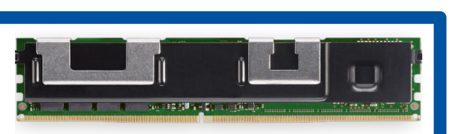
Experience and performance of persistent memory for the DUNE data acquisition system

Adam Abed Abud [1,2]
[1] University of Liverpool, Liverpool, UK [2] CERN, Geneva, Switzerland

Introduction

In this poster we present the performance of persistent memories based on the 3DXP technology. After a short description of the hardware setup, the results obtained in different conditions are discussed: first, raw performance benchmark results are shown for reading and writing; then, the performance of those devices when used within an real-life application are outlined. The workflow of this application mimics a part of the data acquisition system (DAQ) of the DUNE experiment: data are pushed into host memory through custom PCIe cards at a rate of 1 GB/s for each data link. In the target configuration each dual socket server will need to handle 20 data links, and be able to persist 100 s of continuous data upon a specific signal. Therefore, the performance goal is to be able to persist at a rate of 10 GB/s for each CPU socket.

Motivation



DRAM is not a viable solution for the DUNE use-case because it cannot provide storage persistence. The total size (2 TB) needed for 20 data links would also make the system too expensive with only DRAM modules.
To achieve the target goal it is neccessary to use devices capable of providing memory-like bandwidths. Examples of such devices are 3D Xpoint modules such as the **Intel® Optane™ DC Persistent Memory** (**DCPMMs**).
DCPMMs can be configured in multiple ways:

- Memory mode: in this mode the DCPMMs act as a large memory pool alongside the DDR4 memory modules
- App Direct: the DCPMMs provide in-memory persistence by acting as storage devices

Setup and testing strategy

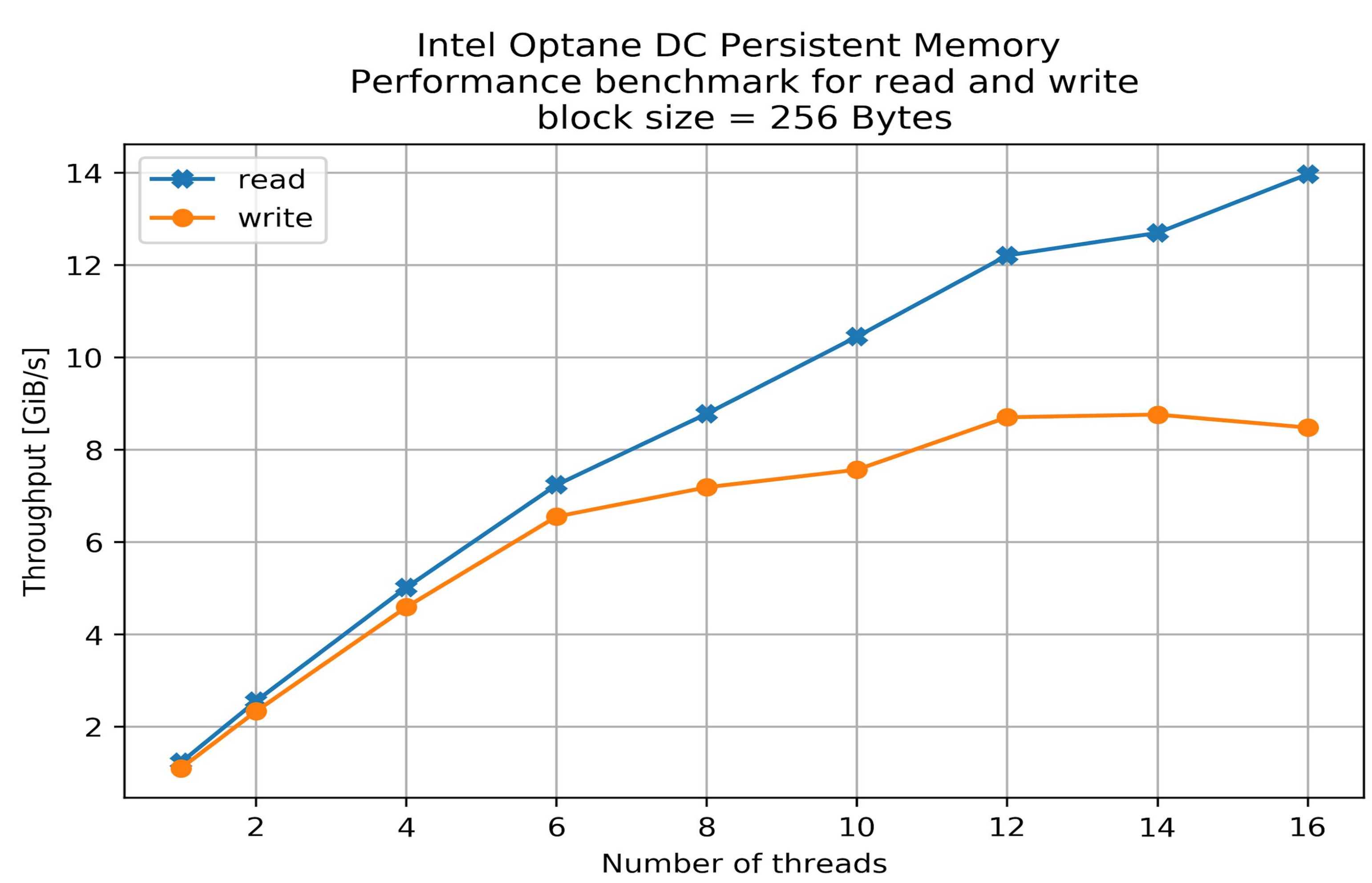
Setup machine:

