GPU Acceleration of the WEST Code for Large-Scale Full-Frequency GW Calculations

Victor Yu and Marco Govoni

Materials Science Division, Argonne National Laboratory

WEST! MICCoM

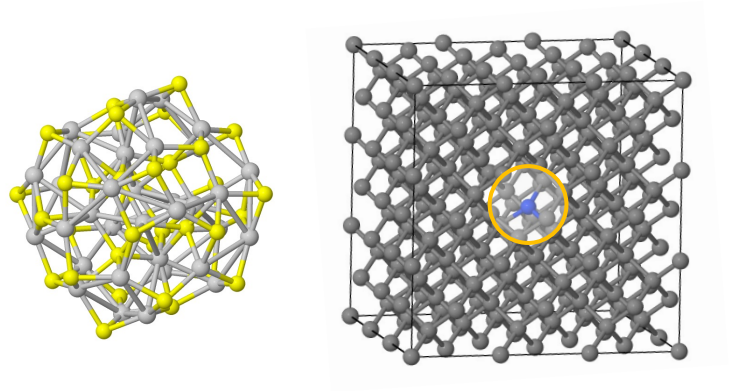
Argonne NATIONAL LABORATORY

Motivation and introduction

http://west-code.org

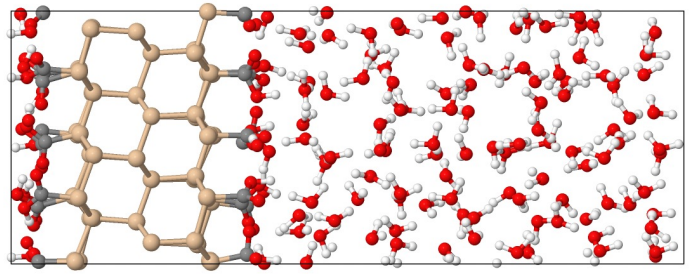
Using first-principles simulations, we study electronic structures, charged and neutral excitations of large heterogeneous systems

- Nanoparticles for energy conversion
- Solid/liquid interfaces for photocatalysis
- Spin defects in semiconductors for quantum information science

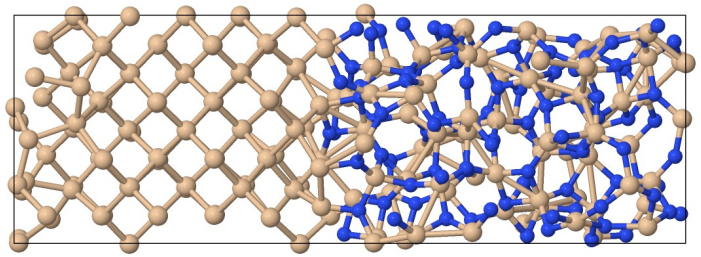


We have developed **WEST**, a massively parallel implementation of many-body perturbation theory (GW/BSE) **without empty states**

- ✓ No explicit summation over virtual orbitals
- ✓ No explicit storage or inversion of large dielectric matrices
- ✓ Full frequency integration with contour deformation

