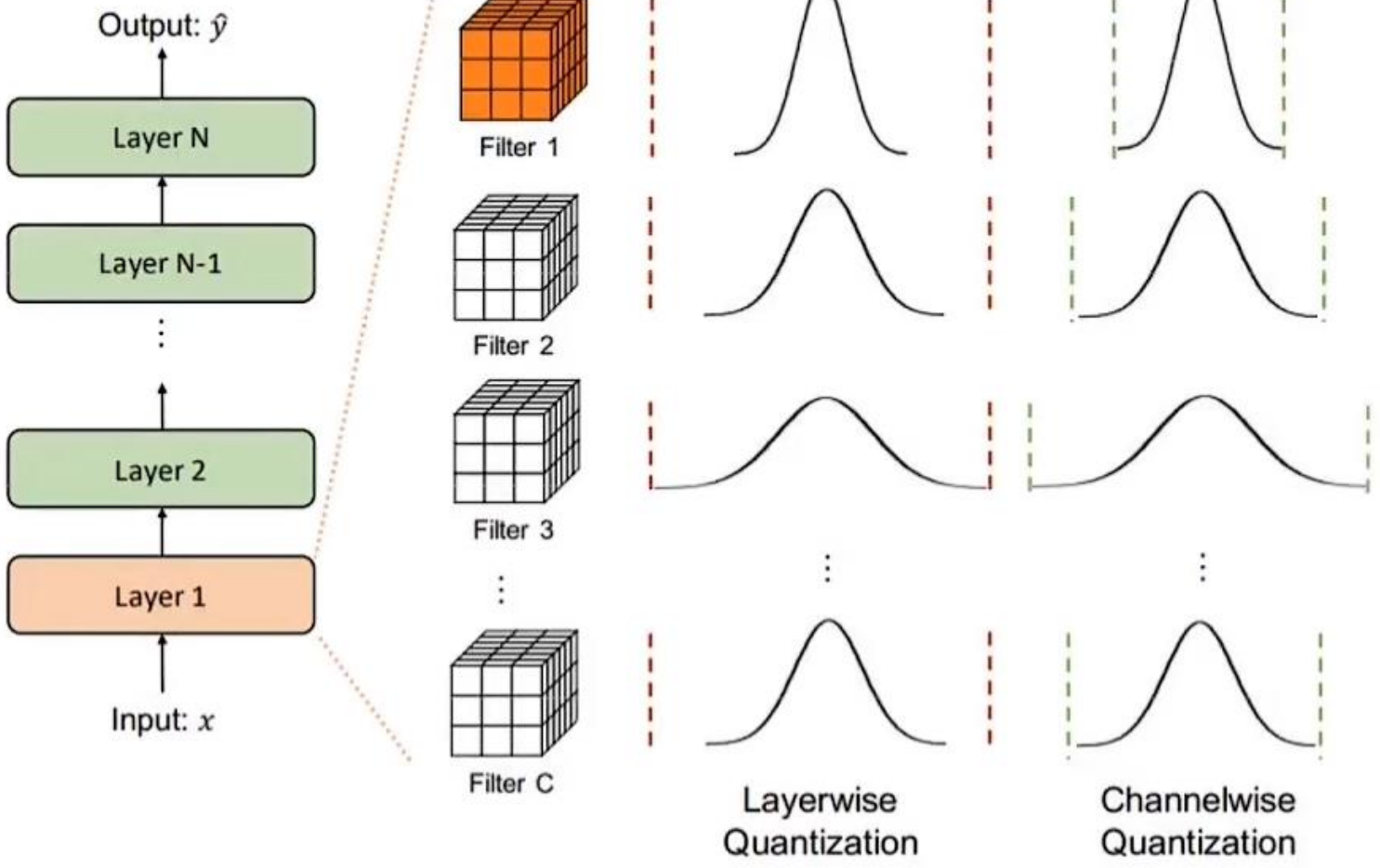
量化的方法：

逐层量化（Per Tensor）

每一层的权重具有相同的$scale$和$zero_point$

逐通道量化（Per Channel）

每通道的权重具有相同的$scale$和$zero_point$



ParticleNet模型量化

power of two 量化：**对称量化**的特殊情况，$scale = 2^x$ 。

采用后训练量化+逐层量化+对称量化的方式对 ParticeNet 模型进行了量化

	Parameters(B)	ACC(%)	Type
ParticleNet	**993KB**	**~93.9%**	**float32**
ParticleNet-quantization	**289KB**	**~93.3%**	**Int8**

```
correct:  0.933899998664856
正计算网络量化误差(SNR)，最后一层的误差应小于 0.1 以保证量化精度:
Analysing Graphwise Quantization Error(Phrase 1):: 100%| | 32/32 [00:00<00:00, 32.11it/s]
Analysing Graphwise Quantization Error(Phrase 2):: 100%| | 32/32 [00:01<00:00, 26.72it/s]
```

