Container-based CTA installation for storing midsize archive data

Victor Kotlyar, Denis Syrko, Dmitry Bolotin

Outline

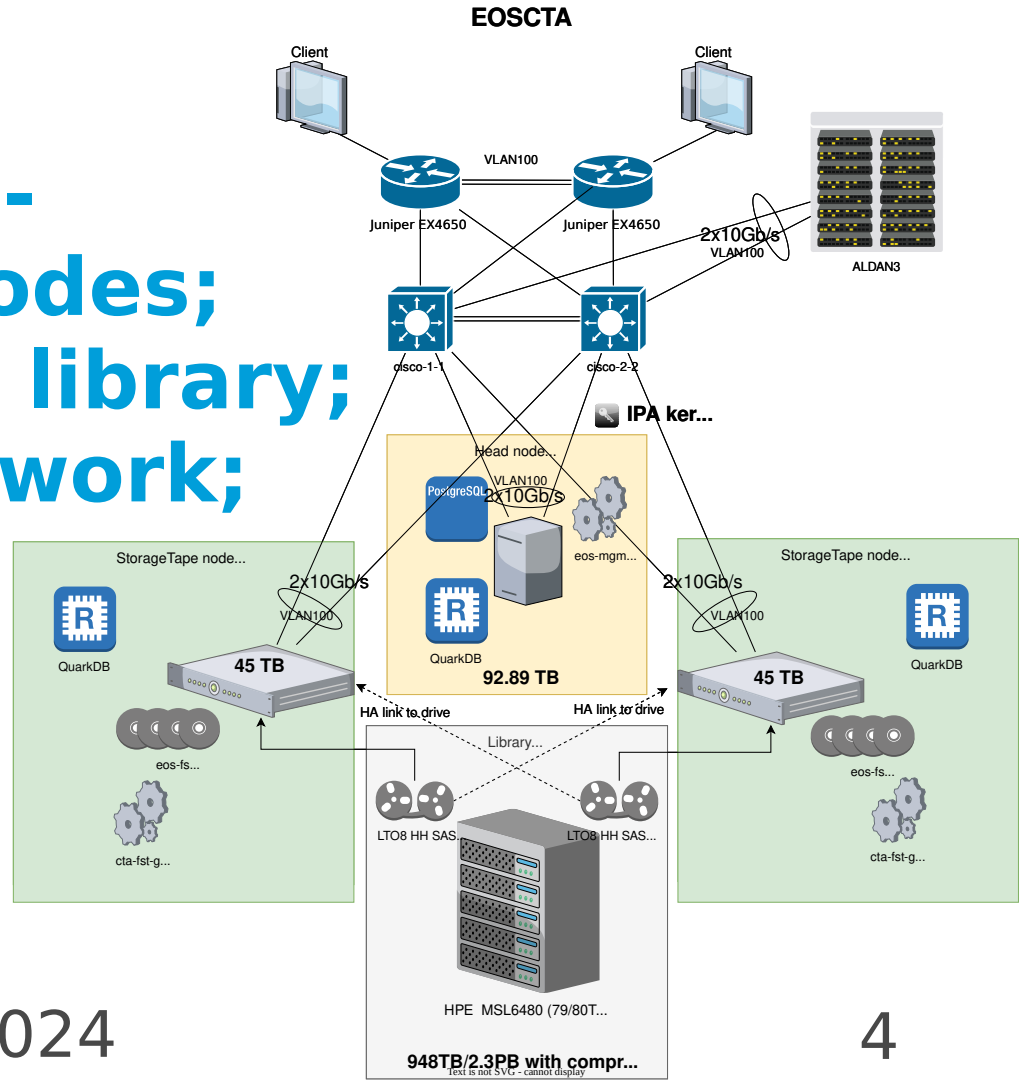
- Generic goal
- System architecture overview
- Hardware overview
- Software overview
- Installation tricks
- System usage
- Need to be done

System requirements:

- **As small and effective as possible;**
- **For store archive data;**
- **For a backup service;**
- **Seamlessly integrated with an existing computing cluster for group based access pattern;**
- **Without any limits in size and features if needed in the future.**

