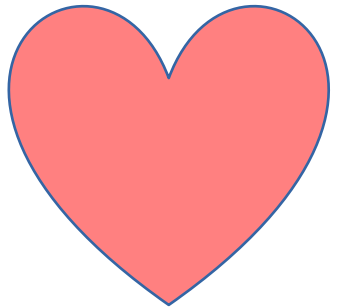
Ceph recherche (unlab-data?)

J-F. Guillaume,Y. Dupont
Équipe projet (Université de Nantes)

Réunion DOMA - 19 janv. 2020

Ceph à Nantes : Une vieille histoire… « Since 2012 »

ceph + UNIVERSITÉ DE NANTES =

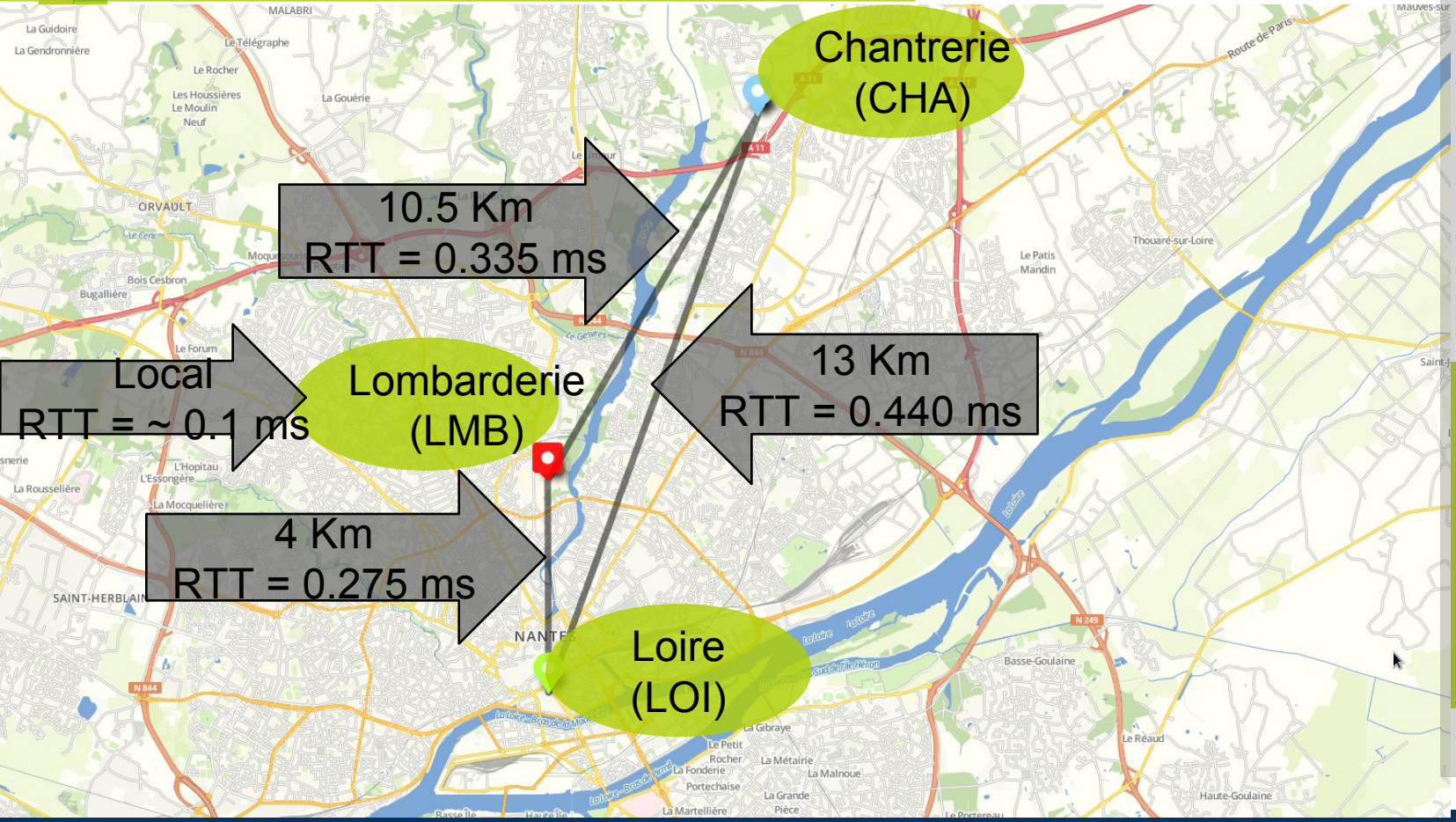


Système de stockage distribué complet, versatile et élastique.

