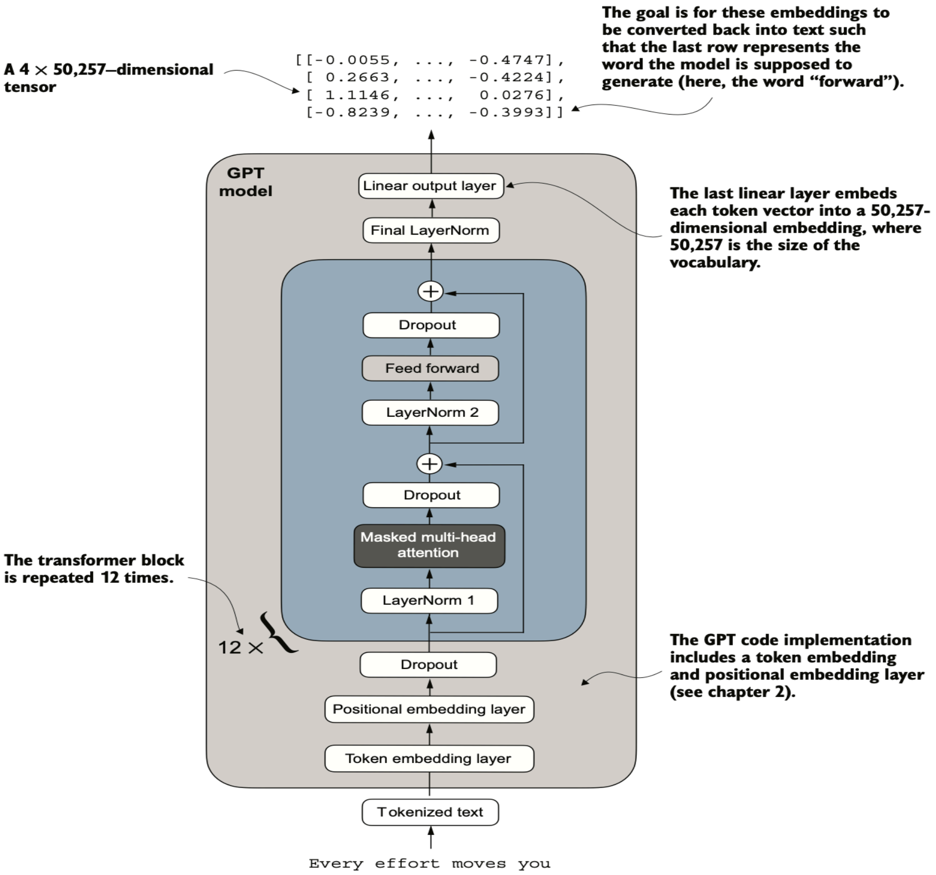
Beating PyTorch: Optimizing GPT-2 Small Model Inference through Custom CUDA Kernel on NVIDIA T4 GPUs

Shubham Ojha (so2754)   Mathew Martin (mm14460)

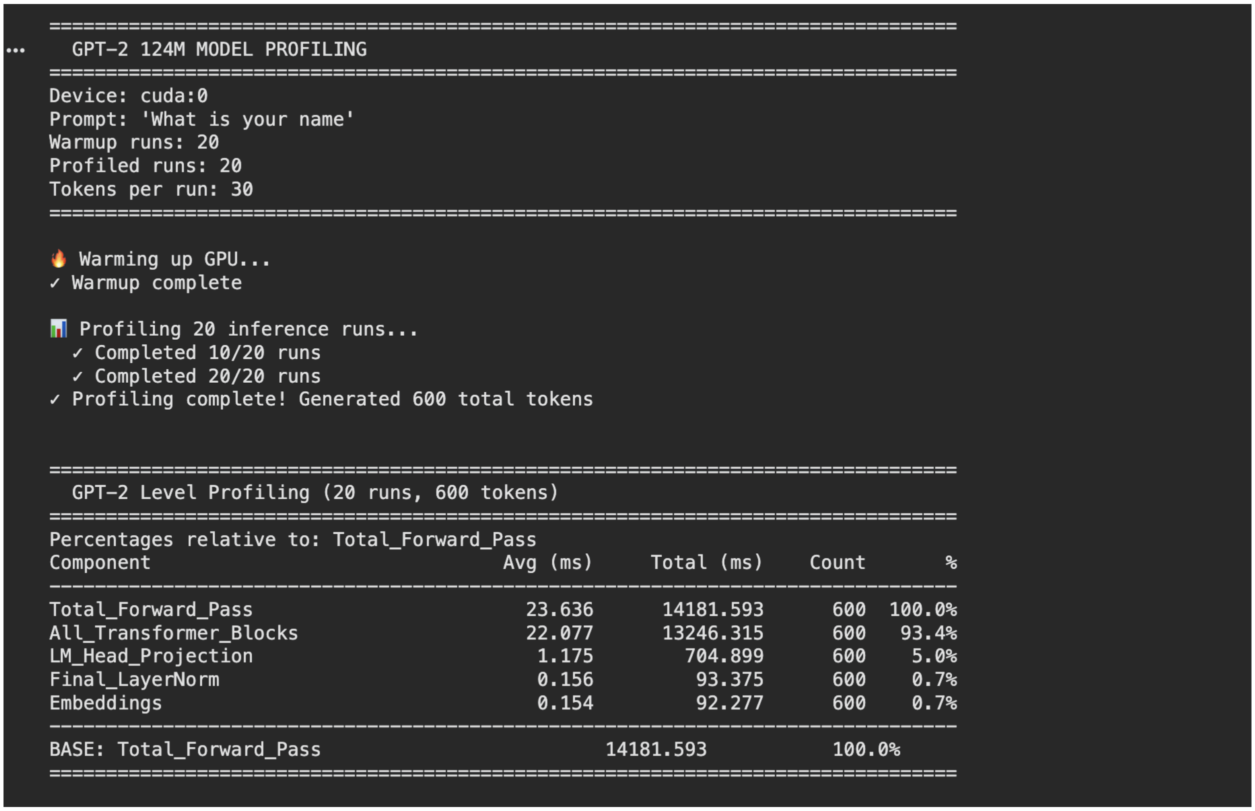
Project milestones completed

1. Implemented the GPT-2 124M parameter model from scratch in PyTorch and successfully loaded the official pretrained weights to enable inference.
2. Performed layer-wise profiling to identify the execution time distribution across model layers. Based on these results, conducted CUDA kernel-level profiling for the top 5 most time-consuming layers.
3. Used insights from both profiling stages to analyze potential optimization opportunities, including identifying which CUDA kernel fusions could yield meaningful performance improvements.

