Modèle de calcul de CMS et Support au CC-IN2P3 T1

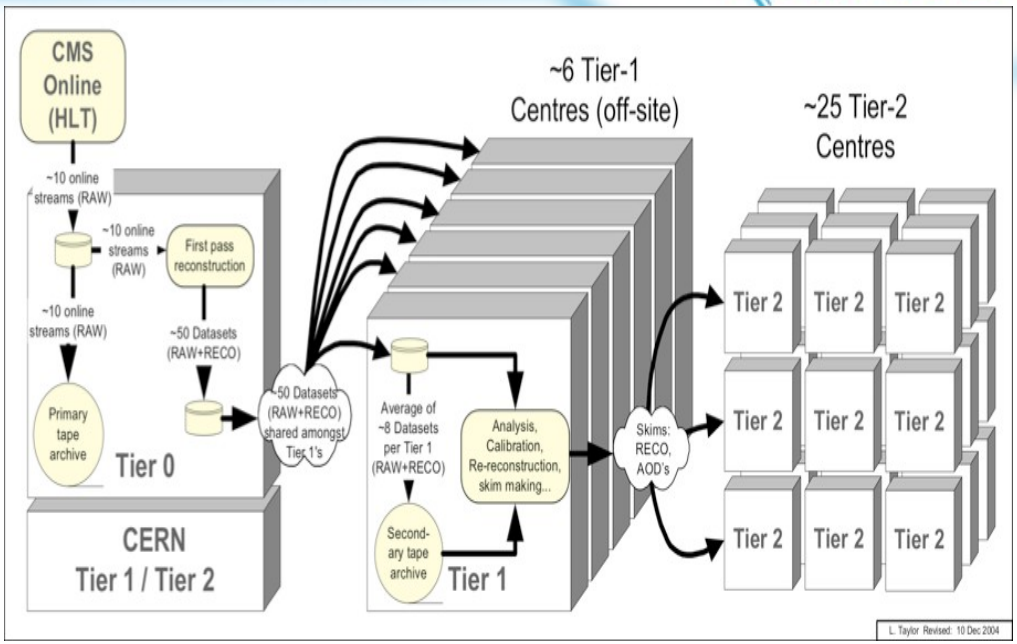
Rencontre LCG-France, 19 septembre 2012

PLAN

- ✓ Présentation du modèle de calcul de CMS
- ✓ Actualités et évolution
- ✓ Activités et support au T1 CC-IN2P3
- ✓ Conclusion

Vue d'ensemble

- Modèle distribué
  - utilisation de toutes les ressources disponibles
  - hiérarchie de sites et d'activités
  - grille comme technologie sous-jacente
  - actuellement 7 Tier1s, 53 Tier2s
- Flux de données considérable
  - taux de déclenchement: 300 Hz
  - taille événement brut: 1.5MB*
  - 0(5PB) / an de données brutes*
- Organisation des données en vue de la physique
  - « data streams » constitués à partir des bits de déclenchement HLT
  - distribution et reprocessing par data stream
  - prioritisation par la physique

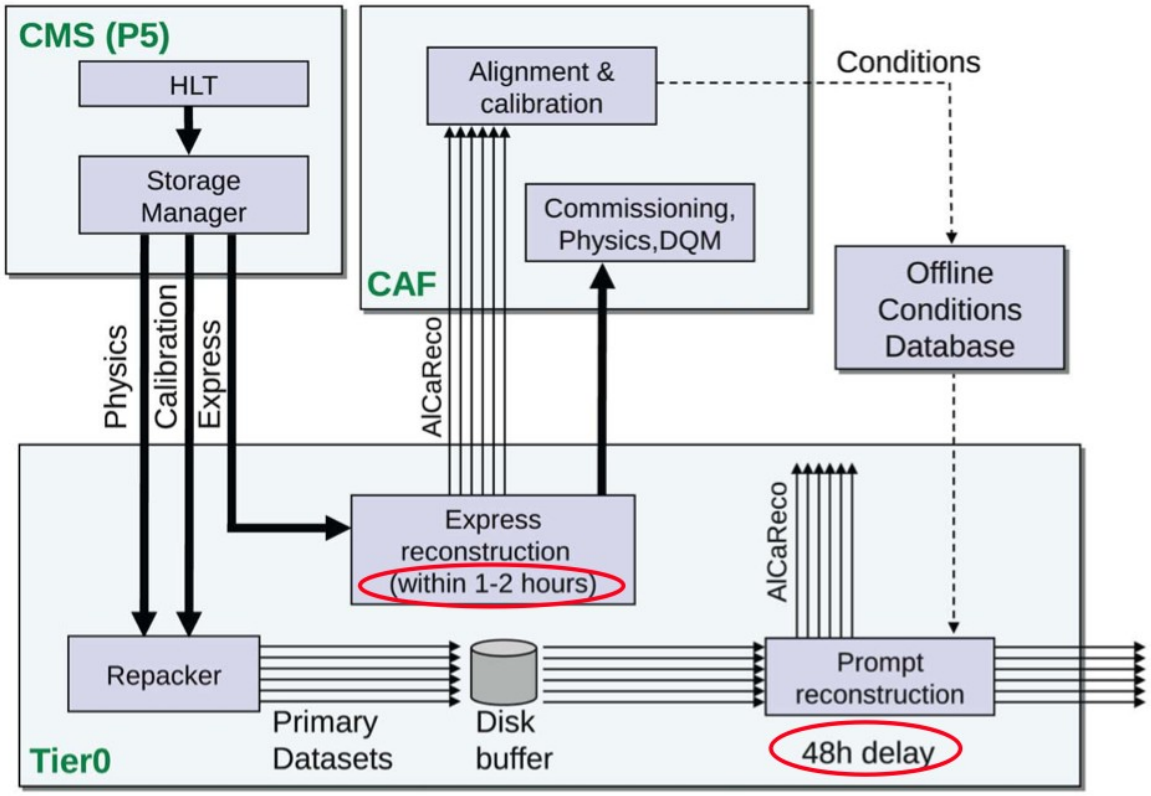


- Software flexible et efficace
  - nombreux reprocessings
  - 1-10 sec/event
- Évolution rapide
  - optimisation (mémoire en particulier)
  - prise en compte des upgrades à venir

