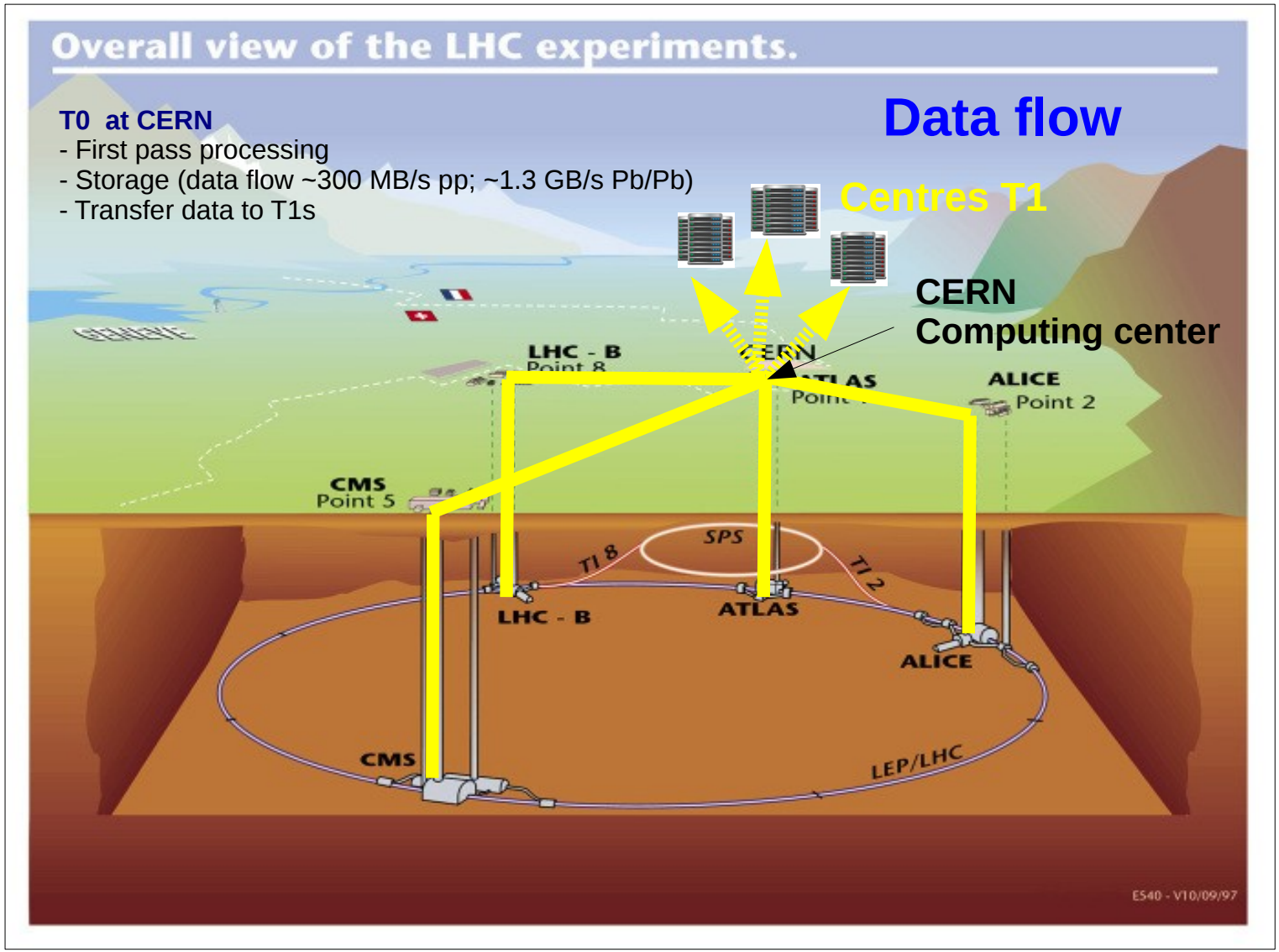
LHC Computing : Stratégies et technologies pour les données (expérience Alice)

Credits : Certaines diapos proviennent de la présentation de Catherine Biscarat aux journées Succes : http://succes2015.sciencesconf.org/

Large Hadron Collider [1] (LHC)



