

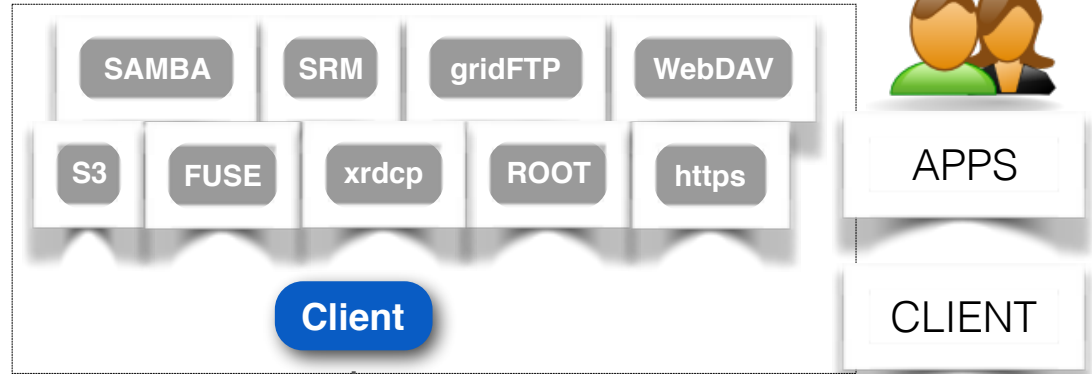




Migration: from EOSUSER to EOSHOME

**Luca Mascetti
CERN IT Storage**

EOS Architecture



EOS Production Releases

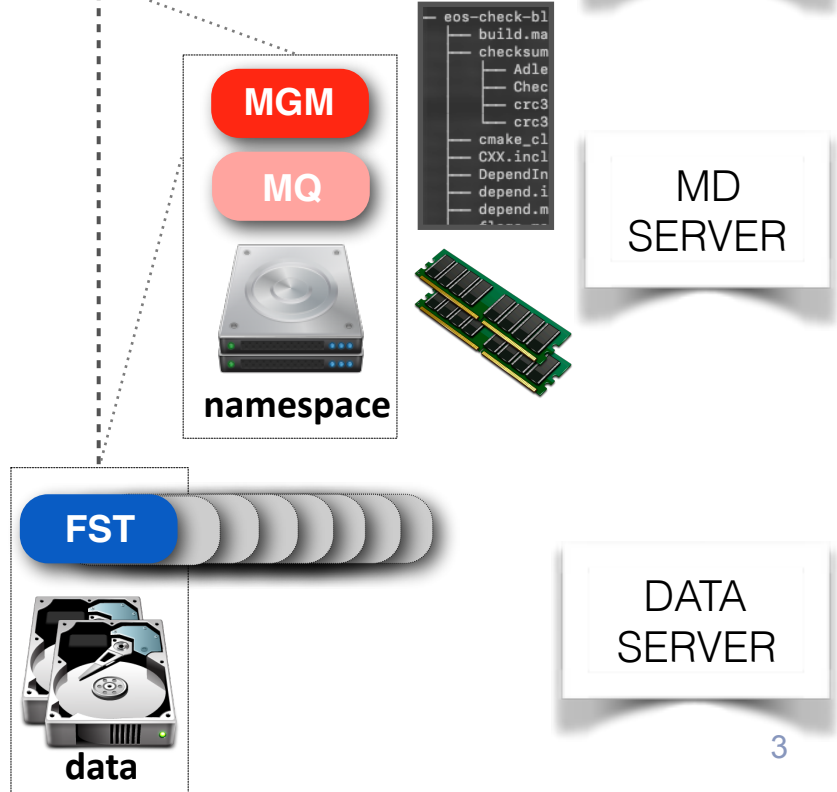


**Aquamarine
V 0.3.X**

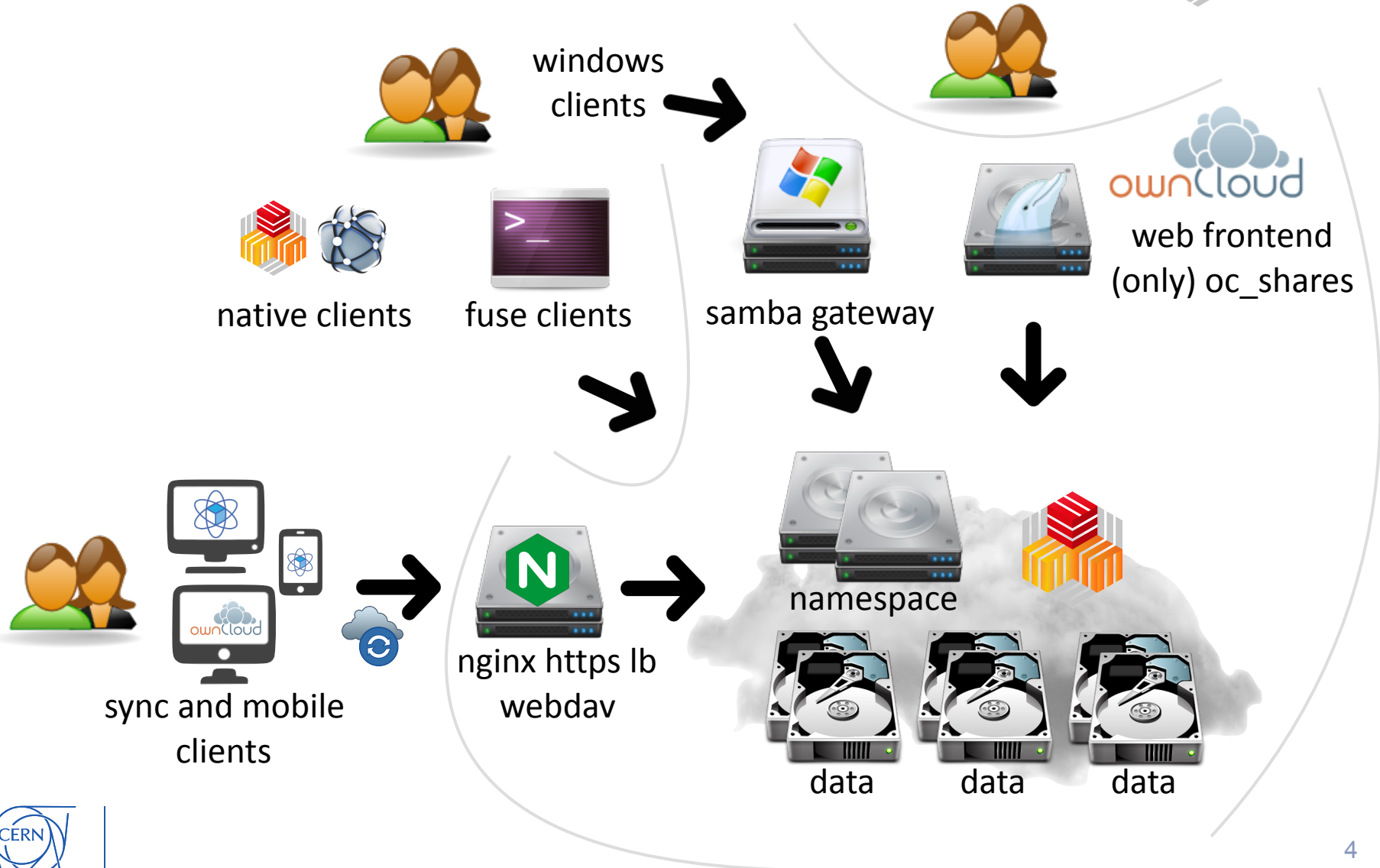


**Citrine
V 4.X**

XRootD V3 IPV4 namespace in-memory data on attached disks	XRootD V4 IPV6 plugins for meta data & data persistency
---	---



