CSE 5449: Intermediate Studies in Scientific Data Management

Lecture 11: Tools for understanding parallel I/O performance - Darshan

Dr. Suren Byna

The Ohio State University

E-mail: byna.1@osu.edu

https://sbyna.github.io

02/14/2023

Today's class

- Any questions?
- Class presentation topic
- Today's class –
  - Tools for understanding parallel I/O performance

Class presentation topic – I/O performance analysis

- Glenn K. Lockwood, Shane Snyder, Teng Wang, Suren Byna, Philip Carns, Nicholas J. Wright. "A Year in the Life of a Parallel File System." In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'18)*. Dallas, TX. November 2018. (Slides)

- Teng Wang, Shane Snyder, Glenn K. Lockwood, Philip Carns, Nicholas J. Wright, and Suren Byna. "IOMiner: Large-Scale Analytics Framework for Gaining Knowledge from I/O Logs." In *Proceedings of the 2018 IEEE International Conference on Cluster Computing (CLUSTER)*. Belfast, UK. September 2018.

- Glenn K. Lockwood, Shane Snyder, George Brown, Kevin Harms, Philip Carns, Nicholas J. Wright. "TOKIO on ClusterStor: Connecting Standard Tools to Enable Holistic I/O Performance Analysis." In *Proceedings of the 2018 Cray User Group*. Stockholm, SE. May 2018. (Slides)

- Glenn K. Lockwood, Wucherl Yoo, Suren Byna, Nicholas J. Wright, Shane Snyder, Kevin Harms, Zachary Nault, Philip Carns. "UMAMI: a recipe for generating meaningful metrics through holistic I/O performance analysis." In *Proceedings of the 2nd Joint International Workshop on Parallel Data Storage & Data Intensive Scalable Computing Systems (PDSW-DISCS'17)*. Denver, CO. November 2017. (Slides)

Links to papers: https://www.nersc.gov/research-and-development/storage-and-i-o-technologies/tokio/

Class presentation format

- Main goal(s)
- Motivation
- Prior work
- Proposed solutions
- How were the solutions evaluated and what was achieved?
- Future work – *You may add your ideas here.*
- Gap analysis – What research gaps are still open?

Presentation on March 9th

Factors that impact parallel I/O performance

Applications

- Number of MPI ranks
- Number of I/O requests
- Size of I/O requests
- Number of files
- Number of metadata calls
  - File open and close requests
- Number of seek operations
- Contiguous / non-contiguous requests
  - Number of seeks
- Alignment of I/O request with
  - File block
  - Sub-files
- Shared file or multiple files
- …

High-level I/O library

- Metadata operations for self-describing property
- Location of metadata
- How many processes are participating in metadata or data operations
- Alignment in file offsets
- Hyperslab selections
  - contiguous / non-contiguous?
  - complex hyperslabs construction cost
- Chunking
  - Chunk size
  - Number of chunks
- Sub-files
  - How many? How's the data aggregated?
- Compression used or not?
  - What's the compression / decompression cost?
  - Where is compression / decompression executed?
- Does a file need to be exact size of data or can it have some gaps?
- Cache metadata or not?

MPI-IO

- Contiguous / non-contiguous accesses
- Number of I/O requests
- Size of I/O requests
- POSIX consistency semantics
- Synchronous / Asynchronous I/O calls
- Collective or independent
- If collective:
  - Number of aggregators
  - Aggregator placement
  - Aggregation buffer size
  - Aggregator to file system mapping – network connections and block sizes

