CASE STUDY

Applying Big Data Analysis Techniques to Simulated DNA Data

Organization

University of Nottingham
School of Pharmacy and Centre for Biomolecular Sciences
Nottingham, UK
www.nottingham.ac.uk/cbs

Application

First-ever analysis of simulation-generated DNA data carried out by running Apache Hadoop® natively on a Cray® XC30™ supercomputer

Cray XC30 "ARCHER" Supercomputer

ARCHER is the latest U.K. National Supercomputing Service. The 2.5 petaflops Cray XC30 system is jointly funded by the EPSRC and NERC research councils, housed at the University of Edinburgh's Advanced Computing Facility, and supported by EPCC and Daresbury Laboratory. ARCHER provides a capability resource allowing researchers to run simulations and calculations requiring large numbers of processing cores working in a tightly coupled, parallel fashion. It features 9,840 12-core Intel® Xeon® processors and a Cray Aries interconnect.

Cray Centre of Excellence

Cray established a Centre of Excellence in 2008 at the University of Edinburgh as a part of HECToR, the predecessor to ARCHER. The center provides in-depth support for current and future Cray supercomputing platforms through advanced research in porting codes, improving scalability of applications, developing algorithms and system software tools and optimizing workload and I/O.

Background

Researchers from the Centre for Biomolecular Sciences at the University of Nottingham along with the Edinburgh-based Cray Centre of Excellence team have been carrying out a pioneering study applying big data analysis techniques to simulation-generated DNA data. Running Apache Hadoop® on "ARCHER," the U.K. National Supercomputing Service's Cray® XC30™ supercomputer, the team has assessed how different DNA sequences vary in shape and flexibility.

From a scientific perspective, this work paves the way for researchers to understand processes such as how genes are switched on and off in both healthy individuals and those with a disease. From a supercomputing perspective, however, this analysis moves the industry closer to data-centric computing. The study represents the first time such work has been carried out by running Hadoop natively on a Cray XC30 system.

Challenge

Big data analysis of simulation-generated scientific data is often carried out using Hadoop or Spark™ frameworks. These frameworks offer a mature, flexible programming environment, and users can find dedicated Hadoop or Spark clusters. However, the datasets are often so large it takes a long time to transfer to such appliances.

The University of Nottingham team uses molecular dynamics (MD) to visualize how the shapes of different DNA sequences evolve over time. Key tasks in understanding this data include identifying structures that remain stable over time and assessing how they match those of other molecules. These tasks require a "similarity analysis" for the molecules' patterns — work well suited to the MapReduce programming approach of Hadoop.

The usual use case for Hadoop involves moving the data to a dedicated cluster and carrying out the analysis there. While this method works well for small datasets, such as 140-character tweets, it is not practical with the gigabytes or terabytes of data generated in a typical MD simulation.

Solution

The Cray team applied a data-centric solution to the researchers' problem. Specifically, they took the compute to the data, rather than the reverse. Cray has been developing a solution where Hadoop can be run natively on the compute nodes of the same XC30 system that generated the data. Besides saving time on data movement, this method is more efficient for supercomputing service providers who no longer need a dedicated Hadoop appliance and can vary the number of nodes running Hadoop to fit the scale of the problem.

For this study, an initial proof-of-concept analysis of simulation data generated on ARCHER was designed and implemented using privileged access to pre-release Cray software. Preliminary data shows that analysis time can be reduced by increasing the number of XC30 system nodes participating in the analysis.

This analysis shows that data-centric computing is both feasible and useful on supercomputing platforms, and allows greater scientific benefits to be gained from simulations through novel analysis techniques.

Cray Inc.
901 Fifth Avenue, Suite 1000
Seattle, WA 98164
Tel: 206.701.2000
Fax: 206.701.2500
www.cray.com

©2015 Cray Inc. All rights reserved. Cray is a registered trademark of Cray Inc.
All other trademarks mentioned herein are the properties of their respective owners. 20151029EMS