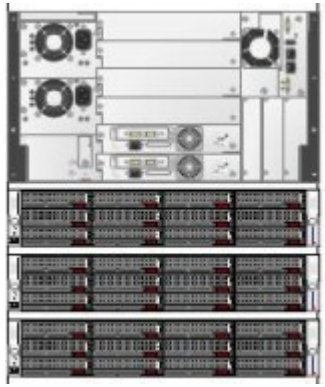
System architecture:

- Three nodes cluster - Head + tape&disk nodes;
- Modular expandable library;
- Generic purpose network;
- External IPA for AA;
- Cluster and clients outside the system.



Hardware overview (12U):

- **Main storage -**
  HPE MSL6480 library with one module;
  79 LTO8 tapes + 1 cleaning;
  940TB (with compression in mind);
  two LTO8 SAS3 HH drives;
- **Disk storage -**
  6x12TB SATAIII per server 87.2TiB
  disks ~ 10% of tapes
- **All nodes are the same: 2x4216(64Th) 96GB**



Hardware overview from inside:

```
ansible@tape-1-3-1:~$ lsscsi -g
[4:0:0:0]    disk    ATA      CT480BX500SSD1   054   /dev/sda   /dev/sg0
[5:0:0:0]    disk    ATA      CT480BX500SSD1   054   /dev/sdb   /dev/sg1
[6:0:0:0]    disk    ATA      TOSHIBA MG07ACA1 0103  /dev/sdc   /dev/sg2
[6:0:1:0]    disk    ATA      TOSHIBA MG07ACA1 0103  /dev/sdd   /dev/sg3
[6:0:2:0]    disk    ATA      TOSHIBA MG07ACA1 0103  /dev/sde   /dev/sg4
[6:0:3:0]    disk    ATA      TOSHIBA MG07ACA1 0103  /dev/sdf   /dev/sg5
[6:0:4:0]    disk    ATA      TOSHIBA MG07ACA1 0103  /dev/sdg   /dev/sg6
[6:0:5:0]    disk    ATA      TOSHIBA MG07ACA1 0103  /dev/sdh   /dev/sg7
[6:0:6:0]    enclosu SMC      SC826N4          100d  -          /dev/sg8
[15:0:0:0]   tape    HPE      Ultrium 8-SCSI   M571  /dev/st0   /dev/sg9
[15:0:1:0]   tape    HPE      Ultrium 8-SCSI   M571  /dev/st1   /dev/sg10
[15:0:1:1]   mediumx HP       MSL6480          6.30  /dev/sch0  /dev/sg11
[15:0:2:0]   enclosu BROADCOM VirtualSES       03    -          /dev/sg12
```

```
ansible@tape-1-3-1:~$ zpool list -vvv
NAME                        SIZE  ALLOC   FREE  CKPOINT  EXPANDSZ   FRAG    CAP  DEDUP    HEALTH  ALTROOT
data                       65.5T  59.3T  6.14T        -         -     1%    90%  1.00x    ONLINE  -
  raidz2                   65.5T  59.3T  6.14T        -         -     1%  90.6%      -  ONLINE
    scsi-35000039b08e2be07     -      -      -        -         -      -      -      -  ONLINE
    scsi-35000039b08e2c970     -      -      -        -         -      -      -      -  ONLINE
    scsi-35000039b08e2c969     -      -      -        -         -      -      -      -  ONLINE
    scsi-35000039b08e2bc8b     -      -      -        -         -      -      -      -  ONLINE
    scsi-35000039b08e2be57     -      -      -        -         -      -      -      -  ONLINE
    scsi-35000039b08e2be24     -      -      -        -         -      -      -      -  ONLINE
```

Software overview:

- **Base operating system Ubuntu 20.04LTS standart setup;**
- **Separate docker container (CentOS7) for each service -**

  head node: eos-mgm, eos-mq, cta-frontend, eos-qdb, PostgreSQL, nfsd;

  tape&disk node: eos-fst, cta-gcd, cta-taped, eos-qdb, cta-rmcd (on master)
- **eos-fst on zfs z2 pool**

