THE VIENNA AB-INITIO SIMULATION PACKAGE

Doris Vogtenhuber

UNIVERSITÄT WIEN

Institut für Materialphysik and
Center for Computational Material Science,
Universität Wien, Sensengasse 8/12, A-1090 Wien, Austria

Overview

- Levels of computer-simulations in materials science
- Basics of ab-initio electronic structure methods
- Pseudopotentials: The PAW and US-PP database
- Controlling the precision of VASP
- electronic optimization
- ionic optimization
- Sampling the Brillouin Zone
- performance

Levels of materials modelling

- Ab-initio techniques
  - Hartree-Fock and post-HF techniques - Quantum chemistry
  - Density functional techniques - Materials science
- Tight-binding techniques
- Force-field simulations
  - Molecular dynamics
  - Monte Carlo

Born-Oppenheimer approximation I

Hamiltonian of the coupled electron-ion system:

N ions, coordinates $\vec{R}_1,\dots,\vec{R}_N \equiv \vec{R}$, momenta $\vec{P}_1,\dots,\vec{P}_N \equiv \vec{P}$, charges $Z_1,\dots,Z_N$, masses $M_I,\dots,M_n$

N_e electrons, coordinates $\vec{r}_1,\dots,\vec{r}_N \equiv \vec{r}$, momenta $\vec{p}_1,\dots,\vec{p}_N \equiv \vec{p}$, mass m

$$\begin{aligned} H &= \sum_{I=1}^{N} \frac{\vec{P}_I^2}{2M_I} + \sum_{i=1}^{N_e} \frac{\vec{p}_i^2}{2m} + \sum_{i>j} \frac{e^2}{|\vec{r}_i-\vec{r}_j|} + \sum_{I>J} \frac{Z_I Z_J e^2}{|\vec{R}_i-\vec{R}_J|} - \sum_{i,I} \frac{Z_I e^2}{|\vec{R}_I-\vec{r}_i|} \\ &= T_N + T_e + V_{ee}(\vec{r}) + V_{NN}(\vec{R}) + V_{Ne}(\vec{r},\vec{R}) \end{aligned} \tag{1}$$

Schrödinger equation

$$[T_N + T_e + V_{ee}(\vec{r}) + V_{NN}(\vec{R}) + V_{Ne}(\vec{r},\vec{R})]\Phi(x,\vec{R}) = E\Phi(x,\vec{R}) \tag{2}$$

$x \equiv (\vec{r},s)$ full set of electronic positions and spin variables

Born-Oppenheimer approximation II

Difference in the time-scales of nuclear and electronic motions $\longrightarrow$ quasi-separable ansatz
decoupled adiabatic Schrödinger equations of electrons and nuclei: $\longrightarrow$ Electronic eigenvalue $\varepsilon_n(\vec{R})$ depends *parametrically* on the ionic positions $\vec{R}$
Adiabatic approximation: Ions move on the potential-energy surface of the electronic ground state only.
Neglect quantum effects in ionic dynamics $\longrightarrow$ replace time-dependent ionic Schrödinger equation by classical Newtonian equation of motion

$$\frac{\partial^2 \vec{P}_I(t)}{\partial t^2} = -\nabla_I E_0(\vec{R}) \tag{3}$$

Force $-\nabla_I E_0(\vec{R})$ contains contributions from the direct ion-ion interaction and a term from the gradient of the electronic total energy

Hellmann-Feynman theorem

$$\begin{aligned}
\nabla_I \varepsilon_0(\vec{R}) &= \frac{\partial}{\partial \vec{R}_I} \langle \Psi_0 \mid H_e(\vec{R}) \mid \Psi_0 \rangle \\
&= \langle \nabla_I \Psi_0 \mid H_e(\vec{R}) \mid \Psi_0 \rangle \\
&\quad + \langle \Psi_0 \mid \nabla_I H_e(\vec{R}) \mid \Psi_0 \rangle \\
&\quad + \langle \Psi_0 \mid H_e(\vec{R}) \mid \nabla_I \Psi_0 \rangle \\
&= \langle \Psi_0(\vec{R}) \mid \nabla_I H_e(\vec{R}) \mid \Psi_0(\vec{R}) \rangle
\end{aligned} \tag{4}$$

First and third terms in the derivative vanish due to variational property of the ground-state $\rightarrow$ Forces acting on the ions are given by the expectation value of the gradient of the electronic Hamiltonian in the ground-state

Ab-initio electronic structure - Density-functional theory I

Hohenberg-Kohn-Sham theorem:

- The ground-state energy of a many-body system is a unique functional of the particle density, $E_0 = E[n(\vec{r})]$.

- The functional $E[n(\vec{r})]$ has its minimum relative to variations $\delta n(\vec{r})$ of the particle density at the equilibrium density $n_0(\vec{r})$,

$$\begin{gathered}
E = E[n_0(\vec{r})] = min\,\{E[n(\vec{r})]\} \\
\frac{\delta E[n(\vec{r})]}{\delta n(\vec{r})}\Big|_{n(\vec{r})=n_0(\vec{r})} = 0
\end{gathered} \tag{5}$$

Density-functional theory II

Total-energy functional

$$E[n] = T[n] + E^H[n] + E^{xc}[n] + \int V(\vec{r})n(\vec{r})d^3r \tag{6}$$

$T[n]\ldots$ kinetic energy,
$E^H[n]\ldots$ Hartree energy (electron-electron repulsion),
$E^{xc}[n]\ldots$ exchange and correlation energies,
$V(\vec{r})$ external potential
- the exact form of $T[n]$ and E_{xc} is unknown !
Local density approximation - "density only":
- Approximate the functionals $T[n]$ and $E_{xc}[n]$ by the corresponding energies of a homogeneous electron gas of the same local density
$\longrightarrow$ Thomas-Fermi theory

Density-functional theory III

Local density approximation - Kohn-Sham theory:

- Parametrize the particle density in terms of a set of one-electron orbitals representing a non-interacting reference system

$$n(\vec{r}) = \sum_i |\phi_i(\vec{r})|^2 \tag{7}$$

- Calculate non-interacting kinetic energy in terms of the $\phi_i(\vec{r})$'s,

$$T[n] = \sum_i \int \phi_i^*(\vec{r}) \left(-\frac{\hbar^2}{2m} \nabla^2 \right) \phi_i(\vec{r}) d^3 r \tag{8}$$

- Determine the optimal one-electron orbitals using the variational condition

$$\frac{\delta E[(n(\vec{r})]}{\delta \phi_i(\vec{r})} = 0 \tag{9}$$

$\longrightarrow$ Kohn-Sham equations

Density-functional theory IV

$$E[n] = T[n] + E^H[n] + E_{xc}[n] + \int V(\vec{r}) n(\vec{r}) d^3 r \tag{10}$$

with the exchange-correlation energy

$$E^{xc}[n(\vec{r})] = \int n(\vec{r}) \varepsilon_{xc}[n(\vec{r})] d^3 r, \tag{11}$$

where $\varepsilon_{xc}[n(\vec{r})]$ is the exchange-correlation energy of a homogeneous electron gas with the local density $n(\vec{r}) \rightarrow$ Kohn-Sham equations:

$$\left\{ -\frac{\hbar^2}{2m} \nabla^2 + \underbrace{V(\vec{r}) + e^2 \int \frac{n(\vec{r}')}{|\vec{r} - \vec{r}'|} d^3 r + \mu_{xc}[n(\vec{r})]}_{V_{eff}(\vec{r})} \right\} \phi_i(\vec{r}) = \varepsilon_i \phi_i(\vec{r}) \tag{12}$$

with the exchange-correlation potential

$$\mu_{xc}[n(\vec{r})] = \frac{\delta E_{xc}[n(\vec{r})]}{\delta n(\vec{r})} = \frac{\delta\{n(\vec{r}) \varepsilon_{xc}[n(\vec{r})]\}}{\delta n(\vec{r})} \tag{13}$$

Solving the Kohn-Sham equations

Choice of a basis set

- Plane waves and related basis functions
  - Plane waves
  - (Linearized) augmented plane waves - (L)APW's
  - (Linearized) muffin-tin orbitals - (L)MTO's
  - Projector augmented waves -PAW's
- Localized orbitals
  - Atomic orbitals - LCAO's
  - Gaussian orbitals

Plane waves (PWs) basis sets

Plane waves (PW's)

- Natural choice for system with periodic boundary conditions
- It is easy to pass from real- to reciprocal space representation (and vice versa) by FFT
- No Pulay correction to forces on atoms
- Basis set convergence easy to control
- Convergence slow $\longrightarrow$
  - Electron-ion interaction must be represented by pseudopotentials or projector-augmented wave (PAW) potentials
  - Use LAPW's or mixed basis sets

Pseudopotentials I

- Slow convergence of PW expansion caused by the necessity to reproduce nodal character of valence orbitals
- Nodes are the consequence of the orthogonality to the tightly-bound core-orbitals $\longrightarrow$
- Eliminate the tightly-bound core states and the strong potential binding these states:
  - Use "frozen-core" approximation
  - Project Kohn-Sham equations onto sub-space orthogonal to core-states $\longrightarrow$ orthogonalized plane waves ...., or
  - Replace strong electron-ion potential by a weak pseudopotential which has the same scattering properties as the all-electron potential beyond a given cut-off radius

