Transistor-Level Gate Modeling for Nano CMOS Circuit Verification Considering Statistical Process Variations

Qin Tang, Amir Zjajo, Michel Berkelaar, and Nick van der Meijs

Circuits and Systems Group, Delft University of Technology Q.tang@tudelft.nl

Abstract. Equation- or table-based gate-level models (GLMs) have been applied in static timing analysis (STA) for decades. In order to evaluate the impact of statistical process variabilities, Monte Carlo (MC) simulations are utilized with GLMs for statistical static timing analysis (SSTA), which requires a massive amount of CPU time. Driven by the challenges associated with CMOS technology scaling to 45nm and below, intensive efforts have been contributed to optimize GLMs for higher accuracy at the expense of enhanced complexity. In order to maintain both accuracy and efficiency at 45nm node and below, in this paper we present a gate model built from a simplified transistor model. Considering the increasing statistical process variabilities, the model is embedded in our new statistical simulation engine, which can do both implicit non-MC statistical as well as deterministic simulations. Results of timing, noise and power grid analysis are presented using a 45nm PTMLP technology.

Key words: gate modeling, transistor-level, non-Monte Carlo, statistical timing analysis

1 Introduction

Nowadays cell-based design flows are still dominant for circuit verification such as timing, noise or power grid analysis. Usually, due to the challenges associated with gate modeling, a unique GLM, such as a noise model or a power droop model, is developed to handle each effect. However, improved based on recent invention of a current source model [8], a unified GLM for timing, noise and power analysis is in sight. Since the analysis is carried out using cell models, the models must accurately represent the behavior of the circuit that makes up the cell for timing, crosstalk, variability calculation, etc. However, the conventional GLMs model every element in the model as a function of input slew and single output effective capacitance (C_{eff}) , and have single-input-switching (SIS) assumption.

Instead of optimizing the GLMs for higher accuracy at the cost of increased complexity and characterization time, we make a case that transistor-level gate models can address most of the limitations of GLMs [5].

With increasing process variations at 45nm and below, the major challenge in timing gate modeling becomes an efficient construction of a parameterized timing model of a design, representing the design characteristics as a function of process variations [6]. The major approaches are Monte Carlo (MC) simulations and computing and propagating statistical arrival times. The MC method suffers from excessive pessimism and poor scalability as the number of process parameters increases. On the other hand, generating statistical arrival time models for all standard cells of a library takes a huge amount of CPU time due to the necessary MC-based simulation.

In this paper, we present a statistical simplified transistor model (SSTM) for cell modeling which is capable of simultaneously handling most of the issues described in section 2. The new non-MC statistical simulation method is introduced in section 4.

2 GLM Limitations and Optimization Trends

By using conventional GLMs, (S)STA provides delay and slew much faster without calculating accurate waveform. In nanometer technology, however, the conventional GLMs become less accurate due to the following intrinsic limitations.

- 1. The simple saturated ramps can no longer represent the input signals, especially if they arise from a complicated driving stage with noise or multipleinput switching (MIS) scenario, or are influenced by process variations or other sources of variabilities [7].
- 2. GLMs fail to work with a multi-port coupled interconnect load since the load is only modeled as an effective capacitance (C_{eff}) . Oversimplification of the interconnect coupling can lead to large errors during timing analysis [1]-[2].
- 3. GLMs are unable to capture MIS and internal charge effects for high-stack and complex cells. The SIS assumption is inherent in all timing tools. In reality, all multiple input cells are subject to delay degradation (or delay improvement for the min-delay STA) due to MIS. Not modeling MIS for timing can result in as much as 100% error in delay and slew calculation [2].
- 4. The increasing modeling complexity required to handle voltage droop effects. In order to account for power supply variations, GLMs are required to be characterized at different supply voltages.

There is a clear trend to optimize GLMs to deal with the limitations listed above. Croix and Wong introduced an input-waveform-independent current source model (CSM) [8] which is essentially a voltage-based, DC-transfer-derived current source with transient effects modeled by a linear capacitance at the output. Many optimized CSMs extend the Croix model to handle other limitations. The Miller capacitance is considered and voltage-based capacitance models are used in [1]-[3] while [9] focuses on waveform models. A non-linear C_{eff} model is described in [4] although its accuracy still needs to be evaluated further. The MIS issue is addressed by modeling every input and output port in the cell [1]-[2]. The internal nodes are also modeled to capture internal charge effects in [1] to obtain higher accuracy. However, they just attempt to optimize GLMs to maintain acceptable accuracy for all types of gates. Unfortunately the fact that GLMs are black-box models where the internal structure of the gates is hidden is the essential root of all these issues. The increasing requirement for accuracy makes the trade-off between better accuracy and shorter runtime a real challenge [6].

