Multi-Scale Light-Matter Dynamics in Quantum Materials on Aurora PVC

Nariman Piroozan (Intel Corporation)

Taufeq Mohammed Razakh (University of Southern California), Thomas Linker (Stanford University), Ye Luo (Argonne National Laboratory), Ken-ichi Nomura and Aiichiro Nakano (University of Southern California)

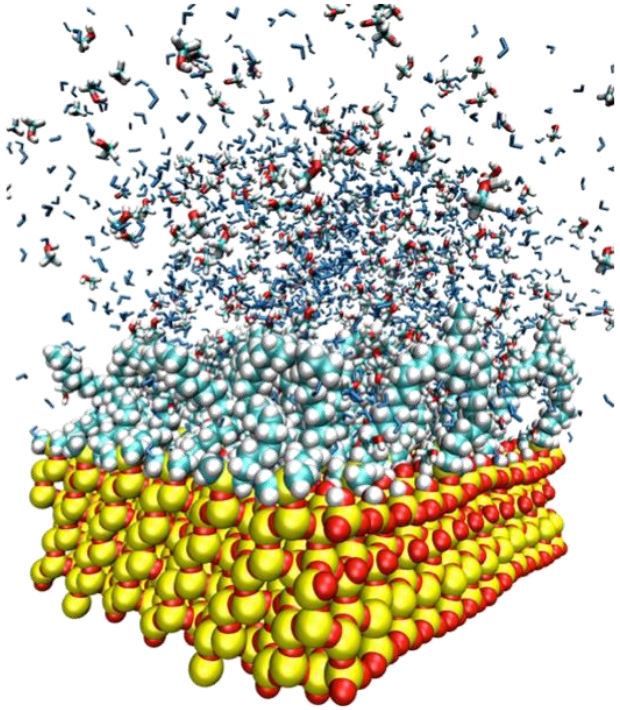
IXPUG Annual Conference 2025

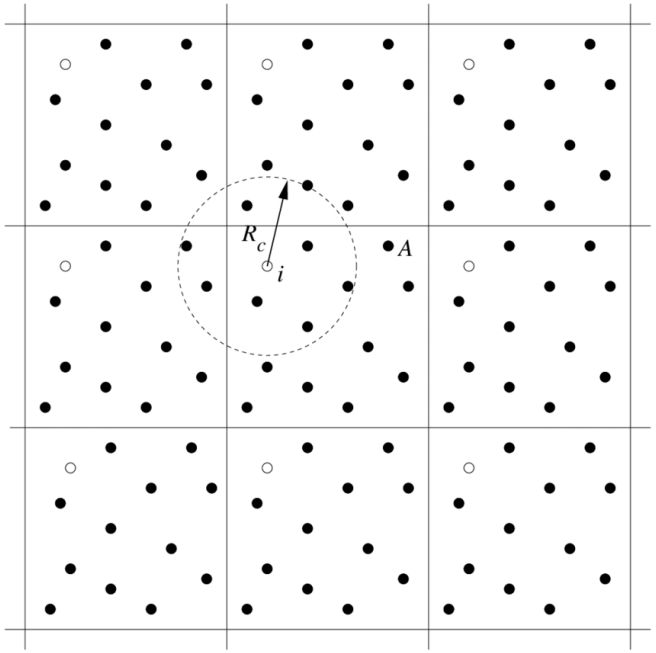
April 15, 2025

intel

Outline

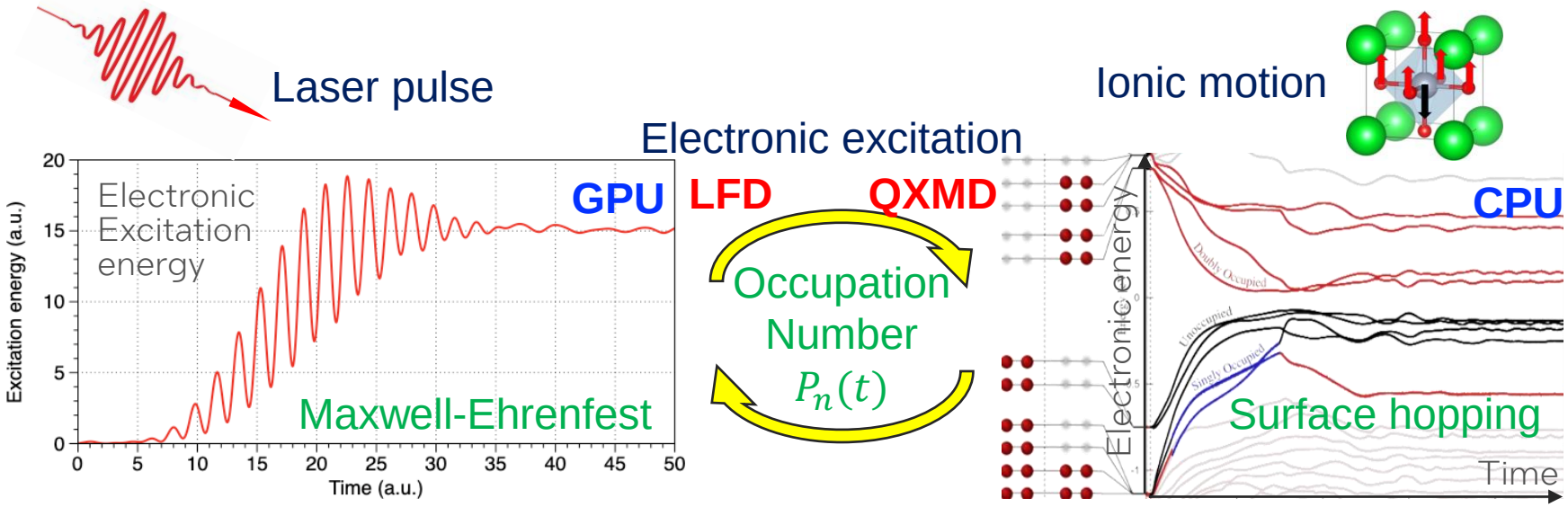
- Application: Introduction to Local Field Dynamics (LFD)
- Transformations to Time Propagation Operator
- Transformations to Electronic Current Computation
- Utilization of Mixed Precision
  - FP32/BF16
- Performance Results
  - Aurora