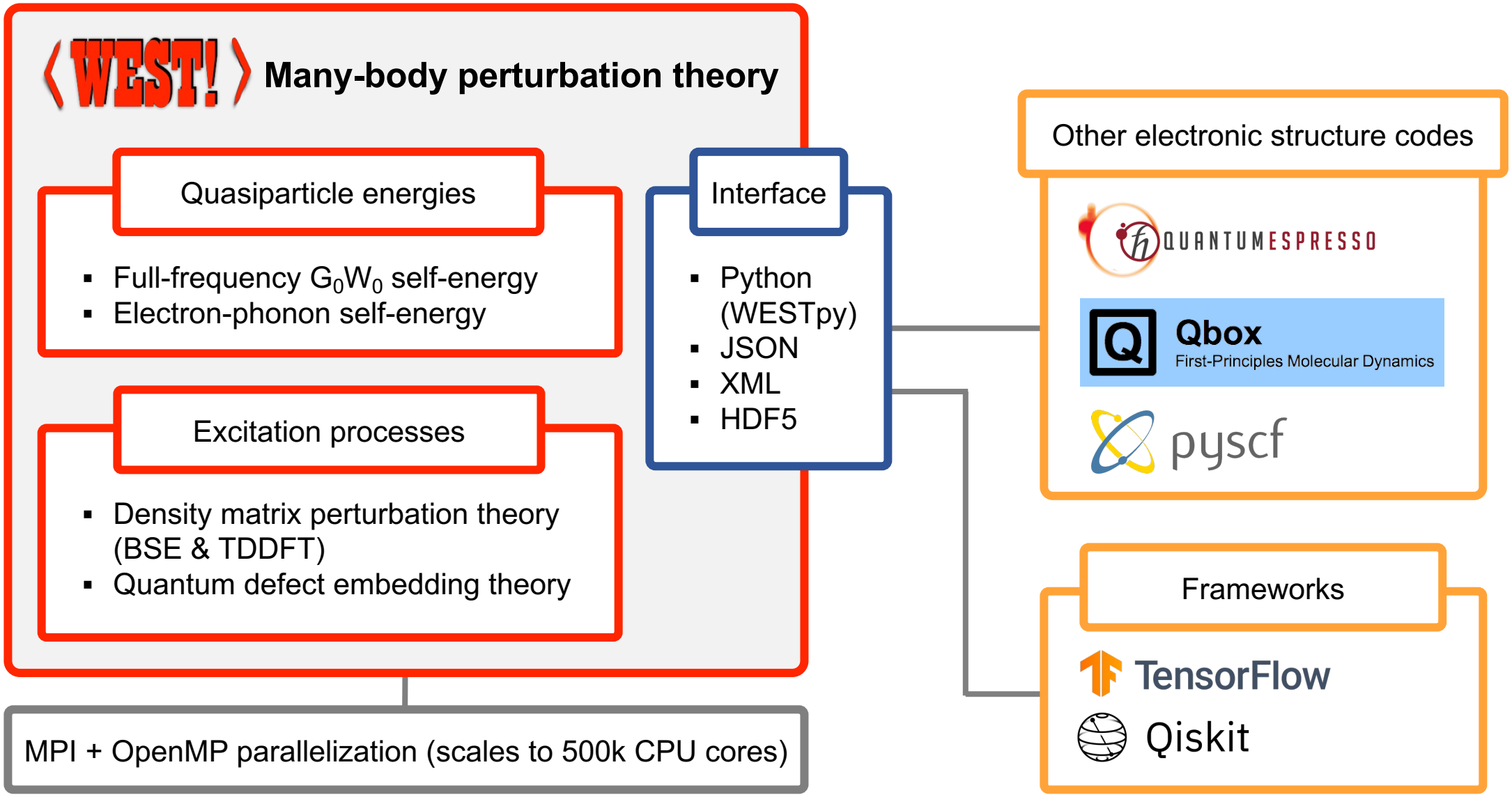


WEST has been successfully applied to large G_0W_0 calculations consisting of ~2,000 electrons on CPUs



Govoni and Galli, J. Chem. Theory Comput.11, 2680-2696 (2015)

The **WEST** code: **W**ithout **E**mpty **ST**ates



Govoni and Galli, J. Chem. Theory Comput.11, 2680-2696 (2015)

GPU acceleration in WEST

WEST has been ported to **NVIDIA GPUs**

Porting strategies

- GPU libraries for performance:
  cuFFT, cuBLAS, cuSOLVER, ...
- Directives for portability:
  CUDA Fortran → OpenACC → OpenMP

Achievements

- Significant speedup over CPU code
- Excellent strong and weak scalability demonstrated on various supercomputers
- Tractable size of full-frequency G_0W_0 pushed to **10k electrons**

✅ ⟨WEST!⟩

OLCF/Summit
200.79 PFLOP/s
NVIDIA V100 GPUs

https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit

✅ ⟨WEST!⟩

NERSC/Perlmutter
93.75 PFLOP/s
NVIDIA A100 GPUs

https://www.nersc.gov/systems/perlmutter

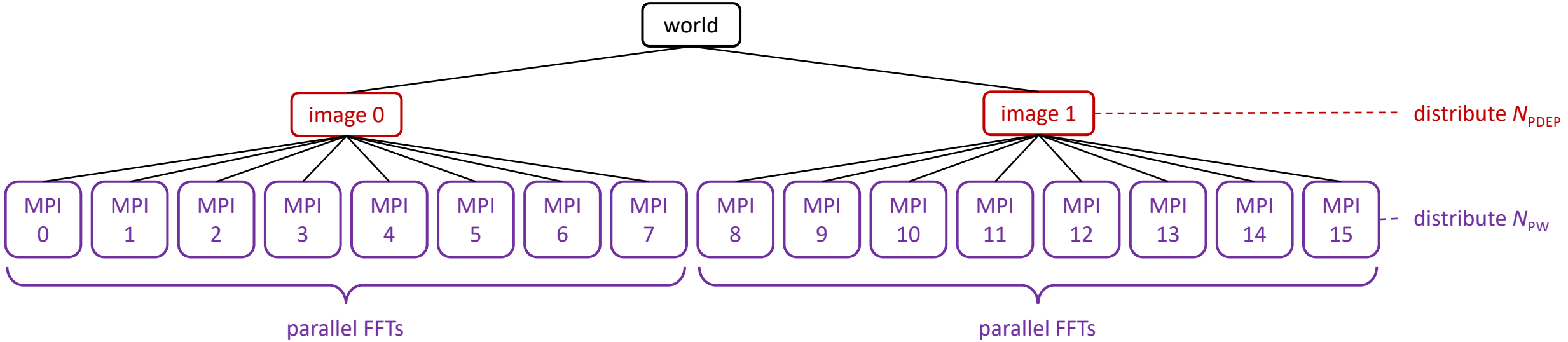
Yu and Govoni, J. Chem. Theory Comput.18, 4690-4707 (2022)

Projective dielectric eigenpotentials (PDEP)

```
davidson loop
  perturbation loop
    spin loop
      band loop
        FFT
        solve linear system
        FFT
    density
  solve eigenproblem
```

N_{PDEP} perturbations distributed across “images”

N_{PW} coefficients distributed across MPI tasks within an image



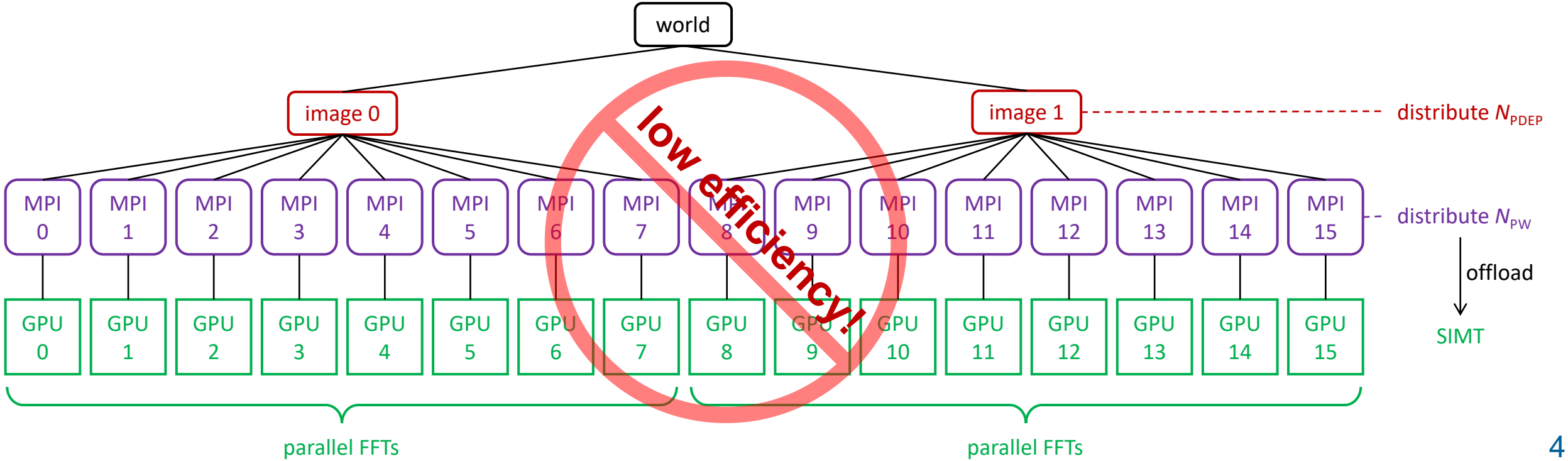
Straightforward GPU offloading?

```
davidson loop
  perturbation loop
    spin loop
      band loop
        FFT
        solve linear system
        FFT
    density
  solve eigenproblem
```

