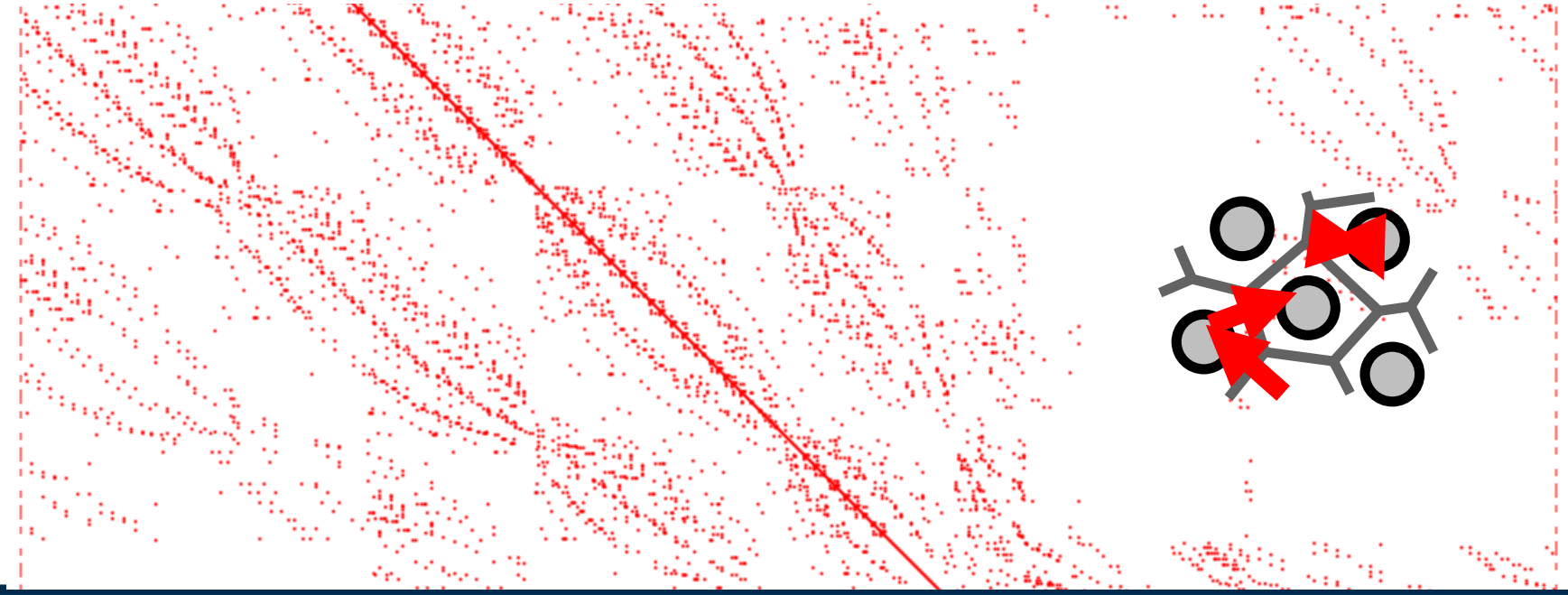


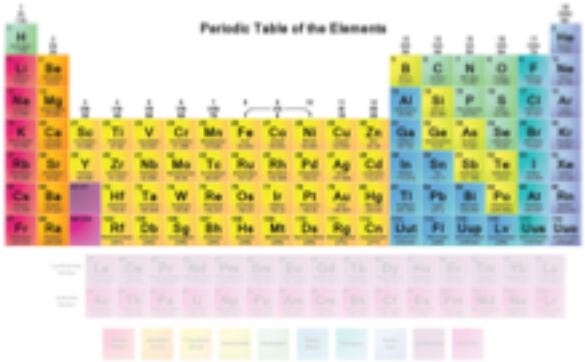
ACCELERATED LARGE-SCALE DFT

On the status of the linear-scaling DFT Application KKRnano

2018 JUNE 21 I PAUL F BAUMEISTER

DFT = Density Functional Theory

- Effective theory for the electronic structure of materials, total energy, atomic forces
- Mostly used for periodic solids



Two major trends:

- Data mining on materials → high-troughput calculations
- Realistic materials → large-scale calculations, $n_{atoms} > 1000$

