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

DOMA : Data Organization Management Access
DOMA-FR : composante française

Contexte

Un constat:

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 de mettre sur la table des approches alternatives aux solutions de stockages qui sont proposées actuellement.

DOMA : Aborder la problématique de la gestion de la donnée-scientifique à l'horizon 2028

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

Pour plus de détail sur le projet lui-même voir la présentation des dernières journées R&T :https://indico.ijclab.in2p3.fr/event/6256/

Les principales étapes du projet:

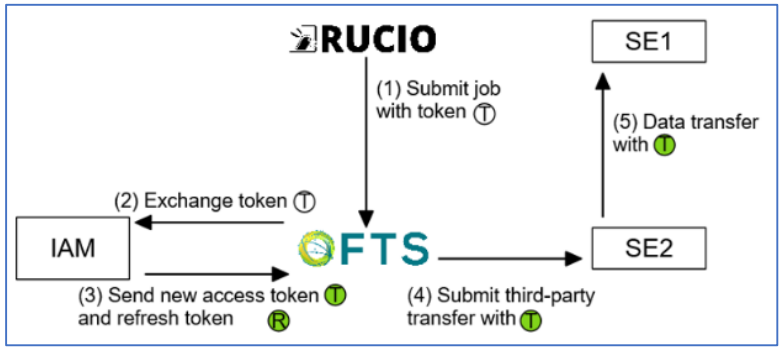
La période 2018-2021 était majoritairement orientée sur la compréhension/évaluation de concepts (datalake, cache « intelligent », protocoles innovants, QoS…)

Depuis 2021 les activités DOMA se focalisent sur la validation (et parfois leur mise en production) des idées et approches.

Feedback sur les activités 2021-2022

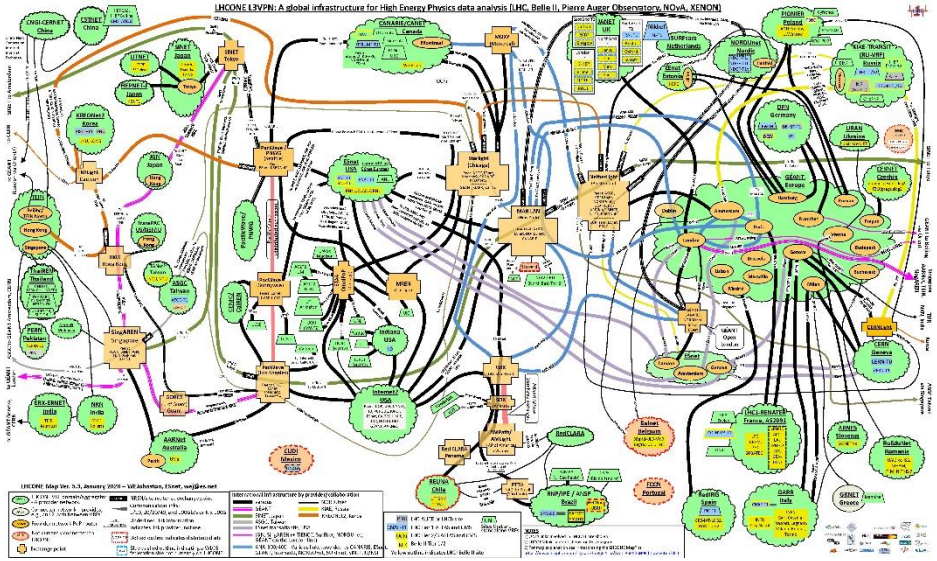
Sur cette période trois principales réalisations ont été faites/drivées dans le cadre du projet DOMA :

- **La validation et la mise en production de la fonctionnalité TPC (Third Party Copy) au niveau de protocole de transfert de fichiers.**
- **Des data challenges permettant de valider la montée en capacité (vers les besoins du HL-LHC) de transfert entre les sites impliqués au niveau international.**
  - **Implication des architectures réseaux et stockage mais aussi des différents services impliqués dans le data management des données**



