L'informatique au service de la physique des hautes énergies : morceaux choisis

Catherine Biscarat
Laboratoire de Physique Subatomique et de Cosmologie
Grenoble, France

Séminaire au LPC de Clermont-Ferrand,
le vendredi 20 janvier 2011

Petite intro sur l'oratrice

- Auparavant : expérimentaliste en physique des particules, sur collisionneur

Détecteur aux EU, à Chicago, collisionneur ppbar Tevatron à √s = 2 TeV
- calorimétrie, production d'événements simulés, recherche SUSY

Détecteur au CERN, à Genève, collisionneur pp LHC à √s = 14 TeV
- calorimétrie, recherche de nouvelles particules
- responsable des activités de calcul ATLAS dans un centre de calcul majeur

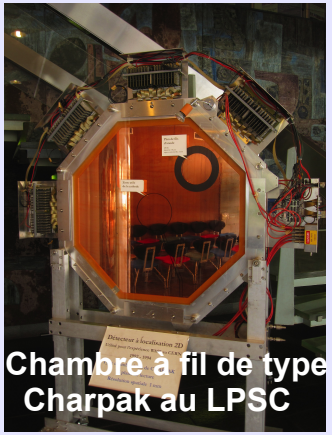
- Ingénieure en informatique au LPSC à Grenoble
  Grilles de calcul (WLCG et grille légère grenobloise CIGRI)

Un grand merci à mes collègues et collaborateurs :

B. Boutherin,
D. Boutigny,
S. Campana,
M. Canehan,
X. Canehan,
F. Chollet,
S. Crépé,
P. Girard,
C. Gondrand,
M. Puel,

pour le coup de pousse que certains de leurs talks ont donné à ce séminaire.

Quelques années en arrière


Chambre à fil de type Charpak au LPSC

- Chambre proportionnelle multifils (G. Charpak, Nobel 1992)
  → éléctronique dans les expériences

- Le volume de données ne cesse d'augmenter
  - Le computing est une part intégrale des expériences (budget)

- D0 Tevatron Run I (1992-1996) :
  - ~50 VAX sur place (acquisition, reconstruction, analyse)

- D0 Tevatron Run II (2001-2011) :
  - sur place traitement des données (600 PCs), centres distants et inhomogènes simulation (et plus tard reprocessing)
    - Centres dédiés, software installé, un user unique local
  - grille : la ferme à Fermilab ne suffit plus → reprocessing 2005 (250 TB de données) et simulation
    - Outils de grille (*metascheduler*); naissance de SAMGrid en 2003.

- Babar (1999-2008) : modèle d'analyse dernières années 1 groupe ↔ 1 site dédié (ssh)

Quantités de données au LHC

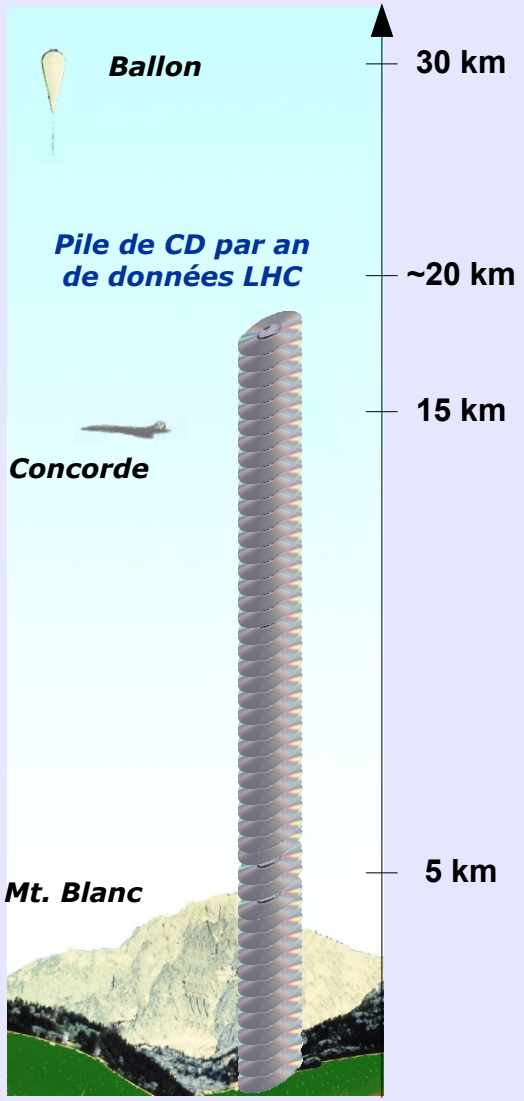
D0 RAW : 250 kB/event

A vos calculettes !

* 100 collisions enregistrées par sec.
* 1-2 MB de données par collision

→ quelques PB de données produites par an
et pour un détecteur LHC (3 pour ATLAS)

Soit 600 000 films DVD produits
par un détecteur en un an



1 MB
1 photo digitale
1 dictionnaire
1 CD = 650 MB

1 GB = 1000 MB
5GB = 1 film DVD

1TB = 1000 GB
Livres produits par an

1PB = 1000 TB
Production par an de
1 expérience LHC

1 EB = 1000 PB
Production mondiale
d'informationen 1 an

Enorme quantité de données
à traiter et à stocker;
aucun centre de calcul
ne peut le faire seul.

La grille de calcul

La grille de calcul

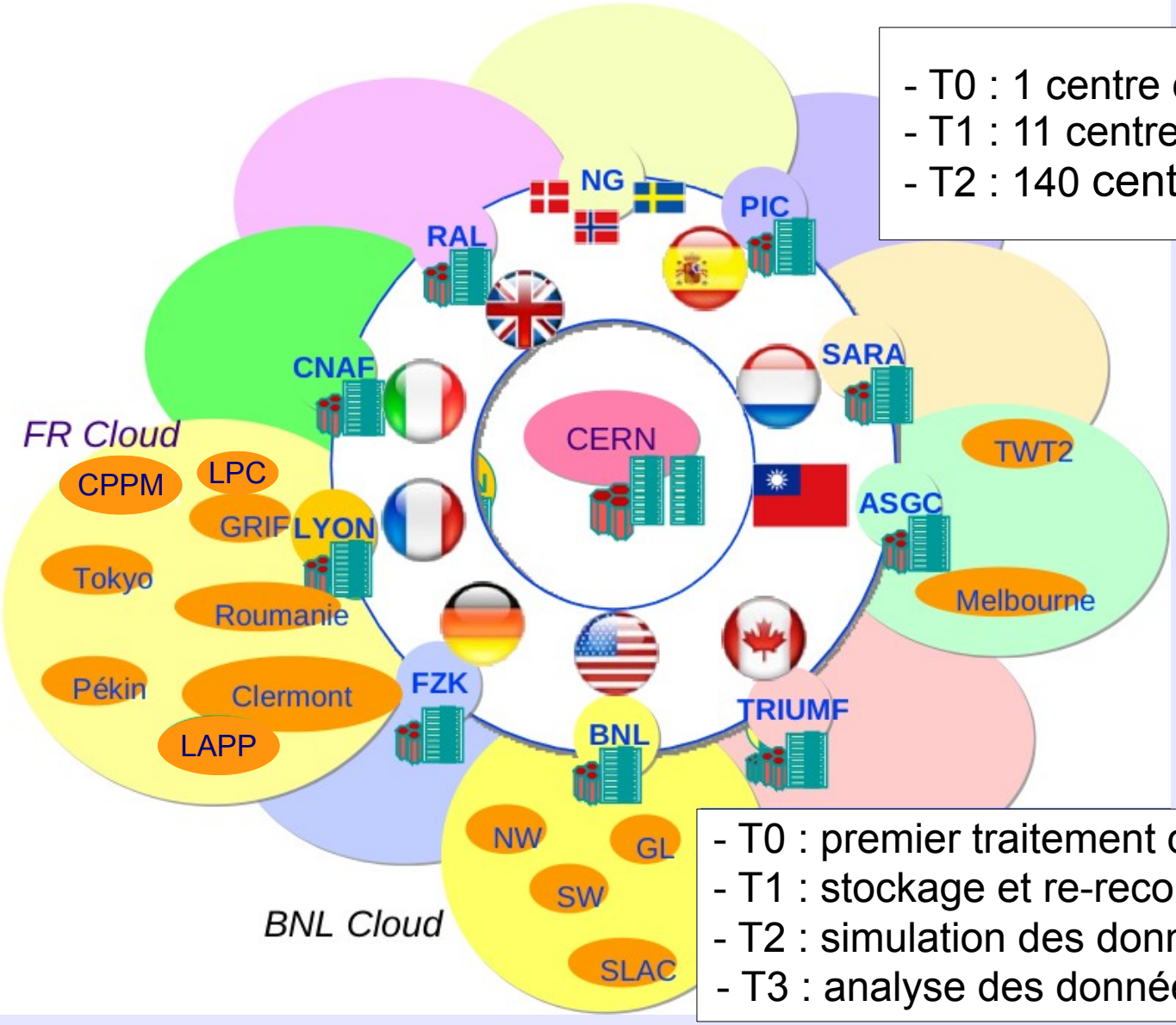
- Mutualisation de ressources de calcul
  de plusieurs unités pour un but commun
- Terme provenant du réseau électrique (*electric power grid*)
  où un appareil électrique est alimenté quelque soit où il est branché.
- L'utilisateur voit un seul « gros ordinateur ».

Worldwide LHC Computing Grid (WLCG)

- 4 expériences LHC
- 150 centres de calcul
- 35 pays
- Réseau privé (LHCOPN)
- Première mondiale

- Utilisateurs distribués partout dans le monde

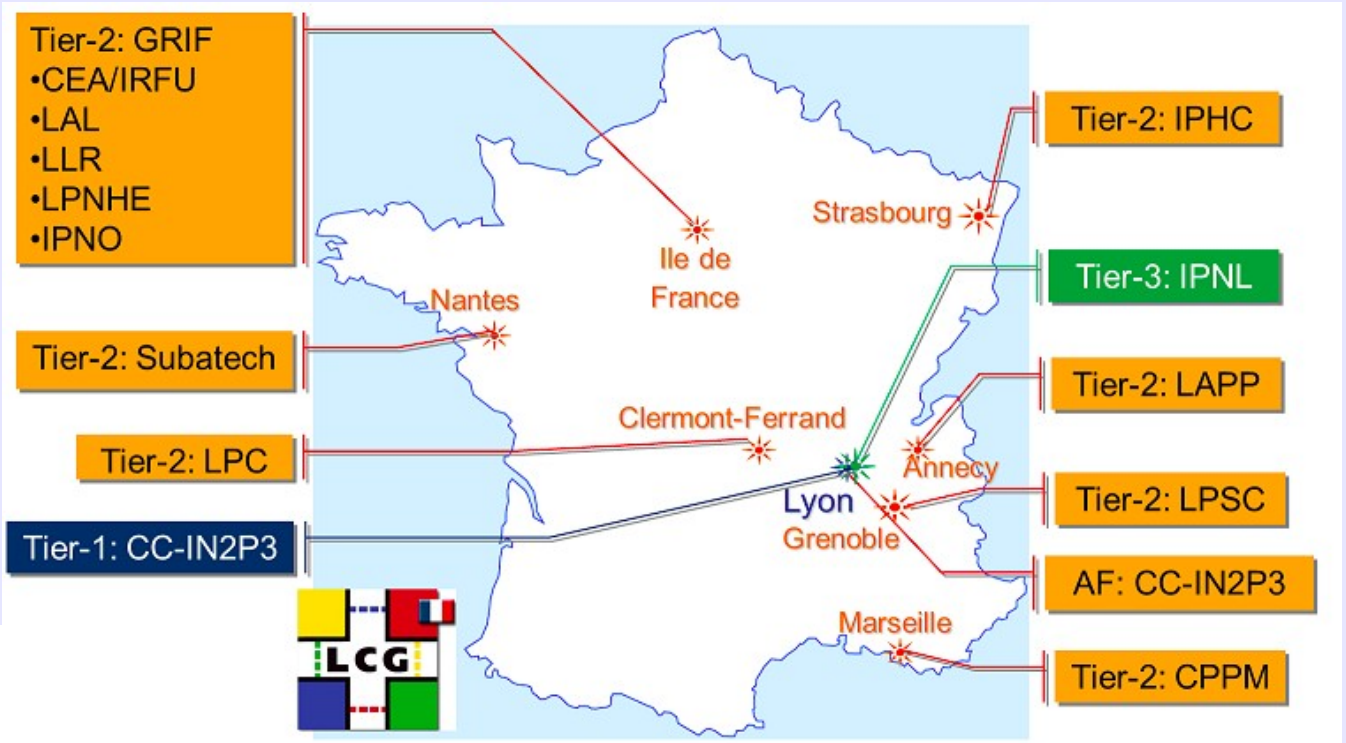
Hiérarchie de centres de calcul

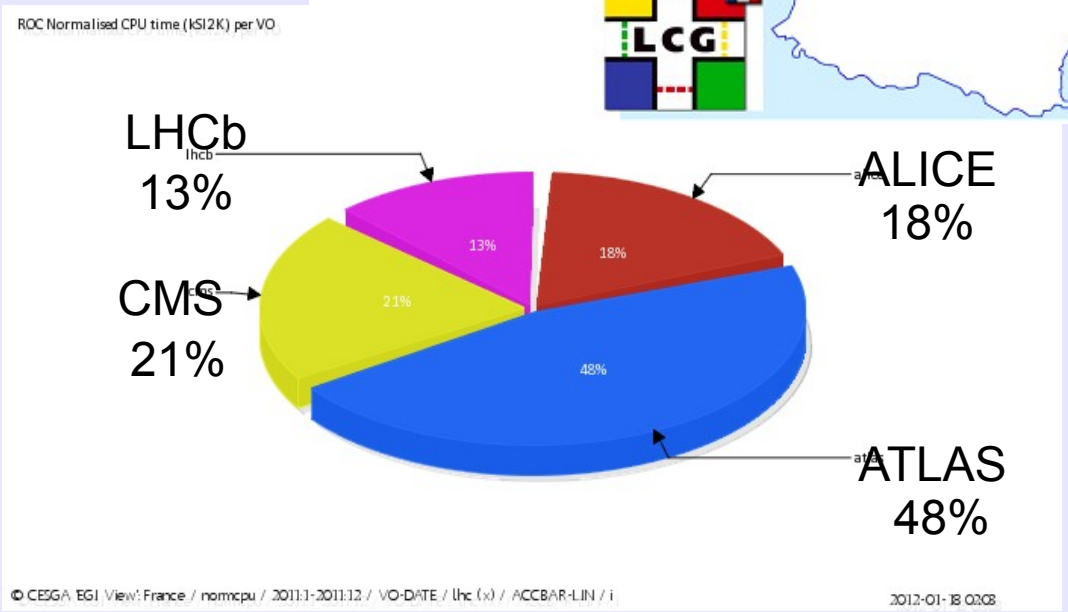


- T0 : 1 centre central (le CERN)
- T1 : 11 centres principaux (réseau dédié)
- T2 : 140 centres secondaires

- T0 : premier traitement des données brutes
- T1 : stockage et re-reconstruction des données brutes
- T2 : simulation des données brutes
- T3 : analyse des données (physiciens)

