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.lin
k.gen.current,temperature.gpu,utilization.gpu,utilization.memory,memory.total,memory.fr
ee,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/BE
RT/bert_wrapper.sh 1 4 fp16 1 prediction

jobstats

$sbatch test_pytorch_bert_large.slurm
```