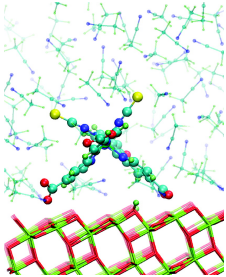
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich
Department of Materials

Accelerated Sparse Matrix Multiplication for Quantum Chemistry with CP2K on Hybrid Supercomputers

Ole Schütt
ole.schuett@mat.ethz.ch
Nanoscale Simulations

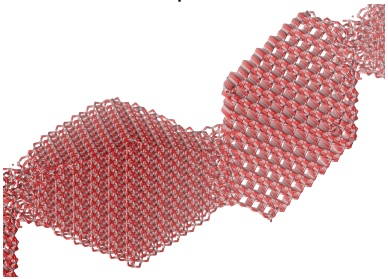
Application: Emerging Photovoltaics

Processes at TiO_2-Interface



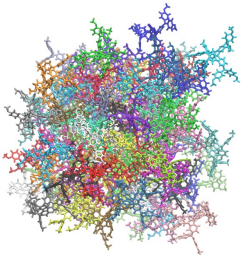
Schiffmann et al. (2010)

Electron Transport across Nanoparticles



17k atoms, 80k electrons

Hole Transporting Material (HTM)

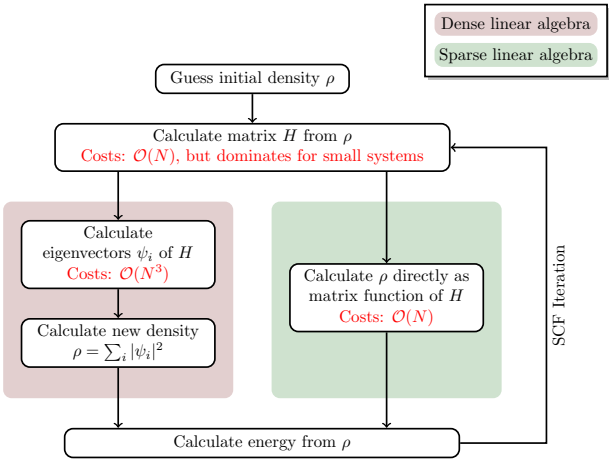


spiro-MeOTAD

Requirements

- electronic properties $\Longrightarrow$ Schrödinger equation ($H\Psi = E\Psi$)
- lack of symmetries $\Longrightarrow$ large simulation cells (> 1000 atoms)

Linear Scaling Self Consistent Field



- Density P as matrix function of P:

$$P = \left[1 + \exp\left(\frac{H - \mu\mathbb{1}}{kT}\right)\right]^{-1}$$
$$= \frac{1}{2}\left[1 - \text{sign}\left(H - \mu\mathbb{1}\right)\right]$$

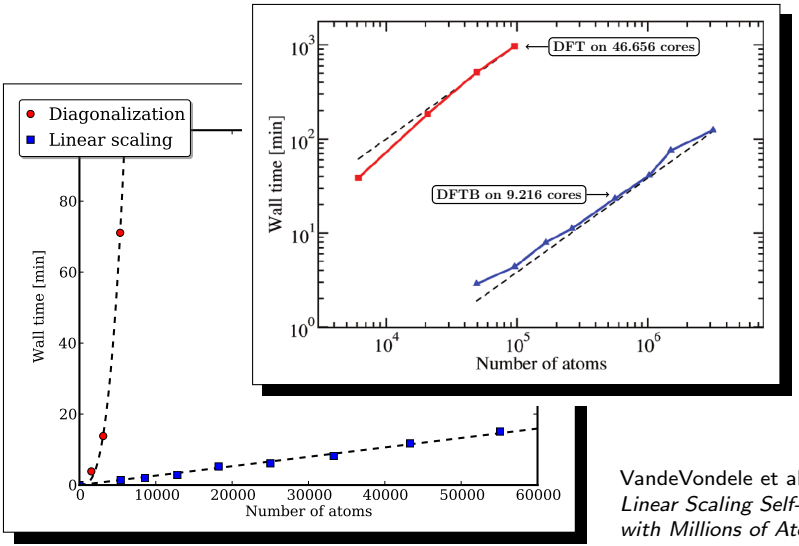
limit of small kT (ground state)

- Evaluate sign() as polynomial series:

$$X_0 = A \cdot ||A||^{-1}$$
$$X_{n+1} = \frac{1}{2} X_n (3\mathbb{1} - X_n^2)$$
$$\text{sign}(A) = X_\infty$$

LS-SCF entirely based on sparse linear algebra.

