Laboratoire d'Annecy-le-Vieux de Physique des Particules

Utilisation de CEPH au LAPP

Mathieu Gauthier-Lafaye, Philippe Séraphin

cnrs

UNIVERSITÉ SAVOIE MONT BLANC

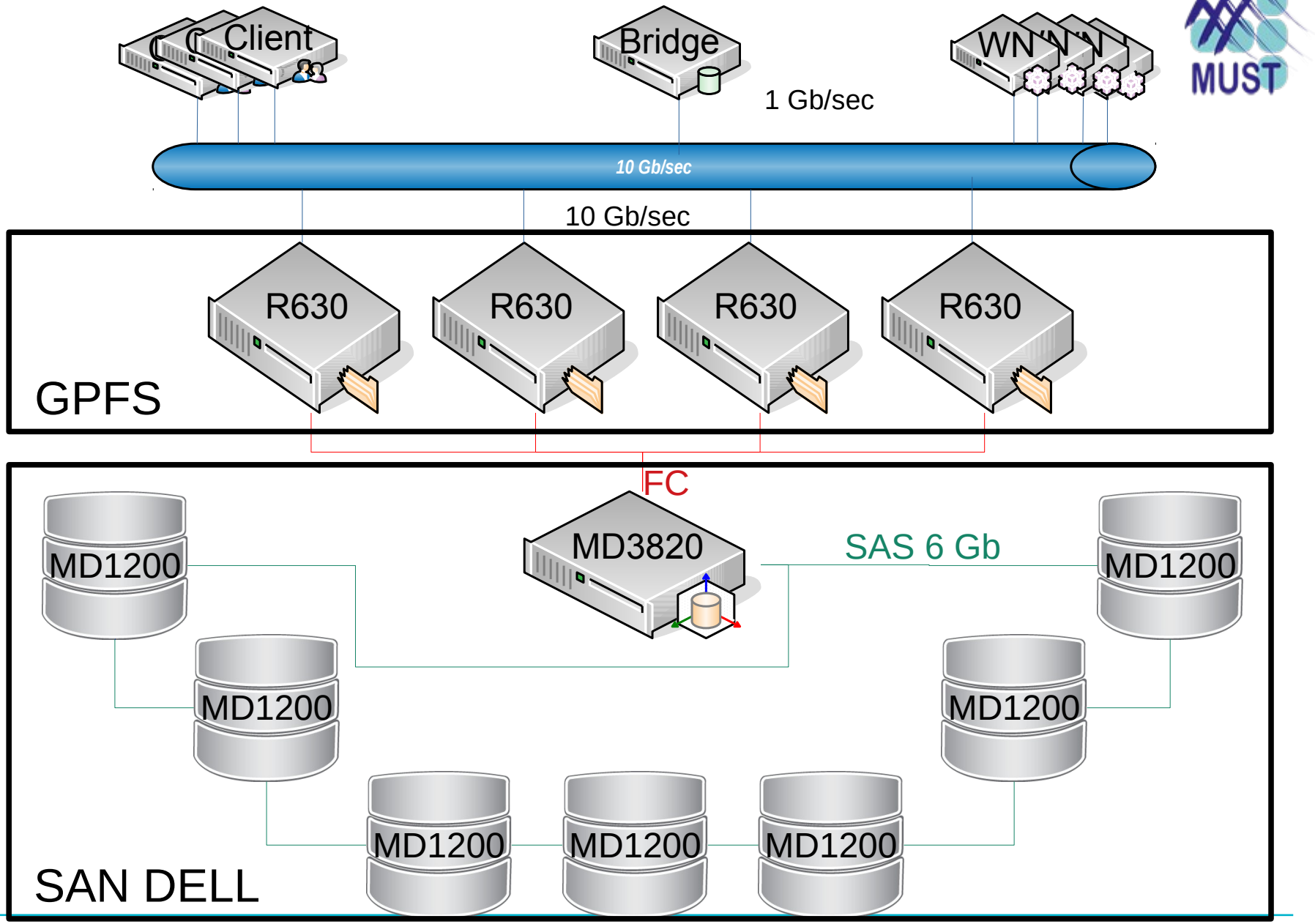
Sommaire

- **Historique**
- **De GPFS à CEPH**
- **Quelques utilisations de Ceph à l'IN2P3**
- **Futur de CEPH au sein de LCG (intérêt)**

LAPP : Historique

- Ceph pour le stockage des VMs
  - Depuis 2014
  - Répartition entre les 2 salles pour reprise d'activité plus rapide
  - Evolution prochaine, séparation du stockage Ceph et des hyperviseurs Promox
- GPFS (Système de fichier distribué d'IBM)
  - Depuis 2006
  - Stockage local pour MUST :
    - Espace partagé entre workers et job manager
    - Dépôt de paquets
    - Stockage données scientifiques locales
  - SAN de 240 To dédié
    - encore sous maintenance pour 3 ans

