Large-scale electronic structure methods

- Introduction
- O(N) Krylov subspace method
- Applications
- Numerically exact low-order scaling method
- Outlook

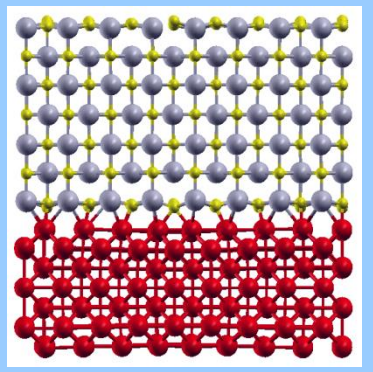
Taisuke Ozaki (ISSP, Univ. of Tokyo)

The Winter School on DFT: Theories and Practical Aspects, Dec. 19-23, CAS.

Towards first-principle studies for industry

10^3 – 10^6 atom

Steel



DFT calculations of thousands atoms is still a grand challenge.

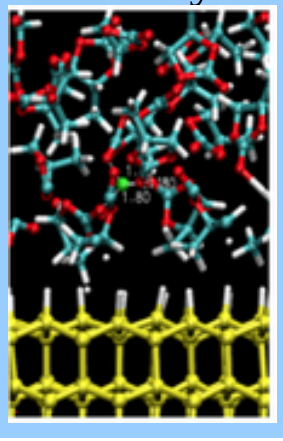
$O(N^3)$ → **Low-order**

10^2 atom

Many applications done.
There are many successes even for material design.

DNA

Battery



Time scale

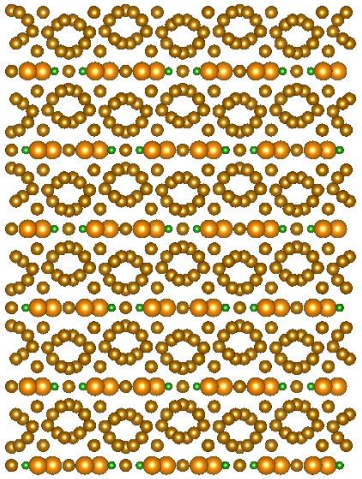
System size

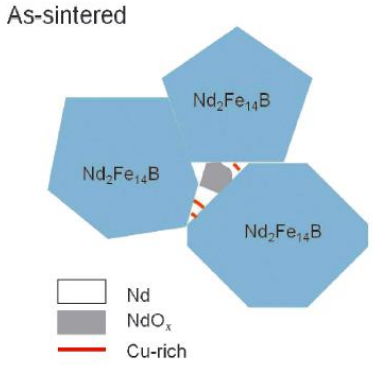
Materials properties

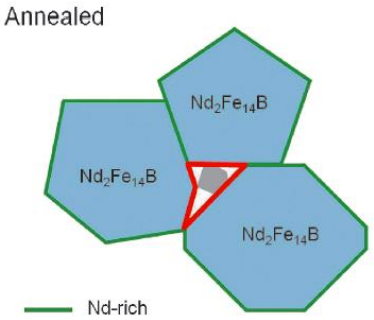
- Materials properties of actual materials are determined by intrinsic properties and secondary properties arising from inhomogeneous structures such as grain size, grain boundary, impurity, and precipitation.
- In use of actual materials, the materials properties can be maximized by carefully designing the crystal structure and higher order of structures .

http://ev.nissan.co.jp/LEAF/PERFORMANCE/

e.g., the coercivity of a permanent magnet of Nd-Fe-B is determined by crystal structure, grain size, and grain boundary.







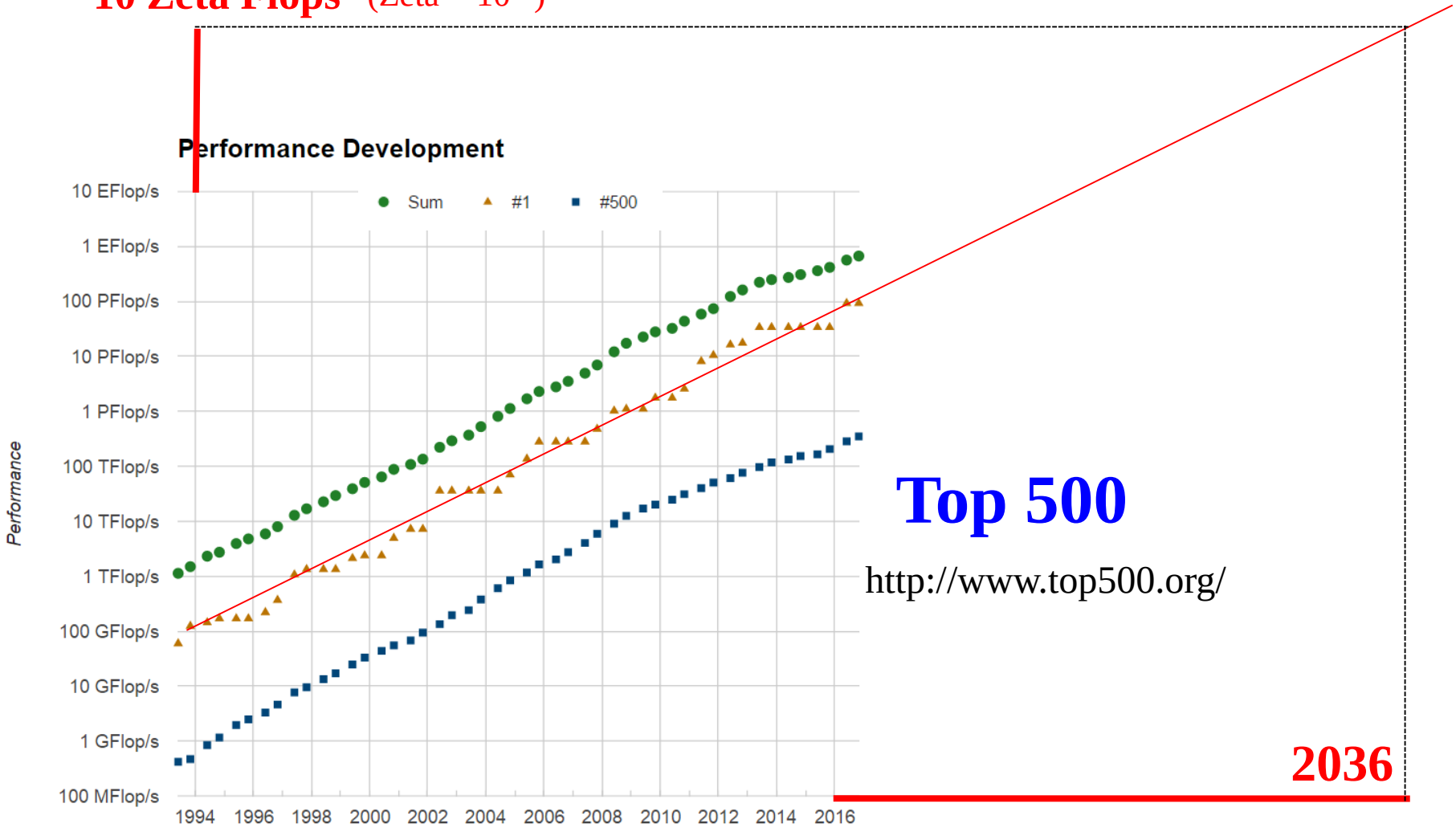
神威·太湖之光：125 Peta flops machine

Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway

NRCPC Cores: 10,649,600
Rmax: 93,014,593.9 (GFLOP/sec.)
Pmax: 125,435,904 (GFLOPS/sec.)

According to Moore's law…

10 Zeta Flops (Zeta = 10^{21})



How large systems can be treated by Exa machines?

The performance increase is only 100 times.

K-computer → Exa machine

10 PFLOPS → 1000 PFLOPS

Computational Scaling $O(N^p)$	Computable size
7	1.7
6	2.2
5	2.5
4	3.2
3	4.6 ← DFT
2	10
1	100

The applicability of the $O(N^3)$ DFT method is extended to only 5 times larger systems.

Linear scaling methods

Two routes towards O(N) DFT

Conventional representation

$$E_{\mathrm{tot}}(\{\Psi\},\rho)$$

Density matrix representation

$$E_{\mathrm{tot}}(n,\rho)$$

Wannier function representation

$$E_{\mathrm{tot}}(\{\phi\},\rho)$$

ψ: KS orbital
ρ: density
φ: Wannier function
n: density matrix

Density functionals as a functional of ρ

Density functionals can be rewritten by the first order reduced density matrix: ρ

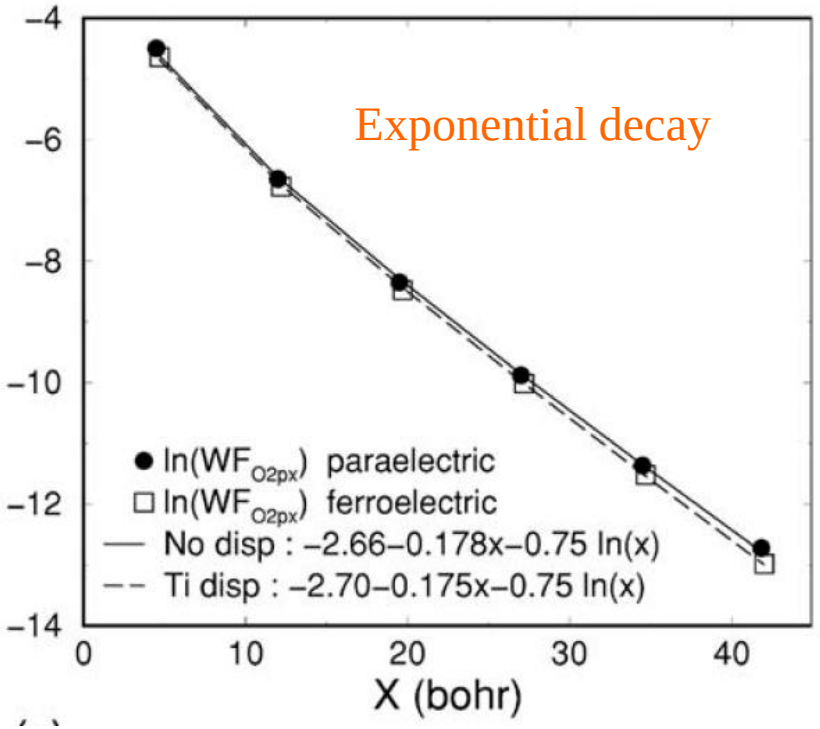
$$E_{\rm tot}[n,\rho] = {\rm Tr}(\rho H_{\rm kin}) + \int d\mathbf{r} n(\mathbf{r}) v_{\rm ext}(\mathbf{r}) + \int\int d\mathbf{r} d\mathbf{r}' \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + E_{\rm xc}[n]$$

where the electron density is given by ρ

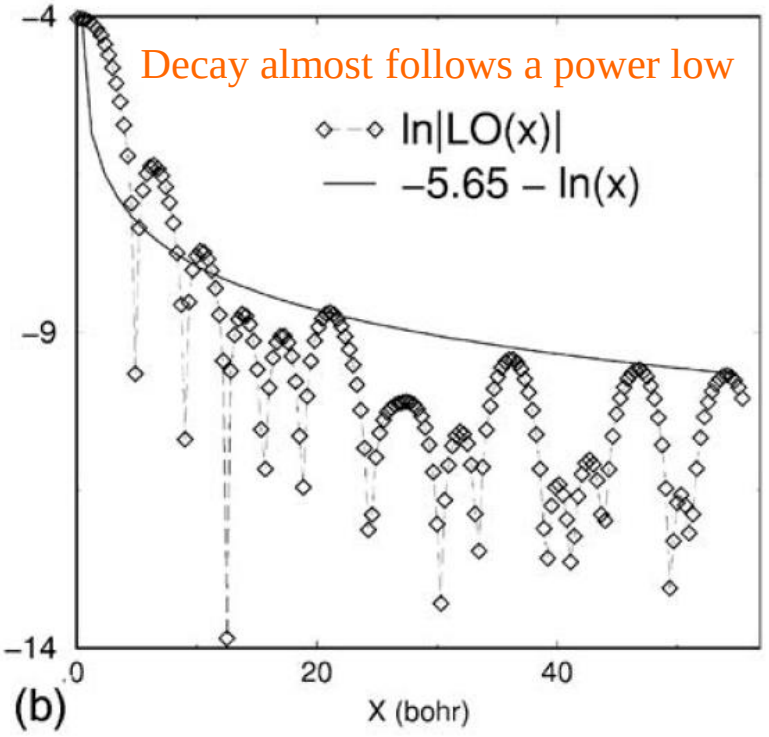
$$n(\mathbf{r}) = \sum_{i,j} \rho_{ij} \chi_j(\mathbf{r}) \chi_i(\mathbf{r})$$

Locality of Wannier functions

O-2px in $PbTiO_3$

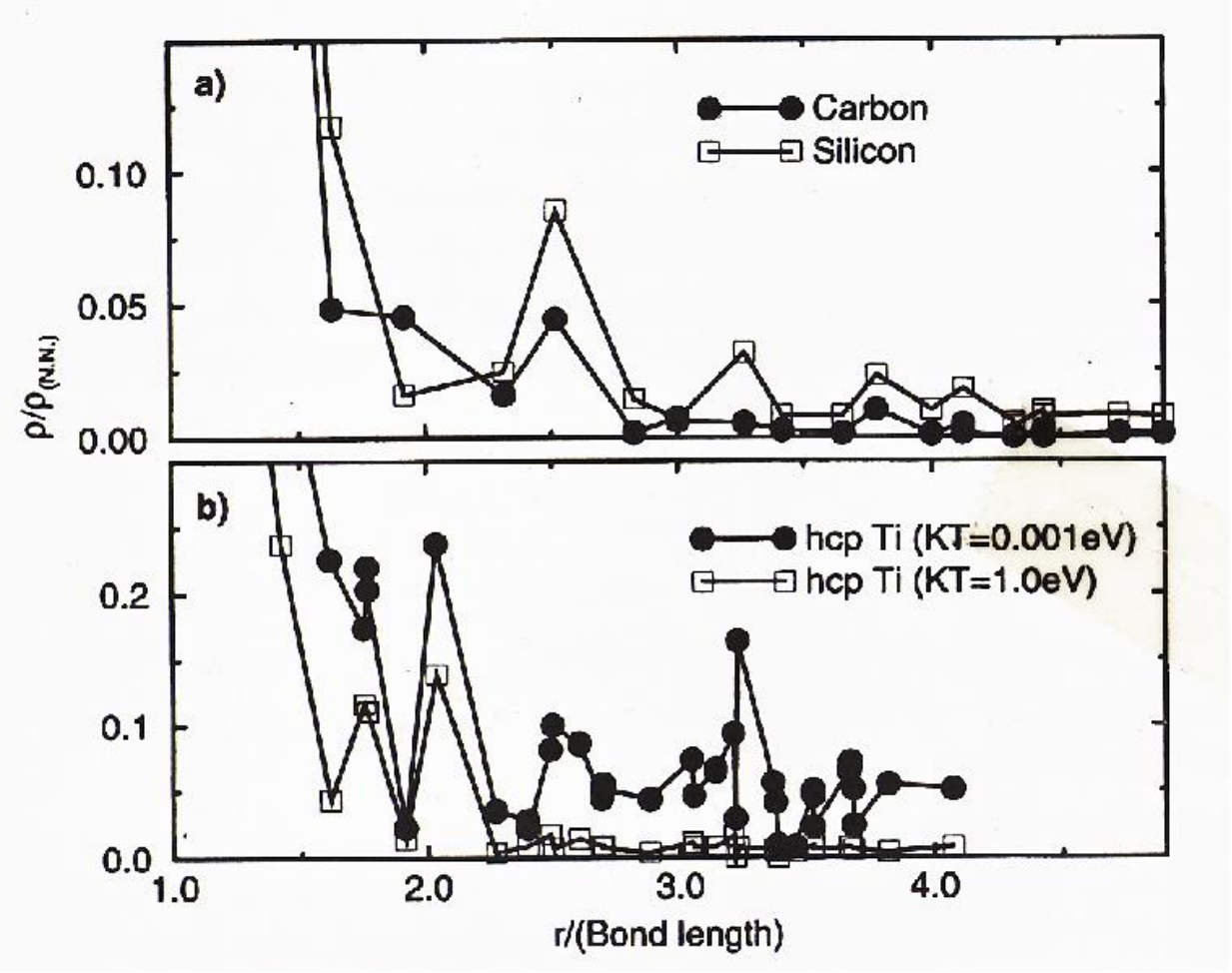


An orbital in Aluminum



J.Battacharjee and U.W.Waghmare, PRB 73, 121102 (2006)

Locality of density matrix



Finite gap systems
exponential decay

Metals
T=0 power law decay
0<T exponential decay

D.R.Bowler et al.,
Modell.Siml.Mater.Sci.Eng.5, 199 (1997)

Various linear scaling methods

Wannier functions (WF)
Density matrix (DM)
} × {
Variational (V)
Perturbative (P)

At least four kinds of linear-scaling methods can be considered as follows:

WF+V	**WF+P**	**DM+V**	**DM+P**
Orbital minimization by Galli, Parrinello, and Ordejon	Hoshi Mostofi	Density matrix by Li and Daw	**Krylov subspace** Divide-conquer Recursion Fermi operator

O(N) DFT codes

OpenMX: (Krylov) Ozaki (U. of Tokyo) et al.

