Modèle de calcul d'ATLAS et Exercices en vraie grandeur de la grille WLCG par l'expérience ATLAS

S. Jézéquel

ANNECY lapp

Journées Informatiques 06

Sommaire

- Expérience ATLAS auprès du LHC
- Modèle de calcul ATLAS (théorie)
  - Taux de production de données
  - Distribution et stockage des données sur la grille
  - Analyse des données
- Manière effective de travailler en Septembre 2006
  - Installation du soft et réplication des données
  - Simulation
  - Analyse (*Collaboration avec C. Bourdarios (LAL)*)

Préambule

Participation à ATLAS/Grille

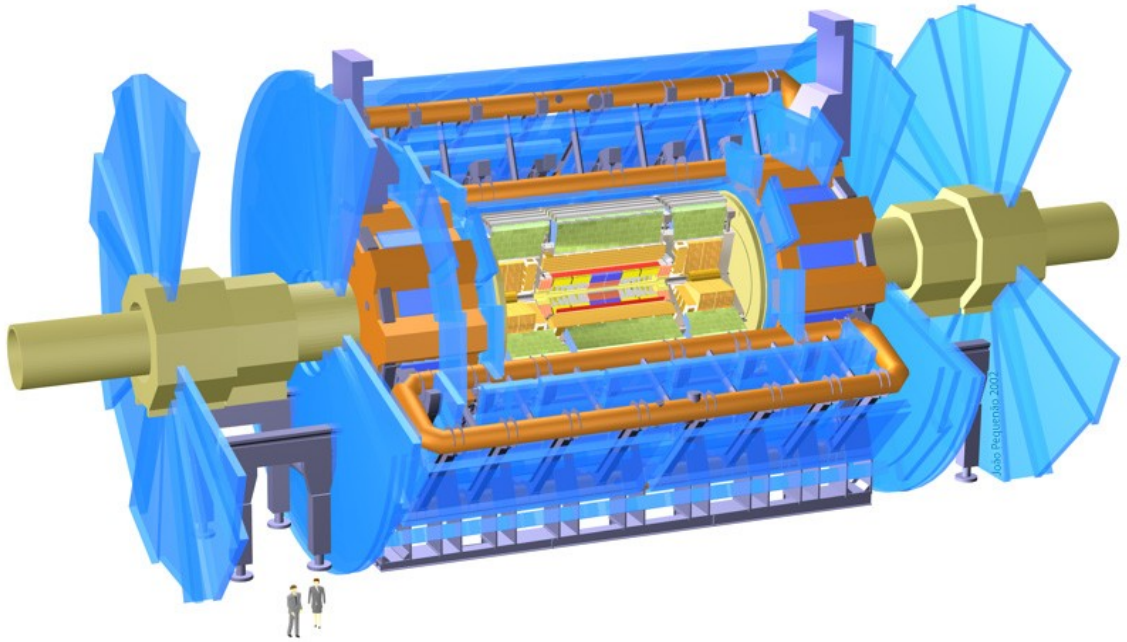
- Depuis seulement 18 mois
- Béneficie de l'aide active:
  - Du support Grille du LAPP
  - Du personnel du CCIN2P3

Outils de grille ont beaucoup changé

(SRM, FTS, LFC, ...) : Pas dans une phase stable

Risque de stigmatiser les problèmes actuels

Expérience ATLAS



Une des 4 expériences LHC auprès du CERN (Genève)
Collisions proton-proton de 14 TeV à 40 MHz

Phase de commisionning de l'accélérateur et détecteur en 2007
Démarrage pour la physique : courant 2008

Montée en charge dans les années 2009/2010

Modèle de calcul

Première version: Publiée mi-2005 (TDR)

Premiers correctifs: courant 2006

Devra encore s'adapter aux réalités

Partie Grille :

Utilisation autant que possible

des outils standards LCG (SRM,LFC,FTS,...)

→ doivent être complètement opérationnels

(faible taux d'erreur en production,

monitoring des erreurs)

sinon ...

LCG Service Model

Tier-0 – Lieu de production des données

- Acquisition et premier traitement
- Stockage à long terme (backup des T1)
- ➢ **Distribution vers les centres Tier-1**

Tier-1 – "online" avec la prise de données
→ fonctionnel 24/7

- Stockage de Masse –
→ Services grille
- ➢ **Tous les re-processings**
- Analyses de beaucoup de données
- Support national et régional

Tier-2 – ~100 centres dans ~40 pays

- Simulation
- Analyse – batch et interactif
- ➢ **Reception des données du Tier-1 associé**

Données réelles: Tier-0

- Après filtrage en ligne,
  - évenements bruts (RAW) : 200 Hz
  - Taille : 1,6 Mo/evt

  ➡ 1 Po/an (1 an=4.10^6s) (2008) 3 Po/an (1 an=10^7s) (2010)
- Première reconstruction des données au CERN:
  - ESD (0,5 Mo/evt)