File systems

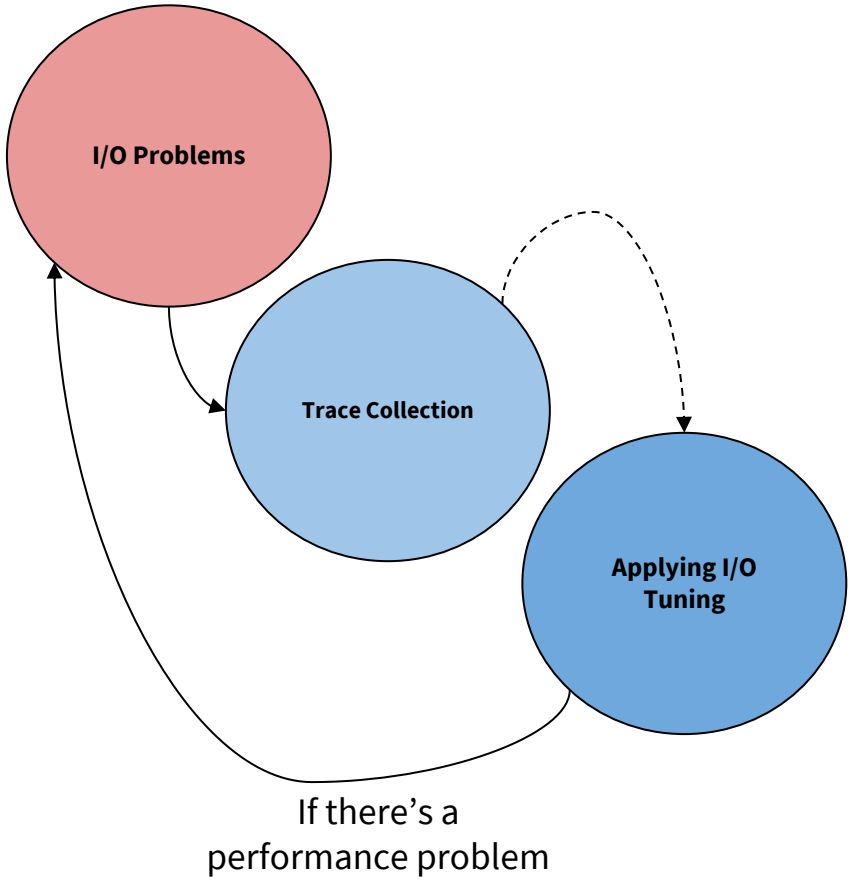
- Number of storage servers
- Number of metadata servers
- Number of storage targets (stripe count)
- Block size on storage server
- Page size on storage target
- Amount of contiguous data stored on a storage target (stripe size)
- Traffic on storage targets
- Fullness of storage targets
- Fragmentation on storage targets

Tools for understanding parallel I/O performance

- Darshan (ANL)
- Darshan Extended Trace (DXT) -- Intel, LBNL, & ANL
- DXT Explorer -- LBNL
- Drishti -- LBNL

Path to understand I/O performance and optimize

- There are several tools available to trace I/O performance
  - Darshan
  - Recorder



Darshan I/O

- A lightweight HPC I/O characterization tool to capture application I/O behavior
  - Provides a summary of I/O calls
    - Counts, histograms, timers, and other statistics
    - Extended tracing for full I/O activity traces

- Developed at Argonne National Laboratory

- Deployed on several supercomputing facilities
- https://www.mcs.anl.gov/research/projects/darshan/

Darshan – How does it work?

- `darshan-runtime` and `darshan-util`

- Instrumentation of I/O calls
  - At link time of application OR
  - At runtime (using LD_PRELOAD)

- Collects file access statistics
  - HDF5, MPI-IO, POSIX-IO, File system layers
  - Computes statistics
  - Compresses the logs and writes