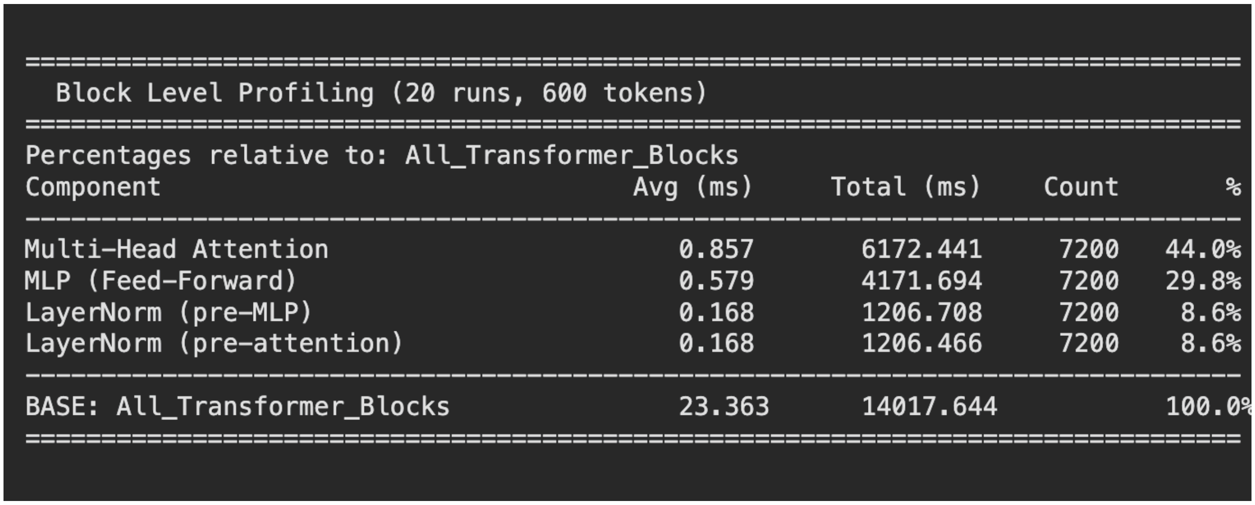

A 4 × 50,257–dimensional tensor
[[-0.0055, ..., -0.4747],
[0.2663, ..., -0.4224],
[1.1146, ..., 0.0276],
[-0.8239, ..., -0.3993]]
The goal is for these embeddings to be converted back into text such that the last row represents the word the model is supposed to generate (here, the word "forward").
GPT model
Linear output layer
Final LayerNorm
The last linear layer embeds each token vector into a 50,257-dimensional embedding, where 50,257 is the size of the vocabulary.
Dropout
Feed forward
LayerNorm 2
Dropout
Masked multi-head attention
LayerNorm 1
The transformer block is repeated 12 times.
12 ×
Dropout
Positional embedding layer
Token embedding layer
The GPT code implementation includes a token embedding and positional embedding layer (see chapter 2).
Tokenized text
Every effort moves you

