Fédération de données pour CMS
Status & perspectives

Rencontres France-Grilles – LCG-France
CC-IN2P3, 26 novembre 2013
Sébastien Gadrat

PLAN

- ✓ Contraintes et difficultés
- ✓ Évolution du modèle de calcul de CMS
- ✓ Fédération de données AAA
- ✓ Status de la fédération & perspectives
- ✓ Placement dynamique des données
- ✓ Conclusion

Status run I et perspectives

✔ Le modèle de calcul a permis de remplir les objectifs en terme de physique pour le run I, néanmoins le besoin en stockage et CPU vont croître énormément...

- Perspectives du run II
  - Luminosité x5 (~100 fb-1)
  - Taux de trigger x2 (soit 2x plus de stockage sur bandes et disques)
  - Durée du run ~2x run I (donc volume ~2x par rapport run I)
  - Ressources CPU pour analyser tout ça...
  - Sans parler de HL-LHC...
- Nécessité d'optimiser l'utilisation des ressources existantes
  - Nouveau modèle de calcul

Contraintes et difficultés

✔ Modèle actuel, côté Tier-1, un seul système pour gérer disques et bandes

- Migration automatique des données sur bandes
  - Effacements fréquents pour libérer de l'espace (et données inutiles)
  - Staging pour mettre sur disque les données requises
- Utilisateurs « bannis » des Tiers-1
  - Pour éviter des stagings intempestifs

✔ Solution proposée : séparation des parties disques et bandes

- Les données sont directement stockées au « bon » endroit
  - Usage limitée (raisonnée) des bandes : moins d'effacements et de staging
- La partie disque peut être fédérée avec le stockage Tiers-2
  - Utilisation alors possible par les utilisateurs

Contraintes et difficultés

✔ Modèle actuel, côté Tier-2, gestion peu efficace de l'espace disque

- Les données de l'espace géré « officiellement » (groupes de physique) sont souvent peu accédées
- La gestion de ces données est chronophage et l'efficacité de l'utilisation du disque est faible

✔ Solution proposée : fédération de stockage/données avec placement dynamique de celles-ci

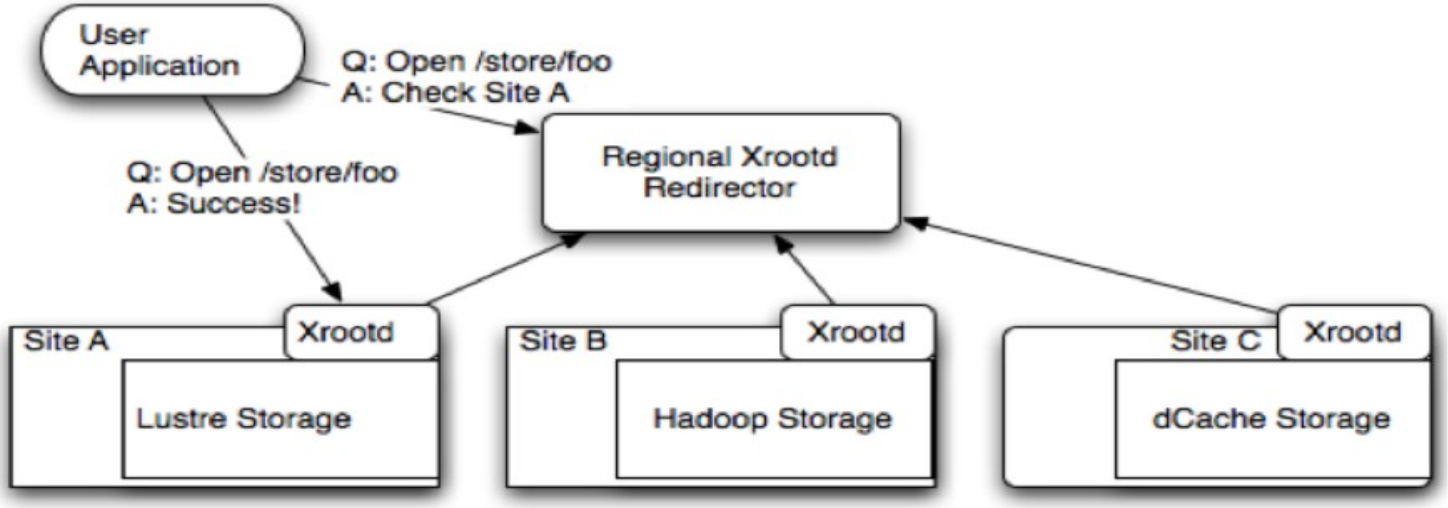
- Accès transparent aux données : « Any data, any time, anywhere »
- Gestion automatisée des données en fonction de leur popularité
- Possible grâce aux bonnes performances réseau et à leur bande passante élevée

