Survol de WLCG/LCG-France
R&D en cours

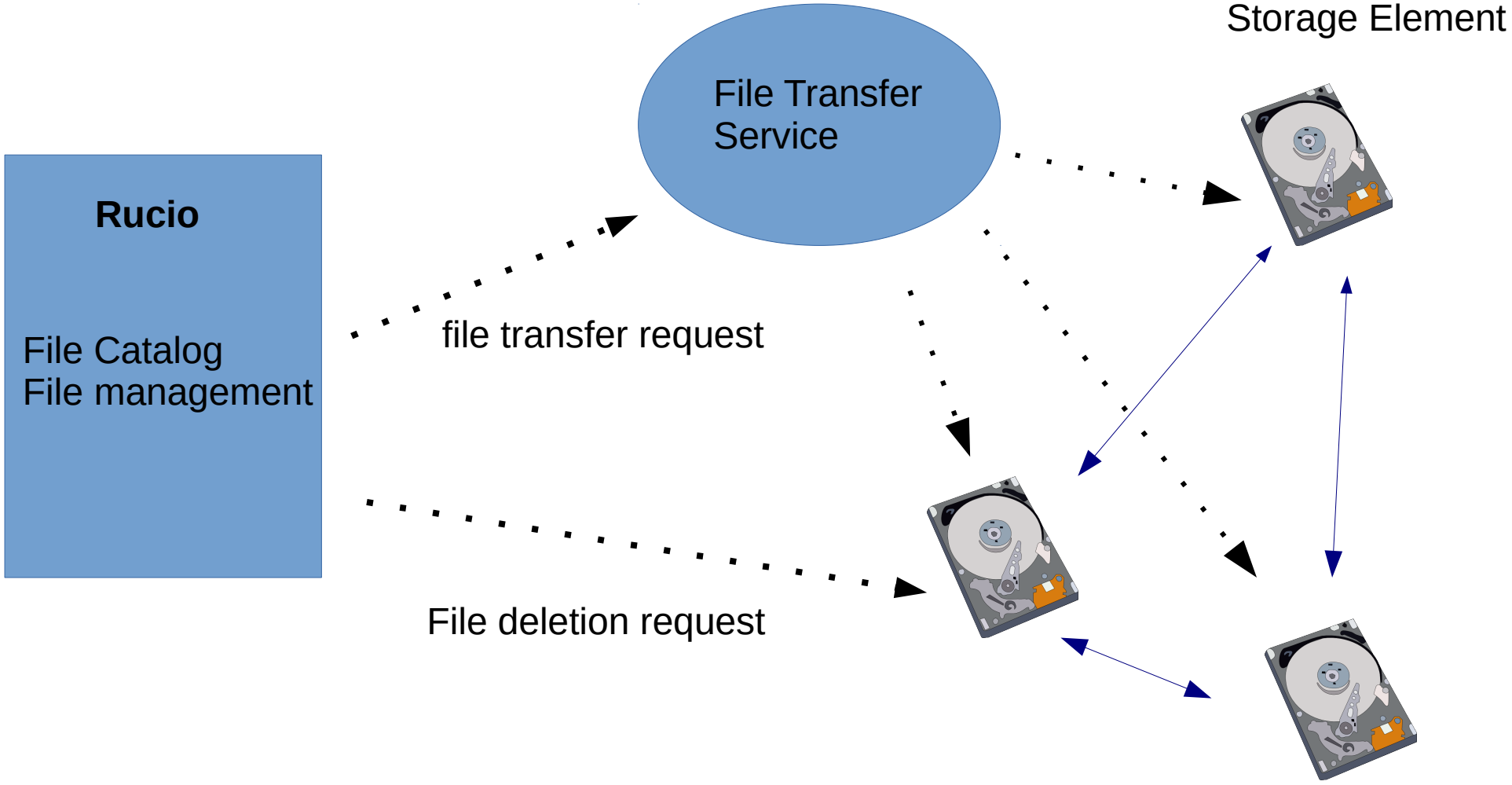
Intro

- Je propose un survol des outils et concept utilisés dans les expériences LHC (ATLAS pour les détails) et les activités actuelles de R&D
- Le but est d'introduire les outils et R&D au cas où certains vous intéresseraient
  - d'où une sélection de sujets potentiellement intéressants autour des données uniquement

