Les nouvelles du Tier 1/Tier 2 CCIN2P3

Journées Grille France

Luisa Arrabito

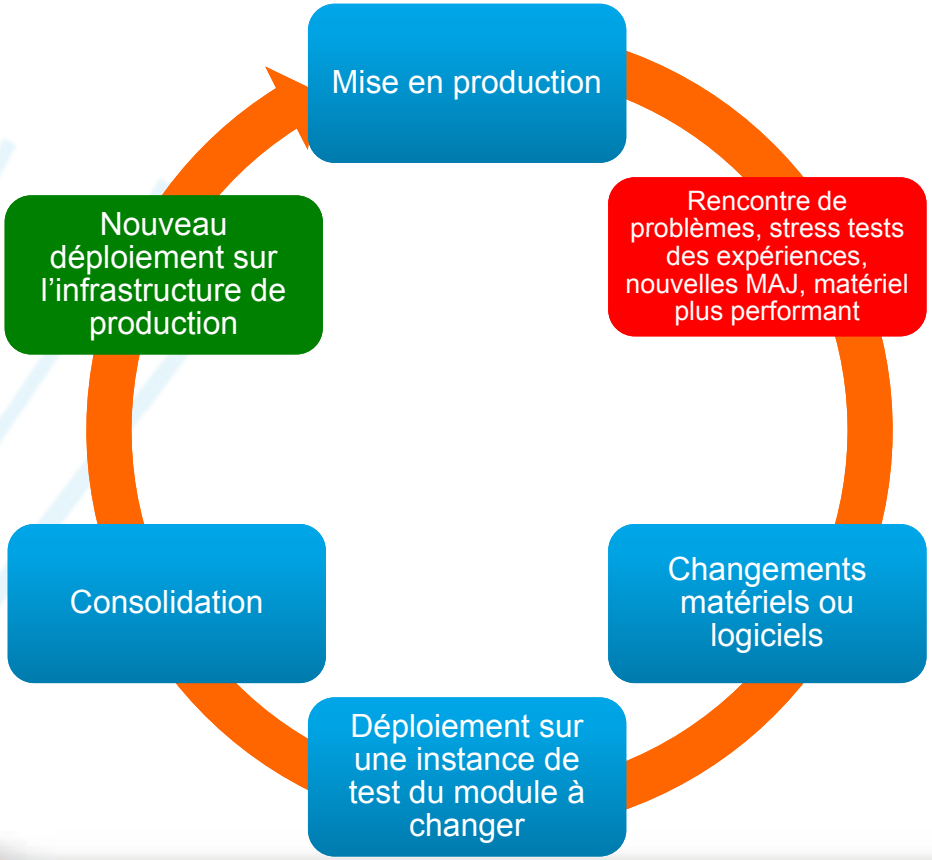
Tout tourne autour des données

Responsabilité du Tier 1 dans la réception, le transfert, le traitement, le stockage et la sécurisation des données

Transferts	Stockage sécurisation	Traitement et utilisation
• réseaux • FTS • ….	• dCache • HPSS • Bases de données • catalogues	• BQS • CE • VOBOX

Vers un Tier 1 performant

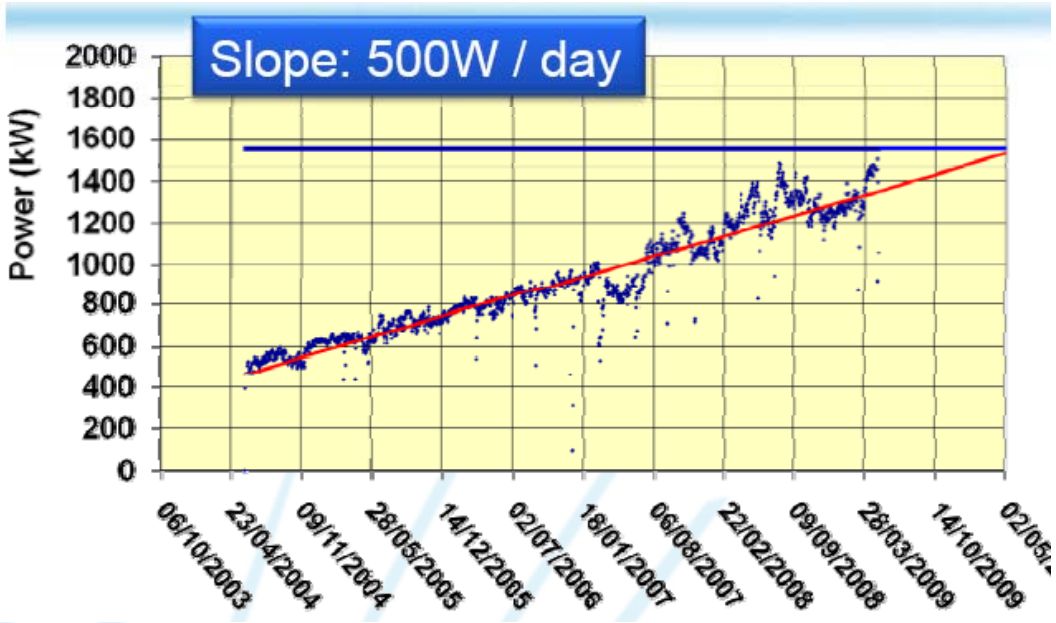
Chaque élément de la chaine doit être robuste, scalable et fournir un service stable et performant



