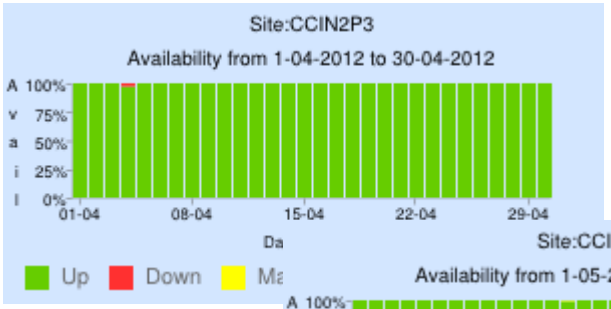
LCG T1/AF : sujet du jour

Renaud Vernet – Jun. 21st 2012

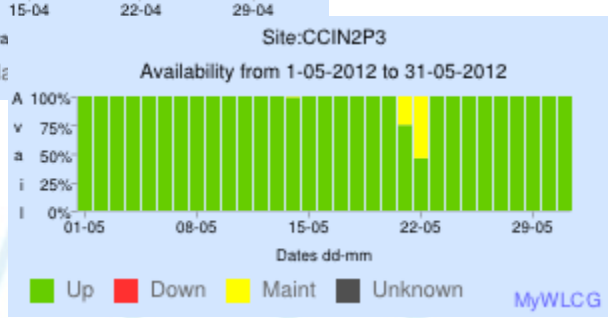
Outline

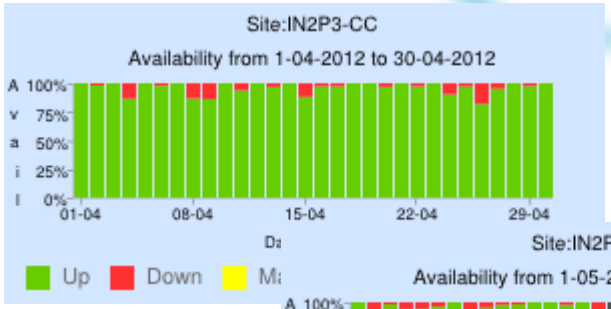
- Site report
- Reseau
- WLCG/GDB news
  - EMI, virtualisation, securite, bdii, federations, multicoeur, cvmfs...
  - GDB
    - *http://indico.cern.ch/conferenceDisplay.py?confId=155069*
  - ATLAS french cloud regional centers meeting
    - *https://indico.cern.ch/conferenceDisplay.py?confId=181944*
- Chantiers
  - Termines, en cours
- Notes de fin en vrac

Site availability (avril + mai)