* valeurs nominales

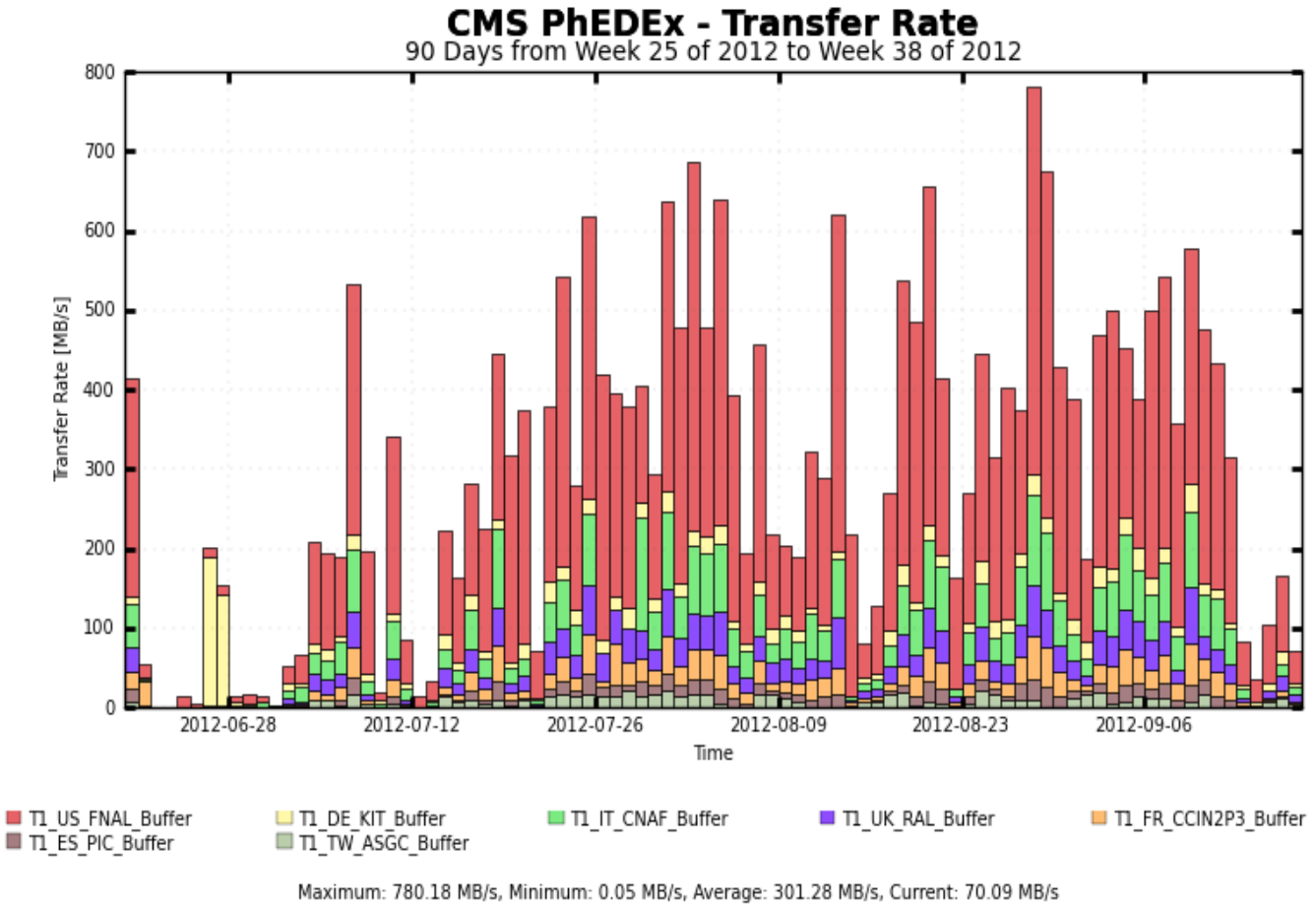
Activités au CERN T0

- Repacker
  - formate les données au format ROOT
  - séparation en datasets suivant les bits du HLT
- Express reconstruction
  - ~10% des données
  - latence ~1-2h
- Prompt reconstruction
  - latence ~48h
  - très stable, peu de crash
- Taille des événements
  - RAW: 0.225MB
  - RECO: 0.4MB



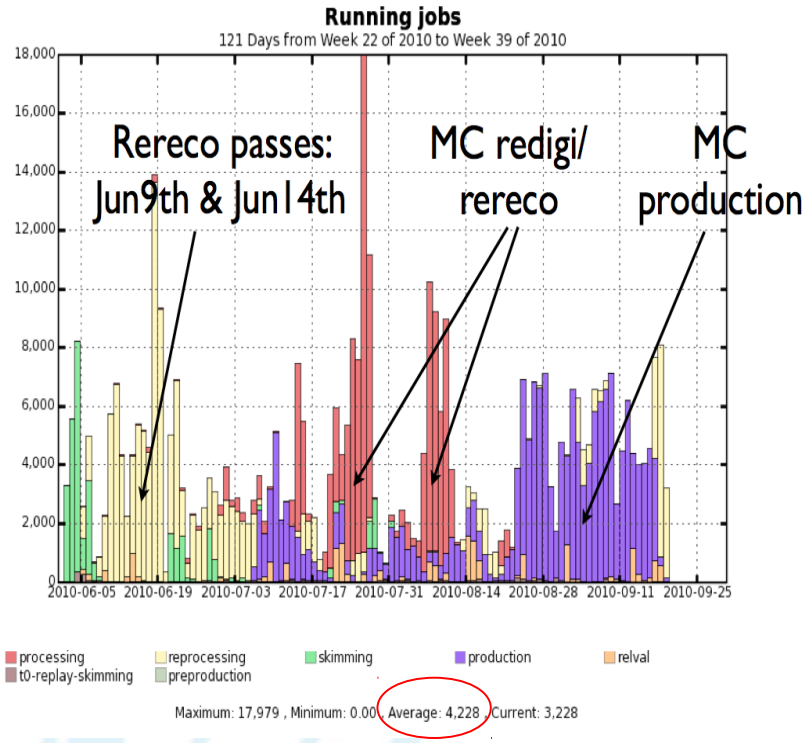
Transferts du T0 vers les T1s

- ✓ Les transferts fiables depuis le CERN
- ✓ Jusqu'à 1 GB/s en export depuis le CERN (valeur maximale prévue dans le modèle)



