A Robust and Physical BSIM3 Non-Quasi-Static Transient and AC Small-Signal Model for Circuit Simulation

Mansun Chan, Member, IEEE, Kelvin Y. Hui, Chenming Hu, Fellow, IEEE, and Ping K. Ko, Fellow, IEEE

Abstract—A new non-quasi-static (NQS) MOSFET model, which is applicable for both large-signal transient and small-signal ac analysis, has been developed. It employs a physical relaxation time approach to take care of the finite channel charging time to reach equilibrium and the effect of instantaneous channel charge redistribution. The NQS model is formulated independently from the dc I–V and the charge-capacitor model, thus can be easily applied to any existing simulators. The model has been implemented in the newly released BSIM3 version 3, and comparison has been made among this model, common quasi-static (QS) SPICE models and PISCES two-dimensional (2-D) numerical device simulator. While predicting accurate NQS behavior, the time penalty for using the new model is only about 20–30% more than the common QS models. It is much less than the time required by other NQS models reported. Limitations and compromises between simplicity, efficiency and accuracy are also discussed.

I. INTRODUCTION

As MOSFET is becoming more performance-driven, it becomes essential to predict the circuit performance when it is operating near the device cut-off frequency. However, most models available in SPICE use the quasi-static (QS) formulation, which has been shown to be inadequate [1]–[4]. In the QS approach, the channel charge is assumed to achieve equilibrium once biases are applied, thus the finite charging time of the carriers in the inversion layer is ignored. This gives erroneous simulation results for signals with rise or fall time comparable to or smaller than the channel transit time. The most common QS model with 40/60 drain/source charge partitioning [1], for example, causes an unrealistic large drain current spike during fast turn on as shown in Fig. 1. Besides affecting the accuracy of the simulation, this nonphysical result can also cause oscillation and convergent problem during subsequent numerical iterations. It is common among circuit designers to bypass the convergent problem by using a 0/100 drain/source charge partitioning ratio [5], which attributes all transient charge to the source side. However, this ad-hoc nonphysical solution merely shifts the problem to the source as shown in Fig. 2, thus only work when the source is grounded. Moreover, none of these QS models can be used to predict the high-frequency transadmittance of a MOSFET correctly as pointed out in [4]. The simulated current response when biasing a 200-μm-long MOSFET in strong inversion saturation appeared completely different from that of its equivalent circuit with two 100-μm-long MOSFETs in series as shown in Fig. 3. The equivalent circuit of breaking down a MOSFET into N equal parts in series (N-lumped model) has been used in high-frequency small-signal network due to the lack of non-quasi-static (NQS) models. The accuracy increases with N, at the expense of simulation time (of the order of N^1.4 times longer [6]). However, the method becomes impractical when the device channel length is scaled because artificial short-
channel effects in the sub-transistors may be activated. It should be noted that NQS behavior is not only confined to long channel transistors. It can be observed in short-channel devices as well, which has been experimentally demonstrated recently [7].

Some NQS models [8]–[10] have been published, but the complexity of the formulations prohibits intuitive insight into the NQS effect. These models usually require four to five times longer circuit simulation time than the common QS models, which makes them unattractive for simulating large circuits. Moreover, most of these models considered the transient and ac small-signal behavior separately, which can lead to inconsistent simulation results in the time domain and frequency domain.

In this work, we have developed a robust physical NQS model for both transient and ac small-signal analysis based on the channel charge relaxation approach. This NQS model is an attempt to compromise between accuracy, efficiency and easy adaptation to the popular approach of physical modeling and to avoid the necessity of solving complicated differential equations. A novel implementation strategy is employed to improve the efficiency as well as the flexibility for incorporating the NQS effect into other existing models. The NQS model has been implemented into the newly released BSIM3 version 3, and the results verified by PISCES. While providing significant improvement in accuracy to the existing models, the time penalty in using this new model is less than 30% for both transient and ac analysis.

