Transient electro-thermal characterization of Si–Ge heterojunction bipolar transistors

Amit Kumar Sahoo *, Mario Weiss, Sébastien Fregonese, Nathalie Malbert, Thomas Zimmer

Laboratoire IMS, CNRS – UMR 5218, Université de Bordeaux 1, Cours de la Liberation, 33405 Talence Cedex, France

Article history:
Available online 26 April 2012

Keywords:
Bipolar transistors
Pulse measurement
Numerical simulation
Thermal resistance
Thermal capacitance
Semiconductor device modeling
Electro-thermal effects

A B S T R A C T

In this paper, a comprehensive evaluation of the transient self-heating in microwave heterojunction bipolar transistors (HBTs) have been carried out through simulations and measurements. Three dimensional thermal TCAD simulations have been performed to investigate precisely the influence of backend metallization on transient thermal behavior of a submicron SiGe:C BiCMOS technology with \( f_T \) and \( f_{max} \) of 230 GHz and 290 GHz, respectively. Transient variation of Collector current caused by self-heating is obtained through pulse measurements. For thermal characterization, different electro-thermal networks have been employed at the temperature node of HiCuM compact model. Thermal parameters have been extracted by means of compact model simulation using a scalable transistor library. It has been shown that, the conventional R–C thermal network is not sufficient to accurately model the transient thermal spreading behavior and therefore a recursive network needs to be used. Recursive network is verified with device simulations as well as measurements and found to be in excellent agreement.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

The downscaling of electronic device dimensions has been expected to be the main way to continue miniaturization as seen by the International Technology Roadmap for Semiconductors ITRSSs to achieve better high frequency performance. Over the past decades this downscaling has allowed the semiconductor industry to gain significant progress in high speed and new circuit applications such as automotive radar and 100 Gb/s data transfer. To achieve higher frequency performance, the transistor is made in such a way that its operating quiescent point is shifted to higher current densities and therefore power consumption of devices has increased significantly, resulting in self-heating effect in power and RF hetero-junction bipolar devices. Moreover, the small geometries in combination with trench isolation reduce parasitic components and provide high-frequency performance [1–3]. The thermal properties of trench isolated devices are quite different than the conventional substrate technology due to the confinement of heat flow through trench wall [3,4]. Also the low thermal conductivity of the oxide used in trench wall would be expected to affect the heat dissipation. Thermal issue is one of the key factors limiting the performance and reliability of the devices and integrated circuits, therefore systematic characterization of thermal effects inside the devices remains very important criteria to explore. A rigorous analysis for device temperature and thermal resistance (\( R_{TH} \)) depending on trench architecture is presented by Walkey et al. [4,5] through an analytical model verified by 3D numerical simulations and measurements. The thermal behavior of devices fabricated in different isolation technologies like STI, DTI and SOI is compared by Haralson et al. [6] through a 2D device simulation. Also a precise analysis for device structure dependent \( R_{TH} \) is given by Marano et al. for trench isolated bipolar device [7] and also fabricated on SOI substrate [8]. In most investigations research has been concentrated only on \( R_{TH} \). However, in this work a methodology in order to extract both \( R_{TH} \) as well as \( C_{TH} \) through 3D TCAD simulations is presented. Thermal simulations are performed for trench isolated device structures including the backend metallization. The influence of the metal contact and the first metallization layer on \( C_{TH} \) is investigated.

In general, the characterization of the self-heating effect is based on steady state condition to extract \( R_{TH} \) and transient condition to extract \( C_{TH} \). A number of attempts have been accomplished to study self-heating effect in steady state conditions. In Refs. [9–16], measurements are performed at different base-plate temperatures using temperature dependence of common-emitter current gain (\( \beta \)) as a thermometer. Another simplified method has been proposed in [17], measuring temperature dependent output characteristic at constant Base current. The transient temperature response can be obtained by pulses applied to the collector node and keeping the Base current constant [18]. Measuring Base-Emitter voltage permits to determine the change in temperature inside the device. There are several pulse measurement techniques presented [19–21], which are based on the assumption...
of isothermal operation during the short pulse and measuring the pulse characteristic at different ambient temperature. However, there are major limitations in these methods. Precise calibrations of the passive elements related to coaxial cables and connectors must be taken into account and should be accurately characterized.

Various analytical models have been developed to investigate self-heating effect with different electro-thermal networks like single RC network, Foster network, Nodal network, Cauer network, recursive network, etc. [22–27]. A systematic study on comparison among these electro-thermal networks in the frequency domain has been presented in [28]. It has been seen that, recursive electro-thermal network provides the best compromise between accuracy, number of model parameters and physical basement.

