

Parallel-in-Time Algorithm for Transient Simulation of Semiconductor Devices

Qingyuan Shi^{1,2}, Bo Lin², Dan Wu², Chijie Zhuang^{1,2*} and Rong Zeng¹

¹Department of Electrical Engineering, Tsinghua University, Beijing 100084, China

²Beijing Huairou Laboratory, Beijing 101499, China

*Correspondence should be addressed to : chijie@tsinghua.edu.cn

This paper explores the application of Parallel-in-Time (PinT) algorithm for transient semiconductor device simulations. We used the Multigrid Reduction in Time (MGRIT) algorithm to address the inefficiencies of conventional sequential time-stepping methods. Experiments on the mixed-mode device-circuit co-simulation of a 2D PiN diode's forward and reverse recovery processes demonstrate gains in computational speed and scalability through temporal parallelism across multiple cores, highlighting the potential of PinT algorithm in semiconductor device simulations. Limitations of the PinT algorithm and potential improvements are also discussed.

Index Terms—Parallel-in-Time (PinT), semiconductor device simulation, transient analysis, parallel computing.

I. INTRODUCTION

SEMICONDUCTOR device simulations play a vital role in modern electronics, enabling the design and optimization of devices ranging from diodes to complex integrated circuits. These simulations provide detailed insights of carrier dynamics, device performance, and reliability, which are critical for semiconductor device analysis [1], [2].

However, the computational cost of device transient simulation remains a significant bottleneck. Conventional time-stepping methods such as backward Euler (BE) or second-order backward differentiation formula (BDF2), process each time step sequentially, leading to inefficiencies, particularly for large-scale simulations with fine temporal resolution. These inefficiencies are complicated by the need to capture intricate multi-scale phenomena inherent to semiconductor devices.

Parallel-in-Time (PinT) algorithms [3] have demonstrated potential in domains like magnetic field [4] and power system analysis [5] by enabling simultaneous computation of multiple time steps to accelerate simulations. However, their application to semiconductor device simulations remains unexplored. This study aims to address this gap by employing the PinT algorithms to overcome the inefficiencies in transient semiconductor analysis.

II. MATHEMATICAL MODEL OF SEMICONDUCTOR DEVICE DYNAMICS

To model the carrier dynamics of semiconductor devices, the electrostatic potential ψ and densities of electrons and holes n, p are solved by Poisson's equation coupled with two continuity equations:

$$\begin{cases} \nabla^2 \psi + \frac{q}{\varepsilon} (p - n + N) = 0 \\ \frac{\partial n}{\partial t} - \frac{1}{q} \nabla \cdot \mathbf{J}_n + R_n - G_n = 0 \\ \frac{\partial p}{\partial t} + \frac{1}{q} \nabla \cdot \mathbf{J}_p + R_p - G_p = 0 \end{cases} \quad (1)$$

where q is the elementary charge and ε is the semiconductor dielectric constant; N denotes the given net doping concentration; G_n, G_p, R_n, R_p denote the generation and recombination

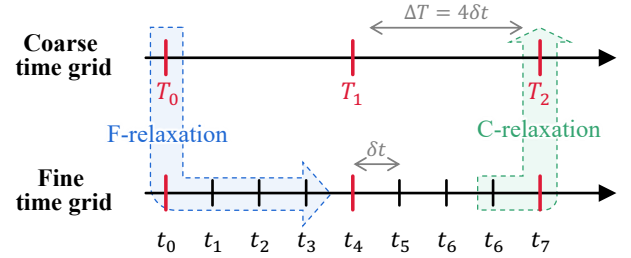


Fig. 1. Illustration of the MGRIT algorithm with 2 levels.

rates of electrons and holes, respectively; the current densities $\mathbf{J}_n, \mathbf{J}_p$ are modeled by drift and diffusion terms

$$\begin{aligned} \mathbf{J}_n &= -q\mu_n n \nabla (\psi + V_T \ln(n_{ie})) + qD_n \nabla n \\ \mathbf{J}_p &= -q\mu_p p \nabla (\psi - V_T \ln(n_{ie})) - qD_p \nabla p \end{aligned} \quad (2)$$

where μ_n, μ_p and D_n, D_p denote mobility and diffusion coefficients of electrons and holes; V_T denotes thermal voltage and n_{ie} denotes effective intrinsic density. Details on these physical models can be found in [1].