  Réduction des données ➡ AOD (0,1 Mo/evt)

  (utilisées par les physiciens pour leur analyse)

Total : 2-5 Po/an

- Transferts des données dans les T1
- Pas de traitement ultérieur au CERN

  mais archivage/backup long terme

Données réelles: Tier-1

T1 recoit les données reconstruite en ligne (~13%)
(~80 Mo/s ATLAS/LYON)

Rôle :

Reconstructions ultérieures (~3/an)
à partir de sa copie de données RAW
Production de nouvelles ESD/AOD

Stockage de toutes les AODs (échange d'AOD entre T1)
Distribution des AODs dans les T2

Redondance du stockage des ESDs
entre paires de T1 (CCIN2P3-BNL)

Etat de la validation

- **Transfert ~OK en phase de test**
- **Reprocessing pas encore tester**

Données réelles: Tier2

T2 recoit une copie des AODs (N*20 Mo/s T1→ tous les T2) (Centre de Calcul Locaux)

Rôle :

Mise à disposition des AODs pour l'analyse des données (= Ferme d'analyse)

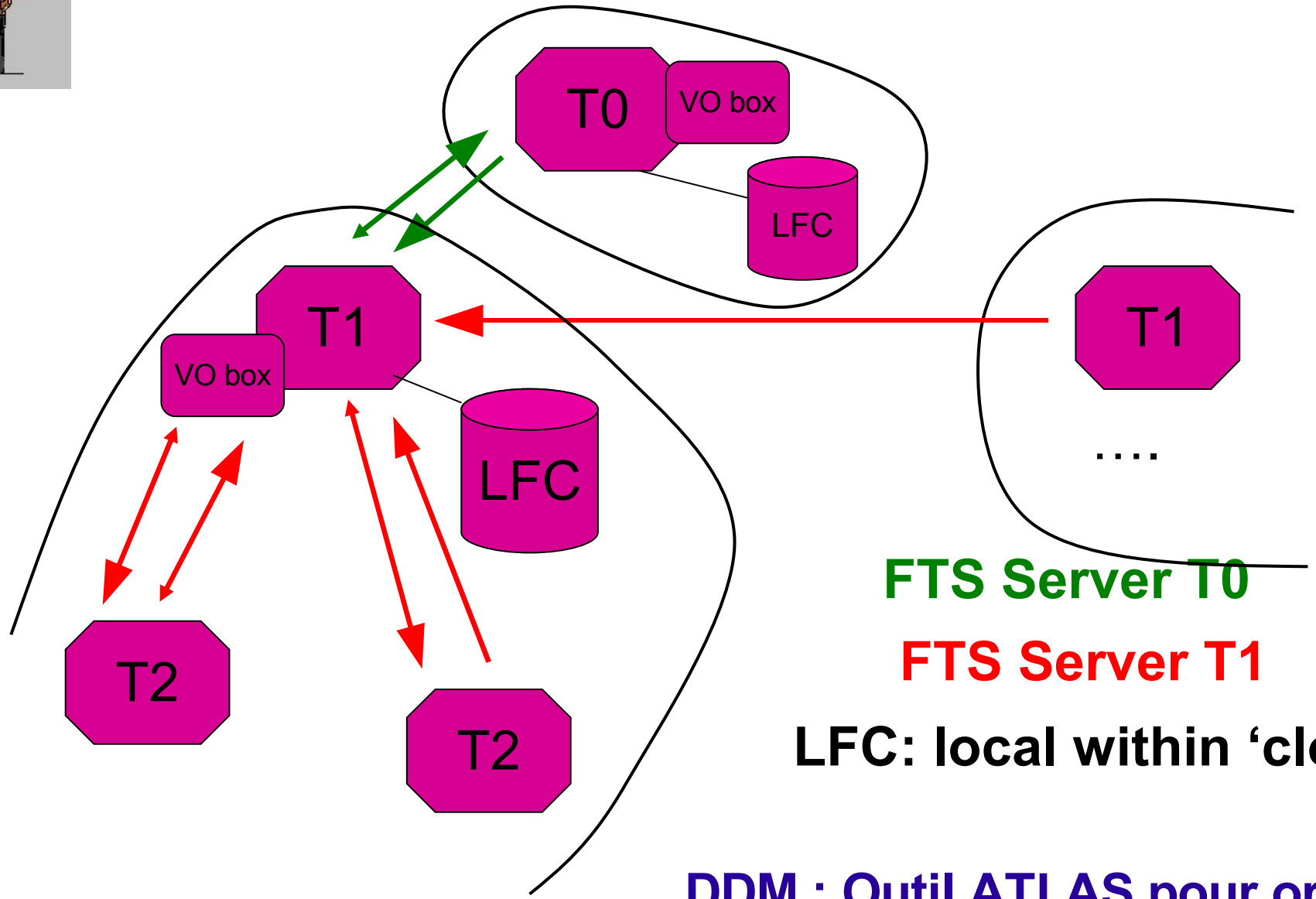
Validation:

- **Transfert commence à marcher sur courte période**
- **Activité d'analyse pas possible sur SE DPM avec les outils Grille (accès uniquement par copie sur le disque local)**

Production de données simulées

- Seule activité opérationnelle et possible à ce jour
- Produites dans les T2/T3 (et T1 actuellement)
- Centralisées dans le T1 (lieu de stockage de masse)
- Réplication des AODs vers les autres T1 (3 To)

Réplication/enregistrement data



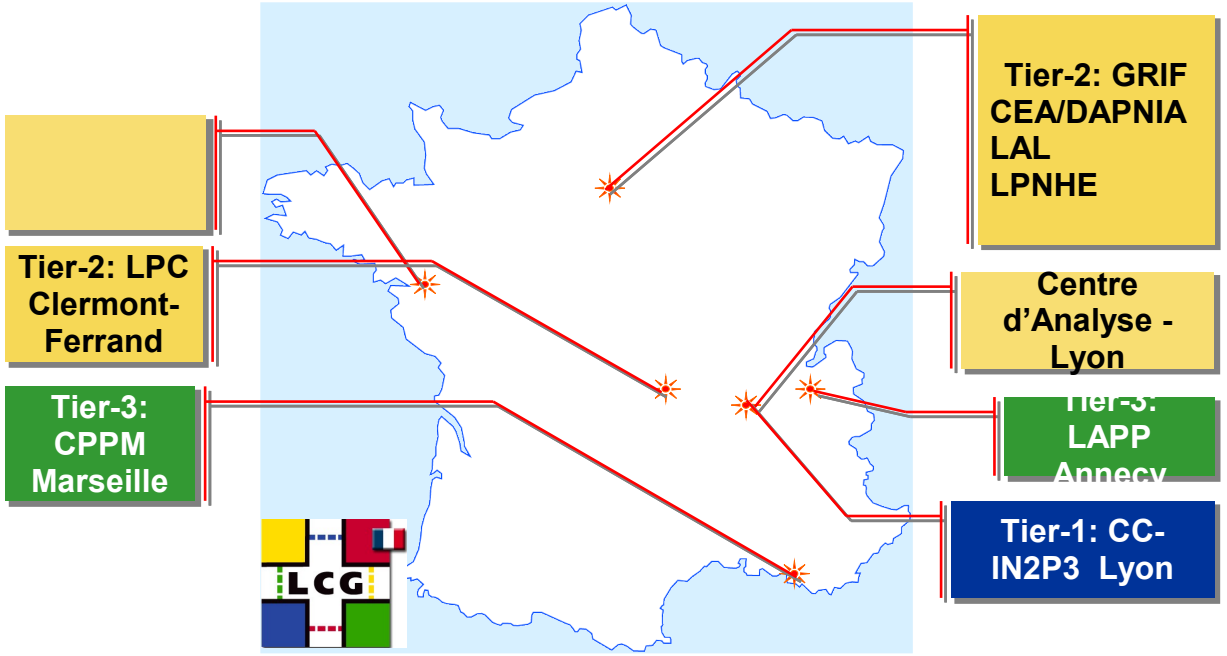
FTS Server T0

FTS Server T1

LFC: local within 'cloud'

DDM : Outil ATLAS pour organiser les transferts

Nuage FR :T1/T2/T3



- BEIJING
- TOKYO

Utilisation actuelle de la grille dans ATLAS

Premier commentaire sur la grille LCG

Quand on utilise la grille quotidiennement sur ATLAS,

il ne se passe pas une journée

sans un nouveau problème

- **Implementation de la grille (certificats,...)**
- **Interface avec les infrastructures matériels**
- **Multiplicité des sites (85% ** 10 sites = 20%)**

Il faut et on peut avancer
mais demande beaucoup d'efforts !!!!

Première solution

Avoir des contacts directs avec les sites

(Si pas de contact, on passe par GGUS)

Essayer de se restreindre à quelques sites

Au niveau T1,

collaboration poussée entre CC, CERN et BNL

(Sites pilotes pour l'analyse de données

recommandés par ATLAS)

