Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Bachelorthesis

Optimization of the Quantum Espresso Density Functional Theory Code for parallel execution on the PHYSnet-Cluster

vorgelegt von

Tjark Sievers

Fakultät: Mathematik, Informatik und Naturwissenschaften
Fachbereich: Physik
Studiengang: Physik
Matrikelnummer: 7147558
Erstgutachter: Prof. Dr. Tim Wehling
Zweitgutachterin: Prof. Dr. Daria Gorelova

Todo list

Kurzzusammenfassung

Abstract

Contents

Motivation

I Many-body physics

I.1 The electronic structure problem

In solid state physics, the Hamiltonian describing the interacting nuclei and electrons in the solid is well known, as all interactions except Coulomb interaction can safely be ignored at the mass and energy scales at which the electrons and nuclei reside. This still very general problem consisting of both the electronic and nuclei degrees of freedom can be simplified in a first step by employing the Born-Oppenheimer approximation. The approximation assumes the nuclei to be fixed point charges which create a potential for the N interacting electrons, so that the electronic part can be solved independently using the nuclei positions $\mathbf{R}_\alpha$ as a parameter.

The problem is then described by the time-independent Schrödinger equation

$$\hat{H}\Psi(\mathbf{r}_1, \ldots, \mathbf{r}_N) = E\Psi(\mathbf{r}_1, \ldots, \mathbf{r}_N) \tag{I.1}$$

with the Hamiltonian in first quantization (i running over the electrons, α, β over the nuclei)

$$\hat{H} = \hat{T}_e + \hat{U}_{e-e} + \hat{V}_{n-e} + \hat{W}_{n-n} \tag{I.2}$$

$$= -\sum_i \frac{1}{2}\nabla_i^2 + \frac{1}{2}\sum_{i \neq j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} - \sum_i \sum_\alpha \frac{Z_\alpha}{|\mathbf{r}_i - \mathbf{R}_\alpha|} + \frac{1}{2}\sum_{\alpha\beta} \frac{Z_\alpha Z_\beta}{R_{\alpha\beta}} \tag{I.3}$$

where:

- $\hat{T}_e$ is the kinetic energy of the electrons
- $\hat{U}_{e-e}$ is the Coulomb interaction between the electrons and
- $\hat{V}_{n-e}$ is the potential energy of the electrons in the field of the nuclei
- $\hat{W}_{n-n}$ is the Coulomb interaction between the nuclei

The terms $\hat{V}_{n-e}$ and $\hat{W}_{n-n}$ can then be combined into an external potential V for the interacting electrons, so that the Hamiltonian reads

$$\hat{H} = \hat{T} + \hat{U} + \hat{V} \tag{I.4}$$

This Hamiltonian will be used in the further development of the underlying theory for this thesis.

I.2 Density Functional Theory

A direct solution to the electronic structure problem, this meaning obtaining the ground-state many-body wave function $\Psi(\mathbf{r_1}, \ldots, \mathbf{r}_N)$ for a given potential is analytically impossible even

for a small number of electrons compared to the number of electrons in a macroscopic crystal. As such, the need for good approximations to obtain results for real world systems is high. One particularly successful approach is Density Functional Theory (DFT). In the following section, the theoretical framework of DFT will be developed, the outline of which can be found in any literature on solid state physics [1].

I.2.1 Hohenberg-Kohn theorems

The starting for DFT is the exact reformulation of the outlined electronic structure problem by Hohenberg and Kohn [2]. This reformulation uses the ground state density of the electronic system $n_o(r)$ as the basic variable. To achieve this, Hohenberg and Kohn [2] formulated two theorems, which demonstrate that the ground state properties of an electronic system can be described using the ground state density (the proof of those theorems is omitted here, but can be found in the original paper [2] or any publication on DFT [1]):

I The external potential (and via the Schrödinger equation also the ground state wave function and the ground state energy) is a unique functional of the ground state density (except for an additive constant).

II The ground state energy minimizes the energy functional,

$$E[n(\mathbf{r})] > E_0 \; \forall \; n(\mathbf{r}) \neq n_0(\mathbf{r})$$

The proof of those theorems show the existence of the energy functional $E[n(\mathbf{r})]$, but a concrete expression for it cannot be given. As the ground state wave function is a functional of the ground state density, a formal definition of the energy functional can be written as

$$\begin{aligned} E[n(\mathbf{r})] &= \langle \Psi | \hat{H} | \Psi \rangle \\ &= \langle \Psi | \hat{T} + \hat{U} + \hat{V} | \Psi \rangle \\ &= \langle \Psi | \hat{T} + \hat{U} | \Psi \rangle + \int \mathrm{d}\mathbf{r}' \, \Psi^*(\mathbf{r}') V(\mathbf{r}') \Psi(\mathbf{r}') \end{aligned}$$

Defining the universal functional $F[n(\mathbf{r})] = \langle \Psi | T + U | \Psi \rangle$, which is material independent and writing $n(\mathbf{r}') = \Psi^*(\mathbf{r}')\Psi(\mathbf{r}')$, the energy functional becomes

$$E[n(\mathbf{r})] = F[n(\mathbf{r})] + \int \mathrm{d}\mathbf{r}' \, V(\mathbf{r}') n(\mathbf{r}') \tag{I.5}$$

This is just a formal definition, as all the formerly mentioned complication of the Hamiltonian I.4 now lie in the functional $F[n(\mathbf{r})]$. With a known or well approximated universal functional $F[n(\mathbf{r})]$, the Hohenberg-Kohn theorems provide a great simplification for finding the ground state properties of a solid state system, as the problem is now only a variational problem with 3 spatial coordinates instead of $3N$ coordinates when trying to solve the full Hamiltonian.

I.2.2 Kohn-Sham equations

One way of approximating the functional $F[n]$ was given by Kohn and Sham [3]. The idea is to use a non-interacting auxiliary system of electrons

$$H_0 = \sum_i^{N_e} \frac{p_i^2}{2m} + v_{KS}(\mathbf{r}_i) \tag{I.6}$$

With a correction potential v_{KS} such that the ground state charge density for the auxiliary and the interacting system are the same. This introduces a new set of orthonormal wave functions, the solutions to the non-interacting problem Ψ_i. The kinetic energy of such a non-interacting problem can be easily calculated as sum over all electrons:

$$T_S[n(\mathbf{r})] = -\frac{1}{2}\sum_{i=1}^{N_e} \int \mathrm{d}^3 r\, \Psi_i^*(\mathbf{r}) \Delta \Psi_i(\mathbf{r}) \tag{I.7}$$

with the density

$$n(\mathbf{r}) = \sum_{i=1}^{N} |\Psi_i(\mathbf{r})|^2 \tag{I.8}$$

With the help of this system and the classical electrostatic energy

$$E_H[n(\mathbf{r})] = \frac{1}{2} \int\int \mathrm{d}^3 r_1\, \mathrm{d}^3 r_2\, \frac{n(\mathbf{r}_1) n(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} \tag{I.9}$$

an ansatz for the universal functional can be written as

$$F[n(\mathbf{r})] = T_S[n(\mathbf{r})] + E_H[n(\mathbf{r})] + E_{XC}[n(\mathbf{r})] \tag{I.10}$$

where now $E_{XC}[n(\mathbf{r})]$ is a functional of the density accounting for all exchange and correlation effects not present in the non-interacting electron system. The success of DFT lies in the fact that E_{XC} only contributes only a small part of the total energy and can be surprisingly well approximated.

Using that form of $F[n(\mathbf{r})]$, from the variational problem (with a Lagrange parameter introduced to ensure the orthonormality of the states Ψ_i)

$$\delta \left(E[n(\mathbf{r})] - \sum_j \lambda_j \left[\int \mathrm{d}^3 r\, |\Psi_j|^2 - 1 \right] \right) = 0 \tag{I.11}$$

single particle, Schrödinger-like equations can be derived:

$$\left(-\frac{1}{2}\Delta + \frac{1}{2} \int \mathrm{d}^3 r'\, \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} + \frac{\delta E_{XC}}{\delta n(\mathbf{r})} + V \right) \Psi_i(\mathbf{r}) = \lambda_i \Psi_i(\mathbf{r}) \tag{I.12}$$

Schrödinger-like in this context means, that with the identification

$$v_{KS} = V_H + V_{XC} + V = \frac{1}{2} \int \mathrm{d}^3 r'\, \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} + \frac{\delta E_{XC}}{\delta n(\mathbf{r})} + V \tag{I.13}$$

eq. I.12 become the KS equations:

$$\left(-\frac{1}{2}\Delta + v_{KS}\right)\Psi_i(\mathbf{r}) = \epsilon_i \Psi_i(\mathbf{r}) \tag{I.14}$$

Importantly, the potential v_{KS} depends on the solutions $\Psi_i(\mathbf{r})$, as the Hartree potential V_H and the XC potential V_{XC} include the density $n(\mathbf{r})$, so the problem becomes a self-consistency problem.

I.2.3 Pseudopotentials and basis set

In order to represent the states and operators in eq. I.12, a basis set has to be chosen. Bloch's theorem states that the in case of a periodic external potential, which makes the Hamiltonian commute with translation operators for translation by a lattice vector, the common eigenstates of these operators are:

$$\Psi(\mathbf{r}) = \Psi_{n\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r}) \tag{I.15}$$

where $\mathbf{k}$ is the quasi-momentum and $u_{n\mathbf{k}}(\mathbf{r})$ has the periodicity of the unit cell. A natural choice to represent $u_{n\mathbf{k}}(\mathbf{r})$ is the discrete set of plane waves $|\mathbf{G}\rangle$

$$\langle \mathbf{r}|\mathbf{G}\rangle = \frac{1}{\sqrt{V}} e^{i\mathbf{G}\mathbf{r}} \tag{I.16}$$

$$\implies \Psi_{n\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} c_{n\mathbf{k},\mathbf{G}} e^{i(\mathbf{k}+\mathbf{G})\mathbf{r}} \tag{I.17}$$

right like that? do the plane waves have the periodicity of the unit cell?

where $\mathbf{G}$ is a reciprocal lattice vector. With that choice of basis set, the kinetic energy is easily calculated:

$$\langle \Psi_{n\mathbf{k}}(\mathbf{r})| -\nabla^2 |\Psi_{n\mathbf{k}}(\mathbf{r})\rangle = \sum_{\mathbf{G}} c^2_{n\mathbf{k},\mathbf{G}} |\mathbf{k}+\mathbf{G}|^2 \tag{I.18}$$

Another important consequence of this choice of basis set is that the electron density (eq. I.8) now becomes an integral over the Brillouin zone, which for numerical computation has to be approximated by a sum over a finite set of k points.

concrete formula for that?

