25/04/2012

LCG-France Tier-1 & AF

Réunion mensuelle de coordination

Pierre Girard

Pierre.girard@in2p3.fr

dapnia

cea

saclay

CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

Plan

- Chaises musicales…
- Nouvelles de LCG
  - CR du dernier T1SCM
  - CR du dernier GDB
- Disponibilité du site
- Dossiers en cours
- Evénements

Chaises musicales…

- Quand la musique donne…
  - Responsable du projet LCG@CC.IN2P3.FR
    - Il a dit oui…
      - Renaud Vernet
    - Passage en première ligne…
      - en juin 2012 (?)
  - Administrateur grille (sysgrid)
    - Embauche de Vanessa Hamar en CDD de 21 mois
    - Concours “IE” avec profil sysgrid
  - Infrastructure CVMFS
    - Client CVMFS (workers/interactives) et serveurs SQUID
    - Administrés par l’équipe système (syslinux pas sysgrid)
  - Publication de l’accounting Grille
    - Qui reprend le travail de Julien ?

Nouvelles de LCG

CR du dernier T1SCM

- T1 Service Coordination Meeting du 5 avril 2012
  - http://indico.cern.ch/conferenceDisplay.py?confId=185372
  - RAS

CR du dernier GDB

- GDB du 18 avril 2012
  - http://indico.cern.ch/conferenceDisplay.py?confId=155067

Future Shape of the GDB (M. Jouvin)

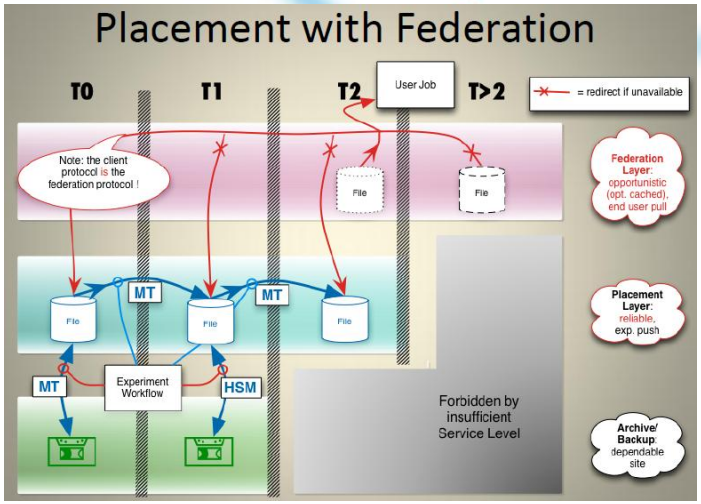
- New chairman: Michel Jouvin
  - http://indico.cern.ch/materialDisplay.py?contribId=1&sessionId=0&materialId=slides&confId=155067
- Nouveautés
  - CR du GDB
  - Agenda du GDB suivant élaboré rapidement à la suite de chaque GDB
- Prochain GDB
  - 9 mai au CERN (début à 9:00 !!!!)
  - http://indico.cern.ch/conferenceDisplay.py?confId=155068
    - Recommendation des TEGs
    - Résumé des C-RSG et CCRB (besoins des expériences dans les prochaines années)
    - Virtualisation des WNs et Cloud

TEG Futures

- The TEG reports
  - https://espace.cern.ch/WLCG-document-repository/Boards/MB
  - Nécessite un compte au CERN

TEG Data Management and Storage

- Placement & Federations
  - Current option is only xrootd
  - Recommendations
    1. HTTP plugin to xrootd
    2. Monitoring of federation network bandwidth
    3. Topical working groups on open questions
       Ex.: separation of read-only and read-write data.
- Point-to-point Protocols
  - GridFTP is ubiquitous and must be supported in medium term
  - Xrootd is currently used alternative
  - HTTP again a serious option
- Managed Transfer (FTS)
  - FTS 3: use of replicas, http transfers, staging from archive
- Separation of archives and disk pools/caches
  - All experiments will split archive (tape) and cache (disk pools)
- Storage Interfaces: SRM and Clouds
  - SRM: Ubiquitous but…
  - Experiment frameworks adapting for alternatives



Source: http://indico.cern.ch/getFile.py/access?contribId=7&sessionId=1&resId=1&materialId=slides&confId=155067

Middleware

WLCG a demandé à EGI/EMI/UMD de clarifier les choses

UMD, EMI, etc

John Gordon, STFC

Content of UMD

- Sites are having to use several repositories (EMI, glite as well as UMD) because required components (eg WMS) have not been included. Can we hear about your policies for inclusion and reasons why certain products are failing. Can WLCG facilitate a plan for resolution?

