Multi-Node Training in PyTorch

NVIDIA Professional Services & Quantiphi

1. Environment Setup

- mkdir demo
- cd demo
- enroot –help
- enroot import docker://quantiphi nvidia practice/multi-node-training:v1
- enroot create --name pytorch

  quantiphinvidiapractice+multi-node-training+v1.sqsh
- enroot list
- Copy the repository to your workspace

2. Download Dataset

- wget

  https://s3.amazonaws.com/research.metamind.io/wikitext/wikitext-2-v1.zip

- unzip wikitext-2-v1.zip

3. Implementation - Multi Node Training

- Create a multi-node job file multi-node.sh to run train_transformer.py
- Multi-node.sh content:

```
#!/bin/bash

#SBATCH --nodes=2
#SBATCH --job-name=multinode-example
#SBATCH --partition=dgxnp
#SBATCH --ntasks=2
#SBATCH --gres=gpu:A100-SXM4:4

nodes=( $( scontrol show hostnames $SLURM_JOB_NODELIST ) )
nodes_array=($nodes)
head_node=${nodes_array[0]}
echo $head_node
head_node_ip=$(srun --nodes=1 --ntasks=1 -w "$head_node" hostname --ip-address)
echo $head_node_ip
echo Node IP: $head_node_ip
export LOGLEVEL=INFO
head_node_array=($head_node_ip)
head_n=${head_node_array[1]}
NUM_NODES=2
NUM_PROC_PER_NODE=2
export NUM_NODES
export NUM_PROC_PER_NODE
export http_proxy=http://proxy-10g.10g.siddhi.param:9090

export https_proxy=http://proxy-10g.10g.siddhi.param:9090
```

```
srun --no-container-entrypoint  --container-image $(pwd)/quantiphinvidiapractice+multi-node-training+v1.sqsh
--container-mounts $(pwd)/gpt/:/workspace/ \
torchrun \
--nnodes 2 \
--nproc_per_node 2 \
--rdzv_id $RANDOM \
--rdzv_backend c10d \
--rdzv_endpoint $head_n:29500 \
/workspace/gpt/train_transformer.py
```

In the above script we can make the following changes :

1. Change the number of nodes by changing
    a. #SBATCH –nodes - *line 4*
    b. #SBATCH –ntasks - line 5
    c. –nnodes - *line 24*
2. Change the number of GPUs by changing
    a. #SBATCH –gpus-per-task - line 6
    b. –nproc_per_node - line 25

- For train_transformerf16.py change the python script name in srun command:

```
srun --no-container-entrypoint  --container-image $(pwd)/quantiphinvidiapractice+multi-node-training+v1.sqsh
--container-mounts $(pwd)/gpt/:/workspace/ \
torchrun \
--nnodes 2 \
--nproc_per_node 2 \
--rdzv_id $RANDOM \
--rdzv_backend c10d \
--rdzv_endpoint $head_n:29500 \
/workspace/gpt/train_transformerf16.py
```

