Contribution ID: **266** Type: **not specified**

Demystifying Linux NPU Subsystem: From Vision to LLM at Edge

Saturday 13 December 2025 17:00 (45 minutes)

Neural Processing Units (NPUs) are becoming as common as GPUs in embedded SoCs, but Linux lacks a unified NPU subsystem. Current drivers are fragmented, vendor-specific, and often only tuned for vision inference (YOLO, ResNet). At the same time, newer workloads such as LLMs and multimodal models demand more flexible memory management, scheduling, and runtime integration.

This talk demystifies how NPUs work at the subsystem level —from DMA mapping and SRAM partitioning to command queue management. It will walk through case studies of deploying both vision models (YOLOv8) and LLMs (LLaMA3-tiny) on NPUs, highlighting where current Linux subsystems (DRM, V4L2, accel, AI/ML proposals) succeed and where they fall short.

Finally, it shows how Edgeble has adapted these learnings in real deployments on SoC and PCIe based NPU engine drivers by adding quantization and model scheduling.

The session aims to start a discussion around a more unified Linux NPU subsystem, drawing parallels to the GPU evolution, and inviting collaboration with kernel developers, hardware vendors, and OSS communities.

Primary author: Mr TEKI, Jagan (Upstream Linux Specialist - Linux DRM Panle/Bridge, U-Boot Allwinner, SPI, SPI Flash Maintainer, NPU Enthusiastic)

Presenter: Mr TEKI, Jagan (Upstream Linux Specialist - Linux DRM Panle/Bridge, U-Boot Allwinner, SPI, SPI Flash Maintainer, NPU Enthusiastic)

Session Classification: LPC Refereed Track

Track Classification: LPC Refereed Track