Jeudi 13 Octobre 2011

Instance dCache LCG: Statut Actuel et Perspectives

Yvan Calas
yvan.calas@cc.in2p3.fr

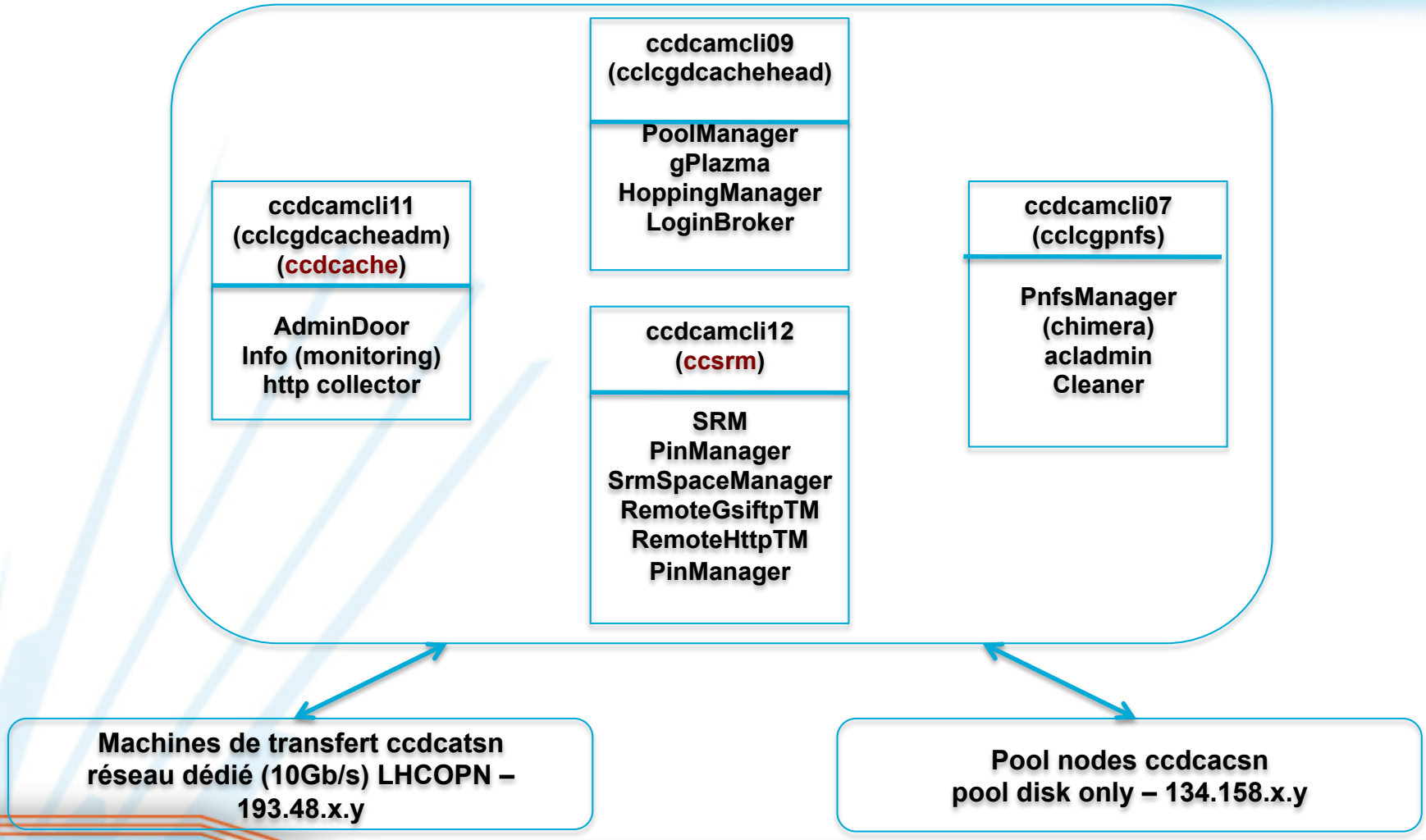
dapnia
cea
saclay

CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

Quelques chiffres pour dCache LCG…

- Support des 3 VOs LHC sur la même instance dCache.
- ~180 machines (core servers + disk servers).
- Core servers: PowerEdge 1950 avec 8 cœurs et 16GB RAM.
- Disk servers:
  - Sun Fire X4540 32TB et 64TB.
  - Dell PowerEdge R510 55TB.
- Capacité actuelle de stockage: ~7.2PiB.
- ~ 25 millions de fichiers référencés dans dCache.
- Version installée: 1.9.5-29 (aka « 1st golden release »). Fin du support dcache.org pour cette version au 1er mars 2012.