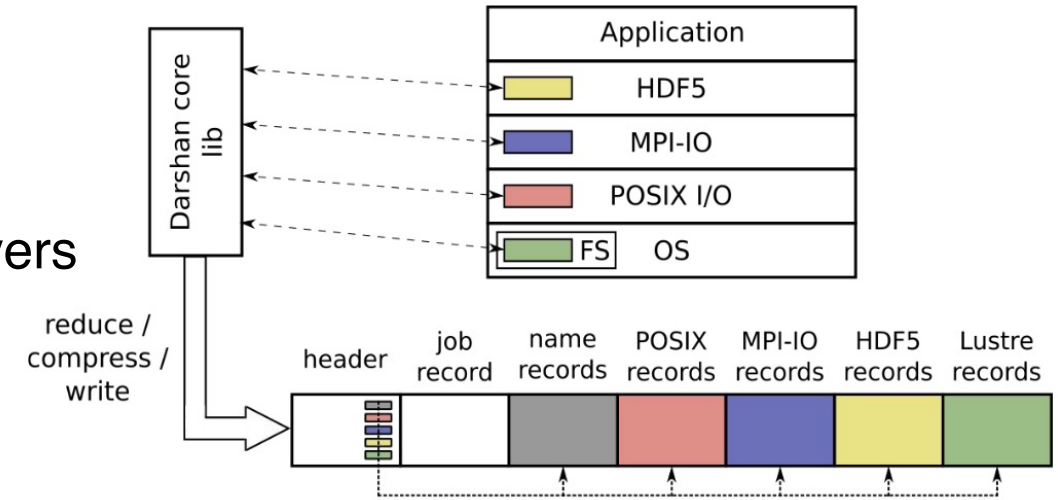


Image from Shane Snyder's Darshan tutorial

Using Darshan on NERSC systems

- Darshan module is available on NERSC systems (may be available at OSC as well)
- Run “module list” command to see if Darshan is loaded and which version
- After compiling and running your job with Darshan loaded, run “‘`darshan-config --log-path`’ to find where the logs are stored

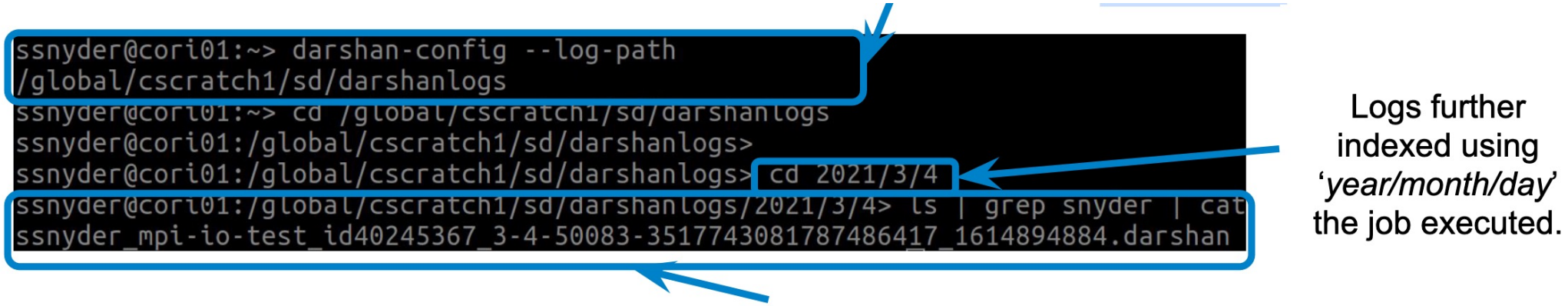


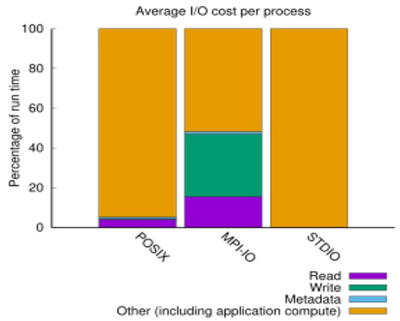
Image from Shane Snyder’s Darshan tutorial

