**Enhancing I/O Observability and Storage Reliability in High-performance
Computing Systems**

Runzhou Han

High-performance computing (HPC) plays an essential role in scientific research, engineering, and business. Due to the complexity of HPC, analysis of HPC workload and HPC system is a both critical and challenging to the community.

We first introduce a data-centric provenance solution to enhance observability of HPC workloads. Data provenance captures the life cycle of data, including its origins and usage patterns in scientific workflows on HPC systems. Existing provenance solutions are inadequate due to incompatible models and system implementations. We collaborate with domain scientists to survey HPC workflows and identify specific provenance requirements. Based on the survey, we propose PROV-IO+, a provenance framework that incorporates an I/O-centric model to precisely describe scientific data, associated I/O operations, and environments. Our experiments with real-world workflows demonstrate that PROV-IO+ effectively meets the provenance needs of domain scientists with reasonable performance.

Then, we introduce our study of the failure recovery and logging mechanisms of parallel file systems (PFSes) which are a major subset of high-performance distributed storage systems (DSSes). We propose a tool called PFault to simulate failures and analyze the behavior of PFSes. We focus on two widely used PFSes (Lustre and BeeGFS) and find areas where they can improve in handling failures. We also investigate the causes of abnormalities, and our analysis has led to a patch for Lustre. Additionally, we characterize the extensive logs generated in the experiments in details and identify the unique patterns and limitations of PFSes in terms of failure logging. Moreover, we further study the capabilities of PFS checkers with more fine-gained fault injection and identify instances of sub-optimal behavior leading to issues like kernel panics and incorrect repairs.

Finally, we present our effort in studying erasure coding (EC) performance in high-performance storage systems. Erasure coding plays a crucial role in the fault tolerance of modern high-performance DSSes. Inspired by recent research on storage configuration, we study the configuration sensitivity of EC in real DSS in this work. Our results show that configurations may affect the EC recovery time significantly, and EC may introduce extra write amplification (WA) than the theoretical expectation. Our work suggests the importance of considering the context of real DSS for EC research and implementation.