LA GRILLE DE CALCUL DU LHC

LAURENT APHECETCHE

SUBATECH
Ecole des Mines de Nantes
4 rue Alfred Kastler
BP 20722
44307 Nantes Cedex 3

Résumé

Nous définissons d'abord quelques principes généraux d'une grille de calcul générique, puis nous nous concentrons ensuite sur une grille particulière, la grille de calcul du LHC (Large Hadron Collider), LCG (« LHC Computing Grid »), qui se doit de répondre au défi que représente le stockage et le traitement des données produites par le plus grand accélérateur au monde. Ceci n'est pas vraiment un cours, mais plutôt une introduction à des concepts (et des acronymes...) qui feront le quotidien informatique du physicien du LHC.

Abstract

We start by defining a few broad principles about a generic computing grid. We then turn our attention to a very specific grid, the LHC (Large Hadron Collider) Computing Grid, LCG, which must address the daunting issue of storing and processing the data generated by the world's largest accelerator. This document is not really a lecture, but a mere introduction to concepts (and acronyms…) that will certainly be part of LHC physicists every day's life.

1 Concept de grille : du rêve à la réalité

1.1 L'ordinateur mondial

On fait souvent l'analogie entre la grille de calcul (« computing grid) et le réseau électrique (« power grid »). Dans un cas comme dans l'autre, l'idée est que l'utilisateur final, vous en l'occurrence, dispose, dès qu'il se branche, d'une certaine puissance, sans avoir à se soucier d'où celle-ci provient. Puissance électrique dans le cas du réseau électrique, afin de faire fonctionner votre grille pain du matin ou votre téléviseur. Puissance de calcul (et de stockage, nous y reviendrons plus tard) dans le cas de la grille de calcul.

Lorsque vous branchez votre grille pain, vous ne vous voulez sans doute pas savoir d'où viennent les électrons qui circulent dans sa résistance. Qu'ils soient d'origine éolienne, hydraulique ou nucléaire, peu vous importe. Seule la puissance compte. De même, lorsque vous avez un calcul à effectuer sur la grille, ou que vous voulez y stocker des informations, peu vous importe où se trouvent le(s) processeur(s) et le(s) disque(s) qui feront le travail, pourvu que ce dernier soit fait.

Vu ainsi, une grille de calcul n'est rien d'autre qu'un immense ordinateur mondial, à disposition de tout un chacun. Cette vision reste pour l'instant du domaine du rêve. Les choses sont, en pratique, un tout petit plus compliquées que cela. Mais il ne faut pas perdre de vue que nous parlons ici d'informatique, et que dans ce domaine tout évolue très vite. Je vous laisse (à titre d'exercice, comme l'on dit souvent) imaginer à quoi un cours sur le web aurait pu ressembler il y a 10 ou 15 ans …

1.2 Une grille, des grilles

En réalité, il n'existe pas à l'heure actuelle une, mais des, grilles de calcul, dont les domaines d'applications et les envergures sont assez variés. Il existe notamment des grilles privées (restreinte d'accès au sein d'une entreprise, par exemple), des grilles de recherches nationales (par exemple Grid5000[1] en France), des grilles commerciales (Sun Grid propose par exemple ses services pour 1$ de l'heure-CPU) et des grilles dites de projet, comme EGEE[2] et LCG[3].

Ce cours va décrire LCG, dont une partie de l'infrastructure s'inscrit dans l'infrastructure « européenne » EGEE. Le projet EGEE regroupe des experts de plus de 50 pays avec comme objectif de capitaliser sur les avancées récentes en matière de technologies de grille et de développer une infrastructure de grille disponible pour les scientifiques, 24 heures sur 24. Le projet fournit aux chercheurs (du monde académique et industriel) l'accès à une grille de production, quelque soit leur localisation géographique. Le projet s'attache aussi à attirer un large panel de nouveaux utilisateurs vers la grille.