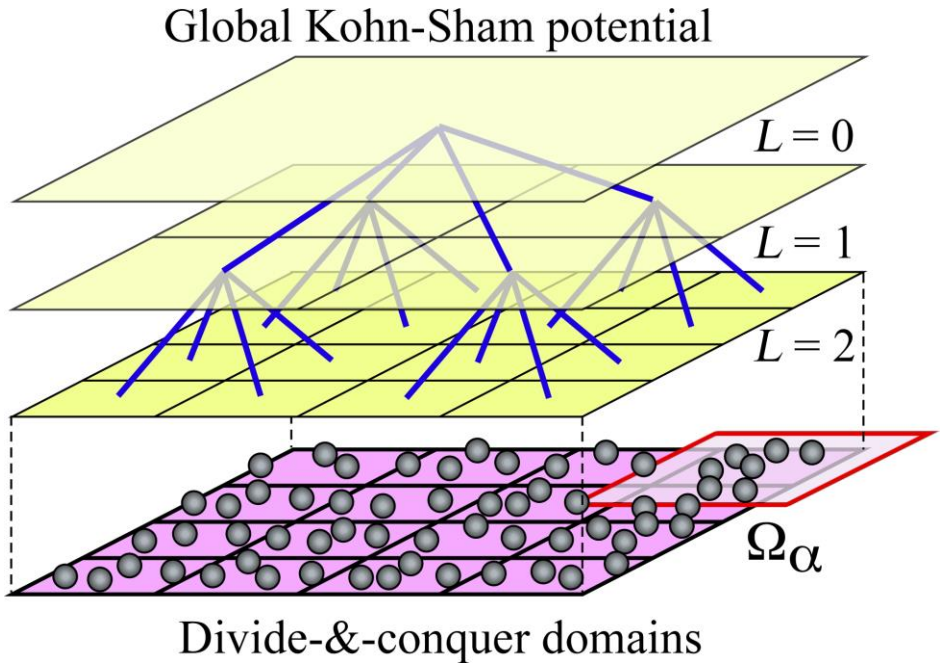
Light-Matter Interaction: DC-MESH

- **DC-MESH (divide-&-conquer Maxwell + Ehrenfest + surface-hopping): O(N) algorithm to simulate photo-induced quantum materials dynamics**
- **LFD (local field dynamics): Maxwell equations for light & real-time time-dependent density functional theory equations for electrons to describe light-matter interaction**
- **QXMD (quantum molecular dynamics with excitation): Nonadiabatic coupling of excited electrons & ionic motions based on surface-hopping approach**
- **"Shadow" LFD (GPU)-QXMD (CPU) handshaking† via electronic occupation numbers with minimal CPU-GPU data transfer**
- **GSLD: Globally sparse (interdomain Hartree coupling via multigrid) & locally dense (intradomain nonlocal exchange-correlation computation via BLAS) solver**



† Niklasson, A.M.N.: 'Extended Lagrangian Born–Oppenheimer molecular dynamics: from density functional theory to charge relaxation models', Euro Phys J B, 2021, 94, (8), pp. 164

Divide-&-Conquer Density Functional Theory



- **Inter Domain - MPI Programming Model - CPU**
  - **Scalability**
- **Intra Domain - OpenMP Programming Model – GPU**
  - **portability**