- Run the script : **sbatch multi-node.sh**
    - Submitted batch job 158173

- Check the job with squeue
    - squeue

158173 benchp multinod sheetals R 0:04 2 scn75-10g,scn76-10g

- Verify the output file - slurm-158173.out

train_transformer.py:

```
scn46-10g
127.0.1.1 172.50.0.46
Node IP: 127.0.1.1 172.50.0.46
master_addr is only used for static rdzv_backend and when rdzv_endpoint is not specified.
WARNING:torch.distributed.run:
*****************************************
Setting OMP_NUM_THREADS environment variable for each process to be 1 in default, to avoid your system being overloaded, please further tune the va>
*****************************************
INFO:torch.distributed.launcher.api:Starting elastic_operator with launch configs:
  entrypoint       : /workspace/gpt/train_transformer.py
  min_nodes        : 2
  max_nodes        : 2
  nproc_per_node   : 2
  run_id           : 17718
  rdzv_backend     : c10d
  rdzv_endpoint    : 172.50.0.46:29500
  rdzv_configs     : {'timeout': 900}
  max_restarts     : 0
  monitor_interval : 5
  log_dir          : None
  metrics_cfg      : {}

INFO:torch.distributed.elastic.agent.server.local_elastic_agent:log directory set to: /tmp/torchelastic_axleuift/17718_diqo22n2
INFO:torch.distributed.elastic.agent.server.api:[default] starting workers for entrypoint: python
INFO:torch.distributed.elastic.agent.server.api:[default] Rendezvous'ing worker group
master_addr is only used for static rdzv_backend and when rdzv_endpoint is not specified.
WARNING:torch.distributed.run:
*****************************************
Setting OMP_NUM_THREADS environment variable for each process to be 1 in default, to avoid your system being overloaded, please further tune the va>
*****************************************
INFO:torch.distributed.launcher.api:Starting elastic_operator with launch configs:
```

```
  GNU nano 4.8                                                    slurm-158339.out
INFO:torch.distributed.elastic.multiprocessing:Setting worker1 reply file to: /tmp/torchelastic_4ldznr0m/17718_pz5mkxv0/attempt_0/1/error.json
Rank: 0
Nodes: 2
Processes per node: 2
GPUs per node: 4
GPUs per process: 2
[0, 1]

Rank: 1
Nodes: 2
Processes per node: 2
GPUs per node: 4
GPUs per process: 2
[2, 3]

Rank: 2
Nodes: 2
Processes per node: 2
GPUs per node: 4
GPUs per process: 2
[0, 1]

Rank: 3
Nodes: 2
Processes per node: 2
GPUs per node: 4
GPUs per process: 2
[2, 3]

Putting block 0 in device 0
Putting block 0 in device 2
```

```
  GNU nano 4.8                                                    slurm-158339.out
)
[RANK 0]: | epoch   1 |    10/   50 batches | lr 0.10 | ms/batch 129.25 | loss 17.86 | ppl 56978492.67
[RANK 1]: | epoch   1 |    10/   50 batches | lr 0.10 | ms/batch 135.15 | loss 18.22 | ppl 82013824.01
