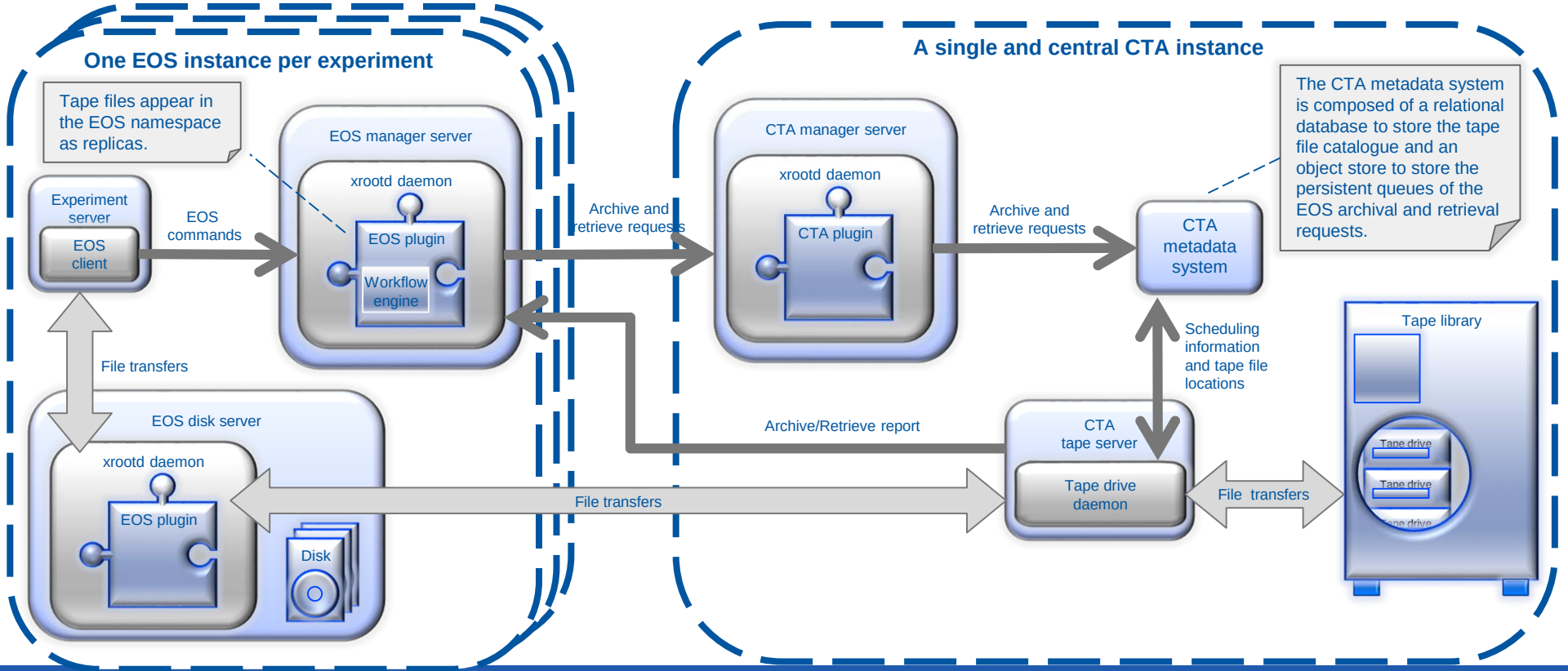
CTA: the tape backend of EOS

EOS Workshop

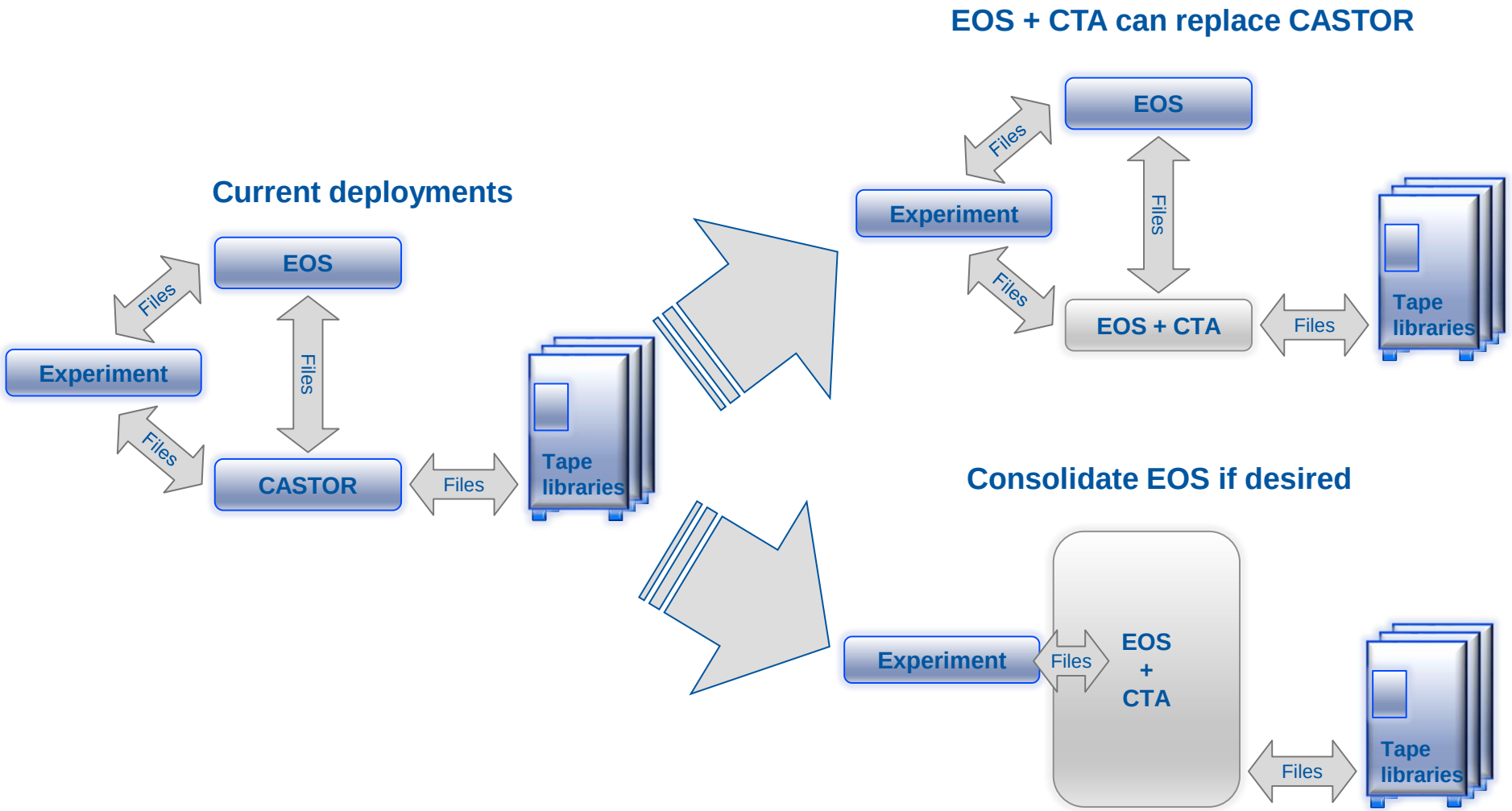
Vladimir Bahyl, Germán Cancio, Eric Cano, Julien Leduc, Steven Murray and Victor Kotlyar

EOS+CTA architecture

- CTA is integrated with EOS: all user interaction via EOS
- CTA tape files appear in the EOS namespace as file replicas
- CTA contains an internal flat catalogue of all tape files



EOS+CTA possible deployments



EOS+CTA File life cycle (user)

- Classic life cycle a seen in CASTOR
- Implicit archive to tape
  - Configured at directory level
- Files are immutable
- Explicit and implicit disk replica deletion
  - Implicit via garbage collection (internal to EOS)
- Explicit and implicit retrieve from tape
  - Implicit via blocking open
- Import from CASTOR
  - No physical data movement
  - Namespace metadata import
  - Transfer of tape ownership

EOS+CTA File life cycle (admin)

- Under the hood operations
- Verification (internal to CTA)
- Repack (internal to CTA)
- Reconciliation
  - Metadata consistency validation between EOS and CTA
- Disaster recovery
  - Re-injects metadata in EOS from CTA

