LARGE SCALE ELECTRONIC STRUCTURE CALCULATIONS ON THETA

Performance optimization of WEST and Qbox

Huihuo Zheng[1], Christopher Knight[1], Giulia Galli[1,2],

Marco Govoni[1,2], and Francois Gygi[3]

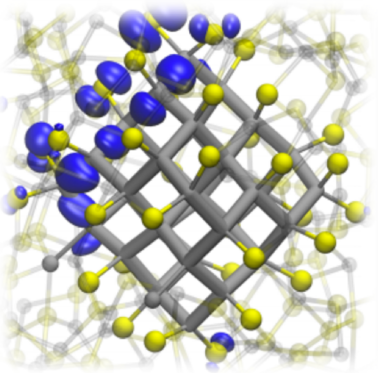
[1]Argonne National Laboratory

[2]University of Chicago

[3]University of California, Davis

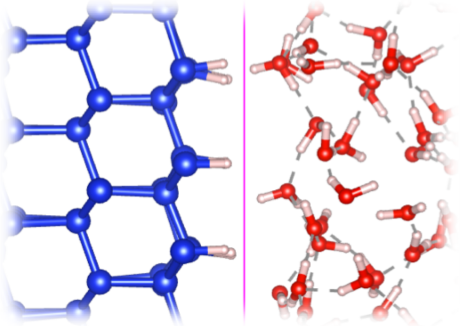
Feb 28th, 2018

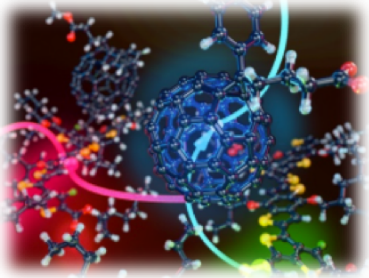
THETA ESP: FIRST-PRINCIPLES SIMULATIONS OF FUNCTIONAL MATERIALS FOR ENERGY CONVERSION



Embedded nanocrystal
T. Li, Phys. Rev. Lett. ***107****, 206805 (2011)*

Heterogeneous interfaces
H. Zheng, APS March meeting, 2018



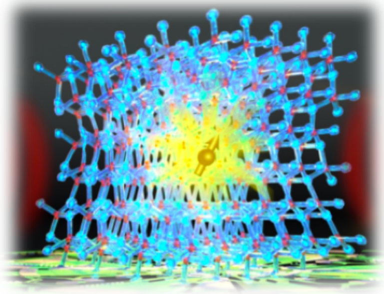


Organic photovoltaics
M. Goldey Phys. Chem. Chem. Phys., Advance Article (2016)

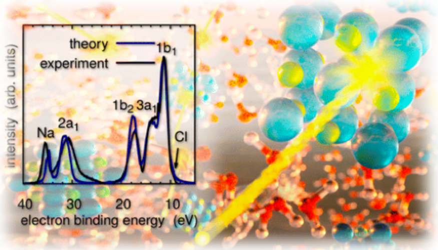
Qbox
First-Principles Molecular Dynamics

WEST!

QUANTUMESPRESSO



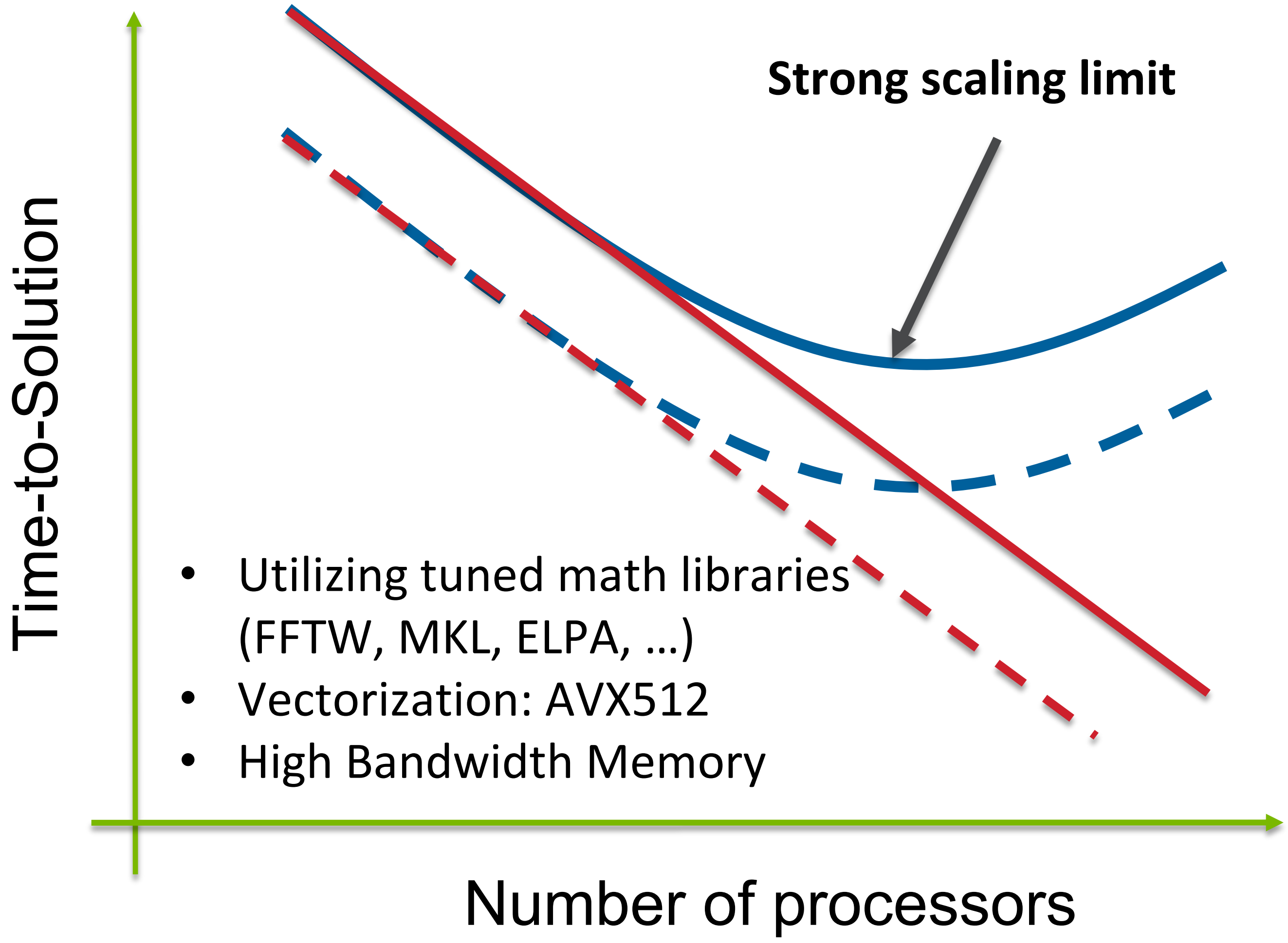
Quantum information
H Seo, Sci Rep. 2016; 6: 20803.