perturbation loop → N_{PDEP} perturbations distributed across "images"

FFT, solve linear system, FFT → N_{PW} coefficients distributed across MPI tasks within an image

Each MPI task offloads to a GPU



Hierarchical parallelization scheme

```
davidson loop
    perturbation loop
        spin loop
            band loop
                FFT
                solve linear system
                FFT
        density
    solve eigenproblem
```

N_{PDEP} perturbations distributed across “images”

N_{PW} coefficients distributed across MPI tasks **within an image**

The calculation of the response can be done independently for each perturbation, band, and spin channel
→ **many layers of parallelism**

```
davidson loop
    copy to GPU
    perturbation loop
        spin loop
            band loop
                GPU FFT
                GPU solve linear system
                GPU FFT
        density
    GPU solve eigenproblem
    copy from GPU
```

N_{PDEP} perturbations distributed across “images”

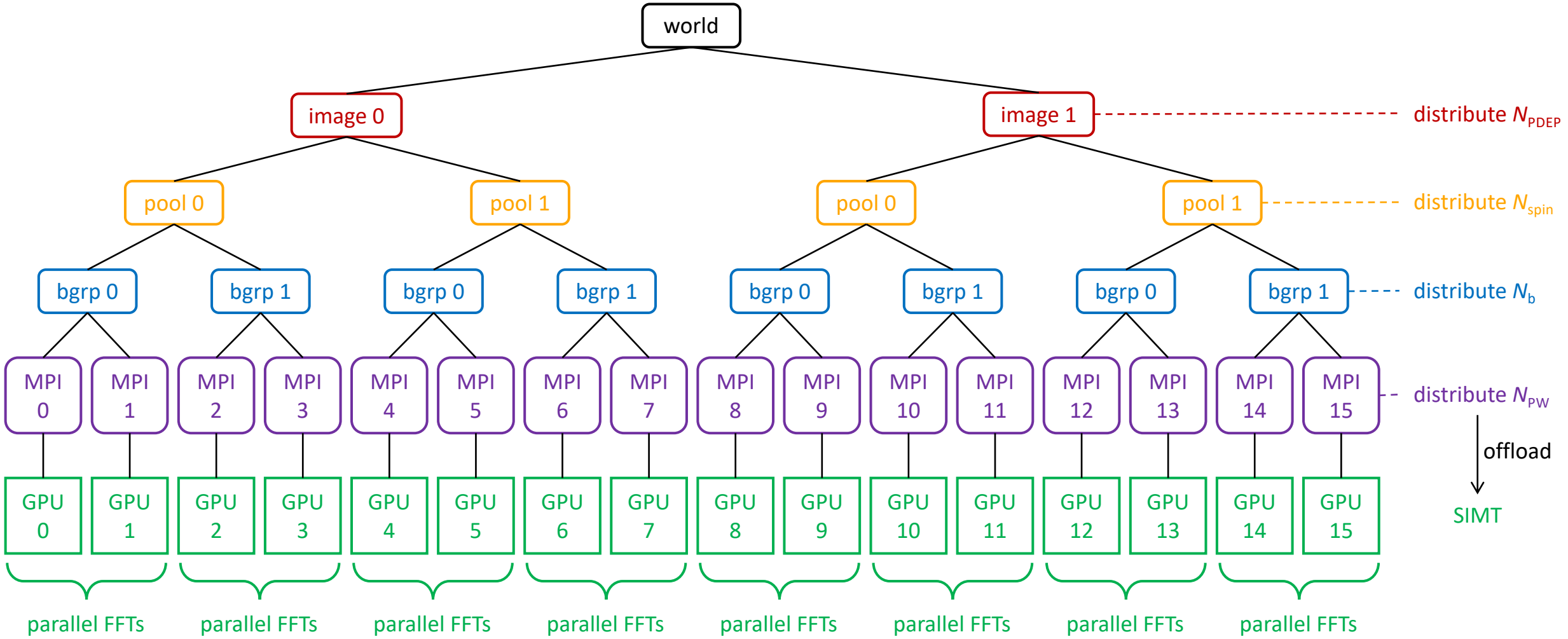
N_{spin} spin channels distributed across “pools” within an image

N_b bands distributed across “band groups” within a pool

N_{PW} coefficients distributed across MPI tasks **within a band group**

Each MPI task offloads to a GPU

Hierarchical parallelization scheme



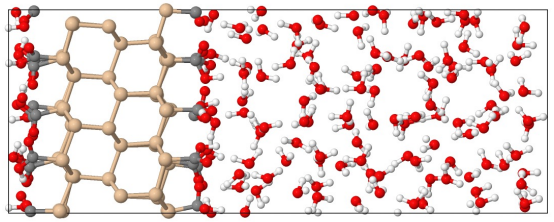
- All levels of parallelism in the PDEP algorithm are fully harnessed
- CPU-GPU and GPU-GPU communication cost is reduced