II. FORMULATION

The channel of a MOSFET can be viewed as to a bias dependent RC distributed transmission line [11] as shown in Fig. 4(a). In the QS approach, the gate capacitors are lumped to the external source and drain nodes as in Fig. 4(b), thereby ignoring the finite delay time required for channel charge build-up. Breaking down the transistor into \( N \) smaller ones connected in series [Fig. 4(c)] gives a better approximation to the RC network but it has the drawbacks mentioned in previous section. A physical approach to model the NQS effect, would be to formulate an estimate for the delay time through the channel RC network and incorporate this time constant into the model equations.

One of the widely used methods to approximate the RC delay was proposed by Elmore [12], [13], by using the mean, or first moment, of the impulse response. Utilizing Elmore’s approach, the RC distributed channel can be approximated by a simple RC equivalent, which retains the lowest frequency pole of the original RC network. The new equivalent circuit is shown in Fig. 4(d). The Elmore resistance \( R_{\text{Elmore}} \) in strong inversion can be calculated from the total channel resistance, which is given by

\[
R_{\text{Elmore}} = \frac{L_{\text{eff}}}{\varepsilon f_{\text{eff}} W_{\text{eff}} Q_{\text{d1}}} \quad (1)
\]

where \( Q_{\text{d1}} \) is the instantaneous channel inversion charge and \( \varepsilon \) is the Elmore constant to match the lowest frequency pole. The value of \( \varepsilon \) is found to be around 3 by matching the output of the equivalent circuit in Fig. 4(a) and (d), and it is invariant with respect to \( W \) and \( L \). The time domain and frequency domain response of the Elmore equivalent approximation [Fig. 4(d)] and the original device with RC distributed channel [Fig. 4(a)] are compared and shown in Figs. 5 and 6, respectively. In the first case, a fast pulse is applied to the gate with both source and drain grounded and the gate current is measured. In the second case, a small-signal voltage is applied to the gate and the voltage fluctuation caused by the excitation is measured at different part of the channel. In both cases, a very good match between the RC Elmore equivalent circuit and the real distributed RC network are observed. However, in the actual BSIM3 implementation, \( \varepsilon \) is chosen to be 5 and the reasons will be given in the discussion section. For subsequent simulations, \( \varepsilon \) will assume the value of 5.

Direct implementation of the model shown in Fig. 4(d) is straightforward, but requires the creation of two additional nodes. This will increase the time to solve the Jacobian Matrix in SPICE by more than 70%. The need to change the device topology also requires modifications to the existing model equations, which is undesirable as well. A simpler way to incorporate the newly formulated NQS model will be employed to provide better robustness, flexibility and extensibility.

To simplify the model implementation, some of the current equations are reviewed. The transient node currents to the gate, drain, and source can be described by the equation

\[
I_{G,D,S}(t) = I_{G,D,S}(t)_{\text{DC}} + X_{G,D,S} \frac{dQ_{\text{d1}}(t)}{dt} \quad (2)
\]

where \( I_{G,D,S}(t) \) are the gate, drain or source current, \( I_{G,D,S}(t)_{\text{DC}} \) are the gate, drain, or source current under dc condition. \( Q_{\text{d1}}(t) \) is the actual channel charge at given time \( t \), and \( X_{G,D,S} \) is the channel charge partition ratio [14], [15] to the gate, drain and source with

\[
X_D + X_S = X_{\text{eq}} = 1 \quad (3)
\]

If the geometrical factor is taken into consideration, \( X_D \) varies from 0.5 at \( V_D = 0 \) V to 0.4 in the saturation region, and \( X_S \) varies from 0.5 to 0.6, respectively [1]. As the 0.4/0.6 partition covered larger voltage operation range and the error introduced
Fig. 4. Different possible equivalent transient and ac small-signal models for a MOS transistor.

by taking a constant $X_D = 0.4$ and $X_S = 0.6$ is less than 5% in the worst case, these values will be adopted to simplify the model.

