ECE 695 Programmable Accelerator Architecture

CUDA Programming Assignment 4 Report

Surya Selvam

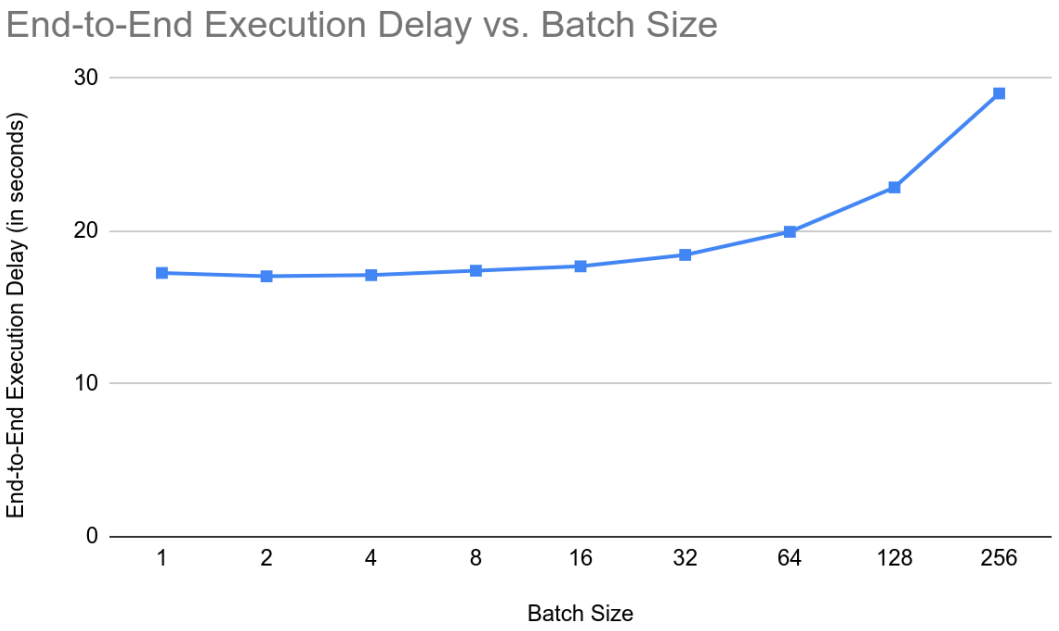
Part 1: AlexNet

Implementation

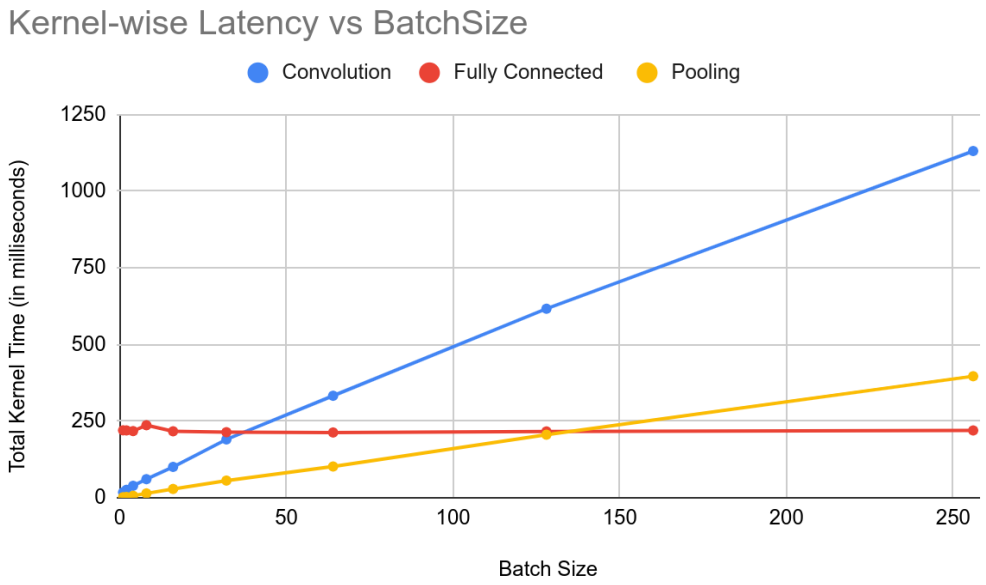
- Kernels of Convolution, Pooling and Fully Connected layers were reused from previous parts of assignment.
- **Unified Memory** was used to program the full network to reduce the explicit memory copies.
- Intermediate activations are stored in a single variable, so each kernel call is followed by device synchronization to maintain RAW dependency.

