



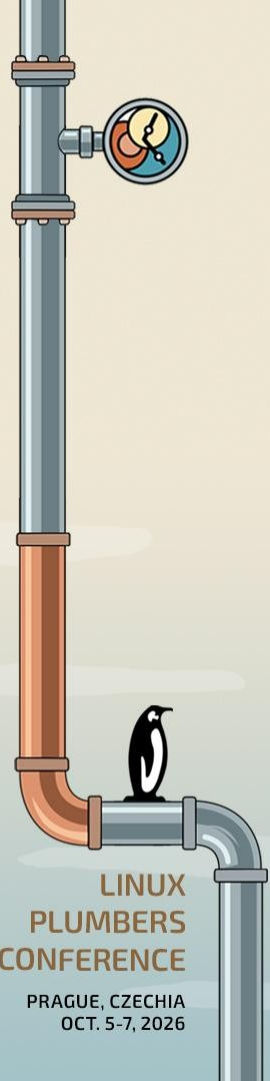
Firmware-mediated accelerators for edge AI

Tomeu Vizoso <tomeu@tomeuvizoso.net>



LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

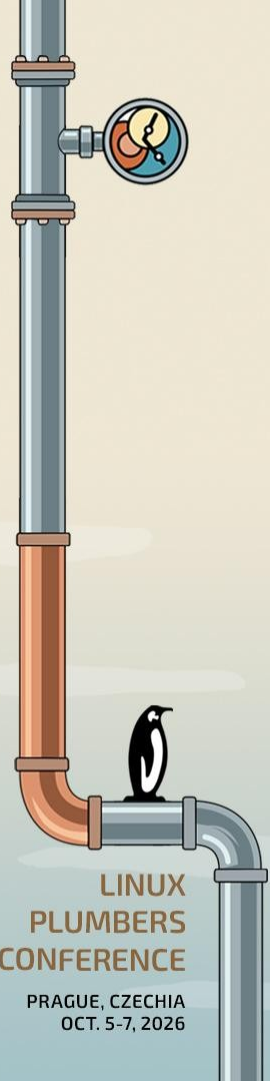


Accelerator drivers already in mainline

- Qualcomm QAIC
- AMD XDNA
- Intel VPU/NPU
- Intel HabanaLabs
- Arm Ethos-U
- Vivante VIP (Etnaviv)
- Rockchip NPU (Rocket)

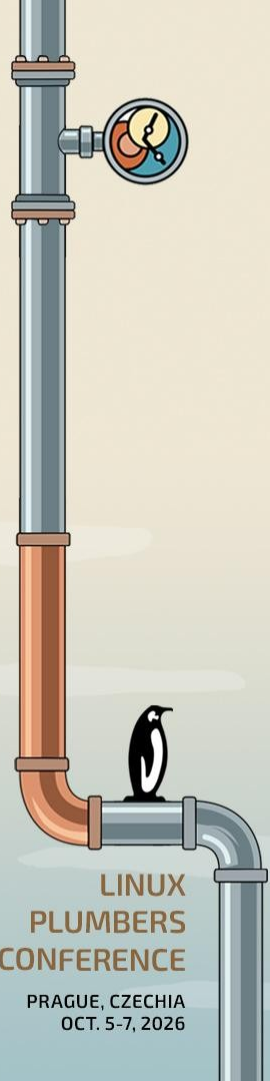
LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026



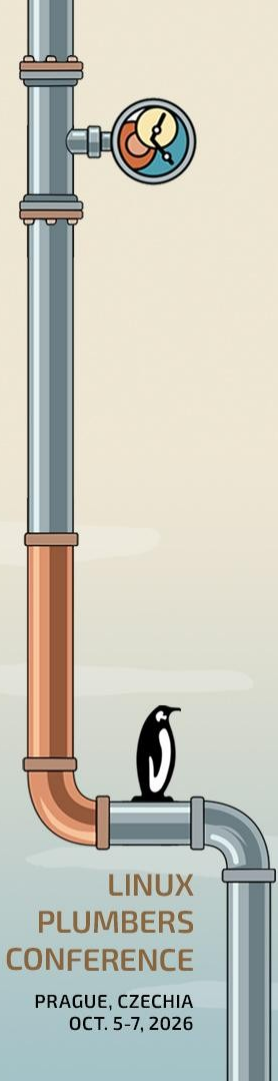
Accelerator drivers under review

- Qualcomm QDA
 - Hexagon DSP
 - HMX and HVX ISA extensions
- NXP Neutron
 - RISC-V core
 - Fixed-function accelerator bolted to the side
- Texas Instruments C7x (Thames)
 - C7x DSP cores
 - SE/SA/MMA ISA extensions



Firmware-mediated accelerators

- Kernel driver interacts with the hardware via IPC with the firmware
- Userspace compiles payloads that run mostly unmodified on the hardware
- Without a front-end processor in silicon, the other components need to pick up its tasks:
 - Firmware uses HW facilities to isolate clients
 - Kernel driver validates what the hardware cannot isolate
 - Userspace takes care of tiling, ordering of operations and planning data movement
 - Userspace coordinates the different accelerator components



