CEPH @ INRA

SCIENCE & IMPACT

Audaces 2019

Sebastien.Cat@inra.fr

INRA
SCIENCE & IMPACT

EPST fondé en 1946

Recherche en agronomie

Ministère de la recherche
Ministère de l'agriculture

5 priorités thématiques et 3 orientations de politique générale

Ambition globale d'atteindre la sécurité alimentaire

Des agricultures diverses et multi-performantes

Des systèmes agricoles et forestiers face au défi climatique

Une alimentation saine et durable

Des bioressources aux usages complémentaires

[#OpenScience]

[#OpenInra]

[#Appui]

7 903 agents titulaires, dont 51,2 % de femmes

1 849 chercheurs titulaires

2 353 stagiaires accueillis & **556** doctorants rémunérés

250 unités de recherche et **45** unités expérimentales

13 départements de recherche et **9** métaprogrammes

17 centres de recherche

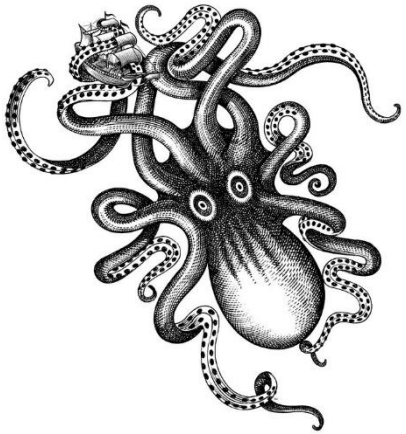
http://www.inra.fr

Points abordés- CEPH@INRA

1- Services et perspectives pour les unités de recherche

2- L’infrastructure et ses composants

3- Retours d’expériences / WIP



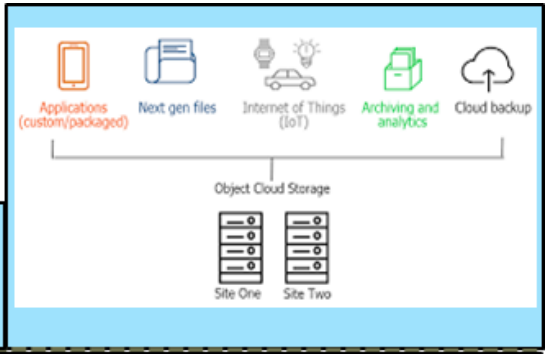
1- CEPH : une infra tout-terrain pour stocker les données

Livraison progressive d'une infra objet, bloc et fichier : V1 =>V4

	Datacenter	Sites	Use case
V1 05/2018	Objet Avec option Réplication	Objet Avec option Réplication	***Stockage objet capacitif avec option PCA*** *(actif/ actif en mode objet) Stockage « froid » d'unité, usage backup possible. Montage d'espace de stockage capacitif sur les serveurs des DC en objet uniquement*
V2 06/2019	Passerelle EFSS vers Objet	Passerelle EFSS vers Objet	*Déploiement du portail Nextcloud - permettant de connecter des « external storage » (S3, SFTP,CIFS..) – portail EFSS - Fonctionnalités de partages avancées / connexion AD – portail web et clients lourds pour PDT*
V3 09/2019	Bloc		*Périmètre DC : services BLOC (RBD)*
V4 01/2020	CEPHFS Snapshots locaux		*Service d'espace fichiers* ***CEPHFS (montage fuse) dans les DC sur les serveurs/VM – possibilité de snapshots sur cephfs***

Mise en exploitation OPEN

Ajout de perf. (NVMe / bluestore



Trajectoire:

OBJET → Passerelle utilisateurs EFSS connectée à CEPH … et aux autres services de stockage DSI → BLOC (RBD) → FICHIER

