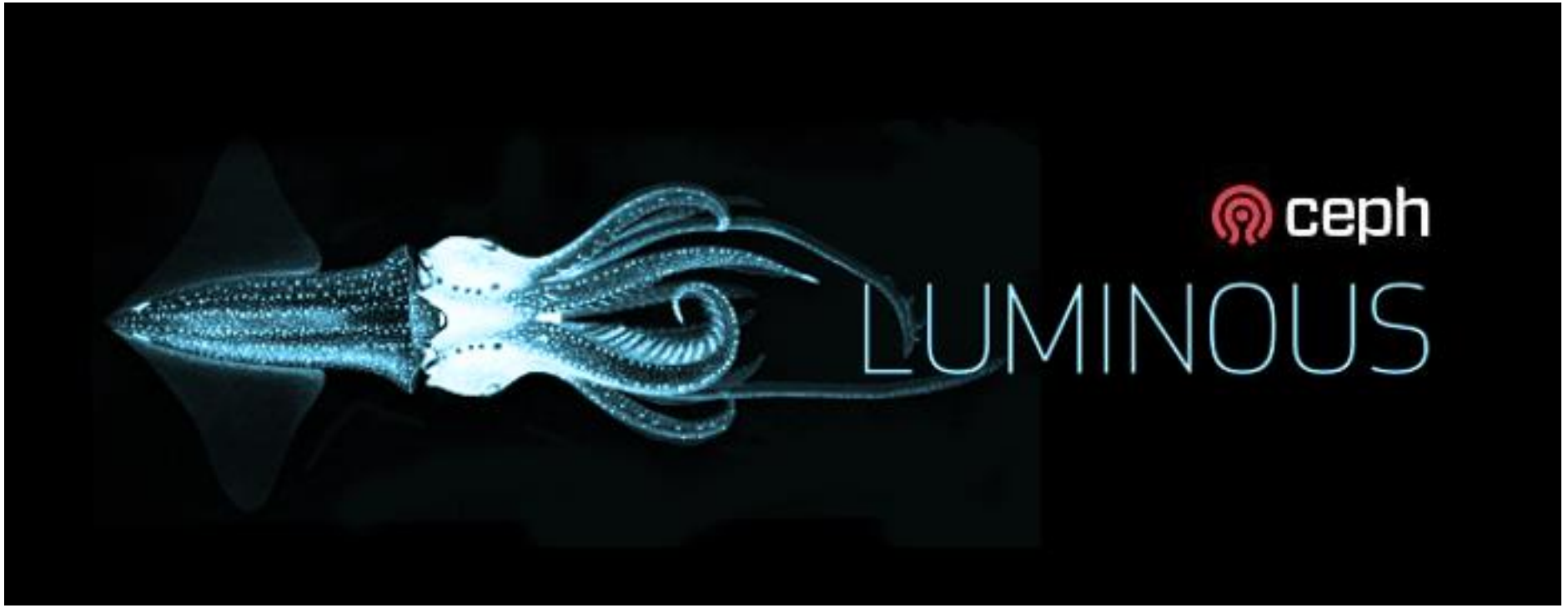
CEPH @ INRA

SCIENCE & IMPACT



JTech CEPH 2018

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

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

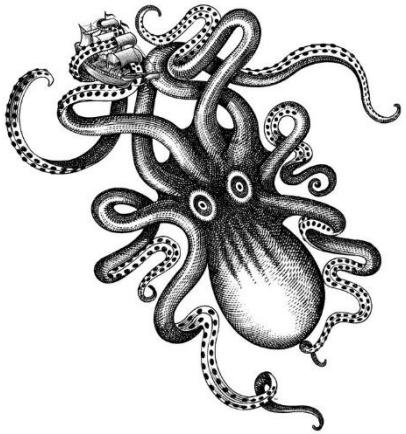
Points abordés- CEPH@INRA

1- Un service basé sur CEPH pour les unités de recherches

2- Synoptique du service

3- Intégration et configuration des composants

4- Retours d'expérience : exploitation, supervision, support



1- Un service basé sur CEPH, pour quels cas d'usages ?