L'ampleur du projet EGEE peut se voir dans les quelques chiffres suivants :

- 259 sites (de 52 pays) connectés à l'infrastructure EGEE
- 72 000 processeurs accessibles, 24h/24
- 20 peta-octets de stockage
- 130 VO (organisations virtuelles) enregistrées
- 7500 utilisateurs enregistrés
- 150 000 jobs/jour

[1] http://www.grid5000.fr
[2] http://www.eu-egee.org/
[3] http://www.cern.ch/lcg

- 15 domaines d'applications

1.3 Quelles applications ?

Historiquement (ce qui est un bien grand mot, j'en conviens, pour un domaine qui n'a que quelques années d'existence), deux disciplines ont été et (restent) pilotes de la grille EGEE. Tout d'abord la physique des hautes énergies avec le LHC et ses dizaines de peta-octets par an de données à stocker et à analyser. Ensuite la biomédecine, avec notamment l'indexation des bases de données médicales dans les hôpitaux (plusieurs tera-octets par an par hôpital), ou encore l'exploration des bases de données génomiques.

Mais au-delà de ces deux disciplines témoins, bien d'autres peuvent bénéficier et bénéficie de fait de l'infrastructure EGEE, comme les sciences de la terre, la fusion, les sciences des matériaux, l'astronomie, l'astrophysique, etc…

1.4 Organisations

Au-delà des importantes ressources informatiques mises en jeu par les grilles de calcul, il s'agit de ne pas négliger l'impact du facteur humain. A la base, l'idée même de grille est une idée de partage de ressources entre individus qui partagent les mêmes intérêts et buts scientifiques. Ces individus sont regroupés en organisations virtuelles (« virtual organization », VO). L'accès à la grille ne se fait que si l'on fait partie d'une VO. Au sein de la physique des hautes énergies, chaque expérience du LHC représente une VO.

Bien sûr, d'un point de vue plus pragmatique, le fonctionnement au jour le jour d'une telle infrastructure serait impossible sans le travail de centaines d'informaticiens de par le monde.

Et la coordination du travail de toutes ces personnes nécessite elle aussi du personnel… Pour ne donner qu'un seul chiffre, l'ensemble du projet EGEE est estimé à 9000 personne-mois.

2 Fonctionnement de la grille

2.1 Les idées-forces

- Partage des ressources : plusieurs ordinateurs, situés partout dans le monde, mettent leur puissance de calcul et de stockage en commun. C'est le principe de base de la grille.

- Sécurité : les fournisseurs et les utilisateurs de la puissance de calcul, qui ignorent leur identité réciproque, doivent pouvoir échanger des données en toute confiance. La sécurité de la grille est notamment basée sur l'utilisation de certificats électroniques (dont je parlerai brièvement à la fin de ce cours).

- Equilibrage de la charge : le mécanisme d'affectation répartit les travaux de manière efficace et équilibrée entre les ressources disponibles.

- Abolition de la distance : le développement des connexions réseau à très haut débit permet d'échanger des données avec un ordinateur situé à l'autre bout du monde.

- Normes ouvertes : des applications faites pour être exécutées sur une grille doivent pouvoir l'être sur toutes les autres. Cela suppose que les normes qui régissent les différentes grilles soient compatibles entre elles.

