Parallel I/O Performance Benchmarking and Investigation on Multiple HPC Architectures

1. Document Information and Version History

Version:	1.0
Status	Release
Author(s):	Bryan Lawrence , Chris Maynard , Andy Turner, Xu Guo, Dominic Sloan-Murphy, Juan Rodriguez Herrera
Reviewer(s)	David Henty

Version	Date	Comments, Changes, Status	Authors, contributors, reviewers
0.1	2016-06-22	Initial draft	Andy Turner
0.2	2016-09-27	Edits to abstract and system introductions	Dominic Sloan-Murphy
0.3	2016-10-31	Additional structure	Dominic Sloan-Murphy
0.4	2017-03-20	Final review before external release	Dominic Sloan-Murphy
0.5	2017-03-24	Reviewed	Andy Turner
1.0	2017-03-28	Updates following review External release	Dominic Sloan-Murphy

2. Abstract

Solving the bottleneck of I/O is a key consideration when optimising application performance, and an essential step in the move towards exascale computing. Users must be informed of the I/O performance of existing HPC resources in order to make best use of the systems and to be able to make decisions about the direction of future software development effort for their application. This paper therefore presents benchmarks for the write capabilities for ARCHER, comparing them with those of the COSMA, UK-RDF DAC, and JASMIN systems, using MPI-IO and, in selected cases, the HDF5 and NetCDF parallel libraries.

We find a reasonable expectation is for approximately 50% of the theoretical system maximum bandwidth to be attainable in practice. Contention is shown to have a dramatic effect on performance. MPI-IO, HDF5 and NetCDF are found to scale similarly but the high-level libraries introduce a small amount of performance overhead.

For the Lustre file system, on a single shared file, maximum performance is found by maximising the stripe count and matching the individual stripe size to the magnitude of I/O operation performed. HDF5 is discovered to scale poorly on Lustre due to an unfavourable interaction with the *H5Fclose()* routine.

3. Introduction

Parallel I/O performance plays a key role in many high performance computing (HPC) applications employed on ARCHER and I/O bottlenecks are an important challenge to understand and eliminate, where possible. It is therefore necessary for users with high I/O requirements to understand the parallel I/O performance of ARCHER, as well as other HPC systems on offer, to be suitably equipped to make informed plans for maximising use of the system and for future software development projects. The results of this work are of particular relevance to ARCHER users currently bottlenecked by I/O performance, but, given the ubiquity of I/O in HPC domains, the findings will be of interest to most researchers and members of the general scientific community. The information here will also be of interest to centres and institutions procuring parallel file systems.

Theoretical performance numbers for parallel file systems are usually easily available but are of limited use as they assume a clean formatted file system with no contention from other users. Obviously, when used in full production, this level of performance will not usually be attained.

The goal of this paper is to provide insight into the performance of parallel file systems in production. To answer questions such as: What is the maximum performance actually experienced? What variation in performance could users experience?

To this end, we detail here the parallel I/O performance of multiple HPC architectures through testing a set of selected I/O benchmarks. Results are presented from the following systems:

- **ARCHER:** the UK national supercomputing service, with a Cray Sonexion Lustre file system.
- **COSMA:** one of the DiRAC UK HPC resources, using a DDN implementation of the IBM GPFS file system.
- **UK-RDF DAC:** the Data Analytic Cluster attached to the UK Research Data Facility, also using DDN GPFS.
- **JASMIN:** a data analysis cluster delivered by the STFC, using the Panasas parallel file system.

We run *benchio*, a parallel benchmarking application which writes a three-dimensional distributed dataset to a single shared file. On all systems, we measure MPI-IO performance and, in selected cases, compare this with HDF5 and NetCDF equivalent implementations.

In the Lustre case, a range of stripe counts and sizes are tested. GPFS file systems do not allow the same level of user configuration so the default configuration as presented to users is employed.

This document is structured as follows: in the subsequent section, we provide detailed specifications on the four chosen benchmark systems and their file systems. We then present our benchio application, highlighting the contrast between its data layout and the layout used by more traditional benchmarks. Results and conclusions follow, and we close by highlighting the opportunities for future work identified during the course of this project.

4. HPC Systems

ARCHER

ARCHER[1] is a Cray XC30-based system and the current UK National Supercomputing Service run by EPCC[2] at the University of Edinburgh[3]. The /work file systems on ARCHER use the Lustre technology in the form of Sonexion parallel file system appliances. The theoretical sustained performance (in terms of bandwidth) of Sonexion Lustre file systems is determined by the number of SSUs (Scalable Storage Units) that make up the file system. ARCHER has three Sonexion file systems available to users:

- fs2: 6 SSU, theoretical sustained = 30 GB/s
- fs3: 6 SSU, theoretical sustained = 30 GB/s
- fs4: 7 SSU, theoretical sustained = 35 GB/s

Each compute node on ARCHER has two Intel Xeon E5-2697 v2 (Ivy Bridge) processors running at 2.7 GHz containing 12 cores each, giving a total of 24 cores per node. Standard compute nodes have 64 GB of memory shared between the two processors. A set of high-memory nodes are offered with 128 GB of available memory but these are not considered in this paper.

Compute nodes are linked via the Cray Aries interconnect[4], a low-latency, high-bandwidth link giving a peak bisection bandwidth of approximately 11,090 GB/s over the entire ARCHER machine. All I/O to the Lustre file systems is routed over the Aries network to dedicated nodes linked to the file systems by Infiniband connections.

COSMA

The Durham-based Cosmology Machine (COSMA)[5] is one of the five systems making up the UK DiRAC facility[6]. Its disks use the IBM General Parallel File System (GPFS) implemented on two DDN SD12K storage controllers. The theoretical maximum performance is 20 GB/s.
Each compute node on COSMA has two 2.6 GHz Intel Xeon E5-2670 CPUs with 8 cores each, i.e. 16 cores per node. 128 GB of RAM is available as standard and the interconnect between node and file system is Mellanox Infiniband FDR10. As for ARCHER, all I/O to the GPFS file system is routed over the Inifiniband compute node network to dedicated nodes linked to the file system by Infiniband connections.

UK-RDF DAC

The UK Research Data Facility (UK-RDF)[7] is a high-volume file storage service collocated with ARCHER. Attached to it is the Data Analytic Cluster (DAC)[8], a system for facilitating the analysis of data held at the RDF. The file system is a DDN GPFS installation and is based on seven DDN 12K couplets. Separate metadata storage is on NetApp EF550/EF540 arrays populated with SSD drives. Three file systems are available to users:

- gpfs1: 6.4 PB storage, mounted as /nerc
- gpfs2: 4.4 PB storage, mounted as /epsrc
- gpfs3: 1.5 PB storage, mounted as /general

The DAC offers two compute node configurations: standard, using two 10-core 2.20 GHz Intel Xeon E5-2660 v2 processors and 128 GB RAM; and high-memory, using four 8-core 2.13 GHz Intel Xeon E7-4830 processors and 2 TB RAM. In this paper, the standard nodes are used exclusively to model the typical use case.

All DAC nodes have direct Infiniband connections to the RDF drives with a maximum theoretical performance of 56 Gbps, or 7 GB/s.

JASMIN

The Joint Analysis System (JASMIN)[9] is an STFC-delivered service providing computing infrastructure for big data analysis.

All tests were run from the Lotus compute cluster on JASMIN on nodes with 2.6 Ghz 8-core Intel Xeon E5-2650 v2 processors and 128 GB memory. The cluster uses the Panasas parallel file system implemented via bladesets connected to compute nodes over a 10 Gbps, i.e. 1.25 GB/s, Ethernet network, the theoretical limit for performance.

5. Parallel I/O benchmark: benchio

The parallel I/O performance of the HPC systems was evaluated by the *benchio* application developed at EPCC. The code is Open Source and is available on GitHub[10]. It was chosen ahead of the popular IOR benchmark for a number of reasons:

- The parallel I/O decomposition can be varied to better model actual user applications.
- The IOR code is very opaque, this makes it very difficult to draw useful conclusions as to what variations in performance are due to.
- benchio is also able to evaluate the performance of HDF5 and NetCDF, two libraries that support parallel I/O and are commonly used by user communities on many HPC services.

Elaborating on the first reason listed, IOR uses an extremely simplistic 1D data decomposition (Figure 1) that does not model user codes and does not test the performance of MPI-IO collective operations that are key to real performance. This is supported by previous work in *Parallel IO Benchmarkin*[11] which found that the optimal MPI-IO write configuration for the IOR layout is to disable collective I/O, a feature essential for achieving speeds beyond that of a few kilobytes-per-second on realistic data layouts.

Figure 1. IOR data layout: simple sequential

The benchio application measures write bandwidth to a single shared file for a given problem size per processor (weak scaling), i.e. the size of the output file scales with the number of processors. We chose to measure write bandwidth as it is the critical consideration of scientific application I/O performance, whereas read performance is traditionally not a factor beyond the initial "one-off" cost of reading input files.

The test data is a series of double precision floating point numbers held in a 3D array and shared over processes in a 3D block decomposition (see Figure 2 and Figure 3). Halos have been added to all dimensions of the local arrays to better approximate the layout of a "real-world" scientific application. By default, each of these local arrays are of size 128^3.