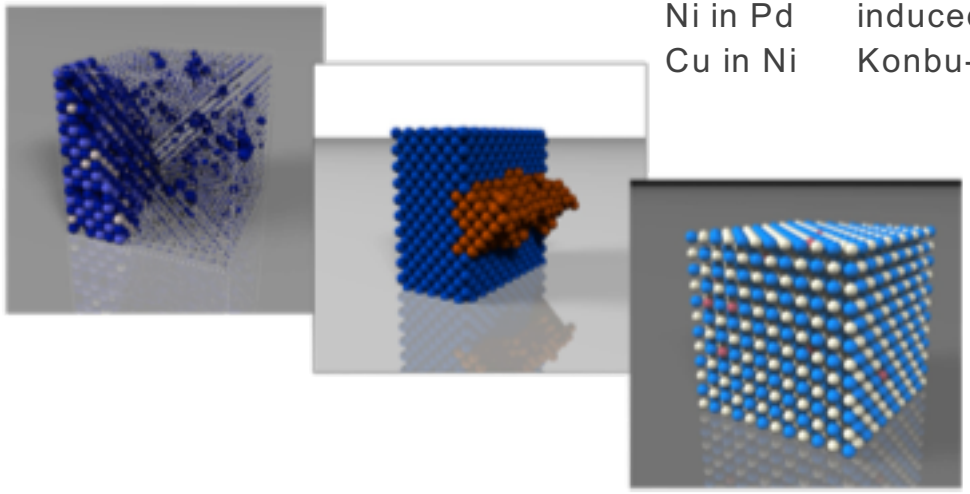
JÜLICH Forschungszentrum

KKRnano – Application to Materials

- Disorder
- Impurities
- Small relaxations
- Grain boundaries
- Metallic glasses
- Magnetic textures

Phase-Change Materials
GeSbTe disorder & localization

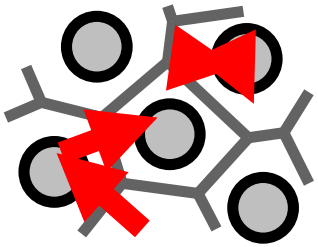
Magnetic Metallic Compounds
Fe_xCo_{1-x} magnetic exchange
Ni in Pd induced moments
Cu in Ni Konbu-phase

