

## Part - III

### Le défi des systèmes distribués



1

## Le défi des systèmes distribués...

- Le développement de l'Internet ces dernières années a permis l'émergence de nouvelles utilisations des systèmes informatiques.
- Internet permet de faire collaborer des machines distribuées dans l'ensemble des bureaux d'une entreprise, dans l'ensemble des bâtiments d'un campus, dans l'ensemble des centres de calcul d'un pays, voire même d'un continent ou de la planète.
- « *A distributed system is a collection of independant computers that appear to the users of the system as a single computer* »  
A. Tanenbaum, Prentice Hall, 1994.
- Plusieurs approches :
  - Le « Métacomputing »,
  - L'Internet Computing & Desktop Grid,
  - Les techniques « P2P : peer to peer »...
  - Le « Grid Computing »
  - ...



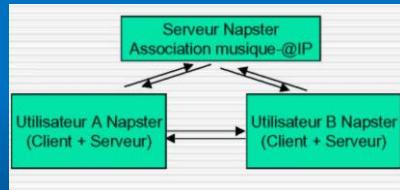
Tout le monde n'a pas un supercalculateur !



3

## Modèles de déploiement

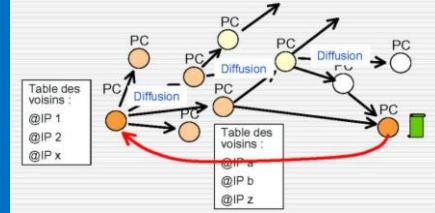
- Client-serveur
  - Centralisé ou distribué
  - Utilisation de cache pour éviter la congestion
  - Information centralisé
- Pair à Pair (P2P)
  - Chaque pair est à la fois client et serveur
  - Distribution de la charge dans le réseau
  - Information distribuée
- Le cas Napster (Hybride)
  - Accès aux données via un site unique contenant un index
  - Stockage et partage des données « inaltérables », copies multiples
  - Serveur vulnérable
  - Entre le client / serveur et le P2P (en penchant client/serveur)



4

# Le modèle de déploiement des grilles : le modèle distribué pair à pair

- Utilisés principalement
  - Pour le stockage, principe :
    - Ne pas être vulnérable
    - Découverte des ressources par diffusion
- Exemples
  - Gnutella (Musique)
  - Freenet, et FreeHaven (Tout type d'information)
  - KaZaA, JXTA (recherche et partage de données même à travers les pare-feux)
  - OceanStore (Global Scale Persistent Data)
- Problèmes crucial : résistance aux pannes des pairs, qu'elles soient volontaires (extinction de la machine), ou non (crash système, panne de courant, etc.), voire malicieuses (piratage du logiciel conduisant à émettre volontairement des informations incorrectes aux autres pairs).



## Desktop Grid ou « *Internet Computing* »



- Principe : récupérer les ressources inutilisées de machines distribuées sur le réseau pour une tâche définie,
  - le plus souvent de manière complètement transparente pour l'utilisateur habituel de ces machines (47% des cycles proc. Inutilisés)
  - Dès que la machine détecte que son utilisateur est inactif, par exemple la nuit ou le week-end, elle signale à un client central qu'elle est disponible comme serveur à sa disposition.
- Spectre étroit : *Embarrassingly parallel Apps*

6

## *Internet Computing (2/2)*

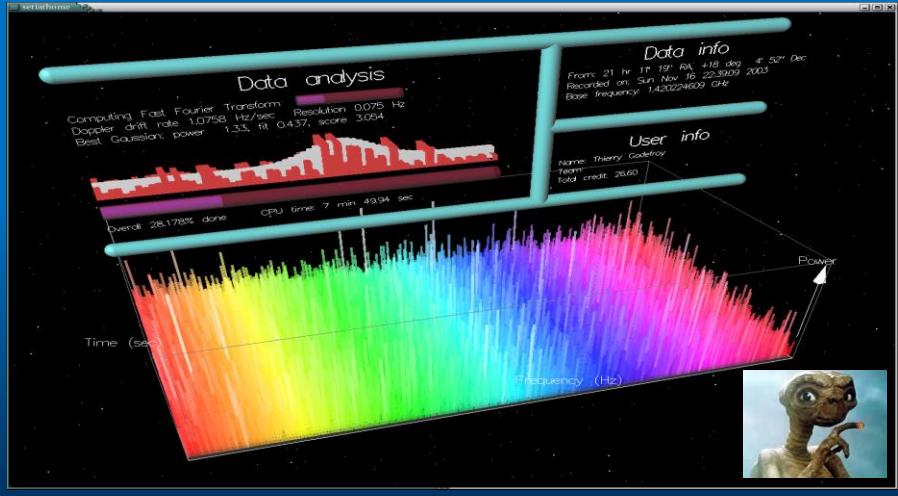
- **Example d'application  
(usage non commercial)**
  - BOINC : open-source software platform for computing using volunteered resources.
  - Became famous with SETI@home project (3 millions PCs workstations distributed all over the world and giving their unused computing power to **Search Extra-Terrestrial Intelligence** (<http://setiathome.ssl.berkeley.edu/>)).
  - The global power was equivalent to 33 TeraFLOP/s around Y2K
  - Others domains : Protein folding, cryptographic security, climate change,...



7



Economiseur d'écran ...  
... réalisant du calcul...  
et cherchant **ET**...



De nombreux projets plus sérieux...



**« Que reste-t-il à l'homme de toute la peine et de tous les calculs pour lesquels il se fatigue sous le soleil ? »**

(L'Ecclésiaste 2,21) 10

# Modèle Client/Serveur pour grille de calcul : « *Le MetaComputing* »

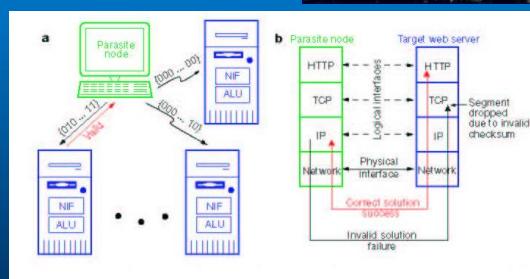
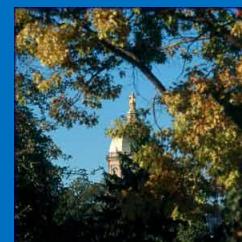
- **Principe : Acheter du service de calcul sur l'Internet**
  - service =
    - application pré-installée
    - +
    - calculateur
- **Ressources :**
  - puissance de calcul
  - Capacité de stockage pour éviter les transferts
- **Exemple :**
  - NetSolve (Université du Tennessee)
  - NINF (Université de Tsukuba)
  - DIET (ENS Lyon et INRIA)
- **Défis : sécurité dans les transferts et le calculs**



## UNIVERSITY OF NOTRE DAME

### Le « *Parasitic computing* » ?

- Publié dans « Nature » par Barabasi et ses collègues
- Utilisation des protocoles comme outils de calcul à l'insu des serveurs contactés !
- Réalisation de calculs NP complets en « parasitant » des serveurs...
- [Barabasi et al. 2001], « *Parasitic computing* », Nature, Volume 412, 30/8/2001



Savez vous vraiment ce qu'est-ce  
une grille de calcul ?



