

CENTRE DE CALCUL DU CNES NOUVELLES PLATEFORMES HPC6G

Journée d'information HPC et Stockage de l'OMP, vendredi 10 février 2023

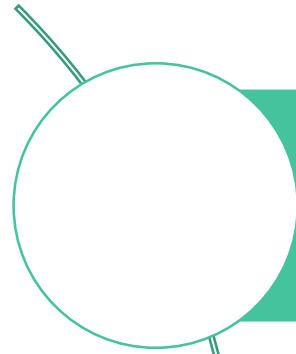
Centre de Calcul du CNES



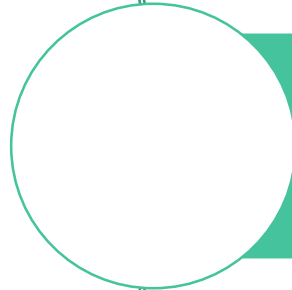
Sommaire



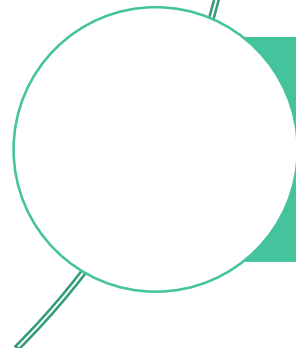
CENTRE DE CALCUL
DU **cnes**



Présentation du Centre de calcul du CNES



Plateformes HPC6G




Plateforme Datalake



CENTRE DE CALCUL DU CNES

Offre de service de l'équipe fondation Calcul



Centre de Calcul | Offre de service

Plateformes de Calcul

Support et Expertise

Datalake

Services et Application

cnes
CENTRE NATIONAL DES ETUDES EPITAXIALES



Offre de service du Centre de Calcul du CNES

- Plateformes de calcul (actuellement HPC5G) :
 - Diffusion limité (HAL) : 12000 cœurs de calcul, 40 GPUs et 8,5 Po de stockage
 - Diffusion Restreinte Special France : 576 cœurs et 130 To de stockage
 - SWOT dédiée au projet (2000 cœurs, 1Po de stockage)
- Plateformes de stockage :
 - Capacitif (Datalake/HPSS/PEPS) : stockage, diffusion et valorisation de données
 - Archivage (STAF) : pérennisation de données.
- Services et Applications :
 - Logiciels scientifiques (Python, C, NetCDF, GDAL, IDEs, etc.)
 - Datalabs (Jupyterhub, VRE, etc.),
 - Conteneurisation (Docker, Singularity, Podman)
 - Intelligence Artificielle (DL, ML, etc.)
- Support utilisateur et Expertise :
 - Support Utilisateurs
 - MCO des plateformes : N1, N2 et N3
 - Métrologie/Monitoring, Calcul Scientifique, Optimisations traitements HPC/Données, IA

Périmètre de l'offre de services

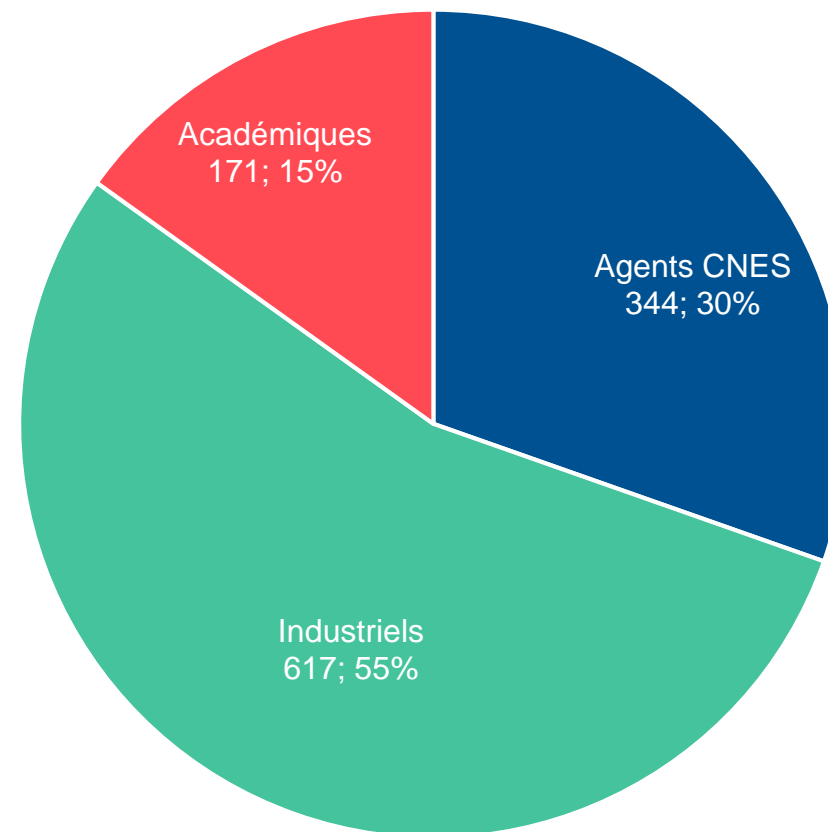
Ces services s'adressent aussi bien aux agents CNES qu'aux partenaires :

- **Tout agent CNES**, dans le cadre d'un projet ou en tant que simple utilisateur métier peut obtenir un compte de connexion à la plateforme de calcul.
- **Toute personne travaillant dans le cadre d'un projet CNES** peut obtenir les accès à la plateforme.
- **Des partenaires extérieurs**, laboratoires ou industriels, peuvent obtenir un accès à condition d'être soutenu par un responsable de projet CNES.

Qui sont les utilisateurs du Centre de Calcul?

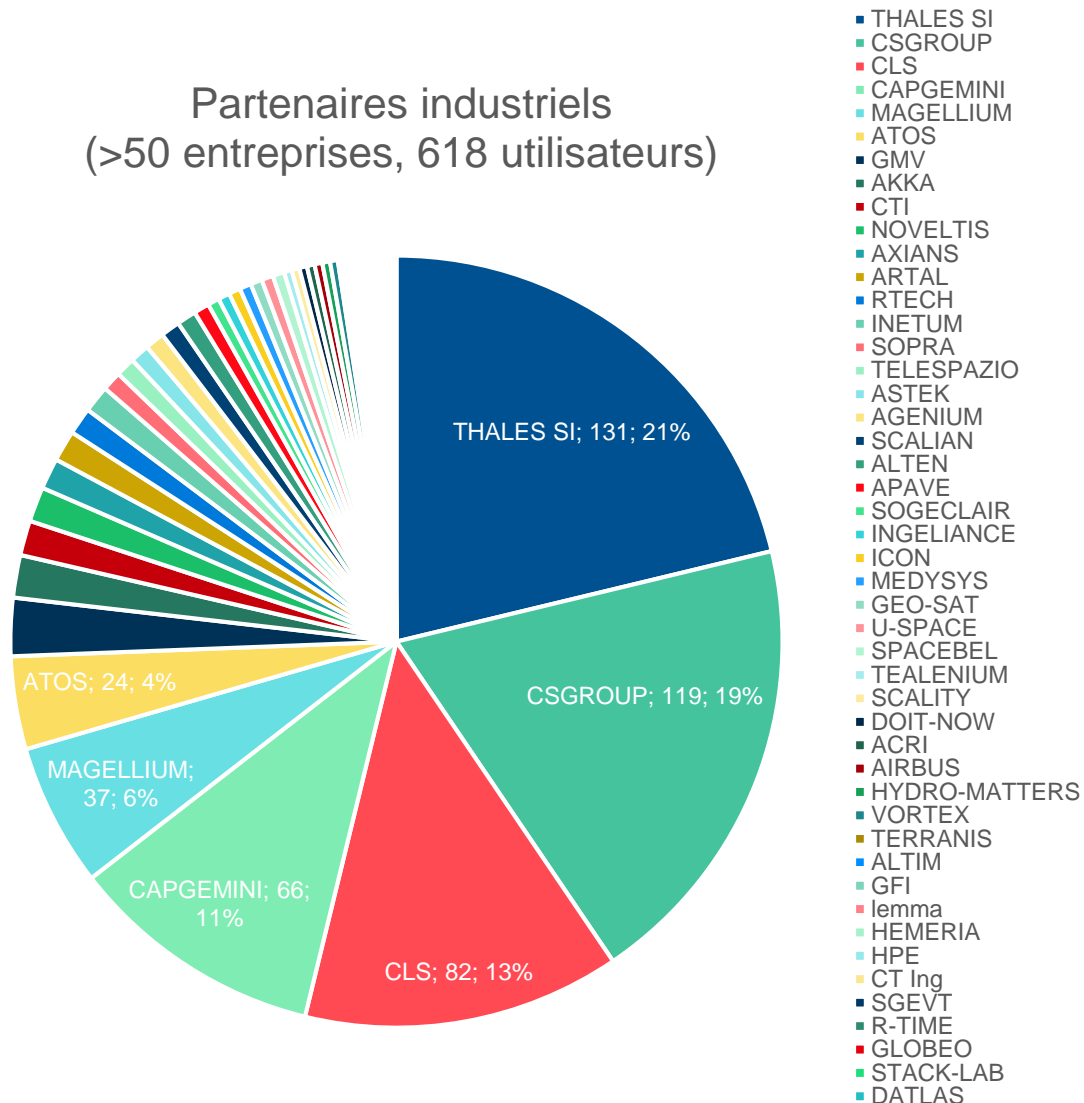
1133 utilisateurs différents se sont connectés au supercalculateur HAL en 2022