In the first part of this paper, we discuss about the transient thermal simulations with TCAD for the three dimensional device structures of a BiCMOS technology. The lattice temperature (\(T_{\text{lattice}}\)) and heat flux (\(P_{\text{heat}}\)) distribution inside the device has been studied. Moreover, the influence of the backend on \(R_{\text{TH}}\) and \(C_{\text{TH}}\) has been examined. In the second part, we discuss transient measurements, applying pulses at Base and Collector terminals simultaneously and measuring the transient response of Collector current increase due to self-heating. In third part, time domain thermal modeling with different electro-thermal networks is presented. A conventional single RC network and a distributed recursive type network are compared where the recursive network presented in [29] is used. The \(R_{\text{TH}}\) extracted from pulse measurements and transient simulations are verified with DC measurements for different geometries. The DC measurements are performed at different ambient temperatures and the \(R_{\text{TH}}\) are extracted with the method described in [17].

2. Thermal TCAD simulations

From the microscopic point of view, the energy is given up by the constituent particle such as atoms or free electrons in a system possessing vibrational motion. The displacement of these particles from their equilibrium position generates heat energy in the system. Conduction is the mode of heat transfer in which energy exchange takes place in the medium from the region of high temperature to the region of low temperature due to the presence of temperature gradient in the body. In an electronic device, the temperature gradient occurs at a given electrical power and the material physical properties changes as they are strongly dependent of temperature [18]. For hetero-junction bipolar transistors the temperature raises at BC junction due to higher localized electric field and it continues to rise with increase in power density. Accordingly a thermal runaway arises at a critical current, strongly dependent of \(R_{\text{TH}}\) which is the ratio between the variation of temperature at Base–Collector junction and the dissipated power. To evaluate the heat diffusion mechanism inside the device, a 3D numerical simulation is performed with Sentaurus device simulator. The model is defined by the lattice heat flow equation, the \(T_{\text{lattice}}\) and the \(P_{\text{heat}}\) distribution due to electro-thermal effect inside the device are calculated by solving lattice heat flow equation,

\[
\frac{\partial T}{\partial t} - \nabla \cdot (\kappa \nabla T) = -\nabla \cdot \left[ (P_f T + \Phi_h) J_n + (P_p T + \Phi_p) J_p \right]
\]

where \(\Phi_h\) and \(\Phi_p\) are the electron and hole quasi-Fermi potentials, \(P_f\) and \(P_p\) are the absolute thermoelectric powers and \(T\) the lattice temperature, \(\kappa\) is the thermal conductivity and \(c_l\) is the lattice heat capacity of the material. The electron \(J_n\) and hole current density \(J_p\) arise due to the non-uniform distribution of temperature inside the device. It gives the flow of current due to the temperature gradient.

**Fig. 1.** (a) The transistor is placed in a semi-infinite Si block, where \(L_z = 300\ \mu m\). (b) Structure of SiGe HBT (\(L_x \times W_x = 9.88 \times 0.15\ \mu m^2\)) with CBE configuration: E, B and C are Emitter, Base and Collector respectively.

<table>
<thead>
<tr>
<th>DC Power</th>
<th>(\Delta T_{\text{DC}} (K))</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 mW</td>
<td>200</td>
</tr>
<tr>
<td>4 mW</td>
<td>150</td>
</tr>
<tr>
<td>2 mW</td>
<td>100</td>
</tr>
</tbody>
</table>

**Fig. 2.** DC simulation for size optimization of semi-infinite Si-block: \(L_x\) and \(L_y\) are described in Fig. 1.

### 2.1. Device under analysis and it’s structure formation

The analyzed NPN transistors are fabricated within the STMicroelectronics with SiGe:C BiCMOS9MW technology which is a quasi self-aligned trench isolated technology. The key figures of this technology are [30]: breakdown voltage, \(BV_{\text{ZET}} = 1.6\ V\), \(BV_{\text{CEO}} = 5.5\ V\), transition frequency (\(f_t\)) = 230 GHz and maximum oscillation frequency (\(f_{\text{max}}\)) = 290 GHz. The geometrical parameters are: shallow trench height (\(D_{\text{ST}}\)) = 0.3 \(\mu m\), deep trench height (\(D_{\text{DT}}\)) = 2.5 \(\mu m\).

