LCG-France Tier-1 & AF

Réunion mensuelle de coordination

Fabio Hernandez
fabio@in2p3.fr

Lyon, 10 décembre 2009

Bienvenue

- Bienvenue à
  - Eric Cogneras (ATLAS)
  - Damien Mercier (CMS)
- … qui nous ont rejoint le 01/12/2009

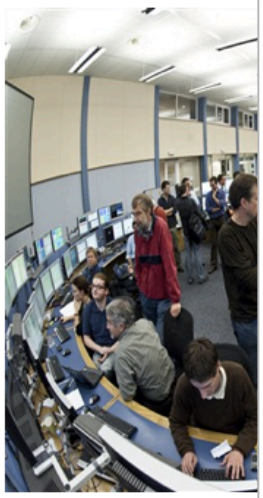
LHC – Redémarrage

Le Monde.fr | CRÉEZ VOTRE BLOG

EN QUÊTE DE SCIENCES

« Une météorite dans le ciel de l'Utah
20 novembre 2009

Les particules sont de



L'heure du week-end n'est pas encore venue pour les employés du CERN. Vendredi en fin d'après-midi, les physiciens ont remis en route le Large hadron collider (LHC), plus grand accélérateur de particules au monde. Pour la première fois depuis son inauguration et ses quelques heures de fonctionnement en septembre 2008, des injections de particules ont eu lieu, l'espace de quelques fractions de seconde, dans les anneaux du LHC.

guardian.co.uk

News | Sport | Comment | Culture | Business | Money | Life & style | Travel | Environment

News › Science › Cern

SCIENCE BLOG

Previous | Blog ho

First image of particle collisi Cern's £6bn atom smasher

The Large Hadron Collider has started crashir together, albeit at low energies. Here is the firs by one of the machine's giant detectors

BBC NEWS | News Front Page

http://news.bbc.co.uk/

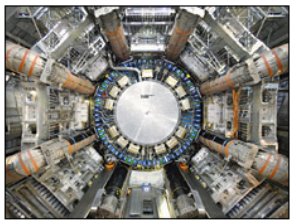
BBC | Low graphics | Help | Search

NEWS | Watch ONE-MINUTE WORLD NEWS

News Front Page

Page last updated at 07:14 GMT, Saturday, 21 November 2009

LATEST: _



Large Hadron Collider works again

The Large Hadron Collider experiment, designed to shed light on the cosmos, restarts after 14 months of repairs.

- Cheers of relief at Cern HQ
- In pictures: Machine reboots
- Large Hadron Collider: Guide

Sri Lanka refugees | Childhood abuse

Near Geneva, Particles Finally Come Together With A Bang - NYTimes.com

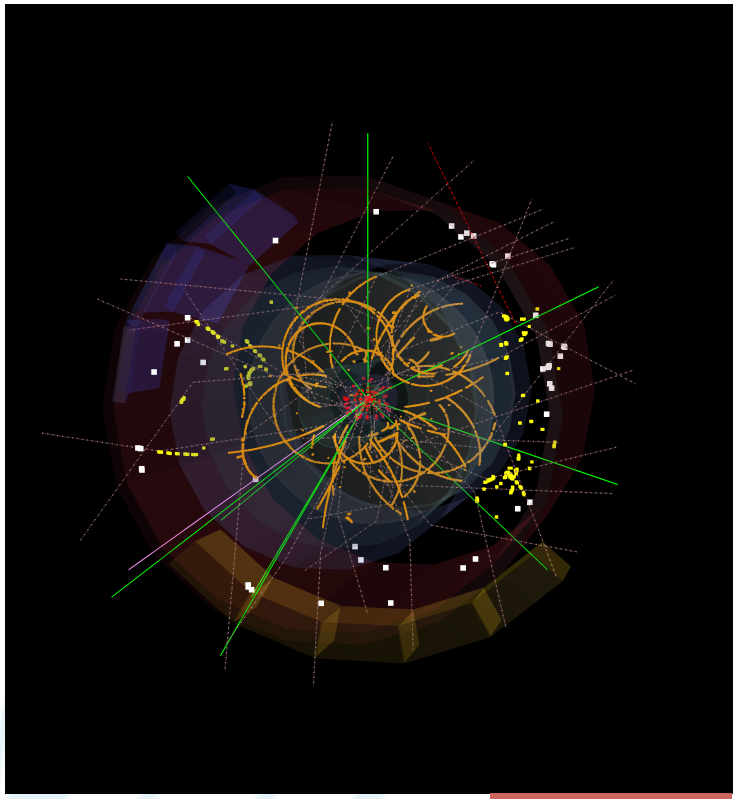
Call it First Bang.

The Large Hadron Collider, the world's biggest and most expensive science experiment, produced its first collisions on Monday, said scientists at CERN, the European Organization for Nuclear Research, outside Geneva.

