Centre de Calcul
de l'Institut National de Physique Nucléaire
et de Physique des Particules

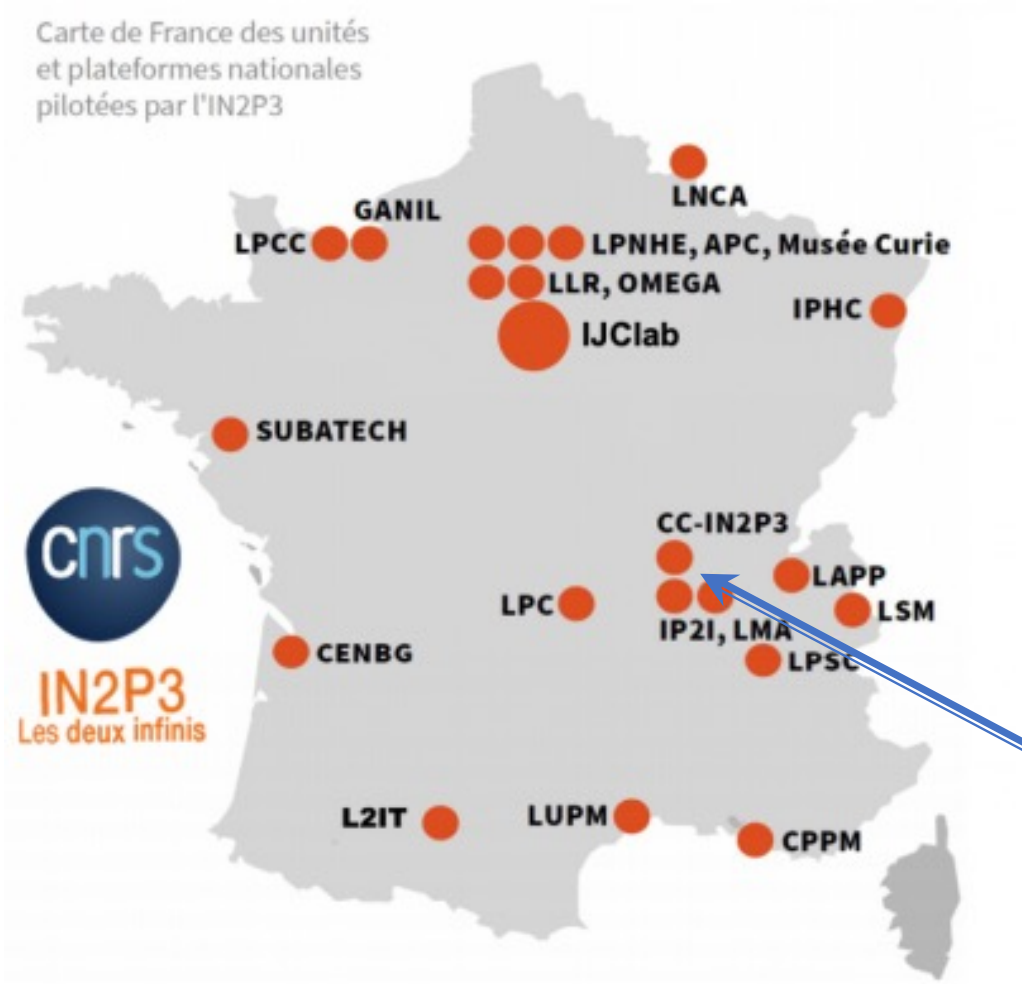
Stockage au CC-IN2P3: migration et valse des octets

Jean-Yves Nief

Aperçu

- CC-IN2P3 : contexte et missions.
- Vue d'ensemble du service de stockage au CC-IN2P3.
- Migration des données : un travail récurrent.
  - Migrer sans arrêter la production.
- Migrations des données sur disque.
- Migrations des données sur bandes magnétiques.
- Bilan et perspectives.

Qu'est-ce que le CC-IN2P3 ?



IN2P3:

- 1 des 10 instituts du CNRS.
- 19 labos dédiés à la rechercher en physique des particules, physique nucléaire, astroparticules et autres disciplines.

CC-IN2P3:

- Ressources informatiques fournies aux projets supportés par l'IN2P3 (collaborations nationales et internationales).
- Ressources ouvertes aux scientifiques français et étrangers.
- 2 salles informatiques (2 x 850 m²):
  - 2000 serveurs.
  - 800 serveurs virtuels.