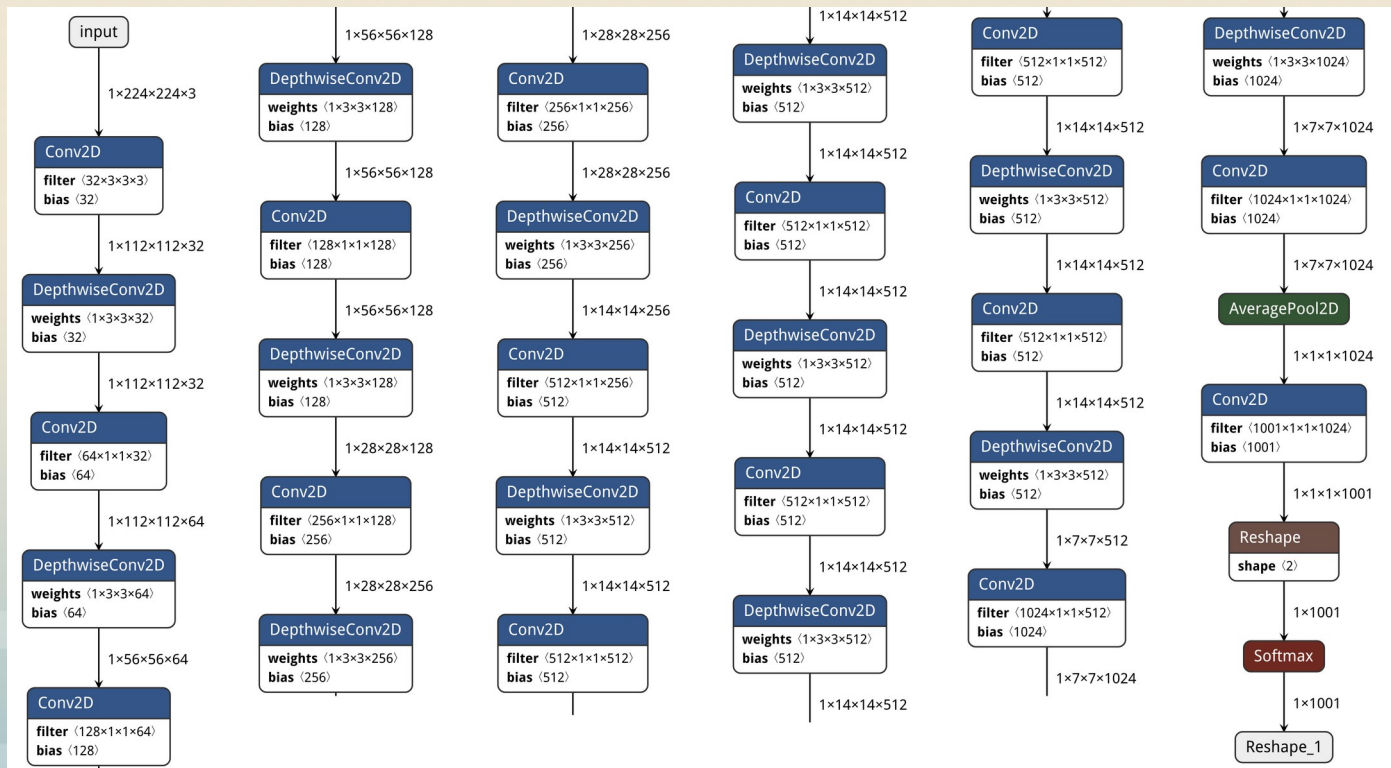
Thames' goals

- Simple UAPI, no hidden or grandfathered requirements
- Security, performance and simplicity, in that order
- Open source firmware
- Everything needed is distributable by Yocto, Debian, Fedora, etc
- Ready for Vulkan and OpenCL
- A model for other drivers

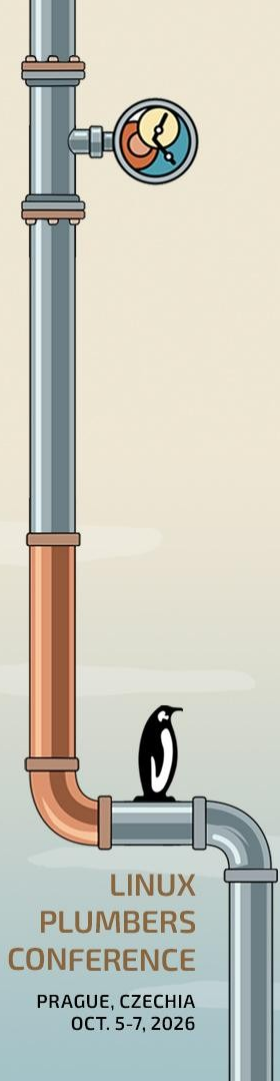
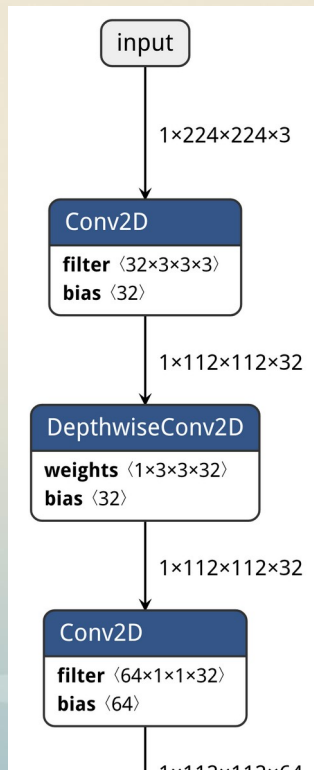
LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026