Software overview from inside:

```
ansible@tape-1-3-3:~$ sudo docker ps
CONTAINER ID   IMAGE                          COMMAND                  CREATED         STATUS        PORTS                                                 NAMES
cb560b8fbcc6   eosmgm:latest                  "/opt/run/bin/eos-mg…"   21 months ago   Up 2 months                                                         eos-mgm-1-3-3
1f8360b6a353   eosmq:latest                   "/opt/run/bin/eos-mq…"   21 months ago   Up 7 months   0.0.0.0:1097->1097/tcp, :::1097->1097/tcp             eos-mq-1-3-3
b95b2e9b5256   ctafrontend:latest             "/opt/run/bin/cta-fr…"   21 months ago   Up 5 days     0.0.0.0:10955->10955/tcp, :::10955->10955/tcp         cta-frontend-1-3-3
31700c4c0caf   quarkdb:latest                 "runuser --user xroo…"   21 months ago   Up 7 months   0.0.0.0:7777->7777/tcp, :::7777->7777/tcp             eos-qdb-1-3-3
1a4c0ad4e0cb   centos/postgresql-13-centos7   "container-entrypoin…"   21 months ago   Up 7 months   0.0.0.0:5432->5432/tcp, :::5432->5432/tcp             postgresql
```

```
ansible@tape-1-3-1:~$ sudo docker ps
CONTAINER ID   IMAGE             COMMAND                  CREATED         STATUS        PORTS                                                 NAMES
ed19748785c2   ctagcd:latest     "/opt/run/bin/cta-fs…"   17 months ago   Up 7 months                                                         cta-gcd-1-3-1
92b420ef45e1   ctarmcd:latest    "/opt/run/bin/cta-rm…"   18 months ago   Up 7 months                                                         cta-rmcd-1-3-1
247b23d892fc   fe0216285abd      "/opt/run/bin/cta-ta…"   20 months ago   Up 7 months                                                         cta-taped-1-3-1
3bea991048c2   eosfst:latest     "/opt/run/bin/eos-fs…"   21 months ago   Up 7 months                                                         eos-fst-1-3-1
601eddeb7f59   quarkdb:latest    "runuser --user xroo…"   21 months ago   Up 7 months   0.0.0.0:7777->7777/tcp, :::7777->7777/tcp             eos-qdb-1-3-1
```