Source: http://indico.cern.ch/getFile.py/access?sessionId=2&resId=1&materialId=0&confId=155067

Contents of UMD (1)

- EGI Policy
  - UMD will contain products coming through releases from EMI, IGE, …
    - Hydra has not been released as part of EMI (yet!)
      - Though planned for a long time
      - Therefore not part of UMD
  - UMD *will not* contain products from other Technology Providers (for Grid Middleware)
    - That have *not signed* an MoU and SLA
    - But EMI still offers maintenance for gLite 3.2
      - One source of confusion!

UMD (comment)

Testing / Software Quality

- EMI testing vs. EGI testing
  - Different scope and goals!
    - EMI tests seek elimination of bugs
    - EGI tests seek continuous service delivery
- EMI testing covers (at EMI expense)
  - Unit Testing
  - Integration Testing
  - System Testing
- EGI testing covers (at EGI expense)
  - (User) Acceptance testing
  - Typically conducted as smoke screen testing

Quality Criteria, and production condition (Staged Rollout)

All provisioning tests are *publically documented*

EMI News

- **EMI 1 – Update 14 (16 March 2012)**
  - **BDII core v. 1.3.0**
    - Cached Top BDII à installer
- **EMI 1 – Update 15 (20 April 2012)**
  - **CREAM**
    - **BLAH, v.1.16.5 (12 April 2012)**
      - Fixes the job registry corruption caused by the purge of the registry under heavy load
  - **UI/WN tarball support**
    - First testing version available
    - Documentation: https://twiki.cern.ch/twiki/bin/view/EMI/EMIuiwnTar
    - Premier test de déploiement
      - Contient des doublons, des choses inutiles…
      - Version de WN pas certifiée par WLCG
- **EMI 2 (Matterhorn) status update**
  - *SL6/x86_64 – 95% successful build rate*
  - Estimated release date – 7th May 2012

Source: http://indico.cern.ch/getFile.py/access?contribId=3&sessionId=2&resId=1&materialId=slides&confId=15506

Retour de glexec (MUPJ)

Test results for CMS

Site	/cms/Role=pilot job + glexec test Apr 17
ASGC	OK
CERN	OK
CNAF	OK
FNAL	OK
IN2P3	error 203
KIT	OK
NDGF	n/a
PIC	OK
RAL	OK

https://sam-cms-prod.cern.ch/nagios/cgi-bin/status.cgi?servicegroup=SERVICE_CE&style=detail

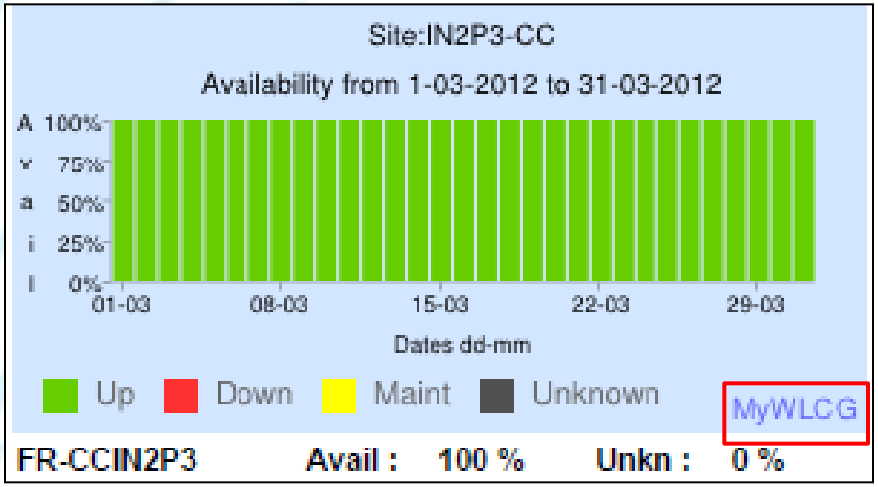
https://sam-cms-prod.cern.ch/nagios/cgi-bin/status.cgi?servicegroup=SERVICE_CREAM-CE&style=detail