File write lifecycle

- Pre-requisite: workflow assign to top-directory (by EOS admins)
- User writes file to EOS as usual
- EOS queues the archival to CTA
- At completion, the tape replica appears in EOS

```
$ eos attr ls /eos/ctaeos/cta/myFile
sys.archiveFileId="2"

$ eos info /eos/ctaeos/cta/myFile
  File: '/eos/ctaeos/cta/myFile'  Flags: 0644
  Size: 406
Modify: Tue Dec 13 11:51:59 2016 Timestamp: 1481626319.474548977
Change: Tue Dec 13 11:51:52 2016 Timestamp: 1481626312.169566904
  CUid: 3 CGid: 4  Fxid: 00000007 Fid: 7    Pid: 16   Pxid: 00000010
XStype: adler    XS: d9 1d 7f 57     ETAG: 1879048192:d91d7f57
replica Stripes: 1 Blocksize: 4k LayoutId: 00100012
  #Rep: 2
 #   fs-id  #.....................................................................................
            #                  host  #     schedgroup #           path #     boot # configstatus #      drain # acti
            #.....................................................................................
  0       1 ...est.svc.cluster.local         default.0             /fst     booted             rw      nodrain   onli
  1   65535                localhost              tape  /does_not_exist                       off      nodrain
```

Disk replica removal

- The disk replica gets deleted by the user

```
$ eos file tag /eos/ctaeos/cta/myFile -1

$ eos attr ls /eos/ctaeos/cta/myFile
sys.archiveFileId="2"

$ eos info /eos/ctaeos/cta/myFile
 File: '/eos/ctaeos/cta/myFile'  Flags: 0644
  Size: 406
Modify: Tue Dec 13 11:51:59 2016 Timestamp: 1481626319.474548977
Change: Tue Dec 13 11:51:52 2016 Timestamp: 1481626312.169566904
  CUid: 3 CGid: 4  Fxid: 00000007 Fid: 7    Pid: 16   Pxid: 00000010
XStype: adler    XS: d9 1d 7f 57     ETAG: 1879048192:d91d7f57
replica Stripes: 1 Blocksize: 4k LayoutId: 00100012
  #Rep: 1
 #   fs-id  #.....................................................................................
            #                  host  #     schedgroup #           path #     boot # configstatus #      drain # activ
            #.....................................................................................
  0   65535                localhost              tape  /does_not_exist                      off      nodrain
```

Retrieve from tape

- The retrieve is triggered by the user

```
xrdfs localhost prepare -s /eos/ctaeos/cta/myFile
```

- The disk replica appears at retrieve completion

```
$ eos attr ls /eos/ctaeos/cta/myFile
CTA_retrieved_timestamp="Tue Dec 13 11:52:29 CET 2016"
sys.archiveFileId="2"

$ eos info /eos/ctaeos/cta/myFile
 File: '/eos/ctaeos/cta/myFile'  Flags: 0644
  Size: 406
Modify: Tue Dec 13 11:52:29 2016 Timestamp: 1481626349.446767653
Change: Tue Dec 13 11:51:52 2016 Timestamp: 1481626312.169566904
  CUid: 3 CGid: 4  Fxid: 00000007 Fid: 7    Pid: 16   Pxid: 00000010
XStype: adler    XS: d9 1d 7f 57     ETAG: 1879048192:d91d7f57
replica Stripes: 1 Blocksize: 4k LayoutId: 00100012
  #Rep: 2
 #   fs-id  #.....................................................................................................
            #                  host  #     schedgroup #           path #     boot # configstatus #      drain # activ
            #.....................................................................................................
  0   65535                localhost             tape  /does_not_exist                      off      nodrain
  1       1 ...est.svc.cluster.local        default.0            /fst     booted             rw      nodrain   onlin
```

Notable new features

- New tape drive scheduling
  - One step with access to all info vs partial steps in CASTOR
  - Pre-emptible, allowing better utilization (fill any gap with repack/verification)
  - Other scheduling improvements looked at
- Additional drive level optimization
  - RAO (Recommended access order = drive assisted read order optimization)
  - Flush optimization
  - Optimized write ordering?

The 2 year development plan