Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

L'expérience au CC-IN2P3 avec le système de fichiers CernVM-FS

Neuvièmes Journées Informatique de l'IN2P3-IRFU

Aurélien GOUNON – Vanessa HAMAR

Introduction

- CernVM-FS
  - Définition
  - Objectives
  - Caractéristiques
- Architecture
- CernVM-FS au CC
- Conclusion


Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

CernVM-FS

Définition

- CernVM-FS est un système de fichiers en lecture seule utilisé pour accéder aux logiciels et aux données des expériences dans les infrastructures de calcul distribué.

- Objectifs:
  - Découpler le logiciel des expériences des disques des machines de calcul.
  - Être utilisé comme remplacement des systèmes des fichiers réseau (AFS, NFS, Lustre)

- Ou simplement :

> Distribuer les logiciels de manière efficace dans diverses infrastructures distribuées

Caractéristiques

- Plateforme virtuelle commune d'applications pour toutes les expériences.
- La procédure de déploiement des applications reste sous la responsabilité de chaque communauté.
- Utilise des technologies standards (fuse, http, SHA-1)
- Les fichiers et les métadonnées sont cachés et téléchargées « on demand »

Caractéristiques

- CernVM-FS est adapté dans les cas d'utilisation des logiciels ou il y a:
  - un grand nombre de petits fichiers à ouvrir et `a lire
  - Des ouvertures fréquentes des fichiers ou plusieurs répertoires sont examinés
  - données stockées en un seul site
  - accédés depuis un grand nombre des machines

Caractéristiques

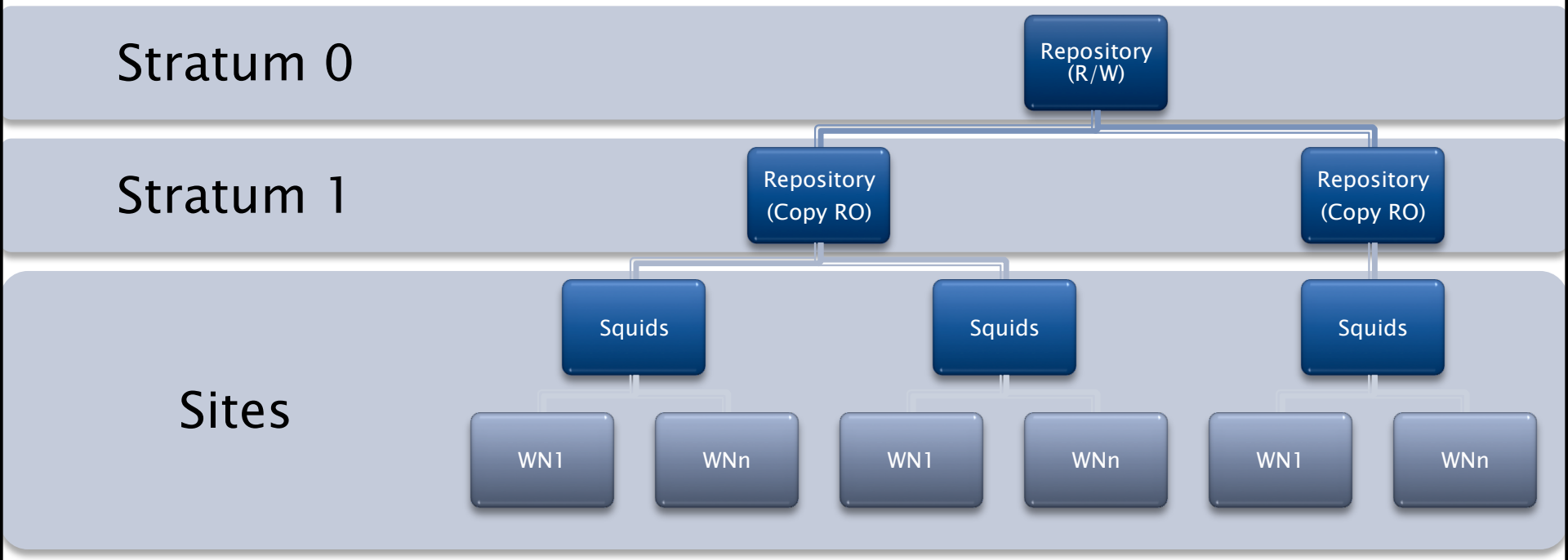
- Les fichiers et les répertoires sont stockés dans des serveurs web standards et accédés dans l'espace universel