Existing numerical solvers typically discretize these equations using finite volume or stabilized finite element methods [6] for spatial terms and BE or BDF2 for time. Transient simulations rely on serial time-stepping, where each time step is solved sequentially using Newton's method. Parallelization in these methods is generally limited to matrix assembly and linear solvers [7], failing to utilize temporal parallelism.

III. THE MULTIGRID REDUCTION IN TIME ALGORITHM

The Multigrid Reduction in Time (MGRIT) [3] is a PinT algorithm that extends the principles of spatial multigrid to time domain, enabling parallel computations by coarsening time grids and iteratively refining solutions. As illustrated in Fig. 1, MGRIT operates on a hierarchy of temporal grids with fine-grid points (F-points, in black) representing all time steps, and coarse-grid points (C-points, in red) forming a subset with reduced resolution. This hierarchy allows for iterative updates that alternate between local relaxation and global corrections.

The core operations of MGRIT are F- and C-relaxations, shown by the blue and green arrows in Fig. 1. In F-relaxation, F-points are updated by the original time-stepping scheme (e.g., BE or BDF2), with Newton's method solving equation(1). C-relaxation restricts the fine time grid solution to the coarse grid, where a reduced system is solved. The coarse grid corrections are then interpolated back to refine the fine-grid solution and achieve global convergence.

The core MGRIT iteration can be expressed as:

$$\mathbf{u}_F^{(k+1)} = \mathbf{u}_F^{(k)} + I_{C \rightarrow F} \left(\mathbf{u}_C^{(k+1)} - R_{F \rightarrow C} \mathbf{u}_F^{(k)} \right) \quad (3)$$

where $\mathbf{u}$ is the solution, k is iteration number, temporal restriction operator $R_{F \rightarrow C}$ restricts the fine time grid solution to the coarser time grid, and interpolation operator $I_{C \rightarrow F}$ interpolates the coarse time grid correction back to the fine time grid.

Algorithm 1 MGRIT

- 1: **Repeat**
- 2: Relax the solution with F- and C-relaxation in parallel in Fig. 1
- 3: Compute the residual on the coarse grid $\mathbf{u}_C^{(k+1)} - R_{F \rightarrow C} \mathbf{u}_F^{(k)}$
- 4: Solve the coarse-grid correction problem sequentially
- 5: Correct the solution at C-points
- 6: Update the solution at the F-points with F-relaxation in Fig. 1
- 7: **until** norm of residual is small enough

The MGRIT efficiently parallelizes steps 2, 5, and 6 in Algorithm 1. Both F- and C-relaxations can be conducted concurrently, enabling parallel updates for local solutions and coarse-grid corrections. This enhances efficiency over sequential time-stepping and makes MGRIT particularly suitable for large-scale simulations requiring fine temporal resolution.

IV. NUMERICAL EXPERIMENTS

To validate the PinT algorithm, transient simulations were performed on a PiN diode coupled with external circuit. The diode's doping profile, computational mesh, and external voltage source are shown in Fig. 2. Using a mesh of 2432 triangular cells, the simulation computed the transient behavior of 10 microseconds, divided into 16384 uniform time steps. The forward and reverse recovery processes of the diode, driven by carrier dynamics, were captured in Fig. 3.

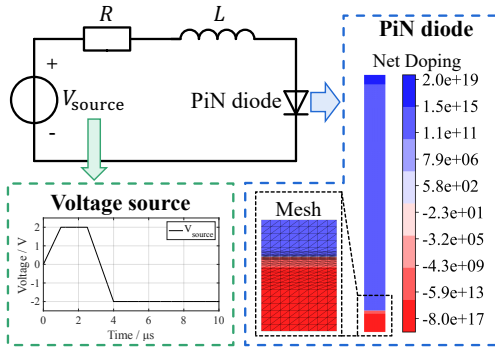


Fig. 2. The V-R-L-D test circuit with a PDE-modeled PiN diode.

