

Parallélisation d'un code de calcul de structure électronique

—
Application aux méthodes norme conservée et PAW

François Bottin^{†,‡,1} et Gilles Zérah^{†,2}

[†]Département de Physique Théorique et Appliquée,
CEA/DAM Ile-de-France,
BP 12, 91680 Bruyères-le-Châtel Cedex, France

[‡]LRC - Centre de Mathématiques et de Leurs Applications,
CNRS (UMR 8536) - ENS Cachan,
61, Avenue du Président Wilson, Cachan Cedex, France

26 janvier 2007

1. Francois.Bottin@cea.fr
2. Gilles.Zerah@cea.fr

Table des matières

Introduction	1
1 Théorie de la Fonctionnelle de la Densité	3
1.1 Hohenberg-Kohn et les fondements de la DFT	3
1.2 Les équations d'Euler et de Kohn et Sham	4
1.3 L'approximation LDA	6
2 Implémentation dans un code ondes planes	9
2.1 Les ondes planes	9
Ondes de Bloch	9
Transformée de Fourier	10
Base variationnelle	10
Quadrillage régulier	10
2.2 Les pseudo-potentiels	11
Pseudisation	11
Densité de cœur	11
Densité de valence	12
Transformation de Kleinman-Bylander: parties locale et non-locale	12
Conservation de la norme	13
2.3 La méthode PAW	13
Théorie	13
Fonction d'onde	14
Densités	15
Charge de compensation	15
L'Énergie	15
Le Hamiltonien	15
3 Mise en œuvre de la parallélisation	17
3.1 Le cycle auto-cohérent	17
3.2 Résolution par blocs d'équations aux valeurs propres	18
Méthode des gradients	18
Méthode LOBPCG	19
3.3 La parallélisation bandFFT	20
3.4 Améliorations et optimisations	22
Généralisation du principe de transposition	22
Librairies mathématiques	23
Diagonalisation dans le sous-espace	23

4	Benchmarks	25
4.1	Détails des calculs	25
4.2	Les calculs NC	26
4.3	Le surcoût PAW	27
	Conclusion	29
	Vers une parallélisation triple n - k - G	29
	Remerciements	29
A	Calcul des termes non-locaux sur une base de PW	31
B	Les harmoniques sphériques complexes et réelles	33
B.1	Définitions	33
B.2	Relations	34

Introduction

Les développements de la physique théorique et les avancées technologiques dans le domaine de la micro informatique ont permis aux calculs de structures électroniques, dits *ab initio*, de connaître un grand essor ces quarante dernières années. Parmi de nombreuses autres méthodes *ab initio*, la théorie de la fonctionnelle de la densité (DFT), élaborée par Hohenberg et Kohn en 1964 (1), présente l'avantage d'être simple à mettre en œuvre, d'être prédictive sur l'ensemble des matériaux, molécules ou espèces atomiques communément étudiées, et enfin d'être utilisable sur des systèmes de très grandes tailles. Cette méthode s'impose donc aujourd'hui au niveau de la recherche comme un outil très puissant, utilisé dans presque tous les domaines de la Physique (voir la Figure 1), mais aussi Chimie et Biologie. Cet engouement se diffuse même jusque dans la recherche-développement, comme dans les laboratoires pharmaceutiques par exemple, où l'achat de programmes DFT "clé en main" permet d'apporter des réponses rapides.

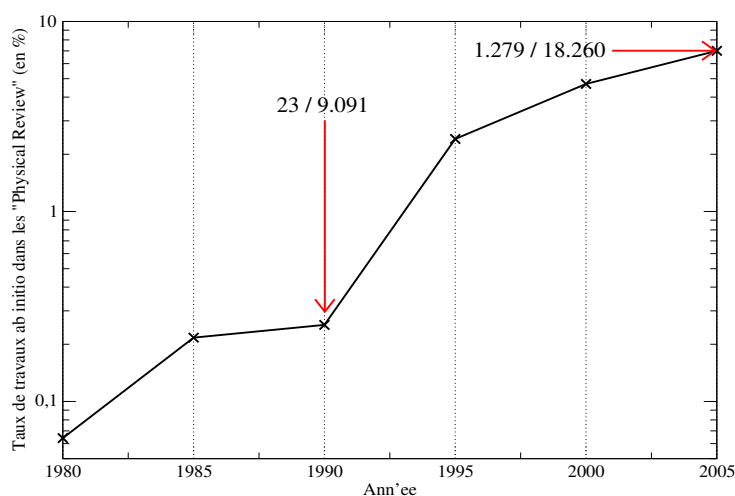


FIG. 1 – Taux d'études *ab initio* dans la revue "Physical Review". Devant les deux flèches sont indiqués le nombre de travaux *ab initio* vs. le nombre de travaux total. Les termes "ab initio", "density functional theory" et "first principles" ont été recherchés dans le titre, le résumé ou les mots clés.

Cette utilisation intensive de la DFT est allée de pair ces dix dernières années avec la construction de machines massivement parallèles. Ces supercalculateurs, permettent aujourd'hui de réaliser des simulations qui demeuraient inenvisageables jusqu'à présent. Certains champs d'investigation, nécessitant des puissances de calcul importantes, peuvent ainsi être explorés. La plupart des domaines sont ainsi concernés, que ce soit la physique de la matière condensée, des liquides, des plasmas.... En sciences des matériaux, par exemple, l'étude de la fusion fait intervenir un grand nombre d'atomes (plusieurs centaines). De même, l'étude des surfaces et des nanostructures nécessite l'utilisation de cellules de simulation de très grandes tailles.

TAB. 1 – Les dix premiers supercalculateurs au monde. Classement du TOP 500 de Novembre 2006.

Organisme	Nom	Processeurs	R_{\max} (Tflops)
DOE/NNSA/LLNL (US)	BlueGene	131072	281
NNSA/Sandia National Laboratories (US)	Red Storm	26544	101
IBM Thomas J. Watson Research Center (US)	BGW	40960	91
DOE/NNSA/LLNL (US)	ASC Purple	12208	76
Bacelona Supercomputing Center (Spain)	Mare Nostrum	10240	63
NNSA/Sandia National Laboratories (US)	Thunderbird	9024	53
CEA/DAM (France)	TERA-10	9968	53
NASA/Ames Research Center/NAS (US)	Columbia	10160	52
GSIC Center (Japan)	TSUBAME	11088	47
Oak Ridge National Laboratories (US)	Jaguar	10424	43

Le CEA, ainsi que quelques laboratoires nationaux (essentiellement américains), sont très engagés au niveau des simulations numériques et se sont en particulier dotés de supercalculateurs possédant plusieurs milliers de processeurs et capables d’atteindre plusieurs dizaines de Teraflops (voir le Tableau 1 et le site <http://www.top500.org>). Dans ces laboratoires, les code de calcul *ab initio*, ont ainsi été profondément remaniés afin d’utiliser à plein les capacités de ces machines. L’introduction d’algorithme de parallélisation plus complexes que précédemment s’est alors avéré nécessaire.

Nous présenterons tout d’abord le cadre théorique dans lequel se place cette étude. Il s’agit de mettre en évidence les principes qui sous-tendent un calcul *ab initio*, fondé sur la DFT, ainsi que d’introduire les méthodes de résolution qui ont été proposées; en particulier les équations de Kohn-Sham. Nous poursuivrons en détaillant les différentes implémentations qui en ont été faites dans le code de calcul ABINIT (2). Celles-ci sont au nombre de deux et se nomment: "Norme Conservée" (NC) et "Projector Augmented-Wave" (PAW). Nous développerons pour chacune les équations définissant les différentes contributions à évaluer. Étant donné que celles-ci diffèrent par le type de base utilisé, cette présentation a pour but d’introduire les différences qui apparaissent entre les deux méthodes. Nous aborderons ensuite l’implémentation proprement-dite de la double parallélisation. En particulier nous nous attacherons à introduire les algorithmes de résolution par blocs ainsi que les transformés de Fourier parallèle. Enfin, nous terminerons par la présentation des résultats obtenus au moyen de cette parallélisation, tant au niveau des méthodes NC que PAW.

Chapitre 1

Théorie de la Fonctionnelle de la Densité

A la fois pour l'intérêt théorique fondamental et pour le développement numérique qui s'en est suivi, Walter Kohn (3) et John A. Pople (4) ont reçu le prix Nobel de chimie en 1998.

to Walter Kohn for his development of the density-functional theory and to John Pople for his development of computational methods in quantum chemistry^{footnotemark} .

Le problème à N corps, tel qu'il est exprimé en physique du solide, correspond au système de N électrons en interaction, eux-mêmes interagissant avec les noyaux. Nous nous plaçons ici dans le cadre de l'approximation de Born-Oppenheimer (5). D'une part, l'énergie cinétique des noyaux est dans un premier temps négligé (les noyaux bougent plus lentement que les électrons), ce qui signifie que les électrons ressentent le potentiel créé par des noyaux placés selon une certaine configuration. D'autre part, si cette géométrie évolue, cela ne peut être fait que par pas de dynamique infinitésimaux (approximation adiabatique), afin de rester sur la même surface d'énergie potentielle. L'énergie cinétique des noyaux n'est alors réintroduite que dans le deuxième cas.

Le Hamiltonien \mathcal{H} du système peut alors s'écrire sous la forme:

$$\mathcal{H} = \mathcal{T}_n + \mathcal{H}_e \quad \text{tel que} \quad \mathcal{H}_e = \mathcal{T}_e + \mathcal{U}_{ne-ne} \quad \text{avec} \quad \mathcal{U}_{ne-ne} = \mathcal{U}_{n-e} + \mathcal{U}_{e-e} + \mathcal{U}_{n-n} \quad (1.1)$$

Les opérateurs énergies cinétiques électroniques et nucléaires sont notées \mathcal{T}_e et \mathcal{T}_n . L'opérateur \mathcal{U}_{ne-ne} traduisant toutes les interactions coulombiennes entre électrons et noyaux a été séparé en trois termes. D'une part, les opérateurs \mathcal{U}_{e-e} et \mathcal{U}_{n-n} traduisant les interactions coulombiennes électron-électron et noyaux-noyaux. D'autre part, l'opérateur \mathcal{U}_{n-e} représentant l'interaction des électrons avec le champ extérieur créé par les noyaux. Il sera noté v_{ext} par la suite.

1.1 Hohenberg-Kohn et les fondements de la DFT

Plusieurs méthodes ont été proposées afin de résoudre ce problème à N corps¹. Les deux théorèmes à la base de la théorie de la fonctionnelle de la densité, ont été établis par Hohenberg et Kohn (1) et sont

0. Nobel citation for chemistry prize.

1. En dehors de la DFT, les méthodes *ab initio* les plus répandues encore aujourd'hui sont: d'une part, la théorie des perturbations (somme de diagrammes de Feynman), la méthode Interaction de Configurations (CI) (développement de la fonction d'onde sur une base de déterminants de Slater) et les simulations Monte-Carlo Quantique qui, en dépit de leurs performances, ne peuvent traiter que des systèmes de petites tailles, et, d'autre part, les méthodes Hartree et Hartree-Fock qui, dans leurs approches initiales, souffrent d'un manque de prédictivité.

les suivants:

Théorème 1 *Le potentiel extérieur $v_{\text{ext}}(\mathbf{r})$ est une fonction biunivoque de la densité électronique $n(\mathbf{r})$.*

Pour un $v_{\text{ext}}(\mathbf{r})$ fixé (i.e.: une configuration de noyaux donnée), l'état fondamental étant univoquement déterminé par sa densité $n(\mathbf{r})$, à partir de l'égalité 1.1 nous pouvons alors écrire l'énergie du fondamental E sous la forme d'une fonctionnelle de la densité:

$$E[n] = \int v_{\text{ext}}(\mathbf{r})n(\mathbf{r})d\mathbf{r} + F[n] \quad (1.2)$$

avec $F[n]$ fonctionnelle universelle de la densité (i.e.: indépendante de $v_{\text{ext}}(\mathbf{r})$), représentant la valeur moyenne des opérateurs $\mathcal{T}_e + \mathcal{U}_{e-e}$ sur l'état de densité $n(\mathbf{r})$. Le terme d'interaction noyaux-noyaux \mathcal{U}_{n-n} ne dépendant pas de la densité électronique, nous n'en tiendrons pas compte dans un premier temps. L'état fondamental du système (densité et énergie) est déterminé par l'intermédiaire du second théorème:

Théorème 2 *Dans un champ extérieur $v_{\text{ext}}(\mathbf{r})$ fixé, l'énergie de l'état fondamental est obtenue en minimisant $E[n]$ par rapport à $n(\mathbf{r})$. La densité électronique $n(\mathbf{r})$ qui correspond au minimum est appelé densité de l'état fondamental.*

Ce théorème définit un principe variationnel, qui peut être écrit sous la forme:

$$E[n] = \min \left[\int v_{\text{ext}}(\mathbf{r})n'(\mathbf{r})d\mathbf{r} + F[n'] \right]_{n'(\mathbf{r})} \quad (1.3)$$

Toute la complexité du problème réside alors dans la définition de la fonctionnelle $F[n]$. Différentes approximations ont été mises en œuvre dont la plus utilisée est sans aucun doute celle de Kohn et Sham.

1.2 Les équations d'Euler et de Kohn et Sham

Soit $F[n]$ la fonctionnelle de la densité précédente. Définissons $T_0[n]$ comme l'énergie cinétique d'un système de N électrons non interagissants et de même densité n . La différence entre l'énergie cinétique du système de N électrons interagissants $T[n]$ et $T_0[n]$ est alors incluse dans le second terme de $F[n]$, appelé $G[n]$, qui contient toutes les interactions entre les N électrons.

$$F[n] = T_0[n] + G[n] \quad (1.4)$$

La fonctionnelle de la densité $E[n]$ peut alors s'écrire:

$$E[n] = \int v_{\text{ext}}(\mathbf{r})n(\mathbf{r})d\mathbf{r} + T_0[n] + G[n] \quad (1.5)$$

En utilisant un formalisme de dérivées fonctionnelles, on peut minimiser l'énergie $E[n]$ (d'après le théorème 2) par rapport à toute variation de la densité $\delta n(\mathbf{r})$, sous la contrainte du nombre d'électrons fixé: $\int n(\mathbf{r})d\mathbf{r} = N$. L'équation d'Euler-Lagrange:

$$\frac{\delta}{\delta n} \left[E[n] - \mu \left(\int n(\mathbf{r})d\mathbf{r} - N \right) \right] = 0 \quad (1.6)$$

devient

$$\begin{aligned} \int \left[v_{\text{ext}}(\mathbf{r}) + \frac{\delta T_0}{\delta n(\mathbf{r})} + \frac{\delta G}{\delta n(\mathbf{r})} \right] \delta n(\mathbf{r})d\mathbf{r} &= \mu \int \delta n(\mathbf{r})d\mathbf{r} \\ v_{\text{ext}}(\mathbf{r}) + \frac{\delta T_0}{\delta n(\mathbf{r})} + \frac{\delta G}{\delta n(\mathbf{r})} &= \mu \end{aligned} \quad (1.7)$$

avec μ le multiplicateur de Lagrange provenant de la contrainte fixant le nombre d'électrons.

