aetina

AI On Prem QSG

Accelerate Every Edge of Business

Agenda

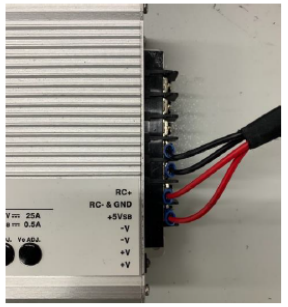
1. Bring up
2. Qualcomm AI100U Card Management
3. SW Stack
4. Ducker Run & Model Deploy
5. Efficient Transformer / Validated AI Model
6. RAG

Bring Up

AI On-Prem Solution

Intel Core i7-13700E
48GB DDR5 UDIMM x4 Total 192GB
M.2 2280 2TB x1 Total 2TB
Qualcomm AI100 Ultra x1
600 W

- Step1: Power supply wiring
- Step2: Press power button
- Step3: Enter Ubuntu OS password : aetina (for username: AI On Prem Solution) or create new account



HEP-600-24 Power supply DC Side wiring

Pin #	Power supply	DC Cable
1	RC+	
2	RC- & GND	
3	+5Vsb	
4	-V	Black
5	-V	Black
6	+V	Red
7	+V	Red

HEP-600-24 Power supply AC Side wiring

Pin #	Power supply	AC Cable
1	GND	Green
2	L	Black
3	N	White

Card Management

- qaic-util command line utility enables developers to query
  - card and SoC(s) health
  - firmware version
  - compute/memory/IO resources available vs in-use
  - card power and temperature
  - status of certain device capabilities like ECC etc

```
sudo /opt/qti-aic/tools/qaic-util -q | grep -e Status -e QID
QID 0
        Status:Ready
QID 1
        Status:Ready
QID 2
        Status:Ready
QID 3
        Status:Ready
```

https://quic.github.io/cloud-ai-sdk-pages/latest/Getting-Started/System-Management/system-management/index.html

Card Management

- sudo /opt/qti-aic/tools/qaic-util --table 1

```
Device Selection  Scroll-Up-Down [                                                        ]
 (*) All-Devices
  o  DeviceID-0
  o  DeviceID-1
  o  DeviceID-2
  o  DeviceID-3

DeviceID-0 --- Status: Ready --- FW Version: 1.19.10.0
NSP Used / Total | NW Loaded / Active | DRAM Used / Total      | DRAM Bandwidth   | NSP Frequency
          0 / 16 |              0 / 0 | 253361 KB / 313917     | 178608.00 KBps   | 825.60 Mhz
DDR Frequency    | Temperature        | Brd. Power / TDP       | Soc Power
    2133.00 Mhz  |            28.00 C | 15.00 W / 75.00 W      |    15.00 W

DeviceID-1 --- Status: Ready --- FW Version: 1.19.10.0
NSP Used / Total | NW Loaded / Active | DRAM Used / Total      | DRAM Bandwidth   | NSP Frequency
          0 / 16 |              0 / 0 | 253361 KB / 313917     | 178608.00 KBps   | 825.60 Mhz
DDR Frequency    | Temperature        | Brd. Power / TDP       | Soc Power
    2133.00 Mhz  |            28.00 C | 19.00 W / 75.00 W      |    19.00 W

DeviceID-2 --- Status: Ready --- FW Version: 1.19.10.0
NSP Used / Total | NW Loaded / Active | DRAM Used / Total      | DRAM Bandwidth   | NSP Frequency
          0 / 16 |              0 / 0 | 253361 KB / 313917     | 178608.00 KBps   | 825.60 Mhz
DDR Frequency    | Temperature        | Brd. Power / TDP       | Soc Power
    2133.00 Mhz  |            28.00 C | 13.00 W / 75.00 W      |    13.00 W

DeviceID-3 --- Status: Ready --- FW Version: 1.19.10.0
NSP Used / Total | NW Loaded / Active | DRAM Used / Total      | DRAM Bandwidth   | NSP Frequency
          0 / 16 |              0 / 0 | 253361 KB / 313917     | 178608.00 KBps   | 825.60 Mhz
DDR Frequency    | Temperature        | Brd. Power / TDP       | Soc Power
    2133.00 Mhz  |            28.00 C | 10.00 W / 75.00 W      |    10.00 W
```

SW Stack

vLLM → Docker → Deploy QPC

Efficient Transformer → Export ONNX, Model Compile

APP SDK 1.19.8

Platform SDK 1.19.8

Ubuntu 22.04
Kernel 6.8 HWE
Sys Fan control utility

AI Model
Hugging Face

Pytorch & TensorRT

Efficient Transformer
Export to ONNX
(w/1TB RAM SYS)

Compile

QPC

Deploy to AIP-FR68

Gen AI Application

Ducker Run

- https://quic.github.io/cloud-ai-sdk-pages/latest/Getting-Started/Installation/vLLM/vLLM/#docker-build-steps

```
docker pull ghcr.io/quic/cloud_ai_inference_ubuntu22:1.19.8.0
docker run -dit \
-p 8000:8000 \
--device=/dev/accel/accel0 \
--device=/dev/accel/accel1 \
--device=/dev/accel/accel2 \
--device=/dev/accel/accel3 \
-v ~/.cache/huggingface:/root/.cache/huggingface
ghcr.io/quic/cloud_ai_inference_ubuntu22:1.19.8.0
```