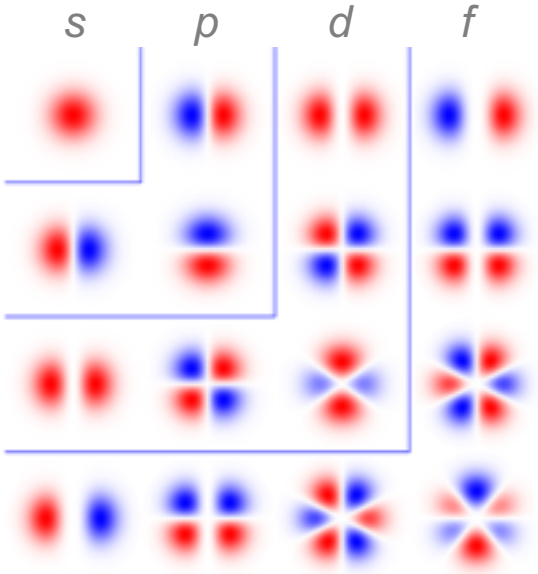


Dilute Magnetic Semiconductors
MgO:N magnetic exchange
GaN:Gd colossal moments

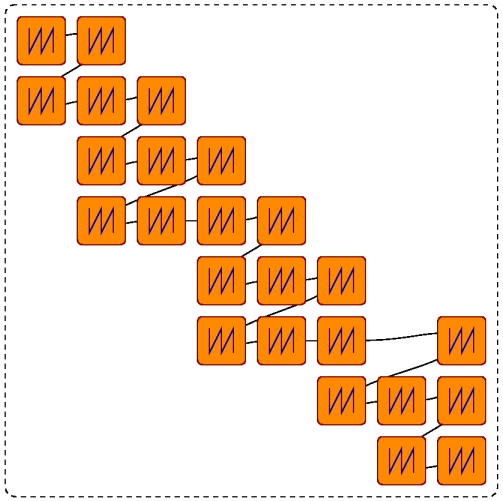
➔ Any sort of broken symmetry

KKRnano – Details I

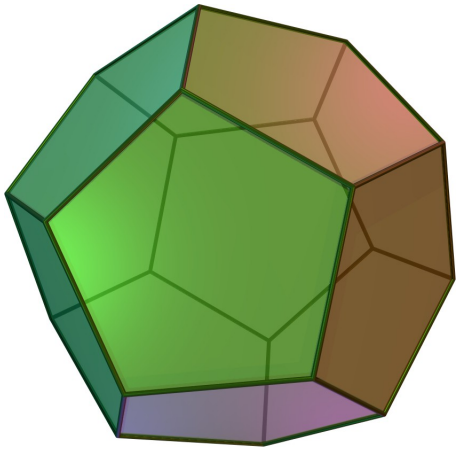




Radial exponential grids plus an expansion in b=16 spherical harmonics



Selected columns of the inverse of a block-sparse complex matrix are sought for

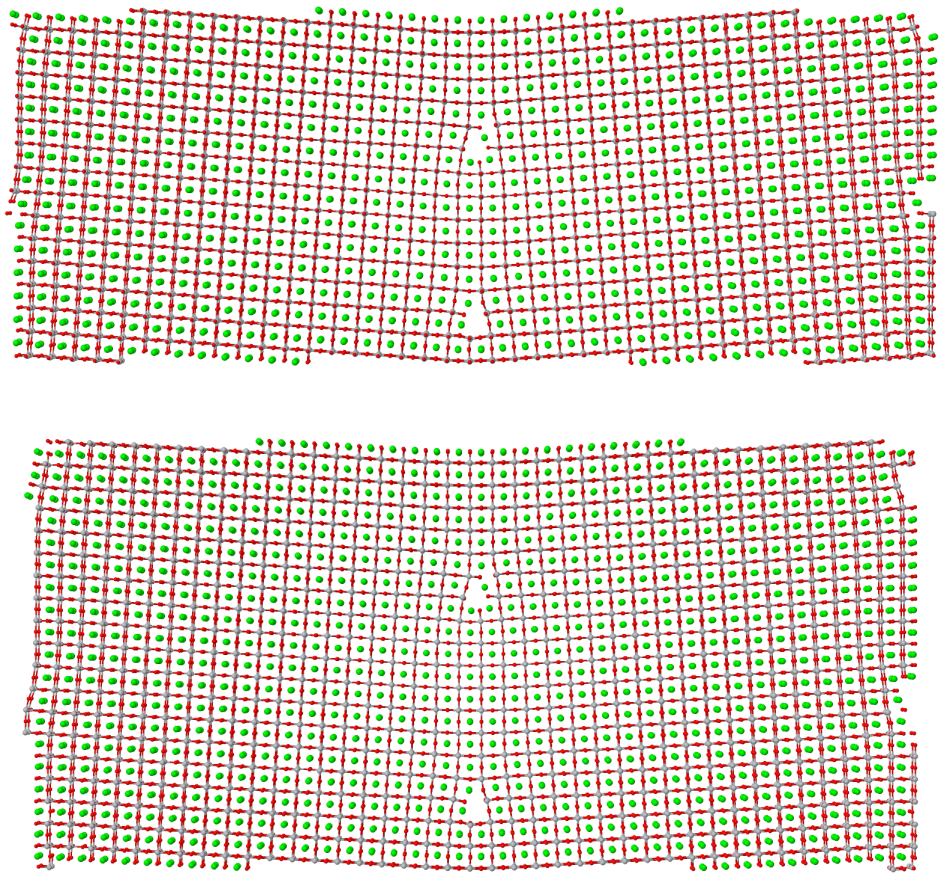


A Voronoi cell around each nucleus limits the radial grid

JÜLICH Forschungszentrum

KKRnano – Application: Grain boundaries

- Small angle grain boundaries in $SrTiO_3$
- Two different setups,
  7.6° → 2950 atoms
  6.0° → 3750 atoms
- Placement of empty cells necessary
- Lloyd's formula needed for a correction of E_{Fermi}



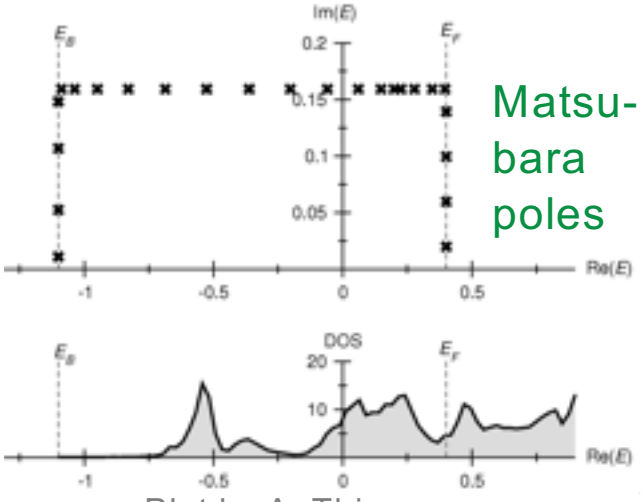
KKRnano – Details II

- Methodology from Korringa-Kohn-Rostoker multiple scattering formalism
- Energy contour integration in the complex plane
- Iterative inversion of a “Hamiltonian” matrix

$$(E - H) \times G == 1$$

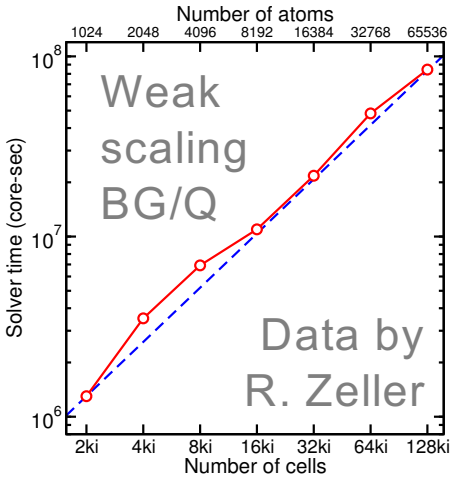
- Multiple right hand sides treated in G
  → Matrix × Matrix instead of Matrix × Vector
- Total energy and atomic forces, non-collinear magnetism, magnetic fields, density of states
- MPI and OpenMP parallel Fortran90 code
- Code generation (replacing C++ templates)