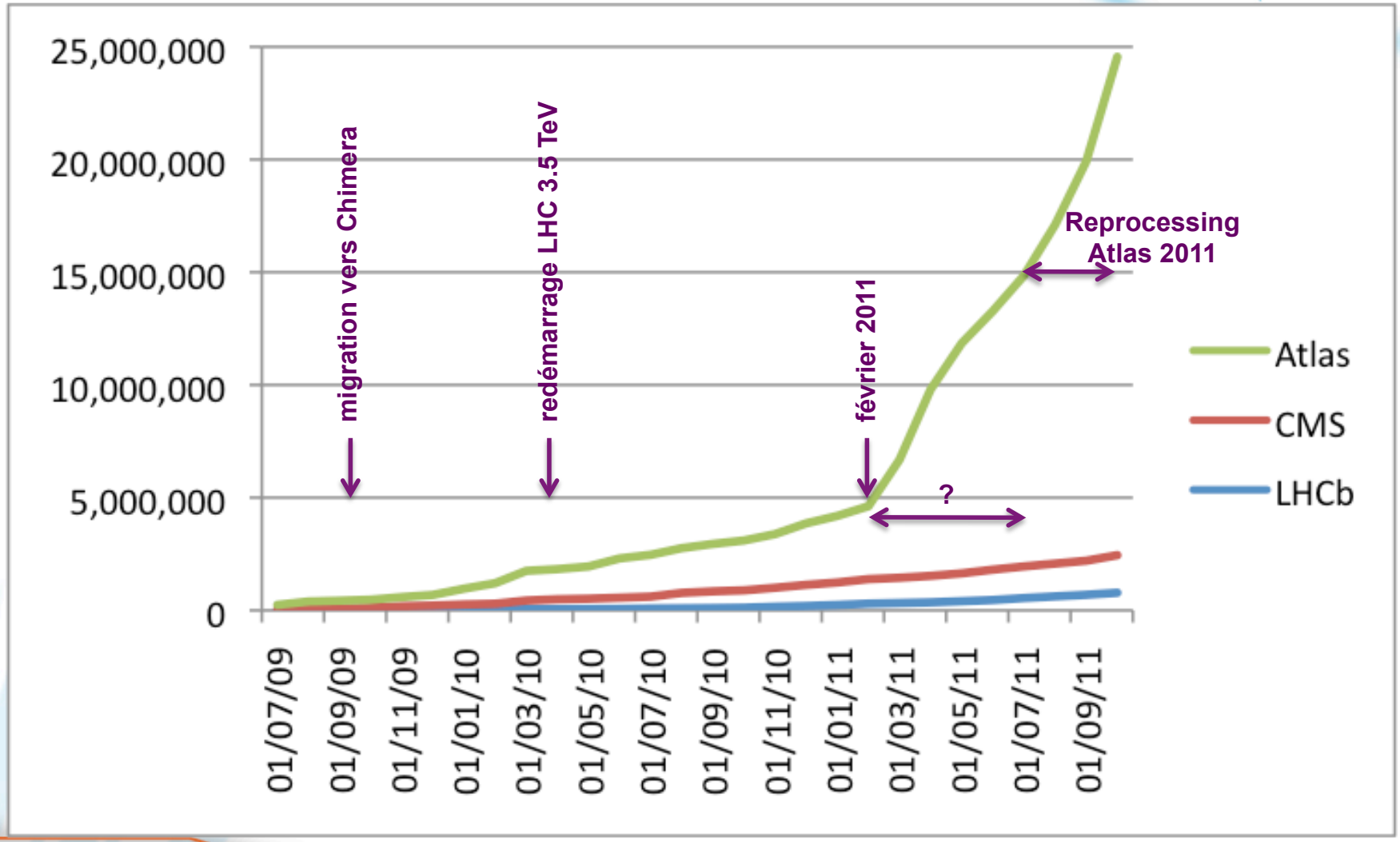
Configuration actuelle (1/2)