Download QPC & Model Deploy

- **Download QPC :**

http://qualcom-qpc-models.s3-website-us-east-1.amazonaws.com/QPC/catalog-index/#sdk-version-12012

Uzip it and add below python to export your QPC folder for vLLM

- i.e., export VLLM_QAIC_QPC_PATH=/qpc_home/qpc-8fb11380ed806bf4/qpc

- **vLLM command for LLAMA 3.3 70B as below :**

python3 -m vllm.entrypoints.api_server --host 0.0.0.0 --port 8000 --model meta-llama/Llama-3.3-70B-Instruct --max-model-len 8192 --max-num-seq 1 --max-seq_len-to-capture 128 --device qaic --quantization mxfp6 --kv-cache-dtype mxint8 --device-group 0,1,2,3

- https://quic.github.io/cloud-ai-sdk-pages/latest/Getting-Started/Installation/vLLM/vLLM/index.html#vllm-input-arguments-for-qaic

Efficient Transformer / Validated AI Model

- **Efficient Transformer**

https://github.com/quic/efficient-transformers/tree/main

https://quic.github.io/efficient-transformers/index.html

- **Validated AI Model**

https://quic.github.io/efficient-transformers/source/validate.html

RAG with ChromaDB

- RAG with ChromaDB — Imagine SDK 0.4.2

- This page shows an example of how the Imagine SDK can be used with LangChain for retrieval-augmented generation. Retrieval augmented generation is a type of information retrieval process. It modifies interactions with a large language model so that it responds to queries with reference to a specified set of documents, using it in preference to information drawn from its own vast, static training data. This allows LLMs to use domain-specific and/or updated information.

Enable Llama3.3 70B

VLLM Server Setup and API Verification Guide

Step 1: Start the VLLM Server

1. Activate Python virtual environment:
   Open your terminal and run the following command to activate the VLLM-specific Python virtual environment:
   source qaic-vllm-venv/bin/activate
2. Navigate to the VLLM directory:
   cd vllm
3. Set QPC path:
   Set the VLLM_QAIC_QPC_PATH environment variable to your QPC path:
   export VLLM_QAIC_QPC_PATH=/home/kq/qpc-8fb11380ed806bf4/qpc
4. Start OpenAI-compatible API server:
   Run this command to start the VLLM API server with the specified model and parameters:
   python3 -m vllm.entrypoints.openai.api_server --host 0.0.0.0 --port 8000 --model meta-llama/Llama-3.3-70B-Instruct --max-model-len 8192 --max-num-seq 1 --max-seq-len-to-capture 128 --device qaic --quantization mxfp6 --kv-cache-dtype mxint8 --device-group 0,1,2,3
5. Verify server startup:
   Once started, you will see logs indicating the model is loading. Do not close this terminal window.

Step 2: Verify if the API Server is Working Properly

1. Open a new terminal window while the server is running.
2. Send a test request:
   Run the following curl command:

```
curl --location --request POST 'http://0.0.0.0:8000/v1/completions' \
--header 'Content-Type: application/json' \
--data '{
  "prompt": "<|user|>\nWhat is 1+1?\n<|assistant|>\n",
  "model": "meta-llama/Llama-3.3-70B-Instruct",
  "max_tokens": 50
}'
```

Step 3: Check the Response

If successful:

You will receive a JSON response containing the model's answer to '1+1' (usually in the 'text' field), confirming the request was handled successfully.

If an error occurs:

Check the following:

- The VLLM server is still running (Step 1).
- The JSON format in the curl command is correct (all strings in double quotes).
- Network connectivity is normal.

Documentation & Image Download

- Login myaetina
- Navigate to Download area > Edge AI Platform >

QSG :

Quick Start Guide > QSG_AI On Prem.pdf file

User manual :

User manual > User Manual_AIP-FR68-A1_AI On-Prem.pdf

Image file :

Image > Image_AIP-FR68_AI On-Prem.tar.gz

*Image recovery SOP please reference AI On Prem user manual

Thank you

DISCLAIMER

This document is for presentation purposes only and should not be construed as an offer, invitation or solicitation to subscribe for, purchase or sell any investment. Neither it nor anything it contains shall form the basis of any contract whatsoever. Opinions expressed herein reflect the current judgement of the management of Aetina Corporation. The presentation contains forward-looking statements that involve risks and uncertainties. The actual results of Aetina Corporation may differ materially from those anticipated in these forward-looking statements and forecasts as a result of a number of factors. The management of GF does not accept any liability whatsoever with respect to the use of this presentation.

At Aetina, we accelerate the pathway to the edge of intelligence – where ambition fuels exploration, progression meets transformation, and knowledge propels innovation.

For those dedicated to discovering new possibilities, we elevate their skills to the cutting edge and push the limits of what's possible.

We navigate businesses toward AI implementation across verticals with our integrated solutions and adaptable services.

Collaborating closely with our partners, we gain actionable insights from the forefront of their fields to tailor our solutions for professional domains.

With streamlined solutions and addressed scenarios, AI deployment is simple and efficient. The dream of an intelligent revolution becomes practical.

WE HAVE YOUR EDGE ON AND FUEL AI TRANSFORMATION BEYOND

aetina

Edge ON AI Beyond