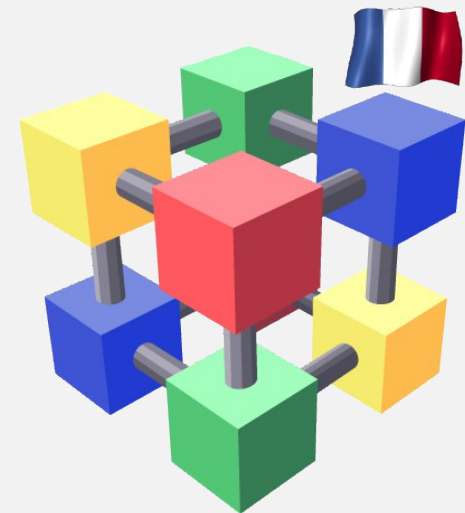


In2p3

Session réseau

Besoin des expériences au Run 3

C. Biscarat (biscarat@lpsc.in2p3.fr)



Journées LCG-France, 27-29 mars, CC-IN2P3

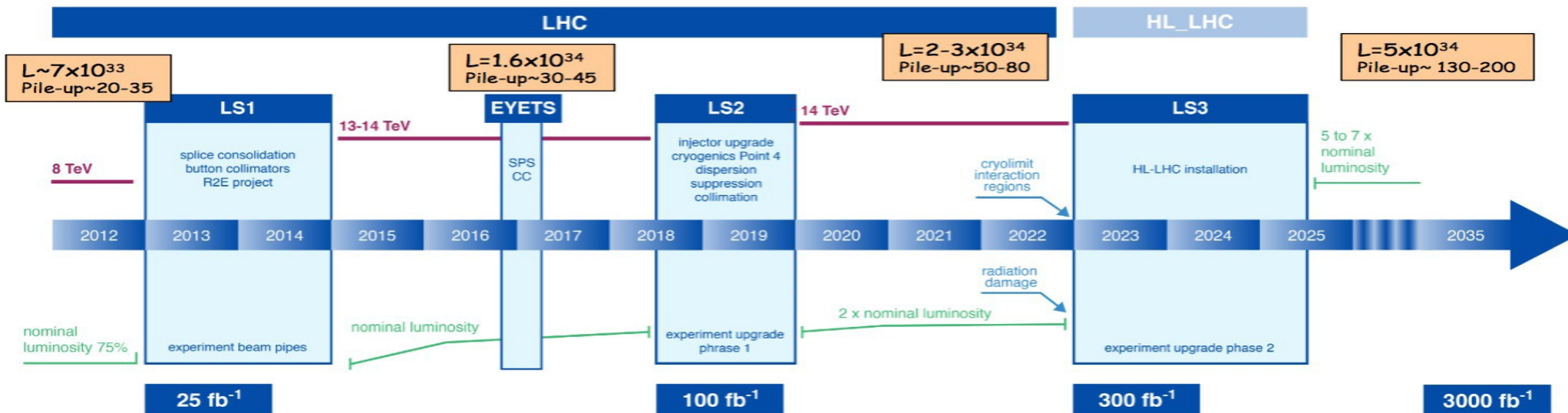
LCG France



LHC planning

New LHC / HL-LHC Plan

L.Rossi



Pile-up : complexité des événements
Lumi. intégrée : nb d'événements produits

Impacts sur le computing

- CPU : augmentation du CPU/event, augmentation de la mémoire / événement
- Stockage : des événements individuels plus gros, plus d'événements



