Panthera: Caching and Cache-based Scheduling in Distributed Computing Systems

Pandya, Dhaivat

The adoption of distributed computing systems has grown massively in the past few years. In particular, Apache Hadoop, which allows developers to create applications that run on a cluster of computers, is currently used throughout academia and industry in areas such as machine learning, medical diagnosis, natural language processing, etc. However, the Hadoop File System fails to effectively utilize random access memory (RAM) and local storage in order to reduce waiting time (i.e. latency). In this project, named Panthera, caching and scheduling systems for Hadoop were developed. Panthera caches both information about files and the file contents, thereby reducing waiting time in accessing the files and related information. It runs with an unmodified version of Hadoop, meaning it can integrate easily into existing architecture. The results showed that data access latency was 9 times lower with Panthera and metadata access latency was 8 times lower. Such drastic decreases in latency can greatly increase the efficiency of existing applications and also allow the creation of algorithms that were previously infeasible on Hadoop. Panthera was tested with existing Hadoop algorithms and running times were up to 3.3 times lower than the control group of a standard Hadoop installation. In addition to the caching system, a scheduler and a scheduling algorithm for Hadoop were also developed. They use information available about the caches to decide the order of execution for computational jobs, thereby further reducing running time. Panthera is widely applicable and can significantly speed up research in a large range of fields ranging from bioinformatics to artificial intelligence. All code associated with this project will be open sourced to further development in the area.

Awards Won:

Third Award of $1,000