Data-Intensive Computing on Commodity Hardware: Hadoop and Beyond

Elmar Kiesling

TU Wien

In recent years, Hadoop has evolved into a mature platform for large-scale data processing on commodity hardware that has quickly gained widespread adoption in many industries. More recently, Hadoop also increasingly opens up new opportunities in data-intense scientific computing applications.

At its core, Hadoop provides a fault-tolerant framework for large-scale batch processing on commodity hardware. This framework abstracts the hardware infrastructure and handles low-level aspects of parallel computing through a distributed file system (HDFS) and a simple parallel programming model (MapReduce). This model makes it very easy to implement robust large-scale data-processing jobs, but also imposes limitations on the types of problems that can be addressed efficiently, i.e., highly parallelizeable analytic tasks with a single synchronization barrier. More recently, frameworks such as Spark have expanded this scope towards more general, iterative, and interactive workloads. Consequently, Hadoop clusters can increasingly tackle more complex data-intensive parallel processing tasks and are now widely used, e.g., in machine learning scenarios.

In this talk, we will discuss how big data challenges differ from the typical requirements in traditional HPC scenarios. Next, we will review how this is reflected in the Hadoop design priciples, architecture, and typical hardware configuration before embarking on a tour of the expanding Hadoop ecosystem. Finally, we will close with a brief outlook on potential future developments of Hadoop as a platform for data-intensive parallel computing.