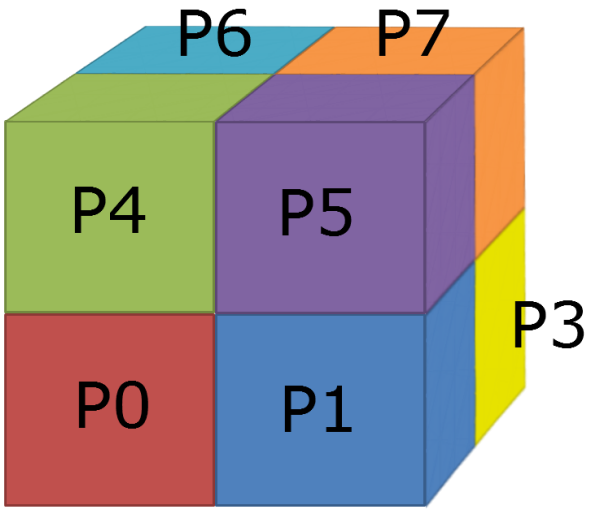

Figure 2. benchio data layout: 3D strided, P2 behind P0

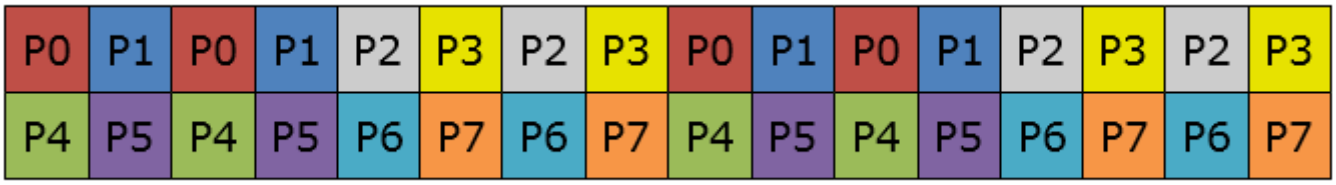

Figure 3. benchio data layout: example 3D decomposition, 2x2x2 grid per processor. Equivalent to layout of output file. Note: data is an entirely contiguous 1x32 array, split into two rows in this figure only for legibility. Contrast with the IOR parallel data layout shown in Figure 1.

6. Results

With benchio, each test is repeated a minimum of ten times and the maximum, minimum and mean bandwidth reported. As I/O is a shared resource on all measured machines, and therefore subject to contention from other users, the maximum attained bandwidth is considered to be most representative of capabilities of a system. In our initial ARCHER results, we present the full range of values to demonstrate the high variance caused by user contention. However, in the results following, we present only the maximum unless otherwise indicated.

ARCHER Performance

benchio was compiled on ARCHER with the following modules loaded:

1) modules/3.2.10.2
2) eswrap/1.3.3-1.020200.1278.0
3) switch/1.0-1.0502.57058.1.58.ari

4) craype-network-aries
5) craype/2.4.2
6) cce/8.4.1
7) cray-libsci/13.2.0
8) udreg/2.3.2-1.0502.9889.2.20.ari
9) ugni/6.0-1.0502.10245.9.9.ari
10) pmi/5.0.7-1.0000.10678.155.25.ari
11) dmapp/7.0.1-1.0502.10246.8.47.ari
12) gni-headers/4.0-1.0502.10317.9.2.ari
13) xpmem/0.1-2.0502.57015.1.15.ari
14) dvs/2.5_0.9.0-1.0502.1958.2.55.ari
15) alps/5.2.3-2.0502.9295.14.14.ari
16) rca/1.0.0-2.0502.57212.2.56.ari
17) atp/1.8.3
18) PrgEnv-cray/5.2.56
19) pbs/12.2.401.141761
20) craype-ivybridge
21) cray-mpich/7.2.6
22) packages-archer
23) bolt/0.6
24) nano/2.2.6
25) leave_time/1.0.0
26) quickstart/1.0
27) ack/2.14
28) xalt/0.6.0
29) epcc-tools/6.0
30) cray-netcdf-hdf5parallel/4.4.0
31) cray-hdf5-parallel/1.8.16

using the Cray Fortran compiler with the default compile flags.

Using the default Lustre settings on ARCHER:

- Stripe size: 1 MiB
- Number of stripes: 4

and running on the fs3 file system, as defined above, we see the performance shown in Figure 4 and listed in Table 1. Recall that each compute node on ARCHER has 24 compute cores and that all cores per node are used when running benchio, giving 24 writers per node.

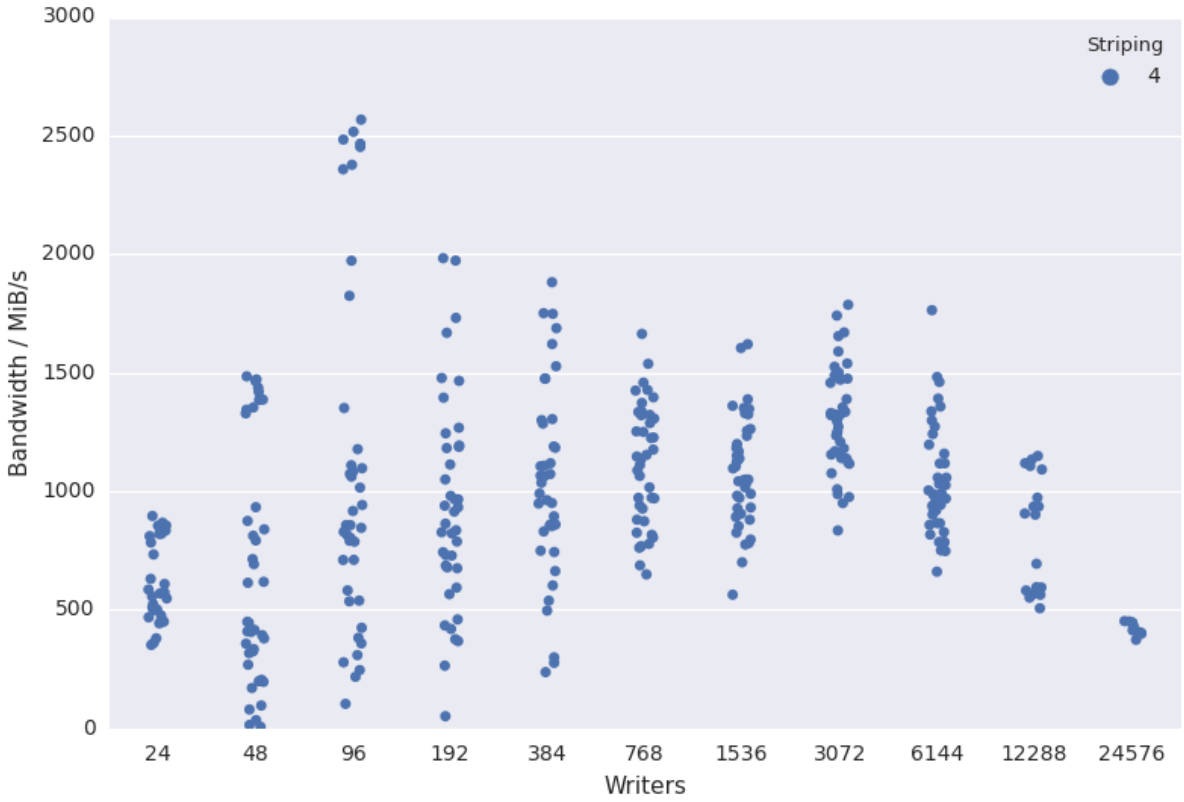

Figure 4. ARCHER MPI-IO default striping (4). A random jitter is applied to the x-axis to better illustrate clusters of similar performance.

		Write Bandwidth (MiB/s)				
Writers	**Total MiB**	**Min.**	**Median**	**Max.**	**Mean**	**Count**
24	384	352	563	896	608	30
48	768	7	448	1485	662	40
96	1536	104	858	2567	1096	40
192	3072	52	889	1983	939	40
384	6144	238	1049	1882	1042	40
768	12288	650	1141	1664	1117	40
1536	24576	564	1049	1620	1081	40
3072	49152	835	1309	1787	1307	40
6144	98304	661	986	1764	1041	40
12288	196608	507	798	1149	803	20
24576	393216	374	423	453	423	10

Table 1. ARCHER MPI-IO default striping (4) raw data.

Using the default stripe settings on ARCHER, the maximum write performance that can be achieved is just over 2,500 MiB/s, just 8.3% of the theoretical sustained performance of 30,000 MiB/s.

In the worst case, 48 writers give a speed of approximately 7 MiB/s, more than a factor of 200 slower than the maximum performance of near 1,500 MiB in that instance. This clearly illustrates the extreme effects file system contention from other users can have on the range of I/O performance.

Lustre Tuning

As described in *Parallel I/O Performance on ARCHER*[12], to get the best parallel write performance for a single-shared file case we must use as many stripes as possible. This is

achieved on Lustre by setting the striping to "-1" which stripes over all available OSTs. We repeated the benchmarks with:

- File system: fs3
- Stripe size: 1 MiB
- Number of stripes: -1 (corresponds to 48 on fs3)

The performance for this configuration is shown in Figure 5 and Table 2.

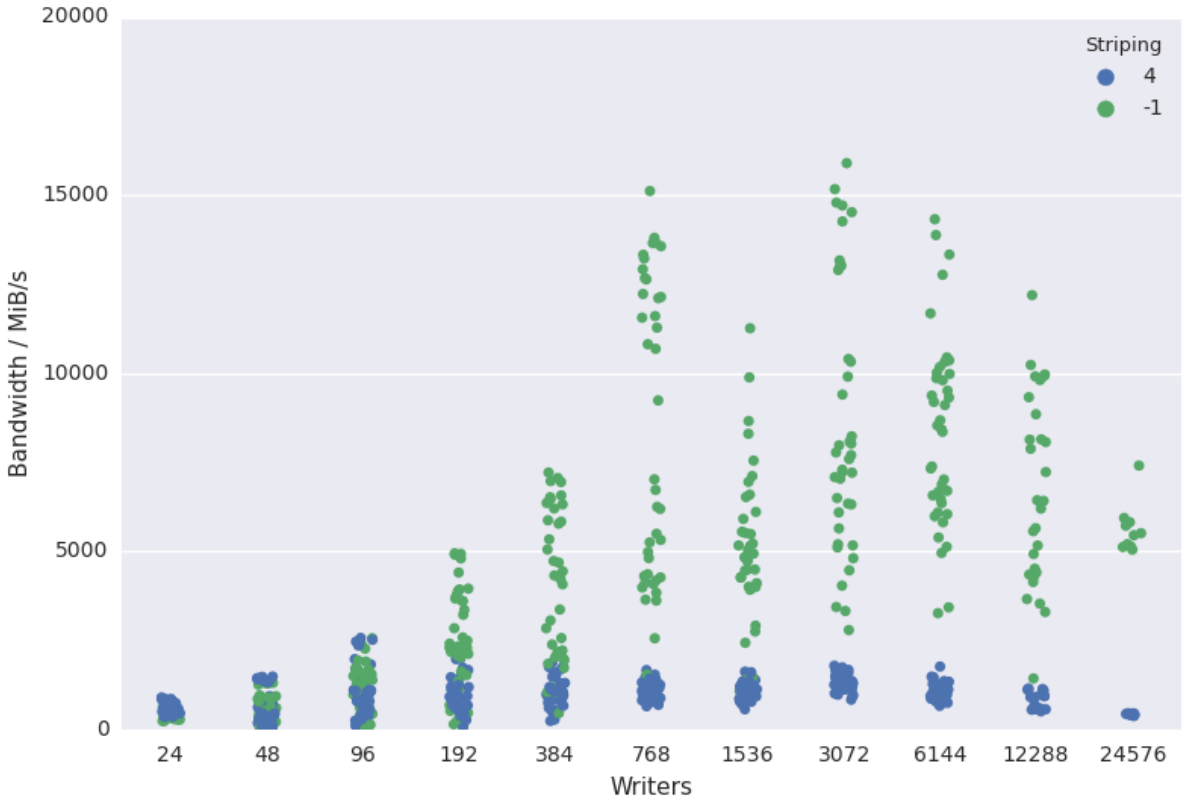

Figure 5. ARCHER MPI-IO maximum striping (-1). Default striping of 4 is plotted for comparison.