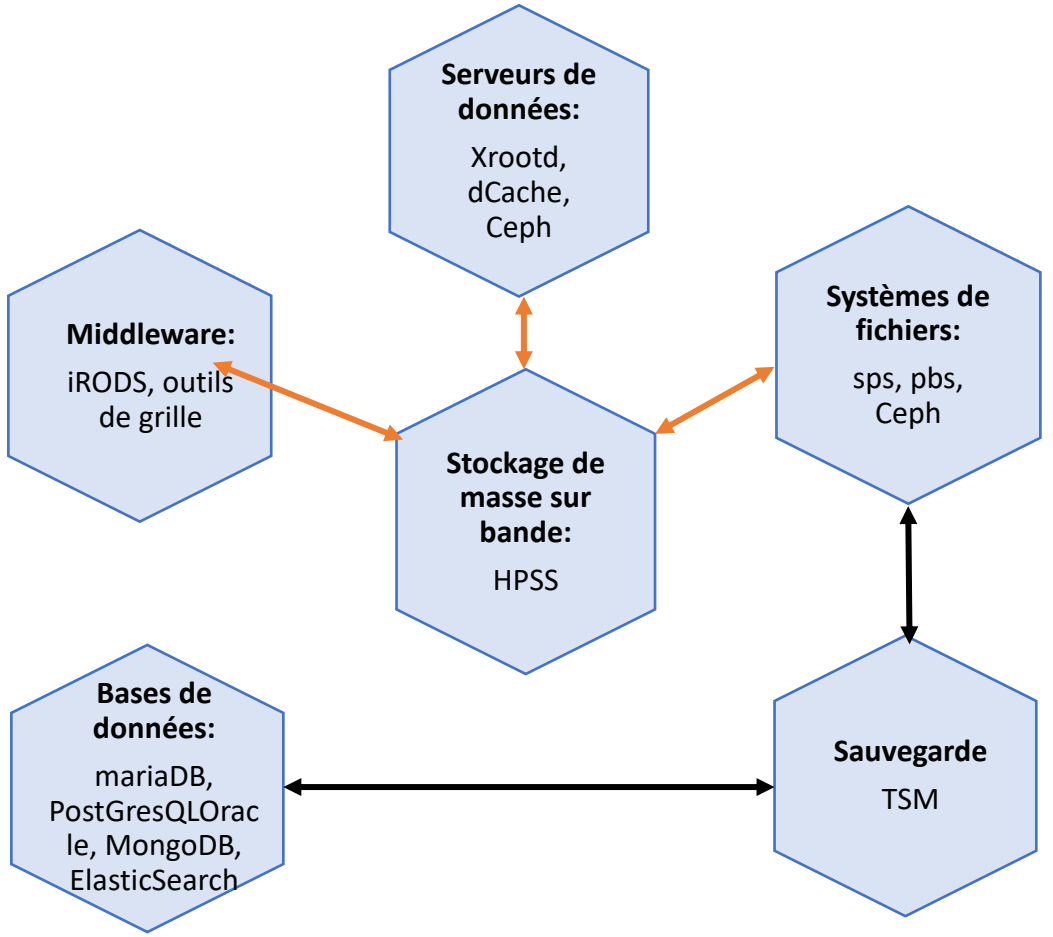
Qui utilise le CC-IN2P3 ?

Le service de stockage

En cours

Fouille de données:
Spark

Archivage



Staging

Sauvegarde

Quelques chiffres

- Équipe de 11 personnes.

- 14 technologies de stockage:
  - Forte hétérogénéité.
  - Pas d'outil universel pour tous les besoins.
  - 3 milliards de fichiers stockés.
  - 195 PiB.
  - Accroissement rapide: + 1-2 PiB / mois
  - Permet de servir ~1200 machines de calcul pour le traitement des données localement et le traitement ou l'échange de données avec des serveurs distants (grilles de calcul).

- Matériel:
  - ~ 500 serveurs.
  - 125 PiB sur bandes.
  - 70 PiB sur disques.
  - Forte hétérogénéité technologique.

