

Resilient HPC Operations through Observability-Aware Predictive Maintenance

Authors:

Michael Bidollahkhani, Julian Kunkel

Predictive maintenance in high-performance computing (HPC) and compute continuum systems commonly relies on telemetry streams, monitoring metrics, and system logs to detect anomalies and anticipate failures. In real operational environments, these data sources can be incomplete, delayed, weakly structured, noisy, or misleading. Current state-of-the-art models can analyze such data and are effective in detecting whether an issue exists or is likely to occur. However, the existing literature shows that methods for identifying cause-effect relationships behind detected issues, as well as assessing the reliability of model outputs during predicted or detected failures, remain relatively underdeveloped.

This limitation reduces the reliability of conventional anomaly detection pipelines and creates uncertainty in maintenance decisions, especially in heterogeneous infrastructures where workloads move across HPC, cloud, and edge resources. To address this gap, this contribution introduces observability as a decision layer for predictive maintenance in HPC and compute continuum systems. The proposed perspective extends direct monitoring by asking what can be inferred about the internal state, degradation behavior, and operational risk of a system when telemetry or logs are missing, incomplete, weak, or inconsistent.

An observability-aware layer can support anomaly detection, failure anticipation, root-cause interpretation, reliability assessment, and risk-aware scheduling in dynamic compute environments. The poster presents observability as a new analytical extension of AI-based predictive maintenance. It shows how observability can bridge raw monitoring data and actionable maintenance decisions, especially when available signals do not fully explain the system state. By positioning observability as a decision-support layer, this contribution highlights a path toward more robust, interpretable, and resilient predictive maintenance frameworks for HPC and compute continuum systems.