Large-Scale Condensed Matter DFT Simulations: Performance and Capabilities of the CRYSTAL Code

(Article begins on next page)

25 April 2024

Large Scale Condensed Matter DFT Simulations: Performance and Capabilities of the CRYSTAL Code

A. Erba,[1, *] J. Baima,[1] I. Bush,[2] R. Orlando,[1] and R. Dovesi[1]

[1]*Dipartimento di Chimica, Università di Torino, Via Giuria 5, 10125 Torino, Italy*
[2]*Oxford e-Research Centre, University of Oxford, 7 Keble Road, OX1 3QG, Oxford, UK*
(Dated: August 30, 2017)

Nowadays, the efficient exploitation of High Performance Computing resources is crucial in order to extend the applicability of first-principles theoretical methods to the description of large, progressively more realistic molecular and condensed matter systems. This can be achieved only by devising effective parallelization strategies for the most time-consuming steps of a calculation, which requires some effort given the usual complexity of quantum-mechanical algorithms, particularly so if parallelization is to be extended to all properties and not just to the basic functionalities of the code. In this paper, the performance and capabilities of the massively-parallel version of the CRYSTAL17 package for first-principles calculations on solids are discussed. In particular, we present: i) recent developments allowing for a further improvement of the code scalability (up to 32768 cores); ii) a quantitative analysis of the scaling and memory requirements of the code when running calculations with several thousands (up to about 14000) of atoms per cell ; iii) a documentation of the high numerical *size consistency* of the code; iv) an overview of recent *ab initio* studies of several physical properties (structural, energetic, electronic, vibrational, spectroscopic, thermodynamic, elastic, piezoelectric, topological) of large systems investigated with the code.

I. INTRODUCTION

Density functional theory (DFT), combined with periodic boundary conditions, is the method of choice for the *ab initio* calculation of a variety of properties of materials, owing to its excellent balance between high accuracy and relatively low computational cost.[1–5] The extension of the application of DFT to large, structurally complex models of real systems (often characterized by a low degree of point- and/or translational-symmetry) is therefore crucial to its future success and popularity. To achieve this the exploitation of high-performance computing (HPC) resources, coupled with the development of reduced system size scaling algorithms, at present constitutes the main route.[6–12] Efficient use of such resources requires the scientific application, CRYSTAL in this case, to exploit the massively parallel and distributed nature of the modern HPC facility; the trend for current machines is a rapid increase in the number of computational cores available coupled with a stagnation in the amount of memory per core and the clock rate. Thus for a modern HPC orientated scientific application, whatever its area, an efficient exploitation of parallel computing, through a hierarchy of strategies, is vital and allows not only for a faster time-to-solution, but also for a reduction of the required memory per process if the data is properly distributed, which makes the study of large systems possible.

Achieving a high degree of parallelization in codes for the *ab initio* description of materials requires a large investment of effort, particularly so if parallelization is to be extended to all algorithms and not just to those referring to basic functionalities (like single-point SCF calculations for energy and electronic structure determinations, for instance). Furthermore, it should be stressed that in order to obtain reliable results which can help in gaining insight on the physics of a large scale system with periodic quantum-mechanical codes, a good speed-up as a function of the number of cores is a necessary but insufficient condition. Indeed, as trivial as it might sound, one has also to ensure the application has a high numerical *size consistency*, that is retention of high accuracy as a calculation is scaled in size. This has seldom been addressed in the literature. In this paper, the parallel efficiency of the CRYSTAL17 code is documented and quantified according to different criteria (strong scaling profiles, system size scaling, memory requirements). Different parallelization strategies are discussed, the numerical *size consistency* documented, and recent applications of the code to large systems reviewed, which show how, nowadays, large scale DFT simulations of many properties of materials can be routinely run on HPC machines.

A replicated-data parallel version of the program, PCRYSTAL using message-passing-interface (MPI) directives, has been available since 1996, in which a complete copy of all necessary data for the calculations is held by every process, but different processes are performing different independent parts of the calculation at a given instant.[13] The evaluation of one- and two-electron integrals is very efficiently parallelized within this scheme, as independent subsets of integrals may be assigned to different processes. For the SCF procedure, parallelism is mainly achieved by exploiting the factorization of many computational tasks in reciprocal space (such as the Fourier transform of the main matrices, Kohn-Sham matrix diagonalization, etc.), into essentially independent **k**-points. When spin-unrestricted formulations are used for open-shell systems, parallelism is also extended to α and β spins. From the CRYSTAL14 version of the program, this kind of parallelism has further been pushed to include point-symmetry factorization within each **k**-point based on the corresponding irreducible representa-

tions (irreps).[14] Over the years, the replicated-data parallelism strategy has been progressively extended to **all** the algorithms of the code,[15–17] thus covering the whole spectrum of available properties (energetic, structural, electronic, vibrational, spectroscopic, thermodynamic, linear and non-linear optical, elastic, piezoelectric, photoelastic, topological, etc.). This parallel mode is particularly effective for systems containing up to several tens of *symmetry-irreducible* atoms per cell (say ≤ 100) and for a standard or dense sampling of reciprocal space. Above this indicative threshold, memory typically becomes the limiting factor of this approach because of data replication; this method of parallelism obviously has very poor memory scalability, and thus more processes do not mean larger systems may be solved. However, given that point-symmetry is fully-exploited at all steps of the calculation in the code (in particular, all the big matrices are stored in their most symmetry-irreducible shape), the overall size of the systems that can be studied with this approach can be much larger if they belong to rich point-symmetry groups.[14,18,19] For instance, many classes of cubic crystals are characterized by 48 point symmetry operators, which makes it possible to study systems with a few hundred atoms per cell. In this respect, single-walled carbon nanotubes and fullerenes represent exceptional cases, where the number of point-symmetry operators can even exceed 48.[20] For instance, a giant (10,10) icosahedral carbon fullerene has been studied with this version of the program, which is characterized by 120 symmetry operators and contains 6000 atoms (out of which only 55 turn out to be symmetry-irreducible).[21]

