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

DOMA : Data Organization Management Access

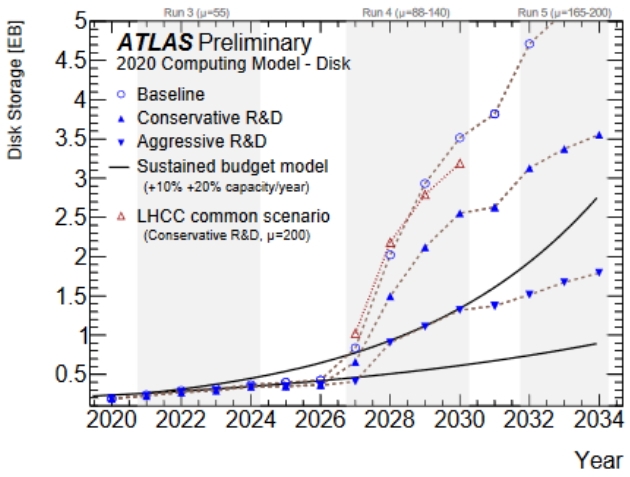
Plan

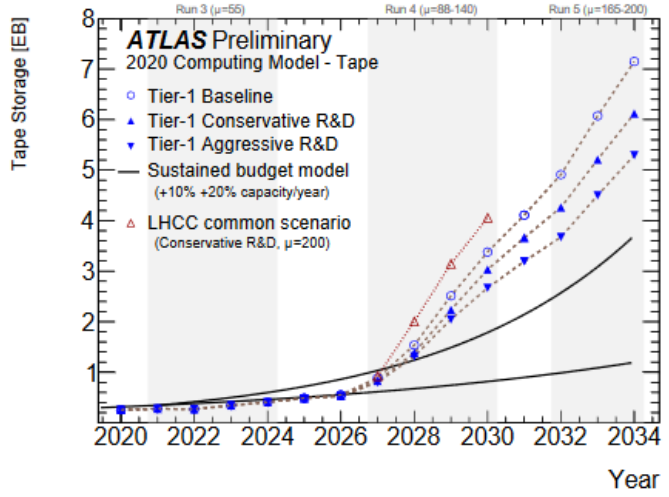
- Un constat.
- Projet DOMA.
  - International.
  - France.
- Activités DOMA.
  - Autour des fédérations de stockage.
  - Autour de l'optimisation d'usage des stockages de masse.
  - Autres.
- Quelques résultats.
- Conclusions et remarques.

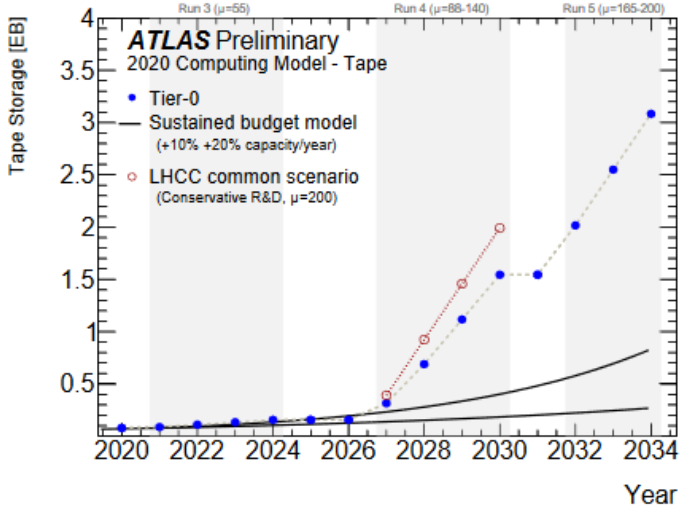
Un constat

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.







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.

Projet DOMA : International

- Suite à cette prise de conscience en Juin 2018 il a été proposé de formaliser les R&D nécessaires à travers un projet : DOMA Data Organization Management Access.
  - kickoff meeting

- Les objectifs initiaux étaient :
  - *keep track of developments and advancements in all DOMA areas (data access, data management, data service).*
  - *provide a forum to discuss ideas.*
  - *foster interoperability of solutions.*
  - *an umbrella for stakeholders, national initiatives, EU projects, already existing working groups.*

- En pratique les activités DOMA cherchent à définir/tester/quantifier des solutions de stockage qui permettraient de satisfaire :
  - Une forte augmentation des capacités de stockage.
  - Une efficacité d'accès aux données compatible avec les besoins expérimentaux.
  - Et cela pour un budget plat/maitrisé, incluant à la fois les aspects techniques et opérationnels.

Projet DOMA : International

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 identification 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.

