Fast and Accurate Timing Simulation with Regionwise Quadratic Models of MOS I-V Characteristics*

A. Dharchoudhury and S. M. Kang
Dept. of Electrical & Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801.

K. H. Kim and S. H. Lee
CAE Department
Samsung Electronics Co.
Seoul, S. Korea.

Abstract

This paper presents a technique called regionwise quadratic (RWQ) modeling that allows highly accurate MOS models, as well as measured I-V data, to be used in fast timing simulation. This technique significantly increases the accuracy of fast timing simulation while maintaining efficiency by permitting analytical solutions of node equations. A fast timing simulator using these RWQ models has been implemented. Several examples of RWQ modeling are provided, and comparisons of simulation results with SPICE3 are shown to demonstrate accuracy and efficiency. Speedups of two to three orders of magnitude for circuits containing up to 2000 transistors are observed.

1 Introduction

The simple but inaccurate Shichman-Hodges MOSFET model is almost universally used in equation-solving fast timing simulators such as IDSIM2 [1] and ILLIADS [2]. The reason for this is that these simulators employ analytical solutions for the node differential equations to achieve high simulation speeds, and more accurate MOSFET models cannot be used since the resultant differential equations are too complex and do not permit analytical solutions. Even though fast timing simulation has been shown to be a viable alternative to circuit simulations for large digital circuits, its application has been limited by its dependence on the Shichman-Hodges model and the deficiencies of that model. In this paper, we present a technique called regionwise quadratic (RWQ) modeling that enables us to use highly accurate analytical and empirical MOS transistor current models, as well as measured I-V data, for fast timing simulation of MOS digital circuits. Since these models are derived from accurate MOS models or measured data, effects ignored in the Shichman-Hodges model are accounted for and simulation results are much more accurate. Further, the form of the differential equations is preserved, thereby allowing analytical solutions to be used and maintaining efficiency. Moreover, since this technique is model-independent, it can be used even when only measured data is available and MOS models have not been fully developed or characterized. Thus, this technique enhances the scope, validity and accuracy of fast timing simulation.

*This research was supported in part by the Joint Services Electronics Program under contract N00014-93-J-1270 and by Samsung Electronics Co.

Permission to copy without fee all or part of this material is granted, provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.

© 1994 ACM 0-89791-690-5/94/0011/0190 $3.50

2 RWQ Modeling

Let us define the effective gate-source voltage of an MOS transistor as $V_{gse} = V_{gs} - V_t$, where V_t is the threshold voltage. The RWQ modeling procedure takes as input a set of data points (V_{ds}, V_{gse}, I_{ds}), which have been obtained either by measurement or by exercising a particular analytical or empirical MOS model. The central idea of RWQ modeling is to optimally partition the (V_{ds}, V_{gse}) plane into a number of regions, and fit a quadratic model of I_{ds} in terms of V_{ds} and V_{gse} in each region. Because of the quadratic nature of the RWQ models, the resultant node differential equations permit analytical solutions and the efficiency of fast timing simulation can be maintained. Further, all major small-geometry effects that are reflected in the data are also captured implicitly in the RWQ models.

The geometry of the regionwise partition of the (V_{ds}, V_{gse}) plane is shown in Fig. 1. Let n_r denote the number of regions and $V_{ds}^{\max}$ denote the maximum values of V_{ds} in the data set. Each boundary line separating adjacent regions has the form $V_{ds} = p + qV_{gse}$, with the left boundary of the first region and the right boundary of the last region being fixed and given by $V_{ds} = 0$ and $V_{ds} = V_{ds}^{\max}$, respectively. We assume that the slopes of all other boundaries are equal to q, and the left boundary of the second region has intercept $p = 0$. Thus, the partition is characterized by one slope and $(n_r - 2)$ intercepts on the V_{ds} axis. Alternatively, a partition is equivalent to a *configuration vector* composed of these $(n_r - 1)$ *configuration parameters*. It has been our experience that increasing the number of configuration parameters (e.g. by allowing boundaries to have different slopes) does not significantly increase the accuracy of the modeling. For a given regionwise partition, the following quadratic model of I_{ds} in terms of V_{ds} and V_{gse} is fitted to the data in the kth region, $k = 1, 2, \ldots, n_r$ (β is the device transconductance):

$$\frac{I_{ds}}{\beta} = \alpha_0^{(k)} + \alpha_1^{(k)} V_{gse} + \alpha_2^{(k)} V_{ds} + \alpha_3^{(k)} V_{gse}^2 + \alpha_4^{(k)} V_{gse} V_{ds} + \alpha_5^{(k)} V_{ds}^2 \qquad (1)$$

