Date 10.05.2022

sivb@qti.qualcomm.com

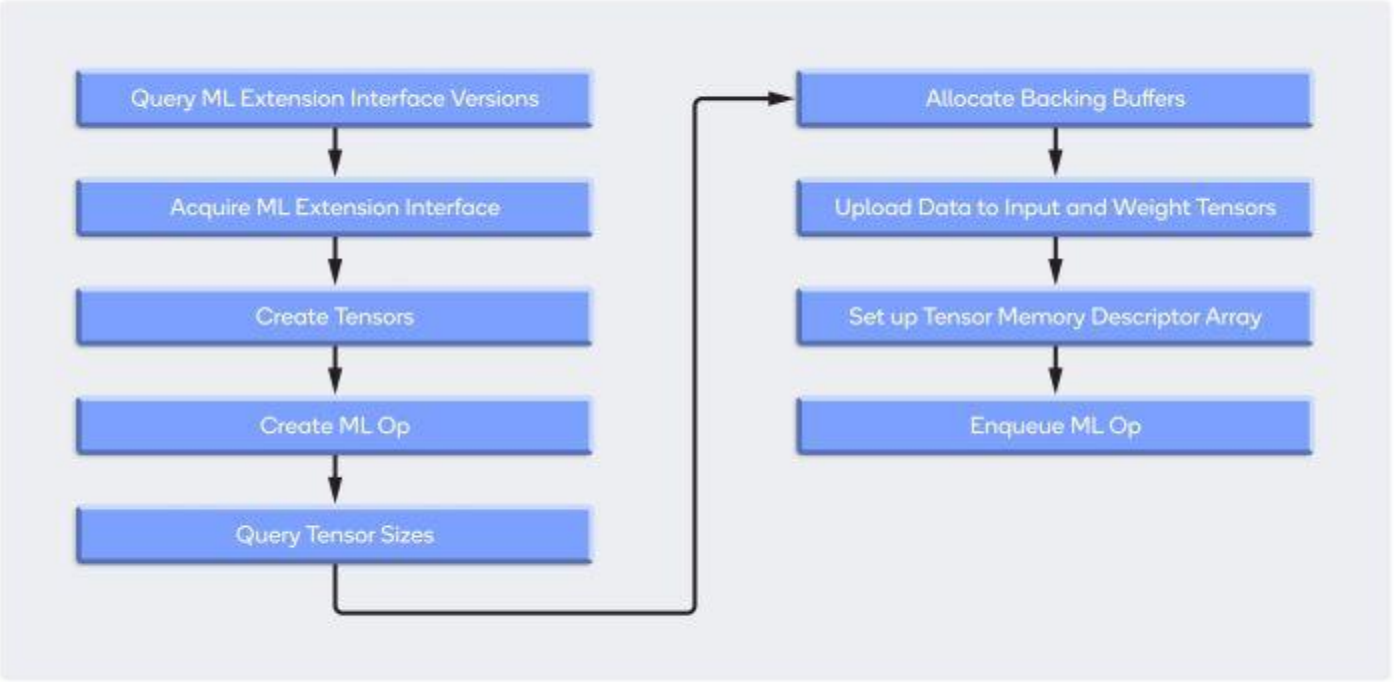
Qualcomm

OpenCLML Integration with TVM

Siva Rama Krishna Reddy .B, Hongqiang Wang, Alex Bourd, Adarsh Golikeri and Balaji Calidas

About OpenCL ML

- An OpenCL extension (cl_qcom_ml_ops) that accelerates Machine Learning at the Op level.
- Leverages deep knowledge of the Adreno GPU for significant performance benefits.
- C based DNN API with compatibility to most of the standard frameworks.
- Uses standard OpenCL features such as command queues, buffers, events. Supports FP16 and FP32 data types.
- Can be interleaved with other OpenCL kernels (i.e. TVM generated kernels) and dispatched to the same command queue.
- Compatible with existing OpenCL extensions for importing memory, controlling performance and controlling data access.