EOSUSER a.k.a. CERNBox



EOS Namespace Challenge

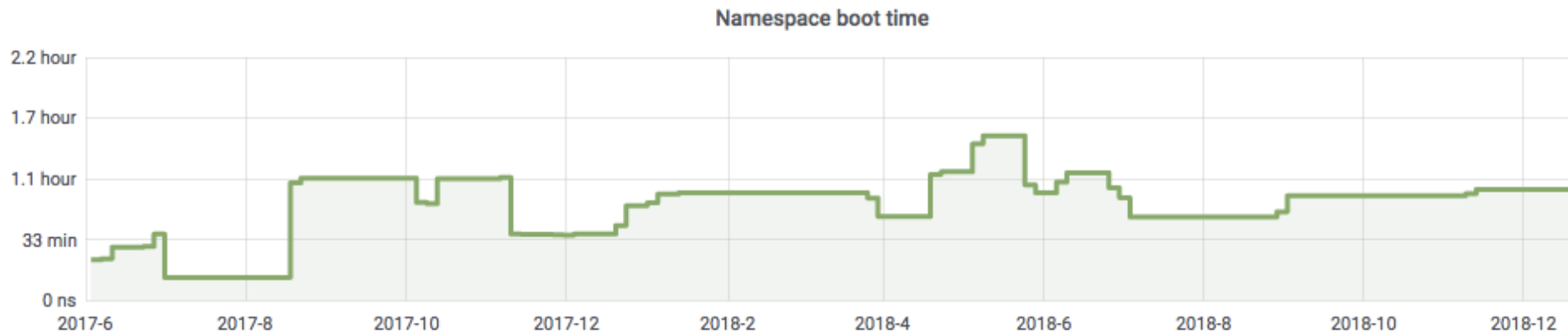
Namespace restart (boot) time

MGM restart:

1. software upgrade
2. memory growth
3. crash or out of memory

While booting:

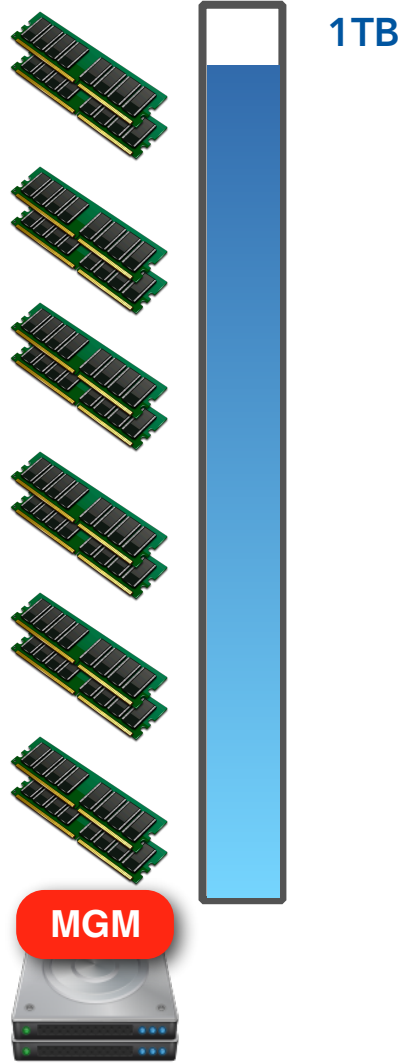
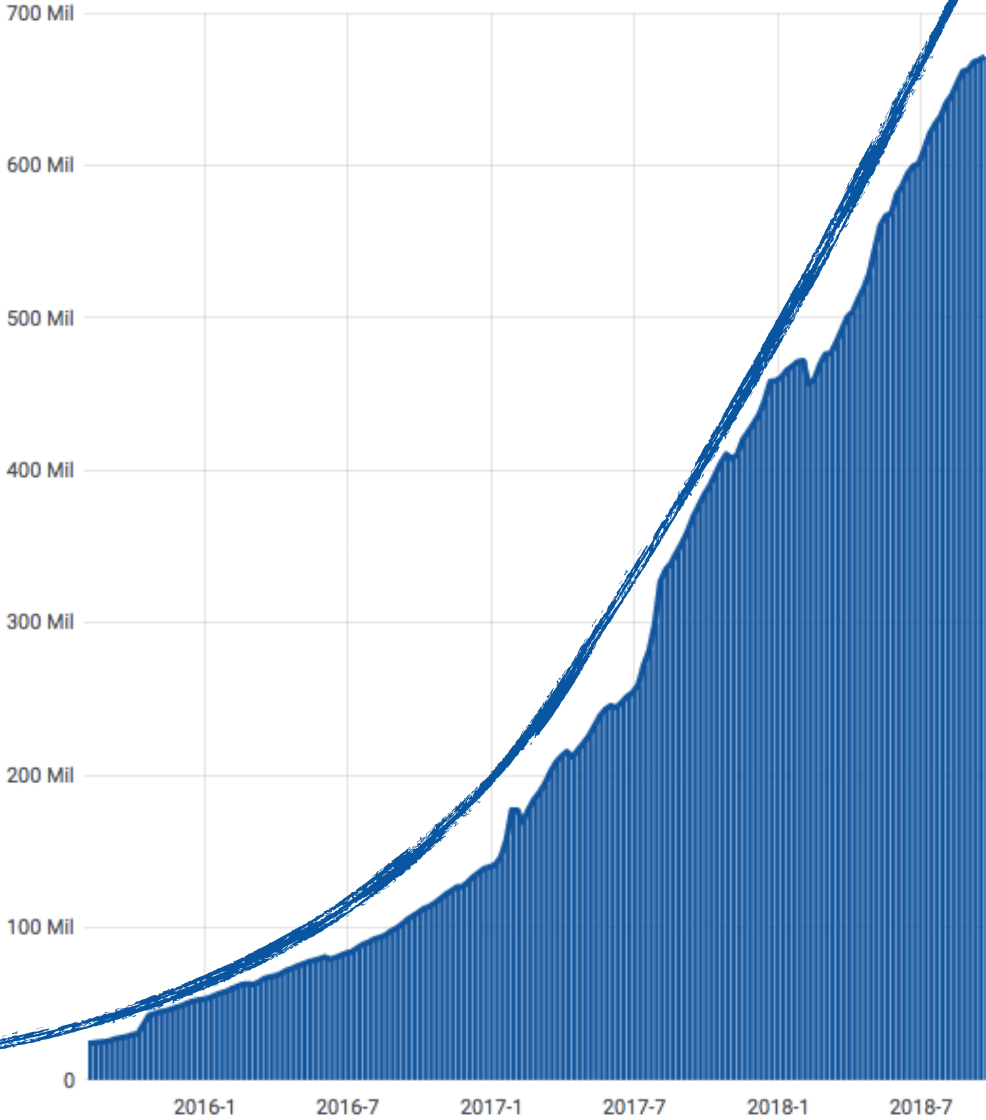
- Instance is in read-only (serving data from MGM Slave)