Results :-

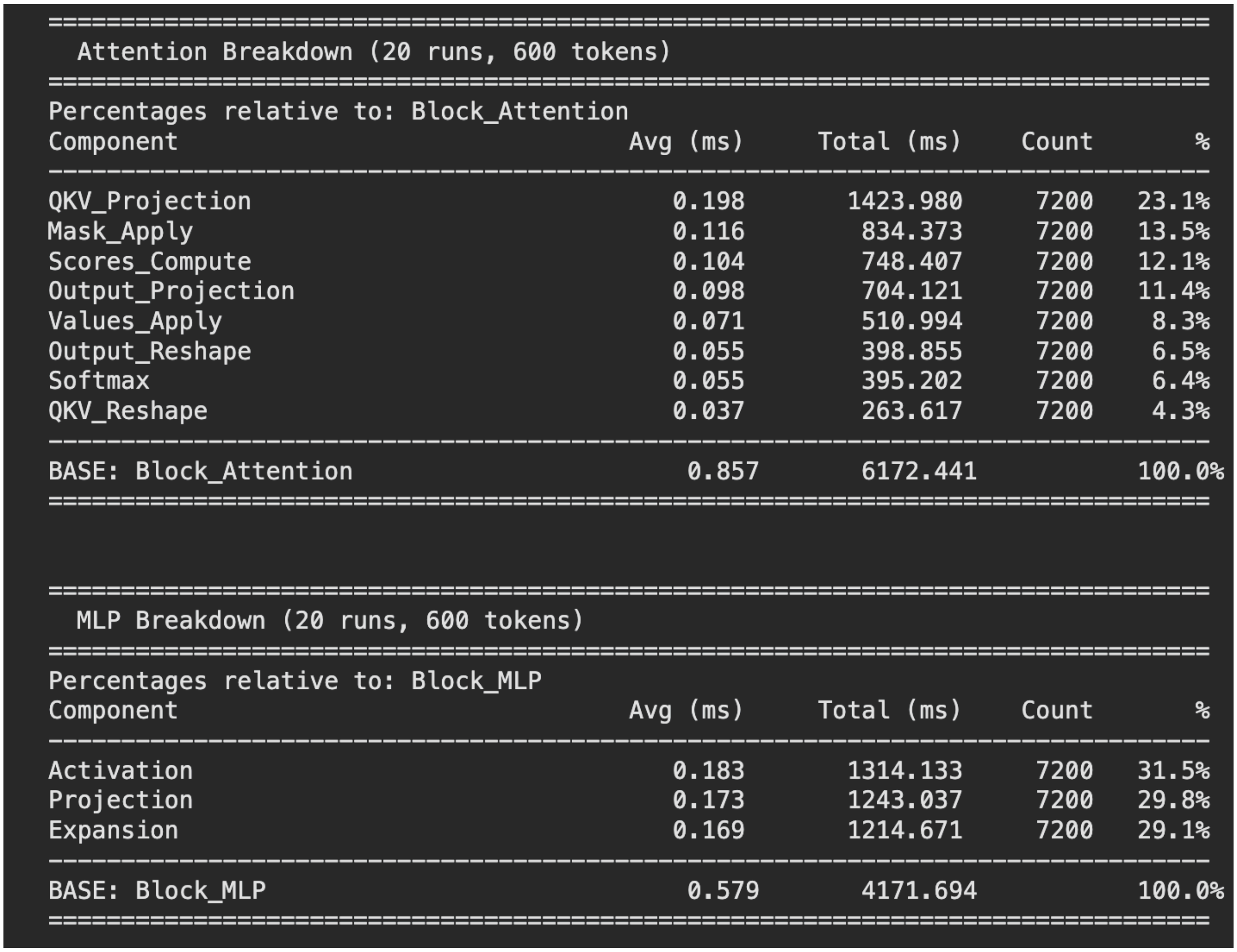
Profiling at model layer level:-



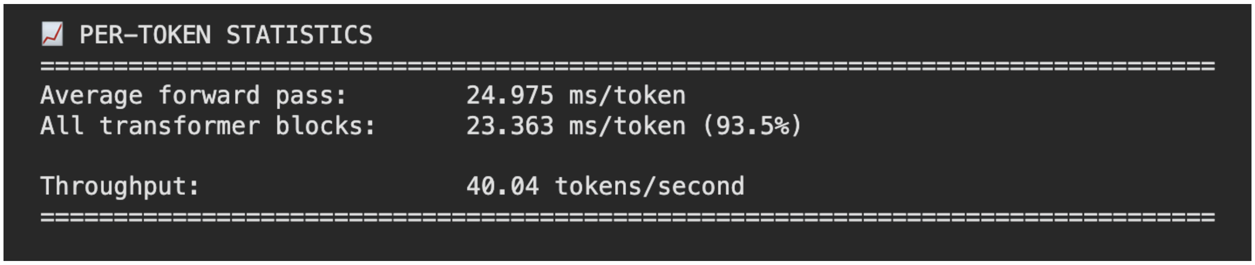
```
================================================================================
  GPT-2 124M MODEL PROFILING
================================================================================
Device: cuda:0
Prompt: 'What is your name'
Warmup runs: 20
Profiled runs: 20
Tokens per run: 30
================================================================================

🔥 Warming up GPU...
✓ Warmup complete

📊 Profiling 20 inference runs...
  ✓ Completed 10/20 runs
  ✓ Completed 20/20 runs
✓ Profiling complete! Generated 600 total tokens


================================================================================
  GPT-2 Level Profiling (20 runs, 600 tokens)
================================================================================
Percentages relative to: Total_Forward_Pass
Component                               Avg (ms)     Total (ms)    Count       %
--------------------------------------------------------------------------------
Total_Forward_Pass                        23.636      14181.593      600  100.0%
All_Transformer_Blocks                    22.077      13246.315      600   93.4%
LM_Head_Projection                         1.175        704.899      600    5.0%
Final_LayerNorm                            0.156         93.375      600    0.7%
Embeddings                                 0.154         92.277      600    0.7%
--------------------------------------------------------------------------------
BASE: Total_Forward_Pass                         14181.593           100.0%
================================================================================
```



```
================================================================================
  Block Level Profiling (20 runs, 600 tokens)
================================================================================
Percentages relative to: All_Transformer_Blocks
Component                               Avg (ms)     Total (ms)    Count       %
--------------------------------------------------------------------------------
Multi-Head Attention                       0.857       6172.441     7200   44.0%
MLP (Feed-Forward)                         0.579       4171.694     7200   29.8%
LayerNorm (pre-MLP)                        0.168       1206.708     7200    8.6%
LayerNorm (pre-attention)                  0.168       1206.466     7200    8.6%
--------------------------------------------------------------------------------
BASE: All_Transformer_Blocks               23.363      14017.644           100.0%
================================================================================
```