Exemple de services impliqués dans le data transferts

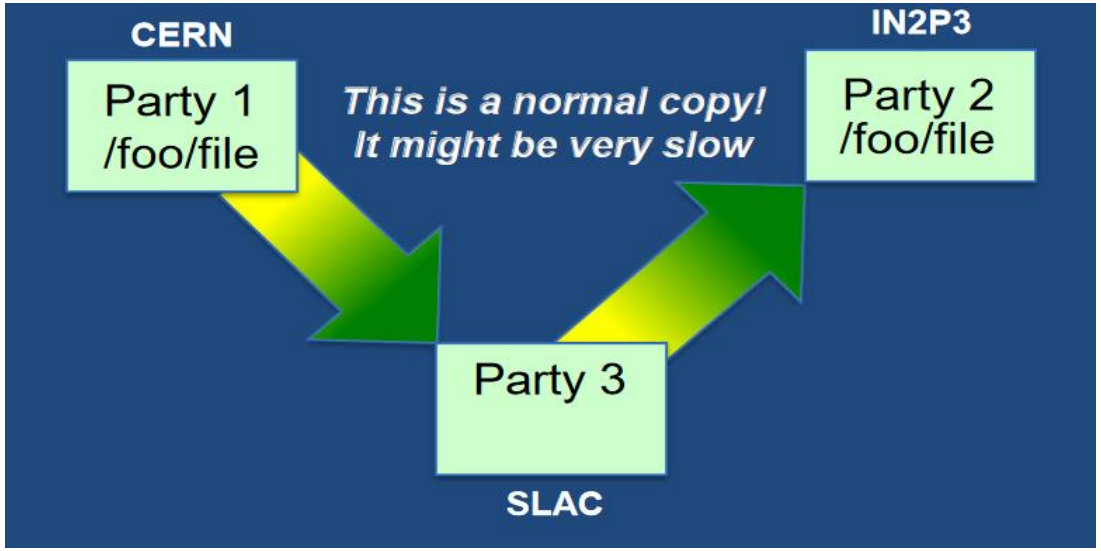
Carte LHCONE



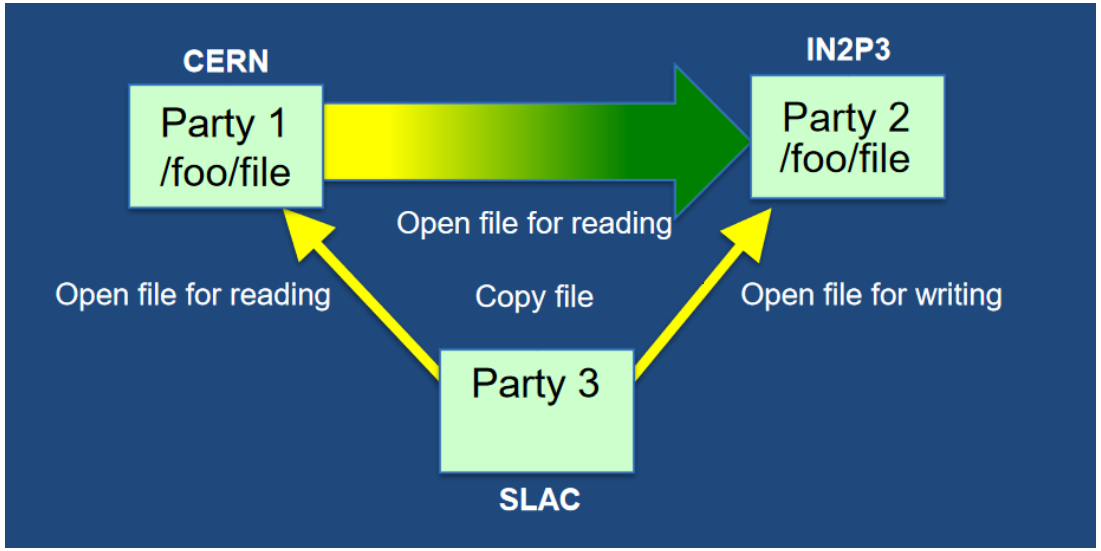
- **Optimisation et « tuning » des infrastructures de stockage sur bande afin d'aller vers une utilisation du stockage sur bande plus dynamique (data carrousel)**

Third Party Copy

Principe assez simple qui consiste à faire de la délégation de transfert entre de multiples partenaires.



Approche classique



Approche TPC

La principale difficulté consiste à gérer le mécanisme d'autorisation entre les partenaires.