At 45nm and below, the propagation of complex signals and accurate modeling for crosstalk effects require accurate cell models. A good cell model for SSTA should be independent of input waveform, output load and circuit structures; should not increase complexity and provide high accuracy and efficiency at the same time compared to SPICE; should have much shorter characterization time, and should be able to capture process variations and be easy to embed in a SPICE-like engine to propagate statistical signal information. By using an efficient transistor model and simulation algorithm, transistor-level gate modeling for timing analysis is a gaining popularity [10]-[12].

3 Statistical Simplified Transistor Model (SSTM)

One extreme way of transistor-level timing analysis is to simply run Spice/Spectre. However, such an approach is computationally impractical due to transistor model (e.g. BSIM4 [13]) evaluation.

Our target is to develop a simplified transistor model which captures sufficient second-order effects and statistical process variations to allow accurate and efficient waveform and delay calculation for (S)STA.

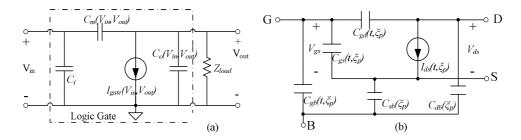


Fig. 1. a) current-source model; b) proposed SSTM

Recently, optimized GLMs typically model every gate by several capacitors and a current source as shown in Fig. 1a [3]. Although the CSM is less accurate for the whole gate representation for nanometer technology, the simple model is, however, appropriate for transistor modeling. The proposed SSTM shown in Fig. 1b represents every transistor by a statistical current source I_{ds} and five parasitic capacitances which also have statistical values as a function of the statistical process parameters of interest.

3.1 Current source modeling

Conventionally, without considering second-order effects of deep-micron MOS-FETs, the Shichman-Hodges model was replaced gradually by Deep Submicron MOSFET Models (DSMM) [14]. Although a DSMM substantially improves accuracy for submicron MOSFET behavior, our experiments in 45nm technology still show significant errors: i) due to channel length modulation (CLM), DIBL and substrate current induced body effect, the CLM parameter λ is a complicated function of V_{gs} and V_{ds} . As a consequence, the method to model saturation current to be a linear function of V_{ds} with constant slope starting from $I_{ds}(V_{dsat})$ is not accurate enough; ii) in the linear region, I_{ds} is no longer proportional to $(V_{gs} - V_{th} - \frac{1}{2}V_{ds})$. In fact the $\frac{1}{2}$ should be replaced by a factor which depends on $V_{gs} - V_{th}$; iii) the cutoff current can not be ignored any more. Simulation results show that when V_{gs} is smaller than V_{th} by a small amount, the current still has similar shape as the current when $V_{gs} > V_{th}$, which cannot be modeled as zero if the input slew and load capacitance are both small.

Similarly, the α -power law MOSFET model [15] is also widely used in digital circuit simulation. This model assumes that near- and sub-threshold region modeling is not important in calculating the delay of digital circuits, so the linear region is just approximated by linear lines and the saturation region current is constant. However, if the load capacitance and input slew are both quite small, the inaccuracy of the linear-region current significantly impacts the output waveform at the end of the transition, which introduces a large error for output slew. Taking these issues into consideration, the proposed BSIM4-based nominal current source I_{ds0} of SSTM in equation form is given as:

$$I_{ds0} = \begin{cases} He^{(V_{gst}/nV_t)}(1 - e^{(-V_{ds}/V_t)}) & V_{gs} \le V_{th} \\ \frac{W}{L} \cdot \left\{ JV_{gst}V_{dseff}\left(1 - \frac{V_{ds}}{2V_b}\right) / \left(1 + \frac{V_{ds}}{V_c}\right) \right\} \cdot \left[1 + \lambda(V_{ds} - V_{dseff})\right] V_{gs} > V_{th} \end{cases}$$

$$\tag{1}$$

where $V_{gst} = V_{gs} - V_{th}$, $V_b = V_{gst} + 2V_t$ and V_t is the thermal voltage. The main components are described as:

$$V_{dseff} = V_{dsat} - \frac{1}{2} \left(V_{dsat} - V_{ds} - \gamma + \sqrt{(V_{dsat} - V_{ds} - \gamma)^2 + 4\gamma V_{dsat}} \right)$$
(2)

$$V_{dsat} = V_c \cdot (V_{gst} + 2V_t) / (V_c + V_{gst} + 2V_t)$$
(3)