3D transient thermal simulations have been performed with Sentaurus device simulator version E-2010.12 [31]. We assume a constant thermal conductivity (\(\kappa\)) (no temperature dependence) and a constant specific heat (\(C_v\)) such that the physical simulation is linear. The device under analysis, fabricated with the Sentaurus Structure Editor, has been “placed” on a semi-infinite Si-block as shown in Fig. 1a. The thickness (\(L_z\)) of the Si-block is 300 \(\mu m\) same as Si-wafer. The transistor configuration is CBE (1 Collector, 1 Base and 1 Emitter) i.e., according to Fig. 1b, the structure is symmetric only with respect to x-axis, therefore \(1/2\) of the device is considered for simulation. The width (\(L_x\) and \(L_y\)) of the 3D simulation domain along the emitter stripe i.e., the dimension of the semi-infinite Si-block is optimized through DC simulation. The surface of the
Si-block is adiabatic and, therefore, no heat flows from the device to outside of the block. In Fig. 2, the device temperature rise ($\Delta T_{DC}$) vs. dimension of $L_x$ and $L_y$ has been plotted when different DC powers are applied at the heat source. It is evident that, nearly above 100 µm the $\Delta T_{DC}s$ are stable at every applied DC power. For the simulation accuracy we have considered $L_x = L_y = 300$ µm.

In a first attempt, the transistor structure is simplified by neglecting the upper part for computation time purpose. We have considered only the lower part starting from BC junction of the device and taking into account the deep trench and shallow trench. The BC junction is considered as the heat source. The dimension of the heat source is same as the Emitter window ($L_x \times W_x$), where $L_x$ and $W_x$ are the length and width of the Emitter respectively. In this simulation the upper surface of BC junction is adiabatic which means that heat will diffuse towards the lower part of the device. In the second part of simulation we have added the backend metal contact for Emitter, Base and Collector with the first metallization layer.

2.2.Transient simulations

In order to perform large signal thermal simulation, a pulse of electric power ($P_{\text{pulse}}$) is applied at the BC junction of the device structure. The transient variations of junction temperature ($\Delta T_J$) are obtained numerically. The $F_{\text{Heat}}$ and $T_{\text{Lattice}}$ distribution inside the complete device structure at pulse time of 5050 ns are shown in Fig. 3. The $\Delta T_J$ vs. time for two different simulations: (i) for the lower part of device where, there is no heat flow above the heat source and (ii) for the complete device including the backend metal region are shown in Fig. 4.

From Fig. 4, it is seen that the transient device temperature change ($\Delta T_J$) decreases when the backend metal layers are added. The influence of this metal layer on $R_{\text{TH}}$ and the $C_{\text{TH}}$ can be obtained from these results.

3. Transient measurements

On chip pulse measurements have been performed with MC2 technology APMS LPM1/HPM1 pulse generator and Agilent 6633B system DC power supply on RF HBTs in Ground–Signal–Ground (GSG) configuration at room temperature. Different Base ($V_{BE}$) and Collector ($V_{CE}$) pulse and bias conditions are used for the measurements. The experimental setup for pulse measurements is described in Fig. 5. Pulses are applied through the DC port of the bias network where RF port is grounded with 50 Ω resistance. Before switching the pulse source, a quiescent bias point is set. In the measurement the device is switched from a state of negligible power dissipation to a high power dissipation using a low-duty-cycle (~5%), fast-rise-time (10 ns) and pulsed Base as well as Collector. The Collector current transient variation is strongly influenced by device self-heating. Since for a very fast pulse, the electrical device response is much faster than the thermal response, there is inadequate time for the temperature to change over the duration of the rise time. Therefore, the transient variations of the Collector current provide an image of the device self-heating. Pulse measurements show often Collector current overshoot when applying the pulse and this is due to the passive elements which may arise from the coaxial cables, connectors, bias-T, etc. It has been seen through measurements that the capacitance ($C_{\text{cab}}$) and inductance ($L_{\text{cab}}$) of cables are most responsible for the noise arising in Base and Collector current pulse profiles and the values of $C_{\text{cab}}$ and $L_{\text{cab}}$ increase with the length of cables. Therefore, it is necessary to optimize these elements to minimize the parasitic effect. These optimizations have been performed through various measurements and compact model simulations. A measurement result in optimized condition for an NPN HBT with the dimension of Emitter window ($L_x \times W_x$) = $9.88 \times 0.15$ µm$^2$ is given in Fig. 6, where we applied 3000 ns $V_{\text{BE}}$ and 5000 ns $V_{\text{CE}}$. The $V_{\text{BE}}$ and $V_{\text{CE}}$ are switched from 0.7 V to 0.9 V and 0.3 V to

![Fig. 3.](image1.png) (a) The $F_{\text{heat}}$ and (b) $T_{\text{lattice}}$ distribution inside the transistor at pulse time of 5050 ns with $P_{\text{pulse}} = 5$ mW: The dimension of the heat source ($L_x \times W_x = 9.88 \times 0.15$ µm$^2$).