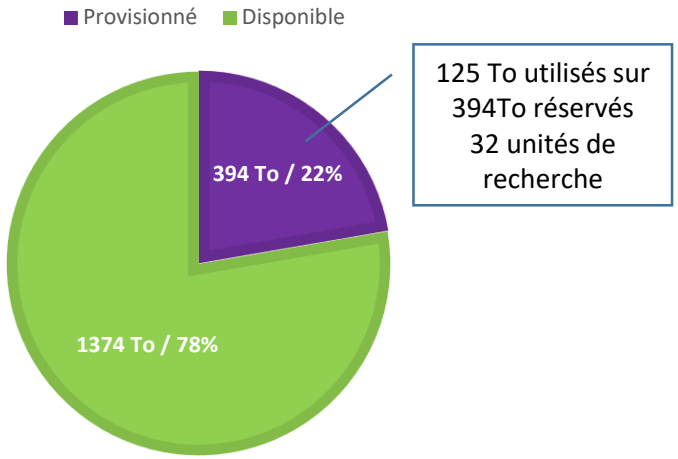
Usage principal : des entrepôts pour uploader et downloader des données

- ✓ Sauvegarde et récupération de données
  - ✓ Décharge ou sauvegarde des espaces de données actives
- ✓ Synchronisation d'espace de données depuis les postes de travail
- ✓ Partage de données via lien HTTPS, usage de sites statiques
- ✓ Entrepôts de données froides ou tièdes
- ✓ Données d'application natives « cloud interne »

Qu'est ce qu'en font les premiers usagers INRA (exemple) ?

- ✓ Sauvegardes de séquences génomiques et données de projet
- ✓ Sauvegarde de données de tomographes
- ✓ Sauvegardes des données de plateformes de phénotypage
- ✓ Sauvegarde de dump de bases de données
- ✓ Sauvegarde de postes de travail
- ✓ Stockage et traitement de données d'images RMN brutes et traitées
- ✓ 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

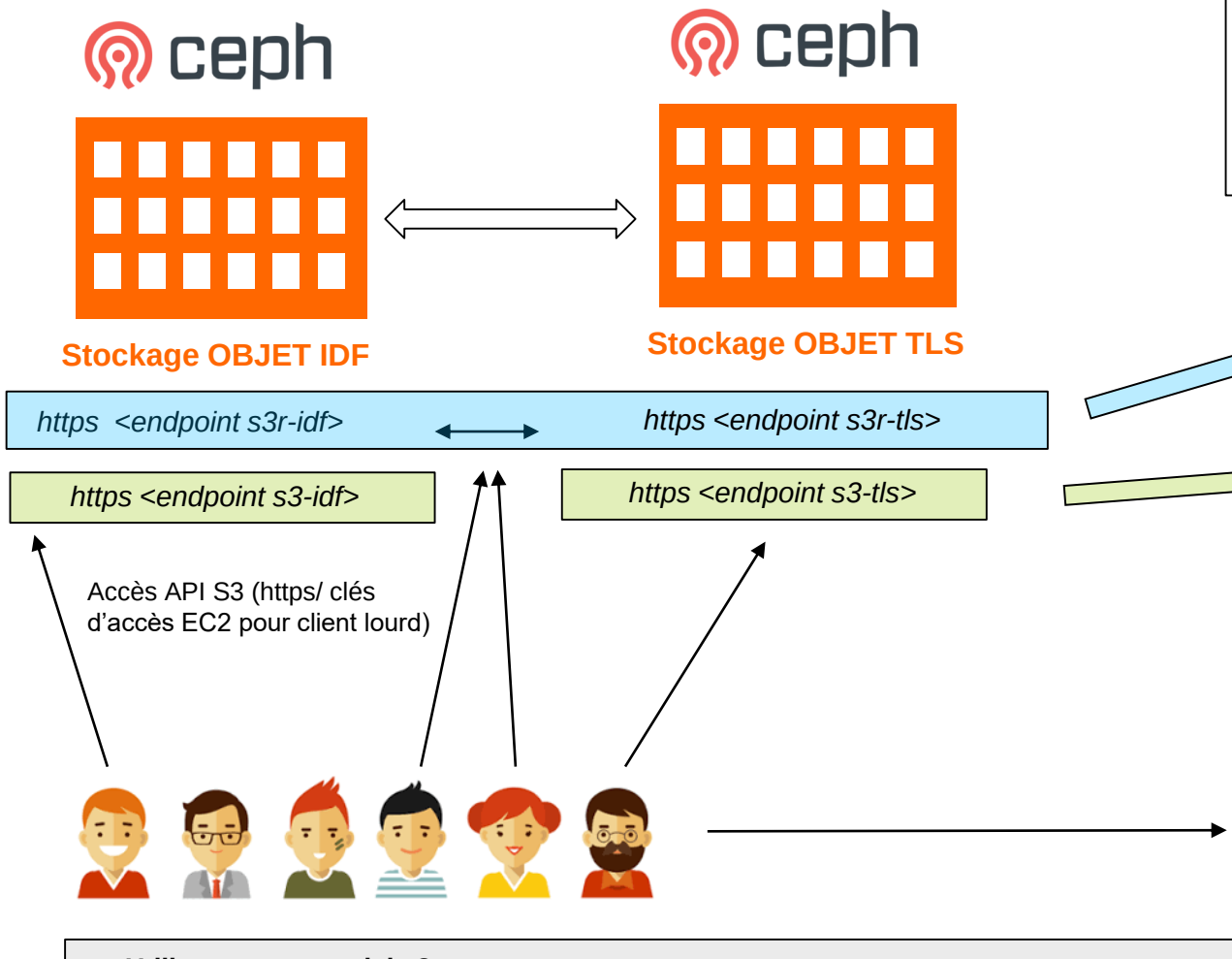
RETOUR DEPUIS MISE EN OEUVRE (05/2018)