		Write Bandwidth (MiB/s)				
Writers	**Total MiB**	**Min.**	**Median**	**Max.**	**Mean**	**Count**
24	384	234	396	616	432	30
48	768	24	581	1356	694	40
96	1536	93	1289	2559	1233	40
192	3072	123	2317	4944	2547	40
384	6144	455	4145	7210	3890	40
768	12288	1541	6872	15116	8318	40
1536	24576	919	4883	11262	5050	40
3072	49152	2789	7645	15898	8547	40
6144	98304	3263	8477	14323	8371	40
12288	196608	1429	6308	12192	6598	30
24576	393216	5046	5480	7407	5634	10

Table 2. ARCHER MPI-IO maximum striping (-1) raw data.

When using the maximum number of stripes, we see much improved performance (compared to the default stripe count of 4) with a maximum write bandwidth of slightly under 16,000 MiB/s with 3072 cores (128 nodes) writing simultaneously. This is a performance of just over 50% of the advertised sustained bandwidth of 30,000 MiB/s for this file system.

The experiments were then repeated, adjusting the size of each Lustre stripe:

- Stripe sizes: 4 MiB and 8 MiB
- Number of stripes: -1 and 4

Maximum measured performance is given in Figure 6 and Figure 7 with the data from the default 1 MiB configuration plotted for comparison. As previously stated, we plot the maximum rather than mean, median or other percentile to account for the high variance in results from contention.

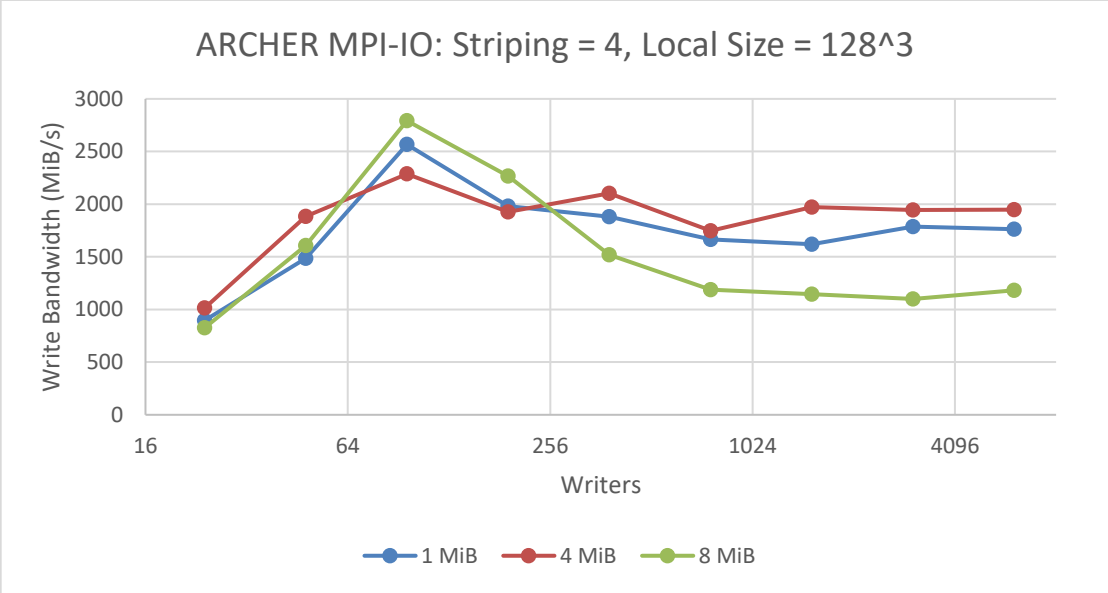

Figure 6. ARCHER stripe size performance, default stripe count

		Max. Write Bandwidth (MiB/s)		
Writers	**Total MiB**	**1 MiB**	**4 MiB**	**8 MiB**
24	384	896.015	1013.819	825.56
48	768	1484.611	1882.74	1606.424
96	1536	2567.143	2287.086	2792.52
192	3072	1982.988	1925.634	2266.698
384	6144	1881.732	2101.862	1520.441
768	12288	1663.967	1747.987	1187.158
1536	24576	1620.391	1971.91	1146.857
3072	49152	1786.612	1944.728	1100.938
6144	98304	1763.888	1947.658	1181.027

Table 3. ARCHER stripe size performance, default stripe count raw data.

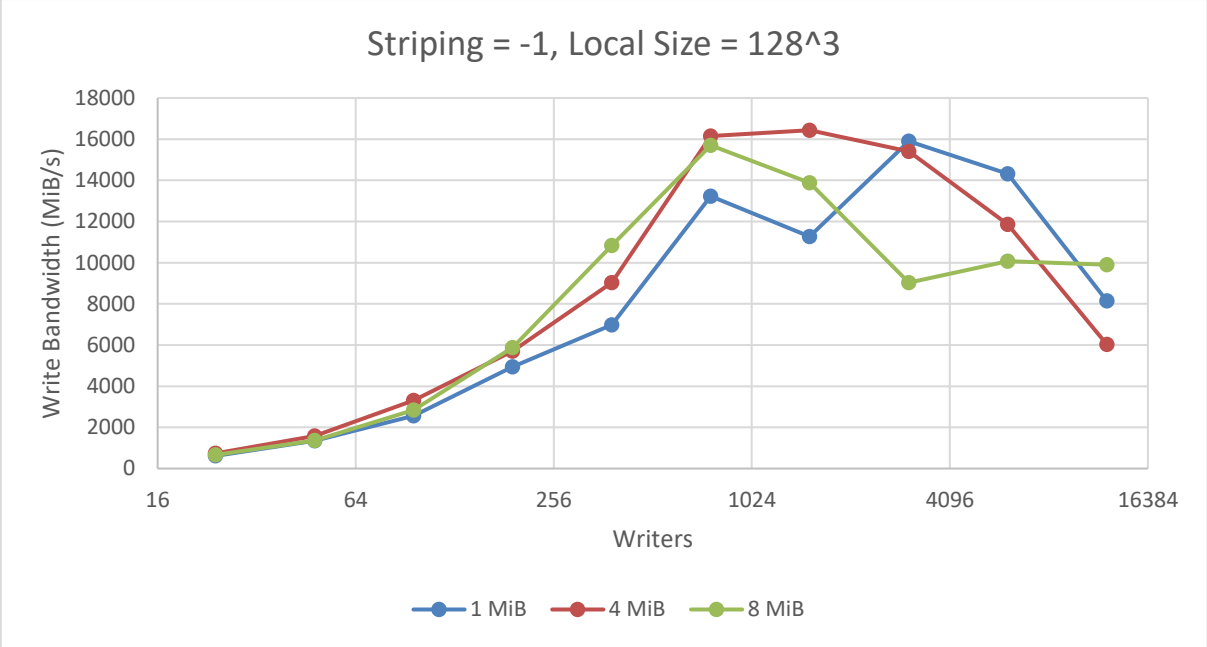

Figure 7. ARCHER stripe size performance, maximum stripe count

	Max. Write Bandwidth (MiB/s)			
Writers	**Total MiB**	**1 MiB**	**4 MiB**	**8 MiB**
24	384	615.717	737.916	660.479
48	768	1355.734	1577.886	1365.949
96	1536	2559.369	3316.318	2840.826
192	3072	4943.626	5707.661	5873.023
384	6144	6971.013	9024.361	10835.89
768	12288	13222.881	16144.447	15697.12
1536	24576	11262.025	16433.642	13874.34
3072	49152	15897.907	15403.649	9037.988
6144	98304	14323.187	11858.55	10073.33
12288	196608	8143.358	6024.108	9907.275

Table 4. ARCHER stripe size performance, maximum stripe count raw data.

Stripe size was found to have a limited effect on the write performance, with the peak for all three sizes being approximately 16,000 MiB/s as before and the measured differences being in-line with the expected variance caused by file system contention. All three settings are shown to be detrimental as core counts increase beyond this performance peak, an effect attributed to increased file locking times and OST contention.

Data Size

All prior experiments were performed with the default local data array of 128^3 double precision values (16 MiB) of data per process. We expected that the benefits of larger stripe sizes would be made apparent with greater volumes of data so repeated the above tests with an increased array size of 256^3 values (128 MiB) per process. Results are given in Figure 8 and Figure 9.