2.2 Les briques de base

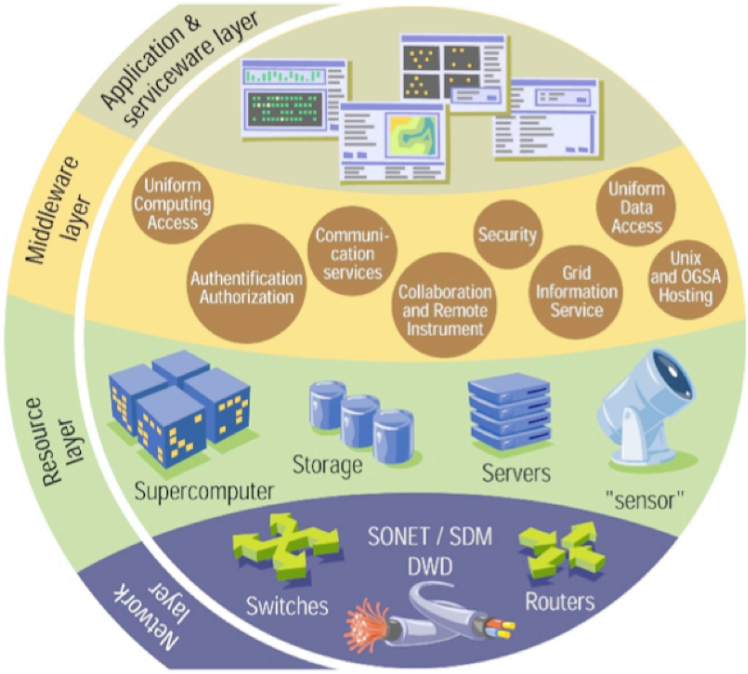


Figure 1 : l'architecture de la grille

La grille de calcul s'appuie sur une infrastructure matérielle et logicielle organisée en couches (figure 1). A la base se trouve l'infrastructure réseau, qui permet à l'information de circuler entre machines. Au-dessus diverses ressources offrent les capacités de calcul et de stockage. Un troisième niveau est celui de l'intergiciel (« middleware »), qui peut être vu comme le système d'exploitation de la grille (i.e. le Windows ou le Mac OS de la grille, ou plutôt le Linux de la grille, considérant qu'une très large majorité des machines de la grille tournent sous Linux). Enfin, au-dessus de tout cela se trouvent les applications des VO.

3 Le défi du calcul au LHC

Je n'ai pas l'intention de présenter ici le LHC, mais plutôt d'insister sur le défi qu'il représente en terme de calcul (il représente bien d'autres défis dans bien d'autres domaines également, mais ce n'est pas le sujet de ce cours…)

Le LHC, lorsqu'il entrera en fonction, génèrera 40 millions de collisions proton-proton par seconde. Après filtrage, les expériences enregistreront de l'ordre de 100 collisions par seconde. L'ordre de grandeur de la taille d'un événement est de 1 à 10 mega-octets. On voit donc que le taux d'enregistrement (par expérience) est de l'ordre de 0,1-1 giga-octets par seconde. Considérant de l'ordre de 10^{10} collisions par an, le volume total de *nouvelles* données produites par le LHC est de quelques dizaines de peta-octets par an. Si l'on devait graver ces données sur CD et que l'on empilait ces CD, on obtiendrait une pile haute de 20 kilomètres ! Voilà qui donne le ton pour la quantité de données au LHC. Reste à considérer le temps de calcul nécessaire au traitement de ces données… Mais avant cela, faisons un petit retour sur les unités utilisées lorsque l'on parle du calcul au LHC.

3.1 Les unités : kSI2K, TB, GB/s

Comme on l'a vu, l'ensemble des données du LHC se chiffre en peta-octets (« PB »). Au niveau des centres de calcul locaux (qui n'ont qu'une partie de ces données), l'unité pertinente est le tera-octets (« TB »). Les petits centres offrent quelques dizaines de tera-octets, alors que les plus gros centres quelques milliers…

Le transport de ces données se fait via le réseau, et dans ce domaine, l'on parle en gigabits par seconde (pas en octets, mais en bits, et un « octet-réseau » est composé de 10 bits). Le CERN est relié aux plus gros centres de LCG par un réseau dédié à 10 Gb/s. Les liaisons entre les autres centres sont typiquement de l'ordre de 1 Gb/s.

