Cluster Computing for Transient Simulations of the Linear Boltzmann Equation on Irregular Three-Dimensional Domains

Matthias K. Gobbert[1], Mark L. Breitenbach[1], and Timothy S. Cale[2]

[1] Department of Mathematics and Statistics,
University of Maryland, Baltimore County,
1000 Hilltop Circle, Baltimore, MD 21250, U.S.A
[2] Focus Center — New York, Rensselaer,
Interconnections for Hyperintegration,
Isermann Department of Chemical and Biological Engineering,
Rensselaer Polytechnic Institute, CII 6015,
110 8th Street, Troy, NY 12180-3590, U.S.A

Abstract. Processes used to manufacture integrated circuits take place at a range of pressures and their models are of interest across a wide range of length scales. We present a kinetic transport and reaction model given by a system of linear Boltzmann equations that is applicable to several important processes that involve contacting in-process wafers with reactive gases. The model is valid for a range of pressures and for length scales from micrometers to centimeters, making it suitable for multiscale models. Since a kinetic model in three dimensions involves discretizations of the three-dimensional position as well as of the three-dimensional velocity space, millions of unknowns result. To efficiently perform transient simulations with many time steps, the size of the problem motivates the use of parallel computing. We present simulation results on an irregular three-dimensional domain that highlights the capabilities of the model and its implementation, as well as parallel performance studies on a distributed-memory cluster show that the computation time scales well with the number of processes.

1 Introduction

Many important manufacturing processes for integrated circuits involve the flow of gaseous reactants at pressures that range from very low to atmospheric. Correspondingly, the mean free path λ (the average distance that a molecule travels before colliding with another molecule) ranges from less than 0.1 micrometers to over 100 micrometers. The typical size of the electronic components (called 'features' during processing) is now below 1 micrometer and the size of the chemical reactor, in which the gas flow takes place, is on the order of decimeters. Thus, models on a range of length scales L^* are of interest, each of which needs to be appropriately selected to be valid on its length scale. These authors have in the

V.S. Sunderam et al. (Eds.): ICCS 2005, LNCS 3516, pp. 41–48, 2005.

past coupled models on several length scales, from feature scale to reactor scale, to form an interactive multiscale reactor simulator [1, 2]. The models used were based on continuum models in all but the feature scale and assumed moderately high pressure to be valid. The current work provides the basis for extending this work to more general pressure regimes. Such multiscale models require well-tested and validated models and numerical methods on every scale of interest. The following results address both the effectiveness of the model and computational efficiency of the numerical method and its parallel implementation.

The appropriate transport model at a given combination of pressure and length scale is determined by the Knudsen number Kn, defined as the ratio of the mean free path and the length scale of interest $\mathrm{Kn} := \lambda/L^*$, which arises as the relevant dimensionless group in kinetic models [3]: (i) For small values $\mathrm{Kn} < 0.01$, continuum models describe the gas flow adequately. (ii) At intermediate values $\mathrm{Kn} \approx 1.0$, kinetic models based on the Boltzmann transport equation capture the influence of both transport of and collisions among the molecules. (iii) For large values $\mathrm{Kn} > 100.0$, kinetic models remain valid with the collision term becoming negligible in the limit as Kn grows.

Our interest includes models on the micro- to meso-scale at a range of pressures, resulting in Knudsen numbers ranging across the wide spectrum from less than $\mathrm{Kn} = 0.1$ to $\mathrm{Kn} \to \infty$, in the regimes (ii) and (iii) above. For flow in a carrier gas, assumed inert, at least an order of magnitude denser than the reactive species, and in spatially uniform steady-state, we have developed a kinetic transport and reaction model (KTRM) [4, 5], given by the system of linear Boltzmann equations for all n_s reactive species in dimensionless form

$$\frac{\partial f^{(i)}}{\partial t} + \mathbf{v} \cdot \nabla_{\mathbf{x}} f^{(i)} = \frac{1}{\mathrm{Kn}}\, Q_i(f^{(i)}), \quad i = 1, \ldots, n_s, \tag{1}$$

