Adv. Radio Sci., 12, 95–101, 2014
www.adv-radio-sci.net/12/95/2014/
doi:10.5194/ars-12-95-2014

Advances in Radio Science
Open Access Proceedings

An enhanced BSIM modeling framework for selfheating aware circuit design

M. Schleyer[1], S. Leuschner[2], P. Baumgartner[2], J.-E. Mueller[2], and H. Klar[1]

[1]Fachgebiet Mikroelektronik, Technische Universität Berlin, Germany
[2]Intel Mobile Communications GmbH, Munich, Germany

Correspondence to: M. Schleyer (martin.schleyer@tu-berlin.de)

Received: 4 February 2014 – Revised: 1 April 2014 – Accepted: 14 April 2014 – Published: 10 November 2014

Abstract. This work proposes a modeling framework to enhance the industry-standard BSIM4 MOSFET models with capabilities for coupled electro-thermal simulations. An automated simulation environment extracts thermal information from model data as provided by the semiconductor foundry. The standard BSIM4 model is enhanced with a Verilog-A based wrapper module, adding thermal nodes which can be connected to a thermal-equivalent RC network. The proposed framework allows a fully automated extraction process based on the netlist of the top-level design and the model library. A numerical analysis tool is used to control the extraction flow and to obtain all required parameters. The framework is used to model self-heating effects on a fully integrated class A/AB power amplifier (PA) designed in a standard 65 nm CMOS process. The PA is driven with +30 dBm output power, leading to an average temperature rise of approximately 40 °C over ambient temperature.

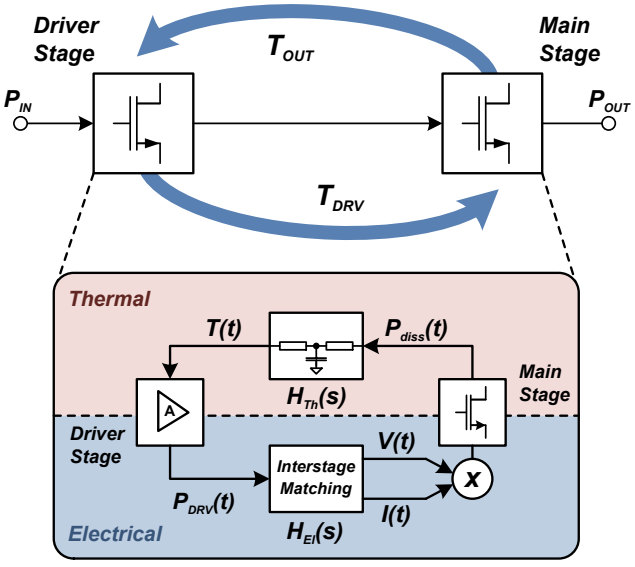


Figure 1. Memory effects due to electro-thermal resonances.

1 Introduction

In most analog and mixed signal radio frequency (RF) designs, static heat distribution is mainly a concern regarding the device matching and circuit performance. While self-heating is immanent in all RF integrated circuits (RFICs), its actual influence is negligible in most cases. As long as dynamic power dissipation is small compared to the dissipated DC power, no changes in the steady-state performance occur. However, the ongoing integration of system-on-chip environments, e.g. by integrating power amplifiers, adds notable dynamic heat sources on the same silicon die and affects the small signal properties other blocks, as heat is conducted through the chip. Therefore self-heating effects arise as new challenge also in RFIC design. For this reason, tools are required to investigate and avoid thermal issues while designing such circuits.

Apart from output power and efficiency degradation due to self-heating, one important phenomenon is the occurrence of thermal memory effects in RF power amplifiers. The Joule effect translates electrical power dissipation into a heat flux Q. The die itself and the packaging have a thermal impedance, which determine the temperature increase ΔT due to Q. As the material has a certain mass and density, the overall thermal impedance is not purely real but has a capacitive component (Vuolevi et al., 2001). Hence, a large

Table 1. Modeling approaches – overview.

