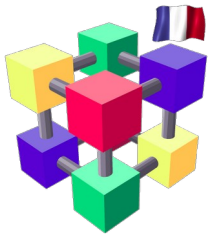
Workshop Data Challenge 2024

David Bouvet, Laurent Duflot, Eric Fede
Journées LCG-France – 29 nov. - 1 déc. - CC-IN2P3



Sommaire

- DC24 : généralités
- Objectifs des expériences
  - ALICE, ATLAS, CMS, LHCb, Dune, Belle II
- Monitoring
- Middleware
- Réseau

Date : 12 février → 23 février

Buts :

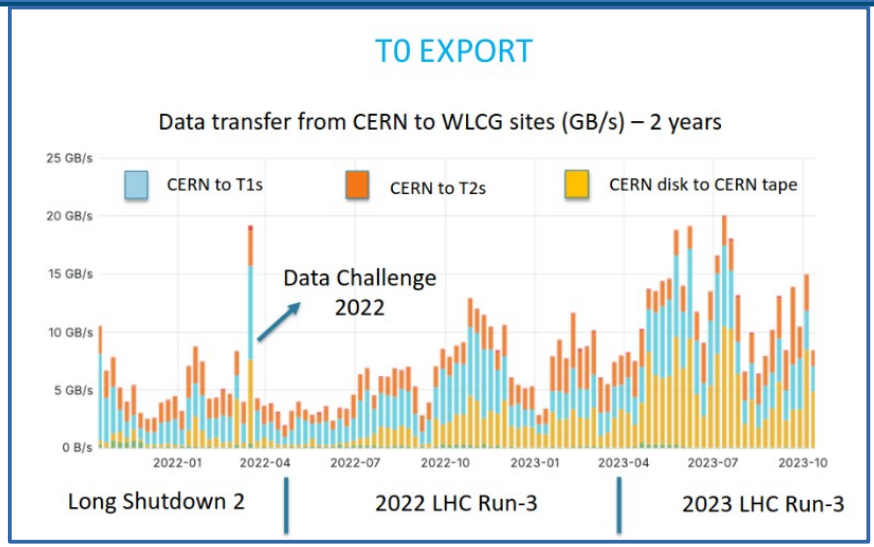
- Valider les capacités de transfert en prévision du HL-LHC :
  - 25 % du HL-LHC
  - maintenir le taux cible sur 48 h
- Augmentation des volumes/débits
  - 4 expériences en même temps
- Valider technologies/solutions ajoutées
  - pile logiciel stockage
  - fonctionnalités réseau

Modalités :

- Infrastructure de production
- Superposition à l'activité normale
  (pas d'interruption de la production habituelle)
- Transferts de vraies données

Rappels : 2 modèles

- Minimal : T0 → T1 et T1 → T2
  ⇒ 4,8 Tbps
- Flexible = 2 x minimal
  (minimal + T1 ↔ T1 et T2 ↔ T2)
  → scénario le plus réaliste
  ⇒ 9,6 Tbps
- Niveau actuel = flexible du DC21



Communications :

- Téléconf. quotidienne : opérateurs des expériences + toute personne intéressée
  - statut, problèmes, plans pour le jour suivant
- Canal Mattermost :
  https://mattermost.web.cern.ch/wlcg-gdb/channels/wlcg-data-challenges

Pas de demande particulière

Sollicitation du T0 et des T1

- Taux proportionnels aux *pledges* des sites
  - CC-IN2P3 : 3,2 Gb/s
- Probable concomitance avec transferts des données du Run Pb-Pb 2023
- Monitoring via MonALISA
  - données peut-être intégrées aussi dans le monitoring commun

