Darshan: state of the project and new features

Phil Carns and Shane Snyder
Mathematics and Computer Science Division
Argonne National Laboratory
carns@mcs.anl.gov, ssnyder@mcs.anl.gov

SC15 BoF: Analyzing Parallel I/O
Austin, Texas
November 2015

http://www.mcs.anl.gov/research/projects/darshan/

What is Darshan?

Darshan (Sanskrit for “sight”) is a scalable HPC I/O characterization tool. It is designed to capture an accurate picture of application I/O behavior with minimum overhead.

- No code changes, easy to use
  - *Negligible performance impact*: just “leave it on”
  - Enabled by default at ALCF, NERSC, and NCSA
  - Used on a case-by-case basis at many other sites

- Produces a summary of I/O activity for each job, including:
  - Counters for file access operations
  - Time stamps and cumulative timers for key operations
  - Histograms of access, stride, datatype, and extent sizes

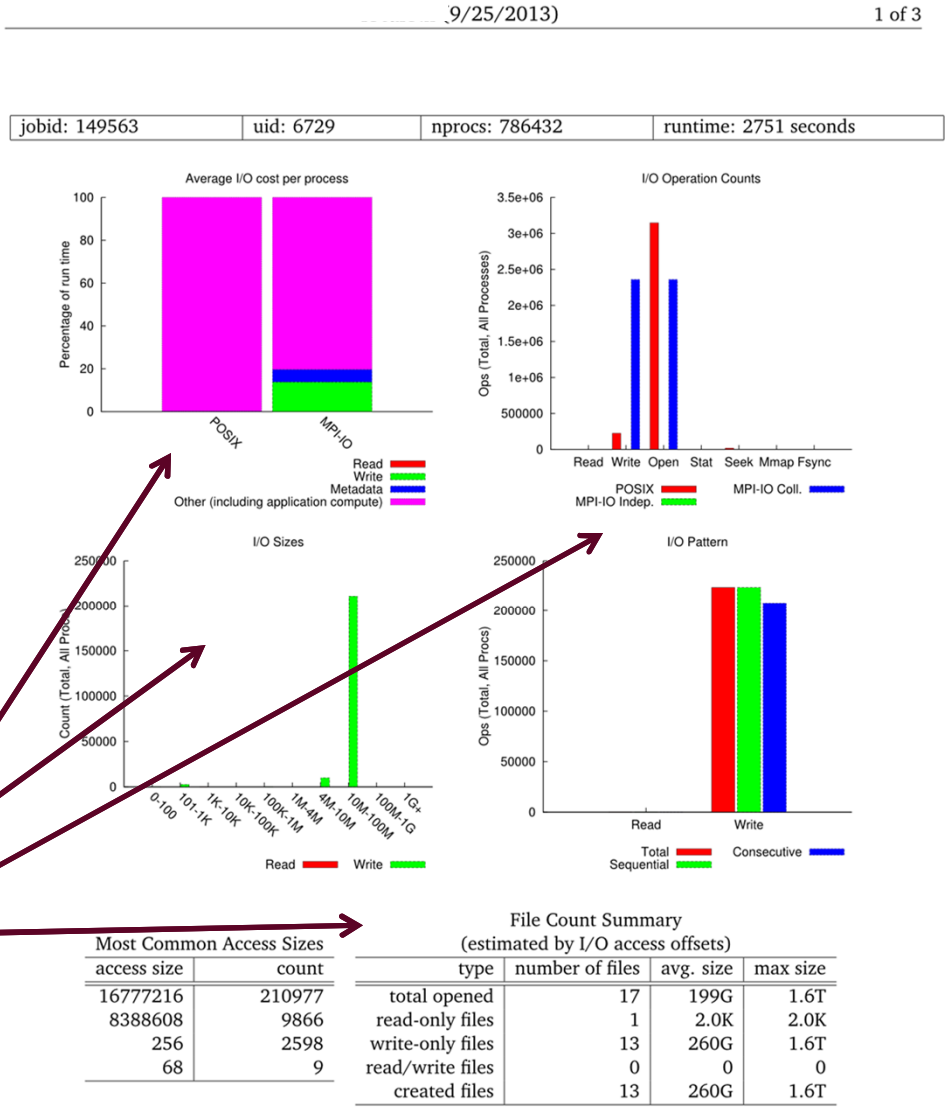
- Can be used to observe and tune individual applications or to capture a broad view of the platform workload