LAPP : GPFS


Client
Bridge
WN
1 Gb/sec
10 Gb/sec
10 Gb/sec
R630
R630
R630
R630
GPFS
FC
MD3820
SAS 6 Gb
MD1200
MD1200
MD1200
MD1200
MD1200
MD1200
MD1200
SAN DELL

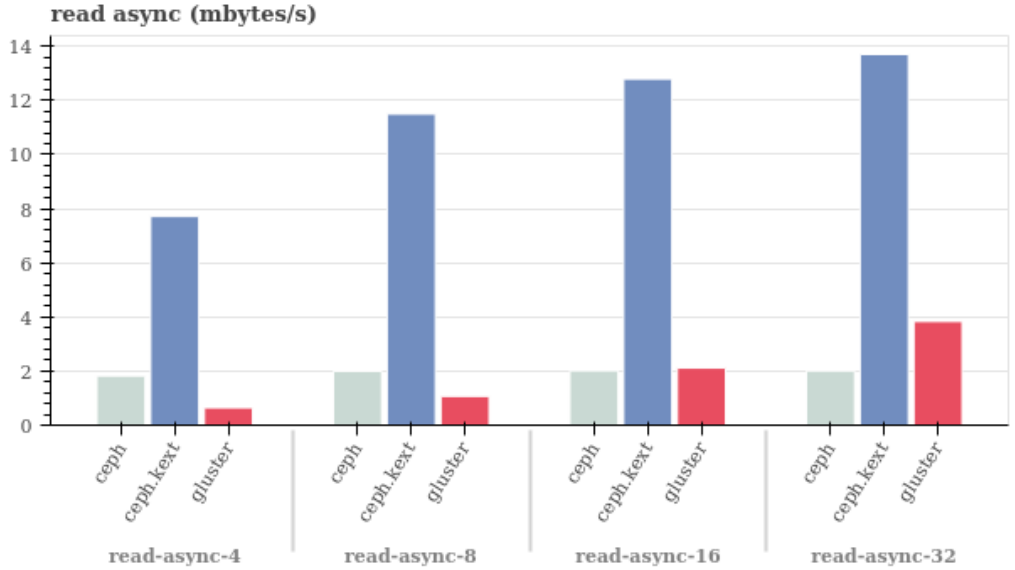
LAPP : De GPFS à CephFS

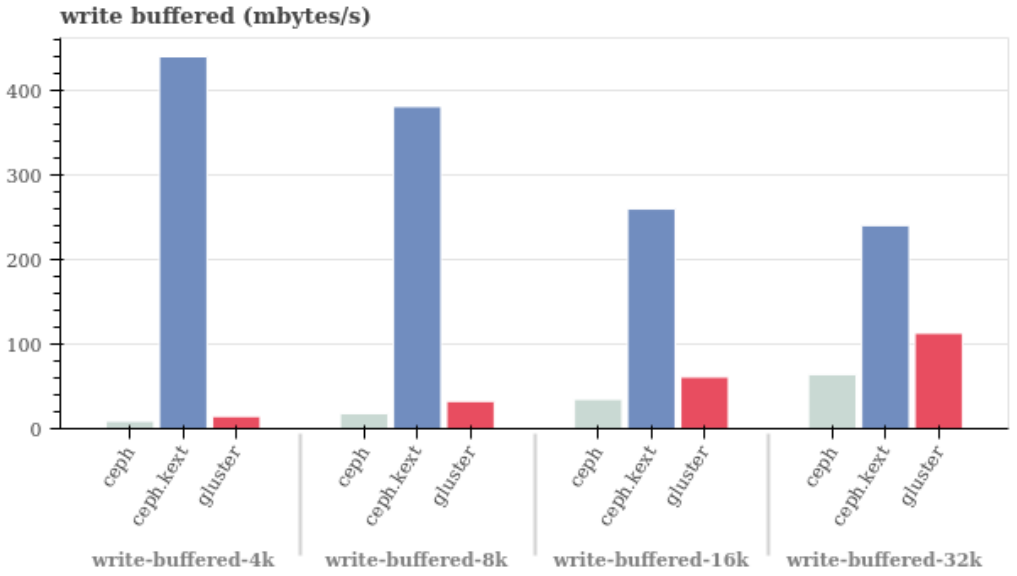
- Étude (Début fin 2017)
  - Problématique
    - plus de licences GPFS suite au changement de politique tarifaire d'IBM
    - réutilisation du SAN sous garantie
  - Comparaison bibliographique
    - GlusterFS,
    - OCFS,
    - GFS2…
  - Maquette de tests
    - GlusterFS
    - CephFS