Les Services et Infrastructure

Travaux électriques

Evolution de la puissance électrique depuis 2003



– Remplacement des transformateurs:

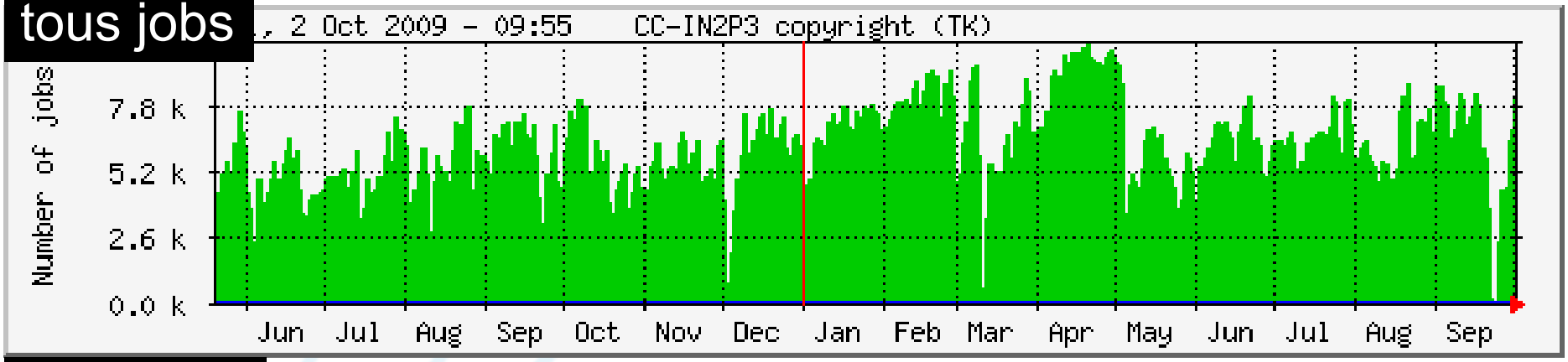
➔ Capacité électrique augmentée de 1550 KW à 3 MW

➔ 3 jours d'arrêt programmé au mois de Septembre

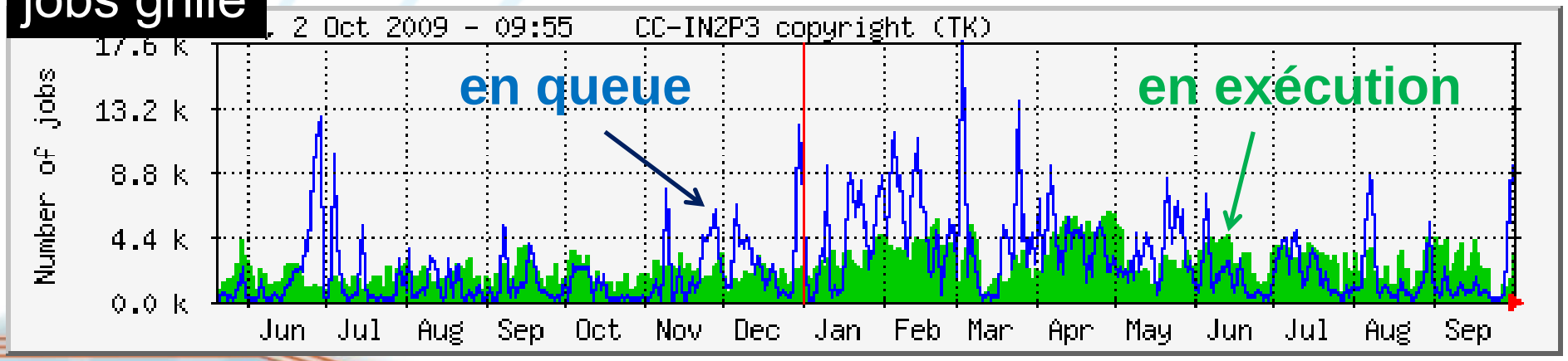
Système de batch BQS et la ferme de calcul (1/4)