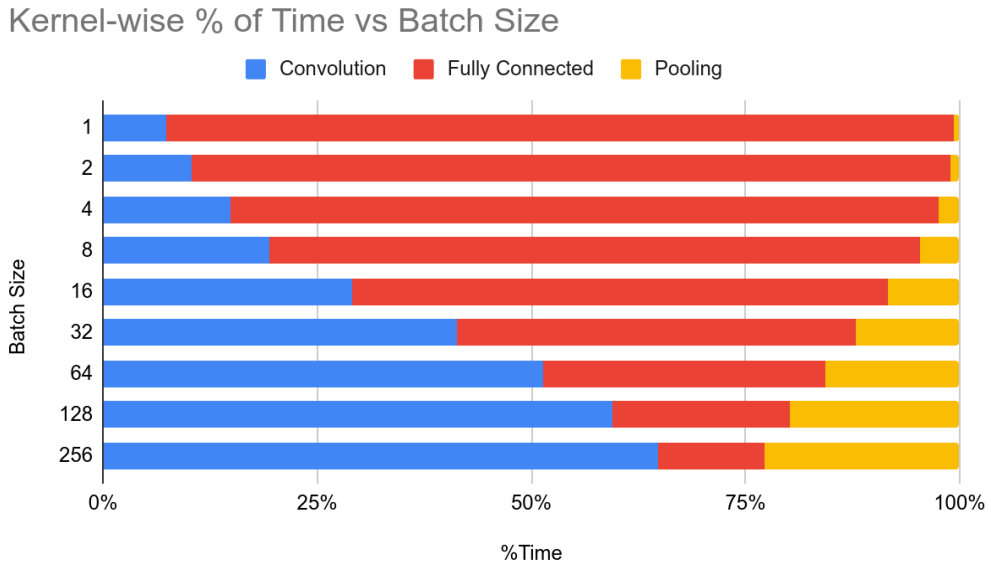
Performance

- **End-to-End Latency:** End-to-End AlexNet Inference latency is measured for various batch sizes. Batch size is varied from 1 to 256 as powers of 2. For smaller batch sizes, even with increase in computation, the *inference latency almost remains the same* because of Memory overheads (cudaMalloc). For higher batch size, with *doubling of batch size, the end-to-end latency also increases*. This is due to the increase in time spent in compute in addition to memory overheads.



- **Kernel-wise Latency:** AlexNet contains 5 convolution operations, 3 Fully connected layers and 3 pooling operations. Convolution and Pooling operations are *compute-bound* whereas FC layer is a *memory-bound* operation.
    - **Convolution and Pooling:** With increase in batch size, the amount of compute increases and total time spent on convolution operations also increases. From the first graph, we can observe that *kernel execution time scales linearly with the batch size*.
    - **Fully connected:** Since FC layer is a memory bound, even with the increase in batch size, the total execution time of FC Kernels remains the same.





Appendix: Table

Batch Size	AlexNet End-to-End Execution Delay (seconds)
1	17.2663
2	17.0471
4	17.1223
8	17.4128
16	17.6991
32	18.4474
64	19.9578
128	22.8655
256	29.0149

	Time Spent in (milliseconds)			Time %		
Batch Size	Convolution	Fully Connected	Pooling	Convolution	Fully Connected	Pooling
1	17.673	219.94	1.6817	7.39	91.91	0.7
2	25.807	219.58	2.5937	10.41	88.55	1.05
4	38.944	217.4	6.4686	14.82	82.72	2.46
8	60.556	236.57	14.142	19.45	76	4.54
16	100.58	216.73	28.659	29.07	62.64	8.28
32	190.22	213.84	55.976	41.35	46.48	12.17
64	332.66	212.86	101.9	51.38	32.88	15.74
128	616.58	216.28	205.72	59.37	20.82	19.81
256	1131.04	219.5	396.37	64.74	12.57	22.69