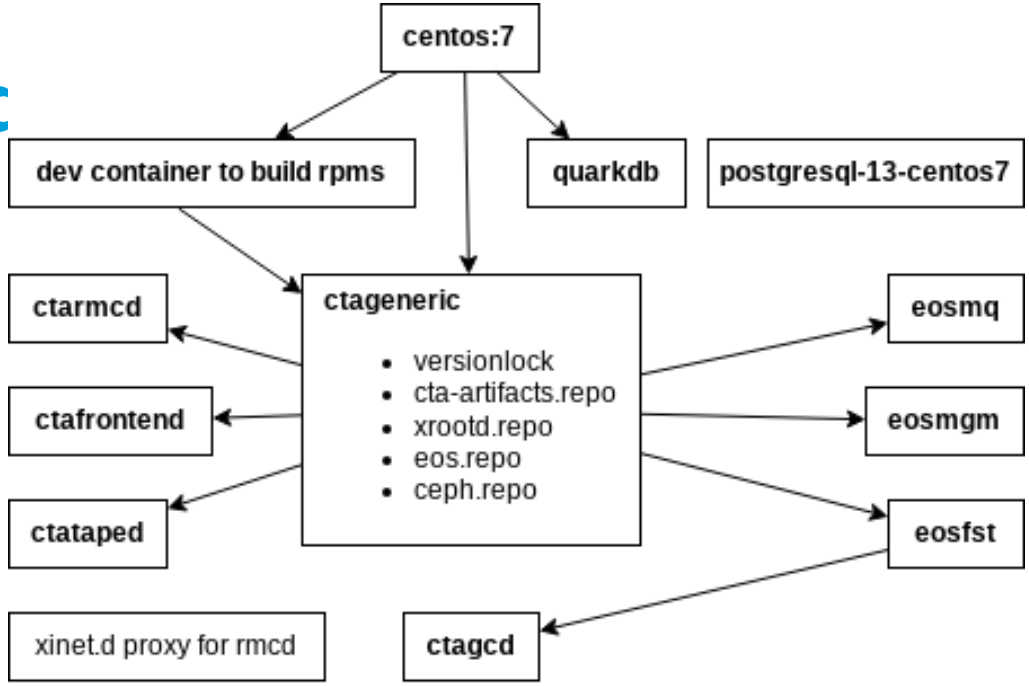
```
ansible@tape-1-3-2:~$ sudo docker ps
CONTAINER ID   IMAGE             COMMAND                  CREATED         STATUS        PORTS                                                 NAMES
2b0f8b7b5208   ctagcd:latest     "/opt/run/bin/cta-fs…"   17 months ago   Up 7 months                                                         cta-gcd-1-3-2
f047470da4c1   ctataped:latest   "/opt/run/bin/cta-ta…"   20 months ago   Up 2 days                                                           cta-taped-1-3-2
4865672bda87   eosfst:latest     "/opt/run/bin/eos-fs…"   20 months ago   Up 7 months                                                         eos-fst-1-3-2
1056cb893330   quarkdb:latest    "runuser --user xroo…"   21 months ago   Up 7 months   0.0.0.0:7777->7777/tcp, :::7777->7777/tcp             eos-qdb-1-3-2
```

Installations tricks:

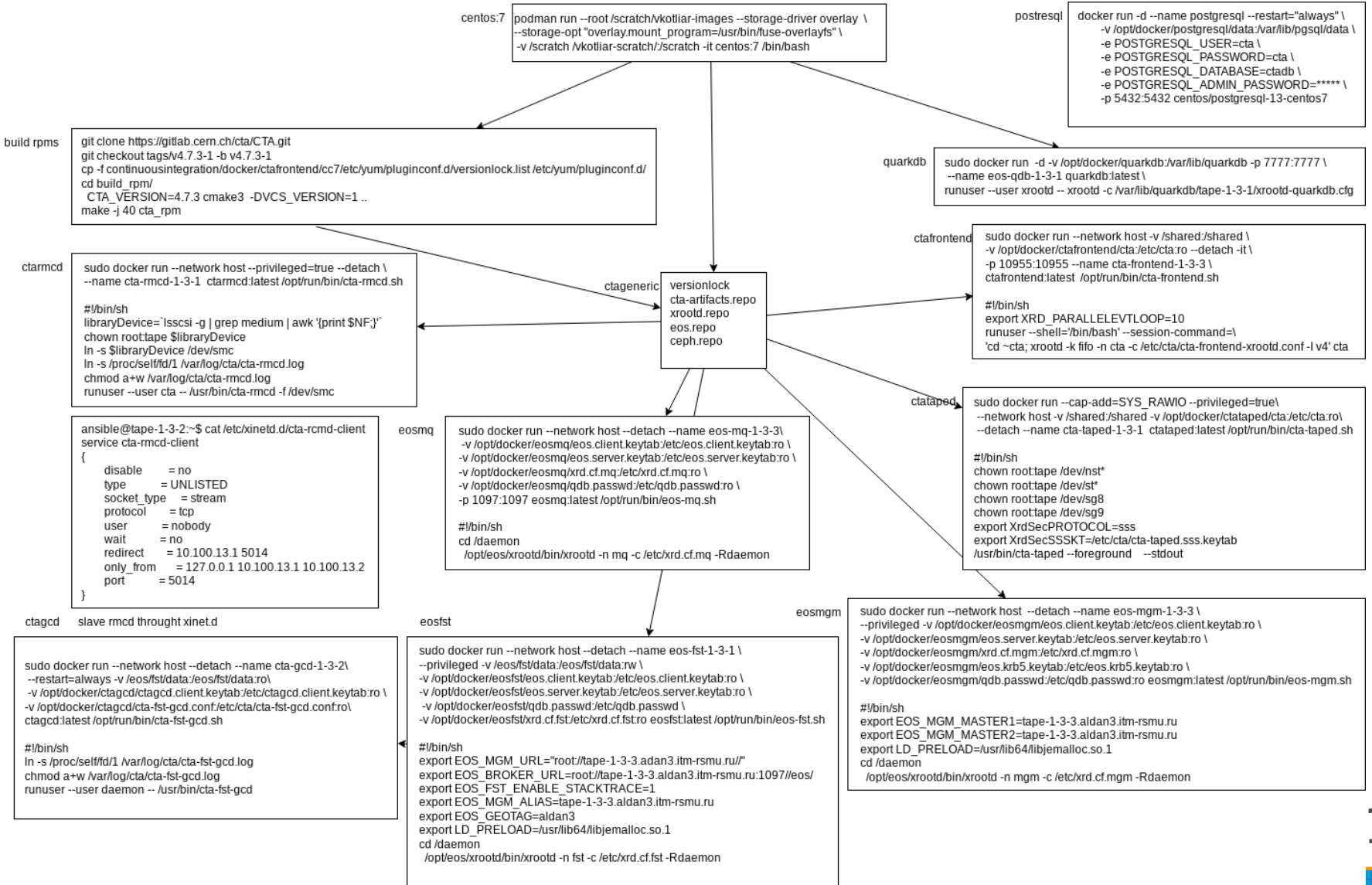
- **All information to build containers has been taken from sources or from the official docs;**
- **Based on images creation -**
  runs on base image;
  make modifictaions inside container;
  save image as a service base image;
  use a service base image to run service.

