14th International Workshop on OpenCL and SYCL

IWOCL 2026

# Opensource Deep Learning Compiler Powering Gen AI on Adreno™ GPU

Siva Rama Krishna Reddy, Qualcomm India Private Limited.

Krishna Raju · Deepanshu Singh · Sanjay Krishnaa · Raman Shinde

May 6-8, 2026 | Heilbronn, Germany | iwocl.org

KHRONOS GROUP

Qualcomm

For 14th International Workshop on OpenCL and SYCL (2026 IWOCL)

# Opensource Deep Learning Compiler Powering Gen AI on Adreno™ GPU

**Speaker: Siva Rama Krishna Reddy**

Krishna Raju · Deepanshu Singh · Sanjay Krishnaa · Raman Shinde

Qualcomm India Private Limited

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries.
Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries.

# Agenda

- About TVM and MLC
- Adreno GPU backend support in TVM
- MLC-LLM support status for Adreno GPU
- Model support & Performance
- Upcoming

# Who are we ?

## Team

Part of Qualcomm GPU Research Team (GRT) from Bangalore

- Work on GPGPU and AI/ML projects for Adreno GPUs,
- Focus on open-source AI solutions for Adreno GPU
- We largely contribute to TVM and MLC communities.
- Long term contributors for IWOCL

## Key contributors for this project

Sr. Staff Engineer

Sr. Lead Engineer

Engineer Senior

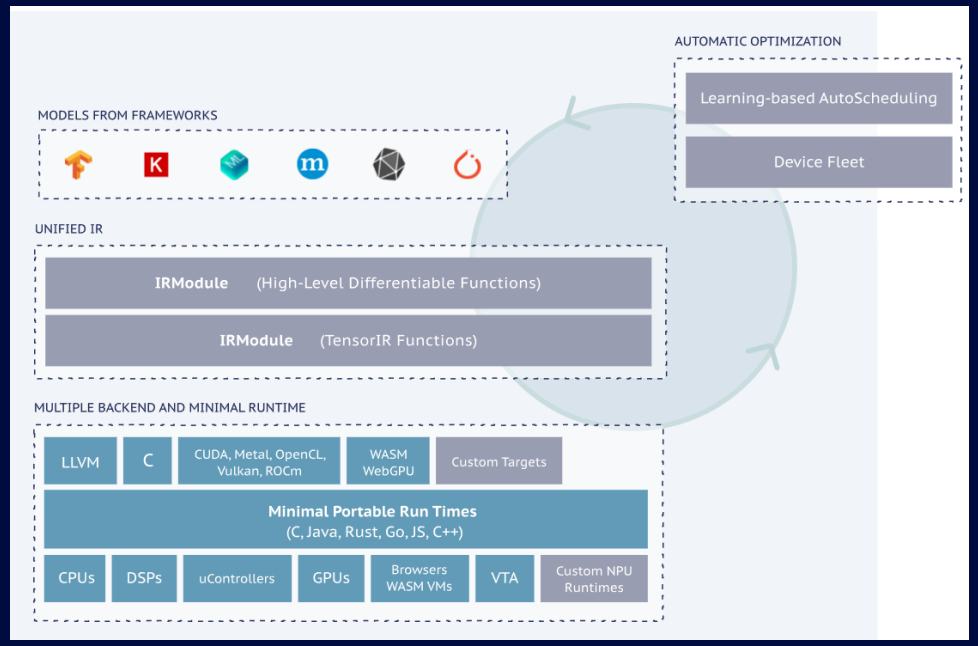
Engineer

# Resources/References on this project

- TVM (Tensor Virtual Machine) : https://tvm.ai
- MLC LLM (Machine Learning Compilation) : https://llm.mlc.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

