Parallel Solutions of Mega-Scale 3D Device Simulations

Ke-Chih Wu, Robert W. Dutton

Integrated Circuits Laboratory
AEL 204, Stanford University, Stanford, CA94305

ABSTRACT

Using 3D device solver STRIDE, large-scale 3D simulations of semiconductor devices have been successfully demonstrated on the massively-parallel computers. Domain decomposition based concurrent computation allows high parallel efficiency for both matrix assembly and matrix solution. Nonlinear scheme adaption and trusted-region damping scheme allows robust convergence performance for the highly nonlinear semiconductor equations. Examples presented illustrates some of the potential practical applications afforded through large-scale 3D simulation capability.

Parallel Solutions of Mega-Scale 3D Device Simulations

Ke-Chih Wu, Robert W. Dutton

Integrated Circuits Laboratory
AEL 204, Stanford University, Stanford, CA94305

The high performance computing power provided by the massively parallel supercomputers holds great promise of removing important barriers, such as long CPU hours, insufficient grid resolution, for the wide spread use of 3D device simulators in the development of VLSI technologies. A prototype 3D device solver, STRIDE (Stanford ThRee dImensional DEvice simulator)[1] has been successfully ported onto the delta machine, a massively parallel supercomputer with 512 i860 CPU nodes and more than 6G bytes of total memory. Running on all 512 CPU nodes, STRIDE sustained 1.7 GFlops in sparse matrix solution corresponding to about 65 percent in parallel efficiency. The CPU time per bias point when calculating the I-V curves of a bipolar transistor consisting of about 5 million grid points, i.e. 15 million variables, is about 30 minutes.

STRIDE solves drift-diffusion model on multiple platforms including multi-processor parallel computers[2]. It applies the finite volume discretization scheme to the non-uniform regular grids representing the device domains. Non-planar structures are supported by allowing certain patterns of different materials inside the brick elements. Polycrystalline material is supported by allowing different mobility and lifetime models.

The domain decomposition principle underlines the schemes for concurrent computation[3]. The entire simulation domain is divided into roughly equal size subdomains and assigned to the processors. The concurrent matrix assembly requires virtually no communication, except a global sum for calculating the infinite norm of the residual vector as each CPU node deal with its own subdomain. To facilitate the concurrent vector and matrix operations, the ordering of the grid points 'seen' by a processor is based on a classification according to how they are shared and with what neighboring processors(s). The operation needed by the iterative matrix algorithms, namely matrix vector multiplication and vector dot product, can therefore be easily constructed. A good preconditioner is the key to success of the iterative algorithms and a challenging issue of parallel computation has been the efficient concurrent implementation of the advanced preconditioners. By dividing the preconditioning process into stages involving grid points with increasing number of shared neighboring processors, concurrent preconditioners such as ILU(0), ILU(1) and ILUV In our experience, while ILU(0) has been adequate for the most problems encountered, ILUV has proven to be very effective in solving otherwise non-convergent problems. In short, based on domain decomposition, high parallel efficiencies are achieved in both the formation and the solution of the matrices.

Robust nonlinear convergence performance is still a challenge for solving the highly nonlinear semiconductor equations. Three approaches are used in STRIDE to attack this issue[1,2]. First, an initial guess scheme is used to spread the voltage steps at the electrodes into the interior of the device, thus avoiding big voltage steps in charge neutral region. Second, a nonlinear algorithm adaptation scheme has been developed to dynamically choose the 'optimal' nonlinear algorithm for the situation. With its widest convergence radius and smallest CPU time per iteration, Gummel iteration is always chosen to start the nonlinear iteration. If the convergence rate is unsatisfactory, a switch into more fragile but faster converging algorithms such as Newton-1C will be made when the solution has settled down sufficiently. The mechanism for switching back into more stable algorithms is provided to deal with the sudden change in the internal dynamics of the device such as the onset of latch-up process. Third, a trusted region approach based algorithm is used to damp the update vector if necessary. The combined impact of these approaches can be seen from the result of a full delta run of a bipolar transistor with an initial bias of 0.9 volt on the base and 5 volts on the collector. For this very difficult nonlinear problem, The solution process converged in thirteen iterations with the last eight being that of Newton-1C.

Figure 1 shows the number of grid points and CPU time per bias point as a function of CPU nodes in the simulation of a non-walled bipolar transistor. The number of grid points used scales linearly with the number CPU nodes as more processors also bring in proportionally more memory. It is well known that the number of linear iterations grows with the one-third power of the number of solution variables in 3D simulation. The fact that CPU time per bias point also grows with the one-third power rule indicates that the parallel algorithms have nearly perfect scalability as the number of processors increase. With the mega-scale simulation capability, various previously impossible issues can now be addressed adequately. For example, by increasing the grid density of the bipolar transistor, it was found that the collector current simulated at a lower grid density was quite adequate while the base current did not approach the asymptotic values until about one million grid points.

