**HPC Tutorial - 13**

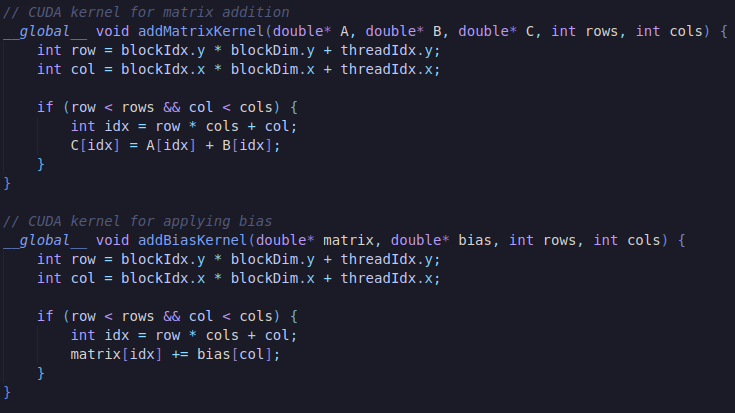
**CUDA - Parallelizing Multi-Head Self Attention Transformers**

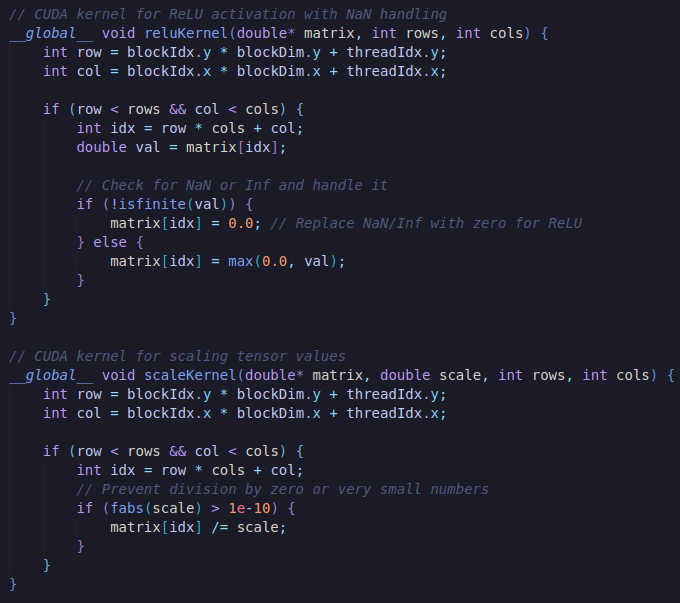
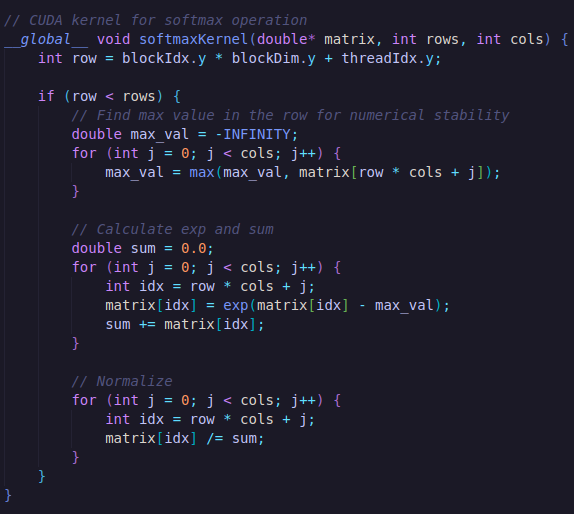
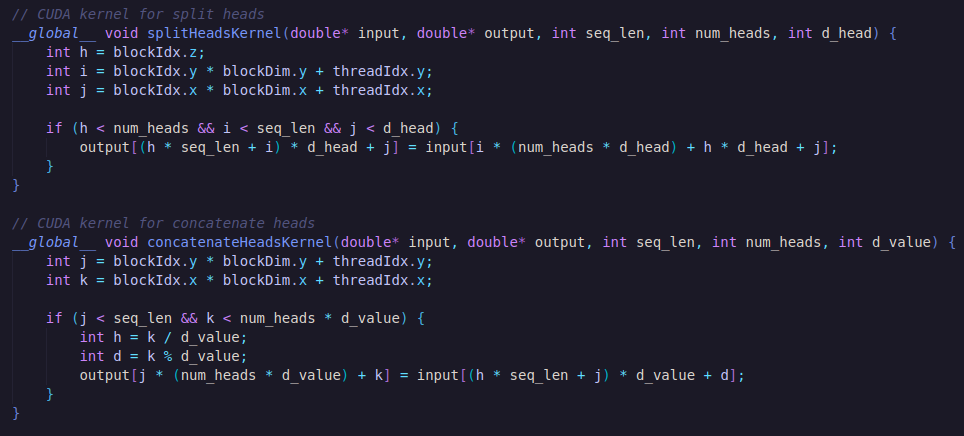
Rohan G