One problem of this choice of basis set lies in the fact that the lower energy core electrons are very localized around the cores and as such need a lot of basis functions. Because the higher orbitals have to be orthogonal to the lower orbitals, they must also have features on the length scale of the core electrons, which means more basis functions are needed to describe them accurately.

Importantly, the core electrons don't contribute much to the electronic structure problem, so an approach to make the calculations more economically is to introduce an effective screened potential which describes the nuclear potential as well as the states close to the nucleus. These potentials are called Pseudopotentials (PP) and the orbitals constructed with them are smooth functions and can be accurately represented with a small number of basis functions. This fact can be used in computation by introducing a *cutoff energy* E_{cutoff} which limits the number of basis functions by using only expansion coefficents with $|\mathbf{k}+\mathbf{G}| \leq E_{cutoff}$. Besides the number of k points, this is the second variable which can influences the computation time of DFT calculations.

should be (k+G) to the power of 2 right? also maybe concrete formula for that

With Fast Fourier Transforms FFT, an efficient algorithm exists to transform from (discrete) real to (discrete) reciprocal space so every expectation value can be calculated in the optimal representation: in reciprocal space for the kinetic energy and in real space for the pseudopotentials.

I.3 Density Functional Perturbation Theory

dfpt!!!

I.4 Examined systems

I.4.1 Silicon

I.4.2 TaS_2

II Computational Details

II.1 Parallel computing

The following section will give an overview of the technical aspects of running computer code (such as QUANTUM ESPRESSO) on massively parallel computing environments (such as the PHYSnet compute cluster). The information presented can be found in any textbook on parallel or high-perfomance computing [4].

II.1.1 On scalabilty

In scientific computing, one can identify two distinct reasons to distribute workloads to multiple processors:

- The execution time on a single core is not sufficient. The definition of sufficient is dependent on the specific task and can range from "over lunch" to "multiple weeks"
- The memory requirements grow outside the capabilities of a single core

In order to judge how well a task can parallelized, usually some sort of scalabilty metric is employed, for example:

- How fast can a problem be solved with N processors instead of one?
- What kind of bigger problem (finer resolution, more particles, etc.) can be solved with N processors?
- How efficiently are the resources utilized?

In this thesis, the main concern is speeding up the calculation of very time expensive calculations with a fixed problem size, so the first metric will be used to judge the quality of parallelization. This metric is called speedup and is defined as $S = \frac{T_1}{T_N}$, where T_1 is the execution time on a single processor and T_N is the execution time on N processors. In the ideal case, where all the work can be perfectly distributed among the processors, all processors need the same time for their respective workloads and don't have to wait for others processors to finish their workload to continue, the execution time on N processors would be $\frac{T_1}{N}$, so the speedup would be $S = \frac{T_1}{\frac{T_1}{N}} = N$.

In reality, there are many factors either limiting or in some cases supporting parallel code scalability. Limiting factors include:

- *Algorithmic limitations*: when parts of a calculation are mutually dependent on each other, the calculation cannot be fully parallelized

- *Bottlenecks*: in any computer system exist resources which are shared between processor cores with limitations on parallel access. This serializes the execution by requiring cores to wait for others to complete the task which uses the shared resources in question
- *Startup Overhead*: introducing parallelization into a programm necessarily introduces an overhead, e.g. for distributing data across all the processors
- *Communication*: often solving a problem requires communication between different cores (e.g. exchange of interim results after a step of the calculation). Communication can be implemented very effectively, but can still introduce a big prize in computation time

On the other hand, faster parallel code execution can come from:

- *Better caching*: optimal performance per core is achieved when all the data can be kept in cache, so distributing data among more processors may lead to better than linear scaling because the data size per processor gets smaller

A simple ansatz for modeling speedup was first derived by Gene Amdahl. Assuming the work that needs to be done is split into a part which cannot be parallelized s and a part which can be parallelized ideally p, we can normalize the serial time to 1:

$$T_1 = s + p = 1 \tag{II.1}$$

The time for solving the problem on N processors is then

$$T_N = s + \frac{p}{N} \tag{II.2}$$

The speedup is now

$$S = \frac{T_s}{T_p} = \frac{1}{s + \frac{p}{N}} = \frac{1}{s + \frac{1-s}{N}} \tag{II.3}$$

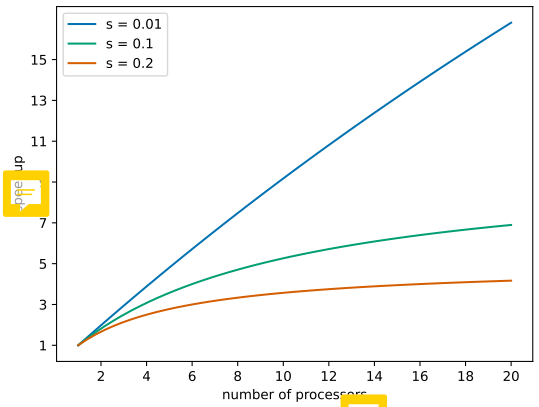


Figure II.1: *Amdahl's law for different portions of not parallelizable workload s*

This equation is called *Amdahl's law* and is plotted in fig. ?? over a range of processors for a few different values of s.

II.1.2 Evaluating the scalability of Quantum ESPRESSO calculations

In the QUANTUM ESPRESSO output, a time report listing is printed at the end. This time report includes cpu time and wall time, from those three different metrics of scalability can be calculated:

- runtime: absolute runtime (wall time) of the compute job
- speedup: runtime divided by runtime of the job on a single core
- wait time: percentage of wall time not used by QUANTUM ESPRESSO process, ~~so~~ writing to disk, waiting for IO devices or other processes, etc. (calculated as (wall time - cpu time) / wall time)

For further analysis mainly speedup will be used to evaluate the scalability of Quantum ESPRESSO calculation, because it makes comparing the scaling of calculations with different absolute runtimes easy: optimal scaling is achieved when the speedup has a slope of one, as discussed in sec. II.1.1, independent of the runtime.

Regardless, the other two parameters should also always be considered: in the end, absolute runtime is the most important factor and should govern the decision of how much resources are used for solving a particular problem. As an example, for a problem with a single core runtime of 600 s, a speedup of 100 would mean a runtime of 6 s, whereas a speedup of 200 would mean a runtime of 3 s. Even with optimal scaling, the 100 processors needed for the speedup of 200 could be considered wasted for just 3 s of saved time. On the other hand, for a problem with a single core runtime of 2400 h, the difference between a speedup of 100 (runtime 24 h) and 200 (runtime 12 h) is the difference between needing to wait a whole day for the calculation and being able to let the job run overnight and continue working in the morning, so using 100 more processors for that can be considered a good use of resources.

As for the wait time, this metric can be used to separate the different factors of poor parallelization discussed in II.1.1. Startup overhead is easy to identify, as this should be a small, near constant percentage of the absolute runtime. This of course can vary depending on how complex data distribution is, but there should at least not be a strong dependence on the number of processors, as only a small amount of communication is needed. Communication and bottlenecks on the other hand both introduce wait time which depends on the number of processors. Differentiating between them ~~is~~ relies on knowledge of the specific hardware of the system running the calculations, so how many cores are on a single chip/motherboard/node, which resources are shared between how many cores, etc.

Importantly, for this interpretation to be meaningful, the cpu and wall times reported by Quantum ESPRESSO have to accurate, because they are trusted without verification. Just as an example for the pitfalls of this approach, when executing programs on multiple processors in parallel, cpu time is measured per processor, so some kind of truncation is done when a single number (such as in Quantum ESPRESSO) is reported. Whether this is taking the average over all processors, just reporting the time for a single processor or any other kind of truncation is unclear.

However, the notion of using the difference between wall time and cpu time for evaluating the quality of parallelization is supported by the user guide for one of the Quantum ESPRESSO modules [5] (sec. 4.5), so it will also be used a qualitative measure of good parallelization in this thesis.

II.2 Quantum ESPRESSO

Quantum ESPRESSO (opEn-Source Package for Research in Electronic Structure, Simulation, and Optimization) [6, 7] is a collection of packages implementing (among others) the techniques described in sec. I.2 and ?? to calculate electronic structure properties (module `PWscf`) as well as phonon frequencies and eigenvectors (module `PHonon`).

II.2.1 Compilation of Quantum ESPRESSO

As the core of this thesis is an in depth examination of the QUANTUM ESPRESSO software and ways its performance can be optimized, a discussion of the way is is compiled is needed. The information in this section is taken from the QUANTUM ESPRESSO 7.0 user guide [8].

The QUANTUM ESPRESSO distribution is packaged with everything needed for simple, non-parallel execution, the only additional software needed are a minimal Unix environment (a shell like `bash` or `sh` as well as the utilities `make awk` and `sed`) and a Fortran compiler compliant with the F2008 standard. For parallel execution, also MPI libraries and an MPI aware compiler need to be provided.

QUANTUM ESPRESSO needs three external mathematical libraries, BLAS and LAPACK for linear algebra as well as an implementation of FFT for fourier transforms. In order to make the installation as easy as possible, QUANTUM ESPRESSO comes with a publicly available reference implementation of the BLAS routines, the publicly available LAPACK package and an older version of FFTW (Fastest Fourier Transform in the West, an open source implementation of FFT). Even though these libraries are already optimized in terms of the algorithms implemented, usage of libraries implementing the same routines which can use more specific CPU optimizations significantly improves performance (e.g. libraries provided by Intel can use CPU optimizations not present on AMD processors).