Activités aux T1s

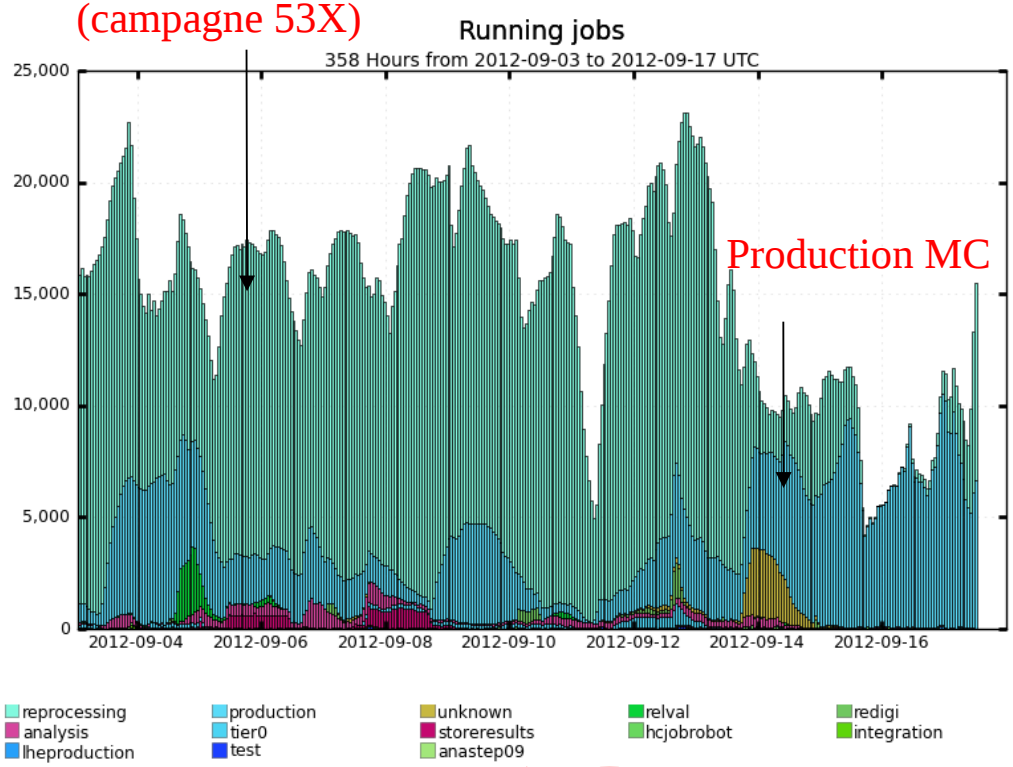
Été 2010



~5000 jobs/jour

Septembre 2012

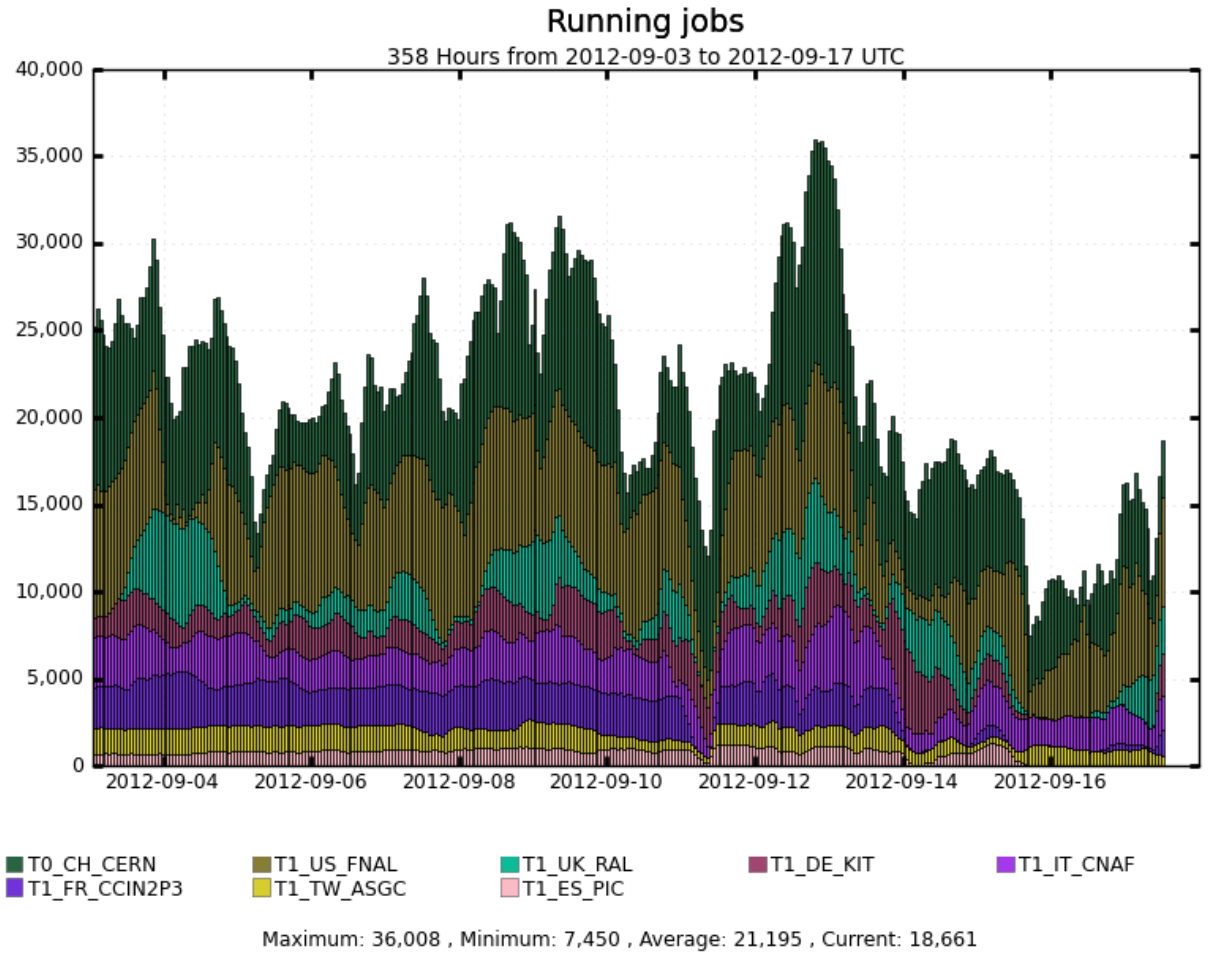
Re-reconstruction (campagne 53X)



