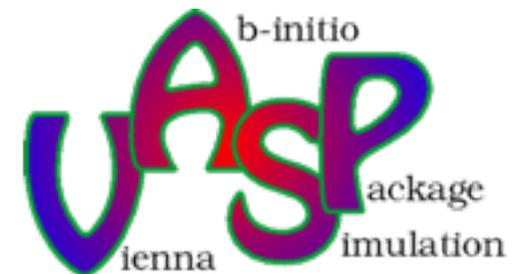# VASP: running on HPC resources

University of Vienna,
Faculty of Physics and Center for Computational Materials Science,
Vienna, Austria

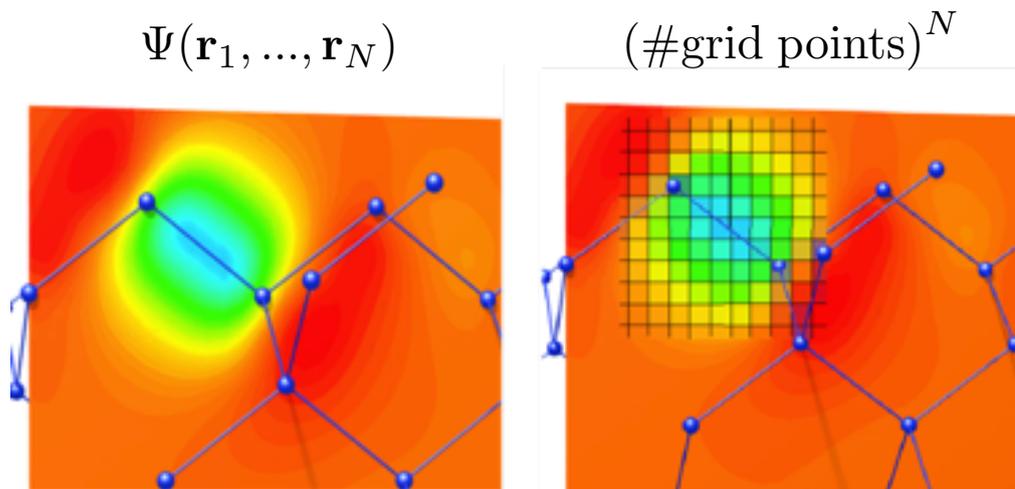# The Many-Body Schrödinger equation

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

$$\left(-\frac{1}{2}\sum_i \Delta_i + \sum_i V(\mathbf{r}_i) + \sum_{i \neq j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}\right)\Psi(\mathbf{r}_1, ..., \mathbf{r}_N) = E\Psi(\mathbf{r}_1, ..., \mathbf{r}_N)$$

For instance, many-body WF storage demands are prohibitive:

$$\Psi(\mathbf{r}_1, ..., \mathbf{r}_N) \qquad (\#\text{grid points})^N$$

5 electrons on a 10×10×10 grid ~ 10 PetaBytes !

A solution: map onto "one-electron" theory:

$$\Psi(\mathbf{r}_1, ..., \mathbf{r}_N) \rightarrow \{\psi_1(\mathbf{r}), \psi_2(\mathbf{r}), ..., \psi_N(\mathbf{r})\}$$

Hohenberg-Kohn-Sham DFT

Map onto "one-electron" theory:

$$\Psi(\mathbf{r}_1, ..., \mathbf{r}_N) \rightarrow \{\psi_1(\mathbf{r}), \psi_2(\mathbf{r}), ..., \psi_N(\mathbf{r})\} \qquad \Psi(\mathbf{r}_1, ..., \mathbf{r}_N) = \prod_i^N \psi_i(\mathbf{r}_i)$$

Total energy is a functional of the density:

$$E[\rho] = T_s[\{\psi_i[\rho]\}] + E_H[\rho] + E_{\mathrm{xc}}[\rho] + E_Z[\rho] + U[Z]$$

The density is computed using the one-electron orbitals:

$$\rho(\mathbf{r}) = \sum_i^N |\psi_i(\mathbf{r})|^2$$

The one-electron orbitals are the solutions of the Kohn-Sham equation:

$$\left(-\frac{1}{2}\Delta + V_Z(\mathbf{r}) + V_H[\rho](\mathbf{r}) + V_{\mathrm{xc}}[\rho](\mathbf{r})\right)\psi_i(\mathbf{r}) = \epsilon_i \psi_i(\mathbf{r})$$

BUT:

$$E_{\mathrm{xc}}[\rho] = ??? \qquad V_{\mathrm{xc}}[\rho](\mathbf{r}) = ???$$

Exchange-Correlation

$$E_{\mathrm{xc}}[\rho] = ??? \qquad V_{\mathrm{xc}}[\rho](\mathbf{r}) = ???$$

- Exchange-Correlation functionals are modeled on the uniform-electron-gas (UEG): The correlation energy (and potential) has been calculated by means of Monte-Carlo methods for a wide range of densities, and has been parametrized to yield a density functional.

- LDA: we simply pretend that an inhomogeneous electronic density locally behaves like a homogeneous electron gas.

- Many, many, many different functionals available:
  LDA, GGA, meta-GGA, van-der-Waals functionals, etc etc

An N-electron system: N = O(10^{23})

Hohenberg-Kohn-Sham DFT takes us a long way:

$$\Psi(\mathbf{r}_1, ..., \mathbf{r}_N) \rightarrow \{\psi_1(\mathbf{r}), \psi_2(\mathbf{r}), ..., \psi_N(\mathbf{r})\}$$

$$(\#\text{grid points})^N \qquad N \times (\#\text{grid points})$$

Nice for atoms and molecules, but in a realistic piece of solid state material N= O(10^{23})!

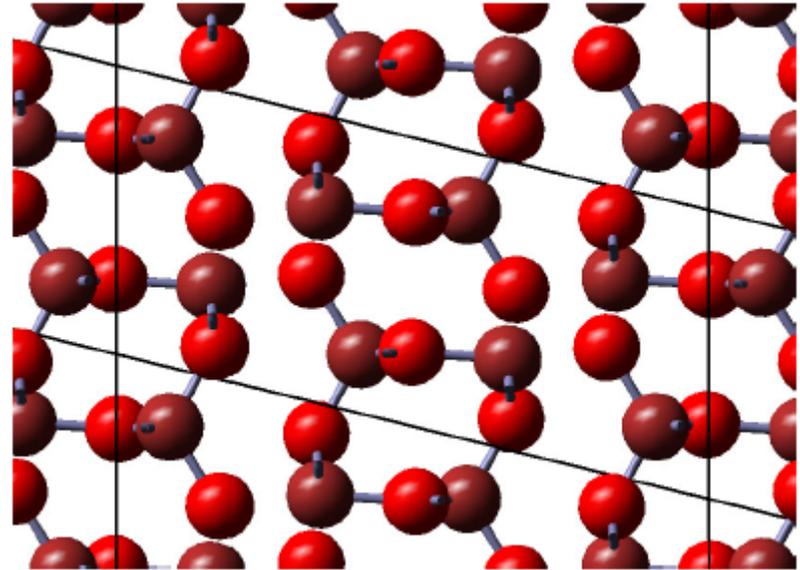# Translational invariance: Periodic Boundary Conditions

Translational invariance implies:

$$\psi_{n\mathbf{k}}(\mathbf{r} + \mathbf{R}) = \psi_{n\mathbf{k}}(\mathbf{r})e^{i\mathbf{k}\mathbf{R}}$$

and

$$\psi_{n\mathbf{k}}(\mathbf{r}) = u_{n\mathbf{k}}(\mathbf{r})e^{i\mathbf{k}\mathbf{r}}$$
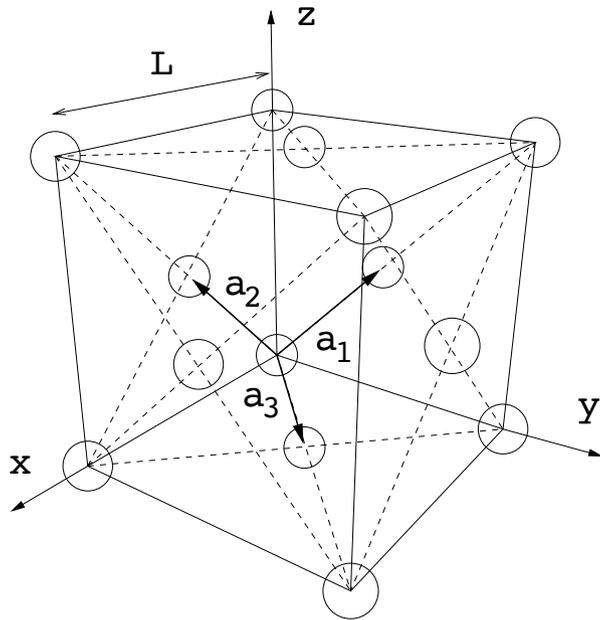
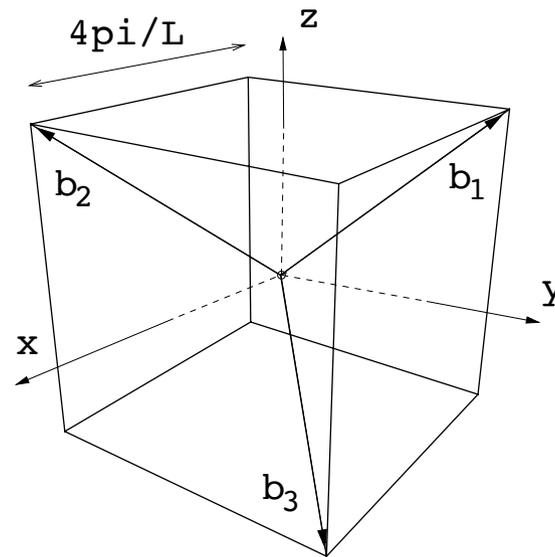$$u_{n\mathbf{k}}(\mathbf{r} + \mathbf{R}) = u_{n\mathbf{k}}(\mathbf{r})$$

All states are labeled by *Bloch vector* **k** and the *band index* n:

- The Bloch vector **k** is usually constrained to lie within the first Brillouin zone of the reciprocal space lattice.