Pseudopotentials II

Scattering approach to pseudopotentials

- Perform all-electron calculation for atom or ion at a reference energy ε
- Define a cut-off radius r_c well outside the node of the highest core-state
- Construct a pseudo valence-orbital $\tilde{\phi}_l$ that is identical to the all-electron orbital ϕ_l for $r \geq r_c$, but nodeless for $r \leq r_c$ and continuous and continuously differentiable at r_c
- The scattering phase-shifts for electrons agree (modulo 2π) if the logarithmic derivatives of ϕ_l and $\tilde{\phi}_l$ agree on the surface of the cut-off sphere:

$$\frac{\partial \, log\phi_l(r,\varepsilon)}{\partial r} = \frac{\partial \, log\tilde{\phi}_l(r,\varepsilon)}{\partial r}, \quad \text{at } r = r_c \tag{14}$$

Pseudopotentials III

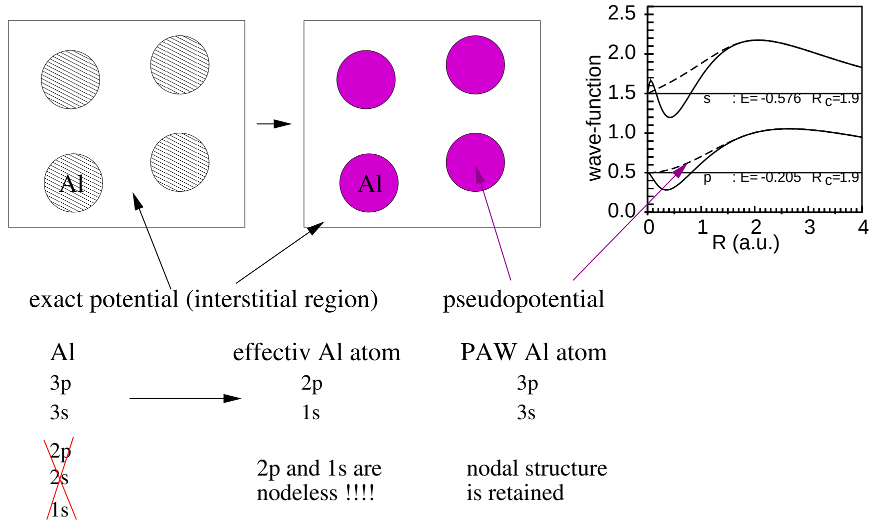
Modern pseudopotentials

- Norm-conserving pseudopotentials (NC-PP)

  Norm-conservation: charge within cut-off sphere fixed

  High cut-off energies for first-row and transition elements
- Ultrasoft pseudopotentials - (US-PP)
  - Norm-conservation relaxed - more freedom for pseudizing $2p$ and $3d$ states
  - Add augmentation charges inside the cut-off sphere to correct charge
  - Multiple reference energies - improved transferability
  - Lower cut-off energies

Pseudopotentials IV

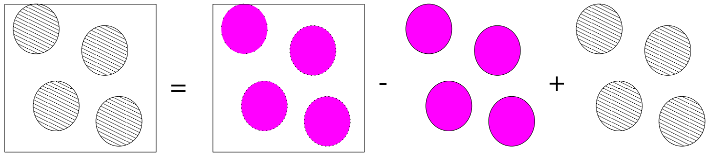


Pseudopotentials V - Projector augmented wave (PAW) method

P.E. Blöchl, Phys. Rev. **B50**, 17953 (1994); G. Kresse, and J. Joubert, Phys. Rev. B **59**, 1758 (1999).

- Pseudization as for ultrasoft potentials
- Reconstruction of exact wavefunction in the core region $\rightarrow$ wave function (and n, T_n, E_{xc}), are decomposed into three terms:

$$|\Phi_n\rangle = |\tilde{\Phi}_n\rangle - \sum_{\text{atoms}} |\tilde{\phi}_{lm\varepsilon}\rangle c_{lm\varepsilon} + \sum_{\text{atoms}} |\phi_{lm\varepsilon}\rangle c_{lm\varepsilon}$$



AE pseudo + compens. pseudo+comp. onsite AE-onsite Pseudo-WF

represented on FFT-grid, on-site terms on atom-centred radial grids

no interaction between different spheres and plane waves $\Rightarrow$ efficient

US-PP or PAW potentials

the general rule is to use PAW potentials wherever possible

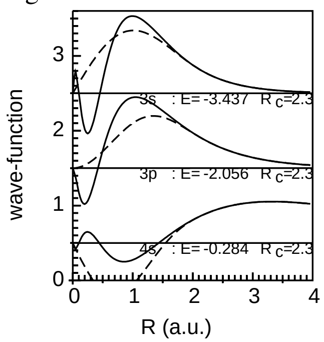
- less parameters involved in the construction of PAW potentials
  - improved accuracy for:
    - magnetic materials
    - alkali and alkali earth elements, early 3d elements to left of periodic table
    - lathanides and actinides
  - generation of datasets is fairly simple (certainly easier than for US-PP)
  - general construction scheme is similar for US-PP and PAW potentials
- most of the PAW potentials were generated 5 years after the US-PP
  $\Rightarrow$ a different and more mature philosophy

Alkali and alkali earth metals

- an accurate treatment of these elements in ionic compounds is very important: oxides e.g. perovskites

  strongly ionised, and small core radii around 1 Å are desirable

- e.g. Ca: one would like to treat 3s, 3p, 4s states as valence states

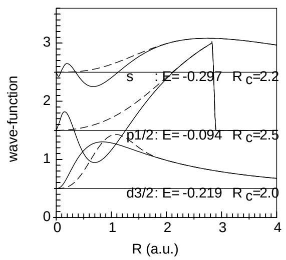


it is very difficult to represent 3s and 4s states equally well in any pseudopotential description (mutual orthogonality)

in the PAW method this is no problem, and the energy cutoffs remain modest (for Ca 200 eV)

Pseudopotentials tend to fail in spin-polarised calculations

- non linear core corrections were included in the US-PP's !
- pseudo-wavefunction for norm-conserving pseudopotentials



— all electron

– – pseudo

the peak in the d-wavefunction is shifted outward to make the PP softer

- similar compromises are made in our US-pseudopotentials (augmentation charge)
- as a result, the valence-core overlap is artificially reduced and the spin enhancement factor $\xi(r) = \frac{m(r)}{n_{\text{valence}}(r)+n_{\text{core}}(r)}$ is overestimated

PAW versus US-PP potentials

- the PAW potentials are generally of similar hardness across the periodic table
- most of the PAW potentials were optimised to work at a cutoff of 250-300 eV
- PAW potentials are usually slightly harder than US-PP
- for the US-PP the core radii were chose according to the covalent radius (periodic table)
- the US-PP become progressively softer when you move down in the periodic table
- for compounds, where often species with very different covalent radii are mixed, the PAW potentials are clearly superior
- for one-component systems the US-PP might be slightly faster (at the price of a somewhat reduced precision)

The PAW potentials

three different flavours, one LDA (CA) and two GGA's (PW91) and new PBE

- recommended: use either the LDA or the PBE potentials
- all LDA potentials apply the exchange correlation form according to Ceperly and Alder as parameterized by Perdew and Zunger
- the PBE implementation adhears strictly to the PBE prescription,
- in VASP the PW91 implementation is sloppy: for the LDA part, the parametrisation of Perdew and Zunger is used, instead of Perdews Pade approximation "VASP bug": different methods to interpolate between non spin polarised and fully spin polarized local density functionals exist
  for the PW91 functional the interpolations should be done according to Vosko Wilk and Nusair

Standard PAW potentials and Energy Cutoffs

B_h	700	C_h	700	N_h	700	O_h	700	F_h	700
B	318	C	400	N	400	O	400	F	400
B_s	250	C_s	273	N_s	250	O_s	250	F_s	250
Al	240	Si	245	P	270	S	280	Cl	280
Al_h	295	Si_h	380	P_h	390	S_h	402	Cl_h	409
Ga	134	Ge	173	As	208	Se	211	Br	216
Ga_d	282	Ge_d	287						
Ga_h	404	Ge_h	410						
In	95	Sn	103	Sb	172	Te	174	I	175
In_d	239	Sn_d	241						
Tl	90	Pb	98	Bi	105				
Tl_d	237	Pb_d	237	Bi_d	242				

Standard PAW potentials, guideline to the table

- the best compromises are marked red
- potentials X_d treat the semi core d electrons as valence
- potentials marked with X_h use very small core radii, and are only required under extreme circumstances
  - strong compression
  - very short bonds
- soft potentials X_s are for less critical calculations

1st row pseudopotentials

B_h	700	C_h	700	N_h	700	O_h	700	F_h	700
B	318	C	400	N	400	O	400	F	400
B_s	250	C_s	273	N_s	250	O_s	250	F_s	250

- B–F are the standard PAW potentials
  even dimers are described very reliably (CO, N_2, -..., 1% error)
- B_h–F_h are only required for exceptional "reference" calculations
  for instance to establish DFT reference results
- the soft potentials (B_s–F_s) are sufficient, when short bonds do not occur

  if there are no bonds between 1st row elements, you can use them

  O_s is sufficiently accurate for most oxides (Al_2O_3, V_xO_y, Fe_xO_y, but not SiO_2)

  C_s is often sufficiently reliable for organic molecules, if some accuracy tradeoffs are allowed (good enough for single and double bonds)

Remaining rows

- potentials with X_d treat the semi core d electrons as valence
  (3d for Ga and Ge, 4d for In and Sn, etc.)
  - for high accuracy, it is generally advisable to treat the semi-core states as valence (hence the corresponding potentials are marked red)
  - in many cases, even with the frozen semi-core d-states good results are obtained

    these potentials often reduce the number of valence electrons (NBANDS) drastically
  - please make tests (it depends on how accurate results you need)
- the hard potentials for Al_h–Cl_h, Ga_h, Ge_h can be used for added accuracy in oxides, when combined with the standard O potential

