CASE STUDY

High Performance Computing (HPC)
2nd Generation Intel® Xeon® Scalable Processors
Intel® Optane™ Persistent Memory

Barcelona Supercomputing Center Research Accelerates HPC Workloads with Heterogeneous Memory

Intelligent allocation of data to Intel Optane persistent memory stores more data close to CPU with fewer power-hungry DIMMs

Title:

- Experimental platform for developing hierarchical memory software tools
- Intel Xeon Platinum 8260L processors
- 6 TB of Intel Optane persistent memory

Executive Summary

Antonio Peña, a senior researcher at Barcelona Supercomputing Center/Centro Nacional de Supercomputación (BSC) is heading development of new software tools and expanding the software ecosystem for 2nd Generation Intel Xeon Scalable processors and Intel Optane persistent memory (Intel® Optane PMem). His work, under the Intel-BSC Exascale Lab, and in collaboration with the EPEEC project (European joint Effort toward a Highly Productive Programming Environment for Heterogeneous Exascale Computing), is helping accelerate High Performance Computing (HPC) applications using heterogeneous memory architectures.

Peña is no stranger to High Performance Computing (HPC). At the U.S. Department of Energy's (DoE) Argonne National Laboratory (ANL) he contributed to projects that included accelerating Message Passing Interface (MPI)-based applications, studying remote virtualized GPUs, and pioneering heterogeneous memory systems in HPC clusters. Under the direction of computer science director professor Jesús Labarta and in close collaboration with Dr. Harald Servat and Marie-Christine Sawley from Intel, he continues to explore how to accelerate large HPC workloads by leveraging hierarchical memory systems. With Intel Optane PMem and 2nd Generation Intel Xeon Scalable processors, he is driving breakthrough architectures that enable high-performance workloads with large data sets on HPC clusters using less power than DRAM.

Challenge

"Right now, HPC applications are constrained by the amount of DRAM in the nodes and cluster," Peña explained. "They need more and more memory but adding larger and more DIMMs with the current technology is not feasible due to the power constraints on the overall system."

In today's large HPC clusters, main memory is typically sized between 2 and 3 GB per core to optimize performance. A study in 2016 showed that scores from High-Performance Linpack (HPL) runs tend to plateau around 2 GB of main memory in typical HPC systems.[1] Indeed, the leading x 86-based clusters on the Top 500 list follow this trend. For example, BSC's MareNostrum 4, with 3,240 nodes (165,888 Intel Xeon Platinum 8160 processor cores), has 2 GB/core on all but a relatively few of the nodes. Two hundred sixteen of MareNostrum 4's nodes offer large memory capacity with 8 GB/core to accommodate much larger data sets. The same study predicted that as data sets expand and systems of the future continue to grow to very large clusters (e.g. one million cores), to reach as much as 99 percent of the ideal HPL performance will require per-core memory of 7 to 16 GB. That will have a large impact on server power consumption.

Traditional memory manufacturers recommend budgeting about three watts per 8 GB of DDR3 or DDR4 memory—and more for RDIMMs.[2] In a large HPC node with 56

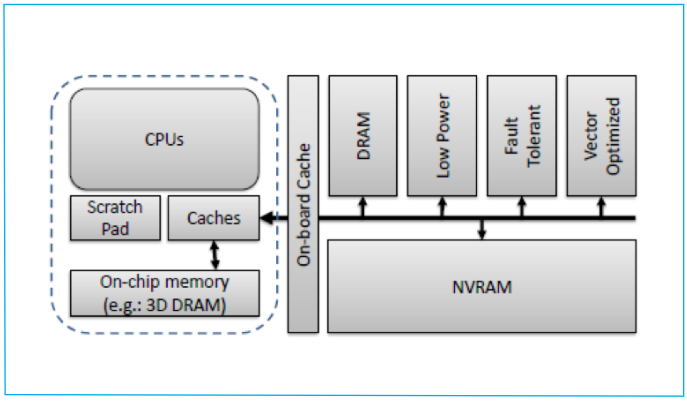
x 86 cores and 112 GB of memory (2 GB/x 86 core), memory budget should be between 42 and 50 watts. Large memory nodes with 8 GB/core could consume 168 watts or more. Looking into the future with massively large systems based on processors with 112 cores per node, for example, memory power itself becomes unwieldly, possibly consuming nearly 700 watts in one node.[3] Hosts of today's HPC systems want to accelerate applications, but reduce power demand, not increase it by expanding memory.

