TVM for micro targets

What's new on microTVM

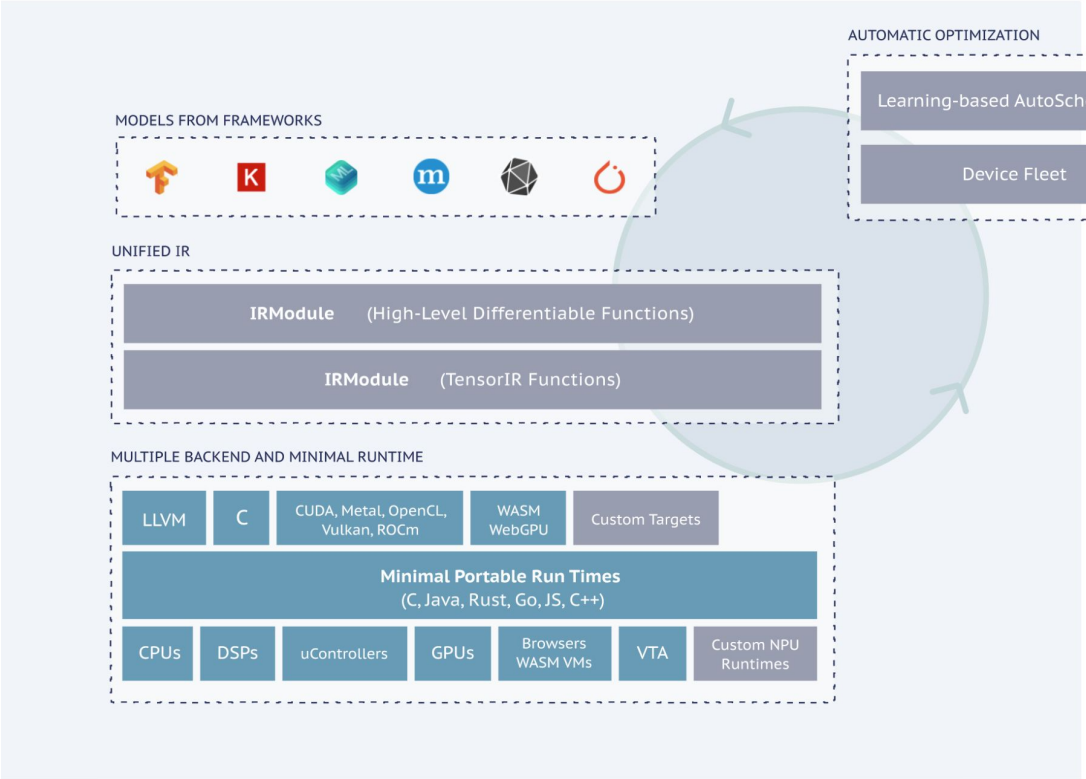
Gustavo Romero
gustavo.romero@linaro.org

Linaro Connect
VIRTUAL • FALL 2021

A bit of context about Apache TVM

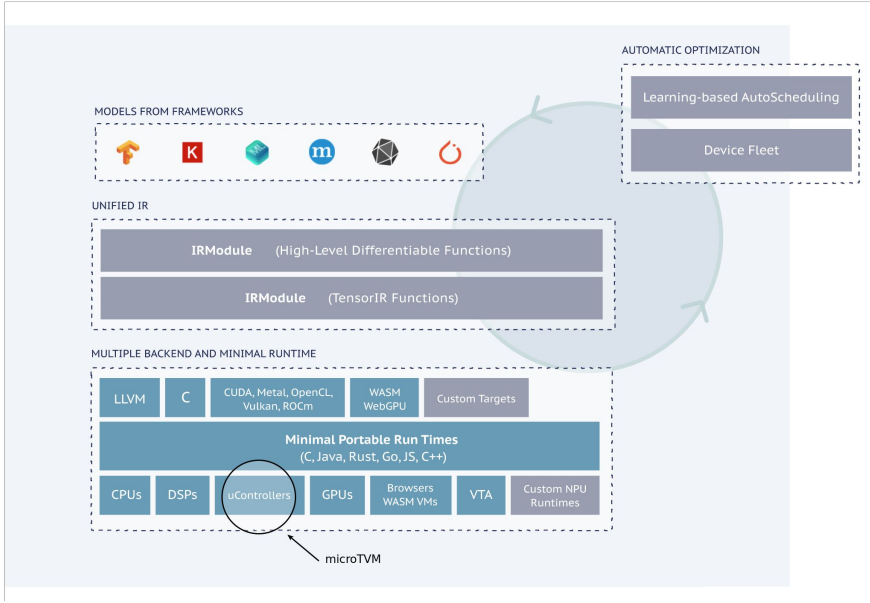
- Apache TVM is a compiler framework for compiling deep learning models for various target devices
- TVM targets CPUs, GPUs, NPUs, browsers, FPGAs, and microcontrollers
- It's not used to train a NN
- However, once you have a trained model TVM will automatically generate and optimize tensor operators for running your model on the specified target
- TVM accepts as input models in various formats, like Keras, MXNet, PyTorch, Tensorflow, Tensorflow Lite, CoreML, DarkNet, ONNX, etc

TVM Architecture



microTVM

- TVM runs on microcontrollers, which are generally quite resource constrained devices, hence TVM has a specific runtime for them (microTVM)
- There are two kinds of so-called 'executors' available on microTVM: *host_driven (graph)* and *aot_driven*
- *host_driven* executor is used for autotuning
- microTVM runs on top of platforms, where platforms can be a RTOS (like Zephyr) or a platform like Arduino. Each platform supports different devices (MCUs).



What's new on microTVM?