![Fig. 4.](image2.png) Transient device temperature variation different device structure with dimension of heat source ($L_x \times W_x = 9.88 \times 0.15$ µm$^2$): transient TCAD simulation with applied $P_{\text{pulse}} = 5$ mW and 5000 ns width.
1.5 V respectively (Fig. 6a). These bias conditions are very close to the peak $f_T$ (about 40 mV above) where self-heating is strongly pronounced. This is necessary for measurement accuracy. The measured Base ($I_B$) and Collector ($I_C$) currents are plotted in Fig. 6b.

4. Electro-thermal modeling

In this section, thermal parameters for two different electro-thermal networks, the conventional single $R_{TH}$–$C_{TH}$ network and the distributed recursive electro-thermal network, are extracted from transient measurements as well as from transient TCAD simulations. The recursive network is a physical approach to characterize the transient thermal spreading behavior of the transistors. The heat diffusion from the active region can be simplified as a spherical heat conduction path [27]. This also shows that every element is equivalent to a heat generating sphere. With further distance from the active region the area of a sphere (equithermal surface) increases. An elementary spherical slice at a distance $z$ (towards thermal ground from heat source) with thickness $\Delta z$ is characterized by a capacitance $\Delta C(z)$ and resistance $\Delta R(z)$ [26] as follows.

$$\Delta R(z) = \frac{1}{kA(z)} \Delta z \quad \text{and} \quad \Delta C(z) = C_vA(z)\Delta z$$  

(2)

Here $k$ is the thermal conductivity of Silicon, $C_v$ is the specific heat at constant volume of Silicon and $A(z)$ is the area of the spherical equithermal surface element at a distance $z$. Therefore the time-dependent diffusion of heat can be modeled by a distributed recursive network, where the respective thermal resistance element decreases and the capacitance element increases, which has been modeled by the multiplication factor $K_r (<1)$ and $K_c (>1)$ with $R$ and $C$ respectively (Fig. 7b). Since several new nodes are added through the self-heating network, simulation time may suffer especially for transient simulations, but higher number of cells may provide better accuracy in short time domain. In our thermal
transient, we have considered nine cells which provide sufficient accuracy in $\Delta T_P$ and $I_C$ modeling. This network also has been verified in a state of the art ring oscillator circuit with 53, 106 and 424 transistors, respectively. In our simulations even for the maximum number of transistors (424) convergence issues have not occurred.

4.1. Transient device temperature ($\Delta T_P$) modeling

The transient variation of $\Delta T_P$ is mainly defined by $C_{TH}$ where the steady state temperature is given by $R_{TH}$. In Fig. 8, $\Delta T_P$ has been modeled with single $R_{TH}$-$C_{TH}$ network. When a transient power is applied to the device, theoretically an infinite number of thermal time constant are necessary to represent the thermal response of the material due to the distributed nature. Thus, dynamic self-heating could not be properly modeled with a conventional $R_{TH}$–$C_{TH}$ network, as it has only a single time constant. But maximum and minimum value of $C_{TH}$ can be obtained through long time and short time extraction as shown in Fig. 8.

The $\Delta T_P$ for both lower part device and with backend metallization has been modeled with the recursive electro-thermal network as given in Fig. 9. A very good agreement has been obtained through modeling the overall transient region with recursive network. The extracted parameters from these thermal modeling provides the influence of backend metallization on thermal $R_{TH}$ and $C_{TH}$ which will be discussed in part-V.

4.2. Transient Collector current ($I_C$) modeling

Transistor models for BiCMOS9MW technology were generated by a scalable HiCuM library provided by XMOD Technologies. In Fig. 10 the output characteristics for a device with an emitter area $A_E$ ($9.88 \times 0.15 \, \mu m^2$) and CBE configuration are shown. The forward Gummel plot for different emitter geometries are displayed in Fig. 11. As shown, the transistor library accurately models the DC behavior of several geometries.

