CONSEIL GENERAL DES TECHNOLOGIES DE L'INFORMATION
139 rue de Bercy
75572 PARIS CEDEX 12

INSPECTION GENERALE DE L'ADMINISTRATION DE L'EDUCATION NATIONALE ET DE LA RECHERCHE
107 rue de Grenelle
75357 PARIS SP 07

Annexe au rapport

sur

La politique française

dans le domaine du calcul scientifique

PROSPECTIVES

Mars 2005

Michel HÉON
Inspecteur général de l'administration
de l'éducation nationale et de la recherche

Emmanuel SARTORIUS
Ingénieur général des Télécommunications

Table des matières

Prospectives

Introduction

Afin de recueillir des éléments de prospective scientifique aptes à éclairer les besoins dans le domaine du calcul intensif, la mission a procédé comme suit, après avoir rencontré les responsables des organismes concernés :

- Elle a demandé au CEA et au CNRS de fournir des prospectives à jour sur le calcul intensif. Une réunion a permis de faire un point à mi-parcours et d'échanger avec les participants au groupe de prospective CNRS réunis par Sylvie Joussaume, Directrice de l'INSU

- Elle a réuni deux groupes ad hoc pour réfléchir sur des domaines horizontaux, en minimisant l'impact des structures verticales *« disciplinaires »* ou liées à la structuration entre organismes. Ces deux groupes ont travaillé sur les sujets suivants :

    1. Le calcul intensif dans les nano-sciences et les nanotechnologies

    2. Le calcul intensif dans le domaine de la biologie.

Etant donné la durée de son intervention, la mission n'a pas tenté d'obtenir une prospective aussi affinée que celle qui serait faite par un conseil scientifique agissant dans le cadre d'une mission permanente définie par le Ministre. Son but à été d'une part de fournir les éléments permettant d'étayer les conclusions et propositions du rapport, d'autre part de se rendre compte de l'intérêt d'institutionnaliser une telle pratique, comme il est fait dans de nombreux pays.

L'objectif de cette consultation n'était pas d'obtenir une évaluation de la demande totale d'heures de calcul mais de repérer les grandes questions scientifiques dont la résolution ne peut être envisagée que si les puissances de calcul sont fortement augmentées. Il a été demandé aux groupes de prospectives de cadrer leurs réflexions dans des scénarios définis a priori par la Direction de la recherche et validés par la mission.

La mission tient à remercier tous les participants pour le travail effectué en préparant les réunions et dans les synthèses. Elle souligne dans son rapport et ses recommandations l'importance de la poursuite de ces activités et de leur intégration dans le dispositif de pilotage scientifique.

Prospective CNRS

Composition du groupe CNRS

Le groupe de travail du CNRS a été animé par Mme Sylvie Joussaume, directrice de l'INSU. Il était composé de :

• Stéphane Cordier	Sciences physiques et de la matière
• François Etienne	IN2P3
• Patrick Le Quéré	Sciences pour l'Ingénieur
• Stratis Manoussis	INSU
• Patrick Mascart	Sciences de l'Univers
• Olivier Pène	Sciences physiques et de la matière
• Serge Petiton	Sciences et technologies de l'information et de la communication
• Marie-Madeleine Rohmer	Sciences chimiques
• Jean-Claude Thierry	Sciences de la vie

Synthèse commune CNRS et CEA/DSM

Suite à la demande des inspecteurs M. Héon et E. Sartorius, le comité d'études prospectives sur le calcul intensif du CNRS, qui vient de se mettre en place, a conduit une première analyse des besoins en matière de calcul intensif. Il ne s'agit pas là d'une étude exhaustive des besoins mais de l'expression des besoins les plus dimensionnants. Ce travail a été mené par le CNRS avec le CEA/DSM pour les applications communes entre les deux organismes.

Le calcul intensif est devenu un outil indispensable pour nombre de domaines permettant de répondre à des enjeux de société, comme le changement climatique, le développement de l'océanographie opérationnelle ou la santé, ou à des enjeux industriels comme la combustion turbulente ou la recherche sur les médicaments. Il permet également d'aborder des questions de recherche fondamentale indispensables au développement de la connaissance comme l'astrophysique, la physique des particules et de la matière.

L'analyse par domaine fait ressortir des besoins d'augmentation en matière de calcul intensif au niveau national, exploitables immédiatement par la communauté scientifique et qui permettraient de suivre le niveau des moyens disponibles par nos partenaires proches voisins en Allemagne et au Royaume-Uni. Elle montre également, pour quelques domaines, le besoin d'avoir accès à des machines de puissance supérieure, qui pourraient être partagées au niveau européen, et permettraient de collaborer avec les USA et le Japon grâce à l'accès à des moyens de calcul comparables.

La recherche, qui s'appuie sur le calcul scientifique, exprime également le besoin d'avoir accès à une panoplie de moyens de calcul : du mésocentre, aux centres nationaux voir à quelques machines au niveau européen, afin de prendre en compte les besoins qui varient selon les domaines et selon les applications.

Le comité a fait également ressortir le rôle clé joué par les mathématiques appliquées et les recherches informatiques avec la nécessité de développer une approche interdisciplinaire combinant chercheurs des applications, mathématiciens appliqués et informaticiens. Plusieurs domaines ressentent également un déficit en soutien informatique de type ingénieur que ce soit de manière ponctuelle pour développer la parallélisation de codes ou de manière soutenue au plus près des applications selon le degré de maturité des développements des codes et l'utilisation par une large communauté.

Besoins les plus dimensionnants issus de différents domaines scientifiques :

Recherches sur le climat (CNRS & CEA/DSM)

Prévoir l'évolution future du climat et répondre aux questions posées à la société sur le réchauffement climatique suite à l'augmentation des gaz à effet de serre est entièrement dépendant de l'accès à des moyens de calcul intensif. Réduire les incertitudes, quantifier la probabilité d'événements extrêmes, quantifier les puits de carbone et leur possible évolution suite au changement du climat, étudier les impacts sur les écosystèmes et la société passe par l'augmentation de la puissance de calcul par un facteur de 10 à 1 000 par rapport aux moyens actuels afin de pouvoir intégrer la complexité du système, augmenter la résolution des modèles, affiner la représentation de processus clés comme les nuages ou les aérosols.

Ce domaine nécessite à la fois une augmentation de moyens au niveau national (5 teraflops soutenus) et l'accès à une machine européenne (50 teraflops soutenus vers 2007-2008) pour certaines expériences numériques les plus poussées en résolution et en complexité. Un projet de calculateur européen pour le climat est soutenu par la communauté scientifique qui pourrait être ouvert à une communauté plus large ayant des besoins similaires. Il peut s'intégrer dans le projet de calculateur européen proposé par l'Allemagne, la France et le Royaume-Uni, si les spécificités de la communauté climat sont bien prises en compte.

Recherches en océanographie et en sciences de la Terre

L'océanographie, en accompagnement du développement de l'océanographie opérationnelle, requiert également une grande capacité de calcul. Comprendre comment fonctionne la variabilité de l'océan, les interactions d'échelle, l'action des tourbillons dans le transport de chaleur et dans les cycles biologiques, nécessite une augmentation de calcul par des facteurs de 10 à 100 et bénéficierait d'une augmentation de calcul au niveau national voire européen.

La modélisation de la dynamique interne de la Terre, la modélisation de la réponse sismique, et des effets de site, lors d'un tremblement de Terre, la modélisation de la rupture sismique et de la génération des ondes courtes associées importantes pour le génie parasismique sont également dépendants des moyens de calcul, même si cette communauté exprime le besoin d'une hiérarchie de moyens de calcul, en particulier la nécessité d'un niveau de type mésocentre.

Recherches en astrophysique (CNRS & CEA/DSM)

La simulation numérique en astrophysique est en pleine expansion grâce au développement des codes et de la puissance de calcul. Les enjeux concernent la compréhension de l'univers qui nous entoure et l'exploitation des données des télescopes sol et des mesures depuis l'espace. Accéder à des moyens de calcul plus puissants par des facteurs 10 à 1000 permettrait de simuler la formation des galaxies, des amas de galaxies, simuler des cartes virtuelles de l'univers afin d'interpréter les grandes expériences de cosmologie comme le futur satellite Planck, de comprendre la dynamique interne du Soleil, la variabilité solaire, les mécanismes physiques de l'explosion des supernovae ou la formation des étoiles.

Recherches en ingénierie

En ingénierie, la simulation numérique en ingénierie est un outil indispensable pour aborder des enjeux industriels dans le domaine des transports, de l'énergie et du développement durable. Réduire la consommation des moteurs, réduire les émissions de gaz à effet de serre, réduire le bruit sont autant d'enjeux sociétaux et industriels qui nécessitent de disposer de capacités prédictives améliorées qui reposent sur la modélisation et simulation numérique intensive. Simulation de la turbulence, des réactions chimiques dans des fluides, des transferts dans des milieux hétérogènes, les interactions entre fluides et matériaux, la réponse des matériaux à différentes contraintes, le développement de nouveaux matériaux s'appuient sur la simulation numérique au sein de projets alliant équipes de recherche en interaction étroite avec les industriels concernés.

Ces domaines d'application sont actuellement limités par l'accès aux moyens de calcul disponibles. Prendre en compte la complexité des phénomènes, celle additionnelle due à leurs couplages, la gamme des échelles spatiales et temporelles impliquées, le grand nombre de réalisations pour tester différents cas et permettre leur optimisation ou leur contrôle, nécessitent une augmentation des moyens de calcul disponibles que ce soit au niveau national et au niveau européen. Pour prendre un des grands défis scientifiques de l'ingénierie scientifique, le problème de la combustion turbulente dans des moteurs, une augmentation par un facteur 10 à 100 permettrait une meilleure description de la turbulence et ainsi d'identifier des zones de gaz frais et de gaz chauds dont le comportement peut être notoirement différent que ce soit en terme de turbulence, d'émissions radiatives ou de formation de polluants.

Recherches en physique fondamentale (CNRS & CEA/DSM)

Comprendre les lois fondamentales de la physique, la composition de la matière, des particules à la composition du cosmos, requiert l'accès à des moyens de calcul. Deux domaines sont

particulièrement demandeurs : la physique des particules en accompagnement des expériences de physique des particules et la chromodynamique quantique.

En physique des particules, les besoins dimensionnants en matière de calcul concernent l'accompagnement des expériences du *Large Hadron Collider* avec des enjeux comme la validation du modèle standard, la prospection de particules supersymétriques, la compréhension de l'absence d'antimatière dans le cosmos ou même l'immense quête de la supposée matière noire galactique. Cette communauté a des besoins très spécifiques liés à la très grande masse de données à traiter qui se compte en petaoctets. Afin d'exploiter au mieux toutes ces données, ils développent une infrastructure répartie au niveau mondial et contribuent fortement au développement de grilles de calcul et grilles de données qui répondent parfaitement à leurs besoins. Ces développements, au départ pour la physique fondamentale, devraient à terme pouvoir servir à de nombreuses autres applications en particulier pour la société et l'industrie.