LAPP : De GPFS à CephFS

- Choix de CephFS (mi 2018)
  - Fonctionnalités
    - Flexibilité très importante dans le placement des données (data+parité)
    - Gestion des quotas sur des répertoires
    - Droits d'accès sur les serveurs clients
    - Scale-out de la gestion des métadonnées
    - Support des snapshots (permet des sauvegardes cohérentes, permet d'offrir un accès à des versions plus anciennes)
  - Maîtrise de la résilience
  - Évolutivité
    - Souplesse d'extension et de renouvellement de l'infrastructure basée sur du matériel standard
  - Performances :
    - Correctes en clients multiples
    - Prometteuses via le module kernel (client) tout en restant nettement inférieures à GPFS
  - Expertise de l'équipe
  - Choix d'avenir
    - Optimisations typ. Cache Tiering, améliorations/évolutions…

LAPP : De GPFS à CephFS



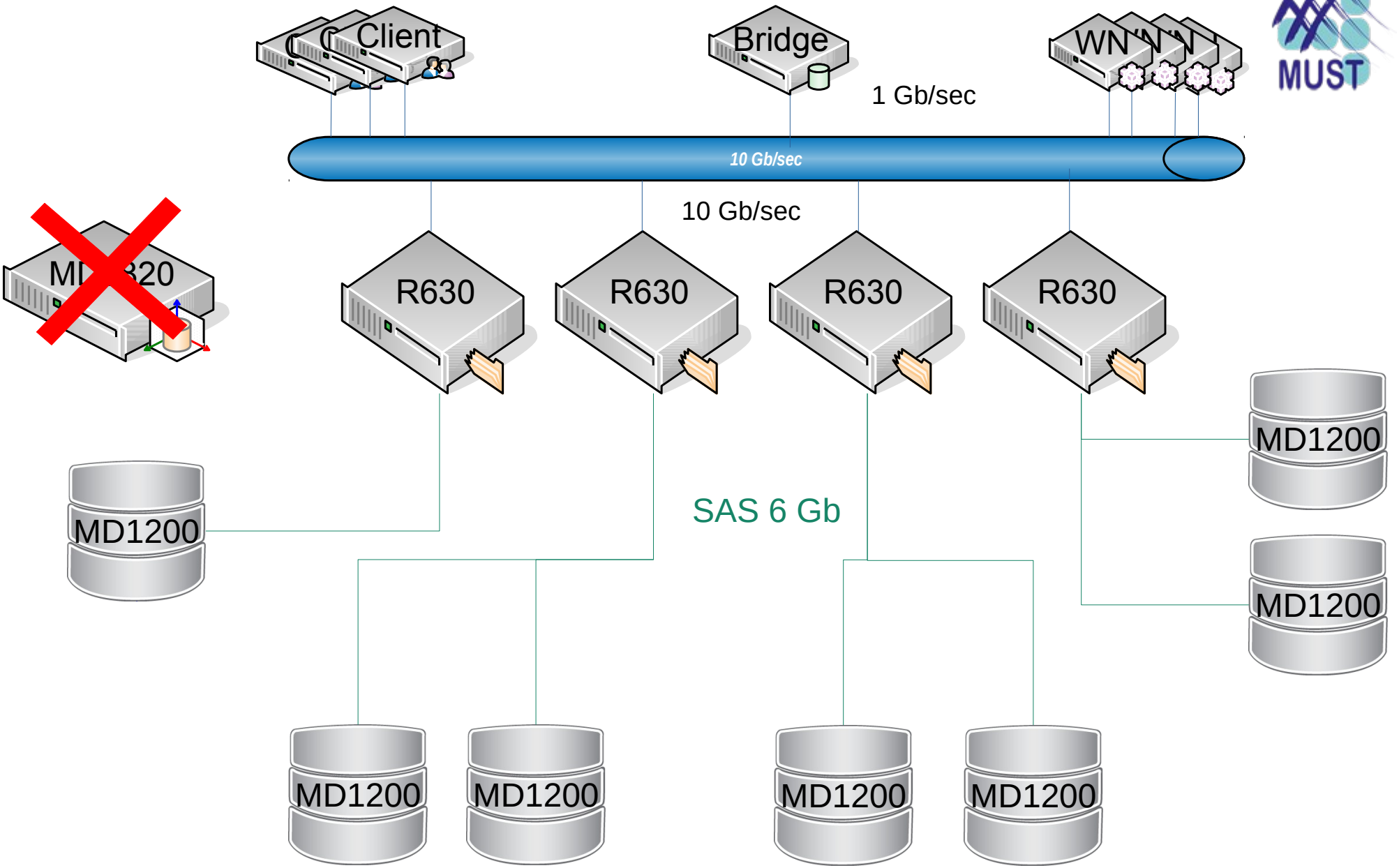


LAPP : De GPFS à CephFS

Infrastructure matérielle cible

- Reconfiguration du matériel du SAN actuel
  - ✗ Abandon du contrôleur SAN
  - ✓ 4 têtes SAN reconfigurées :
    - Ajout de SSD
    - Ajout de cartes RAID pour connecter les tiroirs
    - Disques des tiroirs en attachement direct

