L’UTILISATION DES GRILLES ET DU CLOUD EN SCIENCES DU VIVANT

Dans ce chapitre, nous présentons les acteurs de la recherche de nouveaux médicaments issus de la biodiversité au Vietnam. Ensuite, nous présentons un état de l’art de la recherche de nouveaux médicaments sur la grille. Pour cela, nous présentons les principales infrastructures de grille ainsi que les plates-formes existantes.

2.1 Recherche de nouveaux médicaments issus de la biodiversité au Vietnam

2.1.1 Présentation de la VAST et de l’INPC

L'Académie vietnamienne des sciences et de la technologie (VAST- Vietnam Academy of Science and Technology) est la principale agence gouvernementale scientifique et technologique du Vietnam. Dédiée à l'étude des sciences naturelles et au développement de nouvelles technologies en accord avec les grandes orientations de l'État, son rôle est de fournir une base scientifique et technologique pour la construction de politiques, de stratégies, et de plans de développement socio - économique, ainsi que de former des ressources humaines scientifiques et technologiques de haute qualité pour le pays. La VAST regroupe 30 instituts nationaux et 7 unités non universitaires, 9 entreprises d'État propre, plus de 20

syndicats de production et 35 unités d'instituts, principalement à Hanoi, Ho Chi Minh ville, Haiphong, Nha Trang, Dalat et Hue.

L'Institut de Chimie des Produits Naturels (INPC - Institute of Natural Products Chemistry) de l'Académie Vietnamienne des Sciences et de la Technologie a été créé sur la base du décret 65/CT en date du 5 Mars 1990 par le gouvernement du Vietnam. Les domaines de recherche de l'INPC sont la chimie des produits naturels, la biochimie, la chimie de l'environnement et leurs applications, notamment industrielles. L'INPC s'intéresse notamment à la recherche de composés bioactifs issus de la biodiversité marine et terrestre, dans la perspective de synthétiser des produits valorisables par l'industrie pharmaceutique.

2.1.2 Présentation de la base de données de composés de l'INPC

Au Vietnam, la médecine traditionnelle a une longue histoire de développement. Elle utilise des parties de plantes médicinales naturelles, comme par exemple, les racines, les fleurs, les tiges d'un arbre, les feuilles... Il y a actuellement environ 4000 plantes médicinales enregistrées au Vietnam. L'Institut de Chimie des Produits Naturels de l'Académie des Sciences du Vietnam (INPC) collecte des échantillons issus de la biodiversité locale et détermine la structure tridimensionnelle des molécules isolées. Il y a maintenant environ 200 composés isolés à partir des animaux marins et 300 composés à partir des plantes. L'enjeu pour cet institut est de constituer une base de données des échantillons et de mettre en place une chaîne de traitement des informations structurelles permettant de déterminer sur quelles cibles biologiques les produits isolés sont potentiellement actifs. Comme l'illustre la Figure 2-1, il s'agit d'utiliser des ressources de calcul pour sélectionner parmi les cibles biologiques recensées dans la base de données Protein Data Bank [1] celles sur lesquelles les composés contenus dans la base de données de l'INPC sont les plus prometteurs. Il ne s'agit pas pour l'INPC de développer des médicaments car le coût complet du développement d'un nouveau médicament est typiquement de l'ordre d'un milliard d'euros et requiert l'intervention de laboratoires pharmaceutiques notamment pour toutes les phases d'expérimentation clinique. Par contre, le dépôt d'un brevet international pour une molécule peut intervenir une fois que l'activité inhibitrice de celle-ci sur le virus, la bactérie ou le parasite vecteur d'une maladie est démontrée. Par exemple, les médicaments actuels les plus efficaces dans le traitement du

paludisme utilisent des composés extraits initialement de plantes de la famille de l'artémisine dont l'action était connue dans la médecine traditionnelle chinoise.

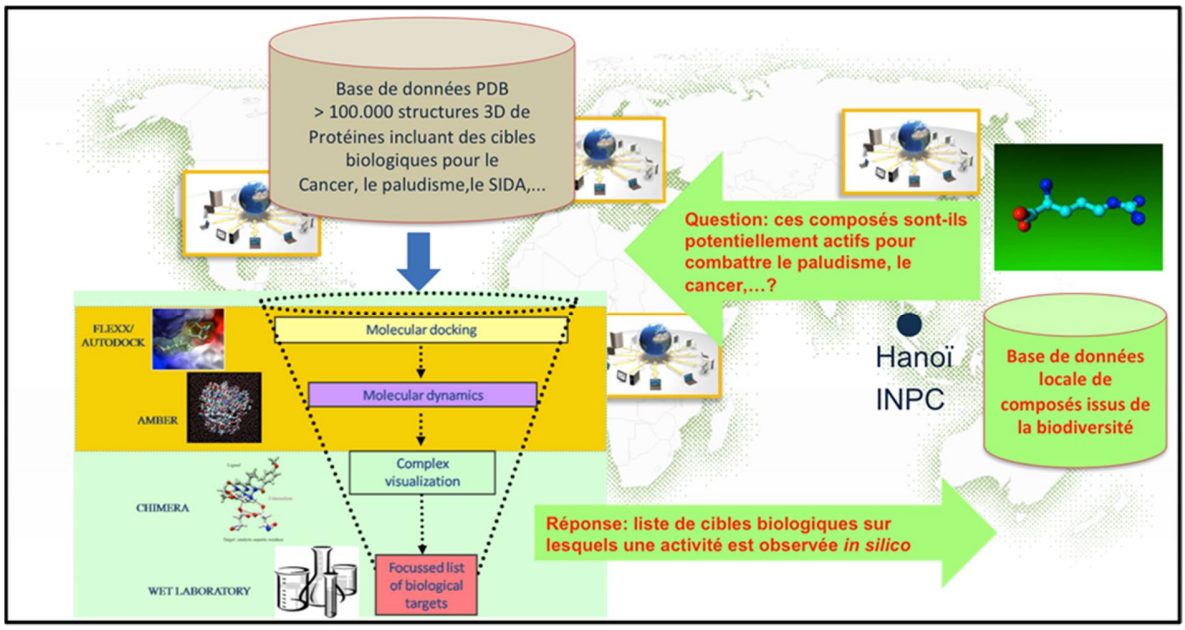

Figure 2-1 : Représentation schématique de la chaine de traitement des composés chimiques isolés par l'INPC

Nous allons maintenant présenter l'état de l'art de l'utilisation des grilles informatiques pour la recherche de nouveaux médicaments.

2.2 Recherche de nouveaux médicaments sur la grille : Etat de l'art

La recherche *in silico* (*in silico* signifie assistée par ordinateur) de nouvelles molécules actives pour combattre une maladie a été identifiée très tôt comme une application prometteuse des grilles de calcul, car la première étape dite de criblage requiert de tester l'action inhibitrice d'un très grand nombre de composés sur une cible donnée [2].

2.2.1 Introduction au criblage virtuel

Le criblage virtuel est la sélection *in silico* des meilleurs candidats médicaments qui agissent sur une protéine cible donnée [3]. Le criblage peut se faire in vitro, à la paillasse, mais son

coût est très élevé : plusieurs euros par composé testé. Multiplié par le nombre de composés ou ligands qui peuvent être synthétisés par les industries chimiques [4], le coût d'un criblage systématique atteint des millions d'euros, somme hors de portée d'un laboratoire de recherche public. Un criblage *in silico* permet de sélectionner un nombre réduit de molécules prometteuses, ramenant le nombre de tests in vitro de quelques millions à quelques centaines.

Le criblage virtuel requiert de connaître la structure tridimensionnelle du site actif de la protéine ciblée, la structure tridimensionnelle du composé chimique ou ligand et un logiciel qui va calculer la probabilité d'ancrage (« docking » en anglais) du ligand sur le site actif de la protéine.

Par exemple, le virus H5N1 de la grippe aviaire porte le nom de deux protéines (H5 et N1) qui sont impliquées dans la réplication des virus et leur propagation de cellule en cellule (Figure 2-2). La neuraminidase N1 est une cible connue des médicaments antiviraux comme le Tamiflu utilisé pour traiter les malades atteints de formes graves de la grippe lors de la pandémie de grippe A de type H1N1 pendant l'été et l'automne 2009.

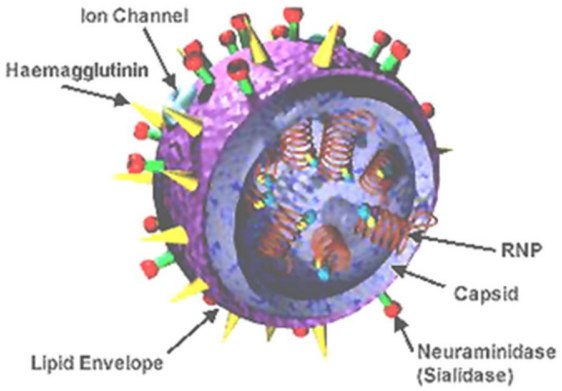

Figure 2-2: Structure du virus influenza

La neuraminidase constitue donc une cible de choix pour la recherche de nouveaux médicaments et plusieurs structures sont documentées dans la Protein Data Bank.

Le criblage virtuel permet la sélection et le classement *in silico* des meilleurs médicaments candidats, c'est à dire les molécules qui pourraient influer sur l'activité biochimique de la cible. Il peut être utilisé comme un filtre pour éliminer les composés toxiques qui sont susceptibles d'échouer dans la phase finale du processus de découverte de médicaments. Le criblage virtuel à haut débit permet ainsi de tester les grandes bibliothèques chimiques où sont enregistrées dans des bases de données les structures de plusieurs millions de composés disponibles dans les laboratoires ou entreprises.

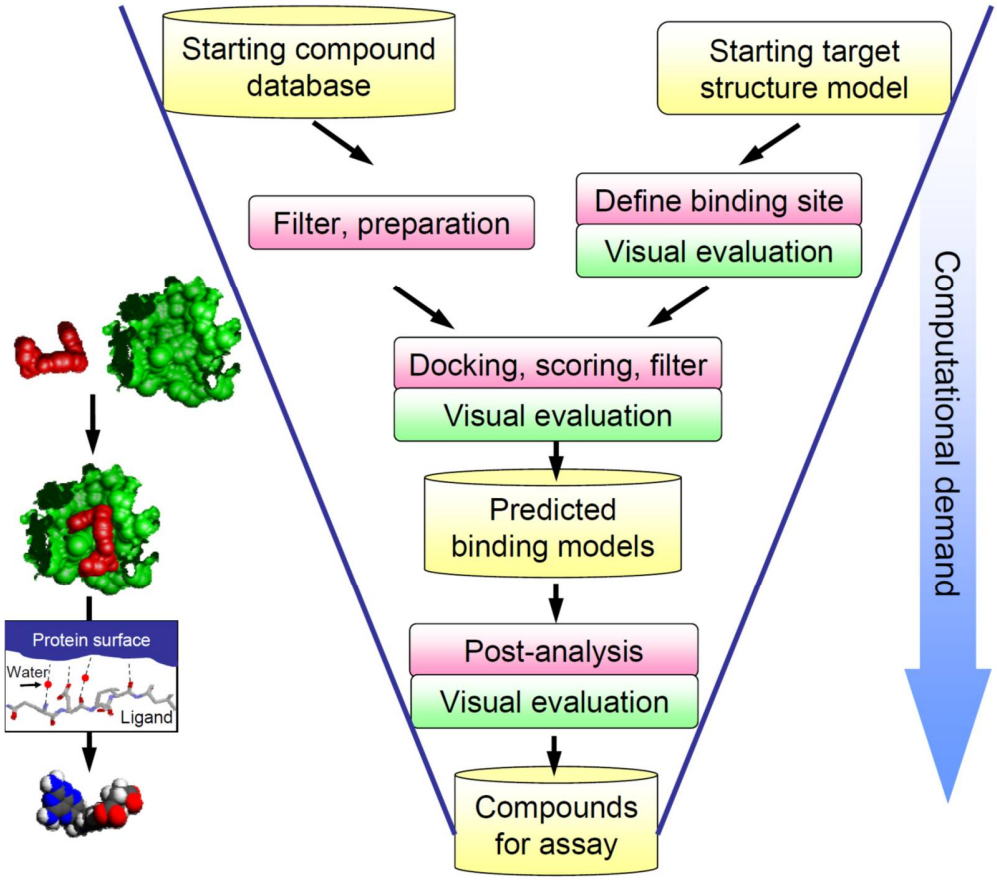

Figure 2-3: Etapes du criblage virtuel à haut débit

La Figure 2-3 décrit les étapes typiques d'un projet de criblage virtuel d'une base de données de composés sur une cible spécifique. Sa mise en œuvre requiert des points de contrôle visuel et éventuellement des cycles automatiques répétés.