  BUT: G.K. found some pecularities and problems with these potentials Al_h still does not work, new Si_h is obviously reliable

Standard PAW potentials for "simple" metals

H	250		
H_h	700		
Li	140	Be	300
Li_sv	271	Be_sv	308
Na	81	Mg	210
Na_pv	300	Mg_pv	265
Na_sv	700		
K_pv	150	Ca_pv	150
K_sv	259	Ca_sv	290
Rb_pv	121	Sr_sv	226
Rb_sv	220		
Cs_sv	220	Ba_sv	187

Simple metal PAW potentials, guideline to the table

- Contrary to the common believe,

  these elements are exceedingly difficult to pseudize
  - in particular in combination with strongly electronegative elements (F)
    errors can be larger then usual
  - the present versions are very precise and should offer a highly reliable description (phonon calculations for alkali-halides by Martijn Marsman)
- for X_pv pseudopotentials the semi core p states are treated as valence (2p in Na and Mg, 3p in K and Ca etc.)

  for X_sv pseudopotentials, the semi core s states are treated as valence (1s in Li and Be, 2s in Na etc.)
- at least relaxing the semi core p states is important, since in strongly ionic environments these elements loose all their "valence" electrons, and the semi-core states relax substantially

Transition metal pseudopotentials

		Ti	178	V	192	Cr	227	Mn	269
Sc_sv	222	Ti_pv	222	V_pv	263	Cr_pv	265	Mn_pv	269
						Mo	224	Tc	228
Y_sv	211	Zr_sv	229	Nb_pv	207	Mo_pv	224	Tc_pv	228
		Hf	220	Ta	223	W	223	Re	226
		Hf_pv	220	Ta_pv	223	W_pv	223	Re_pv	226
Fe	267	Co	267	Ni	269	Cu	273	Zn	276
Fe_pv	293			Ni_pv	367	Cu_pv	368		
Ru	213	Rh	228	Pd	250	Ag	249	Cd	274
Ru_pv	230	Rh_pv	271	Pd_pv	350				
Os	228	Ir	210	Pt	230	Au	229	Hg	233
Os_pv	228								

Transition metal PAW potentials, guideline to the table

- for X_pv pseudopotentials, the semi core p states are treated as valence
  for X_sv pseudopotentials, the semi core s states are treated as valence
- X_pv potentials are required for early transition metals, but one can freeze the semi-core p states for late transition metals (in particular noble metals)
- when to switch from X_pv potentials to the X potentials depends on the required accuracy and the row
  - for the 3d elements, even the Ti, V and Cr potentials give reasonable results
  - 4d elements are most problematic, the X_pv potentials should be used up to Tc_pv
  - 5d elements: 5p states are rather strongly localised (below 3 Ry), since the 4f shell becomes filled
    one can use the standard potentials starting from Hf, but it is recommended to perform test calculations

Actenides and Lathanides

- it is no trouble to generate PAW potentials for these elements, and we have already a rather extensive database

  is standard DFT applicable to these strongly correlated elements ?
  - yes: if the f electrons are itinerant (band like)
  - no: if the f electrons are strongly localised (strong correlation effects, SIC)

  similar problems do occur in some transition metal oxides (NiO, V_2O_3, V_2O_4, Fe_2O_3, FeO etc.)

Pseudopotential that treat the *f*-electrons as valence

Ce	300	Pr	252	Nd	253	Pm	258	Sm	255	Eu	249	Gd	256
								Tm	257	Yb	291	Lu	255
La	219	Ac	169	Th	247	Pa	252	U	252	Np	254	Pu	254
La_s	136	Ac_s	119	Th_s	169	Pa_s	193	U_s	209	Np_s	210	Pu_s	211

- in all cases, the lower semi-core p shell is treated as valence
  for the standard potentials, even the semi-core s states are treated as valence (count the number of electrons, have a look into V_RHFIN)
  the standard versions are recommended (oxides and sulfides)
  except for inter-metallic compounds, where the soft versions might be sufficiently accurate
- the actinides and La PAW have been tested, but the other lathanide PAW potentials are largely untested

f-electrons in the core

standard model for localised f-electrons

the f-electrons are not treated explicitly, but are kept frozen in the core

Ce_3 181	Pr_3 181	Nd_3 182	Pm_3 183	Sm_3 184 Sm_2 183	Eu_2 99	Gd_3 154
Tb_3 155	Dy_3 155	Ho_3 154	Er_3 155 Er_2 119	Tm_3 154	Yb_2 112	Lu_3 155

- X_3 implies that the potential has been created for valency 3

  X_2 implies that the potential has been created for valency 2

- the formal valency in the POTCAR file is larger, since the semi-core p states are also included when the number of valence electrons is determined

  in fact, in all cases, the lower semi-core p shell is treated as valence

US-PP

they come in two different flavours, LDA and GGA

- generally similar naming conventions as for PAW potentials
- generally less accurate, in particular 1st row and 2row standard US-PP are less precise, since they have been generated with slightly larger core radii
- no US-PP for lathanides and actenides are available
- the “simple metal” US-PP’s have rather bad quality

Controlling the precision in VASP

- the plane wave energy cutoff
- technical errors
- the critical parameters `ENCUT, LREAL`
- the “super-flag” `PREC`

Energy cutoff

- controls the completeness of the basis set

  at each k-point only the plane waves that fulfil

$$\frac{\hbar^2}{2m_e}|\mathbf{G}+\mathbf{k}|^2 < E_{\text{cutoff}}$$

  are included

  different number of plane waves at each k-point

- E_{cutoff} is controlled by `ENCUT` in the INCAR file

  the number of plane wave for each k-point is written to the OUTCAR file:

```
k-point  1 :  0.25000.25000.2500  plane waves:   1546
k-point  2 :  -.25000.25000.2500  plane waves:   1557
```

- defaults for `ENCUT` are supplied in the pseudopotential files (POTCAR)

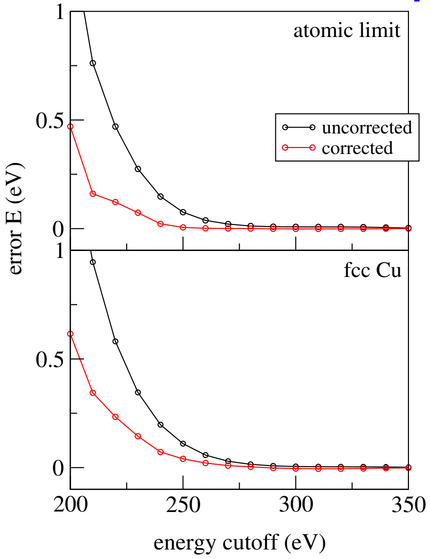
  usually the maximum `ENMAX` of all atoms is chosen as energy cutoff

Convergence correction

- VASP applies an automatic convergence correction based on the kinetic energy of wave-functions in the atomic limit

```
energy of atom 1      EATOM=-1393.0707
kinetic energy error for atom= 0.0229
```

- works well in the atomic limit, and for free electron metals
- corrects for 80 % of the total error
- for d-elements and bulk calculations, corrections are only partial
- correction $\Rightarrow$ the energy may increase, when the cutoff is increased



Can one rely on the default cutoff ?

it depends

- ENCUT is a very reasonable compromise between accuracy and speed
- you can rely on ENCUT, as long as the
  cell-shape and the volume remain unchanged
  – frozen phonon calculations
  – surface and slab calculations
  – adsorption of molecules on surfaces
- otherwise you might need to be rather careful

  the basis set changes discontinuously when the cell-shape is changed, since new plane waves are included when they satisfy the cutoff criterion

$$\frac{\hbar^2}{2m_e}|\mathbf{G}+\mathbf{k}|^2 < E_{\text{cutoff}}$$

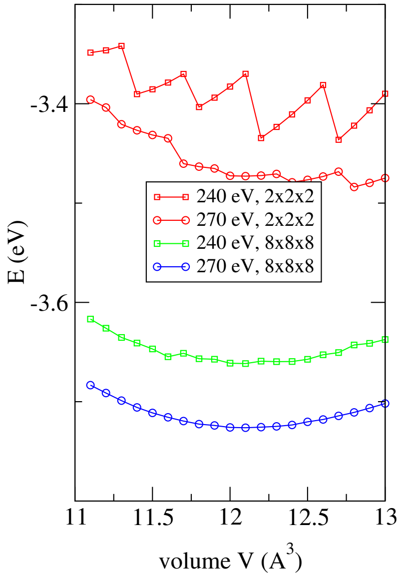
k-points and cutoff

e.g.: energy versus volume for fcc Cu

- by using more k-points or a higher energy cutoff, the energy surface becomes smoother

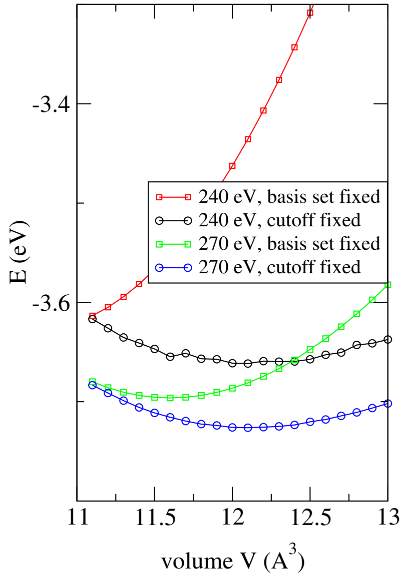
  at 270 eV and using 8x8x8 k-points, the energy varies smoothly
- in general, elastic constants are most prone to such errors

  if you sample the energy surface on a coarse scale, problems are less severe (recommended distortions 1 %)

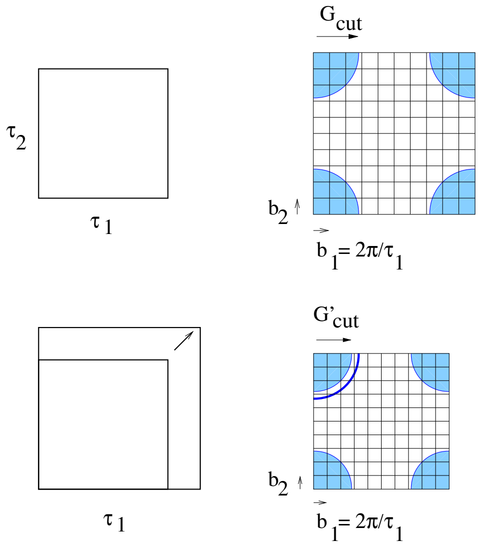


Number of plane wave fixed instead of fixed cutoff?