模型量化部署

浮点计算转定点计算

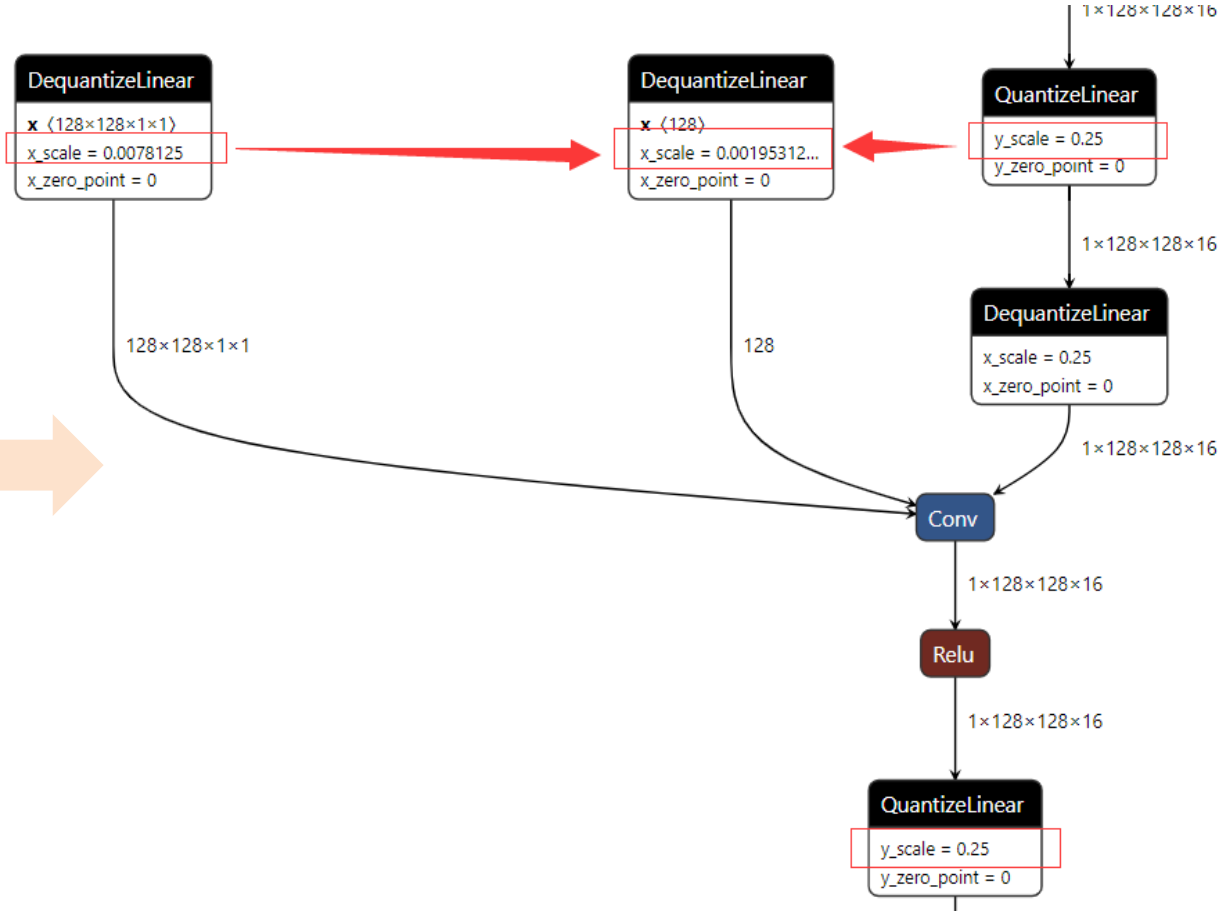
$$A_{n+1} = A_n \odot W_n + B_n$$

$$Aq_{n+1}S_{a_{n+1}} = Aq_nS_{a_n} \odot Wq_nS_{w_n} + Bq_nS_{b_n}$$

$$Aq_{n+1}S_{a_{n+1}} = \left(Aq_n \odot Wq_n\right)S_{a_n}S_{w_n} + Bq_nS_{b_n}$$

$$因为 S_{b_n} = S_{a_n}S_{w_n}$$

$$Aq_{n+1} = \left(Aq_n \odot Wq_n + Bq_n\right)\frac{S_{a_n}S_{w_n}}{S_{a_{n+1}}}$$



浮点计算转定点计算