Déploiement de glexec à revoir au CC

Résultats du site

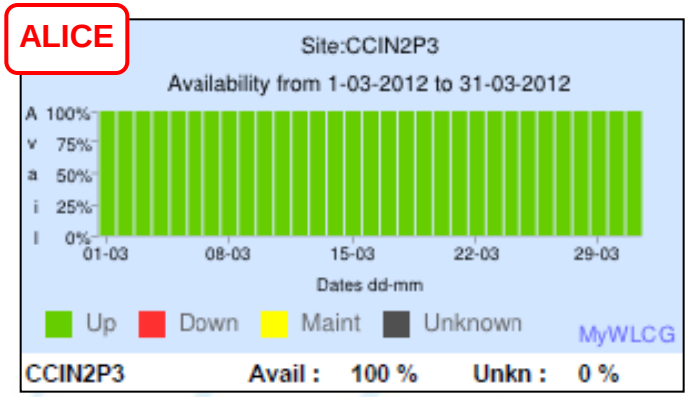
Résultats Mars 2012

- MyWLC IN2P3-CC
  - http://grid-monitoring.cern.ch/mywlcg/sa/?group=1&site=457&graph=1&profile=15
- LCG Office
  - https://espace.cern.ch/WLCG-document-repository/ReliabilityAvailability/Tier-1/2012/tier1_reliab_0312/WLCG_Tier1_Summary_Mar2012.pdf