with the linear collision operators

$$Q_i(f^{(i)})(\mathbf{x}, \mathbf{v}, t) = \int_{\mathbb{R}^3} \sigma_i(\mathbf{v}, \mathbf{v}') \left[M_i(\mathbf{v}) f^{(i)}(\mathbf{x}, \mathbf{v}', t) - M_i(\mathbf{v}') f^{(i)}(\mathbf{x}, \mathbf{v}, t) \right] d\mathbf{v}',$$

where $\sigma_i(\mathbf{v}, \mathbf{v}') = \sigma_i(\mathbf{v}', \mathbf{v}) \geq 0$ is a given collision frequency model and $M_i(\mathbf{v})$ denotes the Maxwellian density of species i. The left-hand side of (1) models the advective transport of molecules of species i (local coupling of spatial variations via the spatial derivatives $\nabla_{\mathbf{x}} f^{(i)}$), while the right-hand side models the effect of collisions (global coupling of all velocities in the integral operators Q_i). The unknown functions $f^{(i)}(\mathbf{x}, \mathbf{v}, t)$ in this kinetic model represent the (scaled) probability density that a molecule of species $i = 1, \ldots, n_s$ at position $\mathbf{x} \in \Omega \subset \mathbb{R}^3$ has velocity $\mathbf{v} \in \mathbb{R}^3$ at time t. Its values need to be determined at all points $\mathbf{x}$ in the three-dimensional spatial domain Ω and for all three-dimensional velocity vectors $\mathbf{v}$ at all times $0 < t \leq t_{\mathrm{fin}}$. This high dimensionality of the space of independent variables is responsible for the numerical complexity of kinetic models, as six dimensions need to be discretized, at every time step for transient simulations. Notice that while the equations in (1) appear decoupled, they actually remain coupled through the boundary condition at the wafer surface that models the surface reactions and is of crucial importance for the applications under consideration.

2 The Numerical Method

The numerical method for (1) needs to discretize the spatial domain $\Omega \subset \mathbb{R}^3$ and the (unbounded) velocity space $\mathbb{R}^3$. We start by approximating each $f^{(i)}(\mathbf{x}, \mathbf{v}, t)$ by an expansion $f_K^{(i)}(\mathbf{x}, \mathbf{v}, t) := \sum_{\ell=0}^{K-1} f_\ell^{(i)}(\mathbf{x}, t)\varphi_\ell(\mathbf{v})$. Here, the basis functions $\varphi_\ell(\mathbf{v})$ in velocity space are chosen such that they form an orthogonal set of basis functions in velocity space with respect to an inner product that arises naturally from entropy considerations for the linear Boltzmann equation [6]. Testing (1) successively against $\varphi_k(\mathbf{v})$ with respect to this inner product approximates (1) by a system of linear hyperbolic equations

$$\frac{\partial F^{(i)}}{\partial t} + A^{(1)}\frac{\partial F^{(i)}}{\partial \mathbf{x}_1} + A^{(2)}\frac{\partial F^{(i)}}{\partial \mathbf{x}_2} + A^{(3)}\frac{\partial F^{(i)}}{\partial \mathbf{x}_3} = \frac{1}{\mathrm{Kn}}\, B^{(i)}\, F^{(i)}, \quad i = 1, \ldots, n_s, \tag{2}$$

where $F^{(i)}(\mathbf{x}, t) := (f_0^{(i)}(\mathbf{x}, t), ..., f_{K-1}^{(i)}(\mathbf{x}, t))^T$ is the vector of the K coefficient functions in the expansion in velocity space. Here, $A^{(1)}$, $A^{(2)}$, $A^{(3)}$, and $B^{(i)}$ are constant $K \times K$ matrices. Picking specially designed basis functions, the coefficient matrices $A^{(\delta)}$, $\delta = 1, 2, 3$, become diagonal matrices [7, 8]. Note again that the equations for all species remain coupled through the crucial reaction boundary condition at the wafer surface.