Time Propagation of Electronic Wave Function is Well-Suited for GPUs

1. Given a wave function ψ for mesh S, quantum time step Δt_{QD}, compute diagonal α, and lower-upper diagonal coefficients β_l, β_u
2. Pick a point $i \in I$ from the $S = \{I, J, K\}$
3. Apply stencil to the grid points $(j, k) \in (J, K)$ to evolve ψ_i over Δt_{QD}
4. Update new wave function at i

$$\psi_i \leftarrow \beta_l \psi_{i-1} + \alpha \psi_i + \beta_u \psi_{i+1}$$

5. Repeat 2 $\forall \{I\}$ with next grid point i_{new}

Re-ordering, Blocking, Hierarchical Parallelism

```
1:  void kin_prop (psi, al, bl, bu, p, d, Norb, Nr, Nx, Ny, Nz) {
2:    complex<float> w;
3:    #pragma omp target teams distribute collapse(3)
4:    for (int j=1; j <= Nr[1]; j++)
5:      for (int k=1; k <= Nr[2]; k++) {
6:        for (int ib=0; ib < (Norb+1)/block_size; ib++) {
7:          complex<float> psi_old[block_size];
8:          int begin = ib*block_size;
            int end = min((ib+1)*block_size, norb);
11:         #pragma omp parallel for simd nowait
9:          for (int n=begin; n < end; n++)
10:           psi_old[n-begin] = psi[0][j][k][n];
8:          for (int i=1; i <= Nr[0]; i++)
9:            #pragma omp parallel for simd nowait
10:           for (int n=begin; n < end; n++) {
11:             w = al*psi[i][j][k][n]
                w += bl[i]*psi_old[n-begin]
12:             …
13:             # update psi_old ← psi [i][j][k][n]
14:             # update psi [i][j][k][n] ← w
15:           }
16:       }
        }
17:   }
```

coarse grain parallelism

override synchronization

fine grain parallelism

Non-local Potential Propagation

The compute-intensive operation arises from solving the following equations†:

- For $Norb$ orbitals solve:

$$0 < \mathrm{n}homo < \ n\mathrm{lumo} \leq Norb$$

$$|\psi_n\rangle -= \frac{i\Delta_{sci}\Delta_{QD}}{2} \underbrace{\sum_{nlumo}^{Norb} |m\rangle\langle m|\psi_{n,}\rangle}_{\text{GEMM operation}} \quad n \in [0, nhomo]$$

- Where:

$$\langle m|\psi_n\rangle = \Delta_x\Delta_y\Delta_z \sum_{ijk} \psi^T_{[0:nlumo],t=0}\psi_{[nlumo:]}$$

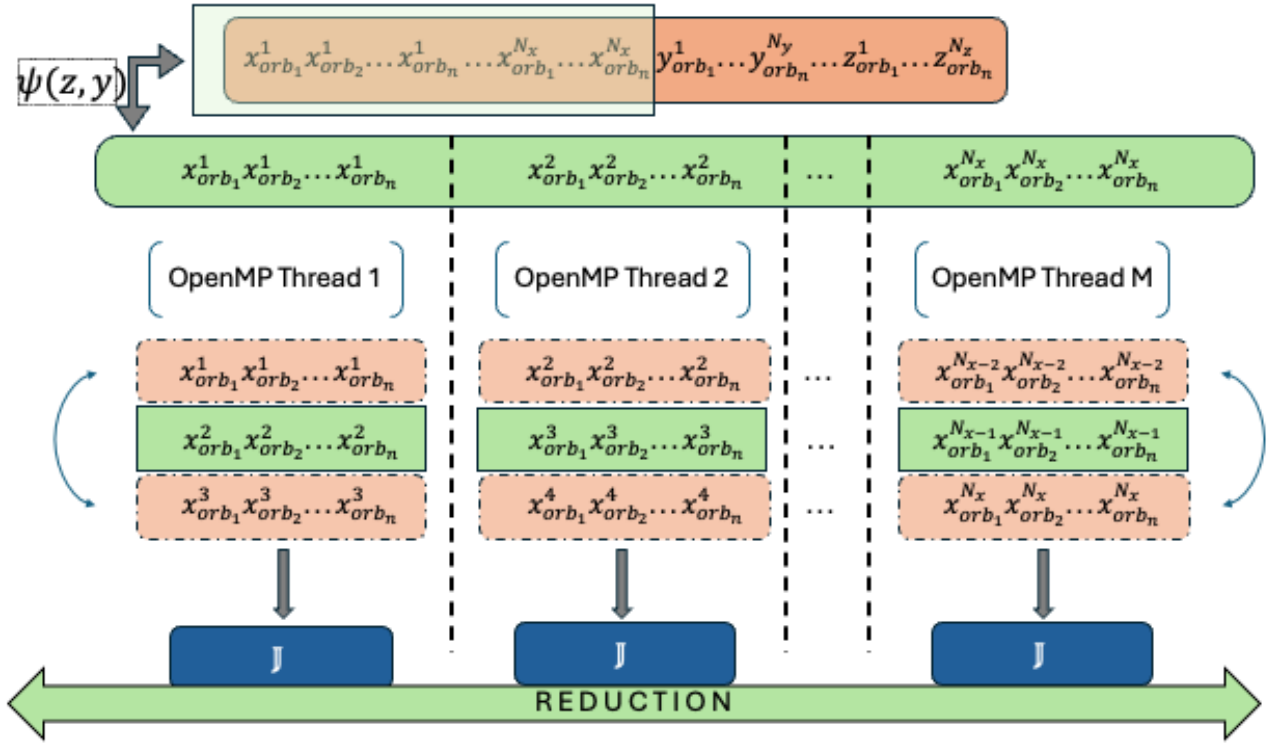
Re-write as two BLAS Level 3 calls (Low Level Matrix-Matrix Op.)

