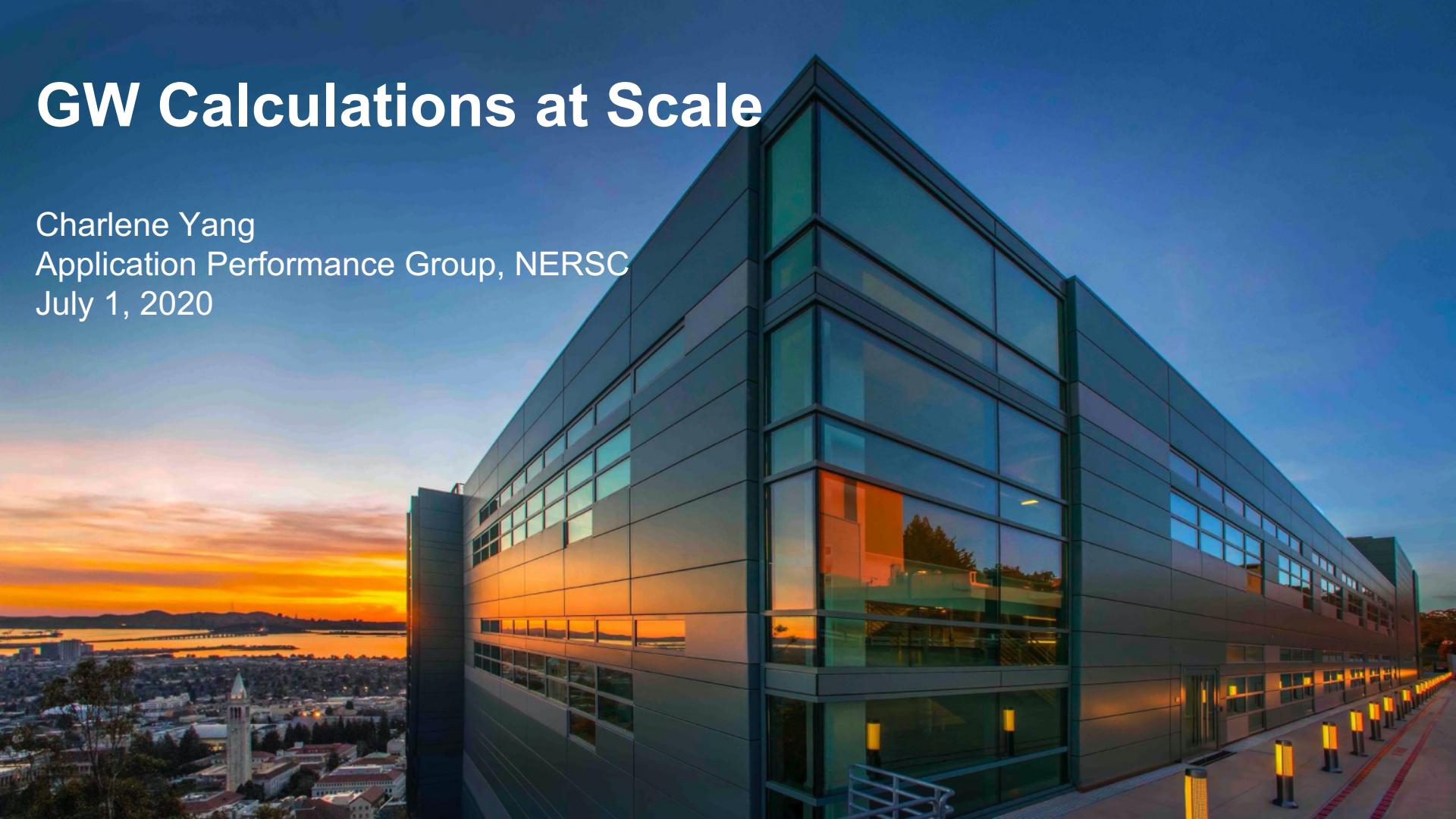
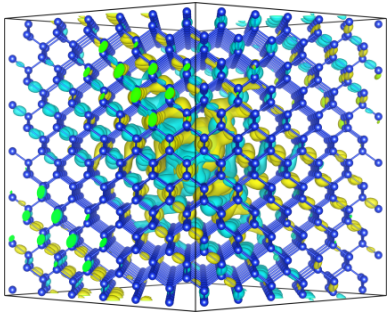
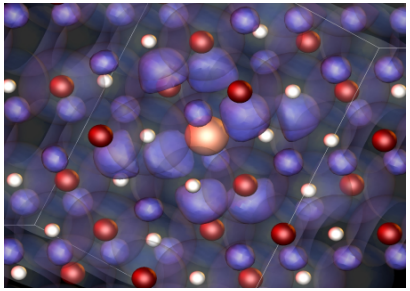
GW Calculations at Scale

