Università degli Studi di Padova

Corso di Laurea Magistrale
in Ingegneria Elettronica

Université de Bordeaux

Laboratoire de l’Intégration
du Matériau au Système

Characterisation of Self-Heating in SiGe:C HBTs and InP HBTs

Supervisors

Prof Enrico Zanoni
DEI – University of Padova

Prof Cristell Maneux
Bordeaux IMS

Co-Supervisor

Dr Sébastien Frégonèse
Bordeaux IMS

Master Candidate

Marco Cabbia

Academic Year 2016/2017

Abstract

In this thesis, the topic of self-heating of heterojunction bipolar transistors (HBTs) for millimetre wave and terahertz circuits is described. The physical background, the importance of pulsed measurements and thermal resistance for the characterisation of thermal phenomena and the components being investigated are extensively introduced. These are based on two different types of compound semiconductor: SiGe:C (a silicon-germanium carbon-doped alloy, used in the BiCMOS fabricated by STMicroelectronics) and InP (indium phosphide, used in the InP/InGaAs HBT fabricated by III-V Lab and in the InP/GaAsSb HBT by ETH Zurich). A program for practical retrieval and elaboration of data from pulsed measurements on these devices has been developed, which allows considerations on their electro-thermal characteristics. Also, an algorithm for the extraction of the thermal resistance has been implemented. It is validated though simulations and used to trace out the resistance alteration due to an internal temperature rise, which has not been widely studied yet.

Sommario

In questa tesi verrà affrontato l'argomento dell'autoriscaldamento dei transistor bipolari a eterogiunzione (HBT) per circuiti a onde millimetriche e terahertz. Dapprima si introdurrà l'argomento dal punto di vista fisico; in seguito, si parlerà estesamente dell'importanza delle misure pulsate e della resistenza termica nella caratterizzazione dei fenomeni termici, presentando i componenti che saranno oggetto dello studio. Essi sono costituiti da due diversi tipi di semiconduttori composti: SiGe:C (una lega di silicio e germanio drogata di carbonio, che è usata nei BiCMOS fabbricati da STMicroelectronics) e InP (fosfuro d'indio, usato invece negli HBT InP/InGaAs prodotti da III-V Lab e negli HBT InP/GaAsSb prodotti dal ETH di Zurigo). Grazie allo sviluppo di un programma per una raccolta e un'elaborazione più pratiche dei dati misurati in questi dispositivi, saranno possibili delle considerazioni sulle loro caratteristiche elettro-termiche. È stato creato, inoltre, un algoritmo per l'estrazione della resistenza termica la cui validità verrà confermata attraverso delle simulazioni e che sarà utilizzato per tracciare l'andamento della resistenza in funzione dell'innalzamento della temperatura interna: un tipo di analisi che non è stata ancora ampiamente condotta in letteratura.

Contents

Chapter 1

Introduction

The issue of thermal spreading resistance as a major limit for the operating speed of electronic devices is no new problem, since it was already envisioned by R. Keyes in 1969 [1]. At that age, the size of devices was not yet so small to achieve high performances in terms of speed and power and the junction thermal resistance was negligible compared to the total thermal resistance.

Modern devices, however, have almost reached the physical limits foreseen by G. Moore and the rise of temperature inside the devices generates ever-growing concern, since it limits performance and reliability of devices, and ultimately of all the radio-frequency circuits (RFICs) that integrate them.

It is important to accurately measure and predict temperature rise phenomena, in particular self-heating, which is dominant, in analogue circuits, over thermal coupling and rise due to chip-package-ambient thermal impedances. This is important for a reliable characterisation of these devices' performance, which is temperature-dependent at steady-state. Furthermore, at high current densities, the proportionality between the temperature rise and the dissipated power cannot be considered as direct, thus making the predictability of their behaviour even more troublesome.

In Chapter 2, a general overview of the devices and their applications in the radio-frequency (RF) domain is carried out. The importance of the so-called “terahertz gap” is outlined, along with the devices which are able to fill this sector of electronics. Those are high electron mobility transistors (HEMTs) and heterojunction bipolar transistors (HBTs) with excellent potential for high-speed, high-power RF applications. The latter will be the subject of this study. Their strong points and the figures of merit for high-frequency (HF) operations will be described.

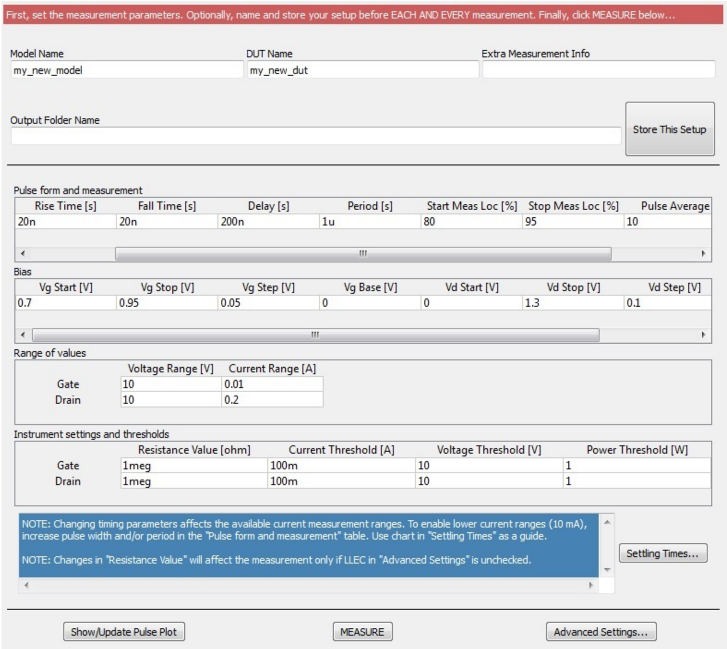
We will go deeper in the description of the technology, briefly explaining the production process, the motivation of their specific use and the difference between them, in their structure, performance and applications. Two different typologies of compound semiconductors are taken into account, a technologically mature Si-based HBT (Si/SiGe:C) and advanced InP-based HBTs (InP/InGaAs and InP/GaAsSb), provided by STMicroelectronics, III-V Lab and the Federal Institute of Technology in Zurich (ETH).

A pulsed measurement-lead evaluation of self-heating will be discussed in Chapter 3, since that kind of measurement allows quasi-isothermal measurements. To do so, a graphical user interface (GUI) has been implemented on the Integrated Circuit Characterization and Analysis Program (IC-CAP) by Keysight (see Fig. 1.1) in order to remotely control – via a GPIB connection – the Keithley automated system for I/V and pulsed-I/V measurements (Keithley 4200-SCS) and retrieve data.

The measures are effectively accomplished by a semi-automatic probe station, which, together with the pulsed measurement system, is part of the NANOCOM platform of the IMS Laboratory of the University of Bordeaux, which is dedicated characterise, model, and extract the fundamental parameters of RFICs [2] (see Fig. 1.2).

Pulse characteristic and bias definitions, as well as practical data manipulation via IC-CAP and storage are effectively implemented for a number of DC pulsed measurements on bipolar transistors and FETs. Thanks to this new software set-up, pulsed measurements will be quickly visualized and analysed for all the aforementioned technologies, and geometries will be compared.

Transient-I/V measurements will be studied to provide a good approx-



First, set the measurement parameters. Optionally, name and store your setup before EACH AND EVERY measurement. Finally, click MEASURE below...

Model Name: my_new_model

DUT Name: my_new_dut

Extra Measurement Info:

Output Folder Name:

Pulse form and measurement

Rise Time [s]	Fall Time [s]	Delay [s]	Period [s]	Start Meas Loc [%]	Stop Meas Loc [%]	Pulse Average
20n	20n	200n	1u	80	95	10

Bias

Vg Start [V]	Vg Stop [V]	Vg Step [V]	Vg Base [V]	Vd Start [V]	Vd Stop [V]	Vd Step [V]
0.7	0.95	0.05	0	0	1.3	0.1

Range of values

	Voltage Range [V]	Current Range [A]
Gate	10	0.01
Drain	10	0.2

Instrument settings and thresholds

	Resistance Value [ohm]	Current Threshold [A]	Voltage Threshold [V]	Power Threshold [W]
Gate	1meg	100m	10	1
Drain	1meg	100m	10	1

NOTE: Changing timing parameters affects the available current measurement ranges. To enable lower current ranges (10 mA), increase pulse width and/or period in the "Pulse form and measurement" table. Use chart in "Settling Times" as a guide.

NOTE: Changes in "Resistance Value" will affect the measurement only if LLEC in "Advanced Settings" is unchecked.

Figure 1.1: The IC-CAP interface used to retrieve pulsed measurements.

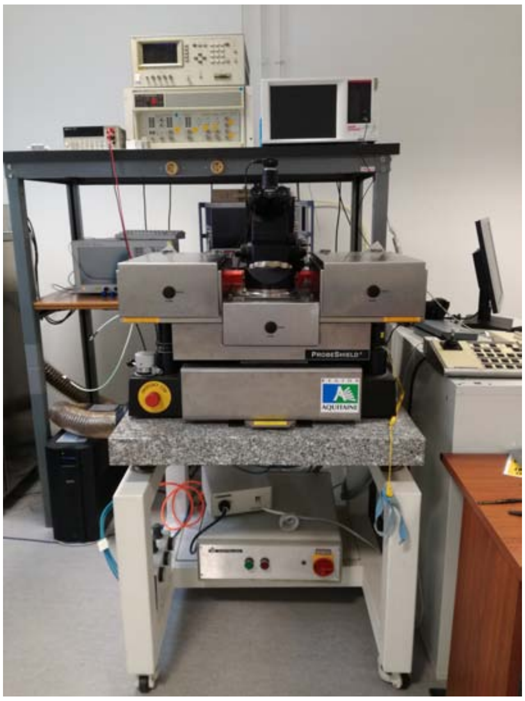


Figure 1.2: The part of the NANOCOM platform which served for the measurements in this thesis.

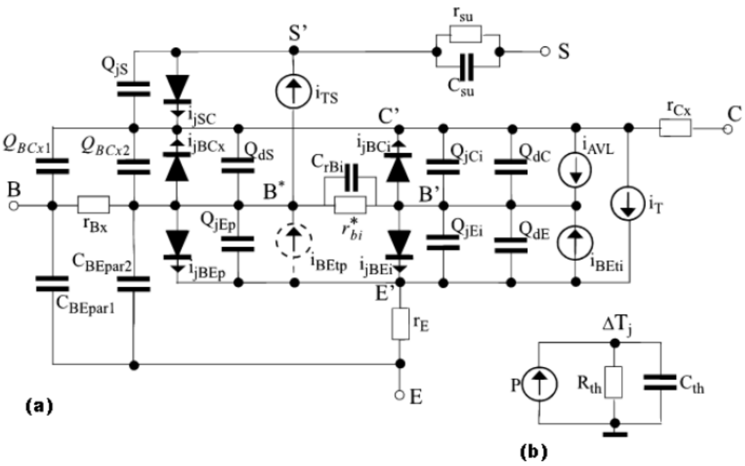


Figure 1.3: (a) Large-signal HICUM/L2 equivalent circuit. (b) Adjunct single-pole network for self-heating.

imation of some of the parameters of the devices, both electrical (the base resistance) and thermal (the thermal time constant), as well as giving an opportunity to analyse the measurement process in detail.

Then, in Chapter 4, classic DC measurement will be considered to extract the thermal resistance R_{TH} with an intuitive and practical method and its dependence on the temperature rise inside the device will be shown.

The extraction method has been fully implemented on IC-CAP and can automatically retrieve pairs of $R_{TH} - T_j$ values. It will be validated by using HICUM Level 2 [3], a circuit model (see Fig. 1.3) implemented in Verilog-A over many years of research on SiGe and InP HBTs.

R_{TH} will be extracted as an example for some of the geometries of our devices, and the thermal resistance of our InP-based devices, of which we did not dispose of the compact model, will be compared to the Si-based, to understand how differently the effect of self-heating impacts on these technologies.

In Chapter 5, the measurement perspectives given by the S-parameters extracted with a network analyser will be mentioned. Finally, in Chapter 6, conclusions will be drawn.

Chapter 2

General Outline

2.1 Millimetre and Terahertz Radiation

Fig. 2.1 shows a representation of the electromagnetic spectrum where the bands are shown as a function of frequency, wavelength and energy. At the lower end of the picture is the radio spectrum, a region of frequency – from few kHz to few GHz – where classical radio systems usually work (FM, AM, cellular radio, etc...). The other side is the domain of optical radiation bands – around $10^{13}\,Hz$ – and it extends up to gamma rays – up to $10^{21}\,Hz$. Optoelectronics deals with light considered as both visible and invisible radiation (photodiodes, lasers, optical fibers, etc...).

Millimetre waves (also abbreviated as *mmW* or *MMW*) and terahertz radiation (also called submillimetre radiation and abbreviated as *T-ray* or

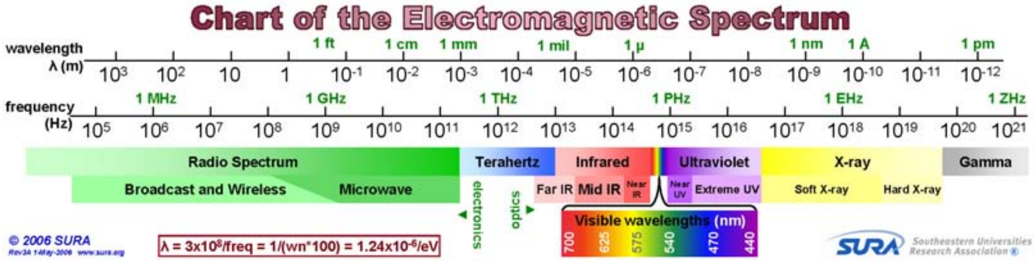


Figure 2.1: The electromagnetic spectrum (after [4]).

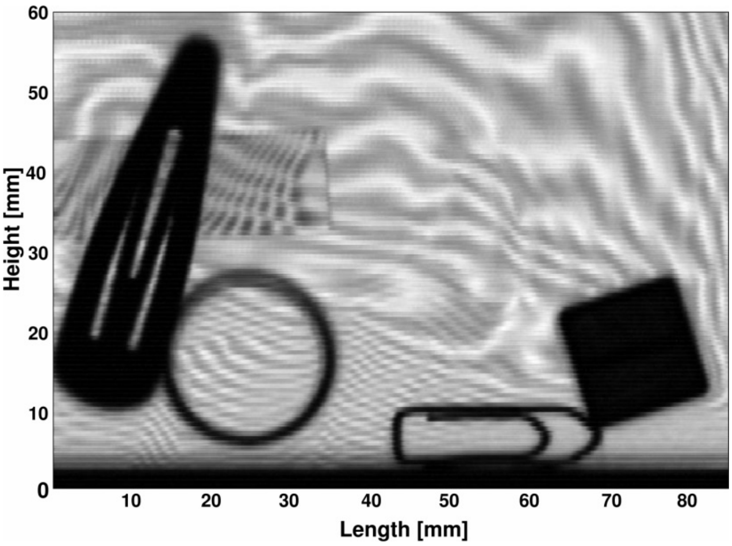


Figure 2.2: Captured 0.6-THz image of a postal envelope, revealing its hidden office supplies (such as paper clips, tape and candy bars). (after [6]).

simply THz) cover the region of the spectrum from microwave to optical frequencies, the so called "terahertz gap". More specifically, the range for the millimetre band is $30 - 300\,GHz$ ($30 - 300 \cdot 10^9\,Hz$ or equivalently $\lambda = 1 - 10\,mm$), while the range for the terahertz band proper is $300\,GHz$ to $3\,THz$ ($0.3 - 3 \cdot 10^{12}\,Hz$ or equivalently $\lambda = 100\,\mu m$ to $1\,mm$) [5]. This regime carries a lot of the benefits of both sides: like radio waves it can penetrate objects and walls, but it has wider bandwidths, and like visible light it has very short wavelengths which give very precise measurements and high quality images (Fig. 2.2).

At the radio side of the spectrum we typically use electronic devices and the power available decreases at higher frequency; while for the upper side of the spectrum photonic devices are used in which the energy per photon decreases at lower frequencies, and so does the available power for these devices. A growing number of devices can now reach enough power while working in the millimetre and terahertz regions and the systems that imply them become more and more scaled, powerful and affordable. This is due to an evolution of the semiconductor technologies.

There are some unique specifications that make terahertz waves attractive for the scientific community. To stimulate a transition between energy states in order to measure the spectrum of molecules (e.g. in the gas state), rotational and vibrational frequencies are considered. Many of those lie in the millimetre and submillimetre region, making terahertz spectroscopy a valuable tool for non-invasive chemical identification. Also, many optically opaque materials are more transparent to terahertz frequencies, and this allows to see through these objects and access environments that are usually inaccessible. Terahertz wave energy is much less than X-rays and even optical waves (photons at $1\,THz$ have an energy of just $4.1\,meV$ [6]). As a result, they are non-ionizing, non-destructive, and they can be used for different medical imaging and sensing applications, as they have less chances of damaging tissues and causing destruction in products.

f_T and f_{max} are two fundamental figures of merit for designing high frequency circuits (they will be considered later in this work), but in systems where high-frequency components – like HBTs – are used, they do not provide an exhaustive characterization of the overall circuit. Amplifier noise figure, oscillator phase noise, linearity and output power are to be taken into account when designing such circuits, of which the actual working frequency will be reduced. However, when employed in circuits based on subharmonic mixing, they enable applications operating at more than $1\,THz$ with decent output power [6].

2.1.1 Some Applications

The electronic devices that can reach such high frequencies and high powers are mainly used to fabricate monolithic microwave integrated circuits (MMICs) that perform the function of power amplifiers, low-noise amplifiers (LNAs), flash ADCs and voltage-controlled oscillators. It has already been mentioned that the panorama of possible applications is vast.

For example, in medical imaging, a skin cancer image obtained by a THz camera gives better contrast compared to classical optical imagining because of the strong absorption by water at these specific wavelengths. Also, in many early stages of organic material decay when physical damage has not happened yet, classical imaging cannot see any kind of erosion even if actually a transformation is happening. Though since this decaying material has a different water content, THz waves prove to be reliable in detecting them.

Like in the commonly experienced airport security screenings, which allow to find potentially dangerous carried objects by seeing through clothing and luggage, many screenings can be made by devices working at millimetre waves. By extending these technologies to higher frequencies, it can be possible to take higher resolution images – even from a large distance – by also applying higher powers.

In the pharmaceutical industry, there are two main issues tied up with medicine production. When a capsule is produced, for instance, one wants to make sure that the right thickness of coating is used because this will affect how long it takes to the drug to be released into the body. Terahertz is good in both penetrating the capsule enough to measure the thickness and at the same time preserving its contents. The other problem is that the chemical inside the drug tends to crystallize in different forms, which again affects its solubility in water and its release time into the body, and with many classical chemical sensors one is not able to sense that. However, terahertz can identify different rotational and vibrational frequencies of molecules and therefore spot in a non-invasive way if there are wrong crystalline forms.

For industrial quality control, there exist many applications to see if the right thickness of a material is used. In food industry, it's possible to determine if crops and grains quality is good and if they are fresh.

In safety systems for automotive, many radars are needed to accurately track the movement of objects all around the vehicle, opening the way to fully autonomous cars with cameras able to detect obstacles and moving objects under poor visibility and with a quick response. Millimetre waves circuits compose radars, as well as wireless and optical transmitters, blazing a trail to fifth generation (5G) and Internet of Things (IoT) networks.

Though, the very first type of application where terahertz waves were applied was space and atmospheric studies: THz sensors have been flying for years now within Earth-observing satellites to detect specific spectral lines of gases which determine the health of the Earth atmosphere and the ozone layer. Nevertheless, these satellites are carrying very large and heavy lasers; by imagining to develop even more these technologies in the future, one will get more compact and more performing sources to fit in a single satellite.

2.2 Bipolar Transistors

Even if the usage of MOSFETs in the industry of semiconductors is predominant nowadays, bipolar transistors have been vastly used in the past and thanks to improvements in bipolar technologies, they are currently maintaining and extending their predominance in many circuit applications, notably in high-speed performing systems.

The bipolar structure provides several natural advantages, such as: a short transit time for electrons flowing from emitter to collector (higher cut-off frequency); higher output current due to electron flowing through the entire emitter area – not just a channel –; direct control of the output current through the input voltage leading to high transconductance; turn-on voltage independent of size; possibility of working either with high and low currents without experiencing considerable delays [7].

In Fig. 2.3 the energy-band diagram of a BJT along the direction of the electrons' travel is shown. Its configuration shows the forward active mode, that is the emitter-base junction (EBJ) is forward biased while the base-collector junction (CBJ) is reverse biased. In this way, electrons can surmount the EBJ barrier potential and they are swept through the CBJ by the strong electric field inside the space charge region (SCR).

The input voltage leads to a flow of holes into the base, mainly composed by holes recombining in the base region (typically small) and holes flowing to the base-emitter depletion region and emitter – we assemble them a single

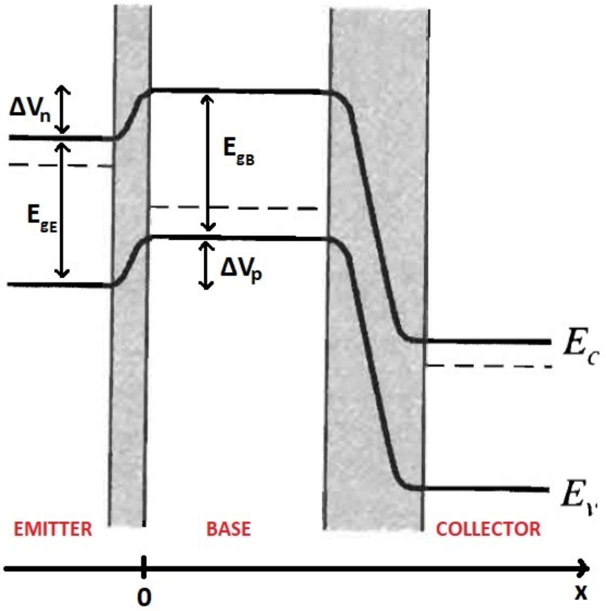


Figure 2.3: Energy-band diagram of a *npn* homojunction bipolar transistor.

component, I_{pE} – where they recombine.

The current gain of a bipolar transistor is calculated as follows [8] [7]. For this calculation, the SCR widths are neglected. The excess concentration of minority carriers (in this case, electrons in the p-doped base) at the EBJ interface (in Fig. 2.3, at $x = 0$) is

$$n_B(x=0) = n_{Eo} \exp\left(-\frac{q V_{bi}}{kT}\right)\left[\exp\left(\frac{q V_{BE}}{kT}\right) - 1\right] \tag{2.1}$$

where n_{Eo} is the emitter free-electron concentration (it corresponds to the emitter n-type dopant concentration), and V_{bi} is the built-in voltage. q is the elementary charge value, k is the Boltzmann constant, T is temperature in kelvin, V_{BE} is the applied bias. This expression could be written

$$n_B(0) = \frac{n_i^2}{p_{Bo}}\left[\exp\left(\frac{q V_{BE}}{kT}\right) - 1\right] \tag{2.2}$$

by exploiting the law of junction. n_i is the intrinsic concentration of carriers and p_{Bo} is the base free-hole concentration.

We can also write that same equation in a different way. We want to highlight the conduction band energy difference between the quasi-neutral emitter and base regions, whether a bias is applied or not (see Fig. 2.3). So we define ΔV_n in order to approximate Eq. (2.1) as

$$n_B(0) \simeq n_{Eo} \exp\left(-\frac{q\,\Delta V_n}{kT}\right) \tag{2.3}$$

In the base region, the electron density of current main component is the diffusion one, which is given by a gradient of concentration

$$J_{nB} = qD_B\frac{dn_B(x)}{dx} \tag{2.4}$$

where D_B is the diffusion coefficient in the base. Since the base is short we can also approximate the minority concentration profile with a linear decreasing trend

$$\frac{dn_B(x)}{dx} \simeq -\frac{n_B(0)}{W_B} \tag{2.5}$$

where W_B is the base width, which approximates the diffusion length of minorities in the base, $L_B \simeq W_B$. Given also that the number of electrons recombining in the base is low, so that $J_{nB} \simeq J_{nC}$, we can combine Eq. (2.3) and (2.4) and obtain

$$I_C = -I_{nC} = \frac{A_E\,q\,D_B\,n_{Eo}}{W_B}\exp\left(-\frac{q\,\Delta V_n}{kT}\right) \tag{2.6}$$

where A_E is the emitter area. We can write I_C in its well-known form by using Eq. (2.2) as the starting point, thus obtaining

$$I_C = I_s\left[\exp\left(\frac{V_{BE}}{V_T}\right) - 1\right] \tag{2.7}$$

where I_s is the subthreshold current and $V_T = kT/q$ is the so-called *thermal voltage.*

Considering the hole current, which is composed of two components, we approximate as: $I_B \simeq I_{pE}$. The component of holes flowing to the emitter will be similar to I_C, with coefficients referring to the emitter instead of the base

$$I_{pE} = \frac{A_E\,q\,D_E\,p_{Bo}}{L_E}\exp\left(-\frac{q\,\Delta V_p}{kT}\right) \tag{2.8}$$

where all the coefficients are analogous as before, p_{Bo} is the base free-hole concentration (it corresponds to the base p-type dopant concentration) and ΔV_p is the valence band energy difference between the quasi-neutral emitter and base regions as shown in Fig. 2.3.

By isolating I_C, we get

$$\begin{aligned} I_B &= I_C \left(\frac{D_E}{D_B} \cdot \frac{p_{Bo}}{n_{Eo}} \cdot \frac{W_B}{L_E}\right) \exp\left(\frac{\Delta V_n - \Delta V_p}{V_T}\right) \\ &= I_C \, \frac{1}{k} \exp\left(\frac{\Delta V_n - \Delta V_p}{V_T}\right) \end{aligned} \tag{2.9}$$

where κ has been defined as

$$k = \frac{D_B}{D_E} \cdot \frac{n_{Eo}}{p_{Bo}} \cdot \frac{L_E}{W_B}$$

and takes into account all the contributes due to the diffusion of carriers, the concentration and the geometry.

Finally, the current gain β_F is defined as the ratio between the two currents, namely $\beta_F = I_C/I_B$. Eq. (2.6) and (2.9) yield

$$\begin{aligned} \beta_F &= k \, \exp\left(\frac{\Delta V_p - \Delta V_n}{V_T}\right) \\ &= k \, \exp\left(\frac{\Delta E_g}{q\,V_T}\right) \end{aligned} \tag{2.10}$$

where $\Delta E_g = E_{gE} - E_{gB} = q\,(\Delta V_p - \Delta V_n)$. So this is the difference between the energy gaps of base and emitter (Fig. 2.3). Some general considerations can be done.

Firstly, the ratio between n_{Eo} and p_{Bo} is to keep high, in order to maintain k high, or equivalently get high current gain. So that is why for a Si homojunction transistor, for example, emitter doping levels of $10^{20}\,cm^{-3}$ and base doping levels on the order of $10^{17} - 10^{18}\,cm^{-3}$ are typically used. Further, Eq. (2.10) motivates why the base width is kept thin: k increases for small W_B. Though the strongest dependence of β_F stems from the argument of the exponential, ΔE_g. In a classical BJT (homojunction transistor) the effective difference between bandgaps is approximately zero, but it can even be a small negative value [7]. Because of the heavy doping of the emitter, it has been proved both theoretically and experimentally [9] that the bandgap on the emitter side even shrinks according to the emitter doping concentration and the gain is reduced. HBTs can provide a positive bandgap difference, as we will discuss in the following.

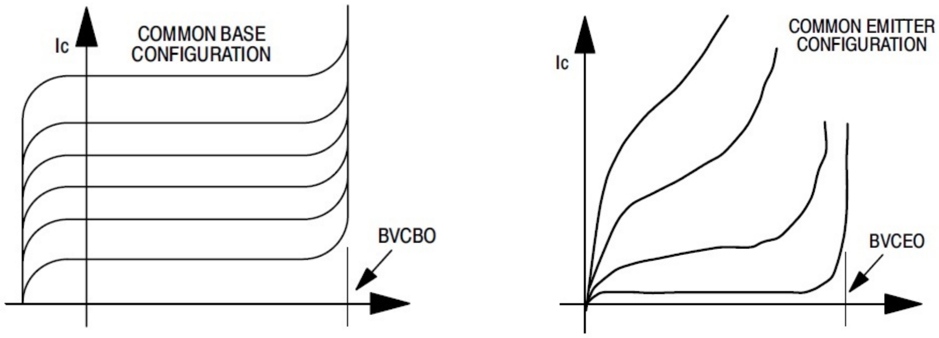
Breakdown Mechanisms

The breakdown is an uncontrolled rapid rise of the current through a junction when a voltage higher than a threshold value is applied as inverse bias. The net effect is to limit the maximum supply voltage we can use for a bipolar transistor, therefore having a huge impact on performance. There are two breakdown phenomena that usually occur in BJT junctions: *breakdown for tunnelling* and *impact ionization.*

They usually both happen at the two junctions of a transistor, but in different moments. So, while breakdown for tunnelling is the predominant mechanism at the EBJ, since the two edge-regions are highly doped and electrons might flow from one side to another if the depletion region is small by surpassing the energy barrier, impact ionization is predominant at the CBJ. Also, in active mode, only the CBJ is inverse biased while the EBJ is forward polarized, so this mechanism is the one which prevails.

When $V_{CB} > BV_{CBO}$, the electric field in the SCR between base and collector is so high that electrons are carried with such an energy to impact with the crystalline lattice. When one electron impacts, a electron-hole couple is created: the two electrons keep going to the collector, while the hole is swept to the base. The electrons may impact again with the lattice, thus provoking the avalanche effect.

In Fig. 2.4 is shown how breakdown manifests itself. In the common-base configuration (a) the breakdown occurs at a well-defined voltage and the current coming from the emitter has little effect on the breakdown, while in the common-emitter configuration (b) the breakdown voltage $BV_{CEO} < BV_{CBO}$ is not as sharp as before and the increase in hole current due to impact ionization reflects on an increase of the emitter current which finally



(a) Breakdown as seen in the common base configuration

(b) Breakdown as seen in the common emitter configuration

Figure 2.4: Breakdown in different configurations (after [10]).

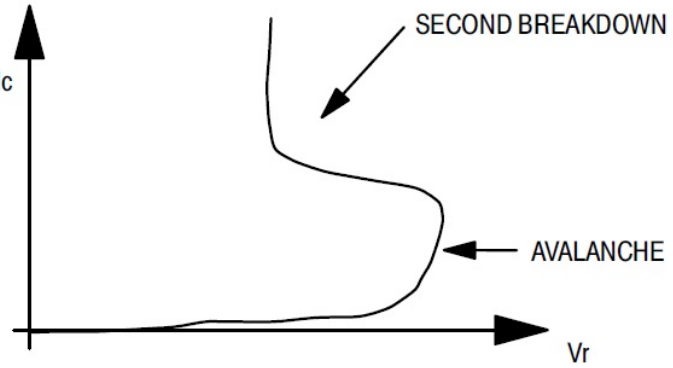


Figure 2.5: Typical second breakdown characteristic (after [10]).

yields to a runaway of I_C [10].

Especially in high current devices, once the first breakdown has happened as described so far, the junction is downgraded and if the current keeps rising. Electrons concentrate in particular spots, and a second breakdown may be activated due to local thermal heating, making silicon melt and destroying the transistor forever (Fig. 2.5).

These two mechanisms limit the transistor safe operating area (SOA) (darker line in Fig. 2.6) namely the values of the (V_{CE}, I_C) couples allowed in order to operate with the device. Two anticipating considerations can

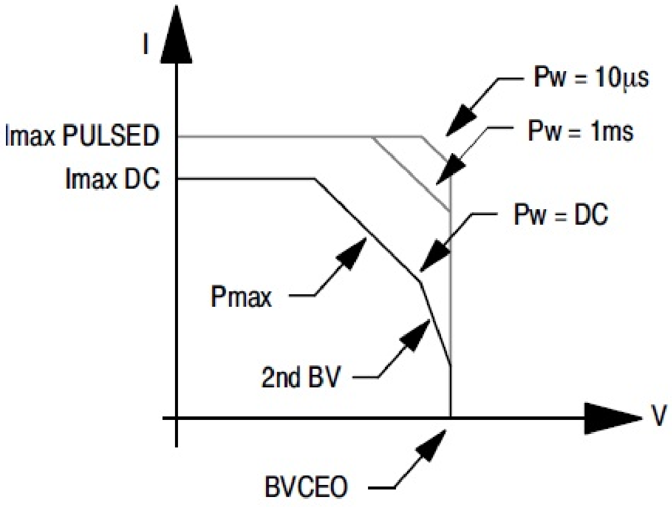


Figure 2.6: Typical safe operating area curve (after [10]).

be done. First, the SOA depends on geometry and fabrication processes. Processes that tend to boost the HF performances of HBTs also tend to decrease the area. Secondly, as it is clearly shown in Fig. 2.6, when biased by pulsed voltages, the device SOA extends.

2.2.1 Heterojunction Bipolar Transistor (HBT)

The idea of bandgap engineering dates back to the 1950s and was first conceived by Shockley and Kroemer [11]. They notices that a dramatic improvement of the current gain in bipolar devices could have been achieved by creating a much wider bandgap for the emitter compared to the base. However, the technology of heterojunction bipolar transistors (HBTs) was developed decades later and this delay was due to the struggle in dealing with contacting interfaces from different materials free of defects and imperfections, which usually come along with lattice mismatch.

In our study on BJTs, Eq. (2.10) showed how the current gain is affected by a possible bandgap difference. In heterojunctions, the change in energy gap is made of two steps that arise at the conduction band and valence band respectively, that is $\Delta E_g = \Delta E_c + \Delta E_v$ (Fig. 2.7).

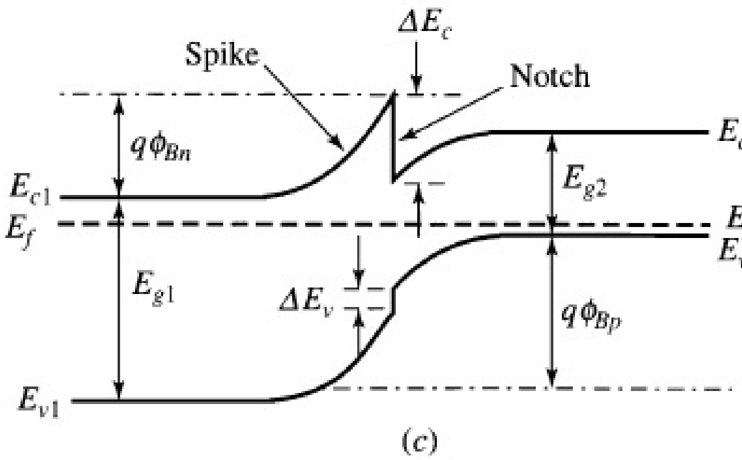


Figure 2.7: A spike appears in the emitter-base heterojunction when different materials are put into contact (after [12]).