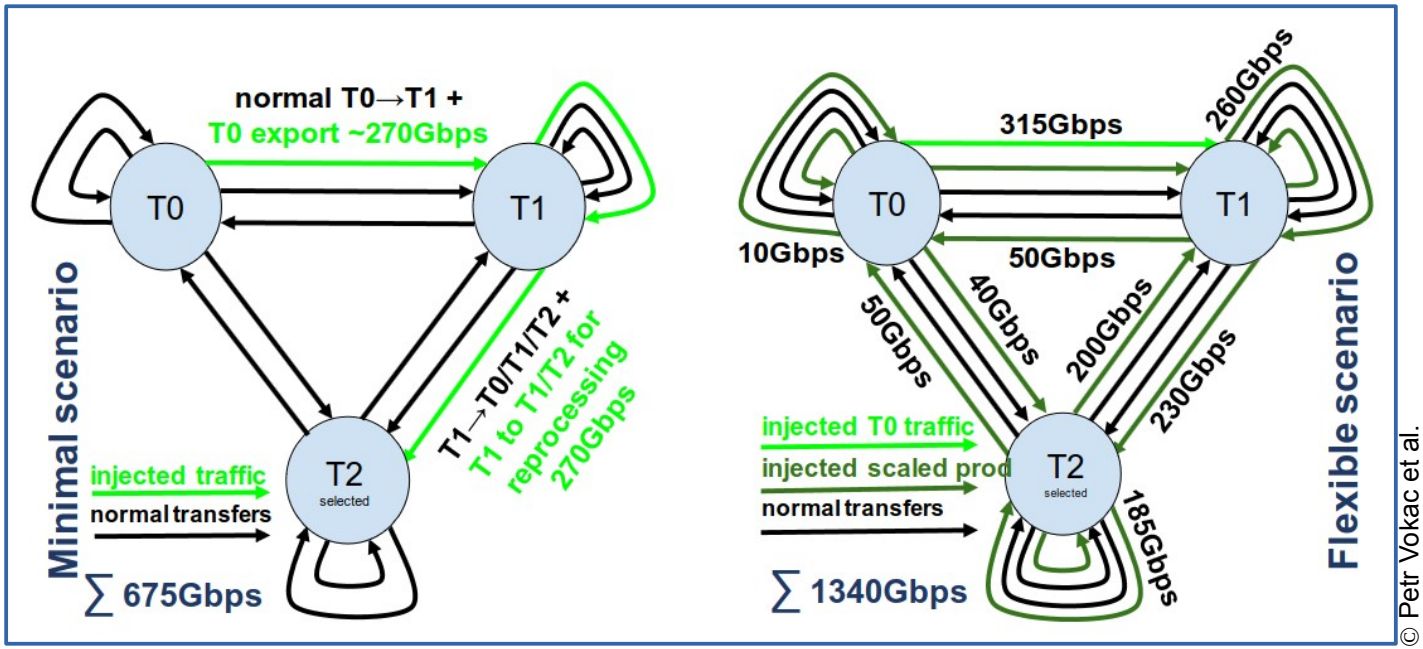
Centre	Target rate GB/s	Achieved rate GB/s
CNAF	0.8	2 (250%)
IN2P3	0.4	0.8 (200%)
KISTI	0.2	1 (500%)
GridKA	0.6	2 (300%)
NDGF	0.3	0.4 (133%)
NL-T1	0.1	0.9 (900%)
~~RRC-KI~~	~~0.4~~	~~0.53 (128%)~~
RAL	0.1	0.7 (700%)
CERN	*10*	*20 (200%)*

Target 2.5GB/s (T1s) + 10GB/s (T0)

T0, T1 et T2 impliqués

Vise le modèle flexible pour être le plus réaliste possible

- T0 → T1 ; T1 → T0,T1,T2 ; T2 → T1
  - T1 : disque uniquement
  - T2 : choix de participer ou non → **date butoir : fin nov.**
- Données réelles stockées sur les sites



Injection

- En plus de la production normale
  - fluctuations quotidiennes de la production d'un facteur 2-3 ⇒ niveau d'injection adaptable
  - trafic additionnel calculé pour chaque lien
- Outil d'injection semi-automatique
  - règles Rucio définies toutes les 15 min pour assurer la stabilité du trafic
  - données supprimées toutes les 2 h → espace à prévoir en conséquence
- Synchrone avec les autres expériences
- **Tests préparatoires dès décembre**

- Tableau des taux de transferts cibles (lien)

