Qualcomm

APPS Training Session

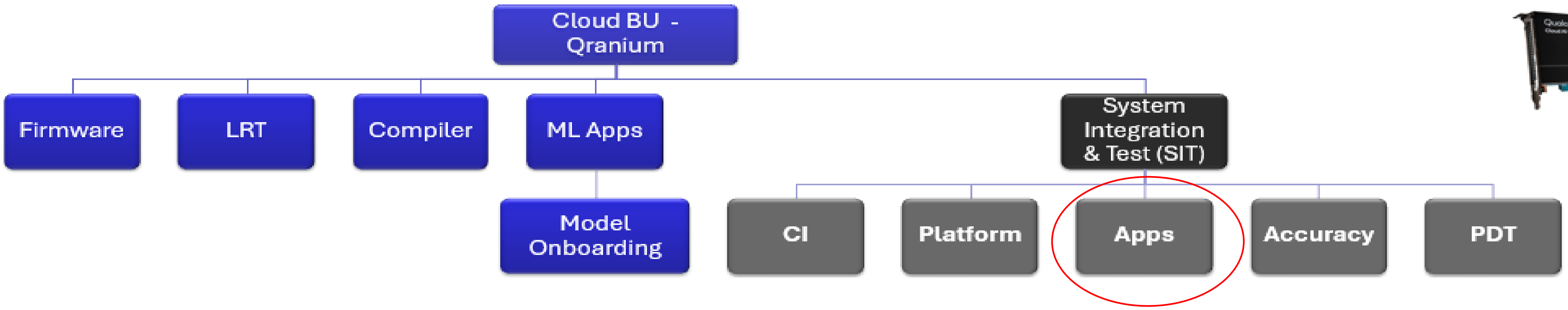
Agenda

Day: 1

- General Overview
- QAIC Apps Overview
- Accuracy tools
- Hands-ON : Performance & Accuracy
  - Compiling and measure performance of a model
  - Getting accuracy metrics of a model

Cloud BU : Qranium

Qranium Team


Cloud BU - Qranium

Firmware

LRT

Compiler

ML Apps

Model Onboarding

System Integration & Test (SIT)

CI

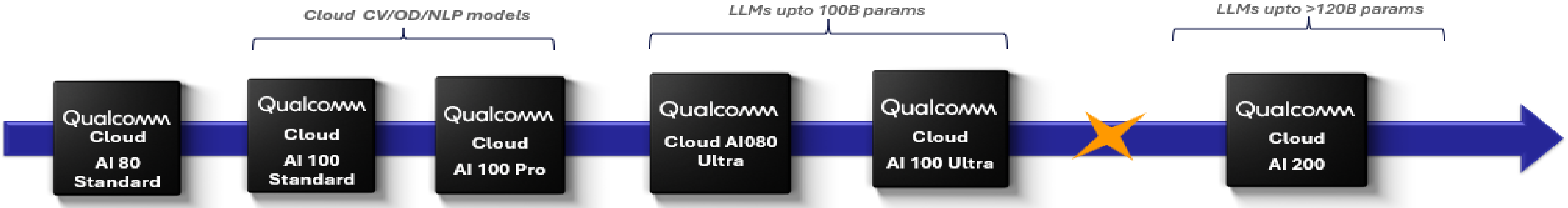
Platform

Apps

Accuracy

PDT

Product SKUs


Cloud CV/OD/NLP models

LLMs upto 100B params

LLMs upto >120B params

Qualcomm Cloud AI 80 Standard

NSP:8
SOC: 1
TDP: 75W
TOPS: 190
DRAM BW: 137 GB/s
Air cooled

Qualcomm Cloud AI 100 Standard

NSP: 14
SOC:1
TDP: 75W
TOPS: 350
DRAM BW: 137 GB/s
Air cooled

Qualcomm Cloud AI 100 Pro

NSP:16
SOC:1
TDP: 75W
TOPS: 400
DRAM BW: 137 GB/s
Air cooled

Qualcomm Cloud AI080 Ultra

NSP: 8
SOC:4
TDP: 150W
TOPS: 618
DRAM BW: 548 GB/s
Air cooled

Qualcomm Cloud AI 100 Ultra

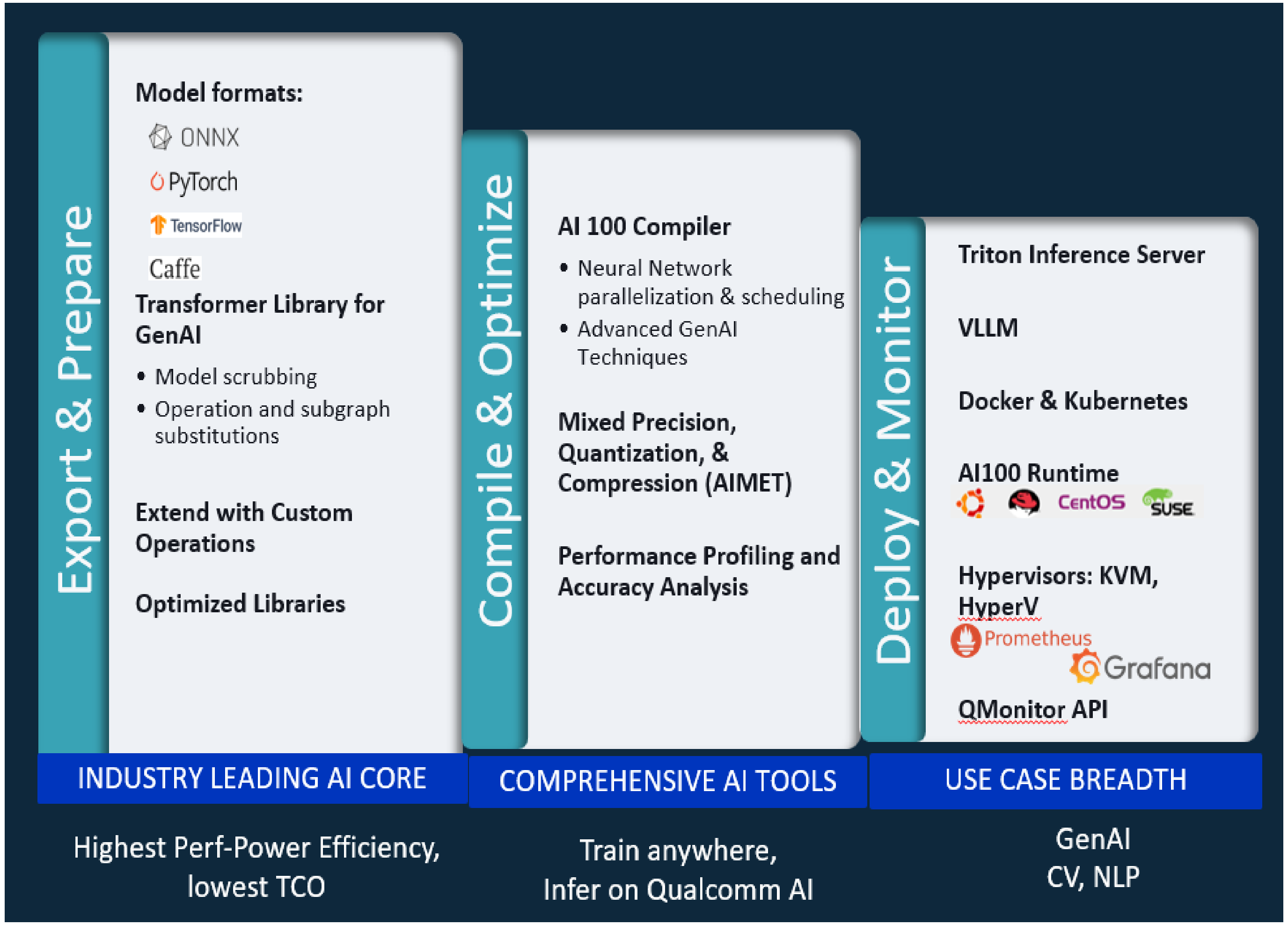
NSP:16
SOC:4
TDP: 150W
TOPS: 870
DRAM BW: 548 GB/s
Air cooled

Qualcomm Cloud AI 200

NSP:4 (Gen6)
SOC: Upto 24
TDP: Upto 2400W
DRAM BW: Upto 42TB/s
Immersion cooled

Cloud AI100

Getting-Started_Inference-Workflow/



Export & Prepare

Model formats:

- ONNX
- PyTorch
- TensorFlow
- Caffe

Transformer Library for GenAI

- Model scrubbing
- Operation and subgraph substitutions

Extend with Custom Operations

Optimized Libraries

INDUSTRY LEADING AI CORE

Highest Perf-Power Efficiency, lowest TCO

Compile & Optimize

AI 100 Compiler

- Neural Network parallelization & scheduling
- Advanced GenAI Techniques

Mixed Precision, Quantization, & Compression (AIMET)

Performance Profiling and Accuracy Analysis

COMPREHENSIVE AI TOOLS

Train anywhere, Infer on Qualcomm AI

Deploy & Monitor

Triton Inference Server

VLLM

Docker & Kubernetes

AI100 Runtime

CentOS SUSE

Hypervisors: KVM, HyperV

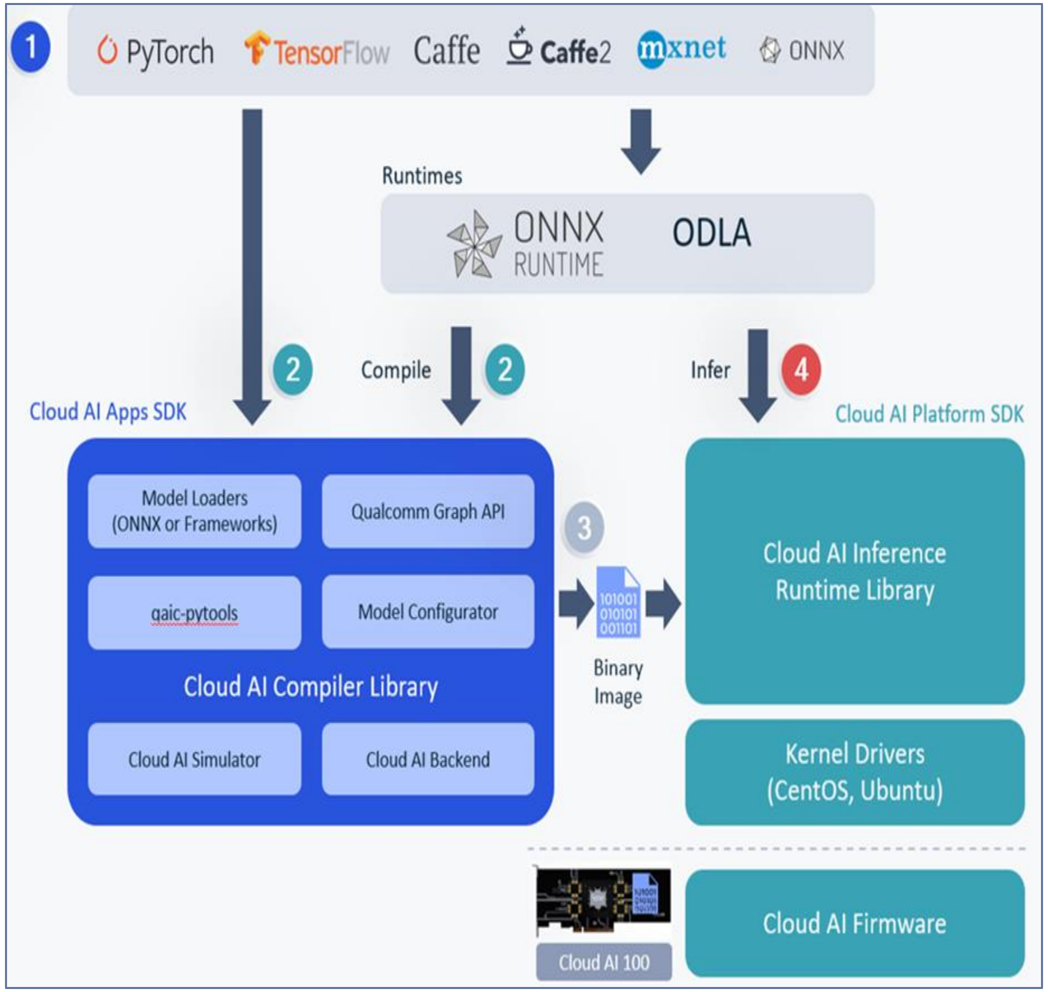
Prometheus

Grafana

QMonitor API

USE CASE BREADTH

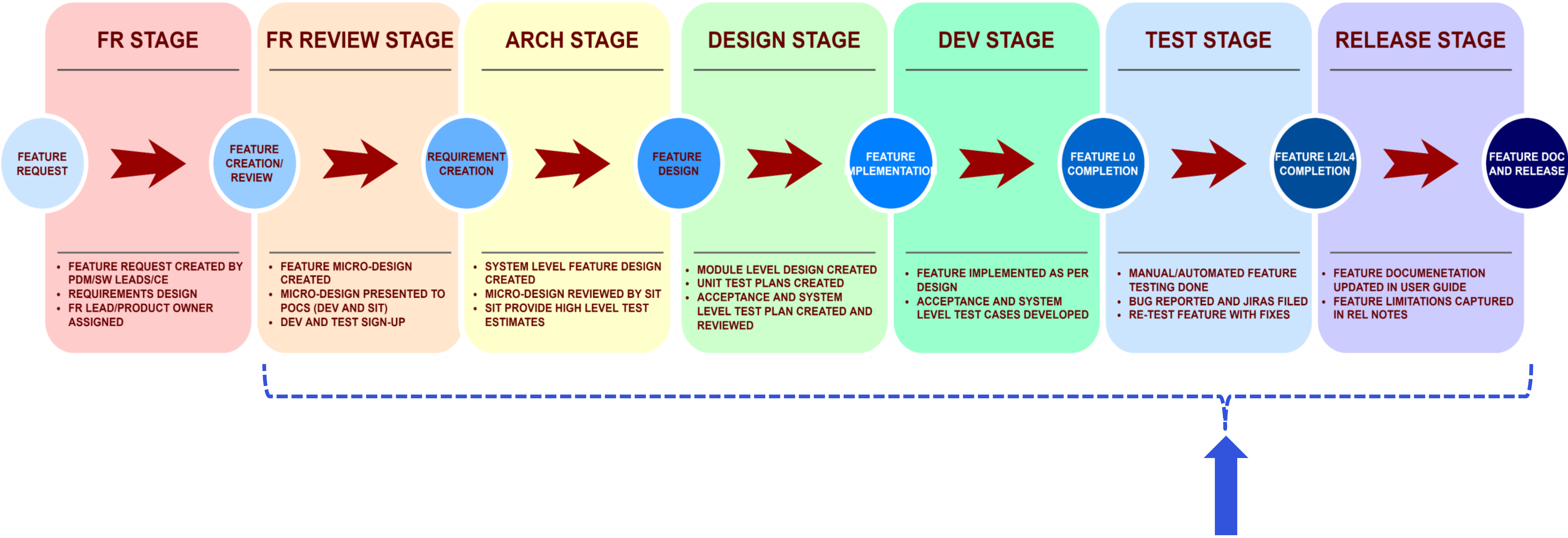
GenAI
CV, NLP



Software Phase

FR (Feature Request) → → → → → → → → → → → → → → → → → → Customer Release

FEATURE PLANNING AND TESTING PHASES



FEATURE REQUEST

FR STAGE

- FEATURE REQUEST CREATED BY PDM/SW LEADS/CE
- REQUIREMENTS DESIGN
- FR LEAD/PRODUCT OWNER ASSIGNED

FEATURE CREATION/ REVIEW

FR REVIEW STAGE

- FEATURE MICRO-DESIGN CREATED
- MICRO-DESIGN PRESENTED TO POCS (DEV AND SIT)
- DEV AND TEST SIGN-UP

REQUIREMENT CREATION

ARCH STAGE

- SYSTEM LEVEL FEATURE DESIGN CREATED
- MICRO-DESIGN REVIEWED BY SIT
- SIT PROVIDE HIGH LEVEL TEST ESTIMATES

FEATURE DESIGN

DESIGN STAGE

- MODULE LEVEL DESIGN CREATED
- UNIT TEST PLANS CREATED
- ACCEPTANCE AND SYSTEM LEVEL TEST PLAN CREATED AND REVIEWED

FEATURE MPLEMENTATIO

DEV STAGE

- FEATURE IMPLEMENTED AS PER DESIGN
- ACCEPTANCE AND SYSTEM LEVEL TEST CASES DEVELOPED

FEATURE L0 COMPLETION

TEST STAGE

- MANUAL/AUTOMATED FEATURE TESTING DONE
- BUG REPORTED AND JIRAS FILED
- RE-TEST FEATURE WITH FIXES

FEATURE L2/L4 COMPLETION

RELEASE STAGE

- FEATURE DOCUMENETATION UPDATED IN USER GUIDE
- FEATURE LIMITATIONS CAPTURED IN REL NOTES

FEATURE DOC AND RELEASE

Reading....

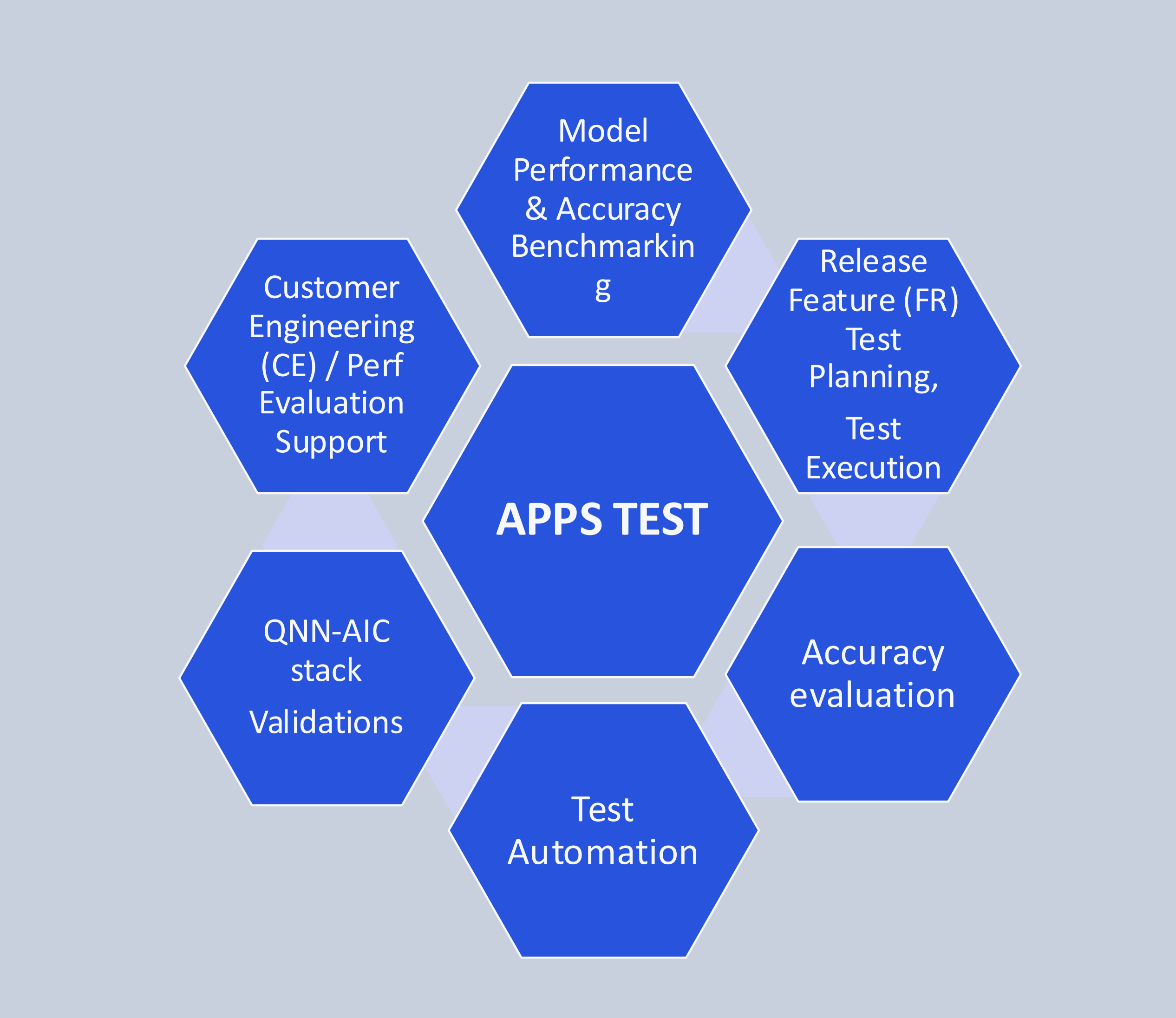
- Compilers
- https://unify.ai/blog/deep-learning-compilers
    - https://medium.com/geekculture/ai-compilers-ae28afbc4907
    - a-friendly-introduction-to-machine-learning-compilers-and-optimizers.html
    - https://mlc.ai/
    - https://www.modular.com/ai-resources/mac

- Deep Learning /LLMs
- https://www.youtube.com/playlist?list=PLqYmG7hTraZCDxZ44o4p3N5Anz3lLRVZF
- https://www.youtube.com/playlist?list=PLqGkIjcOyrGnjyBHl4GE2S9kX47X96FH-

- Quantization
    - https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-quantization

QAIC Apps Overview

APPS SIT : Charter



APPS TEST

Model Performance & Accuracy Benchmarking

Release Feature (FR) Test Planning,

Test Execution

Accuracy evaluation

Test Automation

QNN-AIC stack

Validations

Customer Engineering (CE) / Perf Evaluation Support

Framework : ONNX/Pytorch

ML Models : LLM, CV , OD, NLP, etc..

Precision : fp16, int8, int8_mp, etc..