Peña's research shows promise for using heterogeneous memories, including large banks of non-volatile memory, to help accelerate applications at lower power.

Solution

"We're trying to reduce server power while accelerating applications by using Intel Optane PMem and intelligently managing where the data is located and its movements," Peña said. "We can take advantage of the big memory footprint that the new technology offers and put more data closer to the processor. There is a slightly longer latency than DRAM, but we don't have to pay for the penalty of even more latency going to other storage technologies."

Peña's vision of heterogeneous memory architectures formed while at ANL, eventually working with the Intel® Xeon Phi™ processor. He has continued his work at BSC using Intel Optane PMem.



Simplified view of a heterogeneous memory node sample.[4]

Intel Optane PMem is a high-density 3D memory technology in a DIMM form factor that delivers a unique combination of large capacity, low power, and support for data persistence on the memory bus. 2nd Generation Intel Xeon Scalable processors support the memory modules.

"Large non-volatile memory as a basis for hierarchical memory solutions is a great candidate," Peña added. "It's byte-addressable, so we can use it for regular load-stores, offers large capacity, and uses less power."

Intel Optane persistent memory can operate in different modes—Memory Mode, App Direct Mode, and Storage over App Direct. In Memory Mode, the CPU manages the memory with DRAM transparently used as a cache for the most frequently accessed data. In App Direct Mode it can be used as either volatile or nonvolatile memory where the application intelligently handles allocation. Peña is using the technology in App Direct mode, treating the memory modules as large capacity, manageable system memory with data objects intelligently allocated to optimize the performance of the running application. To optimize usage in this mode, Peña's team develops and runs custom software to both tune where data is placed in the memory structure and directly manage the persistent memory. The team is collaborating with Intel to help build a new software ecosystem around byte-addressable persistent memory to accelerate HPC applications while enabling more power-efficient supercomputers.

Using Intel Optane PMem, power reduction is significant. Compared with the recommended 3 watts per 8 GB (375 mW/GB) for standard DDR3 and DDR4 DIMMs, Intel Optane PMem's 128 GB memory modules consume only 117 mW/GB, and the 512 GB modules use just 35 mW/GB—a 10X reduction in power compared to DRAM DIMMs.

Innovative Data Profiling and Memory Allocation Tools for Intelligent Data Management

To enable his approach with heterogeneous memories, Peña and his team have created several software tools using Extrae, a general-purpose profiler developed by BSC, Intel vTune™ profiler, and Extended Valgrind for Object Differentiated Profiling (EVOP), among others. EVOP was first developed by Peña at ANL and is now maintained at BSC. Their tools first perform what Peña calls data-oriented profiling by running the profiling tools while the application executes normally. The tools analyze the demand and latencies for different objects and create a large file listing all data accesses.

"Knowing how each data object is accessed during execution helps us decide in the optimization step where those have to be allocated in the different memories," Peña described. "In a simplified view, we associate metrics with the different data objects. Then we count the number of accesses or the number of last level cache misses for each object. From this, we can apply different algorithms for memory allocations to maximize the performance."

Once the software knows how and when data objects are accessed and the size of the data object and size of memory tiers, the results are fed into variations of multiple knapsack problems.[5] A knapsack optimization algorithm attempts to fit the most objects of the most value into a given 'container' with a limited capacity. In Peña's case, the memory objects are the items and memories in the system are the knapsacks or containers. The output of the problems provides guidance to allocate the data objects to appropriate memories in the hierarchy.