Fédération de données

- Le but est l'accès le plus simple et le plus fiable aux données
  - Fiable (sans problème d'accès)
  - Transparent (localisation des données)
  - Facile (pas de surcoût pour le physicien)
  - Autorise l'utilisation opportuniste des ressources

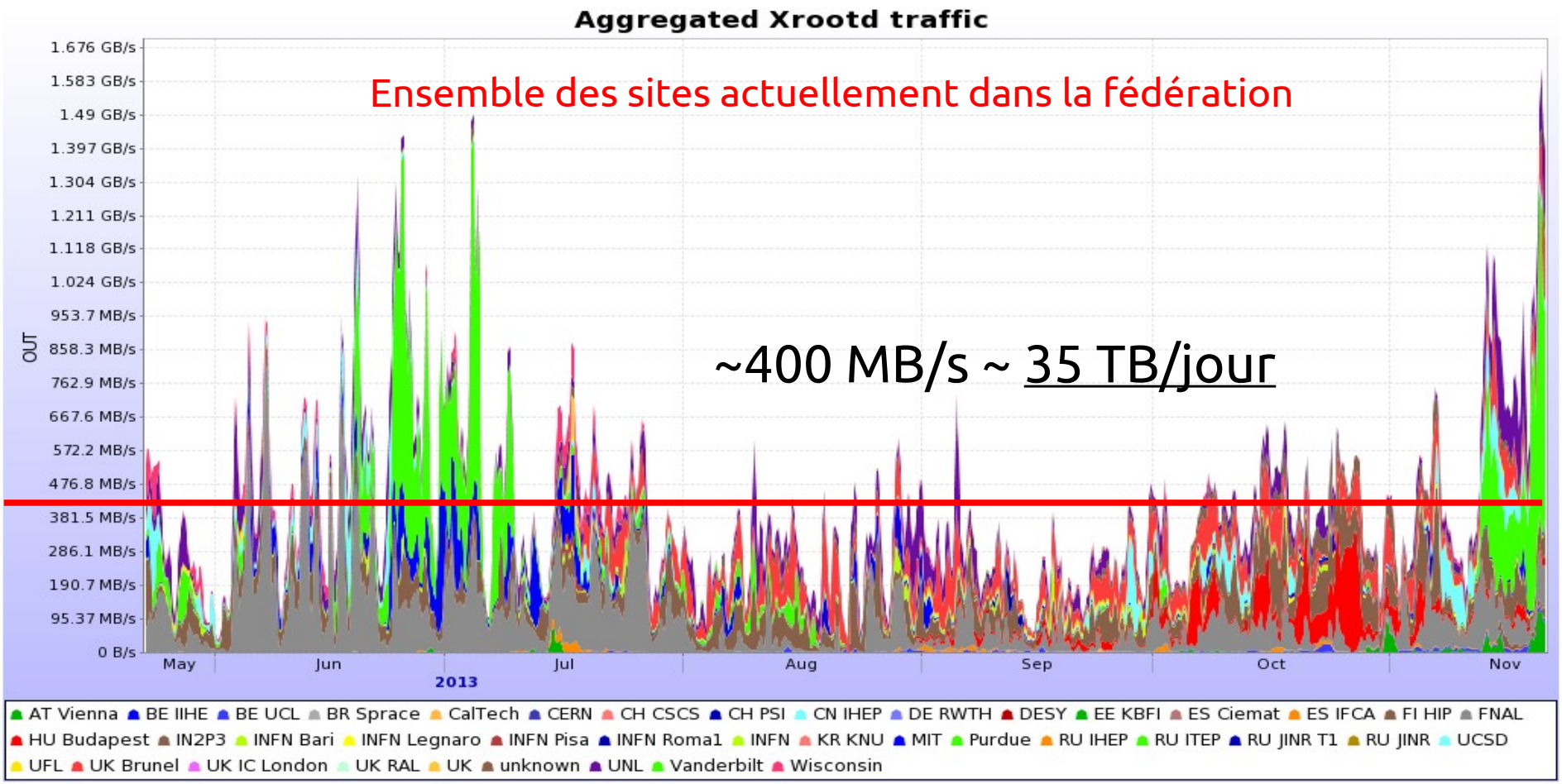
- Réalisation : mise en place d'une fédération de stockage permettant l'accès aux données de manière transparente grâce à un espace de nommage commun à l'ensemble des sites.