La chromodynamique quantique s'attache à comprendre l'interaction forte qui lie ensemble quarks et gluons au sein des protons et neutrons. La France a participé à un projet de R & D de l'ordinateur européen apeNEXT qui est adapté aux besoins de cette discipline. L'enjeu pour cette communauté de plus petite taille que la physique des particules serait de pouvoir contribuer à l'implémentation d'une telle machine au niveau européen de 10 teraflops soutenus dans l'immédiat, 1 petaflop vers 2010. Ces développements devraient avoir des retombées dans l'industrie des ordinateurs.

Recherches en chimie

La chimie est un des domaines utilisateur du calcul intensif que ce soit pour des applications industrielles ou des enjeux de société comme l'industrie pharmaceutique. A titre d'exemple dimensionnant on peut citer l'utilisation des méthodes de la chimie théorique pour étudier les structures, les propriétés et la réactivité des complexes de métaux de transition. Il s'agit de comprendre la spécificité des réactions chimiques et le rôle joué dans une structure et dans une réaction par tous les groupes chimiques constituants les molécules. Ces méthodes ont des applications dans la catalyse ou dans la diffusion des ions lithium dans des batteries. Ils nécessitent de représenter des systèmes de grande taille avec des calculs numériques itératifs très demandeurs en temps de calcul.

Dans le domaine des médicaments, un autre exemple est cité ci-dessous avec les applications biologiques : celui des nanostructures biologiques à l'interface entre chimie et biologie.

Recherches en biologie

La biologie est un domaine en émergence dans le domaine du calcul scientifique. Les besoins concernent tout aussi bien la fabrication de médicaments, avec en particulier l'ingénierie des ligands, la compréhension de mécanismes réactionnels, le repliement de protéines, la simulation de la complexité du vivant, comme les cellules ou les écosystèmes. Un besoin particulier d'accès interactif à de grandes bases de données en réseau est un également un enjeu majeur pour la biologie.

A titre d'exemple dimensionnant on peut citer le domaine à l'interface entre la biologie et la chimie : la simulation de nanostructures biologiques. Celle-ci permet d'étudier l'assemblage et la transformation de macromolécules biologiques, des réactions enzymologiques, à la clé de la fabrication de médicaments ou d'étude de mécanismes biologiques. Le repliement inverse des protéines est également un domaine qui requiert une augmentation de puissance de calcul. L'évolution rapide des besoins de calcul intensif dans ce domaine mérite une analyse approfondie.

Recherches en mathématiques appliquées et en informatique

Recherches en mathématiques appliquées et en informatique jouent un rôle particulier vis-à-vis de l'utilisation du calcul intensif.

Mathématiciens experts en analyse numérique, en modélisation et en algorithmique jouent un rôle important dans un très grand nombre de projets de calcul scientifique. On peut citer à titre d'exemple l'apport des mathématiques dans le développement de modèles du stockage géologique profond des déchets nucléaires. L'algorithmie joue également un rôle clé qui s'amplifiera dans les années à venir à mesure que les progrès du calcul intensif devront reposer de plus en plus sur l'inventivité algorithmique à mesure que les progrès des processeurs dus simplement à des gravures fines vont se tarir.

De nombreuses recherches en STIC peuvent également avoir un impact sur le calcul scientifique intensif. Certaines de ces recherches se dont principalement en amont : sur les architectures, la compilation, la programmation, d'autres sont induites par les applications dimensionnantes comme la visualisation, l'analyse de données. Enfin certaines peuvent demander une collaboration étroite avec les applications : stabilité numérique, arithmétique, algorithmique, complexité, calcul formel. Si la communauté STIC calcule peu sur les centres nationaux, les recherches concernées ont souvent besoin pour effectuer des mesures précises d'avoir accès aux machines en mode dédié. Il est aussi nécessaire de pouvoir avoir accès à différents types de machines et des plates-formes comme Grille 5000.

Prospective CEA

La prospective CEA est issue pour l'essentiel de l'analyse des besoins en calcul scientifique intensif lancée par le Comité directeur de l'informatique scientifique du CEA à la mi 2004. Répondant à la demande de la mission de disposer d'éléments de prospective selon les disciplines, la prospective pour l'astrophysique, la climatologie et la physique à haute énergie a été élaborée avec le CNRS et incluse dans la synthèse commune CNRS et CEA/DSM. S'agissant des domaines « nano-sciences et nanotechnologies » et « biologie », des spécialistes du CEA ont participé aux deux groupes ad hoc réunis par la mission. Enfin les éléments de prospective au CEA sur le calcul intensif dans le domaine de l'énergie nucléaire (fission et fusion) sont fournis ci-dessous.

Contributions

Besoins de calcul intensif de la Direction de l'Energie Nucléaire contribution de la Direction de l'Energie Nucléaire du CEA.

Introduction

Toutes les disciplines de la physique qui entrent en jeu dans le cycle de vie des installations nucléaires de la conception au démantèlement et au stockage des déchets en passant par le fonctionnement normal ou accidentel utilisent des outils de simulation. Il s'agit principalement des matériaux, de la neutronique, de la thermohydraulique, du combustible, du stockage géologique et de la mécanique.

Ces domaines de simulation ne sont évidemment pas indépendants, et la plupart des phénomènes à simuler nécessitent des couplages entre les différentes disciplines et de plus en plus souvent des couplages d'échelles dans une même discipline.

Trois disciplines dominent aujourd'hui et continueront dominer demain les moyens de calcul centralisé : les **matériaux**, la **neutronique** et la **thermohydraulique**. Les autres disciplines qui nécessitent beaucoup moins de puissance de calcul, relèvent davantage de ressources locales que de grands moyens mutualisés, elles ont donc volontairement été exclues de la présente analyse prospective de besoin.

1.1/ Comportement des matériaux sous irradiation

L'accumulation de défauts d'irradiation peut entraîner une modification de la microstructure des matériaux et avoir des effets sur les propriétés macroscopiques, la tenue mécanique par exemple. Cette problématique concerne :

- la prolongation de la durée de vie des installations nucléaires,
- le stockage géologique,
- L'optimisation des combustibles actuels et la conception de combustibles pour les réacteurs futurs…

Les simulations actuelles représentent une centaine atomes, ce qui est suffisant pour la représentation de **défauts** ou **d'amas de défauts isolés.** Elles ont permis le calcul d'un grand nombre de configurations d'amas de défauts dans le fer. Il faut maintenant :

- modéliser des matériaux plus complexes : Fe(C), Fe(Cu).
- modéliser de défauts plus étendus (doublement du nombre actuel d'atomes - multiplication par 4 des ressources de calcul)
- modéliser des dislocations (quadruplement du nombre actuel d'atomes - multiplication par 16 des ressources de calcul)
- automatiser les couplages 2 échelles pour permettre par exemple : des couplages *ab initio* – dynamique moléculaire (pour l'ajustement des potentiels) ; des couplages dynamique moléculaire - Monte Carlo (pour la description de défaut d'irradiation produit par des cascades).
- à plus long terme généraliser le couplage des trois échelles (*ab initio*, dynamique moléculaire et Monte-Carlo).

1.2/ Physique des réacteurs

Cette problématique concerne en particulier :

- la conception et de développement de cœurs nucléaires nouveaux,
- l'optimisation du fonctionnement des réacteurs actuels,
- la sûreté des installations et le démantèlement…

Parmi les principales évolutions prévisibles qui consommeront de grosses ressources de calcul on peut citer :

- Les couplages neutronique – photonique pour la conception et le développement de futurs réacteurs expérimentaux.
- Les couplages neutronique probabiliste – combustible qui devraient rester limitée jusqu'en 2007 puis croître rapidement dans les années suivantes.
- Les calculs 3D fins de cœur d'un cycle complet de type cellule par cellule
- Les calculs de radioprotection d'une installation complète sans approximations et les calculs d'activation des structures pour des applications de démantèlement.

1.3/ Thermohydraulique

Dans un horizon de 3 à 4 ans les besoins en thermohydraulique vont changer brutalement d'ordre de grandeur par la généralisation de calculs de Simulation Numérique Directe. Ces simulations sont et seront de plus en plus utilisées en support à la modélisation physique des échelles macroscopiques en particulier lorsque le support expérimental n'est plus accessible ou trop onéreux.

2/ Quantification des besoins

La présente analyse est limitée aux besoins en ressources de calcul scalaire. Les besoins vectoriels ne sont pas abordés, l'analyse faite montrant une relative stabilité dans l'utilisation de ce type de ressource.

teraflops crête scalaire	2004	2007-2009	2009-2012
Matériaux	**0,50**	**2,00**	**4,00**
Thermohydraulique	**0,25**	**1,00**	**4,00**
Neutronique	**0,25**	**1,00**	**2,00**
Total	**1,00**	**4,00**	**10,00**

Processeurs utilisés	2004	2007	2009
Matériaux	**30**	**60**	**200**
Thermohydraulique	**30**	**60**	**500**
Neutronique	**20**	**60**	**100**

2/ Positionnement par rapport aux scénarios envisagés

L'analyse présentée ci-dessus montre un quadruplement des besoins de puissance de calcul de la DEN entre fin 2004 et début 2007. Le scénario A ne peut d'évidence pas répondre au besoin de disposer de 4 teraflop/s dès 2007 (la DEN consommerait à elle seule 16% des ressources ce qui ne paraît pas envisageable). A court-moyen terme seul le scénario B permettrait de répondre aux besoins de la DEN. A moyen-long terme les 3 scénarios B, C et D répondent à nos besoins en terme de puissance brute (10 teraflop/s dès 2009). La discrimination devra se faire sur d'autres critères en particulier sur le mode d'utilisation. Deux modes sont envisagés :

- Mode « Usage par une communauté »: Utilisation d'au plus 50% de la configuration par un utilisateur à un instant donné.
- Mode « Grand Challenge »: Utilisation d'au moins 50% de la configuration pendant une durée importante. Seul mode disponible dans les scénarios C et D.

Si les besoins en neutronique relèvent du premier mode d'utilisation, les besoins en matériaux et en thermohydraulique relèvent davantage du mode « grand challenge ».

Conclusion

Dans ce qui précède nous avons limité l'analyse au seul niveau de puissance. D'autres aspects sont cependant très importants pour le dimensionnement de moyens de calcul :

- la taille de la mémoire proche (disponible sur une même carte et partagée par un nombre limité de processeurs). L'expérience montre qu'un équilibre doit être recherché entre la taille de cette mémoire et la puissance du processeur. ***Augmenter d'un facteur x la puissance des processeurs nécessite, pour conserver cet équilibre, d'augmenter la taille de la mémoire proche du même facteur*** ;

- les temps d'accès de la mémoire éloignée (disponible sur d'autres cartes et partagée par un grand nombre de processeurs). Compte tenu des méthodes de parallélisation utilisées il est primordial que ces temps d'accès soient les plus petits possibles. ***Les caractéristiques des réseaux d'interconnexion doivent donc être un élément fort de choix.***

- la disponibilité de moyens de pré et post traitement…

  Des besoins de visualisation 3D, de type stéréoscopique, existent par exemple pour suivre la migration d'atomes ; la visualisation des chemins de migration ; etc. La mise à disposition de ce type d'outils serait un plus incontestable, pour voir et éventuellement agir durant un calcul.