By working on the association of materials in HBTs, ΔE_g is typically chosen to be greater than $250\,meV$ and β_F is 10^4 times greater than the homojunction case for the same doping profile. A higher level of doping of the emitter compared to the base is no longer necessary; instead, p_{Bo} can be risen up to $10^{20}\,cm^{-3}$, thus reducing the base resistance. High gain and narrow base are still maintained. Similarly, n_{Eo} can be reduced, thus increasing the SCR on the emitter side and equivalently reducing the emitter junction capacitance [7].

Putting in contact regions with different bandgaps gives rise to discontinuities between the conduction bands and the valence bands as shown in Fig. 2.7. We can see that a spike may appear in front of the path of the electrons flowing from emitter to base (for example in some III-V materials), eventually reducing the injection efficiency of the carriers (abrupt junction). Grading the junction over several hundreds angstroms may solve this issue.

Fig. 2.8 shows what happens if we extend the concept of heterojunctions to the CBJ too and we increase the collector bandgap. When the CBJ is forward biased (saturation mode), fewer holes flow from the base into the

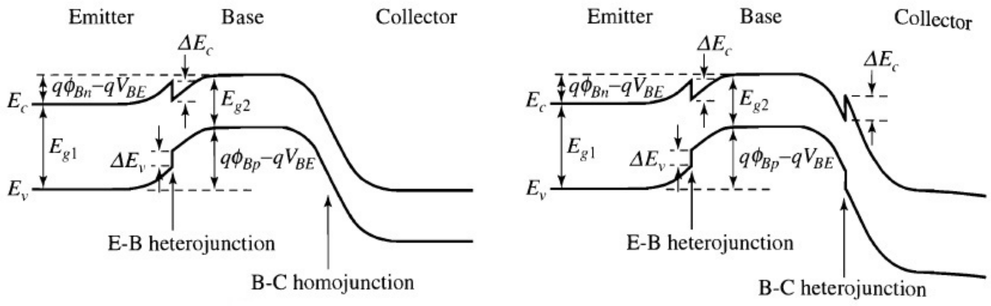


Figure 2.8: Single heterojunction transistor (at the left) and double heterojunction transistor (at the right) (after [12]).

collector and storage in the collector decreases, resulting in a quicker turn-off of the device. In the case of double heterojunction transistors, grading becomes mandatory in order not to suppress collector current [7].

High Frequency Performance

Two are the main figures of merit for high frequency transistor performance: *current-gain cutoff frequency* f_T and the *maximum frequency of oscillation* f_{max} [7].

The cutoff frequency is defined as the highest frequency at which the transistor current gain is equal to one, or more precisely, the frequency at which the incremental current gain equals one [7]. The incremental gain is not constant over frequencies, we can approximate its behaviour to a first-order system, so that, at high frequencies, from the definition of the cutoff frequency, $\left|di_C/di_B\right| \cdot \omega = 1 \cdot \omega_T$. From the well-known equivalence $\omega = 2\pi f$, we can rewrite as

$$\left|\frac{di_C}{di_B}\right| = \frac{2\pi f_T}{2\pi f} \Longrightarrow \left|\frac{di_C}{dq_B}\,\frac{1}{j2\pi f}\right| = \frac{f_T}{f} \tag{2.11}$$

where dq_B is the base charge associated with an increment of the input voltage and di_C is the collector current associated with an increment of the input voltage as well. The cut-off frequency f_T is linked to the emitter-to-collector transit time, τ_{EC} being

$$\tau_{EC} = \frac{1}{2\pi f_T} = \frac{dq_B}{di_C}$$

where the last equation follows from Eq. (2.11).

A more rigorous definition of the emitter-to-collector transit time states that it must be composed of the delays associated with the excess minority carrier charge in the neutral emitter, the emitter-base depletion region (very small and neglected), the neutral base, the collector-base depletion region and the delay to account for charging the base-collector capacitance, respectively

$$\tau_{EC} = \tau_E + \tau_B + \tau_{CBD} + \tau_{CBC}$$

τ_E is mainly associated with holes that are stored in the quasi-neutral emitter $Q_p = \int p(x)dx$. When designing HBTs, the emitter depth W_E should be as small as possible and the emitter doping as high as possible to maintain τ_E low.

The base transit time is defined as

$$\tau_B = \frac{Q_n}{I_C}$$

where Q_n is the total charge due to minorities (electrons) in the base. The total charge can be found thanks to the same linear approximation of the minorities profile in the base as Eq. (2.5). In fact, it is the area subtended by this profile. Hence, it is given that

$$Q_n = A_E \, q \, \frac{n_B(0) W_B}{2}$$

which yields

$$\tau_B = \frac{W_B^2}{2D_B} \tag{2.12}$$

Eq. (2.12) shows a second-order proportionality between τ_B and W_B, therefore explaining why it is essential for high-speed bipolar transistor to come with very small base widths.

τ_{CBD} is the time in which electrons travel across the collector-base SCR by drift. The electric field is very high so the electrons reach their saturated velocity almost immediately. An expression of proportionality for τ_{CBD} is

$$\tau_{CBD} \propto \frac{W_{CBD}}{v_{sat}}$$

where W_{CBD} is the collector-base depletion region width and v_{sat} is the electron saturation velocity. An increased base doping reduces the space-charge width thus improving τ_{CBD}.

τ_{CBC} is the RC time constant for charging the base-collector capacitance following to an incremental change of dv_{BE}, and it is found to be

$$\tau_{CBC} = C_{BC}\left(R_E + R_C + \frac{kT}{qI_C}\right)$$

where C_{BC} is the base-collector depletion capacitance and the resistances correspond to the emitter and collector. The last term is a dynamic resistance. This last term shows a dependence of f_T on the collector current: at low collector currents, f_T rises as I_C rises (or equivalently, τ_{CBC} falls). Then, τ_{CBC} becomes small compared to the other delays and reaches a plateau value

$$f_{TMAX} = \frac{1}{2\pi\left(\tau_E + \tau_B + \tau_{CBD}\right)}$$

At very high collector current, f_T rapidly falls due to high current effects (Kirk effect).

f_{max} is the frequency at which the maximum available power gain G_p of the transistor drops to 1, but can also be defined as the highest frequency an oscillator made with a particular device can achieve, thus explaining its name. For high frequencies, its dependence from frequency is of the kind of

$$G_p \simeq \left(\frac{f_{max}}{f}\right)^2$$

and for bipolar transistor we can approximate f_{max} by the following formula [8]

$$f_{max} = \left(\frac{f_T}{8\pi R_B C_{BC}}\right)^{1/2} \tag{2.13}$$

where R_B is the base resistance and C_{BC} the base-collector capacitance.

As we can clearly see from Eq. (2.13), the great decrease of the base resistance in HBTs mirrors on a higher f_{max}, which is made even bigger by the overall increase of f_T as well due to the decrease of τ_E and τ_{CBD} for these devices.

In order to reduce the transit time τ_B in the base too, a graded base can be made. The thickness of the base bandgap reduces itself gradually from emitter to collector producing an energy gradient. This gradient generates a quasi-electric field that drives electrons by drift as well as diffusion. A formula of τ_B that takes into account the base grading is

$$\tau_B = \frac{W_B^2}{\eta\, D_B}$$

with η being a correction parameter ($\eta = 2$, no grading; $\eta = 4$, very high-graded base).

In practice, parasitic impedances will affect performance by limiting the operation of analogue circuits and slowing the switching of digital circuits. One might try with thinning or doping the collector to obtain higher speed, but this comes with a trade-off on the breakdown voltage, that become unacceptably low for many applications.

The Kirk Effect

Let's focus on the intrinsic ideal bipolar device that we have discussed thus far. The analysis made is based upon the hypothesis of low injection which usually well approximates the behaviour of classical BJTs but no longer proves correct when devices like HBTs are more heavily biased in their active mode. High injection essentially describes a situation when the excess of carrier is comparable to the dopant concentration, for instance where the dopant atoms density is low. This happens mostly in the collector region of HBTs. The collector is lightly doped to reduce the breakdown voltage in the CBJ and the Early effect, thanks to the depletion region extending across the collector.

So, ideally the base minority carriers (electrons) are swept across the SCR

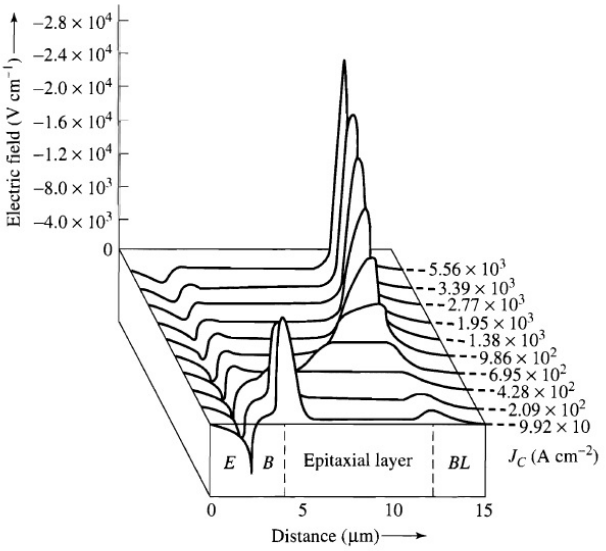


Figure 2.9: Negative electric field as a function of distance for various collector currents (after [8]).

toward the collector at infinite velocity but in a real situation and with high currents, a finite density of carriers appears in the SCR. The charge profile inside the SCR is modified because electrons sums up to the negative charge of fixed ions in the base side of the SCR and reduce the positive fixed charge in the collector side of the SCR, but since the voltage across the region is constant, so is the integral of the electric field: the depletion region width decreases at the base side, and increases at the collector side. So the short quasi-neutral base extends.

When the electron current rises any further, up to $N_{epi} = J_c/qv_s$ where v_s is the saturation velocity and N_{epi} is the dopant concentration in the lightly-doped collector epitaxial layer, charge neutrality imposes that a positive charge appears in the n^+-type subcollector. The electric field extends all the way across the epitaxial layer, because its integral is still to stay the same. If more current flows, the charge inside the depleted epi-layer will be overall negative and the field will rise in the proximity of the subcollector (Fig. 2.9).

The effective width of the base layer equals the width of the base and epi-layer, which increases the transit time substantially, thus both reducing the current gain and degrading the frequency response of the system – in brief, the cutoff frequency.

2.2.2 Si-based and III-V HBTs

Existing semiconductors can be divided into two categories: elemental and compound semiconductors. Among the first group are germanium and silicon. Because of its diamond-like lattice structure and some interesting properties, like high electron mobility, germanium was one of the first semiconductors applied for early transistors. Nevertheless, silicon has a bigger bandgap ($1.12\,eV$ versus $0.67\,eV$), showing less intrinsic-carriers growth with temperature, so it is nowadays mainly used in the semiconductor industry. But still, germanium has been widely applied in the last decades in compounds.

Compound semiconductors are materials exploited for bandgap engineering in devices like HBTs. Those can be alloys of elemental materials, like SiGe, or alloys formed by elements from two different groups of the periodic table, like III-V compounds (GaN, GaAs and InP).

SiGe HBTs

In a SiGe HBT, the base p-type semiconductor is composed of an alloy of silicon and germanium $\mathrm{Si}_{1-x}\mathrm{Ge}_x$, x being the atomic percentage of germanium in the alloy. The electron affinity of SiGe is similar to that of Si, so that the conduction band discontinuity is small, or equivalently we can substitute in Eq. (2.10) ΔE_g with ΔE_v, the difference between the levels of the valence bands. SiGe HBTs have achieved an average percentage of germanium of $x = 0.2$, resulting in $\Delta E_v = 200\,meV$ [7].

From the carriers point of view, the Fermi level in the p-type base is closer to vacuum in HBTs than it is in homojunction BJTs. This means that the displacement of conduction bands when the two regions are in contact is less for heterojunctions, hence the barrier for electrons is lower than in the EBJ of BJTs (Fig. 2.10) [8].

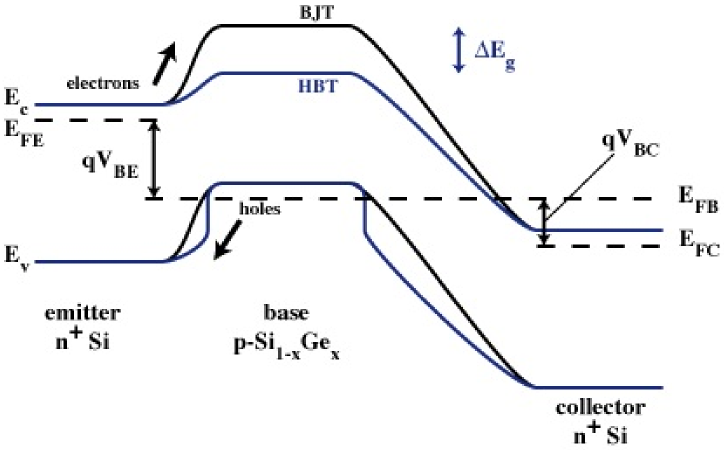


Figure 2.10: Different displacement of the conduction band in BJT and SiGe HBT (after [13]).

When the EBJ is created, dislocations may form because the lattice constant of the alloy is over 4% larger than that of Si, but as seen in [14], if a critical strained-layer thickness is respected by limiting the dose of germanium, the misfit adjusts itself elastically and no defects appear (pseudomorphic growth). The bandgap of SiGe is usually several tenths of an electron-volt smaller than that of Si, but pseudomorphic growth makes it even smaller. In fact, SiGe is now considerably strained and this has a beneficial effect on the transistor properties by further reducing the bandgap of SiGe.

If HBTs are built at low temperatures, dislocations due to incoherently strained layers tend to be less, using the same germanium fraction. This is why for the fabrication process, molecular beam epitaxy (MBE), which allows pseudomorphic growth at relatively low temperatures, is preferred. Other deposition methods are chemical vapour deposition (CVD) and rapid thermal chemical vapour deposition (RTCVD), which allows selective epitaxial growth (SEG) on patterned regions. In Fig. 2.11 an example of SiGe HBT structure is presented, the Silicon-On-Insulator technology being applied. SiGe HBT processes a huge advantage over the III-V HBTs in that the fabrication is more mature and SiGe HBT can be fabricated with the

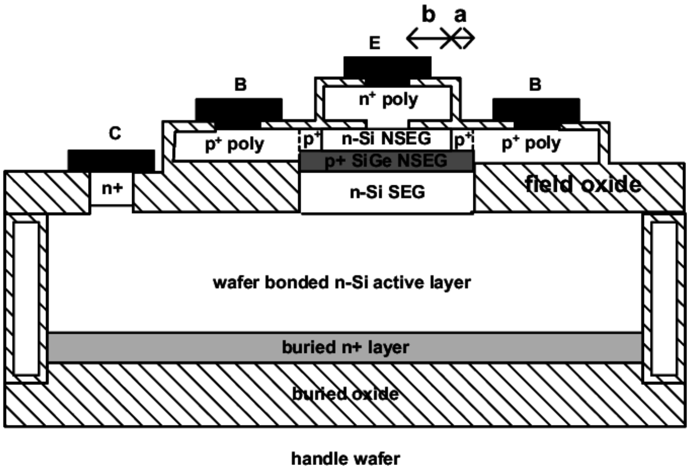


Figure 2.11: A UK consortium's platform SOI Si/SiGe/Si HBT device architecture (after [15]).

existing CMOS technology with only few more steps needed.

BiCMOS technology (Bipolar Complementary Metal Oxide Semiconductor) integrates in the same device both a classical CMOS technology and a bipolar transistor, taking advantage of the properties of the two technologies [16]:

- low power consumption (from CMOS)
- very good analogue amplifier (the CMOS gives high input impedance while the bipolar makes output impedance low)
- low variability in electrical parameters to temperature and process variations
- high current gain (from bipolar), making it suitable for long-lasting remote applications, for example
- higher packaging density for logic (from CMOS)
- in a series configuration, its total capacitance is low (almost as much as the bipolar), this leading to better frequency performance as a broadband amplifier or high switching speed in digital applications

- good fan-out, i.e. it can drive high capacitance load with reduced cycle time (no buffers needed, unlike CMOS)
- latch-up invulnerability

The main drawback is their high complexity in fabrication thus high costs, with respect to pure CMOS technology.

InP HBTs

InP-based HBTs use indium phosphide (InP) as main component and some materials which are lattice-matched to InP [7][12]. In the following, the emitter of our devices will be made of InP; the base will have therefore a shorter bandgap to exploit the properties of heterojunctions. Several ternary alloys can be used as the base, such as $In_{0.52}Al_{0.48}As$ (indium aluminium arsenide, InAlAs), $In_{0.53}Ga_{0.47}As$ (indium gallium arsenide, InGaAs), $GaAs_xSb_{1-x}$ (gallium arsenide antimony, GaAsSb). InGaAs and GaAsSb will be considered in this work. InGaAs has $E_g = 0.75\,eV$, GaAsSb has (on average) $E_g = 0.72\,eV$ [17], while InP has a bandgap $E_g = 1.35\,eV$. This remarkable bandgap difference provides an orders-of-magnitude higher current gain than any other solution.

The collector can be of the same material of the base (single heterojunction, SHBT) or another compound (double heterojunction, DHBT). SHBTs, like the one in Fig. 2.12, show higher C_{BC} and low breakdown voltage at high current densities, so to overcome that, the idea of a DHBT has been implemented.

InP-based HBTs exhibit many attractive properties. For instance, InGaAs has a 9-times-higher electron mobility than elemental Si, and 1.6-times-higher than GaAs, an alternative III-V compound for HBTs; in such a way, higher f_T values are reached. The InGaAs base has smaller bandgap in InP HBTs with respect to pure silicon and both SiGe and GaAs HBTs. Also, InP substrate provides twice the thermal conductivity of GaAs, allowing chips to operate at lower temperature by maintaining the same power dissipation,

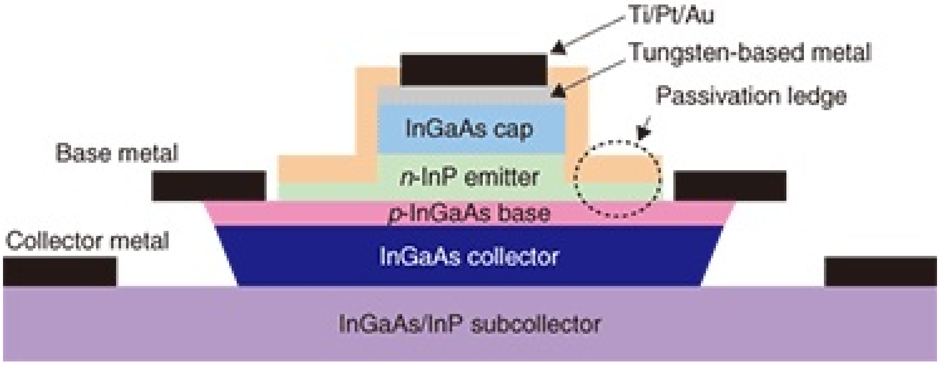


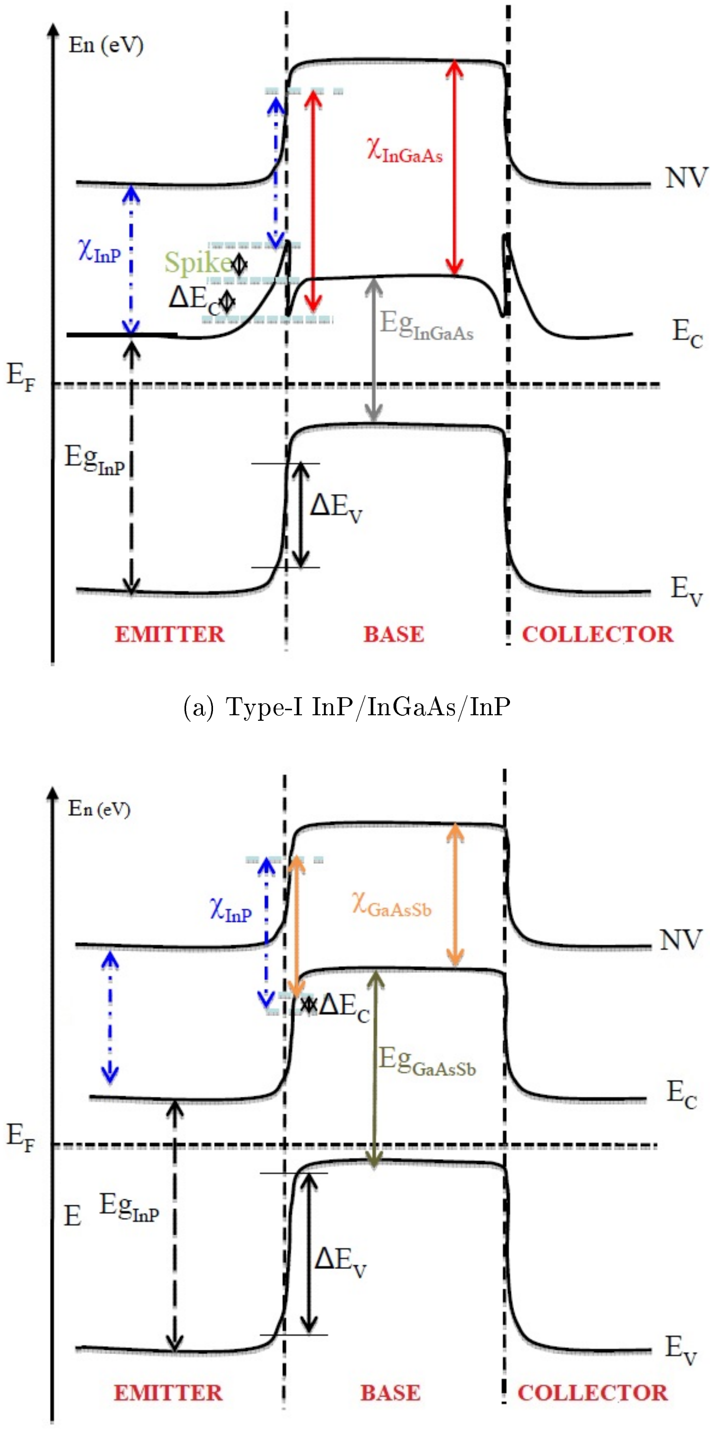
Figure 2.12: Cross-sectional view of an InP/InGaAs/InGaAs HBT (after [18]).

thus extending the device lifetime.

InP/InGaAs HBT provided for many years very good HF performance, but there is one fundamental drawback, that is the presence of a discontinuity (spike) along the path of electrons, both at the EBJ and at the CBJ in a DHBT at zero bias. This spike originates from the different electron affinity χ of InGaAs, $\chi_{InGaAs} > \chi_{InP}$, that leads to a negative discontinuity of the conduction band $\Delta E_c = q(\chi_{InP} - \chi_{InGaAs})$. This kind of heterojunction is called *type-I* [19]. EBJ grading can be done to take advantage of a smaller turn-on V_{BE} thus smaller power consumption; if the conduction step is kept, on the other hand, hot electrons can enter the base region with higher energies, thus reducing base transit time considerably.

Another type of HBT is the *type-II*, which is the case of the InP/GaAsSb heterojunction. The situation here is different since no spike is ever originated because $\chi_{GaAsSb} < \chi_{InP}$ and so $\Delta E_c > 0$. The drawback now is that the electron flux is not facilitated any more and f_T might fall. However, $\Delta E_v > \Delta E_c$ means that the holes are more bounded to the base region and consequently the current gain will rise as well as the breakdown voltage, while the turn-on voltage will drop. A pictorial comparison between the two HBT types and their band diagrams is made in Fig. 2.13.

Handling solid phosphorus is a hard task and this is the reason why InP is



(a) Type-I InP/InGaAs/InP

(b) Type-II InP/GaAsSb/InP

Figure 2.13: Band diagram comparison (after [19]).

difficult to grow with conventional MBE; gas-source MBE (GSMBE) provides a gas source for group V material and solves this issue.

2.2.3 Technology Under Analysis

In brief, their intrinsic characteristics ensure that III-V devices dominate over Si-based technologies because of the lower carrier mobility in silicon. They can reach higher frequency performance for a given BV_{CBO}. Nevertheless, advance in Si technology, growing demand for reliable RFICs, lower manufacture costs and much easier integration have made SiGe HBTs competitive in mmW applications.

STMicroelectronics's SiGe:C BiCMOS

In this work, the 55 nm SiGe:C BiCMOS (BiCMOS055, or B55 for short) by STMicroelectronics will be studied [20] [21]. ST's BiCMOS combines a CMOS and a NpN SiGe HBT, the architecture of which has been developed through generations from the early single-polysilicon quasi self-aligned architectures to modern double-poly fully self-aligned – which technology is used for the B55 too. Owing to their digital density 5 times higher than previous 130 nm technology, B55 well serves optical, wireless and high-performance analogue applications.

B55 is based on a 55 nm triple gate CMOS platform (55 nm being the average distance between identical features in an array of memory cells made with this technology), featuring both Low Power and General Purpose CMOS. This SiGe HBT has an extrinsic collector module (deep trench, sinker and buried layer) and it features the double-poly self-aligned architecture with selective epitaxial growth (DPSA-SEG) of the boron-doped SiGe:C base (C hinders boron diffusion) which provides the self-alignment of the emitter with the base (Fig. 2.14-a) and it is inserted in the fabrication process between gate polysilicon deposition and gate patterning of CMOS to reduce the total amount of thermal energy transferred to the CMOS during elevated temperature operations. The HBT (Fig. 2.14-b) comes in three collector flavours, with different $f_T \times BV_{CEO}$ trade-offs: High Speed (HS), Medium Voltage (MV) and High Voltage (HV). The back-end (made of 8 copper layers and an aluminium cap, Fig. 2.14-c) is fully compatible with CMOS and provides enhanced mmW performance [17].

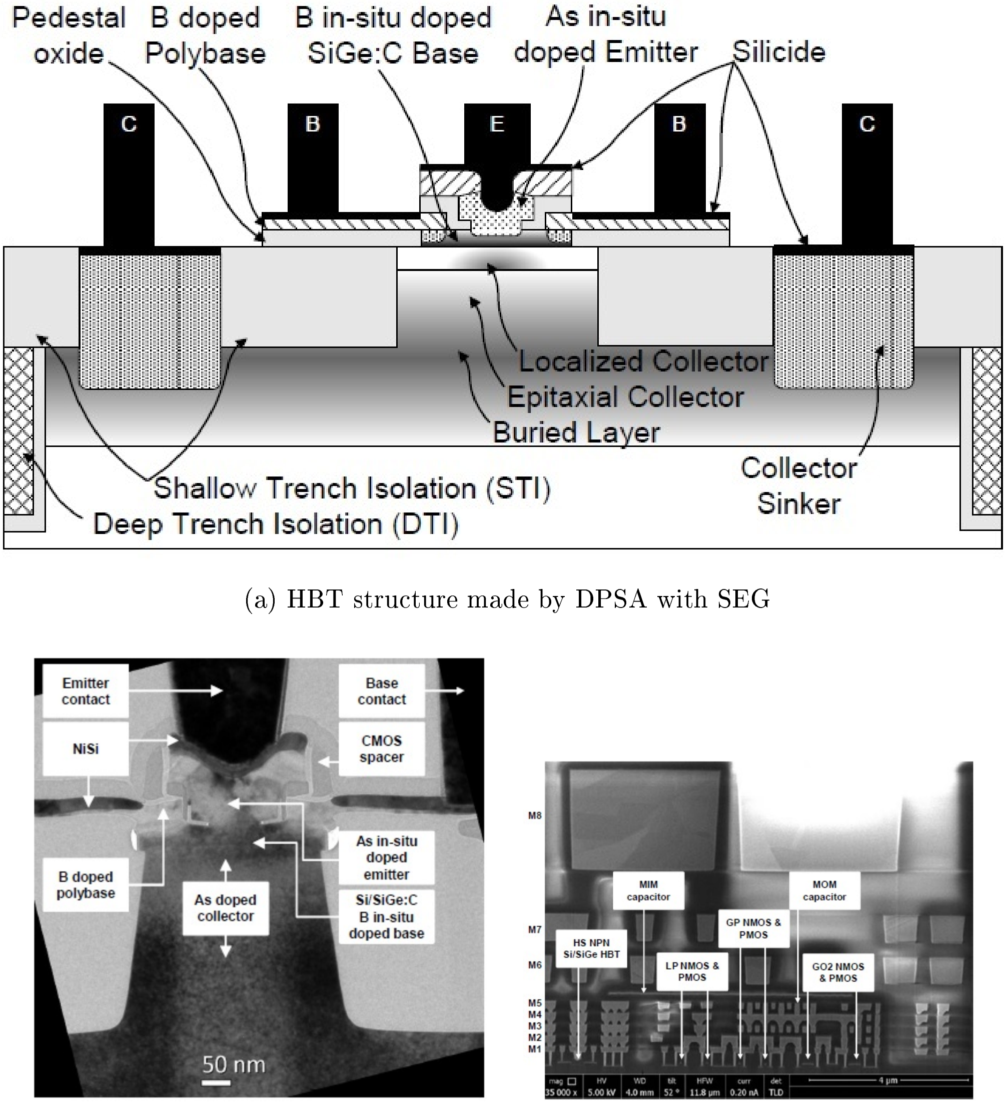
f_T and f_{max} rises are obtained thanks to the higher collector current densities and reduced parasitic resistances and capacitances allowed by the vertical and lateral scaling of transistors. In Fig. 2.15, we can observe such an improvement from older technologies notably at relatively high collector current density, with HF performance standing out for B55. Then the performance is degraded by heavy injection effects. By increasing f_T and f_{max}, the collector-emitter breakdown voltage consequent reduction points out the need for a trade-off. However, the $f_T \times BV_{CEO}$ of B55 is the highest of all ST's BiCMOS.

While the early $90\,nm$ BiCMOS technologies featured HBTs with HF performance that reached $(f_T, f_{max}) = (130, 100)\,GHz$ [22], measured results for the SiGe HS HBT in B55 prove a better HF behaviour, the couple of values being $(f_T, f_{max}) = (326, 376)\,GHz$. Measured HF figures of merit, collector current and breakdown voltages are all listed in Fig. 2.16.

In this thesis, different device geometries will be considered. The reference transistor is a HS $0.2 \cdot 5\,\mu m^2$ while several other combinations of emitter lengths/widths will be studied, such as $W_E = 0.3$, $0.45\,\mu m$ and $L_E = 1$, 3, $10\,\mu m$.

III-V Lab's InP/InGaAs DHBT

The type-I NpN double heterojunction bipolar transistor InP/InGaAs/InP by III-V Lab is the first of the two InP-based HBTs we consider [23]. Our measurements have been carried out on a wafer which represents an improvement on the technology applied by III-V Lab since their early InP/InGaAs HBTs. An In-rich InGaAs emitter cap which provides low contact resistivity, base growth optimization for static current gain increase and $35\,\%$ base resistance reduction are among these technological enhancements.



(a) HBT structure made by DPSA with SEG

(b) TEM cross-section of a $0.1 \times 4.9\,\mu m^2$ HS SiGe HBT (zoom in on Emitter and Base regions)

(c) SEM cross-section showing main devices and back-end up to Metal 8

Figure 2.14: BiCMOS055 by STMicroelectronics (after [20]).

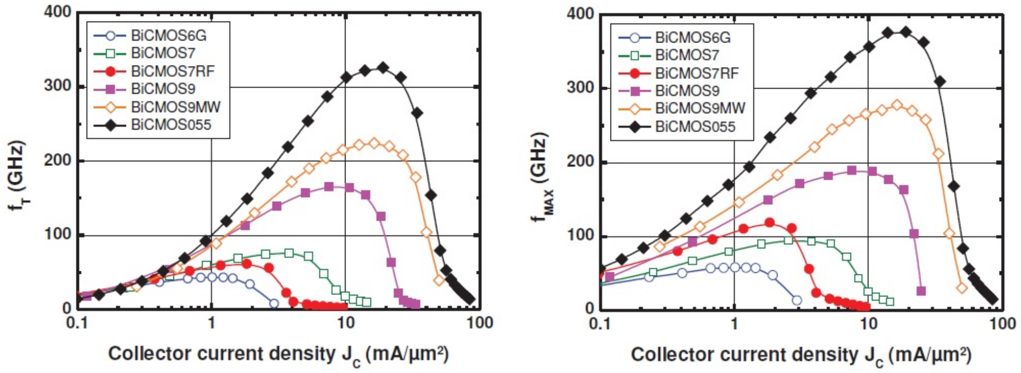
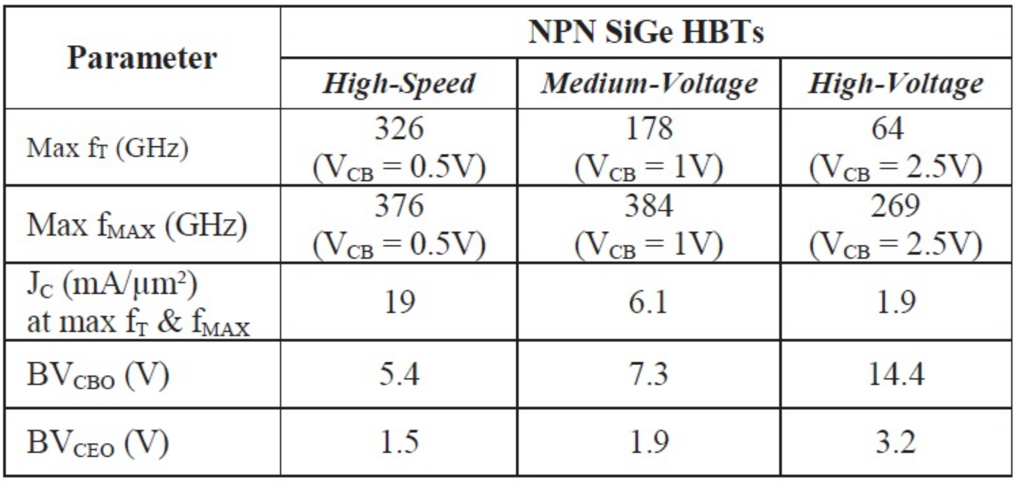


Figure 2.15: Evolution of f_T and f_{max} vs. collector current density in STMicroelectronics SiGe BiCMOS technologies (after [21]).



Parameter	**NPN SiGe HBTs**		
	High-Speed	***Medium-Voltage***	***High-Voltage***
Max f_T (GHz)	326 (V_{CB} = 0.5V)	178 (V_{CB} = 1V)	64 (V_{CB} = 2.5V)
Max f_{MAX} (GHz)	376 (V_{CB} = 0.5V)	384 (V_{CB} = 1V)	269 (V_{CB} = 2.5V)
J_C (mA/µm²) at max f_T & f_{MAX}	19	6.1	1.9
BV_{CBO} (V)	5.4	7.3	14.4
BV_{CEO} (V)	1.5	1.9	3.2

Figure 2.16: Measured results of NpN SiGe HBT available in B55 (after [20]).