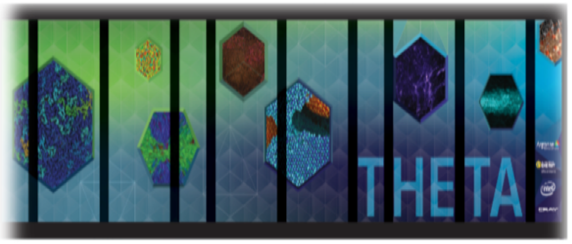
Aqueous solution
A.Gaiduk et al., J. Am. Chem. Soc. Comm. (2016)

http://qboxcode.org/; http://west-code.org; http://www.quantum-espresso.org/
M. Govoni, G. Galli, J. Chem. Theory Comput. 2015, 11, 2680−2696
P. Giannozzi, et al J.Phys.:Condens.Matter, 21, 395502 (2009)

PERFORMANCE OPTIMIZATION



- Utilizing tuned math libraries (FFTW, MKL, ELPA, …)
- Vectorization: AVX512
- High Bandwidth Memory



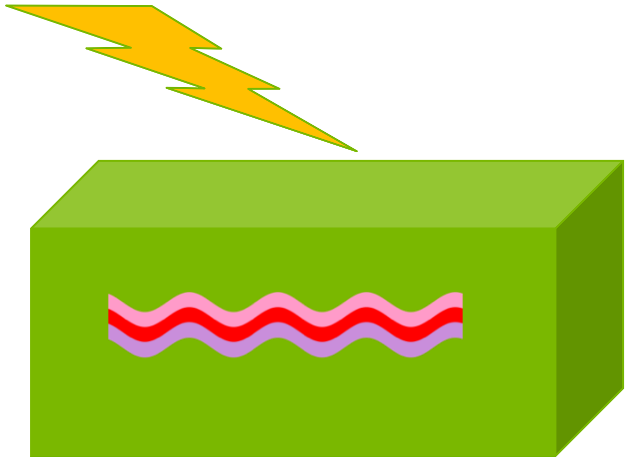
- Adding extra layers of parallelization -> increase intrinsic scaling limit
- Reducing communication overhead to reach the intrinsic limit

OUTLINE

- WEST – adding extra layers of parallelism
  - Addressing bottleneck from I/O
  - Implementing band parallelization
- Qbox – reducing communication overheads of dense linear algebra with on-the-fly data redistribution
  - Gather & scatter remap
  - Transpose remap
- Conclusions and insights

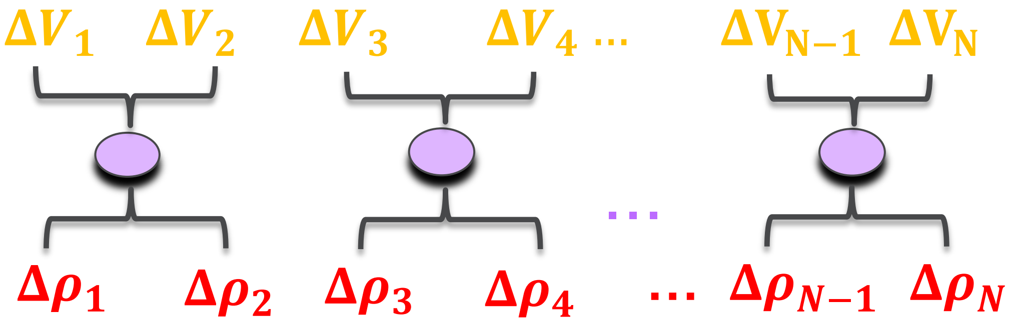
OPTOELECTRONIC CALCULATIONS USING MANY-BODY PERTURBATION THEORY (GW)

Linear response theory



$$\Delta\rho = \chi\,\Delta V_{pert}$$

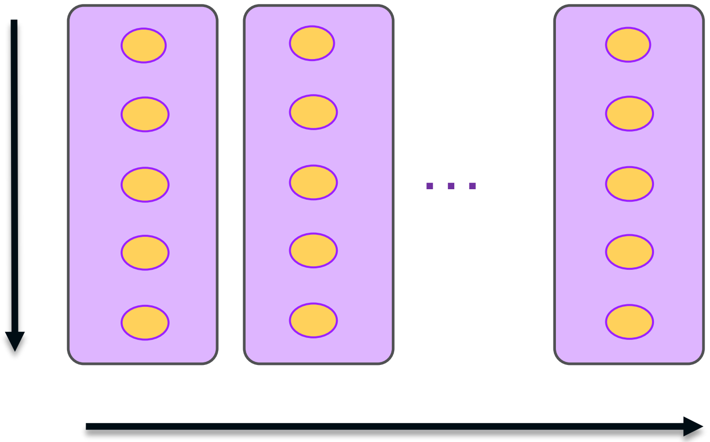
Electronic density — **Response function** — **Perturbation potential**



Massively parallel by distributing perturbations

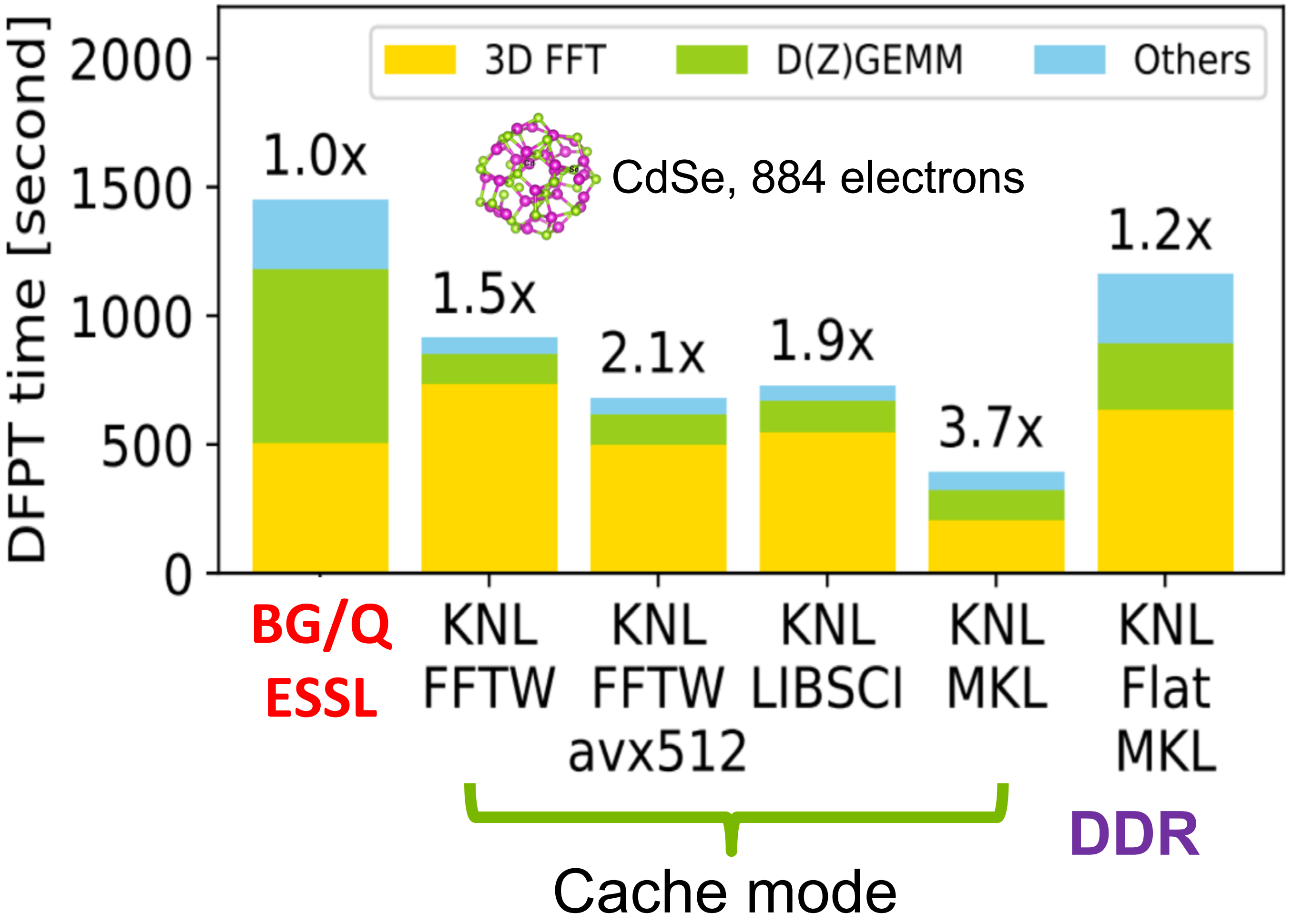
Parallelization scheme (image & plane wave)

