AMD GPU Usage

Using AMD GPUs Via Automation

Demonstrate the feasibility of automatically running existing models and applications that were developed for Nvidia instead on AMD GPUS.

Feasibility

You are to demonstrate the feasibility of automatically re-targeting models and applications, such as those using PyTorch, that have presumably only been run on NViida GPUs, instead on AMD GPUs.

Inputs

Phase 1: demonstrate by using Llama-3.3-70B FP8.

Deliverables

1. Demonstrate that indeed the specified models or programs run correctly on AMD GPUs.
2. Produce Benchmark results in a form identical to NVidia benchmarks to provide easy comparison of performance. Whenever possible use existing benchmarking technology. Develop new technology only after approval by Runara. This project is not to develop new technology but to study the use of existing technology. What may be required for this step is merely reformatting benchmark data to make comparison to Nvidia simpler.
3. A document explaining the tooling you employed with a discussion of the challenges and usage. See examples of tooling below.
4. Deliver the tooling with a documented workflow to allow Runara to automate the process of running models and PyTorch applications on AMD GPUs.
5. Validate, by reproducing, the benchmark results published by AMD.

Tooling

Use existing tools such as MLIR dialects ROCDL, AMDGPU, and GPU Dialect as appropriate. Test options such as “-convert-gpu-to-rocdl”, “-rocdl-attach-target”, and “ranslate --mlir-to-llvmir”.

Also study rocMLIR, IREE, and MLIR-AIR or any other existing technologies related to this process.

Employ other tooling as appropriate.