

GPU-Accelerated Feature-Scale Process Simulation of CFET Structures

Tobias Reiter^{†*} and Lado Filipovic[†]

[†]Christian Doppler Laboratory for Multi-Scale Process Modeling of Semiconductor Devices and Sensors,
Institute for Microelectronics, TU Wien, 1040 Vienna, Austria

*Email: reiter@iue.tuwien.ac.at

Abstract—Accurate feature-scale process simulation is essential for advanced 3D device architectures such as CFETs, but physics-based methods are computationally expensive and time-consuming. In this work, we present GPU-accelerated etching and deposition simulations within the ViennaPS framework and compare them to geometric emulation and CPU-based simulation for a representative CFET process flow. GPU acceleration reduces the runtime for physics-based models from several hours to minutes while employing the same physical models as the CPU implementation. The results show that this approach enables efficient and realistic simulation of complex 3D fabrication processes.

Index Terms—Feature-Scale Process Simulation, Topography Simulation, GPU Acceleration, Monte Carlo Ray Tracing, Etching and Deposition

I. INTRODUCTION

The continuous scaling of semiconductor devices has led to increasingly complex three-dimensional architectures, such as complementary field-effect transistors (CFETs) [1]–[3]. These structures place high demands on process simulation, as accurate prediction of feature evolution requires detailed modeling of transport and surface reactions.

Physics-based feature-scale simulations, such as Monte Carlo ray tracing combined with level-set surface evolution, provide the required physical detail but are computationally expensive. As a result, two main approaches are commonly used. Geometric emulation [4] offers faster turnaround times but neglects important physical effects, often leading to unrealistic structures. In contrast, CPU-based simulations capture the relevant physics but suffer from long runtimes, limiting their practical use in complex 3D process flows.

In this work, we address this trade-off by leveraging GPU acceleration within the ViennaPS framework [5]. We present a systematic comparison of emulation, CPU-based simulation, and GPU-accelerated simulation for a representative CFET process flow. The results demonstrate that GPU acceleration enables physics-based simulations with realistic 3D structures at significantly reduced runtimes, making them practical for advanced process development.

II. NUMERICAL METHODS

In the ViennaPS framework, material interfaces are represented using a level-set function $\phi(\vec{x}, t)$, whose zero contour defines the evolving surface [6]. Surface evolution is governed by

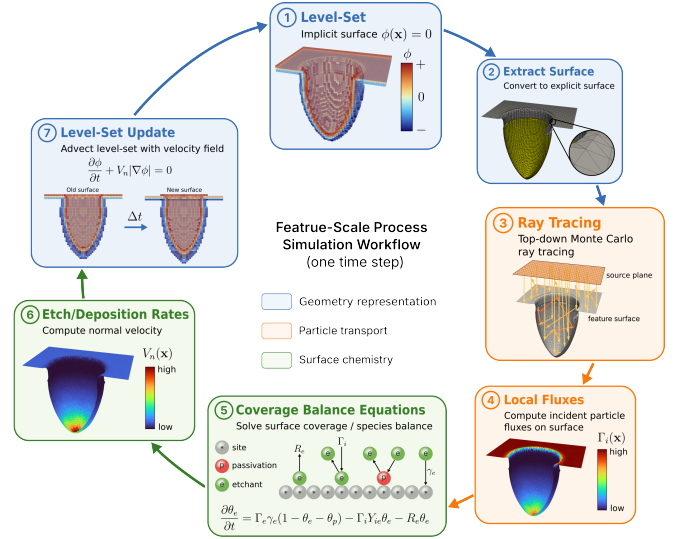


Fig. 1. Simulation workflow in ViennaPS. The level-set surface is converted to an explicit representation for Monte Carlo particle transport. Computed particle fluxes are used to evaluate surface reactions and local etch/deposition rates, which define the normal velocity for level-set advection. GPU acceleration is applied to the ray tracing step using NVIDIA OptiX.

$$\frac{\partial \phi}{\partial t} + V_n(\vec{x}) \|\nabla \phi\| = 0, \quad (1)$$

where V_n denotes the local normal velocity. The simulation workflow is illustrated in Fig. 1. For each time step, the implicit level-set surface is first converted into an explicit surface representation suitable for particle transport calculations. Particle fluxes are then computed using a top-down Monte Carlo ray tracing approach.