Darshan analysis example

Example: Darshan-job-summary.pl produces a 3-page PDF file summarizing various aspects of I/O performance

This figure shows the I/O behavior of a 786,432 process turbulence simulation (production run) on the Mira system at ANL

Percentage of runtime in I/O
Access size histogram
Access type histograms
File usage



Most Common Access Sizes

access size	count
16777216	210977
8388608	9866
256	2598
68	9

File Count Summary
(estimated by I/O access offsets)

type	number of files	avg. size	max size
total opened	17	199G	1.6T
read-only files	1	2.0K	2.0K
write-only files	13	260G	1.6T
read/write files	0	0	0
created files	13	260G	1.6T

/gpfs/mira-fs0/projects/WallModJet/cwhamman/wavelength/./rbturb.x 11 specs.in

What's new?

- Darshan was first released to the public in 2009
- Most recent releases:
  - Stable: Darshan 2.3.1 (March 2015)
  - Preview: Darshan 3.0.0-pre2 (November 2015)

- New features:

  - Integration with CODES workload model API (IOWA)
    COMPLETE (stable release)

  - Modularized library and file format
    COMPLETE (preview release)

  - Ability to capture data even if application exits abruptly
    IN PROGRESS

CODES workload model integration

- CODES exascale storage simulation toolkit includes an interface for generating I/O workloads called *IOWA*
  - Interface allows I/O workloads to be generated from a range of sources (I/O traces, I/O kernels, synthetic I/O descriptions, etc.)
  - I/O analysis tools can leverage workloads from different sources using consistent interface

- We have developed a workload generation method in IOWA for converting Darshan I/O logs into complete I/O workloads

- "Techniques for modeling large-scale HPC I/O workloads"
  - Presented @ PMBS 2015

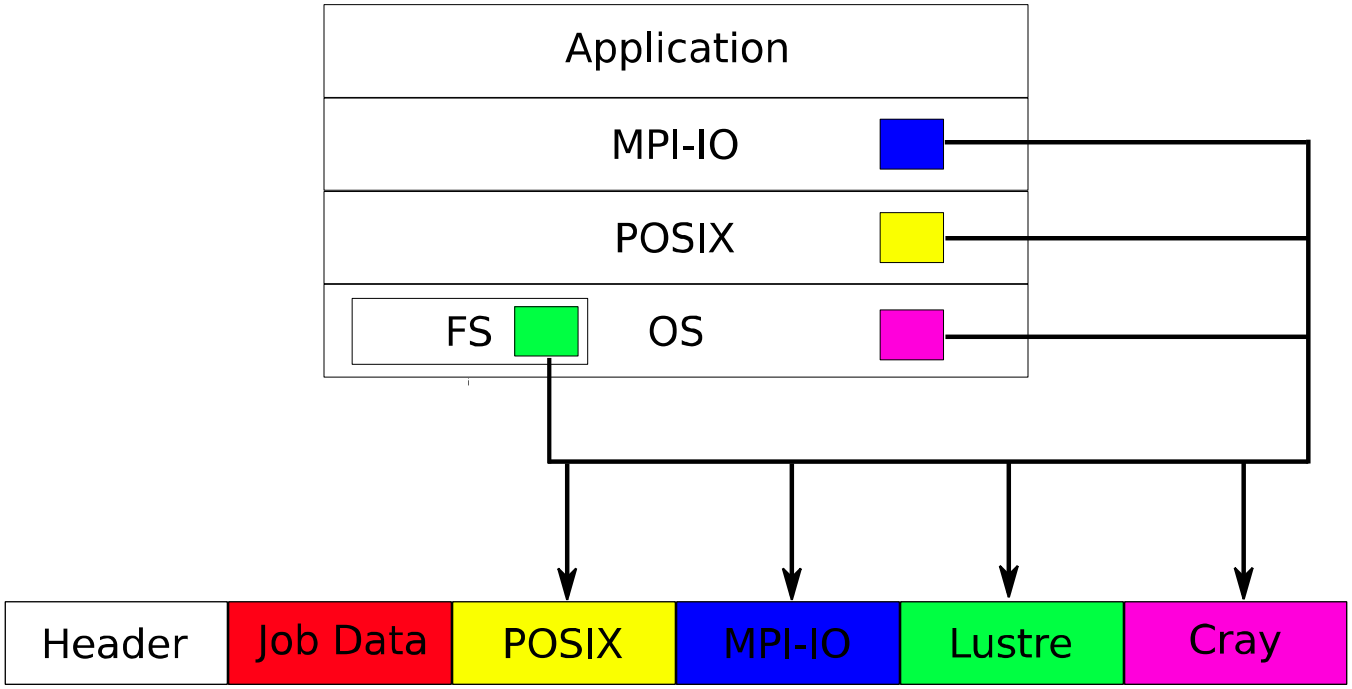
Modularized Darshan version (3.0)