The hyperbolic system (2) is now posed in a standard form as a system of partial differential equations on the spatial domain $\Omega \subset \mathbb{R}^3$ and in time t and amenable to solution by various methods. Figure 1 shows two views of a representative domain $\Omega \subset \mathbb{R}^3$; more precisely, the plots show the solid wafer surface consisting of a 0.3 micrometer deep trench, in which is etched another 0.3 micrometer deep via (round hole). The domain Ω for our model is the gaseous region above the solid wafer surface up to the top of the plot box shown at $\mathbf{x}_3 =$ 0.3 micrometers in Figure 1. Since typical domains in our applications such as this one are of irregular shape, we use the discontinuous Galerkin method (DGM) [9], relying on a finite element discretization into tetrahedra of the domain. Explicit time-stepping is used because of its memory efficiency and cheap cost per time step. We accept at this point the large number of time steps that will be required, because we also wish to control the size of the step size to maintain a high level of accuracy.

The degrees of freedom (DOF) of the finite element method are the values of the n_s species' coefficient functions $f_k^{(i)}(\mathbf{x}, t)$ in the Galerkin expansion at K discrete velocities on the 4 vertices of each of the N_e tetrahedra of the three-dimensional mesh. Hence, the complexity of the computational problem is given by $4\, N_e\, K\, n_s$ at every time step. To appreciate the size of the problem, consider that the mesh of the domain in Figure 1 uses $N_e = 7{,}087$ three-dimensional tetrahedral elements; even in the case of a single-species model ($n_s = 1$) and if we use just $K = 4 \times 4 \times 4 = 64$ discrete velocities in three dimensions, as used for the application results in the following section, the total DOF are $N = 1{,}814{,}272$ or nearly 2 million unknowns to be determined at every time step. Then, to reach a (re-dimensionalized) final time of 30.0 nanoseconds, for instance, requires about 60,000 time steps, with time step Δt selected according to a CFL-condition. This

size of problem at every time step motivates our interest in parallel computing for this problem. For the parallel computations on a distributed-memory cluster, the spatial domain Ω is partitioned in a pre-processing step, and the disjoint subdomains are distributed to separate parallel processes. The discontinuous Galerkin method for (2) needs the flux through the element faces. At the interface from one subdomain to the next, communications are required among those pairs of parallel processes that share a subdomain boundary. Additionally, a number of global reduce operations are needed to compute inner products, norms, and other diagnostic quantities.

3 Application Results

As an application example, we present a model for chemical vapor deposition. In this process, gaseous chemicals are supplied from the gas-phase interface at $\mathbf{x}_3 = 0.3$ in Figure 1. The gaseous chemicals flow downwards throughout the domain Ω until they reach the solid wafer surface in Figure 1, where some of the molecules react to form a solid deposit. The time scale of our simulations corresponds to forming only a very thin layer; hence the surface is not moved within a simulation. We focus on how the flow behaves when starting from no gas present throughout Ω, modeled by initial condition $f^{(i)} \equiv 0$ at $t = 0$. Here, we use a single-species model with one reactive species ($n_s = 1$). The deposition at the wafer surface can then be modeled using a sticking factor $0 \leq \gamma_0 \leq 1$ that represents the fraction of molecules that are modeled to deposit at ("stick to") the wafer surface. The re-emission into Ω of gaseous molecules from the wafer surface is modeled as re-emission with velocity components in Maxwellian form and proportional to the flux to the surface as well as proportional to $1-\gamma_0$. The re-emission is scaled to conserve mass in the absence of deposition ($\gamma_0 = 0$). The studies shown use a sticking factor of $\gamma_0 = 0.01$, that is, most molecules re-emit from the surface, which is a realistic condition. The collision operator uses a relaxation time discretization by choosing $\sigma_1(\mathbf{v}, \mathbf{v}') \equiv 1/\tau_1$ with (dimensionless) relaxation time $\tau_1 = 1.0$.

