

QPO: Accelerating Memory-Efficient DNN Training with Quantization and Pipelining

Xiang Fan and Shaohuai Shi

School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
Shenzhen, China

xiangfan@stu.hit.edu.cn, shaohuais@hit.edu.cn

ICPADS 2025 (The 31st IEEE International Conference on Parallel and Distributed Systems)

Background

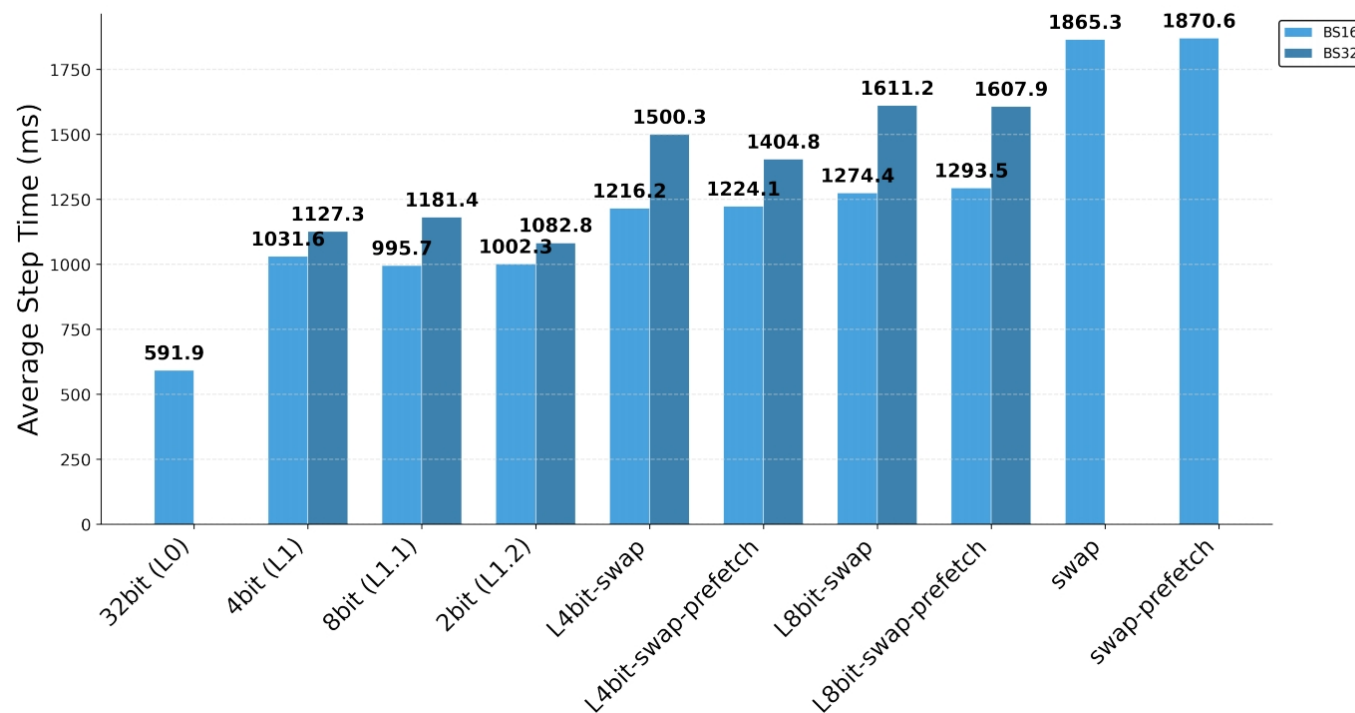
- 深度神经网络（DNN）训练需要大量的计算能力，并且还需要昂贵的设备内存来存储参数、梯度、优化器状态和激活值销。
- 卸载是一种广泛使用的内存高效技术，通过将临时数据从设备内存转移到CPU内存来节省设备内存。

Problem

- 卸载和上传操作容易导致显著的数据传输开销。
- 增加批处理大小和序列长度会导致卸载数据量线性增长，这会显著增加 CPU 到 GPU 和 GPU 到 CPU 的传输延迟，从而在训练性能中引入严重的瓶颈。

