CEPH
3 ans après
(Bilan & Perspectives)

Yann Dupont
DSIN Université de Nantes

www.univ-nantes.fr

19 Novembre 2015

UNIVERSITÉ DE NANTES

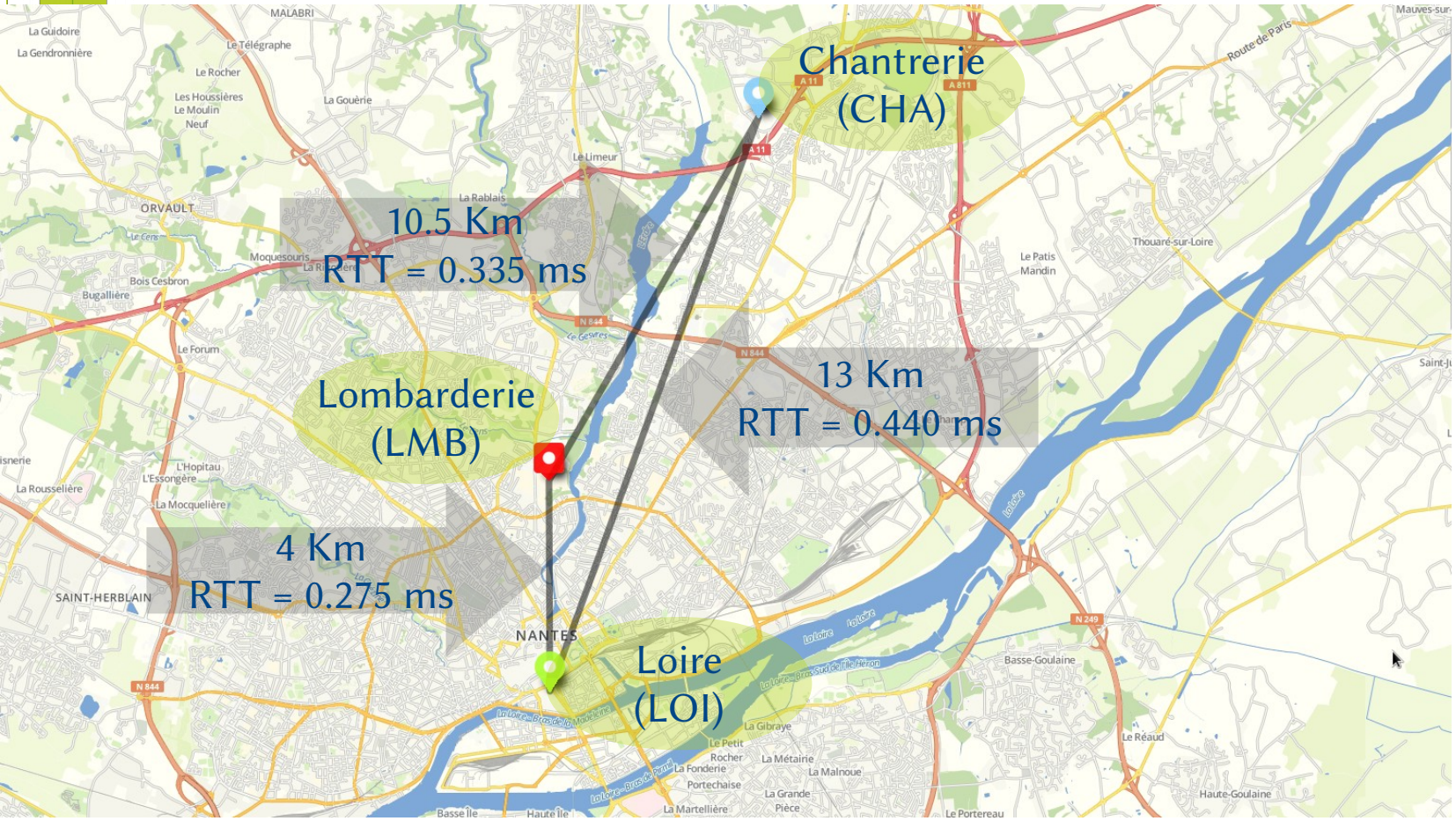
Ceph ?

Ceph est UN système complet de stockage distribué
(**Stockage Objet, Bloc, Système de fichiers**)

Qui fonctionne bien.

Pour une présentation plus complète de Ceph :
cf Jres 2013, Meetup Openstack Paris, Ceph Days Paris 2014, Bio Ouest 2014...

Topologie réseau à Nantes



3 points de présence

Fibre noire (Allumée par la DSIN)

40 Gbit/s entre sites

1 Datacenter

Mise en œuvre optimale

- 2012 : aucun guide de mise en œuvre...
  - Expériences heureuses puis malheureuses !
  - Erreurs de débutant...
  - Peinture pas encore sèche !
- Partage d'expérience
  - Liste de diffusion
  - Canal IRC
  - Meetups, Ceph Days, Ceph Breizh
  - Jres, OpenStack summit
  - Blogs
- 2015 : Plus simple, assez commun

Historique

	47 Préversions (!)	Tests	Début 2012
	Argonaut (v 0.48)	Pré-prod Prod	03 / 07 / 2012
	Bobtail (v 0.56)		01 / 01 / 2013
	Cuttlefish (v 0.61)	Multi clusters	07 / 05 / 2013
	Dumpling (v 0.67) LTS	Archi dédiée	14 / 08 / 2013
	Emperor (v0.72)		09 / 11 / 2013
Majeure	Firefly (v0.80) LTS	Augmentation espace	07 / 05 / 2014
	Giant (v0.87)		29 / 10 / 2014
	Hammer (v0.94) LTS	Tiering/Erasure coding	07/ 04 / 2015
	Infernalis (v9.2.y)		06/ 11 / 2015
renumérotation	Jewel (v10.2.y) LTS		??/ ?? / 2016