Upgrades des expériences

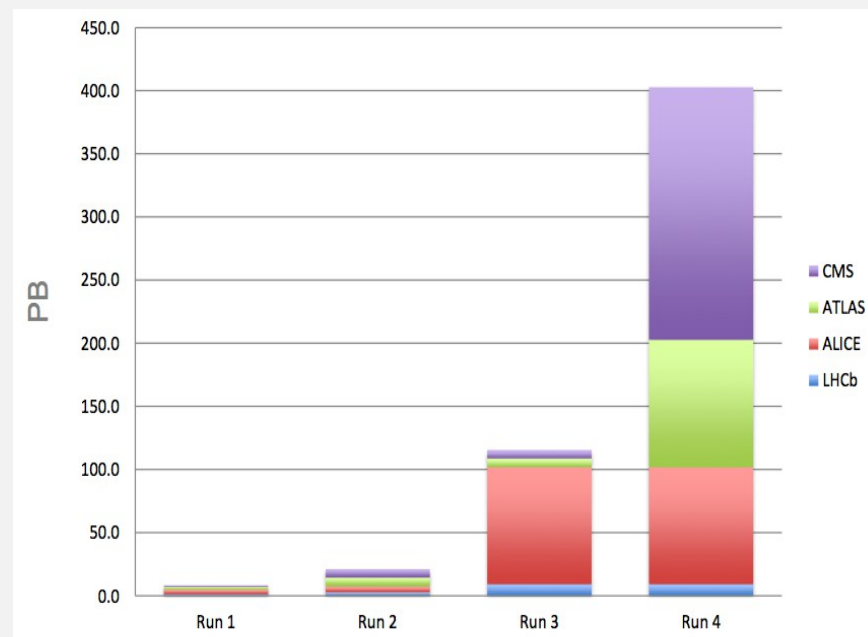
- ATLAS et CMS : Run 4
- ALICE et LHCb : Run 3
 - HLT (kHz) ATLAS, CMS, ALICE, LHCb : 1, 1, 50, 30 000

Déjà des optimisations substantielles

- Modèle des expériences (moins de copies)
- Software (améliorations constantes, simulation rapide)
- Stockage (campagne de déletion, formats d'analyse réduits)
- Utilisation plus agressive du réseau

Changements dans les modèles Run 3

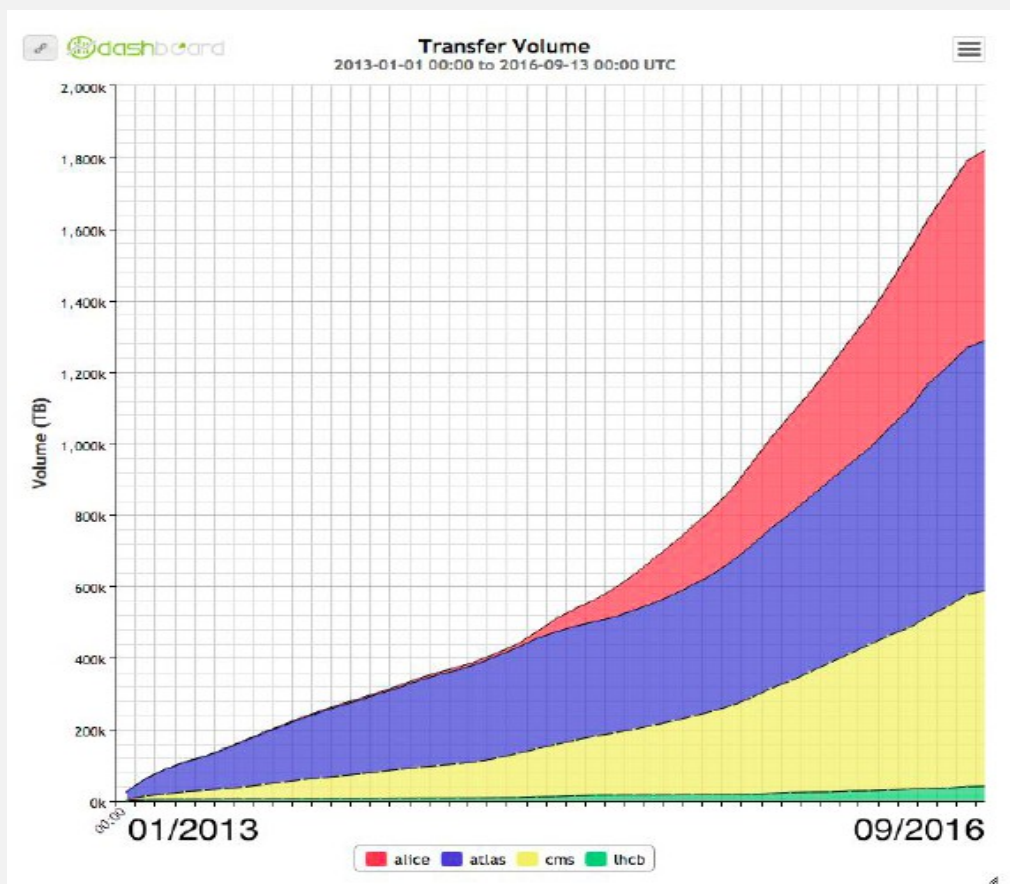
- Calibration et reconstruction en ligne (ALICE et LHCb)
- Rentrer dans des budgets plats