~15000 jobs/heure

Activités aux T1s

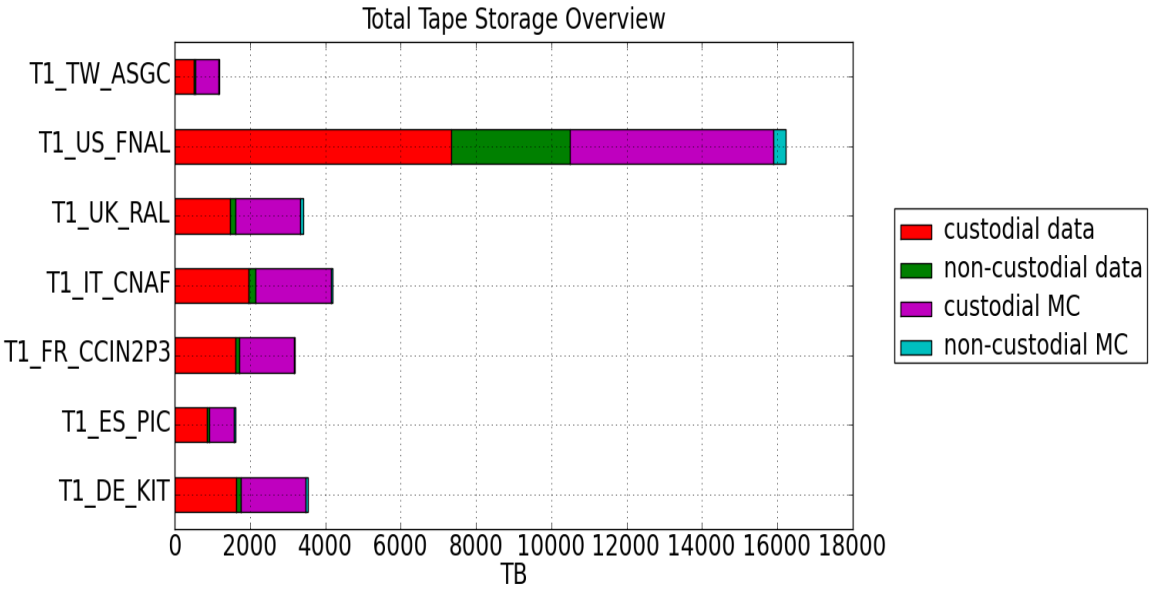
- Évolution de l'utilisation des sites :
  - Modéré (faible) jusqu'à mi 2011 (*~5000 jobs/j en mi 2010*)
  - Grosse augmentation fin 2011
  - Légère augmentation en 2012 (*~15000 jobs/h !*)
  - **Utilisation stable et continue !**
- On atteint les limites de la capacité de calcul
  - Moins de re-reco
  - Prioritisation de plus en plus forte des activités (MC, rereco)
  - Retard sur les re-reco...



Stockage des données

Capacités totales de stockage : 47 PB

- ✓ Bonne utilisation du stockage
- ✓ Campagnes d'effacement régulières : 1 toutes les 6 à 8 semaines (dépend du T1)



70 % 64 % 90 % 66 % 85 % 75 % 48 % 73 %

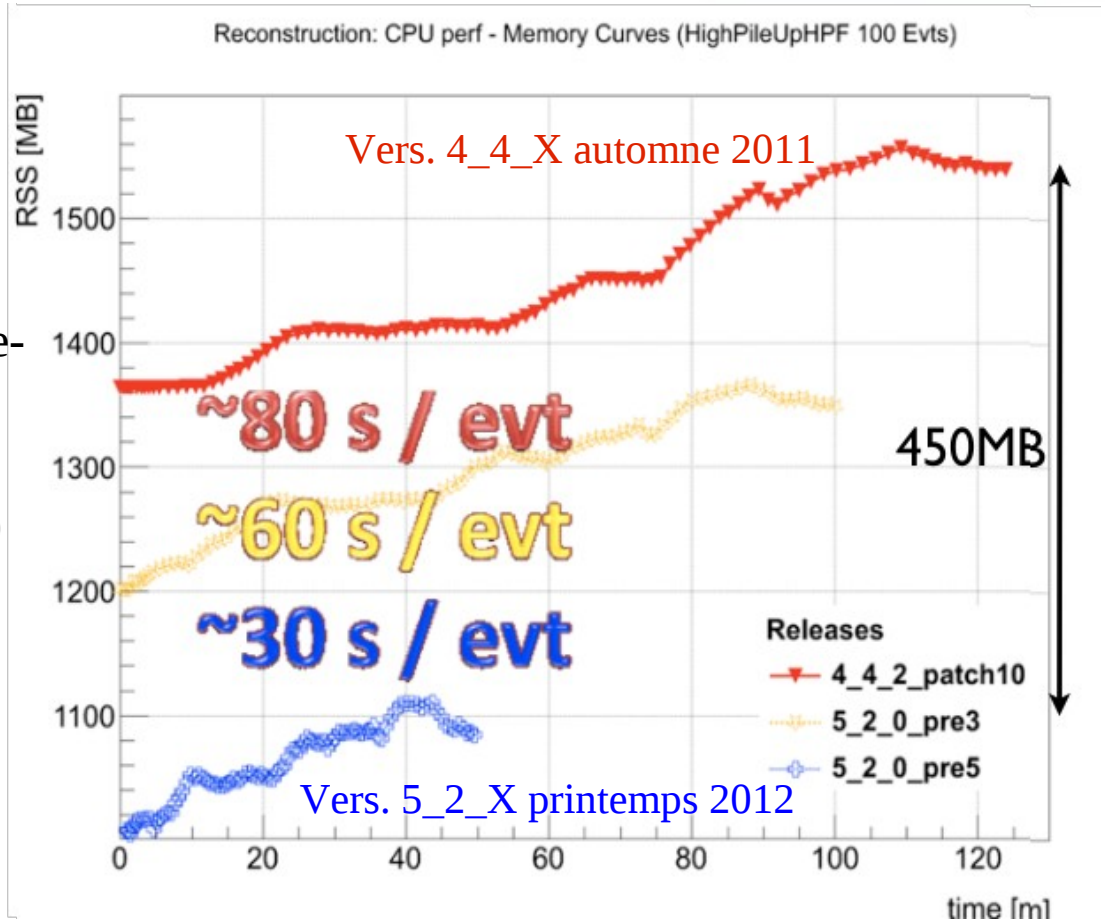
	T1_DE_KIT	T1_ES_PIC	T1_FR_CCIN2P3	T1_IT_CNAF	T1_UK_RAL	T1_US_FNAL	T1_TW_ASGC	All sites
Pledges [TB][*]	5100.000	2601.000	3600.000	6600.000	4080.000	22000.000	2550.000	46731.000
Used (PhEDEx)	3563.855	1651.391	3228.535	4331.748	3490.158	16573.894	1218.885	34058.466

- [*]: Pledges last updated 30th April 2012 from http://gstat-wlcg.cern.ch/apps/pledges/resources/