Table: DC24 (src)			Site WAN (Gb/s)		DC24 minimal scenario				DC24 flexible scenario			
			Total (Gb/s)	Usable by ATLAS	T0 Export	Total Gb/s & bandwidth		Space [TB/24h] (deletions/hour)	T0 Export	Total Gb/s & bandwidth		Space [TB/24h] (deletions/hour)
Site	Tier	Cloud				∑ ingress	∑ egress			∑ ingress	∑ egress	
IN2P3-CC	T1	FR	200	89	38.0	58.4	38.0	484 (7k)	38.0	93.2 - 102.5	76.0	926 (13k)
GRIF	T2	FR	100	100		7.3 - 8.4	6.4 - 6.4	40 (1k)		49.3 - 57.5	50.8 - 50.8	559 (8k)
IN2P3-LPC	T2	FR	100	100		2.1 - 2.4	1.4 - 1.4	9 (0k)		11.2 - 13.0	10.7 - 10.7	120 (2k)
IN2P3-LAPP	T2	FR	20	20		5.7 - 6.2	3.1 - 3.1	21 (0k)		17.4 - 19.9	17.8 - 17.8	165 (2k)
IN2P3-CPPM	T2	FR	100	100		2.5 - 2.8	1.5 - 1.5	12 (0k)		10.2 - 11.7	10.5 - 10.5	106 (2k)


Vise le modèle flexible via 4 flux de transferts

1) Export T0 : T0 → T1 : 250 Gbps
2) *Reprocessing* dans les T2 : T1 → T2 : 250 Gbps
3) *Production output* (production MC) : T1/T2 → T1/T2 : 250 Gbps
4) AAA : cas particulier de transferts FNAL vers T2 américains et du CERN vers T1/T2 d'Europe/Asie : 250 Gbps

⇒ au total un objectif de 750-1000 Gbps (~93-125 GB/s)

Injection

- En plus de la production habituelle
  - ajout de chacun des flux un par un
  - taux fonction des *pledges* des sites :
    1) T1 : *pledge* bande mais usage disque
    2) T1 : *pledge* bande mais usage disque; T2 : *pledge* disque
    3) *pledge* disque
    4) *pledge* CPU
- Même outil d'injection que ATLAS
- Synchrone avec les autres expériences
- **Tests et mini challenges préparatoires entre nov. et fév.**

RSE	Ingress (GB/s)	Egress (GB/s)
T1_FR_CCIN2P3_Disk	5.576	4.494

RSE	Ingress (GB/s)	Egress (GB/s)
T2_FR_GRIF	1.661	0.405
T2_FR_IPHC	1.148	0.315

Plan journalier

Flux

	1		1+2	2	2+3		
Day of challenge	1	2	3	4	5	6	7
Day of week	Monday	Tues	Wed	Thur	Fr	Sat	Sun
Scenario	T0 export	T0 export	**Mixed**	T1 export	**Mixed**	**Mixed**	**Mixed**
			T0 export		T1 export	T1 export	T1 export
			T1 export		Prod. output	Prod. output	Prod. output
Mode	"Data taking"	"Data taking"	T1 read+write	T1s -> T2s	T1s <-> T2s	T1s <-> T2s	T1s <-> T2s
T0->T1s	31	31	31	0	0	0	0
T1s->T2s	0	0	31	31	31	31	31
T2s->T1s	0	0	0	0	31	31	31
AAA	0	0	0	0	0	0	0
Total rate (GB/s)	31	31	62	31	62	62	62
Total rate (Gb/s)	248	248	496	248	496	496	496

Flux

	4	1+2+3	1+2+3+4		
Day of challenge	8	9	10	11	12
Day of week	Mon	Tues	Wed	Thur	Fri
Scenario	AAA	**"Max throughput"**	**"Max throughput"**	Contingency	Contingency
		T0 export	T0 export		
		T1 export	T1 export		
		Prod. output	Prod. output		
			AAA?		
Mode	CERN/FNAL to	Everything	Everything	?	?
	T1s + T2s				
T0->T1s	0	31	31		
T1s->T2s	0	31	31		
T2s->T1s	0	31	31		
AAA	31	0	31		
Total rate (GB/s)	31	93	124		
Total rate (Gb/s)	248	744	992		

Même tests que pour le DC21 avec

- Des débits plus forts
- Utilisation des bandes pour les T1
- *Token* si possible (pas encore prêt)

Injection

Site	shares	Data written (TB)	Export speed	Staging Speed (GB/s)	Staging duration (hours)
IN2P3	10.93%	231.38	1.53	1.20	53.77

© Alexander Rogovskiy et al.