La première étape consiste à choisir la cible biologique dans la Protein Data Bank et à préparer la base de données de composés chimiques ou ligands. Des bibliothèques de structures tridimensionnelles de composés sont disponibles sur internet, soit qu'elles soient issues de la recherche publique et fournies par des laboratoires universitaires, soit qu'elles soient proposées par les industriels qui commercialisent les composés. Il est possible aussi d'utiliser des bases de données privées de ligands comme celle de l'INPC.

Il est possible de cibler la recherche sur des familles de composés préalablement identifiés ou d'avoir une approche très globalisée en testant de façon aléatoire des millions de ligands. Cette deuxième approche est beaucoup plus coûteuse en temps de calcul.

Après le choix des ligands, il est nécessaire de préparer la description de la structure tridimensionnelle du site actif de la protéine cible, en ajoutant notamment des atomes d'hydrogène et des molécules d'eau.

L'étape suivante dite de « docking » est le calcul de l'énergie de liaison du ligand au site actif en utilisant un algorithme d'évaluation (scoring algorithm) par les modèles et les approximations. Beaucoup de logiciels de docking sont disponibles (Table 2-1) : ils diffèrent notamment par les modèles et approximations utilisés pour le calcul de l'énergie de liaison.

Nom	Editeur	Site Internet
Autodock	Scripps	http://www.scripps.edu/mb/olson/doc/autodock/
Surflex	Biopharmics	http://www.biopharmics.com/products.html
FlexX	BioSolveIT	http://www.biosolveit.de/FlexX/
Dock	UCSF	http://dock.compbio.ucsf.edu/
Glide	Schrödinger	http://www.schrodinger.com/Products/glide.html
Gold	CCDC	http://www.ccdc.cam.ac.uk/products/life_sciences/gold/
LigandFit	Accelrys	http://www.accelrys.com/cerius2/c2ligandfit.html
Fred	OpenEyes	http://www.eyesopen.com/products/applications/fred.html
ICM	Molsoft	http://www.molsoft.com/products.html

Table 2-1: Logiciels de « docking »

La post-analyse, la dernière étape, est alors nécessaire avant l'inspection visuelle finale. La notation de consensus [5], l'incorporation d'énergies de solvatation, une meilleure description des interactions électrostatiques [6] et l'analyse géométrique de l'aire de surface composé-protéine peuvent être utilisés pour cette fin.

La dynamique moléculaire [7] permet un traitement souple des complexes formés par la cible et le ligand à température ambiante et est en mesure d'affiner les orientations des ligands afin de trouver des complexes plus stables. Elle permet également la reclassement des molécules sur la base de fonctions d'évaluation plus précise [8].

La plupart des logiciels de docking, notamment les plus populaires comme Autodock, ne requièrent pas de configuration particulière en termes de mémoire vive. A l'inverse, les logiciels de dynamique moléculaire sont très gourmands en mémoire vive et tournent sur des supercalculateurs. Le docking a donc été rapidement identifié comme une application particulièrement adaptée au calcul distribué, notamment sur grille de calcul.

2.2.2 Introduction à la grille de calcul

La grille de calcul consiste en un ensemble de ressources informatiques réparties géographiquement qui sont utilisées pour atteindre un but commun. La grille est donc un

système distribué. Les grilles sont souvent construites à l'aide de bibliothèques de logiciels à usage général appelées communément intergiciels en français ou middleware en anglais. Voici une liste des quelques termes importants qui seront répétés tout au long de cette thèse :

Computing Element (CE) : L'élément de calcul est un service qui donne accès à un gestionnaire de ressources local. Sa fonction principale est la gestion des travaux ou jobs (défini ci-après), par exemple : la soumission des jobs, le contrôle des jobs, etc… Le CE peut être utilisé directement par un utilisateur ou à travers le système de gestion des charges ou Workload Management System (WMS, défini ci-après également) qui soumet un job à un CE approprié en fonction des spécificités de ce job.

Certificat et proxy : Afin de s'authentifier auprès des ressources de la grille, un utilisateur doit avoir un certificat numérique délivré par une autorité de certification et reconnu par une Organisation Virtuelle (VO, définie ci-après également). Le certificat utilisateur, dont la clé privée est protégée par un mot de passe, est utilisé pour générer et signer un certificat temporaire, appelé proxy, qui est utilisé pour l'authentification auprès des services de la grille. Parce qu'un proxy est une preuve d'identité, le fichier qui le contient doit être lisible uniquement par l'utilisateur, et un proxy a, par défaut, une durée de vie courte (généralement 12 heures) pour des raisons de sécurité.

Job : Un job est une unité de travail, un programme avec des paramètres et des entrées. Soumis par un utilisateur, il s'exécute sur les ressources de la grille pour produire des résultats de sorties. Une application de grille peut se résumer à un seul job, ou requérir l'exécution de plusieurs jobs similaires (c'est-à-dire un même programme qui est exécuté plusieurs fois avec des paramètres ou entrées différents, comme dans le cas du docking) ou impliquer un enchaînement (flot) de jobs différents qui s'exécutent sur la grille.

Agent-pilote : Un agent-pilote ou job pilote est un job générique soumis par un utilisateur sur la grille dont le rôle est de réserver une ressource de la grille qui sera utilisée ultérieurement par un programme réel. Un job pilote peut être utilisé pour lancer plusieurs tâches (programmes).

Job Description Language (JDL) : le langage de description de job permet de décrire un job avant de le soumettre. Il permet de préciser par exemple l'exécutable à lancer et ses paramètres, les fichiers à déplacer vers et à partir du WN (Worker Node, défini ci-après) sur

lequel le job est exécuté, les fichiers d'entrée nécessaires sur la grille, et toutes les autres exigences du CE et du WN.

Storage Element (SE): un élément de stockage fournit un accès uniforme à des ressources de stockage de données. L'élément de stockage peut contrôler des serveurs de disques simples, des piles de disques de grande taille ou d'autres moyens de stockage. Le SE peut prendre en charge différents protocoles d'accès aux données.

Tâche : une tâche est une unité de travail, un programme avec des paramètres et des entrées qui s'exécutent sur la grille. Dans cette thèse, nous différencions les tâches des jobs de la façon suivante : une tâche n'est pas soumise directement par l'utilisateur mais est lancée par un job pilote. Une tâche est donc un travail qu'un job pilote peut exécuter au cours de son existence sur la grille.

Worker Node (WN) : c'est la machine où les jobs sont exécutés. Un Computing Element s'appuie sur un ensemble de WNs pour exécuter les jobs.

Workload Management System (WMS) : Le but du système de gestion de charge est d'accepter des jobs utilisateurs, de les assigner au CE le plus approprié, d'enregistrer leur statut et de récupérer leur résultat. L'utilisateur peut soumettre un job à un CE directement mais en général, il laisse le WMS choisir le CE le plus approprié à partir de la description de ce job.

Virtual Organisation (VO) : une organisation virtuelle se réfère à un ensemble dynamique d'individus ou d'institutions définis autour d'un ensemble de règles et de conditions de partage de ressources. L'appartenance à une VO accorde des privilèges spécifiques à un utilisateur dans l'accès aux ressources. Pour devenir membre d'une VO, un utilisateur doit se conformer aux règles de l'utilisation de la VO.

Virtual Organisation Membership Service (VOMS) : VOMS est un système de gestion des données d'autorisation d'accès aux ressources de la grille. VOMS fournit une base de données de rôles et de fonctions d'utilisateur et un ensemble d'outils pour l'accès et la manipulation. VOMS utilise le contenu de cette base de données pour générer des certificats de grille pour les utilisateurs lorsque cela est nécessaire.

2.2.3 Les grilles pour les sciences du vivant en Europe

Les sciences du vivant sont un très vaste domaine scientifique regroupant de nombreuses disciplines. Parmi celles-ci, la bioinformatique, la biologie structurale et l'imagerie médicale ont été pionnières dans l'utilisation des grilles. La bioinformatique constitue l'ensemble des méthodes informatiques pour gérer, organiser et analyser les données biologiques. La biologie structurale est la branche de la biologie qui étudie la structure et l'organisation spatiale des macromolécules biologiques. Un de ses champs d'application privilégiés est la recherche de nouveaux médicaments *in silico* comme nous l'avons vu dans les paragraphes précédents. L'imagerie médicale regroupe les moyens d'acquisition et de restitution d'images du corps humain à partir de différents phénomènes.

Dès le démarrage des projets européens d'infrastructure de grilles, bioinformatique et imagerie médicale ont constitué des domaines d'application [9] privilégiés. Mais l'adoption des grilles à grande échelle dans ces champs disciplinaires s'est heurtée à plusieurs obstacles. Les besoins spécifiques de l'imagerie médicale en matière de sécurité et de la bioinformatique en termes de flexibilité et de gestion des données n'ont pas été satisfaits par les intergiciels de grille déployés sur les infrastructures européennes et l'utilisation de la grille reste encore limitée par rapport aux besoins en croissance exponentielle. Malgré cela, les sciences du vivant constituent la deuxième plus grosse communauté d'utilisateurs de l'infrastructure européenne EGI comme nous allons le voir plus en détails dans le paragraphe suivant.

L'INFRASTRUCTURE DE GRILLE EUROPEENNE EGI

EGI (European Grid Initiative) est une infrastructure européenne distribuée qui a pris la suite des projets européens DataGrid et EGEE (Enabling Grids for E-SciencE). Les pays partenaires d'EGI participent à sa coordination par le biais d'Initiatives de Grilles Nationales (National Grid Initiatives ou NGIs en anglais), organisations mises en place pour coordonner et animer les efforts dans l'opération et l'utilisation de ressources informatiques de grille et de cloud distribuées au niveau national. Elles constituent des points de contact et des structures d'animation pour les communautés de recherche et les centres de ressources. L'Initiative de Grille Européenne EGI fédère les efforts des Initiatives de Grilles Nationales pour fournir une

infrastructure intégrée au niveau européen et ouverte à toutes les disciplines. Les principaux services offerts par EGI à travers les NGIs sont les suivants:

- Services d´opération et de sécurité : opération d'une infrastructure de calcul et de données dans le monde entier, fiable et hautement disponible 24 heures sur 24, 7 jours sur 7
- Services aux utilisateurs: support aux utilisateurs, formations, etc.
- Services de liaisons : connexion avec les autres infrastructures non-européennes

Le soutien de la Commission Européenne à EGI se matérialise par un co-financement du projet EGI-InSPIRE (Integrated Sustainable Pan-European Infrastructure for Researchers in Europe), effort de collaboration impliquant plus de 50 institutions dans plus de 40 pays débuté le 1er mai 2010 et prolongé jusqu'au 31 Décembre 2014. Son objectif est de mettre en place une infrastructure de grille européenne durable (EGI - European Grid Infrastructure). Selon le rapport annuel de EGI pour l'année 2013[1], le nombre total de cœurs fournis par EGI a atteint 347.307, tandis que le nombre total de ressources de calcul des infrastructures intégrées dans EGI est de 373.235, en augmentation de 30,7% depuis l'année précédente. La capacité totale de stockage est de 177 Petaoctets, en augmentation de 25,36% depuis l'an dernier.