A distributed-data massively-parallel version of the program, MPPcrystal, has been available since 2010, where the largest arrays required by a calculation are partitioned and distributed among the processes so that each process stores in memory only a fraction of the total array on which it operates, and ScaLAPACK libraries are used for performing linear algebra tasks.[11,12] In particular, all objects that depend upon the square of the system size, including such entities as the Kohn-Sham matrix and its eigenvectors, are either distributed in this way or have been eliminated from the code. Further, all large objects in direct space (such as Fock and density matrices for each direct lattice cell) are stored in their most compact "irreducible" form and deallocated when not needed (during the reciprocal space operations, for instance), thus reducing the memory footprint of the program. We also note that this compact representation scales linearly with the system size, not quadratically. Owing to the distributed-data strategy, the larger the number of processes, the smaller the required memory per process as each process holds a smaller part of the distributed matrices. Thus the distributed-data parallel version represents the ideal tool to study very large systems (containing several hundreds or thousands of atoms per cell), with little or no point-symmetry, and a coarse reciprocal space sampling. At variance with the replicated-data version, not all algorithms have been extended to the distributed-data strategy yet. Available algorithms include the self-consistent-field procedure, analytical gradient evaluation, the geometry optimizer, the vibrational frequency and infrared intensities (through a Berry-phase approach) calculation, and the elastic and piezoelectric tensor determination. Work is constantly in progress to extend this strategy to other properties.

In this paper, we report on how the massively-parallel version of the program has been improved in the last five years, mainly in three respects: i) the required memory per core has been further reduced by adopting a direct strategy (according to which some arrays are recomputed when needed instead of being stored on memory) for relatively small matrices, which would progressively constitute a memory bottleneck as the size of the system increases; ii) the scalability of the wall-clock time needed to complete the calculation with respect to the number of cores used to run the calculation on has been improved, which required almost all tasks (including those taking very little time for small- or medium-sized systems) to be efficiently parallelized; iii) the scalability with respect to system size has also been improved by means of a restructuring of the routines for the selection of Coulomb and exchange two-electron integrals.

The structure of the paper is as follows: the technical improvements that made it possible to scale up to 32768 CPUs are discussed and documented in Section II; recent applications of the MPPcrystal code to the study of several properties of large condensed matter systems are reviewed in Section III; conclusions and perspectives are drawn in Section IV.

II. TECHNICAL IMPROVEMENTS

In order to document the performance of MPPcrystal, a structural model of the MCM-41 compound (amorphous mesoporous silica) is here taken as a reference, which contains 579 atoms per primitive cell, displays no point-symmetry, and was already used as a benchmark for the documentation of the parallel efficiency of previous versions of the program.[11,16] When systems of different size have to be considered, larger structural models are obtained by growing supercells of this system along the **c** crystallographic axis, which ensures a good structural and chemical consistency. These supercells will be denoted as Xn, where n is an integer defining the order of expansion along **c**. The PBE functional is used for most calculations in combination with a 6-31G** Pople basis set (corresponding to 7756 atomic orbitals, AOs, per primitive cell). All calculations were run on the SuperMUC (LRZ, Germany) HPC IBM iDataPlex machine powered by 16 Intel cores per node running at 2.7 GHz, with 2 GB/core. The largest supercell here considered is X24, which is 24 times larger than the primitive cell and thus contains 13896 atoms/cell and 186144 AOs/cell.

A. Memory Occupancy

As anticipated in the Introduction, memory is progressively becoming the main bottleneck in the quantum-mechanical simulation of large systems given that modern HPC architectures are often characterized by a relatively low memory per core (of about 2 GB/core). Efficient data-distribution strategies thus have to be devised in order to fully-exploit HPC resources. Most of the largest arrays in the MPPCRYSTAL code, and in particular all matrices whose size increases quadratically with the system size, are distributed among processors. A core of initial arrays (related to geometry, symmetry and integral screening), scaling linearly with the system size, are still replicated on each processor. The memory occupancy of the MPPCRYSTAL code is documented in Figure 1 as regards the SCF process. The left-hand side panel reports the memory peak per core (in MB) as a function of the size of the system for a given number of adopted processors. We observe that the memory occupancy grows linearly up to a certain system size and then scales quadratically for larger systems. This transition from a linear to a quadratic regime occurs at a system size that increases with the number of processes: it is about 2000 atoms/cell for 1024 processes, 5000 atoms/cell for 2048, 9000 atoms/cell for 4096, and finally a linear scaling of the memory occupancy is observed up to about 14000 atoms/cell when the calculation is run over 8192 processes. Figure 1 (left panel) roughly documents which systems can be studied, memory-wise, on a HPC machine with 2 GB/core with the MPPCRYSTAL program: the primitive cell of MCM-41 containing 579 atoms can be run on just 32 2 GByte cores; the X8 supercell (4632 atoms/cell) requires 512 such cores; the X16 supercell (9264 atoms/cell) 1024 cores; while the largest system here considered (X24 with 13896 atoms/cell) can be studied by using 4096 cores. Let us note that the memory requirement for a global hybrid SCF calculation would be between twice and three times the occupancy of a pure DFT calculation; a B3LYP calculation could indeed be performed for the X12 system (6948 atoms/cell) over 4096 cores, for instance.
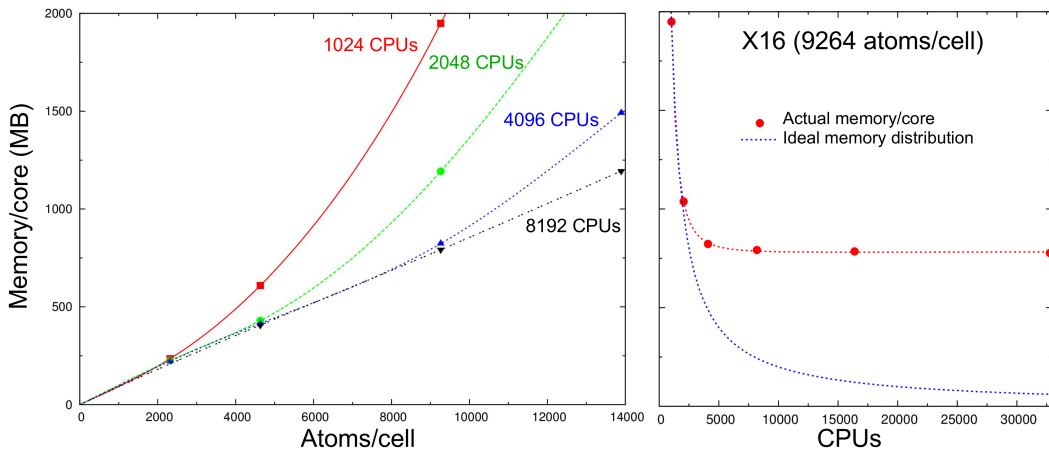

FIG. 1: Maximum memory per core during the SCF step of the calculation as a function of the system size for different numbers of CPUs (left panel) and as a function of the number of CPUs for the X16 system with 9264 atoms/cell (right panel). With the adopted basis set, there are 7756 AOs per primitive cell (i.e. per 579 atoms); 10000 atoms thus correspond to about 134000 AOs.