Env/OS: Ubuntu 22 , KVM, RHEL

Model_zoo: \\blrsweng1\model_zoo\master\model_zoo

Test Reports : Daily Digest : Perf , Accuracy & Functional

Tools: JIRA QRANIUMSW (Bug reporting), Axiom (Test Planning)

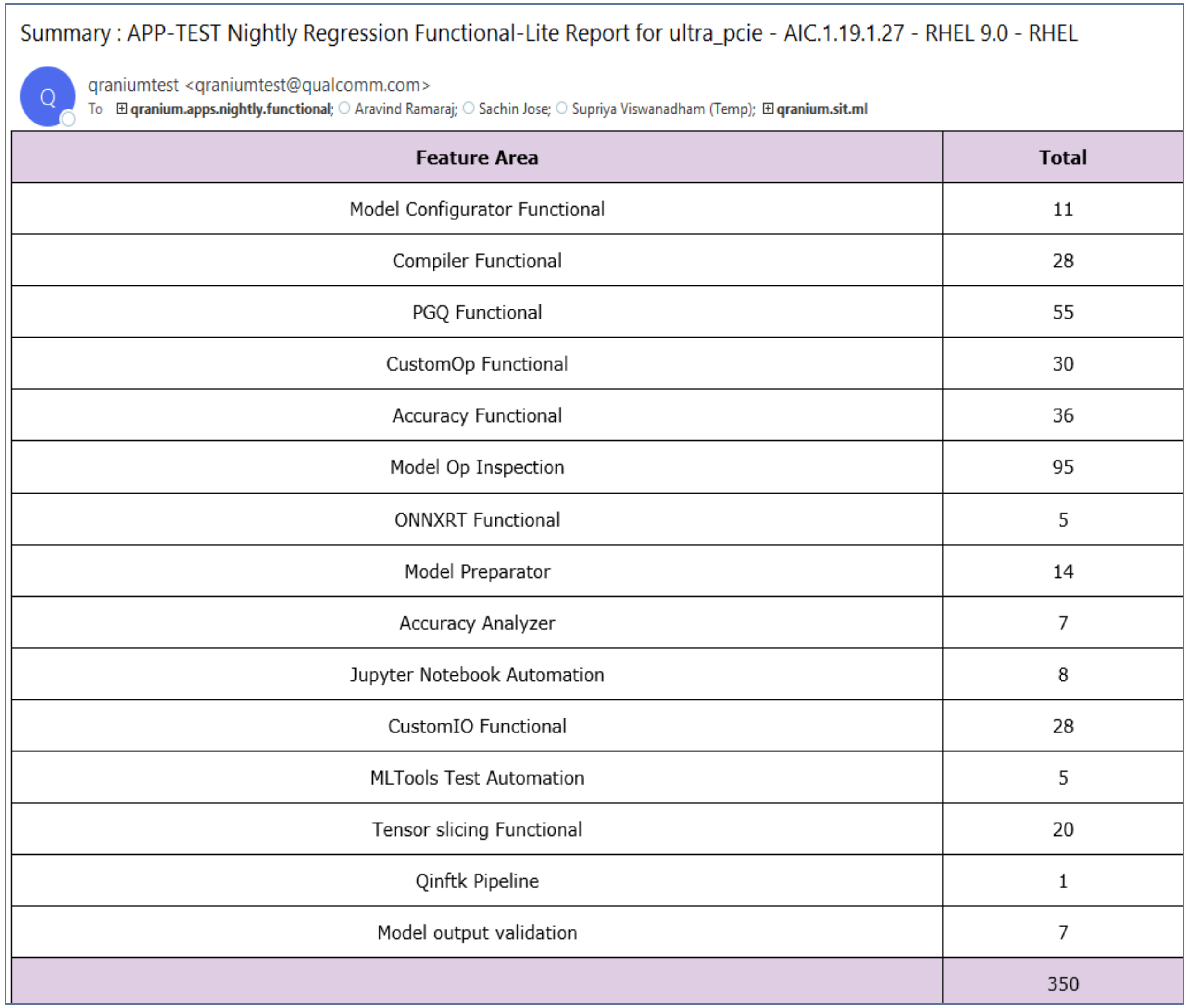
Test metrics (KPIs)

- **Performance :**
  - Model Throughput (Inf/sec)
  - LLM (token/sec)
- **Accuracy:**
  - diff metrics based on the model category:
  - F1, bbox, perplexity, etc..
- **Latency:** Response Times
- **Memory Utilization:** Host Peak memory, DDR required
- **CPU Utilization of Host**
- **Miscellaneous :** HMX and HVX Utilization, TDP etc..

Primary Tools

- Compilation
  - /opt/qti-aic/exec/qaic-exec
- Execution
  - /opt/qti-aic/exec/qaic-runner
- Accuracy Evaluator
  - /opt/qti-aic/tools/qaic-pytools/qaic-acc-evaluator.py
- Accuracy Debugger
  - /opt/qti-aic/tools/qaic-pytools/qaic-acc-analyzer.py
- Model Preparator
  - /opt/qti-aic/tools/qaic-pytools/qaic-model-preparator.py

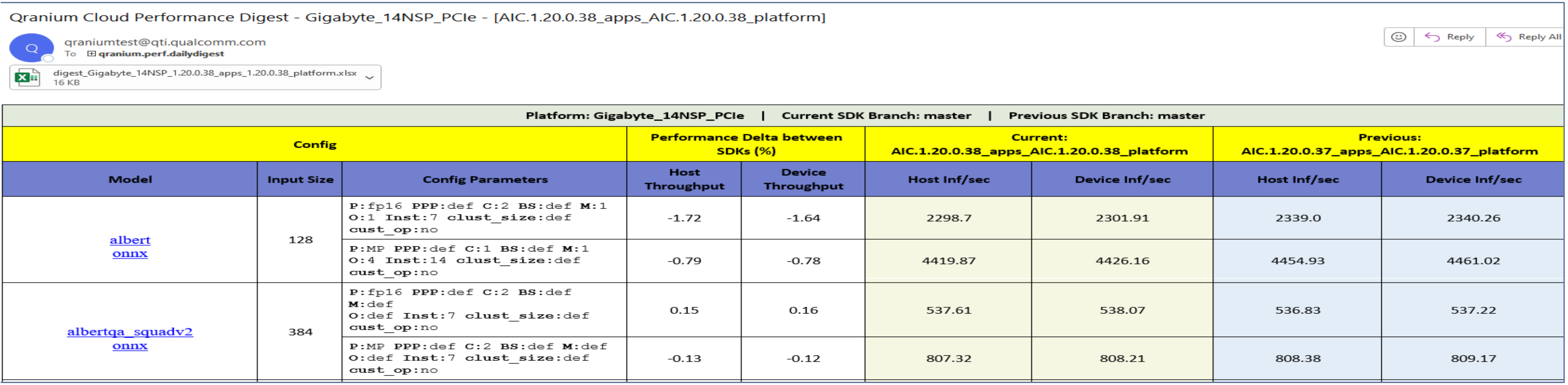
Functional nightly - Test reports



Summary : APP-TEST Nightly Regression Functional-Lite Report for ultra_pcie - AIC.1.19.1.27 - RHEL 9.0 - RHEL

qraniumtest <qraniumtest@qualcomm.com>
To ⊞ **qranium.apps.nightly.functional**; ○ Aravind Ramaraj; ○ Sachin Jose; ○ Supriya Viswanadham (Temp); ⊞ **qranium.sit.ml**

Feature Area	Total
Model Configurator Functional	11
Compiler Functional	28
PGQ Functional	55
CustomOp Functional	30
Accuracy Functional	36
Model Op Inspection	95
ONNXRT Functional	5
Model Preparator	14
Accuracy Analyzer	7
Jupyter Notebook Automation	8
CustomIO Functional	28
MLTools Test Automation	5
Tensor slicing Functional	20
Qinftk Pipeline	1
Model output validation	7
	350

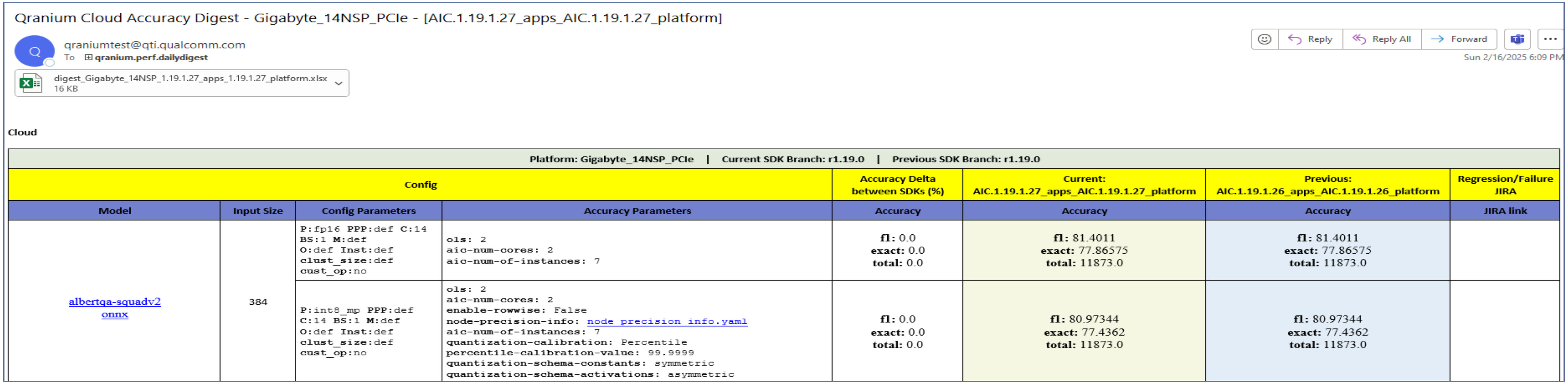


Qranium Cloud Performance Digest - Gigabyte_14NSP_PCIe - [AIC.1.20.0.38_apps_AIC.1.20.0.38_platform]

qraniumtest@qti.qualcomm.com
To ⊞ qranium.perf.dailydigest

digest_Gigabyte_14NSP_1.20.0.38_apps_1.20.0.38_platform.xlsx
16 KB

Platform: Gigabyte_14NSP_PCIe | Current SDK Branch: master | Previous SDK Branch: master								
Config			Performance Delta between SDKs (%)		Current: AIC.1.20.0.38_apps_AIC.1.20.0.38_platform		Previous: AIC.1.20.0.37_apps_AIC.1.20.0.37_platform	
Model	Input Size	Config Parameters	Host Throughput	Device Throughput	Host Inf/sec	Device Inf/sec	Host Inf/sec	Device Inf/sec
albert onnx	128	P:fp16 PPP:def C:2 BS:def M:1 O:1 Inst:7 clust_size:def cust_op:no	-1.72	-1.64	2298.7	2301.91	2339.0	2340.26
		P:MP PPP:def C:1 BS:def M:1 O:4 Inst:14 clust_size:def cust_op:no	-0.79	-0.78	4419.87	4426.16	4454.93	4461.02
albertqa_squadv2 onnx	384	P:fp16 PPP:def C:2 BS:def M:def O:def Inst:7 clust_size:def cust_op:no	0.15	0.16	537.61	538.07	536.83	537.22
		P:MP PPP:def C:2 BS:def M:def O:def Inst:7 clust_size:def cust_op:no	-0.13	-0.12	807.32	808.21	808.38	809.17



Qranium Cloud Accuracy Digest - Gigabyte_14NSP_PCIe - [AIC.1.19.1.27_apps_AIC.1.19.1.27_platform]

qraniumtest@qti.qualcomm.com
To ⊞ qranium.perf.dailydigest

Sun 2/16/2025 6:09 PM

digest_Gigabyte_14NSP_1.19.1.27_apps_1.19.1.27_platform.xlsx
16 KB

Cloud

Platform: Gigabyte_14NSP_PCIe | Current SDK Branch: r1.19.0 | Previous SDK Branch: r1.19.0							
Config				Accuracy Delta between SDKs (%)	Current: AIC.1.19.1.27_apps_AIC.1.19.1.27_platform	Previous: AIC.1.19.1.26_apps_AIC.1.19.1.26_platform	Regression/Failure JIRA
Model	Input Size	Config Parameters	Accuracy Parameters	Accuracy	Accuracy	Accuracy	JIRA link
albertqa-squadv2 onnx	384	P:fp16 PPP:def C:14 BS:1 M:def O:def Inst:def clust_size:def cust_op:no	ols: 2 aic-num-cores: 2 aic-num-of-instances: 7	**f1:** 0.0 **exact:** 0.0 **total:** 0.0	**f1:** 81.4011 **exact:** 77.86575 **total:** 11873.0	**f1:** 81.4011 **exact:** 77.86575 **total:** 11873.0	
		P:int8_mp PPP:def C:14 BS:1 M:def O:def Inst:def clust_size:def cust_op:no	ols: 2 aic-num-cores: 2 enable-rowwise: False node-precision-info: node_precision_info.yaml aic-num-of-instances: 7 quantization-calibration: Percentile percentile-calibration-value: 99.9999 quantization-schema-constants: symmetric quantization-schema-activations: asymmetric	**f1:** 0.0 **exact:** 0.0 **total:** 0.0	**f1:** 80.97344 **exact:** 77.4362 **total:** 11873.0	**f1:** 80.97344 **exact:** 77.4362 **total:** 11873.0	

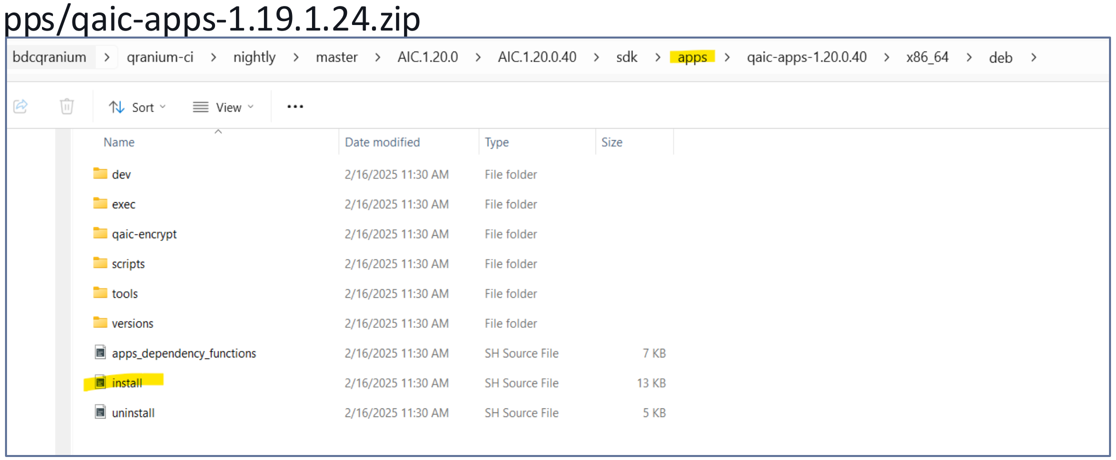
Hands - ON

Outcome: Ability to Install apps SDK , compile and run inference for any ML model

- Documentation :
  - https://quic.github.io/cloud-ai-sdk-pages/latest/Getting-Started/Quick-Start-Guide/
- Install : QAIC APPs and Platform SDK Cloud AI SDK - Cloud AI 100
  - Example:
    - /prj/bdcqranium/qranium-ci/nightly/r1.19.0/AIC.1.19.1/AIC.1.19.1.24/sdk/platform/silicon/release/qaic-platform-sdk-1.19.1.24.zip

    -
      ```
      unzip qaic-platform-sdk-1.19.1.24.zip
      cd qaic-platform-sdk-1.19.1.24/x86_64/deb
      ./uninstall.sh
      ./install.sh
      ```
  - Example:
    - /prj/bdcqranium/qranium-ci/nightly/r1.19.0/AIC.1.19.1/AIC.1.19.1.24/sdk/apps/qaic-apps-1.19.1.24.zip

    -
      ```
      unzip qaic-apps-sdk-1.19.1.24.zip
      cd qaic-apps-sdk-1.19.1.24/x86_64/deb
      ./uninstall.sh
      ./install.sh --enable-qaic-pytools
      ```
  - Example:
    - Qranium Nightly Build 1.20.0 - AIC.1.20.0.38 - [Build Status - SUCCESS]



bdcqranium › qranium-ci › nightly › master › AIC.1.20.0 › AIC.1.20.0.40 › sdk › apps › qaic-apps-1.20.0.40 › x86_64 › deb ›
Sort
View

Name	Date modified	Type	Size
dev	2/16/2025 11:30 AM	File folder	
exec	2/16/2025 11:30 AM	File folder	
qaic-encrypt	2/16/2025 11:30 AM	File folder	
scripts	2/16/2025 11:30 AM	File folder	
tools	2/16/2025 11:30 AM	File folder	
versions	2/16/2025 11:30 AM	File folder	
apps_dependency_functions	2/16/2025 11:30 AM	SH Source File	7 KB
install	2/16/2025 11:30 AM	SH Source File	13 KB
uninstall	2/16/2025 11:30 AM	SH Source File	5 KB

Performance metrics

Performance : Compiling and running a model in QAIC (Non-LLM)

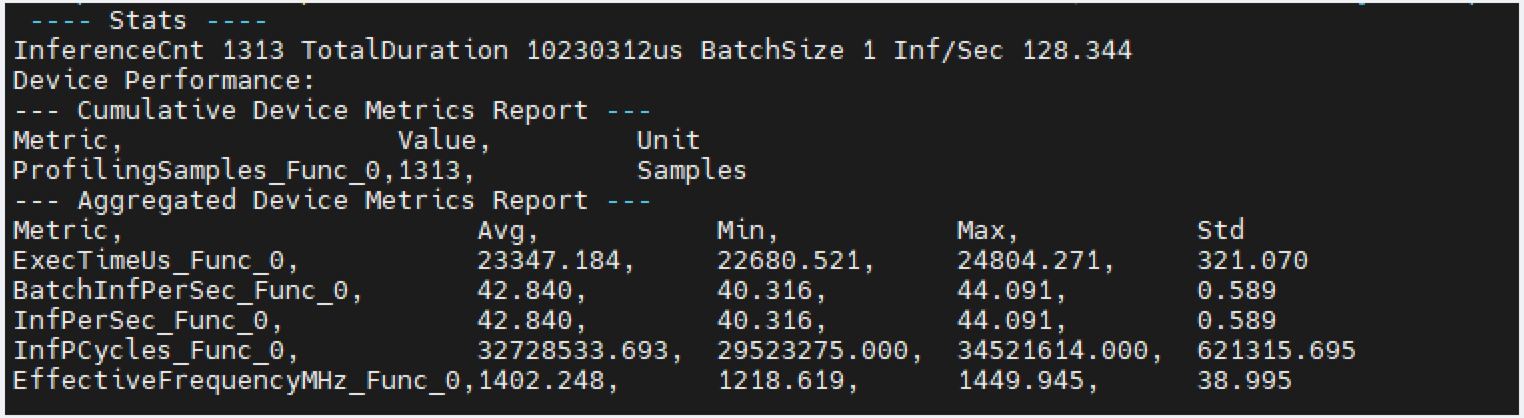
CASE 1: FP16 PRECISON

- **Compile command:**

```
/opt/qti-aic/exec/qaic-exec -aic-num-cores=4 -mos=1 -ols=1 -m=/home/qraniumtest/model_zoo/customer/MSFT/Big_Bird/generatedModels/ms_config/BigBird_bs_2048_msconfig_64blk_blockwiseattn_sim.onnx -input-list-file=/home/qraniumtest/model_zoo/model-inputs/inputs/Big_Bird/SL-2048/batch_size_1/file-list.txt -aic-binary-dir=/home/qraniumtest/binaries/BigBird/aic/ -aic-hw -aic-hw-version=2.0 -convert-to-fp16 -stats-batchsize=1 -onnx-define-symbol=batch_size,1 -multicast-weights -aic-perf-warnings -aic-perf-metrics -stats-level=40 -size-split-granularity=2048 -compile-only
```

- **qaic-runner command:**

```
cd /home/qraniumtest/model_zoo/model-inputs/inputs/Big_Bird/SL-2048/batch_size_1 && /opt/qti-aic/exec/qaic-runner -t /home/qraniumtest/binaries/BigBird/aic/ -a 3 -i ./input_ids.raw -i ./attention_mask.raw -T 4 --time 10 -d 5
```



```
 ---- Stats ----
InferenceCnt 1313 TotalDuration 10230312us BatchSize 1 Inf/Sec 128.344
Device Performance:
--- Cumulative Device Metrics Report ---
Metric,                 Value,        Unit
ProfilingSamples_Func_0,1313,         Samples
--- Aggregated Device Metrics Report ---
Metric,                      Avg,           Min,           Max,           Std
ExecTimeUs_Func_0,           23347.184,     22680.521,     24804.271,     321.070
BatchInfPerSec_Func_0,       42.840,        40.316,        44.091,        0.589
InfPerSec_Func_0,            42.840,        40.316,        44.091,        0.589
InfPCycles_Func_0,           32728533.693,  29523275.000,  34521614.000,  621315.695
EffectiveFrequencyMHz_Func_0,1402.248,      1218.619,      1449.945,      38.995
```

Performance : Compiling and running a model in QAIC (Non-LLM)

CASE 2: INT8

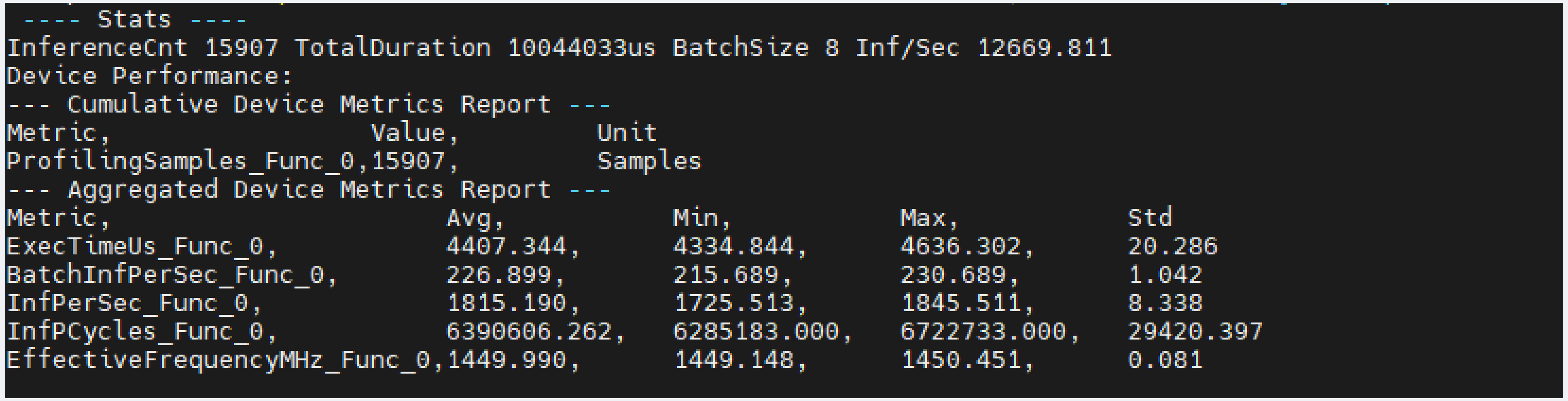
➢ Compile command:

/opt/qti-aic/exec/qaic-exec -aic-num-cores=2 -batchsize=8 -mos=4 -ols=4 -dump-profile=/home/qraniumtest/pgq_profiles/densenet169.yaml -m=/home/qraniumtest/model_zoo/internal/DenseNet169/generatedModels/ONNX/densenet169.onnx -input-list-file=/home/qraniumtest/model_zoo/model-inputs/inputs/224x224/batch_size_8/file-list.txt

/opt/qti-aic/exec/qaic-exec -aic-num-cores=2 -batchsize=8 -mos=4 -ols=4 -load-profile=/home/qraniumtest/pgq_profiles/densenet169.yaml -m=/home/qraniumtest/model_zoo/internal/DenseNet169/generatedModels/ONNX/densenet169.onnx -input-list-file=/home/qraniumtest/model_zoo/model-inputs/inputs/224x224/batch_size_8/file-list.txt -aic-binary-dir=/home/qraniumtest/binaries/DenseNet169/aic/ -aic-hw -aic-hw-version=2.0 -quantization-schema-activations=symmetric_with_uint8 -quantization-schema-constants=symmetric_with_uint8 -quantization-precision=Int8 -aic-perf-warnings -aic-perf-metrics -stats-level=40 -compile-only