Tous les 6 mois

Tous les 3 mois

Tous les 6 mois

Stratégie 2012

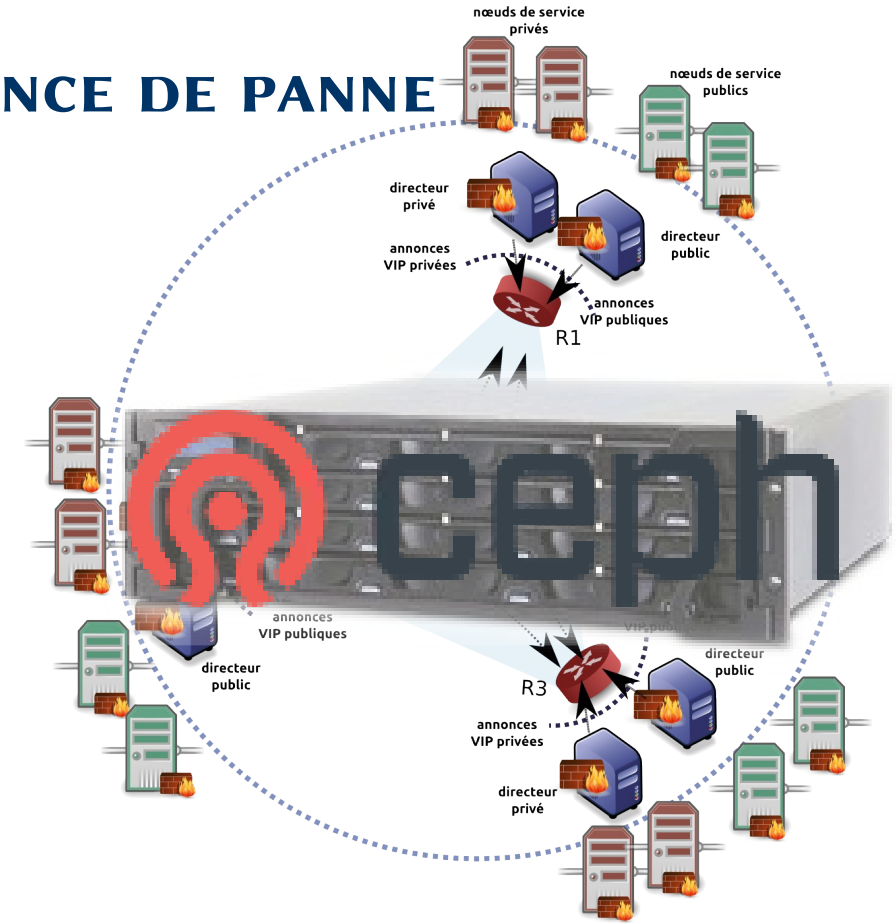
Tester CEPH

Valider la plate-forme

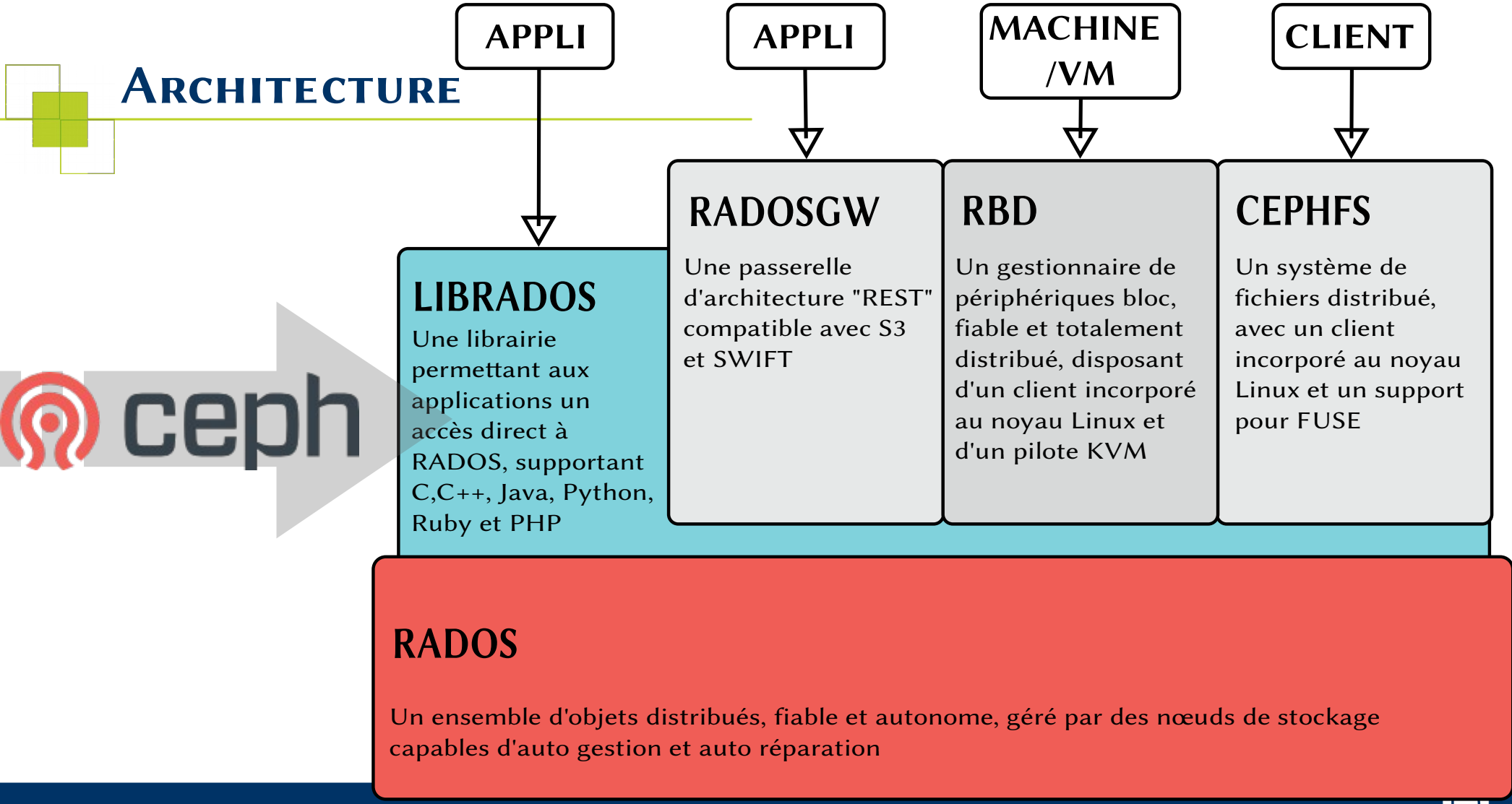
Démarrer des sauvegardes répliquées

Stockage distribué à tolérance de panne

Après beaucoup d'essais (autres solutions),
CEPH est validé
Début 2012



Architecture



Architecture de Ceph

Software Defined Storage

Nombre impair

MON

Vérifie le bon état du cluster
Assure la communication initiale avec les clients
Vérifie les droits d'accès

Nombre impair

MDS

Gère les méta données (uniquement pour CephFS)

OSD

Stocke les objets (généralement des fichiers de 4 Mo sur filesystem local (XFS))
Communique avec les clients

Bugs & déceptions (2012)

Cluster

*Peu d'*OSD très volumineux (baies SAN) :
1 unique GROS cluster CEPH
Journaux sur disques : mauvaise idée, Leeeent !

Kernel

Bug mémoire virtuelle, crash fréquent des OSD

Filesystems

BTRFS (Lent, se fige, plante...)
XFS (1 Bug sévère et dévastateur) (Vite corrigé)

Cluster presque plein + bugs + effet domino à la reconstruction = « On casse l'incassable »

Un désastre peut arriver
Choix architecturaux pour atténuer celà

Ne pas mettre tous ses œufs dans le même panier

Une volumétrie importante

Des besoins différents
Des administrateurs différents

Démarrer plusieurs clusters CEPH mais de façon virtuelle

(Utilisation de virtualisation, conteneurs LXC)