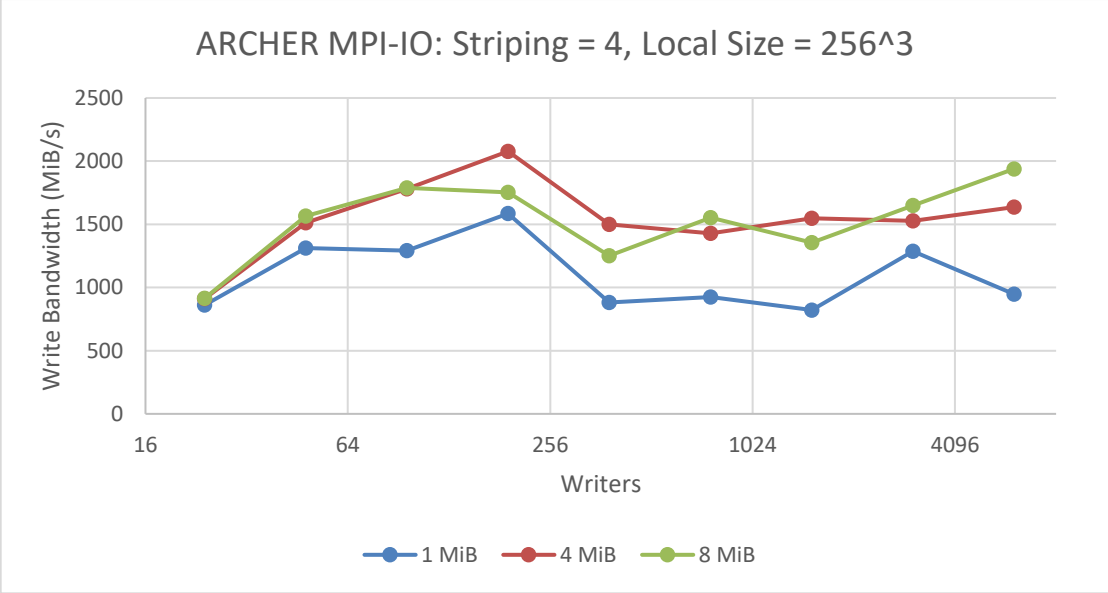

Figure 8. ARCHER large local arrays bandwidth, default stripe count

	Max. Write Bandwidth (MiB/s)			
Writers	Total MiB	1 MiB	4 MiB	8 MiB
24	3072	862.431	911.453	915.131
48	6144	1312.826	1512.007	1565.051
96	12288	1292.413	1781.018	1788.575
192	24576	1584.816	2077.506	1752.687
384	49152	880.738	1499.177	1251.076
768	98304	924.212	1428.858	1553.405
1536	196608	821.884	1548.462	1354.874
3072	393216	1287.72	1527.065	1649.823
6144	786432	946.356	1635.712	1939.126

Table 5. ARCHER large local arrays bandwidth, default stripe count raw data.

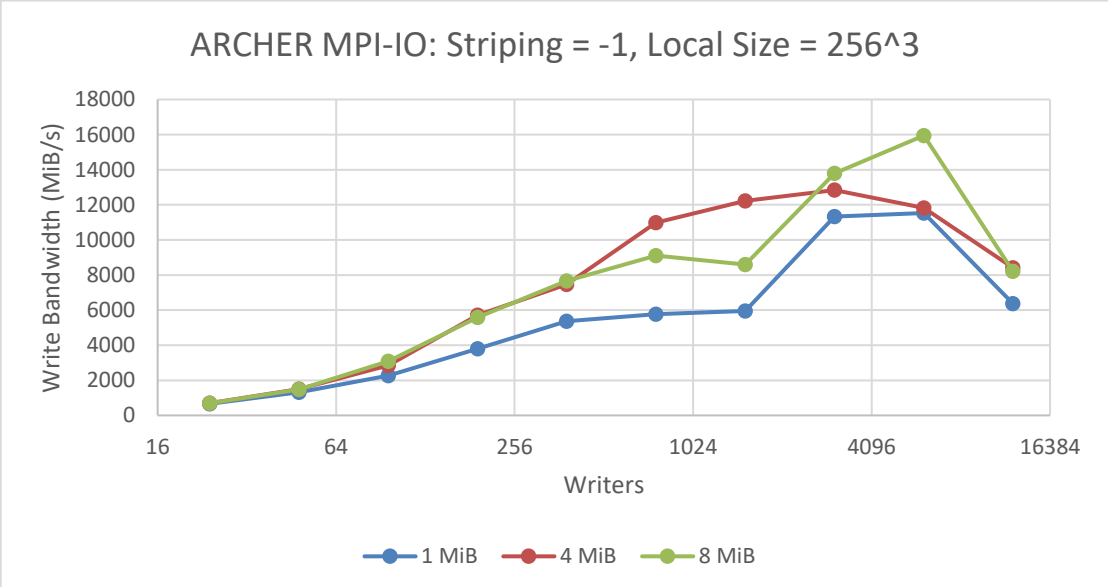

Figure 9. ARCHER large local arrays bandwidth, maximum stripe count

Max. Write Bandwidth (MiB/s)				
Writers	Total MiB	1 MiB	4 MiB	8 MiB
24	3072	661.667	698.198	691.463
48	6144	1310.712	1495.313	1487.371
96	12288	2270.36	2853.872	3084.965
192	24576	3790.924	5716.189	5586.629
384	49152	5359.535	7469.593	7662.899
768	98304	5775.287	10987.14	9107.814
1536	196608	5945.99	12219.962	8598.607
3072	393216	11320.886	12836.212	13784.597
6144	786432	11529.934	11821.306	15946.277
12288	1572864	6367.598	8402.777	8204.55

Table 6. ARCHER large local arrays bandwidth, maximum stripe count raw data.

The larger 4 MiB and 8 MiB stripe sizes give consistently better performance than the default 1 MiB at both 4 and -1 stripe counts. Indeed 8 MiB at 6144 cores is the only configuration to achieve the apparent 16,000 MiB/s limit on ARCHER I/O while the default 1 MiB reaches less than 12,000 MiB/s.

It is apparent that stripe size configuration must be considered in conjunction with I/O operation size to attain maximum performance. In general they must match; lower volume operations should be given smaller stripe sizes, while larger operations require larger stripes.

NetCDF Performance

Optimised installations of NetCDF, backed by parallel HDF5, are provided by Cray as part of the operating system on ARCHER. At time of writing, the default version of this cray-netcdf-hdf5parallel module is 4.3.3.1. However, it was found to give poor performance, failing to demonstrate scalability and instead reaching a peak bandwidth of approximately 1 GiB/s regardless of number of writers or Lustre configuration. We therefore used the more recent NetCDF version 4.4.0 which scales as expected for all benchmarks and recommend to avoid the use of NetCDF versions 4.3.3.1 and below for performance reasons.

Results for version 4.4.0, repeating the stripe and array size experiments performed for MPI-IO, are plotted in Figure 10 to Figure 13.

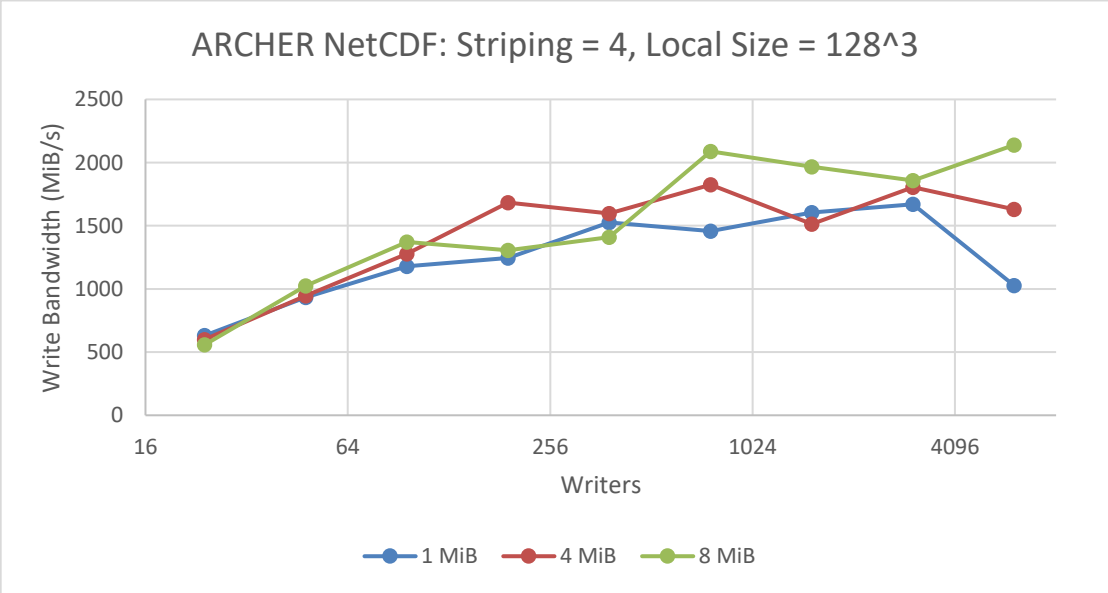

Figure 10. ARCHER NetCDF v4.4.0 performance, default striping, default array sizes

Max. Write Bandwidth (MiB/s)				
Writers	Total MiB	1 MiB	4 MiB	8 MiB
24	384	630.919	600.061	558.083064
48	768	933.118	946.605	1025.44352
96	1536	1177.923	1279.098	1371.81252
192	3072	1244.675	1683.079	1304.60863
384	6144	1527.386	1597.371	1410.57667
768	12288	1458.318	1824.661	2088.70964
1536	24576	1604.539	1512.824	1965.94972
3072	49152	1669.925	1803.806	1858.53392
6144	98304	1026.17	1630.857	2139.2064

Table 7. ARCHER NetCDF v4.4.0 performance, default striping, default array sizes raw data.