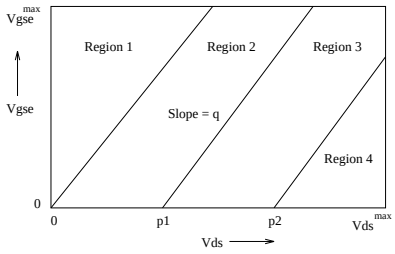


Fig. 1: Regionwise Partition of the (V_{ds}, V_{gse}) plane

2.1 Cost of a Partition

The RWQ modeling procedure seeks to obtain the partition that is optimal for the given data set. This is achieved by associating a *cost* with a partition, and determining the partition that minimizes that cost. For a particular partition, the cost has two components: a fitting cost and a smoothness cost.

Fitting Cost

The coefficients of the quadratic model in (1) are obtained by fitting the model to the data in each region subject to two restrictions: (i) the I_{ds} values should be *continuous* across the region boundaries and (ii) I_{ds} should *monotonically increase* with respect to both V_{ds} and V_{gs} at all points of the (V_{ds}, V_{gse}) plane. The first condition is satisfied by enforcing the continuity of I_{ds} at the left boundary of each region as it is being fitted. Suppose that the kth region is currently being fitted and its left boundary is given by $V_{ds} = p + qV_{gse}$. Then it can be shown that the continuity conditions are equivalent to three equality constraints:

$$\begin{aligned} \alpha_0^{(k)} + \alpha_2^{(k)} p + \alpha_5^{(k)} p^2 &= m_1 \\ \alpha_1^{(k)} + \alpha_2^{(k)} q + \alpha_4^{(k)} p + 2\alpha_5^{(k)} pq &= m_2 \\ \alpha_3^{(k)} + \alpha_4^{(k)} q + \alpha_5^{(k)} q^2 &= m_3, \end{aligned} \qquad (2)$$

where m_1, m_2 and m_3 depend on the coefficients of the $(k-1)$th region and are therefore known. For the first region, we have $I_{ds} \equiv 0$ at the left boundary and $m_1 = m_2 = m_3 = 0$. It can also be shown that to ensure monotonicity, it is sufficient to apply the non-negativity condition on $\partial I_{ds}/\partial V_{ds}$ and $\partial I_{ds}/\partial V_{gse}$ at the n_c distinct "corners" of a region ($3 \le n_c \le 5$ from Fig. 1). Thus, monotonicity imposes $2n_c$ inequality constraints. Therefore, if there are N data points in the kth region, the coefficients $\alpha_i^{(k)}$ are obtained by solving the following quadratic programming problem:

$$\min_{\alpha_i^{(k)}} \quad \epsilon_k^2 = \sum_{j=1}^{N} w_j (I_{ds,j}^{data} - I_{ds,j}^{model})^2$$
$$\text{s.t. continuity and monotonicity constraints.} \qquad (3)$$

The fitting cost C_f of a particular partition is defined as $C_f = \sum_{k=1}^{n_r} \epsilon_k$.