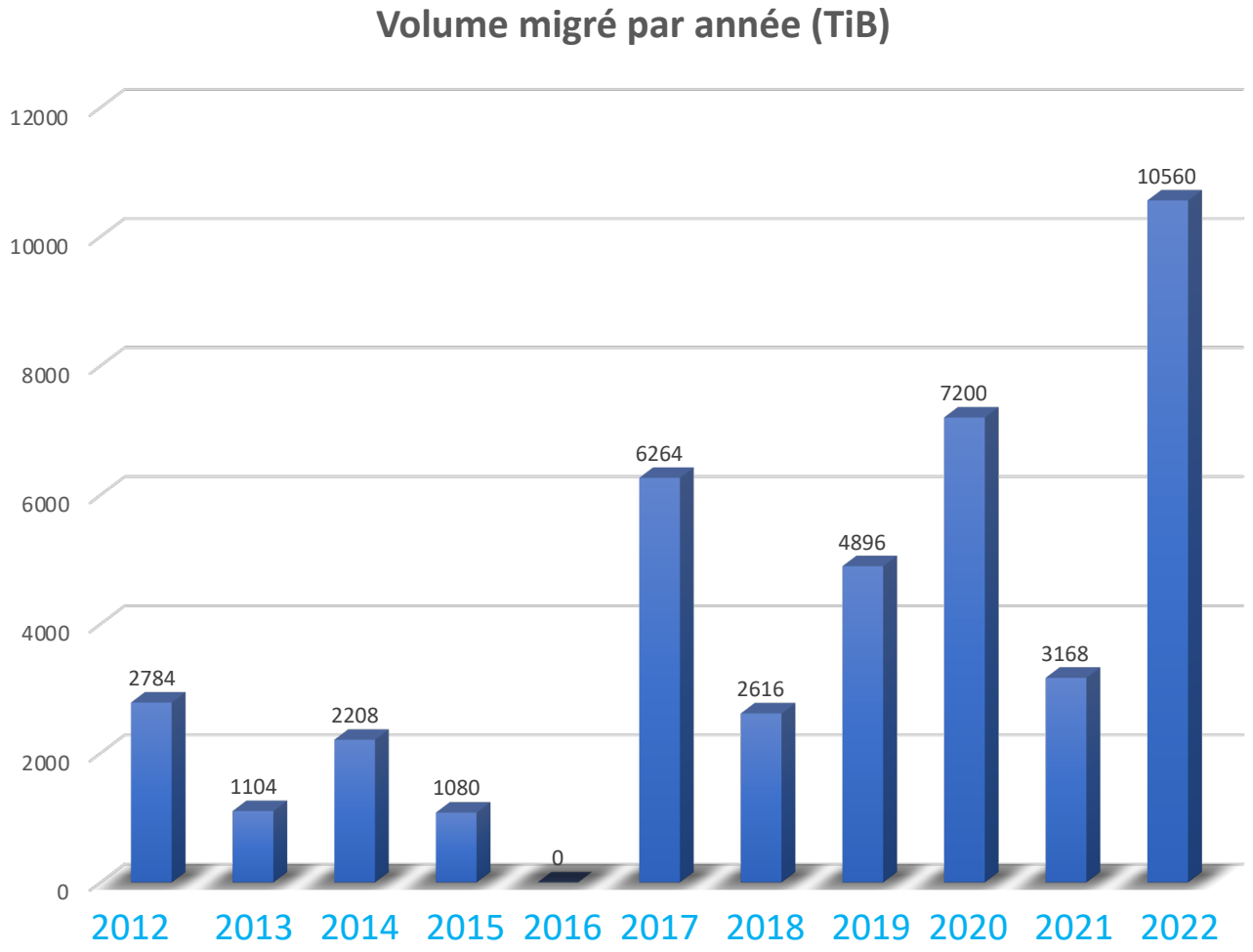
Le stockage sur disque

- Technologies, essentiellement déclinées en HDD, un peu en SSD :
  - DAS.
  - NAS.
  - SAN.
- Cible : systèmes de fichiers, serveurs de fichiers, bases de données.
- Cycle de vie : 5 ans pour le DAS, 7 voire (rarement) 8 ans pour le reste.

Concentrons nous sur les serveurs DAS !

- Migrations à effectuer en production :
  - Fréquence annuelle: entre 20 et 50 serveurs (4U chacun).
- Conditions requises :
  - Besoin des capacités réseaux et de la place en rack (+ alimentation) supplémentaires.
  - Systèmes de stockage doivent être capable de gérer les migrations physiques de façon transparente pour les utilisateurs.