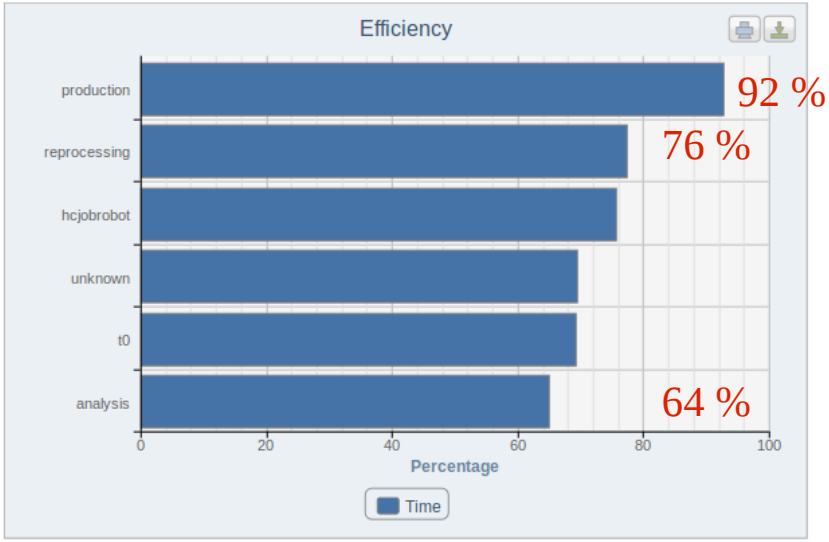
Situation au 17 septembre 2012

Évolution du logiciel

- ✓ Améliorations notables à chaque release :
  - ✓ Technique
  - ✓ Algorithmique
- ✓ Dév. en cours (vers. 6_X_Y) :
  - ✓ Support des upgrades à venir
  - ✓ Nouvelles conditions de run (pile-up, mixing)
  - ✓ Multicore & parallélisme (meilleur partage de la mémoire)
- ➢ Release 6_1_Y en décembre
  - ➢ *Possible rereco de 2011 et 2012*
- ✓ Futur : 7_X_Y (mi 2013?)
  - ✓ nouveau tracking (opt. Niveau CPU)
  - ✓ Multicore
  - ✓ Opt. simu. lente



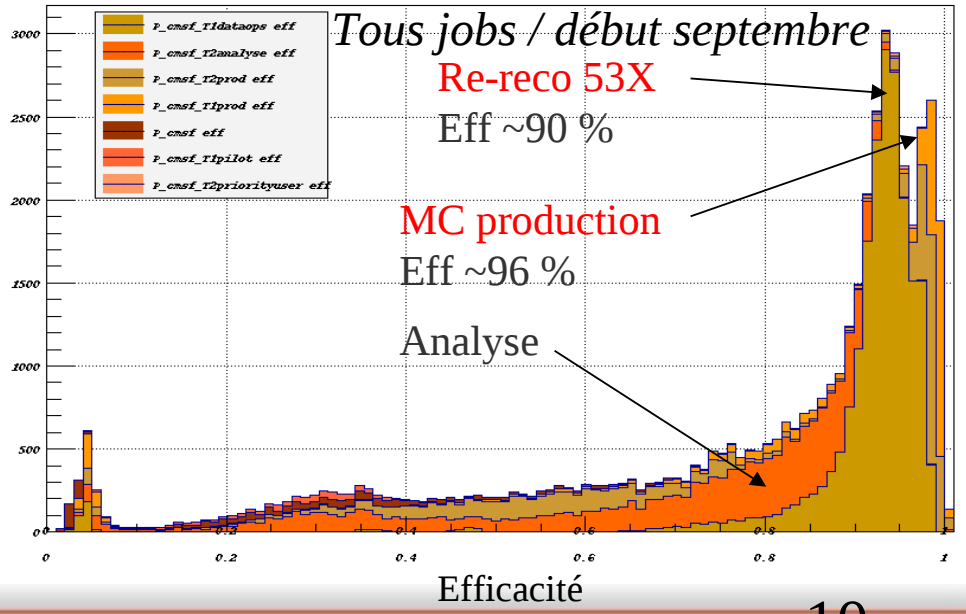
Efficacités des jobs



CMS dashboard

Efficiency	
activity	Time
production	92.94%
reprocessing	77.56%
hcjobrobot	75.78%
unknown	69.45%
t0	69.32%
analysis	65.08%
AVERAGE	75.02%

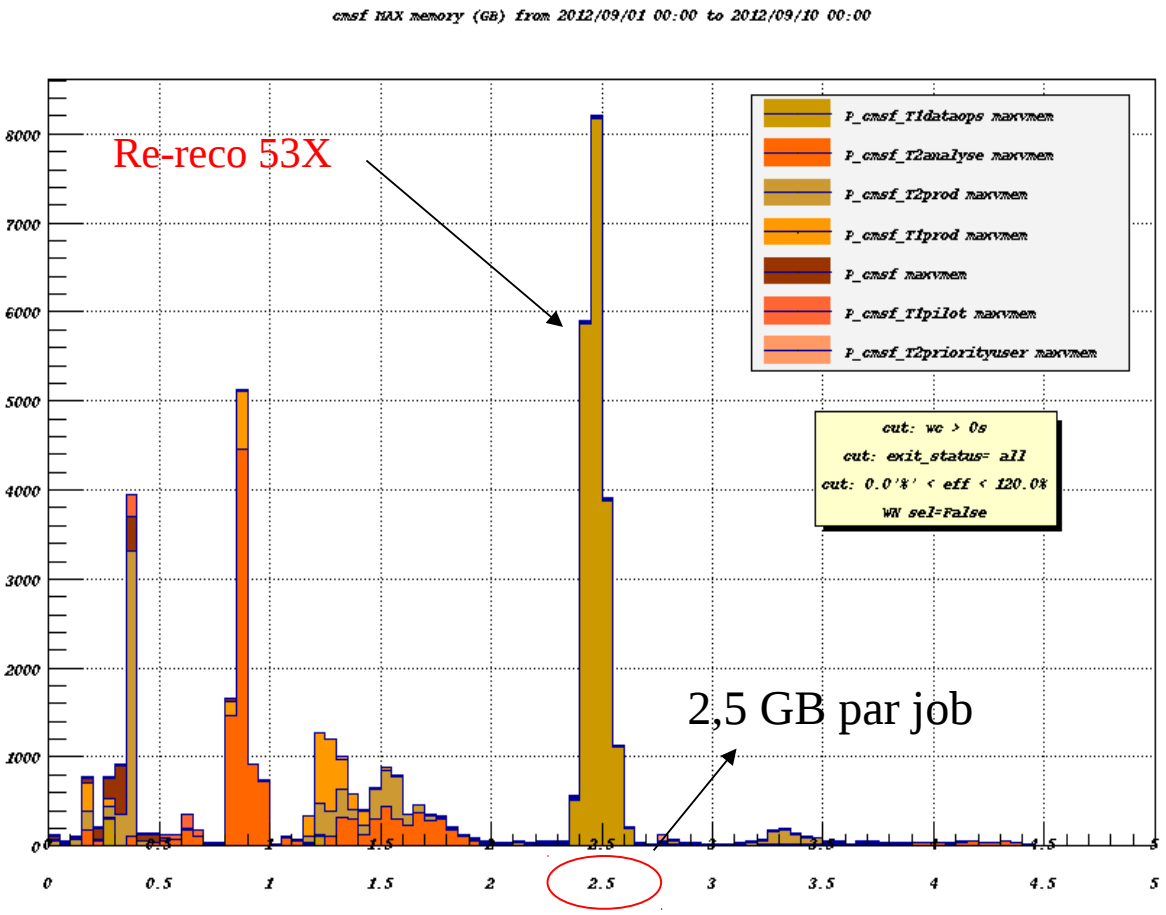
cmsf EFFICIENCY CPU/WC from 2012/09/01 00:00 to 2012/09/10 00:00



Bonne efficacité des jobs de CMS sur l'ensemble des sites.