3D FFTs + D(Z)GEMM



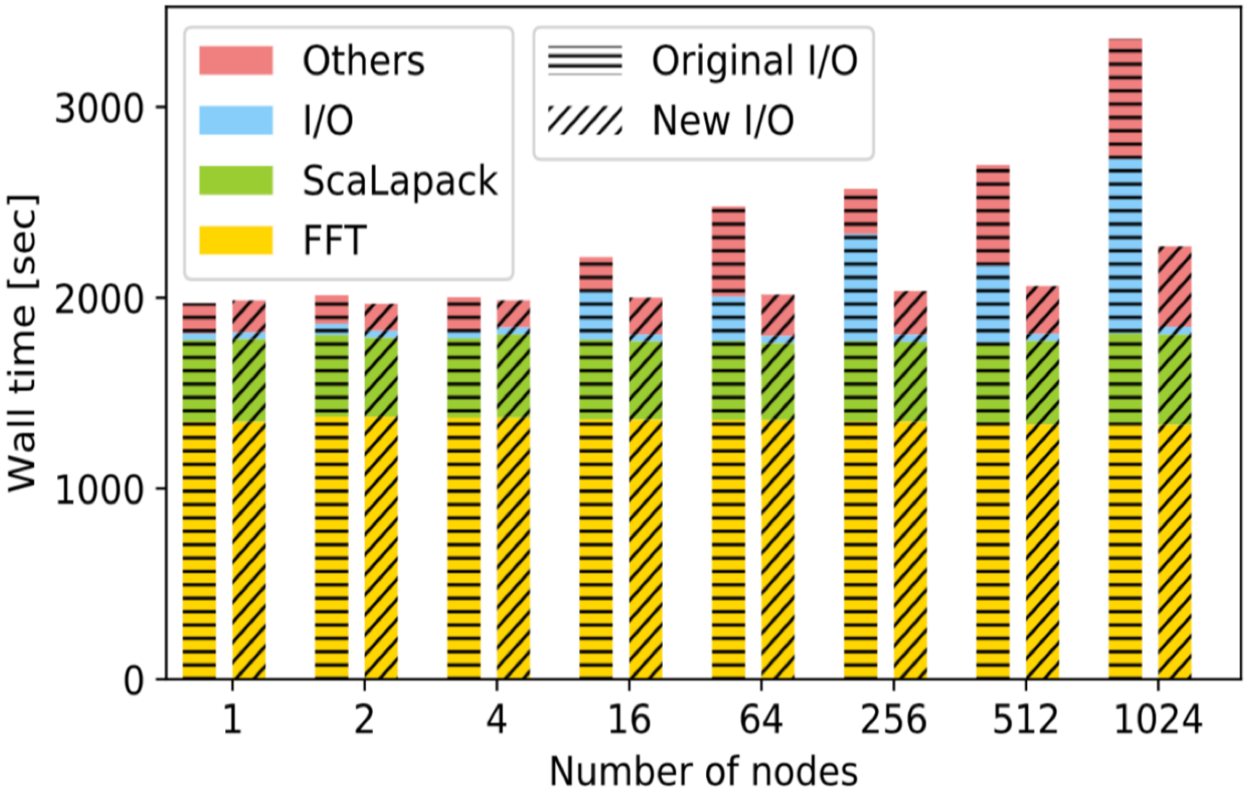
Matrix diagonization (syev, heev, elpa)

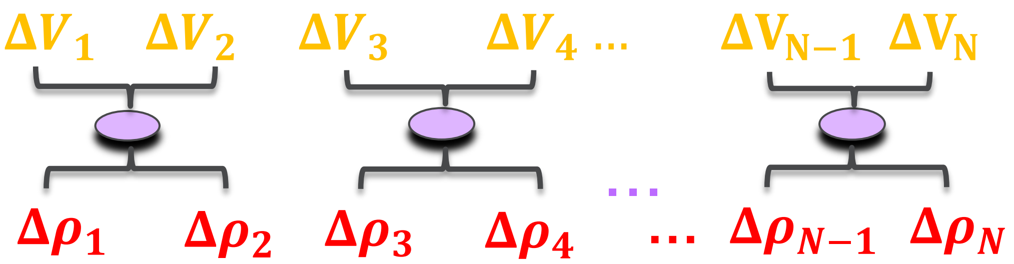
SINGLE NODE RUNTIME ON THETA IN COMPARISION WITH MIRA (1KNL VS 4BG/Q)

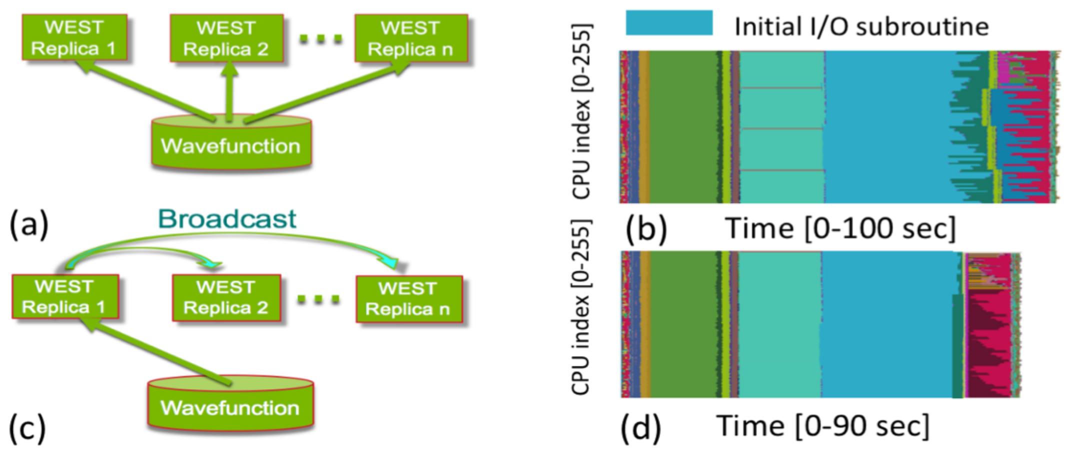


- **80% of runtime is spent in external libraries**
- 3.7x speedup from BG/Q(ESSL) to KNL(MKL)
- High-bandwidth memory on Theta critical for performance (e.g. 3D FFTs): 3.1x speedup