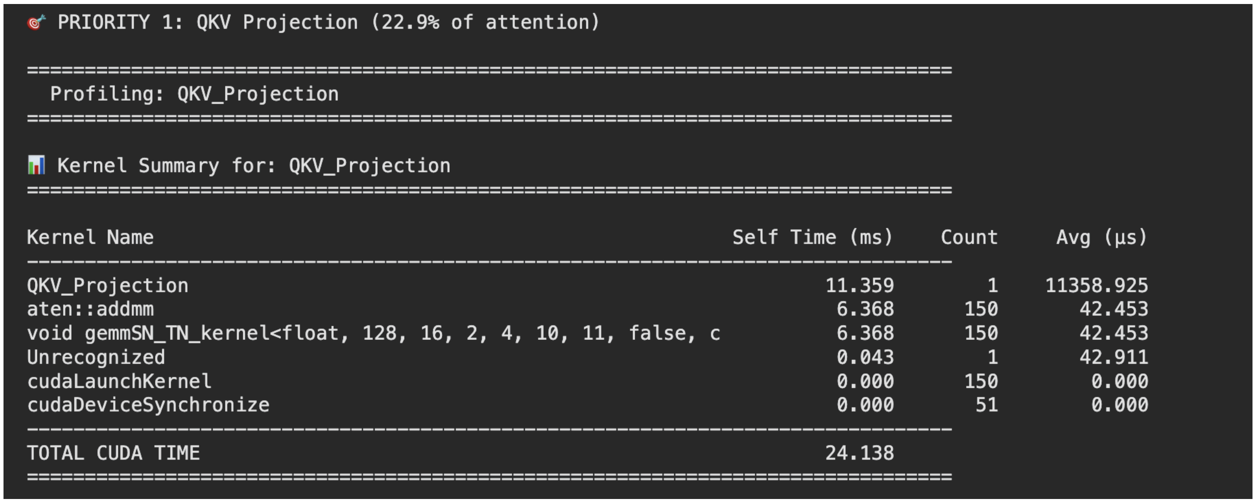
```
================================================================================
  Attention Breakdown (20 runs, 600 tokens)
================================================================================
Percentages relative to: Block_Attention
Component                               Avg (ms)     Total (ms)    Count       %
--------------------------------------------------------------------------------
QKV_Projection                             0.198       1423.980     7200   23.1%
Mask_Apply                                 0.116        834.373     7200   13.5%
Scores_Compute                             0.104        748.407     7200   12.1%
Output_Projection                          0.098        704.121     7200   11.4%
Values_Apply                               0.071        510.994     7200    8.3%
Output_Reshape                             0.055        398.855     7200    6.5%
Softmax                                    0.055        395.202     7200    6.4%
QKV_Reshape                                0.037        263.617     7200    4.3%
--------------------------------------------------------------------------------
BASE: Block_Attention                       0.857       6172.441           100.0%
================================================================================


================================================================================
  MLP Breakdown (20 runs, 600 tokens)
================================================================================
Percentages relative to: Block_MLP
Component                               Avg (ms)     Total (ms)    Count       %
--------------------------------------------------------------------------------
Activation                                 0.183       1314.133     7200   31.5%
Projection                                 0.173       1243.037     7200   29.8%
Expansion                                  0.169       1214.671     7200   29.1%
--------------------------------------------------------------------------------
BASE: Block_MLP                             0.579       4171.694           100.0%
================================================================================
```



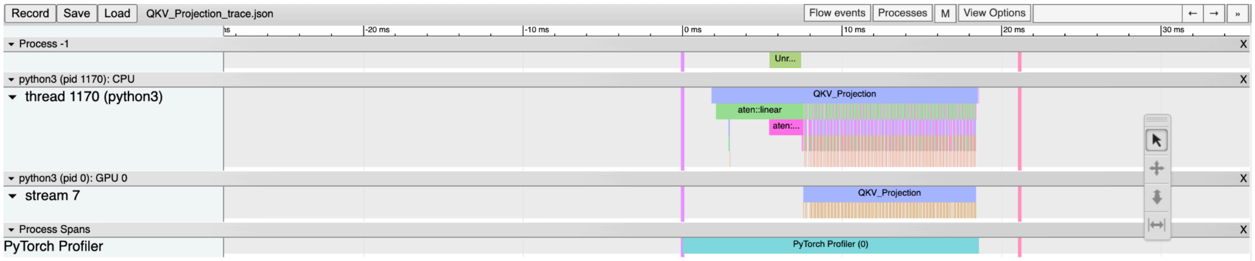
```
📈 PER-TOKEN STATISTICS
================================================================================
Average forward pass:        24.975 ms/token
All transformer blocks:      23.363 ms/token (93.5%)

Throughput:                  40.04 tokens/second
================================================================================
```

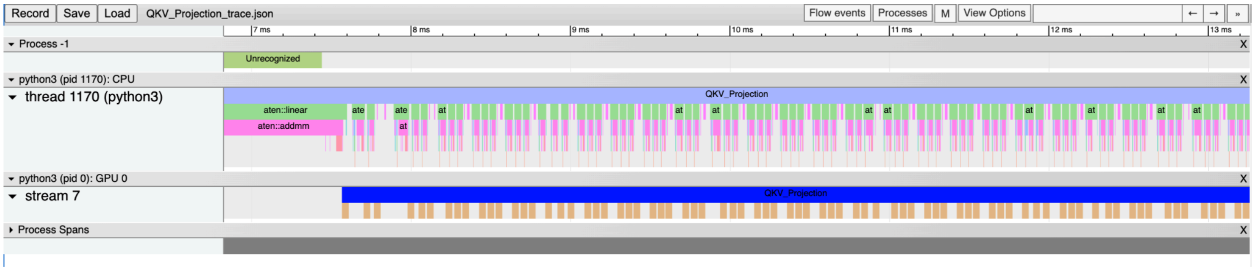
From Model layer level profiling, we get an idea that following layers take max time :- QKV projection, attention score computation, MLP expansion + GELU, MLP projection and softmax + values