Installations tricks: container images hierarchy

- Centos:7 was usec
- CTA rpms were built;
- Ctageneric with version lock as a prototipe for all others;



- Each container is one service ~~systemd~~


centos:7
podman run --root /scratch/vkotliar-images --storage-driver overlay \
--storage-opt "overlay.mount_program=/usr/bin/fuse-overlayfs" \
-v /scratch /vkotliar-scratch/:/scratch -it centos:7 /bin/bash
postresql
docker run -d --name postgresql --restart="always" \
-v /opt/docker/postgresql/data:/var/lib/pgsql/data \
-e POSTGRESQL_USER=cta \
-e POSTGRESQL_PASSWORD=cta \
-e POSTGRESQL_DATABASE=ctadb \
-e POSTGRESQL_ADMIN_PASSWORD=***** \
-p 5432:5432 centos/postgresql-13-centos7
build rpms
git clone https://gitlab.cern.ch/cta/CTA.git
git checkout tags/v4.7.3-1 -b v4.7.3-1
cp -f continuousintegration/docker/ctafrontend/cc7/etc/yum/pluginconf.d/versionlock.list /etc/yum/pluginconf.d/
cd build_rpm/
CTA_VERSION=4.7.3 cmake3 -DVCS_VERSION=1 ..
make -j 40 cta_rpm
quarkdb
sudo docker run -d -v /opt/docker/quarkdb:/var/lib/quarkdb -p 7777:7777 \
--name eos-qdb-1-3-1 quarkdb:latest \
runuser --user xrootd -- xrootd -c /var/lib/quarkdb/tape-1-3-1/xrootd-quarkdb.cfg
ctafrontend
sudo docker run --network host -v /shared:/shared \
-v /opt/docker/ctafrontend/cta:/etc/cta:ro --detach -it \
-p 10955:10955 --name cta-frontend-1-3-3 \
ctafrontend:latest /opt/run/bin/cta-frontend.sh
#!/bin/sh
export XRD_PARALLELEVTLOOP=10
runuser --shell='/bin/bash' --session-command=\
'cd ~cta; xrootd -k fifo -n cta -c /etc/cta/cta-frontend-xrootd.conf -l v4' cta
ctarmcd
sudo docker run --network host --privileged=true --detach \
--name cta-rmcd-1-3-1 ctarmcd:latest /opt/run/bin/cta-rmcd.sh
#!/bin/sh
libraryDevice=`lsscsi -g | grep medium | awk '{print $NF;}'`
chown root:tape $libraryDevice
ln -s $libraryDevice /dev/smc
ln -s /proc/self/fd/1 /var/log/cta/cta-rmcd.log
chmod a+w /var/log/cta/cta-rmcd.log
runuser --user cta -- /usr/bin/cta-rmcd -f /dev/smc
ctageneric
versionlock
cta-artifacts.repo
xrootd.repo
eos.repo
ceph.repo
ctataped
sudo docker run --cap-add=SYS_RAWIO --privileged=true\
--network host -v /shared:/shared -v /opt/docker/ctataped/cta:/etc/cta:ro\
--detach --name cta-taped-1-3-1 ctataped:latest /opt/run/bin/cta-taped.sh
#!/bin/sh
chown root:tape /dev/nst*
chown root:tape /dev/st*
chown root:tape /dev/sg8
chown root:tape /dev/sg9
export XrdSecPROTOCOL=sss
export XrdSecSSSKT=/etc/cta/cta-taped.sss.keytab
/usr/bin/cta-taped --foreground --stdout
ansible@tape-1-3-2:~$ cat /etc/xinetd.d/cta-rcmd-client