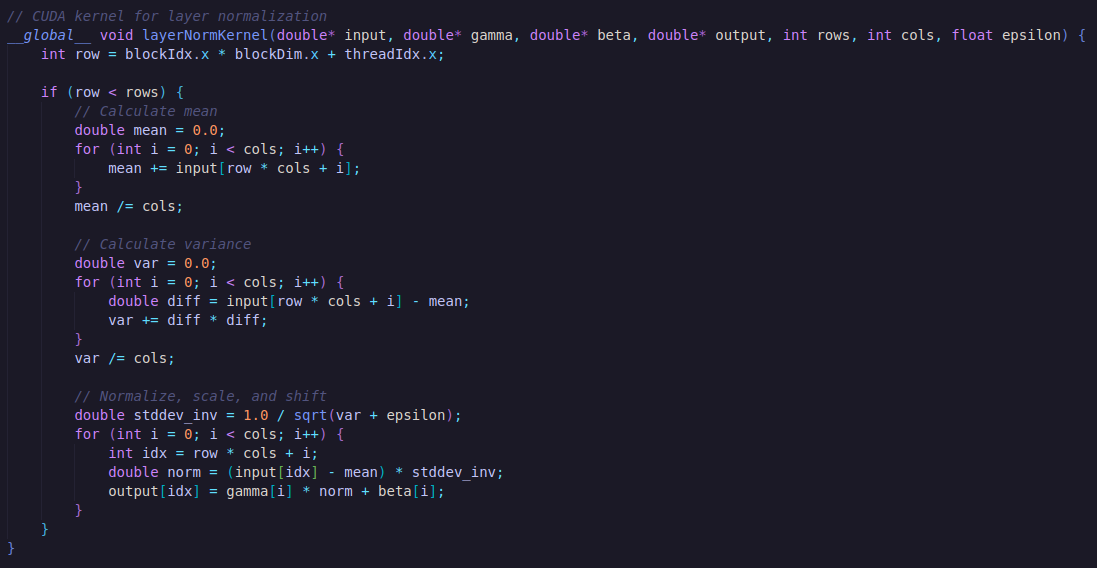
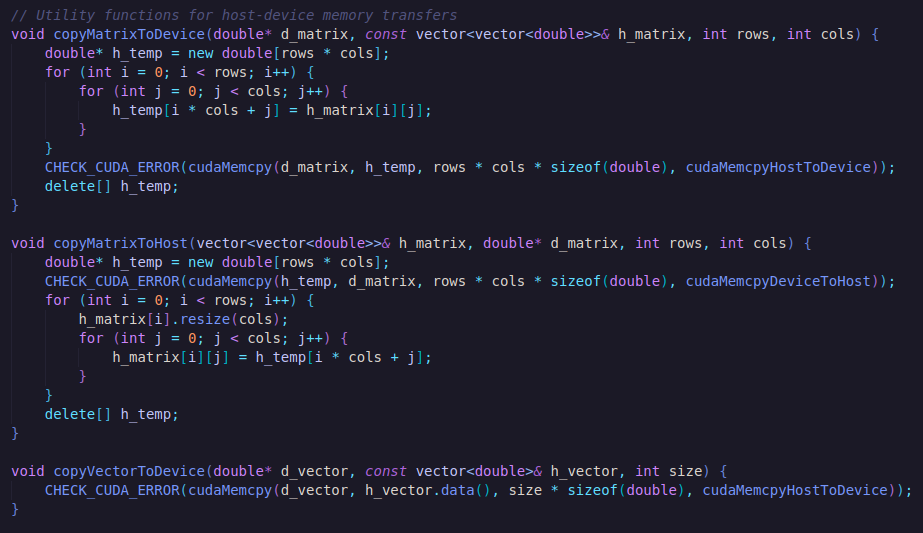