- such calculations clearly yield much too small volumes even at 270 eV (5 % error)
- effectively the cutoff decreases when the volume is increased (since the reciprocal lattice vectors become shorter)
- fixed basis set calculations are obviously a very bad idea



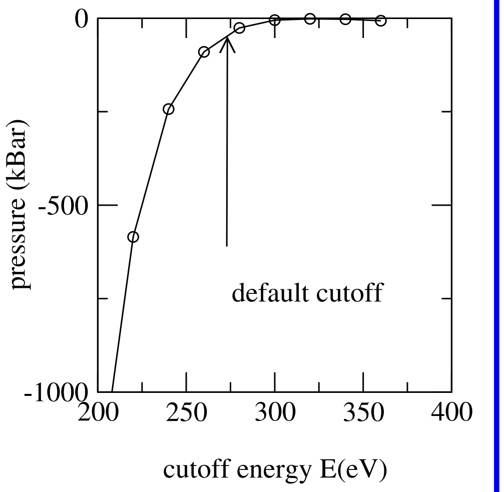
Fixed basis-set calculations



- the cutoff decreases by a factor τ_1/τ_1' when the lattice is expanded from $\tau_1 \to \tau_1'$
- for the expanded lattice the basis set corresponds effectively to a lower cutoff G'_{cut} and therefore a lower quality,
  $\Rightarrow$ the energy is overestimated at larger volumes
  $\Rightarrow$ the volume is underestimated for fixed basis-set calculations

Stress tensor and cell-shape relaxation

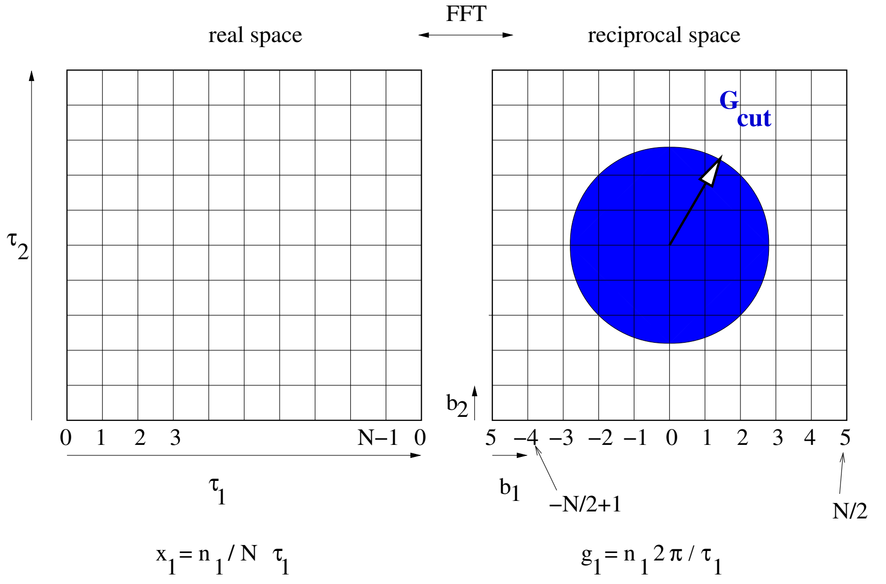
- the stress tensor is implicitly calculated at a fixed basis set
  upon cell-shape or volume relaxation one obtains too small volumes
  (2-5 % errors at the default cutoff)
- cutoff must be increased by 20-30%, when cell relaxations are performed
- calculations at the equilibrium lattice parameter of fcc Cu:
  270 eV: p=−50 kBar (contract)
  350 eV: a few kBar (correct result)



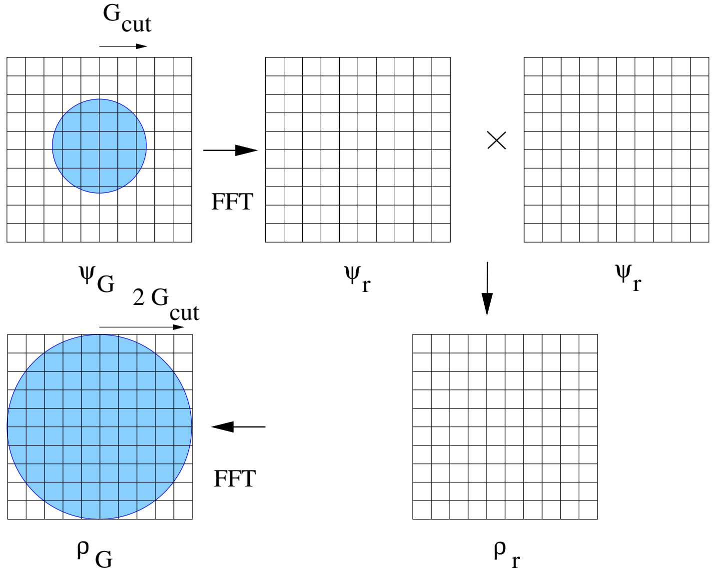
Cell-shape relaxations

- increase the cutoff by 30 %
  and restart the calculations, after the first ionic relaxation has succeeded
  the basis set is then adopted to the new geometry
- alternatively, but not recommended: `PREC= High`

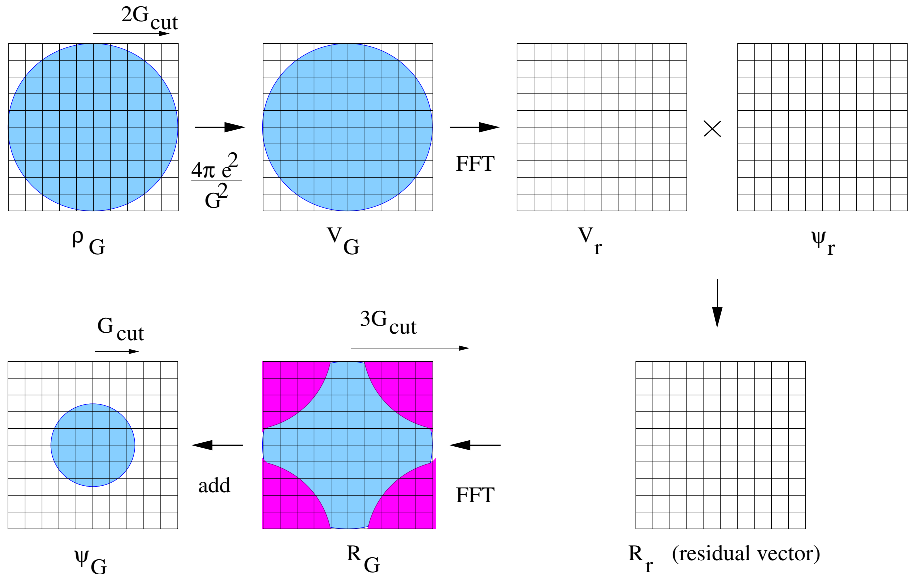
Technical errors related to the truncated FFT mesh



Evaluation of the charge density, folding theorem



Evaluation of the local part of the Hamiltonian $\mathbf{H}\psi$



The FFT grid

- the folding theorem implies that the charge density contains components up to $2 \times G_{\rm cut}$ where

$$\frac{\hbar^2}{2m_e}|G_{\rm cut}|^2 = E_{\rm cut}$$

- the Hartree potential contains Fourier components up to $2 \times G_{\rm cut}$ as well
- the residual vector $\mathbf{H}\psi$ contains Fourier components up to $3 \times G_{\rm cut}$

to avoid any errors, the Fourier grid must contain all wave-vectors up to $2 \times G_{\rm cut}$

this is true for both, the evaluation of the charge-density and the residual vector

- if this is not the case, components in the charge density are wrapped around from the other side of the box: "wrap around errors"

  the proper terminus technicus is aliasing errors

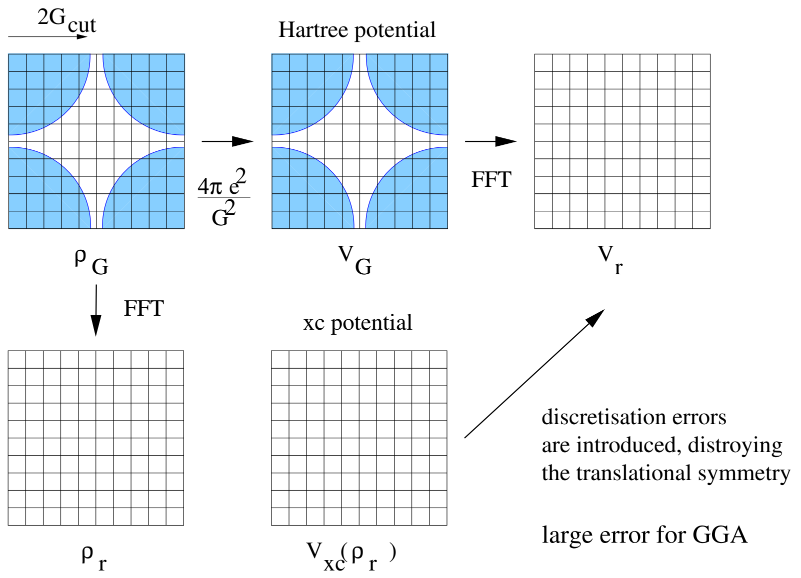
  high frequency components are aliased to low frequency components