In the QS approach, it is assumed that

$$\frac{dQ_{eq}(t)}{dt} = \frac{dQ_{eq}/dt}{dV/dt}$$

where $Q_{eq}(t)$ is the equilibrium channel charge under the biases at time $t$. The assumption of equilibrium at all time gives rise to error in calculating the NQS currents. To account for the NQS effect, a new state variable $Q_{def}$ has been introduced to keep track of the amount of deficit (or surplus) channel charge necessary to achieve equilibrium at a given time where

$$Q_{def}(t) = |Q_{eq}(t) - Q_{eq}(t)|$$

and

$$\frac{dQ_{def}(t)}{dt} = \frac{dQ_{eq}(t)}{dt} - \frac{dQ_{eq}(t)}{dt}$$

$Q_{def}$ is allowed to decay exponentially into the channel with a bias dependent NQS relaxation time $\tau$, which represents the delay due to the RC distributed network in the channel. Thus the charging current can be approximated by

$$\frac{dQ_{def}(t)}{dt} \approx \frac{Q_{def}(t)}{\tau}.$$ (7)

Substituting (7) into (6) we get

$$\frac{dQ_{def}(t)}{dt} = \frac{dQ_{eq}(t)}{dt} - \frac{Q_{def}(t)}{\tau}.$$ (8)

$Q_{def}(t)$ can be calculated from (8) with a subcircuit as shown in Fig. 7, which is a direct translation of (8) into basic circuit elements. The node voltage of the subcircuit gives the value of $Q_{def}(t)$, and the branch current flowing through the resistor with value $\tau$ gives the total charging NQS currents. This charging current is then partitioned to the gate, drain and source with the corresponding charge partition ratios. With this approach, only one addition node is required and the topology of the original transistor is preserved.
The value of the channel relaxation time constant $\tau$ is composed of the diffusion component which dominated in the subthreshold region and by the drift component (calculated from the RC Elmore equivalent discussed at the beginning of this section) dominated in the strong inversion region. Various components of $\tau$ are given by

$$\tau_{\text{diffusion}} = \frac{q(L_{\text{eff}}/A)^2}{\mu_{\text{eff}}kT}$$

$$\tau_{\text{drift}} \approx 0.5(C_{SG} + C_{DG})W_{\text{eff}}L_{\text{eff}}R_{\text{Elmore}}$$

$$= \left(\frac{C_{SG} + C_{DG}}{2}\right)\frac{L_{\text{eff}}}{\mu_{\text{eff}}(Q_{\text{dchq}} - Q_{\text{def}})}$$

$$\approx \frac{C_{SG}L_{\text{eff}}}{2\mu_{\text{eff}}Q_{\text{dchq}}}$$

and

$$\frac{1}{\tau} = \frac{1}{\tau_{\text{diffusion}}} + \frac{1}{\tau_{\text{drift}}}.$$  

In the $\tau_{\text{drift}}$ term, $C_{SG}$ and $C_{DG}$ are the source-to-gate and drain-to-gate capacitance respectively as in Fig. 4(d). Due to the complexity of calculating the derivatives of the capacitances, the total capacitance $(C_{SG} + C_{DG})$ is approximated by $C_{SG}$. Actual value of $(C_{SG} + C_{DG})$ varies between $C_{SG}$ for small $V_D$ and 0.75$C_{SG}$ for $V_D > V_{D_{\text{Drain}}}$ [16]. The error due to this approximation will be discussed in the discussion section. Also in (10), the instantaneous channel charge ($Q_{\text{chq}}$) is approximated by the equilibrium inversion charge ($Q_{\text{dchq}}$) to simplify the calculations. The validity of this approximation has been verified for rise time slower than ten times the cutoff frequency ($f_T$) of the device and is sufficient for most of the practical circuits. Fig. 8 compared the value $\tau$ used in the new model with result obtained from PISCES simulation by applying a small step input at the gate and observing the decay time of the gate current. The comparison shows that the model agrees reasonably well with the two-dimensional (2-D) simulator despite the approximations made.

