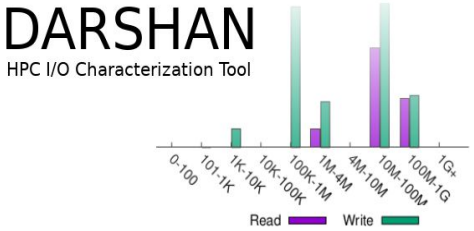
HEP-CCE IOS: Analysis of Root I/O in HEP Workflows with Darshan

Shane Snyder (ANL), Doug Benjamin (ANL), Patrick Gartung (FNAL), Ken Herner (FNAL)

Darshan background

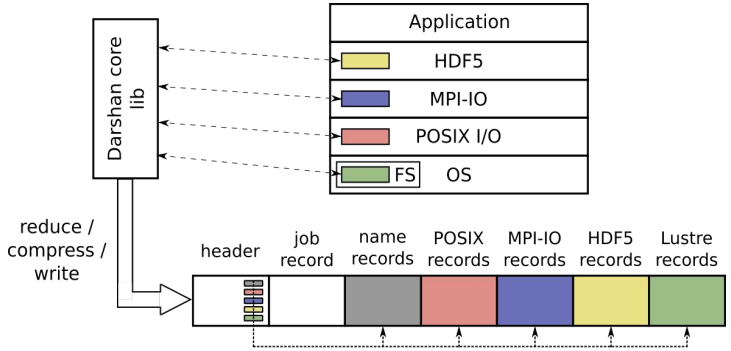


Darshan is a lightweight tool that can capture details about the I/O behavior of applications

- Inform tuning decisions of app scientists
- Gain insight into I/O trends on large-scale computing platforms

Darshan's design geared towards full-time deployment on HPC systems (currently on by default at ALCF, NERSC, OLCF, etc.)

- Transparent – no app changes required
- Low overhead – minimal perturbations to app runtime
- Modular – instrumentation can be extended to account for new I/O technologies



Default mode: capture bounded statistical records of I/O activity for each file accessed by the app

DXT (Darshan eXtended Tracing) mode: high-fidelity tracing of read/write operations

Darshan as a utility for HEP-CCE

Motivation: An ability to instrument the I/O behavior of HEP workflows is critical to characterizing and improving their usage of HPC storage

- ➢ The ongoing shift of HEP workflows to HPC facilities points to potential untapped I/O tuning opportunities here

Recent Darshan enhancements have broken its MPI dependency, enabling its use in new contexts, such as HEP workflow systems (traditionally non-MPI)

The IOS team has been an early power user of Darshan in non-MPI mode – and the use cases have spurred additional enhancements to the Darshan library!

- ➢ Proper handling of apps that `fork()` (e.g., ATLAS AthenaMP)
- ➢ Darshan runtime library config capability to fine-tune internal memory limits, focus instrumentation scope on particular apps/files, etc.

