XROOTD NATIF POUR ALICE : GÉRER LES PERTES DE DONNÉES ET LE DECOMMISSIONING

Journées LCG-France, CC-IN2P3
14-16 Décembre 2015
diarra@ipno.in2p3.fr
D2I-S2I

Unité mixte de recherche
CNRS-IN2P3
Université Paris-Sud

91406 Orsay cedex
Tél. : +33 1 69 15 73 40
Fax : +33 1 69 15 64 70
http://ipnweb.in2p3.fr

BUT DE LA PRESENTATION

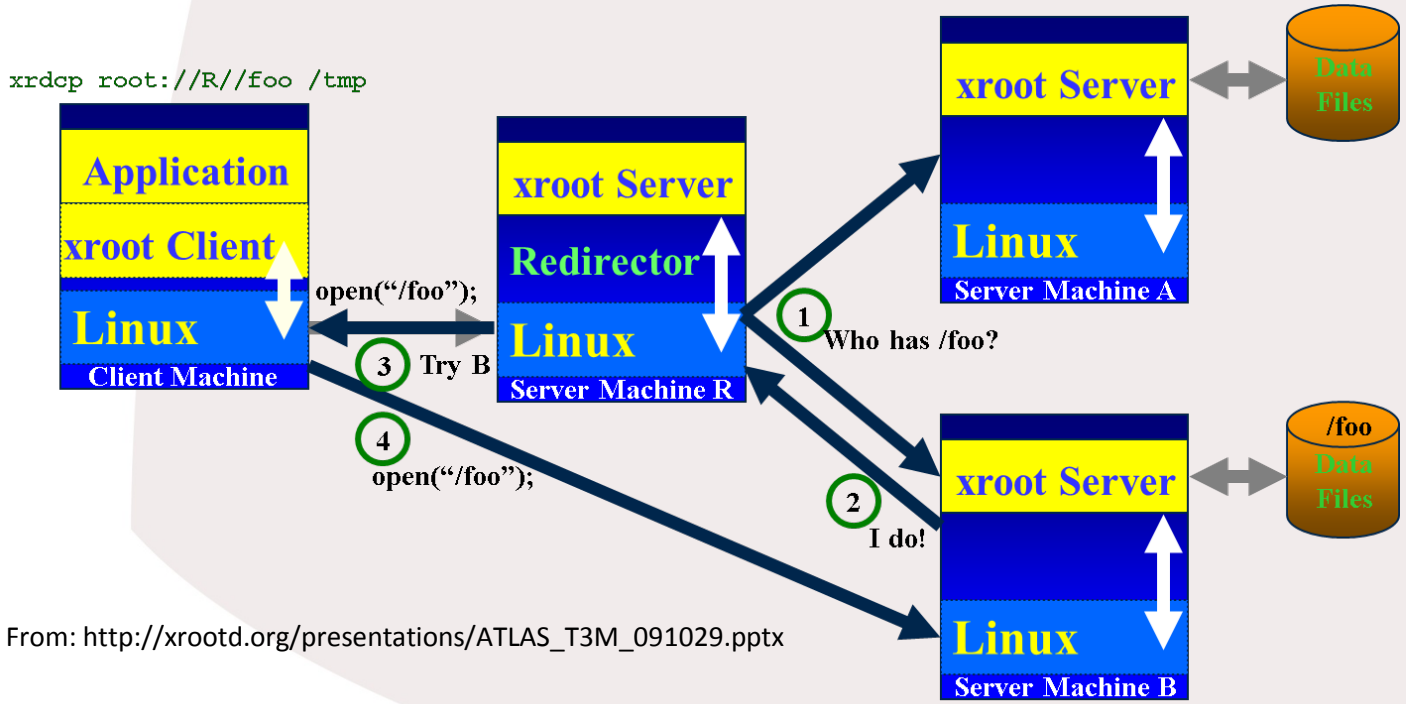
- Partager l'expérience de l'IPNO avec xrootd natif :
  - Dans le decommissioning de serveurs xrootd

  → Par exemple quand on veut arrêter du vieux matériel
  - Sur la procédure à suivre pour restaurer des données perdues

  → Par exemple suite à la perte d'un volume RAID

RAPPEL RAPIDE DU FONCTIONNEMENT DE XROOTD NATIF SUR LES SITES ALICE

- Un SE xrootd natif pour ALICE comprend :
  - un redirecteur (manager)
  - N serveurs xrootd (disk severs ou xrootd servers ou servers)
  - Authentification propre à ALICE pour les 'write' ('read' ouvert à tous)



From: http://xrootd.org/presentations/ATLAS_T3M_091029.pptx

- \+ Il n'y a pas de base de données ➔ gestion simplifiée
- \- Il n' y a pas d'utilitaire comme dpm-drain ➔ decommissioning moins formalisé