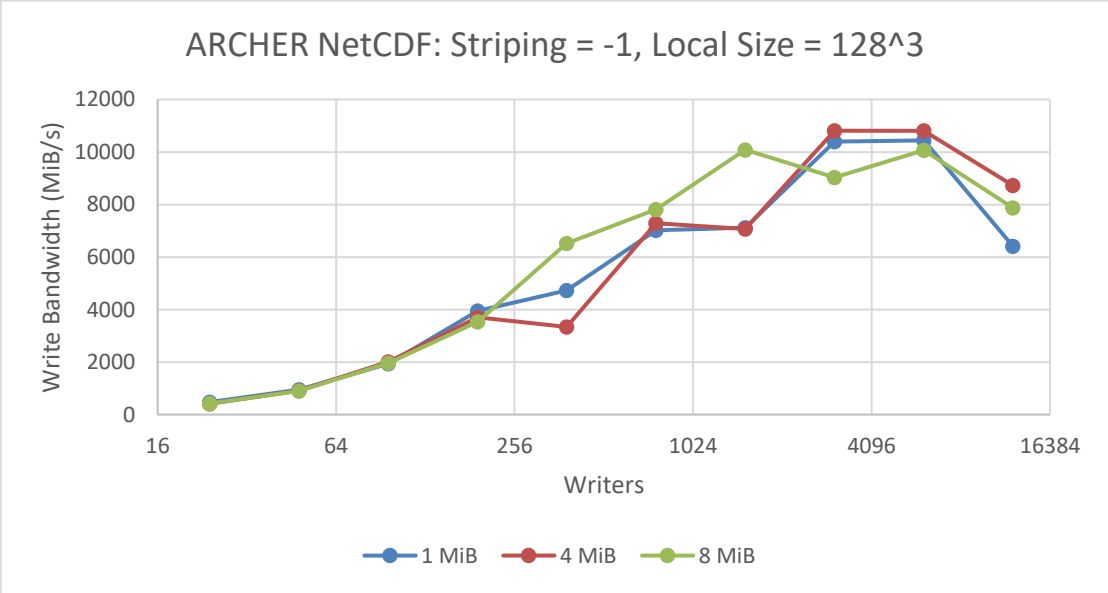

Figure 11. ARCHER NetCDF v4.4.0 performance, maximum striping, default array sizes

Max. Write Bandwidth (MiB/s)				
Writers	Total MiB	1 MiB	4 MiB	8 MiB
24	384	476.357	414.52	425.062
48	768	954.557	911.008	904.397
96	1536	1935.982	2005.05	1957.675
192	3072	3952.425	3710.232	3535.613
384	6144	4728.441	3339.897	6523.897
768	12288	7020.143	7284.373	7810.41
1536	24576	7112.494	7073.241	10085.431
3072	49152	10399.51	10806.77	9037.988
6144	98304	10442.644	10807.797	10073.327
12288	196608	6416.175	8727.265	7877.869

Table 8. ARCHER NetCDF v4.4.0 performance, maximum striping, default array sizes raw data.

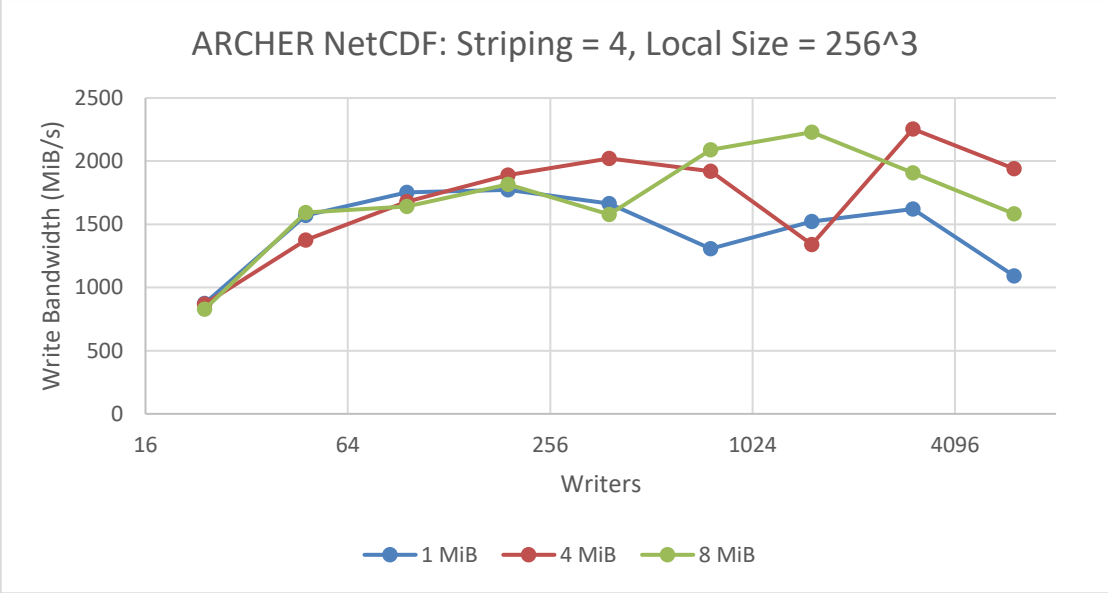

Figure 12. ARCHER NetCDF v4.4.0 performance, default striping, large arrays

		Max. Write Bandwidth (MiB/s)		
Writers	Total MiB	1 MiB	4 MiB	8 MiB
24	3072	874.671	869.748	828.117
48	6144	1569.881	1375.863	1593.592
96	12288	1752.223	1678.23	1640.99
192	24576	1772.707	1889.365	1817.492
384	49152	1664.027	2021.052	1577.356
768	98304	1306.42	1920.333	2088.71
1536	196608	1521.227	1340.975	2229.308
3072	393216	1620.105	2254.418	1906.907
6144	786432	1091.415	1939.353	1583.67

Table 9. ARCHER NetCDF v4.4.0 performance, default striping, large arrays raw data.

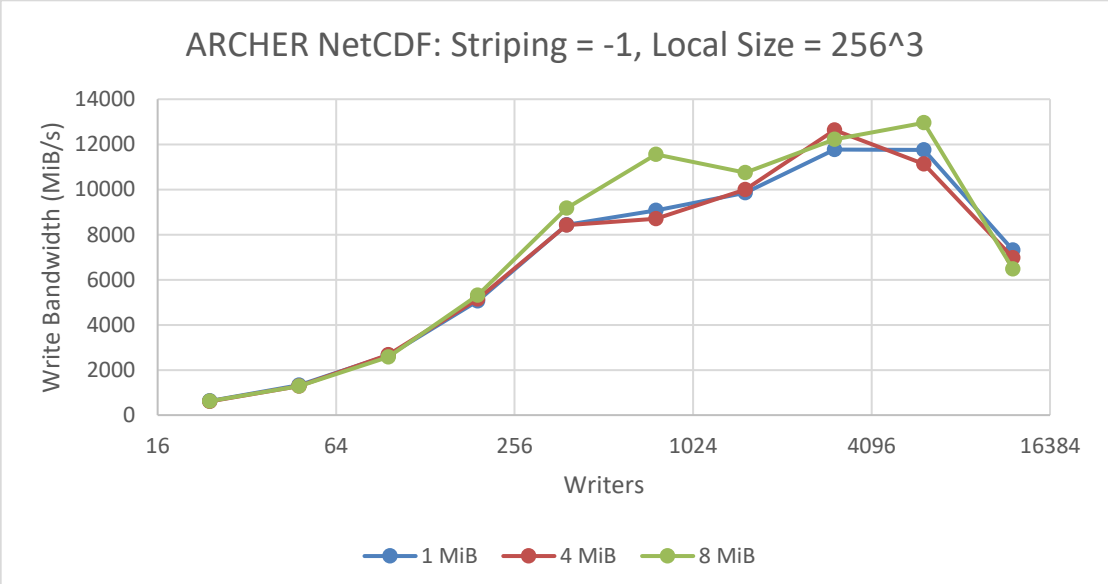

Figure 13. ARCHER NetCDF v4.4.0 performance, maximum striping, large arrays

Max. Write Bandwidth (MiB/s)				
Writers	Total MiB	1 MiB	4 MiB	8 MiB
24	3072	624.57	608.132	619.614
48	6144	1331.474	1279.222	1282.378
96	12288	2652.12	2680.079	2578.753
192	24576	5055.44	5143.392	5318.449
384	49152	8429.176	8416.282	9166.182
768	98304	9069.933	8709.177	11555.443
1536	196608	9858.811	9992.29	10750.391
3072	393216	11769.056	12631.356	12219.471
6144	786432	11746.485	11133.308	12953.971
12288	1572864	7305.898	6978.894	6478.172

Table 10. ARCHER NetCDF v4.4.0 performance, maximum striping, large arrays raw data.

NetCDF performance characteristics were found to be entirely similar to MPI-IO, with variations in stripe count, stripe size and local array size producing the same general trend. This is in line with expectations as NetCDF interfaces to HDF5 for its parallel implementation, which is itself based on MPI-IO.

Peak bandwidth was measured at 13,000 MiB/s, down from the 16,000 MiB/s seen with MPI-IO, i.e NetCDF achieves roughly 80% of MPI-IO performance. This is attributed to the overhead of the NetCDF/HDF5/MPI-IO stack and the additional structuring applied to NetCDF files. To verify this, we examined the write statistics recorded by MPICH, specifically those reported through the *MPICH_MPIIO_STATS* environment variable. Extracts from a simple base case – single writer, maximum striping – are given below:

MPIIO write access patterns for striped/mpiio.dat
independent writes = 0
collective writes = 24

MPIIO write access patterns for striped/hdf5.dat
independent writes = 6
collective writes = 24

MPIIO write access patterns for striped/netcdf.dat
independent writes = 10
collective writes = 24

From this, we can see the actual parallel I/O performed, the collective writes count, is identical between the three libraries, while independent writes increase with the richness of the structural and header information provided. This partially accounts for the lowered performance peak with the remaining deficit being additional time spent in library-specific functions. This last point is of particular relevance in the case of HDF5 on ARCHER, detailed in the following section.

HDF5 Performance

As with NetCDF, Cray provides several pre-installed versions of the HDF5 parallel library on ARCHER. For these library versions (from the default 1.8.14 to the most current 1.10.0), similar performance limitations as for NetCDF 4.3.3.1 were observed. Given the hierarchical nature of the libraries, we theorised that the NetCDF 4.3.3.1 limitations were in reality a manifestation of a bug in the HDF5 layer, and that NetCDF 4.4.0 circumvented the issue by following an alternate code path around the problematic library calls.

Application profiling of benchio with the HDF5 backend, to verify this theory, found the majority of compute time is spent in function *MPI_File_set_size()*, called within the HDF5 library from the user-level *H5Fclose()* routine. Discussions with Cray revealed this to indeed be a known bug specific to the combination of HDF5 with Lustre file systems.

An *MPI_File_set_size()* operation, on a Linux platform like ARCHER, eventually calls the POSIX function: *ftruncate()*. This has an unfavourable interaction with the locking for the series of metadata communications the HDF5 library makes during a file close. In practice, this leads to relatively long close times of tens of seconds and hence the lack of scalability observed.