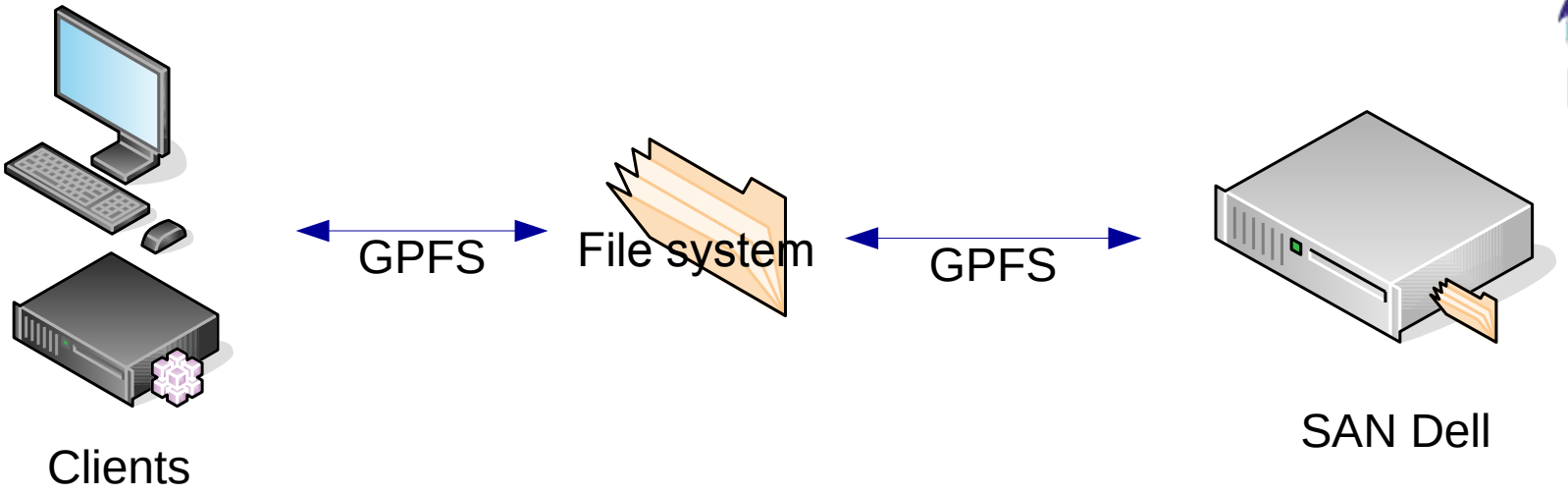
LAPP : CephFS


Client
Bridge
WN
1 Gb/sec
10 Gb/sec
10 Gb/sec
R630
R630
R630
R630
MD1200
MD1200
MD1200
MD1200
MD1200
MD1200
SAS 6 Gb