Projet DOMA : France

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 8 laboratoires participent à DOMA-FR

- CC-IN2P3, LAL/IJCLAB, LLR, IPHC,LAPP, LPSC, CPPM (2020), LPC (2020)

Des contacts (non concrétisés) ont eu lieu avec des structures hors HEP , tel que CNES, Université, INSERM

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

Activité : Fédération de stockage

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 au 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.
  - Des 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.

Activité : Fédération de stockage

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 ?
- ……

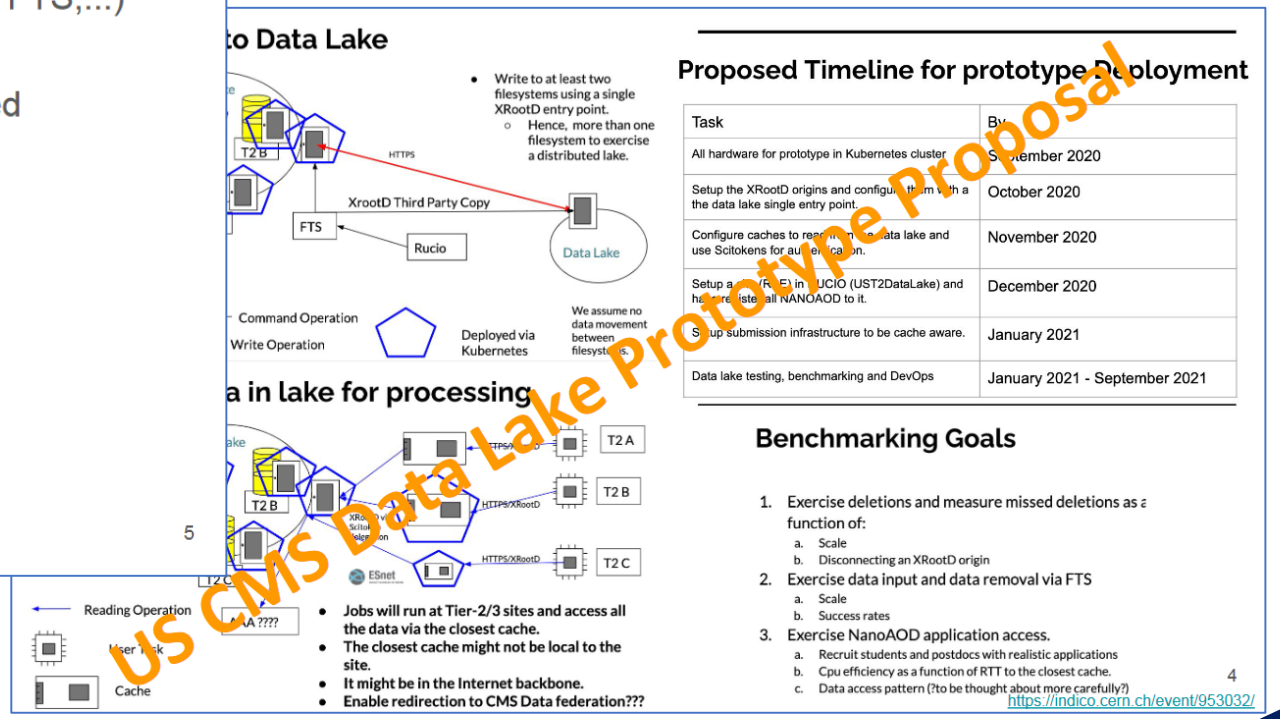
Au cours de 2020 plusieurs initiatives ont été menées pour "construire" de proto datalake et commencer à répondre à certaines de ces questions.

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

Bien sûr des datalake à des échelles plus grandes sont nécessaires/envisagés.

- 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 DOMA ACCESS to be scheduled

ATLAS Data Lake Prototype Proposal



US CMS Data Lake Prototype Proposal

Activité : DATA CAROUSEL

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/réplications 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 effective dès le run3 (2022-…)

Activité : DATA CAROUSEL

Plusieurs campagnes ont été menées en 2020 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".

Activité : Autres

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).

Caractérisation des notions de QoS (Quality of Services).

Quelques résultats

En octobre 2021 deux challenges (2 fois une semaine) ont été réalisés à l'échelle de WLCG pour caractériser les capacités de gestion des stockages.

Ces challenges ont été réalisés sur l'infrastructure de production de WLCG : site, réseau, outils de gestion.

Ces challenges s'étaient fixés des objectifs afin de valider la montée en puissance avec pour objectifs le HL-LHC.

Un des intérêts de ces challenges était de mettre en œuvre des activités des 4 expériences du LHC simultanément et de se superposer aux activités quotidiennes de production.

Quelques résultats : Data challenges

Le but de ces challenges est de valider les capacités de transfert de fichiers (file to file).

- Sont notamment sollicités par cet exercice.
  - Les capacités réseaux.
  - Les protocoles.
  - Les solutions de stockage.
  - Les services de data management (RUCIO,FTS,…)

- Des transferts de fichiers supplémentaires ont été injectés en plus de l'activité habituelle.