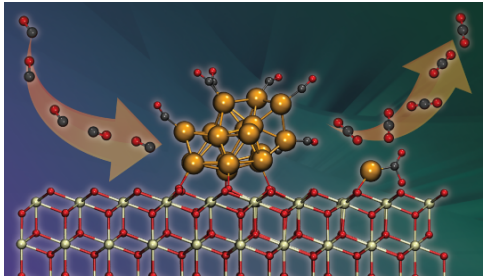
Charlene Yang
Application Performance Group, NERSC
July 1, 2020



Material Science/Chemistry at Exascale

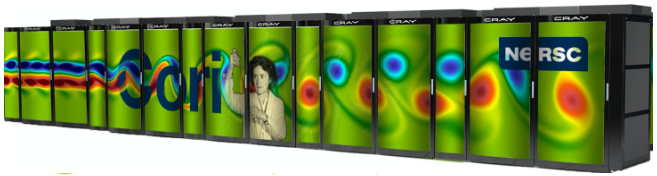






Mat. Sci & Chem apps like **VASP, Quantum ESPRESSO, NWChem, GAMESS, QMCPACK, BerkeleyGW, and CP2K** are some of the most heavily used apps at DOE facilities.

They are being used to design and understand the fundamental components of **Quantum Computers, Solar Cells, OLEDs, Batteries, Catalysts, Bio-Energy, Semiconductors, Sensors, Hydrogen Storage, Carbon Sequestration**