In order to link the continuous linear current with the saturation current, a smooth function (2) based on BSIM4 is used. V_{dseff} enables a unified expression for both linear and saturation currents. The threshold voltage V_{th} divides the I-V plane to two parts, thus accurate V_{th} modeling is important. According to the BSIM4 model, a linear dependence of V_{th} on V_{ds} is a good approximation. We simplify the V_{th} model as:

$$V_{th} = V_{th0} - \alpha \cdot V_{ds} + K_1(\sqrt{\Phi_s - V_{bs}} - \sqrt{\Phi_s}) - K_2 \cdot V_{bs}$$

$$\tag{4}$$

where V_{th0} is the zero-biased long-channel device V_{th} and α is a coefficient for drain/source charge sharing and DIBL effects on V_{th} . The coefficients K_1 , K_2 and surface potential Φ_s are obtained and derived from the technology file.

The model simplification focuses on the following items: i) instead of using complicated expressions, the parameter J considers several effects, including mobility degradation; ii) no consideration for narrow channel effect for V_{th} model; iii) V_{gsteff} model in BSIM4 [13] is replaced by V_{gst} since the unified expression for the current from strong inversion to linear region is not used. As a result, the I_{ds0} model and it's derivative are dramatically simplified. It should be noticed that the cut-off current could simply be modeled as zero if sharp input ramps and extremely small load capacitances rarely occur at the same time. Then the proposed model is simplified further to the 2^{nd} equation in (1) where only J and λ are obtained in the characterization stage.

The statistical description of I-V model is:

$$I_{ds} = I_{ds0}(t) + \sum_{k=1}^{m} \frac{\partial I_{ds}}{\partial p_k} |_{p_k = p_{k0}} (t) \cdot \xi_k = I_{ds0}(t) + \sum_{k=1}^{m} \chi_k(t) \cdot \xi_k$$
(5)

$$p_k = p_{k0} + \xi_k \qquad (k = 1 \sim m)$$
 (6)

where p_k is the k_{th} random process parameter which is the sum of nominal value p_{k0} and random variable ξ_k with zero mean (μ) and the same standard deviation (σ) as p_k . $\chi_k(t)$ is the differential function of I_{ds} by the elements of p_k .

3.2 Intrinsic capacitance modeling

The most accurate way to model non-linear capacitances is to represent them as voltage dependent terminal charge sources [13]. Characterization of such a model would involve generating charge tables for a range of terminal voltages. All capacitances are derived from the charge to ensure charge conservation. Each capacitance is computed by $C_{ij} = \partial Q_i / \partial V_j$ at every time step, where *i* and *j* denote the transistor terminals. Although this approach would be the most accurate, the massive amount of simulation time would be a problem for STA and SSTA.

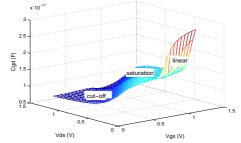


Fig. 2. Cgd variation for a minimum-sized NMOS

Using a single value for all capacitors promises fast simulation, but it results in an overly simple model which produces errors in (S)STA for nanometer technology. Fig. 2 shows the variation of C_{gd} for a minimum-sized NMOS. Clearly, at the 45nm node, the capacitances are too nonlinear to be accurately modeled as a constant value. In order to improve accuracy while maintain good computational efficiency, SSTM treats the five capacitances differently. For gate channel capacitances (GCC) C_{gs} , C_{gd} and C_{gb} , SSTM uses a constant value in the cut-off and saturation regions respectively, while approximates them as a linear function of V_{gs} and V_{ds} in the linear region. For junction depletion capacitances C_{sb} and C_{db} , SSTM uses a single value model since they are 1-2 orders of magnitude smaller than GCCs.

