Qualcomm

Opensource Deep learning Compiler Solution Updates for Qualcomm® Adreno GPU™

Siva Rama Krishna Reddy B

Engineer, Senior Staff/Manager , Qualcomm India Private Limited

Agenda

MLC/TVM Solution for Generative AI

Model support & Performance

Windows on Snapdragon support

Upcoming

Who are we?

Team

Part of Qualcomm GPU Research Team (GRT) from Bangalore

- Focus on open-source AI solutions for Adreno GPU
- We primarily contribute to TVM and MLC communities.
- Long term contributors for IWOCL

Key contributors for this project

Krishna Raju Vegiraju
Staff Engineer

Deepanshu Singh
Engineer Senior

Hongqiang Wang
Engineer, Principal/Manager

Alex Bourd
Senior Director, Technology

Pavan Kumar A
Senior Director, Engineering

Resources/References on this project

- TVM (Tensor Virtual Machine) : https://tvm.ai
- LLM MLC (Machine Learning Compilation) : https://llm.mlc.ai
- Publicly available SDK: https://artifacts.codelinaro.org/ui/native/clo-472-adreno-opensource-ai/
- Qualcomm Developer Network blogs:
  - “Introducing the new OpenCL™ GPU backend in llama.cpp for Qualcomm Adreno GPUs”
  - Blog: How to run DeepSeek models on Windows on Snapdragon – Llama.cpp and MLC-LLM tutorial
  - https://www.qualcomm.com/developer/blog/2025/02/harnessing-qualcomm-adreno-gpu-generative-ai-open-source-approach
  - https://www.qualcomm.com/developer/blog/2022/09/accelerate-your-machine-learning-networks-using-tvm-and-adreno-opencl-ml-apis-adreno-gpus

MLC/TVM Solution for Generative AI

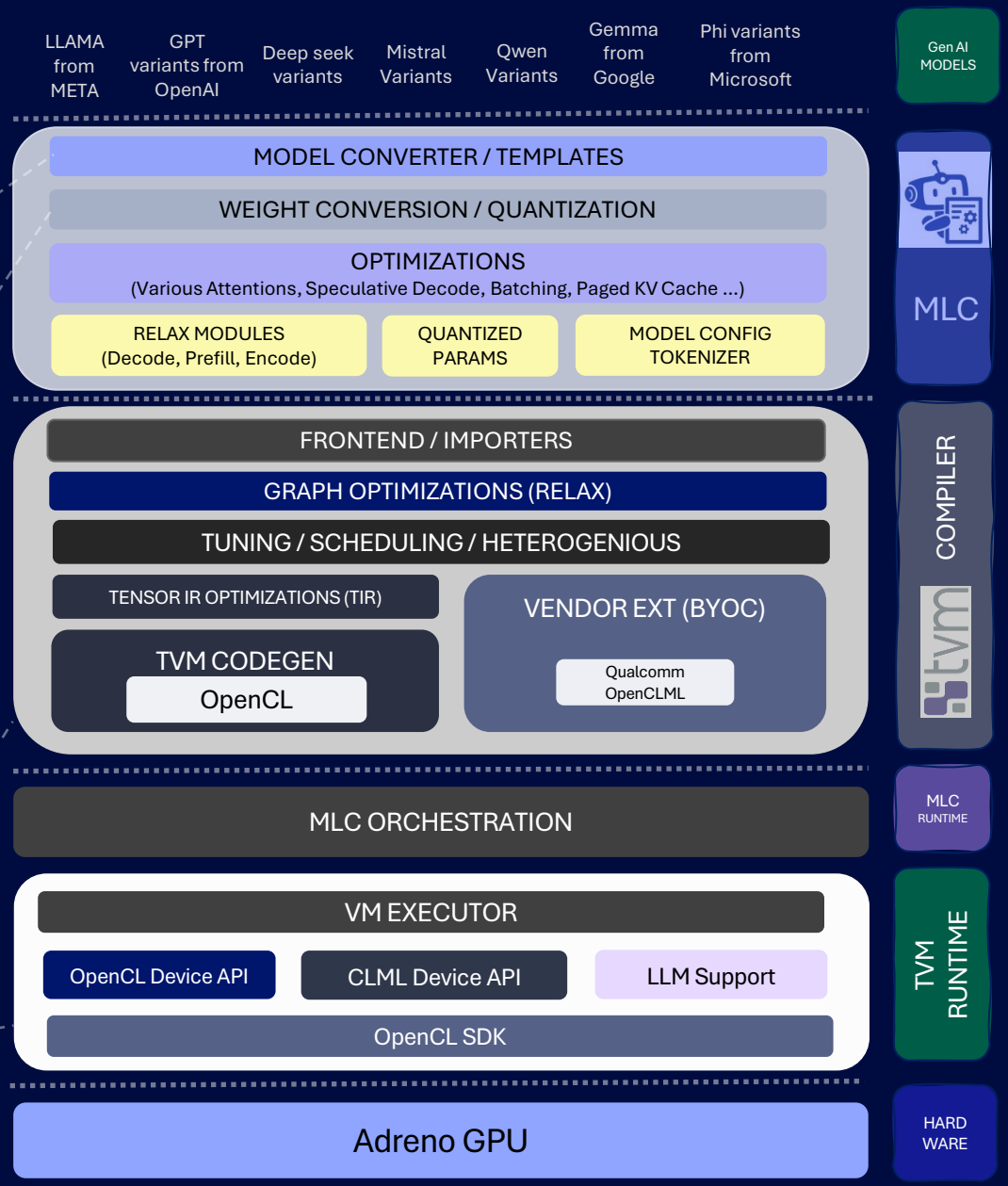
- We know the model architecture hence we build it as a TVM Relax module.
- In general, **community brings it in**. Sometimes we do for non-popular architectures.