III. SIMULATION RESULTS

The new model has been implemented in the BSIM3 version 3 release [17]. A number of benchmark simulations have been performed and the results verified by PISCES 2-D device simulator. Fig. 9 shows the simulated turn-on and turn-off transients in the linear region (small $V_D$). Good agreement between the new model and PISCES is observed in both cases. The small discrepancy at the beginning of the turn-on is mainly due to the inaccuracy of the dc model around the threshold voltage rather than the NQS model. The simulation result in the saturation region (high $V_D$) is shown in Fig. 10. Despite some minor errors at the on-set of the transient (will be discussed in the next section), the model describes the NQS behavior with very high accuracy.

Fig. 11 shows the result of the high-frequency transadmittance test suggested in [4], where the real part of the transadmittance is plotted against the frequency of operation. The discrepancy caused by the QS models in simulating a single transistor and its $N$-lumped equivalent has been resolved by the new NQS model. The NQS model and the $N$-lumped model both predict the transadmittance to fall off around the same frequency. It is also observed that simulation results obtained from the $N$-lumped model go asymptotically to those obtained from the NQS model as $N$ increases. Similar results are also observed in the amplitude and phase plots of a simple resistive load inverter as shown in Fig. 12.

As a practical example to illustrate the importance of the NQS effect in circuit design, a low-voltage, high-speed current output Digital/Analog converter (DAC) cell [18] as shown in Fig. 13 was simulated. In this circuit, $M_1$ and $M_{\text{llb}}$ operate as current sources when turned on, and an output current appears as $I_{\text{out}}$ or $I_{\text{llb}}$. To obtain high output resistance, $M_1$ must be greater than the minimum length (9.6 $\mu$m was used). Current switching was limited by the speed of the switching of the voltage at node 1. Fig. 14 shows a simulation using standard QS models and the new NQS model. The new model indicates slower rise and settling times, limited by NQS effect in the long channel of $M_1$. It illustrates an intrinsic limitation to the speed of this DAC circuit, which is not apparent with the QS models. Table I compares the time required to simulate the

<table>
<thead>
<tr>
<th>Time Required for Simulating the DAC Cell by Different Models</th>
<th>DC</th>
<th>0/100</th>
<th>40/60</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total time (sec)</td>
<td>1.777</td>
<td>1.863</td>
<td>1.906</td>
<td>2.243</td>
</tr>
<tr>
<td># of iteration</td>
<td>732</td>
<td>729</td>
<td>781</td>
<td>848</td>
</tr>
<tr>
<td># of time pt.</td>
<td>173</td>
<td>186</td>
<td>187</td>
<td>252</td>
</tr>
<tr>
<td>accepted time pt.</td>
<td>121</td>
<td>124</td>
<td>127</td>
<td>186</td>
</tr>
</tbody>
</table>
Fig. 9. Simulated $I-V$ characteristics in the linear region (low $V_{DS}$). Excellent match between the model and PISCES is observed.

Fig. 10. Simulated $I-V$ characteristics in the saturation region (high $V_{DS}$). Very good agreement between the model and PISCES is observed.

DAC cell using different models. The overall simulation time penalty introduced by the new NQS model is less than 30%.

IV. DISCUSSION AND LIMITATION OF THE MODEL

In this version of the NQS model, we intend to provide a practical and efficient solution, which can be easily implemented in existing simulators. A lot of compromise has been made between simulation efficiency, complexity and accuracy. A number of physical effects have been ignored, such as the body current. It has little effect on the ac simulation, but can lead to error during transient simulation. The body current can be included by partitioning the gate current from $Q_{G}$ between the gate and the body. As the body current is usually small compared to the gate current as shown in Fig. 15, and the robustness of the model takes higher priority, we chose to leave it aside. Nevertheless, the new NQS model has been shown to provide substantial improvement over existing BSIM3 charge models even without the substrate current.