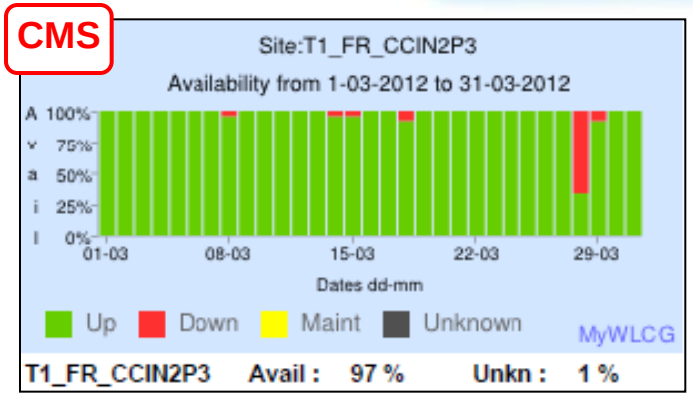


Disponibilités des VOs pour mars 2012

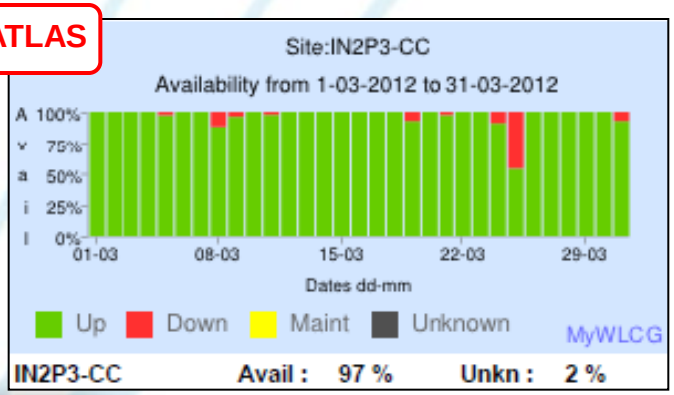
ALICE



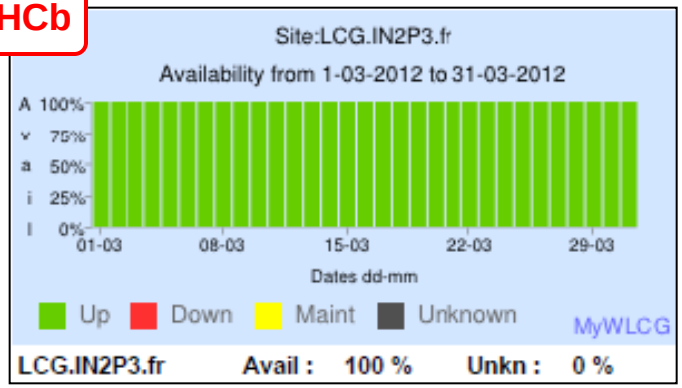
CMS



ATLAS

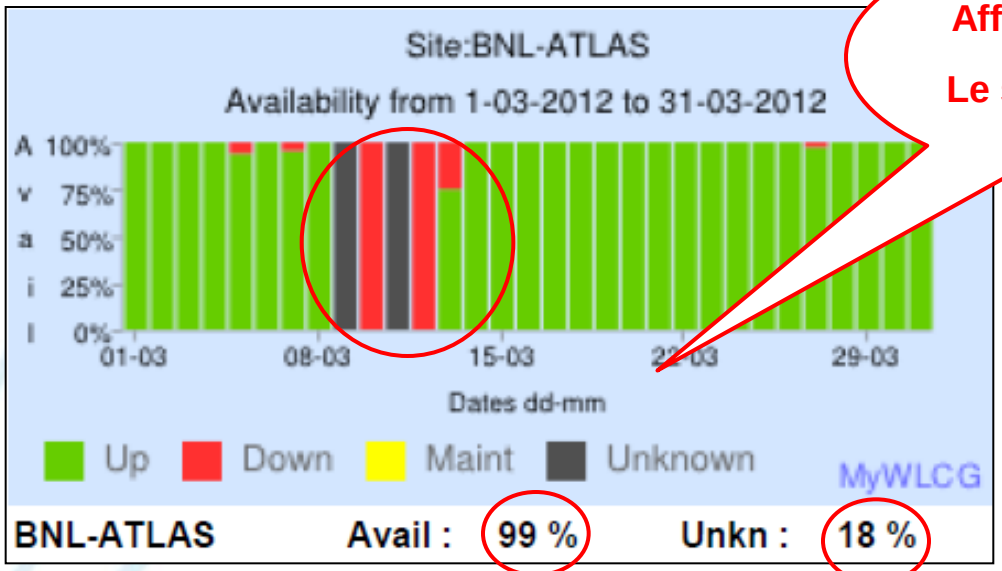


LHCb



Source: https://espace.cern.ch/WLCG-document-repository/ReliabilityAvailability/Forms/AllItems.aspx?RootFolder=%2FWLCG-document-repository%2FReliabilityAvailability%2FTier-1%2F2011%2Ftier1_reliab_122011

Disponibilités de BNL pour Atlas ???



Affichage ou calcul faux ?
Le signaler au MB ?

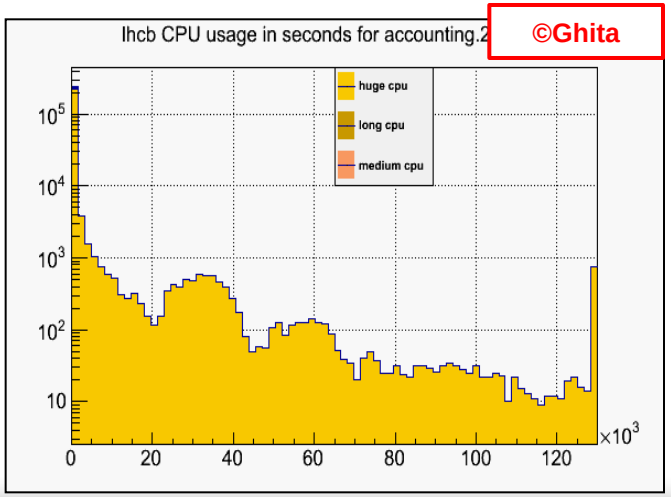
Dossiers en cours

Problèmes avec les jobs

- Multiplication des jobs courts sur GE
  - Passes de scheduling (de + en +) longues
    - Job court = baisse de la production globale
  - ATLAS (résolu)
    - Symptômes
      - Pilot jobs qui tournent à vide
      - Pourtant ATLAS a des « payload » en attente
    - Raison
      - Mauvaise configuration au niveau des queues panda d'atlas
  - LHCb (en cours)
    - Symptômes
      - 90% de jobs qui consomment moins de 400s de CPU
    - Raisons
      - Jobs de certification en pagaille (DN d'Andreï)
      - « Single User Pilot Job » au lieu de « Multi Users Pilot Job »
        - » Pb du pilot job à déterminer le temps CPU restant après un payload
      - Payload avec exit status non nul => fin prématurée du Pilot Job
    - Suivi par cc-lhcb, l'exploitation et LHCb au CERN