Centres de calcul WLCG en france

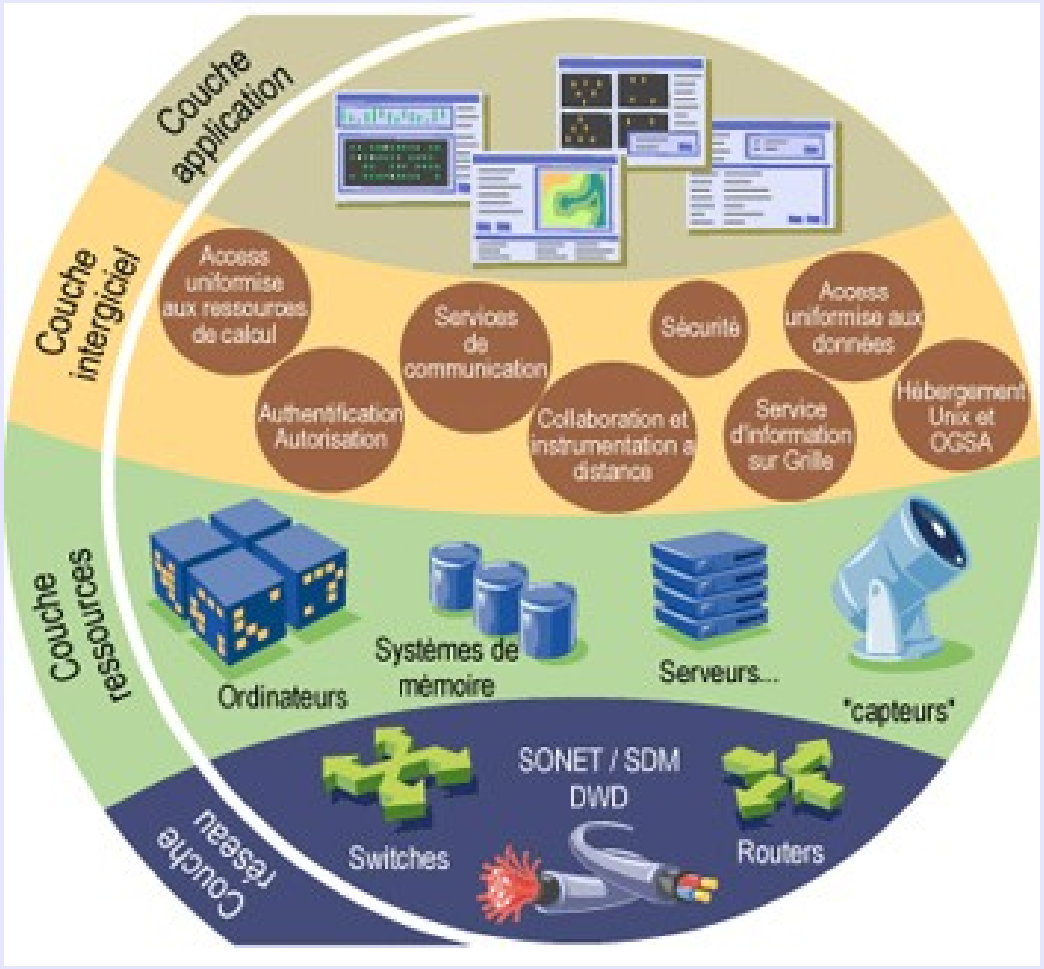




CPU dépensée sur les sites FR par VO et pour l'année 2011

Architecture de la grille

- Architecture complexe
- *Middleware* (intergiciel), le liant
- Virtual Organisation

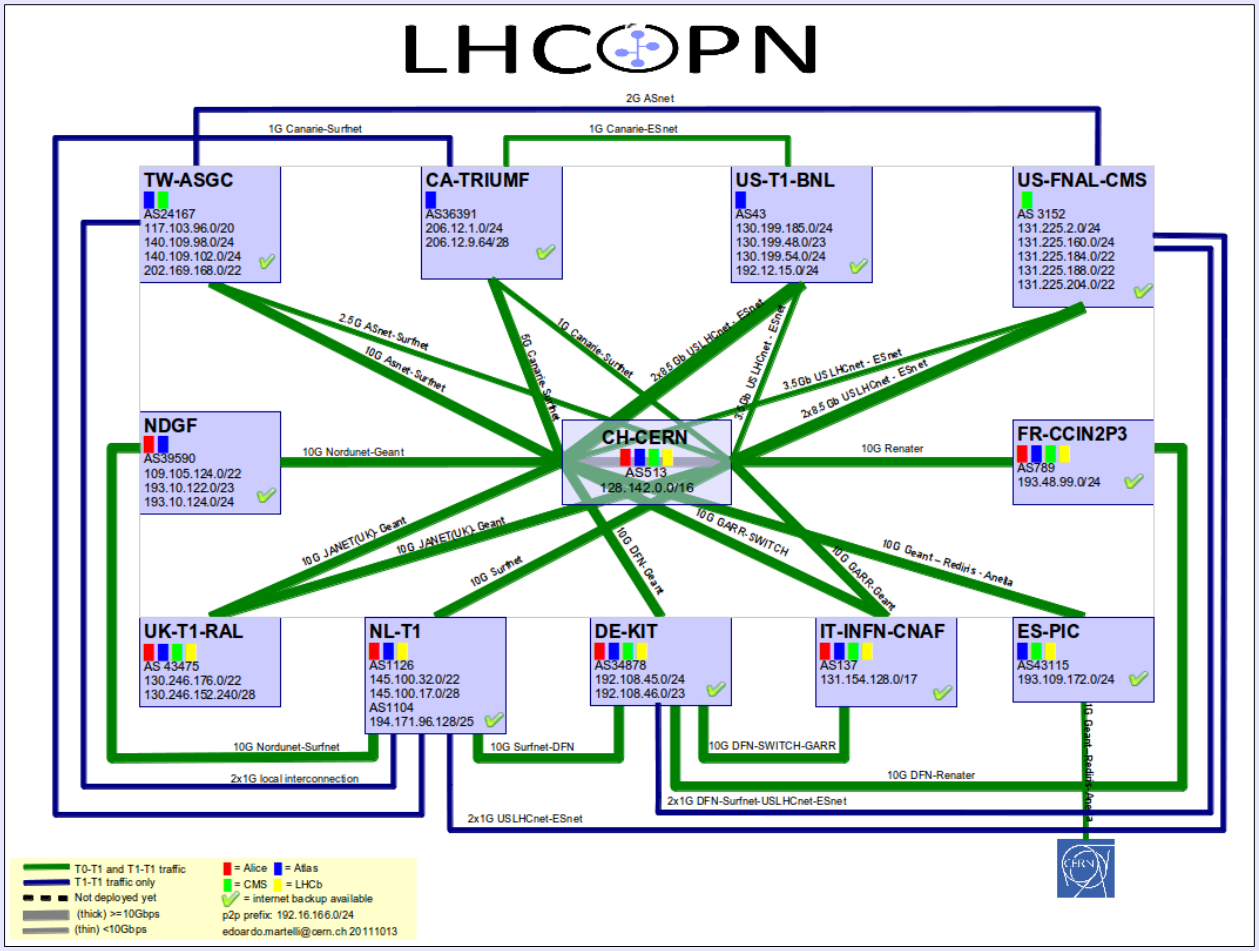


Source : http://lcg.web.cern.ch/lcg/

- Ingrédient crucial et souvent oublié : le réseau !

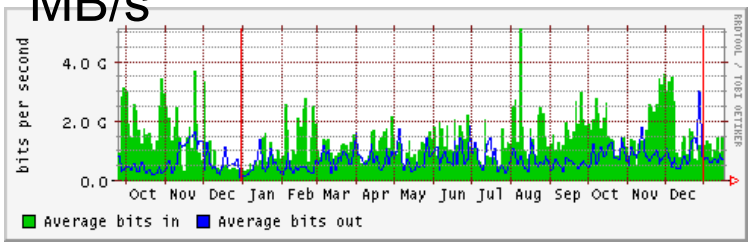
Un réseau international performant

- LHC Optical Private Network
- Réseau privé et parallèle aux autres réseaux pour liaisons T0 - T1s



T0 ↔ CC-IN2P3 (10Gb/s)

T0->CC : nominal ~250 MB/s



Max from CERN 5.6 Gb/s
Ave. from CERN 2.1 Gb/s

Source: http://netstat.in2p3.fr

Un réseau national performant

Renater :

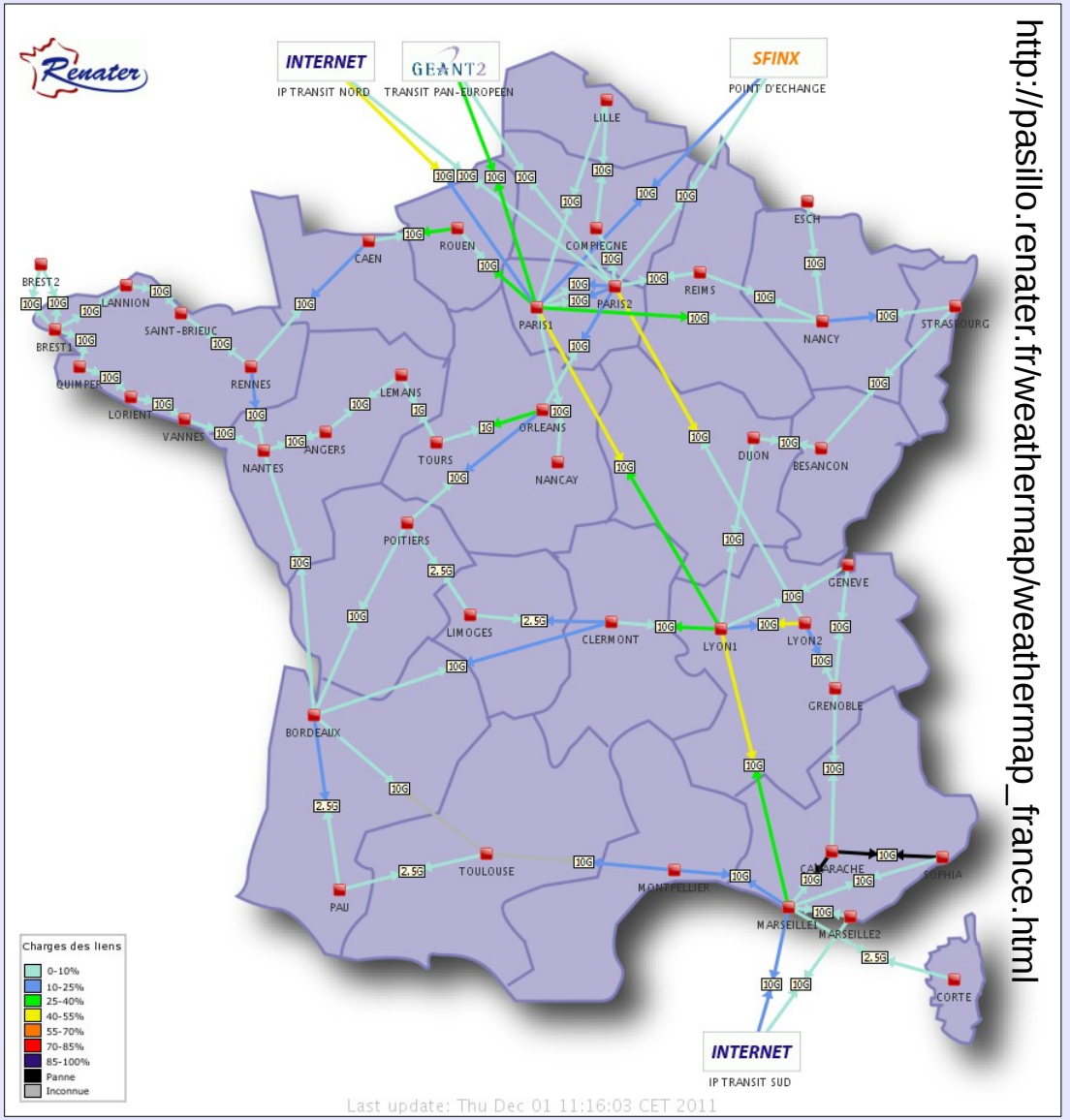
- réseau pour l'éducation et la recherche
- lien 10 Gb/s à des Points De Présence proches des sites

Connectivité des sites :

- dépend de la connection au POP
- p. ex. LPSC : 700 Mb/s

LHC Open Netwrok Environment :

- Projet international, Renater en FR
- Complémente LHCOPN
- Lisaions régionales améliorées et dédiées (“10 Gb/s pour tous ! “)
- Phase de prototype (POP, connection)
- Orsay, Nantes, (+ tard Marseille, ...)



Tests de transmission à 100 Gb/s

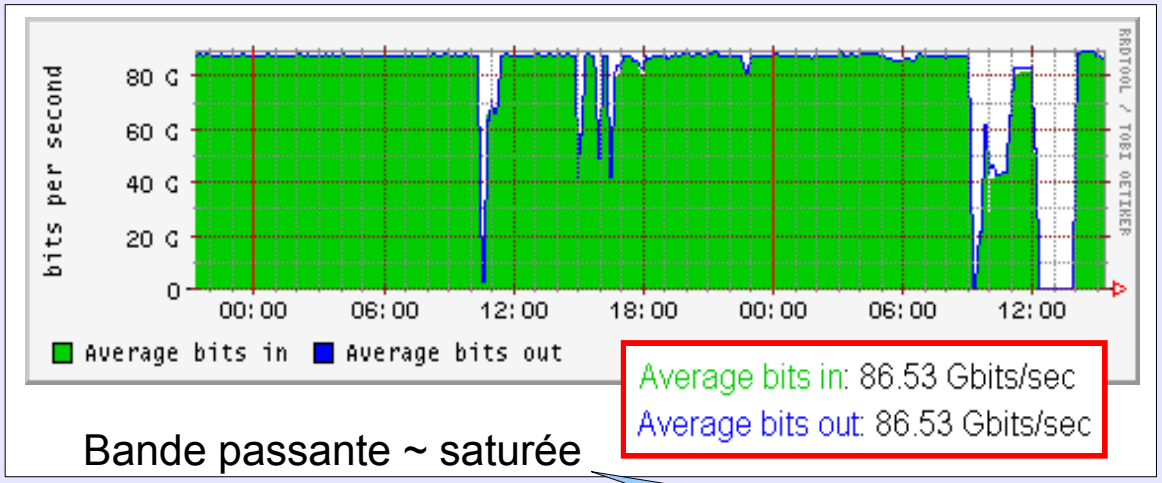
Partenaires : RENATER, CERN, IN2P3-CC and Ciena

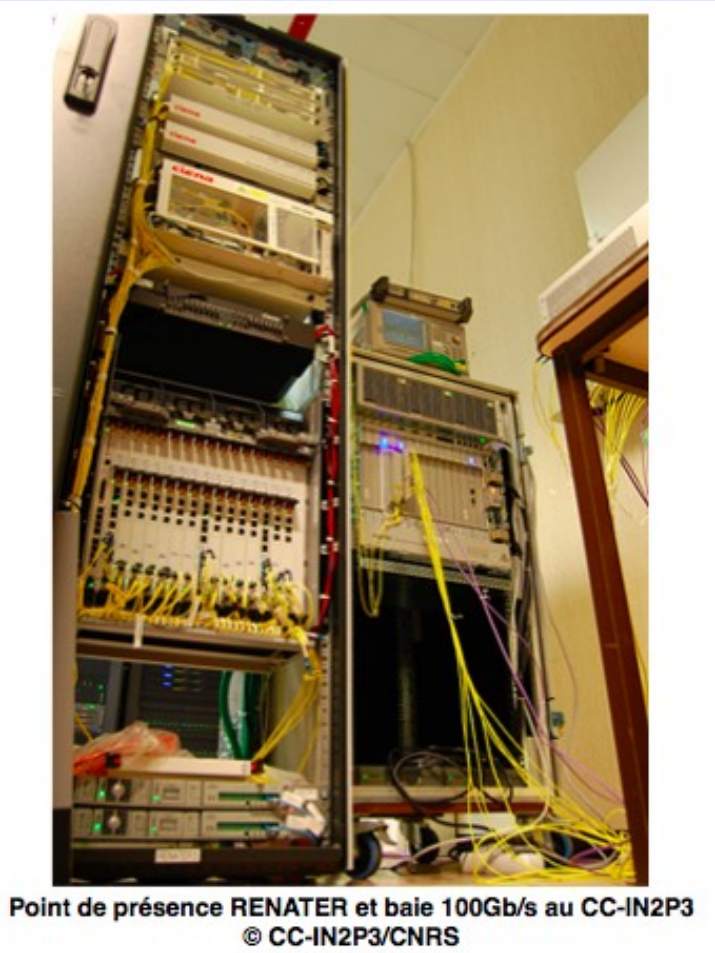
But : valider une possible upgrade de liens RENATER à 100 Gb/s

- Sur une simple longueur d'onde
- Sans changer les infrastructures actuelles

Tests : fin 2010 entre le CC-IN2P3 et le CERN

- Traffic de production
- Bons résultats (912 TB copié en un jour)



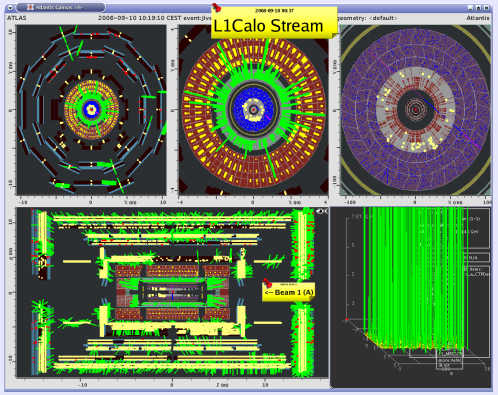


Point de présence RENATER et baie 100Gb/s au CC-IN2P3
© CC-IN2P3/CNRS

D-day « LHC first beam »