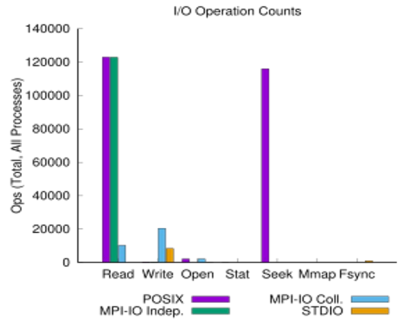
Darshan analysis tools

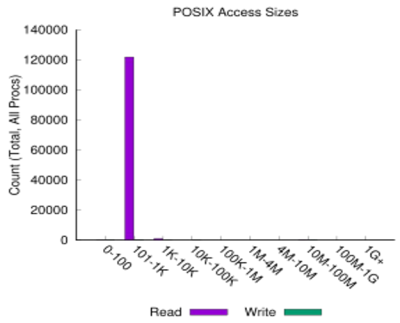
- darshan-util package
- `darshan-job-summary`
  - provides a summary characterizing application I/O behavior in a PDF format

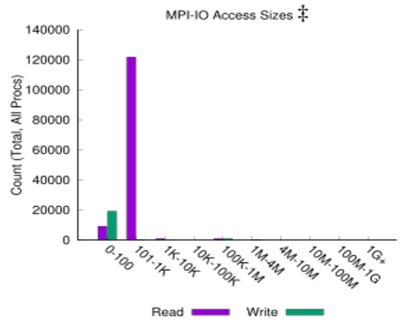
jobid: 45195555	uid: 95230	nprocs: 1024	runtime: 6 seconds

I/O performance *estimate* (at the MPI-IO layer): transferred 5072.0 MiB at 798.82 MiB/s
I/O performance *estimate* (at the STDIO layer): transferred 0.1 MiB at 37.30 MiB/s



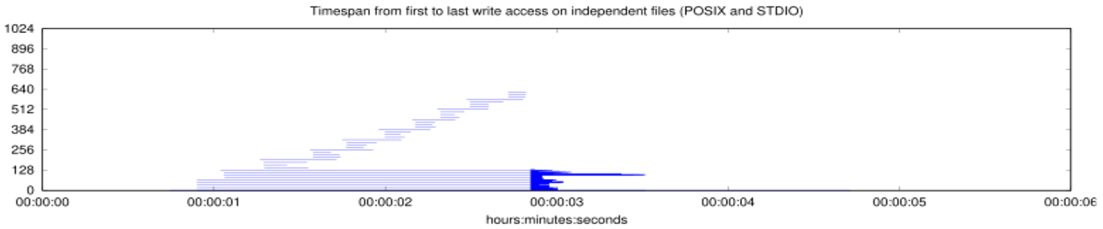


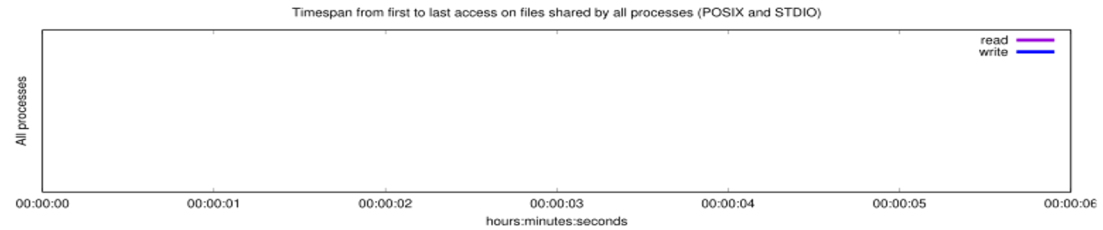




Darshan analysis tools

- `darshan-job-summary`
  - provides a summary characterizing application I/O behavior in a PDF format





Average I/O per process (POSIX and STDIO)

	Cumulative time spent in I/O functions (seconds)	Amount of I/O (MB)
Independent reads	0.243677474609375	2.4876377210021
Independent writes	0.024767330078125	2.46554678305984
Independent metadata	0.042551724609375	N/A
Shared reads	0	0
Shared writes	0	0
Shared metadata	0	N/A