Stratégie 2013

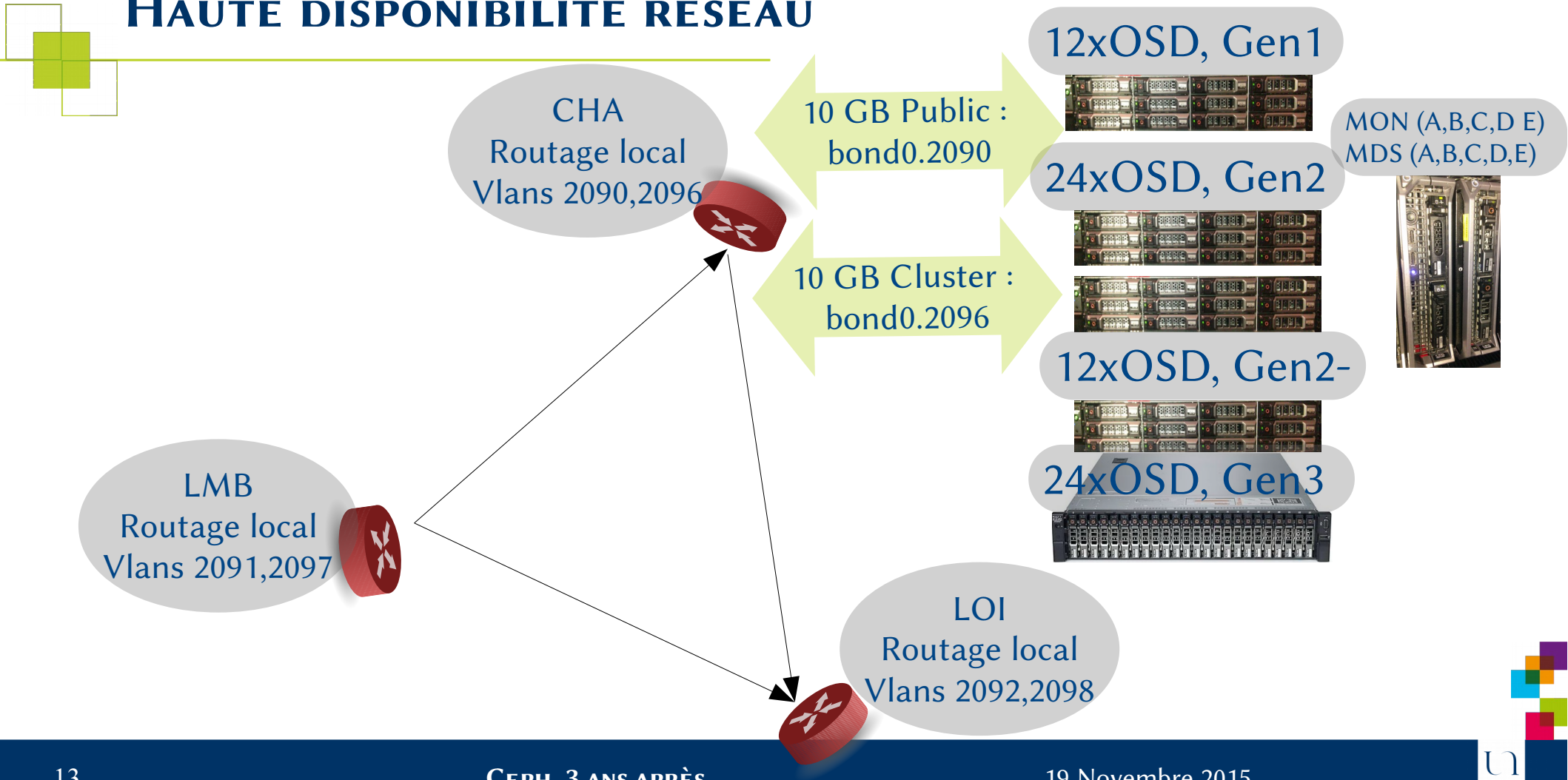
Plate-formes dédiées

Plusieurs clusters distincts
Données placées intelligemment

Virtualisation des clusters

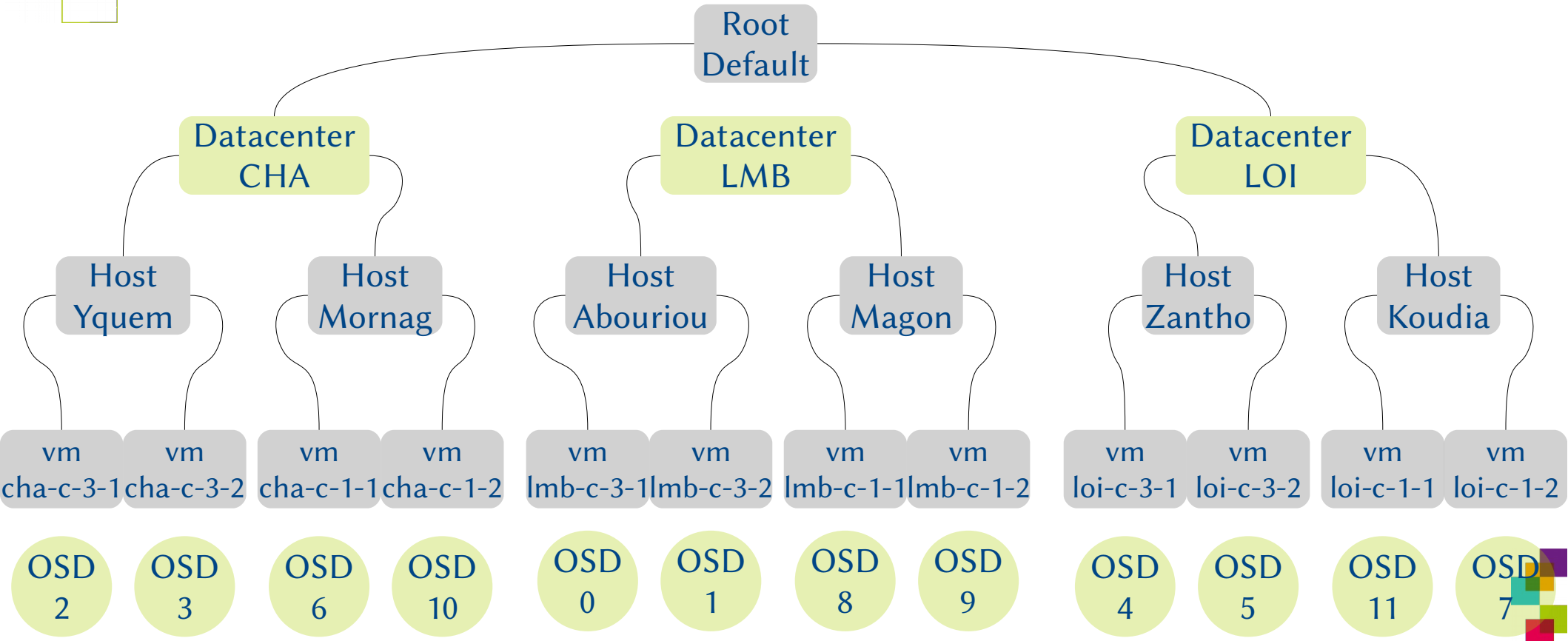
Démarrage production

Haute disponibilité réseau



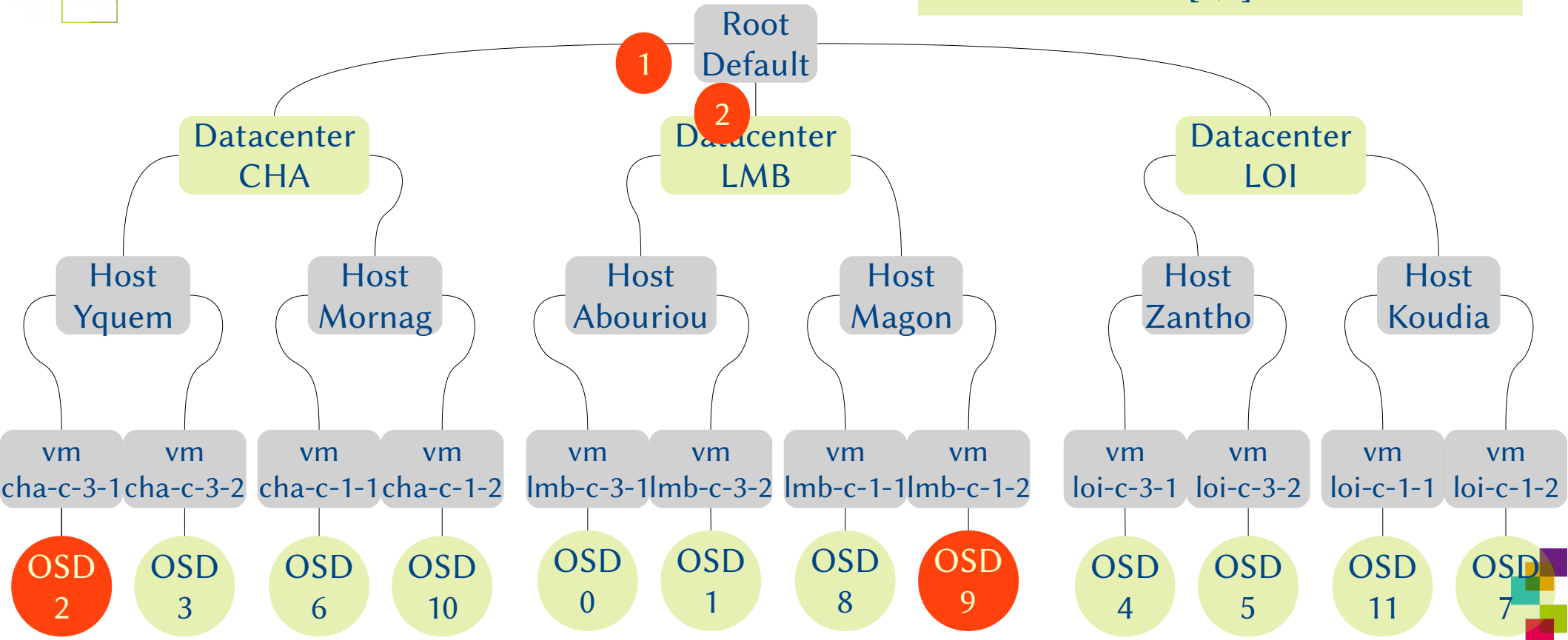
Règles Crush, cluster C

step take default