- **Délégation d'autorisation**

Le second point notable est le fait que ce mécanisme fait partie intégrante du protocole de transfert

- **Lié au protocole**
- **Ne nécessite pas de service particulier supplémentaire**
- **Dans le cadre des transferts WLCG les protocoles webdav et xrootd savent faire du TPC.**

Third Party Copy

Courant 2019 les premiers tests de validation et de montée en charge ont été réalisés sur la plateforme de production de WLCG

TPC Functional tests

	AGLT2	Australia-ATLAS	BNL-ATLAS	CA-VICTORIA-WESTGRID-T2	CERN-PROD	FR-ALPES	FZK-LCG2	GRIF-IRFU	GRIF-LAL	IN2P3-CC	IN2P3-LAPP	IN2P3-LPC	IN2P3-LPSC	INFN-NAPOLI-ATLAS	INFN-T1	NDGF-T1	pic	praguelcg2	RO-07-NIPNE	RRC-KI-T1	SARA-MATRIX	NOIM	UNI-BONN
AGLT2_SCRATCHDISK	-	100%	98%	0%	0%	79%	100%	100%	100%	100%	96%	99%	100%	100%	100%	100%	100%	100%	100%	98%	100%	0%	10
AUSTRALIA-ATLAS_SCRATCHDISK	96%	-	93%	100%	92%	53%	93%	94%	96%	94%	92%	87%	99%	96%	93%	96%	100%	93%	95%	94%	96%	94%	96
BNL-OSG2_SCRATCHDISK	95%	99%	-	0%	0%	19%	95%	100%	99%	99%	21%	97%	97%	98%	100%	95%	100%	99%	17%	100%	99%	0%	95
CA-VICTORIA-WESTGRID-T2_SCRATCHDISK	0%	93%	0%	-	0%	23%	0%	100%	99%	0%	23%	92%	100%	100%	0%	0%	0%	98%	20%	0%	0%	0%	98
CERN-PROD_SCRATCHDISK	96%	62%	0%	0%	-	11%	96%	70%	94%	100%	14%	90%	94%	59%	0%	100%	94%	91%	13%	0%	96%	0%	82
FR-ALPES_SCRATCHDISK_DATALAKES	2%	2%	2%	2%	2%	-	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
FZK-LCG2_SCRATCHDISK	100%	100%	98%	0%	0%	100%	-	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	0%	100%	0%	10
GRIF-IRFU_SCRATCHDISK	98%	86%	70%	65%	62%	87%	100%	-	95%	100%	80%	97%	94%	75%	61%	100%	100%	97%	85%	55%	100%	60%	86
GRIF-LAL_SCRATCHDISK	100%	90%	88%	91%	86%	92%	100%	97%	-	100%	88%	98%	98%	91%	89%	100%	100%	98%	96%	90%	100%	83%	91
IN2P3-CC_SCRATCHDISK	100%	99%	99%	0%	0%	100%	100%	100%	98%	-	100%	100%	100%	100%	100%	100%	100%	100%	97%	100%	100%	0%	98

Les résultats ont été très concluants au point de mettre en production la fonction TPC dans l'ensemble de nos systèmes de stockage courant 2020.

- **Demande forte de la part des expériences**
- **Cela signifie aussi l'abandon du protocole historique gridftp**

Assez caché du point de vue des utilisateurs finaux, la fonction TPC a une impact important dans le transfert des fichiers des modèles de calcul (WLCG mais pas que).

- **Tous les mécanismes de data mangement , ordonnanceur de transferts, data catalogue,…. utilisent cette fonctionnalité.**
- **La TPC a également un impact important dans la mise en place de mécanismes de failover.**

Data challenges

Une série de challenges (un tous les deux ans) est planifiée jusqu'au début du run 4 pour valider plusieurs éléments

- **La montée en capacité des infrastructures réseaux**
  - **Sert aussi à avoir un calendrier des besoins pour interagir avec les prestataires de réseau (RENATER en France)**
- **La capacité des infrastructures de stockage à pouvoir « absorber » et « émettre » d'importants volumes de données.**
- **Valider la chaine fonctionnelle entre outils de data management (RUCIO), service en charge des transferts (FTS), services d'autorisation et authentification et même service en charge de monitorer tout ca.**
- **Valider le passage à l'échelle de fonctionnalités techniques ex: nouveau système d'authentification par token**

Data challenges

Le challenge réalisé en octobre 2021 était caractérisé par :

