Measurements of data access

Team: Frank Wuerthwein, Diego Davila, Jonathan Guiang
Institutions: University of California in San Diego

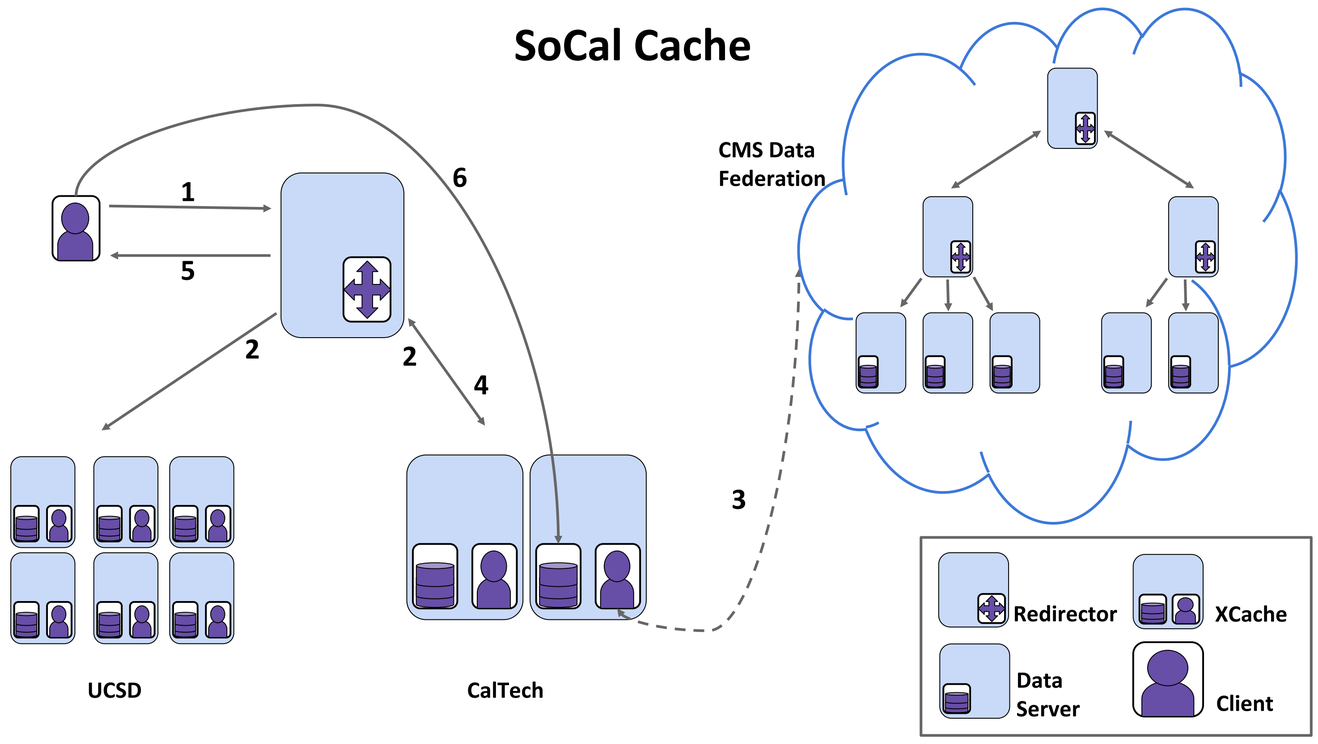
Overview

Significant portions of LHC analyses use the same data, running over each dataset several times. Hence, a cache-based approach can be utilized to to reduce the latency of the applications reading these data by keeping the most popular data closer to computing resources. The “SoCal” cache is the largest cache deployment in production for the LHC and is composed by a combined infrastructure in two sites in Southern California, UCSD and Caltech. The SoCal cache utilization was studied for the period of late March to mid August 2019. During this period, SoCal cached a fraction of the entire CMS “MINIAOD(SIM)” data tier, using approximately 600TB of disk space.

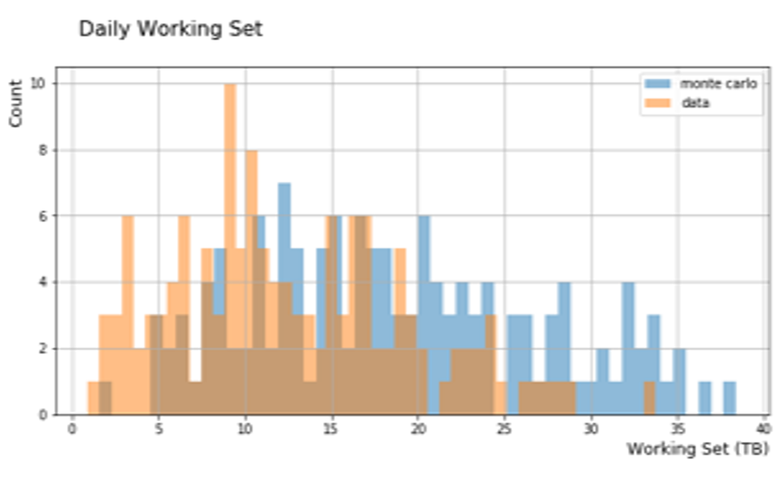
Data Analytics

The diagram on the right provides an overview of the structure of the SoCal cache and its data source, the CMS Data Federation.