Au-delà d'une simple augmentation de puissance, de mémoire et de débits d'entrées sorties, certains des nouveaux besoins, tout particulièrement les couplages de disciplines et d'échelles, vont imposer des choix différents dans : les méthodes de gestion des ressources ; les outils disponibles sur les machines centrales ; les règles de sécurité ; le calcul conjoint sur moyens locaux et centraux, etc.

Le calcul dans le cadre du projet ITER
contribution de J. P. Génin, CEA

Le défi scientifique et technique posé à la communauté internationale par la construction du réacteur expérimental de fusion thermonucléaire par confinement magnétique ITER demande un travail intense au niveau de la théorie et de la modélisation, pour faire avancer non seulement la compréhension des phénomènes de base dans des domaines de la physique extrêmement variés, mais

également pour intégrer l'ensemble de ces modèles dans des codes de calcul capables de prédire avec une fiabilité suffisante le comportement du plasma et des divers composants du réacteur : les « simulateurs tokamak ». Il s'agit d'assurer à la fois le succès de l'expérience et sa sûreté.

ITER représente un véritable tournant pour la modélisation, qui doit désormais réaliser cette intégration à l'horizon 2015-2020. Or les recherches actuelles montrent que les processus physiques qui régissent, par exemple, le comportement du cœur du plasma impliquent de résoudre les équations cinétiques complètes dépendantes du temps (2 ou 3 dimensions dans l'espace des vitesses dans une géométrie principalement à deux dimensions), couplées aux équations de l'électromagnétisme. Les échelles de temps en jeu couvrent 4 à 5 ordres de grandeurs (magnétohydrodynamique, transport de la chaleur et des particules, évolution du profil de courant plasma). L'intégration de cette physique du cœur du plasma aux phénomènes périphériques dits d'interaction plasma-paroi, eux-mêmes influencés par la physico-chimie locale dans une géométrie à trois dimensions spatiales avec de nouvelles échelles de temps requiert des moyens de calcul centralisé hors de la portée actuelle de la communauté fusion magnétique.

La nécessité du travail est reconnue au niveau mondial depuis deux à trois ans, principalement par les partenaires Etats-Unis (Fusion Simulator Project (FSP), ou initiative SciDAC), et Europe (European Task Force on Integrated Tokamak Modelling). Il faut noter que le Japon a également débuté une réflexion dans ce sens depuis quelques mois. Ces initiatives de modélisation intégrée vont cependant se heurter rapidement au problème des ressources de calcul centralisé. Toute augmentation de la puissance de calcul disponible permet à la fois d'affiner les modèles, mais également de les coupler entre eux de manière à gagner sur la cohérence globale. Le projet américain FSP a quantifié ce que l'utilisation des «supercalculateurs » actuels peut apporter. On peut citer quelques chiffres : au niveau des ressources de calcul pur certaines simulations de magnétohydrodynamique des plasmas en ignition approchent la centaine de teraflops. La résolution de certains problèmes de propagation/absorption d'ondes de chauffage additionnel requiert un nombre d'opérations d'environ un millier de teraflops*heure. La résolution de la dynamique des électrons sur l'échelle du temps de confinement en régime turbulent nécessite au moins 10 000 teraflops*heure par simulation. Il faut également souligner les besoins concomitants en matière de stockage de données et de rapidité des réseaux d'échanges liés aux quantités de données traitées. Le projet FSP souhaite disposer de 4 teraflops maintenant, de 10 dans un à trois ans et de 100 au-delà de trois ans.

En résumé, une simulation typique de MHD ou de turbulence nécessitera 103 teraflops*heure en 2008 et 10 fois plus (104 teraflops*heure) à l'horizon 2012 (pour un démarrage prévu d'ITER en 2015), lorsque la physique complète des électrons sera incluse. L'objectif est d'optimiser l'exploitation d'ITER, i.e. de préparer les expériences au mieux par des simulations les plus détaillées possible de la stabilité, du confinement, du chauffage et de l'extraction de puissance des plasmas de fusion. Un projet scientifique, par exemple la préparation d'un scénario dans ITER, requiert un nombre élevé de simulations (de l'ordre de la centaine). L'estimation est donc de l'ordre de 10^5 teraflops*heure à l'horizon 2008 et 10^6 teraflops*heure à l'horizon 2012 pour chaque projet scientifique.

Prospective INRIA

Voir Annexe.

Prospective Calcul Intensif en Biologie

Composition du groupe

Vincent Breton	CNRS / IN2P3
Christophe Combet	CNRS / IBCP
Gilbert Deleage	CNRS /IBCP
Martin Field	CEA
Christine Gaspin	INRA / LBIA
Nicolas Jacq	CNRS /IN2P3 LPC Clermont Ferrand
Richard Lavery	IBPC Paris
Michael Nilges	Institut Pasteur
François Rodolphe	INRA Jouy-en-Josas
William Saurin	Genomining
Thomas Simonson	Ecole Polytechnique
Jean-Claude Thierry	CNRS SdV / IGBMC Strasbourg

Contributions

Quelques enjeux majeurs en biologie computationnelle
contribution de MM. R. Lavery et T. Simonson

Le problème du repliement inverse

Avec l'explosion de l'information génomique, l'écart entre le nombre de séquences et le nombre de structures de protéines connues continue de se creuser. La simulation moléculaire peut aider à résoudre ce problème grâce aux techniques dites de *repliement inverse.* Il s'agit d'identifier les séquences les plus favorables correspondant à une structure tridimensionnelle donnée. Un enjeu actuel est d'appliquer la méthode aux 3000 repliements protéiques connus, en utilisant par exemple l'évolution dirigée *in silico*, qui a atteint un niveau raisonnable de maturité et continue de progresser rapidement. On peut estimer à 5-10 ans le temps monoprocesseur nécessaire avec les modèles actuels. Pour la prochaine génération de modèles théoriques, aujourd'hui en développement, il faudra augmenter la puissance de calcul par 10. La “cartographie des séquences” obtenue permettra d'avancer dans l'attribution des structures (e. g. dans le contexte de la génomique structurale), l'attribution des fonctions (pour l'annotation des génomes et la protéomique), et l'ingénierie de protéines avec des fonctions nouvelles.

La reconnaissance protéine-ligand et l'ingénierie de ligands : l'apport de simulations quantitatives

L'ingénierie de ligands s'appuie aujourd'hui sur un spectre de méthodes de coûts et fiabilités variables. Dans les sociétés pharmaceutiques, des méthodes simples sont employées, mais également des méthodes assez sophistiquées, et une demande existe pour des méthodes véritablement quantitatives. En recherche fondamentale, des méthodes coûteuses mais potentiellement très fiables sont développées depuis dix ans. Les progrès en puissance de calcul et une expérience grandissante sont en train de porter à maturité ces méthodes “exactes”, dites “calculs d'énergie libre”. Pour augmenter leur fiabilité, il faut pouvoir multiplier les simulations de dynamique moléculaire et aussi explorer de nouveaux degrés de liberté, comme le couplage entre fixation de ligand et fixation/libération de protons. Pour appliquer ces méthodes efficacement, nous devons gagner un ordre de grandeur en puissance de calcul. Pour un seul projet (= un chercheur), il faut une puissance dédiée qui équivaut à une grappe de 32 processeurs Intel de dernière génération (= 300 000 heures CPU au CINES). De plus, il faut pouvoir accéder à des pointes de vitesse 2 ou 3 fois supérieures lors des phases critiques d'un projet.

Assemblage de complexes protéiques

La plupart des processus biologiques impliquent l'assemblage de complexes protéiques. Notre connaissance de la gamme de ses assemblages est en train d'exploser grâce à de nouvelles techniques expérimentales. Ces techniques témoignent de l'existence de milliers d'interactions binaires au sein des cellules, et permettent de dresser les premières cartes de "l'interactome". Ce domaine nécessite néanmoins de nouvelles collaborations entre bioinformaticiens et structuralistes pour traduire des informations binaires en modèles moléculaires, permettant ainsi l'obtention des données structurales, thermodynamiques et cinétiques pour des complexes bi- ou multimoléculaires. Grâce à de tels modèles il devient également possible d'étudier la perturbation des complexes suite aux mutations ou aux interactions avec des agents externes. Ce travail implique des efforts significatifs pour améliorer les techniques existantes pour prédire la conformation des complexes protéine-protéine. Il faut notamment pouvoir tenir compte de la flexibilité intrinsèque des protéines et savoir intégrer des informations fournies par l'étude de l'évolution des séquences. La quantité d'information à traiter dans ce domaine doit stimuler l'exploitation des ressources en calcul distribué.

Contrôle de l'expression génétique

Une meilleure compréhension du contrôle de l'expression génétique reste un enjeu majeur pour la biologie. Les zones de fixation des protéines impliquées dans ce contrôle sont plus difficiles à déceler que les gènes correspondants, notamment à cause des séquences consensus mal définies et d'un manque d'information sur l'ensemble des protéines impliquées. Dans le cas des eucaryotes, il faut ajouter la complexité liée à l'empaquetage du génome et notamment au positionnement des nucléosomes. Les études de modélisation peuvent contribuer à résoudre ces problèmes en permettant d'accéder aux facteurs physiques qui sous-tendent les interactions protéine-ADN. De telles données peuvent être exploitées en parallèle avec des informations provenant de létude des séquences pour améliorer les capacités de détection, mais aussi peuvent servir pour aborder la modélisation de protéines homologues pour lesquelles il n'y a pas d'information sur leurs interactions. Comme dans la

plupart des domaines liés à l'analyse des génomes, une étude de la multiplicité des éléments en interaction passe par l'exploitation des moyens de calcul significatifs, mais aussi par l'intégration des informations dans des modèles à différents niveaux qui nécessiteront des collaborations transdisciplinaires efficaces.

Exploitation des données produites par la biologie contribution de C. Gaspin

Les volumes considérables de données produites par la biologie interpellent les mathématiques et l'informatique autour de leur exploitation. Deux familles de problèmes pourraient bénéficier d'une structure offrant des possibilités en calcul intensif. La première famille de problèmes s'appuie sur des algorithmes dont la complexité est «acceptable », typiquement, des algorithmes linéaires ou de faible complexité polynomiale. En général, le volume considérable des données (le volume des données croit de manière exponentielle) à analyser constitue le frein majeur (voir par exemple 1)). La deuxième famille contient les problèmes pour lesquels les algorithmes sont coûteux voire NP-durs. (voir par exemple 2).

