Developing and testing TensorFlow Lite Micro edge AI algorithms on RISC-V and FPGAs

London Open Source Meetup for RISC-V, 2021-01-18
Michael Gielda, mgielda@antmicro.com
ANTMICRO

- Founded in 2009
- Building open source platforms and helping customers develop software-driven products
- Industrial IoT and embedded systems: AI/ML in defense/security, mining, agriculture, autonomous vehicles, robotics, aerospace, industrial automation
- Member of Linux Foundation, Zephyr Project, CHIPS Alliance, OpenPOWER Foundation, Strategic Founding member of RISC-V International
- Introducing new design methodologies and workflows based on open source
WHAT WE DO

See our technology showcase on antmicro.com
WHAT WE DO

**HARDWARE**
Proof of Concepts (PoC), demonstrators, prototyping, open source platforms

**SOFTWARE & AI**
OS porting, building BSPs, build systems, device management, edge & cloud AI

**FPGA & ASIC**
Custom IP blocks, SiP development, soft SoCs, heterogeneous processing systems

**TOOLS**
Tools, new software and hardware development and testing methodologies
WHAT WE DO

HARDWARE
Proof of Concepts (PoC), demonstrators, prototyping, open source platforms

SOFTWARE & AI
OS porting, building BSPs, build systems, device management, edge & cloud AI

FPGA & ASIC
Custom IP blocks, SiP development, soft SoCs, heterogeneous processing systems

TOOLS
Tools, new software and hardware development and testing methodologies
TINY ML - BIG OPPORTUNITIES, BIG CHALLENGES

- Power/performance/memory constraints
- Sourcing hardware
- Testing at scale
- Configuration of complex systems of devices
- Tedious manual testing procedures
- Repeatability/Determinism
RENODE CAN HELP

- Our tools empower organizations like Google, Arm, QuickLogic, Microchip with test-driven open source methodologies of developing TinyML
- Google and Arm use Renode to develop and test TensorFlow Lite
- QuickLogic embraced open FPGA tools, Renode and other open technologies from Antmicro
- Microchip hired us to develop their pre-silicon development platform for RISC-V based PolarFire SoC with Renode
Develop your IoT product with Renode:

GET STARTED
A BIT OF HISTORY

- Simulation framework developed by Antmicro since 2010
- Created in response to our internal needs
- Open source since 2015
- In 2020 used by Arm, NXP, Google, Zephyr, Systerel, QuickLogic, Microchip, Amazon, a number of startups and universities including Poznan University of Technology
Use cases

- Development of complex software for embedded and IoT systems
- Architectural exploration and research
- Pre-silicon prototyping and HW-SW co-development
- Co-simulation with FPGA
Features & use

- Plug-and-play building blocks
- “Batteries included” - lots of demos and binaries
- Flexible, deterministic and software-agnostic
- Continuous Integration-oriented - Robot Framework, Jenkins, GitLab CI, GitHub Actions
- Enables e.g. protocol / stack testing: OPC-UA, TSN, 6lowpan, Thread etc.
Simulate full systems / real devices
Peripherals & sensors

Our focus is not just on SoCs, but also I/O peripherals
- UART, SPI, I2C, RAM, ROM, GPIO, CAN,
  ETH, I2S, PCIE...

Also sensors:
- Thermometers, humidity meters, accelerometers,
  microphones, etc.

This allows end-to-end machine learning
processing in simulation with real and synthetic data
Platform description format

- Human readable
- Modular
- Extendible
- Enable new boards / platforms w/o coding
Platform description format

uart: UART.MiV_CoreUART @ sysbus 0x70001000
  clockFrequency: 66000000

cpu: CPU.RiscV @ sysbus
  cpuType: "rv32g"

plic: Interrupts.PlatformLevelInterruptController @ sysbus 0x40000000
  IRQ -> cpu@1
  numberOfSources: 31 //based on release notes
CO-SIMULATION WITH HDL SIMULATOR

- Provides an integration layer for Verilator
- Enables HW/SW co-simulation with Verilator
- No need to create models of the IP you can “verilate”
- Supports AXI4Lite and Wishbone, full AXI4 under way
- Support for interrupts and external interfaces, like UART Rx/Tx lines
CO-SIMULATION WITH FPGA

- Another alternative that can be used in FPGA co-development is emulating IP on real FPGA HW
- Majority of simulation (e.g. RISC-V SoC) is in Renode, only the part under development is in the FPGA
- Renode supports EtherBone protocol, to interact with Wishbone-connected soft IP
- Connection with the FPGA over a socket - possibly in a remote location
## Supported platforms/vendors