Hiérarchie des sites

- Tier 0 : le CERN
- Tier 1 :
  - Centres nationaux, une dizaine par expérience (certains couvrent plusieurs expériences)
  - Collectent une copie des RAW data, font du reprocessing
  - Ont une bandothèque
  - Réseau dédié LHCOPN vers le CERN et 1/ 2 Tier 1 « ami » (vers le 100 Gb/s)
  - Le CC-IN2P3 pour la France et les 4 expériences LHC
- Tier 2 :
  - Centre régionaux, une douzaine en France
  - Souvent connectés par le réseau overlay LHCOne (10-100 Gb/s), tous les sites FR sont sur LHCOne
- Le modèle était initialement très hiérarchique dans la distribution des données, ce n'est plus le cas pour les données autre que RAW data

Data management



Rucio

- Catalogue de fichiers, datasets [quelques milliards]
- Orchestration de la distribution des fichiers (nombre de copie, où) aussi en interaction avec le système de gestion des calculs (par ex pour amener les données là où il y a du CPU dispo). Durée de vie des fichiers. Effacement des fichiers.
- Gestion des données basée sur la durée de vie et des règles de distribution (# copies, type de stockage, localisation etc...)
- Initialement une application utilisée par ATLAS, maintenant une app open source : site rucio
  - Sera utilisé par la 2eme grosse expérience du LHC
  - Retenue ou étudiée par une vingtaines projets y compris hors physique des particules (Aeneas, LSST, ...)

File Tranfer Service

- Orchestrations de transferts entre sites
  - multi-utilisateurs/experiences, optimisation globale des transferts (3 instances FTS pour tout LHC+qq autres)
- Les données ne transitent pas par le service qui délègue les droits lecture/écriture issus du demandeur
- Protocoles :
  - Gridftp
  - Xrootd (issue de la communauté HEP)
  - Http
- Travail en cours :
  - Au delà de x509 pour authentification/autorisation, par ex web tokens
  - Pouvoir se passer de gridftp (fin support globus)

Solutions de stockage

- Les centres Tier 1 ont des tous des systèmes de bandes et donc souvent des systèmes de stockage plus complexe intégrant la gestion des disques et bandes
- Les Tier 2 sont plus hétérogènes
  - DPM (solution développée pour la grille) : base de donnée, propose une vue « style filesystem », gestion de pools de disques (souvent « bêtes » serveurs de fichiers)
  - Solutions distributed filesystem (HDFS, GPFS,...) + surcouche « grille »
  - EOS : développé par le CERN, en particulier pour facilité l'utilisation de disque « consumer » (réplication et bientôt Erasure Coding)
  - RAL (UK) promeut une solution CEPH + serveurs d'accès

Paysage Français

- Un centre Tier 1 (CC-IN2P3) : environ 100 PB de données dont env 40 PB sur disque
- Une douzaine de centres Tier 2 (total de ~ 20 PB)
  - Typiquement 1-2 PB, le plus gros a environ 5 PB
  - La grande majorité utilise DPM
    - La solution devrait tenir jusqu'à des volumétries de ~100 PB
    - la France est membre de la collaboration DPM et discute de la pérennité de la solution en terme de besoins et disponibilité des développeurs / support
- Solutions techniques :
  - Brique de base pour le disque: serveur disque ~ 200TB/serveur, filesystem (marché MATINFO)
    - Coût total installé (avec rack, switch réseau, etc.) ~ 95 euros/TB
  - Bandes : le coût du média est faible mais le coût des drive est élevé (sans parler du coût des librairies)
    - Bande ~ 22 euros/TB
    - Drive ~ 13 keuros

R&D

- Pas mal de R&D lancé dans la communauté LHC du fait des défis posés par le run démarrant en 2026
- Volonté d'inclure les autres communautés dans les R & D (succès modéré)
  - Lien implicite avec ESCAPE (personnes en commun)
- Quelques sujets :
  - Accès aux données distantes, Data Management, Data Lake : DataOrganisationManagementAccess
  - Protocoles de transfert / techno d'authentification/autorisation
  - Qualité de Service du stockage
  - Technologies de stockage
  - Coûts consolidés

DOMA-FR : contacts initiaux avec CNES, INRIA, INSERM, ..