T1	%ATLAS	%CMS	% Alice	% LHCb	ATLAS+CMS Network Needs (Gbps) Minimal Scenario in 2027	Alice Network Needs (Gbps) Minimal Scenario in 2027	LHCb Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) 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	0	45	0	0	909	0	0	909	1817
(atlantic link)					1362	0	0	1362	2723
Sum	100	100	100	100	4000	400	400	4800	9600

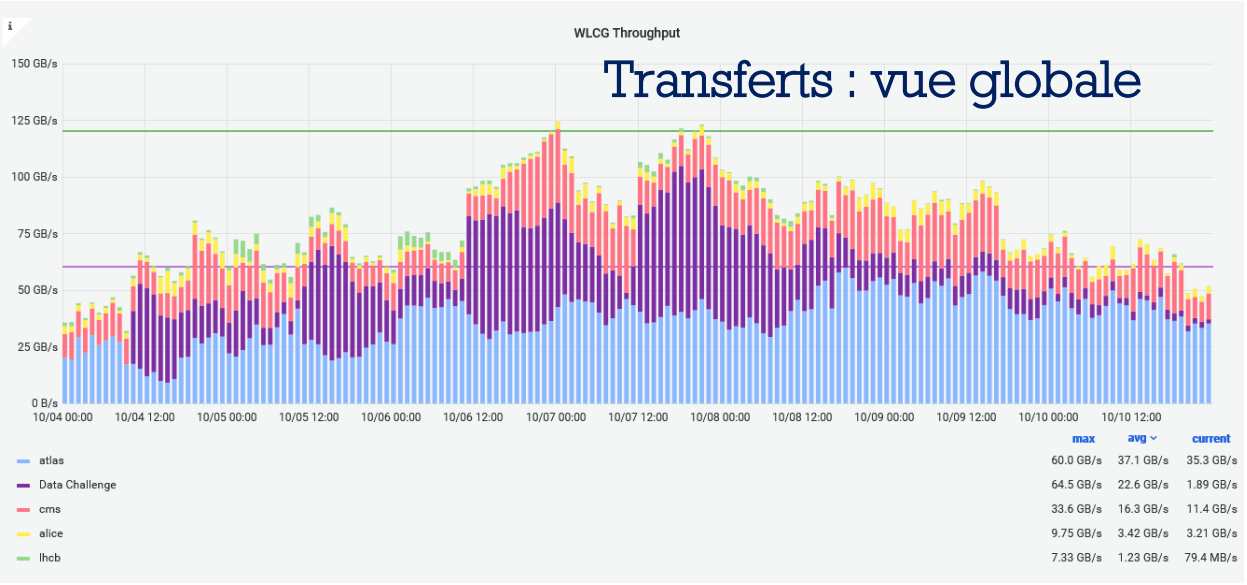
Network bandwidth needs per T1 (or region)

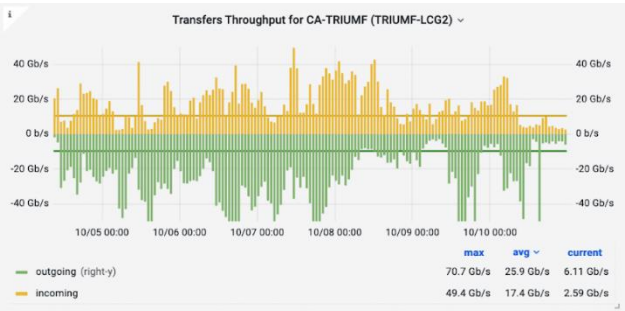
Objectifs des network data challenges à 50 % des capacités réseau (scenario minimal)

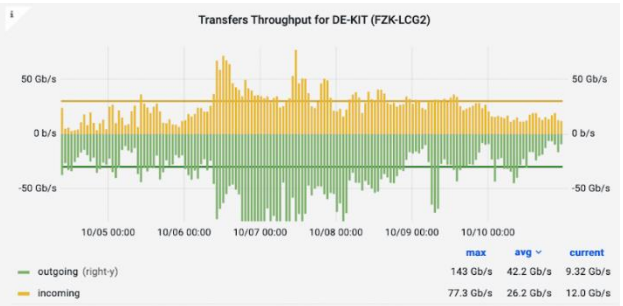
T1	Data Challenge target 2027 (Gbps)	Data Challenge target 2025 (Gbps)	Data Challenge target 2023 (Gbps)	Data Challenge target 2021 (Gbps)
CA-TRIUMF	98	59	29	10
DE-KIT	312	187	94	31
ES-PIC	93	56	28	9
FR-CCIN2P3	281	169	84	28
IT-INFN-CNAF	336	202	101	34
KR-KISTI-GSDC	25	15	7	2
NDGF	71	43	21	7
NL-T1	94	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	454	273	136	45
(atlantic link)	681	408	204	68
Sum	2400	1440	720	240

