HPC-BigData Convergence: What to do when scientific data becomes too big?

Franck Cappello*[1]

[1]Argonne National Laboratory [Lemont] – United States

Abstract

A critical common problem in consumer big data applications and scientific computing (HPC) is the need to communicate, store, compute and analyze extremely large volumes of high velocity and diverse data. For many scientific simulations and instruments, data is already "too big". Architectural and technological trends of systems used in HPC call for a significant reduction of these big scientific datasets that are mainly composed of floating-point data. In this talk, we present experimental results of currently identified use cases of generic lossy compression to address the different limitations related to processing and managing scientific bigdata. We show from a collection of experiments run on parallel systems of a leadership facility that lossy data compression not only can reduce the footprint of big scientific datasets on storage but also can reduce I/O and checkpoint/restart times, accelerate computation, and even allow significantly larger problems to be run than without lossy compression. These results suggest that lossy compression will become an important technology in many aspects of the convergence between HPC and bigdata. This talk is intended to develop discussion between the data compression, the HPC and scheduling communities.

*Speaker