SambaNova SYSTEMS

SambaTune

Measurement, Profiling, and Visualization

April 2024

Agenda

1. Goals
2. Performance Profiling Methodology (Documentation Review)
3. SambaTune Usage
4. Demo

Performance Profiling Methodology

Methodology

- The SambaNova Systems model performance profiling methodology is best described In the SambTune Documentation. So, instead of duplicating the documentation text into the slides, we will refer to the documentation directly.
- The documentation is located at : https://docs.sambanova.ai/sambatune/latest/workflow.html

SambaTune Usage

SambaTune Usage

Note: this does not need to be run as root, but was in this example
Note: SambaTune does not run via slurm so you have to allocate a node to do your work manually

```
# salloc --gres=rdu:8 -n 1  --ntasks-per-node  1 --nodes 1  --cpus-per-task=128 --nodelist $(hostname)
/bin/bash
```

Need to set a DUMP_ROOT and ARTIFACT_ROOT environment variables

```
#(venv) export DUMP_ROOT=`pwd`
```

Alternatively, you can pass artifact-root as a command line argument

```
# sambatune <config.yaml> --artifact-root=$(pwd)/artifact_root
```

Command Execution

```
#(venv) /opt/sambaflow/bin/sambatune
No plugins available at this time.
usage: sambatune [-h] [--artifact-root ARTIFACT_ROOT] [--compile-only | -m MODES [MODES ...]] [--version]
config
sambatune: error: the following arguments are required: config □ This is referring to requiring  a yaml
file
```

SambaTune: linear_net Configuration Files Example: Single Config

```
cat linear_net.yaml
# This would require the sambaflow-apps-micros package to be installed.
app: /opt/sambaflow/apps/micros/linear_net.py
model-args: >
  -b 1024
  -mb 64
  --in-features 5120
  --out-features 512
compile-args: >
  --plot
run-args: --num-iterations 100
env:
  SF_RNT_FSM_POLL_BUSY_WAIT: 1
  SF_RNT_DMA_POLL_BUSY_WAIT: 1
  CONVFUNC_DEBUG_RUN": 0
```

SambaNova SYSTEMS

SambaTune: linear_net Configuration Files Example : Compare Config

```
cat linear_net_compare.yaml
# This would require the sambaflow-apps-micros package to be installed.
app: /opt/sambaflow/apps/micros/linear_net.py
model-args: >
  -b 1024
  -mb 128
  --in-features 5120
  --out-features 512
compile-args: >
  --plot
run-args: --num-iterations 100
env:
  SF_RNT_FSM_POLL_BUSY_WAIT: 1
  SF_RNT_DMA_POLL_BUSY_WAIT: 1
  CONVFUNC_DEBUG_RUN": 0
---
app: /opt/sambaflow/apps/micros/linear_net.py
model-args: >
  -b 512
  -mb 64
  --in-features 2560
  --out-features 256
compile-args: >
  --plot
run-args: --num-iterations 100
env:
  SF_RNT_FSM_POLL_BUSY_WAIT: 1
  SF_RNT_DMA_POLL_BUSY_WAIT: 1
  CONVFUNC_DEBUG_RUN": 0
```

