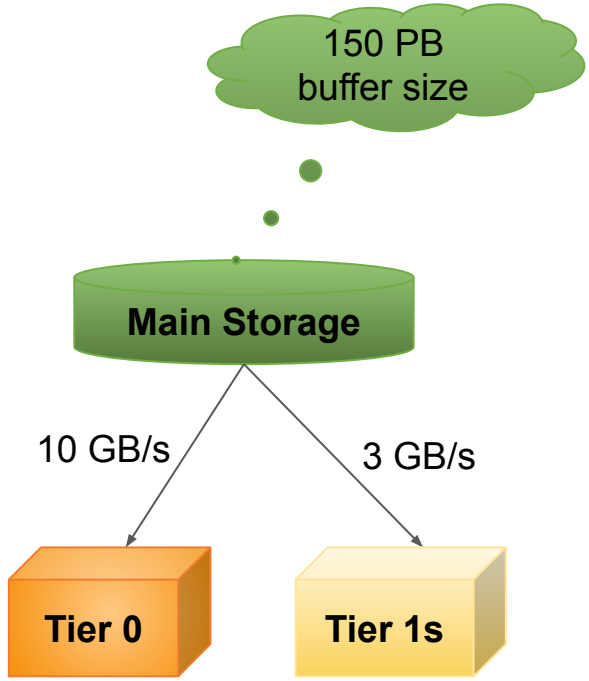
Disk management tools

Alice Suiu - asuiu@cern.ch

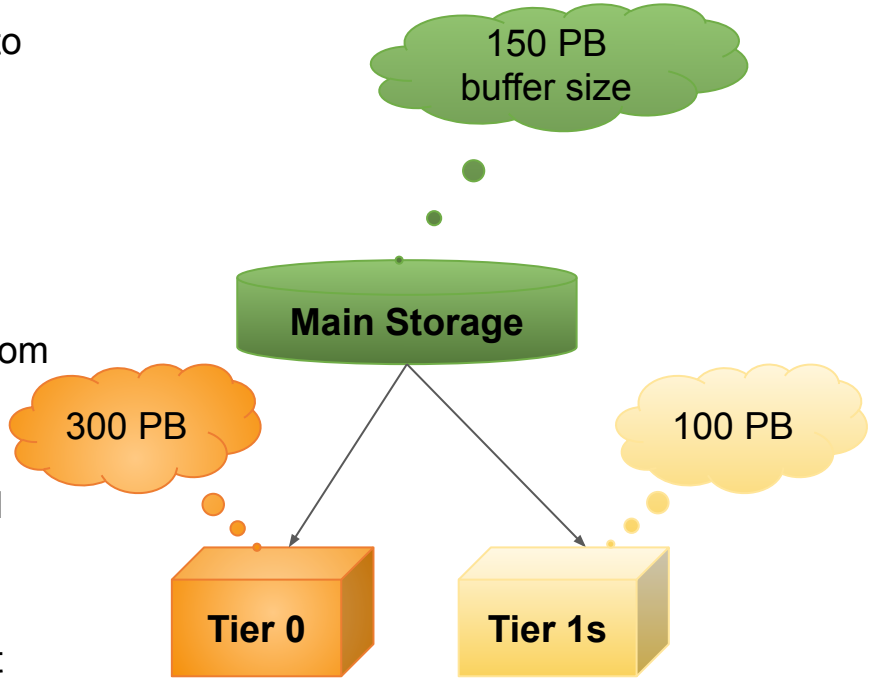
ALICE storage elements

- **Tier 0**
  - Located at CERN
  - Offers storage, CPU, and archiving capabilities
  - A total capacity of approximately **300 PB** (disk and tapes)
- **Tier 1**
  - Mass storage capabilities, and I/O optimized for communications with Tier 0
  - A total capacity of approximately **100 PB** (disk and tapes)
- Ensure long-term storage of the data collected from the experiment as well as the results obtained from the physics analysis

