XDC
2025

Embedded NPU drivers: 2025 update

Tomeu Vizoso <tomeu@tomeuvizoso.net>

ML accelerators in the edge

- All SoC vendors have ML accelerator IP in some of their chips
- Mainly oriented towards accelerating convolution neural networks:
  - Vision: image recognition, object detection, facial recognition, etc.
  - Speech recognition (RNNs)

XDC
2025

Evolution in drivers/accel

- Initial merge: November 2022
- Habanalabs: January 2023
- Intel VPU/NPU: January 2023
- Qualcomm QAIC: April 2023
- Etnaviv (Vivante): January 2024 ← In drivers/gpu/drm, though.
- AMD XDNA: November 2024

XDC 2025

Etnaviv: 2025 update

- Added support for a new generation: VIPNano SI+ as in the NXP i.MX 8MP
  - Very popular SoC in industrial, healthcare, transportation, building automation among others.
  - Long-term availability (10-15 years)
- Added a bunch of new operations:
  - ReLU, Reshape, Abs, Logistic, Subtract and Transpose
  - Added support for YOLOX-based models

XDC 2025

New driver: Rocket

- For the NPUs in recent Rockchip SoCs:
  - RK3588, RK3588S, RK3576, RK3566, RK3568
- In drivers/accel (drm-misc-next atm)
- No programmable cores
- 3 almost identical cores in the RK3588(S)
- drm-gpu-scheduler is used to load-balance among cores

XDC
2025

New driver: Ethos

- For the Ethos U65 and U85 NPUs:
  - The U65 can be found in the NXP i.MX93 SoC
  - The U85 hasn’t made it to public silicon yet, can be tested with Arm’s Fixed Virtual Platforms
- In drivers/accel (undergoing review in LKML atm)
- No programmable cores
- Had already an open source driver stack, but it isn’t suitable for mainline
- Documentation available

XDC 2025

Upcoming work in 2026

- PyTorch/Executorch support
- NIR for optimization passes and lowering models to the hardware
- New driver coming for Q1/Q2 2026
- One more new driver coming for H2 2026

XDC
2025

Questions

- You can ask in #ml-mainline at OFTC
- Or feel free to ask me any time after this talk

XDC
2025