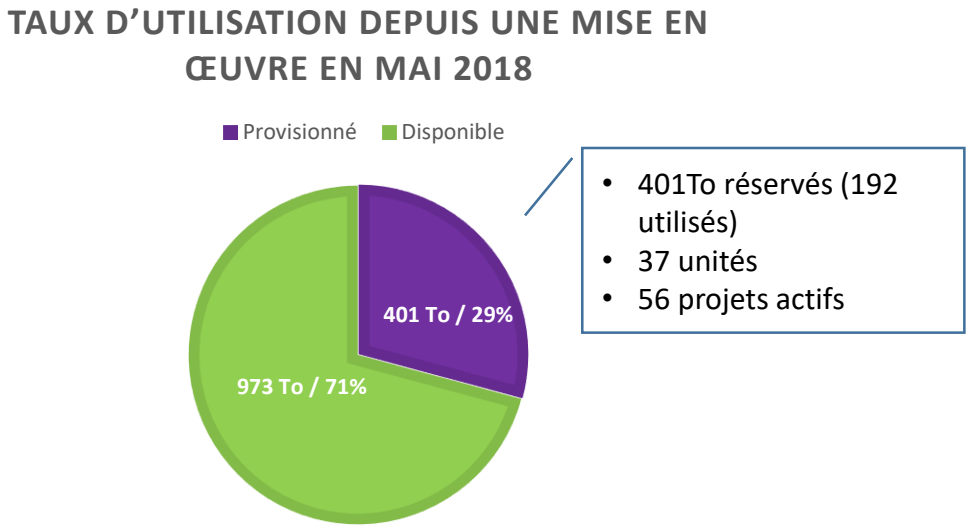
1- Focus sur le stockage objet – service/infra V1

Quelques exemples de cas d'usages à l'INRA

- ✓ Sauvegardes de séquences génomiques et données de projet
- ✓ Sauvegarde de données de dispositifs scientifiques (tomographes..)
- ✓ Sauvegardes des données de plateformes de phénotypage
- ✓ Sauvegarde de dump de bases de données
- ✓ Sauvegarde de postes de travail
- ✓ Sauvegarde de données actives de NAS (Netapp..)

- ✓ Stockage et traitement de données d'images RMN brutes et traitées
- ✓ Usage des espaces au travers de portail web d'unité (Nextcloud / goofys)

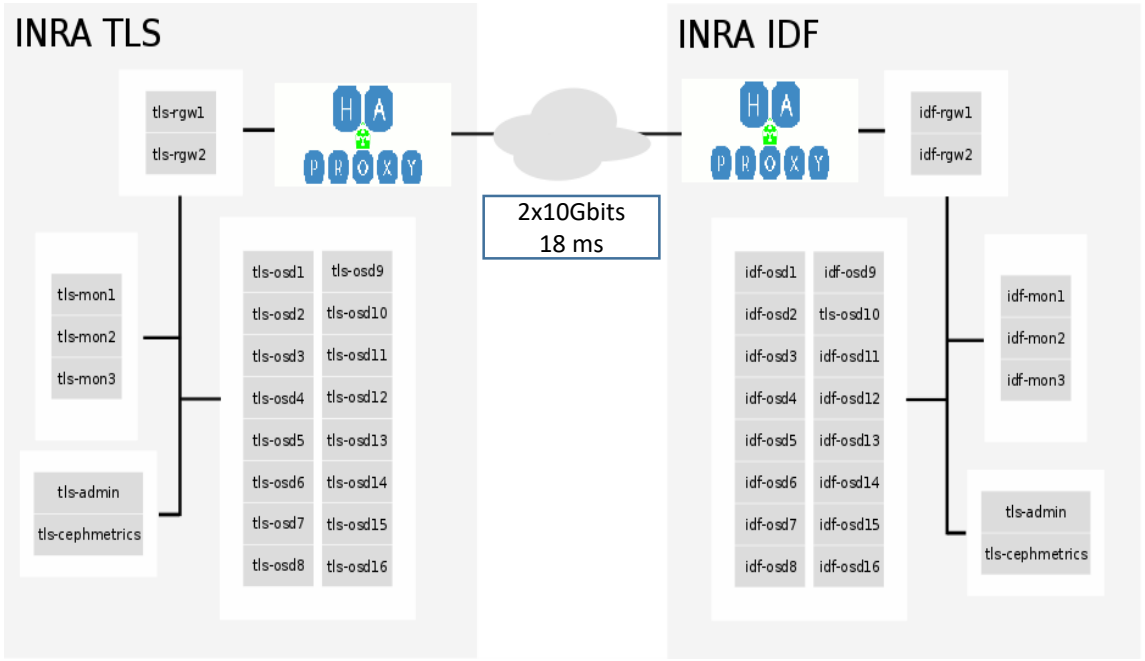
- ✓ En cours : Utilisation pour stocker une partie de la production scientifique INRA qui sera référencée dans HAL (https://hal.archives-ouvertes.fr/)
- ✓ En prévision : stockage V2 pour https://data.inra.fr/ via l'application Dataverse