MobileNetV1: image classification



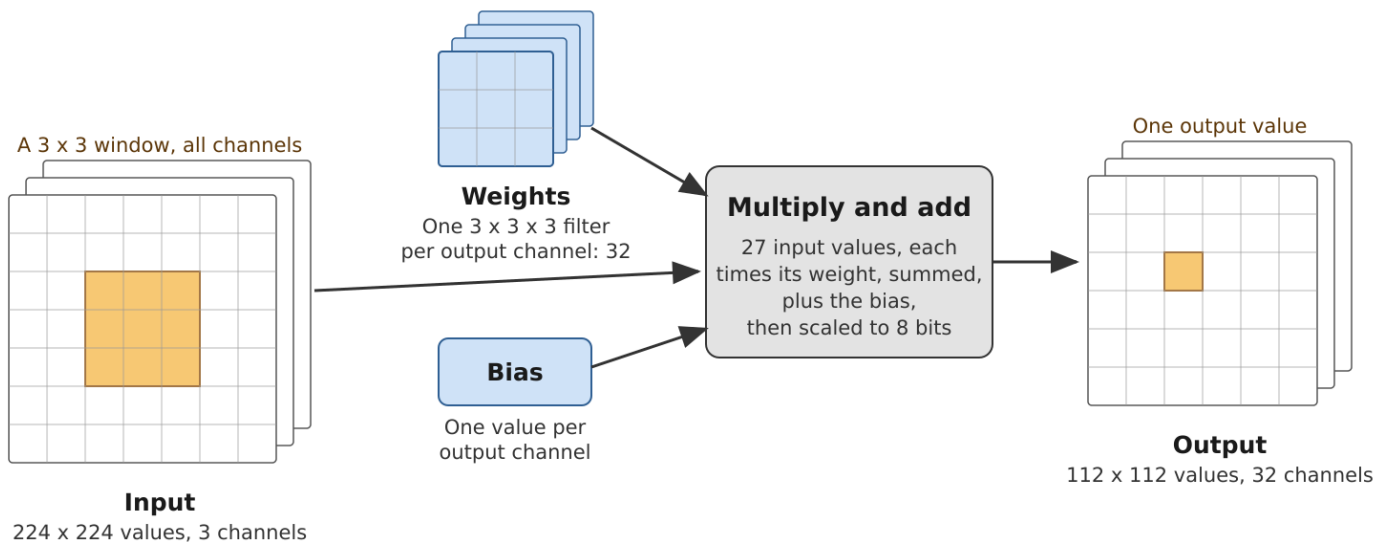
MobileNetV1: zoomed in



LINUX
PLUMBERS
CONFERENCE

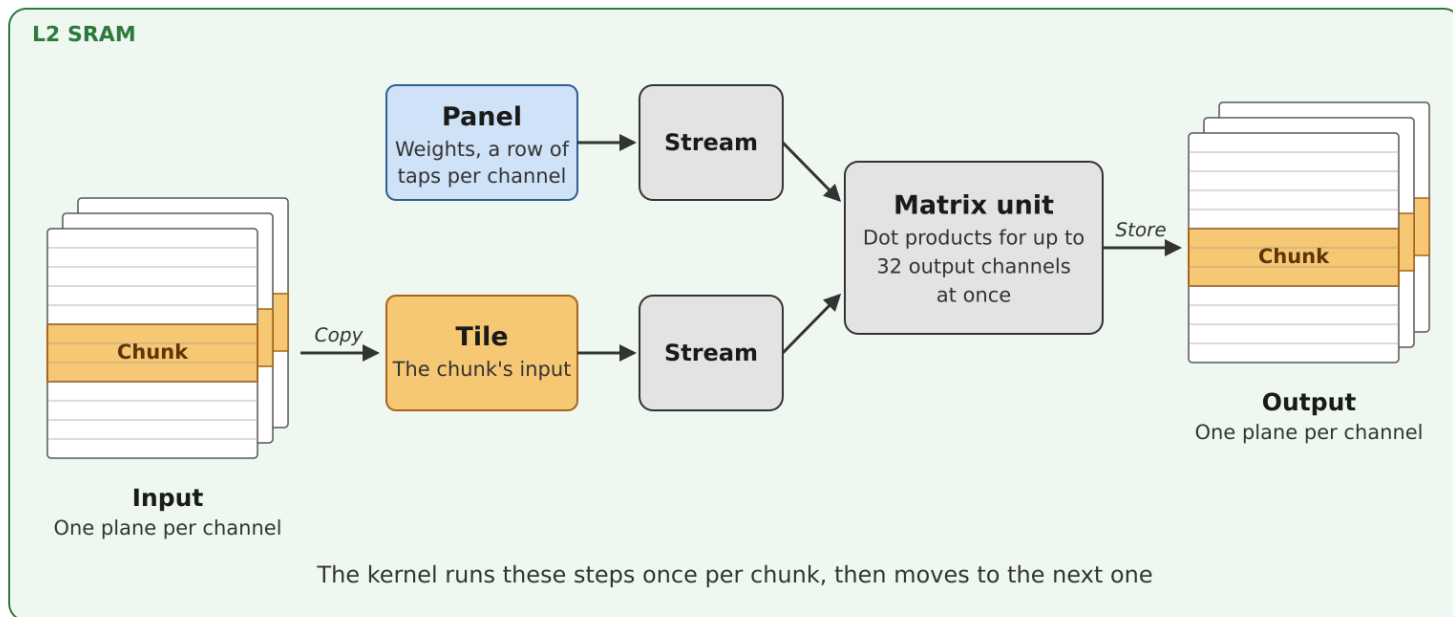
PRAGUE, CZECHIA
OCT. 5-7, 2026

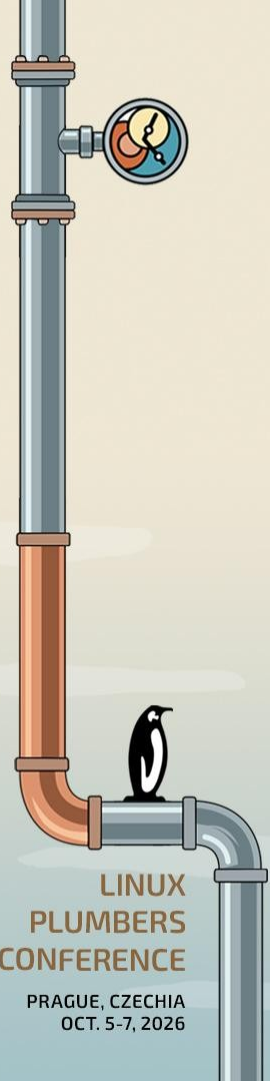
Neural network convolution



Repeated for every output position and every output channel:
112 x 112 x 32 values of 27 multiplications each, 10.8 million for this one layer

