Présentation de la Grille EGEE

- Introduction aux grilles
- La grille EGEE
- Exemples d'applications en physique des particules et en sciences de la vie
- Le cercle vertueux
- Conclusion

Guy Wormser

Directeur de l'Institut des Grilles du CNRS

Une grille ... Pourquoi ?

- Mettre à la disposition des utilisateurs une puissance de calcul et des capacités de stockage importantes
- Garantir l'efficacité et la sécurité de ces calculs et stockages
- Fonctionner sur un mode décentralisé

Principales caractéristiques d'une Grille de calcul

Une grille est constituée d'un ensemble d'ordinateurs et d'outils logiciels destinés à les faire fonctionner de manière cohérente

Les ressources sont en principe hétérogènes

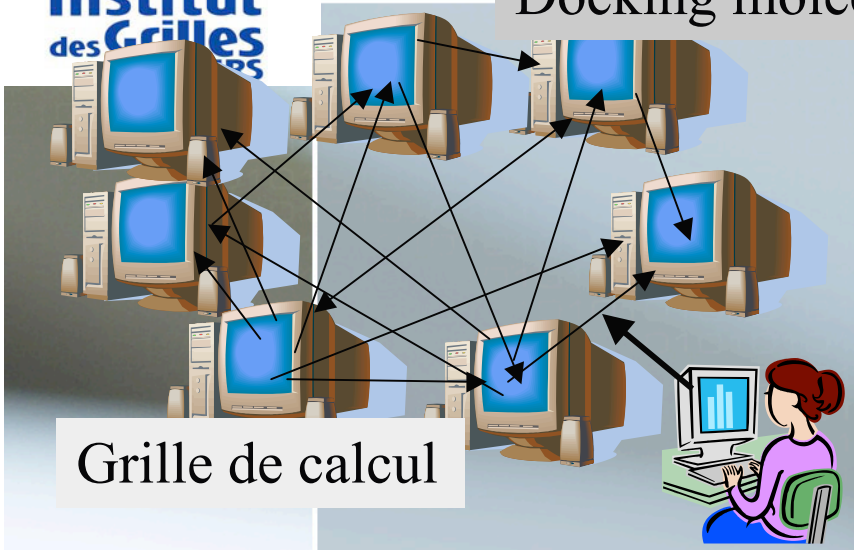
Un système d'information (même très simple) doit être présent pour allouer les ressources informatiques adaptées aux tâches à exécuter

Chaque nœud de la grille est administré localement. Mais une coordination centralisée est indispensable pour garder le système cohérent

Le réseau sur lequel s'appuie la grille est crucial

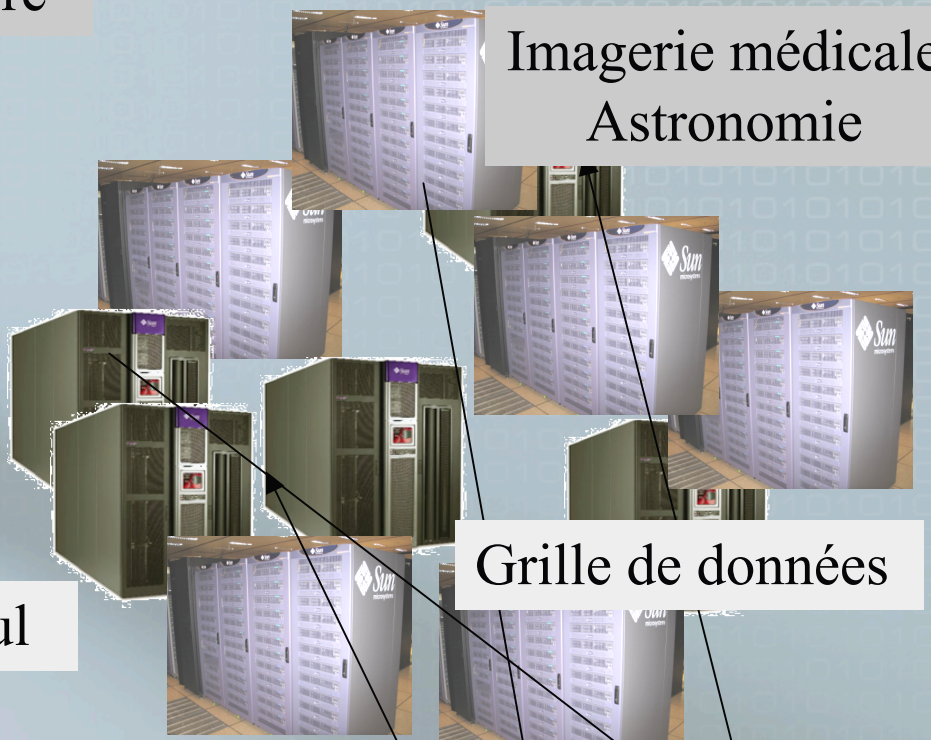
Un système d'authentification et de sécurité doit être présent

Différents types de grilles



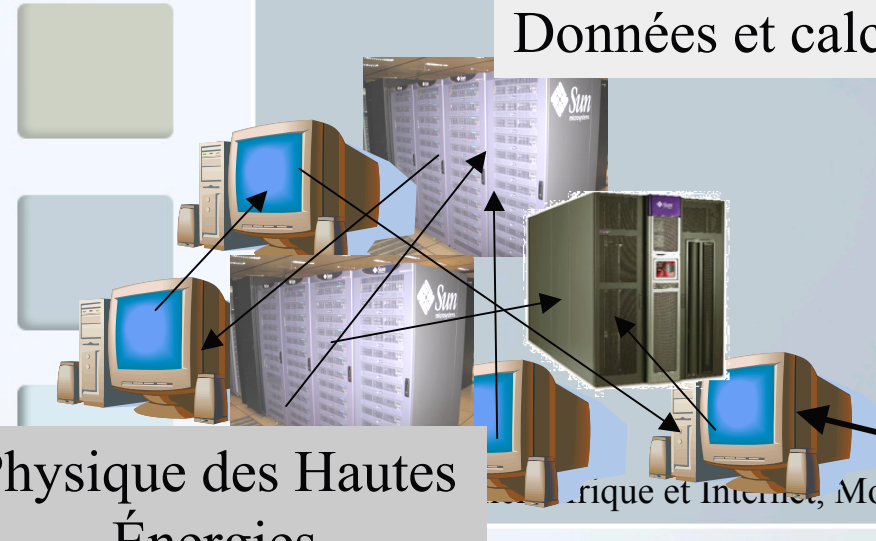
Grille de calcul

Docking moléculaire



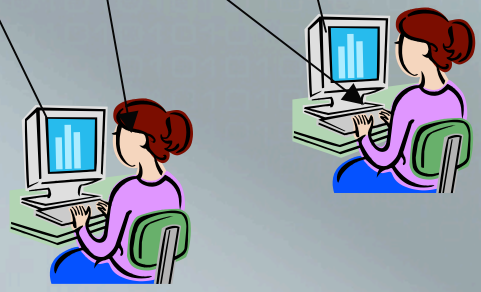
Imagerie médicale
Astronomie

Grille de données

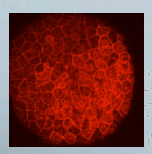


Données et calcul

Physique des Hautes Énergies



Les grilles de données

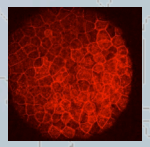





Production de données par de nombreuses équipes distribuées géographiquement et indépendantes

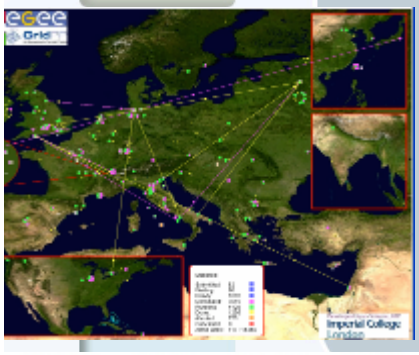
L'ensemble des données doit être accessibles par tous



EGEE – What do we deliver?

