PyTorch LEVEL 5

1. CUDA SEMANTICS

2. TensorFloat-32(TF32) on Ampere devices

3. Asynchronous execution

4. CUDA streams

5. Memory management

6. cuFFT plan cache

7. Best practices

8. Device-agnostic code

9. Use pinned memory buffers

10.   Use nn.parallel.DistributedDataParallel instead of multiprocessing or nn.DataParallel