La gestion des ressources de l'infrastructure de EGI est faite par plusieurs intergiciels de grille. Les intergiciels constituent le lien qui tient ensemble les éléments de la grille. Ils fournissent des services généraux pour l'ensemble de l'infrastructure : gestion de la charge, gestion de données, authentification, surveillance, etc. Les intergiciels déployés sur EGI doivent respecter le cahier des charges de l' « Unified Middleware Distribution » (UMD). Les quatre intergiciels aujourd'hui déployés sur les infrastructures d'EGI sont : gLite [10], UNICORE [11], Globus Toolkit [12] et ARC [13].

[1] Annual Report 2013 : : https://documents.egi.eu/public/ShowDocument?docid=1664

Discipline	May 12–April 13		May 11 – April 12		Jobs	CPU wall time
	% CPU n. wall time (A)	% of Jobs done (B)	% CPU n. wall time (C)	% of jobs done (D)	(yearly increase from May 11) (E)	(yearly increase from May 11) (F)
High-Energy Physics	93.78	89.58	93.60	91.58	**+1.22%**	**+40.97%**
Infrastructure	0.10	2.88	0.20	3.26	-8.70%	-29,67%
Life Sciences	1.52	4.34	1.30	1.75	**+156.79%**	**+65.12%**
Astrophysics	2.82	1.82	2.25	1.58	**+18.57%**	**+76.64%**
Multidisciplinary	0.12	0.17	0.39	0.48	-62.77%	-56.97%
Others Disciplines	0.59	0.45	1.23	0.72	-36.713%	-32.12%
Unknown Discipline	0.43	0.27	0.20	0.29	**-3.08%**	**+199.45%**
Comput. Chemistry	0.48	0.22	0.38	0.03	**+83.04%**	**+78.31%**
Fusion	0.01	0.10	0.37	0.13	-24.56%	-96,98%
Earth Sciences	0.15	0.11	0.10	0.05	**+139.95%**	**+123.45%**
CS and Mathematics	0.00	0.07	0.00	0.03	+170.56%	-68.06%

Table 2-2: L'usage des ressources par discipline

Le Table 2-2 présente l'usage annuel des ressources par disciplines de Mai 2011 à Avril 2013. Les sciences du vivant représentent 1.52% de l'usage de CPU, en troisième position après la physique des hautes énergies et l'astrophysique. Le nombre de jobs soumis représente 4.34% du nombre total de jobs sur la grille EGI, en deuxième position après la physique des hautes énergies.

En termes de nombre d'utilisateurs, la communauté des sciences du vivant est la plus représentée après la physique des particules au niveau national (Figure 2-4).

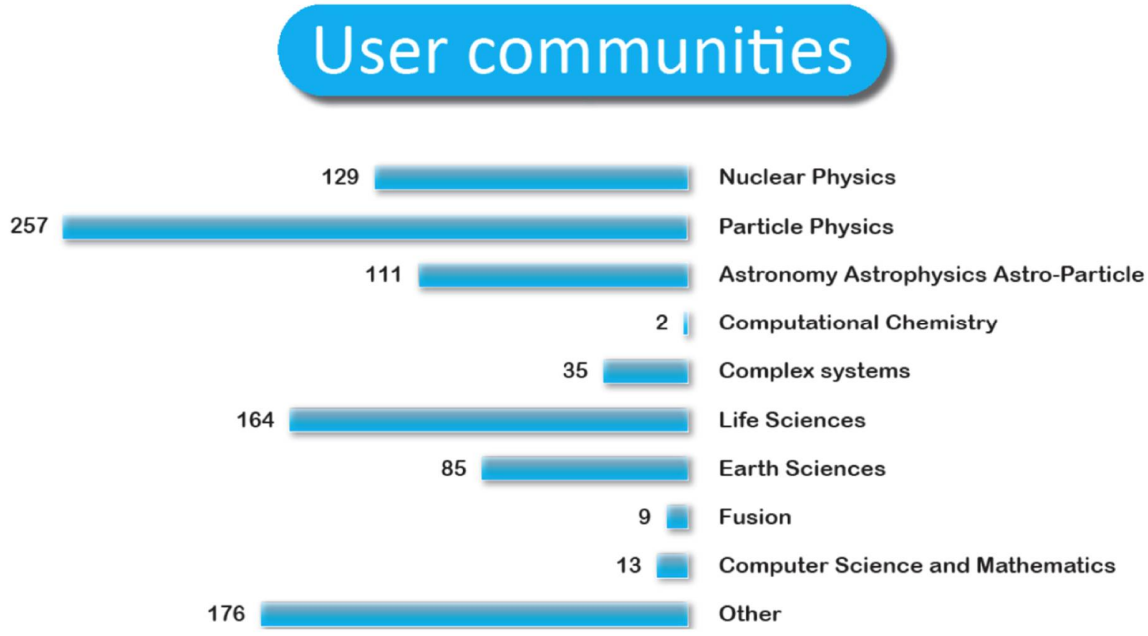

Figure 2-4: Distribution thématique des utilisateurs titulaires d'un certificat émis par France Grilles

ORGANISATION DES SCIENCES DU VIVANT SUR LA GRILLE EGI

Les Organisations virtuelles (Virtual Organization ou VO) sont des groupes de chercheurs ayant des intérêts et des exigences scientifiques similaires, qui souhaitent travailler en collaboration et/ou partager des ressources (par exemple, les données, les logiciels, l'expertise, CPU, espace de stockage), indépendamment de leur emplacement géographique. Les chercheurs doivent adhérer à une VO afin d'utiliser les ressources de l'infrastructure EGI. Chaque VO est autorisée à accéder à un sous–ensemble des ressources d'EGI. Le choix des VOs autorisées à accéder à un site donné de la grille relève de l'entière compétence et autorité de l'administrateur du site. Tout utilisateur muni d´un certificat d´authentification délivré par une autorité de certification acceptée au niveau international peut demander à rejoindre une VO. Chaque organisation virtuelle gère sa propre liste de membres, selon les exigences et les objectifs de la VO. EGI fournit les services centraux permettant aux VOs de se déployer sur des ressources internationales et de tirer le meilleur parti de ces ressources.

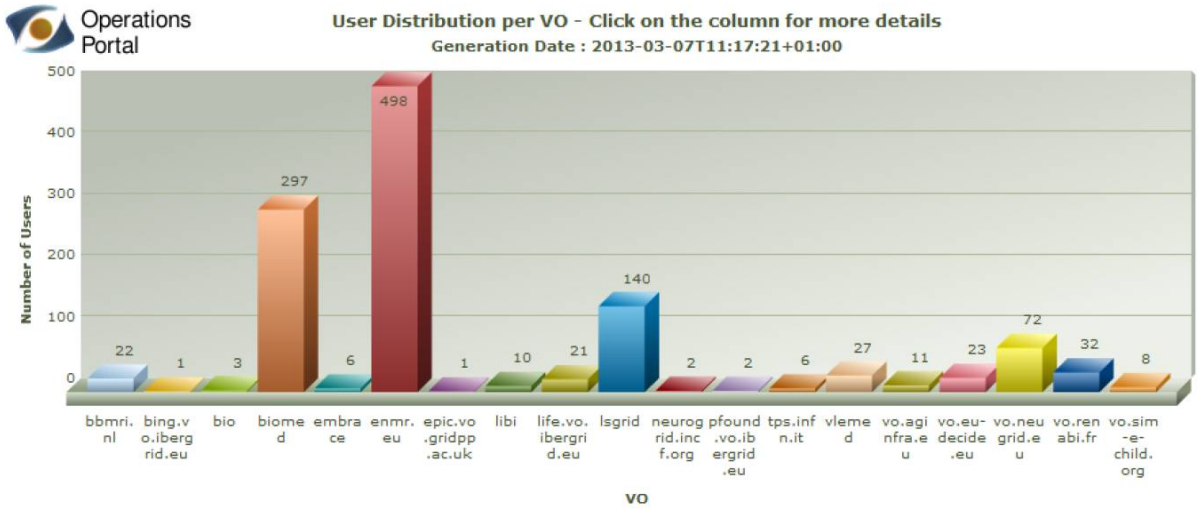

Figure 2-5: Distribution de nombre total d'utilisateurs par organisation virtuelle (VO)

Plusieurs organisations virtuelles (VOs) sont aujourd'hui déployées sur EGI, représentant des communautés d'utilisateurs de tailles variées comme l'illustre la Figure 2-5 qui montre la distribution du nombre total d'utilisateurs par VO du secteur des sciences du vivant. La plus grande par le nombre d'utilisateurs enregistrés est la VO enmr.eu, dédiée à la biologie structurale et plus spécifiquement à la communauté utilisant les techniques de résonance magnétique nucléaire (Nuclear Magnetic Resonance).

Une partie des travaux décrits dans cette thèse a été réalisée sur la VO Biomed qui comptait au 1er Janvier 2014 329 utilisateurs de 20 pays différents. Cette VO a vocation à couvrir de façon très large tous les domaines liés aux sciences du vivant mais la plupart des utilisateurs sont issus de l'imagerie médicale, la bioinformatique et la biologie structurale. Cette VO est librement accessible à des universitaires et à des sociétés privées à des fins non commerciales. La VO Biomed donne accès à des dizaines de milliers de cœurs dans de nombreux sites à travers l'Europe et le monde entier. Ces sites ne sont pour la plupart pas des laboratoires de la discipline.

Face à la multiplication des VOs et pour faciliter la coordination des efforts dans chaque discipline, les Communautés de Recherche Virtuelles (Virtual Research Community - VRC) constituent un mécanisme pour représenter les intérêts d'un domaine de recherche au sein de l'écosystème d'EGI. Ces VRCs peuvent inclure une ou plusieurs organisations virtuelles et ont vocation à être le principal canal de communication entre les chercheurs et EGI. EGI établit des partenariats avec les VRC par le biais d'une convention (Memorandum of Understanding - MoU). Après le processus d'accréditation et l'accord final, les représentants

des VRCs deviennent les interlocuteurs d'EGI dans leur domaine scientifique et profitent des nombreux avantages d'un partenariat solide avec EGI. Ils représentent leur discipline pour négocier des ressources, assurer la liaison avec les Initiatives de Grille Nationales et d'autres fournisseurs de ressources à travers le monde. EGI propose des workshops et des forums pour aider et supporter les communautés sur les problèmes techniques spécifiques, et afin aussi de les impliquer dans l'évolution de l'infrastructure de production d'EGI. Les représentants des VRCs expriment à EGI le cahier des charges de leurs exigences techniques et de service qui est pris en compte dans le développement global de l'infrastructure. Vis-à-vis des communautés qu'ils représentent, ils servent de point de contact pour les nouveaux utilisateurs et apportent de l'aide pour partager les expertises, éviter la réplication des efforts et encourager le partage des ressources, des données et des outils. Ils organisent des événements de formation et fournissent des services pour opérer et supporter les VOs communes, opérer des services partagés, etc. Ils contribuent enfin à diffuser les informations et les connaissances afin de faciliter la communication interne et externe avec d'autres groupes d'intérêt.

La Communauté Virtuelle de Recherche des Sciences de la Vie (Life Science Grid Community - LSGC) rassemble les domaines scientifiques suivants: la bioinformatique, la biologie moléculaire, les bio-banques, l'imagerie médicale, la biologie structurale et les neurosciences.

En résumé, EGI fournit une plate-forme de calcul et de stockage de données pour les communautés scientifiques en Europe. EGI est un environnement où les utilisateurs peuvent se rencontrer et travailler ensemble dans le cadre de collaborations internationales. Aujourd'hui EGI est une infrastructure de ressources en expansion et elle continue à évoluer pour inclure de nouveaux types de ressources et notamment des « nuages » ou « clouds » académiques.