<table>
<thead>
<tr>
<th>LEON3</th>
<th></th>
<th>QuickLogic</th>
</tr>
</thead>
<tbody>
<tr>
<td>NORDIC</td>
<td></td>
<td>MICROCHIP</td>
</tr>
<tr>
<td>SILICON LABS</td>
<td></td>
<td>TEXAS INSTRUMENTS</td>
</tr>
<tr>
<td>SiFive</td>
<td></td>
<td>RISC-V</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OpenPOWER</td>
</tr>
</tbody>
</table>
Some supported RISC-V/FPGA boards
REAL-WORLD USAGE
ML ACCELERATION WITH QUICKLOGIC FPGA PLATFORM

- QuickFeather devkit is a tiny open hardware platform (ARM Cortex-M4F + FPGA) built for QuickLogic by Antmicro with TinyML in mind
- QuickLogic is the 1st ever FPGA vendor to support/sponsor open source FPGA tools
- Developed by Antmicro in joint project with Google
- QL also sponsored Renode simulation and Zephyr RTOS port
NEXT STEP: RISC-V + OPEN SOURCE FPGA

- Next gen platform with RISC-V and larger FPGA
- Also supported in open source FPGA tools developed by Antmicro
- Part of Open Hardware Group’s Core-V project, collaboration with e.g. NXP and Silicon Labs
- Will enable high-perf, low-power TinyML
- Renode support also in progress, co-sponsored by Google and QuickLogic
- Read more
GOOGLE / TF LITE AND ANTMICRO COLLABORATION

- Collaboration started in 2018
- Lots of joint work around open source ASICs, FPGA, software, ML
- Initially enabled running TF Lite on RISC-V PoC, presented at RISC-V Summit 2018
- First integration with Zephyr and co-marketing with Zephyr and RISC-V
- Resulted in a [note on TF Lite blog](#) demoing soft RISC-V MCU on Digilent Arty board
- Turned into large direct collaboration and adoption of Renode by the TF Lite team
THE ORIGINAL DEMO THAT BROUGHT US HERE

- Renode-simulated soft VexRiscv CPU + sensors
- Zephyr + TFLite integration
- Demo originally created for Arty board with ADXL345 accelerometer
RENODE IN EDUCATION

- To be used in Harvard University’s EdX course with over 16000 participants
- Renode also adopted at University of Minnesota, lots of interest from other unis
- PUT course in Collaboration with ST - make TinyML boards available to students virtually during pandemic despite lab closures
VEDLIOT - VERY EFFICIENT DEEP LEARNING IN IOT

- EC-funded edge ML oriented project, coordinated by Bielefeld University, 12 international members
- Aim: scalable, deep-learning capable ML pipelines, also in FPGA
- Antmicro’s role
  - open source RISC-V soft SoC infrastructure
  - Renode used as a simulation platform
- Read more
RENODE - TESTING & METRICS
Renode-based Continuous Integration workflow for IoT systems

- **Company Environment**
- **Local PC**
  - Interactive test and debug in Renode
  - Get help from colleagues
- **Commit code**
- **Develop with favorite IDE/compiler**
- **Tests pass?**
  - Yes: Merge changes
  - No: Go back to development
- **Push to server**
  - CI e.g. with Robot + Renode
  - Test with various configurations
- **Field tests / deployment**
Developing and testing TensorFlow Lite Micro edge AI algorithms on RISC-V and FPGAs

EXAMPLE CI

- You can look at the “original” demo repository - there’s a CI performing the whole “demo” procedure automatically
- A demo of how CI with Renode could work (also possible with GitLab CI, Jenkins, GitHub Actions... We work with all of them)
RENODE METRICS ANALYZER

- Collecting execution data from the simulation
- Representing data as graphs
- Currently supported:
  - Executed instructions
  - Memory access
  - Peripheral access
  - Exceptions
- Virtual and real-time stamps
MSC: TF LITE MACHINE LEARNING
ALGORITHM ANALYSIS

- MSc: Analysis of optimization method - quantization
- Experiments with CNN model architectures
- No instrumentation!
- Hard to get that kind of data from real hardware
- Very good for experimenting with e.g. various amounts of available memory

