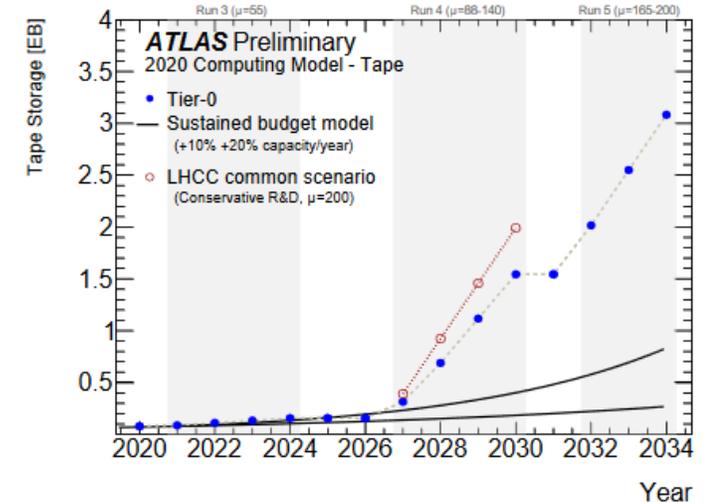
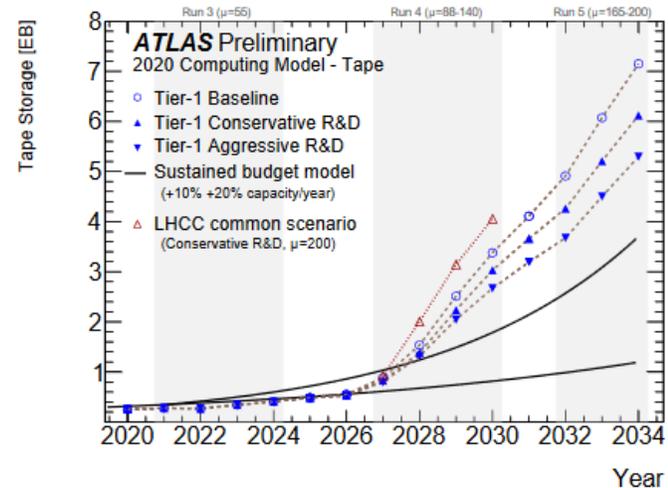
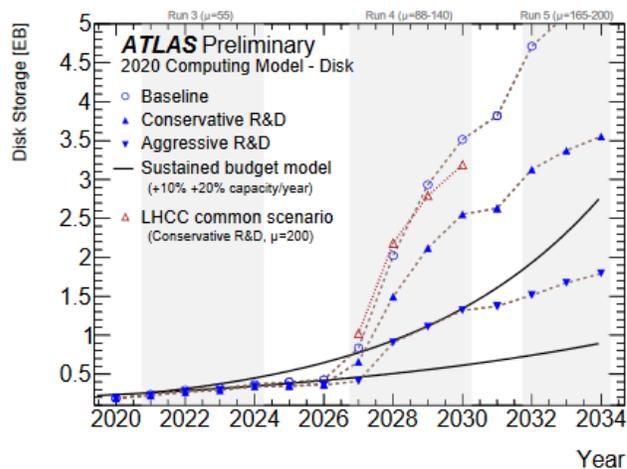


Centre de Calcul
de l'Institut National de Physique Nucléaire
et de Physique des Particules

DOMA : Data Organization Management Access

Aborder la problématique de la gestion de la données scientifiques à l'horizon 2028

- Gestion de la volumétrie.
- Gestion et usage de la donnée (data management)
- Gestion du coût.



From the [ATLAS Computing Conceptual Design Report](#)

L'évolution "naturelle" grâce à la technique et à la baisse des coûts sera insuffisante pour satisfaire les besoins de stockage à l'échéance du HL-LHC.

Il est nécessaire dès aujourd'hui de mettre sur la table des approches alternatives aux solutions de stockages qui sont proposées actuellement.

Le projet DOMA ne couvre que les aspects service de stockage et gestion/accès à la donnée.

DOMA ne se fixe pas comme objectif uniquement la problématique HL-LHC ni même ne se cantonne qu'à la physique des hautes énergies.

Mais aujourd'hui en pratique :

- Le projet DOMA est fortement focalisé sur les besoins du HL-LHC.

Un calendrier mouvant et très dépendant des avancées et approches envisagées.

- 2018-2021: Etude de concept et identifications des axes pertinents
- 2021-..... : Déploiement de démonstrateurs et quantification des performances

Des parties prenantes multiples : les expériences, les concepteurs de piles logicielles et les fournisseurs de services de stockage.