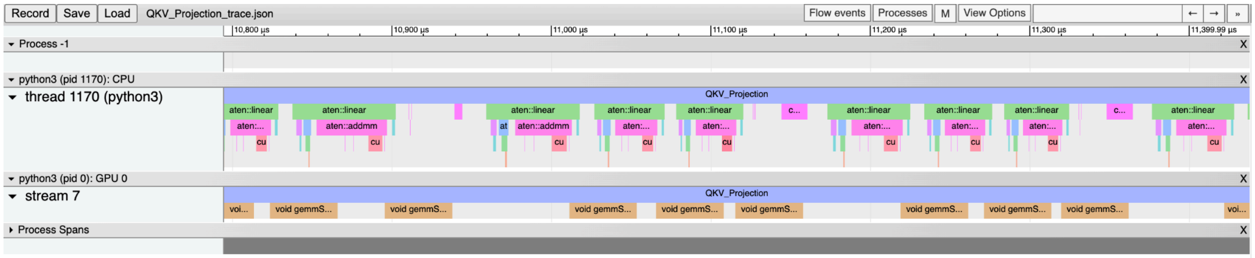
Profiling at Cuda Kernel layer level:-

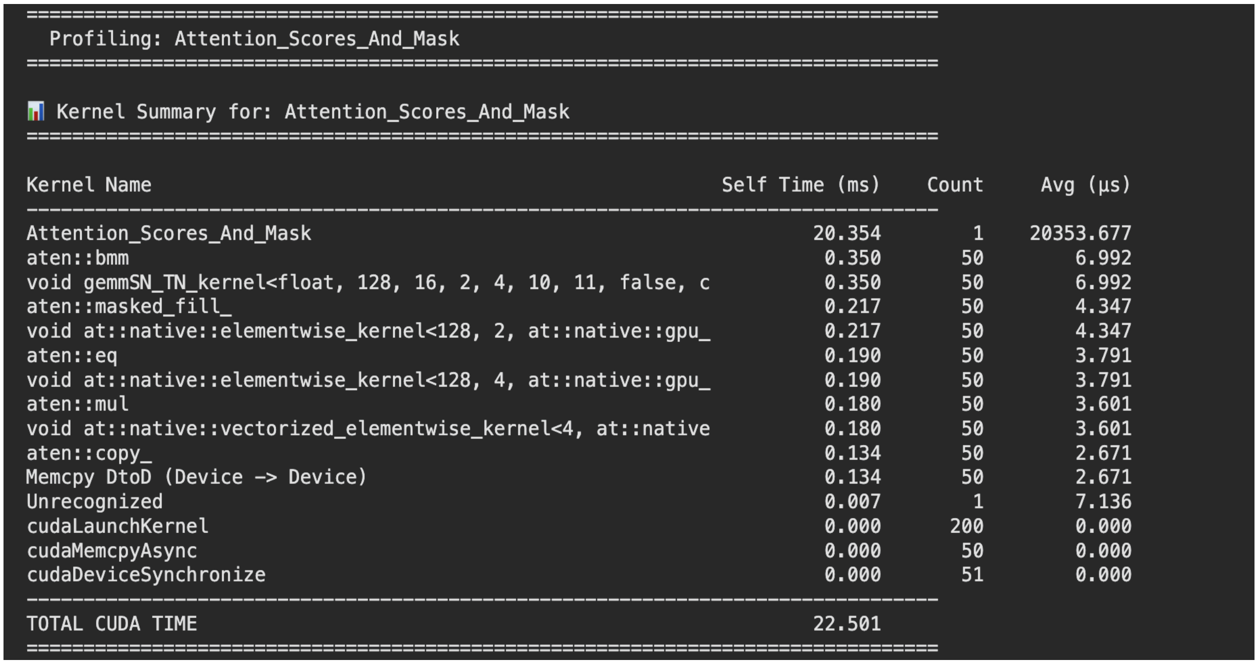


```
🎯 PRIORITY 1: QKV Projection (22.9% of attention)

================================================================================
  Profiling: QKV_Projection
================================================================================

📊 Kernel Summary for: QKV_Projection
================================================================================

Kernel Name                                                 Self Time (ms)   Count     Avg (µs)
--------------------------------------------------------------------------------
QKV_Projection                                                      11.359       1    11358.925
aten::addmm                                                          6.368     150       42.453
void gemmSN_TN_kernel<float, 128, 16, 2, 4, 10, 11, false, c         6.368     150       42.453
Unrecognized                                                         0.043       1       42.911
cudaLaunchKernel                                                     0.000     150        0.000
cudaDeviceSynchronize                                                0.000      51        0.000
--------------------------------------------------------------------------------
TOTAL CUDA TIME                                                     24.138
================================================================================
```