	Volumétrie réservée
Toulouse	136 To
Paris	176 To
Géorépliqué	89 To

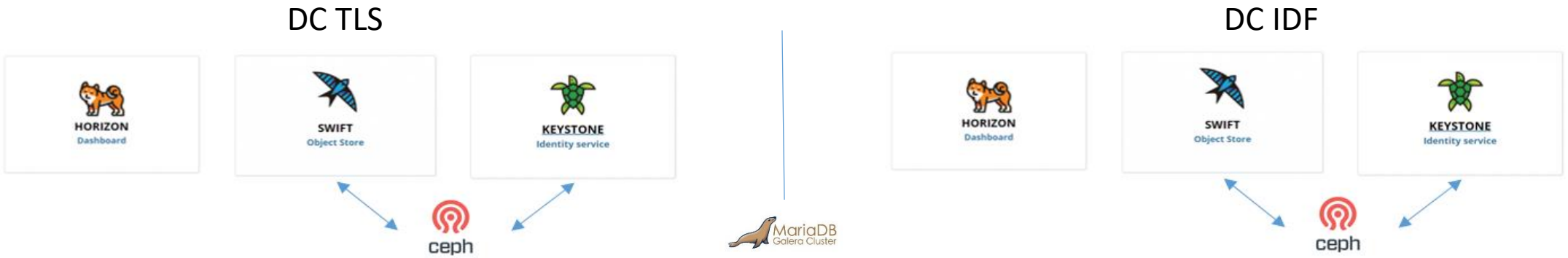
2- Focus sur les composants de l'infrastructure V1

Architecture physique CEPH

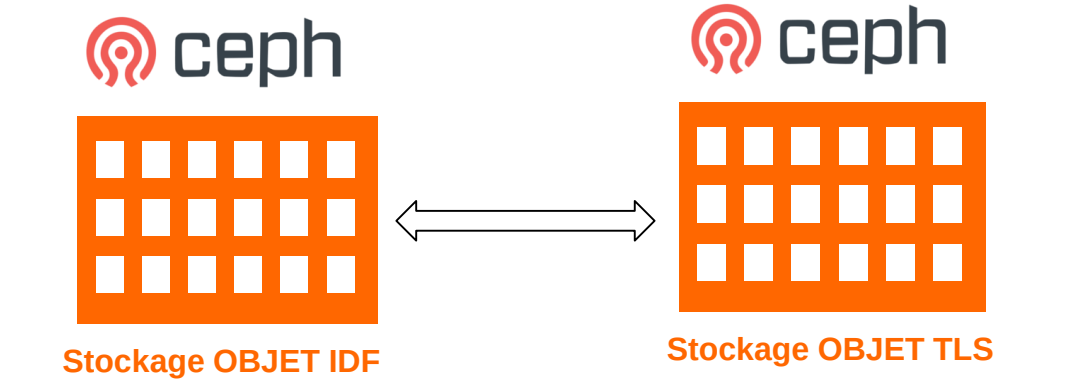


Configuration des « zonegroup » et « zone » RGW

Architecture OpenStack/CEPH



2- Un portail utilisateur : OpenStack



https <endpoint s3r-idf> ⟷ *https <endpoint s3r-tls>*

https <endpoint s3-idf>

https <endpoint s3-tls>

Accès API S3 (https/ clés d'accès EC2 pour client lourd)

2 DC, 1 CLUSTER CEPH /DC

- Stockage objet, en EC 8+4, filestore
- Sur 12 serveurs OSD/DC, perte possible de 2 serveurs en production (et le service reste UP)
- A 3 serveurs perdus, le service s'arrête, mais les données ne sont pas perdues
- 5 R640 : 2 RGW – 96G DDR4 – 3 MON – 32G DDR4
- 12 R740XD : nœuds d'OSD : (16x8T + 4 SSD de 240G) / 192G DDR4

2 OFFRES DE SERVICES

- Endpoint : Stockage « géorépliqué » (PRA/PCA – Actif/ Actif)
  - Stockage possible sur cluster DC TLS et/ou IDF
    - Réplication automatisée
    - RPO/RTO ~ 0
- Endpoint : Stockage « mono-DC »
  - Stockage non répliqué et stocké uniquement sur un DC INRA
    - Choix de localisation: IDF ou TLS