2- Synoptique du service

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 ?

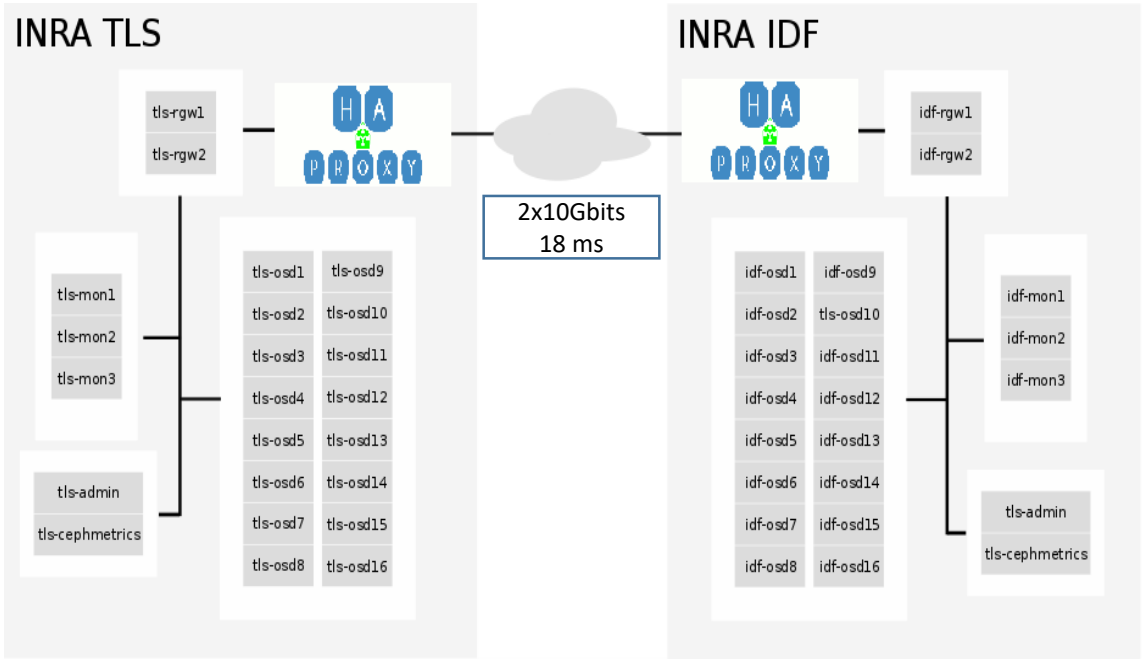
- Gestionnaires de données
- Informaticiens d'unité
- Data scientists
- Développeurs
- Scientifiques ayant besoin de sauvegarder leurs données brutes et traitées

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 (point de montage unix)
- Dataverse / acces via API swift