Conquest: (DM) Bowler(London), Gillan(London), Miyazaki (NIMS)

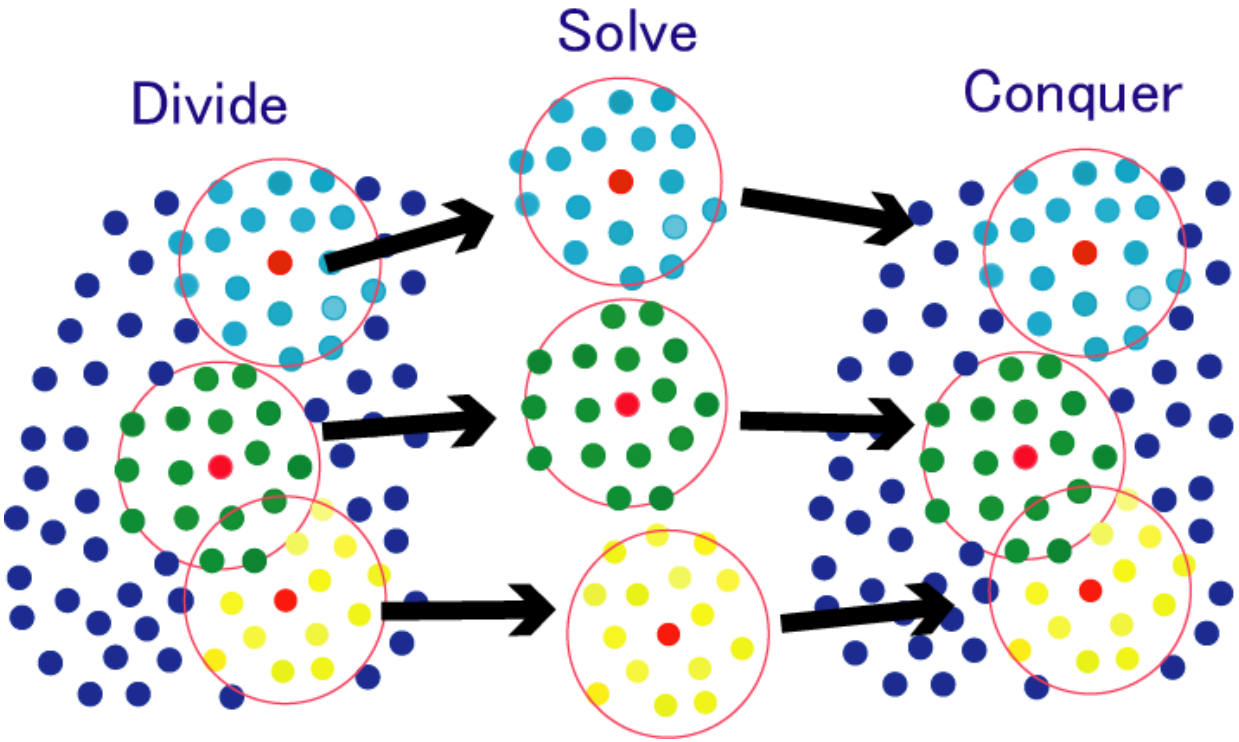
Siesta: (OM) Ordejon et al.(Spain)

ONETEP: (DM) Hayne et al.(Imperial)

FEMTECK: (OM) Tsuchida (AIST)

FreeON: (DM) Challacombe et al.(Minnesota)

Basic idea behind the O(N) method



Assumption

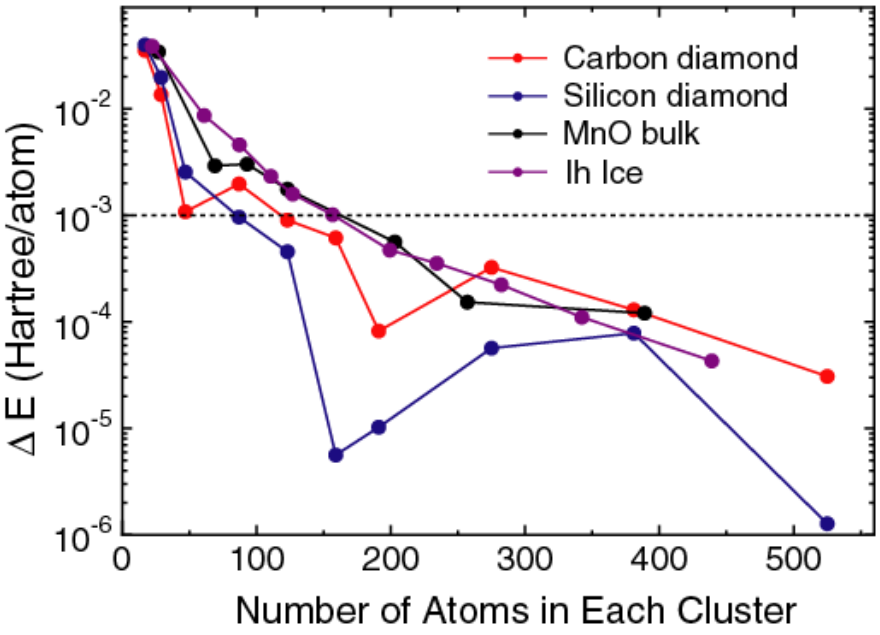
Local electronic structure of each atom is mainly determined by neighboring atomic arrangement producing chemical environment.

Convergence by the DC method

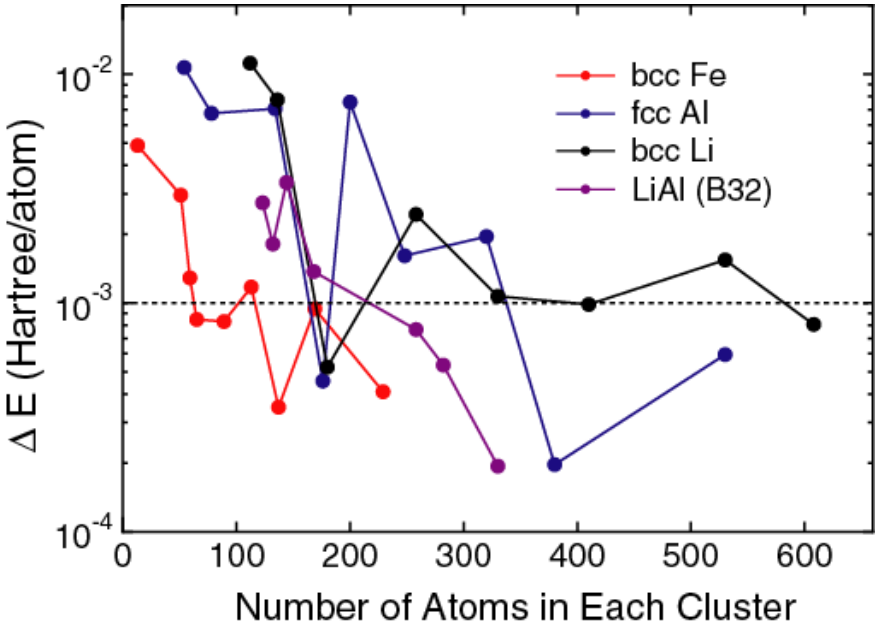
Just solve the truncated clusters → Divide-Conquer method

W.Yang, PRL 66, 1438 (1991)

Insulators, semi-conductors



Metals

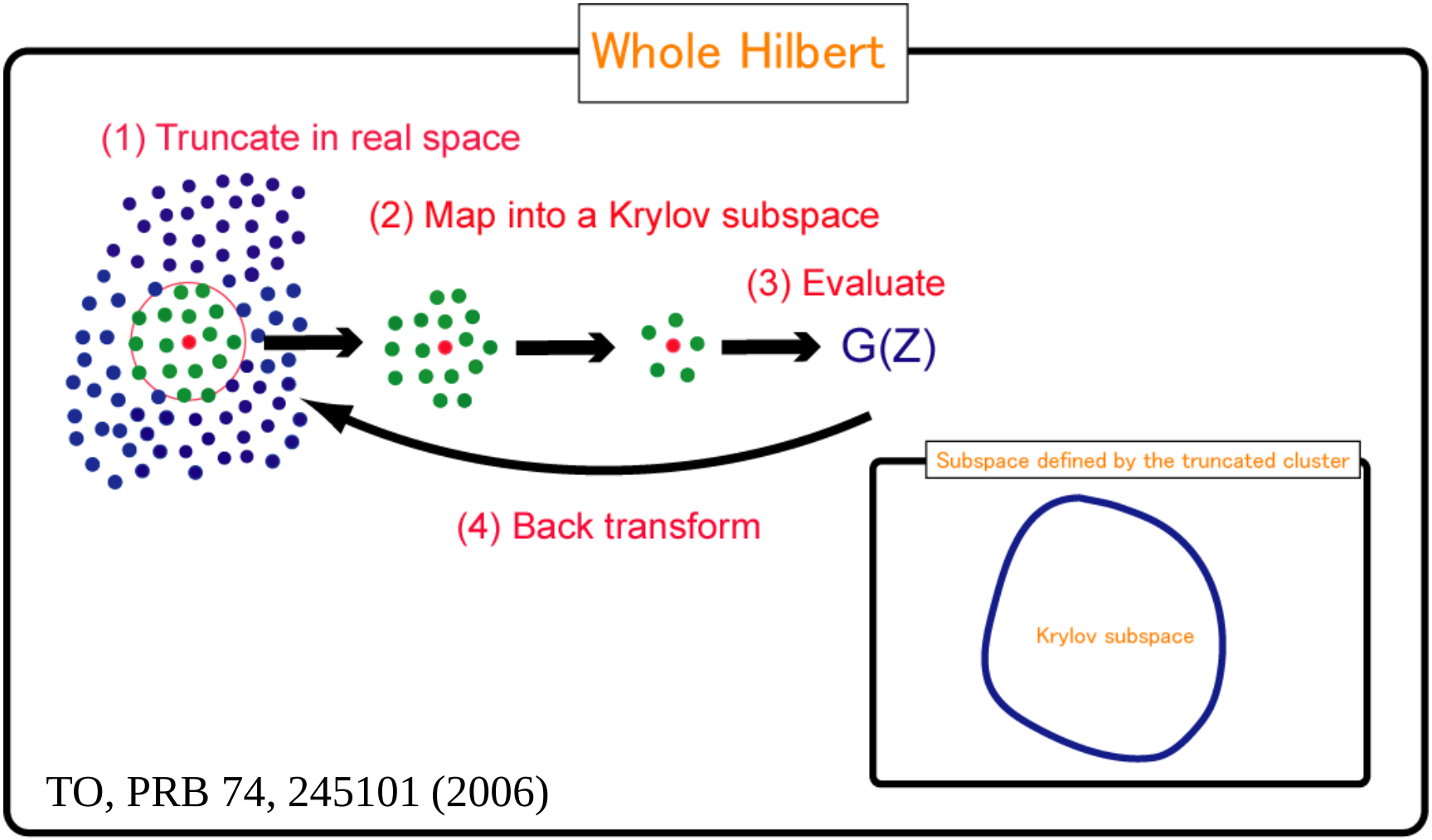


For metals, a large cluster size is required for the convergence.

→ Difficult for direct application of the DC method for metals

O(N) Krylov subspace method

Two step mapping of the whole Hilbert space into subspaces

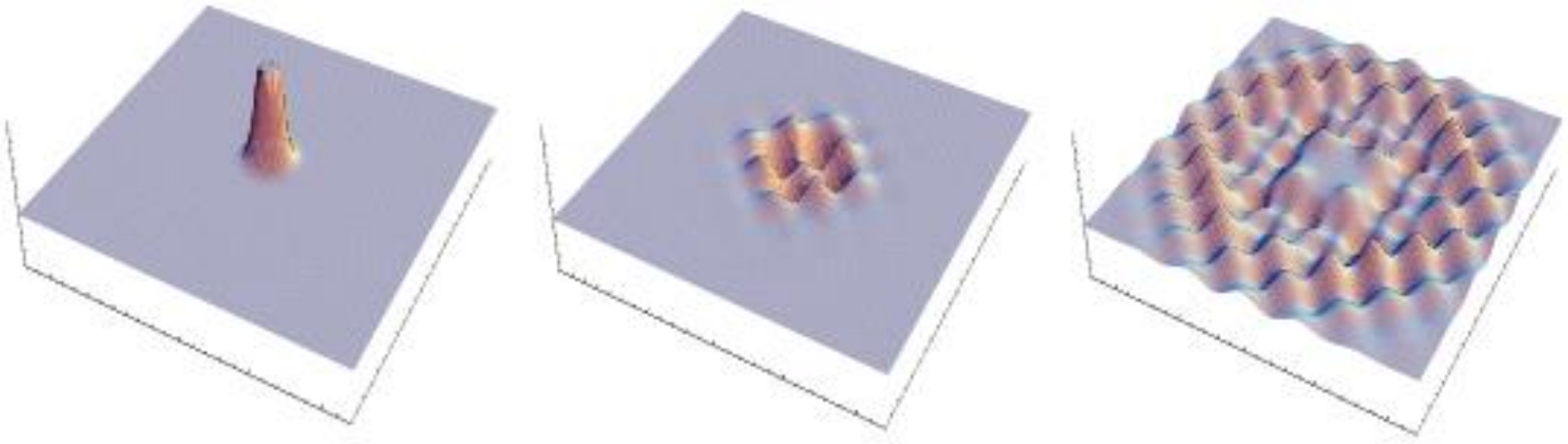


TO, PRB 74, 245101 (2006)

Development of Krylov subspace vectors

The Krylov vector is generated by a multiplication of H by |K>, and the development of the Krylov subspace vectors can be understood as hopping process of electron.

$|K_0\rangle$ $|K_1\rangle$ $|K_5\rangle$



The information on *environment* can be included from near sites step by step, resulting in reduction of the dimension.