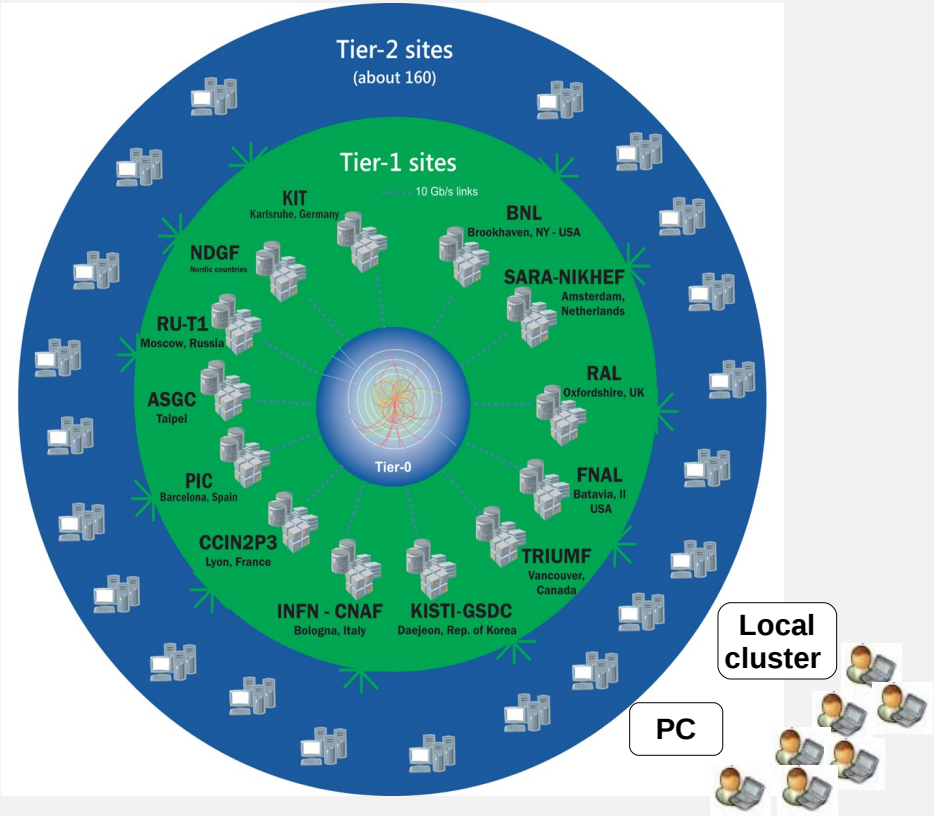
Grille de Calcul

- Avec les volumes de données et les calculs nécessaires, il n'est plus possible de tout faire au CERN
- Par ailleurs, les calculs sont très aisément sécables en petites unités indépendantes les unes des autres sur la base des évènements (intrinsèquement parallèles)
- Ceci explique en partie le choix de s'appuyer sur des technologies de grilles de calcul
- Mais la nécessité de traiter des volumes importants de données également réparties est source de complexité

Hiérarchie des sites / modèle MONARC

Premier modèle pour l'informatique au LHC (1999)

- Modèle en étoile, hiérarchique, distribué
- Focus sur le contrôle du réseau (1Gb/s attendu)



Tier-0 (CERN):
- Raw data storage
- Calibration
- Initial reco
- Data distribution to T1

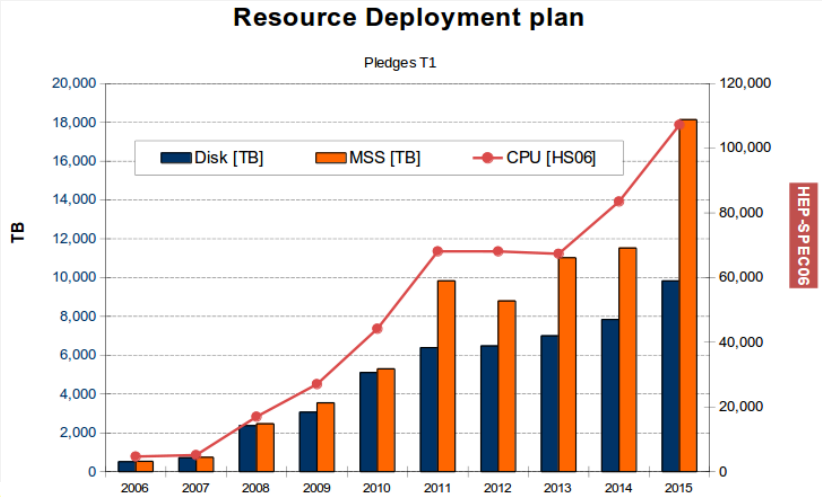
Tier-1:
- Long term archiving
- Subsequent reco passes
- Large scale organised analysis

Tier-2:
- Simulation
- End user analysis

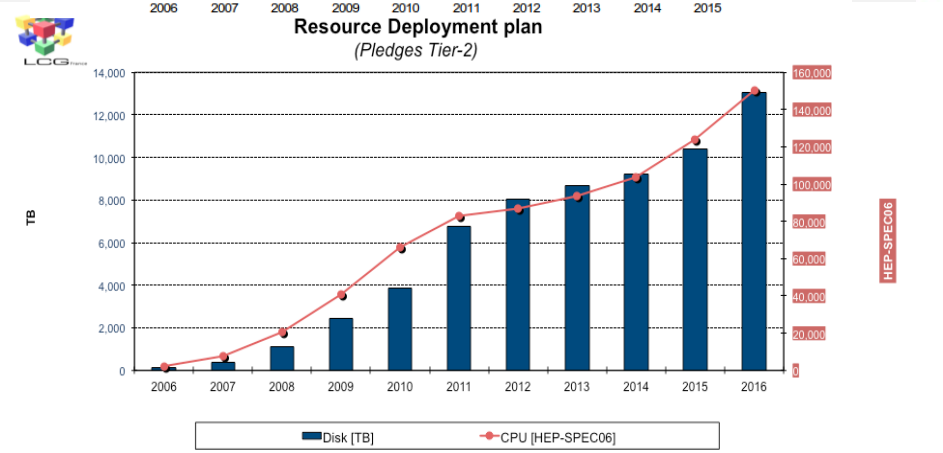
In addition (end user analysis):
- Tier-3
- Local clusters

