SDAV
Scalable Data Management, Analysis, and Visualization

U.S. DEPARTMENT OF ENERGY | Office of Science

I/O CHARACTERIZATION OF LARGE-SCALE APPLICATIONS WITH DARSHAN

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

SDAV All-Hands Meeting
February 25-26, 2014

Motivation: understanding HPC I/O

- Leadership-class computing resources host hundreds of applications from diverse scientific domains
  - Wide variety of I/O strategies, I/O libraries, and I/O requirements
- Questions:
  - How do we tune applications quickly in this environment?
  - What are the overall trends in I/O behavior and system usage?
  - How can we share I/O behavior data with the research community?
- Constraints on methods to answer these questions:
  - *Don't perturb application behavior or performance*
  - Low (or preferably, non-existent) barriers to entry for scientists to participate
  - Portability across platforms
- Darshan was developed to address these issues

Darshan is a lightweight, scalable I/O characterization tool that transparently captures I/O access pattern information from production applications.

Darshan overview

- Runtime library for characterization of application I/O
  - Instrumentation is inserted at build time (for static executables) or at run time (for dynamic executables)
  - Captures POSIX I/O, MPI-IO, and limited HDF5 and PNetCDF functions
- Minimal application impact
  - Bounded memory consumption per process
  - Records strategically chosen counters, timestamps, and histograms
  - Reduces, compresses, and aggregates data at MPI_Finalize() time
- Compatible with IBM BG, Cray, and Linux environments
  - Deployed system-wide or enabled by individual users
  - Instrumentation is enabled via software modules, environment variables, or compiler scripts
  - No source code modifications or changes to build rules
  - No file system dependencies
  - Currently beta testing Cray PE 2.x support for XC30 systems