GPU porting and optimization strategies

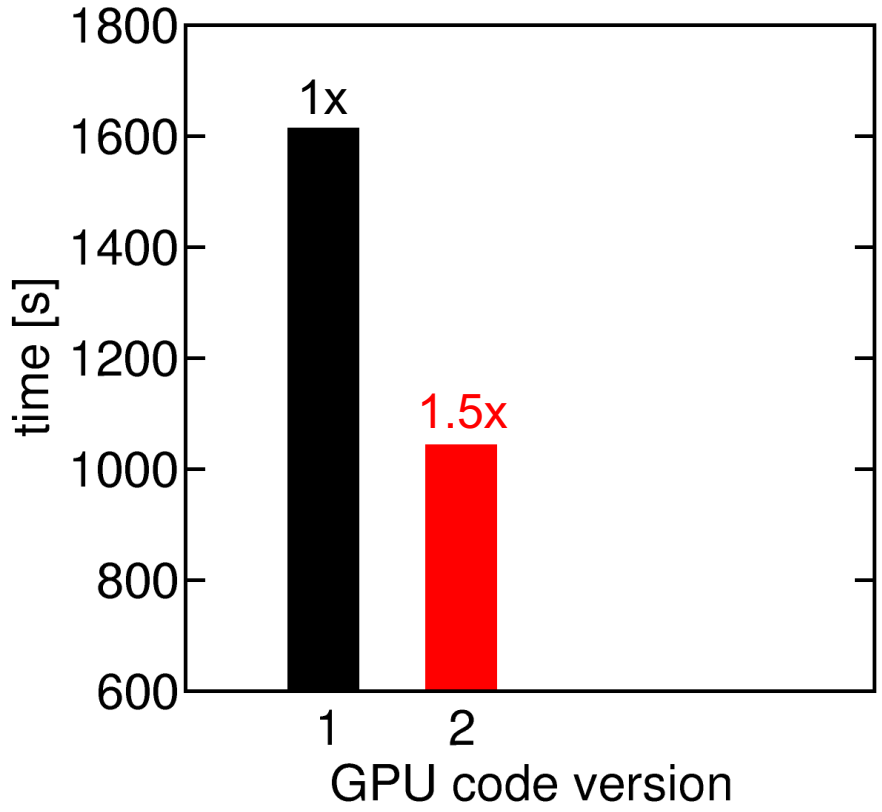
v1: Initial GPU porting (baseline)

v2: Multi-level parallelization and data distribution

- Reduced CPU-GPU and MPI communication
- Improved load balance across GPUs



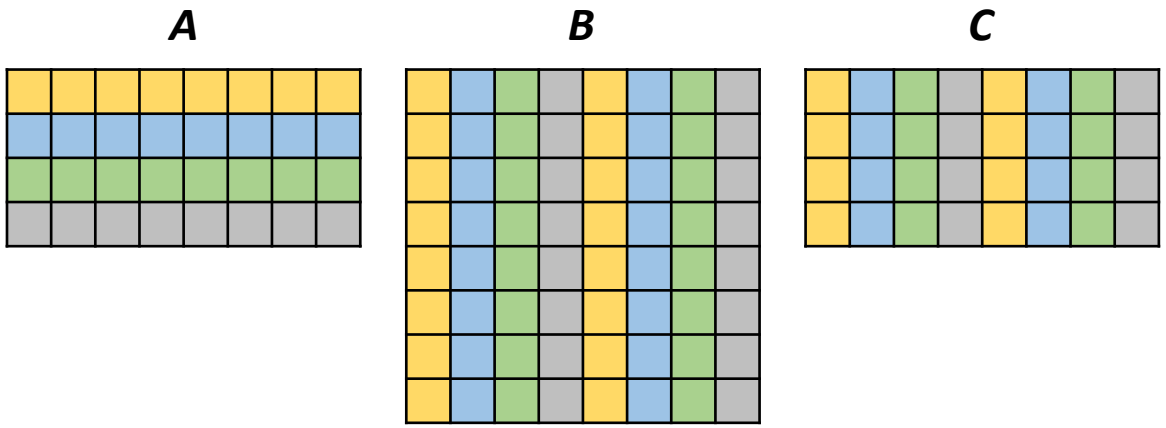
Full-frequency G_0W_0 calculation of COOH-Si/H_2O interface
1560 electrons, cutoff 60 Ry, PBE, ONCV PP
Ground state DFT with Quantum ESPRESSO



Overlapping communication and computation

CPU and GPU operations can be done asynchronously
- MPI communications
- CPU-GPU communications
- GPU computations

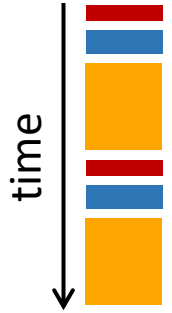
Example:
Distributed matrix multiplication ***A*** x ***B*** = ***C***



(each color represents an MPI process)

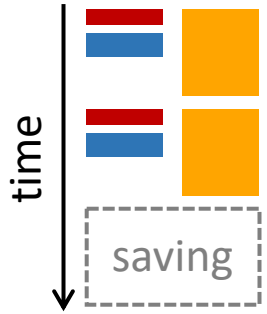
GPU v1

```
1 do i = 1, n_mpi
2   copy a_h to a_d
3   compute c_d(i) = a_d * b_d
4   circular shift a_h
5 end do
```