Dans la foulée de la création de DOMA, fin 2018 un projet IN2P3 porté par le DAS computing a été proposé : DOMA-FR

Les objectifs sont :

- Organiser les activités en France qui rentrent dans le giron de DOMA.
- Valoriser ces activités auprès de DOMA et des expériences/utilisateurs.
- Assurer le partage de connaissances entre les différents acteurs.

En 2021 7 laboratoires sont partenaires de DOMA-FR

- CC-IN2P3, LAL/IJCLAB, LLR, IPHC, LPSC, CPPM (2020), LPC (2020)
- Le LAPP bien que non partenaire du projet a une activité importante dans le projet.

Les laboratoires se positionnent sur les sujets pour lesquels ils ont un intérêt et bien sûr en fonction des moyens humains disponibles (chercheur et IT).

La période 2018-2021 était majoritairement orientée sur la compréhension/évaluation de concepts (datalake, cache « intelligent », protocoles innovants, QoS...)

2021-.... . Depuis cette année les activités DOMA se focalisent sur la validation des idées et approches.

L'objectif est de valider la possibilité d'utiliser les solutions de stockage de masse (Tape/Bande) de façon plus dynamique: Utiliser les tapes comme des disques.

Cette approche n'est envisageable que par un effort commun:

- Des fournisseurs de ressources notamment parce que les performances atteignables sont fortement corrélées aux choix technologique et de configuration.
- Au niveau des services de data management (FTS,Rucio,..). L'organisation des staging, des migrations/répliquions de fichiers dépendent de ces services qui se doivent d'être efficaces et robustes.
- Des expériences. La façon dont la donnée sera accédée et architecturée est un élément important dans la performance de cette approche (dataset schema, event access,....).

Seul un nombre limité de workflows (mais très consommateurs de stockage) sont susceptibles de tirer partie de cette approche.

Cette approche du stockage de masse doit être effectif dès le run3 (2022-...)

Plusieurs campagnes ont été menées avec les expériences du LHC.

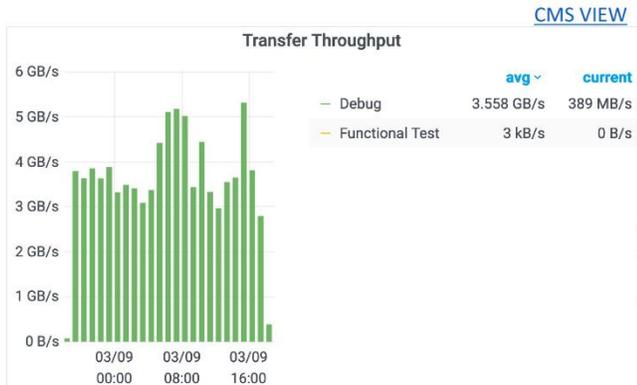
En tant que Tier1, le CC a été impliqué.

Les premiers retours ont permis notamment:

- Aux administrateurs de ressources d'identifier les composants qui sont susceptibles d'être primordiaux dans la recherche de la performance attendue. Cela est non trivial notamment car il y a de nombreuses contraintes sur les solutions de stockage de masse (partagées par diverses expériences, différents types d'usages, des aspects techniques/mécanique à prendre en compte,...).
- De mettre en évidence des limitations mais aussi des manques fonctionnels au niveau des services de data management.
- De confirmer l'importance des patterns de data access (phases d'écriture et lecture) dans l'usage qui peut être fait des données. DATASET schema, event access,....

L'approche DATA CAROUSEL est importante car potentiellement elle permettrait pour un nombre significatif de workflow de fournir un niveau de performance suffisant avec une ressource "bon marché".

Overview: ATLAS vs CMS staging throughput

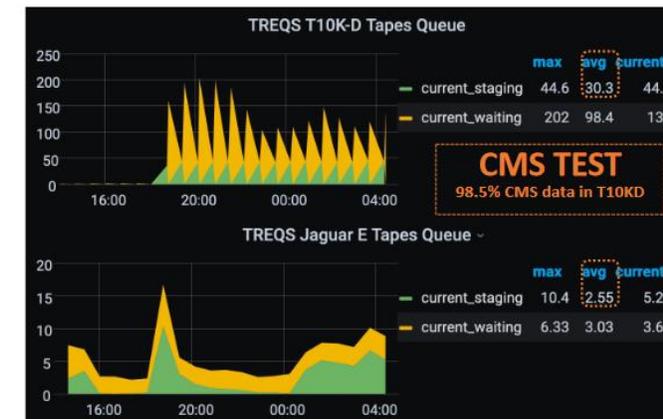
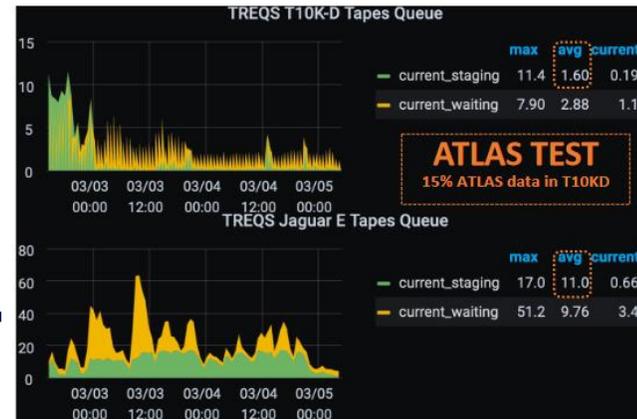