service cta-rmcd-client
{
disable = no
type = UNLISTED
socket_type = stream
protocol = tcp
user = nobody
wait = no
redirect = 10.100.13.1 5014
only_from = 127.0.0.1 10.100.13.1 10.100.13.2
port = 5014
}
eosmq
sudo docker run --network host --detach --name eos-mq-1-3-3\
-v /opt/docker/eosmq/eos.client.keytab:/etc/eos.client.keytab:ro \
-v /opt/docker/eosmq/eos.server.keytab:/etc/eos.server.keytab:ro \
-v /opt/docker/eosmq/xrd.cf.mq:/etc/xrd.cf.mq:ro \
-v /opt/docker/eosmq/qdb.passwd:/etc/qdb.passwd:ro \
-p 1097:1097 eosmq:latest /opt/run/bin/eos-mq.sh
#!/bin/sh
cd /daemon
/opt/eos/xrootd/bin/xrootd -n mq -c /etc/xrd.cf.mq -Rdaemon
eosmgm
sudo docker run --network host --detach --name eos-mgm-1-3-3 \
--privileged -v /opt/docker/eosmgm/eos.client.keytab:/etc/eos.client.keytab:ro \
-v /opt/docker/eosmgm/eos.server.keytab:/etc/eos.server.keytab:ro \
-v /opt/docker/eosmgm/xrd.cf.mgm:/etc/xrd.cf.mgm:ro \
-v /opt/docker/eosmgm/eos.krb5.keytab:/etc/eos.krb5.keytab:ro \
-v /opt/docker/eosmgm/qdb.passwd:/etc/qdb.passwd:ro eosmgm:latest /opt/run/bin/eos-mgm.sh
#!/bin/sh
export EOS_MGM_MASTER1=tape-1-3-3.aldan3.itm-rsmu.ru
export EOS_MGM_MASTER2=tape-1-3-3.aldan3.itm-rsmu.ru
export LD_PRELOAD=/usr/lib64/libjemalloc.so.1
cd /daemon
/opt/eos/xrootd/bin/xrootd -n mgm -c /etc/xrd.cf.mgm -Rdaemon
ctagcd
slave rmcd throught xinet.d
sudo docker run --network host --detach --name cta-gcd-1-3-2\
--restart=always -v /eos/fst/data:/eos/fst/data:ro\
-v /opt/docker/ctagcd/ctagcd.client.keytab:/etc/ctagcd.client.keytab:ro \
-v /opt/docker/ctagcd/cta-fst-gcd.conf:/etc/cta/cta-fst-gcd.conf:ro\
ctagcd:latest /opt/run/bin/cta-fst-gcd.sh
#!/bin/sh
ln -s /proc/self/fd/1 /var/log/cta/cta-fst-gcd.log
chmod a+w /var/log/cta/cta-fst-gcd.log
runuser --user daemon -- /usr/bin/cta-fst-gcd
eosfst
sudo docker run --network host --detach --name eos-fst-1-3-1 \
--privileged -v /eos/fst/data:/eos/fst/data:rw \
-v /opt/docker/eosfst/eos.client.keytab:/etc/eos.client.keytab:ro \
-v /opt/docker/eosfst/eos.server.keytab:/etc/eos.server.keytab:ro \
-v /opt/docker/eosfst/qdb.passwd:/etc/qdb.passwd \
-v /opt/docker/eosfst/xrd.cf.fst:/etc/xrd.cf.fst:ro eosfst:latest /opt/run/bin/eos-fst.sh
#!/bin/sh
export EOS_MGM_URL="root://tape-1-3-3.adan3.itm-rsmu.ru//"
export EOS_BROKER_URL=root://tape-1-3-3.aldan3.itm-rsmu.ru:1097//eos/
export EOS_FST_ENABLE_STACKTRACE=1
export EOS_MGM_ALIAS=tape-1-3-3.aldan3.itm-rsmu.ru
export EOS_GEOTAG=aldan3
export LD_PRELOAD=/usr/lib64/libjemalloc.so.1
cd /daemon
/opt/eos/xrootd/bin/xrootd -n fst -c /etc/xrd.cf.fst -Rdaemon