Benchmarks of Condensed Phase Systems

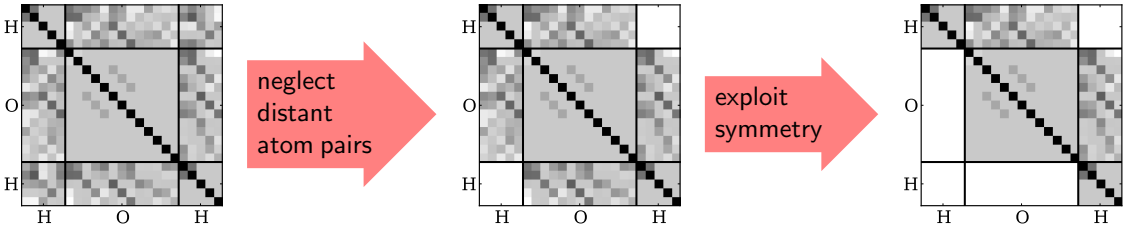


$\mathcal{O}(N)$ methods are inevitable for large systems

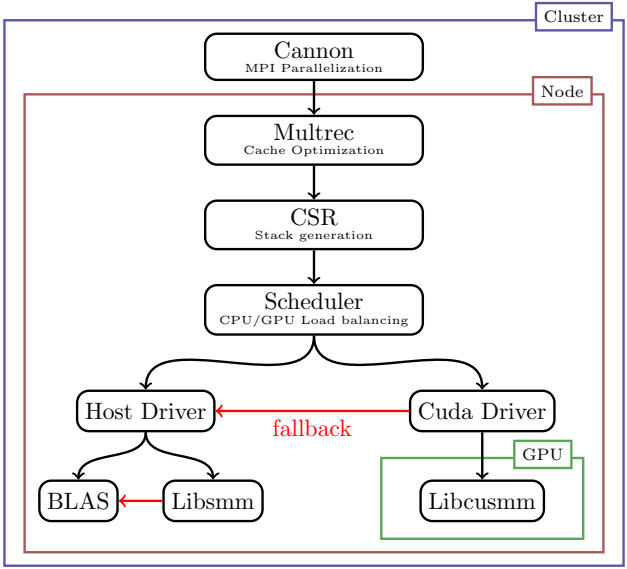
VandeVondele et al. (2012):
Linear Scaling Self-Consistent Field Calculations with Millions of Atoms in the Condensed Phase

The DBCSR Library

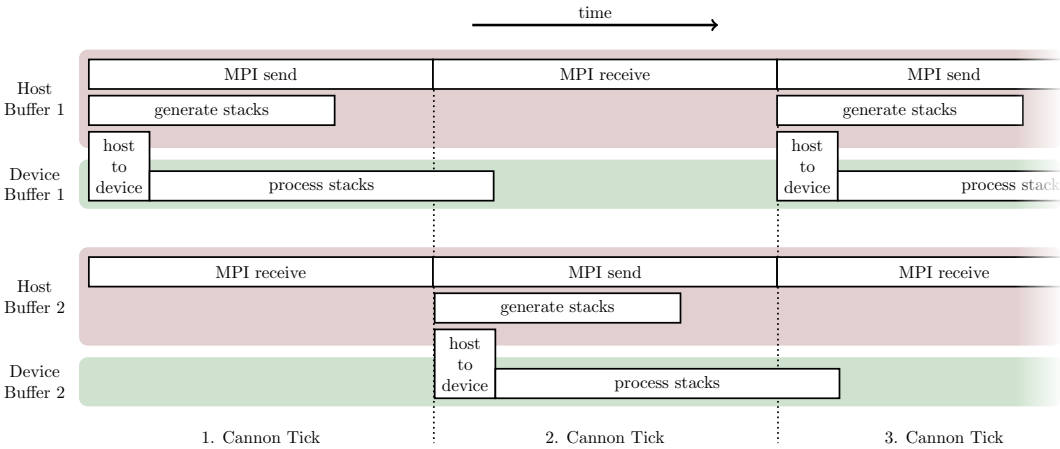
- DBCSR = Distributed Block Compressed Sparse Row
- Working horse of CP2K's linear scaling DFT code
- Non-zero elements are small dense blocks e.g. 13×13
- Each block corresponds to interaction between two atoms
- Additions are *local* operations
- Multiplications are more elaborate...