- The band index *n* is of the order if the number of electrons per unit cell.
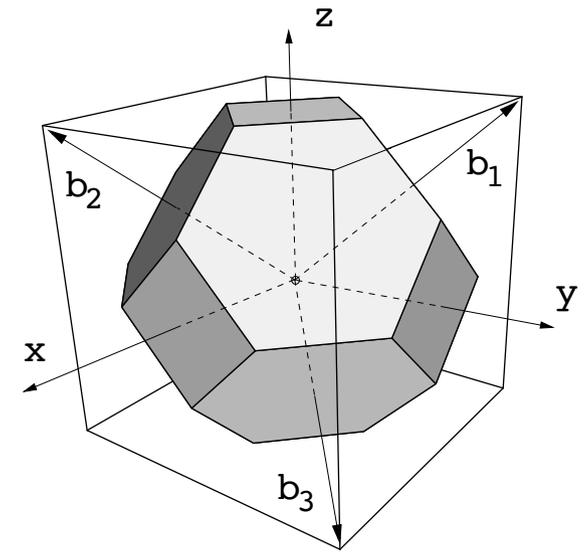
Reciprocal space & the first Brillouin zone

$$\mathbf{b}_1 = \frac{2\pi}{\Omega}\mathbf{a}_2 \times \mathbf{a}_3 \quad \mathbf{b}_2 = \frac{2\pi}{\Omega}\mathbf{a}_3 \times \mathbf{a}_1 \quad \mathbf{b}_3 = \frac{2\pi}{\Omega}\mathbf{a}_1 \times \mathbf{a}_2$$

$$\Omega = \mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3 \qquad \mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}$$

Sampling the 1$^\text{st}$ BZ

The evaluation of many key quantities involves an integral over the 1$^\text{st}$ BZ.
For instance the charge density:

$$\rho(\mathbf{r}) = \frac{1}{\Omega_{\text{BZ}}} \sum_n \int_{\text{BZ}} f_{n\mathbf{k}} |\psi_{n\mathbf{k}}(\mathbf{r})|^2 d\mathbf{k}$$

We exploit the fact that the orbitals at Bloch vectors $\mathbf{k}$ that are close together are almost identical and approximate the integral over the 1$^\text{st}$ BZ by a weighted sum over a discrete set of $\mathbf{k}$-points:

$$\rho(\mathbf{r}) = \sum_{n\mathbf{k}} w_{\mathbf{k}} f_{n\mathbf{k}} |\psi_{n\mathbf{k}}(\mathbf{r})|^2 d\mathbf{k},$$

Fazit: the intractable task of determining $\Psi(\mathbf{r}_1, ..., \mathbf{r}_N)$ with N=10^{23}, has been reduced to calculating $\psi_{n\mathbf{k}}(\mathbf{r})$ at a discrete set of $\mathbf{k}$-points in the 1$^\text{st}$ BZ, for a number of bands that is of the order if the number of electrons in the unit cell.

The total energy

$$E[\rho, \{\mathbf{R}, Z\}] = T_s[\{\psi_{n\mathbf{k}}[\rho]\}] + E_H[\rho, \{\mathbf{R}, Z\}] + E_{\mathrm{xc}}[\rho] + U(\{\mathbf{R}, Z\})$$

The kinetic energy

$$T_s[\{\psi_{n\mathbf{k}}[\rho]\}] = \sum_{n\mathbf{k}} w_{\mathbf{k}} f_{n\mathbf{k}} \langle \psi_{n\mathbf{k}} | -\frac{1}{2}\Delta | \psi_{n\mathbf{k}} \rangle$$

The Hartree energy

$$E_{\mathrm{H}}[\rho, \{\mathbf{R}, Z\}] = \frac{1}{2} \iint \frac{\rho_{eZ}(\mathbf{r})\rho_{eZ}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' d\mathbf{r}$$

where

$$\rho_{eZ}(\mathbf{r}) = \rho(\mathbf{r}) + \sum_i Z_i \delta(\mathbf{r} - \mathbf{R}_i) \qquad \rho(\mathbf{r}) = \sum_{n\mathbf{k}} w_{\mathbf{k}} f_{n\mathbf{k}} |\psi_{n\mathbf{k}}(\mathbf{r})|^2 d\mathbf{k}$$

The Kohn-Sham equations

$$\left(-\frac{1}{2}\Delta + V_H[\rho_{eZ}](\mathbf{r}) + V_{\mathrm{xc}}[\rho](\mathbf{r})\right)\psi_{n\mathbf{k}}(\mathbf{r}) = \epsilon_{n\mathbf{k}}\psi_{n\mathbf{k}}(\mathbf{r})$$

The Hartree potential

$$V_H[\rho_{eZ}](\mathbf{r}) = \int \frac{\rho_{eZ}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}'$$

A plane wave basis set

$$\psi_{n\mathbf{k}}(\mathbf{r}) = u_{n\mathbf{k}}(\mathbf{r})e^{i\mathbf{k}\mathbf{r}} \qquad\qquad u_{n\mathbf{k}}(\mathbf{r}+\mathbf{R}) = u_{n\mathbf{k}}(\mathbf{r})$$

All cell-periodic functions are expanded in plane waves (Fourier analysis):

$$u_{n\mathbf{k}}(\mathbf{r}) = \frac{1}{\Omega^{1/2}}\sum_{\mathbf{G}} C_{\mathbf{G}n\mathbf{k}}e^{i\mathbf{G}\mathbf{r}} \qquad\qquad \psi_{n\mathbf{k}}(\mathbf{r}) = \frac{1}{\Omega^{1/2}}\sum_{\mathbf{G}} C_{\mathbf{G}n\mathbf{k}}e^{i(\mathbf{G}+\mathbf{k})\mathbf{r}}$$

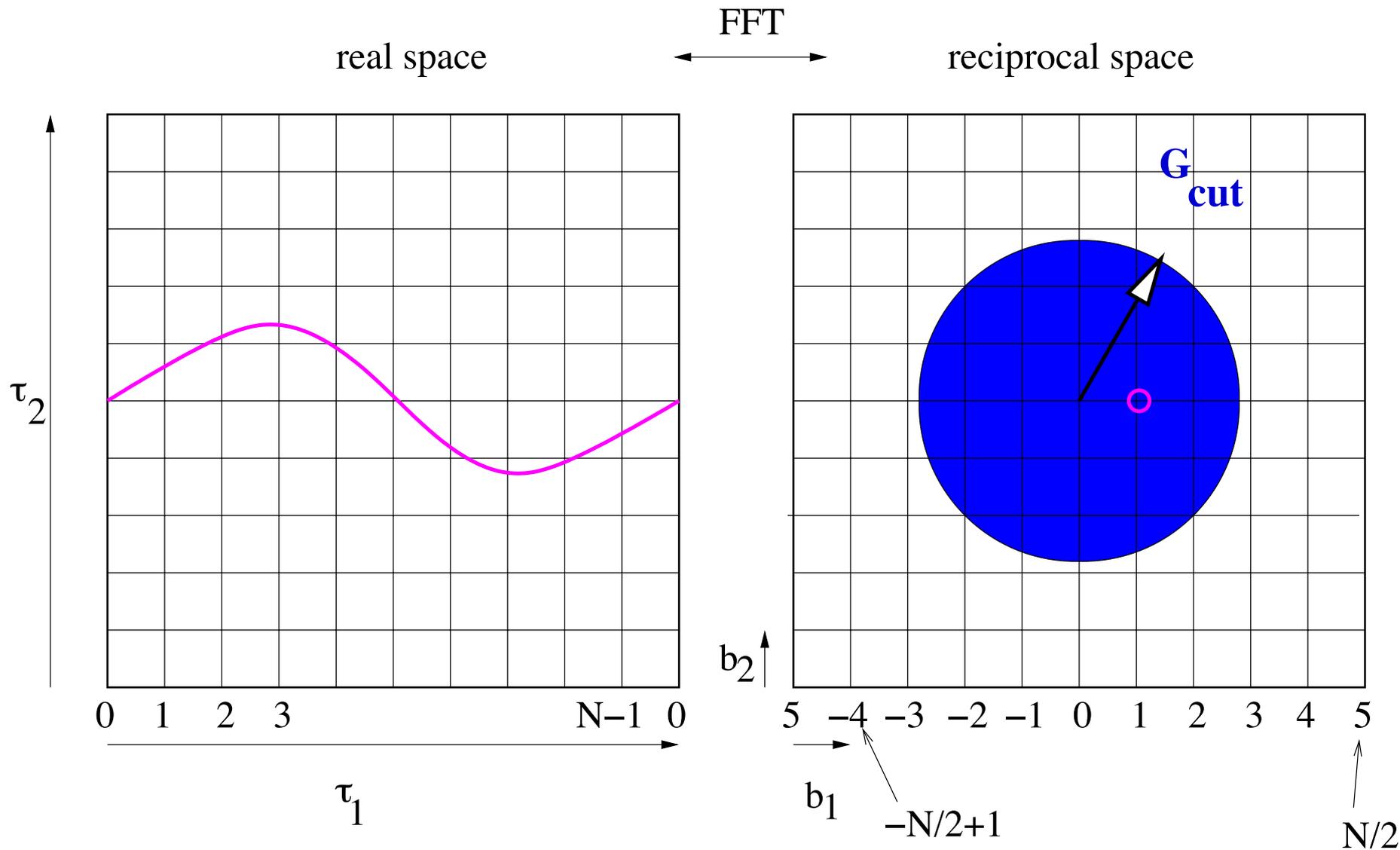$$\rho(\mathbf{r}) = \sum_{\mathbf{G}} \rho_{\mathbf{G}}e^{i\mathbf{G}\mathbf{r}} \qquad\qquad V(\mathbf{r}) = \sum_{\mathbf{G}} V_{\mathbf{G}}e^{i\mathbf{G}\mathbf{r}}$$

The basis set includes all plane waves for which

$$\frac{1}{2}|\mathbf{G}+\mathbf{k}|^2 < E_{\text{cutoff}}$$

Transformation by means of FFT between "real" space and "reciprocal" space:

$$C_{\mathbf{r}n\mathbf{k}} = \sum_{\mathbf{G}} C_{\mathbf{G}n\mathbf{k}}e^{i\mathbf{G}\mathbf{r}} \quad\xleftrightarrow{\text{FFT}}\quad C_{\mathbf{G}n\mathbf{k}} = \frac{1}{N_{\text{FFT}}}\sum_{\mathbf{r}} C_{\mathbf{r}n\mathbf{k}}e^{-i\mathbf{G}\mathbf{r}}$$
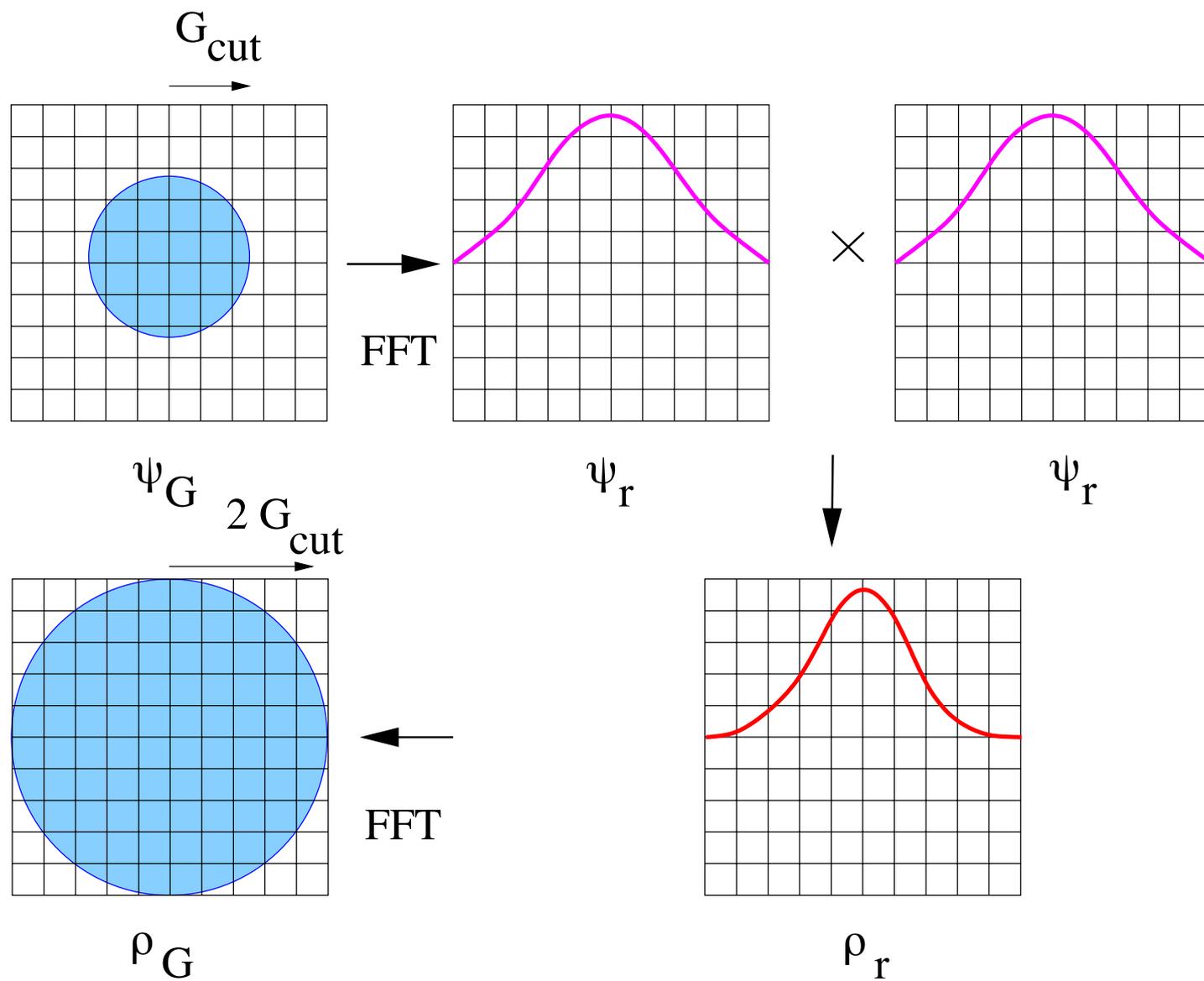
real space — FFT — reciprocal space

$$x_1 = n_1 / N \; \tau_1 \qquad\qquad g_1 = n_1 \, 2\pi / \tau_1$$

$$C_{\mathbf{r}n\mathbf{k}} = \sum_{\mathbf{G}} C_{\mathbf{G}n\mathbf{k}} e^{i\mathbf{G}\mathbf{r}} \quad \xleftrightarrow{\text{FFT}} \quad C_{\mathbf{G}n\mathbf{k}} = \frac{1}{N_{\text{FFT}}} \sum_{\mathbf{r}} C_{\mathbf{r}n\mathbf{k}} e^{-i\mathbf{G}\mathbf{r}}$$
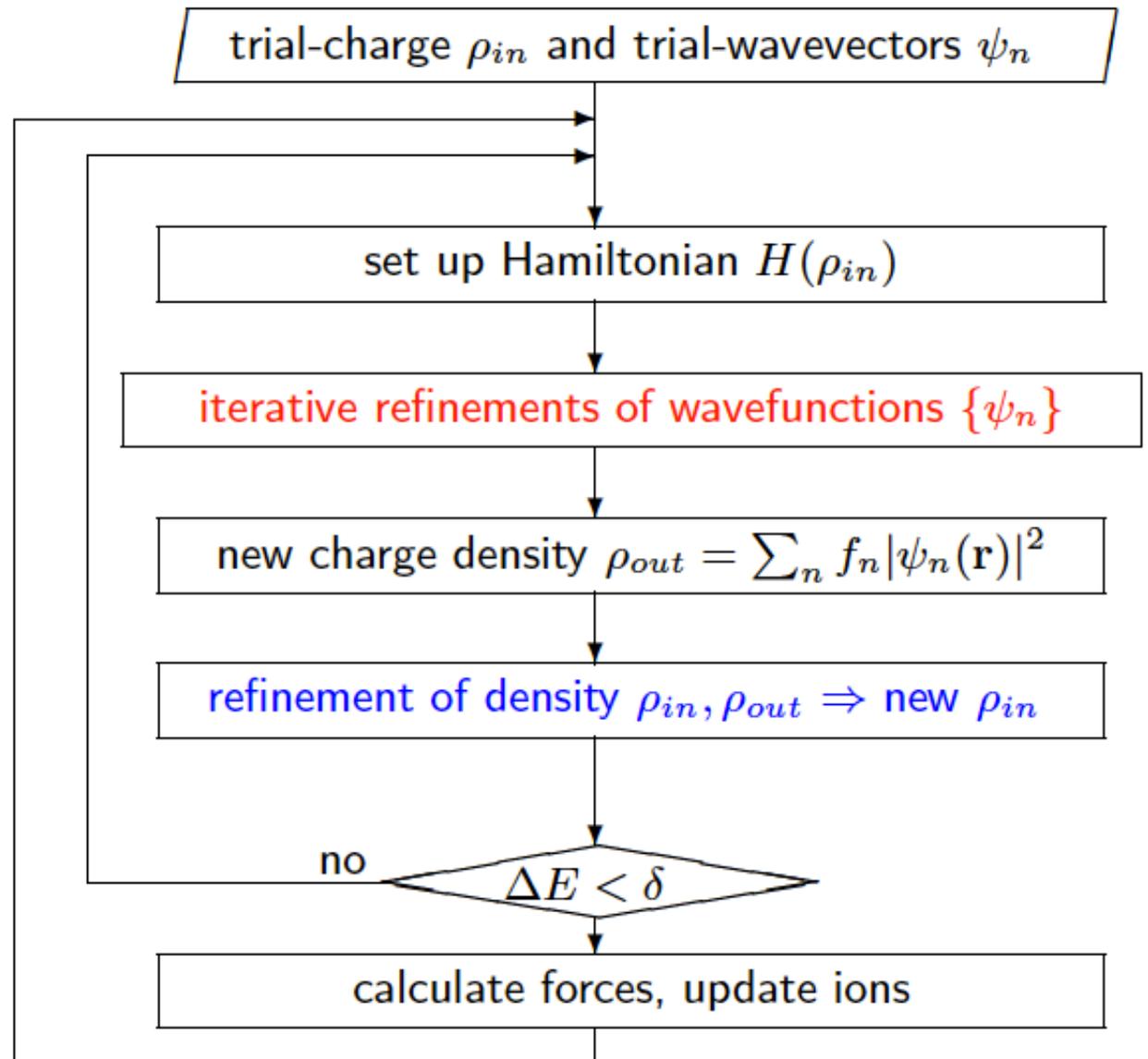
The charge density

The Self-Consistency-Cycle (cont.)

Two sub-problems:

- Optimization of $\{\psi_n\}$
  Iterative Diagonalization
  e.g. RMM-DIIS or
  Blocked Davidson

- Construction of ρ_{in}
  Density Mixing
  e.g. Broyden mixer

The self-Consistency-Cycle

A naïve algorithm: express the Hamilton matrix in a plane wave basis and diagonalize it:

$$\mathbf{H} = \langle \mathbf{G} | \hat{H}[\rho] | \mathbf{G}' \rangle \rightarrow \text{diagonalize } \mathbf{H} \rightarrow \{\psi_i, \epsilon_i\} \ \ i = 1, .., N_{\text{FFT}}$$

Self-consistency-cycle:

$$\rho_0 \rightarrow \mathbf{H}_0 \rightarrow \rho' \rightarrow \rho_1 = f(\rho_0, \rho') \rightarrow \mathbf{H}_1 \rightarrow ...$$

Iterate until: $\rho = \rho'$

BUT: we do not need N_{FFT} eigenvectors of the Hamiltonian (at a cost of $O(N_{\text{FFT}}{}^3)$).
Actually we only the N_b lowest eigenstates of $\mathbf{H}$, where N_b is of the order
of the number of electrons per unit cell ($N_b << N_{\text{FFT}}$).

Solution: use iterative matrix diagonalization techniques to find the N_b lowest
Eigenvector of the Hamiltonian: RMM-DIIS, blocked-Davidson, etc.

Key ingredients: Subspace diagonalization and the Residual

- Rayleigh-Ritz: diagonalization of the N_b x N_b subspace

$$\sum_m \bar{H}_{nm} B_{mk} = \sum_m \epsilon_k^{\text{app}} \bar{S}_{nm} B_{mk}$$

with

$$\bar{H}_{nm} = \langle \psi_n | \hat{H} | \psi_m \rangle \qquad \bar{S}_{nm} = \langle \psi_n | \hat{S} | \psi_m \rangle$$

yields N_b eigenvectors $|\bar{\psi}_k\rangle = \sum_m B_{mk} |\psi_m\rangle$ with eigenvalues ϵ_{app}.

These eigenstates are the best approximation to the exact N_b lowest eigenstates of **H** within the subspace spanned by the current orbitals.

- The Residual:

$$|R(\psi_n)\rangle = (\hat{H} - \epsilon_{\text{app}} \hat{S}) |\psi_n\rangle \qquad \epsilon_{\text{app}} = \frac{\langle \psi_n | \hat{H} | \psi_n \rangle}{\langle \psi_n | \hat{S} | \psi_n \rangle}$$

(its norm is measure for the error in the eigenvector)

Blocked-Davidson

- Take a subset of all bands: $\{\psi_n | n = 1,..,N\} \Rightarrow \{\psi_k^1 | k = 1,..,n_1\}$

  - Extend this subset by adding the (preconditioned) residual vectors to the presently considered subspace:

    $$\{\psi_k^1/g_k^1 = \mathbf{K}(\mathbf{H} - \epsilon_{\text{app}}\mathbf{S})\psi_k^1 | k = 1,..,n_1\}$$

  - Rayleigh-Ritz optimization ("sub-space rotation") in the $2n_1$ dimensional subspace to determine the n_1 lowest eigenvectors:

    $$\text{diag}\{\psi_k^1/g_k^1\} \longrightarrow \{\psi_k^2 | k = 1,..,n_1\}$$

  - Extend subspace with the residuals of $\{\psi_k^2\}$

    $$\{\psi_k^1/g_k^1/g_k^2 = \mathbf{K}(\mathbf{H} - \epsilon_{\text{app}}\mathbf{S})\psi_k^2 | k = 1,..,n_1\}$$

  - Rayleigh-Ritz optimization $\Rightarrow \{\psi_k^3 | k = 1,..,n_1\}$

  - Etc …

  - The optimized set replaces the original subset:

    $$\{\psi_k^m | k = 1,..,n_1\} \longrightarrow \{\psi_n | n = 1,..,n_1\}$$

- Continue with next subset: $\{\psi_k^1 | k = n_1 + 1,..,n_2\}$, etc, …

After treating all bands: Rayleigh-Ritz optimization of $\{\psi_n | n = 1,..,N\}$

The action of the Hamiltonian

The action $\mathbf{H}|\psi_{n\mathbf{k}}\rangle$

$$\left(-\frac{1}{2}\Delta + V(\mathbf{r})\right)\psi_{n\mathbf{k}}(\mathbf{r})$$