† Wang, C.Y., Elliott, P., Sharma, S., and Dewhurst, J.K.: 'Real time scissor correction in TD-DFT', J Phys-Condens Mat, 2019, 31, (21), pp. 214002

Optimizing Data Access for Electronic Current Density

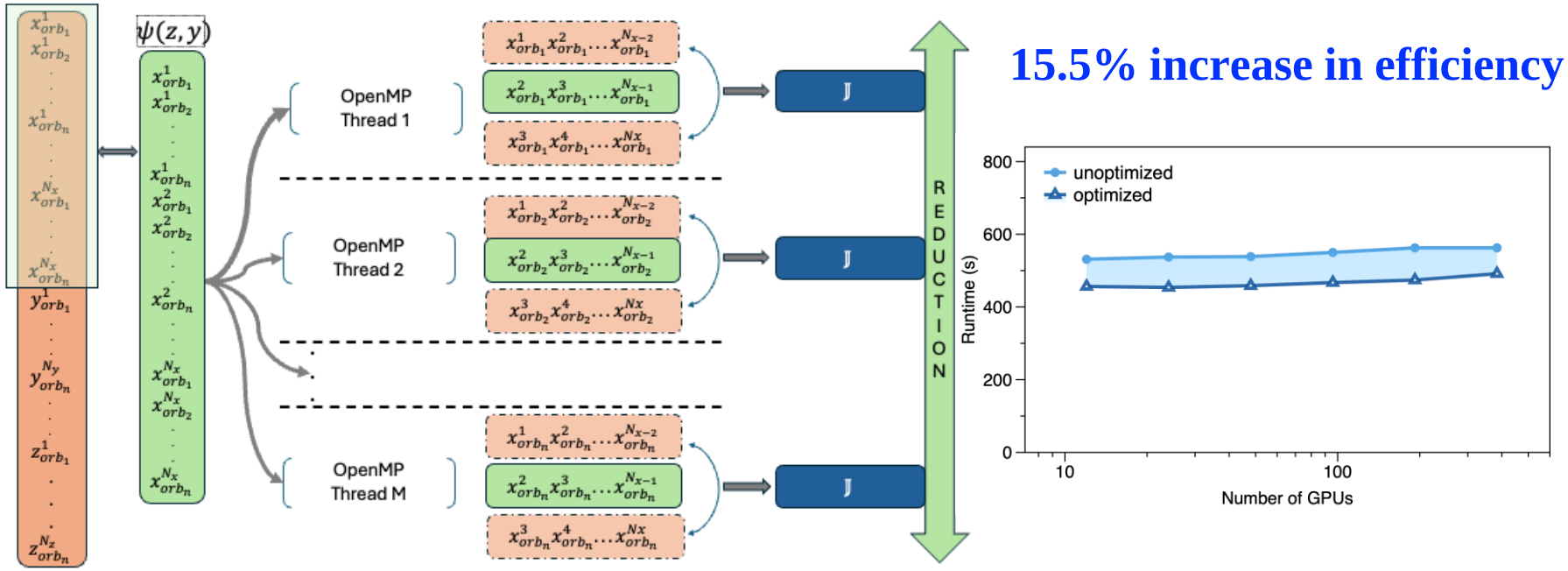
$$\mathbf{j}_{\mathbf{R}(\alpha)}(t) = -\frac{e}{m}\int_{\Omega_\alpha} d\mathbf{r}\left\{\sum_{s\sigma} f_{s\sigma}^{(\alpha)}\,\mathrm{Re}\left[\psi_{s\sigma}^{(\alpha)*}(\mathbf{r},t)\,\frac{\hbar}{i}\nabla\psi_{s\sigma}^{(\alpha)}(\mathbf{r},t)\right] + \frac{e}{c}\mathbf{A}_{\mathbf{R}(\alpha)}(t)\rho(\mathbf{r},t)\right\}$$

Bottleneck: Spatial Derivative of the wave functions during electronic current density



Optimizing Data Access for Electronic Current Density

$$\mathbf{j}_{\mathbf{R}(\alpha)}(t) = -\frac{e}{m}\int_{\Omega_\alpha} d\mathbf{r}\left\{\sum_{s\sigma} f_{s\sigma}^{(\alpha)}\mathrm{Re}\left[\psi_{s\sigma}^{(\alpha)*}(\mathbf{r},t)\frac{\hbar}{i}\nabla\psi_{s\sigma}^{(\alpha)}(\mathbf{r},t)\right] + \frac{e}{c}\mathbf{A}_{\mathbf{R}(\alpha)}(t)\rho(\mathbf{r},t)\right\}$$

