Self-consistent GW method: O(N) algorithm for the polarizability and the self energy

A. L. Kutepov*$

Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Upton, NY 11973

Abstract

An efficient implementation of the self-consistent GW method in the FlapwMBPT code [1] is presented. It features the evaluation of the polarizability and the self-energy which scales linearly with respect to the system size. Altogether the computational time scaling was measured to be between linear and quadratic in the applications to the silicon supercells with up to 72 atoms. Application to such materials as paracostibite CoSbS, supercells of La$_2$CuO$_4$ (up to 56 atoms), and SmB$_6$ illustrate the potential of the approach in computational material science.

Keywords:
self-consistent GW; quasiparticle approximation; electronic structure

1. Introduction

The so called GW method was introduced by Lars Hedin in his seminal paper [2] more than 50 years ago. Its non-self-consistent variant, which often is abbreviated as $G_0W_0$ is probably the most popular method for the evaluation of the electronic structure in weakly and moderately correlated materials. It has been implemented in many computer codes as an efficient perturbative correction to the electronic structure usually calculated at the DFT (Density Functional Theory) level. Since its first pseudopotential-based implementation and applications to simple materials [3, 4] considerable efforts were undertaken to enhance the computational efficiency of the method [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. Also, the $G_0W_0$ approach was adapted to all-electron basis sets such as FLAPW (Full-potential Linearized Augmented Wave) method [18, 19, 20, 21, 22, 23].

Despite its popularity and sufficiently low computational cost, $G_0W_0$ approach is seriously limited when one needs a real prediction of the basic features of the electronic structure. Different starting points (such as LDA, GGA, Hybrid functionals, LDA+U) are required depending on a particular material and it is hard to make a decision on the starting point without having some experimental input available. For instance, using Perdew-Burke-Ernzerhof (PBE) functional [24] as a starting point for the $G_0W_0$ calculations of the band gap in simple covalent semiconductors and insulators might be considered as sufficiently successful strategy. The combination PBE+$G_0W_0$ provides the gaps with typical underestimation (as compared to the experimental

*$Corresponding author.
E-mail address: akutepov@bnl.gov

Preprint submitted to Computer Physics Communications November 14, 2019
gaps) within 10-15% \cite{13}. Application of $G_0W_0$ to d- and f-electron oxides requires the usage of different approximation (LDA+U, \cite{25}) as a starting point \cite{26}. LDA+U method, in its own turn, has an adjustable parameter (Hubbard U) which complicates its usage in the combination LDA+U+$G_0W_0$ even further. Other starting points are also possible. For example, the band gap in the widely studied insulating monoclinic phase of vanadium dioxide VO$_2$ can be reproduced with the COHSEX approach \cite{2} as a starting point, whereas the $G_0W_0$ based on the LDA doesn’t open the gap \cite{27}. Obviously, the development of efficient self-consistent GW-based approaches which eliminate the dependence on the starting point is important.

Several self-consistent implementations of the GW method (scGW) are now available \cite{1,28,29,30}. Their applications to real materials are, however, limited by considerable computational cost and memory requirements. Normally one applies them to materials with less than 10-12 atoms per unit cell. At the same time, a surge in the experimental exploration of new materials which we are witnessing nowadays deals with 20-30 (and more) atoms per unit cell. Thus, an extension of the applicability of scGW (or its highly successful quasiparticle version QSGW \cite{31}) to the solids with large unit cells is mandatory.

High computational cost of the self-consistent GW method originates mostly from the fact that many existing GW codes (both the $G_0W_0$ and scGW) still use a naive implementation of the polarizability $\mathcal{P}$ and the self-energy $\Sigma$. This implementation uses frequencies and momenta as basic variables and scales as $O(N^4N^2_\omega N^2_k)$ (N being the system size, $N_\omega$ is the number of frequencies, and $N_k$ stands for the number of points in the Brillouin zone (momenta)).

Certain progress was achieved during the last 20 years. First, the evaluation of $\mathcal{P}$ and $\Sigma$ is more efficient when one uses imaginary time instead of frequency (scales linearly with respect to the number of time points) \cite{32,23}. Second, the evaluation of $\mathcal{P}$ and $\Sigma$ benefits a lot when the real space is used instead of the $k$-space. The real space implementation scales formally as $O(N^2N_k)$. First implementation of it was announced more than 20 years ago \cite{32} in the context of plane wave basis set and pseudopotential approximation. First implementation of the real space technique for all-electron method and MT geometry (FLAPW basis set) was done in the code FlapwMBPT \cite{1} and published a few years ago \cite{33}. Recently, similar technique was also included in the code VASP \cite{30} in the context of the PAW (Projected Augmented Wave) method \cite{14}.

In this work, the advantages of using the real space in the evaluation of $\mathcal{P}$ ans $\Sigma$ are explored further. Namely, not only the linear scaling with respect to the system size is introduced, but also it is shown that the real space provides greater flexibility for the choice of the basis set which makes calculations faster and dramatically alleviates the memory requirements.

2. Methodology

Let us begin the presentation of the GW implementation with a brief account of the self-consistent GW method in its classic form which is written in momenta+frequency representation. Each iteration of the GW method takes some approximate Green’s function $G$ as an input and proceeds with the following steps. First one evaluates the polarizability

$$P_{IJ}(\nu) = \sum_{\alpha, \lambda, k} \sum_{\lambda', k', q} \langle M_{I}^{\alpha, k} | \Psi_{\alpha, k}^{\nu} | M_{J}^{k'} \rangle \sum_{\lambda''} G_{\lambda', \lambda''}^{\alpha, k, q - \nu} \sum_{\lambda', \lambda, k'} G_{\lambda, \lambda'}^{\alpha, k, q - \nu} \langle \Psi_{\alpha, k}^{\nu} | M_{I}^{\lambda} \rangle,$$  

(1)

where $\alpha$ stands for spin, $k$ and $q$ are momenta (points in the Brillouin zone), $\nu$ and $\omega$ are bosonic and fermionic Matsubara frequencies correspondingly, $\lambda, \lambda', \lambda''$ stand for band indexes, $I$ and
The expression (1) scales as $O(N^2)$ with respect to the number of momenta and frequencies, and as $O(N_k^2)$ with the number of momenta. It is very time consuming in both limits of the very large systems (small $N_k$ but large $N$) and of the small systems (small $N$ but large $N_k$).