On the PHYSnet cluster, a variety of software packages are available as modules. The benchmark in this thesis were made using the following modules:

- `openmpi/4.1.1.gcc10.2-infiniband`: OpenMPI 4.1.0 (implies usage of QUANTUM ESPRESSO provided BLAS/LAPACK)
- `openmpi/4.1.1.gcc10.2-infiniband openblas/0.3.20`: OpenMPI 4.1.0 and OpenBLAS 0.3.20
- `scalapack/2.2.0`: OpenMPI 4.1.0, OpenBLAS 0.3.20 and ScaLAPACK 2.2.0
- `intel/oneAPI-2021.4`: Intel oneAPI 2021.4

QUANTUM ESPRESSO offers an configuration script to automatically find all required libraries. As the default options of the `configure` script work well in the use case of this thesis, all compilations were made using the minimal commands

```
module load <module names>
./configure --with-scalapack=no|yes|intel
```

with the scalapack options yes (when using `scalapack/2.2.0`), intel (when using `intel/oneAPI-2021.4`) and no otherwise.

The output of the configuration script gives information about the libraries it found. In the following output, the Intel Intel oneAPI package was loaded, so BLAS and ScaLAPACK libraries from that package will be used, whereas the included FFT library will be used:

```
The following libraries have been found:
BLAS_LIBS=  -lmkl_intel_lp64  -lmkl_sequential -lmkl_core
LAPACK_LIBS=
SCALAPACK_LIBS=-lmkl_scalapack_lp64 -lmkl_blacs_intelmpi_lp64
FFT_LIBS=
```

II.2.2 Parallelization capabilities offered by Quantum ESPRESSO

QUANTUM ESPRESSO is intended to be used in parallel environments and as such offers possibilities to manage how the work is parallelized. This section introduces the parallelization capabilities of the `PWscf` and `PHonon` modules and explores how they potentially affect the scaling behavior of QUANTUM ESPRESSO .

Fig. ?? shows a possible approach to solving the KS equations. This opens a few possibilities for parallelization of calculations: first of all, the orbitals in the plane wave basis set as well as charges and densities can be distributed among processors. This distribution of data mainly works around memory constraints, as using more processors lowers the memory requirement for every single processor. Going further, QUANTUM ESPRESSO automatically parallelizes all linear algebra operations on real space/reciprocal grid. The price to pay for this parallelization is the need for communication between processors: as an example, fourier transforms always need to collect and distribute contributions from and to the whole reciprocal/real grid in order to transform between them. This kind of parallelization is called *PW (plane wave)* or *R&G (real & reciprocal)* parallelization.

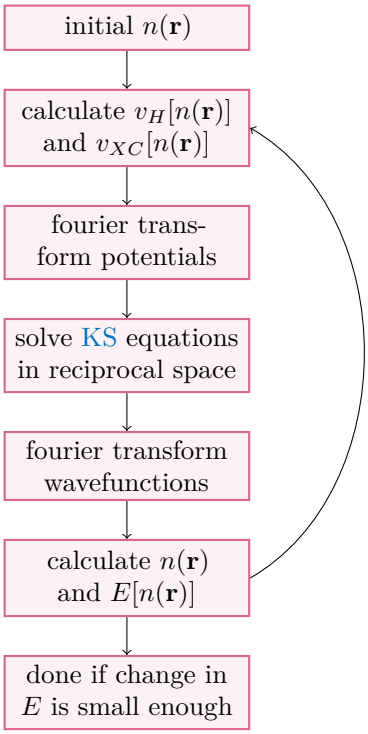


Figure II.2: *Flowchart of an algorithm to solve the KS equations*

As discussed in sec. I.2.3, the density in the plane wave basis set is a sum over different k points, where the calculation for these are independent of each other until calculating the density $n(\mathbf{r})$. QUANTUM ESPRESSO can use this fact and separate the total number of processors into smaller pools, each doing the calculations for a set of k points, so called *k point parallelization*. The CLI parameter `-nk <number of pools>` determines how many pools the total number of processor N is split into, so the resulting number of processors in one pool is $\frac{N}{N_k}$. Within one k point processor pool, the PW parallelization with its heavy communication is automatically applied.

In an level of parallelization independent of that, QUANTUM ESPRESSO can use ScaLAPACK to parallelize (among other things) the iterative orthonormalization of KS states. This parallelization level is called *linear algebra parallelization* and is controlled by the CLI parameter `-nd <number of processors in linear algebra group>`. Importantly, this parameter sets the size for the linear algebra group in every k point processor pool, so the number of processors in the linear algebra group has to be smaller than the number of processors in one pool. Furthermore, the arrays on which the calculations are performed on are

parallelization
ph.x

distributed in a 2D grid among processors, so the number of processors in the linear algebra group has to be a square number.

II.3 Hardware configuration of the PHYSnet cluster

All calculations were run on the `infinix` queue on the PHYSnet compute cluster. As of time of writing, the nodes in this queue are equipped with two Intel Xeon E5-2680 CPUs, as such providing 20 cores per node, 10 per chip.

III Parallelisation of electronic-structure calculations

The `PWscf` (Plane-Wave Self-Consistend Field) package is one of the core modules of QUANTUM ESPRESSO , as many other modules need ground state density and total energy as input. This chapter deals with examining the best ways to run `PWscf` calculations in the `scf` mode.

III.1 First scaling tests

The first step in analysing the scaling of the `PWscf` module is to perform a baseline scaling test without any optimisations appplied. In Fig. III.1 to III.4 two scaling tests on the earlier mentioned benchmarking systems Si and TaS2 are pictured. The tests are run using QUANTUM ESPRESSO 7.0, compiled using the Fortran and C compilers in OpenMPI 4.1.0, without any of the compilation or runtime optimisation parameters mentioned in section II.2 used.

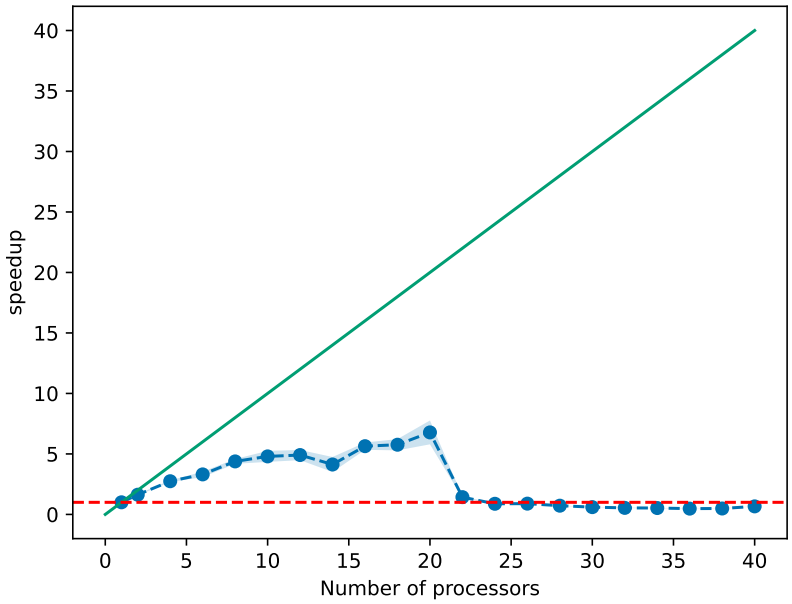


Figure III.1: *Baseline scaling test on the Si benchmarking system* QUANTUM ESPRESSO 7.0, OpenMPI 4.1.0, `nk 1` and `nd 1`

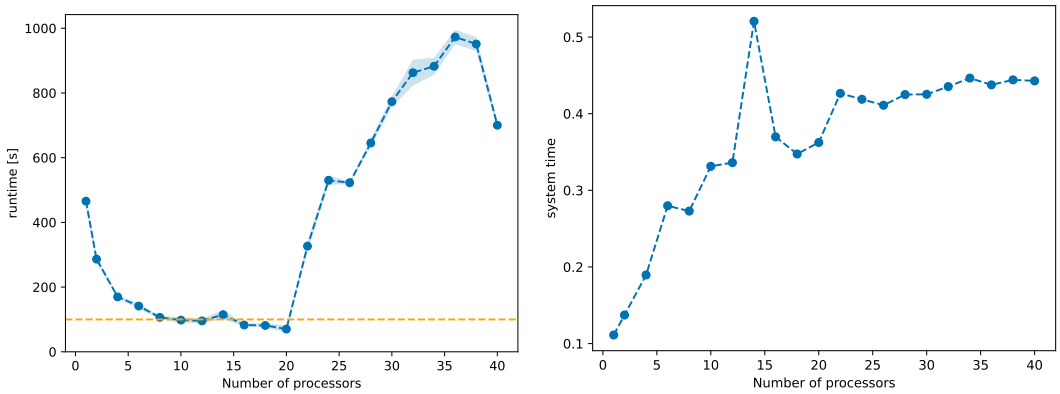


Figure III.2: *Baseline scaling test on the Si benchmarking system* QUANTUM ESPRESSO 7.0, OpenMPI 4.1.0, `nk 1` and `nd 1`

As discussed in sec. II.1.2, three different metrics of scalability can be deduced from the time data given by QUANTUM ESPRESSO :

- runtime: absolute runtime (walltime) of the compute job
- speedup: runtime divided by runtime of the job on a single core
- wait time: percentage of wall time used by system tasks, e.g. writing to disk, etc.

These are pictured in fig. III.1 and III.2 for the silicon benchmarking system.

On a single node, the speedup does scale linearly with the number of processors until around 10 processors, but with a slope of $\frac{1}{2}$ instead of 1 (which would mean ideal scaling). Beyond that number, the slope decreases even more so that a maximal speedup of around 7 is achieved for 20 processors used. One compute node is equipped with 20 cores, so trying to scale the communication intensive calculations beyond that threshold makes the calculations run even slower than on a single core. Interestingly, the wait time plot in III.2 shows that a significant amount (10 % to 40 %) of runtime is taken up by wait time already for less than 20 processors. As discussed in sec. II.1.1, this is a sign of poor parallelization, which can explain the poor scaling seen in fig. III.1.