- efficacité au-dessus de la moyenne au CC-IN2P3.

Utilisation de la mémoire



✓ La consommation de mémoire est exactement celle annoncée (CMSSW 5_3_Y)

Difficultés dues au stockage

Le stockage est un facteur très contraignant au CC-IN2P3, politique de gestion des données aménagée au CC-IN2P3 :

- Les données sur disque (dCache) restent 1 mois (sauf RAW, SIM et AOD) [au lieu des 6 mois recommandés par CMS]
- campagnes d'effacement plus régulières pour HPSS (campagne spécifique en cas de besoin).

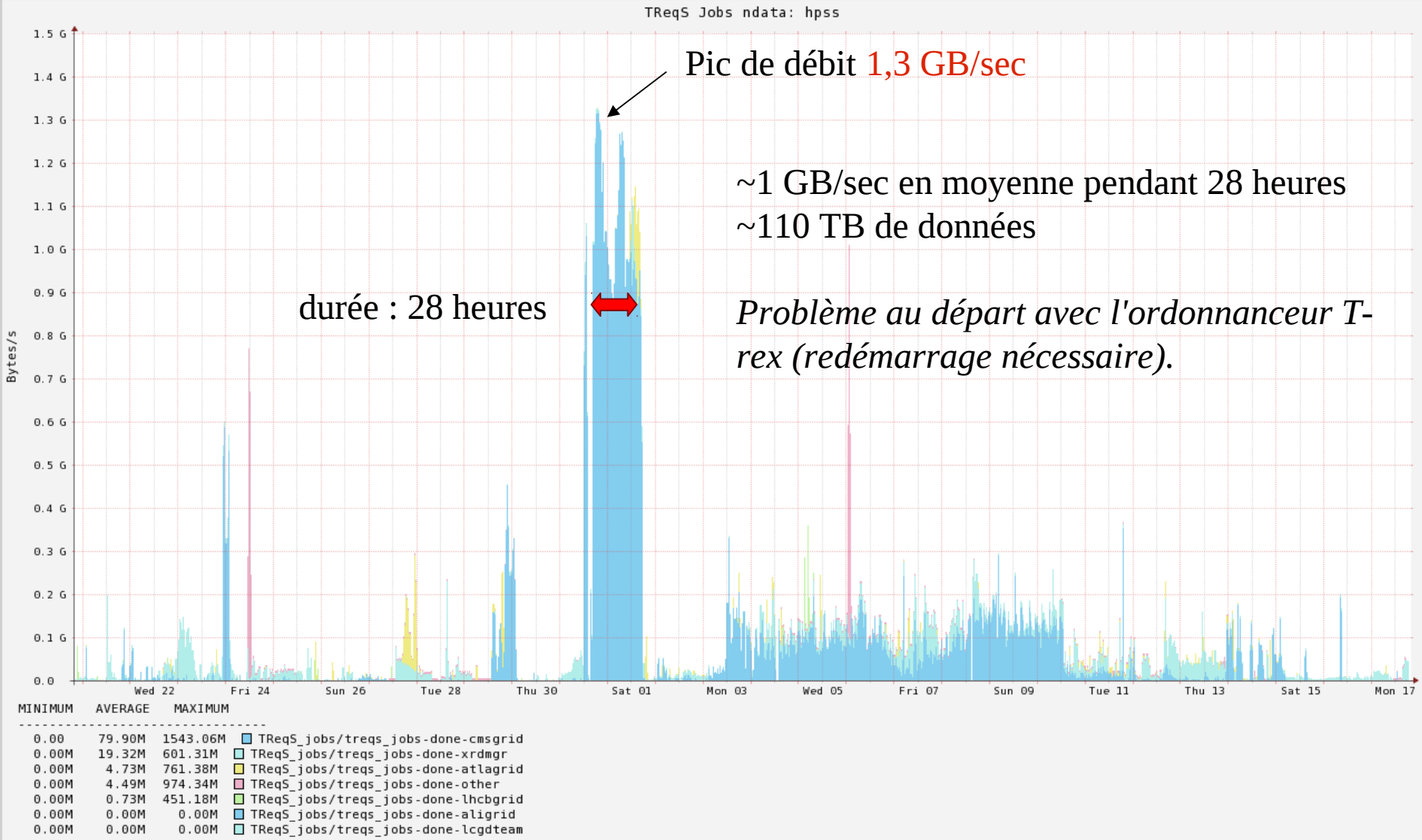
Néanmoins, du fait de la durée moyenne de « vie » des données (6 mois), cela pose des problèmes de temps en temps : « staging » très massif !

Exemple récent, vendredi 31 août, 15h, pour le besoin de la campagne DR53X :

- ~27 000 requêtes de « staging »
- > 100 TB de données

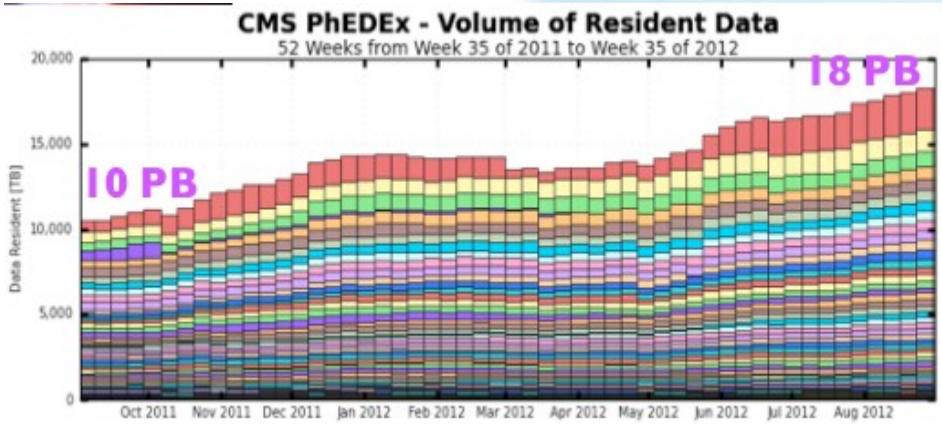
Notes : l'expert HPSS et l'astreint sont « moyennement » contents...