DOMA-FR : CEPH distribué

- Projets de laboratoires sur Paris-Saclay
- Construction d'une instance CEPH distribuée sur 3 labos (séparés de ~5km) pour redondance
  - Évaluation pour l'utilisation dans le cadre de la grille de calcul
- Paris-Saclay + Strasbourg : instance CEPH pour catalogue application cloud
-

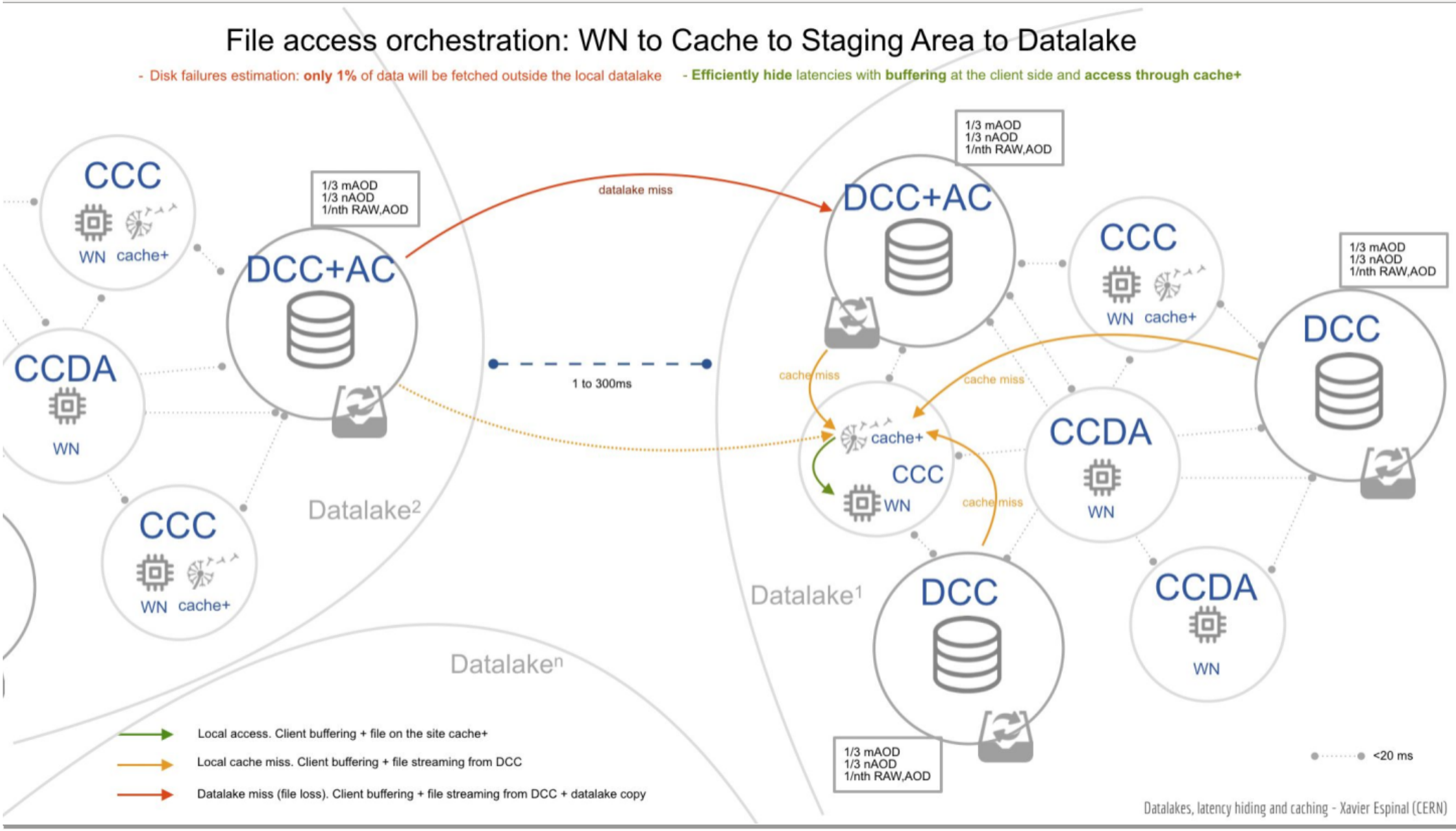
DOMA-FR : accès distant

- Un test-bed est en place pour un stockage unifié multi-site (Annecy – Grenoble – Lyon - Clermont-Ferrand - Marseille) associé aux queues de calcul dans tous ces sites
  - Comparaison de performance entre fichiers locaux vs fichiers distants
  - Futur : test de technologies de cache pour limiter l'impact

DOMA-FR : « data carousel »

- Investigation de l'utilisation intensive des bandes dans le workflow des expériences
  - Plusieurs « challenges » pour estimer le potentiel de « staging » de données « just-in-time » pour processing
    - e.g. dernier challenge : ~ 6 PB depuis une dizaine de Tier 1, quelle bande passante, durée pour 90 %, durée pour obtenir le dernier fichier