Avoir des réunions régulières avec les T2/T3
du nuage francais

Distribution du software ATLAS

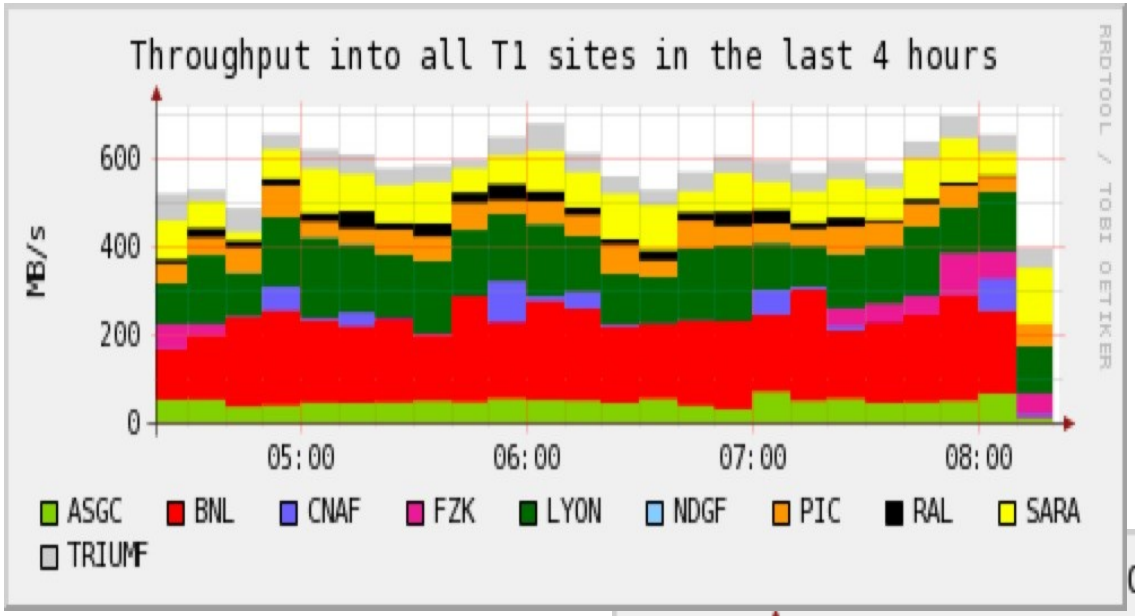
- ~100 sites pour ATLAS
- Installation automatisée/centralisée
  - Marche dans 50% des cas (sites stables)
  - Origines d'erreurs
    - Sites en évolution (changement de SE, d'OS,...)
    - Zones disques affectées au stockage de softs pleine

Besoin de contact direct

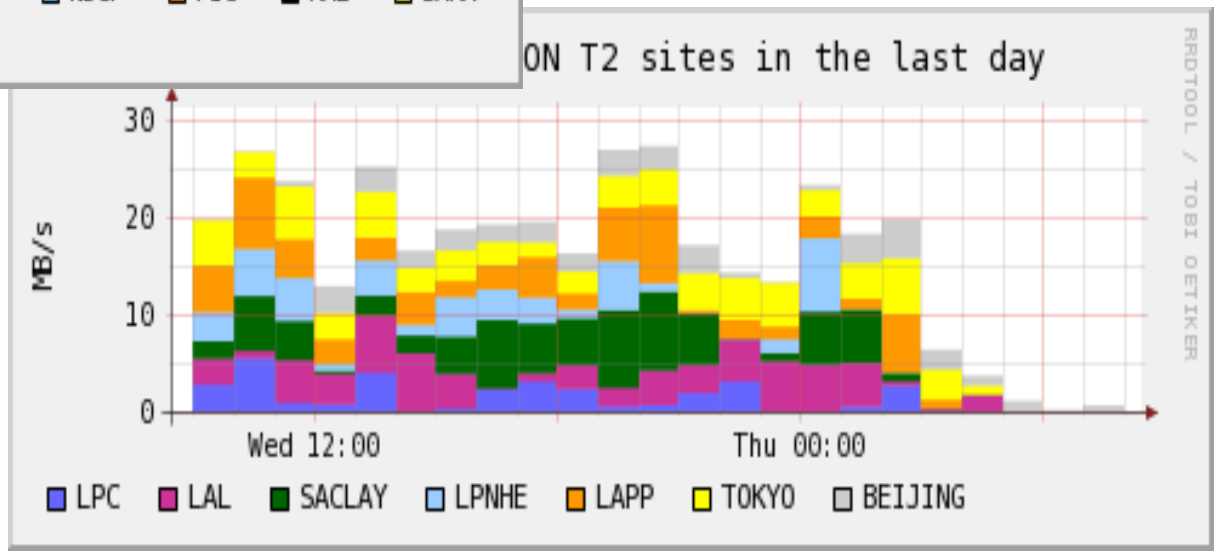
entre le responsable du site et ATLAS

Réplication/Publication des données

- Passage obligé
- Besoin d'outils performants et robustes



OK dans cadre restreint



FR : Premier 'nuage' opérationnel de la grille EGEE

Réplication/Publication des données

- Dans un environnement non dédié :
  - Surcharge du catalogue LFC
  - 10-20% d'erreur sur les transferts FTS (transfert multi-site simultannés)
  - Non optimisation de DDM

Travail en cours
mais
déjà en phase de production dans ATLAS

Interface Athena/Grille

- Soft ATLAS : Athena (aucune interaction directe avec la Grille)
- Enchainement des actions:
  - S'assure que les données sont accessibles localement sinon utilise les commandes grilles pour les repliquer localement
  - Construit le joboption en utilisant les infos des catalogues LFC
  - Tourne un job Athena en accédant directement aux données ou en les copiant sur le disque local
  - Copie les fichiers en sortie localement ou sur un SE grille