/cvmfs

- Les connections sortantes sont seulement des connections HTTP pour éviter des problèmes avec les pare-feux

- La vérification des transferts des données et des métadonnées est faite en utilisant des « hash » cryptographiques

Architecture



Stratum 0 – Serveurs R/W - Release Manager Machine

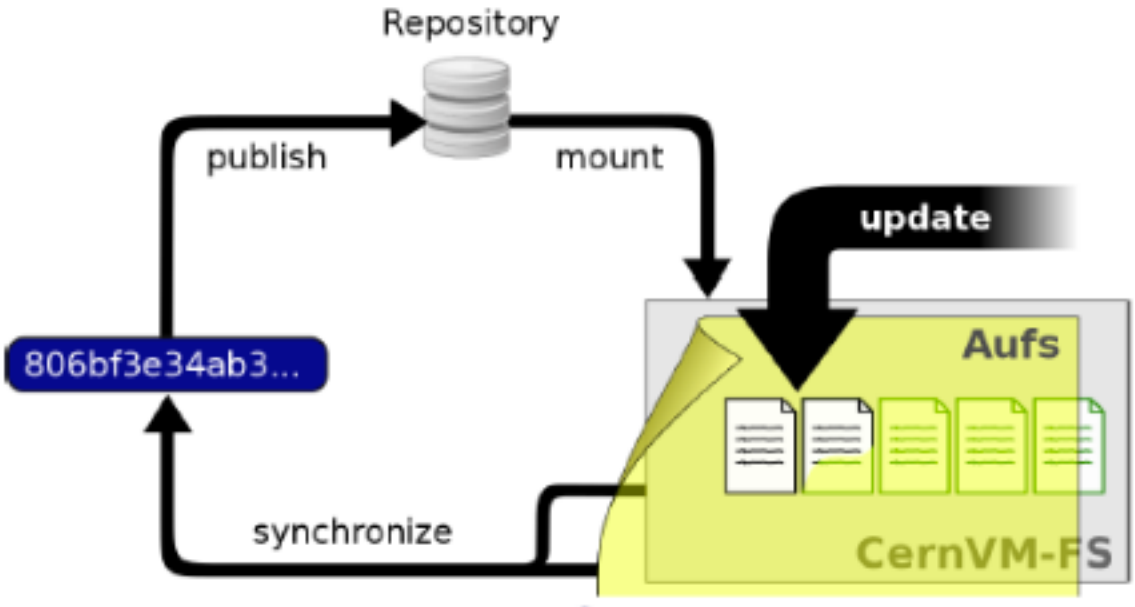
- Unique source des nouvelles données

- ***Repository :*** arborescence des fichiers stockée sur un serveur web (*repository server*) qui est vue comme un système de fichiers em lecture seule du coté client.

  - ***File Catalog :*** base de données SQLite qui contient une seule table avec :
    - Structure répertoires
    - Métadonnées
    - Identifié par la clé SHA-1 du contenu du fichier de la base de donnes SQLite.
    - Contient les clés SHA-1 des contenus des toutes les fichiers réguliers.
    - Liens symboliques

  - ***Data Store*** : Contient les fragments compressés des fichiers (« Compressed Chunk files »),
    - les fichiers dupliqués sont détectés
    - les fichiers ne sont pas jamais effacés.

Repository



La MAJ est faite sur un system de fichiers AUFS avec la permission en écriture
Les changements sont synchronisés avec le *repository* original
Le *file catalog* est génèré, les fichiers sont comprimés et renommés, la clé hash est crée avant de les copier dans le *data store*.