Another limitation for the model is the oversight of the saturation region operation during the formulation. The derivation of $\tau_{\text{drift}}$ in (10) is based on operation in the linear region, thus no velocity saturation is considered. When a MOSFET is operated in the velocity saturation regime, the channel conductivity reduces and the value of $R_{\text{sub}}$ increases. However, this increase is partially compensated by the decrease in the term $(C_{SC} + C_{DC})$ from $C_{DE}W_{eff}L_{eff}$ to about $0.75C_{DE}W_{eff}L_{eff}$ [16] due to nonuniform channel charge...
Fig. 11. Result of high-frequency transadmittance test. The NQS model is capable of predicting the transadmittance fall-off at high frequency, which agrees with the $N$-lumped model. The result from the $N$-lumped model follows asymptotically to the result from the NQS model as $N$ increases.

Fig. 12. Frequency response of a NMOS inverter with a resistive load simulated using the new NQS model and the $N$-lumped model. The result from the $N$-lumped model follows asymptotically to the result from the NQS model as $N$ increases.

Fig. 13. A low-voltage, high-speed current output Digital/Analog Converter (DAC) cell simulated by existing QS models and the new NQS model.

distribution as $V_D$ increases from 0 V to $V_{Dsat}$. The resulting $\tau_{shift}$ increases slightly by about 30% as shown in Fig. 16. The worst case error in node currents when the device is operated in the velocity saturation region is thus less than 20%. This error can be reduced to less than 10% by choosing a compromising Elmore constant ($\varepsilon$) between the linear region and the saturation. Due to this reason, the default $\varepsilon$ in the BSIM3 model is chosen to be five instead of three calculated

Fig. 15. Relaxation time constant as a function of drain bias with the effect of velocity saturation included. Empirical fit given by (12) is superimposed for comparison.

Fig. 16. Simulated gate current and substrate current and the comparison with PISCES result.
in the linear region. When a more accurate result is desired, the drift component of the relaxation time given in (10) can be replaced by the following empirical model

\[
\tau_{\text{drift}} = \begin{cases} 
\frac{(C_{SG} + C_{DG})L_{\text{eff}}}{2\varepsilon_{\text{eff}}(Q_{\text{eq,ch}} - Q_{\text{def}})} & \text{for } V_{DS} \leq V_{\text{Dsat}} \\
\frac{(C_{SG} + C_{DG})L_{\text{eff}}}{\varepsilon_{\text{eff}}(Q_{\text{eq,ch}} - Q_{\text{def}})} & \text{for } V_{DS} > V_{\text{Dsat}}
\end{cases}
\]

Comparison between the empirical model and PISCES is shown in Fig. 15 indicating a very good agreement between them. But this model has been left out again for simplicity reasons.

The effects of high-channel electric field in the velocity saturation region deserve further consideration. At high drain voltage, high electric field is developed near the drain/channel junction. This electric field prevents the carriers to flow into the channel from the drain side during fast turn-on. In this case, all the channel charge will be provided by the source, and the drain current cannot go negative for NMOSFET (vice versa for PMOSFET) as in Fig. 10(a). Note that it is only true for velocity saturation region, and the drain current can actually go negative in the linear region as in Fig. 9. This effect can be taken care simply by restricting the drain current to be positive when the drain voltage is larger than the saturation drain voltage. That is

\[
I_D \geq 0 \quad \text{for} \quad V_D \geq V_{\text{Dsat}}.
\]

During turn-off transient, another restriction has to be imposed due to the fact that the maximum drain current is limited by the number of the carriers controlled by the drain and the maximum speed they can move. Therefore, the maximum current must satisfy

\[
I_D \leq W_{\text{eff}}(Q_{\text{ch}\text{eq}} - X_D Q_{\text{def}}).
\]  \hspace{1cm} (14)

After including the enhancements given by (12)–(14) in the model, simulation has been performed with high \( V_D \). The results of the simulation are shown in Fig. 16, which show a nearly perfect fit. The improvement after adding the enhancement is more obvious after comparing Figs. 17 and 10. However, with the restrictions added, the nice continuous properties in the BSIM3 formulation cannot be preserved and leading to slower convergence as in some older models. Due to this consideration, the restriction is not imposed in the release. Nevertheless, the small negative current, which appears in extremely fast ramping, only shows up as a small overshoot in most simulation. The implemented model is able to predict the delay time and the frequency fall off with very high accuracy in the present form.