La grille en France

France Grilles[2] est l'Initiative de Grille Nationale (National Grid Initiative) Française établie depuis Juin 2010 sous la forme d'un Groupement d'Intérêt Scientifique entre 8 organismes majeurs de recherche :

[2] Site web de Frace Grilles : http://www.france-grilles.fr/

- Le Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche (MENESR)
- Le CNRS
- Le Commissariat à l'Energie Atomique et aux énergies alternatives (CEA)
- L'Institut National de la Recherche Agronomique (INRA)
- L'Institut National de Recherche en Informatique et en Automatique (INRIA)
- L'Institut National de la Santé et de la Recherche Médicale (INSERM)
- La Conférence des Présidents d'Université (CPU)
- Le REseau NAtional de télécommunications pour la Technologie, l'Enseignement et la Recherche (RENATER)

Les missions de France Grilles sont les suivantes :

- Établir et opérer une infrastructure nationale de grille de production, pour le traitement et le stockage de grandes masses de données scientifiques.
- Contribuer avec les autres états membres impliqués au fonctionnement de l'infrastructure européenne EGI.
- Favoriser les rapprochements et les échanges entre les équipes travaillant sur les grilles de production et les grilles de recherche.

Ces missions s'étendent aujourd'hui au domaine du « cloud computing ».

L'infrastructure de production France Grilles intègre plus de 32.000 processeurs et 31 Petaoctets de stockage répartis dans 18 sites à travers le territoire français pour les analyses de données à haut débit.

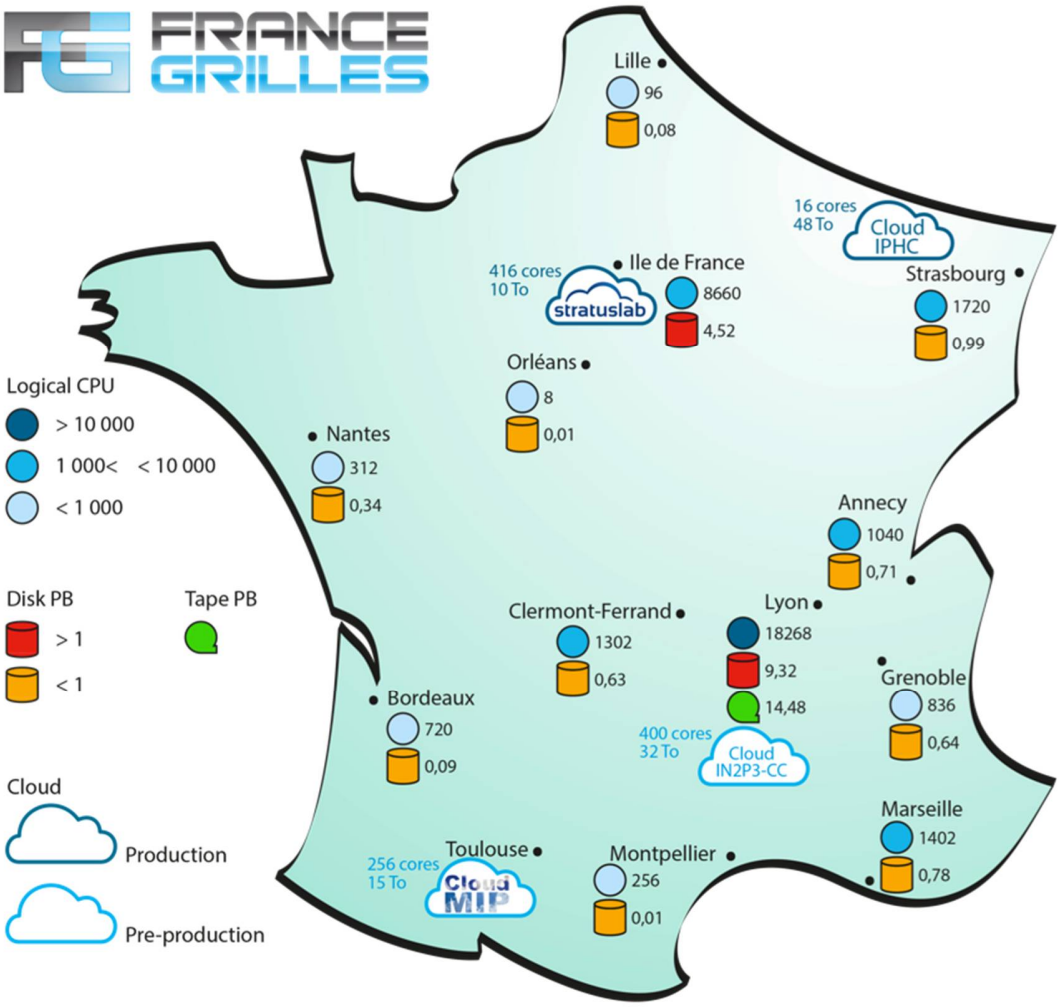

Figure 2-6: Carte des sites fournissant des ressources de grille et de cloud dans France Grilles

Ces ressources sont mises à la disposition des utilisateurs français – il y a aujourd'hui environ 850 titulaires de certificats de grille en France - mais aussi des chercheurs à travers le monde via des organisations virtuelles internationales. France Grilles fait partie des plus gros contributeurs à EGI (Figure 2-7), avec l'Allemagne, l'Italie et le Royaume-Uni.

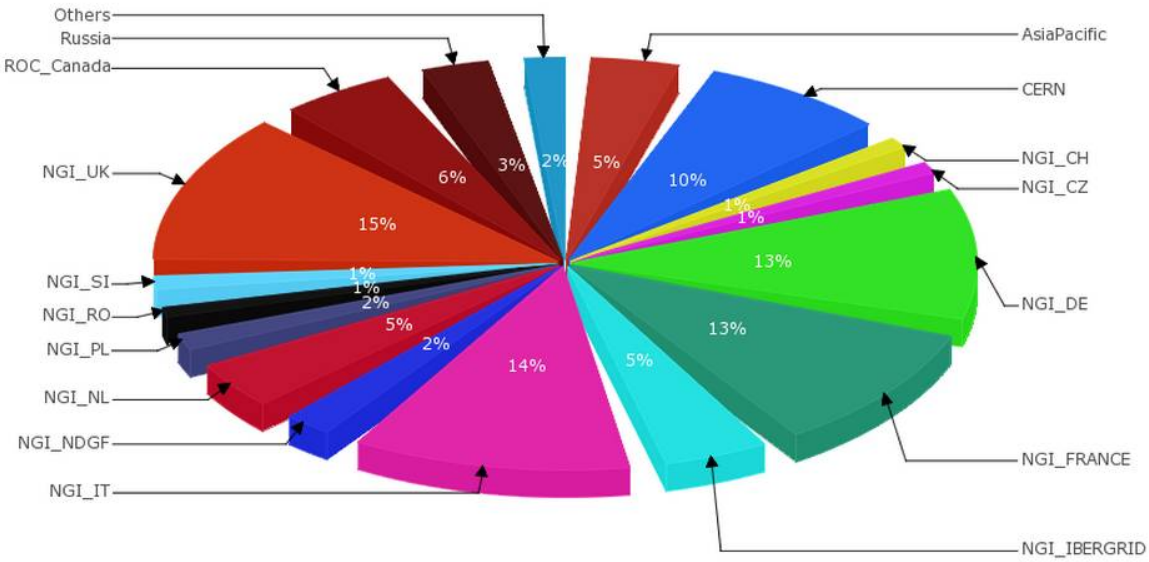

Figure 2-7: Temps de calcul normalisé fourni pendant l'année 2013 par les différentes initiatives de grille nationales

Les communautés de recherche les plus actives en France sur la grille sont les suivantes:

- L'astronomie, l'astrophysique et les astroparticules, principalement pour la simulation Monte-Carlo de nouveaux détecteurs comme CTA (Cerenkov Telescope Array), l'analyse des données issues d'instruments comme AUGER, GLAST ou LSST dans le futur, ou pour la modélisation de systèmes astrophysiques complexes
- La physique des particules, plus gros consommateur de ressources pour l'analyse des données du LHC mais aussi pour d'autres grandes expériences internationales, comme l'expérience D0 à FERMILAB
- Les sciences de la Planète, notamment la sismologie, les sciences de l'atmosphère, l'hydrologie, l'hydrodynamique, l'exploration géophysique et l'étude du climat
- Les sciences du Vivant aujourd'hui très actives sur la grille notamment dans les domaines suivants :
    - la bioinformatique autour du Réseau National des plateformes de Bioinformatique (RENABI), avec des infrastructures dédiées comme Décrypthon.
    - l'imagerie médicale, au sein de la Life Science Grid Community[3]
    - les neurosciences, à travers des projets nationaux et européens

[3] http://www.lsgc.org

Des communautés plus réduites sont très actives dans le domaine des systèmes complexes ou de la chimie tandis que l'observatoire de la grille se propose de collecter et d'analyser les données sur le comportement de la grille[4] pour les chercheurs en informatique. La croissance très rapide des besoins de traitement des données dans de nombreux domaines scientifiques amène régulièrement de nouvelles communautés à frapper à la porte de la grille.

RENABI/GRISBI

ReNaBi (Réseau National des plaques-formes Bioinformatiques - http://www.renabi.fr/) est un réseau de bioinformatique qui a pour objectif de favoriser la coordination de l'activité des nombreuses plates-formes des différents instituts pour mieux répondre aux besoins des équipes de recherche en biologie à l'échelle nationale. Le réseau ReNaBi rassemble 28 plates-formes au sein des centres régionaux suivants: APLIBIO (Paris – Île de France), PRABI (Rhône-Alpes), ReNaBi-GO (Grand Ouest), ReNaBi-GS (Grand Sud), ReNaBi-NE (Nord-Est), ReNaBi-SO (Sud Ouest). Ces plates-formes offrent de nombreuses ressources bioinformatiques pour les communautés scientifiques en sciences de la vie.

Au sein du réseau ReNaBi, l'initiative GRISBI (Grid Support to Bioinformatics) a été lancée pour rassembler les ressources de six plates-formes bioinformatiques françaises dans une grille dédiée pour rendre possibles des applications bioinformatiques à grande échelle: génomique comparative et annotation de génome, biologie des systèmes, prédiction de la fonction des protéines, interaction moléculaire comme protéine-protéine ou d'ADN-protéines. Les six centres de base partagent et mutualisent leurs ressources de stockage et de calcul, mais aussi des ressources comme des bases de données et des logiciels, comme illustré sur la Figure 2-8.