- Jour du grand démarrage du LHC le 10 sept. 2008

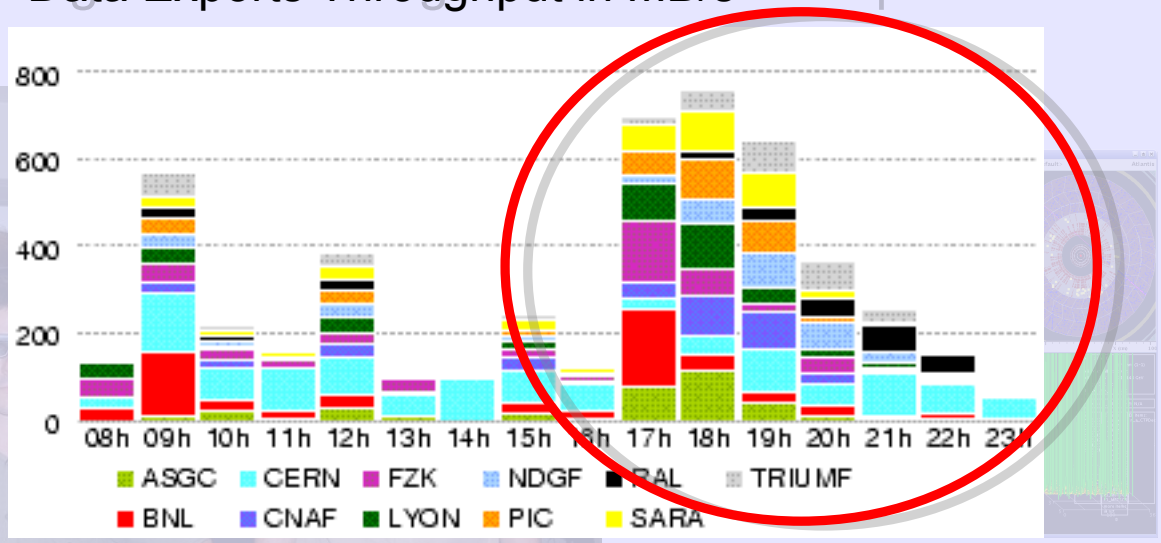
Simone Ale Alexei



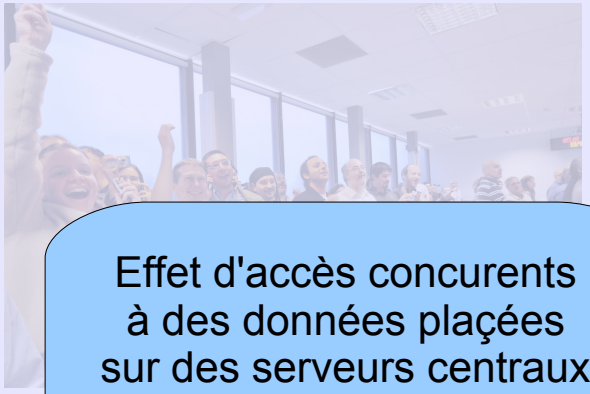
D-day « LHC first beam »

- Jour du grand démarrage du LHC : 10 sept. 2008

Data Exports Throughput in MB/s

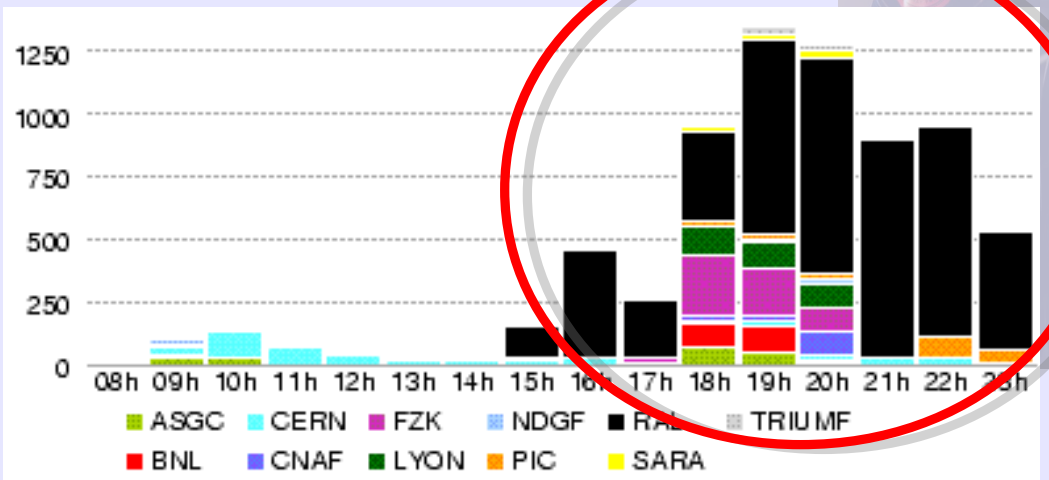


Dans la *control room* du computing (1er étage)



Effet d'accès concurents à des données plaçées sur des serveurs centraux et de l'activité des utilisateurs

Number of errors

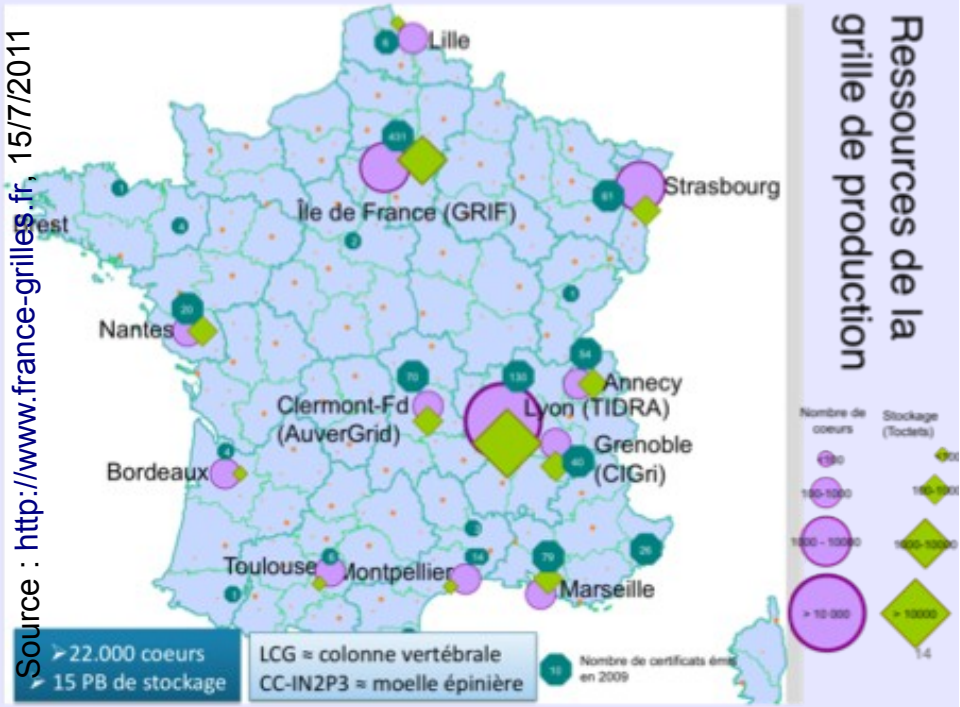


Les communautés hors LHC

Ou le Groupement d'Intérêt Scientifique

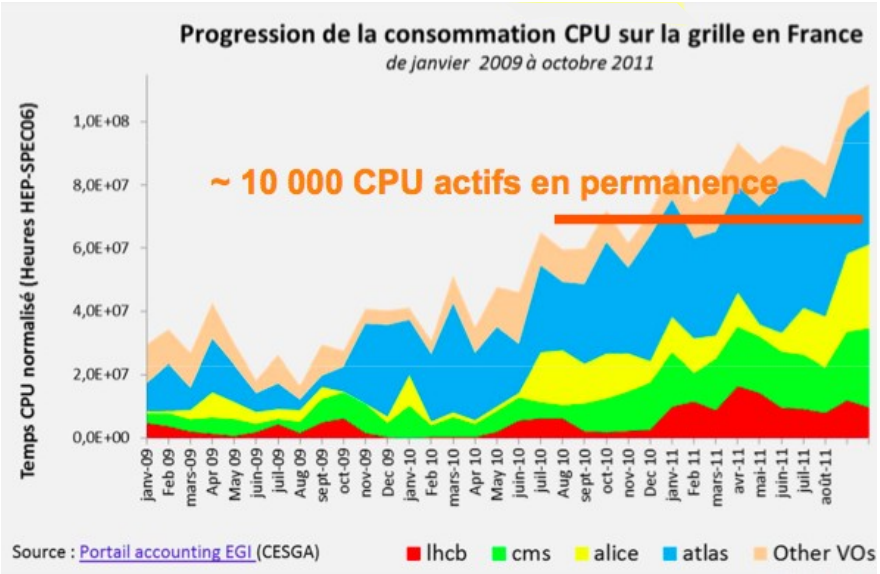
- NGI (Initiative nationale de grille) Française
  - Coordonne, optimise et met à dispoition les ressources des partenaires dans les labos de recherche
  - V. Breton, directeur de l'Institut des Grilles (pilotage)
  - 18 sites en production

- Disciplines : HEP-LHC et beaucoup d'autres
  - Sciences de la planète,
  - Sciences de l'univers
  - Sciences du vivant
  - Imagerie médicale
  - Sciences Humaines et Sociales
  - ...

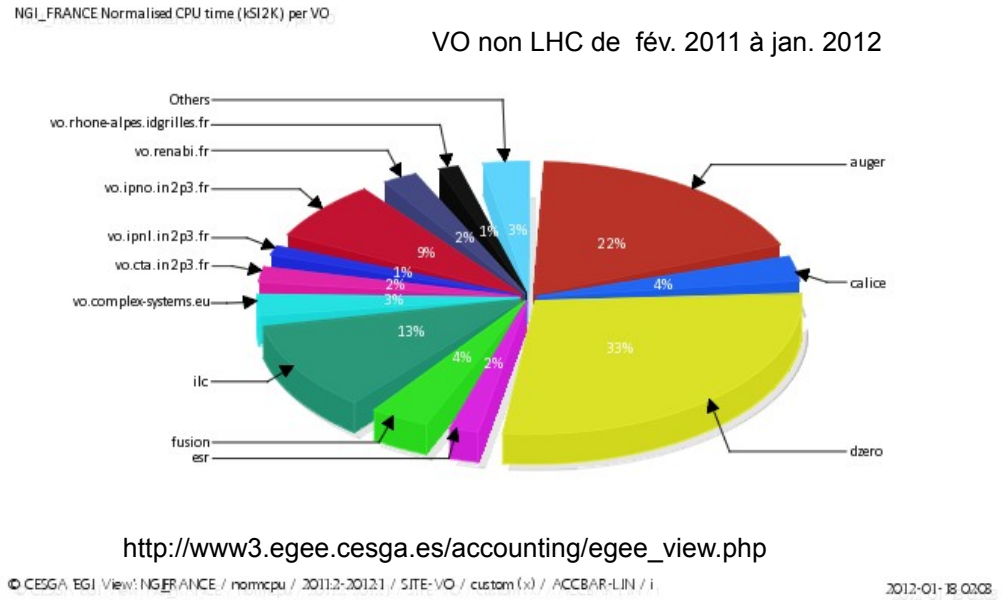


Activités en France

En 2011, le calcul non LHC représente 15% de la consommation CPU en France.



Source : F. Chollet (LAPP), réunion des sites LCG-FR, automne 2011

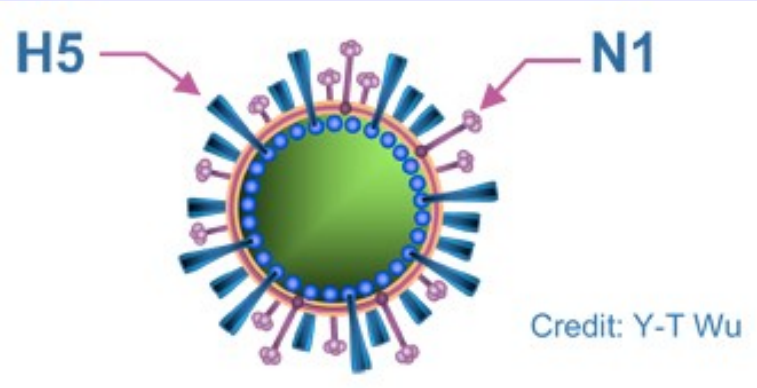


La VO nationale

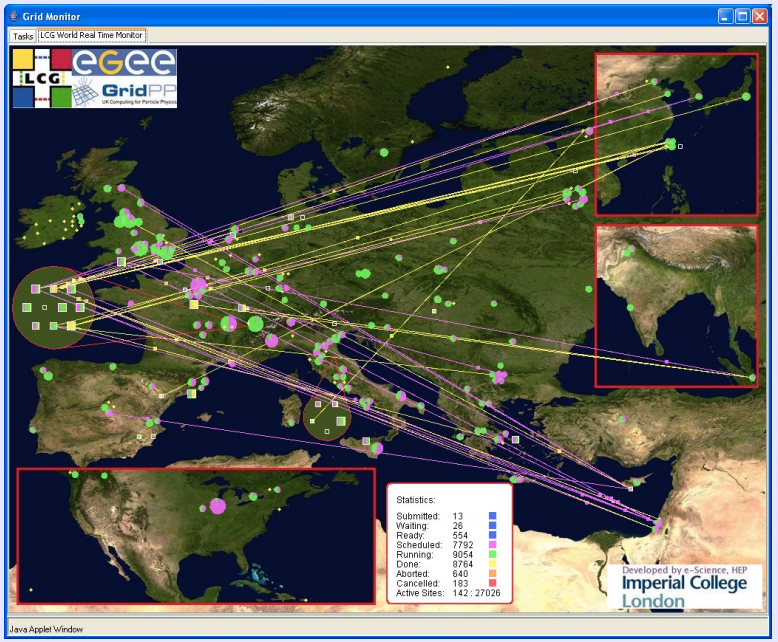
- Fournir un accès à la grille (vo.france-grilles.fr)
- Mise à disposition d'une infrastructure (production et stockage)

Grippe aviaire

- Calculs de probabilité d'accrochage de molécules sur les sites actifs du virus pour le bloquer
- Accélération de la recherche avec utilisation de la Grille de calcul Européene EGEE (EGI)
- Avril 2005 :
  - 300 000 calculs de probabilités
  - 100 années de calcul sur un ordinateur



Sites participants



Emergence du *cloud computing*

Le *cloud* : décorrélation de l'infrastructure machine (réelle) des ressources présentées aux utilisateurs (virtuelle)

- Fourniture des ressources à la demande
- L'utilisateur alloue/ désalloue selon ses besoins
- On ne compte (facture) que les ressources consommées

– Exemple : Amazon (services, CPU, stockage, ...)

Testé par ATLAS (avec succès) pour générer des événements

Bénéfices côté utilisateur (ou d'une expérience)

– Déploiement d'environnement spécifique à la demande sur la durée voulue

- Grande souplesse
- Pics d'analyse avant conférences
- Petite structure de recherche

Bénéfices côté "administration des sites"

– Pas de "1 serveur, 1 service" (optimisation ds ressources, coût, gérance)

– Séparation ressources/applications (plus facile à maintenir)

Le *cloud* dans notre discipline

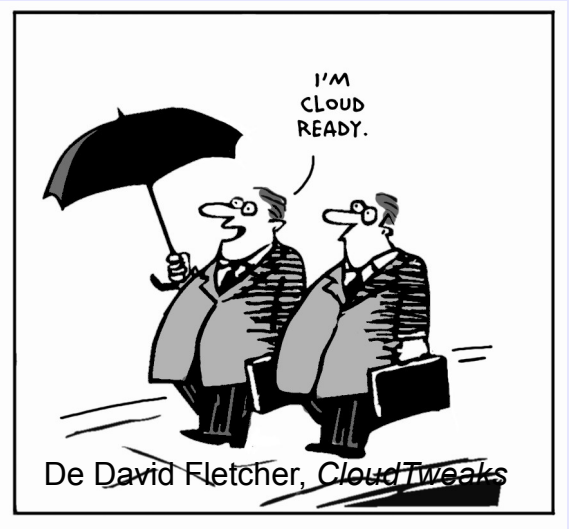
- *Cloud* privé, académique, hybride, fédération de clouds (*sky computing*), …
- *Infrastructure/Plateforme/Software As A Service*

CAPRI : Développement d'un cloud académique (Cloud pour la Recherche et l'Innovation)

- CCIN2P3, LAL (StratusLab), IRFU et IPGP → IAAS
  - "je veux 20 machines avec telle configuration"
  - Aussi des services (sites web, ...)
- Intégration dans projet plus large (EGI) avec fédération de clouds IAAS
  - Le cloud sera intégré dans la grille

Le cloud sans les orages !?