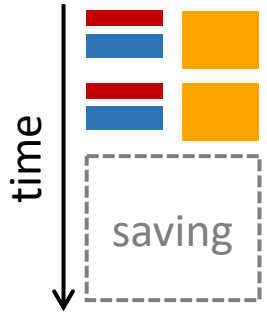
GPU v2 (async)

```
1 do i = 1, n_mpi
2   async circular shift a_h
3   copy a_h to a_d
4   compute c_d(i) = a_d * b_d
5 end do
```



GPU v3 (async + SP)

```
1 convert a_h to a_h_sp
2 do i = 1, n_mpi
3   async circular shift a_h_sp
4   copy a_h_sp to a_d_sp
5   convert a_d_sp to a_d
6   compute c_d(i) = a_d * b_d
7 end do
```



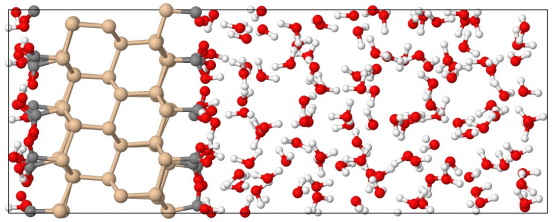
GPU porting and optimization strategies

v1: Initial GPU porting (baseline)

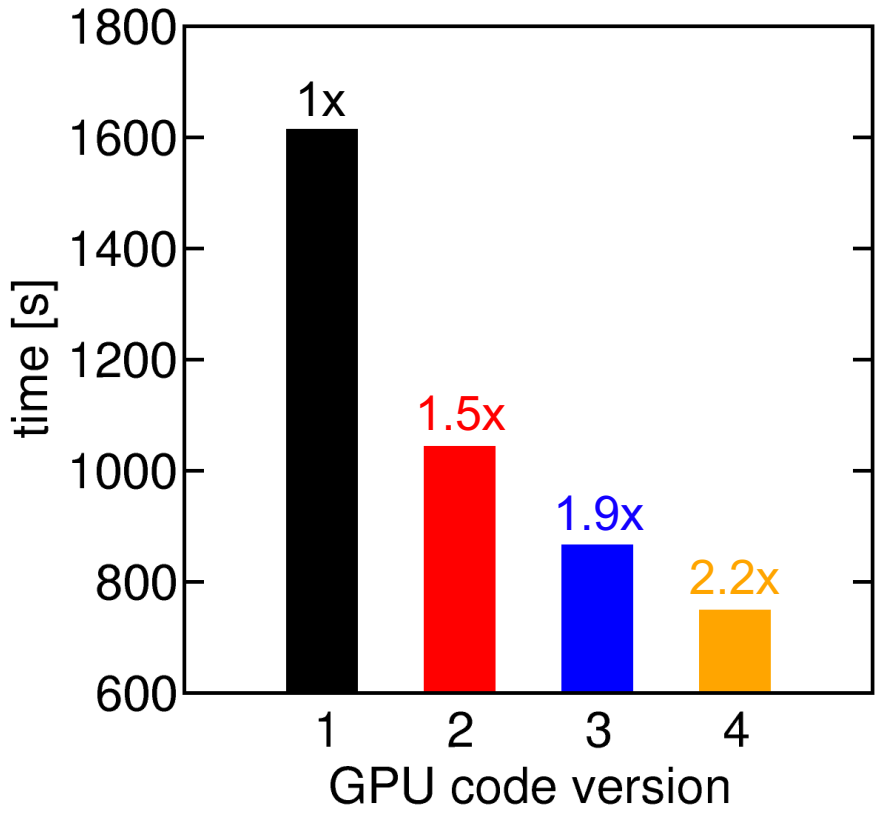
v2: Multi-level parallelization and data distribution
- Reduced CPU-GPU and MPI communication
- Improved load balance across GPUs

v3: Single-precision for selected operations
- Fast Fourier transforms (FFTs)
- Data communication (MPI)
- Quasiparticle energies match double-precision

