PENSER PETAFLOPS

Groupe de travail : Infrastructure du calcul intensif

10/11/2008

Liste des participants au groupe de travail:

Anthony	ASSI	INRIA
Françoise	BERTHOUD	CIMENT
Dominique	BIRMAN	Météo-France
Christophe	BLANCHET	IBCP
Domnique	BOUTIGNY	CNRS / IN2P3
Bruno	BZEZNIK	Université Joseph Fourier
Jaume	CARBONELL	LPSC Grenoble
Claude	COHEN	BULL
Philippe	COUVEE	Bull HPC R&D
Laurent	CROUZET	CEA
Romaric	DAVID	CECPV - Université Louis Pasteur
Sébastien	DENVIL	Institut Pierre Simon Laplace
Laurent	DESBAT	UJF Grenoble 1
Dominique	FOUGERE	M2P2-UMR6181- Université de la Méditerranée
Marie-Alice	FOUJOLS	IPSL - Pôle de modélisation
Erbacci	GIOVANNI	CINECA Supercomputing Center
Michel	KERN	Ministère de la Recherche
Violaine	LOUVET	CNRS/Université Lyon 1
Stratis	MANOUSSIS	CNRS/INSU
Corine	MARCHAND	BULL
Patrick	MASCART	Observatoire Midi-Pyrénée
Alain	MINIUSSI	Observatoire de la Cote d'Azur
Geneviève	MOGUILNY	Institut de Physique du Globe
Alain	PASTUREL	CNRS
Karim	RAMAGE	IPSL/CNRS
Olivier	RICHARD	Laboratoire d'Informatique de Grenoble
Francois	ROBIN	CEA
Paul	ROUSSEAU	ServiWare - Groupe BULL
Dorothée	SENECHAL	UPMC Paris 6 Institut Jean le Rond d'Alembert
Thomas	SIMONSON	Ecole Polytechnique
Pierre	VALIRON	Observatoire de Grenoble
Jean-Pierre	VILOTTE	Institut de Physique du Globe
Laurence	VIRY	CIMENT - UJF Grenoble
Jules	WAKU	

Pierre Valiron du Laboratoire d'Astrophysique de Grenoble était également membre de ce groupe de travail et y avait contribué. Il est décédé le 31 août 2008. Voir : https://ciment.ujf-grenoble.fr/news/pierre-valiron-nous-a-quitte

Sommaire

1. Introduction....3
2. Cas d'utilisation....3
2.1 Calculs sur réseau en Chromodynamique Quantique (LQCD)....4
2.2 Climatologie....6
2.3 Météorologie....7
2.4 Sciences de la terre....8
3. Les données....8
4. Relations avec les Méso-centres....10
5. Relations avec les Grilles....11
5.1 Les grilles de productions :....11
5.2 Relations entre les Grilles et les méso-centres....12
6. Importance du réseau....13
7. Calculateurs et environnement....14
8. Support....15

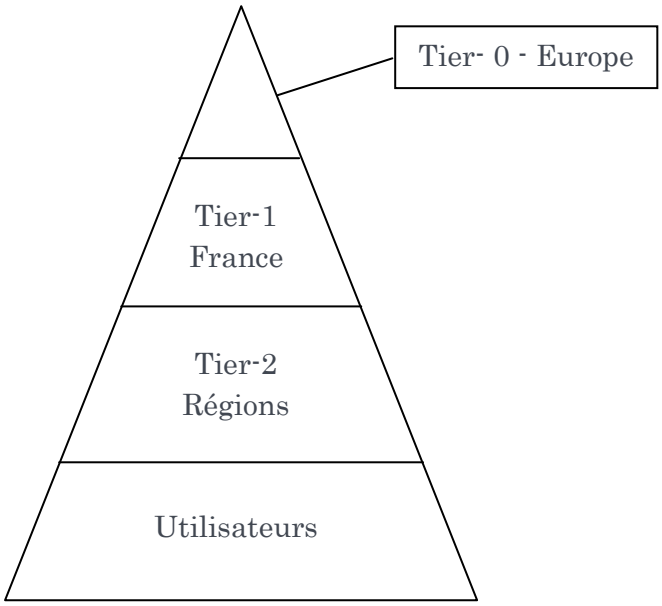
PENSER PETAFLOPS

Infrastructure du calcul intensif

1. INTRODUCTION

L'infrastructure du calcul intensif est l'ensemble des moyens qui doivent être mis en œuvre afin de rendre l'utilisation des très grands calculateurs, souple, efficace et optimale. Il n'est en effet pas envisageable de considérer l'installation en France ou ailleurs en Europe, d'un calculateur de la classe pétaflopique sans avoir conçu, renforcé et coordonné, l'ensemble de la chaîne du calcul intensif. Les principales composantes de cette infrastructure sont : La gestion des données, les méso-centres, les grilles de calcul et de stockage et l'architecture réseau.