	Effort	Speed	Accuracy	Risk	Flexibility
(A) Adaption of BSIM4 C-Code	–	+	+	–	+
(B) Customized Verilog-A model	○	○	+	–	+
(C) Behavioral description	+	○	–	+	○
(D) Table Based Model	+	–	○	+	–
(E) Combined *BSIM2THERM* Model	+	○	○	+	○

thermal time constant is added to the system – typically in the order of a few kilohertz.

This temperature change due to electrical power dissipation directly modifies the properties of active devices: both electron mobility μ_e and threshold voltage V_{th} of a FET device decrease due to the rising temperature. Figure 1 illustrates the electro-thermal interaction in a typical two-stage PA design. Here, thermal time constants resonate with electrical memory effects in the baseband frequency domain and can cause severe memory to the power amplifier (Wolf, 2012, p. 78). These self-heating effects in power amplifiers have been studied extensively using behavioral models. Boumaiza et al. (2003) follow the basic concepts as presented by Vuolevi et al. (2001), and use a thermal network to control a simple model inheriting the gain reduction of an LDMOS amplifier. Boumaiza et al. (2003) verify their model also with measurements for pulsed signals. Mazeau et al. (2007) apply dynamic Volterra series and obtain a coupled behavioral electro-thermal model. These approaches allow an investigation of thermal memory effects based on measurements of an actual implementation and allow to model the influence of self-heating effects on non-linear distortions and spectral regrowth. Anyhow, due to their nature as behavioral models, their use for circuit designers is very limited. As they model the whole block, no actual interaction between individual devices is investigated.

To close this gap, customized device models or simulation tools have been developed. Heo et al. (1999) propose a MOSFET large signal model targeting at LDMOS device design. They extend the default equations by first order temperature dependencies for drain current and threshold voltage in a custom device model. Codecasa et al. (2002) perform a decent analysis on electro-thermal resonance effects. With their results, the SPICE level 3 MOSFET model is extended to incorporate the electro-thermal effects into the circuit design environment. Unfortunately, none of these and other published approaches (Jardel et al., 2006; Du et al., 2008) give a robust and generic CMOS device model as required for self-heating aware design in standard IC design flows: they all require non-standard device models or manual work to find the behavioral descriptions. Actually, the recent PSP Level 103.2 device model (Smit et al., 2013) indeed supports an external thermal equivalent network to anticipate self-heating effects. However, PSP Level 103.2 models are not available for most standard CMOS technology nodes larger than 28nm. In contrast, the BSIM4 model (Xi et al., 2004) is still used in many wide-spread and cost-effective CMOS technologies – but does not allow dissipation-driven temperature changes.

This work proposes an extension to the widely used BSIM4 model. The BSIM4 model is enhanced using a Verilog-A wrapper module. It adds additional temperature and power nodes to convert the dynamically dissipated power of the particular device into a temperature change. This temperature change is used to determine variations of the device characteristics in addition to the original BSIM model equations.

2 Enhanced BSIM Modeling Flow

The BSIM4 model uses temperature-dependent equations to include thermal effects on various device properties. However, all those effects are modeled static, and the models cannot be used for dynamic electro-thermal simulations. The goal of the presented framework is to overcome this issue and to add support for external thermal equivalent networks.

2.1 Modeling Strategies

Several approaches have been already discussed within the introduction – each associated with its own advantages and drawbacks. A short summary is given in Table 1. The most straight forward implementation would be the adaption of the original BSIM4 source code (A). With this approach, a very generic and geometry independent self-heating aware model could be generated. The model is provided as source code using the SPICE API (Quarles, 1989). In total, the model contains approx. 25 000 lines of code. To support dynamic temperature changes, a major overhaul of this code would be required. Altering the model in such an intrusive way includes a severe risk in changing the numerical behavior and can lead to inconsistencies compared to the original models.