Analogie - Power Grid  
Distribution de la puissance électrique



Credit :  
T. Priol

Interconnection de PCs, de Clusters, de Supercalculateurs...  
Distribution de la puissance informatique  
grâce aux lignes réseau à hauts débits



**Elément de base**

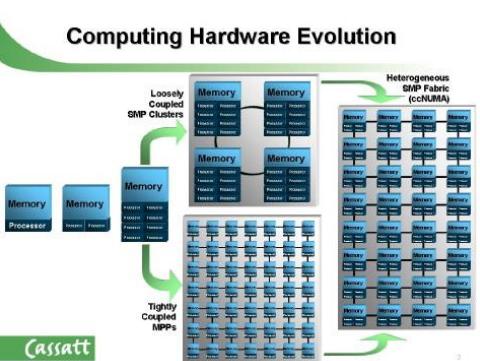
SMP :  
“Symetric MultiProcessing”

Architecture parallèle avec partage de la mémoire par les processeurs

« Clusters » : Assemblage de Machine SMP Fortement couplées.

## Evolution des machines

**Computing Hardware Evolution**



**Cassatt**

**Hank Schiffman**, Mountain View, California



ccNUMA :  
Cache Coherent Nonuniform Memory Access.  
Interconnection de groupes de 4 processeurs pour les applications souffrant peu d'un passage à l'échelle (grd # de CPUs)  
Développé par Data General.<sup>15</sup>

## Modèle de calcul : « *Grid Computing* »

- **Principe : Offrir des ressources de calcul « hors normes »**
- **Moyen :**
  - Faire exécuter ses gros calculs (typiquement des simulations numériques de plusieurs heures) sur des ordinateurs distants
  - Utiliser les ressources lourdes de plusieurs centres de calcul distribués sur l'ensemble de la planète.
  - Les ressources utilisées sont
    - La puissance de calcul,
    - La capacité de stockage (disques, tours, robots, etc.),
    - La capacité d'acquisition de données (accélérateurs de particules, microscopes électroniques, satellites, sismographes, etc.)
    - et surtout d'exploitation des données (visualisation immersive en réalité virtuelle, etc.).

## Le « grid Computing » (suite)

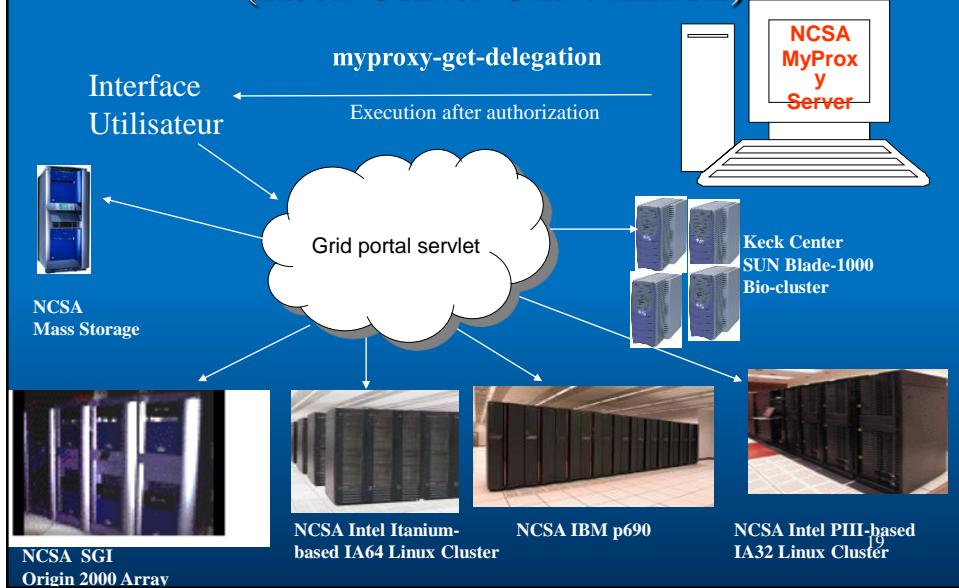
- On peut avoir également une coopération entre serveurs selon un schéma en général statique, par exemple un pipeline:
  - les données sont acquises en un lieu de la planète,
  - traitées en un autre lieu,
  - et finalement visualisées dans un troisième lieu.
- Le principal environnement de gestion de ce type de grille était Globus. La plate-forme GUSTO démontrée en février 2000 réunissait 125 sites dans 23 pays différents (<http://www.globus.org/>). Aujourd'hui la grille EGEE est un autre exemple concret, sinon la plus grande, grille institutionnelle (> 350000 CPUs).
- **Problèmes :** procédures administratives imposées par les centres de calcul (inscription, réservation des ressources, facturation, etc.), de confidentialité des données...

17

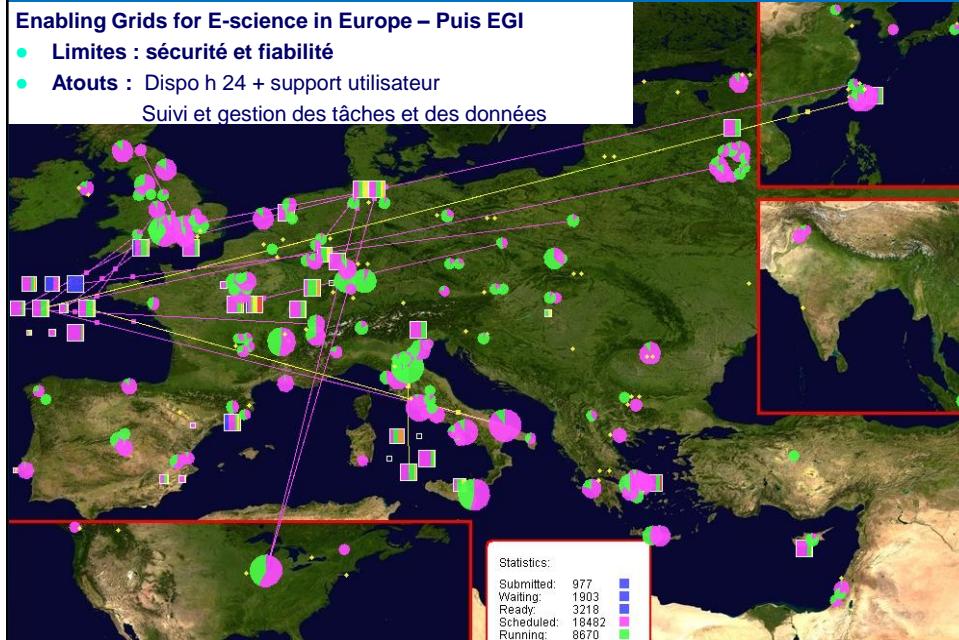
Caractéristiques	Cluster	Grille	P2P
Type d'ordinateur	PC dédiés ou racks de PC	Serveurs, stations de travail et PC dédiés	Feuille du réseau (PC de bureau)
Propriétaire (# entité)	Une seule	Plusieurs	Plusieurs
Gestion des utilisateurs	Centralisée	Décentralisée	Décentralisée
Gestion des Ressources	Centralisée	Distribuée	Distribuée
Allocation/Ordonnancement	Centralisé	Décentralisé	Décentralisé
Inter-Opérabilité	Pas de standard	Pas de standard	Pas de standard
Taille possible	100s	1000?	Millions? [@Home]
Capacité de calcul	Garantie	Elevée mais peut varier	Elevée mais peut varier
Bande passante	Très bonne	Faible	Faible