Next step in the self-consistent GW method is the evaluation of the screened interaction $W$:

$$W^{q}_{IJ}(\nu) = V^{q}_{IJ} + \sum_{KL} V^{q}_{IK} P^{q}_{KL}(\nu) W^{q}_{LJ}(\nu). \quad (2)$$

This expression scales linearly with respect to the number of momenta and frequencies, and its scaling is cubic with respect to the system size. Quantities $V^{q}_{IJ}$ in (2) represent matrix elements of the bare Coulomb interaction.

The third step is the evaluation of the self energy:

$$\Sigma^{q}_{kl}(\omega) = -\sum_{q} \sum_{x',x''} \sum_{IJ} \langle \phi^{q}_{x'} | [\nu | \psi^{q}_{x''} M^{\nu}_{IJ}] G^{\nu}_{x''x} (\omega - \nu) W^{q}_{IJ}(\nu) | \nu \phi^{q}_{x'} \rangle,$$  \quad (3)

which scales similarly to the expression (1) for the polarizability.

Finally one has to solve the Dyson equation for the Green’s function

$$G^{\nu}_{\lambda \delta}(\omega) = G^{0,\nu}_{\lambda \delta}(\omega) + \sum_{x',x''} G^{0,\nu}_{\lambda x'}(\omega) \Sigma^{q}_{x'x''}(\omega) G^{0,\nu}_{x'' \delta}(\omega), \quad (4)$$

which formally scales similarly to the expression (2) for the screened interaction, but in practice is less time consuming because usually the number of band states is a few times smaller than the size of the product basis. Quantities $G^{\nu}_{\lambda \delta}(\omega)$ in (4) are the components of the Green function in the Hartree approximation.

From the comparison of scalings of the above four equations one can easily conclude that in order to make calculations faster one needs to optimize the evaluation of $P$ and $\Sigma$. As it was already mentioned in the Introduction, the first serious step in that direction was undertaken by Rojas et al. \[32\] who implemented GW method in the real space + imaginary time representation with the pseudopotential approximation and plane wave basis set. The scaling of the evaluation of the polarizability and the self energy was improved to $O(N^2 N_\nu N_\lambda)$ with obvious advantages. In our previous paper \[33\] (the corresponding algorithm will be called as the old version here) we implemented the self-consistent GW approach for all electron full potential method (FLAPW+LO basis set) in the real space + imaginary (Matsubara) time representation which resulted in enormous saving of computer time in the applications to realistic materials. Because the basic formulae of the implementation in the Ref. \[33\] are very similar to the ones used in the new methodology which is presented in this work it is convenient to give them here (formulae (5)-(10) below). Muffin-tin geometry associated with the LAPW method breaks the expressions for $P$ and $\Sigma$ into three distinctive cases corresponding to where the two space arguments belong (MT sphere or interstitial (INT)). In the formulas given below, the following notations are used: $t, t'$ for the coordinates of the centers of the MT spheres; $K, K'$ for the product basis set indexes inside the muffin-tin spheres; $\varphi^k_{t}$ for the one-electron orbitals inside the MT spheres with $L, L', L''$, $L'''$ representing all quantum numbers to distinguish them; $\tau$ for the Matsubara (imaginary) time defined in the range $[0: \beta]$; $r, r'$ for the points on the real space grid in the interstitial region; $W$ is
the screening part of the interaction \( W (W = V + \bar{W}) \). Formulæ \([8]-[10]\) represent the evaluation of the dynamic (frequency dependent) part of the self energy. Static (exchange) part is evaluated according to the same expressions but with replacement \( \bar{W}(\tau) \rightarrow V \) and \( G(\tau) \rightarrow -G(\beta) \).

Polarizability (MT-MT):

\[
P_{\alpha L,x}^R(\tau) = -\sum_{\alpha} \sum_{L} \langle \phi_{\alpha L}^L | t^L | \phi_{\alpha L}^R \rangle L \cdot \sum_{\alpha} \sum_{L'} \langle \phi_{\alpha L'}^L | t^L | \phi_{\alpha L'}^R \rangle L' \cdot \sum_{\alpha} \sum_{L''} \langle \phi_{\alpha L''}^L | t^L | \phi_{\alpha L''}^R \rangle L'' \cdot (\beta - \tau) \langle \phi_{\alpha L}^R | M_{\alpha L}^R \rangle.
\]

(INT-INT):

\[
P_{\alpha R}^R(\tau) = -\sum_{\alpha} \sum_{L} \langle \phi_{\alpha L}^L | t^L | \phi_{\alpha L}^R \rangle L \cdot \sum_{\alpha} \sum_{L'} \langle \phi_{\alpha L'}^L | t^L | \phi_{\alpha L'}^R \rangle L' \cdot \sum_{\alpha} \sum_{L''} \langle \phi_{\alpha L''}^L | t^L | \phi_{\alpha L''}^R \rangle L'' \cdot (\beta - \tau) \langle \phi_{\alpha L}^R | M_{\alpha L}^R \rangle.
\]

Self energy (MT-MT):

\[
\Sigma_{\alpha L,x}^R(\tau) = -\sum_{\alpha} \sum_{L} \langle \phi_{\alpha L}^L | t^L | \phi_{\alpha L}^R \rangle L \cdot \sum_{\alpha} \sum_{L'} \langle \phi_{\alpha L'}^L | t^L | \phi_{\alpha L'}^R \rangle L' \cdot \sum_{\alpha} \sum_{L''} \langle \phi_{\alpha L''}^L | t^L | \phi_{\alpha L''}^R \rangle L'' \cdot (\beta - \tau) \langle \phi_{\alpha L}^R | M_{\alpha L}^R \rangle.
\]