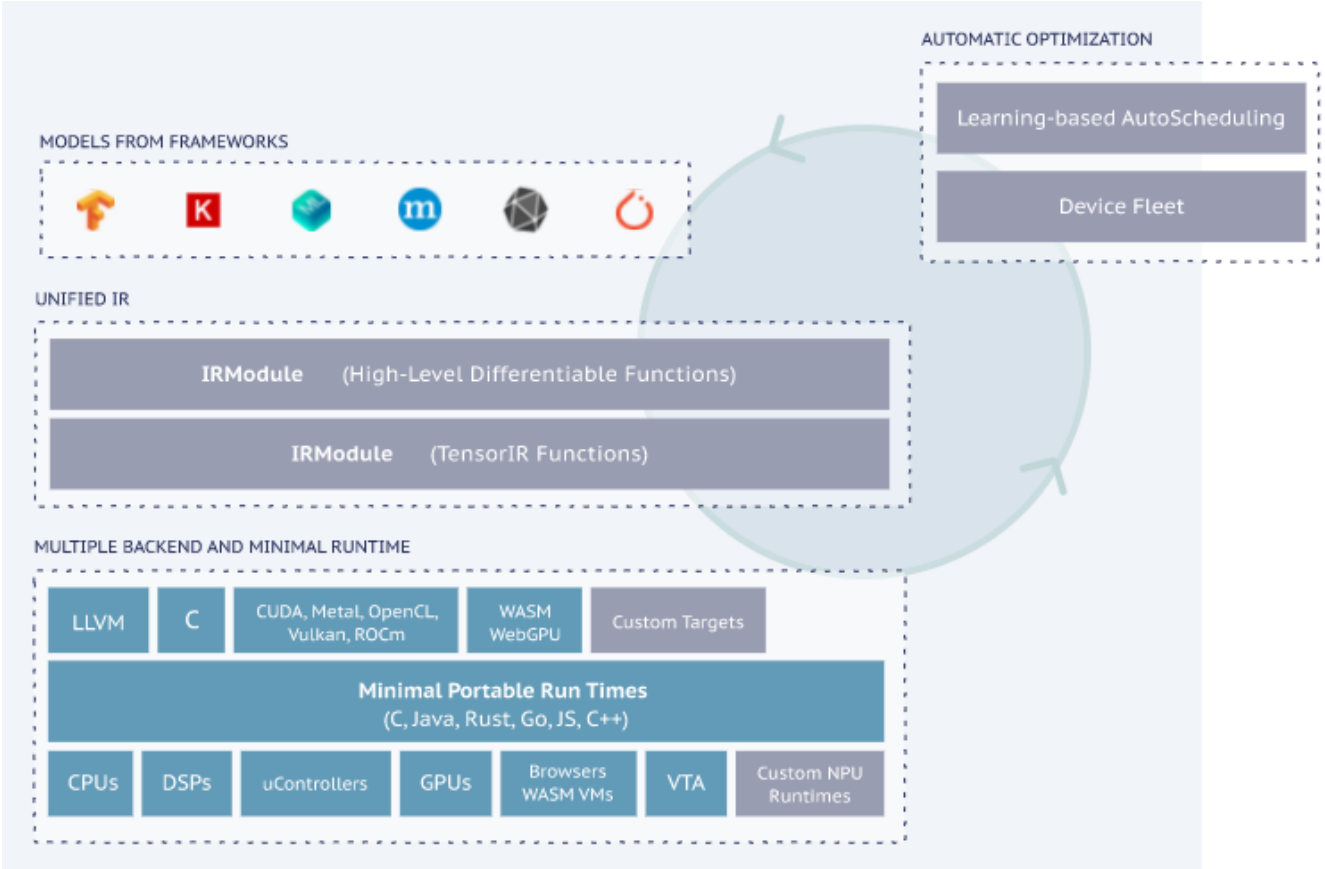
https://developer.qualcomm.com/blog/accelerate-your-models-our-opencl-ml-sdk

- Download the SDK at https://developer.qualcomm.com/blog/accelerate-your-models-our-opencl-ml-sdk
- SDK documentation helps with API details, Data layout information and other tools that helps with model conversion from Tensorflow or Tensorflow Lite.

Introduction

TVM (Tensor Virtual Machine)