Using the convention

$$\langle \mathbf{r}|\mathbf{G}+\mathbf{k}\rangle = \frac{1}{\Omega^{1/2}}e^{i(\mathbf{G}+\mathbf{k})\mathbf{r}} \rightarrow \langle \mathbf{G}+\mathbf{k}|\psi_{n\mathbf{k}}\rangle = C_{\mathbf{G}n\mathbf{k}}$$

- Kinetic energy:

$$\langle \mathbf{G}+\mathbf{k}|-\frac{1}{2}\Delta|\psi_{n\mathbf{k}}\rangle = \frac{1}{2}|\mathbf{G}+\mathbf{k}|^2 C_{\mathbf{G}n\mathbf{k}} \qquad N_{\mathrm{NPLW}}$$

- Local potential: $V = V_{\mathrm{H}}[\rho] + V_{xc}[\rho] + V_{\mathrm{ext}}$
  - Exchange-correlation: easily obtained in real space $V_{\mathrm{xc},\mathbf{r}} = V_{\mathrm{xc}}[\rho_{\mathbf{r}}]$
  - FFT to reciprocal space $\{V_{\mathrm{xc},\mathbf{r}}\} \rightarrow \{V_{\mathrm{xc},\mathbf{G}}\}$
  - Hartree potential: solve Poisson eq. in reciprocal space $V_{\mathrm{H},\mathbf{G}} = \dfrac{4\pi}{|\mathbf{G}|^2}\rho_{\mathbf{G}}$
  - Add all contributions $V_{\mathbf{G}} = V_{\mathrm{H},\mathbf{G}} + V_{\mathrm{xc},\mathbf{G}} + V_{\mathrm{ext},\mathbf{G}}$
  - FFT to real space $\{V_{\mathbf{G}}\} \rightarrow \{V_{\mathbf{r}}\}$

  The action

$$\langle \mathbf{G}+\mathbf{k}|V|\psi_{n\mathbf{k}}\rangle = \frac{1}{N_{\mathrm{FFT}}}\sum_{\mathbf{r}} V_{\mathbf{r}} C_{\mathbf{r}n\mathbf{k}} e^{-i\mathbf{G}\mathbf{r}} \qquad N_{\mathrm{FFT}}\log N_{\mathrm{FFT}}$$

Solving the KS equations

$$\left(-\frac{1}{2}\Delta + V_{\text{ext}}(\mathbf{r}) + V_{\text{H}}(\mathbf{r}) + V_{\text{xc}}(\mathbf{r})\right)\psi_{n\mathbf{k}}(\mathbf{r}) = \epsilon_{n\mathbf{k}}\psi_{n\mathbf{k}}(\mathbf{r})$$

- FFTs extensively used to evaluate $\mathbf{H}|\psi_n\rangle$

- We actually use a mixed basis set (Projector-Augmented-Waves):

$$|\psi_n\rangle = |\widetilde{\psi}_n\rangle + \sum_i (|\phi_i\rangle - |\widetilde{\phi}_i\rangle)\langle\widetilde{p}_i|\widetilde{\psi}_n\rangle$$

  which involves projection of the pseudo-wave functions on local projection operators (DGEMM).

- One needs to keep the solutions (bands) at each k-point orthonormal: essentially done by a Choleski decomposition (LU) of the overlap matrix, followed by an inversion of U (LAPACK/scaLAPACK) and a transformation between the wave function (ZGEMM).

- Diagonalization of the Hamiltonian in the subspace of the current wave functions (LAPACK/scaLAPACK or ELPA), followed by a unitary transformation between the wave function (ZGEMM).
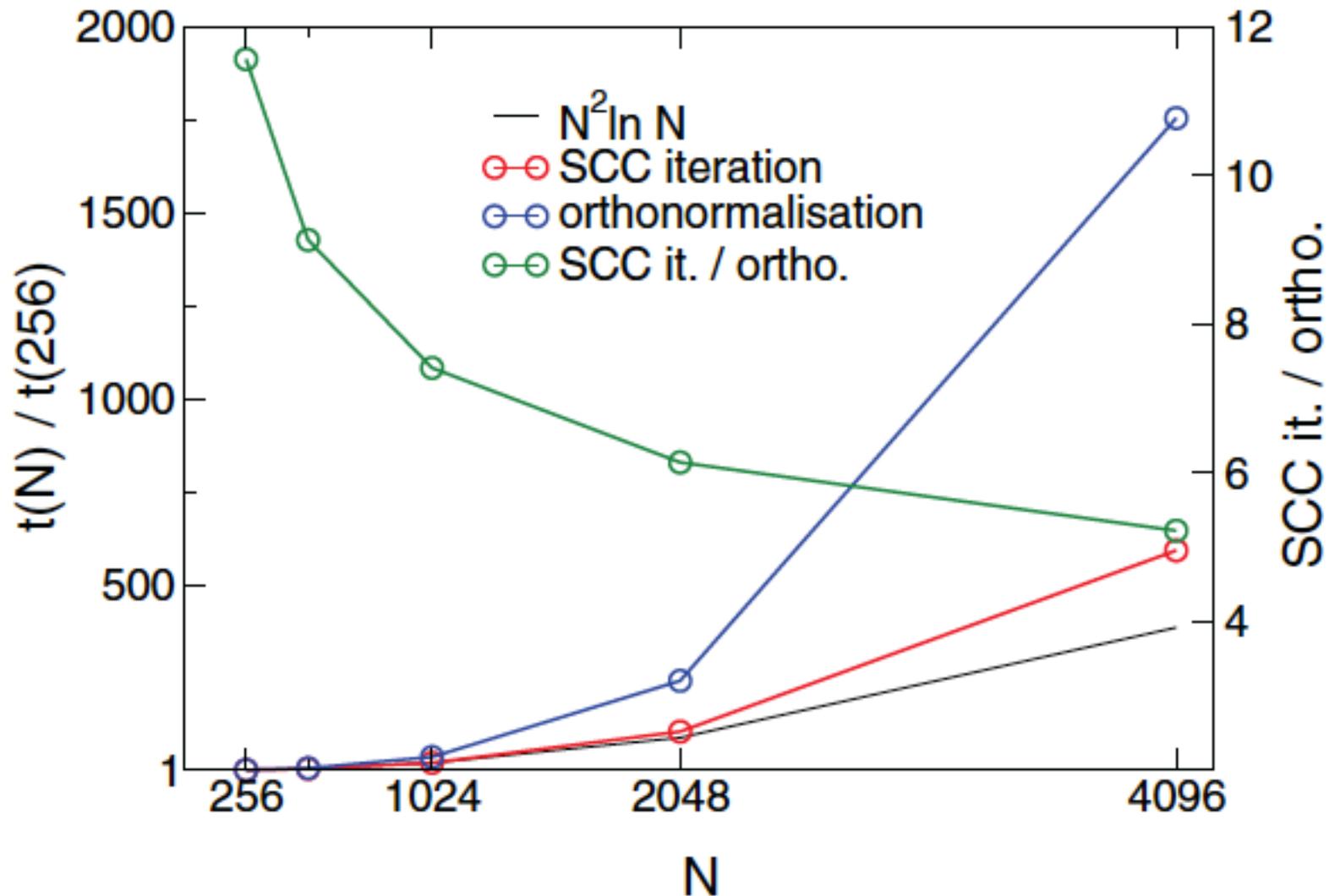
A typical workload

- Action: $H|\tilde{\psi}_n\rangle$

$\mathrm{FFT}(\tilde{\psi}_n)$	$N \ln N$	$\forall\ n$	$N^2 \ln N$		fftw
$V(\mathbf{r})\tilde{\psi}_\mathbf{n}(\mathbf{r})$	N	$\forall\ n$	N^2		
$\langle \tilde{p}_i | \tilde{\psi}_n \rangle$	N const.	$\forall\ i, n$	N^3 N^2	(real sp.) LREAL= A	BLAS3 (DGEMM)

- Subspace rotation:

$H_{nm} = \langle \tilde{\psi}_n | H | \tilde{\psi}_m \rangle$	$\forall\ n, m$	N^3		BLAS3 (ZGEMM)
$\mathrm{diag}(H)$		N^3		(sca)LAPACK
$|\tilde{\psi}'_n\rangle = \sum_m U_{nm}|\tilde{\psi}_m\rangle$	$\forall\ n$	N^3		BLAS3 (ZGEMM)

Scaling with system size (N)

Self-Consistency Cycle (SCC): RMM-DIIS
running on 8 Intel X5550 quadcore procs. (total: 32 x 2.66 GHz cores)

Distribution of work and data
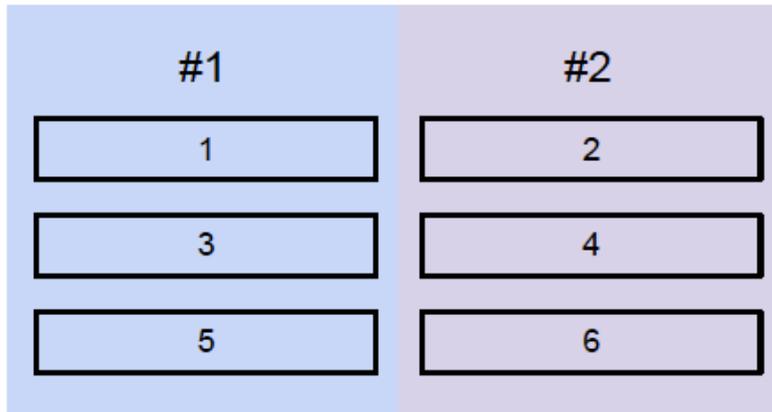
2 MPI-ranks, NCORE=1

#1	#2
1	2
3	4
5	6

Distribute work and data "over-orbitals"
- Default
- NCORE = 1
  (or equivalently: NPAR = #-of-MPI-ranks)
- KPAR = 1