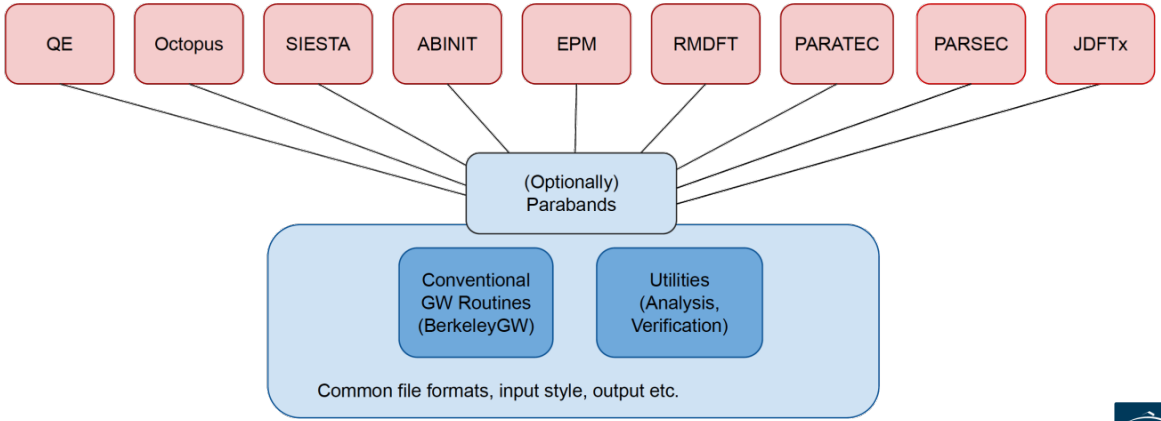


What is GW? Green's function/screened Coulomb interaction

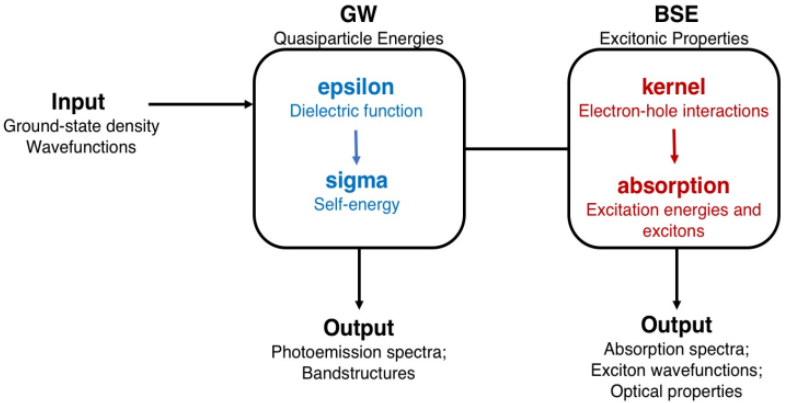
State-of-the-art to **accurately** describe many-body **excited-state** phenomena in complex materials:

- What happens when you add or remove an electron from a system?
- How do electrons behave when you apply a voltage?
- How does the system respond to light or X-rays?

https://berkeleygw.org



BerkeleyGW



Computational motifs:

- Large matrix multiplications (100k's x 10m's!)
- Fourier transforms
- Large low-rank reductions
- Eigen problems
- Matrix inversions

Scaling for computation vs memory:

- Epsilon: $O(N^4)$ vs $O(N^3)$
- Sigma: $O(N^3)$ vs $O(N^2)$

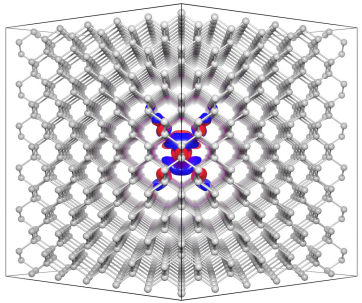
	Kernel	Computation	Memory
Epsilon	MTXEL	$O(N_v N_c N_G^\psi \log N_G^\psi)$	$O(N_v N_c N_G)$
	CHI-0	$O(N_v N_c N_G^2)$	$O(N_v N_c N_G + N_G^2)$
	Inversion	$O(N_G^3)$	$O(N_G^2)$
Sigma	MTXEL	$O(N_\Sigma N_b N_G^\psi \log N_G^\psi)$	$O(N_b N_G)$
	GPP	$O(N_\Sigma N_b N_G^2)$	$O(N_G^2 + N_b N_G)$

BerkeleyGW

Initial Code: ~100k LOC; Fortran; MPI/OpenMP on CPU

GPU porting and optimization:

- CUDA/C++ and OpenACC branches
- cuBLAS/cuFFT libraries and custom codes
- non-blocking cyclic communication scheme (comp./comm. overlapping)
- CUDA streams (comp./D-H transfers overlapping)
- batching mechanism
- pre-computing



Isosurface for one of the in-gap states associated with a divacancy defect in Silicon

Parameters	Si-214	Si-510	Si-998	SiC-998	Si-2742
N_{spin}	1	1	1	2 (↑/↓)	1
N_G^{ψ}	31,463	74,653	145,837	422,789	363,477
N_G	11,075	26,529	51,627	149,397	141,505
N_b	6,397	15,045	29,346	16,153	80,694
N_v	428	1,020	1,996	1,997/1,995	5,484
N_c	5,969	14,025	27,350	14,156/14,158	75,210
N_Σ	Variable, up to 120 per spin				
Epsilon PFLOPs	2.5	80.5	1164	10,091	66,070
Epsilon Memory (TB)	0.45	6.07	45.1	135	934
Epsilon Comm.Vol. (GB)	3.92	22.5	85.3	1428	640
Sigma PFLOPs	0.127	1.71	12.6	58.2	260.7
Sigma Memory (GB)	6.19	34.3	133.8	791.4	1006
Sigma Comm.Vol. (GB)	2.27	12.8	48.5	77.2	365.4