[4] http://www.grid-observatory.org

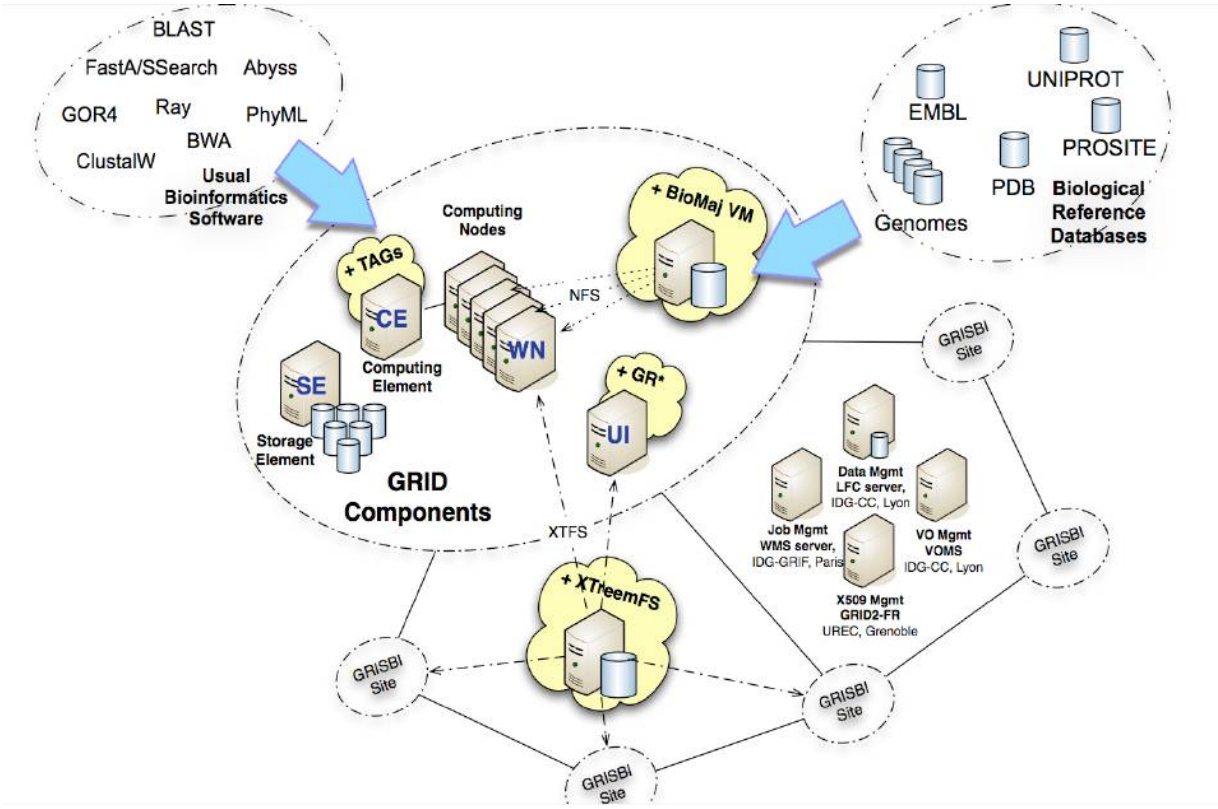

Figure 2-8: Infrastructure GRISBI (Source : http://www.isima.fr/~mephu/FILES/JOBIM12/jobim_actes_2012_clef.pdf)

Cette figure présente l'infrastructure GRISBI avec les composants de gLite standard (en gris) et les composants supplémentaires (en jaune): BioMAJ, XtreemFS [14] et GR * commandes. Différents composants de la grille sont fournis par le middleware gLite [10]: User Interface (UI), Computing Element (CE), Worker Node (WN), Storage Element (SE), Workload Manager System (WMS), Logic File Catalog (LFC), Berkeley Database Information Index (BDII). La plate-forme GRISBI utilise les systèmes de stockage des fichiers de gLite et de XtreemFS [14]. L'outil BioMAJ (http://biomaj.genouest.org/) déployé sur GRISBI permet de synchroniser et de mettre à jour les données sur chaque site. Des logiciels particulièrement utilisés par la communauté de bioinformatique comme ClustalW [15], FastA [16], PhyML [17], sont pré-installés sur les nœuds de la grille GRISBI gère quelques banques de données biologiques : UNIPROT (http://www.uniprot.org/), UNIREF [18] et la Brookhaven Protein Data Base [1]. Pour les services centraux de la grille, GRISBI collabore avec France Grilles en utilisant ses services.

Décrypthon

Fruit d'une collaboration entre le CNRS, l'Association Française contre les Myopathies (AFM) et la société IBM, le programme Décrypthon (http://www.decrypthon.fr) a fourni à partir de 2004 aux équipes de recherche en bioinformatique des ressources de calcul et de stockage pour aider la recherche sur les maladies neuromusculaires et les maladies rares. Il a utilisé pour cela des ressources de calcul distribuées installées par IBM dans six universités françaises (Bordeaux 1, Lille 1, Paris 6 Jussieu, ENS Lyon, Crihan à Rouen et Orsay) et/ou des ordinateurs personnels via le projet World Community Grid (http://www.worldcommunitygrid.org/index.jsp) . Ce projet propose aux internautes de mettre leur ordinateur personnel à la disposition des chercheurs en téléchargeant un logiciel qui permet d'intégrer l'ordinateur à une grille pendant le temps où il est disponible.
Ce sont ainsi des centaines de milliers d'ordinateurs qui ont pu être mobilisés à travers le monde sur des projets scientifiques et notamment la recherche de nouveaux médicaments pour les maladies négligées.

2.2.4 Etat de l'art de la grille au Vietnam

Les acteurs

Le calcul haute performance au Vietnam se concentre dans des centres de calcul dispersés à travers le pays. Ces centres de taille modeste (de quelques dizaines à quelques centaines de cœurs) sont installés dans les universités, les instituts et les organisations et fonctionnent de manière indépendante, pratiquement sans coordination, ni partage des ressources et des services.

On peut citer le Centre de Calcul de Haute Performance de l'Université Polytechnique de Hanoi (HPC-HUST), le Centre de Prévision Hydro-météorologique Nationale, les Centres de Calculs de l'Université des Sciences Naturelles de l'Université Nationale de Hanoi, de l'Université Polytechnique de Hanoï et de l'Université Nationale de Ho Chi Minh-Ville. L'Académie des Sciences et Technologies du Vietnam (VAST) héberge aussi des ressources informatiques à l'Institut de Mathématiques, l'Institut des Technologie de l'Informatique et l'Institut Militaire des Sciences et Techniques.

L'Institut de la Francophonie pour l'Informatique (IFI) est un institut international en informatique créé par l'Agence universitaire de la Francophonie (AUF) en 1995 pour répondre à une demande vietnamienne en formation de cadres supérieurs en informatique et pour aider à l'émergence de professeurs d'informatique vietnamiens de calibre international au niveau universitaire.

Pour le Vietnam, l'IFI représente une aide au développement et à la formation de formateurs et une passerelle vers l'acquisition, l'utilisation et le développement local des techniques avancées d'information et de communication. C'est dans ce contexte que l'IFI s'est intéressé à partir de 2007 à la technologie des grilles de calculs.

Histoire de la grille au vietnam

La première initiative de grille au Vietnam est venue des Etats-Unis à travers le projet PRAGMA (Pacific Rim Application and Grid Middleware Assembly - http://www.pragma-grid.net/) en 2002. Son objectif était d´établir des collaborations et de promouvoir l'utilisation des technologies de grilles entre les communautés de chercheurs des grandes institutions à travers le Pacifique.

La collaboration entre la France et le Vietnam sur la grille a démarré à l'automne 2007 où une première école, baptisée ACGRID, a été organisée dans les locaux de l'Institut des Technologies de l'Informatique (Institute Of Information Technology - IOIT) de l'Académie des Sciences et Technologies du Vietnam (VAST) dans le cadre de la série des écoles Do-Son. Les objectifs de cette école étaient très ambitieux :

- former des administrateurs systèmes vietnamiens à la gestion des services de grille
- démarrer 5 sites de la grille EGEE au Vietnam grâce à du matériel acheté par le CNRS pour la formation
- former des utilisateurs de ces sites pour démarrer le déploiement d'applications scientifiques

L'école n'a pas atteint tous ses objectifs mais 3 sites à Hanoï (Centre de Calcul de Haute Performance de l'Université Polytechnique de Hanoi, Institut de la Francophonie pour l'Informatique et Institut des Technologies de l'Information de la VAST) ont pu effectivement rejoindre la grille dans les mois suivants.

Mais rapidement, l'installation des sites s'est heurtée à des problèmes d'infrastructure réseau. En effet, toutes les connexions entre les sites académiques au Vietnam et en Europe utilisaient le réseau TEIN2 (http://tein2.archive.dante.net/server/show/nav.621.html), projet international financé par la Commission Européenne avec les objectifs suivants:

- Augmenter la connectivité à Internet pour la collaboration dans les domaines de la recherche et l'éducation entre l'Europe et l'Asie
- Améliorer la connectivité intra-régionale en Asie
- Agir comme un catalyseur pour le développement des réseaux de recherche nationaux dans les pays en développement dans la région Asie-Pacifique.

Bien que l'objectif de TEIN2 ait été d'offrir une infrastructure de connexion à haute vitesse qui satisfasse aux conditions requises pour des sites de grille, en pratique, le processus de déploiement des nœuds de la grille au Vietnam s'est heurté à de multiples problèmes d'accès à TEIN2, d'instabilité de la connexion, de capacités limitées de gestion,…

Dans ce contexte difficile, le projet d'Initiative de Grille Nationale VNGrid a été lancé en 2008 avec pour objectif de connecter et de partager les ressources entre les organisations participantes. A cause des difficultés rencontrées, notamment au niveau de l'infrastructure réseau, ce projet n'a pas réussi à poser les fondations d'une infrastructure nationale exploitable et durable. Il a par contre produit des résultats académiques et quelques publications dans le domaine de la recherche et du développement sur cluster de calcul et sur grille.

Malgré ces difficultés, en 2009, l'Institut de la Francophonie pour l'Informatique (IFI) a eu l´opportunité de participer au projet européen EuAsiaGrid (http://www.euasiagrid.org/) et est devenu un des instituts les plus actifs et engagés dans des activités de recherche sur la grille et le cloud. L'IFI a accueilli les écoles ACGRID en 2009 et 2012 et joué un rôle moteur dans l'animation de l'activité scientifique sur la grille et le cloud depuis lors.

Le portail g-INFO de surveillance de la grippe aviaire [19] a été développé dans le cadre du projet EuAsiaGrid par Dr. Doan Trung Tung pendant son travail de thèse. g-INFO est une solution complète pour aider les utilisateurs non experts de la grille à analyser et construire les arbres phylogénétiques sur les séquences de virus influenza. Un deuxième outil de déploiement de logiciel bioinformatique sur la grille est e-Panam, une solution pour le traitement de données métagénomiques issues de séquençage haut débit [19].

D´autres projets de recherche et de développement basés sur la grille et le cloud sont en cours entre l´IFI et les partenaires de la VAST comme le projet sur le criblage virtuel des composants naturels issus de la biodiversité avec l´Institut de Chimie des Produits Naturels (INPC), objet de cette thèse ou le projet entre l'IFI et l´Institut Physique du Globe (IGP) et l´Institut des Technologie de l'information (IOIT) sur la modélisation d'une base de scénarios de tremblement de terre pour le Centre d'Études, de Surveillance et d'Alerte de risque de tsunami dans la mer de l´Est.

2.3 Des grilles au «cloud computing»

2.3.1 Introduction

Bien qu'il y ait un certain flou sur la définition précise, la plupart s'accordent pour définir le Cloud Computing ou informatique en nuage comme un système où les ressources sont exposées comme des services [20]. L'idée de fournir les technologies de l'information comme un service (IT-as-a-service) offre des avantages énormes aux clients qui veulent une puissance de calcul importante sans devoir acheter des ordinateurs coûteux. La capacité à augmenter ou à diminuer la puissance de calcul à la demande en payant seulement la puissance consommée est particulièrement intéressante pour la gestion des pics de charge.

Le cloud computing s'appuie sur l'utilisation d'environnements virtualisés d'exécution (virtualized running environment). De cette façon, via le découplage du matériel, le fournisseur des ressources n'a plus à supporter de multiples exigences des systèmes d'exploitation tandis que le développeur ne dépend plus de la plate-forme du fournisseur, comme c'était le cas pour les grilles. L'utilisateur a donc la possibilité de configurer des environnements de fonctionnement dynamiques sans la médiation du fournisseur de ressources. Un environnement d'exécution encapsulée permet d'éviter une défaillance dans le système d'exploitation des machines virtuelles et de les affecter sur le même hôte.

La technologie de l'informatique en nuage est le résultat d'un processus évolutif qui a commencé il y a 20 ans et dont la grille de calcul a constitué certainement une étape. Elle se situe à la convergence des grilles, de la virtualisation et de l'automatisation. Bien que la

fiabilité et la sécurité des clouds pose encore question, de nombreuses entreprises ont pris les premières mesures en vue de son adoption, soit en utilisant un Software-as-a-Service (SaaS), soit en louant des cycles de calcul à des fournisseurs. Une autre approche dans les grands groupes industriels est le déploiement de clouds internes privés.