V. CONCLUSION

Based on the relaxation time approach, a NQS MOSFET model, valid in both fast transient and ac small-signal analysis, has been developed and implemented in BSIM3 version 3. The formulation is independent of any particular \( I-V \) or \( C-V \) model, and can be implemented with any existing MOS models. Most of the parameters required by the new model can be deduced from the existing dc and charge model, and thus no new parameter extraction scheme is necessary. Extensive simulation has been performed with the new model and results verified by PISCES. The time penalty for using the new model is less than 30%, mainly due to an additional node and more time points are required for convergence during transient simulation. Limitations and enhancements to include the effects such as substrate current and velocity saturation...
are also discussed. Excellent agreement between PISCES and the new model is achieved over a wide range of biases and physical dimensions.

REFERENCES


Mansun Chan (S’92–M’95), for a photograph and biography, see p. 212 of the January 1998 issue of this TRANSACTIONS.

Ping K. Ko (S’78–M’81–SM’93–F’96), for a photograph and biography, see p. 212 of the January 1998 issue of this TRANSACTIONS.

Kelvin Y. Hui received the B.S. degree in 1983 in electrical engineering from the University of Hong Kong, and the M.E. degree in 1984 from the University of California, Los Angeles. He is currently a Ph.D. candidate at the University of California, Berkeley.

After receiving the M.E. degree, he worked for Bell Communications Research on compound semiconductor research and for Advanced Micro Devices on EEPROM. His research areas include SOI device modeling, MOS capacitance modeling, nonquasistatic transient and ac SPICE models. He was with Lattice Semiconductor Corporation for about two years after leaving Berkeley. Currently, he is with Programmable Silicon Solutions, Fremont, CA.

Chenming Hu (S’71–M’76–SM’83–F’90) received the B.S. degree from the National Taiwan University, Taipei, Taiwan, R.O.C., and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1970 and 1973, respectively.

From 1973 to 1976, he was an Assistant Professor at Massachusetts Institute of Technology, Cambridge. In 1976, he joined the University of California, Berkeley, as Professor of Electrical Engineering and Computer Sciences. He is also the Director of Industrial Liaison Program. While on industrial leave from the university in 1980 and 1981, he was Manager of nonvolatile memory development at National Semiconductor. Since 1973, he has served as a consultant to the electronics industry. He has also been an advisor to many government and educational institutions. His present research areas include VLSI devices, silicon-on-insulator devices, hot electron effects, thin dielectrics, electromigration, circuit reliability simulation, and nonvolatile semiconductor memories. He has also conducted research on electro-optics, solar cells, and power electronics. He has been awarded several patents on semiconductor devices and technology. He has authored or coauthored three books and over 400 research papers. He has delivered dozens of keynote addresses and invited papers at scientific conferences.

Dr. Hu is an Honorary Professor of Beijing University, China, of Tsinghua University, China, and of the Chinese Academy of Science. He has been a guest editor of PROCEEDINGS OF THE IEEE and of the IEEE TRANSACTIONS ON ELECTRONIC DEVICES. He serves on an editorial board of SEMICONDUCTOR SCIENCE and TECHNOLOGY, Institute of Physics, U.K. He was Associate Editor of IEEE TRANSACTIONS ON ELECTRONIC DEVICES from 1986 to 1988, and Vice Chairman of IEEE Electron Devices Society, Santa Clara Valley Chapter, from 1980 to 1982. He was the first National Science Council Invited Chair Lecturer, R.O.C., in 1987. He was Board Chairman of East San Francisco Bay Chinese School from 1988 to 1991. He received the 1991 Design News Excellence in Design Award and the 1991 Semiconductor Research Corporation Technical Excellence Award for leading the development of IC reliability simulator, BERT. He has also received five best-paper awards. He has been listed in Who is Who in America, American Men and Women of Science, Who is Who in Science and Engineering, and Who’s Who in American Education.