Pictured in fig. III.3 and III.4 are the same scaling test run for the TaS2 benchmarking system. Here, the speedup is not taken as runtime divided by runtime on a single core, as the memory required is more than what can be accessed by a single core. Instead, an estimate of the single core runtime is made by multiplying the runtime of the job running on 4 cores by 4. This assumes perfect scaling for 1-4 processors, but the relative scaling is accurate, no matter the accuracy of this assumption.

The scaling test on the TaS2 system shows much better scaling. For up to 20 processors, the speedup follows the ideal scaling with a stark decline with more processors. This is also reflected in the wait time in fig. III.4, as it goes from a small constant value for under 20 processors to some kind of dependence of the number of processors, which hints to communication or bottlenecks being a limiting factor here.

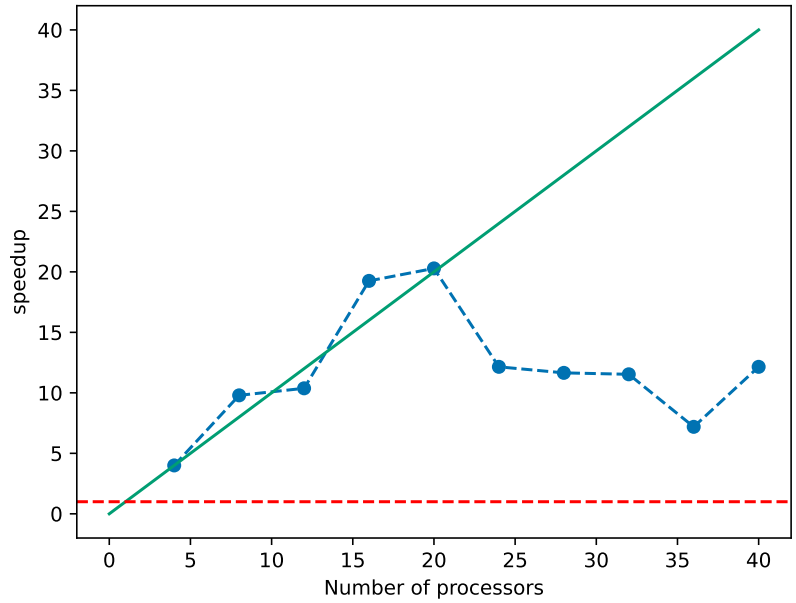


Figure III.3: *Baseline scaling test on the TaS2 benchmarking system* QUANTUM ESPRESSO 7.0, OpenMPI 4.1.0, `nk 1` and `nd 1`

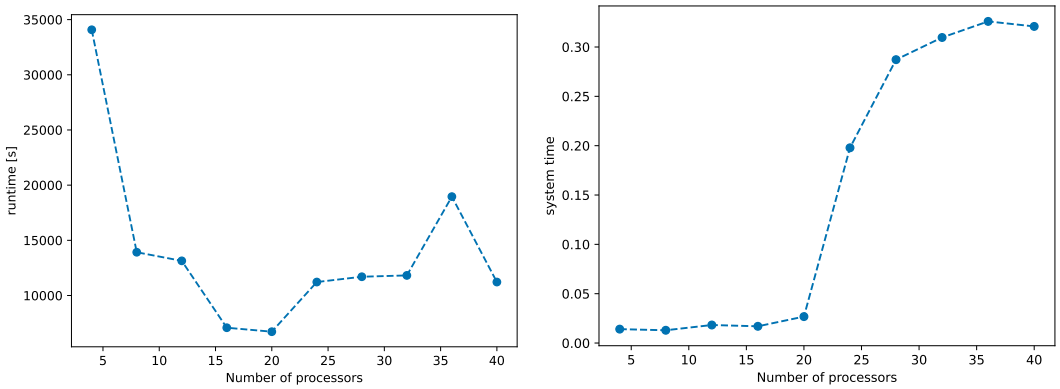


Figure III.4: *Baseline scaling test on the TaS2 benchmarking system* QUANTUM ESPRESSO 7.0, OpenMPI 4.1.0, `nk 1` and `nd 1`

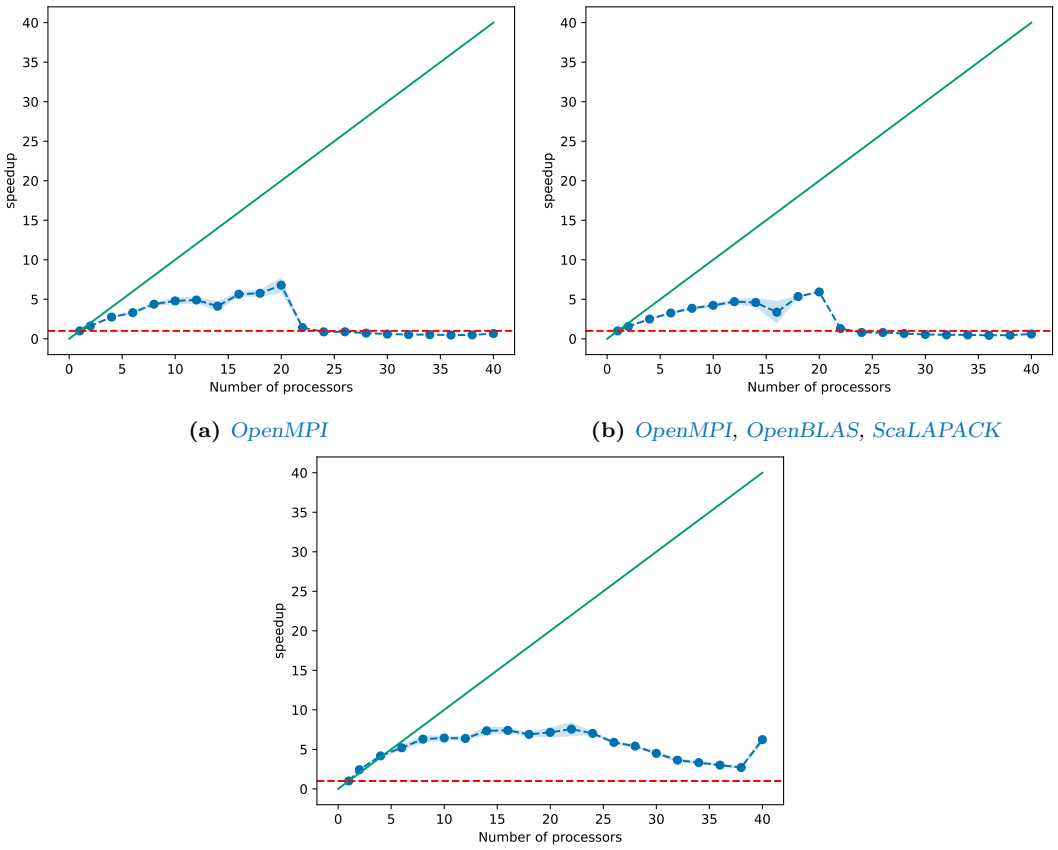
In conclusion, systems with more electrons and by extension bigger matrices and longer iteration times seem to be parallelize better and as such profit more from using more processors than systems with just a few number of electrons.

These scaling tests poses now the question how better scaling over more than one node can be achieved.

III.2 Testing different compilers and mathematical libraries

A first strategy for solving issues with parallelization is trying different compilers and mathematical libraries. As discussed in sec. II.2.1, QUANTUM ESPRESSO can make use of a variety of software packages available on the PHYSnet cluster. The benchmarks in **??** are run with the following software combinations:

- OpenMPI 4.1.0 and QUANTUM ESPRESSO provided BLAS/LAPACK, so the baseline test discussed in sec. III.1
- OpenMPI 4.1.0, OpenBLAS 0.3.20 and ScaLAPACK 2.2.0
- Intel oneAPI 2021.4

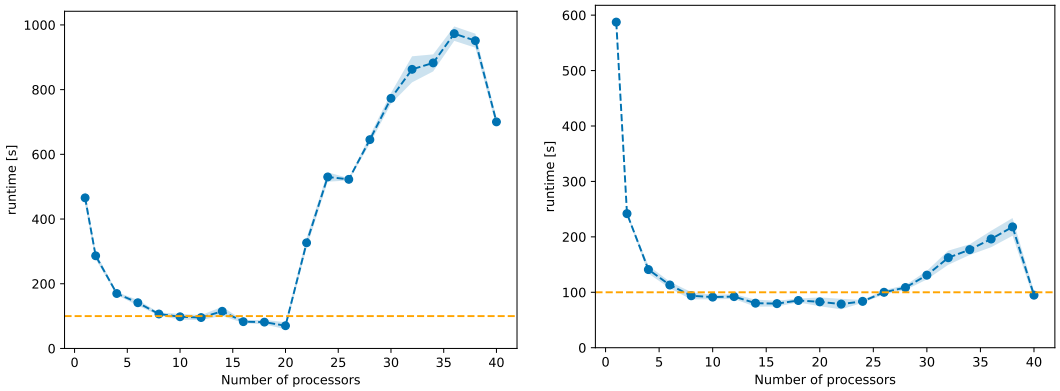


(a) *OpenMPI* **(b)** *OpenMPI*, *OpenBLAS*, *ScaLAPACK*

(c) *Intel oneAPI*

Figure III.5: *Baseline scaling test Si benchmarking system with different combinations of compilers and mathematical libraries*

Fig. **??** shows that just dropping in another BLAS library (OpenBLAS in this case) does not change the scaling behavior, in contrast to using Intels Intel oneAPI packages. Here, optimal scaling behavior is seen for up to 6 processors. It is however important to also look at the total runtime in this context.