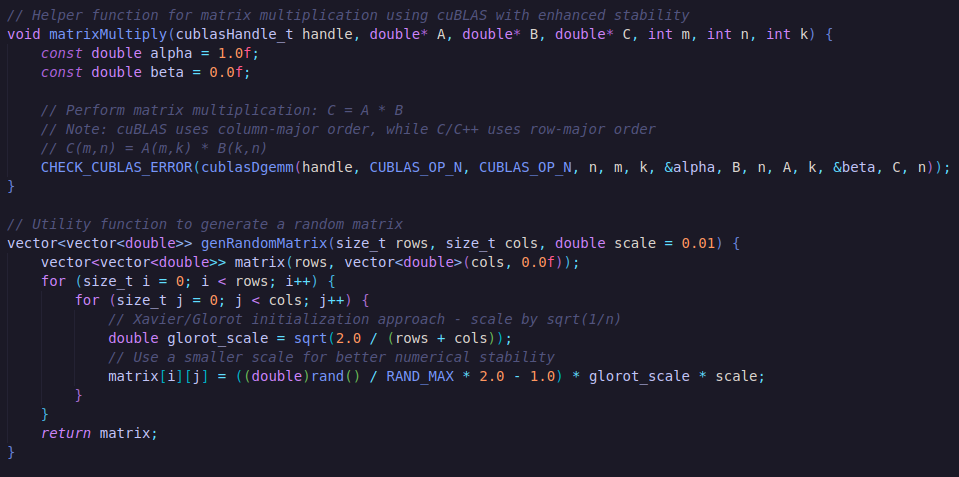
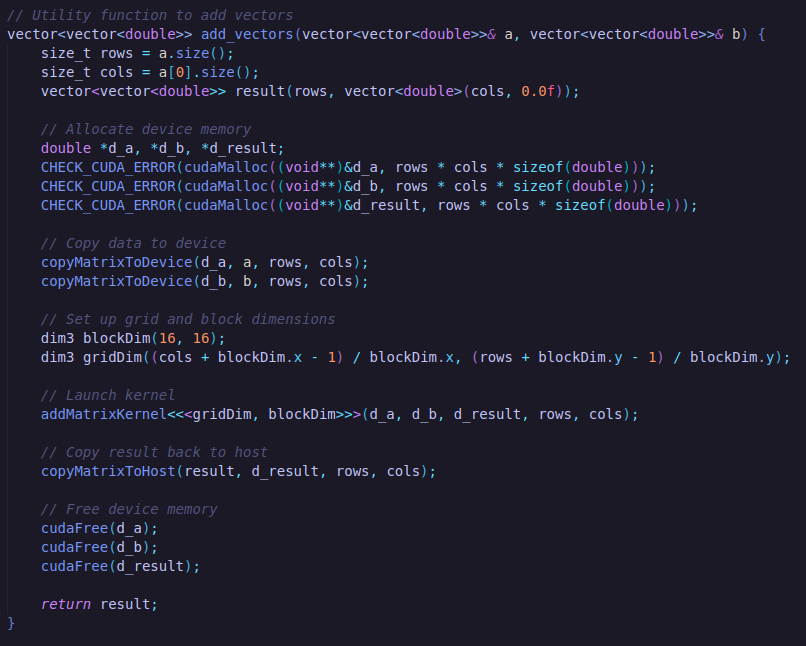