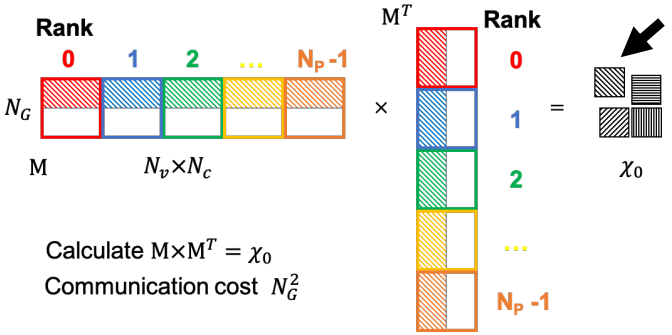
Communication Schemes

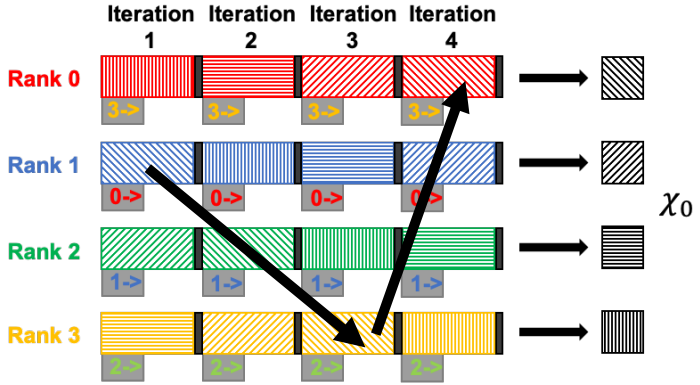
Large distributed GEMMs:

- MPI collectives (MPI_Reduce)

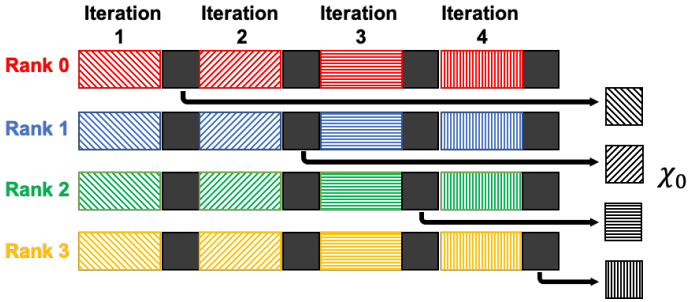
  versus

- point-to-point routines (MPI_Isend/Irecv)





non-blocking cyclic communication scheme



collective-based communication scheme

Large Low-Rank Reductions

Sigma GPP kernel:

- large matrices collapsed to a 3x1 vector
- sum across threads, thread blocks, streams, and MPI ranks
- shared memory, mixed CPU-GPU work

$$\Sigma_n = \sum_{n'} \sum_{\mathbf{GG'}} M^*_{n'n}(-\mathbf{G}) M_{n'n}(-\mathbf{G'}) \frac{\Omega^2_{\mathbf{GG'}}}{\tilde{\omega}_{\mathbf{GG'}}(E - E_{n'} - \tilde{\omega}_{\mathbf{GG'}})} v(\mathbf{G'})$$

$\tilde{\omega}$ and Ω are complex DP arrays over $\mathbf{G},\mathbf{G'}$ from polarizability
M are complex DP arrays for transition probabilities
E_n are DFT orbital energies
E is an array of "response" energies
v is the Coulomb interaction in plane-wave basis $\mathbf{G}$

loop $G' < N_G^{\text{distr}}$
. loop $G < N_G$
. . loop $n < N_{b-\text{block}}^{\text{distr.}}$
. . . Contract $P_{GG'}$ with M^l_{Gn} and $M^m_{G'n}$
. . . Accumulate $\sigma_{b-\text{block}}$ (shared memory)
Reduce σ over GPU thread blocks, CUDA streams, and MPI ranks

Useful optimizations:

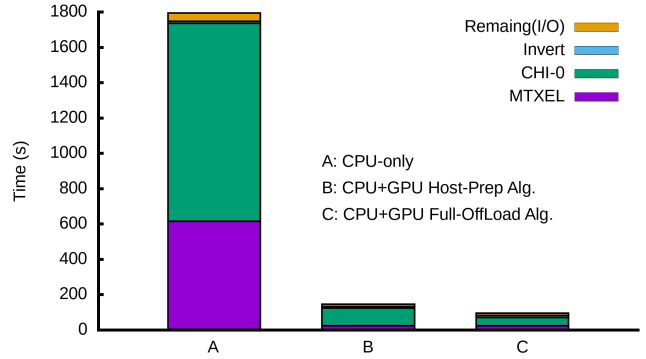
- rearrange loops to transition from bandwidth bound to compute bound region (Roofline)
- replace long latency instructions with shorter ones
- remove excessive branching
- increase occupancy

https://www.nersc.gov/users/training/events/roofline-on-nvidia-gpus-hackathon/

GPU vs CPU Speedup

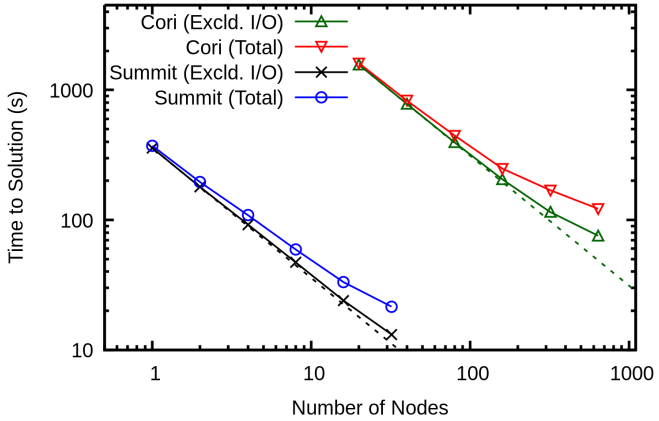
- **[Top]** Runtime comparison of **Epsilon** on Cori between Skylake CPU and V100 GPU for Si-214
- Case B prepares ZGEMM buffer on the host (Host-Prep.) while Case C on the device (Full-Offload)

18x node-to-node speedup!

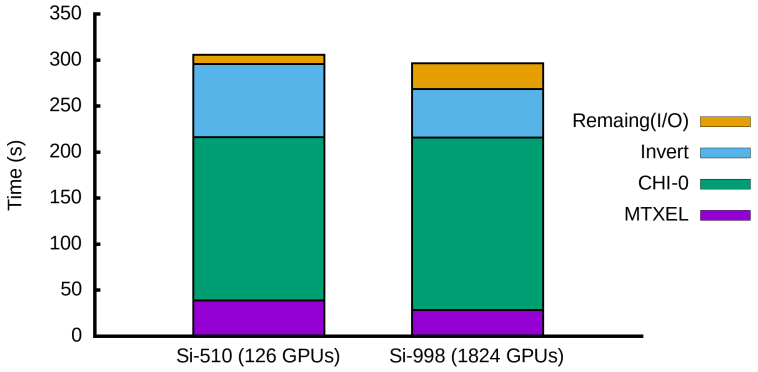


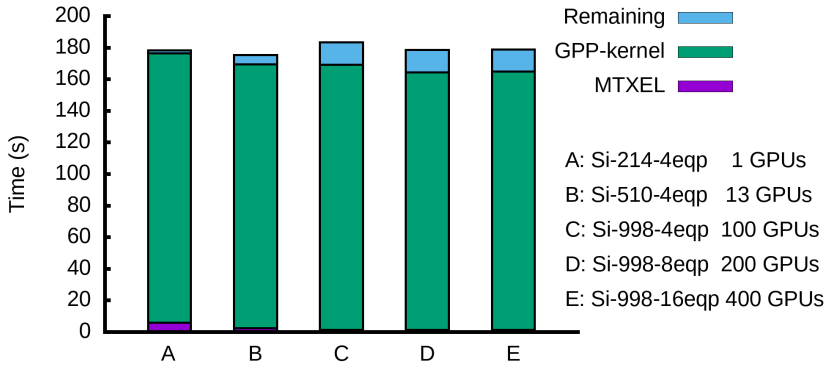
- **[Bottom]** Runtime comparison of **Sigma** between Cori Haswell CPU and Summit V100 GPU for Si-510

86x node-to-node speedup!