Bien que cette équation soit la conséquence directe des deux théorèmes d'Hohenberg-Kohn, elle n'en demeure pas moins inutilisable en pratique car on ne dispose pas d'approximation précise en ce qui concerne par exemple l'énergie cinétique d'un gaz d'électron **inhomogène**. Kohn et Sham (6) ont montré comment déterminer l'état fondamental du système sans résoudre explicitement cette équation fonctionnelle. En procédant en deux étapes:

1. Pour un système de N électrons non interagissant on a $G = 0$. L'équation précédente 1.7 se réduit alors à:

$$\frac{\delta T_0}{\delta n(\mathbf{r})} + v_{\text{ext}}(\mathbf{r}) = \mu \quad (1.8)$$

Dans ce cas, d'une part la densité peut être facilement développée sur la base d'orbitales monoélectroniques Ψ_i , telle que $n(\mathbf{r}) = \sum_i |\Psi_i(\mathbf{r})|^2$, et d'autre part il est possible de développer T_0 en fonction de ces mêmes orbitales, en écrivant²:

$$T_0 = - \sum_i \langle \Psi_i | \frac{\Delta}{2} | \Psi_i \rangle \quad (1.9)$$

Nous sommes ainsi ramenés à la résolution de N équations de Schrödinger à un électron

$$-\frac{\Delta}{2}\Psi_i + v_{\text{ext}}(\mathbf{r})\Psi_i = \epsilon_i\Psi_i \quad (1.10)$$

2. Pour un système de N électrons interagissant, on peut faire une simple analogie avec le cas précédent en introduisant le potentiel effectif $v_{\text{eff}}(\mathbf{r})$

$$v_{\text{eff}}(\mathbf{r}) = v_{\text{ext}}(\mathbf{r}) + \frac{\delta G}{\delta n(\mathbf{r})} \quad (1.11)$$

L'Equation 1.7 devient ainsi de la même forme que l'Équation 1.8:

$$\frac{\delta T_0}{\delta n(\mathbf{r})} + v_{\text{eff}}(\mathbf{r}) = \mu \quad (1.12)$$

En revanche, la densité $n(\mathbf{r})$ est ici celle du système des N électrons interagissants. La suite de l'analogie consiste à faire l'hypothèse qu'il existe un système de N électrons non-interagissants de même densité: $n_0(\mathbf{r}) = n(\mathbf{r})$. Une fois ceci supposé, de la même façon que précédemment la densité peut être développée sur la base d'orbitales monoélectroniques Ψ_i , et il est alors possible d'obtenir N équations à un électron:

$$\mathcal{H}\Psi_i = -\frac{\Delta}{2}\Psi_i + v_{\text{eff}}(\mathbf{r})\Psi_i = \epsilon_i\Psi_i \quad (1.13)$$

Celles-ci sont appelées équations de Kohn et Sham.

Ce sont des équations de Schrödinger à un électron dans lesquelles le potentiel effectif $v_{\text{eff}}(\mathbf{r})$ traduit en pratique toutes les interactions avec les autres électrons. Par conséquent toute la difficulté du problème à N corps a été transférée dans $v_{\text{eff}}(\mathbf{r})$.

Il faut noter qu'au travers de cette transposition physique et mathématique de Kohn et Sham, les ϵ_i et Ψ_i ne sont plus les valeurs propres et les fonctions propres des états i correspondant au système des N électrons interagissants mais celles du système analogique: les N électrons non-interagissants réalisant la même densité. Celles-ci sont par conséquent appelées énergies et fonctions d'ondes de Kohn et Sham, et ne pourront être considérées véritablement comme celles correspondant au système étudié. En principe,

2. Nous utilisons ici les unités atomiques: $\hbar = m_e = e^2/4\pi\epsilon_0 = 1$.

les Ψ_i sont de simples intermédiaires de calculs qui ne sont utiles qu'au calcul de la densité de l'état fondamental du système réel:

$$n(\mathbf{r}) = \sum_i f_i |\Psi_i(\mathbf{r})|^2 \quad (1.14)$$

avec f_i le facteur d'occupation de l'état i . En dépit de ces remarques, les fonctions d'onde Ψ_i et valeurs propres ϵ_i de Kohn et Sham sont en pratique utilisées afin d'avoir une représentation de la nature et de la localisation des états du système étudié.

En revanche, la densité $n(\mathbf{r})$ et l'énergie $E[n]$ (fonctionnelle de cette même densité d'après le théorème 1) ont une réalité physique. Enfin, l'équation de Kohn et Sham doit être résolue de façon auto-cohérente afin de déterminer la densité de l'état fondamental (principe variationnel du théorème 2) puis l'énergie associée.

Remarques: Jusqu'à présent l'approche DFT est **exacte**, pourvu que, d'une part, $n(\mathbf{r})$ soit développable sur les orbitales monoélectroniques (ce qui n'est pas toujours vrai pour certains systèmes pathologiques) et que, d'autre part, l'on ne s'intéresse qu'à l'état fondamental.

1.3 L'approximation LDA

L'énergie $E[n]$ a été décomposée en trois parties: une énergie correspondant à l'interaction de la distribution de charge avec le potentiel extérieur $v_{\text{ext}}(\mathbf{r})$, l'énergie cinétique des N électrons non interagissants $T_0[n]$ et un troisième terme $G[n]$ dans lequel réside toute la difficulté du problème, c'est à dire les interactions entre électrons. Afin de trouver des expressions explicites de G en terme de la densité, séparons à nouveau $G[n]$ en deux parties:

$$G[n] = E_H[n] + E_{XC}[n] \quad (1.15)$$

Le premier terme du membre de droite correspond à la répulsion électrostatique moyenne des N électrons. C'est le terme d'Hartree dans l'approche Hartree-Fock. Celui-ci représente l'interaction directe entre électrons via un potentiel électrostatique créé par la distribution électronique moyenne:

$$E_H = \frac{1}{2} \int \int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}d\mathbf{r}' \quad (1.16)$$

Le second correspond au terme d'échange et corrélation, défini comme la différence entre $G[n]$ et $E_H[n]$, et est noté $E_{XC}[n]$. Celui-ci est composé de deux termes. D'une part, le terme d'échange $E_X[n]$ (appelé aussi terme de Fock) signifie "échange de l'électron entre deux états" et provient de l'antisymétrisation de la fonction d'onde, donc de l'indiscernabilité des électrons dans ces interactions. L'effet produit est d'empêcher la superposition de deux électrons de même spin. Ceci se traduit par l'apparition d'un trou d'échange autour de chaque électron diminuant la répulsion entre les N électrons et stabilisant ainsi la structure. Enfin, notons que le potentiel d'échange est non-local: sa valeur en chaque point dépend de la valeur de la fonction d'onde à laquelle il s'applique en tout point de l'espace. D'autre part, l'énergie de corrélation $E_C[n]$, définie comme la différence $G[n] - E_H[n] - E_X[n]$, trouve son origine physique dans l'impossibilité de décomposer $P(\mathbf{r}_1, \mathbf{r}_2)$ la probabilité de présence conditionnelle d'un électron en \mathbf{r}_1 , compte tenu de la présence d'un autre en \mathbf{r}_2 , en produit de probabilités indépendantes³.

La fonctionnelle de la densité se réécrit alors:

$$E[n] = \int v_{\text{ext}}(\mathbf{r})n(\mathbf{r})d\mathbf{r} + T_0[n] + E_H[n] + E_{XC}[n] \quad (1.17)$$

3. Notons que ce terme n'apparaît pas dans l'approche d'Hartree Fock et que par conséquent ces corrélations ne sont pas prises en compte. Celles-ci pouvant être aussi importantes que l'échange aux faibles densités, ces méthodes ne permettent pas de décrire correctement la liaison chimique. La DFT a été élaborée entre autre dans ce but afin de corriger une partie des erreurs liées à l'approche Hartree-Fock, en incluant de manière approchée les corrélations.

Les trois premiers termes sont donc déterminés (T_0 par l'équation 1.9 et E_H par l'équation 1.16), seul le quatrième est pour l'instant inconnu. L'approximation que nous allons présenter est celle qui a été proposée historiquement et à laquelle les auteurs ne prédisaient pas un grand avenir du fait de sa simplicité. L'approximation locale de la densité ou "Local Density Approximation" (LDA) consiste à prendre la fonctionnelle énergie d'échange et corrélation sous la forme:

$$E_{XC}^{LDA} = \int n(\mathbf{r}) \epsilon_{XC}(n(\mathbf{r})) d\mathbf{r} \quad (1.18)$$

avec $\epsilon_{XC}(n)$ l'énergie d'échange et corrélation par électron d'un gaz d'électrons homogène de densité n . Ceci revient à considérer qu'en chaque point \mathbf{r} de l'espace on suppose que **localement** la distribution électronique $n(\mathbf{r})$ du système d'électrons interagissants puisse être assimilée à celle d'un gaz d'électron homogène de même densité n et non polarisé. Les potentiels correspondant peuvent être obtenus en vue de déterminer le potentiel effectif $v_{\text{eff}}^{LDA}(\mathbf{r})$ qui agit sur les états électroniques du système.

Le potentiel d'Hartree $v_H(\mathbf{r}) = \frac{\delta E_H}{\delta n(\mathbf{r})}$ se calcule facilement à partir de E_H en différenciant fonctionnellement l'Equation 1.16 et devient:

$$v_H(\mathbf{r}) = \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \quad (1.19)$$

En effectuant de même avec l'Equation 1.18, on obtient le potentiel d'échange et corrélation:

$$v_{XC}^{LDA}(\mathbf{r}) = \left[\frac{d}{dn} (n \epsilon_{XC}(n)) \right]_{n=n(\mathbf{r})} \quad (1.20)$$

Finalement, le potentiel effectif de l'équation de Kohn et Sham dans l'approximation LDA s'écrit:

$$v_{\text{eff}}^{LDA}(\mathbf{r}) = v_{\text{ext}}(\mathbf{r}) + v_H(\mathbf{r}) + v_{XC}^{LDA}(\mathbf{r}) \quad (1.21)$$

Plusieurs remarques peuvent être faites au sujet de cette approximation:

1. Tout d'abord, celle-ci suppose (au travers de l'Equation 1.18), que le terme d'échange et corrélation est local. Ceci est en contradiction avec la non-localité de l'échange exact.
2. L'énergie d'échange et corrélation $\epsilon_{XC}(n)$ d'un gaz d'électron homogène n peut être obtenue par une technique de Monte-Carlo (7), et différentes paramétrisations existent. Les plus connues sont celles de Kohn-Sham (6) (première à être introduite), Perdew et Zunger (8) ou Perdew-Wang (9). Toutes vérifient la loi de somme⁴ correcte pour le trou d'échange et corrélation.

4. L'énergie d'échange et corrélation peut se réécrire $E_{XC}^{LDA} = \frac{1}{2} \int \int \frac{n(\mathbf{r})n_{XC}(\mathbf{r},\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' d\mathbf{r}$ avec n_{XC} la densité du trou d'échange et corrélation. Celle-ci vérifie alors la loi de somme $\int n_{XC}(\mathbf{r},\mathbf{r}') d\mathbf{r} = -1$.

Chapitre 2

Implémentation dans un code ondes planes

L'objet de ce chapitre est donner une expression des termes densités, potentiels et énergies précédents tels qu'ils sont implémentés dans un code de calcul ondes-planes (PW). Leur expression nous permettra de mettre en évidence les termes nécessitant une attention particulière, car lourds à évaluer, ainsi que de souligner les différences qui existent entre les deux implémentations numériques existantes.

Les fonctions d'ondes monoélectroniques Ψ_i , solutions des équations de Kohn-Sham, sont développées sur une base. Deux grandes catégories de bases existent: les bases localisées et les bases délocalisées. Certaines méthodes sont exclusives et n'utilisent que l'une ou l'autre, essentiellement des orbitales atomiques ou des gaussiennes dans le premier cas et des PW dans le second. D'autres ne procèdent pas de cette façon et mélangent les deux bases (localisée et délocalisée), la première décrivant le cœur et la seconde la région interstitielle: ce sont les méthodes de type "Linear Muffin-Tin Orbital" (LMTO) et "Linear Augmented Plane-Wave" (LAPW) (10; 11; 12) ou bien "Projector Augmented-Wave" (PAW) (13).

2.1 Les ondes planes

La base d'ondes planes, délocalisée, est la plus couramment utilisée en physique du solide. Elle a en effet prouvé à plusieurs reprises son efficacité et sa facilité d'utilisation (14). Les avantages les plus régulièrement mis en évidence pour une base de PW sont de plusieurs ordres:

Ondes de Bloch Le premier provient de la signification physique d'une base de PW. En effet, selon le théorème de Bloch (15), la périodicité des fonctions d'ondes du système est déterminée par la périodicité du réseau cristallin. Les fonctions d'ondes de Bloch s'écrivent naturellement, sous la forme d'une somme de PW:

$$\Psi_{nk}(\mathbf{r}) = \sqrt{\frac{1}{\Omega}} \sum_{\mathbf{G}} c_{nk}(\mathbf{G}) e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} \quad (2.1)$$

avec \mathbf{G} les vecteurs du réseau réciproque, \mathbf{k} les vecteurs d'onde appartenant à la première zone de Brillouin, n l'index de bande et Ω le volume du système. Au niveau numérique, la périodicité du système réel est prise en compte par l'utilisation conjointe d'une cellule unité de simulation (super-cellule) et des conditions aux bords périodiques (PBC).

Transformée de Fourier Le second provient de la facilité d'utilisation d'une base de PW. En effet, si les fonctions d'ondes sont développées sur une base de PW, il est alors possible de passer facilement de l'espace réel à l'espace réciproque (et inversement) en utilisant les transformées de Fourier rapides tridimensionnelles (3dim-FFT). L'utilité d'une telle procédure (changement d'espace) peut provenir par exemple de la nécessaire transformation d'un produit de convolution en un produit simple ou de la possibilité de transformer une interaction à longue portée dans un espace, donc difficile à calculer, en une interaction à courte portée dans l'autre espace. Par exemple, en définissant $n(\mathbf{G}) = \frac{1}{\Omega} \int e^{-i\mathbf{G}\cdot\mathbf{r}} n(\mathbf{r}) d\mathbf{r}$ comme la transformée de Fourier directe de la densité électronique (voir Équation 1.14)

$$n(\mathbf{r}) = \sum_{\mathbf{nk}} f_{\mathbf{nk}} |\Psi_{\mathbf{nk}}(\mathbf{r})|^2 \quad (2.2)$$

on peut réécrire le potentiel¹ et l'énergie d'Hartree de la manière suivante:

$$E_H = \frac{\Omega}{2} \sum_{\mathbf{G}} \frac{4\pi}{G^2} |n(\mathbf{G})|^2 \quad \text{et} \quad v_H(\mathbf{r}) = \sum_{\mathbf{G}} \frac{4\pi}{G^2} n(\mathbf{G}) e^{i\mathbf{G}\cdot\mathbf{r}} \quad (2.3)$$

L'énergie cinétique s'exprime elle aussi dans l'espace réciproque,

$$E_K = \sum_{\mathbf{nk}} f_{\mathbf{nk}} \langle \tilde{\Psi}_{\mathbf{nk}} | -\frac{\Delta}{2} | \tilde{\Psi}_{\mathbf{nk}} \rangle = \sum_{\mathbf{nk}} f_{\mathbf{nk}} \sum_{\mathbf{G}} \frac{(\mathbf{k} + \mathbf{G})^2}{2} |c_{\mathbf{nk}}(\mathbf{G})|^2 \quad (2.4)$$

alors qu'il est commode de calculer les contributions d'échange et corrélation dans l'espace direct (voir les Équations 1.18 et 1.20).

Base variationnelle Le troisième avantage provient de la simplicité avec laquelle la base de PW utilisée numériquement peut être contrôlée. En effet, dans les codes de calcul, le développement 2.1 doit être tronqué et seul un certain nombre de PW est pris en compte. Cependant, cette méthode (qui consiste à écrire la fonction d'onde comme un développement tronqué) est variationnelle puisqu'il est possible d'augmenter sûrement et régulièrement la précision du calcul en ajoutant des PW et donc de s'approcher de la solution numérique exacte dans un espace à dimension infinie. Le nombre d'ondes planes considéré, déterminant la précision du calcul, peut être défini par l'intermédiaire de l'énergie cinétique d'un électron exprimée dans l'espace réciproque. Seules les PW associées à une énergie cinétique inférieure à une énergie appelée "de cutoff" E_{cut} telle que:

$$\frac{\hbar^2 \mathbf{G}^2}{2m_e} \leq E_{\text{cut}} \quad (2.5)$$

appartiennent à la base. La dernière PW prise en compte, de vecteur du réseau réciproque $|\mathbf{G}_{\text{max}}| = \frac{N_x 2\pi}{L_x}$, avec L_x la taille de la super-cellule dans la direction x , définit le nombre N_x de plans de PW dans cette direction. Le nombre total N_{tot} de PW compris dans la boîte définissant l'espace réciproque s'obtient alors facilement: $N_{\text{tot}} = N_x N_y N_z$.

