



Contribution ID: 20

Type: poster presentation

Extending DIRAC File Management with Erasure-Coding for efficient storage.

The state of the art in Grid style data management is to achieve increased resilience of data via multiple complete replicas of data files across multiple storage endpoints. While this is effective, it is not the most space-efficient approach to resilience, especially when the reliability of individual storage endpoints is sufficiently high that only a few will be inactive at any point in time.

We extended the Dirac File Catalogue and file management interface to allow the placement of *erasure-coded* files: each file distributed as N identically-sized chunks of data striped across a vector of storage endpoints, encoded such that any M chunks can be lost and the original file can be reconstructed.

The tools developed are transparent to the user, and, as well as allowing up and downloading of data to Grid storage, also provide the possibility of parallelizing access across all of the distributed chunks at once, improving data transfer and IO performance.

We expect this approach to be of most interest to smaller VOs, who have tighter bounds on the storage available to them, but larger VOs may be interested as their total data increases during Run 2.

With this in mind, we tested the applicability of our tools to the NA62 analysis model, which already expects data to be distributed across multiple Tier-2 sites.

We provide an analysis of the costs and benefits of the approach, along with future development and implementation plans.

Primary author: Dr SKIPSEY, Samuel Cadellin

Co-authors: PROTOPOPESCU, Dan (University of Glasgow (GB)); Prof. BRITTON, David (University of Glasgow (GB)); CROOKS, David (University of Glasgow (GB)); ROY, Gareth Douglas (University of Glasgow (GB)); MARTYNIK, Janusz (Imperial College London); Mr TODEV, Paulin (University of Glasgow)

Presenter: Dr SKIPSEY, Samuel Cadellin

Track Classification: Track3: Data store and access