  - Intégration dans les workflows des expériences pour limiter le volume de données en permanence sur disque

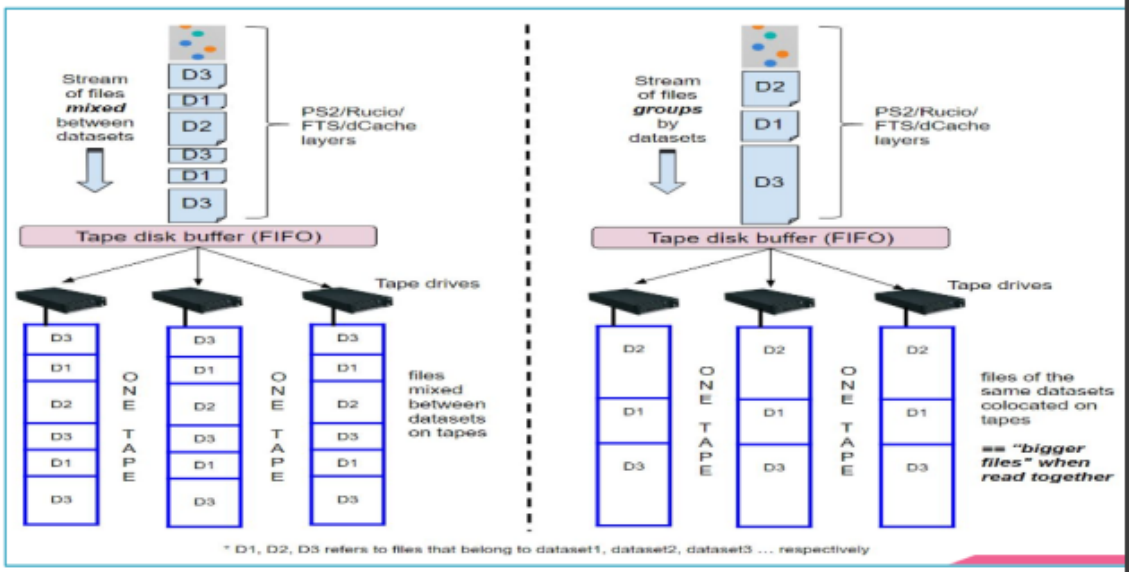
Backup


File access orchestration: WN to Cache to Staging Area to Datalake
- Disk failures estimation: only 1% of data will be fetched outside the local datalake
- Efficiently hide latencies with buffering at the client side and access through cache+
CCC
WN cache+
1/3 mAOD
1/3 nAOD
1/nth RAW,AOD
DCC+AC
datalake miss
1 to 300ms
CCDA
WN
CCC
WN cache+
Datalake2
Datalaken
cache+
CCC
WN
cache miss
cache miss
cache miss
DCC+AC
1/3 mAOD
1/3 nAOD
1/nth RAW,AOD
CCC
WN cache+
DCC
1/3 mAOD
1/3 nAOD
1/nth RAW,AOD
CCDA
WN
Datalake1
DCC
1/3 mAOD
1/3 nAOD
1/nth RAW,AOD
CCDA
WN
<20 ms
Local access. Client buffering + file on the site cache+
Local cache miss. Client buffering + file streaming from DCC
Datalake miss (file loss). Client buffering + file streaming from DCC + datalake copy
Datalakes, latency hiding and caching - Xavier Espinal (CERN)