Problèmes avec les jobs

- Constat partagé
  - Manque de monitoring
    - Ex.: taux de remplissage de la ferme, taux de jobs courts, etc.
  - Compliqué de se rendre compte ou d'expliquer les relations de cause à effet
- Chacun se fait ses propres outils
  - Avec des requêtes plus ou moins similaires
  - Rarement pérennes
- Ghita propose une solution en attendant mieux



Problèmes avec les jobs

- CMS a rencontré plusieurs problèmes
  - Atteinte du seuil des 3000 jobs par compte
    - Workaround au niveau des CREAMs
      - 2 CREAMs/Tier => 2 comptes différents
      - Dépend de la bonne répartition des soumissions par la VO
      - Dépend du bon fonctionnement des CREAMs
    - Solution plus générale à étudier
      - Au niveau GE ? Au niveau CREAM ?
  - Jobs morts avec status REALLY-RUNNING
    - Bloque les soumissions de CMS
    - A nécessité une purge manuelle au niveau des CREAMs
  - Bug CREAM
    - https://savannah.cern.ch/bugs/?83275
    - There is a problem in all the updater with very short jobs and the notifier will not send any notification to cream about this job.
    - Mise en place d'un EMI-CREAM pour corriger le problème
      - Cccreamceli02.in2p3.fr avec profil T1
      - Update de cccreamceli05 et cccreamceli06 à prévoir

Problème avec les jobs

- Utilisation de la mémoire
  - Mémoire nominale de WLCG: 2 GB / Job
  - Mémoire fixée au CC: +4GB / Job
  - Mémoire réellement utilisée / Job
    - Ghita peut fournir des plots basés sur l'accounting de GE
    - Grosses variations (à surveiller)
  - ATLAS / LHCb ont entamé des discussions
    - Remettant en cause installation/comptabilisation de la mémoire par jobs
    - A surveiller

Problème avec les jobs

- Utilisation de la mémoire
  - ATLAS
    - Propose de mettre autant de swap que de RAM pour absorber des pics à 4 GB
    - Nous ne mettons pas de swap mais avons assez de RAM
    - Cette demande s'adresse plutôt aux T2s
  - LHCb prétend n'utiliser que 3 GB / job
    - Certains de leurs jobs sont tués pour dépassement de 4 GB
    - Etude faite avec Aresh montre que ~1 GB est pris par leur wrapper Dirac
    - Utilisation de Tcmalloc (Thread Caching Malloc)
      - Plus rapide mais pré-allocation de la mémoire même si pas utilisée
    - Discussion autour des concepts de VMEM et RSS
      - LHCb (Stefan Roiser): Peu de mémoire résidente (RAM) mais bcp de VMEM
      - CC (Aresh et moi): aucun moyen de savoir que la VMEM ne sera pas utilisée