L'infrastructure du calcul intensif est la charpente de la pyramide :



Le groupe de travail a considéré qu'il était important de mettre l'accent sur les problèmes environnementaux posés par le calcul. Cette dimension devant dorénavant impérativement être prise en compte lors de l'acquisition et de l'installation de nouveaux équipements.

Les méso-centres étant le plus souvent associés à des Universités, le groupe de travail a souligné l'importance de maintenir et de développer un excellent niveau de formation dans les méso-centres, ce thème n'a toutefois pas été développé ici, puisqu'un groupe de travail lui est dédié.

2. CAS D'UTILISATION

Cette section est destinée à montrer par quelques cas d'utilisation l'importance des différents éléments de l'infrastructure du calcul intensif qui entrent en jeu dans le cadre d'applications réelles. Ces cas d'utilisation ne sont que des exemples sélectionnés en fonction de l'expertise des membres du groupe de travail, ils ne prétendent pas être exhaustifs.

2.1 Calculs sur réseau en Chromodynamique Quantique (LQCD)

Les calculs sur réseau constituent à l'heure actuelle le seul moyen pour résoudre les théories quantiques des champs dans le domaine non perturbatif, en particulier la chromodynamique quantique (QCD) ou théorie des interactions fortes. Le but de cette théorie, élaborée dans les années 70, est d'expliquer la cohésion des noyaux ainsi que la structure de leurs composants – protons et neutrons – c'est-à-dire l'essentiel de la matière visible de l'Univers.

L'activité LQCD en France se trouve maintenant en pleine expansion. Elle se développe dans plusieurs laboratoires et s'intègre dans trois des plus importantes collaborations européennes. Ce développement s'accompagne d'un remarquable effort au niveau national avec l'achat de supercalculateurs (BlueGene/P à l'IDRIS, Altix au CINES, Bull au CCRT) qui remettent la France au niveau de ses voisins européens les plus performants.

Le calcul sur réseau a été dès son origine au berceau du Calcul de Haute Performance (HPC), soit en construisant ses propres ordinateurs dédiés (e.g. la saga des APE) soit en inspirant les architectures des supercalculateurs (e.g. le BlueGene/P issue de la collaboration anglo-américaine QCDOC).

Les ingrédients de la QCD sont les quarks – briques élémentaires de la matière – et les gluons – qui en assurent la cohésion – représentés par des "champs quantiques" i.e. des familles d'opérateurs définis en chaque point de l'espace temps.

La dynamique des quarks et des gluons est régie par la théorie quantique des champs dont les équations du mouvement permettent en principe de connaitre leur valeur en chaque point. En pratique une telle résolution est impossible. Il existe cependant un formalisme dû à Feynman qui permet d'obtenir les valeurs moyennes de produits des opérateurs comme une somme sur toutes les réalisations possibles des champs. Un tel calcul n'est possible qu'en discrétisant l'espace-temps, ainsi Le continuum spatio-temporel R^4 devient un réseau ("lattice") de $V=L^3\times T$ sites où L représente le nombre de points dans chacune des 3 dimensions spatiales et T le nombre de point dans la dimension temporelle.

Les **étapes d'un calcul sur réseau** sont bien différenciées quant aux types de ressources informatiques qu'elles requièrent, couvrant l'étendue des moyens offerts par les centres de calcul:

1- Génération des configurations de jauge suivant une loi de probabilité :

C'est, de loin, la partie la plus difficile et la plus coûteuse, elle se fait sur des méga-centres et constitue le fer de lance de la LQCD. Le coût caractéristique pour un réseau L=48 T=96 est d'une heure par configuration avec 10000 CPU (MPI). La communauté française utilise les APEnext pour générer les configurations des petis réseaux ($L \leq 32$) et bénéficie depuis cette année d'un apport essentiel du BlueGene/P d'IDRIS pour accéder aux grandes tailles ($L \geq 32$).

Les méso-centres participent également de façon essentielle à cette étape pour le stockage et la mise en commun des configurations générées. La taille d'une configuration pour L=32 est 1.26 Go et le nombre de configurations pour une taille de réseau donnée est de plusieurs milliers. Il est ici indispensable de pouvoir disposer d'outils permettant des accès rapides et des transferts massifs de fichiers de données. Le CCIN2P3 a mis à la disposition de LQCD le protocole de stockage de données SRB interfacé au système de stockage hiérarchique HPSS.