1 PORTAIL CLOUD

https <portail cloud INRA>

Portail web horizon de gestion/ récupération des clés objets, Visualisation des projets accessibles + buckets – Accès LDAP

openstack™ CLOUD SOFTWARE

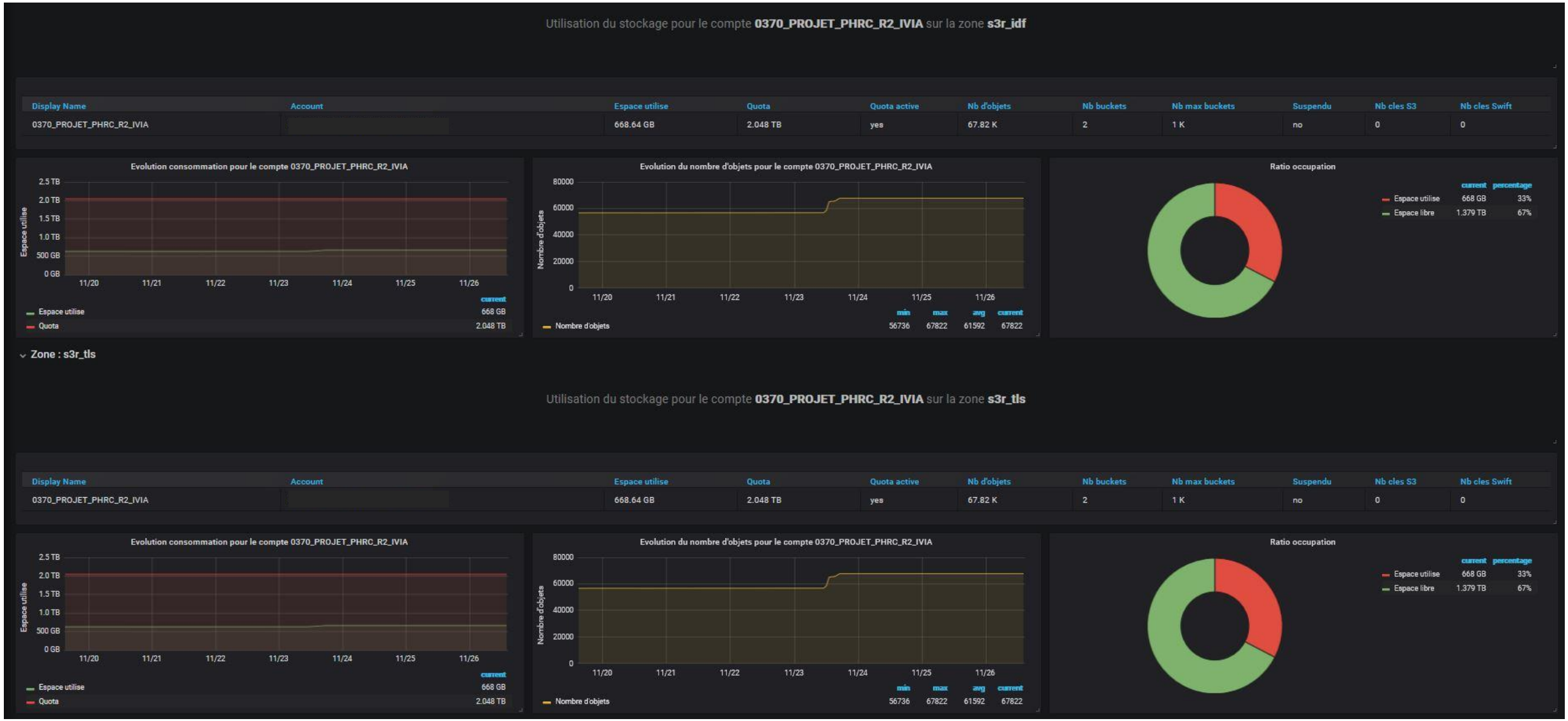
Utilisateurs potentiels ?

- Gestionnaires de données
- Informaticiens d'unité
- Data scientists
- Développeurs
- Scientifiques ayant besoin de sauvegarder leurs données brutes et traitées

Clients lourds d'accès au service ?

- CloudBerry, S3browser (explorateurs)
- Duplicati, Duplicity (backup)
- AWS CLI, S3cmd, rclone (cli muti-usage)
- Webdrive, client « hybride » avec cache, drive windows
- S3FS / goofys (point de montage unix)
- Dataverse / acces via API swift