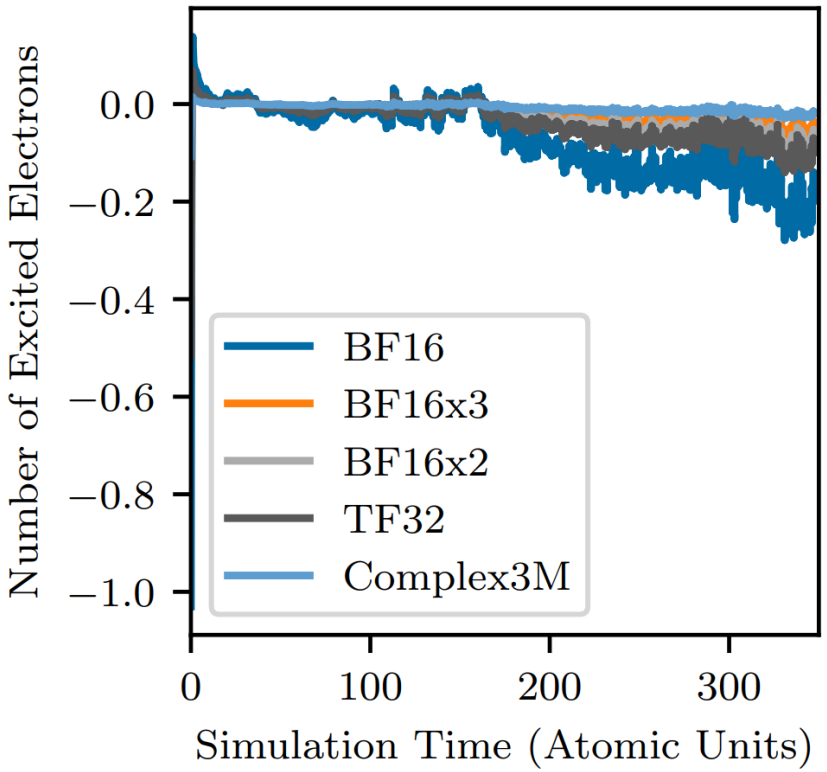


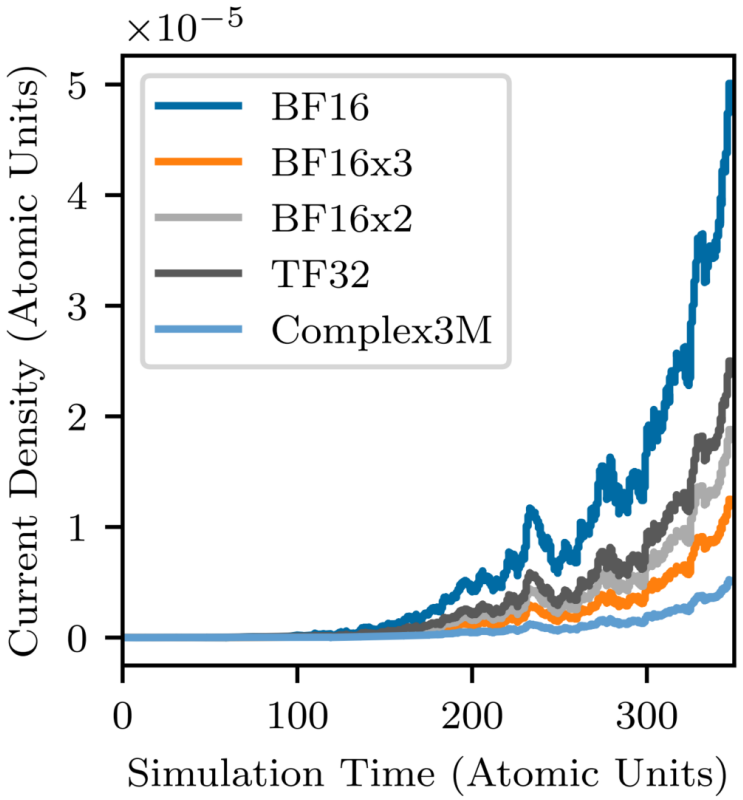
Alternative BLAS Precision Modes

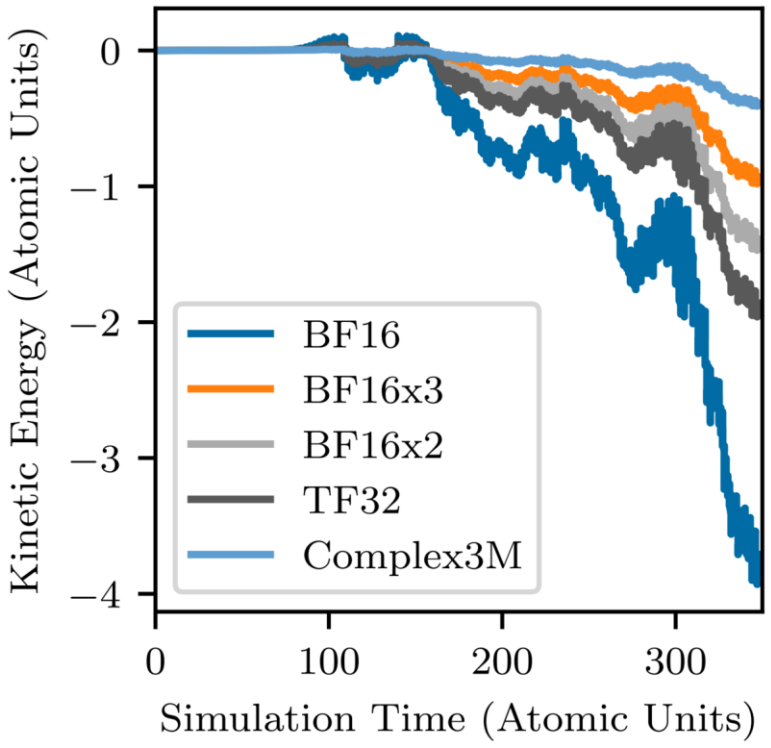
Environmental Variable: MKL_COMPUTE_MODE

Compute Mode	Environmental Variable	Peak Theoretical
BF16	FLOAT_TO_BF16	16x
BF16X2	FLOAT_TO_BF16X2	(16/3)x
BF16X3	FLOAT_TO_BF16X3	(8/3)x
TF32	FLOAT_TO_TF32	8x
Complex_3M	COMPLEX_3M	(4/3)x