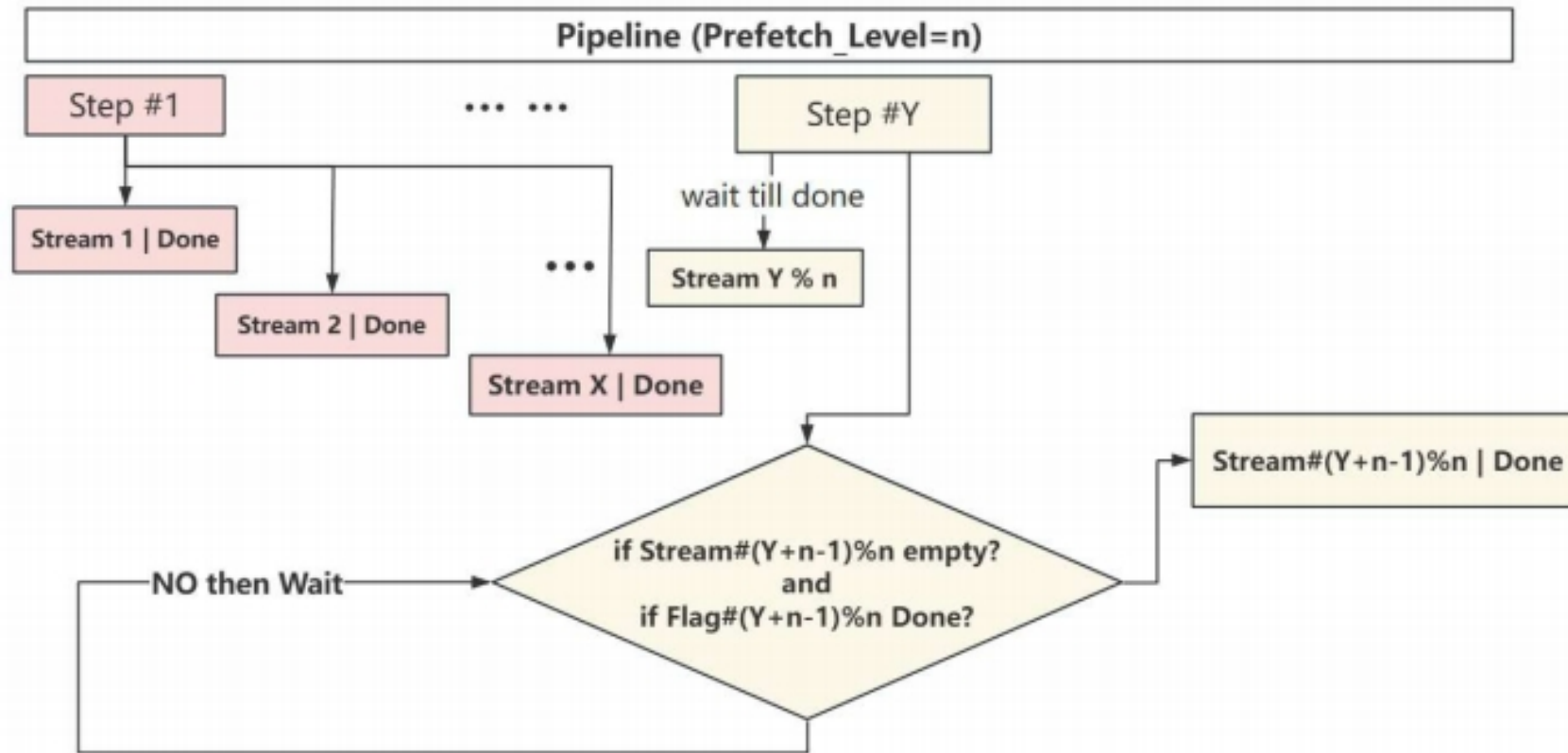
Existing solutions-GACT

Step Time Comparison



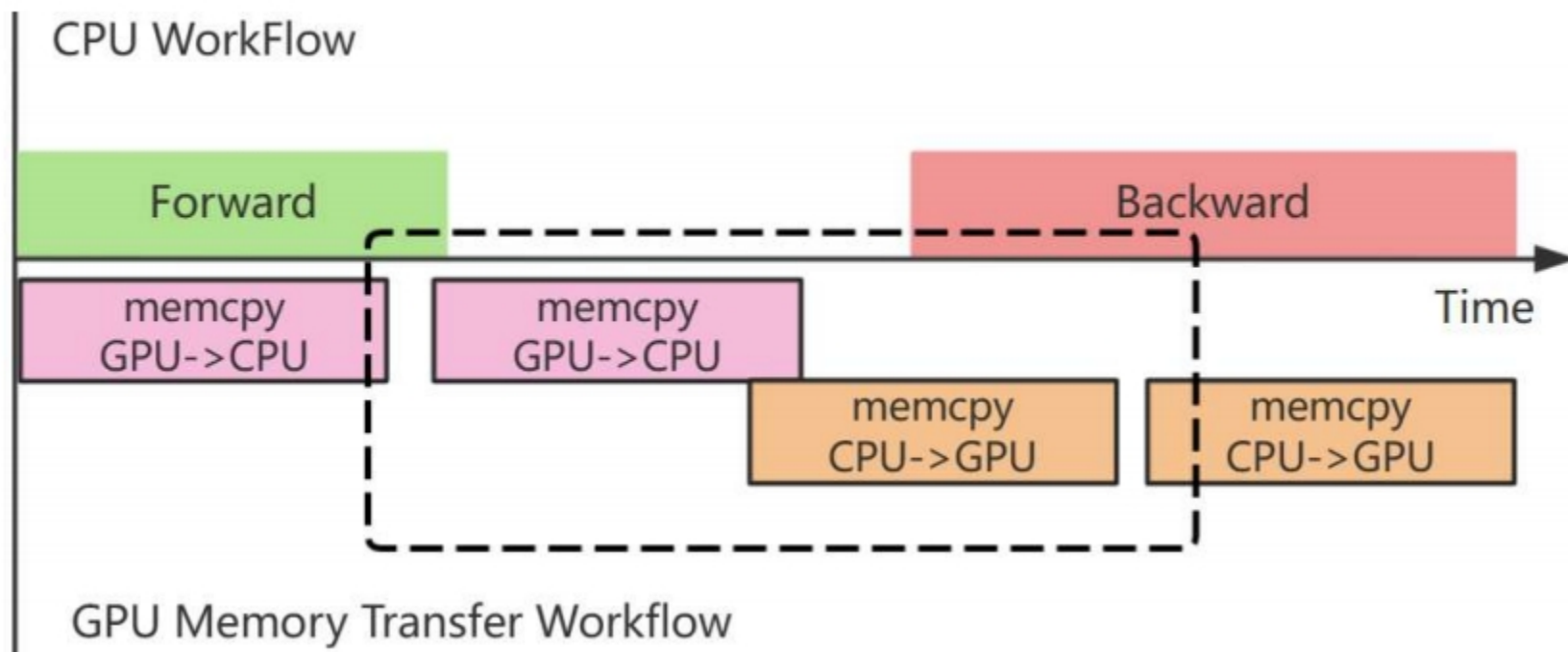
Strategy Abbreviation	Meaning
L0 (Level-0)	No compression
L1 (Level-1)	4-bit quantization
L1.1 (Level-1.1)	8-bit quantization
L1.2 (Level-1.2)	2-bit quantization
L2 (Level-2)	4-bit quantization with tensor size-based auto-adjustment
Swap	No compression + Swapping
Swap-Prefetch	No compression + Swapping + Prefetching
L4bit-swap	4-bit quantization + Swapping
L8bit-swap	8-bit quantization + Swapping
L4bit-swap-prefetch	4-bit quantization + Swapping + Prefetching
L8bit-swap-prefetch	8-bit quantization + Swapping + Prefetching