2- Calculs des propagateurs S de quarks.

Il se ramène à la résolution d'un système linéaire qui se fait par des méthodes itératives. Pour les petites tailles – L=24,32 – la quasi-totalité des inversions se fait dans les mésocentres (CCIN2P3), où les configurations ont été stockées et où seront aussi stockés les propagateurs obtenus. Toutefois pour L=32 les ressources de ce type de centres sont déjà très limitées. Au delà il ne reste que le choix des méga-centres car ces calculs doivent être fait en parallèle sur des centaines de nœuds. Des demandes de calcul sont en cours à IDRIS et au CINES pour réaliser ces inversions assez coûteuses.

3- Calcul des observables physiques.

Ces calculs se ramènent à de simples manipulations algébriques des propagateurs. Ils ne nécessitent pas des tâches très longues. Leur difficulté vient des tailles des propagateurs (15 Go pour un baryon avec L=32) et du fait qu'il faut moyenner chaque fois sur plusieurs centaines de propagateurs.

L'analyse statistique des résultats (moyennes, calcul d'erreurs, ajustement, …) se fait localement dans les mini-centres que constituent les laboratoires.

En conclusion, d'un point de vue historique, la LQCD s'est toujours développée en synergie avec le calcul haute performance. Non seulement elle lui doit sa viabilité mais elle a contribué à définir l'architecture même des ordinateurs. Cette synergie se poursuit tant au niveau national (ANR PetaQCD) qu'européen (PRACE, QPACE).

Les méga-centres (IDRIS, CINES, CCRT) sont réservés aux calculs les plus lourds, essentiellement l'obtention des configurations de jauge. Leur puissance définit en fait les ambitions de ce type de calcul ainsi que celle des pays qui en disposent.

L'articulation avec les méso-centres se fait tout naturellement. Elle est essentielle à plusieurs égards: *(i)* pour une partie moins lourde, mais tout aussi indispensable des calculs (obtention des propagateurs, contractions); *(ii)* en ce qui concerne le stockage et mise en commun et *(iii)* pour permettre l'accès à un nombre – relativement important – des membres impliqués dans une des collaborations de LQCD.

Signalons ici le rôle clé du CCIN2P3 qui, avec sa capacité de stockage inégalée en France et sa puissance de calcul appréciable, a permis qu'une bonne partie des calculs de LQCD puisse être menée à bien.

2.2Climatologie

Les applications en climatologie nécessitent des ressources calcul importantes et ce depuis le début de ce domaine de recherche. Par exemple en 2009, nous estimons à l'équivalent de 80 processeurs NEC SX-8, pendant toute l'année, les ressources calcul nécessaires pour la réalisation des simulations GIEC. Et ceci ne représente que la moitié des ressources nécessaires à nos recherches. Dans le même temps, les évolutions nécessaires tendent toutes à un accroissement conséquent des ressources calcul :

- augmentation de la résolution horizontale
- augmentation de la résolution verticale
- augmentation de la complexité (intégration de toutes les composantes du système climatique)
- augmentation de la qualité (des paramétrisations)
- augmentation du nombre de simulations des ensembles
- augmentation des longueurs des simulations

Celles-ci impliquent des fichiers de plus en plus volumineux, des analyses automatiques et des standardisations (métadonnées).

L'environnement nécessaire à ses simulations impose des ressources calcul importantes, rapides; le système de fichiers doit être à la hauteur; les capacités d'analyses automatiques dimensionnées en conséquence. A côté de centres de calcul équilibrés (calcul, fichiers, post-traitement), le système de diffusion des données doit être dimensionné (réseau, système de distribution de type Earth System Grid) pour permettre la diffusion des données depuis les centres de production des données et l'accès depuis tous les types de postes de travail du scientifique. Le réseau est un point crucial et les systèmes locaux (labos, instituts et méso-centres) doivent être capables de gérer des données en masse pour permettre les analyses in fine en local.

Durant l'année 2009, au cours de laquelle l'équivalent de 80 processeurs NEX SX8 sera dédié aux simulations GIEC, nous estimons à environ 250 TB le

volume de données destiné à être distribuer ; la quantité totale de données produites devant avoisiner les 1000 TB. Ceci met d'autant plus en avant la nécessité d'avoir des centres de calcul équilibrés en terme de serveur de calculs, de serveurs de fichiers et de serveurs dédiés aux post-traitements.