Accuracy Using Alternative Precisions

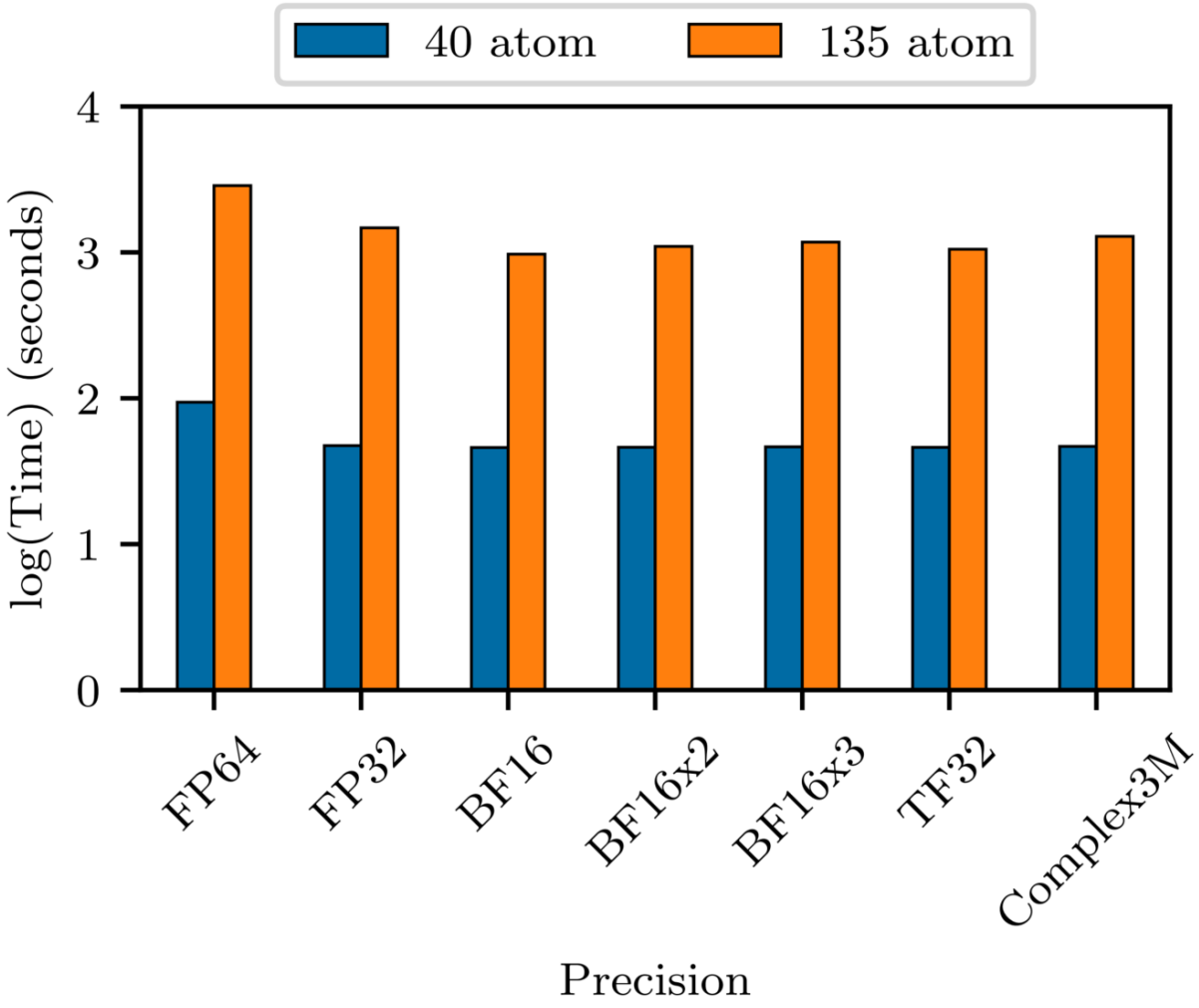