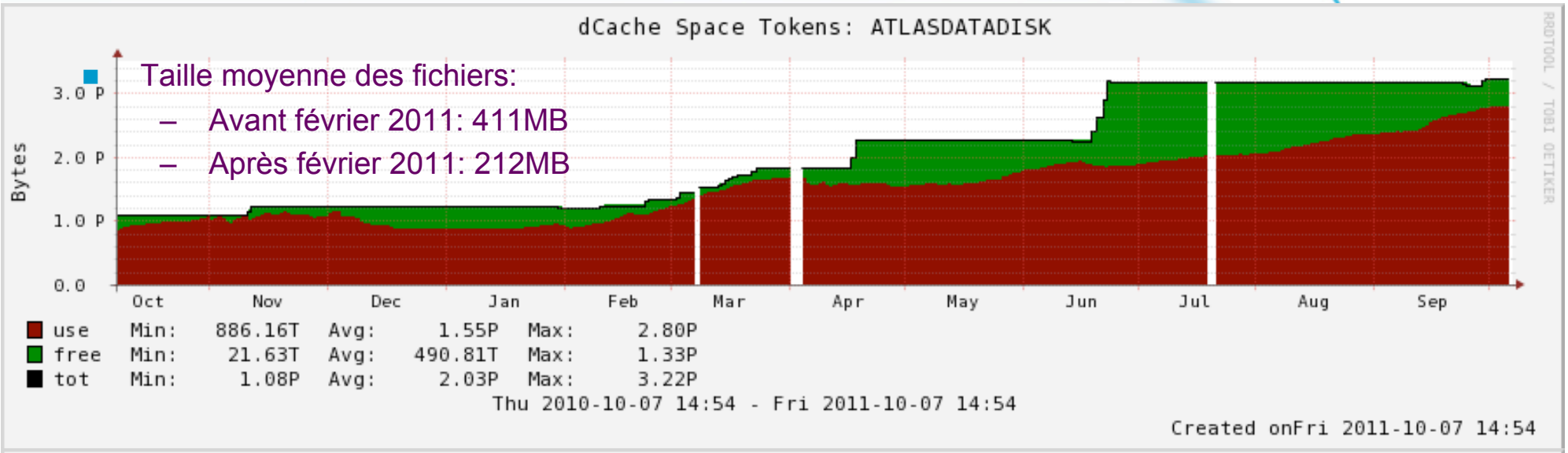
Configuration actuelle (2/2)

- 4 core servers exécutant des services critiques:
  - 2 serveurs utilisés principalement:
    - A l'archivage durant 15 jours des transferts (billing) internes à dCache ainsi que pour l'import/export (Base de données Postgres). Utilisé principalement pour le debug ou l'audit.
    - A la gestion interne des pools et du contrôle des flux internes d'information (PoolManager).
    - ***Statut: OK.***
  - 1 serveur SRM (ccsrm.in2p3.fr) pour:
    - Communications inter-sites (lectures srmget / écritures srmput de fichiers, réservation d'espace, etc.).
    - Base de données Postgres pour la conservation de ces informations durant 15 jours.
    - ***Statut: ~OK*** (rejet des connexions entrantes par TCP backlog durant le dernier reprocessing).
  - Serveur Chimera (ccdcamcli07.in2p3.fr):
    - Elément central de dCache.
    - Toute demande d'écriture / lecture / suppression / migration / staging passe par Chimera.
    - Base de données Postgres (8.3.5) contenant les métadonnées associées à chaque fichier enregistré dans dCache (fichiers stockés sur disques dCache et/ou dans HPSS).
    - 24.9 millions de fichiers au 29/09/2011 (90% pour Atlas, 6.8% pour CMS et 3.2% pour LHCb).
    - Base de taille assez importante: 48GB (au 29/09/2011).
    - Cette base a vocation à grossir « indéfiniment » avec une croissance ***exponentielle***.
    - ***Statut NOK actuellement*** (cf. transparents suivants).