- + de 1000 utilisateurs en 2022
- provenance diversifiée (CNES, industriels, académiques)
- compétences HPC « variées » : du débutant à l'expert
- des usages très divers : HTC, HPC, IA, ... pour de la production, du développement, des Proof of Concept, ...

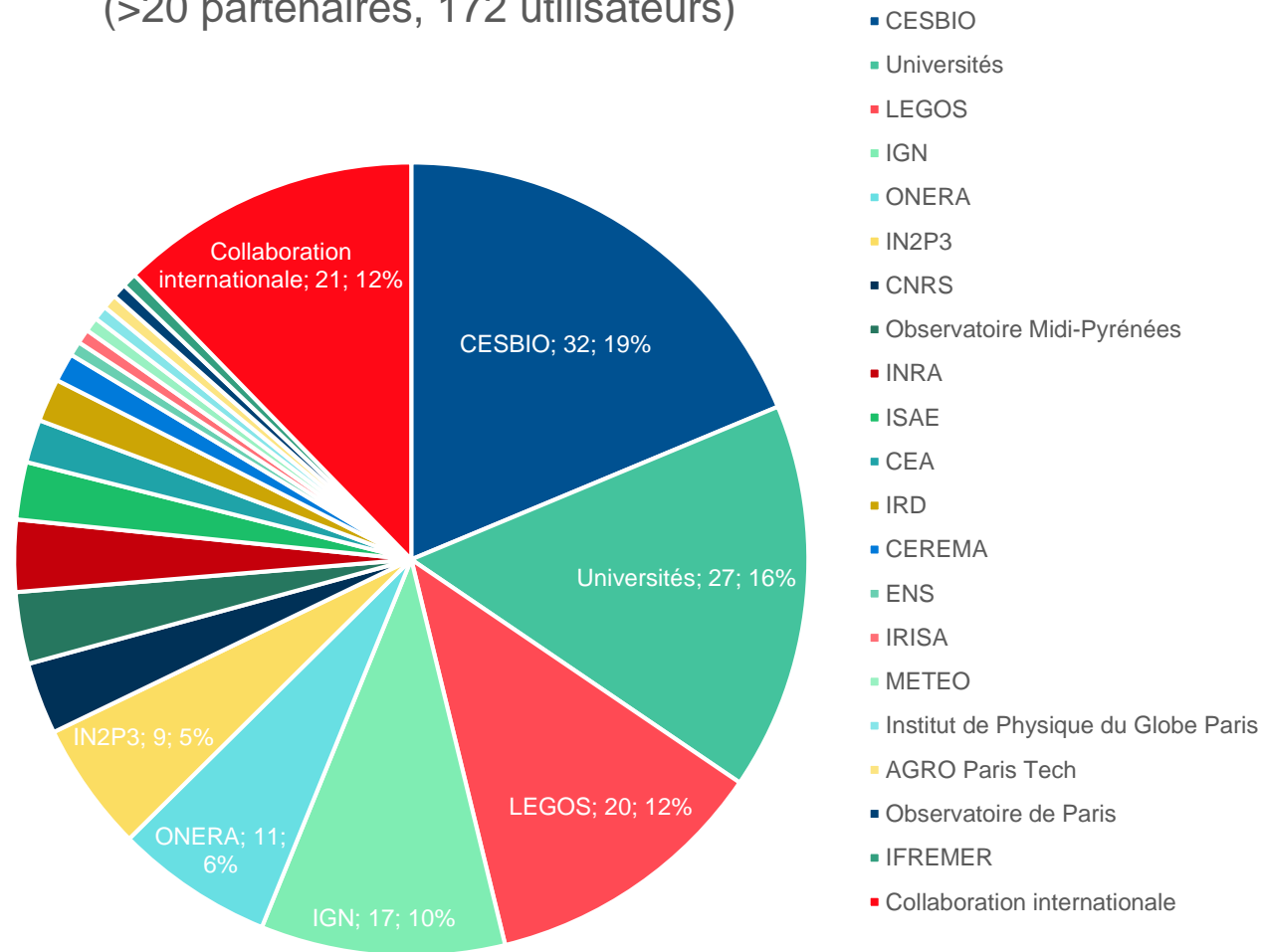


Utilisateurs du Centre de Calcul

Partenaires industriels
(>50 entreprises, 618 utilisateurs)



Partenaires académiques
(>20 partenaires, 172 utilisateurs)



Pour quels usages?

❖ R&D, ÉTUDE, PHASE AMONT

préparer un nouveau lanceur, simuler de la donnée issue d'un nouveau type de capteur ou préparer les algorithmes de traitement de données des futurs instruments lancés à bord des satellites



❖ PRODUCTION ET DIFFUSION DE DONNÉES

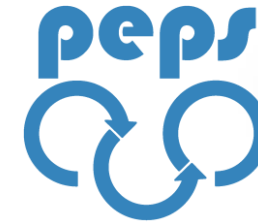
PEPS qui récupère les données Sentinel depuis l'ESA, les diffuse et propose des traitements à la demande
THEIA avec atelier de production Muscate : produits de niveau avancé autour de l'observation de la terre

CFOSat produit des variables sur les surfaces océaniques

SSALTO : le segment sol d'altimétrie multi-mission : retraitements missions Jason 2, Jason 3, ...

SWOT : altimétrie océans, lacs et rivières (plateforme dédiée)

Microcarb : prochainement pour le suivi du CO2 de l'atmosphère.



❖ STOCKAGE, DIFFUSION, VISUALISATION, ET VALORISATION DE DONNÉES POUR LA RECHERCHE

distribuer et exploiter toutes les données produites par des centres de mission

Manipulation de données ou Machine Learning

Les services Datalake et Jupyterhub/Datalabs sont des éléments clefs

Laboratoires sous tutelle du CNES (CESBIO, LEGOS)

Projets CNES : Hysope II, GeoDataHub, AI4Geo.



