

EXHIBIT

11

General Deployment

- Do you have architecture diagrams for your hardware and software systems that use NVIDIA GPUs?
- Do you use PyTorch or TensorRT?
- Which of the following, if any, top-layer applications do you use: Modulus, Maxine, cuQuantum, Merlin, Aerial, Monai, Triton, Nemo, VSS Blueprint, Riva, Metropolis, Holoscan, Clara Parabricks, Rapids, Isaac, Isaac Lab, Drive, DriveWorks, and Morpheus?
- For each NVIDIA software product identified, in the ordinary course, do you download and use it locally, use it via a cloud-based solution offered by NVIDIA, or through a third-party cloud provider?
- Do you make modifications to the source code for PyTorch, TensorRT, or other NVIDIA-provided software when its deployed on NVIDIA GPUs?
- In the ordinary course, approximately how frequently do you run computations utilizing PyTorch or TensorRT on NVIDIA GPUs (e.g., many times per day, every day, every week, or every month)?
- Are the relevant systems operated in the United States or used to support U.S.-directed operations?

Input Data Path

- In the ordinary course, do you use GPUDirect Storage (GDS) to load input data directly from storage into GPU memory, bypassing CPU main memory? Or do you use CPU main memory?
- In the ordinary course, do you use GPUDirect RDMA or any direct NIC-to-GPU memory path for receiving live input data?
- In the ordinary course, do you use unified memory (cudaMallocManaged), Unified Virtual Memory (UVM), or a coherent CPU/GPU memory architecture (e.g., DGX Spark UMA, Grace Hopper coherent memory) for the input data path?
- In the ordinary course, do you use the PyTorch function `torch.cuda.gds.GdsFile` or any GDS-enabled data loader (e.g., DALI GDS, KvikIO) to load data?

Output Data Path and Memory Transfers

- In the ordinary course, are outputs of GPU computations copied back to CPU/main memory? If so, what data is copied (e.g., transcripts, generated tokens, logits, embeddings)?
- In the ordinary course, do GPU-to-CPU (D2H) or CPU-to-GPU (H2D) data transfers occur during or in parallel with GPU computations, or only after each computation completes?
- In the ordinary course, how are your memory partitions configured for GPU computations? Are there separate memory regions for input data, output data, workspace, and intermediate results?
- In the ordinary course, are output buffers from one GPU computation reused as input buffers for a subsequent computation?
- Do you use `torch.compile`, TorchDynamo, Triton, TensorRT-LLM compilation, NVRTC, or PTX JIT to generate GPU programs at runtime, or do you use only precompiled kernels?

TensorRT Usage

- If you use TensorRT or TensorRT-LLM, how do you populate the input buffers and retrieve the outputs? Do you use CPU-side request processing and H2D/D2H copies?
- Are TensorRT engines built by your organization, provided pre-built by NVIDIA, or obtained from another source?