Date: 15.12.2021

sivb@qti.qualcomm.com

TVM at Qualcomm Adreno

Siva

TVM journey with Adreno GPU

Evaluation:

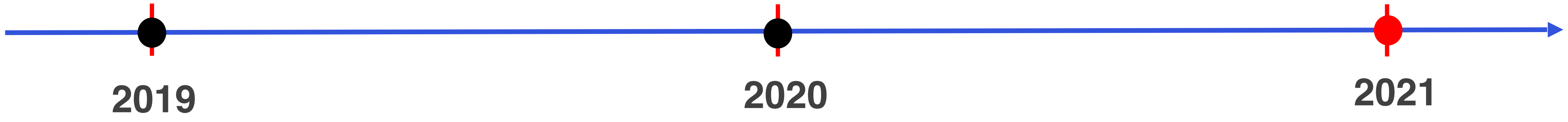
- Basic 1D texture base experiments.
- Auto tuning & Cache strategy
- **2.56x** speedup over vanilla TVM.
- Published with IOWCL (https://www.youtube.com/watch?v=jedW0cjNTDk)

OctoML Collaboration:

- More enhancements for Adreno.

Other Initiatives:

- OpenCL ML with TVM.
- TVM backend for MLPerf.



2019

2020

2021

OctoML Collaboration:

- https://github.com/octoml/qualcomm

DNN Training Evaluation:

- TVM enhancements to enable DNN training.
- Mobilenet V1 training over Adreno is functionally working.
- Published with IOWCL (https://www.youtube.com/watch?v=6pYV7T-Jzi8)

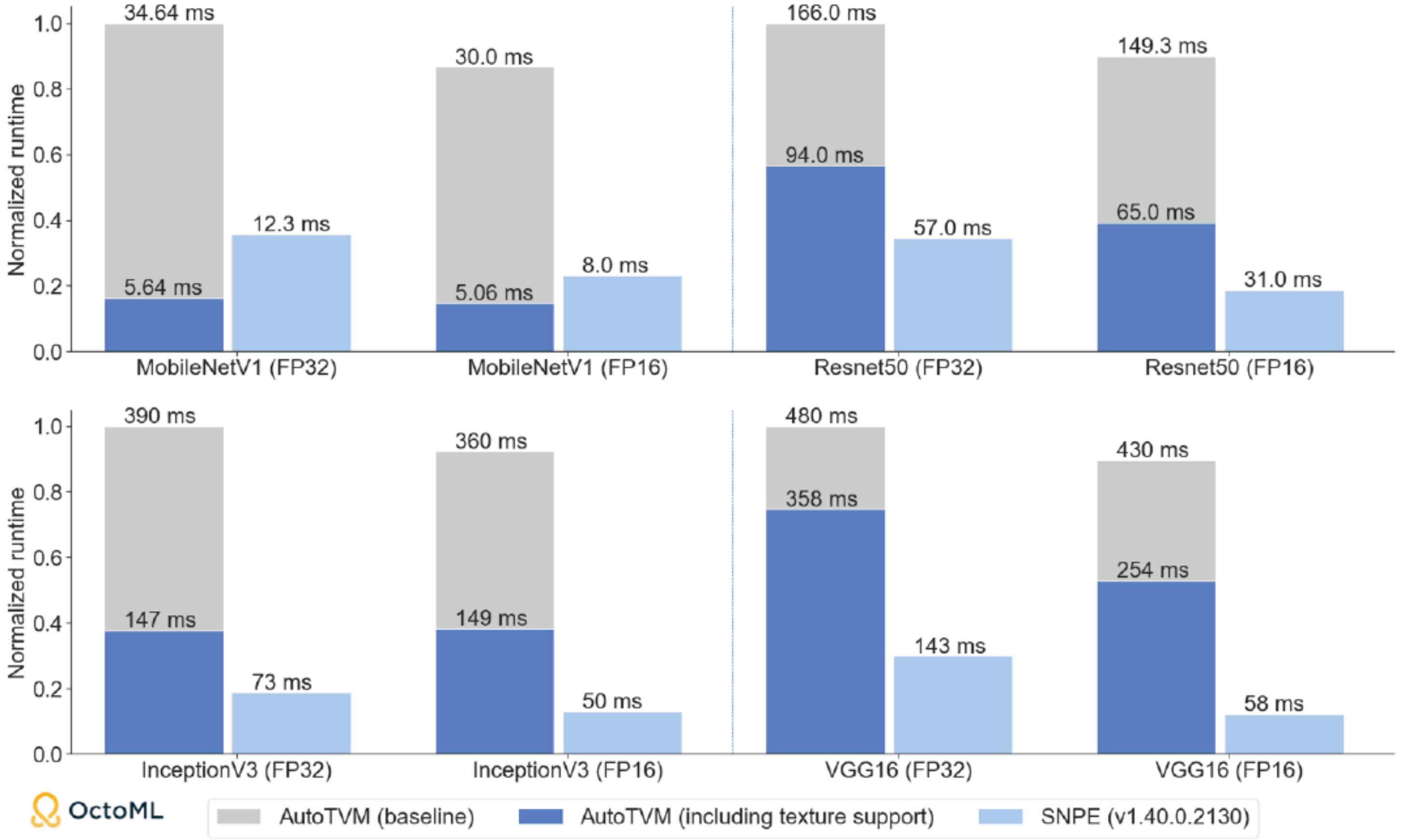
OctoML Collaboration

Enhancements:

- OpenCL image objects are implicitly backed by texture cache.
- Friendly layouts to take advantage of OpenCL vectorization.
- Brand new schedules to drive the codegen with these changes.
- Finally, the magic of AutoTVM to bring out the best possible kernels.

Thanks Thierry Moreau, Chris Sullivan and OctoML Team for making it happen

Normalized runtime performance on Adreno 650 for MobilenetV1, ResNet50, InceptionV3, and VGG16 including FP16 compute and FP16/FP32 accumulation



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 like command queues, buffers, events and 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.

➢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.

OpenCL ML into TVM via BYOC

Efforts:

- Frontend to transform and offload the subgraphs to OpenCL ML path.
- Codegen extended over existing JSON Codegen.
- OpenCL ML runtime for subgraph execution.
- OpenCL workspace reuse across CLML and default OpenCL runtimes.

Plan:

- OpenCL ML SDK 2.1 with more operators and enhancements is planned for release soon.
- Snapdragon 8 Gen 1 devices would be available across vendors in coming months.
- We are working on a contribution plan to land this feature into community.

TVM backend for MLPerf

About MLPerf:

- Driven by mlcommons community (https://mlcommons.org/en/)
- Has got Android APK (https://github.com/mlcommons/mobile_app_open) that can evaluate the platform performance for various use cases like Image Classification, Object detection, Image Segmentation and Language Understanding.
- Uses well standard datasets to evaluate the models.

Efforts:

- Generic TVM backend inline with MLPerf's backend interface definition.

Q&A

Qualcomm

Thank you!

Follow us on: f 🐦 in

For more information, visit us at:

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

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

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 Qualcomm’s licensing business, QTL, and the vast majority of its patent portfolio. Qualcomm Technologies, Inc., a wholly-owned subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of Qualcomm’s engineering, research and development functions, and substantially all of its product and services businesses, including its semiconductor business, QCT.