Contribution ID: **73** Type: **not specified**

Accelerating AI training fleets with sched_ext

We present one of the first deployments of sched_ext to a large fleet of AI training hardware composed of multi CPU socket systems with attached Nvidia GPUs. GPU training workflows run frequent synchronization across all the training processes which makes them extremely sensitive to task scheduling micro-delays that prevent work from being dispatched to the GPUs. In addition, the training systems boxes run several components of the stack like data loading, preprocessing and model checkpointing on the CPUs which increases scheduling congestion. We used sched_ext, a user-space scheduler (scx_layered) and we deployed it to the entire Reality Labs GPU fleet with tens of thousands of GPUs. We were able to improve the GPUs'compute unit utilization on certain model types by 9% and reduce the fleet training cost. The presentation describes our journey in identifying the latency critical system tasks, developing the scheduler, ensuring resource isolation, debugging corner cases and monitoring the performance across the entire fleet.

Primary authors: LU, Patrick; ANDREI, Valentin (Meta)

Presenters: LU, Patrick; ANDREI, Valentin (Meta)

Session Classification: sched_ext: The BPF extensible scheduler class MC

Track Classification: sched_ext: The BPF extensible scheduler class MC