Contribution ID: **74** Type: **Poster**

Memory Power Consumption on Heterogeneous Memory Systems

Tuesday 21 March 2023 18:30 (1 hour)

The architecture of supercomputers over the years has evolved to support different need in applications that seek to solve some human concerns. Heterogeneity role nowadays is important in processors and also in the memory-storage system. In processors, we can observe CPUs, GPUs and other accelerators coexisting. In the same fashion, different kinds of memory have appeared over the years, fulfilling some gaps in the memory-storage continuum. E.g., high bandwidth memory (HBM), that is embedded on the processor package, coexist mainly with dynamic random access memory (DRAM) into Intel Xeon Phi Processors or Knight Landing (KNL). Non-volatile memory (NVM), that can be found with DRAM into the 2nd Generation Intel Xeon Scalable Processors. Nowadays, the upcoming Intel Sapphire Rapids support HBM inside the processor package, DRAM through the memory bus, and also it supports disaggregated memory by the Compute Express Link (CXL) that in principle allows to connect HBM, NVM and DRAM on it.

The task of developers when programming new applications or adapting the existing ones requires full knowledge of the memory system and without a specific strategy it can be very complicated depending on the conditions in which the applications are required to run. Today, for developers is pertinent to prepare their applications so that they adequately face at least the main heterogeneous memory system (HMS) setups. For that reason we consider that every developer should at least understand HMSs in terms of simple and easy metrics such as: bandwidth, latency, capacity, data persistence, power consumption, etc. Especially, it is essential to know how much memory power applications are going to use in a given memory system. It is vital in situations where executions need to be performed with minimal power consumption mode, or when we need to balance power consumption and performance.

In this poster presentation, we focuses on understanding and giving a perspective on how to analyse memory energy consumption metric over different HMS setups.
We consider, identifying and exposing the memory system in the simplest manner developers could access. E.g., a memory system with DRAM and NVM can be exposed as different NUMAs in some systems and their access implies binding the applications process to the kind of memory required. Then, we have selected a some memory-intensive applications that should be profiled. Profiling depends on the expertise of developers and also tools can give more or less information depending on their capabilities. In our case, Intel Performance Counter Monitor (PCM) enables the possibility to get some performance counters related to memory power consumption in between others related to bandwidth. Also we used Linux Perf profiler tool to retrieve relevant information related to cache misses and verify if applications are behaving as a memory-intensive application. The final objective when analysing the power consumption metric is to be able to give a certain ordering for which the developer can look for a memory with very low consumption, as she/he could look for one that allows her/him to have a balance between the performance of the applications and the consumption of memory power. In addition to this analysis, we have sought to provide developers with an early HMS memory power prediction model, which allows getting an idea of the possible consumption of their application towards a given HMS.

JLESC topic

Primary author: Mr RUBIO PROAÑO, Andrès (Riken)

Co-author: SATO, Kento (Riken)

Presenter: Mr RUBIO PROAÑO, Andrès (Riken)

Session Classification: Poster Session

Track Classification: Performance tools