Jobs en exécution entre juin 2008 et septembre 2009

tous jobs



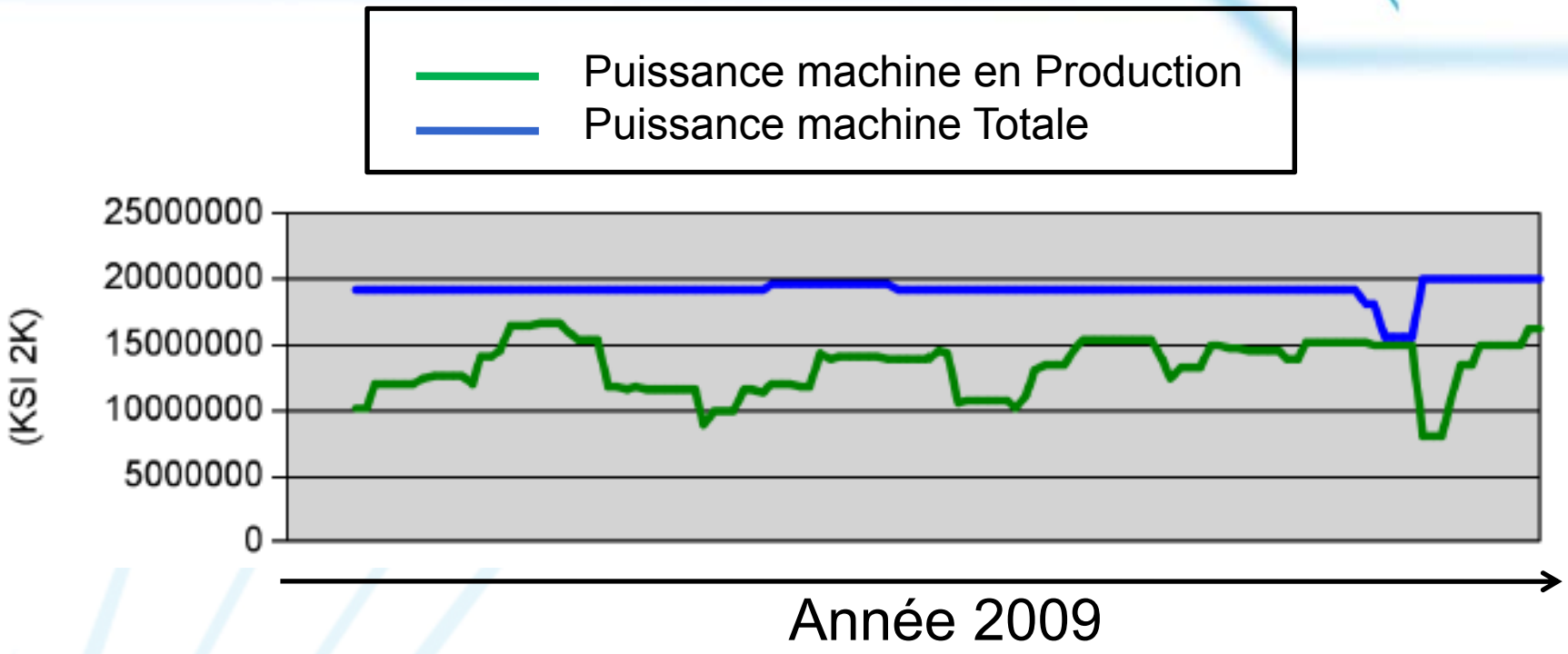
jobs grille



Système de batch BQS et la ferme de calcul (2/4)

- Travaux de Consolidation:
  - Contrôle de l'allocation de CPU « shares » à l'intérieur d'une VO:
    - ➔Modification des allocations par les utilisateurs privilégiés de la VO (utilisé dans Atlas et CMS)
  - Transmission de l'information Grille à BQS
    - ➔ID job grille, nom de la VO, ID de l'utilisateur, type de site (T1, T2,…),
    - ➔Outils pour utiliser cette information (très utile pour le debug)
  - Algorithme plus rapide de classement des jobs en queue pour l'entrée en machine => entrée plus rapide en exécution

Système de batch BQS et la ferme de calcul (3/4)



- Migration de la ferme vers SL5-64b compatible 32b:

  ➔Migration progressive depuis le mois de juillet
  ➔Actuellement 1/3 de la ferme en SL5, mais encore sous utilisée

Système de batch BQS et la ferme de calcul (4/4)

- Perspectives:

  - Etude en cours pour le choix du remplaçant de BQS

    ➔Objectif : Remplacement progressif courant 2010

  - Migration vers SL5:

    ➔80% de la puissance de calcul en SL5 prévue pour mi-Novembre
    ➔Nous encourageons les expériences à utiliser cette plateforme le plus rapidement possible

Computing Elements (1/3)

- Rappel de la configuration:
  - 2 CEs par site pour chaque VO (4 CEs):
    - ➔Redistribution des VOs
    - ➔Robustesse, disponibilité
    - ➔Mises à jour et changements sans arrêt du service
  - Répartition des ressources par rôle VOMS:
    - ➔Restriction d'accès par rôle sur les T1 et T2 (Atlas, CMS)
    - ➔Publication des restrictions (avec CMS)
    - ➔Adopté par les autres sites

Computing Elements (2/3)

- Améliorations en 2009:
  - Upgrade matériel/logiciel (de tous les nœuds grille)
    - Problème hardware des X3550
    - Les CEs migrent vers des « Dell - PowerEdge 1950 »
    - Passage progressif à SL5
  - Installation de 2 nouveaux CEs pour le cluster SL5
  - Amélioration du job manager BQS
    - ➔ Meilleure traçabilité des erreurs
    - ➔ Possibilité de modulation des « shares » à l'intérieur d'une VO au niveau du CE

Computing Elements (3/3)

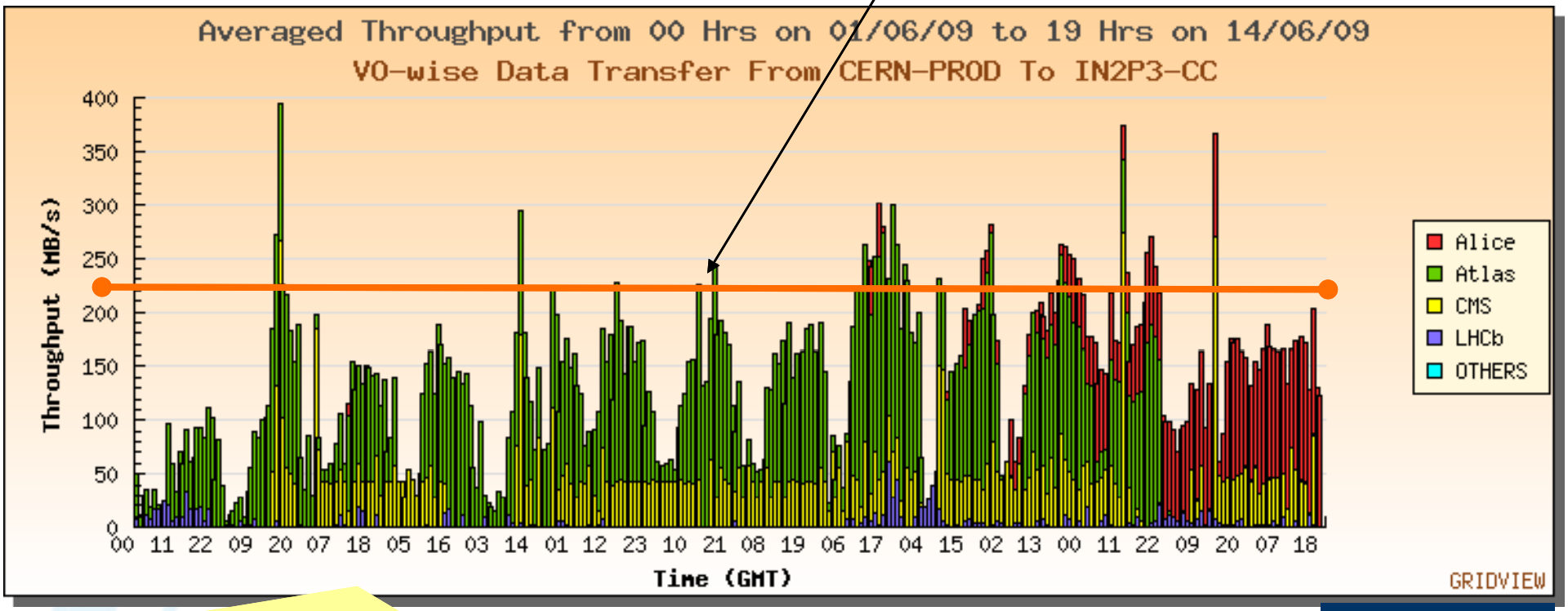
- Travaux en cours et perspectives:
  - Déploiement d'un premier CE CREAM
    - ➔Développement d'une nouvelle interface BQS/Grille
  - Campagne de test sur glexec (production candidate)
  - Séparation des VOs LHC et EGEE sur les CEs du T1
    - ➔Pour améliorer l'exploitation du site
  - Priorité à la mise en place :
    - ➔D'une infrastructure de monitoring
    - ➔Des outils d'exploitation des nœuds grille