Staging massif



Distribution des données et analyse

- ✓ Évolution du modèle de distribution des données
  - ✓ Ajout d'un agent de nettoyage (T2)
  - ✓ Qui s'appuie sur un système « *Dynamic Data Placement* » afin d'automatiser le nettoyage.
- ✓ Le DDP devrait s'inspirer de celui d'ATLAS.
- ✓ Les données sont publiées, peuvent être utilisées, puis peuvent être effacées en fonction *de leur popularité.*
- ✓ Fédération des données, via un service *xrootd* qui permettra d'accéder de manière transparente aux données de sites distants.
- ✓ CRAB3 en phase d'intégration



✓ 18 TB de données réparties sur 53 T2s.
✓ transferts vers T2s, 3 fois le volume de données résident

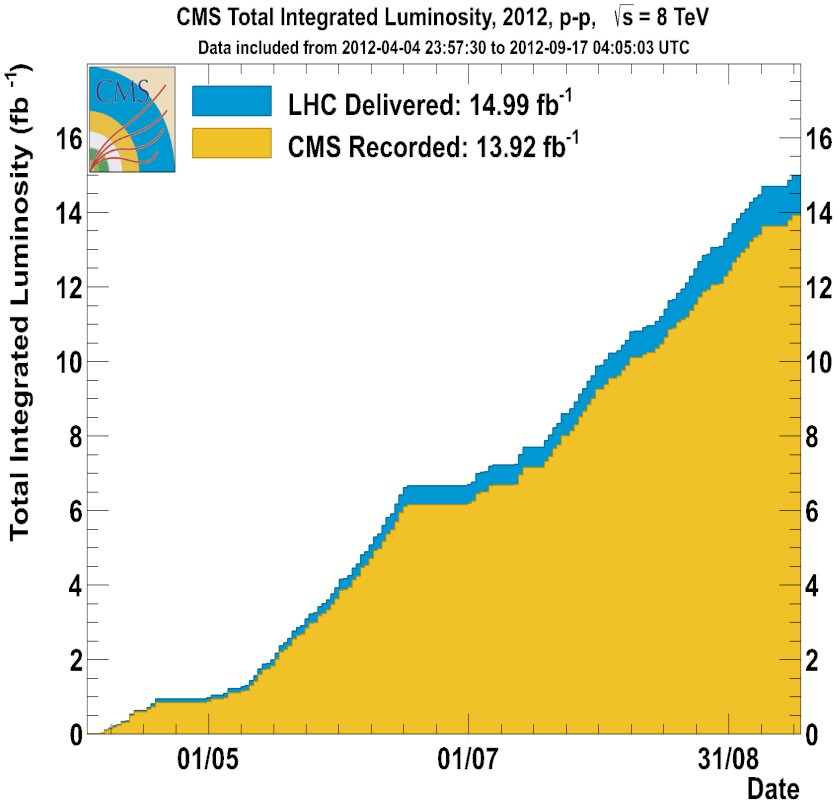
➢ savoir où sont les données pour y accéder facilement : « *Any data, any time, anywhere* »...

Évolution au niveau des T1/T2 CC-IN2P3

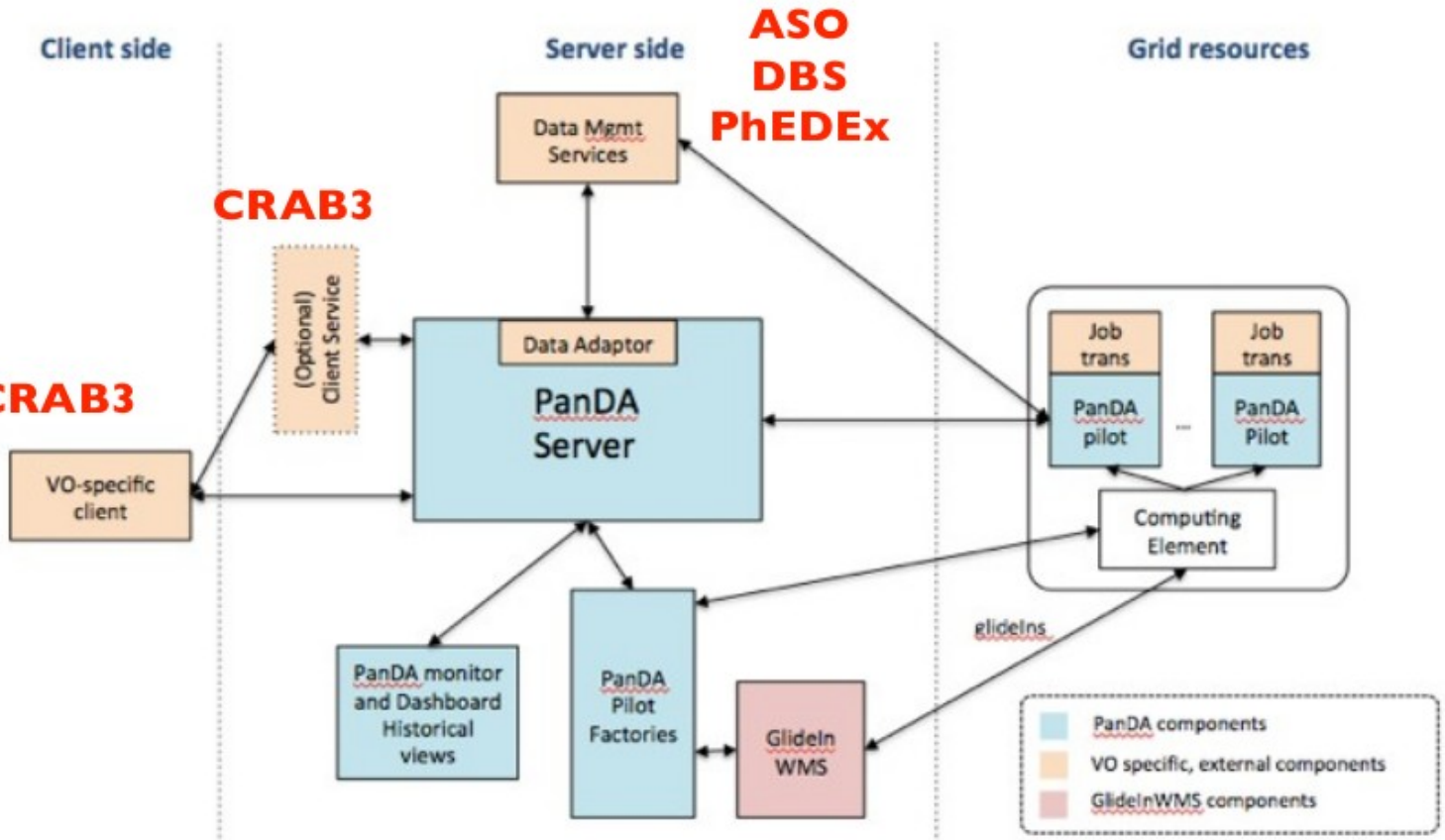
- À la suite des TEGs :
  - CVMFS (installée le 18 septembre, 2nd T1 à le mettre en production)
  - Glexec sur les workers (fonctionne, mais il s'agit pour l'instant d'un « workaround »)
- Fédération des données :
  - Le T2 va participer aux tests (demande de CMS, groupe stockage Ok)
  - Nouveaux tests SAM dédiés
- Problème de transferts vers certains T2s :
  - Installation d'un serveur PerfSonar sur tous les T2 afin de diagnostiquer les problèmes réseau.