Taille des lots de données brutes
Exatrpolation des modèles en cours
ECFA High Lumi. LHC experiments workshop (2013)



Des augmentations continues



Des transferts de données en constante augmentation depuis le début du LHC.

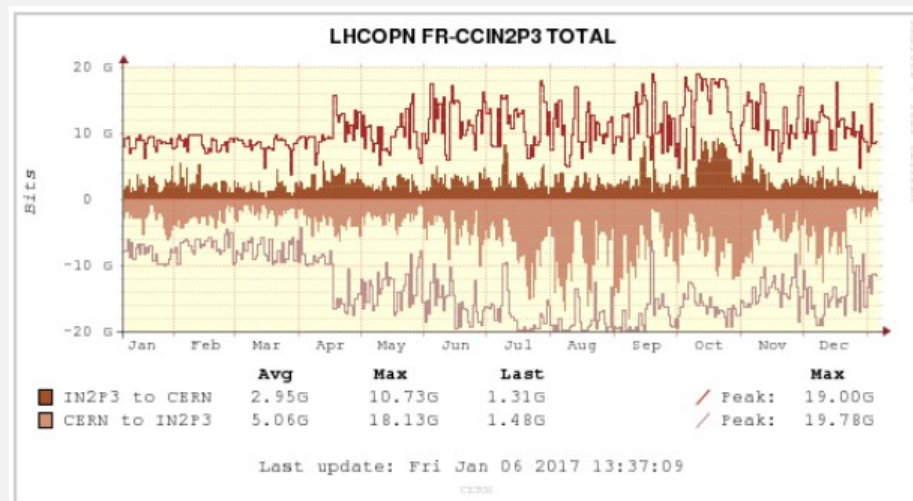
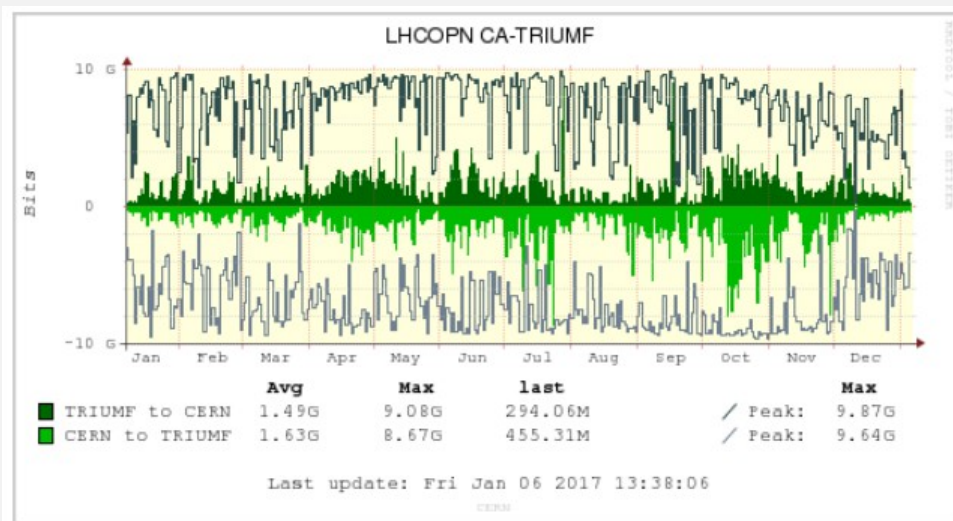
Drivé par le volume de données accumulé