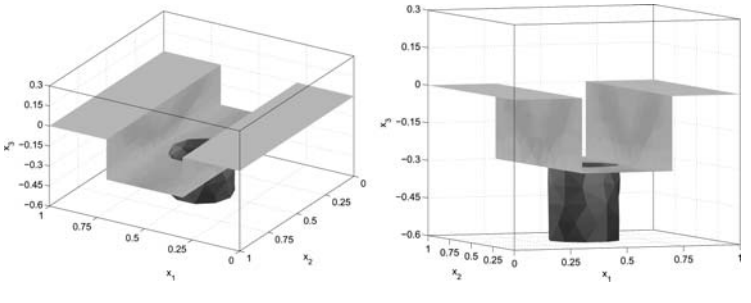

Fig. 1. Two views of the solid wafer surface boundary of the trench/via domain

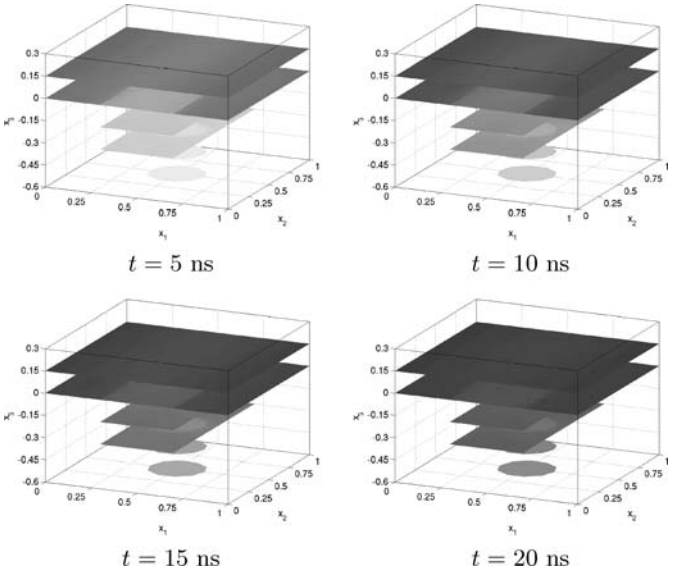

Fig. 2. Slice plots at heights $\mathbf{x}_3 = -0.60, -0.45, -0.30, -0.15, 0.00, 0.15$ at different times t of the dimensionless concentration $c(\mathbf{x}, t)$ for $\mathrm{Kn} = 0.1$ throughout domain Ω. Grayscale from light $\Leftrightarrow$ $c = 0$ to dark $\Leftrightarrow$ $c = 1$

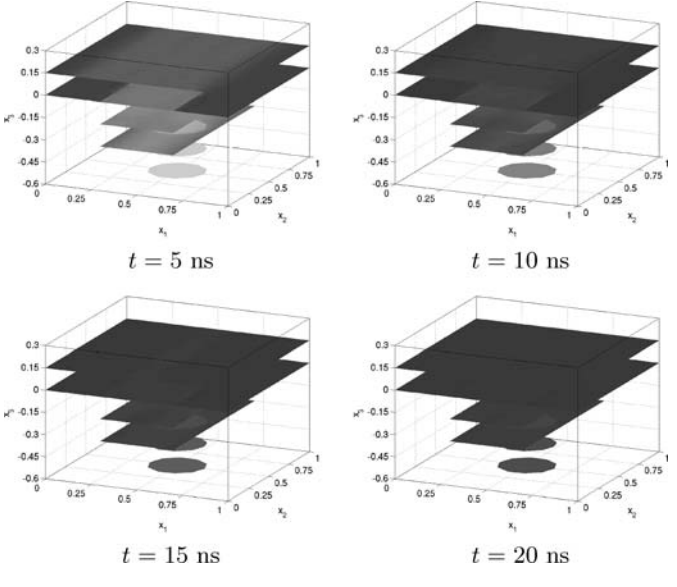

Fig. 3. Slice plots at heights $\mathbf{x}_3 = -0.60, -0.45, -0.30, -0.15, 0.00, 0.15$ at different times t of the dimensionless concentration $c(\mathbf{x}, t)$ for $\mathrm{Kn} = 1.0$ throughout domain Ω. Grayscale from light $\Leftrightarrow$ $c = 0$ to dark $\Leftrightarrow$ $c = 1$

Figures 2 and 3 show the results of transient simulations for the values of the Knudsen number $\mathrm{Kn} = 0.1$ and $\mathrm{Kn} = 1.0$, respectively. The quantity plotted for each (re-dimensionalized) time is the (dimensionless) concentration

$$c(\mathbf{x}, t) := \int_{\mathbb{R}^3} f^{(1)}(\mathbf{x}, \mathbf{v}, t)\, d\mathbf{v}$$

across the domain Ω. The values of the dimensionless concentration $0 \leq c \leq 1$ is represented by the gray-scale color on each of the horizontal slices through Ω at the vertical levels at six values of $\mathbf{x}_3$; the shapes of all slices together indicate the shape of the domain Ω.