What sort of errors does aliasing cause

- None, even translational symmetry is conserved
- the energies are slightly changed by aliasing errors (of the order of 1-2 meV/atom)
- aliasing errors are of course rather dependent on the cell/supercell, and they change when the supercell changes

supercells might not have exactly a multiple of the energy of a primitive cell

Exchange correlation potential



What sort of errors does this cause

- **the translational invariance is destroyed**

  if all atoms are shifted by an arbitrary vector $\vec{\tau}$ the energy should remain exactly identical

  this is however only the case, if discretisation errors are avoided

  equivalently, the sum of all ionic forces should be zero

$$\sum_{i=1}^{N_{\text{atoms}}} \vec{F}_i = 0$$

  offers a convenient way to check for such errors

- **symmetry inequivalent atoms are no longer strictly symmetry equivalent**

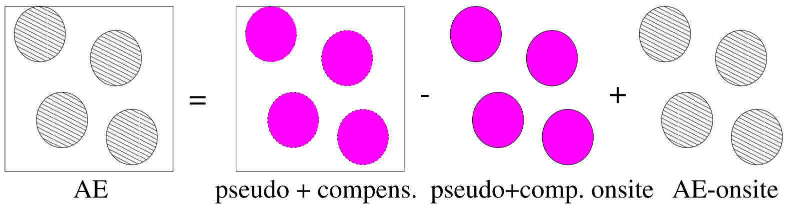
  VASP however symmetrises the charge and the forces explicitly

  to quantify this sort of errors, you need to switch off symmetry `ISYM=0`

The PAW compensation charge on regular grid

- the pseudo-wavefunctions do not have the same norm as the AE wavefunctions inside the spheres
- to deal with long range electrostatic interactions between spheres

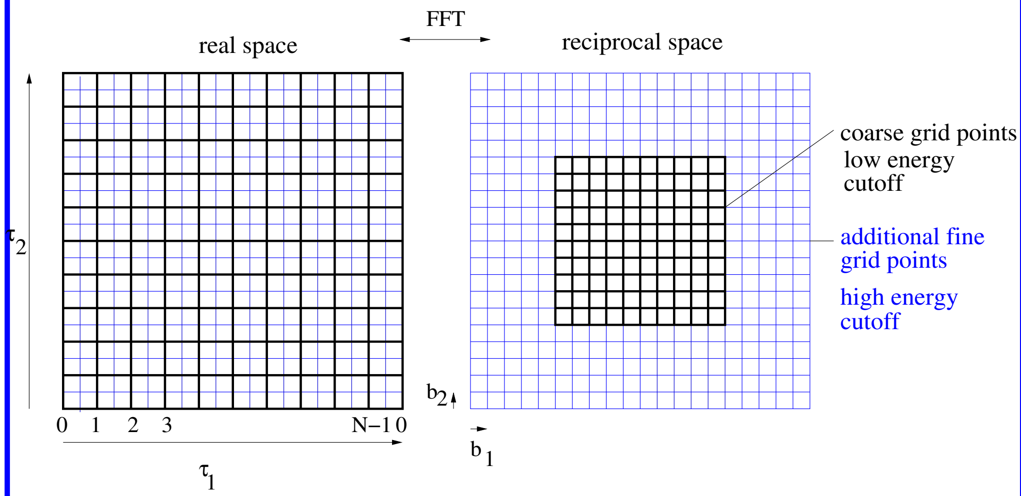
  a "soft" compensation charge $\hat{n}$ is introduced (similar to FLAPW)



- these localised compensation charges can be rather hard and are not well represented on the plane wave grid

  $\Rightarrow$ dual grid technique

Representation of the compensation charge: Dual grid technique



data transfer occurs only in reciprocal space, grids are not necessarily "aligned"

evaluation of the potentials (XC) is also done on the fine grid

Non local part of the potentials

- in the PAW and PP methods, the following expressions occurs in the evaluation of $\mathbf{H}|\psi_n\rangle$

$$\sum_{ij}^{\text{sites}} |\beta_j\rangle D_{ij} \langle\beta_i|\psi_n\rangle$$

- the expression can be evaluated in real space or reciprocal space

$$C_{in} = \langle\beta_i|\psi_n\rangle =$$

$$\sum_{\mathbf{G}} \bar{\beta}(\mathbf{k}+\mathbf{G})\psi_{\mathbf{G}+\mathbf{k}n} = \sum_{\mathbf{r}} \beta_i(\mathbf{r})\psi_{n\mathbf{k}}(\mathbf{r})$$

- in reciprocal space $N_{\text{planewaves}} \times N_{\text{ion}} \times N_{\text{proj}}$ operations are required
  $\mathbf{H}|\psi_n\rangle$ scales quadratically with the number of ions
- in real space $N_{\text{points}} \times N_{\text{ion}} \times N_{\text{proj}}$ operations are required, since $\beta_i(\mathbf{r})$ is localised around ions $\Rightarrow$ $\mathbf{H}|\psi_n\rangle$ scales linearly with the number of ions

Reciprocal / Real space

- reciprocal space (LREAL = False): slower but more accurate
  strongly recommended for phonon calculations (if you can afford it)
- real space (LREAL = True : much much faster, but less accurate
  the absolute energies are also slightly modified, and hence calculations with and without real space optimisation should not be compared

Three sources of inaccuracies

- coarse grid errors, FFT related
  - charge density steming from the soft part of the wavefunctions $\tilde{\Psi}$
  - application of the local part of the potential to the wavefunctions $\tilde{\Psi}$
- errors stemming from the representation of the compensation charges on the second finer grid
  related errors due to the xc-potential
  can be substantial for GGA's
- errors stemming from the non local part of the pseudopotential, when real space projection is selected

`total drift` in forces as written to the OUTCAR file indicates how accurate the calculations are

The `PREC` tag

the `PREC` tag allows to control the behaviour of VASP in a convenient manner, by influencing a number of other parameters

`PREC = Low | Medium | High |            Normal | Accurate`

- `Low`: only recommended for quick and dirty calculations
  e.g. initial relaxations with few k-points
- `Normal`: standard calculations
- `Accurate`: exceptional high accuracy
- the two older settings `Medium` and `High` are no longer recommended, although they are still supported

PREC *and* ENCUT

PREC	ENCUT	NGx	NGxF	ROPT
Normal	max(ENMAX)	3/2 G_{cut}	2 NGx	-5E-4
Accurate	max(ENMAX)	2 G_{cut}	2 NGx	-2.5E-4
Low	max(ENMIN)	3/2 G_{cut}	3 G_{aug}	-1E-2
Med	max(ENMAX)	3/2 G_{cut}	4 G_{aug}	-2E-3
High	max(ENMAX)*1.3	2 G_{cut}	16/3 G_{aug}	-4E-4

$$\frac{\hbar^2}{2m_e}|G_{\text{cut}}|^2 = \texttt{ENCUT} \qquad \frac{\hbar^2}{2m_e}|G_{\text{aug}}|^2 = \texttt{ENAUG}$$

max(ENMAX/ENMIN) corresponds to the maximum ENMAX/ENMIN found in POTCAR

ENAUG defaults to the maximum EAUG found in POTCAR

PREC= Normal *and* Accurate

- for Accurate wrap around errors are avoided, whereas for Normal 3/4 of the required grid dimensions are used

  Normal is an excellent compromise
- the energy cutoff ENCUT should be set manually in any case

  this makes the calculations more concise and better controlled

  for stress calculations and cell shape deformations, one might need to increase ENCUT from the default value
- the grids for the compensations charges have exactly twice the dimension than those of the coarser grids

  (Hartree and XC potentials are also evaluated on those grids)

PREC= Normal offers a very high accuracy at modest computational costs

PREC= Medium *and* High

- for High wrap around errors are avoided as for Accurate

  for Medium 3/4 of the required grid dimensions are used as for Normal

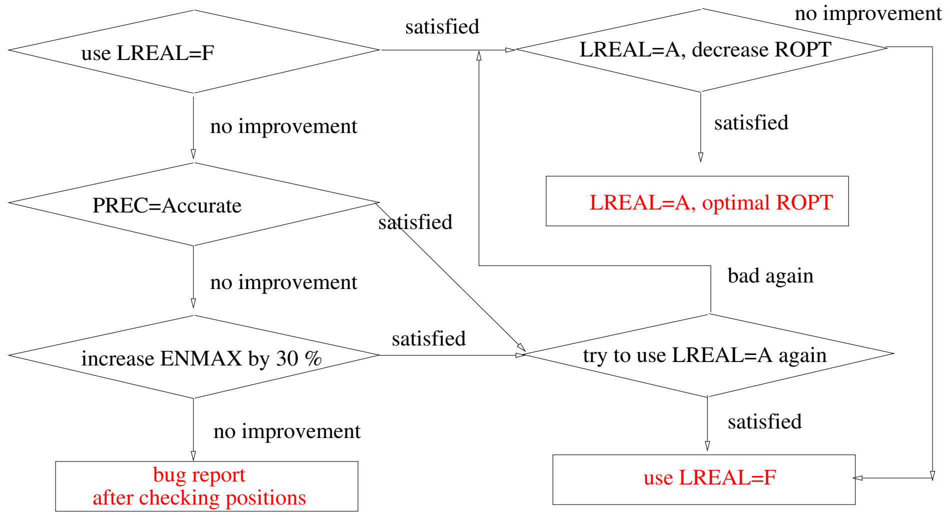
  for High the energy cutoff is increased