At the feature scale, the mean free path of process species is typically comparable to or larger than the device dimensions, allowing transport to be modeled as ballistic [7]. Pseudo-particles are emitted from a source plane and traced until they intersect the surface or leave the computational domain. Different transport models can be applied for neutral and ion species, including directional incidence, diffuse reflection, and angle-dependent surface interactions.

The resulting local particle fluxes are used to evaluate surface coverages and reaction rates. These rates are converted into a normal velocity V_n , which is used to advance the level-

set function and update the surface geometry for the next time step.

The computational bottleneck of this workflow is the repeated ray-surface intersection test during flux evaluation. To accelerate this step, a bounding volume hierarchy is constructed over the extracted surface geometry. CPU simulations employ Intel Embree [8] for ray traversal, while GPU simulations use NVIDIA OptiX [9]. Both implementations use identical physical models and differ only in the hardware backend used for ray tracing.

A. Surface Reaction Models

The plasma etching steps are described by coverage-based surface reaction models [10], [11]. Ions provide directional energy and sputtering, neutral radicals provide chemical etchants, and passivating species inhibit reactions on sidewalls or shadowed regions. The particle fluxes are computed by ray tracing and enter local site-balance equations for the surface coverages.

For a silicon plasma etch, the coverages represent the fraction of available surface sites occupied by reactive or passivating species. Since surface reactions are fast compared with the topography evolution, the coverages are evaluated in a quasi-steady state. A representative balance for an etchant coverage θ_e and a passivation coverage θ_p has the form

$$0 = \Gamma_e \gamma_e (1 - \theta_e - \theta_p) - \Gamma_i Y_{ie} \theta_e - R_e \theta_e, \quad (2)$$

$$0 = \Gamma_p \gamma_p (1 - \theta_e - \theta_p) - \Gamma_i Y_p \theta_p, \quad (3)$$

where γ_e and γ_p are sticking probabilities, Y_{ie} and Y_p describe ion-enhanced removal, and R_e summarizes thermal or chemical desorption. The normal velocity is obtained from chemical, ion-enhanced, and physical sputtering contributions:

$$V_n \propto R_{\text{chem}} \theta_e + \Gamma_i Y_{ie} \theta_e + \Gamma_i Y_{sp} (1 - \theta_e). \quad (4)$$

This formulation captures the competition between radical adsorption, ion-enhanced removal, and passivation.

Conformal deposition and selective material removal are modeled using single-particle flux approaches, which are appropriate when detailed plasma chemistry is not the dominant mechanism. In contrast, selective epitaxial growth is described by an orientation-dependent level-set model [12], as the process is governed primarily by surface growth kinetics rather than ballistic particle transport.

III. SIMULATION METHODOLOGY

A. Process Sequence

The evaluation is performed on a CFET structure consisting of vertically stacked nanosheet devices. A simplified process flow is used, where computationally critical steps are simulated using physics-based models and less critical steps are approximated using geometric operations.

The process sequence starts with the formation of a stacked Si/SiGe multilayer structure, followed by fin patterning and shallow trench isolation (STI). A dummy gate is then introduced to define the gate region. Subsequent steps prepare the

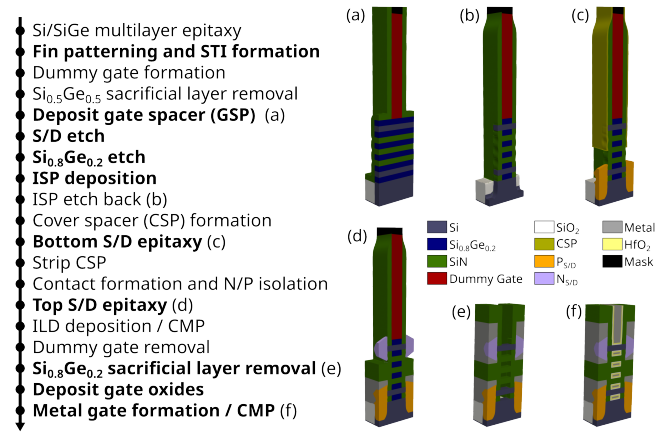


Fig. 2. CFET process flow with intermediate geometries at key stages. Top row: (a) after dummy gate formation and GSP deposition, (b) after S/D recess etch, and (c) after bottom S/D epitaxy. Bottom row: (d) after top S/D epitaxy, (e) after SiGe sacrificial layer removal, and (f) final device structure.

structure for source/drain formation by selectively removing sacrificial layers and forming spacers. The resulting geometry contains narrow cavities, stacked material interfaces, and shadowed regions around the future gate-all-around channels, making it a useful benchmark for comparing geometric emulation with particle-based simulation. The full sequence is shown in Fig. 2.