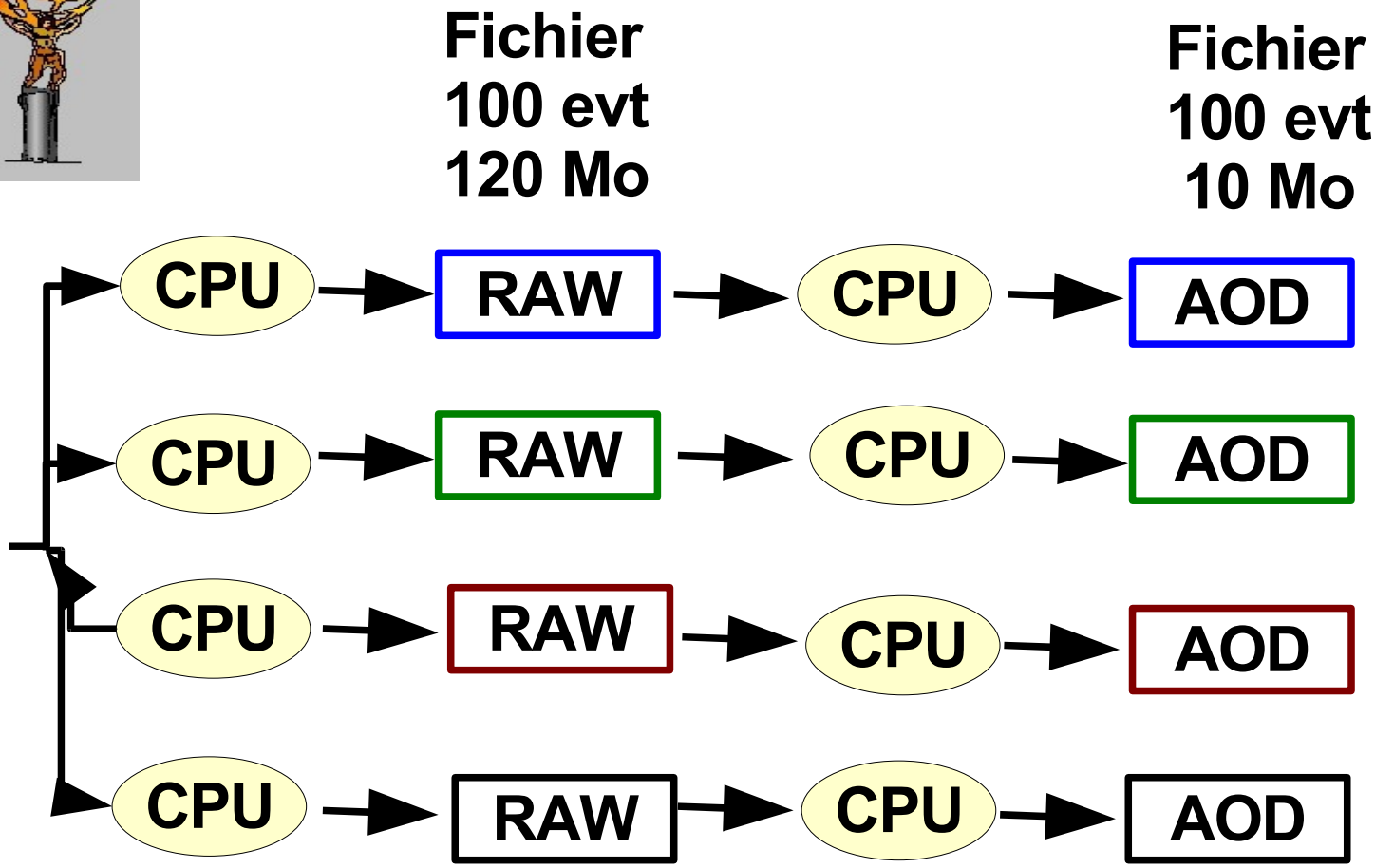
Simulation

Partie de la chaine d'analyse tirant le meilleur profit de la Grille avec un minimum de développement

- Besoin d'avoir un CPU disponible
- Un seul fichier en entrée
- Peu d'accès en lecture/écriture
- Temps de soumission << Temps d'éxecution
- Ne nécessite que qq experts pour lancer les jobs