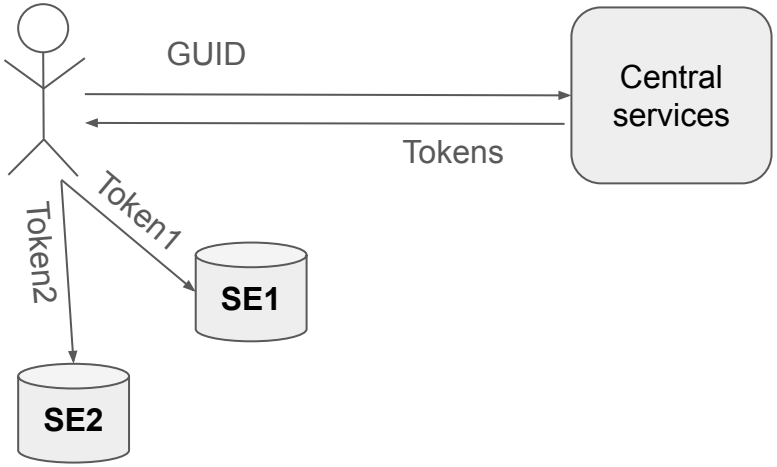


Run actions

- **Move** — means that the data files of a run are moved to another storage element
- **Copy** — implies creating a replica of a run on another storage element
- **Delete replica** — represents erasing a copy of a run from a certain storage element
- **Delete** — means erasing the data files of a run from all storage elements
- **Reconstruction** — means reproducing the events that happened within the experiment based on a configuration and the results obtained after the physical experiment