T = 25000 Kelvin



Matsu-bara poles

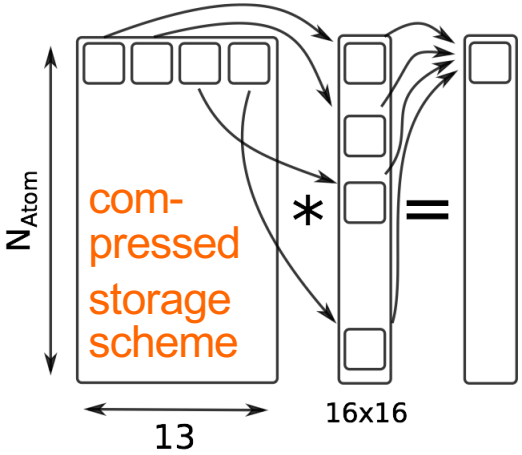
Plot by A. Thiess

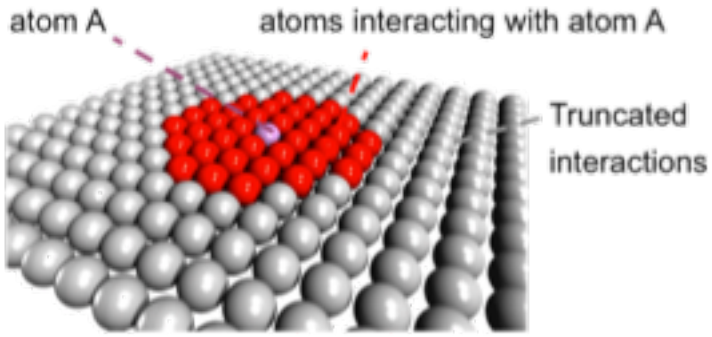


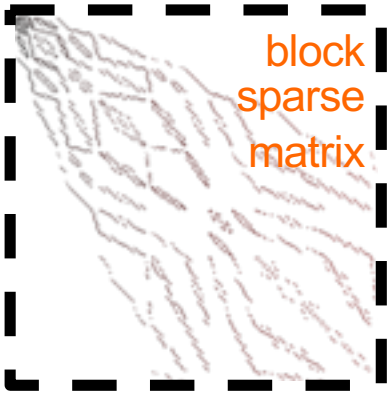
JÜLICH Forschungszentrum

KKRnano – Linear Scaling by Truncation

- Green function elements are truncated when $|R_i - R_j| > R_{cut}$
- Fast convergence for insulators (→ c.f. density matrix methods)
- Kohn-nearsightedness principle allows also metals
- Convergence of relevant quantities for $N_{trunc} \cong 1000$ atoms
- Cost scales as $N_{atoms} \times N_{trunc} \times$ Slowdown