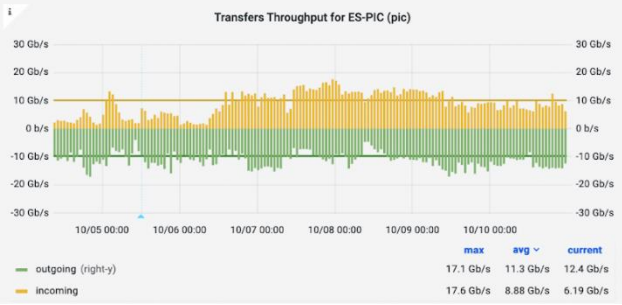
Quelques résultats : Data challenges

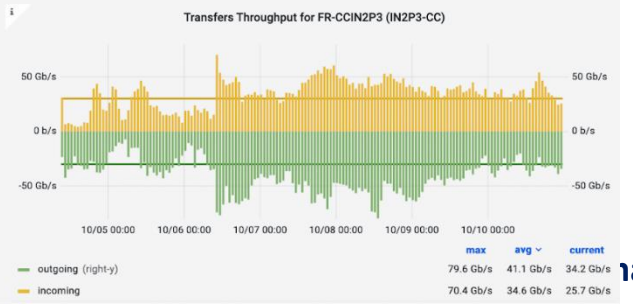
Transferts : vue globale





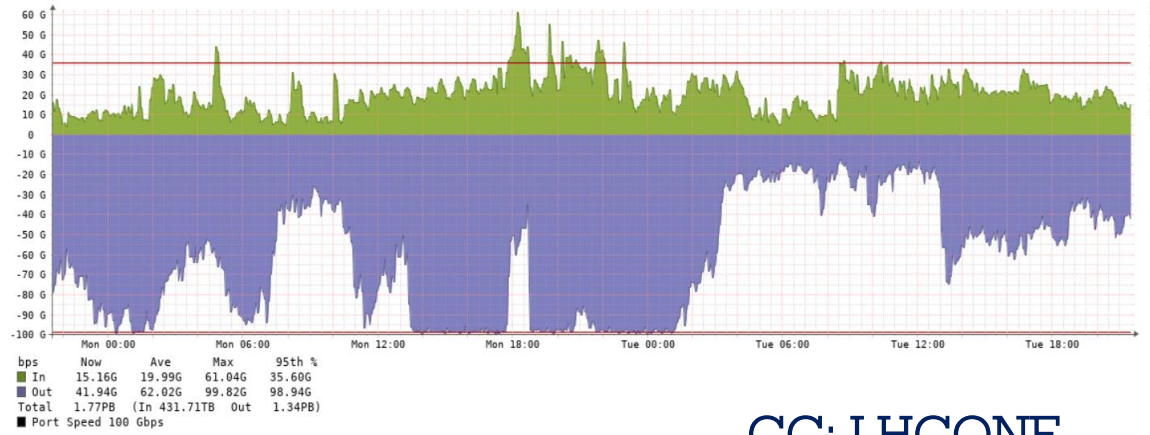






Resultats

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



natiques

CC: LHCONE

Quelques résultats : Data challenges

Globalement les objectifs du challenge ont été atteints.

- L'objectif de satisfaire 10 % des besoins du HL-LHC ont été franchis.
- Les infrastructures actuelles (réseau et stockages) sont en mesures de répondre à 10 % du besoin HL-LHC.
- Les services de data management ont été en mesure de gérer les mouvements de données.
  - 30 Po de data transférés en 5 jours

Quelques points notables tout de même.

- Des disparités dans la façon dont le challenge a été géré par les diverses applications.
  - CMS a eu quelques difficultés dans la mise en œuvre du challenge.
- Une difficulté identifiée dans le suivi (monitoring).
  - Crosscheck entre vue expériences/vue service/vue site.

Quelques résultats : Tape challenges

Le but de ce challenge est de valider les capacités d'usage (read/write) des systèmes de stockage sur bande.

Un élément important dans cet exercice est le fait que les 4 expériences du LHC ont sollicité simultanément les solutions de stockage sur tape.

A garder en tête que les solutions de stockage sur bande sont toutes différentes.

- Configuration.
- Solutions techniques disponibles.
- # d'expériences quelles servent.