  It is recommended to do this manually in the INCAR file

  ENCUT should be specified manually for any calculation
- the defaults for ROPT were not sufficiently accurate for Medium and High
- the grids for the augmentations charges are controlled by ENAUG

  this offers more flexibility, but the "doubled grids" used for Normal and Accurate are more precise and do not cost a noticable amount of computer time

What to do, if you are not satisfied with the forces (drift)



The most common mistakes

- energy differences from calculations with different energy cutoffs

  Pt slab calculations with 3x3x4 atoms at the default cutoff 230 eV
  add CO molecule and calculate adsorption energy (CO default 400 eV)

  errors will be proportional to the number of Pt atoms and around 200 meV per Pt atom

- energy differences from calculations with different KPOINTS

Validating results

- cutoff and aliasing errors:
  - increase the cutoff
  - or try to perform `PREC=Accurate` calculations
  - possibly switch off the real space optimisation
- Related errors:
  - increase the number of k-points
  - increase the slab thickness
  - for defects increase the size of the supercell to remove artificial interactions

  TEST, TEST, TEST ....

Solving the Kohn-Sham equations II

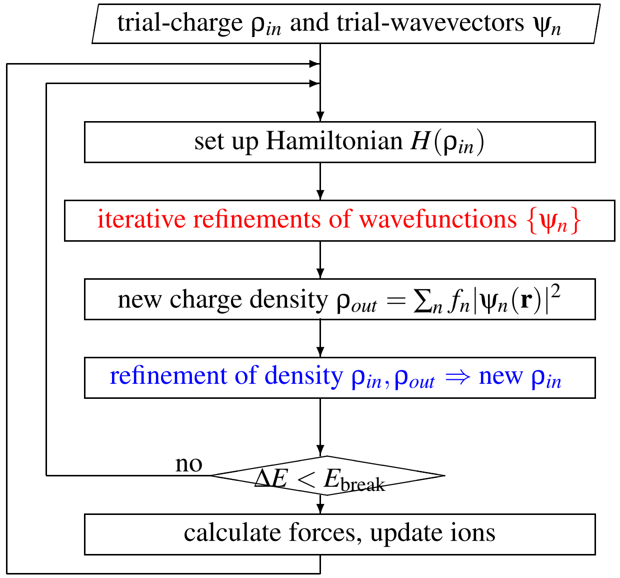
Algorithms implemented in VASP

- Updating the wavefunctions
  - Blocked Davidson algorithm
  - RMM-DIIS: Residuum minimization method - direct inversion in the iterative subspace: minimize norm $\langle R_n \mid R_n \rangle$ of residual vector to each eigenstate (no orthogonality constraint)

$$| R(\phi_n) \rangle = (\mathbf{H} - \varepsilon_n^{app}) \mid \phi_n \rangle, \quad \varepsilon_n^{app} = \langle \phi_n \mid \mathbf{H} \mid \phi_n \rangle \tag{15}$$

- Mixing:
  - DIIS

Selfconsistency Scheme

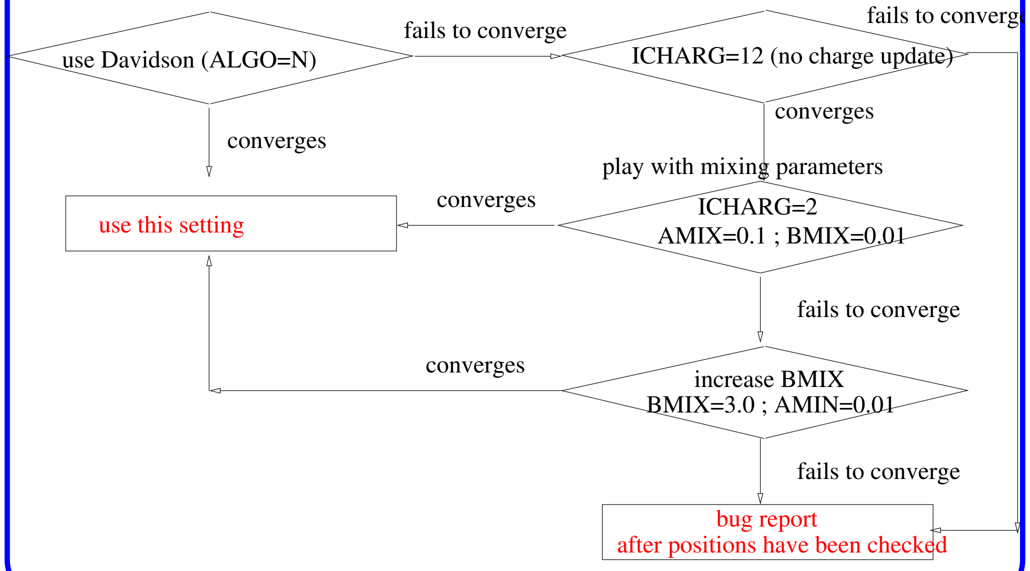


- two subproblems optimization of $\{\psi_n\}$ and ρ_{in}
- refinement of density: DIIS algorithm P. Pulay, Chem. Phys. Lett. **73**, 393 (1980).
- refinement of wavefunctions: DIIS or Davidson algorithm

ALGO flag

- ALGO determines how the wavefunctions are optimized
  all algorithms are fully parallel for any data distribution scheme
  - ALGO= Normal (default): blocked Davidson algorithm
  - ALGO= Very Fast: DIIS algorithm
  - ALGO= Fast: 5 initial steps blocked Davidson, afterwards DIIS algorithm
    after ions are moved: 1 Davidson step, afterwards again DIIS
- RMM-DIIS is 1.5 to 2 times faster, but Davidson is more stable
  ALGO= Fast is a very reasonable compromise, and should be specified for system with more than 10-20 atoms
- generally the user can not influence the behavior of these algorithms (delicately optimized black box routines)

What to do when electronic convergence fails



Ionic relaxation: algorithms implemented in VASP

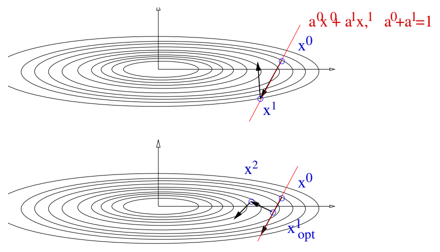
		additional flags	termination
DIIS	IBRION =1	POTIM=0.5, NFREE=5	EDIFFG
CG	IBRION =2	POTIM=0.5	EDIFFG
damped MD	IBRION =3	POTIM=0.15-0.4, SMASS=0.4	EDIFFG

POTIM determines generally the step size
for the CG gradient algorithm, where line minisations are performed, this is the size of the very first trial step

EDIFFG determines when to terminate relaxation
positive values: energy change between steps must be less than EDIFFG
negative values: $|\vec{F}_i| < |\texttt{EDIFFG}| \quad \forall i = 1, N_{ions}$

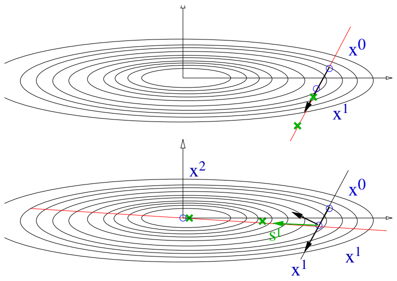
Full DIIS algorithm

1. steepest descent step from $\vec{x}^0$ to $\vec{x}^1$ (arrows correspond to gradients $\vec{g}_0$ and $\vec{g}_1$)
2. gradient along indicated red line is now know, determine optimal position $\vec{x}^1_{opt}$
3. another steepest descent step form $\vec{x}^1_{opt}$ along $\vec{g}_{opt} = \vec{g}(\vec{x}^1_{opt})$
4. calculate gradient $x^2 \Rightarrow$ now the gradient is known in the entire 2 dimensional space (linearity condition) and the function can be minimised exactly



Conjugate gradient

1. steepest descent step from $\vec{x}^0$, search for minimum along $\vec{g}_0$ by performing several trial steps (crosses, at least one trialstep is required) $\rightarrow \vec{x}^1$
2. determine new gradient $\vec{g}_1 = \vec{g}(\vec{x}_1)$ and conjugate it to get $\vec{s}_1$ (green arrow)
   for 2d-functions the gradient points now directly to the minimum
3. minimisation along search direction $\vec{s}_1$



Asymptotic convergence rate

- asymptotic convergence rate is the convergence behaviour for the case that the degrees of freedom are much larger than the number of steps
  e.g. 100 degrees of freedom but you perform only 10-20 steps
- how quickly, do the forces decrease?
- this depends entirely on the eigenvalue spectrum of the Hessian matrix:
  – steepest descent: $\Gamma_{\max}/\Gamma_{\min}$ steps are required to reduce the forces to a fraction ε
  – DIIS, CG, damped MD: $\sqrt{\Gamma_{\max}/\Gamma_{\min}}$ steps are required to reduce the forces to a fraction ε