In the right-hand side panel of Figure 1, the maximum memory per core of the SCF calculation for the X16 system (9264 atoms/cell) is reported as a function of the number of adopted processors, as compared with the ideal case (blue line) in which all arrays are effectively distributed. A rather efficient memory reduction is observed when several hundreds or few thousands cores are used. Above 4096 processors, the required memory per core reaches a plateau of about 750 MB in this case, which is basically due to those initial arrays mentioned above (with a linear dependence on the system size) that are replicated for each core. To address this, a shared-memory strategy is being devised in order to share these arrays among all the processors within the same node, which will significantly decrease the value of this plateau and which will be available in next versions of the code.

B.   Strong Scaling Profiles

In Section II A we have discussed the effect of the use of an increasing number of processes on the reduction of the required memory per process, which ultimately determines the size of the systems that can be studied with MPP CRYSTAL. In this section, we discuss the complementary effect of the use of an increasing number of processors on the reduction of the time needed to complete the calculation. When a faster time-to-solution is sought, it is crucial to estimate the optimal number of processors to run the calculation on: too small a number would not fully exploit the parallel scalability of the code while too large a number would clearly result in an inefficient scalability because the speedup reaches a plateau in those conditions.
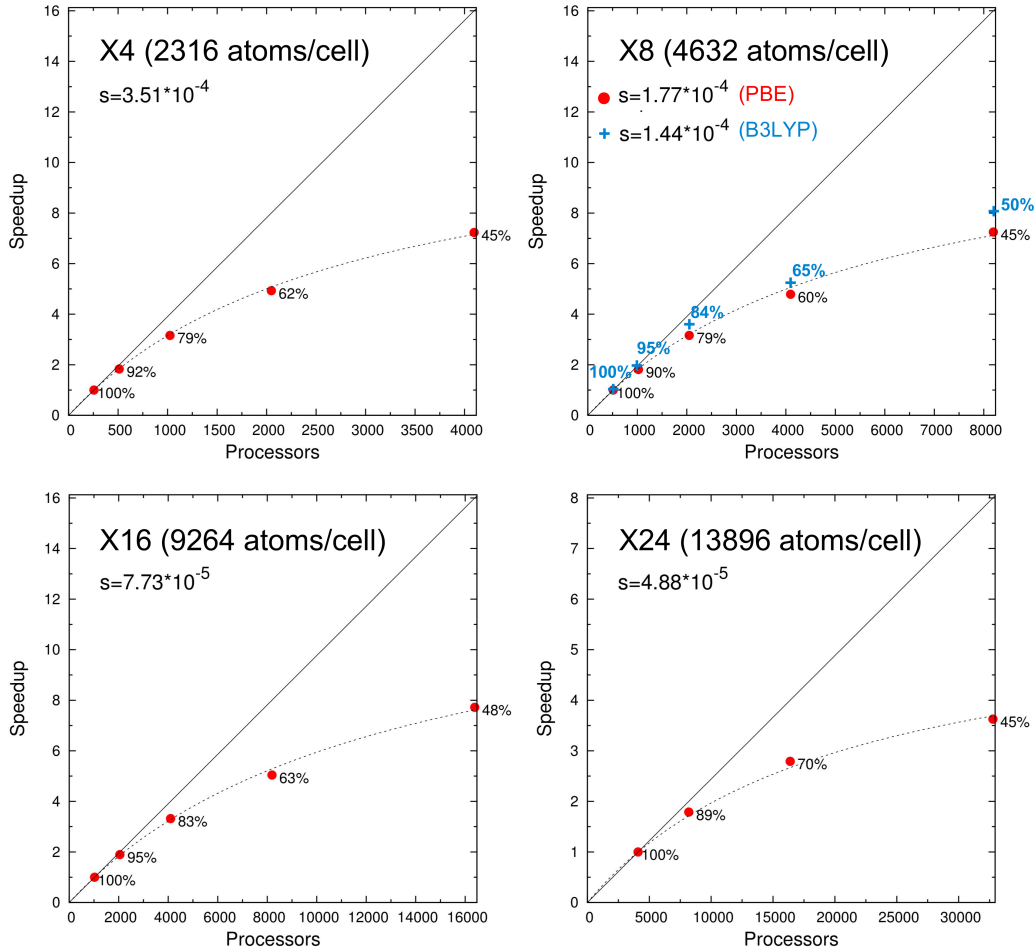

FIG. 2: Wall-clock time speedup of a SCF cycle with MPP CRYSTAL as a function of the number of cores used for the X4 (top left panel), X8 (top right panel), X16 (bottom left panel), and X24 (bottom right panel) supercells of the MCM-41 model (PBE functional, red circles). The baseline used in the definition of the speedup is 256 cores for X4, 512 for X8, 1024 for X16, and 4096 cores for X24. The dashed line shows the fit of the obtained speedup values to Amdahl's law. At each point, the scaling efficiency is reported (in %), where the diagonal of the plot corresponds to the ideal scaling. For X8, the speedup is also reported for B3LYP calculations (blue crosses).

The most useful tool for this kind of analysis are *strong scaling* profiles, which report the wall-clock time speedup as a function of the number of processors used for a system of fixed size. The strong scaling of a SCF cycle with the latest version of MPP CRYSTAL is documented in the four panels of Figure 2 for four systems (X4, X8, X16 and

TABLE I: Time needed t (in minutes) to perform a SCF cycle with the PBE functional for systems of different size (X4, X8, X16 and X24) by running the calculation with different numbers of processors N_{CPU}.