EOS Namespace Challenge

Memory consumption

Number of files



namespace



ATTENZIONE PERICOLO



POSSIBILITÀ DI ONDE DI PIENA IMPROVVISE
ANCHE PER MANOVRE SU OPERE IDRAULICHE

DANGER



POSSIBILITY OF SUDDEN FLOOD WAVES ALSO
BECAUSE OF MANOEUVRES ON HYDRAULIC PLANTS

ATTENTION DANGER

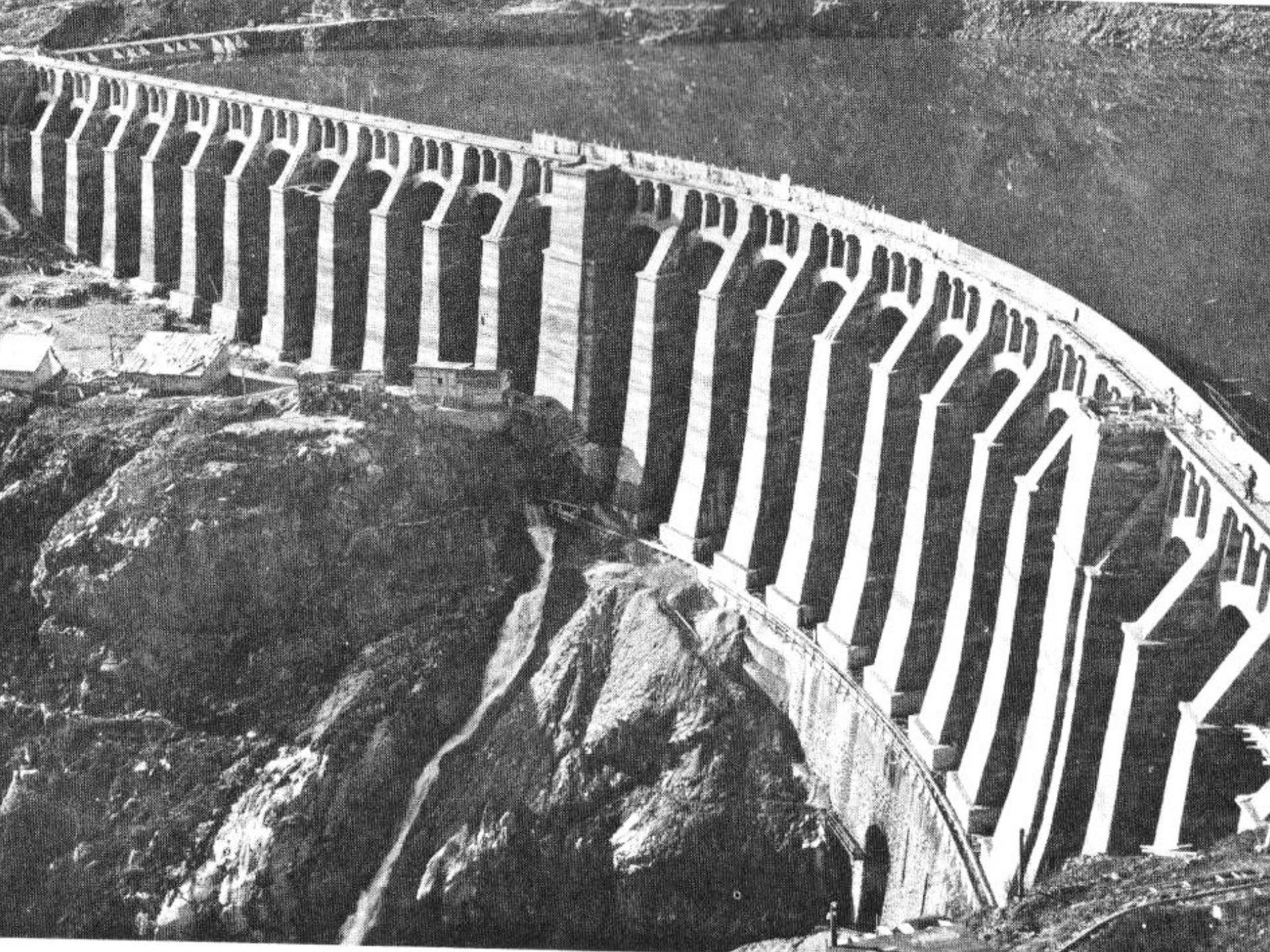


POSSIBILITÉ DE CRUES SOUDAINES À LA SUITE
AUSSI DE MANOEUVRES SUR OUVRAGES HYDRAULIQUES

ACHTUNG GEFAHR



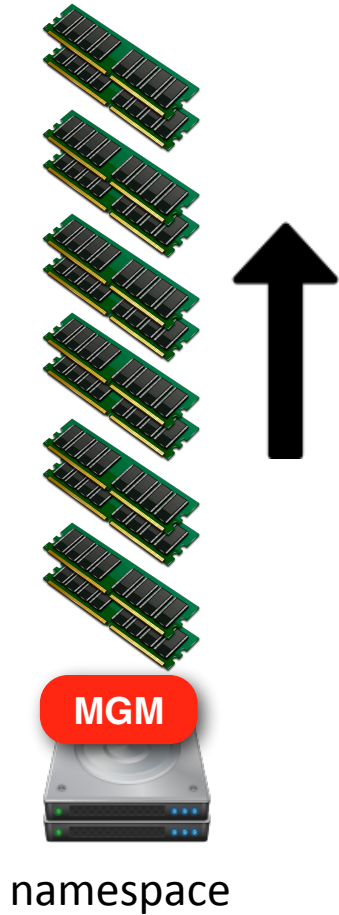
MÖGLICHKEIT PLÖTZLICHER FLUTWELLEN AUCH
ZUFOLGE VON BETÄTIGUNG DER STAUDAMMSCHÜTZE



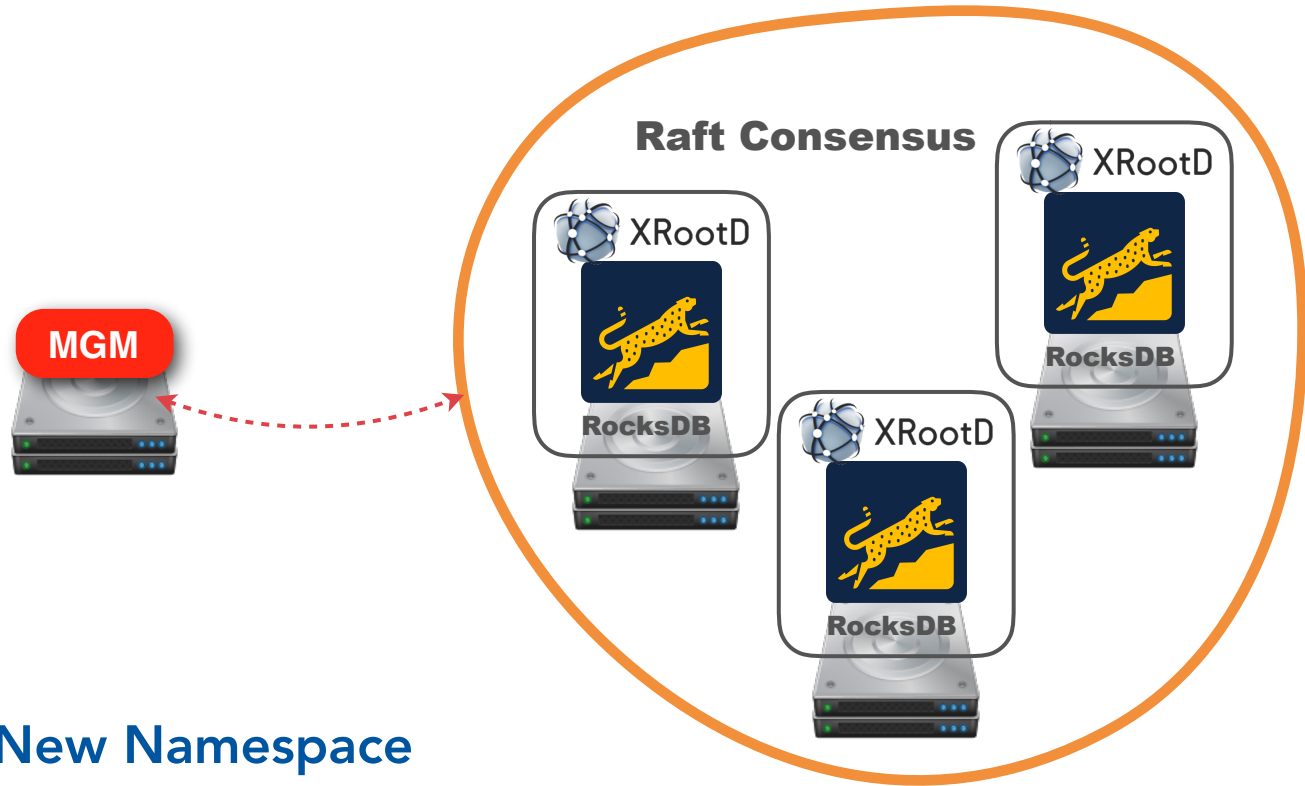


EOS Namespace Challenge

Shifting namespace paradigm: from scale-up to scale-out



QuarkDB namespace



New Namespace

- KV store resident on disk
- **very short restart time!!**
 - not based on #files and #dirs
- namespace caching in MGM memory

Now, how to proceed ...