- Infrastructure operation
  - **Currently includes ~250 sites across 45 countries**
    - **Continuous monitoring of grid services & automated site configuration/management**
    - **Support many Virtual Organisations from diverse research disciplines**
- Middleware
  - **Production quality middleware distributed under business friendly open source licence**
    - Implements a service-oriented architecture that virtualises resources
    - Adheres to recommendations on web service inter-operability and evolving towards emerging standards



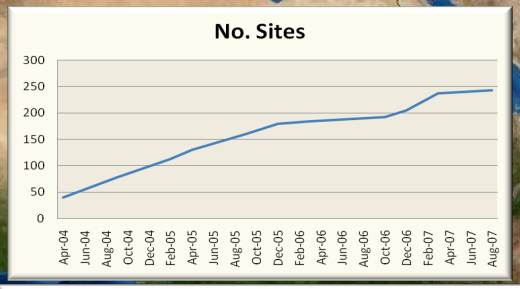
User Support - *Managed process from first contact through to production usage*

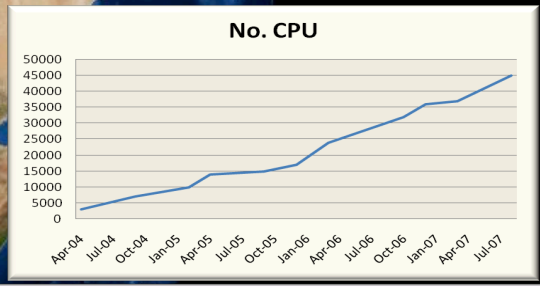
- **Training**
- **Expertise in grid-enabling applications**
- **Online helpdesk**
- **Networking events (User Forum, Conferences etc.)**

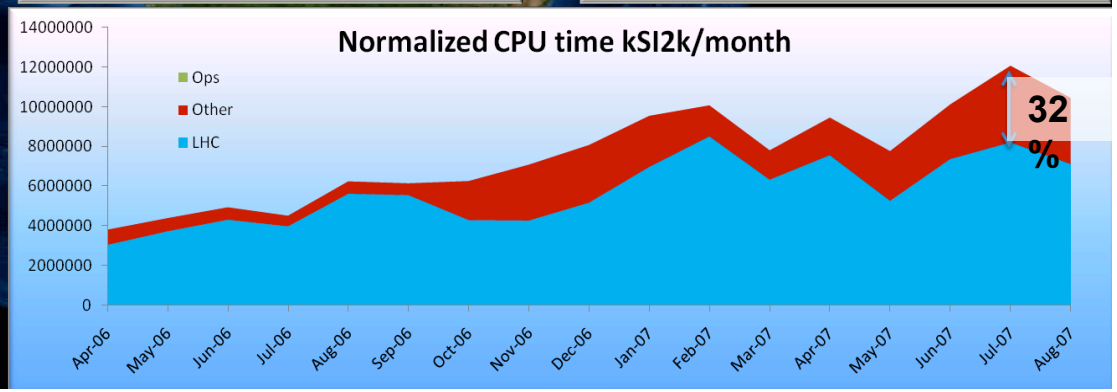
Enabling Grids for E-sciencE

240 sites
45 countries

41,000 CPUs
5 PetaBytes
>5000 users
>100 VOs
>100,000 jobs/day





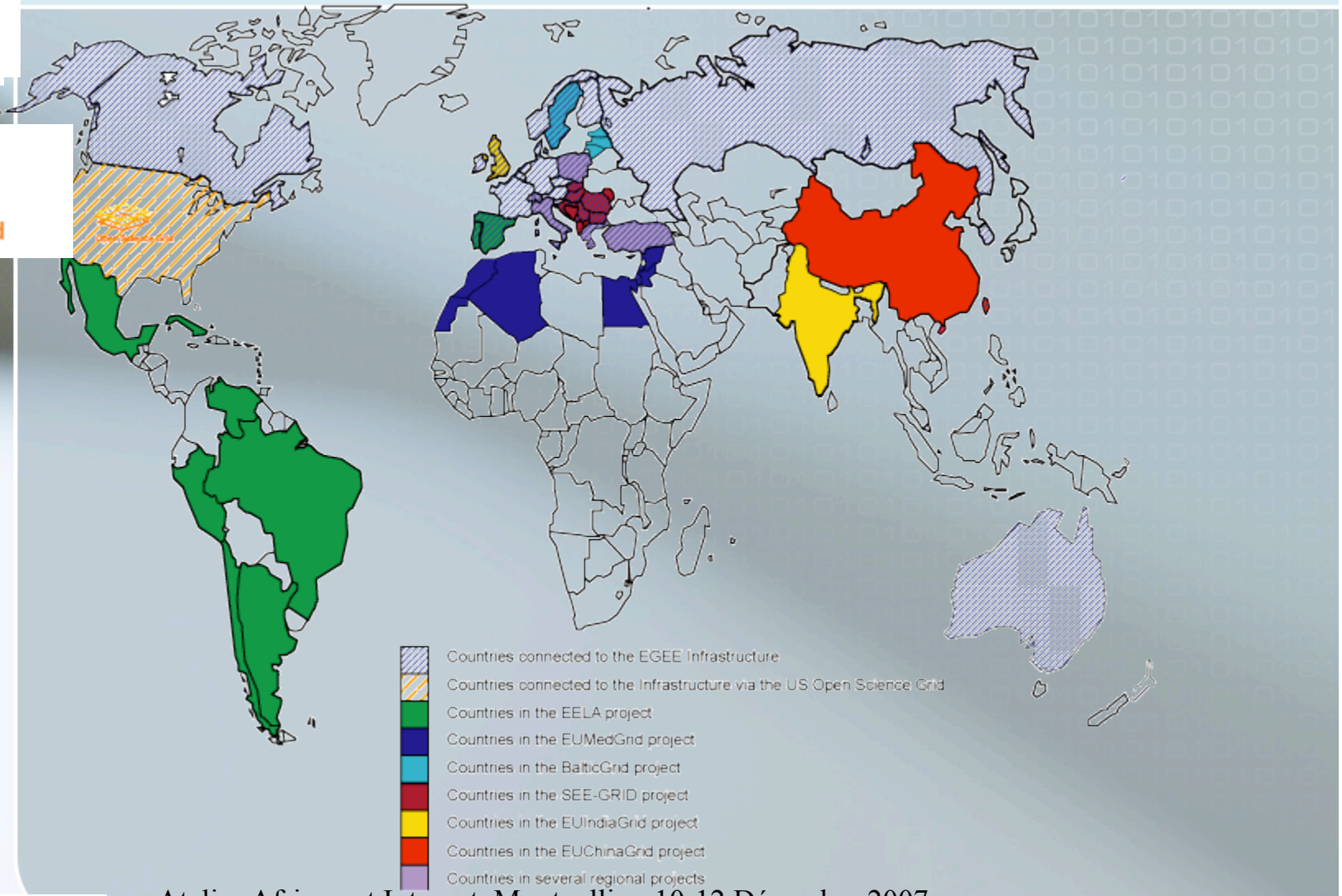


Archeology
Astronomy
Astrophysics
Civil Protection
Comp. Chemistry
Earth Sciences
Finance
Fusion
Geophysics
High Energy Physics
Life Sciences
Multimedia
Material Sciences
…

Types of applications

- Simulation
  - LHC Monte Carlo simulations; Fusion; WISDOM
  - Jobs needing significant processing power; Large number of independent jobs; limited input data; significant output data
- Bulk Processing
  - HEP ; Processing of satellite data
  - Distributed input data; Large amount of input and output data; Job management (WMS); Metadata services; complex data structures
- Parallel Jobs
  - Climate models, computational chemistry
  - Large number of independent but communicating jobs; Need for simultaneous access to large number of CPUs; MPI libraries
- Short-response delays
  - Prototyping new applications; grid Monitoring grid; Interactivity
  - Limited input & output data; processing needs but fast response and quality of service
- Workflow
  - Medical imaging; flood analysis
  - Complex analysis algorithms; complex dependencies between jobs
- Commercial Applications
  - Non-open source software; Geocluster (seismic platform); FlexX (molecular docking); Matlab, Mathematics; Idl, …
  - License server associated to an application deployment model