At $t = 5$ ns, the top-most slice at $\mathbf{x}_3 = 0.15$ is mostly dark-colored, indicating that a relatively high concentration of molecules have reached this level from the inflow at the top of the domain. The slice at $\mathbf{x}_3 = 0$ shows that the concentration at the flat parts of the wafer surface has reached relatively high values, as well, while the lighter color above the mouth of the trench ($0.3 \leq \mathbf{x}_1 \leq 0.7$) is explained by the ongoing flow of molecules into the trench. At the slice for $\mathbf{x}_3 = -0.3$, we observe the same phenomenon where the concentration has reached a higher value in the flat areas of the trench bottom as compared to the opening into the via (round hole) below. Finally, not many molecules have reached the via bottom, yet, indicated by the very light color there.

Comparing the plots at $t = 5$ ns in Figures 2 and 3 with each other, the one for the smaller $\mathrm{Kn} = 0.1$ has generally lighter color indicating a slower fill of the feature with gas. The smaller Knudsen number means more collisions among molecules, leading to a less directional flow than for the larger Knudsen number $\mathrm{Kn} = 1.0$. Since the bulk direction of the flow is downward because of the supply at the top with downward velocity, the feature fills faster with molecules in this case.

The following plots in both figures show how the fill of the entire domain with gaseous molecules continues over time. Figure 3 shows that steady-state of complete fill is reached slightly faster for the larger Knudsen number, which is realistic. Consistent results are obtained by considering a wider range of values of Kn than presented here.

The results validate the effectiveness of the model and its numerical method for multiscale simulations for this application.

4 Parallel Performance Results

In this section, parallel performance results are presented for the code running on a 64-processor Beowulf cluster. This system has 32 dual-processor compute

Table 1. Observed wall clock times in minutes up to 64 processes

K	DOF	1	2	4	8	16	32	64
8	226,784	154	93	48	40	21	8	6
64	1,814,272	2,294	1,252	686	323	160	91	50

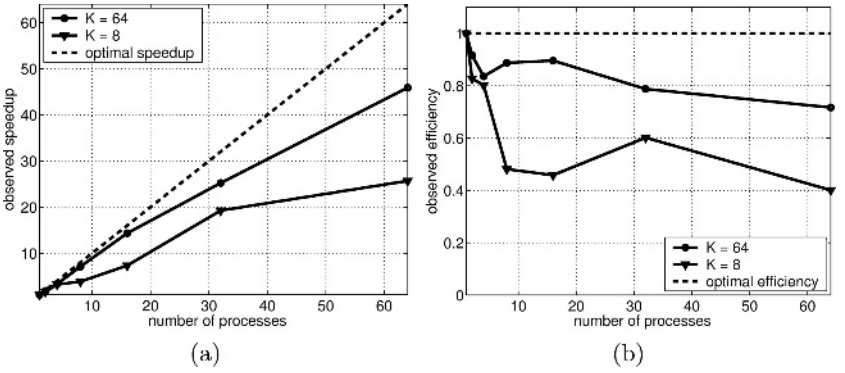

Fig. 4. (a) Observed speedup and (b) observed efficiency up to 64 processes

nodes, each with two Intel Xeon 2.0 GHz (512 kB L2 Cache) chips and 1 GB of memory. The nodes communicate through a high-performance Myrinet interconnect using the Message Passing Interface (MPI).

Table 1 shows observed wall clock times for simulations with Kn = 1.0 up to a final time of 30.0 nanoseconds for the domain with 7,087 elements in Figure 1 for velocity discretizations using $K = 2 \times 2 \times 2 = 8$ and $K = 4 \times 4 \times 4 = 64$ discrete velocities, respectively. The second column in the table lists the number of degrees of freedom (DOF) for reference. The following columns show the observed wall-clock times in minutes for each number of parallel processes 1, 2, 4, ..., 64 used; the runs using 1 process use a serial code of the same algorithm. The wall clock times in Table 1 were obtained by computing the difference of the time stamps on the first and last output file written by the code. Thus, these numbers are the most *pessimistic* measure of parallel performance possible, as various extraneous delays like operating system activity, file serving, I/O delay, etc. are included in this measurement in addition to the actual cost of calculations and communications of the numerical algorithm. Clearly, the computation times for three-dimensional models are quite significant, even using the modest resolution of velocity space presented here.