step chooseleaf firstn 0 type ***datacenter***
step emit

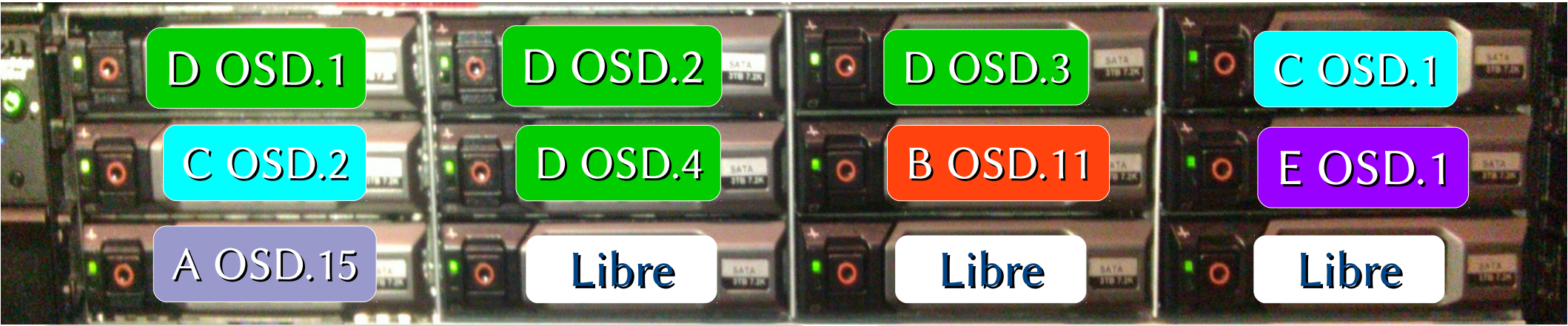


Règles Crush, cluster C

Pool 3 : mirrors (***2 copies***)
PG dump 3.1ed
[2,9]



Déploiements (fin 2013)



1 OSD = 1 Disk = 1 LXC
plus de RAID Hardware
2 SSD SLC partagés (Journaux)
12 OSD / machines
Capacité Brute : 48 To

Déploiements symétriques
par plaque.

3 machines identiques.