Comparaisons de séquences (1)

Les séquences complètes des premiers génomes séquencés ont donné la possibilité d'accéder à des organisations structurales « statiques » des génomes. Le volume déjà considérable des données a très rapidement imposé, dès les premiers lots de séquences disponibles, l'utilisation à grande échelle d'algorithmes efficaces (souvent des heuristiques) pour la comparaison des séquences entre elles (le logiciel BLAST en est un exemple). C'est donc ici la distribution d'une masse considérable de calculs de comparaisons sur un grand nombre de processeurs qui pourra constituer une avancée et donner un avantage à une équipe à un instant donné (ex : assemblage du génome humain). Les enjeux sont par exemple : une prise de position dans un contexte/programme international d'assemblage/annotation avec à la clef, un accès privilégié aux premières séquences annotées. Des architectures spécialisées existent aussi pour ce type de calculs.

Alignement de séquences et phylogénie (2)

Les premiers algorithmes exacts d'alignement de séquences (programmation dynamique), mais aussi les méthodes d'analyse de données, de maximum de vraisemblances, les modèles markoviens et autres... sont des algorithmes dont la complexité (parfois exponentielle) devient un frein prohibitif dès que la taille du problème croit (ex : longueur d'une séquence, nombre de séquences, d'individus….). La reconstruction phylogénétique constitue une illustration de ce type de problème. Une phylogénie représente l'histoire évolutive d'un groupe de gène ou de taxa. Les algorithmes largement utilisés de nos jours s'appuient sur des modèles basés sur l'évaluation des substitutions (mutations) dans les données de séquences. Ils peuvent se révéler très coûteux selon les méthodes utilisées (méthode de parcimonie et de maximum de vraisemblance) et pourraient bénéficier

pleinement d'un centre ouvert de calcul intensif (cf. [1]). On peut citer parmi les enjeux, hors une meilleure connaissance de l'histoire évolutive, une meilleure compréhension des maladies et de leur mode de propagation. Avec l'accès aux génomes complets et à leur annotation, la reconstruction phylogénétique invente aujourd'hui de nouveaux modèles qui tentent de reconstruire l'histoire des génomes à partir de leur organisation structurale et d'un ensemble d'opérations autorisées (perte, duplication, réarrangement, …) sur les gènes. Les algorithmes restent cependant très combinatoires et pourraient aussi pleinement bénéficier de telles infrastructures. D'autres questions sont aussi concernées : analyse des données d'expressions, recherche de motifs, alignement multiple, cartographie des marqueurs moléculaires... Certaines de ces questions sont déjà ouvertes devraient s'ouvrir à des modèles et méthodes considérant les aspects quantitatifs (analyse des données d'expression notamment – transcriptome mais aussi protéome).

Portails intégrés d'analyses informatiques pour la biologie
contribution de C. Combet, laboratoire PBIL-IBCP

A l'ère de la biologie à grande échelle, la quantité toujours croissante de données biologiques hétérogènes nécessite de plus en plus de traitements informatiques pour le stockage et l'exploitation de celles-ci. Devant le grand nombre de données et d'outils de traitements de celles-ci, il est nécessaire de mettre en place pour les utilisateurs finaux que sont les biologistes des services intégrés, conviviaux et interactifs d'analyses, de modélisation et de simulation des macromolécules biologiques qui les aident dans leurs recherches au quotidien. De tels services devront être accessibles par Internet et proposer une interface simple et convivial aux utilisateurs en intégrant dans un enchaînement logique les outils de traitements des données avec un temps de restitution le plus court possible (ce qui nécessite des moyens de calculs importants). Les problématiques adressées dans la mise en place de tels portails sont l'uniformisation des données, la définition de protocoles d'analyses, la "parallélisation" des calculs, l'intégration et la définition des "interfaces homme-machines". La faisabilité et l'utilité de tels portails ont été démontrées au travers de nombreux portails mis en place par les laboratoires nationaux de bioinformatique sur leurs ressources propres et qui sont plus ou moins limités dans leur capacité pour des raisons de temps de calculs.

Calcul intensif et mésocentre en biologie
contribution de V. Breton et N. Jacq, CNRS/LPC

Note liminaire : *La mission a décidé d'intégrer ici cette contribution qui traite de l'organisation requise de la recherche en biologie et bioinformatique pour prendre en compte l'intervention de plus en plus fondamentale de l'information sous des formes extrêmement massives. Ceci faisait partie des questions qui ont été posées au groupe de travail ad hoc. Toutefois, dans ses*

[1] Article de D.A. Bader, "Computational Biology and High Performance Computing", Comm. ACM, Vol 47, N° 11, Nov 2004

conclusions, la mission a estimé préférable de ne pas aborder cette question, qui a son avis doit faire l'objet d'une étude plus complète de la part le ministère en charge de la Recherche.

Introduction

L'objectif de cette note est de présenter une réflexion sur la mise à disposition des biologistes de ressources de calcul et de stockage des données ainsi que de services bioinformatiques intégrant des compétences indissociables en biologie et en informatique, pour la création et la maintenance d'outils informatiques et de banques de données, le traitement de données pour la génération de connaissance, et la formation.

Afin d'en garantir la pertinence biologique, ces infrastructures seront adossées à un ou plusieurs laboratoires de biologie de visibilité nationale et internationale.

Perspective sur les besoins en calcul scientifique et en traitement de données

Les besoins exprimés par les biologistes peuvent se résumer en quelques points essentiels :

- Disposer ou développer des bases de données généralistes ou thématiques, validées, nécessaires à leur recherche, mises à jour en temps réel dans un environnement convivial,
- Disposer des algorithmes et procédures adaptés à une recherche en constante évolution dans un environnement convivial et les faire évoluer en permanence,
- Disposer de ressources de calcul mais surtout humaines suffisantes pour un travail d'annotation et d'analyse,
- Pouvoir combiner complexité informatique et innovation biologique sous-tendue par les demandes exprimées précédemment, les outils génériques nécessitant de plus en plus une utilisation intégrant la complexité spécifique du domaine.

Pour répondre à ce cahier des charges, les centres européens et nationaux de ressource en bioinformatique ont notamment développé des systèmes de mise à jour et d'accès aux bases de données, des boîtes à outils d'algorithmes et des portails web mettant ces outils à la disposition des biologistes via Internet dans un environnement simple et raisonnablement convivial.

Ces portails se sont avérés des moyens puissants pour permettre l'analyse des données biologiques mais ils ont été confrontés à plusieurs défis :

- la complexité et la diversité (séquences, images, données de puces, CGH arrays, données structurales, données bibliographiques, criblages pour le médicament, …) de l'information biologique qui ont suscité une multiplication des bases de données et des algorithmes,
- Le volume exponentiellement croissant des données produites qui a induit une augmentation tout aussi rapide des besoins de calcul, de stockage et de mise à jour des données,
- La multiplication des biologistes produisant et analysant des données de la génomique et de la post-génomique qui a créé un phénomène de goulot d'étranglement dans l'accès aux portails et aux ressources offertes par ces portails,

- Surtout la non-adaptation de ces sites à la spécificité des requêtes et à l'émergence de problèmes multifactoriels impliquant la combinaison de données de natures complémentaires (séquences et puces par exemple) nécessitant autant de nouveaux outils (clustering), conditions incontournables à l'exploitation des données.

C'est ainsi qu'à côté du centre national français Infobiogen, des portails thématiques (génomique, protéomique, …) ont vu le jour dans plusieurs régions françaises notamment dans le cadre des génopôles. Ces portails régionaux ont permis d'absorber une fraction significative des besoins des biologistes pour certains dans un contexte local, régional et international. Cependant, ces portails régionaux ont leur propre architecture, assurent eux-mêmes la mise à jour des bases de données et la croissance de ces bases induit une pression croissante en ressource humaine pour la maintenance du portail.

Nous observons depuis 4 ans dans la communauté des biologistes une approche très pragmatique : ils vont « à la pêche » sur Internet des sites ou des informations qu'ils peuvent exploiter, s'adaptant presque au jour le jour aux temps de réponses offerts par les différents portails qu'ils connaissent.

La situation actuelle laisse ainsi apparaître d'importantes lacunes :

- la politique nationale est seulement en cours de mise en place pour la coordination de la gestion et de la maintenance de ces portails donnant accès à des ressources bioinformatiques dans l'hexagone. Cette action est actuellement coordonnée et des réunions de travail planifiées en 2005 par F. Soubriet et C. Gautier en coordination avec les organismes dont le CNRS (J.C. Thierry).
- Chaque portail fait sa propre recherche de fond et fait évoluer son offre en calcul ou en données en fonction des financements obtenus, de ses propres thèmes de recherche et de ses ressources humaines.

Dans ce contexte, accumuler des teraflops de puissance de calcul n'aura qu'une utilité limitée si ces teraflops ne sont pas proprement interfacés à des services bioinformatiques accessibles de façon conviviale. La croissance exponentielle des données, la multiplication des utilisateurs de ces données et la diversité des compétences requises rend le modèle centralisé de plus en plus difficile à gérer. Les nouvelles technologies, notamment les grilles, permettent de déporter les calculs les plus gourmands en teraflops vers des centres de calcul spécialisés tels que l'IDRIS, le CINES ou ceux de l'IN2P3. Gérer un centre de calcul est un métier différent de celui de gérer un centre de ressources en bioinformatique, qui doit maintenant comme l'EBI ou le NCBI combiner expériences biologiques et informatiques (algorithmique, mathématiques appliquées,...).

Les clefs d'une politique de dynamisation de la recherche en biologie en France nous apparaissent donc les suivantes à travers une décentralisation de l'offre de calculs :

- Mise en réseau de portails offrant de **façons concertées et complémentaires** des services communs et des services spécifiques,
- **mise en place d'une politique commune de gestion de la mise à jour des bases de données** à l'ensemble des prestataires de services,
- généralisation des **accords entre centres de ressources en bioinformatique et centres de calcul intensif et à haute performance** pour que ces derniers soient de plus en plus prestataires de ressources de calcul et de stockage en toute transparence pour l'utilisateur,
- Evolution des centres de calculs intensifs vers une réponse temps réels pour l'indexation des banques (quotidienne) et les requêtes des biologistes.

Propositions d'action

Nous proposons donc de mettre en place un modèle hiérarchique du calcul pour la biologie.

