CENTRE DE CALCUL DU CNES

-

2 NOUVELLES PLATEFORMES DE STOCKAGE ET DE CALCUL POUR RAPPROCHER LES TRAITEMENTS DES DONNÉES

Journées Calcul et Données

REIMS

2 octobre 2023

Sommaire

CENTRE DE CALCUL DU cnes

1

CENTRE DE CALCUL DU CNES

Quelques chiffres

Une équipe Calcul et Données

- 7 personnels CNES
- 5 ETP pour le support quotidien (HPC & Datalake)
- 5 ETP prestations de portage et optimisation
- 2 ETP MCO Datalake
- 2 ETP MCO HPC6G

Utilisateurs

- + de **1100** utilisateurs réguliers
  - 30% Agent CNES
  - 50% Industriels
  - 20% Académiques (en hausse)
- ~100 projets spatiaux

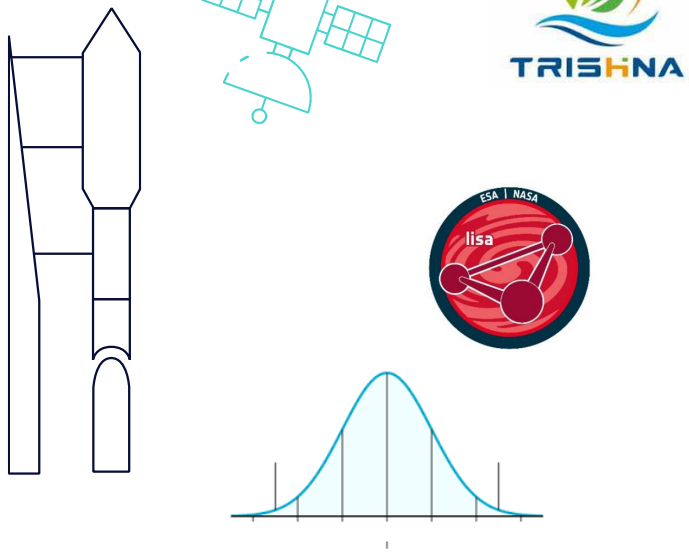
3 plateformes de calcul (HPC/HTC)

- PF diffusion limitée : ~16000 cœurs de calcul, 50 cartes GPU et 11 Po de stockage
- PF Diffusion Restreinte Spécial France : 1028 cœurs et 200 To de stockage
- PF SWOT dédiée au projet (2000 cœurs, 1Po de stockage)

3 plateformes de stockage

- Datalake : stockage objet, diffusion et valorisation de données 35 Po disque et 35 Po bandes
- HPSS données PEPS en fin de vie
- Archivage (STAF) : pérennisation de données

Exemples d'utilisation

R&D, étude, phase amont	Production, stockage, diffusion et visualisation de données (Observation de la Terre, Orbitographie)	Exploitation et valorisation de données
		