➢ qaic-runner command:

cd /home/qraniumtest/model_zoo/model-inputs/inputs/224x224/batch_size_8 && /opt/qti-aic/exec/qaic-runner -t /home/qraniumtest/binaries/DenseNet169/aic/ -a 7 -i ./img_0.raw -T 4 --time 10 -d 5



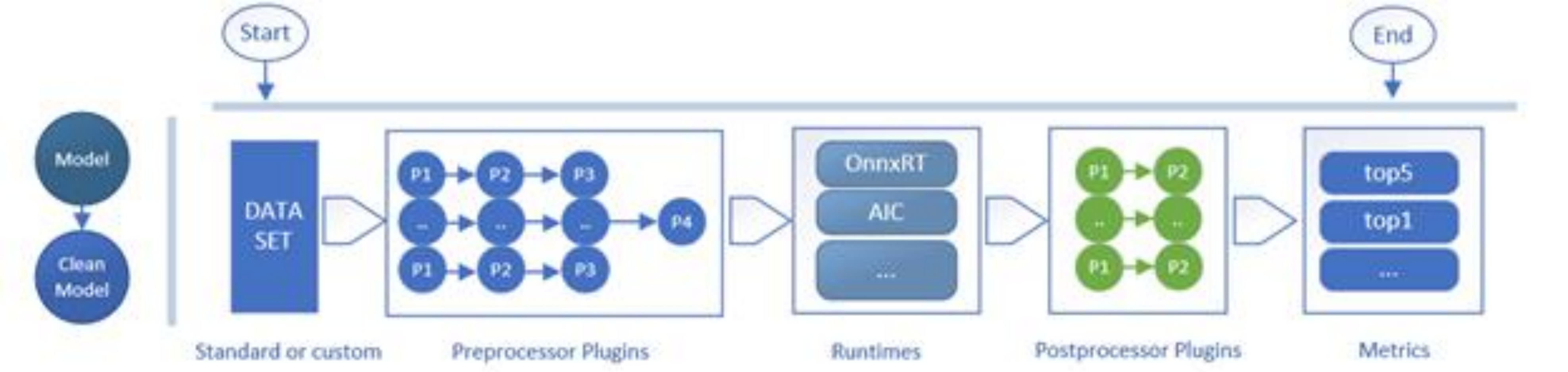
```
 ---- Stats ----
InferenceCnt 15907 TotalDuration 10044033us BatchSize 8 Inf/Sec 12669.811
Device Performance:
--- Cumulative Device Metrics Report ---
Metric,                 Value,         Unit
ProfilingSamples_Func_0,15907,         Samples
--- Aggregated Device Metrics Report ---
Metric,                      Avg,            Min,            Max,            Std
ExecTimeUs_Func_0,           4407.344,       4334.844,       4636.302,       20.286
BatchInfPerSec_Func_0,       226.899,        215.689,        230.689,        1.042
InfPerSec_Func_0,            1815.190,       1725.513,       1845.511,       8.338
InfPCycles_Func_0,           6390606.262,    6285183.000,    6722733.000,    29420.397
EffectiveFrequencyMHz_Func_0,1449.990,       1449.148,       1450.451,       0.081
```

Accuracy evaluator Tool & metrics

Accuracy tools in QAIC

Introduction

- The Accuracy evaluator is a framework to define and execute the end-to-end accuracy evaluation of a given model. The evaluation pipeline is configured in a yaml config file for a model and the tool loads this config file to execute the pipeline.
- The Pipeline consists of the below stages. The tool has options to run the complete pipeline or parts of it.
  - Selecting a Dataset
  - Running Preprocessors on the dataset
  - Running inference on the processed data on one or more platforms.
  - Post processing of inference raw outputs.
  - Accuracy Metrics evaluation



Reference link

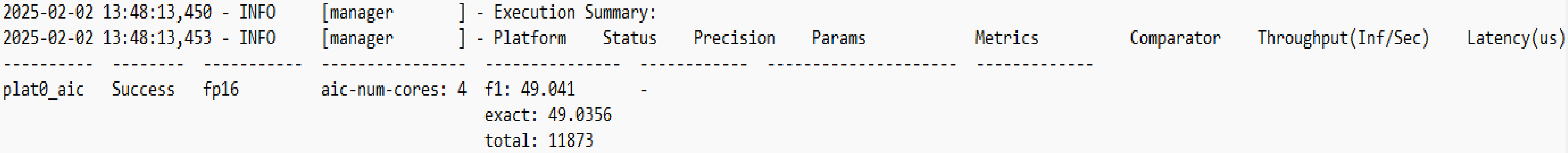
Accuracy tools in QAIC

BigBird – FP16 precision

- **Command used –**

source /opt/qti-aic/dev/python/qaic-env/bin/activate && python3 /opt/qti-aic/tools/qaic-pytools/**qaic-acc-evaluator.py** -onnx-symbol seg_length:2048 -onnx-symbol batch_size:1 -config /local/mnt/workspace/jyashwan/APPS_TRAINING/bigbird_config.yaml -cleanup end -work-dir /local/mnt/workspace/jyashwan/APPS_TRAINING/bigbird_fp16 -silent -platform-tag aic_fp16 -device-id 0

- **Metrics -**

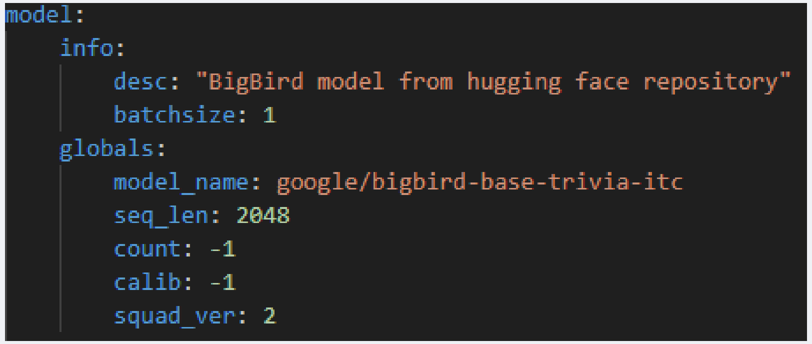


```
2025-02-02 13:48:13,450 - INFO     [manager        ] - Execution Summary:
2025-02-02 13:48:13,453 - INFO     [manager        ] - Platform   Status    Precision    Params            Metrics          Comparator    Throughput(Inf/Sec)    Latency(us)
----------  --------  -----------  ----------------  ---------------  ------------  ---------------------  -------------
plat0_aic   Success   fp16         aic-num-cores: 4  f1: 49.041       -
                                                     exact: 49.0356
                                                     total: 11873
```

Accuracy tools in QAIC

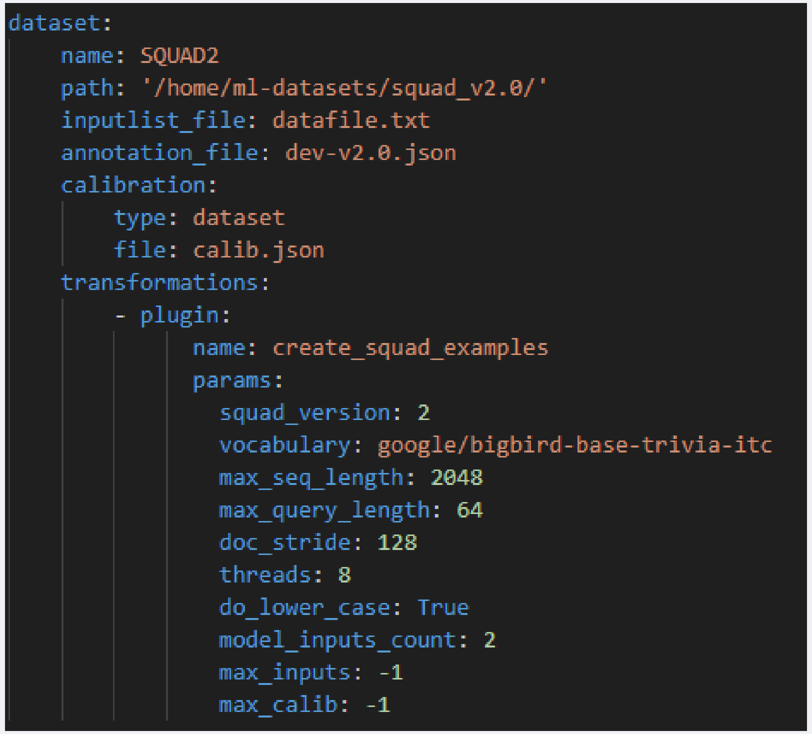
BigBird – FP16 precision model config

1



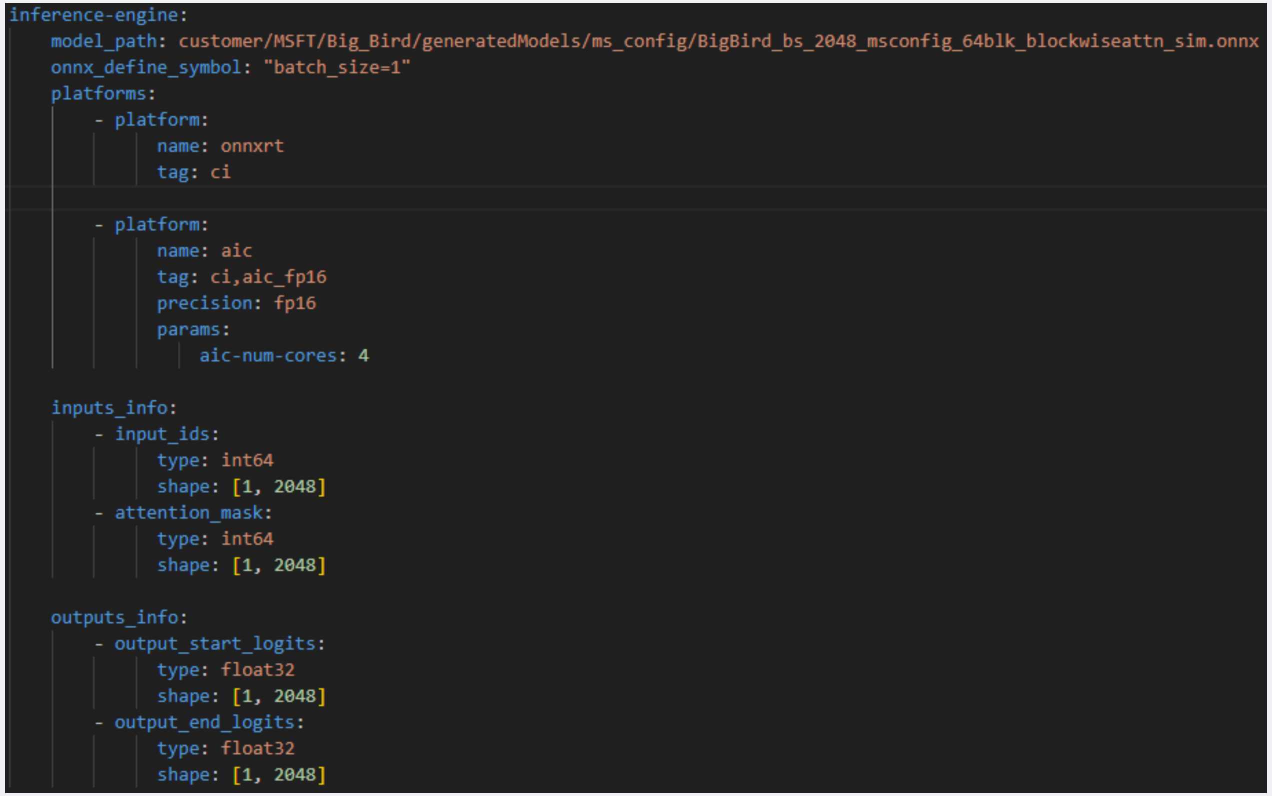
```yaml
model:
    info:
        desc: "BigBird model from hugging face repository"
        batchsize: 1
    globals:
        model_name: google/bigbird-base-trivia-itc
        seq_len: 2048
        count: -1
        calib: -1
        squad_ver: 2
```

2



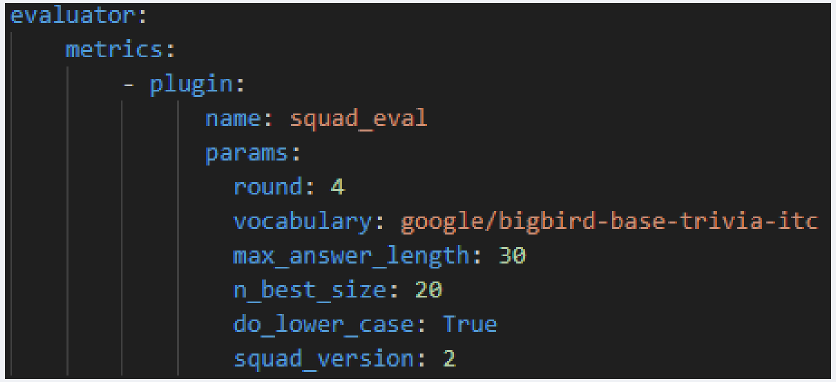
```yaml
dataset:
    name: SQUAD2
    path: '/home/ml-datasets/squad_v2.0/'
    inputlist_file: datafile.txt
    annotation_file: dev-v2.0.json
    calibration:
        type: dataset
        file: calib.json
    transformations:
        - plugin:
            name: create_squad_examples
            params:
              squad_version: 2
              vocabulary: google/bigbird-base-trivia-itc
              max_seq_length: 2048
              max_query_length: 64
              doc_stride: 128
              threads: 8
              do_lower_case: True
              model_inputs_count: 2
              max_inputs: -1
              max_calib: -1
```

3



```yaml
inference-engine:
    model_path: customer/MSFT/Big_Bird/generatedModels/ms_config/BigBird_bs_2048_msconfig_64blk_blockwiseattn_sim.onnx
    onnx_define_symbol: "batch_size=1"
    platforms:
        - platform:
            name: onnxrt
            tag: ci

        - platform:
            name: aic
            tag: ci,aic_fp16
            precision: fp16
            params:
                aic-num-cores: 4

    inputs_info:
        - input_ids:
            type: int64
            shape: [1, 2048]
        - attention_mask:
            type: int64
            shape: [1, 2048]

    outputs_info:
        - output_start_logits:
            type: float32
            shape: [1, 2048]
        - output_end_logits:
            type: float32
            shape: [1, 2048]
```

4



```yaml
evaluator:
    metrics:
        - plugin:
            name: squad_eval
            params:
              round: 4
              vocabulary: google/bigbird-base-trivia-itc
              max_answer_length: 30
              n_best_size: 20
              do_lower_case: True
              squad_version: 2
```

Agenda

Day: 2

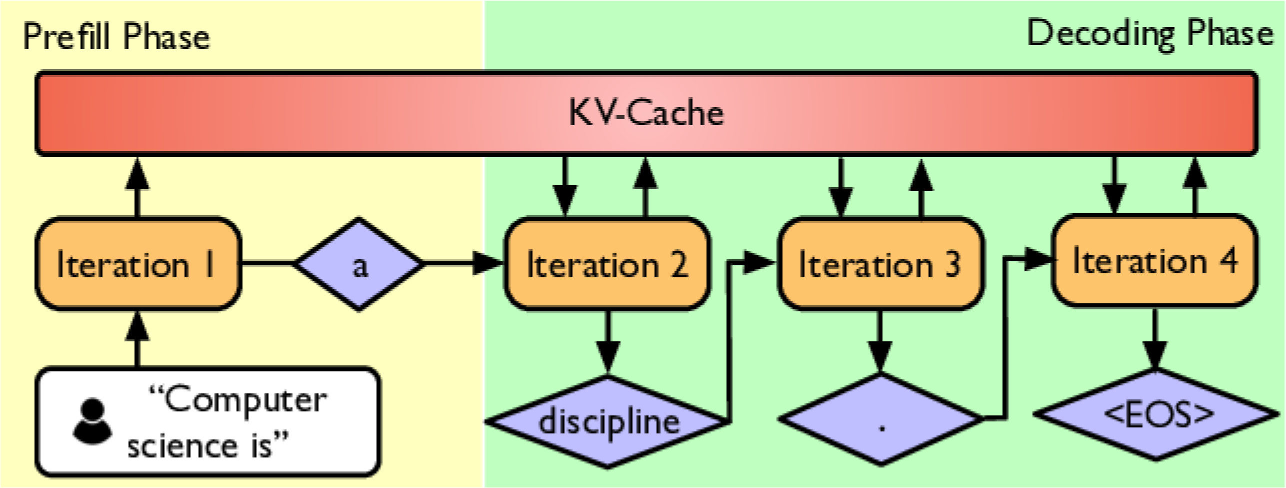
- LLM
- Hands ON
  - Qeff
  - VLLM

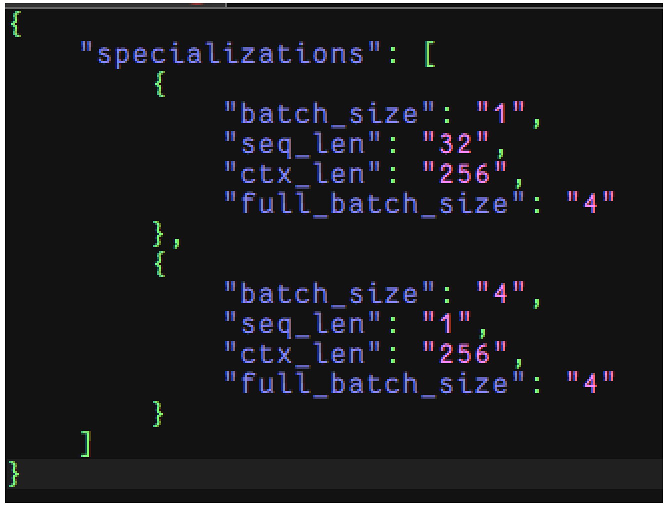
LLM - Serving platforms

LLM

Readings:

- Large Language Models explained briefly
- Modeling | CS324
- LLM Bootcamp - Spring 2023 - The Full Stack
- LLM University (LLMU)
- Generative AI for Beginners
- GitHub - mlabonne/llm-course: Course to get into Large Language Models (LLMs) with roadmaps and Colab notebooks.
- GitHub - Hannibal046/Awesome-LLM: Awesome-LLM: a curated list of Large Language Model





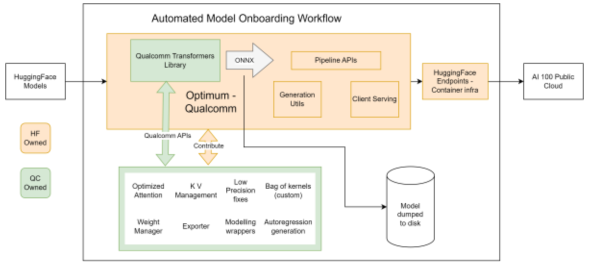
```
{
    "specializations": [
        {
            "batch_size": "1",
            "seq_len": "32",
            "ctx_len": "256",
            "full_batch_size": "4"
        },
        {
            "batch_size": "4",
            "seq_len": "1",
            "ctx_len": "256",
            "full_batch_size": "4"
        }
    ]
}
```

Cloud BU : LLM Serving Platforms & Frameworks

Inference as a Service : IaaS

QEfficient-transformers

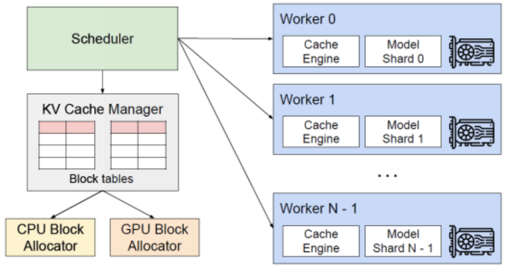
Train anywhere, Infer on Qualcomm Cloud AI with a Developer-centric Toolchain



- library provides reimplemented blocks of LLMs which are used to make the models functional and highly performant on Qualcomm Cloud AI 100.
- support wide range of model architectures, for easy efficient deployment on Cloud AI 100 cards
- Users only need to provide model card from HuggingFace or Path to the local model and the library will take care of transforming model to its efficient implementation for Cloud AI 100.

vLLM

Easy, fast, and cheap LLM serving for everyone
Supported in Qualcomm tool chain



- **vLLM** is a high-throughput LLM serving engine that achieves **near-zero waste** in KV cache memory.
- Implements Paged Attention as its core attention algorithm and Continuous batching.

Key Features
* High Throughput
* Memory Efficiency
* Seamless Integration

* Continuous Batching
* Distributed Inference
* API Support
* Streaming Outputs

NVIDIA and other open-source software supported by Qualcomm

LLM Framework

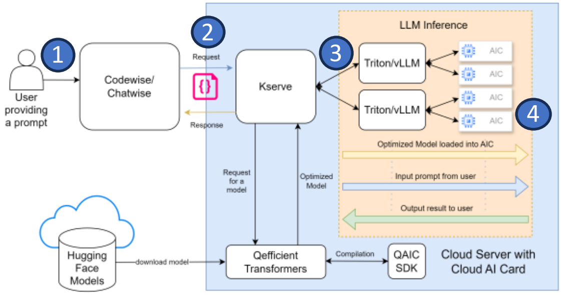
Inference as Service [IaaS]

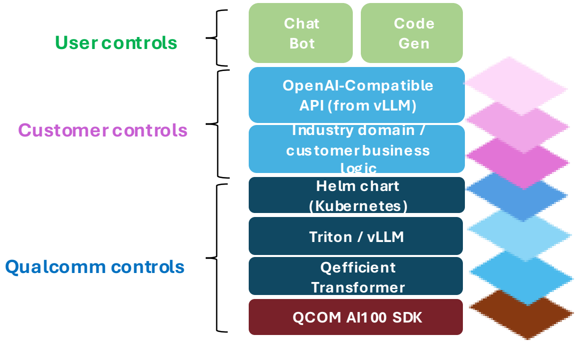
1)User provides a prompt in Codewise/Chatwise UI.

2)The response is passed to the server via APIs.

3)Kserve gets a new request and allocates hardware and software resources (vLLM/Triton) to operate on the input request.

4)vLLM takes optimized model, generated by Qefficient transformers and executes it in Cloud AI 100 card

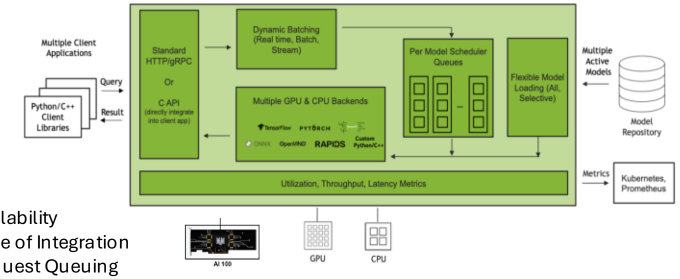




NVIDIA and other open-source software supported by Qualcomm

TRITON (Nvidia)

The Triton Inference Server is an open-source software developed by NVIDIA that standardizes and optimizes the deployment of AI models across various platforms and workloads.