Solution 1: EOSUSER upgrade

Upgrade current production

Two steps upgrade:

1. upgrade from aquamarine to citrine
2. namespace conversion

From past experiences:

- very very very long downtime => just not acceptable
- instabilities in booting filesystems with millions of files

Solution 1: EOSUSER upgrade

Upgrade current production

Two steps upgrade:

1. upgrade from aquamarine to citrine
2. namespace converge

From past experiences:

- very very very long downtime => just not acceptable
- instabilities in booting filesystems => can't read files



Solution 2: EOSUSER2

Deploy a new empty instance with latest namespace technology

New deployment and migration:

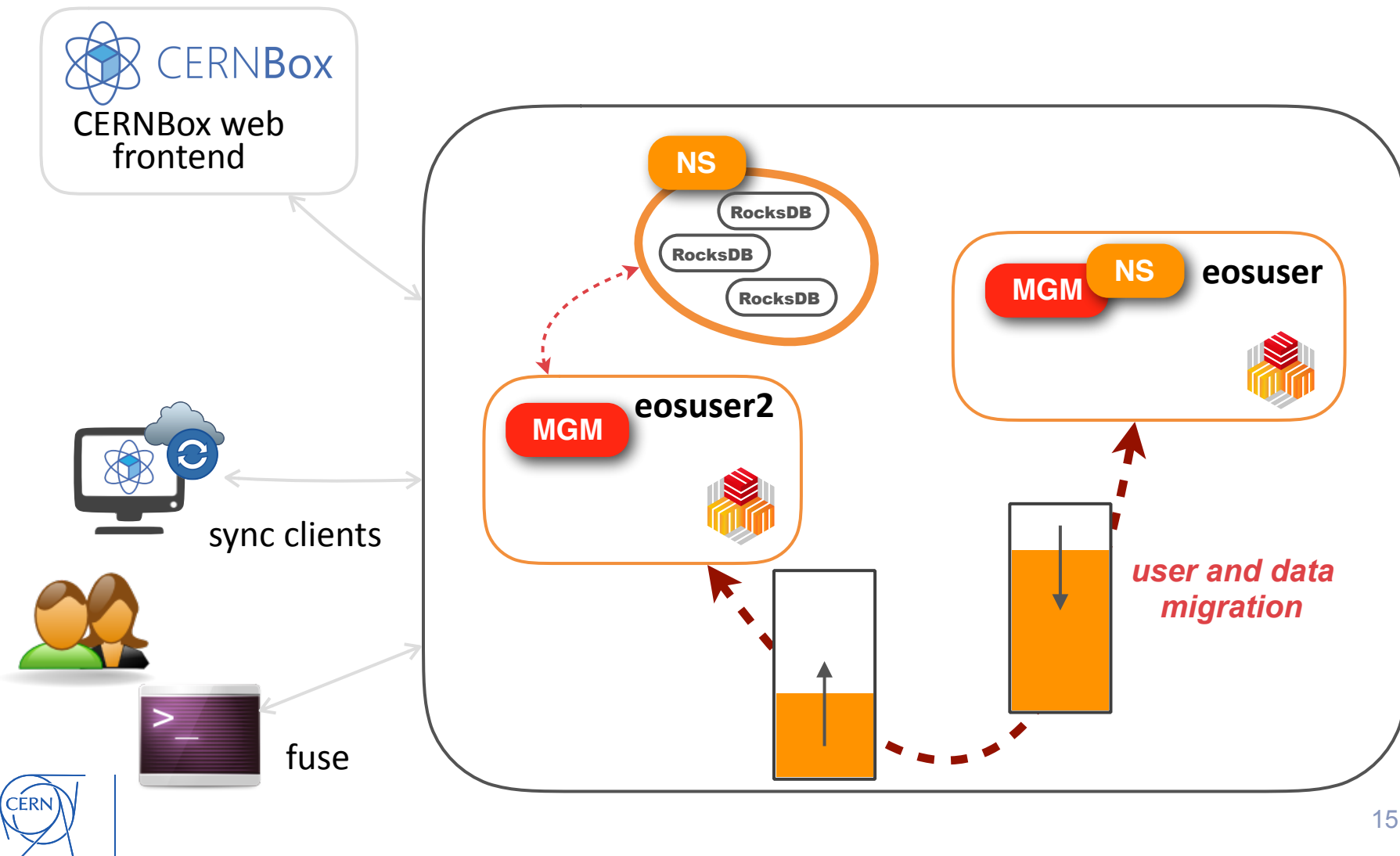
1. build a **new empty EOS instance**
 1. start immediately with QDB namespace
 2. migrate gradually users

From past experiences:

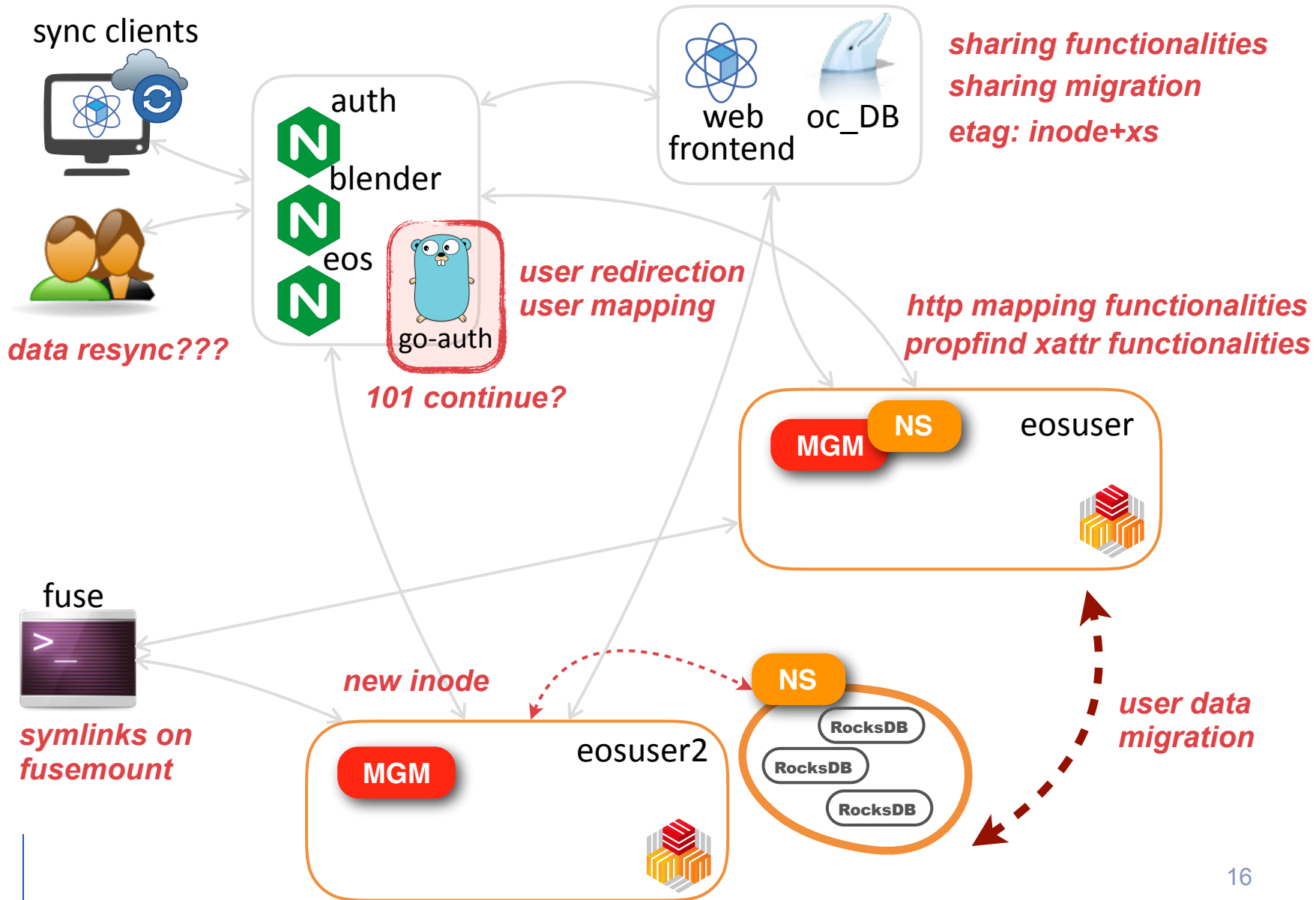
- migration needs to be transparent!!!
- no BIG-BANG! approach
- better load control over time
- better operations “feeling”