2.3.2 Clouds pour les sciences du vivant en France

Le monde académique s'est emparé de cette évolution technologique en parallèle des grands acteurs industriels du secteur. Plusieurs piles logicielles pour l'administration de clouds font l'objet d'un développement collaboratif open source. On peut citer notamment Openstack (Open Source Cloud Computing Software: http://www.openstack.org), OpenNebula (Open Source Data Center Virtualization: http://www.opennebula.org) et StratusLab (http://stratuslab.eu).

France Grilles développe une infrastructure nationale de type IAAS (Infrastructure As A Service) construite sur la base d'une fédération de ressources de calcul et de stockage distribuées sur un ensemble de sites français [21]. Comme illustré sur la Figure 2-6, quatre sites sont aujourd'hui intégrés à cette fédération, à Lyon, Paris, Strasbourg et Toulouse. Trois autres sites sont en cours d'installation à Grenoble, Lille et Montpellier.

Bien que physiquement distribuées, les ressources de cette infrastructure doivent pouvoir être utilisées de manière homogène et transparente. Elles doivent aussi pouvoir faire partie intégrante d'infrastructures plus larges, telles que celle proposée dans le cadre du projet européen EGI-Inspire (INtegrated Sustainable Pan-european Infrastructure for Researchers in Europe). Ce projet coordonne le déploiement d'une offre européenne de type IaaS de calcul et de stockage pour la communauté scientifique. Baptisée EGI FedCloud (EGI Federated Cloud -), cette infrastructure de production est l'agrégation de multiples fournisseurs de ressources Cloud utilisant différents gestionnaires. La fédération de clouds France Grilles doit intégrer cette fédération EGI Fed Cloud dans les prochains mois.

L'utilisation des infrastructures de cloud dans les sciences du vivant n'en est qu'à ses débuts. Quelques groupes ont expérimenté l'utilisation des clouds publics : on peut citer par exemple l'expérience de déploiement du pipeline EOULSAN de traitement de données de séquençage haut débit à travers les solutions Amazon Web Services [22] qui sont flexibles et économiques mais peuvent poser des problèmes en terme de sécurité et de temps de transfert

des données. Les développeurs d'Eoulsan ont déployé l'outil sur la grille EGI avec des performances observées significativement supérieures à celles d'Amazon.

Dans le domaine académique, nous allons présenter plus en détails deux initiatives importantes de mise à disposition de ressources cloud pour les communautés des sciences du vivant : le projet e-biothon [23] et le projet IFB-core. Ces deux projets ont en commun le déploiement d'une ressource cloud à l'IDRIS (Institut du Développement et des Ressources en Informatique Scientifique), un des trois centres de calcul hébergeant des supercalculateurs d'envergure nationale (Tier-1).

E-Biothon

IBM, le CNRS (via l'IDRIS et l'Institut des Grilles et du Cloud avec le soutien de France Grilles), l'InstitutFrançais de Bioinformatique, l'INRIA et SysFera se sont associés pour mettre à disposition des chercheurs la plate-forme E-Biothon qui consiste en deux racks de Blue Gene/P. Ces deux racks offrent une puissance en crête de 28 téraflops et chaque rack compte mille vingt-quatre noeuds, de quatre cœurs chacun. Chaque nœud a une quantité de mémoire RAM partagée de 2 gigaoctets. La Blue Gene est aussi dotée d'une capacité de stockage de 200 téraoctets. Pour accéder à cette puissance de calcul, un serveur dédié est utilisé comme machine frontale. À travers ce portail développé par la société SYSFERA [24], les chercheurs ont accès à tout un environnement de travail leur permettant d'exécuter simplement les traitements informatiques en lien avec les analyses « omiques » à réaliser, puis de gérer les données générées, tout cela à partir d'un simple navigateur web, sans installation locale et avec une sécurité des données garantie. Ainsi, ils peuvent interagir avec une seule interface conviviale plutôt qu'avec des gestionnaires de ressources de Calcul Haute Performance. Le portail s'occupe ainsi d'abstraire et de rendre transparente l'infrastructure de calcul aux utilisateurs afin de leur permettre de se concentrer sur leurs recherches.

La plate-forme E-Biothon est utilisée depuis quelques mois par trois équipes qui y déploient des analyses de phylogénie, d'épidémiologie et de génomique.

Institut français de bioinformatique

L'institut Français de Bioinformatique est une infrastructure de service en bio-informatique issue d¹un appel à propositions « Infrastructures en Biologie et Santé » du programme national des «Investissements d'Avenir».

Ce projet implique des plates-formes de bioinformatique dépendant des cinq principaux organismes publics de recherche, CNRS, INRA, INRIA, CEA et INSERM, des universités et des Instituts Pasteur et Curie. Au total, ce sont une vingtaine de plates-formes bioinformatiques regroupées en six pôles régionaux couvrant le territoire national qui sont ainsi impliquées dans l'IFB.

L'IFB a pour mission principale de fournir les services de base en bioinformatique à la communauté des sciences de la vie dans les domaines de la génomique, de la protéomique, etc. L'IFB est également le nœud français d'ELIXIR, une initiative européenne, menée par l'EBI (European Bioinformatics Institute) qui vise à fournir, au niveau européen, des services similaires à ceux que l'IFB fournit au niveau français.

Au cœur des infrastructures de l'IFB, un cloud académique est en cours de déploiement à l'IDRIS. L'ambition est qu'il fournisse 10.000 cœurs et 1 PetaOctet de stockage pour la gestion et l'analyse des données de sciences du vivant, en complément des 6000 cœurs et 1 PetaOctet de stockage fournis par les plates-formes distribuées sur le territoire national. Après le choix de la pile logicielle pour la gestion du cloud qui s'est porté sur StratusLab, les premiers tests ont démarré au printemps 2014.

2.3.3 Les clouds au Vietnam

Le Vietnam est un marché à forte potentialité pour le cloud computing avec un nombre important d'entreprises qui ont décidé d'utiliser ce type de services [25]. Les organisations au Vietnam adoptent les services de cloud computing pour réduire les coûts, pour accéder à des applications liées à la gouvernance d'entreprise et pour la capacité de test à distance et de développement des logiciels. L'intérêt des entreprises vietnamiennes pour le cloud computing a grandi du fait du volume croissant d'informations auxquelles elles sont confrontées tous les jours. En outre, il y a une tendance croissante au Vietnam de permettre aux employés de travailler à distance (télétravail) avec leurs appareils mobiles personnels. VMWare a conjecturé que le pourcentage d'entreprises au Vietnam utilisant le cloud computing serait plus élevé que les autres pays d'Asie du Sud-Est. Un rapport indique que certains experts estiment que le marché du cloud computing au Vietnam va quadrupler d'ici à 2015 [26]. Cette technologie promet donc d'être utilisée largement dans les entreprises et les instituts de recherche au Vietnam à l'avenir.

2.4 Plates-formes sur grille et cloud pour la recherche de nouveaux médicaments

2.4.1 Introduction

Nous venons de voir dans les paragraphes précédents que des infrastructures distribuées pour le stockage et l'analyse des données scientifiques étaient aujourd'hui disponibles et utilisées par les équipes de recherche en sciences du vivant. Aujourd'hui, les grilles de calcul ont des performances remarquables en termes de fiabilité et de disponibilité tandis que les clouds académiques sont dans une phase de montée en puissance semblable à celle connue par les grilles au début des années 2000. A cette époque, les grilles présentaient des limitations importantes en termes de fiabilité et performances et pour améliorer l'expérience des utilisateurs, l'utilisation de plates-formes s'est développée.

Comme nous l'avons vu aussi précédemment, l'étape de criblage virtuel dans la recherche de nouveaux médicaments *in silico* peut être efficacement accélérée sur la grille en déployant sur un très grand nombre de nœuds les calculs d'ancrage (« docking ») entre le site actif de la cible biologique et les composés chimiques ou ligands. Le calcul de l'énergie de liaison entre le ligand et le site actif utilise des logiciels de chimie computationnelle dont le plus connu est Autodock [27] et requiert typiquement quelques minutes par docking sur un PC traditionnel. Le nombre de ligands recensés dans les chimiothèques pouvant atteindre des millions, un projet de criblage virtuel peut requérir des dizaines de millions de calculs si plusieurs cibles biologiques ou plusieurs configurations des cibles, voire plusieurs jeux de paramètres des logiciels de docking sont testés.

L'utilisation de plates-formes s'est donc imposée naturellement pour les projets de criblage virtuel sur grille. Nous allons présenter de façon détaillée trois plates-formes développées en Europe et en Asie : WISDOM, GVSS et HTCaaS.

2.4.2 WISDOM

La plate-forme WISDOM [28], développée par le Laboratoire de Physique Corpusculaire (LPC) Clermont-Ferrand, France, a été utilisée avec succès dans les projets WISDOM-I (2005) et WISDOM-II (2006) comme une couche entre les utilisateurs et les ressources de la grille. Elle facilite la soumission massive des jobs de docking sur la grille. Baptisée WPE pour WISDOM Production Environment, elle est construite au-dessus de l'intergiciel gLite. L'approche adoptée par la collaboration WISDOM fut de développer un système permettant de centraliser la gestion de plusieurs milliers de tâches soumises simultanément sur la grille.

La **Error! Not a valid bookmark self-reference.** montre l'architecture de WPE. Ce système fut conçu au départ sur une stratégie dite « PULL » : Tout d'abord, le client archive ses fichiers de docking dans un fichier compressé et le sauvegarde sur un élément de stockage (SE) dans la grille. Ensuite, il enregistre la métadonnée de sa tâche de docking dans le Task Manager (TM). Le module Job Manager (JM) soumet des jobs pilotes sur la grille. Les jobs pilotes envoient des requêtes de tâche au Task Manager : en d'autres termes, le job pilote récupère sa tâche à partir du Task Manager. S'il y a des tâches dans la file d'attente du Task Manager, il fournit la métadonnée de tâche au job pilote. Le job pilote télécharge le contenu de la tâche et l'exécute. Lorsque les jobs finissent avec succès, les données sont collectées sur des éléments de stockage de la grille. Si les jobs échouent pour une raison connue ou inconnue, ils sont resoumis. Le module WIS (WISDOM Information System) met à jour constamment l'état des jobs pilotes sur la grille. Cette approche pragmatique a permis de faire le premier déploiement à très grande échelle de calculs et a constitué ainsi le premier déploiement d'un « data challenge » dans un autre domaine autre que la physique des hautes énergies sur la grille EGEE.

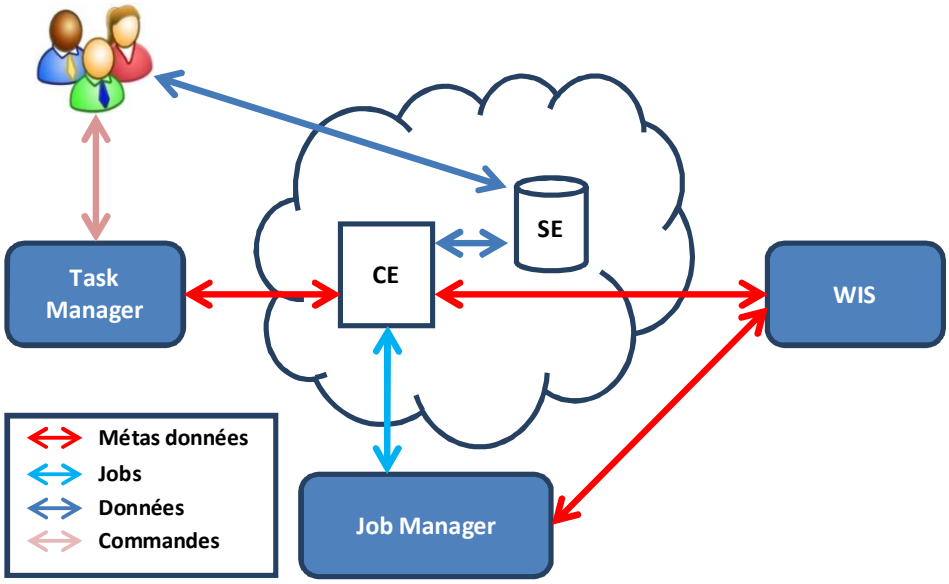

Figure 2-9: L'architecture de WPE

Des déploiements à grande échelle se sont succédés de 2005 à 2011 en ciblant des maladies différentes.