The Kohn-Sham equation:
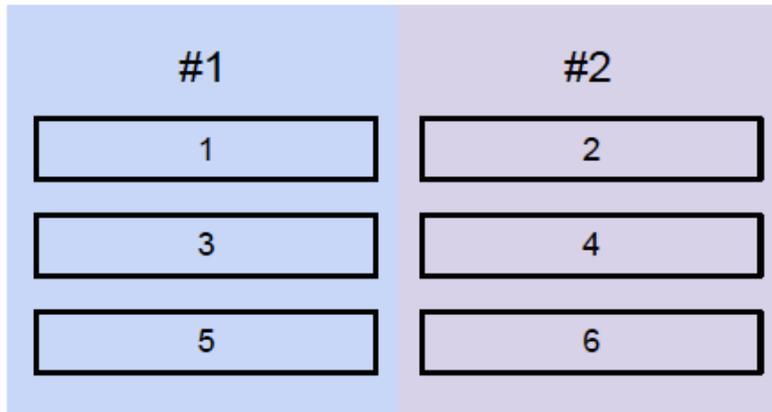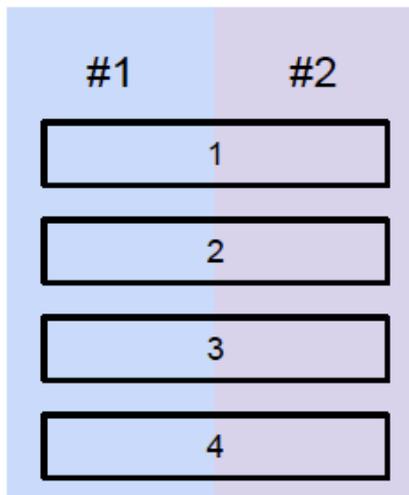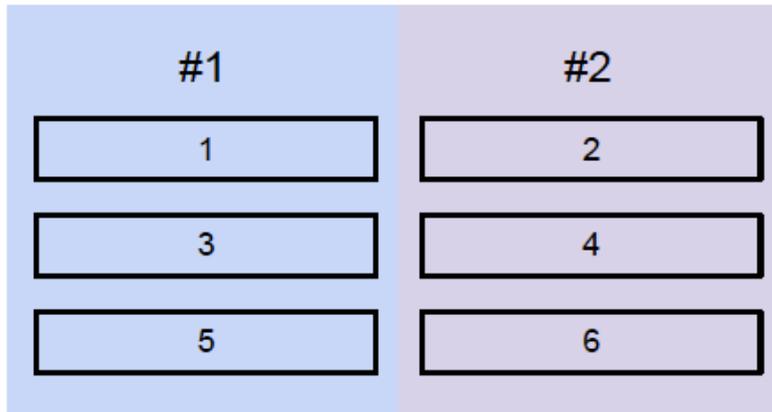
$$\left(-\frac{1}{2}\Delta + V_{\mathrm{ext}}(\mathbf{r}) + V_{\mathrm{H}}(\mathbf{r}) + V_{\mathrm{xc}}(\mathbf{r})\right)\psi_{n\mathbf{k}}(\mathbf{r}) = \epsilon_{n\mathbf{k}}\psi_{n\mathbf{k}}(\mathbf{r})$$

- Orbital index n

Distribution of work and data

2 MPI-ranks, NCORE=1

#1	#2
1	2
3	4
5	6

Distribute work and data "over-orbitals"
- Default
- NCORE = 1
  (or equivalently: NPAR = #-of-MPI-ranks)
- KPAR = 1

2 MPI-ranks, NCORE=2

#1	#2
1	
2	
3	
4	

Distribution work and data "over-plane-waves"
- NCORE = #-of-MPI-ranks
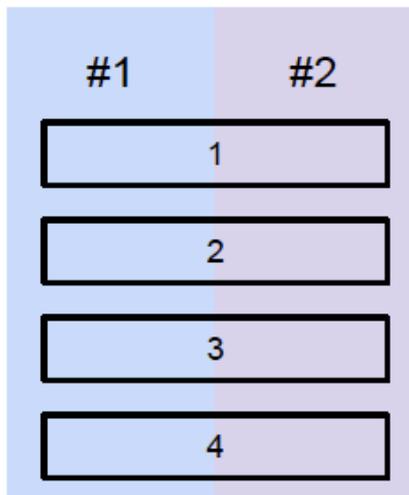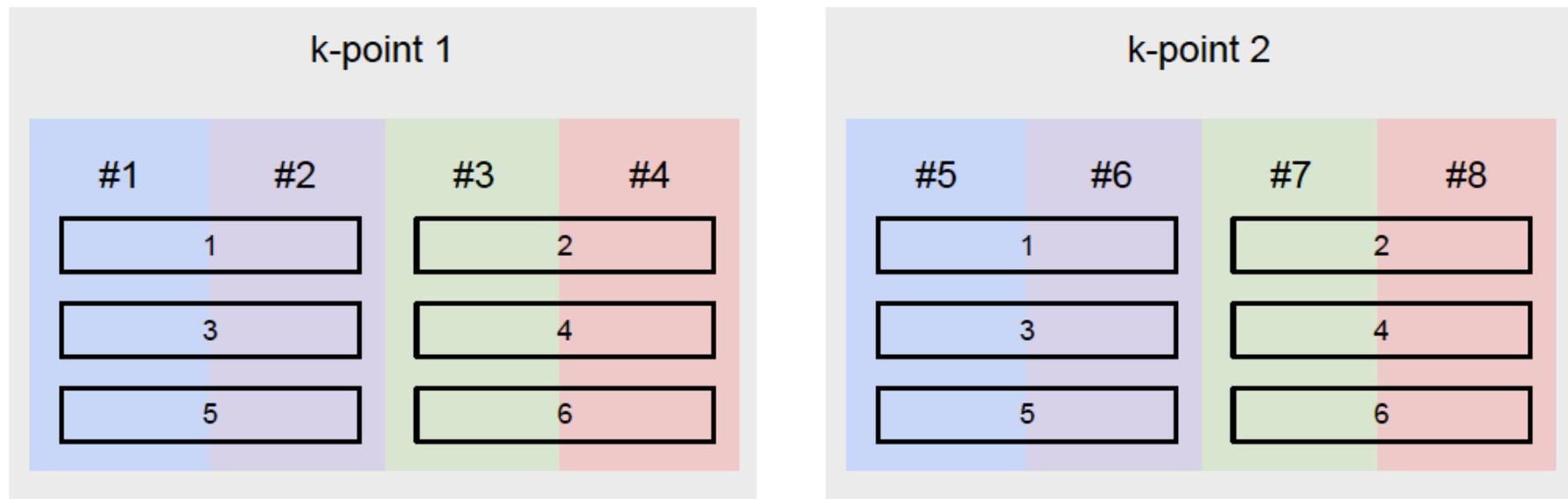 (or equivalently: NPAR = 1)
- KPAR = 1

Distribution of work and data

Additionally work may be distributed "over-**k**-points"

- KPAR = n ($n>1$)
- m = (#-of-MPI-ranks / n) must be an integer
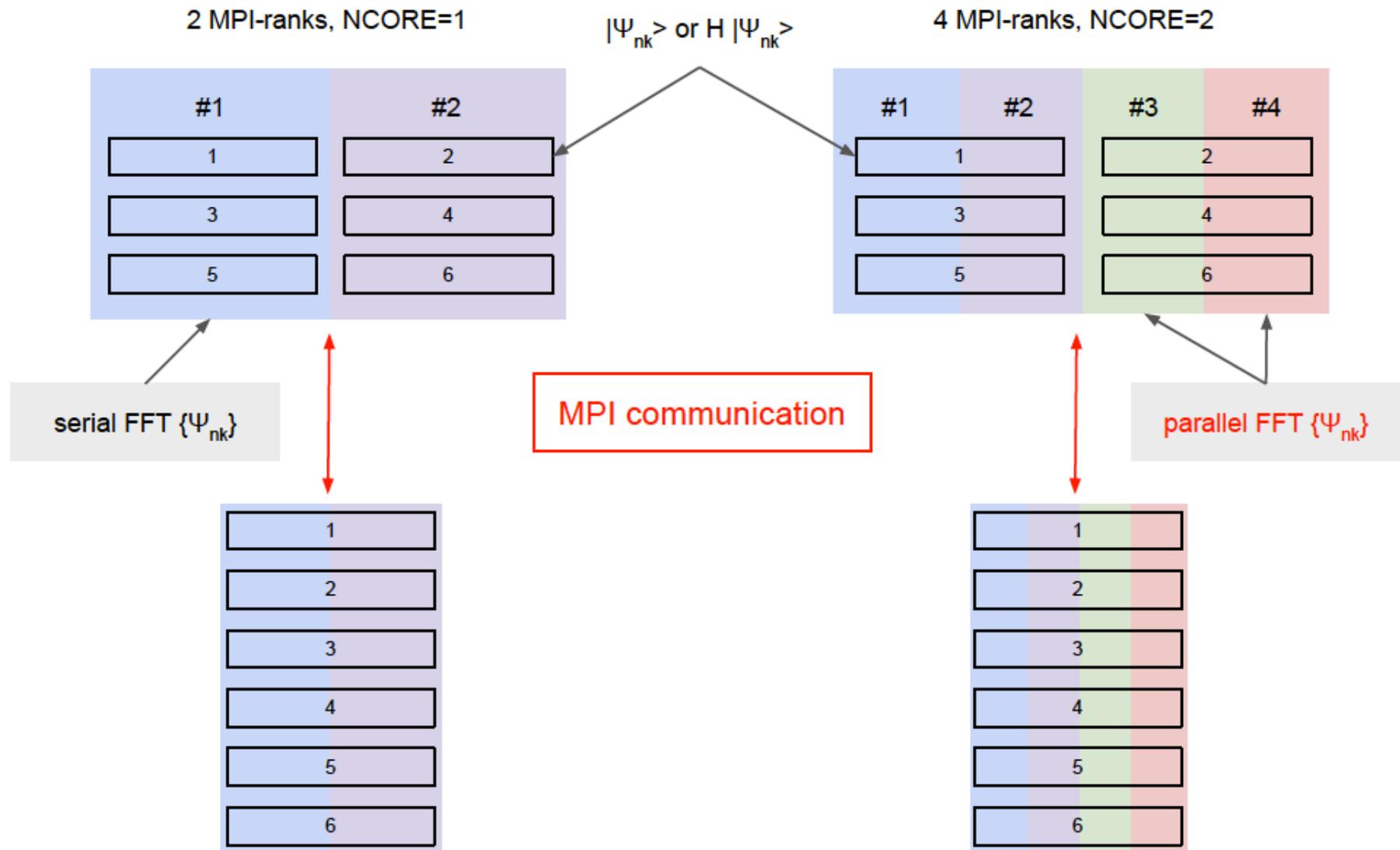- Work is distributed in a round-robin fashion over groups of m MPI-ranks
- Data is duplicated!

$$\left(-\frac{1}{2}\Delta + V_{\text{ext}}(\mathbf{r}) + V_{\text{H}}(\mathbf{r}) + V_{\text{xc}}(\mathbf{r})\right)\psi_{n\mathbf{k}}(\mathbf{r}) = \epsilon_{n\mathbf{k}}\psi_{n\mathbf{k}}(\mathbf{r})$$