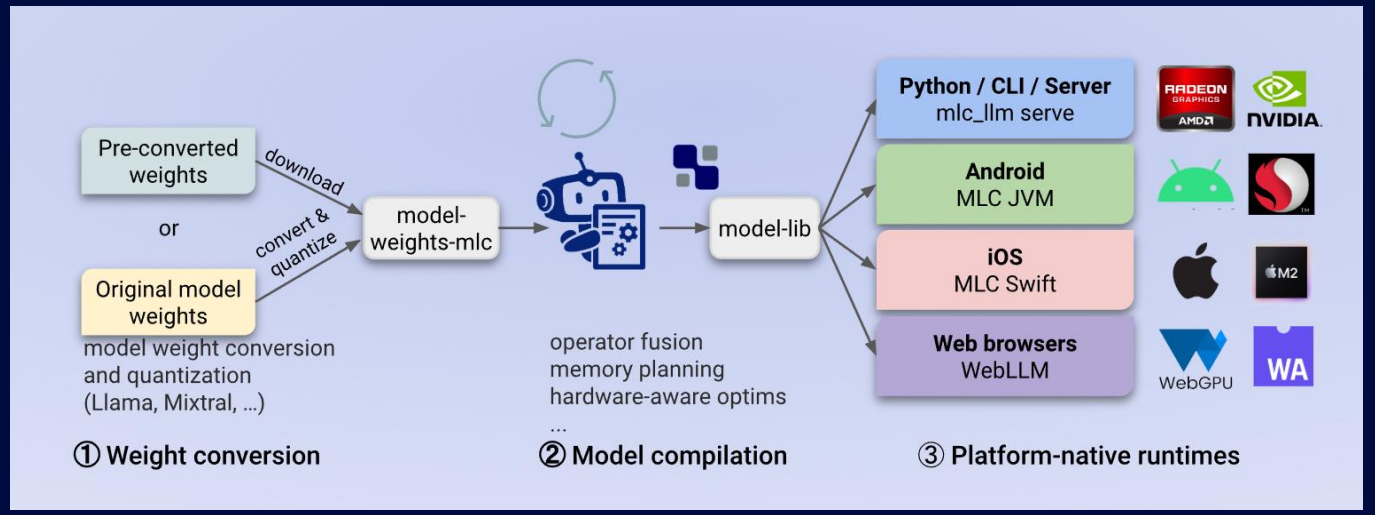
# What is TVM and MLC

## tvm.ai




- An Open Machine Learning Compiler Framework.
- Powerful with features like python first, frontend support, target independent IR, tuning support, vendor integrations using BYOC, wide range of codegen and target support.
- Qualcomm is contributing to OpenCL backend for many years for Adreno GPU

## llm.mlc.ai




- Universal LLM Deployment Engine With ML Compilation
- From the early days we have adopted this project and enhancing for Adreno GPU.
- Can leverage all optimizations of TVM for Generative AI models.