Quelques chiffres-clefs

❖ Fin 2022 (HPC5G + HPSS + STAF)

~ **1000** utilisateurs

725 Tflops, **12 300** cœurs de calcul, 48 cartes GPU, **8 PiB** stockage disque rapide

Datalake v1 = 20 PO Bandes / 2PO Tampon GPFS (valorisation de données)

STAF v3 = 3PO Long terme Archive (CST)

65 millions d'heures CPU consommées en 2022

Big data spatial : **20 Po** de données stockées dont **3 Po** archive pérenne

❖ Fin 2023 (HPC6G + Datalake + STAF)

~ **1200** utilisateurs

1 Pflops, **17 312** cœurs de calcul, 60 cartes GPU, **10 PiB** stockage disque rapide

Datalake v2 = Object Storage (S3 accessible directement par les utilisateurs du Système d'Information Scientifique), 35 PO Disques / 35 PO Bandes

STAF v4 = 5 PO en 2023 Long Terme Archive (évolutif à 15PO sous 5 ans)

142 millions d'heures CPU disponibles par an

Big data spatial / Diffusion de la donnée : **70 Po** Datalake (35 Po Disque et 35 Po Bande) / **7 Po** STAF (Archive perenne)



PLATEFORMES HPC6G

Objectifs du renouvellement des plateformes de calcul

- Gérer l'obsolescence (matériel actuel 2016)
- Augmenter de la capacité
 - **+25% de ressources lié à la croissance du nombre d'utilisateurs**
- Favoriser l'inter-opérabilité (Partenaires, centres de calcul, fournisseurs CLOUD)
- Proposer de nouveaux services
 - **Kubernetes**
 - **Monitoring/Métriologie pour les utilisateurs**
 - **Maîtrise de la consommation d'énergie**

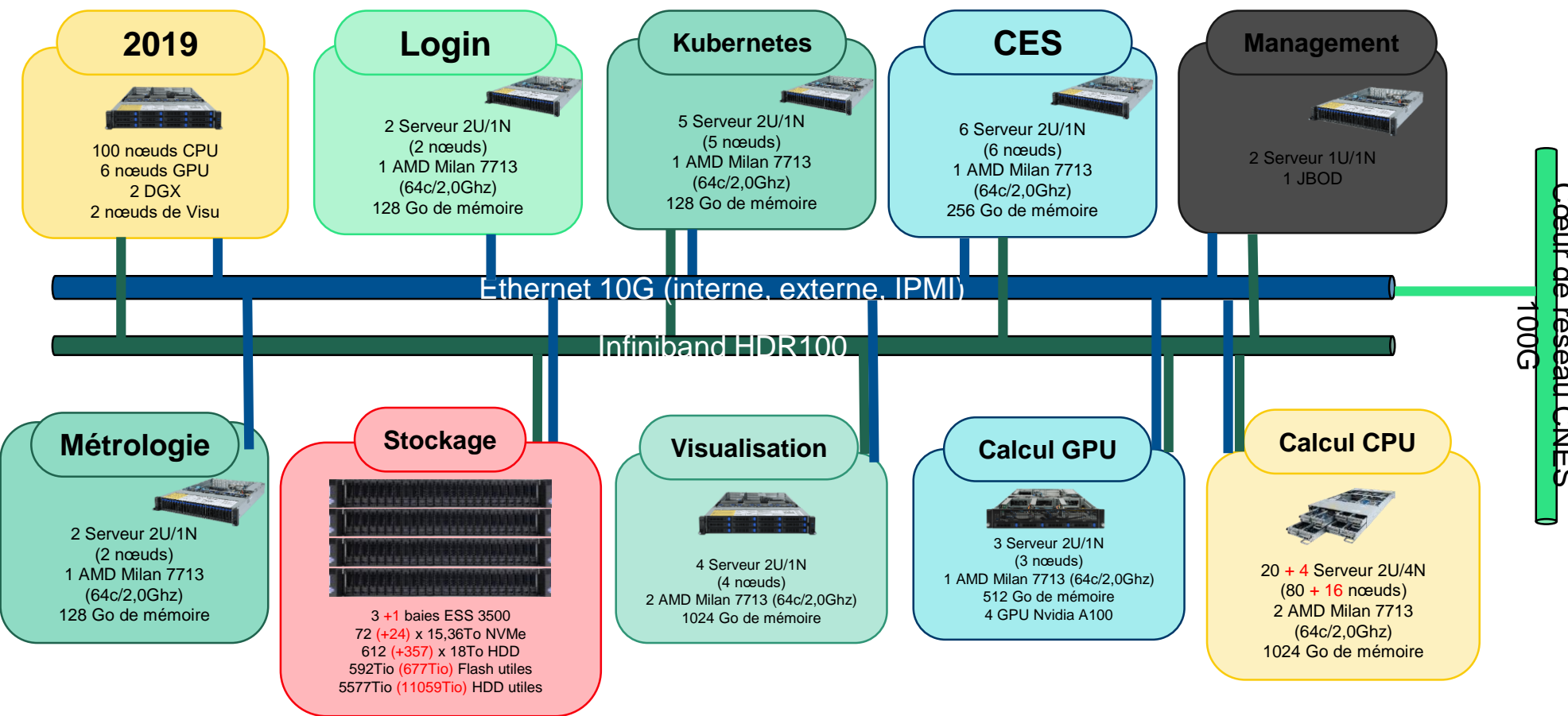
HPC6G Diffusion Limitée : mise en production juin 2023

Plateforme DL	CPU	GPU	Stockage (Po)
Réutilisation de matériel	4000 (Intel CascadLake 40 cores, 2,1 GHz)	40 (24xV100 + 8xA100 + 8xT4)	-
Nouveau matériel	12 288 cores	12 cartes A100	10 Po
Total	16 288 CPU	52 GPUs	10 Po

- **Processeurs AMD** : 96 serveurs bi-socket
- **12 GPU** : **NVIDIA A100** 80Go Pci-e
- **10 PiB** de stockage POSIX : Système de fichier // et distribué **IBM Spectrum Scale** basées sur baies « intelligentes » **ESS-3500**
- Réseau Faible latence **Mellanox Infiniband HDR-100** : **100Gb**
- Réseau **Ethernet 10G**
- Orchestration Conteneurs : **Kubernetes**
- Système d'exploitation : **RedHat 8.X**
- Orchestrateur Jobs HPC : **SLURM**

Nœud de calcul	
Nb de sockets/serveur	2
Nb de coeurs/CPU	64
Type de processeurs	AMD EPYC Milan 7713, 64C (2.0GHz-225W)
Mémoire par noeud	1024 Go mémoire DDR4 (16*64 Go@3200MT/s Dual Rank)
IO/noeud par noeud	1 disque NVMe de 3.2 To U.2

Architecture globale



10 240 CPU : Processeurs AMD
(64 Cœurs par CPU)

12 GPU : NVIDIA A100 80Go Pci-e

10,2 PiB de stockage Posix : Système de fichier parallèle et distribué **IBM Spectrum Scale** basées sur baies intelligentes **ESS-3500**