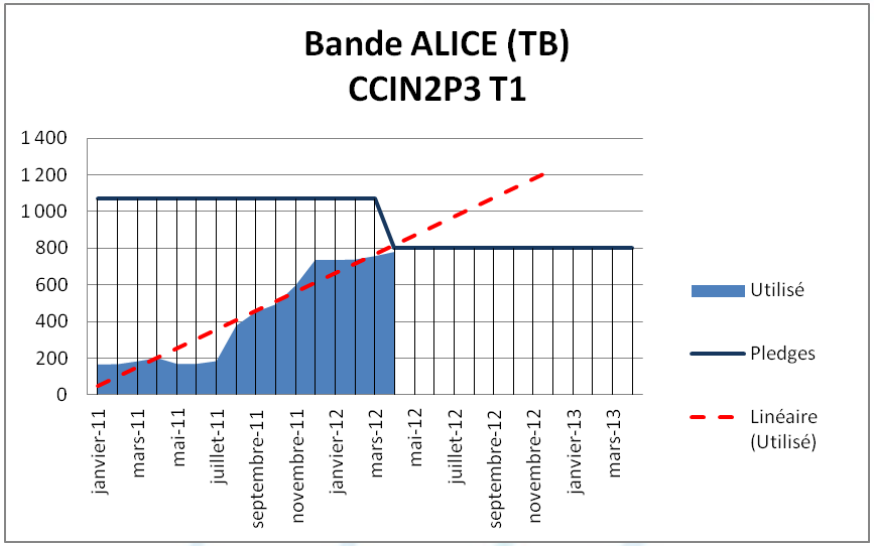
Problème avec les jobs

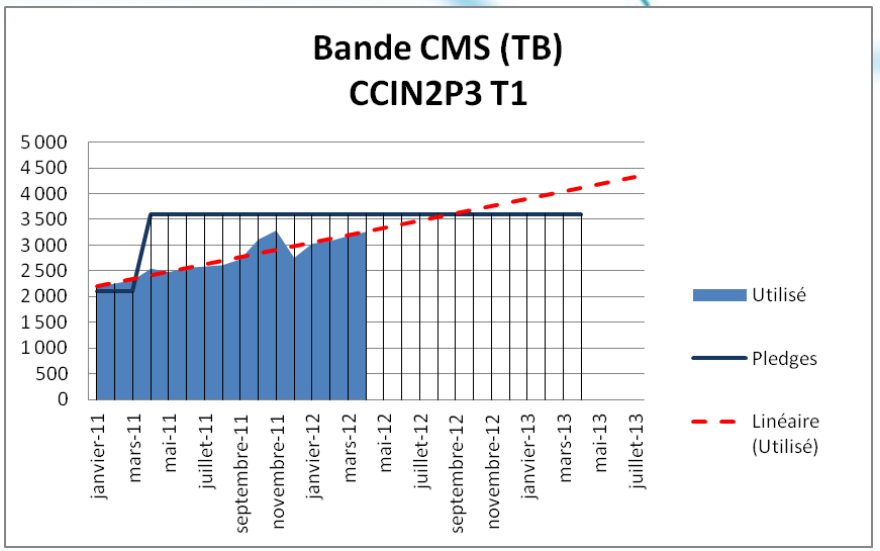
- Utilisation de la mémoire
  - Test naïf pour mieux comprendre le comportement de GE
    - Qsub d'un binaire qui fait un malloc de 4 GB en ne demandant à GE que 3 GB
    - Si on n'adresse pas la mémoire allouée
      - non vu comme vmem
      - le job passe
    - Si on adresse la mémoire allouée
      - vu comme vmem
      - le job est tué par le système (rlimit)
  - A étudier, l'apport de la virtualisation
    - Chaque VO sait ce qu'elle utilise réellement comme RAM par job
      - Allocation de la RAM par VM sur la base des prétentions de la VO par jobs
      - Plus la RAM nécessaire au système
    - Intérêts
      - Taillage au plus juste de ce que demande/prétend la VO
      - Cloisonnement dans la VM impactant pas les autres expériences