Transferts: FTS (1/3)

Goal nominal: 225 Mo/s

- **ALICE: 6 Mo/s**
- **ATLAS: 109 Mo/s**
- **CMS: 100 Mo/s**
- **LHCb: 10 Mo/s**

- Tier-0 → CCIN2P3 (STEP'09)



June 1-14, 2009

Des taux de transfert jusqu'à 250 Mo/sec déjà maintenus pendant plusieurs jours au cours de CCRC'08

Source: Gridview http://gridview.cern.ch

Transferts: FTS (2/3)

- Améliorations en 2008/2009:
  - Matériel:
    - Accroissement de 2 à 4 machines SL4/64 avec mise en place de « load balancing »
    - Machine virtuelle de backup
  - Logiciel:
    - Version v2.1 déployée

Transferts: FTS (3/3)

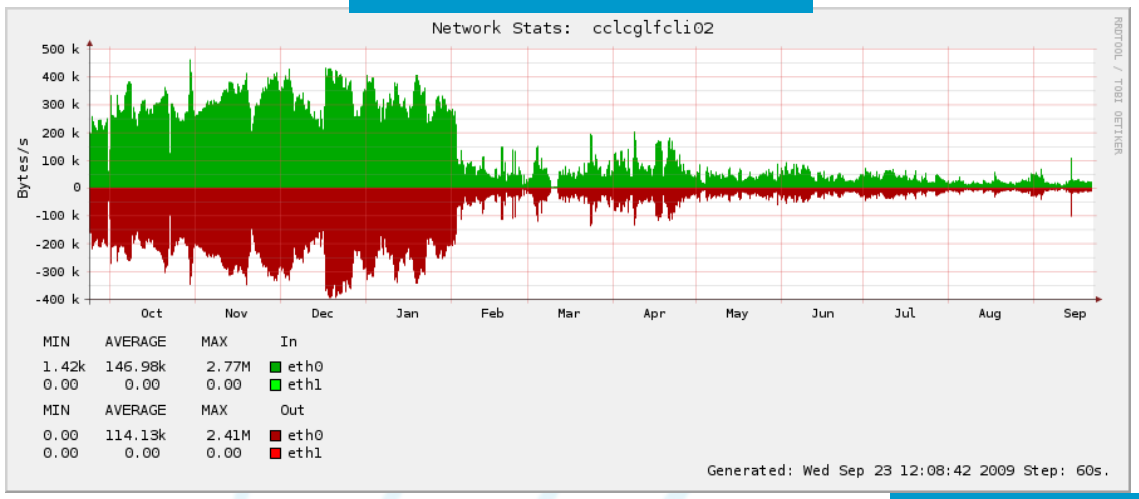
- Perspectives:
  - Meilleur load balancing
  - Monitoring plus global
  - Sonde Nagios sur: connexions oracle, daemons, canaux
  - Réplication des données via LCG-3D?

LFC (1/3)

- Importance du LFC
  - Goulot d'étranglement en cas de problème
  - Bloque les stockages et les transferts
- Infrastructure:
  - LHCb: LFC RO réplica du LFC du CERN
    - Alias sur 1 machine physique SL4/64
    - 1 BD Oracle (streams Oracle avec le LFC central LHCb du CERN via LCG 3D)
    - 1 machine virtuelle de backup
  - ATLAS: LFC local
    - Alias sur 2 machines physiques SL4/64
    - 1 BD Oracle
    - 1 machine virtuelle de backup