18

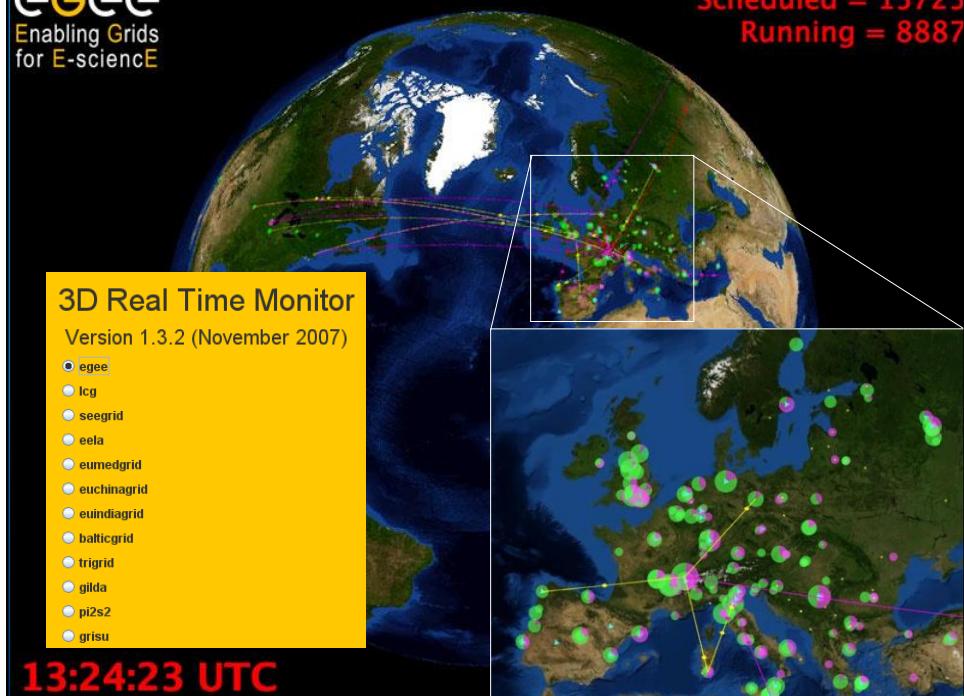
## Exemple de campus “Grid-Aware” (Keck Center Univ. Illinois)



## Grille EGI (> 530 000 lCPUs, > 200 Po)



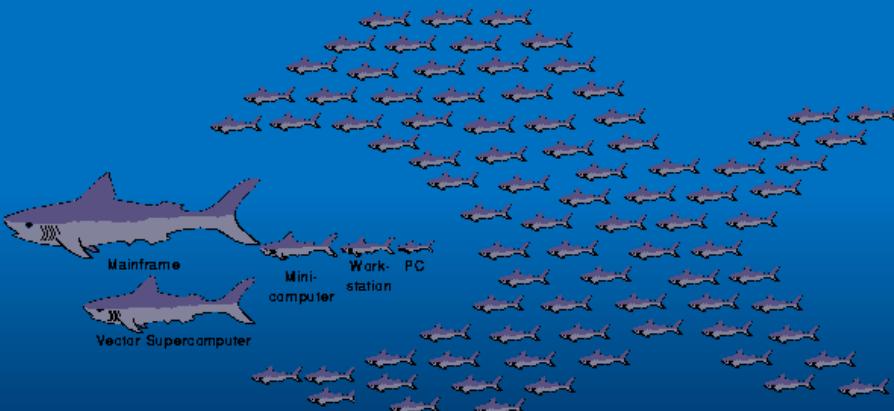
Scheduled = 15725  
Running = 8887



Même ici, des ordinateurs inutilisés...



## La chaîne alimentaire...



23

NOW



Centre de Recherche  
de Jülich  
Allemagne

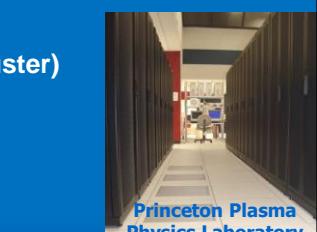
Mare nostrum  
"Chapel Supercomputer"  
Polytechnical University  
in Barcelona, Spain



- Desktop grids – Interconnections de PC distants sur Internet (evol. Windows Comp. Cluster)
- Grilles institutionnelles
  - Grilles de clusters
    - Calculs indépendants, peu de communications, taille mémoire raisonnable
  - Grilles de super calculateurs (DEISA)



Différents  
types  
de grilles



Princeton Plasma  
Physics Laboratory



Earth Simulator Center Japan

# Un projet aux origines : le projet européen DataGrid



## Objectifs

- **LHC (Large Hadron Collider)**
  - LHC : Nom donné à l'accélérateur de particule le plus grand du monde, Situé au CERN à Genève
  - Hadron : Terme employé en physique des hautes énergies pour désigner les particules à interaction fortes.
  - A partir de 2006, Traitement de 100 Mo de données par seconde :
    - Besoin d'une grille de calcul
- D'autres domaines d'applications:
  - Biologie
  - Observation de la Terre
  - Industriels...

25



the globus project™

## Initialement : le logiciel de grille Globus

- Utilisé dans le projet européen DataGrid
- Développé par les universités américaines depuis 10 ans
- Interface entre les éléments de la grille
  - Les systèmes d'exploitation
  - Les réseaux
- Gestionnaire des fonctionnalités basiques de la grille
  - communication,
  - recensement,
  - sécurité,
  - Authentification\*
- Passage à « Glite » puis aux outils intégrés et plus évolués de soumission. Ex: Dirac, OpenMole...

26

<http://www.openmole.org/current>

**Getting Started   Documentation   Who are we?**

Need to tune parameters of your program?  
 Need to execute your program against many datasets?  
 Need to calibrate?  
 Need to optimize?

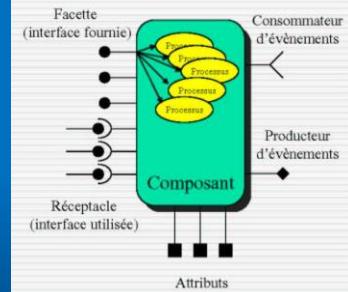
OpenMOLE (Open MOdel Experiment) makes it simple to execute your programs on distributed computing environments. If you want to execute the same program for many different inputs (parameters or datasets), OpenMOLE is the tool that you need. The typical usages of OpenMOLE are high performance model calibration, model exploration, machine learning, optimization, data processing.