L'ARBORESCENCE DES FICHIERS XROOTD SUR LES SERVEURS

- Sur chaque serveur il y a un namespace (un simple directory)
  - ✓ Il contient simplement des symlinks vers les vrais fichiers de data
- Les fichiers sont stockées dans les partitions de données
- Chaque fichier a un GUID généré par ALICE lors de sa création
- Le GUID fait partie du nom du fichier

L'ARBORESCENCE DES FICHIERS SUR LES SERVEURS- EXEMPLE

Fichier de **GUID: b8f9f574-dd42-11e4-a4e6-63e8b3f6492f**

Sur le serveur où se trouve le fichier:

ls -lh /grid/xrddata1/namespace/00/65278/b8f9f574-dd42-11e4-a4e6-63e8b3f6492f

lrwxrwxrwx 1 xrootd xrootd 85 Apr 7 2015
/grid/xrddata1/namespace/00/65278/**b8f9f574-dd42-11e4-a4e6-63e8b3f6492f ->**
/grid/xrddata6/%grid%xrddata1%namespace%00%65278%**b8f9f574-dd42-11e4-a4e6-63e8b3f6492f**

ls -lLh /grid/xrddata1/namespace/00/65278/b8f9f574-dd42-11e4-a4e6-63e8b3f6492f
-rw-rw-r-- 1 xrootd xrootd **3.6M** Apr 7 2015 /grid/xrddata1/namespace/00/65278/b8f9f574-dd42-11e4-a4e6-63e8b3f6492f

Depuis un WN on peut faire un **xrdcp** de ce fichier par :

xrdcp \
root://ipngridxrd0.in2p3.fr:1094//00/65278/**b8f9f574-dd42-11e4-a4e6-63e8b3f6492f** \
/tmp/xrd_test.dat

DECOMMISSIONING PAR MIGRATION DES DONNEES VERS UN AUTRE SERVEUR ET AVEC PRESERVATION DES PATH

C'est le cas le plus pratique pour décommissionner un serveur A. Les étapes:

1. Trouver un serveur B avec assez d'espace disponible
2. Prévenez ALICE (mail à alice-lcg-task-force@cern.ch)
3. Mettre le serveur A en read-only
4. Avec rsync copier chaque partition de A en totalité dans une partition de B

   N.B.: on considère que les 2 partitions portent le même nom

   Ex: rsync –a A:/grid/xrddata${i}/ B:/grid/xrddata1${i}/ i=1..N
5. Merger les namespace: copier avec rsync le namespace de A dans B

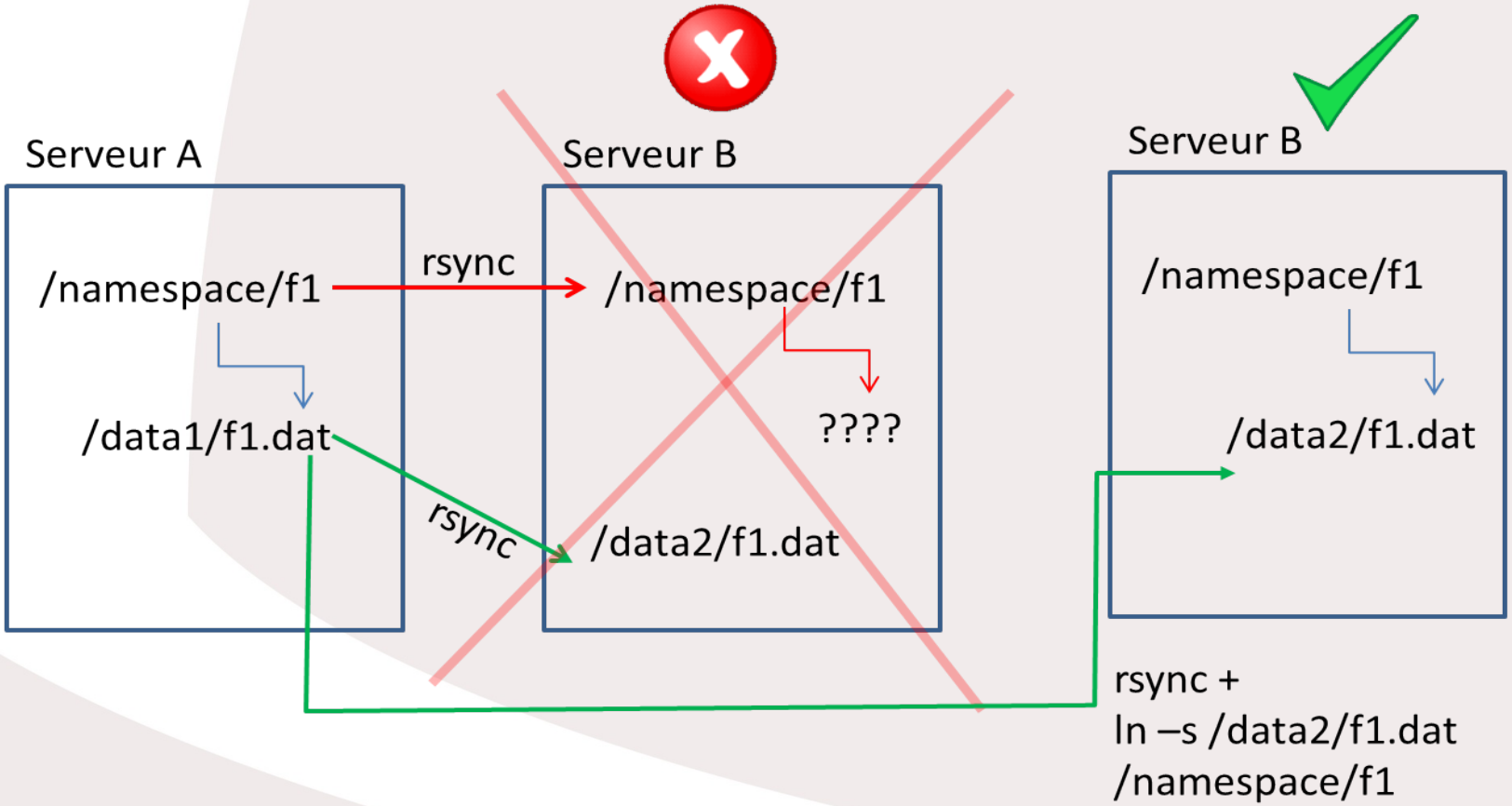
   Ex: rsync –a A:/grid/xrddata1/namespace/ B:/grid/xrddata1/namespace

