Parallel Computing and I/O

System profiling and data aggregation for smart compression in Lustre

- Author: Pablo Correa Gómez
- Type: Master's Thesis
- Date: 2022-08-17
- Reviewers: Jun.-Prof. Dr. Michael Kuhn, Prof. Dr. Thomas Ludwig
- Supervisors: Anna Fuchs, Jun.-Prof. Dr. Michael Kuhn
- Download: PDF

In High-performance computing (HPC) setups, the IO and data transfer can be a big part of the processing requirements of scientific applications. When that is the case and they become a bottleneck, the application performance can degrade. This problem is expected to become more common since CPU processing has been for many years and continues growing at a faste rate than network or storage speed. Moreover, the imbalance between different machines with different roles in the setup and applications' inefficiencies make this problem worse. In this thesis, compression is considered a solution to this problem. Compression allows trading the excess in computation power for a reduction in the data size, both for IO and transfer. However, static compression can potentially result in a similar set of inefficiencies as those that it aims to solve. For this reason, I propose extract and analyze information from the HPC setup, introducing a collection and decision-making process that makes compression smart. The integration point for compression is the parallel filesystem, which is the piece of software that, in HPC, takes care of the data transfer. For this work, Lustre, the most popular filesystem among big HPC deployments, is the filesystem choice. In consequence, this thesis analyses a typical Lustre setup to identify and extract the components that would take part ir the smart compression. Those components are studied to obtain the metrics relevant for compression. Later, the required proce for smart compression is considered, and two relevant decisions, the location of the compression, and the algorithm configuratio to use are analyzed in detail. The analysis assesses the key metrics for each decision and possible ways to integrate their calculation. Finally, to prove the relevance of smart compression, a small set of experiments show both the benefits of compress and the dangers of a wrong configuration.