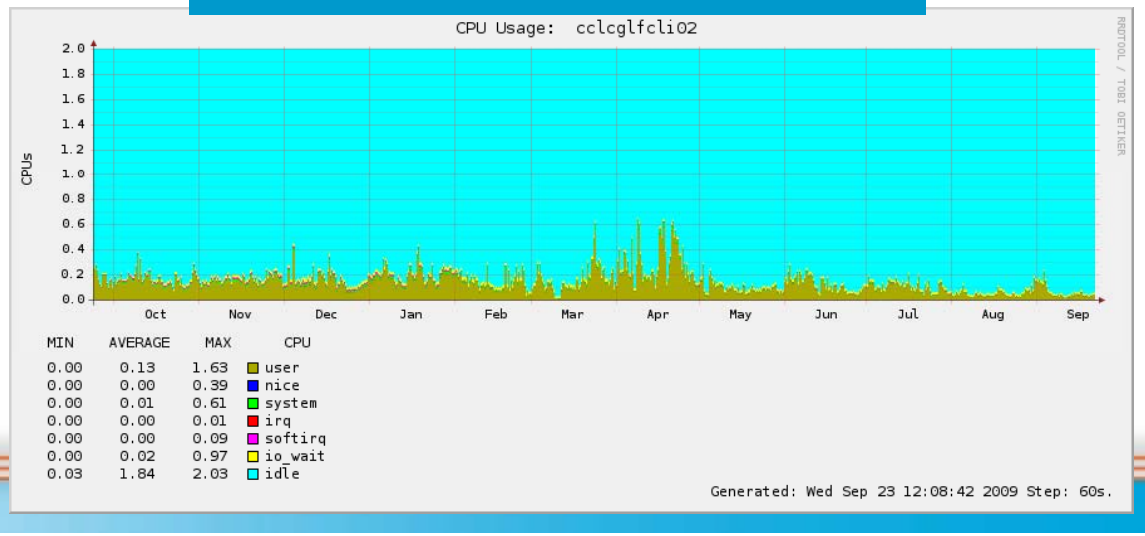
LFC (2/3)

Activité réseau



En moyenne:
- 146 Ko/s en entrée
- 114 Ko/s en sortie

Consommation de CPU



0.13 CPUs
en moyenne

LFC (3/3)

- Améliorations en 2008/2009:
  - Consolidation matérielle: ajout de machines
  - Procédure automatique de vérification du daemon LFC
  - 1 machine virtuelle de backup
- Perspectives:
  - Sondes de surveillance dans Nagios
  - Meilleur Load-Balancing
  - Redondance base de données

dCache/SRM (1/6)

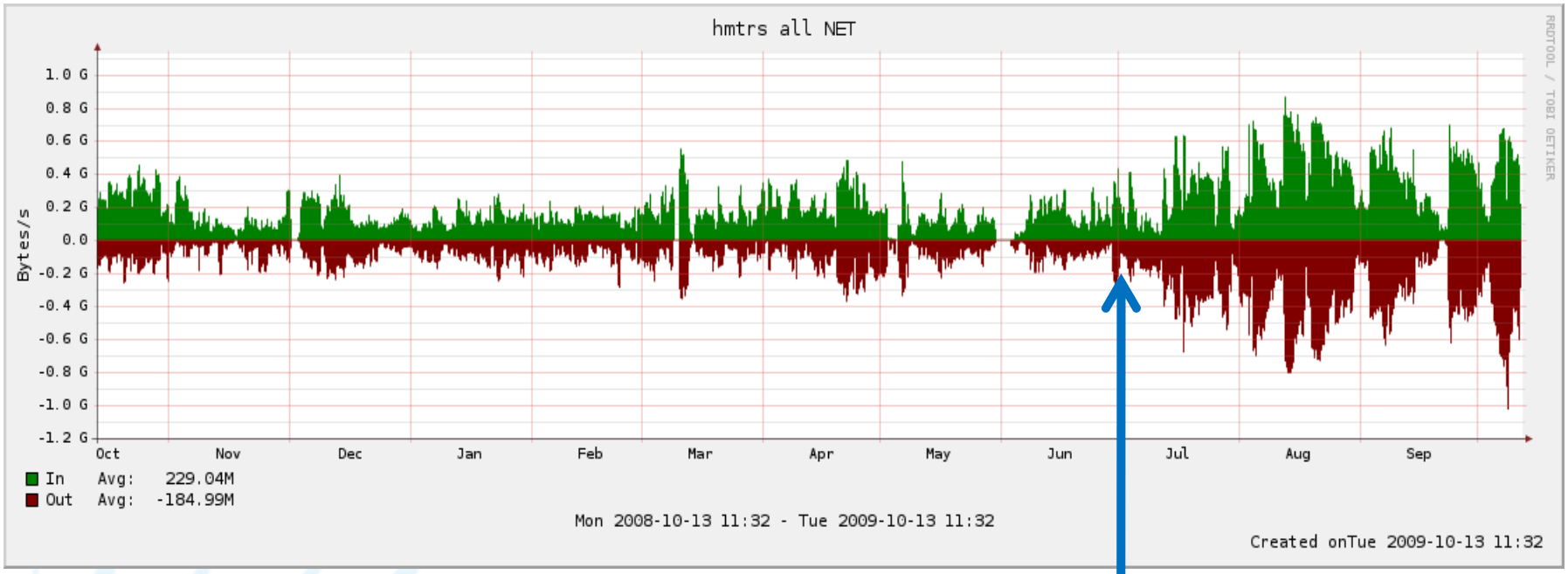
- Actuellement:
  - 10M de fichiers et répertoires
  - 2 Po de disque
  - Débit moyen soutenu:
    - 150 Mo/s en entrée
    - 500 Mo/s en sortie

dCache/SRM (2/6)