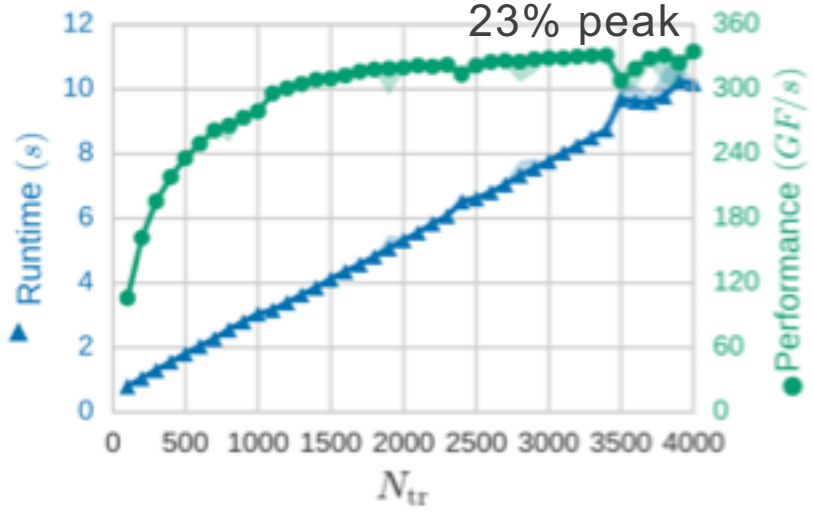


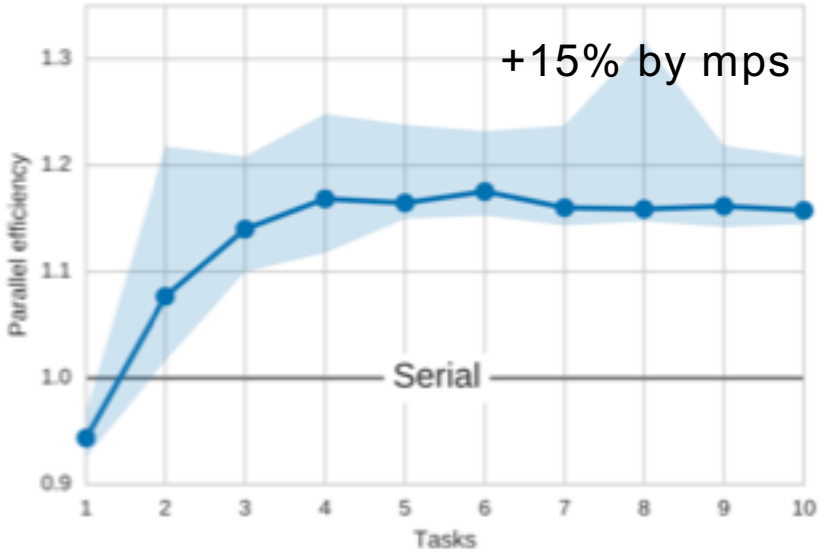


JÜLICH Forschungszentrum

KKRnano – GPU prototype

- Solver based on cuBLAS and cuSPARSE executed on NVIDIA K40m [1]





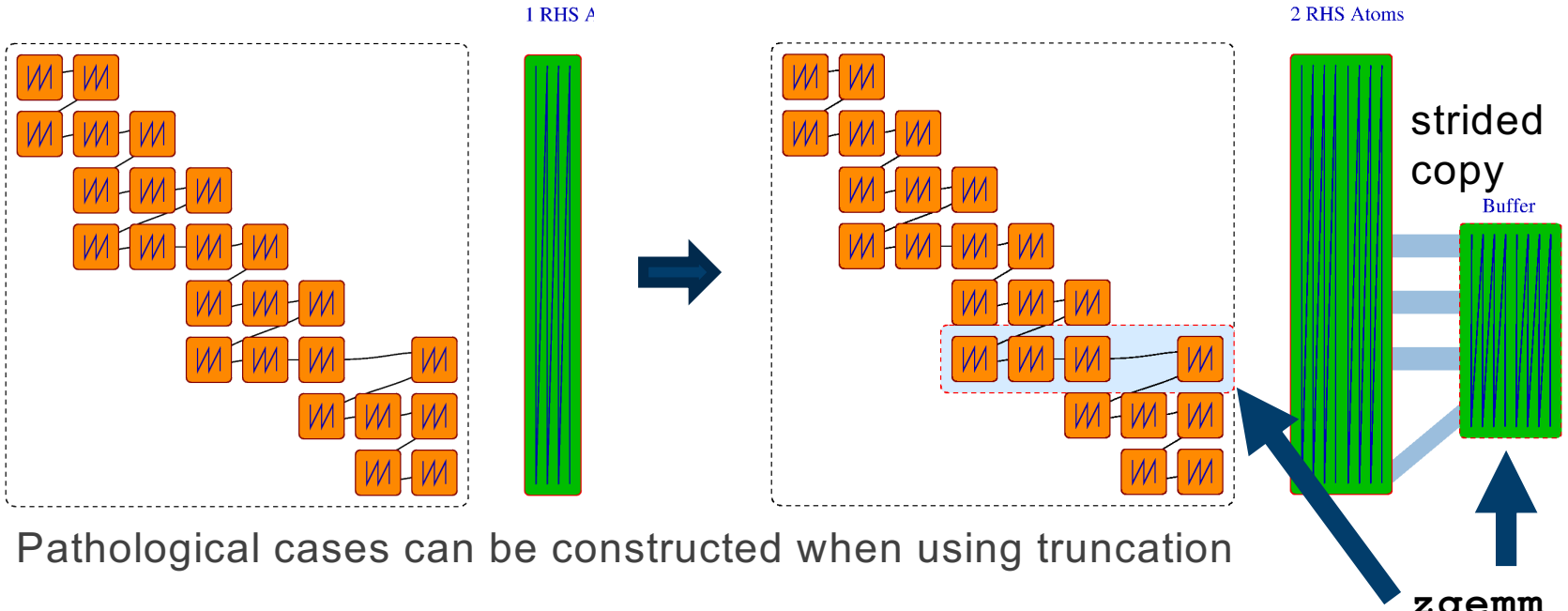
- Library not flexible enough, does not exploit the increased arithmetic intensity for more than 1 atom/MPI process → limited by memory bandwidth