Migration du stockage sur disque : exemple des DAS



Migration du stockage sur disque : écueils et questions

- Même « automatisé », une migration nécessite de la vigilance (perte de données potentielle).
  - Aucun problème rencontré dans le passé, mais le risque existe.
  - Risque augmenté pour migration aussi au niveau logicielle (ex: migration d'un système de fichiers à un nouveau).
  - Opérations réalisées en 1-2 mois.

- Migration des myriades de petits fichiers (< qques MiB) : lent.
  - Parallélisation des processus de migration nécessaire.

- Débat autour de la durée de conservation des serveurs :
  - Allonger la durée de vie des serveurs (de 5 à 7 ans) ?
  - Ou renouveler les serveurs au bout de 5 ans (coût au TiB décroit au fil du temps) ?
  - Tout dépend de l'évolution du coût de l'énergie.

Le stockage sur bande

- Stockage central au CC-IN2P3 :
  - Plus économique que le stockage sur disque.

- 3,7 PiB pour la partie sauvegarde gérée par Spectrum Protect (IBM).

- 122 PiB gérés par HPSS (IBM):
  - Interfacé avec les systèmes de stockage sur disque.
  - Utilisé pour le stockage froid, mais aussi le stockage de données moins actives à un instant t.
  - Stockage online : relecture de bandes à disque jusqu'à **150 TiB / jour**, moyenne de **2 PiB / mois**.
  - Challenge :
    - Demandes de fichiers non programmées à l'avance!
    - Demandes hétérogènes (profils d'accès différents).

Le stockage sur bande en 2022

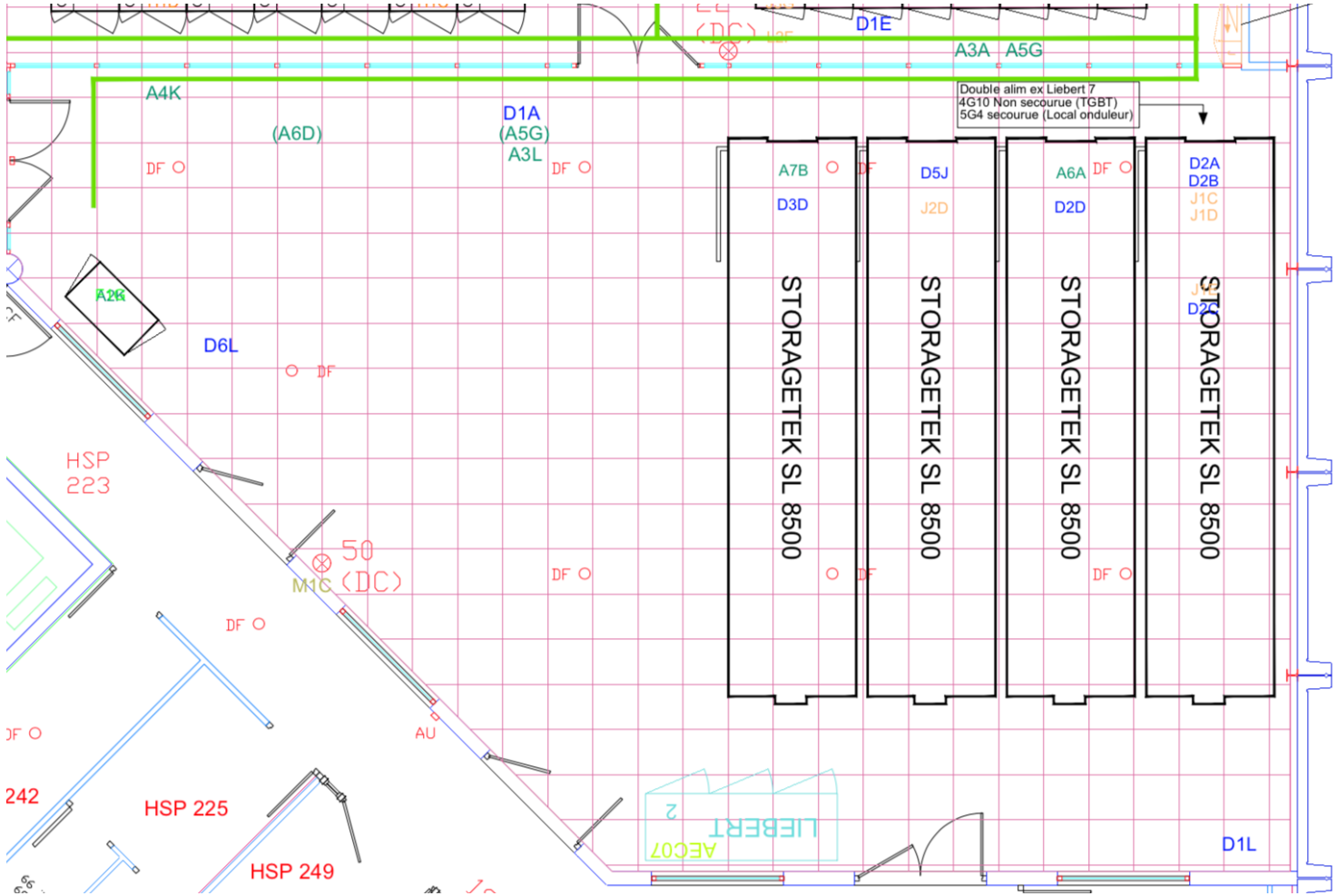
- 3 technologies de bandes :
  - LTO6 (2.4 TiB).
  - Jaguar JE (20 TiB) : IBM.
  - T10KD (8.5 TiB) : Oracle.