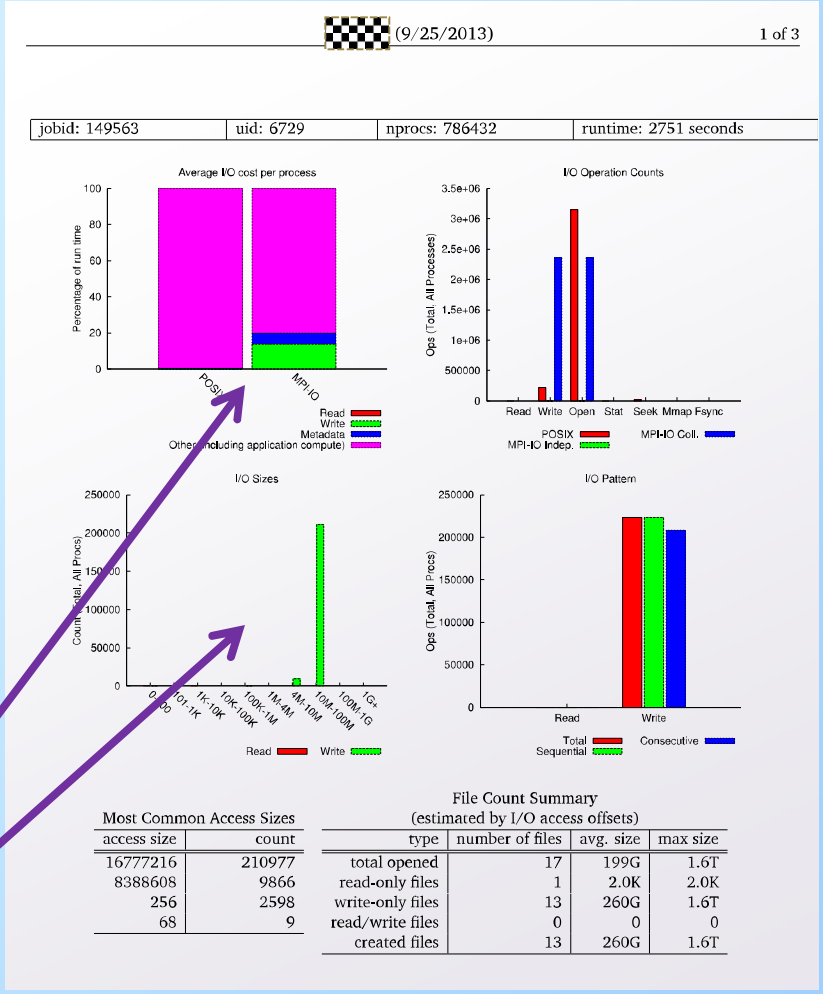
Darshan analysis tools

- Each job instrumented with Darshan produces a compact characterization log file
- Darshan command line utilities can be used to analyze these log files
- 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 Mira
- Application is write intensive and benefits greatly from collective buffering