Computation pipeline

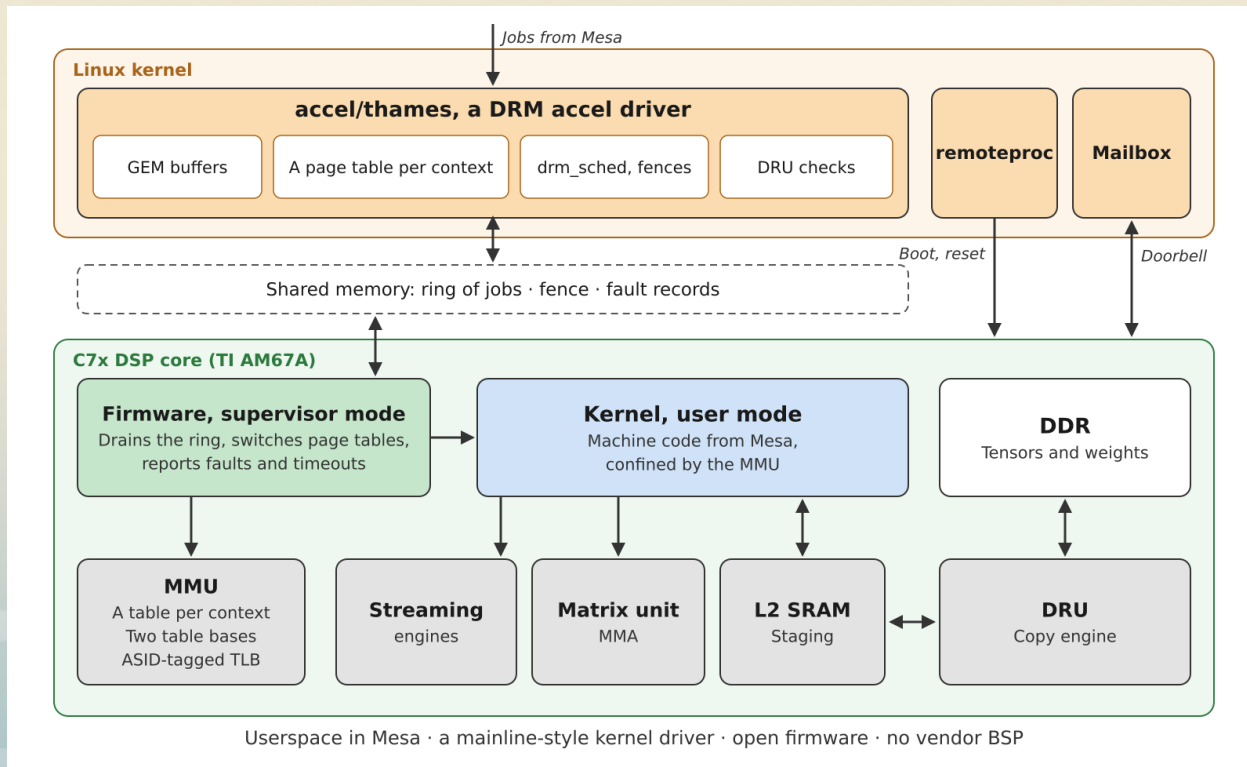


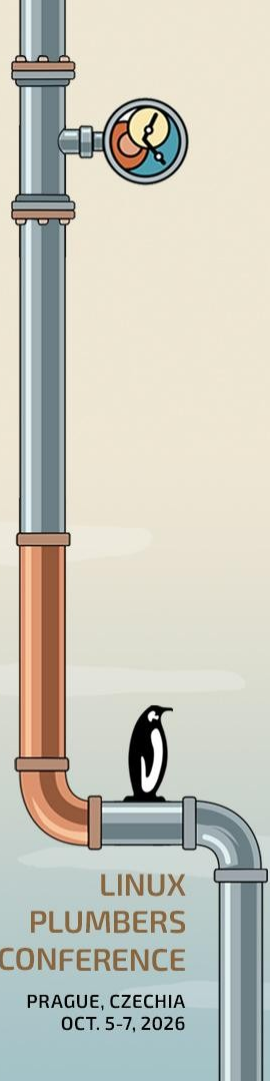


Texas Instruments' J722s and C7524

- 2 C7524 cores at 1 GHz, with each:
 - 32-byte vectors
 - MMA2 (Matrix Multiply Accelerator 2) unit with 32x32 byte tiles
 - 2 Streaming Engines
 - 4 Streaming Address Generators
 - MMU with VMSAv8-64 translation table format, 8-bit ASIDs
 - 2 MiB of main SRAM
 - 256 KiB of AUX SRAM
 - 1 DMA engine: DRU (Data Routing Unit)

The complete pipeline



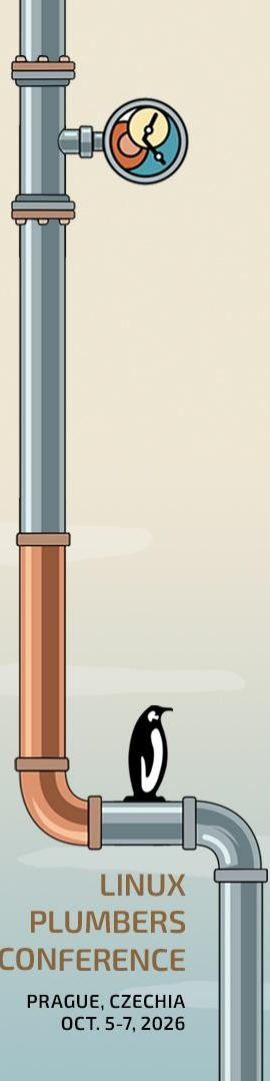


Component responsibilities

- Userspace
 - Compile DSP kernel
 - Fill data structures
 - Plan DMA transfers
- Kernel driver
 - Validate DMA transfers
 - Load balance jobs among cores
 - Manage the DSP address spaces
 - Power management
 - Ban contexts after repeated faults, and reset the cores
- Firmware
 - Accept jobs
 - Context switching
 - Configuration of DMA transfers
 - Kernel execution in user mode
 - Fault handling, job timeouts