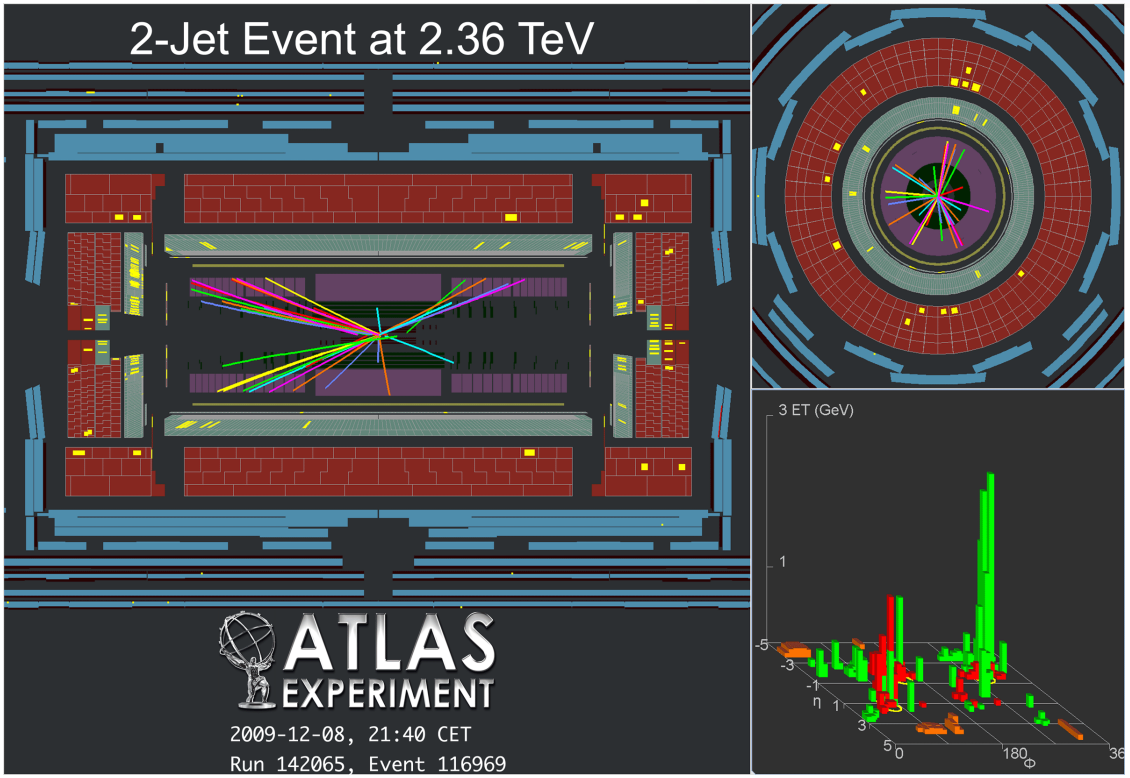
Seemingly making up for lost time after years of disasters and delays, the collisions came only three days after engineers had begun shooting the subatomic particles known as protons around their 17-mile underground racetrack. The physicists announced that they had succeeded in making the beams collide, producing what they called "candidate collision events" in the giant particle detectors in the collider.

The collider has been built over 15 years at a cost of $9 billion to accelerate protons to energies of seven trillion electron volts apiece and then slam them together in an attempt to recreate forces and particles that reigned during the first moments of the Big Bang. But for much of that time, the only things that have gone bang in the collider were magnets and other components, most notably in September 2008 after the first time protons circled the collider.

LHC - Collisions

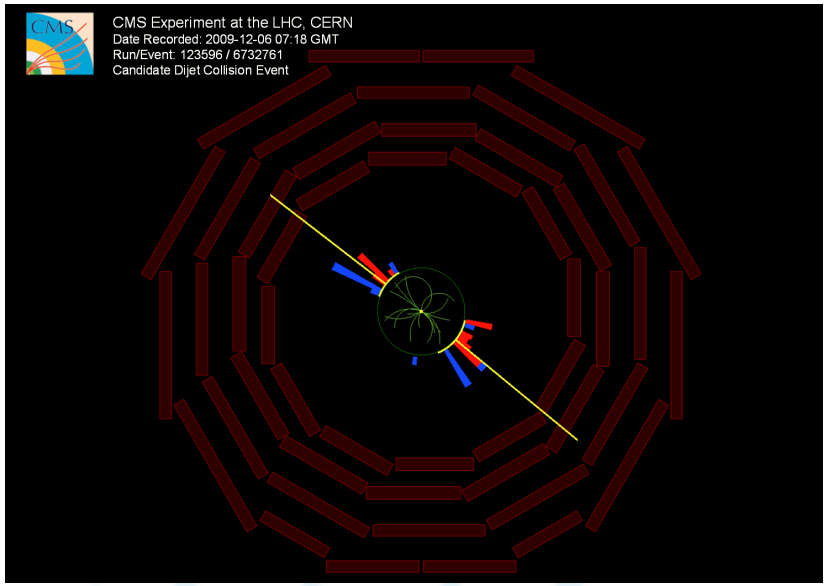


ALICE

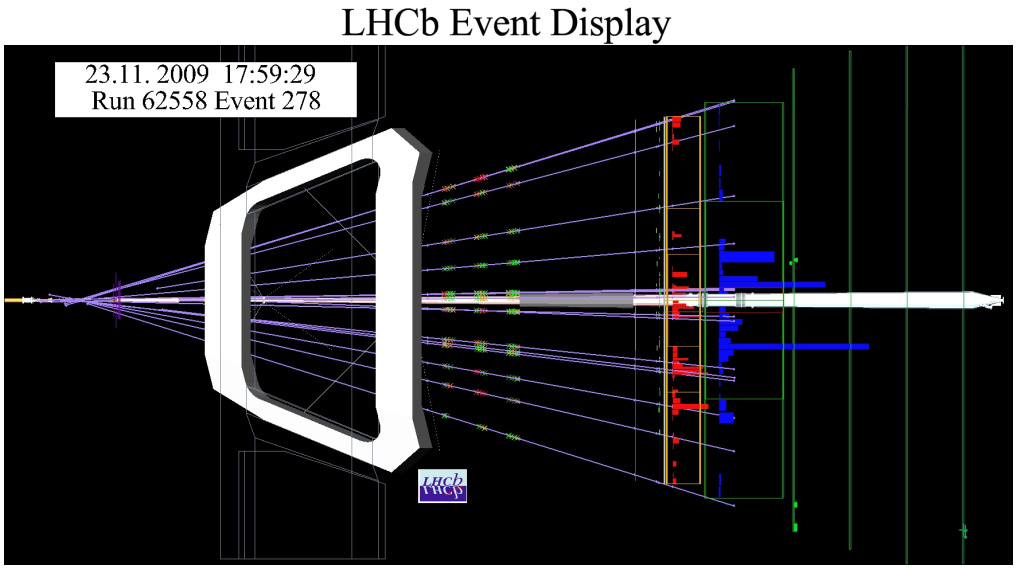


ATLAS

LHC – Collisions (suite)