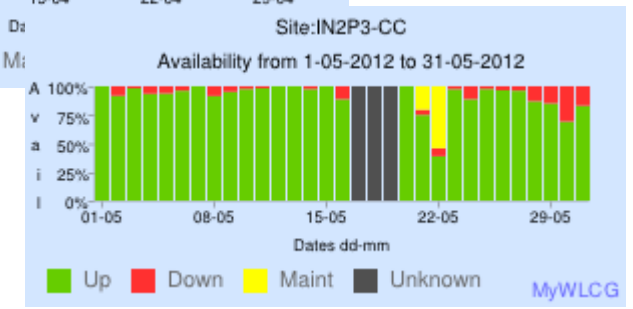


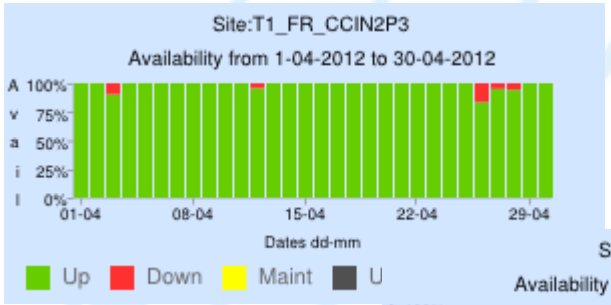
ALICE



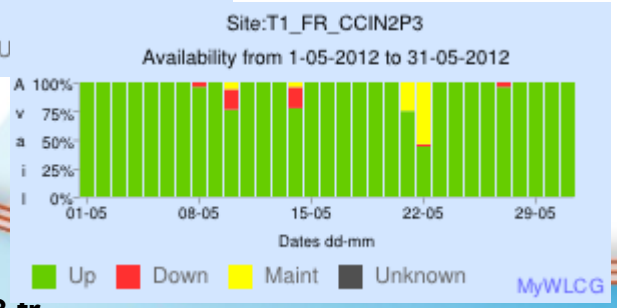


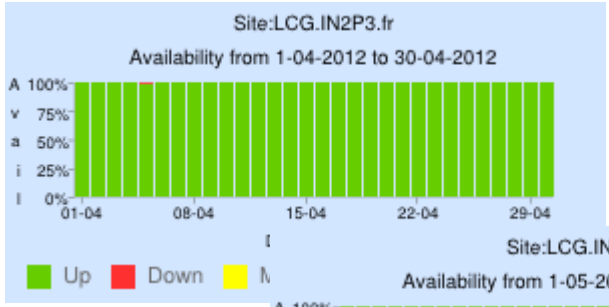
ATLAS



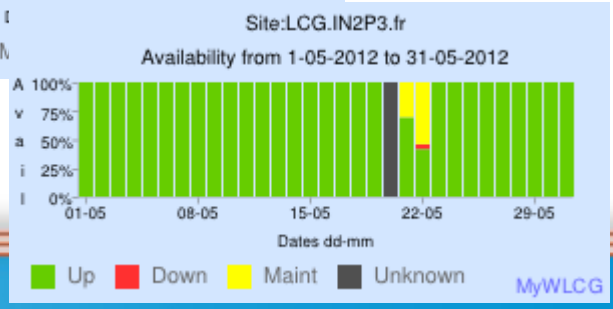


CMS





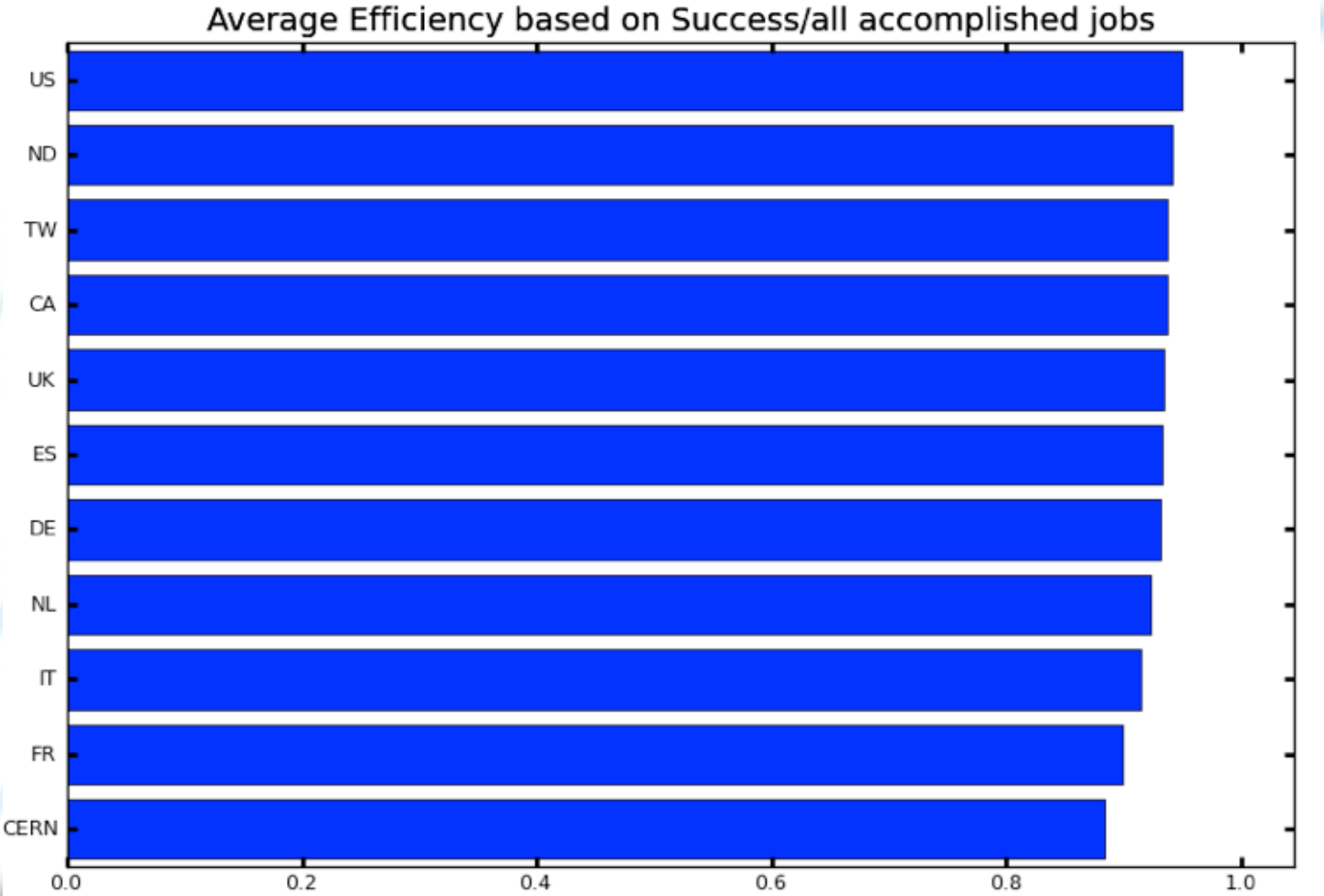
LHCb



renaud.vernet@cc.in2p3.fr

Taux de succes des jobs ATLAS

Annee 2011



Reseau

- Fin de la hierarchisation par Tier
  - Performances sur reseaus lointains plus importants qu'avant
- Transferts inter-sites de differents 'nuages'
  - Tier2 doivent devenir Tier2-D des que possible si queue d'analyse
    - performance reseau > 5MB/s
- ALICE (et peut-etre LHCb) sont deja dans un modele de ce genre
  - Peu de hierarchisation
  - Mais pas de pression sur les sites
- Importance du monitoring reseau (perfSonar)

Reseau

- Performance pas toujours bonnes
  - Depend du site chez qui on envoit
- Difficile de mettre ne place un monitoring global sur toute le chemin reseau
  - En particulier pour les liens hors France
- ATLAS :
  - Vers Beijing : tests dedies a mettre en place par telecom

Reseau

Transfer performances

CCIN2P3 to distant sites

CCIN2P3 to	LAN	LAL Paris	GEANT London	IHEP China	ICEPP Japan
LINUX 2x 1 Gb/s Old algo	830 Mb/s	620 Mb/s	160 Mb/s	15 Mb/s	2 Mb/s
LINUX 2x 1 Gb/s New algo	860 Mb/s	600 Mb/s	400 Mb/s	220 Mb/s	12 Mb/s
SOLARIS 2x 1 Gb/s	940 Mb/s	940 Mb/s	400 Mb/s	200 Mb/s	400 Mb/s