- Infobiogen ou tout autre centre national intéressé *(niveau 0)* pourrait par exemple être le centre de gravité de ce modèle hiérarchique. Une politique commune de gestion des mises à jour des bases de données pourrait ainsi être mise en place et coordonnée par le centre de niveau 0 pour en réduire le coût humain.
- Autour du portail national, seraient créés des ''grands si possible'' *centres régionaux (niveau 1)* disposant de ressources humaines pour la mise à jour des données et mettant à disposition un ou des portails web pour les utilisateurs. Ces centres de niveau 1 disposeraient de ressources de calcul et de stockage significatives et seraient reliés par des réseaux à très haut débit à des centres de calculs (centres nationaux, mésocentres) fournissant des ressources beaucoup plus importantes de calcul et de stockage.
- Autour de ces grands centres régionaux seraient mis en place des centres régionaux plus petits *(niveau 2)* disposant de ressources plus réduites pour mettre à disposition des services plus limités mais reliés aux centres de niveau 1 et de niveau 0 pour les services au-delà de leurs propres capacités.

Un tel modèle intègre la démarche actuelle du Ministère dans le cadre des génopôles. Les candidats naturels pour les centres de niveau 1 seront les centres de ressources bioinformatiques des génopôles. Cependant, ces centres doivent identifier des centres de calcul avec lesquels ils puissent collaborer selon le modèle proposé. Une telle approche hiérarchique permet de mieux absorber la communauté d'utilisateurs et privilégie les ressources de proximité.

Des synergies fortes pourraient être développées avec l'EBI. Un bon exemple de synergie avec l'EBI est le réseau d'excellence Embrace sur les grilles pour la bioinformatique (http://www2.cnrs.fr/presse/communique/610.htm). Son but est de standardiser l'accès aux innombrables données issues des projets de génomique, afin que les chercheurs puissent les consulter et les exploiter facilement. Le CNRS (à travers l'Institut de biologie et de chimie des protéines (IBCP) de Lyon, le centre de calcul de Lyon et le Laboratoire de physique corpusculaire (LPC) de Clermont-Ferrand) est responsable de la veille technologique : il fera en sorte qu'Embrace bénéficie des dernières avancées en matière de grille de calcul et que les choix retenus soient les mieux adaptés aux problématiques de la bioinformatique. La participation du CNRS au projet Embrace permettrait de faire bénéficier l'ensemble des centres des derniers développements (API, standards…) liés à ce réseau d'excellence, ainsi que d'un accès facilité et peu coûteux en ressources humaines à toutes les bases de données et algorithmes de ce réseau.

Dans ce contexte, le CNRS est associé au Généthon pour des premiers essais de gridification d'applications développées au niveau national, montrant ainsi son intérêt pour ce type de développement.

La technologie de grille semble être donc l'un des outils possibles pour mutualiser et partager les ressources (calcul, stockage, algorithmes, bases de données) entre les différents acteurs de ce modèle. Les fermes des centres de calcul de l'IN2P3, de l'IBCP, et bientôt du CINES, sont déjà nœuds du projet européen EGEE (http://public.eu-egee.org), dont le deuxième champ pilote d'applications est le biomédical.

Comme l'a demandé le Ministère à travers les actions génopôles, les centres acteurs de ce modèle doivent s'engager sur un cahier des charges précis des services offerts et doivent faire l'objet

d’une évaluation par rapport à ce cahier des charges. Ce modèle pourrait démarrer avec quelques centres pilotes.

Prospective nanosciences et matériaux

Composition du groupe

Xavier Blase	CNRS/LPMCN Lyon
Christophe Delerue	CNRS/IEMN
Thierry Deutsch	CEA/DSM/DRFMC
Patrice Hesto	IEF CNRS et Université Paris XI
François Jollet	CEA /DAM
Jean-Luc Leray	CEA/DAM
Georges Martin	CEA/HC
Lucia Reining	Ecole Polytechnique.
François Willaime	CEA/DEN
Ludger Wirtz	CNRS/IEMN
Gilles Zerah	CEA/DAM

Note : *La mission remercie Georges Martin d'avoir attiré son attention sur de nombreux aspects liés à la simulation numérique des matériaux intéressant les industries nationales, ainsi que sur le rapport du DoE sur la «Science numérique des matériaux appliquée à la fusion et aux réacteurs de 4ème génération[2] »*

[2] *Workshop on Advanced Computational Materials Science : Application to Fusion and Generation IV Fission Reactors*, DoE, Washington DC, 2004 (http://www.csm.ornl.gov/meetings/SCNEworkshop/Workshop-Report-ORNL-TM-2004-132.pdf).

Contributions

Exemples de besoins en moyens de calcul pour des simulations « Grand Challenge »
Contribution de G. Zerah (CEA/DAM)

Modélisation multi-échelles du comportement des métaux

L'objectif est de parvenir à une modélisation prédictive du comportement des solides, via l'utilisation de modèles décrivant plusieurs échelles spatiales et temporelles

La simulation multi-échelles des matériaux en mécanique vise à définir des modèles de comportement utilisables en calculs de type 'calcul de structures' (typiquement, un calcul d'ingénieur par éléments finis), à partir de données calculées à des échelles inférieures, depuis l'échelle atomique (10 -10 m) jusqu'à l'échelle du cm. L'objectif est de parvenir à effectuer des prédictions de comportement dans des conditions où l'expérience n'est pas réalisable (grandes vitesses de déformation, grandes pressions, grandes variétés de sollicitations, typiquement).

On peut distinguer deux types de propriétés relatives au comportement : Les propriétés de type « équation d'état ». Elles relèvent de la thermodynamique à l'équilibre, et elles permettent de simuler les grandes déformations élastiques (c'est-à-dire réversibles) de divers matériaux, des métaux aux nanotubes de carbone. Les propriétés élastiques utilisées au niveau macroscopique par un code de type « éléments finis » peuvent être calculées « à la volée » au niveau microscopique avec chaînage direct des codes

Les propriétés liées à la présence de défauts. Ce sont les propriétés élastiques affectées par des impuretés ou bien les propriétés de déformation plastique (irréversibles) déterminées par les défauts linaires appelées « dislocations » qui balayent un cristal parfait lors de telles déformations. Ces défauts sont susceptibles d'une traitement à une échelle mésoscopique.

En général, le chaînage des échelles dans les codes s'effectue préférentiellement à travers l'introduction manuelle dans un code traitant l'échelle supérieure, de données calculées une fois pour toutes à l'échelle inférieure. Cependant, dans certains cas de localisation des défauts (rupture par fissures, notamment), on peut raffiner localement la description géométrique pour descendre de l'échelle nanométrique à l'échelle atomique (méthode de quasi-continuum).

Les défis actuels concernent le développement de modèles de comportement macroscopiques de polycristaux capables de prendre en compte la simulation de la déformation en conditions rapides (chocs), le caractère polycristallin des métaux (effets des joints de grains)

Etude de la réactivité chimique par une approche multi-échelles

Les modèles du futur.

La description de l'évolution de milieux réactifs sous l'effet de la pression ou de la température doit reposer sur une analyse fine des réactions chimiques. Cette étude nécessite des modèles de type cinétique chimique qui traitent de manière exacte la zone de réaction.

Possibilités actuelles.

L'outil de base présent et futur pour l'analyse à l'échelle microscopique des propriétés de la zone de réaction chimique est la dynamique moléculaire classique (compte tenu du nombre de molécules à prendre en compte pour cette application spécifique (de l'ordre du million), le calcul *ab initio* reste exclu). Cette méthode nécessite la connaissance du potentiel classique rendant compte des interactions entre les atomes du système. Les outils aujourd'hui disponibles permettent de simuler :

- des systèmes inertes par hypothèse comportant quelques millions de molécules,
- des systèmes réactifs simples de type AB comportant quelques millions de molécules,
- des systèmes réactifs complexes comportant quelques centaines de molécules.

Ces outils permettent la prédiction de quelques-unes unes des données nécessaires à la construction d'un modèle macroscopique de type cinétique (éléments de cinétique chimique) et apportent des informations qualitatives sur la physique de la zone de réactions (structure et propriétés thermodynamiques). Les ressources de calcul nécessaires sont au maximum d'une centaine d'heures sur quelques centaines de processeurs.

Les travaux futurs concerneront, dans une première phase de 3 à 5 ans :

1. Paramétrage de potentiels classiques : calculs *ab initio* standards sur quelques centaines de molécules.

2. Extraction de toutes les informations nécessaires à la construction de modèles de type cinétique *i.e.* calculs des propriétés thermodynamiques et de cinétiques chimiques avec une grande précision : quelques dizaines de milliers de molécules associées à un potentiel évolué, soit quelques dizaines de milliers d'heures de calcul sur un processeur actuel de la machine du CCRT.

3. Simulation d'un front de réaction chimique dans des géométries permettant de mettre à l'épreuve les modèles cinétiques : comparaison directe entre le calcul de dynamique moléculaire et le calcul macroscopique. Le calcul microscopique nécessitera quelques millions de molécules associées à un potentiel évolué soit quelques millions d'heures de calcul sur un processeur actuel.

Au-delà de 5 ans :

1. Calcul en routine de la propagation de réaction chimique dans un environnement complexe (interaction avec différents milieux connexes) avec un modèle de type cinétique.

3. Calcul loupe (défauts, points singuliers, etc.) directement par dynamique moléculaire : quelques centaines de millions de molécules associées à un potentiel évolué, soit quelques centaines de millions d'heures de calcul sur un processeur actuel.

Modélisation multi-échelle des effets d'irradiation

L'objectif est la connaissance de l'évolution des propriétés des alliages d'actinides en fonction du temps sous irradiation.

Les défauts élémentaires issus par exemple de la désintégration d'un actinide radioactif sont produits dans un temps inférieur à la nanoseconde, alors que leur diffusion et l'établissement d'un régime stationnaire se produisent sur des durées allant de la microseconde à la seconde, quant au temps « macroscopique » d'observation il doit s'étendre sur plusieurs décennies.

D'autre part, les défauts ponctuels produits occupent un volume de quelques mailles cristallines, les défauts étendus comme les dislocations concernent le grain alors qu'un changement de phase (de type martensitique) ou un phénomène de gonflement concernent l'échantillon dans sa globalité (le poly cristal).

Il est donc nécessaire d'adopter une démarche multi-échelle pour modéliser cette évolution de manière fiable. Cette démarche peut s'articuler autour de deux objectifs :

- mieux connaître les processus mis en jeu à chaque échelle de temps et d'espace ;
- alimenter en données les domaines les plus macroscopiques.

Les échelles les plus exigeantes sont les échelles atomiques et quantiques.

1- Etudes à l'échelle quantique

L'objectif des études à l'échelle quantique est de mieux connaître les processus élémentaires impliqués dans la stabilité des alliages d'actinides (en particulier le rôle des corrélations électroniques) et d'alimenter les échelles supérieures – en particulier l'échelle atomique- soit par le calcul de données thermodynamiques utiles pour ajuster un potentiel effectif, soit pour élaborer un modèle spécifique du potentiel effectif lui-même.