In the statistical extension of the capacitance model (7), C_{j0} is the nominal value of the j^{th} capacitance in Fig. 1 and the sensitivity ζ is characterized by perturbing the process variables of interest.

$$C_{j}(t,\xi) = C_{j0} + \sum_{k=1}^{m} \frac{\partial C_{j}}{\partial p_{k}} |_{p_{k}=p_{k0}} \cdot \xi_{k} = C_{j0} + \sum_{k=1}^{m} \zeta_{k} \cdot \xi_{k}$$
(7)

The characterization time of GLMs for SSTA is quite long since standard cell libraries consist of hundreds of cells with different sizes and process corners. In contrast, by using transistor-based gate modeling like SSTM, the characterization time is significantly reduced as only the unique transistors used in the cell library need to be characterized.

4 Non-MC Statistical Simulator

The proposed SSTM is embedded in our non-MC statistical simulator [16] for fast statistical timing analysis. In general, for deterministic time-domain analysis, the modified nodal analysis (MNA) equations for any circuit can be expressed in compact form as:

$$F(x', x, t, p_0) = 0 \qquad x(t_0) = x_0 \tag{8}$$

where x is the vector of the circuit state variables consisting of nodal voltages and branch currents and p_0 is the nominal process variable vector with elements p_{k0} introduced in (6). x' denotes the time derivative of x. Let x_s be the solution to (8). Transient analysis in a conventional simulator solves for x_s using numerical integration methods. However, the existence and importance of process variations at 45nm and below result in a random MNA which can be expressed as:

$$F(x', x, t, p) = 0 \qquad x(t_0) = x_0 + \delta_{x0} \tag{9}$$

where p is the statistical process variable vector with elements p_k introduced in (6). δ_{x0} denotes the initial variation caused by p.

It is computationally impracticable to solve (9) due to a large set of correlated random variables and the nonlinearity. Therefore, in order to make the problem manageable, we employ principal component analysis (PCA) to model a large set m in (6) of correlated p to a n-dimensional ($n \ll m$) vector of uncorrelated random variables, and linearize (9) with a truncated Taylor expansion. To avoid notational cluttering, the notation p representing the uncorrelated process variables after PCA is further used in the paper. The linear Taylor expansion is carried out at the point of x'_s , x_s and p_0 . Let's define $y(t) = x(t) - x_s(t)$ as the x(t) variation vector due to process variation ξ with zero μ and finite σ mentioned in (6). Re-organizing the 1st-order Taylor expansion of (9) we can obtain a compact format as:

$$y'(t) = E(t)y(t) + F(t)\xi$$
 $y(t_0) = \delta_{x0}$ (10)

The nonlinear random equation (9) is converted to a linear random differential equation (RDE) in y. According to the mean square (m.s.) integral theorem [17], there exists a unique solution. Assuming the initial condition x_0 is set to a fixed value, the solution is found as $y(t) = \alpha(t) \cdot \xi$. By substituting $y(t) = \alpha(t) \cdot \xi$ in (10), $\alpha(t)$ is easy to calculate by solving the resulting ODE.

Then the mean, variance and covariance of x(t) can be calculated as:

$$E\{x(t)\} = x_s(t) \qquad Var\{x_j(t)\} = \sum_{k=1}^n \alpha_{jk}^2(t) Var\{\xi_k\}$$
(11)

$$Cov(x_a, x_b) = \alpha(t_a) \cdot diag(Var\{\xi_1\}, \cdots, Var\{\xi_n\}) \cdot \alpha^T(t_b)$$
(12)

where $x_j(t)$ is the j_{th} element of vector x(t). As long as $\alpha(t)$ is calculated, y(t) is known, thus the covariance matrix of the solution y(t) at two different time points t_a and t_b can be calculated by (12).