$\Gamma_{\max}, \Gamma_{\min}$ are the maximum and minimal eigenvalue of the Hessian matrix

Damped molecular dynamics

instead of using a fancy minimisation algorithms it is possible to treat the minimisation problem using a simple “simulated annealing algorithm”

- regard the positions as dynamic degrees of freedom
- the forces serve as accelerations and an additional friction term is introduced
- equation of motion ($\vec{x}$ are the positions)

$$\ddot{\vec{x}} = -2 * \alpha \vec{g}(\vec{x}) - \mu \dot{\vec{x}},$$

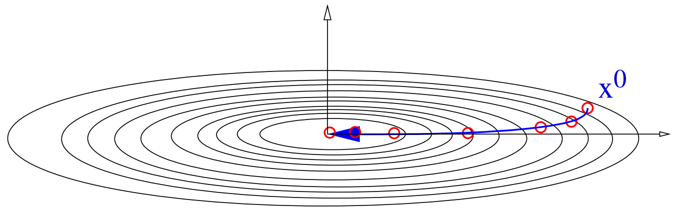
using a velocity Verlet algorithm this becomes

$$\vec{v}_{N+1/2} = \Big((1-\mu/2)\vec{v}_{N-1/2} - 2 * \alpha \vec{F}_N \Big) / (1+\mu/2)$$

$$\vec{x}_{N+1} = \vec{x}_{N+1} + \vec{v}_{N+1/2}$$

for $\mu = 2$, this is equivalent to a simple steepest descent step

- behaves like a rolling ball with a friction
  it will accelerate initially, and then deaccelerate when close to the minimum
- if the optimal friction is chosen the ball will glide right away into the minimum
- for a too small friction it will overshoot the minimum and accelerate back
- for a tool large friction relaxation will also slow down (behaves like a steepest descent)



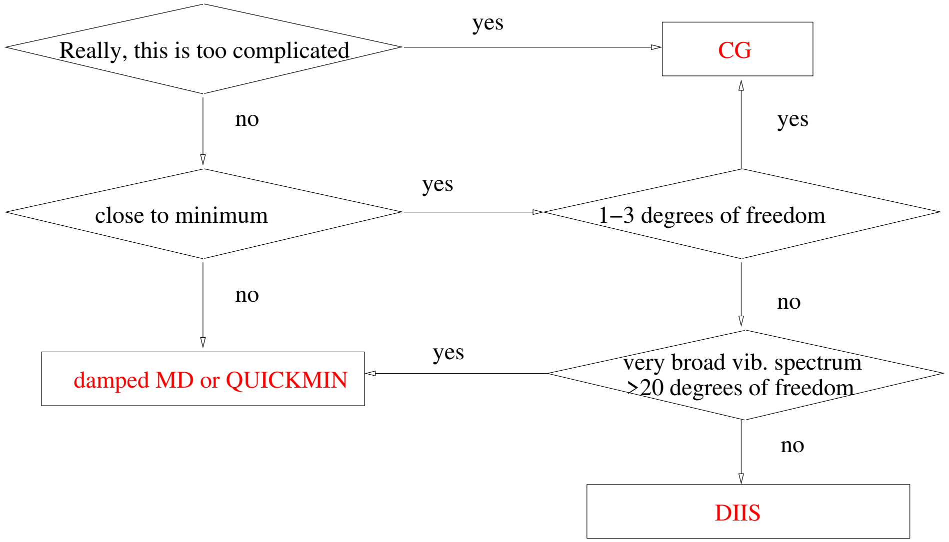
Strength and weakness

- DIIS: Quasi-Newton algorithm, stores an approximation of the Hessian matrix with maximal rank `NFREE`

  very good close to minimum (quadratic functions)
- CG: simple and slow
- damped MD: very efficient when the start guess is bad, but requires one parameter `POTIM`

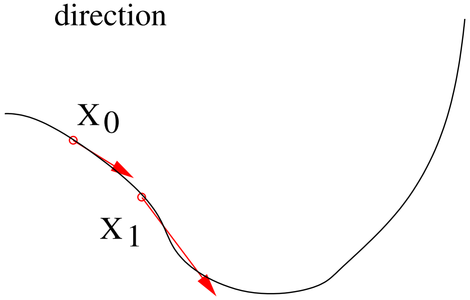
  for too large `POTIM` $\Rightarrow$ divergence; too small `POTIM` $\Rightarrow$ slow

Why so many algorithms :-(... decision chart



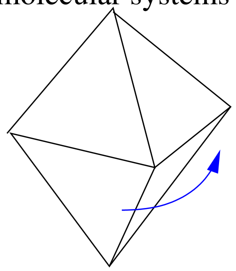
Two cases where the DIIS has huge troubles

force increases along the search direction



DIIS is dead, since it consideres the forces only
it will move uphill instead of down

rigid unit modes i.e. in perovskites (rotation)
molecular systems (rotation)



in cartesian coordinates
the Hessian matrix changes
when the octahedron rotates!

Sampling the Brillouin zone

For many properties
(e.g.: density of states, charge density, matrix elements, response functions, ...)
integrals (I) over the Brillouin-zone are necessary:

$$I(\varepsilon) = \frac{1}{\Omega_{BZ}} \int_{BZ} F(\varepsilon)\delta(\varepsilon_{n\vec{k}} - \varepsilon) d\vec{k}$$

To evaluate computationally
integrals $\Rightarrow$ weighted sum over special k-points

$$\frac{1}{\Omega_{BZ}} \int_{BZ} \quad \Rightarrow \quad \sum_{\vec{k}} \omega_{\vec{k}_i}$$

Monkhorst and Pack (1976):

Idea: equally spaced mesh in Brillouin-zone.
Algorithm:

- calculate equally spaced-mesh
- shift the mesh if desired
- apply all symmetry operations of Bravais lattice to all k-points
- extract the irreducible k-points ($\equiv$ IBZ)
- calculate the proper weighting
- Two choices for the center of the mesh
  - centered on Γ ($\Rightarrow \Gamma$ belongs to mesh).
  - centered around Γ. (can break symmetry !!)

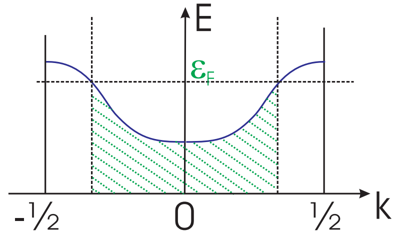
Smearing methods

Problem: in metallic systems Brillouin-zone integrals over functions that are discontinuous at the Fermi-level.
$\Rightarrow$ high Fourier-components $\Rightarrow$ dense grid is necessary.
Solution: replace step function by a smoother function.
Example: bandstructure energy

$$\sum_{n\mathbf{k}} \omega_{\mathbf{k}} \varepsilon_{n\mathbf{k}} \bar{\Theta}(\varepsilon_{n\mathbf{k}} - \mu)$$

$$\text{with:} \bar{\Theta}(x) = \begin{cases} 1 & x \leq 0 \\ 0 & x > 0 \end{cases}$$

$$\Rightarrow \sum_{n\mathbf{k}} \omega_{\mathbf{k}} \varepsilon_{n\mathbf{k}} f\left(\frac{\varepsilon_{n\mathbf{k}} - \mu}{\sigma}\right)$$



necessary: appropriate function $f \Rightarrow f$ equivalent to partial occupancies.

Gaussian smearing

broadening of energy-levels with Gaussian function.
$\Rightarrow f$ becomes an integral of the Gaussian function:
$f\left(\frac{\varepsilon_{n\mathbf{k}}-\mu}{\sigma}\right) = \frac{1}{2}\left[1 - \text{erf}\left(\frac{\varepsilon_{n\mathbf{k}}-\mu}{\sigma}\right)\right]$
no analytical inversion of the error-function erf exists $\Rightarrow$ entropy and free energy cannot be written in terms of f.
$S\left(\frac{\varepsilon-\mu}{\sigma}\right) = \frac{1}{2\sqrt{\pi}} \exp\left[-\left(\frac{\varepsilon-\mu}{\sigma}\right)^2\right]$

- σ has no physical interpretation.
- variational functional $F(\sigma)$ differs from $E(0)$.
- forces are calculated as derivatives of the variational quantity ($F(\sigma)$). $\Rightarrow$ not necessarily equal to forces at $E(0)$.
- **Improvement:** extrapolation to $\sigma \rightarrow 0$.

Method of Methfessel and Paxton (1989)

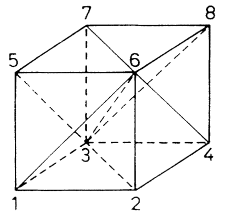
Idea: expansion of stepfunction in a complete set of orthogonal functions
$\Rightarrow$ term of order 0 = integral over Gaussians
$\Rightarrow$ generalization of Gaussian broadening with functions of higher (N) order.
advantages:

- deviation of $F(\sigma)$ from $E(0)$ only of order 2+N in σ
- extrapolation for $\sigma \rightarrow 0$ usually not necessary, but also possible:
- MP of order N leads to a negligible error, if $X(\varepsilon)$ is representable as a polynomial of degree $2N$ around ε_F.
- linewidth σ can be increased for higher order to obtain the same accuracy
- entropy term ($\mathbf{S} = \sigma \sum_n S_N(f_n)$) describes deviation of $F(\sigma)$ from $E(\sigma)$.
  $\Rightarrow$ if $\mathbf{S}<$ few meV, $\hat{E}(\sigma) \approx F(\sigma) \approx E(\sigma) \approx E(0)$.
  $\Rightarrow$ forces correct within that limit.