2- Portail de supervision basé sur Grafana en complément



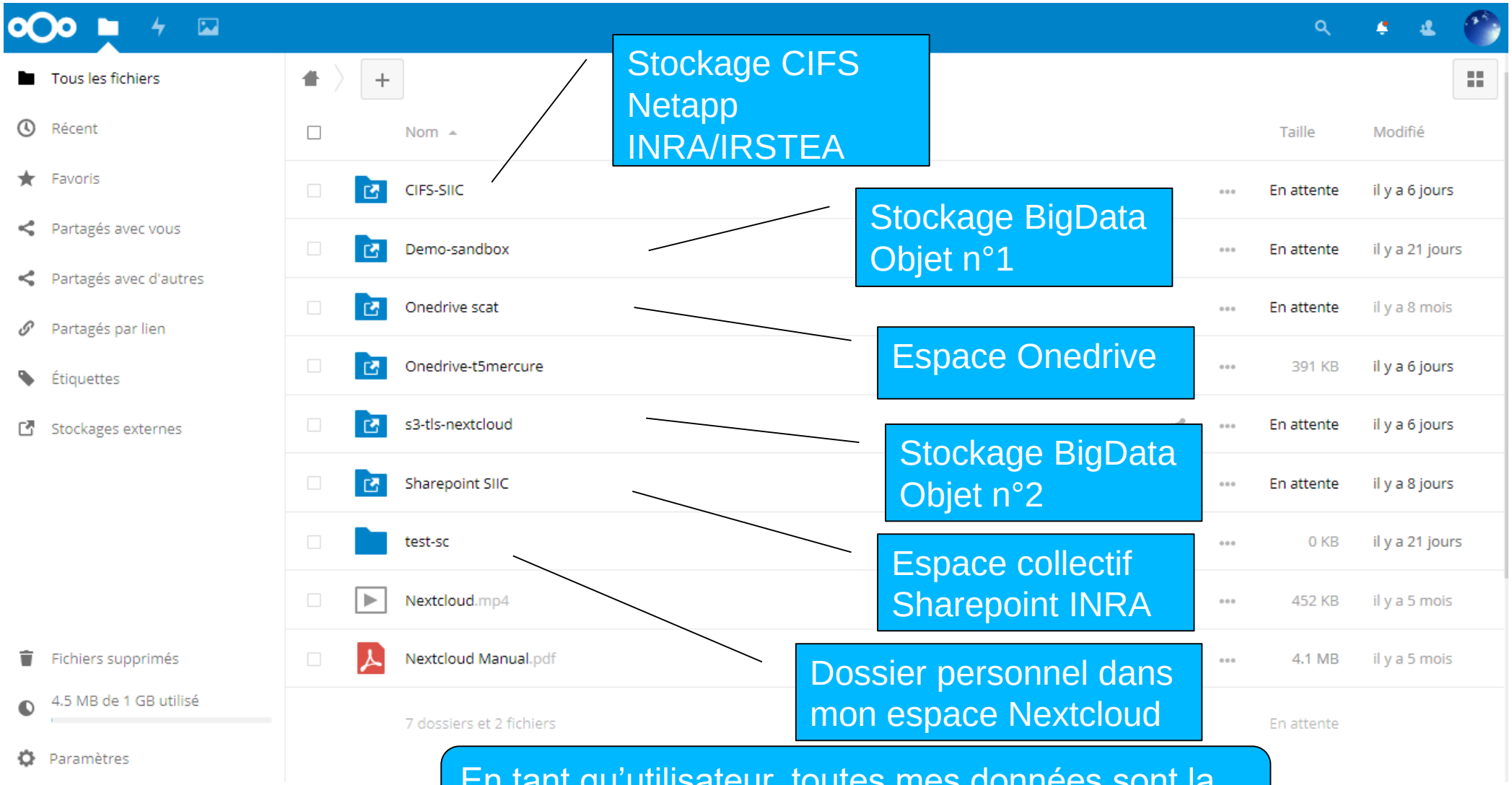
Via openstack :

- ✓ Accéder à tous les projets objets, taille et nombre d'objets dans les buckets, accès API
- ✓ Interface web permettant de manipuler les données de manière limitée

Via Grafana (+ librairie python rgwadmin) :

- ✓ Visibilité sur les taux de montée en charge par projets, permettant à chaque gestionnaire d'espace de faire des prévisions de capacité

Service V2 : ajout d'un portail vers des « stockage externes »



Stockage CIFS Netapp INRA/IRSTEA

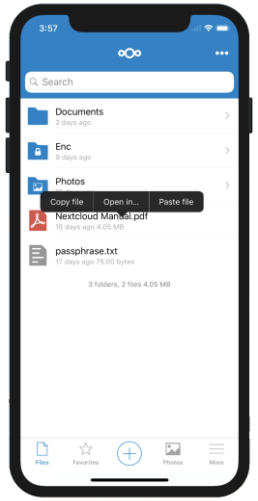
Stockage BigData Objet n°1

Espace Onedrive

Stockage BigData Objet n°2

Espace collectif Sharepoint INRA

Dossier personnel dans mon espace Nextcloud



En tant qu'utilisateur, toutes mes données sont la, et accessibles depuis l'extérieur