Utilisé en production depuis 2012, brique support de nombreux projets de la DSIN.

9 clusters déployés, reposant sur une infrastructure commune :
42 serveurs standards déployés dans 3 sites sécurisés à Nantes.

Ceph à Nantes : Une vieille histoire… « Since 2012 »



Opéré par DSIN

3 points de prés

40 Gbit/s
entre sites, (mai

+ 3 Po bruts

Clusters Ceph et destination (usage * volume * Remarques)

Nom	Usage	Depuis	Taille Brute	Remarques
A	Tiers (labos de recherche)	07/2015	384 To	Données froides, sauf pour celles couplées à l'IaaS
B	Incubation	03/2013	32 To	Pré prod, tests...
C2	Serveurs DSIN	05/2014,2019	21 To	Full SSD depuis 2019
D3	Backups	01/2013 11/2015	1125 To	
E	Expérimental	06/2014	0	Arrêté pour l'instant
F	Data cloud	04/2016	452 To Filestore (+ 500 bluestore provisionné)	Données S3 essentiellement
G	Sécurité, logs archives video-surveillance	05/2016	22 To	
H	Data serveurs DSIN	05/2016	339 To	Mix Sas 7.2k, Sas 15k, SSD Mix use...
I	Iaas (Racines, Images)	05/2016	349 To	Volumétrique + caches SSD disponible

Cluster A DSIN

Fait partie des clusters dédiés et opérés par la DSIN.

Destiné à la recherche

Les composantes payent au To/Mois

Les volumes sont essentiellement disponibles via des serveurs SMB/NFS, ou RBD, pas de cephFS

Le cluster est plutôt balancé pour du volume (non massif) que de la performance

→ Plutôt utilisé pour stocker des données froides

Pas accessible facilement par le cluster de calcul du CCIPL

Pas d'administration fine possible

Autres raisons rationnelles… ou pas.

Il répond à beaucoup des besoins, mais pas à tous.Demandes spécifiques de certains laboratoires :

Investissement en matériel plutôt que payer du fonctionnement

Besoins spécifiques pour la recherche ; données jetables vs Pérennes

Éviter le transit de données permanentes entre équipements scientifiques, serveurs d'équipes, Centre de calcul : CephFS

Sans les clusters de la DSIN

CCIPL : Home (10 To) et Scratch (300 To, BEEGFS) mais pas de stockage pérenne

Moyens de stockage dans les labos

Des baies SAN/NAS

Des synology …. Multiples...

Des disques USB, des clés USB

Sans parachute et sans filet

Ceph-R comme « Recherche »

- Initiative LS2N (Labo des sciences du numérique de Nantes)
  - Achat 7 machines pour un cluster CEPH dédié
    - Stockage données de recherche
    - En particulier bases pour deep-learning
    - MAIS….
    - Nécessité de monter en compétence
    - Problème de gérer la complexité d'un cluster seul

ET …

Interactions LS2N/CCIPL

- Interactions LS2N/BiRD
- Interactions CCIPL/BiRD (rapprochement)

- Mise en place d'un cotech → naissance du projet ceph-recherche

Recherche

Un espace de stockage volumétrique, extensible et fiable, Accessible nativement depuis les labos ET les calculateurs
nécessite du débit réseau (pas d'omnipath)
Ne remplace pas le scratch, vient en complément

Parties prenantes :
LS2N : 6(+1) machines
BiRD : 3 machines en 2019, 6 à 9 à venir en 2020
CCIPL : au moins 1 machine début 2020
Ceisam : au moins 4 machines début 2020
DSIN : moyens humains et techniques (switches)
807 TiB bruts, 270 à 400 To utiles aujourd'hui.
va fortement augmenter en 2020 + partie backup.