Activités notables en France

- Data Carrousel
  - Tests et validation de nouvelles approches dans l'usage des solutions de stockage de masse (les bandes magnétiques).
  - Validation des performances intrinsèques des solutions hardwares et des sites.
  - Validation et optimisations de différents pattern de gestion des données.
  - Couvre un large spectre de compétences (hardware jusqu'à la pile applicative).

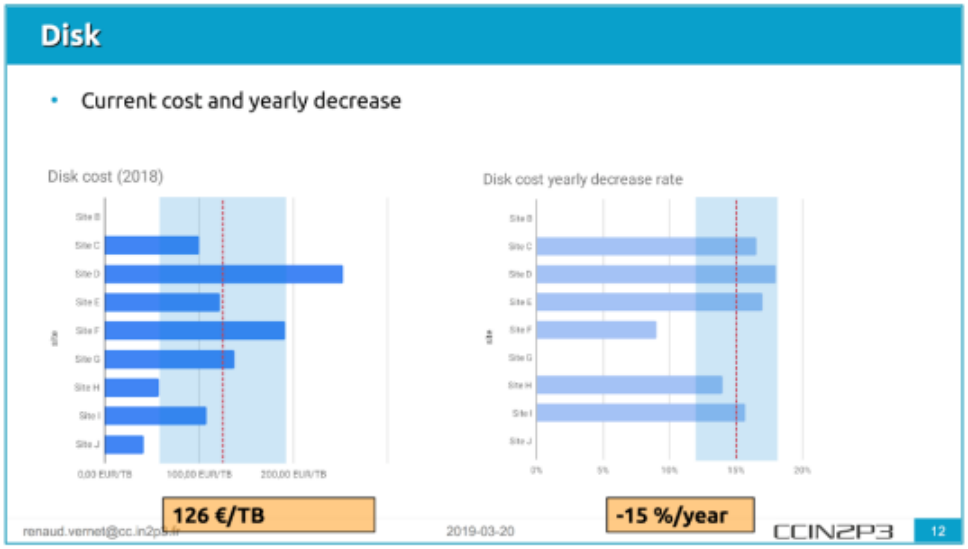
Graph from Xavier Espinal (DOMA)



Tape organization writing

Activités notables en France

- Modélisation des coûts
  - L'objectif de cette activité, plus large que la simple problématique du stockage, est de créer un modèle de coûts réaliste et global qui permette de faire des projections notamment pour anticiper le coût de tel ou tel modèle de stockage.



Coûts et décroissance des coûts disque sur quelques sites important de grille

Change	Effort Sites	Effort Users	Gain
managed storage on 15 sites + caches elsewhere	Some on large sites/gain on small sites	little	40% decrease in operations effort for storage
Caches at most sites (dataLake strawman)	Some everywhere	Frameworks some	15% of storage
Reduced data redundancy	Some large sites	Frameworks some	30-50% disk costs
Reduced data replication and cold data	little	Frameworks some	30% disk costs
Compact data formats for analysis	none	Some	>15% disk costs
Scheduling and site inefficiencies	Some	Some	10-20% gain CPU
Reduced job failure rates	Little	Some-Massive	5-10% CPU
Compiler and build improvements	None	Little-Some	15-20% CPU
Improved memory access/management	None	Realistic	10%-15% CPU
Exploiting modern CPU architectures	None	Massive	100% CPU
Paradigm shift algorithms	Some	Massive-Infinite	Factor 2-100 CPU
Paradigm shift online/offline data	Little	Massive-Infinite	2-10 CPU 10-20 Storage

Potentiels gains

Graph from « systems performance ans cost modeling working group