Deep Learning (DL) applications are becoming an increasingly important workload on HPC systems such as Summit and Frontier.

To efficiently Run and Scale DL Applications to leverage state-of-the-art HPC system remains challenging.

Existing prefetching and caching solutions pose non-trivial challenges such as:
- Fail to fully utilize the compute node-local NVMe.
- Require modifications to applications or input I/O pipeline.
- Incur additional metadata overhead and bottlenecks.

**Motivation and Research Challenges**

- I/O optimization for DL Applications is non-trivial challenge.
- Dataset characteristics, DL Access patterns, and I/O properties.
- MDTest on Summit for DL workload access patterns.
- (Open-Read-Close) with two different file sizes, 32KB and 8MB.

Opportunity to exploit node-local or near-node local storage on compute nodes and solve I/O Scalability limitations by layering a caching system.

- With diverse deployment model, portable and POSIX support, no metadata slowdown, no repeated re-reads from PFS.

**An Overview of High-Velocity AI Caching System**

"To Scale on thousands of compute nodes on leadership-class supercomputer such as Summit and Frontier without modifying DL applications and additional metadata bottlenecks and storage overhead".

- A simple, lightweight and transparent library intended to accelerate I/O access for DL applications that utilize read-only data with a high re-read rate.
- Architecture agnostic and can be deployed on node-local, near-node local or rack-local storage.
- Supports POSIX operations (open, read and close) via LD_PRELOAD and fully portable, requiring no changes to application or PFS.
- Guarantees no repeated re-reads from underlying PFS, files are cached in node-local storage.
- Consists of two components:
  - **Client**: intercepts the application I/O calls.
  - **Server**: processes the I/O operations and cache dataset locally on node-local NVMe.

**ACKNOWLEDGMENTS**

This research used resources of the Oak Ridge Leadership Computing Facility, located in the National Center for Computational Sciences at the Oak Ridge National Laboratory, which is supported by the Office of Science of the DOE under Contract DE-AC05-00OR22725.