Data Transfer Per Filesystem (POSIX and STDIO)

File System	Write		Read	
	MiB	Ratio	MiB	Ratio
UNKNOWN	0.07063	0.00003	0.00000	0.00000
/global/cscratch1	2524.64928	0.99997	2547.34103	1.00000

Darshan analysis tools

- `darshan-parser` – prints all counters in a log file to a text format file

nc5_sml.txt

```
# darshan log version: 3.10
# compression method: ZLIB
# exe: /global/homes/t/tonglin/e2e-hpdc2011/3d//write_3d_nc5 /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes 28 16 16 32 32 32
# uid: 77441
# jobid: 36747741
# start_time: 1606859889
# start_time_asci: Tue Dec  1 13:58:09 2020
# end_time: 1606859895
# end_time_asci: Tue Dec  1 13:58:15 2020
# nprocs: 32
# run time: 7
# metadata: lib_ver = 3.1.7
# metadata: h = romio_no_indep_rw=true;cb_nodes=4

# log file regions
# -------------------------------------------------------
# header: 360 bytes (uncompressed)
# job data: 595 bytes (compressed)
# record table: 137 bytes (compressed)
# POSIX module: 4727 bytes (compressed), ver=4
# MPI-IO module: 3341 bytes (compressed), ver=3
# PNETCDF module: 1398 bytes (compressed), ver=2
# LUSTRE module: 14188 bytes (compressed), ver=1
# STDIO module: 51 bytes (compressed), ver=2
# DXT_POSIX module: 2928986 bytes (compressed), ver=1
# DXT_MPIIO module: 6014 bytes (compressed), ver=1
```

Darshan analysis tools

- `darshan-parser` – prints all counters in a log file to a text format file

```
# description of POSIX counters:
#   POSIX_*: posix operation counts.
#   READS,WRITES,OPENS,SEEKS,STATS,MMAPS,SYNCS,FILENOS,DUPS are types of operations.
#   POSIX_RENAME_SOURCES/TARGETS: total count file was source or target of a rename operation
#   POSIX_RENAMED_FROM: Darshan record ID of the first rename source, if file was a rename target
#   POSIX_MODE: mode that file was opened in.
#   POSIX_BYTES_*: total bytes read and written.
#   POSIX_MAX_BYTE_*: highest offset byte read and written.
#   POSIX_CONSEC_*: number of exactly adjacent reads and writes.
#   POSIX_SEQ_*: number of reads and writes from increasing offsets.
#   POSIX_RW_SWITCHES: number of times access alternated between read and write.
#   POSIX_*_ALIGNMENT: memory and file alignment.
#   POSIX_*_NOT_ALIGNED: number of reads and writes that were not aligned.
#   POSIX_MAX_*_TIME_SIZE: size of the slowest read and write operations.
#   POSIX_SIZE_*_*: histogram of read and write access sizes.
#   POSIX_STRIDE*_STRIDE: the four most common strides detected.
#   POSIX_STRIDE*_COUNT: count of the four most common strides.
#   POSIX_ACCESS*_ACCESS: the four most common access sizes.
#   POSIX_ACCESS*_COUNT: count of the four most common access sizes.
#   POSIX_*_RANK: rank of the processes that were the fastest and slowest at I/O (for shared files).
#   POSIX_*_RANK_BYTES: bytes transferred by the fastest and slowest ranks (for shared files).
#   POSIX_F_*_START_TIMESTAMP: timestamp of first open/read/write/close.
#   POSIX_F_*_END_TIMESTAMP: timestamp of last open/read/write/close.
#   POSIX_F_READ/WRITE/META_TIME: cumulative time spent in read, write, or metadata operations.
#   POSIX_F_MAX_*_TIME: duration of the slowest read and write operations.
#   POSIX_F_*_RANK_TIME: fastest and slowest I/O time for a single rank (for shared files).
#   POSIX_F_VARIANCE_RANK_*: variance of total I/O time and bytes moved for all ranks (for shared files).
```