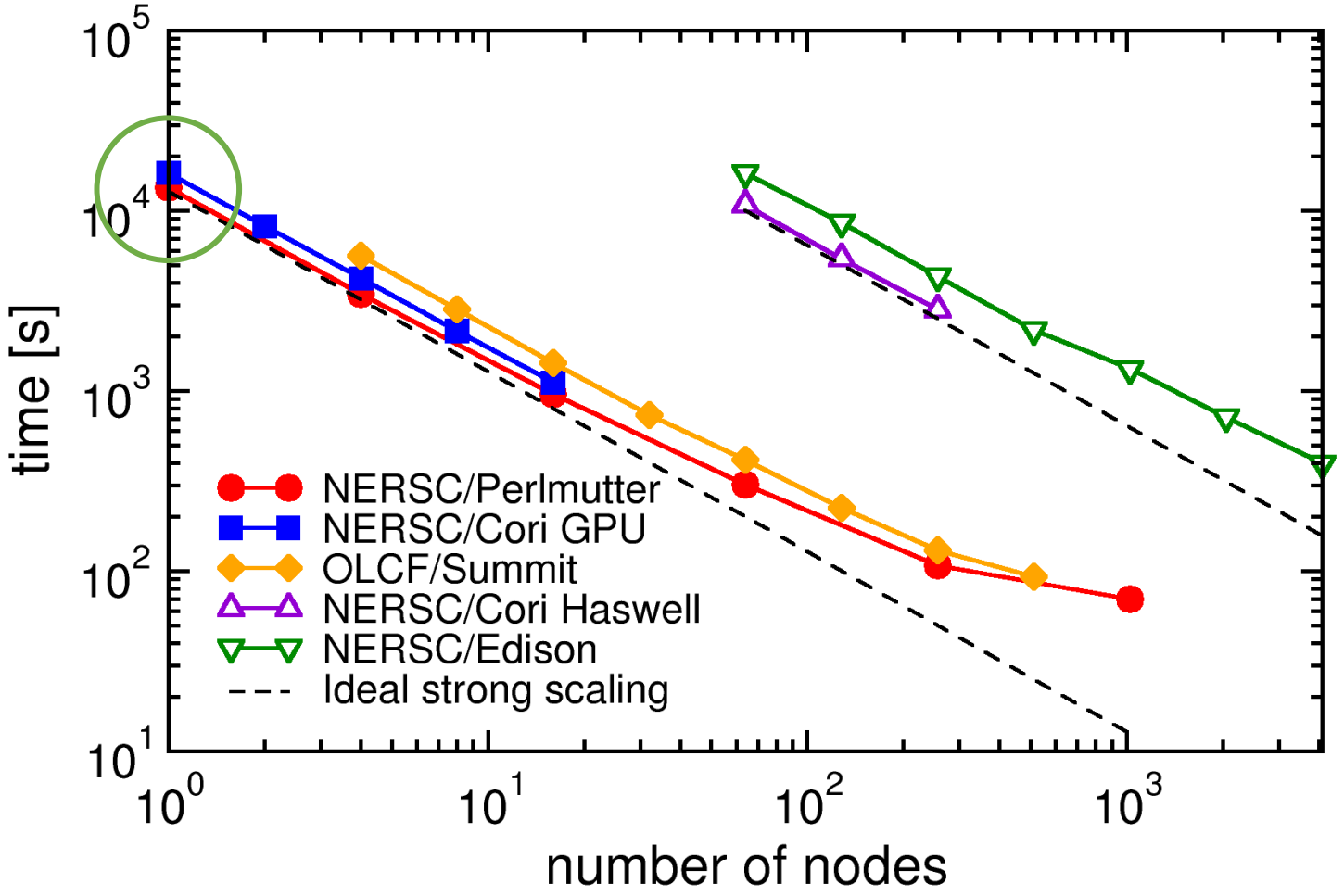
v4: Additional optimizations
- Overlap between communication and computation
- GPU memory access
- MPI I/O



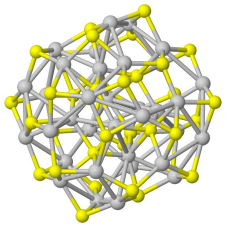
Full-frequency G_0W_0 calculation of COOH-Si/H_2O interface
1560 electrons, cutoff 60 Ry, PBE, ONCV PP
Ground state DFT with Quantum ESPRESSO



Performance benchmark



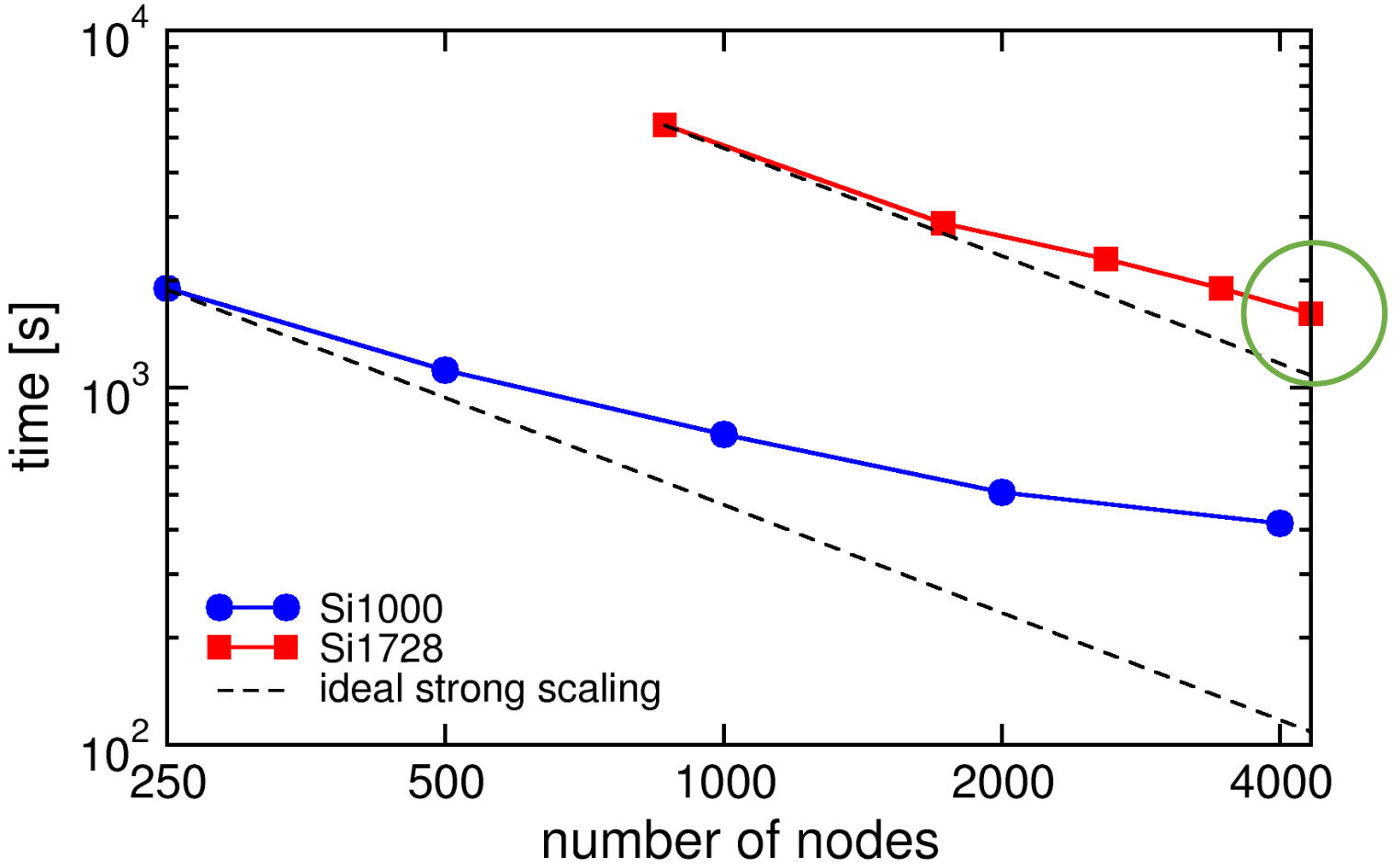
1 NERSC/Perlmutter node = 4 NVIDIA A100 GPUs
1 NERSC/CoriGPU node = 8 NVIDIA V100 GPUs
1 OLCF/Summit node = 6 NVIDIA V100 GPUs
1 NERSC/CoriHaswell node = 32 Intel Haswell CPU cores
1 NERSC/Edison node = 24 Intel Ivy Bridge CPU cores