(INT-INT):

\[
\Sigma_{\alpha R}^R(\tau) = -\sum_{\alpha} \sum_{L} \langle \phi_{\alpha L}^L | t^L | \phi_{\alpha L}^R \rangle L \cdot \sum_{\alpha} \sum_{L'} \langle \phi_{\alpha L'}^L | t^L | \phi_{\alpha L'}^R \rangle L' \cdot \sum_{\alpha} \sum_{L''} \langle \phi_{\alpha L''}^L | t^L | \phi_{\alpha L''}^R \rangle L'' \cdot (\beta - \tau) \langle \phi_{\alpha L}^R | M_{\alpha L}^R \rangle.
\]

Quadratic scaling of the above formulæ follows from the fact that the computations in each of them are proportional to the number of pairs \((t, t')\), \((t, r')\), or \((r, r')\) which grows quadratically when the unit cell increases. Linear scaling with respect to the number of time points and of the \(k\)-points (which is equal to the number of \(R\)-vectors in the old version) is also obvious. The implementation of the self consistent GW method in the real space + Matsubara time representation allowed considerable savings of computer time. However, the need to store quantities like \( F_{\alpha L}^R(\tau) \) (or like \( F_{\alpha L}^k(\tau) \)) where \( F \) stands for any two-point quantity (\( G, P, \Sigma, \) or \( W \)) was limiting the applications to the unit cells with no more than 10-15 atoms. The principal reason for the exceedingly large amount of information for storing is that the approximations (such as the size of the basis set) were introduced in the reciprocal space which didn’t provide a flexibility in the choice of a basis set size. Basis set size was essentially the same for all \(k\)-points. Real space, which formally provides such a flexibility (see below) was used in the old version only as an intermediate step in order to accelerate the calculations.

In this work it is proposed to use real space as a principal representation, where all important cutoffs are introduced. The quantities for storing now look like \( F_{\alpha R}^R(\tau) \equiv F_{\alpha R+R}(\tau) \), where the real space points \(r\) and \(r'\) belong to the unit cells (generally different) and the translation vector \( R \) measures the distance between two unit cells. Essential difference of the new approach as compared to the old version of the code FlapwMBPT consists in the fact that the list of \(R\)-vectors is defined not from the given \(k\)-mesh but from the following consideration which naturally leads to the linear scaling with respect to the system size. Namely, one accepts an approximation
Figure 1: Graph illustrating the basic approximation. Vertical and horizontal lines correspond to the boundaries of unit cells. For any given point \( r' \) (center of the sphere in the figure) the two point functions \( F_{rr'} \) are nonzero only for \( r \) inside the sphere of a certain radius (red circle). Number of unit cells which the limiting sphere encloses or intersects in each direction defines the mesh of \( k \)-points in the Brillouin zone. See the text for more details.

\[
F_{rr'}^{R}(\tau) = 0 \quad \text{for all pairs} \quad R + r, r' \quad \text{with} \quad |R + r - r'| > R_{\text{max}}. \quad R_{\text{max}} \quad \text{here is a parameter eventually defining the density of} \quad k \quad \text{-points in the Brillouin zone. Now the requirement for the number of} \quad k \quad \text{-points is that it should be the same or larger than the number of unit cells which are inside of or cross the sphere} \quad R_{\text{max}}. \quad \text{This ensures the absence of periodic images inside the sphere. For example, in the Figure} \quad 1 \quad \text{the minimal (two-dimensional in this case)} \quad k \quad \text{-mesh should be at least} \quad 3 \times 3 \quad \text{as the sphere} \quad R_{\text{max}} \quad \text{encloses or crosses three unit cells in each direction. Formally, the convergence is achieved when the function at} \quad R_{\text{max}} \quad \text{is very small. By making} \quad R_{\text{max}} \quad \text{independent on the system size (which is physically sound) the scaling becomes} \quad O(mN) \quad \text{where} \quad m \quad \text{is the number of} \quad R + r \quad \text{points inside sphere} \quad R_{\text{max}}. \quad \text{The number} \quad m \quad \text{is independent on the system size and, correspondingly, the approach has linear scaling.}

The principal advantage of the new methodology comes however not from its linear scaling, but from the flexibility in the selection of the basis set size which it provides. The idea is that the change in any two-point function \( F_{rr'}^{R} \) when the vectors \( R + r \) and/or \( r' \) change is much larger when \( R + r \) and \( r' \) are close to each other than when they are distant. In practice, it means that the density of \( r \) and \( r' \) points can be considerably reduced when they belong to the distant unit cells. In the LAPW method, this statement is directly relevant to the interstitial region. Inside the MT spheres, the functions are expanded in spherical harmonics. Again the cutoff \( L_{\text{max}} \), for this expansion can be considerably reduced when two MT spheres are distant.

Present implementation of the above ideas in the code FlapwMBPT uses three levels of approximation related to the basis sets inside muffin spheres. Figure 2 illustrates it. The largest basis set is used when both space arguments belong to the same MT sphere (small sphere A in the middle of the figure). The intermediate size of the basis set is used when the distance between the centers of the MT spheres doesn’t exceed the radius of sphere B in the figure. For example, this situation is realized when the argument \( r' \) belongs to the central (red) MT sphere A but the argument \( r \) belongs to the blue sphere which is one of the MT spheres inside the limiting sphere B. Finally, the smallest basis set is used when the distance between the muffin spheres is larger.
Figure 2: Illustration of the flexibility of the real space for a basis set selection. Dots on the grid correspond to the centers of atoms. When two space arguments of a two point function \( F_{r,r'} \) belong to the same central sphere (sphere A in the figure) the function is expanded in a largest basis set. When one space argument belongs to sphere A but another one belongs to sphere B (for instance blue circle) the basis set can be reduced. Finally, when one space argument belongs to sphere A but another one belongs to sphere C (for instance green circle) the basis set is the smallest.

than the radius of sphere B but smaller than the radius of sphere C. An example would be the pair consisting from the red MT sphere at the center and the green MT sphere inside limiting sphere C. When space argument belongs to the interstitial region, however, there is obviously no central A sphere. In this case only spheres B and C are used when one constructs the grids of \( r \) and \( r' \) points.