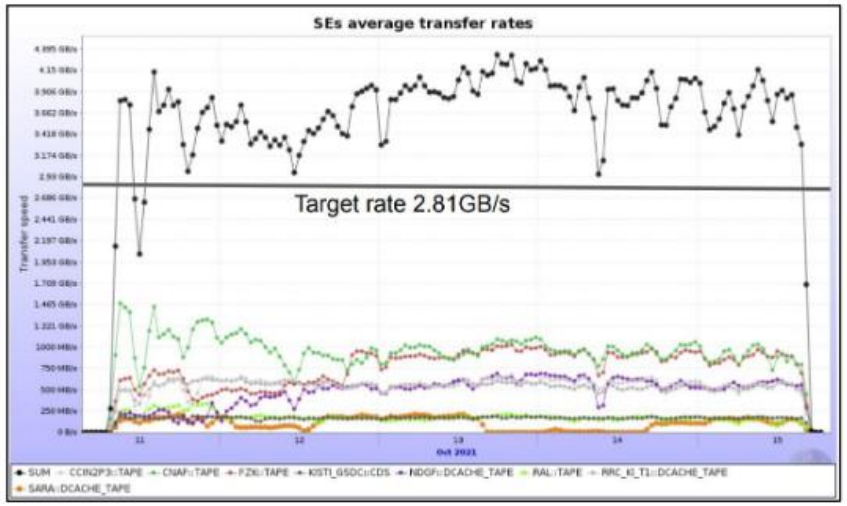
Des objectifs ont également été définis.

Overall T1s objectives for RUN3:

Indicate in this table, the bandwidths required for all T1s for reads and writes during data taking (DT) and right after data taking(A-DT).

VO	Reads (DT) GB/s	Writes (DT) GB/s	Reads (A-DT) GB/s	Writes (A-DT) GB/s
ALICE	0	2.8	1.1	2.8
ATLAS	2.5	9.6	8.4	5.1
CMS	0.8	7.6	12.3	1.1
LHCb		11	3.38	
Total	2.5	24.78	25.18	8.3

Quelques résultats : Tape challenges



T1 Centre	Target rate GB/s	Achieved rate GB/s
CNAF	0.8	0.94 (116%)
IN2P3	0.4	0.54 (130%)
KISTI	0.15	0.16 (106%)
GridKA	0.6	0.76 (123%)
NDGF	0.3	0.47 (144%)
NL-T1	0.08	0.1 (122%)
RRC-KI	0.4	0.53 (128%)
RAL	0.08	0.17 (172%)

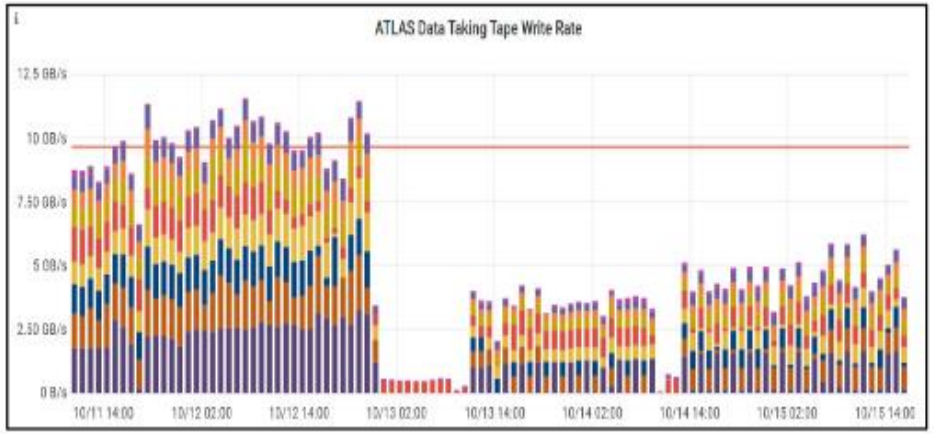
Sum 2.81GB/s

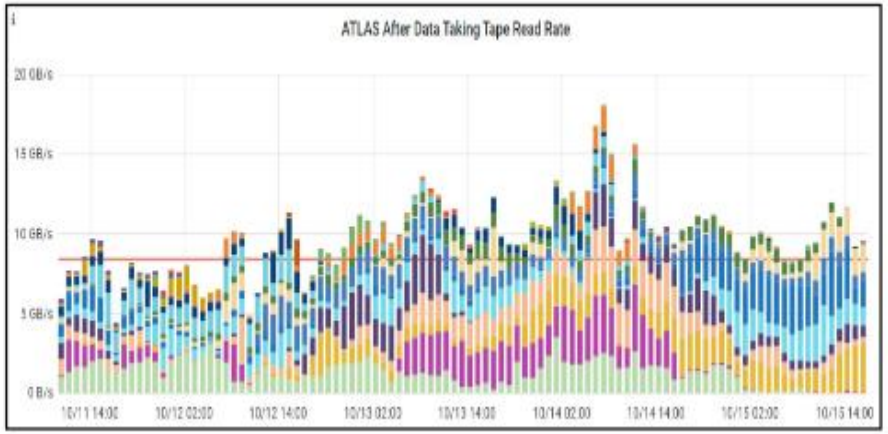
ALICE

- Ecriture seulement.
- Aucune difficulté.