[RANK 2]: | epoch   1 |    10/   50 batches | lr 0.10 | ms/batch 138.27 | loss 17.79 | ppl 53078305.47
[RANK 3]: | epoch   1 |    10/   50 batches | lr 0.10 | ms/batch 140.56 | loss 17.82 | ppl 55079154.61
[RANK 0]: | epoch   1 |    20/   50 batches | lr 0.10 | ms/batch 60.19 | loss 15.33 | ppl 4532149.06
[RANK 1]: | epoch   1 |    20/   50 batches | lr 0.10 | ms/batch 60.25 | loss 15.48 | ppl 5265892.72
[RANK 2]: | epoch   1 |    20/   50 batches | lr 0.10 | ms/batch 60.46 | loss 15.22 | ppl 4060364.61
[RANK 3]: | epoch   1 |    20/   50 batches | lr 0.10 | ms/batch 60.21 | loss 15.21 | ppl 4044844.88
[RANK 0]: | epoch   1 |    30/   50 batches | lr 0.10 | ms/batch 60.24 | loss 14.60 | ppl 2201553.21
[RANK 1]: | epoch   1 |    30/   50 batches | lr 0.10 | ms/batch 60.26 | loss 14.77 | ppl 2592376.49
[RANK 2]: | epoch   1 |    30/   50 batches | lr 0.10 | ms/batch 60.44 | loss 14.58 | ppl 2137491.15
[RANK 3]: | epoch   1 |    30/   50 batches | lr 0.10 | ms/batch 60.21 | loss 14.54 | ppl 2056326.89
[RANK 0]: | epoch   1 |    40/   50 batches | lr 0.10 | ms/batch 60.26 | loss 14.14 | ppl 1378609.32
[RANK 1]: | epoch   1 |    40/   50 batches | lr 0.10 | ms/batch 60.22 | loss 14.30 | ppl 1625521.03
[RANK 2]: | epoch   1 |    40/   50 batches | lr 0.10 | ms/batch 60.46 | loss 14.06 | ppl 1272070.38
[RANK 3]: | epoch   1 |    40/   50 batches | lr 0.10 | ms/batch 60.33 | loss 14.08 | ppl 1307773.90
[RANK 0]: -----------------------------------------------------------------------------------------
[RANK 0]: | end of epoch   1 | time:  3.75s | valid loss   nan | valid ppl      nan
[RANK 0]: -----------------------------------------------------------------------------------------
[RANK 1]: -----------------------------------------------------------------------------------------
[RANK 1]: | end of epoch   1 | time:  3.81s | valid loss   nan | valid ppl      nan
[RANK 1]: -----------------------------------------------------------------------------------------
[RANK 2]: -----------------------------------------------------------------------------------------
[RANK 2]: | end of epoch   1 | time:  3.85s | valid loss   nan | valid ppl      nan
[RANK 2]: -----------------------------------------------------------------------------------------
[RANK 3]: -----------------------------------------------------------------------------------------
[RANK 3]: | end of epoch   1 | time:  3.87s | valid loss   nan | valid ppl      nan
[RANK 3]: -----------------------------------------------------------------------------------------
[RANK 0]: | epoch   2 |    10/   50 batches | lr 0.10 | ms/batch 66.49 | loss 14.96 | ppl 3148387.91
[RANK 1]: | epoch   2 |    10/   50 batches | lr 0.10 | ms/batch 66.52 | loss 15.22 | ppl 4089604.70
```