3- Retours d'expériences sur CEPH et le stockage objet

Coté utilisateurs :

- ✓ Solution OpenSource, interopérable, interfaçable
  - ✓ Ex : Backend IRODS possible
- ✓ Service accessible aux partenaires externes, authentification par clés révocables.
- ✓ Le stockage objet est puissant, mais nécessite un accompagnement
  - ✓ Usages des API S3/Swift
- ✓ Nécessite de faire beaucoup de documentations pour la prise en main
  - ✓ Différents cas d'usages, différents OS, différents souhaits
- ✓ Des clients lourds aux performances et aux fonctionnalités très variables
- ✓ La compatibilité des clients lourds avec CEPH doit être vérifiée
- ✓ Des utilisateurs très/trop habitués aux drives CIFS cherchant à avoir le même « look and feel »
- ✓ Vitesse de transfert : variable selon la taille des fichiers, des débits et latences WAN vers les DC
  - ✓ Acquisition et amélioration du logiciel webdrive et s3browser
  - ✓ Intégration d'une interface web s3explorer
    - ✓ https://github.com/awslabs/aws-js-s3-explorer/tree/v2-alpha
  - ✓ Adaptation de s3-pit-restore
    https://github.com/madisoft/s3-pit-restore

Coté administrateurs:

- ✓ CEPH permet une automatisation poussée, mais il faut prendre le temps de la faire
- ✓ L'intégration avec Openstack facilite vraiment la gestion (distribution des clés, accès à différents projets).
- ✓ La configuration des zones groupes « local » et « replicated » sur une même RGW a nécessité l'expertise de Redhat
- ✓ Nous avons identifié un bug sur la géoréplication asynchrone, corrigée dans CEPH 12.2.5
  - ✓ https://access.redhat.com/errata/RHBA-2018:2375 (BZ#1608977)
- ✓ Quelques paramètres ont dû être ajustés en production : '*Dynamic resharding is not supported in multisite environment*'
- ✓ Le "bucket sharding" sur les index a eu pour conséquence de fausser les stats de volumétrie (correction en 12.2.8)

Coté stockage objet / service:

- ✓ Fiabilité : arrêt de DC : ok, disque HS : ok, nœud « down » : ok
- ✓ PRA / PCA apprécié en géorépliqué
- ✓ Stockage objet : de nombreuses possibilités
  - ✓ Très intéressantes à comprendre pour utiliser au mieux ce type de service
- ✓ Le multipart-upload: attention en géorépliqué
  - ✓ AbortIncompleteMultipartUpload lifecycle policy
- ✓ Le versionning
- ✓ Les règles de cycles de de vie
- ✓ La publication d'objet
  - ✓ Time limited URL
- ✓ Les « external buckets »
- ✓ Les API S3 et/ou Swift

3 – Zoom : migration Filestore => Bluestore

Objectif recherché : apporter de la performance sur le stockage objet et permettre d'autres cas d'usages

✓ Choix initial : nœud d'OSD composé de 16 disques de 8T et pool en EC 8+4 : stockage orienté capacitif
✓ Avec le temps, certains buckets qui comportent des dizaines de millions d'objets sont lents à parcourir
✓ Finalement avoir de la performance sur du stockage objet est aussi nécessaire dans notre cas

Les perspectives qui renforcent le besoin de performances :

- Stockage RBD pour Cinder/Openstack
- Montage CephFS demandé par les unités : cas d'usage calcul/ traitement de données envisagée

Choix fait : acquisition de 2 cartes NVMe 1,6To par nœuds d'OSD pour déporter les DB et WAL Bluestore.

Méthode de migration choisie :

- Validation et vérification sur une pré-production (les gains observés : Latence /4 en particulier)
- Migration nœud par nœud en bluestore + NVMe à l'aide d'un playbook Ansible
- Les données du nœud sont perdues et reconstruites à partir des autres nœuds (backfill)
- Durée de reconstruction d'un nœud ~ 2 jours actuellement avec max_backfill à 7
    - En cours sur la production, jusque la tout va bien ☺
- Phase 2 de la migration : créer un pool SSD pour y mettre les index.

Merci pour votre attention

Questions ?