X4		X8		X16		X24	
N_{CPU}	t (m)	N_{CPU}	t (m)	N_{CPU}	t (m)	N_{CPU}	t (m)
256	6.1	512	17.2	1024	58.7	-	-
512	3.3	1024	9.5	2048	31.0	4096	57.2
1024	1.9	2048	5.4	4096	17.7	8192	32.0
2048	1.2	4096	3.6	8192	11.6	16384	20.5
4096	0.8	8192	2.4	16384	7.6	32768	15.8

X24), when use is made of the PBE functional (red circles). The scaling of X4 (2316 atoms/cell) is documented from 256 to 4096 processors. The maximum number of processors used is then doubled for each system of increasing size, up to 32768 processors for the X24 system, which contains 13896 atoms/cell. The baseline used in the definition of the speedup is 256 cores for X4, 512 for X8, 1024 for X16, and 4096 cores for X24. At each point, the scaling efficiency is reported (in %), where the diagonal of the plot corresponds to the ideal scaling. The resulting plots, on their different scales, look similar to one another and they reach approximately the same efficiency for the largest number of processors considered in each case (45% - 50%). The obvious conclusion is that the optimal number of cores to run the calculation on is proportional to the size of the system. For instance, a speedup efficiency of about 80% is obtained by running X4 (2316 atoms/cell) on 1024 processors, or X8 (4632 atoms/cell) on 2048 processors, or X16 (9264 atoms/cell) on 4096 processors.

Overall, the wall-clock time speedup as a function of the number of adopted processors shows a regular behavior and a high scalability efficiency, due to the very high degree of parallelization of the whole code. By following the prescription by Corsetti,[7] the last statement can be quantified by fitting the speedup values as a function of the number of processors to the popular Amdahl's model, which assumes that the code can be divided into a perfectly parallelized fraction p, and into a complementary sequential fraction $s = 1 - p$.[22] Dashed lines in Figure 2 represent the fitting to Amdahl's law. Despite the oversimplification introduced by this model, the measured speedup values are found to be nicely described by Amdahl's law. In each panel of the figure, the value of the fraction of the non-parallelized code s is reported, as determined from the fitting (the smaller the s values, the better the parallelization of the code). From these results it can be seen that: i) the value of s decreases as the size of the system increases because for large systems the most expensive tasks are dominating, which are typically well-parallelized (general feature of most codes);[23]

ii) very low absolute values of s are obtained, showing the high degree of parallelization of the calculation.

The speedup of the code for hybrid functionals (B3LYP in this case) is documented in the top-right panel of Figure 2 for the X8 system (blue crosses). The scaling of the wall-clock time with respect to the adopted number of processors for hybrid functionals turns out to be slightly more efficient than for pure DFT functionals. This is due to the fact that the typical additional computational cost of the inclusion of exact Fock exchange is localized in those parts of the code which display the most efficient scalability (i.e. integral evaluation). Indeed, the computational cost of Fock matrix diagonalization is the same for pure DFT and hybrid functionals.

The total time needed (per SCF cycle) to run the PBE calculations discussed above is reported in Table I for the four systems X4, X8, X16 and X24 and for the different numbers of processors used.

C. System Size Scaling

In Section II B we have discussed the scalability of the program with respect to the number of adopted processors for a system of fixed size. In order to be able to study large systems, another scaling is crucial: the scaling of the wall-clock time with respect to the system size. Ideally, one would aim at a linear scaling with system size. However, this is not possible for conventional Kohn-Sham DFT codes, which evaluate eigenvalues and eigenvectors either by explicit diagonalization (as done in MPPcrystal via the Divide & Conquer algorithm implemented in ScaLAPACK) or by iterative minimization, as both these algorithms scale cubically with system size. Linear-scaling methods, which make use of approximate spatial truncations based on the principle of electronic nearsightedness, have been proposed in recent years to overcome this limitation.[24,25]

When use is made of a localized basis set (as in Crystal), diagonalization turns out to be fairly inexpensive for relatively small systems so that other tasks possibly characterized by a non-linear scaling would become the limiting steps. Until a few years ago, this was indeed the case in Crystal because of the quadratic scaling of the routines devoted to the selection (i.e pre-screening) of the two-electron integrals to be evaluated. This was due to the fact that some preliminary quantities were always computed before the decision whether an integral should be evaluated or not was made. This inefficiency in the selection implied an almost negligible overhead for small- or medium-sized systems but a progressively more significant one for large systems. The routines for the selection of Coulomb and exchange two-electron integrals have now been revised in such a way to compute the same set of integrals as before (i.e. results are unchanged) with an improved (linear) scaling with respect to system size. This improvement turns out to be particularly relevant for calculations performed with hybrid functionals,
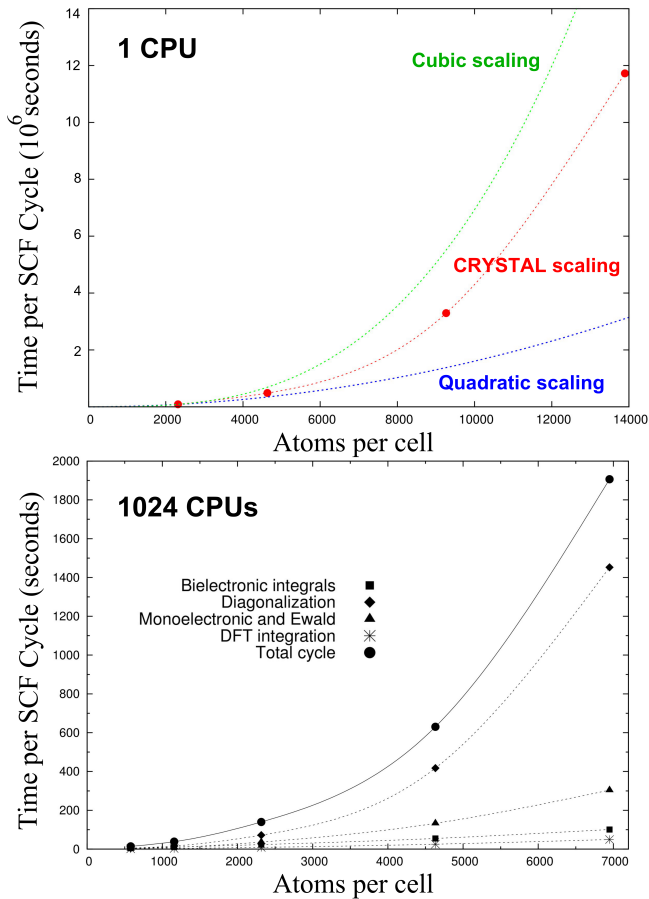
FIG. 3: (Bottom panel) Wall-clock time (in seconds) of a SCF cycle run in parallel over 1024 processors for MCM-41 supercells of increasing size with the hybrid B3LYP functional. The time required by the main computational tasks at each SCF cycle are also separately reported. (Upper panel) The same quantity as extrapolated to a single processor by use of Amdahl's law.

where the evaluation of the exact Fock exchange implies an overall increased computational cost for the integral evaluation (by a factor from 2 to 5 depending on the system) with respect to pure DFT functionals.