(a) *OpenMPI* **(b)** *Intel oneAPI*

Figure III.6: *Baseline scaling test Si benchmarking system with different combinations of compilers and mathematical libraries*

Fig. **??** shows the absolute runtime for both the OpenMPI and Intel oneAPI benchmarks. This explains the difference in scaling seen in the speedup plots: the runtime on a single core is significantly higher for the Intel oneAPI benchmark, so even though the runtime between both benchmarks is about the same starting from around 10 processors there is a difference in speedup. To assess this more quantitatively, tab. III.1 lists the average runtime for some selected number of processors. Importantly, the runtime for the Intel oneAPI benchmark is faster for smaller numbers of processors (except 1), but only 15 % for 2 cores and even smaller differences for more cores, with the OpenMPI calculation being even a little faster for 20 processors.

Table III.1: *CAPTION*

Number of processors	OpenMPI	Intel oneAPI
1	466 s	587 s
2	286 s	242 s
4	170 s	141 s
10	97.9 s	91.3 s
20	70.2 s	82.8 s

The same benchmark with the Intel oneAPI compiled version of Quantum ESPRESSO is shown in fig. III.7. For this system, the speedup follows Amdahl's law, discussed in sec. II.1.1 with a linear growth in speedup up to 32 processors with a saturation and only a small gain in speedup with more processors.

TaS2 intel scaling

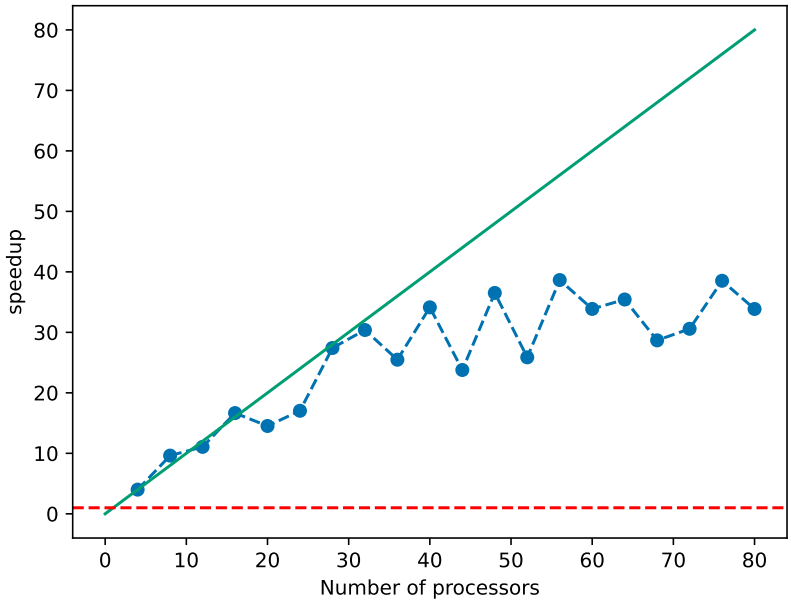


Figure III.7: *CAPTION*

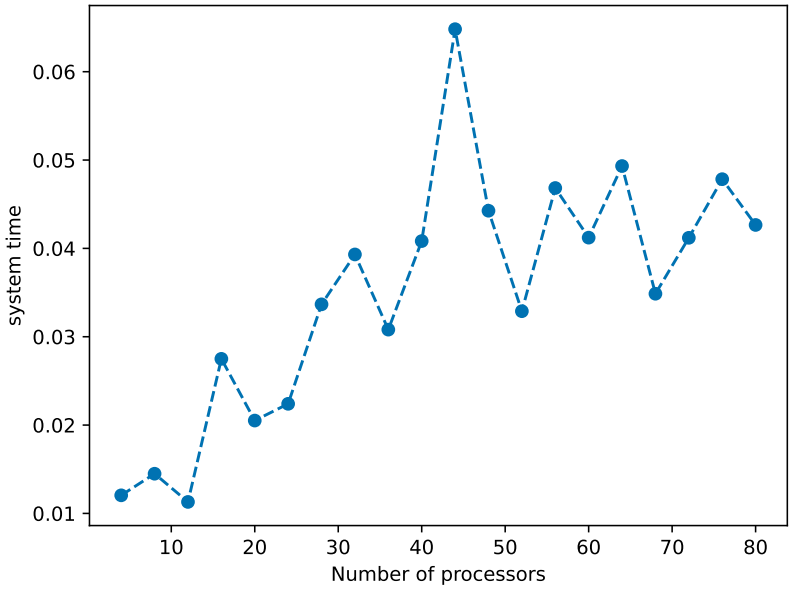


Figure III.8: *CAPTION*

This result not only stands for itself as a statement about scaling on a single node, but also provides a basis for scaling beyond the respective optimal ranges of processors for both systems:

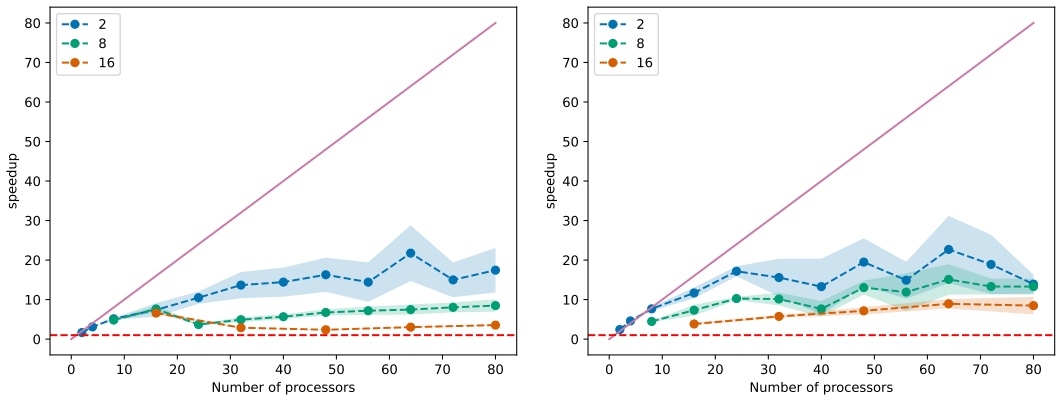
The k point parallelization explained in sec. II.2.2 can distribute the workload in such a way that processor pools of sizes within this range work on individual k points and as such can provide optimal scaling within one pool while also not losing performance because the pools do not need to communicate with each other in the same order of magnitude as the pools have to communicate within themselves.

III.3 Using the parallelization parameters of Quantum ESPRESSO

As detailed in section II.2.2, QUANTUM ESPRESSO offers ways to manage how the workload is distributed among the processors. In `pw.x` the default plane wave parallelization, k-point-parallelization and linear-algebra parallelization are implemented.

III.3.1 k point parallelization

The benchmark pictured in III.9 is set up as follows: for a given number of processors N_p, the parameter N_k splits the N_p processors into N_k processors pools. As the number of processors in one pool has to be a whole number, only certain combinations of N_p and N_k are possible, for example $N_p = 32$ could be split into processor pools of size 2 with $N_k = 16$, size 8 with $N_k = 4$ or size 16 with $N_k = 2$. This leads to choosing the size of the processor pools as a variable, not the parameter `nk`. Fig. III.9 shows the scaling for poolsizes 2, 8 and 16 for QUANTUM ESPRESSO being compiled with OpenMPI/Scalapack and Intel oneAPI. This choice of pool sizes showcases the smallest pool size possibly (namely 2), as well as a bigger pool size with 16, that still gives rise to a few data points over the chosen range of processors.



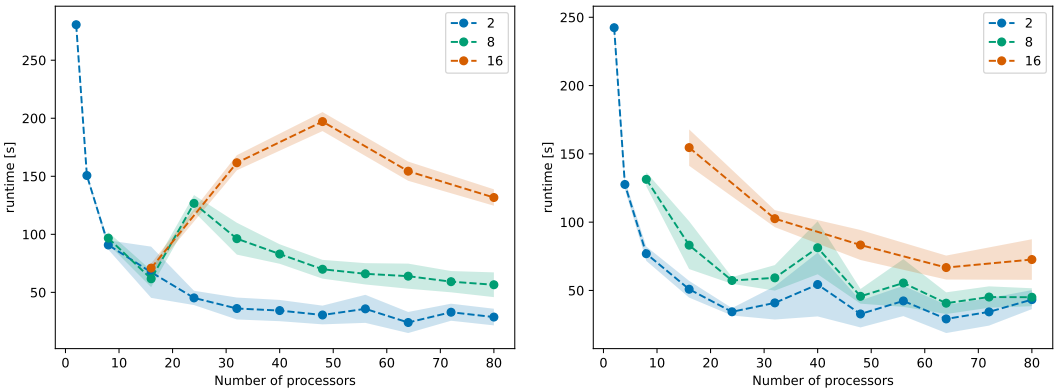
(a) *OpenMPI* **(b)** *Intel oneAPI*

Figure III.9: *Benchmark with k-point parallelization for the Si benchmarking system with 3 different sizes of processor pools*

Fig. III.9 shows that using k parallelization with a pool size of 2 significantly improves the scaling behavior, not only on one node, but especially over more than one node.

more analysis: difference between poolsizes

Another important conclusion to draw out of fig. III.9 is the impact of using Intels compiler instead of OpenMPI, as that factor alone speeds up the calculation by a factor of 2 over the whole range of processors.



(a) *OpenMPI* **(b)** *Intel oneAPI*

Figure III.10: *Benchmark with k-point parallelization for the Si benchmarking system with 3 different sizes of processor pools*

The same scaling test is applied to the TaS2 system in fig. III.11, with a similar list of pool sizes, but over a wider range of processors.