From the waveform modeling point of view, the waveform is modeled as a time-indexed voltage array for STA while the mean, variance and covariance array are used for SSTA. Based on (11)-(12), the probability density function (pdf) of every crossing time for rising and falling transitions can be straightforwardly calculated by (13) and (14) respectively assuming the voltage at any time point is Gaussian distributed [16].

$$Pr(t_{r\eta} = t) = Pr(V_o(t - \Delta t) \le V_\eta) - Pr(V_o(t - \Delta t) \le V_\eta \cap V_o(t) \le V_\eta) (13)$$
$$Pr(t_{f\eta} = t) = Pr(V_o(t) \le V_\eta) - Pr(V_o(t - \Delta t) \le V_\eta \cap V_o(t) \le V_\eta)$$
(14)

where the crossing time t_{η} is the time when the node voltage crosses the corresponding voltage threshold $V_{\eta} = \eta \% \cdot V_{dd}$. $Pr(V_o(t - \Delta t) \leq V_{\eta} \cap V_o(t) \leq V_{\eta})$ is the joint *cdf* of V_o at two time steps. Note that the proposed method calculates the *pdf* directly and considers the correlation of V_o at two time steps in contrast to [18] and [19]. Given mean and variance of crossing time, the mean and variance of delay and slew can be calculated.

5 Experimental Results

The proposed SSTM and non-MC statistical simulation method were evaluated using 45nm PTMLP technology [20] and implemented in MATLAB. For SSTM, the data for characterization were obtained from Spectre using a BSIM4 model and then imported to a characterization algorithm in MATLAB to acquire the required parameters described in section 3. We present the accuracy evaluation of SSTM for minimum-sized cells, arbitrary inputs and MIS and the applicability of SSTM for power grid and signal integrity verification. In the end, the statistical simulation results were presented.

We evaluated the nominal SSTM when process variations are not included in SSTM in minimum-sized inverter and NAND2 cells with different input slew (S_{in}) and capacitive load (C_{load}) . The S_{in} ranges from 1ps to 500ps and the C_{load} spans from 0.5fF to 40fF. In comparison with Spectre using the BSIM4 model, It is clear from Fig. 3 (a)-(b) that the relative error for delay calculation is within 5%. 99.2% of the output rise delay and 93.9% of output fall delay are within 1.6%. The average relative error of output slew calculation is 1.2%. Although the maximum relative error is 3.3% with zero C_{load} , Fig. 3 (c)-(d) show the absolute error is nearly zero.

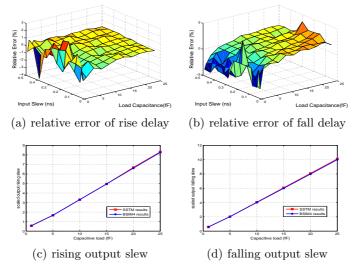


Fig. 3. delay and output slew evaluation

In essence, SSTM is input waveform independent so it can handle arbitrary input waveforms. Certain cells may experience simultaneous MIS and internal charge sharing during some specific input to output transitions. The transistorbased SSTM is able to handle these since every node is considered at the same time. Fig. 4 illustrates the accuracy of the nominal SSTM used in a minimumsized inverter with irregular input and a NAND2 cell in a simultaneous MIS scenario. The results show a very good match between the nominal SSTM and BSIM4 model.

Power supply integrity verification is an essential step in current design flows due to the large currents drawn through an increasingly resistive power supply network.

The models used in power grid analysis must capture the dynamic current characteristics of the cells. Fig. 5(a) shows the current drawn by a cell from

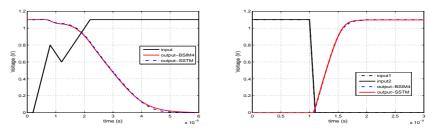
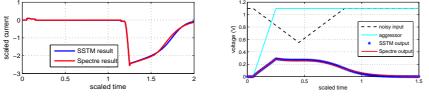


Fig. 4. left: irregular input; right: simultaneous MIS for a NAND2 cell

the power supply at both rising and falling transitions. It is easy for transistorbased gate models to capture the dynamic currents since the desired current is calculated during the simulation.

The primary modeling challenge for on-chip signal integrity verification has been the simulation of a driver (the victim), subject to an input noise, whose interconnect load is capacitively coupled to the output of another driver (the aggressor). In Fig. 5(b) we see the SSTM captures this scenario well. All waveforms in Fig. 5 show SSTM can be applied to power grid and signal integrity verification flows.



(a) SSTM to power grid verification(b) SSTM to signal integrity verificationFig. 5. SSTM's application to power grid and signal integrity verification

We combined SSTM with the proposed non-MC statistical simulation method for a large number of standard cells in a 45nm technology. The uncorrelated process variations are length and width variations with zero μ . The 3σ of length and width are 20% and 15% of the nominal length and the largest width of every cell respectively. In comparison with 1000 Monte Carlo trials in Spectre, the proposed modeling and simulation method achieved relative error within 1.4% for μ and within 6.8% for σ with an average $40 \times$ speedup [16].