Les données provenant des modèles climatiques sont ainsi produites en grande majorité sur les centres de calculs nationaux tels l'IDRIS ou le CCRT, voire internationaux dans certains cas, comme le Earth Simulator japonais. Plusieurs initiatives internationales développent des standards pour décrire les modèles climatiques et les données produites, en particulier le projet européen METAFOR et le projet américain CURATOR, ouvrant la voie à une généralisation de l'approche par base de données des résultats de simulations climatiques. La quantité de données générées devenant conséquente, l'idée de base est d'éviter au maximum de les déplacer, de les laisser dans les centres de calcul, et de leur donner un accès unifié en construisant une « grille de données ». Celle-ci doit impérativement inclure les centres de calcul nationaux (IDRiS, CCRT, CINES et quelques méso-centres).

Le prochain rapport de l'*International Panel on Climate Change* (le cinquième rapport : IPCC AR5) s'appuiera effectivement sur des données distribuées, et non plus centralisées. Dans ce contexte, l'Institut Pierre Simon Laplace étudie de près les potentialités de la grille et son apport possible en terme de gestion de données. Il vise à intégrer l'ensemble de l'architecture (*data nodes*, *gateway*, *global services*), avec la mise en place de *data nodes* dès 2009. L'apport des fonctionnalités des portails devrait permettre d'élargir de façon significative l'éventail des utilisateurs des données de simulation numérique.

2.3Météorologie

ECMWF : Un exemple "qui marche" :

ECMWF (European Centre for Medium-Range Weather Forecast) est une organisation internationale à laquelle adhèrent 31 Etats Européens. Ses objectifs principaux sont la production opérationnelle d'analyses, de prévisions à moyen terme (de 5 à 10 jours, les services météorologiques nationaux se concentrant sur le court terme, de 1 à 5 jours, sur leur zone d'intérêt), du temps et de l'état de la mer et de prévisions saisonnières, ainsi que la recherche scientifique et technique tendant à améliorer la qualité de ces prévisions. Au-delà de sa mission de prévision opérationnelle, ECMWF propose entre autres à ses Etats-Membres l'accès à des données issues de ré-analyses et l'utilisation de ses moyens de calcul, notamment pour des activités de recherche. C'est ce dernier aspect seul que je prends en compte ici.

Le constat est que l'activité de support d'ECMWF, vue par l'utilisateur, fonctionne bien. Elle s'appuie sur une structuration classique en trois niveaux.

- Le niveau expert est assuré par les spécialistes du Centre lui-même et parfois de ses fournisseurs.

- Le niveau intermédiaire est assuré par une équipe locale dédiée de cinq personnes.
- Le premier niveau est partagé entre une équipe locale resserrée (Call-Desk) pour les urgences et un correspondant chez chacun des services météo des Etats-Membres. Ce correspondant connaît bien le Centre, il est le premier contact des utilisateurs nationaux pour tout besoin de support et le premier interlocuteur du Centre pour toute information vers ces utilisateurs nationaux.

Les correspondants sont réunis une fois par an par le Centre pour des réunions d'information et d'échanges. Un utilisateur peut être amené à séjourner au Centre (parfois pour quelques semaines) dans le cas de portages lourds ou d'optimisation délicate.

2.4Sciences de la terre

Voir document en annexe : « **Modélisation numérique et infrastructures en Sciences de la Terre** » ***J.-P. Vilotte , E. Chaljub, E. Dormy, F. Fluteau, A. Fournier, J. Virieux***

3. LES DONNEES

L'ensemble des communautés scientifiques concernées par le calcul intensif, qu'il soit massivement parallèle sur les supers-calculateurs, ou bien séquentiel sur les grilles de calcul, font ressortir l'importance croissante des données.

Traditionnellement la physique des hautes énergies, en raison du nombre de canaux mis en jeu dans les expériences, a toujours été amenée à manipuler des quantités très importantes de données. Par exemple, les expériences installées sur le LHC, vont produire 15 Po de données brutes chaque année, qui vont être distribuées et traitées sur une grille de calcul. L'aspect international de la physique des hautes-énergies a conduit celle-ci à mettre en place des mécanismes permettant de distribuer et de stocker les données de manière très automatisée. Il y a donc là un savoir faire qu'il conviendrait de partager auprès des autres communautés.

En France, le CC-IN2P3 dispose par exemple de plus de 6 Pétaoctets de stockage disque et de 3 silos robotisés d'une capacité théorique totale de 30 Pétaoctets.

D'une certaine manière, les supers-calculateurs peuvent être vu comme de très grands équipements produisant des données au même titre que le LHC. Les simulations et modélisations réalisées sur les grands calculateurs produisent un flot de données qui doivent être mise à la disposition des communautés scientifiques afin d'être analysées et/ou post-traitées. On estime que dans un avenir proche les simulations météo hautes résolutions tournant sur des calculateurs pétaflopiques vont produire des lots de données atteignant quelques Pétaoctets, donc tout à fait comparables à la physique des hautes énergies.