Domaines :

- cern.ch
  - alice.cern.ch
  - alice-ocdb.cern.ch
  - alice-ocdb.cern.ch
  - atlas.cern.ch
  - atlas-condb.cern.ch
  - cms.cern.ch
  - grid.cern.ch
  - lhcb.cern.ch
  - na61.cern.ch
  - sft.cern.ch
  - boss.cern.ch
  - grid.cern.ch
  - sft.cern.ch
  - geant4.cern.ch
  - t2k.egi.eu
  - cernatschool.egi.eu
  - glast.egi.eu
  - snoplus.egi.eu
- egi.eu
  - auger.egi.eu
  - BIOMED.egi.eu
  - km3net.egi.eu
  - mice.egi.eu
  - wenmr.egi.eu
  - pheno.egi.eu
  - phys-ibergrid.egi.eu
  - hyperk.egi.eu
  - t2k.egi.eu
  - cernatschool.egi.eu
  - glast.egi.eu
  - snoplus.egi.eu
- openscience.org
  - oasis

L'utilisation par autres communautés commence a augmenter rapidement !!!

Stratum 1 – Serveurs R/O

- Multiple réplicas a la fin de :
  - Réduire la charge des serveurs
  - Améliorer la fiabilité
  - Protéger les serveurs de connections directes

- Serveur web standard
  - Utilisent les commandes CernVM-FS pour créer et maintenir un miroir des *repositories* servi par un serveur stratum 0.

- Synchronisations périodiques (crons)
  - CernVM-FS a des mécanismes pour vérifier l'intégrité des fichiers, réutilise le client fuse.

- Replicas
  - cvmfs-stratum-one.cern.ch
  - cernvmfs.gridpp.rl.ac.uk
  - cvmfs-atlas-nightlies.cern.ch
  - grid-cvmfs-one.desy.de
  - klei.nikhef.nl
  - cvmfs02.racf.bnl.gov
  - cvmfs03.racf.bnl.gov
  - cvmfs02.grid.sinica.edu.tw
  - cvmfs.fnal.gov

- Matrice de replicas

  http://cernvm-monitor.cern.ch/cvmfs-monitor/matrix/

Sites – Serveurs Squids

- Serveurs « Squids » ou autres proxys http doivent être installés dans les sites.
  - Préférable en redondance.

- La configuration des clients permet d'avoir une liste des serveurs squids.
  - Random « squidA|squidB »
  - Ordered « squidA;squidB »

Sites - Clients

- Connexion sortante HTTP aux serveurs proxys
- Les fichiers seront seulement en lecture
- CernVM-FS cache sur le disque local les données et le catalogue des fichiers.
- Les clients CernVM-FS ne doivent pas se connecter directement aux serveurs du stratum 0.
- Ouverture d'un fichier :
  - Le nom est résolu a travers le catalogue Sqlite
  - Les fichiers sont téléchargés et sa clé hash est vérifiée avec l'entrée correspondante dans le catalogue.

Considerations de securite

- CernVM-FS fait confiance au données du répertoire de cache local.
  - cvmfs_fsck
  - cvmfs et fuse sont les propriétaires du répertoire cache
- Libcurl gère les téléchargements