Solution 2: EOSUSER2

Deploy a new empty instance with latest namespace technology



FYI: Behind the scenes

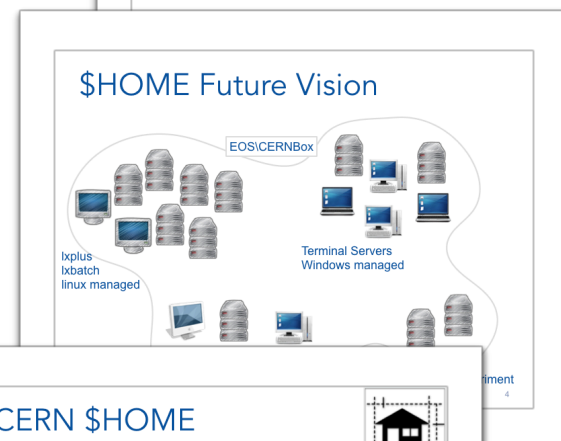


Solution 2: EOSUSER2

Deploy a new empty instance with latest namespace technology

Some additional considerations:

- single instance for all users
- MGMs single point of failures
- Scale metadata performance
- difficult users isolation
- future big upgrade => big bang?
- “plane crash” effect
- Shared Desktop across devices
 - CERN \$HOME future plans
 - DFS and linux home discussions



CERN \$HOME

- Taking advantage from synchronisation
 - local vs. remote access
- Avoid to run separate/isolate storage cluster
 - better system interoperability
 - profit from future synergy (DFS)

\$HOME will be set to `/eos/user/<u>/<username>/`

```
lsroot@lxplus066 ~]$  
lsroot@lxplus066 ~]$ pwd  
/eos/user/jj/doe
```

Solution 2: EOSUSER2

Deploy a new empty instance with latest namespace technology

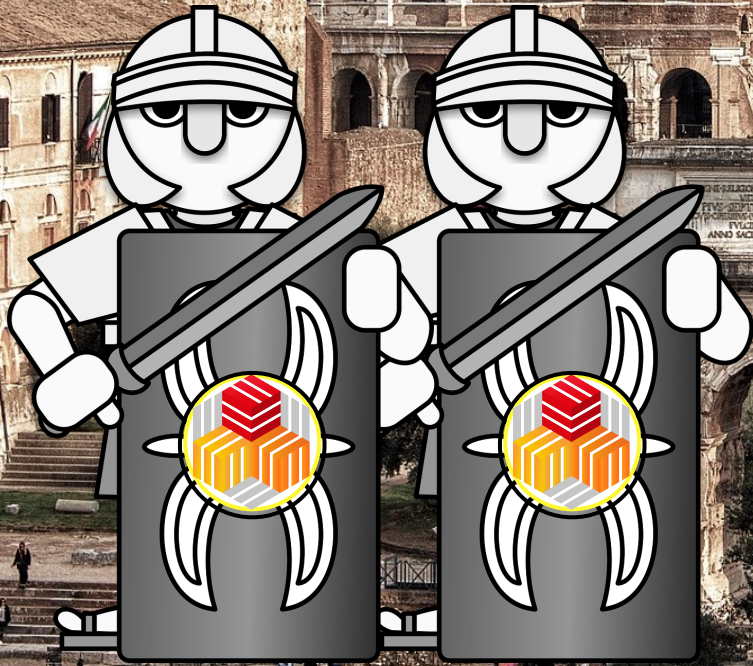
Some additional considerations:

- single instance for all users
- MGMs single point of failures
- Scale metadata performance
- difficult users isolation
- future big upgrade => big barrier
- "plane crash" effect
- Shared Desktop across devices
 - CERN \$HOME future plans
 - DFS and linux home discussions



Solution 3: ...

Divide et Impera



Solution 3: EOSHOME (running out of names...)

Deploy multiple empty instances with latest namespace technology

Architectural review, new deployment and migration:

1. build a **multiple empty EOS instance**
 1. start immediately with QDB namespace
2. add a level of indirection
3. support system expansion and reduction
4. migrate gradually users

From past experiences:

- migrations need to be transparent!!!
- no BIG-BANG! approach
 - gradual (future) software roll-out
- better load control over time
- better operations “feeling”
- better user isolation
- less load/stress per instance

Solution 3: EOSHOME (running out of names...)

Deploy multiple empty instances with latest namespace technology

Architectural review, new deployment and migration:

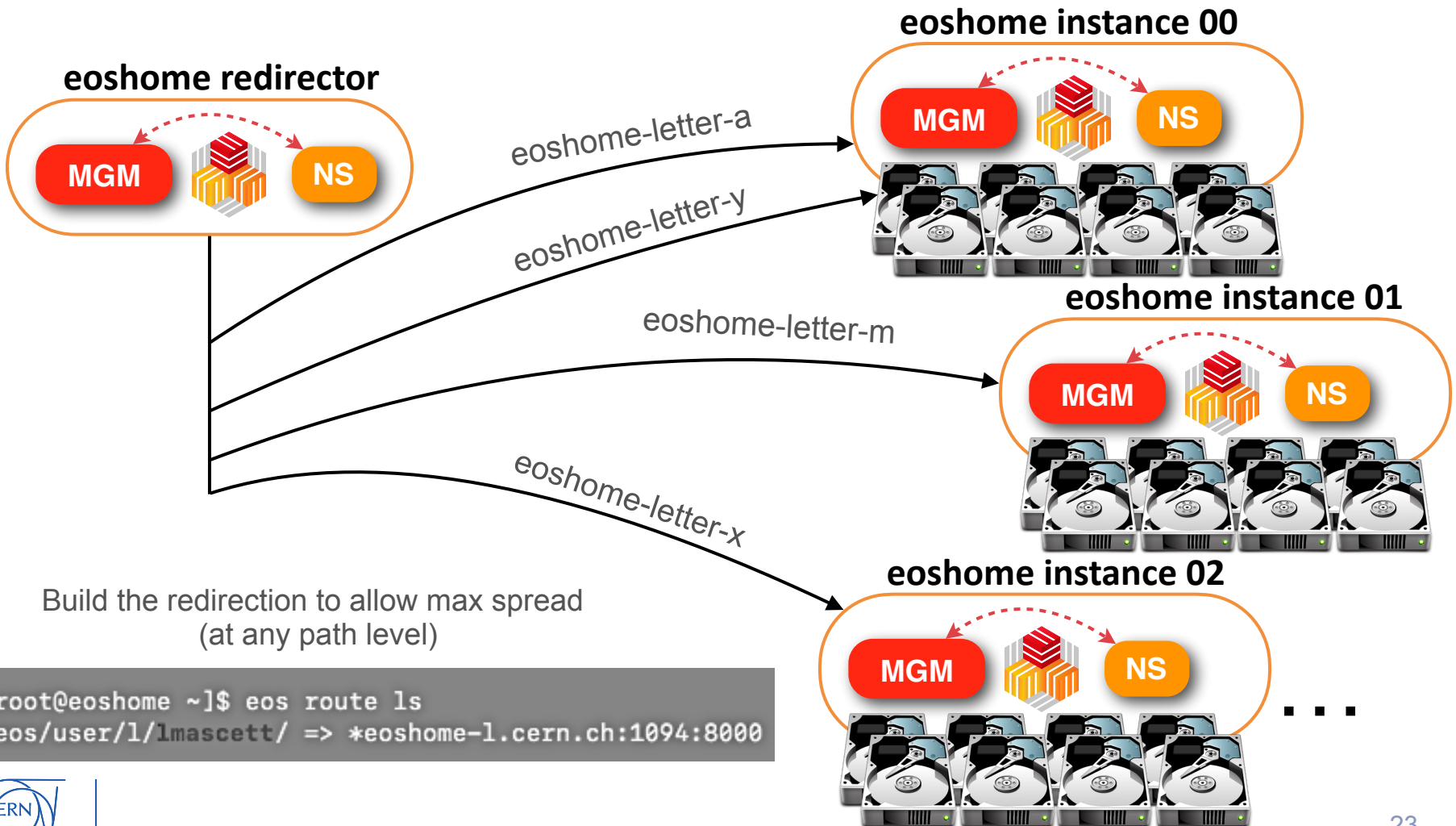
1. build a **multiple empty EOS instance**
 1. start immediately with ODB namespace
2. add a level of indirection
3. support system expansion and production
4. migrate gradually users

From past experiences:

- migrations need to be transparent!!!
- no BIG BANG! approach
 - gradual (future) software roll-out
- better kind control over time
- better operations "feeling"
- better user isolation
- less load/stress per instance

Solution 3: EOSHOME

Deploy multiple empty instances with latest namespace technology



Build the redirection to allow max spread
(at any path level)

```
[root@eoshome ~]$ eos route ls  
/eos/user/l/lmascett/ => *eoshome-1.cern.ch:1094:8000
```

Solution 3: EOSHOME

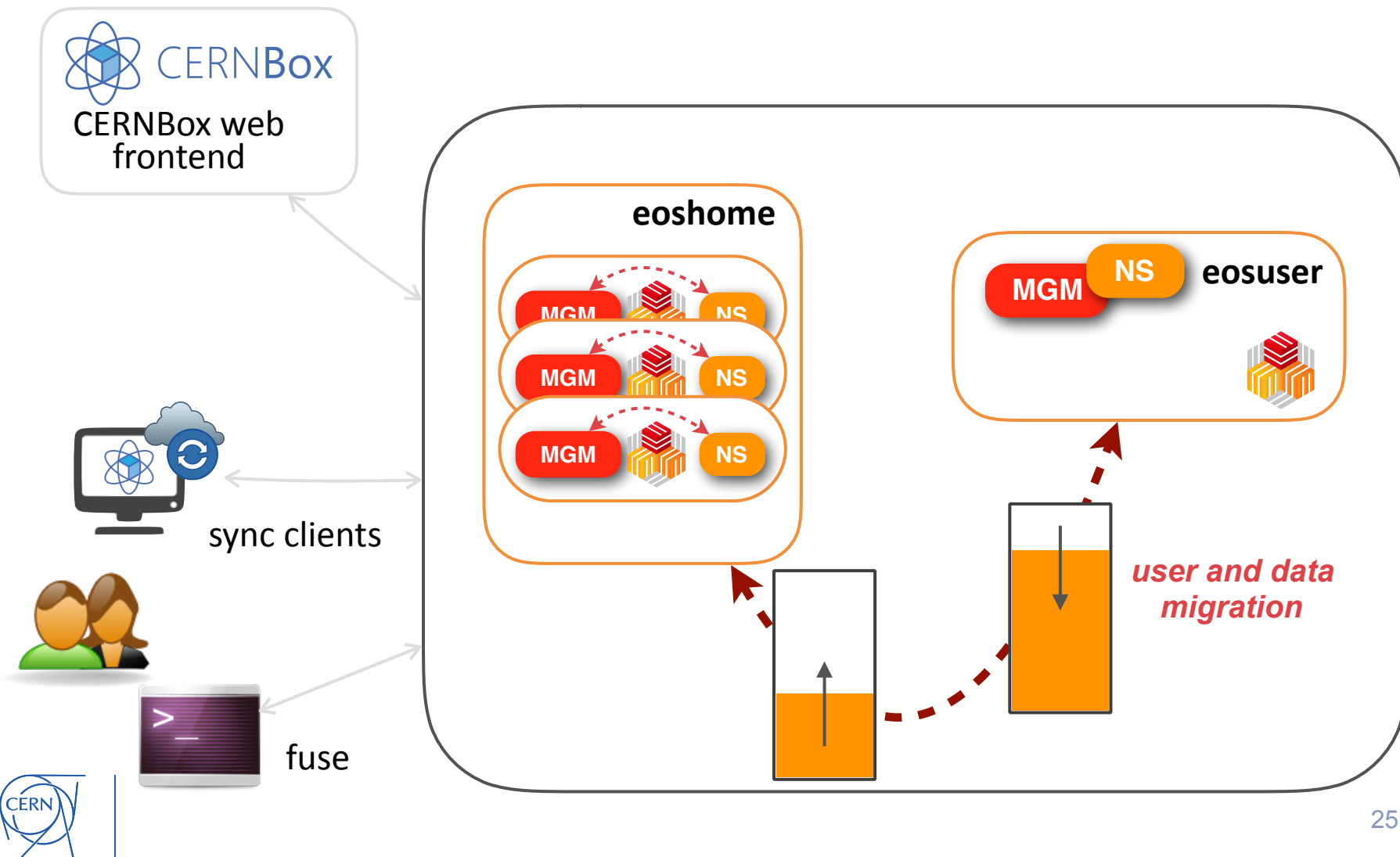
Deploy multiple empty instances with latest namespace technology