Next to the more complex C model, a Verilog-A compact model implementation of BSIM4 was investigated (B). This customized model is less complex, but is per se error-prone as the Verilog-A model cannot directly implement the same routines and calculations as the C code model. Another trade-off to consider is the reduced computational speed of the Verilog-A implementation. While still being compiled before

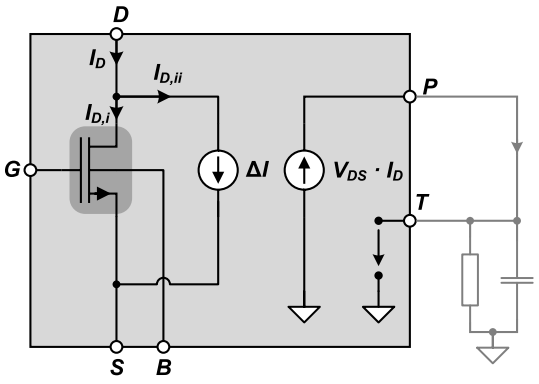


Figure 2. Equivalent circuit of the Verilog-A module.

run-time, it shows rather poor performance in comparison to the highly optimized binary implementation in C.

The danger of possibly deteriorating the model accuracy by altering its source code can be avoided by following a wrapper approach as e.g. proposed by Marbell and Hwang (2005). Behavioral sources are added to the underlying BSIM4 model (C). Although computationally efficient and with only slight implementation effort, a closed form description valid in all operation ranges is hard to find. The contradictory approach would be a table based model (D) which implements a look-up table based method for all operating conditions. The effort in terms of implementation and computations is very low, but a table-based model is generally less flexible and requires a huge amount of input data, if the complete operating range shall be covered. To allow implementation in both a reasonable time frame and with sufficient accuracy, a combined approach (E) is presented in this work, the so called *BSIM2THERM* framework. A behavioral source using a polynomial representation of drain current changes allows accurate modeling without modifying the BSIM4 source code. The coefficients for the polynomial representation are determined using fully automatized simulation and fitting routines.

2.2 *BSIM2THERM* Verilog-A module

In the recent years, the Verilog-A language superseded C and FORTRAN implementations for device models. Verilog-A based models do not require simulator- or vendor-specific coding when creating them (Troyanovsky et al., 2006). Thus, Verilog-A became the de facto language standard for compact device modeling – and has been used to implement e.g. the PSP or EKV device models. The *BSIM2THERM* modeling flow exploits the macro preprocessing capabilities of Verilog-A and splits the module in several parts. The core part of the wrapper contains descriptions of the device terminals and the branches required for current and voltage sensing and the controlled sources connected to the internal transistor device. Figure 2 shows the equivalent circuit

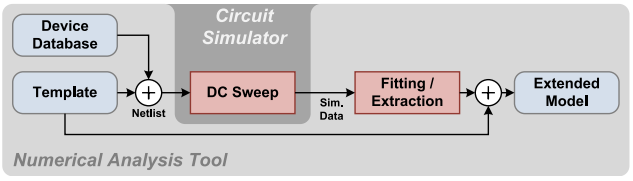


Figure 3. BSIM2THERM model generation flow.

of the overall module, with an additional external thermal-equivalent RC network. The modeling equations, coefficients and intermediate variables are defined in an additional file and referenced with macro statements. Hence, structure and functionality are separated, which allows better maintenance of the model database. The controlled source is dependent of the temperature applied to the T node. The current source $\Delta I(T)$ incorporates current changes on the V/I characteristics of the device due to the dynamic temperature variations. In case of a short-channel CMOS device, the most notable mechanisms are electron mobility reduction and the velocity saturation.

3 Model generation

The frameworks aims to allow easy and fast characterization of the initial BSIM model. Figure 3 shows the basic control flow: A template is combined with device data stored in the *device database* to create the wrapper module with all connections and parameters as used by the original devices. The framework allows to parse foundry-provided model libraries and extract all BSIM4 based devices provided in the technology library. A fully characterized BSIM4 model uses 200+ parameters. The foundry-provided models normally calculate some parameters internally, others are left unaltered. The device database therefore contains a list of the parameters which need to be externally accessible. Furthermore, it holds default values and data types for these parameters, as Verilog-A does not allow undefined or empty values for instance parameters.

