**Guidance to run PyTorch BERT-Large Inference on Intel Max 1100 GPUs**

Login to ACES cluster and run the commands below.

$cd $SCRATCH

$mkdir pvc-benchmarks

$cd pvc-benchmarks

$git clone <https://github.com/IntelAI/models.git>

$module purge

$ml GCCcore/11.2.0 Python/3.9.6

$python3 -m venv bert-large-pt-inference-trial

$source bert-large-pt-inference-trial/bin/activate

$pip install torch==2.1.0.post0 torchvision==0.16.0.post0 torchaudio==2.1.0.post0 intel\_extension\_for\_pytorch==2.1.20+xpu oneccl-bind-pt==2.1.200 deepspeed==0.14.0 --extra-index-url <https://pytorch-extension.intel.com/release-whl-aitools/>

$Download pretrained model as mentioned here:  
<https://github.com/IntelAI/models/tree/master/models_v2/pytorch/bert_large/inference/gpu#pre-trained-model>

$./setup.sh

$deactivate

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

$vim test\_pytorch\_bert\_large\_inference\_squad.slurm

#!/bin/bash

##NECESSARY JOB SPECIFICATIONS

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

#SBATCH --time=10:00:00 # the wallclock time for a job

#SBATCH --nodes=1 # total number of nodes

#SBATCH --ntasks=1 # total number of processes

#SBATCH --output=<your\_job\_name>\_run.%j # output of your slurm job

#SBATCH --gres=gpu:pvc:1 # for 2 gpus, set --gres=gpu:pvc:2

#SBATCH --partition=pvc # partition should be pvc for intel gpus

#SBATCH --mem=60G

ml purge

ml WebProxy

ml GCCcore/11.2.0 Python/3.9.6

source $SCRATCH/pvc-benchmarks/bert-large-pt-inference-trial/bin/activate

source /sw/hprc/sw/oneAPI/2024.1/setvars.sh

export MULTI\_TILE=True

export BERT\_WEIGHT=<path\_to\_BERT\_WEIGHT\_directory>/squad\_large\_finetuned\_checkpoint

export PLATFORM=Max

export DATASET\_DIR=/scratch/data/pytorch-language-modelling-datasets/squad/v1.1

export BATCH\_SIZE=32

export PRECISION=FP16

export OUTPUT\_DIR=$SCRATCH/pvc-benchmarks/output\_logs/bert-large-inference-squad

export CCL\_TOPO\_FABRIC\_VERTEX\_CONNECTION\_CHECK=0

cd

$SCRATCH/pvc-benchmarks/models/models\_v2/pytorch/bert\_large/inference/gpu

bash run\_model.sh

$sbatch test\_pytorch\_bert\_large.slurm