Multiple Client Applications
Python/C++ Client Libraries
Query
Result
Standard HTTP/gRPC
Or
C API (directly integrate into client app)
Dynamic Batching (Real time, Batch, Stream)
Multiple GPU & CPU Backends
TensorFlow
PYTORCH
ONNX
OpenVINO
RAPIDS
Custom Python/C++
Per Model Scheduler Queues
Flexible Model Loading (All, Selective)
Multiple Active Models
Model Repository
Utilization, Throughput, Latency Metrics
Metrics
Kubernetes, Prometheus
AI 100
GPU
CPU

Key Features -

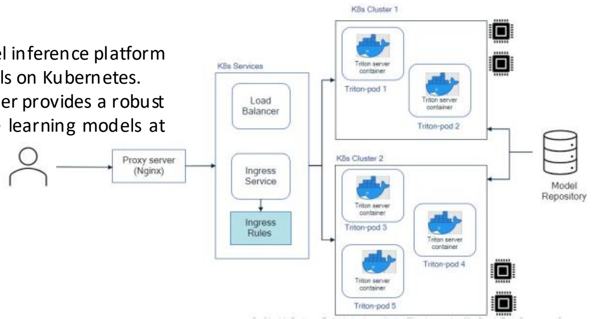
- Concurrent Model Execution
- Dynamic Batching
- Model Ensemble
- Scalability
- Ease of Integration
- Request Queuing

KSERVE / KUBERNETES

- KServe is an open-source, cloud-agnostic model inference platform designed to serve machine learning (ML) models on Kubernetes.
- Using Kserve with NVIDIA Triton Inference Server provides a robust solution for deploying and managing machine learning models at scale.

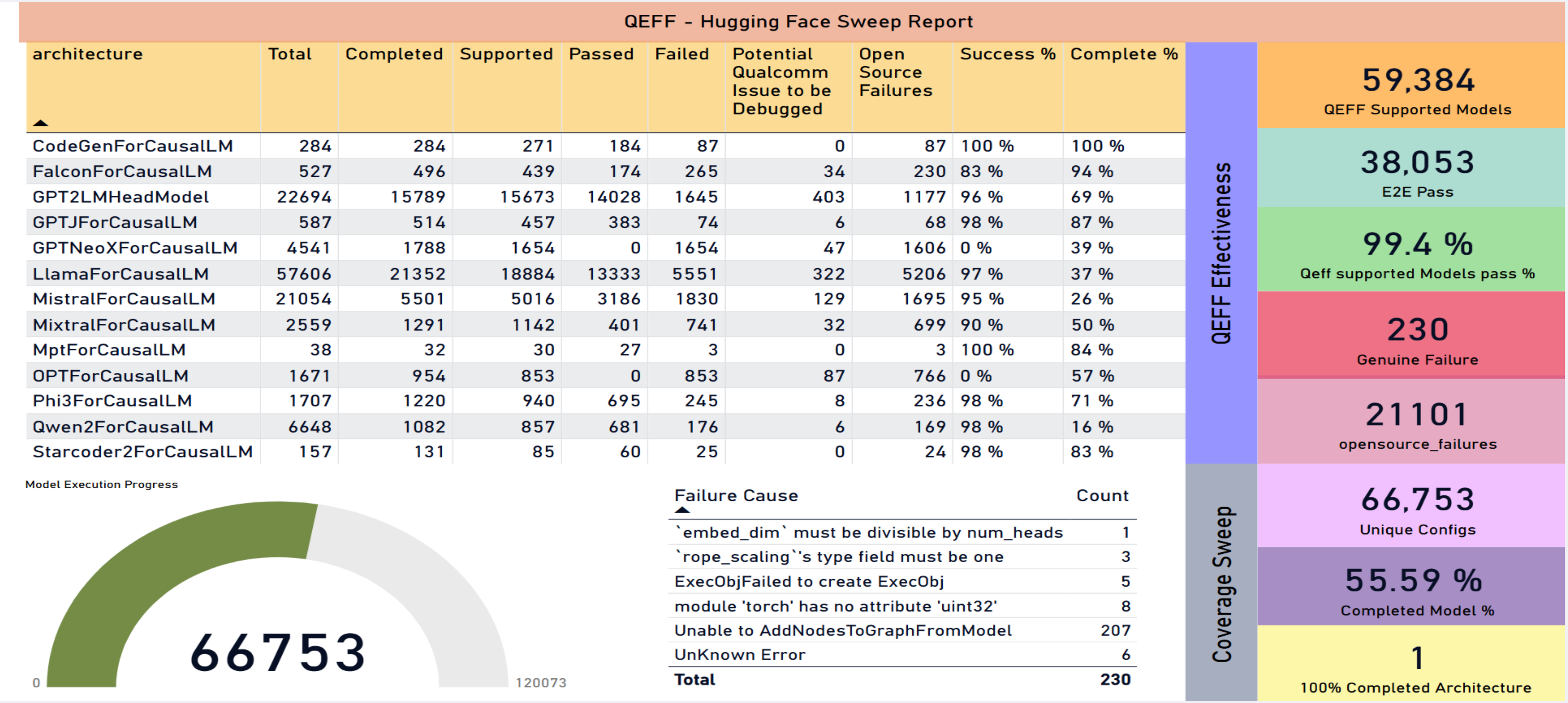
Key Features :

- Supports difference Inference Protocols
- Multi-Model Serving
- Scalability and Efficiency
- Ease of Deployment



Efficient Transformers

HF_Sweep_PowerBi_dashboard



QEFF - Hugging Face Sweep Report

architecture	Total	Completed	Supported	Passed	Failed	Potential Qualcomm Issue to be Debugged	Open Source Failures	Success %	Complete %
CodeGenForCausalLM	284	284	271	184	87	0	87	100 %	100 %
FalconForCausalLM	527	496	439	174	265	34	230	83 %	94 %
GPT2LMHeadModel	22694	15789	15673	14028	1645	403	1177	96 %	69 %
GPTJForCausalLM	587	514	457	383	74	6	68	98 %	87 %
GPTNeoXForCausalLM	4541	1788	1654	0	1654	47	1606	0 %	39 %
LlamaForCausalLM	57606	21352	18884	13333	5551	322	5206	97 %	37 %
MistralForCausalLM	21054	5501	5016	3186	1830	129	1695	95 %	26 %
MixtralForCausalLM	2559	1291	1142	401	741	32	699	90 %	50 %
MptForCausalLM	38	32	30	27	3	0	3	100 %	84 %
OPTForCausalLM	1671	954	853	0	853	87	766	0 %	57 %
Phi3ForCausalLM	1707	1220	940	695	245	8	236	98 %	71 %
Qwen2ForCausalLM	6648	1082	857	681	176	6	169	98 %	16 %
Starcoder2ForCausalLM	157	131	85	60	25	0	24	98 %	83 %

Model Execution Progress

66753

0

120073

Failure Cause	Count
\`embed_dim\` must be divisible by num_heads	1
\`rope_scaling\`'s type field must be one	3
ExecObjFailed to create ExecObj	5
module 'torch' has no attribute 'uint32'	8
Unable to AddNodesToGraphFromModel	207
UnKnown Error	6
Total	**230**

QEFF Effectiveness

59,384
QEFF Supported Models

38,053
E2E Pass

99.4 %
Qeff supported Models pass %

230
Genuine Failure

21101
opensource_failures

Coverage Sweep

66,753
Unique Configs

55.59 %
Completed Model %

1
100% Completed Architecture

Efficient Transformers : Running LLMs in QAIC with Qefficient

- Documentation link-GitHub - quic/efficient-transformers :
- Execution steps:

1. source qeff_env/bin/activate
2. pip install git+https://github.com/quic/efficient-transformers
3. Set export variables
   - ✓ export HF_HOME=/home/qraniumtest/gayav/llm_demo/models/
   - ✓ export QEFF_HOME=//home/qraniumtest/gayav/llm_demo/models/
   - ✓ export HF_TOKEN="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"

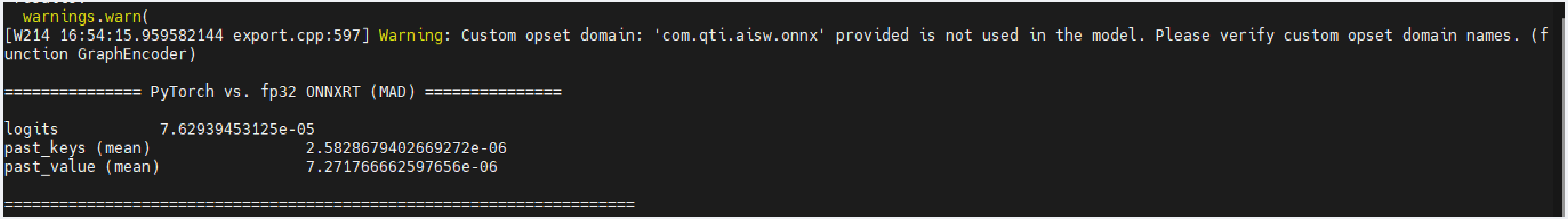
(OR)

1. docker pull docker-registry.qualcomm.com/qraniumtest/qranium:1.19.1.24-ubuntu22-x86_64
2. docker run --privileged -dit -v /home/:/home/ --name qaic_docker docker-registry.qualcomm.com/qraniumtest/qranium:1.19.1.24-ubuntu22-x86_64
3. docker exec -it qaic_docker bash
4. source /opt/qeff-env/bin/activate
5. Set export variables
   - ✓ export HF_HOME=/home/qraniumtest/gayav/llm_demo/models/
   - ✓ export QEFF_HOME=//home/qraniumtest/gayav/llm_demo/models/
   - ✓ export HF_TOKEN="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"

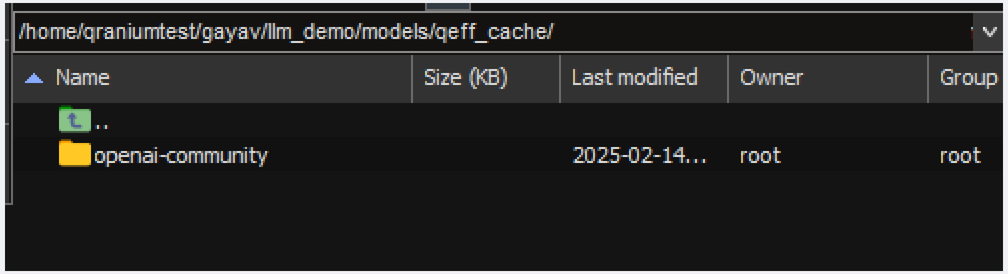
Running LLMs in QAIC using Qefficient

QEfficient.cloud.export

python -m QEfficient.cloud.**export** --model_name openai-community/gpt2 --cache_dir /home/qraniumtest/gayav/llm_demo/models/ --hf-token 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 --full_batch_size 4



```
  warnings.warn(
[W214 16:54:15.959582144 export.cpp:597] Warning: Custom opset domain: 'com.qti.aisw.onnx' provided is not used in the model. Please verify custom opset domain names. (f
unction GraphEncoder)

=============== PyTorch vs. fp32 ONNXRT (MAD) ===============

logits              7.62939453125e-05
past_keys (mean)                 2.5828679402669272e-06
past_value (mean)                7.271766662597656e-06

======================================================================
```



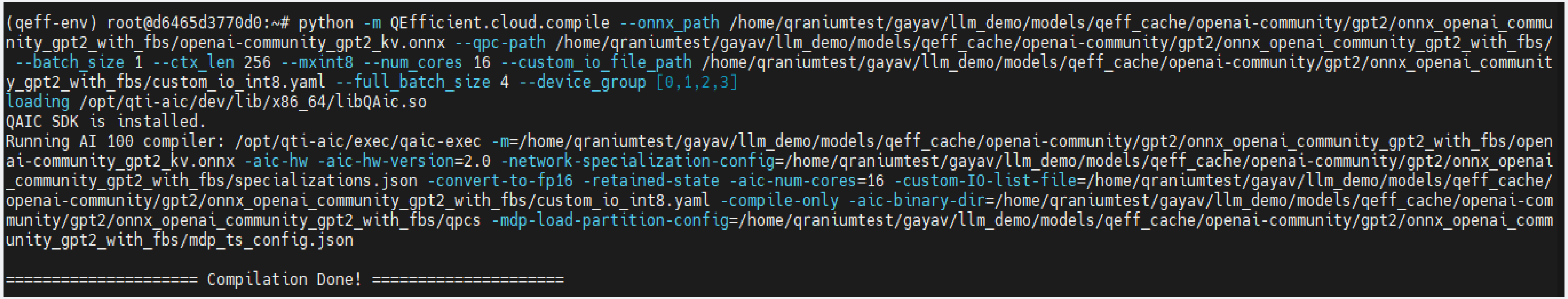
/home/qraniumtest/gayav/llm_demo/models/qeff_cache/

Name	Size (KB)	Last modified	Owner	Group
..				
openai-community		2025-02-14...	root	root

Running LLMs in QAIC with Qefficient

QEfficient.cloud.compile

- python -m QEfficient.cloud.**compile** --onnx_path /home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/onnx_openai_community_gpt2_with_fbs/openai-community_gpt2_kv.onnx --qpc-path /home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/onnx_openai_community_gpt2_with_fbs/ --batch_size 1 --ctx_len 256 --mxint8 --num_cores 16 --custom_io_file_path /home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/onnx_openai_community_gpt2_with_fbs/custom_io_int8.yaml --full_batch_size 4 --device_group [0,1,2,3]

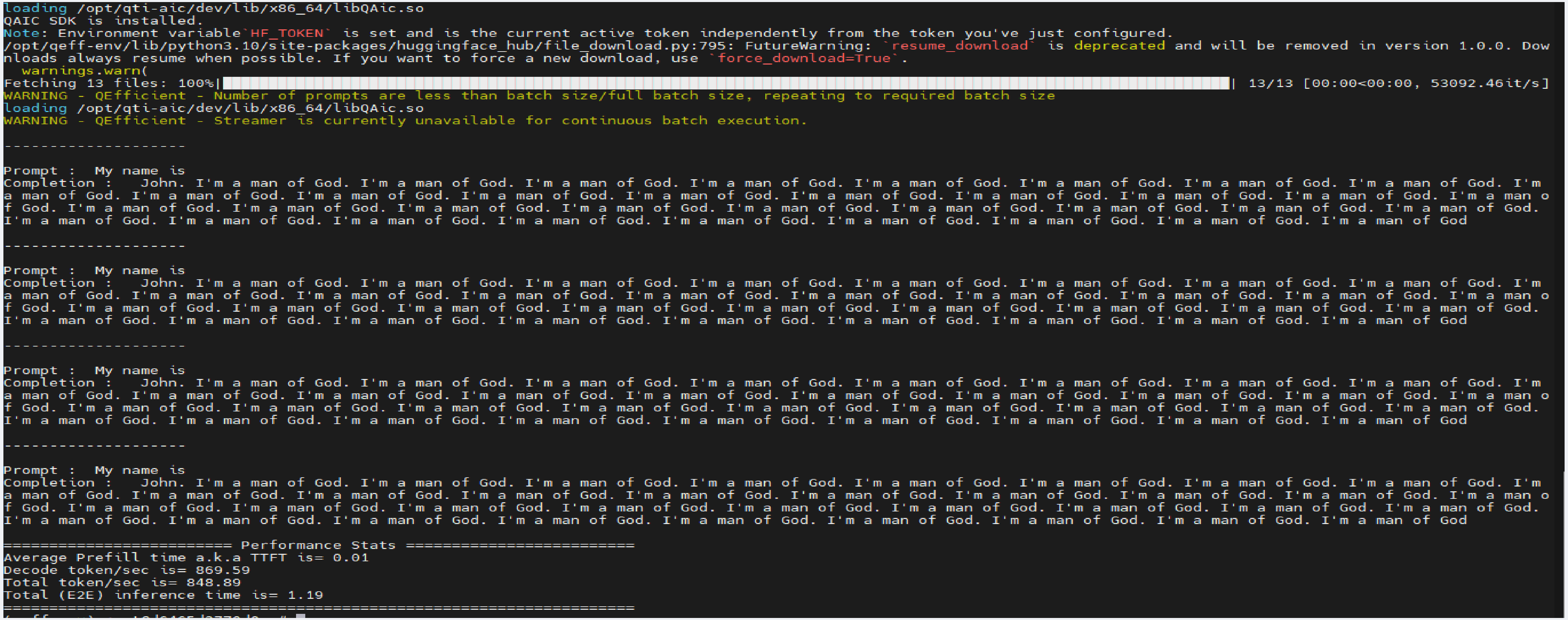


```
(qeff-env) root@d6465d3770d0:~# python -m QEfficient.cloud.compile --onnx_path /home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/onnx_openai_commu
nity_gpt2_with_fbs/openai-community_gpt2_kv.onnx --qpc-path /home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/onnx_openai_community_gpt2_with_fbs/
 --batch_size 1 --ctx_len 256 --mxint8 --num_cores 16 --custom_io_file_path /home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/onnx_openai_communit
y_gpt2_with_fbs/custom_io_int8.yaml --full_batch_size 4 --device_group [0,1,2,3]
loading /opt/qti-aic/dev/lib/x86_64/libQAic.so
QAIC SDK is installed.
Running AI 100 compiler: /opt/qti-aic/exec/qaic-exec -m=/home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/onnx_openai_community_gpt2_with_fbs/open
ai-community_gpt2_kv.onnx -aic-hw -aic-hw-version=2.0 -network-specialization-config=/home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/onnx_openai
_community_gpt2_with_fbs/specializations.json -convert-to-fp16 -retained-state -aic-num-cores=16 -custom-IO-list-file=/home/qraniumtest/gayav/llm_demo/models/qeff_cache/
openai-community/gpt2/onnx_openai_community_gpt2_with_fbs/custom_io_int8.yaml -compile-only -aic-binary-dir=/home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-com
munity/gpt2/onnx_openai_community_gpt2_with_fbs/qpcs -mdp-load-partition-config=/home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/onnx_openai_comm
unity_gpt2_with_fbs/mdp_ts_config.json

===================== Compilation Done! =====================
```

Running LLMs in QAIC with Qefficient

QEfficient.cloud.execute

python -m QEfficient.cloud.execute --model_name openai-community/gpt2 --qpc_path /home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/onnx_openai_community_gpt2_with_fbs/qpcs/ --full_batch_size 4 --cache_dir /home/qraniumtest/gayav/llm_demo/models --hf-token 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 --prompt "My name is" --device_group [0,1,2,3]

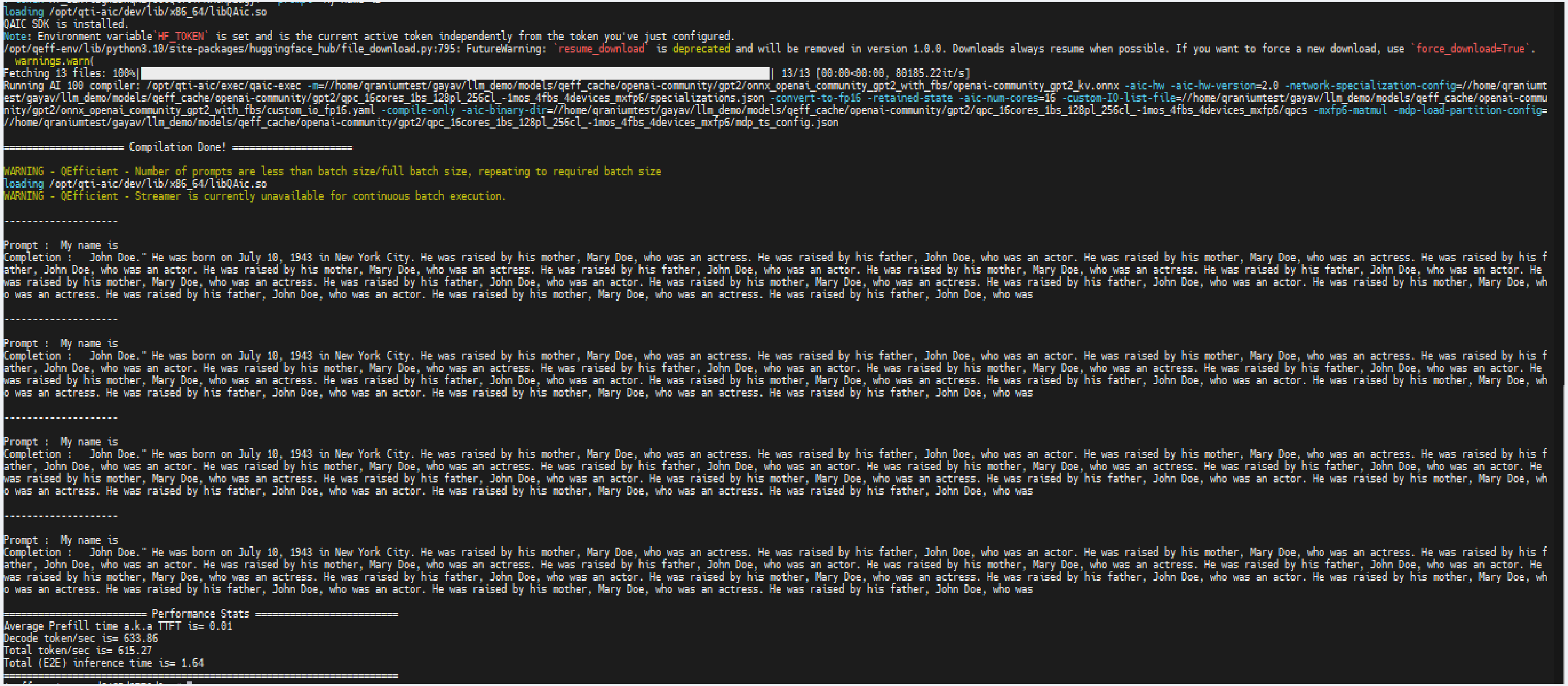


```
loading /opt/qti-aic/dev/lib/x86_64/libQAic.so
QAIC SDK is installed.
Note: Environment variable`HF_TOKEN` is set and is the current active token independently from the token you've just configured.
/opt/qeff-env/lib/python3.10/site-packages/huggingface_hub/file_download.py:795: FutureWarning: `resume_download` is deprecated and will be removed in version 1.0.0. Dow
nloads always resume when possible. If you want to force a new download, use `force_download=True`.
  warnings.warn(
Fetching 13 files: 100%|██████████████████████████████████████████████████████████████████████| 13/13 [00:00<00:00, 53092.46it/s]
WARNING - QEfficient - Number of prompts are less than batch size/full batch size, repeating to required batch size
loading /opt/qti-aic/dev/lib/x86_64/libQAic.so
WARNING - QEfficient - Streamer is currently unavailable for continuous batch execution.

--------------------

Prompt :  My name is
Completion :   John. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm
a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man o
f God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God.
I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God

--------------------

Prompt :  My name is
Completion :   John. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm
a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man o
f God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God.
I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God

--------------------

Prompt :  My name is
Completion :   John. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm
a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man o
f God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God.
I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God

--------------------

Prompt :  My name is
Completion :   John. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm
a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man o
f God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God.
I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God. I'm a man of God

========================= Performance Stats =========================
Average Prefill time a.k.a TTFT is= 0.01
Decode token/sec is= 869.59
Total token/sec is= 848.89
Total (E2E) inference time is= 1.19
=====================================================================
```