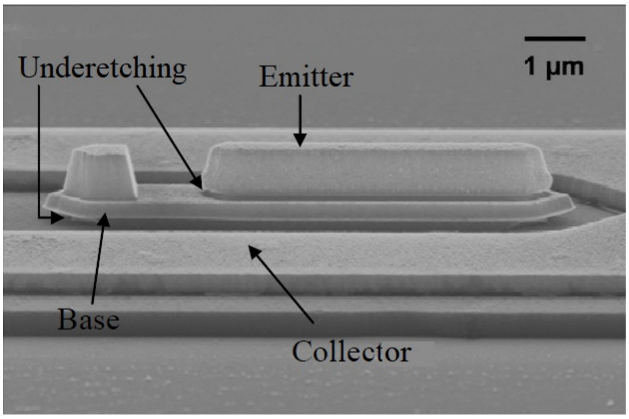


Figure 2.17: TEM view of an hexagonal III-V Lab HBT (modified after [19]).

This structure is composed by two heterojunctions, the emitter and the collector are made of InP, while the base is a lattice-matched InGaAs alloy. The base is highly C-doped. The layers beneath the emitter and collector ohmic contacts are highly doped as well. In order to get rid of the spike, many layers are inserted between the base and the extrinsic InGaAs sub-collector to create a grading. Collector 3 acts like a spacer, while collector 2 and 1 have gradually increasing doping levels. Fig. 2.18 represents the structure [19].

As we can see in Fig. 2.19 where all its performance parameters are listed, this device proves itself suitable for high speed digital and mixed circuits for $> 100\,Gb/s$ application.

In the following we will consider a reference transistor corresponding to an emitter area of $0.7 \cdot 5\,\mu m^2$. $A_E = 0.5 \cdot 7\,\mu m^2$ and $A_E = 1 \cdot 10\,\mu m^2$ will also be taken into account.

ETHZ's InP/GaAsSb DHBT

The InP/GaAsSb represents a III-V compound alternative to InP/InGaAs. The technology that will be used is from the Federal Institute of Technology (Eidgenössische Technische Hochschule, ETH) in Zurich. Unlike III-V Lab's HBTs, this double heterojunction bipolar transistor is of type-II which means

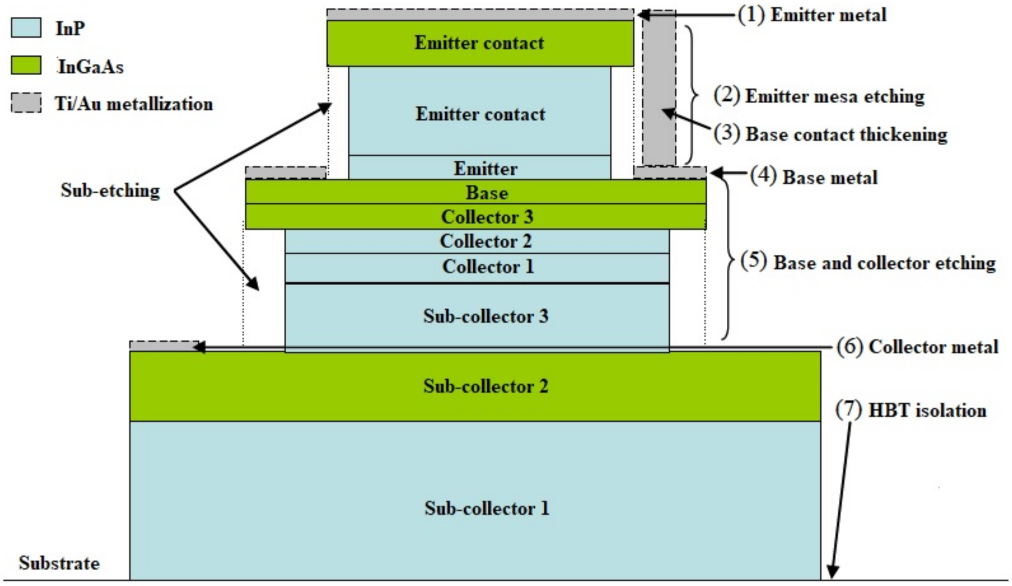
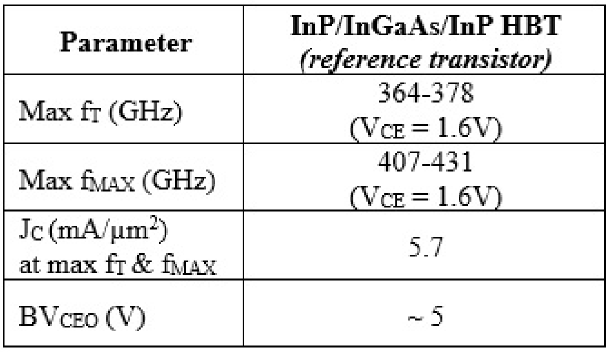


Figure 2.18: III-V Lab triple mesa structure (modified after [19]).



Parameter	InP/InGaAs/InP HBT *(reference transistor)*
Max f_T (GHz)	364-378 (V_{CE} = 1.6V)
Max f_{MAX} (GHz)	407-431 (V_{CE} = 1.6V)
J_C (mA/µm^2) at max f_T & f_{MAX}	5.7
BV_{CEO} (V)	~ 5

Figure 2.19: Measured results of NPN InP/InGaAs HBT III-V Lab reference transistor (data provided by [24]).

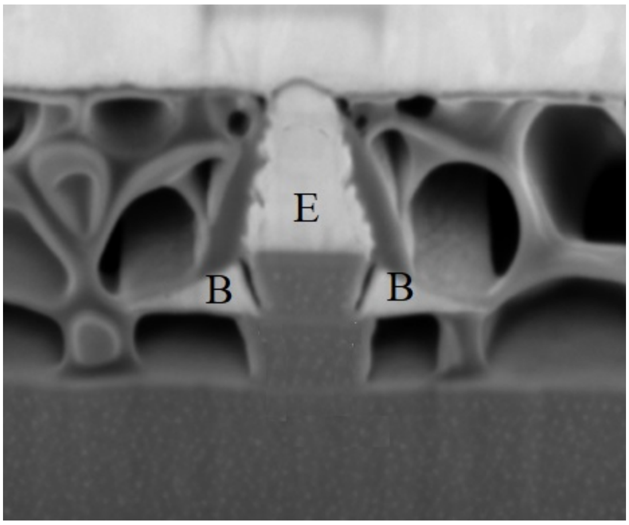


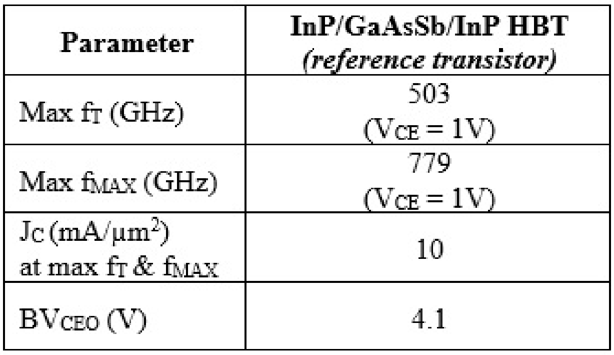
Figure 2.20: TEM image detail of the InP/GaAsSb DHBT (after [25]).

that the abrupt junctions can be kept since no obstacle is formed by the difference of the electron affinities of the materials (Fig. 2.20) [17] [25].

Also, the growth process is different: the epitaxial layers are grown by metal-organic chemical vapour deposition (MOCVD) that, in contrast to MBE, employs chemical reactions instead of a physical deposition. The $\mathrm{GaAs}_x\mathrm{Sb}_{1-x}$ base is highly C-doped and it is graded from $x = 0.41$ at the collector side to $x = 0.61$ at the emitter side. This is done to create a quasi-electric field (the drop is around $50\,meV$) to speed up electrons that otherwise would take longer time to pass, owing to the low electron mobility of GaAsSb. The structure is the usual self-aligned triple mesa.

In Fig. 2.21 all the main AC/DC parameters are listed. To date, this is the highest reported f_{max} for a GaAsSb DHBT [25]. This has been achieved thanks to a reduction of the base access distance that has beneficial consequences both on the access base and base-collector capacitance. Such performances are especially attractive for the implementation of submillimetre-wave integrated circuits.

In the following, only one geometry ($A_E = 0.2 \cdot 4.4\,\mu m^2$) will be analysed, but taking into consideration also some other versions which have a different



Parameter	**InP/GaAsSb/InP HBT** ***(reference transistor)***
Max f_T (GHz)	503 (V_{CE} = 1V)
Max f_{MAX} (GHz)	779 (V_{CE} = 1V)
J_C (mA/µm^2) at max f_T & f_{MAX}	10
BV_{CEO} (V)	4.1

Figure 2.21: Measured results of NPN InP/GaAsSb HBT ETHZ reference transistor (data provided by [25]).

technological dispersion (different process, doping profiles, etc. . .).

2.3 The Self-Heating

As seen, to exploit the highest cut-off frequencies, thus the highest performances, an HBT will experience intense current densities and electric fields inside of it – in the base-collector depletion region particularly.

These high levels of power are dissipated inside the device, the geometry of which is shrunk compared to older technologies, resulting in even greater power densities. High power dissipation leads to higher temperatures at the chip level, a process called *self-heating*, which has negative consequences on reliability and reduces electron mobility, resulting in poor performance and degradation of the electrical characteristics of these devices [26].

2.3.1 Heat Generation Effects

Heat is generated by a heat source – we can assume that it is placed at the CBJ – and is transmitted through solids – the different layers – by conduction. The basic assumption is that, due to the dimensions of the source, the effect of all the boundaries are neglected, but the top surface, which is considered adiabatic: there will be no conduction through the metalization and no convection from the surface [27]. The heat generated by the self-heating of the device flows from the heat source to the backside of the substrate, which is considered to be controlled to the external environment at a substrate temperature T_{sub} [28]. Heat, flowing through materials, produces a temperature drop, that, as a first approximation, can be proportional to the heat flux. The proportionality factor is the thermal resistance R_{TH}.

The analogy with an electrical circuit is straightforward:

- heat (measured in $[J]$) is analogous to the elementary charge,
- the heat flux (measured in $[W] = [J/s]$) is analogous to current,
- temperature (measured in $[K]$) is analogous to voltage,
- the thermal resistance (measured in $[K/W]$) is analogous to the ohmic resistance,
- the thermal capacitance (measured in $[Ws/K]$) is analogous to the electrical capacitance.

In transient conditions, during rapid variations of the dissipated power like pulsed power dissipations, the temperature at the junction does not reach its final value immediately. Any material must absorb a given heat quantity in order to increase its temperature: to take this behaviour into account the thermal capacitance C_{TH} has been introduced.

Unfortunately, some technological improvement like the use of shallow trenches and deep trenches, which have the beneficial consequence of reducing parasitic electrical capacitances and therefore of decreasing the transit time of electrons, tend to confine the heat flux because of the lower thermal conductivity of silicon dioxide, and the thermal impedance (Z_{TH}, in analogy to the electrical impedence) rises [28]. In Fig. 2.22, a TCAD simulation of the temperature profile of a B55 is shown, considering the CBJ as the planar heat source. It is clear by the figure not only that deep trenches confine heat, but also that heat propagates differently in different materials, according to the thermal conductivity of components.

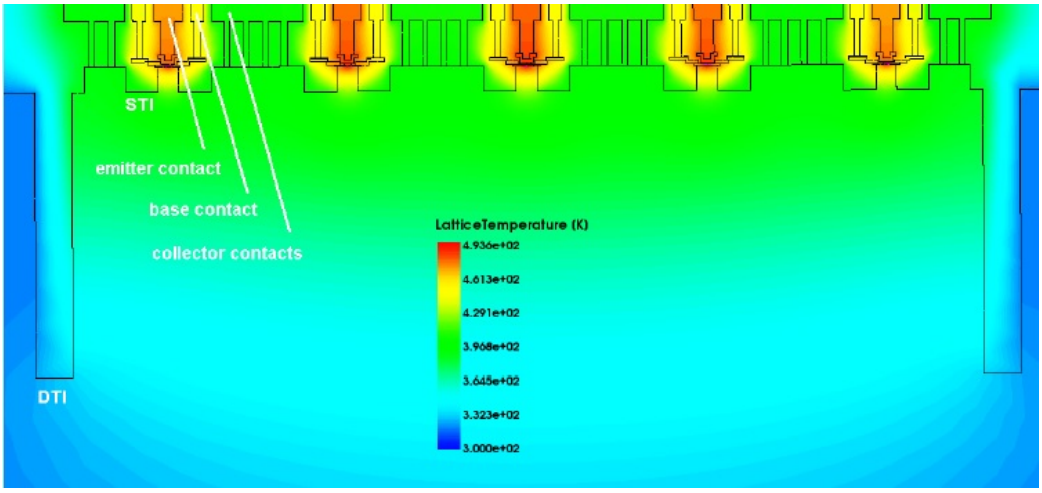


Figure 2.22: Temperature profile simulated in Sentaurus TCAD for a five fingers SiGe HBT realized in ST B55 technology (after [29]).

In certain circumstances, even a positive thermal feedback (*thermal runaway*) may take place. In the case an HBT is biased by a base-emitter voltage, the self-heating led by high power dissipation increases the collector current. To understand this, it is convenient to consider Eq. (2.14) and describe the collector current in the following terms [30]

$$I_C = I_s \left[\exp\left(\frac{V_{BE} + \phi \Delta T}{V_T} \right) - 1 \right] \tag{2.14}$$

where $\phi = -\partial V_{BE}/\partial T|_{I_C,V_{CE}}$ is considered equal to a constant value of $2\,mV/°C$. The dissipated power inside the device is

$$P_d = I_B\,V_{BE} + I_C\,V_{CE} \simeq I_C\,V_{CE} \tag{2.15}$$

So the increase of ΔT as a consequence of self-heating makes I_C rise and this will finally result in a further P_d increase, in a positive-feedback mechanism. In this manner, the $I_C - V_{CE}$ curve might reach a turning (or *flyback*) point which limits the SOA. This risk of instability is completely avoided by biasing an HBT by a constant base current [31].

Let's go back to better explain how temperature changes over time as heat spreads throughout space. To do so, we must consider the three-dimensional heat equation

$$\frac{\partial T}{\partial t} = \frac{\kappa}{c\rho}\nabla^2 T$$

where $T = T(x, y, z, t)$ is the temperature as a function of time and space, κ is the thermal conductivity – in general, a function of temperature –, ρ is the density of the material and c is the specific heat capacitance. By this equation, giving proper boundary conditions, T can be found as a function of time (t) and space ($\vec{x} = (x, y, z)$). The thermal resistance, which is a function of position, is defined as the ratio of the final value of the temperature increase to the dissipated power, namely [27]

$$R_{TH}(\vec{x}) = \frac{T(\vec{x}, \infty) - T(\vec{x}, 0)}{P_d} = \frac{\Delta T(\vec{x})}{P_d} \tag{2.16}$$

It is conventional to consider R_{TH} in the region where the temperature variation is maximum (from ambient temperature T_{amb}, i.e., no power dissipation, to junction temperature T_j, i.e., the maximum experienced temperature rise). The previous equation becomes

$$R_{TH} = \frac{T_j - T_{amb}}{P_d} \tag{2.17}$$

where P_d has been defined in Eq. (2.15).

2.3.2 Heat Flux Modelling

As discussed before, the heat flux from the source to the bulk can be analysed with an analogous electrical circuit. However, heat diffusion phenomena from the heat source to the heat sink show a distributed behaviour and different time constants are necessary to describe this physical phenomenon [29]. Therefore, a propagation line better describes the heat diffusion, with distributed resistances and capacities. Each RC block corresponds to a volume element in which heat is generated. Thus, each volume element is associated to its corresponding lumped R_{TH} and C_{TH} (see Fig. 2.23). The following assumptions are made [28]:

- each layer's thickness is chosen so that the thermal time constant ($\tau_{THi} = R_{THi} C_{THi}$) values are progressively growing,

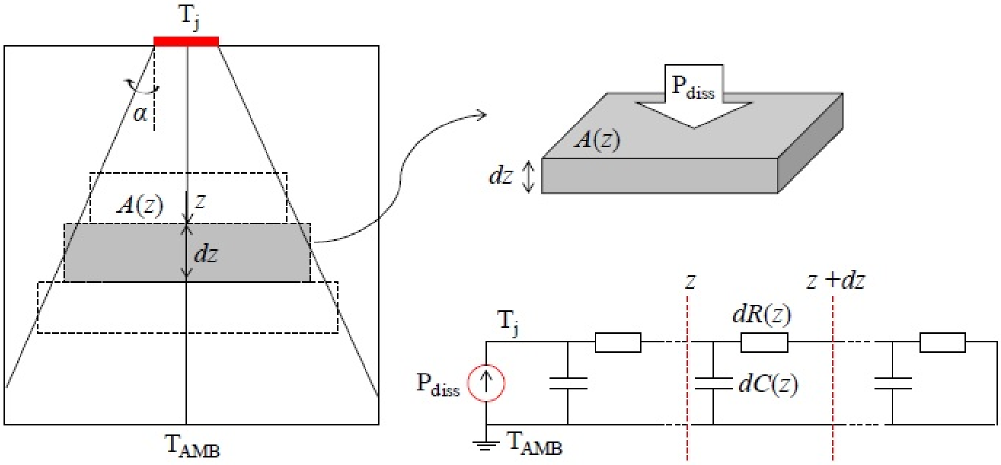


Figure 2.23: Heat diffusion within spreading angle α and its simplified electrical transmission line equivalent circuit representation (after [28]).

- α is assumed to be the spreading angle of heat across its substrate.

As seen in [29], thermal TCAD simulations help to identify the different volumes which are delimited by isothermal contours. In Fig. 2.24 we can see a zoom on one of the five fingers of a B55 device. Considering the heat flowing in a one-dimensional system (along z), the Fourier's law yields

$$\Delta R_{TH}(z) = \frac{1}{\kappa\, A(z)}\Delta z, \quad \Delta C_{TH}(z) = \rho c\, A(z)\, \Delta z \tag{2.18}$$

This means that the values of ΔR_{TH} and ΔC_{TH} to insert into the distributed system can be determined directly by the physical structure. Also, from Fig. 2.24 we can also affirm that the thermal resistance in the region marked by 1 will be greater than that of region 2, because the volume of region 1 is smaller. The opposite happens with the thermal capacitance. In the equivalent circuit, this means that resistances closer to the thermal "power generator" will be greater than the furthest, while the capacitance closer to the generator will be the smallest.

In many compact models that determine the electrical behaviour of components (like in HICUM, for instance), a very basic RC electro-thermal circuit

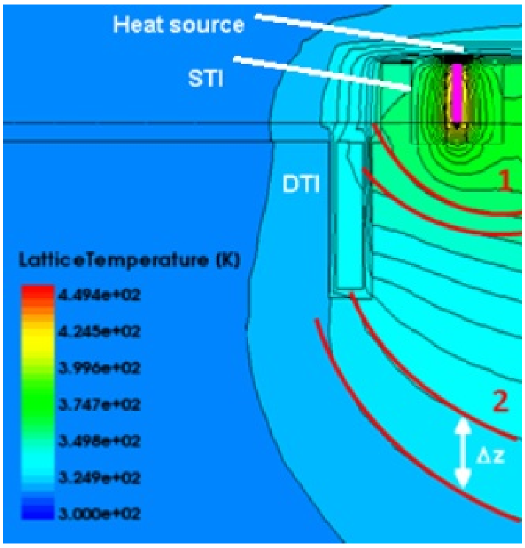


Figure 2.24: Thermal TCAD simulations for a five fingers HBT in ST B55 technology (after [29]).

is used to take into account self-heating (single-pole, see Fig. 2.25). Often, though, this characterisation is not sufficient. A distributed thermal network like the one above is implemented in Verilog-A and used for accurate simulations of the junction temperature rise, with thermal parameters properly extracted by the device's characteristics [28]. In the next, we will analyse these phenomena with a single-pole analogous circuit only. Also, in Chapter 4, the validation of the extraction of R_{TH} will be accomplished by using this simple network.

So, the thermal impedance Z_{TH} in Laplace domain simply is found to be

$$Z_{TH}(s) = \frac{R_{TH}}{1 + s\, R_{TH} C_{TH}} \tag{2.19}$$

while the temperature rise follows an exponential law dependant on the instantaneous power dissipation

$$T_{rise}(t) = P_d^{RISE}\, R_{TH} \left[1 - \exp\left(-\frac{t}{\tau_{TH}}\right)\right] \tag{2.20}$$

where P_d^{RISE} is the difference between the power dissipation in the DC condition and the isothermal condition, and $\tau_{TH} = R_{TH} C_{TH}$ is the thermal time constant.

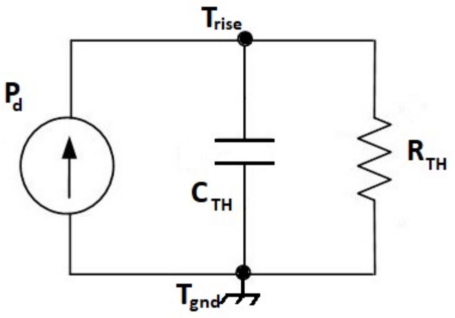


Figure 2.25: A single-pole electro-thermal network used in many simulation models.

2.3.3 Pulsed Measurements

Since temperature has a strong impact on the physics of the semiconductor, the consequences manifest themselves on the measured DC and AC characteristics, which change at different bias. Nevertheless, a transistor model accuracy in the whole domain of operation is an absolutely necessary key point in order to achieve successful RF designs [32].

The strong dependence of the electrical behaviour of the transistor on temperature has repercussions on the extraction of many parameters which are critical for measurement-based models for microwave CAD, such as access resistances, transit times and saturation currents [33] [34].

In particular, for a specific $I - V$ (current vs. voltage) electrical characteristic, the higher the current and/or voltage value is, the more power is dissipated, thus the more heat is self-generated within the device. This temperature change makes the calculation of the parameters difficult since, by definition, it should be computed in isothermal conditions.

Isothermal measurements are therefore an essential requirement to characterize such high-dissipating devices. These data allow to exclude – almost – entirely the effect of self-heating from the extracted parameters, thus showing just the effects of other phenomena such as the Kirk effect.

Several methods can be used to retrieve isothermal characteristics; one of them is to perform measurement in pulsed conditions [35]. The goal is to exploit the thermal capacitance C_{TH} of the substrate to delay the heating enough. In other words, the thermal time constants should be large compared to the pulse width (t_w).

They are short pulses that better remove self-heating effects. According to the device under test, though, some trade-offs must be considered when reducing t_w. Too short pulse widths degrade the quality of measurement, and speaking of duty cycle ($D = t_w/T$, where T is the pulse period), it must be short enough to permit the cooling of the device: in the following, D will be chosen of 10 %.

So, because of the small percentage of application time of pulse over the total period, the stress on the device is lower. As already mentioned before and represented in Fig. 2.6, pulsed measurements extend the safe operating area without causing further damage to the transistor, and the study of high current and breakdown phenomena is therefore possible.

In the next chapter we will explore the results of pulsed-DC measurements on the devices we have introduced in this section, and by comparing their characteristics it will be possible to understand to which extent self-heating affects them.

Chapter 3

Pulsed-DC Measurements

The use of a pulsed measurement system allows to get data in (theoretically) isothermal conditions, that is to reduce the thermal contribution from the electrical behaviour. The thermal phenomenon of self-heating is just one of the many thermal phenomena that, along with electron trapping and impact ionization, cause the dispersion of the characteristic curves. A bias condition that avoids all these effects is called *isodynamic* and the isothermal requirement is a necessary condition for that.

To produce our pulsed-DC measurements we use a set-up in which, while the collector terminal is biased by a DC voltage, the base is biased every period T by a short perturbation – the pulse stimulus –, followed by a longer period when the terminal is unbiased to cool down. The stimulus must be shorter than the associated thermal impedance, which means in the nanosecond range, in order for the internal temperature of the device not to change significantly.

This is technically difficult to achieve and requires a specific costly instrumentation set, which assures the precision and reliability of this technique.

Figure 3.1: Keithley Model 4200-SCS.

3.1 Performing the Measurements

The pulsed-current/voltage (*pulsed-I/V*) characteristic curves are curves determined by short pulses separated by long relaxation periods that allow less strain. In general, short pulses produce quasi-isothermal pulsed-I/V characteristics while very long pulses produce quasi-DC characteristics [36].

Keithley Model 4200 Semiconductor Characterization System (4200-SCS) (Fig. 3.1) is an automated system that provides I/V and pulsed-I/V characterisation of semiconductor devices and test structures. Pulse source and measure tests are provided by the Model 4225-PMU Ultra-Fast IV pulse-measure card. When a synchronisation with a RF pulse during a pulsed measurement is needed, the Model 4220-PGU pulse-generating card is used, as we will see in Chapter 5 (Fig. 3.2).

In Fig. 3.3 a block system of the instrumentation is shown. As we can see, Channel 1 of the PMU is connected through a ground-signal-ground (GSG) pad to the base terminal of the transistor (gate, in MOSFETs) and Channel 2 is connected to the collector (drain). The emitter (source) terminal is grounded. An internal voltmeter-ammeter couple measures the electrical

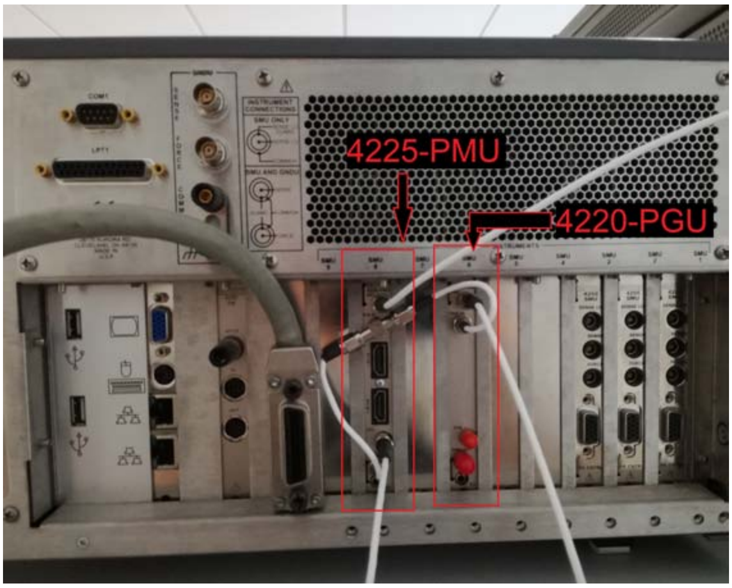


Figure 3.2: Backside of Keithley 4200-SCS pulsed measurement system where the 4225-PMU and 4220-PGU modules are highlighted.

quantities, while the pulse is generated by a generator with a 50 Ω internal impedance. Although many tests are allowed by the model card, in this work only DC-V_C-Sweep/Pulsed-V_B-Step will be performed: while a pulse is applied at Channel 1, a DC sweep is carried out at Channel 2, according to the user's specifications.

The PMU has been used for two types of fast I/V tests: pulsed-I/V and transient-I/V. The pulsed-I/V test is a time-based measurement that provides a DC-like characteristic. Through the IC-CAP interface, the user can define the bias condition – both for the base and collector terminals – and the base pulse form and characteristics. It is therefore necessary to define the starting, stepping and stopping points for the collector DC-sweep and the base pulsed-step, as well as the pulse low level. The pulse low level represents the constant quiescent condition for the base terminal, that in the following we always establish to zero. Please note that in this way the pulse amplitude, which is defined as the difference between the high and low levels of the pulse, corresponds to the high level.

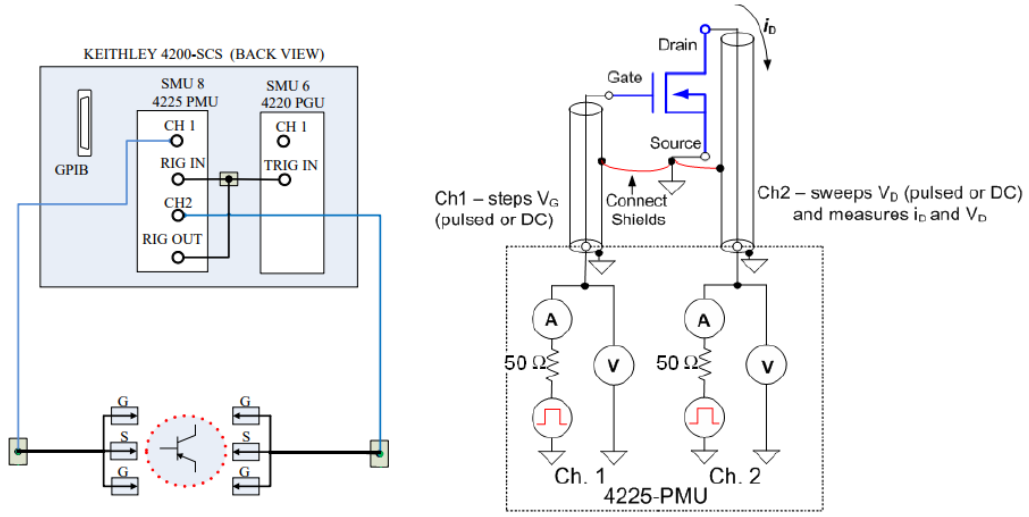


Figure 3.3: Block diagram of Keithley 4200-SCS pulsed measurement system with 4225-PMU and 4220-PGU modules (left) (modified from [28]) and equivalent circuit showing the connections to a three-terminal device (right) (modified from [37]).

Then, all the pulse form parameters need to be defined, i.e.:

- the pulse width (t_w), the interval of time from 50% of the rise time to 50% of the fall time,
- the pulse rise (fall) time (t_R and t_F respectively), the period spent by the pulse to go from 10% (90%) to 90% (10%) of the high value – we chose a fix value, namely the minimum reachable, $20\,ns$,
- the pulse delay time, from the begin of the period (namely from the pulse trigger),
- the total period (T), properly chosen as $T = 10\,t_w$,
- start and stop measurement locations, namely the time interval percentage during which the measurement is carried out. The start and stop location of the measurement is calculated with respect to t_{on}

$$t_{on} = t_w - \frac{1}{2}(t_R + t_F),$$

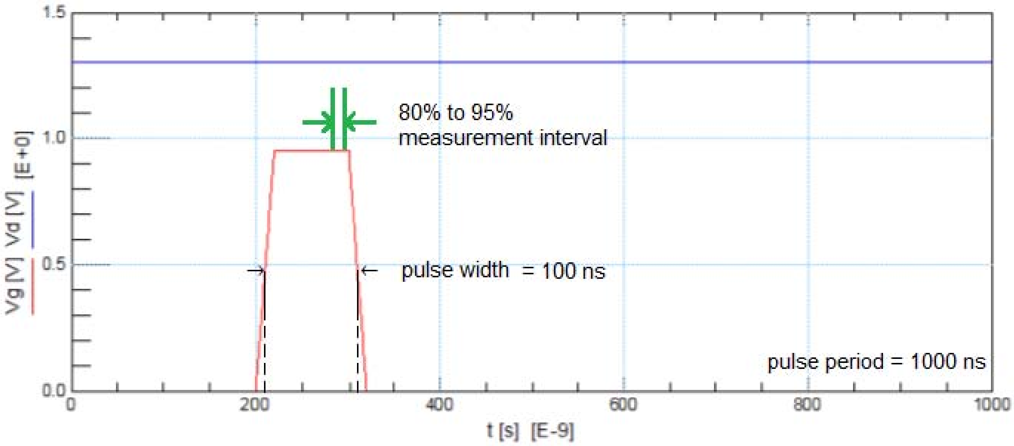


Figure 3.4: Base (V_g in red) and collector (V_d in blue) voltages as a function of time. The measurement interval is highlighted in green.

- the pulse average, the number of pulses that is averaged to get the final value.

It is useful to provide a clarifying example. In Fig. 3.4 the base and collector voltages are shown as a function of time. The settings are as follows:

- $V_{B_{START}} = 0.7\,V$; $V_{B_{STOP}} = 0.95\,V$; $V_{B_{STEP}} = 0.05\,V$; $V_{B_{BASE}} = 0\,V$.
- $V_{C_{START}} = 0\,V$; $V_{C_{STOP}} = 1.3\,V$; $V_{C_{STEP}} = 0.1\,V$.
- $t_w = 100\,ns$; $t_R = t_F = 20\,ns$; $t_{DELAY} = 200\,ns$; $T = 1\,\mu s$; start measurement location = 80%; stop measurement location = 95%; pulse average = 10.

The figure displays in the same plot the final values of both voltages ($V_{B_{STOP}} = 0.95\,V$ and $V_{C_{STOP}} = 1.3\,V$). Also shown are the selected pulse width and period, as well as the measure window. During this interval (in green in the figure) the currents and voltages are measured and the samples are averaged to yield unique current and voltage values (the sample rate is $200 \cdot 10^6\,s^{-1}$, but the limit on the number of samples is 10^6). When a "pulse average" is specified (in this case this number is 10), a burst of pulses is

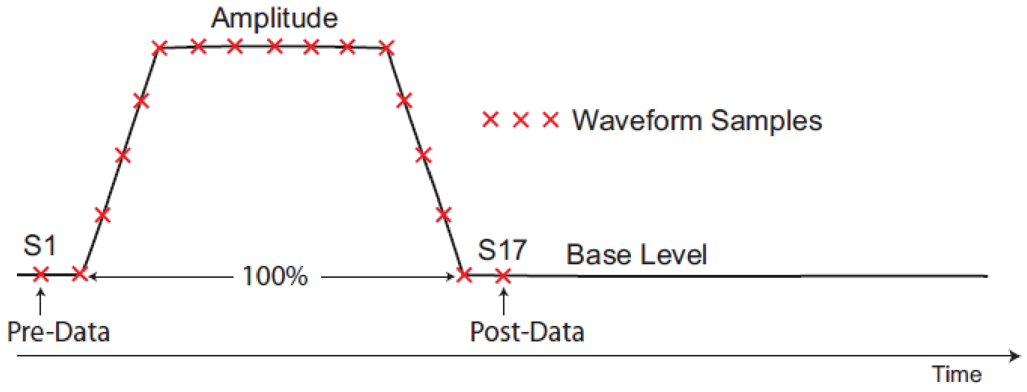


Figure 3.5: Example of a waveform capture of a single pulse (after [37]).

generated and the final current and voltage values are the "mean of means" of the samples. This measure is called *spot mean measurement*.