I/O ISSUE APPEARED IN WEAK SCALING STUDY

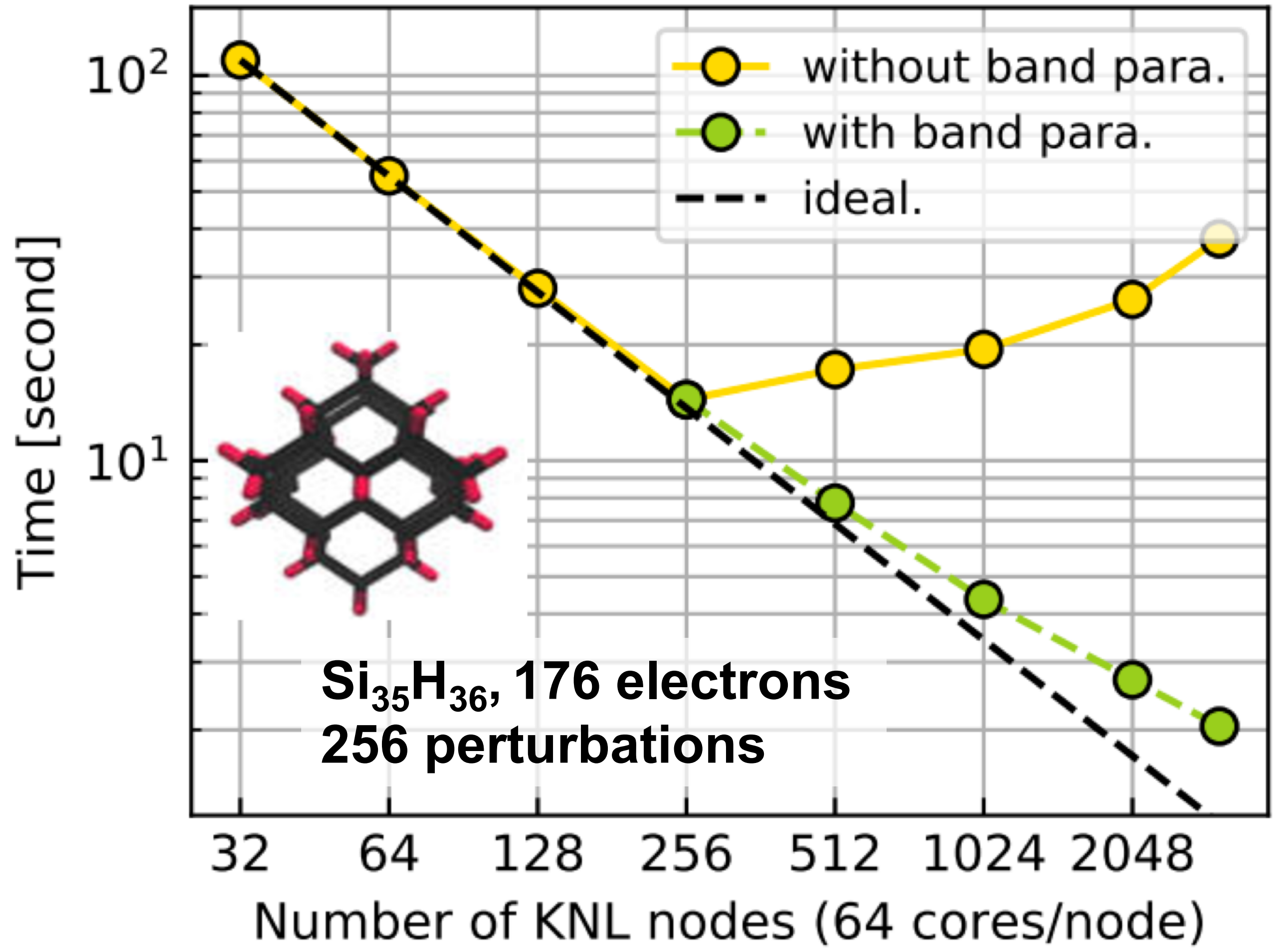


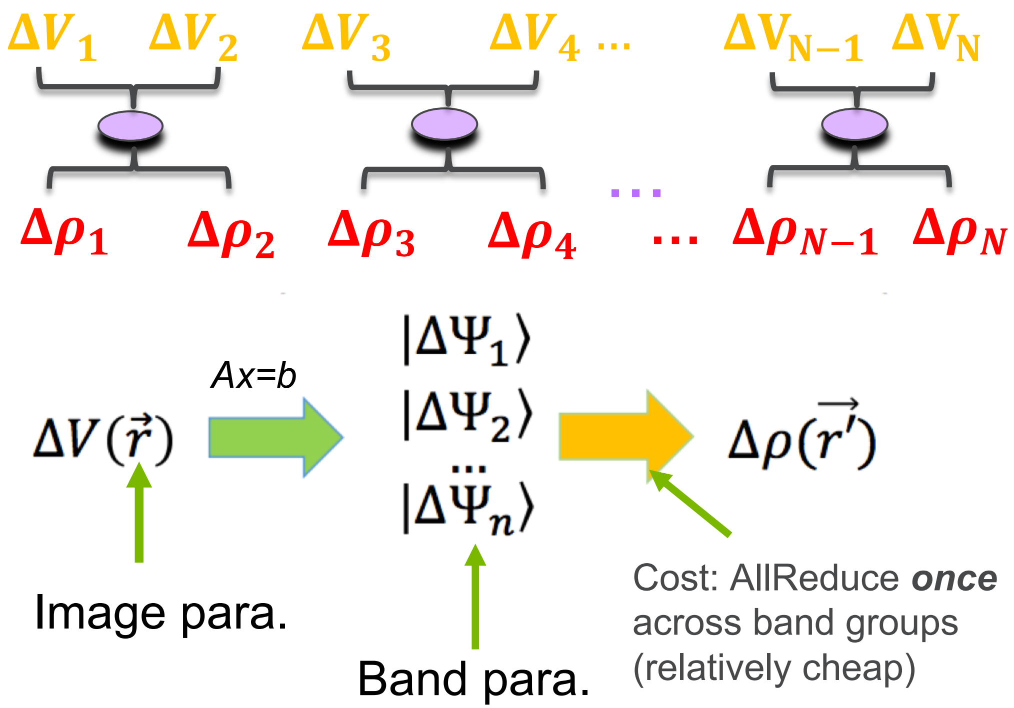




- Original I/O scheme: all replica read the same file; I/O time increased with number of nodes becoming a significant fraction of runtime.
- Time spent in I/O reduced to negligible fraction of runtime on 1-1024 nodes by having master process read and distribute wave function.

IMPROVEMENT OF STRONG SCALING BY BAND PARALLELIZATION – A PATHWAY TO A21





Increased parallelism by arranging the MPI ranks in a 3D grid (perturbations & bands & FFT)

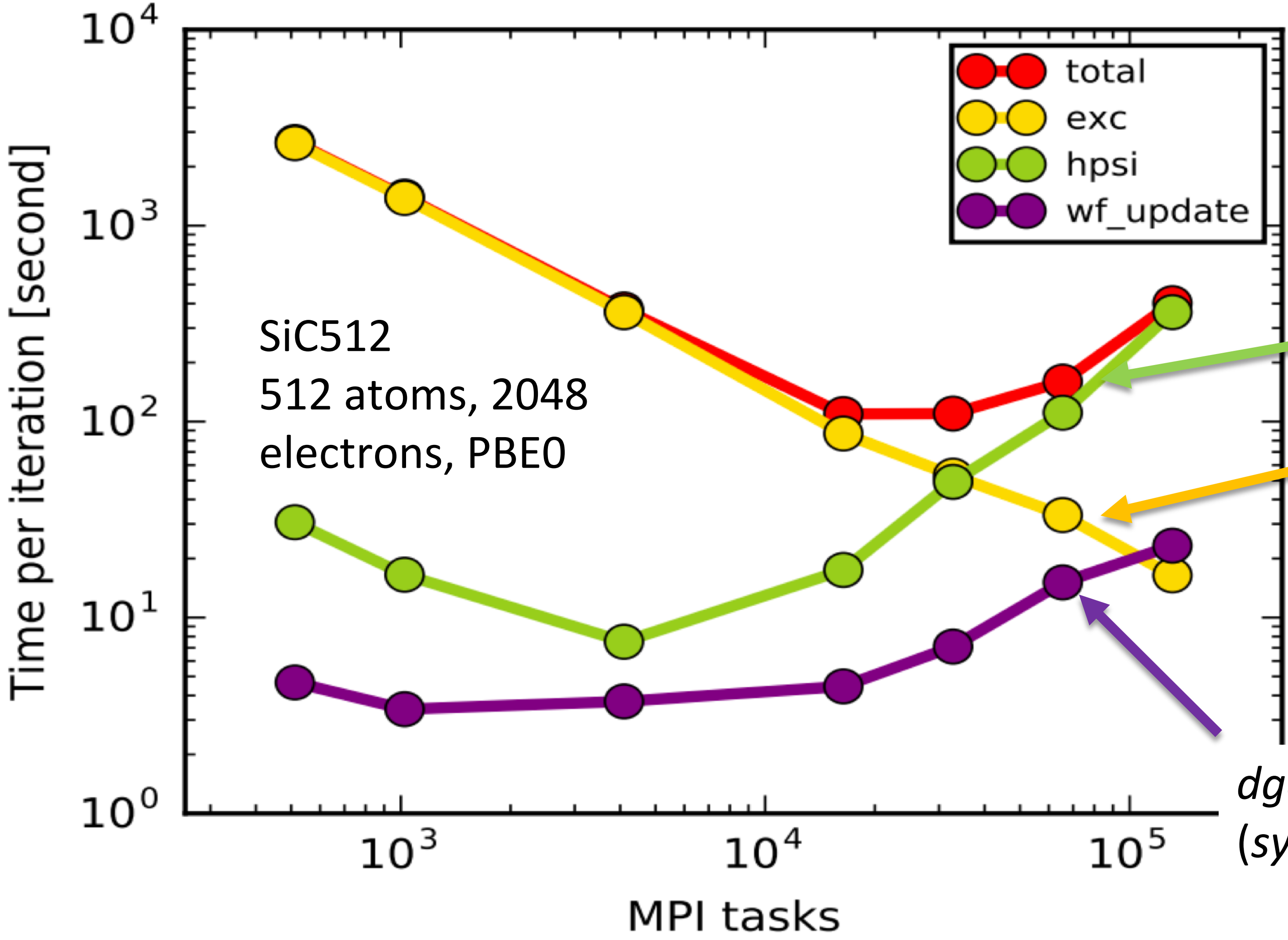
New intrinsic strong scaling limit:

$$nproc = N_{pert} \times \mathbf{N_{band}} \times N_z$$

QBOX

SCALING HYBRID DENSITY FUNCTIONAL CALCULATIONS

STRONG SCALING ANALYSIS OF QBOX FOR HYBRID-DFT CALCULATIONS



dgemm

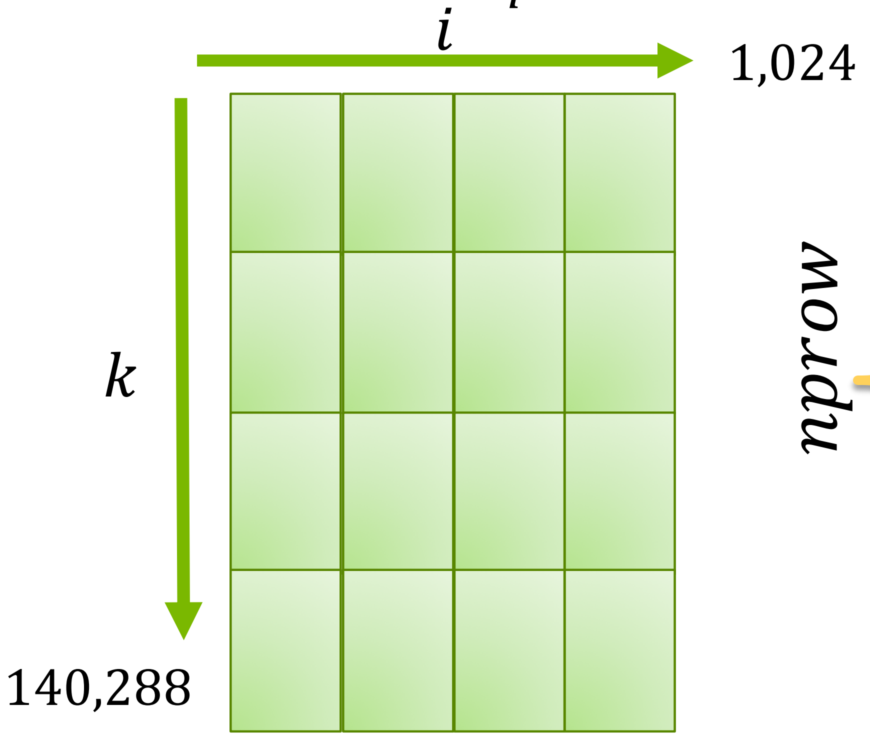
Exact exchange
3D FFTs
$N_{elec}^2 \sim 10^6 - 10^7\ FFTs$
Grid: 256×256×256

dgemm, Gram-Schmidt
(*syrk, potrf, trsm*)

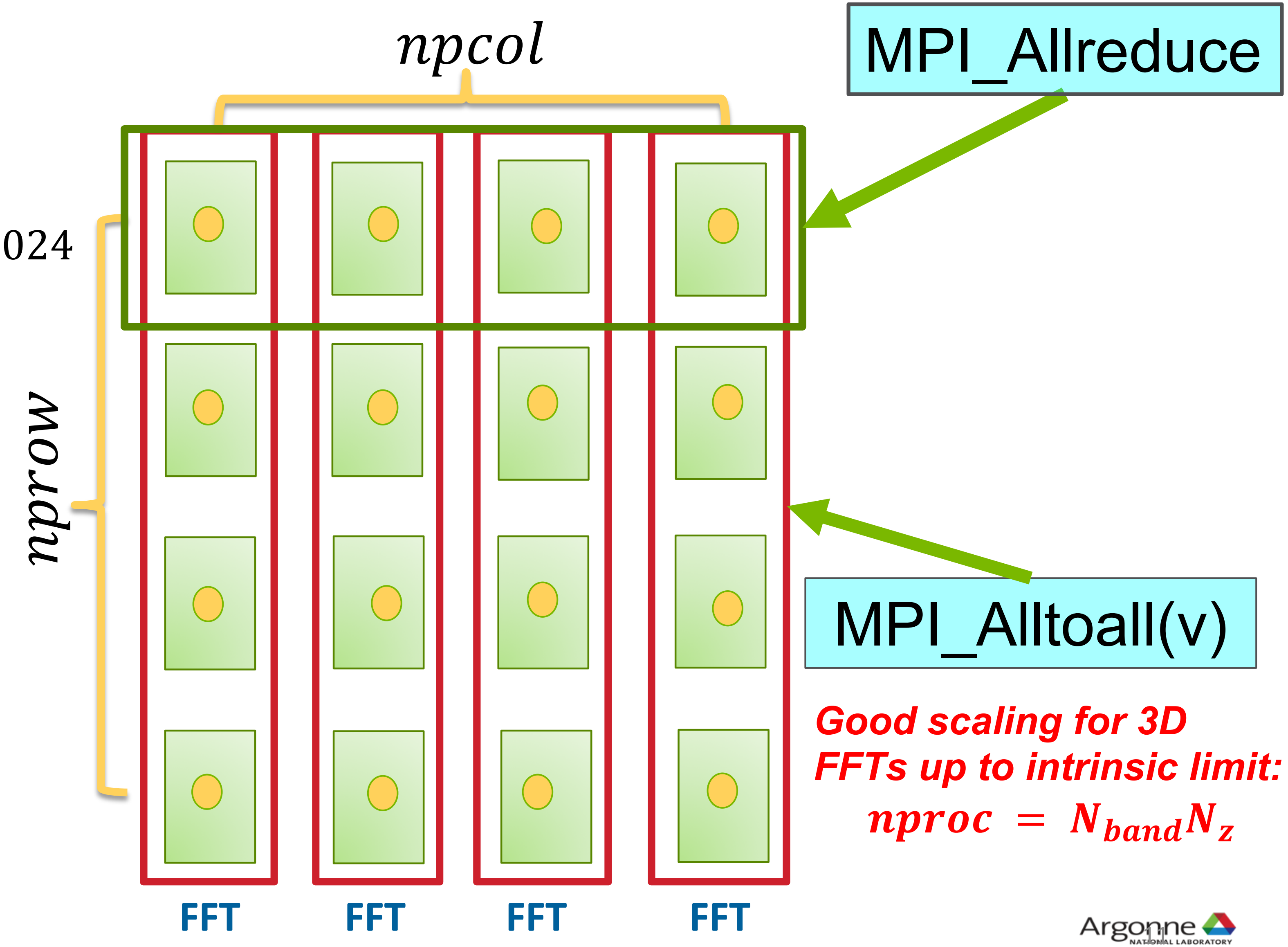
F. Gygi and I. Duchemin *J. Chem. Theory Comput.*, 2013, *9* (1), pp 582–587