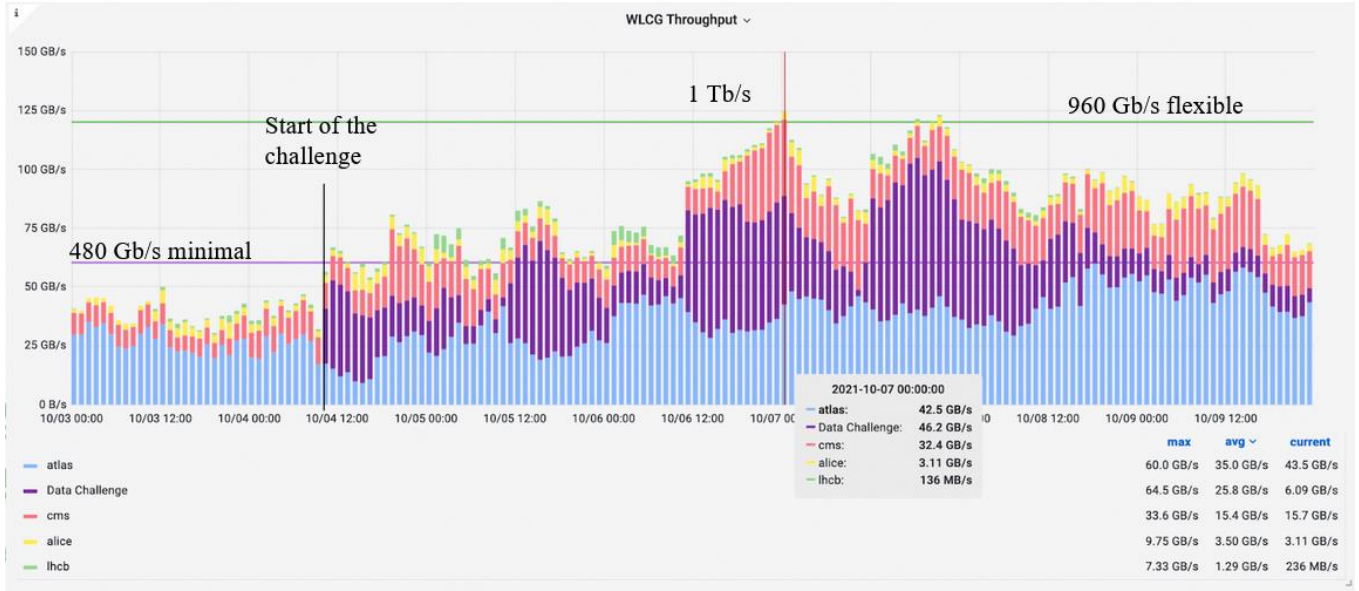
- Deux phases de une semaine chacune
  - La première pour faire un challenge purement réseau/stockage sur disque.
  - Une seconde semaine impliquant également la stockage sur bande magnétique.
- Les 4 expériences du LHC étaient impliquées.
- Un objectif : atteindre 10 % des objectifs 2027 dans le transfert de fichiers sur l'ensemble de l'infrastructure.
- Des challenges qui se déroulent sur l'infrastructure de production et qui se superposent aux activités de production des expériences WLCG.

T1	Minimal Scenario 2027	Flexible scenario 2027	Minimal scenario ingress/egress targets 2021
CA-TRIUMF	200	400	10/10
DE-KIT	600	1200	30/30
ES-PIC	200	400	10/10
FR-CCIN2P3	570	1140	30/30
IT-INFN-CNAF	690	1380	30/30
KR-KISTI-GSDC	50	100	0
NDGF	140	280	10/10
NL-T1 (NIKHEF)	-	-	10/10
NL-T1 (SARA)	180	360	10/10
RU-JINR-T1	200	400	10/10
RU-NRC-KI-T1	120	240	10/10
TW-ASGC	-	-	10/10
UK-T1-RAL	610	1220	30/30
US-FNAL-CMS	800	1600	40/40
US-T1-BNL	450	900	20/20
Atlantic link	1250	2500	60/60
Sum	4810	9620	240/240

Objectifs 2027 et déclinaison en objectifs 2021 pour les sites Tier1 (en Gb/s)

Data challenges

Résultats :



Trafic global durant DC 2021

T1	Minimal Scenario 2027	Flexible scenario 2027	Minimal scenario ingress/egress targets 2021	Ingress (hourly avg/max)	Egress (hourly avg/max)	comments
CA-TRIUMF	200	400	10/10	17/49	25/70	ok
DE-KIT	600	1200	30/30	33/77	52/143	ok
ES-PIC	200	400	10/10	11/18	11/17	ok
FR-CCIN2P3	570	1140	30/30	35/70	41/80	ok
IT-INFN-CNAF	690	1380	30/30	25/57	43/87	sum ok
KR-KISTI-GSDC	50	100	0	0	0	Alice T1
NDGF	140	280	10/10	26/49	27/82	ok
NL-T1 (NIKHEF)	-	-	10/10	10/37	12/53	ok
NL-T1 (SARA)	180	360	10/10	13/51	16/79	ok
RU-JINR-T1	200	400	10/10	11/26	12/31	ok
RU-NRC-KI-T1	120	240	10/10	9/18	12/34	sum ok
TW-ASGC	-	-	10/10	8/16	10/13	explain
UK-T1-RAL	610	1220	30/30	16/41	25/43	explain
US-FNAL-CMS	800	1600	40/40	16/49	19/49	explain
US-T1-BNL	450	900	20/20	29/75	38/117	ok
Atlantic link	1250	2500	60/60			
Sum	4810	9620	240/240	259 avg	343 avg	ok