LAPP : De GPFS à CephFS

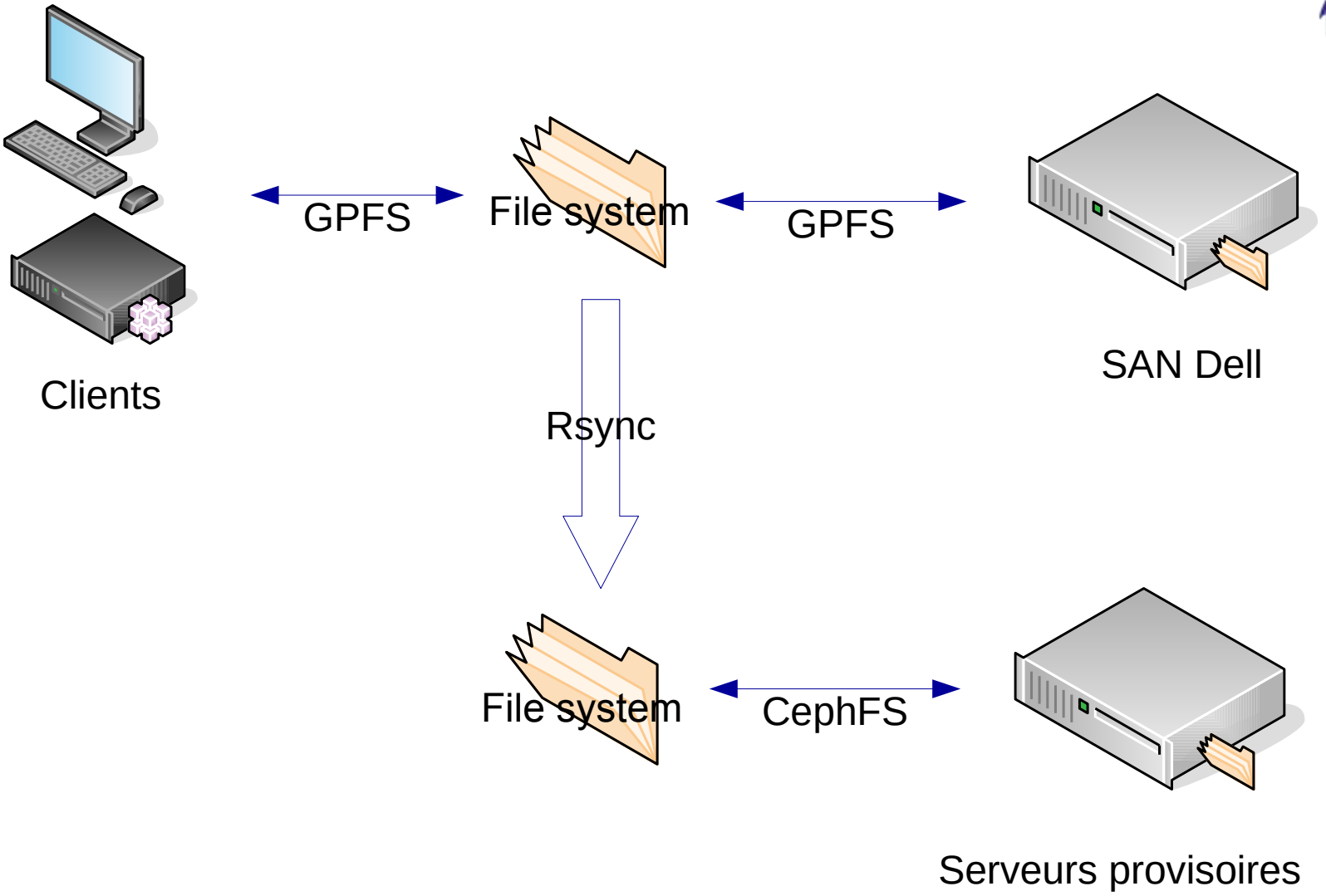
- Migration (depuis fin 2018)
  - Prévision réutilisation ancien SAN HP mais trop de problèmes
  - Configuration de migration
    - Utilisation des 4 R540 prévus pour l'évolution de Ceph RBD pour pouvoir effectuer la migration
    - « RAIN, redondance par noeud » Erasure coding en ($6^{+}3$)
    - 4 services de métadonnées (CEPHFS) dont 1 actif
    - 1 Pool de disques SAS pour les données
    - 1 Pool de disques SSD pour les métadonnées
  - En cours
  - Copie des différents FS de GPFS par rsync
    - Parallèle + finale