CMS



LHCb

Avancement des chantiers

Computing Elements & ferme batch

- Re-configuration des coefficients de la ferme, en prenant en compte la puissance CPU mesurée des worker nodes
- Re-fonte du mécanisme de publication des données d'accounting CPU du site vers le portail EGEE
  - Uniquement les informations relatives aux jobs grid sont publiées
  - Cohérence retrouvée avec les rapports internes
  - L'accounting nominatif (par DN) est en place
- Limite maximum en mémoire de la classe G montée à 2GB
  - Utilisée par les jobs LHCb
- Séparation des CEs pour les VOs LHC et les VOs EGEE
  - Voir slide suivant

Computing Elements

Tier Level	CE hostname [.in2p3.fr]	ALICE	ATLAS	CMS	LHCb
Tier-1	cclcgceli02	✓	✓		
	cclcgceli04			✓	✓
	cclcgceli07	✓	✓		
	cclcgceli08			✓	✓
Tier-2	cclcgceli06	✓	✓	✓	✓
	cclcgceli09	✓	✓	✓	✓

Les computing elements sont configurés pour les expériences LHC exclusivement. Un jeu de CEs séparé est configuré pour le support des VOs EGEE.

Les jobs grid des expériences LHC utilisent exclusivement les worker nodes sous SL5.

Source: https://grid.in2p3.fr/html/Private/index.php?id=Mapping

Avancement – Ferme batch

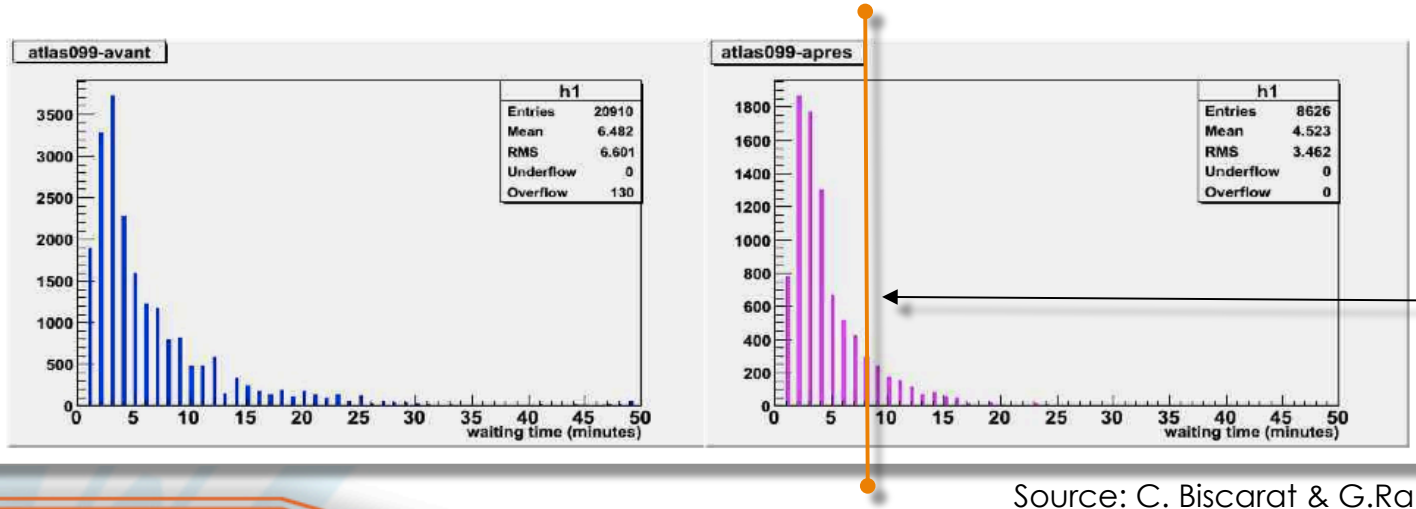
- Modification de la configuration de la ferme
  - Un point d'exécution dédié pour la classe G sur chaque worker node
  - Amélioration mesurée du turnover des jobs d'analyse ATLAS

Avancement – Ferme batch (suite)

Reactivite aux jobs d'analyse (cont'd)

Jobs d'analyse ATLAS099

- Comparaison des temps d'attente des jobs atlas099 pour des jobs tournés avant et après le changement.
  - Résultats:
    - Amélioration très notable: 4.5 minutes d'attente en moyenne après le changement contre 6.5 minutes avant le changement.
    - Le « spread » est bien meilleur: rms de 3.4 après contre 6.6 avant. L'utilisateur récupère plus vite son bloc de jobs.