Example measurements: % of runtime in I/O

access size histogram



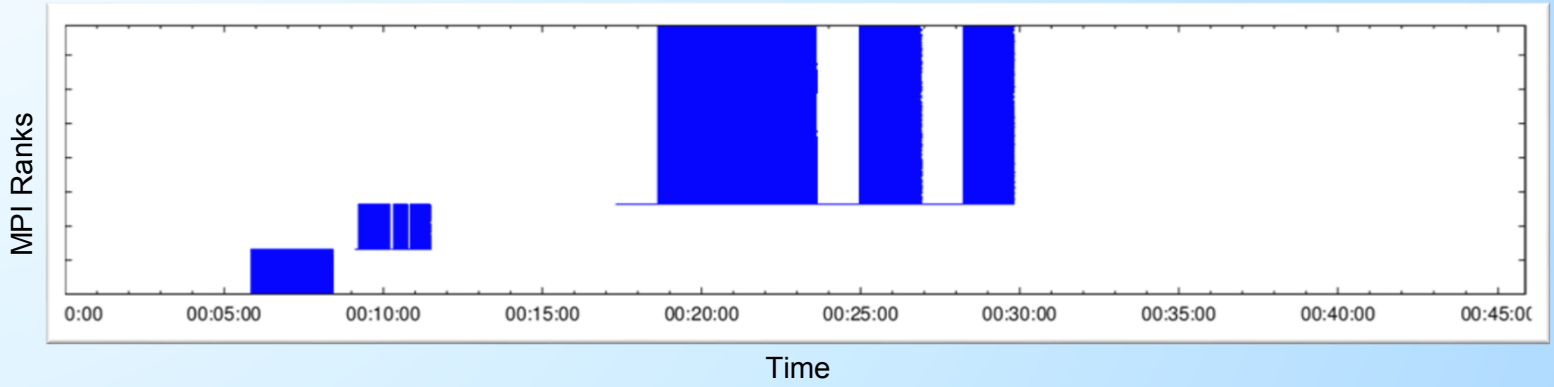
Most Common Access Sizes

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

Darshan analysis tools

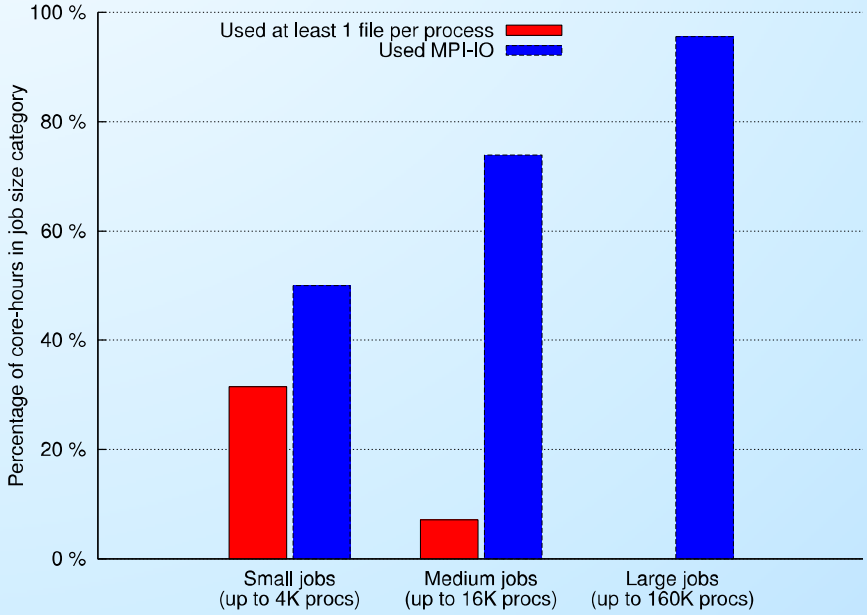


Darshan can also show intervals of I/O activity from each process. In this example we see that subsets of processes write data at different times during execution.

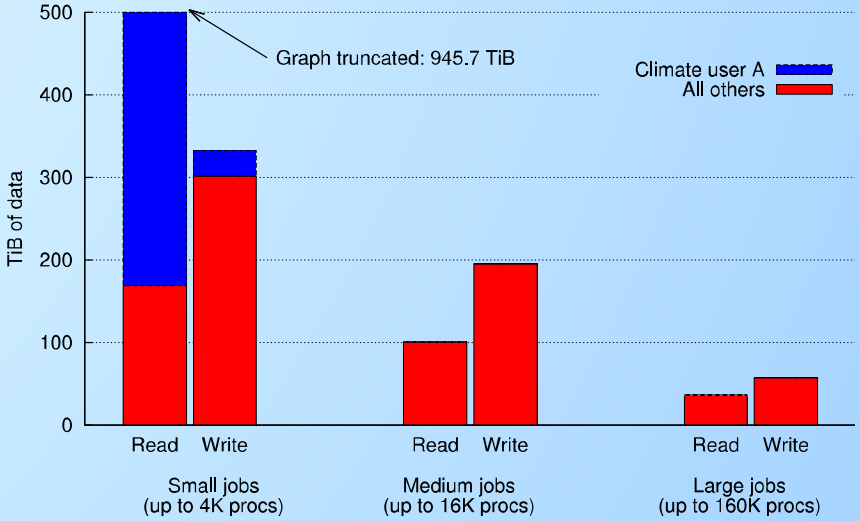
A variety of applications have been tuned with assistance from Darshan, including HSCD, FLASH, NekCEM, PHASTA, and pF3D.

Understanding system usage

- Aggregate data from Darshan can provide a broad view of system usage and application trends
- Examples from 4 months of data collected on Intrepid BG/P system in 2013:



MPI-IO usage becomes more prevalent at scale, while file-per-process access patterns become less prevalent.



This graph shows I/O volume at various scales. Write-intensive behavior is evident in each category. This data also illustrates how a single user on the system can dominate I/O activity in some cases.

Sharing data with the community

- Conversion utilities can anonymize and re-compress data
- Compact data format in conjunction with anonymization makes it possible to share data with the community in bulk
- The ALCF I/O Data Repository provides access to production logs captured on Intrepid, including all 2012 and 2013 activity

ALCF I/O Data Repository Statistics	
Unique log files	152,167
Core-hours instrumented	721 million
Data read	25.2 petabytes
Data written	5.7 petabytes

Current status

- Automatically enabled for all users on
  - Mira IBM BG/Q system at ALCF
    (Kevin Harms)
  - Hopper Cray XE6 system at NERSC
    (Yushu Yao and Katie Antypas)
  - Other facilities are evaluating deployment as well
- Future work
  - 2.2.9 release will focus on minor bug fixes and additional platform support
  - Long term: broader, modular instrumentation
  - More facility deployments
  - Helping more researchers leverage Darshan data
    - Examples: identifying I/O motifs (Huong Luu, UIUC) and modeling I/O workloads for HPC system simulations (Shane Snyder, ANL)