- Darshan redesigned to expose an interface allowing the addition of new *instrumentation modules*
  - An *instrumentation module* is a Darshan component responsible for gathering I/O data from a specific system component
    - I/O libraries (POSIX, MPI-IO, HDF5, PnetCDF, …)
    - FS interfaces (e.g., Lustre API)
    - System-specific data (e.g., BG/Q or Cray)

- Instrumentation modules register with Darshan when they have data to contribute
  - Darshan assigns memory to modules for storing I/O data
  - Instrumentation module provide callback functions so Darshan can interface with them at shutdown time

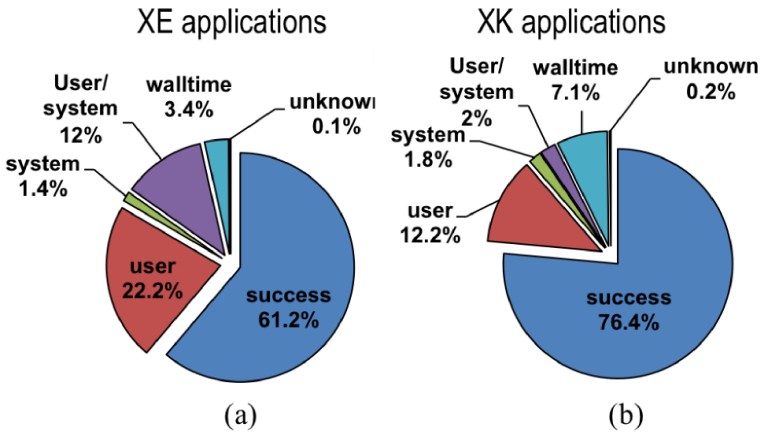
Modularized Darshan version (3.0)

- At shutdown, Darshan:
  - Retrieves I/O data from each module
  - Compresses data
  - Collectively writes data to Darshan log



Robust Darshan data capture

- Darshan coverage reduced by applications that do not shut down properly
  - Darshan's shutdown procedure hooks into MPI_Finalize()
  - Typical causes are running to wall-clock limit or crashing



Martino, Catello Di, et al. "LogDiver: A Tool for Measuring Resilience of Extreme-Scale Systems and Applications." Proceedings of the 5th Workshop on Fault Tolerance for HPC at eXtreme Scale. ACM, 2015.

- Approach:
  - Persist Darshan I/O characterization data structures to storage as application runs
    - I/O strategy?
    - Storage location?
    - Granularity of data?
  - Clean-up scripts to merge I/O characterization data structures into standard Darshan log format

For more information and downloads:

http://www.mcs.anl.gov/research/projects/darshan/

This material is based on work supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computer Research Program under contract DE-AC02-06CH11357. The research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is a DOE Office of Science User Facility.