Running LLMs in QAIC with Qefficient

QEfficient.cloud.infer

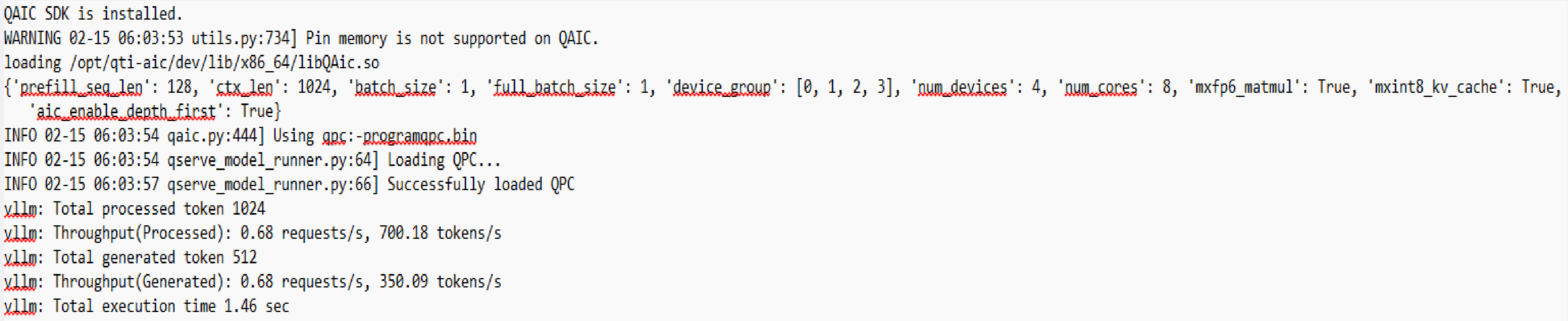
- python -m QEfficient.cloud.infer --model_name openai-community/gpt2 --num_cores 16 --batch_size 1 --prompt_len 128 --ctx_len 256 --mxfp6 --full_batch_size 4 --device_group [0,1,2,3] --cache_dir /home/qraniumtest/gayav/llm_demo/models --hf-token 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 --prompt "My name is"



```
loading /opt/qti-aic/dev/lib/x86_64/libQAic.so
QAIC SDK is installed.
Note: Environment variable`HF_TOKEN` is set and is the current active token independently from the token you've just configured.
/opt/qeff-env/lib/python3.10/site-packages/huggingface_hub/file_download.py:795: FutureWarning: `resume_download` is deprecated and will be removed in version 1.0.0. Downloads always resume when possible. If you want to force a new download, use `force_download=True`.
  warnings.warn(
Fetching 13 files: 100%|████████████████████████████████████████████████████████████| 13/13 [00:00<00:00, 80185.22it/s]
Running AI 100 compiler: /opt/qti-aic/exec/qaic-exec -m=//home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/onnx_openai_community_gpt2_with_fbs/openai-community_gpt2_kv.onnx -aic-hw -aic-hw-version=2.0 -network-specialization-config=//home/qraniumt
est/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/qpc_16cores_1bs_128pl_256cl_-1mos_4fbs_4devices_mxfp6/specializations.json -convert-to-fp16 -retained-state -aic-num-cores=16 -custom-IO-list-file=//home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-commu
nity/gpt2/onnx_openai_community_gpt2_with_fbs/custom_io_fp16.yaml -compile-only -aic-binary-dir=//home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/qpc_16cores_1bs_128pl_256cl_-1mos_4fbs_4devices_mxfp6/qpcs -mxfp6-matmul -mdp-load-partition-config=
//home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/qpc_16cores_1bs_128pl_256cl_-1mos_4fbs_4devices_mxfp6/mdp_ts_config.json

===================== Compilation Done! =====================

WARNING - QEfficient - Number of prompts are less than batch size/full batch size, repeating to required batch size
loading /opt/qti-aic/dev/lib/x86_64/libQAic.so
WARNING - QEfficient - Streamer is currently unavailable for continuous batch execution.

--------------------

Prompt :  My name is
Completion :   John Doe." He was born on July 10, 1943 in New York City. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his f
ather, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He
was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, wh
o was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was

--------------------

Prompt :  My name is
Completion :   John Doe." He was born on July 10, 1943 in New York City. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his f
ather, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He
was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, wh
o was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was

--------------------

Prompt :  My name is
Completion :   John Doe." He was born on July 10, 1943 in New York City. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his f
ather, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He
was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, wh
o was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was

--------------------

Prompt :  My name is
Completion :   John Doe." He was born on July 10, 1943 in New York City. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his f
ather, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He
was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, wh
o was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was

========================== Performance Stats ==========================
Average Prefill time a.k.a TTFT is= 0.01
Decode token/sec is= 633.86
Total token/sec is= 615.27
Total (E2E) inference time is= 1.64
=======================================================================
```

VLLM Serving platform

- Documentation link- VLLM - RadiusSW - Qualcomm Confluence
- vLLM is a fast and easy-to-use library for LLM inference and serving.
- It supports many features that help in serving larger number of users.
- **Example:** python benchmarks/benchmark_throughput.py --max-num-seqs 4 --max-seq_len-to-capture 128 --device qaic --max-model-len 256 --num-prompts 100 --quantization mxfp6 --backend vllm --dataset benchmarks/ShareGPT_V3_unfiltered_cleaned_split.json --model meta-llama/Meta-Llama-3-70B --input-len 128 --seed 20 --temperature 0.0 --device-group 1,2,3,4
  - bs : max-num-seqs
  - cpl : max-seq_len-to-capture
  - cl : max-model-len
  - pl : input-len



```
QAIC SDK is installed.
WARNING 02-15 06:03:53 utils.py:734] Pin memory is not supported on QAIC.
loading /opt/qti-aic/dev/lib/x86_64/libQAic.so
{'prefill_seq_len': 128, 'ctx_len': 1024, 'batch_size': 1, 'full_batch_size': 1, 'device_group': [0, 1, 2, 3], 'num_devices': 4, 'num_cores': 8, 'mxfp6_matmul': True, 'mxint8_kv_cache': True,
   'aic_enable_depth_first': True}
INFO 02-15 06:03:54 qaic.py:444] Using qpc:-programqpc.bin
INFO 02-15 06:03:54 qserve_model_runner.py:64] Loading QPC...
INFO 02-15 06:03:57 qserve_model_runner.py:66] Successfully loaded QPC
vllm: Total processed token 1024
vllm: Throughput(Processed): 0.68 requests/s, 700.18 tokens/s
vllm: Total generated token 512
vllm: Throughput(Generated): 0.68 requests/s, 350.09 tokens/s
vllm: Total execution time 1.46 sec
```

- More reading - Welcome to vLLM — vLLM
- qranium/vllm - Gitiles

Agenda

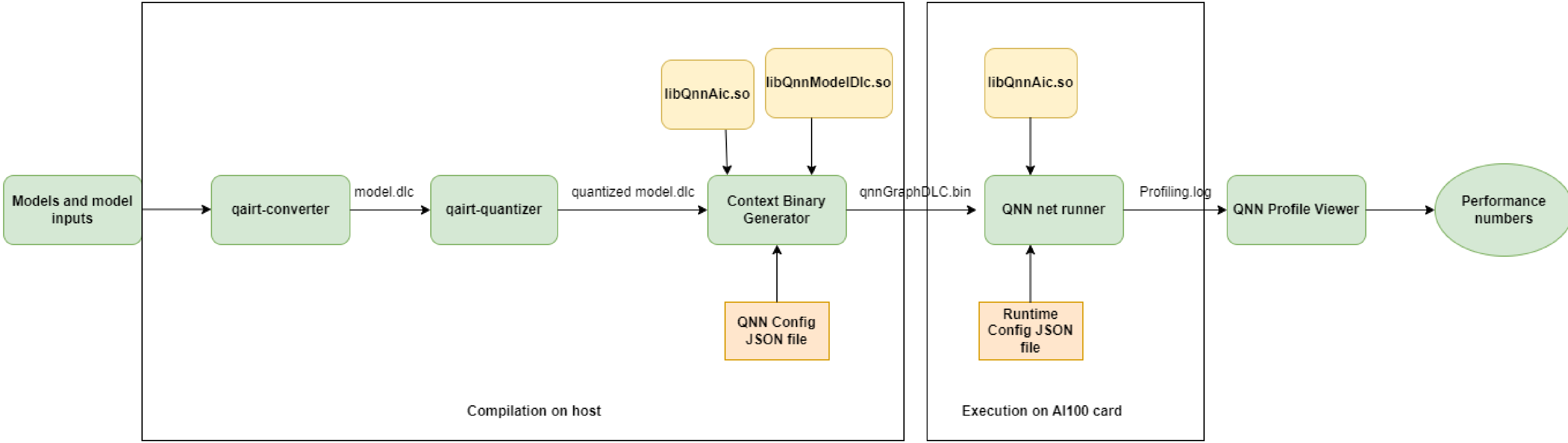
Day: 3

- QNN-AIC Apps
- Compilation & Execution

QNN Apps Overview

QNN Workflow

- The trained models are passed to the converter along with the Op package definition files.
- Op Packages are a collection of operations that are made available to a backend in order to be utilized in creating and executing QNN graphs representing network models.
- The output of converter and quantizer is DLC (a converted graph) which is given to the context binary stage (similar to exec) and this stage produces the compiled binaries.
- The qnn net runner takes these binaries and produces the outputs and profiling data which are then rendered using the qnn-profile viewer tool.



Docker Setup

Download Docker Open-Source Image from Qualcomm Docker registry - [qraniumtest/qranium · Quay]

- Login to Qualcomm Docker registry

```
sudo docker login -u="\$app" -p="HZBWP5R45MMYLJOG6IMM6HGIH8KFVP5IX861XG9UG09WVK5V7ZF7S5T42XITW2FRTC1IMRLC5W6LKALHFQ65S6DXPJZPKLZC7N3KXNR1FSD9QNN6C9M839VG" docker-registry.qualcomm.com
```

- Pull Latest QNN Docker Image.

```
sudo docker pull docker-registry.qualcomm.com/qraniumtest/qranium:QNN-1.19.1.21-ubuntu22-x86_64
```

- Create a container

```
sudo docker run --privileged -dit -v /home/qraniumtest:/home/qraniumtest --name mlg-dev-22.04_root docker-registry.qualcomm.com/qraniumtest/qranium:QNN-1.19.1.21-ubuntu22-x86_64
```

- Go into the container

```
sudo docker exec -it mlg-dev-22.04_root bash
```

- Run the following script to check and install missing dependencies.

```
source /opt/venv_py310/bin/activate
python3 -m pip install --upgrade pip
${QNN_SDK_ROOT}/bin/check-python-dependency
```

Compiling and running a model in QNN (Non-LLM)

CASE 1: FP16 PRECISION

CONVERTER

```
$QNN_SDK_ROOT/bin/x86_64-linux-clang/qairt-converter --input_network /home/qraniumtest/model_zoo/customer/MSFT/Big_Bird/generatedModels/ms_config/BigBird_bs_2048_msconfig_64blk_blockwiseattn_sim.onnx --output_path /home/qraniumtest/binaries/BigBird/model.dlc --float_bitwidth 16 --float_bias_bitwidth 32 --preserve_io_datatype  --onnx_define_symbol batch_size 1 --onnx_skip_simplification  --onnx_defer_loading
```

CONTEXT BINARY

```
$QNN_SDK_ROOT/bin/x86_64-linux-clang/qnn-context-binary-generator --binary_file qnngraph.serialized --backend $QNN_SDK_ROOT/lib/x86_64-linux-clang/libQnnAic.so --output_dir /home/qraniumtest/binaries/BigBird/ --config_file /home/qraniumtest/binaries/BigBird/qnn_config.json --log_level debug  --backend_binary /home/qraniumtest/binaries/BigBird/programqpc_dir/programqpc.bin --model $QNN_SDK_ROOT/lib/x86_64-linux-clang/libQnnModelDlc.so --dlc_path /home/qraniumtest/binaries/BigBird/model.dlc
```

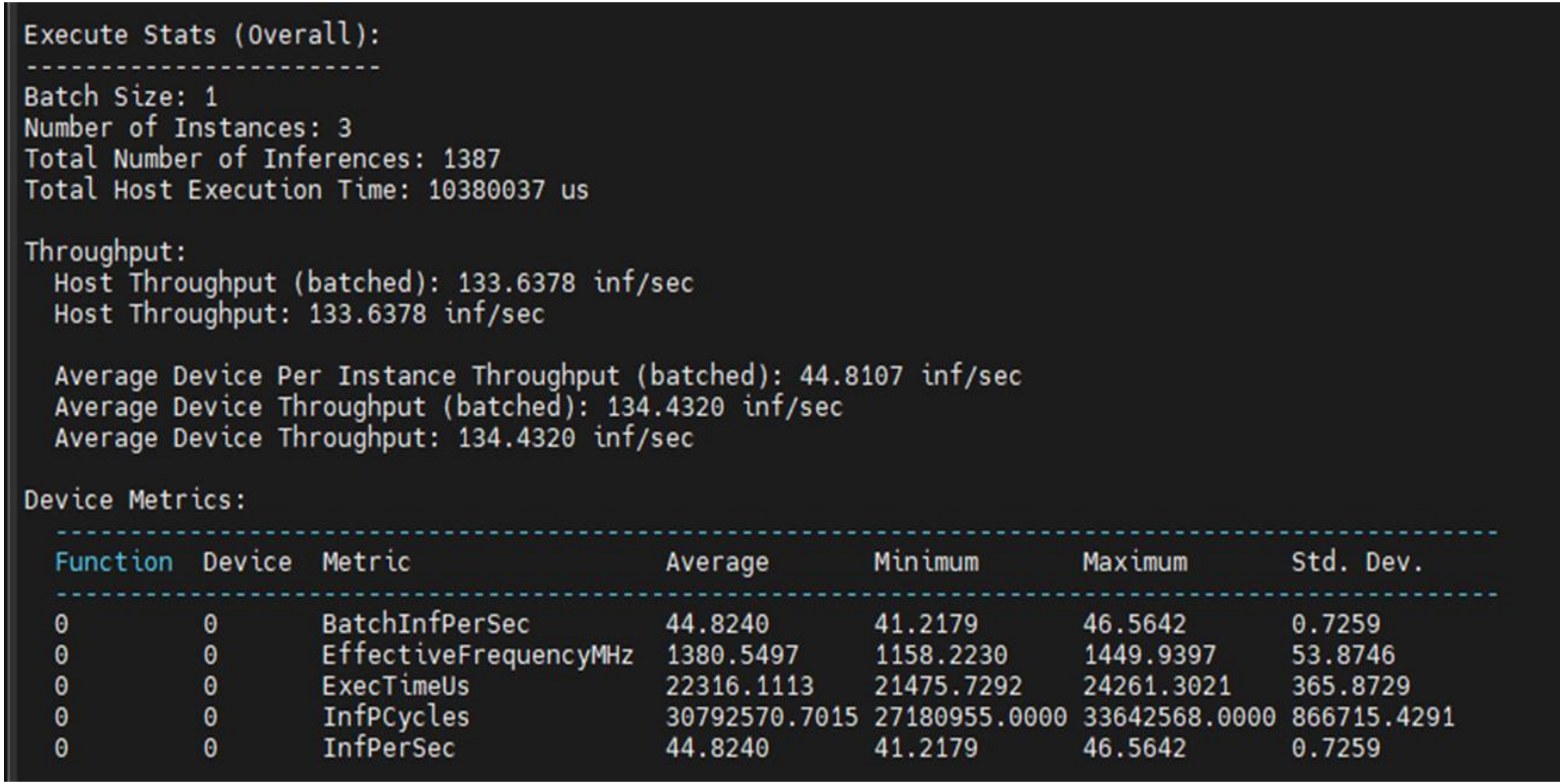
NET RUNNER

```
cd /home/qraniumtest/binaries/BigBird/ && $QNN_SDK_ROOT/bin/x86_64-linux-clang/qnn-net-run --backend $QNN_SDK_ROOT/lib/x86_64-linux-clang/libQnnAic.so --input_list /home/qraniumtest/binaries/BigBird/qnn_list.txt --retrieve_context qnngraph.serialized.bin --log_level info --profiling_level basic --config_file /home/qraniumtest/binaries/BigBird/qnn_net_runner_config.json  --duration 10 --keep_num_outputs 2  --use_native_input_files
```

PROFILE VIEWER

```
$QNN_SDK_ROOT/bin/x86_64-linux-clang/qnn-profile-viewer --input_log output/qnn-profiling-data.log --reader $QNN_SDK_ROOT/lib/x86_64-linux-clang/libQnnAicProfilingReader.so
```

Compiling and running a model in QNN (Non-LLM)



```
Execute Stats (Overall):
------------------------
Batch Size: 1
Number of Instances: 3
Total Number of Inferences: 1387
Total Host Execution Time: 10380037 us

Throughput:
  Host Throughput (batched): 133.6378 inf/sec
  Host Throughput: 133.6378 inf/sec

  Average Device Per Instance Throughput (batched): 44.8107 inf/sec
  Average Device Throughput (batched): 134.4320 inf/sec
  Average Device Throughput: 134.4320 inf/sec

Device Metrics:
  -------------------------------------------------------------------------------------------------
  Function  Device  Metric                 Average       Minimum       Maximum       Std. Dev.
  -------------------------------------------------------------------------------------------------
  0         0       BatchInfPerSec         44.8240       41.2179       46.5642       0.7259
  0         0       EffectiveFrequencyMHz  1380.5497     1158.2230     1449.9397     53.8746
  0         0       ExecTimeUs             22316.1113    21475.7292    24261.3021    365.8729
  0         0       InfPCycles             30792570.7015 27180955.0000 33642568.0000 866715.4291
  0         0       InfPerSec              44.8240       41.2179       46.5642       0.7259
```

Compiling and running a model in QNN (Non-LLM)

CASE 3: INT8

CONVERTER

-
```
$QNN_SDK_ROOT/bin/x86_64-linux-clang/qairt-converter --input_network /home/qraniumtest/model_zoo//internal/DenseNet169/generatedModels/ONNX/densenet169.onnx --output_path /home/qraniumtest/binaries/DenseNet169/model.dlc --float_bias_bitwidth 32 --preserve_io_datatype  --onnx_batch 8  --onnx_skip_simplification  --onnx_defer_loading
```

QUANTIZER

```
$QNN_SDK_ROOT/bin/x86_64-linux-clang/qairt-quantizer --input_dlc /home/qraniumtest/binaries/DenseNet169/model.dlc --output_dlc /home/qraniumtest/binaries/DenseNet169/model_quantized.dlc  --preserve_io_datatype --use_native_input_files  --input_list /home/qraniumtest/binaries/DenseNet169/qnn_list.txt   --act_quantizer_schema unsignedsymmetric --param_quantizer_schema unsignedsymmetric
```

CONTEXT BINARY

```
$QNN_SDK_ROOT/bin/x86_64-linux-clang/qnn-context-binary-generator --binary_file qnngraph.serialized --backend $QNN_SDK_ROOT/lib/x86_64-linux-clang/libQnnAic.so --output_dir /home/qraniumtest/binaries/DenseNet169 --config_file /home/qraniumtest/binaries/DenseNet169/qnn_config.json --log_level debug  --backend_binary /home/qraniumtest/binaries/DenseNet169/programqpc_dir/programqpc.bin --model $QNN_SDK_ROOT//lib/x86_64-linux-clang/libQnnModelDlc.so --dlc_path /home/qraniumtest/binaries/DenseNet169/model_quantized.dlc
```

Compiling and running a model in QNN (Non-LLM)

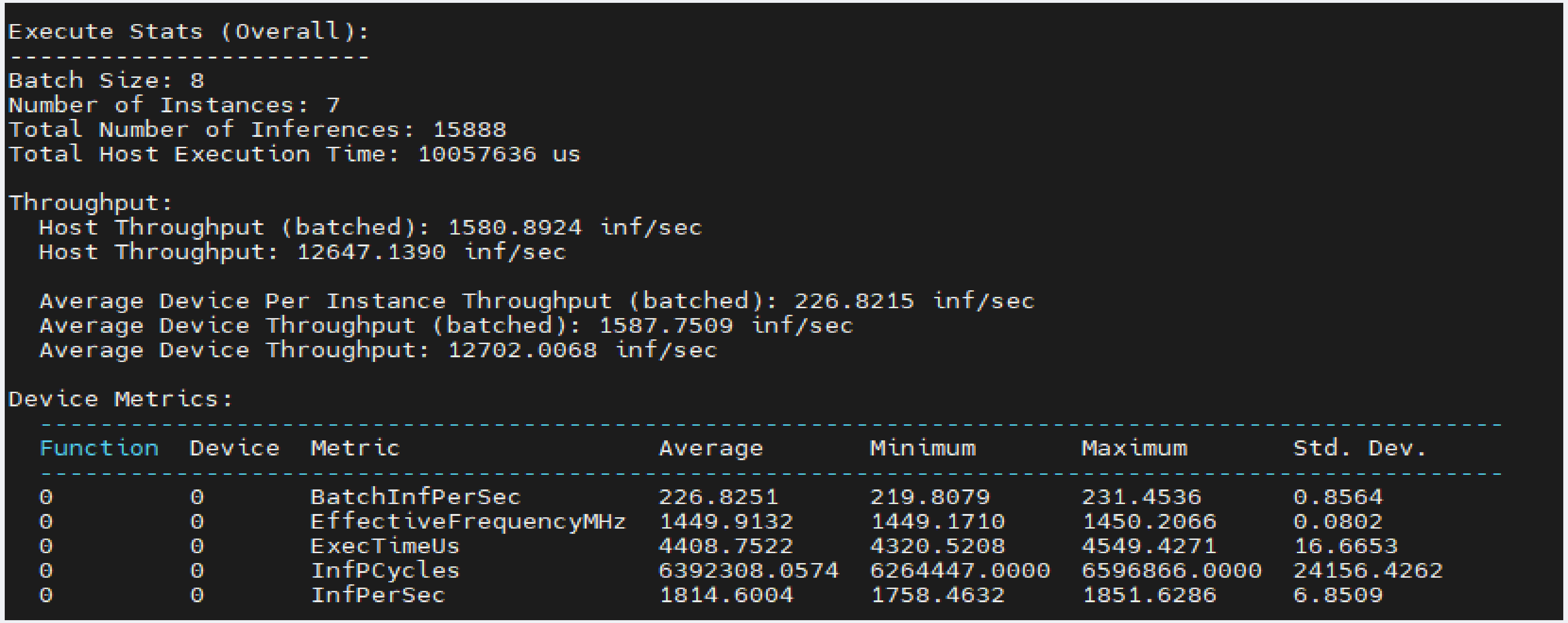
CASE 3: INT8

NET RUNNER

- cd /home/qraniumtest/binaries/DenseNet169 && $QNN_SDK_ROOT//bin/x86_64-linux-clang/qnn-net-run --backend $QNN_SDK_ROOT//lib/x86_64-linux-clang/libQnnAic.so --input_list /home/qraniumtest/binaries/DenseNet169/qnn_list.txt --retrieve_context qnngraph.serialized.bin --log_level info --profiling_level basic --config_file /home/qraniumtest/binaries/DenseNet169/qnn_net_runner_config.json --duration 10 --keep_num_outputs 2 --use_native_input_files