Smoothness Cost

The smoothness cost penalizes the discontinuity in the first partial derivatives of I_{ds} across the boundaries of adjacent regions. Consider the boundary $V_{ds} = p + qV_{gse}$ separating the k and $(k-1)$th regions. Two quantities P and Q are defined as

$$\begin{aligned} P &= \left(\frac{\partial I_{ds}}{\partial V_{ds}}\right)^{(k)} - \left(\frac{\partial I_{ds}}{\partial V_{ds}}\right)^{(k-1)}\Bigg|_{V_{ds}=p+qV_{gse}} \\ Q &= \left(\frac{\partial I_{ds}}{\partial V_{gse}}\right)^{(k)} - \left(\frac{\partial I_{ds}}{\partial V_{gse}}\right)^{(k-1)}\Bigg|_{V_{ds}=p+qV_{gse}} \end{aligned} \qquad (4)$$

Let $\sigma_k = \max\{|P|, |Q|\}$ define the smoothness cost for the kth region. Then, the smoothness cost C_s of a particular partition is $C_s = \sum_{k=2}^{n_r} \sigma_k$.

Optimal Partitioning

The *total cost* C_t for a particular partition is defined as $C_t = \lambda_1 C_f + \lambda_2 C_s$, where λ_1 and λ_2 represent the relative importance or weights of the two costs and can be set by the user. To determine the optimal partition, we minimize the total cost C_t by varying the $(n_r - 1)$ configuration parameters according to one of three optimization strategies: *random search*, *guided random search*, and *simulated annealing*. Two remarks are in order here: (i) the inclusion of C_s makes the total cost C_t non-differentiable (even if C_s is not included, gradients are difficult to obtain), and (ii) the cost of optimal partitioning is not critical since the models, once obtained, are stored and repeatedly used during fast timing simulation.

3 RWQ Modeling Examples

The RWQ modeling technique outlined above is applied to data generated using the SPICE MOS2 and MOS3 models for an NMOS and a PMOS device using the device dimensions and model parameters shown in Table I. The corresponding RWQ models are referred to as RWQ2 and RWQ3, respectively. For both RWQ2 and RWQ3, sufficient accuracy is obtained with $n_r = 3$ (which means that there are two region boundaries). The quality of the final fits are shown are shown in Fig. 2(a)–(d). For lack of space, the coefficients in each region are not shown here, but the final values of the configuration parameters are given: (i) RWQ2 NMOS: [1.08, 1.48], (i) RWQ2 PMOS: [0.98, 1.48], (iii) RWQ3 NMOS: [0.38, 0.90], and (iv) RWQ3 PMOS: [0.43, 0.66]. Note that the first value in each of the above configuration vectors is the slope of the boundaries and the second value is the intercept of the second boundary on the V_{ds}–axis. Further note that the RWQ models implicitly contain the channel-length modulation effect that is seen in the data (LAMBDA for MOS2 and KAPPA for MOS3 are nonzero in Table I). In the next example, the RWQ modeling procedure is applied to NMOS and PMOS test data obtained from a 256 Mbit DRAM chip. The parameters for the test device are W=10μm, L=0.5μm,