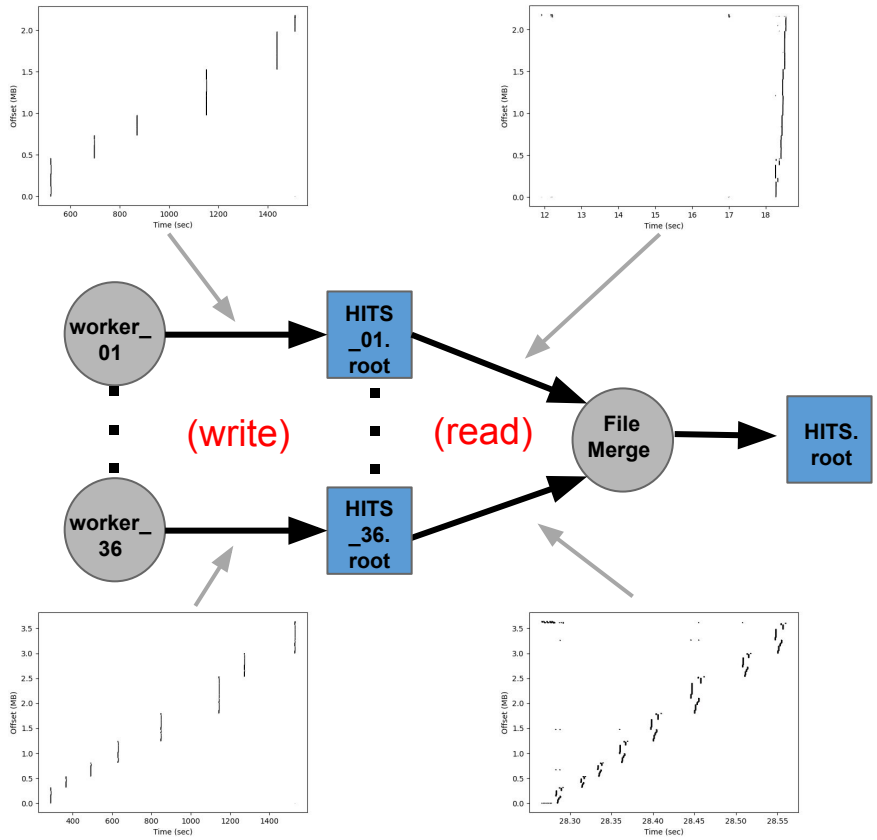
Darshan I/O trace analysis of ATLAS

ATLAS AthenaMP standard simulation, simulating 1000 events using 36 worker subprocesses

- Workers each process a number of events and generate HITS files, which are eventually merged into a single output HITS file

Figures show trace data for a portion of this workflow:

- (write) Workers write individual HITS file
- (read) File merge process reads workers HITS ahead of the merge to a single file



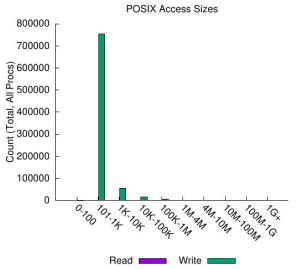
Darshan I/O analysis of CMS

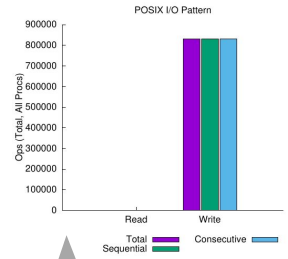
CMS workflow based on event generation, detector simulation, event reconstruction, and analysis

Figures show I/O performance determined by Darshan for various phases, with additional access info on 2 phases (RECO, EVENT SELECT)

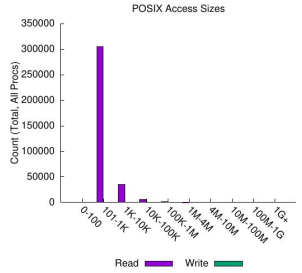
- Small I/O access sizes common across all phases – related to ROOT TTreeCache vector I/O support on certain FSes, worth continued investigation

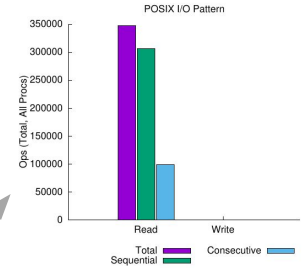
RECO write



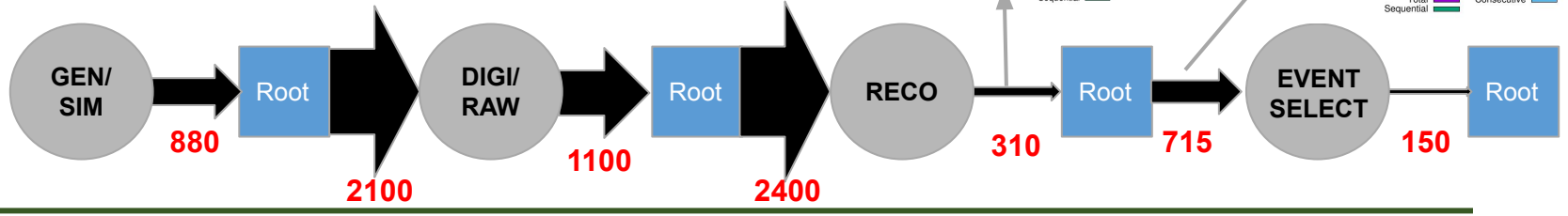


EVENT SELECT read





I/O perf. (MB/sec)



Potential next steps with Darshan in IOS

Utilize new Darshan instrumentation modules to better understand I/O behavior of other IOS activities

- *HDF5 module*: insights into ROOT→HDF5 serialization efforts
- *DAOS module*: insights into ROOT's RNTuple DAOS backend

Workflow-aware Darshan analysis tools to automate association of I/O activity across workflow steps