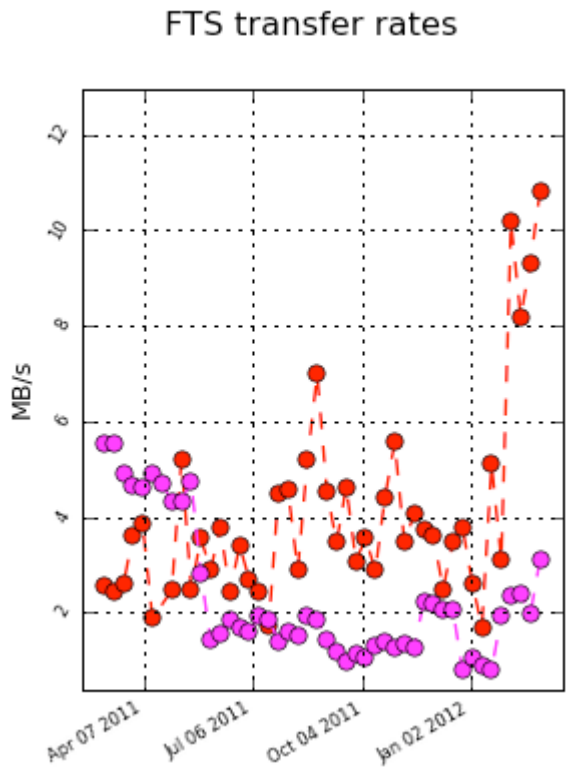
iperf tests 1 stream -w16M

ATLAS FR Cloud June 2012

Test J. Bernier

Reseau

- Debit ATLAS



Beijing → CCIN2P3

CCIN2P3 → Beijing

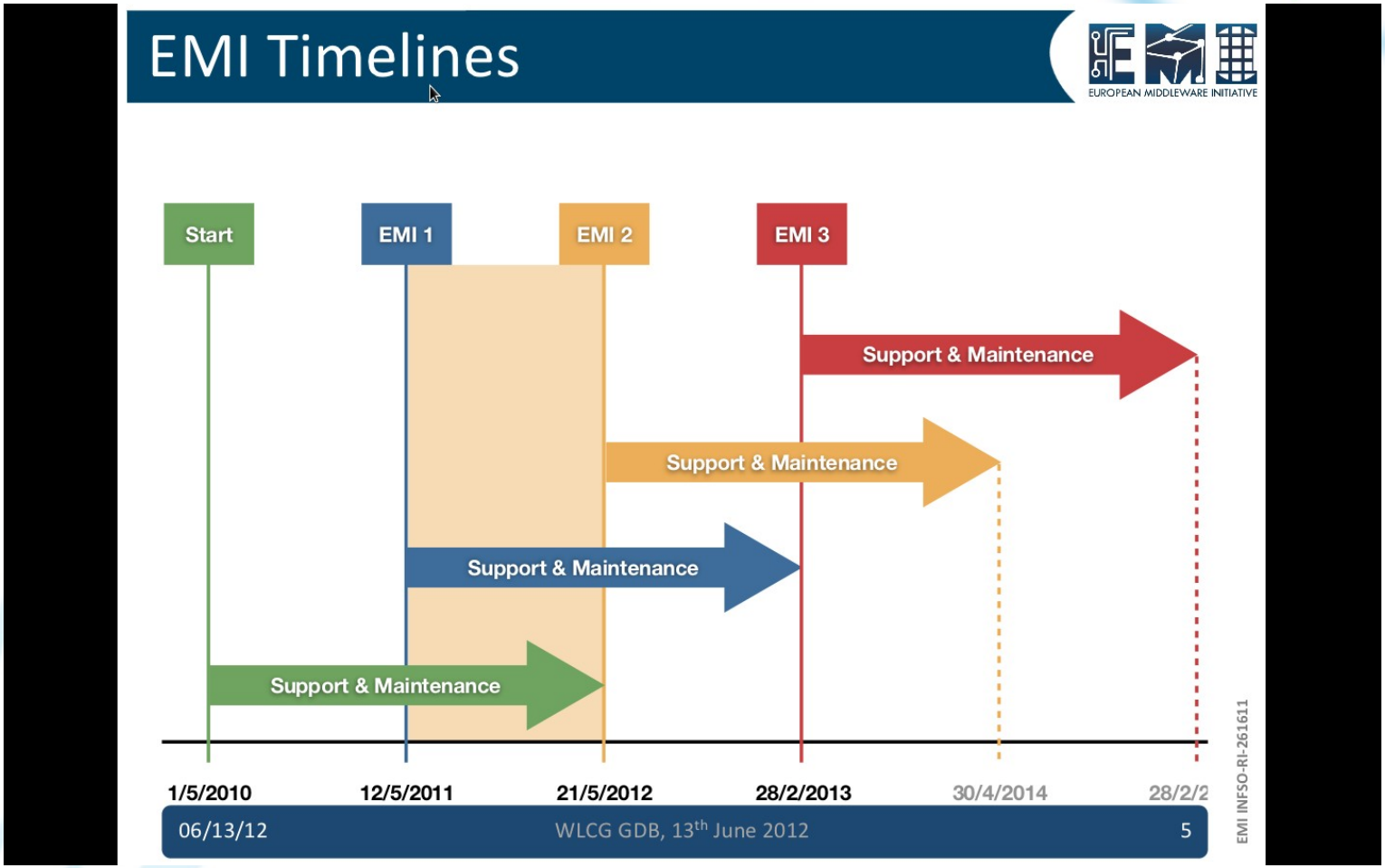
Jerome va suivre ca

Reseau

- PerfSonar !
- Deploiement souhaite au CC
  - 1 machine mise a disposition par telecom
  - ATLAS en souhaite 2
    - 1 monitoring latence, 1 monitoring debit
  - CMS en souhaite 1

→ devra-t-on multiplier ces machines ?

EMI


EMI Timelines
EUROPEAN MIDDLEWARE INITIATIVE
Start
EMI 1
EMI 2
EMI 3
Support & Maintenance
Support & Maintenance
Support & Maintenance
Support & Maintenance
1/5/2010
12/5/2011
21/5/2012
28/2/2013
30/4/2014
28/2/2
06/13/12
WLCG GDB, 13th June 2012
5
EMI INFSO-RI-261611

EMI

EMI 2 highlights

- 5 new products:
  - CANL, EMIR, EMI-Nagios, Pseudonymity, WNoDeS
- Better integration of the **Argus** authorization service in the CEs and SEs
- Better support for **GLUE 2** in all relevant services
- initial implementation of the **EMI Execution Service** interface in all CEs
- Support for **industry-standard protocols** for accessing the EMI SEs (NFS4.1, http, webDAV), a messaging-based Storage Element-File Catalog synchronization service (**SEMsg**)
- Not released: **Hydra, WMS**

Point sur les backward incompatibilities :
http://indico.cern.ch/getFile.py/access?contribId=4&resId=1&materialId=slides&confId=155069

Using external clouds

- When VOs use resources not provided by WLCG sites, or sites choose to expand by instantiating off-site cloud VMs, it is currently not possible to do so in such a way that conforms with WLCG security policies
- As specified in the WLCG risk assessment, there are significant concerns in using external cloud providers and additional work is needed to understand the policy issues it raises. There are also operational issues (including procedures and traceability)
- A working group should be appointed to conduct this work and report back to the GDB or MB as appropriate
- In the meantime VOs or sites instantiating external cloud resources should be aware of these concerns and the responsibilities they accept by using these services

Communication worker-batch

Proposal for real machines

VM & PM

- Physical machines, just as virtual machines, should have in `/etc/machinefeatures` the files
  - `hs06` – the HS06 rating for a single core
  - `shutdowntime` – timestamp for when the node is expected to be rebooted (may be empty)

PM

- Additionally, the batch system should provide each job with an environment variable, `$JOBFEATURES`, pointing to a directory with the files
  - `jobstart` timestamp for the start of job execution
  - `cpu_limit` the allowed number of cpu-seconds that can be used by the job. The value must be comparable directly against system reported cpu consumption.
  - `wall_limit` the allowed number of wall clock seconds. The job will be terminated at time `jobstart+wall_limit` whether or not `cpu_limit` has been reached.

