V25.3

Overview

The Unified training docker aim at making customers in China more easily to use ROCm to train. The v25.3 release contains 2 docker images, share the similar environments, but with different examples.

v25.3-megatronlm

pull tags: docker pull packages.xilinx.com/instinct-china/dev-benchmark-300x:rocm6.3.0_ubuntu22.04_py3.10_megatronlm_v253

doc: https://rocm.docs.amd.com/en/latest/how-to/rocm-for-ai/training/benchmark-docker/megatron-lm.html

v25.3-pytorchtraining

pull tags: docker pull packages.xilinx.com/instinct-china/dev-benchmark-300x:rocm6.3.0_ubuntu22.04_py3.10_pytorchtraining_v253

doc: https://rocm.docs.amd.com/en/latest/how-to/rocm-for-ai/training/benchmark-docker/pytorch-training.html

Basic components

Software component	Version
ROCm	6.3.0
PyTorch	2.7.0a0+git637433
Python	3.10
Transformer Engine	1.11
Flash Attention	3.0.0
hipBLASLt	git258a2162
Triton	3.1

Example model and performance

Docker	code	Model	#Nodes	Seq_Len	MBS	GBS	Data Type	TP	PP	CP	EP	memory %	TFLOPs/s/GPU	MFU	Best 0115/TFLOPS
v25.3	example	LLAMA 3.1-8B	1	8192	2	128	BF16	1	1	1	1	72 %	172	83 %	163

v25.3	example	LLAMA3.1-8B	1	8192	2	128	FP8	1	1	1	1	70%	249	60%	243
v25.3	example	LLAMA3.1-70B	1	2048	4	256	BF16	8	1	1	1	99%	130	63%	123
v25.3	example	LLAMA3.1-70B	1	2048	4	256	FP8	8	1	1	1	99%	205	50%	200
v25.3	example	Qwen2.5-7B	1	2048	10	320	BF16	1	1	1	1	90%	158	77%	146
v25.3	example	Qwen2.5-7B	1	2048	10	320	FP8	1	1	1	1	90%	232	56%	205
v25.3	example	Qwen2.5-72B	1	2048	2	128	BF16	8	1	1	1	99%	124	60%	103
v25.3	example	Qwen2.5-72B	1	2048	2	128	FP8	8	1	1	1	99%	182	44%	173
v25.3	example	Mixtral-7B	1	4096	3	264	BF16	4	1	1	1	99%	111	54%	111
v25.3	example	Mixtral-7B	1	4096	3	264	FP8	4	1	1	1	99%	142	34%	140
v25.3	example	Deepseekv2-16B	1	2048	8	256	BF16	1	1	1	8	90%	66	32%	67
v25.3	example	Deepseekv2-16B	1	2048	8	256	FP8	1	1	1	8	90%	68	17%	68

v25.3	example	Flux	1	512 (imagesize)	--	1	BF16	--	--	--	--	95%	47	23%	40
v25.3	example	Flux	1	512 (imagesize)	--	10	BF16	--	--	--	--	99%	84	41%	80

Key features:

- Transformer Engine (TE)
- APEX
- GEMM tuning
- Torch.compile
- 3D parallelism: TP + SP + CP
- Distributed optimizer
- Flash Attention (FA) 3
- Fused kernels
- Pre-training
- Supported BF16/FP8
- Support Model: LLAMA3.1-8B/70B, Mixtral-7B, Qwen2.5-7B/72B, DeepSeekV2 Lite，Flux

Examples

inside docker, we provides examples with LLaMA3-8B, QWen2.5-7B, Mixtral 8x7B and Deepseekv2 using Megatron-LM.

```

└── workspace
    └── Megatron-LM

cd /workspace/Megatron-LM
# llama3 8B
bash examples/llama/train_llama3.sh TP=1 CP=1 PP=1 MBS=7 BS=280 TE_FP8=0 MODEL_SIZE=8 SEQ_LENGTH=2048 TOTAL_ITER
#  qwen2.5 7b
bash examples/qwen/train_qwen2.sh TP=1 CP=1 PP=1 MBS=10 BS=320 TE_FP8=0 MODEL_SIZE=7 SEQ_LENGTH=2048 TOTAL_ITERS
# mixtral
bash examples/mixtral/train_mixtral_8x7b_distributed_bf16.sh
# deepseekv2
bash  examples/deepseek_v2/train_deepseekv2.sh
```

Pytorch Training

inside docker, we provides examples with Flux and LLama-3.1-70B using pytorch.

```
└── workspace
    ├── MAD
    ├── FluxBenchmark
    ├── torchtitan
    └── torchtune
# Flux
cd /workspace/FluxBenchmark
python3 launch.py
# llama3-70b (torchtitan)
git clone https://github.com/ROCM/MAD.git
cd MAD/scripts/pytorch_train
./pytorch_benchmark_report.sh -t pretrain -p BF16 -m Llama-3.1-70B -s 8192
```