Résultats DC 2021 déclinaison pour les sites Tier1 (en Gb/s)

Data Challenge + production

Attempted Transfers	Successful Transfers (%)	Failed Transfers	Failed Transfers (%)
19.8 Mil	83.75%	3.23 Mil	16.25%
Attempted Transfers (vol.)	**Successful Transfers (vol.)**	**Failed Transfers (vol.)**	**Average Throughput**
37.29 PB	32.12 PB	5.17 PB	73.7 GB/s

Vue macro des volumes transférés durant le DC network 2021

Optimisation de l’usage des bandes magnétiques (DATA Carrousel)

Objectifs :

« Utiliser les bandes magnétiques comme du stockage sur disque « , enfin presque.

Pourquoi ?

Une question de coûts, investissement mais aussi de fonctionnement.

Est-ce un pur problème technique ?

Non, la façon dont les données mises sur bande magnétique et seront utilisées a un impact important sur les usages et performances possibles.

Par conséquent c’est autant une approche technique (sites de stockage) que usage de la donnée (coté application/collaboration)

C’est quoi la difficulté ?

Satisfaire une nécessité d’écriture (raw data), un besoin de lecture (analyse des data), avec un nombre restreint de lecteurs et caches, des temps de latence importants, une redondance importante et tous cela pour une nombre important de collaboration.

Très régulièrement des exercices sont faits pour cross valider les usages (aka comment les collaborations travaillent) et les capacités techniques du stockage sur bande.

DATA Carrousel : exercice 2022

12 Tier1 qui font du stockage sur bande c'est 12 infrastructures différentes avec des contraintes différentes

- Différences de choix technologiques et de dimensionnement.
- Différences dans la granularité/configuration/partitionnement de la solution de stockage
- Différences dans le nombre et type de collaborations à supporter
- Différences dans les approches de caching/redondance

Par conséquence rien de comparable ou opposable, mais des enseignement à tirer du site voisin et des tunings/config qui évoluent de challenge en challenge.

L'intérêt de faire ce type de challenges en impliquant plusieurs collaborations en même temps, c'est introduire le problème de la concurrence d'accès à la ressource.

Voila quelques retours au niveau du CC

- L'exercice 2022 était déjà le troisième exercice depuis 2021 et donc un certains nombres d'optimisations avaient été implémentées, mais on a encore identifié pas mal de points d'optimisation à réaliser.

DATA Carrousel : exercice 2022 au CC

Différentiation entre les phases d'enregistrement des raw data (DT: data taking) et les phases de lecture (A-DT)

