



Contribution ID: 216

Type: Oral

Data Locality via Coordinated Caching for Distributed Processing

Tuesday 19 January 2016 14:25 (25 minutes)

With the increasing data volumes of the second LHC run, analysis groups have to handle unprecedented amounts of data.

This puts many compute clusters relying on network based storage to their limit.

In contrast, data locality based processing enables infrastructure to scale out practically indefinitely.

However, data locality frameworks and infrastructure often add severe constraints and requirements.

To address this, we have developed an approach of adding coordinated caches to existing compute clusters.

Since the data stored locally is volatile and selected dynamically, only a fraction of local storage space is required.

Our approach allows to freely select the degree at which data locality is provided.

It may be used to work in conjunction with large network bandwidths, providing only highly used data to reduce peak loads.

Alternatively, local storage may be scaled up to perform data analysis even with low network bandwidth.

To prove the applicability of our approach, we have developed a prototype implementing all required functionality.

It integrates seamlessly into batch systems, requiring practically no adjustments by users.

We have now been actively using this prototype on a test cluster for HEP analyses.

Specifically, it has been integral to our jet energy calibration analyses for CMS during run 2.

The system has proven to be easily usable, while providing substantial performance improvements.

Since confirming the applicability for our use case, we have investigated the design in a more general way.

Simulations show that many infrastructure setups can benefit from our approach.

For example, it may enable us to dynamically provide data locality in opportunistic cloud resources.

The experience we have gained from our prototype enables us to realistically assess the feasibility for general production use.

Author: FISCHER, Max (KIT - Karlsruhe Institute of Technology (DE))

Co-authors: JUNG, Christopher; KUHN, Eileen (KIT - Karlsruhe Institute of Technology (DE)); GIFFELS, Manuel (KIT - Karlsruhe Institute of Technology (DE))

Presenter: FISCHER, Max (KIT - Karlsruhe Institute of Technology (DE))

Session Classification: Track 1

Track Classification: Computing Technology for Physics Research