2- Etudes à l'échelle atomique par la dynamique moléculaire

Les études à l'échelle quantique sont limitées en taille à moins d'une centaine d'atomes. Pour espérer prédire les conséquences de l'irradiation sur les propriétés des alliages d'actinide il est nécessaire de prendre en compte, au minimum, un volume qui contient la totalité d'une cascade de déplacement induite par les atomes de recul issus de la désintégration des atomes d'actinide. Cette échelle ne peut être atteinte qu'en abandonnant la description quantique et en passant à une description atomique de type dynamique moléculaire. On attend de cette description la connaissance de la nature et de la quantité des défauts produits par auto irradiation.

Nanoélectronique
contribution de J. L. Leray, CEA/DIF/DCRE

Note : *La contribution a été fournie à la mission sous forme de transparents qu'elle a résumés ci-après en raison de l'importance du sujet. Une proposition de programme incitatif fédérateur pluriannuel considérant à la fois la nanoélectronique et les matériaux a été transmise à la direction de la Recherche par le CEA. La suite qui pourrait être donnée relève de l'action de soutien « Calcul Intensif et Grilles » de l'Agence Nationale de la Recherche et du GIP qui doit la préfigurer.*

Les réalisations matérielles relevant des Technologies de l'information et de la communication reposent sur les circuits sur silicium. Ceci peut représenter jusqu'à 30 % de la valeur d'un produit tel qu'un avion militaire ou civil ou qu'une automobile. Deux types de calculs sont requis : 1) matière (nanosciences, nanoprocédés = nanomatériaux « classique ») 2) information (courants d'électrons, de spins etc. = « transport »).

Roadmap » de l'ITRS **(**International Technology Roadmap of Semiconductors), tracée jusqu'en 2016, mais comportant des défis technologiques importants. L'historique récente indique que les années 1998-2000 ont assisté à une accélération du progrès par rapport à la loi empirique de Moore• On est déjà dans le nanométrique, en ce qui concerne les transistors (45-32 nanomètres en xy, 1-2 nanomètres en z).

La simulation des matériaux est un soutien essentiel aux activités d'assemblage des parties du transistor « front end » et « device » (fonctionnement électrique) principalement. Les défis sont : les clusters, l'évolution vers les transistors à faible nombre d'électrons, qui requièrent la parallélisation des calculs. Il faudra également étendre les codes *ab initio* actuels aux modèles de transport quantique et mixer les deux niveaux de calcul : *ab initio* parallélisé / transport électronique en parallèle.

Il faut maîtriser la variabilité pour assurer la fiabilité, et le rendement de la fabrication ce qui requiert des calculs en parallèle et supervisés. Ce type de considération est essentiel pour la réussite du transfert industriel de l'innovation dans ces domaines caractérisés par une grande réactivité dans la compétition et l'intensité capitalistique des investissements sur les nouvelles productions.

Au-delà de ce plan, il faut considérer des plans « révolutionnaires » qui se justifient par l'essoufflement du paradigme de la loi de Moore, le coût des usines de fabrication par rapport à la solvabilité des marchés. Plusieurs concepts de transition ou alternatifs sont à considérer : « nano-inside » et nano tubes de carbone, électronique moléculaire, spintronique. Dans ce cas, on devra affronter la difficulté à simuler l'état amorphe.

Tableaux synthétiques

Les tableaux synthétiques se réfèrent aux questions posées pour l'exercice de prospective :

- Préciser la contribution attendue de la simulation, du calcul intensif et de la manipulation de données massives dans votre domaine scientifique ou technologique ?
- Quel pourra être l'impact de la mise en œuvre des ressources décrites dans les scénarios ci-après ?D'autre part, il a été demandé aux contributeurs d'envisager

les scénarios suivants :

	Scénarios			
Année	**France A**	**France B**	**Europe C**	**Europe D**
2006	25 teraflops	50 teraflops		
2009	90 teraflops	250 teraflops	0,5 petaflops	1 petaflops
2015	0,5 petaflops	2 petaflops	6 petaflops	12 petaflops
2020				100 petaflops

Pour permettre d'intégrer la réflexion sur l'utilisation éventuelle de machines d'architecture vectorielle, les scénarios suivants ont été ajoutés. Toutefois, le caractère spéculatif de l'analyse de l'offre de tels systèmes sur le marché a été précisé aux membres des groupes de réflexion prospective.

	Scénarios		
Année	France A	France B	Vectoriel E
2006	2 teraflops	5 teraflops	15 teraflops
2009	3 teraflops	10 teraflops	100 teraflops

Les scénarios ont été conçus dans la perspective suivante :

Scénario A Evolution à budget constant en se contentant du progrès technologique
Scénario B Rattrapage par la France suivant les recommandations de ce rapport
Scénario C Programme européen, exécution minimale
Scénario D Programme européen, exécution définie dans le projet tripartite

Problème/ Application	**Challenge scientifique et apport scientifique**	**Scénario choisi / Année**	**Ressources requise (teraflops efficaces*heure) / Mode d'exploitation**	**Renvoi table 2**
Fusion Contrôlée (CEA / DRFC)				
Ondes de chauffage dans ITER	Propagation et absorption d'ondes de chauffage additionnel Résolution spatiale fine. Calcul du chauffage dans ITER.	France B / 2010	1 000 teraflops*heure	
Instabilités dans les plasmas	Simulations d'instabilités MHD dans des plasmas en ignition. Résolutions spatiale et temporelle fines. Couplage aux particules rapides. Evolution non linéaire.	Europe C / 2010	100 000 teraflops*heure "Grand Challenge"	
Plasmas turbulents ITER	Transport turbulent dans les plasmas d'ITER. Turbulence développée. Description cinétique nécessaire : 5 D. Résolutions fines en espace, vitesse, et temps.	Europe C-D/2010	1 000 000 teraflops*heure	
Energie nucléaire (CEA /DEN)				
Cycle de vie des installations nucléaires	On résume ici la contribution CEA-DEN figurant ci-dessus. Les trois disciplines qui dominent pour les besoins en calcul intensif sont : les matériaux, la neutronique, la thermohydraulique.	France B Europe C ou D	Matériaux et thermohydraulique considérés « Grand Challenge »	

Climatologie / Institut Pierre Simon Laplace				
Climat futur de la terre	Quels sont les systèmes à prendre en compte pour déterminer le futur de la planète ? Modèle climat système terre avec toutes les composantes interactives permettant les phénomènes non linéaires.	France B Vectoriel parallèle 2006. Niveau de base : besoin dès 2005 Vectoriel Parallèle Scénario E 2009	**En 2005 : 50 % de la machine à 5 teraflops (40 % d'efficacité). Permet 15 simulations de 40 ans en 3 mois sur 40 processeurs vectoriels en parallèle. Mémoire critique : 160 Go** **En 2009 : 25 % de la machine à 100 teraflops (40 % d'efficacité) 15 simulations de 200 ans en 3 mois sur 200 processeurs en parallèle. Mémoire critique : 200 Go**	IPSL-8
Perturbations anthropiques et rétroactions nuageuses	Améliorer la connaissance de la réponse globale du climat aux perturbations anthropiques en améliorant les rétroactions nuageuses. Bonne représentation des nuages bas et de leur sensibilité aux changements de circulation, de température et d'humidité.	France B Vectoriel Parallèle 2006. Croissance CPU facteur 20 par rapport à ligne de base. Besoin dès 2005	Résolution verticale x2 à x4 couche limite. Résolution horizontale x4. CPU 50 % de la machine à 5 teraflops (40 % d'efficacité). 15 simulations de 200 ans en 3 mois sur 20 processeurs en parallèle chacune	IPSL - 1
Effet du couplage climat carbone à l'échelle globale.	Simuler les changements climatiques et les couplages climat-carbone à l'échelle globale. Idem ci-dessus avec en plus les modèles de carbone des surfaces continentales et de l'océan	France B Vectoriel Parallèle 2006. Croissance facteur CPU 30 et mémoire 2 par rapport à ligne de base. Besoin dès 2005	Résolution idem ci-dessus. CPU 50 % de la machine à 5 teraflops (40 % d'efficacité) 10 simulations de 200 ans en 3 mois sur 20 processeurs en parallèle chacune.	IPSL - 2
Quantifier les changements climatiques en Europe	Bien simuler les régimes dépressionnaires des moyennes latitudes. Effets des changements climatiques :tempêtes, inondations, sécheresse	France B Vectoriel Parallèle 2006. Croissance facteur CPU 15 et mémoire 10	Résolution horizontale 1°x1° (300 x 225) soit facteur 9. CPU 50 % de la machine à 5 teraflops (40 % d'efficacité) 10 simulations de 200 ans en 3	IPSL- 3

		par rapport à ligne de base. Besoin dès 2005	mois sur 20 processeurs en parallèle.	
Impact régional du changement climatique sur la chimie de l'atmosphère	Quantifier l'impact régional du changement climatique sur la chimie dans la troposphère et la stratosphère. Modèle couplé avec chimie-aérosols. Ajouter 200 traceurs et encore plus de réactions chimiques. Représentation onde de gravité. Prise en compte cirrus élevés et nuages polaires	France B Vectoriel Parallèle 2006 et Vectoriel Parallèle 2009. Croissance facteur CPU 100 et mémoire 40 par rapport à ligne de base.	Amélioration de la résolution verticale à la tropopause pour atteindre 100 niveaux Facteur 4 en résolution. CPU 50 % de la machine à 5 tera flops (40 % d'efficacité) 15 simulations de 40 ans en 3 mois sur 40 processeurs en parallèle Mémoire critique : 160 Go. En 2009, simulations plus longues.	IPSL - 4
Cyclones tropicaux	Simuler les phénomènes extrêmes dans les régions tropicales et leur évolution Simuler les cyclones tropicaux	France B Vectoriel Parallèle 2006 et Vectoriel Parallèle 2009. Croissance facteur CPU 90 et mémoire 40 par rapport à ligne de base.	Résolution horizontale 0,5°x 0,5° (150 x 110). CPU 50 % de la machine à 5 teraflops (40 % d'efficacité) 15 simulations de 40 ans en 3 mois sur 40 processeurs en parallèle En 2009, simulations plus longues ou à plus haute résolution	IPSL-5
Le devenir des phénomènes tropicaux	Le devenir des phénomènes tropicaux (El Niño, moussons, …). Résoudre les transferts verticaux aux échelles pertinentes. Formulation explicite de certains phénomènes : déferlement ondes internes, turbulences de surface	Vectoriel Parallèle E 2009	Facteur 1 000 par rapport au couplé 1° atmosphère et 0,5° océan. **Donc x 20 000 CPU 50 % de la machine à 100 teraflops (40 % efficacité)** 15 simulations de 20 ans 100 processeurs en parallèle	IPSL-6