Architecture of DBCSR's Multiplication Code

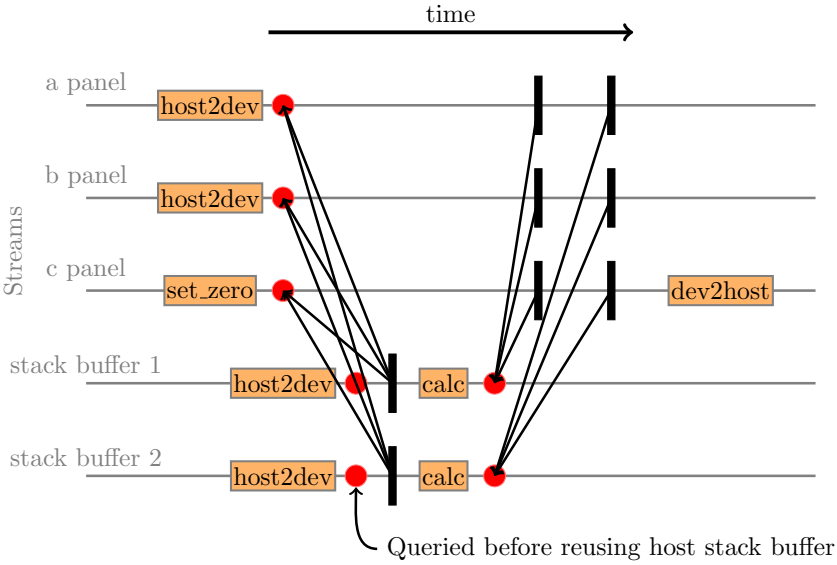


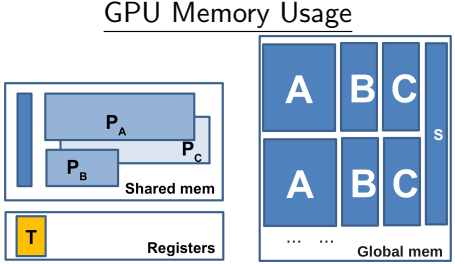
Hiding Communication with Double Buffering



Ideally: Network and GPU always busy

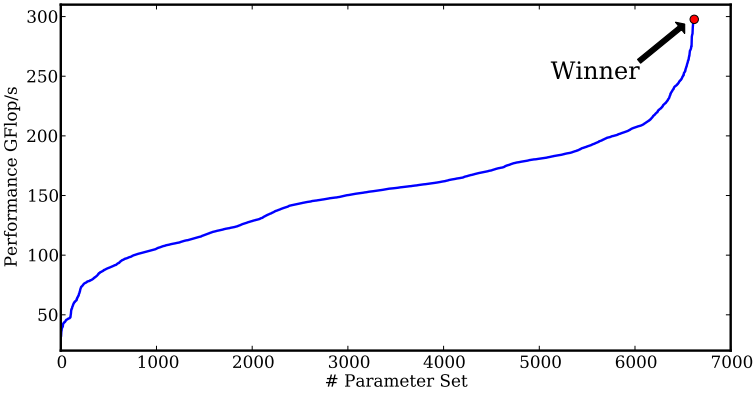
Managing Dependencies with Cuda Events and Streams