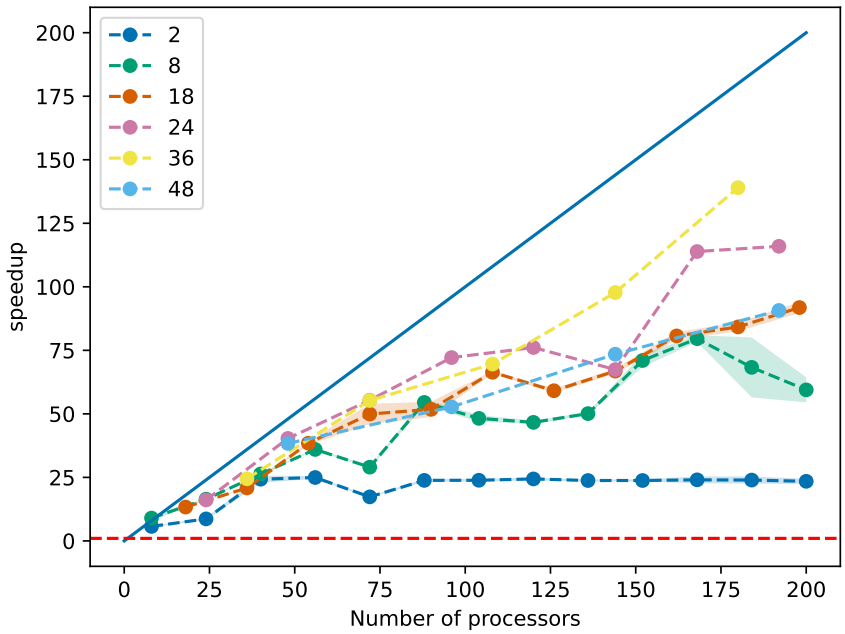


Figure III.11: *Benchmark with k-point parallelization for the TaS2 benchmarking system*

Remarkably, the scaling behavior is swapped in comparison to III.9, as the pool size 2 saturates fast and the bigger pool sizes show way better scaling behavior. Furthermore, there are instances of better than linear scaling, which according to QUANTUM ESPRESSO docs can be attributed to better caching of data.

It can also be instructive to look at the idle time for this benchmark to judge the quality of parallelization.

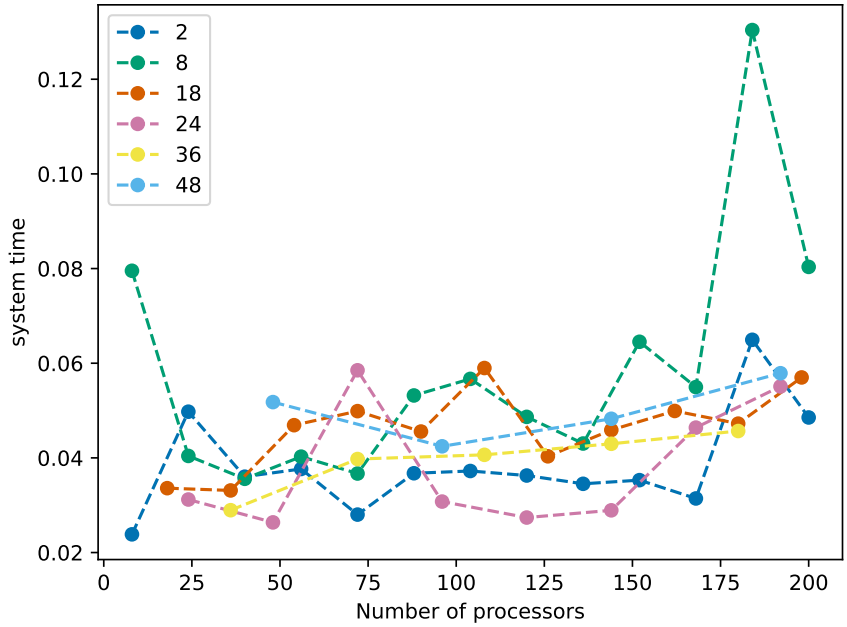


Figure III.12: *Idle time for the k point parallelization benchmark for the TaS2 system*

Fig. III.12 shows a distribution of idle times between about 4 % and 6 % of the whole wall time, without any kind of systemic increase over any range of processors.

more analysis: difference between poolsizes

III.3.2 Linear algebra parallelization

Fig. ?? shows the scaling behavior for different values of the parameter `nd`. Here, nd_auto means that no value for `nd` is specified so QUANTUM ESPRESSO automatically chooses the biggest square number smaller than the number of processors. It is clearly shown that using linear algebra parallelization slows the calculation down significantly for the silicon system.

Interestingly, this again is not reproduced for the more expensive TaS2 benchmarking system. Fig. ?? shows a pretty much consistent times across all values for `nd`.

Those results are already hinted at in the `PWscf` user guide [5]. Here, in the guide for choosing parallelization parameters, using linear algebra parallelization is recommended when the number of KS states is a few hundred or more. The silicon system has 8 electrons and is as such described with 8 Kohn-Sham (KS) states, the TaS2 system has 153 electrons, so QUANTUM ESPRESSO

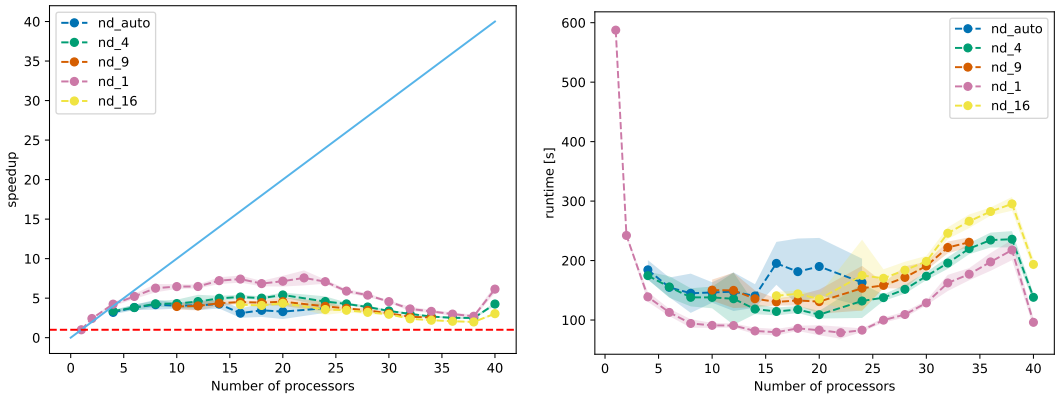


Figure III.13: *Benchmark with linear algebra parallelization for the silicon benchmarking system*

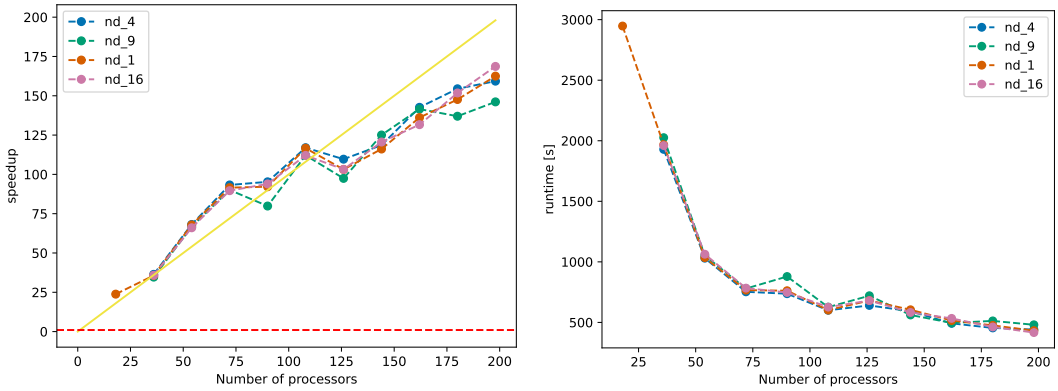


Figure III.14: *Benchmark with linear algebra parallelization for the TaS2 benchmarking system*

uses 92 KS states (in case of metallic materials, the band occupation is smeared around the Fermi energy to avoid level crossings, so more KS states than $\frac{1}{2} *$ (number of electrons) are needed to account for that). Evidently, this number of KS states is on the edge of linear algebra parallelization actually speeding up calculations.