Compact model simulation including parasitic elements is mandatory to achieve an accurate parameter extraction. Therefore, parasitic components like coaxial cables, connectors, bias network and internal resistance of the pulse generator have been characterized. The thermal parameters for both networks are extracted from transient $I_C$ measurements. A comparison between measurement and HiCuM compact model simulation connecting a conventional single $R_{TH}$–$C_{TH}$ network at the temperature node is shown in Fig. 12. Steady state $I_C$ is modeled with $R_{TH}$ where the maximum and minimum limit of $C_{TH}$ is obtained with long time and short time extraction, respectively.

In order to model the whole transient thermal response on $I_C$, a recursive network has been connected at the temperature node of
the compact model. A comparison between compact model simulations and the measurements are shown in Fig. 13. Excellent agreement has been obtained. The quasi-isothermal current is extracted without any thermal network.

5. Results and verification

The variation of $R_{TH}$ vs. Emitter length is shown in Fig. 14. $R_{TH}$ extracted from DC and pulse measurements are very close which confirms the accuracy of both methods.

In Fig. 15 the average value of $C_{TH}$ extracted from pulse measurements and TCAD simulations are plotted for different Emitter lengths. The maximum and minimum limit of $C_{TH}$ is indicated through the error bar. The $R_{TH}$s and $C_{TH}$s extracted from the TCAD simulation (for lower part device structure) are higher than the measurements. This may be due to the adiabatic condition for the heat flux through metal contacts. In this simulation, the heat flow through the metal contact has been neglected. Therefore, in the next simulation step a part of the backend structure is added which consists of SiGe Base, poly Si-Base, Y-shape emitter, via contacts for Emitter, Base, Collector and one metal layer. The device structure with an Emitter window of $(L_E \times W_E) = 9.88 \times 0.15 \text{μm}^2$ has been taken to investigate the influence of upper metallic part on the $R_{TH}$ and $C_{TH}$.

The thermal modeling with recursive network has been described in part IV (A). The extracted values of distributed $C_{TH}$s ($k_n^C C (W s/K)$) of each cell of the recursive network for both structures are
compared in Fig. 16. The extracted $R_{TH}$ and the average thermal capacitance are given in Fig. 17.

In order to verify the accuracy of the extracted thermal capacitance, low frequency S-parameters measurements have been performed on the same device. This method has been proven to give accurate results [28]. The thermal resistance extracted from the DC measurements and the thermal capacitance extracted from the low frequency S-parameters measurements are considered as a references in this study. A statistical comparison among all the measurements and simulations has been given in Fig. 17a and b. It is observed that, by adding the upper part, $R_{TH}$ decreases nearly 4% and the $C_{TH}$ increases nearly 27.4% which is very close to the reference value. This huge influence of the backend layer can be understood as follows: In fact, the heat flux diffusion in any medium depends on the temperature gradient. As shown in Fig. 3, the upper metallic region is much closer to the heat source and therefore, the temperature gradient is much smaller in comparison to the lower part of the device. The thermal capacitance is a function of the temperature rise associated with a given quantity of applied energy. It is also a function of material properties, the product of material’s specific heat, density and volume.

$$C_{TH} = \frac{q}{\Delta T} = c\rho V$$

Here, $q$ is the heat transfer per second, $t$ the time, $\Delta T$ the temperature increase, $c$ the specific heat, $\rho$ the density and $V$ is the volume of the material. The specific heat of Silicon is 0.7 J/g °C, where it is $0.385$ J/g °C for Copper. Therefore (3) gives some reasonable contribution of thermal capacitance for upper metallic region which is not negligible.

6. Conclusion

In this study, a 3D electro-thermal device simulation has been performed to appraise the thermal resistance and capacitance effect inside the device. A complete device structure has been built including contacts and first metallization layer in order to investigate their influence on the device temperature. The heat flow through the metal contacts causes a 4% decrease in $R_{TH}$ which is in agreement with previous studies. Moreover, it has been found that, the backend structure has a major impact on $C_{TH}$. For thermal characterization, two different electro-thermal networks have been examined. In comparison with the conventional $R-C$ network, the recursive network provides the best compromise among accuracy, number of model parameters and physical basement. Moreover, compact model simulation using the recursive network found to be in excellent agreement with pulse measurements. The accuracy in time domain in combination with the simple parameter extraction methodology provides a better prediction of circuit performance due to thermal issues for the design engineers.

Acknowledgements

This work is part of the Rf2THz Project. The Authors would like to thank XMOD TECHNOLOGIES for supplying the compact model parameters and STMicroelectronics for BiCMOS wafer supply.

References