One iteration of the new implementation formally proceeds in precisely the same way as it is in the older version of the FlapwMBPT code. It was described in details in the Ref. [33] (fully relativistic variant) and in the Ref. [34] (non-relativistic variant). Briefly one follows the formulae (1)-(4). However, the equation (1) is replaced with efficient (order \( O(N) \) implementation of equations (5)-(7) and the equation (3) is replaced with the corresponding efficient implementation of the equations (8)-(10). For the evaluation of \( W \) (Eqn. 2) and of the Green’s function (Eqn. 4) the LAPACK or SCALAPACK libraries are used. One more thing needs to be mentioned here. The transformations between \( R \) and \( k \) spaces are also performed more efficiently in the new implementation as compared to the old version. This is because the complexity of these transformations is proportional to the number of pairs \( (R + r, r') \) for instance) which grows as \( O(N) \) in the new algorithm versus \( O(N^2) \) in the old one. Also, which is even more important for the efficiency, the basis set size inside the muffin tin spheres (or the density of grid points in the interstitial region) is more compact in the new implementation.

3. Convergence tests

In this section, a brief account of the convergence of the results with respect to the most important setup parameters of the new implementation is given. They are the radii of spheres B and C (see Fig. 2) and the spherical harmonics cutoff \( L_{\text{max}} \) for the product basis expansions.
(both for atoms in sphere B and in sphere C). The possibility of a reduction in the density of the \( r \)-points in the interstitial region for remoted unit cells is similar to the possibility of a reduction in the \( L_{\text{max}} \) and, correspondingly, will not be presented.

Left part of the figure 3 represents the calculated band gap of silicon as a function of the radius of sphere C. The radius of sphere B was fixed at 8 a.u. The cutoff \( L_{\text{max}} \) was fixed at 6 (spheres A and B) and at 2 (outside of sphere B but inside of sphere C). Other setup parameters, such as the number of functions in the muffin tin spheres and the number of plane waves in the interstitial region were also fixed at the level ensuring that an uncertainty of the calculated band gap was no more than 0.02-0.03 eV. As it was explained earlier, an increasing of the radius of sphere C automatically changes the density of \( k \)-points in the Brillouin zone. This is reflected in the Fig. 3 as well. As one can see, the change in the number of \( k \)-points is not necessary monotonic: it happens only when sphere C crosses additional unit cells (see Fig. 1). The band gap converges within 0.02eV when sphere C radius reaches 14 a.u. It corresponds to 216 \( k \)-points (6 \( \times \) 6 \( \times \) 6 mesh in the Brillouin zone) which reflects usual rate of convergence for this material.

Right part of the figure 3 shows the band gap of Si evaluated as a function of the radius of sphere B which is the internal sphere in Fig. 2. Dependence of the results on this parameter appears only because of the differences in basis set sizes between the interior of sphere B and the region outside of sphere B but inside of sphere C. Convergence is achieved when the basis set outside of sphere B becomes good enough so that its gradual replacement with the basis set from the interior of sphere B (when the radius of sphere B increases) does not change the results.

As one can see from the Figure 3, the convergence in case of Si is achieved when the radius of sphere B reaches value 12 a.u. Even when its value is equal to 8 a.u., the calculated band gap is very close to the final answer. Thus, Fig. 3 shows the possibility to use reduced basis sets for two-point quantities when their space arguments are sufficiently far away from each other. It is interesting also to look at this from slightly different perspective. Namely, if we fix the radii of all spheres and just change the size of the basis sets in different regions of space we will get direct measure of how big those basis sets should be. It would also be another support of the possibility to use compact basis sets for distant cells. Figure 4 represents such exploration. It shows two lines - solid and dashed. Solid line represents the band gap of Si as a function of \( L_C \) (cutoff in the product basis set for the muffin tin spheres in the distant unit cells) with \( L_B = 6 \). Dashed line

![Figure 3: Left: Convergence of the calculated band gap of Si with respect to the radius of sphere C (see Fig. 2). The radius of sphere B was fixed at 8 a.u.. Right: Convergence of the calculated band gap of Si with respect to the radius of sphere B (see Fig. 2). The radius of sphere C was fixed at 24 a.u.](image-url)
shows the band gap of Si as a function of $L_B$ with $L_C$ fixed at $L_C = 2$. The basis set in the central sphere (sphere A in the Fig. 2) was fixed at $L_A = 6$. The radii of spheres B and C were kept unchanged at $R_B = 8\,\text{a.u.}$ and $R_C = 22\,\text{a.u.}$. First, it is clear that the convergence of the calculated band gap is very fast when one varies $L_C$ and its value $L_C = 2$ is sufficient. The convergence of the calculated gap with respect to $L_B$ is slower and is only reached at $L_B = 6$ which is equal to the fixed value of $L_A$. This fact simply indicates that for Si, the physics associated with neighboring atoms is essential. Most important for our study, however, is the fact that the basis set for remoted unit cells can be considerably reduced. The amount of computer memory needed to store two-point quantities is proportional to the number of basis set functions per one pair of atoms and to the number of pairs of atoms. The number of pairs belonging to the region between sphere B and sphere C is usually much larger than the number of pairs belonging to the interior of sphere B as one can conclude from the corresponding ratio $\frac{R_C - R_B}{R_A}$. For our selected geometry ($R_B = 8\,\text{a.u.}$, $R_C = 22\,\text{a.u.}$) this ratio is approximately 20. The storage related to the size of the product basis set is roughly proportional to $(L_B(L_C) + 1)^4$ per one pair which simply tells us that the reduction in memory for remoted cells is $(7/3)^4 \approx 30$. Combined with the fact that the distant cells represent vast majority of all included cells, the advantages of the new implementation are obvious.