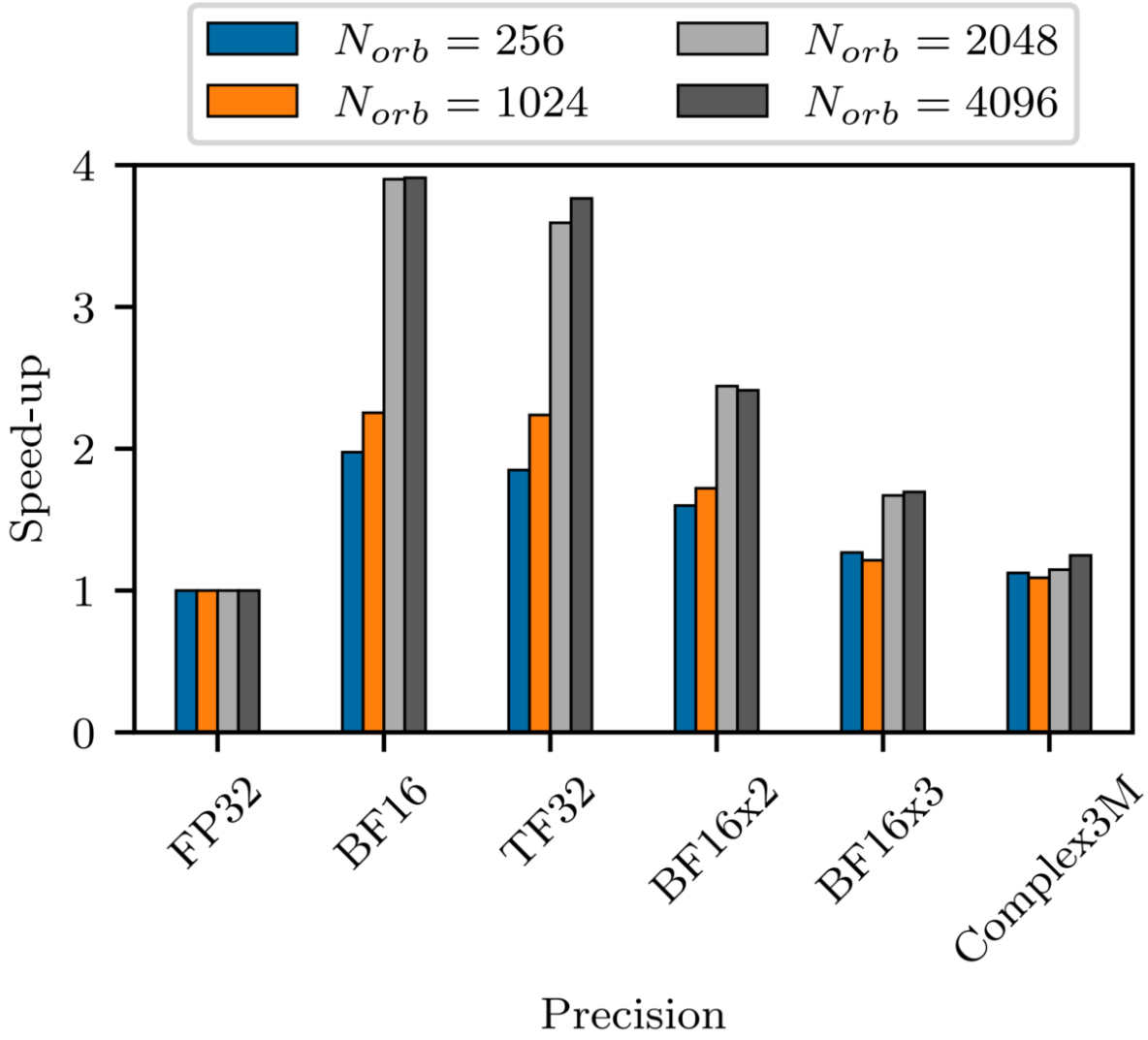




See slide 17 for workloads and configurations. Results may vary.

