Fully GPU-Accelerated Distributed Full Configuration Interaction Beyond Trillion-Determinant Scale

Hong Gao*†, Satoshi Imamura*, Akihiko Kasagi*, Yasuaki Ito†, Koji Nakano†
* Fujitsu Limited, Kawasaki, Japan
† Hiroshima University, Higashi-Hiroshima, Japan
Email: gao.hong@fujitsu.com

***Abstract*—Full configuration interaction (FCI) provides exact reference energies for benchmarking approximate quantum chemistry methods including emerging quantum algorithms, but its applicability is severely limited by exponential computational and memory requirements. To enable practical large-scale FCI calculations, we present a fully GPU-accelerated distributed implementation that combines chunk-based determinant partitioning, a hybrid GPU–CPU memory scheme, and GPU-native communication using NVIDIA NCCL. The proposed implementation demonstrates execution of C_2H_4/6-31G with 2.4 trillion determinants, representing the largest determinant space reported to date for an uncompressed FCI implementation. It scales efficiently on up to 1,024 GPUs, and achieves up to a 58× speedup over our previous CPU-based implementation. These results establish the practicality of trillion-determinant FCI calculations on modern GPU clusters and provide large-scale exact benchmark data for validating approximate quantum chemistry methods.**

***Index Terms*—Full configuration interaction, distributed GPU computing, scalability, hybrid memory, NCCL, quantum chemistry.**

I. Introduction

As quantum chemistry methods scale to larger systems, large-scale exact FCI benchmark calculations become increasingly important for rigorous accuracy assessment of approximate methods. However, the exponential computational cost and memory requirements of FCI limit its application to relatively small systems. Recent GPU-based implementations have substantially reduced execution time, but remain limited by the memory capacity of a single GPU [1]–[3] or a single node [4]. Consequently, the largest GPU-based FCI calculation reported to date is limited to F_2 (18e, 18o) with 2.4×10^9 determinants [4].

In this work, we extend GPU-accelerated FCI beyond the memory limitations of a single GPU and a single node by distributing both determinant data and computational workloads across 1,024 NVIDIA H100 GPUs, enabling 2.4-trillion-determinant FCI calculations.

II. Methods

Figure 1 illustrates the proposed fully GPU-accelerated distributed workflow. The determinant space is row-partitioned across GPUs, while the input vector is further divided into column chunks to reduce the GPU memory required for intermediate data. For each column chunk, the required input vector segments are exchanged through NCCL AllGather. Each GPU constructs an intermediate matrix from its assigned row block and the gathered column chunk. Dense GEMM operations are then performed using cuBLAS [5], and the resulting partial contributions are reduced across GPUs through NCCL ReduceScatter. All other local computations, including dot products, are also executed on GPUs.

To overcome GPU memory limitations, the implementation adopts a hybrid GPU–CPU memory scheme. Performance-critical data remain resident in GPU memory, whereas large auxiliary vectors are stored in CPU memory. This design substantially extends the tractable determinant space while maintaining efficient distributed execution.

TABLE I: Execution time comparison between the prior CPU-based FCI implementation [6] (t_{CPU}) and the proposed GPU-accelerated implementation (t_{GPU}).

System	N_{det}	N_{devices}	t_{CPU} [6]	t_{GPU}	Speedup
LiH/cc-pVQZ	1.3×10^7	4	29.3 m	**30 s**	58×
CH_4/cc-pVDZ	7.7×10^{10}	32	16.1 h	**18.1 m**	53×
C_3H_8/STO-3G	1.3×10^{12}	512	113.6 h	**8.6 h**	13×

III. Results

Table I summarizes the overall performance of the proposed implementation. Compared with our previous distributed CPU implementation [6], the proposed GPU implementation achieves up to a 58× speedup. Even for the largest previously reported benchmark, C_3H_8/STO-3G containing 1.3×10^{12} determinants, the execution time is reduced from 113.6 hours to 8.6 hours, corresponding to a 13× speedup. All benchmark calculations were performed on the ABCI-Q system H (two Intel Xeon Platinum CPUs, 1 TB CPU memory and four NVIDIA H100 GPUs per node) [7], using one MPI process per GPU.