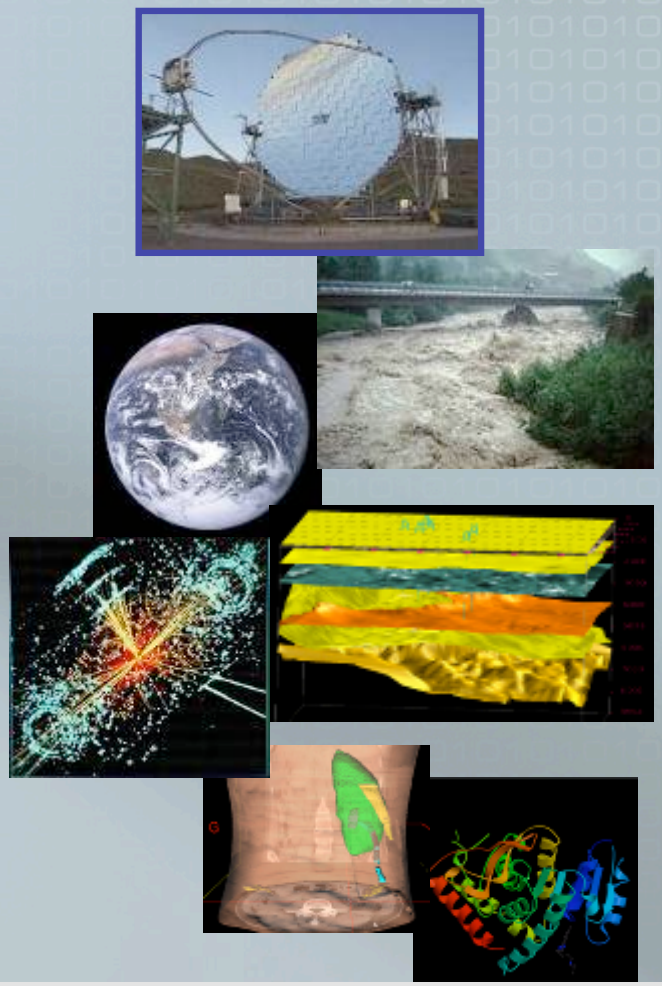
Collaborating infrastructures



EGEE: Une architecture de grille pluridisciplinaire

De très nombreux domaines d'applications

- Archéologie
- Astronomie & Astrophysique
- Protection civile
- Chimie
- Sciences de la Planète
- Simulation Financière
- Fusion
- Géophysique
- Physique des Hautes Energies
- Science de la vie
- MultiMedia
- Science des matériaux
- ...