LAPP : De GPFS à CephFS



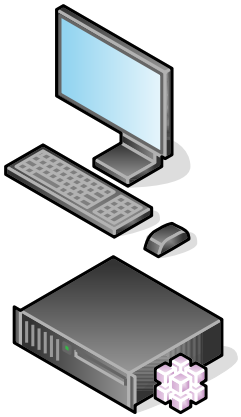
Situation initiale

LAPP : De GPFS à CephFS



Depuis novembre 2018 : Synchronisations continues

LAPP : De GPFS à CephFS

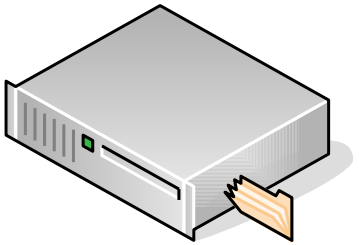


Clients

Reconfiguration

File system

GPFS

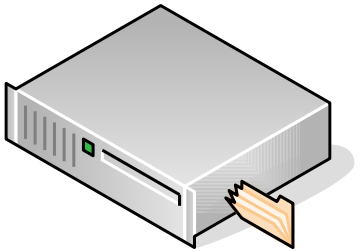


SAN Dell

Rsync finale du FS

File system

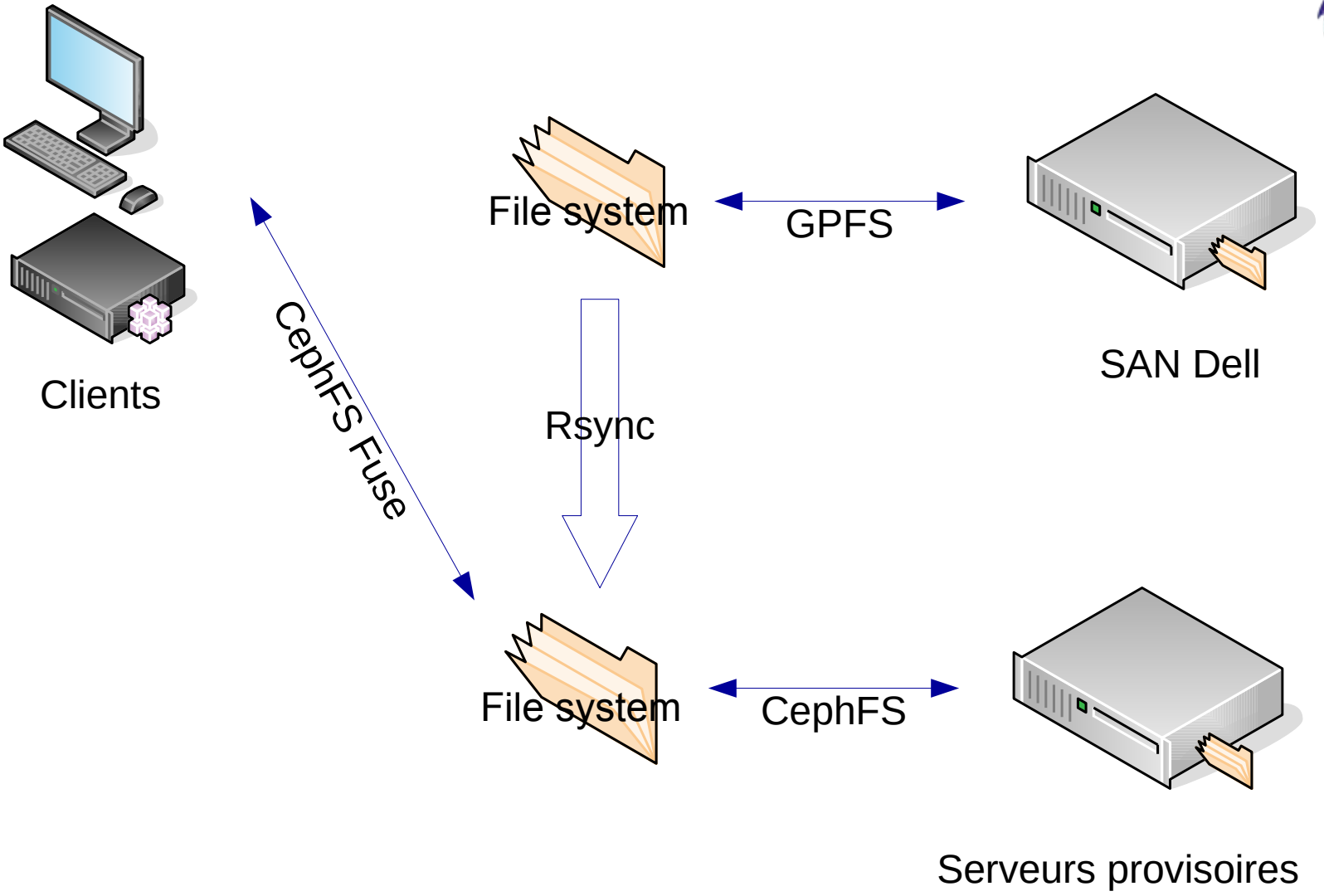
CephFS



Serveurs provisoires

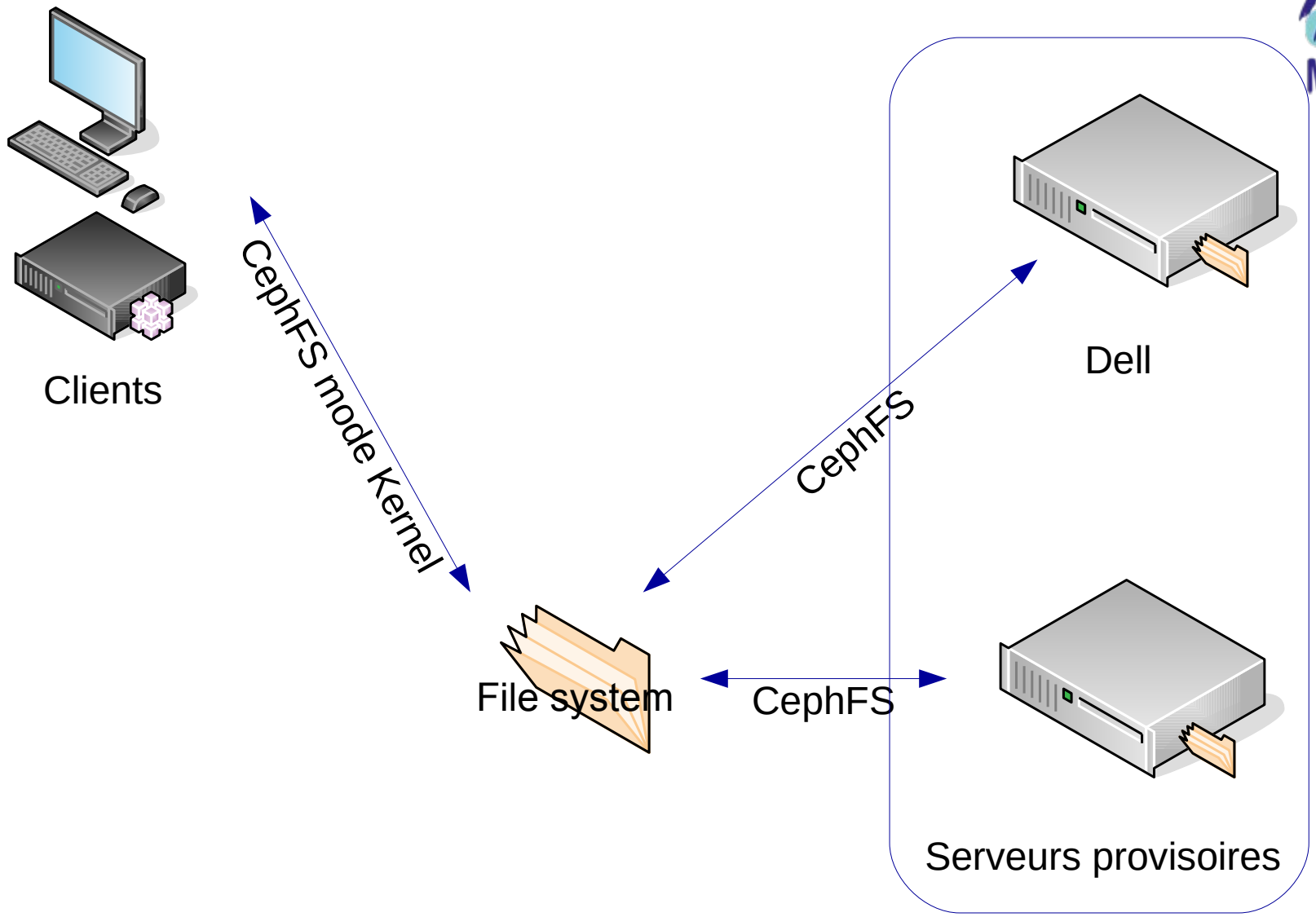
Jour J par File System : synchronisation finale et reconfiguration client