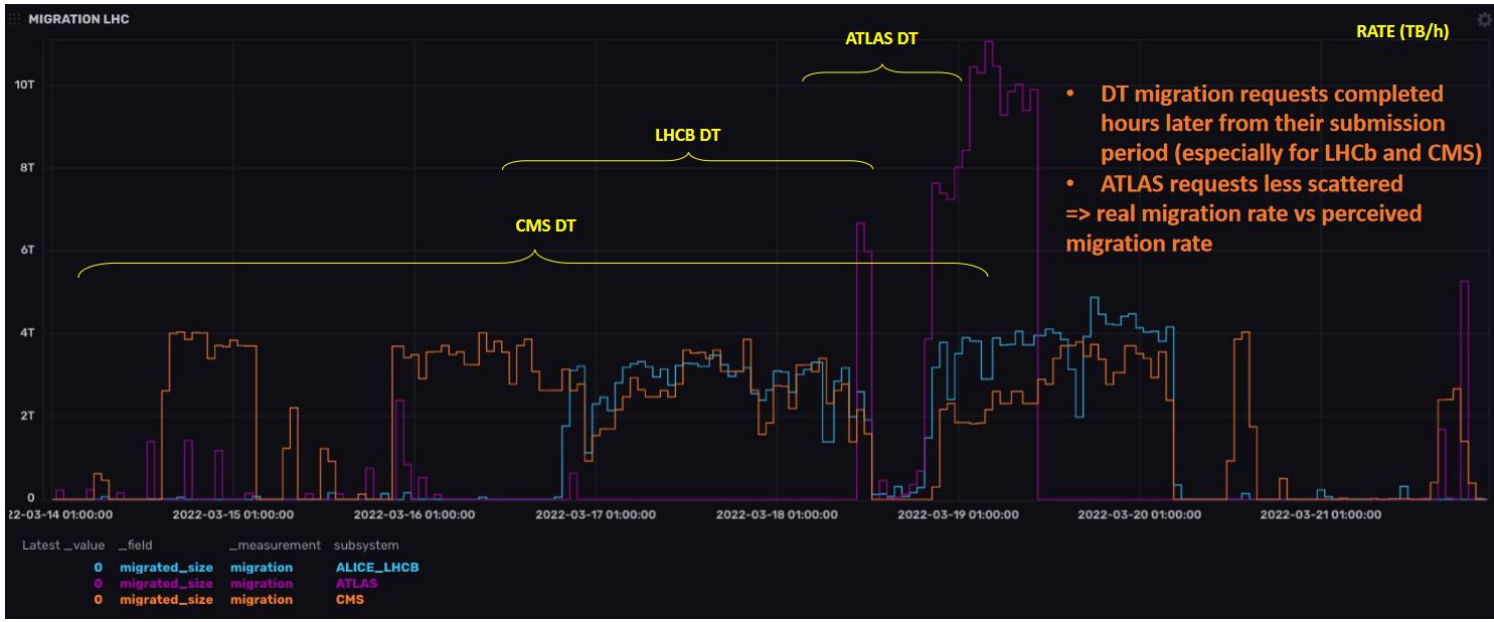
	DT (migrations)			OK?	A-DT (stagings)			OK?
	MAX	AVG	TARGET		MAX	AVG	TARGET	
ATLAS	4GB/s	1.8GB/s	3.5GB/s	✓	6.9GB/s	2.6GB/s	1.2GB/s	✓
CMS	7GB/s	0.9GB/s	0.29GB/s	✓	8.3GB/s	3.4GB/s	?	?
LHCB	4GB/s	1.4GB/s	1.2GB/s	✓	4.5GB/s	2GB/s	0.98GB/s	✓