Generation of Krylov subspaces

The ingredients of generation of Krylov subspaces is to multiply $|W_n)$ by $S^{-1}H$. The other things are made only for stabilization of the calculation.

$$|R_{n+1}) = S^{-1}H|W_n)$$

$$|W'_{n+1}) = |R_{n+1}) - \sum_{m=0}^{n} |W_m)(W_m|\hat{S}|R_{n+1})$$

$$|W_{n+1}) = S - \text{orthonormalized block vector of } |W'_{n+1})$$

Furthermore, in order to assure the S-orthonormality of the Krylov subspace vectors, an orthogonal transformation is performed by

$$\mathbf{U}_\mathrm{K} = \mathbf{W}\mathbf{X}\lambda^{-1}$$

$$\lambda^2 = \mathbf{X}^\dagger\mathbf{W}^\dagger\hat{S}\mathbf{W}\mathbf{X}$$

For numerical stability, it is crucial to generate the Krylov subspace at the first SCF step.

Embedded cluster problem

Taking the Krylov subspace representation, the cluster eigenvalue problem is transformed to a standard eigenvalue problem as:

$$H c_\mu = \varepsilon_\mu S c_\mu \longrightarrow H^{\mathrm{K}} b_\mu = \varepsilon b_\mu$$

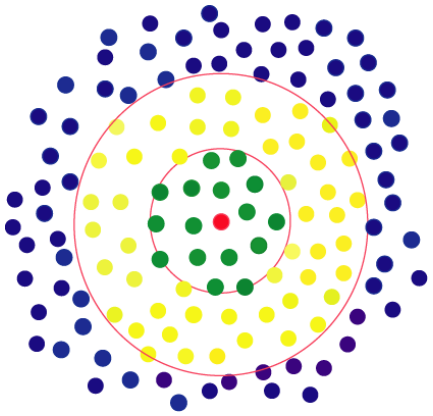
where H^{K} consists of the short and long range contributions.

$$\begin{aligned} H^{\mathrm{K}} &= U^\dagger H U \\ &= u_c^\dagger H_c u_c + u_c^\dagger H_{cb}^\dagger u_b + u_b^\dagger H_{bc} u_c + u_b^\dagger H_b u_b \\ &= H_s^{\mathrm{K}} + H_l^{\mathrm{K}} \end{aligned}$$

updated **fixed**

Green: core region

Yellow: buffer region



- **The embedded cluster is under the Coulomb interaction from the other parts.**
- **The charge flow from one embedded cluster to the others is allowed.**

Relation between the Krylov subspace and Green's funtion

A Krylov subspace is defined by

$$\mathbf{U}_{\mathrm{K}} = \left\{|W_0), (S^{-1}H)|W_0), (S^{-1}H)^2|W_0), ..., (S^{-1}H)^q|W_0)\right\}$$

A set of q-th Krylov vectors contains up to information of (2q+1)th moments.

$$\begin{aligned}\underline{\mathrm{H}}^K_{mn} &= (W_0|(A^\dagger)^m H A^n|W_0) \\ &= (W_0|S(S^{-1}H)^{m+n+1}|W_0), \\ &= (W_0|S\mu^{(m+n+1)}S|W_0)\end{aligned}$$

Definition of moments

$$\begin{aligned}\mu^{(p)} &= c\varepsilon^p c^\dagger, \\ &= cc^\dagger Hcc^\dagger Hc\cdots c^\dagger Hcc^\dagger, \\ &= (S^{-1}H)^p S^{-1}\end{aligned}$$

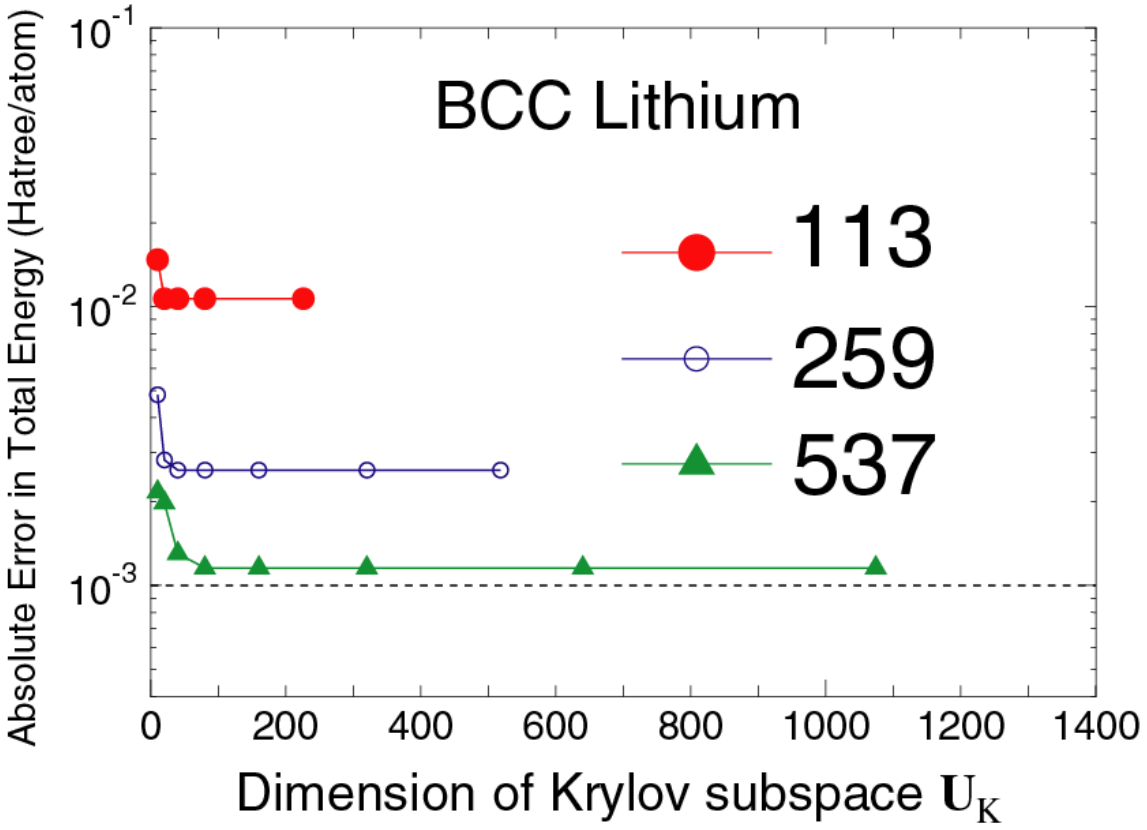
The moment representation of G(Z) gives us the relation.

$$G_{ij}(Z) = \sum_{p=0}^{\infty} \frac{\mu_{ij}^{(p)}}{Z^{p+1}}$$

One-to-one correspondence between the dimension of Krylov subspace and the order of moments can be found from above consideration.

Convergence property

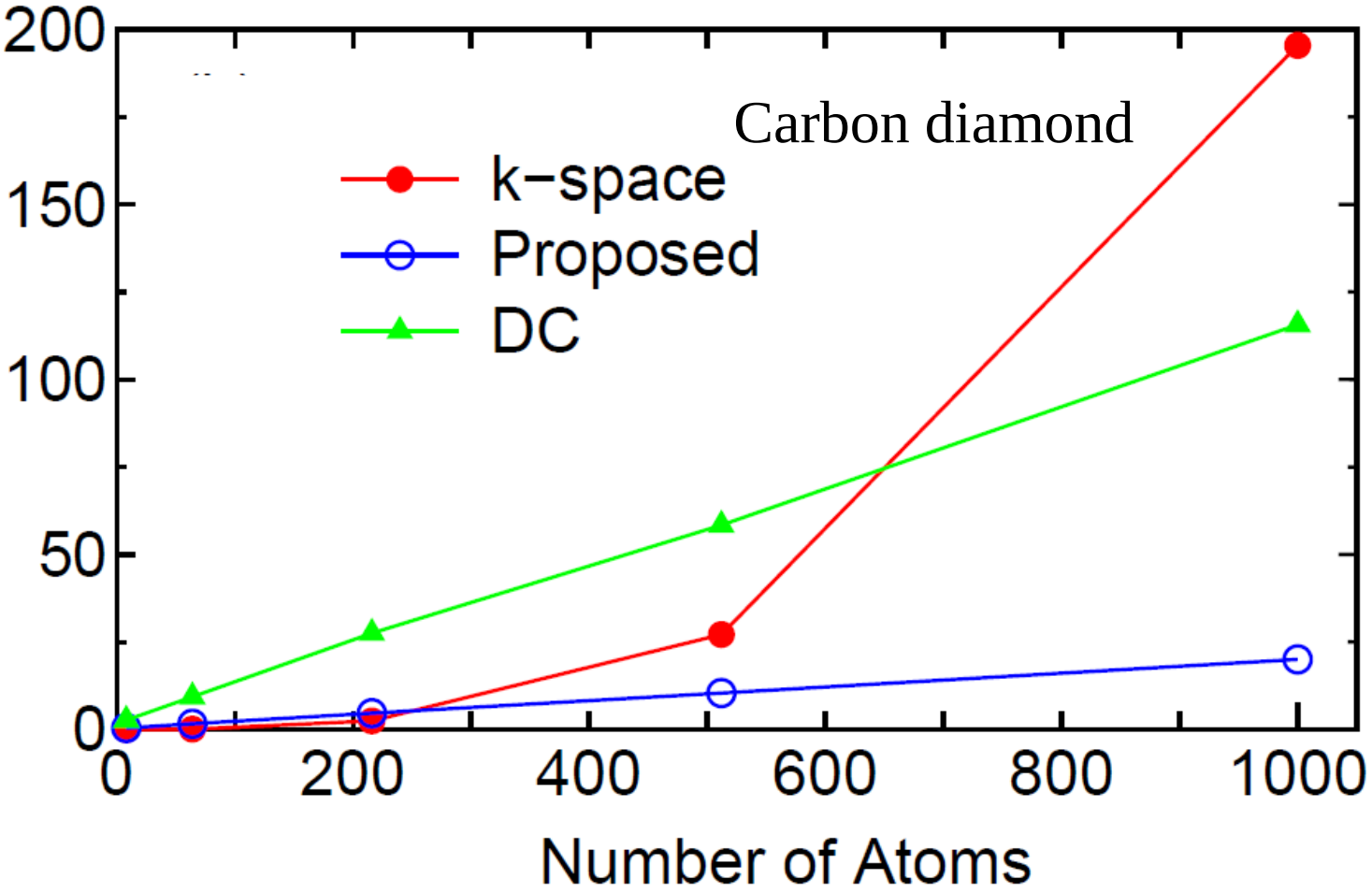
The accuracy and efficiency can be controlled by the size of truncated cluster and dimension of Krylov subspace.



In general, the convergence property is more complicated. See PRB 74, 245101 (2006).

Comparison of computational time

The computational time of calculation for each cluster does not depend on the system size. Thus, the computational time is O(N) in principle.



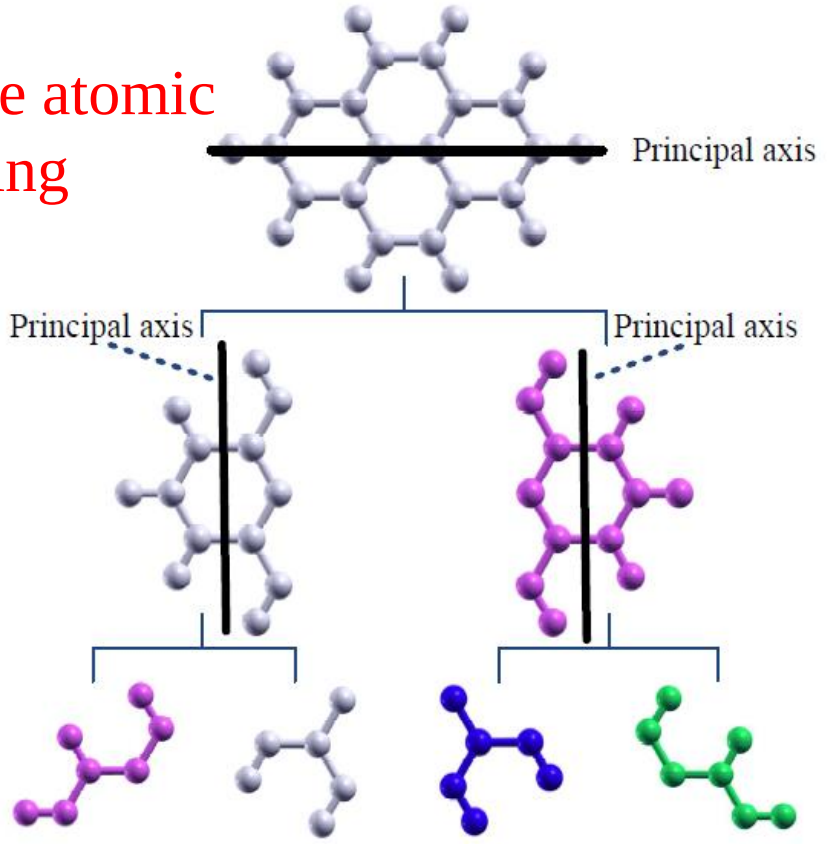
Parallelization

How one can partition atoms to minimize communication and memory usage?

Requirement:

- Locality
- Same computational cost
- Applicable to any systems
- Small computational overhead

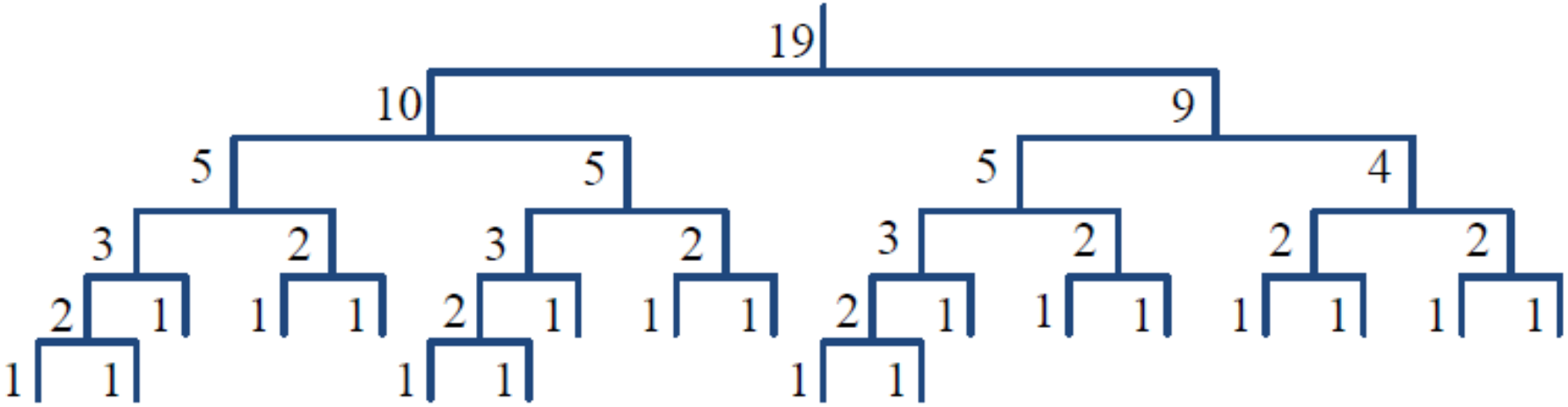
Recursive atomic partitioning



T.V.T. Duy and T. Ozaki, CPC 185, 777 (2014).