Point critique :
Accès aux données au cours de la simulation ou après

Simulation: Implémentation mi-2005



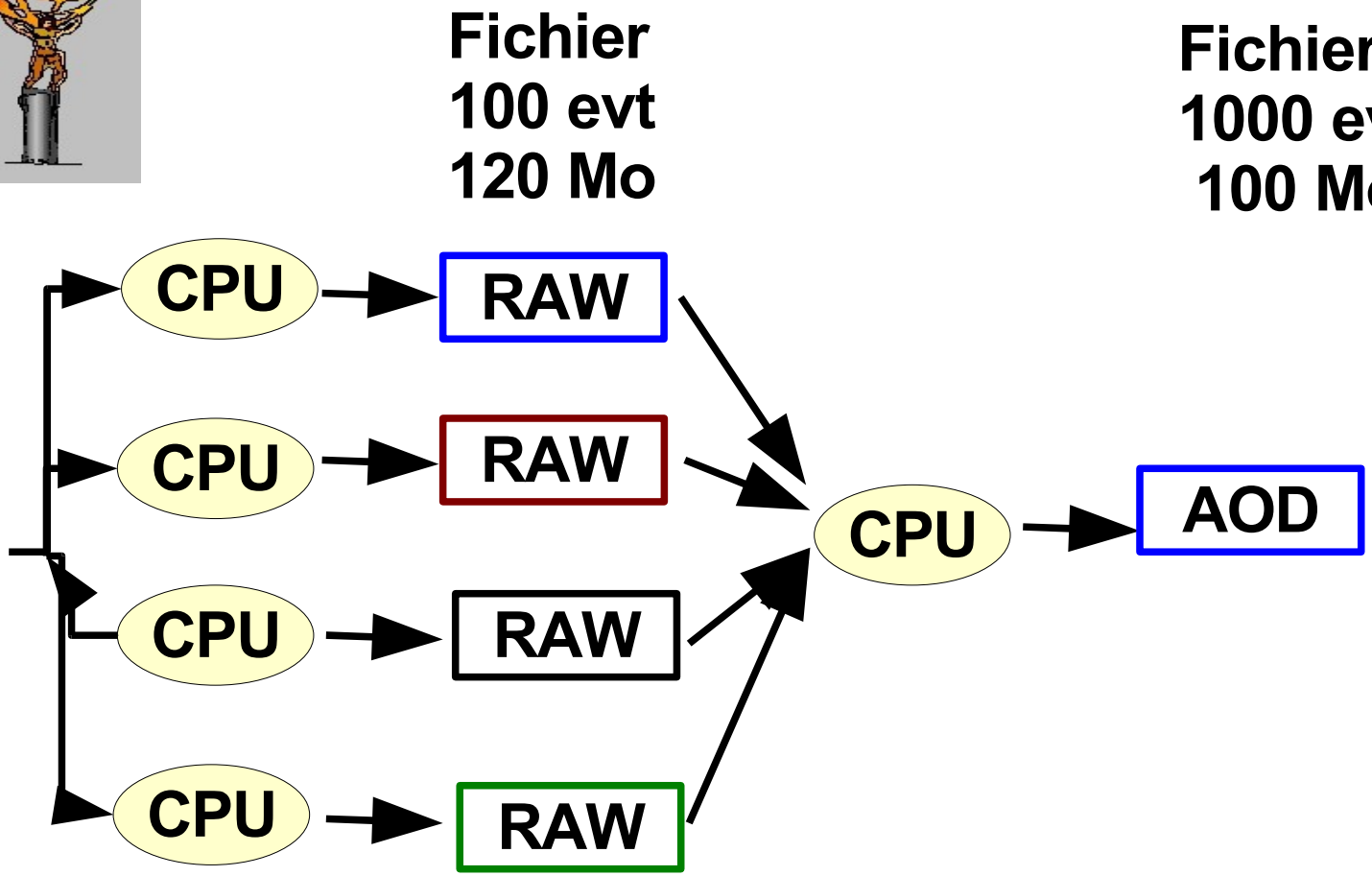
Simulation **Reconstruction**

Trop de petits fichiers : Surcharge des catalogues LFC

Fichiers dispersés sur les sites :

