PLFS and HDFS: Enabling Parallel Filesystem Semantics In The Cloud

Milo Polte[1], Esteban Molina-Estolan[2], John Bent[3], Garth Gibson[1], Carlos Maltzahn[2], Maya B. Gokhale[4], Scott Brandt[2]

1: Carnegie Mellon University
2: University of California, Santa Cruz
3: Los Alamos National Laboratory
4: Lawrence Livermore National Laboratory

Hadoop Distributed File System

- Increasingly wide deployment due to prevalence of Hadoop
  - Facebook, Yahoo, Hulu ....
- Good support for resilience
- But lacks some common file system features
  - No support for concurrent writers
  - No support to re-open files for rewrite

Design

- Augment PLFS to speak to HDFS
- PLFS writes to HDFS files as log files
- Two minor variations from normal PLFS behavior:
  - New log file on each open
  - New index file on every session
- HDFS sees a set of individuals writers accessing exclusive files
- Going through PLFS allows HPC applications to use HDFS as a store

Challenges

- Teaching PLFS to speak HDFS
  - PLFS is designed for POSIX. HDFS is a new API, semantics
- Output from an HPC application should be available to MapReduce
- PLFS could expose a file map to Hadoop applications
  - But with strided writes, sequential data will be small
  - May create too many map jobs or map jobs that mostly read remotely
  - Possible solution: A MapReduce PLFS flattener

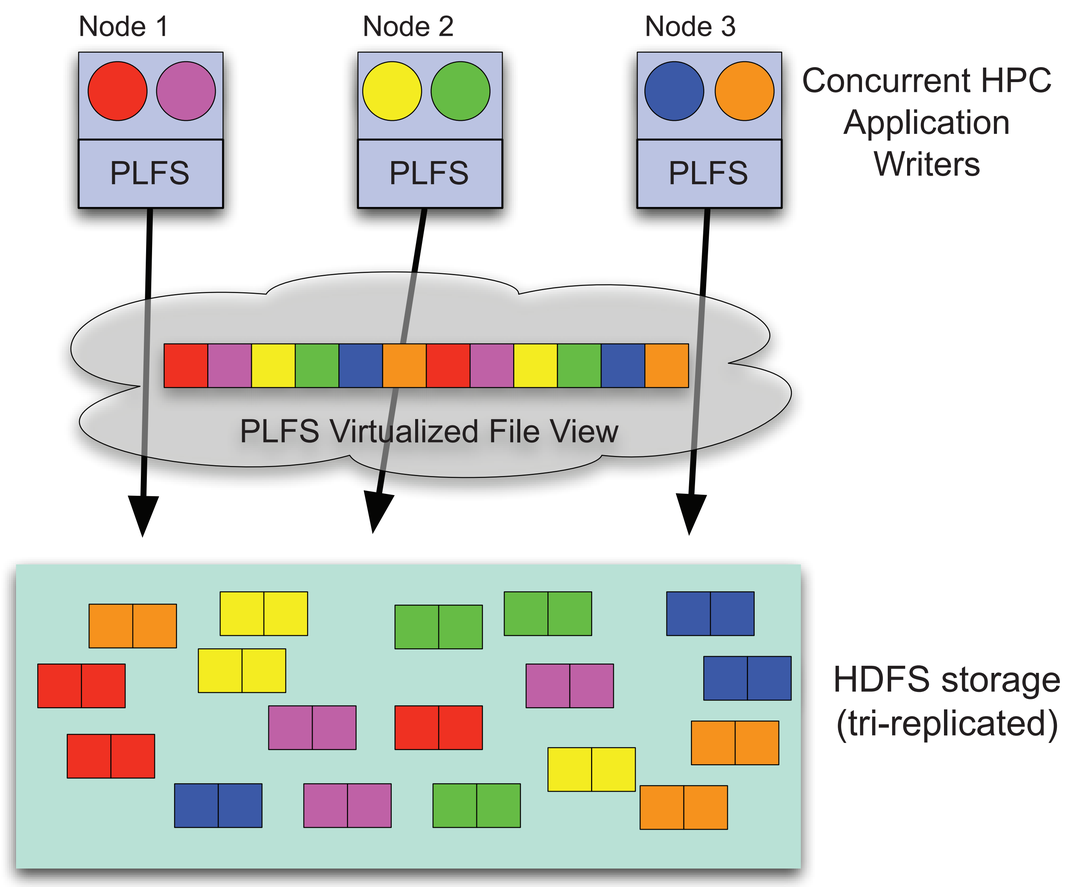
Current Status

- Collaboration between CMU, UCSU, LANL, LLNL
- Interposition layer written
- Current work is on reducing overhead
  - Presently significant overhead on reads
- Multiple test systems
  - OpenCloud, 64 cluster (CMU)
  - Tuson, 140 node data cluster (LLNL)
  - Test clusters (LANL)

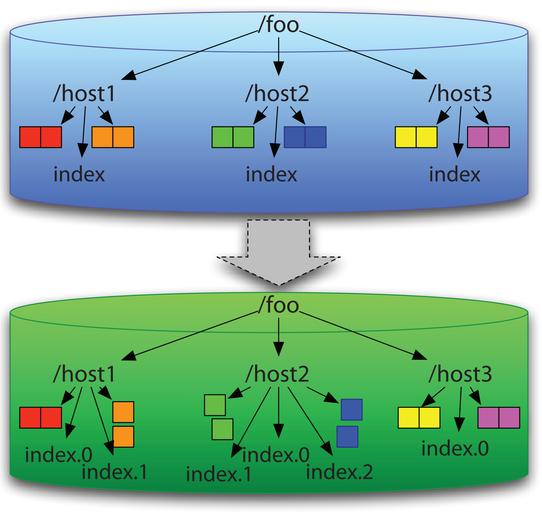
Parallel Log-Structured File System

- Decouples concurrent file access
  - Each writer gets exclusive log file
  - Each node gets exclusive index file
- Designed for checkpointing
- Used for HPC applications
- But we can use this functionality to enrich the semantics of HDFS

HPC APP ON PLFS-HDFS



PLFS CONTAINER FOR PLFS-HDFS



PLFS container for typical parallel file system. Contains a sub-directory per host, each containing an index file and a data log per writer. These files may grow throughout the lifetime of

PLFS container for HDFS. Still contains a sub-directory per host, but data logs and indices may consist of multiple files, one per open.

Carnegie Mellon

pdsi

IRHPIT INSTITUTE FOR RELIABLE HIGH PERFORMANCE INFORMATION TECHNOLOGY

Parallel Data Laboratory SVD10