- Ecriture et lecture séparément
  - écriture → 2 jours
    - EOS → CTA T0
    - EOS → disques T1 → bandes T1
    - suppression des disques T1
  - lecture → à partir du 20 fév.
    - bandes T1 → disques T1

Vise la simulation des conditions de transferts pour la HI

- KEK → T1
  - disque seulement
- Stress de leurs infrastructure et services
  - réseau, SE, FTS, Rucio, IAM (partiellement *token*), protocoles, monitoring

Injection

- Facteur 5 sur leur besoin → 18,5 Gb/s
- Utilisation d'un seul jeu de données plusieurs fois
- Tests préparatoires de nov. à jan.

Site	Country	#5G Files	Replica Factor	Total TB	Ingress (Gb/s)	Egress (Gb/s)
IN2P3CC	FR	1200	5,0	30	2,8	0

© Silvio Pardi

En plus

- Test du marquage de paquets
- Transferts entre T1
  - trafic sur LHCOPN

Vise la simulation de l'archivage de 25 % des données brutes

- BNL (SURF) → FNAL (1 GB/s)
- Réplication sur bandes dans les sites de données
- Réplication sur disques dans les sites de traitement

Injection au moment des pics WLCG du DC24

Pas de site FR impliqué

Monitoring

XRootD

- Difficulté au niveau du protocole
  - remplacer protocole UDP (perte de données)
  - solution : service shoveler (UDP → TCP) envoie données dans MONIT
- Portes XRootD pour dCache, STORM…
  - données issues de Kafka et poussées vers MONIT
    - script d'intégration → dCache 9.2

Réseau

- Description réseau + monitoring dans CRIC
  - sites FR ✔

Expériences

- Réutilisation et extension du dashboard DC21
  - ATLAS, CMS et LHCb
  - intégration données XRootD + réseau

Middleware

XRootD

- Support SciTokens depuis 5.1.0 (02/2021)
- Paquets Firefly depuis de 5.4.0
- SciTag à partir de 5.6.3
- Monitoring : pas mal de possibilités offertes

dCache

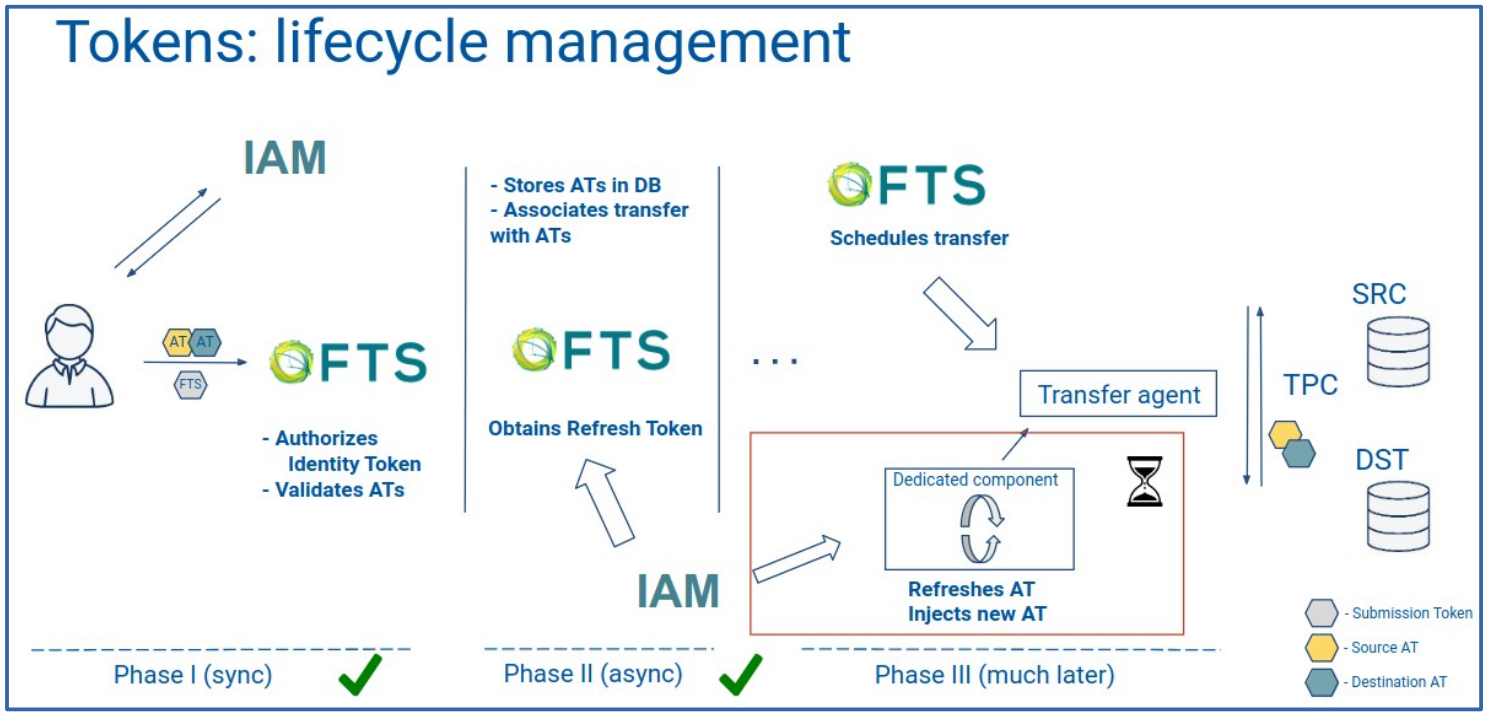
- Java Flight Recorder
  - outil pour post analyse
    - à partir de 7.2, mais plus simple d'utilisation depuis 9.1
- *Tokens*
  - support des ID Tokens (identité et/ou groupe) et Access Tokens (action possible)
- Recommandation de passer à 9.2
  - gain des SciTags (pour marquage paquets)
  - amélioration monitoring : ajout info. source/destination

