Computer Physics Communications 187 (2015) 98–105

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Parallel eigensolvers in plane-wave Density Functional Theory

Antoine Levitt *, Marc Torrent

CEA, DAM, DIF, F-91297, Arpajon, France

ARTICLE INFO

Article history:
Received 19 June 2014
Received in revised form
7 October 2014
Accepted 14 October 2014
Available online 24 October 2014

Keywords:
Density Functional Theory
ABINIT
Projector augmented-wave
Chebyshev filtering
LOBPCG
Woodbury formula

ABSTRACT

We consider the problem of parallelizing electronic structure computations in plane-wave Density Functional Theory. Because of the limited scalability of Fourier transforms, parallelism has to be found at the eigensolver level. We show how a recently proposed algorithm based on Chebyshev polynomials can scale into the tens of thousands of processors, outperforming block conjugate gradient algorithms for large computations.

1. Introduction

Kohn–Sham Density Functional Theory is an efficient way to solve the Schrödinger equation for quantum systems [1,2]. By modelling the correlation between N electrons via exchange–correlation functionals, it leads to the Kohn–Sham system, mathematically formulated as a nonlinear eigenvalue problem. This problem can be discretized and solved numerically, and the result of this computation allows the determination of physical properties of interest via higher-level processing such as geometry optimization, molecular dynamics or response-function computation. Density Functional Theory (DFT) codes can be classified according to the discretization scheme used to represent wavefunctions (plane waves, localized basis functions, finite differences, etc.) and the treatment of core electrons (all-electron computations, pseudopotentials, etc.). We focus on the ABINIT software [3], which uses a plane-wave basis and either norm-conserving pseudopotentials or the Projector Augmented-Wave (PAW) approach [4,5].

The bottleneck of most simulations is the computation of the electronic ground state. This is done by a self-consistent cycle whose inner step is the solution of a linear eigenvalue problem. This step has to be implemented efficiently, taking into account the specificities of the problem at hand, which rules out the use of generic black-box solvers. Furthermore, the growing need for parallelization constrains the choice of the eigensolver. Indeed, one specificity of plane-wave DFT as opposed to real-space codes is that Fourier transforms do not scale beyond about 100 processors: effective parallelization requires eigensolvers that are able to compute several Hamiltonian applications in parallel.

The historic eigensolver used in plane-wave DFT, the conjugate gradient scheme of Refs. [6,7], is inherently sequential, although there are attempts at parallelization by omitting orthogonalizations [8]. Several methods work on blocks of eigenvectors and are more suited for parallelization, such as the residual vector minimization — direct inversion in the iterative subspace (RMM-DIIS) scheme [7], and block Davidson algorithms [9], including the locally optimal block preconditioned conjugate gradient (LOBPCG) algorithm [10], implemented in ABINIT [11].

Parallel implementations of plane-wave DFT codes include Quantum Espresso [12], VASP [7], QBOX [13] or CASTEP [14]. The scalability of these codes is mainly limited by orthogonalizations and the Rayleigh–Ritz step, a dense matrix diagonalization, which is hard to parallelize efficiently, even using state-of-the-art libraries such as ELPA [15] or Elemental [16]. The Rayleigh–Ritz step usually becomes the bottleneck when using more than a thousand processors.

There are two main ways to decrease the cost of this step. One is to use it as rarely as possible. This usually means applying the Hamiltonian more than one time to each vector before applying the Rayleigh–Ritz procedure, in order to speed up convergence. The other is getting rid of it entirely. This requires the independent computation of parts of the spectrum, as in the methods of spectrum slicing [17] or of contour integrals [18,19]. These approaches

* Corresponding author.
E-mail addresses: antoine.levitt@gmail.com (A. Levitt), marc.torrent@cea.fr (M. Torrent).

http://dx.doi.org/10.1016/j.cpc.2014.10.015
0010-4655/

effectively solve an interior eigenvalue problem, which is considerably harder than the original exterior one. The result is that a large number of Hamiltonian applications is needed, to obtain a high-degree polynomial or to solve linear systems.

While these spectrum decomposition techniques will surely become the dominant methods for exascale computing, we address the current generation of supercomputers, on which the decrease in the costs of the Rayleigh–Ritz step is not worth the great increase in the number of Hamiltonian applications. We therefore focus in this paper on the method of Chebyshev filtering, which aims to limit the number of Rayleigh–Ritz steps by applying polynomials of the Hamiltonian to each vector. It can be seen as an accelerated subspace iteration, and dates back to the RITZIT code in 1970 [20]. It has been proposed for use in DFT in Refs. [21,22], and has recently been adopted by several groups [23,24].

The contribution of this paper is twofold. First, we show how to adapt the Chebyshev filtering algorithm of Ref. [21] in the context of generalized eigenproblems, here due to the PAW formalism. By exploiting the particular nature of the PAW overlap matrix (a low-rank perturbation of the identity), we are able to invert it efficiently. Second, we compare the Chebyshev filtering algorithm with CG and LOBPCG, both in terms of convergence and scalability.

2. The eigenvalue problem

2.1. The operators

First, we define some relevant variables. For a system of N_{atoms} atoms in a box, we solve the Kohn–Sham equations in a plane-wave basis. This basis is defined by the set of all plane waves whose kinetic energy is less than a threshold E_{cut}. This yields a sphere of N_{pw} plane waves, upon which the wavefunctions are discretized.

