

# GPU-Accelerated Massively Parallel FDTD Method for Large-Scale Electromagnetic Problems

You-Shen Tian<sup>(1) (2)</sup>, Lixiao Wang<sup>(2)</sup>, Qing Huo Liu<sup>(2)</sup>, Wen Chen<sup>(1)</sup>, and Qingtao Sun<sup>(2)</sup>

(1) Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, China

(2) School of Electronic Science and Technology, Eastern Institute of Technology, Ningbo, China

The combination of Computational Electromagnetic (CEM) methods with High-Performance Computing (HPC) technology has proven to be effective for achieving accurate and efficient electromagnetic simulation. As one of the most widely used CEM methods, the Finite-Difference Time-Domain (FDTD) method has been intensively investigated, and numerous studies have demonstrated that Graphics Processing Unit (GPU) can significantly enhance the computational efficiency of FDTD. However, the relatively limited on-board memory of GPU restricts single GPU or single-node multi-GPU implementations from solving extremely large-scale problems. To overcome this limitation, the multi-node multi-GPU FDTD scheme has become a popular solution using a hybrid MPI-CUDA parallel framework. Nevertheless, these schemes often face challenges such as the massive cross-node data transmission and the workload imbalance caused by different computational loads in GPUs, resulting in low parallel efficiency.

To address the problems associated with parallel efficiency, a series of algorithms for the multi-node multi-GPU FDTD method are developed in this work. Based on a three-dimensional domain decomposition framework, a mathematical communication model is established using the MPI Cartesian topology. This model enables the automatic selection of the optimal topology that minimizes cross-node data transmission. Additionally, a load-balancing algorithm is utilized to redistribute grids dynamically during FDTD iterations, ensuring balanced computational workloads across GPUs. To further enhance parallel efficiency, multiple CUDA streams are employed to parallelize internal field computing and boundary field data transmission, enabling effective communication-computation overlap. The proposed method is evaluated through extensive numerical experiments conducted on the supercomputer at the Computer Network Information Centre (CNIC), Chinese Academy of Sciences (CAS), which is configured with acceleration cards with performance comparable to NVIDIA V100 GPUs. Numerical results demonstrate that the proposed method is accurate and achieves notable parallel efficiency, even when deployed on over one hundred GPUs. This study introduces novel directions and possibilities for the massively parallel electromagnetic simulation.

This work was supported by the National Key R&D Program of China under Grant 2023YFB3002603.