- Intel Xeon Platinum: dual CPU socket, 56 physical cores
- DDR4 16 GB, 12 slots (192 GB in total)
- DDR-T 512 GB DCPMM, 12 slots (6 TB in total)
- OS: Centos 7 and kernel version 4.15

The raw performance was obtained by executing a synthetic benchmark evaluation with the flexible-IO tool (FIO). This was executed with the DCPMMs used as a storage device in **App Direct mode** and mounted with a DAX-enabled ext4 file system. The system was also tested with a high level C++ application that used the DCPMMs as the target storage media.

Benchmarks of DCPMMs

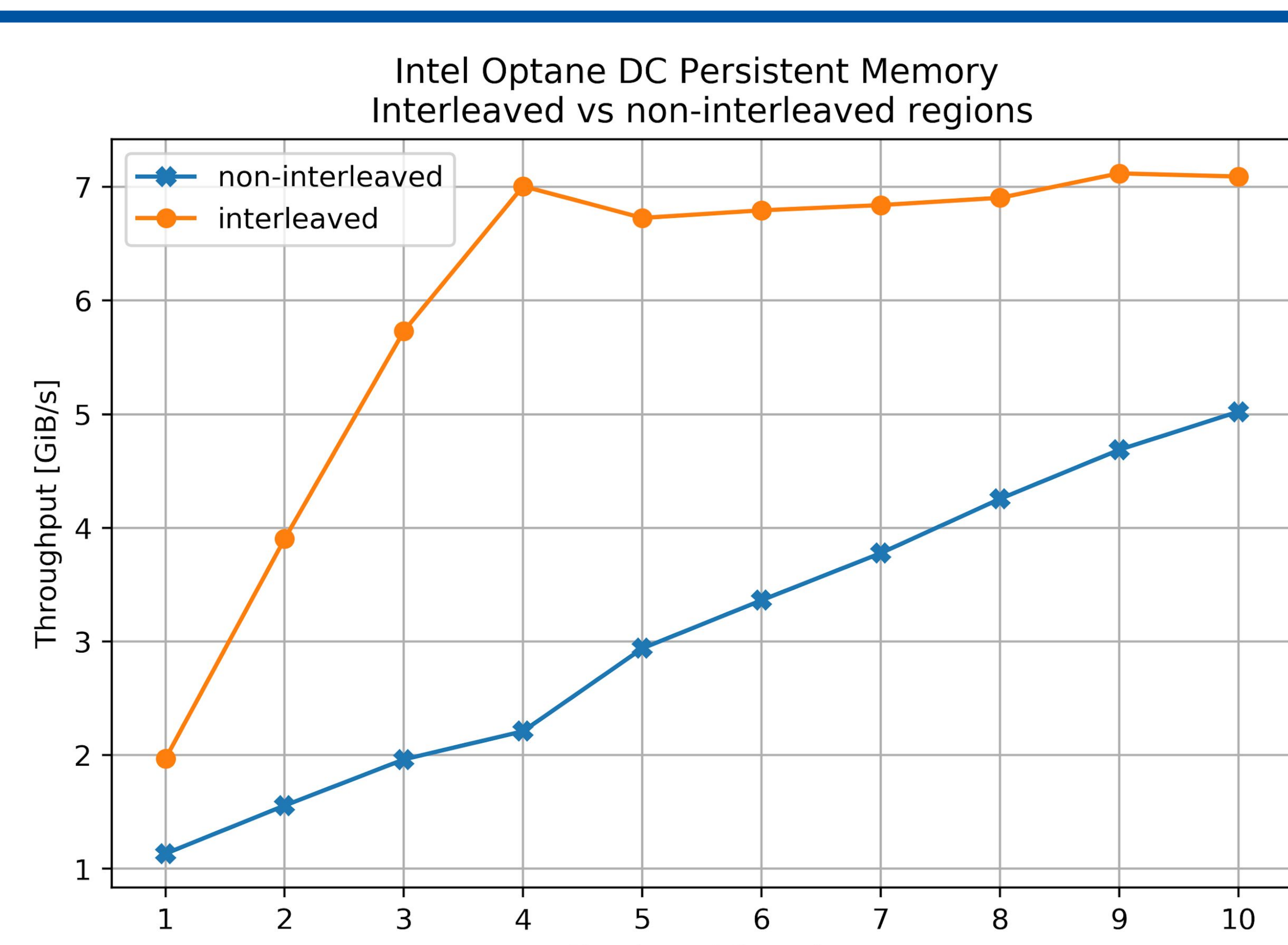
The **maximum** achieved **write throughput** for the DCPMMs used as storage devices is approximately **8.5 GiB/s**. This was achieved by configuring the modules in the **non-interleaved** mode and using a block size of 256 bytes.
CPU affinity was set up to avoid any cross-NUMA access that can lead to an increased latency time and, therefore, a reduced throughput.
Software optimizations were executed by memory mapping the block of data and then using the **MOVNTI non-temporal** SSE **instruction** provided by the processor. In this way, the operation has **no overhead from the file system** because it invalidates the cache line and therefore results in a pure device operation.



Application for DUNE DAQ

Based on the benchmarks of the DCPMMs a more complete application was developed resembling as much as possible the workload needed for the DUNE experiment. This was obtained by memory copying a static buffer and persisting it with the DCPMMs used in App Direct Mode.

The system saturates the available bandwidth with a throughput of 7 GiB/s starting from 4 threads. This means that with the current DCPMMs available today it is possible to sustain, per CPU socket, **80% of the target throughput**.
The non-interleaved configuration was also tested. This provides a block level access for each DCPMM module. However, the performance obtained is 60% lower compared to the interleaved configuration in the case of 5 threads.



Conclusion and outlook

The Intel® Optane™ DC Persistent Memory modules have been tested in detail with synthetic benchmarks. The throughput obtained is approximately 10 GB/s per CPU socket. A high-level application mimicking a part of the data acquisition system of the DUNE experiment was developed and it sustained only 80% of the target throughput per CPU socket. In the future, further optimizations to the test application will be considered. In addition, the new generation of DCPMMs promises the missing bandwidth that is needed to reach the required performance goal.

adam.abed.abud@cern.ch