Darshan analysis tools

- `darshan-parser` – prints all counters in a log file to a text format file

```
# ****************************************************
# MPI-IO module data
# ****************************************************

# description of MPIIO counters:
#   MPIIO_INDEP_*: MPI independent operation counts.
#   MPIIO_COLL_*: MPI collective operation counts.
#   MPIIO_SPLIT_*: MPI split collective operation counts.
#   MPIIO_NB_*: MPI non blocking operation counts.
#   READS,WRITES,and OPENS are types of operations.
#   MPIIO_SYNCS: MPI file sync operation counts.
#   MPIIO_HINTS: number of times MPI hints were used.
#   MPIIO_VIEWS: number of times MPI file views were used.
#   MPIIO_MODE: MPI-IO access mode that file was opened with.
#   MPIIO_BYTES_*: total bytes read and written at MPI-IO layer.
#   MPIIO_RW_SWITCHES: number of times access alternated between read and write.
#   MPIIO_MAX_*_TIME_SIZE: size of the slowest read and write operations.
#   MPIIO_SIZE_*_AGG_*: histogram of MPI datatype total sizes for read and write operations.
#   MPIIO_ACCESS*_ACCESS: the four most common total access sizes.
#   MPIIO_ACCESS*_COUNT: count of the four most common total access sizes.
#   MPIIO_*_RANK: rank of the processes that were the fastest and slowest at I/O (for shared files).
#   MPIIO_*_RANK_BYTES: total bytes transferred at MPI-IO layer by the fastest and slowest ranks (for shared files).
#   MPIIO_F_*_START_TIMESTAMP: timestamp of first MPI-IO open/read/write/close.
#   MPIIO_F_*_END_TIMESTAMP: timestamp of last MPI-IO open/read/write/close.
#   MPIIO_F_READ/WRITE/META_TIME: cumulative time spent in MPI-IO read, write, or metadata operations.
#   MPIIO_F_MAX_*_TIME: duration of the slowest MPI-IO read and write operations.
#   MPIIO_F_*_RANK_TIME: fastest and slowest I/O time for a single rank (for shared files).
#   MPIIO_F_VARIANCE_RANK_*: variance of total I/O time and bytes moved for all ranks (for shared files).
```

```
#<module>      <rank>  <record id>     <counter>       <value> <file name>     <mount pt>      <fs type>
MPI-IO  0       14892389100877985224    MPIIO_INDEP_OPENS       0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_COLL_OPENS        1       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_INDEP_READS       0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_INDEP_WRITES      1       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_COLL_READS        0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_COLL_WRITES       10      /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SPLIT_READS       0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SPLIT_WRITES      0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_NB_READS  0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_NB_WRITES 0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SYNCS     0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_HINTS     1       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_VIEWS     10      /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_MODE      9       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_BYTES_READ        0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_BYTES_WRITTEN     2622032 /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_RW_SWITCHES       0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_MAX_READ_TIME_SIZE        0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_MAX_WRITE_TIME_SIZE       262144  /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_READ_AGG_0_100       0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_READ_AGG_100_1K      0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_READ_AGG_1K_10K      0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_READ_AGG_10K_100K    0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_READ_AGG_100K_1M     0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_READ_AGG_1M_4M       0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_READ_AGG_4M_10M      0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_READ_AGG_10M_100M    0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_READ_AGG_100M_1G     0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_READ_AGG_1G_PLUS     0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_WRITE_AGG_0_100      0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_WRITE_AGG_100_1K     1       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_WRITE_AGG_1K_10K     0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_WRITE_AGG_10K_100K   0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_WRITE_AGG_100K_1M    10      /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_WRITE_AGG_1M_4M      0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_WRITE_AGG_4M_10M     0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_WRITE_AGG_10M_100M   0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_WRITE_AGG_100M_1G    0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SIZE_WRITE_AGG_1G_PLUS    0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_ACCESS1_ACCESS    262144  /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_ACCESS2_ACCESS    592     /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_ACCESS3_ACCESS    0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_ACCESS4_ACCESS    0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_ACCESS1_COUNT     10      /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_ACCESS2_COUNT     1       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_ACCESS3_COUNT     0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_ACCESS4_COUNT     0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_FASTEST_RANK      0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_FASTEST_RANK_BYTES        0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SLOWEST_RANK      0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_SLOWEST_RANK_BYTES        0       /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1       lustre
MPI-IO  0       14892389100877985224    MPIIO_F_OPEN_START_TIMESTAMP    0.027893        /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36747741-1-nodes.nc5 /global/cscratch1
        lustre
```