- 3 technologies de bandothèque :
  - TS3500 : IBM.
  - SL8500 : Oracle.
  - Tfinity : SpectraLogic.

- Pourquoi tout ça ?
  - Distingo LTO / bandes Enterprise : pour sauvegarde vs HPSS.

Situation pour les bandothèques jusqu'en 2019

- 4 SL8500 :
  - Localisation : 1ère salle.
  - HPSS + copie primaire pour le service de sauvegarde.
- TS3500 :
  - Localisation : 2nd salle.
  - Copie secondaire pour le service de sauvegarde.



Quelques temps plus tôt (partie I) ...

- 1er trimestre 2018 : annonce par Oracle de l'abandon de la techno « Enterprise ».
  - Pas de T10K de 3ème génération !!

- Réflexions:
  - Passer sur LTO ?
    - Conserver les librairies Oracle.
    - Capacité : 8,5 TiB (T10K) → 12 TiB (LTO8).
    - Performances LTO8 < T10K.
  - Passer sur les technologies Entreprise IBM ?
    - Capacité : 8,5 TiB -> 20 TiB.
    - Nécessite de remplacer les librairies.

- Décision:
  - Passer sur la techno « Enterprise » Jaguar / IBM ➔ chgt de librairies !

Quelques temps plus tôt (partie II) ...

- Appel d'offre publié en Septembre 2019.
- Petite librairie pour valider les choix technologiques.
- Capacité 20 PiB, 12 lecteurs Enterprise.
- Marché à bon de commandes :
  - Extension de capacité, lecteurs supplémentaires.

- 5 réponses / 2 solutions techniques:
  - TS4500 (IBM) ou Tfinity (SpectraLogic) avec lecteur TS1160.

- Choix :
  - Tfinity.

Objectif initial : migration de 80 PiB d'ici fin 2023.

Configuration initiale de la robotique au CC-IN2P3

- Installé 18-22 février 2020 par les équipes Spectralogic.
- Configuration
  - 5 frames (1 drive / 4 stockage).
    - ~ 3200 bandes
  - 2 teraporter (robot).
- 12 lecteurs IBM TS-1160
  - 400 Mo/s R/W.
- 990 bandes 3592-60F JE
  - 20 TiB / cartouche.
  - Terapack de 9 bandes.
- En production depuis fin mars 2020 (début des migrations, test grandeur nature).

