OAK RIDGE National Laboratory

A Brief Introduction to UnifyFS

OLCF User Meeting – Oct 2022

LLNL: Kathryn Mohror (PI), Adam Moody, Cameron Stanavige

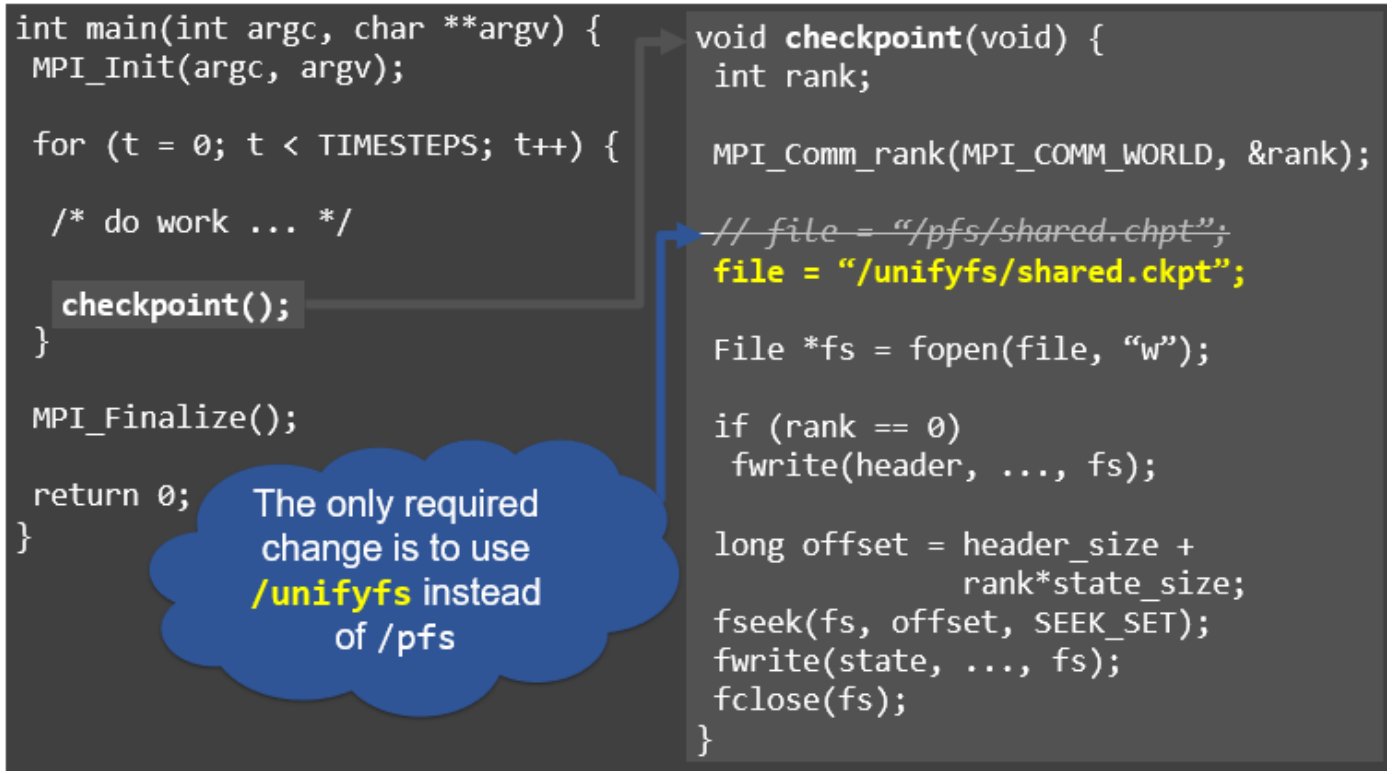
ORNL: Sarp Oral (ORNL-PI), Swen Boehm, Michael Brim, Seung-Hwan LIM, Ross Miller

ORNL is managed by UT-Battelle LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

What Is UnifyFS?

- An ephemeral, user-level shared file system for node-local storage
- Our goal is to make using node-local storage on exascale systems as *easy* as writing to the parallel file system and orders of magnitude *faster*



UnifyFS Overview

- Designed for efficient use of node-local storage for applications with typical HPC bulk-synchronous I/O patterns
  - Provides unified namespace over independent node-local storage
  - supports both shared files (N-1) and file-per-process (N-N)
- Transparently intercepts I/O: POSIX, MPI-IO, HDF5
- Usage: 3 Easy Steps
  1. Update application file paths to use prefix `/unifyfs`
     - often, this does not require code changes
  2. Link application with the UnifyFS client library
  3. Update job script to start/stop UnifyFS servers at beginning/end of job (with optional stage-in and stage-out)

UnifyFS File Sharing Semantics

- Application data written to node-local storage is only visible to remote application processes after an explicit sync operation
  - POSIX I/O: `fsync()` or `close()`
  - MPI-IO: `MPI_File_sync()`
  - HDF5: `H5Fflush()`

- File Lamination: mark file as read-only, no further writes allowed
  - allows UnifyFS to optimize metadata management for faster reads
  - to laminate, use UnifyFS API or `chmod(0444)`

UnifyFS - Summit Quickstart Guide

- `module use /sw/summit/unifyfs/modulefiles`
- `module avail unifyfs/1.0.0`
- `module load unifyfs/1.0.0/xl-16.1.1-10`
- `unifyfs-help # for info on job integration and configuration`

- Modules are also available for gcc 9.1 & 10.2
- For Help/Issues/Suggestions:
  - Do **not** contact OLCF help
  - Send an email to ecp-unifyfs@exascaleproject.org

Summary

- Modules are available on Summit (and will be on Frontier)
  - Remember to run `module use /sw/summit/unifyfs/modulefiles`
- UnifyFS is **EPHEMERAL**!
  - Save any data you want to keep to the parallel filesystem **before your job ends**
  - See the man page for the unifyfs-stage command
- Detailed documentation is available at https://unifyfs.readthedocs.io/en/dev/
  - Can also run the `unifyfs-help` command for more details

Questions?

Acknowledgements:

This research was supported by the Exascale Computing Project (17-SC-20-SC), a joint project of the U.S. Department of Energy's Office of Science and National Nuclear Security Administration, responsible for delivering a capable exascale ecosystem, including software, applications, and hardware technology, to support the nation's exascale computing imperative.

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.