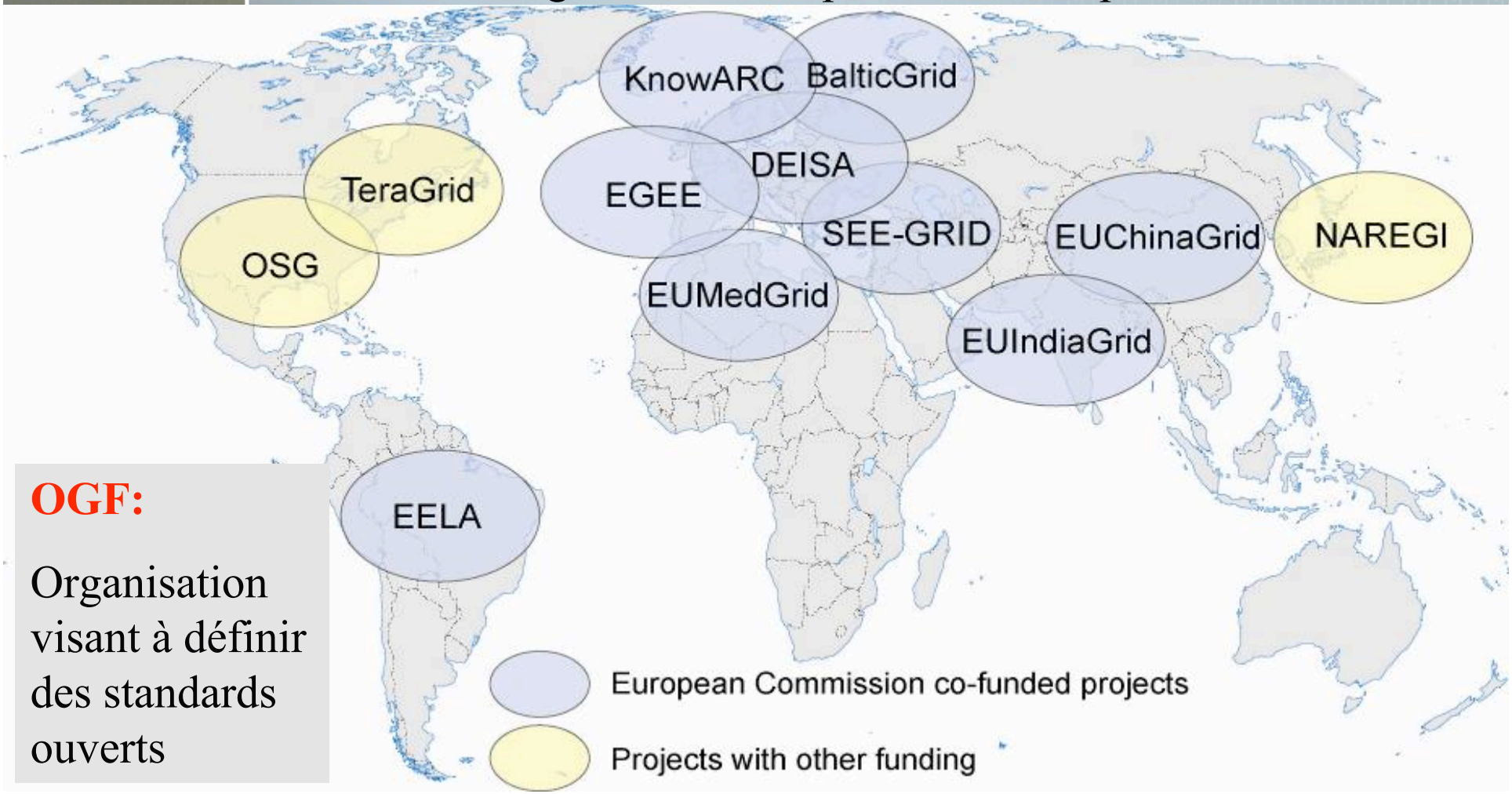


Des milliers d'utilisateurs à travers le monde

Interopérabilité

L'idée d'une grille mondiale unique est un mythe

➔ A terme les grilles doivent pouvoir inter-opérer

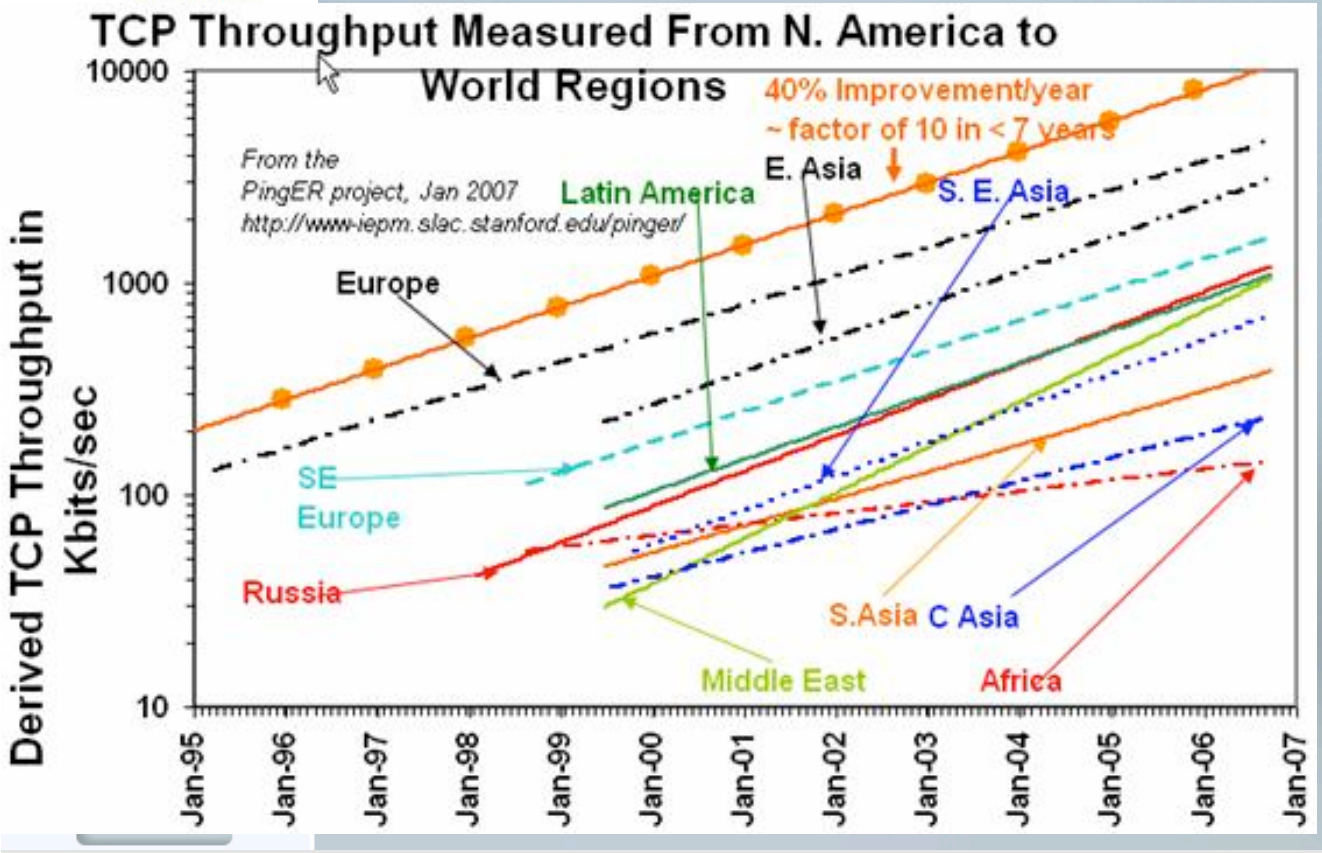


OGF:

Organisation visant à définir des standards ouverts

Les grilles et la fracture numérique

R. L. Cottrell and S. Khan http://www.slac.stanford.edu/xorg/icfa/icfa-net-paper-jan07/



Derrière l'Europe:
6 ans: Russie, Amérique Latine
7 ans: Moyen Orient, Asie du SE
8-9 ans: Asie du SO
11 ans: Asie Centrale
12 ans: Afrique

Les grilles et en particuliers les projets partenaires d'EGEE: EELA, EUChinaGrid, EUMedGrid, EUIndiaGrid contribuent à la réduction de la fracture numérique – Idem pour le modèle de calcul du LHC

Exemples de problèmes à résoudre (1)

Gestion dynamique des ressources:

- Les machines ne sont pas disponibles en permanence, certaines sont en panne,
- Les performances du réseau d'interconnexion fluctuent avec la charge et le nombre d'utilisateurs...

Sécurité et données confidentielles:

- Il faut garantir aux utilisateurs que personne ne pourra interférer dans leur calcul ou accéder à leurs données;
- Cryptographie

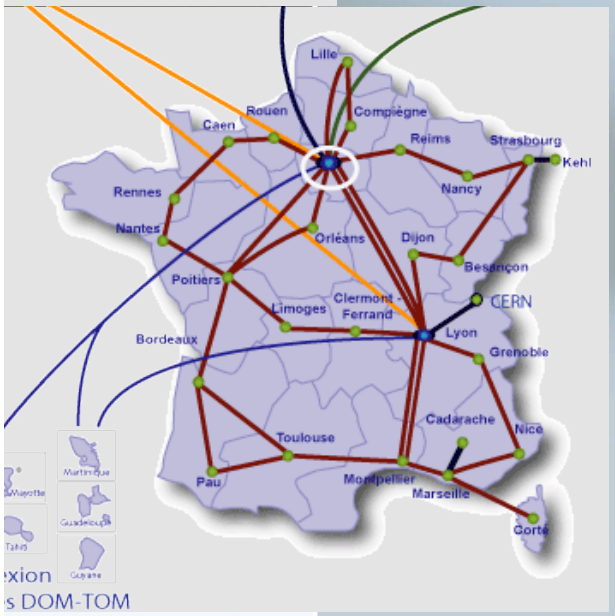
Exemples de problèmes à résoudre (2)

Optimisation des performances:

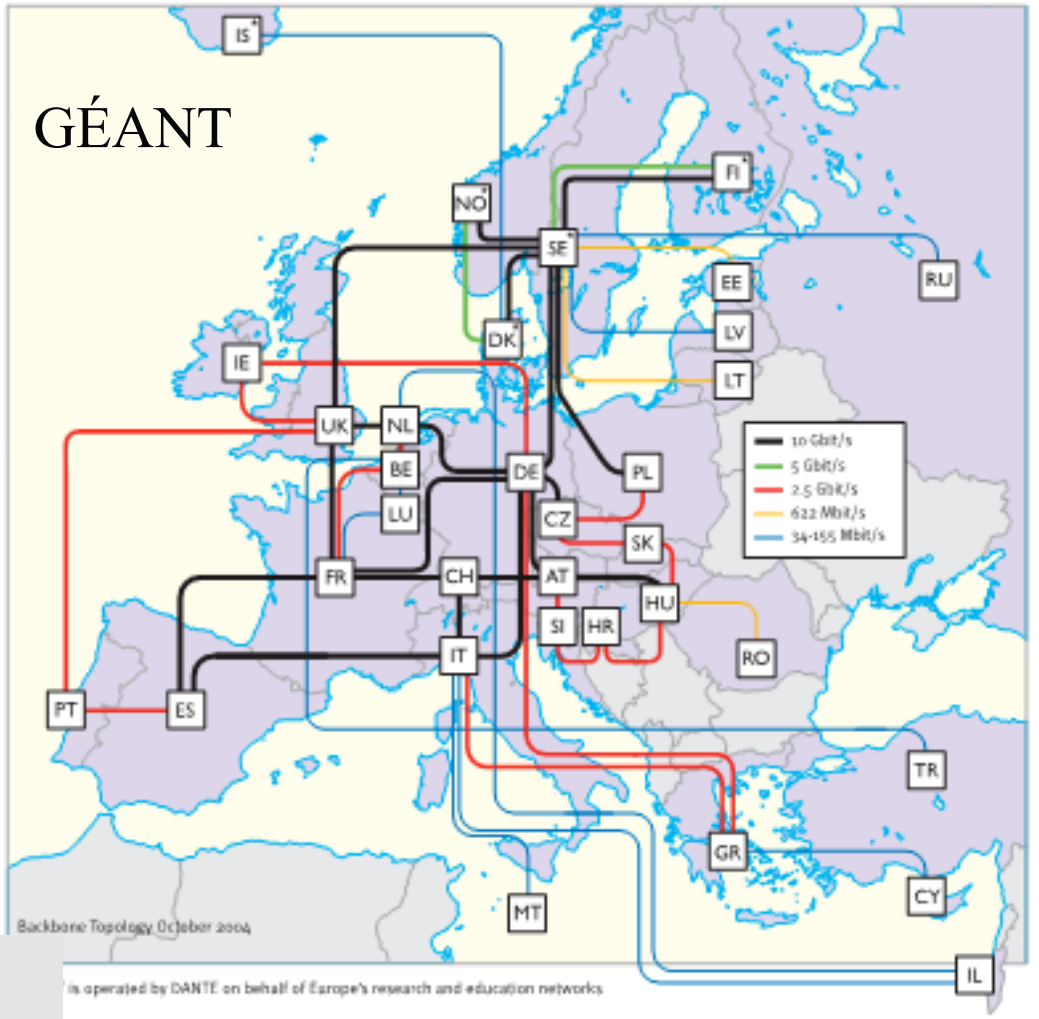
- Des matériels divers sont mis en oeuvre à tous les niveaux (liens de communications, mémoires, disques, processeurs); cacher cette hétérogénéité.
- Faire bon usage de ces matériels pour obtenir leur meilleure performance simultanément est une tâche très complexe.