Notons qu'en réalité toutes les PW de la boîte ne sont pas prises en compte. En effet, seules celles d'énergie inférieure ou égales à l'énergie de cutoff E_{cut} contribuent. Les PW conduisant à un coefficient $c_{\mathbf{nk}}(\mathbf{G})$ non-nul sont comprises dans une sphère de rayon $|\mathbf{G}_{\text{max}}|$ et de centre \mathbf{k} qui est inscrit dans la boîte.

Quadrillage régulier Enfin, le quatrième avantage des PW provient de la très bonne description de l'espace qui est réalisée. En effet, les PW quadrillent entièrement le système étudié, qu'il y ait du vide au pas, alors qu'une base d'orbitale localisée est attachée à l'atome. De plus, contrairement aux bases

1. Le potentiel électrostatique est par définition un potentiel à longue portée dans l'espace réel. Il devient à courte portée dans l'espace réciproque et peut donc être plus facilement intégré.

localisées, il n'est pas nécessaire de prendre en compte, lors de calculs de dynamique moléculaire ou d'optimisation de géométrie, certaines forces supplémentaires appelées "de Pulay" (16).

Nous venons de détailler certains des avantages qui apparaissent dans un code de calcul lorsqu'une fonction d'onde est développée sur une base d'onde plane. Cette dernière, telle qu'elle est écrite dans l'Équation 2.1, caractérise l'état d'indice de bande n et de vecteur de Bloch \mathbf{k} . La densité totale $n(\mathbf{r})$ s'obtenant par l'intégration du module carré de cette quantité sur tous les vecteurs d'onde de la première zone de Brillouin \mathbf{k} , une nouvelle approximation a été nécessaire. L'intégration sur les points \mathbf{k} a été remplacée par une somme finie sur un nombre restreint de points \mathbf{k} (voir l'Équation 2.2) appartenant à la partie irréductible de la première zone de Brillouin (BZI). Ils sont appelés "points spéciaux" et ont été introduits par Chadi-Cohen (17). Une méthode permettant leur génération a ensuite été proposée par Monkhorst et Pack (18).

2.2 Les pseudo-potentiels

L'utilisation d'une base de PW nécessite de recourir à l'utilisation d'un pseudo-potentiel. Leur formulation mathématique ainsi que la procédure de génération dans le cas général n'est pas l'objet de cette sous-section et un article de revue récent peut être consulté dans ce but (19).

Pseudisation Derrière le terme de pseudisation peut se cacher en réalité trois approximations. La première, est la régularisation du potentiel crée par le noyau. En éliminant la divergence du potentiel coulombien à l'origine, et en adoucissant ses variations, l'énergie de cutoff peut être abaissée. La seconde est l'approximation de "cœur gelé" qui correspond au gel des orbitales de cœur trop profondes pour participer à la liaison chimique. La troisième provient de l'intégration de ce cœur gelé à l'intérieur du potentiel crée par le noyau. La région de cœur étant celle qui présente les plus fortes variations de densité, en ne les traitant plus explicitement dans le calcul, ceci engendre une nouvelle réduction de l'énergie de cutoff.

La prodédure de pseudisation est indépendante du code de calcul *ab initio*. Celle-ci le précède et prend en compte tous les électrons de l'atome isolé considéré. Le calcul effectué est sphérique et toutes les contributions sont donc obtenues à une dimension. Ainsi, le potentiel extérieur de l'atome v_{ext} devient un pseudo-potentiel \tilde{v}_{ext} , le Hamiltonien \mathcal{H} un pseudo-Hamiltonien $\tilde{\mathcal{H}}$, les fonctions d'ondes de valence Ψ_{nk} des pseudo-fonctions d'onde $\tilde{\Psi}_{\text{nk}}$ et la densité $n(\mathbf{r})$ une pseudo-densité $\tilde{n}(\mathbf{r})$.

Densité de cœur Il est cependant parfois utile et même souhaitable d'utiliser la densité de cœur, même gelée, dans les calculs *ab initio*. Elle peut avoir un effet important, par exemple en pression. Celle-ci, est alors pseudisée et est évaluée dans l'espace de Fourier en sommant les contributions provenant de tous les atomes I:

$$\tilde{n}_c(\mathbf{G}) = \frac{1}{\Omega} \int e^{-i\mathbf{G}\cdot\mathbf{r}} \tilde{n}_c(\mathbf{r}) d\mathbf{r} = \frac{1}{\Omega} \int e^{-i\mathbf{G}\cdot\mathbf{r}} \sum_I \tilde{n}_c^I(\mathbf{r} - \mathbf{R}_I) d\mathbf{r} \quad (2.6)$$

$$= \frac{1}{\Omega} \sum_I e^{-i\mathbf{G}\cdot\mathbf{R}_I} \int \tilde{n}_c^I(\mathbf{r}) e^{-i\mathbf{G}\cdot\mathbf{r}} d\mathbf{r} = \frac{1}{\Omega} \underbrace{\sum_I e^{-i\mathbf{G}\cdot\mathbf{R}_I}}_{S_{\text{tr}}(\mathbf{G})} \underbrace{\int \frac{\sin(\mathbf{G}\mathbf{r})}{\mathbf{G}\mathbf{r}} \tilde{n}_c^s(\mathbf{r}) 4\pi r^2 d\mathbf{r}}_{n_c^s(\mathbf{G})} \quad (2.7)$$

Nous avons ici supposé que tous les atomes I étaient identiques, de densité de cœur pseudisée sphérique: $\tilde{n}_c^s(\mathbf{r})$. La quantité $S_{\text{tr}}(\mathbf{G})$ correspond au facteur de structure qui lui aussi est évalué dans l'espace réciproque. La densité de cœur s'exprime alors dans l'espace réel sous la forme:

$$\tilde{n}_c(\mathbf{r}) = \sum_{\mathbf{G}} \frac{S_{\text{tr}}(\mathbf{G})}{\Omega} \tilde{n}_c^s(\mathbf{G}) e^{i\mathbf{G}\cdot\mathbf{r}} \quad (2.8)$$

Sa contribution à l'échange et corrélation est alors ajoutée *a posteriori* afin de déterminer l'énergie $E_{XC}[\tilde{n} + \tilde{n}_c]$ et le potentiel $v_{XC}(\mathbf{r})$ qui en dérive. Cette fonctionnelle n'étant pas linéaire en fonction de la densité, une procédure particulière, appelée "corrections non-linéaires de cœur", est alors appliquée (voir S.G. Louie *et al.* (20)).

Densité de valence Les électrons conservés dans la valence remplissent les nouveaux états de Kohn-Sham, et ceci à partir du nouveau fondamental déterminé par le pseudo-Hamiltonien. Soulignons que la pseudo-fonction d'onde correspondant au fondamental est sans nœud, même si l'état physique auquel elle correspond en possède. Par conséquent, les pseudo-fonctions d'onde présentent moins de variations que les véritables fonctions d'onde et permettent ainsi d'utiliser une énergie de cutoff raisonnable. Enfin, en ne conservant pour chaque atome que les électrons de valence, la pseudisation permet de réduire le nombre total des états du système lors de la résolution auto-cohérente des équations de Kohn-Sham et donc de diminuer le coût des calculs de façon significative.

Transformation de Kleinman-Bylander: parties locale et non-locale Décomposons le pseudo-potentiel en somme de contributions provenant de chaque atome I. Cela revient donc à réécrire le potentiel extérieur sous la forme: $\tilde{v}_{\text{ext}}(\mathbf{r}) = \sum_I \tilde{v}_{\text{ext}}^I(\mathbf{r} - \mathbf{R}_I)$. Par la suite, nous traiterons chaque atome de manière indépendant et isolé. Les pseudo-fonctions d'onde $\tilde{\Phi}_{nl}^I(\mathbf{r})$ et le pseudo-potentiel $\tilde{v}_{\text{ext}}^I(\mathbf{r})$ de chacun est obtenu en inversant l'équation de Schrödinger radiale:

$$\left[-\frac{d^2}{2dr^2} + \frac{l(l+1)}{2r^2} + \tilde{v}_{\text{ext}}^I(r) - \epsilon_{nl}^I\right] \frac{\tilde{\Phi}_{nl}^I(r)}{r} = 0 \quad (2.9)$$

Nous avons posé que $\tilde{\Phi}_{nlm}^I(\mathbf{r}) = \tilde{\Phi}_{nl}^I(r)Y_{lm}(\theta, \phi)$, avec Y_{lm} les harmoniques sphériques et (θ, ϕ) les variables angulaires (voir l'Annexe B). Le pseudo-potentiel issu de l'inversion de cette équation est radial et présente une dépendance explicite en l . Les électrons s, p, d et f ressentiront donc un pseudo-potentiel extérieur différent, celui-ci pouvant s'écrire en toute généralité: $\tilde{v}_{\text{ext}}^I(\mathbf{r}) = \sum_l \tilde{v}_{\text{ext},l}^I(r)$. Il est cependant commode de décomposer ce pseudo-potentiel extérieur d'une autre manière:

$$\tilde{v}_{\text{ext}}^I(\mathbf{r}) = \tilde{v}_{\text{loc}}^I(\mathbf{r}) + \sum_l \tilde{v}_{\text{nl},l}^I(r)\tilde{\mathcal{P}}_l^I \quad (2.10)$$

où nous avons défini une partie locale $\tilde{v}_{\text{loc}}^I(\mathbf{r})$ de moment angulaire $l = l_{\text{loc}}$ et des parties angulaires non-locales $\tilde{v}_{\text{nl},l}^I(r) = \tilde{v}_{\text{ext},l}^I(r) - \tilde{v}_{\text{loc}}^I(r)$ obtenues en projetant le pseudo-potentiel sur chaque moment angulaire au moyen du projecteur $\tilde{\mathcal{P}}_l^I$. Le premier terme peut être aisément calculé puisque local. En sommant les contributions provenant de tous les atomes I, nous obtenons l'énergie locale et le potentiel local associé sous la forme:

$$E_{\text{loc}} = \Omega \sum_{\mathbf{G}} \tilde{v}_{\text{loc}}(\mathbf{G})\tilde{n}(\mathbf{G}) \quad \text{et} \quad \tilde{v}_{\text{loc}}(\mathbf{r}) = \sum_{\mathbf{G}} \frac{S_{\text{tr}}(\mathbf{G})}{\Omega} \tilde{v}_{\text{loc}}^s(\mathbf{G})e^{i\mathbf{G}\cdot\mathbf{r}} \quad (2.11)$$

Comme pour la densité de cœur, nous avons ici supposé que tous les atomes I étaient identiques, de pseudo-potentiel local sphérique: $\tilde{v}_{\text{loc}}^s(\mathbf{r})$. Le second terme de l'équation 2.10, que nous nommerons par commodité $\tilde{v}_{\text{NL}}^I(\mathbf{r})$, est plus délicat à traiter car il est local en \mathbf{r} mais non-local en (θ, ϕ) . On parle alors de potentiel semi-local. Ce terme est par définition lourd à évaluer dans un code de calcul puisqu'il faut procéder à un produit de convolution. Une transformation, appelée de Kleinman-Bylander (21), permet de transformer cette partie semi-locale du potentiel en un terme totalement non-local dans l'espace réel (c'est à dire totalement local dans l'espace réciproque). On parle alors de potentiel séparable. Dans le cadre qui nous intéresse, cette partie non-locale du pseudo-potentiel, telle qu'elle est exprimée au travers de l'Équation 2.10, peut se réécrire:

$$\tilde{v}_{\text{NL}}^I(\mathbf{r}) = \sum_{lm \in I} \frac{|\tilde{\mathcal{P}}_{lm}^I\rangle\langle\tilde{\mathcal{P}}_{lm}^I|}{\langle\tilde{\mathcal{P}}_{lm}^I|\tilde{\Phi}_{lm}^I\rangle} \quad (2.12)$$

avec $|\tilde{\mathcal{P}}_{lm}^I\rangle = |\tilde{v}_{\text{loc},I}^I\tilde{\Phi}_{lm}^I\rangle$ les projecteurs². L'énergie non-locale, associé à cette partie du potentiel, s'écrit ainsi:

$$E_{\text{NL}} = \sum_{\text{nk}} \sum_I \sum_{lm \in I} \frac{\langle \tilde{\Psi}_{\text{nk}} | \tilde{\mathcal{P}}_{lm}^I \rangle \langle \tilde{\mathcal{P}}_{lm}^I | \tilde{\Psi}_{\text{nk}} \rangle}{\langle \tilde{\mathcal{P}}_{lm}^I | \tilde{\Phi}_{lm}^I \rangle} \quad (2.13)$$

Ces contributions non-locales à l'énergie totale et au potentiel effectif des équations de Kohn-Sham (voir les équations 1.13 et 1.21) sont très lourdes à évaluer dans un code PW. Le calcul de ces termes est effectué dans l'espace réciproque (voir l'annexe A) et de nombreuses études ont été réalisées afin de rendre ce calcul le plus efficace possible (la Référence (22) peut être consultée à ce sujet). Nous verrons, dans certains cas, que le temps passé à calculer ces contributions correspond à près de 20% du temps total.

Nous concluons ce paragraphe en rassemblant toutes les contributions intervenant dans le Hamiltonien:

$$\tilde{\mathcal{H}} = -\frac{\Delta}{2} + \tilde{v}_{\text{loc}} + v_{\text{XC}} + v_{\text{H}} + \sum_I \sum_{lm \in I} \frac{|\tilde{\mathcal{P}}_{lm}^I\rangle \langle \tilde{\mathcal{P}}_{lm}^I|}{\langle \tilde{\mathcal{P}}_{lm}^I | \tilde{\Phi}_{lm}^I \rangle} \quad \text{avec} \quad \tilde{\mathcal{H}}\tilde{\Psi}_{\text{nk}} = \epsilon_{\text{nk}}\tilde{\Psi}_{\text{nk}} \quad (2.14)$$

Conservation de la norme Dans le cas appelé "à norme conservée", les pseudo-fonctions d'onde doivent respecter plusieurs contraintes au cours de la procédure de pseudisation (voir par exemple les méthodes de Hamann, Schlüter et Chiang (23) ou de Troullier et Martins (24)). D'une part, à l'intérieur d'une sphère de rayon R_1^I , les pseudo-fonctions d'onde doivent donner lieu à la même densité que les vraies fonctions d'ondes, même si elles n'ont pas les mêmes variations spatiales (puisque les nœuds ont été supprimés):

$$\int_0^{R_1^I} |\phi_{nl}(r)|^2 r^2 dr = \int_0^{R_1^I} |\tilde{\phi}_{nl}(r)|^2 r^2 dr \quad (2.15)$$

D'autre part, à l'extérieur de la sphère, les pseudo-fonctions d'onde doivent être identiques aux vraies fonctions d'onde. La pseudo-fonction d'onde et la vraie la fonction d'onde sont ainsi raccordées en R_1^I (ainsi que les quatre premières dérivées) pour chaque moment angulaire l .

Dans le cas où cette contrainte est relâchée (voir par exemple les pseudo-potentiels ultra-soft de Vanderbilt (25) ou la méthode PAW de Blöchl (13; 26)), les densités électroniques sont alors encore plus douces et un gain substantiel en temps de calcul est réalisé puisque la valeur de l'énergie de cutoff est fortement diminuée (le nombre de PW effectivement pris en compte dans le calcul est fortement réduit). Nous présentons cette méthode dans la section suivante.

2.3 La méthode PAW

Le but de la méthode PAW introduite par Blöchl (13), est de reconstruire la vraie fonction d'onde Ψ_{nk} , avec toute sa structure nodale, à partir de la fonction d'onde auxiliaire $\tilde{\Psi}_{\text{nk}}$; cette dernière convergeant rapidement dans un développement en PW.