Offre de service Calcul

Support et expertise HPC

WikiHPC

Description de l'offre de service Calcul

Accompagnement utilisateurs

Ma Vie Numérique

APPTAINER
SINGULARITY
SARUS
Conteneurs

Métrologie

Spack
GDAL
netCDF

Logiciels scientifiques

Kubernetes
docker
Cloud

cnes DATALABS
jupyterhub

Services et applications

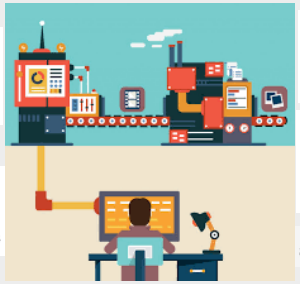
Catalogue

Archivage

Usine Intelligence Artificielle

PostGIS
SGBD

GitLab
Jenkins



JFrog Artifactory
JFrog Xray
sonarqube

Usine Logiciel
Intégration continue

Calcul

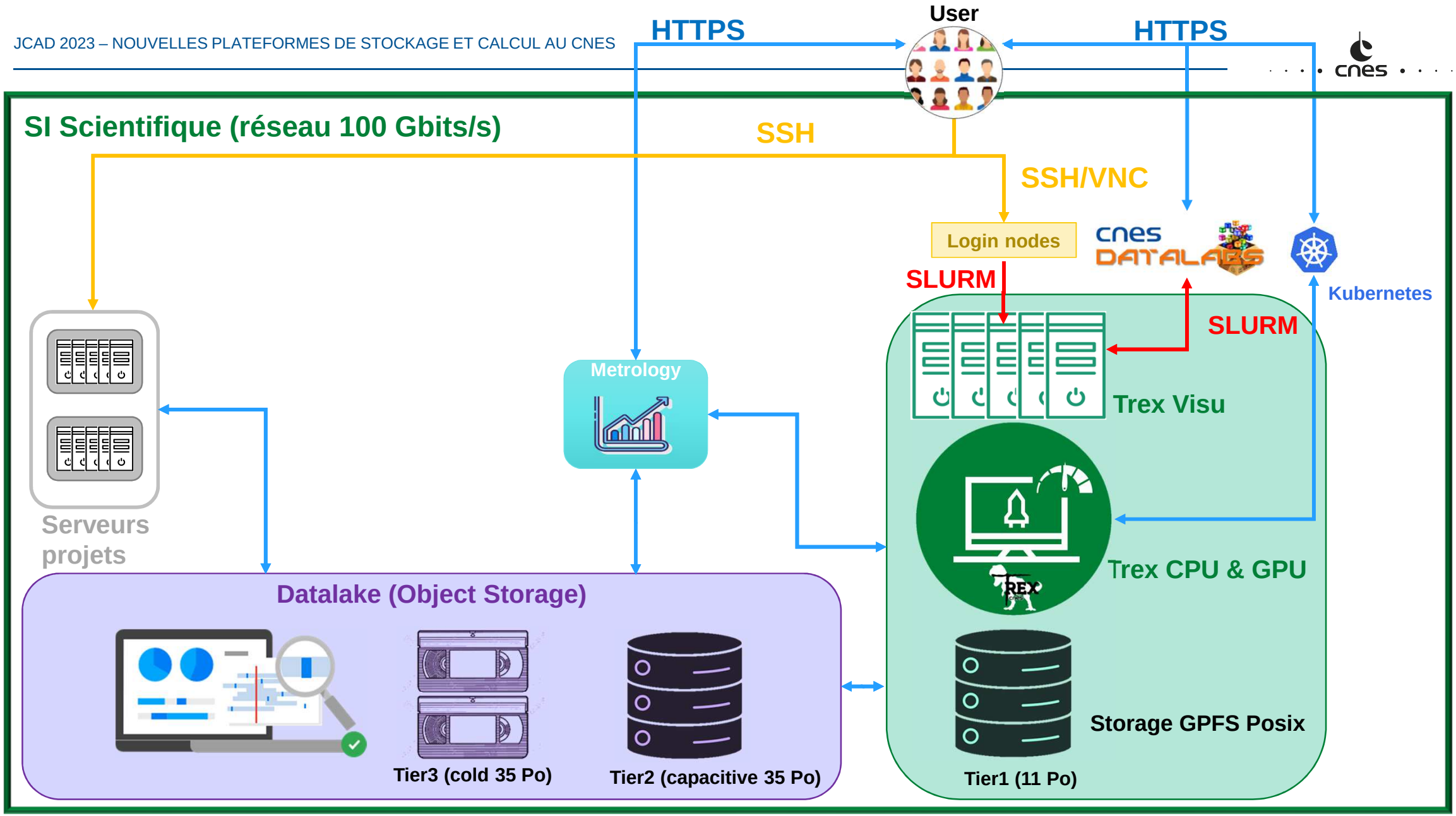
Stockage

Infrastructures

Stockage et sauvegarde NAS (HOME)

Hébergement App et Services

Réseaux et connectivité


User
HTTPS
HTTPS
SI Scientifique (réseau 100 Gbits/s)
SSH
SSH/VNC
Login nodes
cnes DATALABS
Kubernetes
SLURM
SLURM
Metrology
Trex Visu
Trex CPU & GPU
Serveurs projets
Datalake (Object Storage)
Tier3 (cold 35 Po)
Tier2 (capacitive 35 Po)
Storage GPFS Posix
Tier1 (11 Po)