Bilan des résultats:

• Réduction du temps d'attente en queue des jobs d'analyse

•Exécution groupée des jobs d'analyse: l'utilisateur final obtient les résultats de l'ensemble de ses jobs plus rapidement

68% des jobs d'analyse sont mis en exécution en moins de 8 min (μ + RMS) contre 13 min auparavant

Source: C. Biscarat & G.Rahal, Réunion CAF 08/12/2009
http://indico.cern.ch/conferenceDisplay.py?confId=69944

Avancement - Stockage

- Mise à jour de dCache
  - Passage à la version 1.9.5 (a.k.a. Golden Release) le 09/11/2009
  - Problèmes observés avec les outils de monitoring et l'importation de données ATLAS (généralisation de gridFTP2 a causé des problèmes)
- Allocation disque (dCache) en conformité avec les besoins des expériences et les engagements du site
- Métriques de l'activité tape staging publiées via SLS du CERN
  - https://sls.cern.ch/sls/service.php?id=WLCG_FR_CCIN2P3_Tape_Metrics
  - Métriques d'écriture de données LHC dans HPSS en cours de préparation

Avancement - stockage

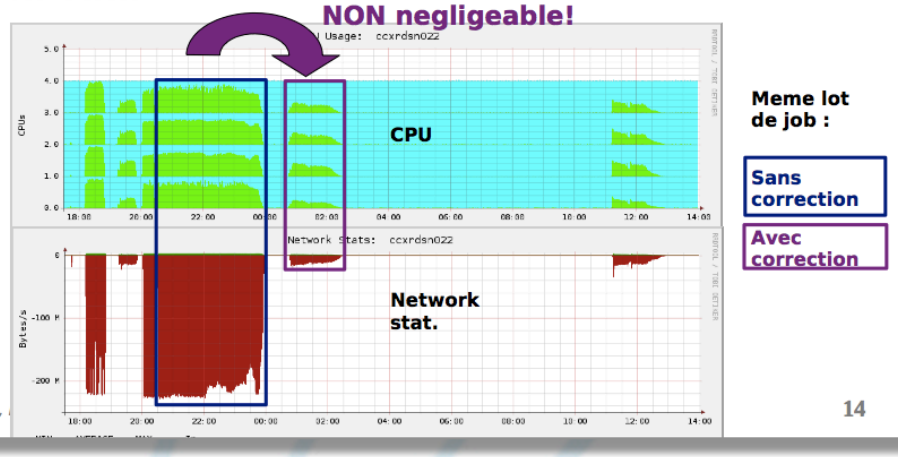
- Procédure de vérification régulière de cohérence du catalogue dCache avec les fichiers effectivement sur disque
  - Fréquence: 1 fois par mois, depuis 01/11/2009
  - ATLAS: 14 fichiers sur le catalogue, non trouvés sur disque
    - *Cause de l'incohérence comprise et correctif implementé*
    - *Détails: http://cctools2.in2p3.fr/elog/dCache/387*
  - Quel est le résultat de cette vérification sur les fichiers des 3 autres expériences?

Avancement – Stockage (suite)

Performances Xrootd

Bug de root transferts network :

- Saturation du reseau et donc ralentissement des jobs xrootd car mauvaise configuration de root dans les releases atlas (aussi reporte a SLAC)
- Solution : corriger a la main les parametres Xnet.Read dans le fichier system.rootrc des releases



CAF,

14

Performances xRootd (cont'd)

AODs modifies :

- AODs modifies de Ilija and co'
  - Re-ordonnacment des baskets
  - Optimisation des basket size
  - ... [1]
- Tester au CC (re-ordonnacement et optimisation des BS) : gain en temps d'execution wall-clock d'un facteur 6 sur le processing des jobs athena
  - Voir talk separe pour les details
- Decision dans ATLAS prise d'implementer ces changements dans les releases pour ecrire les AODs directement avec nouvelle structure :
  - Pour le MC et le reprocessing (rel 15.6.x, 16.X.0)
  - Seulement le re-ordonnacement (majeur partie du gain en temps, pas trop CPU consumming); eventuellement split-0
- [1] EDM meeting, Tuesday December, 1st, 2009

CAF, Nouvelles du CC, 18/11/2009 16

Suite de l'étude des faibles performances xrootd observées pendant STEP'09 pour les jobs d'analyse ATLAS:
• Reconfiguration au niveau du file system (ZFS) des serveurs xrootd
• Mise en évidence d'une mauvaise configuration de ROOT dans le software ATLAS: modification centralisée ou au niveau du site nécessaire
• Décision prise par ATLAS de modifier le format de données pour optimiser la lecture

Source: C. Biscarat & G.Rahal, Réunion CAF 08/12/2009
http://indico.cern.ch/conferenceDisplay.py?confId=69944

Avancement depuis la dernière réunion (suite)

- Utilisation du nouveau mécanisme d'installation et réplication automatique du logiciel ATLAS sur AFS
  - Après un intense travail de débogage de l'outil d'installation (ATLAS) et de réplication des volumes AFS (CC)
- Visite des responsables computing CMS au site
  - Agenda: http://indico.in2p3.fr/conferenceDisplay.py?confId=2357
  - Points identifiés à améliorer: stabilité du site, performances des transferts des données (importation & exportation)
  - Farida et moi suivons les actions identifiées suite à cette réunion: beaucoup de points ont avancé depuis
- CMS: séparation terminée des zones de stockage sur disque (dCache) pour le tier-2
  - Zone disk-only, sans possibilité de staging de fichiers depuis HPSS
  - Tier-1 & tier-2 partagent la même software area (AFS)

Avancement depuis la dernière réunion (suite)

- User Interface
  - Mise à jour de gLite UI pour SL5: toutes les machines interactives du site ont cette nouvelle version
- VO boxes sous SL5
  - ALICE: machine livrée
  - ATLAS: en préparation
    - *Personnalisation de l'installation nécessaire pour les besoins de l'expérience (durée du proxy, nombre d'utilisateurs, …)*
- Utilisation plus systématique des tickets Xoops pour les demandes d'intervention
  - Tel que convenu lors de la réunion de octobre