Théorie Cette reconstruction peut schématiquement s'écrire sous la forme $\Psi_{\text{nk}} = \mathcal{T}\tilde{\Psi}_{\text{nk}}$, avec \mathcal{T} l'opérateur permettant de réaliser cette transformation. On peut aisément se convaincre que la fonction d'onde auxiliaire étant égale à la vraie fonction d'onde en dehors d'une sphère de rayon R_1^I il n'est nécessaire de réaliser cette transformation qu'à l'intérieur. Il s'agit donc d'ajouter pour chaque atome la différence entre la vraie fonction d'onde et la fonction auxiliaire. Les ondes partielles physiques Φ_1^I de chaque atome I , solutions de l'équation de Schrödinger pour l'atome isolé, ainsi que les ondes partielles auxiliaires $\tilde{\Phi}_1^I$, égales aux premières en dehors de la sphère et plus douces en-dedans, sont utilisées comme bases dans ce but.

2. Notons que la sommation intervenant dans l'équation 2.12 n'est réellement effectuée que sur l

On peut montrer que la transformation linéaire permettant de reconstruire la vraie fonction d'onde s'écrit:

$$\mathcal{T} = \mathcal{I} + \sum_{\mathbf{I}} \sum_{\mathbf{i} \in \mathbf{I}} (|\Phi_{\mathbf{i}}^{\mathbf{I}}\rangle - |\tilde{\Phi}_{\mathbf{i}}^{\mathbf{I}}\rangle) \langle \tilde{\mathcal{P}}_{\mathbf{i}}^{\mathbf{I}}| \quad (2.16)$$

avec $\tilde{\mathcal{P}}_{\mathbf{i}}^{\mathbf{I}}$ les fonctions projecteurs assurant le caractère local de la transformation (ils sont localisés à l'intérieur de la sphère). Notons qu'il existe une relation de dualité entre les fonction projecteurs et fonctions d'onde partielles auxiliaires tel que $\langle \tilde{\mathcal{P}}_{\mathbf{i}}^{\mathbf{I}} | \tilde{\Phi}_{\mathbf{j}}^{\mathbf{I}} \rangle = \delta_{\mathbf{ij}}$. L'indice \mathbf{i} se réfère ici aux nombres quantiques $l_{\mathbf{i}}$ et $m_{\mathbf{i}}$ ainsi qu'à un indice supplémentaire $n_{\mathbf{i}}$ différenciant les ondes partielles d'un site ayant les mêmes nombres quantiques angulaires³.

Dans le cadre de cette transformation, la valeur moyenne d'un opérateur \mathcal{A} sur l'état correspondant à la vraie fonction d'onde $\Psi_{\mathbf{nk}}$, devient:

$$\langle \mathcal{A} \rangle = \sum_{\mathbf{nk}} f_{\mathbf{nk}} \langle \tilde{\Psi}_{\mathbf{nk}} | \mathcal{A} | \tilde{\Psi}_{\mathbf{nk}} \rangle + \sum_{\mathbf{I}} \sum_{\mathbf{ij} \in \mathbf{I}} \rho_{\mathbf{ij}}^{\mathbf{I}} (\langle \Phi_{\mathbf{j}}^{\mathbf{I}} | \mathcal{A} | \Phi_{\mathbf{i}}^{\mathbf{I}} \rangle - \langle \tilde{\Phi}_{\mathbf{j}}^{\mathbf{I}} | \mathcal{A} | \tilde{\Phi}_{\mathbf{i}}^{\mathbf{I}} \rangle) \quad (2.17)$$

avec $\rho_{\mathbf{ij}}^{\mathbf{I}}$ la "matrice densité" à un centre définie comme:

$$\rho_{\mathbf{ij}}^{\mathbf{I}} = \sum_{\mathbf{nk}} f_{\mathbf{nk}} \langle \tilde{\Psi}_{\mathbf{nk}} | \tilde{\mathcal{P}}_{\mathbf{i}}^{\mathbf{I}} \rangle \langle \tilde{\mathcal{P}}_{\mathbf{j}}^{\mathbf{I}} | \tilde{\Psi}_{\mathbf{nk}} \rangle \quad (2.18)$$

Cette "matrice densité" met en évidence les facteurs d'occupation des ondes partielles à l'intérieur de la sphère PAW de l'atome \mathbf{I} .

Fonction d'onde En utilisant l'Équation 2.16, la vraie fonction d'onde peut s'écrire:

$$|\Psi_{\mathbf{nk}}\rangle = |\tilde{\Psi}_{\mathbf{nk}}\rangle + \sum_{\mathbf{I}} \sum_{\mathbf{i} \in \mathbf{I}} (|\Phi_{\mathbf{i}}^{\mathbf{I}}\rangle - |\tilde{\Phi}_{\mathbf{i}}^{\mathbf{I}}\rangle) \langle \tilde{\mathcal{P}}_{\mathbf{i}}^{\mathbf{I}} | \tilde{\Psi}_{\mathbf{nk}} \rangle \quad (2.19)$$

Hors de la sphère, la vraie fonction d'onde est égale à la fonction d'onde auxiliaire car $\tilde{\Phi}_{\mathbf{i}}^{\mathbf{I}} = \Phi_{\mathbf{i}}^{\mathbf{I}}$. A l'intérieur, la contribution auxiliaire de la fonction d'onde $\tilde{\Psi}_{\mathbf{nk}}$ est supprimée tandis que la contribution physique, avec tous les nœuds de la fonction d'onde, est ajoutée.

Les fonctions d'ondes physiques $\Psi_{\mathbf{nk}}$ sont ortho-normales et vérifient donc l'égalité $\langle \Psi_{\mathbf{nk}} | \Psi_{\mathbf{n}'\mathbf{k}'} \rangle = \delta_{\mathbf{n},\mathbf{n}'} \delta_{\mathbf{k},\mathbf{k}'}$. En utilisant la loi de transformation (voir l'équation 2.16), cette relation devient:

$$\langle \tilde{\Psi}_{\mathbf{nk}} | \mathcal{T}^{\dagger} \mathcal{T} | \tilde{\Psi}_{\mathbf{n}'\mathbf{k}'} \rangle = \delta_{\mathbf{n},\mathbf{n}'} \delta_{\mathbf{k},\mathbf{k}'} \quad (2.20)$$

Ainsi, les fonctions d'onde auxiliaires $\tilde{\Psi}_{\mathbf{nk}}$ ne sont plus ortho-normales. La relation d'ortho-normalité est recouverte par l'intermédiaire d'un opérateur \mathcal{O} appelé d'"overlap" tel que:

$$\mathcal{O} = \mathcal{T}^{\dagger} \mathcal{T} = \mathcal{I} + \sum_{\mathbf{I}} \sum_{\mathbf{ij} \in \mathbf{I}} |\tilde{\mathcal{P}}_{\mathbf{i}}^{\mathbf{I}}\rangle (\langle \Phi_{\mathbf{i}}^{\mathbf{I}} | \Phi_{\mathbf{j}}^{\mathbf{I}} \rangle) - \langle \tilde{\Phi}_{\mathbf{i}}^{\mathbf{I}} | \tilde{\Phi}_{\mathbf{j}}^{\mathbf{I}} \rangle \langle \tilde{\mathcal{P}}_{\mathbf{j}}^{\mathbf{I}}| \quad (2.21)$$

Cette même transformation peut être effectuée sur les vraies équations de Kohn-Sham (voir l'équation 1.13). Celles-ci deviennent des équations de Kohn-Sham généralisées $\tilde{\mathcal{H}} \tilde{\Psi}_{\mathbf{nk}} = \epsilon_{\mathbf{nk}} \mathcal{O} \tilde{\Psi}_{\mathbf{nk}}$. Contrairement au cas "norme conservée" (voir l'équation 2.14), un opérateur d'"overlap" apparaît dans le membre de droite et ce sont donc des équations de Schrödinger avec second membre qu'il faut résoudre.

3. Contrairement à la transformation de Kleinman-Bylander exposée précédemment, notons qu'il est ici possible d'avoir plusieurs projecteurs par moment angulaire. Ceci permet d'obtenir une meilleur description du système en PAW, le résultat exact n'étant obtenu qu'à complétude de la base, donc avec un nombre de projecteurs infini.

Densités Au moyen de l'équation 2.17 définissant la valeur moyenne d'un opérateur et de l'équation 2.2 définissant la densité, on peut écrire:

$$n(\mathbf{r}) = \tilde{n}(\mathbf{r}) + \sum_{\mathbf{I}} (n^{1,\mathbf{I}}(\mathbf{r}) - \tilde{n}^{1,\mathbf{I}}(\mathbf{r})) \quad (2.22)$$

avec respectivement, $n^{1,\mathbf{I}}$ et $\tilde{n}^{1,\mathbf{I}}$ les densités définies à l'intérieur de la sphère:

$$n^{1,\mathbf{I}}(\mathbf{r}) = \sum_{ij \in \mathbf{I}} \rho_{ij}^{\mathbf{I}} \Phi_i^{\mathbf{I}}(\mathbf{r}) \Phi_j^{\mathbf{I}}(\mathbf{r}) \quad \text{et} \quad \tilde{n}^{1,\mathbf{I}}(\mathbf{r}) = \sum_{ij \in \mathbf{I}} \rho_{ij}^{\mathbf{I}} \tilde{\Phi}_i^{\mathbf{I}}(\mathbf{r}) \tilde{\Phi}_j^{\mathbf{I}}(\mathbf{r}) \quad (2.23)$$

Les fonction d'ondes, auxiliaires ou vraies, ainsi que les fonctions projecteurs utilisées dans ces équations sont développées sur une base d'harmoniques sphériques réelles $S_{lm}(\hat{r})$ (voir l'Annexe B), avec:

$$\Phi_i^{\mathbf{I}}(\mathbf{r}) = \frac{\Phi_{n_i l_i}^{\mathbf{I}}(r)}{r} S_{l_i m_i}(\hat{r}) ; \quad \tilde{\Phi}_i^{\mathbf{I}}(\mathbf{r}) = \frac{\tilde{\Phi}_{n_i l_i}^{\mathbf{I}}(r)}{r} S_{l_i m_i}(\hat{r}) ; \quad \tilde{\mathcal{P}}_i^{\mathbf{I}}(\mathbf{r}) = \frac{\tilde{\mathcal{P}}_{n_i l_i}^{\mathbf{I}}(r)}{r} S_{l_i m_i}(\hat{r}) \quad (2.24)$$

Ainsi, les densités $n^{1,\mathbf{I}}$ et $\tilde{n}^{1,\mathbf{I}}$ sont évaluées sur une grille sphérique, en séparant les parties radiales et angulaires pour alléger les calculs, tandis que la densité auxiliaire \tilde{n} reste calculée sur la grille FFT.

Charge de compensation Les densités auxiliaires \tilde{n} et $\tilde{n}^{1,\mathbf{I}}$ ne respectent pas le critère de conservation de la norme (voir la section précédente). En relâchant cette contrainte, les variations de densités sont rendues plus douces et la base de PW utilisée est ainsi réduite (voir par exemple les pseudo-potentiels ultra-soft de Vanderbilt (25)). Afin de reproduire correctement le développement multipolaire de la vraie densité de charge, il est nécessaire d'introduire une charge de compensation \hat{n} . Celle-ci est ajoutée aux charges auxiliaires à chaque fois que celles-ci interviennent dans le calcul.

La charge de compensation est définie par l'équation intégrale suivante:

$$\int_{\Omega_{\mathbf{R}}} (n^{1,\mathbf{I}} - \tilde{n}^{1,\mathbf{I}} - \hat{n}^{\mathbf{I}}) |\mathbf{r} - \mathbf{R}|^L S_{LM}(\widehat{\mathbf{r} - \mathbf{R}}) d\mathbf{r} = 0 \quad (2.25)$$

avec $\Omega_{\mathbf{R}}$ le volume de la sphère PAW. A chaque atome \mathbf{I} , correspond une charge de compensation $\hat{n}^{\mathbf{I}}$. Nous définirons pour la suite $\hat{n} = \sum_{\mathbf{I}} \hat{n}^{\mathbf{I}}$.

L'Énergie En utilisant une nouvelle fois l'équation 2.17 nous pouvons déterminer l'énergie totale. Comme la vraie fonction d'onde Ψ_{nk} et la vraie densité $n(\mathbf{r})$, cette quantité peut être divisée en trois parties:

$$E = \tilde{E} + \sum_{\mathbf{I}} (E^{1,\mathbf{I}} - \tilde{E}^{1,\mathbf{I}}) \quad (2.26)$$

La terme \tilde{E} correspond à la partie "norme conservée" et est calculé sur la grille FFT, tandis que les deux autres termes sont spécifiques à la méthode PAW et sont évalués sur la grille sphérique. Ces trois termes s'expriment de la façon suivante:

$$\begin{cases} \tilde{E} = \sum_{nk} f_{nk} \langle \tilde{\Psi}_{nk} | -\frac{\Delta}{2} | \tilde{\Psi}_{nk} \rangle + E_{xc}[\tilde{n} + \hat{n} + \tilde{n}_c] + E_H[\tilde{n} + \hat{n}] + \int \tilde{v}_{\text{ext}}^{\mathbf{I}}(\tilde{n} + \hat{n}) d\mathbf{r} + U_{n-n} \\ E^{1,\mathbf{I}} = \sum_{ij \in \mathbf{I}} \rho_{ij}^{\mathbf{I}} \langle \Phi_i^{\mathbf{I}} | -\frac{\Delta}{2} | \Phi_j^{\mathbf{I}} \rangle + E_{xc}[\tilde{n}^{1,\mathbf{I}} + n_c] + E_H[\tilde{n}^{1,\mathbf{I}}] + \int_{\Omega_{\mathbf{R}}} v_{\text{ext}}^{\mathbf{I}} \tilde{n}^{1,\mathbf{I}} d\mathbf{r} \\ \tilde{E}^{1,\mathbf{I}} = \sum_{ij \in \mathbf{I}} \rho_{ij}^{\mathbf{I}} \langle \tilde{\Phi}_i^{\mathbf{I}} | -\frac{\Delta}{2} | \tilde{\Phi}_j^{\mathbf{I}} \rangle + E_{xc}[\tilde{n}^{1,\mathbf{I}} + \hat{n} + \tilde{n}_c] + E_H[\tilde{n}^{1,\mathbf{I}} + \hat{n}] + \int_{\Omega_{\mathbf{R}}} \tilde{v}_{\text{ext}}^{\mathbf{I}}(\tilde{n}^{1,\mathbf{I}} + \hat{n}) d\mathbf{r} \end{cases}$$

Le Hamiltonien En minimisant l'énergie précédente par rapport à toute variation de la fonction d'onde auxiliaire $\delta|\tilde{\Psi}_{nk}\rangle$, sous la contrainte que ces fonctions d'onde auxiliaires soient \mathcal{O} -ortho-normalisées,

$$\frac{\delta}{\delta|\tilde{\Psi}_{nk}\rangle} \left[\tilde{E}[\tilde{n}] - \mu \sum_{nk} \sum_{n'k'} (\langle \tilde{\Psi}_{nk} | \mathcal{O} | \tilde{\Psi}_{n'k'} \rangle - \delta_{n,n'} \delta_{k,k'}) \right] = 0 \quad (2.27)$$

il est alors possible d'obtenir le Hamiltonien qui intervient dans les équations de Kohn-Sham généralisées. Celui-ci a alors la forme suivante:

$$\tilde{\mathcal{H}} = -\frac{\Delta}{2} + \tilde{v}_{\text{ext}} + v_{\text{XC}} + v_{\text{H}} + \sum_{\mathbf{I}} \sum_{ij \in \mathbf{I}} |\tilde{\mathcal{P}}_{\mathbf{i}}^{\mathbf{I}}\rangle D_{ij}^{\mathbf{I}} \langle \tilde{\mathcal{P}}_{\mathbf{j}}^{\mathbf{I}}| \quad \text{avec} \quad \tilde{\mathcal{H}}\tilde{\Psi}_{\mathbf{nk}} = \epsilon_{\mathbf{nk}}\mathcal{O}\tilde{\Psi}_{\mathbf{nk}} \quad (2.28)$$

Nous ne développerons pas ici le terme $D_{ij}^{\mathbf{I}}$ qui apparaît dans l'équation ci-dessus. Au travers de cette équation, le but est de mettre en évidence les similitudes et différences qui apparaissent entre les calculs NC et PAW. (i) Les Hamiltoniens (voir les équations 2.14 et 2.28) ont à première vue la même forme; ils sont composés d'un terme local et d'un autre "non-local". (ii) En revanche, le terme "non-local" PAW comporte une sommation sur $n_i l_i m_i n_j l_j m_j \mathbf{I}$, là où ce terme ne comporte qu'une sommation sur $l m \mathbf{I}$ en NC. (iii) Contrairement au cas NC, le terme $D_{ij}^{\mathbf{I}}$ est non-diagonal et couple différents moments angulaires i et j . (iv) De plus, alors que la transformation de Kleinman-Bylander ne permet d'utiliser qu'un seul projecteur par moment angulaire, il est ici possible d'en introduire plusieurs (les indices n_i et n_j). (v) Ajoutons que, dans les calculs PAW, un second membre est aussi à déterminer (le terme \mathcal{O}) pour résoudre les équations de Kohn-Sham et assurer l'orthonormalisation des vraies fonctions d'onde.