$$Aq_{n+1} = \left(Aq_n \odot Wq_n + Bq_n \right) \frac{S_{a_n} S_{w_n}}{S_{a_{n+1}}}$$

因为 $S = 2^x$，因此

$$\frac{S_{a_n} S_{w_n}}{S_{a_{n+1}}} = 2^{a_n + w_n - a_{n+1}},$$

$$Aq_{n+1} = \left(Aq_n \odot Wq_n + Bq_n \right) 2^{a_n + w_n - a_{n+1}}$$

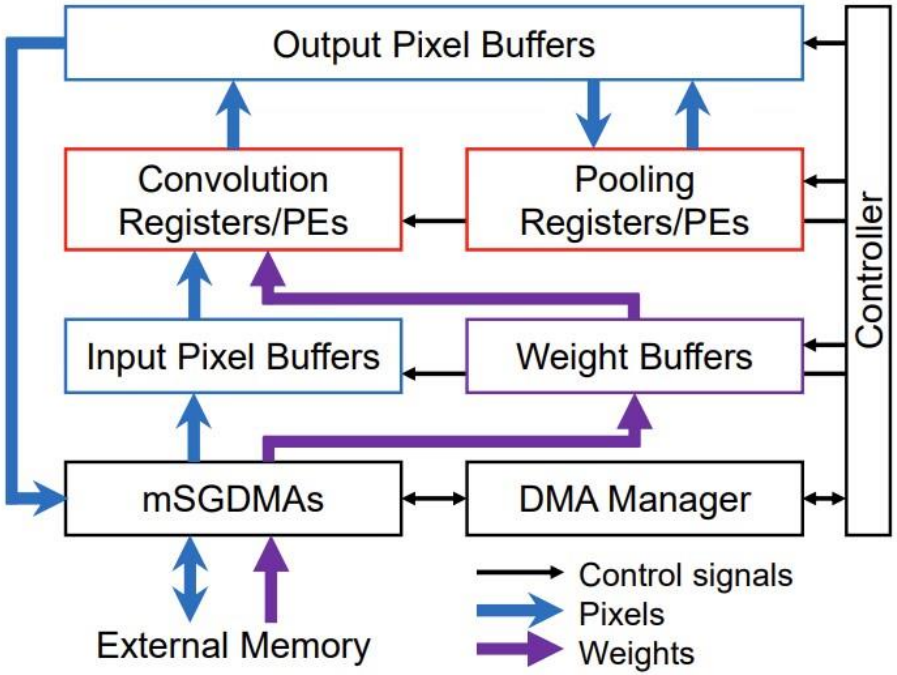
任意2^x运算，在硬件逻辑上可以进一步简化为：

$$\begin{cases} R \times 2^x = R << |x|, x \geq 0 \\ R \times 2^x = R >> |x|, x < 0 \end{cases}$$