Réseau HP Infiniband HDR-100

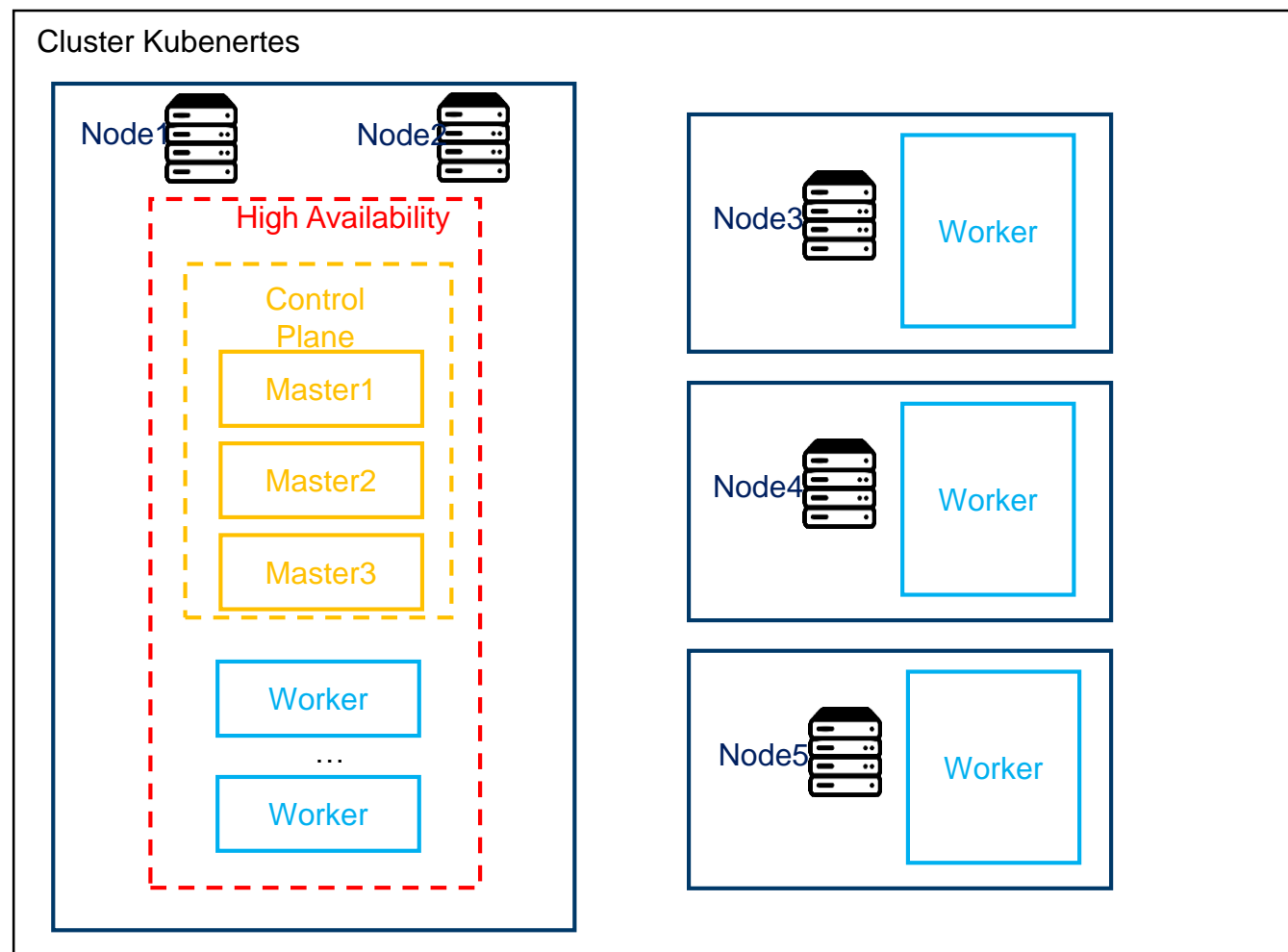
Réseau Ethernet 10G

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

Orchestrateur Jobs : **SLURM**

Cloud – Kubernetes

- Réserver des ressources de calcul pour un temps assez long (plusieurs mois), pour effectuer un traitement continu (par exemple une campagne de retraitement de données annuelles ou encore une phase de recette/qualification applicative)
- Soumettre des jobs conteneurisés et adapter la taille du cluster en fonction de la demande.
- Réserver une machine déconnectée des données pour faire des tests, de la validation ou de l'exploration technologique.
- Réserver des ressources de calcul éphémères pour faire de l'intégration continue.
- Construire et détruire des clusters de calcul type Big Data en utilisant des recettes d'infrastructure as code pour des besoins ponctuels (< 6 mois).



Métrologie et monitoring orienté utilisateurs/projets

Par login :

- Nombre de jobs, utilisateurs uniques, temps elapsed moyen, nombre de nœuds par job
- Tableau de bord permettant d'afficher les métriques d'un job (utilisation CPU, mémoire, kWh et CO2eq)
- Détection des mauvais usages (%utilisation CPU faible, sur-réservation de ressources,)

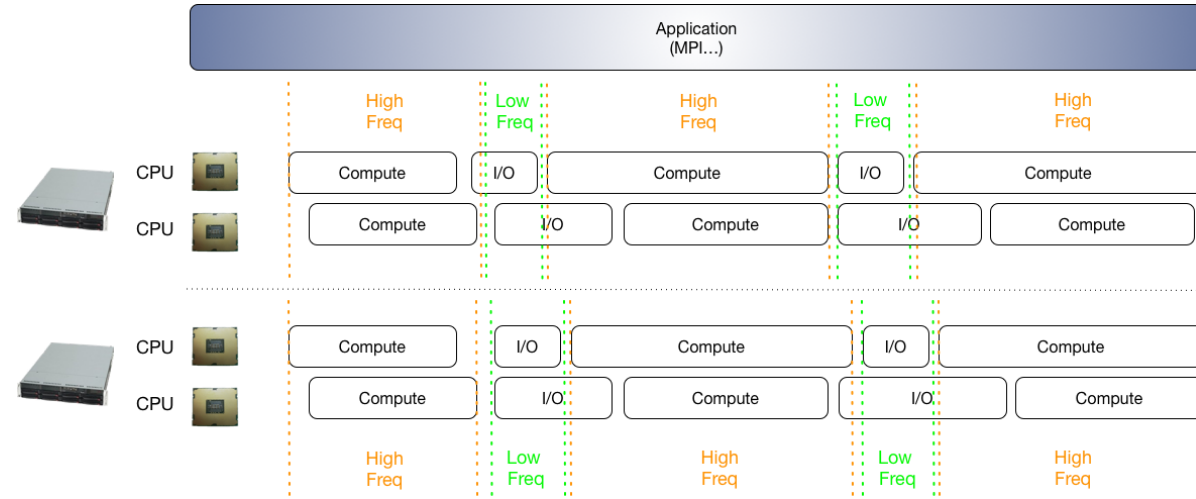
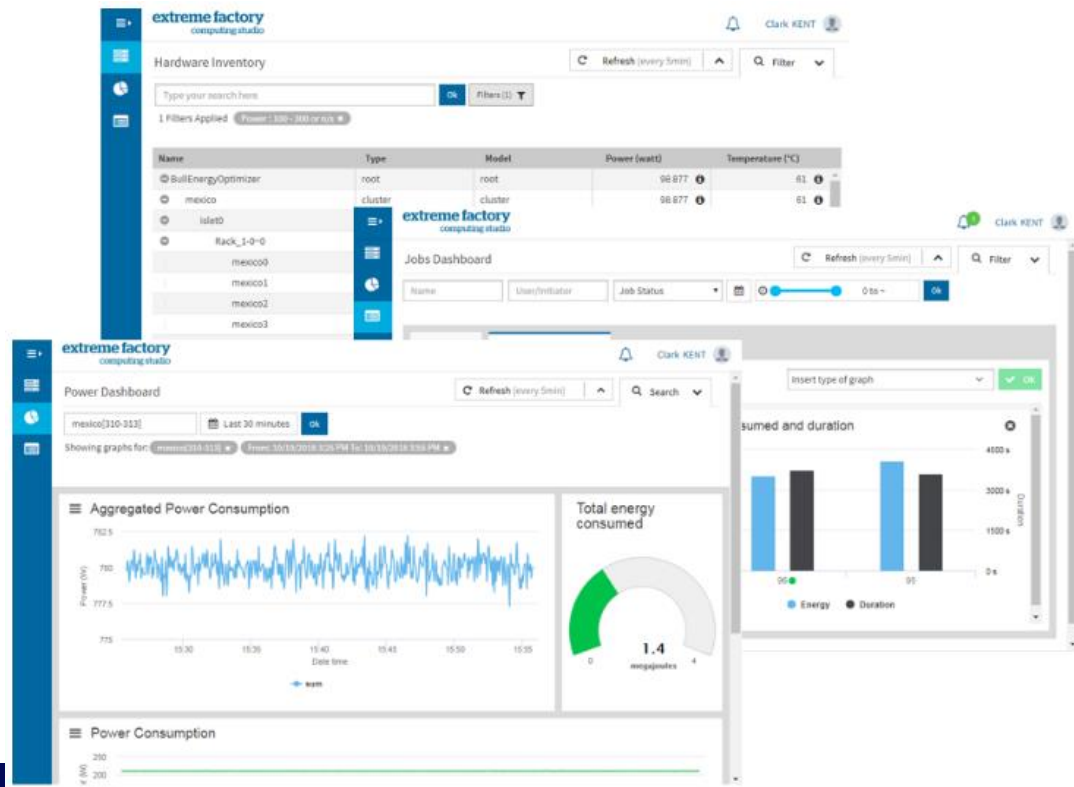
Par projet :