CC-IN2P3 VIEW

- ATLAS TEST: 1.7GB/s
- CMS TEST: 4.6GB/s

CC-IN2P3

Test Data Carousel Mars 2021



Campagne de test à venir la seconde semaine d'octobre (Challenges) :

Les 4 expériences du LHC sont impliquées et vont simultanément effectuer un challenge (avec de vrai data LHC) sur les infrastructures de stockage de masse des Tier1.

L'objectif est de proposer aux utilisateurs/expériences un point entrée (endpoint) unique vers non pas un site de stockage mais vers une entité qui serait composée de plusieurs types de stockage. Les expériences seraient consommatrices de stockage auprès d'une fédération de stockage vs un site. (C'est ce que l'on appelle aussi un datalake)

- **Avantages**

- Beaucoup d'aspects liés aux data management ne seraient plus du ressort des expériences mais de la fédération de stockage (migration des données, réplication, redondance, sécurisation de la donnée,...).
- Plusieurs types de stockage (et donc des coûts différents) peuvent constituer une fédération.
- Les coûts opérationnels (coté expériences et coté fournisseur de ressources) doivent baisser.
- Les notions de QoS (qualité de service) au niveau du stockage seraient disponibles.

- **Contre**

- N'existe pas encore, du moins en production.
- Et donc de nombreuses questions restent ouvertes.

Quelques unes des questions ouvertes:

- Quelle infrastructure réseau est nécessaire pour une fédération de stockage ?
- Quelle est la bonne taille pour un datalake ? Ou plutôt quelle est la taille optimale d'un datalake qui garantit un certain niveau de performance.
 - Echelle nationale? Echelle continentale?
- Quels éléments doivent être constitutifs de cette fédération.
 - Diskless site, service de cache, disk site, mass storage site, service de catalogue,...
- Une expérience (aka ATLAS/CMS) peut elle efficacement utiliser ce service de stockage ?
-

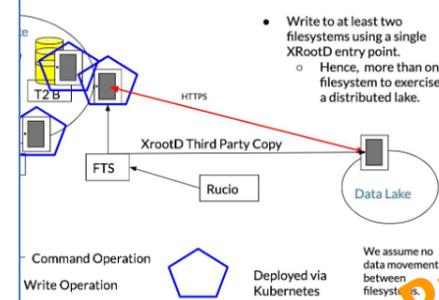
Au cours de 2020 plusieurs initiatives ont été menées pour “construire” de proto datalake et commencer à répondre à certaines de ces questions.

- ALPAMED est une de ces initiatives construites en France autour de plusieurs de nos sites.

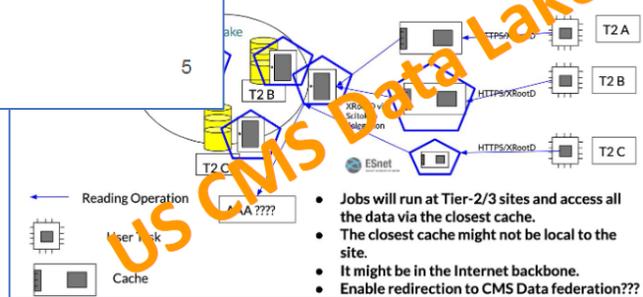
DATALAKE Proposal

- Many datalake components already included in ATLAS Grid (Rucio, FTS,...)
 - Adiabatic changes from existing infrastructure
 - Evolution to be endorsed within ATLAS
- Autumn 2020 : Steps towards datalake (US, EU) carefully monitored
 - Datalake include few sites
 - Existing monitoring to be consolidated (Network)
 - Measure performances :
 - benchmark jobs (local analysis, HammerCloud) : Existing at local level
 - VP within US
- 2021 :
 - Embark more Grid/local sites
 - Larger diversity of jobs
- Presentation in DOMAACCESS to be scheduled