Full-frequency G_0W_0 calculation of CdSe nanoparticle
884 electrons, cutoff 50Ry, PBE, ONCV PP
Ground state DFT with Quantum ESPRESSO

- Nearly ideal strong scaling demonstrated on various machines
- Time to solution on GPU nodes is less than **1/30** of that on CPU nodes
- **2x faster on A100** than on V100 (more memory, higher memory bandwidth, FP64 tensor cores)

Strong scaling to full Summit

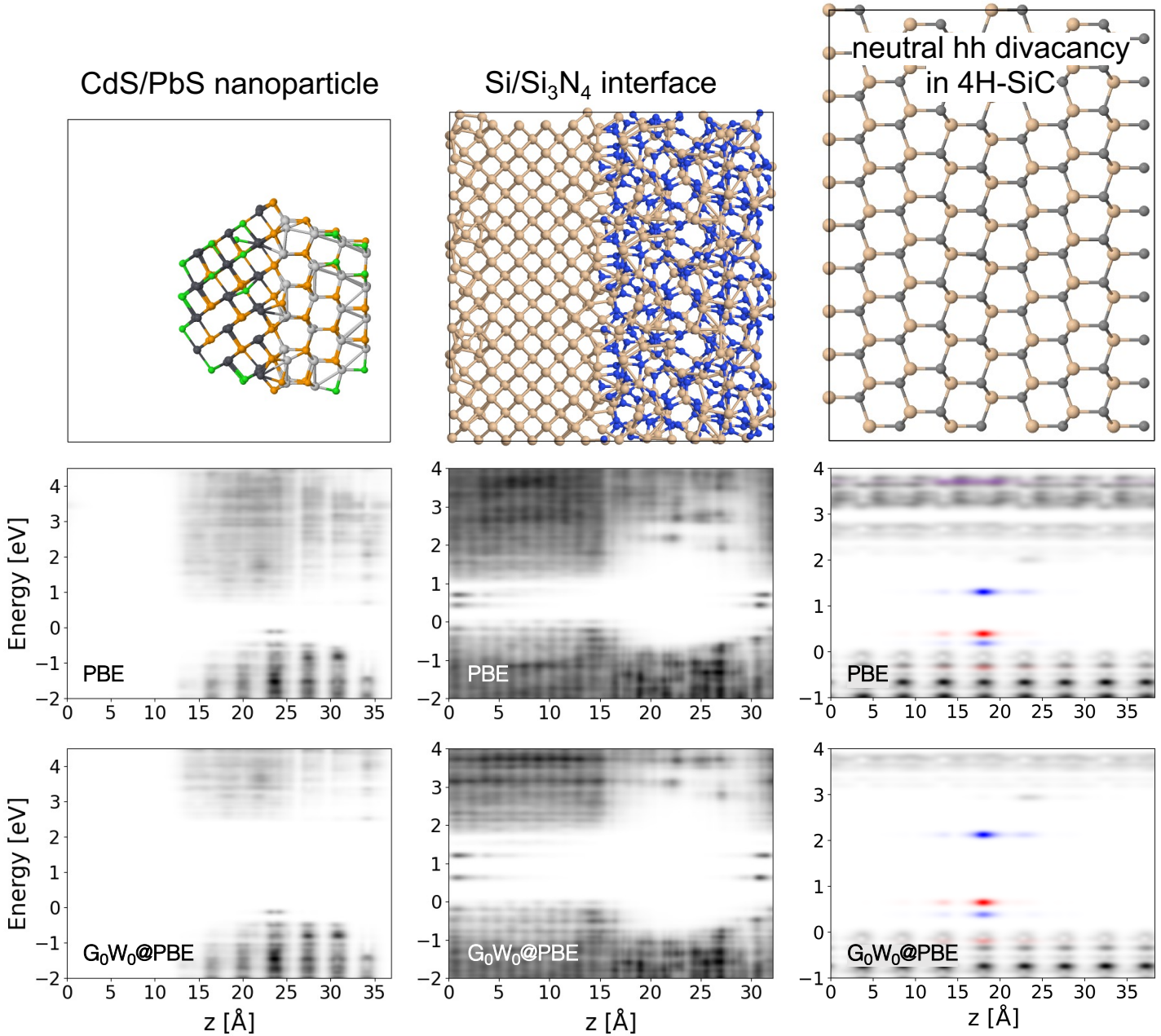


1 Summit node = 2 IBM POWER9 CPUs + 6 NVIDIA V100 GPUs

Full-frequency G_0W_0 calculation of 1000 or 1728 Si atoms
Ground state DFT with Quantum ESPRESSO

- WEST-GPU scales to the entire Summit supercomputer
- Better scalability observed for bigger system (1,728-atom silicon supercell) due to a higher computation-to-communication ratio
- 80 quasiparticle energies of the 1,728-atom silicon supercell solved in ~30 min using **25,920 V100 GPUs** (94% of Summit)