We consider a system where N_{bands} bands are sought. For a simple ground state computation, N_{bands} represents the number of states occupied by valence electrons of the N_{atoms} atoms. For more sophisticated analysis such as Many-Body Perturbation Theory (MBPT), the computation of empty states is necessary, and N_{bands} can be significantly higher. It is also convenient to speed up convergence of ground state computations to use more bands than strictly necessary.

To account for the core electrons, we use pseudopotentials. ABINIT implements both norm-conserving pseudopotentials and the Projector Augmented-Wave (PAW) method. For the purposes of this paper, the main difference is the presence of an overlap matrix in the PAW case, leading to a generalized eigenvalue problem. We will assume in the rest of this paper that we use the PAW method: norm-conserving pseudopotentials follow as a special case.

For simplicity of notation, we consider in this paper the case where periodicity is not taken explicitly into account, and the wavefunctions will be assumed to be real. The following discussion extends to the periodic case by sampling of the Brillouin zone, provided that we consider complex eigenproblems, with the necessary adjustments.

The Kohn–Sham equations for the electronic wavefunctions ψ_n are

$$H\psi_n = \lambda_n S\psi_n, \tag{1}$$

where H is the Hamiltonian, and S the overlap matrix arising from the PAW method ($S = I$ with norm-conserving pseudopotentials). H and S are $N_{\text{pw}} \times N_{\text{pw}}$ Hermitian matrices (although they are never formed explicitly), and Ψ is a $N_{\text{pw}} \times N_{\text{bands}}$ matrix of wavefunctions. The Hamiltonian operator depends self-consistently on the wavefunctions Ψ. It can be written in the form

$$H = K + V_{\text{loc}} + V_{\text{nonloc}}. \tag{2}$$

The kinetic energy operator K is, in our plane wave basis, a simple diagonal matrix. The local operator $V_{\text{loc}} = V_{\text{ext}} + V_{\text{H}} + V_{\text{XC}}$ is a multiplication in real space by a potential determined from atomic data and the wavefunctions Ψ. It can therefore be computed efficiently using a pair of inverse and direct FFTs. The nonlocal operator V_{nonloc} and the overlap matrix S depend on the atomic data used. For both PAW method and norm-conserving pseudopotentials, we introduce a set of n_{lmn} projectors per atom, where n_{lmn} is the number of projectors used to model the core electrons of each atom, and usually varies between 1 and 40 according to the atom and pseudopotential type. Therefore, for a homogeneous system of N_{atoms} atoms we use a total of $N_{\text{projs}} = n_{lmn}N_{\text{atoms}}$ projectors. We have $N_{\text{projs}} \ll N_{\text{pw}}$, but N_{projs} is comparable to N_{bands}.

We gather formally these projectors in a $N_{\text{pw}} \times N_{\text{projs}}$ matrix P. The non-local operator V_{nonloc} is computed as

$$V_{\text{nonloc}} = PD_VP^T. \tag{3}$$

Similarly, the overlap matrix in the PAW formalism is

$$S = I + PD_SP^T. \tag{4}$$

The matrices D_S and D_V do not couple the different atoms in the system: they are block-diagonal. They can be precomputed from atomic data. The matrix D_V additionally depends self-consistently on the wavefunctions Ψ.

Therefore, for a single band ψ, the process of computing $H\psi$ and $S\psi$ can be decomposed as follows:

Input: a wavefunction ψ
Output: $H\psi$, $S\psi$
∘ Compute $K\psi$ by a simple scaling
∘ Apply an inverse FFT to ψ to compute its real-space representation, multiply by V_{loc}, and apply a FFT to get $V_{\text{loc}}\psi$
∘ Compute the N_{projs} projections $p_\psi = P^T\psi$
Apply the block-diagonal matrices D_V and D_S to p_ψ
Compute the contributions PD_Vp_ψ and PD_Sp_ψ to the nonlocal and overlap operator
∘ Assemble $H\psi = K\psi + V_{\text{loc}}\psi + PD_Vp_\psi$
∘ Assemble $S\psi = \psi + PD_Sp_\psi$

Algorithm 1: Computation of $H\psi$, $S\psi$

The total cost of this operation is $O(N_{\text{pw}} \log N_{\text{pw}} + N_{\text{pw}}N_{\text{projs}})$. As N_{pw} and N_{projs} both scale with the number of atoms N_{atoms}, the cost of computing the non-local operator dominates for large systems. However, N_{pw} is usually much greater than N_{projs}, and the prefactor involved in computing FFTs is much greater than the one involved in computing the simple matrix products $P^T\psi$ and Pp_ψ (which can be efficiently implemented as a level-3 BLAS operation). The FFT and non-local operator costs are usually of the same order of magnitude for systems up to about 50 atoms.

2.2. Solving the eigenvalue problem: conjugate gradient

The historical algorithm used to compute the N_{bands} first eigenvectors of (1) in the framework of plane-wave DFT is the conjugate gradient algorithm, described in [6,7]. It is mathematically based on the following variational characterization of the n-th eigenvector of the eigenproblem $H\psi = \lambda S\psi$:

$$\psi_n = \underset{\langle \psi_i,\, S\psi\rangle = \delta_{i,n},\ i=1,\ldots,N_{\text{pw}}}{\arg\min} \langle \psi, H\psi \rangle .$$

The conjugate gradient method of Refs. [6,7] consists of minimizing this functional by a projected conjugate gradient method. Note that, because of the constraints, this is a nonlinear conjugate gradient problem, to which classical (linear) results cannot be