The transient-I/V measurement (also called *waveform capture*) is a dynamic representation of the electrical quantities. This is performed by the internal 4200-SCS software KITE (Keithley Interactive Test Environment) by using some Interactive Test Modules (ITMs). The result is a number of waveforms as a function of time, where we can observe the evolution of currents and voltages at the channels at each moment of time. The sample rate is the same and the measurement mechanism too: an average is made on the sample means of every pulse in a burst of pulses. Pre-data and post-data are collected as well (Fig. 3.5).

3.2 Measurement Results

Some measures have been carried out on the devices we have previously introduced – the HBT technologies from STMicroelectronics, III-V Lab and ETH Zurich. We will refer to them as B55, III-V Lab and ETHZ respectively. The measures are both pulsed-I/V and transient-I/V and they permit considerations on the devices themselves and specifically on self-heating.

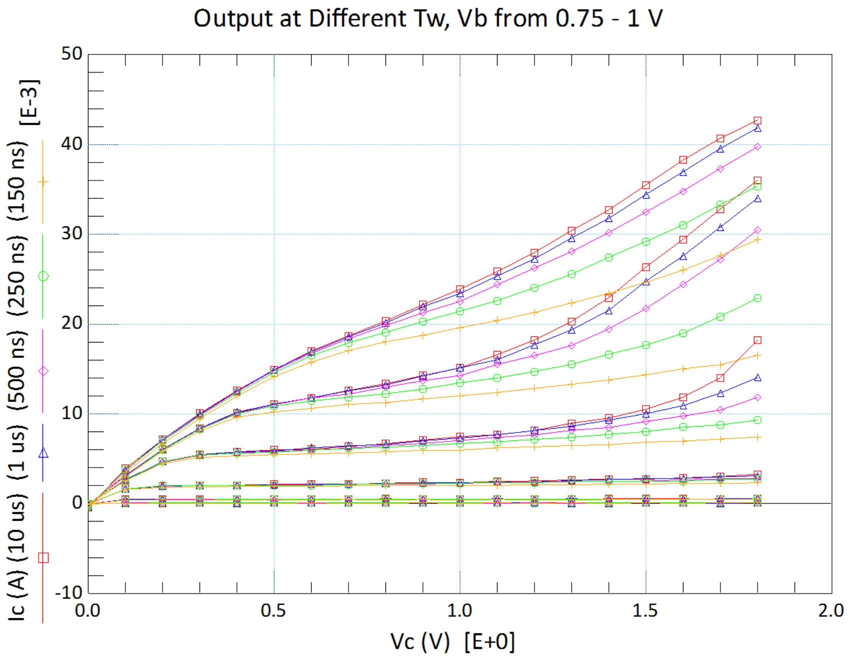


Figure 3.6: $I_C - V_{CE}$ characteristic; sweeping V_{CE} every $100\,mV$, stepping V_{BE} every $50\,mV$. B55: emitter area $A_E = 0.2 \cdot 5\,\mu m^2$.

3.2.1 Pulsed-I/V

Characteristics of the Reference B55 Device

In Fig. 3.6 a standard pulsed-I/V measure is shown for a B55 device. I_C vs. V_{CE} curves are obtained through a DC-V_C-sweep/pulsed-V_B-step measurement. Each colour represents a different pulse width, the shortest being $t_w = 150\,ns$, the longest $t_w = 10\,\mu s$. V_{BE} is stepped from 0.75 to $1\,V$ each $50\,mV$.

From a simple inspection of this graph, some remarks can be done straight away. For each V_{BE} value, all curves superpose at low V_{CE} (saturation region) but then separate as the DC voltage rises. The difference is directly linked to self-heating, whose effect is preponderant when the pulse is so long that the dissipated DC power in the device is as much as in the case of a non-pulsed DC voltage at the base terminal.

The yellow curves (extracted at $t_w = 150\,ns$) appear to be nearly horizontal because the effect of self-heating is almost excluded (quasi-isothermal condition), while the red curves (extracted at $t_w = 10\,\mu s$) diverge from the classical $I_C - V_{CE}$ plot appearance, providing values of the collector current which are altered by heat almost as in a DC condition (quasi-DC condition). This modification of the collector current is so strong that for the highest swept collector voltage, $V_{CE} = 1.8\,V$, pulses greater than $500\,ns$ and amplitude $V_{BE} = 0.95\,V$ alter the collector current so much that it appears greater than the same current when the device is pulsed at $V_{BE} = 1\,V$ but, on the contrary, is unaffected by self heating ($150\,ns$).

Self-heating does not affect the device in the same way at every bias condition: the higher V_{BE} is, the more the curves diverge from the quasi-isothermal one, at high V_{CE}. For low V_{BE}, the rise of the collector current is almost absent (see the third and forth "group of curves" in Fig. 3.6, corresponding to $V_{BE} = 0.9$ and $0.85\,V$).

Let's "zoom" on the curves for low V_{BE} amplitude (Fig. 3.7. The y-axis is in log scale). While in Fig. 3.6 the I_C values, for $V_{BE} = 0.8\,V$ for example, are barely visible because the device is almost in the cut-off region, we can now see that the current reaches $3 \cdot 10^{-4}\,A$ and there is no difference at all in its shape by changing the pulse width.

In Fig. 3.7 we can investigate what happens when a $100\,ns$ pulse is produced. While the 4200-SCS is technically able to reach that small value, an important difference is seen between the $100\,ns$ I_C curve and the rest, taken at $t_w \geq 150\,ns$. This difference is roughly constant as V_{BE} rises, and that means, considering the semi-logarithmic scale, that its significance increases exponentially as the base voltage grows.

To justify this, we need to consider the evolution of the pulses over time, in the cases where the applied pulse are $100\,ns$ (Fig. 3.8) and $150\,ns$ (Fig. 3.9). The measure window is based on t_{on}, the interval when V_{BE} reaches its high level.

Because of a transient delay in the collector current growth present in

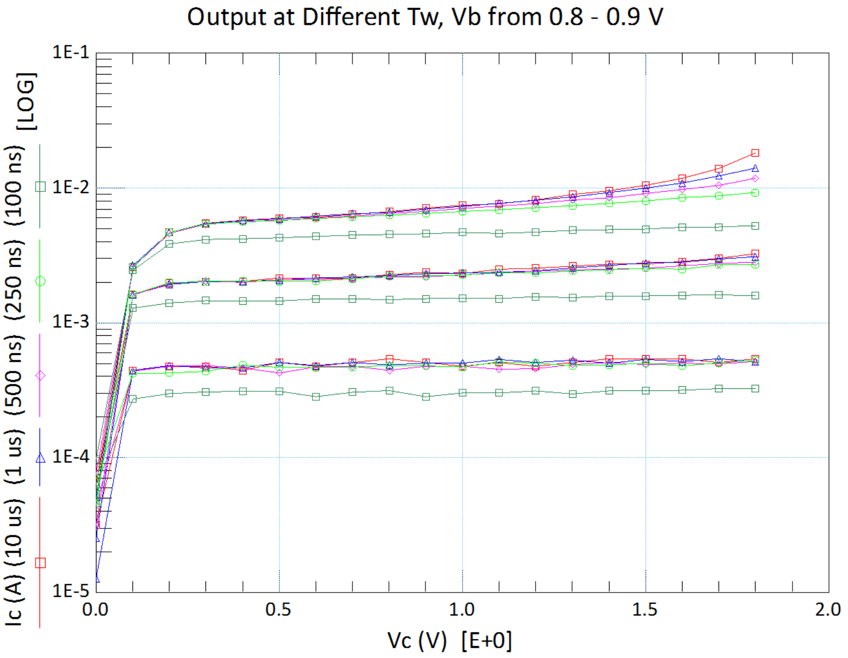


Figure 3.7: $I_C - V_{CE}$ characteristic (semilog scale); sweeping V_{CE} every $100\,mV$, stepping V_{BE} every $50\,mV$. B55: emitter area $A_E = 0.2 \cdot 5\,\mu m^2$.

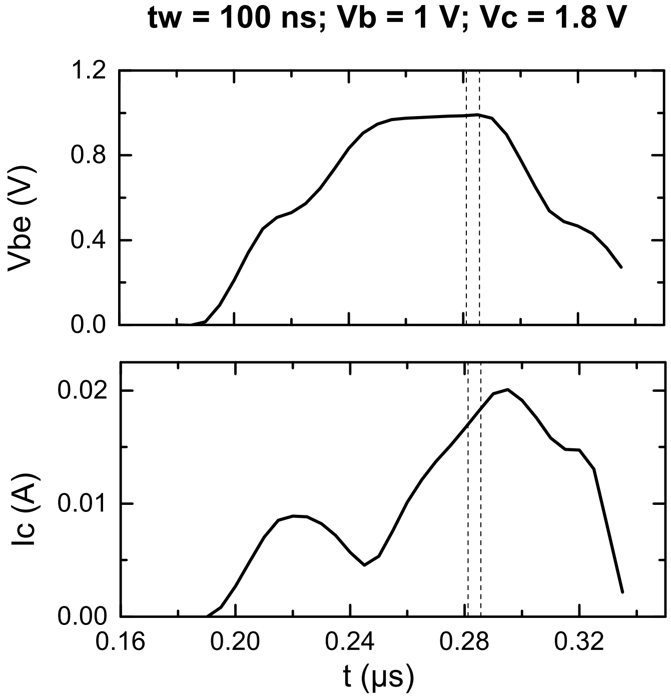


Figure 3.8: Waveform of the base voltage and collector current when a 100 ns pulse is applied. The measurement window is shown. B55: $A_E = 0.2 \cdot 5\,\mu m^2$.

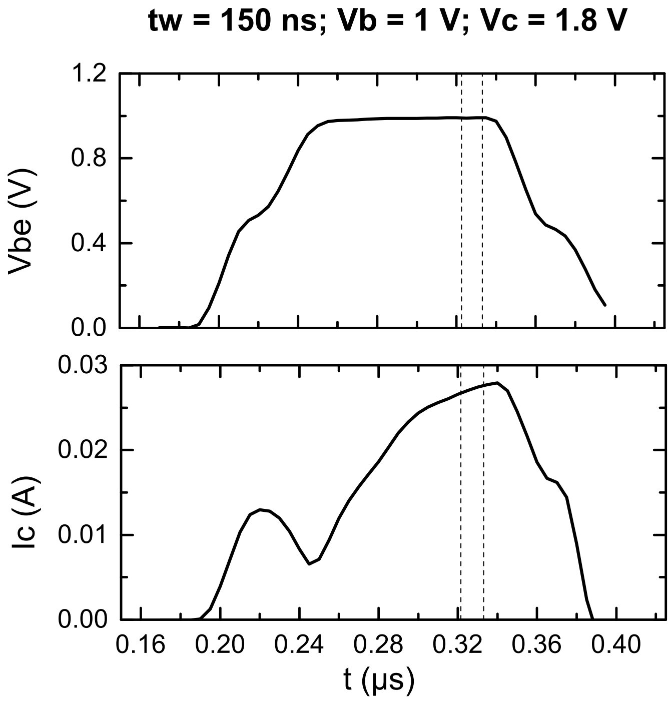


Figure 3.9: Waveform of the base voltage and collector current when a $150\,ns$ pulse is applied. The measurement window is shown. B55: $A_E = 0.2 \cdot 5\,\mu m^2$.

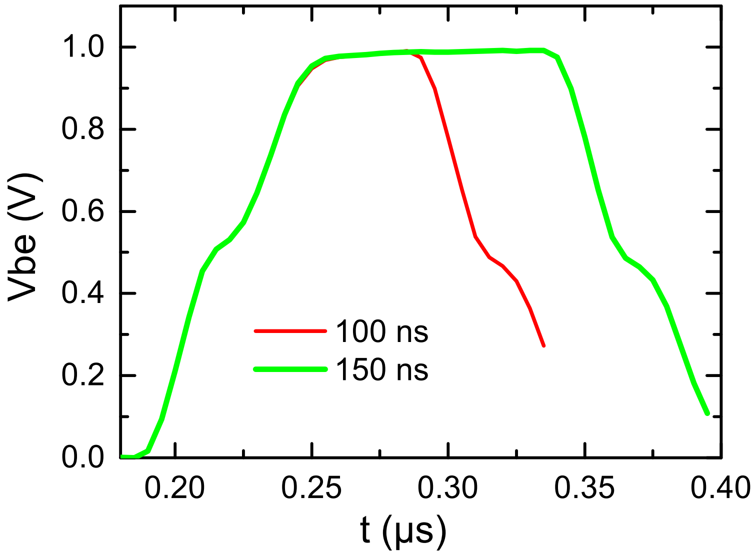


Figure 3.10: Comparison between the V_{BE} evolution over time in the case of $t_w = 100\,ns$ and $t_w = 150\,ns$. B55: $A_E = 0.2 \cdot 5\,\mu m^2$.

both waveforms between 0.20 and $0.24\,\mu s$ mainly due to parasitics on cables (we will better discuss about its origin later on), $I_C(t)$ appears to be still on that transient part due to parasitics at $0.28\,\mu s$, when the sampling for the smaller pulse begins. On the contrary, in Fig. 3.9 $I_C(t)$ reaches a steadier level (the slope of $I_C(t)$ decreases from approximately $0.3\,\mu s$ on even though, due to the bias, a small influence of self-heating is present), right where the sampling is done.

In Fig. 3.10 we directly compare the waveforms of the base voltage and we see that they are exactly the same for the first period corresponding to the $100\,ns$ pulse.

Fig. 3.11 shows the collector currents of which we have discussed so far, remarking the level where $dI_C(t)/dt$ decreases. This is why, in the following, only pulses greater or equal to $150\,ns$ will be applied to the base terminal.

Fig. 3.12 and Fig. 3.13 show the increase of I_C with respect to the pulse width: those plots have been extracted from Fig. 3.6 with vertical sections at

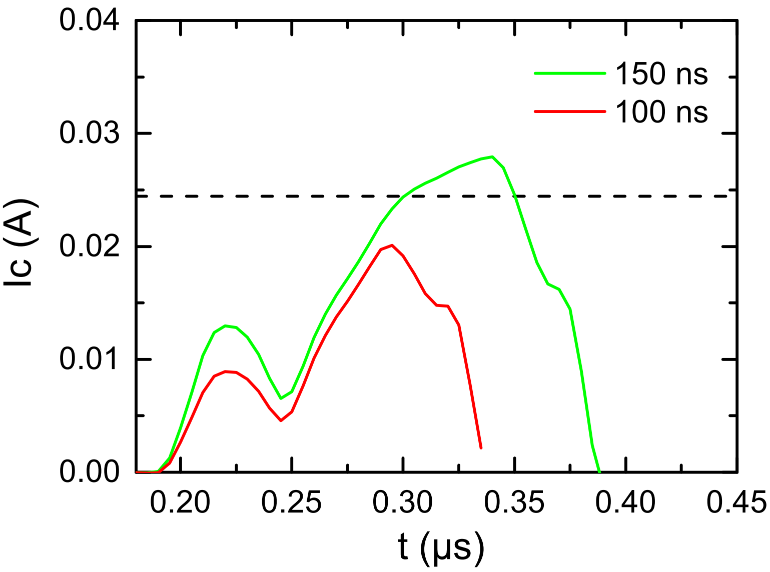


Figure 3.11: Comparison between the I_C evolution over time in the case of $t_w = 100\, ns$ and $t_w = 150\, ns$. B55: $A_E = 0.2 \cdot 5\, \mu m^2$.

fixed V_{CE} values. While in Fig. 3.12 it is not well delineated (owing to V_{BE} close to the turn-on voltage), Fig. 3.13 shows a growing trend of I_C which reaches a saturation level for high t_w. This means that self-heating does not alter the value of I_C any further when very long pulses are applied – namely long pulses well approach the DC behaviour.

Geometrical Comparison of SiGe HBTs

Let's have a look to other geometries from the same B55 technology. Fig. 3.14 and Fig. 3.15 show as an example the smallest and biggest (respectively) geometries considered in this thesis. Of course, I_C is much greater in the biggest devices (the current is two orders of magnitude higher at $V_{BE} = 0.95\, V$, for example), so for sake of visualisation $J_C = I_C/A_E$ will be considered in the following instead and, to avoid as much as possible high current effects, it is convenient that the bias is low.

Focusing on Fig. 3.15, at high collector and base voltage levels, we can

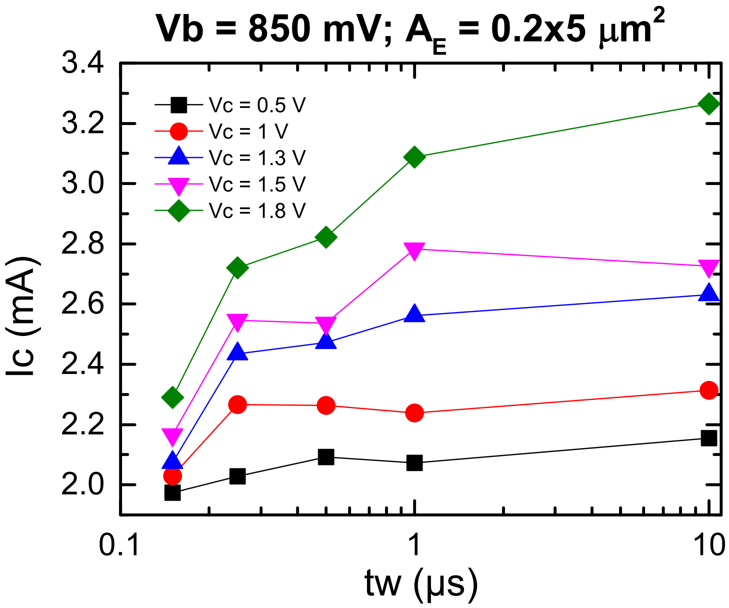


Figure 3.12: $I_C - t_w$ characteristic (semilog scale) at different V_{CE}, $V_{BE} = 850\,mV$; B55: $A_E = 0.2 \cdot 5\,\mu m^2$.

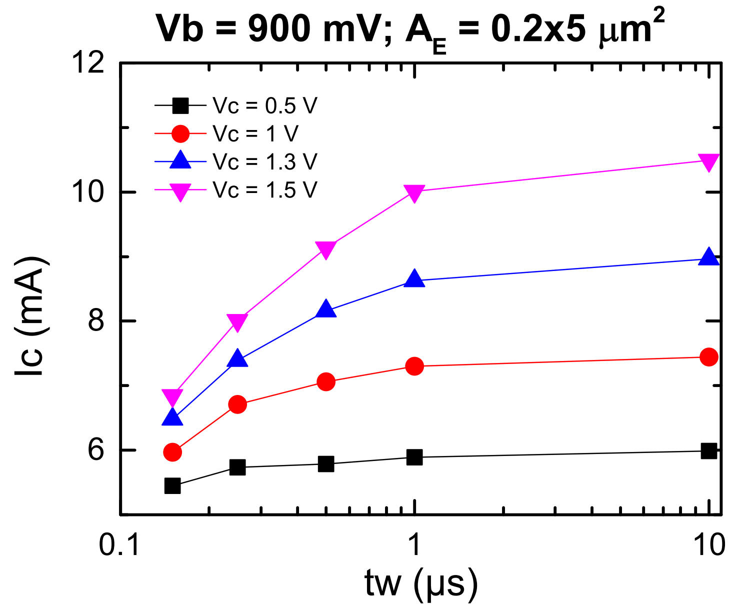


Figure 3.13: $I_C - t_w$ characteristic (semilog scale) at different V_{CE}, $V_{BE} = 900\,mV$; B55: $A_E = 0.2 \cdot 5\,\mu m^2$.

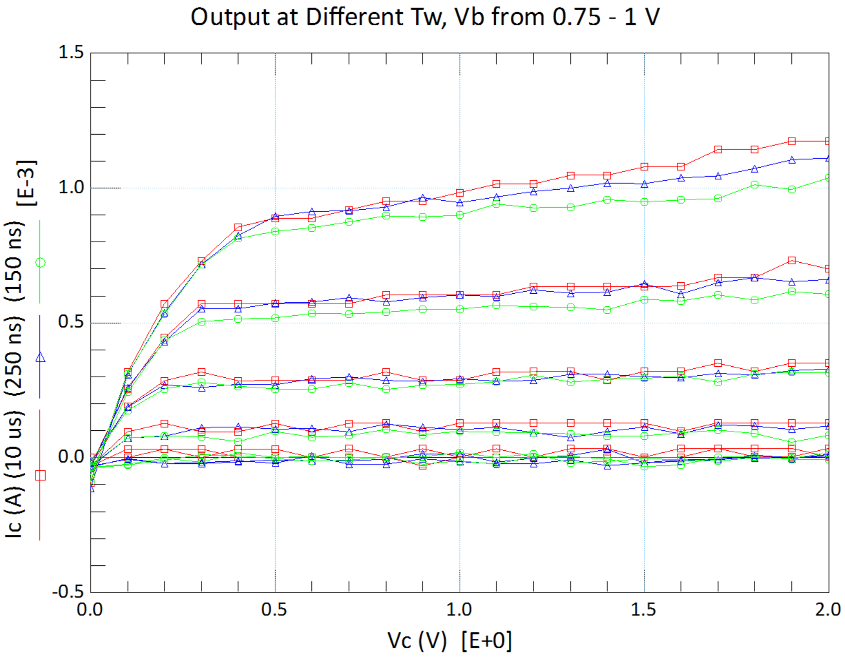


Figure 3.14: $I_C - V_{CE}$ characteristic; sweeping V_{CE} every $100\,mV$, stepping V_{BE} every $50\,mV$. B55: emitter area $A_E = 0.2 \cdot 0.45\,\mu m^2$.

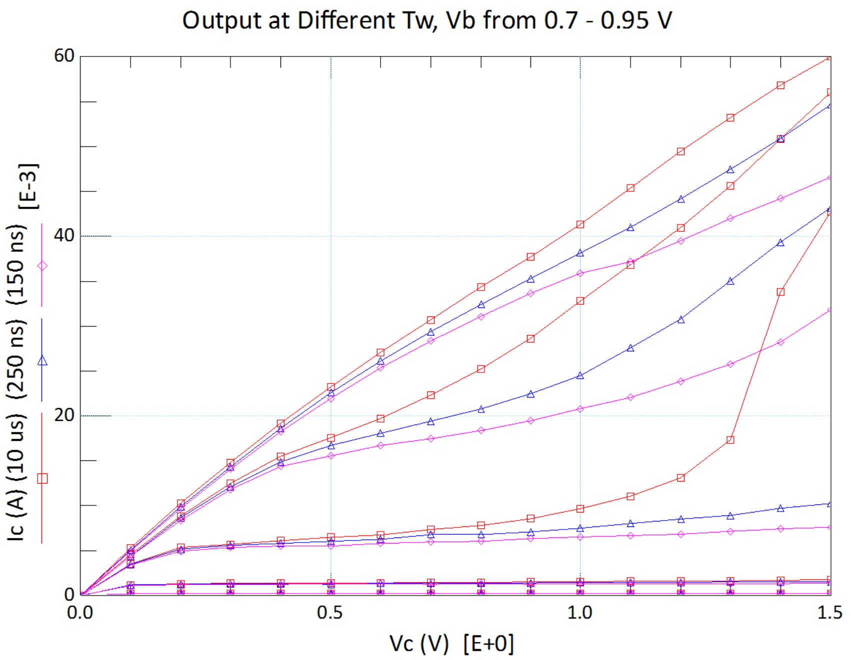


Figure 3.15: $I_C - V_{CE}$ characteristic; sweeping V_{CE} every $100\,mV$, stepping V_{BE} every $50\,mV$. B55: emitter area $A_E = 0.42 \cdot 5\,\mu m^2$.

clearly see a deformation of the standard I/V characteristic. The cause is the joint presence of high-current effects (Kirk effect) and self-heating. At first glimpse, self-heating seems almost absent in the smallest device even at high bias, while it is evident in the latter case. Let's carry out a more rigorous analysis.

In Fig. 3.16 and Fig. 3.17 a comparison is made. We consider here the two extreme cases of quasi-isothermal and quasi-DC conditions, and different geometries provided by the B55 technology.

From Fig. 3.16 (a) we can easily assert that, independently of the width of the device, self-heating does not appear other than as a small variation at high V_{CE}. The red curve stands for the device with $A_E = 0.2 \cdot 5\,\mu m^2$ and it is approximately 0.8 to $1\,mA/\mu m^2$ distant from the other two curves: this parasitic behaviour is not due to self-heating since the deviation is present from early V_{CE} values. This is due to peripheral effects (for example a slightly different doping distribution) which are dominant in small transistors compared to large ones.

The same geometries are shown in Fig. 3.16 (b) but in this case we are in a quasi-DC condition. Here, the effect of self-heating appears: the purple curve diverges from the blue one at high V_{CE}. Also, the blue one diverges because of self-heating from the red one and we observe that the aforementioned deviation of 0.8 to $1\,mA/\mu m^2$ for low V_{CE} is still present like in the quasi-isothermal case. From these considerations we can affirm that the more the device is wide, the more it heats up.

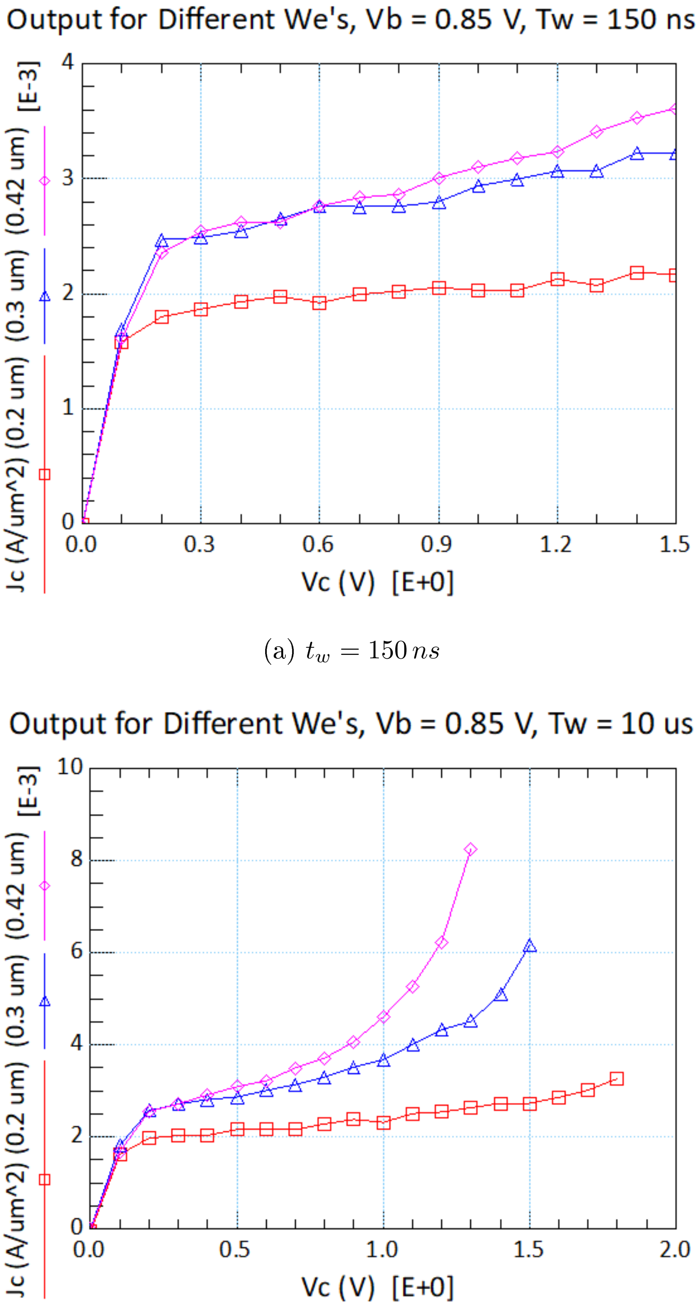
Let's move to make similar consideration about the emitter length variations (see Fig. 3.17). In Fig. 3.17 (a), although the measure for the smallest device is rather noisy, we see that there is no self-heating effect at all, even with a big variation of L_E (from blue to green we change to a ten times longer emitter). In Fig. 3.17 (b) this variation is minimal and it is more remarkable by passing from the noisy blue trace to the purple one (from $L_E = 1$ to $5\,\mu m$) than from the purple one to the green one (from $L_E = 5$ to $10\,\mu m$).

A more complete comparison is made in Fig. 3.18: J_C is plotted versus the pulse width for different emitter widths. The measures are extracted at the same V_{BE} value for two different V_{CE}: 1 and $1.3\,V$ – two values for which the transistor is in the active region but the Kirk effect is not strong.

The samples are taken by applying three pulses with different widths to the base pin: $150\,ns$, that stands for the quasi-isothermal condition, $250\,ns$, an intermediate value, $10\,\mu s$ for the quasi-DC. The curves all show a growing trend due to self-heating, as expected.

At the quasi-isothermal condition, the values of the samples are not the same, but grow as the width grows. This is due to the fact that the thermal resistance does not scale well for the largest transistor meaning that there is in fact some current rise due to self-heating even for short pulses. Whereas, for the smallest, the difference is mainly due to peripheral effects.

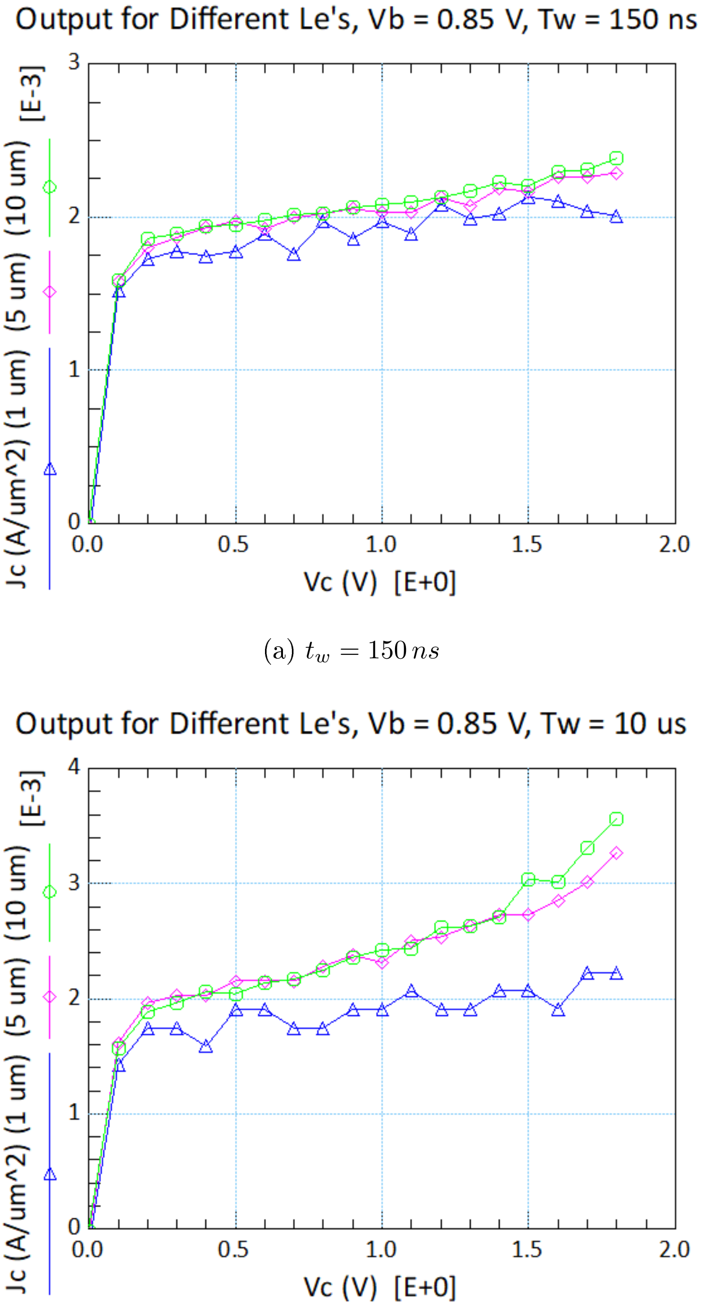
A larger pulse width has a different effect on the rise of J_C, as expected. We can see clearly in both Fig. 3.18 (a) and (b) that passing to a $250\,ns$



(a) $t_w = 150\,ns$

(b) $t_w = 10\,\mu s$

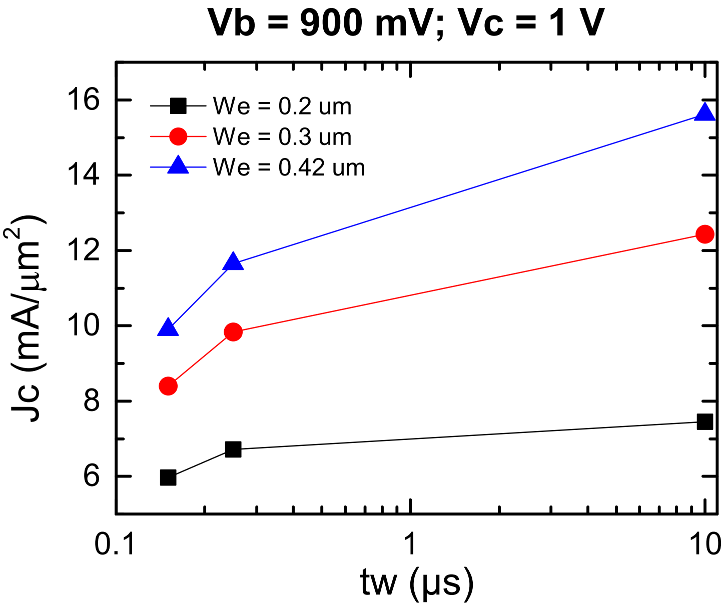
Figure 3.16: $J_C - V_{CE}$ characteristic; sweeping V_{CE} every $100\,mV$, $V_{BE} = 850\,mV$; B55: $L_E = 5\,\mu m$, $W_E = 0.2, 0.3, 0.42\,\mu m$



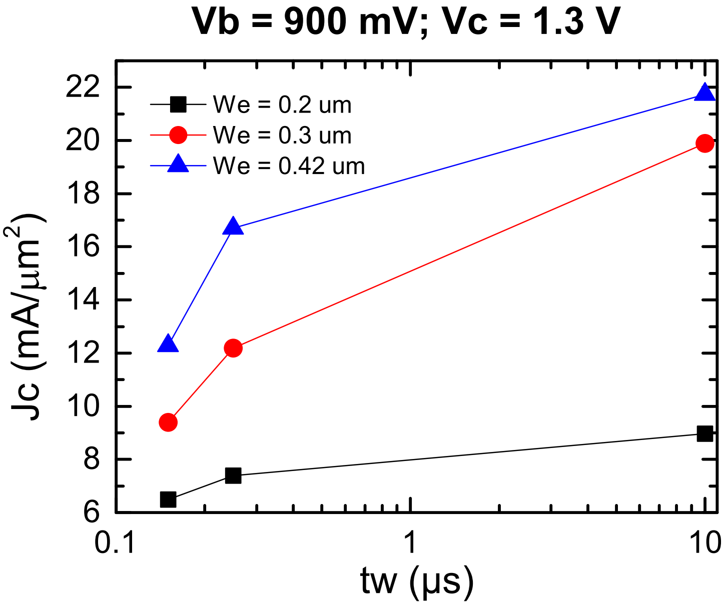
(a) $t_w = 150\,ns$

(b) $t_w = 10\,\mu s$

Figure 3.17: $J_C - V_{CE}$ characteristic; sweeping V_{CE} every $100\,mV$, $V_{BE} = 850\,mV$; B55: $W_E = 0.2\,\mu m$, $L_E = 1, 5, 10\,\mu m$



(a) $V_{CE} = 1\,V$



(b) $V_{CE} = 1.3\,V$

Figure 3.18: $J_C - t_w$ characteristic (semilog scale) at different W_E; $V_{BE} = 900\,mV$; $t_w = 0.15, 0.25, 10\,\mu s$. B55: $L_E = 5\,\mu m$.

pulse ensures that the wider device heats up more – the slope is greater as the geometry is bigger. The same happens passing to $10\,\mu s$, even though, in Fig. 3.18 (b), the red curve has a greater slope than the blue one: that is because, at that bias, the high current effects become non-negligible.