Weak Scaling

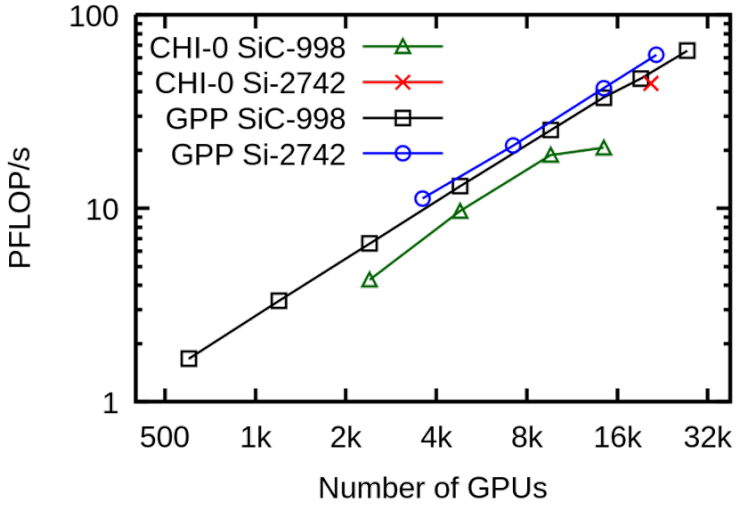


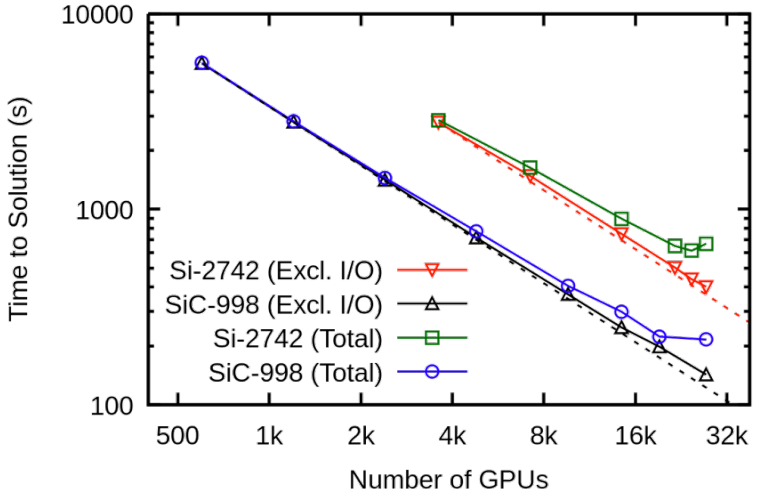


- **[Left]** Weak scaling of **Epsilon** on Summit
- The number of GPUs is scaled according to the computational complexity $O(N^4)$.

- **[Right]** Weak scaling of **Sigma** on Summit
- The number of GPUs is scaled according to the $O(N^3)$ computational complexity in Cases A, B and C, and to the number of quasiparticles in Cases C, D and E.

Strong Scaling and Best Performance





	# of GPUs	# of Pools	GPUs per Pool	Compute (s)	IO (s)	Throughput (PFLOP/s)	% of Peak
Si-2742	27,360	120	228	401	226	78.0	39.2
SiC-998	27,360	80	342	142	71	65.3	32.9

- **[Top left]** Throughput of Epsilon CHI-0 and Sigma GPP for SiC-998 and Si-2742 on Summit
- **[Top right and Bottom]** Strong scaling and best performance (PFLOP/s) of Sigma on Summit

Acknowledgement

- This research used resources of the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-05CH11231.

- This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory (ORNL), which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

- This work is supported by the Center for Computational Study of Excited-State Phenomena in Energy Materials (C2SEPEM), which is funded by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Thank You