Interventions Programmées

Interventions

- Les interventions programmées et déclarées AT_RISK ne sont pas prises en compte dans le score de disponibilité du site par GridView
  - Ceci impacte la disponibilité du site principalement lors des interventions sur l'instance dCache/SRM pour LHC (puisqu'il n'y a pas de redondance)
  - Source "EGEE Intervention Procedures": https://edms.cern.ch/document/1032984

- En attendant une éventuelle clarification des procédures, les interventions sur dCache doivent être déclarées comme OUTAGE, même si le service sera maintenu en mode dégradé

Interventions (suite)

- 22/12/2009 Robotique
  - Interruption du service pour un nettoyage complet des robots
  - Durée: 6 heures
- 22/12/2009 Maintenance HPSS
  - Interruption de service
  - Durée: 5 à 6 heures
- 04/01/2009 Mise à jour dCache LHC
  - Mise à jour système des 4 serveurs centraux
  - Mise à jour des certificats des 4 serveurs centraux
  - Mise à jour logiciel: v1.9.5-6 à v1.9.5-9
  - Durée: “quelques heures”
- 2ème moitié de janvier
  - HPSS: reconfiguration pour introduction de noveaux sous-systèmes

Chantiers en cours ou à venir

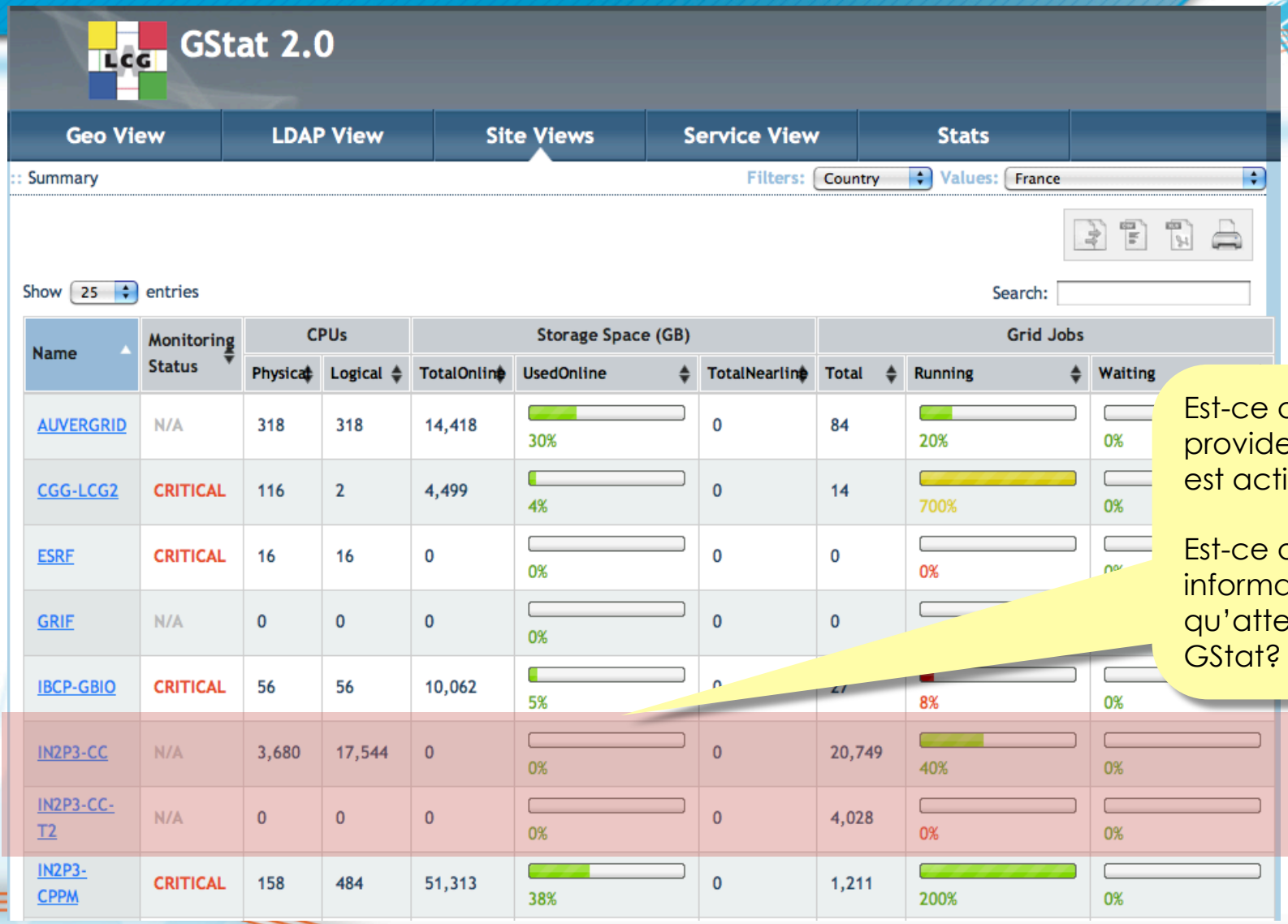
Stockage

- Protection de données des expériences contre les effacement involontaires
  - ACLs?
  - Read-only?
- Renommage (ou plutôt mapping) des espaces de stockage dCache en fonction de l'utilisation actuelle
  - Ex. Production, analyse, merge, import-export, tape-buffer, …
- ATLAS: déploiement de FroNTier + Squid pour l'accès aux bases de données de conditions par les jobs des sites tier-2s du nuage

Computing Element

- Finalisation de la mise en production d'un CE CREAM interfacé avec BQS
  - Détails: http://cctools2.in2p3.fr/baseConnaissance/php/affichage/FAQDetails.php?id_faq=930&intranet=cc_bc&langue=fr
  - Objectif initial: 1 CE CREAM pour le tier-1 et un autre pour le tier-2, les 2 supportant les 4 expériences LHC
  - Période de cohabitation nécessaire avec les CEs LCG estimée à plusieurs mois
- Augmentation significative des comptes d'exécution des jobs grid
  - En particulier pour le tier-2
- Système d'information du site
  - Information publiée sur la puissance CPU installée, le nombre de processeurs, etc. à stabiliser
  - Information Provider BQS à remettre d'applomb

Site Information System



Est-ce que l'information provider de dCache est activé?