Fig. 3.19 shows similar curves for a change in L_E, keeping W_E constant: the slopes of all curves follow the usual trend due to self-heating and, at DC, the highest collector current density is in the biggest devices. The smallest device curve again stays sensibly low with respect to the others. So, while the thermal resistance seems to scale better by changing the emitter length with respect to the emitter width, the lateral effects still remain.

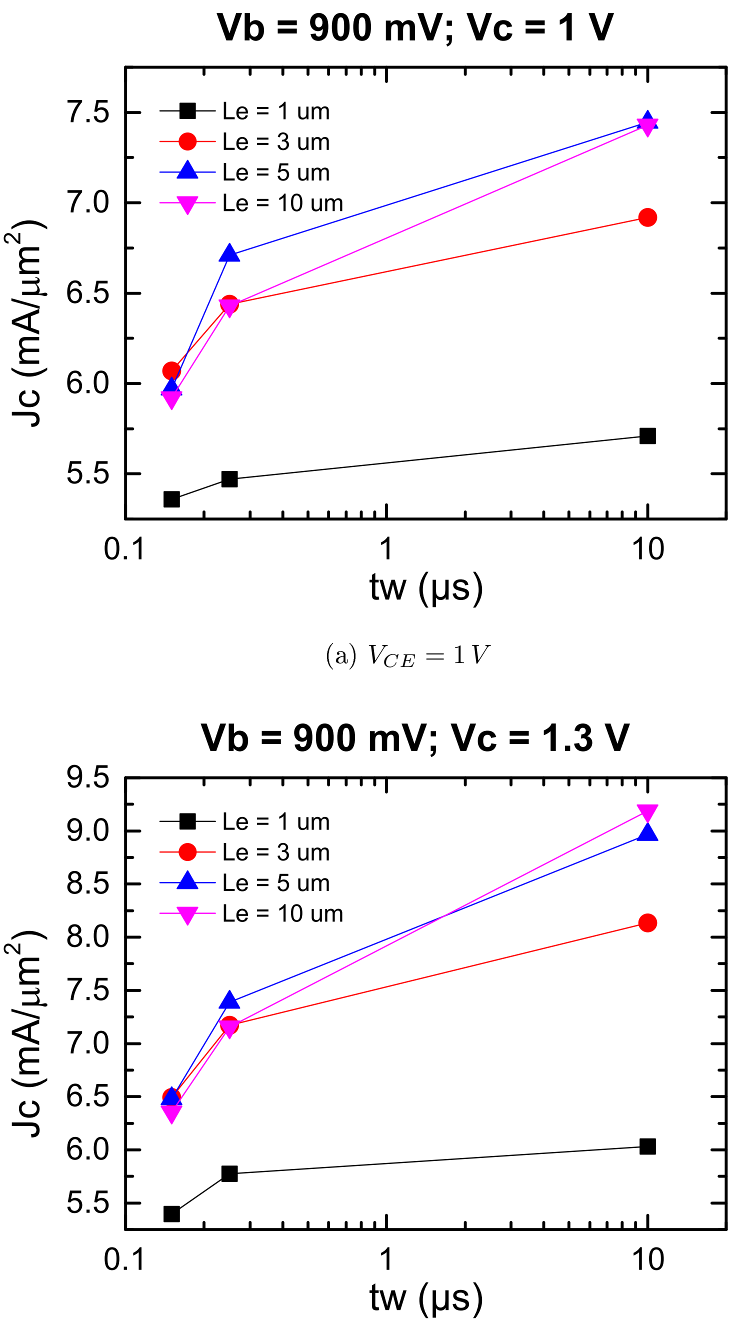
Characteristics of the InP-based HBTs

We now consider a different technology, the InP/InGaAs HBT from III-V Lab. Fig. 3.20 shows our reference device ($A_E = 0.7 \cdot 5\,\mu m^2$) when stimulated by different pulse widths at the base pin, at different V_{BE} amplitude values. Self-heating appears to affect the device in a similar way as depicted for the B55. In Fig. 3.21, the extraction of samples at different V_{CE} has been made analogously to before.

Let's now consider the J_C vs. t_w curves like for the B55 (Fig. 3.22) – similar remarks can be written. We see here very clearly (from the black and the red curves particularly) that the trend of J_C is almost logarithmic, namely $J_C \propto \log(\log t_w)$ at fixed bias.

Finally, a comparison is made between the two technologies as far as the geometry is concerned. In the y-axis of the graph in Fig. 3.23, the percentage increase of the quasi-DC value of J_C with respect to the quasi-isothermal is shown. The samples are grouped according to the dimension changing and the technology. A small change of the area due to an increase of the width has the strongest effect on self-heating (in red). When changing the length of the emitter, the rise of the collector current is rather constant independently of the area, except for small A_E (peripheral effects).

In Fig. 3.24 yet another InP-based technology's behaviour in pulsed con-



(a) $V_{CE} = 1\,V$

(b) $V_{CE} = 1.3\,V$

Figure 3.19: $J_C - t_w$ characteristic (semilog scale) at different L_E; $V_{BE} = 900\,mV$; $t_w = 0.15, 0.25, 10\,\mu s$. B55: $W_E = 0.2\,\mu m$.

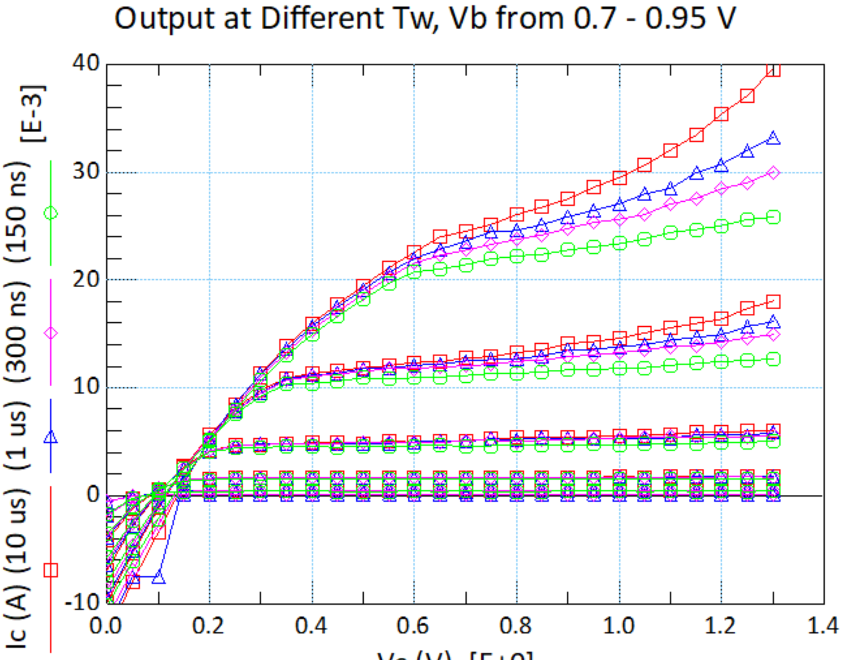


Figure 3.20: $I_C - V_{CE}$ characteristic; sweeping V_{CE} every $50\,mV$, stepping V_{BE} every $50\,mV$. III-V Lab: $A_E = 0.7 \cdot 5\,\mu m^2$.

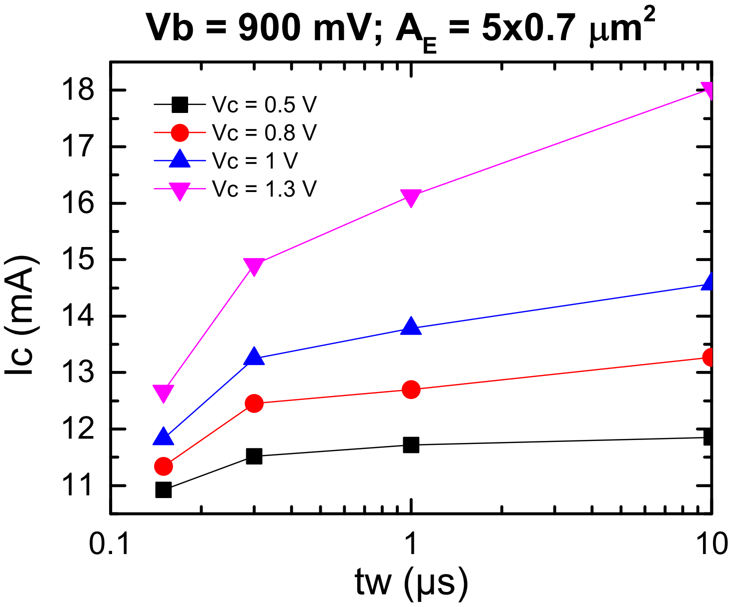
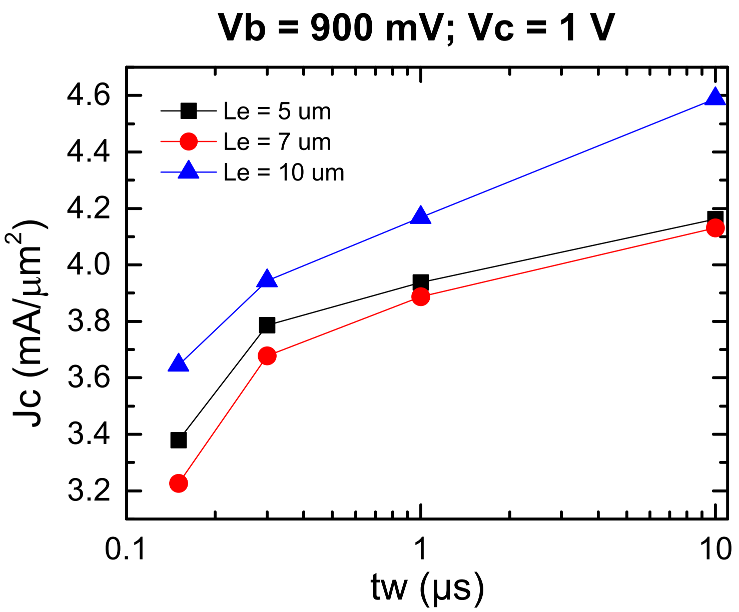


Figure 3.21: $I_C - t_w$ characteristic (semilog scale) at different V_{CE}, $V_{BE} = 900\,mV$; III-V Lab: $A_E = 0.7 \cdot 5\,\mu m^2$.

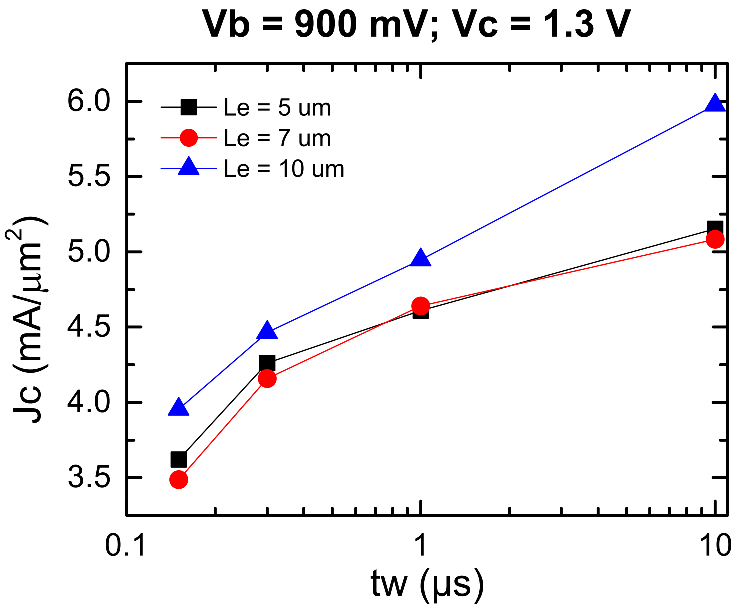
ditions is shown, that is the ETHZ technology (InP/GaAsSb). For these devices we do not dispose of different geometries. The difference between the chips lies in the technological dispersion, that means different processes are involved, different doping is used, etc...

It is still interesting to compare the graphs in Fig. 3.24 (a) and (b) with Fig. 3.6, from which we have developed our discussion in this chapter. These devices present approximately the same emitter area, and even though the formers come from a different technology than the latter – so we may actually expect differences in the behaviour when applying different pulses –, the first two $I_C - V_{CE}$ characteristics deviate between them.

The device from chip J11 – our reference – carries much less current than that of chip F7 at the same bias, and it heats up more, consequently reaching even higher I_C levels. Furthermore, the device of chip F7 seems not to reach a saturated quasi-DC level of current unlike J11, and for this reason a 50 μs pulse has been applied for testing, and it differs substantially from the one at



(a) $V_{CE} = 1\,V$



(b) $V_{CE} = 1.3\,V$

Figure 3.22: $J_C - t_w$ characteristic (semilog scale) at different L_E; $V_{BE} = 900\,mV$; $t_w = 0.15, 0.3, 1, 10\,\mu s$. III-V Lab: $W_E = 0.7\,\mu m$.

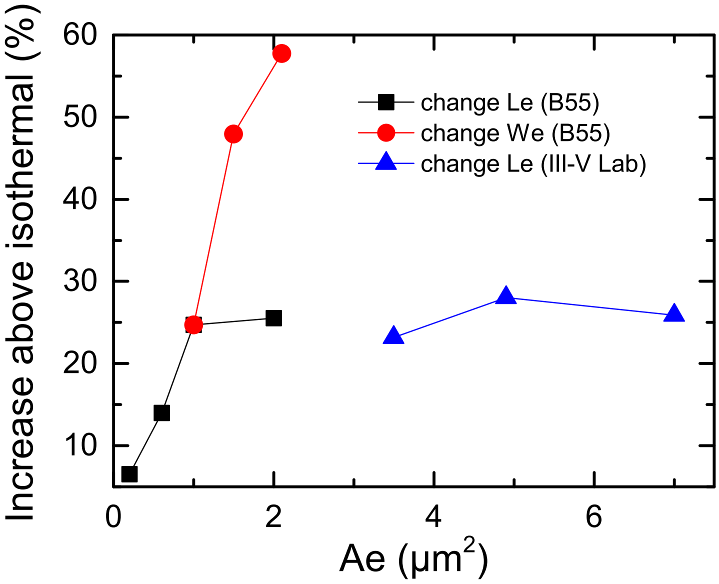
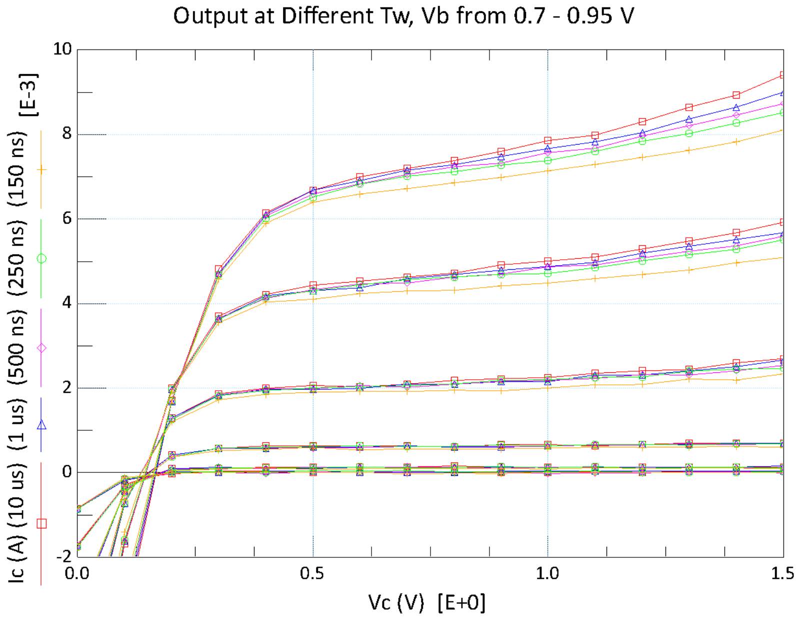
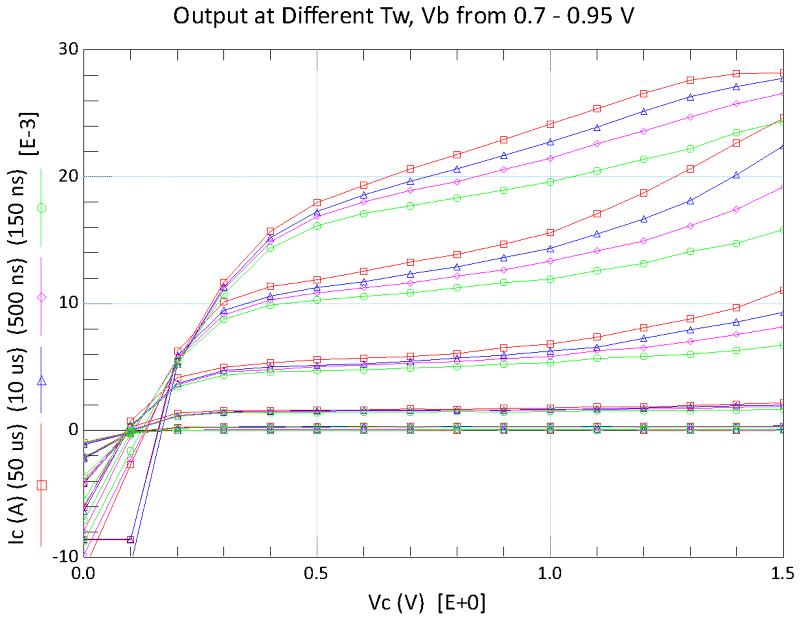


Figure 3.23: Normalized increase of the sample value of J_C measured with a pulse of $10\,\mu s$ with respect to one of $150\,ns$; $V_{BE} = 900\,mV$, $V_{CE} = 1\,V$; SiGe HBT (B55) and InP HBT (III-V Lab).



(a) Chip J11



(b) Chip F7

Figure 3.24: $I_C - V_{CE}$ characteristic; sweeping V_{CE} every $100\,mV$, stepping V_{BE} every $50\,mV$. ETHZ: $A_E = 0.2 \cdot 4.4\,\mu m^2$.

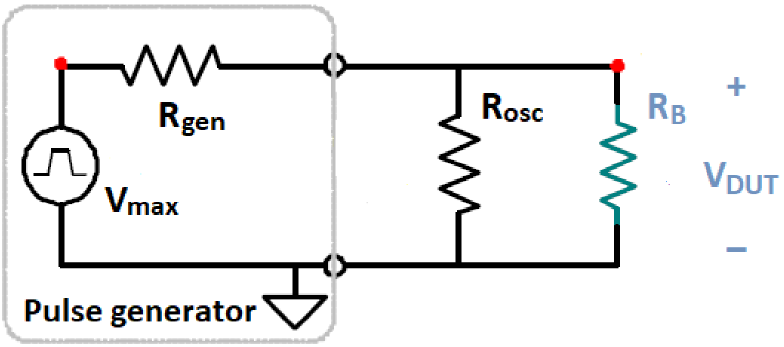


Figure 3.25: The schematic of the base channel and DUT when an oscillator is connected.

10 μs, even at moderate bias. None of the two characteristics is considerably deformed by high currents, unlike the case of B55.

3.2.2 Transient-I/V

The time-domain characterisation provides an useful insight of the gradual effect of the rise of temperature in the device due to the dissipated power, as well as allowing to extract information on the measurement execution and on self-heating itself. In order to deduce any information, we first need to investigate the measurement process.

The Load-Line Effect

Let us consider again Fig. 3.3, that is a situation where the 4220-SCS is connected to the device under test (DUT), and let's add to both channels an oscilloscope to show the voltage waveform evolution "in real time". The equivalent circuit of this configuration for the base terminal is depicted in Fig. 3.25. The circuit is similar for the collector terminal, except for a DC voltage source instead of the pulse generator.

This analysis has been simplified. First, we suppose to be in a steady-state condition, namely we do not consider the reactive components ($t \rightarrow \infty$,

$f = 0$). This hypothesis is of course not fully verified, all the more so because we will apply in this part a pulse width of $t_w = 150\,ns$ with period $T = 2\,t_w$, in order to clearly display the pulse shape on the oscilloscope screen. However, the measurement, as we have already said, is an average carried out from $0.8\,t_{on}$ to $0.95\,t_{on}$, namely in a rather steady situation. Also, we neglect the resistance of the coaxial cables (less than $1\,\Omega$) and its reactive components as well. A full design of this passive circuit has been carried out by [28]: far from providing a rigorous parameter extraction, the following observations just aim to give an explanation of the time-domain curves.

So as far as our analysis is concerned, the generator is made of an ideal voltage pulse source and a series resistance ($R_{gen} = 50\,\Omega$). The load is represented by two shunt resistances, the oscillator (selected $50\,\Omega$) and R_B. R_B is the total base resistance and it is composed by the series of R_{Bx}, the external base resistance dominated by the contact resistance and R_{iB}, the internal base resistance.

By using the π-hybrid model of a bipolar transistor, R_{iB} can be written as [38]

$$R_{iB} = r_\pi + (\beta_o + 1)\,R_E \simeq r_\pi + \beta_o\,R_E$$

where r_π is the differential input resistance, defined at the quiescent point as

$$r_\pi = \left.\frac{\partial v_{BE}}{\partial i_B}\right|_Q = \frac{V_T\,\beta_o}{I_C}$$

while V_T is the thermal voltage, β_o is the internal current gain and R_E is the emitter resistance. We can thus rewrite

$$R_B = R_{Bx} + \beta_o\left(\frac{V_T}{I_C} + R_E\right) \tag{3.1}$$

We will now derive an approximate value of R_B just by inspection of the voltage waveforms, and simultaneously understand how the PMU carries the measurements out. To do so, we consider again Fig. 3.25. When the DUT is changed, R_B changes according to Eq. (3.1) and so does the resistance seen from the pulse generator. This is the so-called *load-line effect.*

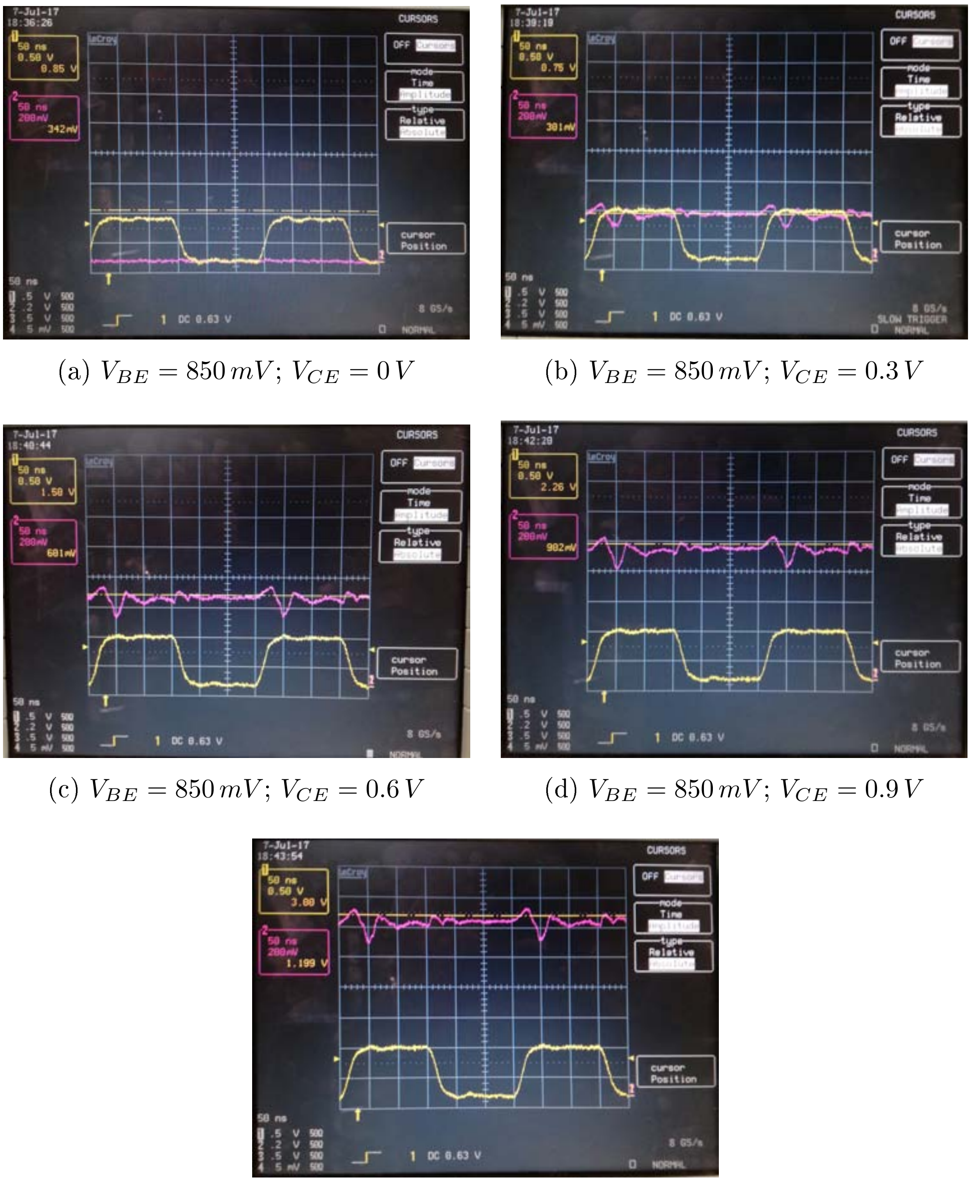
To obviate the problem, a compensation has to be made in such a way that any DUT can be measured independently of the the resistance it shows. There are two ways to perform it: manually or automatically. The manual way consists in using an oscilloscope to target the desired voltage levels and continuously increasing the input settings above the programmed values. The other way lies in exploiting an algorithm to do exactly the same.

The load-line effect compensation (LLEC) implemented by the measuring software used in our discussion adjusts the value of V_{max} (the amplitude of the pulse), but analogously and simultaneously the DC voltage at the collector, in order to match the targeted value inserted by the user. This is done through an algorithm that performs a set number of iteration from a low level of voltage, then measures and averages after every cycle the base and collector temporary values. When the target is reached, the current and voltage data at the base and the collector are retrieved and used for the pulsed-I/V characteristic.

The ongoing of this procedure is shown in Fig. 3.26: here are the waveforms a moment before a new cycle of the LLEC is performed, i.e. these are the curves on which the measures are carried out. The input set-up is the following:

- $V_{B_{START}} = 0.85\,V$; $V_{B_{STOP}} = 0.85\,V$; $V_{B_{STEP}} = 0\,V$; $V_{B_{BASE}} = 0\,V$.
- $V_{C_{START}} = 0\,V$; $V_{C_{STOP}} = 1.2\,V$; $V_{C_{STEP}} = 0.3\,V$.
- $t_w = 150\,ns$; $t_R = t_F = 20\,ns$; $t_{DELAY} = 50\,ns$; $T = 300\,ns$; start measurement location = 80%; stop measurement location = 95%; pulse average = 1000.

The voltage outputs during a full iteration have been noted down and from them we will calculate R_B. The DUT we used is the largest from B55 ($A_E = 0.42 \cdot 5\,\mu m^2$). $V_{DUT}(t_i)$ (the voltage at the end of the first iteration, see Fig. 3.25) is found by inspection to be $V_{DUT}(t_i) = 0.6465\,V$. Our programmed voltage is $V_{prog} = V_{DUT}(t_f) = 0.85\,V$ and corresponds to the final desired value of V_{DUT}. $V_{max}(t_i)$ is set by default to $2\,V_{prog}$ [37], this because



(a) $V_{BE} = 850\,mV$; $V_{CE} = 0\,V$

(b) $V_{BE} = 850\,mV$; $V_{CE} = 0.3\,V$

(c) $V_{BE} = 850\,mV$; $V_{CE} = 0.6\,V$

(d) $V_{BE} = 850\,mV$; $V_{CE} = 0.9\,V$

(e) $V_{BE} = 850\,mV$; $V_{CE} = 1.2\,V$

Figure 3.26: Oscilloscope pictures displaying the actual level reached by the voltages applied to the base (yellow) and the collector (purple) at the end of each LLEC iteration.

the default setting of the resistance seen from the pulse generator is $50\,\Omega$ (impedance matching).

We use the following formula, derived by the circuit, to find the only unknown, namely R_B. The formula is

$$V_{DUT}(t_i) = V_{max}(t_i) \cdot \frac{R_{osc} \| R_B}{R_{osc} \| R_B + R_{gen}}$$

After few algebraic steps, $R_B = 80\,\Omega$ is found. Please note that this value is bias-dependant (thus also time-dependant since the bias changes over time) and this is also clear from Eq. (3.1). Also note that, by this calculation by inspection, the constant value of R_{Bx} cannot be found since it's impossible to measure $I_C(t_i)$, even though an "upper bound" can be indicated.

To do so, we observe the following inequality

$$R_{iB} \gg \frac{V_T}{I_B} = \frac{V_T}{V_{DUT}(t_i)} R_B$$

that stems from Eq. (3.1) and the fact that $\beta_o > \beta_F$. $I_C = \beta_F \, I_B$ and $I_B = V_{DUT}/R_B$ have been used. By substituting all terms we find $R_{iB} \gg 3\,\Omega$. From the HICUM compact model of this HBT we know that $R_{Bx} = 18.9\,\Omega$ in accordance to what we have said so far.

I/V Performance Over Time

After having understood how the measurement process breaks out, we consider again the kind of measure we discussed in the previous section, namely a pulse generated in such a way to either heat up the device or not.

In Fig. 3.27 time-based I/V measurements are show. We are in the quasi-DC case, where a $10\,\mu s$-wide pulse is applied to the base (seen in the fourth curve). This kind of measure is a waveform capture and it has been carried out directly from the KITE interface, which sampled the entire $V_{BE}(t)$ pulse and the correspondent values of $I_B(t)$, $V_{CE}(t)$ and $I_C(t)$ (from bottom to top, in the figure) every $50\,ns$. The pulse average has been chosen of 15, so that every sample has been averaged by correspondent samples from 15 different sets: the curves in Fig. 3.27 are made of the resultant samples.

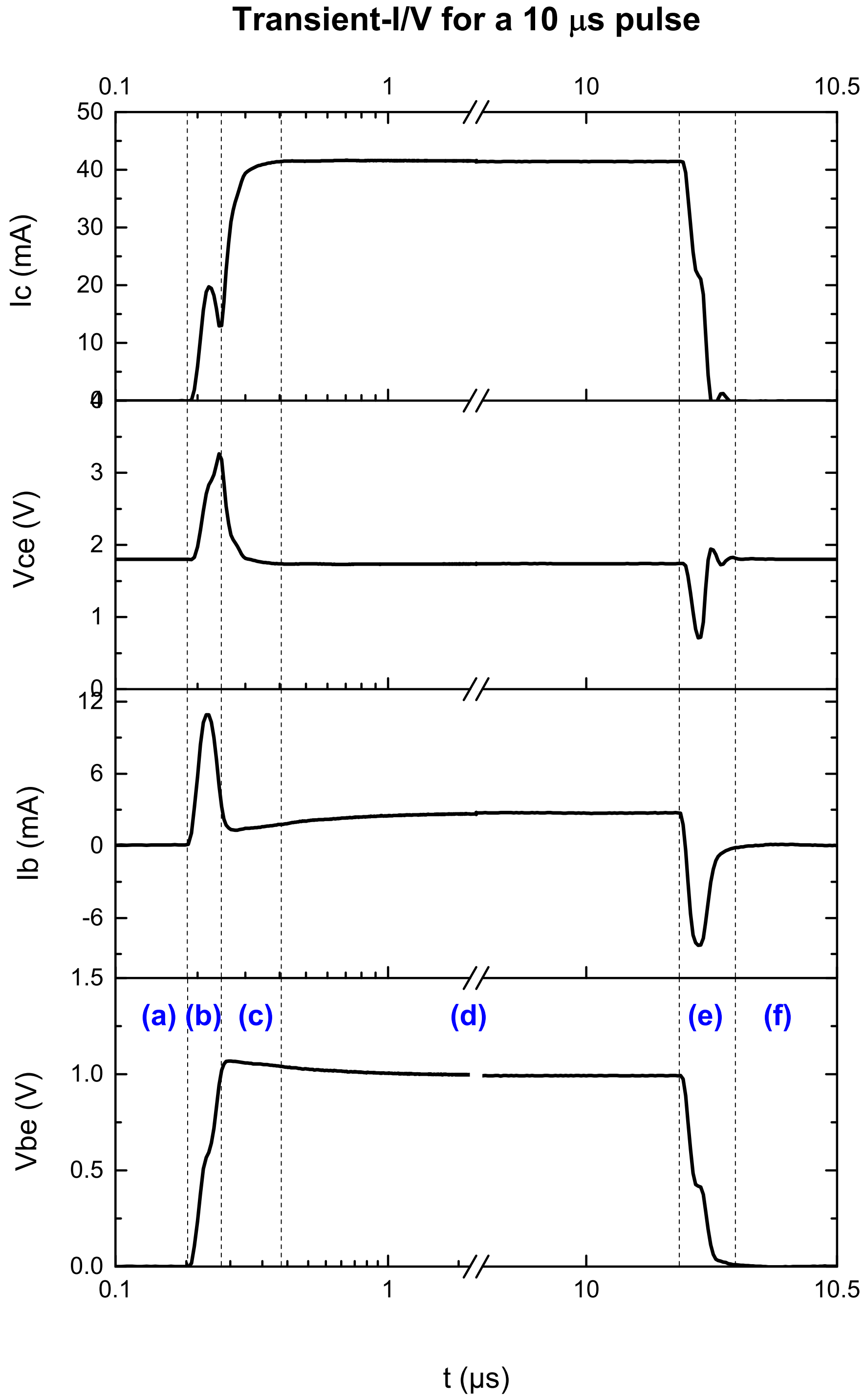


Figure 3.27: Current and voltage waveforms when a pulse is applied (10 μs). The plot has been divided into regions emphasizing the pulse transitions.