Risque de surprise climatique liée à la circulation thermohaline.	Quantifier le risque de surprise climatique liée à la circulation thermohaline. Modélisation des interactions océan-glace de mer-atmosphère à échelle de la dizaine de km, simulations longues, approche statistique	Vectoriel Parallèle E 2009	Etape intermédiaire : 50 % de la machine à 5 teraflops (40 % d'efficacité) **8 simulations de 400 ans** en 3 mois **sur 40 processeurs** en parallèle	IPSL-7
Effet de l'utilisation des sols et de l'urbanisation sur le climat.	Comment évaluer l'effet des changements d'utilisation des sols et de l'urbanisation sur l'évolution du climat ? Modèle climat système terre avec toutes les composantes interactives permettant les phénomènes non linéaires	Idem IPSL-8 ci-dessus	Idem IPSL- 8	IPSL-9
Savoir utiliser les prochaines générations de machines.	Nouveau modèle avec bases numériques différentes pour parallélisation sur milliers de processeurs. Basculer la production scientifique sur ce modèle lorsqu'il permettra les mêmes études que le précédent.	France B Scalaire Parallèle	Accès à un grand nombre de processeurs pour développement, avec temps de restitution court.	IPSL-10
Océanographie à haute résolution				
Variabilité de l'océan à l'échelle globale	Variabilité de l'océan à l'échelle globale : sensibilité aux forçages atmosphériques. Comment fonctionne la variabilité de l'océan telle qu'on l'observe aujourd'hui ? Modèle ORCA – LIM résolution 1/4° 46 niveaux	2005-2007	2 400 heures/an simulé sur 186 processeurs IBM (IDRIS) Besoin: 150 ans simulés/année civile, 2005-2007	OCEANS-1
Variabilité de l'océan à l'échelle régionale	Variabilité de l'océan à l'échelle régionale : physique des tourbillons et courants de bord, transports de polluants, interaction turbulente avec l'atmosphère Compréhension des interactions d'échelles fondamentales pour le climat, compréhension des mécanismes générateurs des courants océaniques. Modèle ORCA -LIM, résolution 1/4° et 46 niveaux + zooms dans régions clés	2006-2009	Coût triplé par rapport à ORCA025 seul. Utilisation de calculateurs parallèles scalaires ou vectoriels	OCEANS-2
Circulation océanique à l'échelle globale.	Circulation océanique à l'échelle globale : vers la résolution des échelles importantes pour lever les incertitudes liées aux paramétrisations Défi de la représentation explicite des tourbillons et de la formation des eaux profondes.	2008 et au-delà	Facteur ~ 40 en CPU par rapport à ORCA 1/4° et Facteur ~ 14 en Mémoire	OCEANS-1

	Mise au point d'un modèle global au 1/12° Modèle ORCA –LIM, résolution 1/12° et 46 niveaux			
Compréhension du rôle de l'océan comme puits et source de CO2	Modélisation directe (non paramétrisée) de l'action de la mésoéchelle océanique sur les cycles biologiques et la séquestration du carbone. Elaboration des modèles physiques et des modèles de biogéochimie	2005-2009	Le coût du modèle biogéochimique représente 1 à 5 fois le coût du modèle physique..	OCEANS-3
Meilleure prise en compte des mesures.	Ré-analyses globales Modèle opérationnel avec assimilation Intégrer les observations des dernières décades dans un système dynamique cohérent pour la compréhension du système climatique. Modèle ORCA-LIM , résolution 1/12° et 46 niveaux Assimilation de données Schéma SEEK	2008 et au-delà	Facteur ~ 80 en CPU et facteur ~ 20 en mémoire	OCEANS- 4
Méthodes multi résolution matériaux métaux et biologie				
Chimie fine et catalyse	Modélisation de complexes organométalliques pour des applications à la chimie fine et à la catalyse : effet des ligands et des pairs d'ions Représenter des systèmes de grande taille, ioniques ou non en solution. Etudier la réactivité des différentes parties et leur influence réciproque.			MMR-Metaux
Catalyseurs greffés sur silice.	Modélisation de catalyseurs greffés sur surface de silice. Comparaison catalyse homogène/ catalyse supportée Difficulté de représenter un greffage non régulier sur une surface amorphe. Influence de la surface sur l'efficacité de catalyseurs.			MMR-Metaux

Batteries au lithium	Diffusion des ions lithiums dans les solides. Réorganisation du solide/ application aux batteries au lithium Représentation de la restructuration du solide. Résilience du matériau. Propriétés électroniques			MMR-Metaux
Matériaux nucléaires de structure et de stockage	Couplage entre les briques de base ayant individuellement atteint la maturité en 2005 : *Ab initio*, Dynamique moléculaire, Monte Carlo. Exemples : 1)matériaux de structure : aide au choix des matériaux des centrales du futur (fission 4ème génération). Matériaux de stockage (oxydes-verres)	France B complémentarité scalaire vectoriel appréciée		DEN-MMR
Modélisation prédictive du comportement des solides	Il s'agit de prédire le comportement macroscopique des solides à partir d'études multi-échelles partant des principes de base. Prise en compte des défauts et dislocations. L'intérêt est de traiter des cas qui ne se prêtent pas à l'expérimentation (rayonnement, longue durée, autres situations extrêmes)	France B Europe C	En 2006-2010 : 1 petaflops*heure par calcul, plus si anisotropie	DAM-MMR
Réactivité chimique	Description des l'évolution des milieux réactifs sous l'effet de la pression ou de la température à partir de l'analyse des réactions chimiques. Prise en compte d'environnement complexe. Construction de modèles cinétiques.	France B Europe C Europe D	5 petaflops*heure à 500 petaflops*heure	DAM-MMR
Modélisation des effets d'irradiation sur les alliages	Il s'agit de comprendre l'évolution des propriétés des alliages d'actinides dans le temps sous irradiation. Les phénomènes se déroulent sur des échelles de temps allant de la nanoseconde pour la désintégration d'un actinide radioactif, à la seconde pour l'établissement d'un régime stationnaire. La prédiction doit se faire sur plusieurs décennies.	France B Europe C	5 petaflops*heure par calcul en 2006-2010	DAM-MMR
Nano-électronique	Il s'agit de se donner les moyens de suivre l'évolution des nouveaux dispositifs électroniques, qui peut se faire de manière évolutive dans le cadre la « Roadmap ITRS » ou révolutionnaire. Dans les deux cas des défis importants de modélisation et de simulation se présentent.			CEA-Nanoélec

Senseurs et commutateurs biologiques, micro livraison de médicaments	Simuler la dynamique de l'auto-assemblage de nanostructures biologiques. Conception rationnelle d'outils nanotechnologiques tels que des senseurs et des commutateurs biologiques, des systèmes de micro-livraisons de médicaments.	France A		MMR - biologie
Mécanismes moléculaires de la biologie, des maladies. Conception de nouveaux médicaments	Simuler des systèmes biologiques : simuler la dynamique de l'assemblage et des transformations de macromolécules biologiques, et ensuite d'organelles et de cellules entières (organisation de la chromatine, repliement des protéines, assemblage des membranes, transduction du signal, formation des complexes). Comprendre la formation des complexes et la transmission d'information. Comprendre les mécanismes moléculaires des processus biologiques. Description au niveau moléculaire des maladies. Conception rationnelle de nouveaux médicaments.	France B		MMR - biologie
Enzymologie moléculaire.	Simulation de réactions enzymatiques. Prédictions métaboliques.	France B		MMR - biologie
Astrophysique : Projet HORIZON				
Galaxies	Physique du gaz à petite échelle, galaxies spirales et elliptiques, interactions, étoiles et vents	France B 2006 France B 2009 Europe D 2009	1000 / communauté 5000 / communauté 20000 / grand challenge	Horizon Galaxies

Amas de galaxies	Physique du gaz à grande échelle, amas de galaxies, structure interne et distribution à grande échelle	France B 2006 France B 2009 Europe D 2009	500/communauté 2500/communauté 10000/grand challenge	Horizon Amas de galaxies
Grandes structures	Formation des halos de matière noire et distribution à grande échelle, modèles semi-analytiques, surveys	France B 2006 France B 2009 Europe D 2009	250/communauté 1250/commu nauté 5000/grand challenge	Horizon Grandes structures
Astrophysique CEA				
Vision MHD 3D Soleil/ Etoiles	Comprendre la dynamique interne des étoiles et le Soleil archétype	2000-2005: CCRT/CEA 2005-2010: France B 2006 2010-2015: Europe D	2005-2010: 800 heures par run, machine scalaire (effectif 5 teraflops, dizaine de runs) 2010-2015: machine scalaire , 1500 heures/run (quelques dizaines de runs, résolution au moins 5 fois supérieures par dimension)	
Relation Soleil-Terre, interaction entre étoiles	Comprendre la variabilité du flux solaire et ses interactions avec la Terre	2005-2008 : calcul 1D/3D France B 2008-2012: observations / prédictions	Machine vectorielle Quelques runs de 20 heures (effectif 5 tera flops) Au-delà 2010 : Quelques centaines d'heures/run ?	
Combustion & Supernovae	Comprendre les mécanismes de fragmentation du milieu interstellaire jusqu'à la formation de cœurs denses préstellaires	France-B: 2006/2015Europe-D: 2009	2005-2010: Environ 30 teraflops*heure par simulation. 2010-2015: De 300 à 1 000 teraflops*heure selon la physique implémentée.	

Milieu Interstellaire	Comprendre les mécanismes de fragmentation du milieu interstellaire jusqu'à la formation de cœurs denses préstellaires	France-B: 2006/2015 Europe-D: 2009	2005-2010: Environ 30 teraflops*heure par simulation.2010-2015 : De 300 à 1 000 teraflops*heure selon la physique implémentée.	