Stratégie 2014

Augmenter la volumétrie,
Replication x3

Profils d'usages,
selon cluster, génération de machines

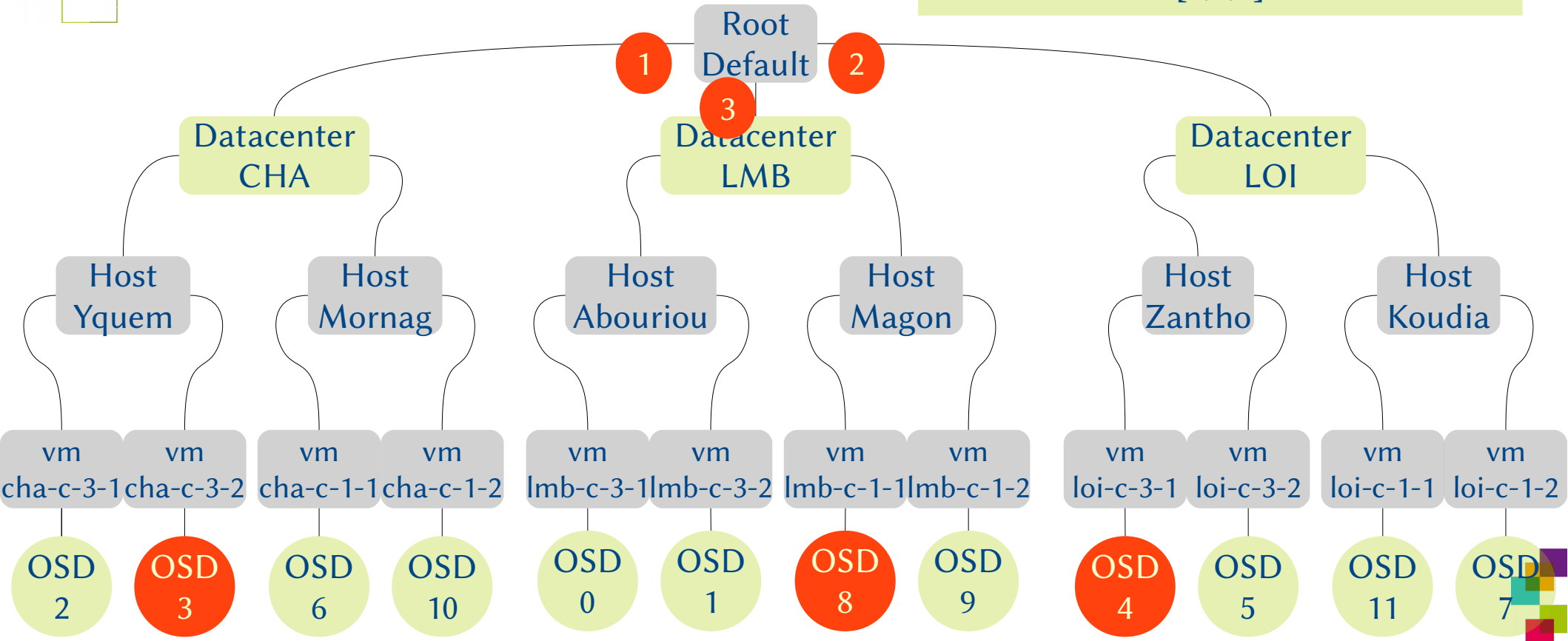
Démarrage des tests IaaS

Commencer à se défaire
des vieilles baies SAN hors garanties

Simplifier les déploiements
(plus de 200 LXC)

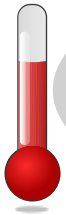
Règles Crush, cluster C

Pool 6 : OS templates (*3 copies*)
PG dump 6.3fb
[3,4,8]

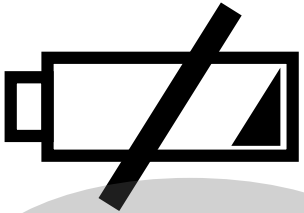


Tolérance à la panne : Testé !

(Fibres!)
CHA



Panne climatiseur
& arrêt d'urgence
LMB



Panne d'onduleur
et disjoncteur
LOI

Les problèmes sur machine physique sont rares
En général, une salle machine complète est impactée
3 répliquas permettent d'éviter une reconstruction

Clusters Ceph et destination

Nom	Usage	Depuis	Taille	Version
A	Tiers (labos de recherche)	07/2015	Pourrait être très importante	Stable LTS (Hammer)
B	Incubation	03/2013	Petit	Stable Hammer→Infernalis
C	Production vSAN	05/2014	Moyenne	Stable LTS (Firefly)
D D2	Backups	01/2013 11/2015	Très importante	Stable LTS : Firefly→ D2 : Hammer
E	Expérimental	06/2014	Petit	Versions de développement
F	IaaS	Q4/2015	Moyen à important	Stable LTS (Firefly)
G	Data / Home dirs	Q2/2016	Pourrait être très importante	Stable LTS (Jewel?)

Stratégie 2015

Augmenter la volumétrie...

Racine de serveurs KVM en prod,
Serveurs logs, miroirs, ucloud

Se défaire de toutes les baies SAN
Volumétriques

Augmenter performance
Ajout de caches (local datacenter)
baisser les latences
Augmenter les débits (4x10 Gb)

Capacité globale

Gén	Destination	Nb	Taille	Version	Volume
1	Mixé	3	12 x 3 To 2 SSD MLC	R720xd Juin 2013	108
2	Mixé/Perf	6	12 x 4 To 2 SSD SLC	R720xd, Nov 2013 Juin 2014	288
2-- 2-	Stockage de masse, BUDGET	3 3	12x2 (utilisés), 0SSD 12x4 (utilisés), 0SSD	R720xd Juin 2014	72 144
3	Perf lecture	3	24x1 (2,5" sas 10k) pas de SSD.	R720xd Nov 2014	72
4	Stockage de masse	9	16 x 8 To Pas de SSD.	R730xd Nov 2015	1152 (!)
4c	PERF !! Cache, écriture, latence.	2	10x400 Gb SSD write intensive	R630 Nov 2015	0

Évolution (Mi 2015)

1 OSD = 1 Disk
1 seul LXC par cluster
Pas de SSD. Journal en tête des disques.
24 OSD / machines
Capacité Brute : ~ 24To

«Seulement» 5 LXC max/machine.
Beaucoup d'axes. Orientation performances en lecture (Racines IaaS)

Cluster C

```
GLOBAL:
    SIZE       AVAIL      RAW USED     %RAW USED
    43483G     11839G       31644G         72.77
POOLS:
    NAME              ID     USED       %USED     MAX AVAIL     OBJECTS
    data              0          0          0         5055G           0
    metadata          1       9470          0         5055G          21
    rbd               2          0          0         5055G           2
    mirrors           3       3038G       6.99         5055G      785006
    cloud-perso       4        817G       1.88         3370G      232739
    openstack         5        301G       0.69         5055G       79617
    os-patrons        6        169G       0.39         3370G       45347
    os-dsin-prv       7       2347G       5.40         3370G      605041
    os-dsin-pub       8          0          0         3370G           0
    logs              9       2460G       5.66         3370G      637182
    os-dsin-sig       10      3368M          0         3370G        1189
    data-dsin         11      2200G       5.06         3370G      564697
    data-tiers        12       180G       0.41         3370G       46167
    opennebula        13     26555M       0.06         3370G        6734
    esxi-backup       14      8973M       0.02         3370G        2913
```

14 ⟶ 43 To : Replicat à 3

Mirroirs Ubuntu, Debian...

Racines de machines virtuelles
Dédup (COW)

Vers cluster F

Snapshots , layering

```
ceph-mon-lmb-C-1:~#
rbd info os-dsin-prv/lubuntu-xs2
rbd image 'lubuntu-xs2':
size 20000 MB in 5000 objects
order 22 (4096 kB objects)
block_name_prefix: rbd_data.3e1292eb141f2
format: 2
features: layering

parent: os-patrons/
lubuntu-2014-10-beta1@20140922

overlap: 20000 MB
```