Choix techniques particuliers

- Une administration ‘collégiale’ avec des cercles de compétence
- Accessible simultanément « nativement » depuis les labos et le CCIPL

Utilisation d’une VRF au niveau de l’Université (étanchéité)

Nombreux VLANs

Serveurs CEPH : 1 backend, 1 frontend (classique)

Chaque « labo » dispose de son vlan client (1 ceisam, 1 LS2N, 1 Bird, 1 CCIPL…)

IPv6 « Only »

Motivé par le nombre de serveurs du CCIPL en accès simultané

Machines uniquement le CAMPUS de l’UFR Sciences

Pour les soucis de latence

Côté CEPH : Nautilus ;

Deux types de stockages : Répliqués et Erasure code (5+4)

Des « pools » par labo pour RBD, des namespaces pour CephFS

Exemple

```
root@mon-1-r1-lmb:~# ceph df
RAW STORAGE:
    CLASS     SIZE        AVAIL       USED        RAW USED     %RAW USED
    hdd       807 TiB     610 TiB     197 TiB      197 TiB         24.45
    TOTAL     807 TiB     610 TiB     197 TiB      197 TiB         24.45

POOLS:
    POOL                   ID     STORED      OBJECTS     USED        %USED     MAX AVAIL
    RBD_3R_BiRD             5      10 MiB          16      33 MiB         0       161 TiB
    RBD_3R_CCIPL            8     3.9 TiB       1.02M      12 TiB      2.34       161 TiB
    RBD_3R_CEISAM          11         0 B           0         0 B         0       161 TiB
    RBD_3R_LS2N            14     3.7 TiB       1.05M      11 TiB      2.27       161 TiB
    CEPHFS_3R_DATA         15     1.5 TiB     390.80k     4.5 TiB      0.91       161 TiB
    CEPHFS_3R_METADATA     16     258 MiB          85     582 MiB         0       161 TiB
    RBD_EC_BiRD            18      16 KiB           1     384 KiB         0       323 TiB
    RBD_EC_CCIPL           19         0 B           0         0 B         0       323 TiB
    RBD_EC_CEISAM          20         0 B           0         0 B         0       323 TiB
    RBD_EC_LS2N            21         0 B           0         0 B         0       323 TiB
    CEPHFS_EC_DATA         22      78 TiB      34.95M     121 TiB     20.05       323 TiB
    CEPHFS_EC_METADATA     23     6.9 GiB     773.38k     7.3 GiB         0       161 TiB
```

Exemple

```
Last login: Wed Jan 15 09:53:15 2020 from 2001:660:7220:0:b8be:bbff:fe71:1101
root@mon-1-r1-lmb:~# ceph osd tree
ID  CLASS WEIGHT    TYPE NAME                STATUS REWEIGHT PRI-AFF
 -1       806.91406 root default
 -5       806.91406     datacenter lmb
 -8       230.54688         room ccipl
 -7       115.27344             rack silverado
 -6       115.27344                 host osd-6
  4   hdd   7.20459                     osd.4       up  1.00000 1.00000
  6   hdd   7.20459                     osd.6       up  1.00000 1.00000
…
-10       115.27344             rack stonehedge
 -9       115.27344                 host osd-5
  3   hdd   7.20459                     osd.3       up  1.00000 1.00000
  8   hdd   7.20459                     osd.8       up  1.00000 1.00000
…
576.36719         room datacenter
 -3       115.27344             rack andes
 -2       115.27344                 host osd-3
  1   hdd   7.20459                     osd.1       up  1.00000 1.00000
  7   hdd   7.20459                     osd.7       up  1.00000 1.00000
 13   hdd   7.20459                     osd.13      up  1.00000 1.00000
```

Exemple

```
[root@budbud002 ~]# df
Sys. de fichiers
shipsterns.opa:/home     10735331296   6747367968   3987963328  63% /home
beegfs_nodev            292987914240 138908550144 154079364096 48% /scratch
[fdb0:cafe:d0d0:ceff:b8be:bbff:fe56:101],[fdb0:cafe:d0d0:ceff:b8be:bbff:fe54:101],
[fdb0:cafe:d0d0:ceff:b8be:bbff:fe55:101]:/CEISAM_EC
                         430641655808  83865186304 346776469504  20% /mnt/cephfs-ec
tmpfs             13193572           0     13193572   0% /run/user/0
[root@budbud002 ~]#
```

En détail...

LS2N

Données deep Learning

VM de calcul

Ceisam : UMR chimie

Besoin de stocker de gros volumes (molécules de chimie) + calcul CCIPL

BiRD (Plateforme Bio-Informatiques)

Génomes

Calculs sur plateforme dédiée (VM) ET CCIPL

CCIPL (Centre de Calcul Intensif des Pays de la Loire) :

Accès aux espaces depuis le cluster (270 nœuds, ~4500 cœurs, réseau OPA/IB/10G)

- Et autres à venir (à ouvrir après phase pilote)

« Backup »

Pour sauvegarder les données non « sacrifiables ».

Projet 2020. (comme ce slide).

Merci !

Questions ?

Crédits : Opencliparts / Openstreetmap / Ceph.com