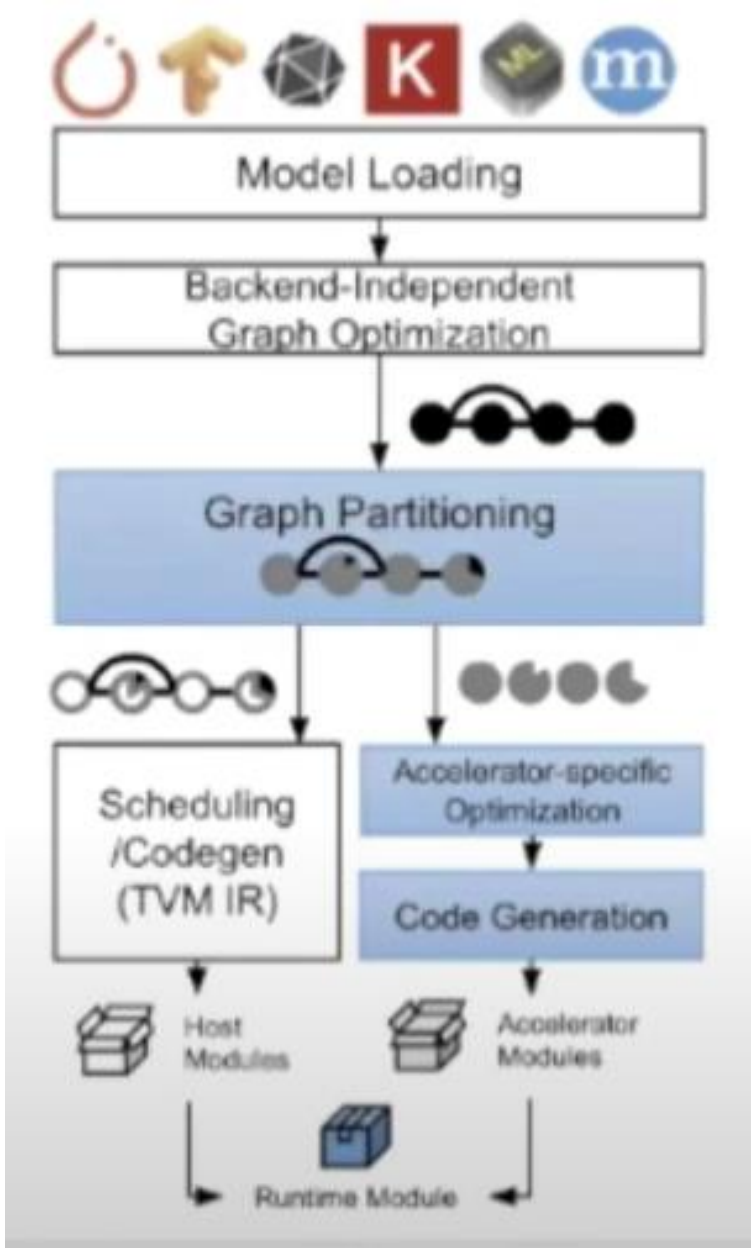
- An Apache project (https://tvm.apache.org/).
- Compiles deep learning models from various frames into minimum deployable modules.
- TVM Features
  - Frontend framework support.
  - Range of hardware.
  - Operating system and programming language compatibility.
  - AutoTVM
  - Auto Scheduler
  - Bring Your Own Codegen (BYOC)



https://tvm.apache.org/

BYOC (Bring Your Own Codegen)

- Graph Partitioning
  - Pattern based partitioning
  - Operator based partitioning
- Greedy based partitioning
- Reuse all graph level optimizations offered by TVM
- Accelerator specific optimizations
  - Constant processing
  - Subgraph pruning
  - Calibration & Quantization
- Codegen flow
  - TVM general codegen flow
  - Accelerator codegen flow