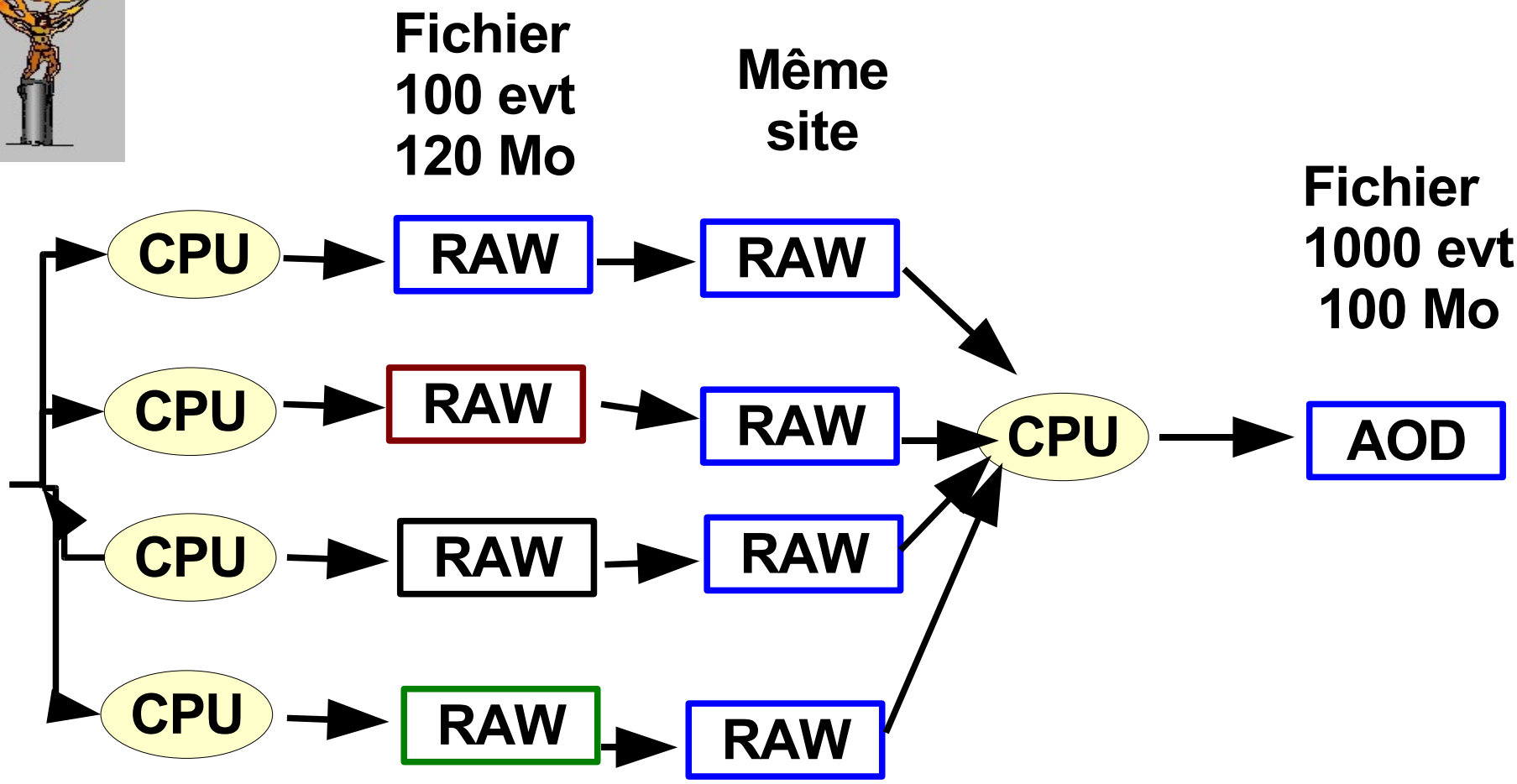
Maximise la probabilité d'indisponibilité

Simulation: Implémentation fin-2005



Fichiers RAW dispersés sur les sites :
Maximise la probabilité d'indisponibilité

Simulation: Implémentation 2006



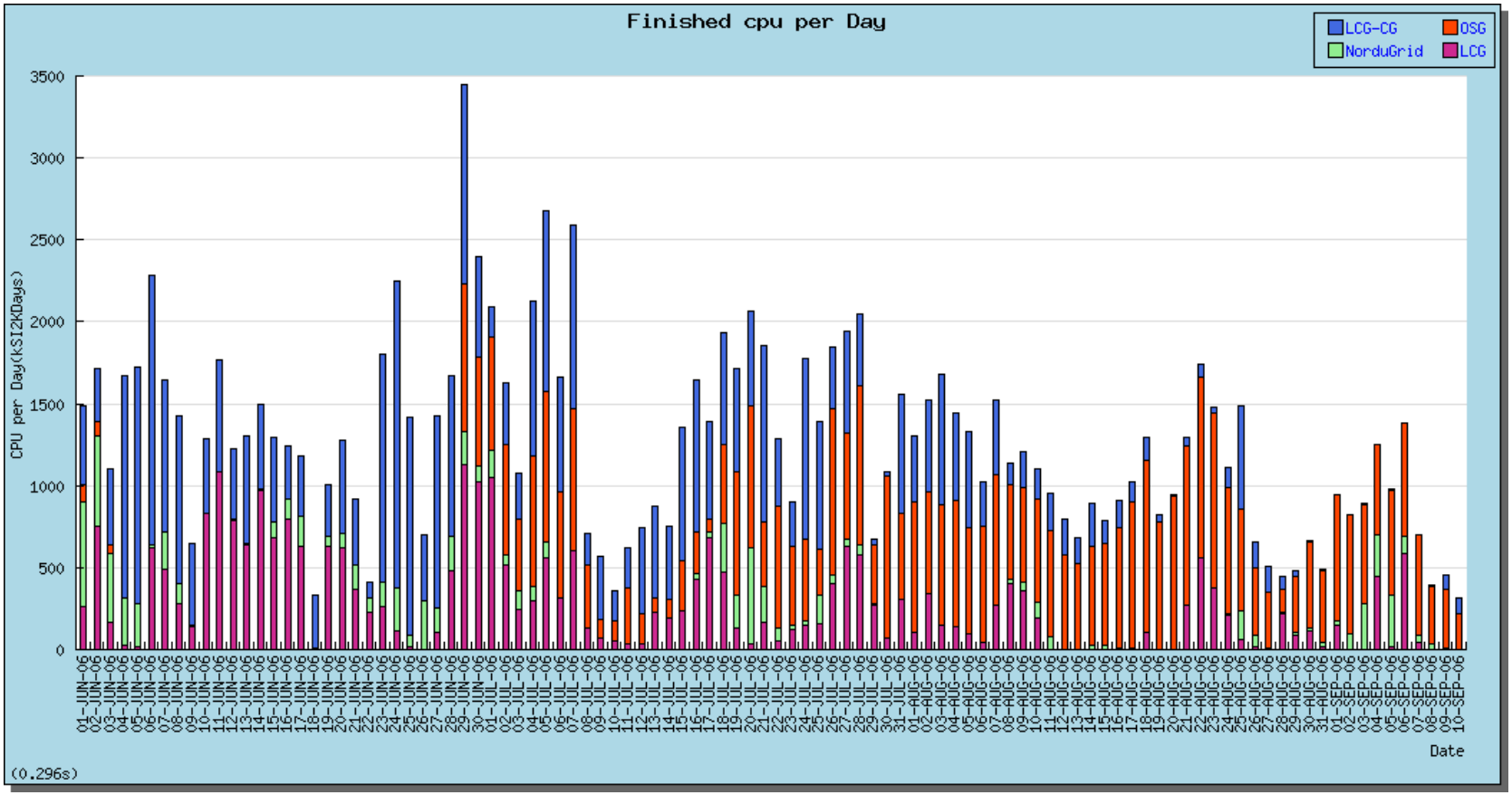
Production d'un type de données
sous la responsabilité d'un 'nuage'