En 2005, le premier déploiement à grande échelle sur l'infrastructure de la grille européenne EGEE, baptisé data challenge WISDOM-I, a utilisé avec succès plus de 80 années de temps CPU en 6 semaines pour une première expérience « *in silico* » [28]–[30]. Il s'agissait de tester environ un million de composés chimiques (médicaments potentiels) avec deux logiciels de docking sur 5 protéines de la famille des plasmepsines, cible biologique potentielle pour combattre la malaria. Cette première étape a permis de sélectionner environ 100.000 composés qui ont fait l'objet d'une deuxième étape de tri [31] par une analyse de dynamique moléculaire. Celle-ci a permis de réduire à 100 le nombre de composés qui ont fait l'objet d'une expérimentation biologique in vitro et finalement in vivo.

En 2006, au moment où les premiers cas de grippe aviaire étaient confirmés en France, une collaboration internationale [32] étudiait l'impact de certaines mutations de la neuraminidase N1 sur l'efficacité des traitements de la grippe H5N1 sur plusieurs milliers de processeurs appartenant à trois infrastructures de grille différentes (Auvergrid-France, EGEE -Europe et TWGrid-Taiwan). Ces études montrèrent que certaines mutations impactaient très significativement l'efficacité du Tamiflu. Elles permirent aussi de tester 300.000 composés chimiques et d'en sélectionner environ 2000 dont les plus prometteurs ont fait l'objet de tests in vitro en Corée du Sud.

En 2007, le data challenge WISDOM-II [33] mobilisait environ 4 siècles de temps CPU pour le criblage virtuel de quatre autres cibles biologiques d'intérêt pour la lutte contre le paludisme, élargissant la recherche à des cibles biologiques du parasite Plasmodium vivax, vecteur d'une forme atténuée de paludisme.

En parallèle de l'analyse des résultats obtenus au cours de ces différentes campagnes de calcul, l'activité dans le cadre de la collaboration WISDOM s'est développée en Corée du Sud où plusieurs centaines de milliers de composés chimiques ont été testés sur des cibles biologiques du SARS [34] et du diabète en 2010 et 2011 [35].

2.4.3 GVSS

Développé à l'Academia Sinica Grid Computing Center (ASGC) de Taïwan, acteur majeur de la grille en Asie, GVSS-1 (Grid-enabled Virtual Screening Services 1 - http://gap.grid.sinica.edu.tw/) est une plate-forme dédiée à la biologie structurale et plus spécifiquement au « docking » pour la recherche de nouveaux médicaments [36]. La plate-forme GVSS-1 offre à l'utilisateur la possibilité de personnaliser la préparation de la base de données ligand et donne accès à plus de 700 protéines de la PDB pour soumettre la simulation de docking. Après la simulation, l'utilisateur peut visualiser les meilleures conformations et des résultats tels que l'histogramme d'énergie et l'analyse en composante principale 2D/3D.

GVSS-1 est basé sur la plate-forme GAP (Grid Application Platform) et utilise l'outil de déploiement DIANE [32] qui déploie des jobs pilotes sur la grille. Les données biochimiques sont stockées en utilisant le catalogue de métadonnées AMGA [37]. L'architecture du système de GVSS est illustrée dans la figure 2-10.

Grâce à DIANE, GVSS-1 peut diviser et exécuter les jobs soumis en plusieurs sous-tâches indépendantes. L'exécution de ces sous-tâches indépendantes tire profit de la disponibilité de la ressource de la grille. GVSS-1 a été utilisé pour la recherche de médicaments *in silico* sur les maladies négligées et émergentes [36].

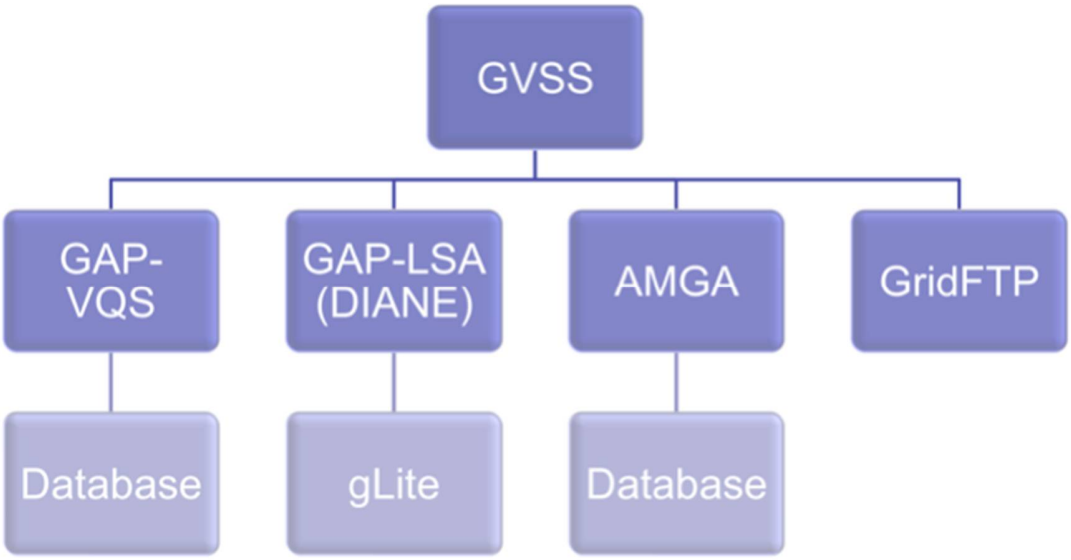

Figure 2-10: Structure de GVSS-1

Après le succès de GVSS-1, ASGC a développé GVSS-2 basé sur la technologie Web [38], [39]. Au lieu de déployer et d'exécuter l'application sur l'ordinateur, ASGC a déployé le système sur un portail web. Un utilisateur muni de son certificat peut participer à la VO EUASIA et peut ainsi accéder à ce portail. Il peut gérer les différentes étapes de préparation du ligand et de la protéine, de déploiement du docking et de traitement des résultats sur le portail, ainsi que de visualisation de ces résultats. Et tout cela se fait à travers une interface web conviviale et facile à utiliser.

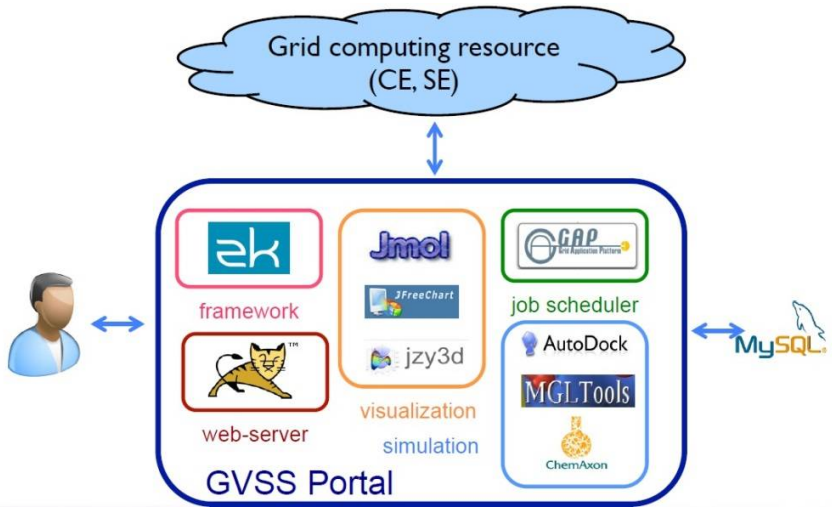

Figure 2-11: Structure du portail GVSS

La Figure 2-11 présente la structure de GVSS-2. Il est basé sur des modules de portail existants (ZK Framework,Tomcat web-server). Les modules de criblage virtuel intègrent des outils de simulation (Autodock, MGL Tools, ChemAxon) et de visualisation (Jmol, jzy3d, JfreeChart).

Bien qu'il soit encore dans un processus de test à petite échelle, GVSS-2 fournit déjà une interface conviviale pour les utilisateurs non spécialistes de l'informatique.

2.4.4 HTCaaS

HTCaaS (High Throughput Computing as a Service) est une plate-forme récemment développée au KISTI pour supporter le déploiement d'applications scientifiques dans le domaine de la pharmacie et de la physique des hautes énergies sur l'infrastructure nationale de calcul intensif en Corée [40]. Elle vise à fournir aux chercheurs des outils pour une exploration complexe et à grande échelle des problèmes scientifiques. HTCaaS permet aux utilisateurs de soumettre efficacement un grand nombre de jobs à la fois par la gestion efficace et l'exploitation de toutes les ressources de calcul disponibles.

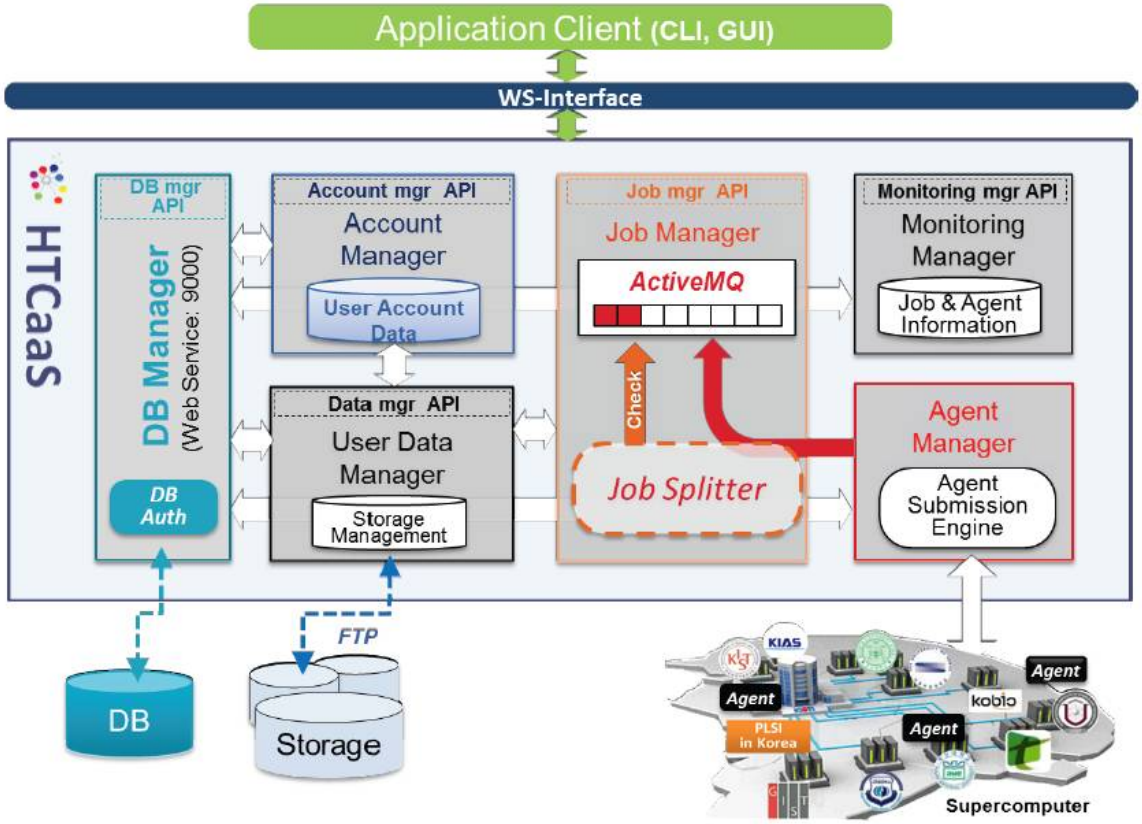

Figure 2-12: La structure de HTCaaS (Source : HTCaaS: Leveraging Distributed Supercomputing Infrastructures for Large-Scale Scientific Computing [41])

La Figure 2-12 montre l'architecture générale du système HTCaaS. Le gestionnaire de comptes (Account Manager) gère les informations de l'utilisateur et fournit des mécanismes d'authentification et d'autorisation intégrés pour accéder à diverses infrastructures informatiques. Le gestionnaire des données des utilisateurs (User Data Manager) est responsable de la gestion des données de l'utilisateur pendant l'exécution des jobs. Le gestionnaire de tâches (Job Manager) effectue l'essentiel de la gestion de jobs. Il maintient les files d'attente séparées par utilisateur, reçoit une métadescription des jobs ou « Meta-job » (écrit en JSDL [42]) qui peut être composée de plusieurs tâches d'un utilisateur, valide le « Meta-job », divise automatiquement le «Meta-job» en de multiples tâches, et contrôle l'exécution de chaque tâche. Le « Monitoring Manager » vérifie périodiquement les exécutions de jobs et des agents actifs en interagissant avec « DB Manager », et si nécessaire, il déclenche des mécanismes de récupération d'échec pour les agents. Une fois que les jobs sont soumis dans le système HTCaaS, les agents (mis en œuvre en Java) sont soumis à partir de l'« Agent Manager » et exécute les tâches sur les ressources de calcul distribuées géographiquement. Une fois déployé, chaque agent appelle activement les tâches, les exécute

et enregistre les statistiques de façon indépendante sur le « DB Manager » qui peut être utilisé pour la surveillance des tâches et des agents par le « Monitoring Manager ».