   N.B.: A et B servent maintenant les mêmes fichiers en lecture
6. Sur A: arrêter le service xrootd, désinstaller xrootd et arrêter A.

DECOMMISSIONING PAR MIGRATION DES DONNEES VERS UN AUTRE SERVEUR ET SANS PRESERVATION DES PATH

Ce cas est moins simple: des partitions de A et B n'ont pas les mêmes noms

- le merge des namespace sur B va nécessiter la mise à jour des symlinks (ln –s)
- une simple copie par rsync du namespace de A vers B ne suffit pas



DECOMMISSIONING PAR MIGRATION DES DONNEES VERS UN AUTRE SERVEUR ET SANS PRESERVATION DES PATH (SUITE)

Après avoir trouvé un serveur B avec de l'espace disponible, prévenu ALICE et mis le serveur A en read-only, faire:

1. Pour chaque partition à migrer, collecter la liste des symlinks du namespace avec les noms des fichiers de données associés
2. Avec rsync copier chaque partition de A en totalité dans une partition de B
3. Pour chaque partition dont le nom a été préservé:
   merger les namespace en copiant avec rsync le namespace de A dans le namespace de B pour uniquement les fichiers de cette partition
4. Pour chaque partition dont le nom n'a pas été préservé:
   refaire les symlinks (nous avions collecté la liste sur A) sur B avec 'ln –s' en actualisant le nom du fichier de données
5. Sur A: arrêter le service xrootd, désinstaller xrootd et arrêter A.

N.B: on peut ignorer 3. et faire 4. pour toutes les partitions copiées

DECOMMISSIONING PAR MIGRATION DES DONNEES VERS UN AUTRE SERVEUR ET SANS PRESERVATION DES PATH (SUITE)

Remarque: en réalité on peut migrer les data du serveur A vers plusieurs serveurs du moment que les symlinks sont correctement mis à jour après la copies des fichiers de data

DECOMMISSIONING PAR MIGRATION DES DONNEES VERS UN SE PAR LES ALICE EXPERTS

Lors du decommissioning d'un serveur A, si vous n'avez pas assez d'espace dans les partitions des autres serveurs, la copie par rsync ne sera pas utilisable même s'il reste assez d'espace sur votre SE. Les étapes à suivre seront donc:

1. Prévenez ALICE (mail à alice-lcg-task-force@cern.ch)
2. Mettre le serveur A en read-only
3. Collecter les GUIDs+la taille des fichiers présents sur le serveur A
4. Envoyer ces informations à ALICE qui fera le transfert par xrootd:
   - soit sur votre SE, s'il y a assez d'espace disponible
   - soit sur un autre SE d'ALICE
   - ✓ ALICE fournira un URL dans Monalisa pour suivre le transfert
5. A la fin du transfert, arrêter xrootd sur A, désinstaller xrootd et arrêter A.

