

Addressing a billion-entries multi-petabyte distributed filesystem backup problem with *cback*: from files to objects

Wednesday 19 May 2021 11:55 (13 minutes)

CERNBox is the cloud collaboration hub at CERN. The service has more than 37,000 user accounts. The backup of user and project data is critical for the service. The underlying storage system hosts over a billion files which amount to 12PB of storage distributed over several hundred disks with a two-replica RAIN layout. Performing a backup operation over this vast amount of data is a non-trivial task.

The original CERNBox backup system (an in-house event-driven file-level system) has been reconsidered and replaced by a new distributed and scalable backup infrastructure based on the open source tool *restic*. The new system, codenamed *cback*, provides features needed in the HEP community to guarantee data safety and smooth operation from the system administrators. Daily snapshot-based backups of all our user and project areas along with automatic verification and restores are possible with this the new development.

The backup data is also de-duplicated in blocks and stored as objects in a disk-based S3 cluster in another geographical location on the CERN campus, reducing storage costs and protecting critical data from major catastrophic events. We report on the design and operational experience of running the system and future improvement possibilities.

Primary authors: VALVERDE CAMESELLE, Roberto (CERN); GONZALEZ LABRADOR, Hugo (CERN)

Presenter: VALVERDE CAMESELLE, Roberto (CERN)

Session Classification: Storage

Track Classification: Distributed Computing, Data Management and Facilities