EOS

- WLCG et ScitTokens supportés par HTTPS, XRootD et EOS
- ALICE Token supportés par HTTPS et XRootD, mais pas compatibles EOS
- Pas de support des *tokens* ZTN demandés par CMS pour AAA
- Support *token* manquant pour *tape REST API* y compris pour les API génériques

FTS

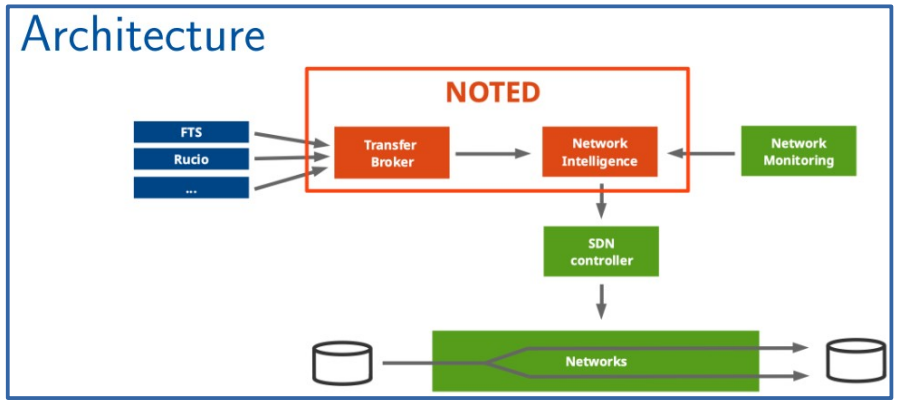
- Propagation SciTag possible jusqu'au SE
  - Support SciTag depuis FTS 3.12.11 (en-tête HTTP) et GFAL2 2.22.0
- *Token*
  - support prêt pour le DC24 (token-exchange + refresh) standard OAuth2 → sera testé pendant le DC24 (manque le mécanisme de rafraîchissement du *token*)
  - fts-token (1 *token* de plus pour identification/autorisation) + refresh-token + src-token + dest-token

