Contribution ID: **391** Type: **Poster**

ATLAS Data Caching based on the Probability of Data Popularity

Tuesday 22 May 2012 13:30 (4h 45m)

Efficient distribution of physics data over ATLAS grid sites is one of the most important tasks for user data processing. ATLAS' initial static data distribution model over-replicated some unpopular data and under-replicated popular data, creating heavy disk space loads while under-utilizing some processing resources due to low data availability. Thus, a new data distribution mechanism was implemented, PD2P (PanDA Dynamic Data Placement) within the production and distributed analysis system PanDA that dynamically reacts to user data needs [1], basing dataset distribution principally on user demand. Data deletion is also demand driven, reducing replica counts for unpopular data [2]. This dynamic model has led to substantial improvements in efficient utilization of storage and processing resources.
Based on this experience, in this work we seek to further improve data placement policy by investigating in detail how data popularity is calculated. For this it is necessary to precisely define what data popularity means, what types of data popularity exist, how it can be measured, and most importantly, how the history of the data can help to predict the popularity of derived data. We introduce locality of the popularity: a dataset may be only of local interest to a subset of clouds/sites or may have a wide (global) interest. We also extend the idea of the "data temperature scale"model [3] and a popularity measure.
Using the ATLAS data replication history, we devise data distribution algorithms based on popularity measures and past history. Based on this work we will describe how to explicitly identify why and how datasets become popular and how such information can be used to predict future popularity.

[1] Kaushik De, Tadashi Maeno, Torre Wenaus, Alexei Klimentov, Rodney Walker, Graeme Stewart, "PD2P – PanDA Dynamic Data Placement", ATLAS Notes, CERN
[2] Angelos Molfetas, Fernando Barreiro Megino, Andrii Tykhonov, Vincent Garonne, Simone Campana, Mario Lassnig, Martin Barisits, Gancho Dimitrov, Florbela Tique Aires Viegas, "Popularity framework to process dataset tracers and its application on dynamic replica reduction in the ATLAS experiment", CHEP, Taipei, Taiwan, October 18-22, 2010
[3] Alexei Klimentov, "ATLAS data over Grid (data replication, placement and deletion policy)", ATLAS Notes, CERN, March 17, 2009

Student? Enter 'yes'. See http://goo.gl/MVv53

yes

Authors: Dr KLIMENTOV, Alexei (Brookhaven National Laboratory (US)); Dr ZARUBA, Gergely (University of Texas at Arlington (US)); Dr DE, Kaushik (University of Texas at Arlington (US)); TITOV, Mikhail (University of Texas at Arlington (US))

Presenter: TITOV, Mikhail (University of Texas at Arlington (US))

Session Classification: Poster Session

Track Classification: Distributed Processing and Analysis on Grids and Clouds (track 3)