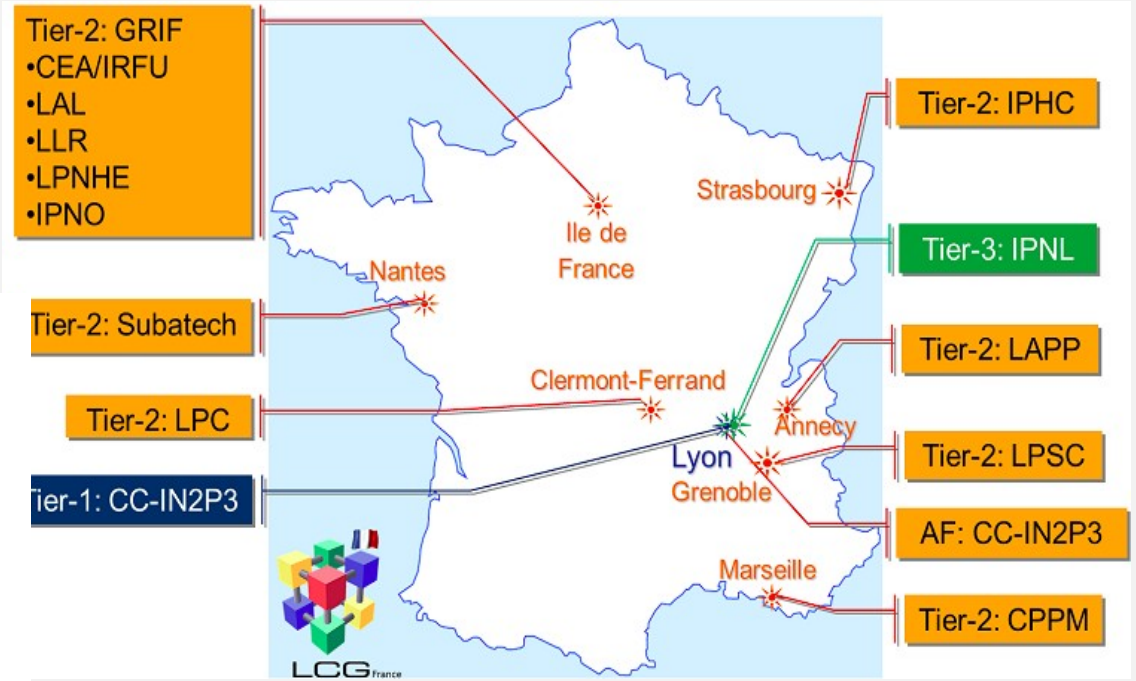
Les sites en France

- Organisés avec les expériences dans le projet « LCG-France » [3] - CNRS/IN2P3 et CEA/IRFU
- Fournir ~10% des ressources informatiques mondiales aux expériences LHC (Tier-1 et Tier-2)



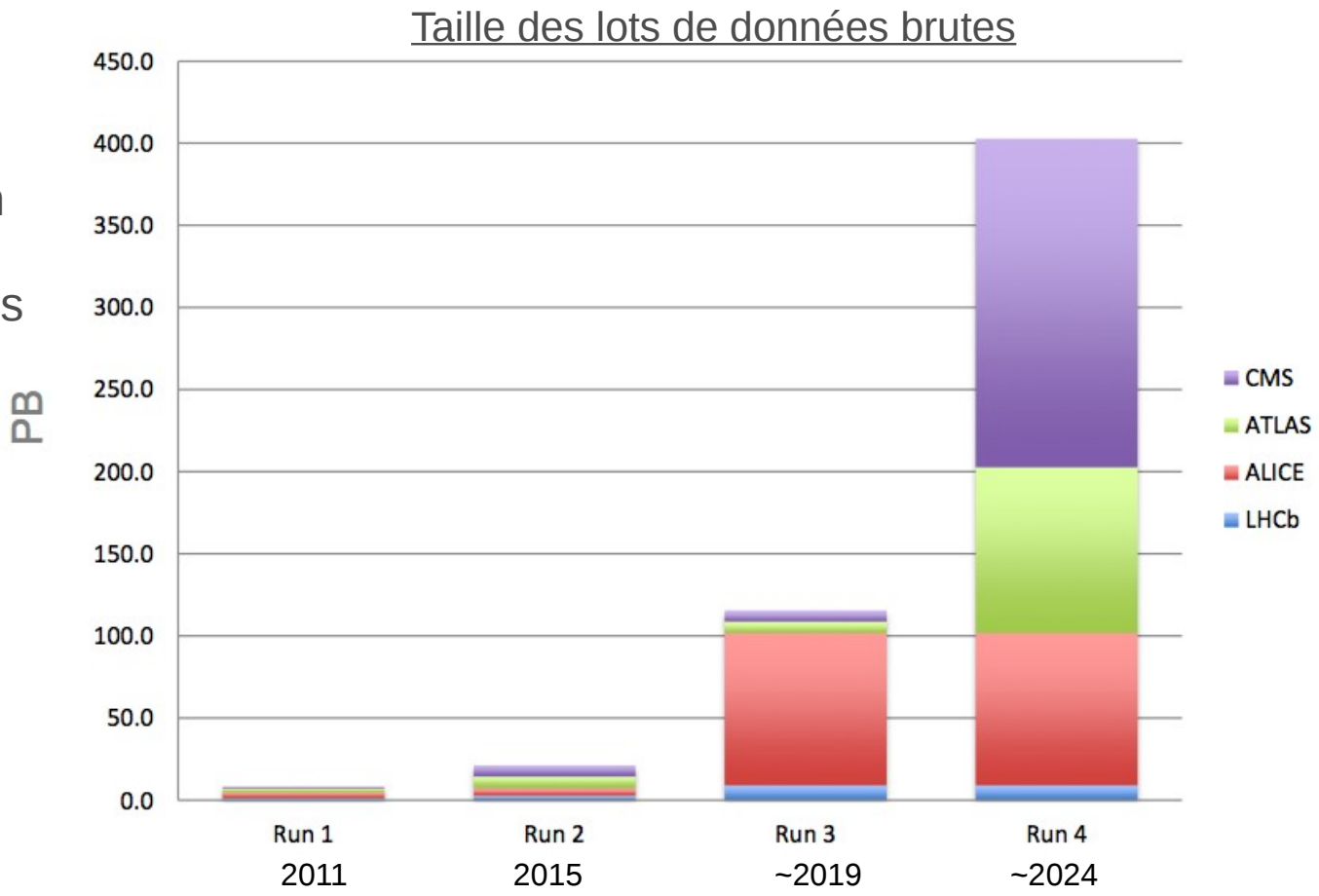
En 2015 au T1 : 18 PB bande, 10 PB disk, 110 kHS06
Les T2 doublent les ressources disk et CPU du T1.