III.4 Comparison with calculations on the HLRN cluster

III.5 Conclusion: Parameters for optimal scaling

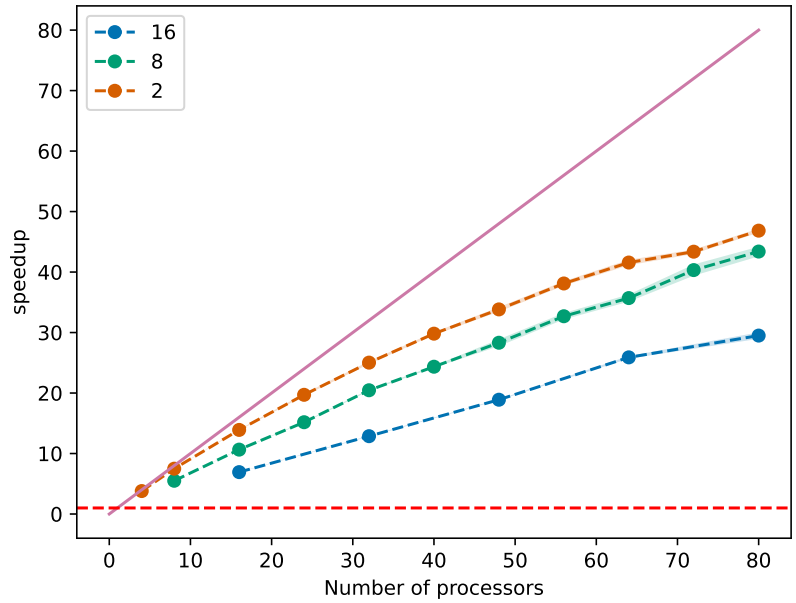


Figure III.15: *CAPTION*

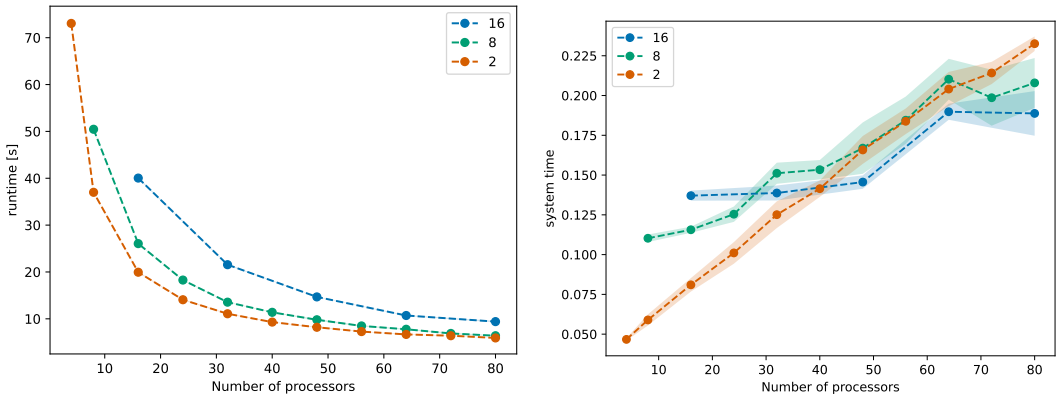


Figure III.16: *CAPTION*

IV Parallelization of DFPT calculations

IV.1 Optimal parallelization parameters for DFPT calculations

IV.1.1 k point parallelization

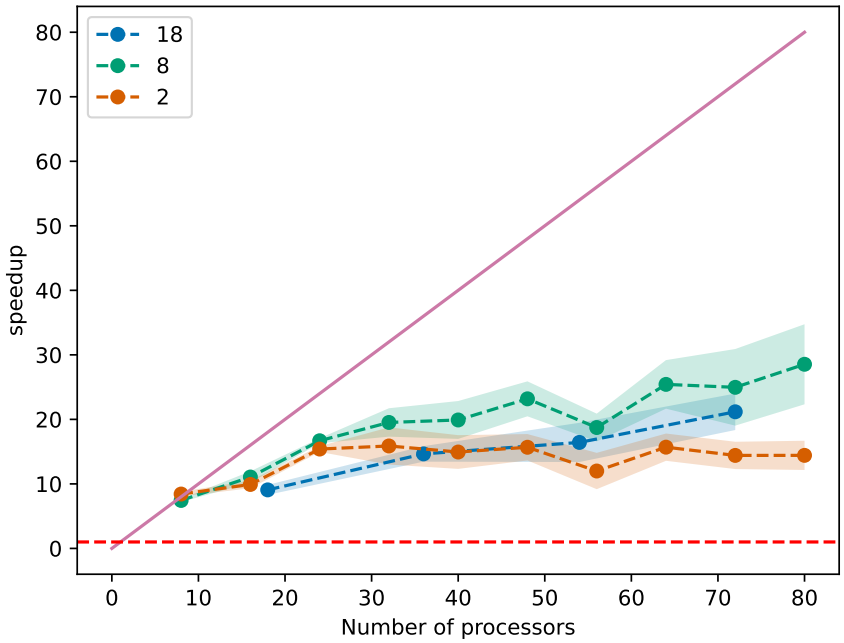


Figure IV.1: *CAPTION*

IV.1.2 Linear algebra parallelization

IV.2 Image parallelization

IV.3 Conclusion: Parameters for optimal scaling

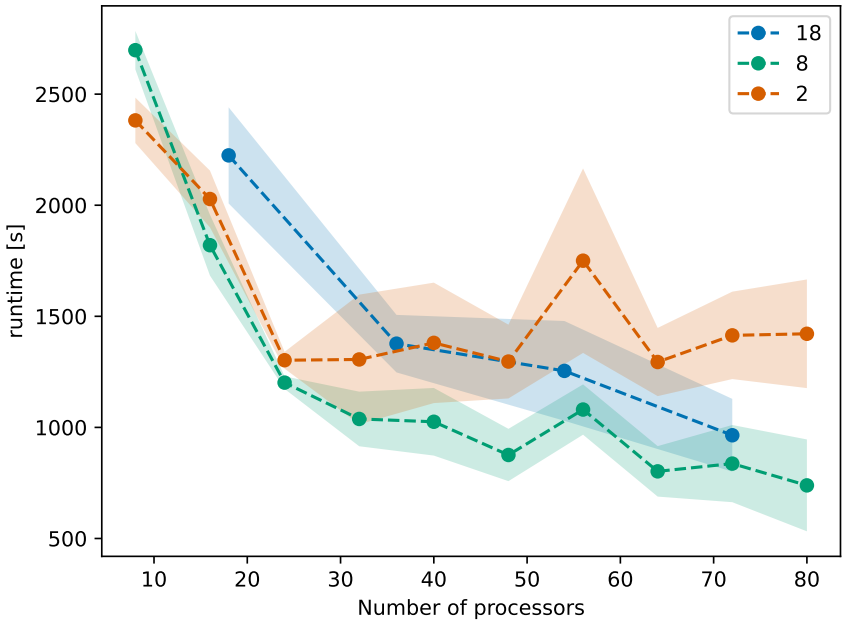


Figure IV.2: *CAPTION*

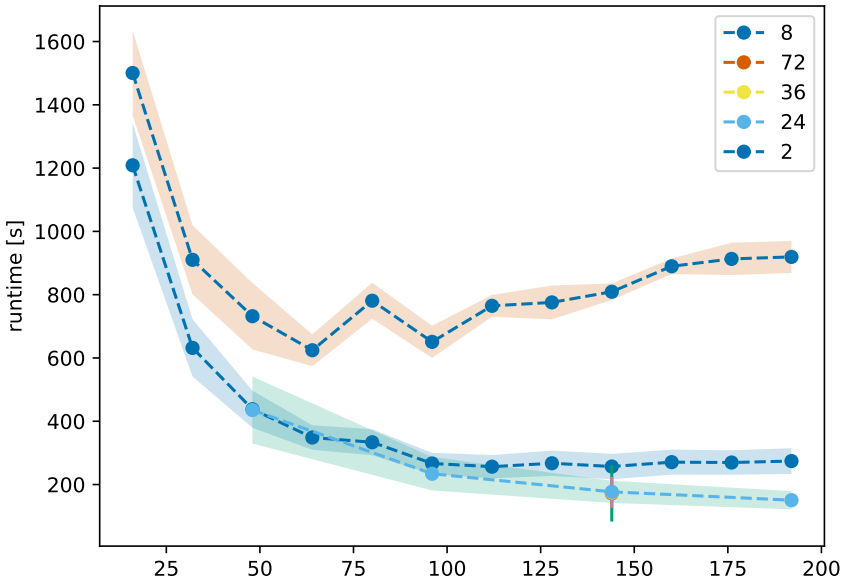


Figure IV.3: *CAPTION*

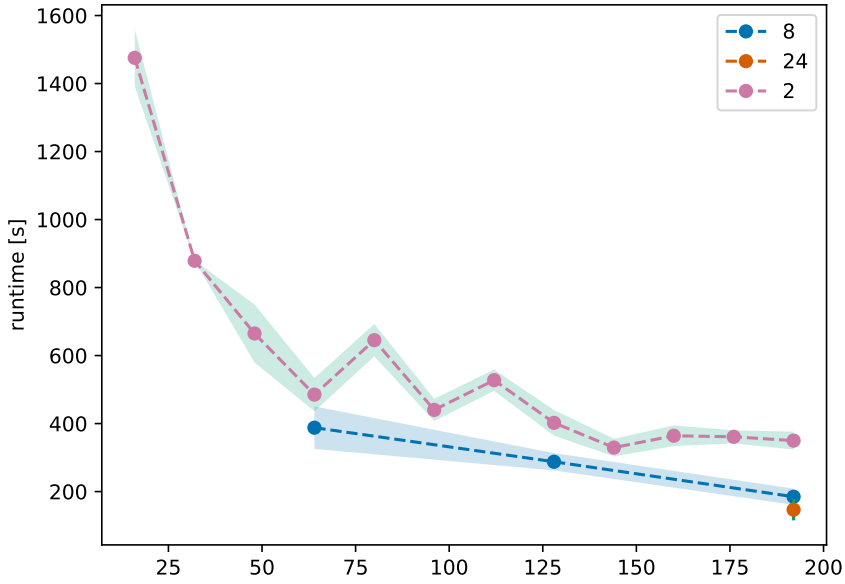


Figure IV.4: *CAPTION*

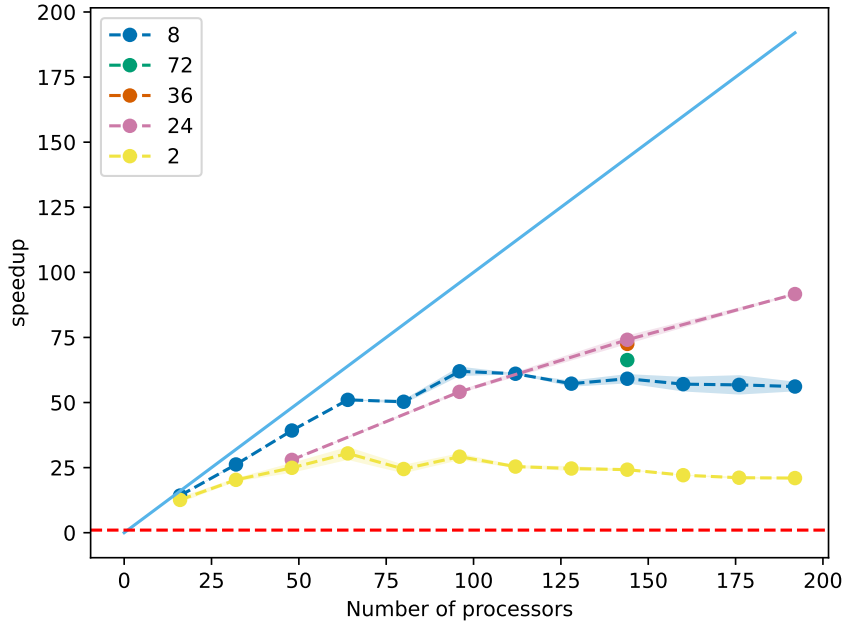


Figure IV.5: *CAPTION*

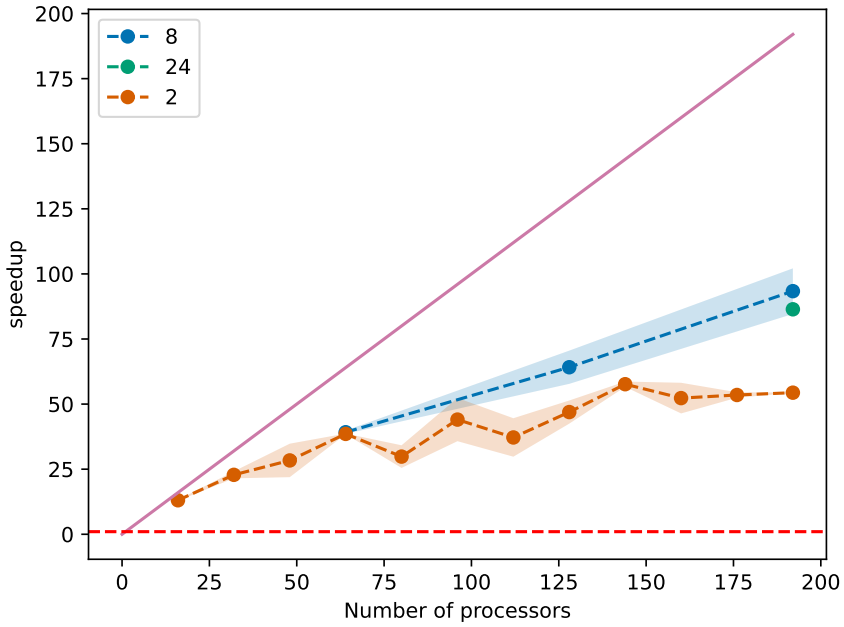


Figure IV.6: *CAPTION*

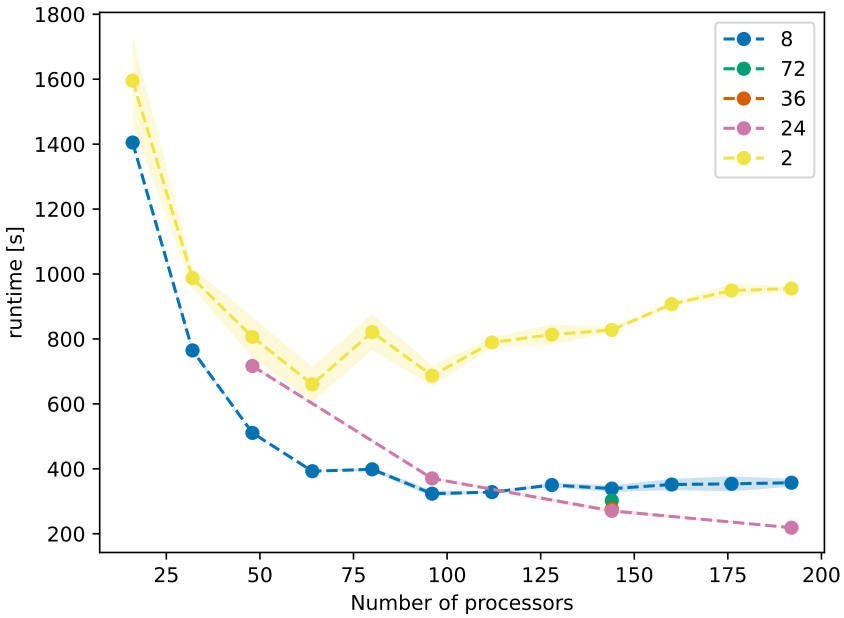


Figure IV.7: *CAPTION*

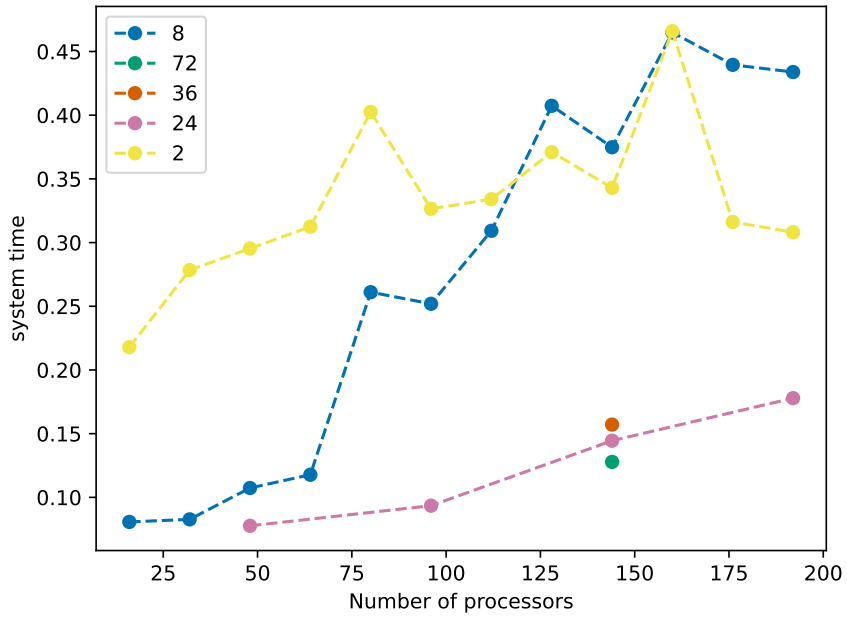


Figure IV.8: *CAPTION*

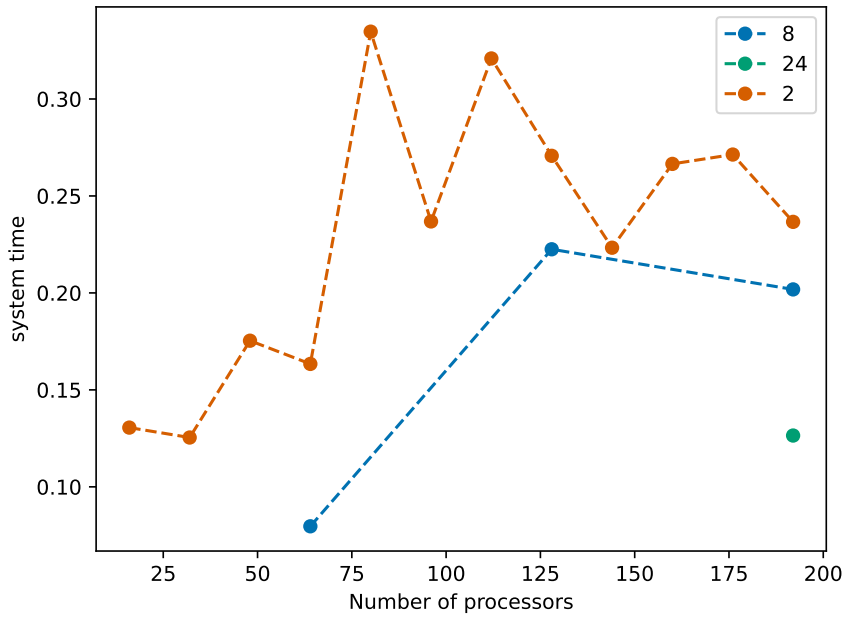


Figure IV.9: *CAPTION*

Bibliography

[1] N. Marzari. "Ab-initio Molecular Dynamics for Metallic Systems". PhD thesis. University of Cambridge, 1996.

[2] P. Hohenberg and W. Kohn. "Inhomogeneous Electron Gas". In: *Phys. Rev.* 136.3 (Nov. 1964). Publisher: American Physical Society, B864–B871. DOI: 10.1103/PhysRev.136.B864.

[3] W. Kohn and L. J. Sham. "Self-Consistent Equations Including Exchange and Correlation Effects". In: *Phys. Rev.* 140.4 (Nov. 1965). Publisher: American Physical Society, A1133–A1138. DOI: 10.1103/PhysRev.140.A1133.

[4] G. Hager and G. Wellein. *Introduction to High Performance Computing for Scientists and Engineers*. 0th ed. CRC Press, July 2, 2010. ISBN: 978-1-4398-1193-1. DOI: 10.1201/EBK1439811924.