The key simulation steps focus on processes that strongly affect the final geometry. Plasma etching is used for pattern transfer and source/drain recess formation and is modeled using coverage-based plasma etching models. Conformal deposition and selective material removal are represented by single-particle-flux models, while selective epitaxial growth of the source/drain regions is captured using a level-set based growth model. All remaining steps, such as spacer definition, contact formation, and planarization, are treated using geometric emulation to reduce computational cost.

The choice of physical simulation steps is based on expected sensitivity to nonlocal transport. Plasma etches and conformal depositions alter the exposed surface area and can therefore change the boundary conditions for later selective steps. In contrast, steps that mainly define nominal layer thicknesses or remove open regions are less sensitive to shadowing and are represented geometrically. This separation keeps the process flow computationally manageable while preserving the transport-limited mechanisms most likely to affect the final CFET geometry.

B. Simulation Modes

Three simulation modes are used to study the trade-off between runtime and physical accuracy.

a) Emulation: In the emulation mode, all process steps are performed using purely geometric operations. No physical transport or surface reaction mechanisms are considered. This approach provides the fastest execution times but does not

capture important effects such as shadowing or aspect-ratio-dependent transport.

b) *CPU Simulation*: In the CPU-based simulation mode, all selected process steps are modeled using physically based models. Particle transport is computed via Monte Carlo ray tracing, and surface evolution is performed using the level-set method. All computations are executed on the CPU.

c) *CPU + GPU Simulation*: The GPU-based simulation mode uses the same physical models as the CPU simulation. The only difference lies in the computational backend, where ray tracing and associated calculations are executed on the GPU. This enables a high degree of parallelism and significantly reduces computation time while maintaining identical physical behavior.

IV. RESULTS

A. Numerical Setup

The simulations are performed on a three-dimensional computational domain with lateral dimensions of $42 \times 40 \text{ nm}^2$ and a vertical extent of up to 200 nm. The domain is discretized using a uniform grid spacing of 0.43 nm, which provides sufficient resolution to accurately capture the evolving surface geometry.

The number of rays per time step is adapted to the surface complexity and scales with the number of surface elements. Between 10 million and 150 million rays are traced per level-set advection step, and each process step consists of several tens to a few hundred time steps. The level-set time step is selected based on the maximum local velocity to satisfy the advection stability condition, while flux smoothing is applied locally to reduce Monte Carlo noise without changing the mean rate.

Simulations are performed on a system with an *Intel Core i7-12700K* CPU using 16 threads and an *NVIDIA GeForce RTX 4070 Ti Super* GPU. Performance is evaluated using wall-clock time for individual process steps and the complete process flow. The measured time includes source sampling, ray traversal, coverage evaluation, velocity-field construction, level-set advection, surface extraction, acceleration-structure construction, and host-device synchronization. This definition captures the runtime relevant for process-flow studies rather than isolated kernel throughput.

B. Runtime Comparison

Table I summarizes the wall-clock time for the three modes. Emulation is fastest, requiring 2 min 44 s, while the full CPU-based simulation requires 3 h 57 min. GPU acceleration reduces the runtime to 16 min 56 s, corresponding to an overall speedup of approximately $14\times$ compared to the CPU implementation. The largest improvements are observed in transport-dominated steps such as S/D etching and spacer deposition. Epitaxy steps show only minor differences because they are not dominated by ray tracing.

The runtime distribution reflects the numerical cost of the individual process models. The S/D etch, gate spacer deposition, and inner spacer deposition require many repeated

TABLE I
EXECUTION TIME OF SELECTED PROCESS STEPS. STEP TIMES ARE GIVEN IN SECONDS.