- Améliorations récentes:
  - Migration des serveurs de disque (Thumpers → Thors):
    - ➔ 80 serveurs de disque migrés (plus de 600 To)
    - ➔ double de capacité pour la même consommation électrique
    - ➔ durée de l'opération : 2 – 3 mois
  - Interfaçage avec HPSS (TReqS) :
    - ➔ mise en production d'un ordonnanceur de requêtes intelligent vers HPSS
    - ➔ les requêtes de staging sont ordonnées par TReqS en fonction de l'emplacement des fichiers sur les cartouches

dCache/SRM (3/6): TReqS

Débit en entrée et en sortie du tape mover HPSS



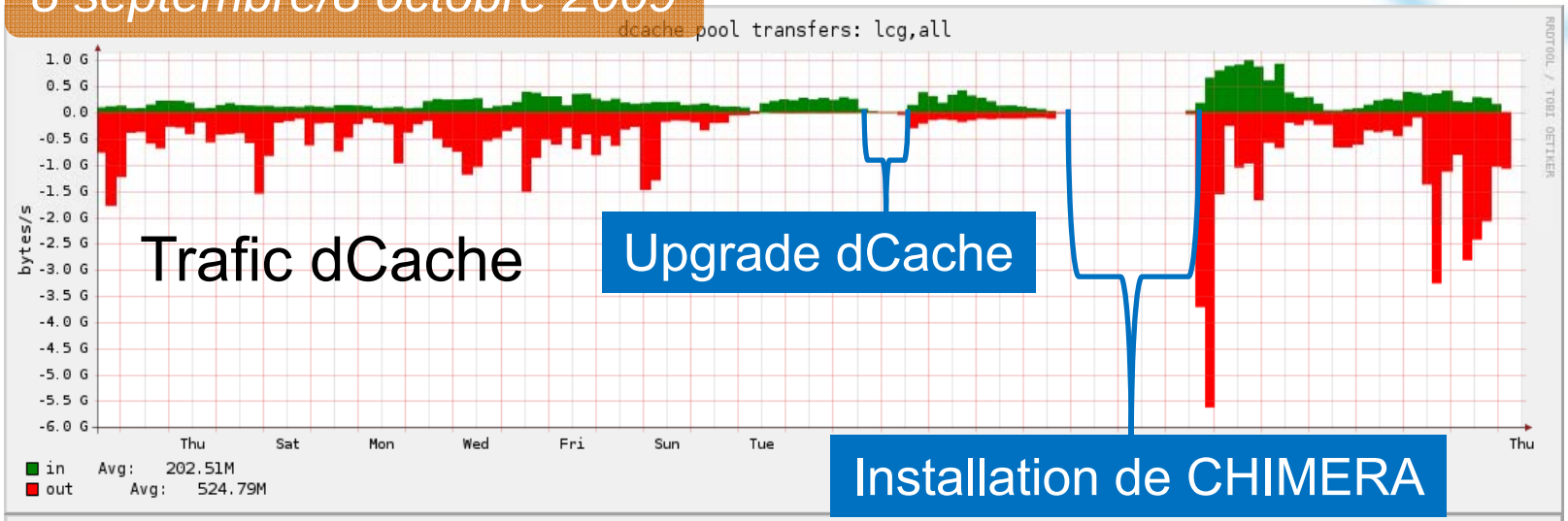
Mise en production de TReqS

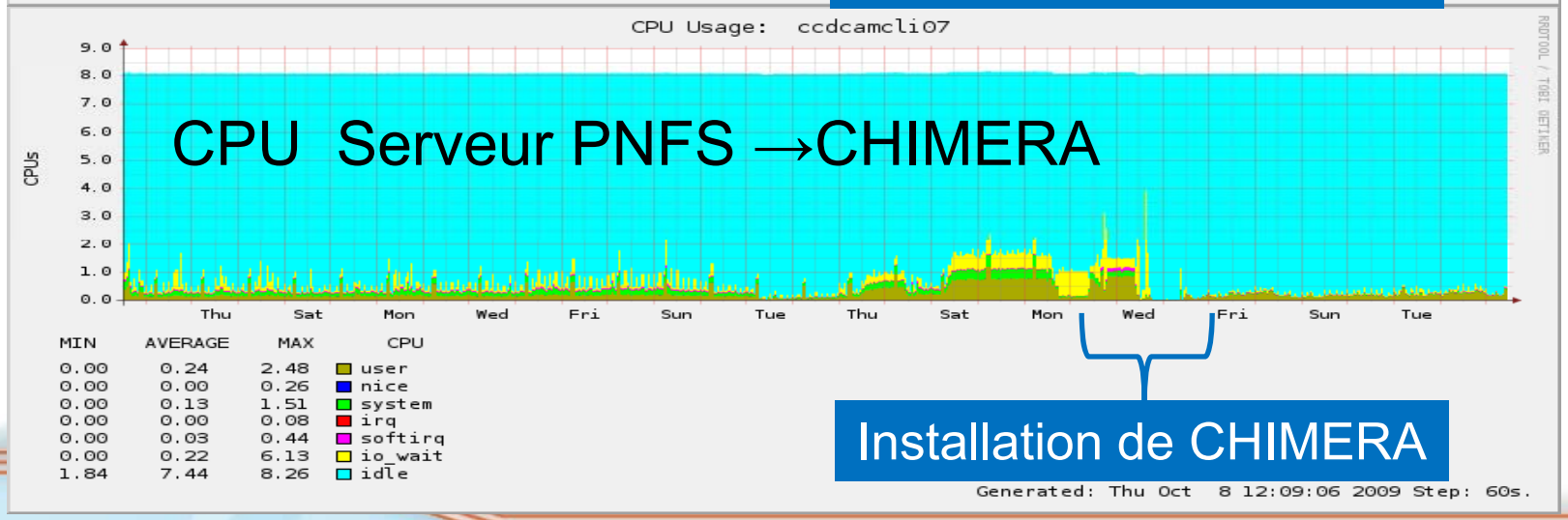
dCache/SRM (2/6)

- Améliorations récentes:
  - Migration des serveurs de disque (Thumpers → Thors):
    - ➔ 80 serveurs de disque migrés (plus de 600 To)
    - ➔ double de capacité pour la même consommation électrique
    - ➔ durée de l'opération : 2 – 3 mois
  - Interfaçage avec HPSS (TReqS) :
    - ➔ mise en production d'un ordonnanceur de requêtes intelligent vers HPSS
    - ➔ les requêtes de staging sont ordonnées par TReqS en fonction de l'emplacement des fichiers sur les cartouches
  - Upgrade de dCache → 1.9.4
  - Remplacement de PNFS par CHIMERA:
    - ➔ scalabilité
    - ➔ meilleurs monitoring et accounting

dCache/SRM (5/6): de PNFS vers CHIMERA

8 septembre/8 octobre 2009





dCache/SRM (6/6)

- Perspectives:
  - Accès à HPSS limité à des membres autorisés des VOs
  - Contrôle d'accès aux données:
    - ➔ CMS : Séparation de l'espace de stockage T1/T2
    - ➔ protection ACL SRM