The selection routines have also been optimized for systems with no point-symmetry so as to reduce the number of integrals to be evaluated. The CRYSTAL program was originally implemented in such a way to take full advantage of point-symmetry equivalences rather than permutation equivalences among two-electron integrals. It follows that, when no point-symmetry was present, this strategy provided no computational gain. In the latest version of the program, two selection routines are available: one for systems with rich point-symmetry where symmetry equivalences are exploited and one for systems with no point-symmetry where advantage is now taken of permutational equivalences. As a consequence, the revised selection not only improves the system size scaling but also reduces the absolute cost of the procedure for systems with no point-symmetry. For example, for

MCM-41, the computational cost for integral evaluation with a hybrid functional was 4.5 times larger than with a pure functional, which is now reduced to a factor of 2.6 by the new strategy.

The wall-clock time required for a SCF cycle of the MCM-41 system with the hybrid B3LYP functional is reported in Figure 3 as a function of the size of the adopted supercell. Timings of different computational subtasks at each SCF cycle are also reported separately. The overall evaluation of two-electron integrals shows a linear scaling, thus documenting the effectiveness of the improvements discussed above. While integral evaluation still represents the most demanding task (taking 36% of the total time) for the primitive cell X1 (579 atoms/cell), it progressively becomes less critical as it falls to just 6% for the X12 case (6948 atoms/cell). As expected, for large systems the time required by Fock matrix diagonalization now dominates the overall scaling. The evaluation of one-electron integrals and long-range Coulomb interactions through the Ewald method (almost negligible for small systems) scales quadratically with the system size and thus becomes the second most expensive task for large cells.

Let us note that, generally, system size scalings as the one reported in Figure 3 might be affected by the choice made for the number of processes used to run the calculations on (1024 in this case) through the parallel scaling of the code, which typically corresponds to different efficiencies for systems of different size. Strong scaling profiles (as discussed in Section II B) come to the rescue as they allow us to extrapolate each system to the single-process wall-clock time (characterized by an efficiency of 100% by definition for all systems). The upper panel of Figure 3 reports the total wall-clock time of a SCF cycle as a function of the system size, as extrapolated to the single-process case. The overall scaling of MPPCRYSTAL is seen to be somehow between quadratic and cubic (closer to quadratic for systems smaller than about 8000 atoms/cell in this case and closer to cubic for larger systems).

III. OVERVIEW OF STUDIED SYSTEMS

The parallel versions of CRYSTAL have been extensively used in recent years to perform quantum-mechanical simulations of structural, energetic, electronic, thermodynamic, spectroscopic, optical, elastic, piezoelectric, topological properties of a large variety of systems characterized by different chemical features and periodicity (0D, 1D, 2D and 3D). In this section, we shall briefly review those studies where the largest systems have been investigated, so as to provide an overview of the current capabilities of the code. A graphical representation of the structure of the systems to be discussed below is given in Figure 4, where their size is also indicated.
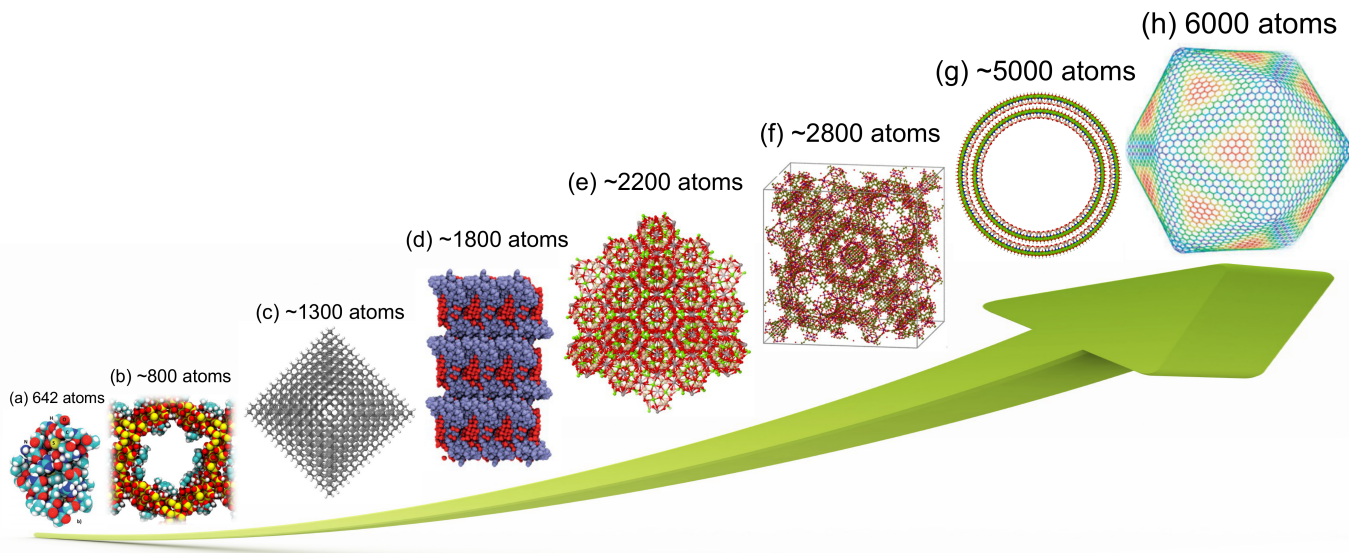
FIG. 4: Selected large-scale systems studied with the parallel versions of the CRYSTAL program in recent years. The size of each system is reported in terms of number of atoms per cell (or per molecule/cluster). See the text for a brief account of the respective studies.