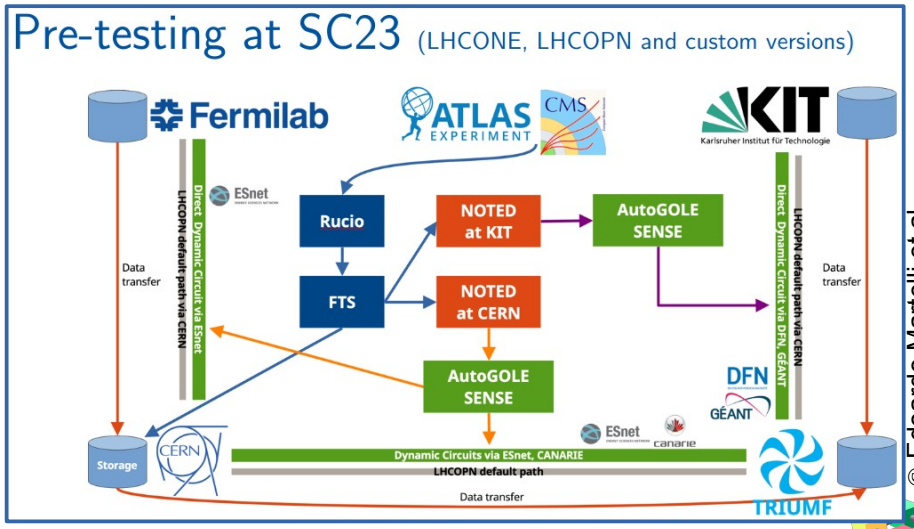


© Mihai Patrascoiu

NOTED

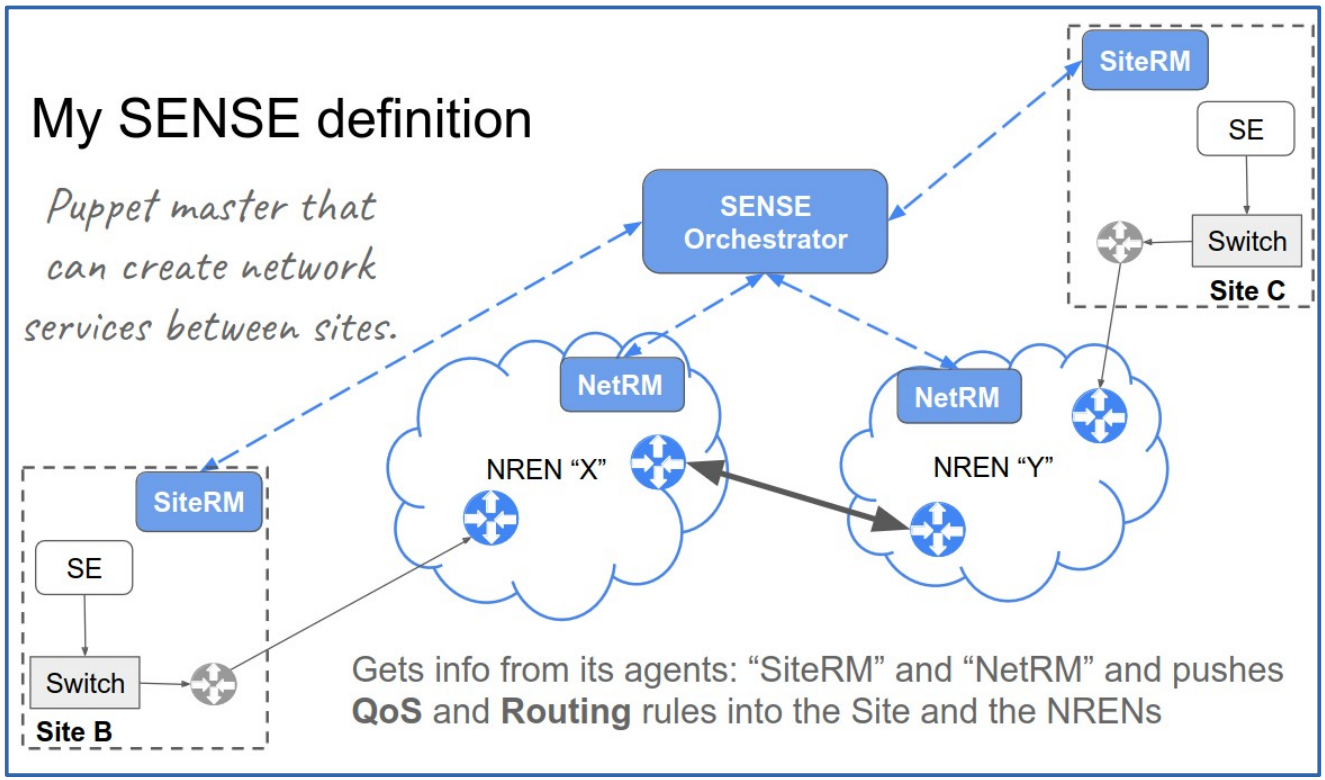
- Contrôleur de trafic pour « reconfiguration » réseau à la demande
  - Interaction avec
    - FTS → analyse des transferts
    - CRIC → connaître la topologie réseau
    - Contrôleur SDN → action de « reconfiguration »
  - déclenchement d'actions du contrôleur SDN suite à alarme de saturation réseau
- testé à SC23