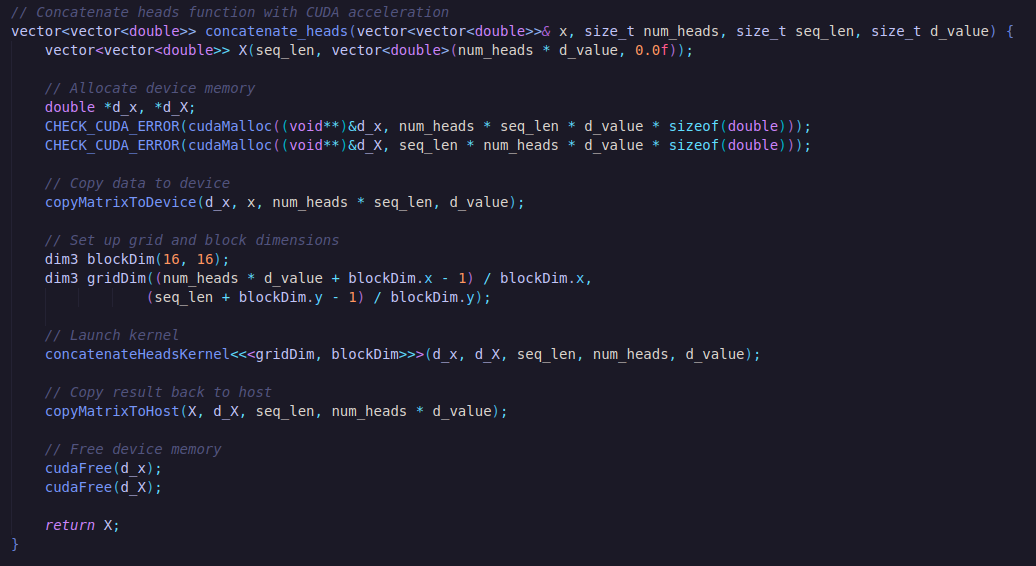
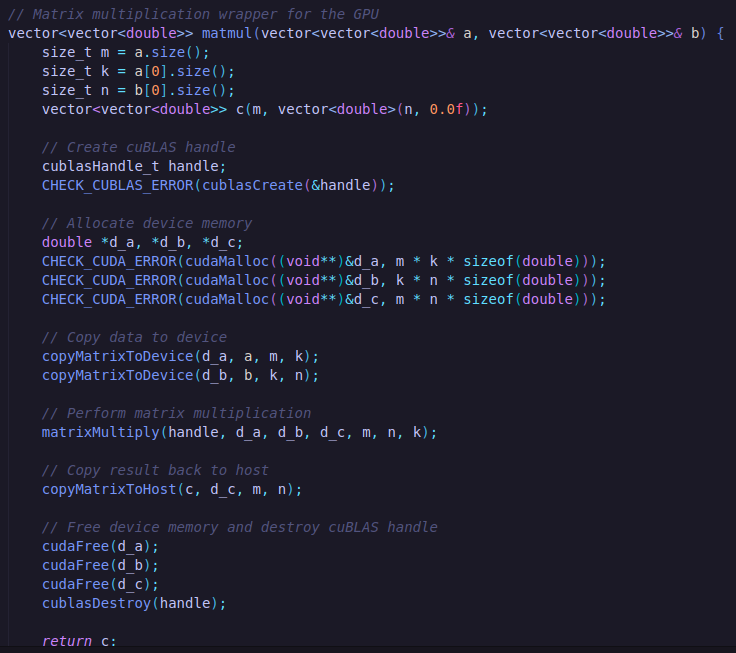
CS22B1093