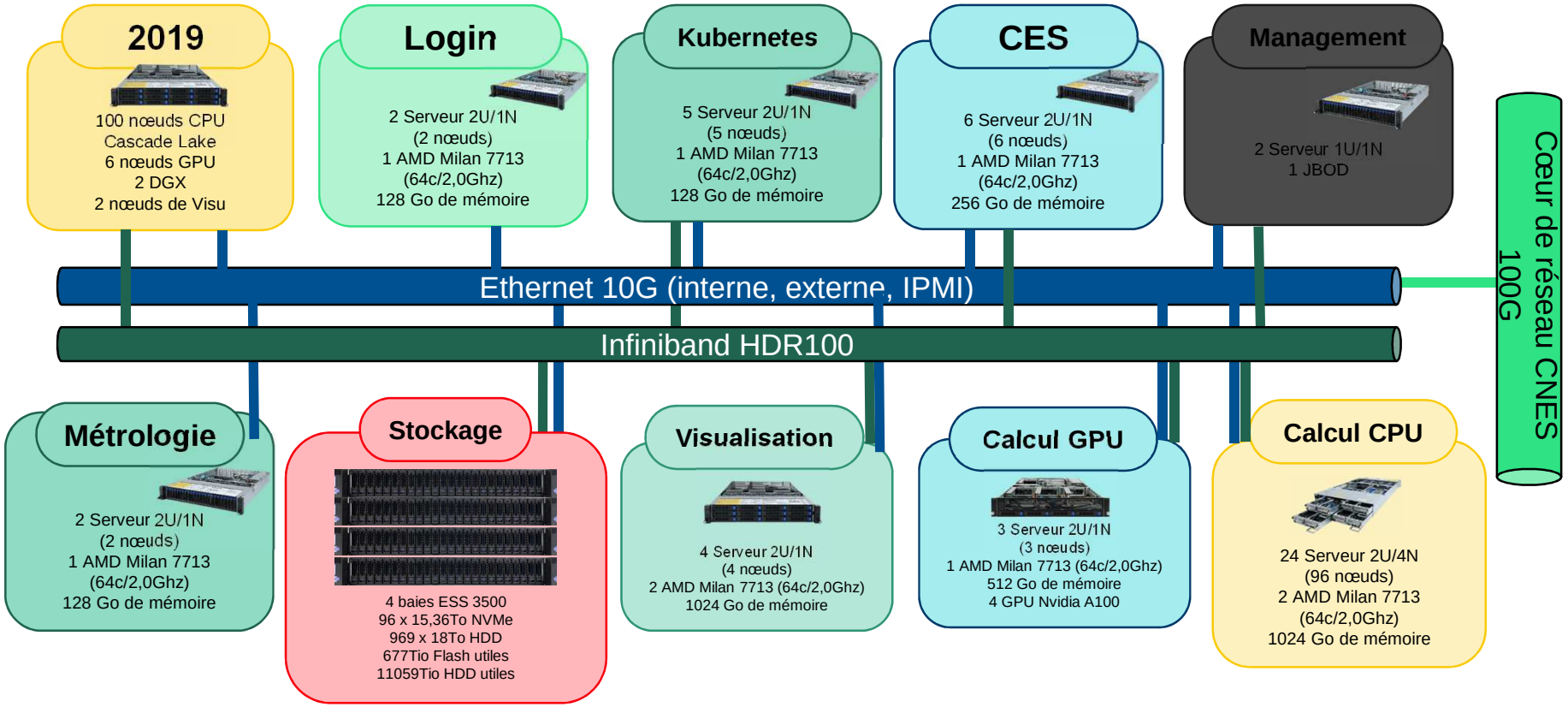
2

PLATEFORME DE CALCUL

Objectifs du renouvellement de la plateforme de calcul

- **Gérer l'obsolescence (matériel actuel 2016)**
- **Augmenter de la capacité**
  - +25% de ressources lié à la croissance du nombre d'utilisateurs
- **Maîtriser la consommation énergétique**
  - Choix des processeurs AMD + surveillance/optimisation de la consommation énergétique
- **Préparer l'inter-opérabilité**
  - Ordonnanceur Slurm
- **Proposer de nouveaux services**
  - Kubernetes (Cloud native)
  - Monitoring/Métrologie accessibles aux utilisateurs

Architecture globale TREX



Nouveau matériel

12 480 CPU : **Processeurs AMD** (64 Cœurs par CPU)