<table>
<thead>
<tr>
<th>Metric</th>
<th>2 pooling layers</th>
<th>1 pooling layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.966887417219</td>
<td>0.940397350993</td>
</tr>
<tr>
<td>Size</td>
<td>19616 bytes</td>
<td>137180 bytes</td>
</tr>
<tr>
<td>Execution time</td>
<td>2425.8 ms</td>
<td>20206.2 ms</td>
</tr>
<tr>
<td>Memory reads</td>
<td>4740 operations</td>
<td>35106 operations</td>
</tr>
<tr>
<td>Memory writes</td>
<td>888 operations</td>
<td>4777 operations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>Base TensorFlow model</th>
<th>Convert model</th>
<th>Quantized model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model size</td>
<td>94 744 bytes</td>
<td>19 616 bytes</td>
<td>8 896 bytes</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.970198690891</td>
<td>0.970198675497</td>
<td>0.94701986755</td>
</tr>
<tr>
<td>Execution time</td>
<td>N/A</td>
<td>2280.7 mc</td>
<td>121.9 ms</td>
</tr>
<tr>
<td>Memory reads</td>
<td>N/A</td>
<td>4406 operations</td>
<td>306 operations</td>
</tr>
<tr>
<td>Memory writes</td>
<td>N/A</td>
<td>834 operations</td>
<td>78 operations</td>
</tr>
</tbody>
</table>
RENODE + TF LITE IN 5 MINUTES: GOOGLE COLAB

```
[10] @timeout 120 tail -c+2 -f renode/uart.dmp | sed '/"Person score: ".*/q'
  tell_renode('q')
  expect cli(\'Renode is quitting\')
  shutdown_renode()

  Attempting to start Arducam
  Starting capture
  Image captured
  Reading 2375 bytes from Arducam
  Finished reading
  Decoding JPG and converting to greyscale
  Image decoded and processed
  Person score: 116 No person score: -116
```

```
[6] from renode.tools.metrics_analyzer.metrics_parser import MetricsParser
    init_notebook_mode(connected=False)
    parser = MetricsParser(\'renode/metrics.dump\')

configure_plotly_browser_state()
show_executed_instructions(parser)
```

Executed Instructions
RENODE + TF LITE IN 5 MINUTES: GOOGLE COLAB

- Install requirements
- Take a photo
- Convert the photo, required size $< 4096$ bytes
- Run a person-detection example with a captured photo in Renode
  - Renode metrics analysis
  - Section
RENODE + TF LITE IN 5 MINUTES: GOOGLE COLAB

- A (mostly) Python workspace in the cloud
- Built with Jupyter notebooks
- Allows you to run arbitrary scripts and share the results online
- Fantastic tool for presentation purposes
- Very popular among students and academics
- Together with Google we implemented colabs which can process video and audio recorded from your computer
- To be used in EdX course with over 16000 participants
- See example colab
TOWARDS ML FLEXIBILITY

- With the Core-V-MCU support, Renode introduces custom instruction and register support
- Core-V-MCU and CV32E40P - Pulpissimo-based platform with many custom extensions
- Ideal for ML accelerator prototyping

```
RegisterCSR((ulong)CustomCSR.HardwareLoop1End, LoopEnd);
RegisterCSR((ulong)CustomCSR.HardwareLoop1Counter, LoopCounter);

InstallCustomInstruction(pattern: "FFFFFFFFFFFFBBBBB000DDDDD0001011",
handler: opcode => LoadRegisterImmediate(opcode, Width.Byte,
BitExtension.Sign, "p.lb rD, Imm(rs1!)"));

InstallCustomInstruction(pattern: "FFFFFFFFFFFFBBBBB100DDDDD0001011",
handler: opcode => LoadRegisterImmediate(opcode, Width.Byte,
BitExtension.Zero, "p.lbu rD, Imm(rs1!)"));
```
TOWARDS ML FLEXIBILITY

- For fast prototyping - Python hooks support
- No need for recompilation - write them directly in a script
- Access to all CPU details

```python
set $xadd

""
src_reg_a = instruction & 0xF
src_reg_b = (instruction >> 12) & 0xF
state['res'] = res
""

sysbus.cpu InstallCustomInstructionHandlerFromString
"1011001110001111bbbb11111000aaaa" "$xadd"
```
<table>
<thead>
<tr>
<th>Project Name</th>
<th>Status</th>
<th>Last Run</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>LITEX MICROPYTHON TUTORIAL TEST</td>
<td>Pass</td>
<td>2021-01-10 04:19:54</td>
<td>0:04:07</td>
</tr>
<tr>
<td>ZEPHYR TSN/GPTP ON SAM E70</td>
<td>Pass</td>
<td>2020-12-06 04:21:21</td>
<td>0:10:04</td>
</tr>
<tr>
<td>STM32F7 MBED OS HELLO WORLD EXAMPLE</td>
<td>Pass</td>
<td>2020-11-13 14:03:40</td>
<td>0:02:45</td>
</tr>
<tr>
<td>RPL UDP IN CONTIKI-NG ON CC2538DK</td>
<td>Pass</td>
<td>2020-12-20 04:19:55</td>
<td>0:03:54</td>
</tr>
</tbody>
</table>
## Invocation