Latch-up analysis is an area where 3-D simulation is essential to obtain realistic results. Figure 2 shows the layout of a CMOS cell with the susceptible area for latchup when the well contacts (not shown) is taken into account. Using STRIDE, the latchup trigger current was analyzed in this area with the well contacts included. By butting the n-well contact with V_{DD}, trigger current is increased by more than a hundred percent. With the mega-scale simulation capability afforded by high performance computing, it is now feasible to analyze the latchup with the actual circuit layout and provide more relevant information to the circuit designers.

In summary, high performance computing capability provided by the massively parallel computer is making the mega-scale 3D device simulations feasible. Results obtained from delta runs of STRIDE demonstrates parallel algorithms based on domain decomposition is very promising in delivering scalable performances on the massively parallel computers.

References

1. K.-C. Wu, etc, *IEEE Trans. CAD of Integrated Circuits and Systems*, p. 1132, 1991.
2. K.-C. Wu, etc, *IEEE Trans. CAD of Integrated Circuits and Systems*, p. 528, 1989.
3. R. F. Lucas, etc, *ICCAD Proceedings*, p. 442, 1987.

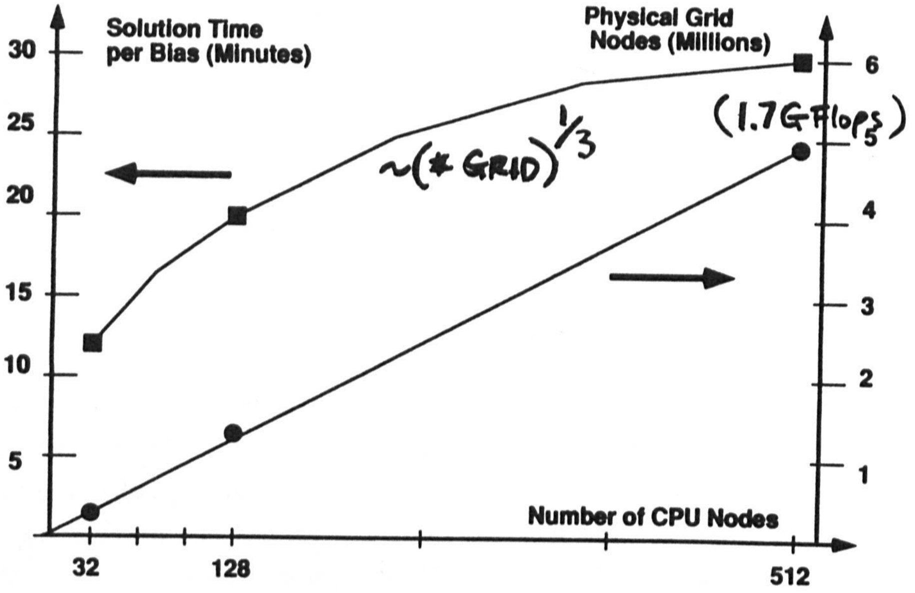
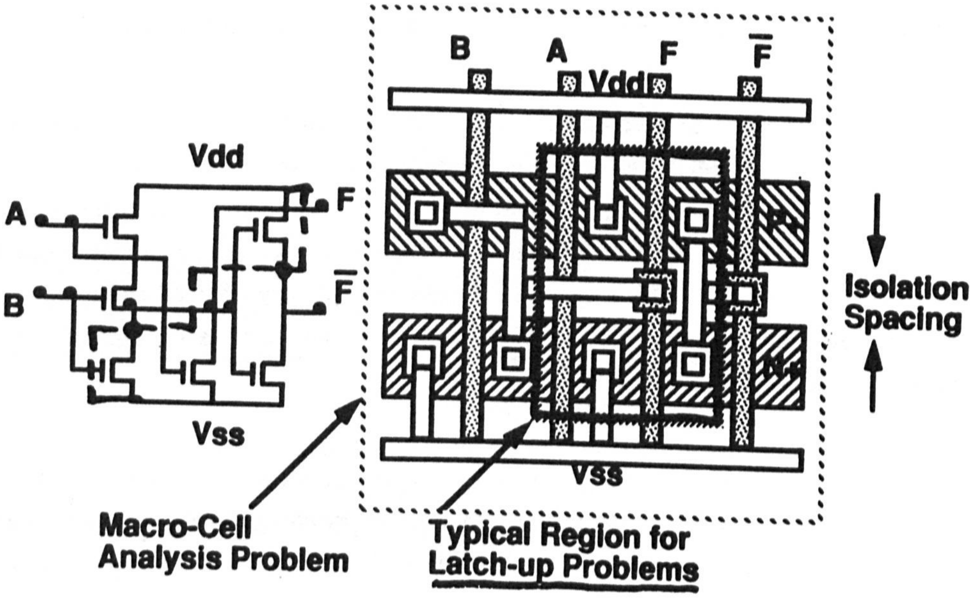


Figure 1 3D Bipolar Simulation on DELTA



Vdd
A
B
F
F̄
Vss
B
A
F
F̄
Vdd
P+
N+
Vss
Isolation Spacing
Macro-Cell Analysis Problem
Typical Region for Latch-up Problems

Figure 2 Latchup Problem Area in A CMOS Cell