ATLAS - présent

Run 2

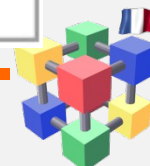
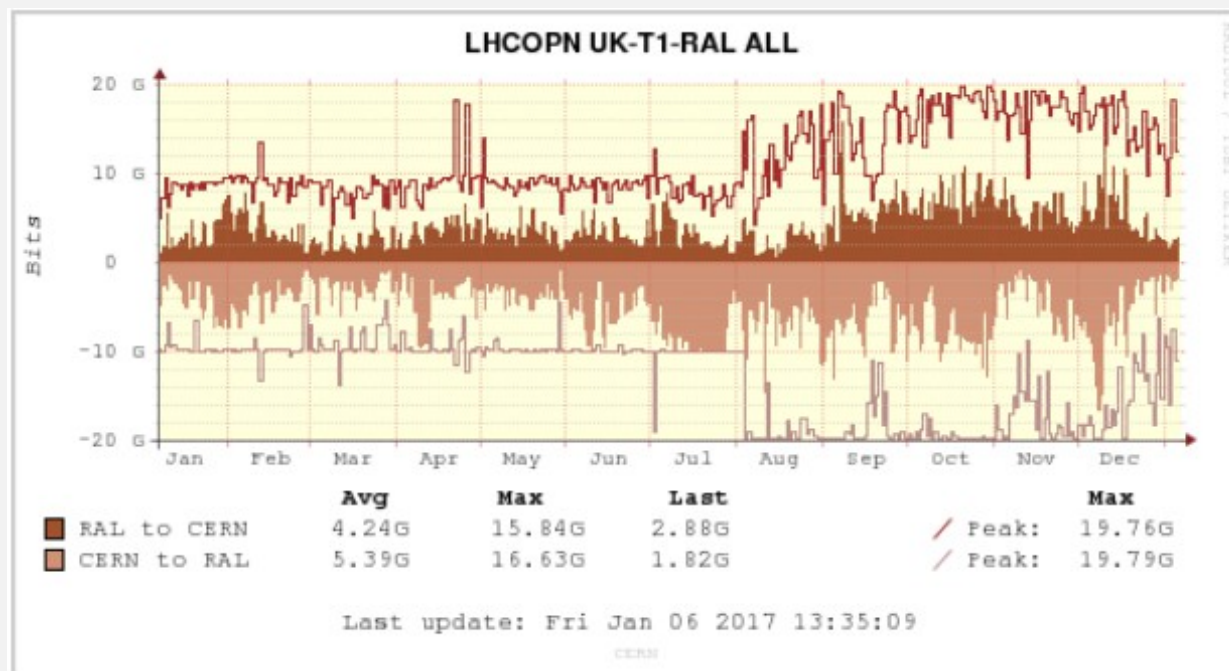
- Tous les transferts de ATLAS : ~stable à 20 GB/s avec des pics à 50 GB/s
- La contribution majeure vient des dérivations (pas des mouvements des données brutes)
- ATLAS ne pense pas que cela change bcp sur le long du Run2
 - Bémol : plus d'accès distants avec la mise en place de sites « storage-less »
- Des saturations observées dans plusieurs liens T0->T1 (10 Gbps) : PIC, Taiwan, Triumph