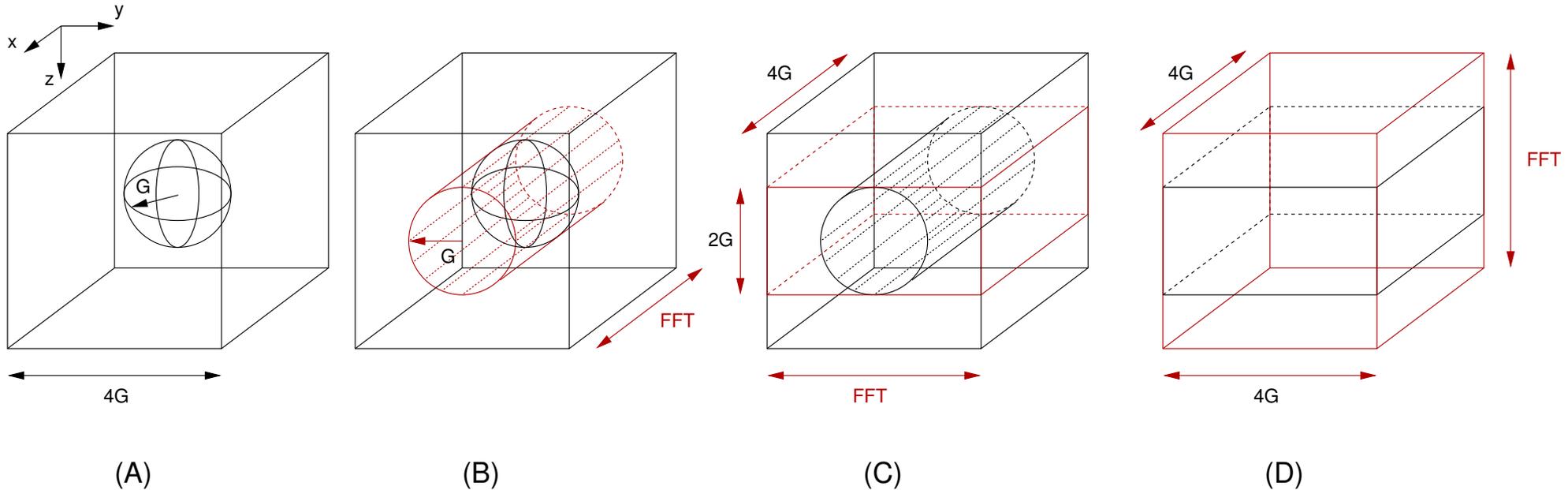- Orbital index n, k-point index **k**

8 MPI-ranks, KPAR=2, NCORE=2

All-2-All communication

2 MPI-ranks, NCORE=1

$|\Psi_{nk}\rangle$ or $H\,|\Psi_{nk}\rangle$

4 MPI-ranks, NCORE=2

serial FFT $\{\Psi_{nk}\}$

MPI communication

parallel FFT $\{\Psi_{nk}\}$

$$C_{nm} = \langle\Psi_{mk}|\,H\,|\Psi_{nk}\rangle$$

Each MPI-rank contracts over its subset of G-vector
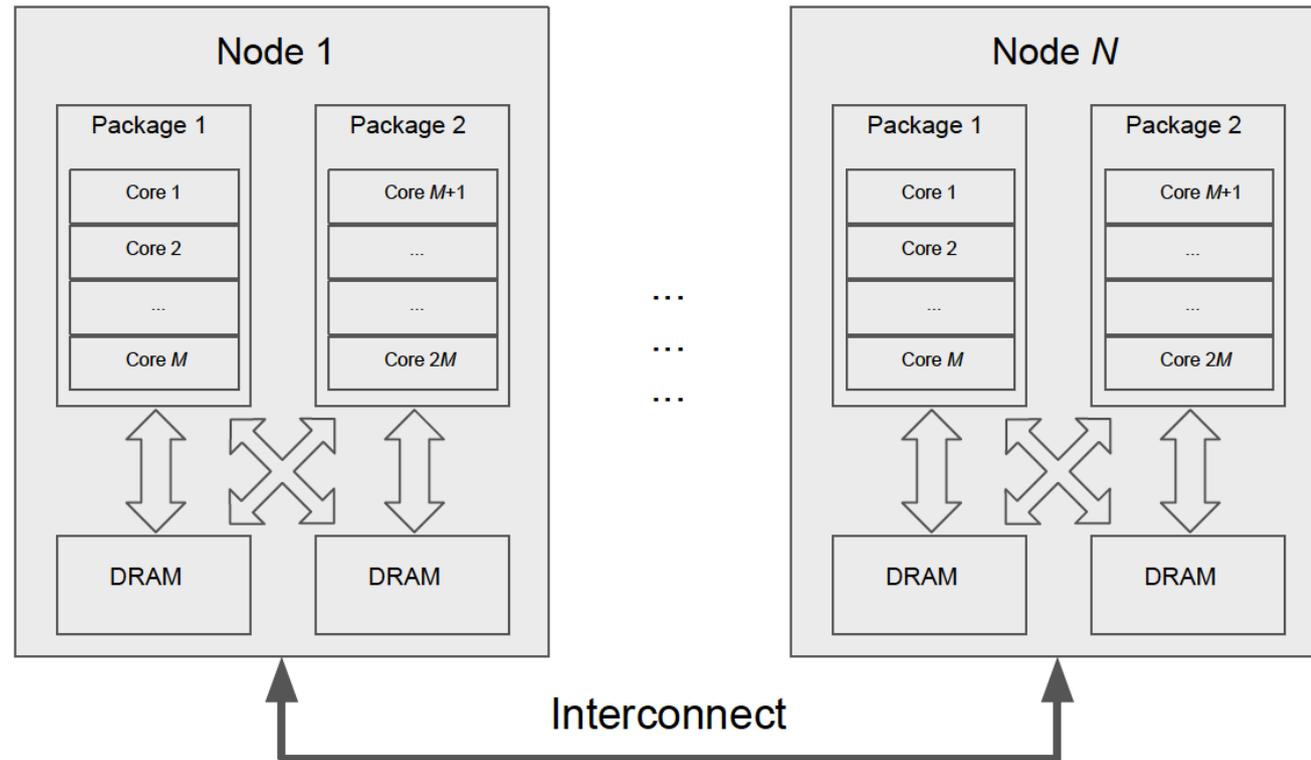followed by a global sum of C_{nm} over all MPI-ranks

Parallel FFT: Ball-2-Cube

(A) (B) (C) (D)

In-house parallel Ball-2-Cube FFT:

- Less 1D-FFTs (reduction: $\approx 1.76 \times$)
- BUT: communication from (B) → (C) and (C) → (D)
- For small to medium sized FFT grids a highly optimized 3D-FFT (e.g. from Intel's mkl) is equally fast

Hardware considerations

Typical configuration:

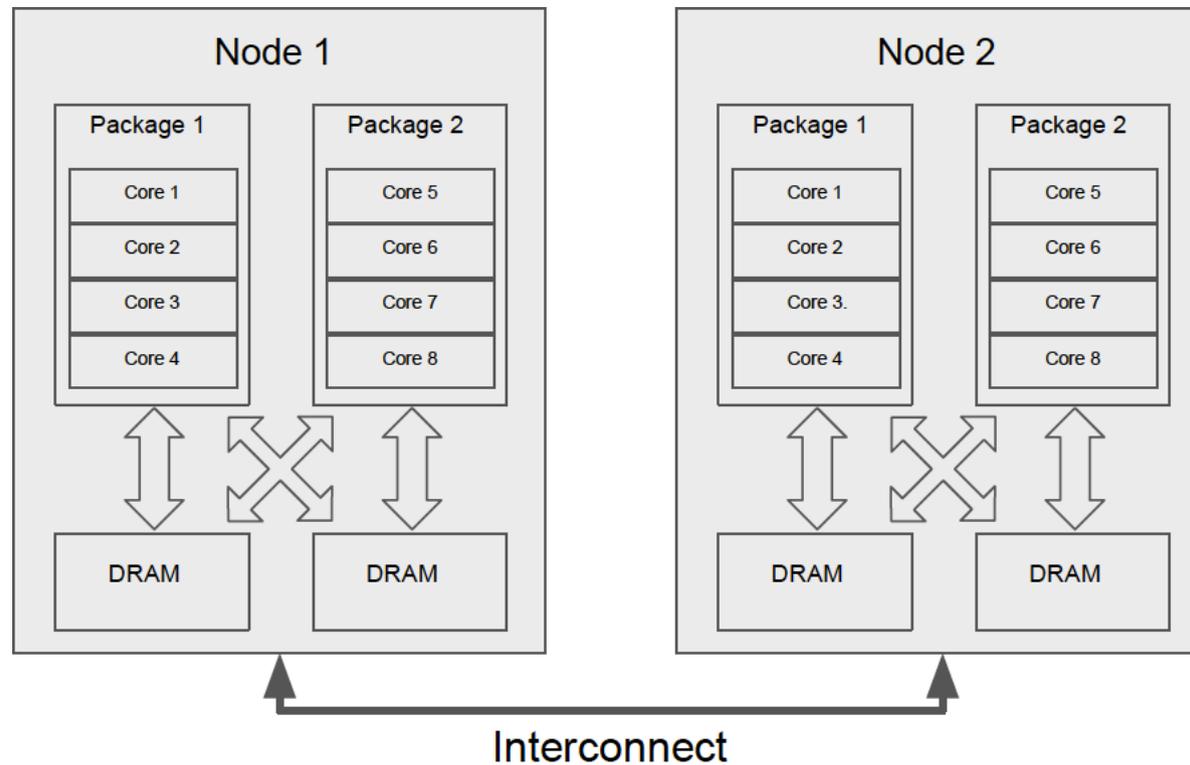- N interconnected nodes
- 2 packages/node
- M cores/package

Node 1

Package 1
Core 1
Core 2
...
Core M

Package 2
Core $M+1$
...
...
Core $2M$

DRAM

DRAM

...
...
...

Node N

Package 1
Core 1
Core 2
...
Core M

Package 2
Core $M+1$
...
...
Core $2M$

DRAM

DRAM

Interconnect

Distribution "over-plane-waves":
MPI-ranks that share an orbital should reside on the same node (better even on the same package).

- NCORE = $n \leq 2M$
- ($2M / n$) should be an integer
- Typically: $n = M$ or $n = M/2$

Distribution "over-**k**-points":

- KPAR = $n \leq$ #-of-k-points (NKPTS)
- (#-of-MPI-ranks / n) should be an integer
- If memory allows: KPAR = NKPTS

Default placement of MPI-ranks on the nodes/packages/cores depends on the particulars of the MPI implementation and its configuration!