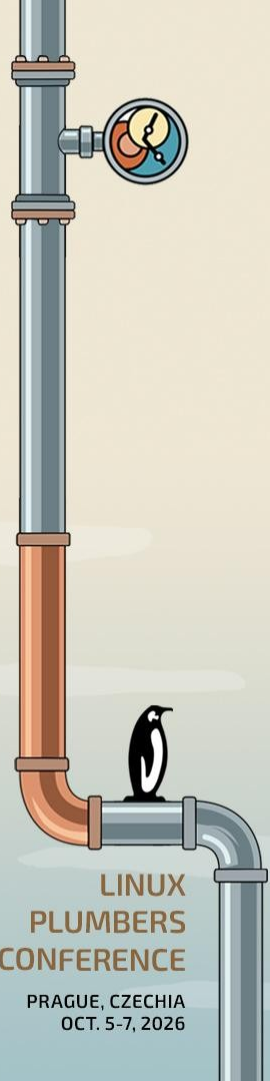
LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026



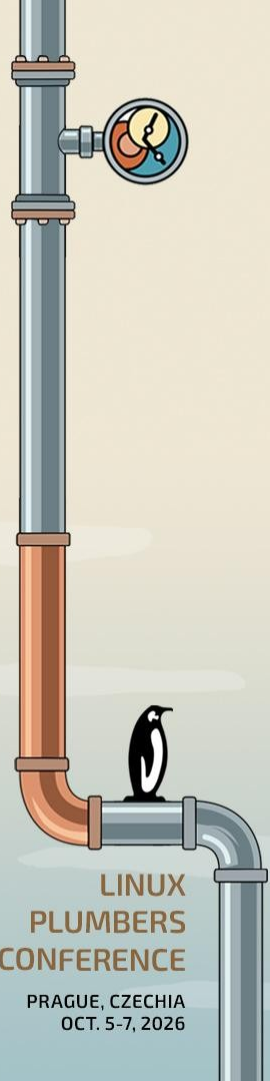
Isolation among clients

- A client cannot access other clients' memory, except those explicitly shared
- Also applies to SRAM, registers and memory internal to the compute units
- A client cannot disrupt other clients' execution
 - Max job timeout
 - Fault recovery



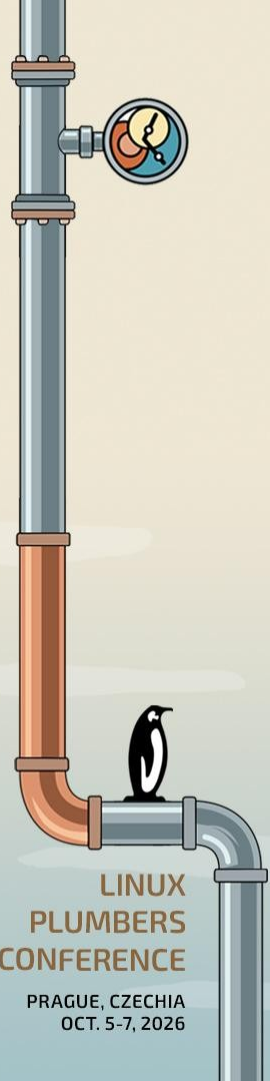
Firmware lifecycle

- Remoteproc:
 - Powers the device on, as described in the DT
 - Loads the right firmware from the filesystem
- Kernel driver:
 - Gets the shared memory carveout from the firmware via remoteproc
 - Handshakes with the firmware
 - IPC version
 - Capabilities
 - Restarts the core after a FW fault or a hang



Graph compilation

- Gallium driver receives a graph with a struct per operation
- Each operation gets compiled into a subkernel, merged into a single one
- Defines DMA transfers:
 - Coefficients
 - Input tiles too, where it pays
- DSP kernel uses the Streaming Engines to stream the inputs in
- Computation uses the MMA if suitable
- Output will be placed in memory using the Streaming Address Generators
- One DSP kernel, its parameters and its DMA plan per subgraph

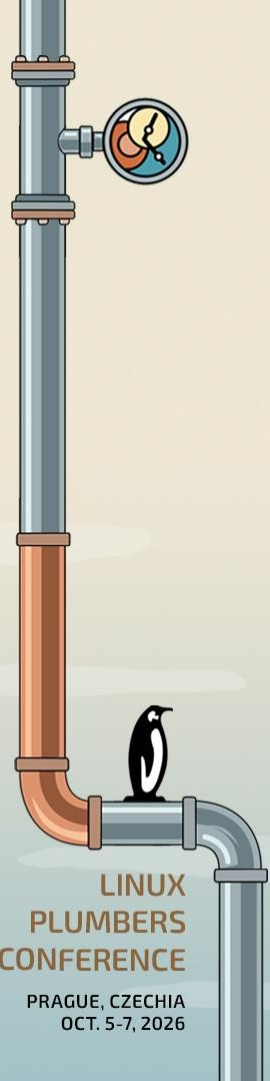


UAPI at a glance

- 9 IOCTLs
- Buffer create, mmap and destroy
- Job create, submit and destroy
- Device queries: capabilities and fault record
- Buffer synchronization: prep and fini