ATLAS – saturations et impact sur les opérations

RAL

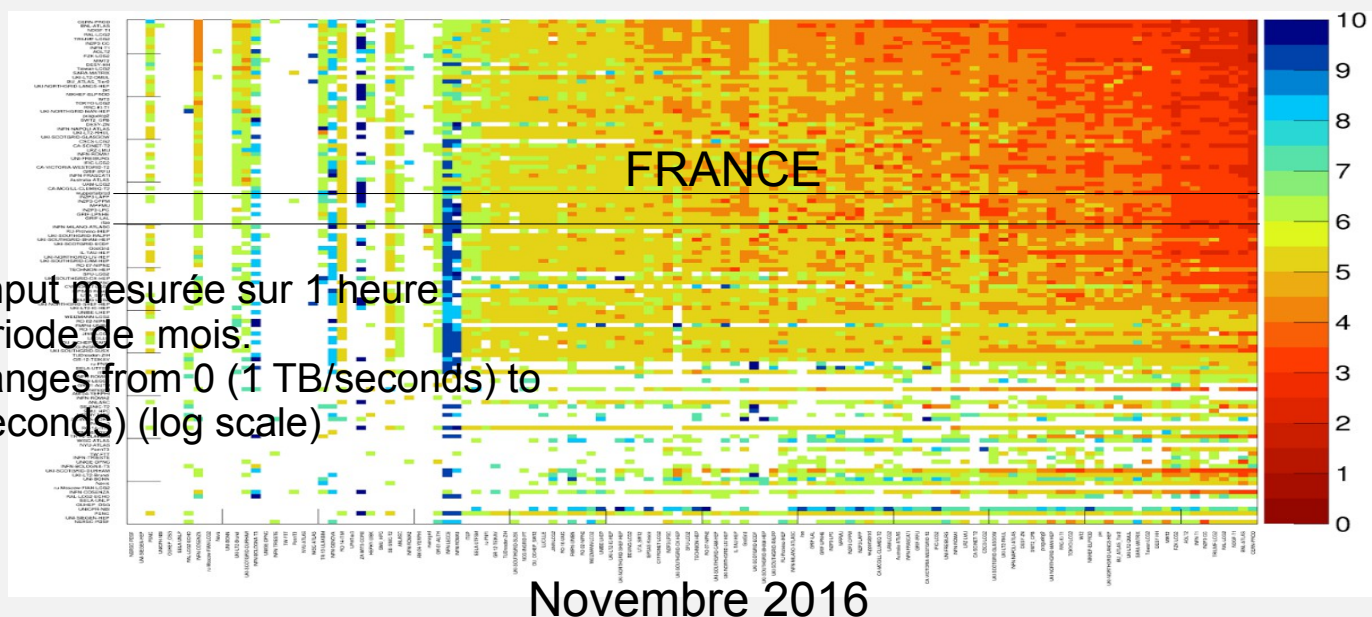
- Saturation du lien en juin 2016
- Le LHC a bien marché, bcp de transferts de données aux T1
- CERN->RAL, 6-7 jours de backlog
- Les données ont été distribuées sur d'autres sites



Le réseau est un atout majeur de nos sites

- Les jobs de ATLAS ne sont pas destinées à des régions particulières
 - Full mesh entre les sites
- Distribution dynamique des données et des jobs basée sur le réseau
 - Perfsonar (latence et bandwidth), throughput des dernières heures/jours
 - Création d'une matrice de « closeness » pour optimiser le job brokering

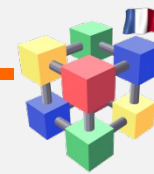
Max. throughput mesurée sur 1 heure
dans une période de mois.
Matrix that ranges from 0 (1 TB/seconds)
to 10 (0.1 kB/seconds) (log scale)



▲ 10: 0,1 kB/s

0: 1TB/s

Novembre 2016



- ATLAS attend :
 - Plus d'événements intéressants pour la physique
 - Plus d'événements à traiter, soit plus de CPU à servir
 - Une évolution des modèles de calcul
 - Moins de répliques (contraintes de budget)
 - Plus de mouvement de données
 - Plus grande exploitation du réseau WAN
 - Accès distant aux données (actuellement 10-20%)
 - T2 sans disk, sites nucleus (serveur de données)
 - Utilisation plus grande des tapes pour stockage et re-lecture (utilisation WAN x2)
- Résultat pour les besoins en réseau :
 - Un facteur entre 5 et 10
 - les sites les plus gros en stockage devraient être à 100 Gbps.