Comment ça marche ?



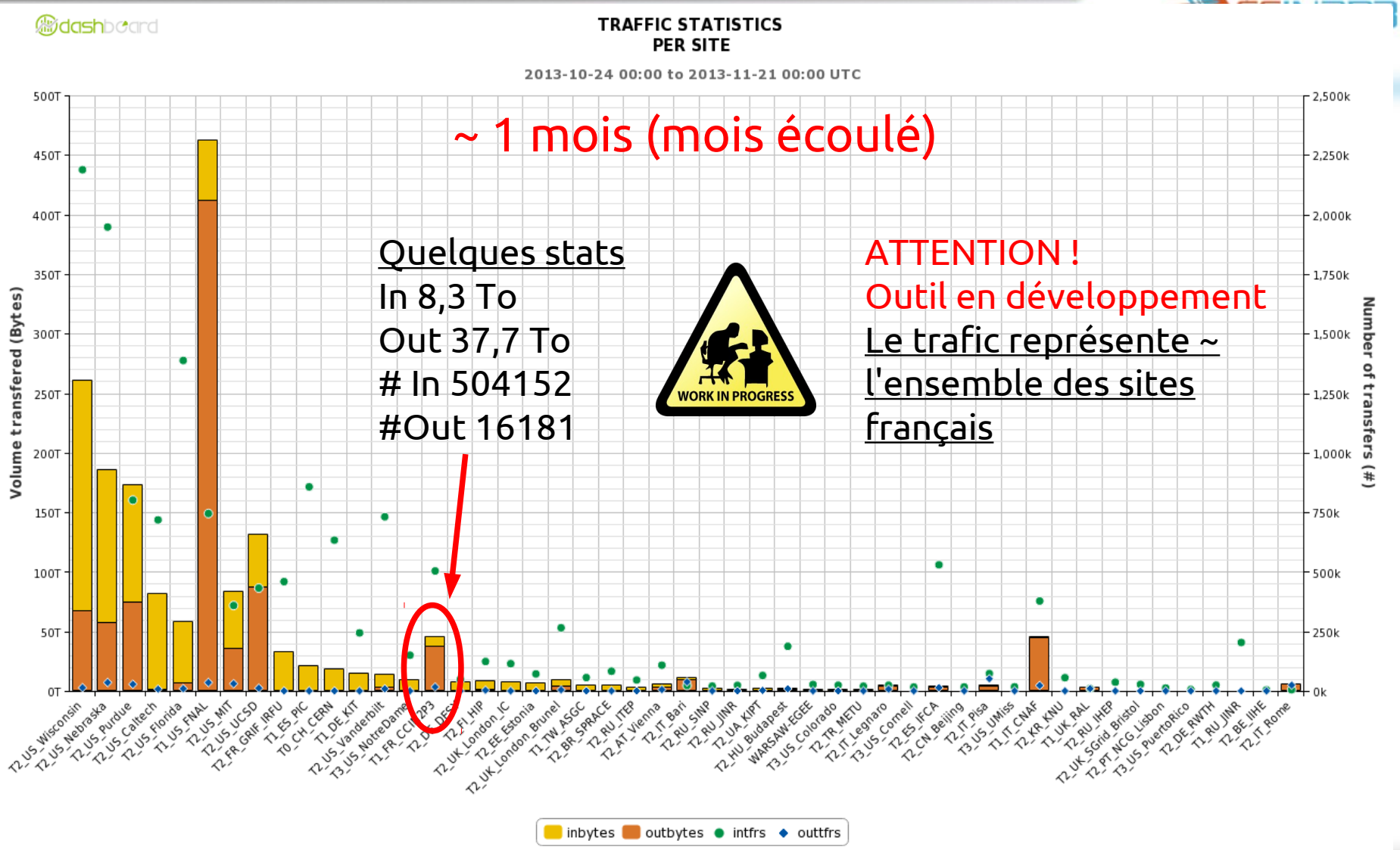
- ✔ Technologie sous-jacente : Xrootd
  - Interface homogène aux stockages hétérogènes
- ✔ Publication des données au niveau du redirecteur (nécessite serveur Xrootd)
  - Permet de trouver où les données recherchées sont disponibles
- ✔ Accès par authentification