Importance du réseau

RENATER

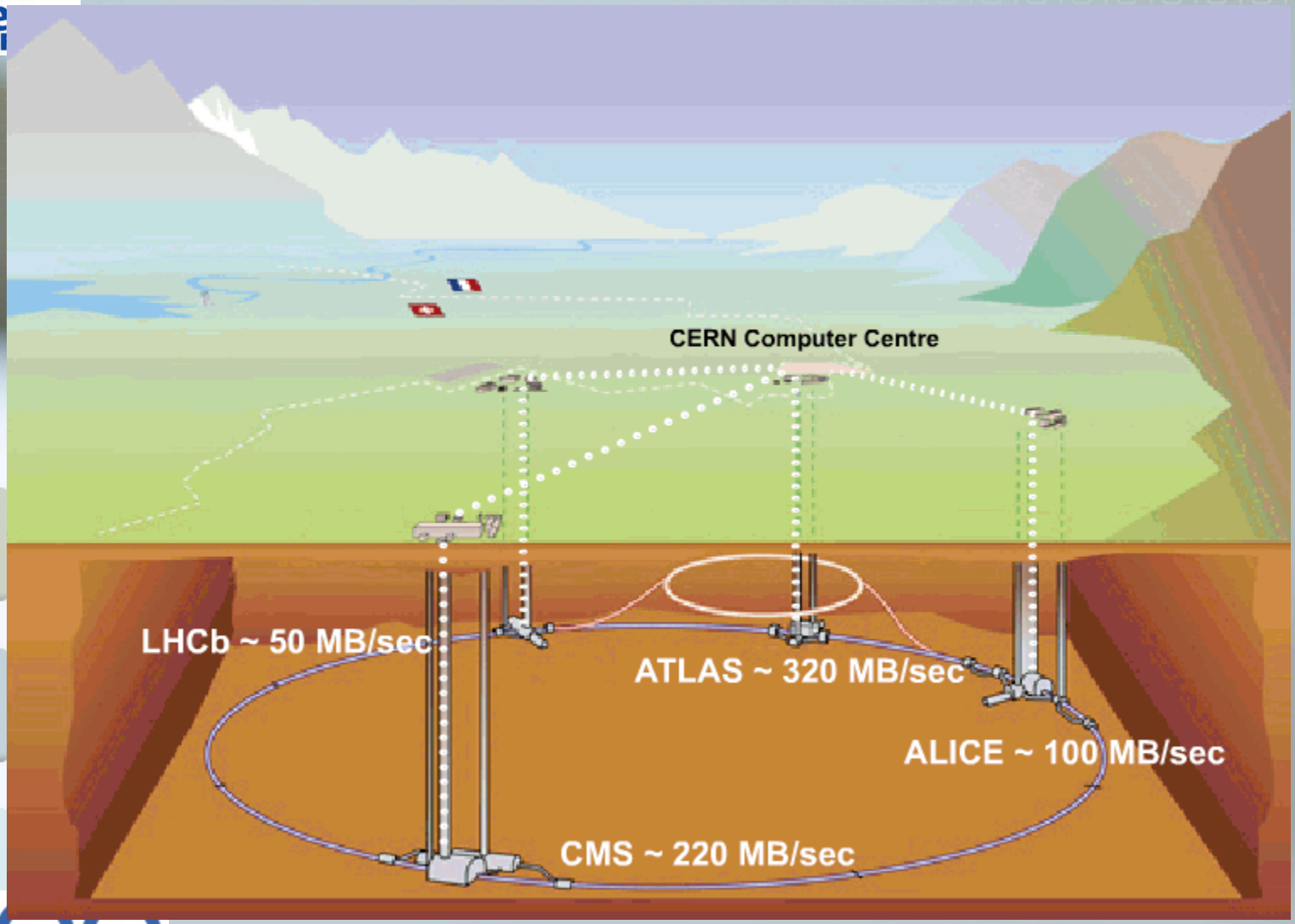


GÉANT

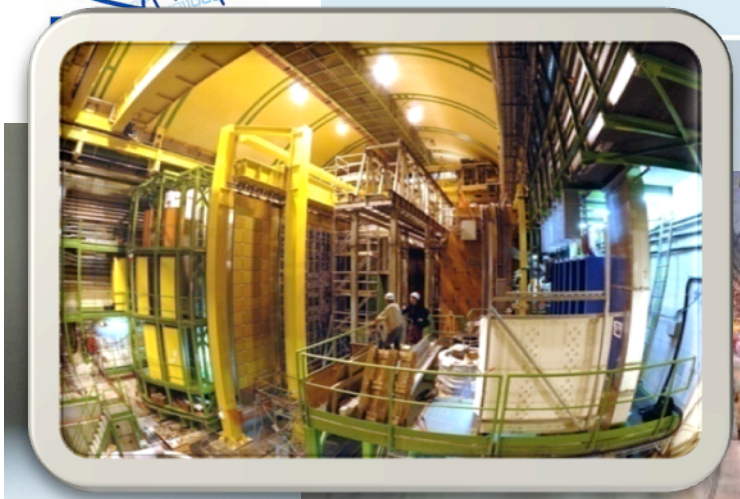


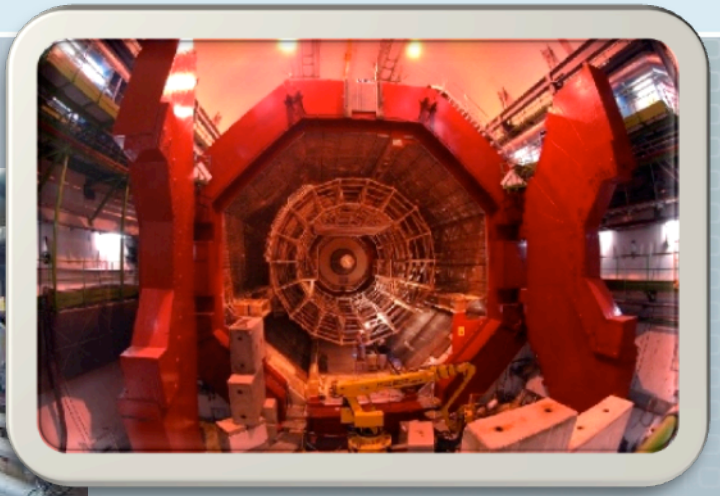
La qualité du réseau est primordiale pour le fonctionnement des grilles de calcul