As an example of non-periodic 0D systems, the structural, electronic and energetic properties of a family of (n,n) giant icosahedral carbon fullerenes have been investigated as a function of n (an integer number that determines the size of the fullerene). The largest system that was studied corresponds to $n = 10$ and is constituted by 6000 atoms. A basis set of double-zeta quality with 14 atomic orbitals per carbon atom was used (corresponding to a total of 84000 AOs for the largest system) and the global hybrid B3LYP functional adopted. Because of the particularly rich symmetry of this family of fullerenes (120 symmetry operators), a single-point SCF calculation for this system could be run on a single processor with 2 GB of memory in about 24 hours. The shape of giant carbon fullerenes was shown to present hybrid features, being closer to a sphere for small values of n and closer to a polyhedron for larger values of n. At small n, hexagonal rings at face centers are strained with respect to the graphene reference while as n increases, they recover a planar, graphene-like, configuration. Figure 4 (h) shows a graphical representation of the atomic structure of the giant (10,10) icosahedral carbon fullerene, where the color scale measures the local strain in the bond lengths. The electronic energy gap E_g is predicted to become zero for $n \geq 34$.[21]

In a recent study, several features (structural, energetic, electronic, vibrational spectroscopic) of non-periodic 0D clusters of defective diamond have been investigated at the hybrid B3LYP level as a function of their increasing size (containing up to 1293 atoms). A graphical representation of the atomic structure of the largest cluster is given in Figure 4 (c). Two different defects were put in the center of the clusters: a vacancy and the $\langle 100 \rangle$ split self-interstitial, whose structural, electronic and vibrational features were recently investigated with periodic calculations.[26,27] The SCF energy calculation for the largest cluster took 1.7 hours when run in parallel over 256 processors. The convergence of the computed properties of clusters of increasing size to the predictions of periodic calculations based on the supercell approach was discussed. While structural and energetic features of the defects are found to be rather local and show a fast convergence with respect to the size of the model, other properties, such as the electronic band gap and the spectroscopic Raman fingerprint of such a covalently-bonded system, are much more collective in nature and do show a very slow convergence to the bulk values.[28]

As an example of a 1D periodic system, we refer to the recent investigation of structural and energetic properties of single-walled chrysotile, $Mg_3Si_2O_5(OH)_4$, nanotubes of increasing size, containing up to 5004 atoms per unit cell and corresponding to a tube radius up to 205 Å. A graphical representation of the atomic structure of such a nanotube, as obtained by rolling up a layered mineral, is given in Figure 4 (g). Hybrid B3LYP calculations were performed to study their structural and energetic properties. The stability with respect to the flat slab of lizardite of the rolled structures was investigated as a function of n. Two nanotube families, $(n,-n)$ and (n,n), were con-

sidered with different chirality. The appearance of clear minima in the energy profile as a function of the tube radius at about 89 Å was observed for both families.[29]

A cutting-edge study has recently been performed, by exploiting the capabilities of MPPCRYSTAL, of structural and energetic properties of the crystalline form of a "small" structural protein: crambin, which is a thionin hydrophobic protein with 46 aminoacids with a distinct secondary structure characterized by both α-helices and β-sheets. The molecular structure of the crambin protein (642 atoms) is graphically given in Figure 4 (a). The unit cell of its crystalline form hosts two protein molecules. A single-point SCF plus gradient calculation on this system takes about 36 minutes if run in parallel over 240 processors, about 21 minutes over 480 processors, about 10 minutes over 1440 processors. Two models of further structural complexity were considered by including an increasing number of explicit solvating water molecules, up to 172, which brought the system to contain up to about 1800 atoms and 16482 AOs per cell. A graphical representation of this system is given in Figure 4 (d). Use of the B3LYP global hybrid functional was made in combination with a semi-empirical correction for dispersive interactions. The structure was fully-relaxed at the quantum-mechanical level and found to agree to a large extent to that experimentally determined. The crystal formation, protein-water, and protein-protein interaction energies could also be computed.[30]

When the thermodynamic properties a material have to be accurately determined, phonon dispersion must be computed, which requires the use of fairly large supercells if a "direct-space" approach is adopted, as done in the CRYSTAL program. In a recent investigation, fully-converged thermodynamic properties of two end-members of the silicate garnet family of rock-forming minerals (Pyrope $Mg_3Al_2Si_3O_{12}$ and Grossular $Ca_3Al_2Si_3O_{12}$) have been computed from harmonic phonon dispersion.[31] The hybrid B3LYP functional has been used. The primitive cell of these cubic systems contains 80 atoms; supercells Xn of increasing size have been built and vibration frequencies computed for each of them. The largest system, X27, corresponds to a $3\times3\times3$ super-cell containing 2160 atoms (see Figure 4 (e) for a graphical representation of its structure) and sampling the phonon dispersion over 27 **k**-points within the Brillouin zone in reciprocal space. DFT methods are intrinsically size-consistent. However, accuracy of the results of a complex numerical procedure strongly depends on the implementation robustness and the choice of the computational parameters. In order to check the size-consistency of the implemented algorithms in the CRYSTAL code, two properties that are invariant with respect to the unit cell size (provided that a consistently scaled grid is used for reciprocal space sampling) were considered: the total energy per formula unit and the full set of wave-numbers associated with the normal vibration modes at the Γ point. They were computed for six different supercells, namely X1, X2, X4, X8, X16 and X27.

TABLE II: Consistency of the total energy and vibration wave-numbers of Pyrope in supercells of increasing number of atoms (N) with reference to X27. ΔE is the total energy difference per formula unit (in micro-hartree); $\Delta\nu_{Max}$ and RMS (in cm^{-1}) are the maximum difference and the Root Mean Square of the full set of wave-numbers obtained in Γ. The various supercells belong to different lattice types: primitive (P), body-centered (I) and face-centered (F). s denotes the shrinking factor defining the Pack-Monkhorst net for each cell.