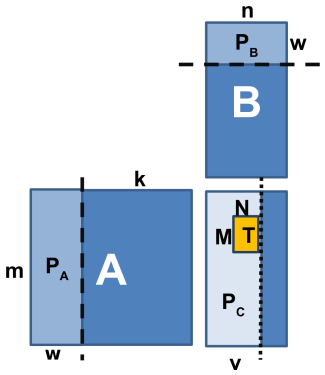




- Larger matrices are processed in slabs P_A, P_B, P_C
- Each thread computes a tile T of the result slab P_C
- Results T are keept in thread's registers
- Outer-product style multiplication reduces access to P_A and P_B
- P_B is stored transposed to coalesced memory access
- Write back to global memory uses *Compare-and-Swap*

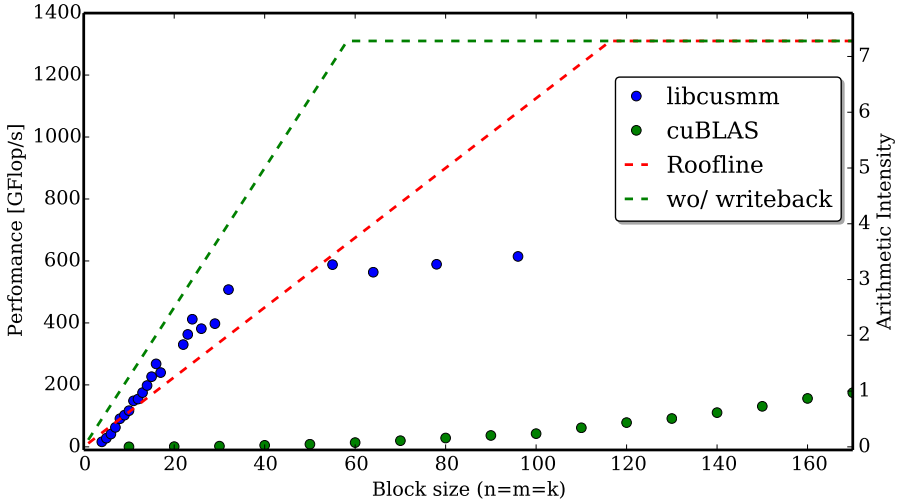
Cuda Kernel Auto-Tuning





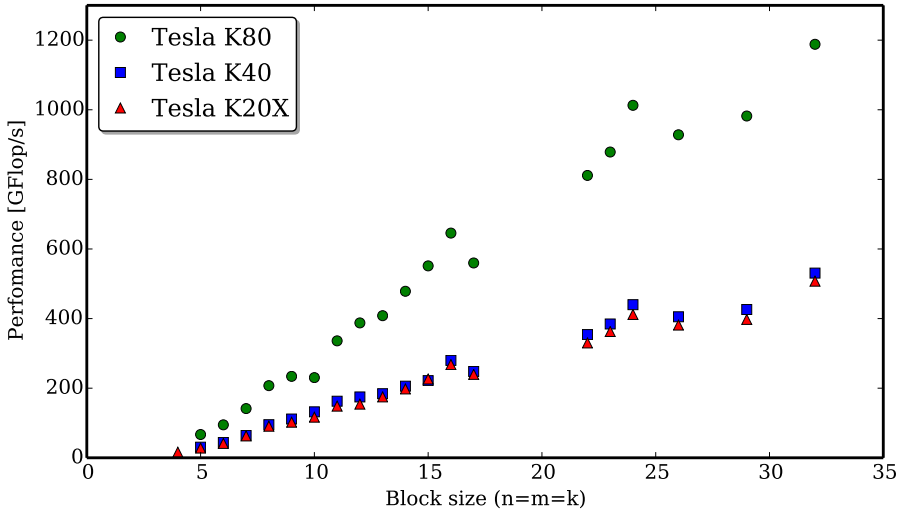
- Six parameters to optimize: v, w, N, M, #threads, #minBlocksPerSM
- On average $>$ 8500 parameters-sets per kernel (heuristically pruned)
- Number of kernels optimized so far: 2349