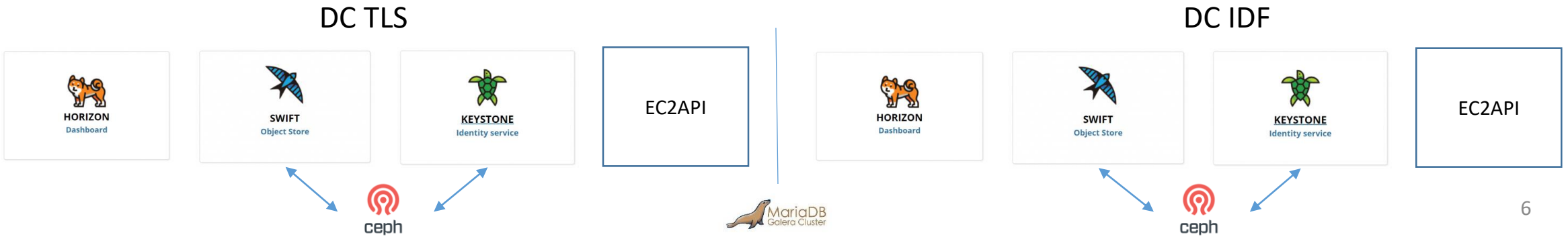
3- Intégration et configuration des composants

Architecture physique CEPH



Configuration des « zonegroup » et « zone » RGW

Architecture OpenStack/CEPH



4- Exploitation des clusters CEPH et du service

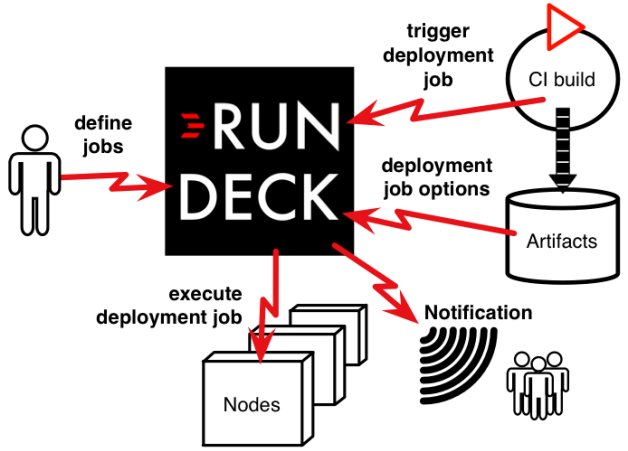
Déploiement des OS - nœuds du cluster via « Foreman »

- ✓ BootPXE / Scripts de post-installation des OS
- ✓ Déploiement automatisé des nœuds du clusters : réutilisable pour le déploiement de prochains nœuds

Déploiement et configuration de CEPH

- ✓ Installation logiciels et configuration des nœuds via Ansible
  - ➢ # ansible-playbook site.yml => déploiement sur tous les nœuds de CEPH / Update
  - ➢ Playbook utilisable lors de remplacement de nœuds ou modification de configurations
- ✓ Description au préalable des rôles et caractéristiques des nœuds dans des templates
- ✓ Validation des montées de version sur un environnement de pré-production