- Décompte des nombres d'heures de calcul consommées
- Décompte du stockage utilisé (Go/mois)
- Interfaçage avec le Centreon du CNES pour les tableaux de bord des projets opérationnels

Gestion de l'énergie – BEO et BDPO

Bull Energy Optimizer (BEO) :

- Informations détaillées de puissance et de consommation électrique
- Au niveau des applications (Utilisateurs finaux et développeurs)
- Au niveau plus global d'un groupe ou de la totalité des infrastructures de calcul
- Associé à SLURM
- Fourni la consommation d'énergie associée à chaque job en incluant la consommation des nœuds de calcul, des commutateurs et des baies de stockage.



Bull Dynamic Power Optimizer (BDPO) :

- Ajustement dynamique du niveau de puissance électrique délivrée au processeur en fonction des types de calculs effectués par l'application.
- Amélioration du « Energy to Solution » sans perte significative sur le « Time to Solution » en fonction du profil des applications.
- Définir des seuils afin de faire varier la fréquence/voltage des processeurs via DVFS en fonction de la charge CPU (analyse basée sur les IPCs).

permet de réduire la consommation des travaux tout en minimisant l'impact sur le temps d'exécution.

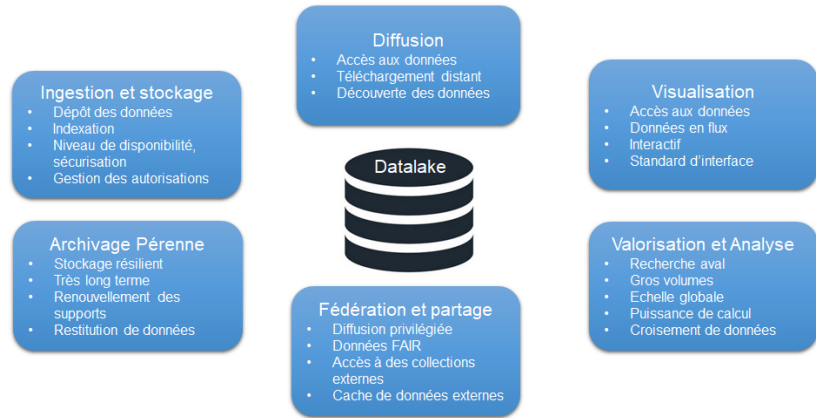
L'utilisation de BDPO est transparente pour l'utilisateur => une option sur la ligne de soumission Slurm.



PLATEFORME DATALAKE

Projet Datalake : Refonte Infrastructure de stockage des données scientifique

❖ Usages multiples

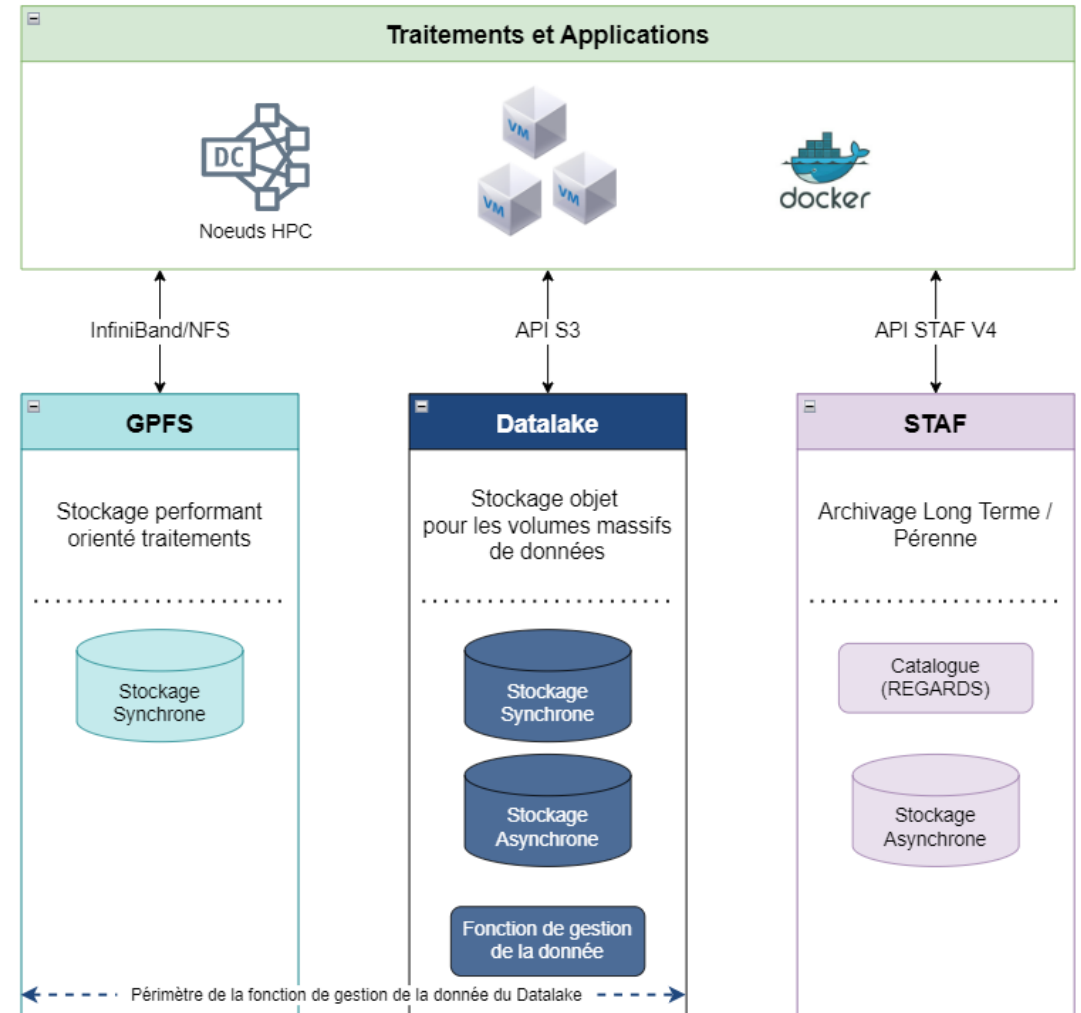


❖ Deux services différents:

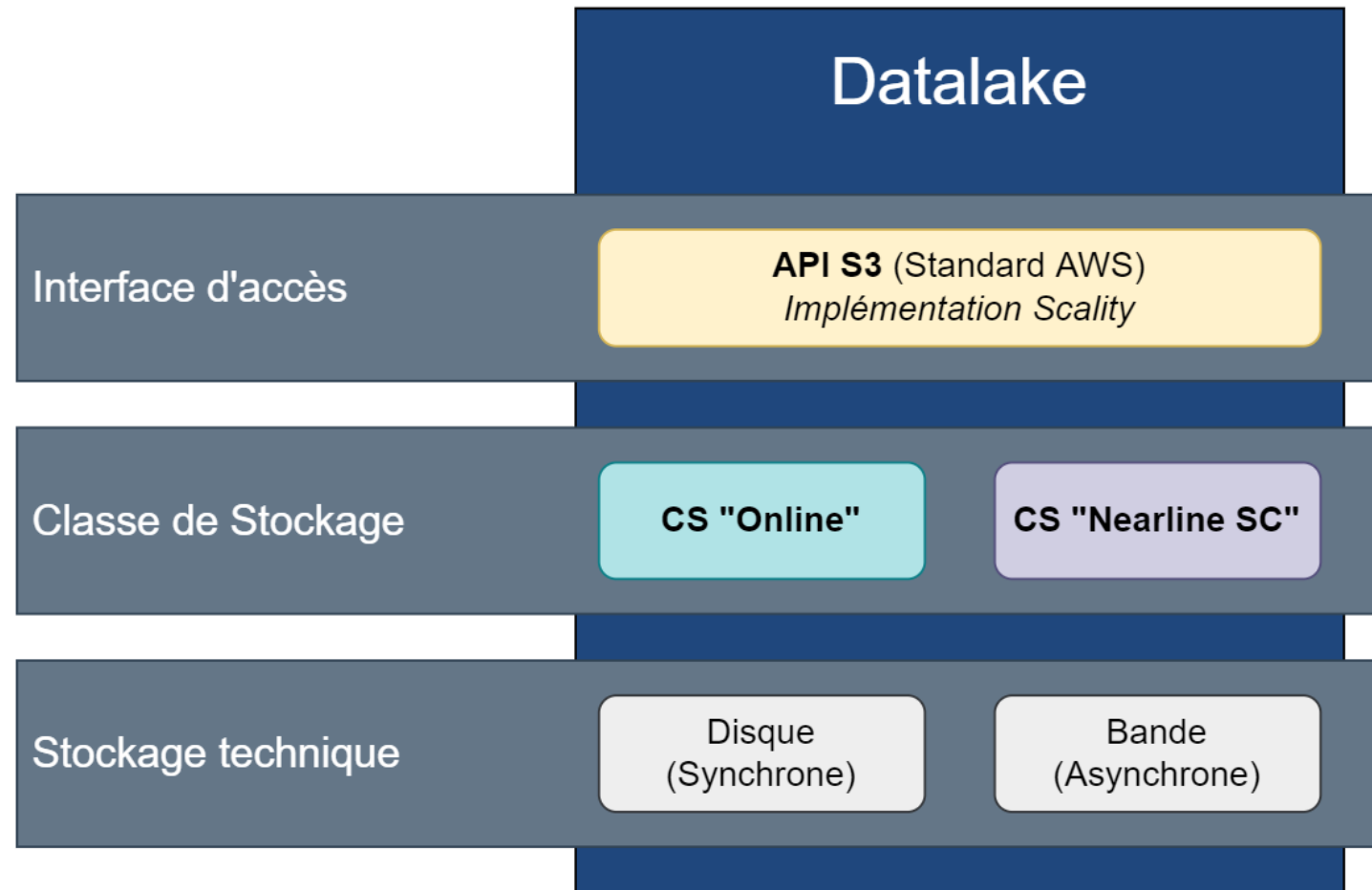
- Datalake = Stockage de masse (accès S3)
- STAF = Archivage pérenne de données
 - API Spécifique (HTTP REST)
 - Fonction catalogue / Intégrité des données
 - Partage l'infrastructure de stockage datalake

❖ Accès S3 et Stockage disque en prod = **10/11/2022**

❖ Qualification globale = **Prévue au 01/09/2023**



Le Datalake : Principes de bases



❖ Interface S3 « AWS »:

- 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)

❖ 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 pour le moment
 - Double copie réservé à l'usage STAF (au départ)

❖ Quota et Métrologie:

- Quota Soft (En cours de dev) : Niveau Bucket
- Quota Hard (Roadmap) : AWS niveau account
- Métrologie des différents composants

Le Datalake : Principes de bases

Des nouveaux concepts :

❖ Organisation des données

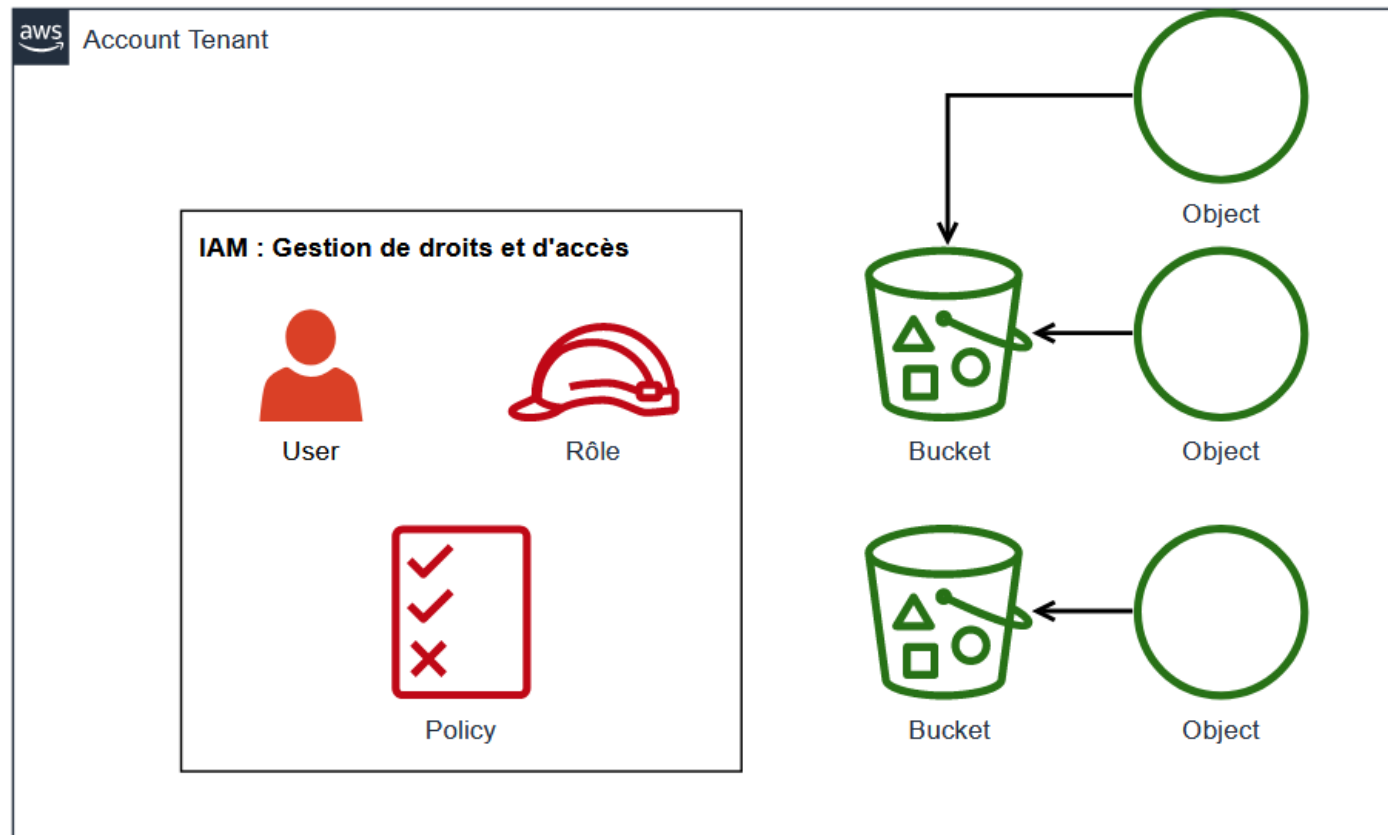
- Tenant / Projet (Account au sens AWS)
- Buckets à l'intérieur des tenants
- Des objets placés à plat au sein des buckets
- Les Préfixes (Pseudo arborescence)

❖ Les Comptes utilisateurs:

- User/Group : utilisateurs/groupes IAM authentifiés avec une paire de clé
- Compte IPA
 - En endossant un rôle
 - Des credentials (paire de clé + token) tmp

❖ Les Accès:

- Policy : les politiques d'accès
- Rôle : Entités détenant des droits d'accès à des ressources (bucket/objet) et pouvant être endossé temporairement



Le Datalake : Principes de bases

Le Datalake respecte pour les APIs les plus répandues, le standard AWS.
La [documentation AWS S3](#) est une excellente base pour commencer !