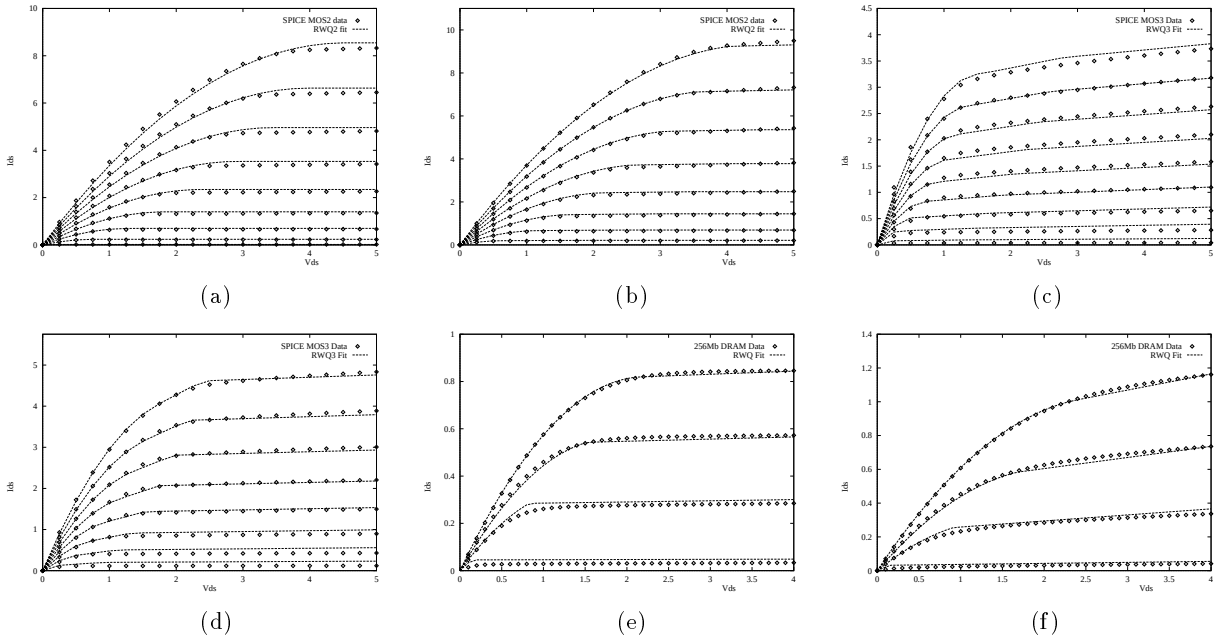


Fig. 2: (a) NMOS level 2, (b) PMOS level 2, (c) NMOS level 3, (d) PMOS level 3, (e) NMOS 256 Mbit DRAM, (f) PMOS 256 Mbit DRAM

Table I: SPICE model parameters and device dimensions

Parameter Name	Level 2		Level 3	
	NMOS	PMOS	NMOS	PMOS
L (μm)	0.8	0.8	0.8	0.8
W (μm)	10.0	20.0	10.0	20.0
TOX (A^0)	250	250	203	203
UO (cm^2/V.s)	600	300	762	254
VTO (V)	0.8	-0.9	0.73	-0.97
NSUB ($1/cm^3$)	9.63e14	1.05e16	1.75e16	2.15e16
LAMBDA (1/V)	0.015	0.026	–	–
VMAX (m/s)	–	–	1.49e5	1.82e5
KAPPA	–	–	9.51e-2	3.22e-2

t_{ox}=80A^0, and N_{sub}=$10^{17}/cm^3$. In this example, two regions are sufficient to fit both data sets; the final fits are shown in Fig. 2(e) and (f). The values of the configuration parameter (slope) for the NMOS and PMOS RWQ models are 0.69 and 0.76, respectively.

4 Fast Timing Simulation with RWQ models

The regionwise quadratic MOS I-V models have been incorporated into a fast timing simulator called ILLIADS2. In ILLIADS2, each simulated node is mapped into the generic MOS circuit primitive [3] shown in Fig. 3. Piecewise-linear input signals are applied at the terminals D_i and G_i, the parasitic resistive and capacitive elements are assumed to be linear, and the output node capacitance is denoted by C_L. If each MOS transistor in the primitive has a RWQ model, it can be easily shown that the differential equation for the output node voltage is a Riccati differential equation of the form

$$\frac{dV}{d\tau} = KV^2 + (p_1\tau + p_0)V + (q_2\tau^2 + q_1\tau + q_0), V(0) = V_0. \quad (5)$$

This differential equation permits analytical solutions which are computationally efficient and numerically stable [3]. To simulate charge-sharing between two nodes, a primitive similar to the one above may be used. The resultant differential equation is of the same form as (5) if RWQ MOS models are used. During the course of simulation, a transistor may change its region of operation. Since the coefficients of the node differential equation depend on the region of operation of each MOS transistor in the primitive, determining the time of region crossings is important. This involves solving a nonlinear algebraic

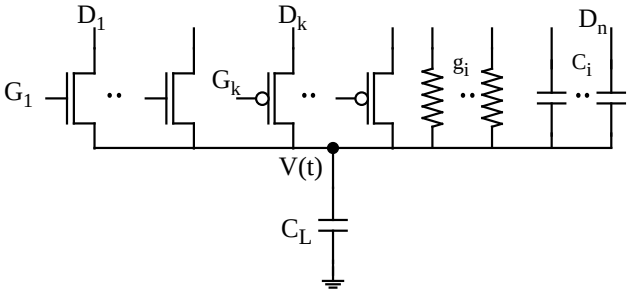


Fig. 3: Circuit Primitive

equation. This overhead is kept to a minimum in RWQ modeling by using the smallest number of regions that can accurately fit the given data. The procedure in [4] can also be viewed as a special case of RWQ modeling, but since it uses a large number of regions, many region crossings will occur even if the timestep or the voltage change is small, thereby making that method inefficient.