Extensions

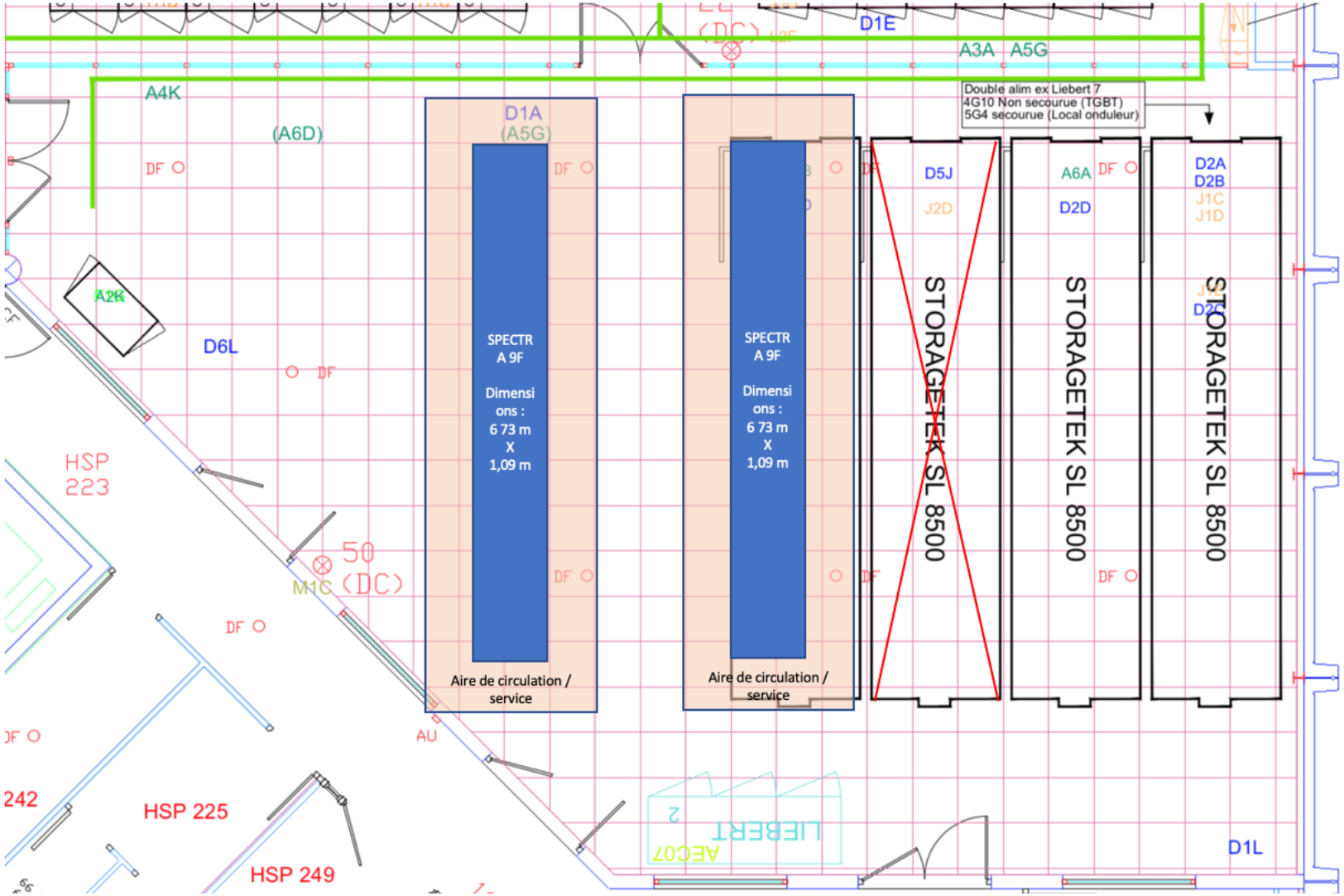
- Septembre 2020 :
  - + 12 lecteurs.
- Avril 2021 (migration et prod perturbée pour 3-4 jours):
  - + 4 armoires.
  - + 12 lecteurs.
- Septembre 2021 :
  - + 12 lecteurs.
- Total :
  - 6795 bandes.
  - 117 PiB utilisées.
  - 136 PiB disponible.

Eté 2022 – Installation de la nouvelle librairie

2021 : démantèlement de la 1ère SL8500.

Avril 2022 : démantèlement de la 2ème SL8500.

Juin 2022: installation de la deuxième bandothèque Spectra en lieu et place de la 1ère SL8500.



Vue d'ensemble

Pendant ce temps, les migrations…

- … continuent.