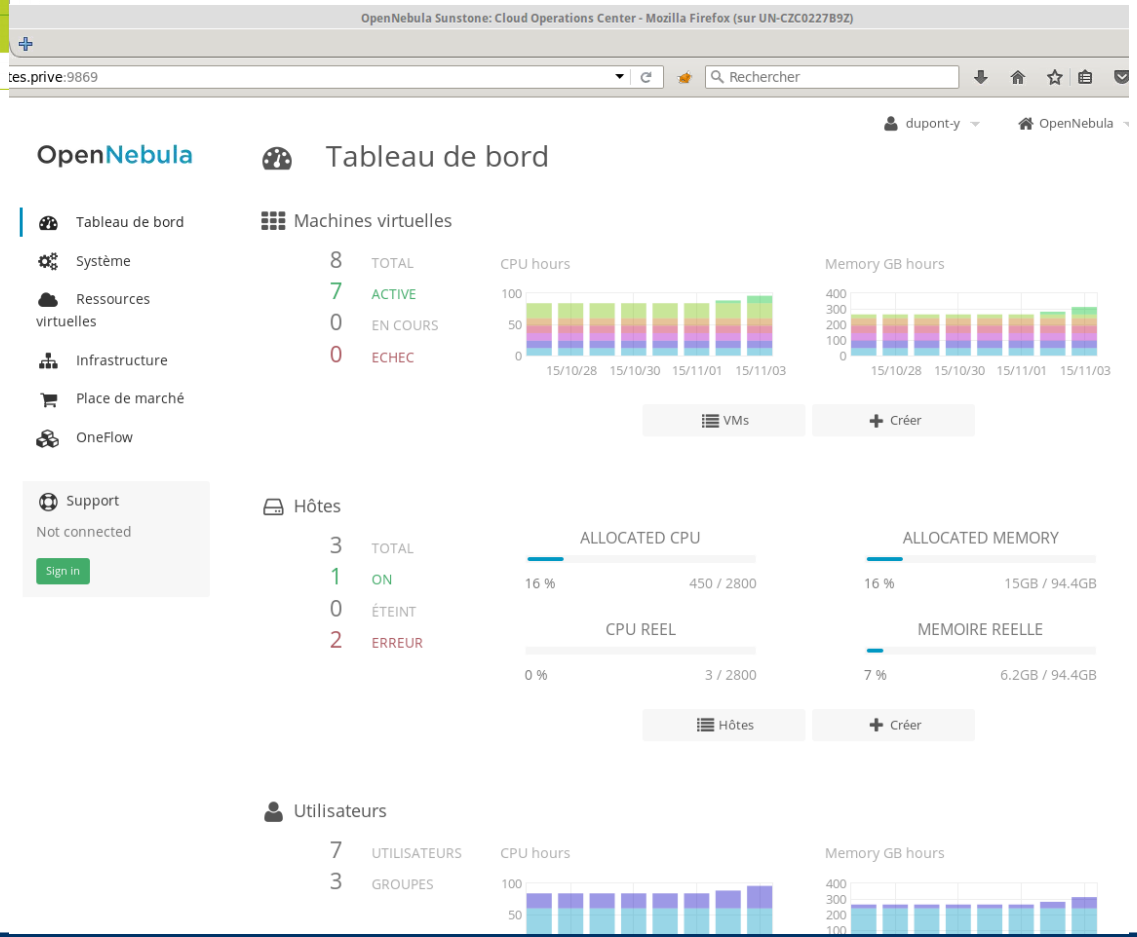
```
ceph-mon-lmb-C-1:~#
rbd snap ls os-dsin-prv/lubuntu-xs2

SNAPID NAME                    SIZE
    9 apres-upgrade-wily    20000 MB
    6 avant-upgrade-vivid   20000 MB
    7 avant-upgrade-vivid-b 20000 MB
    8 avant-upgrade-willy   20000 MB
```

Usage direct avec KVM (librbd)

```
<disk type='network' device='disk'>
    <driver name='qemu' type='raw' cache='writeback' iothread='8'/>
    <auth username='dsin-prv'>
      <secret type='ceph' uuid='d0a44325-05cb-4bc2-9d95-53af647fe78e'/>
    </auth>
    <source protocol='rbd' name='os-dsin-prv/lubuntu-xs2'>
      <host name='172.20.106.85' port='6789'/>
      <host name='172.20.107.85' port='6789'/>
      <host name='172.20.108.85' port='6789'/>
    </source>
    <target dev='vda' bus='virtio'/>
    <address type='pci' domain='0x0000' bus='0x00' slot='0x05' function='0x0'/>
  </disk>
```

OpenNebula (IaaS)



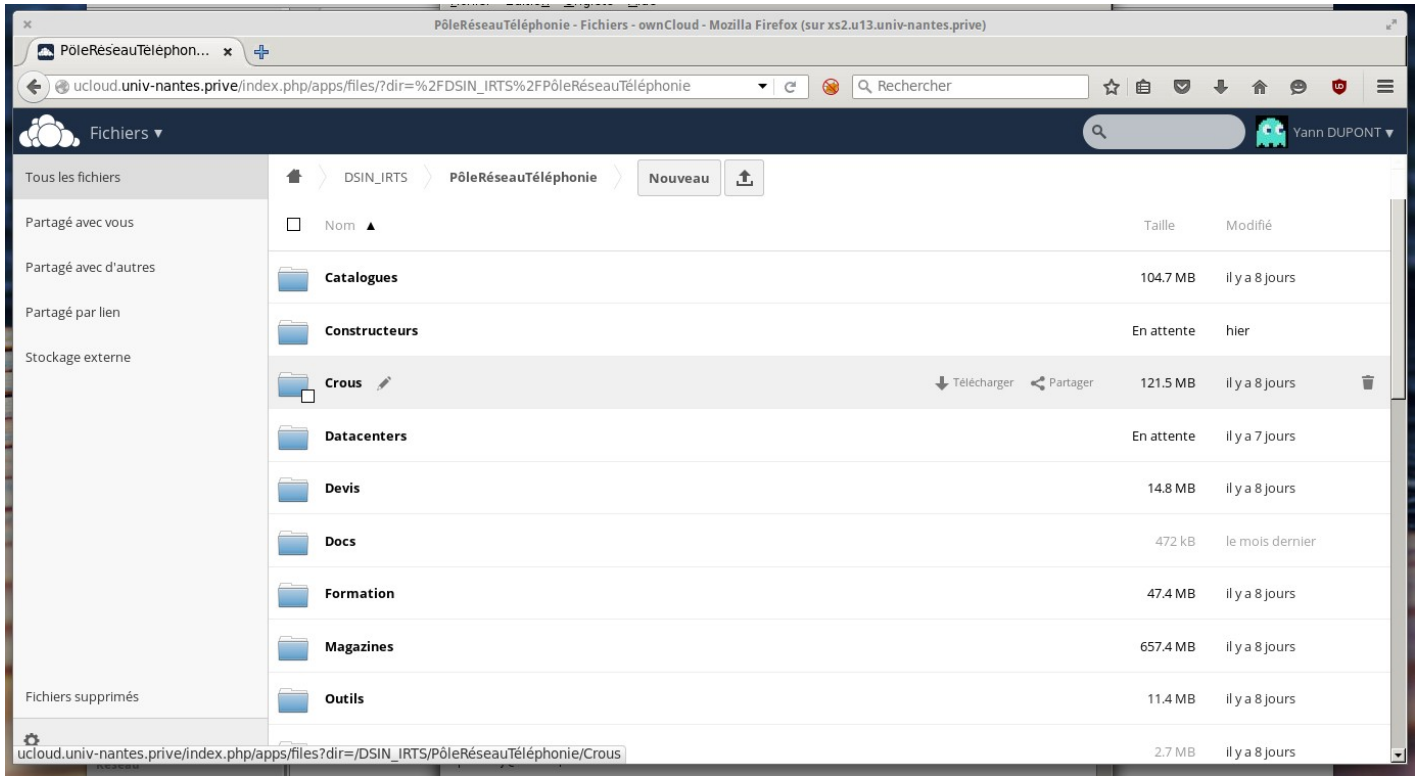
OpenNebula dans un Premier temps

Openstack éventuellement Comme seconde offre.

Les deux reposent sur Les mêmes bases

(Libvirt + KVM + Ceph)

OwnCloud (sync & share)



Actuellement, utilisation
En mode bloc.