DATA LAYOUT: BLOCK DISTRIBUTION OF WAVE FUNCTIONS TO 2D PROCESS GRID

$\psi_i(k), i = 1, 2, \ldots N_{band},$
$k = 0, 1, \ldots, n_{pw} - 1$

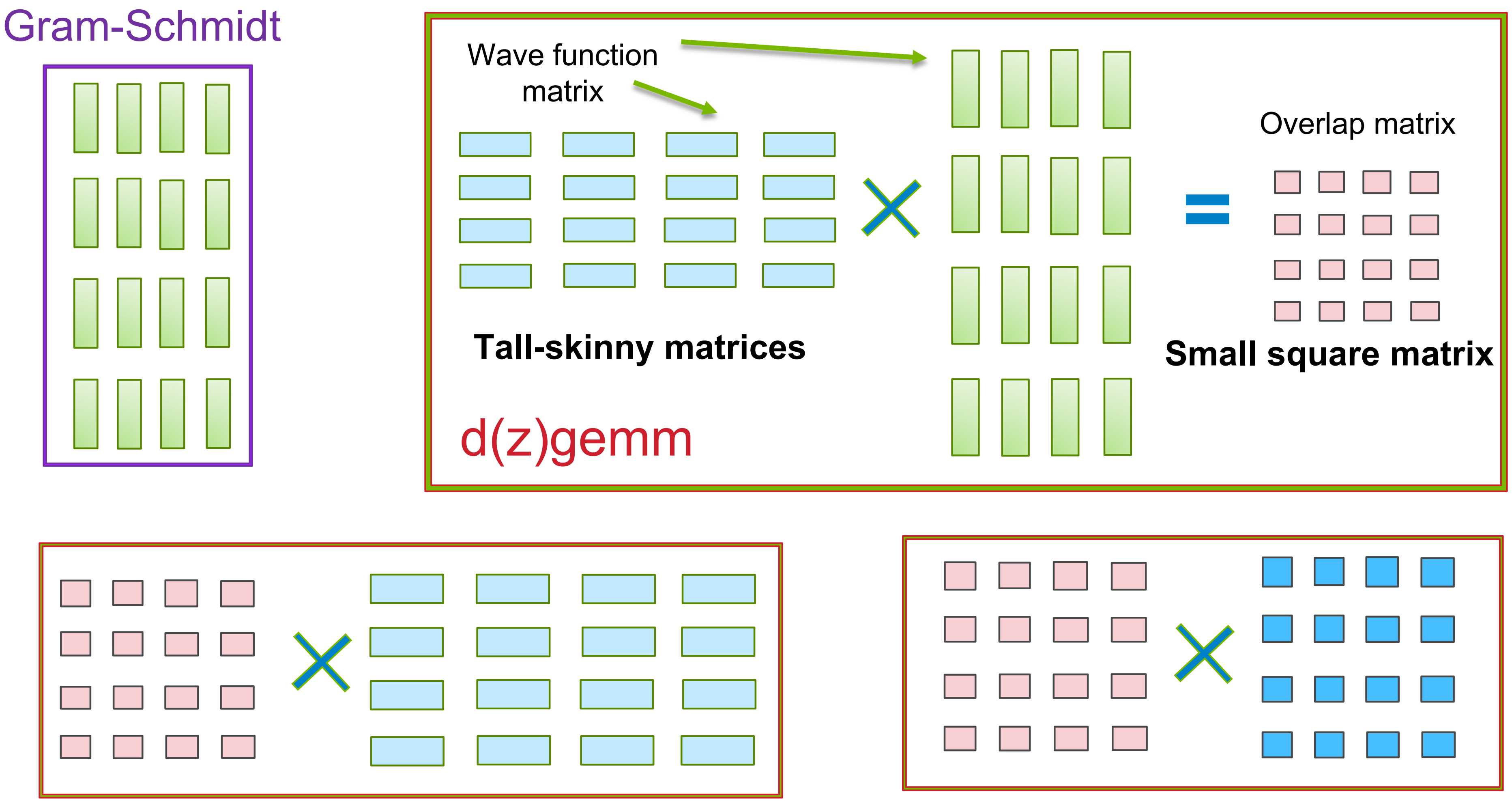


SiC512: 140,288 × 1,024

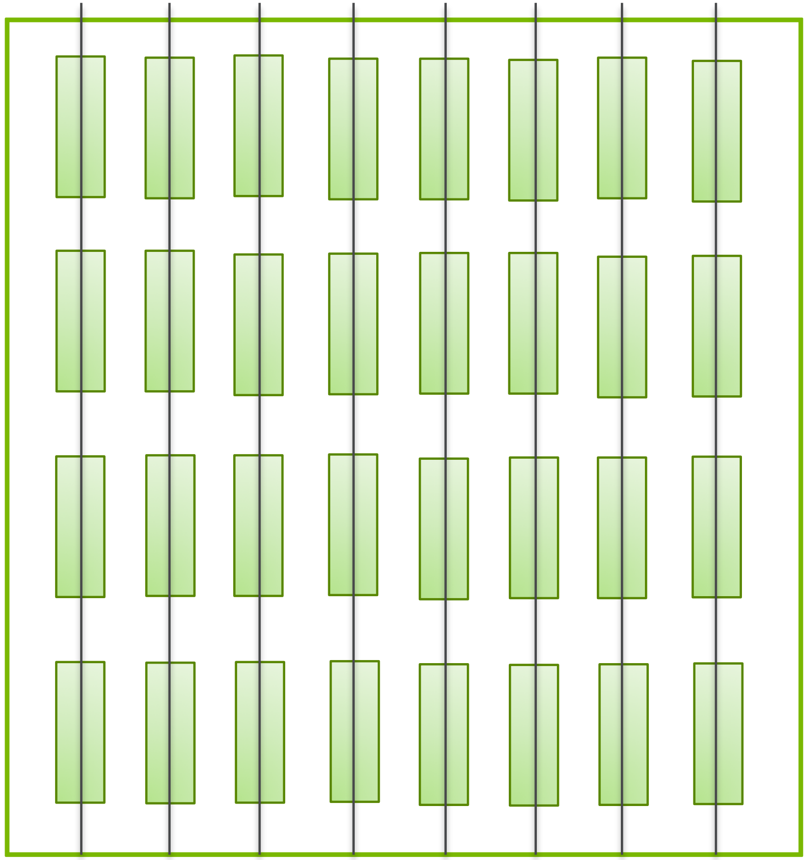


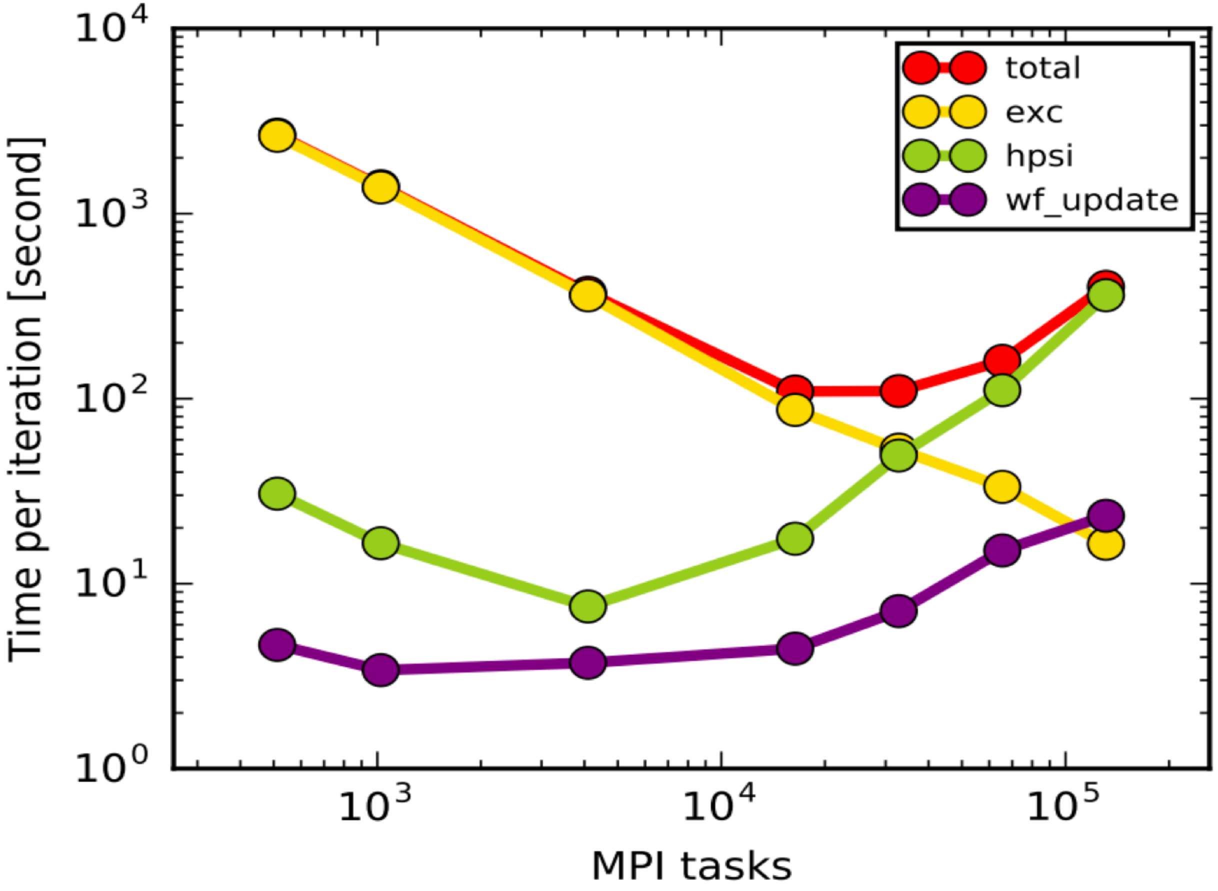
Good scaling for 3D FFTs up to intrinsic limit:
$nproc = N_{band} N_z$