In order to reduce the number of simulated nodes, internal nodes with small loading capacitances are removed by merging serial and parallel transistors. This merging is accomplished by combining the device transconductances to obtain an equivalent transconductance and by combining the gate signals to obtain an equivalent gate signal. In this procedures, there is a tacit assumption that the RWQ models of the merged transistors are the same; the same assumption is also made for conventional MOS models. When capacitances of internal nodes are not negligible, or if the waveforms at these nodes are desired, internal nodes are also simulated along with the driven and pass nodes of the circuit. The simulation of internal nodes and complicated charge-sharing effects are not discussed here. In brief, we employ a waveform relaxation algorithm in which the primitive of each simulated node is solved in turn by linearizing the waveforms at the other nodes.

Body-effect in MOS transistors was not considered during RWQ modeling, but it is accounted for during simulation by updating the value of the threshold voltage using the latest values of the source-bulk voltage V_{sb}. For example, body effect in the upper NMOS transistor of a NAND2 gate can be simulated accurately since the internal node simulation capability provides us with accurate values of the internal node voltage. The accuracy of body-effect simulation is demonstrated through an example in the next section.

5 Simulation Results with RWQ Models

In this section, we demonstrate the accuracy and speed advantage of fast timing simulation with RWQ models by comparing SPICE3 simulations with MOS2 (MOS3) models against ILLIADS2 simulations with RWQ2 (RWQ3) models of Section 3. Simulation results of a CMOS inverter circuit for various input rise-time and output capacitance values are shown in Fig. 4(a) and (b). Next, a CMOS NAND2 gate is simulated using the MOS2 and RWQ2 models and the output and internal node waveforms are shown in Fig. 5(a). The waveforms which result when body-effect is not considered are also shown. Another example of the accuracy of ILLIADS2 is given in Fig. 5(b) which shows the waveforms at one node in a CMOS four-bit full adder circuit. The above examples demonstrate that by using RWQ models, ILLIADS2 can provide simulation results which are very accurate compared to electrical-level circuit simulation with sophisticated high-level MOS models. In order to demonstrate the efficiency of ILLIADS2 with respect to SPICE3, the run-times for several benchmark circuits are collected and compared. The results are shown in Table II, where N refers to the transistor count of the circuit. For the ISCAS85 combinational benchmark circuits (named `c*`), the input patterns were chosen randomly, whereas for the ISCAS89 sequential benchmark circuits (named `s*`), we use test vectors obtained from STG [5] as the input sequences. The run-time data shows that ILLIADS2 achieves two to three orders of magnitude speedup over SPICE3 for circuits with up to 2000 transistors. Moreover, the speedup factor is seen to increase with the number of transistors in the circuit. Hence, we expect that the speed advantage will be even more substantial for larger VLSI circuits.

6 Conclusions

In this paper, we have described a regionwise quadratic (RWQ) modeling technique that provides the ability to perform fast and accurate timing simulation of MOS digital circuits using highly accurate MOS models. The accuracy of the modeling technique has been demonstrated by applying it to high-level SPICE models and measured I-V characteristic data. The implementation of these models in a fast timing simulator, ILLIADS2, has also been described. ILLIADS2 has been tested on several benchmark circuits and the simulation results have been compared with SPICE3. The results show that (i) the use of RWQ models provides very accurate results, and (ii) fast timing simulation with RWQ models is two to three orders of magnitude faster than SPICE3, and the speedup increases with circuit size. Future work in this area will focus on developing better equivalent capacitance models for submicron MOS devices for use in fast timing simulation.