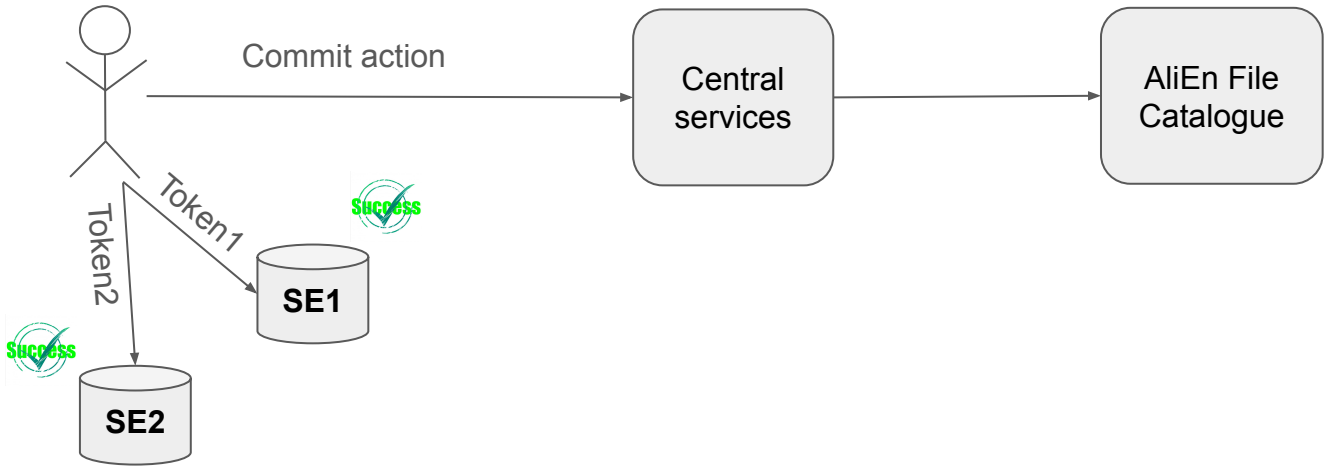


Job writing



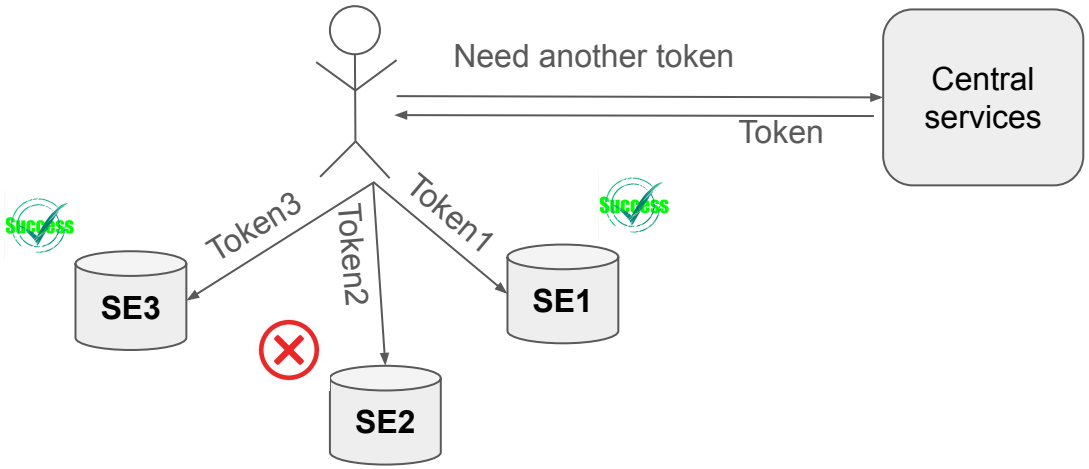
- A job attempts to write a file on two storage elements
- The job connects to the central services, sending a GUID for the file it wants to write
- The central services send a list of two tokens associated with two storage elements
- The job receives the tokens which it uses to identify the storage elements where it will write the file, authenticate, and start the writing process

Successful writing



- If the file writing on both storage elements is successful, the central services are notified, resulting in the file being registered in the catalogue
- The file then has an LFN pointing to two PFNs, since it was written on two storage elements

Failed writing



- If one of the two file writes fails, a request is made to write to another storage element
- The central services send a new token to write to a different storage element

- The file whose write failed on the second storage element is marked as invalid and placed in a queue for deletion
- If the job fails in the meantime, after 24 hours, the files that were attempted to be written are considered invalid and also placed in the queue for deletion

Dark-data removal

- Dark data is considered to be any write attempt by a job that did not complete successfully, meaning it did not result in the files being registered in the catalogue
- Any failed write is marked as invalid, with the file being placed in a queue for deletion
- XrootdCleanupSingle <SE> [1]:
  - Performs a recursive list on the <SE>
  - Retrieves a list of GUIDs associated with the files found on the <SE>
  - Checks if there is an LFN associated with each GUID in the catalogue and if there is a registered copy of it on the <SE>
  - If any of the previously mentioned information is missing from the catalogue, each GUID found is marked as invalid, and the associated file is placed in the queue for deletion

File crawler

- Analyzes data integrity and the quality of a storage element
- Scans periodically a random selection of files from each storage element
- Collects statistics: number of inaccessible or corrupted files, throughput, and download latency on each storage element
- The file selection is processed by jobs running on computational nodes associated with each site
- The number of files in the selection is proportional to the capacity of each storage element
- Each job receives a subset of files from the selection, downloads them, computes the MD5 checksum for each file, and compares it with the value stored in the catalogue
- Each job also compares the downloaded file's size with the size reported in the catalogue
- If either of these values differs, the file is marked as corrupted
- See [2] for more details

Data recovery tool

- Removes the damaged entry from the catalogue to prevent users or jobs from accessing lost or invalid data
- If the catalogue indicates that replicas of that file are available on other storage elements, the tool creates a new replica from the existing ones and updates the catalogue with a corresponding entry
- If there is no replica for the corrupted or invalid file, it is placed in a queue for deletion
- See [3] for more details

Storage element replication

- Scan recursively the storage element targeted for replication
- Obtain a list of all files present on the storage element
- Copy each file to the designated storage element
- Replication constraint:
  - If a file has fewer than 3 replicas (default value), it is also copied to an additional secondary storage element
- Three storage elements:
  - Source storage element where the data is being copied from
  - Destination storage element where the data is being transferred to
  - Secondary storage element where an extra copy of files that do not meet the replication criteria is created