4. Scalability tests and timings

New implementation of the scGW/QSGW approach combines the linear scaling parts (polarizability and self energy) and cubic scaling parts (evaluation of W and G as well as transforms between R and k spaces). It is interesting how it scales as a whole. For instance, how the time per full iteration depends on the number of atoms in the unit cell? Figure 5 provides this measurement performed on the supercells of silicon atoms. The sizes of the supercells are varied from 2 atoms (one unit cell) to 72 atoms ($4 \times 3 \times 3$ supercell). In all cases the same number of computer
cores (288) was used. In order to facilitate the analysis, the coefficients $\alpha, \beta, \gamma$ in assumed linear, quadratic, and cubic scaling (i.e. assuming that the time scales as $\alpha N$, or as $\beta N^2$, or as $\gamma N^3$) were found based on the measured time for 2-atom unit cell. Corresponding lines have been added to the graph. In order to make the test more transparent, also the symmetry was switched off. Quick observation from the Fig. 5 is that up to about 20 atoms the scaling is perfectly linear. Even when the size of the supercell becomes 72 atoms the scaling is still between linear and quadratic, demonstrating the efficiency of the algorithm. A couple of comments are, however, needed. First, as it was explained above, the increase in the size of the supercell automatically results in the reduction of the number of $k$-points in the Brillouin zone (as $1/N$) until only one $k$-point is left. The results on the graph were obtained with the number of $k$-points in the range from 125 (2 atoms) to 18 (72 atoms). In other words, even for 72-atom supercell the limit of just one $k$-point is reached yet. Cubic scaling parts of the algorithm always come with the prefactor $O(N_k)$ and the $1/N$-type reduction in the number $N_k$ makes them scale quadratically until the limit of only one $k$-point is reached. Another comment is related to the parallelization strategy which was optimized for large supercells whereas for small supercells it was accepted to be the same (in order to make sure that only the number of atoms changes). Consequently, the parallelization strategy for small supercells was not necessarily optimal. The optimization would, probably, push the QSGW curve on the graph towards the quadratic scaling a bit. However, it will remain lower than pure quadratic scaling because the ratio of the two last points on QSGW curve still is smaller than the corresponding ratio on the pure $O(N^2)$ curve.

It is interesting to know, at least approximately, how the performance of the new algorithm compares to the performances of other known codes. Figure 6 provides such comparison. As the authors of the work [14] did (using the VASP code [30]), the 16-atom supercell of Si atoms was used in QSGW calculations with 64 cores. One can see, that for this case, the FlapwMBPT code is a few times faster than the VASP code especially for larger number of $k$-points. A few comments, however, are to be said about the comparison. Certain (essential) difference comes from the fact, that in [14] the time corresponds to $G_0W_0$ calculations, whereas in this
Figure 6: Computation time as a function of the number of $k$-points in the Brillouin zone. 16 Si atom supercell and 64 cores are used. Solid lines represent calculations with the symmetry switched off, whereas the dashed lines represent the calculations performed with full symmetry use. FlapwMBPT timings are given in black color. VASP timings are also given for comparison (in red color).

work the time per one QSGW iteration was measured. Full QSGW iteration is more costly than $G_0W_0$ because of the need to update the chemical potential and to evaluate full Green’s function which was expanded in approximately 2000 band states for the 16-atom supercell. In the $G_0W_0$ calculations \cite{14} the correction was done only for 64 band states. On the other hand, measured $G_0W_0$ times (possibly) include the times for initial LDA iterations. In QSGW, the time associated with a few initial LDA iterations was not included, but it comprised in every case less than 10% of the time per one GW iteration and, consequently, would not change the graphs noticeably. Next factor, potentially influencing on the comparison is the differences in the type of computer nodes. Whereas this factor is important, it should not have strong impact on the comparison. The speed of single cores did not change much since 2016 (the year when the work \cite{14} was published). Most important impact may come from the differences in the size of the basis set which is attributed to an intrinsic approach-related part (APW+lo versus PAW) and a part related to the degree of convergence. In this work, the sizes of the basis sets (fermionic and bosonic) were selected such that the convergence of the band energies near the Fermi level was at the level of 0.02eV. This resulted in the sizes of about 2400 (size of APW+lo basis set) and about 3500 (size of the product basis set). Unfortunately, authors of \cite{14} do not provide an information about their basis set sizes used for the time measurements which, of course, complicates the comparison. In any case, despite the above issues related to the direct comparison of the performance, the efficiency of the new implementation of scGW/QSGW in the FlapwMBPT code can be considered to be competitive with the efficiency of the widely used codes.

5. Applications to real materials

Application of the new scGW algorithm to real materials did not pursue the goal of performing thorough study of any specific solid. Rather, the goal of this work was to demonstrate a potential
of the code as a tool to study sufficiently big systems containing atoms with d or f electrons. Despite the fact that the choice of materials was a bit random, one can hope that the first application of the new scGW methodology to the selected compounds will open a path for more detailed studying of them (and of other materials of similar complexity) in the future.

5.1. La$_2$CuO$_4$

Principal goal of this study was to explore the ability of the new version of the FlapwMBPT code to run sufficiently large unit cells of more complicated materials (as compared to the silicon) with d- or f-electrons. This ability could be useful when one is interested in (for instance) the substitution effects. Calculations for the supercells of La$_2$CuO$_4$ in the body-centered-tetragonal (bct) structure were performed in scalar-relativistic approximation assuming anti-ferromagnetic (AFM) ordering of spin moments. Structural parameters for this study were taken from the Ref. [40]. AFM ordering dictated that the minimal size of the supercell has two atoms of Cu in it, i.e. 14 atoms total. This minimal supercell was used as a building block for constructing the 28-atom supercell (by doubling the original supercell in Z direction) and the 56-atom supercell (by doubling the original supercell in X and Y directions). Setup parameters were selected based on a few LDA runs for the minimal 14-atom supercell. Most important setup parameters as well as main results (magnetic moments and band gaps) are collected in the Table 1. The number of computer cores as well as the time per one iteration are also given. As the scGW calculations for such large systems are sufficiently time consuming (even with the new algorithm), scGW runs were performed only once without studying the convergence with respect to the parameters specific for the GW part (such as the product basis set size). Nevertheless, small differences in the results obtained with the supercells of different size may serve as an evidence of sufficiently good convergence. An essential result of this study is that now it is possible to use self consistent GW approach to study the electronic structure of complex materials with about 50 atoms (and may be more) per unit cell. Particularly for the La$_2$CuO$_4$ compound, the ability to study supercells opens a path to study the effects of a substitution of La atoms with other elements (such as strontium) which affects the superconducting properties.