- PROFILE VIEWER

$QNN_SDK_ROOT//bin/x86_64-linux-clang/qnn-profile-viewer --input_log output/qnn-profiling-data.log --reader $QNN_SDK_ROOT//lib/x86_64-linux-clang/libQnnAicProfilingReader.so



```
Execute Stats (Overall):
------------------------
Batch Size: 8
Number of Instances: 7
Total Number of Inferences: 15888
Total Host Execution Time: 10057636 us

Throughput:
  Host Throughput (batched): 1580.8924 inf/sec
  Host Throughput: 12647.1390 inf/sec

  Average Device Per Instance Throughput (batched): 226.8215 inf/sec
  Average Device Throughput (batched): 1587.7509 inf/sec
  Average Device Throughput: 12702.0068 inf/sec

Device Metrics:
  -------------------------------------------------------------------------------------------------
  Function  Device  Metric                  Average       Minimum       Maximum       Std. Dev.
  -------------------------------------------------------------------------------------------------
  0         0       BatchInfPerSec          226.8251      219.8079      231.4536      0.8564
  0         0       EffectiveFrequencyMHz   1449.9132     1449.1710     1450.2066     0.0802
  0         0       ExecTimeUs              4408.7522     4320.5208     4549.4271     16.6653
  0         0       InfPCycles              6392308.0574  6264447.0000  6596866.0000  24156.4262
  0         0       InfPerSec               1814.6004     1758.4632     1851.6286     6.8509
```

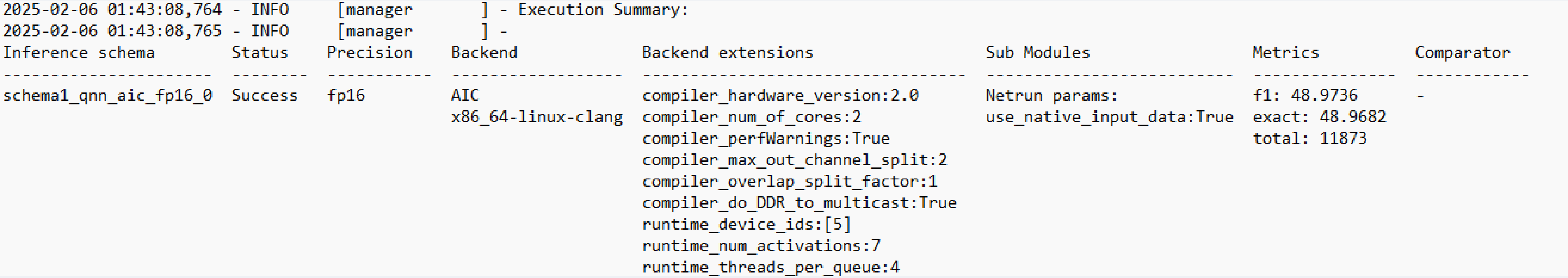
QNN Accuracy Evaluator

BigBird – FP16 precision

- **Command used –**

```
docker exec mlg-dev-22.04_qraniumtest bash -c "source /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/qnn_qairt_accuracy/venv/bin/activate; source /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/aic_perf/qraft/src/tools/qairt_acc/scripts/export_qnn_variables.sh /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/qnn-aic-sdk-V2.31.0-RC5/ /local/mnt/workspace/jyashwan/model_zoo; /usr/bin/time -v /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/qnn-aic-sdk-V2.31.0-RC5//bin/x86_64-linux-clang/qairt-accuracy-evaluator -config /local/mnt/workspace/accuracy//qnn_new_model_configs/customers/MSFT/bigbird/bigbird_config.yaml  -cleanup intermediate -inference_schema_tag qnn_fp16 -work_dir  /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/qnn_qairt_accuracy/BigBird/fp16/acc_log -silent -device_id 5 "
```

- **Metrics -**



```
2025-02-06 01:43:08,764 - INFO     [manager        ] - Execution Summary:
2025-02-06 01:43:08,765 - INFO     [manager        ] -
Inference schema        Status    Precision    Backend             Backend extensions                  Sub Modules                 Metrics          Comparator
----------------------  --------  -----------  ------------------  ----------------------------------  --------------------------  ---------------  ------------
schema1_qnn_aic_fp16_0  Success   fp16         AIC                 compiler_hardware_version:2.0       Netrun params:              f1: 48.9736      -
                                               x86_64-linux-clang  compiler_num_of_cores:2             use_native_input_data:True  exact: 48.9682
                                                                   compiler_perfWarnings:True                                      total: 11873
                                                                   compiler_max_out_channel_split:2
                                                                   compiler_overlap_split_factor:1
                                                                   compiler_do_DDR_to_multicast:True
                                                                   runtime_device_ids:[5]
                                                                   runtime_num_activations:7
                                                                   runtime_threads_per_queue:4
```

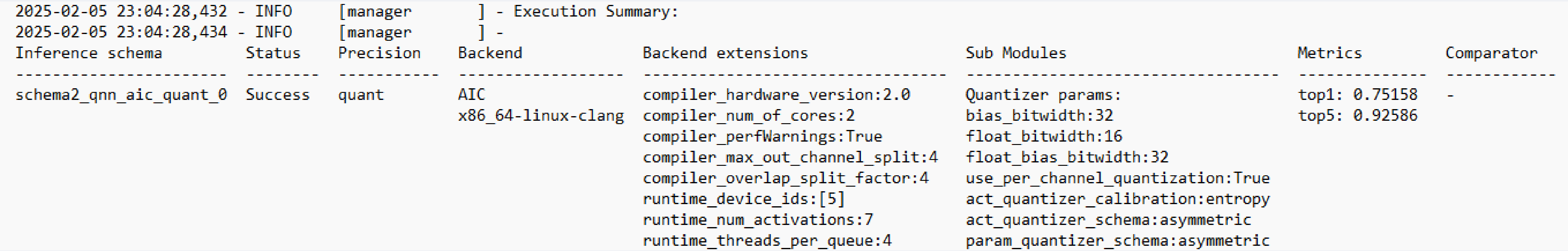
QNN Accuracy Evaluator

DenseNet169 – INT8 precision

- **Command used –**

docker exec mlg-dev-22.04_qraniumtest bash -c "source /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/qnn_qairt_accuracy/venv/bin/activate; source /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/aic_perf/qraft/src/tools/qairt_acc/scripts/export_qnn_variables.sh /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/qnn-aic-sdk-V2.31.0-RC5/ /local/mnt/workspace/jyashwan/model_zoo; /usr/bin/time -v /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/qnn-aic-sdk-V2.31.0-RC5//bin/x86_64-linux-clang/**qairt-accuracy-evaluator** -config /local/mnt/workspace/accuracy//qnn_new_model_configs/public/densenet169/densenet169_config.yaml -cleanup intermediate -inference_schema_tag qnn_int8 -work_dir /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/qnn_qairt_accuracy/DenseNet169/int8/acc_log -silent -device_id 5 “

Metrics -



```
2025-02-05 23:04:28,432 - INFO     [manager       ] - Execution Summary:
2025-02-05 23:04:28,434 - INFO     [manager       ] -
Inference schema         Status    Precision    Backend             Backend extensions                 Sub Modules                         Metrics         Comparator
-----------------------  --------  -----------  ------------------  ---------------------------------  ----------------------------------  --------------  ------------
schema2_qnn_aic_quant_0  Success   quant        AIC                 compiler_hardware_version:2.0      Quantizer params:                   top1: 0.75158   -
                                                x86_64-linux-clang  compiler_num_of_cores:2            bias_bitwidth:32                    top5: 0.92586
                                                                    compiler_perfWarnings:True         float_bitwidth:16
                                                                    compiler_max_out_channel_split:4   float_bias_bitwidth:32
                                                                    compiler_overlap_split_factor:4    use_per_channel_quantization:True
                                                                    runtime_device_ids:[5]             act_quantizer_calibration:entropy
                                                                    runtime_num_activations:7          act_quantizer_schema:asymmetric
                                                                    runtime_threads_per_queue:4        param_quantizer_schema:asymmetric
```

Agenda

Day: 4

- QNN-AIC LLM
- Automation Tools and Dashboards

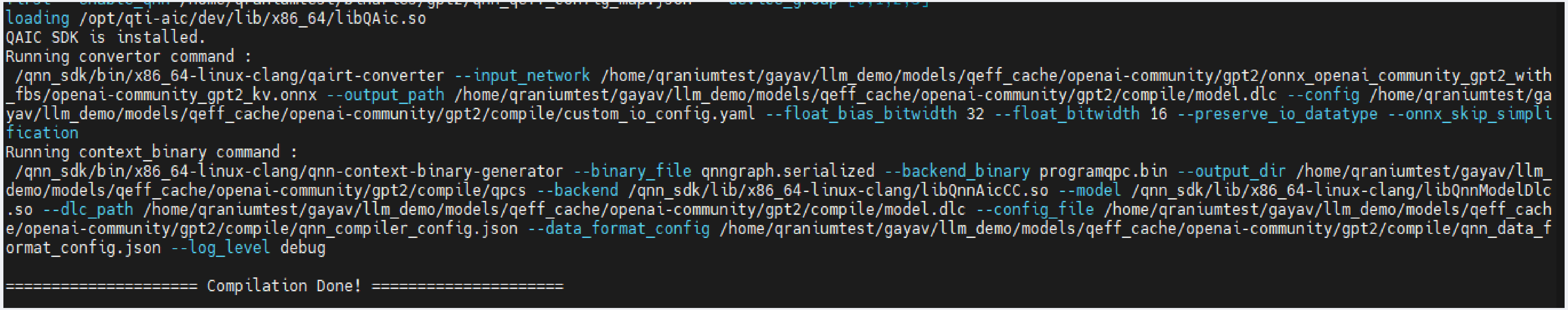
Running LLMs in QNN using Qefficient

- Setup the docker environment.
- Clone Qefficient repo using git clone --branch release/v1.19 https://github.com/quic/efficient-transformers
- Create a virtual env using python3.10 -m venv qeff-env
- Activate the env source qeff-env/bin/activate
- Install qefficient dependencies – cd efficient-transformers; python3 -m pip install -e .
- export HF_HOME=/home/qraniumtest/gayav/llm_demo/models/
- export QEFF_HOME=/home/qraniumtest/gayav/llm_demo/models/
- export HF_TOKEN="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“

Running LLMs in QNN using Qefficient

QEfficient.cloud.compile

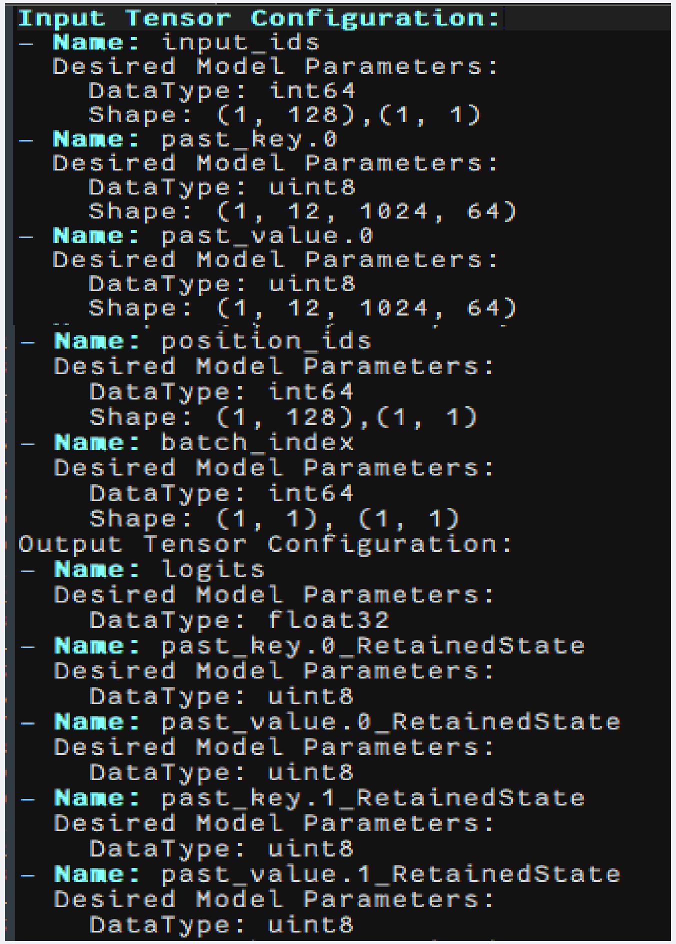
```
python -m QEfficient.cloud.compile --onnx_path /home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/onnx_openai_community_gpt2_with_fbs/openai-community_gpt2_kv.onnx --qpc_path /home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/compile/ --num_cores 16 --full_batch_size 1 --prompt_len 128 --ctx_len 1024 --mxfp6 --mxint8 --allow-mxint8-mdp-io --aic_enable_depth_first --enable_qnn /home/qraniumtest/binaries/gpt2/qnn_qeff_config_map.json --device_group [0,1,2,3]
```



```
loading /opt/qti-aic/dev/lib/x86_64/libQAic.so
QAIC SDK is installed.
Running convertor command :
 /qnn_sdk/bin/x86_64-linux-clang/qairt-converter --input_network /home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/onnx_openai_community_gpt2_with
_fbs/openai-community_gpt2_kv.onnx --output_path /home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/compile/model.dlc --config /home/qraniumtest/ga
yav/llm_demo/models/qeff_cache/openai-community/gpt2/compile/custom_io_config.yaml --float_bias_bitwidth 32 --float_bitwidth 16 --preserve_io_datatype --onnx_skip_simpli
fication
Running context_binary command :
 /qnn_sdk/bin/x86_64-linux-clang/qnn-context-binary-generator --binary_file qnngraph.serialized --backend_binary programqpc.bin --output_dir /home/qraniumtest/gayav/llm_
demo/models/qeff_cache/openai-community/gpt2/compile/qpcs --backend /qnn_sdk/lib/x86_64-linux-clang/libQnnAicCC.so --model /qnn_sdk/lib/x86_64-linux-clang/libQnnModelDlc
.so --dlc_path /home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/compile/model.dlc --config_file /home/qraniumtest/gayav/llm_demo/models/qeff_cach
e/openai-community/gpt2/compile/qnn_compiler_config.json --data_format_config /home/qraniumtest/gayav/llm_demo/models/qeff_cache/openai-community/gpt2/compile/qnn_data_f
ormat_config.json --log_level debug

===================== Compilation Done! =====================
```

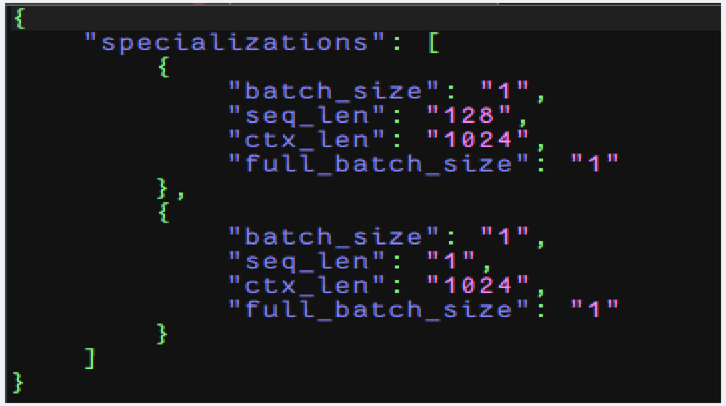
Network Specialization – QNN & QAIC

QNN

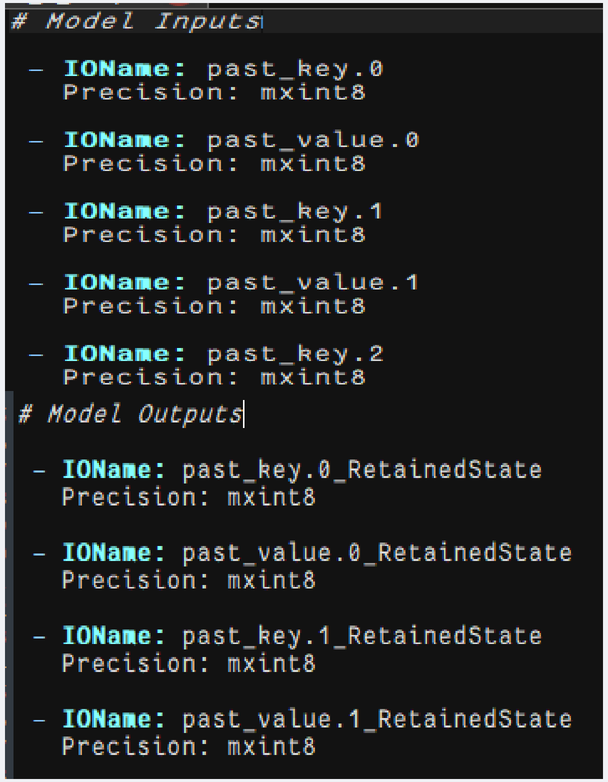


```
Input Tensor Configuration:
– Name: input_ids
  Desired Model Parameters:
    DataType: int64
    Shape: (1, 128),(1, 1)
– Name: past_key.0
  Desired Model Parameters:
    DataType: uint8
    Shape: (1, 12, 1024, 64)
– Name: past_value.0
  Desired Model Parameters:
    DataType: uint8
    Shape: (1, 12, 1024, 64)
– Name: position_ids
  Desired Model Parameters:
    DataType: int64
    Shape: (1, 128),(1, 1)
– Name: batch_index
  Desired Model Parameters:
    DataType: int64
    Shape: (1, 1), (1, 1)
Output Tensor Configuration:
– Name: logits
  Desired Model Parameters:
    DataType: float32
– Name: past_key.0_RetainedState
  Desired Model Parameters:
    DataType: uint8
– Name: past_value.0_RetainedState
  Desired Model Parameters:
    DataType: uint8
– Name: past_key.1_RetainedState
  Desired Model Parameters:
    DataType: uint8
– Name: past_value.1_RetainedState
  Desired Model Parameters:
    DataType: uint8
```

QAIC



```
{
    "specializations": [
        {
            "batch_size": "1",
            "seq_len": "128",
            "ctx_len": "1024",
            "full_batch_size": "1"
        },
        {
            "batch_size": "1",
            "seq_len": "1",
            "ctx_len": "1024",
            "full_batch_size": "1"
        }
    ]
}
```

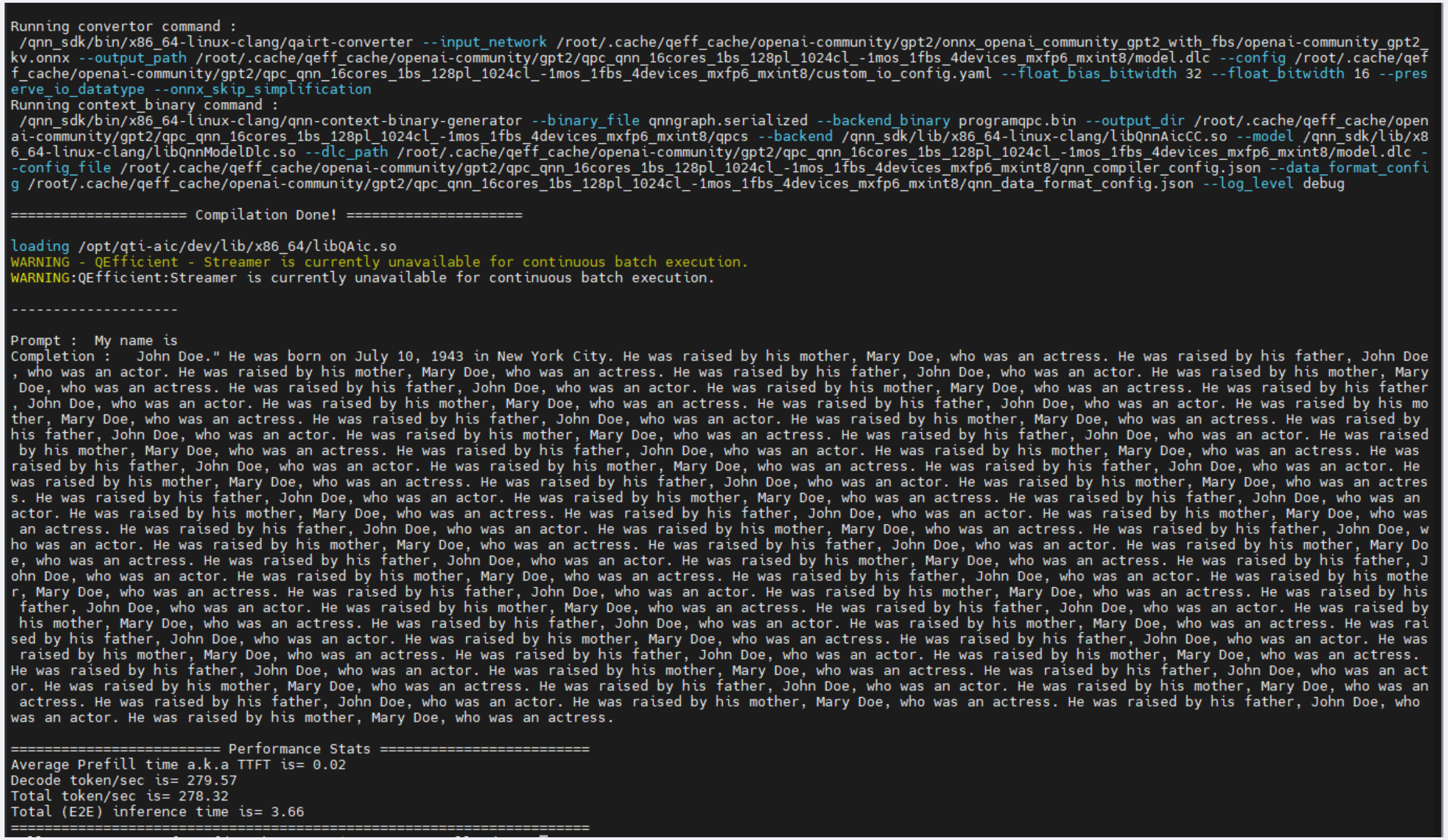


```
# Model Inputs

 – IOName: past_key.0
   Precision: mxint8

 – IOName: past_value.0
   Precision: mxint8

 – IOName: past_key.1
   Precision: mxint8

 – IOName: past_value.1
   Precision: mxint8

 – IOName: past_key.2
   Precision: mxint8
# Model Outputs

 – IOName: past_key.0_RetainedState
   Precision: mxint8

 – IOName: past_value.0_RetainedState
   Precision: mxint8

 – IOName: past_key.1_RetainedState
   Precision: mxint8

 – IOName: past_value.1_RetainedState
   Precision: mxint8
```