SENSE

- Fournit à Rucio les QoS et les règles de routage des sites/NREN
  - interrogation SiteRM/NetRM
- Sélectionne un chemin prioritaire différent pour un transfert si besoin
  - pousse la nouvelle règle de routage via SiteRM/NetRM



perfSONAR

- Prochaine version devrait permettre une meilleure stabilité
  - version 5.0.5 en cours
- Campagne de mise à jour des serveurs à venir
- Création d'outils d'analyse
  - AAAS : système de souscription aux alertes
  - pS-Dash : interface web pour explorer les alertes

⇒ calendrier serré pour mise à jour des instances perfSONAR avant le DC24

Espacement paquets, gestion congestion TCP et Jumbo Frame

- Algorithme de gestion de congestion TCP
  - changement : CUBIC → BBRv3 (développé par Google)
    - taille des trames et espacement des paquets
- Tests Jumbo Frame : amélioration performances transferts

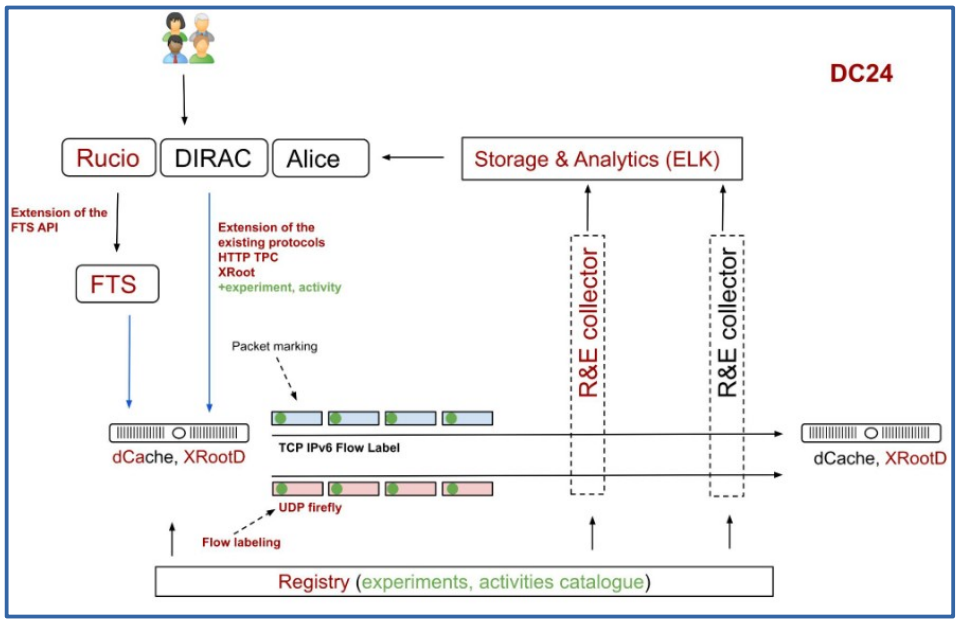
Marquage des paquets

- LHCOPN/LHCONE représentent 40 % du réseau Recherche Education

⇒ besoin d'identifier les différents flux au niveau réseau c-a-d quelle expérience/communauté

- 2 possibilités
  - SciTag dans en-tête des paquets (IPv6 seulement)
    - pas de perte
    - dépôt RFC pour « Flow Label »
  - UDP Firefly (IPv4 et IPv6)
    - insertion paquet UDP dans le flux de transfert
    - perte de paquet UDP possible

⇒ compatibilité middleware nécessaire



© Marian Babik et al.