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

WPxxx: *(Will be assigned by PMO)*
Authors: B. Lawrence (University of Reading), C. Maynard (Met Office), A. Turner (EPCC), X. Guo (EPCC), D. Sloan-Murphy (EPCC)
HPC Tool/Technique: MPI-IO, HDF5, NetCDF, Lustre, GPFS, Panasas
Application: High performance parallel input/output
Exascale, CoEs and Scientific communities: ESiWACE, all communities with significant I/O requirements
Person Months:

Parallel I/O performance plays a key role in many high-performance computing (HPC) applications. I/O bottlenecks are an important challenge to understand and, where possible, eliminate on both current, petascale resources and looking forward to exascale computing[1]. It is therefore necessary for research communities with high I/O requirements to understand the parallel I/O performance of existing HPC systems and applications to be suitably equipped to make informed plans for future procurements and software development projects. The results of this work are of particular relevance to the ESiWACE Centre of Excellence (CoE), as the originators of this project, but, given the ubiquity of I/O in HPC domains, the findings will be of interest to most researchers and members of the European scientific computing community.

This paper presents benchmarks for the write capabilities of the following HPC systems:

- **ARCHER**: the UK national supercomputing service, with a Cray Sonexion Lustre file system[2].

- **COSMA**: one of the DiRAC UK HPC resources, using a DDN implementation of the IBM GPFS file system[3].

- **UK-RDF DAC**: the Data Analytic Cluster attached to the UK Research Data Facility, also using DDN GPFS[4].

- **JASMIN**: a data analysis cluster delivered by the STFC, using the Panasas parallel file system[5].

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 select cases, compare this with HDF5 and NetCDF equivalent implementations.

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.

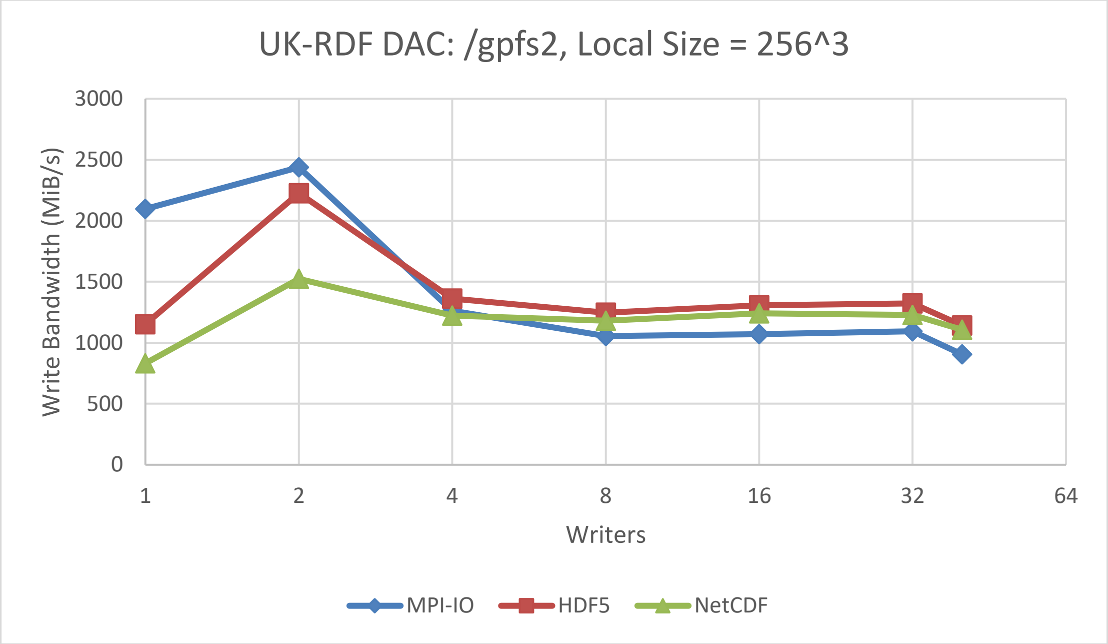

Figure 16. All backends bandwidth for UK-RDF DAC. File system: 4.4PB /gpfs2 mounted as /epsrc. MPI-IO, HDF5 and NetCDF display identical scaling characteristics with their peak bandwidths at 2 writers reflecting the arrangement of their hierarchy

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.