Des moyens de stockage performants et importants sont déjà disponibles au plus près des supers-calculateurs. L'IDRIS, par exemple dispose de 800 To de disques partagés entre le BlueGene/P et le système Power6, le CINES quant à lui, met à la disposition de ses utilisateurs plusieurs centaines de To. Ce stockage proche et donc performant en termes d'entrée/sortie vers les processeurs devra évoluer dans le futur, en même temps que la puissance des machines, tant en terme de capacité qu'en terme de bande passante

Il n'est, en général, pas optimal de conserver les données produites par les grandes simulations au plus près des calculateurs. Les données doivent pouvoir être distribuée vers les utilisateurs distants. Ce schéma permettra une utilisation souple et optimale des données et déchargera les grands centres plus orientés vers la performance que vers la capacité. Le cas où les données produites doivent faire l'objet de post-traitements sur le super-calculateur qui les a produites, constitue bien sûr, une exception à cette rêgle.

Le type de stockage dépend de la nature et de la destination scientifique des données ainsi que de la répartition géographique des utilisateurs. Les grilles de données (i-RODS par exemple) sont bien adaptées pour des données très distribuées et sujettes à des mouvements fréquents. Des capacités de stockage importantes dans les méso-centres sont également indispensables, soit pour servir la communauté locale, soit en tant que nœud de grille de données.

A l'inverse de la puissance de calcul, il est probable que les capacités de stockage des grilles et des méso-centres deviennent à terme plus importantes que celles disponibles dans les très grands centres de calcul.

L'utilisation à grande échelle de données, suppose une compétence importante dans le domaine des systèmes de fichiers, des protocoles de stockage et de transfert ainsi que dans le secteur des très grandes bases de données. Sur ce dernier point, il est intéressant de noter que des sociétés comme Google sont parties prenantes dans des projets scientifiques tels que le Large Synoptic Survey Telescope (LSST) qui produira des dizaines de Pétaoctets de données et des milliards d'objets.[1]

Recommandations :

- **Faire évoluer les capacités de stockage locales auprès des supers-calculateurs en proportion des augmentations de puissance CPU, tant en terme de capacité qu'en terme de débit.**
- **Mettre en place une politique cohérente de développement du stockage dans les méso-centres et sur les grilles de production.**

[1] http://arxiv.org/pdf/cs/0604112 et http://www.adass.org/adass2006/presentations/kantorj.pdf

- **Renforcer les liens avec la communauté de physique des hautes énergies pour profiter de son savoir faire dans le domaine de la répartition et du traitement distribué des données issus des grands instruments de physique.**
- **Développer les compétences dans le domaine des très grandes bases de données et des très grands systèmes de fichiers distribués**

4. Relations avec les Meso-centres

Les méso-centres sont des structures intermédiaires de calcul et souvent de stockage, dont la puissance, en général, se situe entre celle des centres de calcul nationaux et celle des moyens des équipes ou laboratoires. Ils sont caractérisés par une grande proximité des équipes de recherches d'une part et par la concentration locale, le partage d'expertise, de moyens de calcul et de stockage d'autre part, ce qui permet des économies d'échelle et une performance accrue. Leur accès souple, leur taille intermédiaire, sont essentiels pour la préparation des codes et des calculs qui tourneront sur les très grands calculateurs. Le passage au parallèle massif est souvent difficile, procéder par étapes est indispensable surtout dans une logique de réservation des très grands calculateurs aux très grands calculs. Sur le site du groupe calcul (http://calcul.math.cnrs.fr/), les méso-centres en France sont définis par:

- "Un ensemble de moyens humains, de ressources matérielles et logicielles à destination d'une ou plusieurs communautés scientifiques, issus de plusieurs entités (EPST, Universités, Industriels) en général d'une même région, doté de sources de financement propres, destiné à fournir un environnement scientifique et technique propice au calcul haute performance.
- C'est une structure pilotée par un comité scientifique [...]

Un recensement des différents méso-centres a été effectué par Françoise Berthoud et Violaine Louvet début 2008 et est présenté dans cette page. Les méso-centres, par leur localisation régionale et leur proximité des équipes de recherche des universités, EPST, de l'industrie, accompagnent les projets et la politique scientifique de ces établissements. Le partage de moyens et d'expertise est un support à des collaborations entre différentes disciplines scientifiques et des collaborations entre le monde la recherche scientifique et l'industrie, en particulier les PME. Ce partage permet de concentrer localement de l'expertise en modélisation numérique et calcul intensif et de la diffuser à travers des formations doctorales et permanentes.

