SAND2015-8351R

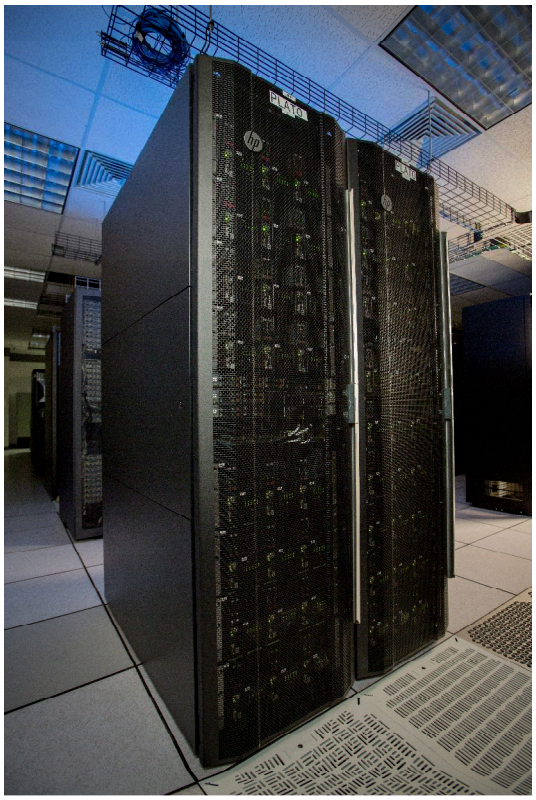
High Performance Data Analytics on a Commodity Private Cloud

Team: Christopher Beggio. PI

Contact cabeggi@sandia.gov

Sandia is researching novel methods of analysis for large datasets. The term “Big Data” has been used frequently in recent years to describe a discipline of computer science that attempts to distil an increasing quantity of data into a result or meaningful subset that can be easily visualized and acted upon by human consumers. These data are frequently characterized by one or more of the following: high volume, velocity, variety, and veracity. Sandia scientists and researchers are no strangers to large datasets, which often typify the result of high-performance computing workloads. The challenge for Sandia data scientists, however, comes at the arrival of exascale computing, which has become mandated by executive order for national computing centers. Because of power and networking limitations, exascale computing precludes the movement of datasets at exascale from data production to data analysis. Such analysis must be done in-situ, to mitigate the time and power costs of network transport. Currently, many of Sandia’s data analytics environments exist as independent entities, often specialized for individual use cases, interconnected to other data analytics, high-performance computing, and storage systems.

Sandia is procuring, installing, and researching future platforms for such data analytics. A memory footprint of 256 GB per node, fourteen data rate (FDR) InfiniBand networking, PCIe connected non-volatile memory (NVMe) data storage, and an object data store will characterize the coming institutional data analytics platforms. The combined systems will contain 300 compute nodes. A cloud-inspired software stack resulting from collaboration between the Sandia centers will allow the systems to support a flexible mix of user and system-facing workloads and use cases including code development and testing in R and Python, Apache Hadoop and Spark, distributed in-memory databases, graph analytics, emulytics, system log and predictive analytics, and data visualization for institutional and mission customers. These systems will arrive at the end of FY 2015 and production is planned by the spring of FY 2016.



Plato, a pilot system for Sandia's institutional data analytics effort, runs primarily Hadoop and Spark.