[1] Addressing Materials Science Challenges Using GPU-accelerated POWER8 Nodes, P. F. Baumeister, M. Bornemann, M. Bühler, Th. Hater, B. Krill, D. Pleiter, R. Zeller, EuroPar 2016 proceedings

JÜLICH Forschungszentrum

KKRnano – Kernel Architecture

- Fast application of A × X required
- For the simplicity of bookkeeping, 1 atom per MPI process was fixed
- Needed to release constraint due to MPI overheads and arithmetic intensity

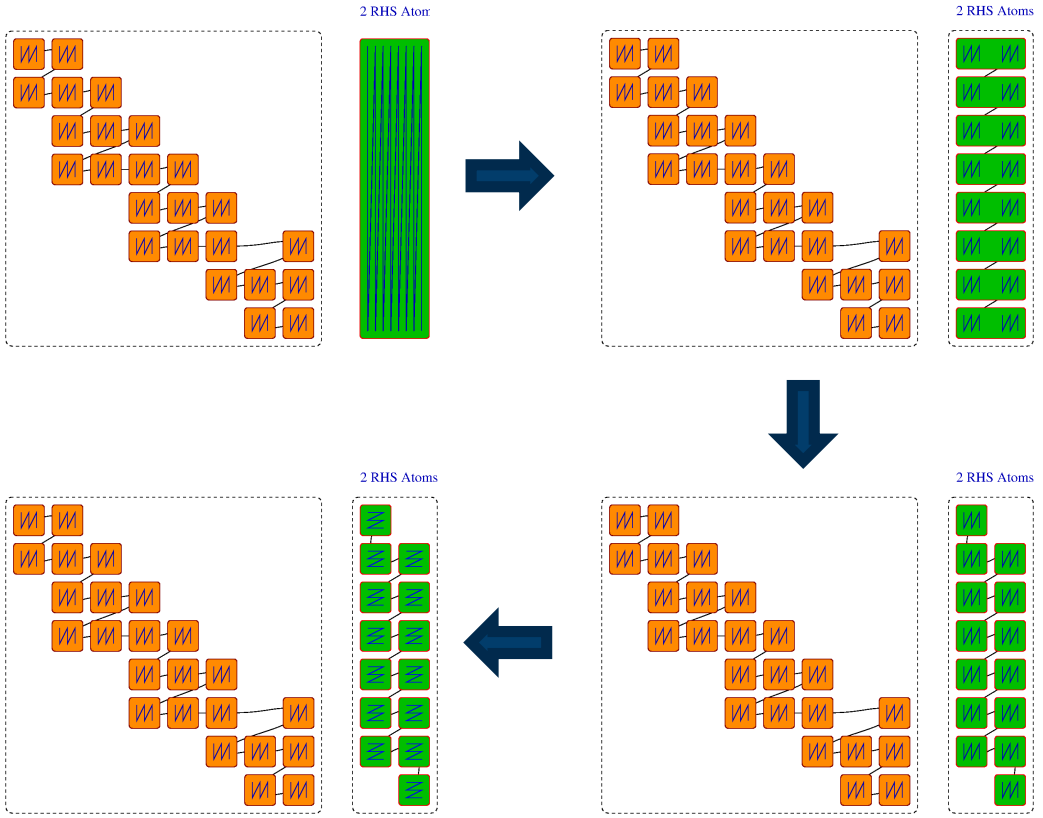


- Pathological cases can be constructed when using truncation

JÜLICH Forschungszentrum

KKRnano - Kernel Restructuring

- Truncation leads to a different view of the operator A for each RHS atom
- Split long vectors X into blocks
- Sparse storage of blocks of X in truncation mode
- Improved vectorization possible with transposed blocks of X → RowMajor



JÜLICH Forschungszentrum

KKRnano – New Kernel Architecture

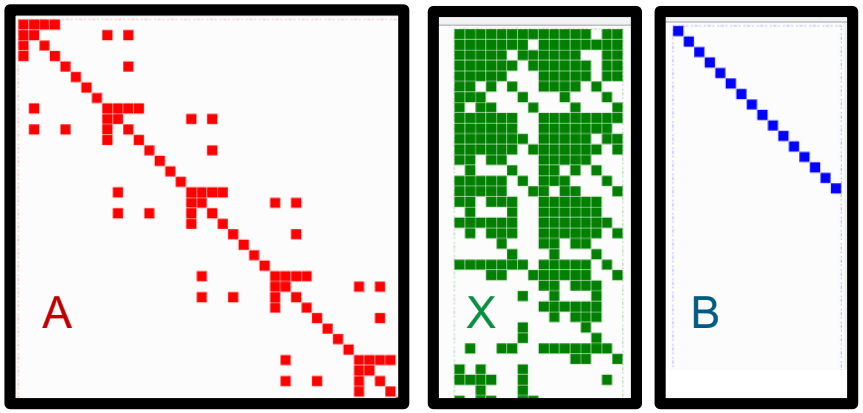
- Restructured KKRnano to use **B**lock**S**parse**R**ow representations for the matrices A, B and X in

