**Guidance to run PyTorch BERT-Large Inference on NVIDIA H100 GPUs**

Login to ACES cluster and run the commands below.

$cd $SCRATCH

$mkdir h100-benchmarks

$cd h100-benchmarks

$git clone <https://github.com/NVIDIA/DeepLearningExamples.git>

$cd DeepLearningExamples/PyTorch/LanguageModeling/BERT

$Get bert\_wrapper.sh from utils/bert\_wrapper.sh

# create a slurm job file test\_pytorch\_bert\_large.slurm and copy and paste the content below to it.

$vim test\_pytorch\_bert\_large.slurm

#!/bin/bash

##ESSARY JOB SPECIFICATIONS

#SBATCH --job-name=<your\_job>

#SBATCH --time=05:00:00 #Set the wall clock limit to 5hr

#SBATCH --nodes=1

#SBATCH --ntasks=1 #Request 1 task

#SBATCH --mem=80G

#SBATCH --output=<your\_job>\_run.%j #Send stdout/err to "Example4Out.[jobID]"

#SBATCH --gres=gpu:h100:1 #Request 1 GPU per node can be 1 or 2

#SBATCH --partition=gpu #Request 1 GPU per node can be 1 or

export SINGULARITY\_BINDPATH="$SCRATCH/h100-benchmarks/DeepLearningExamples/PyTorch/LanguageModeling/BERT/:/workspace/bert,$SCRATCH/h100-benchmarks/DeepLearningExamples/PyTorch/LanguageModeling/BERT/results/:/results, /scratch/data/pytorch-language-modelling-datasets:/shared\_space\_datasets"

export BERT\_PREP\_WORKING\_DIR="/shared\_space\_datasets/squad"

#This command is used to get stats of H100 GPU utilization

nvidia-smi --query-gpu=timestamp,name,pci.bus\_id,driver\_version,pstate,pcie.link.gen.max,pcie.link.gen.current,temperature.gpu,utilization.gpu,utilization.memory,memory.total,memory.free,memory.used --format=csv -l 1 > <your\_job>\_GPU\_stats.log &

watch -n 5 ps -u $USER > <your\_job>\_CPU\_stats.log &

echo $SCRATCH

echo $BERT\_PREP\_WORKING\_DIR

module load WebProxy

jobstats &

singularity exec --nv /scratch/data/containers/nvidia-containers/pytorch-23.06-py3.sif bash $SCRATCH/h100-benchmarks/DeepLearningExamples/PyTorch/LanguageModeling/BERT/bert\_wrapper.sh 1 4 fp16 1 prediction

jobstats

$sbatch test\_pytorch\_bert\_large.slurm