Optimising I/O using non-volatile memory

Adrian Jackson
EPCC, The University of Edinburgh, Edinburgh, United Kingdom, a.jackson@epcc.ed.ac.uk

***Index Terms*—B-APM, NVRAM, I/O**

I. Introduction

I/O, the storage and retrieval of data, is a key, and growing, performance bottleneck for a range of application, especially for newer application areas such as large scale data analytics and machine learning. All applications must have some level of I/O, even if it is simply saving results or reading in initial conditions. Therefore, enabling improved I/O performance, either through software optimisations or exploiting new hardware, is important in ensuring applications can efficiently exploit large scale systems.

n3d is a CFD (computational fluid dynamics) simulation code, built on the SEMTEX simulation package [1], designed for investigating wake effects and other turbulent flow features for wind turbines. Turbulence, and wake effects, can have significant impacts on the power that a wind farm can generate for any particular wind condition, so accurately simulating these wind flow features can help to optimize the layout of wind farms and the designs of wind turbine configurations.

n3d utilises a method that involves the integration of the full three-dimensional Navier-Stokes (NS) equation and the adjoint equation iteratively. However, to undertake the adjoint approach the full direct numerical simulation (DNS) simulation output is required, necessitating storage and reading/writing of very large amounts of data. An average simulation could easily require 30TB (Terabytes) of data to be stored between the algorithmic phases, with larger simulations generating hundreds of TB of intermediate data, even at a moderate Reynolds number several orders smaller than the real one. This represents both a large requirement on storage capacity for HPC systems, and a large cost in terms of the time required to first write this data during the forward phase of the simulation (the DNS part), and then read that data back in again for the inverse (adjoint) part of the simulation workflow.

Therefore, whilst the adjoint approach does allow efficient simulation techniques to be utilised, and high fidelity simulations to be undertaken, addressing this I/O bottleneck is important in ensuring that the simulations are as efficient as possible, and the HPC systems running the simulations are utilised effectively. We have ported the n3d application to utilise Intel Optane DCPMM memory within a system that contains 3TB of this non-volatile memory per node. We exploit this in-node storage capability for the transfer of data between the forward and adjoint phases of the simulation, re-writing some of the I/O functionality to target this optimised hardware directly.

We benchmarked the n3d application with a range of different test cases, one that requires only 8 processes to run and generates 600MB of data for the adjoint phase (the small benchmark), one that requires 72 processes to run and generates around 6TB of data (the medium benchmark), and one that requires 512 processes to run and generates around 40TB of data (the large benchmark). For the small benchmark, there was no noticeable benefit from using the non-volatile memory for transferring data between the different algorithmic phases, and profiling confirmed I/O was not a significant performance bottleneck for this benchmark configuration.

However, when moving to the medium and large benchmarks we observed very significant performance improvements using this new memory technology, with around a 15% runtime reduction for the medium case, and an over 80% performance reduction for the large case, compared to transferring the data through a traditional Lustre filesystem, as shown in Figure1.

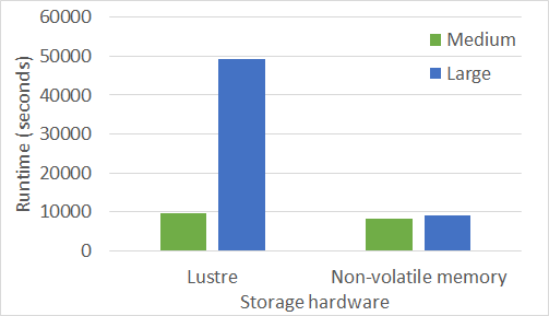

Fig. 1. Performance of Lustre vs non-volatile memory for the medium and large benchmark cases. Total runtime of the application is reported.

These tests were run using 3 compute nodes for the medium case and 22 compute nodes for the large case, on a system with two 24-core Intel Xeon Platinum 8260M processors per node, running at 2.4GHz, with a total of 192 GB of DDR4 Memory shared between the processors, connected together with Intel's OmniPath interconnect. We have also implemented direct memory use through the PMDK library, enabling another 5-10% performance improvement to the overall application runtime through bypassing the block interface used for traditional I/O operations.

This work was funded through the HPCWE project, which received funding from the European Union Horizon 2020 Framework Programme (H2020) under grant agreement number 828799

References

[1] Semtex : a spectral element–Fourier solver for the incompressible Navier–Stokes equations in cylindrical or Cartesian coordinates. Blackburn, H. M.; Lee, D.; Albrecht, T.; Singh, J.