DENSE LINEAR ALGEBRA INVOLVED FOR TALL-SKINNY MATRICES AND SMALL SQUARE MATRICES

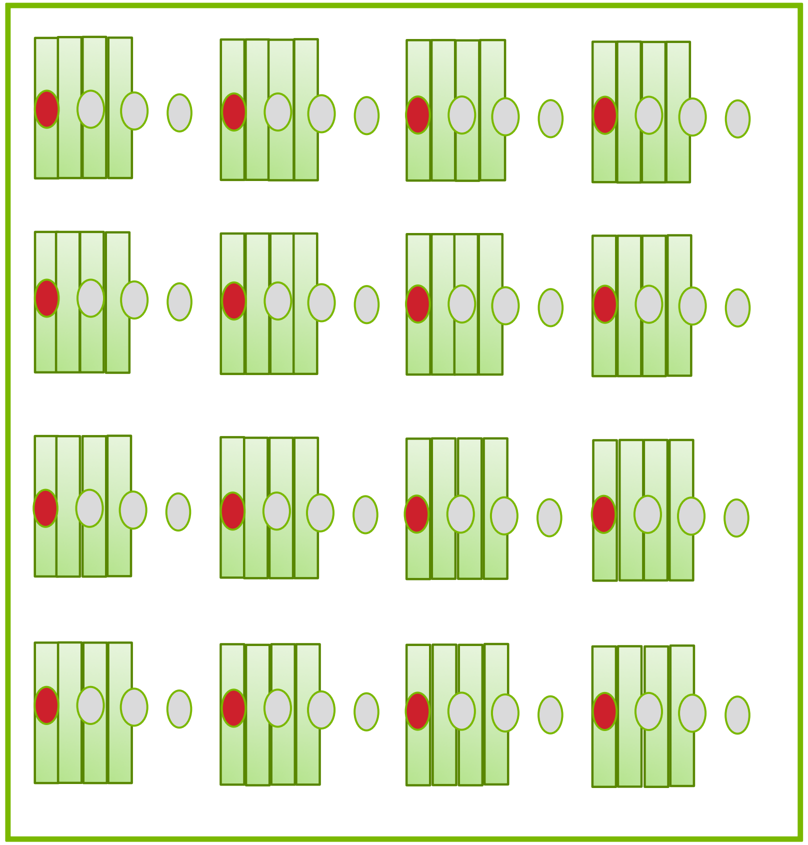


INCREASING OF COMMUNICATION OVERHEAD FROM SCALAPACK SUBROUTINES





REDUCING COMMUNICATION OVERLAP BY ON-THE-FLY REDISTRIBUTING DATA WITH REMAP METHOD



Increasing npcol →

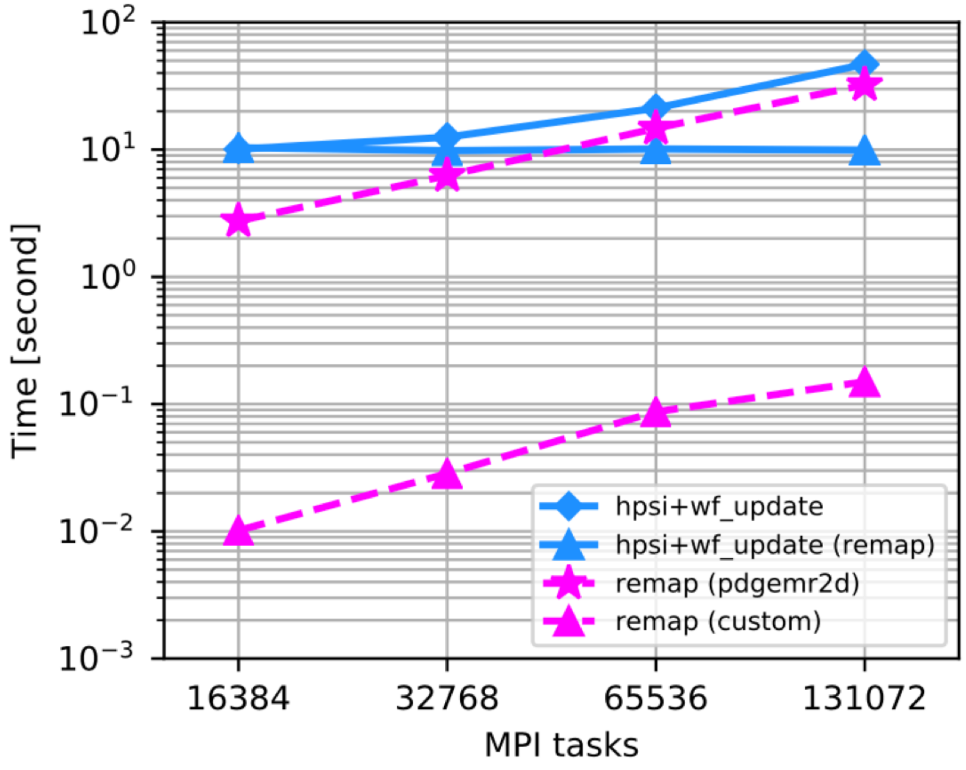
- local computing time decreases,
- communication time increases → Performance degradation

Solution: let a smaller group of processors do ScaLAPACK

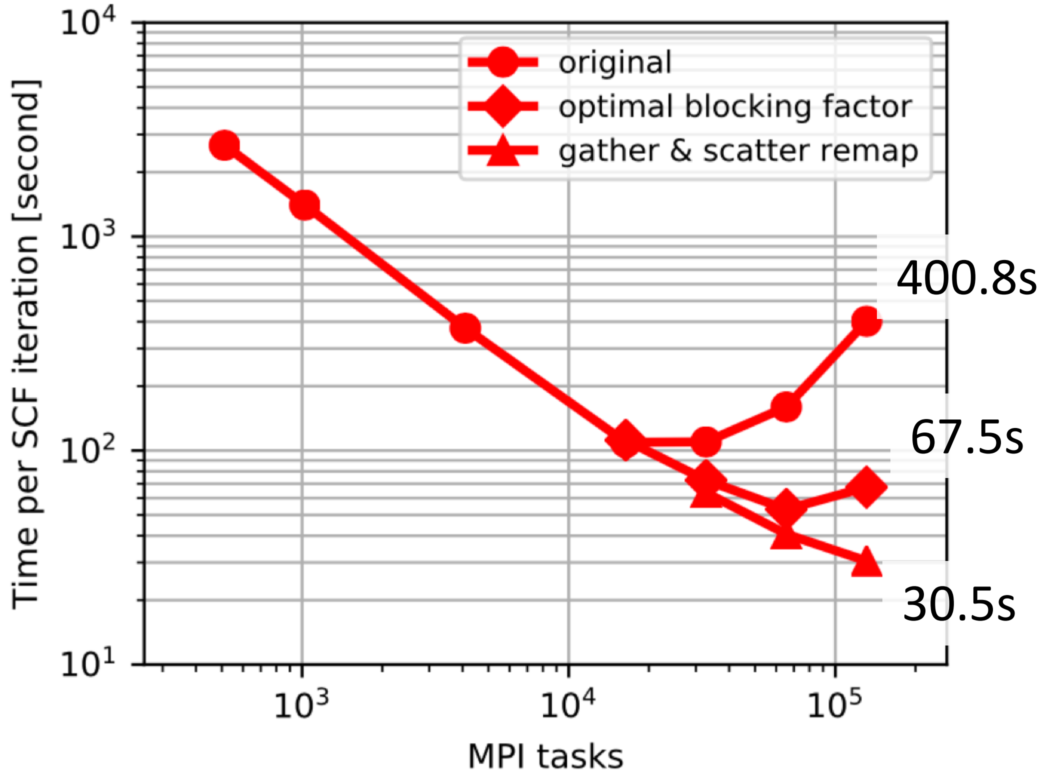
- Do FFT on the original grid
- Gather data to the smaller grid
- Do ScaLAPACK on the smaller grid
- Scatter data back to original grid