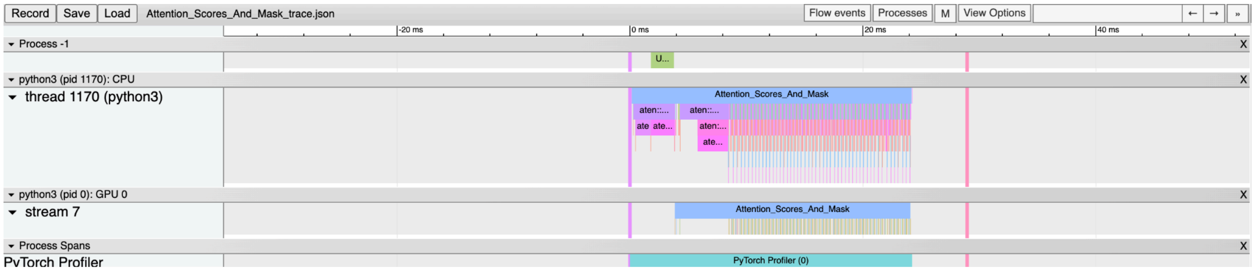


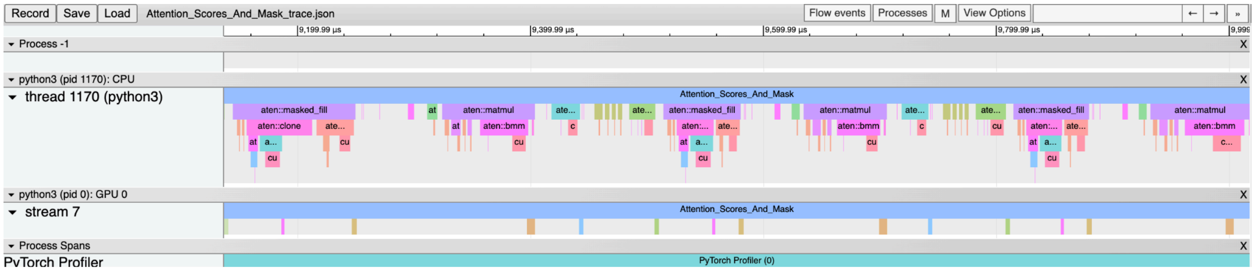


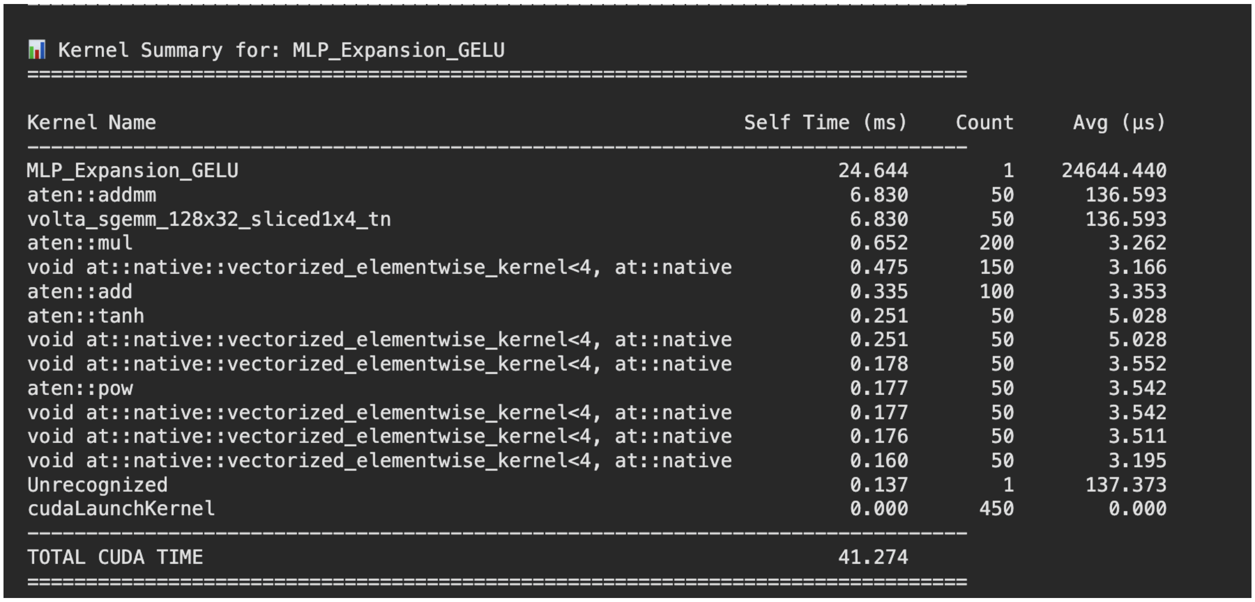




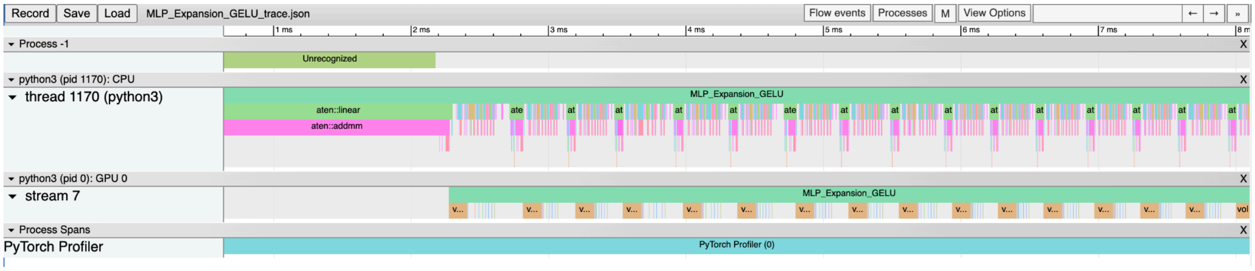
```
================================================================================
  Profiling: Attention_Scores_And_Mask
================================================================================

📊 Kernel Summary for: Attention_Scores_And_Mask
================================================================================

Kernel Name                                                Self Time (ms)    Count     Avg (µs)
--------------------------------------------------------------------------------
Attention_Scores_And_Mask                                          20.354        1    20353.677
aten::bmm                                                           0.350       50        6.992
void gemmSN_TN_kernel<float, 128, 16, 2, 4, 10, 11, false, c        0.350       50        6.992
aten::masked_fill_                                                  0.217       50        4.347
void at::native::elementwise_kernel<128, 2, at::native::gpu_        0.217       50        4.347
aten::eq                                                            0.190       50        3.791
void at::native::elementwise_kernel<128, 4, at::native::gpu_        0.190       50        3.791
aten::mul                                                           0.180       50        3.601
void at::native::vectorized_elementwise_kernel<4, at::native        0.180       50        3.601
aten::copy_                                                         0.134       50        2.671
Memcpy DtoD (Device -> Device)                                      0.134       50        2.671
Unrecognized                                                        0.007        1        7.136
cudaLaunchKernel                                                    0.000      200        0.000
cudaMemcpyAsync                                                     0.000       50        0.000
cudaDeviceSynchronize                                               0.000       51        0.000
--------------------------------------------------------------------------------
TOTAL CUDA TIME                                                    22.501
================================================================================
```