Les méso-centres jouent donc un rôle essentiel dans la pyramide du calcul haute performance car ils permettent de diffuser localement l'expertise du calcul intensif, tant au niveau des ingénieurs qui mettent en œuvre et administrent des solutions et des services HPC, des ingénieurs et chercheurs qui forment et se forment aux méthodes du HPC.

Lors de la journée du 13 février 2008 dédiée aux méso-centres et couplée à la réunion d'information sur les grilles de calcul organisée par la CPU (cf. http://calcul.math.cnrs.fr/spip.php?article10), de nombreux représentants des universités ont demandé des meilleurs liens entre les méso-centres et le sommet de la pyramide calcul en France. La construction du GIS Calcul entre le CNRS et la CPU, en relation étroite avec GENCI, constituera, entre autres, une réponse à cette demande. Sa vocation première est de poursuivre les activités « calcul » du groupe calcul (http://calcul.math.cnrs.fr/) mais il offrira aussi une mise en réseau des méso-centre et une interface naturelle avec les GENCI, les centres de calcul nationaux et internationaux.

Nous recommandons un soutien fort à cette mise en réseau des méso-centres afin qu'ils participent, en lien avec les structures nationales et locales, à la diffusion cohérente du HPC en France, dans les Universités, les laboratoires de recherche et les industries locales. La mise en réseaux des méso-centres devrait permettre qu'un utilisateur d'une région puisse accéder au moyens d'une autres région car mieux adapter à la préparation de ses calculs sur une centre de calcul national par exemple. Par l'expertise du calcul intensif qu'ils développent, par les formations qu'ils mettent en œuvre dans les universités, en particulier dans les écoles doctorales à destination des jeunes chercheurs, mais aussi les formations permanentes pour les ingénieurs et chercheurs du secteur public et du privé, par l'accès souple à des moyens de calcul performants et proches des projets de recherches locaux, ils sont un maillon essentiel de la diffusion du HPC.

Nous recommandons la participation du GIS Calcul à la structuration du HPC en France.

5. RELATIONS AVEC LES GRILLES

5.1 Les grilles de productions :

Les Grilles sont des structures de calcul et/ou de stockage distribuées qui permettent de mettre en commun des ressources informatiques hétérogènes. En Europe, les Grilles de production sont souvent liées au projet Européen multidisciplinaire EGEE (Enabling Grid for E-sciencE)[2]. Les architecture de grilles, en plus des ressources informatiques, fournissent une couche de logiciels pour les accéder (intergiciels) ainsi qu'une structure opérationnelle plus ou moins développée, qui garanti la bonne coordination et le bon fonctionnement de l'ensemble. Les grilles de calcul ne sont en principe pas une alternative pour les supercalculateurs qui nécessitent des réseaux à très faibles latences. Par contre, elles offrent une solution efficace pour le stockage, la distribution et la mise à disposition des données issues des grandes simulations. Les moyens de calcul des nœuds de grille sont également bien adaptés pour assurer les pré et post-traitements qui doivent être appliqués sur ces données.

[2] http://www.eu-egee.org/

La physique des hautes énergies qui, avec la mise en route prochaine du LHC, doit faire face à des masses de données considérables (15 PétaOctets/an) a déployé une grille de calcul mondiale (LHC Computing Grid ou LCG) basée en Europe sur l'intergiciel fournit par EGEE. Cette grille est maintenant opérationnelle et compte plus de 80 000 cœurs de CPU et plusieurs dizaines de PétaOctets de stockage sur disques.

Le projet européen EGEE arrive au terme de sa troisième phase en 2010, d'ici là une structure pérenne devrait se mettre en place au niveau européen (European Grid Initiative ou EGI) en s'appuyant sur des grilles nationales (National Grid Initiative ou NGI). La phase préparatoire d'EGI est actuellement financée par l'Europe (EGI Design Study).

Parallèlement à ces grilles dites de production, il existe des grilles de recherche destinées à mettre au point les logiciels et les architectures des grilles de production de demain. En France, le projet Grid5000 et son successeur ALLADIN, qui interconnecte près de 5000 processeurs répartis dans différents sites, fédère l'essentiel des recherches sur les grilles.

En 2008, le CNRS a créé l'Institut des Grilles du CNRS qui rassemble les communautés CNRS travaillant sur les grilles de production et les grilles de recherche.

En dehors d'EGEE, il existe plusieurs grilles de productions basées sur des architectures plus légères et servant des communautés régionales ou thématiques (par exemple CiGri[3] et DIET[4] en Rhône-Alpes).