Ces différences auront leur importance sur les temps de calcul que nous présenterons (voir le chapitre 4).

Chapitre 3

Mise en œuvre de la parallélisation

Si les calculs *ab initio*, réalisés sur des cellules de simulation comprenant une dizaine d’atomes, sont effectivement possibles sur un seul processeur, en revanche les études conjuguant dynamiques moléculaires et système comprenant plusieurs centaines d’atomes restent aujourd’hui hors d’atteinte en mono-processeur. Le nombre de bandes (donc d’équations de Kohn-Sham) et de PW rendent ces calculs très gourmands en temps de simulation et deviennent les principaux facteurs limitants. Avec l’augmentation de la puissance et du nombre de processeurs des super-calculateurs, la parallélisation des calculs de structure électronique devient donc un véritable enjeu. En particulier, un effort a été consacré, ces dernières années, à la résolution des équations de Kohn-Sham ainsi qu’aux transformés de Fourier, afin d’obtenir des algorithmes efficaces sur des machines massivement parallélisées.

En ce qui concerne le premier, il s’agit d’algorithmes de résolution des équations aux valeurs propres fonctionnant par blocs et permettant donc de résoudre simultanément (en parallèle) un certain nombre d’équations de Kohn-Sham, (voir les équations 2.14 en NC et 2.28 en PAW). L’utilisation de ce type d’algorithme rend ainsi possible l’utilisation d’une **parallélisation sur les bandes**, chaque bloc de bandes étant ainsi distribué sur plusieurs processeurs. Dans le cas du second, les 3dim-FFT parallèles permettent de distribuer aussi sur plusieurs processeurs toutes les quantités (potentiels, énergies...) qui s’expriment sur une base de PW et d’effectuer leur calcul simultanément. Nous renvoyons le lecteur au chapitre précédent afin de décompter le nombre de termes qui sont développés sur une telle base de PW, et donc de mesurer l’importance que peut avoir une telle **parallélisation sur les FFT**.

Afin de tirer avantage de ces deux parallélisations, nous avons mis en œuvre dans le code de calcul ABINIT (2) une **double parallélisation** (27). Le principal enjeu est de les faire coexister sans perdre d’efficacité; i.e.: sans y sacrifier trop de communications. Nous développerons donc par la suite l’implémentation qui a été choisie, en détaillant chacune des parallélisations séparément, et mettrons enfin en évidence les différentes astuces utilisées pour rendre ces deux parallélisations compatibles.

3.1 Le cycle auto-cohérent

Les équations de Kohn-Sham, (voir les équations 2.14 en NC et 2.28 en PAW) sont couplées par l’intermédiaire de la densité électronique. Les fonctions d’ondes auxiliaires $\tilde{\Psi}_{nk}$ sont à la fois sorties et entrées (par l’intermédiaire de la densité) de ces équations. Ce n’est donc qu’au travers d’une boucle auto-cohérente (SCF) qu’elles peuvent être résolues.

Dans le but de garder le problème aussi général que possible nous traiterons le cas PAW. Le calcul NC se déduit alors en remplaçant le terme d’"overlap" \mathcal{O} par l’opérateur identité, en changeant de "terme non-local", en n’introduisant plus de charge de compensation \hat{n} et en ne calculant plus la "matrice densité" ρ_{ij} .

Dans le diagramme suivant, nous indiquons schématiquement les étapes essentielles du cycle SCF:

$$\begin{array}{ccc}
 \tilde{\Psi}_{\mathbf{nk}}(\mathbf{r}) = \sum_{\mathbf{G}} c_{\mathbf{nk}}(\mathbf{G}) e^{i(\mathbf{k}+\mathbf{G}) \cdot \mathbf{r}} & & \\
 \downarrow & & \\
 [\tilde{n} + \hat{n}](\mathbf{r}) \quad \text{et} \quad \rho_{ij} & \longleftarrow & \tilde{\Psi}_{\text{prev}} + \kappa \tilde{\Psi}_{\text{new}} \\
 \downarrow & & \uparrow \\
 \tilde{v}_{\text{ext}} + v_{\text{XC}} + v_{\text{H}} + \sum_{\text{I}} \sum_{ij \in \text{I}} |\tilde{\mathcal{P}}_i^{\text{I}}\rangle D_{ij}^{\text{I}} \langle \tilde{\mathcal{P}}_j^{\text{I}}| & & \left\{ \tilde{\mathcal{H}} | \tilde{\Psi}_{\mathbf{nk}} \rangle = \epsilon_{\mathbf{nk}} \mathcal{O} | \tilde{\Psi}_{\mathbf{nk}} \rangle \right\}_{\mathbf{nk}} \\
 \downarrow & & \uparrow \\
 \langle e^{i(\mathbf{k}+\mathbf{G}) \cdot \mathbf{r}} | \tilde{\mathcal{H}} | \tilde{\Psi}_{\mathbf{nk}} \rangle = \epsilon_{\mathbf{nk}} \langle e^{i(\mathbf{k}+\mathbf{G}) \cdot \mathbf{r}} | \mathcal{O} | \tilde{\Psi}_{\mathbf{nk}} \rangle & \longrightarrow & c_{\mathbf{nk}}(\mathbf{G})
 \end{array}$$

Le calcul débute par l'initialisation de la fonction d'onde $\tilde{\Psi}_{\mathbf{nk}}(\mathbf{r})$. Celle-ci est ensuite utilisée pour déterminer la densité $[\tilde{n} + \hat{n}](\mathbf{r})$ et la matrice densité ρ_{ij} . Ces quantités sont nécessaires pour évaluer l'ensemble des potentiels dits locaux¹ $\tilde{v}_{\text{ext}} + v_{\text{XC}} + v_{\text{H}}$ et le terme non-local $|\tilde{\mathcal{P}}_i^{\text{I}}\rangle D_{ij}^{\text{I}} \langle \tilde{\mathcal{P}}_j^{\text{I}}|$. Les équations de Kohn-Sham généralisées, projetées sur les PW, sont ensuite résolues afin d'obtenir les coefficients $c_{\mathbf{nk}}(\mathbf{g})$ des fonctions d'ondes auxiliaires. Une diagonalisation dans le sous-espace $\{\mathbf{nk}\}$, de dimension $\mathbf{nk} \times \mathbf{nk}$, est alors réalisée afin d'obtenir les vecteurs-propres et valeurs-propres exactes dans cette espace et non-plus approchée (voir l'algorithme LOBPCG). Les fonctions d'ondes auxiliaires entrantes et sortantes de ce cycle sont ensuite mélangées afin d'éliminer les nombreuses instabilités numériques inhérentes à ce type de résolution. La minimisation cesse quand une erreur inférieure à un critère de tolérance choisi est obtenue.

Lorsque la taille du système augmente, plusieurs parties de ce cycle SCF deviennent très gourmandes en temps de calcul. Il s'agit: (i) du calcul des vecteurs-propres et valeurs-propres dans le le sous-espace, (ii) de l'appel aux routines qui effectuent les 3dim-FFT (calcul de la densité et application de l'ensemble des potentiels dits locaux aux fonctions d'onde auxiliaires), (iii) de l'évaluation du terme non-local et (iv) enfin, de la résolution proprement dite des équations de Kohn-Sham.

Presque tous ces calculs peuvent être traités point- \mathbf{k} par point- \mathbf{k} (voir le cycle SCF). Seuls les calculs de densités et de potentiels nécessitent de connaître tous les points- \mathbf{k} . Une importante partie du code peut donc être parallélisée sur les points- \mathbf{k} . Ceux-ci sont nombreux lorsqu'on considère un métal et une boîte de simulation petite (la BZI est alors très grande²) mais se réduisent au point Γ lorsque le système est très grand ou désordonné. La parallélisation sur les points- \mathbf{k} , bien que très efficace lorsqu'on peut l'utiliser, s'avère inutile dès que la taille du système augmente.

3.2 Résolution par blocs d'équations aux valeurs propres

Il s'agit, dans cette section, de déterminer l'ensemble des vecteurs-propres Ψ_m et valeurs-propres ϵ_m (avec $1 \leq m \leq M$ et M le nombre de bandes du système) en résolvant le problème aux valeurs-propres généralisé suivant:

$$\mathcal{H}\Psi = \epsilon \mathcal{O}\Psi \quad (3.1)$$

avec \mathcal{H} et \mathcal{O} des matrices $M \times M$ symétriques définies positives.

Méthode des gradients La résolution de ces équations dans ABINIT était jusqu'à présent effectuée au moyen d'une méthode appelée de Gradient Conjuguée (28) (voir aussi (29) et les références à l'intérieur). Si l'on ne cherche que la valeur-propre ϵ_1 de plus basse énergie, le problème consiste à minimiser la grandeur:

$$\epsilon(\Psi) = \frac{\langle \Psi | \mathcal{H} | \Psi \rangle}{\langle \Psi | \mathcal{O} | \Psi \rangle} \quad (3.2)$$

1. Cette terminologie définit le potentiel local mais aussi l'échange et corrélation ainsi que Hartree.

2. Voir la section 2.1.

appelée quotient de Rayleigh. La méthode de gradient conjugué (CG) est itérative (avec i l'indice repérant l'itération) et fait appel à la méthode de gradient à pas optimal:

algorithme 1 Méthode de gradient à pas optimal

Entrées: Soit $\Psi_1^{(0)}$ une approximation du vecteur-propre Ψ_1 .

- 1: **pour** $i=0,1,\dots$ **faire**
 - 2: Calcul de $\nabla\epsilon(\Psi_1^{(i)})$.
 - 3: Minimisation de ϵ sur $\Xi = \{\Psi_1^{(i)}, \nabla\epsilon(\Psi_1^{(i)})\}$.
 - 4: $\Psi_1^{(i+1)}$ est la nouvelle approximation de Ψ_1 .
 - 5: **fin pour**
-

La minimisation de ϵ est effectuée au moyen de la méthode de Rayleigh-Ritz (30; 31). Celle-ci permet de déterminer, dans une base ortho-normale approchée Ξ , les approximations $\Psi_1^{(i)}, \dots, \Psi_n^{(i)}$ de Ψ_1, \dots, Ψ_n . En ce qui concerne la méthode de gradient conjugué, on obtient donc:

algorithme 2 Méthode de gradient conjugué

Entrées: Soit $\Psi_1^{(0)}$ une approximation du vecteur-propre Ψ_1 . Après une itération de la méthode de gradient à pas optimal,

- 1: **pour** $i=1,2,\dots$ **faire**
 - 2: Calcul de $\nabla\epsilon(\Psi_1^{(i)})$.
 - 3: Minimisation de ϵ sur $\Xi = \{\Psi_1^{(i-1)}, \Psi_1^{(i)}, \nabla\epsilon(\Psi_1^{(i)})\}$.
 - 4: $\Psi_1^{(i+1)}$ est la nouvelle approximation de Ψ_1 .
 - 5: **fin pour**
-

Ces deux méthodes utilisent le gradient du quotient de Rayleigh $\nabla\epsilon(\Psi)$, que l'on peut facilement obtenir en différenciant $\epsilon(\Psi)$ par rapport à $|\Psi\rangle$ et $\langle\Psi|$:

$$\nabla\epsilon(\Psi) = 2 \frac{\mathcal{H}\Psi - \epsilon(\Psi)\mathcal{O}\Psi}{\langle\Psi|\mathcal{O}|\Psi\rangle} \quad (3.3)$$

Comme nous venons de le montrer cette méthode ne permet de déterminer qu'un vecteur propre à la fois, c'est à dire bande-par-bande. L'algorithme est adapté aux calculs séquentiels mais ne permet pas de paralléliser le calcul sur les bandes.

Méthode lobpcg Afin de résoudre plusieurs équations de Kohn-Sham simultanément, une nouvelle méthode a du être implémentée. Il s'agit de la méthode LOBPCG (Locally Optimal Block Preconditioned Conjugate Gradient) introduite par Knyazev (32). Celle-ci s'appuie sur l'algorithme CG pour déterminer les valeurs-propres ϵ .

Knyazev part de la constatation suivante. La base employée dans CG est mal-conditionnée, $\Psi^{(i-1)}$ devenant proche de $\Psi^{(i)}$ lorsque i tend vers l'infini. Un nouveau vecteur $P^{(i)}$, défini à partir du résidu $R^{(i)}$ et permettant d'obtenir une meilleure description de Ξ lors de la minimisation, a donc été introduit par Knyazev.

$$P^{(i+1)} = \lambda^{(i)}R^{(i)} + \gamma^{(i)}P^{(i)} \quad \text{avec} \quad R^{(i)} = \mathcal{H}\Psi^{(i)} - \epsilon^{(i)}\mathcal{O}\Psi^{(i)} \quad (3.4)$$

Dans le sous-espace $\Xi = \{\mathbf{P}^{(i)}, \Psi^{(i)}, \nabla \epsilon(\Psi^{(i)})\}$, les nouveaux vecteurs de Ritz deviennent ainsi:

$$\Psi^{(i+1)} = \delta^{(i)} \Psi^{(i)} + \lambda^{(i)} \mathbf{R}^{(i)} + \gamma^{(i)} \mathbf{P}^{(i)} \quad (3.5)$$

avec $\delta^{(i)}$, $\lambda^{(i)}$ et $\gamma^{(i)}$ les coefficients évalués au cours de la minimisation. Dans le but d'améliorer les performances (d'augmenter la rapidité de convergence), Knyazev a recours à un préconditionnement. Ce procédé n'est pas propre à LOBPCG et peut être aussi employé dans l'algorithme CG. Un preconditionneur \mathbf{K} , optimisé pour les PW, est donc introduit. Le résidu préconditionné

$$\mathbf{W}^{(i)} = \mathbf{K} \mathbf{R}^{(i)} = \mathbf{K}(\mathcal{H} \Psi^{(i)} - \epsilon^{(i)} \mathcal{O} \Psi^{(i)}) \quad (3.6)$$

est alors utilisé à la place de $\mathbf{R}^{(i)}$. Celui-ci remplace aussi le gradient du quotient de Rayleigh³ pour décrire le sous-espace Ξ .

Nous venons ici de décrire la méthode pour une bande d'indice m et de vecteur-propre Ψ_m . Celle-ci peut être facilement généralisée pour un nombre arbitraire m de bandes, avec M le nombre total de bandes (commensurable à m). Les M fonctions d'onde sont ainsi scindées en $\frac{M}{m}$ blocs de taille m s'écrivant $\Psi = \{\Psi_1, \dots, \Psi_m\}$. À l'intérieur de la boucle sur les blocs, chaque bloc est contraint d'être orthogonal aux précédents et est itéré un nombre fixé de fois κ . L'algorithme devient ainsi:

algorithme 3 Algorithme LOBPCG

Entrées: Soient $\Psi^0 = \{\Psi_1^0, \dots, \Psi_m^0\}$ un bloc de fonctions d'onde proches du minimum et un preconditionneur \mathbf{K} ; le bloc $\mathbf{P} = \{\mathbf{P}_1^{(0)}, \dots, \mathbf{P}_m^{(0)}\}$ est initialisé à 0.