$$log_2 \frac{S_{a_n} S_{w_n}}{S_{a_{n+1}}} = a_n + w_n - a_{n+1}$$

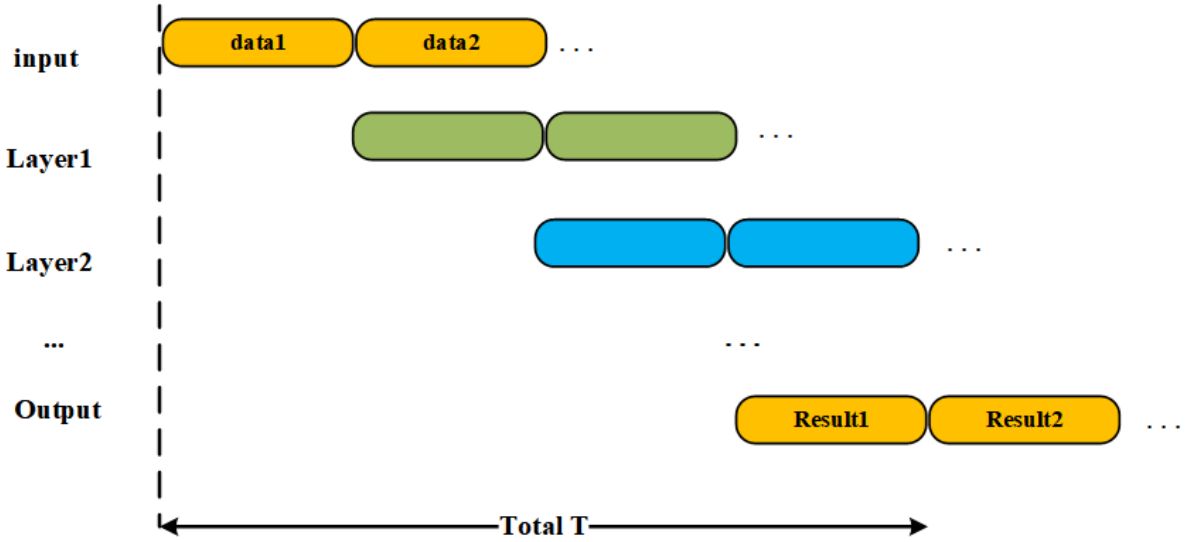
FPGA设计架构

- **通用架构：**
  - **优点：**能够进行通用加速，可针对不同的模型，不同的卷积核计算进行加速
  - **缺点：**每次从外部读取大量数据，每次将计算中间结果写回到外部内存或者主机端，会增加硬件平台功耗以及占用带宽
- **专用架构：**
  - **优点：**充分利用硬件平台的资源，对每个卷积层并行计算，无需将中间结果写回外部内存
  - **缺点：**若加速其他含有卷积的模型，需要重新配置，重新生成内核