Write to Data Lake



Read in lake for processing



Proposed Timeline for prototype Deployment

Task	By
All hardware for prototype in Kubernetes cluster	September 2020
Setup the XRootD origins and configure them with a the data lake single entry point.	October 2020
Configure caches to read from the data lake and use Scitokens for authentication.	November 2020
Setup a (R, E) in RUCIO (UST2DataLake) and have it register all NANO AOD to it.	December 2020
Setup submission infrastructure to be cache aware.	January 2021
Data lake testing, benchmarking and DevOps	January 2021 - September 2021

Benchmarking Goals

1. Exercise deletions and measure missed deletions as a function of:
 - a. Scale
 - b. Disconnecting an XRootD origin
2. Exercise data input and data removal via FTS
 - a. Scale
 - b. Success rates
3. Exercise NanoAOD application access.
 - a. Recruit students and postdocs with realistic applications
 - b. Cpu efficiency as a function of RTT to the closest cache.
 - c. Data access pattern (?to be thought about more carefully?)

<https://indico.cern.ch/event/953032/>

Quelques éléments ont été identifiés comme éléments fondamentaux dans la construction de ces fédérations de stockage.

- Les réseaux
 - Cette semaine a lieu un challenge network impliquant de nombreux sites avec pour objectif de valider le rampup nécessaire pour satisfaire les besoins HL-LHC.
- Les protocoles
 - Une volonté de converger vers un seul protocole de transfert (Webdav) en utilisant les fonctionnalités pertinentes (TPC,.....)
- Les approches de caching
 - Xcache
- Identification par tokens

T1	%ATLAS	%CMS	% Alice	% LHCb	ATLAS+CMS	Alice	LHCb	LHC Network Needs (Gbps)	LHC Network Needs (Gbps)
					Network Needs (Gbps)	Network Needs (Gbps)	Network Needs (Gbps)	Minimal Scenario in 2027	Flexible Scenario in 2027
CA-TRIUMF	10	0	0	0	196	0	0	196	393
DE-KIT	12	11	21	17	473	85	66	624	1247
ES-PIC	4	5	0	4	169	0	16	185	370
FR-CCIN2P3	13	10	14	15	448	56	58	562	1124
IT-INFN-CNAF	9	15	26	24	472	105	95	673	1345
KR-KISTI-GSDC	0	0	12	0	0	50	0	50	99
NDGF	6	0	8	0	111	31	0	142	285
NL-T1	7	0	3	8	143	12	34	188	376
NRC-KI-T1	3	0	13	5	51	51	21	123	247
UK-T1-RAL	15	9	3	27	472	10	109	591	1183
RU-JINR-T1	0	5	0	0	103	0	0	103	207
US-T1-BNL	23	0	0	0	453	0	0	453	906
US-FNAL-CMS (atlantic link)	0	45	0	0	909	0	0	909	1817
					1362	0	0	1362	2723
Sum	100	100	100	100	4000	400	400	4800	9600

Network bandwidth needs per T1 (or region)

T1	Data Challenge target 2027 (Gbps)	Data Challenge target 2025 (Gbps)	Data Challenge target 2023 (Gbps)	Data Challenge target 2021 (Gbps)
CA-TRIUMF	98	59	29	10
DE-KIT	312	187	94	31
ES-PIC	93	56	28	9
FR-CCIN2P3	281	169	84	28
IT-INFN-CNAF	336	202	101	34
KR-KISTI-GSDC	25	15	7	2
NDGF	71	43	21	7
NL-T1	94	56	28	9
NRC-KI-T1	62	37	19	6
UK-T1-RAL	296	177	89	30
RU-JINR-T1	52	31	15	5
US-T1-BNL	227	136	68	23
US-FNAL-CMS (atlantic link)	454	273	136	45
	681	408	204	68
Sum	2400	1440	720	240

Objectifs des network data challenges à 50 % des capacités réseau (scenario minimal)

Essentiellement techniques.

- Caractérisation de solutions de backend de stockage alternatifs.
 - Construire à l'échelle régionale une fédération de stockage basée sur CEPH.
 - SSD rapide
 - Cloud externes
- Monitoring réseau
 - Marquage des flux réseaux
- Généralisation du système d'adressage IPV6 dans les éléments réseaux et piles logicielles

Modèle de coûts (activité en standby).

La difficulté autour de la gestion et de l'accès à la donnée à l'échéance du HL-LHC est clairement identifiée .

Une pure approche technique ne permettra pas de résoudre le problème.

Un/des nouveaux services de stockage doivent être considérés et validés en prenant en compte tous les aspects :

- Techniques.
- Fonctionnels.
- Opérationnels.
- Usage (accès à la donnée).
- Coûts.