Volumes de données du LHC

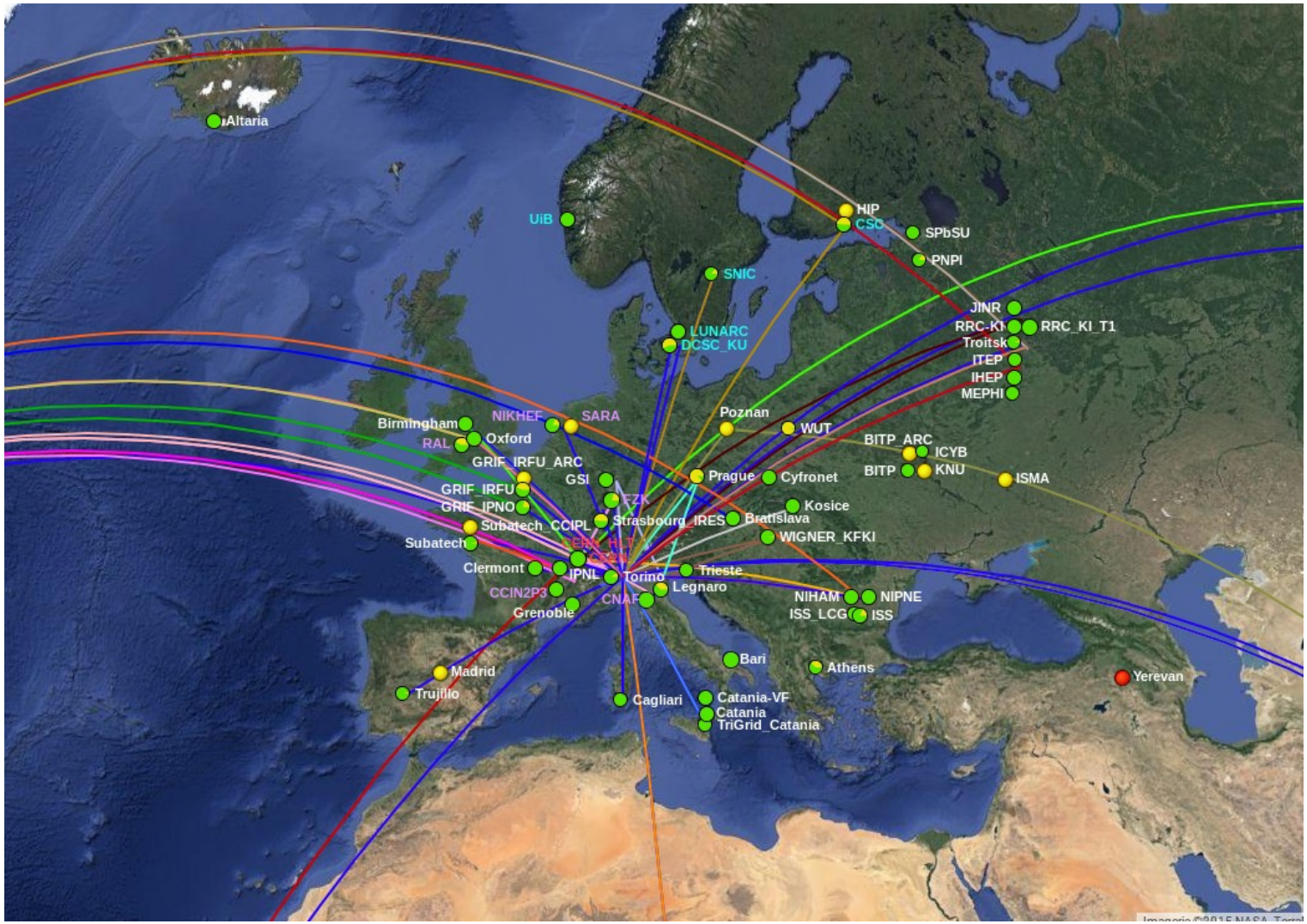
- Des besoins gigantesques
- Dépasse une extrapolation simple à budget plat
- Il faut repenser les modèles
- S'adapter aux nouvelles architectures et revoir les logiciels
- WLCG Technical Forum
- HEP Software Fundation
  - http://hepsoftwarefoundation.org/



Alice et AliEn

- AliEn (Alice Environment)
  - Framework de grille bati au dessus de briques OpenSource
  - Capable d'utiliser la grille EGI (mais aussi NorduGrid, OSG)
  - Pour gérer les différentes activités de calcul
  - Annuaire pour l'authentification et la configuration
  - Catalogue des fichiers pour les données (File Catalog)
  - Queue centrale pour les tâches de calcul
  - Monitoring MonaLisa

Alice en action (*)



(*) Vue limitée à l'Europe. Nombreux sites ailleurs dans le monde : Asie, Amériques,...