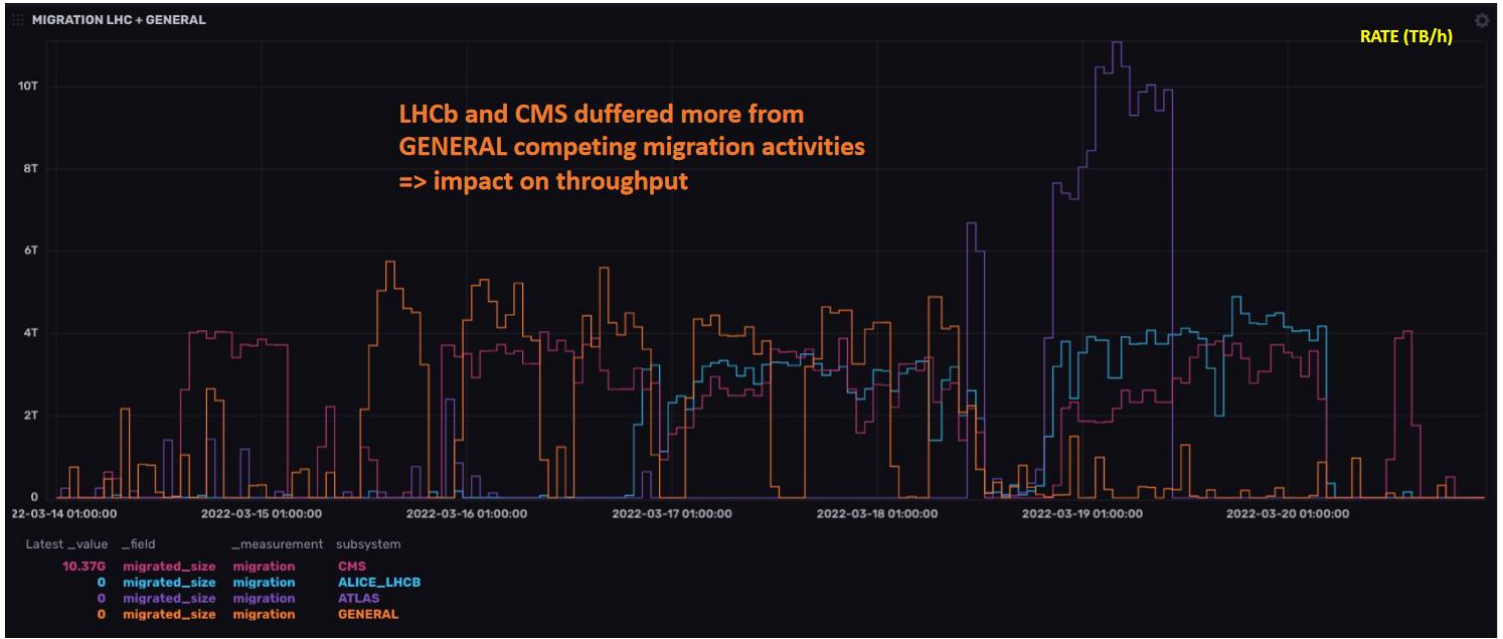
Débit atteints au CC

DATA Carrousel : exercice 2022 au CC

Compétition entre les expériences ATLAS,CMS,LHCB durant la phase de DT

Vues débit



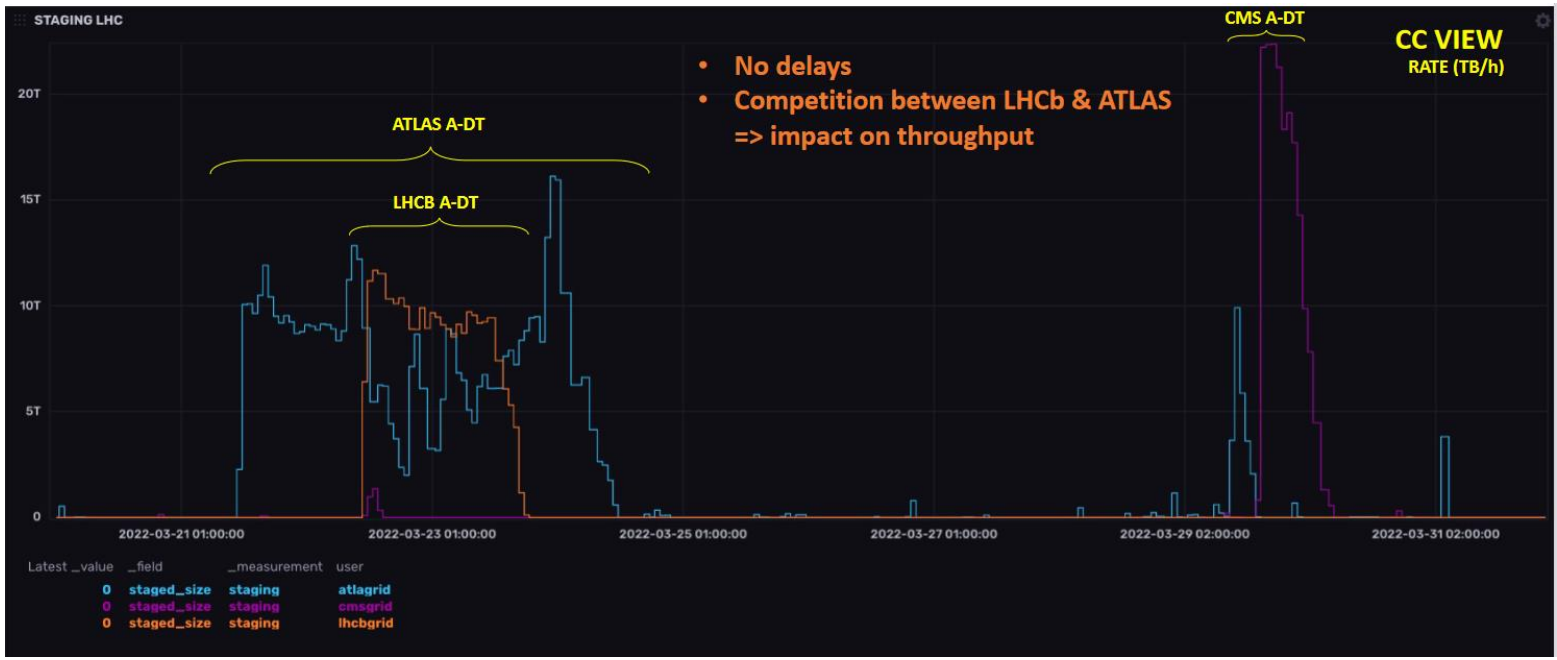


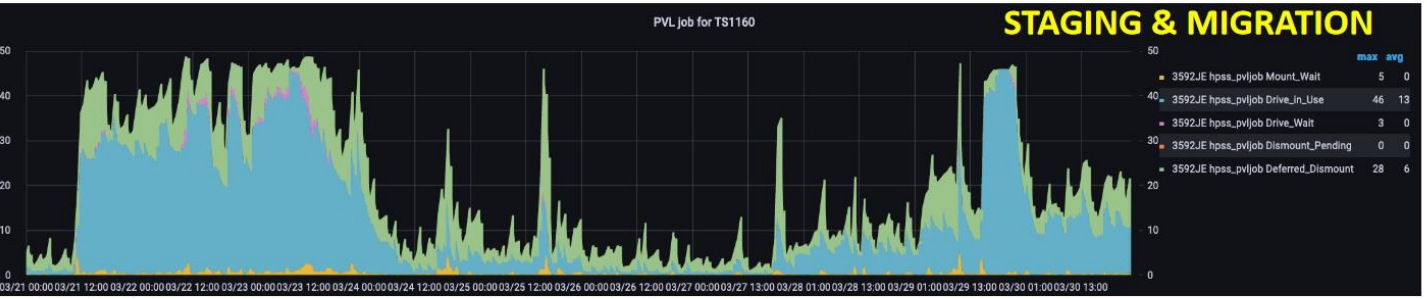
DATA Carrousel : exercice 2022 au CC

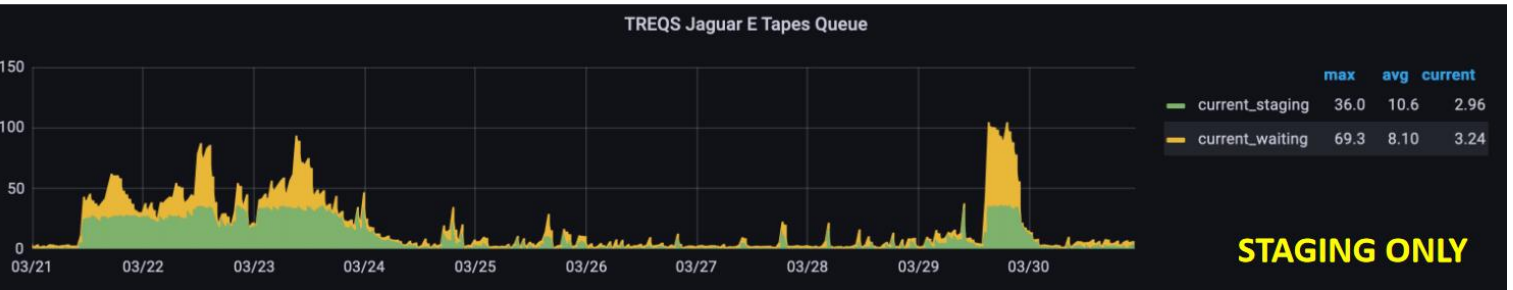
Compétition entre les expériences ATLAS,CMS,LHCB durant la phase de A-DT

- Vue débit

- Vue lecteurs







Ce type d'exercice permet

- De tuner et modifier notre infrastructure
- De coordonner avec les expériences certains dimensionnements

Conclusion

La phase initiale (2018-2021) du projet DOMA était très orientée R & T et foisonnant d'idées et d'approches pour satisfaire les besoins en gestion de la données à l'horizon du run 4

- La France a participé à cette phase avec une implication significative notamment dans le concept de datalake (aka fédération de stockage)

Depuis 2021 le projet DOMA est devenu technique et orienté implémentation et validation de solutions

- Aujourd'hui c'est surtout autour des sites Tier1 , donc CC, que se focalisent les activités et actions du projet.
- Des aspects techniques restent encore à valider.
  - Authentification
  - Protocole de staging des bandes
  - Marquage réseau
  - ….

Conclusion

DOMA a 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.

DOMA quitte (mais il reste encore des sujets intéressants) sa phase de R&T pour rentrer dans sa phase implémentation et mise en production

- Le run3 et la nécessité d'aller de l'avant y est pour beaucoup

?

Backup

Tape challenge: CC setup

CC-IN2P3 setup

- Tape/drive resources are shared by all VOs (LHC and non-LHC)
- HPSS T10K-D Media migration (repacks) suspended during the TC
- HPSS Staging Configuration (based on TREQS staging scheduler):
  - Jaguar-E/TS1160 (Drive nominal speed 450MB/s)
    - **46 drives available for staging and migration:** staging scheduler (TREQS) requests max 32 drives at each staging pass (so max 14 drives are left for migrations)
  - T10KD (Drive nominal speed 240MB/s)
    - **48 drives only for staging**
  - Pending time for staging requests set up to minimum 4min (but there is no max and the staging file can be served after hours)
  - Migration cycle is every 6h (it applies to files written > 2h ago)
  - File size class setup (more relevant than file family): it determines the number of drives used on migration. For LHC VOs:
    - COS 12 (64MB - 2GB): 5 drives
    - COS 14 (> 2GB): 6 drives