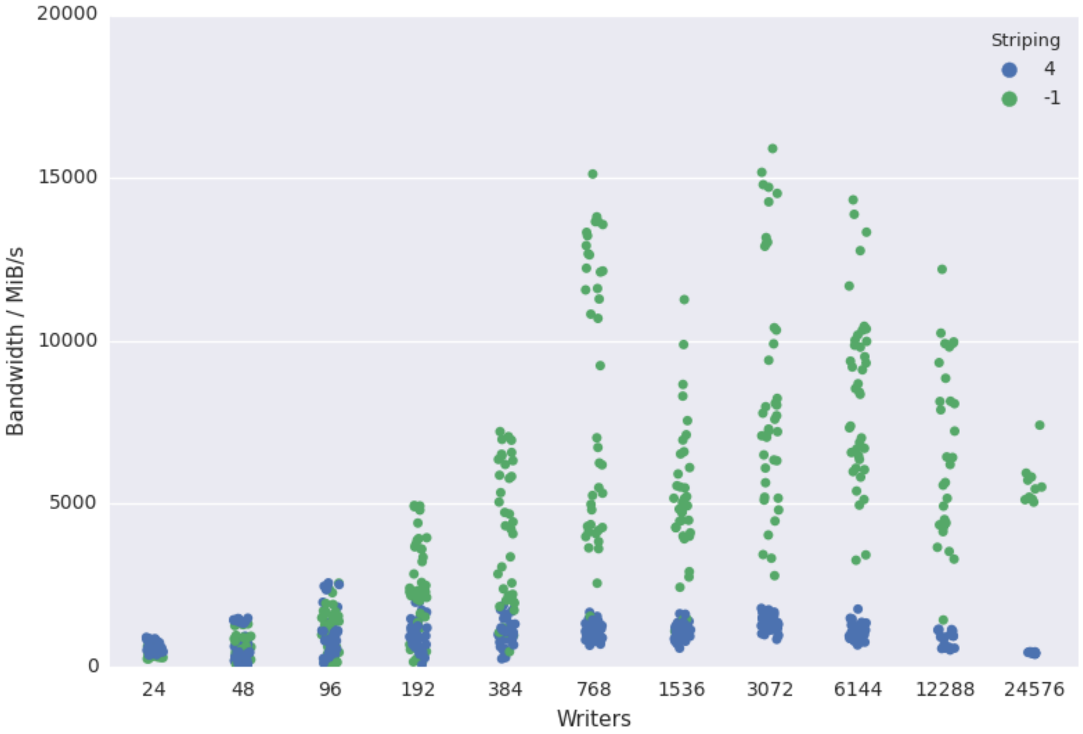

Figure 5. Results spread for ARCHER (Lustre) MPI-IO maximum striping (-1). Default striping of 4 is plotted for comparison.

References

[1] NEXTGenIO | Next Generation I/O for the Exascale, http://www.nextgenio.eu/, retrieved 01 Dec 2016

[2] ARCHER HPC Resource, http://www.archer.ac.uk/, retrieved 28 Nov 2016

[3] DiRAC Distributed Research utilising Advanced Computing, https://www.dirac.ac.uk/, retrieved 28 Nov 2016

[4] ARCHER » 5. UK-RDF Data Analytic Cluster (DAC), http://www.archer.ac.uk/documentation/rdf-guide/cluster.php, retrieved 28 Nov 2016

[5] home | JASMIN, http://www.jasmin.ac.uk/, retrieved 28 Nov 2016