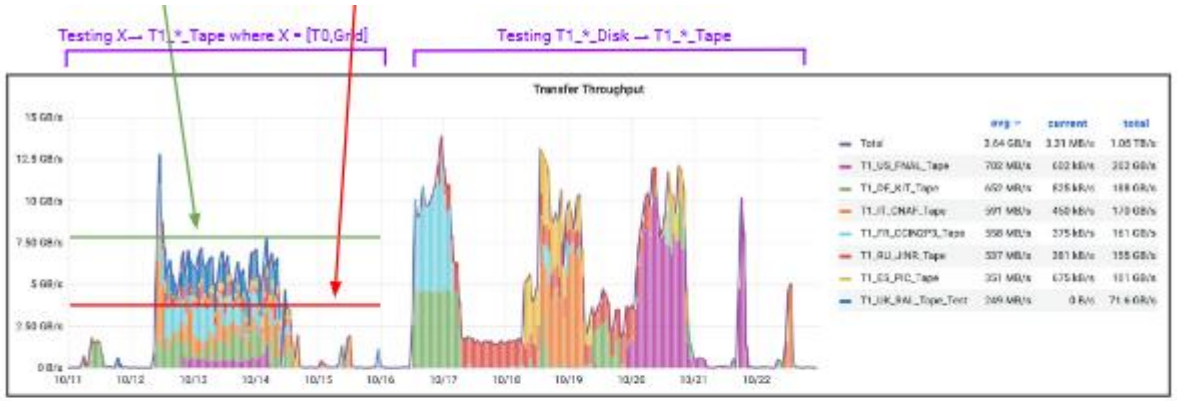
ATLAS

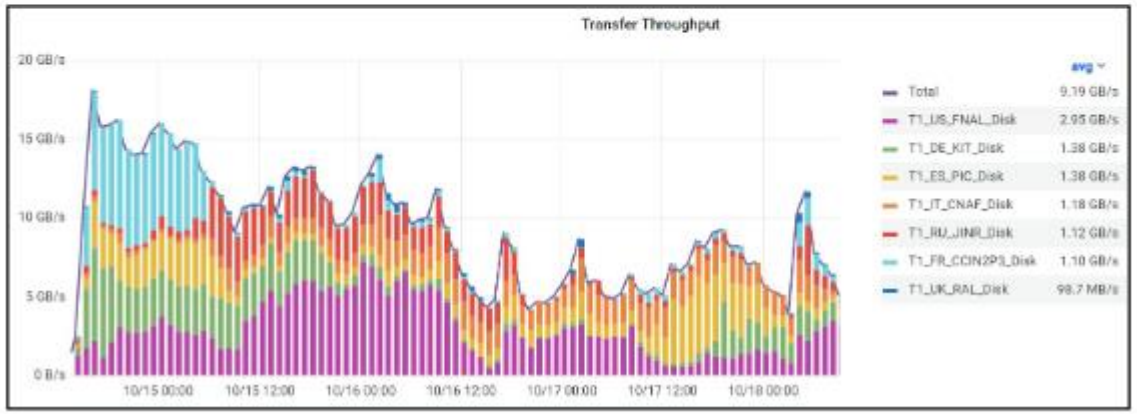
- Read/Write.
- Objectifs atteints.
- Une dépendance au taux d'activité du système de stockage.





Quelques résultats : Tape challenges



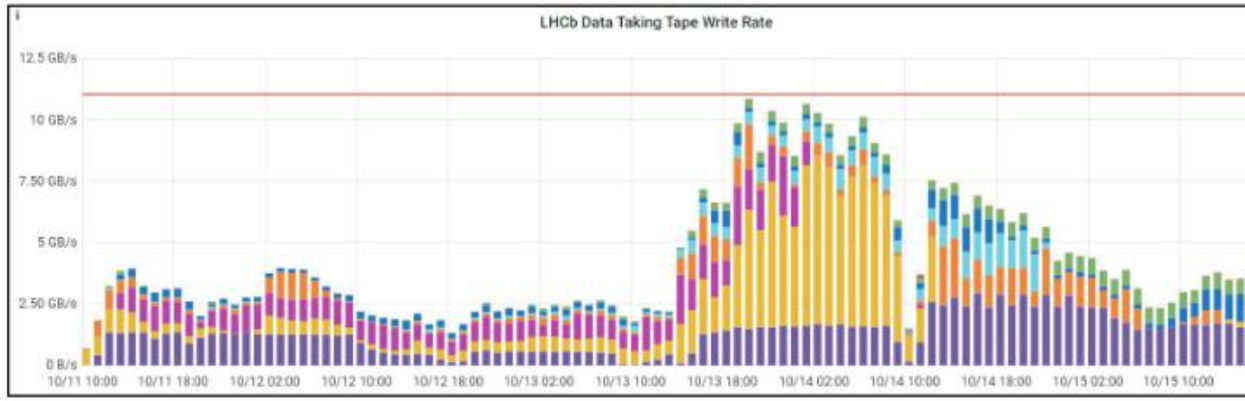


CMS

- Read/Write.
- Objectifs pas atteints.
- Des difficultés au niveau opérationnel (séquençage des demandes de stagging).