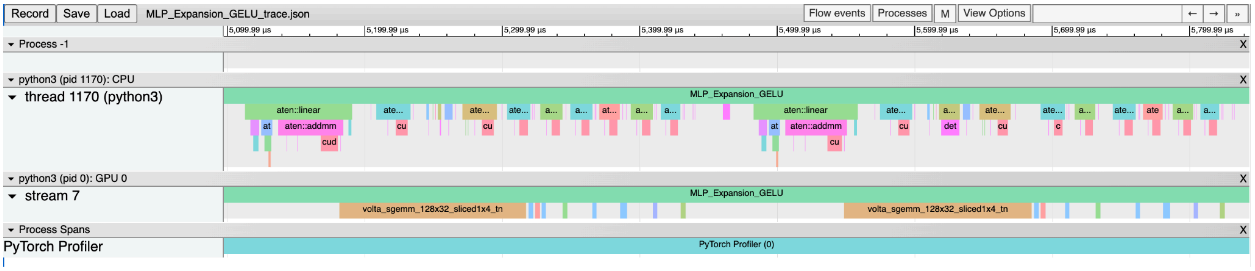


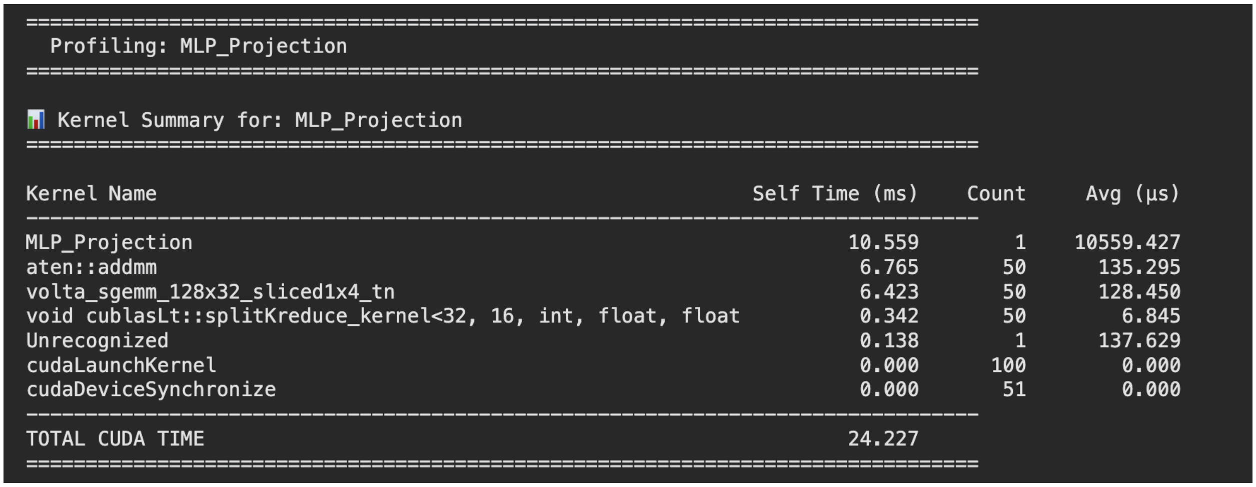




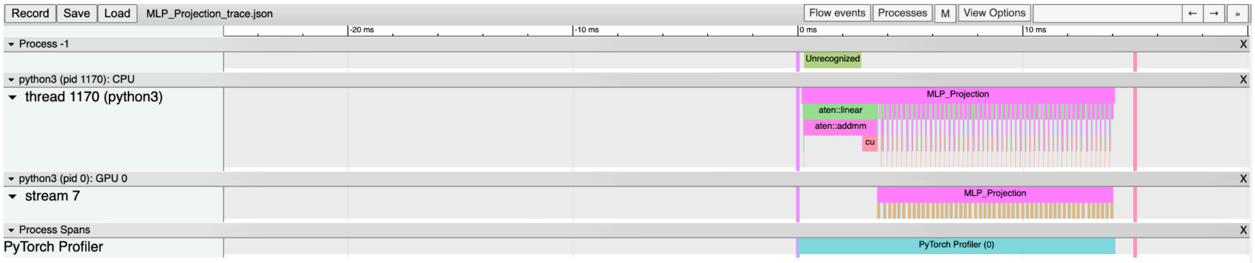
```
--------------------------------------------------------------------------------

📊 Kernel Summary for: MLP_Expansion_GELU
================================================================================

Kernel Name                                                 Self Time (ms)    Count     Avg (µs)
--------------------------------------------------------------------------------
MLP_Expansion_GELU                                                  24.644        1    24644.440
aten::addmm                                                          6.830       50      136.593
volta_sgemm_128x32_sliced1x4_tn                                      6.830       50      136.593
aten::mul                                                            0.652      200        3.262
void at::native::vectorized_elementwise_kernel<4, at::native         0.475      150        3.166
aten::add                                                            0.335      100        3.353
aten::tanh                                                           0.251       50        5.028
void at::native::vectorized_elementwise_kernel<4, at::native         0.251       50        5.028
void at::native::vectorized_elementwise_kernel<4, at::native         0.178       50        3.552
aten::pow                                                            0.177       50        3.542
void at::native::vectorized_elementwise_kernel<4, at::native         0.177       50        3.542
void at::native::vectorized_elementwise_kernel<4, at::native         0.176       50        3.511
void at::native::vectorized_elementwise_kernel<4, at::native         0.160       50        3.195
Unrecognized                                                         0.137        1      137.373
cudaLaunchKernel                                                     0.000      450        0.000
--------------------------------------------------------------------------------
TOTAL CUDA TIME                                                     41.274
================================================================================
```