Est-ce que ça marche (vraiment) ?

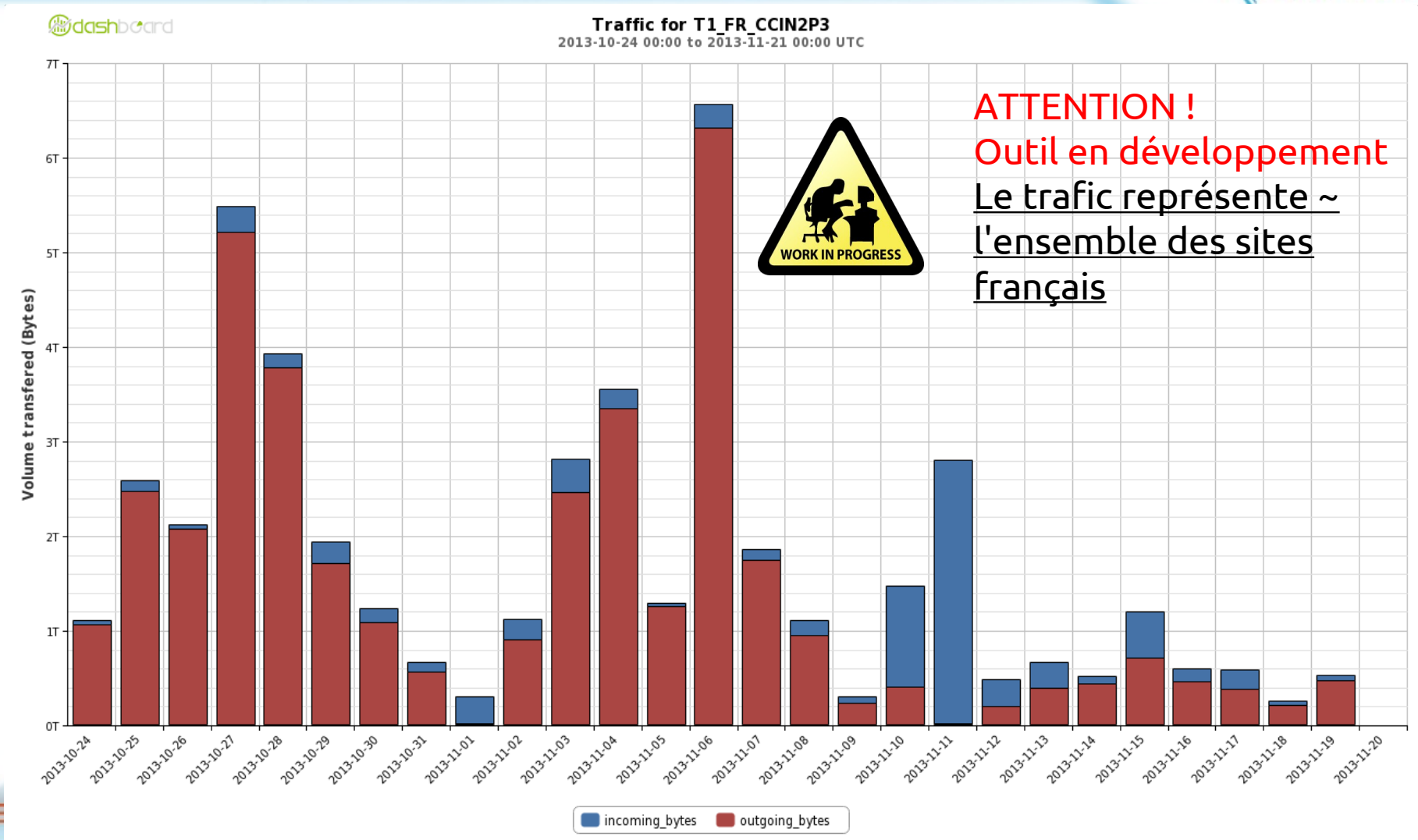


À titre de comparaison, les transferts PhEDEx représentent ~80 To par jour.