❖ Utiliser des clients S3 qui permettent de manipuler les données

- Des APIs haut-niveaux :
 - Simuler les commandes POSIX (ls, cp ...)
 - Uploads/downloads partitionnés (MPU et range request)
- Des APIs bas niveaux : get-object, put-object, head-object, list-object

❖ Prendre en compte les caractéristiques du stockage objet

- très performant en lecture mais comparativement peu performant en écriture.
- organisation des données : metadata personnalisable, tags ...
- formats de données plus ou moins adaptés (en terme de performance) pour le S3
- Bibliothèques de donnée sans support natif S3: Exemple : Netcdf (évolution en cours chez Unidata / Non testé au CNES)



SDKs



Android



iOS



Java



JavaScript



.NET



Node.js



PHP



Python (boto)



Ruby



Xamarin



AWS CLI



AWS Toolkit for Eclipse



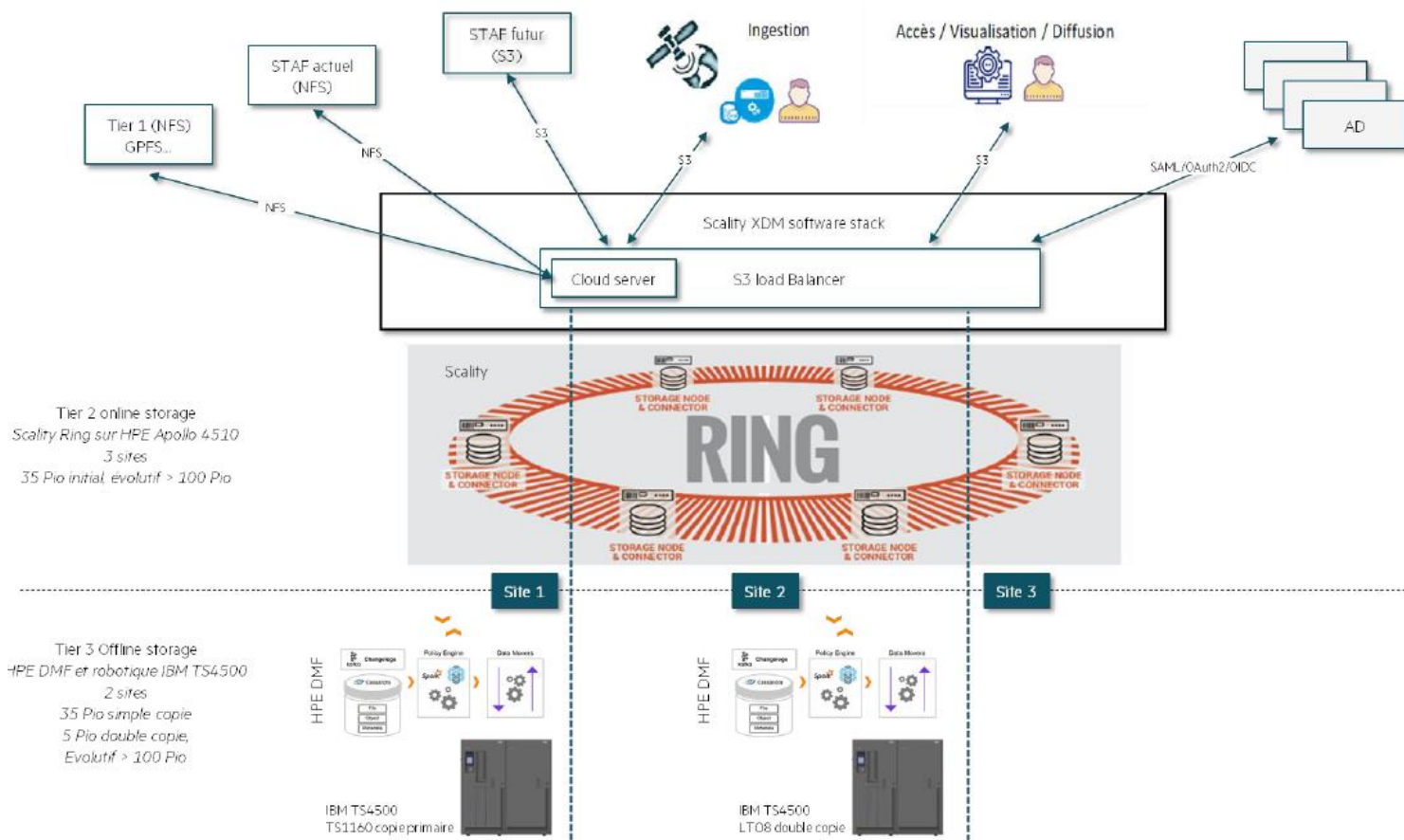
AWS Toolkit for Visual Studio



AWS Tools for Windows PowerShell



Le Datalake : Architecture Technique



❖ Composants Techniques:

➤ Tier 2 = Stockage en ligne

- Solution commerciale « software defined storage » (SDS)
- Scality Ring = Stockage distribué Scality Artesca (XDM) = S3 / ILM
- Concentre tous les accès utilisateurs (pas d'accès direct au tier3)
- Beaucoup de « COTS » : KVM, K8s, ... (packagé dans la solution scality)