Step	Emulation [s]	CPU [s]	CPU + GPU [s]
Fin patterning and STI	2.016	633.429	67.699
Deposit gate spacer	6.797	1933.883	101.860
S/D etch	49.964	7819.898	529.718
SiGe ₂ etch	1.043	1350.101	71.887
Inner spacer deposition	6.776	1454.158	77.446
Bottom S/D epitaxy	41.096	44.677	49.059
Top S/D epitaxy	26.566	25.838	28.475
SiGe sacrificial removal	0.408	611.240	36.696
Deposit gate oxides	1.552	117.580	7.101
Metal gate formation	6.215	237.794	23.264
Total	2 min 44 s	3 h 57 min 32 s	16 min 56 s

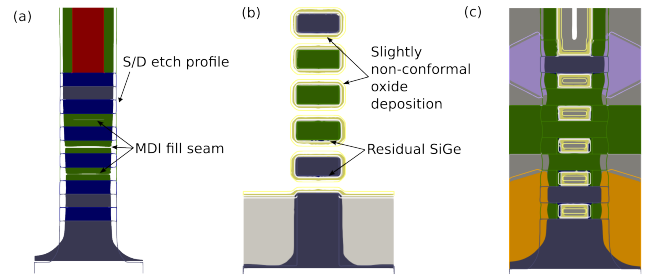


Fig. 3. Comparison of simulated (filled) and emulated (lines) structures. (a) Channel slice after S/D plasma etching. (b) Gate slice after oxide deposition. (c) Final channel slice, showing accumulated differences.

flux evaluations on geometries with narrow openings and shadowed sidewalls. In these steps, only a small fraction of the emitted particle flux reaches deeply recessed surfaces, so many rays are required to obtain smooth local rates. GPU acceleration reduces this dominant cost directly. In contrast, epitaxial growth is governed mainly by level-set advection and orientation-dependent growth rates, so replacing the ray-tracing backend has little influence.

C. Structural Comparison

Figure 3 compares geometries obtained from geometric emulation and physics-based simulation. Clear deviations appear after the S/D plasma etching step and accumulate in later steps. The simulated structure shows non-uniform deposition and etch profiles due to limited transport, surface visibility, and coverage-dependent surface chemistry, whereas the emulated result assumes idealized behavior. These differences affect the available space for channel release, source/drain epitaxy, and gate-stack formation, showing that geometric emulation can miss effects that are critical in advanced 3D structures.

For the coverage-based plasma etch model, the relevant quantity is not only the total incoming flux but the local ratio between ion, etchant, and passivating species. A region that is visible to neutral radicals but weakly exposed to ions can accumulate passivation and etch more slowly, while surfaces receiving a stronger ion flux experience enhanced removal.

This mechanism leads to non-uniform recess profiles and modified corner evolution. Once such deviations are created, subsequent deposition or epitaxy steps operate on a different exposed surface area and can amplify the difference between emulation and simulation.

The structural comparison, therefore, highlights the role of coupled process steps in CFET fabrication. A change in the S/D recess profile modifies the surface exposed to subsequent epitaxy, and deviations in spacer or oxide thickness alter access paths for later selective removal and gate-stack formation. The final geometry contains accumulated differences rather than an isolated error from a single etch step, which is why the complete process flow is simulated instead of comparing individual steps only.

V. DISCUSSION

The comparison illustrates why geometric emulation remains useful but insufficient on its own. Emulation provides fast estimates of the nominal process sequence, but it assumes that process rates can be represented by prescribed displacements or visibility rules. This assumption becomes restrictive for certain structures, where high-aspect-ratio access and material selectivity produce strongly local process conditions. A practical strategy is therefore to use emulation for steps with weak transport sensitivity and to reserve physical simulation for steps where local flux balance determines the final geometry.

The GPU backend does not change the underlying surface reaction model, but it reduces the cost of evaluating the fluxes required by that model. Predictive accuracy still depends on calibrated model parameters, such as sticking probabilities, sputter yields, and reaction-rate parameters. The remaining uncertainty is therefore model-related rather than backend-related, while the acceleration expands the number of process variants and parameter sensitivities that can be evaluated with physics-based models.