Gestion des données d'une expérience LHC

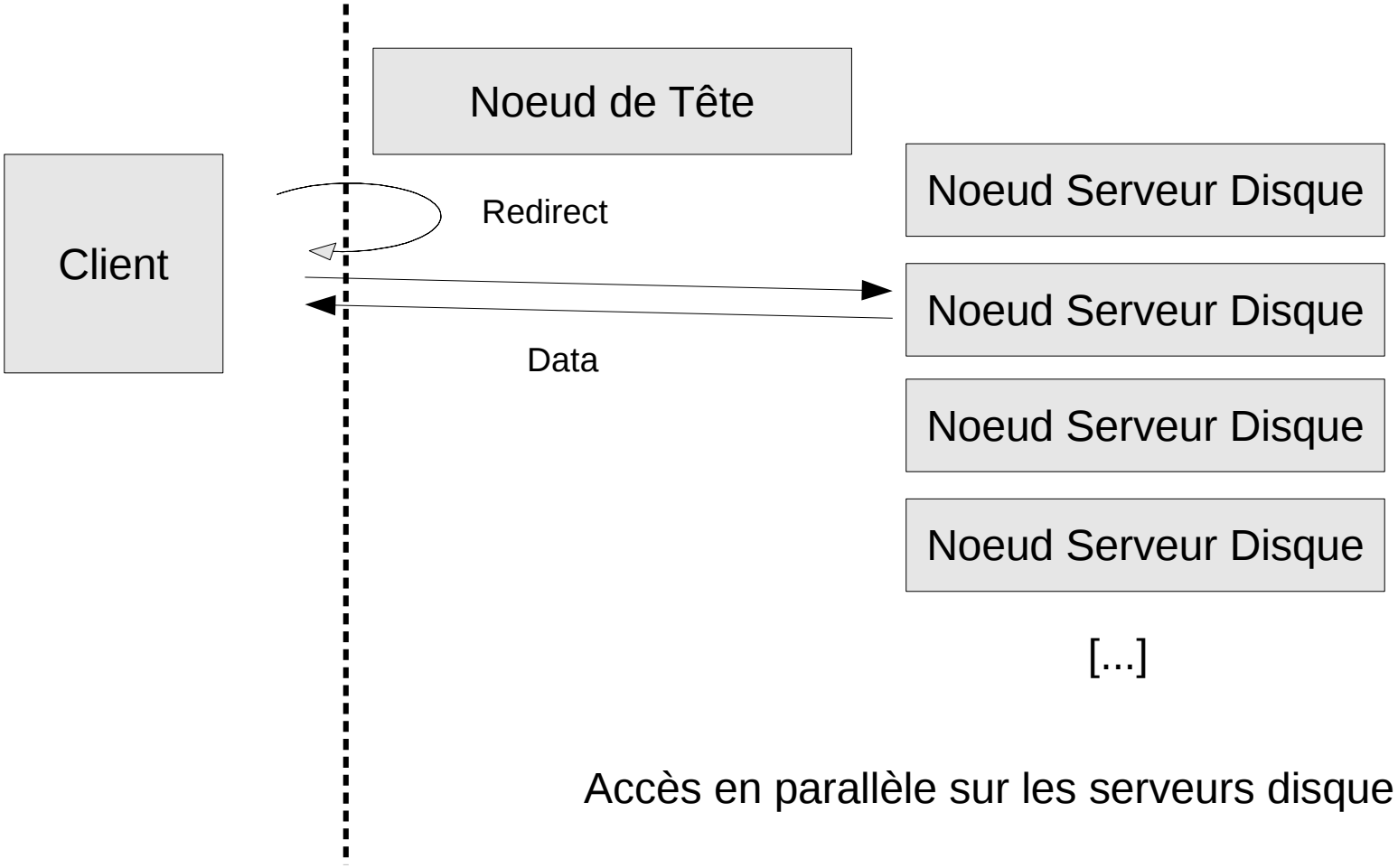
- Catalogue global des données de l'expérience
  - Nom logique (LFN) :
    /alice/data/2015/LHC15i/000236203/muon_calo_pass1/AOD/072/AliAOD.Muons.root
  - Nom physique (PFN) :
    protocole://serveur/ec14d492-646a-11e5-9a08-13df08824a8a?ZIP=AliAOD.Muons.root
  - Exemple :

```
whereis /alice/data/2015/LHC15i/000236203/muon_calo_pass1/AOD/072/aod_archive.zip
Oct 23 08:46:02  info   The file data/2015/LHC15i/000236203/muon_calo_pass1/AOD/072/aod_archive.zip is in
            SE => ALICE::GSI::SE2  pfn
=>root://lxalird1.gsi.de:1094//07/07154/ec14d492-646a-11e5-9a08-13df08824a8a

            SE => ALICE::Catania::SE  pfn
=>root://prod-se-01.ct.infn.it:1094//07/07154/ec14d492-646a-11e5-9a08-13df08824a8a

            SE => ALICE::CERN::EOS  pfn
=>root://eosalice.cern.ch:1094//07/07154/ec14d492-646a-11e5-9a08-13df08824a8a
```

Element de stockage type



Accès en parallèle sur les serveurs disque

Principales technologies

- Tier-0 (CERN)
  - Castor (pour les bandes)
  - EOS (xrootd)
- Tier-1
  - EOS (xrootd)
  - dCache
- Tier-2
  - Disk Pool Manager (DPM)
  - Xrootd natif (sans EOS)
  - EOS (xrootd)

Briques de base stockage

- Noeud « redirecteur » : 1 serveur Linux standard
- Serveurs « Direct Attached Storage » avec 12 à 16 disques en RAID6, 1 volume RAID, 2 partitions (12 à 20 To)
- Scientific Linux 6 x86_64
- Filesystems : ext4 et maintenant xfs
- Double alimentation, partition système en RAID1
- Cinq ans de garantie sur site
- Disques qualité « Enterprise »

Le stockage actuel @Subatech

Round-Robin DNS Alias

EOS Manager 1 (master)

EOS Manager 2 (slave)

2 Serveurs Dell PE1950
8 cores Xeon E5420
16GB RAM
Eth 1Gbits/s
Double alim / RAID1

EOS server
EOS server
EOS server
EOS server
EOS server
EOS server
EOS server
EOS server
EOS server
EOS server
EOS server
EOS server

[...]

