Contribution ID: **106** Type: **Talk**

The Data Deluge: Overcoming the Barriers to Extreme Scale Science

Friday, 2 December 2022 12:00 (30 minutes)

The rapid growth in technology is providing unprecedented opportunities for scientific inquiry. However, dealing with the data produced has resulted in a crisis. Computer speeds are increasing much faster than are storage technology capacities and I/O rates. This ratio is also getting worse for experimental and observational facilities, where for example, the Legacy Survey of Space and Time (LSST) observatory will collect up to 20 TB per night in 2022, yet the Square Kilometre Array will generate over 2 PB per night in 2028. This reality makes it critical for our community to 1) Create efficient mechanisms to move and store the data in a Findable, Addressable, Interoperable, and Reproducible (FAIR) fashion; 2) Create efficient abstractions so that scientists can perform both online and offline analysis in an efficient fashion; 3) Create new reduction algorithms which can be trusted by the scientific community, and which can allow for new ways to not only reduce/compress the data but also to reduce the memory footprint and the overall time spent in analysis.

To tackle these goals, My group had worked closely with many large-scale applications and researchers to co-design critical software infrastructure for these communities. These research artifacts have been fully integrated into many of the largest simulations and experiments, and have increased the performance of these codes by over 10X. This impact was recognized with an R&D 100 award in 2013 and was highlighted in the 2020 US Department of Energy (DOE) Advanced Scientific Computing Research (ASCR) @40 report. In this presentation, I will discuss the research details on three major contributions I have led: large-scale self-describing parallel I/O (ADIOS), in situ/streaming data (SST), and data refactoring (MGARD). I will introduce the overall concepts and present several results from our research, which has been applied and fully integrated into many of the world's largest scientific applications.

Primary author: Dr KLASKY , Scott (Oak Ridge National Laboratory)

Presenter: Dr KLASKY , Scott (Oak Ridge National Laboratory)

Session Classification: HPC

Track Classification: Storage and IO