Est-ce qu'il publie les informations telles qu'attendues par GStat?

Source: Gstat 2.0, 09/12/2009
http://gstat-prod.cern.ch/gstat/summary/Country/France/

Computing Element (suite)

- gLexec & SCAS
  - Dossier du déploiement de gLexec sur les worker nodes, pour le support des jobs pilots à rouvrir

Support de Jobs Pilote

MB Statement

GDB

- It will be acceptable for the experiments to run multi-user pilot jobs **without requiring identity switching for a period of 3 months** (i.e. until end of February 2010)
  - o This means that we are temporarily and exceptionally **suspending the identity-switching requirement of the existing JSPG Policy on Grid Multi-User Pilot Jobs:** https://edms.cern.ch/file/855383/2/PilotJobsPolicy-v1.0.pdf;
  - o During this time problems with workloads at a site will be the responsibility of the VO i.e. **The entire VO could be banned from a site;**
  - o The situation will be reviewed after **3 months, or earlier** if needed due to operational or other circumstances.
- **The deployment of glexec and SCAS (or equivalent) should proceed as rapidly as possible at all sites**. The versions of both components for SL5 are now available. Note that SCAS is the solution for EGEE sites, other implementations of this function may be implemented elsewhere, but glexec must be deployed as the interface for the pilot jobs to use.
- The experiments that propose to submit multi-user pilot jobs should endeavour to ensure that their **frameworks make use of these tools on this same timescale.**
- We will implement **a test to validate** the availability and usability of a glexec installation at a site.
- The long term policy requirements of traceability and the ability to ban individual users from a site remain unchanged, but we agreed to **start a review** of how these requirements could be better managed and implemented in the long term to satisfy the needs and constraints of sites and experiments.

Science & Technology Facilities Council

2

Source: John Gordon, GDB 02/12/2009
http://indico.cern.ch/conferenceDisplay.py?confId=64669

Analyse du service batch pour LHC

- Analyse systématique du service batch rendu pour chacune des expériences LHC
  - Objectifs:
    - *identifier les paramètres de la configuration de la ferme à modifier pour améliorer le service*
    - *nous assurer que la configuration de la ferme s'ajuste le mieux possible à la demande*
  - Mesurer le temps d'attente et le temps de service (temps en queue et temps d'exécution)
  - Mesurer la mémoire et le temps CPU effectivement consommé par les jobs
  - Différentier les jobs tier-1 et tier-2
  - Point mensuel dans le cadre de cette réunion

Ferme Batch – accounting et monitoring

- Depuis 25/11/2009, les jobs grid sont étiquettés dans BQS
  - Attribut '`sitetype`' qui prend la valeur '`tier1`' ou '`tier2`' en fonction du CE de soumission
  - Nous pourrons exploiter cette information pour:
    - *Accounting mensuel CPU*
    - *Monitoring batch (jobs en queue et jobs en exécution) par site (tier-1 ou tier-2)*

CREAM CE

Coming soon

Enabling Grids for E-sciencE

CREAM CE for sl5_x86_64 / gLite 3.2

- Patch #3260 (current state "Rolling-out")
- For what concerns the CREAM/CEMon/BLAH software it is basically the same software used for gLite 3.1 / sl4_i386
  - There is just an extra fix required because of a different behavior of the delegation software between gLite 3.1 and gLite 3.2
- Will be released along with gLite-Torque-Server and gLite-torque-utils
- gLite-LSF-utils for gLite 3.2 / SL5_x86_64 in preparation
  - Found an agreement between INFN and CERN for its support

Coming soon (cont.ed)

Enabling Grids for E-sciencE

- **Fix for a vulnerability problem in BLAH regarding the "forwarding of requirements to the batch system" feature**
  - Risk (as classified by GSVG): Moderate
  - Patch #3289 (current state "Rolling-out")
- **New yaim-cream-ce addressing a couple of problems for which a patch was asked**
  - In particular to allow the check for "the installed capacity"
  - Patch #3438 for gLite 3.1 / sl4_i386 (current state "Rolling-out")
  - Patch #3439 for gLite 3.2 / sl5_x86_64 (current state "Rolling-out")

D'autres détailes à ce sujet dans la présentation de Massimo. Ils devraient intéresser Sylvain et les CE Masters.

Source: Massimo Sgaravato, GDB 02/12/2009
http://indico.cern.ch/conferenceDisplay.py?confId=64669

EGI – Réponse à l'appel à projets soumise à l'UE

EGI Global Tasks

- GOCDB
  - UK
- APEL & Accounting portal
  - UK & Spain
- CIC Portal
  - France
- EGI Helpdesk (GGUS)
  - Germany
- Monitoring framework
  - CERN, Greece & Croatia

GDB - December 2009 6

Operational Tools Development

- Coordinated by Italy
- Background maintenance activity
  - CIC Portal, Helpdesk, GOCDB, Accounting Server & Portal, SAM, Metrics Portal
- Year 1: Continued regionalisation
  - CIC portal, GOCDB, Accounting, SAM
- Year 1-3: CIC Portal
  - Regionalisation, new resource types, ...
- Year 2-4: Accounting for new resources
  - Cloud resources, applications, etc.