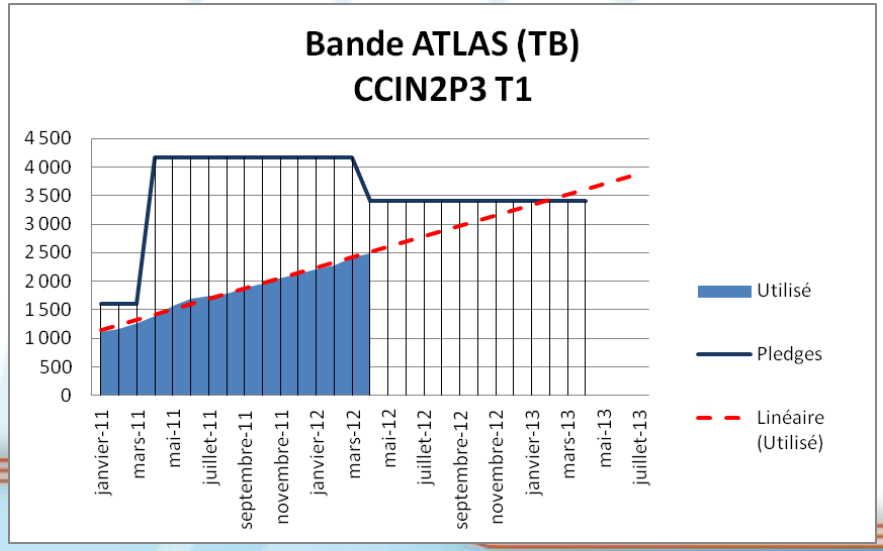
Utilisation du stockage

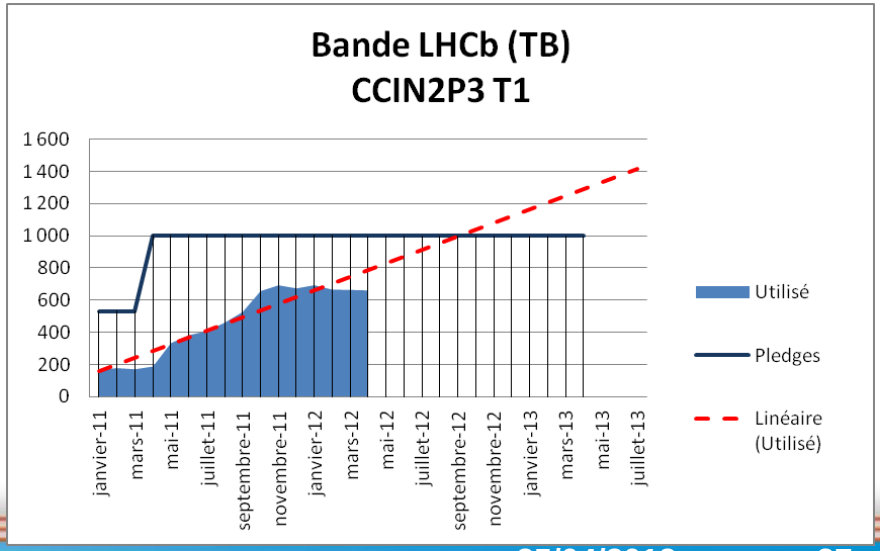
- Accounting LCG-Fr@CC.IN2P3.FR
  - En attendant mieux
    - https://grid.in2p3.fr/LCGFrAccounting/
    - Mais discussion en cours
      - Décisionnel pour convergence sur les valeurs du CPU
      - Groupe stockage
  - On y trouve
    - Les Pledges (à jour) depuis 2009
      - Attention: en TB pas en TiB
    - Accounting
      - Pour les années 2011 et 2012
      - Pour CPU, Disque et Bande
      - Pour les activités T1/T2/T3(local+grid)

Utilisation du stockage / Bande









Utilisation du stockage / Bande

- Mais…
  - Vérifications faites avec Sébastien pour CMS
    - L'activité T3 (locale) utilise de la bande
      - Directement
      - Ou via xrootd
    - Non négligeable en temps de restriction
      - ~100 TB
  - Même phénomène pour les autres VOs
  - A venir
    - Estimation de la part T1 / T2
    - Discussion sur l'utilisation de la bande dans le cadre du T3
      - A priori que du disque pledgé
      - Mais moindre cout pour la bande
      - Diminution possible du disque au profit de la bande ?

