Trying Darshan on a DUNE Workflow

Ken Herner

HEP-CCE IOS

9 Sep 2020

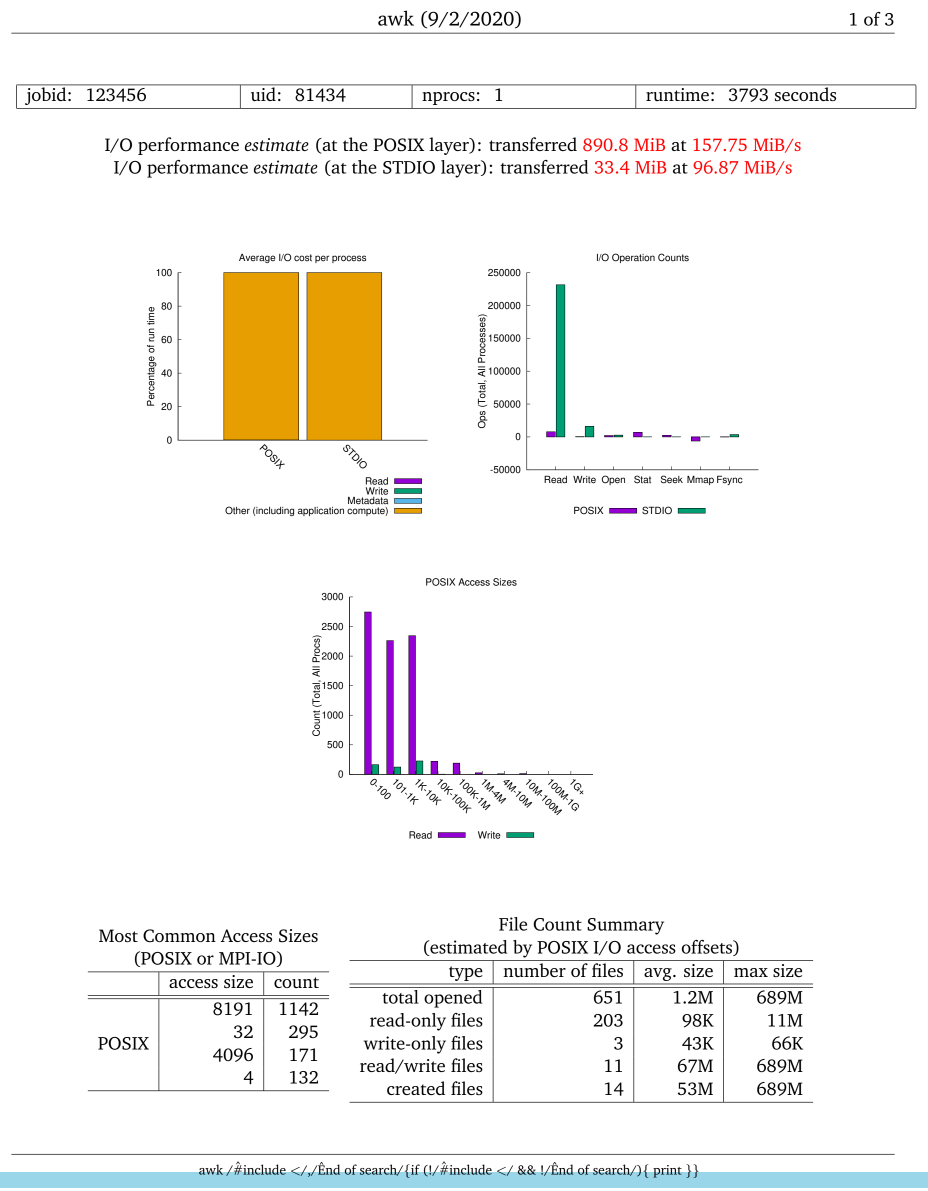
Quick DUNE SW Primer

- DUNE primarily uses LArSoft, shared suite based on Art
  - Reminder: Art was forked from CMSSW many moons ago
- More or less standard simulation workflow chain:
  - Event generation -> Geant4 -> detector sim/noise -> reco
  - Each stage run as separate lar executable with different config file (.fcl files); outputs are inputs to next stage. For this test run all stages in the same "job"
- SW lives in CVMFS, even at NERSC

First pass at Darshan

- Install v3.2.1 at NERSC in DUNE area in non-MPI mode
  - Built w/ gcc 8.2.0 inside usual FNAL SL7 Shifter container
- Make simple bash script to run each of the stages serially; do 5 events only for speed (run on Cori login node inside usual Shifter container)
- Copy Darshan files to laptop, run darshan-merge, then job summary perl script
- This is all VERY preliminary
- Feedback/interpretation help is of course appreciated
- **Q:** *job was running in the /global/cscratch1/sd/dunepro area but I when I ran configure for building Darshan it didn't get compiled with Lustre support. Could we be missing some IO in that case?*

