Evaluating the impact of cache control technologies on HPC workloads

Keywords: High Performance Computing, Applications Instrumentation, Cache and Performance Modeling, Traces Analysis.

1 Description

Large scientific simulations are often made of several components that match distinct phases and distinct parts of the computation. Such components perform computations that include, linear algebra, iterative schemes, data aggregation and postprocessing, information aggregation, visualization, and so on. These computations have typically different requirements that translate into various load levels on the hardware resources. Overall, they can be qualified as highly heterogeneous, some of them requiring large memory pools, while other taking advantage of high floating point computation capabilities or larger caches.

Current high performance computers being rather homogeneous, it is often necessary to map several of these simulation components on the same hardware resource in order to take advantage of the overlapping of their requirements while maximizing parallelism. However, doing this blindly might result in component hindering each other by competing for access to memory, computationnal resources or cache.

Regarding shared resources like caches in many-cores architectures, recent technologies have been made available to partition them among distinct groups of processes, for example with the Cache Allocation Technology on modern Intel processors. Using an interface similar to low-level container mechanims, one can use this technology to split shared caches between competing workloads.

Unfortunately, the impact of such technology on the performance of HPC workloads is still not well understood, making it difficult to choose the right allocation policy, or to tune it on the fly during execution.

2 Expectations

To address this issue, the purpose of this internship is to perform a wide experimental study of the impact of cache controls on HPC workloads with a focus on its impact on runtime in the context of multiple workloads competing for resources. This study will result in the design of predictive models for how a cache control policy would impact a known workload at runtime, focusing on either cache capacity or cache bandwidth control. The evaluation will use currently available processors and existing software allowing the control of Intel's CAT.

3 Contact

- Guillaume Huard <Guillaume.Huard@imag.fr>
- Swann Perarnau <swann@anl.gov>