Figure 2 demonstrates the strong scalability of the proposed implementation for two representative systems. The execution time decreases consistently with increasing GPU count, indicating efficient distributed execution up to 1,024 GPUs.

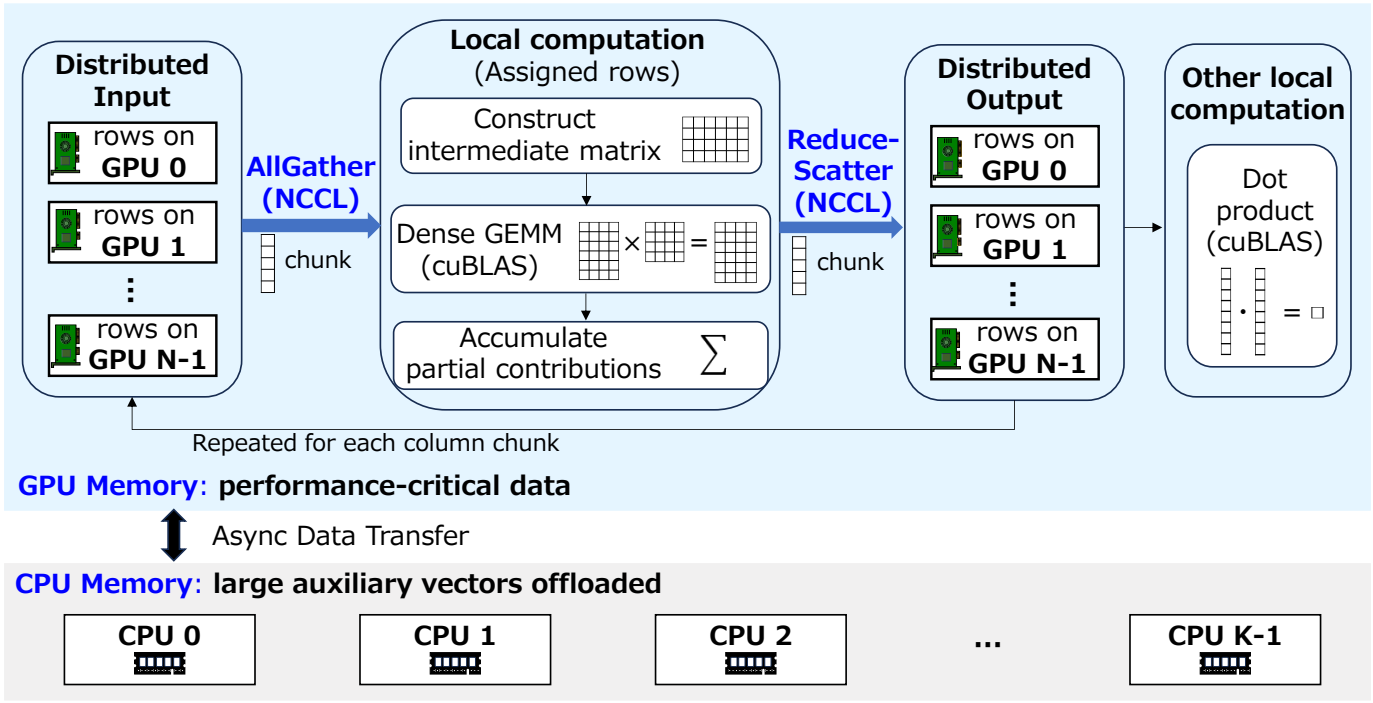


Fig. 1: Distributed workflow for processing one column chunk. All computations are performed on GPUs, while large auxiliary vectors are stored in CPU memory and transferred asynchronously.