- LxCloud au CERN : 400 VMs (ATLAS, CMS, ...)
  - Intégré au batch !
- Cloud dans les services informatiques au CNRS, au CC-IN2P3 (CAPRI), dans la NGI (cloud task force), ...



De David Fletcher, *CloudTweaks*

Le Centre de Calcul de l'IN2P3

- Installé sur le campus de la Doua en 1986
  - Depuis Jussieu où il a été crée dans les années 60

- Site principal de la grille
  - 4 fois site T1 du LHC et T2
  - Seul site WLCG si complet dans le monde
- Récemment ouvert à :
  - L'astrophysique (~40% des ressources du CC)
  - La biologie
- Services associés : hébergements de sites internet (p.ex. le palais de la découverte), mails, visio-conférences, etc

dont les machines de services centraux qui calculent les payes des agents CNRS

Principaux services au CC-IN2P3

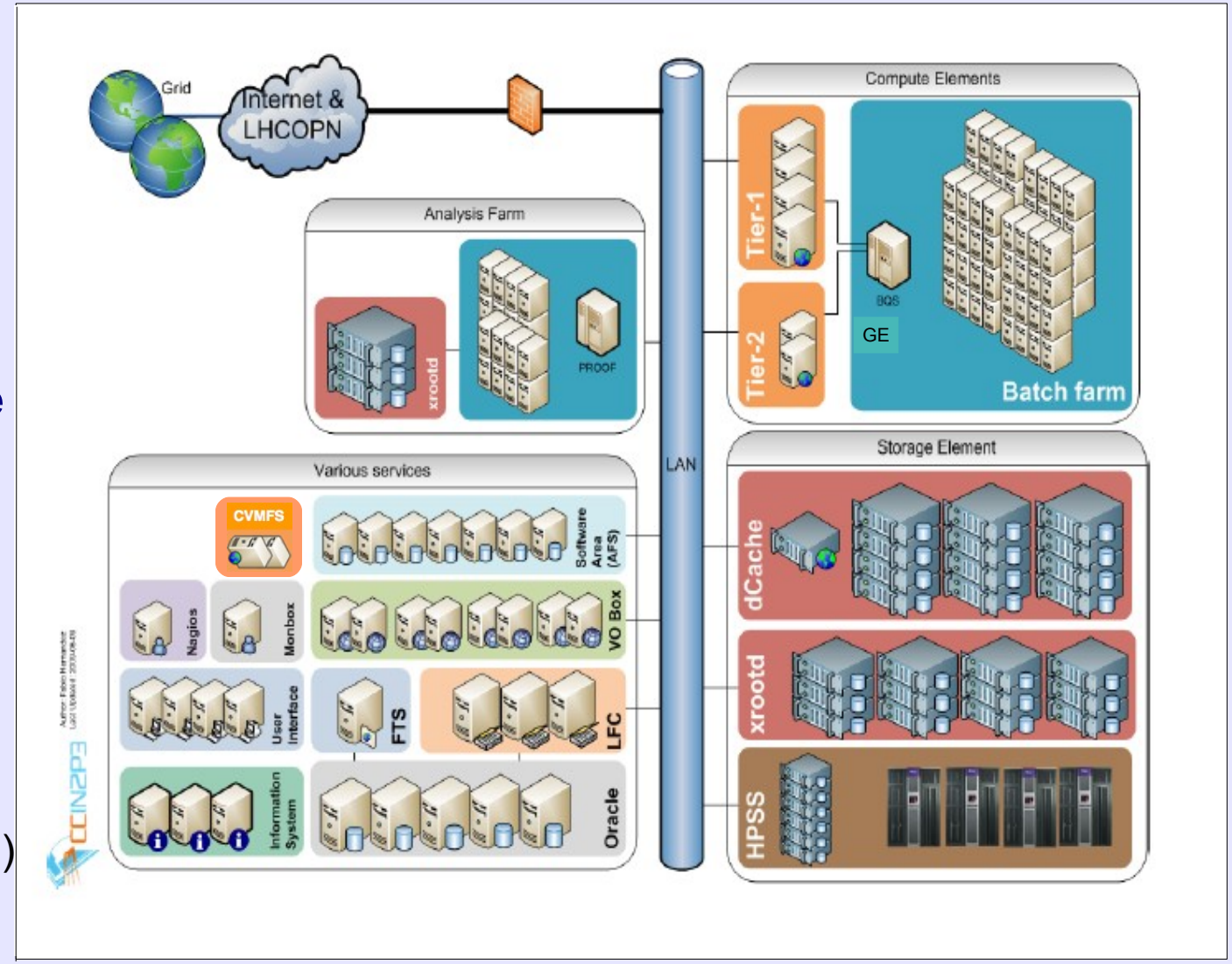
Infrastructure mutualisée

- Grille
- Local

Services de grille

Services locaux

- Zones AFS individuelles
- Stockage semi-perm. (sps)
- Ferme parallèle
- ...



Le CC-IN2P3 en quelques images

Accès doit être < 12 min.
sans casque anti-bruit !

Petit lecteur de bandes
pour les données des télé-
scopes HESS (arrivent
par valise de la Namibie) !

20°C

17°C

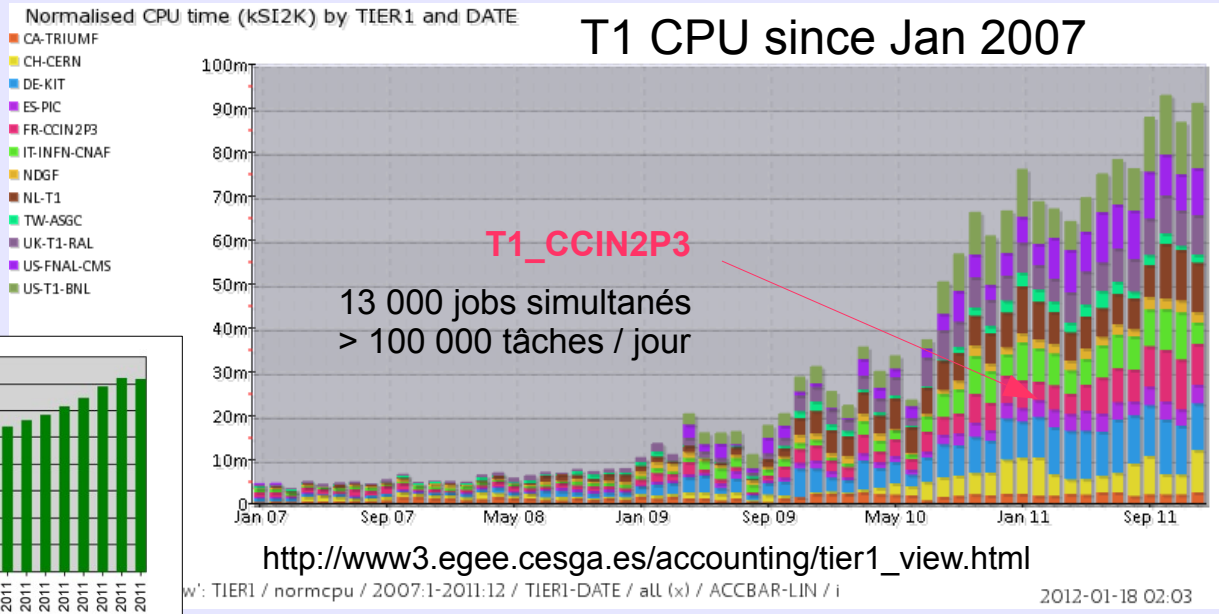
Soufflement d'air
froid au sol

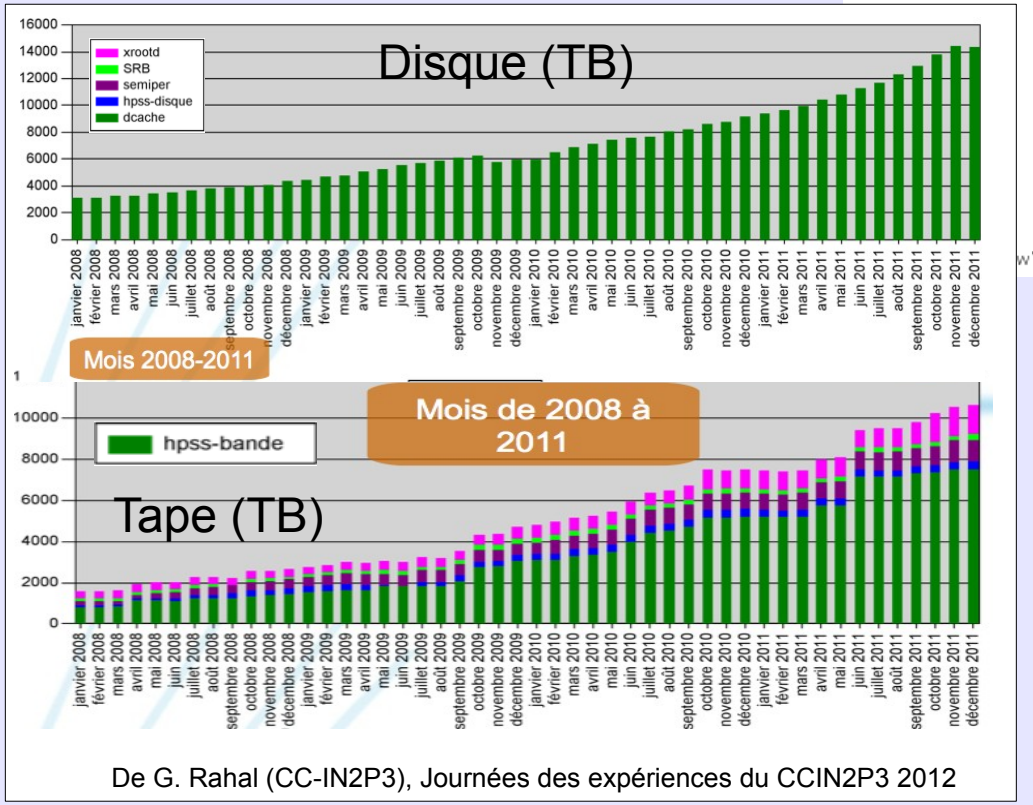
Vue aérienne de la salle machine (800m^2)
centrée sur des serveurs de disks

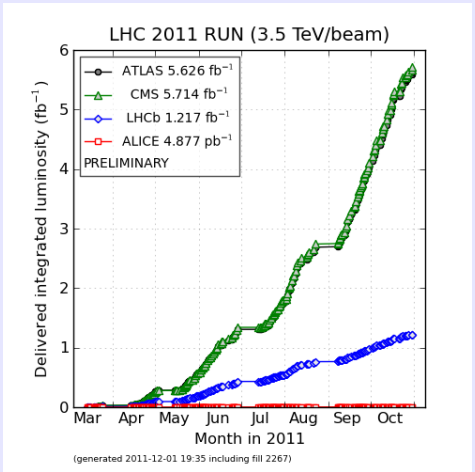
Librairie de bandes de 10 000 slots;
1(5)TB/slot, lecture 100 MB/s

Avant 1991, bandes
plaçées à la main
à la demande des users

Activités en forte croissance au CC-IN2P3







Operations at CC-IN2P3

- MOU requirements for T1
- Rythme 24/7
  - Système de tickets
  - GGUS
- Système de couverture hors “heures de bureau”

MOU between CERN and the institutions participating in WLCG

(CERN-C-RRB-2005-1/Rev. April 2009)

Service	Maximum delay in responding to operational problems			Average availability[5] measured on an annual basis	
	Service interruption	Degradation of the capacity of the service by more than 50%	Degradation of the capacity of the service by more than 20%	During accelerator operation	At all other times
Acceptance of data from the Tier-0 Centre during accelerator operation	12 hours	12 hours	24 hours	99%	n/a
Networking service to the Tier-0 Centre during accelerator operation	12 hours	24 hours	48 hours	98%	n/a
Data-intensive analysis services, including networking to Tier-0, Tier-1 Centres outwith accelerator operation	24 hours	48 hours	48 hours	n/a	98%
All other services[6] – prime service hours[9]	2 hour	2 hour	4 hours	98%	98%
All other services[6] – outwith prime service hours[9]	24 hours	48 hours	48 hours	97%	97%

The response times in the above table refer only to the maximum delay before action is taken to repair the problem. The mean time to repair is also a very important factor that is only covered in this table indirectly through the availability targets. All of these parameters will require an adequate level of staffing of the services, including on-call coverage outside of prime shift.

Les physico-informaticiens du CC-IN2P3

- Intensification des activités grille
  - Éloignement des utilisateurs
  - Complexification des outils

Des jobs sont vus comme *successful* aux sites alors qu'ils sont *failed* pour les VOs !

- Création de passerelles directes entre l'informatique et la physique
  - Acceuil de post-doctorants chercheurs en physique (50% soutien informatique)
  - Chaque expérience LHC

- CC-IN2P3 signataire des publications LHC !
  - 1ere publication ALICE

Catherine Biscarat, physico-informaticienne au CC-IN2P3

Les informaticiens font partie d'une espèce à part entière, diverse et variée, c'est bien connu. Car derrière leur code,

Lettre IN2P3 Informatique, mai 2009

Le CC-IN2P3 chauffe

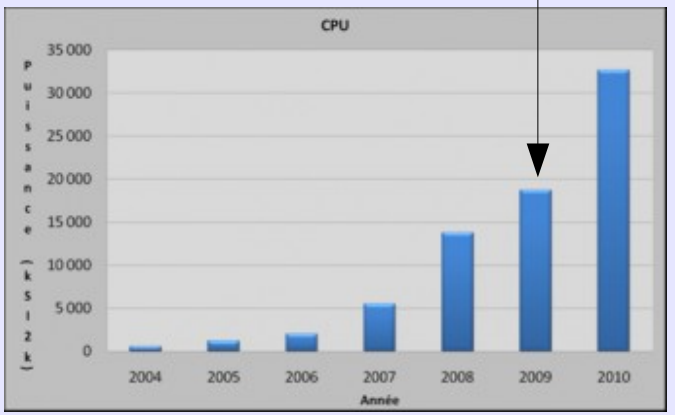
Dissipation
900W/m^2

Puissance électrique

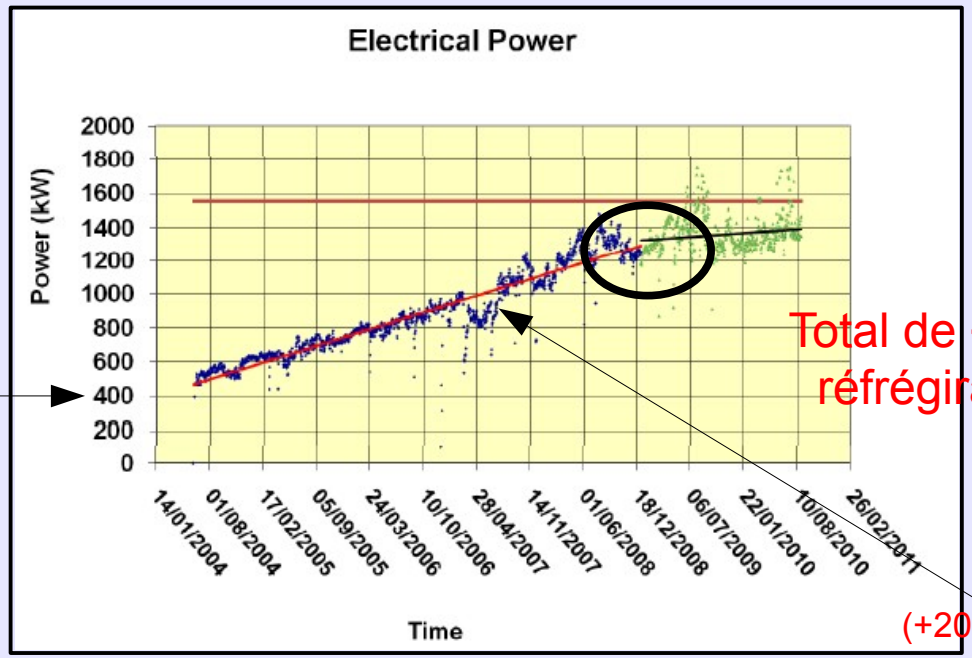
- Difficulté majeure dans les centres de calcul
- Assurer le refroidissement des machines

→ améliorer :

- les infrastructures (ajout de groupes froids)
- la dissipation de chaleur (soufflage d'air vs. confinement et climatisation dans les baies)



Puissance CPU au CC-IN2P3 depuis 2004



Puissance électrique en 1986

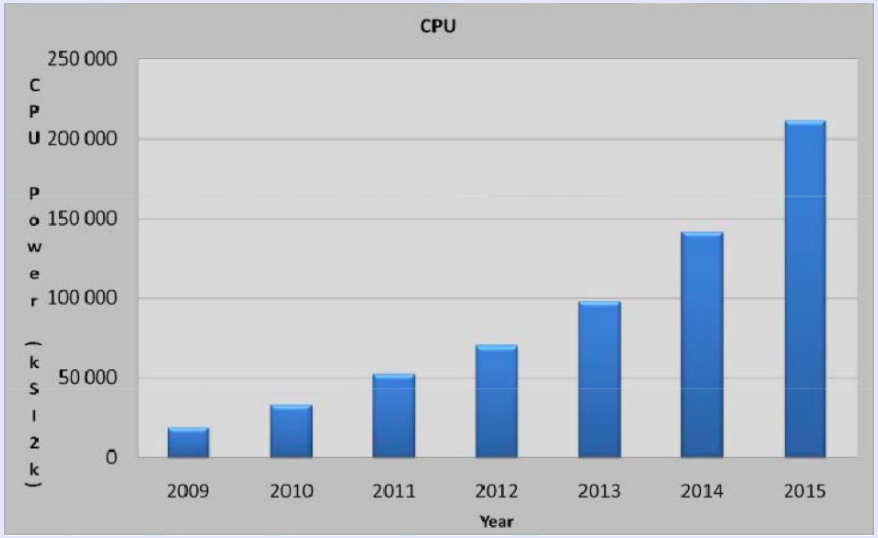
Total de ~500 000 réfrégirateurs !!!