À terme,
multiple frontaux +
Stockage objet

Cluster D

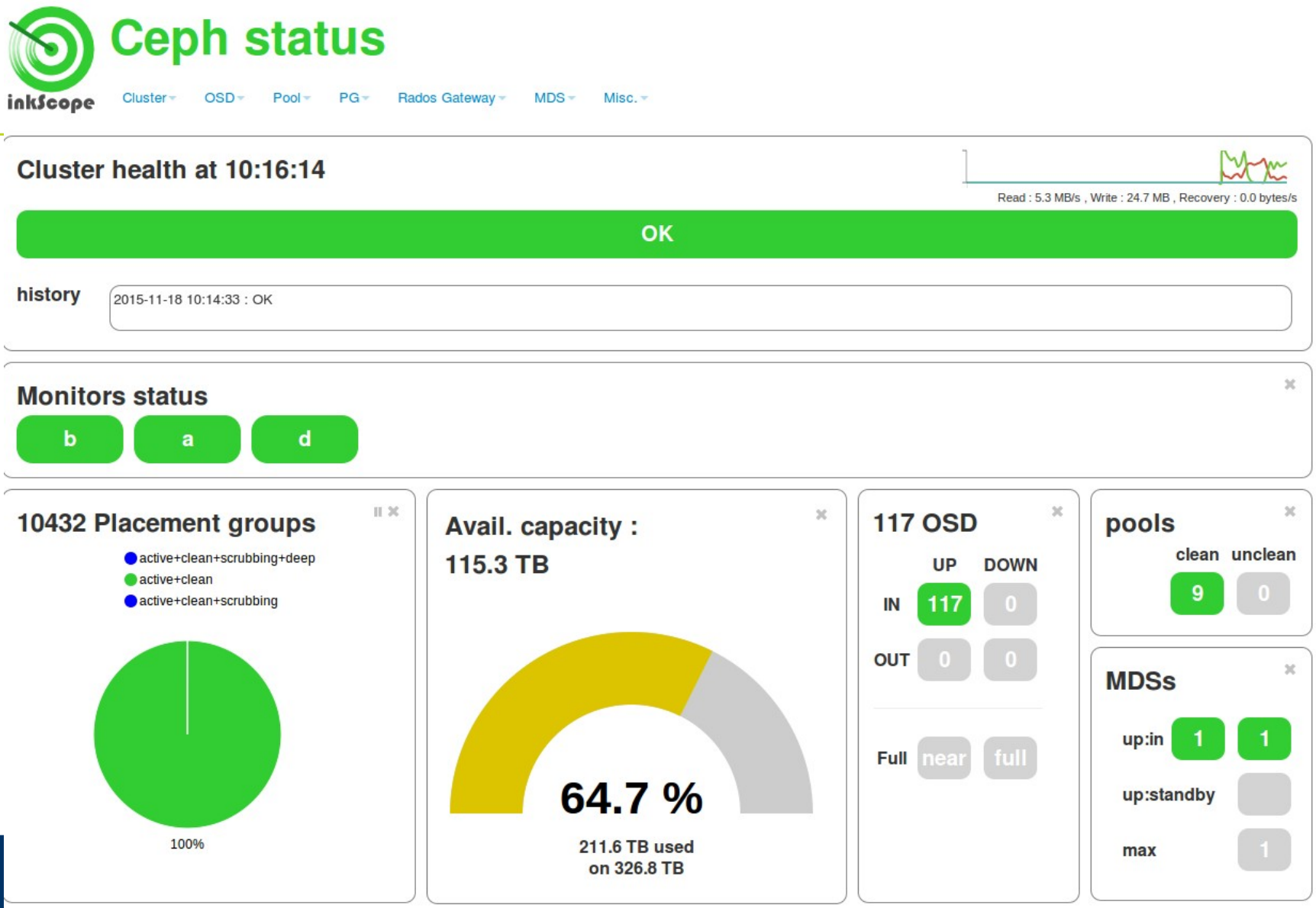
```
GLOBAL:
    SIZE      AVAIL      RAW USED     %RAW USED
    326T       115T         211T         64.72
POOLS:
    NAME             ID     USED       %USED     MAX AVAIL     OBJECTS
    data             0          0         0        38561G           0
    metadata         1       9470         0        38561G          21
    rbd              2          0         0        38561G           0
    bacula-D         3      26623G      7.95        25707G     6815746
    NFS              4       7480G      2.24        38561G     1961891
    bacula-SIG       5      16259G      4.86        25707G     4162837
    bacula-NB        6      23495G      7.02        25707G     6027262
    MIGRATIONS       7          16         0        25707G           3
    backuppc         8        897G      0.27        25707G      229890
```

326 To, 110 To utiles !!
(~75 To utilisés)

Rep = 2

Rep = 3

11/2015 : Cluster D2 = Migration vers Hammer (LTS) + Erasure coding (2+1)
Choix de ne pas faire de migration à chaud. On ne joue pas avec les données.