For instance:
- *2* interconnected nodes
- 2 packages/node
- 4 cores/package

Default placement of MPI-ranks on the nodes/packages/cores depends on the particulars of the MPI implementation and its configuration!
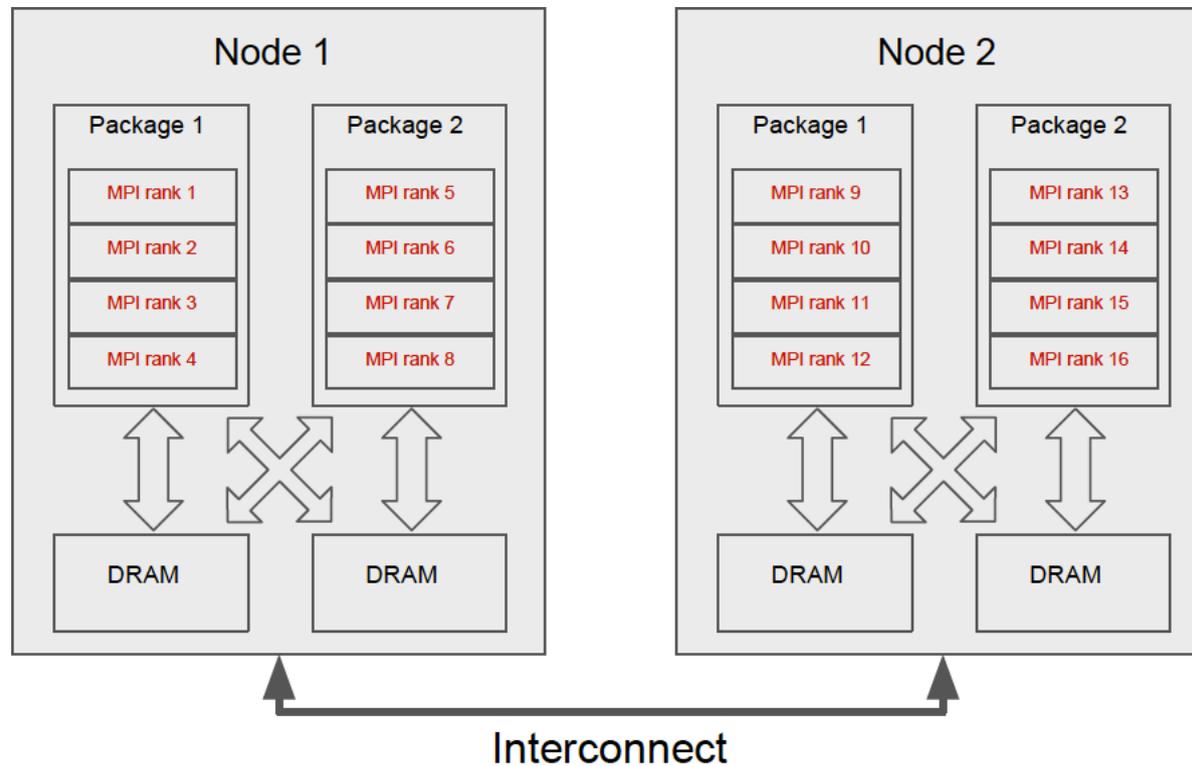
For instance:
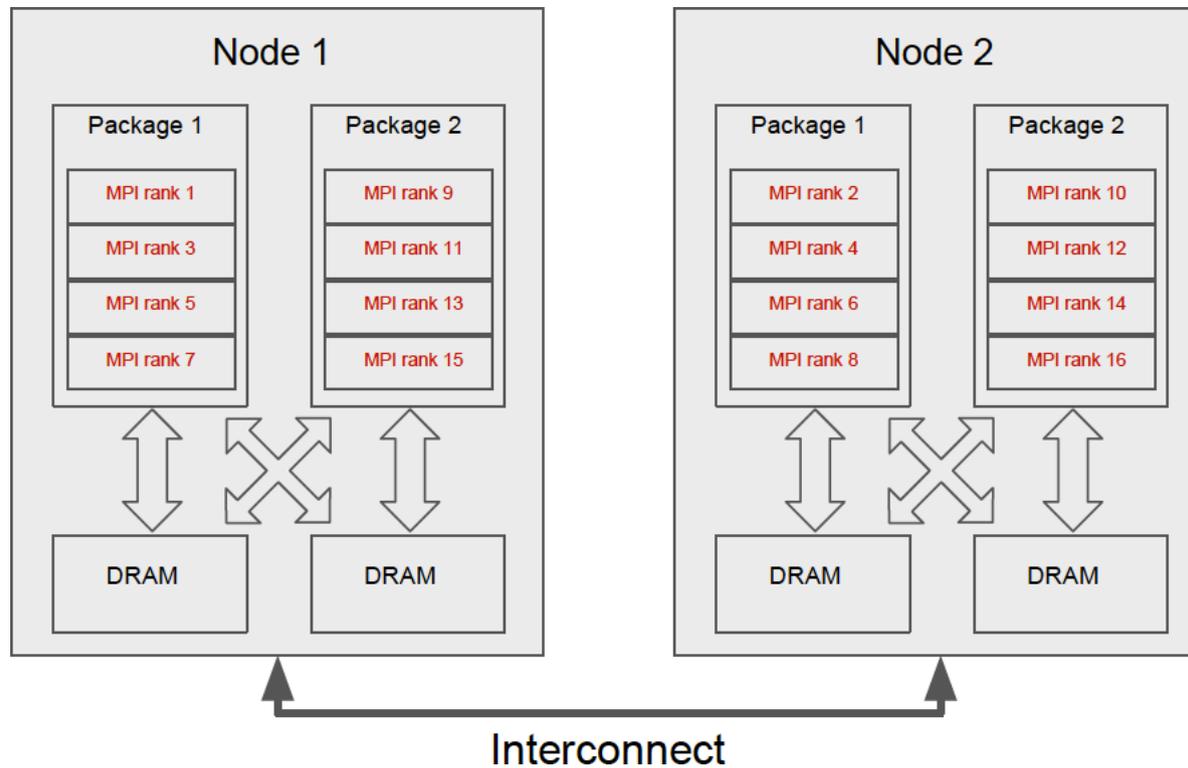- *2* interconnected nodes
- 2 packages/node
- 4 cores/package

Good: Place subsequent MPI-ranks close together, i.e., first on subsequent cores of the same package, then moving on the second package of the same node, before starting to fill the next node.
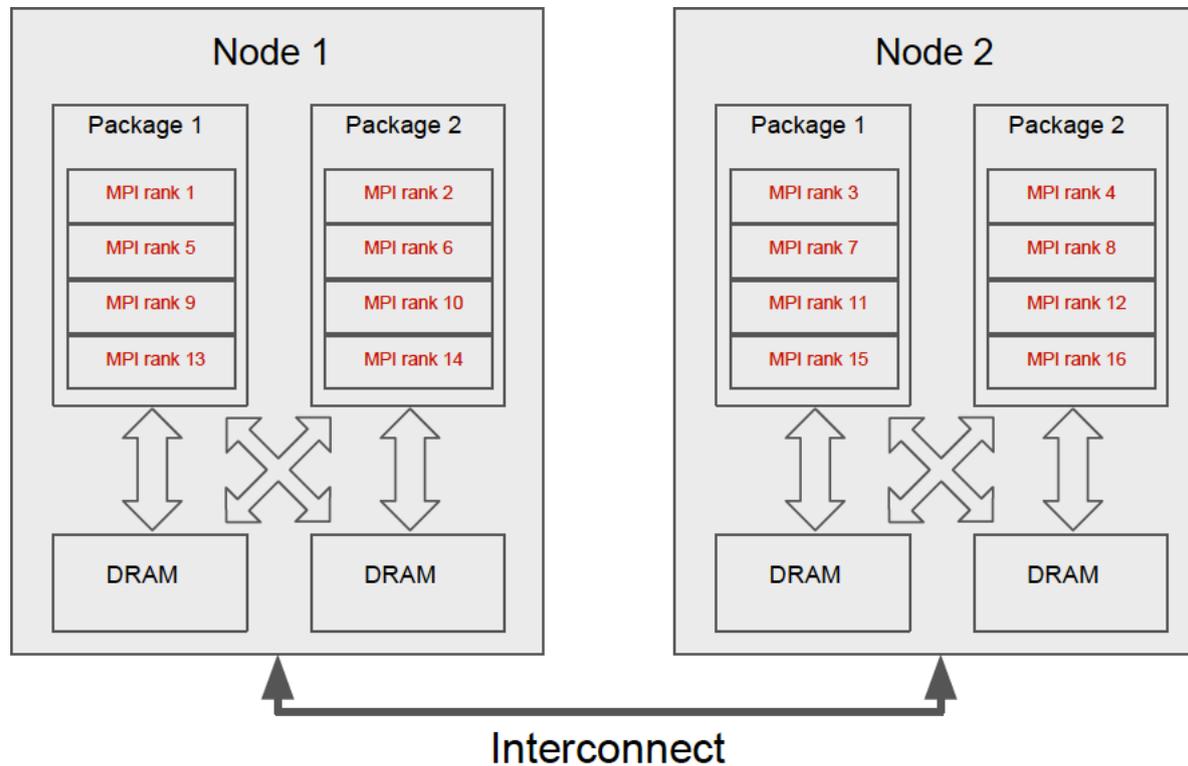
For instance:
- *2* interconnected nodes
- 2 packages/node
- 4 cores/package

Bad: Distribute subsequent MPI-ranks in a round-robin fashion over the packages.

For instance:
- *2* interconnected nodes
- 2 packages/node
- 4 cores/package

Worse: Distribute subsequent MPI-ranks in a round-robin fashion over the nodes.

NCORE, KPAR, and the #-of-MPI-ranks

- N interconnected nodes
- 2 packages/node
- M cores/package
- $2NM$ MPI-ranks

Distribution "over-plane-waves": MPI-ranks that share an orbital should reside on the same node (better even on the same package).

- NCORE = $n \leq 2M$
- ($2M / n$) should be an integer
- Typically: $n = M$ or $n = M/2$

Distribution "over-**k**-points":
- KPAR = $n \leq$ #-of-k-points (NKPTS)
- (#-of-MPI-ranks / n) should be an integer
- If memory allows: KPAR = NKPTS

Distribution "over-orbitals":
- The number of orbitals (NBANDS) is such that NBANDS / ($2NM$ / NCORE / KPAR) is an integer number
- ⇒ Increasing #-of-MPI-ranks may lead to an unnecessarily large NBANDS (*i.e.,* adding "empty" orbitals)

- Some algorithms converge faster when each MPI rank owns (part of) a few orbitals (*e.g.* blocked-Davidson)

- Generally speaking: having lots of MPI ranks and very little work/data per rank is never a good idea since (all-2-all) communication becomes unreasonably expensive.