- Adreno uses q4f16_0. One can choose from various quantization available and bring in new if required.

"Our deep understanding of Adreno GPU combined with Open-source advantage drives rapid development of Generative AI model enablement"

- Schedules that favor Qualcomm® Adreno™ GPU with Q4f16_0 are developed and contributed.
- Development in progress to inject specialized kernels for critical layers via OpenCLML offloading.

- CPU access of KVCache memory objects is accommodated by zero copy or memory mapping from OpenCL memory objects.



Model support and Performance

- MLC as a solution supports nearly 220+ models covering 60+ architectures from various researchers like Meta, Google, Qwen, Mistral, OpenAI, Deep seek and Microsoft - https://mlc.ai/models
- Depending parameter size most of them are compatible to run on Adreno GPU targets across Mobile (Android), Compute(Windows), Auto(Linux) and IoT (Custom Linux) targets.
- Performance data for few prominent models on Android platforms

Model	Snapdragon® 8 Gen 3 (Android)			Snapdragon 8 Elite (Android)		
	Decode (toks/sec) for 100	Encode for 256 prompt		Decode (toks/sec) for 100	Encode for 256 prompt	
		toks / sec	Time To First Token (sec)		toks / sec	Time To First Token (sec)
DeepSeek-R1-Distill-Llama-8B	11.1	64	4	12.5	95	2.69
DeepSeek-R1-Distill-Qwen-1.5B	22.2	232	1.1	50	413	0.62
DeepSeek-R1-Distill-Qwen-7B	11.6	65	3.94	13	101.5	2.52
Llama-2-7b-chat-hf	14	88	2.89	14.34	92.5	2.7
Meta-Llama-3-8B-Instruct	11.1	80.5	3.16	12.5	84.6	3.01
gemma-2b-it	14.77	244.7	1.02	33.25	293.5	0.85
Mistral-7B-Instruct-v0.2	10.85	64.1	3.9	11.5	71	3.5
phi-2	28.9	137.7	1.98	30	145.4	1.86
Phi-3-mini-4k-instruct	22.6	145.5	1.75	24.88	156.358	1.64
Qwen-7B-Chat	12.3	72.3	3.6	11.9	72.5	3.59
llava-1.5-7b-hf	13.3	85.8	2.9	13.1	85.6	2.9

*** Performance numbers are produced under ideal conditions. There might be a slight difference when reproduced on different commercial devices.*

Developer blogs for extended information and resources

Harnessing Qualcomm Adreno GPU for Generative AI: Open-source Approach

How to tun Deep Seek models on Windows on Snapdragon - MLC-LLM tutorial

Windows on Snapdragon support

- Copilot PC powered by the Windows on Snapdragon (WoS) platform, running Adreno GPUs
- Single solution for all platforms.

	Snapdragon X Elite		
Model	Decode (toks/sec) for 100	Encode for 256 prompt	
		toks / sec	Time To Fist Token (sec)
DeepSeek-R1-Distill-Llama-8B	16.8	88	2.91
DeepSeek-R1-Distill-Qwen-1.5B	54.8	413	0.62
DeepSeek-R1-Distill-Qwen-7B	17	95	2.69
Llama-2-7b-chat-hf	20	104	2.45
Meta-Llama-3-8B-Instruct	17	95	2.7
gemma-2b-it	41	330	0.77
Mistral-7B-Instruct-v0.2	16.3	75	3.3
phi-2	42.8	203.5	1.25
Phi-3-mini-4k-instruct	33.5	169.3	1.5
Qwen-7B-Chat	15.7	101.7	2.5
llava-1.5-7b-hf	21	119	2.15

*** Performance numbers are produced under ideal conditions. There might be a slight difference when reproduced on different commercial devices.*

Upcoming

OpenCLML in action for Generative AI

- Ops in prompt processing are ALU intensive.
- We know how to accelerate leveraging our deep understanding of Adreno GPU.
- OpenCLML is going to bring in accelerated API for these ALU intensive ops.
- Near future release of Open CLML will have significant performance boost for prefill.

Towards Relax

- TVM community has retired Relay and Relax is the future.
- MLC is built over Relax.
- Focused towards bringing in all the support we used to have in Relay to Relax.
- Early CLML offload is already mainlined.

Vulkan ML Backend

- TVM already supports Vulkan backend.
- Not efficient enough for Adreno GPU.
- We are going to improve it to deliver similar performance as OpenCL.
- Thanks to TVM's target independent nature that accommodates majority of optimizations being common for both backends.

Disclaimer

- *OpenCL and the OpenCL logo are trademarks of Apple Inc. used by permission by Khronos.*
- *Vulkan and the Vulkan logo are registered trademarks of the Khronos Group Inc.*
- *SYCL and the SYCL logo are trademarks of the Khronos Group Inc.*
- *Khronos and the Khronos Group logo are registered trademarks of the Khronos Group Inc.*

Thank you

Follow us on:

For more information, visit us at qualcomm.com & qualcomm.com/blog