Enfin, le traitement de toutes ces données (voir ci-après) requiert du temps de calcul. Pour comparer des ordinateurs différents, un ensemble de programmes de test (« benchmark ») est exécuté sur les différents types de machine. Cet ensemble de programmes va attribuer une « note » à chaque type. Plus la note est élevée, plus la machine fera de calculs par seconde. Cette note est à l'heure actuelle en kSI2K (kilo specInt 2000). LCG est en train de migrer vers le SI2006 (nouvelle norme de spec.org pour mesurer les performances des machines). En bref, un processeur actuel (un cœur) vaut à peu près 1 à 2 kSI2K.

3.2 Du capteur à la publication

Je rappelle ici (très) brièvement les étapes qui permettent d'aller des données brutes jusqu'aux publications scientifiques.

Les données brutes (RAW) proviennent soit réellement du dispositif expérimental, soit de simulations. Simuler une expérience du LHC est un processus très gourmand en CPU. Il faut générer des particules (en utilisant des codes théoriques), puis transporter ces particules dans les détecteurs en déterminant leurs interactions avec la matière, et enfin simuler la réponse des détecteurs à ces interactions.

Les données brutes doivent ensuite être reconstruites, c'est-à-dire que les signaux électriques enregistrés par les détecteurs doivent être convertis en informations physiques exploitables : énergies, temps de vol, impulsions, identification des particules, etc.… Les données reconstruites sont stockées dans des fichiers ESD (« Event Summary Data »), dont la taille est une fraction de celle des RAW. Cette étape de reconstruction est à la fois gourmande en CPU (les algorithmes de trajectographie notamment) et en stockage (puisqu'il faut lire les données brutes)

Enfin, les ESD (et leurs cousines, les AOD, « Analysis Object Data », qui sont des ESD filtrées et compactées si possible) sont analysées afin d'en extraire les résultats physiques. D'une façon générale, les analyses sont gourmandes en entrées-sorties, plus qu'en CPU (des exceptions existent naturellement), et vont donc très fortement solliciter les espaces de stockage de la grille.

3.3 L'évolution des besoins

Chaque année, la collaboration WLCG demande aux quatre expériences du LHC quels sont leurs besoins (en CPU et en stockage) pour les années à venir. La figure 2 présente ainsi l'ensemble des besoins exprimés (mi 2008) pour la période 2008-2013. On notera encore une fois l'ampleur de la tâche que représente le calcul au LHC : en 2013, il faudra 700 peta-octets et 700 MSI2K (soit 700 000 cœurs actuels) pour accomplir les programmes de physique des expériences !