Cuda Kernel Performance

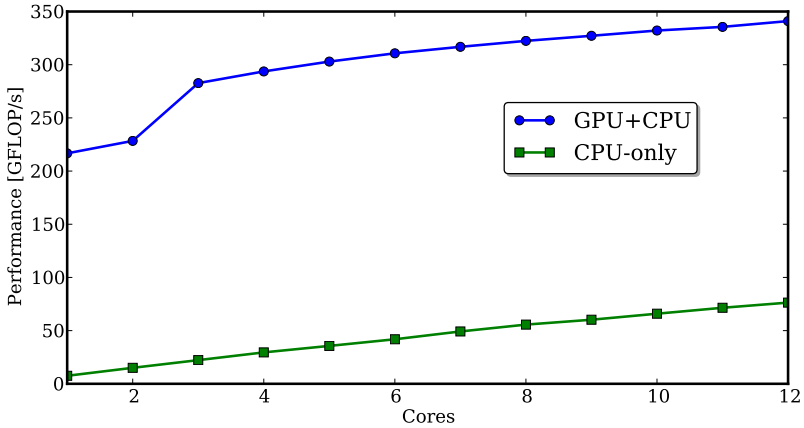


K20X GPU has 1.3TFlop/s and 180GB/s memory bandwidth with ECC

GPU Model Comparison



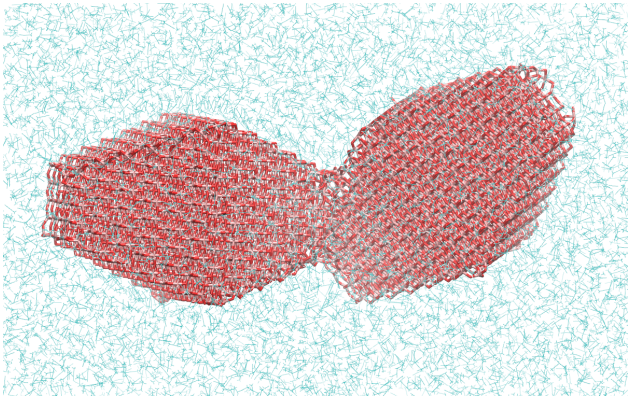
Single Node Performance



4.5x Speedup GPU+CPU vs CPU-only

Artifical benchmark with favorable 23x23 block-size; Dual Sandy Bridge (E5-2620, 2.0GHz, 6 cores); Nvidia K20 GPU.

Full Daint System Science Case

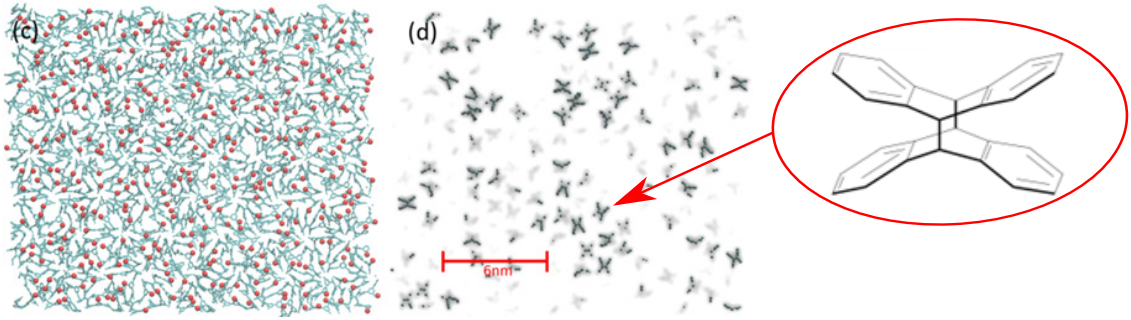


80'000 atoms DFT, high accuracy settings
Aggregated nanoparticles in explicit solution
Relevant for 3rd generation solar cells

