Carnegie Mellon University
Computer Science Department

THESIS PROPOSAL

Qing Zheng

Distributed Metadata and Streaming Data Indexing as Scalable Filesystem Services

Friday, June 14, 2019 – 3:00 p.m. – GHC 8102

As people build larger and more powerful supercomputers, the sheer size of future machines will bring unprecedented levels of concurrency. For applications that write one file per process, increased concurrency will cause more files to be accessed simultaneously and this requires the metadata information of these files to be managed more efficiently. An important factor preventing existing HPC filesystems from being able to more efficiently absorb filesystem metadata mutations is the continued use of a single, globally consistent filesystem namespace to serve all applications running on a single computing environment. Having a shared filesystem namespace accessible from anywhere in a computing environment has many welcome benefits, but it increases each application process' communication with the filesystem's metadata servers for ordering concurrent filesystem metadata changes. This is especially the case when all the metadata synchronization and serialization work is coordinated by a small, fixed set of filesystem metadata servers as we see in many HPC platforms today. Since scientific applications are typically self-coordinated batch programs, the first theme of this thesis is about taking advantage of knowledge about the system and scientific applications to drastically reduce, and in extreme cases, remove unnecessary filesystem metadata synchronization and serialization, enabling HPC applications to better enjoy the increasing level of concurrency in future HPC platforms.

Overcoming filesystem metadata bottlenecks during simulation I/O is important. Achieving efficient analysis of large-scale simulation output is an even more important enabler for fast scientific discovery. With future machines, simulations' output will only become larger and more detailed than it is today. To prevent analysis queries from experiencing excessive I/O delays, the simulation's output must be carefully reorganized for efficient retrieval. Data reorganization is necessary because simulation output is not always written in the optimal order for analysis queries. Data reorganization can be prohibitively time-consuming when its process requires data to be readback from storage in large volumes. The second theme of this thesis is about leveraging idle CPU cycles on the compute nodes of an application to perform data reorganization and indexing, enabling data to be transformed to a read-optimized format without undergoing expensive readbacks.

Thesis Committee:
Garth Gibson, Co-Chair
George Amvrosiadis, Co-Chair
Gregory Ganger,
Bradley Settlemyer, Los Alamos National Laboratory

Thesis Summary: http://www.cs.cmu.edu/~qingzhen/thesis_proposal.pdf