Darshan Extended Tracing - DXT

- Enhance Darshan to (optionally) report every intercepted call
- Traces appear as a time series and can be post-processed offline
- Provide tools for applying different types of analyses to the logs
- Aggregate statistics and/or drill down to any level of granularity

DXT components

- Logging
  - Records each intercepted I/O call
  - Request offset, length, start time, end time, MPI rank and the hostname
  - Can be switched on or off at runtime using an environment variable
  - Log buffer starts small and expands gradually as needed
  - Uses compression to limit the size of the output log file
- Analysis
  - Correlates traces with Lustre striping information
  - Group/filter requests by rank, host or Lustre OST

Cong Xu, Shane Snyder, Omkar Kulkarni, Vishwanath Venkatesan, Philip Carns, Suren Byna, Robert Sisneros, and Kalyana Chadalavada, "DXT: Darshan eXtended Tracing", Cray User Group Conference 2017 (CUG 2017)

dxt-parser

- Enable DXT
  - `setenv DXT_ENABLE_IO_TRACE 1`
- Copy the Darshan file to your directory
- Run DXT parser
  - `darshan-dxt-parser some-darshan-log-file.darshan > ~/trace.txt`

```
# ***************************************************
# DXT_POSIX module data
# ***************************************************

# DXT, file_id: 5076057741753365924, file_name: /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36745115-1-nodes.nc4
# DXT, rank: 0, hostname: nid00604
# DXT, write_count: 10249, read_count: 0
# DXT, mnt_pt: /global/cscratch1, fs_type: lustre
# DXT, Lustre stripe_size: 16777216, Lustre stripe_count: 244
# DXT, Lustre OST obdidx: 132 52 146 214 86 200 176 24 16 6 224 76 90 198 190 112 114 58 78 102 74 32 68 36 48 208 30 194 238 182 126 96
28 142 188 34 44 22 164 54 140 92 110 20 156 62 72 150 84 144 94 128 38 202 8 148 134 158 186 98 46 138 154 168 108 82 106 80 0 136 210
118 4 10 40 14 184 196 172 18 12 174 116 162 64 120 50 166 56 26 192 180 178 104 170 124 42 122 152 130 70 100 160 88 247 243 227 219 215
177 233 221 223 207 89 229 91 213 237 199 205 245 209 193 155 189 123 149 211 169 235 145 201 81 157 21 97 165 175 179 143 161 31 53 41
181 231 225 183 67 129 119 85 71 77 5 29 107 61 9 113 11 147 103 13 111 133 33 63 121 127 141 35 93 101 109 75 23 99 117 167 49 185 115 7
135 3 57 95 43 27 191 1 163 51 15 153 187 55 151 239 79 25 137 47 217 17 39 59 171 69 173 37 203 125 131 87 19 195 65 45 139 105 241 83
159 73 197 2 216 234 218 222 246 220 226 232 244 206 212 236 228 240 242
# Module    Rank  Wt/Rd  Segment          Offset       Length    Start(s)      End(s)  [OST]
 X_POSIX       0  write        0               0         1955      0.1331      0.1359  [132]
 X_POSIX       0  write        1      3758106112         2038      0.1360      0.1372  [ 83]
 X_POSIX       0  write        2     11274300928         1978      0.1372      0.1384  [  7]
 X_POSIX       0  write        3     18790497792         4096      0.1384      0.1391  [ 41]
 X_POSIX       0  write        4     18790501888          366      0.1391      0.1391  [ 41]
 X_POSIX       0  write        5     18790502254          366      0.1391      0.1392  [ 41]
 X_POSIX       0  write        6     18790502922          150      0.1392      0.1392  [ 41]
