CME 213, Introduction to parallel computing
Eric Darve and Vikul Gupta
Spring 2021

Stanford University

Neural Networks using CUDA
Overview of the profiler

Most of the profiling you need to do for this project can be done using Nsight Systems. The command to run is:

```
nsys profile -o nsys --trace cuda,mpi mpirun -n 4 ./main
```

Figure 1 is an example of what a profiling looks like for 1 iteration. We can analyze such a profile by looking at the amount of time spent in kernels (90.6% in stream 7) and memory (9.4% in stream 7). We can also look at which kernels are taking the most amount of time (GEMM here, at 76.8%). This tells us which parts of the code we should focus on improving first.

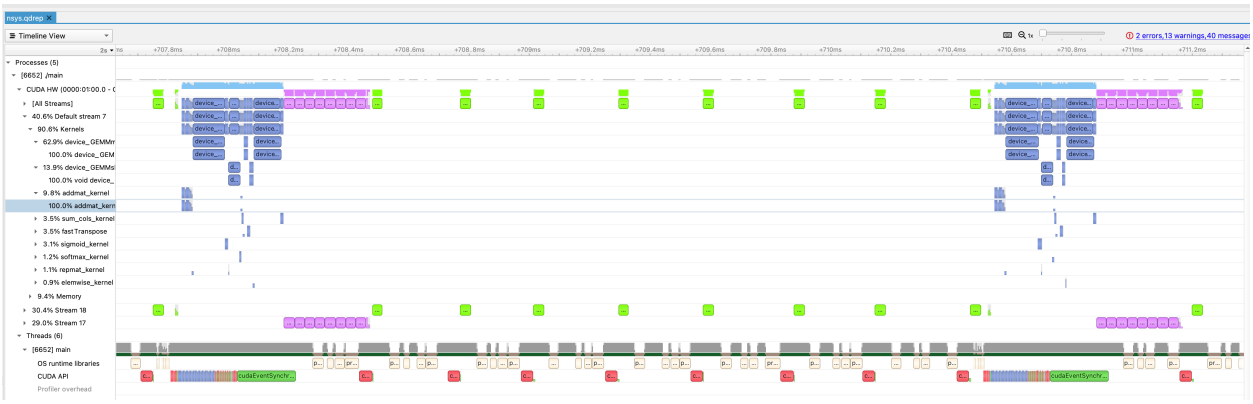

Figure 1: Nsight Systems profile for 1 iteration of the training

Nsight Compute is a more complicated tool to use. Although it is providing useful statistics on several GPU counters, interpreting these statistics and connecting them back to algorithmic changes is more challenging. Guidelines for improving your GPU code were provided in Part 1 of the instructions.