Maximise l'utilisation du CPU

Ne permet pas réallocation de charge hors d'un nuage

Vitesse de production

- Encore fortement instable
- En deçà des attentes d'ATLAS (facteur 3)



Prochaines évolutions:

- Etiqueter un job de production centralisée par un rôle VOMS
- Donner la priorité dans les queues des sites à ces jobs

Analyse des données

En collaboration avec C. Bourdarios

- **Travail de tous les physiciens**

  (ca doit marcher comme sur la doc)

Commandes grilles de bas niveau inconnues

et/ou limitations grille/infrastructure

➡ **Bonne fiabilité et stabilité avant diffusion**

Analyse des données:Standard

- **En 2006:**
  - **Travail sur des données simulées avec**
    l'équivalent de qq jours de données
  - **Utilisent les réplications des données locales**

(commencent à se former aux outils de réplications)

  - **Tourne les jobs analyses en interactif**
    ou avec qq batchs BQS/LSF

Analyse des données:Prospective

Analyse avec infrastucture Grille :
Utile lorsque beaucoup de données seront disponibles

En phase d'évaluation et debugging des outils

Préalable: nécessite de rassembler sur un/plusieurs sites tous les fichiers d'un même job

2 interfaces utilisateurs sur le marché:

- **Panda (made in BNL)**
  Ne tournait qu'à BNL jusqu'à récemment mais avec succès
  Volonté de le porter sur LCG (1 job a déjà tourné à Lyon)

- **Ganga (made in Europe: LHCb/ATLAS)**
  Commence juste à être opérationnel dans ATLAS
  CCIN2P3 est un des sites de validation

Analyse des données:Prospective(2)

Quid de l'analyse sur les T2 avec SE DPM?

- **Envoyer des jobs sur ces sites : OK**
- **Possibilité d'écrire son script pour:**
  Copies des fichiers sur disque scratch local
  Tourner Athena
- **Pas encore d'accès direct aux fichiers par rfio**
  (annoncé pour Octobre)
- **Important d'aboutir pour avoir un intérêt local à faire fonctionner un T2/T3**

Analyse des données:Prospective(3)

Pas encore d'implémentation d'algorithme sur choix des sites au delà de la seule disponibilité de CPU

Exemples:

- **Disponibilité des données (bientôt)**
- **Disponibilité effective du site**
- **Vitesse d'accès aux données à un instant t**
- **Répartition de la charge entre sites (>1000 jobs)**

Rapidité de soumission de job (Ressource Broker)

Conclusion

- Commence la validation concrète du modèle de calcul
- Mise en place pas finie des outils Grille indispensables
- Point de passage obligé : outil de réplication performant et robuste (ATLAS/Grille)
  - Simulation sur la grille en cours de stabilisation
  - Démarrage de la validation de l'analyse sur la grille (job Grille/ accès aux données sur un SE)

Beaucoup de travail encore en perspective

mais effort français visible dans ATLAS