Basic idea



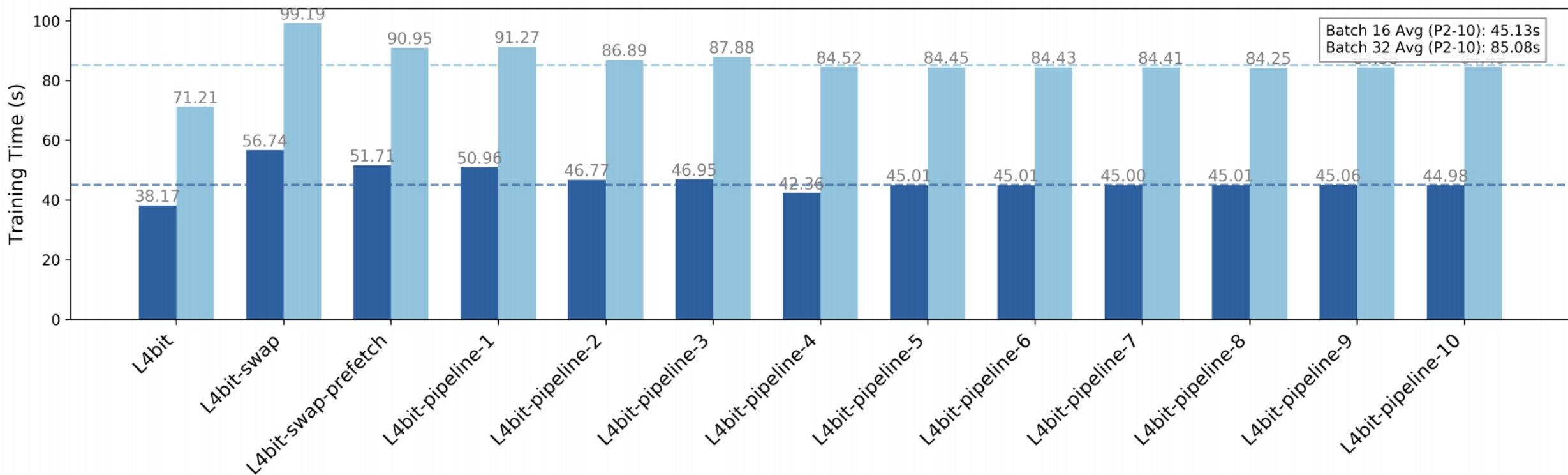
Method

解决前向后几层是反向前几层的等待问题，取消这几层的转移：



Experiments

25 Steps Training Time Comparison (On A100 32G)



Experiments

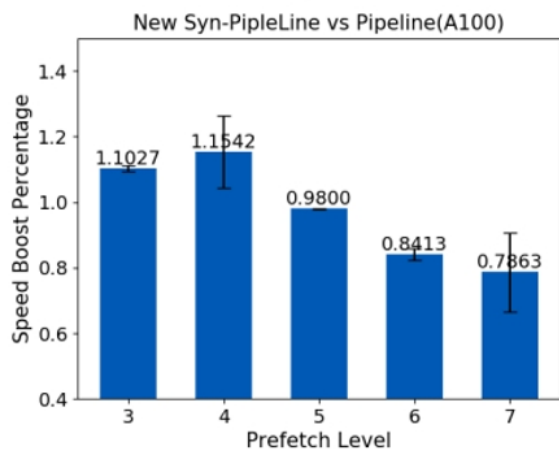


Fig. 12. Performance Comparison Between the Improved Pipeline and the Original Pipeline on A100 (GPT-2)

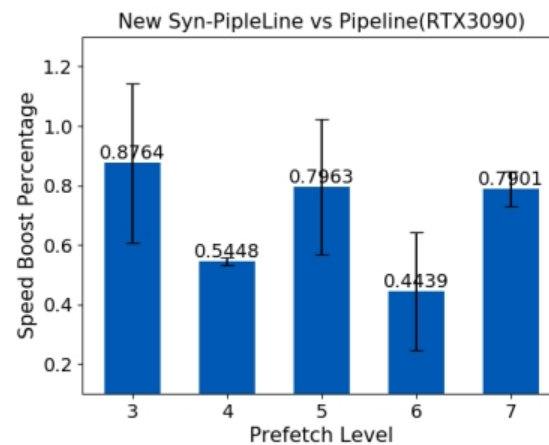


Fig. 13. Comparison Between the Improved Pipeline and the Original Pipeline on RTX 3090 (GPT-2)

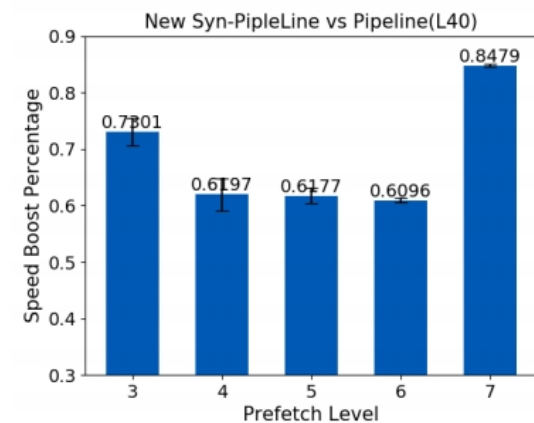


Fig. 14. Comparison Between the Improved Pipeline and the Original Pipeline on L40 (GPT-2)

思考

- 从内存优化这方面考虑下一步改进。
 - 激活值的卸载
 - 权重量化
 - 残差补偿低精度表示

双击此处添加标题

THANK

S