



## CEPH

Du laboratoire GREYC à Normandie Université en passant par l'Université Caen Normandie

Pierre BLONDEAU, Davy GIGAN

UFR Sciences UNICAEN,  
Laboratoire GREYC, CRNS UMR 6072, UNICAEN, ENSICAEN  
pierre.blondeau@unicaen.fr

4 Mai 2016

- 1 Besoins initiaux
- 2 Stockage résultats de recherche
- 3 Réutilisation pour la virtualisation
- 4 Évolutions
- 5 Université de Caen Normandie
- 6 COMUE Normandie Université

- Laboratoire de recherche en informatique
  - environ 200 membres
  - 7 équipes de recherche en informatique, électronique et automatique
  - environ 100 serveurs dont 6 dédiés aux calculs

- «Grosse volumetrie» de stockage partagée pour le calcul ( 12 To déjà pleins )
- Pas trop cher
- Plus flexible :
  - maintenance
  - dépendance matérielle ( raid propriétaires )
  - évolutivité
- Tolérant aux pannes
  - disques durs
  - contrôleurs, cm, ram, cpu, etc ...



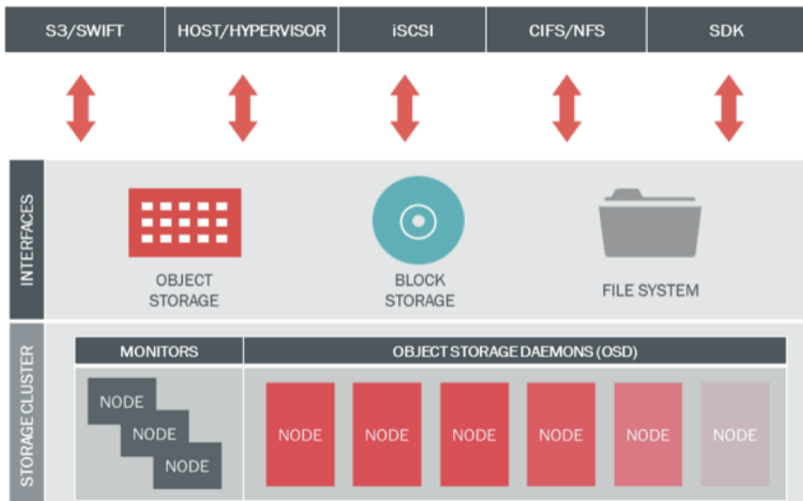
- Architecture distribuée
- SDS ( Software Defined Storage ) : indépendant matériel
- Réplication des données
- Plusieurs mode d'accès aux données
  - RBD : mode block ( type iSCSI )
  - CephFS : système de fichiers distribué ( type Lustre, GFS )
  - RadosGW : stockage object ( type Amazone S3, Swift )



# Besoins initiaux

## La solution CEPH

**Besoins initiaux**  
Stockage résultats de recherche  
Réutilisation pour la virtualisation  
Évolutions  
Université de Caen Normandie  
COMUE Normandie Université





- CephFS : Nos usagers aiment les systèmes de fichiers (ex : 20M pour 20T)
- RBD : un seul point d'accès aux données
- Rados : gestion compliquée pour notre usage ( authentification, espace utilisé, etc ... )
  
- Mise en place d'une plateforme CEPH avec CephFS sous Emperor en 2013.



- Achat de 3 serveurs avec 12 disques de 4To
  - Capacité brute : 130 To
  - Réplication à 3
  - Capacité utile : 43 To
- Mise à disposition avec CephFS : des contraintes et quelques problèmes
  - Nécessité d'avoir un noyau extrêmement récent ( mainline )
  - Pas de quota
  - Des fichiers Matlab corrompus ( Présence de blocs de 0 )
  - Perte de l'accès au système de fichier lors de la mise à jour en firefly
- Après récupération des données, passage en NFS over RBD ( Retour d'un SPOF )