- 1: **pour** $i=0, 1, \dots, \kappa$ **faire**
 - 2: $\Upsilon^{(i)} = \Upsilon(\Psi^{(i)})$
 - 3: $\mathbf{R}^{(i)} = \mathcal{H} \Psi^{(i)} - \Upsilon^{(i)} \mathcal{O} \Psi^{(i)}$
 - 4: $\mathbf{W}^{(i)} = \mathbf{K} \mathbf{R}^{(i)}$
 - 5: La méthode de Rayleigh-Ritz est ensuite appliquée à l'intérieur du sous-espace Ξ décrit par $\{\mathbf{P}_1^{(i)}, \dots, \mathbf{P}_m^{(i)}, \Psi_1^{(i)}, \dots, \Psi_m^{(i)}, \mathbf{W}_1^{(i)}, \dots, \mathbf{W}_m^{(i)}\}$
 - 6: $\Psi^{(i+1)} = \Delta^{(i)} \Psi^{(i)} + \Lambda^{(i)} \mathbf{W}^{(i)} + \Gamma^{(i)} \mathbf{P}^{(i)}$
 - 7: $\mathbf{P}^{(i+1)} = \Lambda^{(i)} \mathbf{W}^{(i)} + \Gamma^{(i)} \mathbf{P}^{(i)}$
 - 8: **fin pour**
-

Nous avons défini différents blocs dans cet algorithme: les lettres grecques majuscules $\Upsilon = \{\epsilon_1, \dots, \epsilon_m\}$, $\Lambda = \{\lambda_1, \dots, \lambda_m\}$, $\Delta = \{\delta_1, \dots, \delta_m\}$, $\Gamma = \{\gamma_1, \dots, \gamma_m\}$ ainsi que toutes les quantités en gras.

Nous venons donc de présenter l'algorithme LOBPCG. Les premières études que nous avons réalisées en séquentiel montrent que celui-ci, en plus d'être parallélisable sur les bandes, présente une rapidité de convergence plus importante que la méthode CG. Ce résultat est empirique et ne s'appuie sur aucune considération théorique. Certaines études sont actuellement en cours pour améliorer cet algorithme⁴ (temps de calcul plus faible, meilleure parallélisation...).

3.3 La parallélisation bandFFT

Nous présentons dans cette section, la parallélisation proprement dite. Celle-ci est intimement liée à la distribution des coefficients des fonctions d'onde $c_{nk}(\mathbf{G})$ sur les processeurs.

3. $\nabla \epsilon(\Psi^{(i)})$ fait en réalité intervenir le résidu (voir équation 3.4).

4. La version que nous avons décrite de cet algorithme est nommée LOBPCG I. Il existe par exemple une autre version nommée LOBPCG II qui est l'objet d'une étude.

Soit une grille bi-dimensionnelle de processeurs, avec les processeurs nommés "FFT" le long de la direction x et les processeur nommés "bandes" le long de la direction y. Nous supposons que le système est décrit par P plans FFT le long d'une des trois directions de la grille du réseau réciproque $\{\mathbf{G}\}$ et par M bandes. Nous omettrons dans cette présentation l'indice k pour ne pas allourdir la notation.

Considérons p "processors FFT" (tel que p soit commensurable avec P) et m "processeurs bandes" (comme précédemment, m est commensurable avec M). Le nombre total de processeurs utilisés est donc: $m \times p$. Supposons qu'au départ, la fonction d'onde ne soit définie que sur un seul processeur; i.e.: tous les coefficients $c_n(\mathbf{G})$ des M bandes (avec $1 \leq n \leq M$) sont sur le processeur 0×0 :

$$\left\{ \begin{array}{cccc} c_n(\mathbf{G}) & - & - & - \\ - & - & - & - \\ - & - & - & - \\ - & - & - & - \end{array} \right\}$$

En premier lieu, les P plans de la grille $\{\mathbf{G}\}$, donc les coefficients $c_n(\mathbf{G})$, sont successivement distribués sur les p "processeurs FFT".

$$\left\{ \begin{array}{cccc} c_n(\mathbf{G}_1) & c_n(\mathbf{G}_2) & \dots & c_n(\mathbf{G}_p) \\ - & - & - & - \\ - & - & - & - \\ - & - & - & - \end{array} \right\}$$

Les ensembles de plans $\{\mathbf{G}_a\}$ avec $a=1,2,\dots,p$ rétablissent la grille du réseau réciproque $\{\mathbf{G}\}$. Ces plans sont ensuite sub-divisés une nouvelle fois en distribuant les coefficients PW sur les m "processeurs bandes":

$$\left\{ \begin{array}{cccc} c_n(\mathbf{G}_{11}) & c_n(\mathbf{G}_{21}) & \dots & c_n(\mathbf{G}_{p1}) \\ c_n(\mathbf{G}_{12}) & c_n(\mathbf{G}_{22}) & \dots & c_n(\mathbf{G}_{p2}) \\ \vdots & \vdots & \ddots & \vdots \\ c_n(\mathbf{G}_{1m}) & \dots & \dots & c_n(\mathbf{G}_{pm}) \end{array} \right\}$$

Les ensembles de sous-plans $\{\mathbf{G}_{ab}\}$ avec $b=1,2,\dots,m$ permettent de recouvrir les ensembles de plans $\{\mathbf{G}_a\}$. C'est la distribution en PW utilisée pour calculer les potentiels lors de la première étape du cycle SCF (voir la section 3.1). Ces quantités sont toutes développées sur une base de PW (voir la section 2.1) et sont donc facilement évaluées. Chaque processeur calcule ainsi sa contribution aux équations de Kohn-Sham. Comme nous allons le voir, il n'est pas utile de la communiquer aux autres processeurs.

Dans la routine LOBPCG, cette distribution ne change pas et les coefficients sont scindés en blocs de taille m (il y a donc $\frac{M}{m}$ blocs).

$$\left\{ \begin{array}{cccc} c_{1:m}(\mathbf{G}_{11}) & c_{1:m}(\mathbf{G}_{21}) & \dots & c_{1:m}(\mathbf{G}_{p1}) \\ c_{1:m}(\mathbf{G}_{12}) & c_{1:m}(\mathbf{G}_{22}) & \dots & c_{1:m}(\mathbf{G}_{p2}) \\ \vdots & \vdots & \ddots & \vdots \\ c_{1:m}(\mathbf{G}_{1m}) & \dots & \dots & c_{1:m}(\mathbf{G}_{pm}) \end{array} \right\}$$

Ainsi les équations de Kohn-Sham sont résolues sur une portion de l'espace réciproque et non pas dans l'espace réciproque tout entier. Cette distribution est adaptée à LOBPCG car elle permet d'effectuer, sur chaque processeur (sans communication avec les autres), les produits matrice-matrice et matrice-vecteur intervenant dans la méthode (voir l'algorithme 3). Nous la nommerons "distribution LOBPCG".

Nous pouvons cependant remarquer que cette distribution n'est pas appropriée au calcul des transformés de Fourier parallèles. Celles-ci interviennent lors de la résolution des équations de Kohn-Sham puisque les "potentiels locaux" doivent être appliqués aux fonctions d'ondes auxiliaires (voir la section 3.1). Notons qu'il en va de même lors du calcul de la densité en début de cycle SCF.

Les 3dim-FFT parallèles exigent que toute la grille du réseau réciproque $\{\mathbf{G}\}$ soit connue sur une ligne de processeur. Celle-ci est pour l'instant répartie sur l'ensemble de la grille bi-dimensionnelle de processeurs. La distribution des coefficients est donc modifiée localement pour effectuer ce calcul. Nous

transposons les éléments placés à l'intérieur de chaque colonne, ce qui nous permet d'obtenir la "distribution FFT" suivante:

$$\begin{pmatrix} c_1(\mathbf{G}_1) & c_1(\mathbf{G}_2) & \dots & c_1(\mathbf{G}_p) \\ c_2(\mathbf{G}_1) & c_2(\mathbf{G}_2) & \dots & c_2(\mathbf{G}_p) \\ \vdots & \vdots & \ddots & \vdots \\ c_m(\mathbf{G}_1) & \dots & \dots & c_m(\mathbf{G}_p) \end{pmatrix}$$

Ainsi chaque bande du bloc est distribuée sur les p "processeurs FFT". Notons que la communication effectuée correspond à un `MPI_ALLTOALL` en utilisant le communicateur des "processeurs bandes". Cette transformation n'est donc pas globale à la grille bi-dimensionnelle de processeurs, mais locale à chacune des colonnes. Ceci à un effet positif non négligeable sur le temps de communication.

La distribution des PW est à présent compatible avec l'utilisation de 3dim-FFT parallèles. Celle de Goedecker *et al.* (33), implémentée dans le code de calcul ABINIT, est adaptée pour réaliser ce type de calcul. Cette approche a en effet fait ses preuves sur des super-calculateurs massivement parallélisés car elle présente plusieurs avantages: (i) une variable, définissant le cache du processeur, peut être défini en dur afin d'optimiser l'algorithme, (ii) le nombre de communications intervenant dans la méthode a été réduit de 2 à 1, de plus sur une quantité de données plus faible, (iii) les coefficients $c_{nk}(\mathbf{G})$ qui sont identiquement nuls, car au-delà de la sphère de cutoff (voir la discussion suivant l'équation 2.5), ne sont pas pris en compte dans le calcul. Notons enfin que les communications intervenant dans la méthode de Goedecker *et al.* sont locales à chacune des lignes (et non pas globales à la grille bi-dimensionnelle de processeurs).

Ainsi, m 3dim-FFT sont donc appliquées simultanément sur les m lignes de la grille bi-dimensionnelle de processeurs. La distribution initiale peut ensuite être aisément rétablie en effectuant la transformation inverse.

En conclusion, aucune communication globale n'a été effectuée. Seules deux communications locales sont nécessaires: une à l'intérieur de chaque ligne lors de la transformée de Fourier et une autre à l'intérieur de chaque colonne pour adapter la distribution aux 3dim-FFT. Cette présentation de la parallélisation correspond à une implémentation standard, celle que nous avons initialement réalisée. D'autres améliorations ont été effectuées par la suite. C'est l'objet de la section suivante.

3.4 Améliorations et optimisations

Cette section n'a pas pour but de détailler toutes les routines qui ont été améliorées/optimisées mais simplement de mentionner trois grands champs d'investigation qui vont au-delà de la version standard.

Généralisation du principe de transposition Nous avons remarqué que le calcul des termes non-locaux (voir l'Annexe A) était plus efficace si nous adoptions la "distribution FFT" plutôt que la "distribution LOBPCG". L'origine d'un tel gain est encore mal comprise. Nous pouvons simplement supposer que cela provient d'une réduction des communications ainsi que d'une meilleure répartition (load-balancing) des données.

En effet, d'après l'équation 2.5, le calcul des termes non-locaux nécessite le calcul des termes $gx_{i,nk}^I$, réduction des $\text{ffnl}_i^I(\mathbf{k} + \mathbf{G})$ sur les PW. Dans le cadre d'une "distribution LOBPCG", il s'agit de faire une communication sur toute la grille bidimensionnelle des processeurs. Dans le cadre d'une "distribution FFT", les communications ne sont plus alors effectuées que sur les lignes.

A ceci s'ajoute un effet de load-balancing. En effet, comme nous l'avons indiqué précédemment, l'algorithme FFT de Goedecker *et al.* est optimisé. Il ne prend pas en compte les coefficients qui sont nuls car au-delà de la sphère de cutoff définie par l'équation 2.5. Or, les plans sont distribués successivement, quel que soit le nombre de coefficients $c_{nk}(\mathbf{G})$ non-nuls. Il se peut donc que la répartition effectuée lors d'une "distribution LOBPCG" soit très inégale. En passant à une "distribution FFT" les plans ne sont plus sub-divisés et la répartition des données devient moins mauvaise; à défaut de devenir excellente.

Librairies mathématiques De nombreux produits matrice-matrice et matrice-vecteur interviennent dans LOBPCG (voir l'algorithme 3), mais aussi dans d'autres parties du code. Afin d'optimiser ces calculs nous avons fait appel, de manière intensive, aux routines des librairies mathématiques LAPACK et BLAS (xgemm, xhemm, xdotc avec $x=z$ ou d selon que le calcul est complexe ou réel). Ces librairies sont distribuées avec le programme ABINIT. Néanmoins, l'utilisation de librairies "optimisées machines", disponibles sur certaines plateformes, s'est par la suite avérée essentiel. En effet, d'une part celles-ci se révèlent plus efficaces que les non-optimisées en séquentiel et, d'autre part, elles présentent un comportement sur-linéaire en parallèle (effet provenant du cache du processeur). Nous présenterons les résultats concernant ce point dans le chapitre suivant.

Diagonalisation dans le sous-espace Nous avons indiqué dans la section 3.1 qu'une diagonalisation dans le sous-espace des états était nécessaire après la résolution des équations de Kohn-Sham. Dans le cas qui nous intéresse, il s'agit donc de diagonaliser une matrice $M \times M$. La parallélisation bandFFT ne peut rien pour changer cela et le calcul est donc fait en séquentiel (tous les processeurs font le même calcul). Celui-ci se révèle être très lourd lorsque le nombre de bandes M ou de processeur $n \times p$ augmente.

Nous avons donc utilisé la librairie SCALAPACK, celle-ci permettant de résoudre des problèmes d'algèbre linéaire sur des machines parallèles, en continuité avec ce que propose LAPACK pour des programmes séquentiels. Les calculs réalisés, une fois le SCALAPACK introduit, montrent que ce goulot d'étranglement disparaît, et ceci, même pour des systèmes comprenant 2000 bandes.

Chapitre 4

Benchmarks

En Français le titre devient beaucoup plus long. Il s'agit de résultats de temps de calcul caractérisant le code et provenant de la mise au "banc d'essai".

4.1 Détails des calculs

Les calculs ont été réalisés sur deux super-calculateurs auxquels nous avons accès au Commissariat à l'énergie atomique (CEA). Le premier (TERA-10), propre au CEA/DAM Ile-de-France, est composé de 4352 processeurs double-cœurs Intel Montecito, interconnectés par l'intermédiaire d'un réseau Quadrics haute performance. Le second (TANTALE), commun à différentes directions du CEA ainsi que plusieurs entreprises publiques, comprend 138 nœuds de 4 processeurs AMD OPTERON 2.4/1.8GHz avec 4 ou 32 Gbytes de mémoire partagée, interconnectés en utilisant la technologie Infiniband (faible temps de latence et grande largeur de bande). Pour ces deux super-calculateurs, les temps de communications et de calculs obtenus sont sensiblement identiques. Nous ne nous référerons plus à l'une ou l'autre des architectures par la suite.

La simulation a été effectuée sur une supercellule (voir la section 2.1) de 108 atomes d'or, dans la phase cubique face-centrée (fcc). Les calculs ont été faits sans prendre en compte le spin. Nous avons donc considéré 648 bandes nous permettant d'obtenir un facteur d'occupation f_{nk} quasi-nul sur la dernière bande. Une énergie de cutoff (voir l'équation 2.5) de 24 Ha. a été utilisée dans le développement des fonctions d'ondes en PW. Celui-ci conduit à prendre en compte 108^3 points pour la grille tri-dimensionnelle du réseau réciproque.

Afin d'évaluer les performances de cette double-parallélisation, nous avons effectué un grand nombre de calculs sur ce système. Pour des nombres de processeurs $n_{proc}=m \times p$ égaux à 1, 4, 18, 54, 108, 162 et 216 nous avons envisagé toutes les distributions $m \times p$ possibles. Nous rappelons que toutes les valeurs de m et de p ne sont pas accessibles. Dans le cas qui nous intéresse, p varie entre 1 et $108/2=54$ non-compris. De plus m et p doivent rester commensurables avec $M=648$ et $P=108$, respectivement.