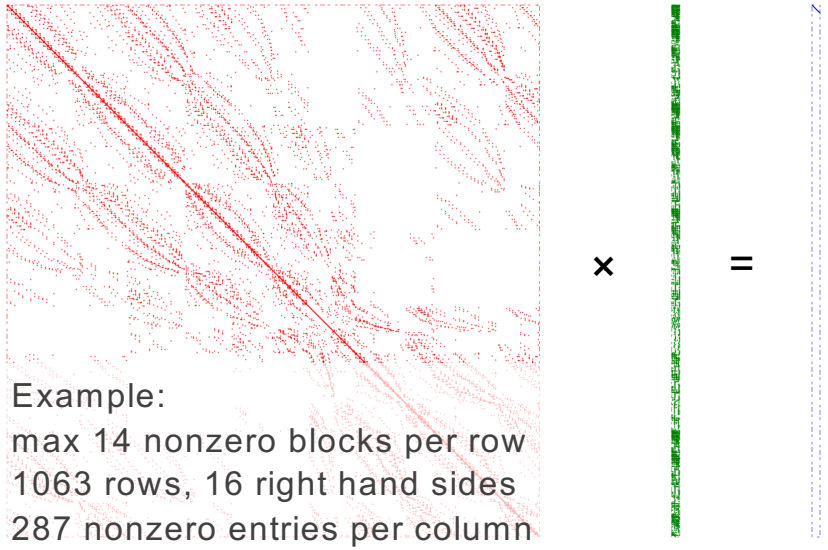
  $$A \times X - B = 0$$

- BSR-BSR multiplication Y = A × X planning performed ahead,

  shape(Y) == shape(X)

- Relies on the performance of **`zgemm`** for many small block matrices, usually 16×16 or 32×32 double precision complex, hand written version may outperform libraries





Example:
max 14 nonzero blocks per row
1063 rows, 16 right hand sides
287 nonzero entries per column