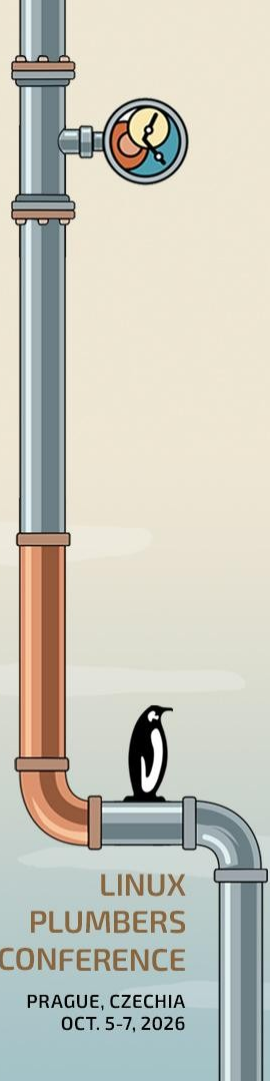
LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026



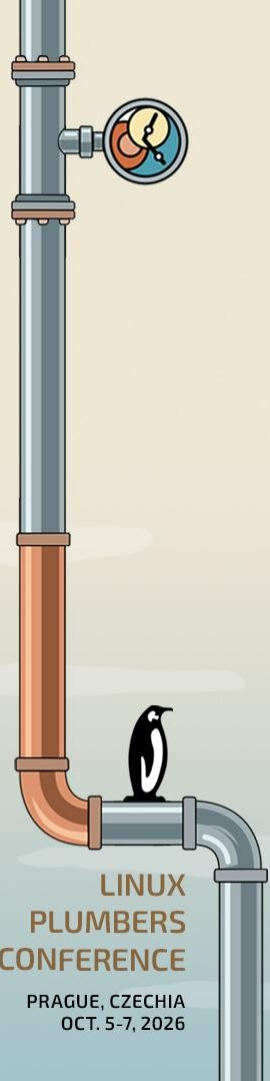
Buffers in a job

- DSP kernel (read-only, executable)
- Parameters (read-only)
- Constants (read-only)
- Intermediate tensors (read-write)
- Status (read-write)
- Graph inputs (read-only)
- Graph outputs (read-write)



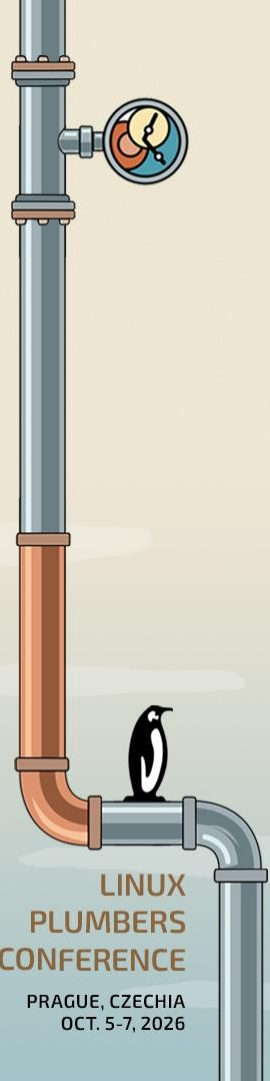
DRM_IOCTL_THAMES_JOB_CREATE

- Job descriptor
 - DSP kernel
 - Params BO
- DMA transfers
- `int kernel(void *params)`
- Created once and submitted many times



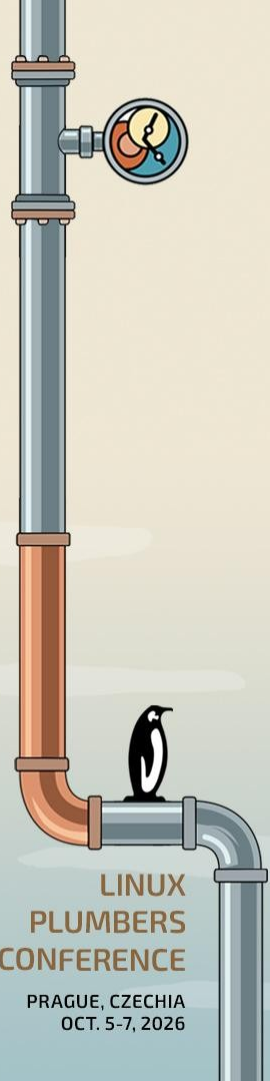
DRM_IOCTL_THAMES_JOB_SUBMIT

- Contents of params BO or other BOs might have changed
- Handles of all BOs above, for implicit synchronization
- In and out fences
- Implicit wait when reading one of the writable BOs
- Once the status BO becomes unblocked for reading, any faults can be processed



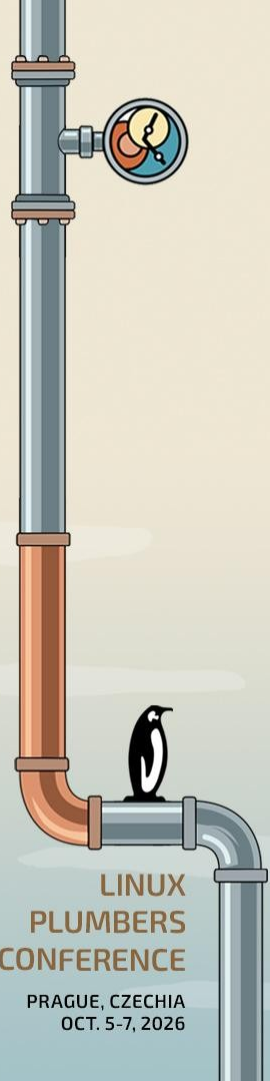
Memory management

- Buffers are mapped to the DSP on creation
- DSP virtual address is given to userspace
 - To be referenced in the DSP kernel
 - And in DMA transfers
- Contiguous flag for DRU transfers
- Alignment also important for performance
- Access permissions on the DSP side