Darshan PDF (should be merged)



awk (9/2/2020) 1 of 3

jobid: 123456	uid: 81434	nprocs: 1	runtime: 3793 seconds

I/O performance *estimate* (at the POSIX layer): transferred 890.8 MiB at 157.75 MiB/s
I/O performance *estimate* (at the STDIO layer): transferred 33.4 MiB at 96.87 MiB/s

Most Common Access Sizes
(POSIX or MPI-IO)

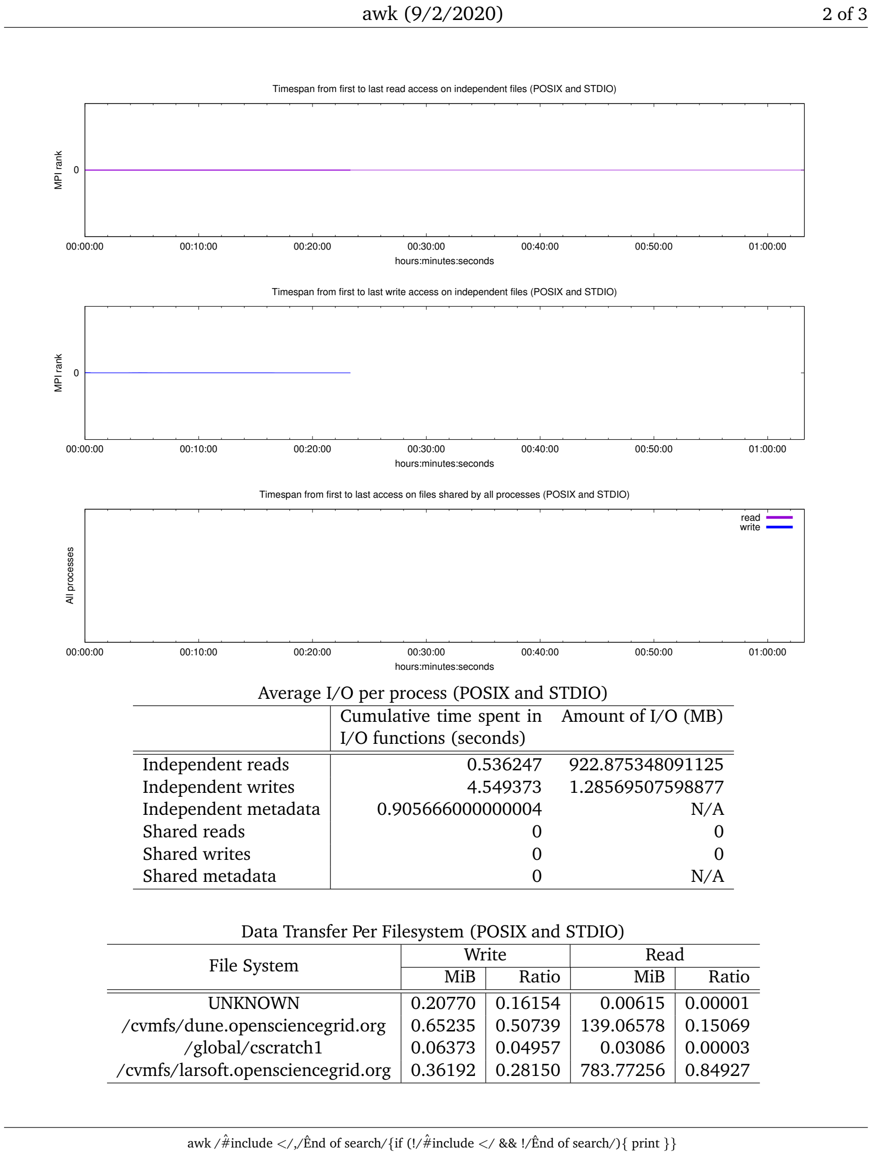
	access size	count
POSIX	8191	1142
	32	295
	4096	171
	4	132

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

type	number of files	avg. size	max size
total opened	651	1.2M	689M
read-only files	203	98K	11M
write-only files	3	43K	66K
read/write files	11	67M	689M
created files	14	53M	689M

awk /#̂include </,/Ênd of search/{if (!/#̂include </ && !/Ênd of search/){ print }}