LAPP : De GPFS à CephFS



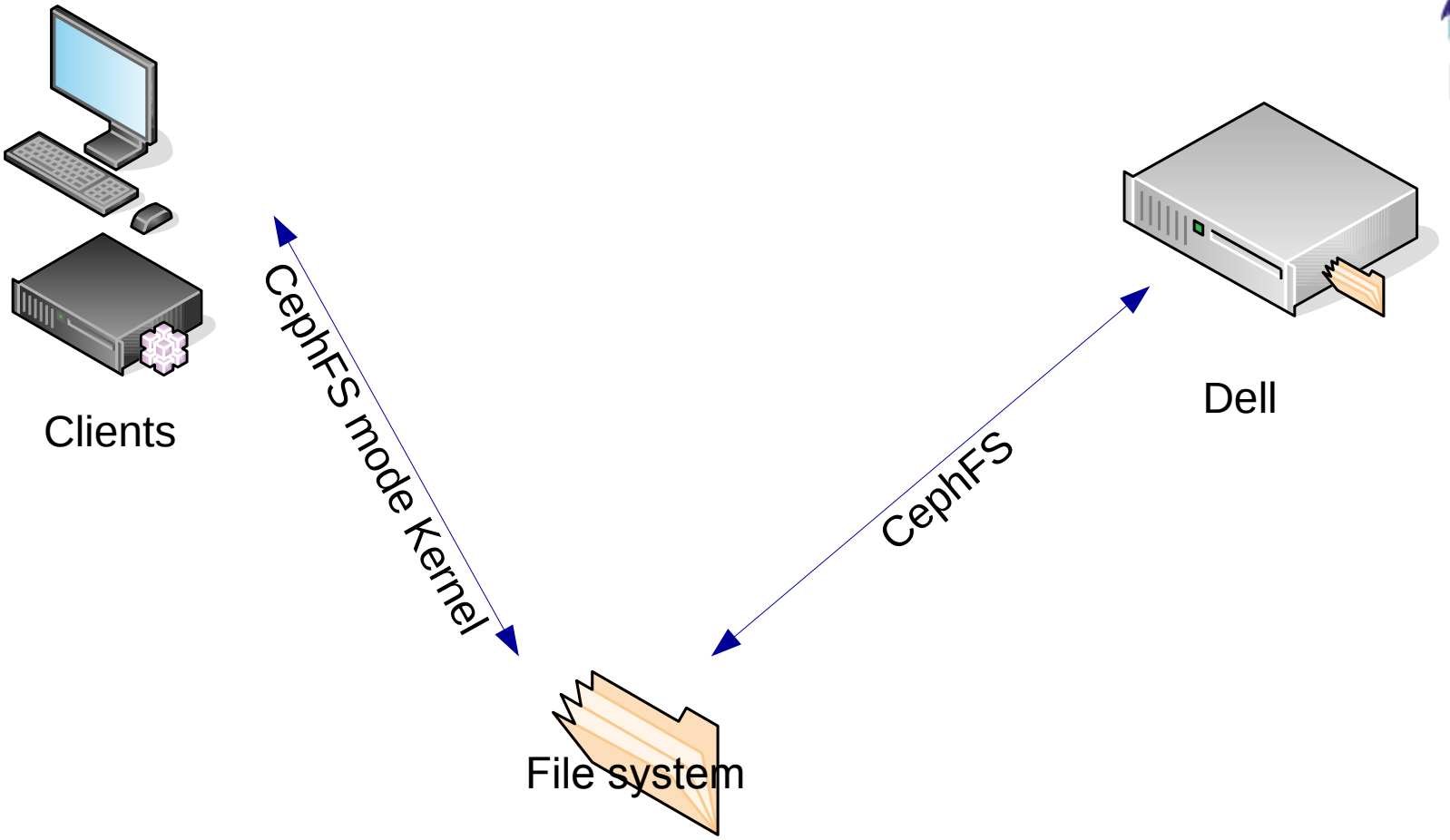
Jour J par File System : remontage sur CephFS

LAPP : De GPFS à CephFS



A la fin de la migration de tous les FS

LAPP : De GPFS à CephFS



Sutuation finale

Problèmes rencontrés

- Strays (égarés) :
  - Limite de 1 million avant : « Suppression du fichier impossible par manque d'espace libre »
  - en cas de suppression d'un hardlink, CephFS conserve une référence au fichier dans les strays jusqu'à ce qu'il y ait un accès à l'un des hardlinks restants (utilisation pour les dépôts)
  - Le même phénomène se produit avec les snapshots (utilisation d'une sorte de hardlink)
  - Problème important sur les zones home des jobs

Problèmes rencontrés

Performance FUSE

- Avec les petits fichiers : jusqu'à 30 fois plus lent que GPFS
- Débit en utilisation réelle sur des petits fichiers :
  - GPFS : 123.081 Mo/sec
  - CephFS Kernel : 34.791 Mo/sec
  - CephFS Fuse : 4.663 Mo/sec

- Actuellement, clients en mode FUSE car il faut avoir un Kernel compatible GPFS durant la migration
- 1 client en mode Kernel pour les sauvegardes au CC
- A terme, tous les clients en mode Kernel pour les performances

Ceph à l'IN2P3

- Utilisation de RBD au centre de calcul et dans beaucoup de laboratoires de l'IN2P3
- Au LPNHE : même approche pour le remplacement de GPFS
  - Choix initial de Gluster
    - Problème de performances
  - Configuration CEPH :
    - Erasure Coding 3+2
    - CEPHFS NFS Ganesha
    - Services CEPH dans des machines séparées
- Groupe de travail RI3 / RESINFO « systèmes de stockage distribué » :
  - Organisation d'une journée autour de CEPH en 2018
  - On souhaite organiser en 2019 des échanges par Visio autour d'une présentation de la communauté

Le futur

- Renouvellement du matériel de l'infrastructure CephFS
  - Étalements…
- Réactivation des snapshots post migration
- Utilisation de Ceph Objet + S3 en backend de SPARK (R&D LAL+IPHC)
- Étude en vue de l'utilisation de Ceph Objet en backend de xRootD (R&D LAL, intérêt du LAPP)

Questions ?