3.1 Device characterization

Based on the top-level netlist of a design, the framework determines all BSIM4 instances and their instance properties, such as geometry or device stress information. To complete the input netlist for the model characterization, the user needs to set limits and step sizes for the individual input variables. Typically, this would be a range from V_{GS} and V_{DS} = 0 V ... V_{DD}. If the range is chosen too large, the fitting algorithm might not be able to properly fit sensitive areas – typically the transition between sub-threshold and linear region or linear and saturation region. The devices of interest are added to a Verilog-A based test bench. The simulator performs a nested DC sweep in the user-defined operating

region. Investigations showed that a decent coverage and accuracy is reached with simulation times of 20 min, executed single-threaded on a 2.9 GHz Intel® Xeon E5-2690 machine.

3.2 Polynomial model representation

The simulation data obtained by DC characterization is processed within a numerical analysis tool to obtain a closed form expression of the V/I characteristics. A direct approach maps the current change due to $T \neq T_{\mathrm{Nom}}$ into the source $\Delta I(T)$, such that

$$\begin{aligned}\Delta I(T) &= I_{\mathrm{D}}(T) - I_{\mathrm{D},i}(T = T_{\mathrm{Nom}}) \\ &\widehat{=} f\left(T, V_{\mathrm{DS}}, V_{\mathrm{GS}}, V_{\mathrm{BS}}, I_{\mathrm{D},i}\right). \end{aligned} \quad (1)$$

Hence, a polynomial expression $P : X \rightarrow \Delta I$ with tupel $X = \left\{V_{\mathrm{DS}}, V_{\mathrm{GS}}, V_{\mathrm{BS}}, I_{\mathrm{D},i}, T\right\}$ can be used to approximate ΔI. If P is of degree N, and has $n = |X| = 5$ variables, it will require $k = \binom{N+n}{n}$ coefficients. For a higher-order degree calculation, this results in a high number of arithmetical calculations performed at each solver iteration step. The Verilog-A interpreter and compiler only does very basic code optimizations. It is therefore inevitable to reduce the arithmetic operations in forehand. A very simple technique is the *precalculation* of certain intermediate variables at runtime.

A second step is *iterative coefficient pruning*. Per default, the degree N is used for all input variables of X. If the degree N of one of the determinants is too high, the underlying QR decomposition delivers very small coefficients for high-order terms. As those coefficients c_k do not significantly contribute to the overall current ΔI, setting all $|c_k| < \varepsilon$ to zero directly reduces the computational effort while only marginally reducing the accuracy. The fitting algorithm removes those terms from the design matrix of P_i and repeats the QR decomposition delivering P_{i+1}, where i denotes the number of iterations starting from 0. If the approximation error $A_{i+1} = \left|1 - \frac{P(X}{\Delta I(X)}\right|$ is not increased by more than an arbitrary chosen boundary E, a new acceptable representation $P_{i+1}(X)$ has been found. This iteration is repeated until E is finally crossed, and then P_i is kept as final representation for ΔI. With varying ε and E, the trade-off *accuracy vs. speed* can be set to an optimal point by empirical investigations. An additional weighting algorithm on the approximation error A further improves the overall model quality, as it penalizes errors in critical domains and adds relaxations in other regions. The representation P states that $I_{\mathrm{D},i} \in X$. While not obvious in the regression model, $I_{\mathrm{D},i}$ is actually dependent on all voltages as this is a boundary condition from the BSIM4 model. Anyhow, this redundancy has the advantage that the polynomial scales with $I_{\mathrm{D},i}$. With $\widetilde{X} = X \setminus I_{\mathrm{D},i}$, where $|\widetilde{X}| = \tilde{n} = n - 1$, the complexity can be reduced by costs of loosing the scaling property.