- **Works with your programs** - Java, Binary exe, NetLogo, R, Scilab, Python, C++...
- **Distributed computing** - Works on your multi-core machines, clusters, grids, desktop grid.
- **Expressive** - Workflow language to describe your naturally parallel processes.
- **Scalable** - Handles millions of tasks, years of computation, and GBs of data.
- **Mature** - Developed since 2008 and widely used.
- **Open** - AGPLv3 free software license.

## Gestion des « workflow » scientifiques et des plans d’expérience

## Défi : vers une approche plus « moderne »?

- **Elargir le spectre des applications**
  - De vrais applications parallèles : SMA ?
- **Construction d'applications à l'aide d'Objets distribués (et d'Agents ?)**
  - Structuration de l'application
  - Encapsulation des codes parallèle
- **Couplage des codes parallèles**
- **Deux aspects s'opposent**
  - Cacher la complexité dans un composant
  - Connecter des composants tout en autorisant des communications inter-échelles



29

## Défis : des agents sur les grilles ?



- **Projet D-Agent pour le Méta-Computing**
  - Les agents mobiles achètent des droits d'accès aux ressources de machines hôtes et établissent un marché des ressources de calcul.
  - <http://agent.cs.dartmouth.edu/papers/#market>
- **Projet Echelon : Agent Based Grid Computing Architecture**
  - <http://www.geocities.com/echelongrid/>
- **Agent Based computational Economics**
  - <http://www.econ.iastate.edu/tesfatsi/ace.htm>

30



## Le besoin de simuler les grilles

- GridSim : Boîte à outils pour l'évaluation de l'ordonnancement des applications sur une grille.
  - <http://www.cs.mu.oz.au/~rai/gridsim/>
- Il existe des limites pour l'accès aux machines sur une grille. Avec un peu de chance on récupère 20 machines
- Pour l'exploration scientifique, il faut tester des algorithmes d'ordonnancement sur des milliers de machines avec des conditions de charge différentes, des scénarios utilisateurs différents,...
- En utilisant un simulateur développé avec GridSim, on peut créer des millions de machines virtuelles, d'utilisateurs et des scénarios d'application pour évaluer les performances d'une grille.

31

## Quelques problèmes

- **Middleware et système : vers un « Grid-aware OS » ?**
  - Gestion dynamique des ressources
  - Equilibrage dynamique par redistribution à la volée des données
  - Non prédictibilité des performances des réseaux (WAN notamment)
- **La programmation**
  - Repenser l'algorithme parallèle (conçue habituellement pour des architectures parallèles régulières et à configuration statique)
  - Couplage des codes : une application est un assemblage de plusieurs codes de calcul.
- **Croissances des besoins en « agents mobiles »**
- **Défis pour la simulation : distribuée et stochastique**

# Des solutions ?



**the globus project™**

**GLite**

## La conception d'outils logiciels

- **Middleware ou Intergiciels**
  - Legion
  - Condor
  - Globus
  - Glite
  - ...
- **Unicore :**
  - outil permettant la construction, la soumission, le contrôle et la surveillance de jobs batch de façon uniforme au sein d'un environnement hétérogène
  - Permet à l'utilisateur de s'affranchir des problèmes d'implémentation relatifs à chaque environnement
  - Description abstraite d'un job (job portable d'un environnement à un autre)
- **Upperware : Proactive**

34

**OMG**  
OBJECT MANAGEMENT GROUP  
**LSR**  
LIFE SCIENCES RESEARCH

## Et les sciences de la vie ?

- Besoins d'Interopérabilité
- Besoin de Sécurité
- Tolérance aux fautes
- Hétérogénéité des réseaux et des ordinateurs
  - SAN, LAN, WAN
- Identifier ce qui est technologique
  - Effet de mode
- Identifier ce qui est plus fondamental
  - Les nouveaux concepts
- Humains
  - Faire collaborer des chercheurs de disciplines différentes (physiciens, biologistes, informaticiens, mathématiciens,...)

Molécule

Protéine

Travail concret à grande échelle pour des applications innovantes

35

## Les défis « urgents » lancés par la Bioinformatique

**nature**  
10 June 1999 Volume 399 Issue no 6736  
Help! The data are coming

	complets	en cours
archébactéries	16	21
bactéries	73	295
eucaryotes	9	207

<http://wit.integratedgenomics.com/GOLD/>

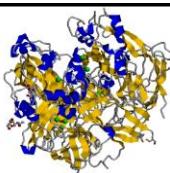
Growth of GenBank

EMBL/GenBank/DDBJ

NCBI

GenBank  
August 2002  
Release 131.0

18 197 119 séquences  
22 616 937 182 bases



## Urgences et prudence...

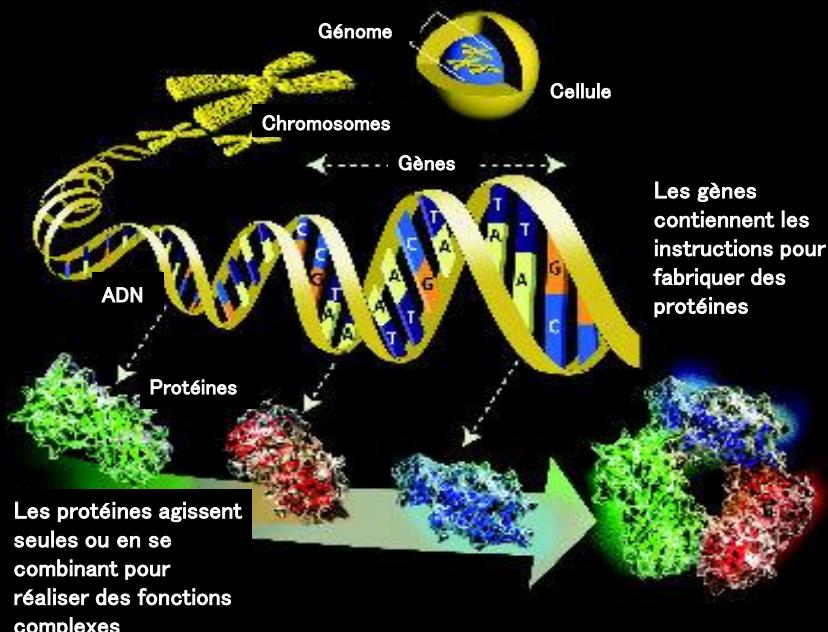
Le Monde 17/9/2002 : Katherine Nelson, était une des responsables du grand projet international de séquençage du génome humain à Berkeley avant de rejoindre le secteur privé. Elle est maintenant catégorique :