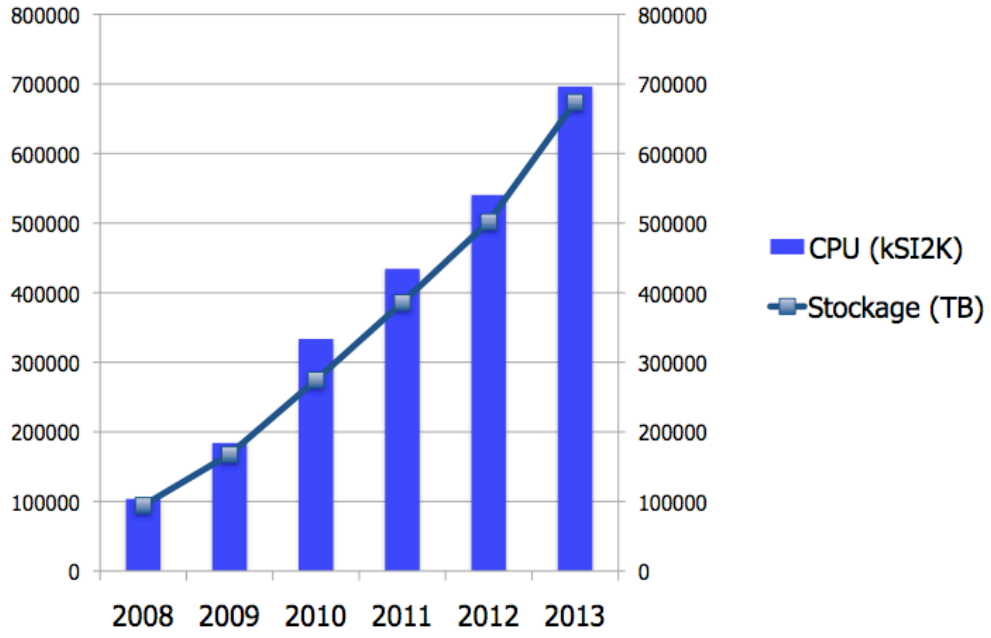


Figure 2 : les besoins de calcul du LHC

Il faut noter également que les expériences ont des besoins différents : les deux « grosses » expériences, ATLAS et CMS, représentent les trois quarts des besoins environ.

4 LCG : la grille du LHC

Les centres de ressources de la grille du LHC, qui est à ce jour la plus grosse grille de production au monde, sont (plus ou moins) hiérarchisés.

Au sommet se trouve le centre de calcul du CERN, le TIER0, qui reçoit les données brutes des expériences, et les enregistre sur un stockage de masse permanent (des bandes magnétiques). Le TIER0 effectue également une première reconstruction des données[4], et exporte les données (RAW et ESD de la première reconstruction) vers les TIER1.

Les TIER1 (une dizaine dans le monde) sont de (très) gros centres de calculs qui se partagent (sur stockage de masse) une copie des données brutes en provenance du TIER0. C'est-à-dire qu'à terme, toute donnée brute se trouve à deux endroits : au TIER0 et dans l'un des TIER1. Les TIER1 sont chargés de la (des) reconstruction(s) des données, pour produire des ESD (et AOD), et également de l'analyse de ces données. Les ESD (et AOD) produites sont exportées vers les TIER2. Les TIER1 peuvent aussi participer à la production des simulations.

Les TIER2 (une à deux centaines dans le monde) sont des centres (généralement) plus petits, qui ne possèdent (généralement) pas de stockage sur bande (mais uniquement sur disque dur). Leur rôle est de produire des données simulées (dites données Monte-Carlo) et d'analyser les données AOD produites par les TIER1. Il faut bien réaliser que leur rôle est capital puisque la moitié du CPU de LCG est fourni par les TIER2.

5 Fonctionnement de LCG

[4] Les modèles de calcul des quatre expériences sont parfois assez différents. Il faut donc considérer que ce qui est dit ici est « vrai en moyenne ». Les détails peuvent varier d'une expérience à l'autre.

La grille fonctionne par la mise à disposition de différents services, qui sont accessibles soit à l'utilisateur final (c'est-à-dire vous, le physicien assis devant son ordinateur portable), soit à des utilisateurs intermédiaires (le plus souvent d'autres services ou machines de grille). Le *détail* du fonctionnement de la grille est bien au-delà de ce cours. Je voudrais juste montrer quelques éléments fondamentaux, qui peuvent évoluer (entre autre le RB, voir ci-dessous) dans leur forme, mais leur raison d'être devrait perdurer.

5.1 Calcul et stockage

Ce sont les services de base de la grille, ceux auxquels on pense le plus souvent, qui permettent de calculer et de stocker.

L'élément de calcul (« Computing Element », CE) est un ensemble de serveurs de calcul (« Worker Nodes »), sur lesquels les jobs sont exécutés.

L'élément de stockage (« Storage Element », SE) est une machine de grille permettant de stocker des fichiers. Le SE est accessible depuis tous les serveurs de calcul et depuis toutes les machines interactives (UI, voir ci-après). Cependant, les fichiers stockés sur cette machine ne sont accessibles que par l'utilisation de commandes spécifiques.

5.2 Informations

Les SE et les CE ne peuvent pas vivre seuls au milieu de nulle part, sans être au courant de rien. Piloter un job sur la grille revient à prendre des décisions, et pour prendre des décisions judicieuses il convient d'être correctement informé.

Le service journal de bord (« Log and Bookeeping », LB) garde l'historique de la gestion des jobs (i.e. lorsqu'ils sont soumis, en exécution, terminés, en erreur, etc…).