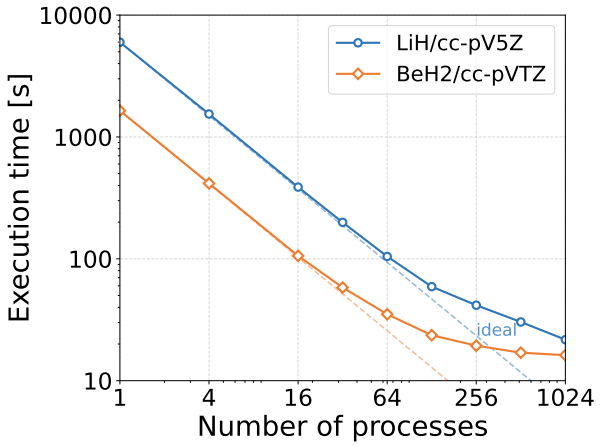


Fig. 2: Strong scalability of the proposed distributed GPU implementation.

Beyond the speedup, the proposed implementation also substantially extends the tractable determinant space. We demonstrate a 2.4×10^{12}-determinant C_2H_4/6-31G calculation in 3 hours through two Davidson iterations [8], representing the largest determinant space reported to date for an uncompressed FCI implementation. This determinant space is over three orders of magnitude larger than the largest previously reported GPU-based FCI calculation [4].

Overall, the proposed implementation substantially reduces execution time and extends the tractable determinant space far beyond previous CPU- and GPU-based implementations. These results demonstrate that practical trillion-determinant uncompressed FCI calculations are now feasible on modern GPU clusters.

IV. Conclusion

We demonstrated that practical trillion-determinant FCI calculations are feasible on 1,024 H100 GPUs. The proposed implementation achieves up to a 58× speedup over our previous CPU implementation, and extends the tractable determinant space to 2.4 trillion determinants. These advances provide large-scale exact benchmark data for developing next-generation quantum chemistry methods and quantum computing.

Acknowledgment

This work was supported by computational resources of the Quantum-AI Hybrid Computing Infrastructure (ABCI-Q), awarded under the "ABCI-Q Grand Challenge" Program of the National Institute of Advanced Industrial Science and Technology (AIST).

References

[1] B. S. Fales and B. G. Levine, "Nanoscale multireference quantum chemistry: Full configuration interaction on graphical processing units," *Journal of Chemical Theory and Computation*, vol. 11, pp. 4708–4716, 9 2015.

[2] Q. Sun, "GPU Acceleration (GPU4PySCF)," https://pyscf.org/user/gpu.html, 2026, [Online; accessed 2026].

[3] H. Gao, Y. Ito, K. Nakano, S. Imamura, A. Kasagi, and S. Tsuji, "Fully GPU-accelerated full configuration interaction for exact molecular ground-state energy calculation," in *2025 Thirteenth International Symposium on Computing and Networking (CANDAR)*. IEEE, 2025, pp. 48–57.

[4] B. S. Fales and T. J. Martínez, "Efficient treatment of large active spaces through multi-GPU parallel implementation of direct configuration interaction," *Journal of Chemical Theory and Computation*, vol. 16, no. 3, pp. 1586–1596, 2020.
[5] NVIDIA Corporation, "cuBLAS: CUDA basic linear algebra subroutines," https://developer.nvidia.com/cublas, 2026, [Online; accessed 2026].
[6] H. Gao, S. Imamura, A. Kasagi, and E. Yoshida, "Distributed implementation of full configuration interaction for one trillion determinants," *Journal of Chemical Theory and Computation*, vol. 20, no. 3, pp. 1185–1192, 2024.
[7] G-QuAT, "Quantum-AI Technology," [Online]. Available: https://unit.aist.go.jp/g-quat/G-QuAT_0617_ol_PDF%E7%94%A8r.pdf, 2025.
[8] E. R. Davidson, "The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices," *Journal of Computational Physics*, vol. 17, pp. 87–94, 1975.