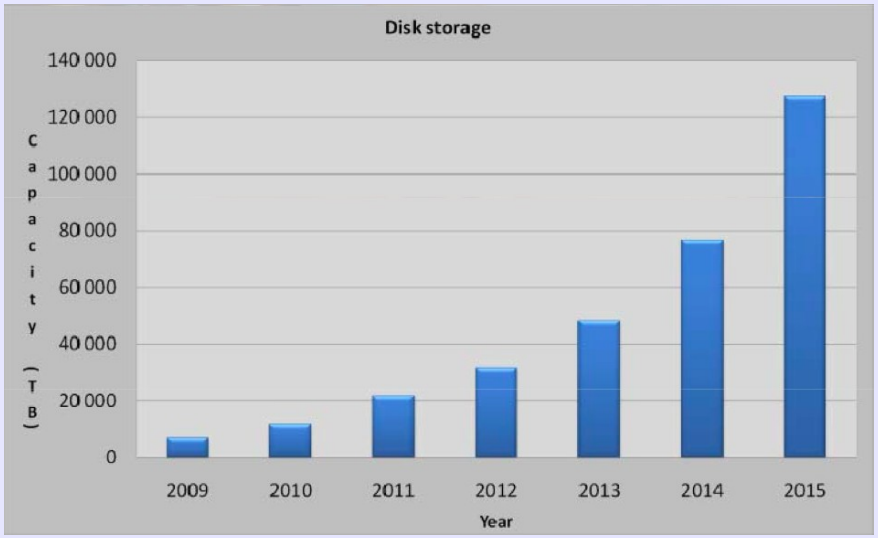
(+20 par jour)

Capacités de la nouvelle salle

Projections

- Besoins de ~40 groupes standards
- Engagements auprès du LHC
- Prise en compte d'un upgrade du LHC
- Augmentation importante des ressources pour l'astroparticule (LSST)





- Salle modulaire :
  - → 2011 : 50 racks et 600 kW (capacité ~ à l'ancienne salle)
  - → 2015 : 125 racks et 1.5 MW
  - → 2019 : 6 blocks de 40 racks et 3.2 MW

Construction de la nouvelle salle

- Financement : CNRS et région Rhône-Alpes
- Saga de cessation du terrain par l'université
- Début des travaux en avril 2010
- Livrée en temps et disponible en mars 2011

850m^2

Mise en production de la nouvelle salle

- Première salle toujours en production
- Actuellement :
  - Ligne de courant directe de la ville (1x20kV)
  - Redondance minimale → migration de WNs (pas de services critiques)
  - 150 kW en moins dans la première salle (retour dans une zone de confort)

- *green grid* : PUE = 1.47 (1.7 pour l'ancienne)

  PUE = (énergie consomée par le centre) / (énergie consomée par les serveurs)

- *green campus* : circuit de distribution d'eau chaude pour l'université (projet de cantine)

- Peu de centre académique offre ces capacités → carte maitresse pour notre communauté

Allées chaudes confinées de racks réfrigérés (1unité de cooling/rack)

Changement d'échelle : le site du LPSC

Le Laboratoire de Physique Subatomique et de Cosmologie de Grenoble

- UMR : IN2P3, INSIS, UJF et l'INPG
- 220 membres (50% enseignants et chercheurs)

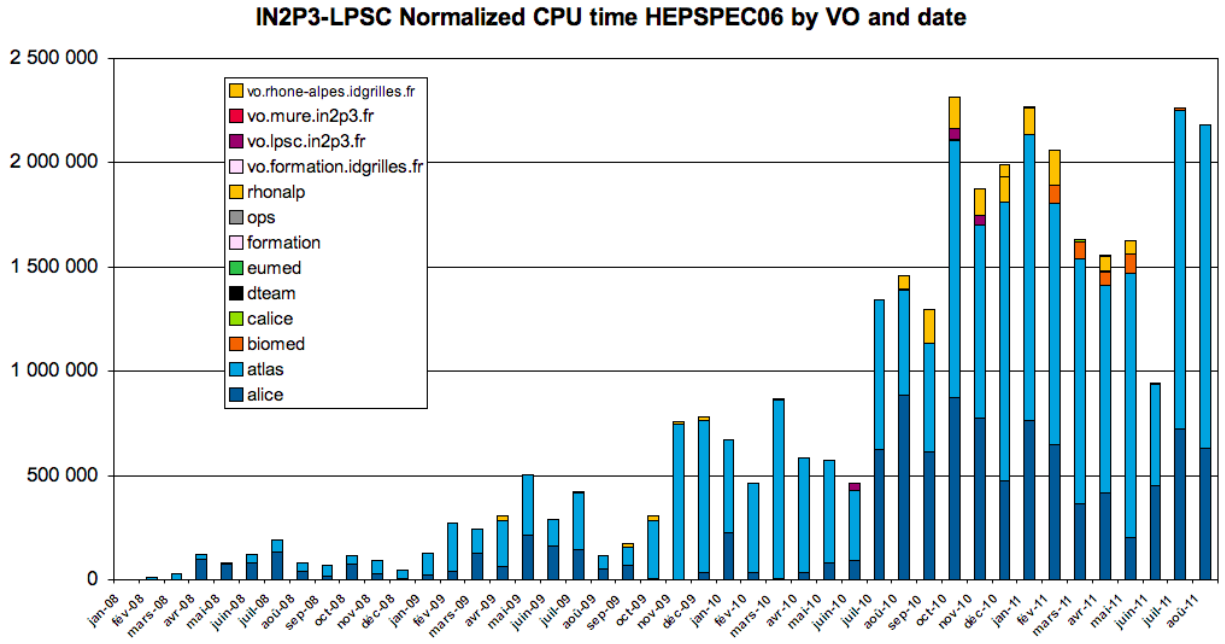
Le site grille

- Nov. 2007 : naissance du site WLCG
- Jan. 2008 : site en production
- Juill. 2011 : T2 de LCG (ALICE ¼ et ATLAS ¾)
- VOs : ALICE, ATLAS , Rhône-Alpes, biomed, calice, lpsc, eumed, MURE, fr-grilles
- Collaborations : CIMENT, TIDRA
- ~2 FTE

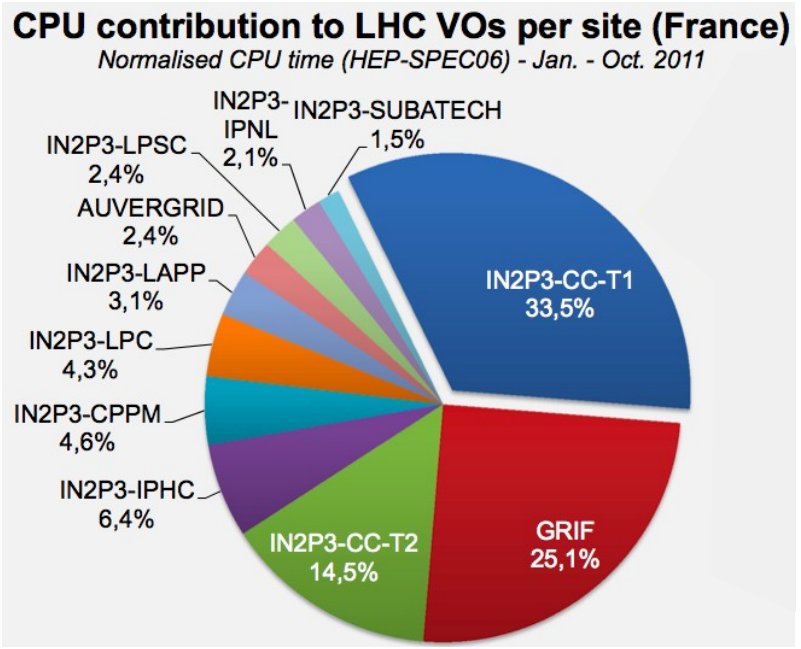
Salle informatique

- 8 baies (42 U) dont 5 pour le T2
- 120 kW électrique disponible
- 100 kW de refroidissement
- ~800 cœurs (Torque Maui)
- ~600 To net (DPM, xROOTd)
- Connexion réseau : 700 Mo/s

Utilisation du site du LPSC



Jan. 2008



Août 2011

Source : F. Chollet (LAPP),
réunion des sites LCG-FR, automne 2011

Requis pour un T2 :

- Disponibilité > 90% — LPSC
- Fiabilité > 95% — LPSC

Une initiative *verte*

Le contexte :

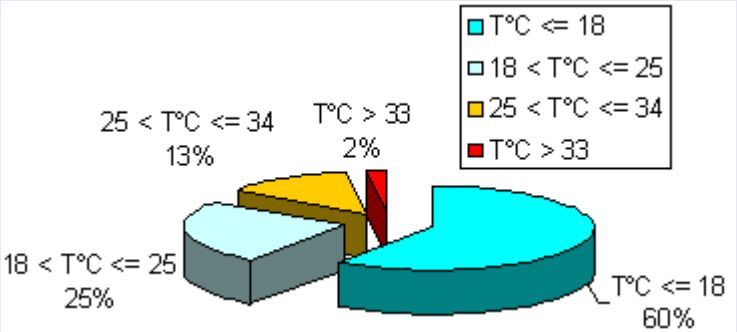
- En 2008, limites de climatisation atteintes
  → il a fallu repenser la salle informatique
- Coût électrique très important (60k€/an pour nos 8 baies), dont 30-50% pour la climatisation
- Choix d'une climatisation robuste, économique et verte :

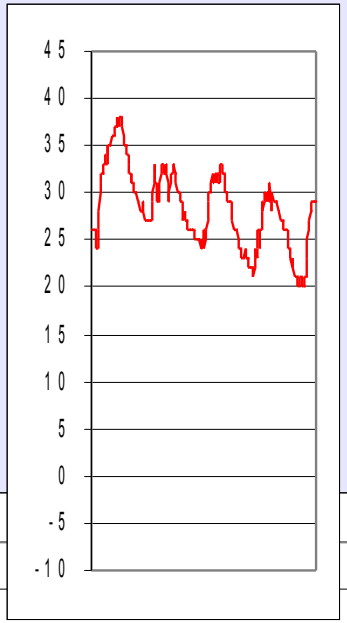
le free cooling à air direct

Les ingrédients :

- Un contexte météorologique attractif
- Une source d'eau industrielle à 16°C

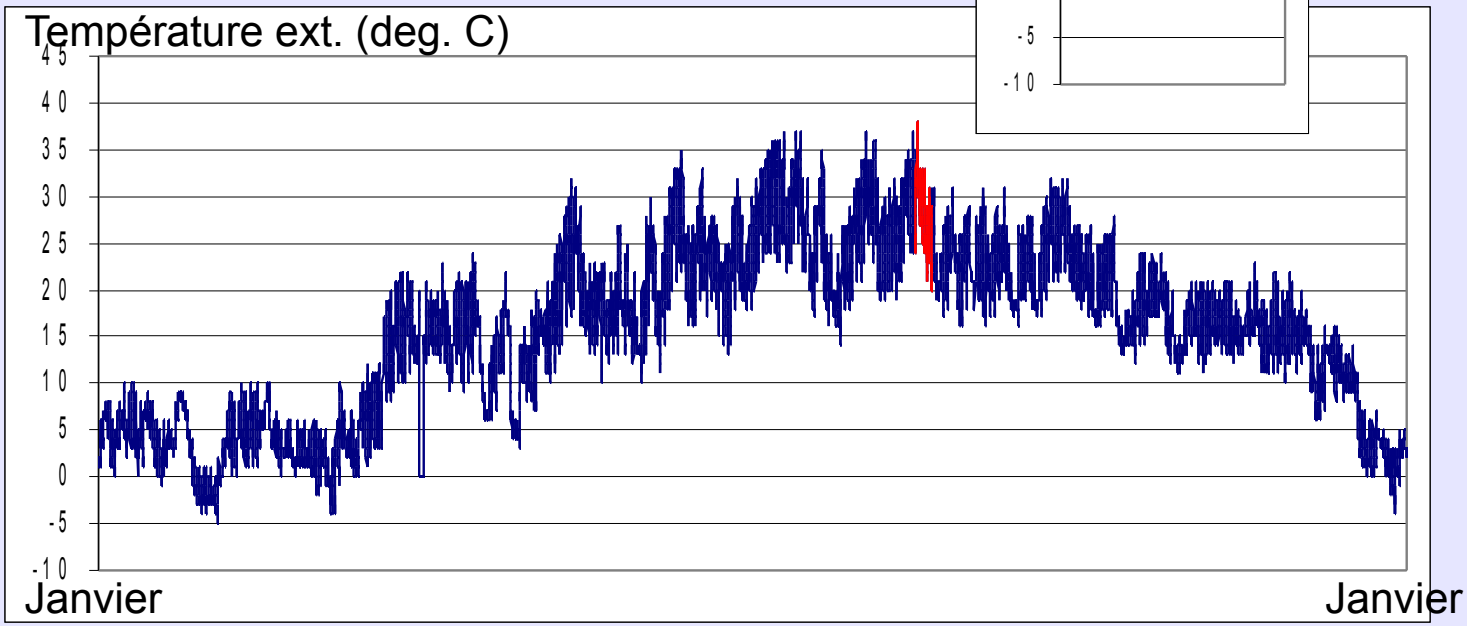
Données météorologiques au LPSC



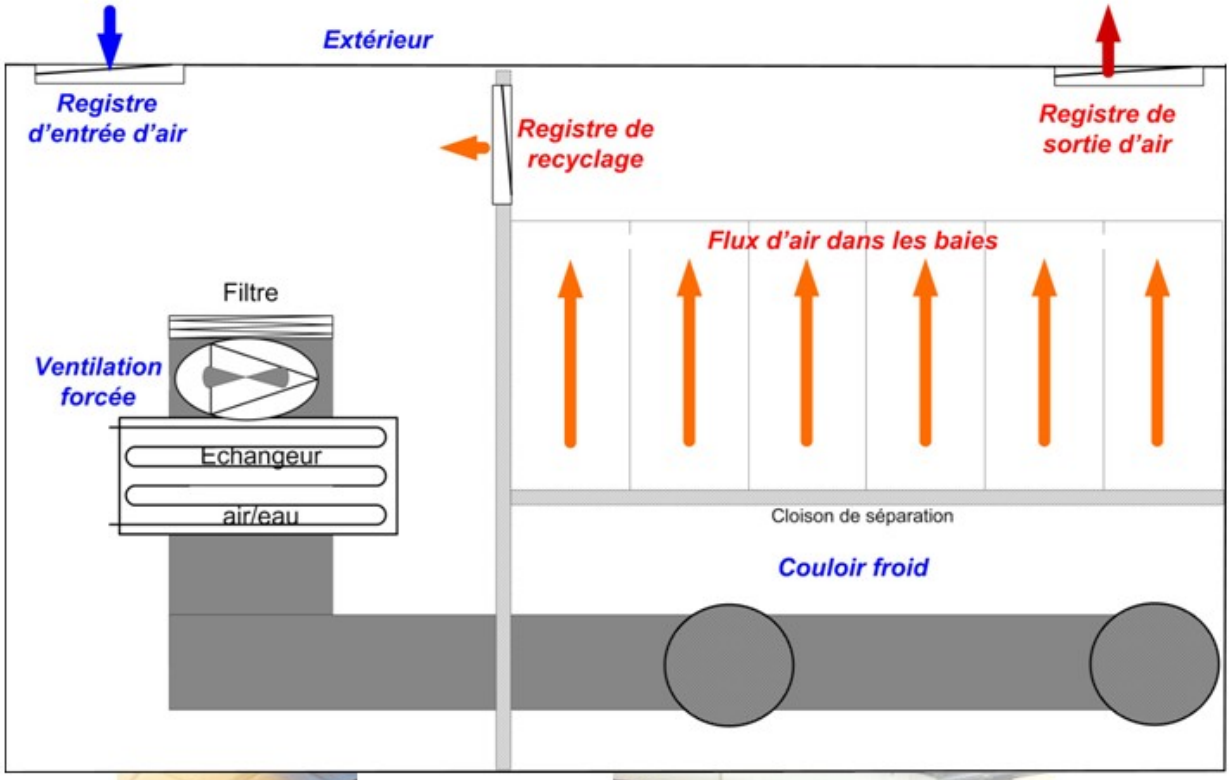


T_{ext} < 25° : 85% du temps

(chaleur des faces arrières)