815ae119-2110-4606-8836-51596c15880d (January 18th, 2021, 5:34:26 pm)

179 targets evaluated on January 18th, 2021 at 5:34:26 pm for 20.5 m

Evaluation started by distant on runner-53b050fc-project-11-concurrent-1

<table>
<thead>
<tr>
<th>Invocation</th>
<th>Targets affected</th>
<th>Broken</th>
<th>Failed</th>
<th>Failed (non-critical)</th>
<th>Successful</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Succeeded</strong></td>
<td>179</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>177</td>
</tr>
</tbody>
</table>

### Targets

#### 2 targets failed (non-critical)
- QuarkC1000.Should Serve Webpage Using Tap
- SiFive-FU540.Should Generate Proper PWM Pulses

#### 177 targets passed
- robot-integration.Should Fail On Loading Nonexisting Script
- robot-integration.Should Fail On Builtin With Invalid Parameters
- robot-integration.Should Fail On Peripheral Method With Invalid Parameters
- robot-integration.Should Fail On Python Command With Invalid Parameters
- robot-integration.Should Fail On Command Error
- emulation-environment.Should List Sensor Once
- emulation-environment.Should Set Temperature On Single Sensor
- emulation-environment.Should Set Temperature On Single Sensor Twice
Developing and testing TensorFlow Lite Micro edge AI algorithms on RISC-V and FPGAs

Build logs

INFO:Sc: Creating Sc... (2021-05-18 04:23:45)
INFO:Sc:-----------------------------------------------
INFO:Sc: FPGA device : xc7a35t-csg324-1.
INFO:Sc: System clocks 100.000Hz.
INFO:Sc:BusHandler:Creating Bus Handler...
INFO:Sc:BusHandler:Adding reserved Bus Regions...
INFO:Sc:BusHandler:Bus Handler created.
INFO:Sc:SRRHandler:Creating CSR Handler...
INFO:Sc:SRRHandler:8-bit CSR Bus, 32-bit aligned, 16.0kIB Address Space, 28448 Paging, big Ordering (Up to 32 Locations).
INFO:Sc:SRRHandler:Adding reserved CSRs...
INFO:Sc:SRRHandler:CSR Handler created.
INFO:Sc:IQHandler:Creating IRQ Handler...
INFO:Sc:IQHandler:IRQ Handler (up to 32 Locations).
INFO:Sc:IQHandler:Adding reserved IRQs...
INFO:Sc:IQHandler:IRQ Handler created.
INFO:Sc:-----------------------------------------------
INFO:Sc:Initial Sc:
INFO:Sc:-----------------------------------------------
INFO:Sc:32-bit wishbone Bus, 4.096kB Address Space.
INFO:Sc:8-bit CSR Bus, 32-bit aligned, 16.0kIB Address Space, 28448 Paging, big Ordering (Up to 32 Locations).
INFO:Sc:IRQ Handler (up to 32 Locations).
INFO:Sc:-----------------------------------------------
INFO:Sc:SRRHandler:ctl1 CSR allocated at Location 8
INFO:Sc:BusHandler:cpu_bus added as Bus Master.
INFO:Sc:BusHandler:cpu_bus added as Bus Slave.
INFO:Sc:SRRHandler:ctl1 CSR allocated at Location 1
INFO:Sc:BusHandler:rom Region added at Origin: 0x00000000, Size: 0x00010000, Mode: R, Cached: True, Linker: False.
INFO:Sc:BusHandler:rom added as Bus Slave.
INFO:Sc:RAM rom added: Origin: 0x00000000, Size: 0x00010000, Mode: R, Cached: True, Linker: False.
WHAT NEXT?
SOME ONGOING DEVELOPMENTS WITH TF LITE MICRO TEAM

- Work going on in TF Lite Micro mainline (and branches/PRs)
- Arduino Nano 33 BLE Sense with Nordic NRF52840 Cortex-M4F + ArduCam Mini OV2640 SPI camera - used in Harvard TinyML course
- Also adding support for more platforms, including QuickLogic RISC-V-based FPGA SoC
- Revamping entire TF Lite Micro CI to use Renode
- CI for Zephyr integration also in scope
- Additional work around Google Colab and demonstration of camera/audio enabled demos
WE CAN HELP YOU

- Build CI-driven and measurable TinyML pipelines
- Implement practical ML scenarios - open HW, SW as well as FPGA IPs (and tools)
- Adopt new, RISC-V and FPGA based approaches to ML. HW-SW co-design
GETTING STARTED WITH RENODE

- **Packages** (.deb, .rpm, .pkg.tar.xz, .dmg)
- **Renode portable**
- **Docker**
- **Conda**
- **Build** from source
THANK YOU FOR YOUR ATTENTION!