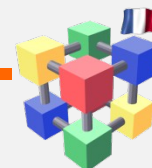


Sa philosophie aujourd'hui

- Utiliser le réseau comme une **ressource infinie**
- Transferts dynamiques de données qui reposent sur un réseau fiable et consistant
- Pas de distinction en terme de trafic réseau entre $T0 \leftrightarrow T1, T1 \leftrightarrow T2, T2 \leftrightarrow T2$

Fédération AAA

- Ingrédient majeur du computing model de CMS
 - De plus en plus utilisée
 - AAA=storage-to-WN, GridFTP=storage-to-storage
- Données lues sur le WAN par « overflow de jobs » :
 - les jobs sont envoyés ailleurs, les data lu sur le WAN
 - Actuellement fait dans des régions (au niveau de 20%), bientôt entre US et EU (à 50%)
 - CMS choisit les sites en fonction de leur connexion réseau
- Les sites doivent avoir 10 Gbps actuellement



Fédérations AAA

- Evolutions run3
 - Pledges des T1 < demandes de CMS : il faut réviser le modèle
 - Nouveau groupe de travail ECOM2017
 - Effet sur le mouvement des données à évaluer
- Conclusion pour CMS
 - CMS fait plus de transferts que ATLAS car ils ont moins de ressources CPU/disk/tape
 - On pourrait imaginer que l'augmentation de la demande atteigne x5-10 pour le réseau
- CMS voudrait explorer de nouvelles fonctionnalités réseau
 - Workload management system < - > réseau



Actuellement

- Utilisation beaucoup plus modeste du réseau que ATLAS ou CMS
- Presque exclusivement des sites Européens
- Tier-1 : total=500 MB/s (mostly inbound traffic via LHCOPN), moyenne/site=70 MB/s
- Tier-2 : total=140 MB/s (mostly outbound traffic via LHCONE), Inbound throughput for T2-Storage : 50 MB/s

Au run 3

- Les demandes en terme de réseau seront toujours plus modestes que ATLAS ou CMS
- LHCb prévoit une augmentation des trafics d'un facteur ~ 10 .
- Sans changement dans la distribution des T0/T1/T2, avec peu (< 20) T2 avec du disk

Network monitoring

- LHCb intègre des nouvelles fonctions de monitoring réseau dans DIRAC (Perfsonar)
 - Visualisation & corrélation avec les opérations de LHCb
- But : voir si un problème vient du réseau ou non & optimisation des transferts de données



ALICE - actuellement

- LHCOPN :
 - RAW data T0->T1 (90% EU, 10% Asie) : 250 (190) MB/s en moyenne en 2015 (2016)
 - Pas de changement pour le reste du Run2
- Distribution et accès aux données
 - Automatique, basée sur la topologie du réseau (mesurée dans les Vobox)
- Accès aux données
 - Le job va aux données
 - Faible taux d'accès remote
 - > 98% des données lues depuis le site même

2016 Avg	Written		Read	
Total volume	46PB		466PB	
Local	50%	0.7GB/s	98%	15.2GB/s
Remote	50%	0.7GB/s	2%	0.3GB/s

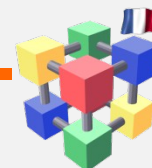


Des changements majeurs dans le modèle de computing

- Nouvelle ferme O2 pour la reco « on-line »
- Nouvelles Analysis Facility (5-10 PB 20-30 k cores) – projet non concrétisé
- Pendant le Run 3, croissance de 20%/an des sites (balance CPU/disk non figée)
- Activités des Tiers
 - O2/T0/T1 : reco et calibration ;
 - T2 & HPC : Monte Carlo jobs
- Organisation en régions, ~10 (simplifier les op. ALICE et raccourcir les chemins réseau)
 - Les données resteraient dans ces « régions », T0->T1 (LHCOPN), T1-> T2 associés
 - 1/3 des données seraient copiées sur les T1
- Peu d'accès remote (2-3%)

Besoins réseau

- Changement d'échelle dans les données enregistrées
- Mais compensée par le nouveau modèle O2
- ALICE reste à une fraction constante de ATLAS/CMS



Pour le Run 3 (2020)

- ATLAS : x 5-10
- CMS : x 5-10
- LHCb x 10 (reste petit consommateur)
- ALICE : reste fans les même proportions que maintenant

En France :

- On peut imaginer pour les plus gros sites FR de devoir passer à 100 Gbps pour le Run 3
- Adéquation entre
 - Utilisation des réseaux par les sites
 - Backbone RENATER
 - Branches régionales
 - Boucles locales (last-mile)
 - Router d'entrée des sites
 - Coeur de réseau des sites, leurs évolutions
- Augmentations des besoins continues : évolutions intermédiaires