Trafic par sites



Historique du trafic



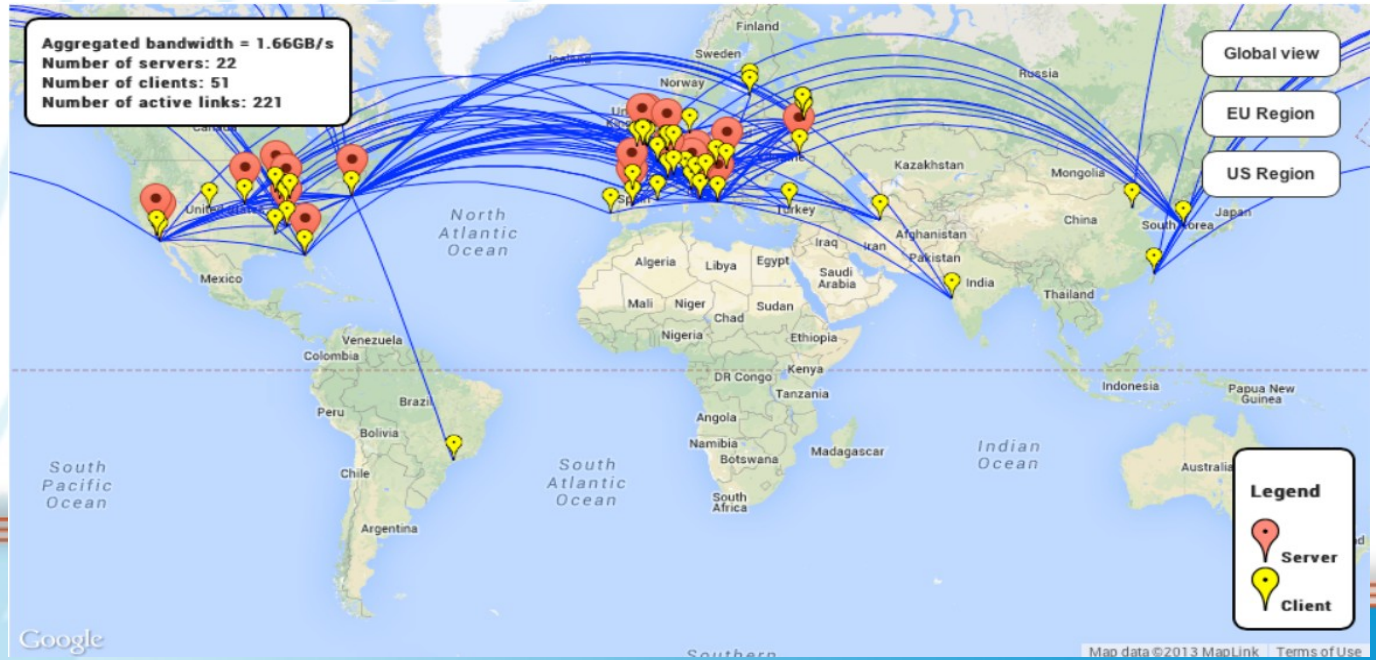
ATTENTION !
Outil en développement
Le trafic représente ~ l'ensemble des sites français