Serveurs disque
Managers

Total net : 435To/396Tio

2 Serveurs Alineos
8 cores Xeon E5620
16GB RAM
Eth 2x1Gbits/s (bond)
Double alim
14 disques 2To/RAID6

6 Serveurs Dell R720xd
4 cores Xeon E5-2603
16GB RAM
Eth 2x1Gbits/s (bond)
Double alim
12 disques 3To/RAID6

5 Serveurs Alineos
6 cores Xeon E5-2603
16GB RAM
Eth 2x1Gbits/s (bond)
Double alim
16 disques 3To/RAID6

Qualité de service

- L'expérience définit le nombre de réplicats de chaque fichier (entre 1 et 3). Il n'est pas demandé de sauvegarde aux sites Tier2.
- Malgré cela, l'indisponibilité d'un élément de stockage pénalise toutes les activités de calcul sur la grille.
- La redondance permet de pallier à un certain nombre de pannes :
  - Panne de disques (RAID6 : jusqu'à deux simultanément)
  - Panne d'alimentation
  - Panne d'un redirecteur (EOS dual manager configuration)
- Il reste quelques SPOF : panne d'un serveur, panne d'un équipement réseau central (difficile à redonder)

Disponibilité

Host 'naneosmgr01'

01-01-2015 00:00:00 to 10-14-2015 16:19:25
Duration: 286d 15h 19m 25s

[Availability report completed in 0 min 5 sec]

Host State Breakdowns:

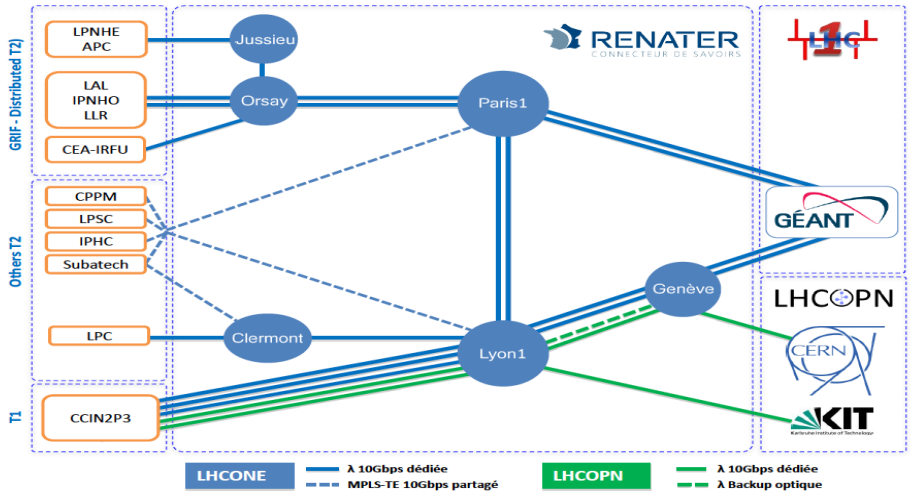
State	Type / Reason	Time	% Total Time	% Known Time
UP	Unscheduled	283d 21h 24m 10s	99.042%	99.042%
	Scheduled	0d 0h 0m 0s	0.000%	0.000%
	Total	**283d 21h 24m 10s**	**99.042%**	**99.042%**
DOWN	Unscheduled	2d 17h 55m 15s	0.958%	0.958%
	Scheduled	0d 0h 0m 0s	0.000%	0.000%
	Total	**2d 17h 55m 15s**	**0.958%**	**0.958%**
UNREACHABLE	Unscheduled	0d 0h 0m 0s	0.000%	0.000%
	Scheduled	0d 0h 0m 0s	0.000%	0.000%
	Total	**0d 0h 0m 0s**	**0.000%**	**0.000%**
Undetermined	Nagios Not Running	0d 0h 0m 0s	0.000%	
	Insufficient Data	0d 0h 0m 0s	0.000%	
	Total	0d 0h 0m 0s	0.000%	
All	Total	286d 15h 19m 25s	100.000%	100.000%

Taux de disponibilité 2015 > 99 %

Il suffit d'un arrêt électrique pendant un WE (66h) pour que la disponibilité passe de 100 % à 99,25 %

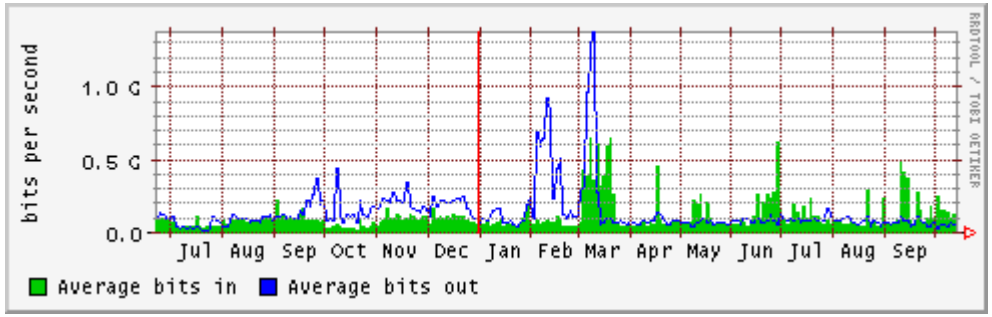
L'importance du réseau

- Grâce à la qualité du réseau, le modèle hierarchique original tend à disparaître
- Il existe un réseau de fibres dédiées entre le Tier0 et les Tier1 : le réseau LHC/OPN
- Un autre réseau de liens virtuels dédiés existe entre les autres sites : LHC/ONE