train_transformerf16.py:

```
  GNU nano 4.8                                                    slurm-158340.out
scn46-10g
127.0.1.1 172.50.0.46
Node IP: 127.0.1.1 172.50.0.46
master_addr is only used for static rdzv_backend and when rdzv_endpoint is not specified.
WARNING:torch.distributed.run:
*****************************************
Setting OMP_NUM_THREADS environment variable for each process to be 1 in default, to avoid your system being overloaded, please further tune the va>
*****************************************
INFO:torch.distributed.launcher.api:Starting elastic_operator with launch configs:
  entrypoint       : /workspace/gpt/train_transformerf16.py
  min_nodes        : 2
  max_nodes        : 2
  nproc_per_node   : 2
  run_id           : 31307
  rdzv_backend     : c10d
  rdzv_endpoint    : 172.50.0.46:29500
  rdzv_configs     : {'timeout': 900}
  max_restarts     : 0
  monitor_interval : 5
  log_dir          : None
  metrics_cfg      : {}

INFO:torch.distributed.elastic.agent.server.local_elastic_agent:log directory set to: /tmp/torchelastic_pa1i_3ey/31307_ue8c1hzz
INFO:torch.distributed.elastic.agent.server.api:[default] starting workers for entrypoint: python
INFO:torch.distributed.elastic.agent.server.api:[default] Rendezvous'ing worker group
master_addr is only used for static rdzv_backend and when rdzv_endpoint is not specified.
WARNING:torch.distributed.run:
*****************************************
Setting OMP_NUM_THREADS environment variable for each process to be 1 in default, to avoid your system being overloaded, please further tune the va>
*****************************************
INFO:torch.distributed.launcher.api:Starting elastic_operator with launch configs:
```

```
 GNU nano 4.8                                                    slurm-158340.out
[2, 3]

Rank: 2
Nodes: 2
Processes per node: 2
GPUs per node: 4
GPUs per process: 2
[0, 1]

Putting block 0 in device 2
Putting block 0 in device 2
Putting block 0 in device 0
Putting block 0 in device 0
Putting block 1 in device 2
Putting block 1 in device 0
Putting block 1 in device 2
Putting block 1 in device 0
Putting block 2 in device 2
Putting block 2 in device 2
Putting block 2 in device 0
Putting block 2 in device 0
Putting block 3 in device 2
Putting block 3 in device 2
Putting block 3 in device 0
Putting block 3 in device 0
Putting block 4 in device 3
Putting block 4 in device 1
Putting block 4 in device 3
Putting block 4 in device 1
Putting block 5 in device 3
Putting block 5 in device 1
```

```
 GNU nano 4.8                                                    slurm-158340.out
    (decoder): Decoder(
      (decoder): Linear(in_features=4096, out_features=28782, bias=True)
    )
  )
)
[RANK 2]: | epoch   1 |    10/   50 batches | lr 0.10 | ms/batch 125.91 | loss 15.75 | ppl 6892684.59
[RANK 3]: | epoch   1 |    10/   50 batches | lr 0.10 | ms/batch 127.44 | loss 16.02 | ppl 9092353.60
[RANK 0]: | epoch   1 |    10/   50 batches | lr 0.10 | ms/batch 129.32 | loss 15.76 | ppl 6958134.22
[RANK 1]: | epoch   1 |    10/   50 batches | lr 0.10 | ms/batch 135.52 | loss 15.93 | ppl 8258140.00
[RANK 2]: | epoch   1 |    20/   50 batches | lr 0.10 | ms/batch 48.94 | loss 13.21 | ppl 548135.58
[RANK 3]: | epoch   1 |    20/   50 batches | lr 0.10 | ms/batch 48.88 | loss 12.98 | ppl 431739.40
[RANK 0]: | epoch   1 |    20/   50 batches | lr 0.10 | ms/batch 48.88 | loss 12.89 | ppl 397521.13
[RANK 1]: | epoch   1 |    20/   50 batches | lr 0.10 | ms/batch 49.15 | loss 13.10 | ppl 491336.28
[RANK 2]: | epoch   1 |    30/   50 batches | lr 0.10 | ms/batch 48.99 | loss 12.85 | ppl 381412.23
[RANK 3]: | epoch   1 |    30/   50 batches | lr 0.10 | ms/batch 48.89 | loss 12.91 | ppl 406095.89
[RANK 0]: | epoch   1 |    30/   50 batches | lr 0.10 | ms/batch 48.86 | loss 12.73 | ppl 337121.84
[RANK 1]: | epoch   1 |    30/   50 batches | lr 0.10 | ms/batch 48.88 | loss 12.65 | ppl 312638.62
[RANK 2]: | epoch   1 |    40/   50 batches | lr 0.10 | ms/batch 49.01 | loss 12.73 | ppl 337699.82
[RANK 3]: | epoch   1 |    40/   50 batches | lr 0.10 | ms/batch 48.89 | loss 12.64 | ppl 308418.23
[RANK 0]: | epoch   1 |    40/   50 batches | lr 0.10 | ms/batch 48.83 | loss 12.54 | ppl 279134.10
[RANK 1]: | epoch   1 |    40/   50 batches | lr 0.10 | ms/batch 48.82 | loss 12.56 | ppl 283906.82
[RANK 2]: -----------------------------------------------------------------------------------------
[RANK 2]: | end of epoch   1 | time:  3.30s | valid loss   nan | valid ppl      nan
[RANK 2]: -----------------------------------------------------------------------------------------
[RANK 3]: -----------------------------------------------------------------------------------------
[RANK 3]: | end of epoch   1 | time:  3.31s | valid loss   nan | valid ppl      nan
[RANK 3]: -----------------------------------------------------------------------------------------
[RANK 0]: -----------------------------------------------------------------------------------------
[RANK 0]: | end of epoch   1 | time:  3.34s | valid loss   nan | valid ppl      nan
[RANK 0]: -----------------------------------------------------------------------------------------
[RANK 1]: -----------------------------------------------------------------------------------------
```