Utilisation de la fédération

- Utilisation en mode « Fallback »
  - Problème de lecture d'un fichier ? Le logiciel demande la lecture à partir d'un site distant
  - Processus transparent pour l'utilisateur, et qui réduit le taux d'échec des jobs
- Devrait permettre une utilisation « opportuniste » des ressources
  - Jobs, qui restent en queue sur un site, peuvent être migrés vers un second dont les ressources sont de suite disponibles
  - Perspective : devrait même permettre d'utiliser des sites avec peu ou pas de stockage propre
- Tests Tiers-3 : analyse de données à partir de la fédération (pendant une semaine ~800 jobs simultanés, 2-3 Go/s WAN, 99 % de taux de réussite)

Status de la fédération

- Déploiement en 2 étapes
  - Mécanisme de « Fallback »
    - 6/7 Tiers-1
    - 44/52 Tiers-2 (tous les Tiers-2 français)
  - Publication des fichiers au niveau de la fédération (par le biais d'un serveur Xrootd)
    - 3/7 Tiers-1 (nécessite la séparation disque-bandes)
    - 39/52 Tiers-2 (tous les Tiers-2 français)



Placement dynamique des données

✔ Afin de répondre au problème de gestion de l'espace disque au niveau des tiers-2, et d'optimiser son utilisation, CMS met en place un système automatique de gestion (transferts, effacements) des données sur les sites

- L'espace disque géré centralement par CMS devient un « cache » (~60 % du disque pour un Tier-2)
- Basé sur un service de « popularité » des données
- Effacement du cache, des données, en fonction de leur popularité

Service de popularité

- C'est la pièce centrale du PDD
- Il s'appuie (est alimenté par) :
  - CRAB (système d'analyse pour utilisateurs)
  - Xrootd (CERN)
- À venir
  - Surveillance Xrootd globale
  - Logiciel CMS de reconstruction et d'analyse des données

CMS POPULARITY

VICTOR LE NETTOYEUR

CERN IT ES

Vidage du cache

- Quand la popularité diminue, les données seront effacées :
  - Diminution du nombre de copies au niveau Tiers-2
  - Effacement de toutes les copies au niveau Tiers-2
  - Effacement de la/les copies au niveau Tiers-1
- Status : un algorithme existant en cours d'optimisation
- Plan : ajout de plusieurs algorithmes

Placement dynamique des données

- Nouvel échantillon de données
  - Répliquer sur 1 Tier-1 et 1 ou 2 Tiers-2
  - Ces premiers sites pourront être choisis en fonction de plusieurs critères pertinents
- En fonction de la popularité
  - Augmentation du nombre de copies au sein de la fédération
  - Localisation des copies en fonction de plusieurs critères dont les ressources disponibles au niveau des sites
- PDD couplé à la fédération
  - L'algorithme devant déterminer le meilleur compromis entre plus de copies, et accès par le réseau (fédération AAA)