Modified recursive bisection

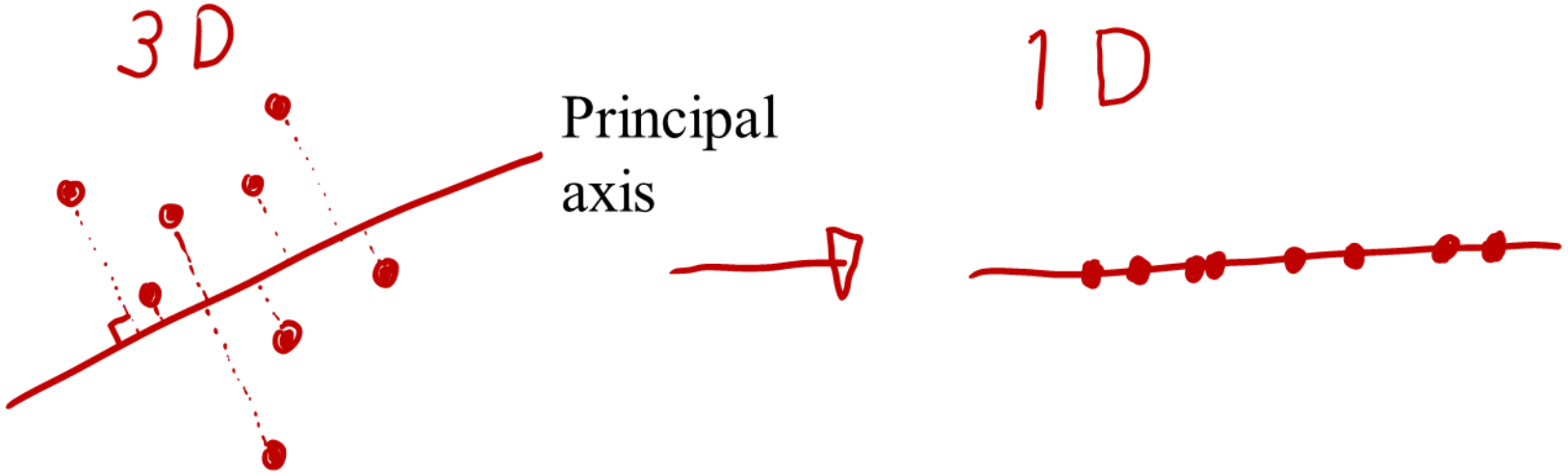
If the number of MPI processes is 19, then the following binary tree structure is constructed.



In the conventional recursive bisection, the bisection is made so that a same number can be assigned to each region. However, the modified version bisects with weights as shown above.

Reordering of atoms by an inertia tensor

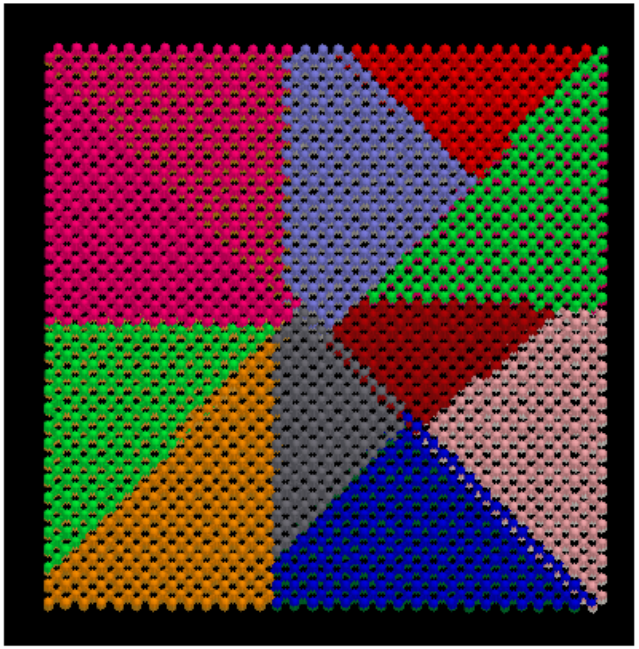
Atoms in an interested region are reordered by projecting them onto a principal axis calculated by an inertia tensor.



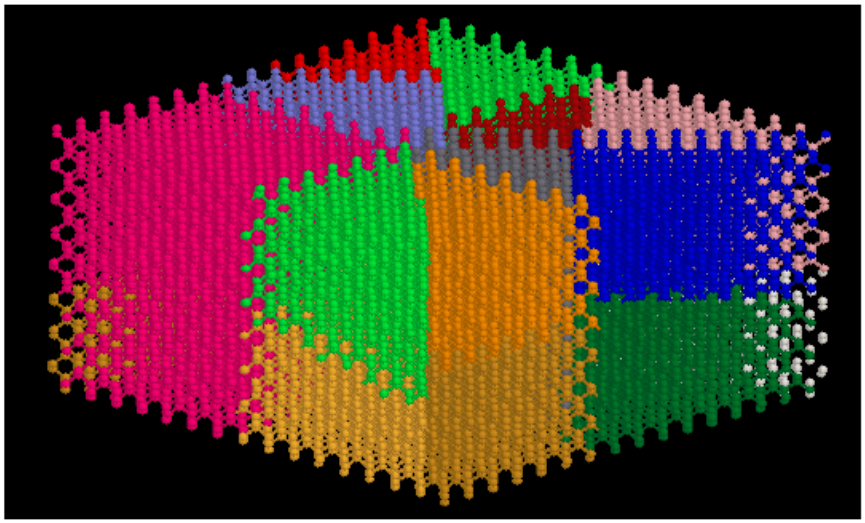
The principal axis is calculated by solving an eigenvalue problem with an inertia tensor:

$$\begin{pmatrix} \sum_i w_i (Y_i^2 + Z_i^2) & -\sum_i w_i X_i Y_i & -\sum_i w_i X_i Z_i \\ -\sum_i w_i Y_i X_i & \sum_i w_i (X_i^2 + Z_i^2) & -\sum_i w_i Y_i Z_i \\ -\sum_i w_i Z_i X_i & -\sum_i w_i Z_i Y_i & -\sum_i w_i (X_i^2 + Y_i \end{pmatrix} \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} = -\lambda \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix}$$

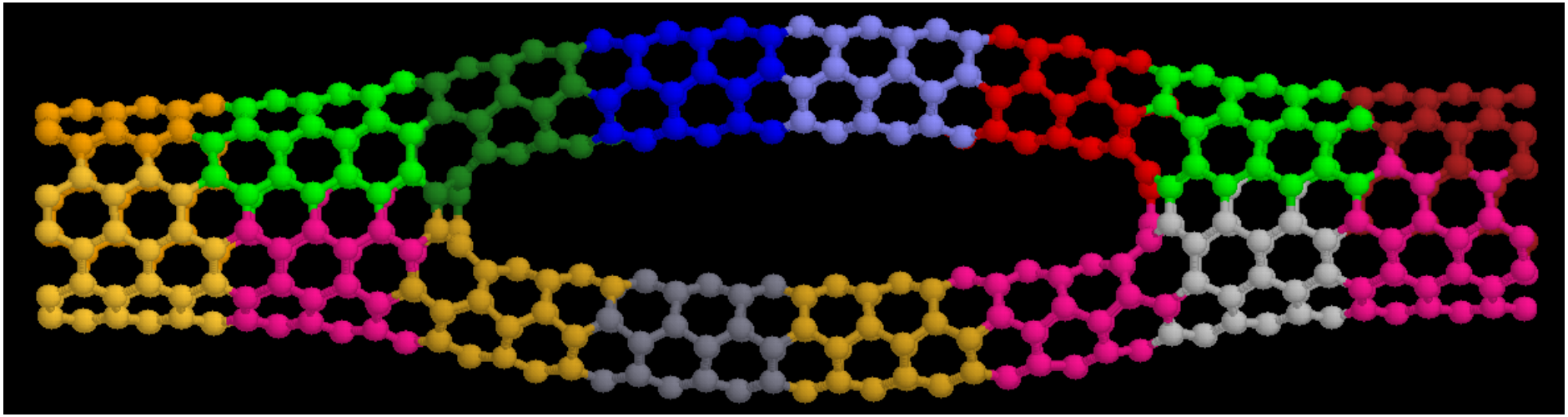
Allocation of atoms to processes



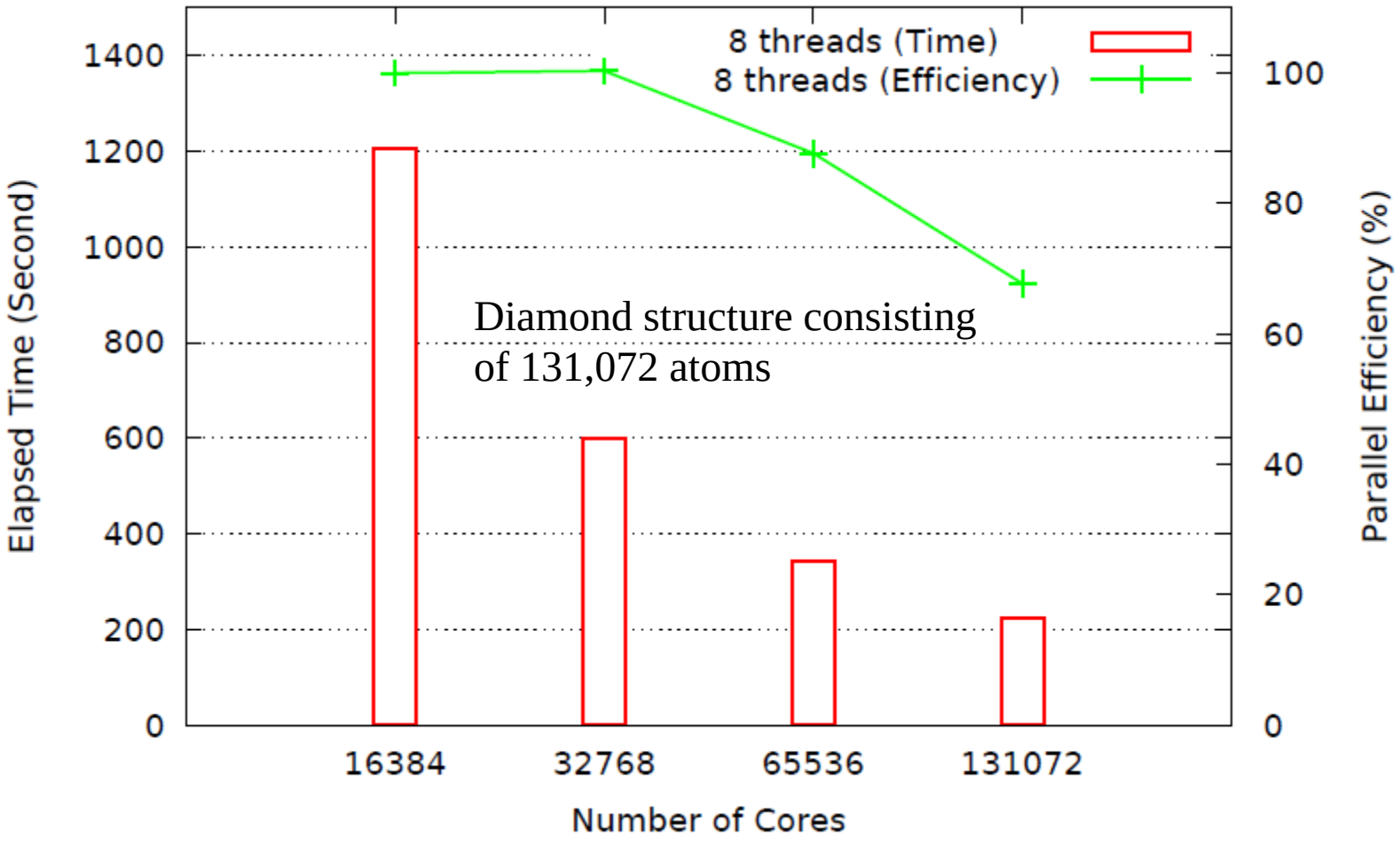
Diamond 16384 atoms, 19 processes



Multiply connected CNT, 16 processes



Parallel efficiency on K



The parallel efficiency is 68 % using 131,072 cores.

Applications of the O(N) method

1. Interface structure between BCC Iron and carbides

   H. Sawada et al., Modelling Simul. Mater. Sci. Eng. 21, 045012 (2013).

2. Desolvation of Li^+

   T. Ohwaki et al., J. Chem. Phys. 136, 134101 (2012).
   T. Ohwaki et al., J. Chem. Phys. 140, 244105 (2014).

3. Electronic transport of graphene nanoribbon

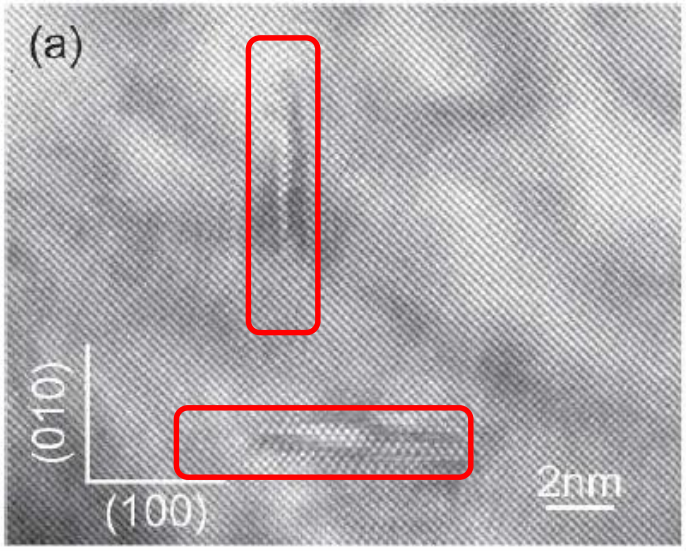
   M. Ohfuchi et al., Appl. Phys. Express 7, 025101 (2014).
   H Jippo, T Ozaki, S Okada, M Ohfuchi, J. Appl. Phys. 120, 154301 (2016).

Precipitation in bcc-Fe

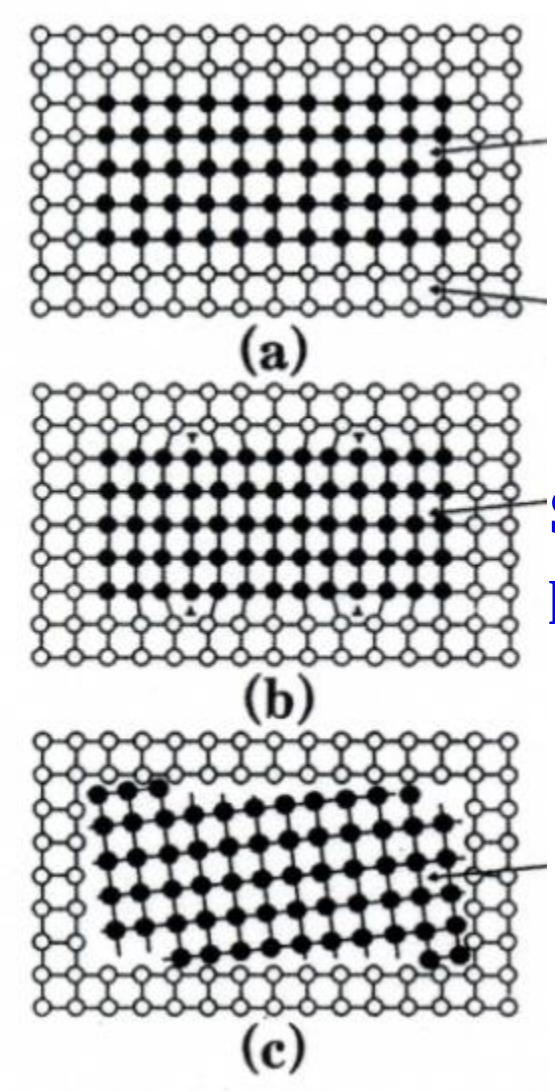
In collaboration wit Dr. Sawada (Nippon Steel)

Pure iron is too soft as structural material. Precipitation of carbide can be used to control the hardness of iron.