HTCaaS maintient des files d'attente et gère un pool d’agents propre à chaque utilisateur. Chaque agent tire activement les tâches de sa file d'attente qui correspond à un utilisateur spécifique, et s'il n'y a pas plus de tâches à traiter, il libère automatiquement les ressources et se termine.

2.4.5 DIRAC

DIRAC est défini par ses créateurs [43] comme un intergiciel. La plate-forme DIRAC facilite le calcul scientifique en exposant les ressources de calcul distribué d’une manière transparente aux utilisateurs. Par rapport aux autres plates-formes décrites dans ce chapitre, DIRAC n’a pas été utilisé pour des déploiements à grande échelle de calculs de criblage virtuel. Il est par contre très utilisé en imagerie médicale à travers la plate-forme VIP et plus généralement à travers le serveur DIRAC-FG opéré par France Grilles pour servir de nombreuses communautés d’utilisateurs.

Développée initialement au CPPM à Marseille puis dans la collaboration LHCb au CERN pour satisfaire les besoins liés à la génération d'un grand volume de données de simulation Monte-Carlo, la plate-forme DIRAC (Distributed Infrastructure with Remote Agent Control) est une solution complète pour la communauté d'utilisateurs ayant besoin d'accéder aux ressources de calcul distribuées. DIRAC forme une couche entre une communauté particulière et diverses ressources informatiques pour permettre une utilisation optimisée, transparente et fiable[43]. Il a été le premier système de soumission de jobs pilotes dès le projet DataGrid au début des années 2000.

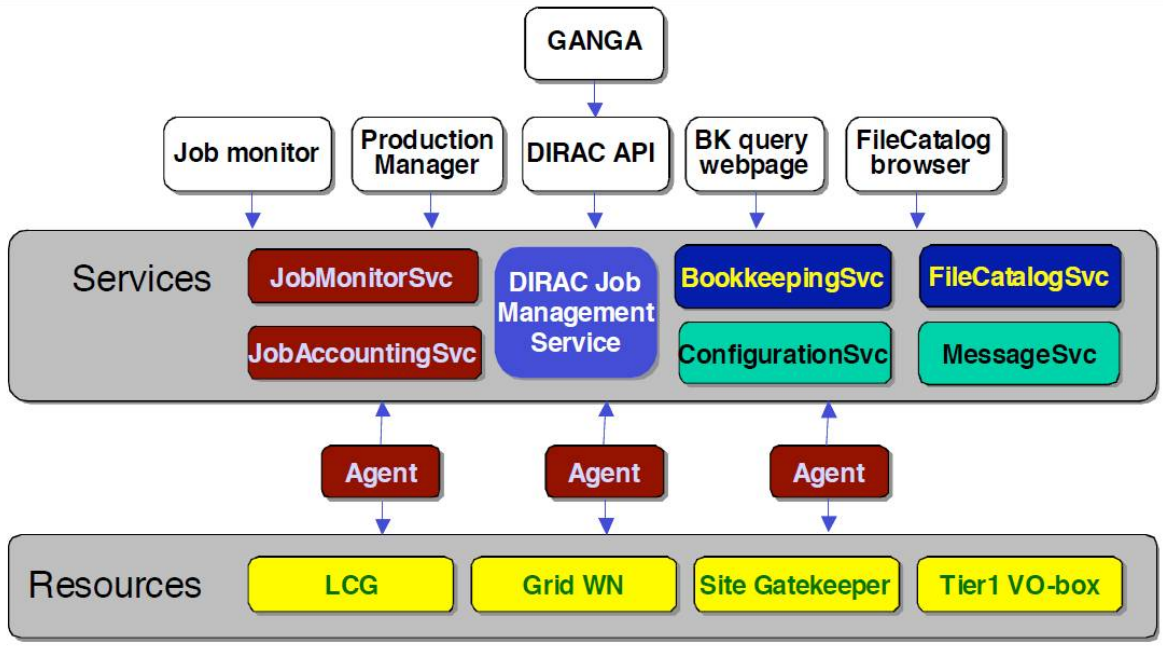

Figure 2-13: L'architecture de la plate-forme DIRAC (Source : [43])

DIRAC suit le paradigme de l'architecture orientée services (SOA). Ses principales composantes sont présentées dans la Figure 2-13. Les composants de DIRAC peuvent être regroupés en quatre groupes (ressources, services, agents et Interfaces):

- les ressources: Les ressources sont des composants qui permettent d'accéder au calcul et au stockage disponibles de DIRAC
- Service: Le système de DIRAC est construit autour d'un ensemble de services faiblement couplés qui maintiennent l'état du système et aident à la gestion de la charge et des données. Les services sont des composants passifs qui ne réagissent aux demandes de leurs clients pour solliciter d'autres services afin d'accomplir les demandes.
- Agent : les agents sont des composants logiciels légers et faciles à déployer qui exécutent des processus indépendants.
- Interface: les fonctionnalités de la plate-forme DIRAC sont exposées aux développeurs du système et aux utilisateurs de différentes façons. Pour les développeurs faisant usage du système de DIRAC, la fonctionnalité est fournie comme une API Python. Pour les utilisateurs du système de DIRAC la fonctionnalité est accessible par une interface en ligne de commande.

DIRAC fournit également des interfaces Web pour les utilisateurs et les gestionnaires du système pour surveiller le comportement du système et pour contrôler les tâches en cours.

DIRAC déploie des agents pilotes sur des nœuds de calcul (Worker Nodes) comme des jobs réguliers en utilisant le mécanisme de soumission des jobs standard de la grille. Il forme un système de gestion de la charge distribuée et réserve les ressources pour une utilisation immédiate.

Le système de production distribué DIRAC fournit les services suivants:

- Définition des tâches de production
- Installation de logiciels sur des sites de production
- Planification et surveillance des jobsGestion et réplication de données

Pour un système de production, le composant en charge de la planification des jobs (job scheduling) est très important. Deux paradigmes différents peuvent être choisis :

Le paradigme PUSH : le planificateur (scheduler) utilise les informations sur la disponibilité et l'état des ressources de calcul afin de trouver le meilleur CE selon les exigences particulières d'un job.

Dans le paradigme PULL, en revanche, c'est la ressource de calcul qui cherche des tâches à exécuter. Les jobs sont d'abord accumulés par un service de production, validés et mis dans une file d'attente. Lorsqu'une ressource de calcul est disponible, elle envoie une requête pour un job à exécuter au service de production. Le service de production choisit un job, selon les capacités de la ressource et le lui envoie pour qu'il puisse s'y exécuter.

Il y a des avantages et des inconvénients dans les deux approches. Dans le paradigme PUSH, l'information sur l'état dynamique de toutes les ressources est généralement collectée dans un seul endroit. Pour chaque job, le meilleur choix possible peut être fait.

Dans le paradigme PULL, la ressource la plus performante est la plus sollicitée. Par conséquent, l'équilibrage de la charge est atteint plus aisément.

DIRAC, comme d'autres systèmes de production de la grille de calcul (DIANE ou WPE) est développé avec le paradigme PULL. L'architecture de DIRAC se compose de services centraux: Production, Monitoring et Bookkeeping. Les agents de production s´exécutent en permanence sur chaque site de production. Les services centraux ont pour rôle de préparer les productions, surveiller l´exécution des jobs et gérer les paramètres des jobs. Les agents examinent l'état des queues de production locales. Si les ressources locales sont capables

d'accepter la charge, l'agent envoie une demande de job au service de production centrale et assure l'exécution et la surveillance de ce job. Après l´exécution du job, l'agent transfère les données de sortie et met à jour le catalogue de métadonnées.

Tous les utilisateurs sont affectés aux groupes de DIRAC et les groupes peuvent être affectés à une Organisation Virtuelle. Les utilisateurs d'un groupe partagent le quota de ressources de leur groupe. Lorsqu'il y a plusieurs lots de tâches à exécuter au sein d'un groupe, la politique de l'ordonnancement entre les utilisateurs d'un groupe est de type Round Robin.

DIRAC est maintenant largement utilisé sur de nombreuses infrastructures dans le monde entier. Son concept permet d'agréger dans un seul système informatique des ressources de nature différente, tels que des grilles de calcul, des clouds ou des clusters, et ce de façon transparente aux utilisateurs.

2.5 Conclusion

Dans ce chapitre, nous avons présenté les enjeux de la recherche de nouveaux médicaments issus de la biodiversité au Vietnam. Nous avons montré comment les infrastructures distribuées de grille constituaient des environnements pertinents pour rechercher *in silico* les cibles biologiques sur lesquelles les composés isolés par l'INPC pouvaient présenter une activité. Nous avons ensuite présenté comment l'utilisation de plates-formes avait permis l'utilisation massive des grilles dans le déploiement de calculs d'ancrage à grande échelle.

L'arrivée de la technologie des clouds apporte de nouvelles opportunités, notamment en termes de flexibilité des systèmes d'exploitation ou de gestion des pics de charge. Mais l'offre des clouds académiques est encore limitée et un long chemin reste à parcourir avant que celle-ci ne soit comparable à celle des grilles. Des efforts importants sont notamment nécessaires pour fédérer les ressources des clouds académiques au niveau national et international.

Dans ce contexte, l'utilisation de plates-formes pour gérer les projets de recherche de nouveaux médicaments semblables à celui de l'INPC demeure une voie privilégiée. Ces plates-formes constituent des couches intermédiaires entre les utilisateurs et les ressources de la grille et des clouds qui permettent de masquer les évolutions technologiques. De plus en plus d'équipes de recherche accèdent aux ressources de la grille à travers de telles plates-

formes et elles sont identifiées comme des outils privilégiés pour exposer les ressources d'une fédération de clouds de façon homogène aux utilisateurs.

L'opération d'une telle plate-forme à l'INPC ne semble pas une approche pertinente car son installation, sa configuration et sa maintenance requièrent des compétences pointues en informatique. Les plates-formes utilisées actuellement en France, en Corée et à Taïwan sont opérées dans des centres de calcul nationaux. Cependant, il peut être envisagé que les chercheurs de l'INPC soient clients de telles plates-formes, mais se pose alors la question de la gestion des clients de ces plates-formes.

Dans le prochain chapitre, nous allons donc nous intéresser à l'ordonnancement des tâches soumises à travers des plates-formes sur grille et cloud.