5.2. CoSbS

Next compound studied in this work is paracostibite CoSbS which is an environmetally friendly thermoelectric material [41]. It crystallizes in orthorhombic structure (space group

| supercell size (atoms) | 14  | 28  | 56  |
|------------------------|-----|-----|-----|
| computer cores         | 320 | 320 | 384 |
| APW+lo basis set       | 1250| 2500| 4920|
| Product Basis set      | 4950| 9880| 19730|
| k-mesh                 | $6 \times 6 \times 3$ | $6 \times 6 \times 2$ | $4 \times 4 \times 3$ |
| time per iteration (min)| 5  | 29  | 137 |
| Magnetic moment (Cu)   | 0.81| 0.80| 0.82|
| Band gap               | 1.91| 1.88| 1.96|
61). Electronic structure of CoSbS was calculated using three different approximations: LDA, QSGW, and scGW. The experimental structural parameters were used in all calculations. Numerical setup (cutoffs) for the calculations were selected such that the principal features of the calculated electronic structure (such as band gap) were converged at the level 1-2% or better. Principal results (band structures) of this study are presented in Figure 7. As one can see, LDA seriously underestimates the band gap (more than 50%), whereas scGW overestimates it (almost by the factor of 2). The best agreement with the experimental value is obtained when one uses QSGW approximation. Thus, at the moment, QSGW can be recommended for more detailed studying of this material. However, QSGW has a disadvantage (similarly to LDA) of being non-diagrammatic, which represents an obstacle for systematic improvements of these two approaches. In this respect, fully self consistent GW (which is diagrammatic) represents more fundamental approach and a better platform for improvements. For instance, diagrammatic vertex corrections which have shown a great success in simple (primarily consisting from elements with sp-electrons) semiconductors/insulators can also be very useful for CoSbS. The support for this statement comes from the fact that CoSbS is also a weakly correlated material as its renormalization Z-factor (Figure 8) has values of about 0.87 (and higher) in scGW and 0.81 (and higher) in QSGW. These values of the renormalization factor are typical for weakly correlated semiconductors. The optimization of the algorithm of vertex corrections represents the next step in the nearest future development of the FlapwMBPT code and it is planned to be accomplished along the same lines as the optimization of the GW part presented in this work.

Plotting the bands associated with the scGW approach needs to be clarified. Strictly speaking,
the one-electron features (band dispersion) in this approach should be obtained as the positions of the peaks of the $k$-resolved spectral functions. These spectral functions include the analytical continuation of the correlation part of the self energy from imaginary to real axis of frequencies. However, as it was demonstrated in the Ref. [33], the positions of the peaks in spectral function near the chemical potential can often be accurately reproduced by a simplified procedure. This procedure (see Appendix D in the Ref. [33] where it was called as QP-III) involves the linearization of the frequency dependence of the self energy near the chemical potential and, consequently, results in effective one-electron energies. Figure 8 demonstrates, that this simplified procedure is quite accurate also in the case of CoSbS, at least for the bands surrounding the Fermi energy. One-electron energies, such obtained, can obviously be used for plotting purposes.

5.3. SmB$_6$

Mixed valence insulator SmB$_6$ attracts considerable attention lately in view of its topological properties [44, 45, 46]. SmB$_6$ belongs to the class of strongly correlated materials and, as such, GW approximation itself cannot be sufficient for a thorough study of its electronic structure. Previous theoretical studies of this compound were performed using the LDA [47, 48, 49], LDA+U [49], onsite approximation to the hybrid functional [50], and LDA+Gutzwiller approach [44]. LDA approximation is considered as inadequate for this material because it places the f-states near the Fermi level with a splitting between the occupied $f_{5/2}$ and unoccupied $f_{7/2}$ states of only
0.7eV as compared to the 7eV splitting found in the photoemission experiment [51]. LDA+U approximation misplaces the position of the occupied f-states shifting them to about 5eV below the Fermi level. LDA+Gutzwiller approximation is designed to take into account strong electronic correlations (though on the model level) but, similar to the LDA or LDA+U, it neglects by non-local effects - exchange and screening. It increases the splitting between the occupied and unoccupied f-states to 4eV, which still is about twice less than in the experiment. In the onsite approximation to the hybrid functional [50], the exchange is taken into account (but only onsite) but the screening is totally missing. Aforementioned splitting between occupied and unoccupied f-states is again about 4eV. Both effects (exchange and screening), however, are included automatically on the ab-initio level in the GW-based approaches. Thus, application of the scGW method to SmB$_6$ could serve as an additional insight from different perspective on the calculated electronic structure of this interesting material. The calculations for SmB$_6$ in this work were performed for experimental structural parameters [52] using fully relativistic approach to account for strong spin-orbit interaction in this material. Principal calculational setup parameters were the following: $8 \times 8 \times 8$ k-mesh, about 1500 basis functions in the APW+lo basis set (slightly different numbers depending on the k-point), and about 2500 functions in the product basis set.