Strong/Weak scaling (Si$_N$)
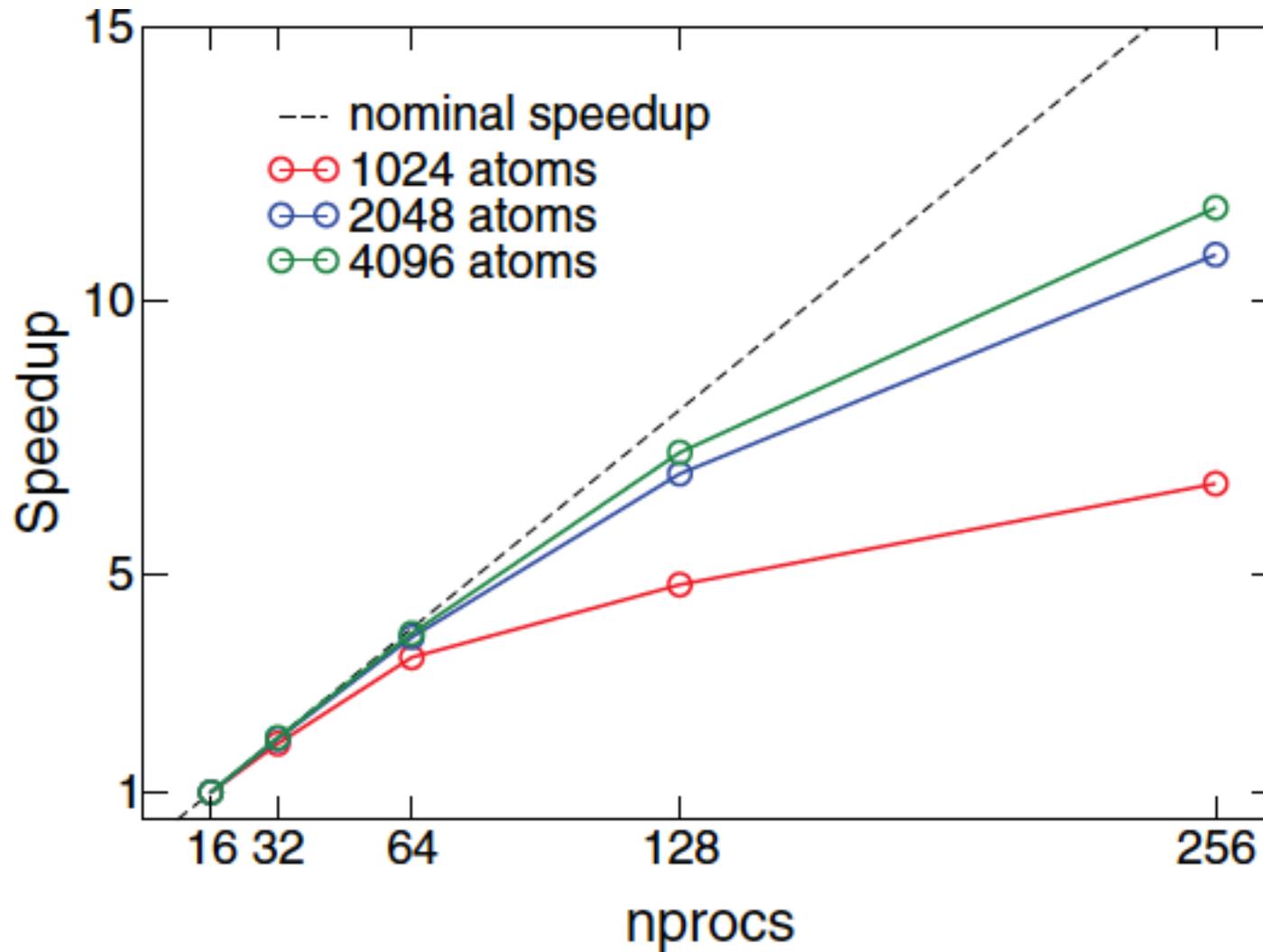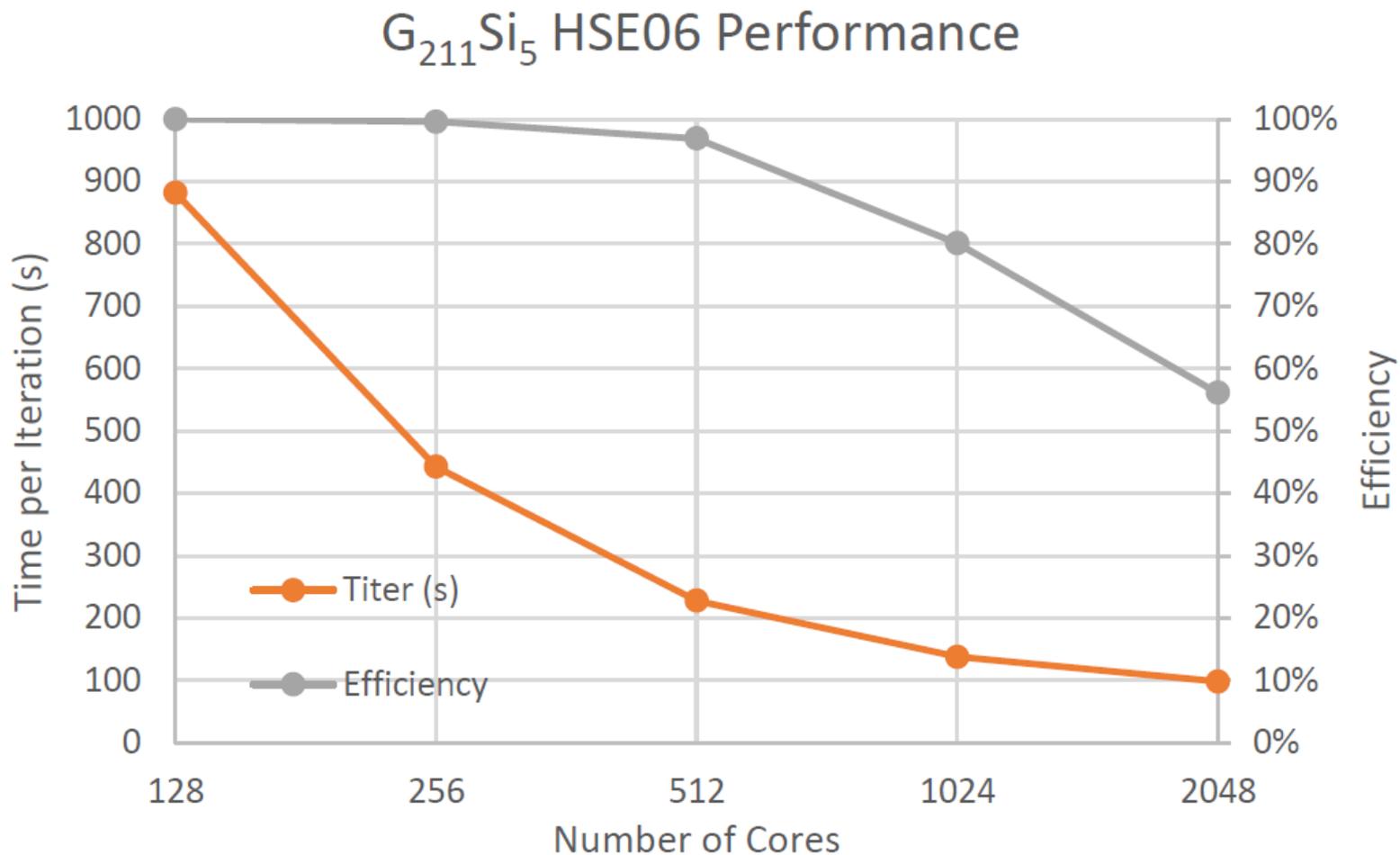
Scaling under MPI (on a Cray XC-40)

- 5713 e-
- 2 k-points
- RMM-DIIS
- PBE
- KPAR=2
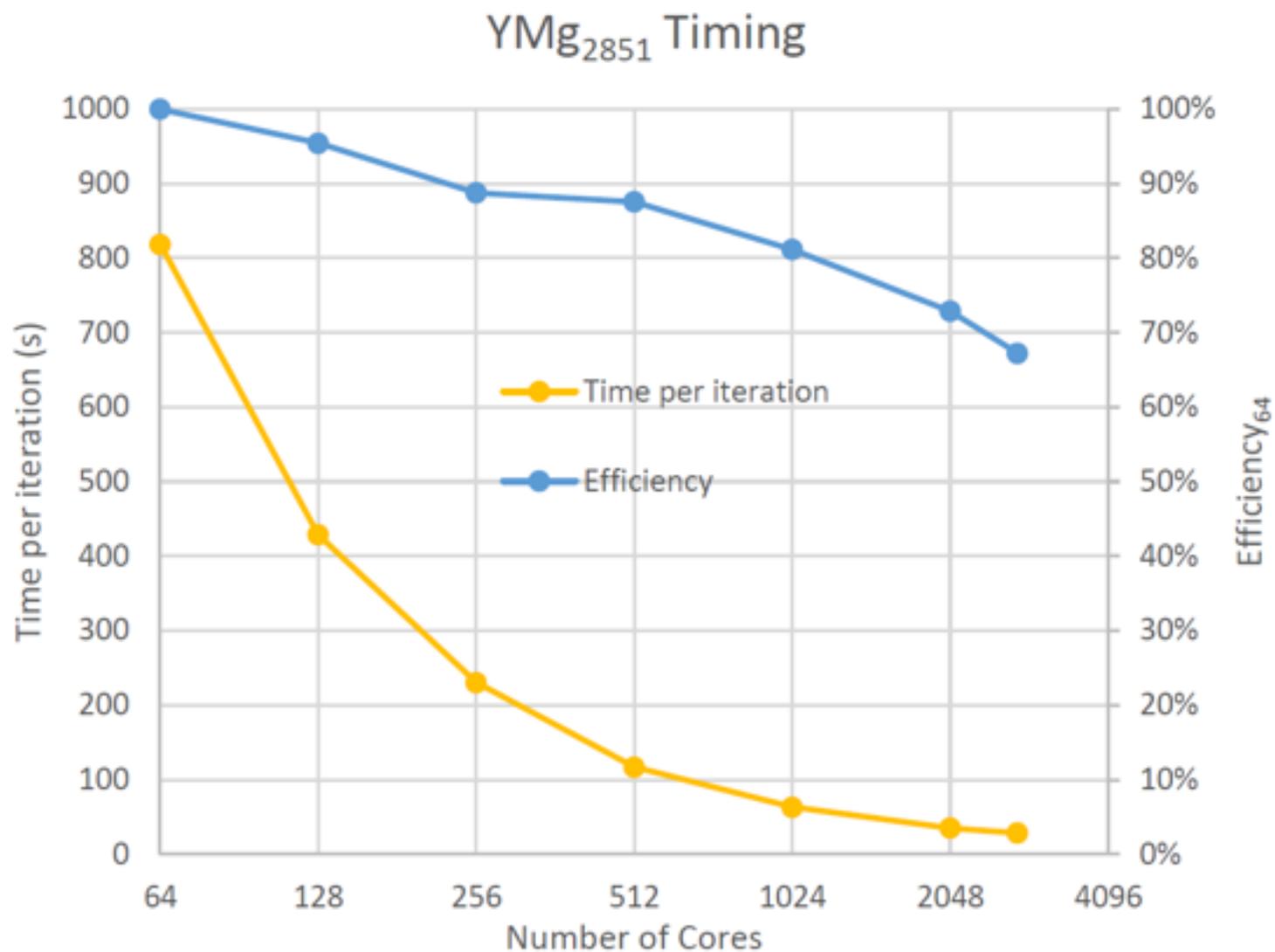
Courtesy of P. Saxe, Materials Design Inc. (and Cray).

Scaling under MPI (on a Cray XC-40)

- 864 e-
- 8 k-points
- HSE06
- Davidson
- KPAR=8

$G_{211}Si_5$ HSE06 Performance

Courtesy of P. Saxe, Materials Design Inc. (and Cray).

Large systems

- Use real-space PAW projectors (instead of reciprocal space projectors)

$$\langle \tilde{p}_i | \tilde{\psi}_n \rangle \qquad \begin{array}{c} N \\ \text{const.} \end{array} \qquad \forall\ i, n \qquad \begin{array}{c} N^3 \\ N^2 \end{array} \qquad \text{(real sp.)} \qquad \text{BLAS3 (DGEMM)}$$

$$\text{LREAL= A}$$

- If you can limit $\mathbf{k}$-point sampling to the Gamma ($\mathbf{\Gamma} \equiv \mathbf{k} = \mathbf{0}$) point: use the "gamma-only" version of VASP

  In the "gamma-only" version of VASP the orbitals are stored as real quantities in real-space:
  - real-2-complex FFTs
  - DGEMMs instead of ZGEMMs

Compilation

- Fast FFTs: the FFTs from Intel's mkl-library seem to be unbeatably fast …
- scaLAPACK for large systems
- A compiler that effectively generates AVX2 instructions and libraries (*e.g.* BLAS) that are optimized for AVX2 (up to 20% performance gain)

The End

Thank you!