Please notice the x-axis scale (the time scale): the first part is semi-logarithmic (from $t = 0.1\,\mu s$ to $2\,\mu s$), then after the break it is linear, up to $10.5\,\mu s$. This has been done for sake of visualisation of the transients, both at the beginning and the end of the pulse. A $0.2\,\mu s$ pulse delay has been applied, too. The total period is $T = 10\,t_w = 100\,\mu s$ and the rise and fall times are $20\,ns$. The imposed bias is $V_{BE} = 1\,V$ and $V_{CE} = 1.8\,V$.

The plot has been divided into regions (from (a) to (f)) for a qualitative analysis:

(a) In this region, only the DC bias at the collector is applied, because of the pulse delay. The device is off and the current flows neither in the base nor in the collector.

(b) The rise period begins and the first transient happens. The capacitance in the system – consisting of the cable capacitance, PMU capacitance, and device capacitance, but mostly due to the cable – is charged by a parasitic current. This current is equal to $C\,dV_{BE}(t)/dt$, so even if the capacitance is on the order of some picofarads, it still is some milliamp high, since this transient lasts just $20\,ns$ and the voltage goes within this period from 0 to $1\,V$. The total $I_B(t)$ current thus shows a peak. This high injection of holes in the base recalls some electrons that manage to flow from emitter to collector, even with a low current gain, since the device is not fully on. Owing to the capacitance at the collector (Channel 2), $V_{CE}(t)$ rises too in this short period.

(c) The base voltage is nearly constant now (it is actually recovering from a moderate overshoot), and so is the base current, stabilizing to few microamps. The collector current, on the contrary, begins to rise up with exponential trend to its final value in (d) instead of sticking to the level imposed by the current gain: this further growth is the effect of self-heating. At the same time, $V_{CE}(t)$ is brought back to its DC value by the current variation through the cable inductance ($L\,dI_C(t)/dt$).

(d) In this phase, the established bias is reached, and the measure is carried

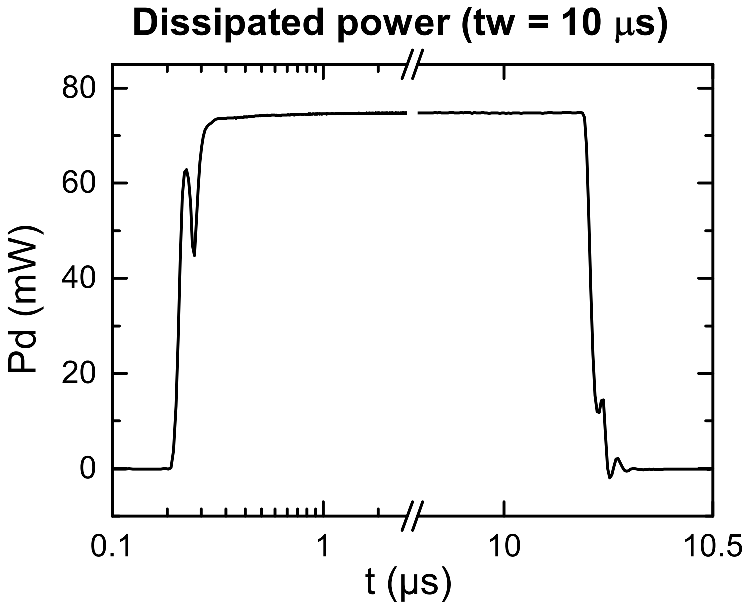


Figure 3.28: Instantaneous dissipated power when a pulse is applied ($10\,\mu s$).

out between 80 and 95 % of the total period given by (c) and (d). We can see that the measurement is retrieved during a completely steady-state condition.

(e) During the V_{BE} fall time, the inverse I_B spike is due to a discharge current in the base capacitance. The base current is null and so is the collector current: the device is off.

(f) We find ourselves in the same situation as (a). The pulse is over and this condition continues for $T - t_w - t_{DELAY}$. Then another period begins.

Fig. 3.28 stems from Fig. 3.27. It represents the instantaneous dissipated power $P_d(t)$ inside the device during the application of the pulse. It has been calculated by using the definition of dissipated power

$$P_d(t) = I_B(t)\,V_{BE}(t) + I_C(t)\,V_{CE}(t)$$

From the retrieved data we can compute the value of the average dissipated power by passing from integration to summation, namely

$$P_{d,avg} = \frac{1}{T}\int_{t_a}^{t_b} p(\tau)\,d\tau = \frac{\Delta t}{T}\sum_{t=t_a}^{t_b} p(t)$$

where $t_b = t_a + t_w$ and $t_a = t_{DELAY}$, and Δt, the sampling step, must be short to verify the equality. In our case, $\Delta t = 0.005\,\mu s$. The average power dissipation when a $10\,\mu s$ pulse is applied to the reference B55 HBT is $P_{d,avg} = 7.46\,mW$.

The waveform in Fig. 3.28 also represents the temperature rise within the device, with only a scaling correction by R_{TH} needed ($\Delta T(t) = P_d(t) \cdot R_{TH}$). It is important to remark that this temperature rise is partly due to the self-heating.

Fig. 3.29, similarly to before, shows the I/V waveforms when applying a pulse such that the temperature rise is moderate. No steady level is reached by I_C unlike in Fig. 3.27 (where it has reached a final value of approx. $43\,mA$), but the qualitative analysis is still valid (from (a) to (f), excluding (d)). Not even the dissipated power (Fig. 3.30) reaches the saturation level and the average dissipated power is $P_{d,avg} = 4.1\,mW$ in this case.

Thermal Time Constant Extraction

By comparing the two cases from above, we see that a duty cycle of $10\,\%$ engenders both the average dissipated powers being one order of magnitude lower than the maximum dissipated during the period, less stressing the devices. Though in the former case, the measurement has been carried out long after the thermal time constant and the self-heating is not negligible.

We will exploit now the I_C waveforms related to the $10\,\mu s$ pulses to get a valid approximation of the value of constant τ_{TH}, as it has been defined in the previous chapter. As seen in the previous qualitative analysis, we can think of I_C as a curve that exponentially increases above its isothermal value because of the heating that is generated within the HBT, up to a steady-state value.

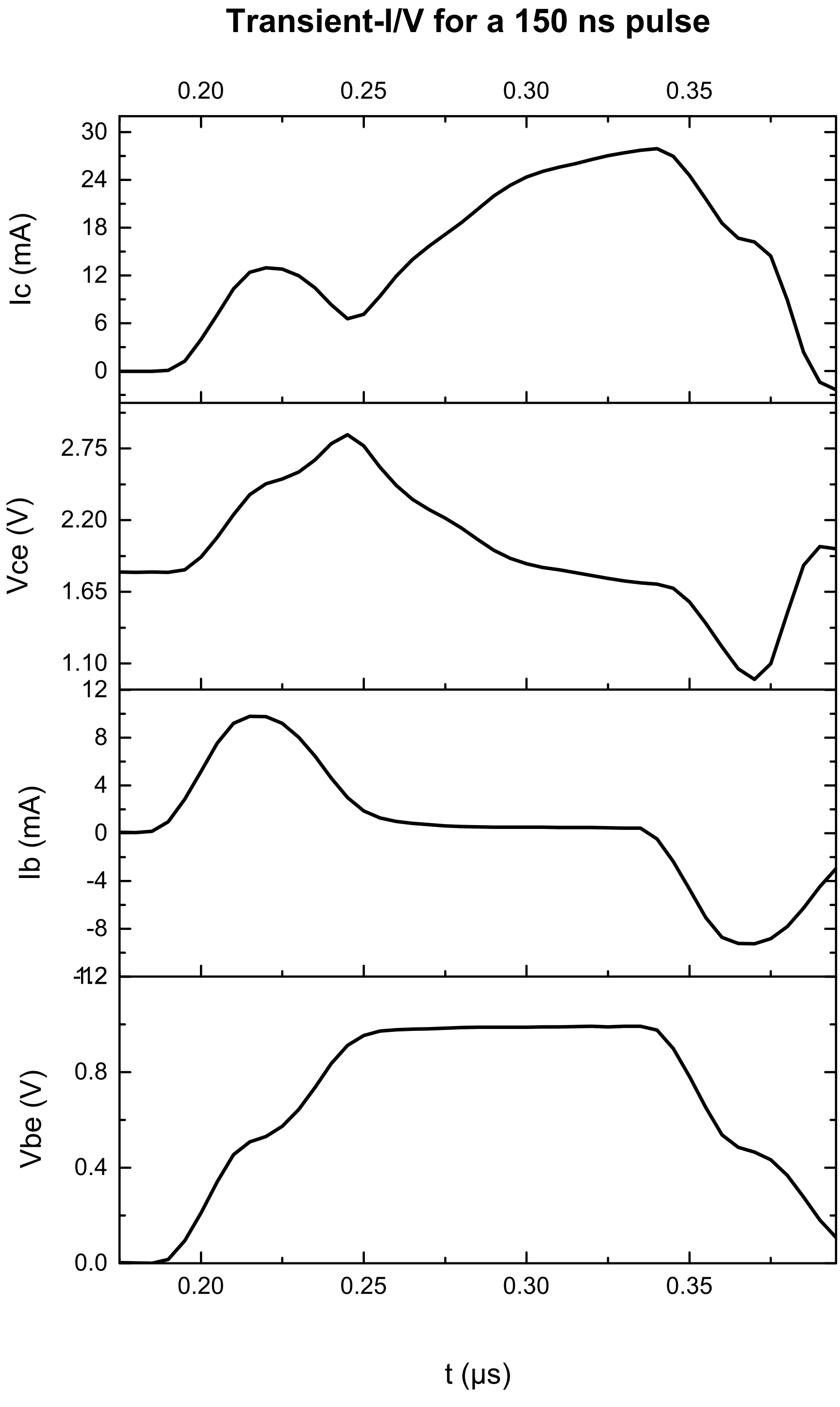


Figure 3.29: Current and voltage waveforms when a pulse is applied ($150\,ns$). The plot has been divided into regions emphasizing the pulse transitions.

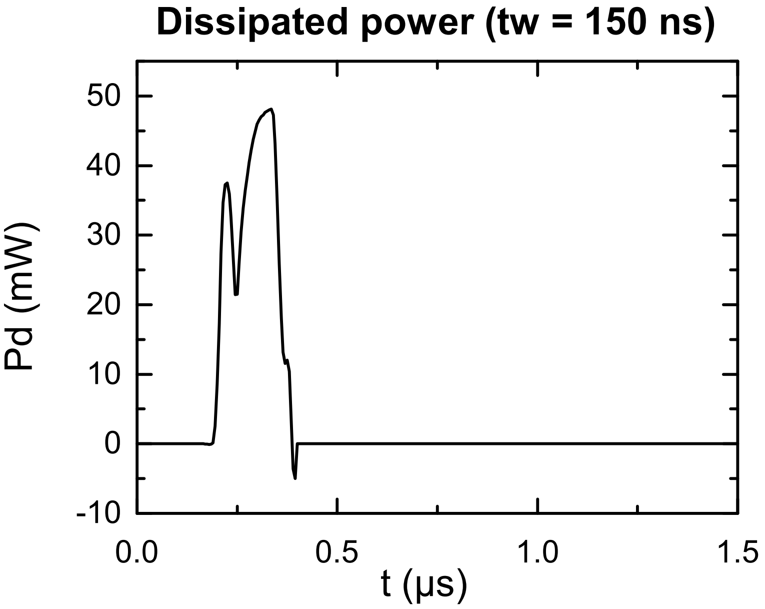


Figure 3.30: Instantaneous dissipated power when a pulse is applied ($150\,ns$). The entire period is shown.

We can thus write the current trend as

$$I_C(t) = I_{sat} - (I_{sat} - I_{iso}) \cdot \exp\left(-\frac{t - t_o}{\tau_{TH}}\right)$$

and use this function as our fitting function. In Fig. 3.31 a detail of the waveform of the reference B55 ($A_E = 0.2 \cdot 5\,\mu m^2$) reaching its saturated level is shown along with the fitting curve described by the previous function (in red). Please note that the fit models the behaviour at the steady condition correctly (the red curve superposes to the black one already from $t = 0.5\,\mu s$, and sticks to it for more than other $9\,\mu s$), but it does not precisely follow the black curve very close to the isothermal I_{iso} value: the thermal constant we get is actually an estimation, that is affected by an error.

Fig. 3.32 and Fig. 3.33 show the same procedure for the standard devices from III-V Lab and ETHZ. In these cases we observe that the fitting function provides a poorer estimation – in particular for III-V Lab. The main cause appears to be that the current continues to slightly increase even after several

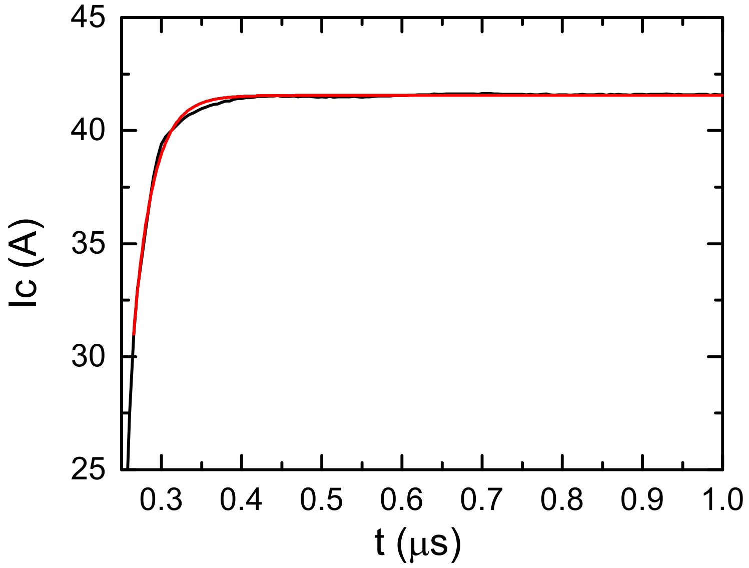


Figure 3.31: Fitting for the extraction of the thermal constant for the reference B55 (in red). $V_{BE} = 1\,V$, $V_{CE} = 1.8\,V$.

microseconds, not reaching a true steady value indeed. This error can be viewed as a marker for the insufficient description of the thermal behaviour through a single RC network.

In Table 3.1 are shown the estimated value of τ_{TH} and its standard error for each case. The results confirm what has been said qualitatively. The error is minimum for B55, namely we got a good estimate of the thermal constant through a single-pole equivalent thermal network, whereas the largest is for III-V Lab.

As to the value of τ_{TH}, for the SiGe HBT the parameter is definitely lower that the InP-based devices, which are at least on the same order of magnitude instead. A big thermal constant means that the rise of I_C due to the heat related to that device is slower than another with a smaller constant. This is again in accordance with what has been aforementioned.

This important result allows us to retrieve the value of C_{TH} for a single-pole thermal network by knowledge of R_{TH}, which will be extracted in the

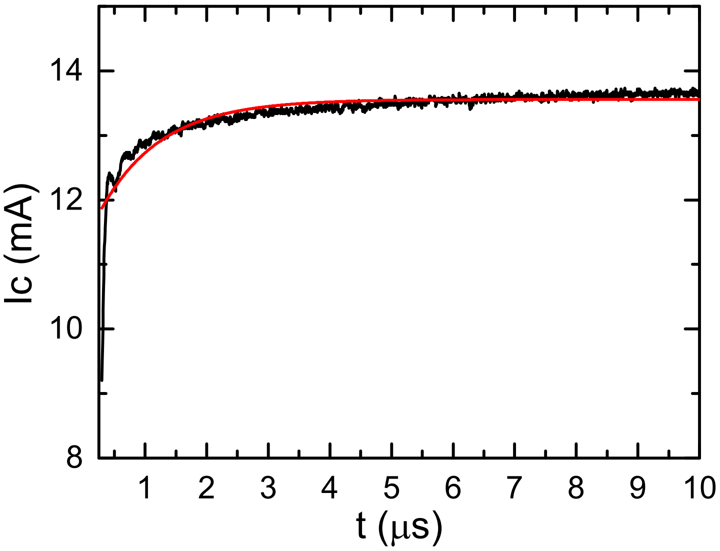


Figure 3.32: Fitting for the extraction of the thermal constant for the reference III-V Lab (in red). $V_{BE} = 0.9\,V$, $V_{CE} = 1.3\,V$.

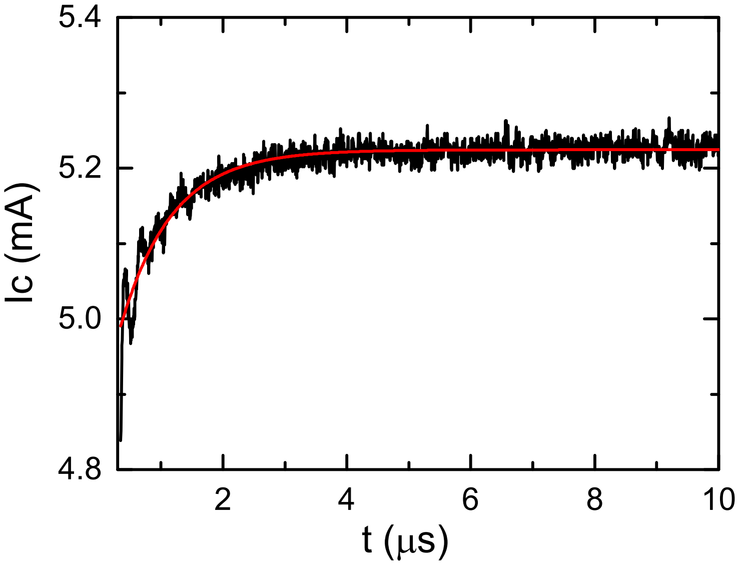


Figure 3.33: Fitting for the extraction of the thermal constant for the reference ETHZ (in red). $V_{BE} = 0.9\,V$, $V_{CE} = 1.5\,V$.

B55		**III-V Lab**		**ETHZ**	
$\boldsymbol{\tau_{TH}}$	**s.e.**	$\boldsymbol{\tau_{TH}}$	**s.e.**	$\boldsymbol{\tau_{TH}}$	**s.e.**
0.025	$2 \cdot 10^{-4}$	0.985	0.021	0.83	0.013

Table 3.1: Extracted thermal constants (in μs) and the respective standard errors. All τ_{TH} refers to the reference devices.

next chapter.

Chapter 4

Thermal Resistance Extraction and Analysis

We have learned that the self-heating thermal resistance is a fundamental parameter of the characterisation of the thermal behaviour of our devices. In this chapter, we will detail an extraction method of R_{TH}, that has been used on the different geometries and technologies.

The extraction method, based on [39], will first be simulated and validated by comparison with the HICUM model temperature-dependent R_{TH}. Then it will be used for drawing the thermal resistance changing over the junction temperature from actual measures.

4.1 Definition of the Thermal Resistance

Although there have been several methods in the past for extracting the thermal resistance out of a bunch of straightforward measurements of the DC electrical characteristics [40], it is more cumbersome in small modern devices to retrieve an accurate estimate of it, since when dealing with high-current HBTs, the junction temperature rise cannot be neglected in determining the thermal resistance, namely $R_{TH} = R_{TH}(T_j)$.

When the current density is low, so is P_d, and the approximation $T_j \simeq$

T_{amb} is valid; R_{TH} is some hundred kelvin over watt, depending on the type of device, and with a small dependence on temperature. Though, as soon as the power dissipated within the device grows and the temperature as well, following $\Delta T = R_{TH}\, P_d$, as defined in Eq. (2.17), the previous approximation of κ, the thermal conductivity, as temperature-independent parameter proves no longer valid. Its non-linear changing makes the thermal resistance deviate substantially from its rather constant value, effectively increasing R_{TH} with temperature or dissipated power (equivalently).

For the moment, let's discuss some results on the thermal resistance stemming from the definitions made in Chapter 2. In Joy and Schlig's work [27], it has been derived from the definition of R_{TH} in Eq. (2.16) that the time-independent thermal resistance at the top center of the heat source can be written as

$$R_{TH} = \frac{1}{2\kappa\sqrt{L_E\, W_E}}\, f(d)\, f(a)\, f(s) \tag{4.1}$$

In that paper, the heat source (that is located at the CBJ) is modelled as a solid rectangular parallelepiped at a distance d from the surface, the thickness of which (the space charge region) is s and with an aspect ratio equal to that of the emitter area ($a = W_E/L_E$). $f(d)$, $f(a)$ and $f(s)$ are functions of those geometrical parameters, respectively.

This result has been further simplified, since for many devices the dimensional relations lead to a constant, namely

$$R_{TH} = \frac{1}{4\kappa\sqrt{L_E\, W_E}} \tag{4.2}$$

It must be noted that some assumptions have been taken by the authors. The top surface, as already said before, is adiabatic; no effect of conduction through interconnect metallization is considered, nor shallow trenches limiting the heat flux; the power dissipation inside the heat source is independent of the position; the medium is homogeneous. We are therefore dealing with a heat source submerged inside a semi-infinite block of semiconductor.

Now, let's expand a discussion partially begun in [41]. If we consider the model depicted in Fig. 2.23, that we used to introduce the electrical equivalent circuit, and we sum up all the discrete contributions of $R_{TH}(z)$ as defined in Eq. (2.18), but in its differential form

$$dR_{TH}(z) = \frac{1}{\kappa\, A(z)}\, dz$$

we get the thermal resistance of the entire semi-space of the semiconductor block in which heat flows. In other words, we do not take into account the distributed behaviour of the heat diffusion from the source at temperature T_j to the substrate at T_{amb}, but this is needed to find a single value of R_{TH} that sticks to the definition $(T_j - T_{amb} = R_{TH}\, P_d)$.

So R_{TH} is simply the integral of the infinitesimal contributes

$$R_{TH} = \int_0^h dR_{TH}(z) = \int_0^h \frac{1}{\kappa\, A(z)}\, dz$$

or, more explicitly,

$$R_{TH} = \frac{1}{\kappa} \int_0^h \frac{1}{(L_E + 2z\, \tan\beta)\,(W_E + 2z\, \tan\alpha)}\, dz \tag{4.3}$$

where h is the depth of the substrate, from source to sink, and α and β are the angles by which heat spreads in the emitter length and width directions, respectively (Fig. 4.1).

For brevity, the algebraic steps are omitted. The result is

$$R_{TH} = \frac{1}{2\kappa\,(W_E\, \tan\beta - L_E\, \tan\alpha)}\, \ln\left(\frac{A_E + 2h\, W_E\, \tan\beta}{A_E + 2h\, L_E\, \tan\alpha}\right)$$

by which, when considering a semi-infinite space (which effectively means $h \gg L_E/2\tan\beta$ and also $h \gg W_E/2\tan\alpha$), we get

$$R_{TH} = \frac{1}{2\kappa\,(W_E\, \tan\beta - L_E\, \tan\alpha)}\, \ln\left(\frac{W_E\, \tan\beta}{L_E\, \tan\alpha}\right) \tag{4.4}$$

In Fig. 4.2 we compare the calculated R_{TH} in Eq. (4.2) and Eq. (4.4) – normalized in such a way that it is only a function of the emitter area – for some typical dimensional values we analyse in this thesis. We notice that for the chosen couple of angles, we obtain a good agreement between the two derivations of R_{TH}.

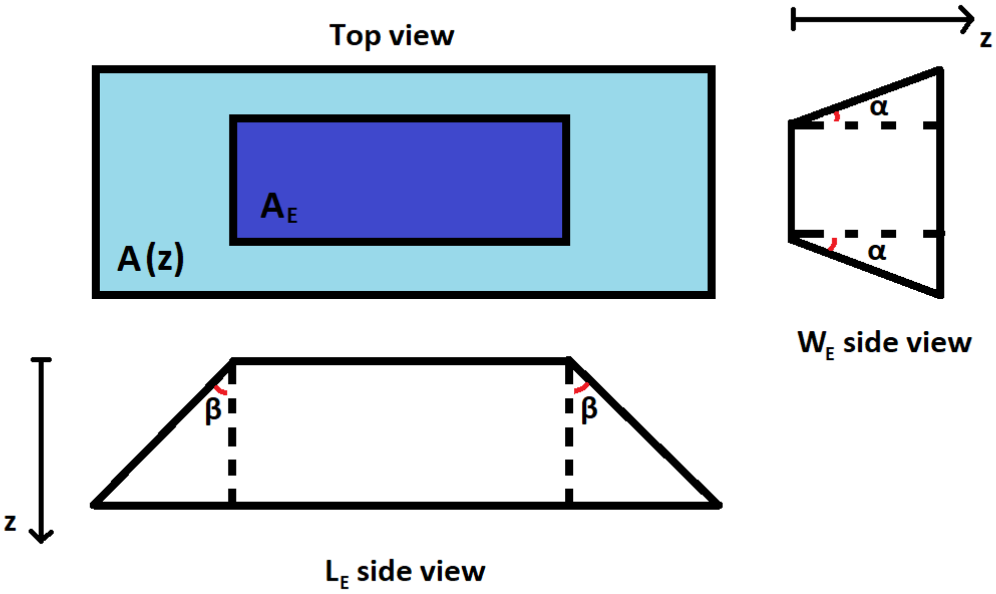


Figure 4.1: Top and side views of the area $A(z)$ and the angles for its calculation. $A(z)$ is the whole area (light and dark coloured).

As a matter of fact, though, this approach is incomplete because the thermal conductivity depends on temperature, $\kappa = \kappa(T)$. A more realistic $\kappa(T)$ is defined in [42]

$$\kappa(T) = \kappa_{ref}\left(\frac{T}{T_{ref}}\right)^{\lambda} \tag{4.5}$$

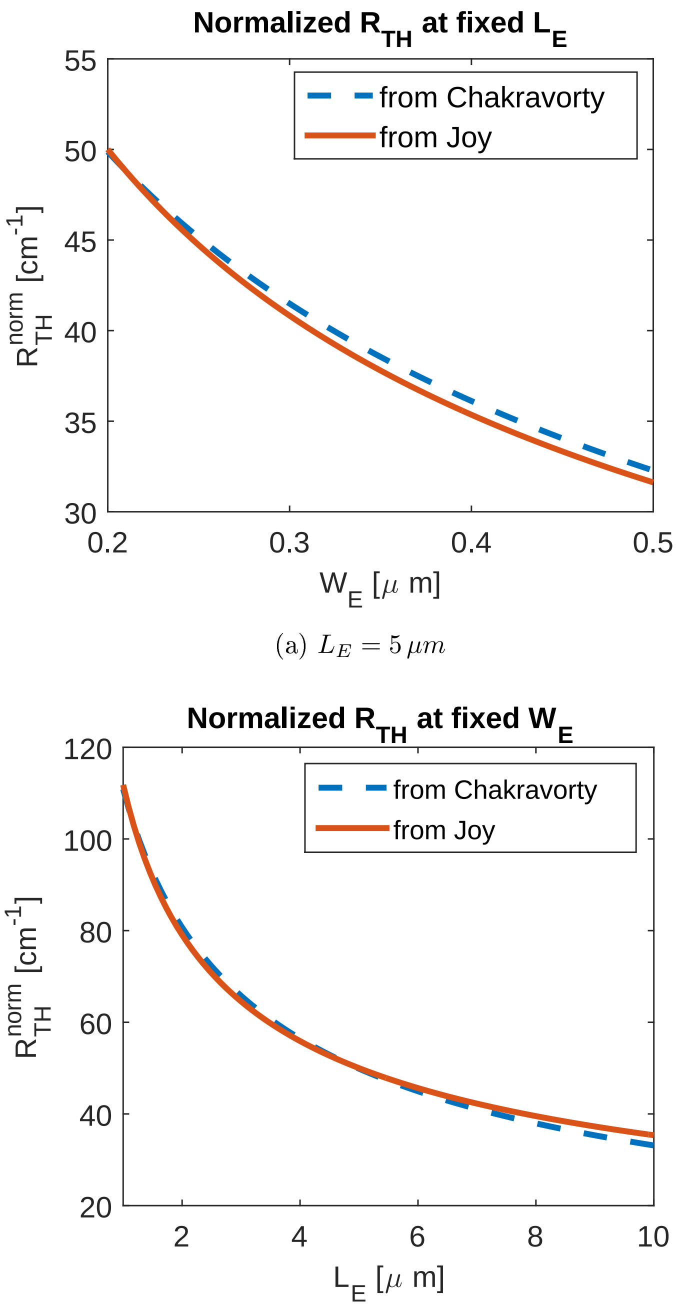
where κ_{ref} and λ are material-dependent constants and $T_{ref} = T_{amb} = 300\,K$. From this, R_{TH} turns out to be a function of the temperature rise or the dissipated power. Its dependence can be considered like in [43] as

$$R_{TH} = R_{THo}\,(1 + \epsilon\,P_d) \tag{4.6}$$

or, like done in the HICUM model [3] as

$$R_{TH} = R_{THo}\,[1 + \alpha\,(T_j - T_{amb})] \tag{4.7}$$

where R_{THo} is the temperature-independent thermal resistance (found at $T_j = T_{amb}$), α and ϵ depending on the material alone. Eq. (4.6) and Eq. (4.7) are substantially equal since both express the same idea: more dissipated power has as a consequence a temperature rise.



(a) $L_E = 5\,\mu m$

(b) $W_E = 0.2\,\mu m$

Figure 4.2: Normalized thermal resistance ($R_{TH}^{norm} = 2R_{TH}\,\kappa$) vs. a typical range of value of L_E and W_E. Joy [27] refers to Eq. (4.2); Chakravorty [41] refers to Eq. (4.4). $\alpha = 30°$, $\beta = 80°$.

A precise expression of the temperature rise is more complicated [43] and would require some iterations to find the R_{TH}. It can be expressed as

$$\Delta T = T_{amb} \left[\left(1 - R_{THo}\, P_d\, \frac{m-1}{T_{amb}} \right)^{1/1-m} - 1 \right] \tag{4.8}$$

where $m \leq 0$ is a material-dependent parameter. Please notice that if $\kappa(T) = \kappa$, $m = 0$, and ΔT is again like in Eq. (2.17) – that is also the same of making the first-order approximation of the Taylor series of Eq. (4.8). The second-order approximation of Eq. (4.8), on the other hand, yields and justifies Eq. (4.6), with $\epsilon = mR_{THo}/2T_{amb}$. The Taylor series can be expanded only if

$$m > 1 - \frac{T_{amb}}{P_d\, R_{THo}}$$

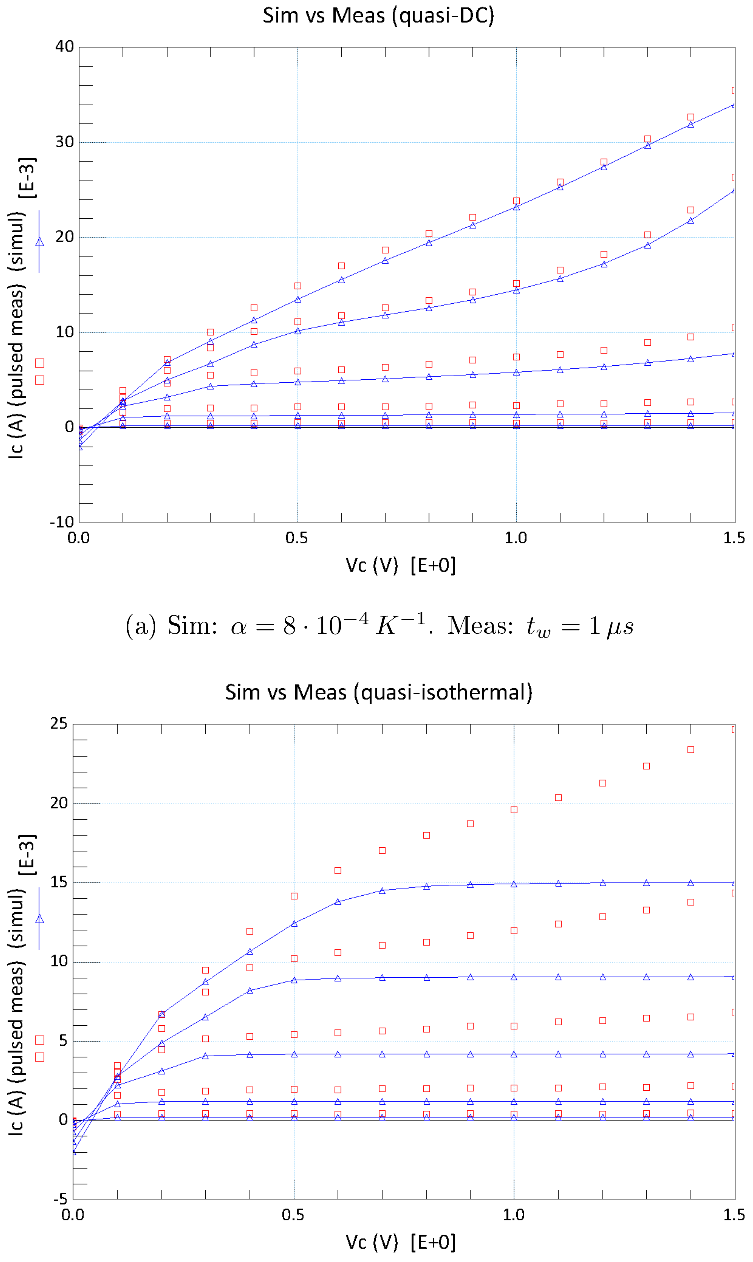
α in Eq. (4.7) is chosen by adjusting the simulated curves to measurements. Fig. 4.3 (a) shows that there is good correspondence between our HICUM model, simulated with $\alpha = 8 \cdot 10^{-4}\, K^{-1}$ and real pulsed measurements on B55, with $t_w = 1\, \mu s$. In Fig. 4.3 (b) is what happens when comparing the simulation with no thermal resistance at all, $R_{THo} = 0$, with real measures where self-heating is strongly reduced, $t_w = 150\, ns$. The discrepancy is remarkable, meaning that the thermal resistance can not be neglected completely – isothermal measures are impossible.

4.2 The Intersection Method

The intersection method for extracting R_{TH} has been already discussed in [39] and is now being implemented on IC-CAP for a practical elaboration of both simulations and measurements. The algorithm follows the theoretical strategy which will be detailed below. Both the electrical feedback (Early effect) and avalanche effects are neglected: for our devices and bias conditions, this is a legitimate choice.