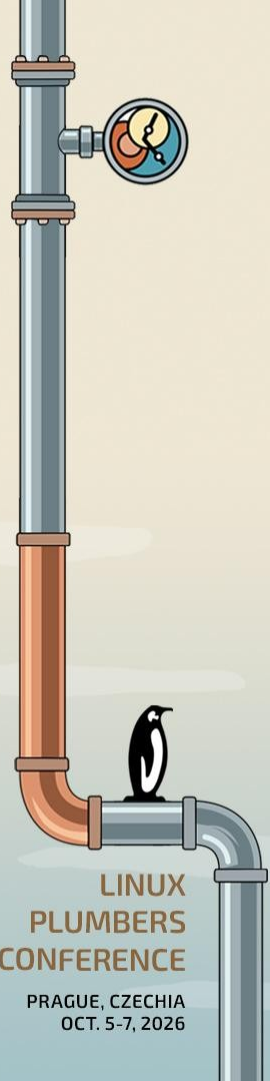
DSP address spaces

- The supervisor in the firmware manages its own address space
- Kernel driver manages the address spaces of clients
- Each client has its own page table
- ASIDs make context switching cheaper



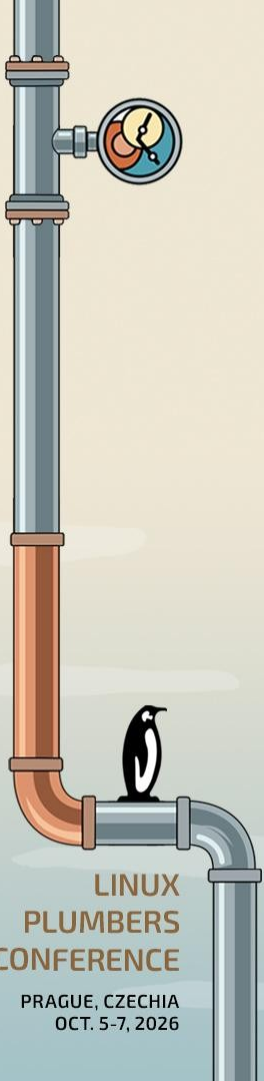
Context switching

- Switch page tables (172 cycles)
- No TLB invalidation needed thanks to ASIDs
- Clear SRAM (223k cycles)
- Close streams and purge MMA unit (2k cycles max)
- Registers wiped at every entry (18 cycles)



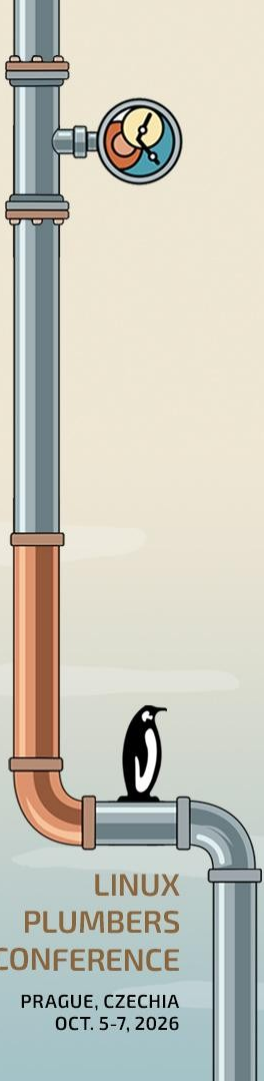
SRAM

- Mapped at a fixed address
- AUX and main
- Much faster than DDR
- Used for staging and intermediate buffers



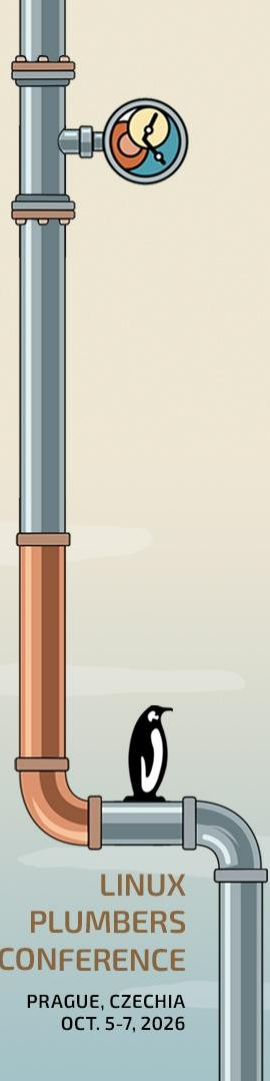
DMA engine

- Asynchronous data moves
- Doesn't go via the MMU
- On some SoCs it can decompress weights on the fly
- Configuration apart from triggering
- DDR -> SRAM: 12 bytes per cycle
- SRAM -> SRAM: 31 bytes per cycle
- Errors don't raise any faults: must be validated



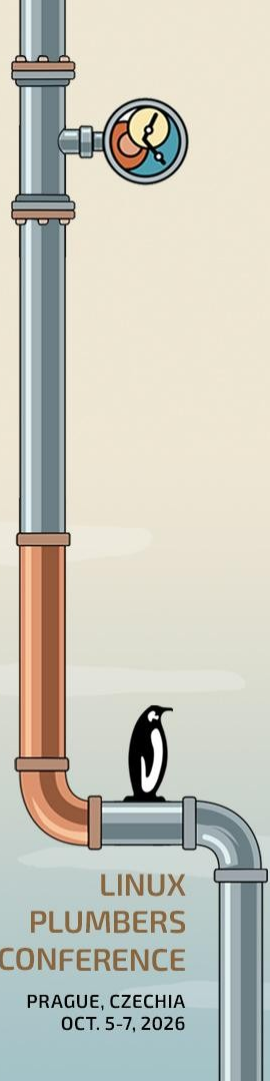
Job validation

- DMA commands
 - DRU bypasses the MMU, so no isolation
 - Userspace programs the actual DRU records
 - Records are validated for correctness
 - Kernel driver translates the addresses from virtual to physical
 - Raise an error on invalid addresses and lengths