Limitations dans les contenus des *repositories*

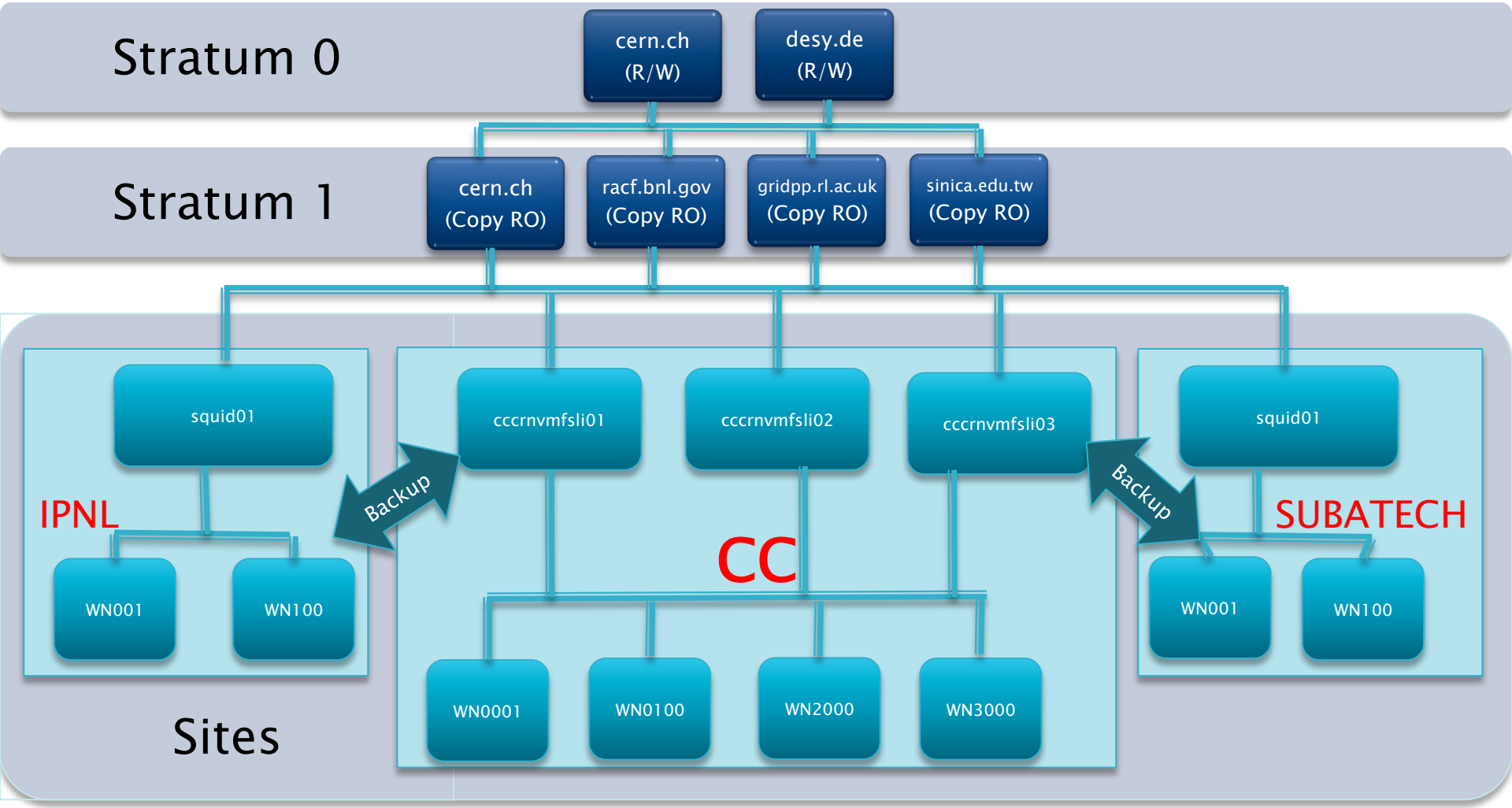
- Fichiers de données : la quantité total de données lue par job est plus ou moins la quantité disponible de RAM par job.

- Création rapide de « données déchets » qui sont maintenues dans les *repositories*.

- On s'attend a ce que :
  - Le même fichier soit lu plusieurs fois pour le même WN
  - Le même fichier soit stocké par multiples WNs qui exécutent un type de job particulier
  - Le même fichier puisse être demandé en même temps par de multiples WNs aux serveurs Squids

CernVM-FS au CC

Architecture



CernVM-FS au CC

- 9 Repositories supportés
- WNs
  - Entre 40GB et 70GB de cache
- Squids
  - 3 serveurs squids avec 170GB de espace pour le répertoire cache
  - Configurés de manière aléatoire dans les WNs pour le CC et backup pour SUBATECH et IPNL.