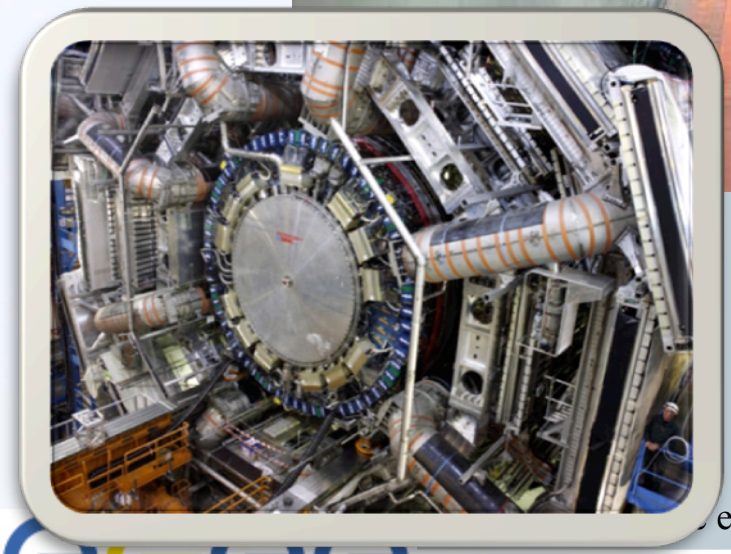
L'accélérateur LHC du CERN

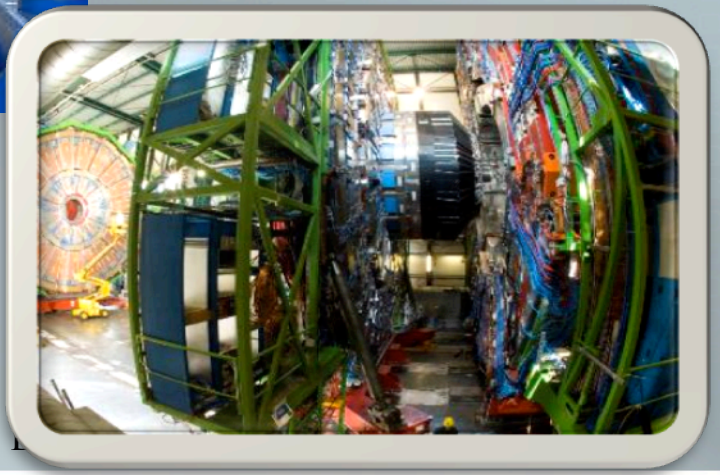


Les 4 expériences du LHC

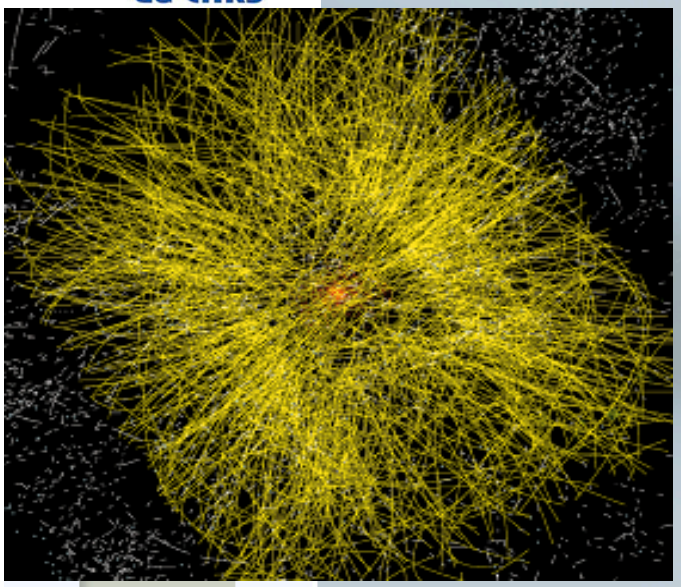








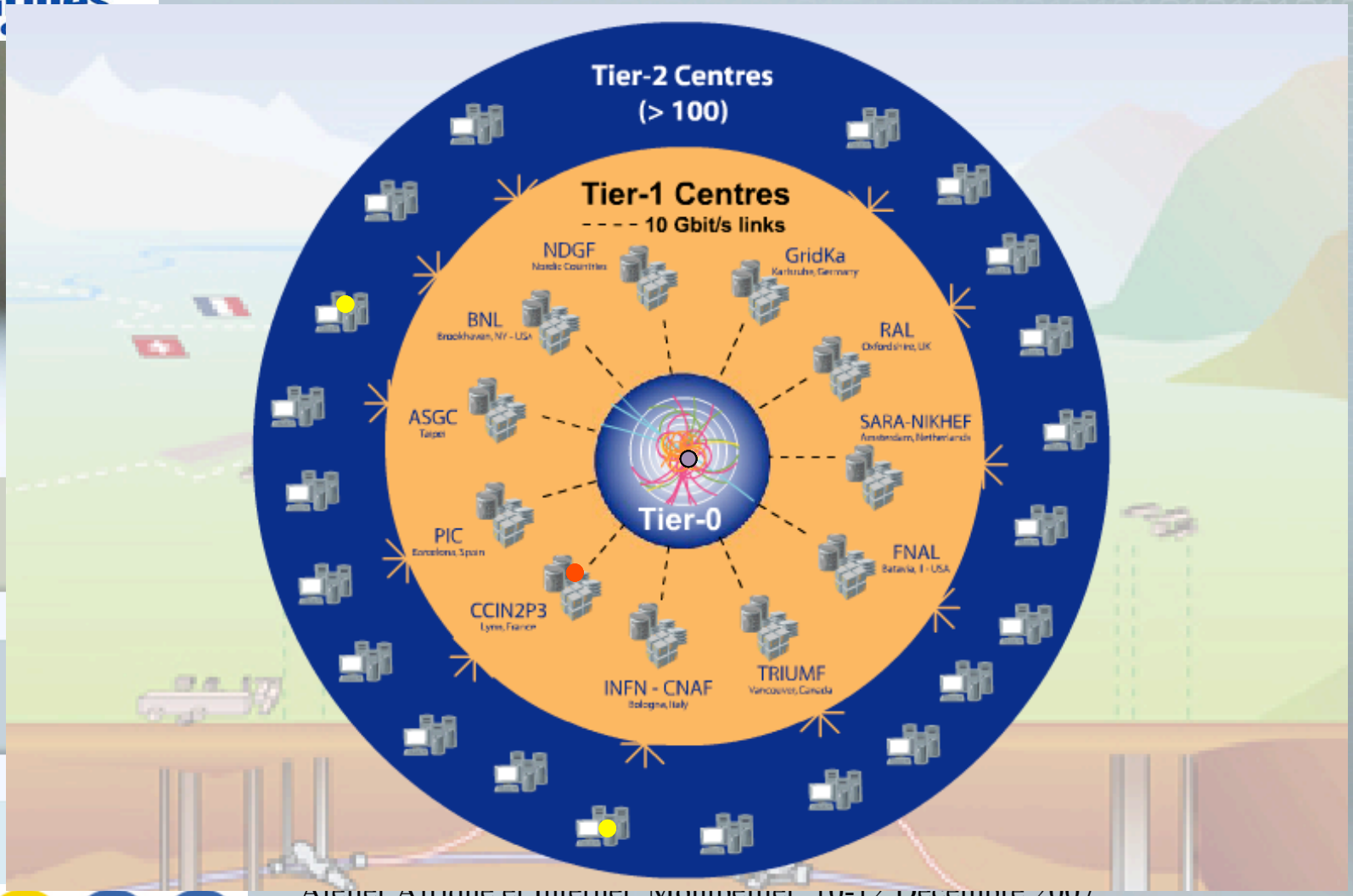
Le calcul au LHC: Un énorme défi!



- Signal/Bruit 10^{-9}
- Volume de données
  - Taux élevé * grand nbre de canaux * 4 expériences
  - ➔ **15 PetaBytes de données nouvelles par an**
- Puissance de calcul (100 MSI2k)
  - Complexité de l'évenement * Nb. événements * milliers d'utilisateurs
  - ➔ **50 k fastest CPUs (d'aujourd'hui)**

Le modèle par étages

Tier-0 -1 -2




Enabling Grids for E-sciencE
Scheduled = 21539
Running = 25374
La grille est une réalité opérationnelle pour le LHC
La Mise à l'échelle est un succès
Elle est Fonctionnelle
Stable
Fiable
Indispensable
21:13:50 UTC
GridPP
UK Computing for Particle Physics

Enjeux et intérêt des grilles en sciences du vivant

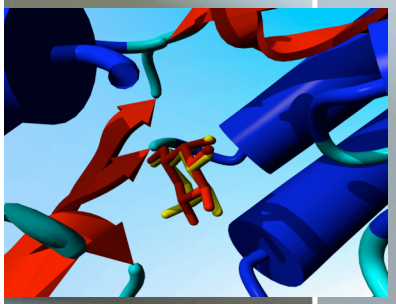
- Les enjeux
  - L'avalanche des données a bouleversé les stratégies de recherche en biologie moléculaire
  - La médecine doit évoluer vers une science exacte exploitant toutes les données de la génomique à l'épidémiologie
- L'apport des grilles
  - La grille fournit **aujourd'hui** les siècles de cycles CPU requis pour les calculs massifs
  - La grille fournit **aujourd'hui** les services de gestion sécurisée pour stocker et copier les données biologiques et médicales
  - La grille offrira **à terme** l'environnement collaboratif pour l'intégration et le partage des données dans les communautés de recherche

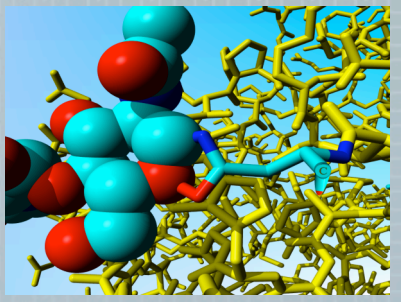
Comment utilise-t-on les grilles aujourd'hui en sciences du vivant ?

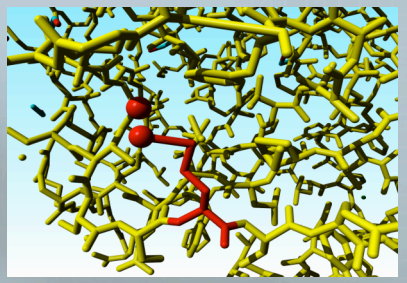
- Pour déployer des calculs à très grande échelle
  - Exemple en bioinformatique: raffinement des structures de la PDB
  - Exemple en imagerie médicale: simulateur de RMN
  - Exemple pour la recherche de nouveaux médicaments: WISDOM
- Pour l'analyse interactive de données de plus en plus volumineuses
  - Exemple en bioinformatique: portail GPS@
  - Exemple en Imagerie médicale: GPTM3D
- Pour mutualiser de nouveaux services et compétences

