EOS+CTA Workflows :

Tape Archival and Retrieval

Michael Davis, Vladimír Bahyl, Cédric Caffy, Eric Cano, David Fernandez Alvarez, Aurelien Gounon, Oliver Keeble, Julien Leduc, Steven Murray, Volodymyr Yurchenko

The archival storage solution from the CERN IT Storage Group

CERN

Tape Archive

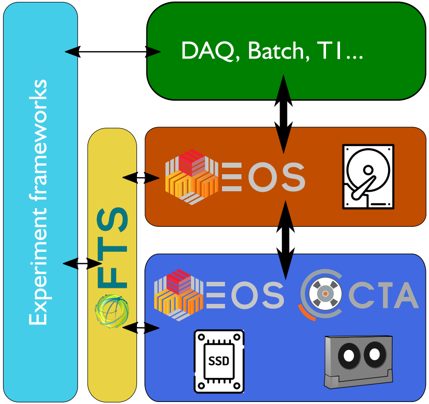
The archival storage solution from the CERN IT Storage Group

EOS + CTA

CTA is the tape back-end to EOS

The archival storage solution from the CERN IT Storage Group

CERN Tier-0: "Big EOS" and "Little EOS"



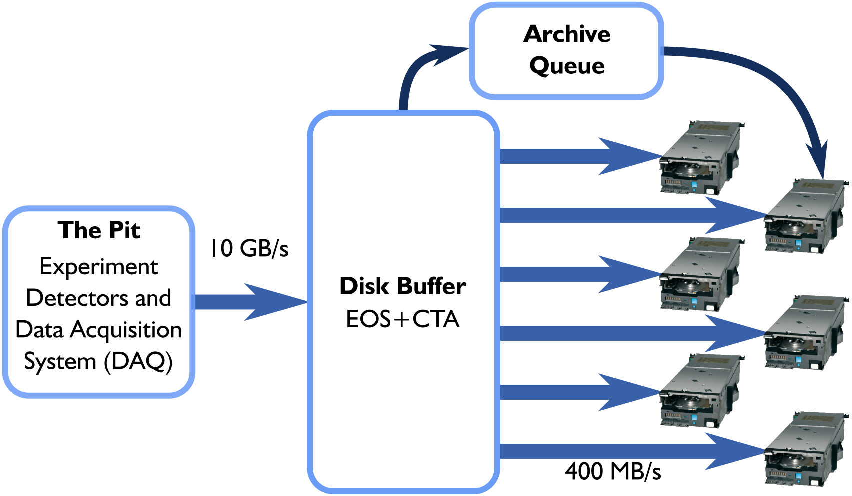
"Big EOS"

- Tens of PB of storage for physics reconstruction and analysis jobs.
- No tape on the back.

"Little EOS"

- Small, fast buffer with tape on the back.
- Copying a file into a tape-backed directory triggers an archival request.
- Files are recalled from tape with two-step *stage-and-transfer* semantics.

Archive Workflow



Archive Queue

The Pit

Experiment Detectors and Data Acquisition System (DAQ)

10 GB/s

Disk Buffer

EOS+CTA

400 MB/s

EOS Configuration

Set an extended attribute on tape-backed directories:

```
sys.archive.storage_class="single"
```

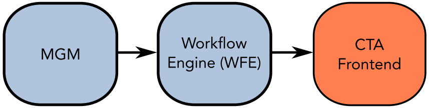
- **Storage Class**

  Specifies how many copies of the file should be archived to tape and the **Tape Pool** for each copy

- **Tape Pool**

  Logical collection of tapes used to manage (a) file ownership; (b) where the file should be physically stored

EOS WorkFlow Engine: CREATE Event



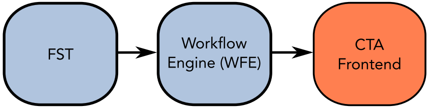
- Validate **Storage Class**
- Allocate **Archive ID**

  Unique, monotonic number allocated by CTA Frontend. Stored as an extended attribute of the file:

  ```
  sys.archive.file_id="4294967296"
  ```

- EOS WFE events are synchronous. If CTA raises an error, EOS deletes the file and reports the error to the client.
- CTA Frontend will reject operations on files which do not have a valid Archive ID.