- Matrix dims: 772868×772868
- Filter threshold: 10^{-6}
- Matrix occupation $\approx 4\%$
- SCF steps ≈ 50
- # multiplies needed ≈ 2000
- Dense flops needed: 1846613343679824128000
- Actual flops needed: 849928403736295802
- Sparsity boost: $2172\times$
- GPU flop share: 99.4%
- Walltime on 5184 nodes: 6264s

Bridging from Linear Scaling SCF to Materials Properties

2D polymers: synthetically tailored 2D materials beyond graphene

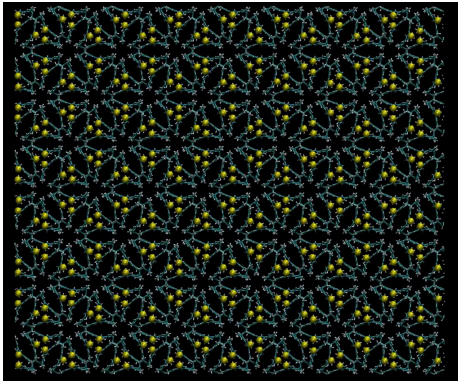


Based on linear scaling MD simulations for 10'000s of atoms,
the morphology and properties of the proposed 2D polymer sheets has been investigated

Payamyar et al., (2013) ADVANCED MATERIALS, DOI: 10.1002/adma.201304705

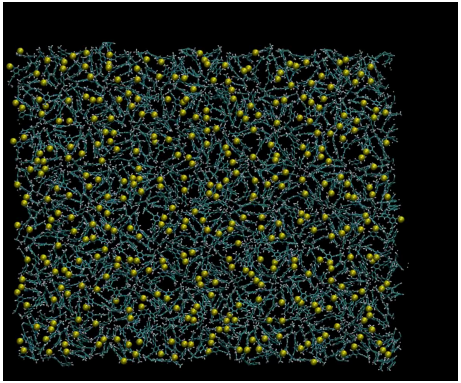
Bridging from Linear Scaling SCF to Materials Properties

Area: 223Å^2



2ps of MD

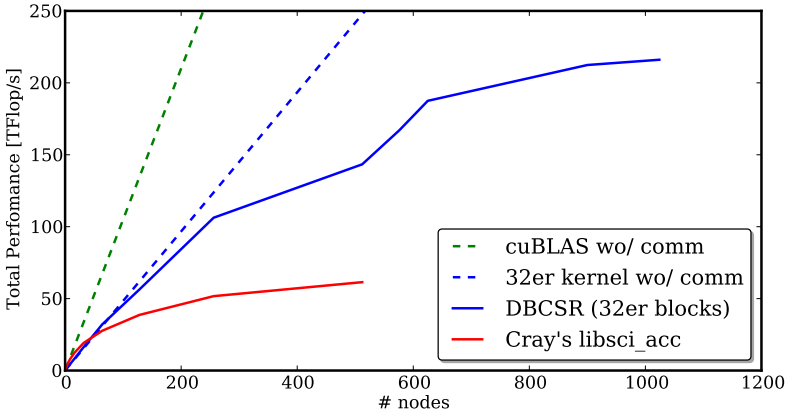
Area: 168Å^2



Payamyar et al., (2013) ADVANCED MATERIALS, DOI: 10.1002/adma.201304705

Outlook: Strong Scaling of Dense Matrix Multiplications

- Matrix Functions: Diagonalization → Taylor series
- Matrix Inverse: Cholesky → Hotteling



Benchmark of pdgemm, 32kx32k double precision matrix

Conclusion

Our DBCSR library enables $\mathcal{O}(N)$ quantum chemistry methods, which allow for novel science.

Lessons learned

- Overlapping communication with computation is key
- Auto-tuning is the way to go
- Avoid manual scheduling, use Cuda events

Acknowledgements

- Joost VandeVondele
- Florian Schiffmann
- Urban Borstnik
- Peter Messmer

Contacts

- mailto:ole.schuett@mat.ethz.ch
- http://nanosim.ethz.ch
- http://dbcsr.cp2k.org
- http://cp2k.org

Please complete the Presenter Evaluation sent to you by email or through the GTC Mobile App.
Your feedback is important!