"After profiling, a script we call the hmem advisor, for heterogeneous memory advisor, parses the large data object profile and generates a distribution list of different objects to the memories," Pena continued. "Then we can run the application binary without changes and, as regular mallocs are called, we have a runtime library, an interposer, that intercepts these calls to allocate the different data objects to the appropriate memory tier."

For their research, Peña's team uses a system at BSC with 2nd Gen Intel Xeon Scalable processors and 6 TB of Intel Optane PMem to do their development and run their testing. The system is part of the Intel-BSC Exascale Lab, where, since 2011, innovative research and collaborative development projects have been carried out to enable next-generation Exascale supercomputing.

Result

Heterogeneous memory research is ongoing at BSC. Currently, Peña's team's code runs statically. It allocates memory objects based on a profiling run and optimization steps. But the team's plans are to make it much more dynamic to accommodate changes at run time and user-specific marking of data.

"Today, the tool focuses on allocating data for optimized performance," Pena explained. "If there are standard DIMMs, ECC DIMMs, and Intel Optane PMem modules in a system, we would allocate the most called data to the standard DIMMs, then less frequently demanded data to ECC, and even less demanded objects to NVRAM. But we envision the tool to not only be dynamic at run time but be able to respond to user marks of data as well. For example, if the user wants to ensure data is protected, we will allocate it to ECC instead of standard DRAM. Or, if it has certain access patterns, such as many writes, we will allocate it to regular memory instead of NVRAM, which has slower write speeds."

Peña's work targets large HPC workloads seen in the world's supercomputing centers, but they are also running it on smaller applications, such as Intel® Distribution for HPCG, Lulesh, miniFE and SNAP. His team is benchmarking their code performance against using Intel Optane PMem in Memory Mode, where the processor itself manages the data.

"We are testing our code on most of the applications from the U.S. DoE, like LAMMPS, OpenFOAM, and NWChem. A key goal of this project is to enable large applications to run with high performance on systems with large NVRAM capacities and smaller amounts of DRAM. We are currently seeing performance improvements in many mini-applications, plus up to 18 percent in OpenFOAM and 10 percent on LAMMPS compared to Memory Mode," Peña concluded.

Peña's approach optimizes performance and power for supercomputers where larger and larger problems on bigger and bigger machines means more and more memory.

Solution Summary

Barcelona Supercomputing Center and the Intel-BSC Exascale Lab are at the heart of innovative hierarchical memory research using 2nd Gen Intel Xeon Scalable processors and Intel Optane persistent memory in App Direct Mode. The center hosts a team lead by Antonio Peña to develop software tools that intelligently allocate data to multiple memory tiers, including Intel Optane PMem, that help optimize performance of large supercomputing applications. Their ongoing research for both large and small HPC applications is showing up to a 2x performance speedup in some mini-applications such as MiniFE compared to Memory Mode according to Peña.

Where to Get More Information

Learn more about Barcelona Supercomputing Center.

Find out more about the 2nd Generation Intel Xeon Scalable processors.

Learn more about Intel Optane persistent memory.

Solution Ingredients

- Experimental platform for developing hierarchical memory software tools
- Intel Xeon Platinum 8260L processors
- 6 TB of Intel Optane persistent memory

[1] https://www.researchgate.net/publication/314290425_Main_memory_in_HPC:_Do_we_need_more_or_could_we_live_with_less

[2] https://www.crucial.com/support/articles-faq-memory/how-much-power-does-memory-use

[3] Intel Xeon 9282 processor with 56 cores/socket * 2 sockets/node = 112 cores * (16 GB/core ÷ 8 GB * 3 watts/GB)

[4] Based on Intel Optane persistent memory brief at https://ark.intel.com

[5] https://en.wikipedia.org/wiki/Knapsack_problem

062020/RJMJ/RL/PDF Please Recycle 42568-001US