TABLEAU II : "Pré-requis, Blocages à lever"

Référence table 1	**Pré-requis / Difficulté scientifique / point bloquant / besoin formation**	**Action à entreprendre**	**Planification souhaitable de l'action (Années début/fin)**
Astrophysique : Projet HORIZON			
Horizon **Galaxies**	Une meilleure résolution implique une physique à petite échelle plus riche. Stockage : schéma d'indexation des données Scalabilité pour nombre CPU > 1000 Technique de parallélisation massive et de gestion des grosses bases de données	1- Développement de routines de physique (atomique, moléculaire, modèle sous-maille, diffusion des métaux, vents, trous noirs super-massifs) 2- Développement des logiciels d'analyse et de visu adaptés aux très gros runs 3- Tests des algorithmes de décomposition de domaine pour NCPU > 1000	1- 2005-2007 2- 2005- 2007 3- 2005-2007
Horizon **Amas de galaxies**	idem	4- Développement de logiciels d'analyse en temps réels (cartes virtuelles)	4- 4- 2005-2007
Horizon **Grandes structures**	idem	5- Développement de modèles semi-analytique de post-traitement parallèles	5- 5- 2005-2007

Climatologie / Institut Pierre Simon Laplace			
IPSL-1 à 6	• Poursuite du développement de la physique des modèles • Augmentation de la résolution verticale et horizontale • Parallélisation du modèle sur 10aine de processeurs • Banaliser les simulations d'ensembles • Formation au calcul parallèle de tous les étudiants • Besoin d'organisation pour préparer les • Organisation autour des ressources matérielles et humaines	• Maintenir de bons temps de réponse pour les développements du modèle • Maintenir de bons temps de réponse pour les ajustements des paramétrisations lors de la réalisation des expériences de référence • Finaliser la parallélisation du couplé actuel • Mise en place outils pour simulations d'ensemble, travail main dans la main avec centres de calcul pour mise en place grosse machinerie incluant pré et post-traitements, soumissions enchaînées de jobs, … • Maintenir ressources stockage et post-traitement • Faciliter exploitation résultats sur les différents sites de calcul	Au plus tôt Et à maintenir dans la durée
IPSL 7-à 9	Faire des simulations climat avec le système complet impose des contraintes lourdes en CPU, mémoire, espaces fichiers, possibilités post-traitements…	Décision de s'équiper du calculateur permettant ce type d'étude à prendre maintenant	
IPSL- 10	Equipe pluridisciplinaire • Numéricien • Algorithmicien • Physicien • Informaticien • Chimiste	• Compléter équipe actuelle pour avoir le « répondant » nécessaire aux sollicitations pour construire ce nouveau modèle • Détecter 4 ans à l'avance la disparition des calculateurs vectoriels pour fixer le rendez-vous avec le modèle climat suivant capable de performances sur des milliers de processeurs	• A monter dès maintenant • Basculer production scientifique sur ce modèle lorsqu'il permettra des avancées scientifiques climat. • J-4ans
Océanographie à haute résolution			
OCEANS-1	Prérequis : modèle océanique performant pour la représentation de la mésoéchelle et des interactions courant/topographie : OPA 9 est un outil performant à continuer à développer. Point bloquant : disposer des heures de calcul nécessaires	1 - optimiser l'accès aux ressources de l'IDRIS et aux autres calculateurs européens pour les gros projets 2 - faire évoluer rapidement les moyens de calcul (scalaire et vectoriels) disponibles en France.	Immédiatement, développement à long terme : 2005-2012 et au-delà

OCEANS-2	Pré-requis: outils pour la régionalisation (données aux frontières, logiciel tel que AGRIF (développé au LMC) permettant d'effectuer des zooms régionaux. Outils existants mais à développer fortement	Outre le problème de temps calcul, il y a un manque de personnel permanent (ingénieur) sur ces questions en France. Le travail est effectué trop souvent par du personnel temporaire (grande perte de temps en formation)	Immédiatement, développement à long terme : 2005-2012 et au-delà
OCEANS- 3	Prérequis : modèles bio-géochimiques performants	Renforcer pluridisciplinarité bio-géochimie / dynamique	Immédiatement, développement à long terme: 2005-2012 et au-delà
Méthodes multi résolution matériaux métaux et biologie			
MMR - Métaux	Tous les processus de calculs mentionnés sont itératifs et la durée d'une itération doit être nettement inférieure à la limite de temps de chaque soumission. Pour des systèmes de grandes tailles, des limites ont été atteintes et le calcul devient inefficace.	Autorisation d'avoir des durées de calculs supérieures à 24h car l'augmentation du nombre de processeurs semble ne pas être toujours la solution efficace.	
MMR-biologie	Continuer le développement de modèles physiques. Continuer la parallélisation des logiciels Validation des modèles. Applications pour valider.		
DEN-MMR	Couplage entre briques de base Calculs de forces dans l'état excité Cascades de déplacement *ab initio* Conserver leadership européen dans les codes *ab initio*	Poursuivre investigation des méthodes de calcul de forces dans l'état excité Mettre en place une structure nationale d'accueil pour l'aide au développement. Formation en sciences « numériques » dans la filière « matériaux » Participation aux projets de logiciel libre (ABINIT, PWSCF, FPMD). Améliorer position dans les réseaux européens : Psi-k, CECAM (Lyon), ICTP (Trieste)	
DAM-MMR CEA-Nanoélec	Programme fédérateur **« maîtrise du logiciel et du calcul intensif pour la simulation des matériaux et des nanostructures » soumis au Ministère de la Recherche et à l'ANR**	- rendre plus performant les logiciels de base pour chacune des échelles, - faciliter l'intégration horizontale pour le calcul multi-échelles, -organiser des workshops et des formations spécialisées.	Action à lancer immédiatement, ne produira son effet que progressivement sur 5 ans, ce qui est l'échelle de temps la plus courte à considérer pour les développements de codes et de méthodes ambitieuses.
Astrophysique CEA			
Astro CEA -	Contraintes expériences Spatiales/ Problème des	-Renforcer assistance programmation	**2000-2005:** Mise en place équations, validation

I	instabilités numériques et physiques/ Topologie des champs magnétiques, adéquation des méthodes numériques à la complexité grandissante des problèmes / Formation aux techniques numériques, support	-Amélioration des techniques numériques pour résolution spatiale au-delà de 1000^3 et de la puissance de calculs pour simulations plus réalistes	code **2005-2010:** Etudes des interactions convection-rotation- turbulence-magnétisme **2010-2015:** calculs réalistes, physique plus riche (atmosphère, multifluides, conditions aux limites) **2015-2020:** Evolution stellaire (sur des Gyr) d'étoiles 3D MHD, phases précoces et ultimes, liens dynamiques entre différentes phases, Machines scalaires + vectorielles ?
Astro CEA - II	Description atmosphère,flux solaire=f(t,?)./ transition champ B fort à faible/Formation aux techniques numériques, support	-Portabilité des codes locaux 1D enrichis sur machines vectorielles pour grands nombres de pas temporels -- Mise en place d'un protocole faisant le pont entre la dynamo solaire interne et le code magnéto-sphérique de la terre.	**2005-2008:** Mise en place du problème ; simulations prévisionnelles des observations **2008-2012 :** extension au climat du passé : 1D, 3D ? **2012-2015:** dépendra des précédentes étapes : non défini aujourd'hui
Astro CEA - III	-Étudier et maîtriser la micro-physique en œuvre dans les SNIa. -Réunir des experts des nombreuses disciplines impliquées (hydrodynamique, combustion, physique nucléaire, physique stellaire…) - Trouver des algorithmes parallèles performants pour modéliser cette physique sur des machines à plus de 1000 processeurs.	- Poursuivre et intensifier les collaborations transdisciplinaires, notamment à travers le projet « Supernovae et Combustion ». -Former des étudiants à ces disciplines et au numérique - Développer un support technique de haut niveau proche des physiciens.	**2005-2010:** Analyse spécifique des différents processus physiques et modélisation mono-dimensionelle des supernovae **2007-2015:** calculs multi-dimensionnels incluant au mieux les différents processus microphysiques.
Astro CEA - IV	L'objectif est de faire le lien entre les échelles galactiques et les cœurs denses de formation stellaire	- Poursuivre le développement d'un code spécifique et des outils d'analyse associés - Développer un support technique de haut niveau proche des physiciens	**2005-2010:** Simulation hydrodynamique avec de la gravite et de la chimie. **2007-2015:** Simulation MHD. Prise en compte de manière réaliste de la turbulence.

Formation au calcul scientifique

Note sur la formation en calcul scientifique
contribution de Serge Petiton, CNRS /STIC

Cette note synthétise certains points évoqués par le groupe de travail et propose quelques pistes possibles ; indépendamment de l'existant. Depuis de nombreuses années la question de la formation en calcul scientifique intensif est discutée régulièrement sans vraiment de réponses claires. A notre avis, le thème de la formation au calcul scientifique devrait faire l'objet d'une réflexion approfondie.

Les formations actuelles en calcul scientifique, intensif ou non, sont difficiles à lister et synthétiser. Elles sont rattachées à différents départements selon les universités ou écoles et n'apparaissent pas toujours en tant que telles dans les cursus. La mise en place récente des LMD ayant restructuré aussi ces enseignements, il est difficile de bien connaître et évaluer les nouveaux. De plus, les mêmes intitulés de cours et des contenus similaires peuvent correspondre à des formations très différentes selon les domaines de recherches des enseignants.

Ces formations doivent s'inscrirent complètement dans le paysage du calcul scientifique en France. Si elles ne sont pas à la hauteur des ambitions affichées par ailleurs, l'efficacité des efforts faits sera moindre.

Formation en Licence et première année de Master

Constatant que de nombreux étudiants des filières universitaires scientifiques ne connaissent que très peu les bases en calcul scientifique en sortie des masters, le plus important semble être d'enseigner un « ***socle commun de connaissance en calcul scientifique*** » aux étudiants de Licence ou de première année de Master (dans le cadre de la structure LMD). On peut penser qu'il n'est pas nécessaire de faire ces cours en Licence mais uniquement en première année de Master. Néanmoins, les connaissances scientifiques nécessaires à l'acquisition de ce socle commun sont suffisantes dès la Licence pour presque l'ensemble des étudiants des disciplines scientifiques. Les élèves-ingénieurs ayant généralement une formation en calcul numérique et en informatique plus complète sont souvent mieux préparés pour le calcul scientifique. Le contenu de ce socle commun est à préciser mais devrait inclure des notions de méthodes numériques, d'informatique scientifique (au moins sur l'arithmétique flottante et les fonctions élémentaires), de génie logiciel et de formation à quelques outils classiques liés au calcul scientifique. Un nombre d'heure suffisant de travaux pratiques doit être proposé car ce domaine ne peut être uniquement enseigné magistralement.

Formation en seconde année de Master

Sur ce socle commun de connaissance en calcul scientifique doit ensuite reposer la formation en calcul scientifique de seconde année de Master (*professionnel* ou *recherche*). Celle-ci doit prendre des formes variées selon la discipline. Des masters à vocations purement « calcul scientifique » ou « calcul haute performance » doivent se développer, mais l'ensemble des étudiants en master scientifique doit pouvoir suivre des cours sur ces deux thèmes au sein des écoles doctorales ; dans des enseignements transversaux ou optionnels, par exemple. Ces enseignements doivent être pluridisciplinaires et non concerner que certains aspects du calcul scientifique. Des expérimentations sur machines adéquates doivent être proposées ; au niveau des mésocentres ou des centres de calcul des universités et écoles. Ces formations concernent principalement les masters *recherches* mais il est aussi nécessaire de développer des enseignements adéquats dans les masters *professionnels* pour former les ingénieurs en « génie calcul scientifique » nécessaires dans ce domaine.

Formation doctorale et post-doctorale

Les formations proposées doivent être aussi accessibles aux doctorants. Des enseignements plus « techniques », comme ceux enseignés à présent par les centres de calculs nationaux, doivent être proposés aux doctorants qui doivent utiliser des machines hautes performances. Un label «calcul scientifique » ou « calcul haute performance » peut être donné à certaines thèses de doctorat. Ce type de label a été proposé lors de discussions au sein d'ORAP (mais sans recommandations particulières pour l'instant). Ce label pourrait, par exemple, être demandé pour certains types de financement post-doctoral qui serait mis en place pour favoriser des recherches pluridisciplinaires.

Annexe : prospective INRIA