In Eq. (1), ΔI is a polynomial of the current change. It is obvious that $P(X)|_{T=T_{\mathrm{Nom}}}$ results in $\Delta I = 0$. This property can be exploited to further reduce the computational effort. Instead of P, a new polynomial model $Q(\tilde{X})$ is defined,

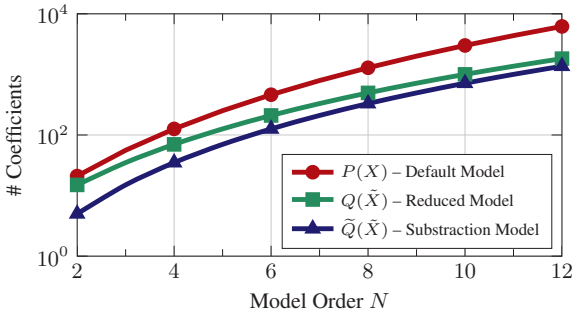


Figure 4. Comparison of different models.

where Q is a model of $I_{\mathrm{D},i}(\widetilde{X})$, i.e. a polynomial representation of the underlying BISM4 model itself under bias and temperature conditions $\tilde{X}$. Now, Eq. (1) can be expressed as

$$\begin{aligned}\Delta I(T) &= I_{\mathrm{D}}(T) - I_{\mathrm{D},i}(T = T_{\mathrm{Nom}}) \\ &= Q(\widetilde{X}) - Q(\widetilde{X},\ \ T = T_{\mathrm{Nom}}) = \widetilde{Q}(\widetilde{X}, T_{\mathrm{Nom}}) \end{aligned} \quad (2)$$

The polynomial $\widetilde{Q}$ is pre-calculated within the numerical analysis tool. A characteristic of $\widetilde{Q}$ lays in the subtraction: all terms of Q which are not related to T are equal in both sub-terms of $\widetilde{Q}$. It can be considered as

$$\widetilde{Q}(\widetilde{X}, T_{\mathrm{Nom}}) = \sum_{m=1}^{N} \left(T^m - T_{\mathrm{Nom}}^m\right) \cdot \bar{Q}(\bar{X}) \,, \quad (3)$$

where $\bar{X} = \{V_{\mathrm{DS}}, V_{\mathrm{GS}}, V_{\mathrm{BS}}\}$ is now independent of T. Thus $|\bar{X}|$ is reduced to $|\bar{X}| = \bar{n} = n - 2$. Hence, the polynomial $\widetilde{Q}$ has a reduced number of coefficients $\tilde{l}$ compared to l of the original $Q(\widetilde{X})$:

$$\tilde{l} = l - \binom{\tilde{n} + N - 1}{\tilde{n} - 1} = \binom{\tilde{n} + N - 1}{\tilde{n}} = \frac{N}{N + \tilde{n}} \cdot \binom{N + \tilde{n}}{\tilde{n}} . \quad (4)$$