Trois types de "benchmarks" ont été envisagés. Le premier consiste à ne faire varier que p en posant $m=1$. Il s'agit d'une "parallélisation FFT" pure que nous dénoterons " $1 \times p$ ". La seconde correspond au système inverse: $p=1$ et $m=n_{proc}$. C'est une "parallélisation bandes" pure que nous dénoterons " $m \times 1$ ". Dans les deux cas la grille de processeur est monodimensionnelle. Le troisième combine les deux. C'est la "double parallélisation" qui nous préoccupe et que nous nommerons " $m \times p$ ". Pour chaque nombre de processeurs total, nous cherchons le couple ($m;p$) qui minimise le temps de calcul. Seul ce résultat est ici présenté.

4.2 Les calculs NC

Pour chaque valeur de nproc, et pour chaque type de parallélisation ($m \times p$, $m \times 1$ et $1 \times p$), nous présentons dans la figure 4.1 le scaling obtenu. Ce dernier, appelé aussi speedup, est le gain réalisé en temps humain entre un calcul séquentiel et un calcul parallèle. Pour bien saisir les comparaisons, il faut se référer au scaling linéaire. Celui-ci correspond au gain que l'on obtiendrait si le temps de calcul physique décroissait de manière inversement proportionnel au nombre de processeurs mis en œuvre.

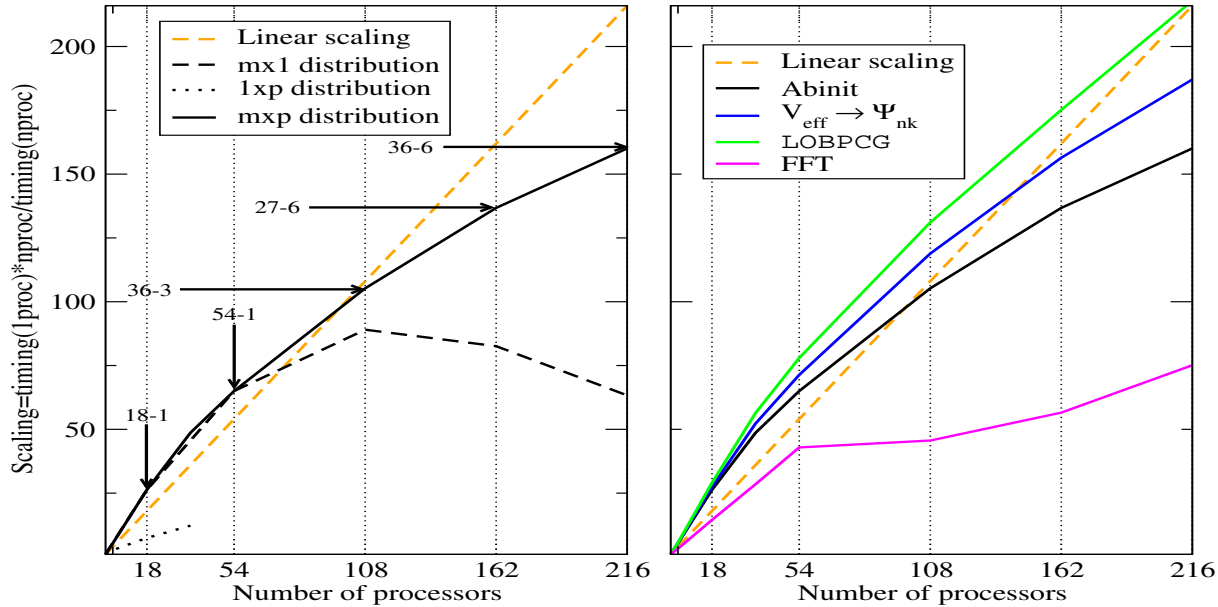


FIG. 4.1 – Double parallélisation du code de calcul ABINIT: résultats NC. Figure de gauche: gain obtenu pour trois types de distributions ($m \times p$, $m \times 1$ et $1 \times p$) et comparaison avec un scaling linéaire. A chaque flèche correspond les nombres m et p de processeurs utilisé dans le cadre de la "double parallélisation" $m \times p$. Figure de droite: distinction des différentes parties du cycle SCF (le calcul des potentiels en partant des fonctions d'ondes, l'algorithme LOBPCG utilisé pour résoudre les équations de Kohn-Sham et la partie FFT).

Comme nous pouvons le voir sur le panneau de gauche de la Figure 4.1, les deux parallélisations mono-dimensionnelles $1 \times p$ et $m \times 1$ perdent leur efficacité lorsque le nombre processeurs augmente. La "parallélisation bandes" chute à partir de 108 processeurs, ce qui signifie qu'au delà les calculs mettront plus de temps à s'exécuter qu'à $nproc=108$. Quant à la distribution $1 \times p$, le gain est trop faible par rapport à celui espéré (scaling linéaire), et ceci même pour un faible nombre de processeurs. Notons que le nombre de processeurs maximal accessible dans le cadre de la "parallélisation FFT" est en pratique très faible (il est ici égal à 36).

Contrairement aux parallélisations mono-dimensionnelles, le scaling de la distribution $m \times p$ (pour une grille bi-dimensionnelle optimisée de processeurs) augmente monotoniquement jusqu'à 216 processeurs. Ce point constitue le principal résultat de cette étude. En combinant astucieusement les parallélisations "bandes" et "FFT" nous obtenons un gain sur-linéaire jusqu'à 100 processeurs et égal à 160 pour 200 processeurs. Notons qu'à 216 processeurs, le temps de communication obtenu dans le cadre de la "parallélisation bandes" devient très important. Les communications effectuées dans ce cadre sont alors globales puisque la grille de processeurs est mono-dimensionnelle, là où la "double parallélisation", en répartissant les données sur 36 "processeurs bandes" et 6 "processeurs FFT", ne fait intervenir que des communications sur les lignes et colonnes.

Sur le panneau de droite de la figure 4.1 nous présentons le scaling des différentes parties du cycle SCF. Cette séparation nous permet de mettre en évidence le gain réalisé par les algorithmes 3dim-FFT et LOBPCG indépendamment.

En ce qui concerne le premier, le gain est presque linéaire jusqu'à 54 processeurs puis n'augmente alors que très peu pour atteindre un gain de 75 sur 216 processeurs. Cette perte est entièrement due aux communications, la contribution provenant des calculs, dans la FFT, n'augmentant pas du tout. La part de la FFT passe ainsi de 10% à 25% du temps total entre 1 et 216 processeurs. Celle-ci est presque entièrement responsable de la perte de gain observée dans le code entre 54 et 216 processeurs (voir la courbe Abinit).

En effet, l'algorithme LOBPCG est quant à lui sur-linéaire en fonction du nombre de processeurs. Cette routine représentant 90% du temps total en séquentiel, le bon comportement du code est en grande partie dû à son scaling sur-linéaire jusqu'à 216 processeurs. Les produits matrice-matrice et matrice-vecteur intervenant dans cet algorithme, effectués au moyen des bibliothèques mathématiques BLAS et LAPACK optimisées machines, sont responsables du bon comportement de LOBPCG. En augmentant le nombre de processeurs, la taille des vecteurs et matrices mises en jeu dans l'algorithme diminue et un effet de cache est alors observé entre 1 et 18 processeurs. Au delà le temps passé dans ces calculs reste constant.

4.3 Le surcoût PAW

Nous conservons le système décrit précédemment (108 atomes d'or). Notons qu'en pratique:

1. l'énergie de cutoff, égale à 24 Ha. en NC, devrait être réduite d'un facteur 2 en PAW. En effet cette méthode ne nécessite pas d'utiliser une telle précision sur la base de PW. C'est tout son intérêt. Nous garderons ici, pour les besoins des benchmarks, l'énergie de cutoff de 24 Ha. et la grille tridimensionnelle de 108^3 points.
2. le nombre de projecteurs par moment angulaire intervenant dans le terme non-local est de l'ordre de 2 en PAW, alors qu'il n'y en a qu'un en NC. Pour les besoins du test, nous avons généré un pseudo-potentiel PAW avec un projecteur par moment angulaire.

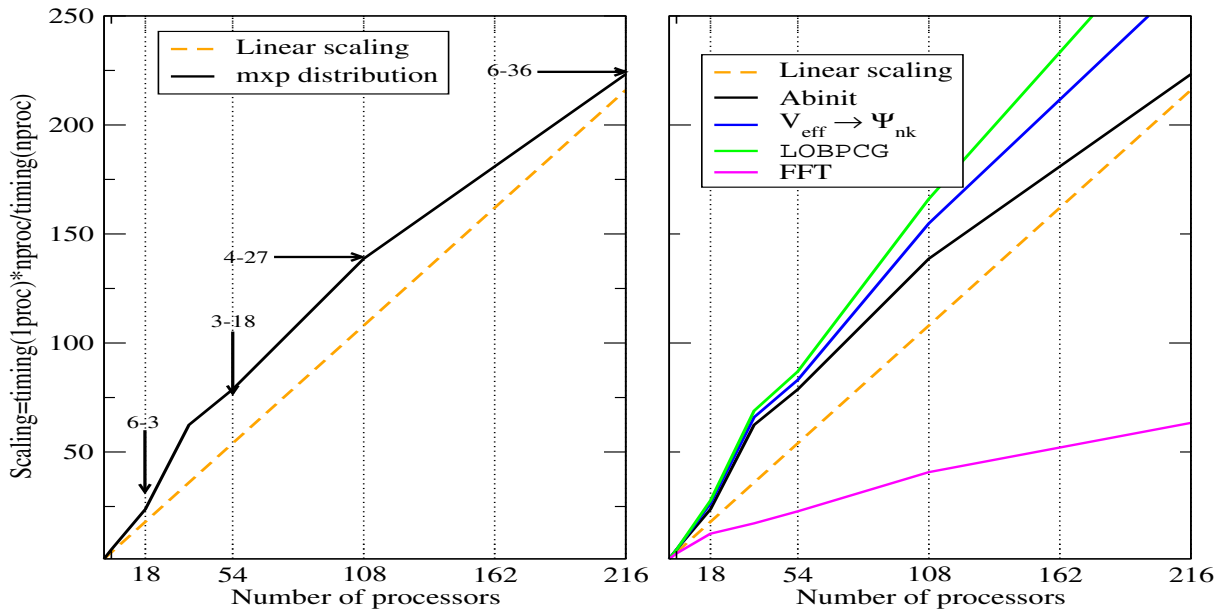


FIG. 4.2 – Double parallélisation du code de calcul ABINIT: résultats NC. Même légende que la Figure 4.1.

Les résultats obtenus en PAW sont présentés dans la figure 4.2. Si les scalings du code ABINIT et de LOBPCG semblent parfait, et même biens meilleurs qu'en NC, plusieurs différences importantes sont à signaler:

- Le temps de calcul obtenu en séquentiel est anormalement long. Il est deux fois et demi plus long qu'en NC. Celui-ci, utilisé comme référence pour déterminer le gain, biaise ainsi les résultats de scaling que nous présentons.
- Un temps de calcul beaucoup plus important est observé lors du calcul du terme non-local, et ceci, même en séquentiel. Ce terme est évalué au moyen de l'équation A.2. Contrairement à la méthode NC, le calcul de cette grandeur en PAW fait intervenir les harmoniques sphériques réelles $S_{l_i, m_i}(\widehat{\mathbf{k} + \mathbf{G}})$. Ainsi, pour chaque moment angulaire l_i il y a m_i calculs supplémentaires de la quantité $g_{i, nk}^I$ (voir l'annexe A).
- La taille mémoire utilisée sur chaque processeur est bien plus importante qu'en norme conservé. Celle-ci est par exemple supérieure à 10 Go en séquentiel. Nous avons identifié les quantités responsables d'une telle augmentation. Ce sont les contributions provenant de chaque bande n et chaque point- \mathbf{k} à la "matrice densité" ρ_{ij} (voir l'équation 2.18). Celles-ci sont bien-sûr spécifiques à la méthode PAW. Nous envisageons, pour remédier à cette augmentation de la taille mémoire, de les distribuer sur la grille de processeurs. En parallélisant ces calculs, nous espérons réduire, dans le même temps, le temps nécessaire à l'évaluation de ces contributions.

Les problèmes que nous rencontrons ici ne sont donc pas dûs à la parallélisation bandFFT. Ceux-ci proviennent de la méthode PAW. Le problème de l'évaluation du "terme non-local" en PAW doit avant tout être réglé en séquentiel. Un travail préliminaire est ici nécessaire afin d'identifier plus précisément les causes. Cette étape nous permettra ensuite d'envisager les actions à mener pour y remédier. Quant au problème de l'augmentation de la taille mémoire, il pourra sans-doute être adouci, à défaut d'être éliminé, par la distribution et la parallélisation du calcul de la "matrice densité" ρ_{ij} .

Notons enfin que le temps passé en dehors de la résolution des équations de Kohn-Sham est identique à celui observé en NC. Ce temps comprend le calcul des nombreux termes sphériques qui apparaissent en PAW (les densités définies dans l'équation 2.23 ainsi que les potentiels associés, ...) et qui ne sont pas parallélisées puisqu'ils ne font intervenir ni bandes ni PW. Ces calculs, que nous pensions être les principaux facteurs limitant en PAW, ne semblent donc pas affecter fortement le temps de calcul en parallèle (et ceci jusqu'à 200 processeurs).

Conclusion

Nous avons mis en œuvre dans le code de calcul de structure électronique ABINIT une double parallélisation, combinant astucieusement parallélisation sur les bandes et parallélisation sur les PW. Les données sont alors réparties sur une grille bi-dimensionnelle de processeurs, les communications ne s'effectuant que sur ses lignes ou ses colonnes. Ceci nous a donc permis d'atteindre des gains importants en NC, avec un comportement sur-linéaire jusqu'à 100 processeurs et un gain de 160 à 200 processeurs.

Cette parallélisation a nécessité l'introduction de deux algorithmes: une transformée de Fourier rapide parallèle et une méthode de résolution des équations de Kohn-Sham fonctionnant par blocs. La première présente une saturation du gain au delà de 54 processeurs tandis que la seconde reste sur-linéaire jusqu'à 216 processeurs.

En ce qui concerne la méthode PAW, les premiers tests mettent en évidence un gain sur-linéaire jusqu'à 216 processeurs. Ce résultat doit cependant être tempéré. Les augmentations de temps de calcul et de taille mémoire que nous observons en séquentiel introduisent un biais dans le calcul du gain. De nombreuses améliorations/optimisations envisagées sont actuellement en cours pour remédier à ces problèmes.

Vers une parallélisation triple n-k-G Afin d'obtenir un code de calcul de structure électronique efficace sur un plus grand nombre de processeurs, nous avons regroupé les parallélisations "bandFFT" et "points-k" sous une même "triple parallélisation". La parallélisation points-k fut chronologiquement la première implémentée, car très simple à mettre en œuvre et très efficace. L'implémentation de cette "triple parallélisation" est aujourd'hui terminée et les premiers tests effectués montrent que nous pouvons obtenir un scaling sur-linéaire jusqu'à 1000 processeurs. Cette unification des parallélisations est à notre connaissance une première mondiale. Celle-ci nous permet d'obtenir des performances jamais atteintes sur un code de calcul de structure électronique.

Remerciements Ces travaux ont été réalisés dans le cadre du projet IOLS grâce au soutien financier du pôle de compétitivité Systém@tic. En ce qui concerne les calculs, ceux-ci ont été effectués sur les super-calculateurs du CEA: sur la plateforme TANTALE au Centre de Calcul Recherche et Technologie et sur la machine TERA-10 du CEA/DAM Ile-de-France. Nous remercions A. Kniazhev et A. Curioni pour les fructueuses discussions eues au sujet de l'algorithme de résolution LOBPCG et de la technique de double parallélisation, respectivement, ainsi que François Jollet et Marc Torrent pour leurs conseils concernant la parallélisation PAW.

Notons que tous les résultats de cette étude ont été obtenus en utilisant le code de calcul ABINIT, un projet commun de l'Université Catholique de Louvain, Corning Incorporated, le Commissariat à l'Energie Atomique d'autres contributeurs (URL <http://www.abinit.org>).