12 GPU : **NVIDIA A100** 80Go Pci-e

11 PiB de stockage Posix : Système de fichier parallèle et distribué **IBM Spectrum Scale**
baies **ESS-3500**
Engagement de performance IBM
Bandwidth : 50 GB/s
IOPS : 159 kiops

Réseau HP **Infiniband HDR-100**

Réseau **Ethernet 10G**

Système d'exploitation : **RedHat 8.X**

Ordonnanceur Jobs : **SLURM**

EVIDEN
an atos business

Métrologie et monitoring orienté utilisateurs/projets

Par login :

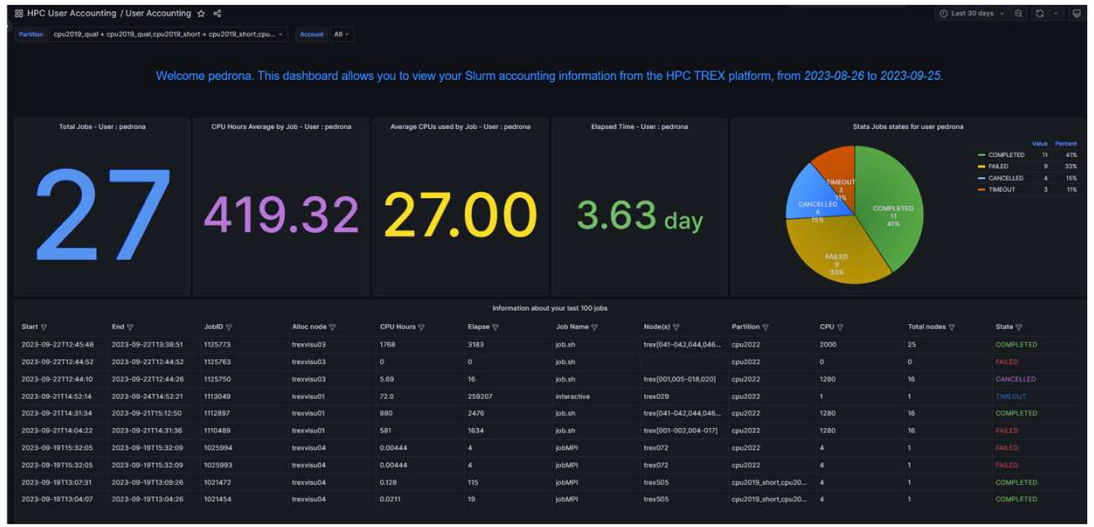
- Nombre de jobs, temps elapsed moyen, nombre de cores/nœuds par job
- Tableau de bord permettant d'afficher les métriques d'un job (utilisation CPU, mémoire, kWh)

Par projet :

- Décompte des nombres d'heures de calcul consommées
- Evolution temporelle du stockage utilisé (Go/mois)

Réalisation en cours ou à l'étude :

- Détection des mauvais usages (% utilisation CPU faible, surréservation de ressources, ….)
- Mesure et optimisation de l'énergie avec les outils BEO / BDPO
- Interfaçage avec le Centreon du CNES pour les tableaux de bord des projets opérationnels



Contraintes et difficultés

- **Réintégrer du matériel existant (lié au mode du financement du CNES) :**
  - 100 nœuds Cascade Lake 2019 sont en cours de réintégration dans la nouvelle plateforme

- **Assurer la continuité pour les projets producteurs de données en coordonnant le passage de chaque projet de HAL vers TREX**
  - Migration des données : 4 Po du GPFS de HAL (HPC5G) vers TREX (HPC6G) (solution IBM AFM)
  - Support projets pour la transition PBS => Slurm, RH7 => RH8, archi Intel => AMD
  - Garder les 100 nœuds de calcul Intel en RH7 jusqu'à juin 2024