Le catalogue des répliques (« Replica Catalog », RC) est le service des pages jaunes de la grille : il permet de localiser les données, qui peuvent par ailleurs être présentes à plusieurs endroits.

Enfin, le répertoire d'information (« Information Service », IS) est l'Agence France Presse de la grille : il permet d'obtenir en temps réel l'état des services de grille de tous les sites.

5.3 Choix

Pour soumettre un job sur la grille, il faut avoir accès à une interface utilisateur (« User Interface », UI). Il s'agit généralement de l'ordinateur du physicien, et plus généralement, de tout ordinateur possédant le logiciel nécessaire pour accéder à la grille (obtention des autorisations, soumission de jobs, etc…).

Enfin, l'intermédiaire ultime dans ce jeu de relation est le courtier en ressources (« Resource Broker », RB), qui sélectionne le CE le mieux adapté à un job donné.

5.4 Push vs pull

Sans trop rentrer dans les détails, je voudrais signaler deux points importants. D'abord, le fonctionnement décrit ici est celui de l'intergiciel LCG. A l'heure actuelle, LCG utilise une version plus récente, dénommé gLite. Ensuite, ce même gLite évolue avec le temps, et les dernières versions ont introduit un double mode de fonctionnement pour l'attribution des jobs, ce que l'on appelle le « pull » en opposition au « push ». Dans le mode « push » (qui est celui utilisé par l'ancien RB LCG), c'est le RB qui décide de prendre un job dans sa liste de job en attente et de l'envoyer vers un CE disponible (il le « pousse » vers le CE). Le RB prend cette décision en fonction des informations disponibles sur ce CE à un instant donné. Si pour une raison quelconque l'état du CE change entre le moment de la décision et le moment où le job démarre sur ce CE, ou si l'information disponible n'est pas à jour, le job peut échouer. Depuis le tout début

de LCG, conscientes du taux d'échec relativement élevé que le mode push peut entraîner, certaines VO ont « contourné » ce problème en lançant non pas de vrais jobs, mais des jobs pilotes, dont la seule tâche est de se lancer sur le CE, de vérifier que l'environnement de la machine correspond aux attentes du vrai job, et ensuite laissent la place au vrai job, qui est récupéré (tiré) depuis la liste des jobs en attente (c'est le mode « pull »). Les dernières versions de gLite rendent ce mode de fonctionnement « officiel » et parfaitement supporté.

5.5 La vie d'un job sur la grille

Pour terminer ce cours, voyons comment partir de rien et lancer un job sur la grille…

La toute première chose à faire est de rejoindre une VO (organisation virtuelle)[5]

Vous devez ensuite (ou en parallèle), demander un certificat électronique à une autorité de certification (pour la France, l'UREC, une unité du CNRS, joue ce rôle). Ce certificat est une véritable carte d'identité électronique. Il vous permet de prouver (à un programme ou service) que vous êtes vous. Il ne vous donne aucun droit en soit. Autrement dit, il vous authentifie, mais ne vous autorise pas. Ce certificat doit être installé sur l'interface utilisateur (UI, machine contenant les logiciels nécessaires à l'accès à la grille). Il est valable un an. Il doit être correctement protégé (afin que personne ne puisse se faire passer pour vous) et renouvelé en temps utiles.

Compte-tenu que de nombreux programmes ou services de grille vont devoir, à un moment ou à un autre, vérifier votre identité, votre certificat devrait beaucoup circuler et *pourrait* être compromis. Afin de limiter cette possibilité, vous vous connectez à la grille non pas avec votre certificat, mais avec une autorisation temporaire (un proxy), émise à partir de votre certificat (une sorte de carte d'identité temporaire), et valable quelques heures (au lieu d'un an pour votre certificat). Typiquement, cette autorisation s'obtient en utilisant une commande du type « grid-proxy-init ».