SambaNova SYSTEMS

SambaTune: Example of an LLM with Compiler Configs and HD Files

Full files is located at /opt/sambaflow/sambatune/configs/gpt13b_sn30.yaml

```
# This would require the sambaflow-apps-datascale-language-gpt13b and sambaflow-apps-datascale-language-blocksparse packages
to be installed.
app: /opt/sambaflow/apps/nlp/transformers_on_rdu/transformers_hook.py

compile-args: compile --module_name gpt2_pretrain --task_name clm --max_seq_length 2048
  -b 16 --output_dir=${HF_OUTPUT} --overwrite_output_dir
  --per_device_train_batch_size 16  --tokenizer_name gpt2
  --model_name gpt2 --mac-v2 --non_split_head --skip_broadcast_patch
  --config_name ${CONFIG_NAME} --data-parallel -ws 2
  --compiler-configs-file ${COMPILER_CONFIGS_FILE}  --no_index_select_patch --weight_decay 0.1
  --model-parallel --n-chips 2 --mac-human-decision ${HD_FILE}
  --generate_par_factors --max_grad_norm_clip 1.0 --optimize-concat-split --enable-hypersection
  --use_3d_attention_mask --article_attention --disable-strict-conversion
…
env:
  ENABLE_TEMPORAL_ACCUM_STOC: 1
  ENABLE_EMBEDDINGBAG_REDUCE_STOC: 1
  ENABLE_ADDBIAS_GRAD_ACCUM_STOC: 1
  ENABLE_SCATTERND_STOC: 1
  HF_OUTPUT: HF_OUTPUT_DIR
  COMPILER_CONFIGS_FILE:
/opt/sambaflow/apps/nlp/transformers_on_rdu/gpt13b/sn30/cc/compiler_configs_gpt3_13b_tgm_mp2_bs16_training_3D_attn_cls_p2p_rc.
json
  CONFIG_NAME: /opt/sambaflow/apps/nlp/transformers_on_rdu/gpt13b/common/configs/gpt3_13b_config_40enc_50260.json
  HD_FILE:
/opt/sambaflow/apps/nlp/transformers_on_rdu/gpt13b/sn30/hd/gpt3_13b_tgm_mp2_hs_bs16_40enc_3d_attn_load_from_dram.json
```

SambaNova SYSTEMS

SambaTune Run Example

Available modes
run – Full training run
benchmark -Section times
instrument - Fine grained performance counters on hardware per section and stage
parameter-sweep – use when using a parameter sweep config

Note: benchmark requires the model to have a measure performance section
Note: When running a sweep, you typically will only pass in the parameter-sweep mode

Example static config with instrumentation and all performance counters command
sambatune ./sambatune_configs/linear_net.yaml -m benchmark instrument run

Example sweep command :
sambatune ./sambatune_configs/linear-samba-sweep.yaml -m 'parameter-sweep'

SambaTune Reports

Reports are generated in $ARTIFACT_ROOT/sambatune_gen/<mm_dd_hh_s_(model_name)>/reports

```
/nvmedata/scratch/cobya/artifact_root/sambatune_gen/<mm_dd_hh_s_(model_name)>/reports#
ls *
collated_report.xlsx  full_stack_chrome_tracing.json  performance_summary.json
section_report.csv
analysis:
summary.json  summary.ui.json
benchmark-data:
<mm_dd_hh_s_(model_name)>_linear_net_measure_performance.json
<mm_dd_hh_s_(model_name)>_linear_net_measure_sections___debug.json
<mm_dd_hh_s_(model_name)>_linear_net_measure_performance___run_graph_only.json
…
snprof:
2764746_timeline_profiling.json  per_section.csv
<mm_dd_hh_s_(model_name)>_linear_net_measure_performance_snprof.json
2824840_timeline_profiling.json  per_tensor.csv   summary.csv
```

SambaTune UI

Available as a flask http sever using gunicorn from the bare-metal install or as an installable whl file for linux and mac. (Windows port is on the roadmap)

```
sambatune_ui --directory /nvmedata/scratch/cobya/artifact_root/sambatune_gen --port 8576
username: "admin", password: ”<unique every run>"
[2023-07-11 23:04:47 -0700] [647810] [INFO] Starting gunicorn 20.1.0
[2023-07-11 23:04:47 -0700] [647810] [INFO] Listening at: http://0.0.0.0:8576 (647810)
[2023-07-11 23:04:47 -0700] [647810] [INFO] Using worker: sync
[2023-07-11 23:04:47 -0700] [648624] [INFO] Booting worker with pid: 648624
[2023-07-11 23:04:47 -0700] [648655] [INFO] Booting worker with pid: 648655
```

SambaTune Demo