- New boards added to microTVM/Zephyr platform:
  1. mps2_an521 (Arm-Arm Cortex-m33)
  2. nRF5340 DK (Nordic-Arm Cortex-m33)
  3. ZynqMP | QEMU (Xilinx-Arm Cortex-r5)
  4. RISC-V 32 and 64 | QEMU

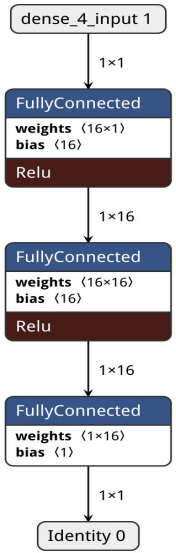
- New platform: Arduino (ARM32 and ESP32 boards)
- AOT runtime
- MLF (Model Library Format)
- **Project API - generate_project, build, flash, and transport (read/write)**

Project API + TVMC

- Currently TVMC (TVM's cli tool) doesn't support micro targets
- Hence if one wants to run a model on a micro target it's necessary to write some Python code to achieve it
- However Project API now eases the integration of TVMC with micro targets, so it can now be done entirely from the command line
- Demo: run a model on a micro target using only the cli

Demo

- Build, flash, and run a simple inference model to infer sin(x) using TVMC
- STM32F746G Discovery board attached to the USB



MODEL PROPERTIES

format	TensorFlow Lite v3
description	TOCO Converted.

INPUTS

dense_4_input1	name: **dense_4_input**
	type: **float32[1,1]**
	quantization: **0 ≤ q ≤ 255**
	location: **1**

OUTPUTS

Identity0	name: **Identity**
	type: **float32[1,1]**
	location: **0**

Demo

$ wget https://people.linaro.org/~tom.gall/sine_model.tflite

$ tvmc compile ./sine_model.tflite --target="c -keys=cpu -link-params=0 -march=armv7e-m -mcpu=cortex-m7 -model=stm32f746xx -runtime=c -system-lib=1" --output sine_model.tar --output-format mlf --pass-config tir.disable_vectorize=1 --disabled-pass="FuseOps,AlterOpLayout"

$ tvmc micro create-project /tmp/x22 ./sine_model.tar zephyr --project-type host_driven --board stm32f746g_disco

$ tvmc micro build /tmp/x22 zephyr --board stm32f746g_disco

$ tvmc micro flash /tmp/x22 zephyr --board stm32f746g_disco

$ tvmc run /tmp/x22 --device micro --print-top 1

How to get involved

Source - https://github.com/apache/tvm

Contributor Guide - https://tvm.apache.org/docs/contribute/

Discord - https://discord.com/invite/77Hh4jVhbM

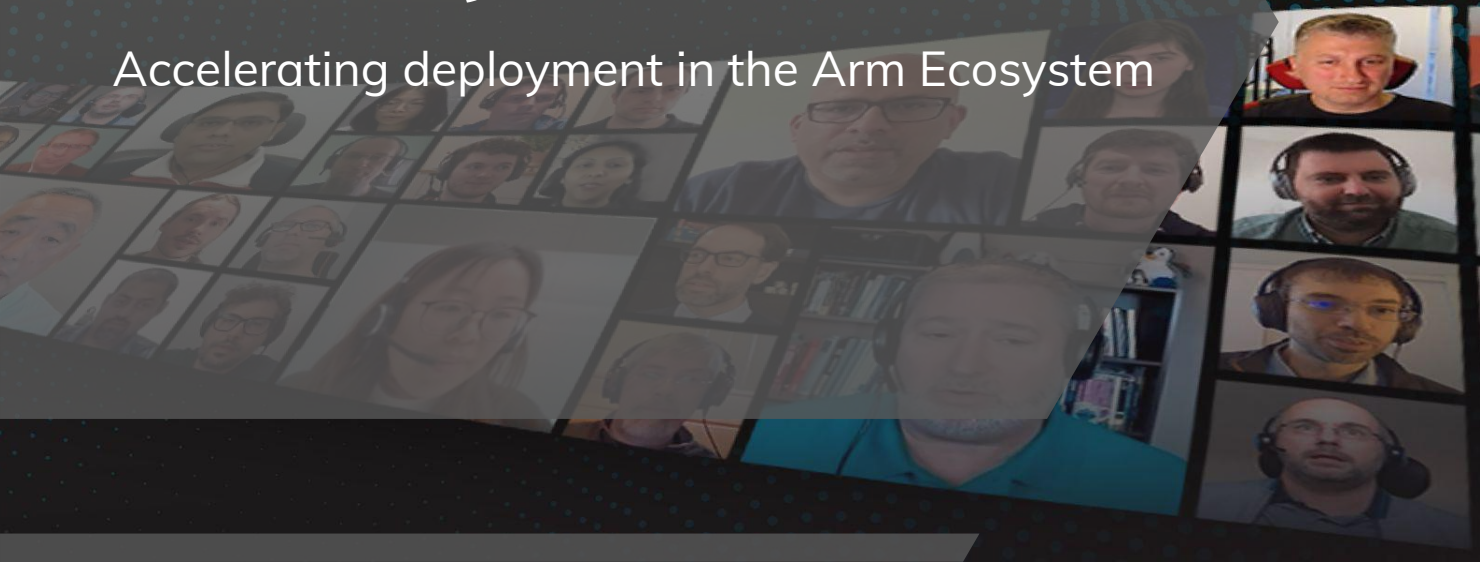
Forum - https://discuss.tvm.ai

Bugs / Issues - https://github.com/apache/tvm/issues

More info about TVM Conf., Calendar, etc, please see: https://tvm.apache.org/community

Thank you

Accelerating deployment in the Arm Ecosystem



Linaro Connect
VIRTUAL • FALL 2021