RESTAURATION DES DONNEES SUITE A UNE PERTE DE PARTITIONS SANS PERTE DU NAMESPACE

Si vous perdez une ou plusieurs partitions sur un serveur A avec un namespace en bon état, ALICE peut restaurer les données perdues sur votre SE ou un autre SE.

Les étapes à suivre seront :

1. Prévenez ALICE (alice-lcg-task-force@cern.ch), mettre le serveur en read-only
2. Collecter les GUIDs des fichiers qui étaient sur les partitions perdues
   - ✓ les (broken) symlinks sont encore dans le namespace
3. Supprimer sur A les broken symlinks pointant sur les partitions perdues, remettre le serveur en read-write
4. Envoyer les GUIDs collectés à ALICE qui fera le transfert par xrootd:
   - Soit sur votre SE, s'il y a de l'espace disponible
   - Soit sur un autre SE d'ALICE
   - ✓ ALICE fournira un URL dans Monalisa pour suivre l'avancement du transfert

RESTAURATION DES DONNEES SUITE A UNE PERTE DES PARTITIONS ET DU NAMESPACE

Si vous perdez un serveur (toutes les partitions + le namespace) , ALICE peut restaurer les données perdues sur votre SE ou un autre SE.

Les étapes à suivre seront :

1. Prévenez ALICE (alice-lcg-task-force@cern.ch)
2. Collecter les GUIDs+la taille des fichiers de tous les serveurs
3. Envoyer ces informations à ALICE qui fera un diff avec la liste des GUIDs présents sur votre SE (grâce au catalogue ALICE) pour identifier les fichiers perdus. ALICE fera ensuite le transfert par xrootd:
   - soit sur votre SE, s'il y a de l'espace disponible
   - soit sur un autre SE d'ALICE
   - ✓ ALICE fournira un URL dans Monalisa pour suivre l'avancement du transfert

RESTABLISSEMENT D'UN SERVEUR SUITE A UNE PERTE DU NAMESPACE DANS UNE PARTITION SEPAREE

Sur un serveur A, si vous perdez le namespace et s'il était dans une partition séparée, il suffira de refaire les bons symlinks à partir des noms des fichiers pour restaurer le namespace.

Les étapes à suivre seront :

1. Prévenez ALICE (alice-lcg-task-force@cern.ch)
2. Arrêter le service xrootd sur A
3. Créer un nouveau namespace sur une partition séparée ou comme subdir dans une partition de data
4. Parcourir toutes les partitions et pour chaque fichier de data, refaire le symlink dans le namespace
5. Redémarrer le service xrootd sur A

RESTABLISSEMENT D'UN SERVEUR SUITE A UNE PERTE DU NAMESPACE QUI EST UN SUBDIR D'UNE PARTITION DE DATA

Sur un serveur A, si vous perdez le namespace et qu'il était dans un subdir d'une partition séparée, il suffira de refaire les bons symlinks pour restaurer le namespace. Quand ça arrive ?

1. Soit par 'rm –ef' du namespace : pas réaliste.
2. Soit la partition de data est perdue (donc les data + le namespace)

Le cas '1.' se résout comme dans le slide précédent. Nous considérons donc le cas '2.' Les étapes à suivre seront :

1. Prévenez ALICE (alice-lcg-task-force@cern.ch), arrêter le service xrootd sur A
2. Recréer le namespace dans la partition réparée ou dans une autre
3. Parcourir toutes les partitions et pour chaque fichier de data, refaire dans le namespace les symlinks
4. Collecter les GUIDs de tous les fichiers de A puis redémarrer xrootd sur A

RESTABLISSEMENT D'UN SERVEUR SUITE A UNE PERTE DU NAMESPACE QUI EST UN SUBDIR D'UNE PARTITION DE DATA (SUITE)

Ensuite, il faut faire restaurer par ALICE les data de la partition perdue dont on ne connaît pas les GUIDs des fichiers. Donc:

1. Collecter les GUIDs+la taille des fichiers de tous les autres serveurs
2. Envoyer ces informations à ALICE qui fera un diff avec la liste des GUIDs présents sur votre SE (grâce au catalogue ALICE) pour identifier les fichiers perdus. ALICE fera ensuite le transfert par xrootd:

- soit sur votre SE, s'il y a de l'espace disponible
- soit sur un autre SE d'ALICE
  - ✓ ALICE fournira un URL dans Monalisa pour suivre l'avancement du transfert

LIENS

- Decommissioning ALICE native xrootd servers and dealing with data loss http://lcg.in2p3.fr/wiki/index.php?title=ALICE_native_xrootd
- (http://alien2.cern.ch/index.php?option=com_content&view=article&id=56&Itemid=96)