Evolution du nombre de fichiers

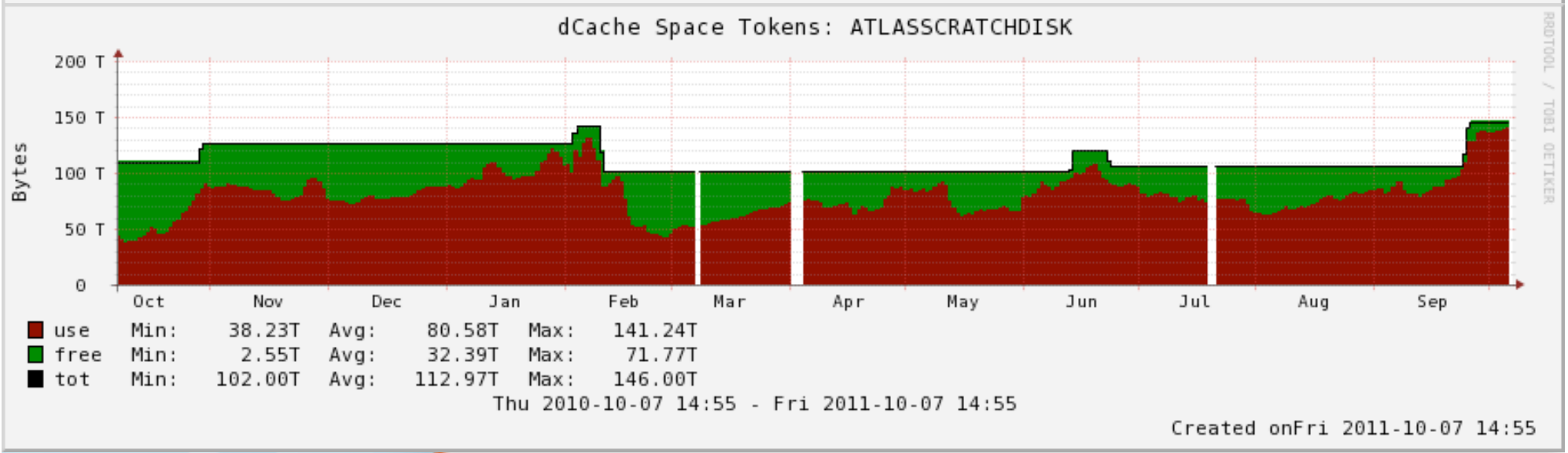


Space token ATLASDATADISK



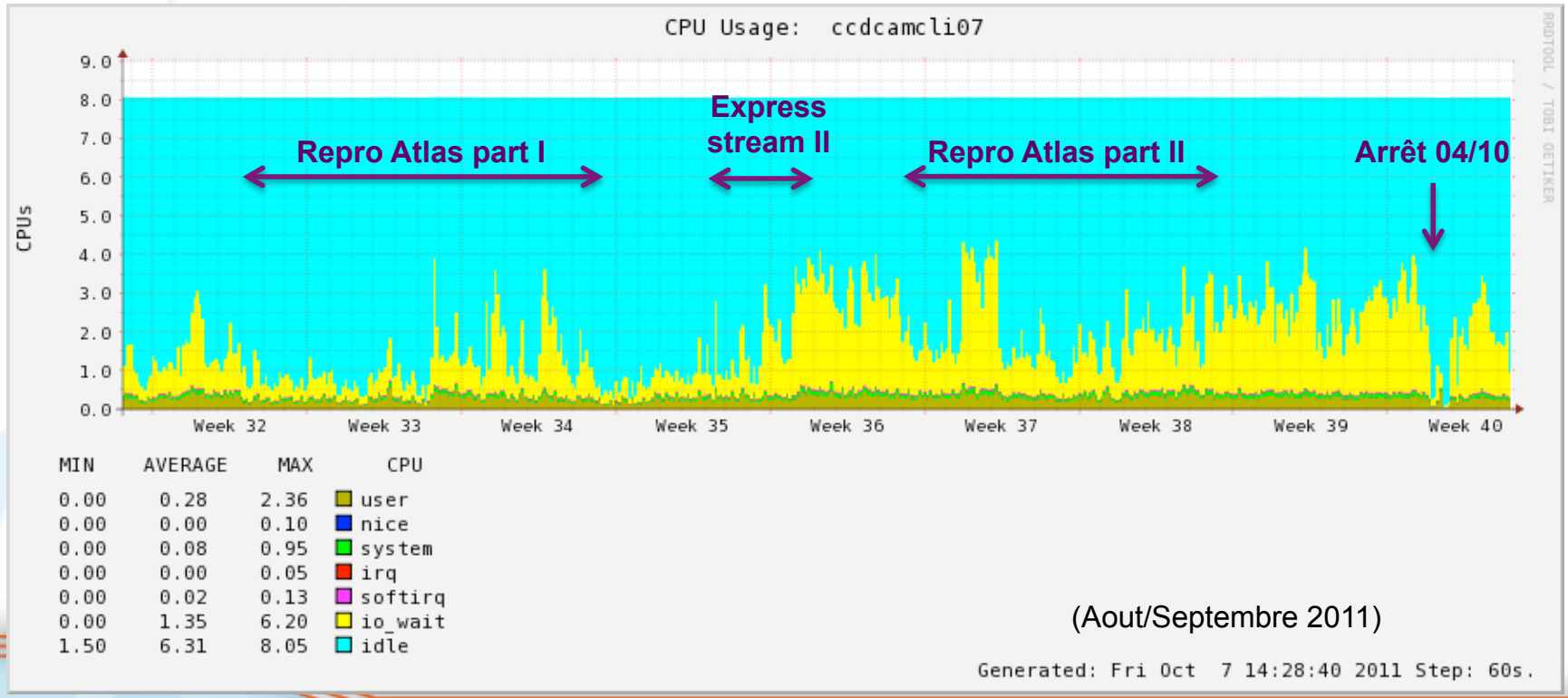
Taille moyenne des fichiers:

- Avant février 2011: 411MB
- Après février 2011: 212MB



Problème actuel avec Chimera

- Forte activité d'IOWait rendant les réponses aux requêtes internes à dCache plus lentes.
- S'est remarqué notamment durant la période de reprocessing Atlas 2011 (erreurs lors des écritures de fichiers à partir des WNs).



Problème actuel avec Chimera

- Actions durant le reprocessing:
  - 1 full dump de Chimera (au lieu de 2).
  - Désactivation de scripts de nettoyage.
  - Désactivation de crons faisant de l'accounting (demande des VOs).
  - Marge de manœuvre restreinte…
- Actions lors de l'arrêt du 4 octobre 2011:
  - Full dump/restore de Chimera + régénération des index.
  - ~~Doublement de la RAM (32GB à présent).~~
- Bilan:
  - ~~Réduction des IOWait et donc amélioration des performances de Chimera comme prévu…~~
  - Jusqu'à quand pourrons nous tenir?

Problème actuel avec Chimera

- Actions à moyen terme (d'ici début 2012?):
  - Achat d'une machine mieux profilée pour optimiser les performances de la base Chimera, notamment:
    - Disques haute performance avec configuration en RAID 10 + « battery backed write-back cache ».
    - RAM (>32GB).
  - Migration de dCache vers la 2ème golden release (1.9.12):
    - Amélioration des performances si l'on en croit les développeurs dCache.
- Action à long terme: quid des autres T1?
  - @FZK,PIC: 2 instances distinctes de dCache (Atlas, CMS).
  - @CC-IN2P3: plusieurs possibilités…
    - Instance actuelle dCache dédiée à Atlas + 1 nouvelle instance pour CMS et LHCb.
    - 1 nouvelle instance pour Atlas, instance « historique » inchangée.
  - Faisabilité?
  - Coût matériel, humain, en temps?

Une instance dédiée à CMS et LHCb

- 3 (ou 4) core servers nécessaires (Chimera, SRM et billing/PoolManager).
- Profonde réorganisation des pools sur l'instance actuelle:
  - Migration interne des données entre pools.
  - Libération progressive des machines en fonction des migrations.
  - Opération délicate.
  - En profiter pour améliorer la configuration interne CMS et LHCb.
- Migration des fichiers CMS et LHCb vers la nouvelle instance:
  - Utilisation d'un canal FTS IN2P3-IN2P3 dédié.
  - Plus simple et plus sûr…
  - 2.6 PiB de données.
  - Temps de migration important (difficulté à en prévoir la durée?).

Une instance dédiée à CMS et LHCb

- Machines de transfert: plus de machines avec moins d'espace disque par machine. Portes GFTP sur machines virtuelles?
- Accès à HPSS inchangé.
- Modification ou ajout de canaux dans FTS (instances FTS locales + externes).
- Dashboard dCache à dupliquer.
- Publication à modifier.

➔ Importance de l'intervention du 4 octobre 2011 pour avoir une idée précise de l'amélioration apportée.

Conclusions

- Forte activité de Atlas:
  - En import (écriture) et en export (lecture).
  - Durant les périodes de reprocessing.
  - Croissance exponentielle de la base Chimera.
  - Problème de passage à l'échelle:
    - Ce qui marchait jusqu'à présent ne semble plus convenir, du moins pour Atlas et avec la configuration actuelle.
    - Ralentissement inexorable des performances.
    - Demander à Atlas de supprimer les fichiers « inutiles »? Instance dCache dédiée à Atlas pour archivage des « vieux » fichiers?
    - Travail de prospection à effectuer avec les autres T1.
- Dédier une instance dCache commune à CMS et LHCb?
  - Faisable, mais avec un coût certain.
  - Si oui, quand effectuer une telle migration? Quelle échéance?
  - Suffisant? ~~Faire le point lors de l'arrêt du 4 octobre 2011…~~