HRTEM image



Precipitating materials:
TiC, VC, NbC

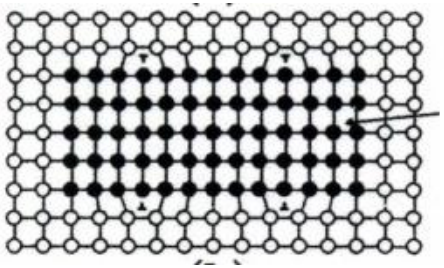


Coherent precipitation

Semicoherent precipitation

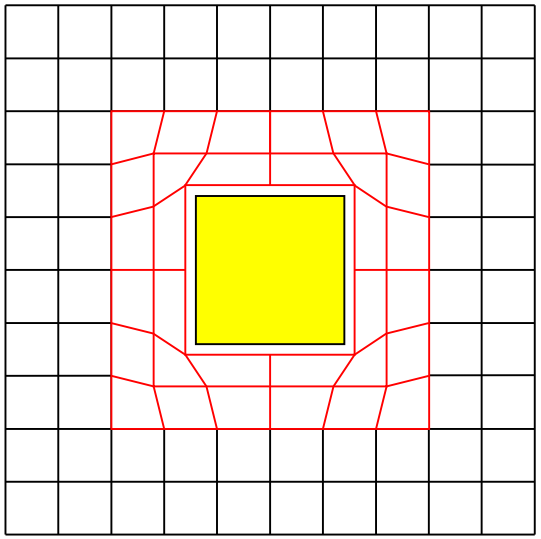
Incoherent precipitation

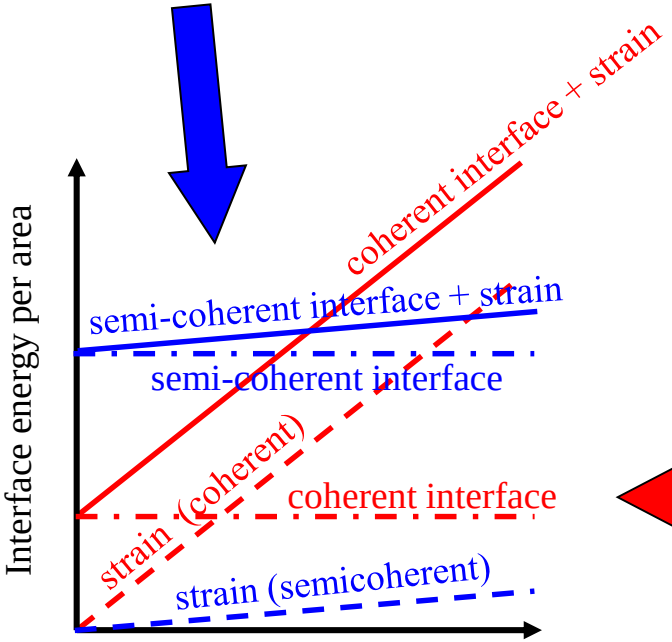
Interface and strain energies

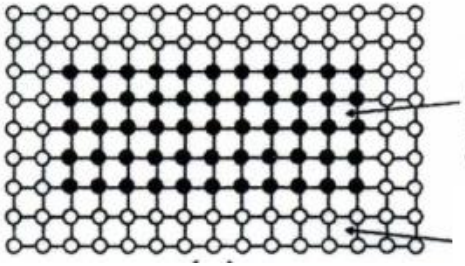


Semi-coherent case

Strain field





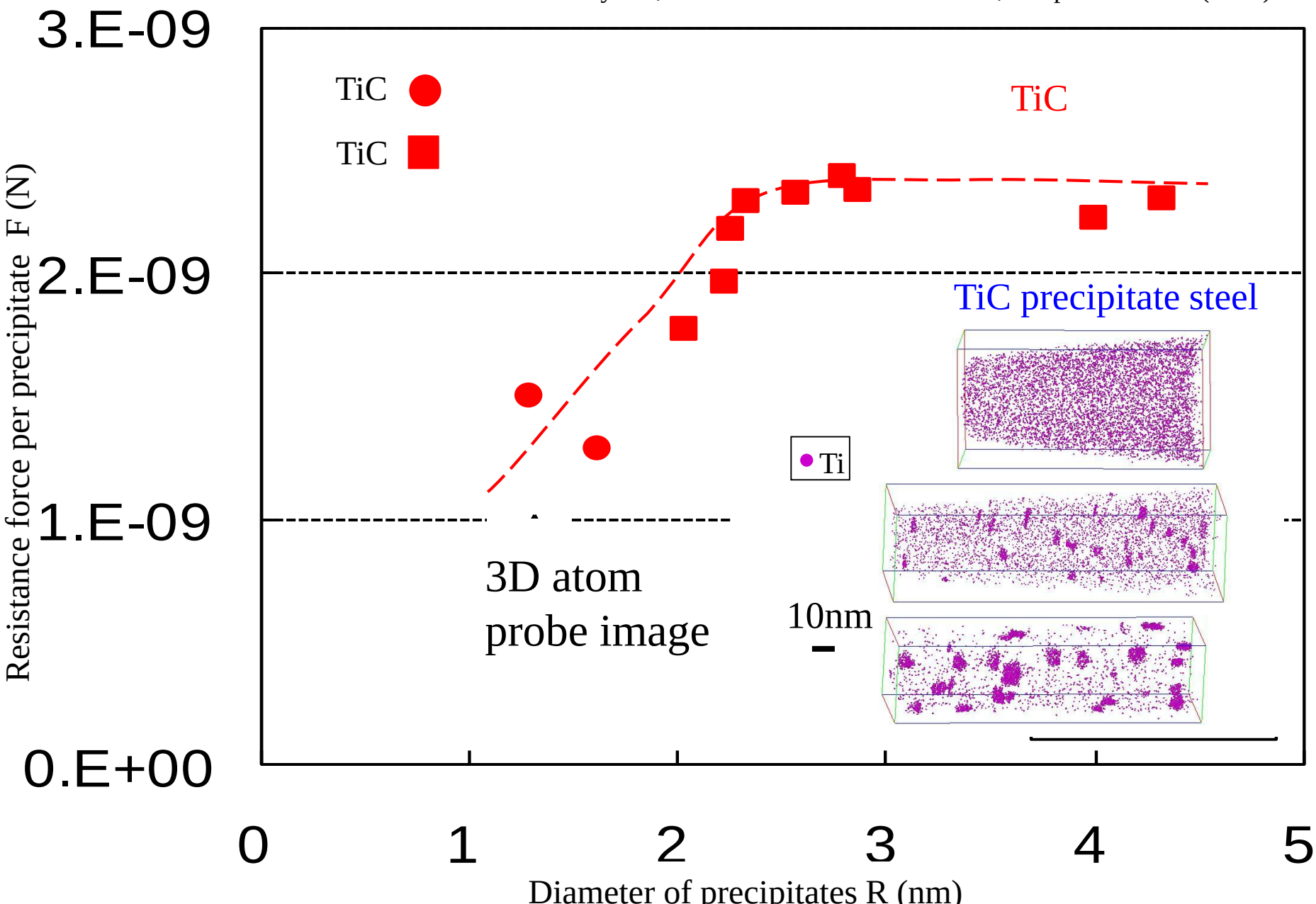


coherent case

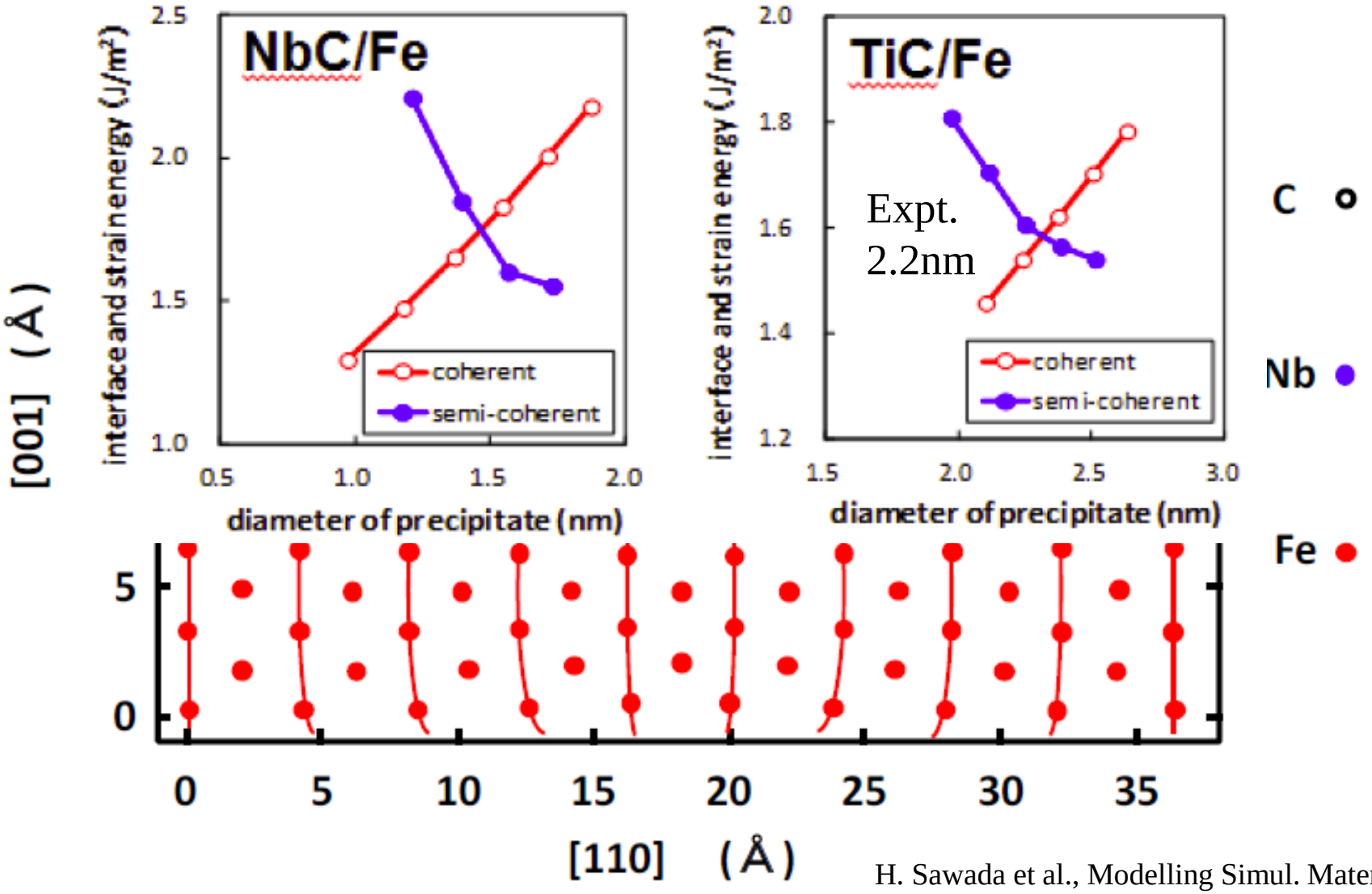
Iron

Resistance force and precipitate diameter

Y. Kobayashi, J. Takahashi and K. Kawakami, Scripta Mater. 67 (2012) 854



Crossover from coherent to semi-coherent



H. Sawada et al., Modelling Simul. Mater. Sci. Eng. 21, 045012 (2013).

Numerically exact low-order scaling method

Main difficulty: 'diagonalization'

$O(N^3)$ method - Numerically exact diagonalization

Householder+QR method

Conjugate gradient (CG) method

Davidson method

Even if basis functions are localized in real space, Gram-Shmidt (GS) type method is needed to satisfy orthonormality among eigenstates, which results in $O(N^3)$ for the computational time.

O(N) method - can be achieved in exchange for accuracy.

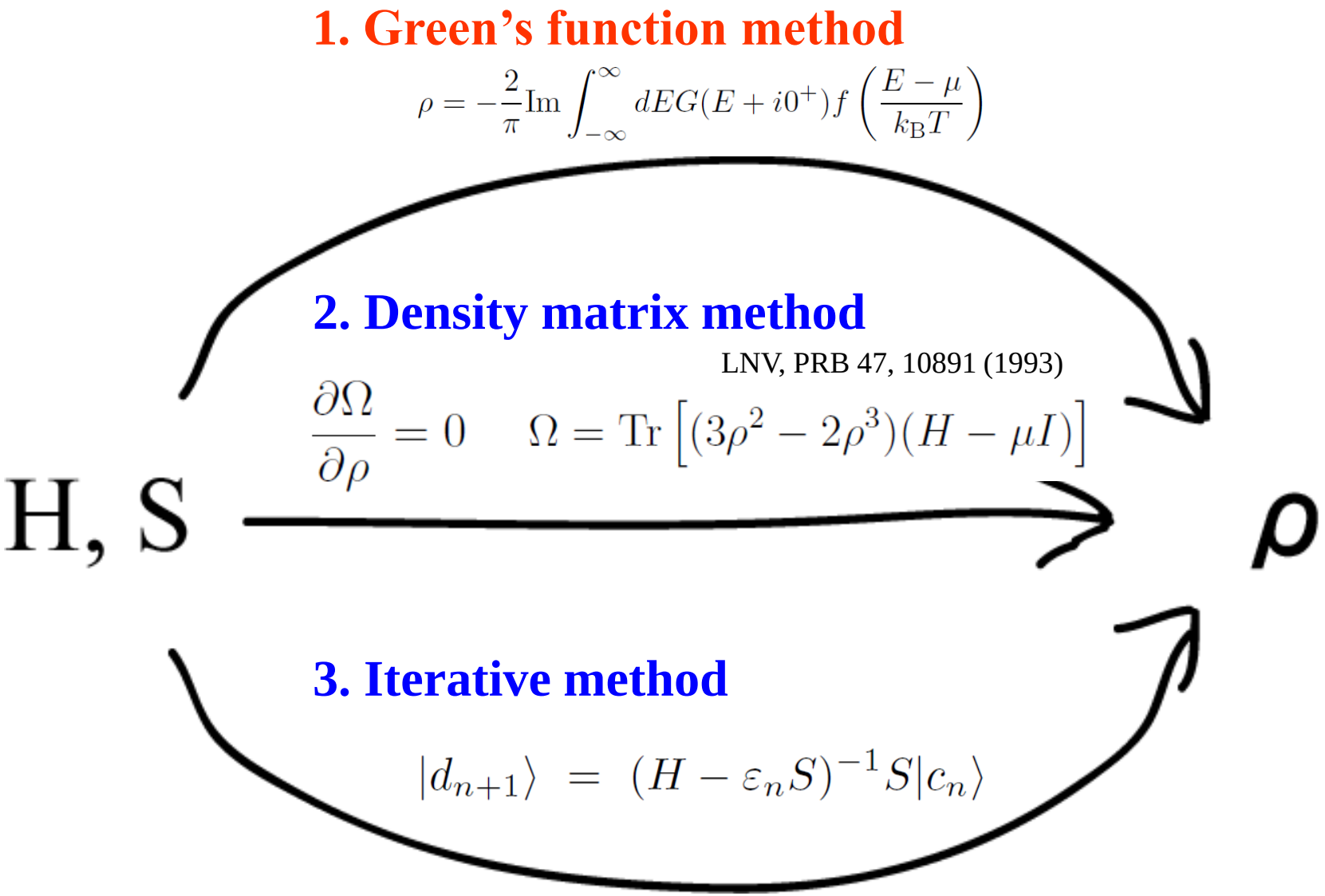
O(N) Krylov subspace method,

DC, DM, OM methods, etc..