Il existe également des intergiciels de Grilles orientés vers le transport et le stockage des données telle que SRB[5] et son successeur i-RODS[6] développés à San Diego. Aux États-Unis, le projet BIRN, basé sur SRB / iRODS permet à des laboratoires répartis sur tout le territoire, de mettre en commun des données de Bioinformatiques.

Les communautés des supercalculateurs et des grilles de production sont actuellement assez disjointes. Vu la complémentarité des approches et en particulier le fait que les grilles de production pourraient offrir une solution viable au problème des masses de données produites par les grandes simulations, **nous recommandons de faire en sorte qu'un rapprochement des communautés s'opère**.

5.2Relations entre les Grilles et les méso-centres

Comme indiqué plus haut, les grilles de productions sont essentiellement dédiées au traitement séquentiel des données ainsi qu'à la soumission de tâches

[3] https://ciment.ujf-grenoble.fr/cigri
[4] http://graal.ens-lyon.fr/~diet/
[5] http://www.sdsc.edu/srb/index.php/Main_Page
[6] https://www.irods.org/

paramétriques (très grand nombres de tâches identiques pour lesquelles on fait varier un jeu de paramètres). Rien ne s'oppose en principe à ce qu'une machine à architecture parallèle soit considérée comme un nœud de la grille, le problème principal étant de faire en sorte que le système d'information de la Grille soit capable de caractériser le nœud avec suffisamment de précision afin qu'il puisse être ciblé automatiquement par le mécanisme de soumission de tâches..

Ce problème est en parti résolu dans la grille DEISA qui met en relation les grands calculateurs européens et permet de distribuer la charge. **Nous recommandons de profiter de l'expérience de DEISA afin de permettre aux méso-centres de s'interconnecter et de former une Grille de méso-centres.** Une telle réalisation aurait un rôle structurant dans le sens ou elle favoriserait la coordination des méso-centres au niveau français.

Un modèle dans lequel l'utilisateur a accès à une diversité de calculateurs (du Teraflops au Petaflops) tous distants, dispersés et souvent éloignés, mais doit se préoccuper explicitement de la localisation de ses données, ne semble pas viable à terme. Les technologies de grilles sont les seules à adresser ce problème aujourd'hui. Le rapprochement des communautés supercalculateur et grilles semble inévitable

6. IMPORTANCE DU RESEAU

Les sections précédentes montrent l'importance des données dans l'infrastructure du calcul intensif. Ces données intéressent diverses communautés et ne peuvent matériellement pas restées localisées près des supercalculateurs. La distribution et le traitement des données supposent l'existence d'un réseau performant. A l'inverse, un réseau médiocre rendra caduque tous les efforts réalisés au niveau de la mise en place de supercalculateurs.

Actuellement le réseau de l'enseignement et de la recherche en France est géré par le GIP RENATER. La structure du réseau RENATER 5 est basée sur une ossature de fibres noires (fibres optiques louées et équipées par RENATER) qui permet une grande souplesse dans l'allocation de la bande passante. Les projets scientifiques peuvent ainsi disposer de bandes passantes dédiées si les débits recherchés le justifient. Dans le cadre du calcul intensif, il convient toutefois de prendre garde que la bande passante nécessaire soit disponible de bout en bout jusqu'à l'utilisateur final et non pas seulement entre les grands centres.

Au niveau européen, il est indispensable de maintenir une infrastructure réseau de très haut niveau capable de faire transiter les flux de données entre les supercalculateurs et les centres régionaux. Actuellement le réseau GÉANT construit et opéré par DANTE est d'excellente qualité, il convient de maintenir celle-ci et de s'assurer qu'une bonne coordination entre GÉANT/DANTE et les projets scientifiques persiste.

Nous recommandons :

1. **Que les organismes de recherche maintiennent un bon niveau de financement de RENATER.**

2. **Que les projets liés au calcul intensif évaluent systématiquement et très en amont, leurs besoins en termes de réseaux et maintiennent un haut niveau de coordination avec RENATER.**

3. **Maintenir la qualité du réseau Européen GÉANT et s'assurer que celui-ci évolue en harmonie avec les projets scientifiques.**

7. CALCULATEURS ET ENVIRONNEMENT

La puissance électrique nécessaire pour alimenter les calculateurs n'a fait que croitre durant les dernières années, à tel point que c'est maintenant un problème majeur pour l'ensemble des centres de calcul. Au-delà de l'aspect purement économique, l'impact environnemental est très fort et dois être pris en compte afin de le minimiser.

L'aspect environnemental est évidemment primordial pour les grands centres nationaux et européens qui doivent pour cela mettre en œuvre des solutions techniques sur mesures.