**CUDA Version of project code -**

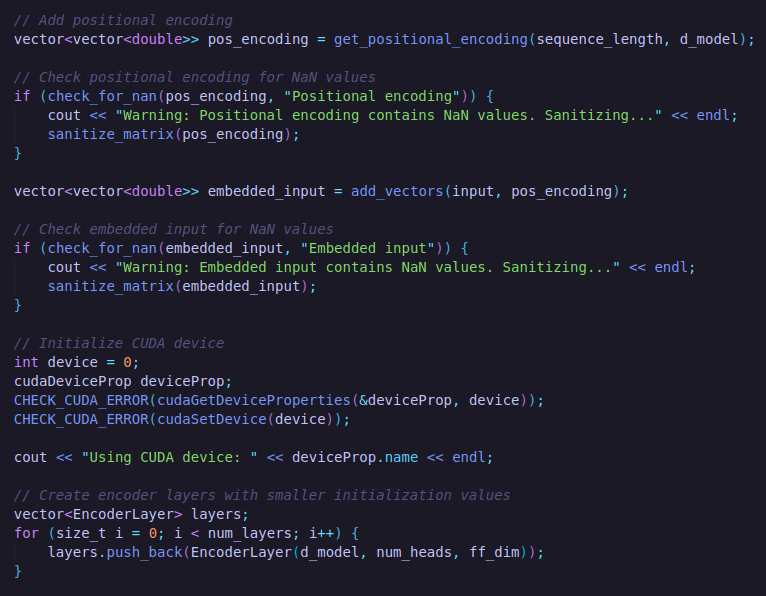
**Device Code -**

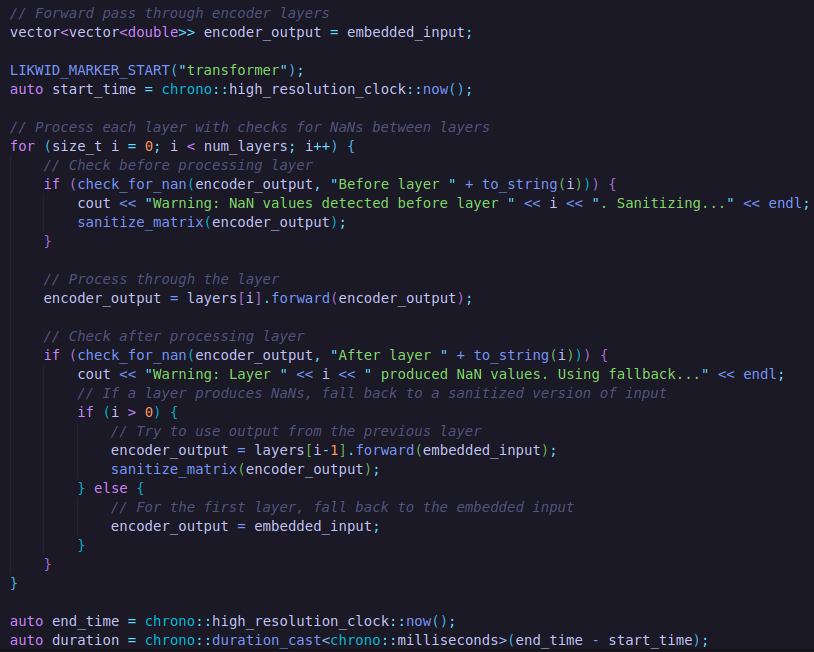


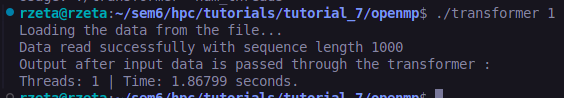
  
  
  
  
  
  