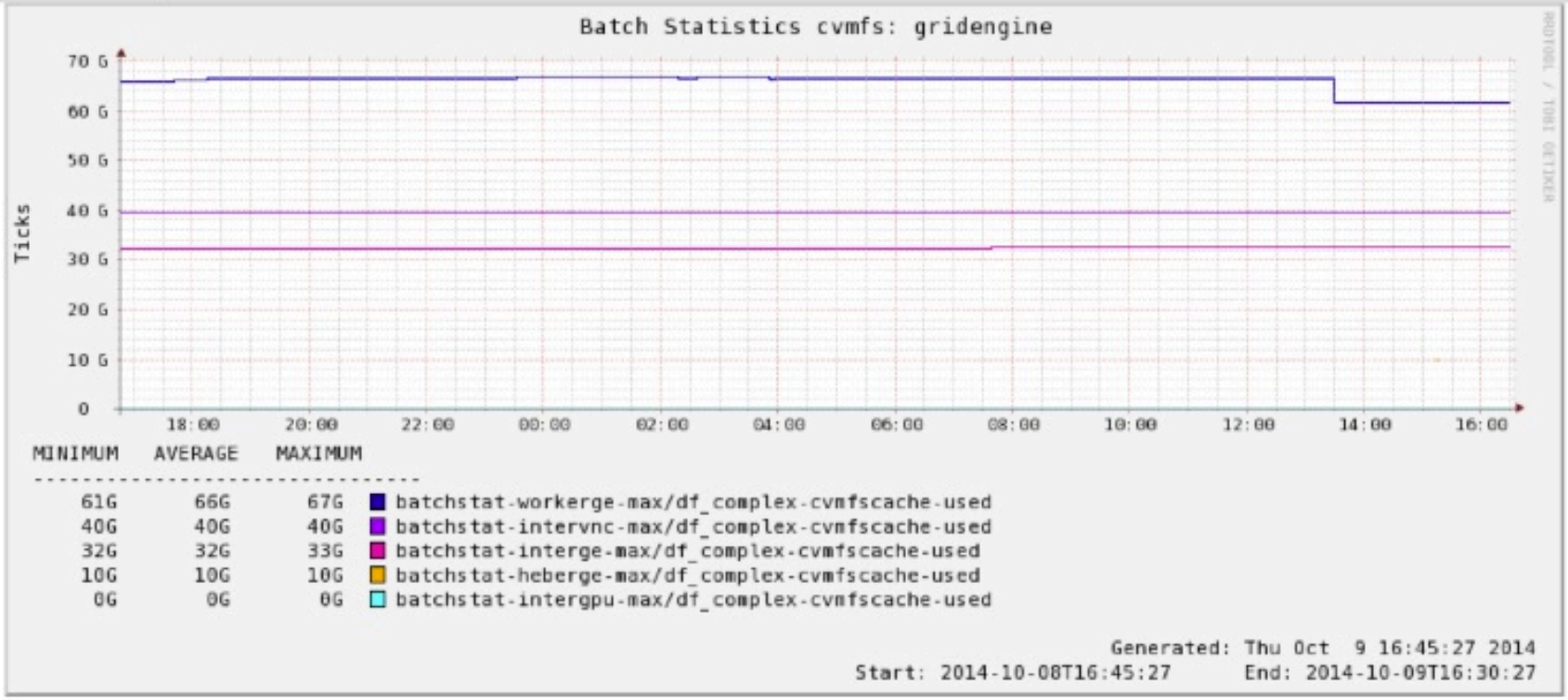
alice.cern.ch
ams.cern.ch
atlas.cern.ch,
atlas-condb.cern.ch
atlas-nightlies.cern.ch
cms.cern.ch
ilc.desy.de
lhcb.cern.ch
sft.cern.ch

ccsmurf.in2p3.fr/graph/machine/gridengine/batchstat-cvmfscache?filematch=used&filematch=max