Au niveau des méso-centres et des salles informatiques des laboratoires il convient de suivre les recommandations suivantes :

1 – concernant le choix du matériel

L'objectif ici est de choisir un matériel performant, efficace sur les plans énergétique (VA/Flops) et environnemental (labels éco-conception et politique environnementale du constructeur) : exiger au minimum une efficacité de 90% sur les boitiers d'alimentation, un équivalent « energy star V4 », imposer un critère environnemental parmi les trois premiers critères de choix des appels d'offre qui prenne en compte la dissipation thermique, la puissance effective dans des conditions précises, …

2 – concernant l'emplacement du matériel

L'objectif est ici d'organiser les serveurs de calcul de façon à optimiser les coûts financiers et environnementaux liés à la climatisation. En premier lieu, il s'agira de diminuer le nombre de salles informatiques en mutualisant les salles, en concentrant les serveurs de calculs en optimisant les volumes de serveurs. Les salles choisies devront être « pensées » dans un soucis d'optimisation : équilibrer la répartition des lames et modules dans les châssis, éviter les grands espaces vides dans les armoires, envoyer l'air froid en direction des faces avant des armoires, serrer les armoires pour éviter les tourbillons d'air chaud, concentrer les flux d'air chaud afin d'en faciliter l'évacuation, veiller au bon dimensionnement du système de climatisation, ne pas refroidir la salle informatique en dessous de 25 ° (si la température de la salle est homogène) etc. et dans la mesure du possible coupler le système de refroidissement avec le

chauffage et choisir une orientation nord de la salle machine. Aujourd'hui les constructeurs de baies proposent un refroidissement par eau au plus prêt des serveurs de calcul : cette approche peut être très pertinente. L'utilisation de groupes froids et de clim efficaces en terme énergétique, l'utilisation au plus juste d'onduleurs, ... sont des pistes importantes aussi.

On peut également noter des initiatives innovantes comme le projet ECOCLIM[7] au LPSC (CNRS/IN2P3) qui utilise un système utilisant l'air frais extérieur pour refroidir la salle informatique (« free cooling »).

8. SUPPORT

On pourra prochainement distinguer au moins trois niveaux de centres de calcul : les Tiers0 européens (projet PRACE), les Tiers1 nationaux (GENCI) et les Tiers2 régionaux ou locaux (méso-centres ou centres Départementaux). Quelque soit le niveau considéré, les utilisateurs de ces centres ont besoin de support, et bénéficier ainsi de différents types de services comme :

- L'aide à l'utilisation des moyens de calcul (compilateur, outils de soumission des travaux,…)
- L'optimisation des codes sur les moyens de calcul (utilisation de bibliothèques spécialisées, examen détaillé des performances du code grâce à l'utilisation d'outils spécifiques,…)
- L'aide à la parallélisation et à la parallélisation massive

Le premier service nécessite une connaissance fine du système de calcul que l'on souhaite utiliser, le deuxième également avec en plus la connaissance de techniques génériques d'optimisation. Le troisième service est le plus complexe, il requiert des connaissances solides en techniques numériques et/ou informatiques mais c'est celui qui est requis pour un passage à l'échelle et l'utilisation de supercalculateurs pétaflopiques. Compte tenu de l'ampleur de la tâche (complexité et durée), cela nécessite clairement une implication du ou développeurs du code de calcul, mais l'équipe support peut avantageusement accompagner la démarche grâce à la capitalisation des expériences passées.

Le lien entre Support et Formation est également un élément important : outre sa connaissance pratique du calcul intensif, le support est un puissant vecteur d'accroissement de l'utilisation du calcul intensif en participant à la diffusion du savoir : les centres de calcul nationaux dispensent déjà des formations ouvertes à leurs utilisateurs et à toute la communauté académique.

Le support est déjà bien organisé autour des centres de calcul et le rôle de celui des centres nationaux ne peut être amené qu'à prendre de l'importance suite à l'augmentation de puissance de calcul que l'on connaît, notamment avec la

[7] http://***lpsc***.*in2p3.fr/images-site/****Ecoclim****.pdf*

création de GENCI. Néanmoins, il reste certainement à lui donner une dimension supplémentaire :

- d'une part en prolongeant l'effort fourni par le support des méso-centres en aval,
- d'autre part en organisant l'accélération que va produire l'arrivée programmée des supercalculateurs du projet PRACE en amont.

Proposition : organiser une « chaîne support » du calcul intensif des Tiers2 au Tiers0 en créant un point de rencontre national, intégrant l'ensemble des connaissances et de l'expérience des supports des centres de calcul et participant ainsi à la promotion de l'utilisation du calcul intensif jusque sur les supercalculateurs les plus puissants.