The HDF5 developers have noted this behaviour in the past where it manifested in *H5Fflush()*, the function for flushing write buffers associated with a file to disk: "when operating in a parallel application, this operation resulted in a call to *MPI_File_set_size*, which currently has very poor performance characteristics on Lustre file systems. Because an HDF5 file's size is not required to be accurately set until the file is closed, this operation was removed from *H5Fflush* and added to the code for closing a file"[13] hence leading to the behaviour currently observed in *H5Fclose()*.

Cray's investigations on this bug are on-going and, at present, no known work-around or mitigation is provided for end users. The recommendation for users is to be aware of this interaction and inform research communities as the issue is observed.

Impact of System Load

To better understand the impact of file system contention, we simulated different degrees of load by running multiple instances of the benchio MPI-IO test in parallel. Figure 14 shows the aggregate mean performance of one, two and four benchio instances writing concurrently to independent files with the default stripe size (1 MiB).

Note that here we use aggregate mean performance, rather than maximum performance, as, in the given setup, often a single benchio instance would be performing I/O while the other instances were preparing to start, had already finished or were otherwise between iterations. The maximum bandwidth achieved during such a test is essentially the same as the maximum bandwidth when running just a single benchio instance and is therefore not representative of the impact of system load.

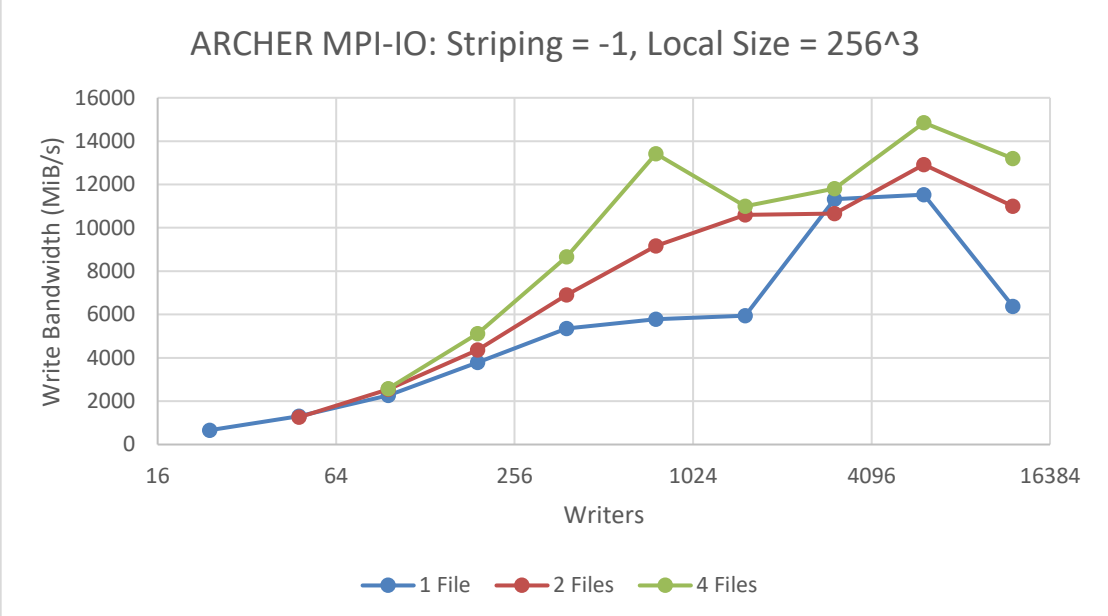

Figure 14. Effect of I/O load on ARCHER

Writers	Total MiB	Average Write Bandwidth (MiB/s) 1 File	2 Files Instance 1	2 Files Instance 2	2 Files Aggregate
24	3072	661.667			
48	6144	1310.712	630.159	627.403	1257.562

96	12288	2270.36	1269.261	1279.809	2549.07
192	24576	3790.924	2170.566	2190.197	4360.763
384	49152	5359.535	3459.273	3441.909	6901.182
768	98304	5775.287	4503.837	4657.415	9161.252
1536	196608	5945.99	5589.881	5009.902	10599.783
3072	393216	11320.886	5373.545	5278.291	10651.836
6144	786432	11529.934	6250.276	6669.603	12919.879
12288	1572864	6367.598	5901.083	5092.51	10993.593

Table 11. Effect of I/O load on ARCHER. 1 and 2 files.

		Average Write Bandwidth (MiB/s)				
Writers	**Total MiB**	**4 Files Instance 1**	**4 Files Instance 2**	**4 Files Instance 3**	**4 Files Instance 4**	**4 Files Aggregate**
96	12288	627.204	664.347	644.53	648.275	2584.356
192	24576	1261.765	1278.588	1278.203	1292.701	5111.257
384	49152	2176.97	2188.784	2126.32	2167.82	8659.894
768	98304	3384.657	3306.514	3360.021	3366.43	13417.622
1536	196608	2786.927	2746.021	2723.053	2745.284	11001.285
3072	393216	2944.726	2954.255	2954.338	2958.292	11811.611
6144	786432	4290.212	3533.092	3513.919	3511.966	14849.189
12288	1572864	2962.049	3469.729	3395.223	3373.497	13200.498

Table 12. Effect of I/O load on ARCHER. 4 files.

At core counts below 96, the data trends are reasonably similar and we see that bandwidth is on average divided equally between writers. E.g. the aggregate bandwidth of two benchio instances, each with 24 writers putting data to independent files, is roughly equivalent to the bandwidth of a single instance with 48 writers. However, as number of writers increase, there is a definite trend that multiple files give better performance than a single file. This is particularly apparent in the 768 writers case where a single file sees approximately 5800 MiB/s while four files achieves near 1000 MiB/s, more than a factor of two difference. In further work, investigations into using varying numbers of files, from the current findings on a single shared file to the extreme case of a single file per process, could be done to further explore the results seen here.

COSMA Performance

The GPFS file system employed by the DiRAC COSMA service does not facilitate user tuning like Lustre. GPFS settings are fixed at installation and cannot be adjusted at run time. We therefore ran a single set of benchmarks to determine the peak bandwidth of the system, presented in Figure 15. NetCDF and HDF5 results were not gathered in this case, due to time constraints. We will investigate the performance of HDF5 and NetCDF on GPFS in a future update to this work but expect to see similar trends to that seen for ARCHER (although HDF5 performance may be improved on GPFS over Lustre because of the particular issues with Lustre described above).

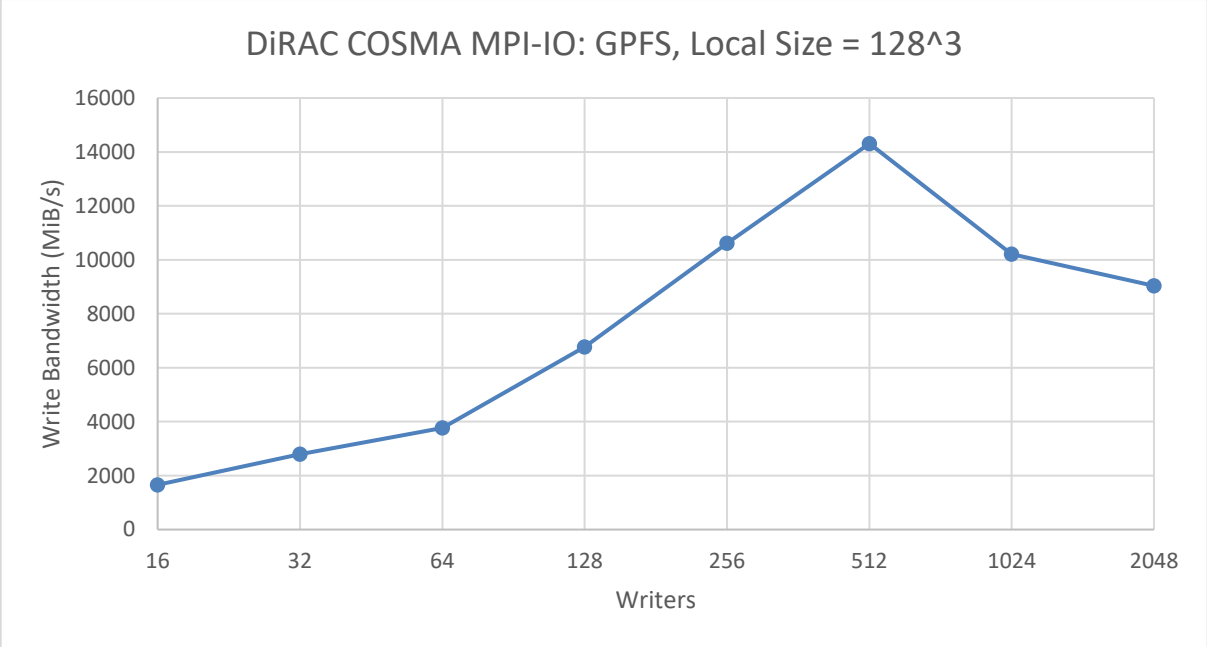

Figure 15. MPI-IO bandwidth for DiRAC COSMA

Max. Write Bandwidth (MiB/s)		
Writers	**Total MiB**	**Bandwidth**
16	384	1658.71634
32	768	2796.63964
64	1536	3771.1242
128	3072	6771.9717
256	6144	10619.3772
512	12288	14308.9679
1024	24576	10214.0822
2048	49152	9039.30916

Table 13. MPI-IO bandwidth for DiRAC COSMA raw data.

Best performance is seen at 512 writers, which attain marginally more than 14000 MiB/s or approximately 68% of the rated maximum, before parallel efficiency drops. As with ARCHER, this is attributed to file and disk contention.

UK-RDF DAC Performance

The UK-RDF DAC supports only on-node parallelism; jobs cannot span multiple nodes. All tests were therefore run on a single, standard compute node offering 40 CPU cores.

We benchmarked two of the three GPFS file systems and examined the performance of each of the benchio parallel backends. Comparisons are given in Figure 16 and Figure 17.