$O(N^{2\sim})$ method

Is it possible to develop $O(N^{2\sim})$ methods without introducing approximations? → No more GS process.

Possible ways to avoid orthogonalization



Numerically exact low-order scaling method

- ✓ Numerically exact
- ✓ Applicable to insulators and metals
- ✓ Suitable for parallel computation
- ✓ Applicable to 1D, 2D, 3D systems
- ✓ Applicable to any local basis functions

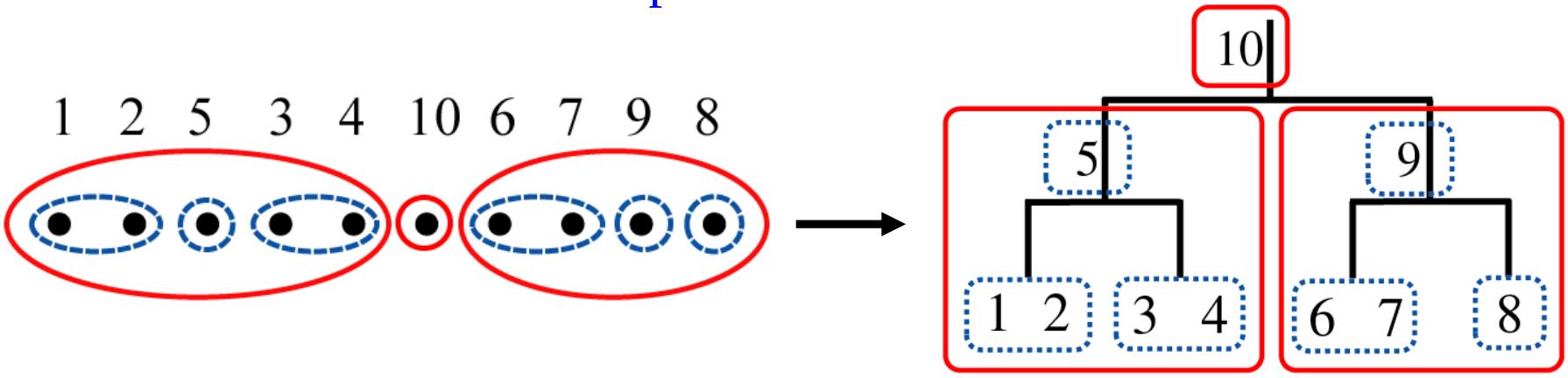
TO, PRB **82**, 075131 (2010)

Numerically exact low-order scaling method

1. Direct evaluation of the selected elements of ρ via a contour integration of the Green's function

$$\rho = M^{(0)} + \mathrm{Im}\left(-\frac{4i}{\beta}\sum_{p=1}^{\infty} G(\alpha_p) R_p\right)$$

2. Nested dissection of sparse matrix



TO, PRB **82**, 075131 (2010)

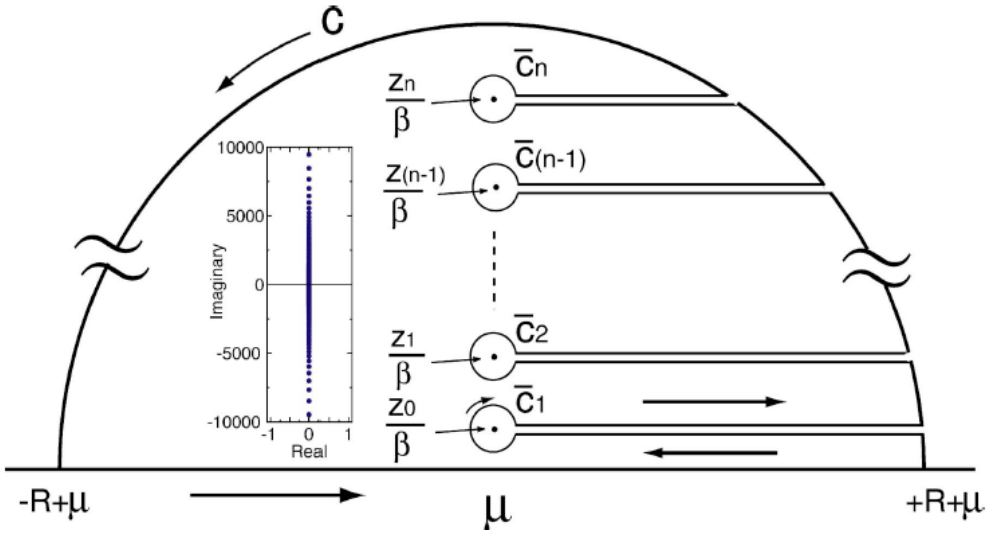
Continued fraction rep. of Fermi function

$$\frac{1}{1+\exp(x)} = \frac{1}{2} - \frac{x}{4}\left(\cfrac{1}{1+\cfrac{\left(\frac{x}{2}\right)^2}{3+\cfrac{\left(\frac{x}{2}\right)^2}{5+\cfrac{\left(\frac{x}{2}\right)^2}{\cfrac{\cdots}{(2M-1)+}\ddots,}}}}\right)$$

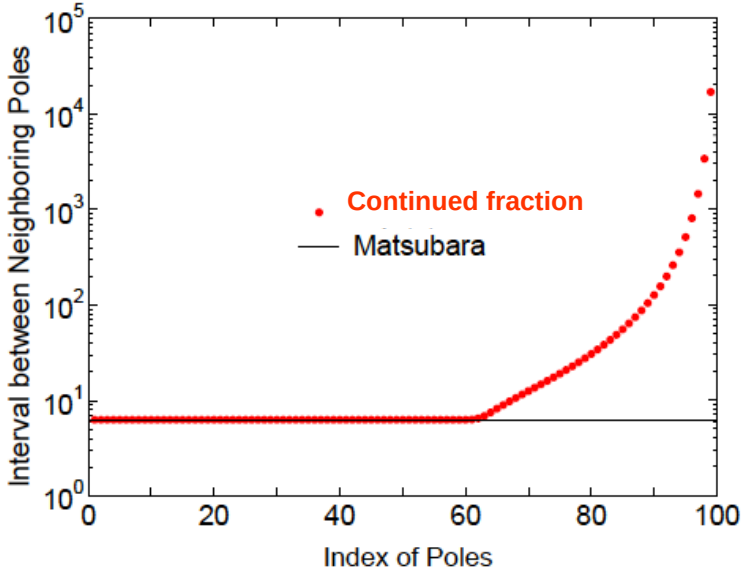
TO, PBR **75**, 035123 (2007)

Contour integration

All the poles are located on the imaginary axis.



The form has a special pole structure, that is, the interval between neighboring poles increases in a faraway region from the real axis, which is very advantageous for the contour integration of Green's function.



Convergence of ρ w.r.t. poles

The calculation of ρ can be expressed by a contour integration:

$$
\begin{aligned}
\rho_{ij} &= \sum_k f(\frac{\varepsilon_k-\mu}{k_{\rm B}T})\langle\chi_i|\phi_k\rangle\langle\phi_k|\chi_j\rangle, \\
&= -\frac{2}{\pi}{\rm Im}\int_{-\infty}^{\infty} dE f(\frac{E-\mu}{k_{\rm B}T})G_{ij}(E+i0^+), \\
&= M_{ij}^{(0)} + {\rm Im}\left[-\frac{4i}{\beta}\sum_{p=1}^{\infty} G_{ij}(\alpha_p)R_p\right], \qquad M_{ij}^{(0)} = {\rm Im}\left[-\frac{1}{\pi}\int_{-\infty}^{\infty} dE G_{ij}(E+i0^+)\right] \simeq iR\,G(iR) \\
&= M_{ij}^{(0)} + {\rm Im}\left[-\frac{4i}{\beta}\sum_{p=1}^{\infty}\sum_k \frac{\langle\chi_i|\phi_k\rangle\langle\phi_k|\chi_j\rangle}{\alpha_p-\varepsilon_k}R_p\right], \qquad \text{Lehmann rep.} \\
&= M_{ij}^{(0)} + \sum_k {\rm Im}\left[-\frac{4i}{\beta}\sum_{p=1}^{\infty} \frac{\langle\chi_i|\phi_k\rangle\langle\phi_k|\chi_j\rangle}{\alpha_p-\varepsilon_k}R_p\right],
\end{aligned}
$$

$\alpha_p = \mu_0 + i\frac{z_p}{\beta}$

The analysis shows that the number of poles for each eigenstate for a sufficient convergence does not depend on the size of system if the spectrum radius does not change. → The scaling property is governed by the calculation of G.

Convergence property of the contour integration

Total energy of aluminum as a function of the number of poles by a recursion method at 600 K.

Nicholson et al., PRB **50**, 14686 (1994).

Poles	Proposed	$\frac{1}{1+\left(1+\frac{x}{n}\right)^n}$	Matsubara
10	−42.933903047211	−33.734015919550	−39.612354360046
20	−47.224346653790	−33.623477214678	−39.849746603905
40	−48.323790725570	−33.346245616679	−40.216055898502
60	−48.324441992259	−33.143128624551	−39.676965494522
80	−48.324441994952	−32.870752577236	−43.523770052176
150	−48.324441994952	−33.837428496424	−41.836938942518
200		−33.418012271726	−42.543354202255
250		−34.003411636691	−43.024756221080
300		−34.003236479262	−43.466729654170
350		−48.324440028792	−43.834528739677
400		−48.324440274509	−44.185100655185
600		−48.324440847749	−45.233651519749
1000		−48.324441306517	−46.331692884149
2000		−48.324441650693	−47.202779497545
5000		−48.324441857239	−47.921384128418
10000		−48.324441926094	−48.122496320516

The energy completely converges using only 80 poles within double precision.

How can Green's funtion be evaluated ?

- The Green's function is the inverse of a sparse matrix *(ZS-H)*.

$$G(Z) = (ZS - H)^{-1}$$

- Selected elements of *G(Z)*, which correspond to non-zero elements of the overlap matrix *S*, are needed to calculate physical properties.

Our idea

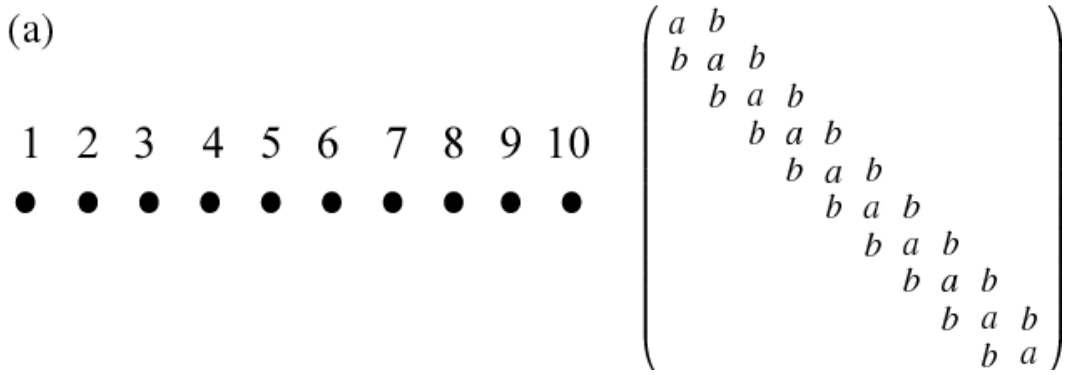
1. Nested dissection of *(ZS-H)*
2. LDL^T decomposition for the structured matrix

→ a set of recurrence relations

TO, PRB **82**, 075131 (2010)

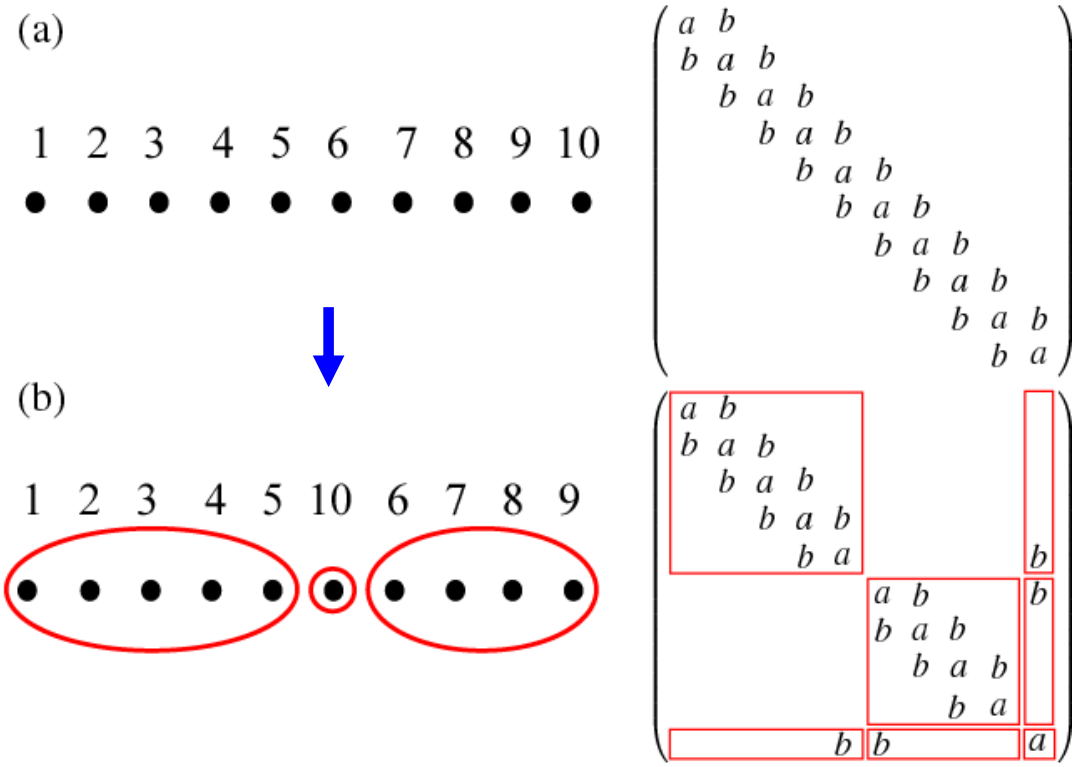
Nested dissection of a sparse matrix

George, SIAM J. Numer. Anal. 10, 345 (1973).