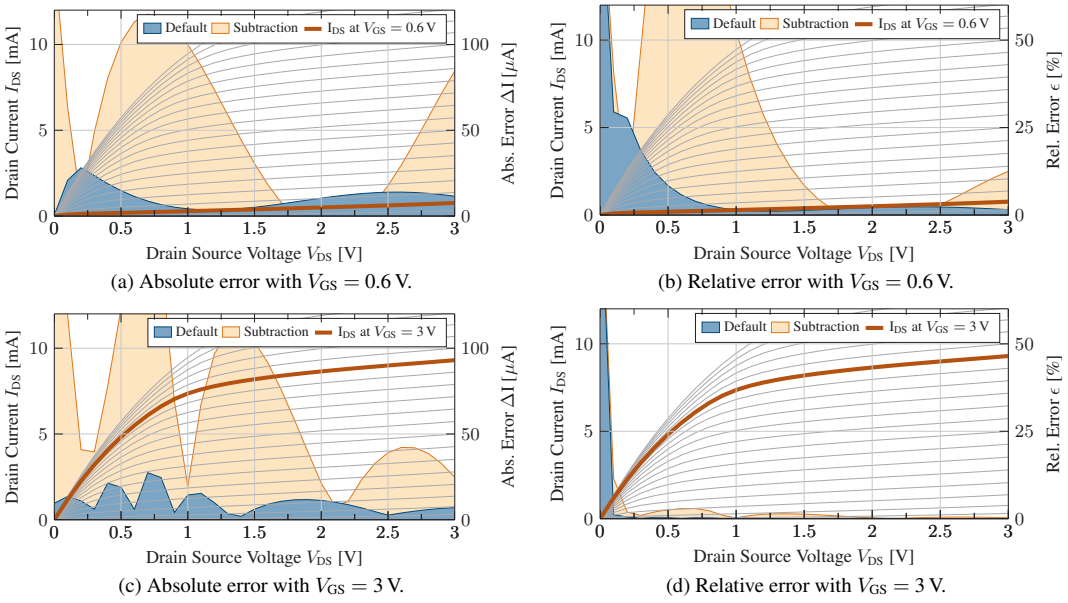
The advantage of this modeling method is directly stated in Eq. (4). Either the number of coefficients is reduced to $\tilde{l} = (l \cdot N)/(N + \tilde{n})$, or the model order can be increased to $\tilde{N} = N + 1$ while $l = \tilde{l}$. If compared to k, the advantage is even bigger. It can be proven that $l = (k \cdot n)/(N + n)$ and finally

$$\tilde{l} = k \cdot \frac{n \cdot N}{(N + n - 1)\ (N + n)} . \quad (5)$$

Figure 4 shows the number of coefficients N for the different modeling strategies. The model based on $P(X)$ is from hereon named *default model*, whereas the model based on $\widetilde{Q}(\tilde{X})$ is named *subtraction model*. The plot also shows the $Q(\tilde{X})$ model, here stated as *reduced model*. In the following, the two first mentioned will be investigated further by employing the modeling concept on an exemplary design.

4 Application example: CMOS RF Power Amplifier

To evaluate the capabilities of the framework, a fully integrated RF CMOS power amplifier for WCDMA operation



(a) Absolute error with $V_{GS} = 0.6$ V.
(b) Relative error with $V_{GS} = 0.6$ V.
(c) Absolute error with $V_{GS} = 3$ V.
(d) Relative error with $V_{GS} = 3$ V.

Figure 5. Output characteristics and modeling error at $V_{GS} = 0.6$ V and 3 V.

is analyzed regarding its electro-thermal properties. The PA is biased in class AB operation. It is built in a differential two-stage stacked-cascode structure with on-chip matching networks (Leuschner et al., 2011). The maximum linear output power $P_{Out,max}$ is + 27.9 dBm with a PAE of ≈ 48 %. The dissipated DC power $P_{Diss,DC}$ is 340 mW typical and rises to $P_{Diss,Max} \approx 1.3$ W for large signal operation.

While not in the scope of this work, a decent modeling of the thermal properties is an important issue to achieve accurate simulation results. For the evaluation at hand, input and output stage have been connected to a single-stage RC network, based on estimations regarding thermal impedance of the package, giving a first approximation with acceptable modeling effort.