Ref. https://www.youtube.com/watch?v=DD8GdZ_OKco

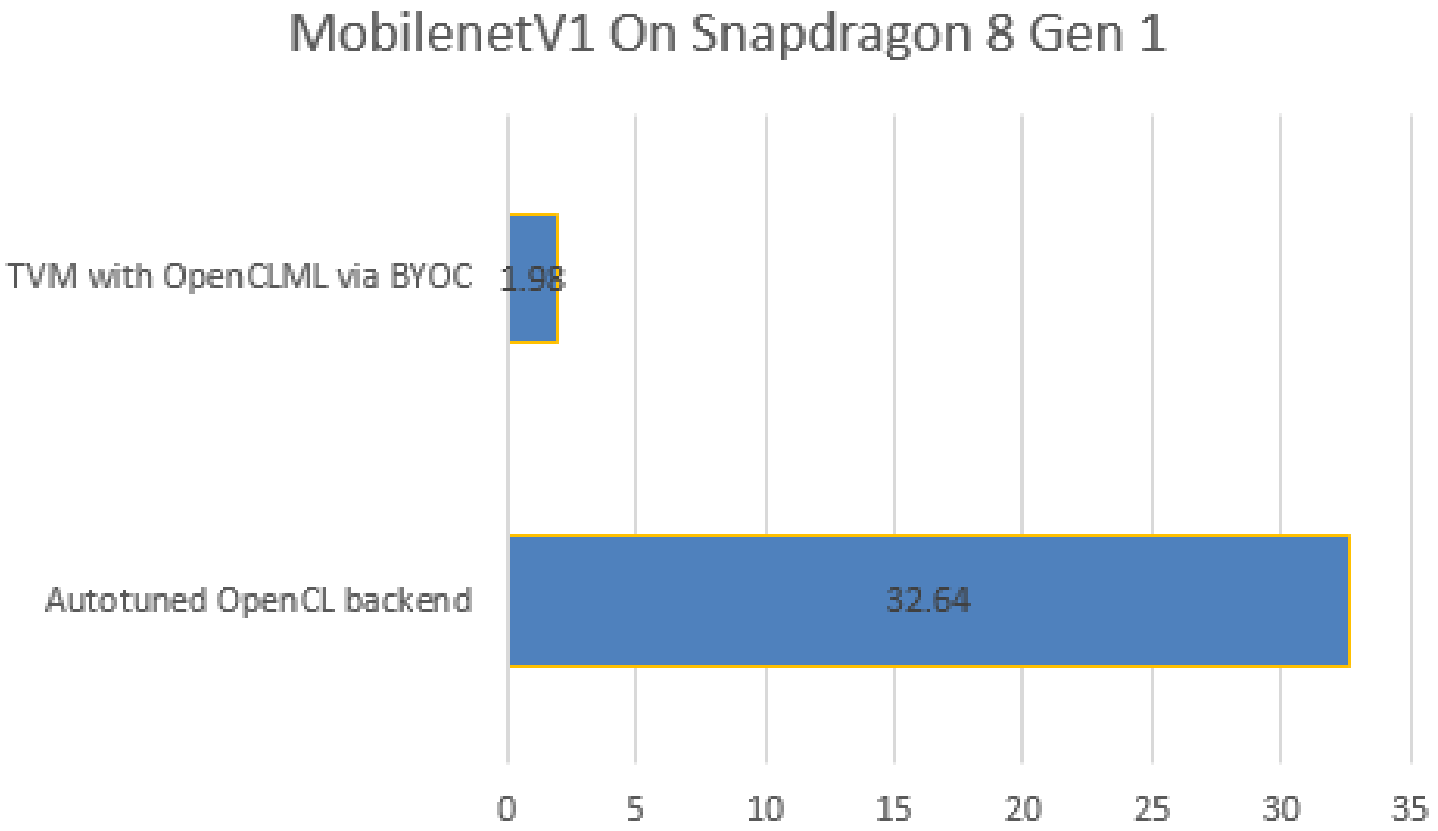
Motivation

- OpenCL backend of TVM is generic and uses standard specification only.
- OpenCL ML offers Adreno GPU accelerated ML ops via OpenCL extension interface.
- TVM has an end-to-end compiler framework with strong frontend support and well-defined high-level graph optimization passes.
- TVM's BYOC approach allows offloading parts or full graph to execute via vendor acceleration library with a fallback option on TVM's default runtimes.

"With TVM having the entire framework of frontends, graph level optimizations and OpenCL ML having kernels that perform best on Adreno GPU, in this work we aim to integrate OpenCLML extension into TVM as a BYOC. This effort brings best of both worlds where TVM handling high level optimizations, sub graphs are scheduled on OpenCL ML based on the support and the operators not supported by OpenCL ML will take TVM's default OpenCL path"

Mobilenet Performance

On Snapdragon 8 Gen 1 here is a comparison between TVM with OpenCL backend auto tuned vs TVM OpenCLML backend (with completed network offloaded to OpenCLML path)



** Mobilenet V1 is hand crafted to make sure the entire models does as one sub graph on OpenCLML path. The same is tuned for TVM + OpenCL path.*

Upstreaming Status

- RFC: https://github.com/apache/tvm-rfcs/blob/main/rfcs/0052-OpenCLML-integratio-as-BYOC.md
- Pull Request: https://github.com/apache/tvm/pull/10243 (In progress)

Conclusions and Future work

- Enhance OpenCLML operator coverage and network support.
- Will keep OpenCLML version support up to date with release cycles.
- OpenCLML SDK 2.0 supports DNN training as well.

Qualcomm

Thank you!

Follow us on:

For more information, visit us at:

www.qualcomm.com & www.qualcomm.com/blog