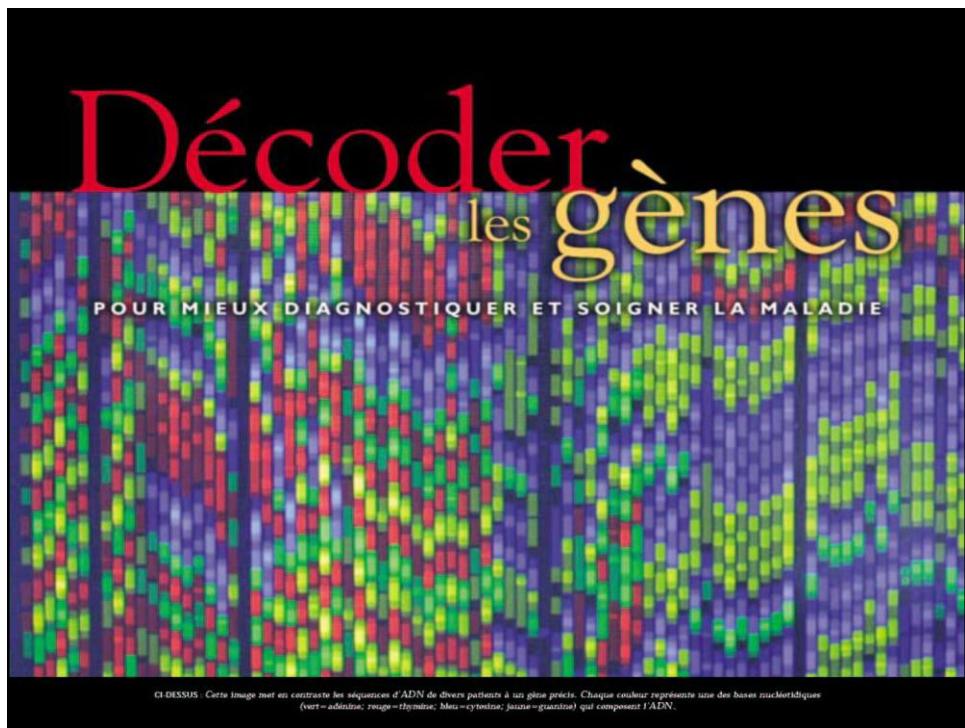
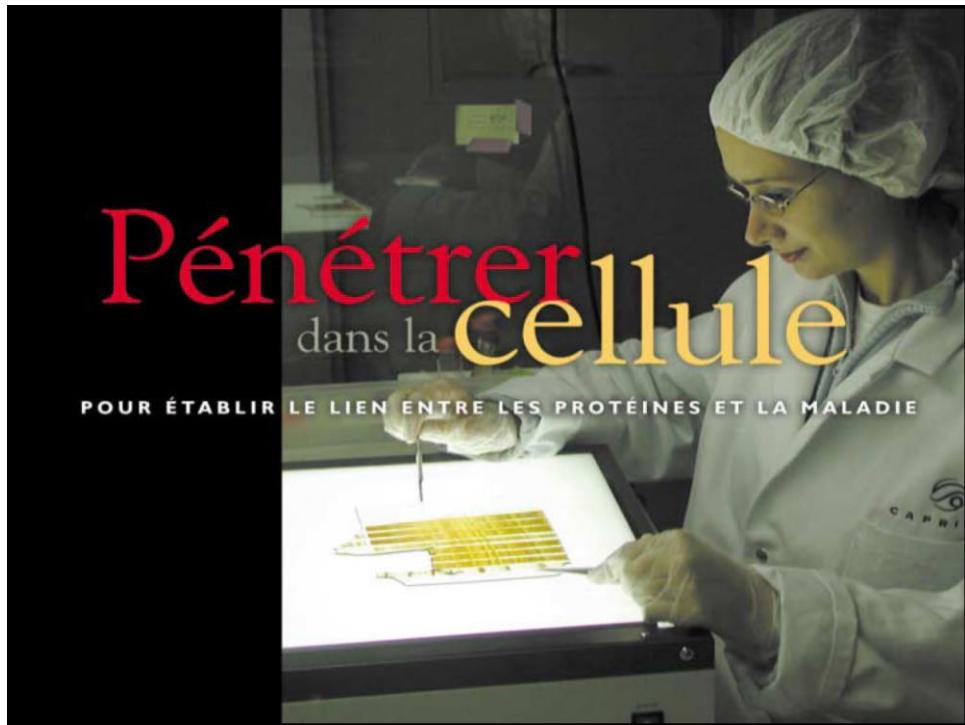
*"Nos patrons se fichent éperdument de guérir le cancer, ils veulent gagner beaucoup d'argent très vite, c'est tout. Notre entreprise a breveté 800 gènes responsables de certains cancers, et désormais elle confisque cette information pour son seul usage.*

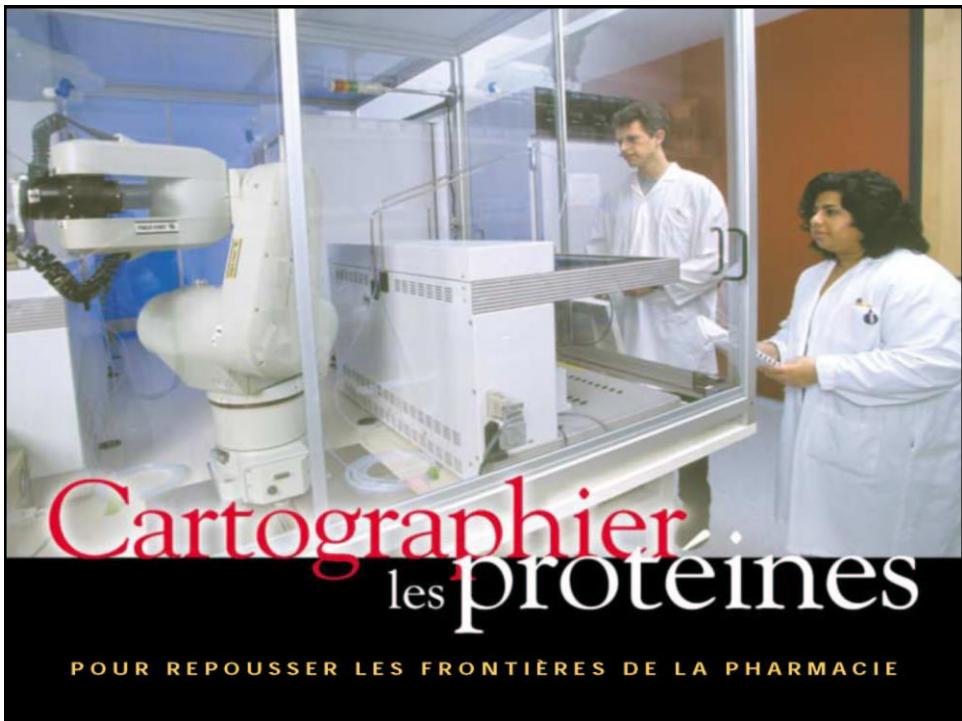
*Si nous partagions nos résultats, d'autres labos se joindraient à nous, et ensemble, nous trouverions des remèdes plus rapidement, mais on nous l'interdit. Au contraire, nos chefs nous ordonnent souvent d'abandonner des pistes prometteuses parce qu'ils ont peur que ce ne soit pas rentable.*