**Host Code -**

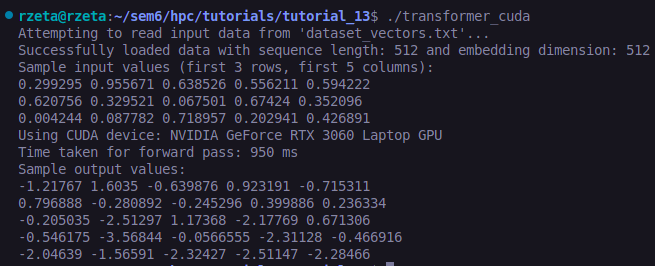
****

****

**Serial Code Output -**

****

**Parallelized CUDA Output -**

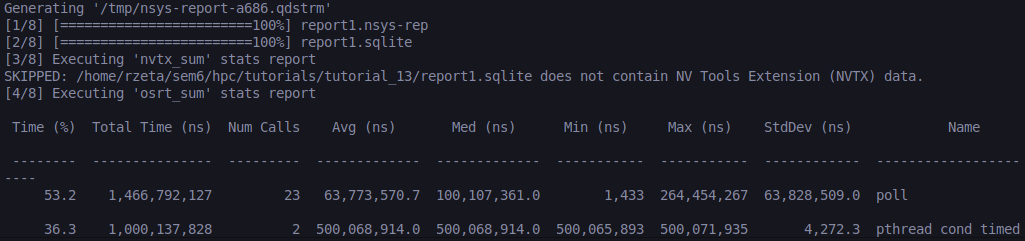
****

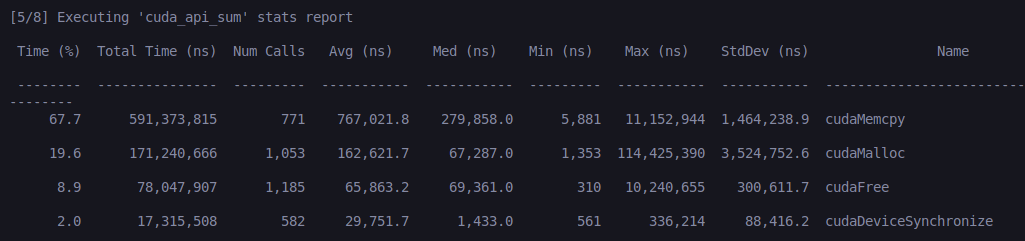
**Speedup -**

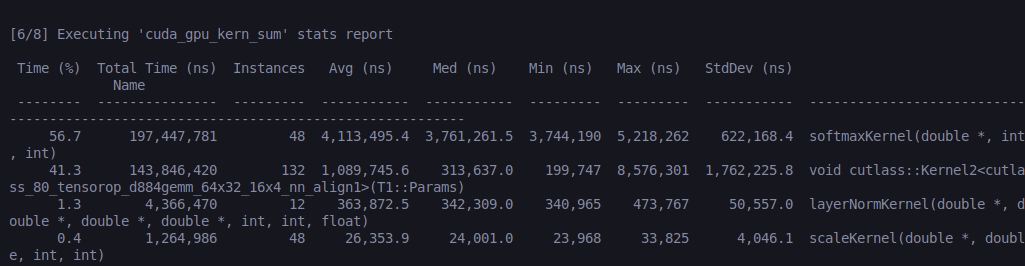
**S = T(1) / T(P) = 1.87 / 0.95 = 1.968**

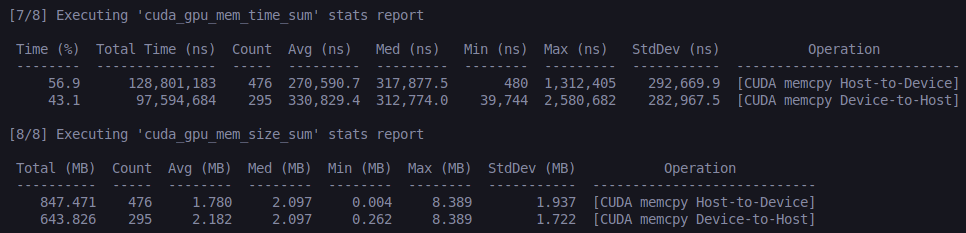
**Speedup Achieved ~ 2x**

**CUDA Profiling -**

****

****

****

****

**Observations from CUDA Nsight Profiling -**

* Batch kernel launches to reduce CPU-GPU synchronization overhead.
* Optimize data transfers with pinned memory and asynchronous streaming.
* Increase block sizes to maximize SM occupancy and warp efficiency.