Big Data research for the Exascale era and beyond

Saba Sehrish, Marc Paterno, Jim Kowalkowski

April 20, 2018

Data format

- HDF5 for optimized parallel IO on HPC
  - Writing data from existing framework
  - Parallel reading for parallel analysis
- Why?
  - The Exascale Computing Project is “a collaborative effort of two U.S. Department of Energy organizations – the Office of Science (DOE-SC) and the National Nuclear Security Administration (NNSA).” The ECP includes the EXAHDF5 project, which states: “This project is endeavoring to develop optimally performing parallel I/O strategies on upcoming exascale architectures, to maintain and optimize existing HDF5 features for ECP applications, and to release new features in HDF5 for broad deployment on HPC systems.”

Data processing

- Use of high level APIs to define data processing tasks
- Use of functional programming approach
- Define operations on the whole data-set
- Using distributed processing framework and libraries
  - Abstract parallel constructs from the user code
- Examples
  - Python/mpi4py/h5py
  - Do-It-Yourself (DIY) framework – C++/MPI
- Aurora ESP proposal submitted for 2B core hours to do NOvA science using this technology (Aurora, the first exascale machine in mid-2021)