Fonctionnement de l'écoclim du LPSC



T°consigne=[13-25]°C; face avant des machines

Régime canicule T°ext > T°arrière (ne s'est pas produit)

T°consigne = 25°C régulation par eau
avec l'échangeur et le recyclage d'air

Régime chaud 25° <T°ext < Tarrière 15% du temps

T°consigne = 25°C régulation par eau
Avec l'échangeur et l'air du dehors

Régime normal 13° < T°ext <= 25°C 45% du temps

T°consigne = T°ext
avec l'air du dehors

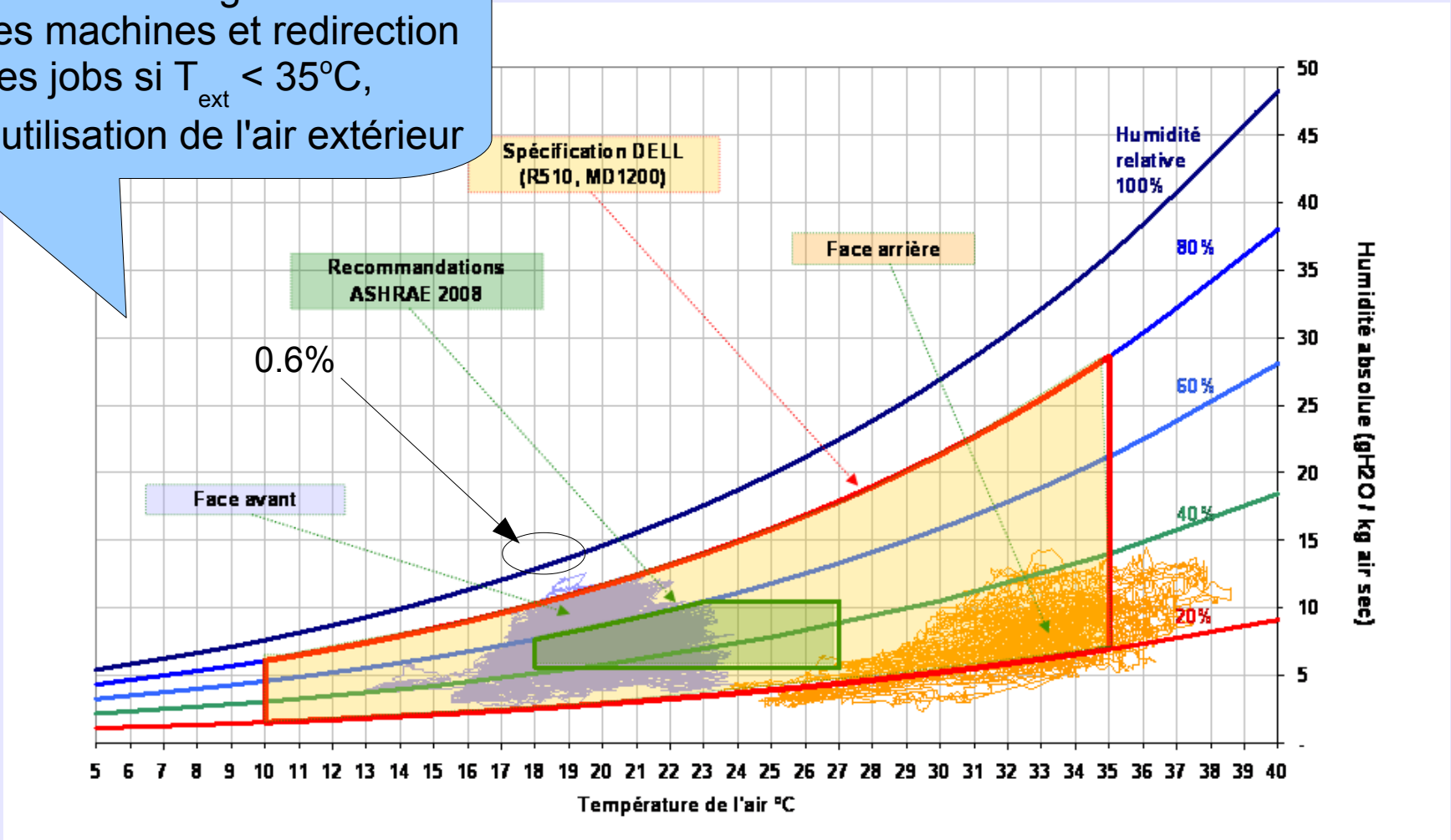
Régime froid T° ext <= 13°C 40% du temps

T°consigne = 13°C régulation par air recyclé
pour réchauffer une part d'air extérieur

Recommandations de fonctionnement

ASHRAE (American Society of Heating, Refrigerating and Air-conditioning Engineers)

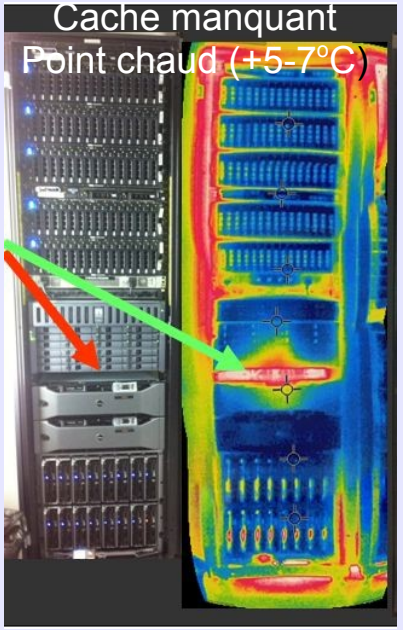
"Extreme freecooling" à CIMENT : arrêt des machines et redirection des jobs si T_{ext} < 35°C, sinon, utilisation de l'air extérieur



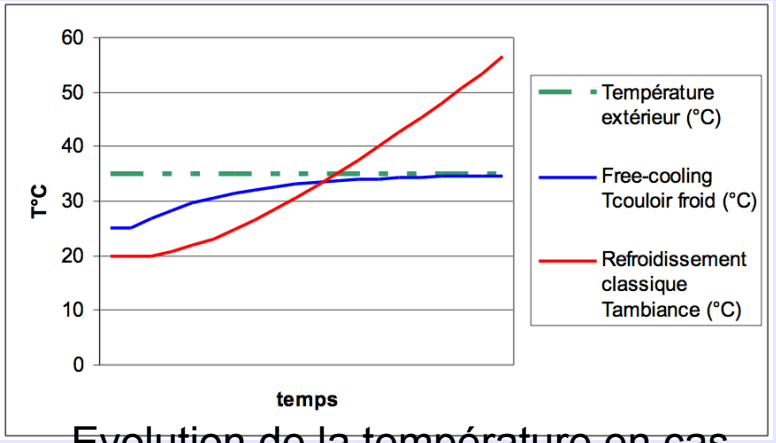
Pas de bandes magnétiques dans notre salle machine (sensibles à l'hygrométrie)

Observations sur cette solution

- Aucun arrêt de production dû à la climatisation depuis 2008
- Encrassement des filtres à changer souvent
  - Aucun incident sur le bon fonctionnement des machines
- Nécessité d'avoir une bonne maitrise des flux d'airs
  - Calfeutrage des allées
- Système ouvert qui évite un augmentation dramatique de T°
  - +1°C / min. dans des circuits fermés
  - < ½ heure pour atteindre ToC de sécurité des serveurs



- La puissance électrique nécessaire au refroidissement des servers est de 3kW, pour 60kW dépensée pour les serveurs.
  - PUE = 1.04
  - Pertes : ~10 fois moins qu'un système classique



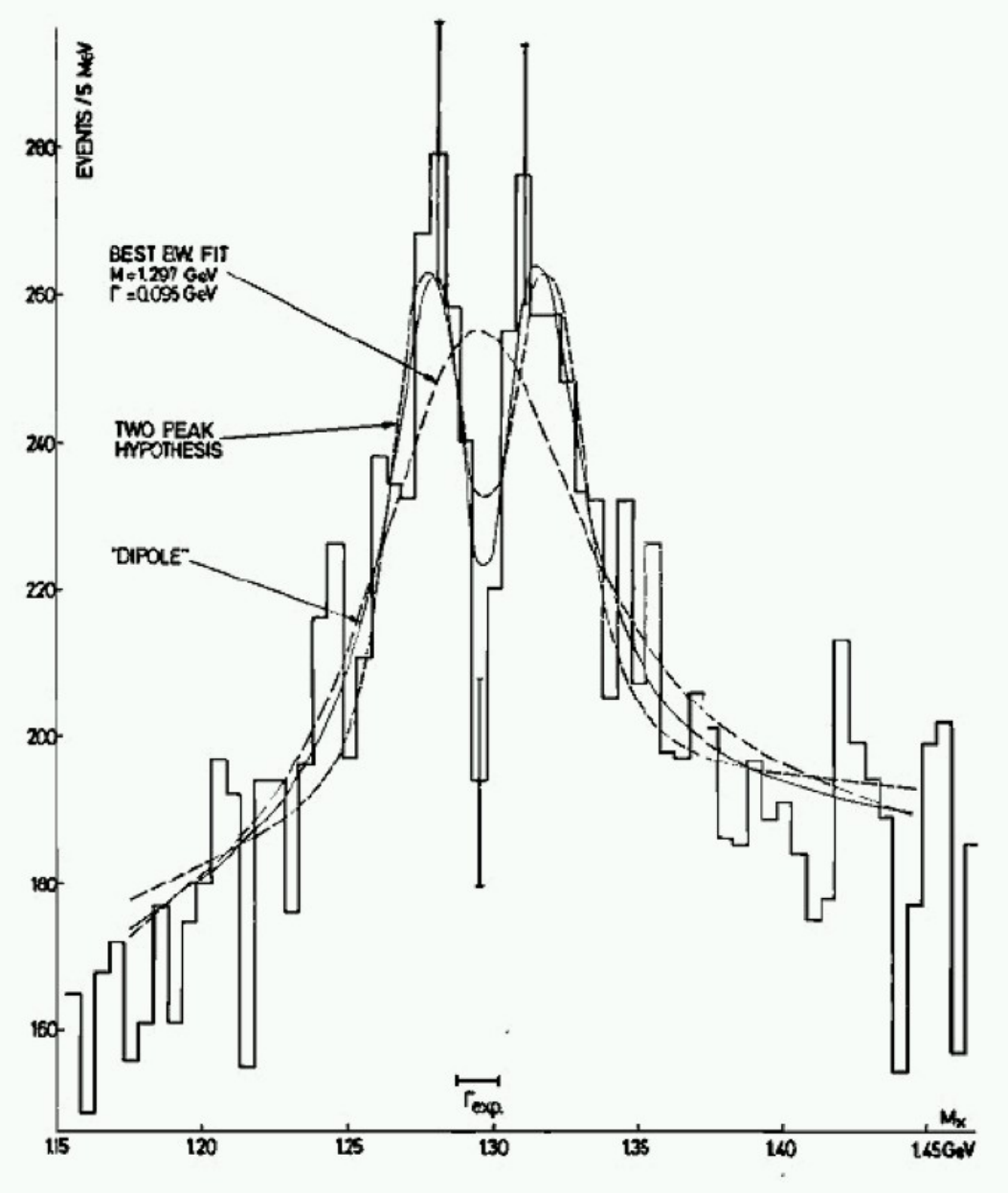
Evolution de la température en cas de coupure de climatisation

“Accidents de parcours”

En 1968, deux expériences au CERN ont « découvert » le *splitting* (séparation en deux) de la particule a_2 (l'une des nombreuses particules du modèle standard) . Une « découverte » majeure.

Après des années d'excitation, il était clair que cette découverte était fausse - la particule a_2 donne un seul pic, comme toutes les autres particules.

J'ai entendu plusieurs rumeurs sur l'origine du double pic dans la publication initiale, dont une qui dit que les auteurs auraient accidentellement mélangé deux versions de leurs fichiers avec des calibrations différentes (et deux positions différentes pour le même pic) ...