Figure 4 (a) and (b) show the observed speedup and efficiency, respectively, computed from the wall clock times given in Table 1. For $K = 8$, the speedup in Figure 4 (a) appears good up to 4 processes, then deteriorates, but continues to improve again for larger number of processes. This behavior is not unexpected, as 226,784 degrees of freedom are not a particularly large number. The speedup for the more complex case $K = 64$ is clearly significantly better. It is near-optimal up to 16 processes, then drops off slightly. Figure 4 (b) shows the observed efficiency. Both lines reveal a noticeable drop of efficiency from the serial to the 2-process run that was not easily visible in the speedup plot. We explain this drop by overhead of the parallel version of the code as compared to the serial code used as 1-process code. The efficiency shows some additional slight decline up to 4 processes. This poorer efficiency might be due to the particular partitioning of the domain Ω among the processes which is computed independently for

each number of processes, thus can result in a particularly inefficient partition in some cases. But it is remarkable that the efficiency does not continue to decrease significantly if more than 4 processes are used. In particular, the more complex case of $K = 64$ maintains its efficiency above 70% nearly all the way to 64 processes. This is a very good result for a distributed-memory cluster and justifies the use of relatively large numbers of parallel processes in more complex cases, e.g., for $K = 8 \times 8 \times 8 = 512$ resulting in 14,514,176 or more than 14.5 million degrees of freedom at every time step.

Acknowledgments

The hardware used in the computational studies was partially supported by the SCREMS grant DMS–0215373 from the U.S. National Science Foundation with additional support from the University of Maryland, Baltimore County. See `www.math.umbc.edu/~gobbert/kali` for more information on the machine and the projects using it. Prof. Gobbert also wishes to thank the Institute for Mathematics and its Applications (IMA) at the University of Minnesota for its hospitality during Fall 2004. The IMA is supported by funds provided by the U.S. National Science Foundation. Prof. Cale acknowledges support from MARCO, DARPA, and NYSTAR through the Interconnect Focus Center. We also wish to thank Max O. Bloomfield for supplying the original mesh of the domain.

References

1. Gobbert, M.K., Merchant, T.P., Borucki, L.J., Cale, T.S.: A multiscale simulator for low pressure chemical vapor deposition. J. Electrochem. Soc. **144** (1997) 3945–3951
2. Merchant, T.P., Gobbert, M.K., Cale, T.S., Borucki, L.J.: Multiple scale integrated modeling of deposition processes. Thin Solid Films **365** (2000) 368–375
3. Kersch, A., Morokoff, W.J.: Transport Simulation in Microelectronics. Volume 3 of Progress in Numerical Simulation for Microelectronics. Birkhäuser Verlag, Basel (1995)
4. Gobbert, M.K., Cale, T.S.: A feature scale transport and reaction model for atomic layer deposition. In Swihart, M.T., Allendorf, M.D., Meyyappan, M., eds.: Fundamental Gas-Phase and Surface Chemistry of Vapor-Phase Deposition II. Volume 2001–13., The Electrochemical Society Proceedings Series (2001) 316–323
5. Gobbert, M.K., Webster, S.G., Cale, T.S.: Transient adsorption and desorption in micrometer scale features. J. Electrochem. Soc. **149** (2002) G461–G473
6. Ringhofer, C., Schmeiser, C., Zwirchmayr, A.: Moment methods for the semiconductor Boltzmann equation on bounded position domains. SIAM J. Numer. Anal. **39** (2001) 1078–1095
7. Webster, S.G.: Stability and convergence of a spectral Galerkin method for the linear Boltzmann equation. Ph.D. thesis, University of Maryland, Baltimore County (2004)
8. Gobbert, M.K., Webster, S.G., Cale, T.S.: A galerkin method for the simulation of the transient 2-D/2-D and 3-D/3-D linear Boltzmann equation. Submitted (2005)
9. Remacle, J.F., Flaherty, J.E., Shephard, M.S.: An adaptive discontinuous Galerkin technique with an orthogonal basis applied to compressible flow problems. SIAM Rev. **45** (2003) 53–72