GDB - December 2009 12

Parmi les tâches internationales de EGI, le développement et l'opération du portail d'opérations grid est la seule responsabilité pour laquelle la France s'est portée candidate

Source: Steven Newhouse, GDB 02/12/2009
http://indico.cern.ch/conferenceDisplay.py?confId=64669

Alertes sécurité

- La procédure de déploiement des palliatifs et correctifs suite aux alertes sécurité n'est pas suffisamment claire
  - Les informations diffusées en interne au site et communiquées à WLCG/EGEE ne sont pas toujours concordantes et sont souvent partielles

- En comité de direction, il a été demandé que les actions déclenchées par ces alertes soient consignées dans l'intranet jusqu'à ce que les correctifs soient intégralement appliqués
  - Cela devrait permettre aux parties prenantes de partager la même information

Incident DNS du 08/12/2009

- Démon de résolution de noms des machines derrière un alias en prenant en compte la charge (lbnamed) dévenu inopérant sans raison apparent

- Analyse de l'incident en cours
  - https://cctools2.in2p3.fr/operations/wiki/doku.php?id=incidents:incidentreseau0812009

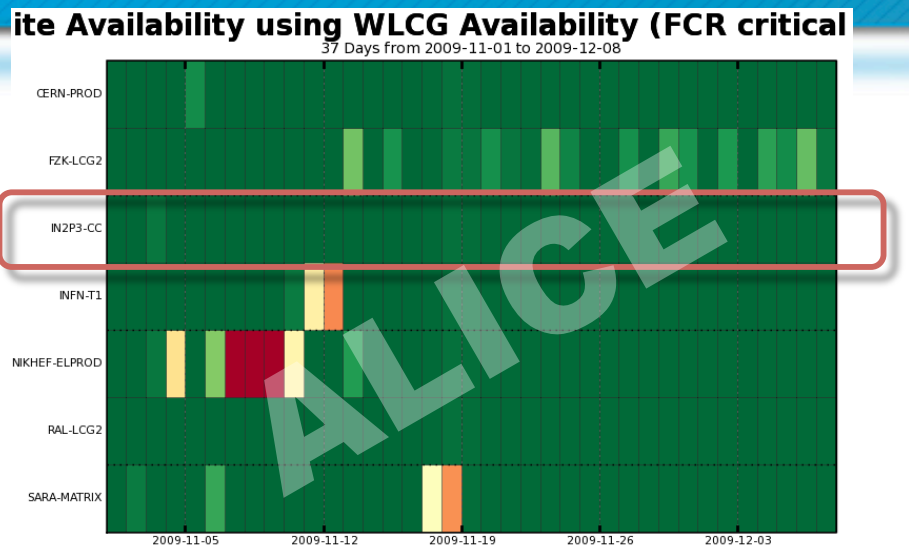
Documentation

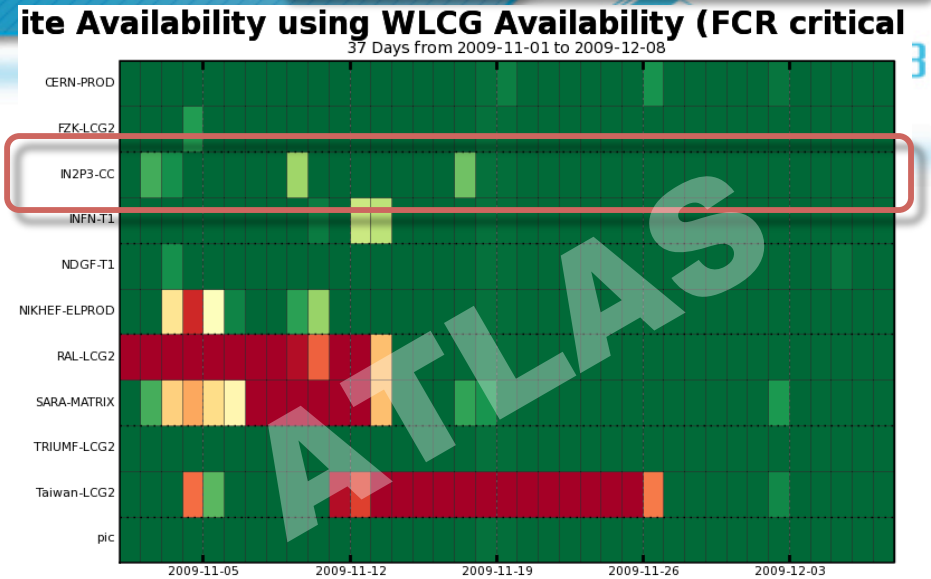
- Description des principes de fonctionnement des différents services grid et de l'utilisation qui en faite par les expériences LHC
  - VO boxes, dCache, FTS, LFC, CEs, software area, système d'information, …
  - Point d'entrée pour avoir une une vue d'ensemble des services
    - *Par exemple, pour les nouveaux arrivants, pour le support général, pour l'exploitation, etc.*
  - Complémentaire des fiches rédigées pour l'exploitation
  - Support: wiki Opérations