Performance Using Alternative Precisions

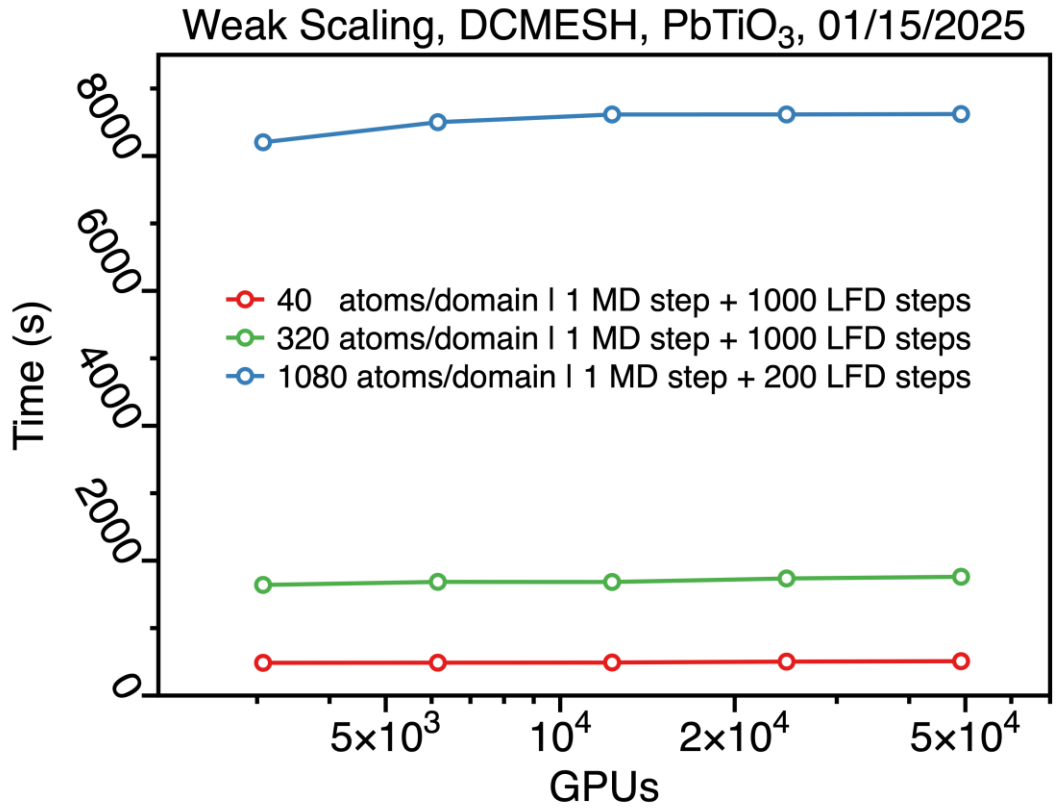


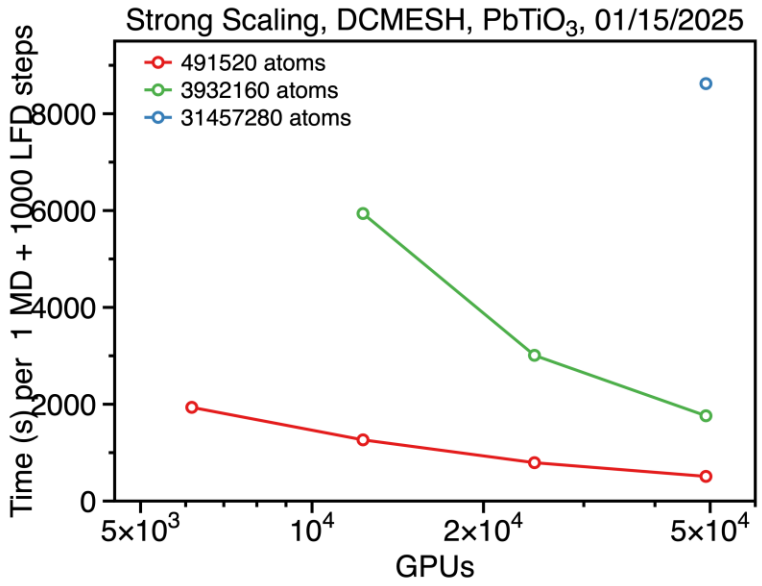


See slide 17 for workloads and configurations. Results may vary.

Performance on Aurora

Achieved 35% of peak during early access with FP 32





Number of atoms	Number of KS orbitals	TFlop/s	% of peak
40	256	3.76 (FP32)	16.34
135	864	7.56 (FP32)	32.86
135	1024	8.10 (FP32)	35.22
135	1024	12.71 (FP32/BLAS – BF16)	55.26

Conclusions

- **We describe how we do light-matter interaction in DC-MESH**
- **Efficient O(N) stencil computation for GPU's when doing electronic time-propagation**
- **Nonlocal Correction as 2 GEMM operations**
- **Streamlined Stencil Kernel computation for current density calculation**
- **More than 50% of peak with Mixed Precision**

Acknowledgements

- Department of Energy (DOE), Office of Science, Basic Energy Sciences, award DE-SC0000267409. K.N
- NSF grant OAC- 2118061
- Aurora ESP program
- DOE Innovative and Novel Computational Impact on Theory and Experiment (INCITE) Program

Questions?

Intel technologies may require enabled hardware, software or service activation.

No product or component can be absolutely secure. Your costs and results may vary.

Performance varies by use, configuration and other factors. Learn more on the Performance Index Site.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel® Data Center GPU Max Series 1550 : 1-node, Total Memory 128 GB, kernel 5.14.21-150500.55.52-default, compiler gcc 7.5.0 20210514, Intel® oneAPI2025.1, DC-MESH, Intel® oneMKL2025.1, Intel® pti-gpu.


intel