Darshan PDF (should be merged)



awk (9/2/2020) 2 of 3

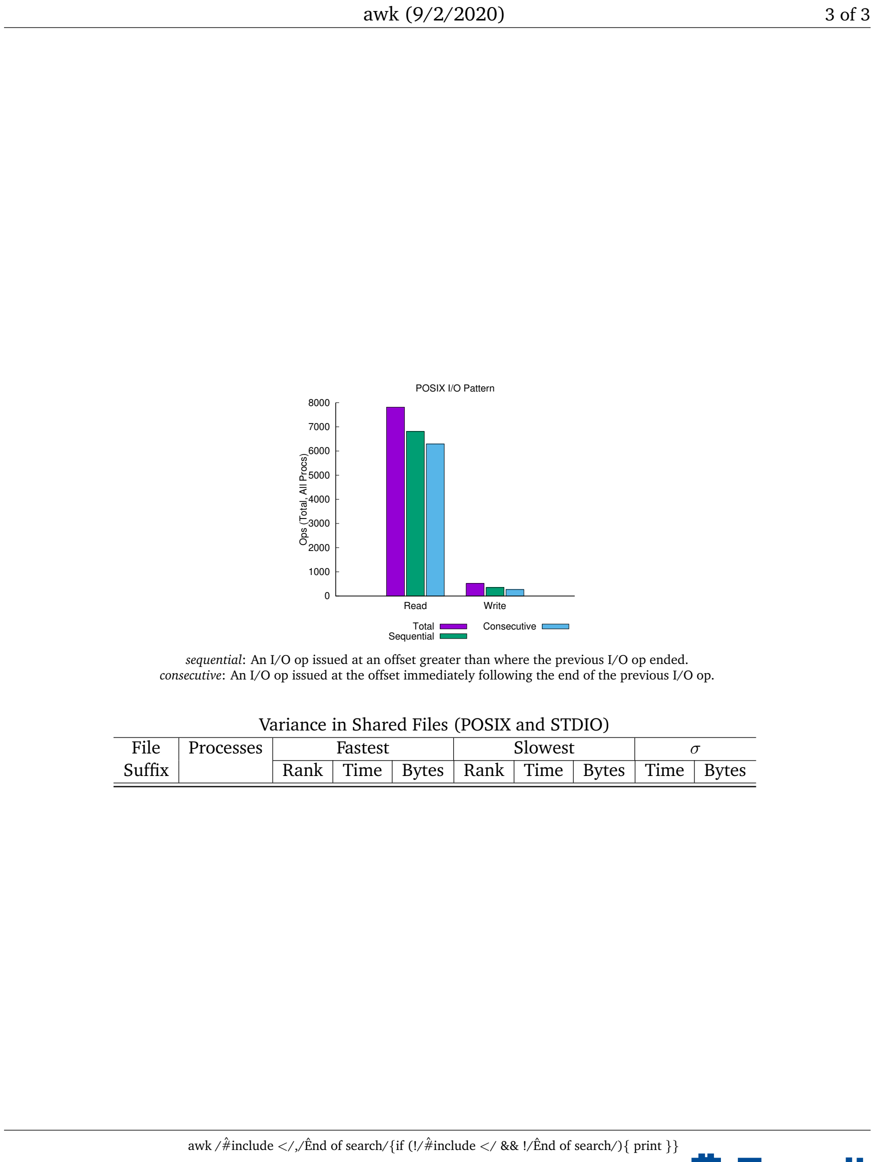
Average I/O per process (POSIX and STDIO)

	Cumulative time spent in I/O functions (seconds)	Amount of I/O (MB)
Independent reads	0.536247	922.875348091125
Independent writes	4.549373	1.28569507598877
Independent metadata	0.905666000000004	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	0.20770	0.16154	0.00615	0.00001
/cvmfs/dune.opensciencegrid.org	0.65235	0.50739	139.06578	0.15069
/global/cscratch1	0.06373	0.04957	0.03086	0.00003
/cvmfs/larsoft.opensciencegrid.org	0.36192	0.28150	783.77256	0.84927

awk /#̂include </,/Ênd of search/{if (!/#̂include </ && !/Ênd of search/){ print }}



awk (9/2/2020) 3 of 3

sequential: An I/O op issued at an offset greater than where the previous I/O op ended.
consecutive: An I/O op issued at the offset immediately following the end of the previous I/O op.

Variance in Shared Files (POSIX and STDIO)

File Suffix	Processes	Fastest			Slowest			σ	
		Rank	Time	Bytes	Rank	Time	Bytes	Time	Bytes

awk /#̂include </,/Ênd of search/{if (!/#̂include </ && !/Ênd of search/){ print }}

Next steps

- Assuming this all looks reasonable,
- Run in a job with a more realistic event count
- Run on a standard worker node as well