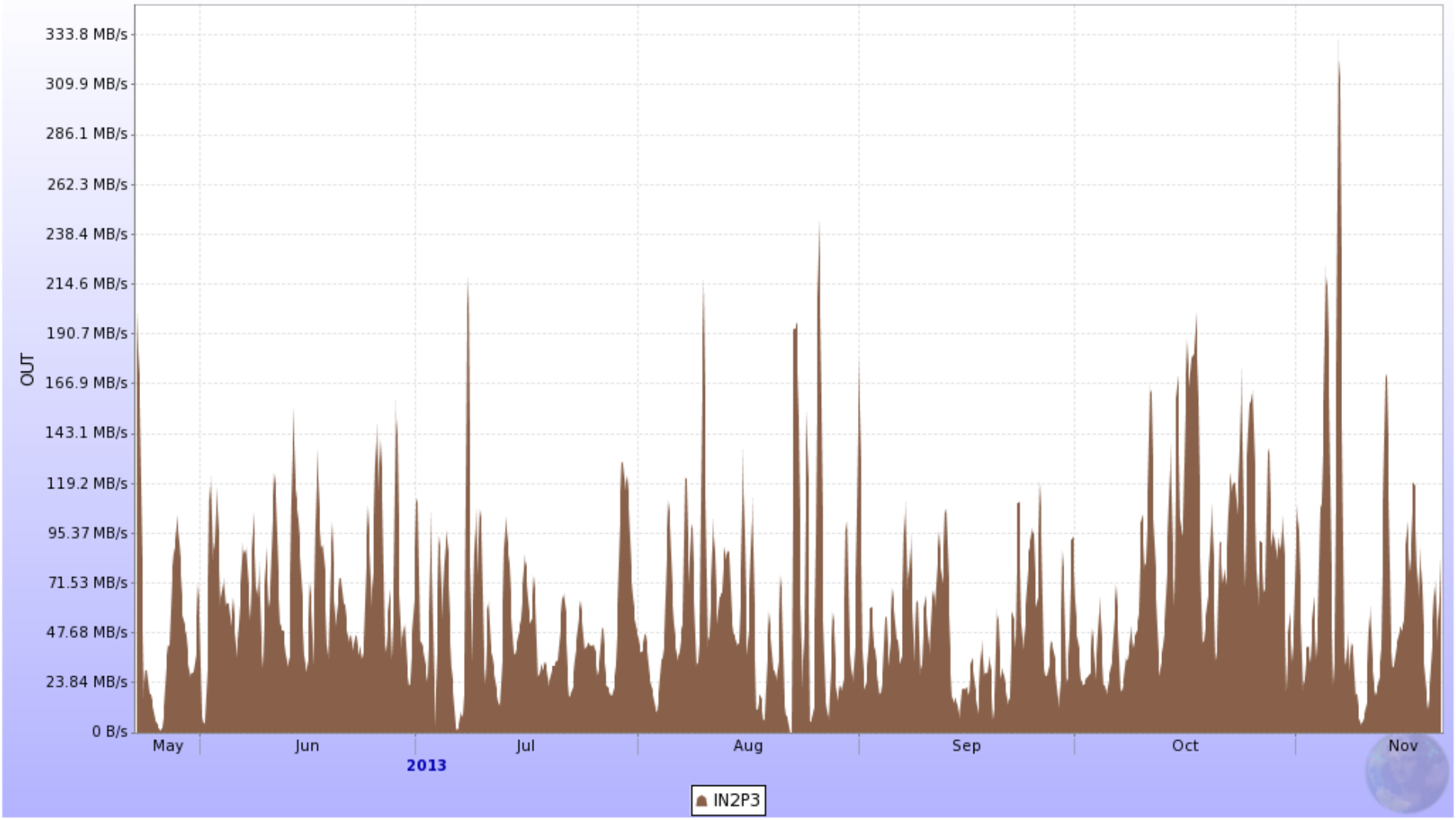
Conclusions

- CMS propose de profondément modifier son modèle de calcul
  - Meilleure souplesse et flexibilité pour la gestion des activités (aussi bien Tiers-1 que Tiers-2)
  - Accès transparent et fiable aux données (où qu'elles soient)
- Premiers résultats très encourageants
  - Sous-tendus par les bonnes performances réseau
- « Roadmap »
  - Séparation disque/bandes : fin 2013
  - Fédération AAA : ~ terminée
  - DDP : fin prévu début run 2 (~début 2015)

Accès local VS réseau

- Les accès locaux restent meilleurs que ceux à distance (jobs d'analyse au niveau Tiers-2)
  - Local : efficacité CPU 92 % avec 0,48 s/événement
  - Distant : efficacité CPU 86 % avec 0,65 s/événement
- Bonne performance des réseaux, et bande passante en augmentation

Aggregated Xrootd traffic



Statistics

TotalTraffic OUT

	Series	Last value	Min	Avg	Max	Total
1.	IN2P3	103.4 MB/s	0 B/s	64.15 MB/s	5.985 GB/s	963.2 TB
	Total	103.4 MB/s		64.15 MB/s		963.2 TB