eoshome instance XY



Solution 3: EOSHOME

Deploy multiple empty instances with latest namespace technology

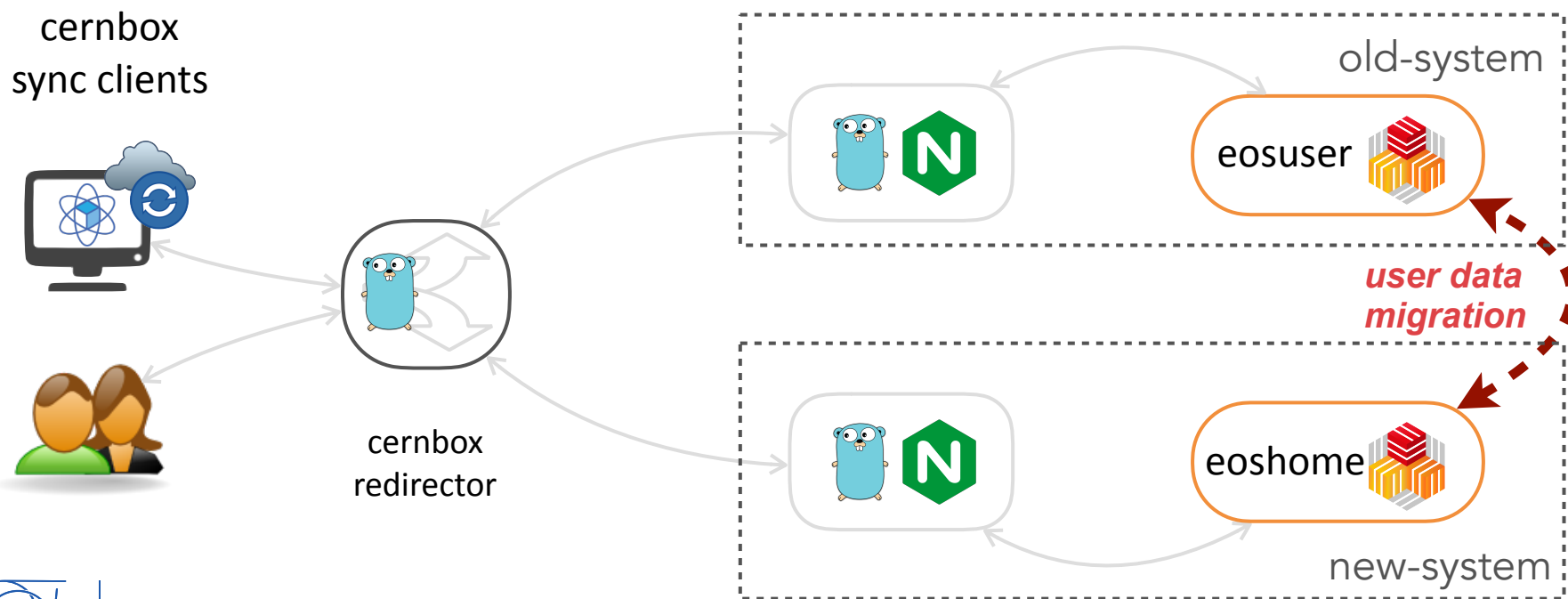


Solution 3: EOSHOME

Deploy multiple empty instances with latest namespace technology

Migration scenario similar to **Solution 2**

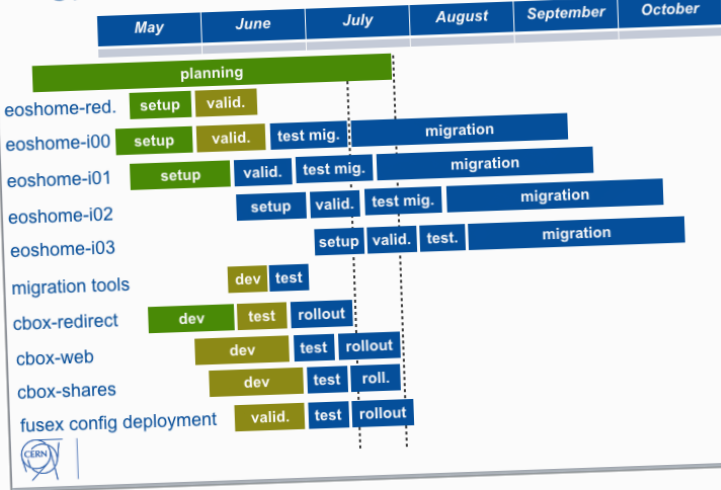
- same requirements on the CERNBox side
- same requirements on the migration tools



Let's go ...



Current Status and Roadmap

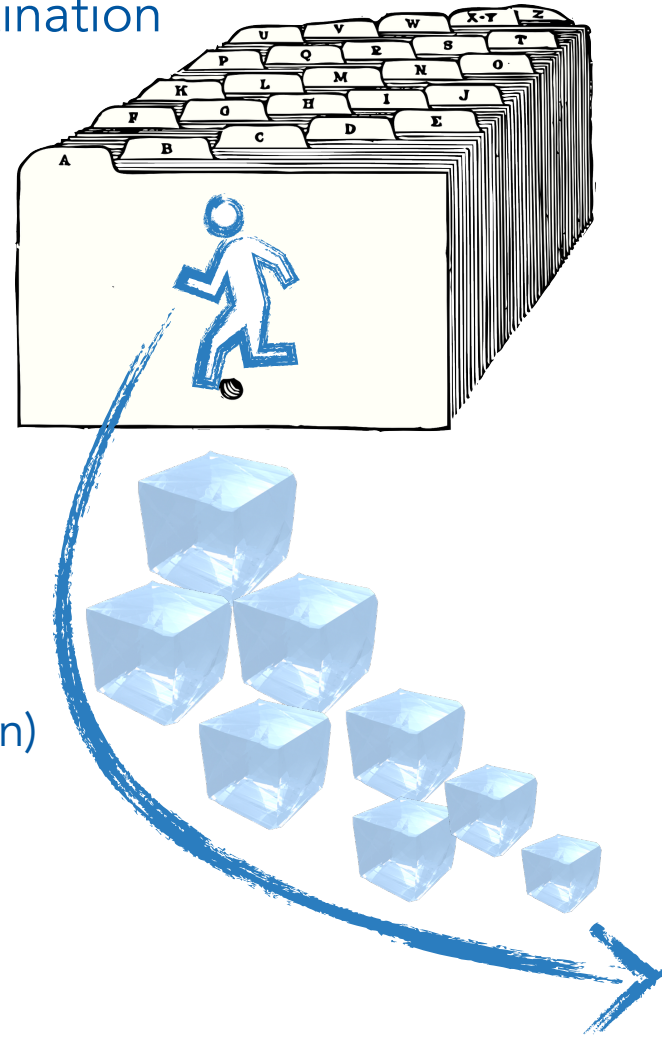


15k Users Data Migration

Data is mirrored in an hidden folder on destination

Migration pseudo-script:

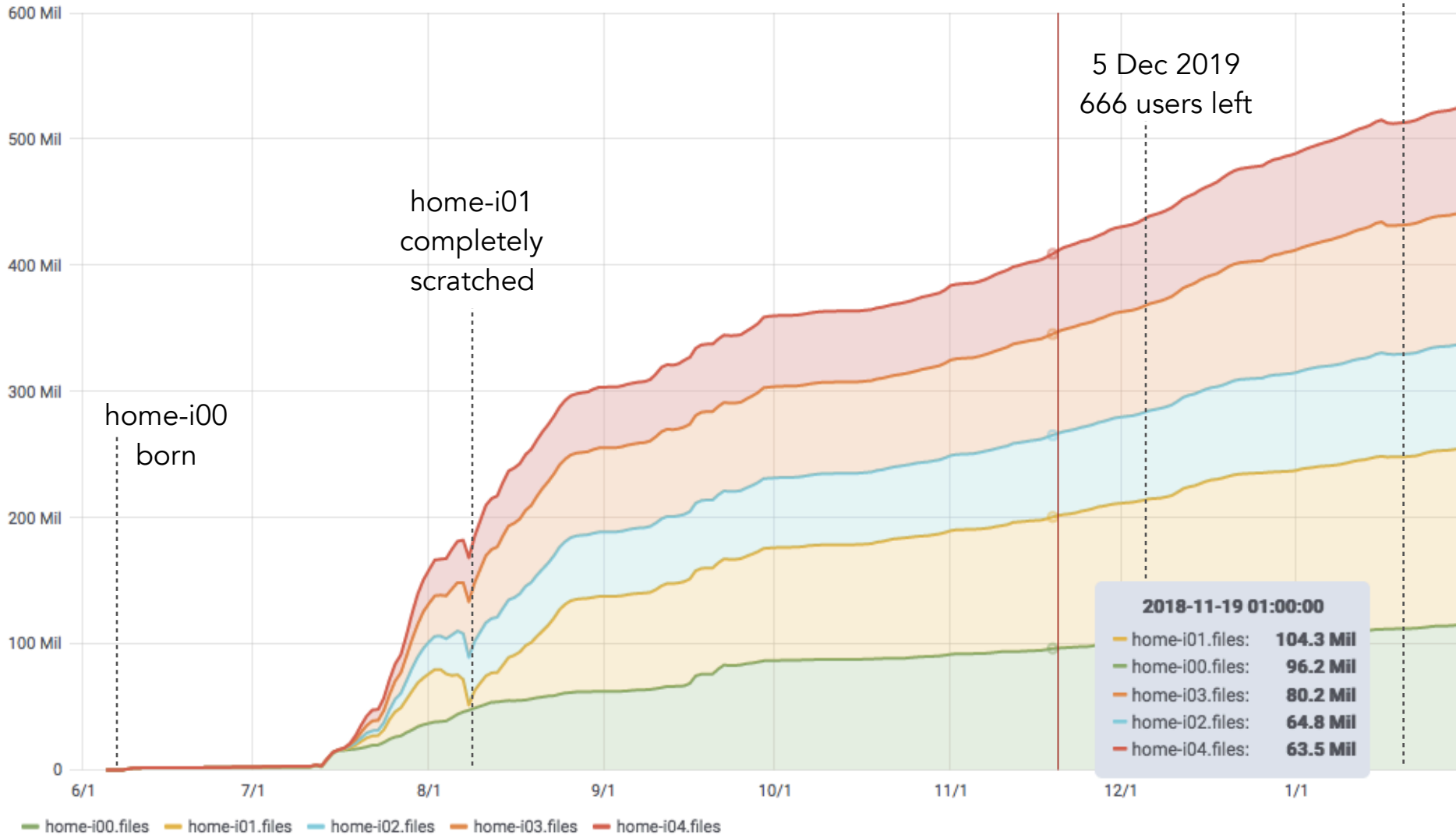
- check user exists
- check user migration status
- block user quota and sync client
- migrate data (MGM query)
 - create folders + ACLs
 - copy files (XS+size match)
 - set correct ETAG for each file
- rsync -av --delete + rsync -av
- sync client verification (propfind comparison)
 - XS+SIZE+ETAG matching
 - No duplicated file & dir IDs
- MGM query comparison
- CERNBox shares migration
- account swap + re-enable