- Infrastructure virtualisation
  - 8 hyperviseurs XEN
  - Environ 50 machines virtuelles
  - Système de fichiers des VMs ( domU ) sur chaque hyperviseur ( dom0 ) en LVM
  - Pas de chiffrement des données des serveurs
- Infrastructure de stockage pour les machines virtuelles
  - Deux serveurs de stockage prévus pour les machines virtuelles non utilisés
  - Tentative infructueuse avec iSCSI / DRBD / Pacemaker



- Mise en place de Ceph en mode RBD
  - Deux serveurs d'OSD avec une réplication à 2
    - recommandation à 3 pour éviter les problèmes en cas de corruption de données
  - Installation initiale en firefly
  - Chiffrement intégral des serveurs ( OS + OSD )
  - Couplage avec XEN
    - Mappage dynamique des images RBD
    - Configuration des VMs en CephFS
    - Migration à chaud des VMs entre les hyperviseurs
    - Utilisation de la fonction de clonage pour dupliquer des serveurs
- Essai concluant : achat d'une troisième machine
  - Capacité brute : 15 To
  - Réplication à 3
  - Capacité utile : 5 To



- Casse disques durs : facile à gérer.
- Atteinte du seuil `full_ratio_threshold` d'un OSD ( 85% par défaut )
  - Arrêt des écritures de l'ensemble des machines virtuelles
  - Régulé à court terme par augmentation de plusieurs seuils ( `nearfull_ratio` à 0.9 et `full_ratio` à 0.93 )
  - Rééquilibrage de l'emplacement des données
  - Régulé à long terme par l'augmentation globale de la capacité du cluster
- Problème de démarrage à froid d'un dom0 (coupure de courant)
  - Trop de demandes de lectures en parallèle
  - Le programme `udev` fait un timeout sur la détection de son disque dur
  - réglé par l'introduction d'un léger délai entre le démarrage de chaque VM



- Très récemment un bug dans le noyau officiel debian Jessie ( Bug CEPH #15302 )
  - Plus de montage du dossier de configuration xen au démarrage de nos dom0
  - kernel panic : le dom0 ne démarre plus
  - fix rapide : passage à la version précédente du noyau

- ajout d'une quatrième machine au cluster recherche : 152To bruts
  - la réplication reste à 3 : augmentation de capacité
- chiffrement de tous les OSD sans interruption de service
- Ajout d'un pool spécifique de SSD pour quelques applications virtualisée
  - Supervision, centralisation de logs
  - 1 SSD de 250 GB par noeud du cluster
- à venir
  - séparation des réseaux client network et cluster network
  - mise à jour en jewel
  - cache tier ssd au lieu des journaux SSD (c.f. présentation de Yann)
  - nouvel essai avec CephFS devenu officiellement stable pour le cluster recherche



- Intêret émis par la DSI de l'université pour la mise à disposition de stockage pour la recherche
  - Mise à profit l'expertise aqoise dans le laboratoire
  - Retour d'expérience de l'Université de Nantes ( Yann Dupont )
- Transfert de compétences ( 3 ingénieurs supplémentaires )
- Contraintes d'exploitation supplémentaires
  - Environnement de préproduction et production
  - Sauvegardes
  - Public plus large et pas forcément informaticien
- Echelle différente
  - 6 serveurs pour le cluster de production ( 260 To brut )
  - 3 serveurs pour la copie des snapshots ( 130 To brut )
  - 3 serveurs de préproduction ( 16 To )



- Réalisation de service en Python
- Les snapshots
  - Conservation sur le cluster de production
  - Création, mise à disposition et suppression automatisées
  - Définition d'un intervalle pour chaque image ( similaire cron )
  - Définition du nombre à conserver par image
- Les copies
  - Machine virtuelle avec un accès sur les deux clusters
  - Transfert des snapshots depuis le cluster de production vers le cluster de copie
  - Réalisée au fur et à mesure de la création des snapshots
  - Utilisation des fonctions d'import / export de ceph



- Projet de réalisation d'un «cloud souverain»
- 6 EPST de Normandie et un opérateur régional
- Utilisation de CEPH pour le stockage
- Echelle envisagée entre 1 et 10 Po
- Répartition sur plusieurs sites
  - Un seul cluster ?
  - Plusieurs clusters en copie croisée ?
- Place au tests !