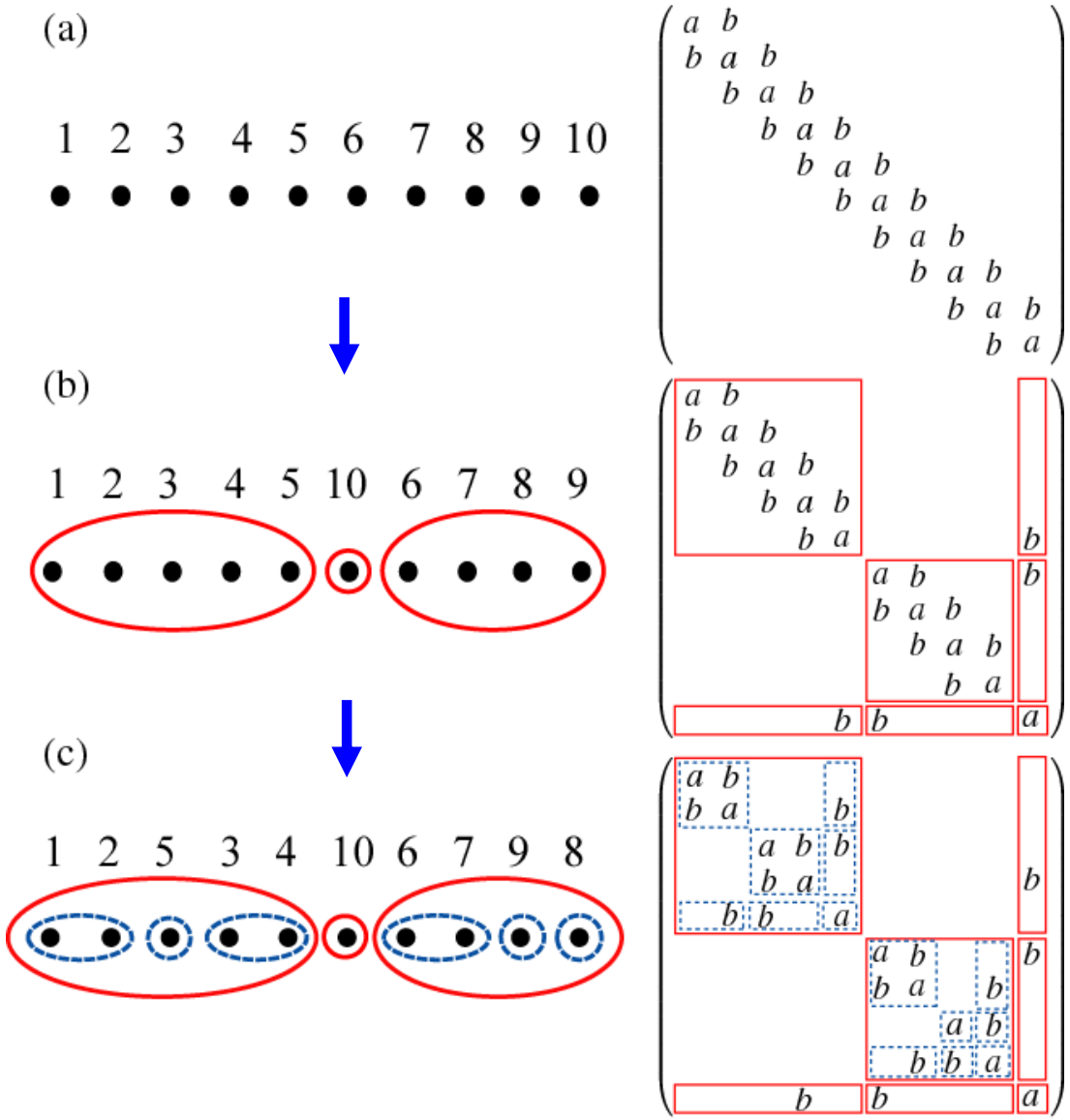
Nested dissection of a sparse matrix

George, SIAM J. Numer. Anal. 10, 345 (1973).



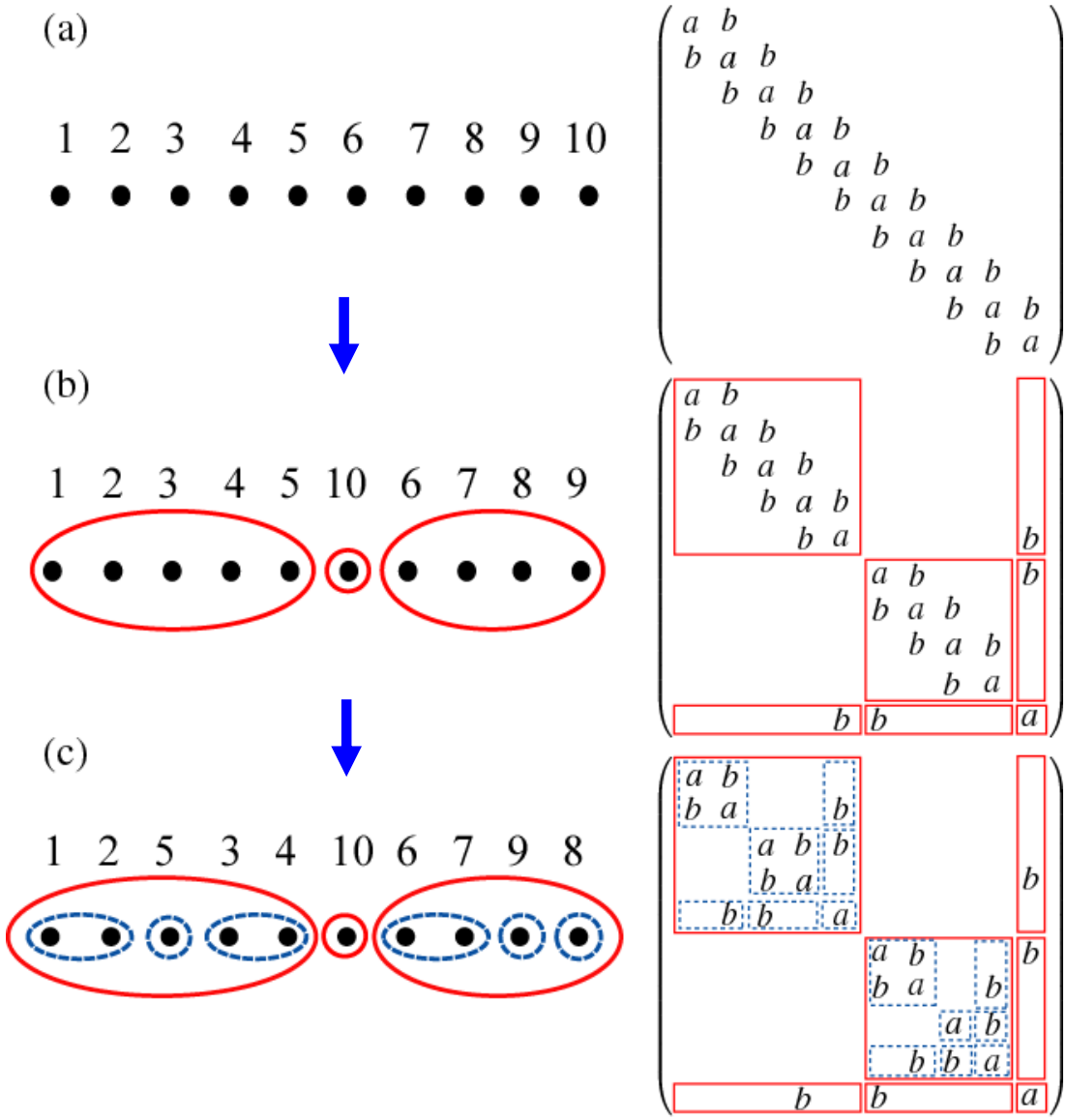
Nested dissection of a sparse matrix

George, SIAM J. Numer. Anal. 10, 345 (1973).

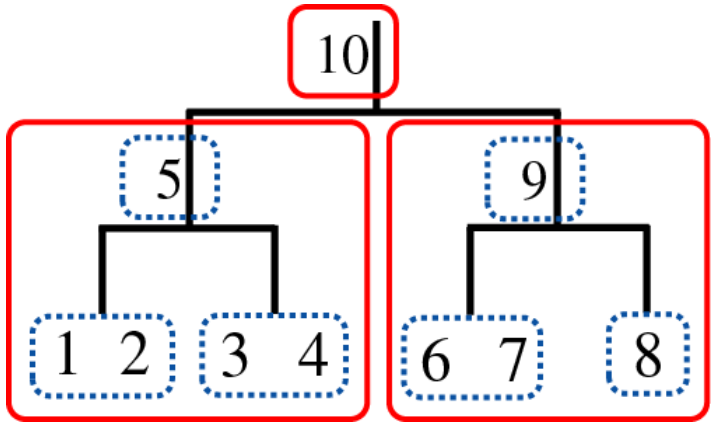


Nested dissection of a sparse matrix

George, SIAM J. Numer. Anal. 10, 345 (1973).



The hierarchical structure of interactions of domains.



Nested dissection of a sparse matrix

The processes (i)-(v) are recursively applied to each domains with computational cost of $O(N(\log_2 N)^2)$ in total.

(i) Ordering:

The basis functions are ordered by coordinates along each direction.

(ii) Screening:

The basis functions with a long tail are assigned as part of the separator.

(iii) Finding of a starting nucleus:

Find a basis function having the smallest number of nonzero overlaps.

(iv) Growth of the nucleus:

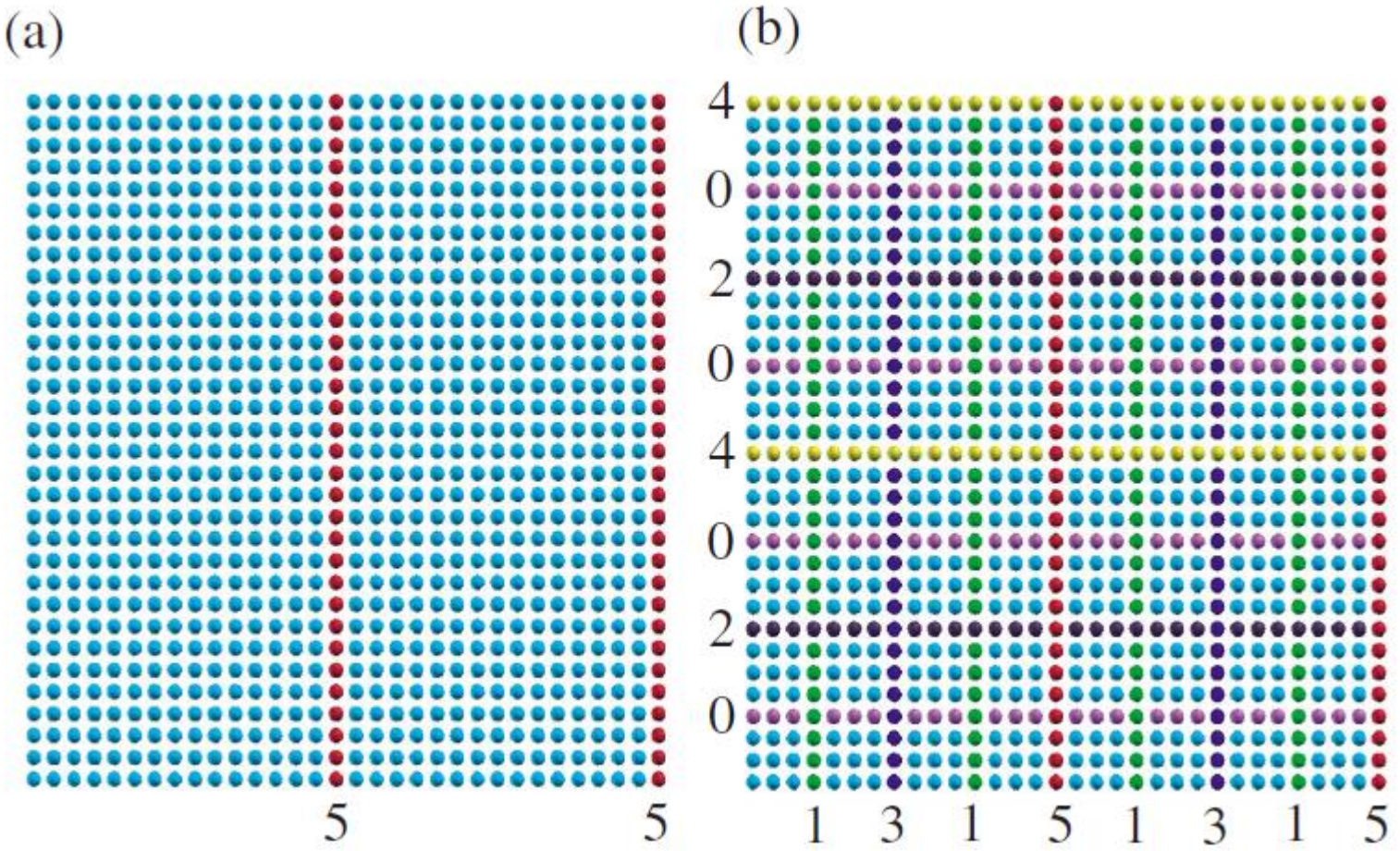
Minimize $|N_0-N_1| + N_s$ by the growth of the nucleus.

N_0: # of bases in domain 0
N_1: # of bases in domain 1
N_s: # of bases in separator

(v) Dissection:

Find a direction with the smallest $|N_0-N_1| + N_s$, make the dissection along the direction.

Square lattice for the nested dissection



Inverse by LDL^T block factorization

A matrix X can be factorized using a Schur complement into a LDL^T form.

$$X = \begin{pmatrix} A & B^T \\ B & C \end{pmatrix} = \begin{pmatrix} I & \\ L & I \end{pmatrix} \begin{pmatrix} A & \\ & S \end{pmatrix} \begin{pmatrix} I & L^T \\ & I \end{pmatrix}$$

$$L = BA^{-1}$$

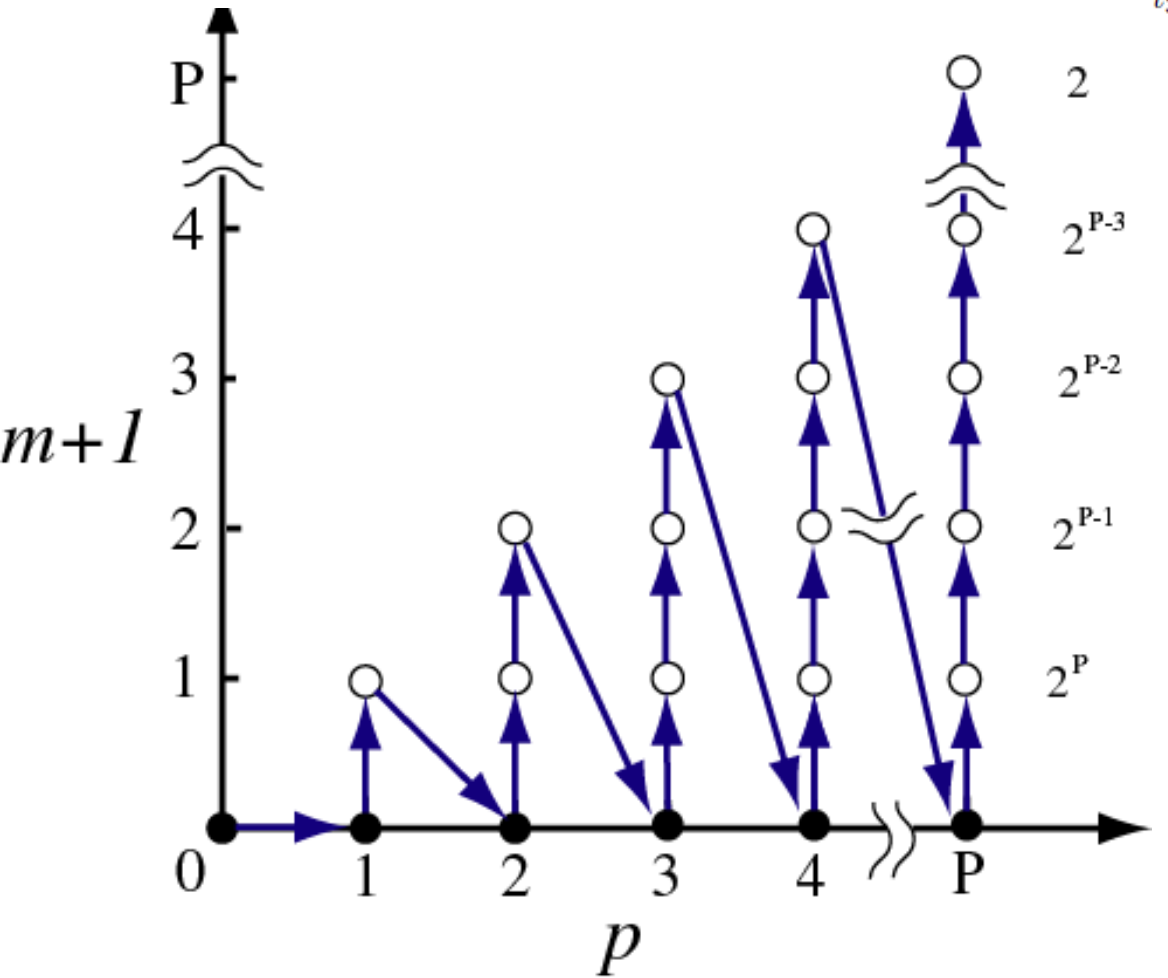
$$S = C - BA^{-1}B^T$$

Then, the inverse of X is given by

$$X^{-1} = \begin{pmatrix} A^{-1} + L^T S^{-1} L & -L^T S^{-1} \\ -S^{-1} L & S^{-1} \end{pmatrix}$$