*Tout le système est pervers : les laboratoires privés collectent des informations scientifiques du domaine public, ils y rajoutent un petit quelque chose, puis ils déposent un brevet couvrant la totalité des données. C'est du vol légalisé."*

37

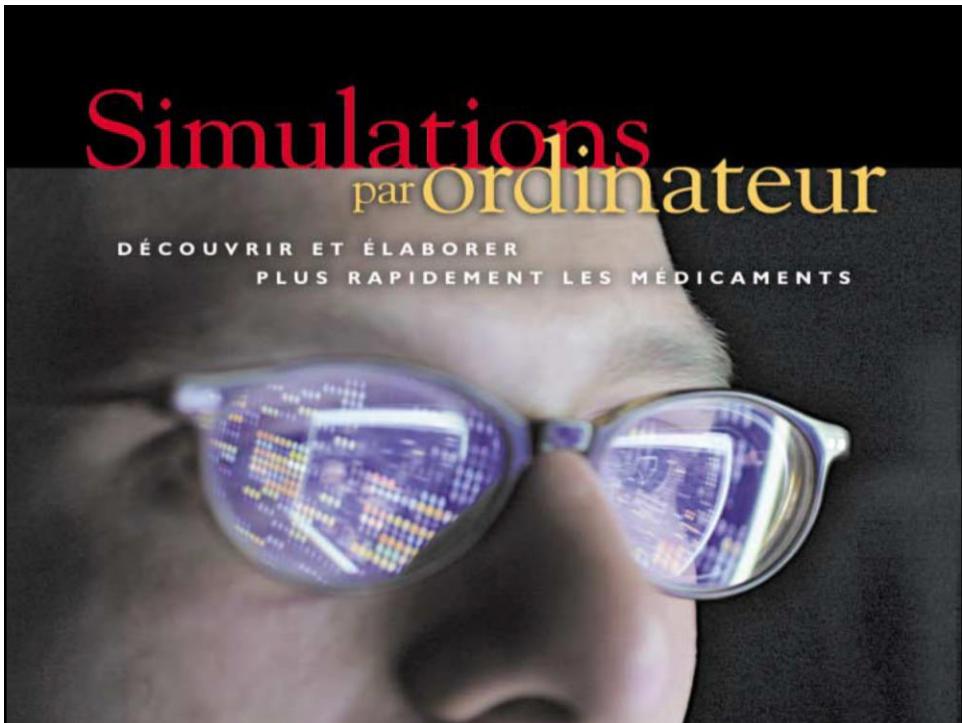






# Cartographier les protéines

POUR REPOUSSER LES FRONTIÈRES DE LA PHARMACIE



# Simulations par ordinateur

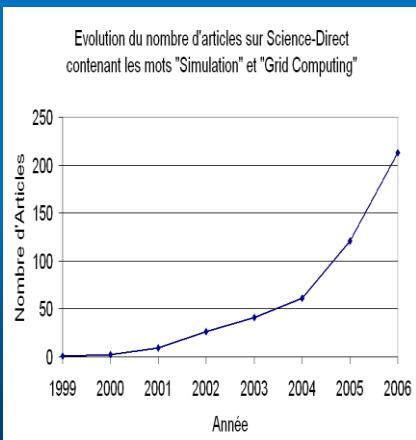
DÉCOUVRIR ET ÉLABORER  
PLUS RAPIDEMENT LES MÉDICAMENTS

## Quelques pistes...

- Analyse et annotation autonome de génomes : BioMAS, RETSINA (<http://citeseer.nj.nec.com/445758.html>)
- Architectures multi-agents pour applications génomiques
- Approches multi-agent pour la classification et le traitement des EST (Expressed Sequence Tags)
- Apprentissage et découverte de la connaissance distribuée de manière autonome
- Approches multi-agent pour l'analyse de l'expression des gènes (utilisation des biopuces)
- Coordination et Contrôle de Systèmes Multi-Agents pour la collecte de données bioinformatique distribuées (<http://www.cs.iastate.edu/~honavar/ailab/projects/control.html>)
- Simulation multi-agents du fonctionnement de l'interaction moléculaire des protéines, des mécanismes de cancérisation...

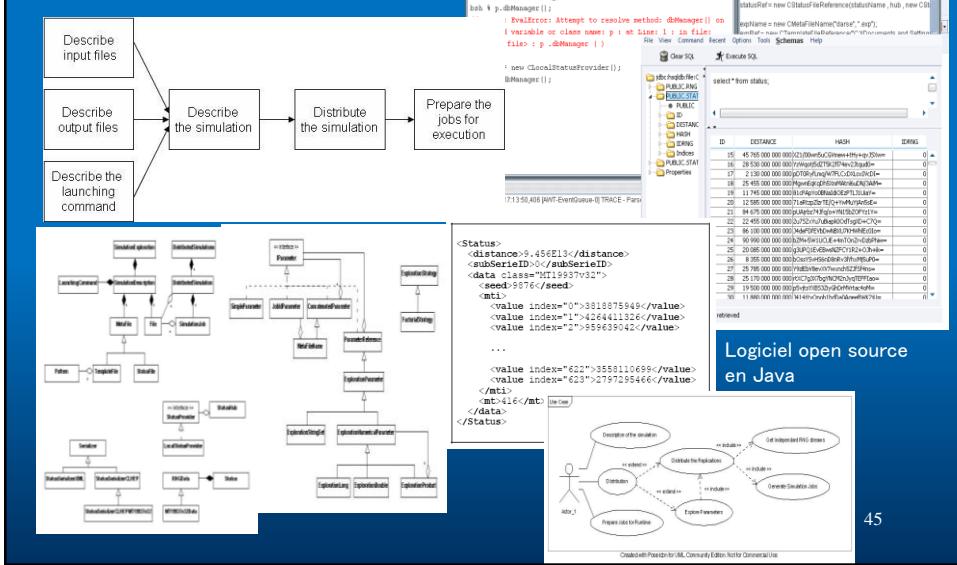


## DistMe un exemple d'outil pour ce contexte



- **Besoin d'outil pour automatiser autant que possible la parallélisation de simulations stochastiques quelque soit l'environnement d'exécution cible.**
  - Par la distribution des réplications
  - Par la distribution de plan d'expériences
  - Par les deux
- **L'objectif est de générer des simulations distribuées prêtées à l'exécution pour des environnements de calcul réparti divers :**
  - Cluster PBS, Grilles de calcul, Internet-Computing...<sup>44</sup>

# Projet DistMe – Approche Ingénierie des modèles



45

## Application à la simulation stochastique à grande échelle pour l'environnement

- Application: dispersion de l'algue *Caulerpa taxifolia*
- Multi-modèle :
  - Simulation individu centrée
  - Simulation combinée continue et discrète de certains processus
  - Interactions locales et stochastiques à distance
  - Simulation couplée à un système d'informations géographique
  - Travail à différentes échelles spatiales
- Simulation distribuée
  - réplications (distribution normale)
  - Intervales de confiance pour distributions asymétriques
  - Vérification et Validation



# Distribution de calculs pour analyses spectrales de résultats de simulations stochastiques sous contraintes spatiales

- Contexte : Simulation pour l'environnement.  
Propagation de l'algue *Caulerpa taxifolia*  
(Méditerranée et maintenant USA).
- Utilisation de métamodèle et métaprogrammation  
pour la distribution de réplications et de plans  
d'expériences visant à fournir des cartes de  
Probabilité de colonisation.