EOS WorkFlow Engine: CLOSEW Event



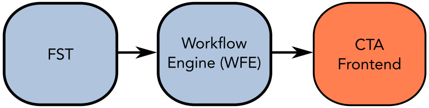
- Create an **Archive Request** for the file.
  The Archive Request ID is stored as an extended attribute of the file:

```
sys.cta.archive.objectstore.id="ArchiveRequest-
  Frontend-ctatest.cern.ch-14148-20200518-14:16:34-0-0"
```

- Once the item is queued, archival proceeds asynchronously:

```
d1::t0  -rw-r----- myuser mygroup 1TB my-important-file
```

EOS WorkFlow Engine: CLOSEW Event



- Create an **Archive Request** for the file.

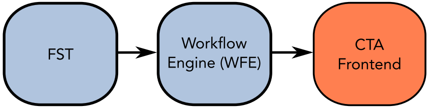
  The Archive Request ID is stored as an extended attribute of the file:

```
sys.cta.archive.objectstore.id="ArchiveRequest-
  Frontend-ctatest.cern.ch-14148-20200518-14:16:34-0-0"
```

- Once the item is queued, archival proceeds asynchronously:

```
d1::t1  -rw-r----- myuser mygroup 1TB my-important-file
```

EOS WorkFlow Engine: CLOSEW Event



- Create an **Archive Request** for the file.

  The Archive Request ID is stored as an extended attribute of the file:

```
sys.cta.archive.objectstore.id="ArchiveRequest-
  Frontend-ctatest.cern.ch-14148-20200518-14:16:34-0-0"
```

- Once the item is queued, archival proceeds asynchronously:

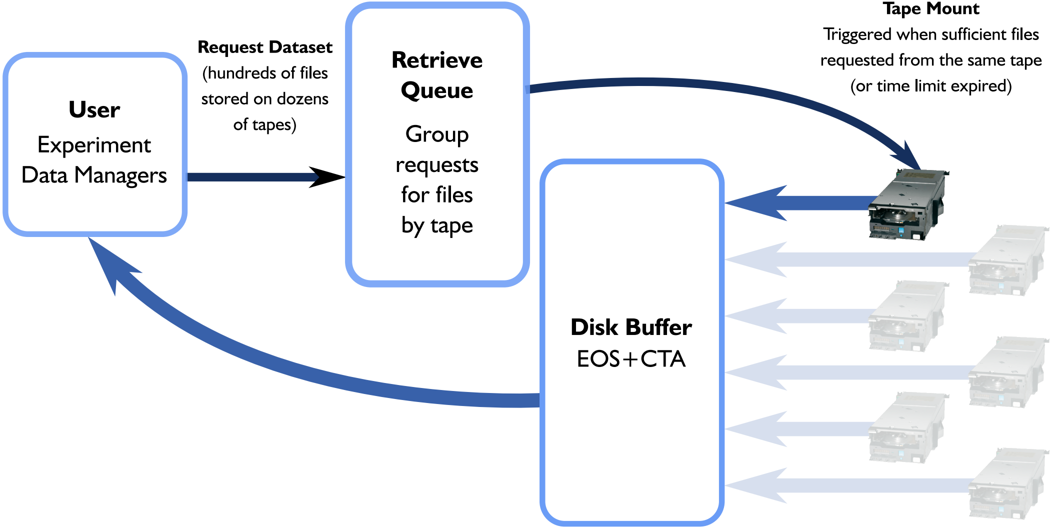
```
d0::t1  -rw-r----- myuser mygroup 1TB my-important-file
```

Other Archival Events

- **archived** event (on success) removes the disk replica from the buffer.
- **archive_failed** event sets `sys.archive.error`

- CTA does not handle **OPENW** events, because files on tape are immutable.
  - Enforced by adding immutable flag (`!u`) to the ACL of the tape-backed directories in EOS, or as a rule.