Latest Headlines | Most Visited

gridengine: batchstat-cvmfscache

Select
List
Matrix
Multi



Max cvmfscache 67 GB
Med cvmfscache 40 GB
Min cvmfscache 33GB

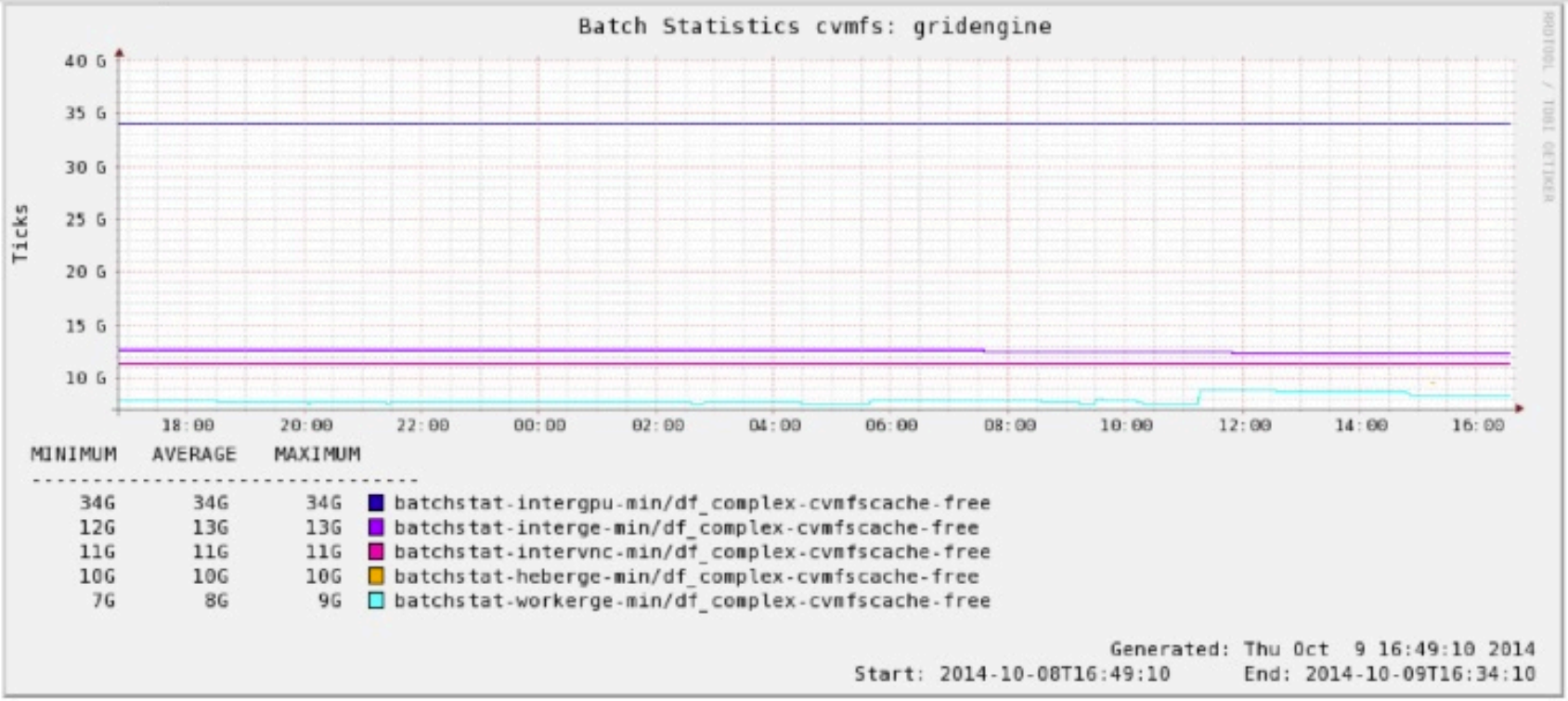
Smurf::Server (0.1034)
Smurf::RRD (0.2020)
Smurf::DB (0.4007)
Smurf::Util (0.1010)

ccsmurf.in2p3.fr/graph/machine/gridengine/batchstat-cvmfscache?filematch=free&filematch=min