Beaucoup de sites Français (Tier1, Tier2) sont partie prenante dans DOMA et y contribuent (DOMA-FR coordination en France).

Participer à la R&D DOMA est important pour les sites mais l'implication des utilisateurs/expériences est absolument nécessaire.

Un impact fondamental dans la définition du HL-LHC (et autres) computing

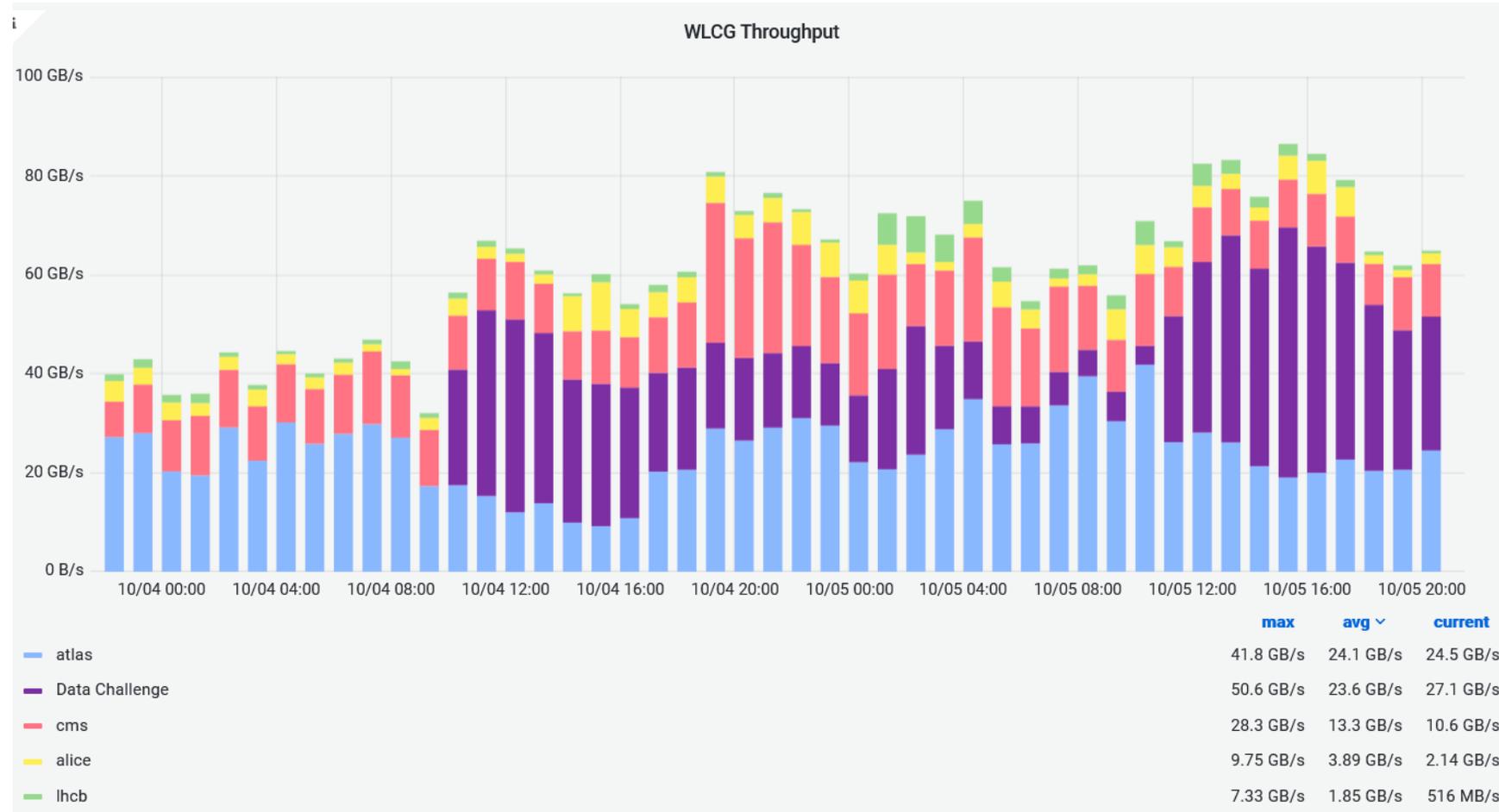
- HL-LHC Computing review
- HL-LHC TDR

HL-LHC n'est pas si loin.

- Les ressources que l'on achète en ce moment seront encore en production en 2027.

?

Backup : Network data challenge



Backup : Network data challenge Ihcone at CC IN2P3

