Contribution ID: **41** Type: **not specified**

Accelerating ML with mainline

Friday, 20 September 2024 10:45 (45 minutes)

As of today, the vast majority of accelerators for machine learning (NPUs, TPUs, DLAs, etc) lack a presence in the mainline kernel.

These accelerators can be used only with out-of-tree kernel drivers and binary-only userspace stacks, often forks of one or more open-source machine-learning frameworks. Companies are prey to vendor lock-in.

Companies selling accelerators are starting to react to the pressure from their customers and are exploring ways to mainline the drivers for their hardware.

Four drivers have been mainlined as of 6.10, but at least four other vendors have tried to mainline their drivers and seemingly abandoned the effort.

At this BoF we will discuss the challenges that existing drivers face, and how to make it easier for other vendors to mainline their drivers.

Agenda:

- What is stopping vendors from mainlining their drivers and how could we make it easier for them?
- Userspace API: how close are we from a common API that we can ask userspace drivers to implement? What can be done to further this goal?
- Automated testing: DRM CI can be used, but would be good to have a common test suite to run there. This is probably dependent on a common userspace API.
- Other shared userspace infrastructure (compiler, execution, synchronization, virtualization, ...)
- Firmware-mediated IP: Can these drivers share a single codebase for their firmware?
- Any standing issues in DRM infrastructure (GEM, gpu scheduler, DMABuf, etc) that are hurting accel drivers?
- GPU and accelerator interoperability: pipelines with graphics, compute and machine learning components, and also offloading portions of a model to a GPU and others to an accelerator.

Primary author: VIZOSO, Tomeu (Independent contractor)

Presenter: VIZOSO, Tomeu (Independent contractor)

Session Classification: Birds of a Feather (BoF)

Track Classification: Birds of a Feather (BoF)