- FP16 `python ./main.py --mode=inference_benchmark --use_tf_amp --warmup_steps 20 --train_iter 100 --iter_unit batch --batch_size <batch size> --data_dir=<path to imagenet> --log_dir=<path to results directory>`

Each of these scripts, by default runs 20 warm-up iterations and measures the next 80 iterations.

To control warm-up and benchmark length, use `--warmup_steps`, `--num_iter` and `--iter_unit` flags.

Results

The following sections provide details on how we achieved our results in training accuracy, performance and inference performance.

Training accuracy results

Our results were obtained by running the `./scripts/RN50_{FP16, FP32}_{1, 4, 8}GPU.sh` script in the tensorflow-19.02-py3 Docker container on NVIDIA DGX-1 with 8 V100 16G GPUs.

number of GPUs	mixed precision top1	mixed precision training time	FP32 top1	FP32 training time
1	76.18	41.3h	76.38	89.4h
4	76.30	10.5h	76.30	22.4h
8	76.18	5.6h	76.26	11.5h

Training performance results

Our results were obtained by running the `./scripts/benchmarking/DGX1V_trainbench_fp16.sh` and `./scripts/benchmarking/DGX1V_trainbench_fp32.sh` scripts in the tensorflow-19.02-py3 Docker container on NVIDIA DGX-1 with 8 V100 16G GPUs.

number of GPUs	mixed precision img/s	FP32 img/s	mixed precision speedup	mixed precision weak scaling	FP32 weak scaling
1	818.3	362.5	2.25	1.00	1.00
4	3276.6	1419.4	2.30	4.00	3.92
8	6508.4	2832.2	2.30	7.95	7.81

Our results were obtained by running the `./scripts/benchmarking/DGX1V_inferbench_fp16.sh` and `./scripts/benchmarking/DGX1V_inferbench_fp32.sh` scripts in the tensorflow-19.02-py3 Docker container on NVIDIA DGX-1 with 8 V100 16G GPUs.

Inference performance results

batch size	mixed precision img/s	FP32 img/s
1	177.2	170.8
2	325.7	308.4
4	587.0	499.4
8	1002.9	688.3
16	1408.5	854.9