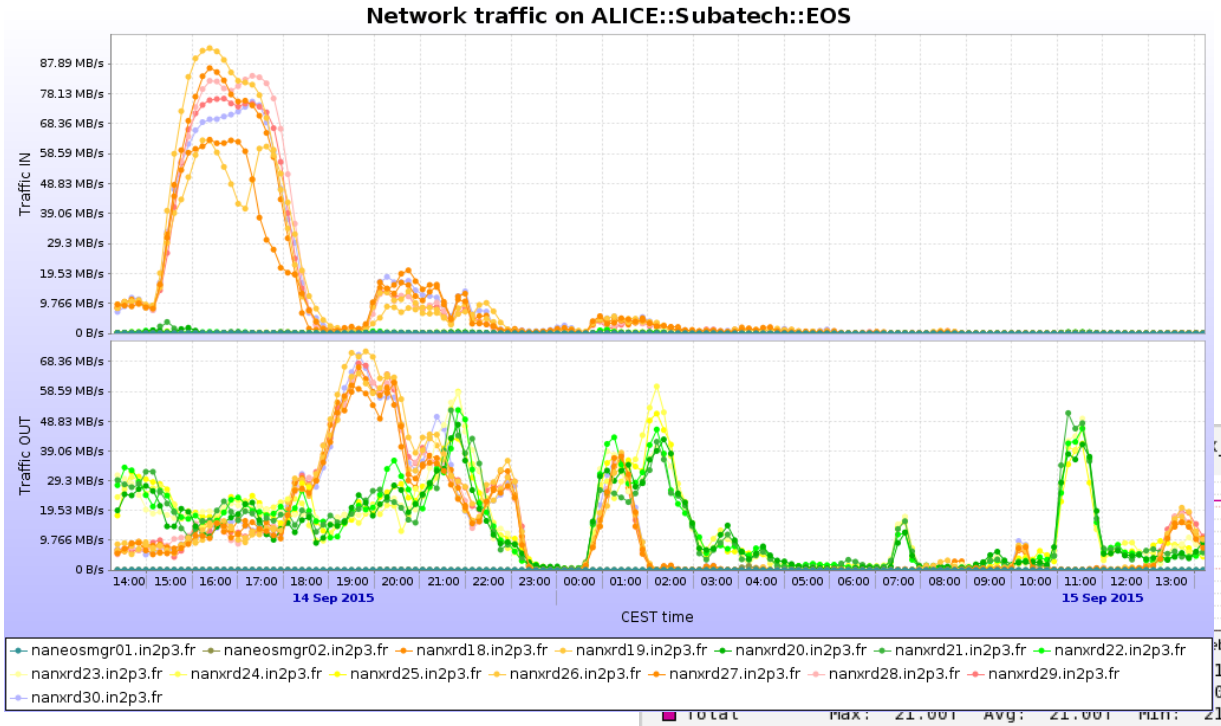


Le réseau LHC/ONE en France

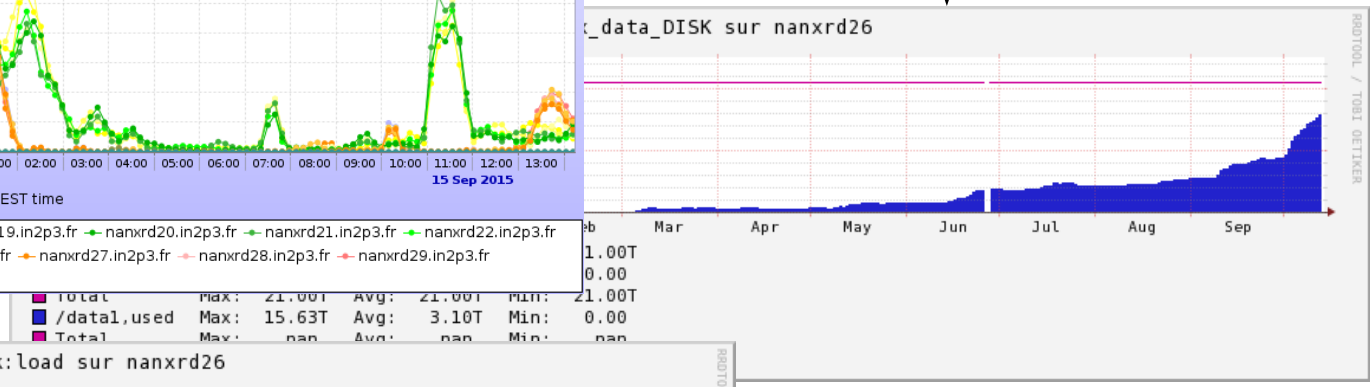
La bande passante 10Gbits/s du réseau « Tier2 » du laboratoire Subatech vers Renater a déjà été utilisée à plus de 30 % soit ~3Gbits/s.