Installations wish list (maybe some already addressed):

- Would be better to be a container friendly (no file to stdout redirect);
- Allow to use only config files without env variables;
- Restrict as much as possible privileges and avoid their excalations;
- One rmcd daemon per tape library;
- Allow to use tape drive with disks (one node transfer optimisations)

System usage from cluster:

- **For backup and archiving we use direct eos and xrootd commands (tried webdav to put files but there is no use case in our setup);**
- **Store files as long as possible on disks;**
- **Authorisation group based so in EOS there is a role for each tapeops group;**
- **To prepare archives we use 4G tar files;**
- **CEPH – ZFS – LTO8 all of the use similar comression LZ4-LZ4-LZS very convinient**

System usage from inside:

```
# eos-mgm-1-3-3
  eos mkdir /eos/cta/aldan3
  eos chown vkotliar:aldan3_tapeops /eos/cta/aldan3
  eos chmod 770 /eos/cta/aldan3
  eos attr set sys.acl=u:0:rwx+dp,g:aldan3_tapeops:rwx+dp,z:'!'u,u:0:+u /eos/cta/aldan3
  eos vid ls
    eos vid set membership vkotliar -gids 1516200179,1516200054
    eos vid set membership dbolotin -gids 1516200179,1516200054
    eos vid set membership taped -gids 1516200179,1516200054,1516200081
  eos vid ls
```

```
# name: Copy to EOS
  rc=3
  export EOS_MGM_URL="root://tape-1-3-3.aldan3.itm-rsmu.ru"
  for HDIR in `ls ./` ;
    do kinit -k -t /home/vkotliar/krb5.keytab  vkotliar ;
    set +e
      eos --role vkotliar aldan3_tapeops mkdir $eosDirForBackup;
    set -e
      eos --role vkotliar aldan3_tapeops cp -r $HDIR $eosDirForBackup;
  done
```

```
# name: Create archive files
cat /root/fstab | \
  grep home | \
  grep ceph | \
  grep "\/$userNameForBackup " | \
    awk '{print $2;}' | \
    xargs -iDIR /bin/bash -c 'mkdir $tempDirForBackup/DIR ; cd $tempDirForBackup/DIR ; DNAME=DIR; NAME="${DNAME////_}";
 $TAR_PATH -c --ignore-failed-read --multi-volume --tape-length=4G --file=$NAME.tar.{0..10000} DIR'
```

Needs to be done:

- **Upgrade to the “latest greatest”**
  AL9
  EOS5
  CTA5

Thank you!

Questions?