Muni de votre autorisation temporaire d'accès, vous devez ensuite décrire le job que vous voulez effectuer sur la grille. Vous utilisez pour cela un langage très simple, le JDL (« Job Description Langage »), qui décrit notamment le programme à lancer, les données à utiliser, où écrire les résultats, etc… dans un simple fichier texte.

Vous soumettez enfin votre JDL à la grille, par une commande du type « job-submit job.jdl »[6].

Quelque soit la méthode utilisée par votre VO (push ou pull), un service de grille, quelque part, faisant usage des services d'informations, va faire qu'une décision va être prise sur la destination de votre job (en fonction de critères géographiques, de priorités internes à la VO, de disponibilité des données, etc…), et il va être envoyé sur ce CE choisi.

Après la soumission de votre job, vous pouvez à tout moment interroger la grille pour savoir où en est votre job (« waiting », « executing », « crashing », « saving », « done »). Lorsque

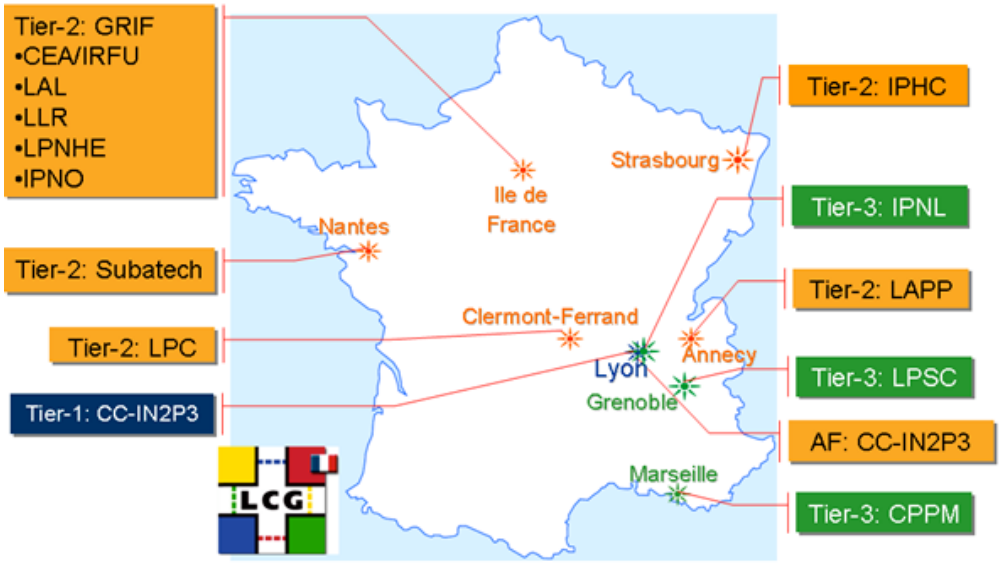
[5] Si vous faites partie d'une expérience du LHC, vous faites déjà partie d'une VO (mais vous devrez sans doute vous enregistrer comme utilisateur de LCG).

[6] Certaines organisations virtuelles, et notamment celles du LHC, utilisent une couche logicielle par dessus la couche gLite (l'intergiciel de grille), afin de mieux adapter leur logiciel spécifique (auquel leurs utilisateurs sont habitués) à l'intergiciel. La commande effective de soumission peut donc varier d'une VO à l'autre (de même que les autres commandes décrites ici), mais l'intention ou la fonction reste la même.

votre job est terminé, vous récupérez la sortie soit directement dans un SE soit par des commandes adéquates (en fonction de votre VO, principalement).

6 Conclusion

En guise de conclusion, je voudrais juste montrer la carte des sites français qui contribuent à LCG, et vous livre quelques liens qui m'ont servi à préparer cette présentation et qui pourraient vous permettre d'aller plus loin. Car la seule vraie façon de connaître la grille, c'est de l'utiliser !



http://gridcafe.org

http://cern.ch/lcg

http://lcg.in2p3.fr