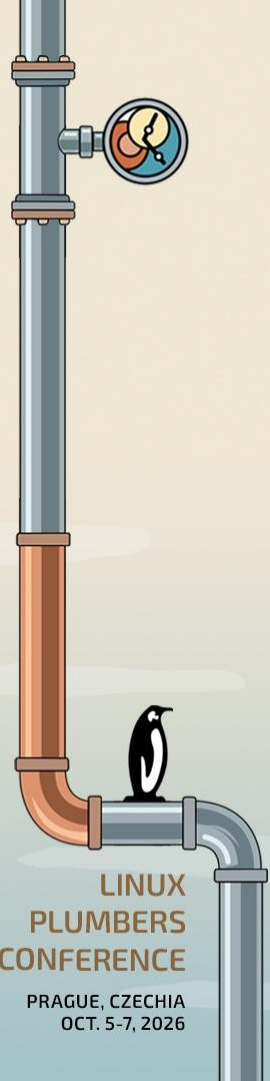
Job dispatch

- `drm_sched` orders jobs between clients and chooses a core to execute them
- One command ring per core in memory shared with the firmware
- A mailbox doorbell interrupt wakes up the core when it's idle on an empty ring (3.5 us)
- Completion is a fence
 - A sequence number the device writes and the host reads
- The firmware looks one entry ahead
 - So the next job's first transfer can start while the current job still runs



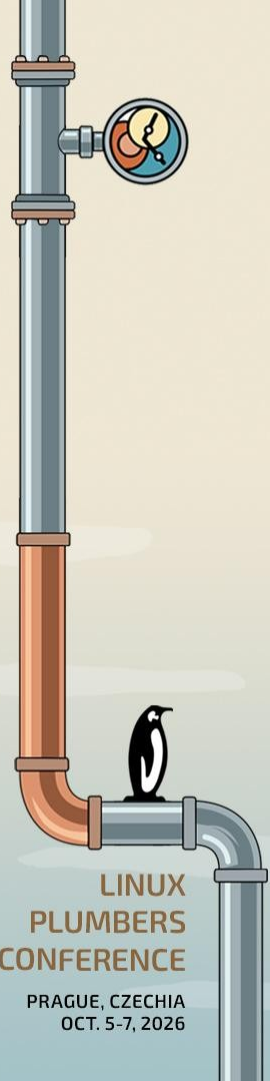
Execution of a kernel

- User execution mode:
 - No access to other page tables
 - No access to privileged registers
 - Clean faults on privilege violations
- DRU transfers:
 - next chunk syscall
 - transfer wait syscall
- Kernel ABI: The only input argument is passed in A4, as is the return value
- Exit syscall



Fault recovery

- Faults caught by the supervisor, state cleared and fault record reported to the kernel driver
- Same for timed out DSP kernels
- Kernel driver will ban a context after faulting three times in the last 10 jobs
- DSP core resets are expensive (94 ms)
 - Only on faults and hangs from the firmware itself

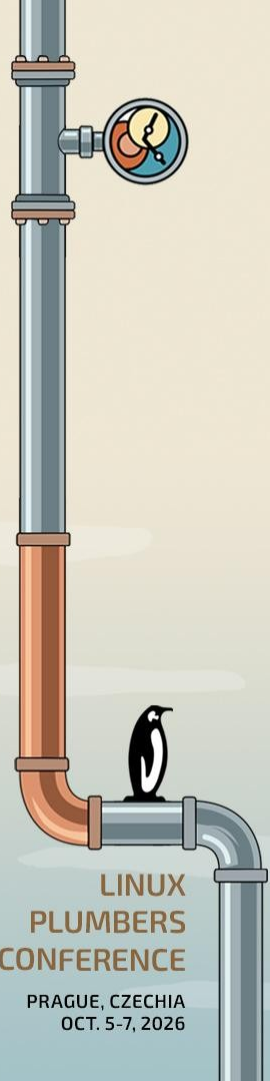


Ready for OpenCL and Vulkan Compute

- UAPI and IPC know nothing about ML
- Kernels are untrusted
- Per-job completion fence
- Extensible UAPI, allows for adding:
 - Input sync objects
 - Timeline points

LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026

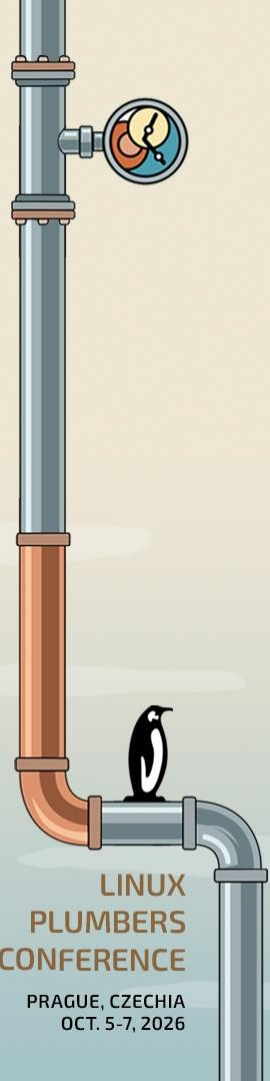


Current status

- Firmware:
 - Open-source
 - Redistributable
 - Needs closed source compiler
- Fulfills DRM mainlining requirements
- SSDLite MobileDet with TensorFlow Lite at 14.4 ms per frame

LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026



What is next

- Power management
 - Suspend/resume
 - Runtime PM
- Other hardware to be supported:
 - TDA4VM (J721E)
 - AM68 (J721S2)
 - AM69 (J784S4)
 - AM62A7
- Performance work in userspace (target is ~5 ms for MobileDet)
- Implement more operations so more models are accelerated
- Test with PyTorch/Executorch
- OpenCL, Vulkan

LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA
OCT. 5-7, 2026