References

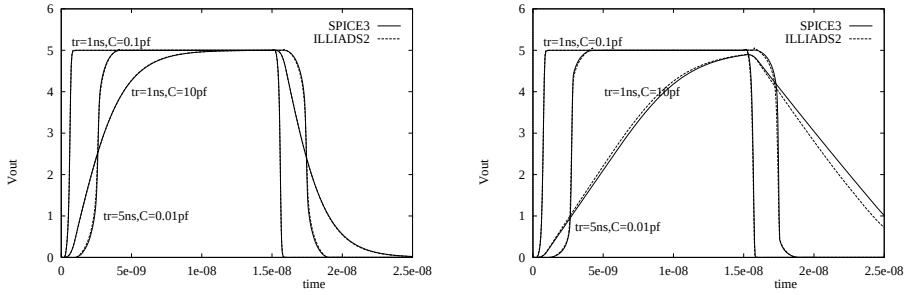
[1] D. Overhauser, *Fast Timing Simulation of MOS VLSI Circuits*. PhD thesis, University of Illinois at Urbana-Champaign, 1989.

[2] Y.-H. Shih, Y. Leblebici, and S. M. Kang, "ILLIADS: A fast timing and reliability simulator for digital MOS circuits," *IEEE Trans. Computer-Aided Design*, vol. 12(9), pp. 1387–1402, Sept. 1993.

[3] Y. H. Shih and S. M. Kang, "Analytic transient solution of general MOS circuit primitives," *IEEE Trans. Computer-Aided Design*, vol. 11(6), pp. 719–731, June 1992.

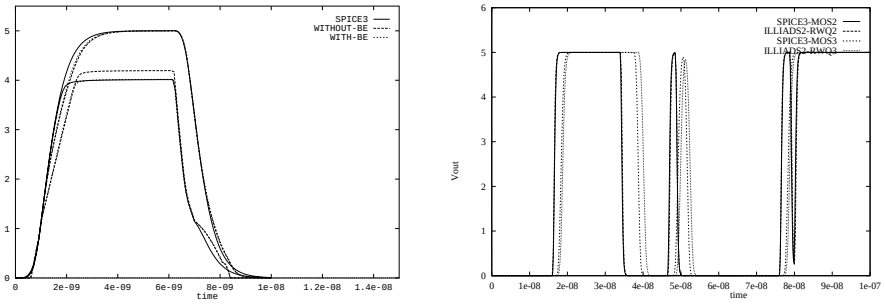
[4] Y. H. Chang and A. T. Yang, "Analytic macromodeling and simulation of strongly coupled mixed analog-digital circuits," *Proc. IEEE Int. Conf. on Computer-Aided Design*, pp. 244–247, Nov. 1992.

[5] W. T. Cheng and S. Davidson, "Sequential circuit test generator (STG) benchmark results," *Proc. IEEE Int. Symp. on Circuits and Systems*, pp. 1938–1941, May 1989.



(a) (b)

Fig. 4: CMOS inverter (a) RWQ2/MOS2 models, (b) RWQ3/MOS3 models



(a) (b)

Fig. 5: (a) Output and internal node waveforms of CMOS NAND2, (b) One node of a 4-bit CMOS full adder

Table II: Speedup Measurements for ILLIADS2

Ckt. Name	N	Level 2			Level 3		
		SPICE3	ILLIADS2	Speedup	SPICE3	ILLIADS2	SPEEDUP
c17	24	7.6	0.4	19	6.2	0.4	15.5
adder4	144	356.5	4.2	84.88	276.2	3.0	92.1
alu4	470	802.9	8.1	99.1	987.2	8.2	120.4
c432	1004	4030.9	39.0	103.4	4115.4	38.4	107.2
c499	2356	10386.4	58.5	177.5	14747.2	45.7	322.7
s27	114	70.4	4.2	16.8	298.1	3.7	80.6
s208.1	624	330.5	12.7	26.0	686.8	10.6	64.8
s641	1740	3376.4	36.7	92	23477.7*	35.0	670.8

$\star$: Large run-time due in part to DC convergence problems