Évolution de l'activité de calcul T0/T1

- Augmentation de la luminosité
  - bonne efficacité d'enregistrement des données ~96 %
  - nécessite beaucoup de travail sur le logiciel (en particulier au niveau du pile-up, et du tracking) afin de maintenir de bonnes performances
- La politique des re-reconstructions va évoluer : limite de calcul atteinte
  - ~1 milliards de données re-reco par mois
  - re-reco des données 2010/2011 requiert maintenant 8 mois
- Moins de release de code (moins de re-reco), prioritisation de plus en plus forte.



Schematic



- Common Engine
- Reuse existing software and tools
- Do not change the user interfaces

CERN IT Department
CH-1211 Geneva 23
Switzerland
www.cern.ch/it

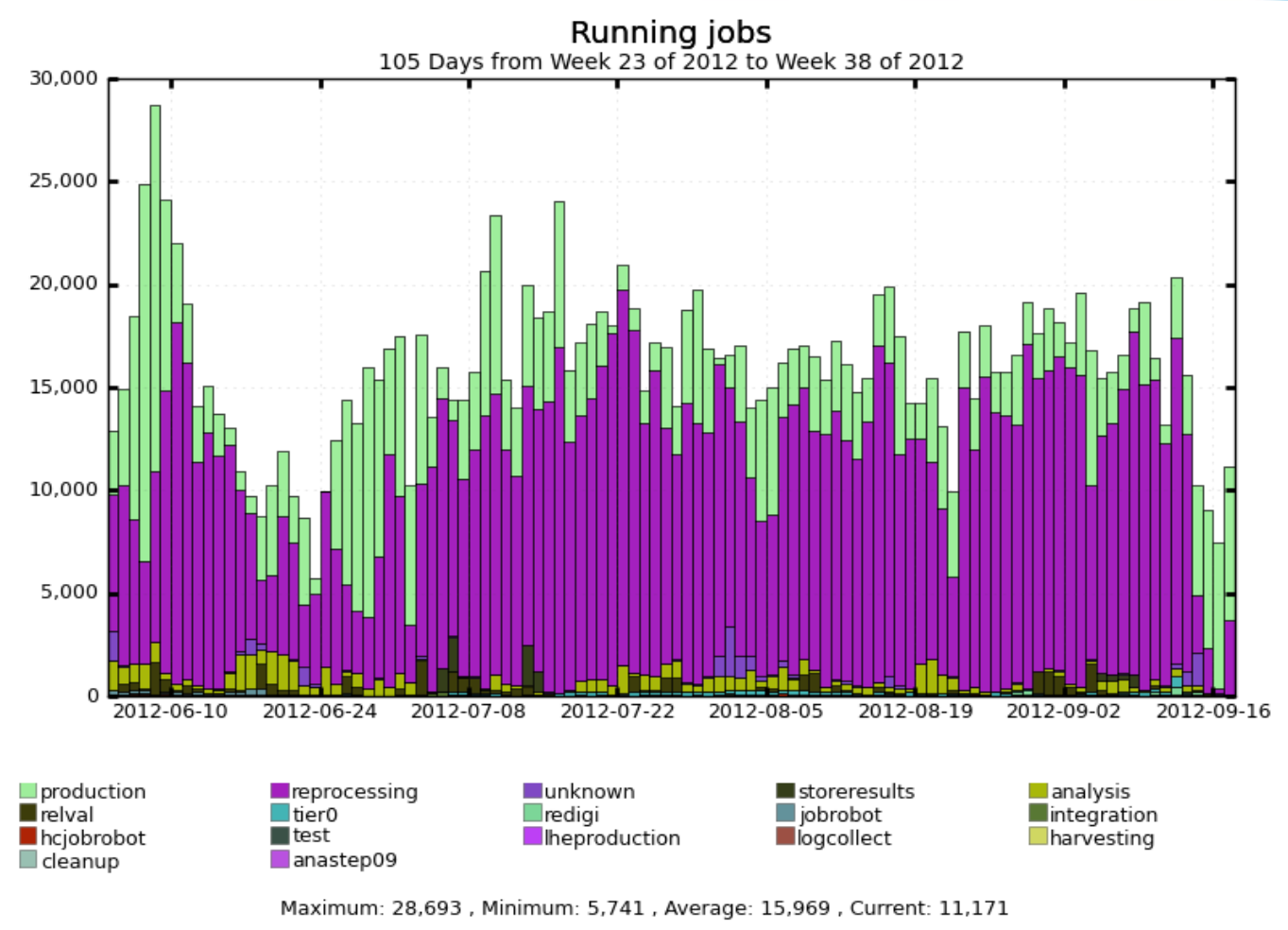
Conclusion

- Activités stables et continues (T1s)
  - Très bonne efficacité des jobs (en particulier pour les jobs officiels)
  - Gestion des données minutieuses
  - Communication rapide avec les gens du CERN (jabber) en cas de besoin

- Évolution du site :
  - CVMFS et glexec déployés, en production pour glexec
  - Fédération des données avec xrootd va être testé sur le T2
  - Multicore sur queue et machines dédiées, à transférer sur la ferme générale

- Mais *problèmes au niveau du stockage* (disque et tape) :
  - Pas assez de disque, donc effacement des données nécessaire (stockées dans HPSS)
  - L'effacement des données trop tôt dans dCache (tous les mois environ), entraine des « staging » massifs fréquents
  - Remplissage important de HPSS (>90%) conduit CMS à faire des campagnes d'effacement plus importante.

Activités T1



Framework support for multicore

- A better solution would be to implement fine-grained parallelism with the framework managing the scheduling (e.g. running independent modules or events in parallel) or within particular algorithms (like tracking). This would improve scalability for total memory use, performance on the memory hierarchy in the CPU and also simplify the I/O.
- The Framework work has recently completed a technology evaluation of various technologies which could support fine-grained parallelism/threads under the hood in the Framework (OpenMP, libdispatch, Intel TBB). Intel TBB was chosen as the best match for our needs. Work is ongoing to understand how to evolve the Framework design to provide this functionality.