LHCB

- Ecriture seulement.
- Objectifs atteints en fin de challenge.



Quelques résultats : Tape challenges

Les objectifs du challenge n'ont pas toujours été atteints, mais ont mis en évidence surtout des difficultés d'ordonnancement entre les expériences pour un accès à une ressource concurrentielle.

Ce type de challenge doit s'inscrire dans un temps long pour s'affranchir des effets asynchrones entre les sites.

Comme dans le data challenge, une difficulté identifiée dans le suivi (monitoring).

- Exacerbée par le fait que les système de stockage sur bande sont susceptibles de fournir un nombre important de métriques significatives pour caractériser le système et ses performances.

Conclusion et remarques

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.

Conclusion et remarques

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.

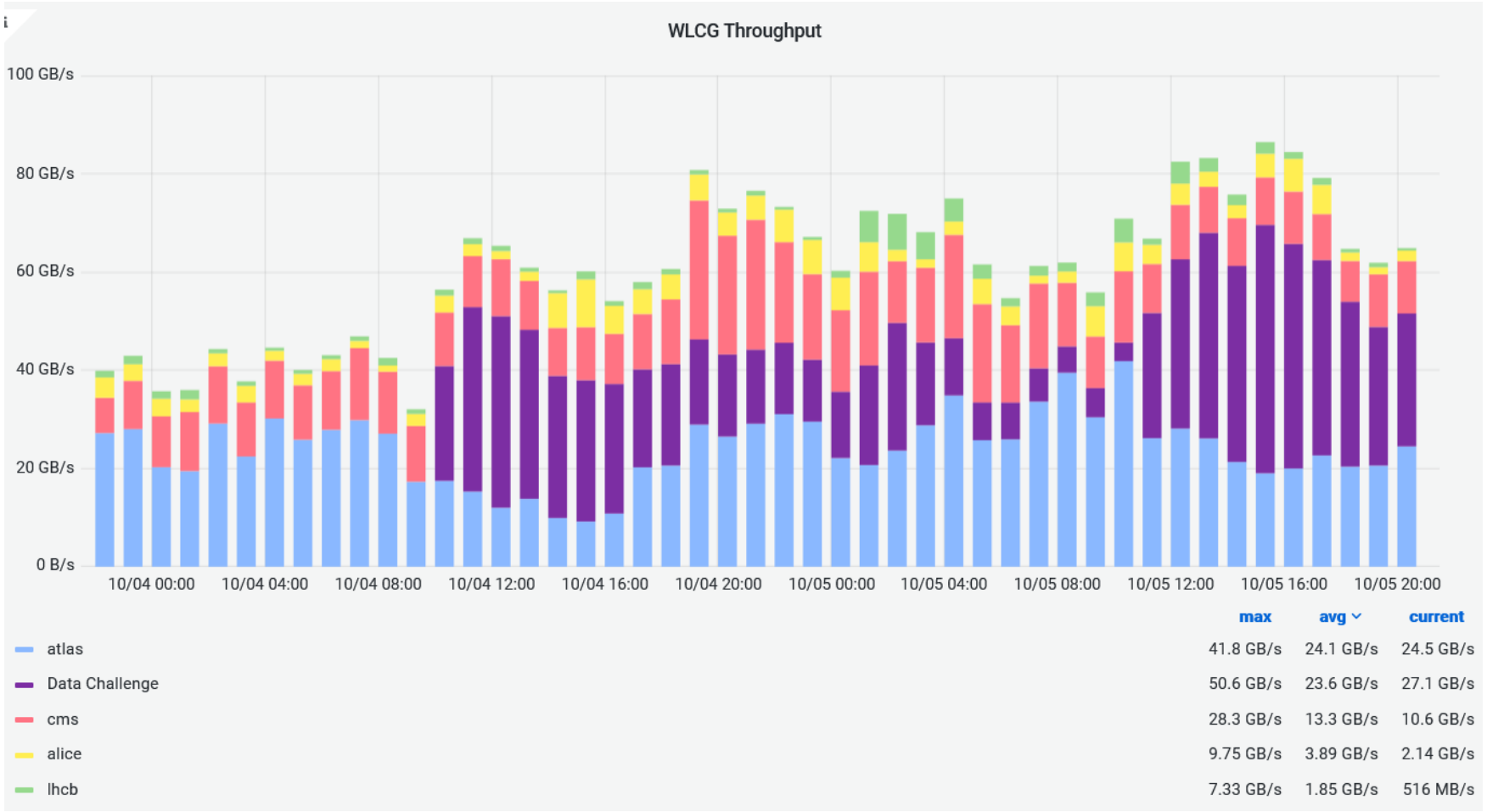
- HL-LHC Computing review.
- HL-LHC TDR.