Large-scale full-frequency G_0W_0 calculations



Local density of states computed for prototypical systems representing our target applications:

- Large nanoparticles and interfaces
  → materials for energy conversion
- Defects in semiconductors
  → quantum information science (quantum computation, communication, and sensing)

system	N_{atom}	$N_{electron}$	N_{spin}	N_{PW}
CdS/PbS	301	2,816	1	948,557
Si/Si_3N_4	2,376	10,368	1	638,633
VV^0	1,598	6,392	2	314,653

Summary

The GW code in WEST has been ported to **NVIDIA GPUs**, with excellent **performance & scalability** achieved on various supercomputers including **Perlmutter**

Large-scale full-frequency G_0W_0 calculations demonstrated on OLCF/Summit

- **25k+** NVIDIA V100 GPUs
- **10k+** valence electrons

Next steps

- Demonstrate scalability to GPUs for quantum defect embedding theory (**QDET**) calculations (newly ported)
- Expand the GPU porting to cover BSE and electron-phonon **without empty states**
- Achieve performance portability targeting **exascale** systems (ALCF/Aurora, OLCF/Frontier)

http://west-code.org

Yu and Govoni, J. Chem. Theory Comput.18, 4690-4707 (2022)

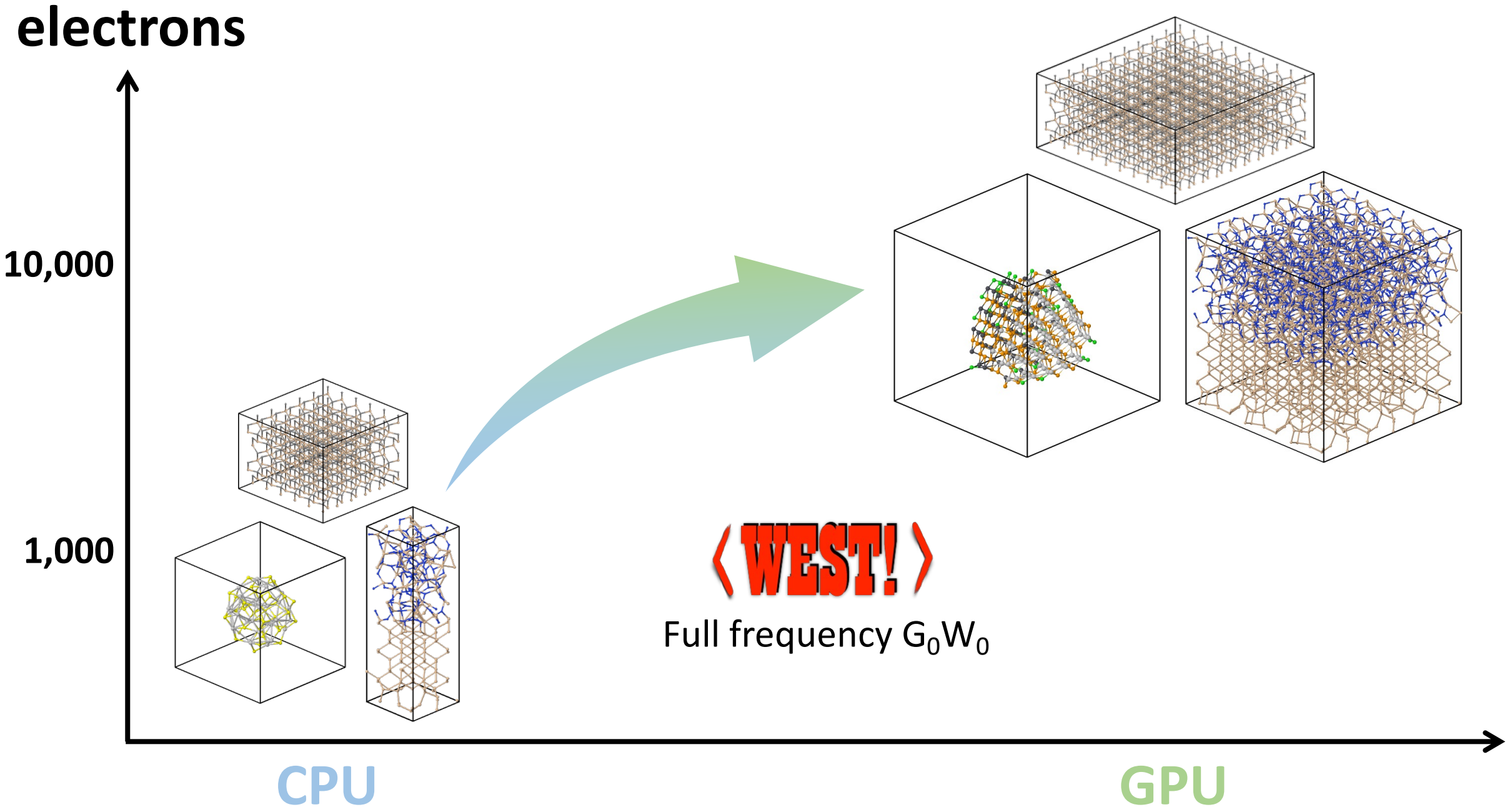
Acknowledgments

NERSC NESAP Tier-1
- Dr. Brandon Cook (LBNL)
- Dr. Soham Ghosh (LBNL)
- NERSC/Cori-GPU
- NERSC/Perlmutter

Computational resources
- OLCF/Summit (ALCC and INCITE)
- ALCF/Theta-GPU (discretionary allocation)

ANL GPU Hackathon (April 2021)
- Dr. William Huhn (ANL; now Intel)
- Dr. Kristopher Keipert (NVIDIA)

We acknowledge support by MICCoM, as part of the Computational Materials Sciences Program funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division.


electrons
10,000
1,000
CPU
GPU
WEST!
Full frequency G_0W_0