High Performance Data Analytics for Numerical Simulations (mésocentre CIMENT)

Bruno Raffin*[1]

[1]Laboratoire d'Informatique de Grenoble (LIG) – CNRS : UMR5217, Université Pierre-Mendès-France - Grenoble II, Institut polytechnique de Grenoble (Grenoble INP), Université Joseph Fourier - Grenoble I – UMR 5217 - Laboratoire LIG - 38041 Grenoble cedex 9 - France Tél. : +33 (0)4 76 51 43 61 - Fax : +33 (0)4 76 51 49 85, France

Résumé

Large scale numerical simulations are producing an ever growing amount of data that represent a double challenge. First, these amounts of data are becoming increasingly difficult to analyse relying on the traditional tools. Next, moving these data from the simulation to disks, to latter retrieve them from disks to the analysis machine is becoming increasingly costly in term of time and energy. And this situation is expected to worsen as supercomputer I/Os and more generally data movements capabilities are progressing more slowly than compute capabilities. While the simulation was at the center of all attentions, it is now time to focus on high performance data analysis. This integration of data analytics with large scale simulations represents a new kind of workflow that needs adapted software solutions. In this talk we will survey two directions: big data like solutions and in-situ analysis. Big Data Analytics solutions like Google MapReduce, Spark or Flink were developed to answer the needs for analyzing large amount of data from the web, social networks, or generated by business applications on cloud infrastructures. We will give an overview of some research work that either developed their specific map/reduce stack for analyzing scientific data or relied on classical Big Data stacks like the Velassco project. In-situ analysis attempt to more specifically address the reduction of data movements and data storage. In-situ analysis proposes to start processing data as soon as made available by the simulation in the memories of the compute nodes. Raw data produced by the simulation can start to be reduced before moving out of the compute nodes, thus saving on data movements and on the amount of data to store to disk. Part of data analysis can be performed on the same supercomputer than the one booked for the simulation. The process can be massively parallelized, reading data from memory and not from disk, reducing the time for performing these tasks. We will give an overview of in-situ approaches with some examples. The conclusion will summarize the main challenges that still need to be addressed before high performance data analysis tools become a commodity in the scientist toolbox.

*Intervenant