Contribution ID: **182**

Type: **not specified**

Lessons from scaling BPF to detect RDMA Device Drivers Bugs in real time

Training large models requires significant resources and failure of any GPU or Host can significantly prolong training times. At Meta, we observed that 17% of our jobs fail due to RDMA-related syscall errors which arise due to bugs in the RDMA driver code. Unlike other parts of the Kernel RDMA-related syscalls are opaque and the errors create a mismatched application/kernel view of hardware resources. As a result of this opacity and mismatch existing observability tools provided limited visibility and DevOps found it challenging to triage – we required a new scalable framework to analyze kernel state and identify the cause of this mismatch.

Direct approaches like tracing the kernel calls and capturing meta involved in the systems turned out to be prohibitively expensive. In this talk, we will describe the set of optimizations used to scale tracking kernel state and the map-based systems designed to efficiently export relevant state without impacting production workloads.

Primary authors: SAMOYLOV, Maxim; GUPTA, Prankur (Meta); BENSON, Theophilus (Carnegie Mellon University)

Presenter: GUPTA, Prankur (Meta)

Session Classification: eBPF Track

Track Classification: eBPF Track