Latest Headlines | Most Visited

gridengine: batchstat-cvmfscache

Select
List
Matrix
Multi



Max cvmfscache free 34 GB
Med cvmfscache 13 GB
Min cvmfscache 11 GB

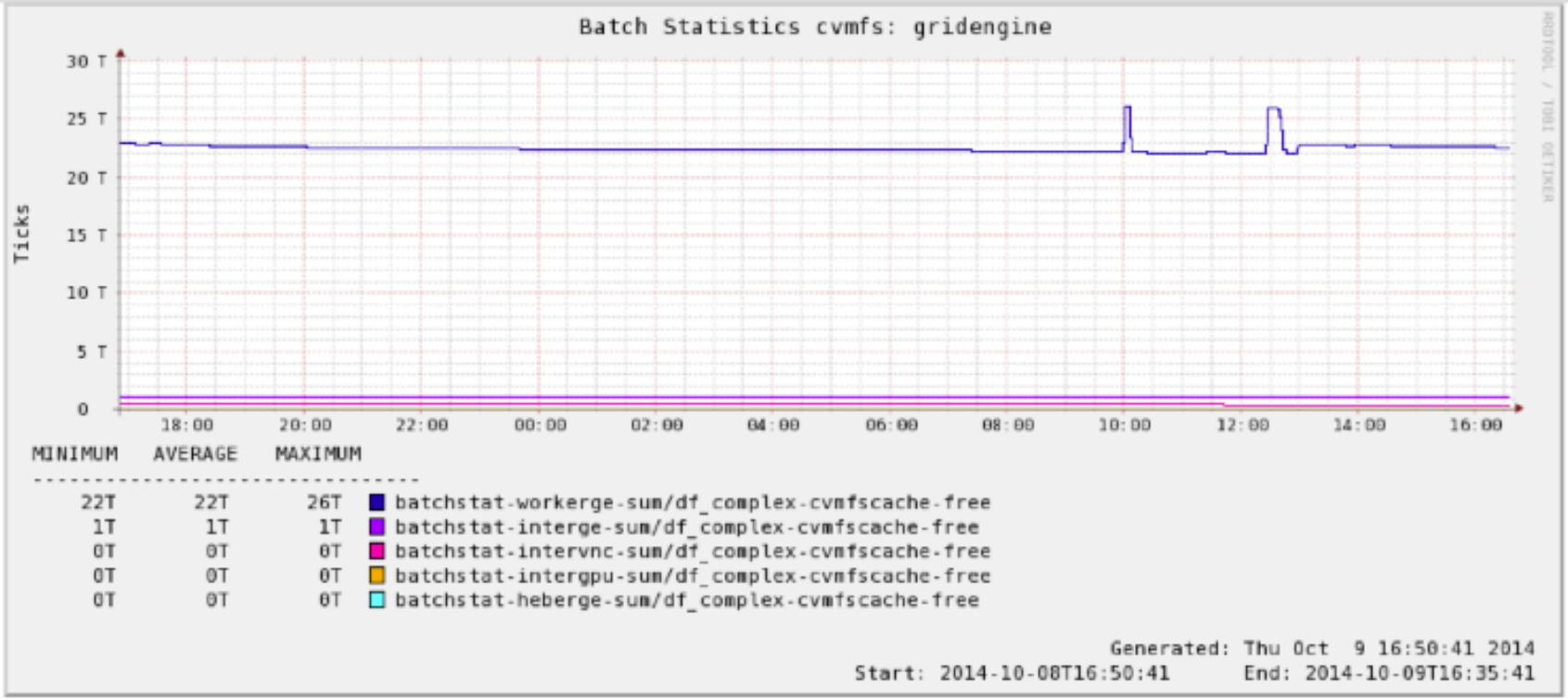
Smurf::Server (0.1034)
Smurf::RRD (0.2020)
Smurf::DB (0.4007)
Smurf::Util (0.1010)

ccsmurf.in2p3.fr/graph/machine/gridengine/batchstat-cvmfscache?filematch=free&filematch=sum

Latest Headlines | Most Visited

gridengine: batchstat-cvmfscache

Select
List
Matrix
Multi



Smurf::Server (0.1034)
Smurf::RRD (0.2020)
Smurf::DB (0.4007)
Smurf::Util (0.1010)

Total cvmfscache 26TB

Conclusions

- Les logiciels des expériences ont des changements fréquents, CernVM-FS met a disposition des utilisateurs de tous les sites ces versions immédiatement quand elles sont publiées.

- Pour des raisons de performance également, CernVM-FS a remplacé des systemes de fichiers distribués hébergeant des zones communes des logiciels d'expériences.

- Trouver la balance entre le nombre des *repositories* supportés, l'espace cache dans les WNs et la vitesse de téléchargement des fichiers dans la cache