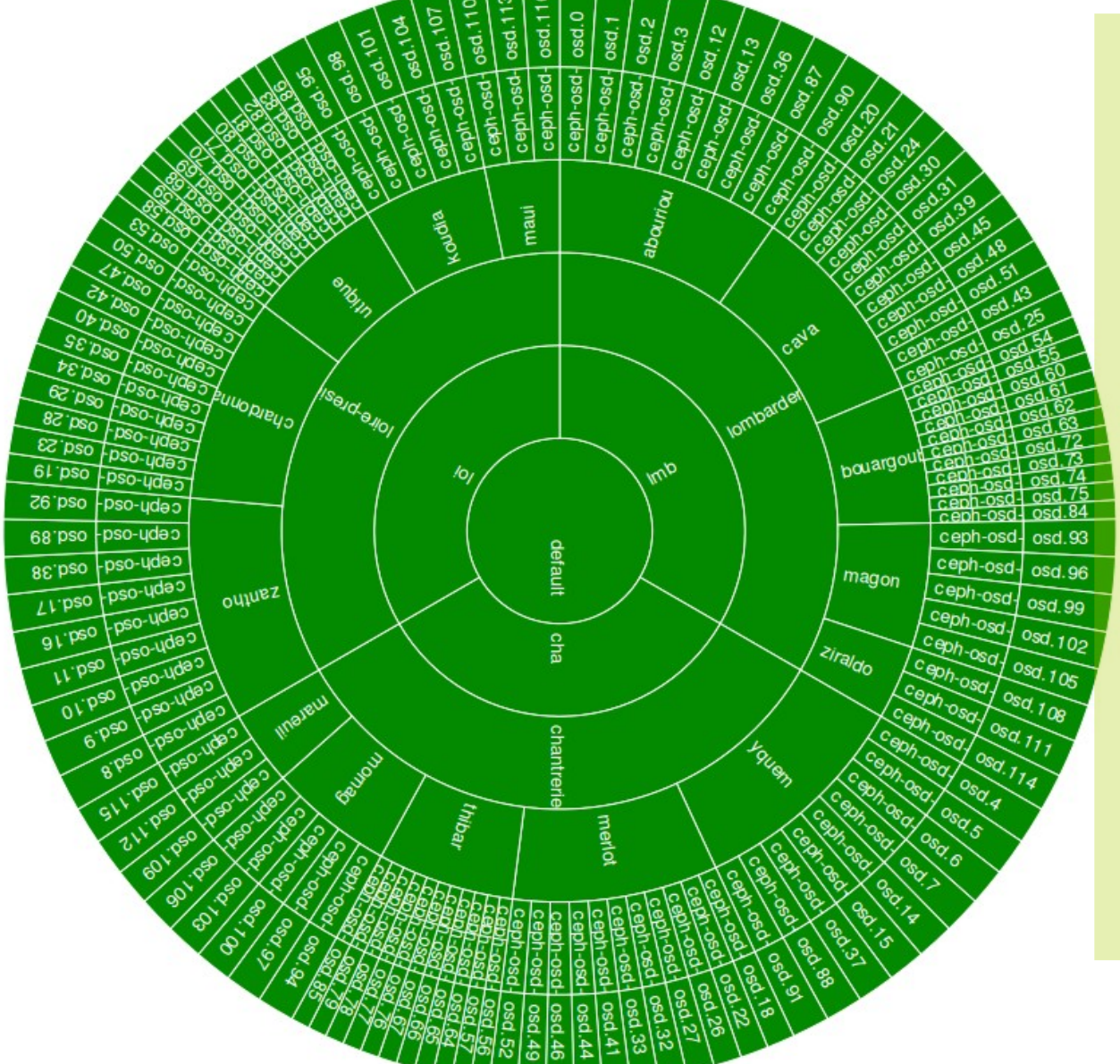


Cluster B
(vue inkscope)

Répliqua x3

~75 To
utiles

9 pools



Cluster B
(vue inkscope)

117 OSD
répartis hiérarchiquement

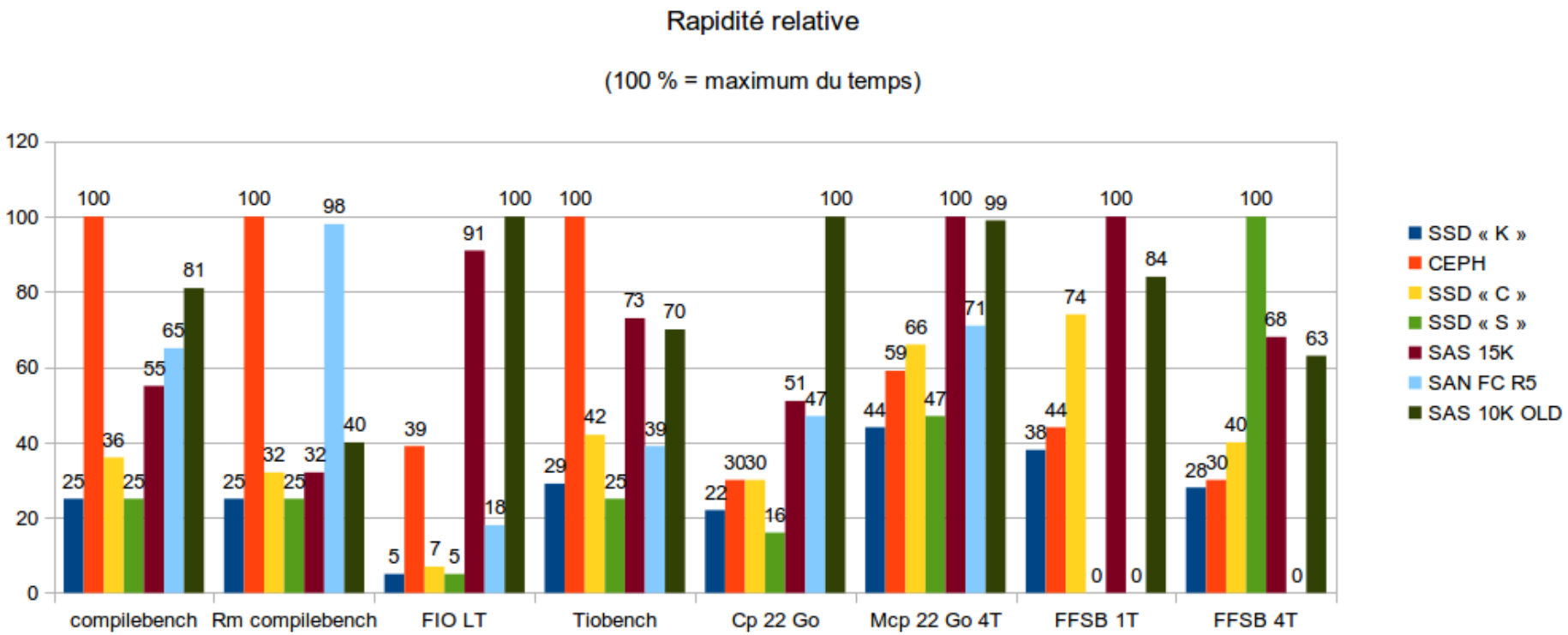
3 lieux distincts
1 salle par lieu
5 serveurs physiques par salle

2 à 9 OSD par serveurs volumétriques
(performance non cruciale)

Accès possibles

Type	Bloc	Fichier	Objet
Linux	Krbd (kernel) Kvm+librbd (userspace/kernel)	Cephfs (kernel)	API Swift API S3 API rados
Autres	Rbd-fuse (*BSD ?) Kvm+librbd (*BSD ?) Portages Windows (? ? ?).	Cephfs-fuse	
Via un Tiers	ISCSI : tgt (module ceph, multipath)	NFS : Linux Knfs Ganesha pNFS (HA) Samba : Linux + krbd/librbd	

Performances (dumpling)



Attention aux benchmarks !!! (Tuning, monothread, RAM, etc...)
À refaire avec Erasure coding, cache/Tiering SSD et Hammer/Infernalis.

Stratégie 2016, perspectives

Retour sur investissement :

- Erasure coding → espace utile x2
- Tiering → Utiliser au mieux les machines

Nouveaux usages → performance, volumétrie

- Support direct de l'IaaS, cloud → SSD
- Nouveau cluster : Logs & traitements, sécu

Rationaliser/Automatiser les déploiements (Ansible, Saltstack, Puppet, Foreman)

Maintenance clusters, évolutions en douceur.

- Clonage du cluster ou évolution à chaud

Meilleures pratiques

- Pas de RAID !
- XFS comme FS de base
- Un Linux de confiance
- Tous les SSD ne se valent pas
- Plusieurs salles machines
- Bien tailler les machines
- Plusieurs clusters : production/sauvegarde/tests
- Les machines doivent êtres considérées comme des boîtes noires
- Le réseau est déterminant
- Suivre le développement, se documenter, se faire épauler

Conclusions

Très stable : plus de 3 années de production et marche remarquablement bien.
Des possibilités d'évolution infinies, simples à mettre en place. Le Po est facilement atteignable.

Mais c'est encore un travail d'expert, en particulier au niveau Debug.

Beaucoup de nouveautés excitantes arrivent avec Infernalis :
(stabilisation de cephfs, amélioration des performances, nouvelles fonctionalitées....)

Prochaine étape : Jewel

Merci !

Questions ?

Crédits : Opencliparts / Openstreetmap / Ceph.com