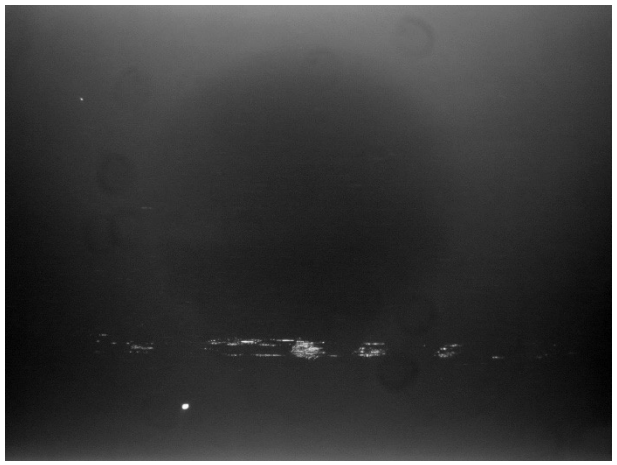
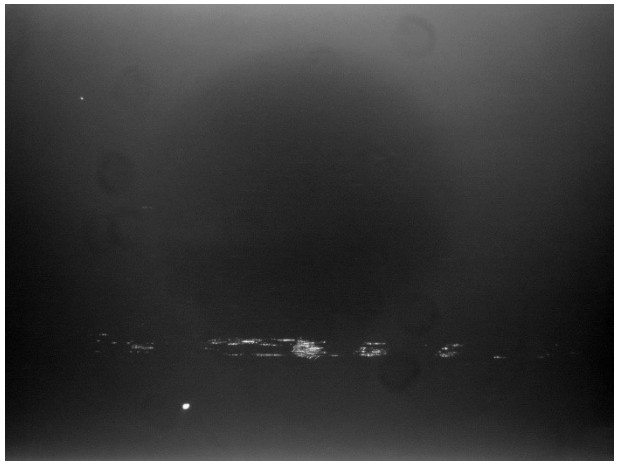
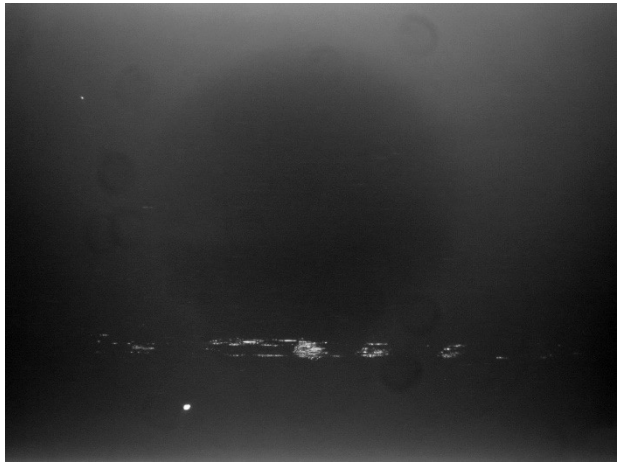
- 2 bandothèques Oracle mise hors prod.

- Avec les coûts de maintenance et désengagement de Oracle:
  - Accélération des migrations.
  - Objectif final: fin 2022 (sera atteint).

- Rythme de migration sur 30 mois :
  - 2,5 PiB par mois.
  - Équilibre à trouver avec la production:
    - pas de monopolisation des lecteurs (T10K et Jaguar).

Migration du stockage sur bande : écueils et questions

- Même « automatisé », une migration nécessite de la vigilance (perte de données potentielle).
  - Bandes coincés dans des lecteurs, bandes endommagées.
  - ~ 60 bandes écrites en 2016 avec quelques fichiers illisibles / bande : lecteurs défectueux.

- Migration des myriades de petits fichiers (< 64 MiB) : lent.

Migration du stockage sur bande : écueils et questions

- Durée d'utilisation des bandes :
  - 8 – 10 ans (de la mise en place jusqu'à l'arrêt complet).
  - Pourrait être plus long, mais course à l'échalote avec l'accroissement des besoins.

- Durée d'utilisation des bandothèques :
  - En pratique 15 - 20 ans, en réalité moins… (changements de politiques commerciales etc).

- Besoins d'espace physique pour effectuer la migration.

- Besoins de ressources physiques supplémentaires (lecteurs de bandes, serveurs, réseau …).

Conclusion

- Nbre de migrations de technologie matérielles et logicielles effectuées en 20 ans.
- Est et doit rester non chronophage pour les administrateurs de service.
- Risques technologiques liés aux migrations.
- Le volume à migrer pas nécessairement la chose la plus importante :
  - Petits volumes (qques TiB) mais des centaines de millions de fichiers.
  - Quantité de métadonnées à migrer critique.
  - Nature des données (expérimentales, simulations ou …. tout simplement du code).
- Migrations toujours techniquement possibles mais…
  - Écueils financiers, changements de stratégie commerciale subie.
  - Écueils matériels (espace physique disponible etc…).

- La clé :
  - Améliorer le cycle de vie des données.
  - Plan de gestion des données (vrai sujet bien avant que ça devienne la mode).

Moins on a de données à migrer, mieux on se porte !