# Adreno GPU  backend support for TVM

## Generic Contributions

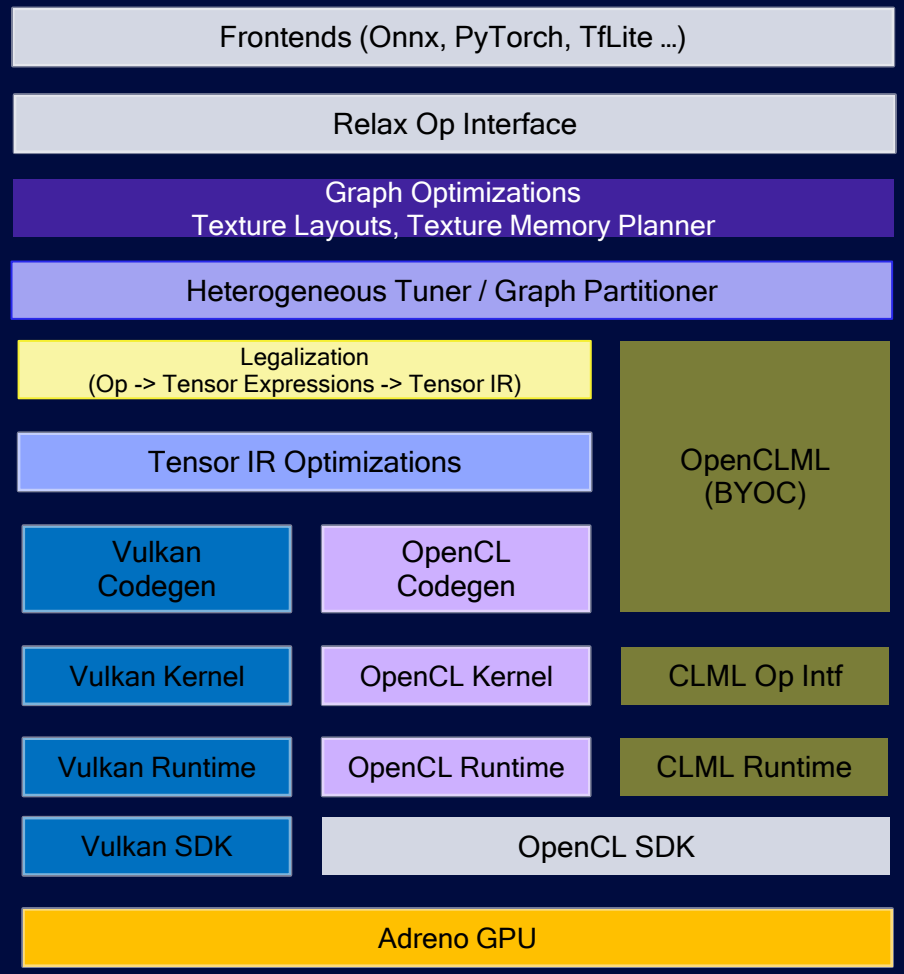
- Dedicated target registry for Adreno GPU - qcom/adreno-*
- Adreno GPU components are moved under python/tvm/relax/backend/adreno
- Specialized Relax pipeline for Adreno GPU.
- TFLite frontend – revived from Relay
- Dynamic Layout support – Sub-indexing like NCHW[4c4n].

## OpenCL

- Native Texture support across Graph, TIR and Runtime
- Use clImage backed by clBuffer for optimal memory usage
- Support for Qualcomm extensions
- Memory mapping (clEnqueMapBUffer) for Gen AI KV Cache zero-copy.
- OpenCLML Offload – v5.0 for Gen AI. OpenCLMLOffLoadForLLM to offload optimized Dequant-Matmul to CLML.

## Vulkan

- OpenCL comparable functional support with Texture support – PR in progress.
- Shader level profiling – Vulkan Timer
- Cooperative Matmul & Conv2D – Ready to upstream.



# MLC-LLM support status for Adreno GPU

- MLC LLM supports models from various researchers like Meta, Google, Qwen, Mistral, OpenAI, Qwen, Deep seek and Microsoft.
- 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.

## Adeno GPU Enhancements

- Android & Windows (A64, X64) target support.
- Adreno Schedules
- Gated Delta Networks (GDN) optimization for Adreno GPU
- Vulkan co-op support for Dequant-MatMul, Flash-Attn & MoE
- A wide range of  MoE models also supported.

# Upcoming

## TVM

- Vulkan cooperative matrix upstreaming – soon
- Rtvm – Android cli from Relay
- Documentation for Windows platforms (X64, A64) covering Vision and Gen AI
- Meta scheduler specialization for Adreno GPU
- OpenCL cooperative matrix support – along with Qualcomm official extn

## MLC

- Schedule changes for Adreno GPU  - Upstream
- mlc_cli – Cli for no python environments