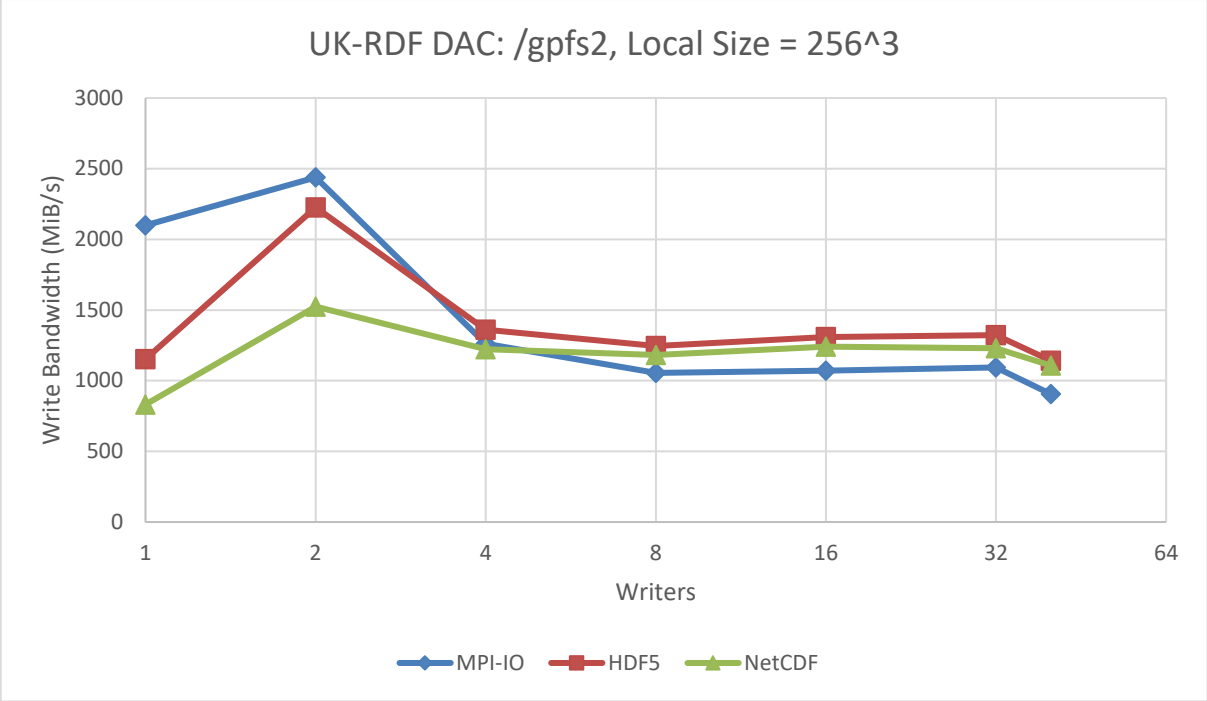

Figure 16. All backends bandwidth for UK-RDF DAC. File system: 4.4PB /gpfs2 mounted as /epsrc.

	Max. Write Bandwidth (MiB/s)			
Writers	**Total MiB**	**MPI-IO**	**HDF5**	**NetCDF**
1	3072	2098.365	1153.154	831.168
2	6144	2438.094	2226.086	1523.809
4	12288	1261.083	1361.701	1221.957
8	24576	1055.67	1245.742	1181.084
16	49152	1070.569	1307.79	1240.46
32	98304	1094.017	1321.716	1228.554
40	196608	905.553	1142.092	1106.309

Table 14. All backends bandwidth for UK-RDF DAC raw data. File system: 4.4PB /gpfs2

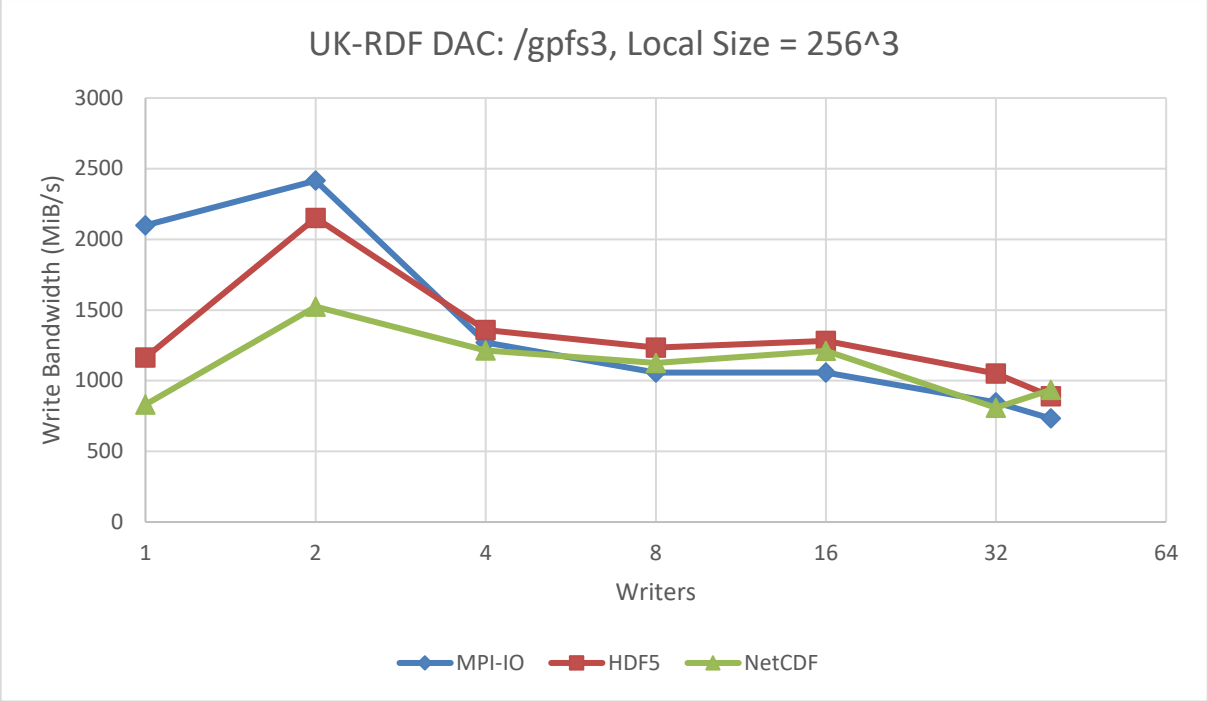

Figure 17. All backends bandwidth for UK-RDF DAC. File system: 1.5 PB /gpfs3 mounted as /general.

		Max. Write Bandwidth (MiB/s)		
Writers	**Total MiB**	**MPI-IO**	**HDF5**	**NetCDF**
1	3072	2098.357	1163.637	831.1686
2	6144	2415.096	2151.261	1523.81
4	12288	1270.471	1358.09	1213.27
8	24576	1056.759	1233.735	1125.275
16	49152	1057.305	1282.404	1210.402
32	98304	845.9315	1051.335	808.5274
40	196608	732.475	889.3521	936.0146

Table 15. All backends bandwidth for UK-RDF DAC raw data. File system: 1.5 PB /gpfs3

No difference in performance was measured between the /gpfs2 and /gpfs3 file systems. Both achieved the same peak performance of approximately 2500 MiB/s, or approximately 35% of the theoretical maximum of 7000 MiB/s. Hence file system storage capacity was found to have no bearing on overall write speed in this instance, contrary to the case of Sonexion Lustre (see the *HPC Systems* section above for an illustration of how additional storage hardware/SSUs influence the maximum potential performance of the fs4 Lustre file system on ARCHER, in comparison to fs2 and fs3).

MPI-IO, HDF5 and NetCDF displayed identical scaling characteristics with their peak bandwidths reflecting the arrangement of their hierarchy. HDF5 reached 2200 MiB/s while NetCDF performed at 1500 MiB/s, or 88% and 60% of MPI-IO respectively.

Scope for parallelisation is limited on this system with performance dropping significantly at 4 writers and above. Previous work in *Investigating Read Performance of Python and NetCDF when using HPC Parallel Filesystems*[14] on the RDF DAC supports these findings, showing sequential serial read performance to peak at roughly 1400 MiB/s, i.e. the same performance level seen from 4 to 40 writers in Figure 16 and Figure 17. Further work is needed to precisely identify the bottleneck limiting the scalability on this system.

JASMIN Performance

As with the RDF DAC, JASMIN is intended for analysis of large volumes of data. However, in contrast to the DAC, jobs can be run across multiple nodes in the cluster, potentially increasing the ceiling for parallelisation. Results were gathered from 1 to 32 writers and are presented in Figure 18.

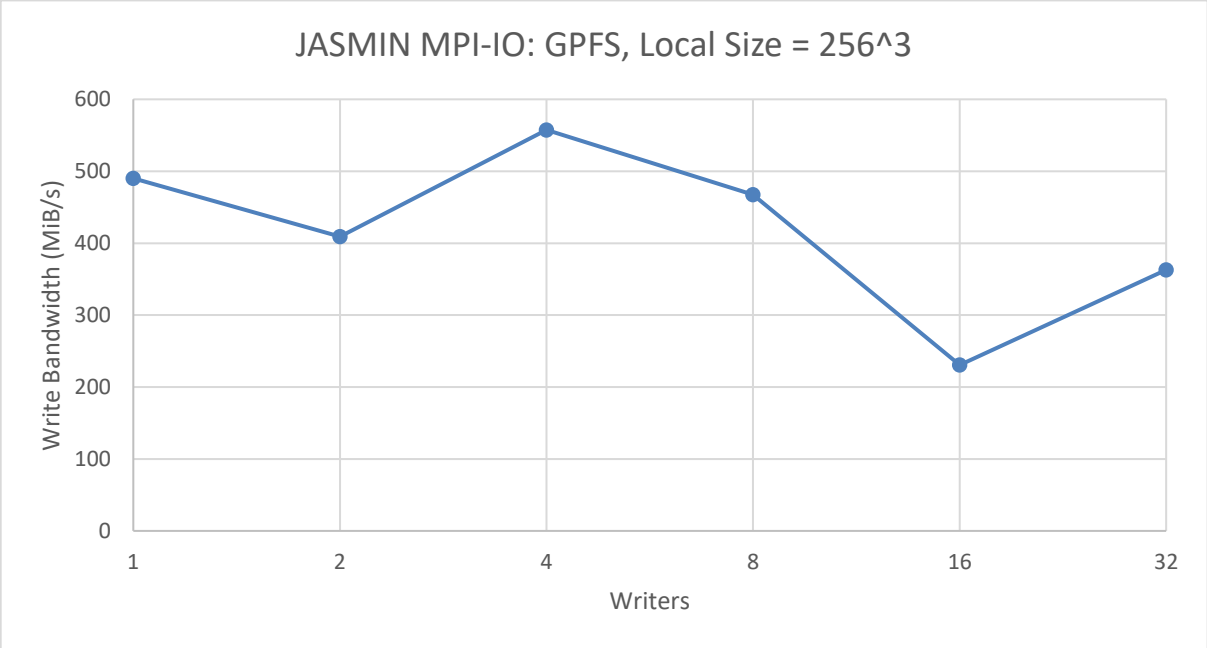

Figure 18. MPI-IO bandwidth for JASMIN

Max. Write Bandwidth (MiB/s)		
Writers	Total MiB	Bandwidth
1	128	490.264
2	256	409.038
4	512	557.479
8	1024	467.489
16	2048	230.663
32	4096	362.702

Table 16. MPI-IO bandwidth for JASMIN raw data.