6 Conclusion

At 45nm and below the gate models for circuit verification should account for increasing accuracy requirements and process variations. In this paper, a statistical simplified transistor model (SSTM) for transistor-level gate modeling which is embedded in our non-MC statistical simulator is presented. The SSTM-based gate model is independent of input waveform and output load, easy to characterize and suitable for SSTA, and accurate compared to Spice/Spectre for standard cells. We show that, in addition to handling accuracy limitations associated with conventional gate-level models for STA like arbitrary input, multi-input switching, etc., it is possible to be applied to power grid verification and noise verification flows. The statistical results show that our transistor-level timing analysis methodology achieves both high accuracy and efficiency.

References

- N. Menezes, C. Kashyap and C. Amin, "A "true" electrical cell model for timing, noise, and power grid verification," *Proc. of DAC*, pp. 462-467, 2008.
- C. Amin, C. Kashyap, N. Menezes and K. Killpack, "A multi-port current source model for multiple-input switching effects in CMOS library cells," *Proc. of DAC*, pp. 247-252, 2006.
- 3. A. Goel and S. Vrudhula, "Statistical waveform and current source based standard cell models for accurate timing analysis," *Proc. of DAC*, pp. 227-230, 2008.
- 4. P. Li and E. Acar, "Waveform independent gate models for accurate timing analysis," *Proc. of ICCD*, pp. 617-622, 1996.
- Qin Tang, A. Zjajo, M. Berkelaar and N. van der Meijs, "A simplified transistor model for CMOS timing analysis," *Proc. of ProRISC*, pp. 1-6, 2009.
- I. Keller, K. H. Tarn and V. Kariat, "Challenges in gate level modeling for delay and SI at 65nm and below," *Proc. of DAC*, pp. 468-473, 2008.
- S. Nazarian, M. Pedram, E. Tuncer and T. Lin, "Sensitivity-based gate delay propagation in static timing analysis," *Proc. of ISQED*, pp. 536-541, 2005.
- 8. J. F. Croix and D. F. Wong, "Blade and Razor: cell and interconnect delay analysis using current-based models," *Proc. of DAC*, pp. 386-389, 2003.
- C.S. Amin, F. Dartu and Y.I. Ismail, "Weibull based analytical waveform model," *IEEE Trans. on CAD*, vol. 24, pp. 1156-1168, 2005.
- S. Raja, F. Varadi, M. Becer and J. Geada, "Transistor level gate modeling for accurate and fast timing, noise, and power analysis," *Proc. of DAC*, pp. 456-461, 2008.
- 11. P. kulshrehtha, R. Palermo and M. Mortazavi, "Transistor-level timing analysis using embedded simulation," *Proc. of ICCAD*, pp. 344-348, 2000.
- Z. Li and S. Chen, "Transistor level timing analysis considering multiple inputs simultaneous switching," *Proc. of CADCG*, pp. 315-320, 2007.
- 13. BSIM4 Home Page In http://www-device.eecs.berkeley.edu/bsim3/bsim4.hml
- J.M. Rabaey, Digital integrated circuit: A design perspective. pp. 96-100, Upper Saddle River: Prentice Hall, 1996.
- T. Sakural and A.R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE JSSC*, vol. 25, no. 2, pp. 584-594, 1990.
- Qin Tang, A. Zjajo, M. Berkelaar and N. van der Meijs, "RDE-based transistorlevel gate simulation for statistical static timing analysis," *Proc. of DAC*, pp. 787-792, 2010.
- 17. T.T. Soong, Random differential equations in science and engineering, New York: Academic Press, 1973.
- H. Fatemi, S. Nazarian and M. Pedram, "Statistical logic cell delay analysis using a current-based model," *Proc. of DAC*, pp. 253-256, 2006.
- B. Liu and A.B. Kahng, "Statistical gate level simulation via voltage controlled current models," *IEEE Proc. of MBAS*, pp. 23-27, 2006.
- Predictive Technology Model for Low-power Applications (PTMLP), Nov. 2008, [Online], Available: http://www.eas.asu.edu/~ptm/modelcard/LP/45nm_LP.pm