Retrieve Workflow



User

Experiment Data Managers

Request Dataset (hundreds of files stored on dozens of tapes)

Retrieve Queue

Group requests for files by tape

Tape Mount

Triggered when sufficient files requested from the same tape (or time limit expired)

Disk Buffer

EOS+CTA

EOS WorkFlow Engine: Retrieve Events

- PREPARE

  Recall (stage) a file from tape to the EOS disk buffer

- QUERY_PREPARE

  Query the status of a file:

  disk residency/tape residency/request status/error messages

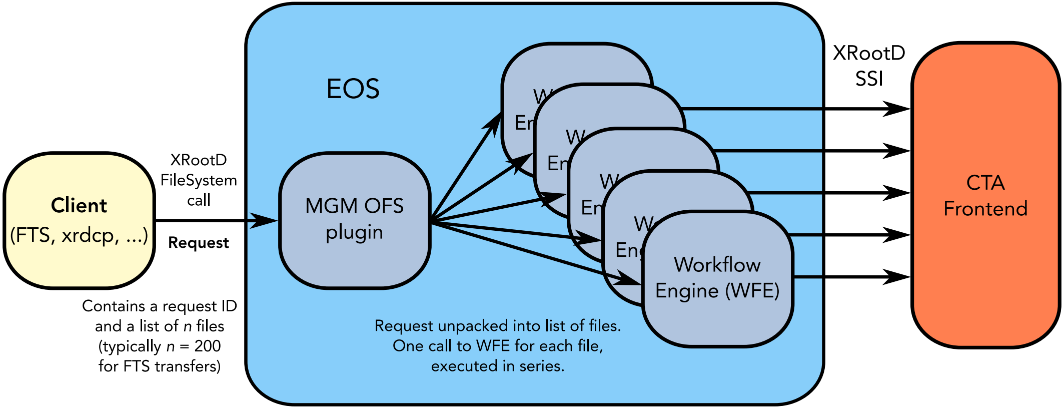
- ABORT_PREPARE

  Cancel a previous PREPARE request

- EVICT_PREPARE

  Remove the disk replica of a previously-retrieved file from the EOS disk buffer

EOS WorkFlow Engine: PREPARE Event



EOS WorkFlow Engine: PREPARE Event

```
$ xrdfs prepare -s /path/to/file1 /path/to/file2 ...
0446898a2076:cd20c08c.60070221:14:1612351205
```

- Provide a list of files and get back a Prepare Request ID
- The Request ID is stored as an extended attribute in every file in the request
- The **PREPARE** request will always succeed; error messages are also stored as extended attributes on each file

QUERY_PREPARE Event

```
$ xrdfs query prepare 0446898a2076:cd20c08c.60070221:14:1612351205
    /path/to/file1 /path/to/file2 ...

  "responses": [
    {
      "path": "/path/to/file1",
      "path_exists": true,
      "on_tape": true,
      "online": false,
      "requested": true,
      "has_reqid": true,
      "req_time": "1612356259",
      "error_text": ""
    },
    ...
  ]
```

EOS WorkFlow Engine: CLOSEW.retrieve_written Event

- Executed when file has been successfully recalled to disk
- Clears the extended attributes:

```
sys.retrieve.req_id
sys.retrieve.req_time
sys.retrieve.error
```

- `xrdfs query prepare` now returns:

```
"online": true
```

EOS WorkFlow Engine: retrieve_failed Event

CTA may be unable to retrieve the file due to:

- an error reading the tape
- an error writing to the disk buffer

CTA will retry three times per mount session for two separate tape mounts. If the file still cannot be recalled:

- Record the error in `sys.retrieve.error`
- Clear the list of pending retrieve requests in `sys.retrieve.req_id`
- Clear `sys.retrieve.req_time`

EOS + CERN Tape Archive : Workflows

Overview of the two primary EOS+CTA workflows:

- **ARCHIVE** and **PREPARE** WorkFlow events

Some topics not covered:

- The **DELETE** Event, including aborting in-flight archival requests
- The **ABORT_PREPARE** Event, to cancel in-flight retrieval requests
- The **EVICT_PREPARE** Event to remove the disk replica of a file on tape from the EOS disk buffer
- The “File Is Safely On Tape” handshake (used by FTS Archive Monitoring feature, see CTA Best Practices for Data Taking Workflows)
- Garbage Collection (see ALICE and the CTA Garbage Collectors)

CERN

home.cern