Exploiting persistent memory for workflows and computational simulation

Wednesday, 25 September 2019 16:45 (45 minutes)

NEXTGenIO project, which started in 2015 and is co-funded under the European Horizon 2020 R&D funding scheme, was one of the very first projects to investigate the use of Optane DC PMM for the HPC segment in detail. Fujitsu have built up a 34-node prototype Cluster at EPCC using Intel Xeon Scalable CPUs (Cascade Lake generation), DC PMM (3 TBytes per dual-socket node), and Intel Omni-Path Architecture (a dual-rail fabric across the 32 nodes). A selection of eight pilot applications ranging from an OpenFOAM use case to the Halvade genomic processing workflow have been studied in detail, and suitable middleware components for the effective use of DC PMM by these applications were created. Actual benchmarking with DC PMM is now possible, and this talk will discuss the architecture, the use of memory and app-direct DC PMM modes, and give first results on achieved performance.

Using DC PMM as local storage targets, OpenFOAM and Halvade workflows show a very significant reduction in I/O times required by passing data between workflow steps, and consequently, significantly reduced runtimes and increased strong scaling. Taking this further, a prototype setup of ECMWF’s IFS forecasting system, which combines the actual weather forecast with several dozens of post-processing steps, does show the vast potential of DC PMM: forecast data is stored in DC PMM on the nodes running the forecast, while post-processing steps can quickly access this data via the OPA network fabric, and a meteorological archive pulls the data into long-term storage. Compared to the traditional system configurations, this scheme brings significant savings in time to completion for the full workflow.

Both of the above do use app-direct mode; the impact and value of memory mode is shown by a key materials science application (CASTEP), the memory requirements of which far exceed the usual HPC system configuration of approx. 4 GByte/core. In current EPCC practice, CASTEP uses only a fraction of the cores on each Cluster node – DC PMM in memory mode with up to 3 TBytes of capacity on the NEXTGenIO prototype enables use of all cores, and even with the unavoidable slowdown of execution compared to a DRAM-only configuration, the cost of running a CASTEP simulation is reduced, and the scientific throughput of a given number of nodes is increased commensurately.

Presenter: Dr JACKSON, Adrian (The University of Edinburgh)
Session Classification: Session 4: Memory and Storage