En guise de conclusion

- L'informatique en HEP a beaucoup évoluée (croissance quantité de données)
- Actuellement : la grille est notre outil
  - Beaucoup de travail sur le derrière de la scène
  - Ca marche (reprocessing en temps, papiers en temps)
  - Mutualisation des ressources (biologie, SHS, etc)
- Evolution croissante : vers le *cloud* (self-service pour les expériences)

Dans les centres :

- Le facteur limitant : infrastructure électique/refoidissement
- France très bien placée (CC-IN2P3)
- Approches *vertes* (p.ex. ecoclim LPSC)

La communauté :

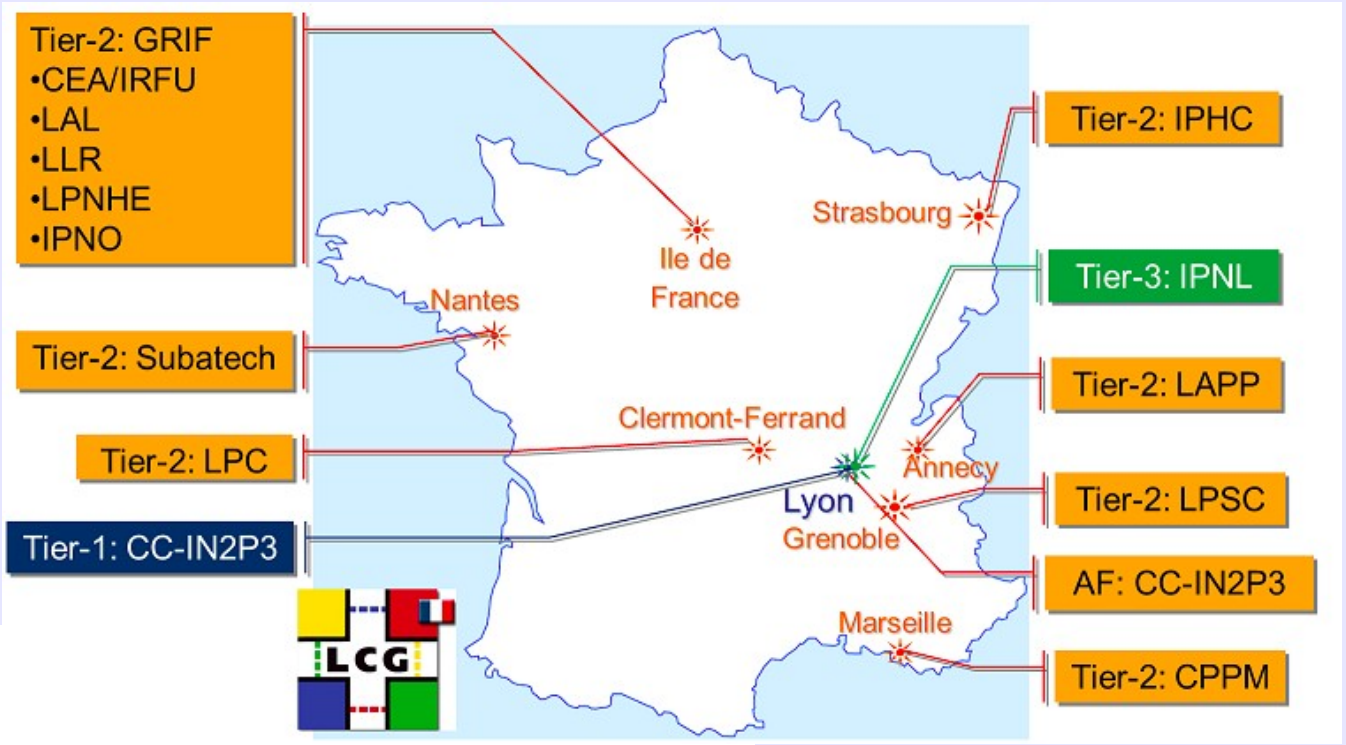
- Éloignement des utilisateurs/administrateurs
- Bonnes relations entre les sites FR

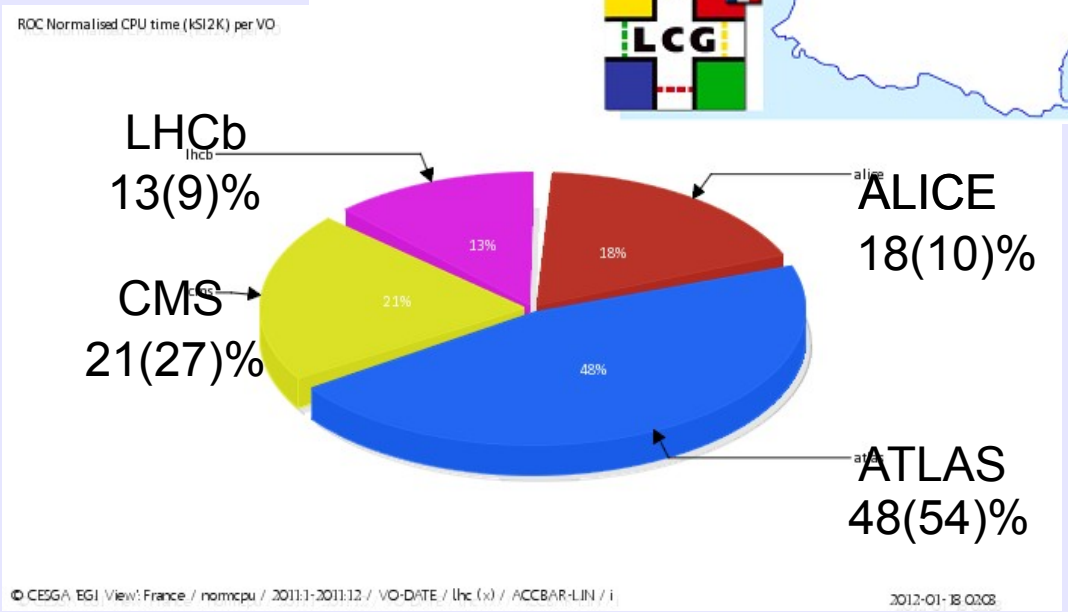
Back-up

"I think there is a world market for maybe five computers."
– Thomas Watson, chairman of IBM, 1943.

"There is no reason anyone would want a computer in their home."
-- Ken Olson, president, chairman and founder of Digital Equipment Corp., 1977.

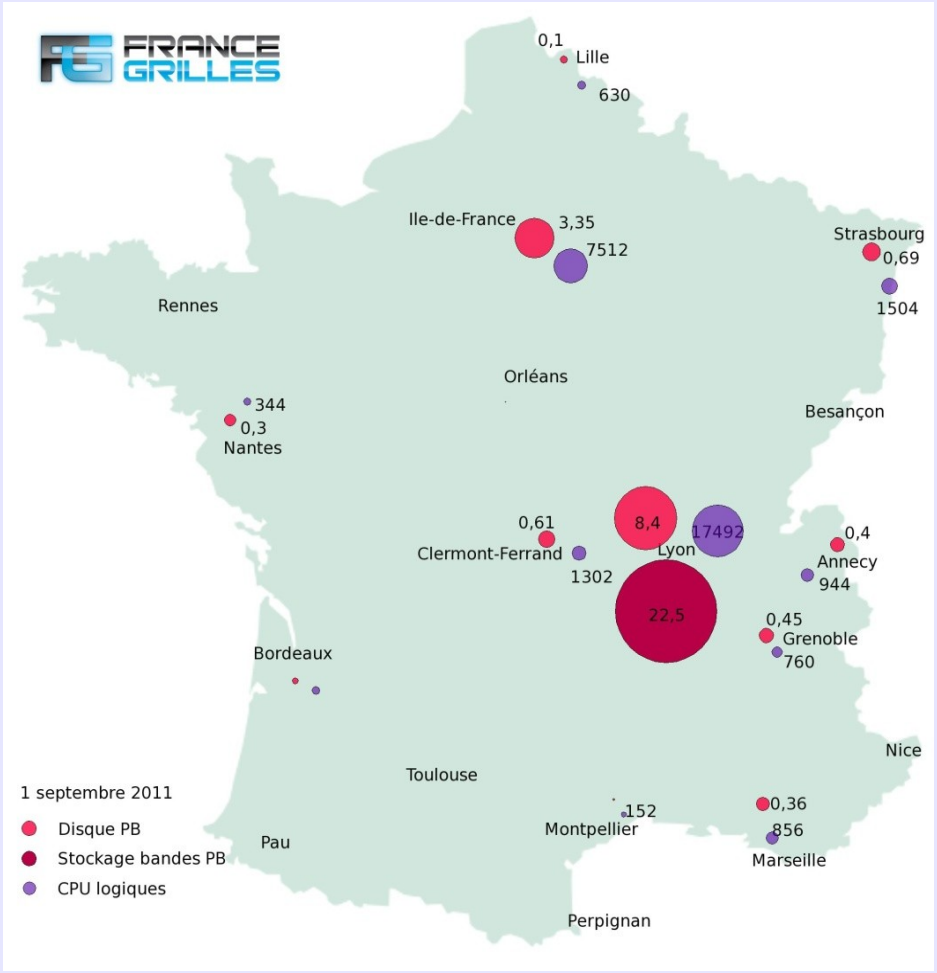
Centres de calcul WLCG en france





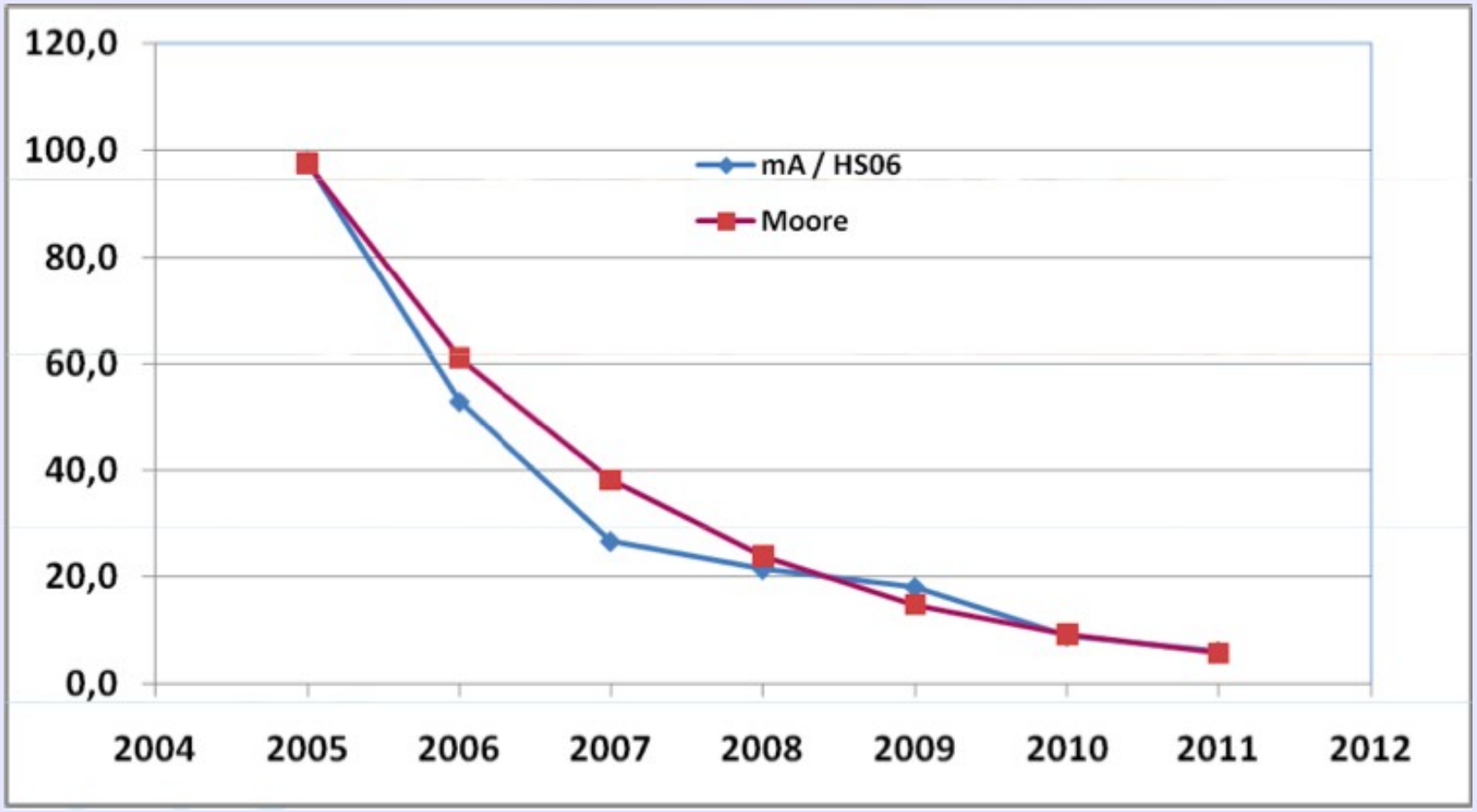
CPU dépensée sur les sites FR
par VO et pour l'année 2011
(pour tous les sites WLCG)

France Grilles



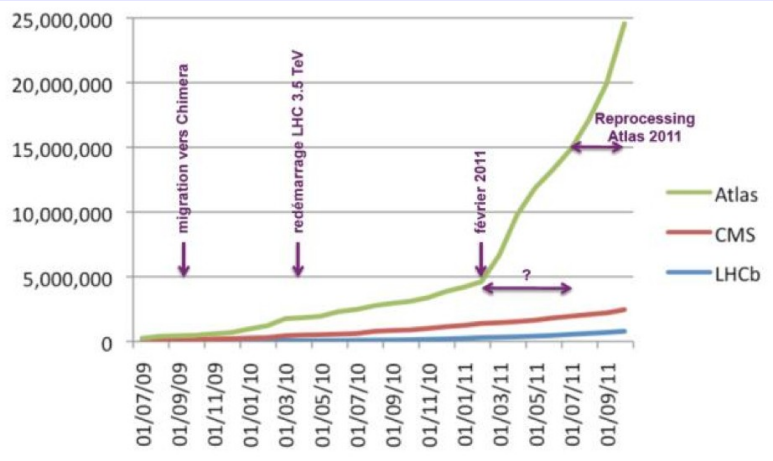
Source : G. Romier

Infrastructure au CC-IN2P3



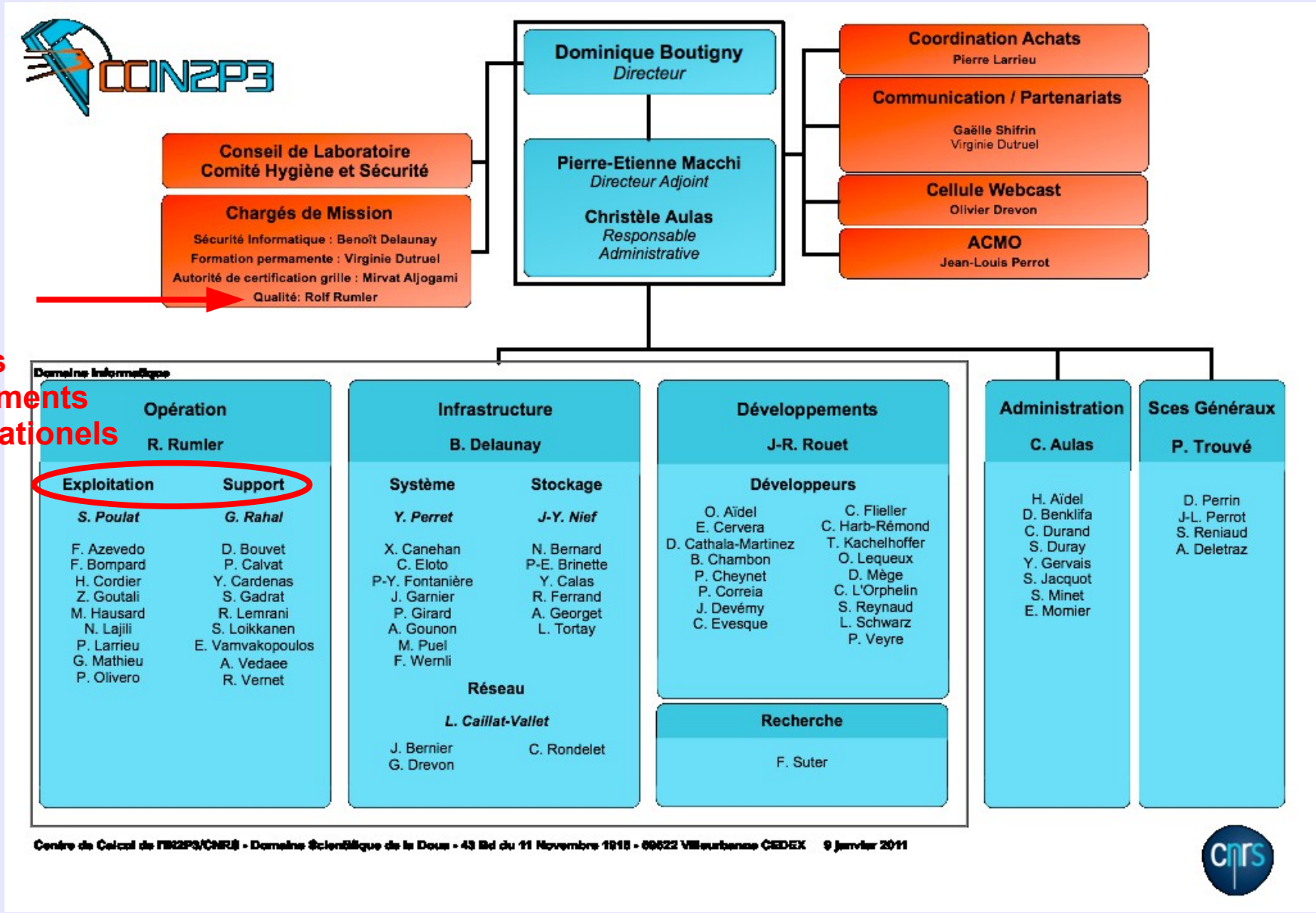
Ref. : D. Boutigny (CC-IN2P3), “réunion des expériences du CC-IN2P3 2011”