 X_POSIX       0  write        7     18790503072          366      0.1392      0.1392  [ 41]
 X_POSIX       0  write        8            9728          256      0.6889      0.6894  [132]
 X_POSIX       0  write        9           13824          256      0.6894      0.6922  [132]
 X_POSIX       0  write       10           17920          256      0.6922      0.6926  [132]
 X_POSIX       0  write       11           22016          256      0.6926      0.6930  [132]
 X_POSIX       0  write       12           26112          256      0.6930      0.6937  [132]
 X_POSIX       0  write       13           30208          256      0.6937      0.6942  [132]
 X_POSIX       0  write       14           34304          256      0.6943      0.6946  [132]
 X_POSIX       0  write       15           38400          256      0.6946      0.6951  [132]
 X_POSIX       0  write       16           42496          256      0.6951      0.6956  [132]
 X_POSIX       0  write       17           46592          256      0.6956      0.6961  [132]
 X_POSIX       0  write       18           50688          256      0.6961      0.6966  [132]
 X_POSIX       0  write       19           54784          256      0.6966      0.6970  [132]
 X_POSIX       0  write       20           58880          256      0.6970      0.6974  [132]
```

```
# ***************************************************
# DXT_MPIIO module data
# ***************************************************

# DXT, file_id: 5076057741753365924, file_name: /global/cscratch1/sd/tonglin/data_e2e/3d_28_16_16_32_32_32-36745115-1-nodes.nc4
# DXT, rank: 0, hostname: nid00604
# DXT, write_count: 12, read_count: 0
# DXT, mnt_pt: /global/cscratch1, fs_type: lustre
# Module    Rank  Wt/Rd  Segment       Length    Start(s)      End(s)
 X_MPIIO       0  write        0          331      0.1315      0.1392
 X_MPIIO       0  write        1       262144      0.6802      1.1448
 X_MPIIO       0  write        2       262144      1.1451      1.7255
 X_MPIIO       0  write        3       262144      1.7257      2.3791
 X_MPIIO       0  write        4       262144      2.3794      3.0459
 X_MPIIO       0  write        5       262144      3.0462      4.2975
 X_MPIIO       0  write        6       262144      4.2978      5.4152
 X_MPIIO       0  write        7       262144      5.4154      6.3356
 X_MPIIO       0  write        8       262144      6.3358      7.0600
 X_MPIIO       0  write        9       262144      7.0602      7.8717
 X_MPIIO       0  write       10       262144      7.8719      8.7132
 X_MPIIO       0  write       11           96      9.5743      9.5746
```

More details on Darshan Utilities:
https://www.mcs.anl.gov/research/projects/darshan/docs/darshan-util.html

Darshan - Homework

- On NERSC's Cori system, perform Darshan analysis
  - For the h5bench write pattern, collect Darshan log
  - Run
    - `darshan-parser`
    - `darshan-job-summary`
    - `darshan-dxt-parser`

Due on: Feb 20th

Summary of today's class

- Parallel I/O performance factors and some application tuning examples
- Next Class – Tracing parallel I/O performance, visualizing
- Class presentation on March 9th
- Homework: Darshan analysis
  - Due on: Feb 20th