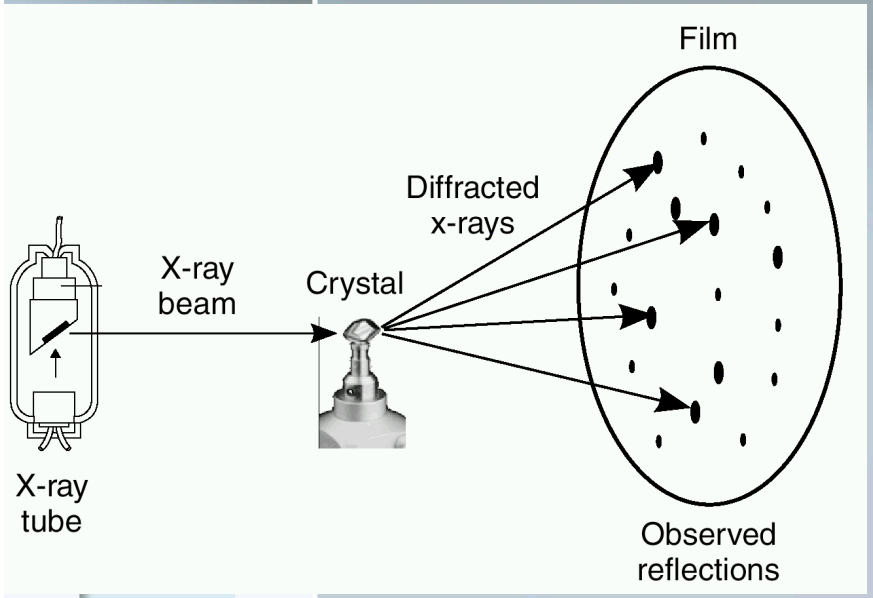
Bioinformatics: recalculating protein 3D structures in PDB

- The PDB data base gathers publicly available 3D protein structures
  - Full of bugs
- Project: redo the structures by recalculating the diffraction patterns





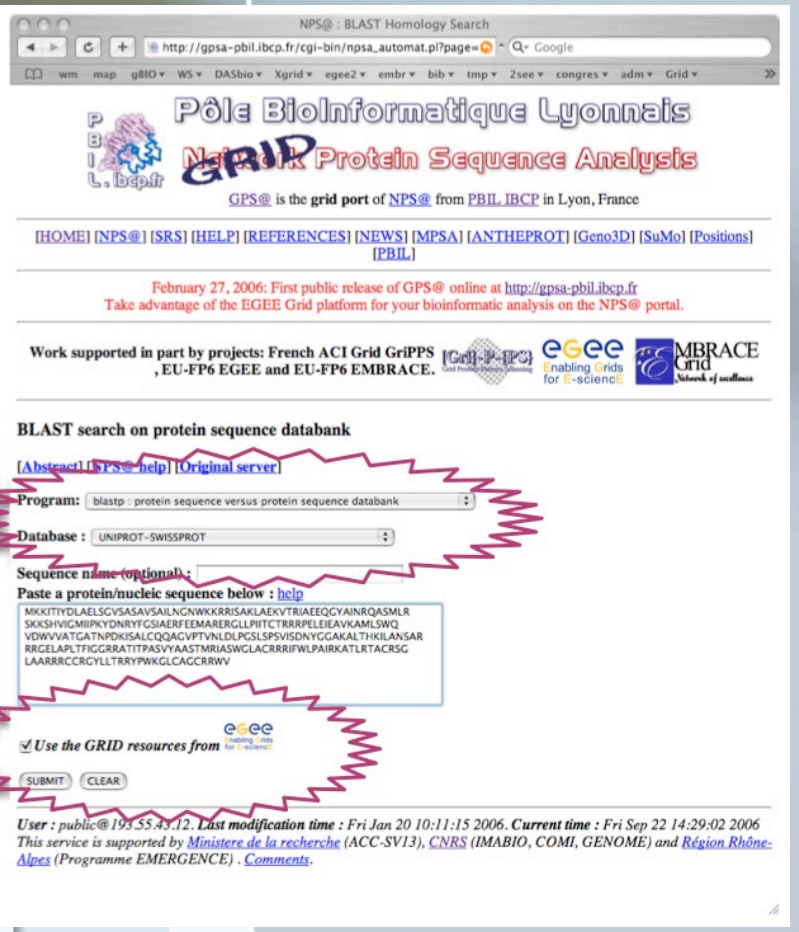




PDB-files	**42.752**
X-ray structures	**36.124**
Successfully recalculated	**~36.000**
Improved R-free	**12.500/17000**
CPU time estimate	**21.7 CPU years**
Real time estimate	**1 month on Embrace Virtual Organization on EGEE**

Credit: G. Vriend, CMBI

GPSA: Bioinformatics Grid Portal



- Scientific objectives
  - Molecular Bioinformatics of proteins
  - Analyze data from high-throughput Biology: complete genome projects, EST, complete proteomes, structural biology, ….
  - Integration of biological data and tools
- Method
  - Provide Biologists with an usual Web interface for Bioinformatics: NPS@
    - NPS@ Web portal online since 1998
    - 46 tools & 12 updated databases
    - + **10,000,000** jobs & 5,000 jobs/day
  - Ease the access to updated databases and algorithms.
    - Protein databases are stored on the grid storage as flat files, encrypted if needed.
    - Wrapping legacy bioinformatics applications
    - Transparent remote access through local file-system accesses
  - Display results in graphical Web interface.
- Status: **Prototype**

Example: radiology analysis

- Scientific objectives

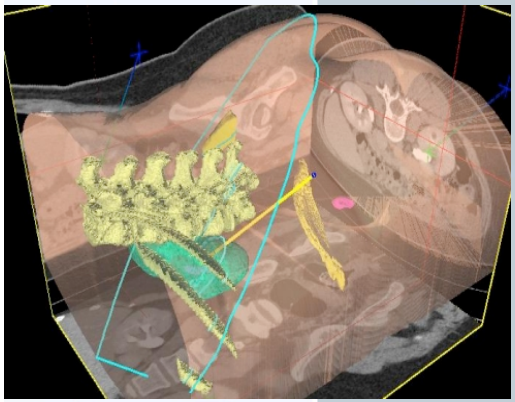
  Interactive volume reconstruction on large radiological data.

  PTM3D is an interactive tool for performing computer-assisted 3D segmentation and volume reconstruction and measurement (RSNA 2004)

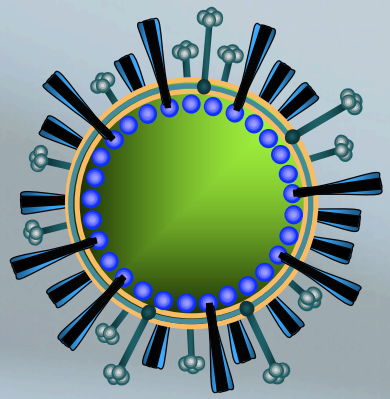
  Reconstruction of complex organs (e.g. lung) or entire body from modern CT-scans is involved in augmented reality use case e.g. therapy planning.

- Method

  Starting from an hand-made rough Initialization,a snake-based algorithm segments each slice of a medical volume. 3D reconstruction is achieved in parallel by triangulating contours from consecutive slices.