The investigated flow is simplified and uses emulation for several integration steps. More detailed process descriptions would likely increase the impact of local flux balance on the final geometry. GPU acceleration therefore not only reproduces CPU results faster but also enables more detailed physics-based simulations for complex CFET process studies.

VI. CONCLUSION

In this work, we presented a GPU-accelerated approach to feature-scale process simulation using the ViennaPS framework and evaluated its performance on a representative CFET process flow. A direct comparison between geometric emulation, CPU-based simulation, and GPU-accelerated simulation was carried out to assess both runtime and structural accuracy.

The results show that GPU acceleration reduces the runtime of physics-based simulations from several hours to less than 20 minutes, corresponding to a speedup of more than one order of magnitude. This improvement is most pronounced for transport-dominated processes such as plasma etching

and conformal deposition, where Monte Carlo ray tracing constitutes the main computational bottleneck.

The presented approach enables efficient and realistic process simulation for complex device architectures. This opens the door for routine use of physics-based models in process development and optimization workflows, where both accuracy and turnaround time are critical.

ACKNOWLEDGMENT

Financial support by the Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development, and the Christian Doppler Research Association is gratefully acknowledged.

REFERENCES

- [1] J. Ryckaert, P. Schuddinck, P. Weckx, G. Bouche, B. Vincent, J. Smith *et al.*, "The Complementary FET (CFET) for CMOS scaling beyond N3," *Proc. IEEE Symp. VLSI Technol.*, pp. 141–142, 2018.
- [2] S. Demuynck, V. Vega Gonzalez, C. Cavalcante, L.P.B. Lima, K. Stiers, C. Sheng *et al.*, "Monolithic Complementary Field Effect Transistors (CFET) Demonstrated using Middle Dielectric Isolation and Stacked Contacts," *Proc. IEEE Symp. VLSI Technol. Circuits*, pp. 1–2, 2024.
- [3] C. Cavalcante, S. Demuynck, C. Sheng, D. Batuk, M. Hosseini, A. Peng *et al.*, "Monolithic CFET Flow Improvements Integrating Cover Spacer and Dual-WF RMG," *Proc. IEEE Symp. VLSI Technol. Circuits*, pp. 1–3, 2025.
- [4] X. Klemenschits, L. Filipovic and S. Selberherr, "Geometric advection and its application in the emulation of high aspect ratio structures," *Comput. Methods Appl. Mech. Eng.*, vol. 386, 114196, 2021.
- [5] T. Reiter and L. Filipovic, "ViennaPS: A flexible framework for semiconductor process simulation," *SoftwareX*, vol. 32, 102453, 2025.
- [6] J. A. Sethian, "Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science," 2nd Edition, Cambridge University Press, Cambridge, U.K., 1999.
- [7] D.J. Economou, "Modeling and simulation of plasma etching reactors for microelectronics," *Thin Solid Films*, vol. 365(2), pp. 348–367, 2000.
- [8] I. Wald, S. Woop, C. Benthin, G.S. Johnson, M. Ernst, "Embree: a kernel framework for efficient CPU ray tracing," *ACM Trans. Graph.*, vol. 33(4), pp. 1–8, 2014.
- [9] S. Parker, J. Bigler, A. Dietrich, H. Friedrih, J. Hoberbock *et al.*, "OptiX: a general purpose ray tracing engine," *ACM Trans. Graph.*, vol. 29(4), pp. 1–13, 2010.
- [10] R. J. Belen, S. Gomez, M. Kiehlbauch, D. Cooperberg, and E. S. Aydil, "Feature-scale model of Si etching in SF₆ plasma and comparison with experiments," *J. Vac. Sci. Technol. A: Vac. Surf. Films*, vol. 23, 2005.
- [11] R. J. Belen, S. Gomez, D. Cooperberg, M. Kiehlbauch, and E. S. Aydil, "Feature-scale model of Si etching in SF₆/O₂ plasma and comparison with experiments," *J. Vac. Sci. Technol. A: Vac. Surf. Films*, vol. 23, 2005.
- [12] A. Toifi, M. Quell, X. Klemenschits, P. Manstetten, A. Hössinger, S. Selberherr *et al.*, "The level-set method for multi-material wet etching and non-planar selective epitaxy," *IEEE Access*, vol. 8, pp. 115406–115422, 2020.