Accelerating ML with mainline

Linux Plumbers 2024 - Vienna

Tomeu Vizoso <tomeu@tomeuvizoso.net>

What is stopping vendors from mainlining their drivers?

- NXP tried to upstream a driver for the Ethos-U NPU, gave up on v1
- Amlogic tested the waters, balked off at the userspace requirement
- Samsung submitted a driver for their Trinity NPU, but gave up at v2
- Baylibre submitted a driver for Mediatek's APU, balked off at the userspace requirement
- Arm China started upstreaming a kernel driver for their Zhouyi NPU driver, gave up on v1
- ITRI tried to upstream a kernel driver for their NVDLA derivative
- Probably most other SoC and IP vendors also wish they had a kernel driver in mainline, but haven't found a way so far to do it.

Tomeu Vizoso <tomeu@tomeuvizoso.net>

How can we help vendors to mainline their drivers?

- Are there any blockers on the technical side?
- Or is it that they just don’t know how to do it?
- Or they haven’t been able to agree on releasing the userspace portion of the driver?

Tomeu Vizoso <tomeu@tomeuvizoso.net>

How close are we to having a common userspace API?

- GPU drivers have OpenGL and Vulkan
- Compute has OpenCL
- Application developers use ML accelerators via:
  - TensorFlow (Lite)
  - PyTorch/Executorch
  - ONNX
  - TVM
  - Vendor-specific runtimes, after having converted a model from one of the above frameworks
- OpenXLA, MLIR, IREE, …
- What can we do to further this goal?

Tomeu Vizoso <tomeu@tomeuvizoso.net>

Automated testing

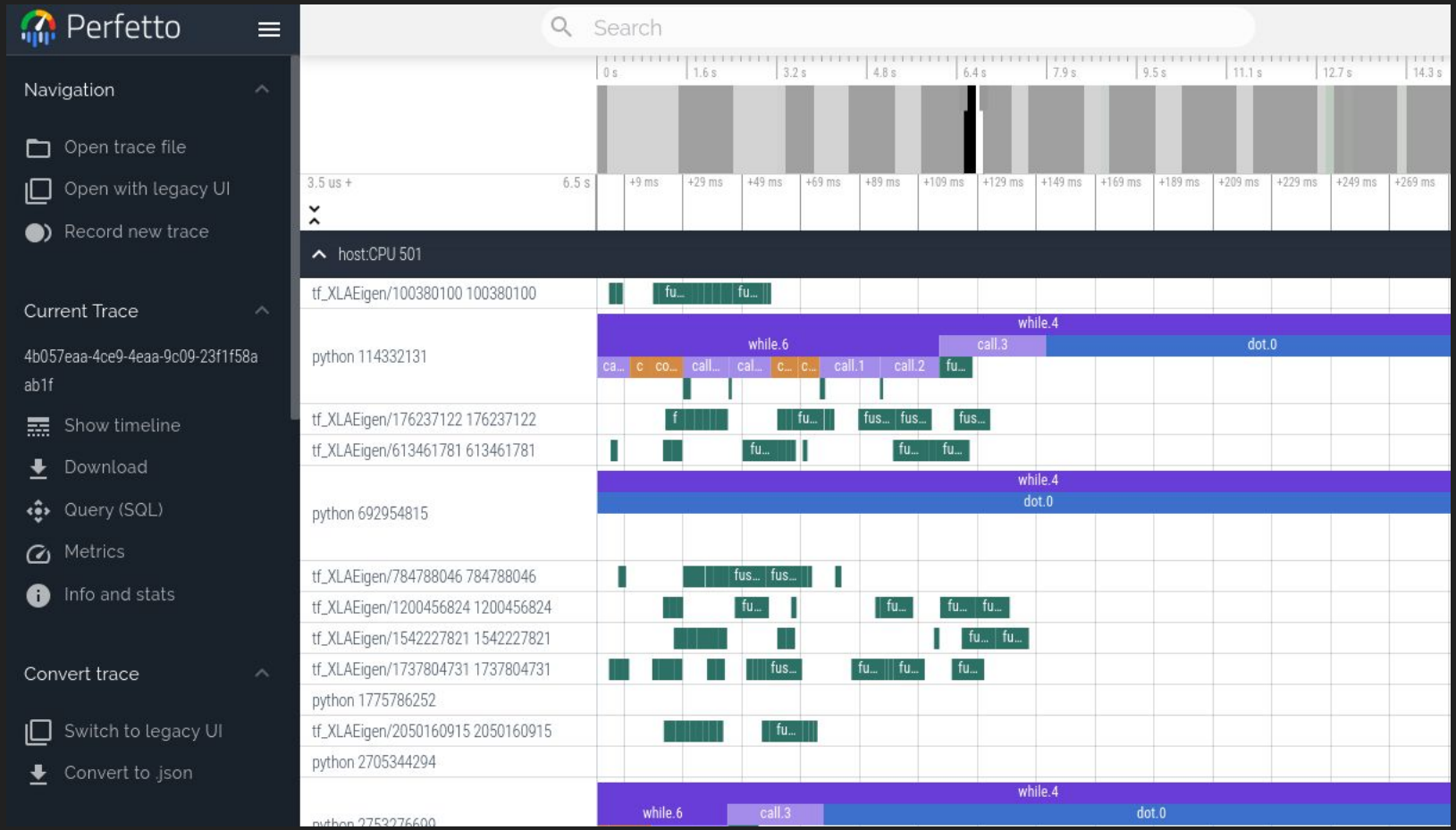
- DRM CI and KernelCI exist today
- Would be good to have a common test suite
- This is probably dependent on a common userspace API

Tomeu Vizoso <tomeu@tomeuvizoso.net>

Shared userspace infrastructure

- Compiler
  - MLIR to NIR pass
  - Shared compiler passes
    - Per-channel quantization to per-tensor
    - Signed to unsigned
    - Fuse activations
    - Unstride convolutions
    - Depthwise to normal convolutions
- Execution runtime
  - Resource management
  - Synchronization
- Virtualization
- Performance tools (Perfetto, …)

Tomeu Vizoso <tomeu@tomeuvizoso.net>



Tomeu Vizoso <tomeu@tomeuvizoso.net>

Firmware-mediated IP

- Quite a few out there:
  - Texas Instruments
  - Arm
  - Cadence
  - Huawei
  - Samsung
  - Intel
  - …
- Could these drivers share a single codebase for their firmware?

Tomeu Vizoso <tomeu@tomeuvizoso.net>

Interoperability

- Multimedia pipelines:
  - Camera/ISP
  - GPU
  - Compute
  - Machine learning
  - Video codec
- Cooperative execution by graph partitioning:
  - CPU
  - GPU
  - Accelerator

Tomeu Vizoso <tomeu@tomeuvizoso.net>

Standing issues in DRM infrastructure

- cgroups
  - Core mask
  - Internal memory allocation
- GPU scheduler
- GEM
- DMABuf
- DMA fences

Tomeu Vizoso <tomeu@tomeuvizoso.net>