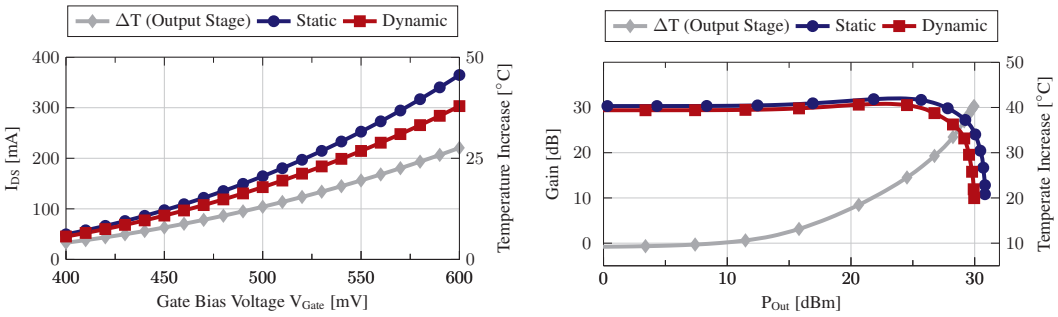
4.1 Model accuracy

The model accuracy is investigated by evaluating a thin-oxide I/O NMOS device with $W \approx 10$ µm and $L = 190$ nm. Both models use a polynomial of order $N = 5$. *Iterative coefficient pruning* was enabled with a boundary of 10^{-9}. For the *default model*, 125 additions and 237 multiplications are required to calculate the drain current change within the Verilog-A module. Due to the reduced number of coefficients and the iterative pruning, only 52 additions and 104 multiplications are required for the *subtraction model*.

The model was evaluated in the overall characterization range. Figure 5 shows absolute and relative errors at two different operating points – one closer to $V_{GS} = V_{th}$, the other with the device fully open. In both cases, $V_{BS} = 0$ V is applied. The shaded areas show the maximal errors over the complete temperature range from 0 . . . 100 °C. In Fig. 5, the $I_{D,i}$ scaling effect of the polynomial is clearly visible, as the default model shows significantly reduced absolute errors for low currents. Furthermore, the trade-off between no. of coefficients and accuracy is present: the subtraction model reduced the computational effort significantly less than ≈ 0.5 of the default model, but shows less accuracy especially in low current regions with small V_{th}.

4.2 Simulation results and performance

The obtained models where used to simulate the presented transistor stack and to investigate typical thermal issues. First, the operating point due to thermal runaway is simulated. Figure 6 shows that the bias current has an increasing offset with higher gate voltages. As expected, the additional self-heating reduces the overall current due to changes in electron mobility and velocity saturation. The gray line (right ordinate) implies that the die temperature increases by ≈ 17.8 °C for a DC operating point of $I_{DC} \approx 200$ mA. These numbers illustrate that the framework is a valuable enhancement for bias design and temperature-independent biasing structures. A commonly seen issue in power amplifier design is the reduced saturation power due to self-heating of the power transistors. For continuous-wave operation, the power amplifier has a reduced gain compared to pulsed operation with small duty cycles. Figure 6 shows that this effect can be foreseen in simulations. Using the results of the DC simulation, the PA has been biased to a DC current of ≈ 100 mA in the output stage. The large signal behavior is evaluated in a single-tone *harmonic balance* simulation to obtain the AM/AM characteristics of the PA. The expected drop in output power is ≈ 0.9 dB at an output power level of +20 dBm. Here, the temperature increase is estimated to 38.8 °C.



(a) DC bias shift.

(b) Overall PA gain, I_{DC} = 100 mA.

Figure 6. Simulated power amplifier characteristics with subtraction model.

Both simulations were executed multi-threaded on a 2.9 GHz Intel® Xeon E5-2690 machine. The time for the combined DC and HB simulations decreased slightly from 395.4 s to 429.4 s. Thus, it can be stated that the additional time constants due to the thermal RC network does not have a major impact on the overall simulation performance and the circuit showed good convergence.

5 Conclusions

The modeling framework presented in this work uses model data provided by the semiconductor foundries to obtain information about the thermal dependency of the drain current in the desired operation regions. An automated simulation procedure generates input data for a fitting process. The final result provides equations describing the dynamic temperature behavior of the device without altering the original model data. Hence, the approach is applicable for CMOS technologies where no foundry or vendor support for other models is available. An integrated CMOS power amplifier was evaluated and a good estimation of the self-heating behavior could be found with good simulation performance. Based on the framework, counter-measures to thermal-related issues can be taken before a first silicon is available for measurements.

Acknowledgements. The authors gratefully acknowledge research funding provided by the DFG (German Research Foundation), grant no. KL 918/8-1.