Utilisation du stockage / Disque

Avril 2012
T1+T2+T3

Total	alice	Usage	installed	935,68	898,30	898,30	898,30
			used	805,14	825,61	781,94	818,02
			occupancy	86,05%	91,91%	87,05%	91,06%
		Pledges	capacity	915	915	915	920
			%installed	102,26%	98,17%	98,17%	97,64%
	atlas	Usage	installed	5 340,20	4 908,21	5 314,71	5 279,53
			used	3 905,34	3 891,26	3 984,74	4 148,46
			occupancy	73,13%	79,28%	74,98%	78,58%
		Pledges	capacity	5 316	5 316	5 316	5 320
			%installed	100,46%	92,33%	99,98%	99,24%
	cms	Usage	installed	1 850,35	1 791,64	1 844,96	1 844,96
			used	1 063,19	1 095,18	1 028,80	1 139,43
			occupancy	57,46%	61,13%	55,76%	61,76%
		Pledges	capacity	1 872	1 872	1 872	1 875
			%installed	98,84%	95,71%	98,56%	98,40%
	lhcb	Usage	installed	1 096,87	1 096,87	1 096,98	1 096,98
			used	495,44	505,78	581,42	612,21
			occupancy	45,17%	46,11%	53,00%	55,81%
		Pledges	capacity	1 090	1 090	1 090	1 090
			%installed	100,63%	100,63%	100,64%	100,64%

Utilisation du stockage / Disque

- Attention
  - Nous utilisons un modèle de coût basé sur l'achat de disque DAS (dCache, xrootd)
  - Pour le T3, nous déployons aussi des solutions plus chères (GPFS)
  - Nous ne pouvons/devons plus/pas assurer le volume fixé par LCG France
- Discussion avec Atlas
  - Souhaitait 300 TB de GPFS cette année
  - Compromis
    - Décommisionning des machines XROOTD Atlas (100TB)
    - Ajout de 30 TB dans GPFS
    - Test d'un serveur SUN Thor dans GPFS
      - Avec une politique de stockage des fichiers les moins accédés

Problème réseau

- Tests en cours entre CC et IHEP
  - Assymétrie
    - IHEP->CC « excellents »
    - CC->IHEP « mediocres »
  - Yvan, Jérôme et Fabio
  - Mise en cause de l'IP Bonding
    - Aggrégation (2 x 1Gbps) des cartes réseau sur les serveurs
    - Test en supprimant une carte réseau
      - Pas d'amélioration
      - Hypothèse invalidée
  - Analyse des paramètres TCP (en cours) de part et d'autre
    - Yvan ?

Tests en cours / Veille technologique

- Soumission Multi-Cœur
  - cccreamceli01
  - Machines réelles et virtuelles
    - Contact « syslinux » le tuning de la virtualisation
      - Aurélien Gounon
  - CMS et ATLAS
    - ATLAS ne fait pas réellement de tests et demande de mettre en production le CREAM
    - Proposition:
      - Arrêter les tests avec ATLAS ?
      - Faire entrer ALICE dans les tests
  - Suspicion de problème GE avec les jobs MC ?
    - Status ?

Tests en cours / Veille technologique

- Serveur CHIRP pour ATLAS
  - Installé sur une VO Box
  - Backend GPFS
- Worker SL6
  - Leader: groupe système
  - Contact syslinux: Yannick Perret
  - Utilisant la couche glite-WN 3.2 (SL5)
  - SAM test ok
- Introduction d'un Serveur Thor dans GPFS pour Atlas
  - Leader: groupe stockage (J.Y. Nief)
  - Contacts: Yannick Perret et Loïc Tortay
  - Vu en production par Atlas T3

Déploiement Grille (production)

- Recommendations WLCG
  - https://twiki.cern.ch/twiki/bin/view/LCG/WLCGBaselineVersions
- Migration vers EMI à faire (Urgent)
  - Argus (pour glexec)
  - BDII_site et BDII_top
  - CREAM (à finaliser et à updater)
- Update glite à faire (Urgent)
  - UI/WN
  - VOBOX
    - Demande d'Alice
    - Nouveau service proxy_renewal

Evénements

- A venir
  - HEPiX Spring Meeting, Prague, 23-27 April 2012
  - EMI AHM, DESY, 8-10 May
  - WLCG Workshop, NYC, 19-20th May 2012
  - CHEP2012, NYC, 21-25th May 2012
  - LCG France, Nantes (SUBATECH), 6-7 June 2012
    - http://indico.in2p3.fr/conferenceDisplay.py?confId=6455
  - EGI Technical Forum, Prague, 17-21September 2012
  - HEPiX Fall Meeting, IHEP, 15-19 October 2012
  - JI IN2P3/IRFU, La Londe les Maures, du 22 au 25 octobre 2012
  - *Workshop « Placement & Federations » (xrootd), à l'automne 2012, au CCIN2P3*