Reveal: Hardware-Centric Detection of Silent Inefficiencies in Machine Learning Systems

Ziji Chen
University of Oxford

Steven Chien
University of St Andrews

Peng Qian
University of Oxford

Noa Zilberman
University of Oxford

Modern machine learning (ML) workloads execute on complex stacks involving CPUs, GPUs, memory hierarchies, operating systems, and high-speed interconnects. In production clusters and cloud environments, developers often lack direct visibility into the underlying infrastructure, while operators typically do not have access to application internals. This separation creates an observability gap that makes performance regressions difficult to diagnose.

Importantly, many production performance degradations are not caused by explicit faults or resource exhaustion of infrastructure. Instead, they arise from subtle interactions between workloads and hardware configurations. Examples include NUMA misplacement, communication misconfiguration (e.g., NCCL queue-pair settings), interrupt imbalance, or memory configuration artifacts. These issues degrade end-to-end performance while leaving coarse system metrics such as utilization within normal ranges. As a result, conventional monitoring and anomaly detection tools often fail to identify them.

In this work, we present *Reveal*, a hardware-centric framework that detects and attributes such silent inefficiencies using only low-level telemetry available to system operators. *Reveal* collects host-level metrics from standard Linux interfaces and hardware performance counters, derives workload-specific behavioral features, and applies an unsupervised anomaly detection pipeline to identify abnormal subsystem interactions across CPU, memory, network, storage, and GPUs. The framework requires no application instrumentation and can operate continuously in production environments.

We evaluate *Reveal* across more than 30 representative ML workloads running on both GPU and CPU clusters. Our results show that *Reveal* surfaces several performance-limiting conditions that remain invisible to traditional monitoring tools. Acting on these signals enables practical remediation, including a 5.97% reduction in runtime for a production DeepSeek fine-tuning workload. The full version of this work is publicly available as an arXiv preprint: https://arxiv.org/abs/2510.26008.