Running LLMs in QNN using Qefficient

QEfficient.cloud.infer

- python -m QEfficient.cloud.infer --model-name openai-community/gpt2 --full_batch_size 1 --prompt_len 128 --ctx_len 1024 --mxfp6 --mxint8 --num_cores 16 --device_group [0,1,2,3] --prompt "My name is" --allow-mxint8-mdp-io --aic_enable_depth_first --enable_qnn /home/qraniumtest/binaries/gpt2/qnn_qeff_config_map.json



```
Running convertor command :
 /qnn_sdk/bin/x86_64-linux-clang/qairt-converter --input_network /root/.cache/qeff_cache/openai-community/gpt2/onnx_openai_community_gpt2_with_fbs/openai-community_gpt2_
kv.onnx --output_path /root/.cache/qeff_cache/openai-community/gpt2/qpc_qnn_16cores_1bs_128pl_1024cl_-1mos_1fbs_4devices_mxfp6_mxint8/model.dlc --config /root/.cache/qef
f_cache/openai-community/gpt2/qpc_qnn_16cores_1bs_128pl_1024cl_-1mos_1fbs_4devices_mxfp6_mxint8/custom_io_config.yaml --float_bias_bitwidth 32 --float_bitwidth 16 --pres
erve_io_datatype --onnx_skip_simplification
Running context_binary command :
 /qnn_sdk/bin/x86_64-linux-clang/qnn-context-binary-generator --binary_file qnngraph.serialized --backend_binary programqpc.bin --output_dir /root/.cache/qeff_cache/open
ai-community/gpt2/qpc_qnn_16cores_1bs_128pl_1024cl_-1mos_1fbs_4devices_mxfp6_mxint8/qpcs --backend /qnn_sdk/lib/x86_64-linux-clang/libQnnAicCC.so --model /qnn_sdk/lib/x8
6_64-linux-clang/libQnnModelDlc.so --dlc_path /root/.cache/qeff_cache/openai-community/gpt2/qpc_qnn_16cores_1bs_128pl_1024cl_-1mos_1fbs_4devices_mxfp6_mxint8/model.dlc -
-config_file /root/.cache/qeff_cache/openai-community/gpt2/qpc_qnn_16cores_1bs_128pl_1024cl_-1mos_1fbs_4devices_mxfp6_mxint8/qnn_compiler_config.json --data_format_confi
g /root/.cache/qeff_cache/openai-community/gpt2/qpc_qnn_16cores_1bs_128pl_1024cl_-1mos_1fbs_4devices_mxfp6_mxint8/qnn_data_format_config.json --log_level debug

===================== Compilation Done! =====================

loading /opt/qti-aic/dev/lib/x86_64/libQAic.so
WARNING - QEfficient - Streamer is currently unavailable for continuous batch execution.
WARNING:QEfficient:Streamer is currently unavailable for continuous batch execution.

--------------------

Prompt :  My name is
Completion :   John Doe." He was born on July 10, 1943 in New York City. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe
, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary
 Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father
, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mo
ther, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by
his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised
 by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was
raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He
was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actres
s. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an
actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was
 an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, w
ho was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Do
e, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, J
ohn Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mothe
r, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his
 father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by
 his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was rai
sed by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was
 raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress.
He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an act
or. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an
 actress. He was raised by his father, John Doe, who was an actor. He was raised by his mother, Mary Doe, who was an actress. He was raised by his father, John Doe, who
was an actor. He was raised by his mother, Mary Doe, who was an actress.

========================= Performance Stats =========================
Average Prefill time a.k.a TTFT is= 0.02
Decode token/sec is= 279.57
Total token/sec is= 278.32
Total (E2E) inference time is= 3.66
=====================================================================
```

Automation Support

Source (part of aic_perf repo) – [qraft - qranium/aic_perf - Gitiles]

Command usage –

Sample QAIC Command

python3 run_inference.py aic-perf --config </path/to/config.csv> --work_dir </path/to/workdir> -r 6 --device_ids 1 --execute --num_paral_comp 1

The above command runs aic-perf (qaic performance) module, from given config file, with a given work dir, with row number as 6. Here parallel compilation is disable, device id chosen as qid 1, and only execute is being performed.

Sample QNN Command

python3 run_inference.py qnn-qairt-aic-perf --config </path/to/config.csv> --enable_docker --work_dir </path/to/workdir> --qnn_sdk_path </path/to/qnn_sdk> --docker_container_name <docker_name> --venv_python "python3.10" -r 6 --device_ids 1 --execute --compile

Here, we are running qnn performance module, with docker flags to run the commands inside docker and inside python3.10 virtual env. QNN sdk path is mandatory for qnn testing. Here end to end execution is performed.

REFERENCES

- ➢Inference workflow on Cloud AI - Cloud AI 100
- ➢ Qualcomm Documentation
- ➢GitHub - quic/efficient-transformers
- ➢QNN AIC hands-on - Qranium - Qualcomm Confluence
- ➢QNN and QAIC Options Mapping - Qranium - Qualcomm Confluence
- ➢QEfficient for QNN-AIC - Qranium - Qualcomm Confluence
- ➢VLLM via QNN Compilation - Qranium - Qualcomm Confluence

BACKUP

Compiling and running a model in QAIC (Non-LLM)

CASE 2: MIXED PRECISON

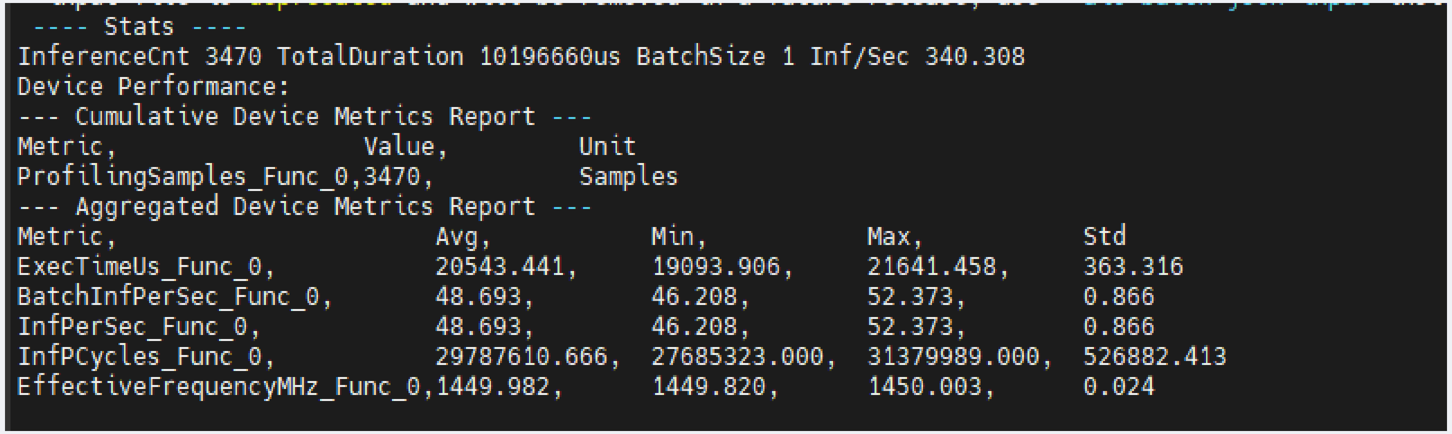
➢ Compile command:

```
/opt/qti-aic/exec/qaic-exec -dump-profile=/home/qraniumtest/pgq_profiles/BERT_Large_Packed_Boolean_Mask_PGQ.yaml -m=/home/qraniumtest/model_zoo/MLPerfModels/BertLarge/generatedModels/ONNX/BERT_MLCommons_Flexible_BS_SL_Packed_BoolMask.onnx  -onnx-define-symbol=batch_size,1 -onnx-define-symbol=seg_length,384 -input-list-file=/home/qraniumtest/model_zoo/model-inputs/inputs/Bert/SL-384/batch_size_1/file-list.txt
```

```
/opt/qti-aic/exec/qaic-exec -aic-num-cores=2 -mos=2 -ols=1 -load-profile=/home/qraniumtest/pgq_profiles/BERT_Large_Packed_Boolean_Mask_PGQ.yaml -m=/home/qraniumtest/model_zoo/MLPerfModels/BertLarge/generatedModels/ONNX/BERT_MLCommons_Flexible_BS_SL_Packed_BoolMask.onnx -input-list-file=/home/qraniumtest/model_zoo/model-inputs/inputs/Bert/SL-384/batch_size_1/file-list.txt -aic-binary-dir=/home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/aic/ -aic-hw -aic-hw-version=2.0 -quantization-schema-activations=symmetric_with_uint8 -quantization-schema-constants=symmetric_with_uint8 -quantization-precision=Int8 -stats-batchsize=1 -onnx-define-symbol=batch_size,1 -onnx-define-symbol=seg_length,384 -aic-perf-metrics -aic-perf-warnings -stats-level=40 -node-precision-info=/home/qraniumtest/model_zoo/MLPerfModels/BertLarge/bert_packing_strategy_boolean_mask_node_precision_file.yaml -compile-only
```

➢ qaic-runner command:

```
cd /home/qraniumtest/model_zoo/model-inputs/inputs/Bert/SL-384/batch_size_1 &&  /opt/qti-aic/exec/qaic-runner -t /home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/aic/ -i ./input_ids.raw -i ./input_mask.raw -i ./segment_ids.raw -i ./input_position_ids.raw -T 4 --time 10 -a 7 -d 5
```



```
 ---- Stats ----
InferenceCnt 3470 TotalDuration 10196660us BatchSize 1 Inf/Sec 340.308
Device Performance:
--- Cumulative Device Metrics Report ---
Metric,                  Value,         Unit
ProfilingSamples_Func_0,3470,           Samples
--- Aggregated Device Metrics Report ---
Metric,                       Avg,           Min,           Max,           Std
ExecTimeUs_Func_0,            20543.441,     19093.906,     21641.458,     363.316
BatchInfPerSec_Func_0,        48.693,        46.208,        52.373,        0.866
InfPerSec_Func_0,             48.693,        46.208,        52.373,        0.866
InfPCycles_Func_0,            29787610.666,  27685323.000,  31379989.000,  526882.413
EffectiveFrequencyMHz_Func_0,1449.982,       1449.820,      1450.003,      0.024
```

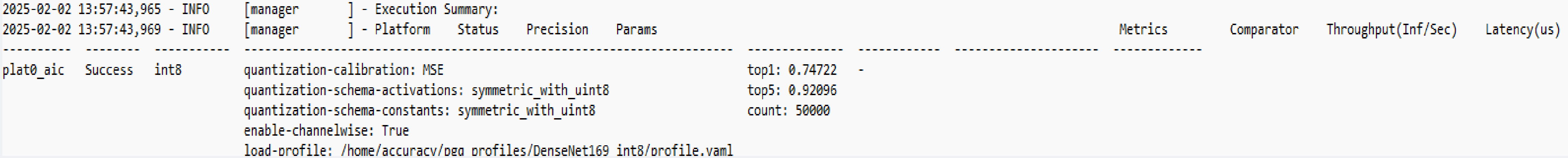
Accuracy tools in QAIC

DenseNet169 – INT8 precision

- **Command used –**

source /opt/qti-aic/dev/python/qaic-env/bin/activate && python3 /opt/qti-aic/tools/qaic-pytools/**qaic-acc-evaluator.py** -config /home/qraniumtest/ml-tools/configs/accuracy_evaluator/CV/densenet169/densenet169_config.yaml -platform-tag-params=aic_int8,load-profile:/home/accuracy/pgq_profiles/DenseNet169_int8/profile.yaml -cleanup end -work-dir /home/qraniumtest/logs/qacc_logs/qacc_evaluator/DenseNet169/int8 -silent -platform-tag aic_int8 -device-id 4

- **Metrics -**



```
2025-02-02 13:57:43,965 - INFO     [manager       ] - Execution Summary:
2025-02-02 13:57:43,969 - INFO     [manager       ] - Platform    Status    Precision    Params                                                                   Metrics         Comparator    Throughput(Inf/Sec)    Latency(us)
----------  --------  -----------  ---------------------------------------------------------------------  ----------------  ------------  ---------------------  -------------
plat0_aic   Success   int8         quantization-calibration: MSE                                          top1: 0.74722    -
                                   quantization-schema-activations: symmetric_with_uint8                  top5: 0.92096
                                   quantization-schema-constants: symmetric_with_uint8                    count: 50000
                                   enable-channelwise: True
                                   load-profile: /home/accuracy/pgq_profiles/DenseNet169_int8/profile.yaml
```

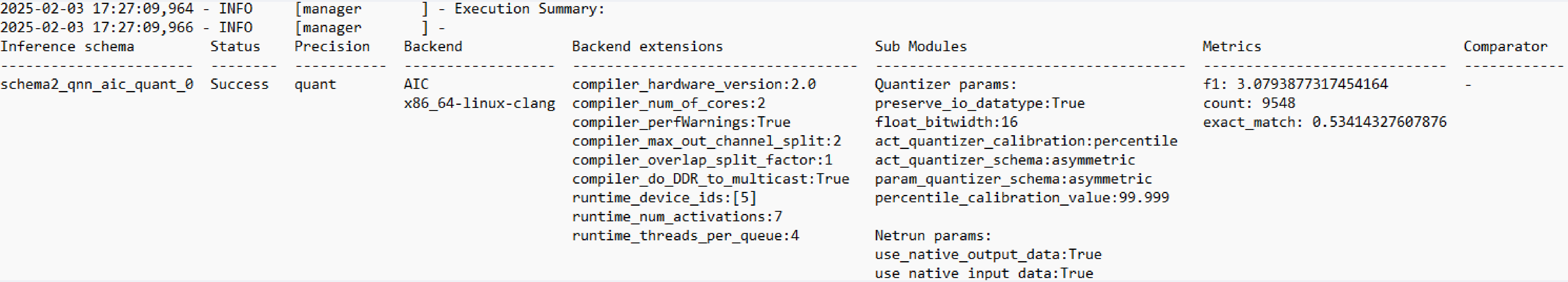
QNN Accuracy Evaluator

BERT Large Packed Boolean Mask – INT8_MP precision

- **Command used –**

docker exec mlg-dev-22.04_qraniumtest bash -c "source /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/qnn_qairt_accuracy/venv/bin/activate; source /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/aic_perf/qraft/src/tools/qairt_acc/scripts/export_qnn_variables.sh /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/qnn-aic-sdk-V2.31.0-RC5/ /local/mnt/workspace/jyashwan/model_zoo; /usr/bin/time -v /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/qnn-aic-sdk-V2.31.0-RC5//bin/x86_64-linux-clang/qairt-accuracy-evaluator -config /local/mnt/workspace/accuracy//qnn_new_model_configs/public/bert/bert_packing_strategy_boolean_mask.yaml -cleanup intermediate -inference_schema_tag qnn_int8_mp -work_dir /local/mnt/workspace/qnn-aic-sdk-V2.31.0-RC5/Gigabyte_14NSP/PCIe/128/qnn_qairt_accuracy/BERT_Large_Packed_Boolean_Mask/int8_mp/acc_log -silent -device_id 5 “

Metrics -



```
2025-02-03 17:27:09,964 - INFO     [manager       ] - Execution Summary:
2025-02-03 17:27:09,966 - INFO     [manager       ] -
Inference schema         Status    Precision    Backend             Backend extensions                  Sub Modules                            Metrics                        Comparator
-----------------------  --------  -----------  ------------------  ----------------------------------  -------------------------------------  -----------------------------  ------------
schema2_qnn_aic_quant_0  Success   quant        AIC                 compiler_hardware_version:2.0       Quantizer params:                      f1: 3.0793877317454164         -
                                                x86_64-linux-clang  compiler_num_of_cores:2             preserve_io_datatype:True              count: 9548
                                                                    compiler_perfWarnings:True          float_bitwidth:16                      exact_match: 0.53414327607876
                                                                    compiler_max_out_channel_split:2    act_quantizer_calibration:percentile
                                                                    compiler_overlap_split_factor:1     act_quantizer_schema:asymmetric
                                                                    compiler_do_DDR_to_multicast:True   param_quantizer_schema:asymmetric
                                                                    runtime_device_ids:[5]              percentile_calibration_value:99.999
                                                                    runtime_num_activations:7
                                                                    runtime_threads_per_queue:4         Netrun params:
                                                                                                        use_native_output_data:True
                                                                                                        use native input data:True
```

Compiling and running a model in QNN (Non-LLM)

CASE 2: MIXED PRECISON

CONVERTER

```
$QNN_SDK_ROOT/bin/x86_64-linux-clang/qairt-converter --input_network /home/qraniumtest/model_zoo//MLPerfModels/BertLarge/generatedModels/ONNX/BERT_MLCommons_Flexible_BS_SL_Packed_BoolMask.onnx --output_path /home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/model.dlc  --onnx_define_symbol batch_size 1 --onnx_define_symbol seg_length 384 --onnx_no_simplification --float_bias_bitwidth 32 --preserve_io_datatype  --onnx_defer_loading  --quantization_overrides /home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/bert_packing_strategy_boolean_mask_node_precision_file.json
```

QUANTIZER

```
$QNN_SDK_ROOT/bin/x86_64-linux-clang/qairt-quantizer --input_dlc /home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/model.dlc --output_dlc /home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/model_quantized.dlc  --preserve_io_datatype --use_native_input_files  --float_bitwidth 16 --input_list /home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/qnn_list.txt  --act_quantizer_schema unsignedsymmetric --param_quantizer_schema unsignedsymmetric --backend AIC --float_bitwidth 16 --float_bitwidth 16
```

CONTEXT BINARY

```
$QNN_SDK_ROOT/bin/x86_64-linux-clang/qnn-context-binary-generator --binary_file qnnGraphDLC --model $QNN_SDK_ROOT/lib/x86_64-linux-clang/libQnnModelDlc.so --backend $QNN_SDK_ROOT/lib/x86_64-linux-clang/libQnnAic.so --output_dir /home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/ --config_file /home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/qnn_config.json --dlc_path /home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/model_quantized.dlc --log_level debug --backend_binary /home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/programqpc_dir/programqpc.bin
```

Compiling and running a model in QNN (Non-LLM)

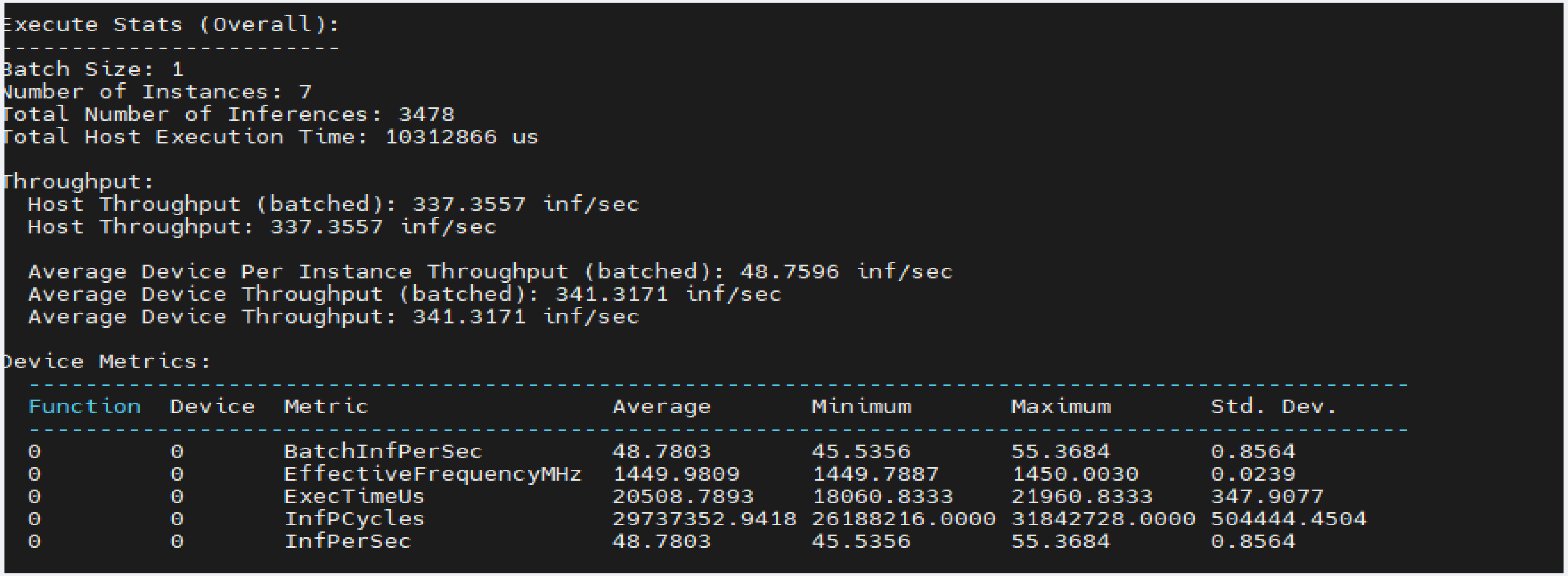
CASE 2: MIXED PRECISON

NET RUNNER

cd /home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/ && $QNN_SDK_ROOT/bin/x86_64-linux-clang/qnn-net-run --backend $QNN_SDK_ROOT/lib/x86_64-linux-clang/libQnnAic.so --input_list /home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/qnn_list.txt --log_level error --profiling_level basic --retrieve_context qnnGraphDLC.bin --config_file /home/qraniumtest/binaries/BERT_Large_Packed_Boolean_Mask/qnn_net_runner_config.json --duration 10 --keep_num_outputs 2 --use_native_input_files

PROFILE VIEWER

- $QNN_SDK_ROOT/bin/x86_64-linux-clang/qnn-profile-viewer --input_log output/qnn-profiling-data.log --reader $QNN_SDK_ROOT/lib/x86_64-linux-clang/libQnnAicProfilingReader.so



```
Execute Stats (Overall):
------------------------
Batch Size: 1
Number of Instances: 7
Total Number of Inferences: 3478
Total Host Execution Time: 10312866 us

Throughput:
  Host Throughput (batched): 337.3557 inf/sec
  Host Throughput: 337.3557 inf/sec

  Average Device Per Instance Throughput (batched): 48.7596 inf/sec
  Average Device Throughput (batched): 341.3171 inf/sec
  Average Device Throughput: 341.3171 inf/sec

Device Metrics:
  -------------------------------------------------------------------------------------------------
  Function  Device  Metric                 Average       Minimum       Maximum       Std. Dev.
  -------------------------------------------------------------------------------------------------
  0         0       BatchInfPerSec         48.7803       45.5356       55.3684       0.8564
  0         0       EffectiveFrequencyMHz  1449.9809     1449.7887     1450.0030     0.0239
  0         0       ExecTimeUs             20508.7893    18060.8333    21960.8333    347.9077
  0         0       InfPCycles             29737352.9418 26188216.0000 31842728.0000 504444.4504
  0         0       InfPerSec              48.7803       45.5356       55.3684       0.8564
```

QNN- Tools

Accuracy Evaluator:

- The qairt-accuracy-evaluator tool provides a framework to evaluate end-to-end accuracy metrics for a model on a given dataset. In addition, the tool can be used to identify the best quantization options for a model on a given set of inputs.

  -
    ```
    qairt-accuracy-evaluator -config efficientNet_b0_config.yaml -cleanup intermediate -inference_schema_tag qnn_fp16 -work_dir WORKING_DIR_PATH -silent -device_id 5
    ```
  - Documentation: Qualcomm® AI Engine Direct

Accuracy Debugger

- The accuracy-debugger tool finds inaccuracies in a neural-network at the layer level. The tool compares the golden outputs produced by running a model through a specific ML framework (ie. Tensorflow, Onnx, TFlite) with the results produced by running the same model through Qualcomm's QNN Inference Engine. The inference engine can be run on a variety of computing mediums including GPU, CPU and AIC.