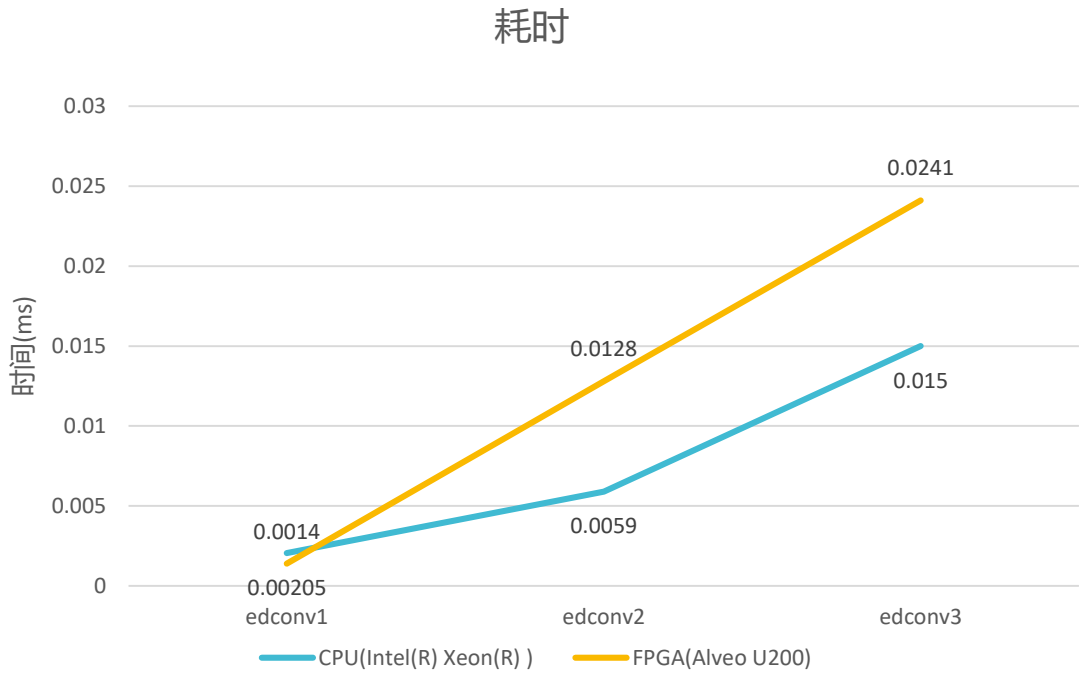
通用架构

FPGA专用架构实现

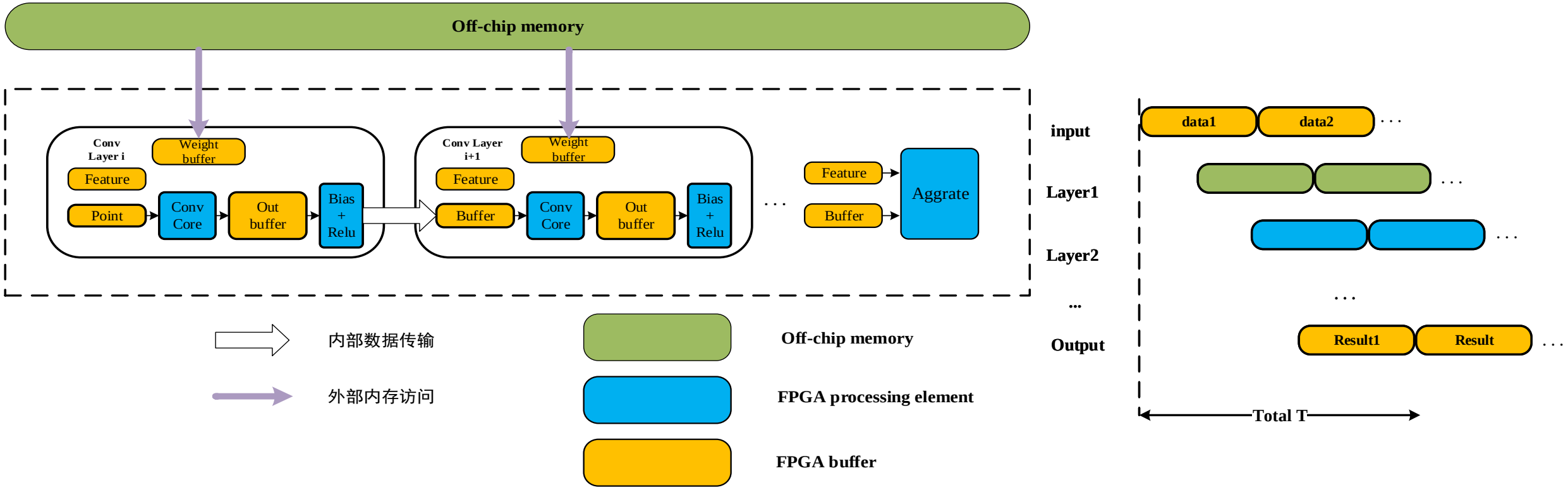


存在问题：

- 需等待上一层运算结束才开始下一层运算
- 已实现的性能存在进一步优化空间

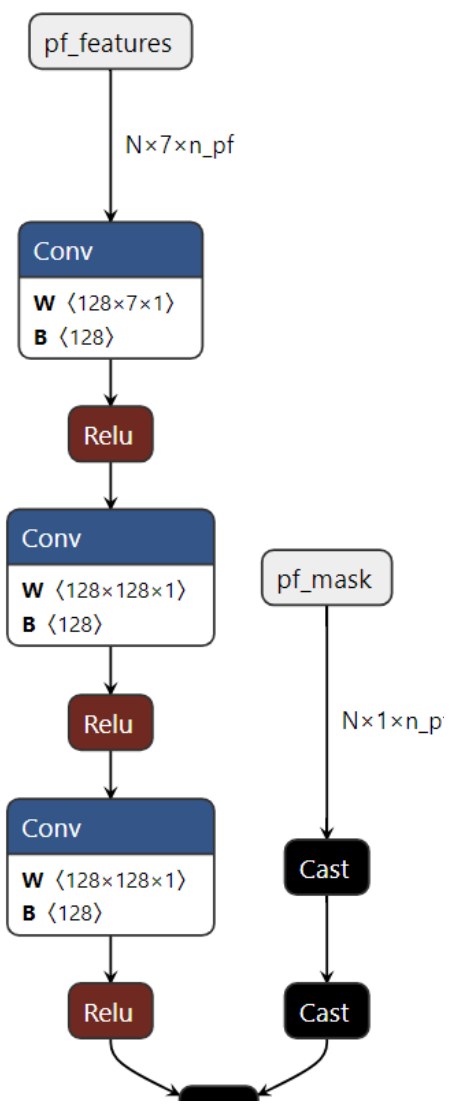


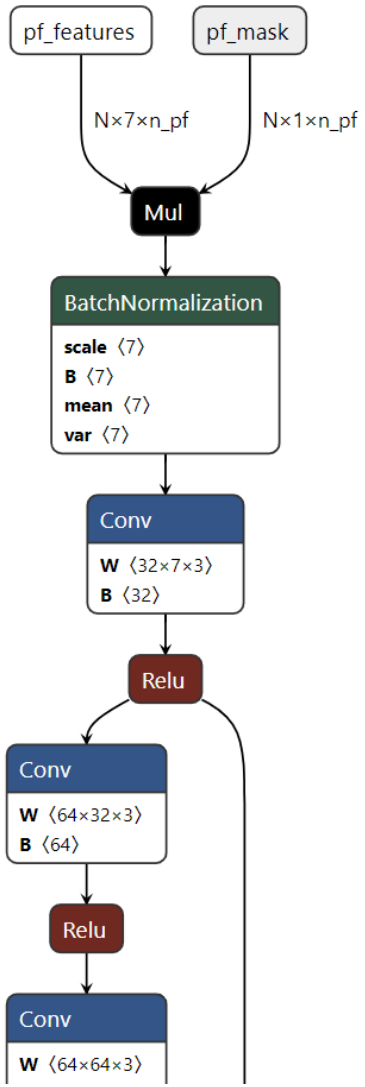
FPGA进一步优化思路



无需等待上一层计算出完整结果，即可开始下一层运算。

其他模型量化





	Parameters(B)	ACC(%)	Type
PFN	330KB	~89.6%	float32
PFN-quantization	94KB	~88.1%	Int8
PCNN	1.4MB	~89.6%	float32
PCNN-quantization	367KB	~89.4%	Int8

总结与展望

- 探索基于FPGA的ParticleNet，并研究了模型量化
- 将其他粒子物理模型进行量化，探索精度损失
- ParticleNet的实现需要进一步验证和优化

谢谢大家！