HL-LHC n'est pas si loin.

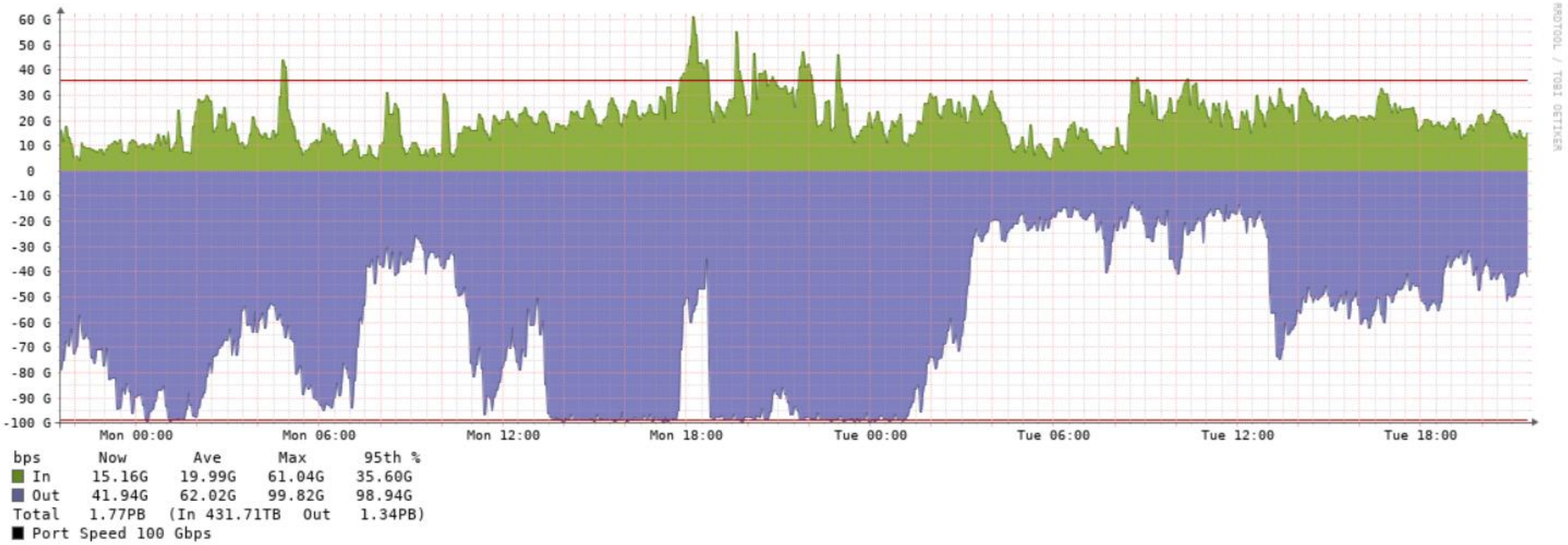
- Les ressources que l'on achète en ce moment seront encore en production en 2027.

?

Backup : Network data challenge



Backup : Network data challenge lhcone at CC IN2P3

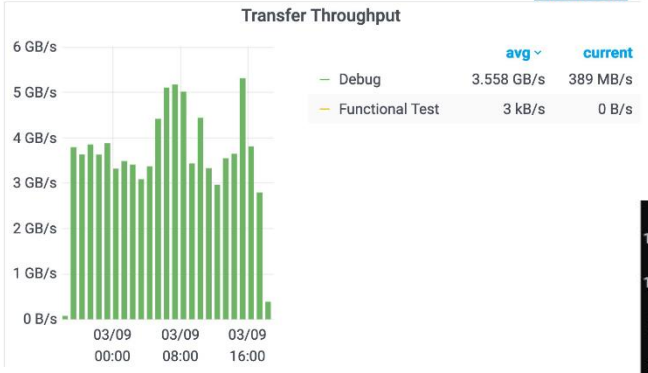


Backup : DATA CAROUSEL

Overview: ATLAS vs CMS staging throughput

ATLAS VIEW

CMS VIEW



CC-IN2P3 VIEW
- **ATLAS TEST: 1.7GB/s**
- **CMS TEST: 4.6GB/s**

CC-IN2P3
Test Data Carrousel Mars 2021