3

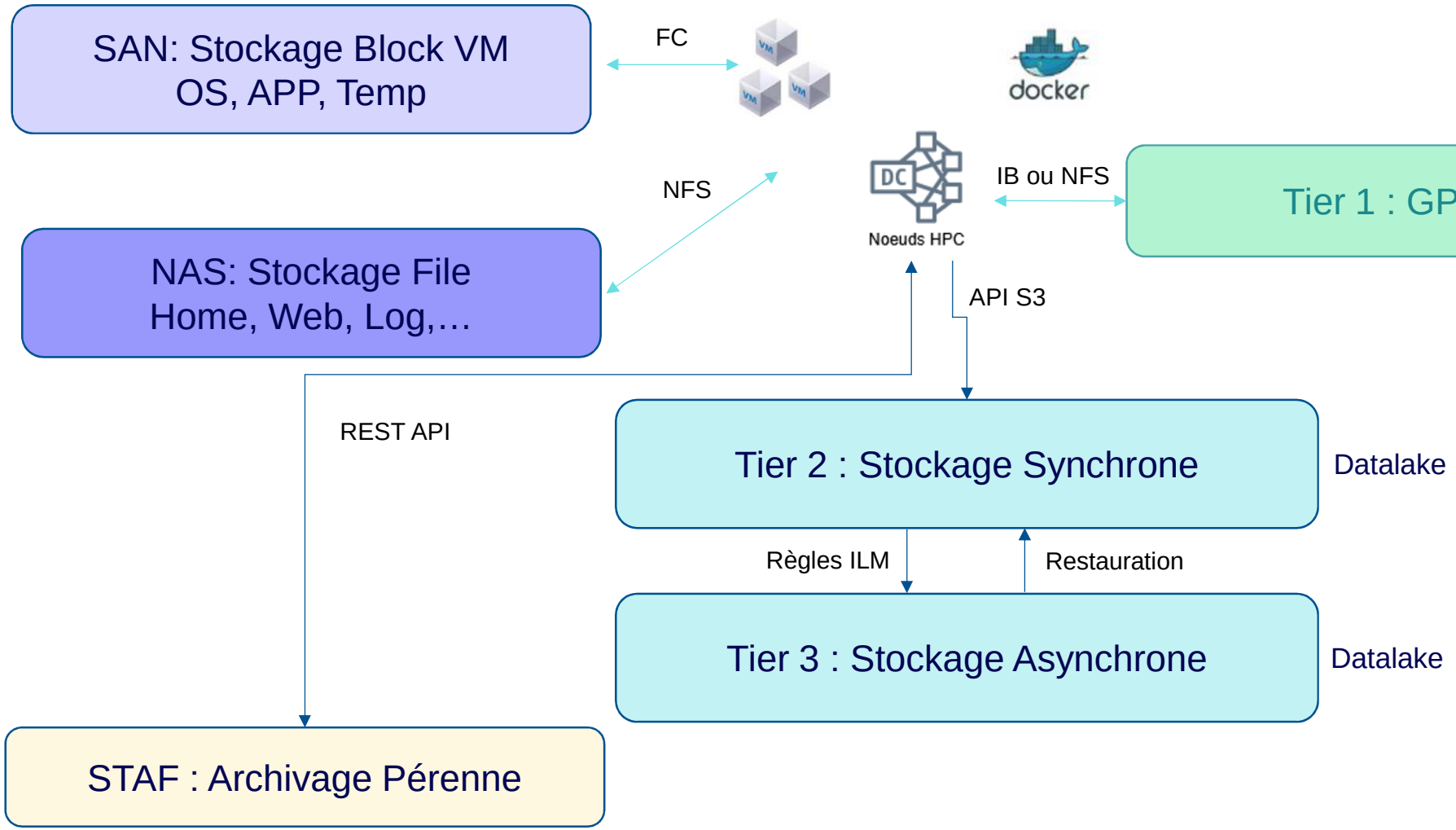
PLATEFORME DATALAKE

Stockage des données spatiales et scientifiques

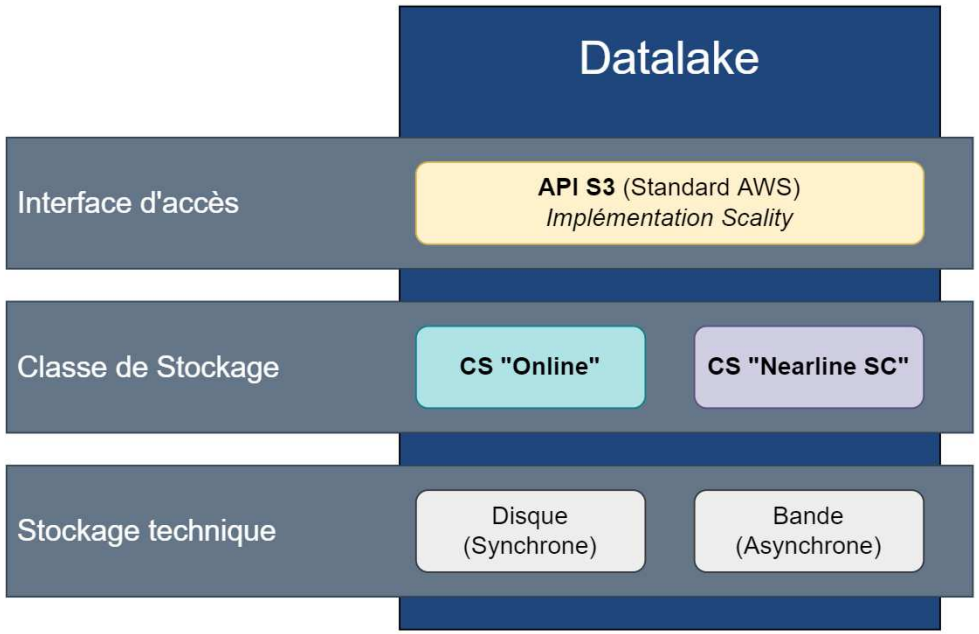
- **Les données :**
  - Données structurées fichier (raster, vectoriel, tableaux, binaires): NetCDF, Tiff (dont COG), JPEG, …
  - Associées à des fichiers Métadonnées (XML, JSON,…)
  - Catalogues multiples (axe d'amélioration majeur / Reste à faire)
  - 40 PO en 2023 -> 85PO (min) en 2027

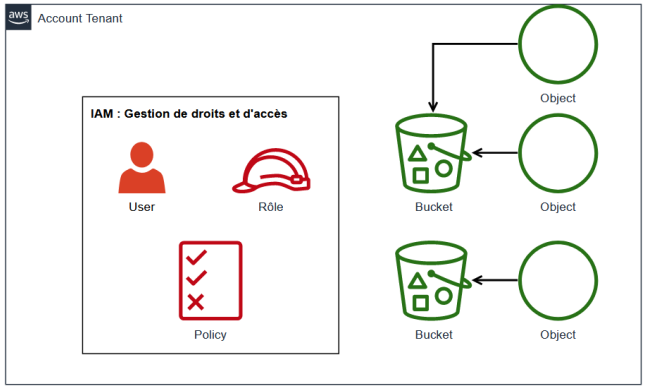
- **Pourquoi Disques et Bandes:**
  - Volume important / Pas besoin de toutes les données en ligne
  - Répondre aux besoins de performances des projets (Profondeur spatio temporelle) et des enjeux de maîtrise de l'impact environnemental (Stockage Froid à faible consommation électrique)
- **Object Storage:**
  - Object Storage: Changement de paradigme par rapport à l'historique POSIX pour préparer l'interopérabilité avec les approches Cloud / ESA)
  - Volume de stockage important à travers deux classes de stockage Disque et Bandes via une API conforme au standard AWS S3 et Glacier (35PO Disque et 35PO Glacier) = Même chemin d'accès
- **Intégration de la solution dans l'écosystème de traitement (Calcul) et diffusion (majoritairement Web)**
  - API S3 beaucoup plus mature au niveau des Framework de traitement (par rapport au début du projet en 2020)
  - Plus efficace pour diffusion Web (Milieu du chemin car Datalake CNES non exposé en dehors du SI Scientifique)

Le Datalake : Positionnement dans le SIS (Sys Information Scientifique)