  -
    ```
    qairt-accuracy-debugger --inference_engine --model_path efficientnet-b0.onnx --runtime aic --architecture x86_64-linux-clang --input_list qnn_efficientNet_b0_list.txt --calibration_input_list qnn_efficientNet_b0_list.txt --working_dir INF_WORKING_DIR --output_dirname InferenceResults --executor_type QNN --engine_path SDK_PATH --verbose --host_device x86 --profiling_level basic --log_level error --debug_mode_off --bias_bitwidth 32 --param_quantizer_schema symmetric --act_quantizer_schema symmetric --param_quantizer_calibration min-max --use_per_channel_quantization --input_tensor 'input.1' 1,3,224,224 input.raw float32 --output_tensor '666'  "
    ```
  - Documentation: Qualcomm® AI Engine Direct

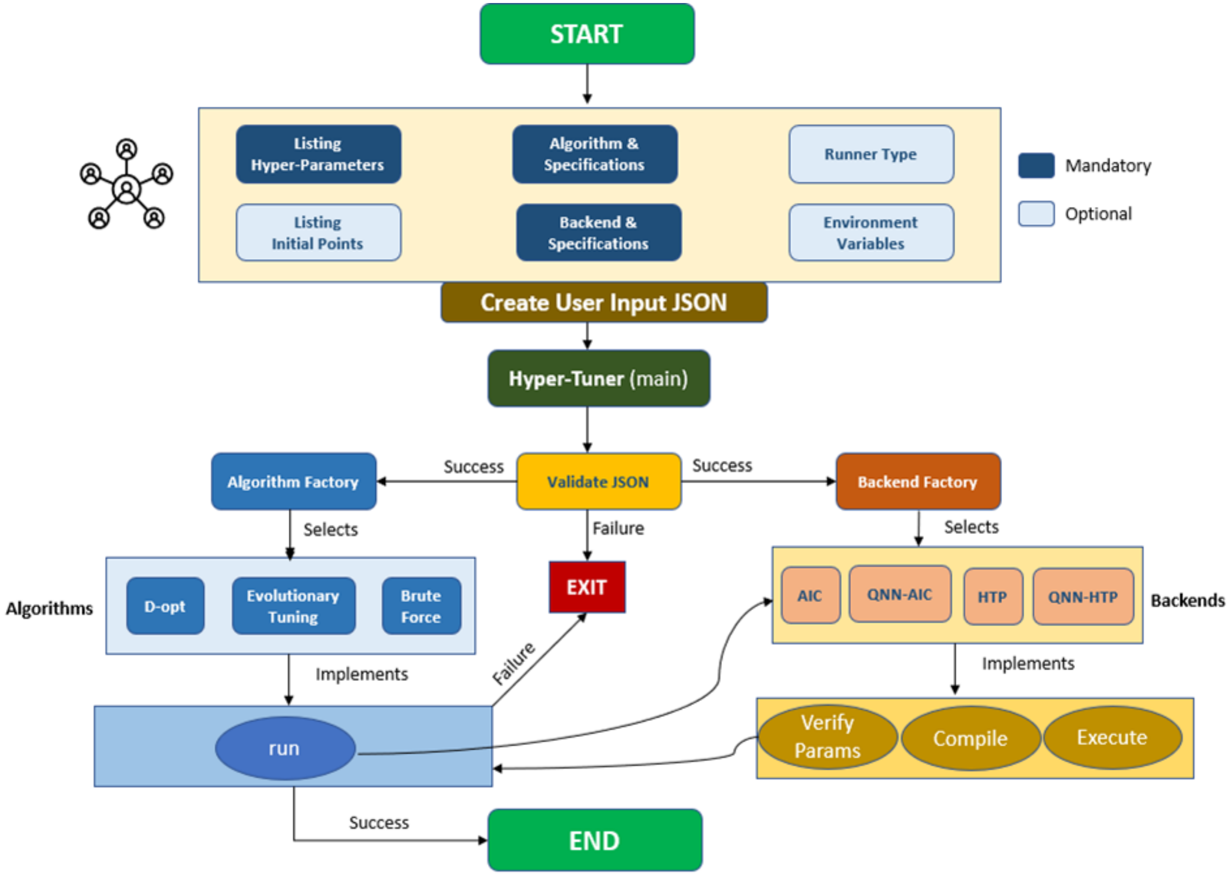
Hyper Tuner

- qnn-hypertuner, also referred to as Hypertuner, is a performance tuning tool that provides an optimal combination of compiler parameters. The Hypertuner takes, as input, a JSON configuration file containing the name of the deep learning model, hyper parameters, search algorithm, and backend. It then performs a search over the hyperspace as defined by the input parameters and outputs an optimal parameter set for use by downstream tasks or applications.

QNN-AIC : Documentation: **8. QNN SDK — Cloud AI 100 documentation**

Hyper Tuner

- Performance tuning tool that provides optimal combination of compiler parameters.



hyper-parameters for QNN-AIC backend

Hyper-parameter name	Data type	Valid value
compiler_num_of_cores	integral	1 - #NSP
compiler_max_out_channel_split	integral	1 - #NSP
compiler_overlap_split_factor	integral	1 - 4
compiler_stats_batch_size	integral	1 - 8
compiler_buffer_dealloc_delay	integral	0 - 4
compiler_size_split_granularity	integral	512 - 2048
compiler_enable_depth_first	string	“True”, “False”
compiler_VTCM_working_set_limit_ratio	float	0.25 - 1.0
runtime_num_activations	integral	1 - #NSP
compiler_depth_first_mem	integral	8 - 16
compiler_do_DDR_to_multicast	string	“True”, “False”
compiler_userDMAProducerDMAEnabled	string	“True”, “False”
compiler_combine_inputs	string	“True”, “False”
compiler_combine_outputs	string	“True”, “False”

Example: example/example_qnn_aic.json - qctaisw/HyperTuner - Gitiles

refs/heads/master - qctaisw/HyperTuner - Gitiles

Agenda

Apps Overview

Install SDKs

Compiling and running a model in QAIC

Accuracy tools in QAIC

Running LLMs in QAIC using Qefficient

QNN Workflow

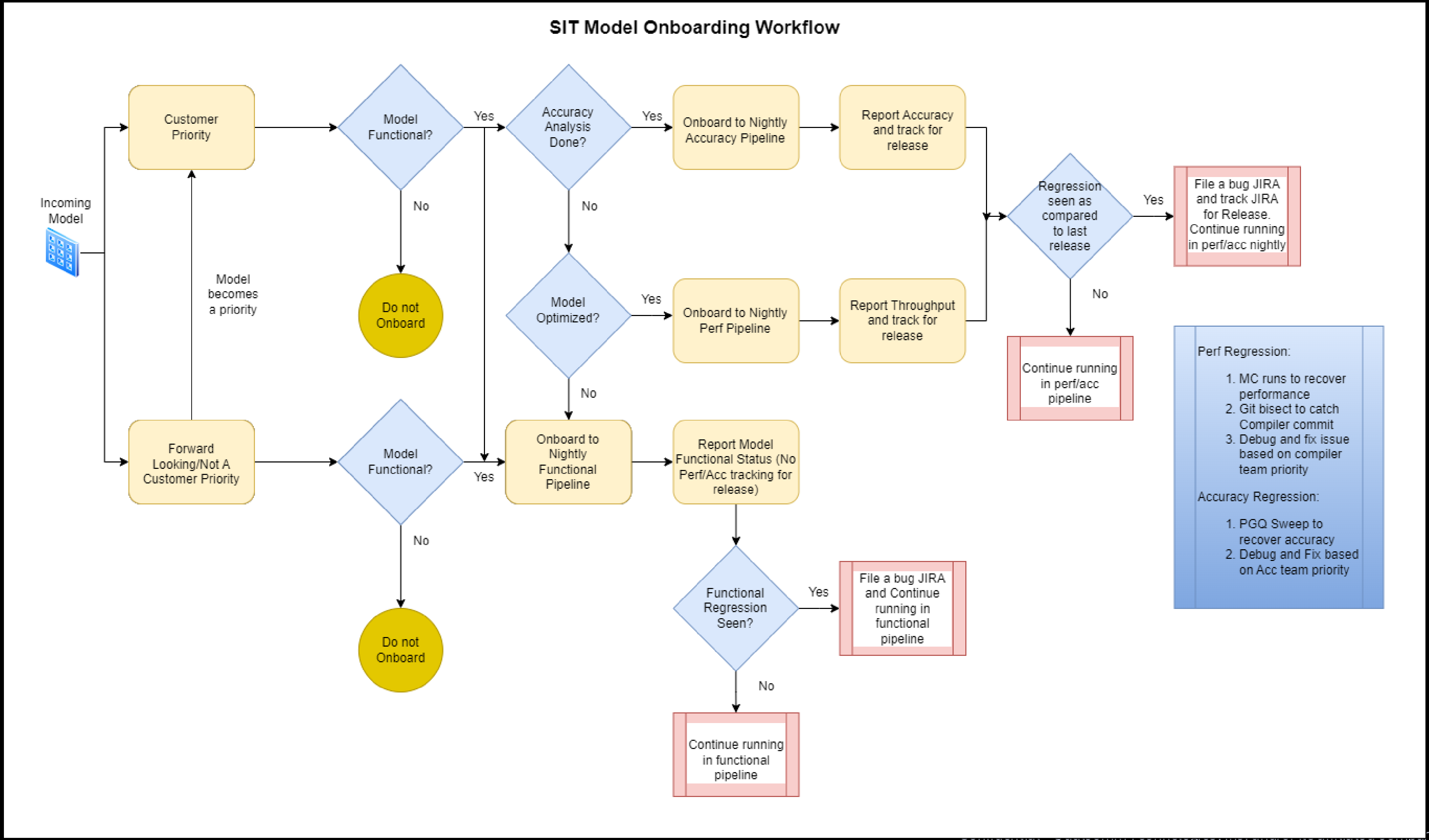
Compiling and running a model in QNN

Running LLMs in QNN using Qefficient

Accuracy tools in QNN

Automation Tools

APPS SIT – Model Onboarding workflow


SIT Model Onboarding Workflow
Incoming Model
Customer Priority
Model becomes a priority
Forward Looking/Not A Customer Priority
Model Functional?
Yes
No
Do not Onboard
Model Functional?
Yes
No
Do not Onboard
Accuracy Analysis Done?
Yes
No
Onboard to Nightly Accuracy Pipeline
Report Accuracy and track for release
Model Optimized?
Yes
No
Onboard to Nightly Perf Pipeline
Report Throughput and track for release
Onboard to Nightly Functional Pipeline
Report Model Functional Status (No Perf/Acc tracking for release)
Functional Regression Seen?
Yes
File a bug JIRA and Continue running in functional pipeline
No
Continue running in functional pipeline
Regression seen as compared to last release
Yes
File a bug JIRA and track JIRA for Release. Continue running in perf/acc nightly
No
Continue running in perf/acc pipeline

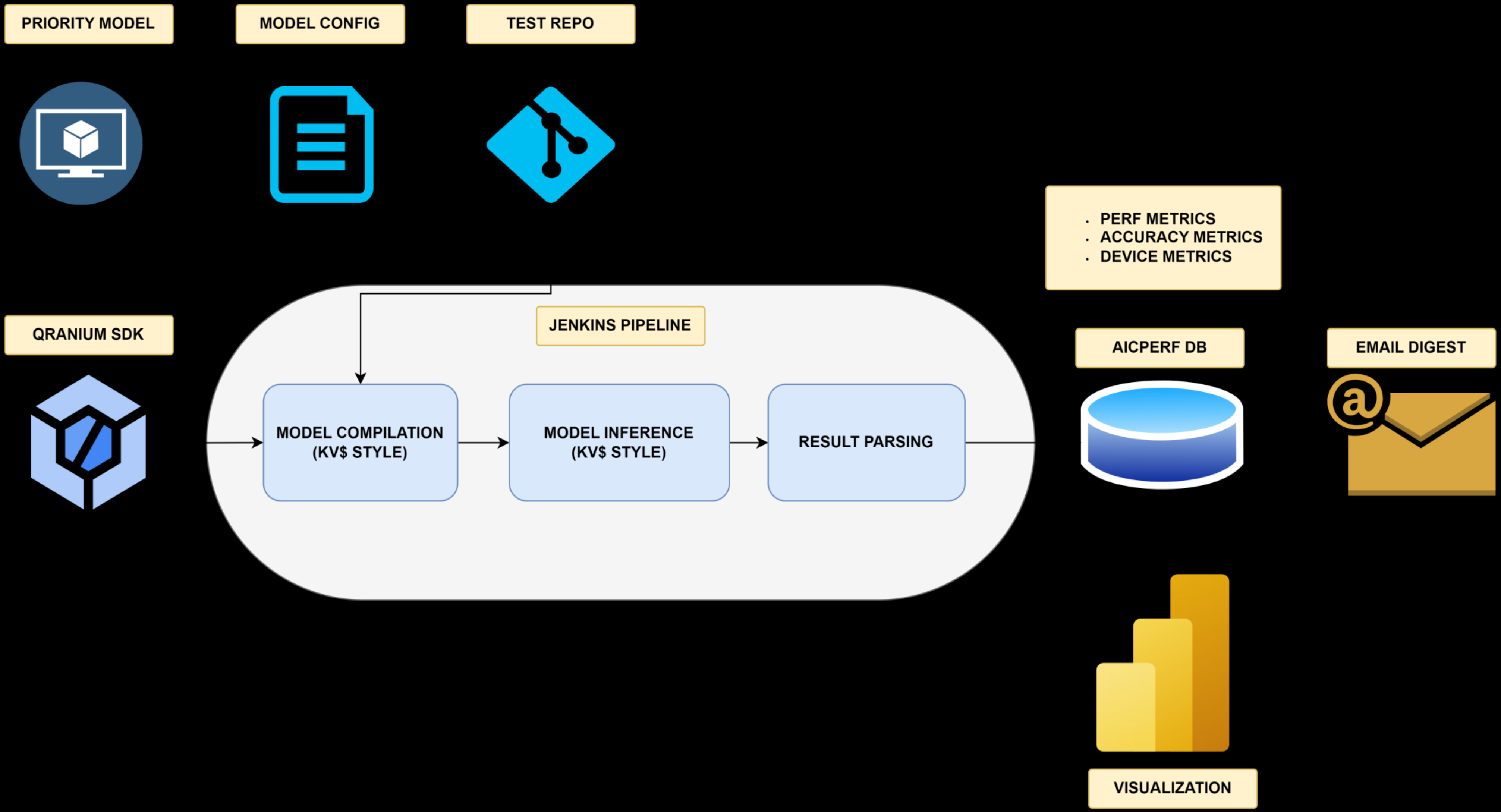
Perf Regression:

1. MC runs to recover performance
2. Git bisect to catch Compiler commit
3. Debug and fix issue based on compiler team priority

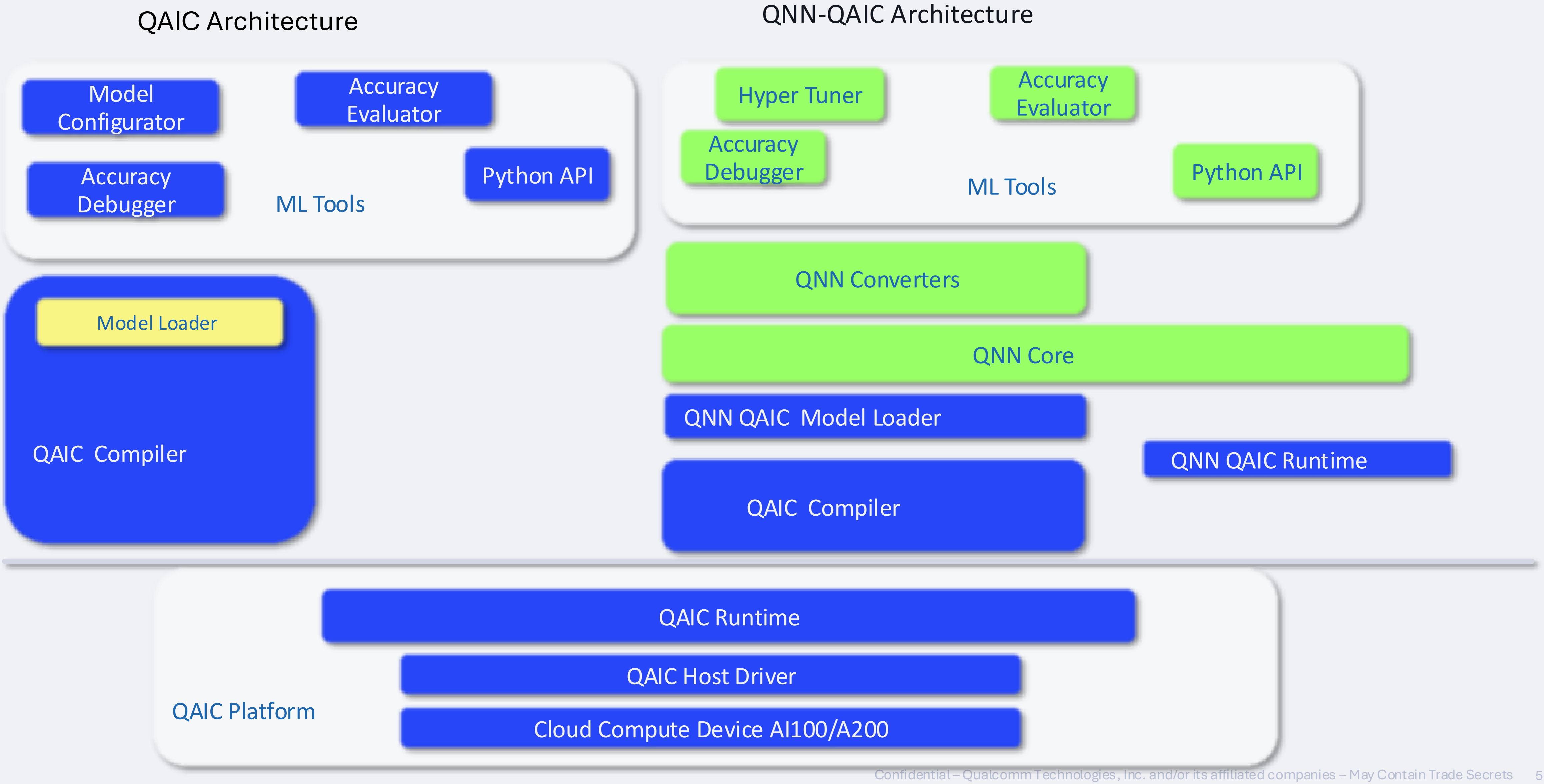
Accuracy Regression:

1. PGQ Sweep to recover accuracy
2. Debug and Fix based on Acc team priority

APPS SIT – LLM Benchmarking workflow



Architecture



QAIC Architecture

QNN-QAIC Architecture

Model Configurator

Accuracy Evaluator

Accuracy Debugger

Python API

ML Tools

Hyper Tuner

Accuracy Evaluator

Accuracy Debugger

ML Tools

Python API

Model Loader

QAIC Compiler

QNN Converters

QNN Core

QNN QAIC Model Loader

QNN QAIC Runtime

QAIC Compiler

QAIC Runtime

QAIC Host Driver

Cloud Compute Device AI100/A200

QAIC Platform

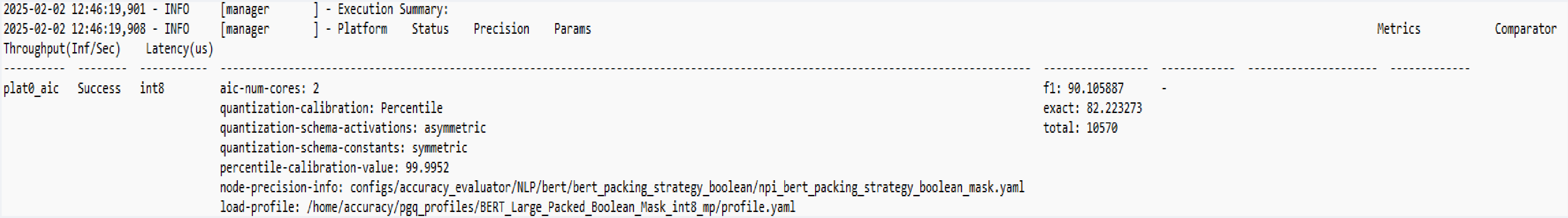
Accuracy tools in QAIC

BERT Large Packed Boolean Mask – INT8_MP precision

- **Command used –**

/opt/qti-aic/dev/python/qaic-env/bin/python3 qaic-acc-evaluator.py -config /home/qraniumtest/ml-tools/configs/accuracy_evaluator/NLP/bert/bert_packing_strategy_boolean/bert_packing_strategy_boolean_mask.yaml -pipeline_start preproc -pipeline_end metric -cleanup end -platform_tag aic_int8_mp -silent True -device_id 0 -work_dir /home/qraniumtest/logs/qacc_logs/qacc_evaluator/BERT_Large_Packed_Boolean_Mask/int8_mp -platform_tag_params [['aic_int8_mp,load-profile:/home/accuracy/pgq_profiles/BERT_Large_Packed_Boolean_Mask_int8_mp/profile.yaml']] -perf_iter_count 200

- **Metrics -**



```
2025-02-02 12:46:19,901 - INFO     [manager       ] - Execution Summary:
2025-02-02 12:46:19,908 - INFO     [manager       ] - Platform    Status    Precision    Params                                                                                                                                 Metrics            Comparator
Throughput(Inf/Sec)    Latency(us)
----------  --------  -----------  -----------------------------------------------------------------------------------------------------------------------------------  -----------------  ------------  ---------------------  -------------
plat0_aic   Success   int8         aic-num-cores: 2                                                                                                                     f1: 90.105887      -
                                   quantization-calibration: Percentile                                                                                                 exact: 82.223273
                                   quantization-schema-activations: asymmetric                                                                                          total: 10570
                                   quantization-schema-constants: symmetric
                                   percentile-calibration-value: 99.9952
                                   node-precision-info: configs/accuracy_evaluator/NLP/bert/bert_packing_strategy_boolean/npi_bert_packing_strategy_boolean_mask.yaml
                                   load-profile: /home/accuracy/pgq_profiles/BERT_Large_Packed_Boolean_Mask_int8_mp/profile.yaml
```

Thank you

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

Qualcomm and Snapdragon are trademarks or registered trademarks of Qualcomm Incorporated.
Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to "Qualcomm" may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Follow us on: in 𝕏 ◎ ▶ f
For more information, visit us at qualcomm.com & qualcomm.com/blog