Automatiser autant que possible

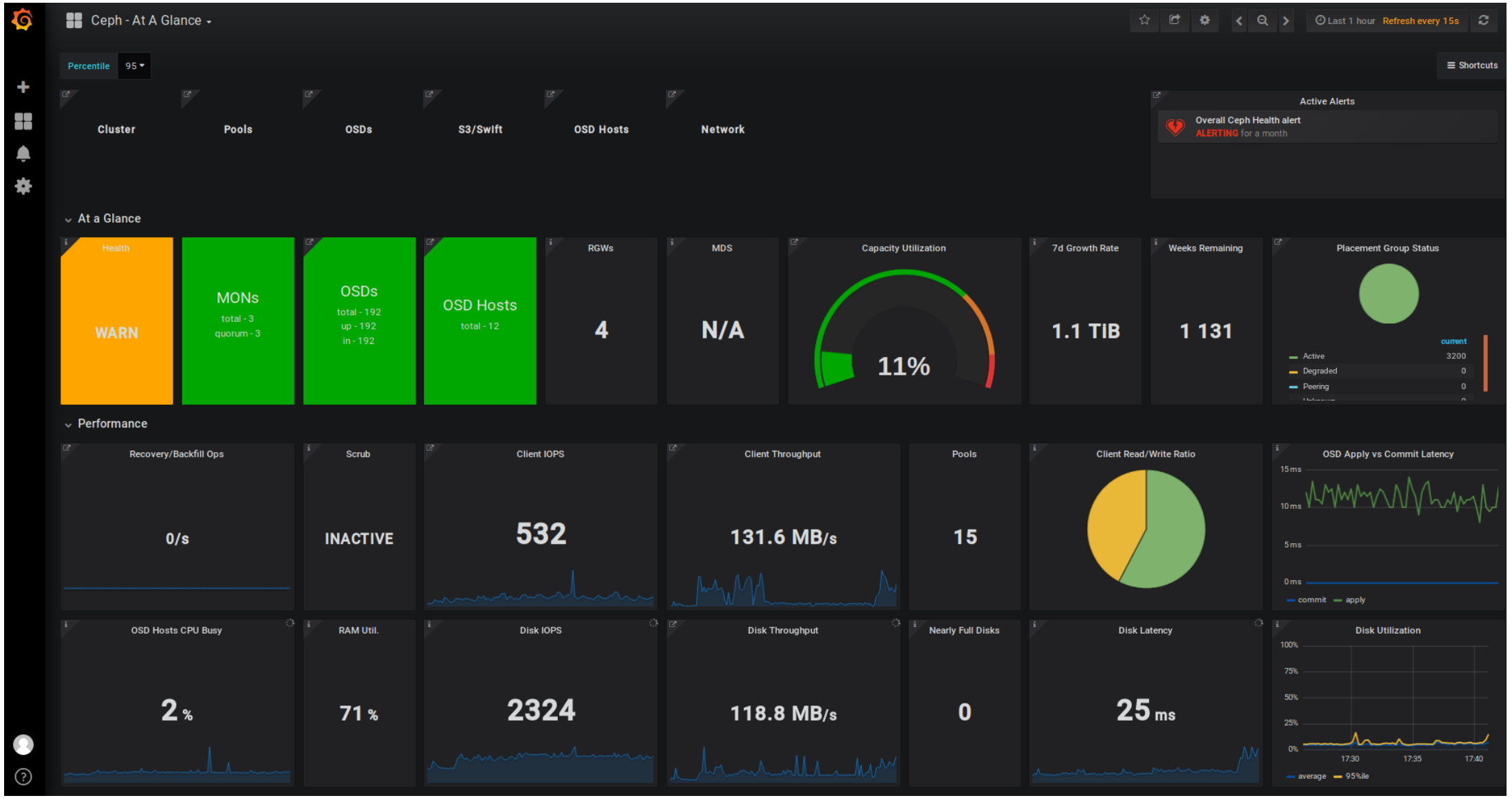


- ✓ Réalisation de scripts pour faciliter l'exploitation
- ✓ Interface d'automatisation RUNDECK interfacée au portail de services INRA