Remapping time (gather + scatter) should be small.

IMPROVEMENT OF STRONG SCALING USING "GATHER & SCATTER" REMAP



hpsi + wf_update time remains minimal relatively flat with remap, and the remap time (custom) is two orders of magnitude smaller than ***hpsi + wf_update* time**.



Improvement of Qbox's strong scaling after optimizations; runtime of improves from ~400 to ~30s per SCF iteration (13x speedup) on 131,072 ranks for 2048 electrons.

Custom remap function is 1000x faster than ScaLAPACK's pdgemr2d.

FURTHER IMPROVEMENT OF DGEMM RUNTIME BY “TRANSPOSE” REMAP

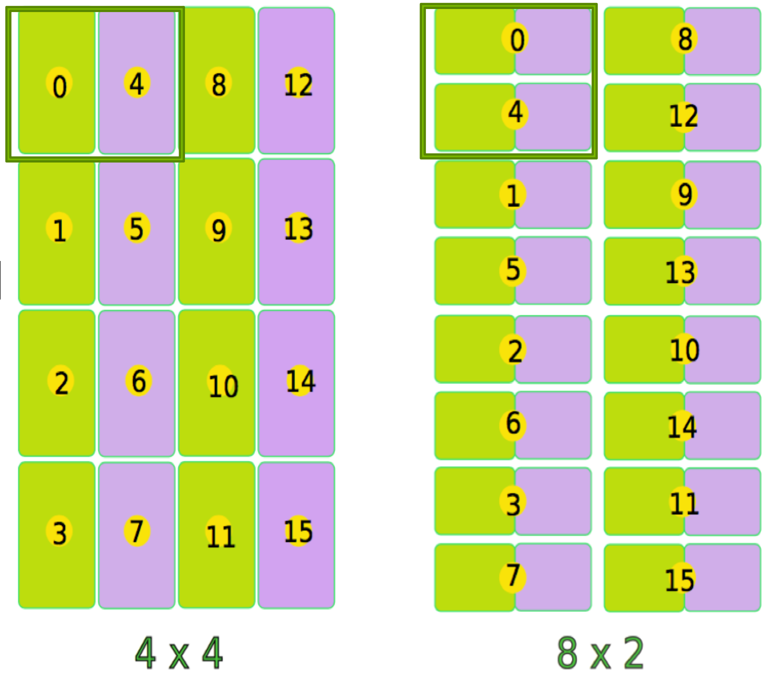
Problem of “gather & scatter”:
Idle processes.
How to utilize them? Assign idle processes to active columns.

Transpose remap:

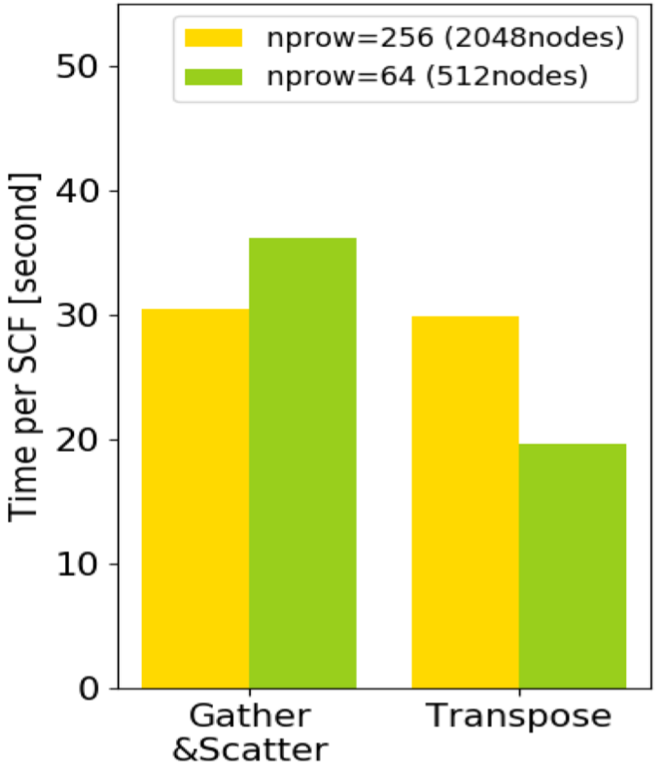
- Perform 3D FFTs in the original context.
- Transfer data through a series of **local regional transposes**
- Run ScaLAPACK in the new context

Key concept for remap: creating different contexts that are optimal for different kernels redistributing the data on-the-fly

Transpose communication pattern



Process rearrangement and data movement of transpose remap



Improvement of runtime by remap methods

$$(1)\ npcol' = \frac{npcol}{8},\ nprow' = nprow$$

$$(2)\ npcol' = \frac{npcol}{8},\ nprow' = 8 \times nprow$$

CONCLUSION AND INSIGHTS

- Band parallelization reduces the internode communication overhead and improves strong scaling of WEST up to $N_{\mathrm{FFT}} N_{pert} N_{band}$ cores.
- Optimal remapping of data for matrix operations in Qbox reduces ScaLAPACK communication overhead at large scale, and makes hybrid- DFT calculation scale to $N_{\mathrm{FFT}} N_{band}$ cores.
- Given the increased computational performance relative to network bandwidths, it is crucial to reduce and/or hide inter-node communication costs.

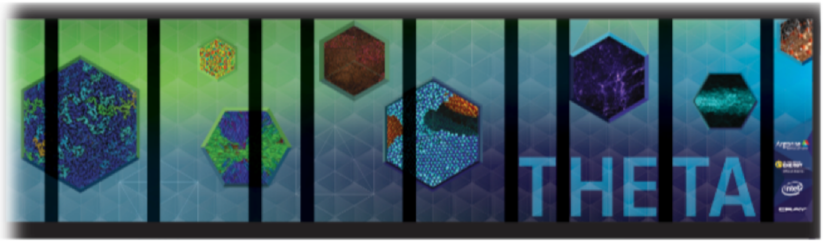
Guiding principles for developing codes in many-core architecture:

1) Fixing non-scalable bottleneck (e.g., Parallel I/O)
2) Parallelizing independent, fine-grain units of work, reducing inter-node communication, and maximizing utilization of on-node resources.
3) Optimizing data layout: optimizing communication patterns for performance critical kernels with on-the-fly data redistribution and process reconfiguration.

ACKNOWLEDGEMENT

- This research is part of Theta Early Science Project at Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility under Contract DE-AC02-06CH11357.

- This work was supported by MICCoM, as part of Comp. Mats. Sci. Program funded by the U.S. DOE, Office of Science, BES, MSE Division.



U.S. DEPARTMENT OF ENERGY

Office of Science

MICCoM

THANK YOU!