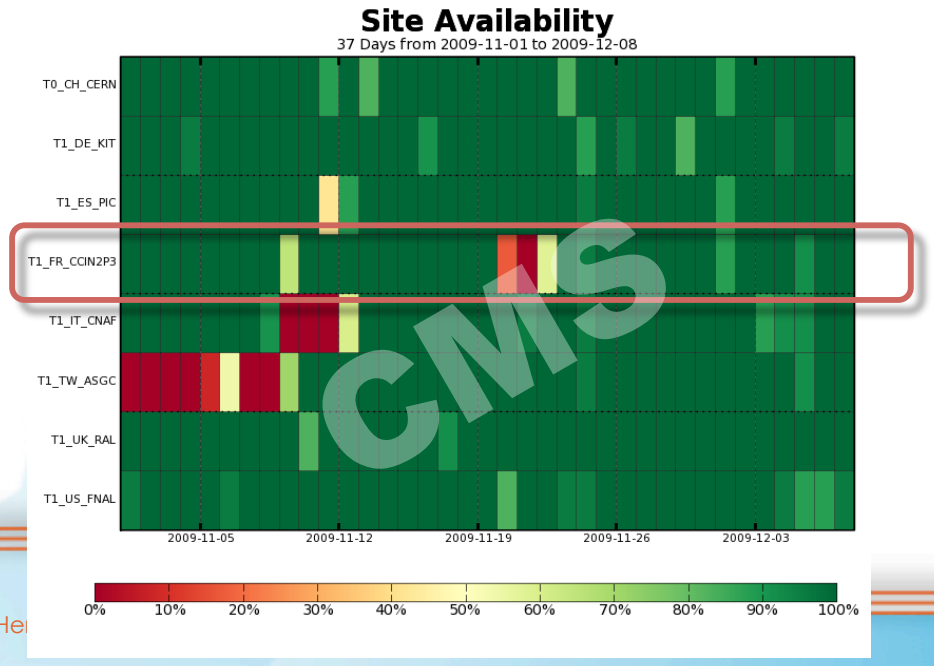
Disponibilité du site

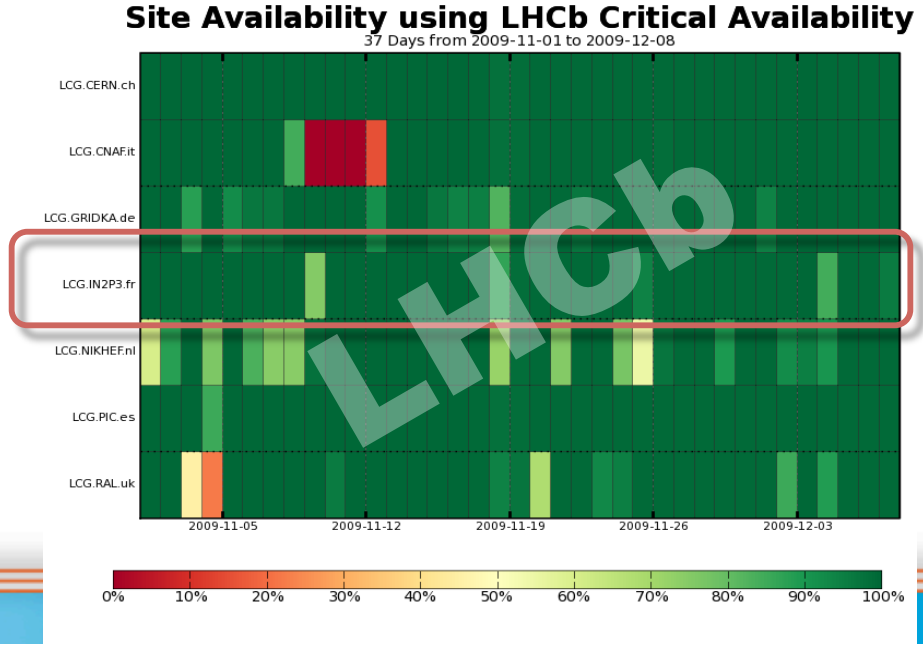
Disponibilité tier-1

Période: Nov 1 – Dec 8, 2009









Source: LHC experiments dashboard http://dashboard.cern.ch/

Remerciements

ATLAS-France

- Pas de problèmes au CC pendant la première période de prise de données
- Aboutissement de la procédure d'installation des softs dans AFS
- La vérification mensuelle des fichiers dans dCache et la compréhension de la perte de 14 fichiers en novembre
- Le nombre de jobs en exécution (~50% des jobs en exécution)
- Efficacité des jobs du nuage français est 93.8% pour la dernière période, ce qui nous place en 2eme position
- L'excellent travail de Catherine, Loic et Xavier sur l'utilisation Xrootd, de l'optimisation des serveurs et de la modification du format des AODs. Les jobs sont 6 fois plus performants
- Le fait que les jobs d'analyse passe rapidement en machine (300 jobs en exécution en moins d'une heure et tous jobs d'un lot de tests terminés en moins de 2h). Ceci est du au fait que nous avons dédié un point d'exécution aux jobs de classe G

- Il reste bien sur des points a améliorer et a garder notre niveau d'efficacité, mais ils tiennent a souligner les résultats du travail effectué au CC.

Petites Annonces

Événements à venir

- Réunion des représentants des expériences utilisatrices des services du CC-IN2P3
  - Lyon, 18 janvier 2010
  - Agenda: à venir

Aujourd'hui et à venir

- Aujourd'hui
  - NAGIOS pour la surveillance des services du site

- Réunions premier semestre 2010
  - 14 janvier
  - 11 février
  - 25 mars
  - 15 avril
  - 27 mai
  - 17 juin
  - 22 juillet

  - Agendas: http://indico.in2p3.fr/categoryDisplay.py?categId=102

Questions/Commentaires