Figure 11 shows the band structure of SmB$_6$ evaluated with both the LDA and the scGW approximations.
proaches. As it was anticipated, the principal difference between two is that $f_{7/2}$ bands which were situated right above the Fermi level in the LDA calculations have been pushed up to about 7.5-8 eV above the Fermi level in the scGW calculations (see Figure 11 which shows the calculated PDOS). This effect is similar to the one observed earlier in the LDA+Gutzwiller calculations [44] and in the calculations based on the onsite hybrid functional [59] where, however, the splitting was insufficient (4 eV). From this comparison one can conclude that the non-local effects included in the scGW are important. Small (0.5-1 eV) overestimation of the splitting in the scGW calculation should obviously be corrected when the vertex corrections are taken into account. Low energy part of the calculated band structure (Figure 10) shows some similarities of the scGW results with the LDA+Gutzwiller band structure obtained in the Ref.[44]. There are differences as well, the principal one being the wider range (beginning at about -0.4eV) of $f_{5/2}$ bands in scGW whereas in LDA+Gutzwiller approach the $f_{5/2}$ bands are in the narrow window (less than 0.1 eV wide) under the Fermi level. Obviously, adding of diagrams to the scGW should shrink the bands but to what degree is hard to estimate without actual calculations. On the other hand, experimental photoemission result [53] indicates the presence of strong $^5H$ multiplets in the range from -0.2 eV to almost the Fermi level which could be more consistent with scGW results, especially if one takes into account the probable ”shrinkage” of the bands associated with the vertex corrections.

More serious insufficiency of the scGW approximation is obvious from the fact, that the experimental flat band representing the $^6F$ multiplet [54] at about 1 eV below the Fermi level is totally missing. It is interesting, however, that in the LDA+Gurzwiller [44] approximation this flat band is also missing. Thus, it looks like the proper on-site physics (supposedly being captured in the Gutzwiller approximation) and the non-local effects (captured in the scGW) should be combined together in order to properly address the physics of SmB$_6$.

Figure 12 (left) demonstrates that SmB$_6$ has renormalization Z-factor of about 0.6 (lowest value) at scGW level. This value will naturally be reduced when higher order vertex corrections are included, but even at the scGW level one can see the difference in the strength of correlations in this material as compared to, for instance, CoSbS (Fig. 8) where the lowest value of the renormalization factor is about 0.87 (in scGW).

Figure 12 (right side) demonstrates that also in the case of SmB$_6$ the linearization procedure used to plot the bands obtained in the scGW runs is quite accurate. As one can see, all features on the proper spectral function (i.e. the one obtained after the analytical continuation of the self energy) are very accurately reproduced by the linearization procedure.

Conclusions

New optimized algorithms for the polarizability and the self energy in the scGW/QSGW method for solids have been developed. Their implementation in the code FlapwMBPT shows linear scaling with respect to the number of atoms in the unit cell and allows a flexible choice of the size of the basis set. Tests and applications to the real materials demonstrate the efficiency of the new approach and its usefulness in material science.

Acknowledgments

This work was supported by the U.S. Department of energy, Office of Science, Basic Energy Sciences as a part of the Computational Materials Science Program.
References

[1] https://www.bnl.gov/cmpmsd/flapwmbpt/.
[2] L. Hedin, New Method for Calculating the One-Particle Green’s Function with Application to the Electron-Gas Problem, Phys. Rev. 139 (1965) A796.
[3] G. Strinati, H. J. Metcalf, and W. Hanke, Dynamical aspects of correlation corrections in a covalent crystal, Phys. Rev. B 25 (1982) 2867.
[4] M. S. Hybertsen and S. G. Louie, Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies, Phys. Rev.B 34 (1986) 5390.
[5] H. F. Wilson, F. Gygi, and G. Galli, Efficient iterative method for calculations of dielectric matrices, Phys. Rev. B 78 (2008) 113303.
[6] D. Roccali, D. Lu, and G. Galli1, Ab initio calculations of optical absorption spectra: Solution of the Bethe-Salpeter equation within density matrix perturbation theory, J. of Chem. Phys. 133 (2010) 164109.
[7] H. F. Wilson, D. Lu, F. Gygi, and G. Galli, Iterative calculations of dielectric eigenvalue spectra, Phys. Rev. B 79 (2009) 245106.
[8] D. Lu, F. Gygi, and G. Galli, Dielectric Properties of Ice and Liquid Water from First-Principles Calculations, Phys. Rev. Lett. 100 (2008) 147601.
[9] P. Umari, G. Strinati, and S. Baroni, GW quasiparticle spectra from occupied states only, Phys. Rev. B 81 (2010) 115104.
[10] F. Giustino, M. L. Cohen, and S. G. Louie, GW method with the self-consistent Sternheimer equation, Phys. Rev. B 81 (2010) 115105.
[11] F. Bruneval and X. Gonze, Accurate GW self-energies in a plane-wave basis using only a few empty states: Towards large systems, Phys. Rev. B 78 (2008) 085125.
[12] J. A. Berger, L. Reining, and F. Sottile, Ab initio calculations of electronic excitations: Collapsing spectral sums, Phys. Rev. B 82 (2010) 041103.
[13] D. Foerster, P. Koval, and D. Sanchez-Portal, An O(N3) implementation of Hedin’s GW approximation for molecules, J. of Chem. Phys. 135 (2011) 074105.
[14] P. Liu, M. Kaltak, J. Klimes, and G. Kresse, Cubic scaling GW: Towards fast quasiparticle calculations, Phys. Rev. B 94 (2016) 165109.
[15] D. Neuhauser, Y. Gao, C. Arntsen, E. Rabani, and R. Baer, Breaking the Theoretical Scaling Limit for Predicting Quasiparticle Energies: The Stochastic GW Approach, Phys. Rev. Lett. 113 (2014) 076402.
[16] M. Kim, G. J. Martyna, S. Ismail-Beigi, Complex time, shredded propagator method for large-scale GW calculations, arXiv:cond.mat.:1904.10512.
[17] W. Gao, W. Xia, X. Gao and P. Zhang, Speeding up GW Calculations to Meet the Challenge of Large Scale Quasiparticle Predictions, Scientific Reports 6 (2016) 36849.
[18] H. Jiang, and P. Blaha, GW with linearized augmented plane waves extended by high-energy local orbitals, Phys. Rev. B 93 (2016) 115203.
[19] C. Friedrich, A. Schindlmayr, and S. Blügel, T. Kotani, Elimination of the linearization error in GW calculations based on the linearized augmented-plane-wave method, Phys. Rev. B 74 (2006) 045104.
[20] C. Friedrich, M. C. Müller, and S. Blügel, Band convergence and linearization error correction of all-electron GW calculations: The extreme case of zinc oxide, Phys. Rev. B 83 (2011) 081101.
[21] C. Friedrich, M. C. Müller, and S. Blügel, Erratum: Band convergence and linearization error correction of all-electron GW calculations: The extreme case of zinc oxide [Phys. Rev. B 83, 081101(R) (2011)], Phys. Rev. B 84 (2011) 039906(E).
[22] D. Nabok, A. Gulans, and C. Draxl, Accurate all-electron G0W0 quasiparticle energies employing the full-potential augmented plane-wave method, Phys. Rev. B 94 (2016) 035118.
[23] A. Kutepov, S. Y. Savrasov, and G. Kotliar, Ground-state properties of simple elements from GW calculations, Phys. Rev. B 80 (2009) 041103.
[24] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized Gradient Approximation Made Simple, Phys. Rev. Lett. 77 (1996) 3865.
[25] V. I. Anisimov, J. Zaanen, and O. K. Andersen, Band theory and Mott insulators: Hubbard U instead of Stoner I, Phys. Rev. B 44 (1991) 943.
[26] H. Jiang, Revisiting the GW approach to d- and f-electron oxides, Phys. Rev. B 97 (2018) 245132.
[27] M. Gatti, F. Bruneval, V. Olevano, and L. Reining, Understanding Correlations in Vanadium Dioxide from First Principles, Phys. Rev. Lett. 99 (2007) 266402.
[28] [http://github.com/kotani/ecalj](http://github.com/kotani/ecalj).
[29] [https://www.questaal.org/](https://www.questaal.org/).
[30] [https://www.vasp.at/](https://www.vasp.at/).
[31] T. Kotani and M. van Schilfgaarde, S. V. Faleev, Quasiparticle self-consistent GW method: A basis for the independent-particle approximation, Phys. Rev. B 76 (2007) 165106.