Analysis of the computational cost

$$V^T_{p,m+1,n} = \begin{pmatrix} V^T_{p,m,2n} \\ V^T_{p,m,2n+1} \\ 0 \end{pmatrix} + \begin{pmatrix} L^T_{m,2n} \\ L^T_{m,2n+1} \\ -I \end{pmatrix} Q^T_{p,m+1,n}$$



$$t_{3D} \propto \sum_{p=1}^{P} \sum_{m=0}^{p-1} \sum_{n=0}^{2^{P-m}-1} \frac{N}{2^{P-m}} N^{(2)}_{p,m,n} N^{(3)}_{p,m,n}$$

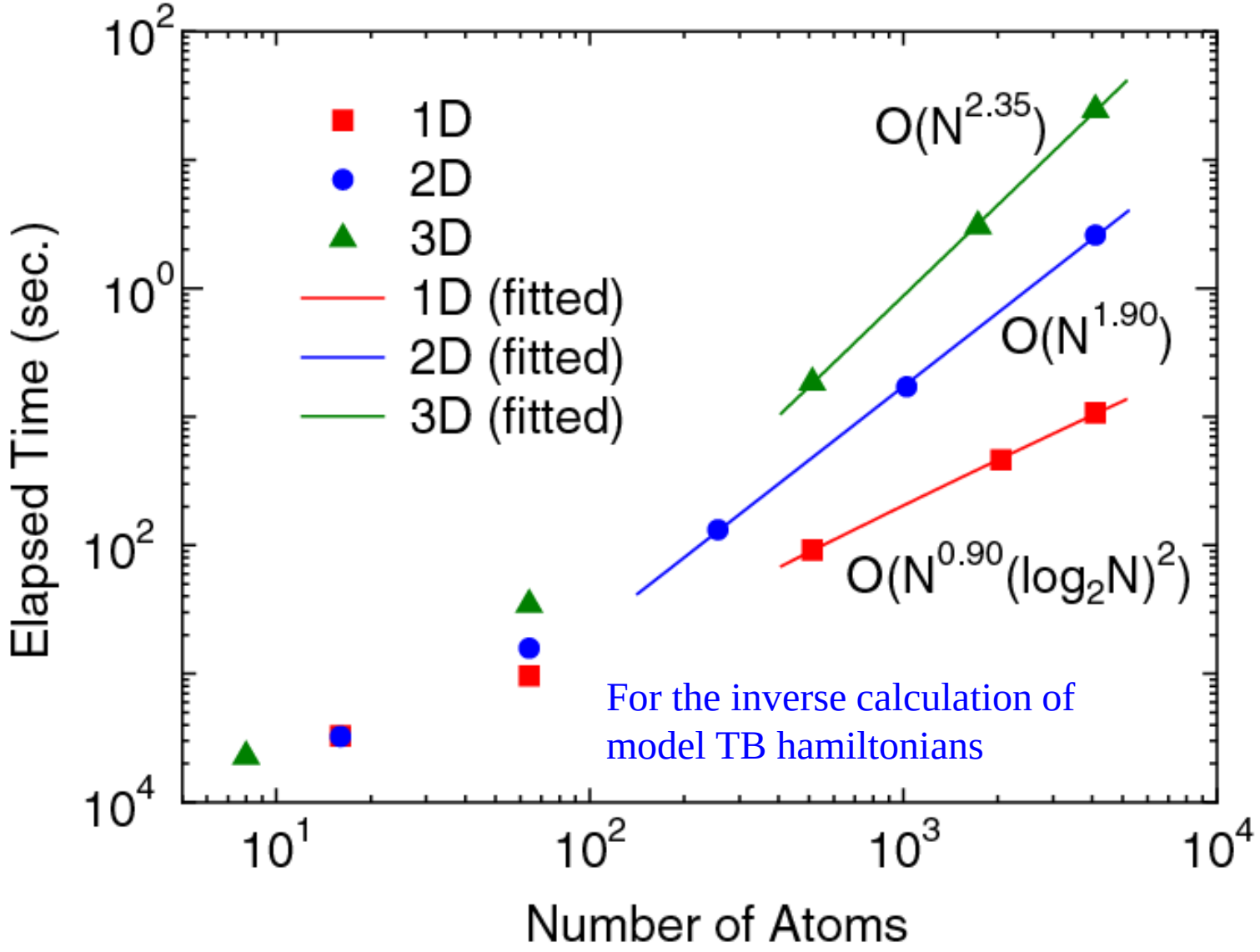
$$< \sum_{p=1}^{P} \sum_{m=0}^{p-1} \sum_{n=0}^{2^{P-m}-1} \frac{N}{2^{P-m}} \frac{N^{2/3}}{2^{\frac{2}{3}(P-m-1)}} \frac{N^{2/3}}{2^{\frac{2}{3}(P-p)}}$$

$$= \frac{4N^{7/3}}{2^{2/3}6-9} \left(-1 + 2^{2/3} - \frac{1}{2^{2/3}2^{4P/3}} + \frac{1}{2^{2/3}2^{2P/3}} - \frac{2^{2/3}}{2^{2P/3}} + \frac{1}{2^{4P/3}} \right).$$

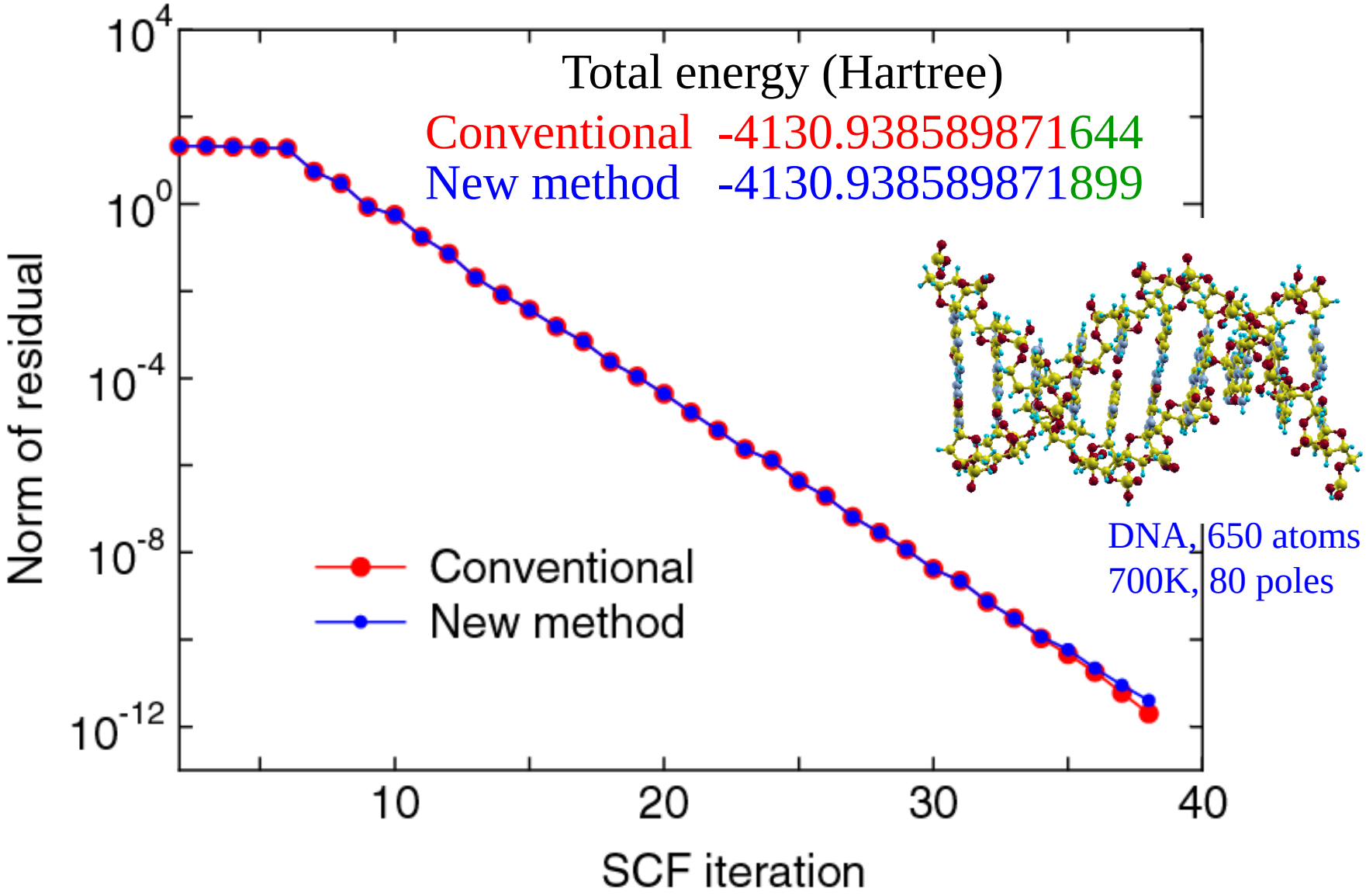
→ $O(N^{7/3})$

1D $O(N(\log_2 N)^2)$
2D $O(N^2)$
3D $O(N^{7/3})$

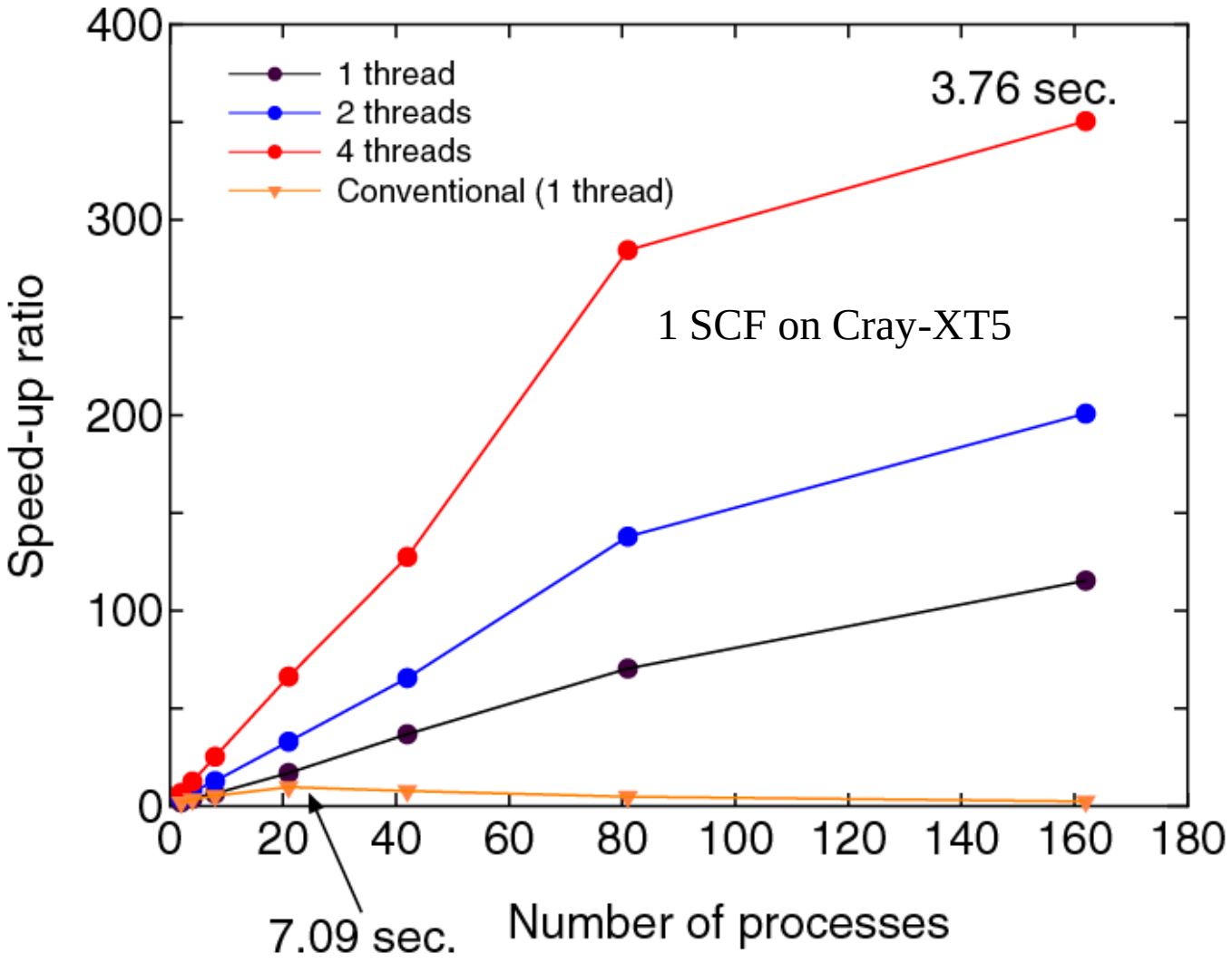
Timing result

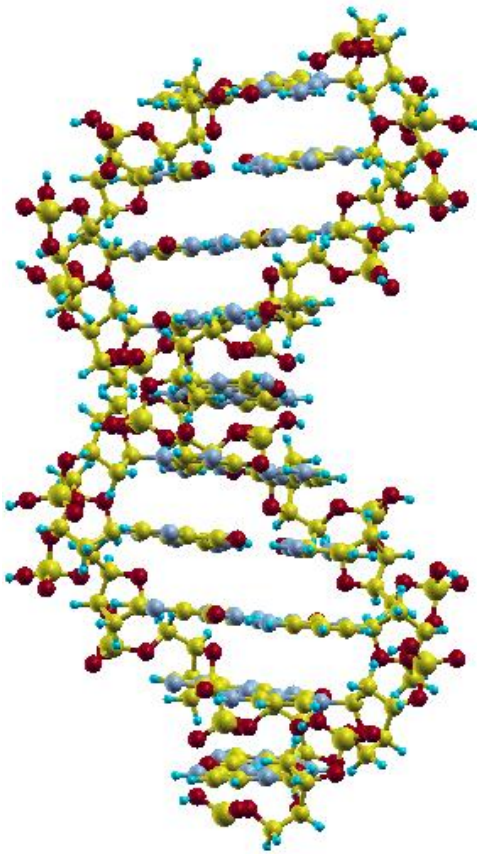


SCF convergence



Parallel efficiency





Outlook

The locality of density matrix and basis function is a key to develop a wide variety of efficient electronic structure methods.

We have demonstrated three methods:

- O(N) Krylov subspace method
- Low-order scaling exact method
- O(N) *exact* exchange method

Plenty of developments of new efficient methods might be still possible.