15k Users Data Migration

Number of files ▼

15 Jan 2019
>200 users left



15k Users Data Migration

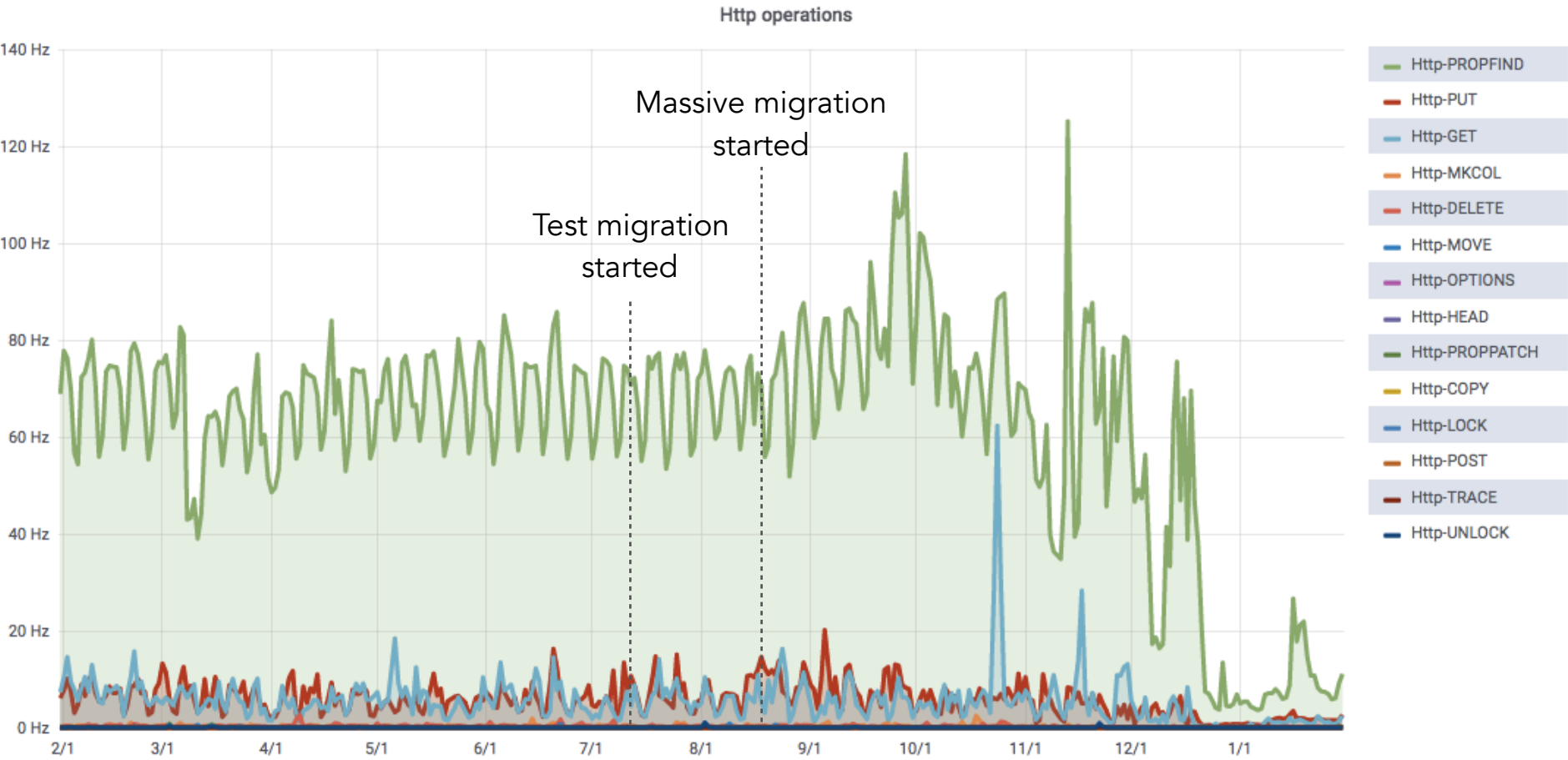
... and as well infinite loops ...

```
'/cernbox/desktop/remote.php/webdav/home/atlas/nosyn/analysis/ttH/tH_prod/tH_4fl/dyn2/clean_d2/shower2/madevent/  
/P0_tx_bxwm_wm_scx/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check  
heck_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/che  
/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/c  
.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f  
sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa  
k_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check  
eck_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/ched  
check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/ch  
f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/  
a.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.  
_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_s  
ck_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check  
heck_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/che  
/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/c  
.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f  
sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa  
k_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check  
eck_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/ched  
check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/ch  
f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/  
{'DAV:}getcontentlength': '5469', '{http://owncloud.org/ns}permissions': 'RWCKNVD', '{DAV:}getlastmodified': '  
one, '{DAV:}getetag': '"461062971846557696:e4da9d7f"', '{http://owncloud.org/ns}id': 'FFF', '{http://owncloud.or  
'/cernbox/desktop/remote.php/webdav/home/atlas/nosyn/analysis/ttH/tH_prod/tH_4fl/dyn2/clean_d2/shower2/madevent/  
/P0_tx_bxwm_wm_scx/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check  
heck_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/che  
/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/c  
.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f  
sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa  
k_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check  
eck_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/check_sa.f/ched
```





EOSUSER HTTP Activity



Summary and Outlook



Thanks to all the people involved in this activity!
Thanks for the hard work!

General improvement of **EOS\CERNBox** architecture

- removing SPOFs
- improving metadata performance
- reducing drastically downtimes
 - less user impacted
 - almost zero restart time
- flexibility to scale up and out at the same time
- removing big-bang upgrades
 - simplify small scale testing and software rollout

Thanks for the attention!



www.cern.ch

Questions?

mixed eoshome instances over hw