With further reference to *Investigating Read Performance of Python and NetCDF when using HPC Parallel Filesystems*[14], sequential serial performance on JASMIN has been measured at approximately 500 MiB/s, the same level of performance observed in these parallel I/O tests. From this, we conclude that there is no scope for improvement with parallelisation on this system under the default configuration. However, at time of writing, additional work is underway from Jones *et al.* to expand their investigation to include multi-threaded performance and examine parallelism on JASMIN in greater detail. Results are expected to be published at a later date.

Comparative System Performance

Figure 19 gives an overview of all four benchmark systems and compares their overall performance.

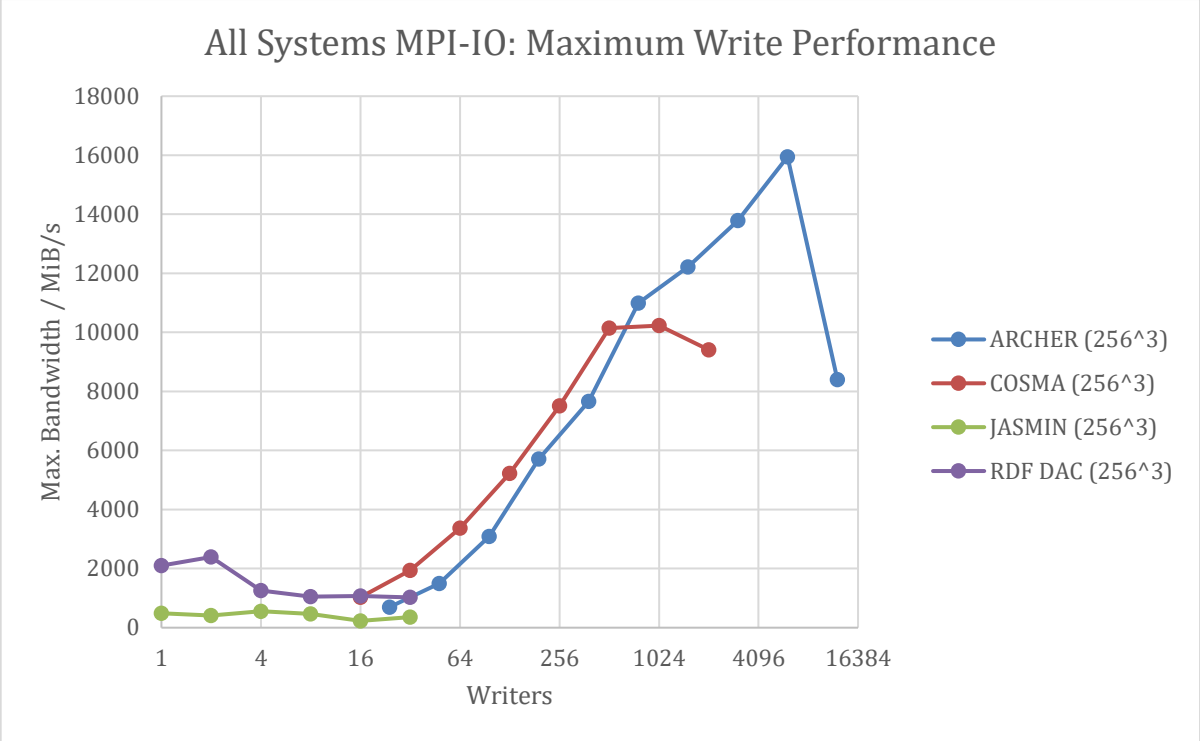

Figure 19. Comparison of maximum write performance between benchmark systems

	Max. Write Bandwidth (MiB/s)			
Writers	**ARCHER**	**COSMA**	**JASMIN**	**RDF-DAC**
1			490.264	2098.365
2			409.038	2392.526
4			557.479	1261.083
8			467.489	1050.257
16		1031.078	230.663	1070.570
24	698.198			
32		1940.589	362.702	1025.025
48	1495.313			
64		3367.448		
96	3084.965			
128		5225.449		
192	5716.189			
256		7514.007		
384	7662.899			
512		10145.69		
768	10987.14			
1024		10229.963		
1536	12219.962			
2048		9405.305		
3072	13784.597			
6144	15946.277			
12288	8402.777			

Table 17. Comparison of maximum write performance between benchmark systems raw data

The two systems intended for high-performance parallel simulations, ARCHER and COSMA, are broadly comparable, as are the two data analysis systems. The scope for parallelism is simply lower on JASMIN and the RDF DAC and users should not expect compute and analysis platforms to have similar performance.

7. Conclusions

Our findings for write performance can be summarised as follows: approximately 50% of the theoretical maximum write performance on a system should be expected to be attainable in production, with dramatic variance due to user contention – a factor of 200 difference in the worst case. We additionally verified that systems designed for parallel simulations offer much higher performance than data analysis platforms.

The three parallel libraries, MPI-IO, HDF5 and NetCDF, share the same performance characteristics but the higher level APIs introduce additional overhead. A reasonable expectation is 10% and 30% overhead for HDF5 and NetCDF respectively.

Tests on Lustre file systems found the optimal configuration for a single shared output file was to use maximum striping and ensure I/O operation and stripe sizes are in accordance. Generally the larger the amount of data written per writer, the larger the stripe size that should be used. Considering peak performances, improvements of approximately 10% and 35% were seen when using 4 MiB and 8 MiB stripe sizes rather than the default 1 MiB, when using large enough data sets (i.e. 256^3 array elements, or 128 MiB per writer).

Further relating to Lustre systems, users should be aware of the HDF5 performance issue and should note that versions of NetCDF below 4.4.0 should be avoided on Cray Systems as they are affected by this issue.

Finally, in contrast to Lustre, we found GPFS file system capacity to have no bearing on overall parallel I/O performance.

8. Future Work

Various opportunities for further investigation were identified during the production of this white paper. In particular, benchio could be extended to support the file-per-process I/O pattern, to complement the current work done on the single-shared-file strategy and follow-up on the bandwidth improvements in the load test shown in Figure 14. Additionally, write performance has been the exclusive focus of this work due to its relative importance in typical HPC workflows but there is scope for considering the equivalent read performance.

These topics are currently being investigated by the authors and will be included in a forthcoming update of this paper.

References

[1] ARCHER HPC Resource, http://www.archer.ac.uk/, retrieved 28 Nov 2016
[2] EPCC at The University of Edinburgh | EPCC, https://www.epcc.ed.ac.uk/, retrieved 28 Nov 2016
[3] The University of Edinburgh, http://www.ed.ac.uk/, retrieved 28 Nov 2016
[4] Performance Computer, XC Series Supercomputers - Technology | Cray, http://www.cray.com/products/computing/xc-series?tab=technology, retrieved 28 Nov 2016
[5] Institute for Computational Cosmology Durham University - PhD and postgraduate research in astronomy, astrophysics and cosmology, http://icc.dur.ac.uk/index.php?content=Computing/Cosma, retrieved 28 Nov 2016
[6] DiRAC Distributed Research utilising Advanced Computing, https://www.dirac.ac.uk/, retrieved 28 Nov 2016
[7] RDF » UK Research Data Facility (UK-RDF), http://www.rdf.ac.uk/, retrieved 28 Nov 2016
[8] ARCHER » 5. UK-RDF Data Analytic Cluster (DAC), http://www.archer.ac.uk/documentation/rdf-guide/cluster.php, retrieved 28 Nov 2016
[9] home | JASMIN, http://www.jasmin.ac.uk/, retrieved 28 Nov 2016

[10] EPCCed/benchio: EPCC I/O benchmarking applications, https://github.com/EPCCed/benchio, retrieved 01 Nov 2016
[11] Jia-Ying Wu, Parallel IO Benchmarking, https://static.ph.ed.ac.uk/dissertations/hpc-msc/2015-2016/Jia-ying_Wu-MSc-dissertation-Parallel_IO_Benchmarking.pdf, retrieved 22 Nov 2016
[12] David Henty, Adrian Jackson, Charles Moulinec, Vendel Szeremi: Performance of Parallel IO on ARCHER Version 1.1, http://www.archer.ac.uk/documentation/white-papers/parallelIO/ARCHER_wp_parallelIO.pdf, retrieved 01 Nov 2016
[13] Mark Howison, Quincey Koziol, David Knaak, John Mainzer, John Shalf: Tuning HDF5 for Lustre File Systems, https://support.hdfgroup.org/pubs/papers/howison_hdf5_lustre_iasds2010.pdf , retrieved 03 Nov 2016
[14] Matthew Jones, Jon Blower, Bryan Lawrence, Annette Osprey: Investigating Read Performance of Python and NetCDF When Using HPC Parallel Filesystems, http://link.springer.com/chapter/10.1007%2F978-3-319-46079-6_12, retrieved 24 Nov 2016

Acknowledgements

The authors would like to thank Harvey Richardson of Cray Inc. for his invaluable advice on the ARCHER file systems and software. We would also like to thank the DiRAC and JASMIN facilities for providing time on their systems to run the benchmarks.