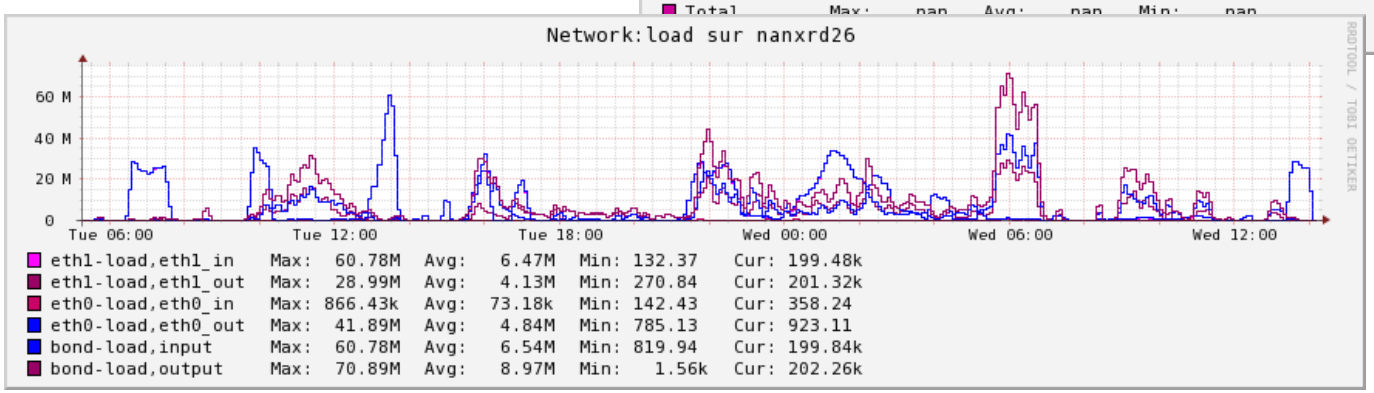


Quelques vues



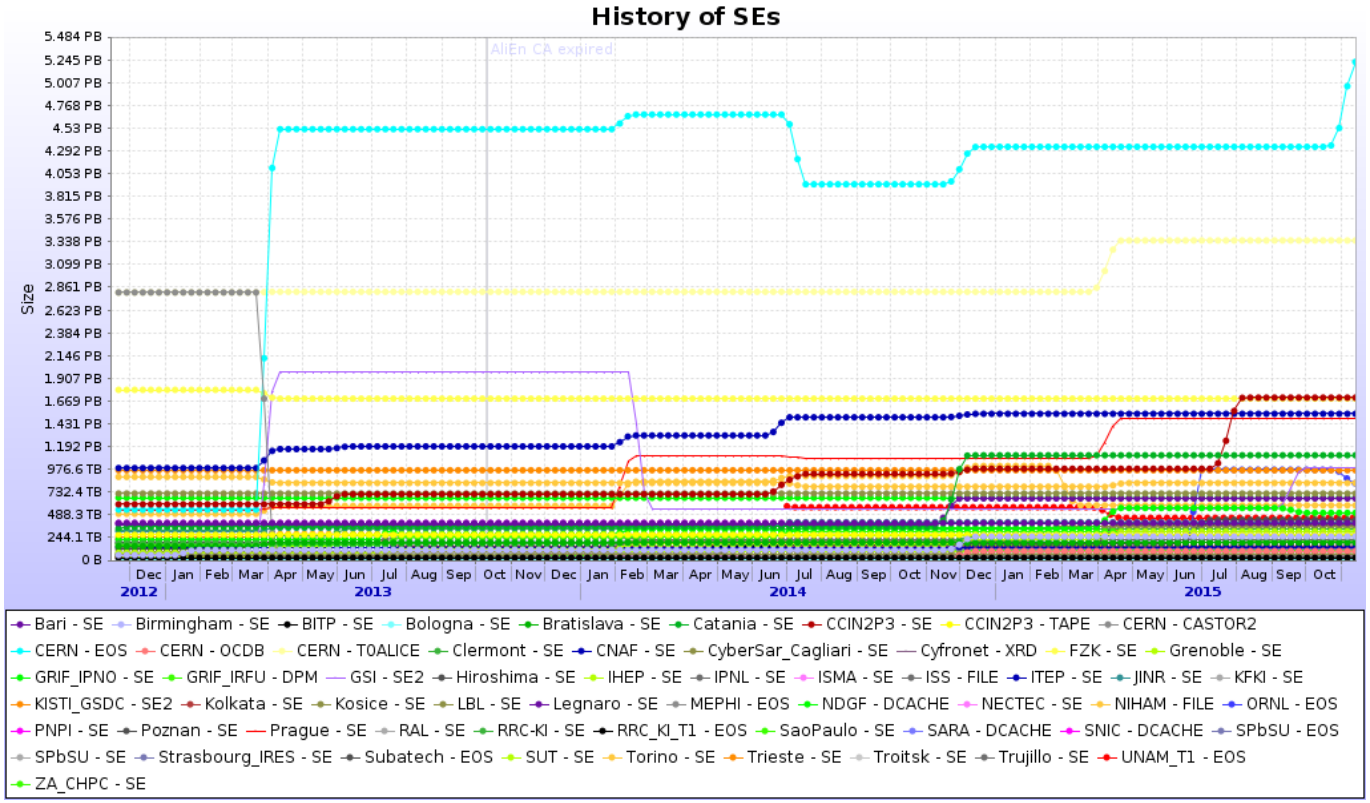
Nouveaux serveurs installés en mars 2015





Données Alice

- Plus de 50 éléments de stockage
- Total : environ 50 Petaoctets (bandes incluses)
- Taux de remplissage : 75 %



Conclusion

- La gestion des données du LHC a déjà subi pas mal d'évolutions depuis le début (de SRM vers xrootd/EOS et la fédération des clusters de stockage), ceci pour les 4 expériences.
- La qualité du réseau a été meilleure que ce qui avait été anticipé, menant à la dissolution progressive du modèle très hiérarchisé du début.
- Il faut s'attendre à d'autres évolutions afin de s'attaquer au défi que représentent les volumes de données des prochaines campagnes du LHC (RUNS 3 et 4).

Liens et références

- [1] LHC : https://en.wikipedia.org/wiki/Large_Hadron_Collider
- [2] LCG : http://wlcg.web.cern.ch/
- [3] LCG-France : http://lcg.in2p3.fr
- [4] Laboratoire SUBATECH : http://www-subatech.in2p3.fr
- [5] Journées SUCCES 2015 : : http://succes2015.sciencesconf.org/
- [6] The data management lifecycle at the Large Hadron Collider, Andrew Washbrook http://datablog.is.ed.ac.uk/files/2015/09/ajw-DWD-final.pdf
- [7] Data Handling for LHC: Plans and Reality, Tony Cass http://web.stanford.edu/group/mmds/slides2012/s-cass.pdf
- [8] Introduction to Big Data http://www.it.northwestern.edu/bin/docs/TT_BigData.pdf