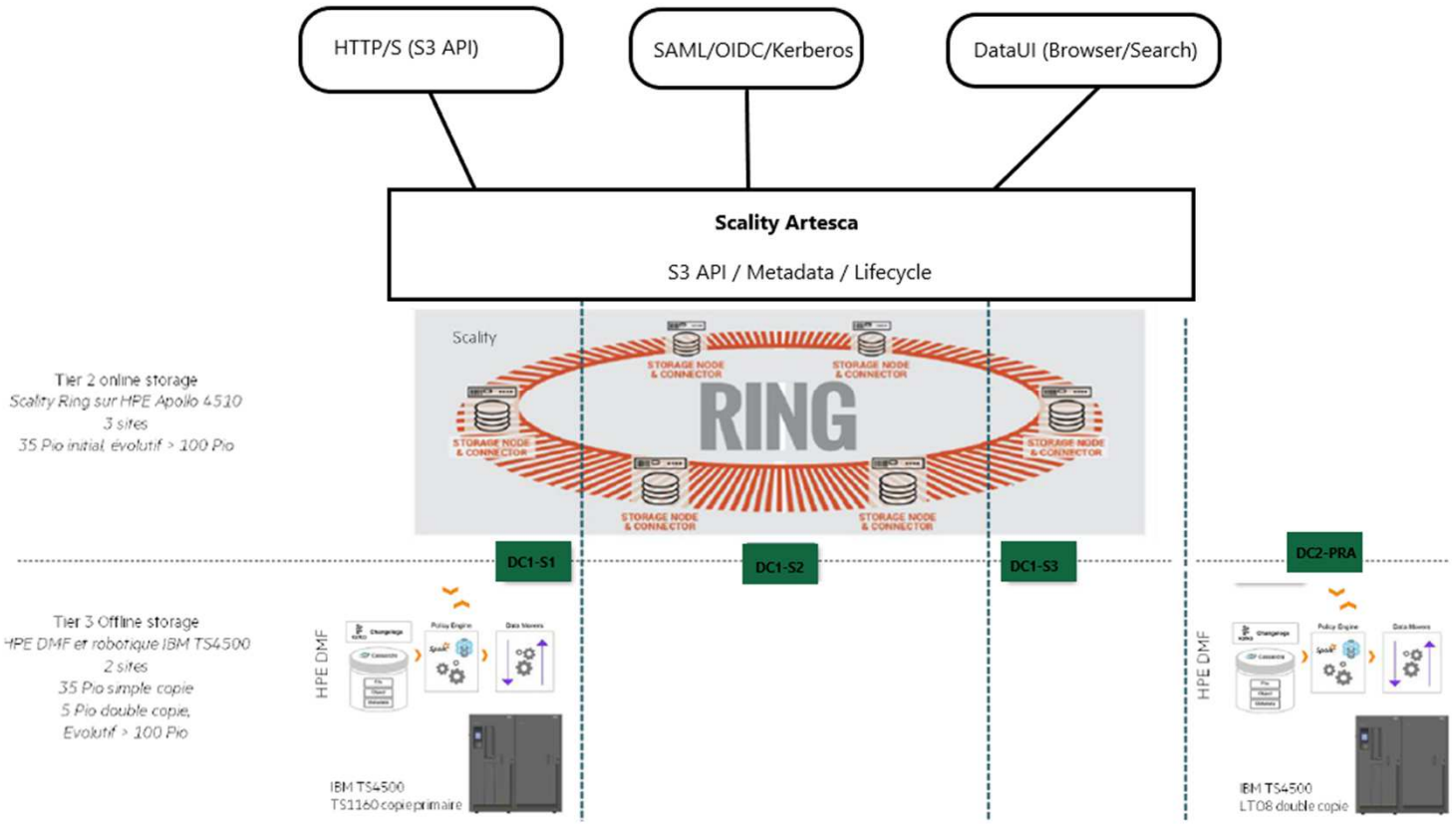
Le Datalake : Description Système





- **Interface S3 « AWS »:**
  - Implémentation « OnPremise » API Standard S3 (Ou presque)
  - Fonction « Glacier » type AWS pour Tape:
    - Objet Unique (Quels que soit la classe stockage, Même identifiant, path)
    - Metadata « storage class »
    - RestoreObject préalable si Tape
  - Pas de POSIX « natif »
  - Accès S3 = Token (access/secret key S3) + AWS STS (IPA CNES)
- **Politique d'accès:**
  - Policy : les politiques d'accès
  - Rôle : Entités détenant des droits d'accès à des ressources (bucket/objet) et pouvant être endossé temporairement
- **Tiering**
  - ILM (Information Lifecycle Management)
    - Règle de transition (sur metadata des objets ou tag spécifique)
    - Opérateurs S3 (CopyObject, …)
  - Classes de stockage fonction du besoin
    - Chaud (disque)
    - Froid (Tape) :
      - Simple Copie (majoritairement = 30PO)
      - Double copie réservé à l'usage « Pérennisation » = 5PO (au départ)
- **Fonctions manquantes (à ce stade):**
  - Quota au niveau Account ou Bucket (soft et hard)
  - Suivi de capacité « fin »
  - Métrologie des différents composants (pas que global plateforme)
  - Tiering en cours de déploiement (anomalies)