Sauvegarder les configurations et docs

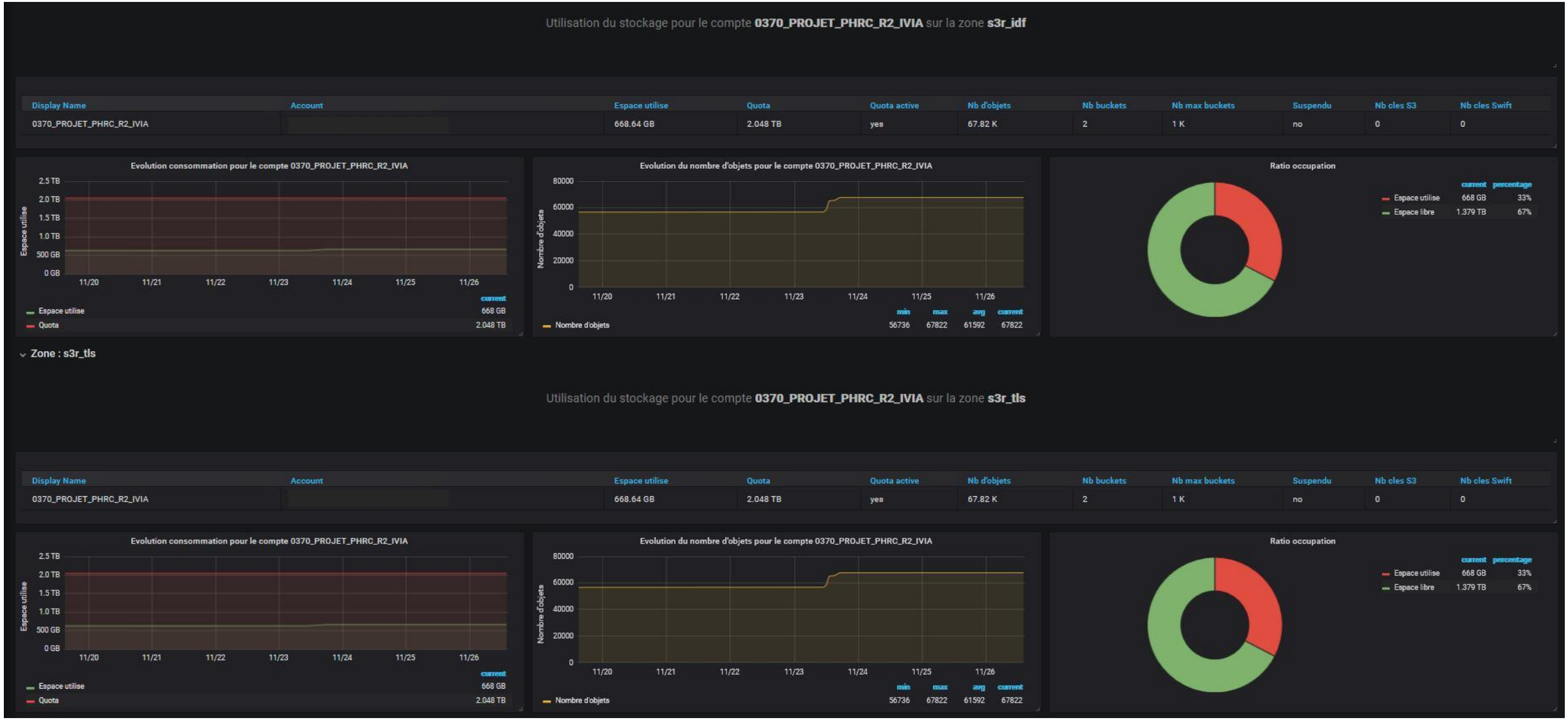
- ✓ Usage de GitLab

4- Supervision et métrologie du cluster et du service



✓ CEPHMETRICS : Etat des nœuds, disques, métrologie performances, alerting de l'équipe d'exploitation..

4- Métrologie et gestion du service pour les utilisateurs



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

4- Support et exploitation

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

Coté administrateurs:

- ✓ CEPH permet une automatisation poussée
- ✓ 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 cours)
  - ✓ workaround : unlink bucket from user / relink and reset stats.

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 versionnement
- ✓ Les règles de cycles de de vie
- ✓ La publication d'objet
  - ✓ Time limited URL
- ✓ Les « external buckets »
- ✓ Les API S3 et/ou Swift

Les prochaines étapes

De nouveaux projets à suivre … ☺

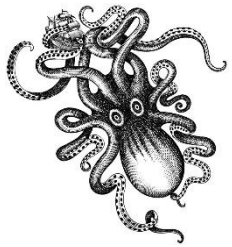
Projet 1 : Déployer un EFSS fédérateur des offres (NextCloud)

- ✓ CEPH en backend de stockage (mais pas que CEPH)

Nextcloud

Projet 2 : Stockage bloc et fichier avec CEPH

- ✓ Evolution : passer à Bluestore, tester les performances apportées par les cartes NVME
- ✓ Déployer CEPHFS : montage de type fuse / NFS
- ✓ Déployer du stockage BLOC (RBD/ISCSI) pour les infrastructures openstack ou vmware

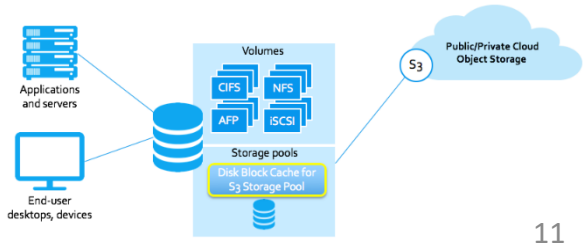


Projet 3 : Compléter les services possibles avec le mode objet

- ✓ Passerelle via « HAPROXY » pour héberger des sites web statique (issu de buckets)
- ✓ Indexer les métadonnées dans Elasticsearch ? « RGW Metadata Search »

Projet 4: CEPH en « Edge » ?

- ✓ Apporter de meilleures réponses pour nos sites distants
- ✓ Cluster CEPH en conteneurs
- ✓ S3FS (usage en cache possible)
- ✓ Proxy S3



Remerciements

A l'équipe « poulpe » ☺

Prototypes, ateliers multiples jusqu'à 20h pour les choix d'archi et le déploiement, la mise en œuvre de l'automatisation en mode devops …

- Etienne Chabrerie
- Mikael Loaec
- Matthieu Simon
- Sébastien Reboux
- Yves Civil