In the active region, the T_j-dependent collector current can be written as [39]

$$I_C(V_{BE}, T_j) = I_s \, \exp \left(\frac{q\, V_{BE}}{k\, T_j} \right) \tag{4.9}$$



(a) Sim: $\alpha = 8 \cdot 10^{-4}\,K^{-1}$. Meas: $t_w = 1\,\mu s$

(b) Sim: $R_{TH} = 0\,K/W$. Meas: $t_w = 150\,ns$

Figure 4.3: Effect of α in the HICUM model. $V_{BE} = 0.75 - 0.95\,V$, $V_{CE} = 0 - 1.5\,V$. Reference B55.

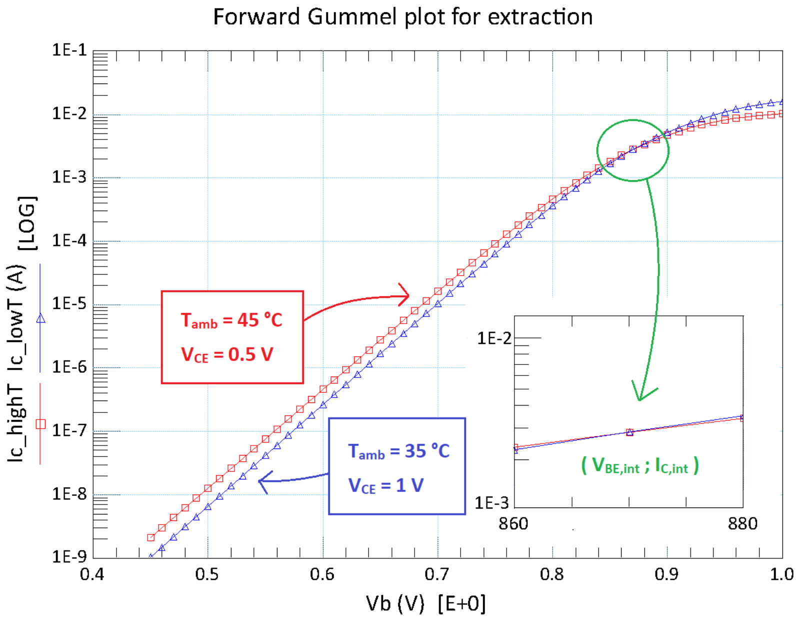


Figure 4.4: High-temperature current: $I_{C,H}$, low-temperature current: $I_{C,L}$. Simulation of the reference B55 at different T_{amb} and V_{CE}. The intersection point is shown.

The strategy consists in taking into account two forward Gummel plots ($I_C - V_{BE}$ curves, where I_C is displayed in logarithmic scale) at two different thermal chuck temperatures T_{amb} and collector-emitter voltages V_{CE}.

Fig. 4.4 shows that an intersection will effectively take place when the Gummel curve simulated – or measured – at lower temperature has the higher V_{CE} bias, and vice versa. In the following, the "$_H$" and "$_L$" subscripts stand for the high and low chuck – ambient – temperatures, respectively.

The T_{amb} step and the two values of V_{CE} need to be chosen properly in order not to push the intersection to a point where Kirk effect dominates. Fig. 4.5 shows what happens when at least one of the collector voltages is too high. This situation should be avoided.

In the Appendix, the IC-CAP code used to find the wanted intersection is detailed. Of course, the samples of the measured current do not allow an exact extraction of the intersection point. What is done is to calculate the

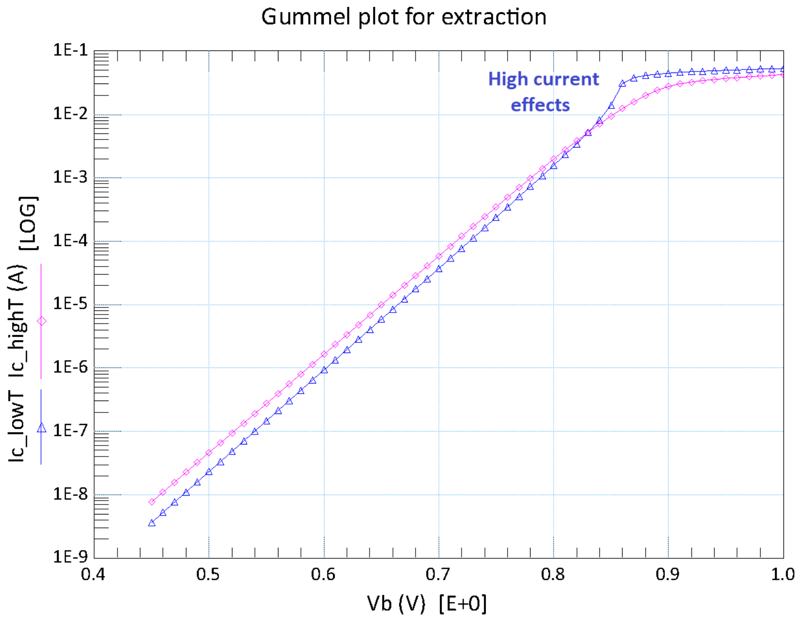


Figure 4.5: $T_{amb,H} = 45°C$, $V_{CE,H} = 1\,V$; $T_{amb,L} = 35°C$, $V_{CE,L} = 1.5\,V$.

absolute value of the difference between the current at high temperature and the one at low temperature, namely $\Delta I_C = |I_{C,H} - I_{C,L}|$ (it is plotted in Fig. 4.6).

Since the two plots intersect, there must be a local minimum, corresponding to the smallest difference between the samples of the two curves (see Fig. 4.6). The algorithm finds the corresponding V_{BE} value ($V_{BE,min}$), alongside the two I_C's ($I_{C,min}$).

At this point, the algorithm searches for the second local minimum, i.e. the second smallest value of ΔI_C, right before or after the newly found $V_{BE,min}$, and compares the two values. The idea is that the real intersection point should lay between the minimum and the second minimum points.

Once the corresponding $V_{BE,min,2nd}$, $I_{C,H,min,2nd}$ and $I_{C,L,min,2nd}$ values are retrieved, a linear interpolation is performed between the two extracted consecutive samples. That means, a line is drawn to connect the minimum and the second minimum of the $I_{C,H} - V_{BE}$ curve. The minimum and the second minimum of the $I_{C,L} - V_{BE}$ curve are also connected. An intersection

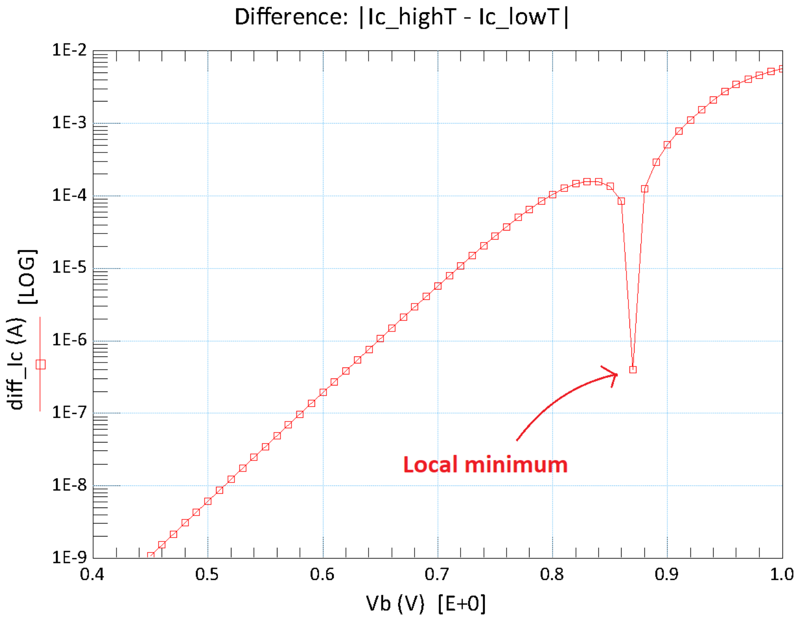


Figure 4.6: ΔI_C vs. V_{BE}. The minimum is indicated.

point – which does not correspond to any of the samples – is finally found.

So a single common point, the "real" intersection point, $(V_{BE,int}; I_{C,int})$, is finally calculated. From Eq. (4.9), since both I_C and V_{BE} are the same for the two plots, so must be T_j and consequently $R_{TH}(T_j)$.

By equating

$$\begin{aligned} T_j &= T_{amb,H} + P_{d,H}\, R_{TH} = T_{amb,H} + I_{C,H}\, V_{CE,H}\, R_{TH} \\ T_j &= T_{amb,L} + P_{d,L}\, R_{TH} = T_{amb,L} + I_{C,L}\, V_{CE,L}\, R_{TH} \end{aligned} \tag{4.10}$$

where at this point $I_{C,int} \simeq I_{C,H} \simeq I_{C,L}$, then we can compute and extract R_{TH} as

$$R_{TH} = \frac{T_{amb,L} - T_{amb,H}}{P_{d,H} - P_{d,L}} \tag{4.11}$$

To find the T_j which corresponds to this retrieved value of R_{TH}, one simply needs to substitute it in Eq. (4.10). This process is automatically repeated by our algorithm for all the input temperatures – stepped by $10°C$, conventionally –, giving a $R_{TH} - T_j$ curve, which is not constant, as foreseen.

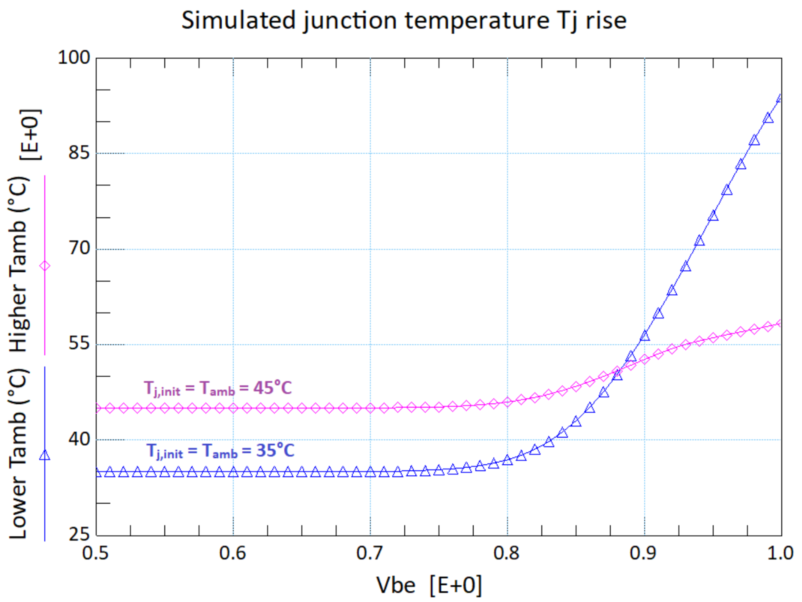


Figure 4.7: Simulated T_j vs. V_{BE} with the same set-up used for finding the intersection (Fig. 4.4). Blue: $T_{amb} = 35°C$, $V_{CE} = 1\,V$; purple: $T_{amb} = 45°C$, $V_{CE} = 0.5\,V$. The simulation is carried out using a distributed network.

A simulation of the junction temperature rise is plotted in Fig. 4.7; we can see that, when the device is off, $T_j = T_{amb}$ and that, since the bias for the low-temperature curve is higher ($1\,V$), T_j grows more when the device enters the active region.

4.2.1 Validation of the Method

First it is necessary to simulate the HICUM model in the IC-CAP environment, plot the Gummel curves at different pairs of ambient temperature and collector-emitter voltage and retrieve some couples of $(R_{TH}; T_j)$ as exposed so far. Please notice that the HICUM model simulates the $I_C - V_{BE}$ performance considering a thermal resistance with a dependence of temperature of the kind of the one described in Eq. (4.7) for a realistic thermal feedback.

Then, we simply plot the $R_{TH}(T_j)$ function as denoted in Eq. (4.7), with $T_{amb} = 300\,K$, $\alpha = 8 \cdot 10^{-4}\,K^{-1}$ and R_{THo}, the nominal thermal resistance,

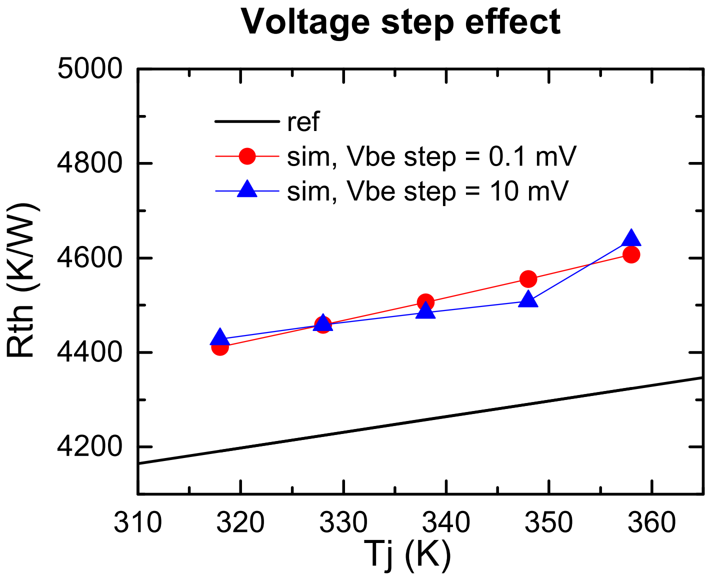


Figure 4.8: Comparison between the intersection method and the reference, the simulations being performed on differently stepped Gummel plots. Reference B55.

depending on the DUT.

In Fig. 4.8 we can see the reference line plotted along with the simulation dots extracted with our method. The blue dots differ from the red ones in that a different sampling step is applied to the Gummel curves. The red curve derives from a $V_{BE} - I_C$ plot with a V_{BE} step of $10\,mV$. This is a realistic step and in fact is the same achieved by our measure equipment. The red curve, on the other hand, is traced out with a simulated $V_{BE} - I_C$ sampled every $100\,\mu V$, a step which is not possible to reach with our instrumentation.

This also shows the limit of the linear interpolation, since the error in the sample values in blue comes from an inaccurate determination of the real intersection point. A more advanced interpolation technique, such as Spline interpolation, would give as a result a line of the kind of the red one.

We can see, however, that the percentage difference is minimal, and limited to only few points (in this case, the error is just 1% for the sample at $348\,K$). The overestimate of the simulation with our extraction method is

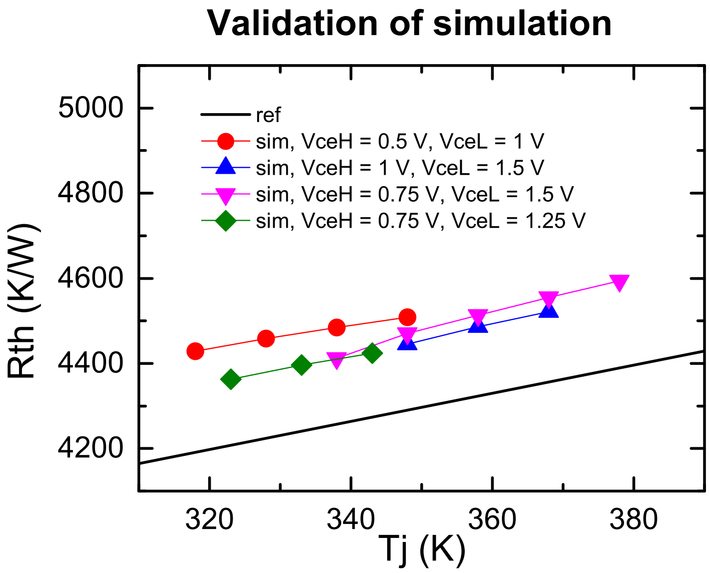


Figure 4.9: Comparison between the intersection method and the reference, with different choices of V_{CE}. Reference B55.

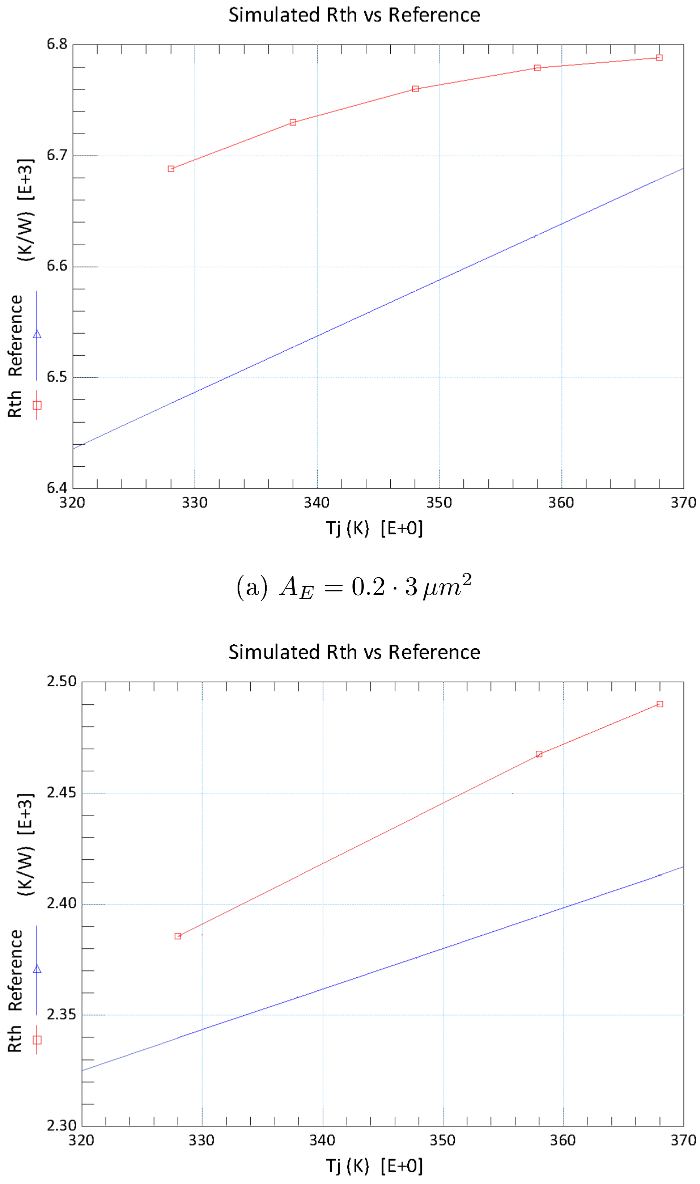
definitely greater (about 5% more than the reference), and this is because of the assumptions involved.

In Fig. 4.9 the slight dependence of the choice of V_{CE} in our method is shown. Even though high current effects may alter the measures in some cases, thus forcing us to exclude some extracted R_{TH} values, choosing a high $V_{CE,L}$ value with a small difference between $V_{CE,H}$ and $V_{CE,L}$ (see the blue dots) seems to provide the most accurate result. This is due to a reduced presence of the Early effect in our retrieval method.

Two different geometries are simulated and the extraction technique is carried out too, in Fig. 4.10, just to make the validation general.

4.2.2 Extraction from Measurements

Thanks to the good results that this strategy provides, we are now able to apply the method to real measurements. We did so for a number of different geometries, as usual. In Fig. 4.11 the symbols represent the retrieved



(a) $A_E = 0.2 \cdot 3\,\mu m^2$

(b) $A_E = 0.2 \cdot 10\,\mu m^2$

Figure 4.10: Comparison between the intersection method and the reference, for two different B55 technology geometries.

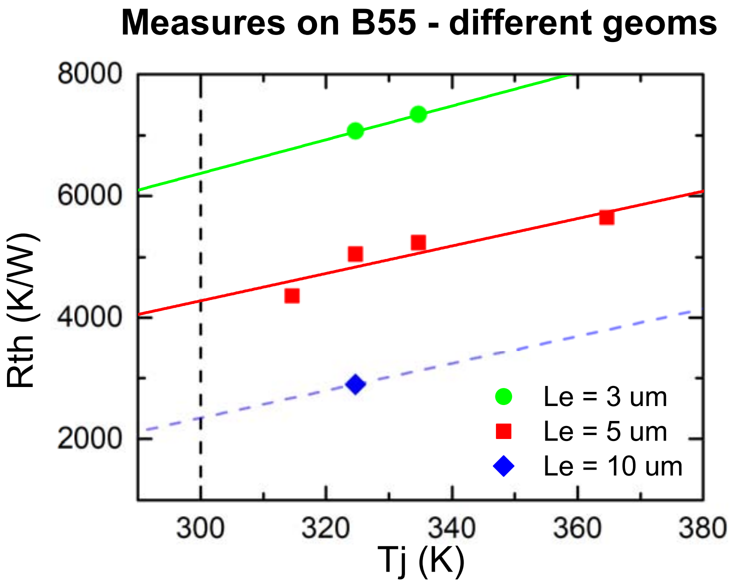


Figure 4.11: Measured T_j dependence of R_{TH} for different B55 geometries. Dashed line at $300\,K$ for extracting R_{TH} at ambient temperature. $W_E = 0.2\,\mu m$.

$(R_{TH}; T_j)$ couples, while the lines come from a linear interpolation of the measured points themselves.

It is clear that, with the method we are applying, n temperature measurements will provide $n-1$ symbols. From these $n-1$ symbols, then, we need to select and keep the ones which lay on an ideal straight line and ignore unlikely overestimates or underestimates. The more measures we make, the more accurate will be the results.

We dispose of many temperature measures for our reference device ($L_E = 5\,\mu m$) – from $15°C$ to $95°C$ every $10°C$, thus 7 points, of which just 4 have a reasonable position. On the other hand, for the others we only dispose of 4 measurements (3 symbols).

A problem arises for the device with $L_E = 10\,\mu m$: only one symbol can be chosen. The situation is explained in Fig. 4.12. In (a), all the symbols are shown, and we can see that the excluded symbols effectively have unrealistic R_{TH} values. This situation happens because of the linearity of the interpolation we have used and the quality of measures themselves. As shown in (b) where ΔI_C is plotted against V_{BE}, we see that no local minimum can be found by the algorithm: the sampling is too large and does not get close to the real intersection.

So just one single point, for which this situation does not happen, is relevant. We also notice that the slope of the $R_{TH}-T_j$ curves decreases as the dimensions increase, the reason of which is clear from Eq. (4.7). According to that equation, the slope is given by $\alpha\, R_{THo}$, α being practically constant since it depends on the material, so the slope grows proportionally to R_{THo} (even though the proportionality is not direct, due to the approximations involved on Eq. (4.7)).

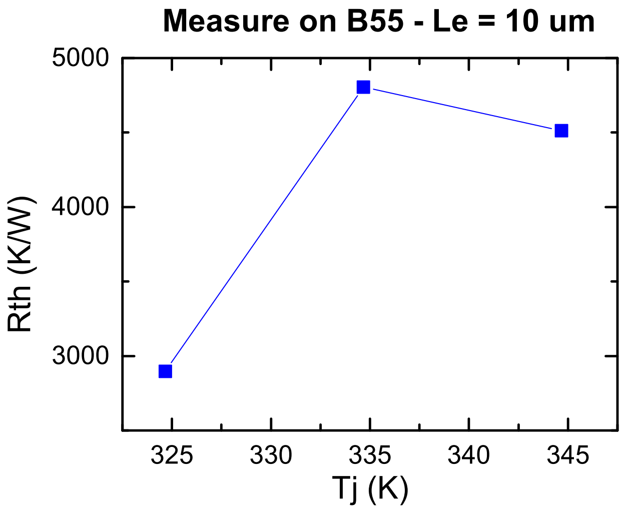
Speaking of the slope furthermore, the use of this intersection method proves, for high values of T_j, a more modest deviation of R_{TH} than other methods like [44] in which the extracted R_{TH} also includes the effect of other resistances, such as the collector and the emitter resistances.

The extraction of R_{TH} at ambient temperature – i.e. R_{THo}, extracted at $300\,K$ in Fig. 4.11 – can be carried out for the three geometries. The idea pursued for the device of which we only dispose of one measure, is to give at least an upper bound (dashed blue line) for the slope, since it should lean less than the red line.

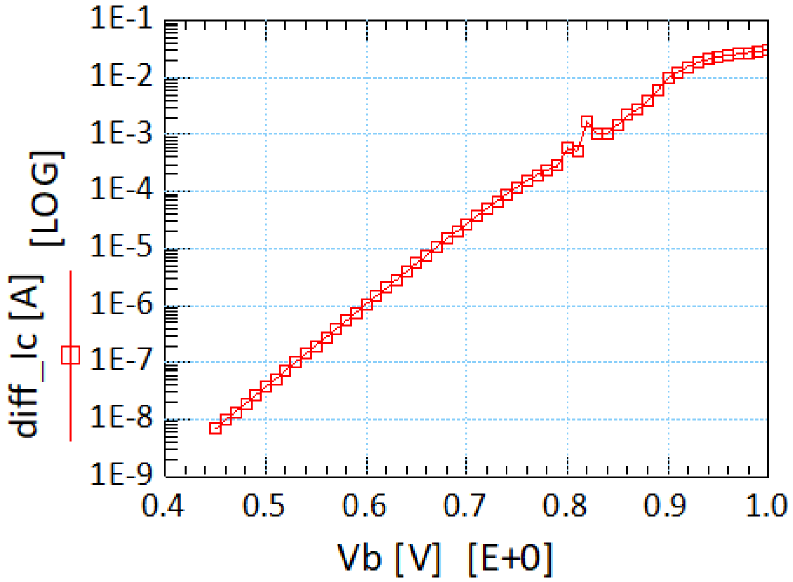
In Fig. 4.13 we plot R_{TH} at room temperature as a function of the length of the emitter, and compare the value retrieved by our method with the nominal R_{THo} used in the HICUM model. Also the relative error is shown in figure. In all cases, there is an overestimate, as foreseen by the simulations.

For the III-V Lab devices we dispose of 5 temperature measures. As we see in Fig. 4.14, only one measure has been excluded for each of the two largest devices of this technology: the measures are more accurate. Again, as the dimensions increase, R_{THo} lowers as well as the slope.

We performed the extraction at $300\,K$ once again and the results are shown in Fig. 4.15. In general, the thermal resistance is smaller in these InP/InGaAs devices than the SiGe:C HBTs. For $L_E = 5\,\mu m$, R_{TH} is $1\,kK/W$



(a) Measured T_j dependence of R_{TH}.



(b) ΔI_C vs. V_{BE}. No local minimum is found.

Figure 4.12: Bad result of the intersection method. $\Delta I_C - V_{BE}$ in (b) was used to find R_{TH} at $T_j = 334.7\,K$ in (a). A similar $\Delta I_C - V_{BE}$ is for R_{TH} at $T_j = 344.7\,K$. B55: $L_E = 10\,\mu m$, $W_E = 0.2\,\mu m$.

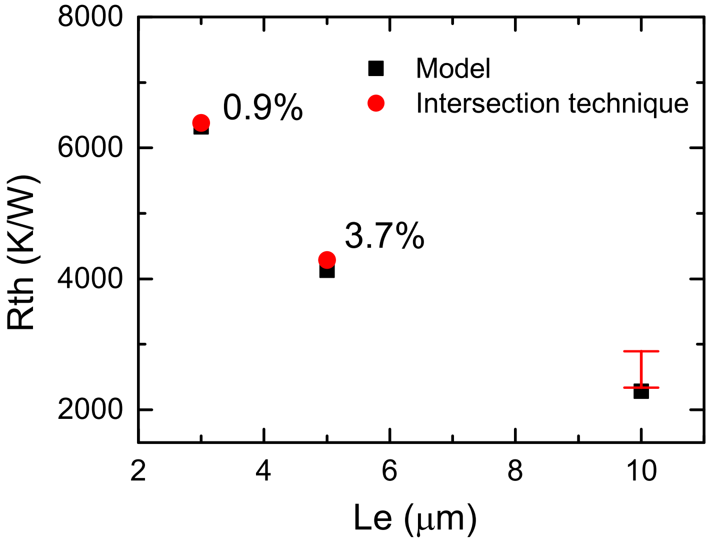


Figure 4.13: R_{TH} extracted at $300\,K$ for all geometries (comparison between the one from the intersection method and the one used in the HICUM model). Relative error is displayed. B55: $W_E = 0.2\,\mu m$.

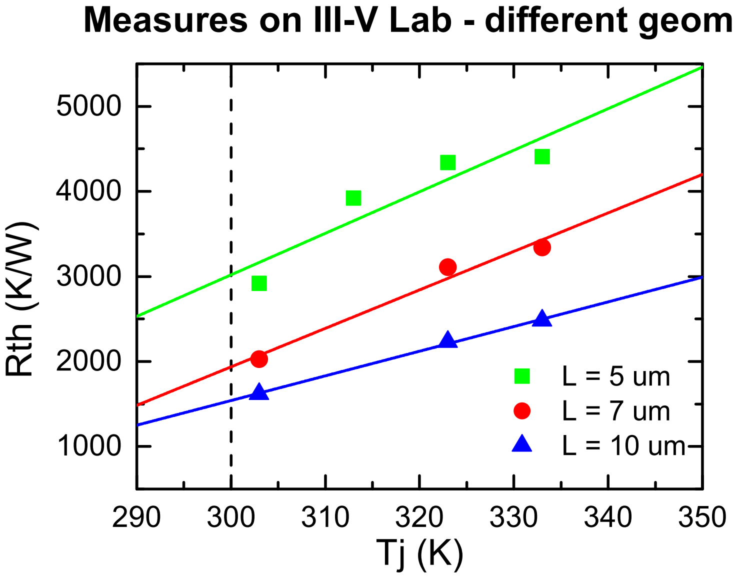


Figure 4.14: Measured T_j dependence of R_{TH} for different III-V Lab geometries. Dashed line at $300\,K$ for extracting R_{TH} at ambient temperature. $W_E = 0.7\,\mu m$.

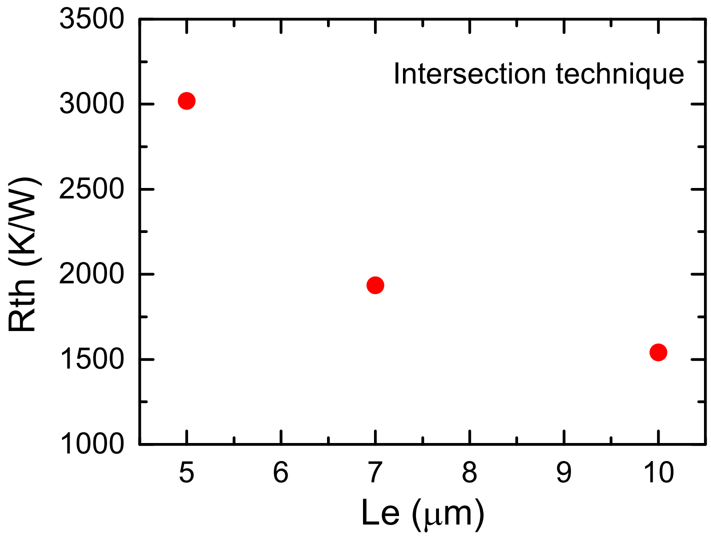


Figure 4.15: R_{TH} extracted at $300\,K$ for all geometries. III-V Lab: $W_E = 0.7\,\mu m$.

smaller in III-V Lab than B55, whereas the width is 3.5 times larger.

4.2.3 Limits of the Analysis

We might want to go a step further and find the value of κ_{ref} from our results, for example by reversing Eq. (4.2) now that we have extracted R_{THo}. Besides the lack of absolute precision of our measurement and extraction procedure, we should also remark that the equation we would apply is an approximation of the more general one, Eq. (4.1), and it is much more cumbersome to compute the values of the geometrical functions for each specific device.

In Fig. 4.16 we tried an even more radical approach, namely to use Eq. (4.2) to calculate each $\kappa(T_j)$ corresponding to the $R_{TH}(T_j)$'s extracted with our intersection method. Then, we interpolated the dots with a non-linear equation of the kind of Eq. (4.5) to find κ_{ref} and λ. In figure, we have also plotted the real $\kappa - T_j$, according to [42].

κ_{ref} might be found by reading the value of the interpolation curves at

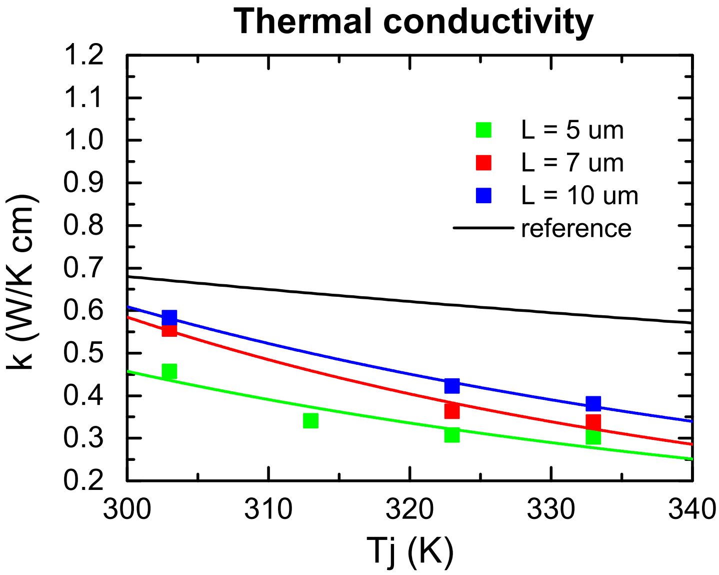


Figure 4.16: κ values calculated by reversing Eq. (4.2) for all geometries. III-V Lab: $W_E = 0.7\,\mu m$. Reference is the real κ trend according to [42].

$300\,K$. That confirms the lack of precision we have already anticipated with respect to the actual one. The decreasing rate of the curves represents λ, the absolute value of which is much higher than the reference one. Also note that κ and λ should be geometry-independent, whereas all the interpolation curves differ one from another.

The comparison between technologies (Fig. 4.17) also shows its limits. The extracted κ_{ref} for III-V Lab is $0.61\,W/Kcm$, instead of $0.68\,W/Kcm$ and the extracted λ is -4.66 instead of -1.4. The values of κ and λ in B55 depend on x, the concentration of germanium of $\mathrm{Si}_{1-x}\mathrm{Ge}_x$. The extracted parameters are $\kappa_{ref} = 0.59\,W/Kcm$ and $\lambda = -1.64$. According to [45], our value of κ_{ref} would lead to a concentration $x < 0.2$ (if our SiGe:C HBT was made of an undoped alloy, which is also not true since it is carbon-doped).