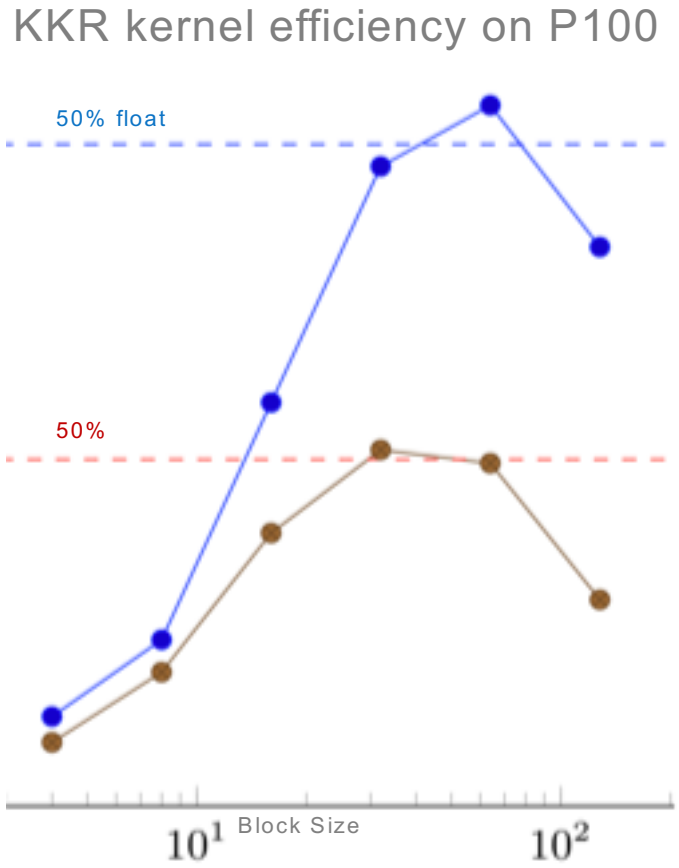
JÜLICH Forschungszentrum

Block-matrix times block-matrix on P100

- Small complex block matrices
- Scan performance for different block sizes

- Around 50% efficiency on P100
- Full tfQMR solver ~ 33% efficiency

- GPU kernel in CUDA
  needs interface from C++ to Fortran



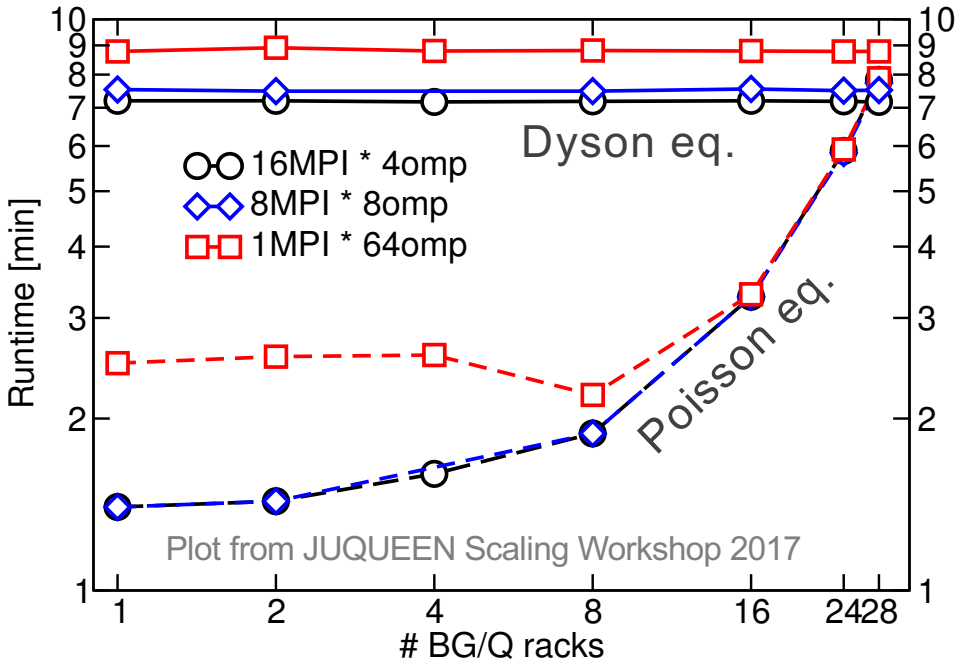
JÜLICH Forschungszentrum

Outlook I – Future Work

- For the Green function solver:
- Interface for the GPU solver
- ~~Improve KnL performance (< 20%)~~
- Preconditioning

- For other code parts:
- Multipole summation technique for electrostatics scale O(n^2_{atoms})
  → Fast multipole method?

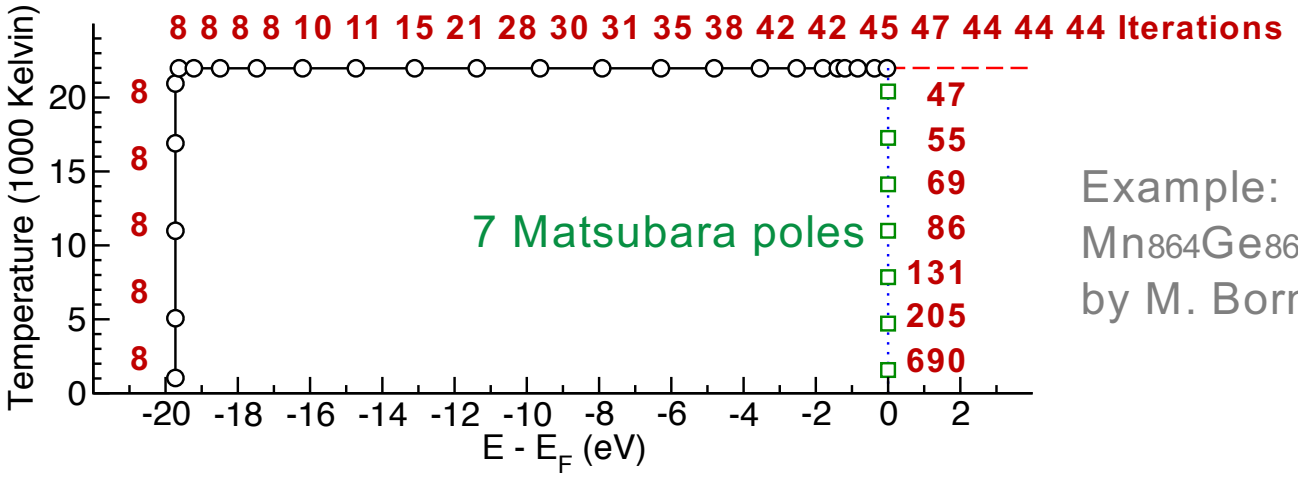
- I/O formats and routines



JÜLICH Forschungszentrum

Outlook II – Preconditioning

- KKRnano: State of the code: preconditioning is implemented, but no general efficient preconditioner available



Example:
$Mn_{864}Ge_{864}$ (B20)
by M. Bornemann

- Idea: Use different truncation radii, find short ranged preconditioner by solving with $R_{trunc} = R_{ref}$, then MPI distribute data

JÜLICH Forschungszentrum

Summary

- DFT application KKRnano based on Green‘s function
- Linear-scaling method for large problem sizes
- Kernel restructuring:
  lower memory capacity requirement
  exploits the memory BW better
- Block-spare times block-sparse multiplications ~50% on P100
- Interface between Fortran and GPU-solver required
- ToDo: Improvements of electrostatics and the preconditioner