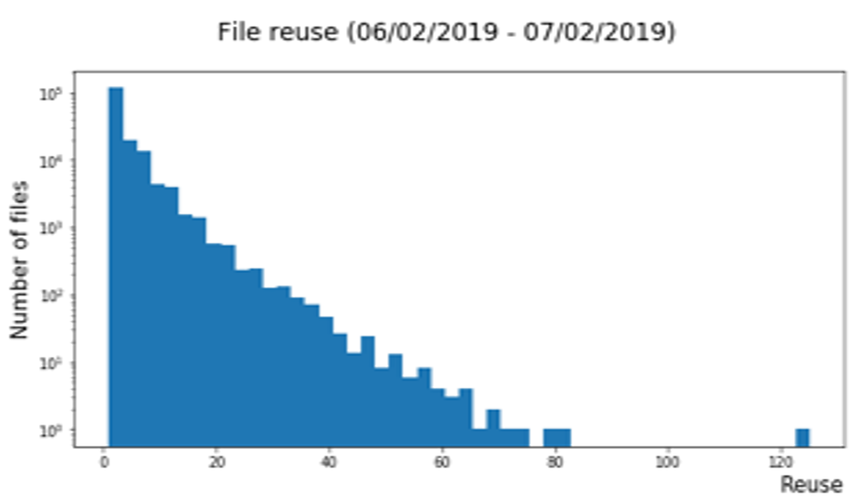
The SoCal Data Servers send information about the file accesses to the central monitoring infrastructure at CERN (“MonIT”), where it is stored in a *ElasticSearch* data base and ingested into *HDFS* for a long-term analytics. Using the the *Apache Spark* instance at CERN we have extracted this data and analyzed it using the *pandas* library to understand the usage of the cache.



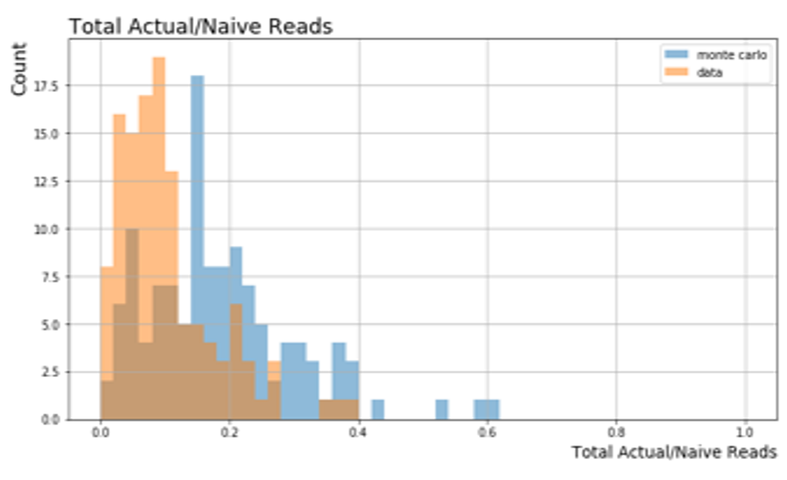
1. A file is requested to the cache, 2. The redirector asks for the file to its Data Servers, 3. If the file is not found it a request is made to the Data Federation, 4. One DataServer claims to have the file locally, 5. The redirector informs the client which DataServer to contact, 6. The client request the file to the correct Data Server.



The amount of unique data accessed daily is small (10-60TB) in comparison with the size of the full namespace available through the cache (1PB).



Most files are reused frequently, a characteristic of a cache-friendly workflow. This allows the cache to provide significant savings in WAN transfers.



This plot shows the ratio between the file size (naive read) and the portion of the file actually being read (actual read). Most jobs only read a small portion of the file (10-20%)

Continuous Monitoring

Making aggregations of this monitoring data over large periods of time is of great interest for the operators of these caches but designing queries over the detailed data consumes significant effort. *Monicron* is a new monitoring system designed to periodically calculate similar metrics to the ones above over defined intervals of time. Its ultimate goal is to provide a simplified view of the health and performance of the cache at a glance. The development of this system is complete and it is anticipated to be fully operational at the end of February, 2020

This project is supported by National Science Foundation under Cooperative Agreement OAC-1836650. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.