SC	Lattice	N	s	ΔE	$\Delta\nu_{Max}$	RMS(ν)
X1	I	80	3	1.22	0.357	0.050
X2	P	160	3	1.64	0.066	0.016
X4	F	320	2	2.15	0.032	0.009
X8	I	640	2	0.99	0.025	0.007
X16	P	1280	1	0.74	0.012	0.001
X27	I	2160	1	-	-	-

Results are reported in Table II. The fifth column in the table reports the total energy difference per formula unit (ΔE, in μHa) between X27 and the smaller cells. In spite of the fairly large size of the cells, the total energy per formula unit can be computed with extreme accuracy in all cases, fluctuations being in the order of the μHa. It must be remarked that the total energy is not only almost size-invariant, it is also shape-invariant as the various supercells along the series span different kinds of cubic lattices: body-centered (I), primitive (P) and face-centered (F). Similar considerations hold true also for the full set of wave-numbers at the Γ point. The two columns on the right of Table II give the maximum differences in the computed wave-numbers observed along the series in the range between X1 and X27 and their root mean square. All such indices correspond to just a fraction of 1 cm^{-1}: a tiny uncertainty. Let us stress that the documentation of evidence of the numerical size-consistency for implemented quantum-mechanical methodologies is seldom reported and yet is crucial to any reliable large-scale simulation.

The capabilities of MPPCRYSTAL have also been exploited to shed some light on the atomistic details of the interaction between ibuprofen (one of the most common non-steroidal anti-inflammatory drugs) and a realistic model of MCM-41 (one of the most studied mesoporous silica materials for drug delivery). A structural model of about 800 atoms/cell was designed, which is graphically reported in Figure 4 (b). Quantum-mechanical calculations (at the B3LYP-D level of theory) were performed to fully-relax the structure and to simulate the infrared spectrum of the system.[32]

Structural, energetic and host/guest features of a giant metal-organic framework, MIL-100, containing 2812 atoms per unit cell (corresponding to 50256 AOs per cell) were also investigated. A graphical representation of the structure of this system is given in Figure 4 (f). In particular, the interaction energy for CO_2 adsorption at dif-

ferent sites of the cavities was computed.

IV. CONCLUSIONS AND PERSPECTIVES

The performance of the latest version of CRYSTAL, an *ab initio* DFT code for condensed mater systems, has been documented with particular emphasis on the study of large systems containing hundreds or thousands of atoms per cell. In particular: i) strong scaling profiles have been reported up to 32768 cores; ii) memory requirements have been quantified, which show that calculations on systems containing several thousands atoms (up to 14000 atoms/cell) can be run with less than 2 GB/core in parallel over just a few thousand cores (1024, 2048 or 4096 CPUs) on HPC machines; iii) the system size scaling has been reported, where the contribution from the two-electron integrals (including integral selection) has been documented to be linear; iv) the high numerical size-consistency of the implemented algorithms has been quantified; v) recent applications of the various parallel strategies of the program have been reviewed.

In order to further improve the performance and capabilities of the massively-parallel version of CRYSTAL, work is currently in progress in different respects: i) the memory occupancy is being further reduced by the use of shared memory to store a number of "write once, read many" arrays, which are currently replicated. This is implemented by use of the facilities provided within the MPI-3 standard, and will allow storing of just one copy of each array per node, rather than many identical copies as at present; ii) the coupled-perturbed-Hartree-Fock/Kohn-Sham procedure is being parallelized according to a distributed-data strategy, which would make it possible to study many properties of large systems (linear and non-linear optical properties, Raman intensities, second-harmonic generation, Pockels effect, and others); iii) a massively-parallel version is being finalized of the module that computes the electronic band structure and the density-of-states.

[*] Electronic address: `alessandro.erba@unito.it`

[1] Hafner, J.; Wolverton, C.; Ceder, G. Toward Computational Materials Design: The Impact of Density Functional Theory on Materials Research. *MRS Bulletin* **2006**, *31*, 659–668.

[2] Baroni, S.; de Gironcoli, S.; Corso, A. D.; Giannozzi, P. Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.* **2001**, *73*, 515.

[3] Karki, B. B.; Stixrude, L.; Wentzcovitch, R. M. High-pressure elastic properties of major materials of earth's mantle from first principles. *Rev. Geophys.* **2001**, *39*, 507–534.

[4] Landers, J.; Gor, G. Y.; Neimark, A. V. Density functional theory methods for characterization of porous materials. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* **2013**, *437*, 3 – 32.

[5] Jain, A.; Hautier, G.; Moore, C. J.; Ong, S. P.; Fischer, C. C.; Mueller, T.; Persson, K. A.; Ceder, G. A high-throughput infrastructure for density functional theory calculations. *Comput. Mater. Sci.* **2011**, *50*, 2295 – 2310.

[6] Varini, N.; Ceresoli, D.; Martin-Samos, L.; Girotto, I.; Cavazzoni, C. Enhancement of DFT-calculations at petascale: Nuclear Magnetic Resonance, Hybrid Density Functional Theory and Car-Parrinello calculations. *Comput. Phys. Commun.* **2013**, *184*, 1827 – 1833.

[7] Corsetti, F. Performance Analysis of Electronic Structure Codes on HPC Systems: A Case Study of SIESTA. *PLoS ONE* **2014**, *9*, e95390.

[8] Hutter, J.; Iannuzzi, M.; Schiffmann, F.; VandeVondele, J. CP2K: atomistic simulations of condensed matter systems. *WIREs: Comput. Mol. Sci.* **2014**, *4*, 15–25.

[9] Maniopoulou, A.; Davidson, E. R.; Grau-Crespo, R.; Walsh, A.; Bush, I. J.; Catlow, C. R. A.; Woodley, S. M. Introducing k-point parallelism into VASP. *Comput. Phys. Commun.* **2012**, *183*, 1696–1701.

[10] Kendall, R. A.; Aprá, E.; Bernholdt, D. E.; Bylaska, E. J.; Dupuis, M.; Fann, G. I.; Harrison, R. J.; Ju, J.; Nichols, J. A.; Nieplocha, J.; Straatsma, T.; Windus, T. L.; Wong, A. T. High performance computational chemistry: An overview of NWChem a distributed parallel application. *Comput. Phys. Commun.* **2000**, *128*, 260 – 283.