Activités en forte croissance au CC-IN2P3



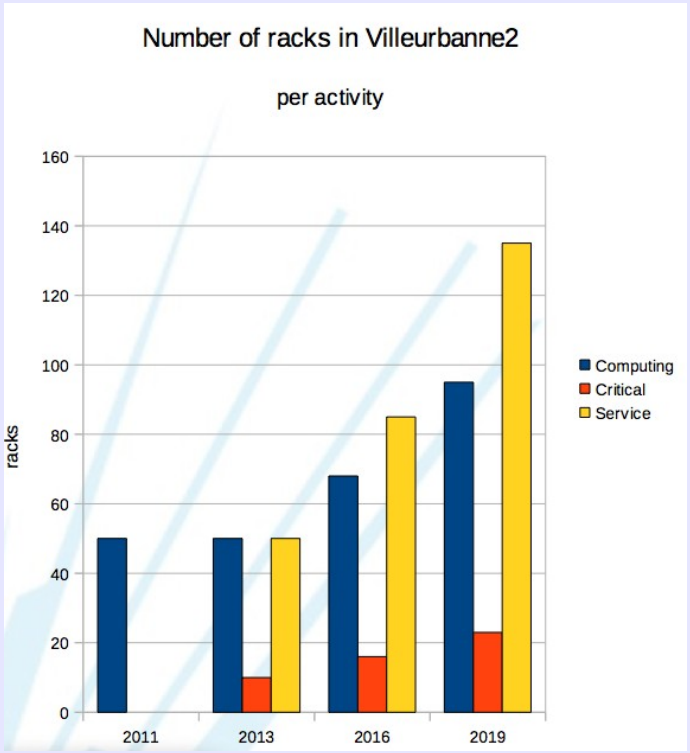
P. Girard (CC-IN2P3), réunion des sites LCG-FR, automne 2011

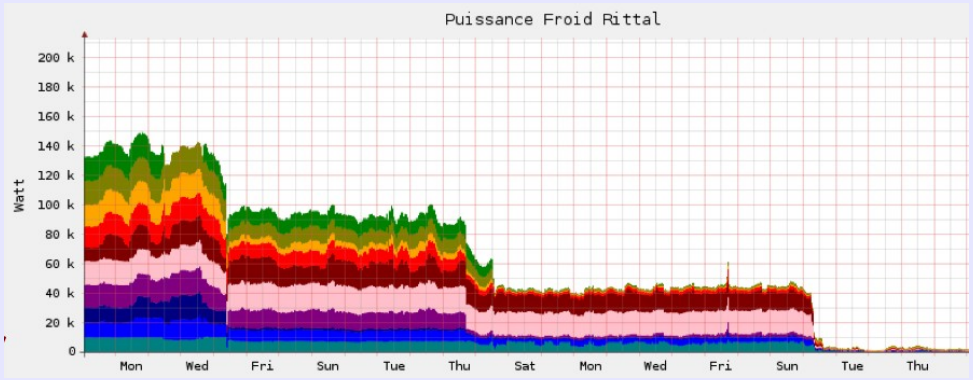
Organisation au CC-IN2P3



Récents changements organisationels

Nouvelle salle machine du CC-IN2P3

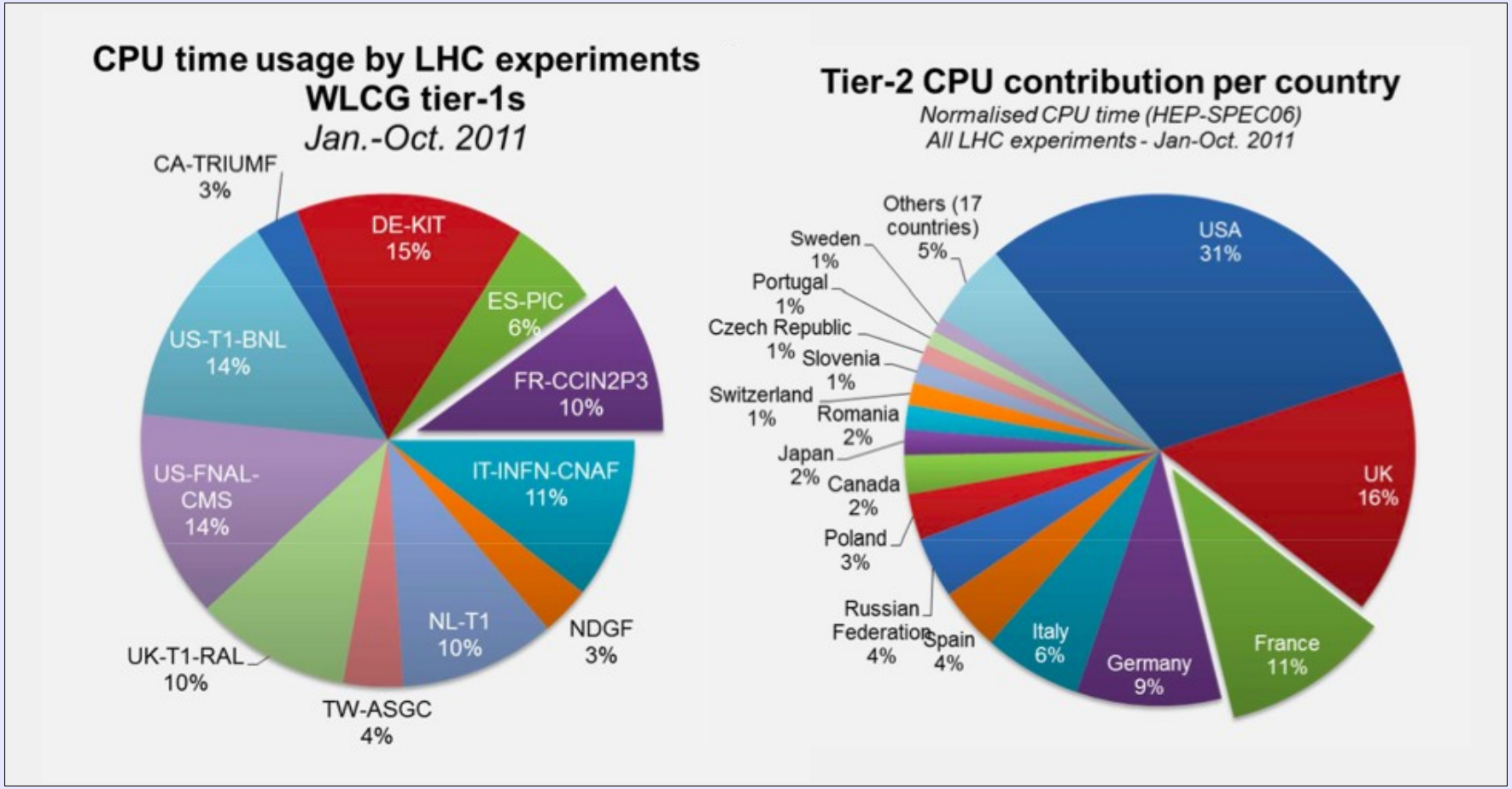




Décroissance de puissance consommée dans la première salle
lors de la migration des premiers racks dans la nouvelle salle en mai 2011

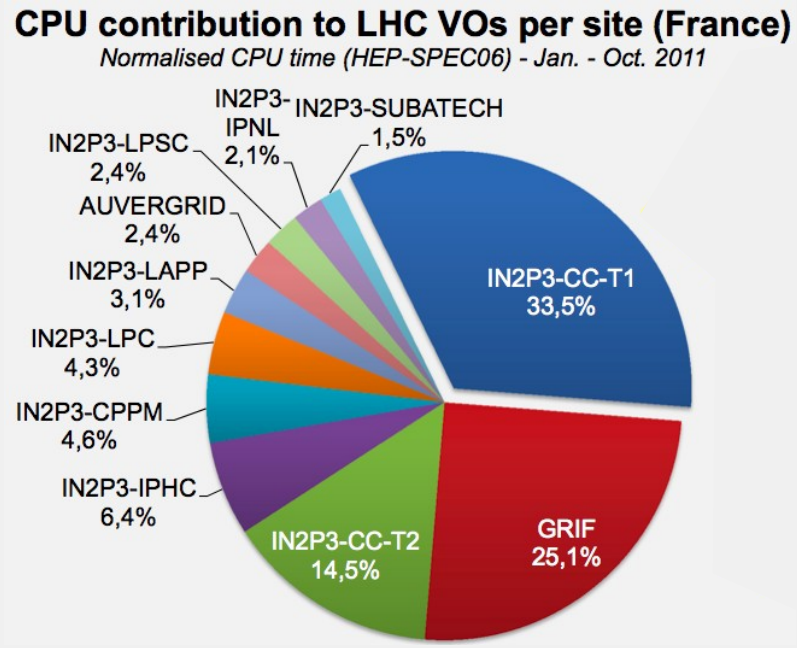
Ref. : X. Canehan (CC-IN2P3), “réunion des sites LCG-FR”, automne 2011

Contribution française à WLCG



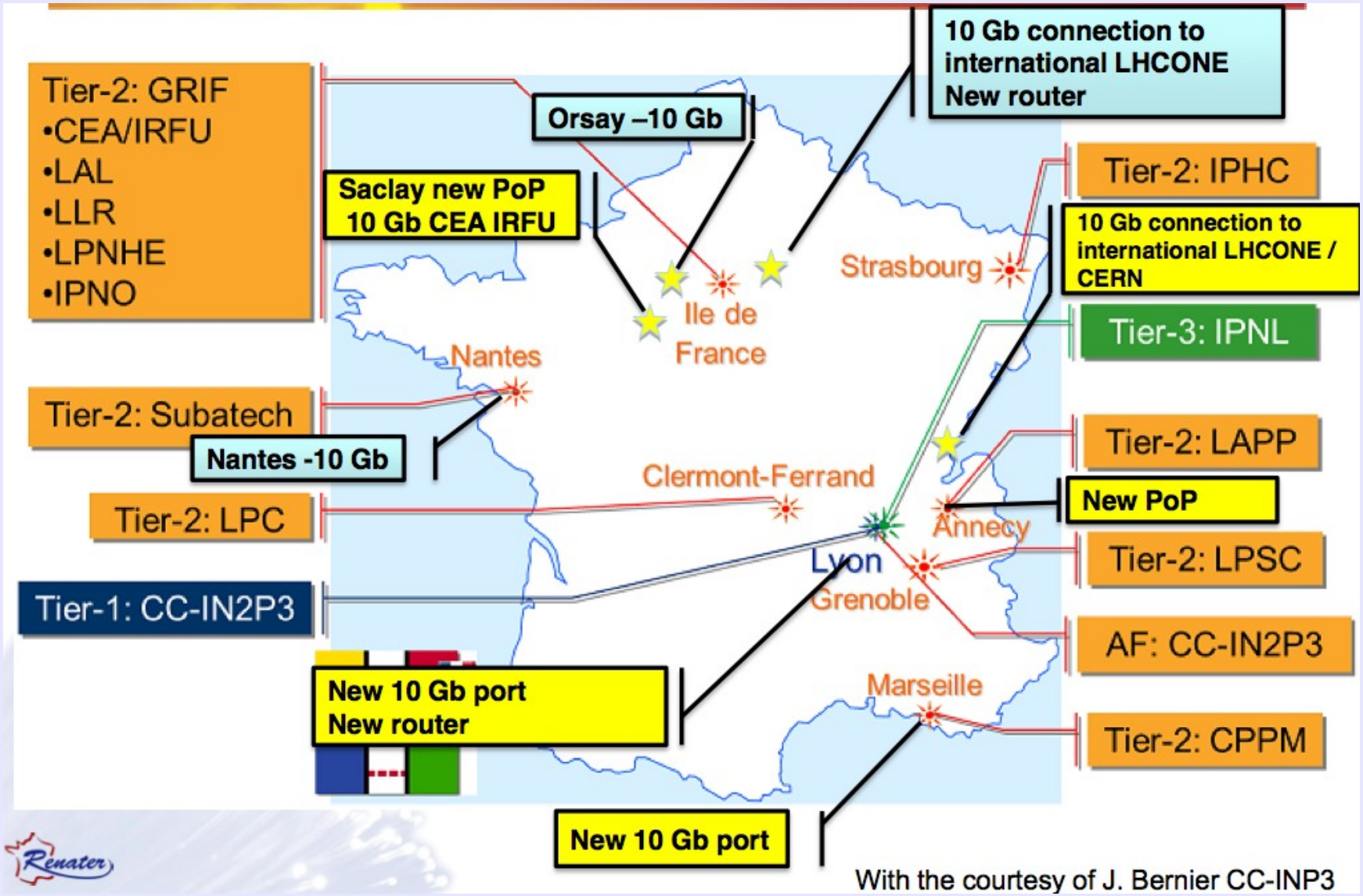
Source : F. Chollet (LAPP), réunion des sites LCG-FR, automne 2011

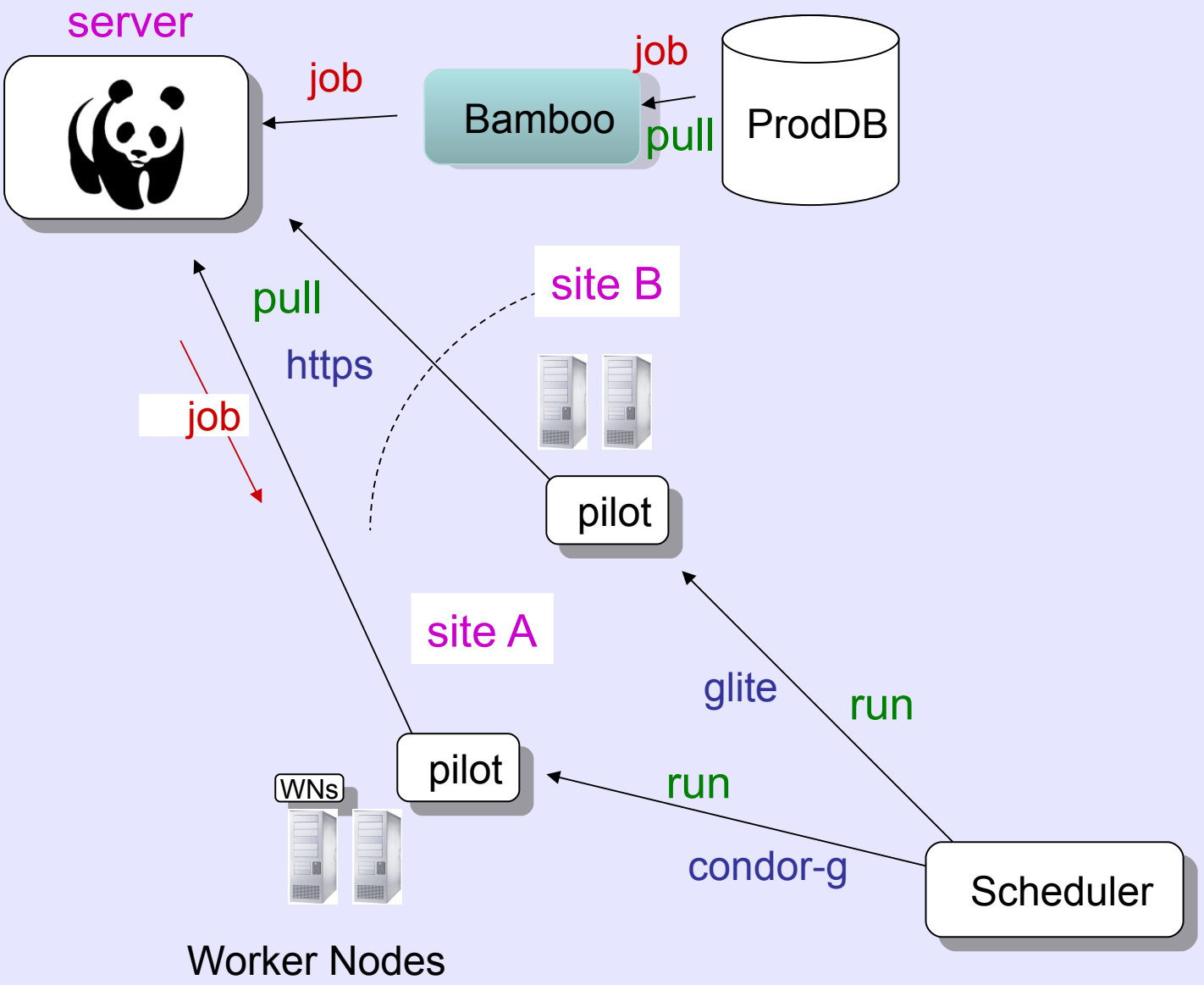
Contribution des sites Français



Source : F. Chollet (LAPP), réunion des sites LCG-FR, automne 2011

LHCONE



Source : X. Jeannin (CC-IN2P3), “réunion des sites LCG-FR”, automne 2011

ATLAS PANDA production scheme