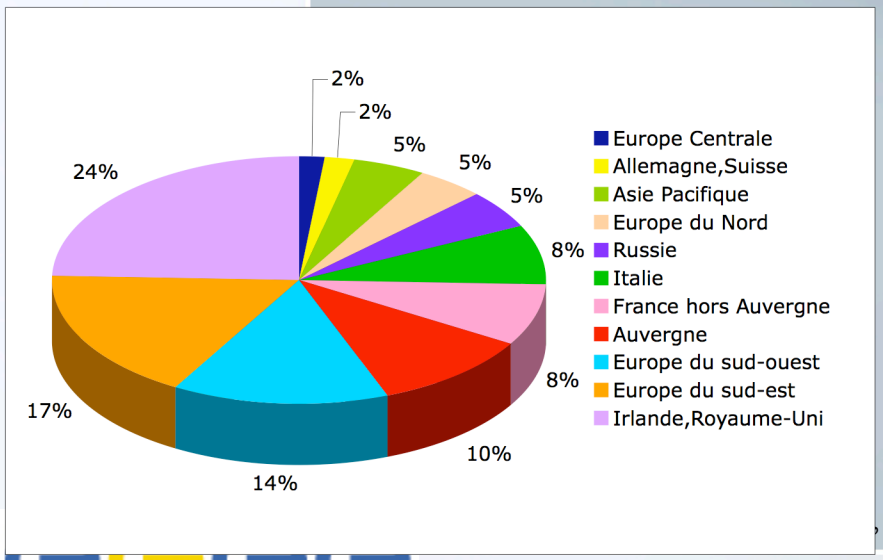


Recherche de nouveaux médicaments contre la grippe aviaire



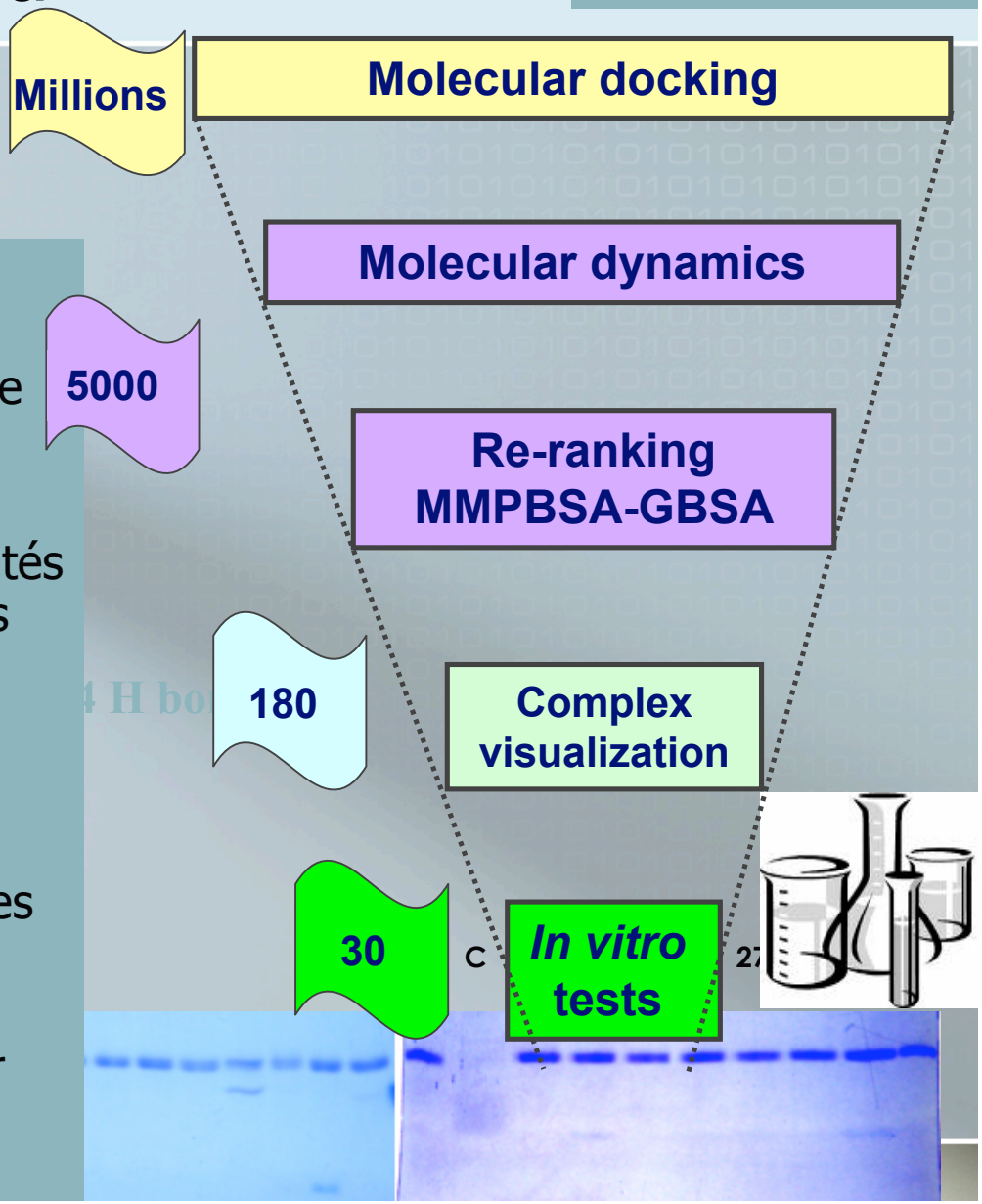
Credit: Y-T Wu – D. Kim

- Objectifs:
  - étudier l'impact de mutations de la neuraminidase N1 sur l'efficacité des médicaments actuels (Tamiflu)
  - Identifier de nouvelles molécules actives
- Méthode:
  - Calculs sur ordinateur des probabilités d'accrochage des molécules sur la neuraminidase mutée
- Résultats expérimentaux
  - 20% des 300 molécules sélectionnées in vitro et testées in vivo sont plus actives que le tamiflu
    - Facteur 200 d'amélioration des résuttats des tests in vitro



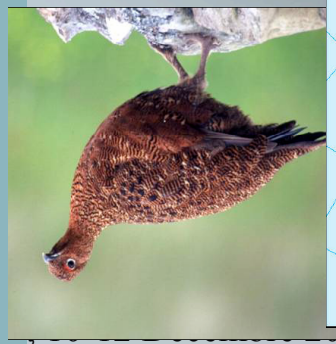
Recherche de nouveaux médicaments contre la malaria

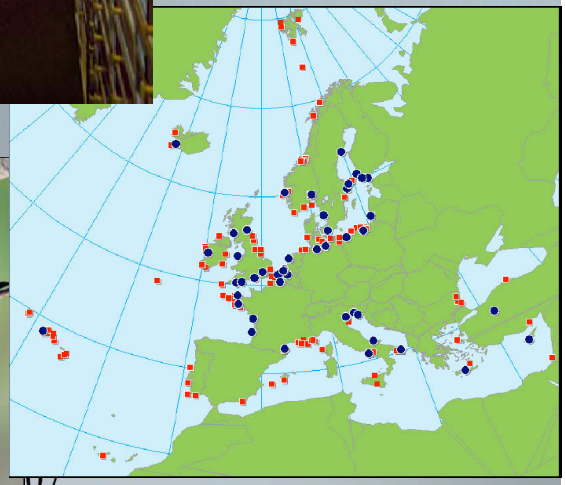
- Objectifs:
  - Identifier de nouvelles molécules actives sur des cibles biologiques de la malaria
- Méthode:
  - Calculs sur ordinateur des probabilités d'accrochage des molécules sur ces cibles
  - Raffinement des calculs par dynamique moléculaire
- Résultats expérimentaux
  - 20% des 30 molécules sélectionnées in vitro et testées in vivo sont des inhibiteurs actifs
  - Tests in vivo en cours à Montpellier



Biodiversity infrastructure: the LifeWatch project

- Life Watch: e-Science and Technology Infrastructure for bioversity data and observatories
  - Several thousand sites collecting data of ecological interest
  - The challenges
    - Distributed data generation
    - Common mechanisms for sharing, analyzing and synthesizing these data
- Building a international infrastructure is easier on a grid foundation.
  - Use case under study (EGEE – HealthGrid)





L'intérêt scientifique des grilles

- Accès transparent à des données distribuées
  - Exemples Sciences de la terre, sciences de la vie
- Manipulation de très grands volumes de données
  - Physique des particules, astrophysique, sciences humaines
- Très grande flexibilité des ressources de calcul
  - Gestion des catastrophes
  - Challenge grippe aviaire , malaria

Le cercle vertueux

- Collaborations scientifiques internationales, réseaux rapides et grilles de calcul sont les trois ressorts d'un progrès scientifique rapide
- Ce tryptique permet aux scientifiques des différents pays de s'intégrer au mieux dans les grandes collaborations internationales
- L'idée est de progresser en parallèle sur ces trois axes qui se renforcent mutuellement

L' Institut des Grilles du CNRS

- L'activité au CNRS sur les grilles a atteint en 2007 une importance considérable en volume et en impact
- Fédérer l'ensemble des activités du CNRS dans le domaine des grilles de recherche et des grilles de production
  - Meilleure visibilité
  - Meilleure efficacité
  - Renforcer l'interaction entre ces deux domaines
- Point de contact pour les partenariats nationaux et internationaux
  - Représentant du CNRS pour les contrats européens, pour les discussions auprès du Ministère
  - Noyau central pour la « National Grid Initiative »
  - Partenariat à nouer avec l'INRIA et les autres organismes pour la recherche sur les grilles de calcul
- « Evangélisation » auprès de nouvelles communautés scientifiques
- Actions d'animation, de formation et de dissémination
- Partenaire fort d'une future structure de type « Institut Français des Grilles »

Conclusion

- L'Europe s'est dotée d'une infrastructure de grille de production la plus importante au monde
- De nombreux résultats scientifiques uniques y sont obtenus
- Ce modèle peut bénéficier aux universités africaines si un réseau rapide peut y être installé
- De très nombreuses collaborations internationales pourront se nouer autour des grilles