  - Améliorations pour:
    - ➔ diagnostics et détections d'erreurs
    - ➔ procédures et outils d'installation
    - ➔ réplication des bases PostgresSQL
    - ➔ généralisation GFTP2 sur l'ensemble du parc

Stockage de Masse : HPSS (1/3)

- Améliorations récentes:
  - Migration de HPSS → 6.2
  - Amélioration de l'infrastructure
    - ➔Evolution matérielle du Core serveur : 16 cœurs Power 6/64 Go de RAM
    - ➔Mise en service des lecteurs T10kB (bandes 1 To)
    - ➔Mise en service de nouveaux tape mover avec interface 10Gbits
  - Migration des gros fichiers vers T10kB (déjà en cours depuis 4 mois)
  - La plupart des données des VOs LHC sont maintenant écrites sur les T10kB

Stockage de Masse: HPSS (2/3)

VO LHC	Nombre Total de fichiers	Fichiers n'ayant JAMAIS été lus
Alice	40.400	99.2%
Atlas	437.000	73%
CMS	986.000	53%
LHCb	15.400	70%

La migration vers les T10kB serait accélérée si les expériences faisaient du nettoyage dans leur données

Stockage de Masse: HPSS (3/3)

- Perspectives:
  - Fin 2009/début 2010 : mise en production nouveau robot + lecteurs T10kB
  - Migration vers 7.1 prévue en 2010
  - Evolution du cache disque

Conclusions

- Beaucoup d'améliorations et des changements majeurs appliqués aux différents services avant le démarrage du LHC pour assurer leur fiabilité et leur scalabilité pendant la prise de données
- Les tests et les exercices faits en collaboration avec les expériences sont très importants pour ajuster les services au computing model des expériences
- Les améliorations bénéficient à l'ensemble des expériences