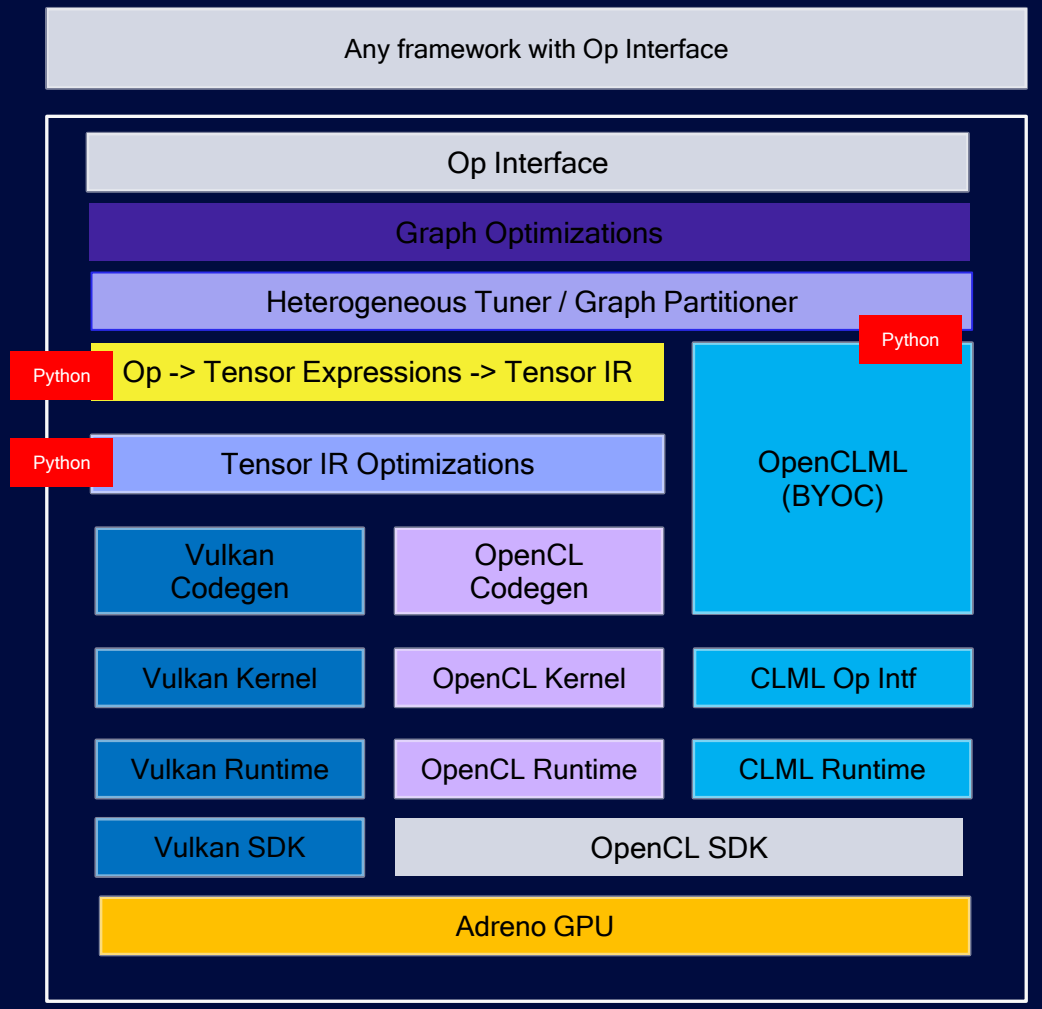
# Online Compiler – Towards Cpp

## What is done

- Complete legalization is in cpp.
- OpenCLML offload infra (Partitioner & passes) in CPP
- Adreno GPU Texture schedules and generic gpu schedules in cpp now.

## TODO

- Upstreaming
  - CPP Ports alone into apache-tvm ?
  - Contrib cpp-compiler ?
  - New project with tvm as 3rdparty ?
- Framework Integrations - tvm-cpp backend for Llama.cpp ?



# 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.*
- *TVM and MLC logo is registered trademarks of respective communities.*

# Thank you

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

© Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm, Snapdragon, and Adreno 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:

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