Linear tetrahedron method

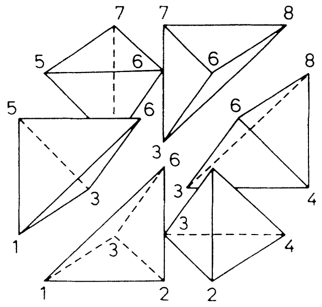
Idea:

1. dividing up the Brillouin-zone into tetrahedra
2. Linear interpolation of the function to be integrated X_n within these tetrahedra
3. integration of the interpolated function $\bar{X}_n$



Drawbacks:

- tetrahedra can break the symmetry of the Bravaislattice
- at least 4 k-points are necessary
- Γ must be included
- linear interpolation under- or overestimates the real curve



Corrections by Blöchl (1993)

Idea:

- linear interpolation under- or overestimates the real curve
- for full-bands or insulators these errors cancel
- for metals: correction of quadratic errors

Result:

- best k-point convergence for energy
- forces:
  - with Blöchl corrections the new effective partial occupancies do not minimize the groundstate total energy
  - variation of occupancies $\omega_{n\vec{k}}$ w.r.t. the ionic positions would be necessary (with US-PP and PAW practically impossible)

What to do in practice

energy/DOS calculations:
linear tetrahedron method with Blöchl corrections (ISMEAR=-5)
calculation of forces:

- semiconductors: Gaussian smearing (ISMEAR=0; SIGMA=0.1)
- metals : Methfessel-Paxton (N=1 or 2)
- always: test for energy with LT+Blöchl-corr.

in any case: careful checks for k-point convergence are necessary

- determine the number of intersections in each direction
- longer axes in real-space $\Leftrightarrow$ shorter axes in k-space $\Rightarrow$ less intersections necessary for equally spaced mesh

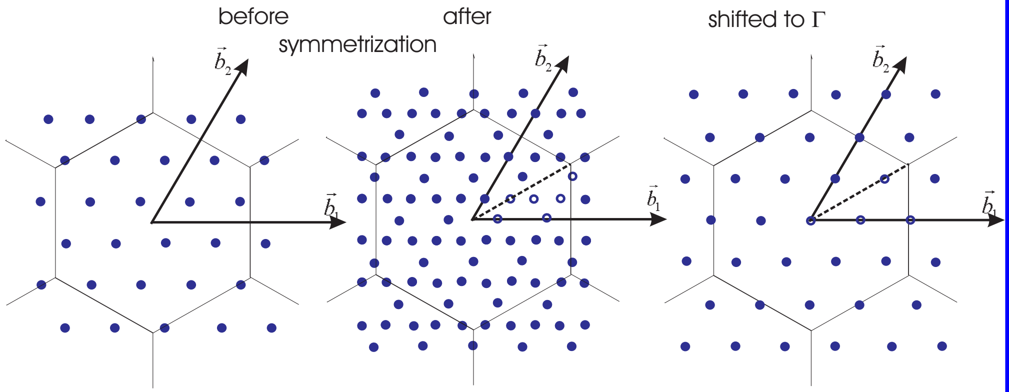
Consequences:

- molecules, atoms (large supercells): $(1 \times 1 \times 1)(\equiv \Gamma)$ is enough.
- surfaces (one long direction $\Rightarrow$ 2-D Brillouin-zone): $(x \times y \times 1)$
- typical values (never trust them!):

metals:	$(9 \times 9 \times 9)$/atom
semiconductors:	$(4 \times 4 \times 4)$/atom

- doubling the cell in real space halves the reciprocal cell
  $\Rightarrow$ zone boundary is folded back to Γ
- same sampling is achieved with halved mesh parameter
- TEST, TEST, TEST....

Example - hexagonal cell



- in certain cell geometries (e.g. hex) even meshes break the symmetry
- symmetrization results in non equally distributed k-points
- Gamma point centered mesh preserves symmetry

Platforms for VASP

important hardware parameters

cpu	the cpu throughput is very important for the vasp performance: many time consuming routines in vasp use BLAS/LAPACK, streaming methods give an enormous performance boost (Intel's SSE2)
memory	vasp requires 512 - 1024 MByte / cpu; computational speed of vasp ≈ sustained memory bandwidth
HD	non critical, large enough to hold `WAVECAR` and `CHG*`
parallel	fast but cheap network

Sofware

- in any case..
  Compiler and Libraries
  - FORTRAN 90 compiler
  - BLAS, LAPACK
- parallelisation:
  - parallel: MPI software
  - Message Passing Interface (MPI), Open-Source (lam, mpich)
  - scheduler with parallel support
  - cluster administration

Performance I

some remarks

- Clusters of PCs are an attractive platform for parallel applications because of their cost effectiveness
- VASP supports parallel execution on commodity components using MPI for parallel communication

performance benchmarks

using vasp as a benchmark

- always use the same vasp version, otherwise the timings are not comparable
- use the best compiler options and the fastest available blas for the test platform

Performance II

some remarks

- How to improve performance?
  - on the supported platforms we have included high performing settings (makefile, algorithm)
  - new compiler & new hardware:
    - ∗ start from a makefile for a similar system
    - ∗ check compiler flags
    - ∗ use vendor BLAS/LAPACK or generate a highly optimized atlas
    - ∗ try different `CACHE_SIZE` settings
- different results on different hardware/software
  - small differences caused by:
    - ∗ different implementations of BLAS/LAPACK
    - ∗ different optimizations of the compiler
    - ∗ different behavior of the FPU (pc80, rounding)

  - large errors and/or crashes
    - ∗ hardware failure (check your memory)
    - ∗ too high optimization
- vasp running on system XXX ?
  - FORTRAN 90 compiler supporting the standard F90 language
  - BLAS/LAPACK `http://www.netlib.org` it's free, but usually slower than vendor supplied
  - some scripting language
  - parallel execution: standard MPI V1 implementation

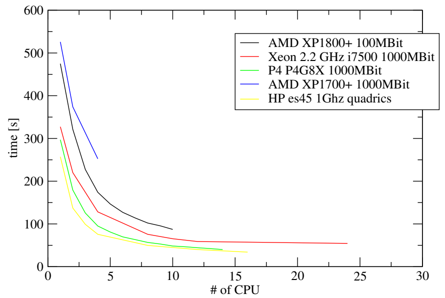
parallel LINUX clusters II

some important notes

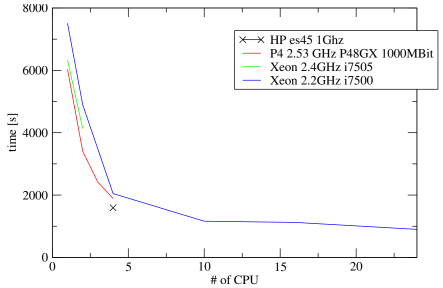
```
FC=mpif77
CPP    = $(CPP_) -DMPI \
          ....
         -DNGZhalf \
         -DMPI_BLOCK=2000 \
          ....
FFT3D   = fftmpi.o fftmpi_map.o \
           fft3dlib.o
```

- use the wrapper from your MPI installation as compiler
- `-DMPI` uses the vasp mpi support
- use `-DNGZhalf` instead of `-DNGXhalf` the MPI blocksize influences the communication time on the MPI net
- vasp supports lam and mpich MPI implementations, but lam is preferred

parallel LINUX clusters III



benchmark: bench_Hg
small system
# of CPU's > 8
Latency bound



benchmark: bench_PdO
large system
# of CPU's > 16
Latency bound

parallel LINUX clusters IV

important parameters for cluster running vasp

- **network**
  - **latency bound** gives a lower bound for the MPI communication blocksize. limits the maximum number of nodes. a typical Ethernet based network (100MBit, 1000MBit) shows around 90μs latency. optimizations on modern Gigabit cards reduce the latency to 30μs
  - **bandwidth bound** the maximum transfer rate limits the communication speed. vasp uses all–to–all communication $\rightarrow$ collision bound
- **cpu**
  - overall performance is limited by the network efficiency $\rightarrow$ increasing the cpu speed alone in not enough
  - fast data-path from network device to cpu (memory)

Amdahl's Law 1967

speedup:

$s = \frac{t(1)}{t(N)}$ $t(1)$ ... serial time $t(N)$ ... time for N CPU's

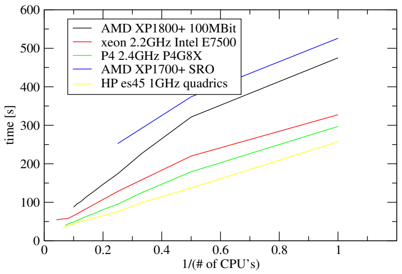
Amdahl's Law

$s = \frac{N}{(B\times N)+(1-B)}$ B ... % of algorithm that is serial

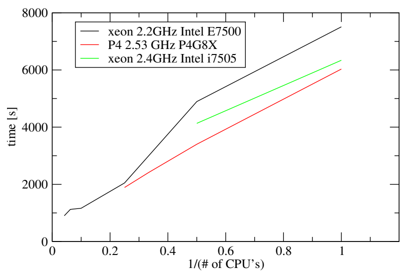
$\rightarrow$ it exists an upper limit on the number of CPU's for a given problemsize

$\rightarrow$ scale the problem with the number of CPU's

$\rightarrow$ use fast Computers with a good price performance ratio



small bench_Hg



big bench_PdO

The VASP team...

- Prof. Jürgen Hafner
- Prof. Georg Kresse
- Orest Dubay
- (Andreas Eichler)
- (Robin Hirschl)
- (Robert Lorenz)
- Martijn Marsman
- Florian Mittendorfer
- Joachim Paier