Extraction of the Thermal Capacitance

Using a single-pole thermal network, it is trivial for us to find C_{TH}, when knowing R_{TH}, since it's sufficient to take τ_{TH} from Table 3.1 and compute

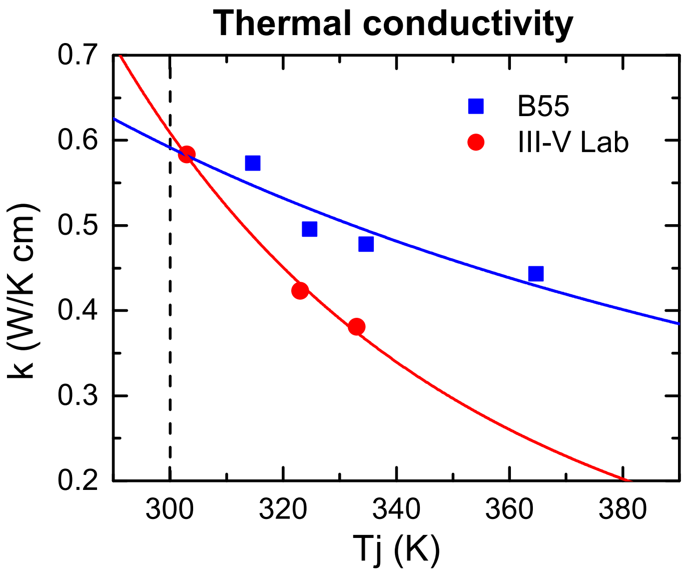


Figure 4.17: κ values calculated by reversing Eq. (4.2). One selected device for each technology.

$C_{TH} = \tau_{TH}/R_{TH}$. Since R_{TH} depends on geometry, so does the capacitance.

Nevertheless, as explained in this chapter, R_{TH} depends on the dissipated power, too. When it is small, we may think of approximating $R_{TH} \simeq R_{THo}$. In this case, what we said in the previous chapter, that is, for instance, $\Delta T(t)$ is just $P_d(t)$ scaled by R_{THo} would be true (Fig. 3.28).

But in the examples from the previous chapter that we used to find the thermal constant, P_d is far from being small. At the steady state, the dissipated powers for the reference devices are $P_d(B55) = 72\,mW$ and $P_d(IIIV) = 17.5\,mW$ at the following bias conditions: $V_{BE} = 1\,V$, $V_{CE} = 1.8\,V$ (B55) and $V_{BE} = 0.9\,V$, $V_{CE} = 1.3\,V$ (III-V Lab).

So, at steady state, $R_{TH} = \Delta T/P_d$ with ΔT being like in Eq. (4.8). But with our extraction method we cannot find m of that equation, since it would be necessary to extract R_{TH} vs. P_d. Thus, we cannot compute the temperature rise at $P_d(B55)$ and $P_d(IIIV)$ nor, consequently, find the corresponding thermal resistances. Any C_{TH} we calculate with the information at our disposal would be imprecise for the case considered. Anyway, this can be done within the compact model.

Chapter 5

Further Thermal Analysis

5.1 Thermal Impedance Extraction

In the last chapter, we have come up with a way to calculate the thermal impedance once the thermal resistance was extracted by using the previously estimated τ_{TH}. Thus, the thermal resistance and capacitance of the single-pole thermal network were defined.

We have seen though that the value of C_{TH} cannot be properly found in some situations, like the case of a highly temperature-dependent R_{TH}. The other limit that exists even when we actually manage to calculate, is that a conventional single-pole network does not provide a sufficient approximation of the thermal behaviour of a device for compact model simulations.

As widely discussed in [28], a recursive network (Fig. 5.1) proves to be much more reliable in terms of accuracy of description of the thermal phenomena. The dynamic self-heating should be modelled as at least a nine-stage recursive network [33]. Here, the resistances and capacitances are distributed; the series of all resistances at $f = 0\,Hz$ yields R_{TH}.

The normalized thermal impedance $Z_{TH}(s)$ has a much more complex expression [28] than Eq. (2.19) and is computed using y-parameters. Those, in turn, can be retrieved by measuring the low frequency S-parameters using a vector network analyser (in a frequency range that goes from few kilohertz

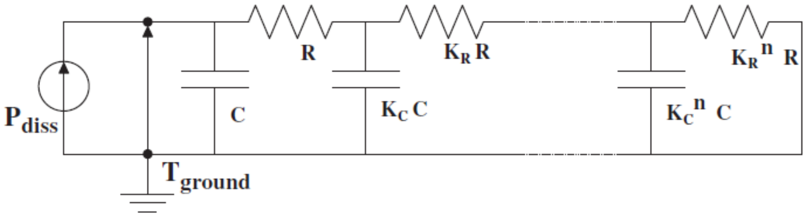


Figure 5.1: A n-cell recursive thermal network.

to few gigahertz).

An on-wafer SOLT calibration is first done. A calibration is made on devices with known properties in order to know the error terms (correction data) and define an error network (linear error model) separated from the ideal network analyser [46].

In particular, the SOLT calibration on a two-port coplanar technology is achieved by first contacting the GSG probe on a short that provides total reflection, which is actually modelled in the network analyser by a series of parasitic inductances for frequencies above a few multiples of $10\,GHz$. Then an open structure is contacted, and again parasitic elements – capacitances – have to be considered. A device with two $100\,\Omega$ impedances between the signal and ground pins at each port is used as a load. Finally, the GSG fingers at port one and two are shorted in a device providing a "through".

The measurements are performed at room temperature by applying an input RF power of $-30\,dBm = 1\,\mu W$. The effect of self-heating is visible (with a proper bias) particularly on $|s_{12}|$ and $|s_{22}|$ [28], having a variable magnitude in this low-frequency range. The conversion to y-parameters is necessary for an easier evaluation of the effect and finally allows to extract the parameters of the recursive network. The y-parameters are defined – at $f = 0$ – as

$$y_{11} = \left.\frac{dI_B}{dV_{BE}}\right|_{V_{CE}=const} \qquad y_{21} = \left.\frac{dI_C}{dV_{BE}}\right|_{V_{CE}=const}$$

$$y_{12} = \left.\frac{dI_B}{dV_{CE}}\right|_{V_{BE}=const} \qquad y_{22} = \left.\frac{dI_C}{dV_{CE}}\right|_{V_{BE}=const}$$

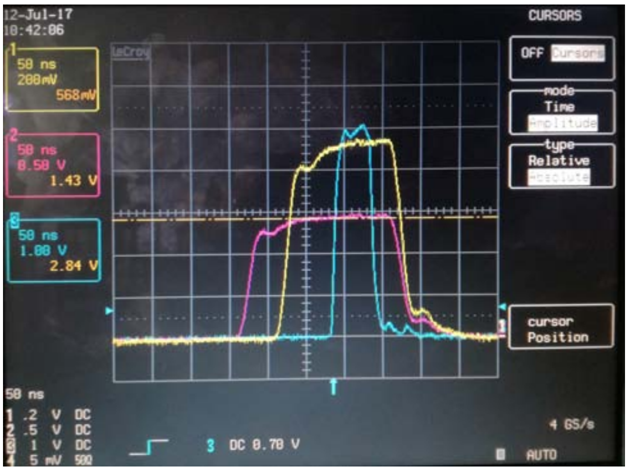


Figure 5.2: Synchronisation of DC and RF pulses. A RF stimulus (blue, $t_w = 50\,ns$) is superposed to the DC base pulse $V_{BE}(t)$ (yellow, $t_w = 150\,ns$) within the measurement window.

The most self-heating-sensitive parameters are $y_{12}(f)$ and $y_{22}(f)$ [28]; we have experienced this for y_{22} in the pulsed I/V analysis in Chapter 3.

Dynamic self-heating (modelled by the distributed thermal capacitance) is present only above a certain frequency (which, of course, corresponds to $1/2\pi\,\tau_{TH}$), up to a point where, at higher frequencies, the junction temperature variation no longer follows the instant power dissipation, i.e. the device behaviour is dominated by its electrical characteristics only ($Z_{TH}(f) \rightarrow 0$). In this range of frequencies, gradually every capacitive reactance decreases and becomes finite until eventually they become negligible.

5.2 Pulsed S-parameters, a Measurement Perspective

Pulsed S-parameter measurements are performed by superposing an RF pulse to the measurement of the pulsed-I/V (Fig. 5.2). For this kind of measurement a VNA is necessary, and a different set-up with respect to Fig. 3.3 has to be built (Fig. 5.3).

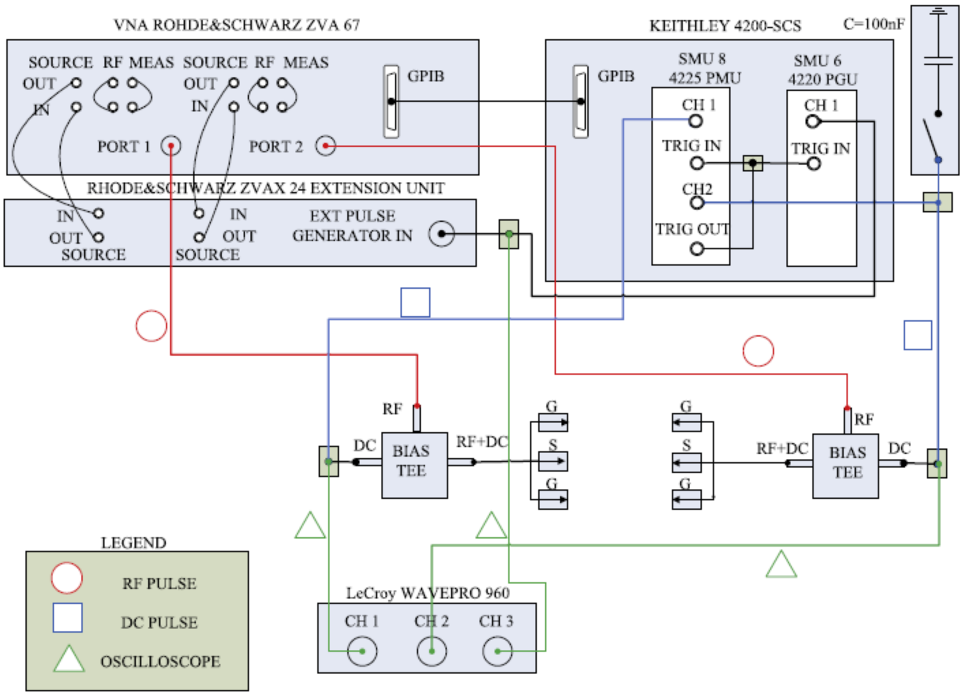


Figure 5.3: Pulsed measurement system where DC and RF pulses are applied through the DC and RF ports of two bias tees to the base and collector of the DUT (after [33]).

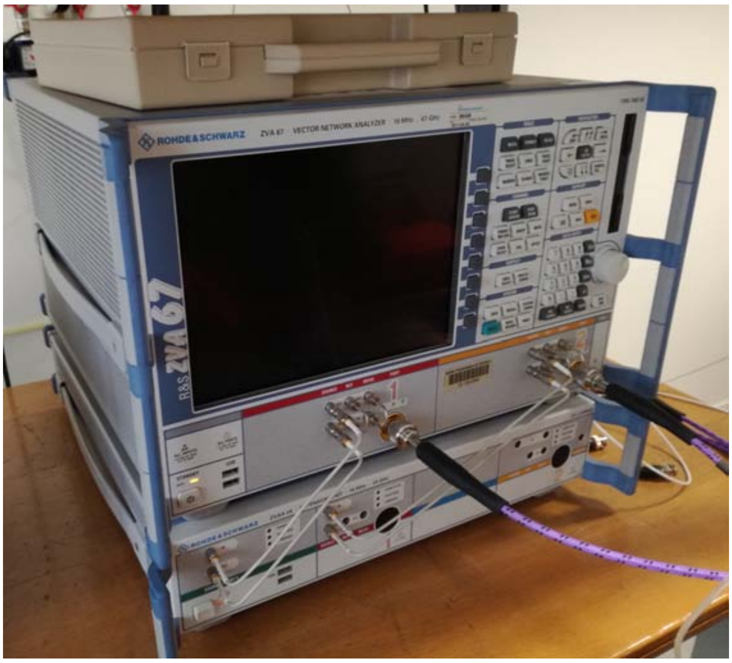


Figure 5.4: Rohde & Schwarz ZVA67 vector network analyser with ZVAX extension unit.

The Keithley 4200-SCS is used in combination with the VNA Rohde & Schwarz ZVA67 with a ZVAX extension unit (Fig. 5.4). In order to synchronize the DC and RF pulses, the 4220 PGU pulse generator unit sends a trigger to the ZVA67 and the user can verify the synchronization with a LeCroy Wavepro 960 oscilloscope (same used previously in this work). The bias tees superpose the two signals.

The duty cycle of the RF signal cannot be smaller than 1% of the total pulse period because this level might be close to the VNA noise level: a good trade-off has been reached with a width around 5% [33].

The effect of self-heating can be visualized also in the modification of the f_T curve when a bias that causes less power dissipation is applied. A technique to extract f_T from measurements is briefly described.

Once the S-parameters are extracted by our VNA, they are converted to

h-parameters. We focus on $|h_{21}(f)|$, which is defined as

$$|h_{21}(f)| = \left|\frac{i_C(f)}{i_B(f)}\right|$$

and is traced in a Bode plot. The cut-off frequency can be estimated by extrapolating the unity gain frequency. In fact, Eq. (2.11), for high frequency, can be rewritten as

$$f_T = |h_{21}(f)|\, f$$

assuming that above a certain frequency, the slope is about $20\, dB/dec$. This provides a practical extraction method.

In Fig. 5.5 we have made a linear interpolation above a given frequency ($f = 15\, GHz$) and extracted the changing of f_T on V_{BE} when a pulsed-RF signal is superposed to the base signal (see Fig. 5.6). The device we took those measures on is the reference ETHZ, at $V_{CE} = 1.25\, V$. We see that the peak is around $330\, GHz$, well below the nominal peak f_T at $V_{CE} = 1\, V$ (Fig. 2.21), which is partly normal, partly due to the imprecision of the instrumentation.

Fig. 5.7 finally shows the effect of heat on f_T on a different device. At a given high V_{BE} value and for high collector voltage, the cut-off frequency is damped down with respect to the same V_{BE} for $V_{CE} = 0.5\, V$ when increasing the pulse width.

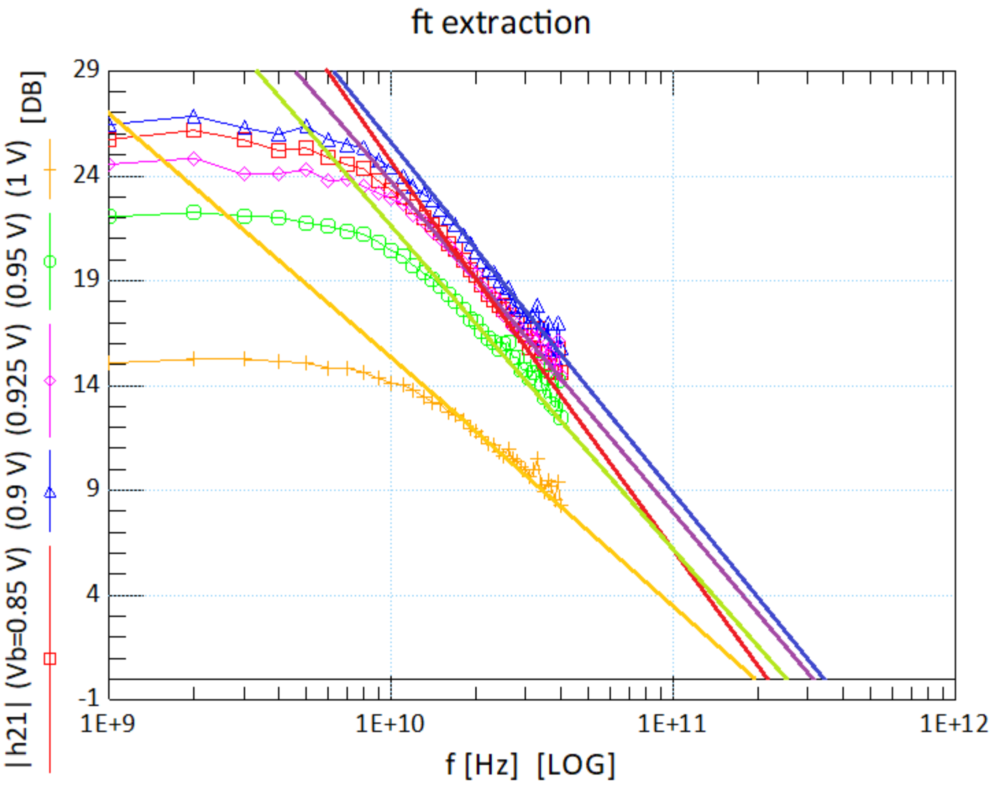


Figure 5.5: Cut-off frequency f_T extraction method at different V_{BE} for the reference ETHZ at $V_{CE} = 1.25\,V$ by pulsed S-parameters ($t_w = 150\,ns$, RF pulse: $t_w = 50\,ns$).

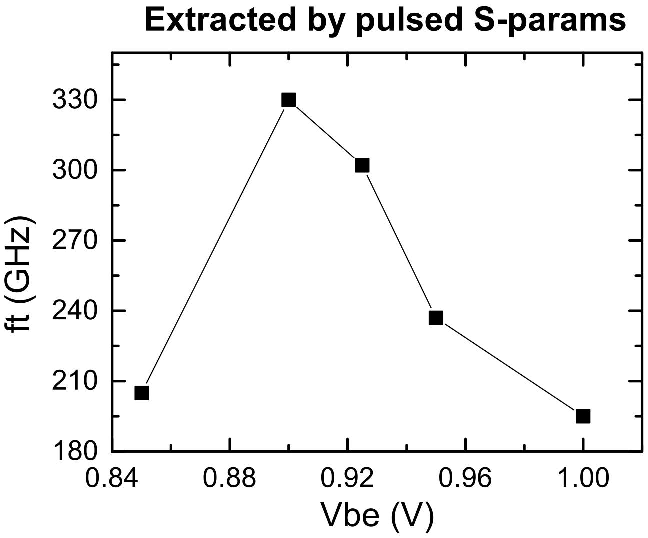


Figure 5.6: Extracted values of f_T vs. V_{BE} for the reference ETHZ at $V_{CE} = 1.25\,V$ by pulsed S-parameters ($t_w = 150\,ns$, RF pulse: $t_w = 50\,ns$).

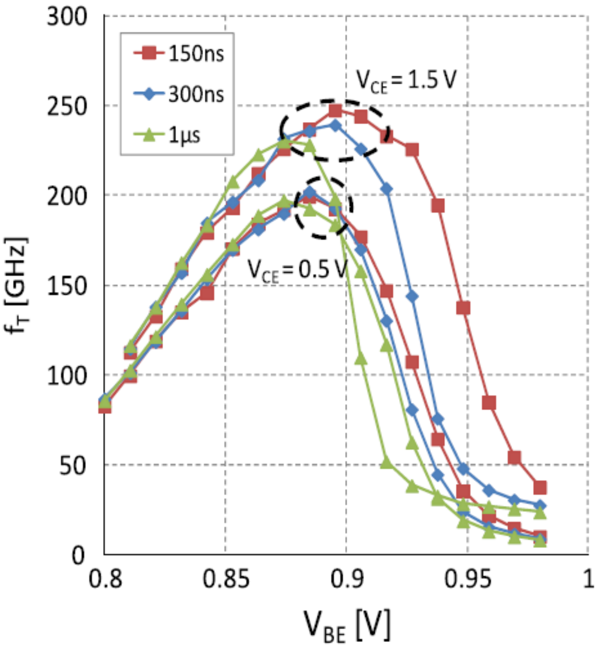


Figure 5.7: Cut-off frequency f_T vs. V_{BE} measured with $t_w = 150\,ns$, $300\,ns$, $1\,\mu s$ at constant $V_{CE} = 0.5\,V$ and $1.5\,V$ for an HBT with $A_E = 0.5 \cdot 5\,\mu m^2$ (after [33]).

Chapter 6

Conclusion and Future Work

In conclusion, in this work we have retrieved many information concerning the self-heating and the set-up for pulsed measurements. The new IC-CAP interface allows a whole set of pulsed-I/V-based analysis on many HF devices, some of which we carried out in Chapter 3, such as collector current versus pulse width and collector current versus geometry.

The comparison among technologies and geometries allowed to determine directly in IC-CAP the exact effect of a longer or wider emitter on the output current, proving a much greater impact on the current due to the width scaling. We have seen that the length scaling do not heat sensibly the device, the percentage increase above isothermal being rather constant also in the InP/InGaAs device.

The transient-I/V analysis was useful on one hand to intuitively motivate the lower bound on the applied pulse width, on the other for determining a valid approximation of the base electrical resistance, but more importantly to extract the thermal time constant for the three types of devices, and discovering that the rise of the current due to self-heating is slower for the InP-based devices, implying a better response to the internal temperature rise. This is a fundamental result for the thermal characterisation of these advanced technologies.

Later on, Chapter 4 discussed about the thermal resistance, compared the theoretical definitions and found accordance. Then the intersection method has been presented and validated by association to IC-CAP simulations. With a fully automated IC-CAP program the R_{TH} extraction at room temperature was made, as well as the tracing of its dependence on the junction temperature.

The results, although the extension of the analysis is limited by the applied strategy, found good accordance with the model parameters for the SiGe:C technology and were well-defined for the InP/InGaAs technology too.

The next steps for an easy and complete thermal description will be the implementation of an GUI interface in IC-CAP for pulsed S-parameters extraction quite similar to the existing one for pulsed-DC. Also, the extraction method might be refined by using a polynomial interpolation such as a Spine.

The challenges related to these technologies will increase in the next generation of Si-based and particularly of InP-based devices and their thermal behaviour will be one of the main factors slacking their vast and affordable commercial diffusion and the consequent creation of related applications.

Appendix

Intersection Point Extraction Algorithm

In the following, the algorithm to find the intersection point of two forward Gummel plots through a linear interpolation is presented. It is iterated over many couples of Gummel plots at different T_{amb}'s.

```
! declare the array for the difference between the Ic's
COMPLEX diff[sizeof(Ic1)]



! finds the minimum difference between the sampled values of Ic,
! ic_diff_min=0 ideally corresponding to two samples positioned at
! the intersection of IC1 and IC2
i = 0
WHILE i<sizeof(Vb)
  x = Vb[i]
  IF x > Vb_threshold THEN
    diff[i] = abs(Ic2[i]-Ic1[i])
  ELSE
    diff[i] = 1
  END IF
  i = i+1
END WHILE
```

```
ic_diff_min = MIN(diff)


! finds the Vbe corresponding to ic_diff_min, namely the closest
! Vbe value to the intersection, vb_min; the correspondent Ic values
! are also found
index_min = 0
i = 0
WHILE i<sizeof(Vb)
  x = diff[i]
  IF ic_diff_min==x THEN
    index_min=i
  END IF
  i = i+1
END WHILE
vb_min = Vb[index_min]
IC1_min = Ic1[index_min]
IC2_min = Ic2[index_min]


! finds the second closest Vbe values to the intersection and the
! correspondent Ic's
vb_min_2nd = 0
index_min_2nd = 0
IF diff[index_min-1]<diff[index_min+1] THEN
  vb_min_2nd = Vb[index_min-1]
  IC1_min_2nd = Ic1[index_min-1]
  IC2_min_2nd = Ic2[index_min-1]
ELSE
  vb_min_2nd = Vb[index_min+1]
  IC1_min_2nd = Ic1[index_min+1]
  IC2_min_2nd = Ic2[index_min+1]
END IF


! finds the parameters of the line connecting the samples around the
! intersection
IF vb_min<vb_min_2nd THEN
  q1 = IC1_min
  q2 = IC2_min
ELSE
  q1 = IC1_min_2nd
  q2 = IC2_min_2nd
END IF
s1 = abs(IC1_min-IC1_min_2nd)/(Vb[1]-Vb[0])
s2 = abs(IC2_min-IC2_min_2nd)/(Vb[1]-Vb[0])


! calculates Vbe and Ic intersection values
x = (q1-q2)/(s2-s1)
IF vb_min<vb_min_2nd THEN
  int_vb = vb_min + x
ELSE
  int_vb = vb_min_2nd + x
END IF
  int_ic = q1 + (s1*x)

Ic_intersection[t] = int_ic
```

Bibliography

[1] R. W. Keyes. Physical problems and limits in computer logic. *IEEE Spectrum*, 6(5):36–45, May 1969.

[2] IMS Laboratoire de l'Intégration du Matériau au Système. NANOCOM platform. `https://www.ims-bordeaux.fr/fr/plateformes/centrale-d-analyse-et-caracterisation/44-NANOCOM`.

[3] M. Schröter and A. Chakravorty. *Compact hierarchical modeling of bipolar transistors with HICUM.* World Scientific, Singapore, 2010.

[4] SURA. Terahertz applications symposium, 2009. `http://www.sura.org/commercialization/terahertz.html`.

[5] G. M. Rebeiz. Millimeter-wave and terahertz integrated circuit antennas. *Proceedings of the IEEE*, 80(11):1748–1770, November 1992.

[6] U. R. Pfeiffer, E. Ojefors, A. Lisauskas, and H. G. Roskos. Opportunities for silicon at mmWave and Terahertz frequencies. In *2008 IEEE Bipolar/BiCMOS Circuits and Technology Meeting*, pages 149–156, October 2008.

[7] S. M. Sze. *High-Speed Semiconductor Devices.* J. Wiley, September 1990.

[8] R. S. Muller and T. I. Kamins. *Device Electronics for Integrated Circuits.* John Wiley & Sons Inc, 2003.

[9] Der Sun Lee and J. G. Fossum. Energy-band distortion in highly doped silicon. *IEEE Transactions on Electron Devices*, 30(6):626–634, June 1983.

[10] M. Bairanzade. Understanding power transistors breakdown parameters. OnSemi application node AN1628/D, 2003. `http://www.onsemi.com/pub/Collateral/AN1628-D.PDF`.

[11] H. Kroemer. Theory of a wide-gap emitter for transistors. *Proceedings of the IRE*, 45(11):1535–1537, November 1957.

[12] O. Wada and H. Hasegawa. *InP-Based Materials and Devices: Physics and Technology.* Wiley Series in Microwave and Optical Engineering. Wiley-Interscience, 1999.

[13] D. J. Paul. Si/SiGe Heterojunction Bipolar Transistors, 2004. `http://userweb.eng.gla.ac.uk/douglas.paul/SiGe/HBT.html`.

[14] J. W. Matthews. Defects associated with the accommodation of misfit between crystals. *Journal of Vacuum Science and Technology 12, 126*, 1975.

[15] I.Z. Mitrovic, O. Buiu, S. Hall, D.M. Bagnall, and P. Ashburn. Review of SiGe HBTs on SOI. *Solid-State Electronics*, 49:1556–1567, September 2005.

[16] T. Agarwal. BiCMOS Technology: Fabrication and Applications, 2015. `https://www.elprocus.com/bicmos-technology-fabrication-and-applications/`.

[17] P. Chevalier, M. Schröter, C. R. Bolognesi, V. d'Alessandro, M. Alexandrova, J. Böck, R. Flückiger, S. Fregonese, B. Heinemann, C. Jungemann, R. Lövblom, C. Maneux, O. Ostinelli, A. Pawlak, N. Rinaldi, H. Rücker, G. Wedel, and T. Zimmer. Si/SiGe:C and InP/GaAsSb Heterojunction Bipolar Transistors for THz Applications. *Proceedings of the IEEE*, 105(6):1035–1050, June 2017.

[18] N. Kashio, K. Kurishima, Y. K. Fukai, and S. Yamahata. High-speed, high-reliability 0.5-um-emitter InP-based heterojunction bipolar transistors. *NTT Technical Review*, 7, December 2009.

[19] G. A. Koné. *Caractérisation des effets thermiques et des mécanismes de défaillance spécifiques aux transistors bipolaires submicroniques sur substrat InP dédiés aux transmissions optiques Ethernet à 112 Gb/s.* PhD thesis, Université de Bordeaux I, 2011.

[20] P. Chevalier, G. Avenier, G. Ribes, A. Montagné, E. Canderle, D. Céli, N. Derrier, C. Deglise, C. Durand, T. Quémerais, M. Buczko, D. Gloria, O. Robin, S. Petitdidier, Y. Campidelli, F. Abbate, M. Gros-Jean, L. Berthier, J. D. Chapon, F. Leverd, C. Jenny, C. Richard, O. Gourhant, C. De-Buttet, R. Beneyton, P. Maury, S. Joblot, L. Favennec, M. Guillermet, P. Brun, K. Courouble, K. Haxaire, G. Imbert, E. Gourvest, J. Cossalter, O. Saxod, C. Tavernier, F. Foussadier, B. Ramadout, R. Bianchini, C. Julien, D. Ney, J. Rosa, S. Haendler, Y. Carminati, and B. Borot. A 55 nm triple gate oxide 9 metal layers SiGe BiCMOS technology featuring 320 GHz fT / 370 GHz fMAX HBT and high-Q millimeter-wave passives. In *2014 IEEE International Electron Devices Meeting*, pages 3.9.1–3.9.3, December 2014.

[21] P. Chevalier, G. Avenier, E. Canderle, A. Montagné, G. Ribes, and V. T. Vu. Nanoscale SiGe BiCMOS technologies: From 55 nm reality to 14 nm opportunities and challenges. In *2015 IEEE Bipolar/BiCMOS Circuits and Technology Meeting - BCTM*, pages 80–87, October 2015.

[22] K. Kuhn, M. Agostinelli, S. Ahmed, S. Chambers, S. Cea, S. Christensen, P. Fischer, J. Gong, C. Kardas, T. Letson, L. Henning, A. Murthy, H. Muthali, B. Obradovic, P. Packan, S. W. Pae, I. Post, S. Putna, K. Raol, A. Roskowski, R. Soman, T. Thomas, P. Vandervoorn, M. Weiss, and I. Young. A 90 nm communication technology featuring SiGe HBT transistors, RF CMOS, precision R-L-C RF el-

ements and 1 /spl mu/m2 6-T SRAM cell. In *Digest. International Electron Devices Meeting*, pages 73–76, December 2002.

[23] V. Nodjiadjim. ANR ULTIMATE - Task 2.2, January 2017. `http://www.agence-nationale-recherche.fr/Project-ANR-16-CE93-0007`.

[24] C. Mukherjee and M. Deng. ANR ULTIMATE - III-V Lab InP HBT RF characterization up to 220 GHz, May 2017. `http://www.agence-nationale-recherche.fr/Project-ANR-16-CE93-0007`.

[25] M. Alexandrova, R. Flüeckiger, R. Lövblom, O. Ostinelli, and C. R. Bolognesi. GaAsSb-Based DHBTs With a Reduced Base Access Distance and $f_{\mathrm{T}}/f_{\mathrm{MAX}} = 503/780$ GHz. *IEEE Electron Device Letters*, 35(12):1218–1220, December 2014.

[26] D. Vasileska, K. Raleva, and S. M. Goodnick. Modeling heating effects in nanoscale devices: The present and the future. *Journal of Computational Electronics*, 7:66–93, June 2008.

[27] R. C. Joy and E. S. Schlig. Thermal properties of very fast transistors. *IEEE Transactions on Electron Devices*, 17(8):586–594, August 1970.

[28] A. K. Sahoo. *Electro-thermal Characterization, Compact Modeling and TCAD based Device Simulations of advanced SiGe:C BiCMOS HBTs and of nanometric CMOS FET*. PhD thesis, Université Bordeaux I, 2012.

[29] R. D'Esposito. *Electro-thermal Characterization, TCAD Simulations and Compact Modeling of Advanced SiGe HBTs at Device and Circuit Level*. PhD thesis, Université de Bordeaux, 2016.

[30] R. H. Winkler. Thermal properties of high-power transistors. *IEEE Transactions on Electron Devices*, 14(5):260–263, May 1967.

[31] N. Rinaldi and V. d'Alessandro. Theory of electrothermal behavior of bipolar transistors: Part I -single-finger devices. *IEEE Transactions on Electron Devices*, 52(9):2009–2021, September 2005.

[32] J. P. Teyssier, P. Bouysse, Z. Ouarch, D. Barataud, T. Peyretaillade, and R. Quere. 40-GHz/150-ns versatile pulsed measurement system for microwave transistor isothermal characterization. *IEEE Transactions on Microwave Theory and Techniques*, 46(12):2043–2052, December 1998.

[33] M. Weiss, S. Fregonese, M. Santorelli, A. Sahoo, C. Maneux, and T. Zimmer. 80ns/45GHz Pulsed measurement system for DC and RF characterization of high speed microwave devices. *Solid State Electronics*, 84, June 2013.

[34] S. Fregonese, T. Zimmer, H. Mnif, P. Baureis, and C. Maneux. Obtaining isothermal data for HBT. *IEEE Transactions on Electron Devices*, 51(7):1211–1214, July 2004.

[35] B. Schaefer and M. Dunn. Pulsed measurements and modeling for electro-thermal effects. In *Proceedings of the 1996 BIPOLAR/BiCMOS Circuits and Technology Meeting*, pages 110–117, September 1996.

[36] M. Golio and J. Golio. *RF and Microwave Circuits, Measurements, and Modeling (The RF and Microwave Handbook, Second Edition)*. CRC Press, 2007.

[37] Keithley Instruments Inc. *Model 4200-SCS Semiconductor Characterization System - Reference Manual*, 2011.

[38] R.C. Jaeger and T. N. Blablock. *Microelectronic Circuit Design*. McGraw-Hill Education, 4 edition, 2011.

[39] J. Berkner. Extraction of thermal resistance and its temperature dependance using DC methods, 2007. `https://www.iee.et.tu-dresden.de/iee/eb/forsch/Models/workshop0607/contr/Berkner_Infineon_HICUM_WS_2007_Dresden_070621s.pdf`.

[40] S. Russo, V. d'Alessandro, L. La Spina, N. Rinaldi, and L. K. Nanver. Evaluating the self-heating thermal resistance of bipolar transistors by DC measurements: A critical review and update. In *2009 IEEE Bipolar/BiCMOS Circuits and Technology Meeting*, pages 95–98, October 2009.

[41] A. Chakravorty, R. D'Esposito, S. Balanethiram, S. Frégonèse, and T. Zimmer. Analytic Estimation of Thermal Resistance in HBTs. *IEEE Transactions on Electron Devices*, 63(8):2994–2998, August 2016.

[42] V. Palankovski. Thermal conductivity, 2001. `http://www.iue.tuwien.ac.at/phd/palankovski/node34.html`.

[43] N. Rinaldi and V. d'Alessandro. Theory of electrothermal behavior of bipolar transistors: part II-two-finger devices. *IEEE Transactions on Electron Devices*, 52(9):2022–2033, September 2005.

[44] M. Pfost, V. Kubrak, and P. Brenner. A practical method to extract the thermal resistance for heterojunction bipolar transistors. In *33rd Conference on European Solid-State Device Research. ESSDERC '03.*, pages 335–338, September 2003.

[45] F. Schaffler, M.E. Levinshtein, S.L. Rumyantsev, and M.S. Shur. *Properties of Advanced Semiconductor Materials GaN, AlN, InN, BN, SiC, SiGe.* Wiley Interscience, 2001.

[46] M. Hiebel. *Fundamentals of Vector Network Analysis.* Rohde & Schwarz GmbH & Co., 2008.