Strong scaling tests were conducted to evaluate performance. Using the same spatial mesh, Fig. 4 demonstrates the efficiency of MGRIT compared to conventional serial time-stepping methods. The backward Euler method, combined with multi-threaded parallel matrix assembly and parallel PARDISO direct

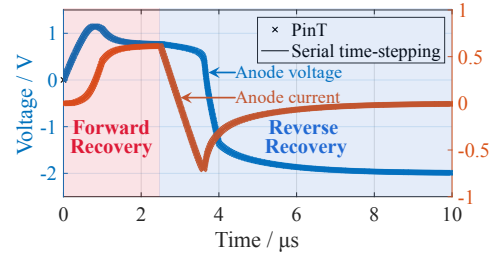


Fig. 3. Transient simulation result of the PiN diode.

solver, exhibited limited scalability. Speedup saturated at 4 CPU cores due to thread-switching overheads, with performance declining as more cores were added.

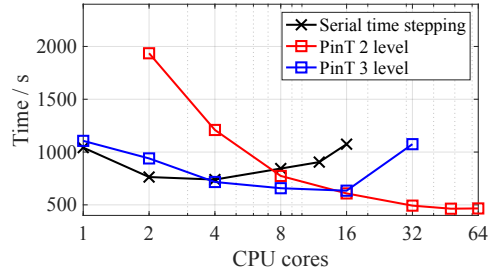


Fig. 4. Parallel speedup comparison between PinT and conventional method.

In contrast, the MGRIT demonstrated better scalability, efficiently utilizing up to 64 cores. This scalability highlights its advantage in simulations requiring high temporal resolution.

V. SUMMARY

This work marks the first step in employing PinT method in semiconductor device modeling. By exploiting temporal parallelization, MGRIT algorithm accelerates devices' transient simulations, although the parallel efficiency is not optimal. Future work will explore adaptive time-stepping and better multigrid relaxation techniques to improve parallel efficiency.

ACKNOWLEDGMENT

This work is supported in part by National Science Foundation of China under Grant U2166602 and Grant 524B2107.

REFERENCES

- [1] S. Selberherr, *Analysis and simulation of semiconductor devices*. Springer Science & Business Media, 1984.
- [2] Q. Shi, C. Zhuang, B. Lin, D. Wu, L. Li and R. Zeng "Adaptive Local Mesh Refinement in 2-D Finite Volume Method for Steady State and Transient Simulations of Semiconductor Devices," *IEEE Trans. Magn.*, vol. 60, Dec. 2024, Art. no. 7202604.
- [3] A. J. Howse, H. D. Sterck, R. D. Falgout, S. MacLachlan, and J. Schroder, "Parallel-in-Time Multigrid with Adaptive Spatial Coarsening for the Linear Advection and Inviscid Burgers Equations," *SIAM J. Sci. Comput.*, vol. 41, pp. A538-A565, 2019.
- [4] Y. Takahashi, K. Fujiwara, T. Iwashita and H. Nakashima, "Parallel Finite-Element Method Based on Space-Time Domain Decomposition for Magnetic Field Analysis of Electric Machines," *IEEE Trans. Magn.*, vol. 55, pp. 1-4, Jun. 2019.
- [5] T. Cheng, N. Lin and V. Dinavahi, "Hybrid parallel-in-time-and-space transient stability simulation of large-scale AC/DC grids," *IEEE Trans. Power Syst.*, vol. 37, pp. 4709-4719, Nov. 2022.
- [6] Q. Shi, Y. Cai, C. Zhuang, B. Lin, D. Wu, R. Zeng and W. Bao, "A robust hybridizable discontinuous Galerkin scheme with harmonic averaging technique for steady state of real-world semiconductor devices," *J. Comput. Phys.*, vol. 519, Dec. 2024, Art. no. 113422.
- [7] P. T. Lin and J. N. Shadid, "Towards large-scale multi-socket, multicore parallel simulations: Performance of an MPI-only semiconductor device simulator," *J. Comput. Phys.*, vol. 229, pp. 6804-6818, Sept. 2010.