With thanks to Steve Traylen and Ulrich Schwickerath

Tony.Cass@CERN.ch 3

gLExec

- Tests Nagios OK au CC
  - Sauf ATLAS
- Problemes vus par CMS
  - Ticket en cours avec Sebastien + Pierre + CMS
  - https://ggus.eu/ws/ticket_info.php?ticket=83087
- ATLAS et ALICE ne poussent pas specialement pour avoir gLExec rapidement

Caching BDII

- How to configure caching mode?
  - Use variable BDII_DELETE_DELAY in /etc/bdii/bdii.conf
    - Default is caching on (12h)
      - BDII_DELETE_DELAY=43200
    - For caching off
      - BDII_DELETE_DELAY=0
  - If you are using YAIM
    - In /opt/glite/yaim/defaults/glite-bdii_top.pre
      - BDII_DELETE_DELAY=43200
    - Redefine BDII_DELETE_DELAY in your site-info.def
- Documentation improved
  - https://tomtools.cern.ch/confluence/display/IS

Calcul parallele / multicore

- ATLAS & ALICE font des etudes de performance et integration de hardware type GPU pour leurs applications

- Grid5000 testbed et (bientôt) machine GPU au CC
  - Propose pour ALICE
  - Pas de nouvelles
  - ATLAS interesse ?
- CMS & LHCb ?

- Discussion au prochain pre-GDB (10 juillet)
  - Tentative de vraie integration du Multicore dans wlcg
  - Contribution developpeurs CE
  - → nouveaux attriuts JDLs a venir

Federations

- ATLAS : xrootd semble la semble issue aujourd'hui
- Besoin de WAN performant
  - Equilibre performance reseau / etendue geographique
  - Evolution de ROOT dans ce sens
    - Plugins lecture distante
    - TTreeCache
-

Avenir de SRM

- Pour ATLAS
  - Au moins pour acces MSS
  - Mais de moins en moins pour le disk
- Autres technologies en vue
  - Xrootd, http, gridFTP

CVMFS

- Interessant pour toutes les VOs
- LHCb et ATLAS poussent pour que tous les sites en soient equipes
- CMS : pour bientôt

- Nouvelles demandes LHCb + bientôt CMS
  - → reconfiguration de la partition sur workers
  - → prevu pour le prochain arret

Databases WLCG

- Orienter vers modele NoSQL
  - Plusieurs solutions examinees
  - → Hadoop est la solution retenue

Chantiers termines *(du moins pour le moment)*

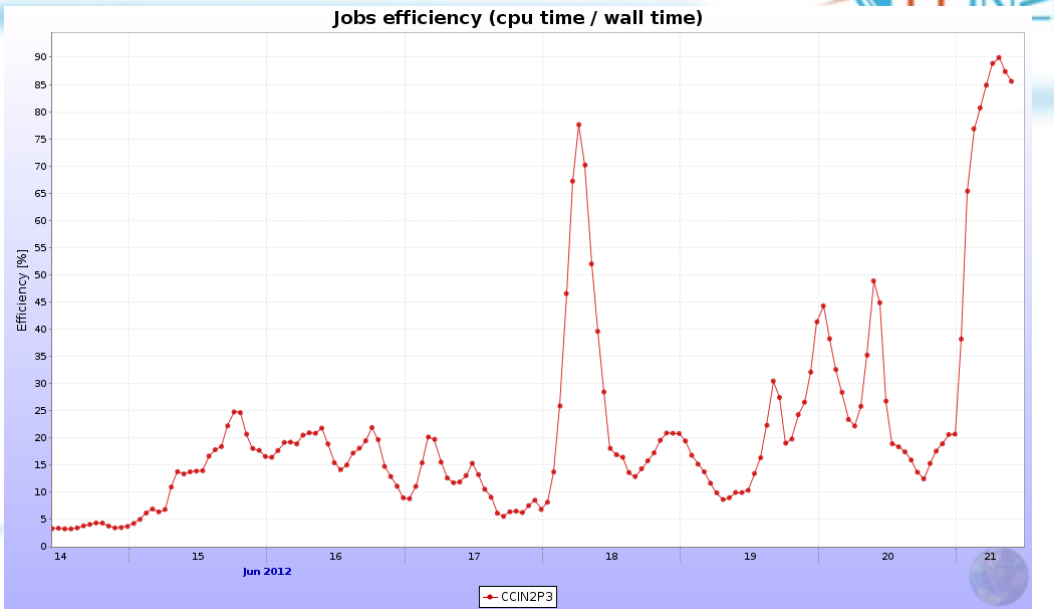
- Fichiers inaccessibles LHCb, probleme hardware dCache
  - https://ggus.eu/ws/ticket_info.php?ticket=82751
- Soumission pilotes LHCb
  - Chute nette du nombre de pilotes par jour
  - Changement dans Dirac ?
- Fichiers corrompus LHCb
  - Ticket clos, pas d'explication robuste
    - https://ggus.eu/ws/ticket_info.php?ticket=82247
  - LHCb s'engage a activer la verification du checksum lors de l'ecriture des fichiers sur les serveurs

