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



## 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.

nsys.qdrep ×		
		Q terrors,13 warnings,40 message
25	na +707.8ms +708.ms +708.2ms +708.4ms +708.4ms +708.8ms +708.9ms +709.2ms +709.4ms +709.4ms +709.8ms +709.8ms +710.2ms +710.4ms	+710.6ms +710.8ms +711ms +711.2ms
<ul> <li>Processes (5)</li> </ul>		
* CUDA HW (0000:01:00.0		
> [All Streams]		device
+ 40.6% Default stream 7	devicedevice	device
+ 90.6% Kernels	(dev/ce) () (dev/ce)	device
+ 62.9% device_GEMM	device	device
100.0% device_GE	divite	device
<ul> <li>13.9% device_GEMM</li> </ul>		
100.0% void device		
<ul> <li>9.8% addmat_kernel</li> </ul>		
100.0% addmat_ke		· · · · · · · · · · · · · · · · · · ·
<ul> <li>3.5% sum_cols_kern</li> </ul>		
→ 3.5% fast Transpose	.I.	
<ul> <li>3.1% sigmoid_kernel</li> </ul>		
1.2% softmax_kernel		
1.1% repmat_kernel		1
0.9% elemwise_kern		
▶ 9.4% Memory		
→ 30.4% Stream 18		-
▶ 29.0% Stream 17		
<ul> <li>Threads (6)</li> </ul>		
<ul> <li>[6652] main</li> </ul>		
OS runtime libraries		p., ., p., pf.,
CUDA API		cudaEventSynchr
Profiler overhead		

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.