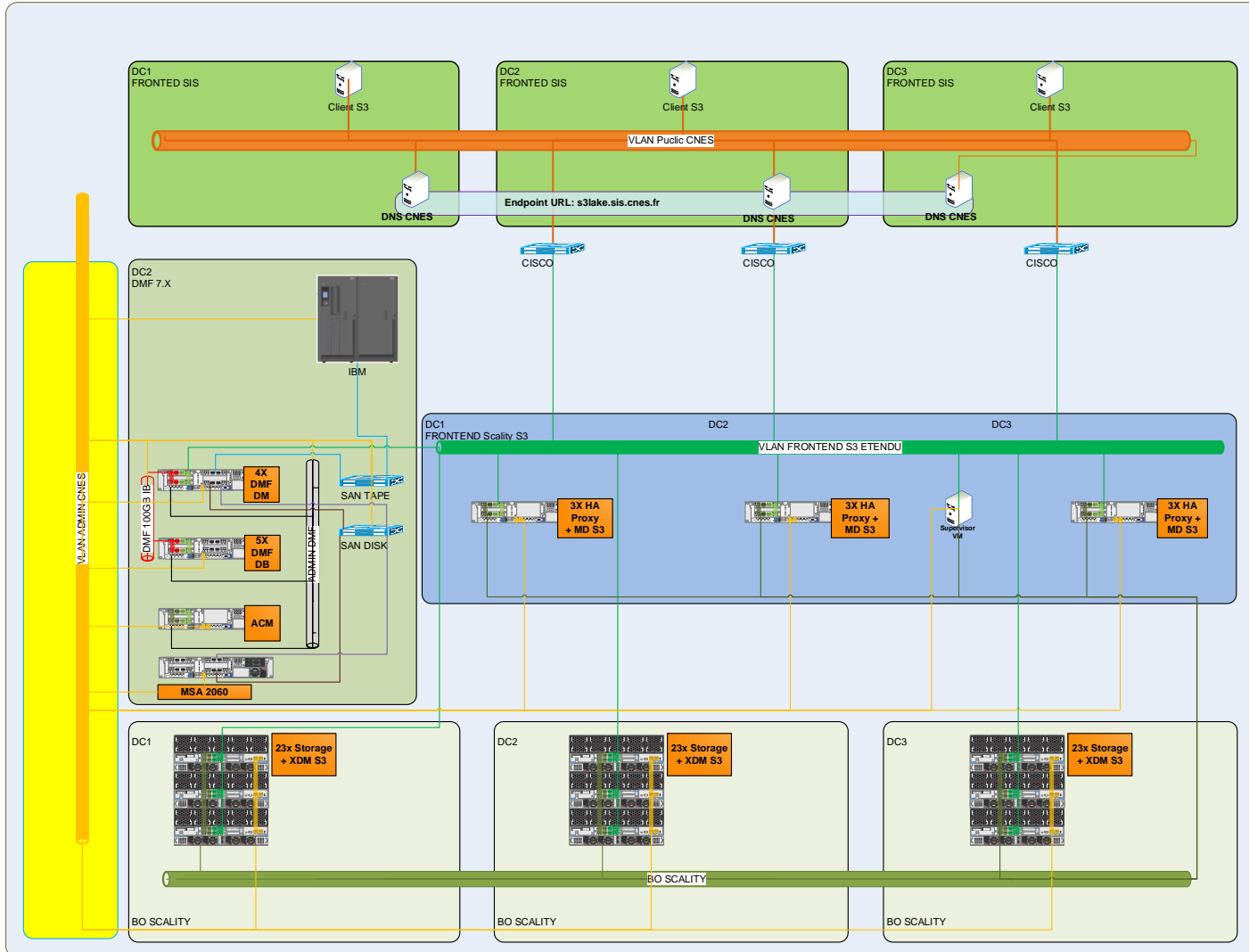
➤ Tier3 = Stockage Bande

- HPe DMF (7.6)
- Totalement asservie par les ILM Artesca
- API S3 pour interface avec le tier2 (Client S3)
- En cours de déploiement...

➤ Métrologie:

- Grafana intégré Scality
- + Stack ELK/Prometheus/Grafana

Le Datalake : Architecture Technique



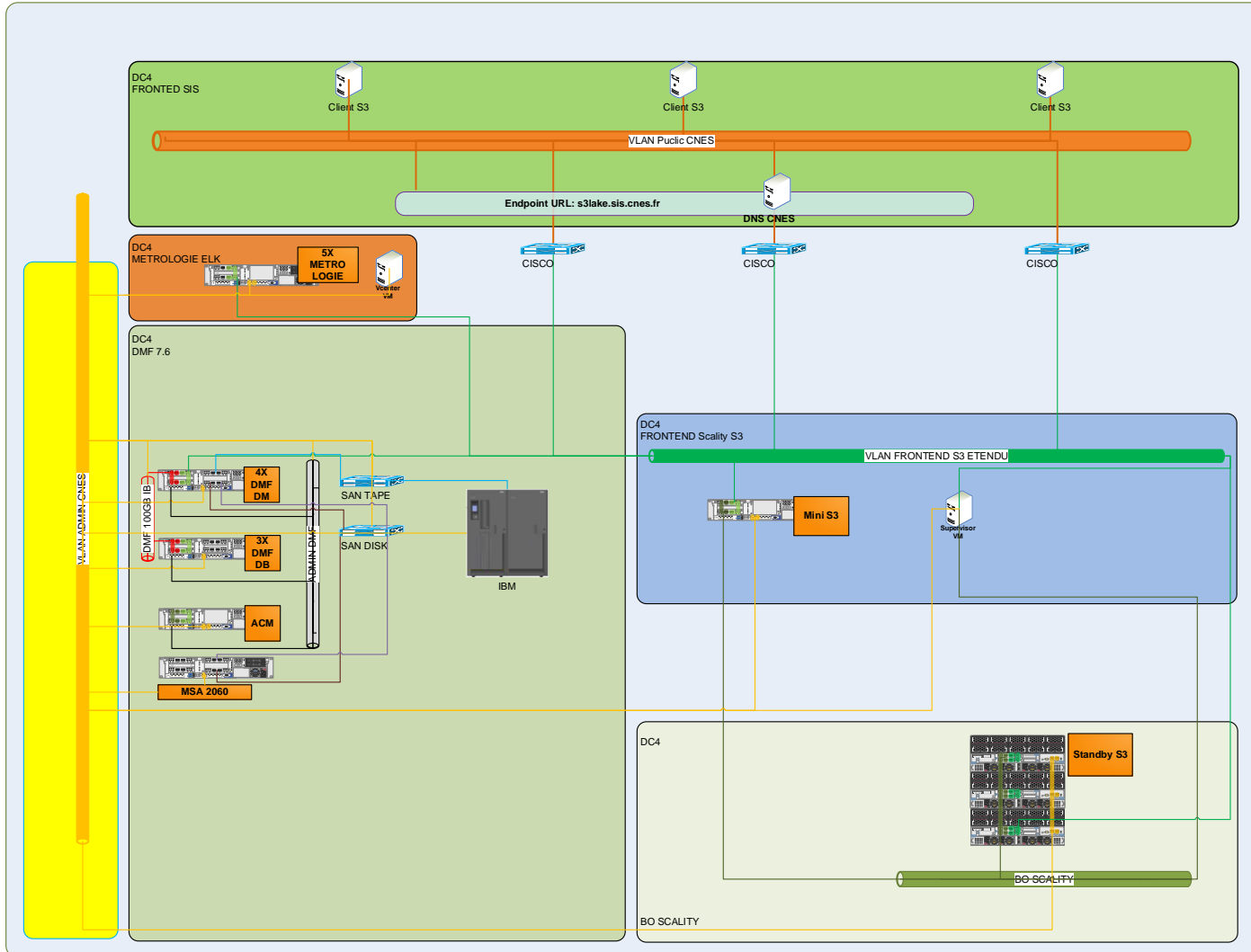
❖ Batiment Galois:

- 3 Datacenter (3 Salles contigus !)
- Scalcity:
 - 69 Serveurs Hpe Apollo 4510 (60 HDD 10TO / NVMe Metadata Ring)
 - 9 HA Proxy et Metadata S3
 - Redondance Scalcity Ring/Artesca à la perte d'une salle
 - Performance
 - 8GBs Read (44 GBs pic)
 - 5GBs Write (24GBs pic)

➤ DMF Principal:

- Tape :
 - TS4500
 - Drives TS1160 (JD/JE)
 - 5 Frames
 - >35 PO
- Performance:
 - Simple Copie: 1,5GBs Read / 1,5GBs Write
 - Double Copie: 0,5GBs Read / 1GBs Write
- DMF PRA avec second Bâtiment

Le Datalake : Architecture Technique



❖ Bâtiment Fermat:

- Datacenter PRA (peu de m2)
- DMF Secondaire
 - Double Copie
 - Tape
 - IBM TS4500
 - Drives LTO8
 - >5PO
- Scality PRA:
 - Uniquement pour maintenir l'accès à la double copie Bande
 - Point d'accès S3
 - Metadata S3
- Un design qui doit évoluer...

QUESTIONS

ANSWERS

Merci pour votre
attention!