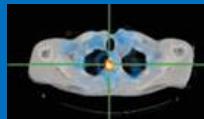
Credit : D. Hill, P. Coquillard,  
J. De Vaugelas, A. Meinesz

## Caulerpa 2...

# of workstations with 10000 parcels	# of logical processes / workstation	Performance gain obtained by adding the other workstations
1	10	
1=>2	5	20%
2=>5	2	13%
5=>10	1	30%

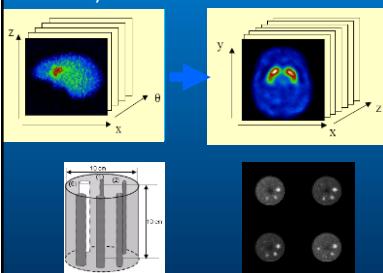


# DistMe : Application à la médecine nucléaire



Credit : I. Buvat, Z. El Bitar

V. Breton, D. Hill



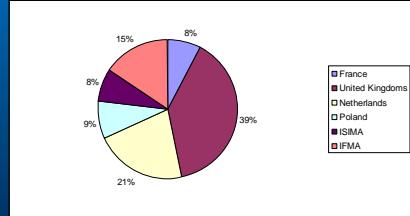
Calcul sur grille et cluster – amélioration d'images de tomographie par simulation de Monte-Carlo

Parallélisation du générateur aléatoire :

- 80 jours de calcul séquentiel pour la parallélisation
- 35 jours de calcul sur un cluster de 28 Xeon 3 GHz (= 3 ans / CPU) pour les tests

Calcul pour la reconstruction :

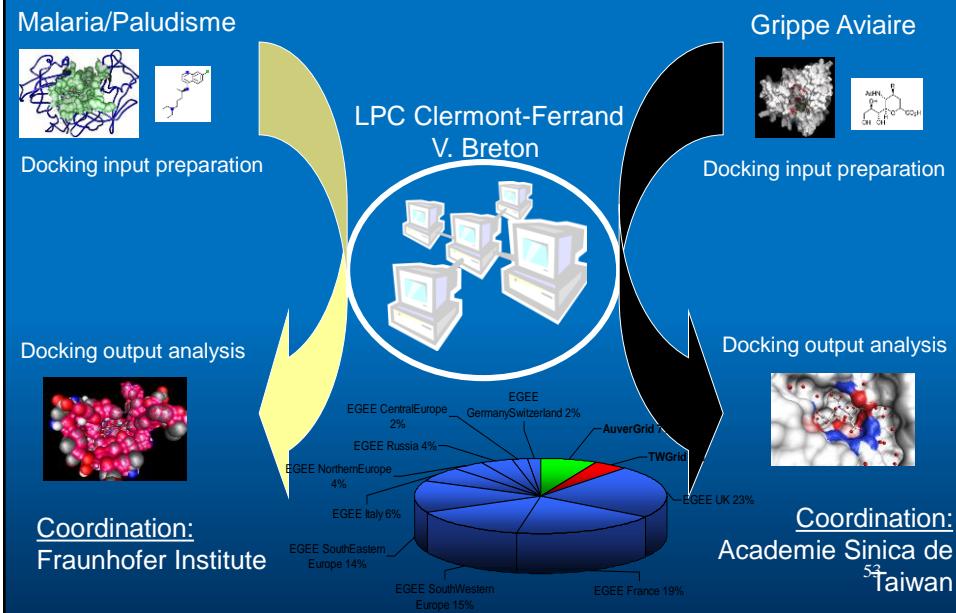
- 2 jours utilisant 600 nœuds de la grille de calcul EGGE (= 3 ans CPU) et deux clusters (avec 84 unités de calcul type Xeon 3 GHz)