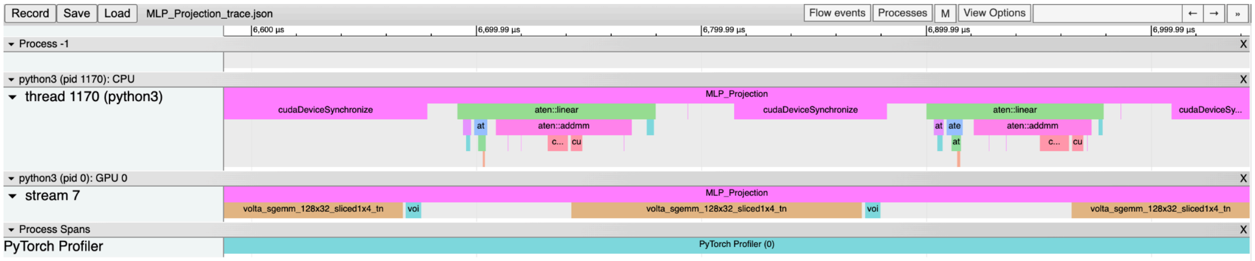


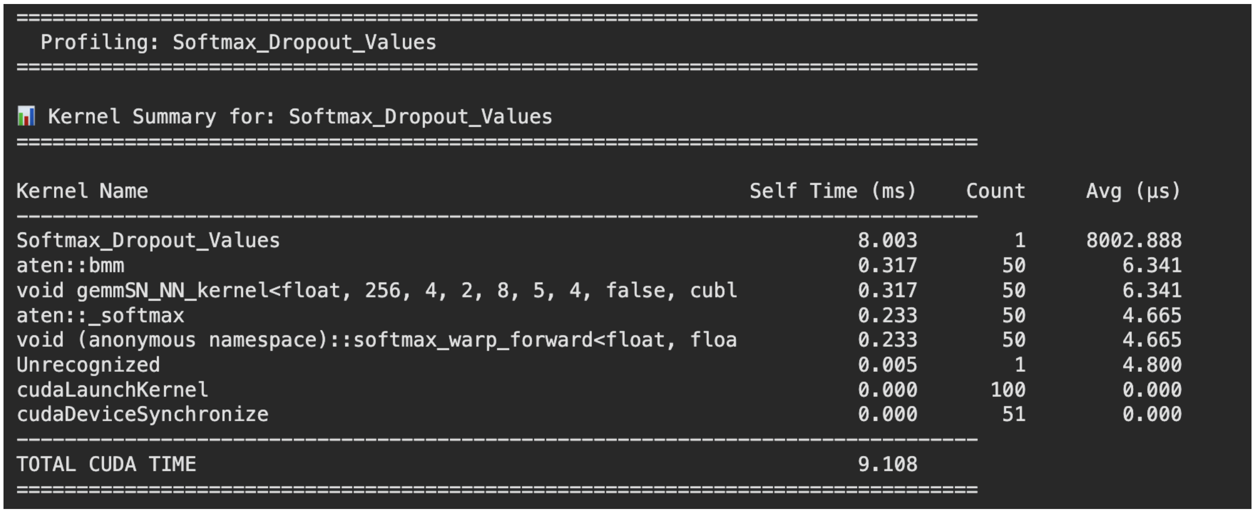




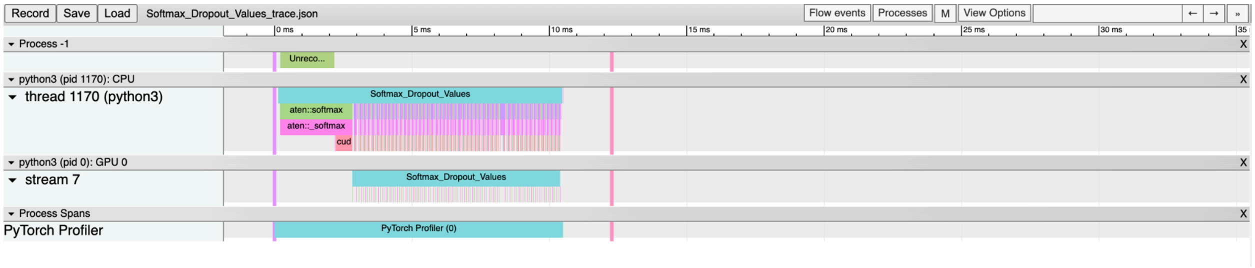
```
================================================================================
  Profiling: MLP_Projection
================================================================================

📊 Kernel Summary for: MLP_Projection
================================================================================

Kernel Name                                            Self Time (ms)   Count     Avg (µs)
--------------------------------------------------------------------------------
MLP_Projection                                                 10.559       1    10559.427
aten::addmm                                                     6.765      50      135.295
volta_sgemm_128x32_sliced1x4_tn                                 6.423      50      128.450
void cublasLt::splitKreduce_kernel<32, 16, int, float, float    0.342      50        6.845
Unrecognized                                                    0.138       1      137.629
cudaLaunchKernel                                                0.000     100        0.000
cudaDeviceSynchronize                                           0.000      51        0.000
--------------------------------------------------------------------------------
TOTAL CUDA TIME                                                24.227
================================================================================
```

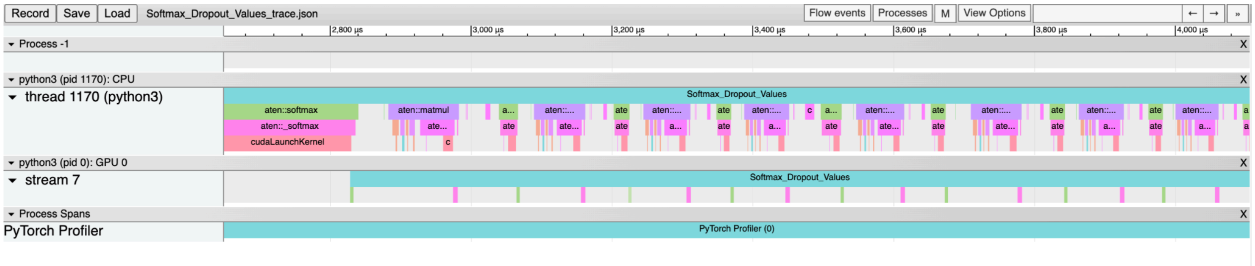






```
================================================================================
  Profiling: Softmax_Dropout_Values
================================================================================

📊 Kernel Summary for: Softmax_Dropout_Values
================================================================================

Kernel Name                                            Self Time (ms)   Count     Avg (µs)
--------------------------------------------------------------------------------
Softmax_Dropout_Values                                          8.003       1     8002.888
aten::bmm                                                       0.317      50        6.341
void gemmSN_NN_kernel<float, 256, 4, 2, 8, 5, 4, false, cubl    0.317      50        6.341
aten::_softmax                                                  0.233      50        4.665
void (anonymous namespace)::softmax_warp_forward<float, floa    0.233      50        4.665
Unrecognized                                                    0.005       1        4.800
cudaLaunchKernel                                                0.000     100        0.000
cudaDeviceSynchronize                                           0.000      51        0.000
--------------------------------------------------------------------------------
TOTAL CUDA TIME                                                 9.108
================================================================================
```





Bottlenecks in completing remaining milestones

1. Still analyzing which CUDA kernel fusions will yield the highest performance gains.
2. Actively studying and working to fully understand the FlashAttention algorithm.

Work contributed by each team member

1. Both team members participated equally in all phases of this project.