Annexe A

Calcul des termes non-locaux sur une base de PW

Le but de cette annexe est de présenter succinctement l'implémentation du terme $\langle \tilde{\Psi}_{\mathbf{nk}} | \tilde{\mathcal{P}}_i^I \rangle$ dans un code de calcul PW. Cette "brique élémentaire" intervient de nombreuses fois, que ce soit en NC (dans le calcul des énergie et potentiel non-locaux) ou en PAW (dans le calcul de la "matrice densité" ρ_{ij}^I , du terme "non-local" faisant intervenir D_{ij}^I ou de l'opérateur d'"overlap" \mathcal{O}). Ce terme est d'une importance capitale puisque, étant très lourd à évaluer, il représente une grande partie du temps de calcul.

Débutons par le calcul du terme $\langle e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} | \tilde{\mathcal{P}}_i^I \rangle$. En utilisant le développement d'une PW sur la base des harmoniques sphériques réelles (voir l'équation B.8 de l'annexe B.2), celui-ci peut s'exprimer de la manière suivante:

$$\langle e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} | \tilde{\mathcal{P}}_i^I \rangle = 4\pi i^{l_i} \text{ffnl}_i^I(\mathbf{k} + \mathbf{G}) \quad (\text{A.1})$$

avec

$$\text{ffnl}_i^I(\mathbf{k} + \mathbf{G}) = \tilde{\mathcal{P}}_{n_i l_i}^I(|\mathbf{k} + \mathbf{G}|) S_{l_i m_i}(\widehat{\mathbf{k} + \mathbf{G}}) \quad (\text{A.2})$$

et

$$\tilde{\mathcal{P}}_{n_i l_i}^I(|\mathbf{k} + \mathbf{G}|) = \int_0^{R_i^I} r dr \tilde{\mathcal{P}}_{n_i l_i}^I(r) j_{l_i}(|\mathbf{k} + \mathbf{G}| r) \quad (\text{A.3})$$

Les quantités $\text{ffnl}_i^I(\mathbf{k} + \mathbf{G})$ sont calculées à l'initialisation et sont ensuite utilisées au cours du calcul pour déterminer $\text{ffnl}_i^I(\mathbf{k} + \mathbf{G})$. Cette procédure permet d'éviter un grand nombre de calculs inutiles. Notons aussi, qu'en NC, les harmoniques sphériques $S_{l_i m_i}(\widehat{\mathbf{k} + \mathbf{G}})$ n'entrent pas en compte dans le calcul du terme $\text{ffnl}_i^I(\mathbf{k} + \mathbf{G})$ (voir l'équation A.2). Celles-ci ne sont évaluées et introduites qu'à la fin, lors du calcul du terme non-local proprement dit, et apparaissent sous la forme de la somme d'un produit d'harmoniques sphériques, c'est à dire d'un polynôme de Legendre. Cette astuce n'a pu être utilisée en PAW puisque les termes D_{ij}^I et \mathcal{O} couplent les nombres quantiques m_i et m_j . Ceci conduit à considérer en PAW, pour chaque moment angulaire l_i , m_i fois plus de $\text{ffnl}_i^I(\mathbf{k} + \mathbf{G})$ que par rapport au même calcul en NC.

La projection de la fonction d'onde peut à présent être évaluée:

$$\langle \tilde{\Psi}_{\mathbf{nk}} | \tilde{\mathcal{P}}_i^I \rangle = \frac{1}{\sqrt{\Omega}} 4\pi i^{l_i} g_{\mathbf{x}_{i,\mathbf{nk}}}^I \quad (\text{A.4})$$

avec

$$g_{\mathbf{x}_{i,\mathbf{nk}}}^I = \sum_{\mathbf{G}} c_{\mathbf{nk}}(\mathbf{G}) \text{ffnl}_i^I(\mathbf{k} + \mathbf{G}) \quad (\text{A.5})$$

Les termes "non-locaux", ρ_{ij}^I et d'"overlap" \mathcal{O} sont ainsi calculés comme somme de produits de $g_{\mathbf{x}_{i,\mathbf{nk}}}^I$.

Annexe B

Les harmoniques sphériques complexes et réelles

B.1 Définitions

Les harmoniques sphériques complexes s'écrivent:

$$Y_{lm}(\theta, \phi) = \sqrt{\frac{(2l+1)}{(4\pi)}} \sqrt{\frac{(l-m)!}{(l+m)!}} (-1)^m P_l^m(\cos(\theta)) e^{im\phi} \quad (\text{B.1})$$

avec P_l^m les polynômes de Legendre définis par:

$$(x^2 - 1) \frac{dP_l^m}{dx} = lxP_l^m - (l+m)P_{l-1}^m \quad \text{avec } 0 \leq m \leq l-1 \text{ et } x = \cos \theta \quad (\text{B.2})$$

Pour $m=1$, les polynômes de Legendre deviennent $P_l^1(\cos \theta) = (2l-1) \sin \theta P_{l-1}^1(\cos \theta)$. La relation permettant de déterminer les harmoniques sphériques réelles S_{lm} à partir des harmoniques sphériques complexes Y_{lm} est la suivante:

$$S_{lm}(\hat{r}) = \begin{cases} \frac{1}{\sqrt{2}}(Y_{l\bar{m}} + (-1)^m Y_{lm}) & \text{pour } m > 0 \\ Y_{l0} & \\ \frac{i}{\sqrt{2}}(Y_{lm} - (-1)^m Y_{l\bar{m}}) & \text{pour } m < 0 \end{cases} \quad (\text{B.3})$$

Ces fonctions vérifient les mêmes relations que harmoniques sphériques:

$$\int_{\Omega} S_{lm}(\hat{r}) S_{l'm'}(\hat{r}) d\Omega = \delta_{ll'} \delta_{mm'} \quad , \quad S_{lm}^*(\hat{r}) = S_{lm}(\hat{r}) \quad (\text{B.4})$$

De même, on peut procéder à une décomposition angulaire sur les angles ϕ et θ tel que:

$$S_{lm}(\theta, \phi) = \Theta_{l|m|}(\theta) \Phi_m(\phi) \quad \text{avec} \quad \Phi_m(\phi) = \begin{cases} \sqrt{2} \cos m\phi \\ 1 \\ \sqrt{2} \sin |m| \phi \end{cases} \quad (\text{B.5})$$

La fonction $\Theta_{l|m|}(\theta)$ est définie par:

$$\Theta_{l|m|}(\theta) = \sqrt{\frac{(2l+1)}{(4\pi)}} \sqrt{\frac{(l-|m|)!}{(l+|m|)!}} P_l^{|m|}(\cos(\theta)) \quad (\text{B.6})$$

Enfin, le produit de trois harmoniques sphériques (réelles ou complexes) est appelé coefficient de Gaunt (RG):

$$RG_{l_1 m_1 l_2 m_2 l_3 m_3}^{LM} = \frac{1}{\sqrt{4\pi}} \int_{\Omega} S_{l_1 m_1}(\hat{r}) S_{LM}(\hat{r}) S_{l_2 m_2}(\hat{r}) d\Omega \quad (\text{B.7})$$

B.2 Relations

Les PW et le potentiel électrostatique peuvent être développés sur une base d'harmoniques sphériques (voir l'Annexe B) de la manière suivante:

$$e^{i\mathbf{r}\cdot\mathbf{k}} = 4\pi \sum_{l=0}^{\infty} \sum_{m=-l}^l i^l S_{lm}(\hat{\mathbf{r}}) S_{lm}(\hat{\mathbf{k}}) j_l(kr) \quad (\text{B.8})$$

$$\frac{1}{|\mathbf{r} - \mathbf{r}'|} = \sum_{l=0}^{\infty} \sum_{m=-l}^l \frac{4\pi}{2l+1} \frac{r_{<}^l}{r_{>}^{l+1}} S_{lm}(\hat{\mathbf{r}}) S_{lm}(\hat{\mathbf{r}}') \quad (\text{B.9})$$

avec $j_l(kr)$ les fonctions de Bessel sphériques, $r_{<} = \inf(r, r')$ et $r_{>} = \max(r, r')$.

Bibliographie

- [1] P. Hohenberg and W. Kohn, *Inhomogeneous Electron Gas*, Phys. Rev. **136**, B864 (1964).
- [2] X. Gonze, J.-M. Beuken, R. Caracas, F. Detraux, M. Fuchs, G.-M. Rignanese, L. Sindic, M. Verstraete, G. Zerah, F. Jollet, M. Torrent, A. Roy, M. Mikami, P. Ghosez, J.-Y. Raty, and D. Allan, *First-principles computation of material properties: the ABINIT software project*, Comput. Mater. Sci **25**, 478 (2002), see the URL <http://www.abinit.org>.
- [3] W. Kohn, *Nobel Lecture: Electronic structure of matter – wave functions and density functionals*, Rev. of Mod. Phys. **71**, 1253 (1999).
- [4] J. Pople, *Nobel Lecture: Quantum chemical models*, Rev. of Mod. Phys. **71**, 1267 (1999).
- [5] M. Born and R. Oppenheimer, *Zur Quantentheorie der Molekeln*, Annalen der Physik **84**, 457 (1927).
- [6] W. Kohn and L. J. Sham, *Self-Consistent Equations Including Exchange and Correlation Effects*, Phys. Rev. **140**, A1133 (1965).
- [7] D. Ceperley and B. Alder, *Ground State of the Electron Gas by a Stochastic Method*, Phys. Rev. Lett. **45**, 566 (1980).
- [8] J. Perdew and A. Zunger, *Self-interaction correction to density-functional approximations for many-electrons systems*, Phys. Rev. B **23**, 5048 (1981).
- [9] J. Perdew and Y. Wang, *Accurate and simple analytic representation of the electron-gas correlation energy*, Phys. Rev. B **45**, 13244 (1992).
- [10] O. Andersen, *Linear methods in band theory*, Phys. Rev. B **12**, 3060 (1975).
- [11] D. Singh, *Planewaves, Pseudopotentials and the LAPW Method* (Kluwer Academic Publishers, Netherlands, 1994).
- [12] E. Wimmer, H. Krakauer, M. Weinert, and A. Freeman, *Full-potential self-consistent linearized-augmented-plane-wave method for calculating the electronic structure of molecules and surfaces: O₂ molecule*, Phys. Rev. B **24**, 864 (1981).
- [13] P. Blöchl, *Projector augmented-wave method*, Phys. Rev. B **50**, 17953 (1994).
- [14] M. Payne, M. Teter, D. Allan, T. Arias, and J. Joannopoulos, *Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients*, Rev. of Mod. Phys. **64**, 1045 (1992).
- [15] N. Ashcroft and N. Mermin, *Solid State Physics* (Saunders College, Philadelphia, 1976).
- [16] P. Pulay, *Ab initio calculation of force constants and equilibrium geometries in polyatomic molecules. I. Theory*, Mol. Phys. **17**, 197 (1969).
- [17] D. Chadi and M. Cohen, *Special Points in the Brillouin Zone*, Phys. Rev. B **8**, 5747 (1973).
- [18] H. J. Monkhorst and J. D. Pack, *Special points for Brillouin-zone integrations*, Phys. Rev. B **13**, 5188 (1976).
- [19] M. Fuchs and M. Scheffler, *Ab initio pseudopotentials for electronic structure calculations of polyatomic systems using density-functional theory*, Comp. Phys. Comm. **119**, 67 (1999).
- [20] S. Louie, S. Froyen, and M. Cohen, *Nonlinear ionic pseudopotentials in spin-density-functional calculations*, Phys. Rev. B **26**, 1738 (1982).
- [21] L. Kleinman and D. M. Bylander, *Efficacious Form for Model Pseudopotentials*, Phys. Rev. Lett. **48**, 1425 (1982).

- [22] S. P. Lewis, C. Y. Wei, E. J. Mele, and A. M. Rappe, *Efficient scaling of calculations involving separable nonlocal potentials*, Phys. Rev. B **58**, 3482 (1998).
- [23] D. Hamann, M. Schlüter, and C. Chiang, *Norm-Conserving Pseudopotentials*, Phys. Rev. Lett. **43**, 1494 (1979).
- [24] N. Troullier and J. Martins, *Efficient pseudopotentials for plane-wave calculations*, Phys. Rev. B **43**, 1993 (1991).
- [25] D. Vanderbilt, *Soft self-consistent pseudopotentials in a generalized eigenvalue formalism*, Phys. Rev. B **41**, 7892 (1990).
- [26] G. Kresse and D. Joubert, *From ultrasoft pseudopotentials to the projector augmented-wave method*, Phys. Rev. B **59**, 1758 (1999).
- [27] J. Hutter and A. Curioni, *Dual-level parallelism for ab initio molecular dynamics: reaching teraflop performance with the CPMD code*, Parallel Computing **31**, 1 (2005).
- [28] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards **49**, 409 (1952).
- [29] J. Demmel, M. Heath, and H. van der Vorst, in *Parallel Numerical Linear Algebra*, Vol. 2 of *Acta Numerica*, edited by D. R. Lide (Cambridge University Press, Cambridge, 1993).
- [30] J. Rayleigh, *In Finding the Correction for the Open End of an Organ-Pip*, Phil. Trans. **161**, 77 (1870).
- [31] W. Ritz, *Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik*, J. reine angew. Math. **135**, 1 (1908).
- [32] A. Knyazev, *Toward the Optimal Preconditioned Eigensolver: Locally Optimal Block Preconditioned Conjugate Gradient Method*, SIAM Journal on Scientific Computing **23**, 517 (2001).
- [33] S. Goedecker, M. Boulet, and T. Deutsch, *An efficient 3-dim FFT for plane wave electronic structure calculations on massively parallel machines composed of multiprocessor nodes*, Comput. Phys. Comm. **154**, 105 (2003).

Résumé

Nous présentons dans ce rapport une étude consacrée à la parallélisation du code de calcul *ab initio* ABINIT. Celui-ci permet de déterminer, à l'échelle atomique, l'état fondamental d'un système réel, ainsi que les réponses de cet état à des sollicitations extérieures. Deux méthodes sont aujourd'hui utilisées dans ce code de calcul: dans la première, nommée "Norme Conservée" (NC), les données sont uniquement développées sur une base d'onde-planes, et dans la seconde, appelée "Projector Augmented-Wave" (PAW), les données sont à la fois développées sur une base d'onde-planes et d'orbitales atomiques.

Dans le cadre de ces deux méthodes, une double parallélisation a été mise en œuvre. Celle-ci s'appuie sur une transformée de Fourier rapide tridimensionnelle (3dim-FFT) ainsi que sur un algorithme de résolution des équations aux valeurs propres fonctionnant par blocs (de bandes). L'originalité de cette parallélisation repose sur l'utilisation d'une double grille de processeurs: celle-ci permet, d'une part, de distribuer les données conformément au type de calcul à réaliser, et, d'autre part, d'effectuer des communications, nombreuses, mais non globales, au cours du calcul.

Au niveau des performances liées à l'introduction de cette parallélisation, il nous faut séparer les deux méthodes. Lors d'un calcul NC, la scalabilité obtenue est excellente. Un comportement surlinéaire est observé jusqu'à 100 processeurs et, en dépit d'une perte de performance au delà, un gain de 160 est tout de même atteint sur 200 processeurs. Notons qu'en particulier l'algorithme de résolution des équations aux valeurs propres présente un comportement linéaire jusqu'à 200 processeurs. Lors d'un calcul PAW, si cette scalabilité présente un comportement linéaire jusqu'à 200 processeurs, de nombreuses précautions doivent entourer ce résultat. D'une part la scalabilité est faussé par un calcul séquentiel qui devient en PAW anormalement long. D'autre part, l'inter-opérabilité de ces deux fonctionnalisés, telles qu'elles sont implémentées aujourd'hui, n'est pas encore optimal.

En ce qui concerne ces deux méthodes de calcul, différentes sources de pertes ont été identifiées (communications, calculs séquentiels...). Notons que l'effort à réaliser porte essentiellement sur la méthode PAW pour laquelle un gain important est espéré en temps de calcul, et ceci même en séquentiel. En outre, une triple parallélisation est dès à présent mise en œuvre et permet d'atteindre un comportement linéaire jusqu'à 1000 processeurs.