[32] H. N. Rojas, R. W. Godby, and R. J. Needs, Space-Time Method for Ab Initio Calculations of Self-Energies and Dielectric Response Functions of Solids, Phys. Rev. Lett. 74 (1985) 1827.

[33] A. Kutelev, K. Haule, S. Y. Savrasov, and G. Kotliar, Electronic structure of Pu and Am metals by self-consistent relativistic GW method, Phys. Rev. B 85 (2012) 155129.

[34] A. L. Kutelev, V. S. Oudovenko, G. Kotliar, Linearized self-consistent quasiparticle GW method: Application to semiconductors and simple metals, Comp. Phys. Comm. 219 (2017) 407.

[35] J. M. Ginder, M. G. Roe, Y. Song, R. P. McCall, J. R. Gaines, E. Ehrenfreund, and A. J. Epstein, Photoexcitations in La2CuO4: 2-eV energy gap and long-lived defect states, Phys. Rev. B 37 (1988) 7506.

[36] S. L. Cooper, G. A. Thomas, A. J. Millis, P. E. Sulewski, J. Orenstein, D. H. Rapkine, S-W. Cheong, and P. L. Trever, Optical studies of gap, exchange, and hopping energies in the insulating cuprates, Phys. Rev. B 42 (1990) 10785.

[37] F. Borsa, P. Carretta, J. H. Cho, F. C. Chou, Q. Hu, D. C. Johnston, A. Lascialfari, D. R. Torgeson, R. J. Gooding, N. M. Salem, and K. J. E. Vos, Staggered magnetization in La2+ xSr+xCuO4 from 139La NQR and µSR: Effects of Sr doping in the range 0 < x < 0.2, Phys. Rev. B 52 (1995) 7334.

[38] M. Reehuis, C. Ulrich, K. Prokes, A. Gozar, G. Blumberg, S. Komiya, Y. Ando, P. Pattison, and B. Keimer, Crystal structure and high-field magnetism of La2CuO4, Phys. Rev. B 73 (2006) 144513.

[39] J. Stankiewicz, M. Evangelisti, P. F. S. Rosa, P. Schlottmann, and Z. Fisk, Physical properties of Sm1−xLa+xB6 single crystals, Phys. Rev. B 94 (2016) 155101.

[40] F. Lu, J.Z. Zhao, H. Weng, Z. Fang, and X. Dai, Correlated Topological Insulators with Mixed Valence, Phys. Rev. B 86 (2012) 075105.

[41] A. Yanase and H. Harima, Band Calculations on YbB12, SmB6 and CeNiSn, Prog. Theor. Phys. Suppl. 108 (1992) 19.

[42] C.-J. Kang, J. Kim, K. Kim, J. Kang, J. D. Denlinger, and B. Il Min, Band Symmetries of Mixed-Valence Topological Insulator: SmB6, J. of Phys. Soc. of Japan 84 (2015) 024722.

[43] V. N. Antonov, B. N. Harmon, and A. N. Yaresko, Electronic structure of mixed-valence semiconductors in the LSDA+U approximation. II. SmB6 and YbB12, Phys. Rev. B 66 (2002) 165209.

[44] M. Gmitra, H. Cencarikova, P. Farkasovsky, First-Principles Study of Kondo Insulator SmB6, Acta Physica Polonica A 126 (2014) 298.

[45] J. D. Denlinger, G.-H. Oweon, J. W. Allen, C. G. Olson, Y. Dalichaouch, B.-W. Lee, M. B. Maple, Z. Fisk, P. C. Canfield, P. E. Armstrong, Advances in photoemission spectroscopy of f-electron materials, Physica B 281 (2000) 716.