Chantiers en cours

- Nettoyage bandes T3
  - Voir mail envoye par Pierre au support et representant VO fr
  - Point a faire la prochaine fois
- PerfSonar
  - ATLAS : demande une deuxieme instance
    - Mieux si la machine est sur le reseau des serveurs
  - CMS : Sebastien doit faire ca avec Laurent
- CVMFS
  - LHCb veut un nouvel espace pour con-db
    - Necessite arret des machines → prochain downtime (septembre?)
    - LHCb est OK
  - CMS va bientôt arriver → important d'avoir les specs en septembre
- gLExec
  - Marche moyen
  - Nous devons faire un point technique avec M. Litmaath
- Augmentation memoire LHCb
  - La memoire utilisee est elle comptee correctement ?
  - LHCb veut des jobs a 6GB VMEM
  - Necessite de reduire les ressources en fonction ? Autres solutions a voir

Chantiers en cours (suite)

- Efficacite ALICE

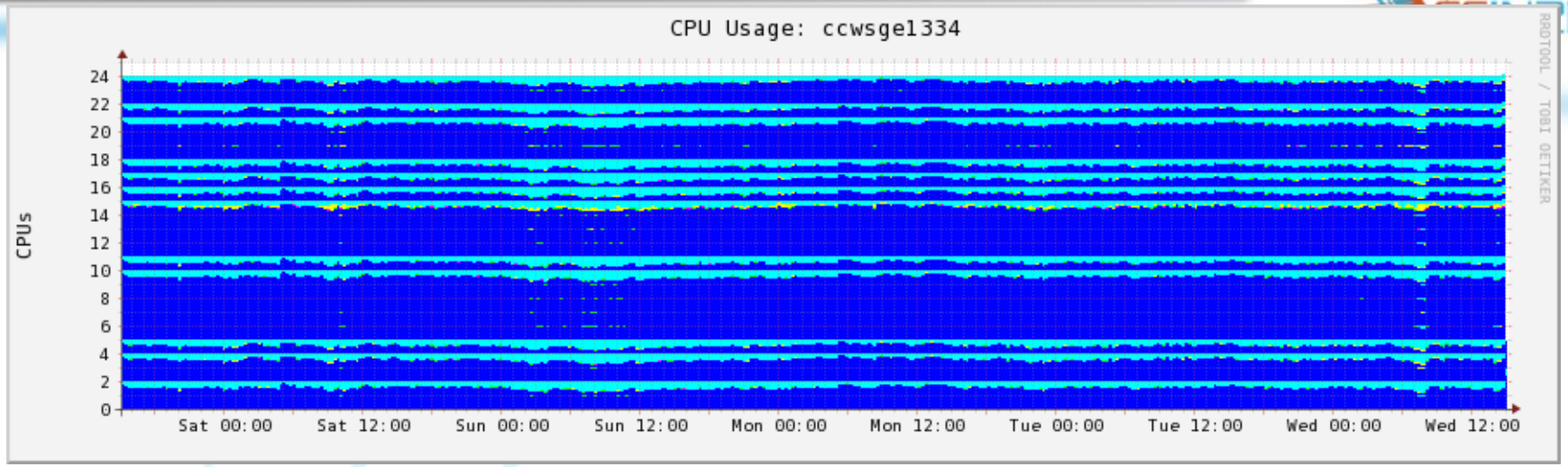


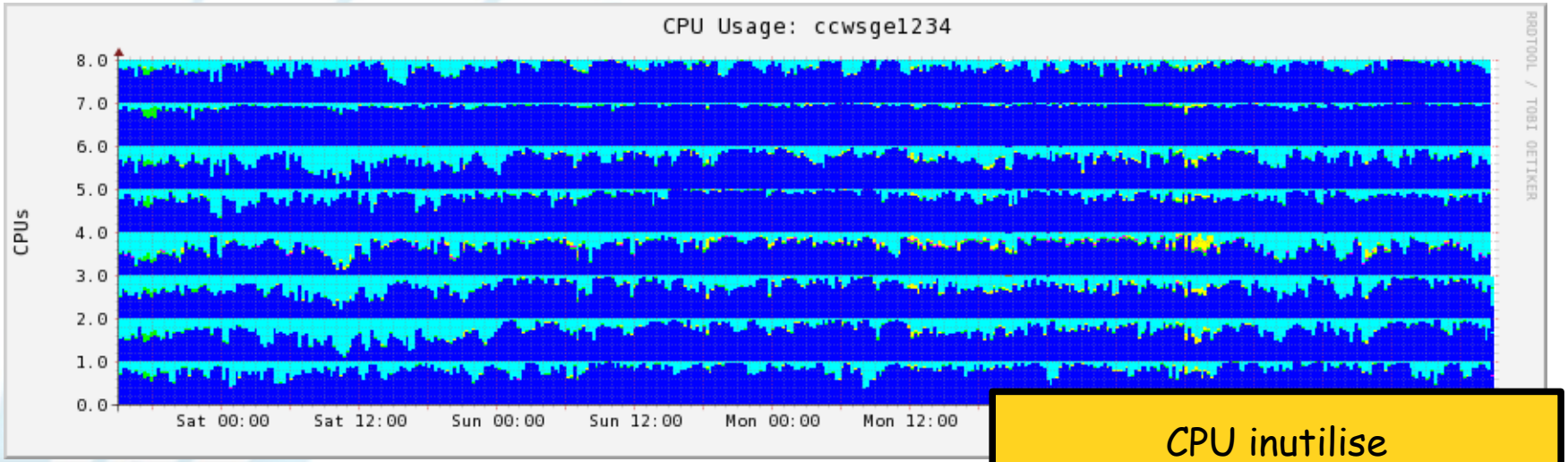
- Tests en cours pour comprendre l'origine
- Ce qui est certain : un job d'analyse lisant des ESD peut avoir une efficacite < 10% (raisonnable pour ALICE)
- Test nuit derniere : seulement des jobs de simu, l'efficacite remonte a 90%

Chantiers en cours (fin)

- Jobs perdus
  - Perte communication de certains jobs CE ↔ WN
  - Souleve regulierement par LHCb → Christelle nettoie manuellement
  - Christelle se charge de faire une etude de ce phenomene
  - Autres VO impactees ?

Utilisation workers





CPU inutilise
Suzanne augmente sueil
suivre deroulement jobs

Notes

- Changement GE :
  - Augmentation des thresholds pour faire passer + de jobs (Suzanne)
  - Sysunix doivent regarder le taux de plantage et cpu_wait
  - Supports LHC : verifions le taux d'echec de nos jobs !
- Fin des iDataplex (8 cœurs)
  - Remplaces la semaine prochaine
    - C6220 → eq. 32 cœurs / machine
    - Meilleure consommation/puissance, disque
- ALICE : manque encore un serveur xrootd
  - Normalement semaine prochaine