Edited by: D. Killat
Reviewed by: two anonymous referees

References

Boumaiza, S., Gauthier, J., and Ghannouchi, F.: Dynamic electro-thermal behavioral model for RF power amplifiers, in: Microwave Symposium Digest, 2003 IEEE MTT-S International, 1, 351–354, doi:10.1109/MWSYM.2003.1210950, 2003.

Codecasa, L., D'Amore, D., and Maffezzoni, P.: Modeling the thermal response of semiconductor devices through equivalent electrical networks, IEEE T. Circuits Syst., 49, 1187–1197, doi:10.1109/TCSI.2002.801279, 2002.

Du, B., Hudgins, J., Santi, E., Bryant, A., Palmer, P., and Mantooth, H.: Transient thermal analysis of power devices based on Fourier-series thermal model, in: Power Electronics Specialists Conference (PESC), 2008 IEEE, 3129–3135, doi:10.1109/PESC.2008.4592433, 2008.

Heo, D., Chen, E., Gebara, E., Yoo, S., Laskar, J., and Anderson, T.: Temperature dependent MOSFET RF large signal model incorporating self heating effects, in: Microwave Symposium Digest, 1999 IEEE MTT-S International, 2, 415–418, doi:10.1109/MWSYM.1999.779791, 1999.

Jardel, O., Quere, R., Heckmann, S., Bousbia, H., Barataud, D., Chartier, E., and Floriot, D.: An Electrothermal Model for GaInP/GaAs Power HBTs with Enhanced Convergence Capabilities, in: European Microwave Integrated Circuits Conference, The 1st, 296–299, doi:10.1109/EMICC.2006.282811, 2006.

Leuschner, S., Mueller, J.-E., and Klar, H.: A 1.8 GHz wideband stacked-cascode CMOS power amplifier for WCDMA applications in 65 nm standard CMOS, in: Radio Frequency Integrated Circuits Symposium (RFIC), 2011 IEEE, doi:10.1109/RFIC.2011.5940623, 2011.

Marbell, M. N. and Hwang, J.: A Verilog-based temperature-dependent BSIM4 model for RF power LDMOSFETs, in: Microwave Symposium Digest, 2005 IEEE MTT-S International, doi:10.1109/MWSYM.2005.1516882, 2005.

Mazeau, J., Sommet, R., Caban-Chastas, D., Gatard, E., Quere, R., and Mancuso, Y.: Behavioral Thermal Modeling for Microwave Power Amplifier Design, IEEE T. Microw. Theory, 55, 2290–2297, doi:10.1109/TMTT.2007.907715, 2007.

Quarles, T.: Adding Devices to SPICE3, Tech. Rep. UCB/ERL M89/45, EECS Department, Univ. of California, Berkeley, 1989.

Smit, G. D. J., Scholten, A. J., Klaassen, D. B. M., and van der Toorn, R.: PSP 103.2, Technical Note NXP-TN-2012-0080, 2013.

Troyanovsky, B., O'Halloran, P., and Mierzwinski, M.: Compact modeling in Verilog-A, in: Transistor Level Modeling For Analog/RF IC Design, edited by Grabinski, W., Nauwelaers, B., and Schreurs, D., 271–291, Springer Netherlands, doi:10.1007/1-4020-4556-5_10, 2006.

Vuolevi, J., Rahkonen, T., and Manninen, J.: Measurement technique for characterizing memory effects in RF power amplifiers, IEEE T. Microw. Theory, 49, 1383–1389, doi:10.1109/22.939917, 2001.

Wolf, N.: Charakterisierung von Leistungsverstärkern für die Entwicklung neuer und einfacher Vorverzerrungssysteme, Ph.D. thesis, Technische Universität Berlin, 2012.

Xi, X., Dunga, M., He, J., Liu, W., Cao, K. M., Jin, X., Ou, J. J., Chan, M., Niknejad, A. M., and Hu, C.: BSIM4.5.0 MOSFET Model, User Manual, 2004.