[5] *PWscf User's Guide (v. 7.0)*. URL: https://www.quantum-espresso.org/documentation/package-specific-documentation/ (visited on 05/23/2022).

[6] P. Giannozzi et al. "QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials". In: *Journal of Physics: Condensed Matter* 21.39 (Sept. 2009). Publisher: IOP Publishing, p. 395502. DOI: 10.1088/0953-8984/21/39/395502.

[7] P. Giannozzi et al. "Advanced capabilities for materials modelling with Quantum ESPRESSO". In: *Journal of Physics: Condensed Matter* 29.46 (Oct. 2017). Publisher: IOP Publishing, p. 465901. DOI: 10.1088/1361-648x/aa8f79.

[8] *Quantum ESPRESSO User's Guide (v. 7.0)*. URL: https://www.quantum-espresso.org/documentation/ (visited on 05/23/2022).

Listings

List of Figures

List of Tables

Acknowledgement

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Eidesstattliche Erklärung

Ich versichere, dass ich die beigefügte schriftliche Bachelorarbeit/Masterarbeit selbstständig angefertigt und keine anderen als die angegebenen Hilfsmittel benutzt habe. Alle Stellen, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, habe ich in jedem einzelnen Fall unter genauer Angabe der Quelle deutlich als Entlehnung kenntlich gemacht. Dies gilt auch für alle Informationen, die dem Internet oder anderer elektronischer Datensammlungen entnommen wurden. Ich erkläre ferner, dass die von mir angefertigte Bachelorarbeit/Masterarbeit in gleicher oder ähnlicher Fassung noch nicht Bestandteil einer Studien- oder Prüfungsleistung im Rahmen meines Studiums war. Die von mir eingereichte schriftliche Fassung entspricht jener auf dem elektronischen Speichermedium.

Ich bin damit (nicht) einverstanden, dass die Bachelorarbeit/Masterarbeit veröffentlicht wird.

_____

Ort, Datum Unterschrift