[11] Orlando, R.; Delle Piane, M.; Bush, I. J.; Ugliengo, P.; Ferrabone, M.; Dovesi, R. A new massively parallel version of CRYSTAL for large systems on high performance computing architectures. *J. Comput. Chem.* **2012**, *33*, 2276–2284.

[12] Bush, I. J.; Tomic, S.; Searle, B. G.; Mallia, G.; Bailey, C. L.; Montanari, B.; Bernasconi, L.; Carr, J. M.; Harrison, N. M. Parallel implementation of the ab initio CRYSTAL program: electronic structure calculations for periodic systems. *Proc. R. Soc. A: Math. Phys. Eng. Sci* **2011**, *467*, 2112.

[13] Dovesi, R.; Civalleri, B.; Orlando, R.; Roetti, C.; Saunders, V. R. CRYSTAL: a computational tool for the ab initio study of the electronic properties of crystals. *Rev. Comp. Chem.* **2005**, *21*, 1.

[14] Orlando, R.; De La Pierre, M.; Zicovich-Wilson, C. M.; Erba, A.; Dovesi, R. On the Full Exploitation of Symmetry in Periodic (as well as Molecular) Self-Consistent-Field Ab initio Calculations. *J. Chem. Phys.* **2014**, *141*, 104108.

[15] Dovesi, R.; Orlando, R.; Civalleri, B.; Roetti, C.; Saunders, V. R.; Zicovich-Wilson, C. M. CRYSTAL: a computational tool for the ab initio study of the electronic properties of crystals. *Z. Kristallogr.* **2005**, *220*, 571–573.

[16] Dovesi, R.; Orlando, R.; Erba, A.; Zicovich-Wilson, C. M.; Civalleri, B.; Casassa, S.; Maschio, L.; Ferrabone, M.; De La Pierre, M.; D'Arco, Ph.; Noël, Y.; Causá, M.; Rérat, M.; Kirtman, B. CRYSTAL14: A Program for the Ab initio Investigation of Crystalline Solids. *Int. J. Quantum Chem.* **2014**, *114*, 1287–1317.

[17] Casassa, S.; Erba, A.; Baima, J.; Orlando, R. Electron Density Analysis of Large (Molecular and Periodic) Sys-

(17) tems: A Parallel Implementation. *J. Comput. Chem.* **2015**, *36*, 1940–1946.

(18) Dovesi, R. On the Role of Symmetry in the Ab Initio Hartree-Fock Linear-Combination-of-Atomic-Orbitals Treatment of Periodic Systems. *Int. J. Quantum Chem.* **1986**, *29*, 1755–1774.

(19) Zicovich-Wilson, C.; Dovesi, R. On the Use of Symmetry-Adapted Crystalline Orbitals in SCF-LCAO Periodic Calculations. I. The Construction of the Symmetrized Orbitals. *Int. J. Quantum Chem.* **1998**, *67*, 299–309.

(20) Zicovich-Wilson, C. M.; Noël, Y.; Ferrari, A. M.; Orlando, R.; De La Pierre, M.; Dovesi, R. On the use of symmetry in SCF calculations. The case of fullerenes and nanotubes. *AIP Conf. Proc.* **2012**, *1456*, 248–255.

(21) Noel, Y.; De La Pierre, M.; Zicovich-Wilson, C. M.; Orlando, R.; Dovesi, R. Structural, electronic and energetic properties of giant icosahedral fullerenes up to C6000: insights from an ab initio hybrid DFT study. *Phys. Chem. Chem. Phys.* **2014**, *16*, 13390–13401.

(22) Amdahl, G. M. Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities. Proceedings of the April 18-20, 1967, Spring Joint Computer Conference. New York, NY, USA, 1967; pp 483–485.

(23) Gustafson, J. L. Reevaluating Amdahl's Law. *Commun. ACM* **1988**, *31*, 532–533.

(24) Bowler, D.; Miyazaki, T. $\mathcal{O}$(N) methods in electronic structure calculations. *Rep. Progr. Phys.* **2012**, *75*, 036503.

(25) VandeVondele, J.; Borŝtnik, U.; Hutter, J. Linear scaling self-consistent field calculations with millions of atoms in the condensed phase. *J. Chem. Theor. Comput.* **2012**, *8*, 3565–3573.

(26) Baima, J.; Zelferino, A.; Olivero, P.; Erba, A.; Dovesi, R. Raman Spectroscopic Features of the Neutral Vacancy in Diamond from Ab Initio Quantum-mechanical Calculations. *Phys. Chem. Chem. Phys.* **2016**, *18*, 1961–1968.

(27) Salustro, S.; Erba, A.; Zicovich-Wilson, C.; Nöel, Y.; Maschio, L.; Dovesi, R. Infrared and Raman Spectroscopic Features of the Self-Interstitial Defect in Diamond from Exact-Exchange Hybrid DFT Calculations. *Phys. Chem. Chem. Phys.* **2016**, *18*, 21288–21295.

(28) Salustro, S.; Ferrari, A.; Orlando, R.; Dovesi, R. Comparison Between Cluster and Supercell Approaches: The Case of Defects in Diamond. *Theor. Chem. Acc.* **2017**, *136*, 42.

(29) Demichelis, R.; De La Pierre, M.; Mookherjee, M.; Zicovich-Wilson, C. M.; Orlando, R. Serpentine polymorphism: a quantitative insight from first-principles calculations. *CrystEngComm* **2016**, *18*, 4412–4419.

(30) Delle Piane, M.; Corno, M.; Orlando, R.; Dovesi, R.; Ugliengo, P. Elucidating the fundamental forces in protein crystal formation: the case of crambin. *Chem. Sci.* **2016**, *7*, 1496–1507.

(31) Baima, J.; Ferrabone, M.; Orlando, R.; Erba, A.; Dovesi, R. Thermodynamics and Phonon Dispersion of Pyrope and Grossular Silicate Garnets from Ab initio Simulations. *Phys. Chem. Minerals* **2016**, *43*, 137–149.

(32) Delle Piane, M.; Corno, M.; Pedone, A.; Dovesi, R.; Ugliengo, P. Large-Scale B3LYP Simulations of Ibuprofen Adsorbed in MCM-41 Mesoporous Silica as Drug Delivery System. *J. Phys. Chem. C* **2014**, *118*, 26737–26749.