Le Datalake : Architecture Technique



❖ **Composants Techniques:**

- ➢ **Tier2** = Stockage en ligne
  - o Solution commerciale « software defined storage » (SDS)
  - o Scality Ring: Stockage distribué / Scality Artesca (XDM): S3 / ILM
  - o Concentre tous les accès utilisateurs (pas d'accès direct au tier3 bande)
  - o Beaucoup de « COTS » : KVM, K8s, … (packagé dans la solution scality)
  - o Protocole Sproxyd entre Artesca/Ring
  - o Basé sur serveur forte densité HDD (60HDD dans 4U)
- ➢ **Tier3** = Stockage Bande
  - o HPe DMF (7.6)
  - o Totalement asservie par les ILM Artesca
  - o API S3 pour interface avec le tier2 (Client S3 / Pas Serveur)
  - o En cours de déploiement (Qual only)
- ➢ **Métrologie**:
  - o Grafana intégré Scality
  - o + Stack ELK/Prometheus/Grafana (En cours)
- ➢ **Performance:**
  - o Tier2 : 8GBs Read/5GBs Write
  - o Tier3 :
    - Simple Copie: 1,5GBs Read / 1,5GBs Write
    - Double Copie: 0,5GBs Read / 1GBs Write

Retour d'expérience

Système et Infrastructure

- **Système**
  - Globalement mature sur les « fonctions » de base S3 (Des limites, anomalies sur le lifecycle et les policy)
  - Peu de solutions open source avec un niveau de compatibilité élevé à la spec aws
  - Interface Glacier non mature et toujours en cours de développement (REX Appel offre : Peu de solutions disponibles, même si progresse: Miria, PAG,…) = HPSS conservé > 12 mois supplémentaires

- **Constats:**
  - Quota (régression par rapport au POSIX)
  - Métrologie limitée / Supervision complexe
  - Nombre de composants (dans le contexte de cette installation: KVM, K8s, Kafka, Cassandra…) = Complexité de mise à jour / Faible automatisation (axe de travail prioritaire en exploit)
  - Scalabilité (pas si linéaire / Impact rebalancing de données)
  - Gestion des exceptions de transitions (Composants différents Scality / DMF)
  - Beaucoup de développement toujours en cours ou correctifs (Agile, mais impact projet)
  - Contrainte Datacenter = Régression vers approche PRA via « custom script », complexe

Coté Utilisateur

- **API S3:**
  - Beaucoup de clients différents avec des performances différentes

  (Demande un plus gros accompagnement / Activités additionnelles pour l'équipe calcul : Aller davantage vers des couches applicatives)
  - Privilégier les API haut niveau (et client) Haut niveau (Gestions des exceptions, MultipartUpload, reprise,…)
  - Métadonnées (Plutôt au niveau catalogue type OpenSearch/STAC versus S3) = Travail en cours

- **Données:**
  - Importance des formats de données pour optimisation cloud (accès par chunck, zip, Taille de chunk compatible avec la bandes)

  => Travail en cours sur Zar, COG, … versus GeoTiff, NetCDF / quid impact bande
  - Pas pour tous les cas d'usage: Pas IO « Random » sur S3 (Accès HTTP/Latences)
  - Librairies de donnée pas toutes matures pour accès S3 direct (Ex:netcdf même si avance)

- **Points Forts:**
  - Accompagnement des utilisateurs (Equipe Support Calcul et Usage)

  => Vraiment important pour assurer l'adoption du système / Succès dans notre contexte avec beaucoup de demandes projet (Charge environ 2 ETP)
  - Protocole S3 Natif / Bonne performance
  - Travaux de portage vers API S3 = Meilleure interopérabilité avec les tendances Cloud / ESA en cours (prépare les segments CNES)

Merci pour votre attention!

Contact : L-DTN-ISA-Calcul@cnes.fr