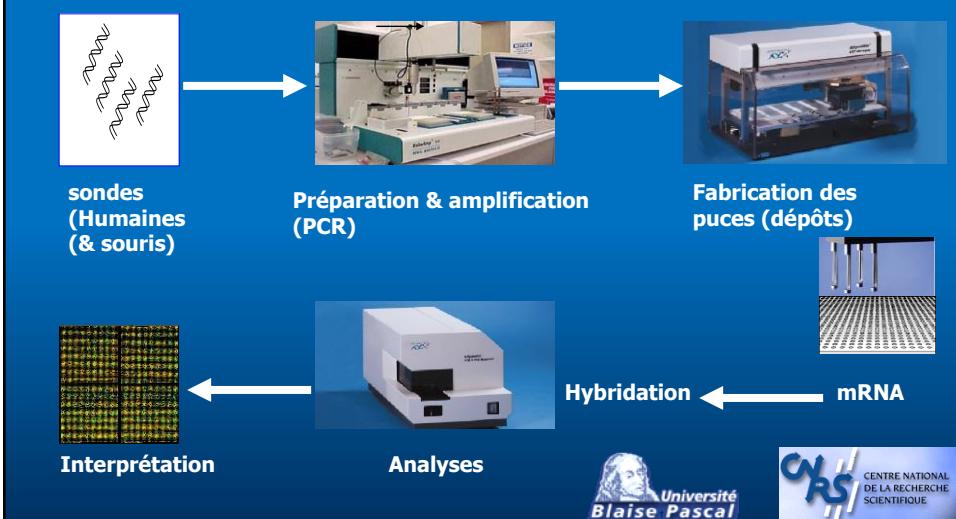
## DistMe – Perspectives

- Mise à disposition d'un environnement pour l'exploitation de simulations stochastiques parallèles – distributions de réplications et de plans d'expériences. (DistMe)
- Ingénierie des modèles – Métamodélisation & Introspection pour la production de générateurs paramétrables par des sérialisations (DistRNG)
- Tests de séquences pseudo-aléatoires
  - Tests d'indépendance des flux de nombres pseudo-aléatoires sur une plateforme BOINC (exécution au minimum d'une batterie de tests par couple de flux = minimum 22000 ans / CPU pour 10 000 flux)
  - Web-Services de mise à disposition des données
- Références
  - REUILLOU R., "Etude empirique de l'hybridation par brassage d'un générateur de nombres quasi aléatoires", *Technique et Science Informatiques*, numéro spécial "Jeunes Chercheurs en STIC", 23 pages, in press, 2007.
  - MAGNÉ L., HILL D., BRETON V., REUILLOU R., CALVAT P., LAZARO D., LEGRE Y., DONNARIEIX D., "Parallelization Of Monte Carlo Simulations And Submission To A Grid Environment", *Parallel processing letters*, Volume 14, N° 2, pp. 177-196, 2004.
  - HILL D., KOMATSU T. (Eds), "Environmental Modelling and Simulation", Special Issue of the *Simulation Journal* & Guest Editor's Introduction, Volume 82, N° 7, pp. 425-495.
  - EL BITAR Z., LAZARO D., BRETON V., HILL D., BUVAT I., Fully 3D Monte Carlo image reconstruction in SPECT using functional regions. *Nucl. Instr. Meth. Phys. Res.* in press, 2006.
  - LAZARO D., EL BITAR Z., BRETON V., HILL D., BUVAT I., "Fully 3D Monte Carlo reconstruction in SPECT : a feasibility study", *Phys. Med. Biol.* Volume 50, 2005, pp. 3739 - 3754
  - REUILLOU R., HILL D. R. C., "Faster Quasi Random Number Generator for Most Monte Carlo Simulations", *Mathematics and Computers in Simulation*, submitted.

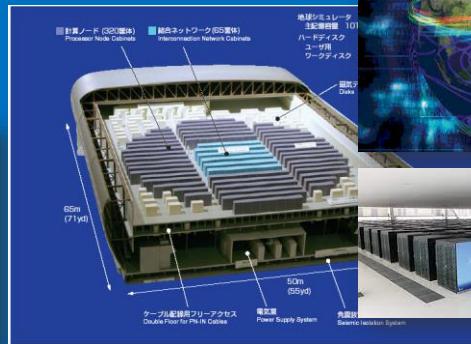
## Calcul massif pour les maladies émergentes ou négligées – Credit : N. Jacq – PCSV – Clermont-Fd



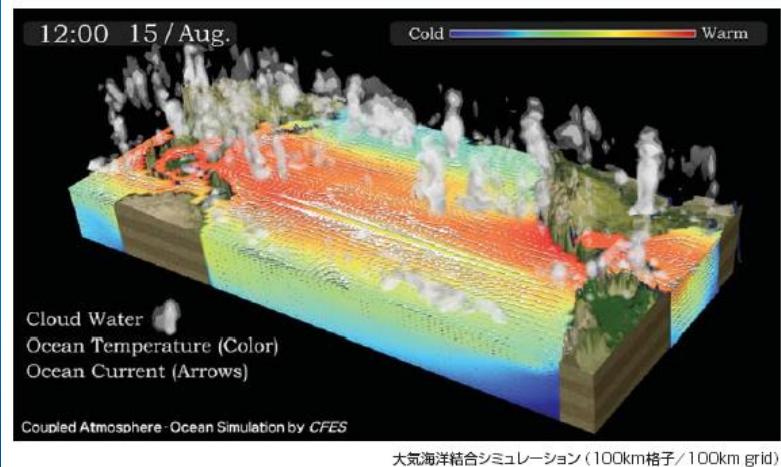
## Conception et exploitation de biopuces à ADN pour analyser l'expression des gènes



# Approche Holistique Nippone



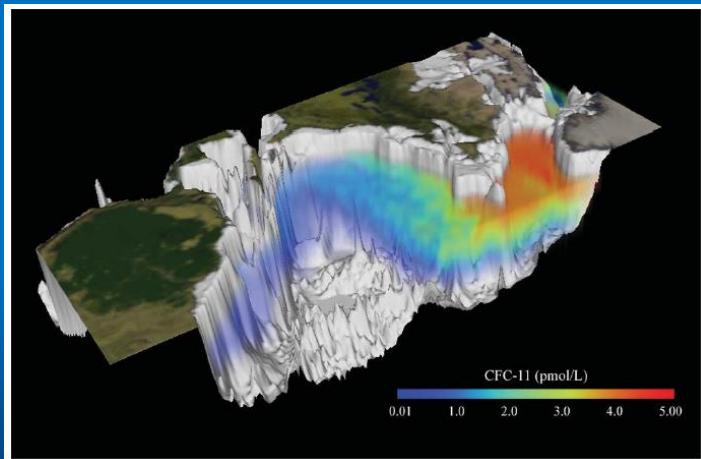
## Interaction Océan - Atmosphère



L'océan réchauffe l'atmosphère ce qui génère du vent. Et le vent conduit les courants de l'océan. Mise en évidence des interactions étroites entre les courants de l'océan et l'atmosphère.

56

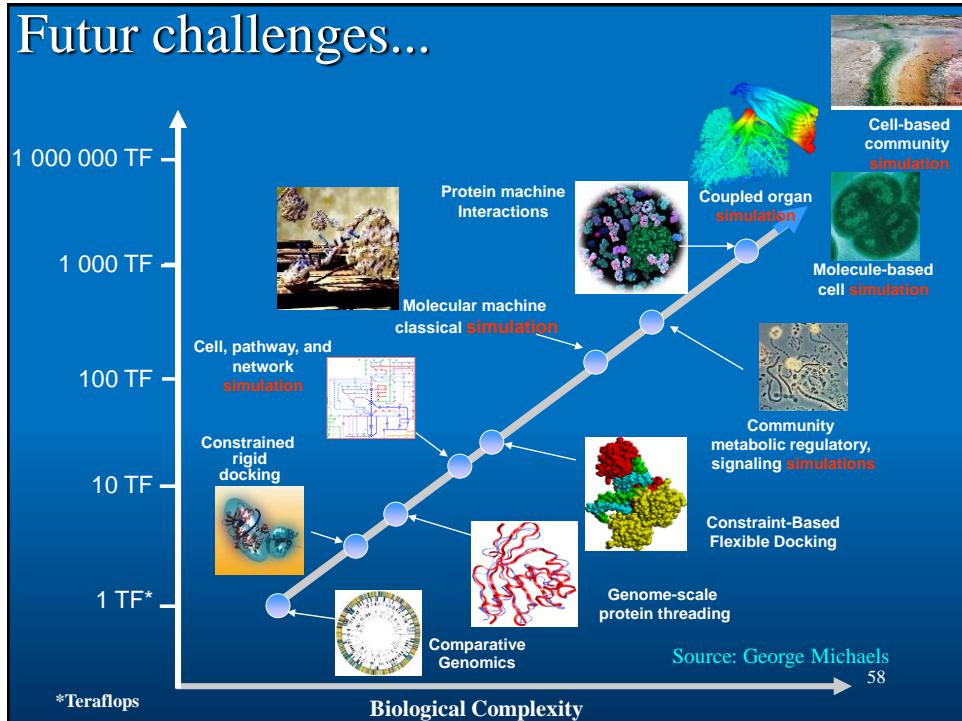
## Simulation de courants profonds dans l'Océan Atlantique



Le Fréon a été transporté vers le Sud par le courant profond littoral ouest.

57

## Futur challenges...



That's all folks...



Benny Hill  
21 Janvier 1924 – 18 Avril 1992

59